

POLITECNICO DI TORINO

Corso di Laurea Magistrale in Ingegneria Informatica



**Politecnico
di Torino**

Tesi di Laurea Magistrale

Analisi dei dati sulle segnalazioni di non emergenza in ambito urbano mediante tecniche di Data Mining

Relatore

Prof.ssa Silvia Chiusano

Candidato

Rocco Collura

Correlatori

Prof. Luca Cagliero

Dot.ssa Elena Daraio

ANNO ACCADEMICO 2021-2022

Sommario

Negli anni la capacità di raccogliere grandi quantità di dati è andata sempre crescendo; il grande numero di dati raccolti costituiscono un notevole potenziale informativo. Grazie anche alla disponibilità di open data al giorno d'oggi c'è molta attenzione al concetto di smart city, ossia alla realizzazione di città intelligenti che permettano di migliorare le condizioni urbane e venire incontro alle esigenze dei cittadini. Nella gestione di una città è un punto importante la capacità di rilevare le situazioni critiche e monitorare la percezione dei cittadini sulla sicurezza urbana. A tal fine l'amministrazione cittadina o la polizia locale acquisiscono comunemente questa tipologia di segnalazioni dai cittadini. L'analisi di questo tipo di dati mediante tecniche di machine learning può fornire informazioni utili per eseguire azioni mirate e di fatto migliorare la sicurezza e la qualità delle città nella percezione del cittadino. In questo lavoro di tesi è stato definito un framework che include una prima fase di preprocessing al fine di migliorare la qualità e l'utilità dei dati, ed una successiva seconda fase di analisi dei dati per l'estrazione di informazione utile rappresentata mediante pattern di facile interpretazione. Come casi di studio di riferimento sono state considerate le segnalazioni di non emergenza legate alla sicurezza urbana in due città diverse, Torino e New York. Per la città Piemontese sono stati esplorati i dati delle segnalazioni inviate dai cittadini al Contact Center della Polizia Municipale di Torino. Per la città Americana è stato utilizzato il dataset relativo alle segnalazioni urbane denominate 311. L'obiettivo è svolgere una analisi di dati attraverso tecniche di data mining che permettano di trovare possibili correlazioni tra i tipi di segnalazione e gli attributi geografici e temporali. I risultati di queste analisi possono fornire informazioni utili da utilizzare nel contesto di una smart city per migliorare la qualità di vita dei cittadini e la percezione di sicurezza in città, ma anche contribuire a rendere il cittadino sempre più partecipe del contesto urbano in cui vive.

Indice

Elenco delle figure	v
1 Introduzione	1
2 Metodologia	3
2.1 Data Acquisition and Collection	4
2.2 Data Mining	4
2.3 Data Preprocessing	5
2.4 Association rules	6
2.4.1 FP-Growth	7
2.4.2 Flipping Correlation	8
2.5 Interpretazione dei risultati ottenuti	10
2.6 Tecnologie utilizzate	10
3 Dataset segnalazioni urbane	13
3.1 Dataset della città di Torino	13
3.2 Dataset della città di New York	14
4 Esplorazione e Preprocessing	19
4.1 Dataset Torino	19
4.2 Dataset New York City	39
4.3 Analogie e Differenze	61
5 Sezione sperimentale	63
5.1 Struttura pattern delle regole associative	63
5.2 Generazione di regole di associazione	64
5.3 Analisi regole associative estratte dalle segnalazioni di Torino	65
5.4 Analisi regole associative estratte dalle segnalazioni di New York	81
5.5 Flipping correlation	119
5.6 Flipping correlation sulle segnalazioni di Torino	120
5.7 Flipping correlation sulle segnalazioni di New York	126

6 Conclusioni e sviluppi futuri	139
Bibliografia	141

Elenco delle figure

2.1	Framework	3
2.2	Flipping Correlation	8
4.1	Segnalazioni divise per Categoria	24
4.2	Segnalazioni divise per Sottocategorie	25
4.3	Segnalazioni divise per Circoscrizione	26
4.4	Segnalazioni in Area verde	27
4.5	Segnalazioni distribuite negli anni	28
4.6	Segnalazioni distribuite nei mesi	29
4.7	Segnalazioni per giorno settimanale	30
4.8	Distribuzione delle segnalazioni nei giorni Feriali e Festivi	31
4.9	Distribuzione delle segnalazioni nelle stagioni	32
4.10	Distribuzione delle segnalazioni per ogni categoria negli anni	33
4.11	Distribuzione delle segnalazioni nelle circoscrizioni per ogni categoria	34
4.12	Distribuzione delle segnalazioni per Area Verde per ogni categoria	35
4.13	Numero segnalazioni per le sottocategorie di Qualità Urbana in Area Verde	35
4.14	Numero segnalazioni per categoria nei mesi	36
4.15	Numero segnalazioni per categoria nelle stagioni	37
4.16	Numero segnalazioni per categoria durante i giorni settimanali	38
4.17	Distribuzione delle segnalazioni per Categoria	49
4.18	Distribuzione delle segnalazioni per Complaint Type	52
4.19	Distribuzione delle segnalazioni nei Distretti	53
4.20	Distribuzione delle segnalazioni nei vari Community Board	54
4.21	Distribuzione delle segnalazioni nei vari anni	55
4.22	Distribuzione delle segnalazioni durante i mesi	56
4.23	Distribuzione delle segnalazioni durante le stagioni	56
4.24	Distribuzione delle segnalazioni durante i giorni della settimana	57
4.25	Distribuzione delle segnalazioni durante i giorni Festivi e Feriali	58
4.26	Distribuzione delle segnalazioni negli anni per categoria	58
4.27	Distribuzione delle segnalazioni nei Distretti per categoria	59

4.28	Distribuzione delle segnalazioni per categoria nelle stagioni	60
4.29	Distribuzione delle segnalazioni per categoria nei giorni settimanali	61
5.1	Regole associative Contesto spaziale → Categoria	66
5.2	Regole associative Contesto temporale → Categoria Convivenza Civile	67
5.3	Regole associative Contesto temporale → Categoria Qualità Urbana	68
5.4	Regole associative Contesto temporale e spaziale → Categoria Con- vivenza Civile	69
5.5	Regole associative Contesto temporale e spaziale → Categoria Qua- lità Urbana	70
5.6	Regole associative Sottocategoria Convivenza Civile→ contesto spaziale	72
5.7	Regole associative Sottocategoria Qualità Urbana→ contesto spaziale	73
5.8	Regole associative Sottocategoria Convivenza Civile → contesto temporale	75
5.9	Regole associative Sottocategoria Qualità Urbana → contesto tem- porale	76
5.10	Regole associative Sottocategoria Convivenza Civile → contesto spaziale e temporale	78
5.11	Regole associative sottocategoria Qualità Urbana → contesto spaziale e temporale	79
5.12	Regole associative contesto spaziale → Categoria Cleaning & Garbage	81
5.13	Regole associative contesto temporale → Categoria Cleaning & Garbage	82
5.14	Regole associative contesto temporale e spaziale → Categoria Clea- ning & Garbage	83
5.15	Regole associative contesto spaziale → Categoria Complaint	84
5.16	Regole associative contesto temporale → Categoria Complaint . . .	85
5.17	Regole associative contesto temporale e spaziale → Categoria Com- plaint	88
5.18	Regole associative contesto spaziale → Categoria Information	89
5.19	Regole associative contesto temporale e spaziale → Categoria Infor- mation	89
5.20	Regole associative contesto spaziale → Categoria Maintenance . . .	90
5.21	Regole associative contesto temporale → Categoria Maintenance . .	92
5.22	Regole associative contesto temporale e spaziale → Categoria Main- tenance	94
5.23	Regole associative contesto spaziale → Categoria Park & Trees . . .	95
5.24	Regole associative contesto temporale → Categoria Park & Trees .	95
5.25	Regole associative contesto spaziale e temporale → Categoria Park & Trees	96
5.26	Regole associative contesto spaziale → Categoria Streets & Sidewalks	96

5.27	Regole associative contesto temporale → Categoria Streets& Sidewalks	97
5.28	Regole associative contesto temporale e spaziale → Categoria Streets & Sidewalks	99
5.29	Regole associative contesto spaziale → Categoria Transportation . .	100
5.30	Regole associative contesto temporale → Categoria Transportation .	101
5.31	Regole associative contesto temporale e spaziale → Categoria Transportation	102
5.32	Regole associative sottocategoria Cleaning & Garbage → contesto spaziale	103
5.33	Regole associative sottocategoria Cleaning & Garbage → contesto temporale e spaziale	104
5.34	Regole associative sottocategoria Complaint → contesto spaziale . .	105
5.35	Regole associative sottocategoria Complaint → contesto temporale .	106
5.36	Regole associative sottocategoria Complaint → contesto temporale e spaziale	107
5.37	Regole associative sottocategoria Information → contesto spaziale .	108
5.38	Regole associative sottocategoria Information → contesto temporale	108
5.39	Regole associative sottocategoria Information → contesto temporale e spaziale	108
5.40	Regole associative sottocategoria Maintenance → contesto spaziale .	109
5.41	Regole associative sottocategoria Maintenance → contesto temporale	111
5.42	Regole associative sottocategoria Maintenance → contesto temporale e spaziale	112
5.43	Regole associative sottocategoria Park & Trees → contesto spaziale	113
5.44	Regole associative sottocategoria Park & Trees → contesto temporale	113
5.45	Regole associative sottocategoria Park & Trees → contesto temporale e spaziale	114
5.46	Regole associative sottocategoria Streets & Sidewalks → contesto spaziale	115
5.47	Regole associative sottocategoria Streets & Sidewalks → contesto temporale	115
5.48	Regole associative sottocategoria Streets & Sidewalks → contesto temporale e spaziale	116
5.49	Regole associative sottocategoria Transportation → contesto spaziale	117
5.50	Regole associative sottocategoria Transportaion → contesto temporale	117
5.51	Regole associative sottocategoria Transportaion → contesto temporale e spaziale	118
5.52	Correlazione Flipping Allarme Sociale e sottocategoria con anno . .	120
5.53	Correlazione Flipping Convivenza Civile e sottocategoria con anno .	121
5.54	Correlazione Flipping Qualità Urbana e sottocategoria con anno . .	121
5.55	Correlazione Flipping Convivenza Civile e sottocategoria con Stagione	122

5.56	Correlazioni Flipping	123
5.57	Correlazioni Flipping Categoria e sottocategoria con Circoscrizione .	123
5.58	Correlazioni Flipping Categoria con Anno e mese	123
5.59	Correlazioni Flipping Sottocategoria con Anno e stagione	124
5.60	Correlazioni Flipping Sottocategoria con Stagione e Mese	125
5.61	Correlazioni Flipping, Sottocategoria con Mese e Giorno	125
5.62	Correlazioni Flipping su 3 livelli	126
5.63	Correlazioni Flipping Categoria e sottocategoria con Anno	127
5.64	Correlazioni Flipping, Categoria e sottocategoria con Stagione . . .	127
5.65	Correlazioni Flipping, son Categoria, sottocategoria e Mese	128
5.66	Correlazioni Flipping, Categoria e sottocategoria con Borough . . .	129
5.67	Correlazioni Flipping, Categoria e sottocategoria con Community Board	129
5.68	Correlazioni Flipping, Categoria e sottocategoria con Luogo	130
5.69	Correlazioni Flipping, Categoria con Stagione e Mese	131
5.70	Correlazioni Flipping, Categoria con Borough e Community Board .	131
5.71	Correlazioni Flipping, Categoria con Borough e Location Type . . .	132
5.72	Correlazioni Flipping, Sottocategoria con stagione e mese	133
5.73	Correlazioni Flipping, Sottocategoria con Borough e Community Board	134
5.74	Correlazioni Flipping, Sottocategoria con Borough e Location Type	135
5.75	Correlazioni Flipping, Categoria con Borough, Community Board e Location Type	136
5.76	Correlazioni Flipping, Sottocategoria con Borough, Community Board e Location Type	137

Capitolo 1

Introduzione

La tecnologia ha acquisito notevole importanza nella vita dell'uomo, dandoci la possibilità di rendere tutto più intelligente e più semplice. Anche in ambito urbano si sfruttano le nuove tecnologie per migliorare le prestazioni dei servizi sanitari, dei trasporti, dell'energia, dell'istruzione portando il comfort dei cittadini a livelli più elevati. L'obiettivo è quello di creare città intelligenti tecnologicamente con il fine di migliorare le condizioni urbane e le esigenze dei cittadini. Il lavoro svolto utilizza open data in ambito urbano da poter utilizzare sfruttando diverse tecniche di data mining con il fine di fornire informazioni utili per migliorare la qualità della vita urbana dei cittadini, infatti nella gestione della città diventa importante rilevare le situazioni critiche e la percezione dei cittadini sulla sicurezza urbana. Analizzare questo tipo di dati può aiutare certamente ad eseguire azioni mirate e di fatto migliorare la sicurezza e la qualità delle città, ad esempio con l'individuazione di zone a rischio e/o pericolose. La metodologia che nell'ambito Data Science permette di sfruttare il potenziale informativo dei dataset fa capo a quello che viene chiamato Data Mining. Ci sono diverse tecniche, racchiuse sotto il termine di Data Mining, che si pongono l'obiettivo di estrapolare informazioni utili implicitamente contenute nei dati presentandole poi sotto forma di pattern utili e di facile comprensione. In questo lavoro di tesi è stata definita una metodologia di lavoro ben precisa che include una prima fase di preprocessing al fine di migliorare la qualità e l'utilità dei dati, ed una successiva seconda fase di analisi dei dati per l'estrazione di informazione utile rappresentata mediante pattern di facile interpretazione. Partendo da questa metodologia il presente lavoro di tesi utilizza degli open data relativi a segnalazioni di non emergenza legati alla sicurezza urbana in due contesti diversi. Una città italiana, Torino, di cui sono stati esplorati i dati delle segnalazioni inviate dai cittadini al Contact Center della Polizia Municipale di Torino, ed una città americana, New York, di cui è stato utilizzato il dataset relativo alle segnalazioni urbane denominate 311. Questi dati sulla sicurezza urbana sono stati esplorati, puliti, riorganizzati e arricchiti e sono stati confrontati tra loro;

si è evidenziato cosa li rende particolari, cosa li accomuna, cosa li differisce. Ogni segnalazione è associata ad una categoria e sottocategoria specifica, ed inoltre sono presenti attributi geografici legati alla città ed attributi temporali legati al contesto temporale delle segnalazioni fatte dai cittadini. L'obiettivo è quello di svolgere una analisi di dati attraverso tecniche diverse di data mining che permettano di trovare possibili correlazioni interessanti tra i tipi di segnalazione e gli attributi geografici e temporali, in modo da poter estrarre informazioni utili e fare delle opportune valutazioni, cercando di poter rispondere a domande del tipo "Dato un luogo, cosa succede in quel luogo? Cosa è accaduto in un determinato periodo di tempo in quel luogo?". La tecnica di Data Mining ampiamente diffusa ed usata in diversi domini e anche in questo lavoro è l'estrazione delle regole associative che permette di evidenziare relazioni interessanti tra le variabili all'interno di un dataset.

La struttura di questa tesi è la seguente:

- Capitolo 2 introduce la metodologia usata e le librerie utilizzate durante l'analisi
- Capitolo 3 fornisce una descrizione sul contenuto dei dataset utilizzati
- Capitolo 4 descrive la fase di preprocessing ed esplorazione dei dati
- Capitolo 5 mostra i risultati ottenuti dalle due tecniche utilizzate e il relativo contenuto informativo estratto
- Capitolo 6 definisce le conclusioni del seguente lavoro e suggerisce eventuali sviluppi futuri

Capitolo 2

Metodologia

L'approccio metodologico relativo all'analisi dei dati usato in questo progetto di tesi verrà descritto in questo capitolo. Tipicamente il processo di analisi dei dati per l'estrazione della conoscenza dai dataset (*Knowledge Discovery in Databases*) è formato da diverse fasi che analizzeremo meglio nei successivi paragrafi. La metodologia usata è rappresentata nella figura 2.1.

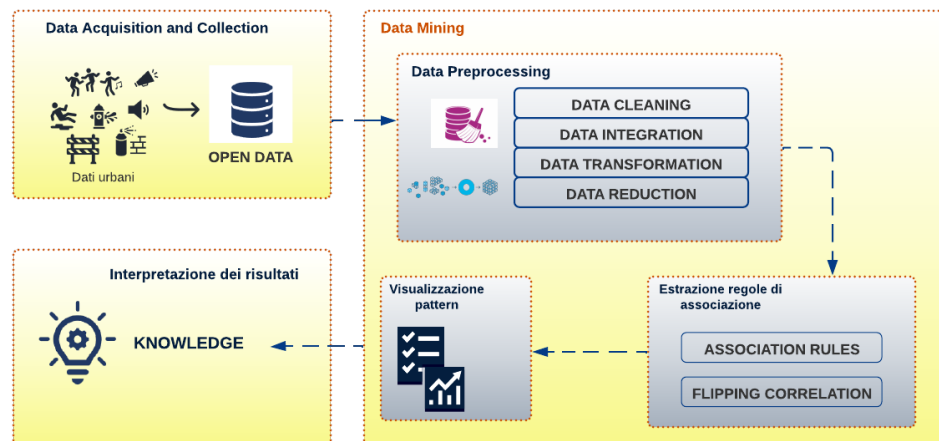


Figura 2.1: Framework

Essa è definita dai seguenti step:

- **Data Acquisition and Collection:** le infinite quantità di open data disponibili rappresentano una potente fonte di conoscenza e il seguente processo di analisi dei dati ci permette attraverso diversi strumenti di scoprire informazioni non note a priori.

- **Data preprocessing:** fase in cui si effettua il controllo dell'integrità e della coerenza dei dati applicando opportuni accorgimenti al fine di migliorare la qualità e l'utilità dei dati.
- **Estrazione di regole di associazione:** si selezionano tecniche di Data Mining per l'estrazione di conoscenza dai dati.
- **Visualizzazione dei pattern:** l'informazione utile è estratta sottoforma di pattern di facile interpretazione.
- **Interpretazione dei risultati:** step finale in cui vengono interpretati i risultati ottenuti al fine di fornire informazioni utili da poter utilizzare.

2.1 Data Acquisition and Collection

La fase di acquisizione dei dati è la fase di raccolta di dati di qualsiasi genere che porta alla formazione di dataset, spesso di tipo open, a cui attingere per poter estrarre un potenziale informativo. La crescita esponenziale di dati memorizzati ha generato negli anni lo sviluppo di strumenti di analisi in modo da poter estrarre conoscenza non nota a priori.

2.2 Data Mining

Il *Data Mining* è l'insieme di tecniche e metodi che ha come obiettivo l'estrazione di informazioni utili implicitamente contenute all'interno di grandi quantità di dati. L'informazione estratta è rappresentata sottoforma di pattern di facile interpretazione.

Le attività di data mining sono principalmente di due tipi:

- **Predizione:** si utilizzano alcune variabili note per predire il valore incognito o futuro di altre variabili. Per la predizione dei dati si usano le tecniche di classificazione
- **Descrizione:** il fine è di trovare modelli interpretabili che descrivano i dati. Per la descrizione dei dati si usano principalmente tecniche di clustering e regole di associazione.

In base al tipo di obiettivo da raggiungere vi sono diverse tecniche principali di data mining da poter utilizzare:

- **Classificazione:** permette di individuare la classe di appartenenza di un dato record nel modo più accurato possibile.

- **Clustering:** tecniche che individuano gruppi simili tra di loro all'interno di un insieme di dati.
- **Association rules:** tecniche che sono utilizzate per trovare correlazioni all'interno di un insieme di dati.

Ci sono anche altre tecniche di Data Mining: regressione, serie temporali, sequence discovery.

2.3 Data Preprocessing

Prima di applicare le tecniche specifiche bisogna effettuare una attenta esplorazione del dataset. Infatti uno step importante ed essenziale nel Data Mining è la fase di *preprocessing* dei dati. I dataset dovrebbero essere analizzati scrupolosamente per capire in primis che tipo di attributi abbiamo in nostro possesso, successivamente questi dovrebbero essere adeguatamente puliti, trattando in maniera opportuna i dati mancanti e anomali, eliminando eventuali valori errati o non utili. Se necessario bisogna anche arricchire i dati con informazioni che potrebbero risultare utili alla nostra analisi. Tutte queste operazioni sono un passo importante nel processo di *data analysis* poiché non possiamo lavorare con dati grezzi e soprattutto questi step fondamentali permettono di migliorare la qualità e l'utilità dei dati in nostro possesso per ottenere risultati più accurati. Sono 4 le principali attività di preprocessing svolte:

- Pulizia dei dati (Data Cleaning)
- Integrazione dei dati (Data Integration)
- Trasformazione dei dati (Data Transformation)
- Riduzione dei dati (Data Reduction)

La pulizia dei dati permette di gestire i dati errati, i dati incompleti e imprecisi, e di gestire eventuali valori mancanti o di risolvere incongruenze nei dati. Il motivo alla base della pulizia dei dati è avere informazioni complete e accurate.

Una volta eseguita una opportuna e attenta pulizia dei dati, è bene verificare la possibilità di integrare tali dati ad altre fonti. Dataset con informazioni simili possono essere uniti, facendo attenzione ad eventuali ridondanze, o è possibile integrare alcuni attributi provenienti da dataset differenti. Inoltre è necessario anche effettuare una strategia di trasformazione dei dati in modo da consolidare i dati di qualità modificandone il valore, la struttura o il formato. Vengono create nuove proprietà dei dati da attributi esistenti in modo da facilitare il processo di data mining. Può essere eseguita l'operazione chiamata discretizzazione per sostituire

i valori di un attributo numerico con livelli di intervallo o livelli concettuali, ad esempio, invece di specificare l'ora, possiamo impostare una fascia oraria (mattino, pomeriggio, sera, notte). Altre possibili operazioni sono l'aggregazione tra due o più attributi in un singolo attributo o la trasformazione dei dati in informazioni di livello superiore. Molto importante è anche lo step di riduzione dei dati che permette di diminuire la quantità dei dati, si punta a ridurre il volume mantenendone la propria integrità. A questa fase può essere associato anche il processo di selezione delle caratteristiche o attributi più utili alla propria analisi. Comprendere come i dati possono essere utilizzati al meglio, ed ottimizzarli per migliorare il bagaglio informativo da estrarre è di fondamentale importanza per trarne vantaggio ai fini del risultato finale dell'analisi.

2.4 Association rules

Le regole di associazione sono una tecnica di *Data Mining* che può essere applicata a qualsiasi tipo di dato e permette di trovare relazioni nascoste tra i dati [16].

Un classico esempio è quello di trovare associazioni analizzando le vendite di un supermercato per capire quali prodotti vengono solitamente comprati insieme così da prevedere il comportamento dei clienti. Una regola di associazione è rappresentata nella seguente forma: $\mathbf{X} \rightarrow \mathbf{Y}$, dove X e Y sono insiemi di elementi disgiunti. X viene detto antecedente e Y conseguente, noti anche come *itemset*.

Se per esempio prendiamo in considerazione un dataset relativo alle vendite in una cartoleria ed estraiamo la seguente regola: $\{ \mathbf{Matita} \} \rightarrow \{ \mathbf{Gomma} \}$ non significa altro che l'acquisto di una gomma è strettamente correlato a quello di una matita.

Un importante concetto in riferimento agli itemset è il *support count*, che indica il numero di transazioni che contengono un dato itemset.

La "forza" di una regola è misurata da 3 parametri principali:

- **Supporto:** fornisce un'indicazione della frequenza con cui un determinato itemset appare nel database. In termini matematici il supporto è la frazione delle transazioni che contengono sia X che Y.

$$Supporto(\{X\} \rightarrow \{Y\}) = \frac{\text{Transazioni contenenti sia X che Y}}{\text{Numero totale di transazioni}} \quad (2.1)$$

Un itemset è detto frequente se ha un supporto maggiore o uguale ad una soglia fissata (*minsup*).

- **Confidence:** indica il numero di volte in cui una determinata regola risulta essere vera. In termini matematici indica la frequenza di Y nelle transazioni

che contengono X.

$$Confidence(\{X\} \rightarrow \{Y\}) = \frac{\text{Transazioni contenenti sia X che Y}}{\text{Transazioni contenenti X}} \quad (2.2)$$

- **Lift:** Indica in che misura X e Y sono correlate. Se il valore del parametro lift è negativo, esiste una correlazione negativa; se il valore è positivo, c'è una correlazione positiva e se il rapporto è uguale a 1, non c'è correlazione.

$$Lift(\{X\} \rightarrow \{Y\}) = \frac{Conf(X \rightarrow Y)}{Support(Y)} \quad (2.3)$$

L'uso di queste misure aiutano a valutare correttamente una determinata regola. Il passaggio successivo consiste nel capire come generare regole dall'intero set di dati e identificare quelle più importanti.

La strategia adottata per l'estrazione delle regole di associazione consiste principalmente in due attività:

- **Generazione itemset frequenti:** ci sono diversi algoritmi che permettono di trovare gli itemset frequenti in funzione del valore di soglia *minsup* scelto.
- **Generazione di regole:** si estraggono tutte le regole di associazione dagli itemset frequenti ottenuti dallo step precedente, filtrando per esempio anche per un valore di confidence superiore ad una soglia scelta.

2.4.1 FP-Growth

FP-Growth è un algoritmo utilizzato per l'estrazione di itemset frequenti. Questo algoritmo utilizza una struttura dati ad albero chiamata FP-tree da cui estrae gli itemset frequenti. Inizialmente, l'albero FP contiene solo il nodo radice rappresentato dal simbolo null. Ogni nodo dell'albero FP rappresenterà un elemento dell'insieme di elementi ed un contatore con il numero di transazioni mappate sul dato percorso. L'associazione dei nodi è l'insieme di elementi con l'ordine mantenuto durante la formazione dell'albero. La costruzione dell'algoritmo avviene nel modo seguente:

1. Il dataset viene scansionato per determinare il *support count* di ogni elemento e gli item che sono al di sotto del supporto minimo vengono rimossi.
2. Il secondo passaggio consiste nella scansione dei dati per costruire l'albero FP ed ordinare gli item in ordine di supporto decrescente.
3. Il passaggio successivo esamina le transazioni. Per ogni transazione si scopre il set di elementi in essa contenuto, e si ordina con supporto decrescente e

si inserisce la transazione nell'albero FP, se esiste già l'elemento nel ramo, condivide lo stesso nodo. Inoltre, il conteggio dell'item viene incrementato man mano che si verifica nelle transazioni. Durante questa procedura viene costruita la tabella di intestazione che permette di trovare l'occorrenza dell'elemento sull'albero in poco tempo senza attraversare l'albero.

4. Il prossimo step è l'estrazione dell'albero FP creato. Per primo viene scansionato il nodo con il supporto più basso insieme agli elementi che lo precedono, e lo stesso viene fatto per ogni elemento, così facendo si ottengono alberi FP condizionali, infine si estraggono in maniera ricorsiva gli itemset frequenti dagli alberi condizionali.

2.4.2 Flipping Correlation

Un set di item può essere modellato come un albero gerarchico, partendo da questa asserzione, l'algoritmo FLIPPER [2] permette di estrarre correlazioni di tipo flipping, cioè correlazioni contrastanti (da positive a negative o viceversa) a differenti livelli di astrazione nella tassonomia in set di dati con supporto medio-basso. Infatti potrebbe risultare interessante osservare come la correlazione può cambiare a diversi livelli di astrazione, vedendo alla fine che la correlazione tra alcuni elementi può essere positiva, mentre la correlazione tra alcuni sottotipi può essere negativa o viceversa. Per esempio se prendiamo un dataset relativo a diversi generi di film creando una gerarchia (genere film, titolo film) si potrebbe ottenere una correlazione negativa al primo livello gerarchico, tra genere romantico e genere western, ma scendendo di livello si potrebbe ottenere una correlazione positiva tra 2 film di generi precedentemente correlati negativamente.

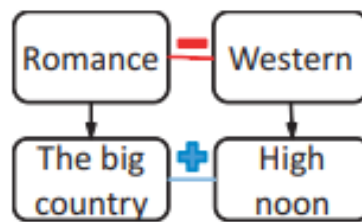


Figura 2.2: Flipping Correlation

Questo tipo di algoritmo può essere molto utile non solo nel vedere questo particolare tipo di correlazioni non banali ma anche per trovare possibili dati assegnati a categorie errate o per scoprire correlazioni di dati sottorappresentati o sovrarappresentati e per ultimo per estrarre relazioni specifiche per alcune sottocategorie. Il

ruolo del flipper è di trovare forti correlazioni contrastanti in itemset con supporto medio-basso, dato che le attuali tecniche non possono gestire la scoperta di pattern in questo intervallo di valori, gli autori di questo algoritmo trovano le correlazioni flipping senza generare tutti gli itemset frequenti ma sviluppando nuovi metodi efficienti di pruning ad hoc. Come prima cosa per affrontare il problema di trovare correlazioni di tipo flipping è fondamentale adottare una metrica di correlazione affidabile al tipo di problema e che non si preoccupi del numero di *null transaction* presenti nel dataset (si intendono le transazioni che non contengono item con una data correlazione misurata). Per tale motivo la metrica di correlazione utilizzata in questo algoritmo è la misura di *Kulczynsky* che fa parte delle misure *null invariant* che hanno la proprietà di essere indipendenti dal numero totale di transazioni del dataset. In termini matematici la correlazione di *Kulczynsky* tra gli item in un itemset A può assumere valori da 0 a 1, ed è calcolata come:

$$Kulc(A) = \frac{1}{k} \sum_{i=1}^k \frac{sup(A)}{sup(a_i)} \quad (2.4)$$

Definizione 1 *Correlazione null-invariant.* Due item sono correlati positivamente se la metrica di correlazione scelta è maggiore di una soglia γ , mentre sono negativamente correlati se tale metrica è minore o uguale ad una soglia ϵ

Per comodità, l'itemset in cui gli item sono correlati positivamente è chiamato itemset positivo, mentre se gli item sono correlati negativamente è chiamato itemset negativo.

Non possono essere estratte correlazioni flipping post elaborando tutte le correlazioni positive e negative poiché non possiamo calcolare tutte le correlazioni positive e negative, la soluzione è di inserire i vincoli di flipping in un processo di mining attraverso tecniche di potatura. L'algoritmo Flipper si affida ad alcune proprietà matematiche comuni a tutte le misure di correlazione null-invariant per costituire la base di un efficiente algoritmo di flipping pattern mining che consiste in tre principali tecniche di potatura:

- Potatura di insiemi di oggetti non flipping: se sia l'itemset dei padri (a, b) che l'itemset dei figli (a_1, b_2) hanno lo stesso segno di correlazioni, allora i figli non possono far parte del modello flipping. Tuttavia, gli itemset figli (a_{12}, b_{12}) possono ancora far parte di un pattern flipping, poiché non possiamo ancora prevedere il valore di correlazione.
- Terminazione dell'intero growth pattern: si dimostra che per qualsiasi misura di correlazione null-invariant, la correlazione dell'itemset preso in considerazione non può essere maggiore del massimo delle correlazioni dei suoi sottoinsiemi. Se aggiungiamo elementi all'itemset e scopriamo che tutte le correlazioni nelle

due celle verticalmente consecutive sono negative allora non ci sono più flipping da estrarre e possiamo interrompere la ricerca.

- Potatura di singoli item e loro superset: se tutti gli item contenenti l'elemento a_1 non sono positivi e tutti gli itemset contenenti il relativo item a non sono positivi, l'item a_1 e tutti i suoi superset possono essere potati senza ulteriori considerazioni

Per utilizzare questi principi di potatura, i risultati si confrontano per due celle verticalmente consecutive.

Gli input dell'algoritmo sono:

- un set di transazioni
- il rispettivo albero di tassonomia
- i due valori soglia γ e ϵ
- una sequenza decrescente di supporti per ogni livello gerarchico

L'output sarà dato da tutte le correlazioni di tipo flipping estratte dall'algoritmo e che soddisfano le soglie definite in input.

2.5 Interpretazione dei risultati ottenuti

In questa fase si valutano e interpretano i risultati estratti dalle tecniche di Data Mining utilizzate, mettendo in risalto le regole di associazione ritenute di fondamentale importanza e interesse sia dal punto di vista delle metriche che del contenuto. Il fine di questa fase finale dell'analisi è quello di concludere che dai risultati si possono estrarre informazioni utili e di mettere in risalto le considerazioni più interessanti.

2.6 Tecnologie utilizzate

In questo paragrafo vengono descritte le tecnologie utilizzate durante l'analisi. Per il progetto è stata utilizzata la piattaforma Jupyter Notebook ed il linguaggio di programmazione Python con le relative librerie.

Librerie Python per l'analisi dei dati

Sono state utilizzate le seguenti librerie Python per l'analisi e la manipolazione dei dati:

- **Pandas** [13]: libreria che fornisce strumenti per l'analisi e la manipolazione dei dati. In particolare, di questa libreria è stata utilizzata la struttura dati DataFrame che permette di importare i dati da diversi formati, quali il CSV e permette di effettuare operazioni di preprocessing dei dati.
- **NumPy** [10]: libreria che fornisce funzioni matematiche per lavorare con matrici e vettori.
- **Matplotlib** [8]: libreria che permette la creazione di grafici statici e interattivi in Python.
- **Seaborn** [15]: libreria di visualizzazione dati Python basata su matplotlib. Fornisce un'interfaccia di alto livello per disegnare grafici statistici interessanti e informativi.
- **Datetime** [4]: libreria che fornisce classi per la manipolazione di date e orari.
- **MLxtend** [9]: libreria che contiene implementazioni di algoritmi di Machine Learning, tra cui l'implementazione di algoritmi per l'estrazione delle regole di associazione.
- **Itertools** [7]: libreria che fornisce varie funzioni sugli iteratori per lavorare con sequenze di dati per produrre iteratori complessi.

Per l'utilizzo dell'algoritmo FLIPPER che estrae le correlazioni di tipo flipping è stata usata l'applicazione presente in [5] ideata in linguaggio Java.

Capitolo 3

Dataset segnalazioni urbane

In questo capitolo faremo una prima esplorazione degli open data utilizzati in questo lavoro di tesi. Descriveremo il contenuto dei dataset utilizzati e la loro struttura.

3.1 Dataset della città di Torino

Per il seguente progetto di tesi sono stati utilizzati i dataset disponibili sul sito AperTO [1] in cui sono presenti diversi open data relativi alla città di Torino e tra questi si possono trovare i dataset sulle segnalazioni cittadine fatte al Contact Center della Polizia Municipale della città di Torino. Questa tipologia di segnalazioni sono dette di non emergenza perché relative a delle violazioni di non immediato intervento ma legate comunque alla sicurezza urbana. Per esempio, i cittadini possono segnalare il rumore proveniente a causa della presenza di un gruppo di giovani che alzano la voce di notte oppure veicoli abbandonati in alcuni tratti di strada della città. Queste segnalazioni sono memorizzate per gli anni che vanno dal 2012 al 2019 e possono presentarsi contenute in un dataset unico per un determinato anno oppure in due dataset semestrali per anno.

I dataset sono tutti organizzati allo stesso modo e sono costituiti da 7 attributi differenti:

- **Categoria criminologa:** indica la categoria associata al tipo di segnalazione. Le tre possibili categorie sono: *Allarme Sociale*, *Convivenza Civile*, *Qualità Urbana*
- **Sottocategoria Criminologica:** indica la sottocategoria associata alla rispettiva categoria. Per la categoria *Allarme Sociale* le possibili sottocategorie sono:
 - *Atti di vandalismo*

- *Altro*

Per la categoria *Convivenza Civile* le possibili sottocategorie sono:

- *Aggregazioni Giovanili*
- *Comportamenti molesti*
- *Disturbi animali*
- *Disturbi cani*
- *Disturbo da locali*
- *Rumori molesti*
- *Uso improprio parti comuni*
- *Altro*

Per la categoria *Qualità Urbana* le possibili sottocategorie sono:

- *Decoro e degrado urbano*
- *Veicoli abbandonati*
- *Altro*

- **Circoscrizione:** indica la circoscrizione della città in cui è avvenuta la segnalazione. Nella città di Torino le circoscrizioni sono dieci e sono indicate attraverso un numero che va da 1 a 10.
- **Località:** indica la via della città in cui è avvenuta la segnalazione.
- **Area Verde:** indica se la segnalazione è avvenuta in un'area verde o meno.
- **Data:** indica la data della segnalazione nel formato gg/mm/aaaa.
- **Ora:** indica l'ora in cui è stata fatta la segnalazione nel formato hh.mm.

3.2 Dataset della città di New York

Se spostiamo l'attenzione fuori dai confini italiani troviamo in diverse città Americane molta attenzione a questo tipologia di dati. Negli Stati Uniti un certo numero di città hanno linee telefoniche governative non di emergenza, accessibili al numero di telefono 311. Di conseguenza, "311" non è solo una semplice linea telefonica ma è diventato l'intero processo di gestione delle richieste di servizio da parte dei cittadini su tutta una serie di questioni non urgenti, dalla spazzatura ai vicini rumorosi, anche se purtroppo non vi è nessuno standard di riferimento ed ogni città utilizza una diversa convenzione di codifica di reclamo/richiesta. Per questa

analisi dei dati si è scelto quindi di usare gli open data relativi alle richieste di servizio 311 della città di New York. Questo tipo di richieste sono segnalazioni sulla sicurezza urbana fatte dai cittadini per richiedere assistenza o risolvere un problema non particolarmente urgente. Rispetto al tipo di segnalazioni in Italia, nel contesto americano sono accettate oltre 500 tipi di segnalazioni, tra queste figurano segnalazioni per riparare una buca in una data zona della città, problematiche legate al riscaldamento in un appartamento o anche rimborsi per un biglietto del parcheggio pagato in eccesso. Sicuramente racchiudono una serie di richieste di servizio che coprono più contesti urbani rispetto a quello della città di Torino e successivamente esamineremo nel dettaglio le differenze facendo gli opportuni confronti. Queste segnalazioni sono open data che possiamo trovare sul sito NYC Open Data [11] per gli anni che vanno dal 2010 ad oggi, infatti queste segnalazioni sono aggiornate quotidianamente. I dati sono tutti organizzati in unico dataset continuamente aggiornato, strutturato in 41 attributi differenti:

- **Unique key:** identificatore univoco del numero di richiesta.
- **Created Date:** indica la data (compresa di ora) in cui è stata fatta la segnalazione.
- **Closed Date:** indica la data di chiusura della segnalazione.
- **Agency:** indica l'acronimo dell'agenzia di governo che ha ricevuto la segnalazione.
- **Agency Name:** indica il nome completo dell'agenzia di governo.
- **Complaint Type:** indica il primo livello gerarchico che identifica il tipo di segnalazione.
- **Descriptor:** descrive la segnalazione fornendo dettagli maggiori.
- **Location Type:** indica il tipo di luogo in cui è avvenuta la richiesta di segnalazione.
- **Incident Zip:** indica lo zip code legato al luogo della segnalazione.
- **Incident Address:** indica l'indirizzo completo del luogo in cui c'è la segnalazione.
- **Street Name:** indica il nome della via in cui è avvenuta la segnalazione.
- **Cross Street 1 e Cross Street 2:** sono due attributi che indicano le due strade trasversali rispetto alla strada della segnalazione.

- **Intersection Street 1 e Intersection Street 2:** sono due attributi che indicano le strade che incrociano la strada in cui è avvenuta la segnalazione.
- **Address Type:** indica il tipo di informazioni disponibili sulla posizione della segnalazione.
- **City:** indica la città della segnalazione.
- **Landmark:** indica l'eventuale punto di riferimento, il nome del punto di riferimento del luogo della segnalazione.
- **Facility Type:** se disponibile, questo campo indica il tipo di struttura cittadina associata alla segnalazione.
- **Status:** indica lo stato della segnalazione.
- **Due Date:** indica la possibile data in cui ci si aspetta che l'agenzia risponda alla segnalazione fatta.
- **Resolution Description:** descrive l'ultima azione intrapresa sulla segnalazione dall'agenzia che risponde. Può descrivere i passi successivi.
- **Resolution Action Updated Date:** data dell'ultimo aggiornamento sulla segnalazione.
- **Community Board:** indica il community board della città di NY.
- **BBL:** serve per identificare la posizione di edifici e proprietà a NY.
- **Borough:** indica il distretto di NY in cui è avvenuta la segnalazione.
- **X Coordinate e Y Coordinate:** sono due attributi che indicano le coordinate legate al luogo della segnalazione.
- **Open Data Channel Type:** indica la modalità di presentazione della segnalazione.
- **Park Facility Name:** indica l'eventuale nome del parco associato alla segnalazione.
- **Park Borough:** indica il distretto in cui si trova l'eventuale parco associato alla segnalazione.
- **Vehicle Type:** se la segnalazione è associata ad un incidente questo attributo serve ad indicare il tipo di veicolo coinvolto.

- **Taxi Company Borough:** se l'incidente è identificato come un taxi, questo campo mostrerà il distretto della compagnia di taxi.
- **Taxi Pick Up Location:** se l'incidente è identificato come un taxi, questo campo visualizza il luogo di prelievo del taxi.
- **Bridge Highway Name:** se l'incidente è identificato come un Ponte/Autostrada, viene in questo campo indicato il nome.
- **Bridge Highway Direction:** se l'incidente è identificato come un Ponte/Autostrada, la direzione in cui ha avuto luogo verrebbe visualizzata qui.
- **Road Ramp:** Se il luogo dell'incidente era Ponte/Autostrada questa colonna indica se la segnalazione era sulla strada o sulla rampa.
- **Bridge Highway Segment:** ulteriori informazioni sulla sezione del Ponte/Autostrada dove è avvenuto l'incidente.
- **Latitude:** indica la latitudine della segnalazione.
- **Longitude:** indica la longitudine della segnalazione.
- **Location:** combinazione della latitudine e della longitudine.

Capitolo 4

Esplorazione e Preprocessing

In questo capitolo mostreremo l'importanza della fase di esplorazione e di preprocessing durante una analisi dei dati. Prima di iniziare l'analisi dei dati è di fondamentale importanza svolgere in primo luogo una esplorazione dei dati contenuti all'interno dei dataset per capire con che tipo di dati svolgeremo la nostra analisi e successivamente valuteremo l'applicazione delle opportune tecniche di preprocessing ai dataset usati in questo lavoro di tesi al fine di migliorare la qualità e l'utilità dei dati in nostro possesso.

4.1 Dataset Torino

Come abbiamo visto nel precedente capitolo il dataset sulle segnalazioni di non emergenza della città di Torino è costituito da una collezione di record relativi alle segnalazioni con i rispettivi attributi. Per prima cosa sono stati scaricati tutti i file di tipo csv relativi alle segnalazioni fatte dai cittadini alla polizia municipale nella città di Torino dal 2012 al 2019, e successivamente fatto una primissima esplorazione. Essendo alcuni dataset separati per anno e altri per semestre di ogni anno è stato deciso di riunire il tutto in un unico dataset. Come prima cosa sono stati esplorati i singoli attributi per verificarne l'integrità. È stato notato il basso numero di segnalazioni per l'anno 2013 e si è scoperto che nel dataset relativo all'anno 2012 vi erano segnalazioni con data relativa al 2013. Quindi sono stati corretti i record errati ed si è deciso di inserire l'attributo anno in ciascun dataset. Sono state ottenute 17738 segnalazioni. Le prime considerazioni sugli attributi erano principalmente legate all'attributo Area Verde, che presenta solamente 836 segnalazioni su aree verdi durante gli anni presi in esame e l'attributo legato all'ora che risulta essere quello di minor contenuto informativo poiché molte segnalazioni non presentano nessuna associazione con l'orario.

Dopo questa prima esplorazione del dataset è iniziata la vera fase di preprocessing

del dataset. La scarsa qualità dei dati influisce negativamente in un lavoro di data mining in cui vogliamo estrapolare contenuto informativo, per questo in prima battuta sono state individuate le problematiche legate alla qualità dei dati in nostro possesso. Durante questa fase è stata data molta attenzione a rintracciare i possibili:

- Valori mancanti
- Dati duplicati
- Dati errati

DATI MANCANTI

Categoria criminologa	0
Sottocategoria criminologica	0
Circoscrizione	348
Località	348
Area verde	16902
Data	0
Ora	13136

Dalla tabella sopra che mostra i dati mancanti individuati nel dataset si evince che per 348 segnalazioni manca sia l'attributo Circoscrizione che l'attributo Località e per 13136 segnalazioni manca l'attributo Ora. Per gli attributi Circoscrizione e Località i valori mancanti riguardano le stesse segnalazioni. Inoltre per molte segnalazioni (13136 su 17738) manca l'attributo temporale (Ora) e questo pone meno l'attenzione su un'analisi temporale nel corso della giornata. Le segnalazioni che non riguardano un'area verde sono presenti nel dataset con valore NaN per questo c'è un numero alto di dati mancanti per quanto riguarda l'attributo Area Verde. Si è scelto di non usare questi 348 record (Circoscrizione e Località) e quindi sono state rimosse queste segnalazioni dalla nostra analisi in quanto risulterebbe impossibile trovare un'associazione spaziale con queste segnalazioni. Inoltre il fatto che solo 4602 record presentano l'attributo temporale legato all'ora si è deciso di rimuovere tale attributo dalla nostra analisi dal momento che il suo utilizzo risulta irrilevante visto che non può essere applicato a tutte le segnalazioni presenti. Successivamente si sono analizzati i vari record per trovare eventuali dati mancanti, duplicati ed errati. Le uniche problematiche riguardavano alcuni dati errati per gli anni 2013 che avevo risolto immediatamente durante la prima esplorazione del dataset. Per il resto gli attributi presenti non riscontravano altre particolari anomalie. Il dataset finale così strutturato ha un totale di 17390 segnalazioni per gli anni che vanno dal 2012 al 2019. In particolar modo è stato ritenuto problematico

il valore dell'attributo sottocategoria criminologica *Altro* che è comune alle tre categorie, di fatto non si era in grado di distinguere a quale macrocategoria appartenesse e per questo motivo si è cercato un modo per evidenziare la voce “Altro” al fine di non essere qualcosa di generico ma qualcosa di più specifico per ogni categoria. Si è deciso di distinguere le tre voci, indicando con una lettera a quale categoria appartenesse così da tenere separate le diverse tipologie di Altro.

- Altro (A) riferito alla categoria Allarme Sociale.
- Altro (C) riferito alla categoria Convivenza Civile.
- Altro (Q) riferito alla categoria Qualità Urbana.

L'attributo “Area Verde”, come abbiamo visto in precedenza, è un campo nullo se la segnalazione non è avvenuta in un'area considerata verde all'interno della città di Torino, altrimenti presenta l'indicazione “Area Verde”. Ogni valore nullo verrà interpretato come informazione mancante, pertanto si è modificato il contenuto di tale attributo in questo modo:

- Si è modificata la scritta “Area Verde” con “Si”
- Si è modificato il valore nullo "NaN" con “No”

Successivamente dato che l'obiettivo della nostra analisi è rivolto a trovare delle correlazioni spaziali e temporali tra le segnalazioni risulta molto importante arricchire il dataset con questa tipologia di attributi. Essendoci pochi attributi temporali si è deciso quindi di arricchire il dataset con nuove proprietà che possono essere estratte da attributi già esistenti in modo da ottenere informazioni di livello superiore. Visto che l'attributo data è nel formato gg/mm/aaaa sono stati ricavati da questa informazione nuovi attributi temporali.

Per prima cosa risulta interessante poter inserire l'attributo *Stagione* al dataset come mostrato nella tabella sotto.

Primavera	Dal 21/03 al 20/06
Estate	Dal 21/06 al 22/09
Autunno	Dal 23/09 al 20/12
Inverno	Dal 21/12 al 20/03

Attraverso alcune funzionalità delle librerie Python si riesce ad associare ciascuna data al giorno della settimana e al mese corrispondente e quindi aggiungiamo al dataset anche l'attributo *Giorno settimanale* e *Mese*. A seguire si è deciso di inserire ai fini dell'analisi un attributo giorno *Feriale/Festivo* al dataset considerando come *Festivo* le seguenti date:

- Domenica
- 1 Gennaio, Capodanno
- 6 Gennaio, Epifania
- Pasqua e Pasquetta filtrando in base all'anno di riferimento
- 25 Aprile, Anniversario della Liberazione
- 1 Maggio, Festa dei lavoratori
- 2 Giugno, Festa della Repubblica
- 24 Giugno, San Giovanni, patrono di Torino
- 15 Agosto, Ferragosto
- 1 Novembre, Tutti i Santi
- 8 Dicembre, Immacolata Concezione
- 25 e 26 Dicembre, Natale e S.Stefano

Come si può notare sono state aggiunte le festività nazionali, mentre tutte le altre date sono etichettate come *Feriale*. Le festività di Pasqua e Pasquetta hanno date differenti negli anni, per questo si sono ricercate le date corrispondenti a ciascun anno così da identificare la data esatta ed etichettarla correttamente. Si è pensato inoltre di aggiungere la ricorrenza del 24 Giugno, San Giovanni, festività patronale della città di Torino.

Il dataset opportunamente sottoposto a preprocessing risulta così strutturato:

- Categoria criminologa
- Sottocategoria criminologica
- Circoscrizione
- Località
- Area Verde
- Data
- Anno
- Stagione
- Giorno settimanale
- Mese
- Feriale/Festivo

Dopo le opportune operazioni di preprocessing metteremo in risalto nel prosieguo di questo paragrafo contenuto del dataset che è stato usato nel seguente lavoro di data mining. Analizzeremo i vari attributi e come questi sono distribuiti con il fine di trarre alcune prime considerazioni iniziali. Il numero totale di segnalazioni è di 17390, e coprono gli anni che vanno dal 2012 al 2019. Il numero di occorrenze per l'attributo *Categoria* è di:

- **Convivenza Civile:** 8963
- **Qualità Urbana:** 7859
- **Allarme Sociale :** 568

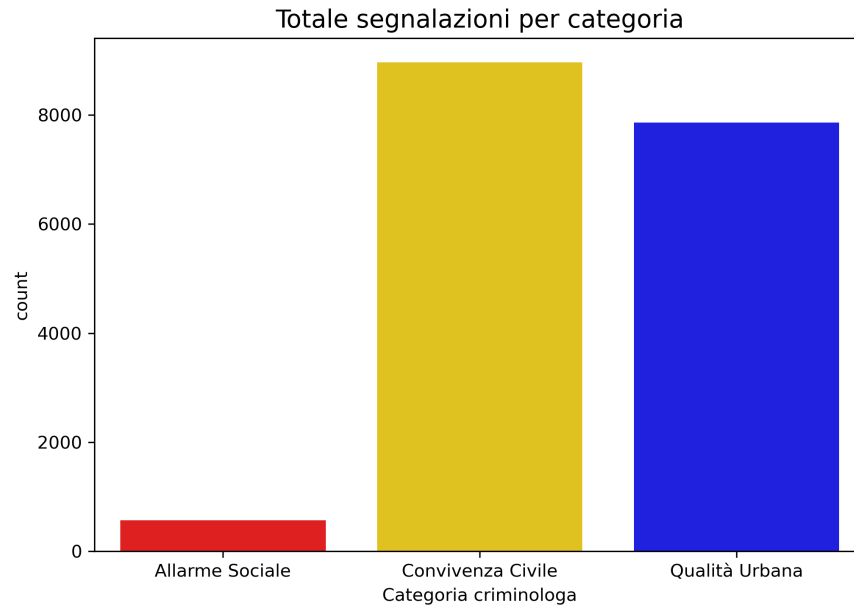


Figura 4.1: Segnalazioni divise per Categoria

Dando uno sguardo al grafico della fig.4.1 si può notare come la categoria *Allarme Sociale* è poco presente rispetto alle altre due categorie, che risultano invece più omogenee in termini quantitativi.

L'attributo *Sottocategoria criminologica* è così distribuito:

- *Allarme Sociale*
 - Atti di vandalismo: 65
 - Altro (A): 503
- *Convivenza Civile*
 - Altro (C): 4429
 - Aggregazioni giovanili: 101
 - Comportamenti molesti: 643
 - Disturbi Cani: 542
 - Disturbi altri animali: 23
 - Disturbi da locali: 1931
 - Rumori molesti: 1219
 - Uso improprio parti comuni: 75

- *Qualità Urbana*
 - Altro (Q): 2628
 - Decoro e degrado urbano: 4170
 - Veicoli abbandonati: 1061

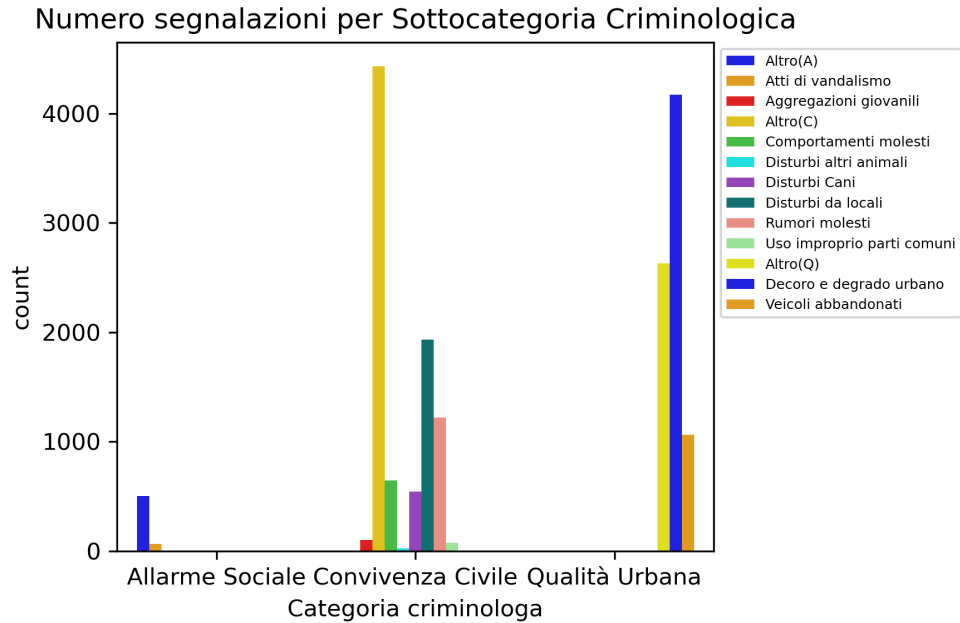


Figura 4.2: Segnalazioni divise per Sottocategorie

Possiamo osservare che nelle categorie Allarme Sociale e Convivenza Civile, la rispettiva sottocategoria *Altro* è quella che presenta un maggior numero di segnalazioni. Mentre nella categoria Qualità Urbana la sottocategoria *Decoro e degrado urbano* è quella che prevale rispetto alle altre. Quest'ultima categoria risulta essere anche leggermente più equilibrata per distribuzione di sottocategorie rispetto alle altre. Il grafico in figura 4.2 riesce a darci un'indicazione di come le sottocategorie riguardanti le segnalazioni sono distribuite nel dataset.

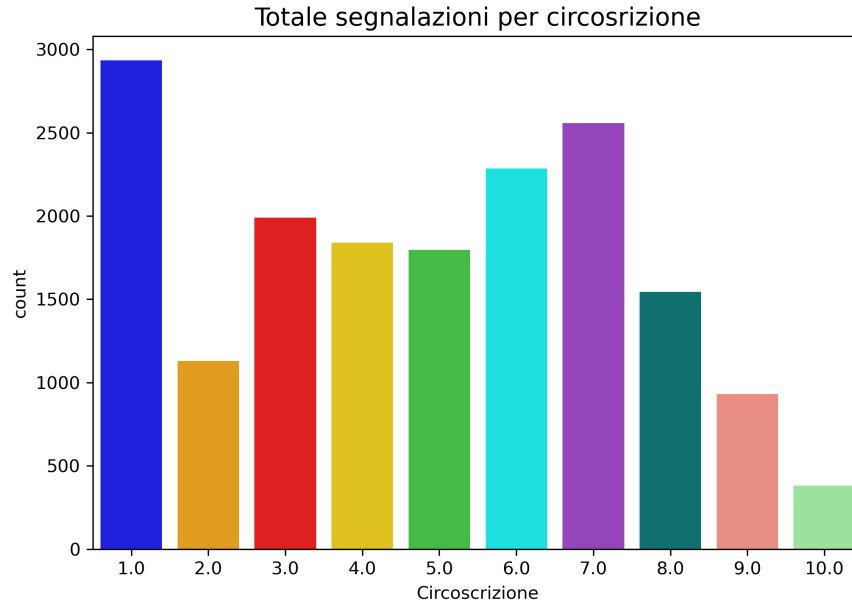


Figura 4.3: Segnalazioni divise per Circoscrizione

Interessante notare come le segnalazioni sono distribuite lungo l'attributo *Circoscrizione*. Dal grafico in figura 4.3 si osserva che risultano esserci maggiori segnalazioni nella Circoscrizione 1, successivamente abbiamo la Circoscrizione 7 e la Circoscrizione 6 in termini strettamente numerici, più o meno simile l'andamento delle segnalazioni nelle Circoscrizioni 3,4,5,8. Da una prima analisi del seguente grafico notiamo inoltre che un numero bassissimo di segnalazioni si registrano per la Circoscrizione 10. Il numero esatto di segnalazioni per ogni circoscrizione della città di Torino è qui indicato:

- Circoscrizione 1: 2933
- Circoscrizione 2: 1129
- Circoscrizione 3: 1991
- Circoscrizione 4: 1840
- Circoscrizione 5: 1796
- Circoscrizione 6: 2286
- Circoscrizione 7: 2557
- Circoscrizione 8: 1545

- Circoscrizione 9: 932
- Circoscrizione 10: 381

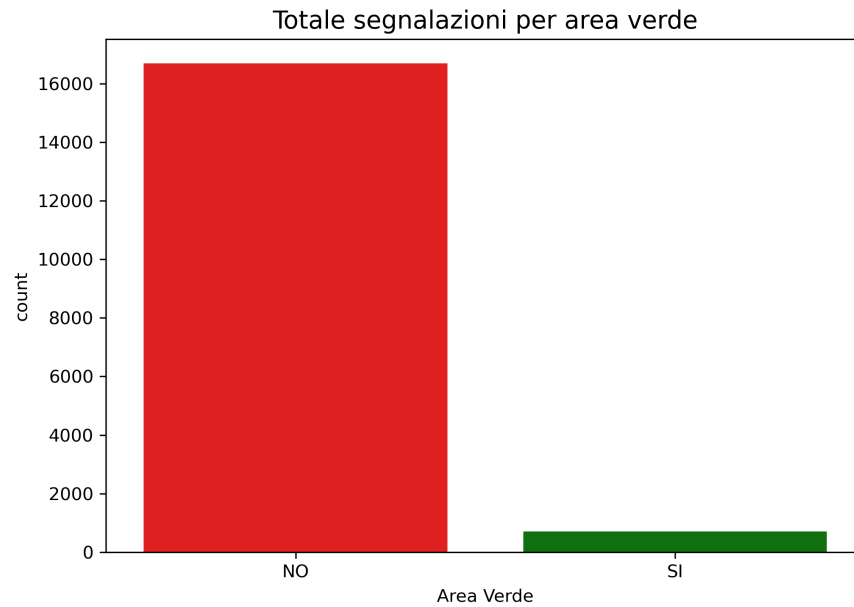


Figura 4.4: Segnalazioni in Area verde

La figura 4.4 mostra il totale di segnalazioni urbane fatte in luoghi considerati *Area Verde*. Il grafico mostra come le segnalazioni in aree verdi sono nettamente inferiori. Non si sa se il motivo di questa differenza sia legata ad una mancata registrazione del tipo di luogo nel momento in cui è stata fatta la segnalazione. Delle segnalazioni totali solo 705 sono registrate come *Area Verde*.

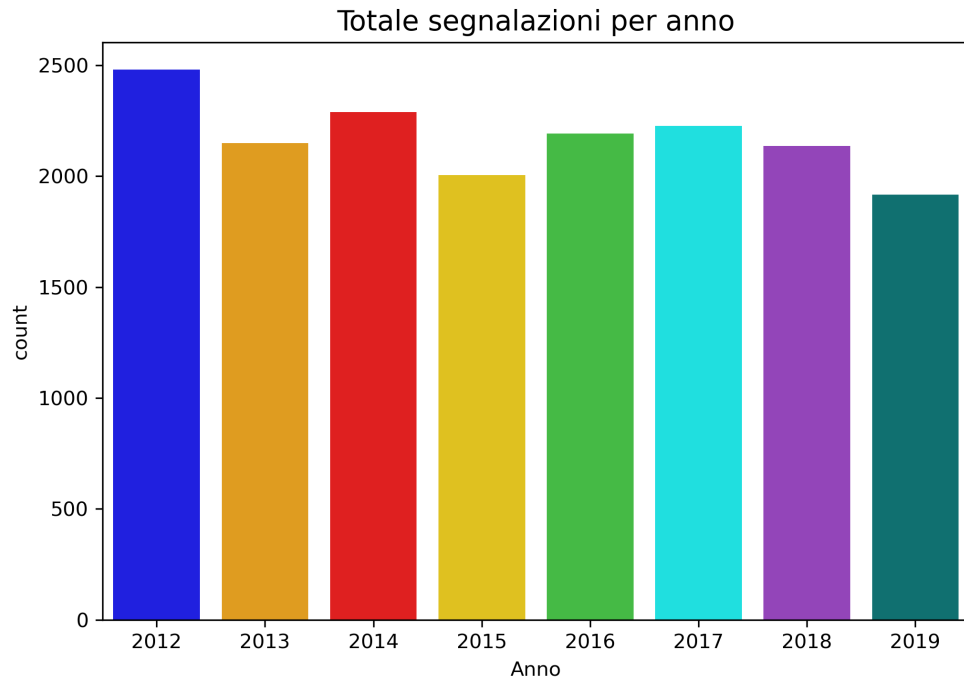


Figura 4.5: Segnalazioni distribuite negli anni

Il dataset così strutturato ha un numero piuttosto omogeneo di segnalazioni nel corso degli anni presi in esame. La figura 4.5 mostra la distribuzione delle segnalazioni durante gli anni. L'anno 2012 è quello che ha un numero maggiore di dati, mentre il 2019 ne presenta meno se si prendono in considerazione gli altri anni.

- 2012: 2479
- 2013: 2148
- 2014: 2288
- 2015: 2004
- 2016: 2191
- 2017: 2227
- 2018: 2136
- 2019: 1917

Interessante osservare l'andamento del numero di segnalazioni nei mesi. Dal grafico in figura 4.6 si evince come le segnalazioni sono distribuite con una frequenza piuttosto simile. I mesi che hanno una distribuzione inferiore sono il mese di Agosto e di Dicembre.

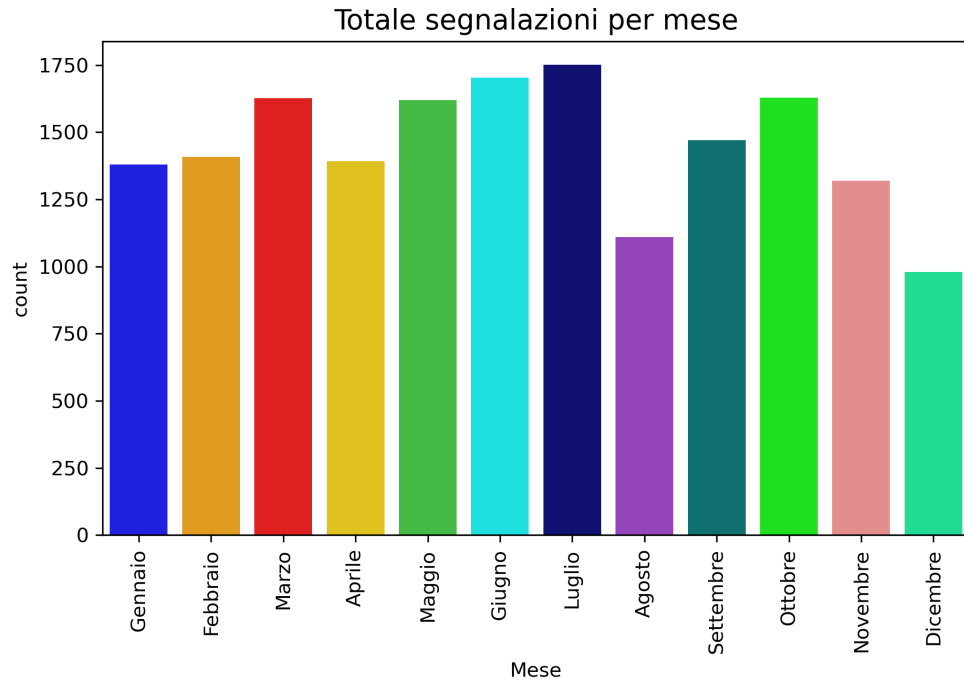


Figura 4.6: Segnalazioni distribuite nei mesi

L'attributo *Mese* è così distribuito:

- Gennaio: 1380
- Febbraio: 1408
- Marzo: 1627
- Aprile: 1392
- Maggio: 1620
- Giugno: 1703
- Luglio: 1751
- Agosto: 1110

- Settembre: 1471
- Ottobre: 1629
- Novembre: 1319
- Dicembre: 980

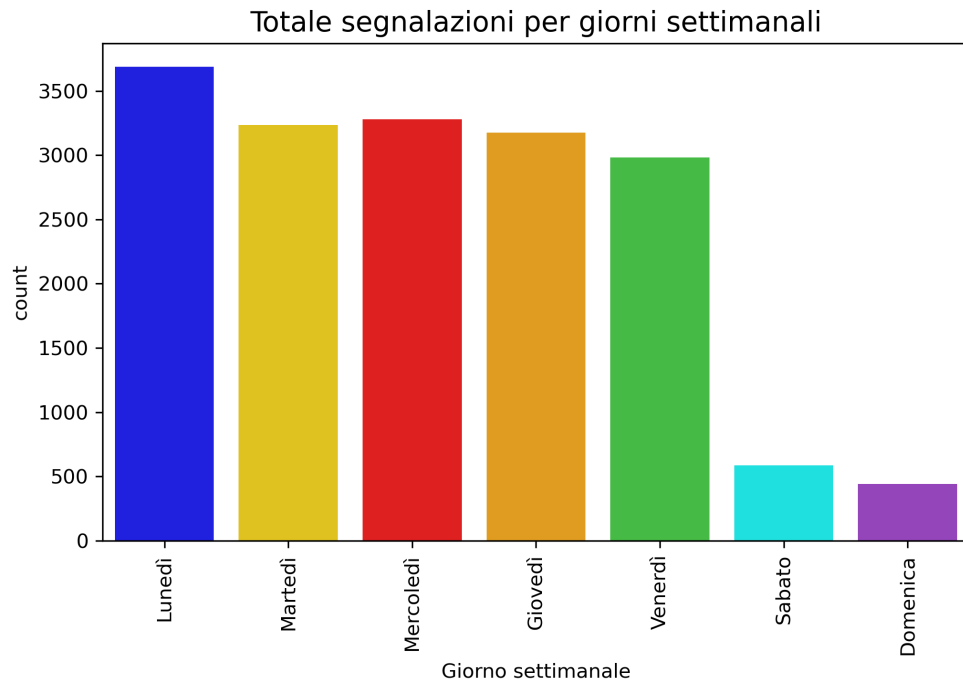


Figura 4.7: Segnalazioni per giorno settimanale

Dal grafico in figura 4.7 si evince la distribuzione delle segnalazioni nel corso dei giorni della settimana. Si osserva come il numero di segnalazioni nei weekend è nettamente inferiore rispetto agli altri giorni settimanali. Il numero esatto di segnalazioni per ogni giorno della settimana è il seguente:

- Lunedì: 3687
- Martedì: 3237
- Mercoledì: 3280
- Giovedì: 3176
- Venerdì: 2982

- Sabato: 586
- Domenica: 442



Figura 4.8: Distribuzione delle segnalazioni nei giorni Feriali e Festivi

Anche il numero di segnalazioni nei giorni considerati *Festivi* è basso, sono solamente 490; mentre ci sono 16900 segnalazioni fatte nei giorni *Feriali*.

Durante l'analisi è stato preso in considerazione anche l'attributo legato alla *Stagione* in cui è registrata la segnalazione da parte di un cittadino. Il grafico in figura 4.9 ne mostra l'andamento; si osserva una simile frequenza di distribuzione.

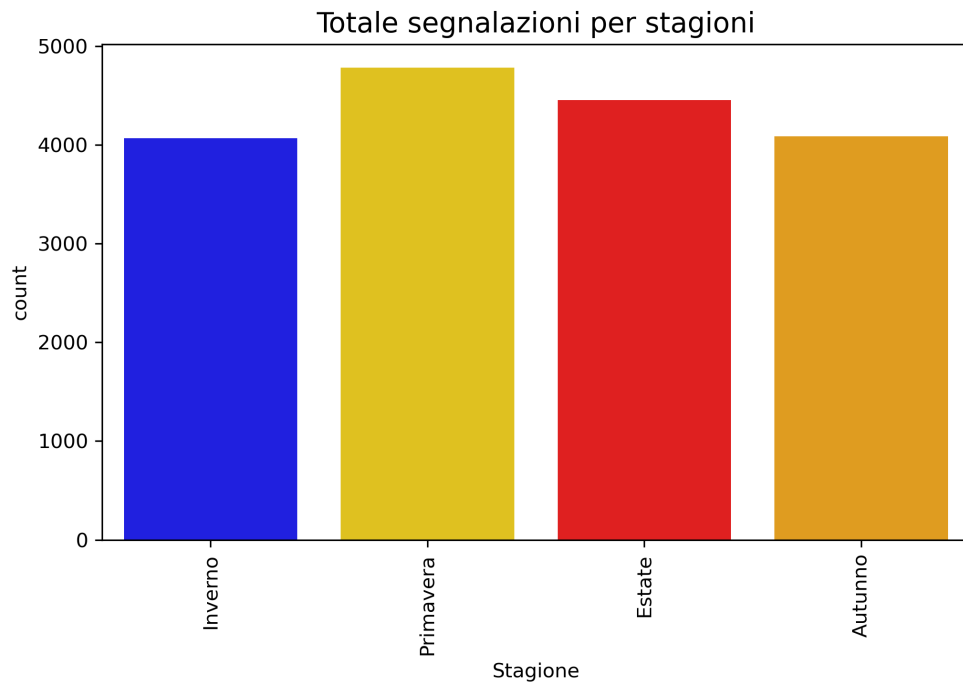


Figura 4.9: Distribuzione delle segnalazioni nelle stagioni

L'attributo *Stagione* è così distribuito all'interno del dataset:

- Inverno: 4067
- Primavera: 4781
- Estate: 4454
- Autunno: 4088

In questa fase di analisi diventa interessante notare l'andamento di ogni categoria nel corso degli anni. La figura 4.10 ci mostra per ogni categoria il totale delle segnalazioni per ciascun anno; si evince che la categoria *Convivenza Civile* ha una distribuzione simile negli anni, e inoltre risulta la categoria che presenta un maggior numero di segnalazioni durante gli anni, tranne per l'anno 2019 in cui si ha un maggior numero di segnalazioni per la categoria *Qualità Urbana*. Si nota anche che nel corso degli ultimi due anni (2018-2019) c'è una diminuzione delle segnalazioni per la categoria *Allarme Sociale*. La distribuzione della categoria *Qualità Urbana* è piuttosto costante. Dista interesse osservare che per l'anno 2012 si ha il maggior numero di segnalazioni sia per la categoria *Convivenza Civile* e sia per *Qualità Urbana*. Nell'anno 2014 si registra il maggior numero di segnalazioni per la categoria *Allarme Sociale*.

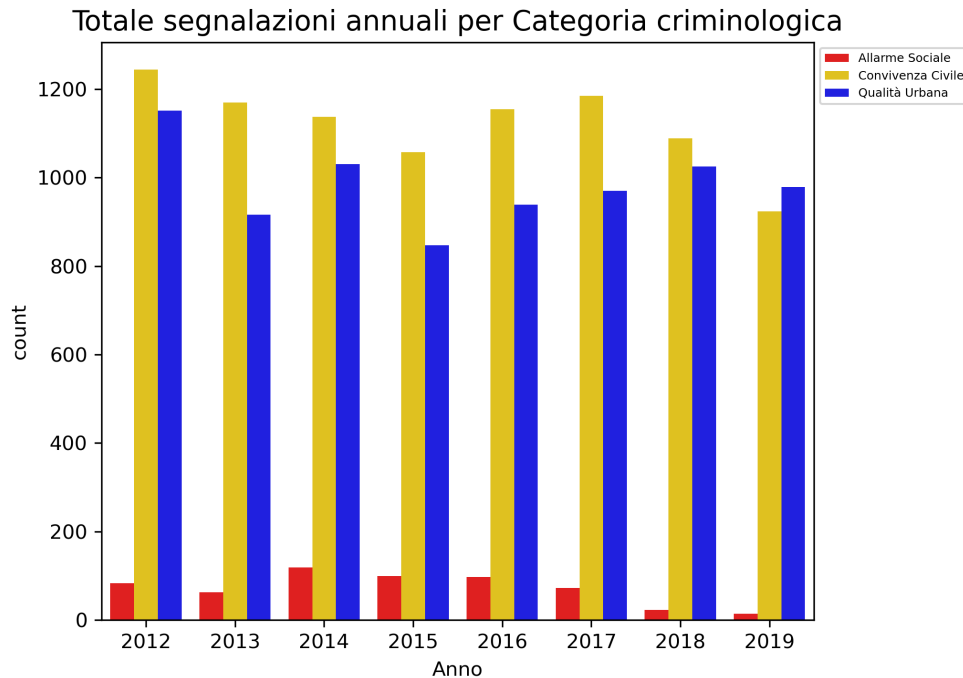


Figura 4.10: Distribuzione delle segnalazioni per ogni categoria negli anni

Il grafico in figura 4.11 espone come la categoria criminologica è distribuita lungo le circoscrizioni della città di Torino. Possiamo rilevare che la circoscrizione 1 presenta il maggior numero di segnalazioni legati alla categoria *Convivenza Civile*. La categoria *Qualità Urbana* ha invece nella circoscrizione 7 il maggior numero di segnalazioni. Questa è anche la circoscrizione che registra la quantità di segnalazioni maggiore relative alla categoria *Allarme Sociale*. Inoltre 6 circoscrizioni su 10 (1,2,5,7,8,9) registrano più segnalazioni per la categoria *Convivenza Civile*.

rispetto alle altre, diversamente le circoscrizioni 3 e 6 presentano lo stesso numero di segnalazioni per queste due categorie, mentre la circoscrizione 4 e 10 riportano un maggior numero di segnalazioni registrate come *Qualità Urbana*. La circoscrizione 10 risulta anche quella con il più basso numero di segnalazioni per tutte le categorie analizzate.

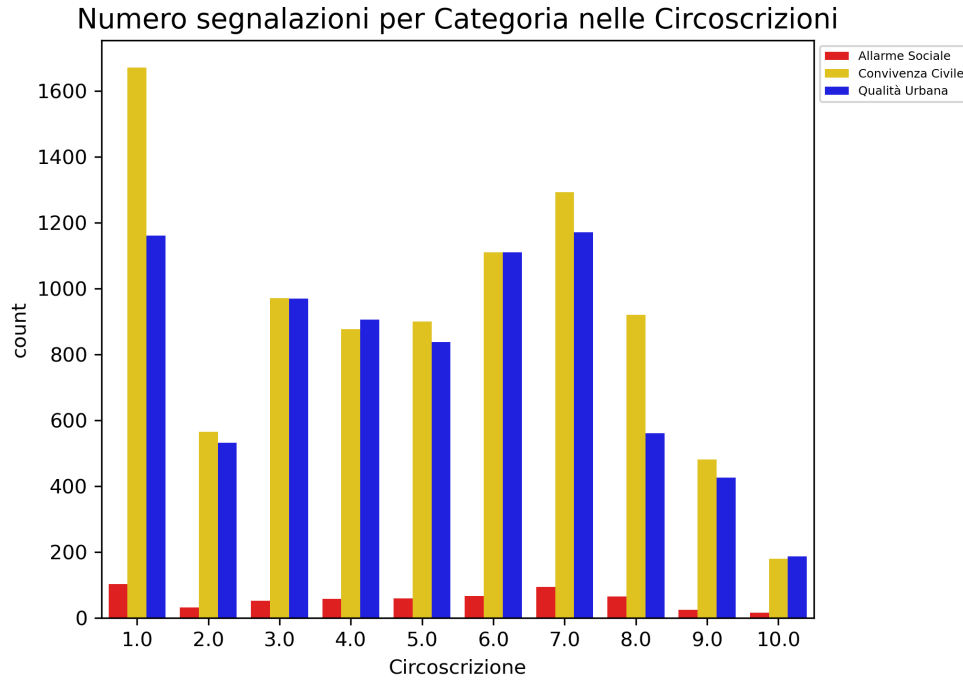


Figura 4.11: Distribuzione delle segnalazioni nelle circoscrizioni per ogni categoria

È stato anche creato un grafico (figura 4.12) per mostrare come l'attributo *Area Verde* è distribuito nelle varie categorie criminologiche, risulta interessante che le poche segnalazioni registrate in luoghi considerati come *Area Verde* sono in maggioranza legate alla categoria *Qualità Urbana*. Nel successivo grafico viene mostrato questo aspetto e si nota che le maggiori richieste fatte per problematiche di *Qualità Urbana* sono relative alla sottocategoria *Decoro e degrado urbano*.

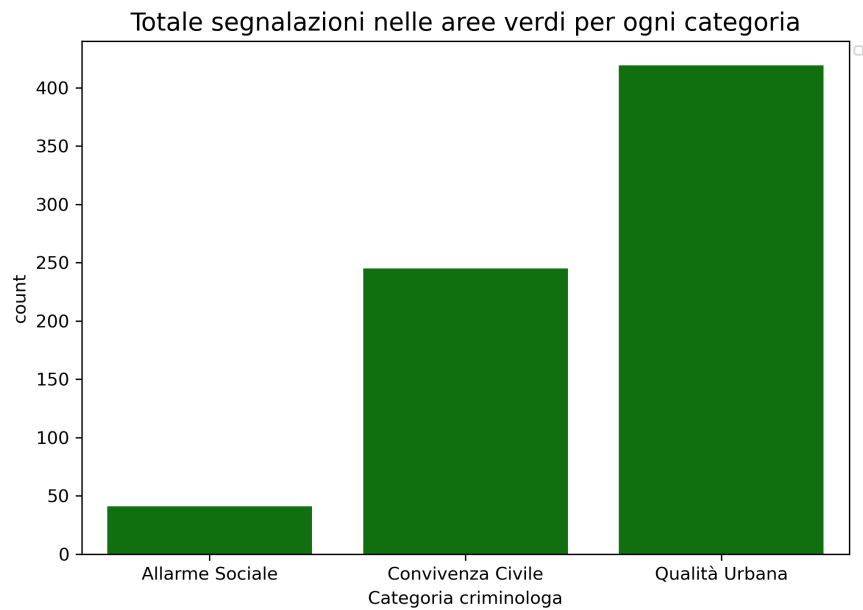


Figura 4.12: Distribuzione delle segnalazioni per Area Verde per ogni categoria

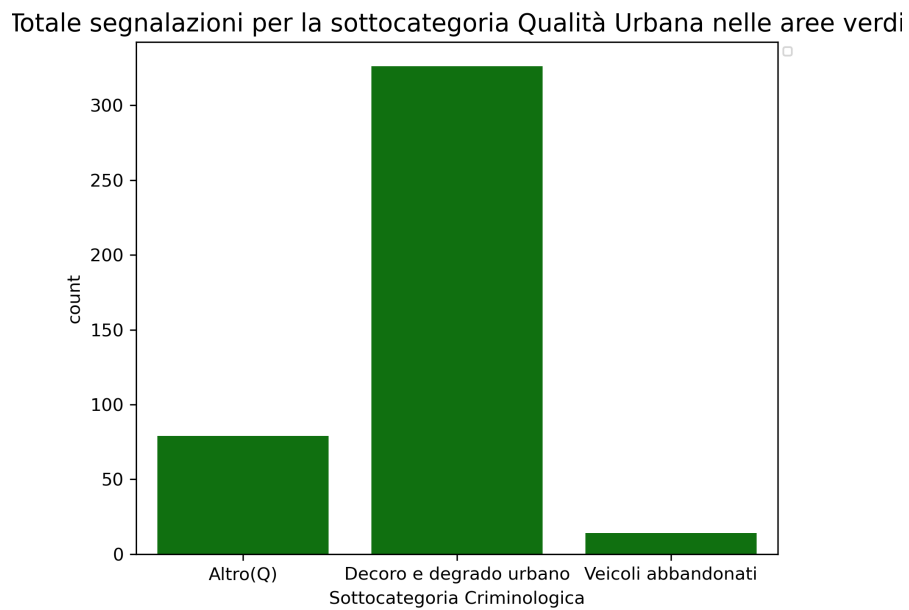


Figura 4.13: Numero segnalazioni per le sottocategorie di Qualità Urbana in Area Verde

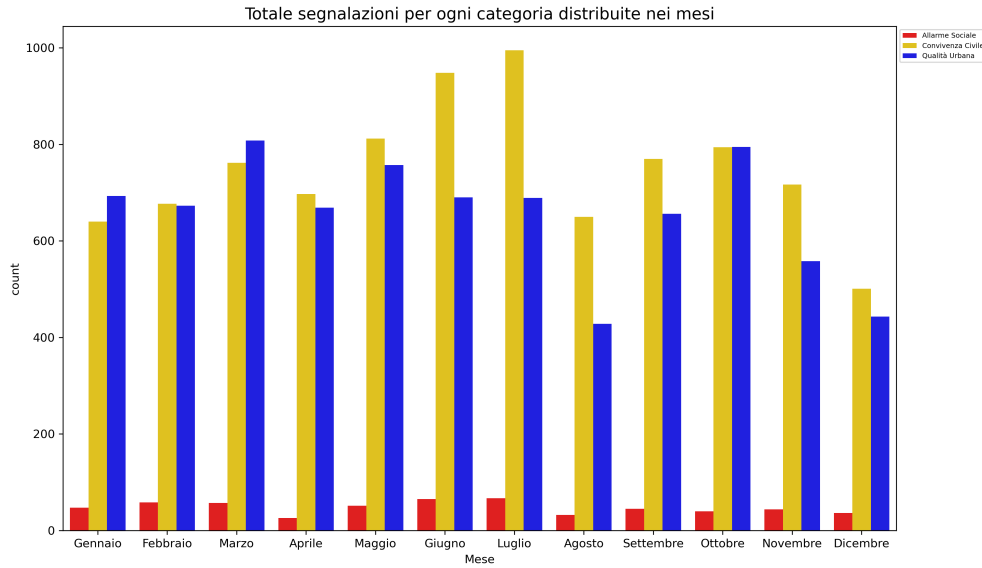


Figura 4.14: Numero segnalazioni per categoria nei mesi

Si sono osservate anche le distribuzioni delle segnalazioni per ogni categoria nei vari mesi. La categoria *Convivenza Civile* registra il maggior numero di segnalazioni nel mese di Luglio, di poco inferiore la frequenza nel mese di Giugno. La categoria *Qualità Urbana* ha un andamento piuttosto lineare, e segna la quantità maggiore di segnalazioni nel mese di Marzo. Il mese di Agosto registra il mese con il minor numero di segnalazioni di tipo *Qualità Urbana*, invece per la categoria *Convivenza Civile* è il mese di Dicembre ad avere il più basso numero di segnalazioni. La categoria *Allarme Sociale* ha una frequenza molto simile nei dodici mesi, il mese di Aprile è quello che presenta meno segnalazioni di questa tipologia categorica. In quasi tutti i mesi la categoria *Convivenza Civile* è quella che primeggia per numero di segnalazioni, tranne nei mesi di Gennaio e Marzo in cui si registrano un maggior numero di segnalazioni relative alla categoria *Qualità Urbana*, oltre a ciò, particolare è il mese Ottobre in cui ci sono lo stesso numero di richieste per le categorie *Convivenza Civile* e *Qualità Urbana*.

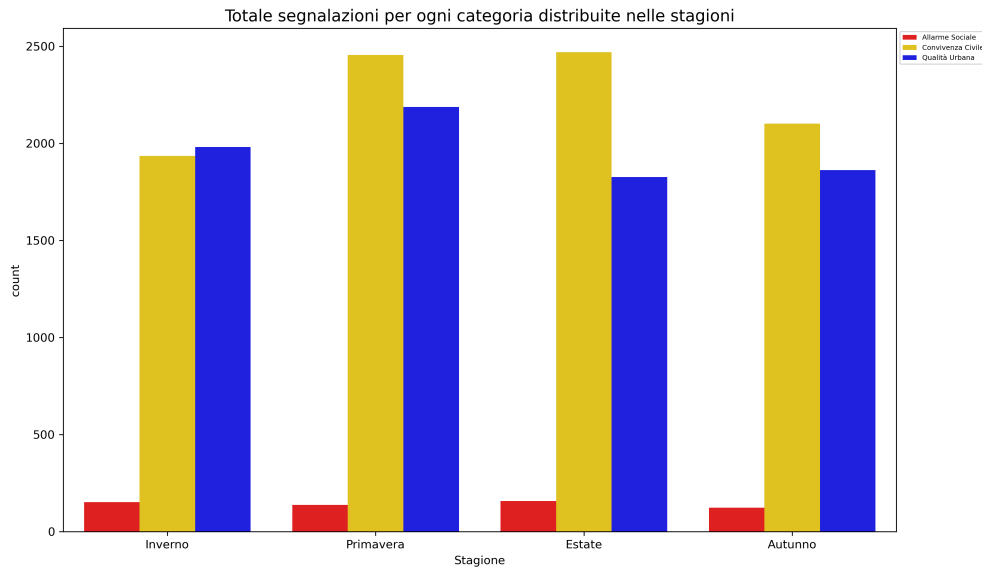


Figura 4.15: Numero segnalazioni per categoria nelle stagioni

Si è ritenuto interessante dare uno sguardo anche al numero di segnalazioni per categoria criminologica nelle varie stagioni. Il grafico della figura 4.15 mostra l'andamento di questa osservazione nel dataset. Le considerazioni emerse sono una distribuzione più o meno uguale per le stagioni Primavera ed Estate per la categoria *Convivenza Civile*; per di più questa risulta essere anche la categoria dominante per numero di segnalazioni prendendo in considerazioni le quattro stagioni, fatta eccezione per la stagione invernale in cui si ha un numero leggermente inferiore di segnalazioni rispetto alle problematiche riguardanti la categoria *Qualità Urbana*. Una distribuzione lineare emerge per la categoria *Allarme Sociale*.

Per ultimo si è analizzata la frequenza per ogni categoria nei giorni settimanali.

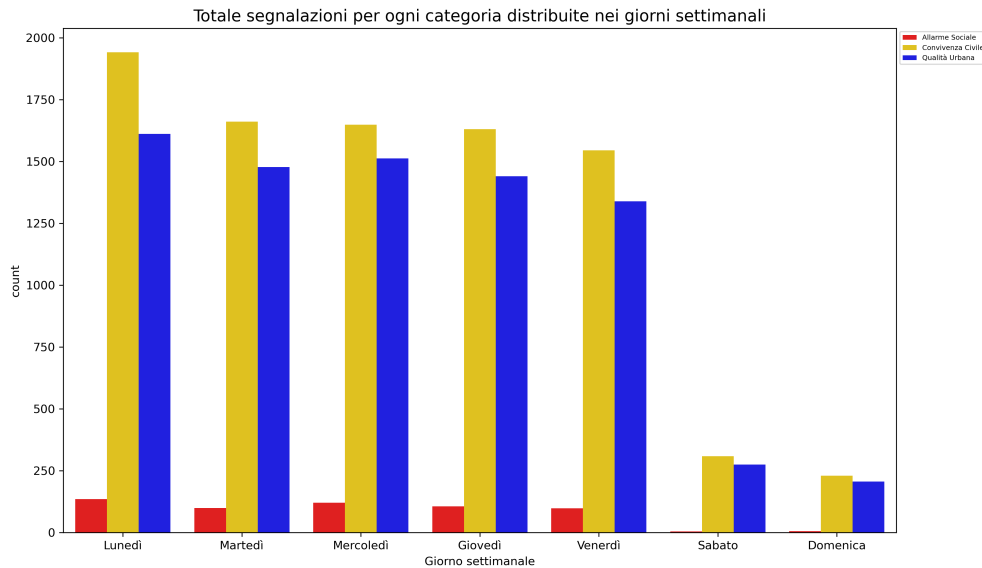


Figura 4.16: Numero segnalazioni per categoria durante i giorni settimanali

Dal grafico in figura 4.16 si osserva che prendendo in esame ciascun giorno della settimana si ha una quantità superiore di segnalazioni per la categoria *Convivenza Civile*. Come si era osservato in precedenza nel corso del fine settimana diminuiscono le segnalazioni registrate, e piuttosto rilevante risulta la categoria *Allarme Sociale* che registra un numero veramente basso nelle giornate di Sabato (4) e Domenica (6). Inoltre solo queste 6 segnalazioni saranno etichettate come *Festivo* per la seguente categoria criminologica.

Le primissime considerazioni emerse da questa analisi esplorativa sono che la categoria *Allarme Sociale* presenta un numero poco rilevante di richieste di non emergenza, soprattutto se paragonate alle altre due tipologie di categoria. In aggiunta si è osservato che gli attributi temporali aggiunti nel corso della fase di preprocessing saranno molto utili ai fini degli obiettivi della nostra analisi dei dati insieme a quelli geografici. Si sono viste alcune caratterizzazioni al variare degli attributi presenti nel dataset ed in particolar modo come le varie categorie cambiano nel tempo e nelle varie zone della città. Oltre a ciò si è notato anche che le sottocategorie presenti risultano piuttosto *sparse*, congiuntamente agli attributi spaziali Area Verde e Feriale/Festivo.

4.2 Dataset New York City

Il dataset sulle segnalazioni di non emergenza della città di New York è strutturato in maniera più dettagliata rispetto a quello Torinese. Inoltre, come si è notato nel capitolo precedente, presenta un numero di segnalazioni e attributi nettamente maggiore. Nonostante questo si è cercato di esplorare nel dettaglio il dataset per trovare possibili analogie che potessero organizzare il dataset in modo simile per effettuare un lavoro affine e infine mettere in risalto eventuali confronti e considerazioni. Dal sito NYC Open Data [11] è possibile scaricare il file csv relativo alle richieste di segnalazioni nella città di New York dal 2010 ad oggi. Il dataset è singolo e registra circa 28 milioni di record. La rispettiva fase di preprocessing visto l'alto numero di segnalazioni e attributi è stata molto lunga e sono stati analizzati diversi aspetti per ottenere una struttura conforme del dataset. Innanzitutto si è avviata una fase di riduzione dei dati in cui sono stati analizzati i vari attributi per capire quali potessero essere utili alla nostra analisi e quali no, così da avere un dataset più compatto su cui lavorare e su cui fare determinate operazioni di pulizia dei dati. L'attributo *Unique key* fornisce una chiave per ogni segnalazione registrata nel dataset. Ai fini della nostra analisi il seguente contenuto informativo dato da questo attributo è stato considerato nullo, pertanto non è stato preso in esame. L'attributo *Created Data* fornisce data e ora della segnalazione, questo contrariamente al precedente risulta rilevante in quanto fornisce un'indicazione temporale sulla segnalazione. L'attributo *Closed Data* registra ora e data della chiusura della segnalazione, sicuramente è un dato aggiuntivo che ai fini informativi però si è scelto di non utilizzare. Gli attributi *Agency* e *Agency Name* indicano quale agenzia governativa di sicurezza pubblica ha trattato la richiesta di assistenza, non essendoci un riferimento nel dataset di Torino e poichè si è constatato che molti enti governativi risultassero simili per determinate zone della città si è scelto di accantonare tali attributi. Non si poteva fare a meno invece dell'attributo *Complaint Type* che identifica la categoria e/o argomento della relativa segnalazione. Questo attributo diventa tra i principali in quanto permette una possibile associazione con la sottocategoria del dataset sulla città italiana. Dall'esplorazione di questo attributo si evince che ci sono più di 200 diversi Complaint Type. L'attributo *Descriptor* fornisce dei dettagli aggiuntivi sulla segnalazione fatta, descrivendo meglio il contesto. Si è notato che non c'è un'indicazione precisa su come fornire la descrizione, in alcuni casi sembra ben delimitata, in altri sembra lasciata libera scelta al cittadino su cosa indicare. Non rari i casi in cui questo attributo risulta mancante. Visto il potenziale informativo che può fornire si è deciso di non eliminare questo attributo in quanto poteva risultare utile nella fase di preprocessing dei dati. Molto interessante l'attributo *Location Type* che indica il contesto spaziale in cui viene registrata la segnalazione. Questo tipo di attributo manca nel dataset della città di Torino, o

meglio è presente solo in parte (attributo Area Verde o meno). L'informazione geografica data da questo attributo risulta di particolare rilievo nel prosieguo del lavoro, e quindi questo dato nonostante siano molti i casi in cui l'attributo è nullo, vedremo come sarà opportunamente trattato nel corso delle successive fasi di preprocessing per organizzarlo adeguatamente. L'attributo *Incident Zip* nonostante fornisca un'indicazione spaziale legata alla città è stato accantonato in quanto è stata data più importanza ad altri attributi che rappresentano meglio i contesti geografici cittadini. *Incident Address* e *Street Name* sono attributi molto simili. Il primo registra l'indirizzo completo di nome e numero civico. Il secondo indica solo il nome della via in cui è registrata la segnalazione. Tra i due si è scelto di utilizzare l'attributo *Street Name* in quanto analogo all'attributo *Località* presente nel dataset di Torino. I quattro attributi *Cross Street 1*, *Cross Street 2*, *Intersection Street 1*, *Intersection Street 2* relativi alle strade incrociate dalla via in cui si esegue la segnalazione sono stati valutati superflui. Anche l'attributo *Address Type* è stato ritenuto poco interessante ai fini informativi. L'attributo *City* è un attributo controverso per il tipo di divisione territoriale della città di New York, in quanto questo è legato allo zip code di una determinata zona della città. Per tale motivo si è preferito non prenderlo in considerazione, visto che sono presenti nel dataset attributi geografici che identificano meglio le zone territoriali di New York City. L'attributo *Landmark* denota un eventuale punto di riferimento. Questo attributo non è presente in tutti i record, inoltre spesso riporta nuovamente il nome della via e fornisce una informazione poco precisa. Sono stati ritenuti poco attinenti al nostro lavoro gli attributi *Status*, *Due Date*, *Resolution Description*, *Resolution Action Updated Date*. L'attributo *Facility Type* è raramente presente nel dataset, ed in aggiunta non ha molto contenuto informativo per il nostro tipo di analisi. Sono stati rimossi dal dataset anche gli attributi *Open Data Channel Type*, *BBL* e i due attributi relativi alle coordinate geografiche. L'attributo *Borough* indica il distretto in cui avviene la segnalazione. Questo ci permette di avere un riferimento geografico importante per la nostra analisi. Un altro attributo legato gerarchicamente al precedente è *Community Board*, questo fornisce una ulteriore divisione territoriale all'interno del relativo distretto. Questo attributo può fornirci un importante contenuto informativo dal punto di vista geografico. Gli attributi *Park Facility Name* e *Park Borough* dovrebbero darci informazioni sui luoghi considerati Area Verde, i parchi in questo preciso caso. Purtroppo il primo è quasi sempre registrato come "non specificato", mentre il secondo, se presente, ripete l'informazione contenuta già nell'attributo *Borough*. Per tale motivazione non sono stati presi in considerazione ai fini della struttura finale del dataset da utilizzare per il progetto. Ci sono attributi che sono in larga quantità nulli all'interno del dataset di NY, per questo si è deciso di rimuoverli. Rientrano in questa tipologia gli attributi *Vehicle Type*, *Taxi Company Borough*, *Taxi Pick Up Location*, *Bridge Highway Name*, *Bridge Highway Direction*, *Road*

Ramp, Bridge Highway Segment. Analogamente gli attributi *Latitude, Longitude e Location* sono stati rimossi dal dataset nonostante abbiano un contenuto informativo molto dettagliato spazialmente, ma più vicino ad una analisi di geolocalizzazione attraverso coordinate geografiche.

Il dataset risulta a questo stato così composto da un punto di vista di attributi:

- Created Date
- Complaint Type
- Descriptor
- Location Type
- Street Name
- Borough
- Community Board

Successivamente è stato svolto un lavoro di pulizia dei dati, sottoponendo ad attenta considerazione eventuali valori nulli, incompleti e/o errati.

Si è scelto di svolgere l'analisi sugli interi anni a nostra disposizione, quindi sono stati selezionati gli anni dal 2010 al 2021. Le segnalazioni riguardanti i primi mesi del 2022 sono stati rimossi dal dataset.

Il numero di valori mancanti individuati per ogni attributo sono qui riportati:

DATI MANCANTI

Created Data	0
Complaint Type	0
Descriptor	438069
Location Type	6419730
Street Name	120212
Borough	45035
Community Board	45035

Le segnalazioni che non registravano Borough e Community Board sono stati rimossi dal dataset in quanto non è stato possibile trovare una associazione geografica corretta anche a causa della mancata registrazione della via di riferimento. Mentre gli altri attributi mancanti per il momento sono rimasti intatti in vista di un futuro step di Data Transformation.

Analizzando l'attributo legato alla data e ora della segnalazioni si è scelto di formattarlo adeguatamente, scegliendo il seguente formato datetime della libreria Python: "aaaa-mm-gg hh:mm:ss".

Visto l'enorme quantità di record si sono analizzati anche eventuali dati incompleti e imprecisi provando a risolvere queste incognuenze. È stata constatata la correttezza dei record per l'attributo che indica il distretto territoriale di appartenenza (Borough) che nella città di NY sono 5: *Bronx*, *Brooklyn*, *Manhattan*, *Queens* e *Staten Island*. Contrariamente l'attributo *Community Board* registrava alcuni errori e imprecisioni. Ci sono stati casi in cui il dato non era pertinente con l'attributo o grammaticalmente errato. I record che presentavano valori completamente errati sono stati rimossi, i secondi dove si è potuto, è stata effettuata una opportuna correzione. Vi erano record che comparivano incompleti, con l'indicazione "Unspecified + Borough di riferimento", per questi dati non è stato possibile identificarli correttamente, per questo sono stati rimossi dal set di dati. Si è deciso di tenere solo valori esattamente associati ad un corrispondente Community Board. L'attributo più ostico è stato quello che permette una identificazione categorica, *Complaint Type*. In primis per l'alto numero di valori "sottocategorie" presenti, circa 230. D'altro lato per le similarità esistenti tra alcune sottocategorie. Alcuni record completamente errati da un punto di vista significativo sono stati rimossi. Al fine di ottimizzare al meglio il contenuto del dataset si è deciso di lavorare opportunamente questo tipo di attributo seguendo una appropriata strategia di trasformazione ed integrazione dei dati. In rete è stata effettuata una scrupolosa ricerca su dataset 311 in America e sono stati presi in considerazione tre diversi studi. La prima [17] considera uno studio su 3 città americane, nel periodo che va dal 2012 al 2015, usando open data 311, effettuando degli ulteriori adeguamenti al fine di riclassificare i tipi di reclamo in categorie standardizzate nelle diverse città e nel periodo di tempo dell'analisi. La loro analisi verte su un aspetto socio-economico dei quartieri delle varie città, in modo da mostrare differenze socio-economiche esistenti al fine di identificare dei modelli di cluster. Una seconda ricerca [12] in cui vengono usati dataset 311 della città di New York per gli anni che vanno dal 2010 al 2014 ha evidenziato spunti interessanti da poter integrare nel seguente lavoro di tesi. Questo lavoro fornisce una gerarchia superiore all'attributo Complaint Type, infatti sono creati 16 Complaint Category a cui i vari Complaint Type vengono associati. Questa creazione del doppio livello è molto vicina alla nostra idea di divisione del dataset di Torino in Categoria Criminologica e Sottocategoria criminologica, nonostante la divisione delle Categorie sembra essere molto diversa rispetto alle tre categorie del dataset della città di Torino e malgrado il numero dei possibili valori dell'attributo *Complaint Type* risulti molto alto. Il terzo lavoro [6] è un'analisi molto recente di tipo clustering sulle città americane per frequenza delle categorie. Vengono presi in considerazione dataset 311 di 20 città americane (42 milioni di record su 20 città dal 2006 al 2019), per ogni anno si effettua clustering sulle città che risultano simili in base al tipo di richiesta. Cosa molto interessante è che in questa analisi per superare il problema delle categorie non standard per le diverse città è stato deciso di standardizzare i tipi di richiesta in 79 tipi di richiesta raggruppando queste

ultime in 12 categorie generali, così da avere uno standard fisso per tutte le città prese in considerazione nello studio. Anche qui si ha un doppio livello gerarchico che richiama la tassonomia categoria/sottocategoria del dataset di Torino.

Partendo dalle idee estrapolate in questi diversi studi, si è deciso di organizzare il dataset di NY con una struttura gerarchica che permetta di avere due livelli il più possibile affini a quelli del dataset della città di Torino. Quindi si è deciso di utilizzare l'attributo *Complaint Type* come sottocategoria, e di conseguenza creare un nuovo attributo, *Complaint Category*, prendendo spunto principalmente dalla tassonomia progettata nell'ultima ricerca [6] in quanto permette di avere un numero contenuto sia di possibili categorie che di sottocategorie.

Per ottenere dei dati ben strutturati ed avere così questa suddivisione gerarchica si è svolto un lavoro accurato in cui sono stati esaminati gli attributi *Complaint Type* e *Descriptor*. Sono stati presi in considerazione eventuali informazioni simili e/o eventuali ridondanze presenti nell'attributo *Complaint Type*, se due caratteristiche rappresentano lo stesso concetto, dovrebbero essere identificate dallo stesso nome. Parallelamente alcune sottocategorie generiche sono state etichettate dettagliatamente e correttamente grazie al contenuto informativo dell'attributo *Descriptor*. È stato fatto uno sforzo enorme per standardizzare questi concetti e si è usato tutto il potenziale informativo degli attributi prendendo spunto dalla standardizzazione fornita dal lavoro precedentemente citato. Prima vi erano più di 230 tipi di *Complaint Type* nell'intero set di dati, dopo la standardizzazione, questo numero si è ridotto a 60. A sua volta questi elementi sono stati raggruppati in 11 categorie generali sotto l'attributo *Complaint Category*. Quanto ottenuto permette di avere un quadro generale che copre in maniera gerarchica tutte le segnalazioni di non emergenza della città di New York nella sua interezza.

La struttura gerarchica ottenuta è la seguente:

- **Air & Water Quality**
 - *Air quality complaint*
 - *Request Water Test Kit*
 - *Water quality complaint*
 - *Asbestos complaint*
- **Animal**
 - *Animal control/care*
 - *Dead animal*
- **Cleaning & Garbage**
 - *Bulky item collection*

- *Garbage*
- *Graffiti*
- *Illegal dumping*
- *Hazardous waste*
- *Missed Recycling collection*
- *Sewage/Storm water*
- *Unsanitary condition*
- *Weeds*

- **Complaint**

- *Residential noise*
- *Street noise*
- *Nuisance complaint*
- *Rodent*
- *Mold*
- *Consumer complaint*
- *Illegal building*
- *Smoking*
- *Illegal Posting/Signs*
- *Aircraft/Helicopter noise*
- *Drinking*
- *Drug Activity*
- *Vaccine complaint*
- *Property Violations*

- **Information**

- *Information/Request*
- *DOF Property*
- *Payment Problem*
- *License*
- *Inspection*
- *Lost and Found*

- *Refund request*

- **Maintenance**

- *Heat and Hot Water system*
- *Construction and Plumbing*
- *Building repair*
- *Water system*
- *Electrical*
- *Elevator*
- *Vacant Property*

- **Parks & Trees**

- *Tree condition*
- *New Tree Request*
- *Park maintenance*

- **Poverty**

- *Homeless/Encampment*
- *Housing*
- *Panhandling*

- **Public Mobility Services**

- *Taxi complaint*
- *Shared vehicle complaint*

- **Streets & Sidewalks**

- *Street/Sidewalks repair*
- *Street/Road/Sidewalk obstruction*
- *Street light repair*
- *Street/Road sign repair*
- *Snow or Ice removal*

- **Transportation**

- *Illegal Parking*
- *Derelect vehicles*
- *Driving violations*
- *Derelect Bicycle*

L'attributo *Location Type* ha molti attributi nulli, circa 6 milioni. Essendo un attributo particolarmente utile a livello informativo è utile conservarlo e non perdere questo contenuto ai fini dell'analisi. Per questo motivo si è cercato di comprendere come poter utilizzare anche gli attributi nulli così da poterli utilizzare al meglio. Aiutandoci con il rispettivo attributo *Descriptor* spesso si è riusciti ad estrapolare il tipo di luogo corrispondente alla segnalazione. In più molti elementi già etichettati rappresentavano lo stesso concetto, ma questo era espresso attraverso sostantivi simili. Per questi attributi si è scelto di assegnare un singolo valore che descrivesse il luogo in maniera corretta. Diversamente gli elementi a cui non si è potuto assegnare un tipo di luogo sono stati identificati con "Unspecified". Ottimizzare questo attributo ci ha permesso di ottenere 17 tipologie di luogo differenti, qui schematizzati:

- Street/Sidewalk
- Residential Building/House
- Store/Commercial
- Club/Bar/Restaurant
- Park
- School
- Vacant Building
- House of Worship
- Pool/Beach/Sport center
- Taxi
- Kennel/Animal Shelter
- Hospital
- Office Building
- Pier
- Senior Center
- Petting Zoo/Animal Exhibit
- Unspecified

A questo punto dell'analisi l'attributo *Descriptor* essendo molto sparso ed avendolo utilizzato spesso per ottimizzare gli attributi già citati in questo paragrafo si è scelto di rimuoverlo dal dataset.

Successivamente, come fatto in precedenza per il dataset sulle segnalazioni nella città di Torino, si è arricchito il dataset di informazioni temporali che saranno poi utili per trovare interessanti relazioni tra i dati. Anche per questo dataset è stato utilizzato l'attributo relativo alla data *Created Date* per estrarre nuovi attributi temporali. In primo luogo è stato aggiunto l'attributo relativo all'anno, opportunamente identificato come *Year*. Si è verificato poi che le stagioni per la città di New York coincidessero con i periodi stagionali italiani, e si sono scelti quindi gli stessi intervalli di tempo usati per il dataset precedentemente analizzato, e si è quindi inserito nel dataset l'attributo *Season*.

Primavera	Dal 21/03 al 20/06
Estate	Dal 21/06 al 22/09
Autunno	Dal 23/09 al 20/12
Inverno	Dal 21/12 al 20/03

In seguito si è ricavato dalla data il mese e il giorno della settimana corrispondente, rispettivamente indicati nel dataset come *Month* e *Week day*. Oltre a ciò si è deciso di inserire ai fini dell'analisi un attributo che indica se il giorno preso in esame è Feriale oppure Festivo. Si è quindi aggiunto un attributo *Weekday/Holiday* per indicare ciò. Oltre al giorno settimanale Domenica considerato come festivo, sono state ricercate le festività celebrate nella città di New York. Da questo sito [14] è stato possibile estrapolare le ricorrenze festive per ogni anno che sono state identificate ed etichettate correttamente, queste vengono qui riportate:

- Domenica
- 1 Gennaio, Capodanno
- Giorno di Martin Luther King, viene celebrato il terzo lunedì di Gennaio.
- Giorno del Presidente, si celebra il terzo lunedì di Febbraio.
- Giorno della Memoria, si celebra l'ultimo lunedì di Maggio.
- 19 Giugno, Juneteenth
- 4 Luglio, giorno dell'Indipendenza.
- Giorno del Lavoro, primo lunedì di Settembre.
- Giorno della Scoperta dell'America, viene celebrato il secondo lunedì di Ottobre.

- Election Day, martedì successivo al primo lunedì del mese di Novembre
- 11 Novembre, giorno del Veterano.
- Giorno del Ringraziamento, si celebra l'ultimo giovedì di Novembre.
- 25 Dicembre, Natale

Il dataset opportunamente sottoposto a preprocessing presenta 11 attributi e risulta così strutturato:

- Created Date
- Complaint Category
- Complaint Type
- Location Type
- Borough
- Community Board
- Year
- Month
- Season
- Week day
- Weekday/Holiday

Adesso verrà messo in risalto il contenuto del dataset dopo aver svolto la fase di preprocessing, osservando le distribuzioni di ciascun attributo con il fine di trarre alcune prime considerazioni iniziali e valutare l'andamento di alcuni di questi lungo gli elementi geografici e temporali. Il numero totale di segnalazioni è di circa 24,4 milioni, e comprendono gli anni che vanno dal 2010 al 2021.

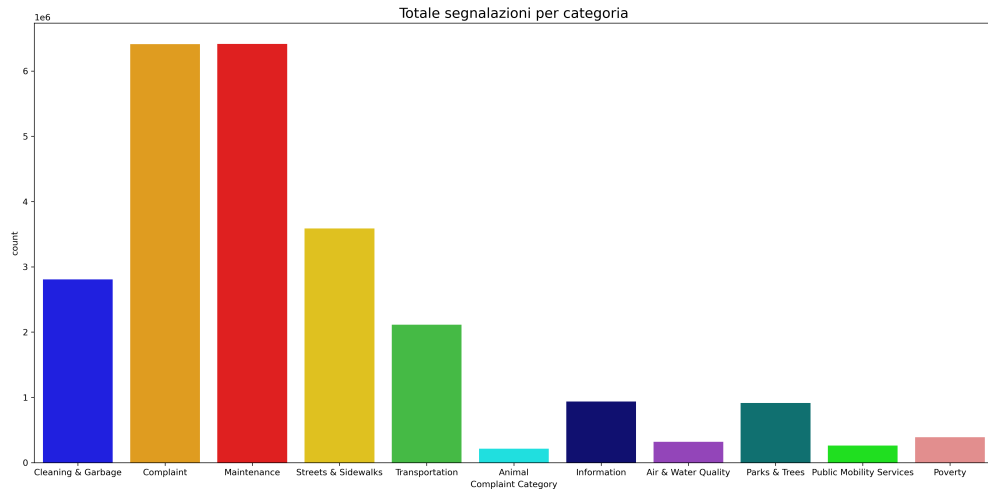


Figura 4.17: Distribuzione delle segnalazioni per Categoria

Dal grafico in figura 4.17 si osserva che le categorie a cui sono associate un maggior numero di segnalazioni sono *Maintenance* e *Complaint* con circa 6.4 milioni di richieste. Poi troviamo in ordine *Streets & Sidewalks*, *Cleaning & Garbage* e *Transportation*, rispettivamente 3.5 milioni, 2.8 milioni e 2.1 milioni. Le categorie *Information* e *Parks & Trees* hanno occorrenze nell'ordine di circa 937 mila e 914 mila segnalazioni. La categoria *Poverty* ha pressappoco 391 mila richieste, mentre la categoria *Air & Water Quality* presenta intorno alle 319 mila segnalazioni. Ci sono circa 260 mila segnalazioni etichettate come *Public Mobility Services*, la categoria meno frequente all'interno del dataset è *Animal* con solo 214 mila segnalazioni.

Per quanto riguarda le 60 sottocategorie sono così distribuite:

- **Air & Water Quality**
 - *Air quality complaint*: 158958
 - *Request Water Test Kit*: 74332
 - *Water quality complaint*: 63579
 - *Asbestos complaint*: 22298
- **Animal**
 - *Animal control/care*: 126717
 - *Dead animal*: 88260
- **Cleaning & Garbage**
 - *Bulky item collection*: 328449

- *Garbage*: 564900
- *Graffiti*: 188066
- *Illegal dumping*: 73255
- *Hazardous waste*: 105113
- *Missed Recycling collection*: 383962
- *Sewage/Storm water*: 460809
- *Unsanitary condition*: 669225
- *Weeds*: 33596

- **Complaint**

- *Residential noise*: 2582250
- *Street noise*: 2310821
- *Nuisance complaint*: 319523
- *Consumer complaint*: 217446
- *Illegal building*: 207102
- *Smoking*: 77766
- *Mold*: 246780
- *Rodent*: 292836
- *Illegal Posting/Signs*: 63888
- *Aircraft/Helicopter noise*: 48582
- *Drinking*: 20227
- *Drug Activity*: 16567
- *Vaccine complaint*: 6132
- *Property Violations*: 665

- **Information**

- *Information/Request*: 113173
- *DOF Property*: 322511
- *Payment Problem*: 229059
- *License*: 158403
- *Inspection*: 94535
- *Lost and Found*: 19132

- *Refund request*: 485
- **Maintenance**
 - *Heat and Hot Water system*: 2144493
 - *Construction and Plumbing*: 1554370
 - *Building repair*: 1167080
 - *Water system*: 950842
 - *Electrical*: 437182
 - *Elevator*: 139858
 - *Vacant Property*: 19933
- **Parks & Trees**
 - *Tree condition*: 711972
 - *New Tree Request*: 183380
 - *Park maintenance*: 19602
- **Poverty**
 - *Homeless/Encampment*: 237178
 - *Housing*: 145728
 - *Panhandling*: 8164
- **Public Mobility Services**
 - *Taxi complaint*: 181940
 - *Shared vehicle complaint*: 181940
- **Streets & Sidewalks**
 - *Street/Sidewalks repair*: 1252985
 - *Street/Road/Sidewalk obstruction*: 1215234
 - *Street light repair*: 558731
 - *Street/Road sign repair*: 464877
 - *Snow or Ice removal*: 97206
- **Transportation**
 - *Illegal Parking*: 1424434
 - *Derelect vehicles*: 581178
 - *Driving violations*: 91662
 - *Derelect Bicycle*: 16331

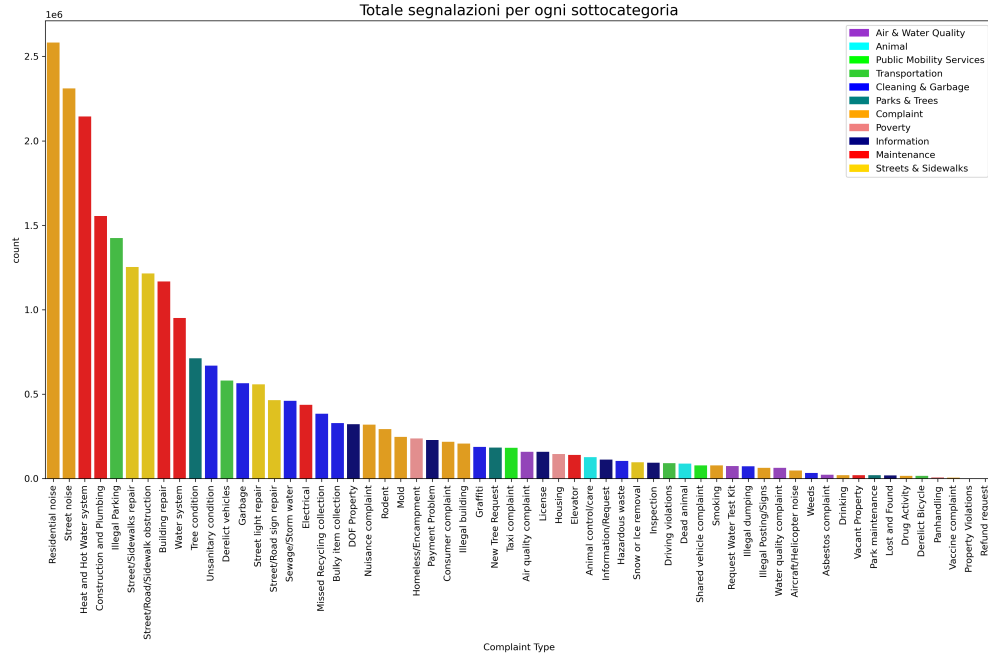


Figura 4.18: Distribuzione delle segnalazioni per Complaint Type

Se ispezioniamo le categorie maggiormente frequenti troviamo per Complaint la sottocategoria *Residential noise* e *Street noise*, rispettivamente con all'incirca 2,5 milioni e 2,3 milioni di segnalazioni appartenenti a questa tipologia. Mentre per la categoria Maintenance si registrano indicativamente 2,1 milioni di segnalazioni etichettate come *Heat and Hot Water system* e circa 1,5 milioni appartenenti a *Construction and Plumbing*. Il tipo di sottocategoria *Illegal Parking* risulta essere la più frequente tra la categoria Transportation con circa 1,4 milioni. La categoria Streets & Sidewalks presenta relativamente alla propria tipologia un numero simile di segnalazioni per le sottocategorie più rappresentative, che risultano essere indicativamente con 1,2 milioni di segnalazioni: *Street/Sidewalks repair* e *Street/-Road/Sidewalk obstruction*.

Le segnalazioni meno frequenti relativamente all'attributo *Complaint Type* riguardano *Property Violations* (665) e *Refund request* (485), appartenenti rispettivamente alle categorie Complaint e Information. Il grafico in figura 4.18 illustra la distribuzione in ordine decrescente dell'attributo *Complaint Type* nel dataset.

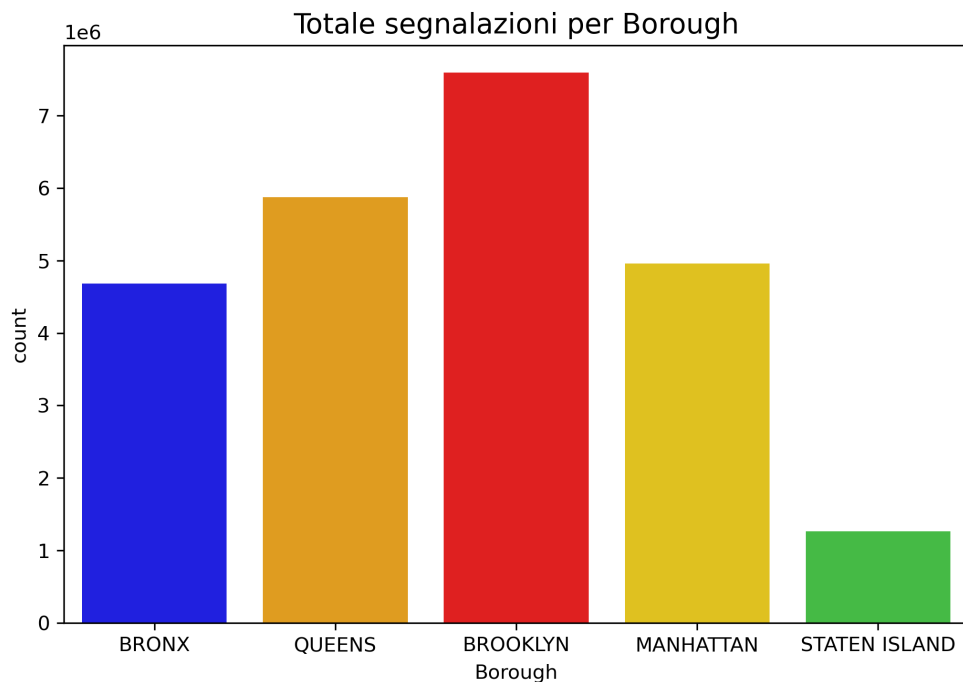


Figura 4.19: Distribuzione delle segnalazioni nei Distretti

Nel dataset troviamo un attributo geografico che indica il distretto della città in cui viene fatta la segnalazione, *Borough*.

I distretti nella città di New York sono 5:

- Bronx
- Brooklyn
- Manhattan
- Staten Island
- Queens

Dalla figura 4.19 possiamo vedere la distribuzione di questi attributi nel dataset; il distretto con più segnalazioni è quello di Brooklyn con approssimativamente 7,5 milioni, a seguire c'è il distretto di Queens con all'incirca 5,8 milioni, poi Manhattan con circa 4,9 milioni e Bronx con circa 4,6 milioni. Il distretto di Staten Island è quello che registra meno segnalazioni, indicativamente 1,2 milioni.

Altro attributo territoriale che identifica delle zone precise all'interno della città di New York è *Community Board*, esso rappresenta una ulteriore divisione all'interno di ciascun distretto. In totale se ne registrano 71 diversi all'interno del set di dati. Nel grafico in figura 4.20 si può apprezzare la distribuzione delle segnalazioni per *Community Board*. Le maggiori segnalazioni sono registrate nel 12 Manhattan community board, per un totale di 833629 richieste di non emergenza. Successivamente troviamo 12 Queens e 12 Bronx. Possiamo apprezzare che a partire da questi due si possono via via identificare dei gruppi in cui il numero di segnalazioni risulta indicativamente simile. Risultano esserci delle community board che hanno invece un numero veramente basso di segnalazioni, queste si trovano all'interno di un determinato distretto ma sono al di fuori della competenza governativa e giuridica dei singoli distretti comunitari; vengono chiamate Joint Interest Areas (JIA).

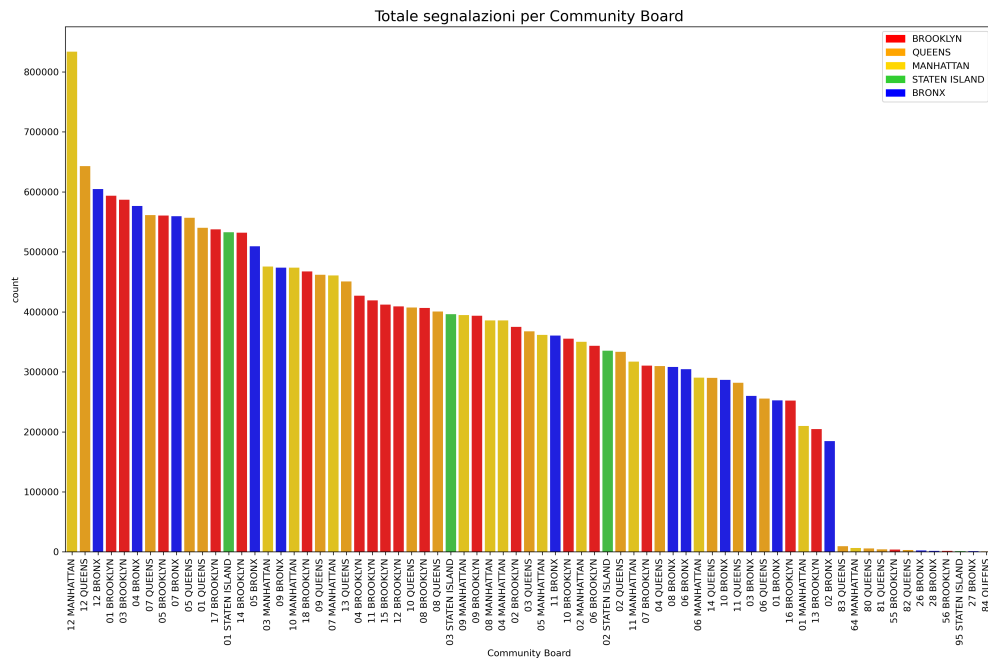


Figura 4.20: Distribuzione delle segnalazioni nei vari Community Board

L'attributo *Location Type* che indica la tipologia di luogo in cui è stata fatta la segnalazione registra il maggior numero di occorrenze per il luogo identificato come Street/Sidewalk con circa 11 milioni. Poi troviamo Residential Building/House con indicativamente 9,6 milioni e il tipo Unspecified registra 2,2 milioni. Molto più bassa la frequenza di luoghi come Store/Commercial (750715), Club/Bar/Restaurant (351491) e Park (254693). All'incirca 25 mila segnalazioni indicate in luoghi scolastici (School), e circa 15 mila in luoghi religiosi (House of Worship). Le altre tipologie specifiche di questo set di attributo presentano valori piuttosto bassi.

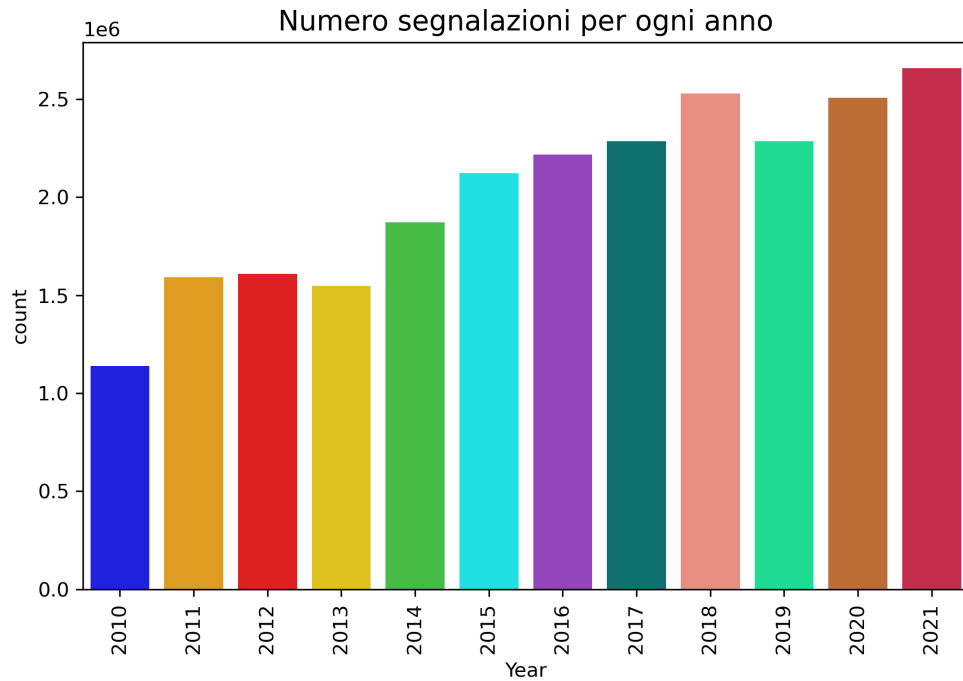


Figura 4.21: Distribuzione delle segnalazioni nei vari anni

Dal grafico in figura 4.21 possiamo notare l'andamento delle segnalazioni con il passare degli anni. Ne deduciamo che con il passare del tempo i cittadini hanno maggiormente utilizzato questo servizio per segnalare delle problematiche urbane, infatti dal grafico si osserva l'andamento crescente del numero di segnalazioni nel corso degli anni. Solo nel 2019 c'è stata una diminuzione di segnalazioni registrate rispetto all'anno precedente se si prendono in considerazione gli anni più recenti. Si è passati da circa 1,1 milioni di segnalazioni nell'anno 2010 a indicativamente 2,6 milioni di richieste di non emergenza nell'anno appena trascorso, che risulta anche il valore più alto di segnalazioni registrate relativamente all'attributo anno.

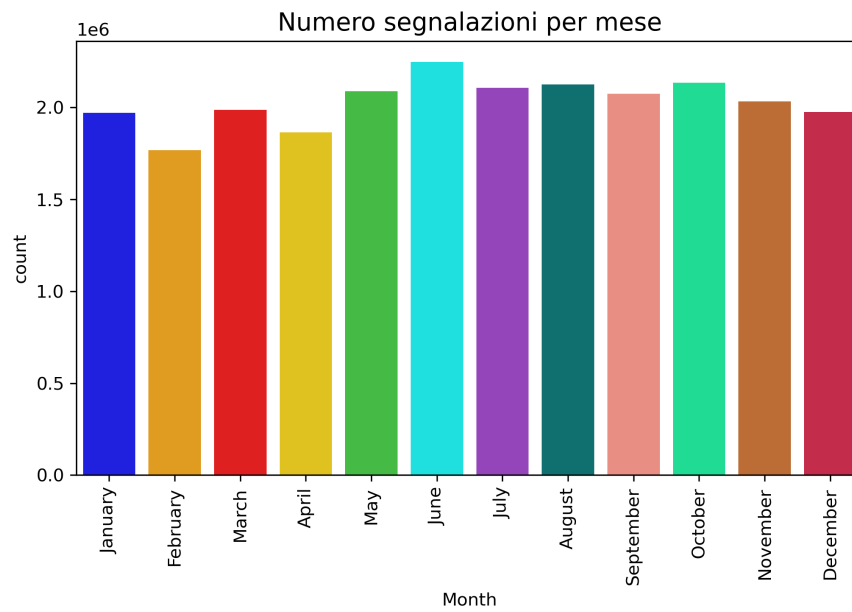


Figura 4.22: Distribuzione delle segnalazioni durante i mesi

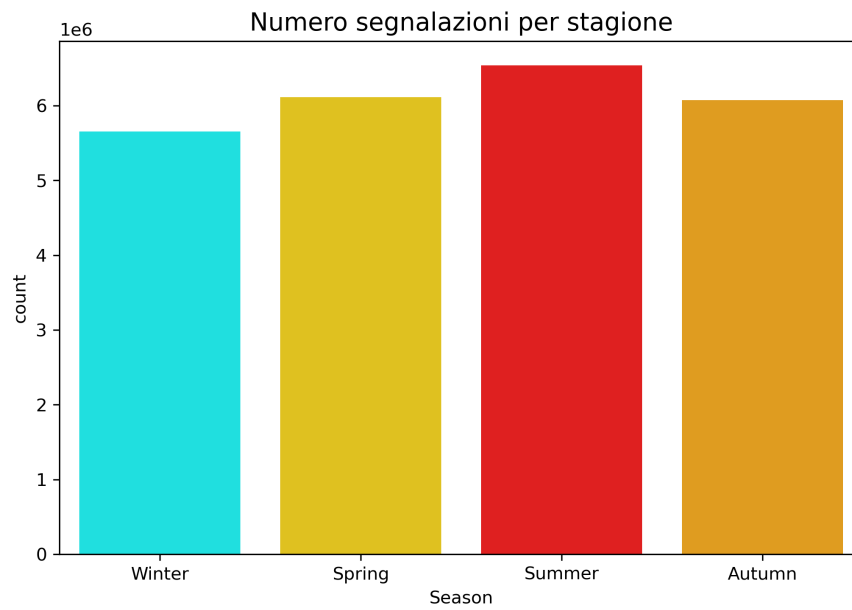


Figura 4.23: Distribuzione delle segnalazioni durante le stagioni

Se prendiamo in considerazione l'attributo *Month* e ne osserviamo l'andamento del numero di segnalazioni riscontriamo una distribuzione piuttosto equa durante i mesi. Il mese in cui si registra il maggior numero di segnalazioni è Giugno con circa 2,2 milioni, contrariamente Febbraio con indicativamente 1,8 milioni di richieste è il mese in cui si effettuano meno segnalazioni.

Di conseguenza anche il numero di segnalazioni durante i mesi dell'anno risultano equamente distribuite. Dalla figura 4.23 è messo in evidenza l'andamento dell'attributo *Season*, la stagione estiva risulta il periodo in cui sono effettuate una maggiore quantità di richieste, all'incirca 6,5 milioni. Mentre la stagione invernale registra un numero più basso rispetto alle altre stagioni, 5,6 milioni in modo approssimativo.

Risulta interessante osservare anche la distribuzione nel corso dei giorni settimanali, per questo il grafico in figura 4.24 prende in considerazione l'attributo *Week day* ed evidenzia l'andamento del numero di segnalazioni nell'arco temporale indicato. Particolare il fatto che nel corso dei giorni "lavorativi" l'andamento è piuttosto simile, mentre si nota poi una diminuzione nel corso del weekend. Infatti il sabato e la domenica sono registrate meno segnalazioni rispetto ai giorni precedenti. Domenica è il giorno con meno segnalazioni effettuate, circa 2,9 milioni. Martedì è invece il giorno in cui se ne registrano di più, orientativamente 3,8 milioni.

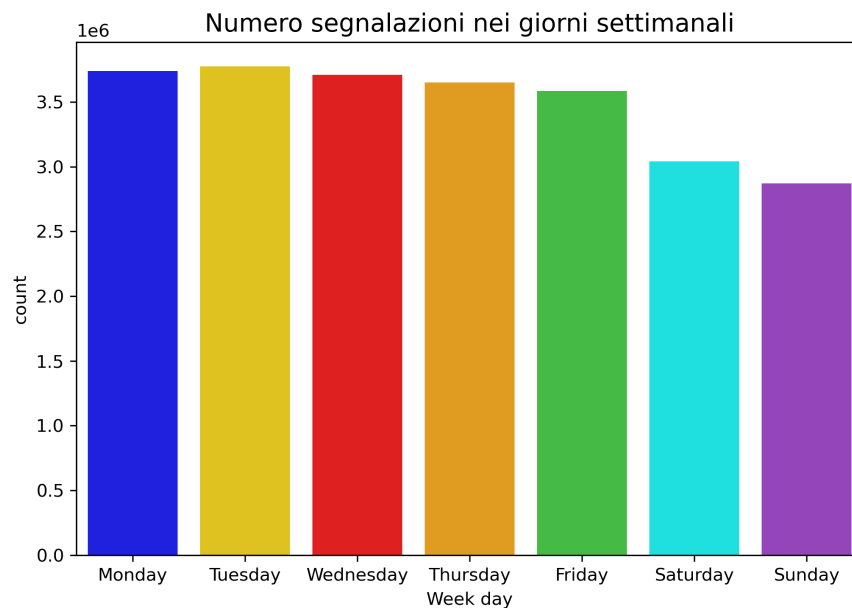


Figura 4.24: Distribuzione delle segnalazioni durante i giorni della settimana

Il grafico successivo presenta la divisione tra giorno feriale e festivo, indicando la quantità di segnalazioni per ciascun tipo. Nel dataset ci sono circa 21 milioni di segnalazioni effettuate nei giorni feriali, mentre 3,6 milioni etichettate nei giorni festivi.

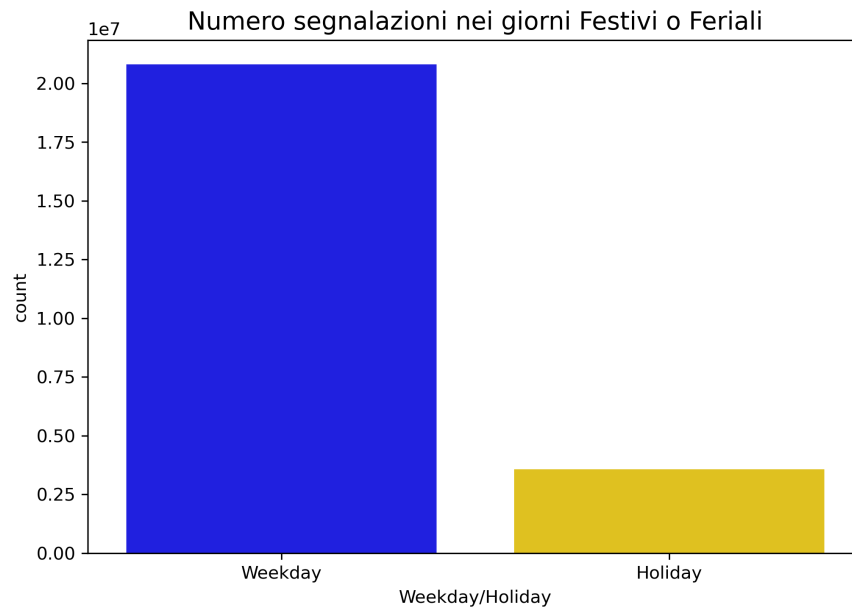


Figura 4.25: Distribuzione delle segnalazioni durante i giorni Festivi e Feriali

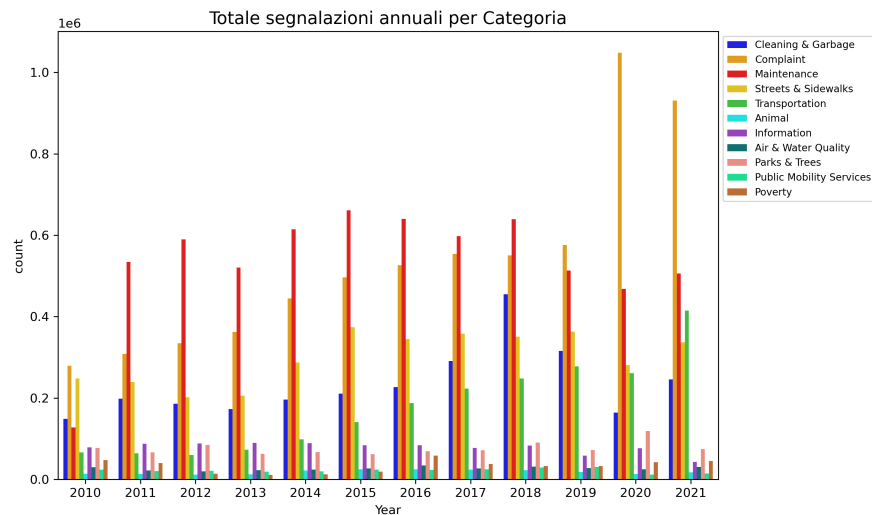


Figura 4.26: Distribuzione delle segnalazioni negli anni per categoria

Osservando il grafico in figura 4.26 si evince come le categorie sono distribuite lungo gli anni. Da questo è possibile rilevare spunti interessanti, permette di capire come è cambiato l'andamento di ciascuna categoria nel corso del tempo, per esempio dal 2011 al 2018 le maggiori segnalazioni erano associate alla categoria Maintenance, mentre negli anni più recenti, dal 2019 al 2021, si è assistito ad un rapido aumento di segnalazioni relative alla categoria Complaint, che risulta la più frequente. Nell'ultimo anno appena trascorso si segnala anche un aumento delle segnalazione per la categoria Transportation. Invece la categoria Cleaning & Garbage che aveva fatto registrare nell'anno 2018 un elevato numero di segnalazioni, negli ultimi due anni la frequenza sembra diminuire.

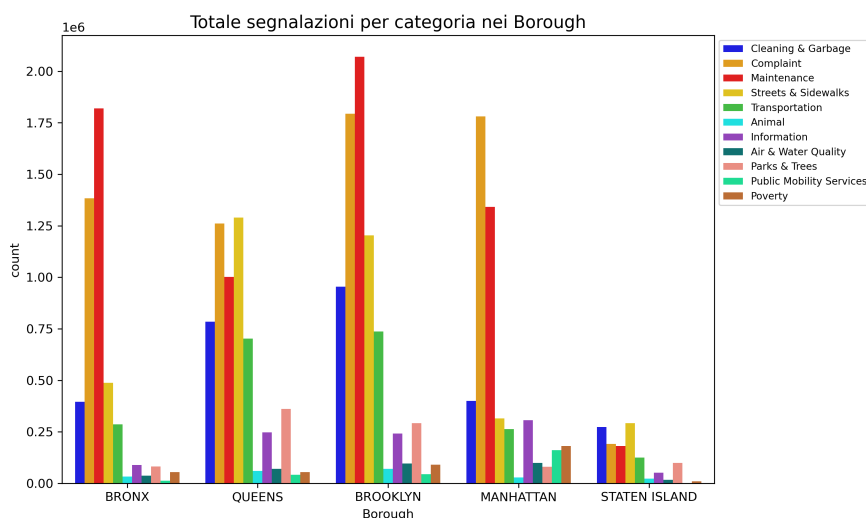


Figura 4.27: Distribuzione delle segnalazioni nei Distretti per categoria

Molto interessante analizzare l'andamento distributivo delle categorie nei vari distretti della città di New York, il grafico 4.27 permette di fare delle valutazioni in merito. I distretti di Bronx e Brooklyn registrano le maggiori segnalazioni relativamente alla categoria Maintenance, mentre nel distretto di Manhattan si effettuano più segnalazioni per la categoria Complaint. Nella zona di Queens invece molte sono le segnalazioni associate alla categoria Streets & Sidewalks, in aggiunta questa categoria registra in questo distretto il maggior numero di segnalazioni, questa è seguita con valori pressapoco uguali dalla categoria Complaint. Questo distretto è anche quello a cui si associano il maggior numero di segnalazioni per la categoria Parks & Trees. Mentre nel distretto Staten Island risultano più frequenti le segnalazioni in ambito Streets & Sidewalks e Cleaning & Garbage. Questa ultima categoria registra le maggiori segnalazioni nel distretto di Brooklyn. Interessante anche la categoria Poverty che risulta maggiormente distribuita nel distretto di Manhattan.

Nel corso delle stagioni possiamo vedere che durante il periodo estivo e primaverile si registrano maggiori segnalazioni attribuite alla categoria Complaint, mentre nei periodi autunnali e invernali sono effettuate un numero superiore di segnalazioni per la categoria Maintenance. La categoria Streets & Sidewalks ha una distribuzione piuttosto uniforme nel corso di ogni singola stagione, come anche le categorie Transportation e Cleaning & Garbage. Durante l'estate si registra anche un aumento del numero di segnalazioni relativamente alla categoria Parks & Trees.

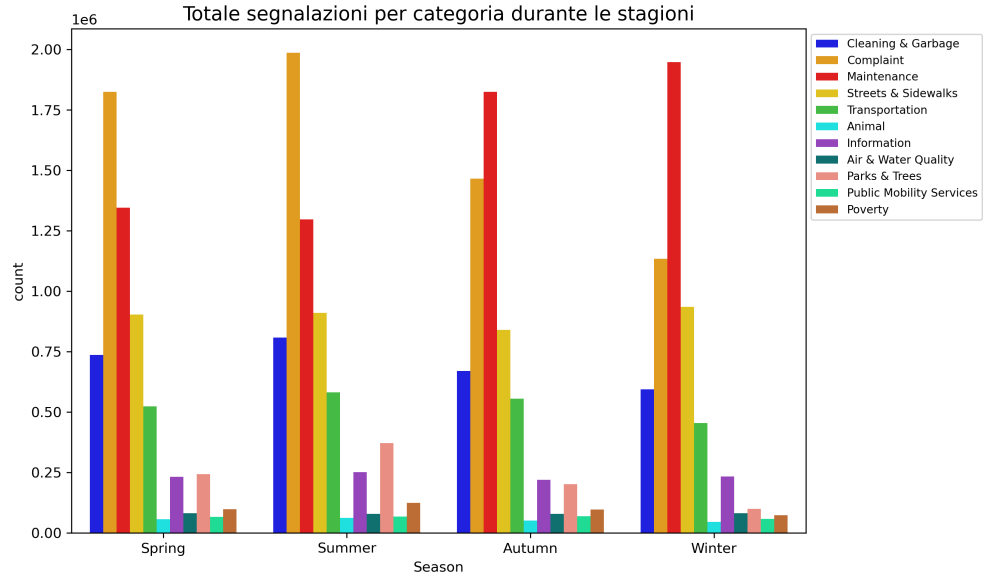


Figura 4.28: Distribuzione delle segnalazioni per categoria nelle stagioni

Nel grafico in figura 4.29 possiamo osservare come cambiano le segnalazioni a livello di categoria nel corso della settimana. Particolare è la considerazione che evidenzia la maggior frequenza delle categoria Maintenance durante i giorni che vanno dal Lunedì al Venerdì, mentre nelle giornate di Sabato e Domenica si effettuano più segnalazioni di tipo Complaint. Diminuiscono anche durante il weekend il numero di segnalazioni per le categorie Cleaning & Garbage e Streets & Sidewalks.

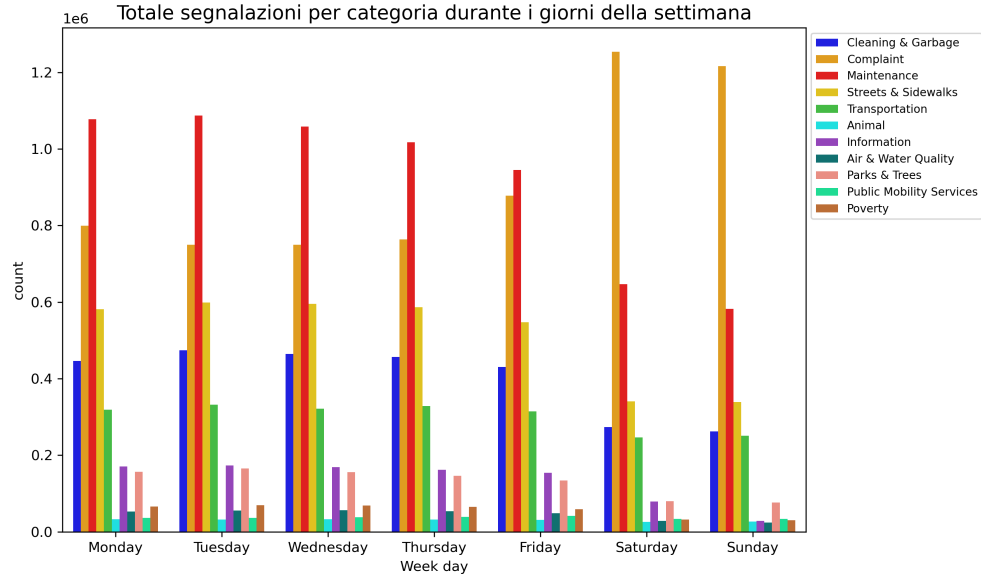


Figura 4.29: Distribuzione delle segnalazioni per categoria nei giorni settimanali

4.3 Analogie e Differenze

Conclusa la fase di preprocessing e di esplorazione dei dataset finali, si può fare un piccolo sunto considerando differenze e analogie emerse. Innanzitutto una delle principali differenze sta nel contenuto quantitativo dei due dataset. Nonostante ciò, partendo dalla struttura del dataset della città di Torino si è cercato di creare una simile struttura gerarchica anche per il dataset americano, provando a renderlo il più possibile simile. Si è mantenuta l'organizzazione a due livelli categorici per identificare il contesto e il tipo di segnalazione. La differenza principale è che la gerarchia creata per le segnalazioni di non emergenza della città di New York risulta più dettagliata e vasta. È stato mantenuto l'assetto geografico-temporale attraverso i vari attributi presenti nel dataset. A livello temporale gli attributi presenti sono identici, mentre gli attributi spaziali riportano la divisione territoriale delle due città. New York presenta due attributi geografici, Torino solamente quello relativo alla circoscrizione. Inoltre mentre per il dataset di Torino troviamo un attributo che identifica se il luogo della segnalazione è un area verde o meno, per il dataset americano abbiamo un attributo che identifica, se possibile, il luogo della segnalazione in maniera precisa.

Capitolo 5

Sezione sperimentale

In questo capitolo si descrive la soluzione proposta per lo svolgimento di questo progetto in cui vengono usate determinate tecniche di data mining per poter estrarre un prezioso contenuto informativo. Si estraggono regole associative attraverso due differenti algoritmi e si analizzano le regole generate valutandole in base a delle opportune metriche oggettive. Infine si proverà a dare anche un'interpretazione soggettiva a tali risultati.

5.1 Struttura pattern delle regole associative

L'obiettivo di questa analisi è di estrarre delle regole associative che riescano a fornire delle informazioni su questo tipo di segnalazioni in ambito urbano. Avendo a disposizione degli attributi di due tipi: spaziale e temporale, si è scelto di ottenere delle regole che avessero come premessa o conclusione il tipo di segnalazione (categoria o sottocategoria), così da trovare una correlazione tra il tipo di categoria e/o sottocategoria e l'informazione spaziale, temporale o entrambe.

Nel dettaglio si è interessati a regole strutturate in questo modo:

- Categoria o Sottocategoria $\rightarrow$ Luogo
- Categoria o Sottocategoria $\rightarrow$ Tempo
- Categoria o Sottocategoria $\rightarrow$ Luogo, Tempo
- Luogo $\rightarrow$ Categoria o Sottocategoria
- Tempo $\rightarrow$ Categoria o Sottocategoria
- Luogo, Tempo $\rightarrow$ Categoria o Sottocategoria

Dove per Luogo si intendono gli attributi geografici contenuti nei dataset, mentre per Tempo gli attributi temporali presenti nei dataset. In questo modo si generano delle regole che possono risultare molto interessanti da un punto di vista del contenuto estratto, infatti potrebbero aiutarci a capire cosa per esempio succede in un luogo, oppure data una certa segnalazione quando o dove potrebbe verificarsi un determinato evento. Per ottenere quanto detto si danno in input agli algoritmi parti di dataset, prima si estraggono le regole di associazione relative alla categoria di riferimento, successivamente quelle relative alla sottocategoria.

5.2 Generazione di regole di associazione

L'intero dataset viene suddiviso in due dataframe, uno per la categoria e uno per la sottocategoria, da dare in input all'algoritmo FP-Growth per il corretto svolgimento dell'analisi con il fine di ottenere regole associative come descritte nel precedente paragrafo. In primo luogo il dataset è stato strutturato adeguatamente attraverso l'uso di alcune librerie Python per essere opportunamente organizzato per la corretta esecuzione dell'algoritmo. Successivamente sono stati formati 6 dataframe, 3 in riferimento all'attributo categoria e 3 all'attributo sottocategoria, divisi rispettivamente per la tipologia di attributi in esso contenuti (spaziale, temporale e spaziale-temporale). Quindi si è poi scelto il supporto minimo per estrarre gli itemset frequenti grazie all'algoritmo FP-Growth, più è grande il valore del supporto più frequentemente l'itemset compare nel dataset. Nonostante questo, il valore scelto per il minimo supporto è stato 0.01, un valore piuttosto basso. Questa scelta permette di avere un ampio margine e di estrarre quindi un numero maggiore di regole associative, inoltre ci permette comunque di selezionare successivamente quelle ritenute più importanti. Di fatto scegliendo una soglia alta di supporto si corre il rischio di perdere regole che hanno una correlazione abbastanza forte ma frequenza bassa.

Quindi sono stati dati in input i vari dataframe e sono stati generati dall'algoritmo FP-Growth i relativi itemset frequenti sulla base del supporto scelto. Dopo sono state generate le regole di associazione da questi itemset frequenti, e sono state filtrate per antecedente e conseguente in base al nostro interesse e in base alle metriche qualitative a nostra disposizione. Sono state ricercate tutte le regole di associazione dove la categoria e/o sottocategoria era una premessa e poi, sono state filtrate le regole al contrario, ovvero dove la categoria e/o sottocategoria era la conclusione. Tutto ciò per ogni dataset in input contenente gli attributi spaziali, temporali o entrambi contemporaneamente.

Dopo la generazione delle regole è importante valutarne la qualità attraverso le metriche opportune presentate nel capitolo 2 che ci permettono di avere una analisi oggettiva delle informazioni estratte.

Oltre al supporto si utilizza la metrica confidence che ci fornisce una indicazione di quanto affidabile è la regola; maggiore sono i valori di supporto e confidence migliore è la regola di associazione. Oltre a queste due è stata utilizzata la metrica lift che indica in che misura antecedente e conseguente dipendono l'una dall'altra, ci dice quanto forte è la regola. Se il valore è maggiore di 1 allora gli itemset sono positivamente correlati, negativamente altrimenti. Se uguale a 1 le due ricorrenze risultano indipendenti.

Le regole generate in nostro possesso sono state poi filtrate in base al proprio valore di lift. Di fatto si sono scelte solo le regole con un lift maggiore di uno, in quanto è evidente che le correlazioni debbano essere tutte positive per avere un senso, e quindi non sono state considerate interessanti le correlazioni negative. Arrivati a questo punto si sono ordinate le regole estratte per lift e confidence, e si è cercato di interpretare i risultati ottenuti e individuare le associazioni ritenute più valide. Le regole verranno presentate sottoforma di grafico Heatmap, in cui verrà visualizzata nella rispettiva barra il valore di confidence, e la regola verrà colorata in base al valore del lift, alle regole con lift maggiore per quel determinato pattern si associa un colore più scuro.

5.3 Analisi regole associative estratte dalle segnalazioni di Torino

In questo paragrafo vengono messi in rassegna le regole associative relative al dataset sulle segnalazioni di non emergenza della città di Torino. In base ai tipi di pattern scelti si sono ottenute regole con contenuti informativi identici, infatti invertendo le regole, e quindi antecedente e conseguente, si ottengono metriche di lift e supporto uguali, ciò che cambia sono i valori ottenuti dalla metrica confidence in quanto quest'ultima essendo una probabilità condizionata cambia il rispettivo valore. Vengono presentati i risultati derivanti da dati di non emergenza quando una specifica zona o periodo o una data combinazione di essi sembra essere fortemente correlata con una data categoria o sottocategoria. Prima si riportano le regole in cui la categoria è il conseguente, successivamente quelle relative alla sottocategorie sono presentate inversamente. Queste regole possono essere sfruttate per evidenziare delle problematiche relativi a una specifica area urbana e in un determinato lasso di tempo.

Prese in esame le regole di tipo Contesto spaziale $\rightarrow$ Categoria, figura 5.1, notiamo subito che la categoria *Allarme Sociale* non compare tra le regole generate. Sono 5 le regole estratte, 4 appartenenti alla categoria Convivenza Civile, 1 a Qualità Urbana. Se si osservano le regole relative alla categoria Convivenza Civile troviamo correlazione con la circoscrizione 1 e 8, e con luoghi considerati non area verde. Per queste regole il supporto va da un minimo di 0.02 ad un massimo di 0.05. Il valore

di lift è piuttosto simile, approssimativamente 1.1. Mentre il valore della confidence va da un minimo di 0.57 ad un massimo di 0.60. I risultati ottenuti evidenziano che nella circoscrizione numero 1 e 8 si potrebbero verificare segnalazioni di tipo convivenza civile.

Per quanto riguarda invece l'unica regola relativa alla categoria Qualità Urbana denota una correlazione tra le aree verdi della città e la rispettiva categoria. Questa regola presenta valore di supporto di 0.024, confidence di 0.59, e valore di lift 1.31. Si deduce che esiste una relazione tra le aree verde della città di Torino e le segnalazioni nel contesto della qualità urbana.

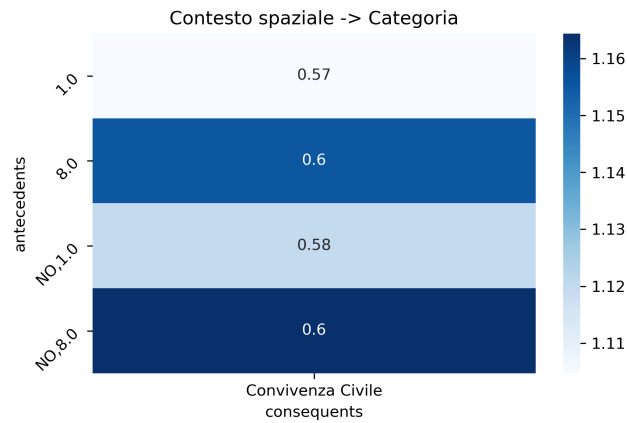


Figura 5.1: Regole associative Contesto spaziale $\rightarrow$ Categoria

Nella figura 5.2 possiamo apprezzare le 30 regole estratte di tipo Contesto temporale $\rightarrow$ Convivenza Civile. I valori di supporto non sono altissimi, si va da 0.01 a 0.037. Le seguenti regole hanno tutte lift di poco superiore a 1. Possiamo notare che gli anni presenti sono relativi al 2013 e 2016, ciò sta a significare che dati questi anni si sono verificati con una certa frequenza segnalazioni di tipo convivenza civile. Sono perlopiù presenti mesi relativi alla stagione estate, infatti troviamo maggiormente Giugno, Luglio e Agosto; inoltre si nota in diverse regole una correlazione tra questo tipo di segnalazioni e Primavera, Giugno, probabilmente ciò significa che c'è una correlazione con le prime settimane di Giugno. Interessanti alcune regole che riguardano l'associazione Luglio, Estate $\rightarrow$ Convivenza Civile. Questa regola sottolinea che questi tipi di segnalazione potrebbero verificarsi in periodi estivi dell'anno in particolare durante il mese di Luglio. Altre regole generate sono relazionate con alcuni giorni settimanali, Lunedì, Mercoledì, Venerdì. Per esempio la regola Lunedì, Giugno $\rightarrow$ Convivenza Civile che denota la possibilità di verificarsi di una determinata categoria di segnalazioni nel periodo di Giugno ed in particolar modo nel giorno di Lunedì.

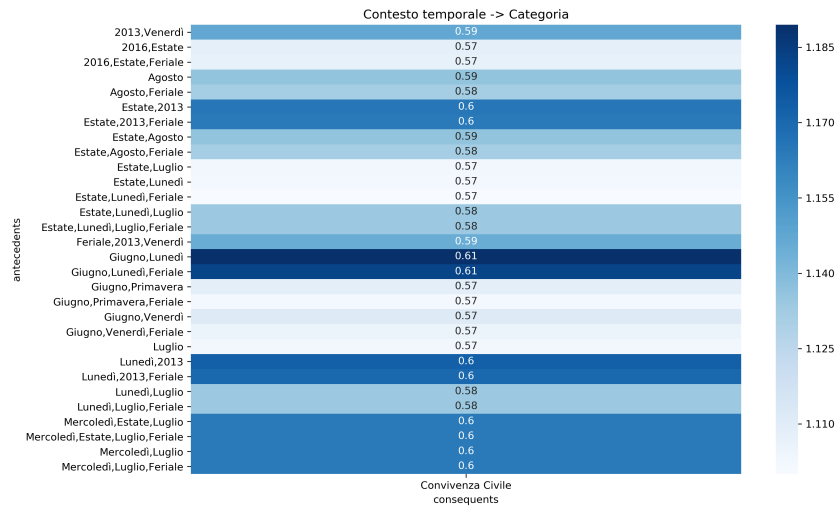


Figura 5.2: Regole associative Contesto temporale $\rightarrow$ Categoria Convivenza Civile

Per quanto riguarda la categoria Qualità Urbana vengono generate 33 regole che sono presenti nella figura 5.3. La prima è riferita alla tipologia contesto temporale $\rightarrow$ categoria. Sono presenti gli anni 2019 (12 regole), 2018 (6 regole), 2014 (1 regola) e 2012 (4 regole). Mentre come mesi troviamo Gennaio (4) e Marzo (2). Si possono notare 10 correlazioni con la stagione invernale, 2 per la stagione Autunno e 5 nella stagione Primavera. Troviamo correlazione in 3 giorni settimanali, Lunedì, Mercoledì e Venerdì. Nelle seguenti regole abbiamo valori di supporto non molto alti, si va da 0.01 a 0.056. La regola con il valore più alto di lift è 1.33, quella con il

valore più basso è 1.1. Se si prende in considerazione il valore della confidence varia da un minimo di 0.5 ad un massimo di 0.6. In linea generale possiamo affermare che questa categoria di segnalazioni risulta maggiormente presente negli ultimi anni in esame, 2018 e 2019, quindi si può affermare che con molta probabilità vi è stato un incremento nel corso degli anni di questo tipo di segnalazioni. Inoltre i mesi "freddi" dell'anno sono quelli in cui queste problematiche si presentano maggiormente.

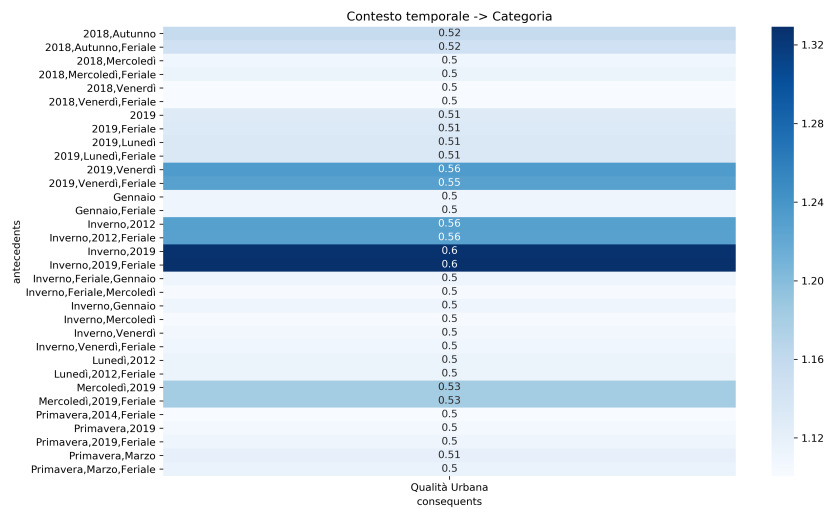


Figura 5.3: Regole associative Contesto temporale → Categoria Qualità Urbana

Dopo si è esplorata la tipologia di regola associativa che contiene sia il contesto temporale e il contesto spaziale. Nella figura 5.4 si è vista la tipologia di regola riferita al contesto spaziale e temporale → Categoria convivenza civile. Sono state generate 100 regole, tra queste si rimarca che la circoscrizione 1 e 8 sono strettamente correlate a questa categoria di segnalazione. Le metriche presentano valori indicativamente simili a quanto trovato precedentemente. Risulta interessante che le circoscrizioni numero 8 e 1 sono correlate a questa categoria insieme a 3 stagioni: Estate, Primavera ed Autunno. Anche se maggiore interesse è dato dal fatto che ben 27 regole trovano una correlazione tra la stagione estiva e un certo insieme di attributi spaziali con la categoria esaminata, mentre solo 7 con Autunno e 6 con Primavera. I valori di confidence si attestano nel range 0.56-0.66.

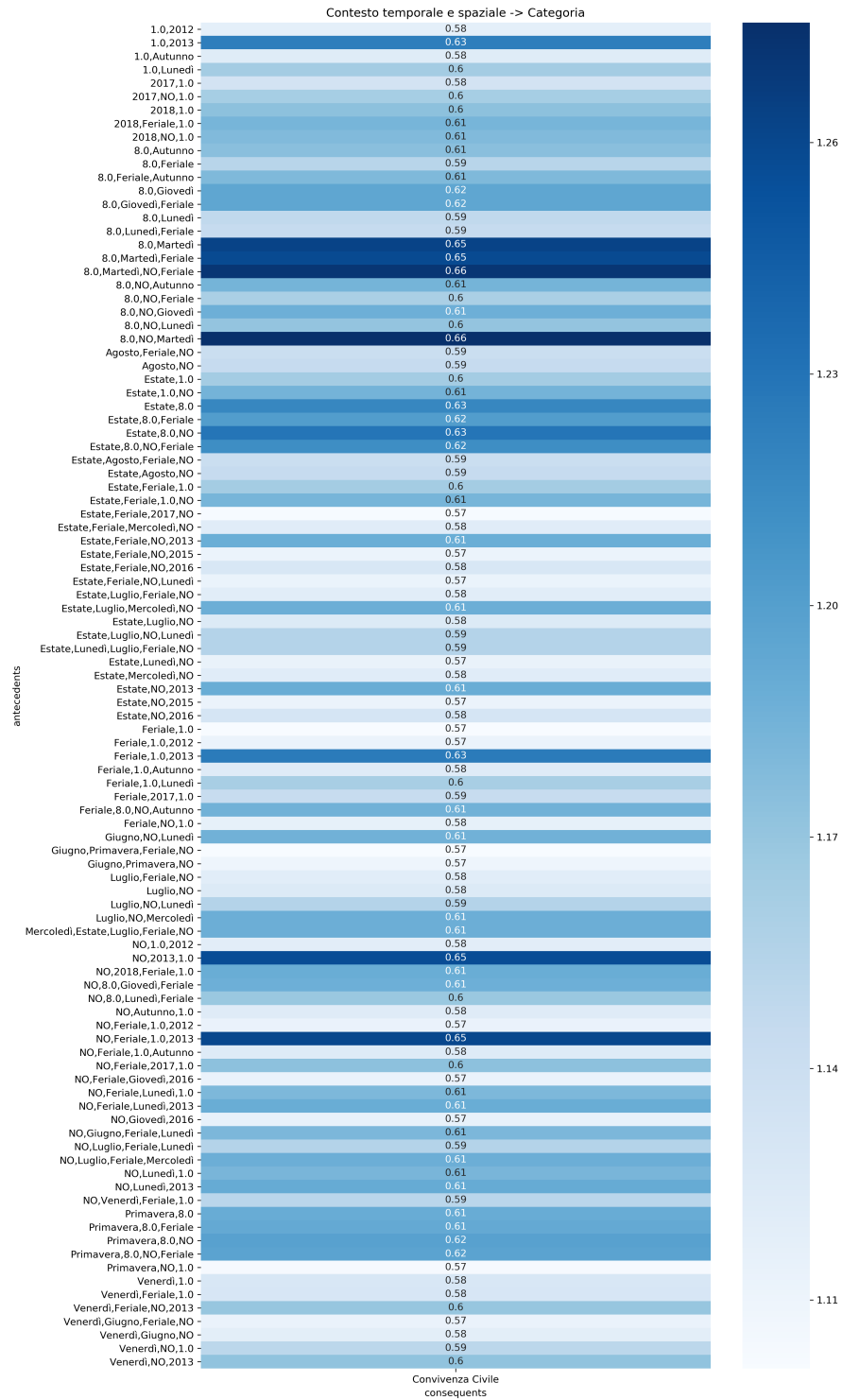


Figura 5.4: Regole associative Contesto temporale e spaziale → Categoria Convivenza Civile

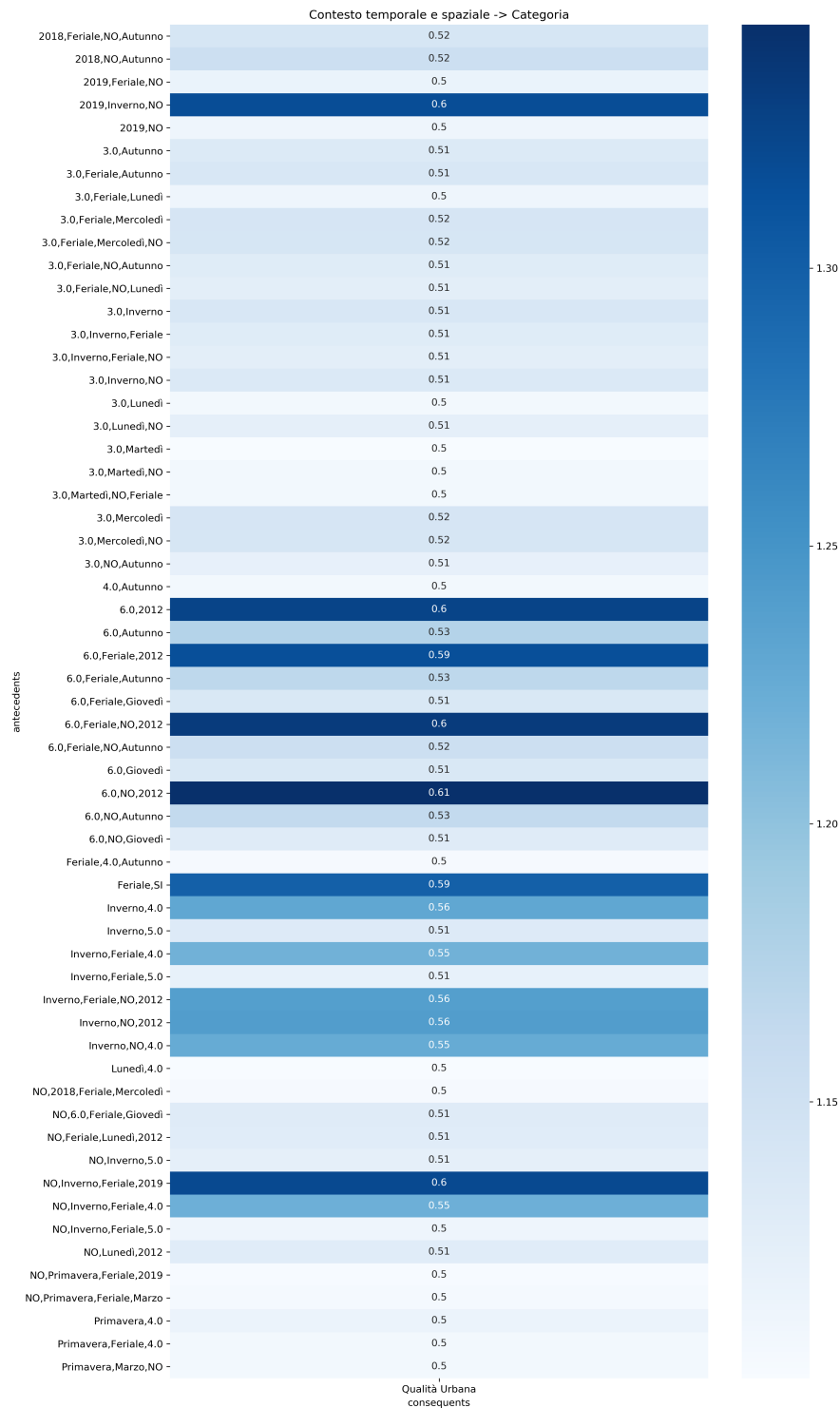


Figura 5.5: Regole associative Contesto temporale e spaziale → Categoria Qualità Urbana

La figura 5.5 si riferisce alla tipologia di regole contesto temporale e spaziale → Qualità urbana. Queste regole sottolineano la relazione tra le stagioni Inverno e Autunno e tale categoria. Inoltre in queste regole si trovano più attributi geografici correlati, in particolare le circoscrizioni 3, 4, 5 e 6. Deduciamo che ci sono delle zone di Torino dove si potrebbero verificare questi tipi di segnalazione, e addirittura in una data finestra temporale. Questo tipo di considerazioni sono ricavate da regole come Autunno,6 oppure Inverno, 3 → Qualità Urbana. Le regole citate sopra raggiungono valori di confidence rispettivamente del 53% e del 51%. Interessante anche l'informazione sul giorno settimanale in cui la maggior parte delle segnalazioni di tipo qualità urbana sono state ricevute in determinate zone, di Mercoledì nella circoscrizione 3, di Martedì e Giovedì nella 8, di Lunedì nella 1. Come si può notare non è mai presente la categoria Allarme sociale nelle regole estratte, probabilmente il motivo è dovuto al numero relativamente basso di segnalazioni associate a tale categoria. Inoltre non si riescono ad estrarre correlazioni per quanto riguarda l'attributo relativo all'indirizzo della segnalazione. Si è anche constatato che non siamo riusciti ad estrarre informazioni su tutte le circoscrizioni di Torino.

Successivamente è stato dato in input il dataset relativo alla sottocategoria, possiamo apprezzare che scendendo di livello gerarchico le regole generate risultano qualitativamente migliori ed inoltre riusciamo a dettagliare meglio le motivazioni che hanno portato i cittadini a fare certe segnalazioni. Anche queste regole sono state valutate prima in maniera oggettiva tenendo conto dei valori delle misure usate, e dopo si è cercato di dare una interpretazione in base all'utilità dell'informazione estratta.

La figura 5.6 si riferisce alla tipologia di regole generate Sottocategoria Convivenza Civile → Contesto spaziale, possiamo osservare nonostante valori di supporto non particolarmente alti che si ottengono comunque 8 regole interessanti con valori di confidence tra 0.16 e 0.22, e lift minimo 1.11 e massimo 2. Le sottocategorie presenti sono Disturbi da locali, Rumori molesti e Altro(C). Particolarmente interessante il fatto che nel 17% dei casi la sottocategoria Disturbi da locali è relazionata alla circoscrizione 8 e ad un posto che non è una area verde con un valore di lift uguale a 2. Questo risulta interessante in quanto proprio il quartiere San Salvario della circoscrizione 8 è tra i luoghi della movida torinese, e non sono rari i frastuoni, specialmente durante il weekend. Ma troviamo anche ulteriori tipologie di correlazioni, precisamente le sottocategorie Disturbi da locali, Rumori Molesti e Altro (C) si potrebbero verificare nella circoscrizione numero 1 della città di Torino, problematiche alquanto plausibili visto che è il centro città. Secondo le relazioni estratte da queste regole si potrebbe ancora assistere a spiacevoli disturbi dovuti ai locali nella circoscrizione 7, in cui vi è zona Vanchiglia, anche questo luogo di movida. Questi risultati evidenziano che ci sono zone della città in cui si potrebbero verificare determinati eventi e magari potrebbe essere interessante

utilizzare questi dati per risolvere queste problematiche o semplicemente identificare le cause e i luoghi.

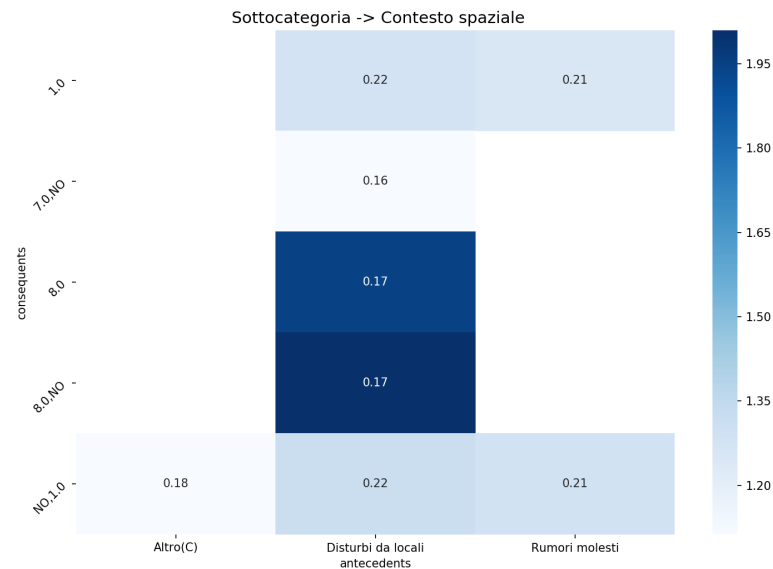


Figura 5.6: Regole associative Sottocategoria Convivenza Civile→ contesto spaziale

Per la tipologia di regole in cui una sottocategoria di tipo Qualità Urbana è correlata con determinati contesti geografici sono estratte 6 regole associative, riportate nella figura 5.7. Non si trovano supporti alti ma possiamo identificare regole con confidence tra 0.078 e 0.25, e con valori di lift tra 1.1 e 1.99. Veicoli abbandonati, decoro e degrado urbano sono le sottocategorie associate a queste regole. Molto interessante la regola che associa segnalazioni di decoro e degrado urbano ad aree verdi con un valore di confidence del 7.8% e lift 1.93. Inoltre per la sottocategoria decoro e degrado urbano c'è una correlazione con due differenti circoscrizioni, 4 e 5 con confidence indicativamente dell'11%. In aggiunta otteniamo che date le segnalazioni riguardanti la sottocategoria veicoli abbandonati queste si verificano maggiormente nella circoscrizione 6 ed in particolare in luoghi che non sono aree verdi, con una frequenza di circa il 25% e lift quasi 2.

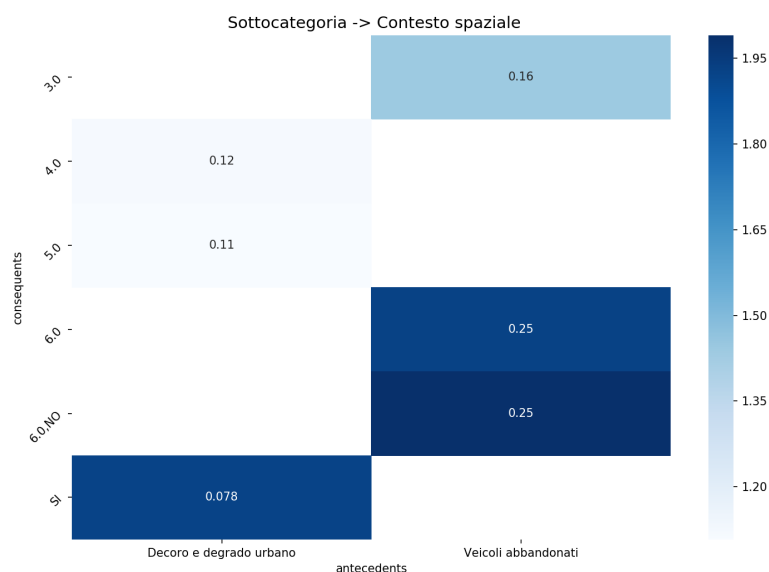


Figura 5.7: Regole associative Sottocategoria Qualità Urbana → contesto spaziale

Le regole di tipo sottocategoria Convivenza Civile → contesto temporale sono annotate nella figura 5.8. Si sono ottenute 40 regole con misure di supporto non alte ma possiamo trovare comunque regole apprezzabili, non solo per il contenuto informativo espresso ma anche da un punto di vista di metriche confidence e lift. Tra le regole estratte alcune raggiungono un valore di confidence massimo di circa 0.35, e regole che vanno da lift di 1.1 a 1.43. Le sottocategorie che hanno una relazione con determinati periodi temporali sono: Disturbi da locali, Comportamenti molesti, Altro(C) e Rumori molesti. Disturbi da locali e Altro(C) compaiono rispettivamente in 18 e 17 regole, le altre solamente in 3 (Rumori molesti) e in 2 (Comportamenti molesti). Come archi temporali si trovano corrispondenze nei mesi di Giugno, Luglio, Novembre e Dicembre. Inoltre sono presenti regole associative per tutte le stagioni dell'anno. Il contenuto informativo ottenuto è particolarmente interessante, troviamo associazioni tra le segnalazioni di tipo disturbi da locali nel periodo Giugno e Luglio, infatti è alquanto probabile la formazione di disturbi legati ai locali durante la bella stagione dovuti anche con molta probabilità alla formazione di locali all'esterno. L'anno 2012 è inoltre correlato con questo sottotipo della categoria Convivenza Civile. Al periodo estivo sono anche associate segnalazioni causate da comportamenti molesti, Comportamenti molesti → Estate ha una confidence del 35% con valori di lift di 1.37. La sottocategoria Altro(C) nel periodo 2017 e 2018 ha probabilmente avuto un incremento, visto le regole estratte, ed in particolare queste segnalazioni sono da seguire attentamente durante le stagioni Inverno e Autunno. Segnalazioni dovuti a rumori molesti, secondo le regole, hanno un'alta probabilità di verificarsi durante la stagione estiva, si registra infatti una

confidence del 31% e lift 1.25. Si nota anche che ci sono molte segnalazioni relative ai disturbi da locali nel giorno Lunedì, si è data come interpretazione che molto probabilmente questo tipo di segnalazioni sono fatte di lunedì ma in relazione ai giorni del weekend in cui è più plausibile la formazione di assembramenti e schiamazzi nei dintorni di alcuni locali anche fino a tarda notte.

Otteniamo 49 regole per la tipologia Qualità Urbana → contesto temporale, inserita nella figura 5.9. In queste regole generate possiamo notare l'incremento del valore del lift (>2) e valori di confidence per diverse regole maggiore del 30%. Da alcune regole possiamo dedurre importanti considerazioni, problemi legati ai veicoli abbandonati si sono avute nell'anno 2012, 2013, 2014, principalmente durante l'estate e la primavera. Troviamo infatti tra questa tipologia di regole Veicoli abbandonati → 2014 con confidence di 33% e valori di lift più alti, 2.5. Veicoli abbandonati → Estate ha valori di confidence del 28% e lift 1.1. Di poco più alto il valore di confidence (circa 31%) per la regola veicoli abbandonati → Primavera, Feriale. Particolare anche la regola che mette in relazione veicoli abbandonati con il giorno lunedì. Segnalazioni di tipo Decoro e degrado urbano hanno una certa correlazione con gli anni 2016, 2018, 2019 e anche con il mese di Ottobre, e la stagione Autunno. A titolo esemplificativo la regola decoro e degrado urbano → ottobre, autunno con valori di confidence del 10% e lift 1.1. La sottocategoria che numericamente compare di più tra queste regole è Altro(Q), ci sono interessanti relazioni tra questo tipo di sottocategoria ed il mese di Gennaio, Febbraio, Marzo, ed inoltre è ben correlata con la stagione invernale. Alcuni anni, in particolare 2017, 2018, 2019, risultano frequenti per questa tipologia di richieste. Altro(Q) in relazione con Inverno registra 27% di confidence e indicativamente 1.16 di lift. Queste regole ci forniscono nel complesso una buona indicazione temporale associata a determinate sottocategorie di tipo qualità urbana.

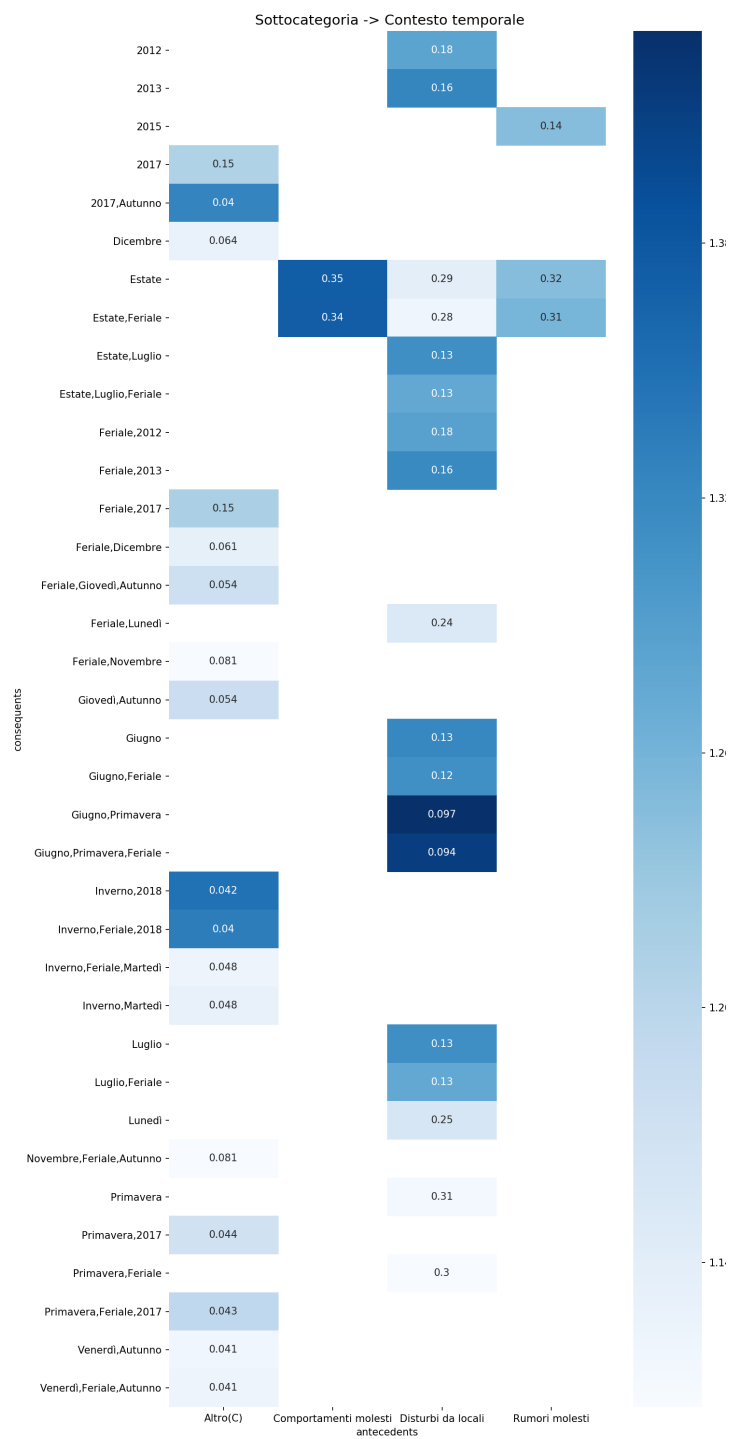


Figura 5.8: Regole associative Sottocategoria Convivenza Civile → contesto temporale

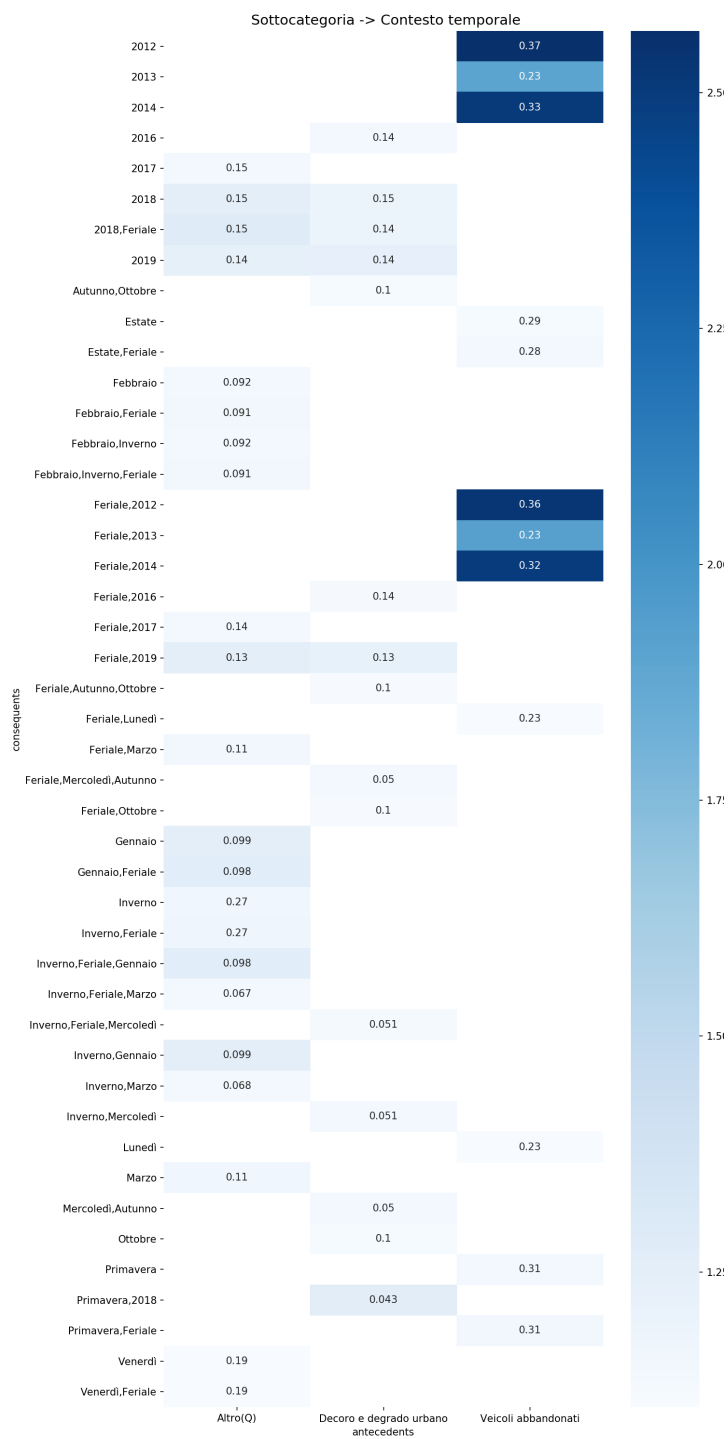


Figura 5.9: Regole associative Sottocategoria Qualità Urbana → contesto temporale

Le regole di tipo sottocategoria Convivenza Civile $\rightarrow$ Contesto spaziale e temporale estratte sono ben 61, riportate nella figura 5.10. I valori di supporto non sono particolarmente rilevanti in un termine puramente metrico ma troviamo comunque regole che ci forniscono determinate informazioni da poter analizzare. Troviamo tra le regole generate un valore di confidence massimo di 32%, e valore massimo di lift indicativamente di 1.96. Le sottocategorie trovate in relazione contemporaneamente a determinati attributi spaziali e temporali sono: Disturbi da locali (24 regole), Altro(C) (31 regole), Comportamenti molesti (2), Rumori molesti (4). Si trovano segnalazioni relative a disturbi da locali nella circoscrizione 8 in giorni feriali e in posti non aree verdi, e si rimarca nuovamente la correlazione di questa segnalazione con i mesi di Giugno e Luglio. Quindi potrebbe risultare probabile che queste segnalazioni sono effettuate maggiormente nei dintorni dei luoghi della movida nel quartiere San Salvario e in linea generale questa problematica si verifica maggiormente nella stagione estiva. Possiamo apprezzare che rispetto alle regole estratte precedentemente le segnalazioni di non emergenza registrate da questa tipologia di sottocategoria (disturbi da locali) sono associate a livello temporale a giorni non feriali. Purtroppo questa risulta un'informazione generale e le regole non riescono a registrare una relazione con un determinato giorno della settimana. Dalla sottocategoria Altro(A), frequente in determinate stagioni (Autunno, Inverno), notiamo l'aggregazione con l'informazione spaziale non aree verdi. In base al contenuto informativo estratto da queste regole possiamo affermare anche che segnalazioni di rumori molesti si potrebbero registrare nella circoscrizione 1 in giorni feriali e in aree non verdi, come anche nei periodi estivi. Comportamenti molesti si potrebbero verificare in estate e in luoghi non verdi e feriali. Possiamo affermare che il contenuto informativo rimarca quanto detto nelle precedenti regole sebbene con un maggiore rilievo.



Figura 5.10: Regole associative Sottocategoria Convivenza Civile → contesto spaziale e temporale

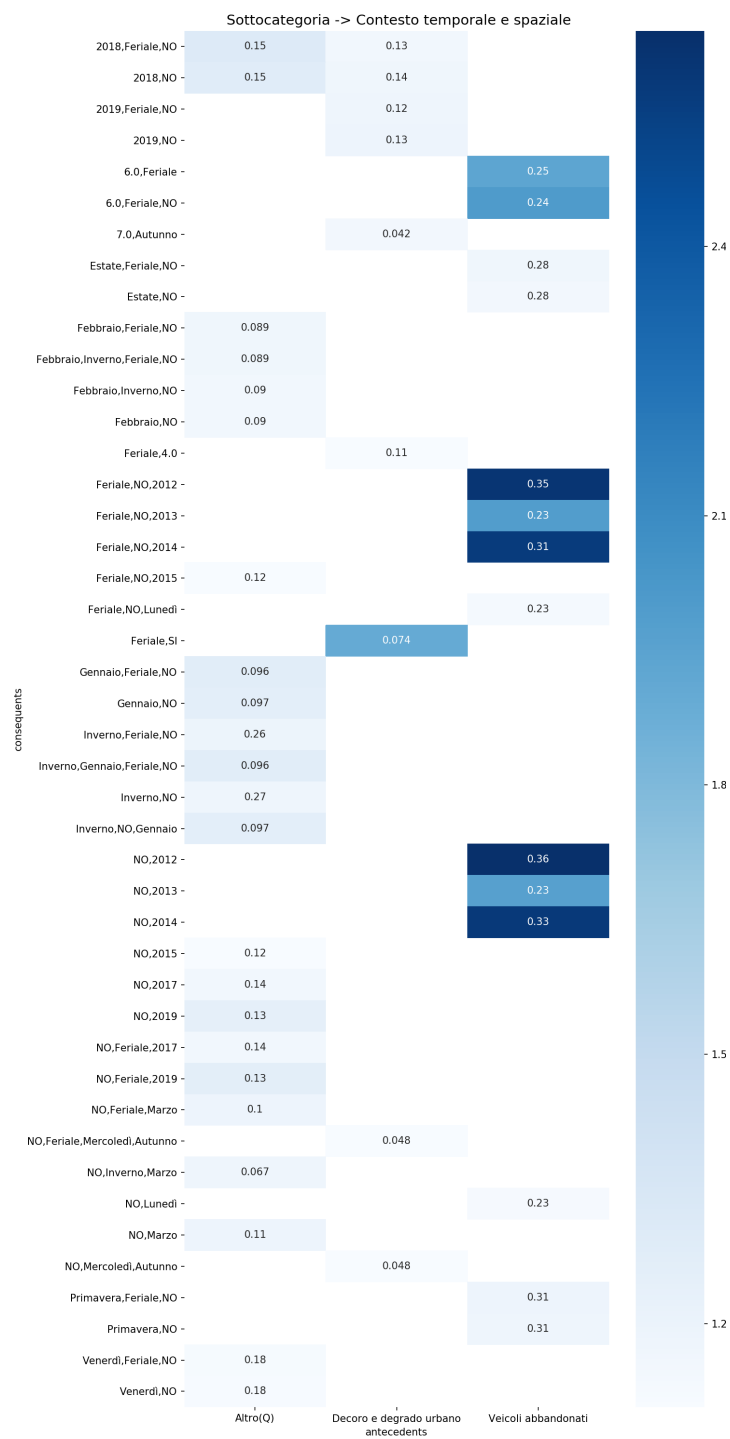


Figura 5.11: Regole associative sottocategoria Qualità Urbana → contesto spaziale e temporale

La figura 5.11 riporta le regole associative di tipo sottocategorie di tipo Qualità Urbana → contesto temporale e spaziale. Le sottocategorie che possiamo osservare in queste regole sono: Altro(Q) (45 regole), Veicoli abbandonati (29 regole), Decoro e degrado urbano (27 regole). Da sottolineare che tra le regole compaiono tutti gli anni dal 2012 al 2019, e ci sono informazioni anche per tutte le stagioni. Mentre a livello informativo se teniamo in considerazione i mesi presenti riguardano solo i mesi di Gennaio, Febbraio, Marzo e Ottobre. A livello di informazioni aggiuntive troviamo tra queste correlazioni segnalazioni di decoro e degrado urbano anche nella circoscrizione 7 e durante la stagione autunnale. Questa informazione spaziale non era contenuta tra le regole precedentemente generate, ed associata al periodo stagionale suggerisce una informazione di particolare importanza, nonostante i valori di confidence non siano altissimi, all'incirca il 4% e lift 1.15. Le regole più forti sono relative a segnalazioni sui veicoli abbandonati e in determinati anni, come anche visto in precedenza, la novità è l'associazione temporale con queste circoscrizione e tra aree della città di Torino non considerate aree verdi. Osservando questa tipologia di regole associate alla sottocategoria, si può dire che otteniamo approfondimenti più specifici che ci permette di delineare una problematica ben precisa.

I risultati analizzati in questo paragrafo ci hanno permesso di capire quanto è importante analizzare questa tipologia di dati, sia per provare ad identificare delle situazioni allarmanti di sicurezza e disturbo in determinate aree della città di Torino, sia per osservare come determinate problematiche sono cambiate nel tempo. Il contenuto informativo estratto è molto interessante, e ci ha permesso di fare delle determinate considerazioni. Per due delle tre macrocategorie che identificano determinate tipologie di segnalazioni si sono estratte informazioni che potrebbero risultare utili in ambito urbano, purtroppo per la categoria Allarme Sociale le regole non sono riuscite a fornire alcun contenuto, molto probabilmente le segnalazioni presenti per questo tipo di categoria risultano parecchio sparse e numericamente poche. Nemmeno per la tipologia sottocategoria si sono trovate correlazioni con le vie cittadine e con alcune circoscrizioni. Mancano anche i riferimenti ad alcuni attributi temporali. Inoltre questa analisi ci permette di capire quanto potenziale informativo possano avere questi dati urbani per monitorare la percezione di sicurezza in città dei cittadini, e sicuramente avere dataset più corposi ci permetterebbe di avere risultati maggiormente precisi e dettagliati. Attraverso i risultati provenienti da analisi simili, per esempio la polizia locale potrebbe aumentare la sorveglianza notturna vicino a ristoranti, pub e discoteche situati nel centro della città o nelle zone tipiche di assembramenti notturni da parte dei giovani, o andare ad individuare circoscrizioni o quartieri che peccano di decoro urbano, e adottare delle misure per risolvere problematiche di degrado all'interno della città.

5.4 Analisi regole associative estratte dalle segnalazioni di New York

In questo paragrafo verranno presentati i risultati dell'analisi dei dati sulle segnalazioni urbane della città di New York attraverso la tecnica di data mining scelta. Saranno analizzate le regole associative generate seguendo la struttura già usata per il dataset di Torino. Riporteremo prima le regole in cui l'attributo relativo alla categoria, Complaint Category, è il conseguente, dopo le regole di associazione relative al sottotipo Complaint Type, in cui il contesto spaziale/temporale è il conseguente, e la sottocategoria l'antecedente.

Per primo vengono presentati i risultati ottenuti dalla generazione di regole di tipo: contesto spaziale $\rightarrow$ Categoria. Nel dataset si contano 11 tipologie di categorie, di queste solo 7 categorie trovano delle relazioni. Infatti vedremo che le categorie e i rispettivi sottotipi che riportano informazioni estrapolate dalle regole associative sono: Cleaning & Garbage, Complaint, Information, Maintenance, Park & Tree, Streets & Sidewalks e Transportation.

Per quanto riguarda la categoria Cleaning & Garbage, le regole in cui la categoria è conseguente ed il contesto spaziale è antecedente sono riportate nella figura 5.12.

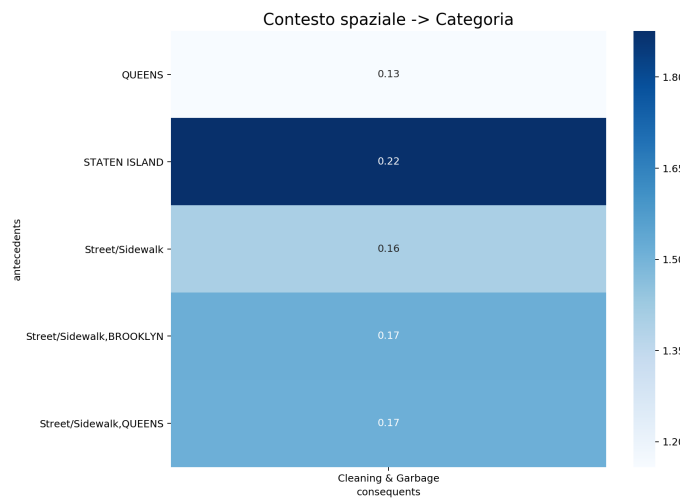


Figura 5.12: Regole associative contesto spaziale $\rightarrow$ Categoria Cleaning & Garbage

Queste 5 regole relative a segnalazioni per problematiche legate alla pulizia e ai rifiuti mostrano delle correlazioni con determinati luoghi e distretti della città di New York. Troviamo la regola Streets/Sidewalks $\rightarrow$ Cleaning & Garbage che indica che nei luoghi stradali il 16% di segnalazioni appartiene a questa tipologia di problemi. Il lift di questa regola è di 1.4. Le altre regole mostrano delle correlazioni con 3 differenti distretti della città: Staten Island, Brooklyn e Queens. Inoltre in 2

regole vi è rispettivamente l'associazione tra il distretto di Brooklyn e Queens e il luogo della segnalazione che risulta strada o marciapiede, queste due regole hanno una confidence del 17% e lift 1.51. La regola più forte in termini di lift è quella che indica il verificarsi nel distretto di Staten Island di segnalazioni di questa categoria. Il contenuto informativo di queste regole di associazione ci permette di identificare determinate aree della città di New York che dovrebbero risolvere problematiche legate ai rifiuti e alla pulizia della città.

Le regole categoria contesto temporale → Cleaning & Garbage sono riportate nella figura 5.13. Si ottengono 8 regole con informazioni temporali. Sono presenti gli anni 2017, 2018 e 2019, si deduce che probabilmente in questi anni si è assistito ad un incremento di segnalazioni di questo tipo, inoltre si verificano perlopiù nei giorni feriali. Particolare il fatto che l'unica correlazione con periodi dell'anno è il mese di Giugno in una regola, e la stagione estiva in un'altra. Precisamente si ha una frequenza del 13% di questa categoria di problemi urbani in Estate e nei giorni feriali, con lift 1.12. Le regole più forti in termini di lift (circa 1.5) sono correlate all'anno 2018. In definitiva possiamo indicativamente dire che a livello temporale si potrebbe verificare un aumento di queste segnalazioni in determinati periodi estivi, e negli anni dal 2017 al 2019 la situazione probabilmente risultava abbastanza critica rispetto ad altri anni.

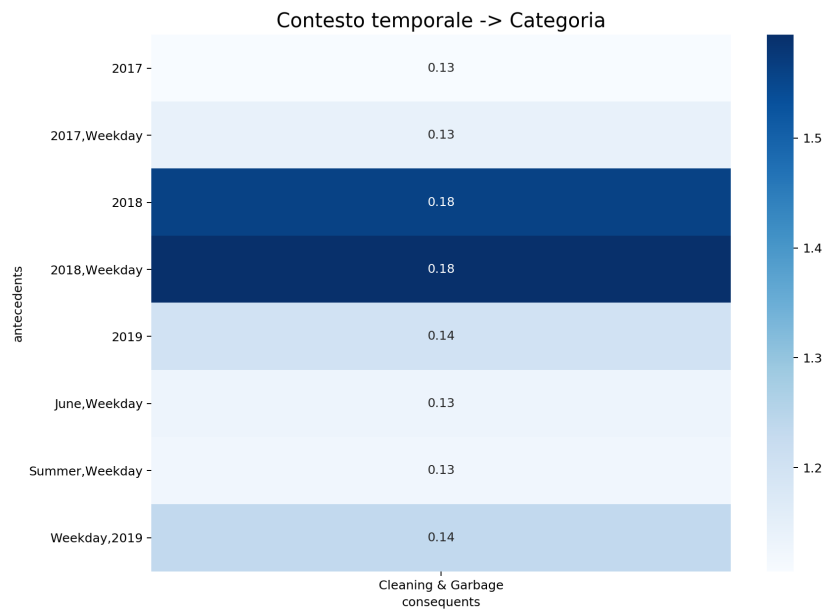


Figura 5.13: Regole associative contesto temporale → Categoria Cleaning & Garbage

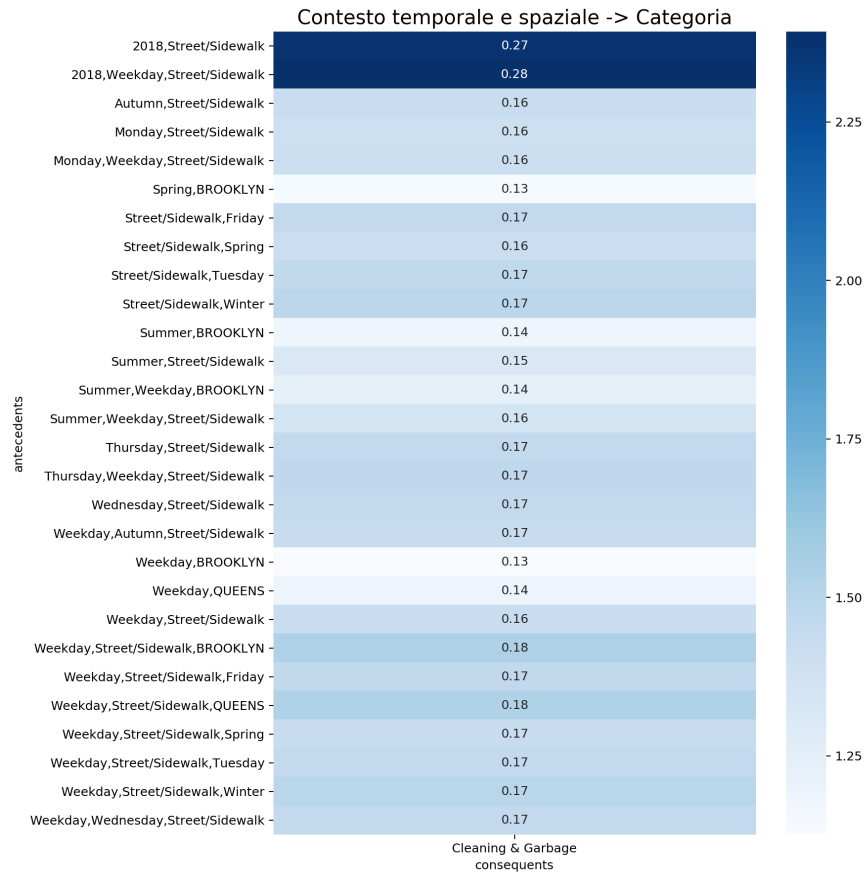


Figura 5.14: Regole associative contesto temporale e spaziale → Categoria Cleaning & Garbage

Molto interessanti le regole contesto temporale e spaziale → Cleaning & Garbage, figura 5.14. Queste 28 regole ci forniscono informazioni aggiuntive tramite l'associazione di questa tipologia categorica sia con attributi spaziali e temporali. Probabilmente l'anno 2018 visti i valori delle metriche, si trovano valori di lift superiore a 2, è stato problematico per quanto riguarda problemi di pulizia e rifiuti, inoltre se ne deduce che queste segnalazioni erano relative a giorni feriali e in zone stradali. In aggiunta si ottengono regole in cui compare la stagione inverno, vi è una confidence di circa 17% per la regola Winter, Streets/Sidewalks → Cleaning & Garbage, con lift 1.48. Quindi rispetto alle regole precedenti, attraverso questa regola si evidenziano problematiche di questo tipo sulle strade e marciapiedi e durante l'Inverno. Ma non solo, ci sono correlazioni anche nel periodo autunnale, primaverile ed estivo. In particolare il distretto di Brooklyn durante la stagione primaverile ed estiva registra segnalazioni etichettate come categoria Cleaning &

Garbage. Particolare la presenza di determinati giorni della settimana, si evidenziano segnalazioni di questo tipo Lunedì, Martedì, Giovedì e Venerdì in luoghi stradali. Non a caso una regola associativa indica che il 16% di segnalazioni su strada e marciapiede riguardano questa categoria di segnalazioni. Inoltre è stato osservato che in queste regole non sono presenti gli anni 2017 e 2019 come nelle precedenti regole in cui venivano presi in esame solamente gli attributi temporali, questo è stato interpretato come un possibile processo di degrado che è iniziato nel 2017 ed ha trovato il suo culmine nell'anno 2018, come dimostrato dai valori di robustezza delle regole.

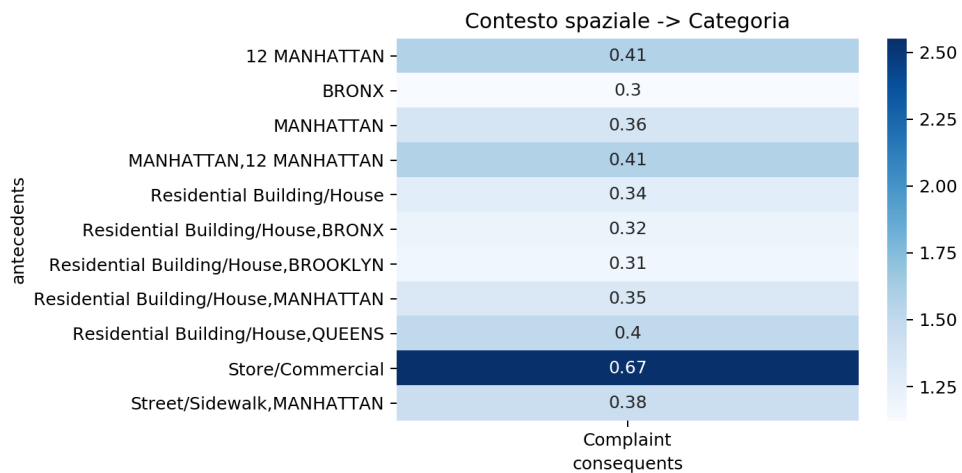


Figura 5.15: Regole associative contesto spaziale → Categoria Complaint

Successivamente si sono esaminate le regole associative per la categoria Complaint. Sono segnalazioni in cui si fanno dei reclami ben precisi. Nella figura 5.15 troviamo le regole di tipo contesto spaziale → Complaint. La regola con il lift più elevato (2.55) è quella in cui si osserva che preso un luogo commerciale nel 67% di casi si registrano dei reclami di tipo Complaint. Mentre circa il 33% di confidence è la regola che associa luoghi residenziali a questa tipologia di segnalazioni. Tra i distretti osserviamo che nel distretto di Bronx sono registrate circa il 29% di queste problematiche, mentre il 36% in quello di Manhattan. Troviamo anche delle regole che evidenziano il verificarsi di eventi di questo genere in alcuni distretti ed in particolare in edifici residenziali. È il caso dei distretti di Manhattan, Queens, Bronx e Brooklyn, una frequenza di 35% si ha per le abitazioni di Manhattan, del 32% nei distretti di Bronx, e circa il 30% a Brooklyn. Si deduce che questo tipo di segnalazioni si registrano in quasi tutte le zone di New York. Particolarmente interessante la regola 12 Manhattan → Complaint che ci fornisce un dettaglio ancora più specifico, cioè che diverse tipologie di reclami vengono fatte nella zona 12 di

Manhattan. Altra possibile informazione è estraibile dalla regola Streets/Sidewalks, Manhattan → Complaint che ha un valore di confidence pari a circa il 38%, e da cui deduciamo che nelle strade di Manhattan si registrano diversi reclami.

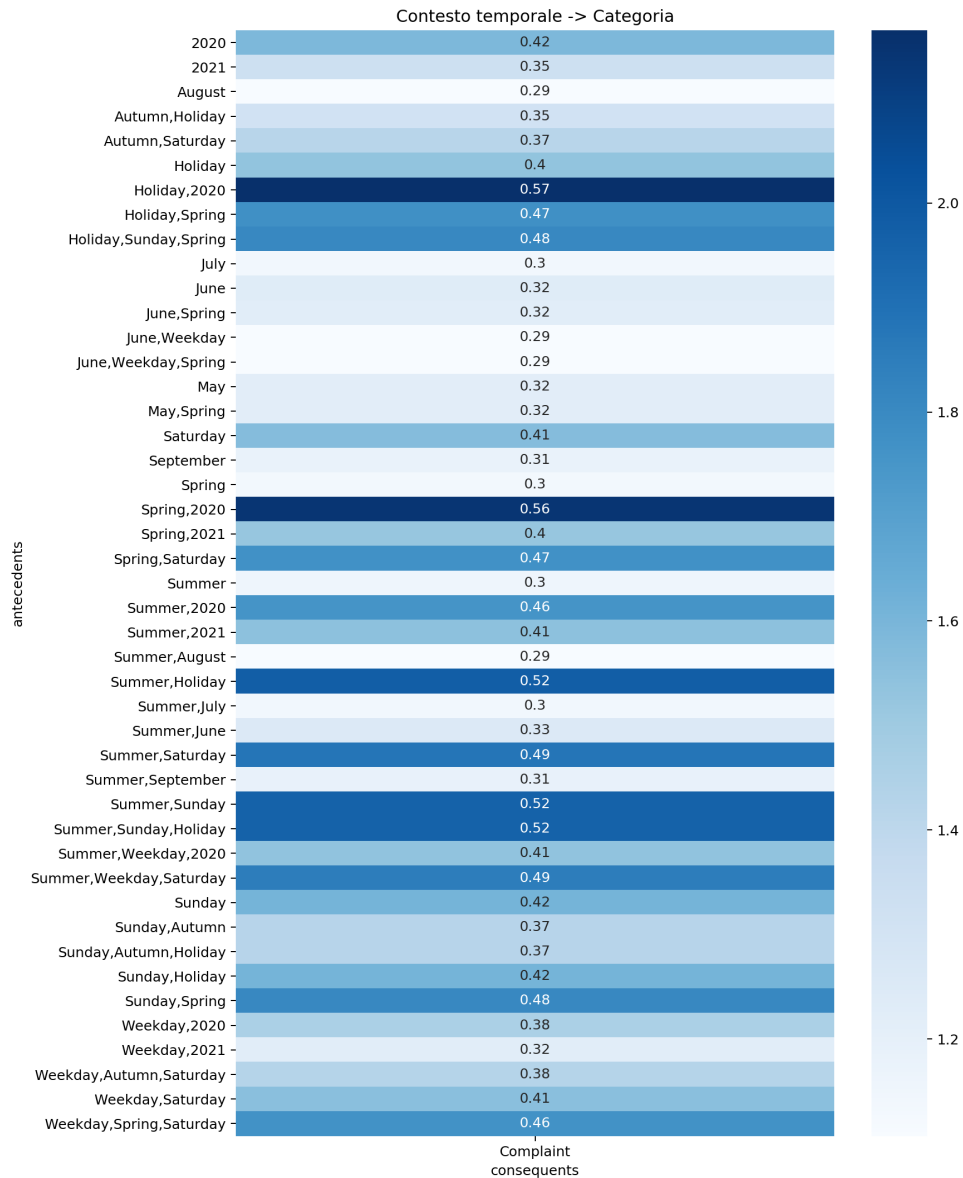


Figura 5.16: Regole associative contesto temporale → Categoria Complaint

Le regole associative di tipo Complaint → contesto temporale sono indicate nella figura 5.16. Ci sono osservazioni molto interessanti anche a livello temporale con questa tipologia di segnalazione. Troviamo una regola forte, lift 2.16, che associa i giorni festivi dell'anno 2020 a questa tipologia categorica. Inoltre oltre all'anno 2020 in diverse regole è presente anche l'anno 2021, questo ci porta a dedurre che gli ultimi anni si è magari verificata una maggiore attenzione a questa tipologia di richieste. Molto interessante anche le regole che ci forniscono l'informazione che questa categoria si verifica con una certa frequenza il sabato e la domenica e durante la stagione estiva, primaverile e autunnale. Non a caso si trovano anche correlazioni con mesi appartenenti alle stagioni appena indicate. Questa informazione è particolarmente preziosa, ci permette di ipotizzare il verificarsi di particolari forme di reclamo durante i weekend di particolari periodi dell'anno. In questo modo abbiamo un riferimento ben preciso a livello temporale su una categoria di segnalazioni. Possiamo dire che nel 2020 il 57% delle segnalazioni sollevate dai cittadini nella giornata di domenica appartengono alla categoria Complaint, come anche la regola Holiday → Complaint che ha un valore di confidence del 40% ci suggerisce che nei giorni festivi determinate problematiche sono da monitorare nella città. La conoscenza estratta da queste regole può essere utile nel contesto urbano per capire le esigenze dei cittadini in determinati archi temporali.

Informazioni molto più dettagliate si generano unendo il contesto spaziale e temporale a questo tipo di segnalazioni, La figura 5.17 riesce a chiarire meglio le possibili intuizioni da poter estrarre da questi risultati. Vengono qui riassunte le principali in modo da poter capire quali meglio contestualizzano questo tipo di problema. Riusciamo a precisare che le segnalazioni di questa tipologia sono principalmente effettuate nel periodo estivo, nei giorni festivi e in centri abitati, mentre segnalazioni legati alle attività commerciali sono fatte nei giorni feriali. Si riesce anche a dare un'associazione temporale ai distretti, per esempio la domenica e il sabato nel distretto di Manhattan, di Bronx, di Brooklyn e di Queens. Mentre nella community board numero 12 di Manhattan maggiori richieste di reclamo sono fatte nei giorni settimanali. Riusciamo anche a capire il luogo di queste segnalazioni in determinati mesi dell'anno, spiccano il mese di Maggio, Luglio, Agosto e Ottobre in zone abitative. Riusciamo ad aggiungere alle informazioni estratte dalle regole estratte in precedenza che gli anni 2020 e 2021 relative alle segnalazioni appartenenti alla categoria Complaint hanno una buona correlazione in particolar modo con il distretto di Manhattan. Una regola in particolare ci indica che circa il 9,5% di segnalazioni di tipo Complaint fatte durante i giorni feriali appartengono a zone stradali di Manhattan, da altre regole si osserva che le strade e i marciapiedi hanno una certa rilevanza anche durante l'anno 2020 per queste richieste di segnalazione. Inoltre per quanto riguarda il distretto di Manhattan riusciamo anche a trovare correlazioni con determinati mesi dell'anno, in particolare Autunno e Primavera ed Estate. Il distretto di Bronx ha una certa rilevanza durante

la stagiona primaverile ed estiva. Sicuramente il potenziale informativo estratto è molto valido seppur non stiamo ancora scendendo nel dettaglio dei sottotipi di segnalazione che meglio precisano la problematica. Nelle regole inverse possiamo ammirare livelli di confidence più alti. Quindi questo ci suggerisce come dato un luogo e un arco temporale preciso si riesce ad individuare quanto è frequente una determinata segnalazione.

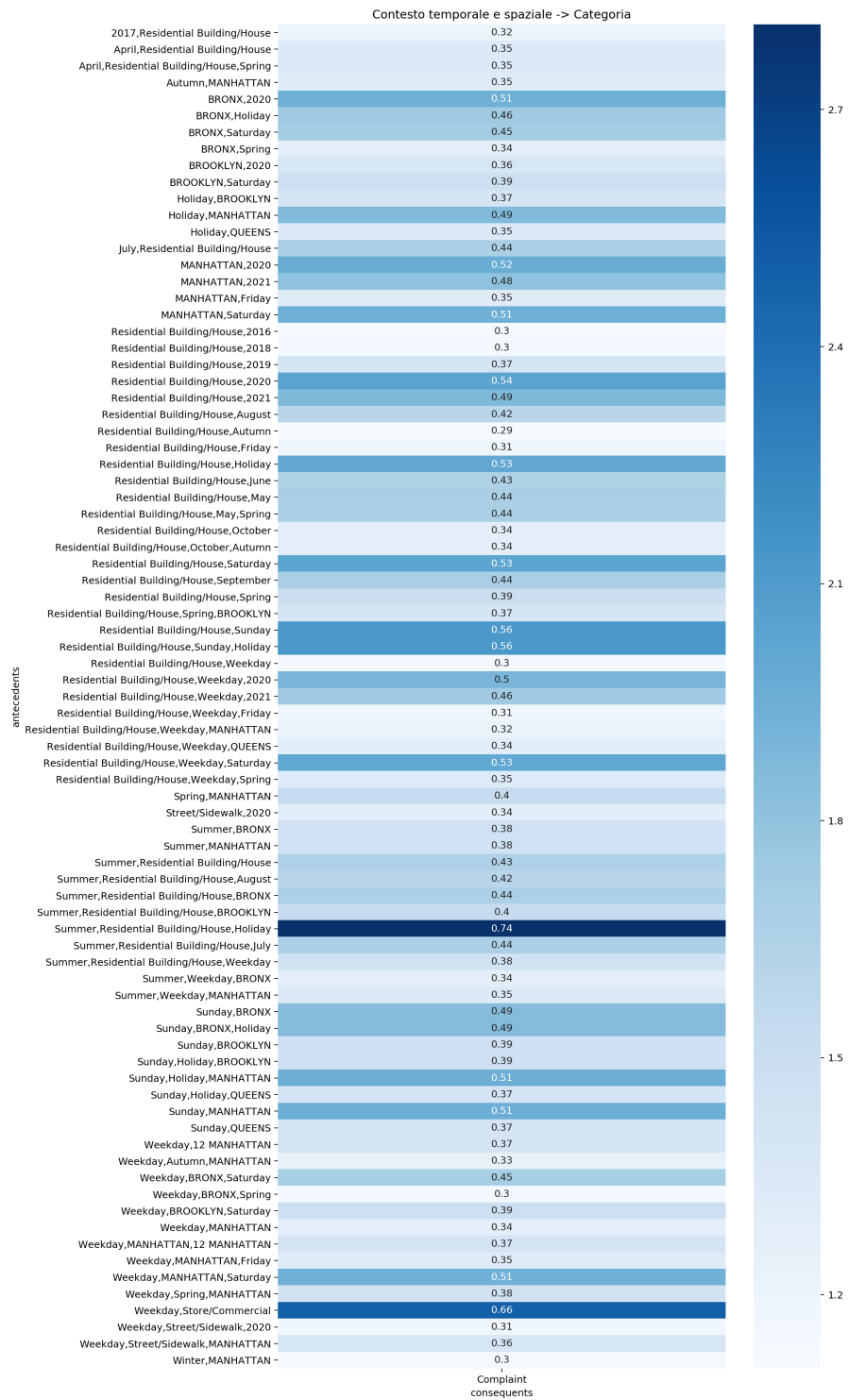


Figura 5.17: Regole associative contesto temporale e spaziale → Categoria Complaint

Adesso si riportano i risultati ottenuti dalla generazione di regole associative per la categoria Information. Questa categoria è molto generica rispetto alle altre e si riferisce ad eventuali informazioni utili per risolvere una determinata problematica. Sono solamente 2 le regole estratte di tipo contesto spaziale → Information.

La prima informazione è che nel distretto di Manhattan nel 6% dei casi si registrano richieste di informazioni, mentre nell'altra regola si deduce che sono relativi a case/abitazioni.

L'unica informazione temporale è invece legata ai giorni feriali. Mentre se si prendono insieme i contesti spaziali e temporali come era prevedibile si deduce la presenza di certe segnalazioni informative nel distretto di Manhattan durante i giorni feriali.

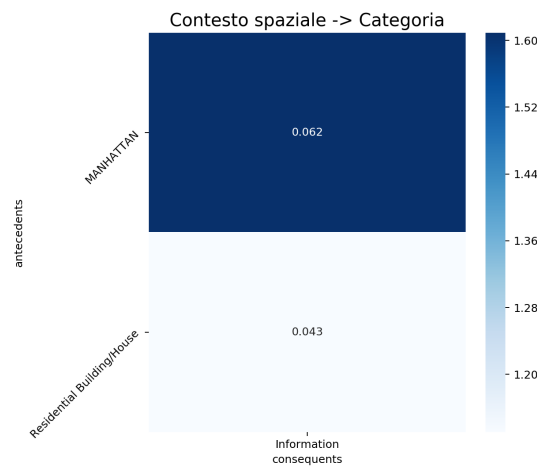


Figura 5.18: Regole associative contesto spaziale → Categoria Information

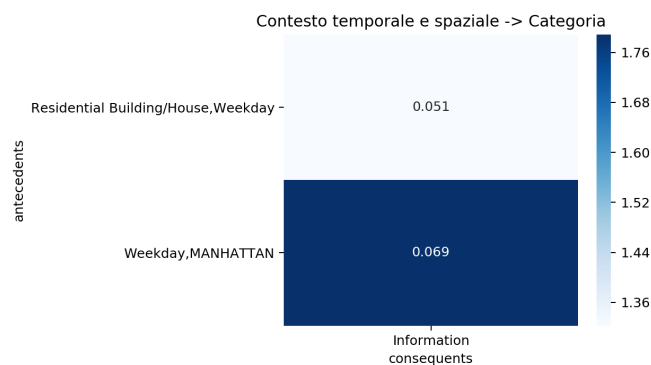


Figura 5.19: Regole associative contesto temporale e spaziale → Categoria Information

Adesso si osserveranno le regole generate per la categoria Maintenance, una categoria che racchiude tutte le segnalazioni relative a problemi di manutenzione. Si presentano in un primo momento le correlazioni contesto spaziale $\rightarrow$ Maintenance nella figura 5.20.

Si notano regole con confidence comprese tra 0.38 e 0.64, e lift tra 1.47 e 2.44. Si deduce che nel distretto di Bronx ed in particolare nel community Board 4, e in edifici residenziali si potrebbero verificare problematiche di manutenzione. Questa risulta essere la regola più forte in termini di lift. Situazione potenzialmente simile anche nell'area 12 di Manhattan. Anche il distretto di Brooklyn ha una frequenza di segnalazioni appartenenti a questa categoria ma in luoghi non specifici. Nel distretto di Bronx, e specificamente nella community board 5 e 7 si verificano rispettivamente nel 50% e nel 48% dei casi problematiche di manutenzione. Vi è anche una certa correlazione con il distretto di Queens sempre in abitazioni.

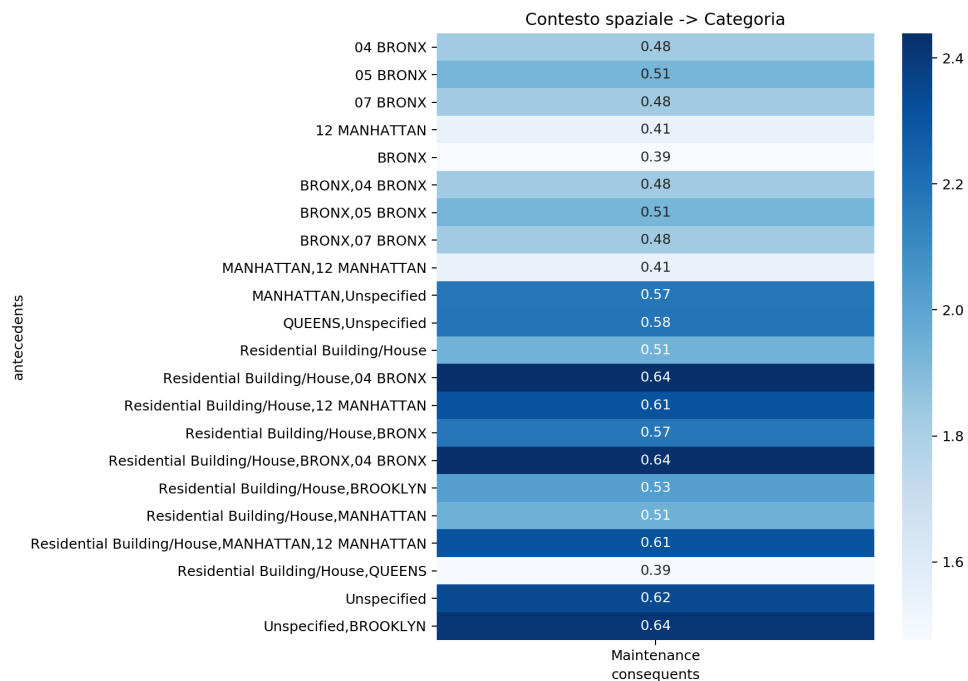


Figura 5.20: Regole associative contesto spaziale $\rightarrow$ Categoria Maintenance

Nella figura 5.21 sono indicate le regole generate di tipo contesto temporale → Maintenance. Le regole più robuste sono quelle riferite a mesi invernali ed in particolare Gennaio, circa il 38% delle segnalazioni in quel periodo è riferita a problemi di manutenzione. Interessante il dato sugli anni, si trovano correlazioni dall'anno 2011 al 2016, questo potrebbe anche essere potenzialmente indice di un possibile miglioramento di questa problematica negli ultimi anni. Si trovano anche correlazione nei mesi di Dicembre, Marzo, Febbraio e Novembre seppur con livelli di lift e confidence minori. Si evince che nei giorni feriali si verificano questo genere di segnalazioni. Molto dettagliata in questo aspetto le regole che associano anche il dettaglio del giorno alla segnalazione, come Winter, Monday → Maintenance che denota il verificarsi con una confidence del 36% di richieste di manutenzione nei lunedì della stagione invernale. Vi è una correlazione simile anche per altri giorni settimanali, Martedì, Mercoledì, Giovedì e Venerdì. Questo ci indica che comunque è abbastanza costante segnalare questi problemi nel corso della settimana. Inoltre, un pò a sorpresa, si trova una certa relazione tra i giorni festivi nel periodo invernale e problematiche di manutenzione. Ciò che evince maggiormente è anche la non correlazione nei periodi primaverili ed estivi con questa tipologia di richieste.

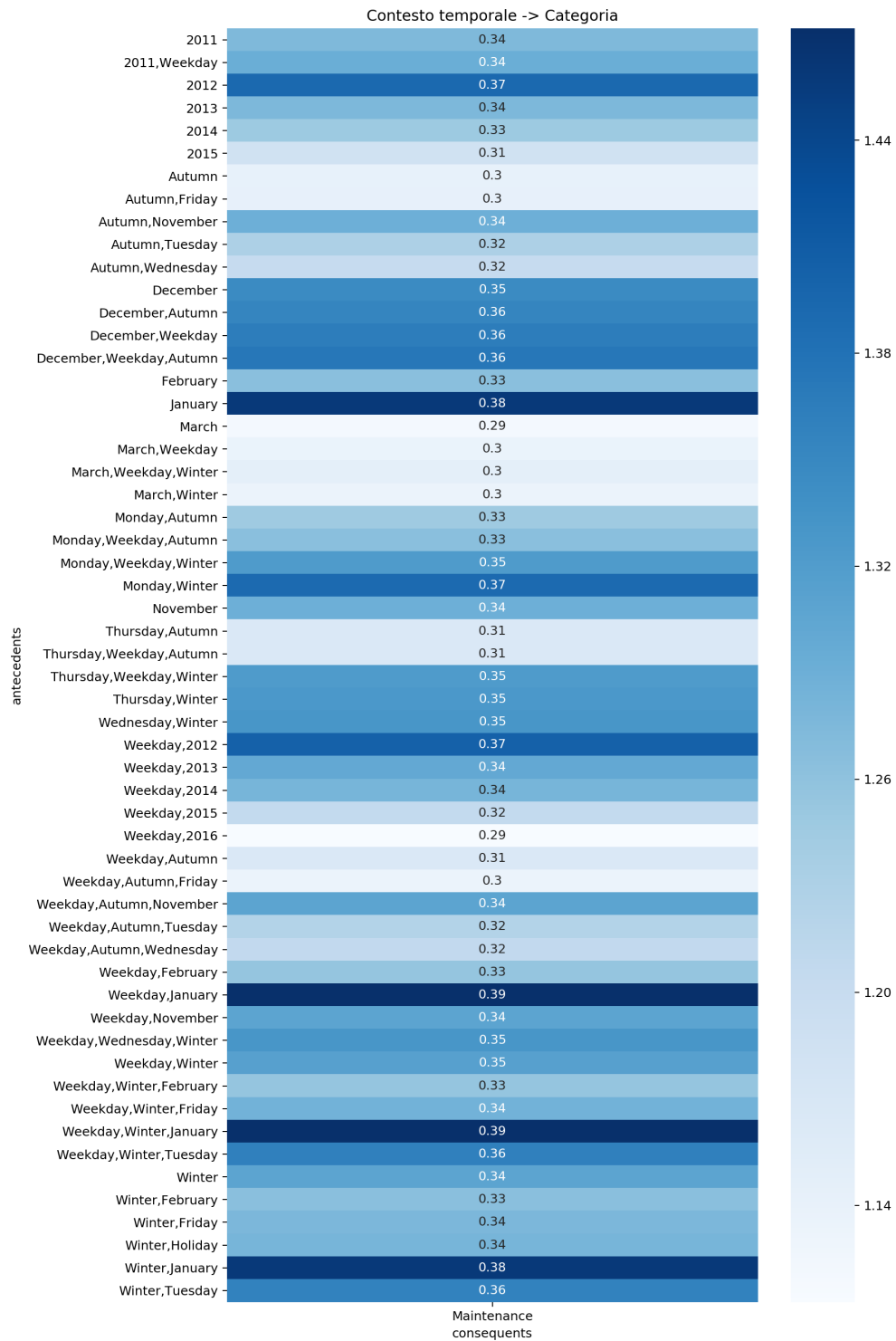


Figura 5.21: Regole associative contesto temporale → Categoria Maintenance

Prendendo in considerazione contemporaneamente il contesto spaziale e temporale riusciamo ad estrarre un contenuto informativo veramente valido per questa tipologia di segnalazioni. Troviamo 179 regole, tra cui, alcune molto affidabili e forti. Le seguenti regole sono descritte nella figura 5.22. Molte sono le regole con lift maggiore di 2. Il valore della confidence varia tra le regole da un minimo di 0.29 ad un massimo di 0.69. Riusciamo ad estrapolare informazioni sulla correlazione di zone urbane e archi temporali per questa categoria. Troviamo nel distretto di Manhattan durante la stagione invernale e in luoghi abitativi il verificarsi di problemi legati alla manutenzione. Forti anche le regole che implicano la registrazione di questa tipologia di segnalazione in determinati giorni settimanali e nel distretto di Brooklyn, e in zone abitative. Ricorre spesso la correlazione tra diversi distretti e abitazioni in determinati giorni della settimana. Non solo si deduce anche l'associazione di segnalazione in determinate stagioni e mesi nei vari distretti della città, a titolo di esempio: Bronx, Spring, Residential/House $\rightarrow$ Maintenance con un valore di confidence del 51% e lift 1.94. Inoltre questa è un'informazione aggiuntiva rispetto alle precedenti regole in cui veniva preso in considerazione solo l'aspetto temporale, precedentemente infatti tra le regole vi era l'associazione solamente con Autunno e Inverno. Inoltre nella stagione primaverile in determinati luoghi non specificati nel 59% potrebbero verificarsi problemi di manutenzione. Non solo regole robuste si hanno anche con la stagione Estate, si rileva una confidence del 65% con lift 2.46 per la regola Unspecified place, Summer $\rightarrow$ Maintenance. Se quindi in precedenza sembrava che il verificarsi in determinate stagioni di questa problematica era improbabile, si deduce contrariamente che in determinati posti può invece registrarsi. I mesi che rappresentano una nuova conoscenza sono Agosto, Aprile e Maggio. Inoltre diversamente da come si pensava, anche negli anni recenti, 2019, 2020, 2021, si associano richieste di manutenzione. Questo ci fa pensare che questa tipologia di problema non ha un arco temporale ristretto ma si verifica in maniera costante, probabilmente sono segnalazioni che possono essere all'ordine del giorno e verificarsi quotidianamente. Riusciamo a guadagnare qualche altra informazione interessante, come il fatto che nei giorni festivi durante la stagione invernale e in luoghi abitati si registrano segnalazioni relative a questa categoria con valori di confidence del 57% e con lift 2.18. Nonostante ciò comunque si possono estrapolare dei luoghi e delle aree ben precise da monitorare. L'unico distretto a non comparire tra le regole è quello di Staten Island.

Sezione sperimentale

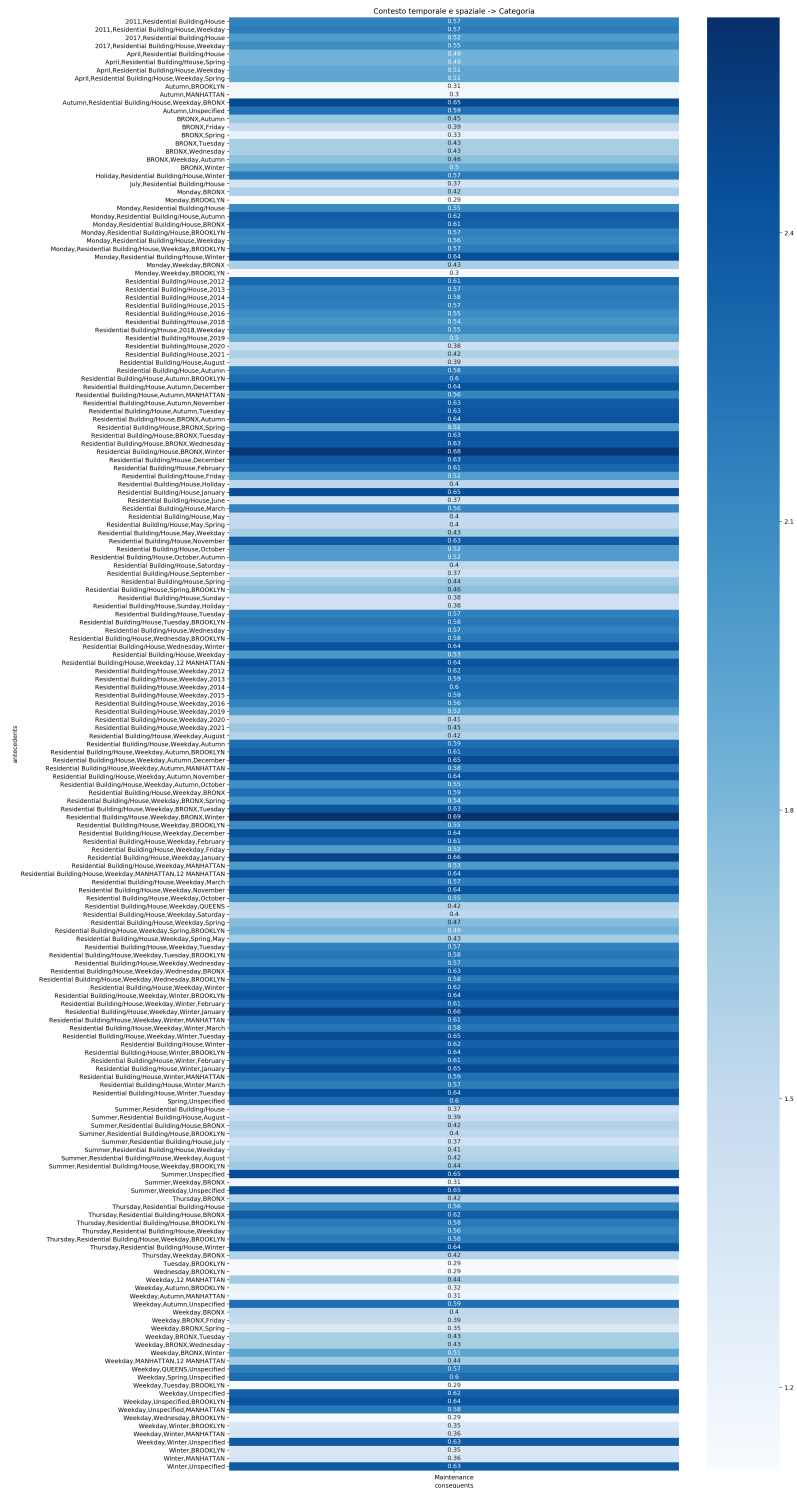


Figura 5.22: Regole associative contesto temporale e spaziale → Categoria Maintenance

Le figure sotto sono relative alla categoria Park & Trees, è facile capire che si tratta di segnalazioni di non emergenza relativi a parchi ed aree verdi. Principalmente nelle strade o marciapiedi e nei distretti di Queens e Brooklyn si potrebbero verificare richieste legate alla categoria Park & Trees. Per quanto riguarda l'aspetto temporale, troviamo correlazioni molto forti con i giorni feriali ed in particolar modo nel periodo estivo. Queste considerazioni sono confermate anche nelle regole sul contesto temporale e spaziale.

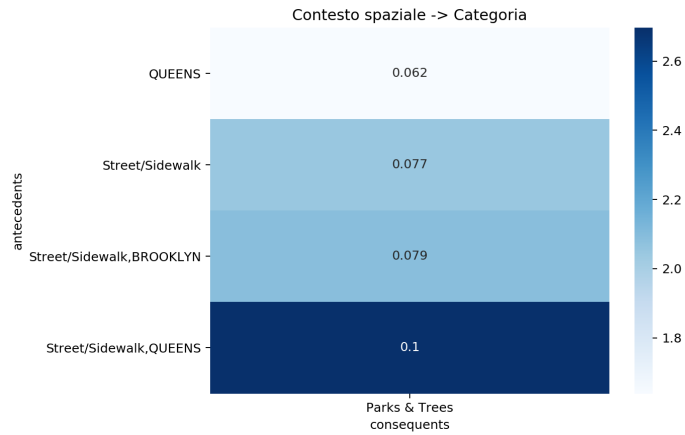


Figura 5.23: Regole associative contesto spaziale → Categoria Park & Trees

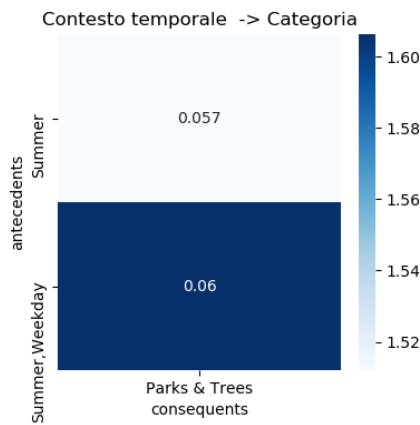


Figura 5.24: Regole associative contesto temporale → Categoria Park & Trees

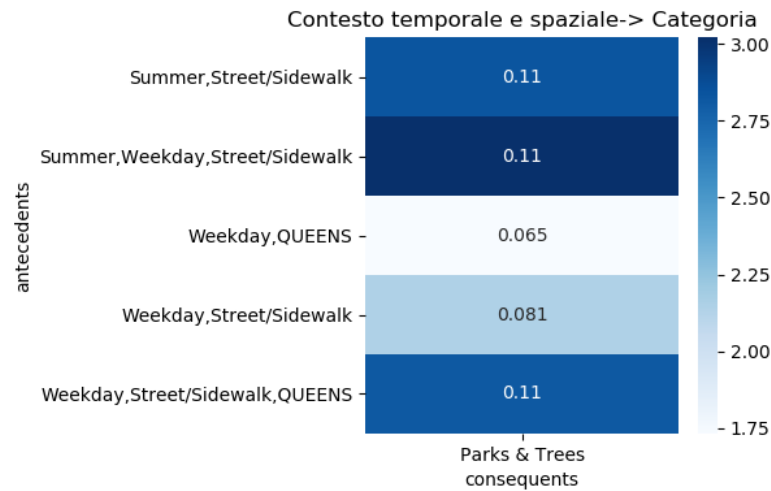


Figura 5.25: Regole associative contesto spaziale e temporale → Categoria Park & Trees

La figura 5.26 mostra 7 le regole generate filtrando la categoria Streets & Sidewalks. Ci sono riferimenti ai distretti e all'attributo Location Type. Dalle regole si evince che nelle strade e marciapiedi dei distretti di Queens, Staten Island, Brooklyn e Bronx potrebbero verificarsi segnalazioni urbane in riferimento alla categoria Streets & Sidewalks. Vi sono anche due regole, rispettivamente per Queens e Staten Island, che implicano il verificarsi di questa categoria di segnalazioni indipendentemente dal luogo specifico.

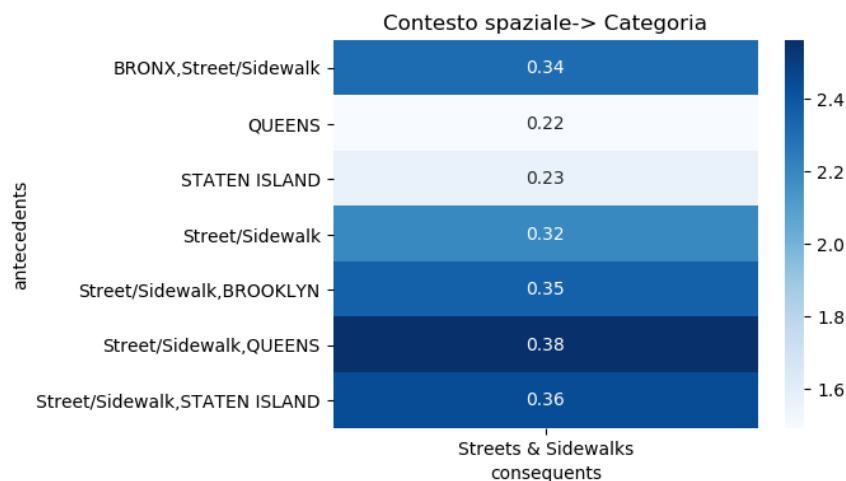


Figura 5.26: Regole associative contesto spaziale → Categoria Streets & Sidewalks

Ci sono 12 regole estratte di tipo contesto temporale → categoria Streets & Sidewalks, figura 5.27. Una forte correlazione con questa tipologia è dato dall'anno 2010. Ci sono anche ulteriori correlazioni con gli anni 2015 e 2019. La principale correlazione con il mese è data dalla regola Weekday, February, Winter → categoria Streets & Sidewalks. Anche il mese di Febbraio è correlato a segnalazioni di tipo stradali.

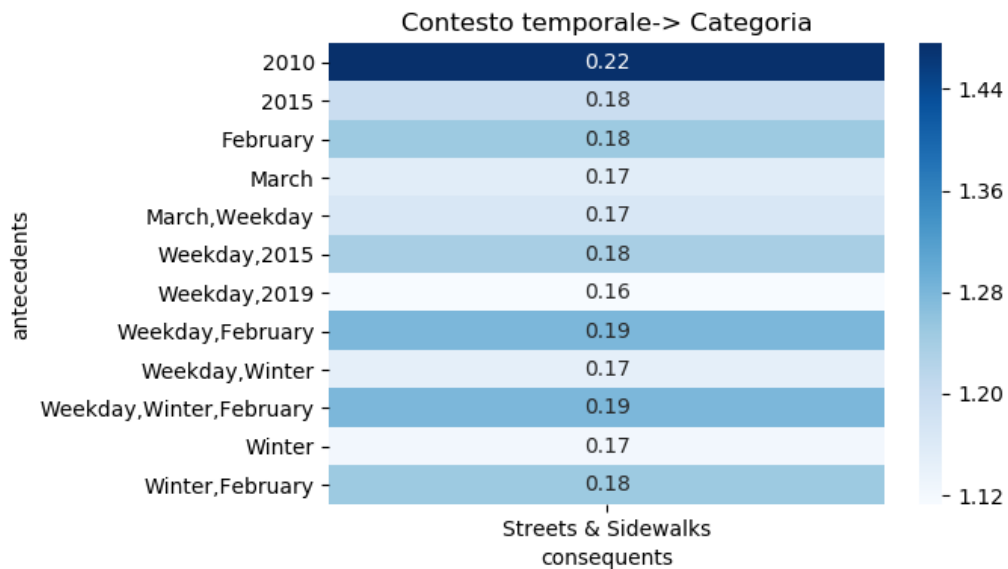


Figura 5.27: Regole associative contesto temporale → Categoria Streets& Sidewalks

Nella figura successiva, 5.28, si descrivono le regole da un punto di vista puramente metrico per il contesto spaziale e temporale che implica una data segnalazione stradale. Sono estratte ben 112 regole. Cercheremo di analizzare i valori informativi salienti. I valori di confidence per queste regole variano nel range di 0.22 a 0.46. La regola più forte ha lift di circa 3, la meno forte ha un lift di 1.46. Si possono estrapolare diverse considerazioni interessanti relativamente a questa segnalazione di non emergenza. Se già prima avevamo detto che il distretto di Queens implicava il verificarsi di questa tipologia di segnalazione urbana, qui la cosa è rimarcata nuovamente, possiamo aggiungere a tale informazione geografica che segnalazioni di questo tipo nelle strade di questo distretto sono correlate anche al periodo invernale e ai relativi mesi, in maniera più robusta con il mese di Febbraio. Stessa cosa seppur con valori leggermente diversi di confidence e lift, possiamo dire per il distretto di Brooklyn. Inoltre il distretto di Brooklyn implica anche questa tipologia di segnalazioni nelle altre stagioni. Mentre il distretto di Queens associa queste segnalazioni maggiormente a mesi primaverili ed autunnali, ma si trovano anche ottime correlazioni nei mesi estivi e invernali. Nelle regole associative in cui compare il distretto di Staten Island non si osserva nessuna associazione con mesi o stagioni, bensì con i giorni settimanali. Nessuna regola si trova per questo problema generale per quanto riguarda il distretto di Manhattan. Gli anni maggiormente più forti in termini di correlazione sono in ordine decrescente 2015, 2010, 2014, 2017, 2018, questo attributo si trova correlate insieme al luogo specifico, strada o marciapiede, e giorni feriali. Diverse sono le correlazioni con i giorni feriali, possiamo affermare che vedendo l'andamento delle regole da un punto di vista metrico, i vari giorni settimanali e mesi annuali sono presenti in maniera piuttosto regolare, si potrebbe concludere che non vi è un particolare periodo in linea generale da monitorare, tranne se questo risulta associato a qualche precisa zona della città. Vi è anche una regola relativa ai giorni festivi, che mostra che il 28% delle segnalazioni fatte nelle giornate festive sono riferite a luoghi stradali e riguardano problemi di categoria Streets & Sidewalks.

Sezione sperimentale

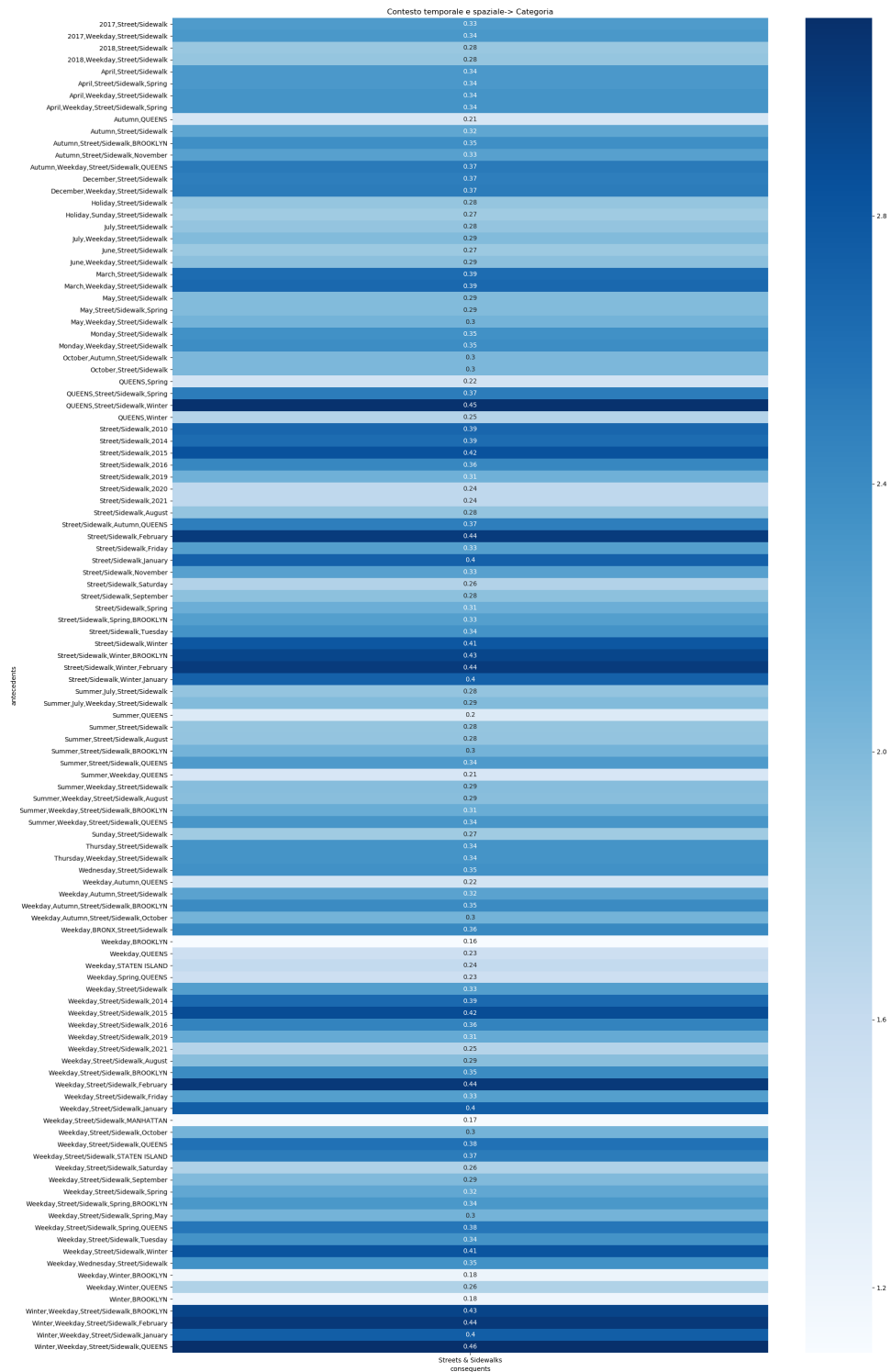


Figura 5.28: Regole associative contesto temporale e spaziale → Categoria Streets & Sidewalks

Ultima categoria che presenta delle regole associative è la categoria Transportation che riguarda segnalazioni urbane relativamente ai mezzi di trasporto. Nella figura 5.29 troviamo le 7 regole associative di tipo contesto spaziale → Transportation. Troviamo che indicativamente nel 20% delle segnalazioni fatte in luoghi stradali delle aree di Brooklyn, Queens e Bronx si verificano nella categoria Transportation. La regola più "forte" risulta essere quella relativa al distretto di Brooklyn, anche se per le altre due regole il lift si abbassa di poco. Il distretto di Manhattan implica anche questa tipologia di segnalazioni ma con valori di confidence e lift più bassi.

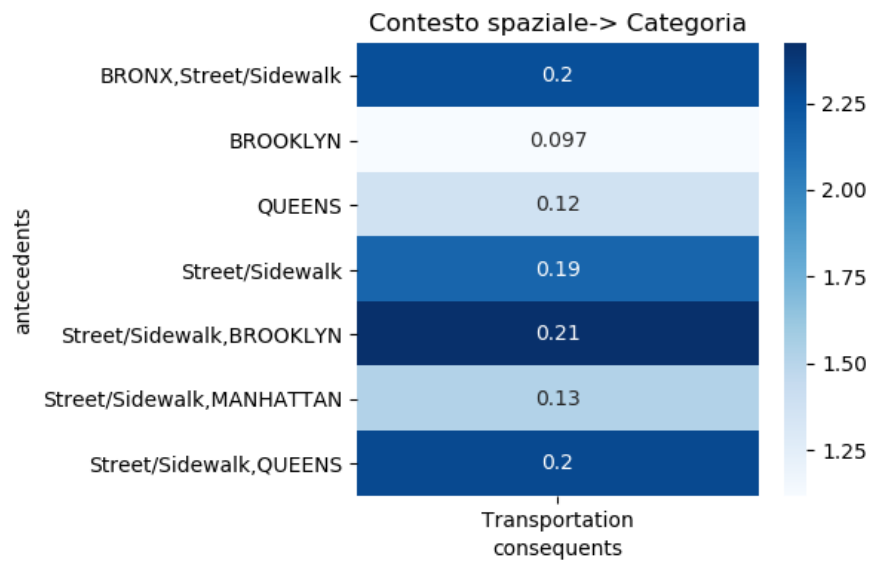


Figura 5.29: Regole associative contesto spaziale → Categoria Transportation

Nella figura 5.30 troviamo le regole per contesto temporale per la categoria Transportation. Particolarmente interessante la presenza degli anni dal 2018 al 2021. Questo ci porta a pensare che probabilmente negli ultimi anni c'è una maggiore attenzione a questa tipologia di segnalazione oppure si è registrato un aumento di problemi legate a tale categoria. Inoltre per l'anno 2021 vi è una associazione insieme al valore giorni feriali.

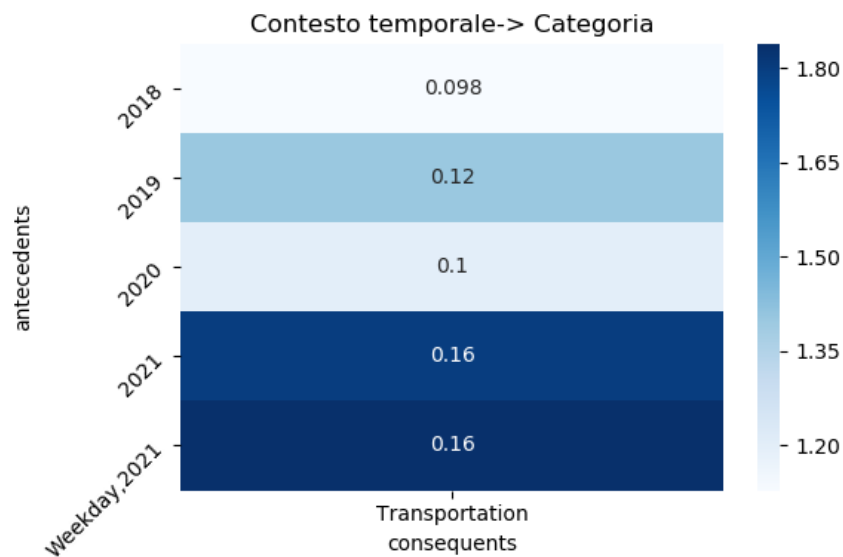


Figura 5.30: Regole associative contesto temporale → Categoria Transportation

Le regole generate prendendo in considerazione gli attributi spaziali e temporali contemporaneamente offrono un maggior contenuto informativo. Otteniamo 31 regole, inserite nella tabella 5.31, con lift compreso tra 0.1 e 0.3, e valore di lift nel range 1.1 - 3.4. Troviamo associazione con gli anni precedentemente osservati, 2018 al 2021, e qui sono accomunati con luoghi stradali. Raggiungendo valori di confidence intorno a 29% e un'ottima robustezza, valori di lift > 3 per gli anni 2021 e 2020, > 2 per i restanti. A livello di distretti, troviamo correlazioni per questa categoria con Brooklyn e Bronx in aree stradali e nei giorni feriali. Molto interessante anche l'associazione Autumn, Street/Sidewalks → Transportation. Buona correlazione anche con lo stesso pattern di regola ma con le altre tre stagioni. Particolare il fatto che la domenica è il giorno che associa con maggiore robustezza questa tipologia di segnalazione urbana.

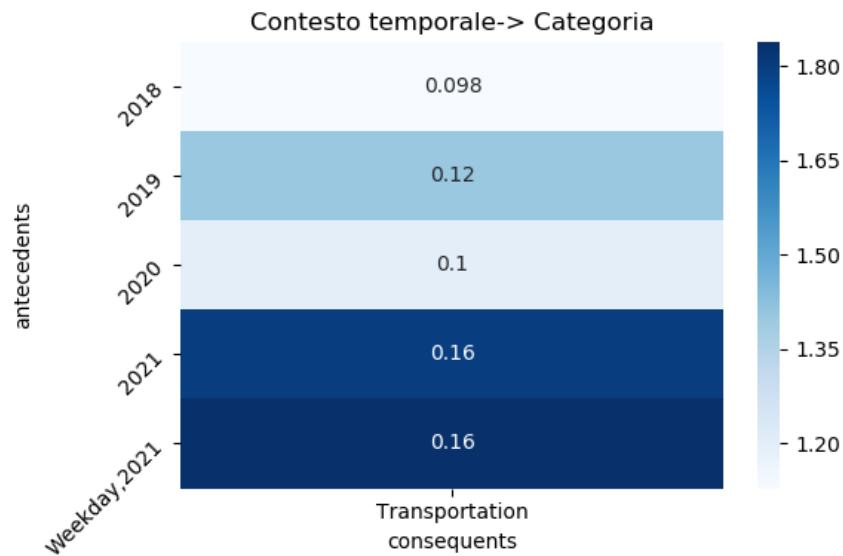


Figura 5.31: Regole associative contesto temporale e spaziale → Categoria Transportation

Quello che emerge da questa prima fase di estrazione di regole associative è già particolarmente rilevante. Anche se vi sono attributi spaziali che non sono emersi in questa analisi, per esempio per alcune tipologie di luoghi e per la maggior parte delle community board non sono state trovate correlazioni. Inoltre come era già stato detto alcune categorie non presentano relazioni rilevanti tra gli attributi del dataset. Inoltre si ipotizza che considerando un livello gerarchico più dettagliato si possano trovare correlazioni più precise e migliori in modo da poter apprezzare correlazioni maggiormente dettagliate.

Nel prosieguo di questo paragrafo analizzeremo i risultati delle regole in cui si considera l'attributo Complaint Type, che è relativo alla sottocategoria. I risultati presentati riportano la struttura sottocategoria → contesto spaziale/temporale o entrambi. Anche in questo caso osserveremo che le sottocategorie estratte sono relative alle stesse categorie esaminate in precedenza.

La figura 5.32 riporta le regole associative delle sottocategorie relative a Cleaning & Garbage.

Troviamo 5 regole in cui compaiono cinque differenti sottotipi, e tutte associate ad attributo Location Type. Si osservano regole molto valide, in tutte si osservano valori di confidence abbastanza alti, in due casi lift > 2. Troviamo segnalazioni di rifiuti ingombranti che si verificano con una frequenza del 99% su strada o marciapiede, e lift 2.19. Altra regola riguarda la mancata raccolta sempre su strada o marciapiede, con confidence del 98% e lift 2.1. Problemi di acqua piovana e relativa

rete fognaria correlata a luogo stradale, altra regola ci fornisce l'informazione che è frequente la presenza di rifiuti nelle strade. Mentre problemi di poca condizione sanitaria si associano a luoghi abitativi. Per quanto riguarda regole che implicano un contesto temporale non si riescono ad ottenere regole particolarmente valide.

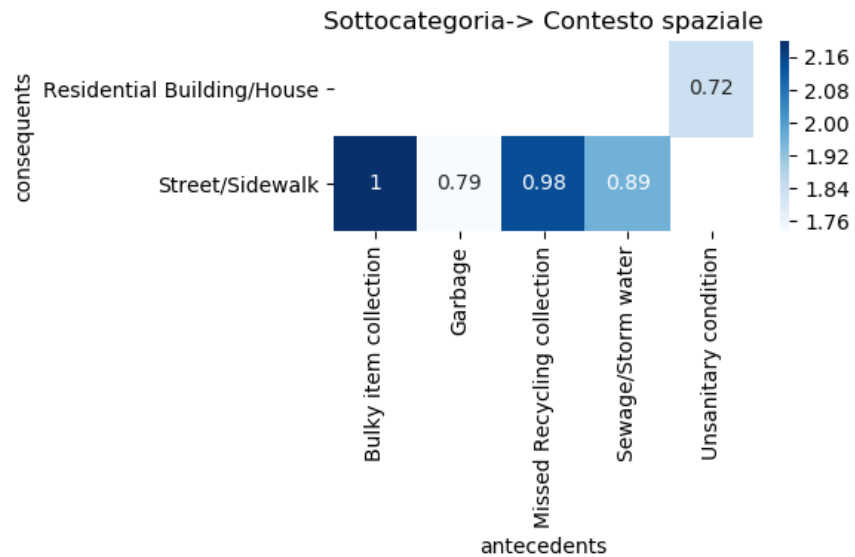


Figura 5.32: Regole associative sottocategoria Cleaning & Garbage → contesto spaziale

Le regole sottocategoria Cleaning & Garbage → contesto spaziale e temporale ci forniscono 5 regole, che sono a livello di contenuto spaziale uguale alle precedenti regole estratte, in aggiunta qui sono correlate non solo con il luogo ma anche con l'attributo che indica giorno feriale. Da un punto di vista metrico rimangono oggettivamente regole forti.

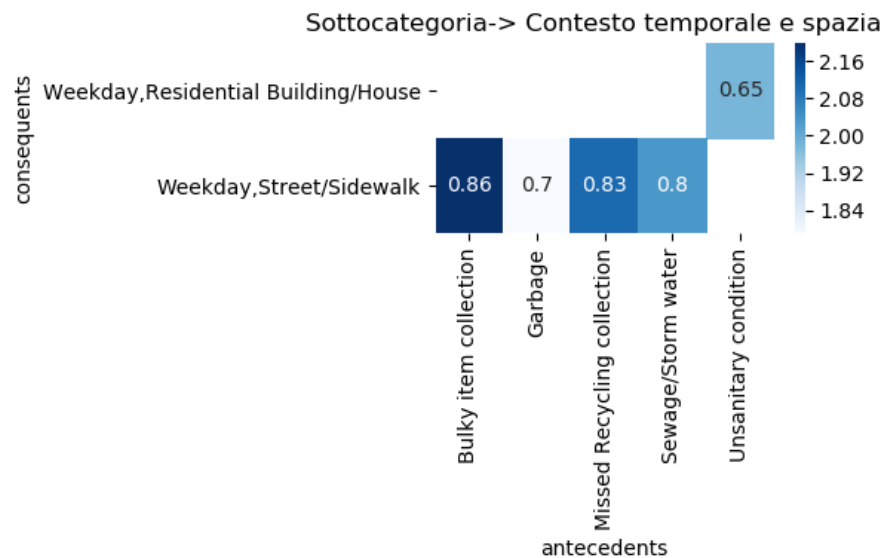


Figura 5.33: Regole associative sottocategoria Cleaning & Garbage → contesto temporale e spaziale

Le prossime sottocategorie ad essere esaminate sono quelle della categoria Complaint. Nella figura 5.34 si rappresentano le regole associative estratte con le rispettive metriche di valutazione. Le sottocategorie presenti in queste regole sono: Street noise e Residential noise. Problemi di rumore stradale, secondo questa regola, si potrebbero verificare in strada e nel distretto di Manhattan con una percentuale di confidence del 30% e lift decisamente maggiore di 1, circa 3.7. Questo ci fornisce come informazione che non potrebbe essere raro il verificarsi di fastidi a livello sonoro tra le strade di Manhattan, da queste informazioni si potrebbe etichettare come un distretto caotico. Stesso pattern di regola ma metricamente meno robusta è la relazione con il distretto di Bronx e poi Brooklyn. Particolare anche se abbastanza ovvia la regola che associa problemi sonori residenziali a luoghi appunto residenziali, troviamo confidence 1 e lift di circa 2.5. Problemi di vicinato rumorosi si trovano correlati anche a Manhattan, con un livello di confidence pari a 21% e lift circa 2.59. La presenza di disturbo della quiete si verifica anche in distretti come Bronx e Brooklyn.

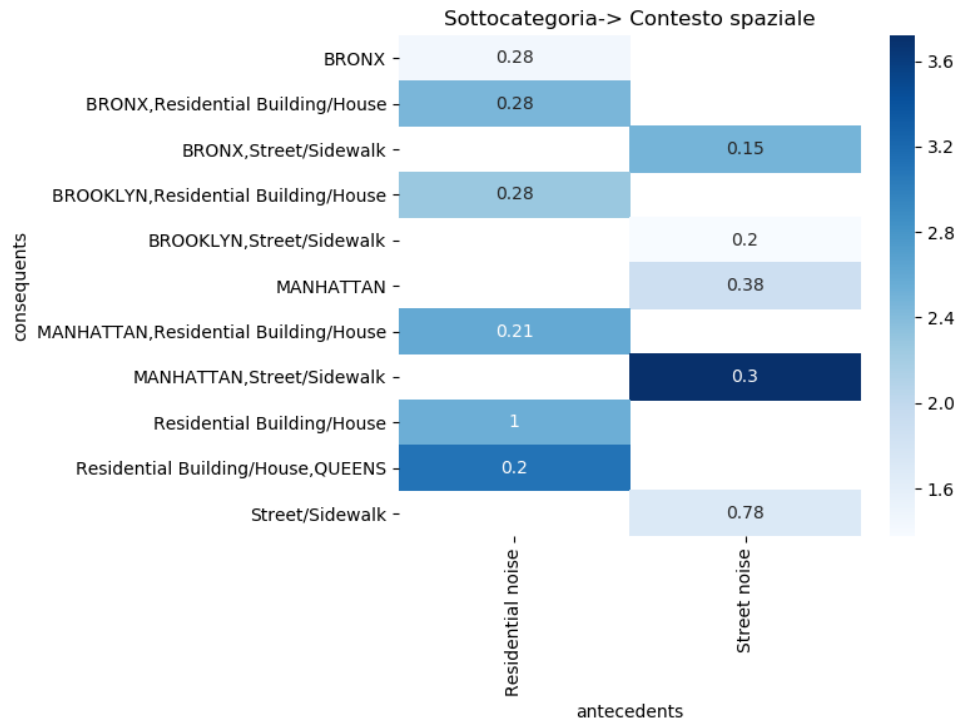


Figura 5.34: Regole associative sottocategoria Complaint → contesto spaziale

Per quanto riguarda le regole sottocategoria Complaint → contesto temporale troviamo il tutto riportato nella tabella 5.35. Le sottocategorie osservate sono identiche alle precedenti regole estratte, si trovano maggior frequenza di problemi di disturbi della quiete residenziale precisamente durante la Domenica e il Sabato. Troviamo parecchie regole che associano lo stesso problema agli anni 2020 e 2021, anche se principalmente nei giorni feriali. Particolare che il verificarsi di questo tipo di segnalazioni residenziali è associato al mese di Settembre e Maggio, anche se i valori di confidence sono bassi e lift di poco superiore a 1. Problematiche su rumori in strada potrebbero verificarsi, guardando le regole, il Sabato e la Domenica. Forse la causa potrebbe essere dovuta a possibili rumori dovuti a disturbi provenienti da locali e pub, in quanto principalmente nei weekend questi posti sono maggiormente frequentati, soprattutto dai giovani. Inoltre si scopre che si verificano con una certa frequenza a Maggio, Giugno, Agosto e Settembre. Questo ci fa intuire che nei periodi dell'anno maggiormente "freddi" probabilmente ci sono meno problematiche dovute ai rumori/disturbi stradali.

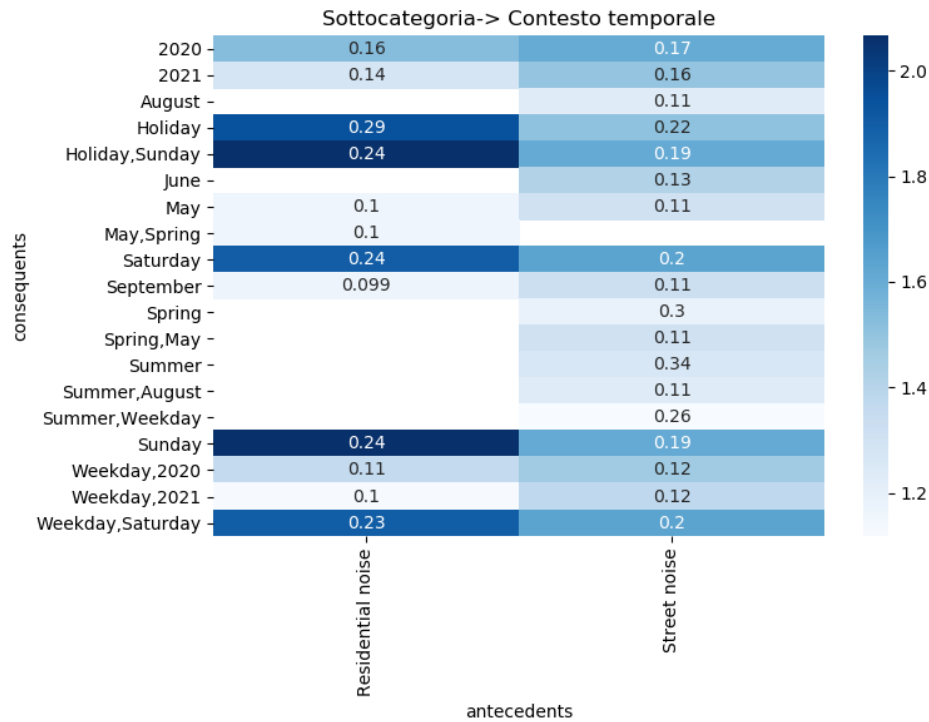


Figura 5.35: Regole associative sottocategoria Complaint → contesto temporale

Nella tabella 5.36 sono presenti le regole associative di tipo sottocategorie di tipo Complaint → contesto temporale e spaziale. Troviamo diverse relazioni forti e robuste, come la regola Residential noise → sunday,residential buliding che ha una confidence del 24% e lift alto, 4.81. Valori anche simili per la regola in cui invece è presente il giorno Sabato. Altra regola abbastanza forte è quella che associa problemi di rumori in strada nelle strade di Manhattan nei giorni feriali. Ci sono regole che mostrano anche una relazione tra questo distretto, la stagione estiva o primaverile e i rumori stradali. Si potrebbero verificare disturbi nelle abitazioni nel periodo primaverile e precisamente nel mese di Maggio, ma anche nel mese di Settembre e quindi stagione Autunno, queste regole raggiungono un valore di lift superiore a 3. In linea generale troviamo segnalazioni su rumori in zone residenziali nei giorni feriali nei distretti di Queens, Bronx, Brooklyn e Manhattan, e Brooklyn a livello di metriche risulta la meno "forte" tra queste regole. Rumori stradali, anche se meno regole meno robuste, si associano ai distretti di Bronx e Brooklyn. Vi è una relazione piuttosto buona che associa i rumori provenienti dalle abitazioni della città il Venerdì, meno forte la regola che associa questo giorno ai rumori provenienti dalle vie della città di New York.



Figura 5.36: Regole associative sottocategoria Complaint → contesto temporale e spaziale

Nelle figure sotto sono riportate le regole generate per le sottocategorie di tipo Information. La sottocategoria presente nell'unica regola per il pattern contesto spaziale è relativa al dipartimento di finanza per quanto riguarda tasse e pagamenti. Mostra una correlazione tra questa tipologia di segnalazione e luoghi abitativi, con confidence parecchio alta 96% e lift 2.44. Se si prende il contesto temporale questo sottotipo di segnalazione è effettuata con maggior frequenza durante i giorni feriali, nelle regole in cui si prendono sia gli attributi spaziali e temporali come conseguente si può notare che si associano, ad entrambi, gli attributi spaziali e temporali emersi dalle prime regole valutate singolarmente.

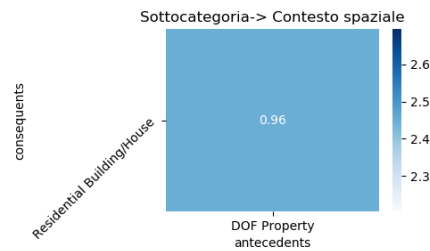


Figura 5.37: Regole associative sottocategoria Information → contesto spaziale

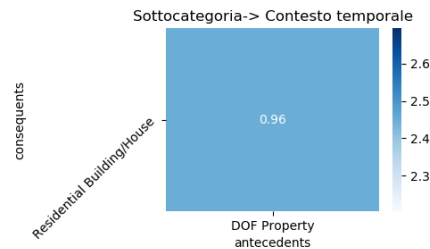


Figura 5.38: Regole associative sottocategoria Information → contesto temporale

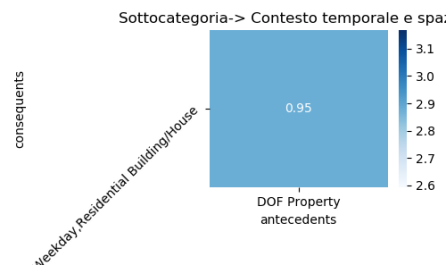


Figura 5.39: Regole associative sottocategoria Information → contesto temporale e spaziale

Si analizzano i risultati per le sottocategorie di Maintenance, figura 5.40. Si ottengono regole con confidence che variano tra 0.21 e 0.75, lift massimo ottenuto da una regola è pari a 8, ci sono anche diverse regole con buoni valori di lift. La regola che risulta più forte è quella che associa problemi ai sistemi dell'acqua in luoghi non specifici, questo ci permette di dedurre che spesso quando si effettuano segnalazioni ai problemi del sistema idrico pubblico non vi è un riscontro certo con la tipologia di luogo in cui si verifica il problema, anche se intuitivamente si potrebbe pensare a luoghi stradali. Stessa tipo di associazione per la sottocategoria Construction and Plumbing, anche se a livello di metriche leggermente minore. A livello di informazione, seppur meno forti, sono le regole che mostrano il verificarsi con una certa frequenza di problemi di riscaldamento del sistema dell'acqua nelle case e residenze dei distretti di Bronx, Manhattan, Brooklyn e Queens. Inoltre questa tipologia di segnalazione insieme a problematiche di riparazione e di tipo elettrico sono maggiormente frequenti nelle abitazioni, abbiamo valori di confidence parecchio alte, 0.98 per la prima, 0.92 per la seconda, 0.91 per l'altra, lift superiori a 2. Segnalazioni causate da problemi edilizi e idrici potrebbero verificarsi, in ordine di confidence e lift, nelle abitazioni situate nei distretti di Brooklyn e Bronx. Altro contenuto emerso dalle regole è la correlazione di segnalazioni di richieste di possibili interventi di riparazione negli edifici con i distretti di Bronx e Brooklyn.

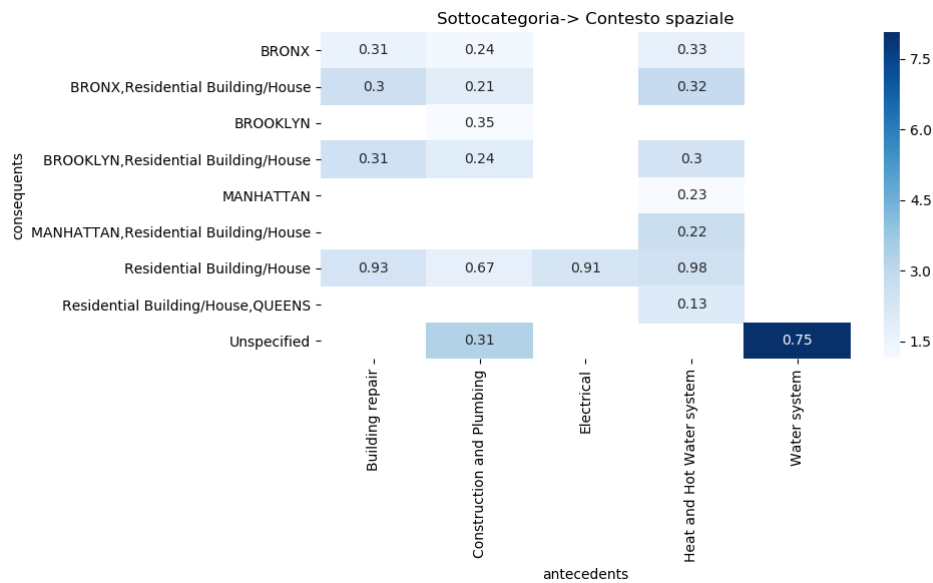


Figura 5.40: Regole associative sottocategoria Maintenance → contesto spaziale

Si trovano 32 possibili correlazioni tra questo sottotipo di segnalazioni e il contesto temporale. Le associazioni evidenziano correlazioni con determinati periodi dell'anno. Problemi di riscaldamento potrebbero registrarsi nei mesi invernali, con maggiore probabilità durante il mese di Gennaio e Dicembre, rispetto a Novembre e Febbraio. Principalmente le regole più forti associano anche queste segnalazioni con più frequenza nei giorni feriali, tanto è vero che ci sono correlazioni meno forti anche con i giorni festivi, principalmente Domenica. Segnalazioni su problemi alla rete idrica pubblica si verificano con un valore di confidence vicino al 31% durante i giorni feriali estivi, con lift 1.33. Problematiche edilizie e idriche sono correlate a determinati giorni della settimana che vanno da Lunedì a Giovedì. Inoltre si osserva una regola che implica richieste di questo tipo nei giorni feriali di Primavera.

Se si prendono in considerazione le regole che associano le sottocategorie di Maintenance a contesti spaziale e temporali, troviamo molte regole robuste, in cui si trovano valori di lift superiore a 2. La mappa in figura 5.42 riporta le regole associative e i valori delle metriche per quanto riguarda questo tipo di pattern. La regola più forte mette in relazione problemi al sistema idrico pubblico durante l'Estate e in posti non specificati. Si ottiene un lift di 9.9. Mentre a seguire abbiamo la regola che associa questo tipo di segnalazione (Water System) ai giorni settimanali, lift 7.9. Inoltre notiamo problemi di riscaldamento principalmente nelle abitazioni e il mese di Gennaio, ma ci sono regole altrettanto forti che descrivono problemi di riscaldamento negli edifici residenziali in tutto il periodo invernale, e quindi nei rispettivi mesi. Anche se Gennaio risulta l'associazione di più forte. In linea generale vi è correlazione tra queste richieste di non emergenza nelle case e nei giorni settimanali e in diversi distretti della città. Quanto ottenuto è in linea con il periodo maggiormente freddo dell'anno in cui nelle abitazioni sono accesi i mezzi di riscaldamento, e quindi potrebbe crescere la probabilità di maggiori segnalazioni. Vi è una forte relazione tra queste segnalazioni sul sistema di riscaldamento dell'acqua durante l'Inverno nelle case degli abitanti del distretto di Bronx, lift 4.7. Ma si osservano anche regole altrettanto robuste anche in diversi altri distretti cittadini, Brooklyn, Manhattan. Inoltre non rare le possibilità di avere queste problematiche nei giorni festivi e nelle abitazioni (lift 2.7). Segnalazioni per quanto riguarda riparazioni negli edifici si potrebbero verificare principalmente nelle abitazioni del distretto di Brooklyn durante l'Estate. Vi è anche comunque una correlazione tra le stagioni Autunno e Inverno nelle abitazioni e le segnalazioni Building repair. I problemi di riparazione nei distretti di Bronx si verificano principalmente nei giorni feriali. Altre informazioni, seppur meno robuste, sono ottenute dalle regole che mostrano la correlazione tra la sottocategoria Construction and Plumbing e i distretti di Bronx e Brooklyn, principalmente alle abitazioni e nei giorni feriali.

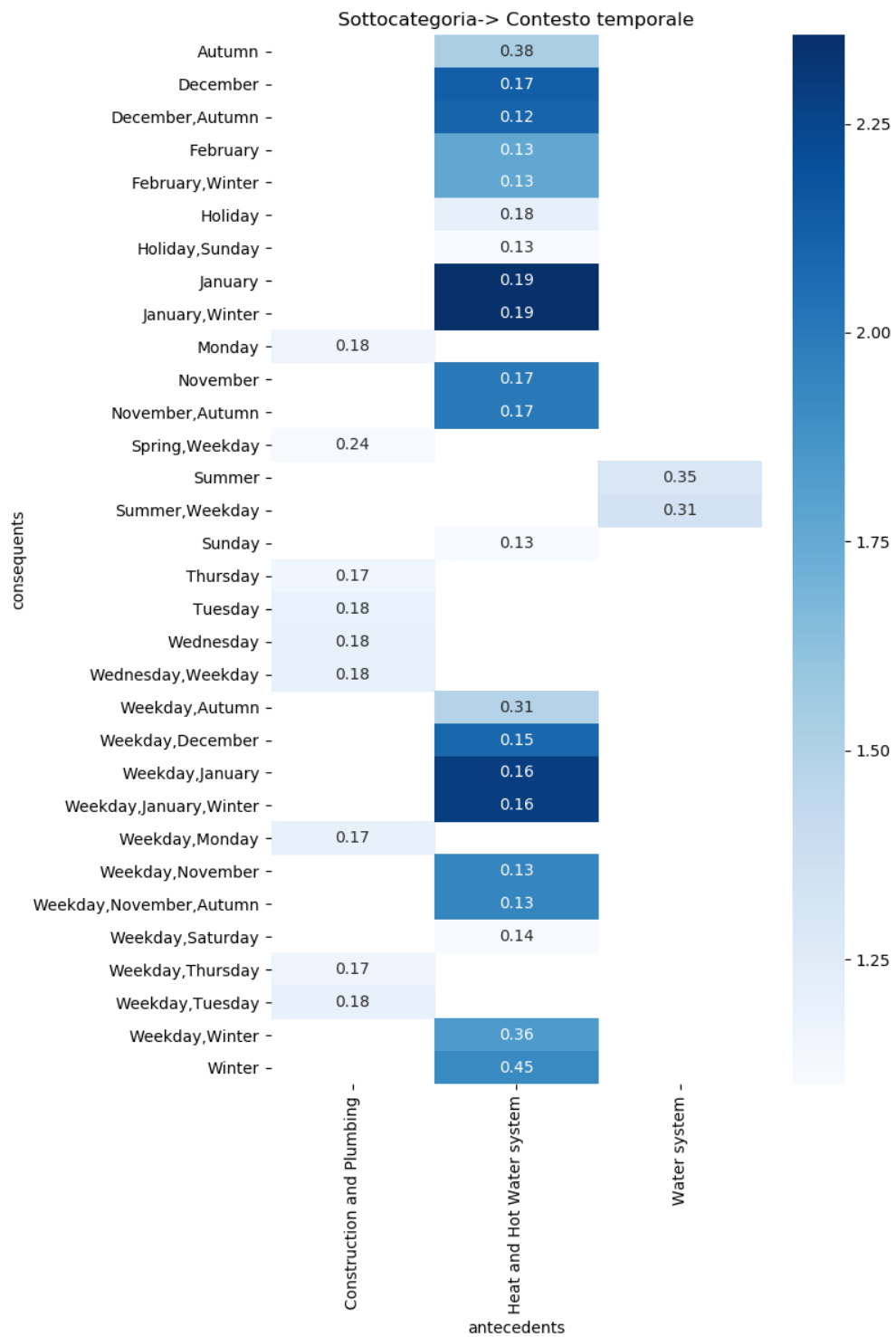


Figura 5.41: Regole associative sottocategoria Maintenance → contesto temporale

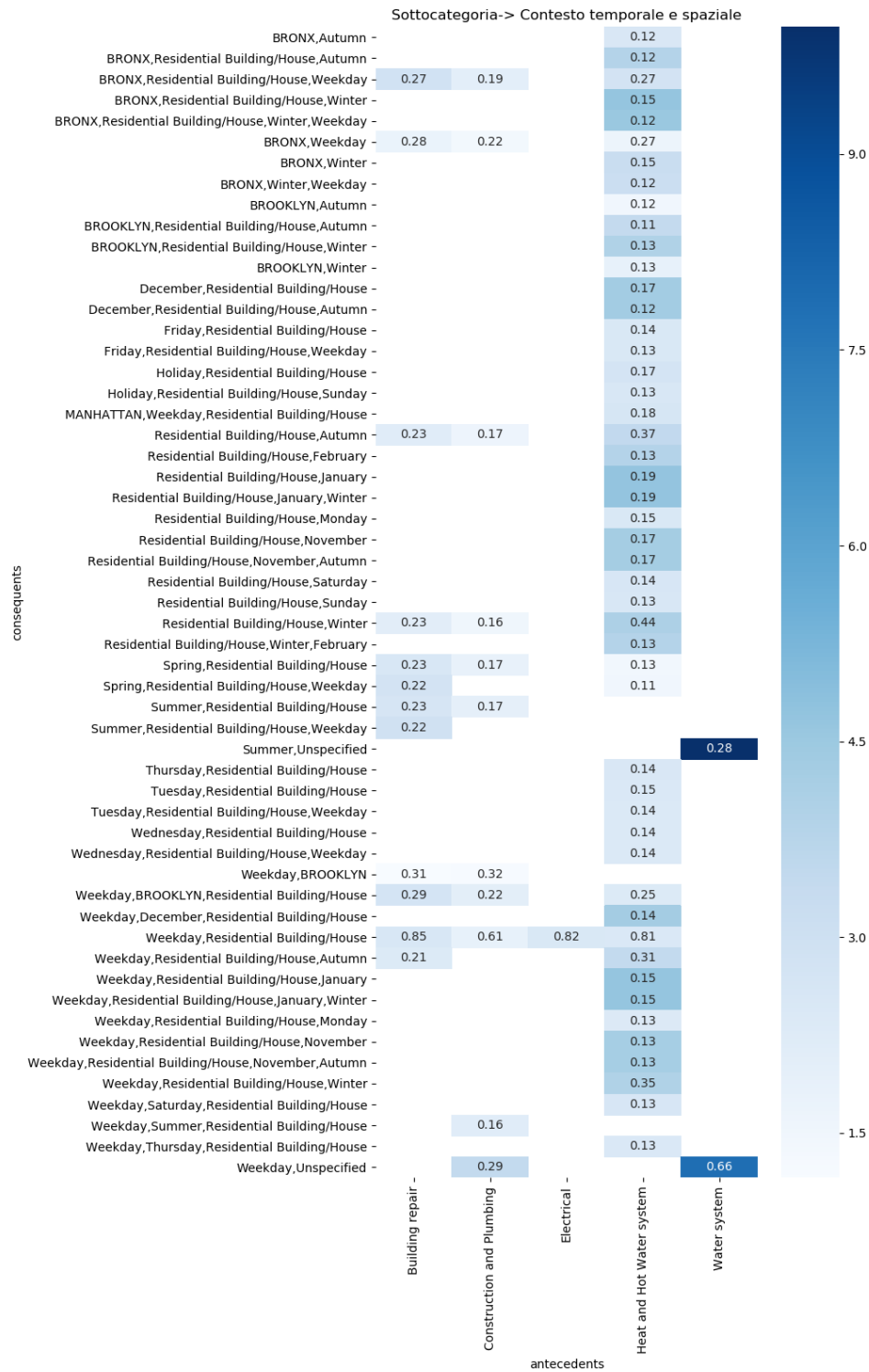


Figura 5.42: Regole associative sottocategoria Maintenance → contesto temporale e spaziale

Analizzeremo ora le regole estratte per le sottocategorie appartenenti a Park and Trees. Dalle regole nella figura 5.43 riusciamo ad estrarre informazioni con il contesto spaziale della città, dalle 3 regole l'informazione collettiva è che problemi legati alla condizione degli alberi potrebbero essere segnalati tra le strade del distretto di Queens. La regola presenta una confidence del 41% e valore di lift pari a 2.98.

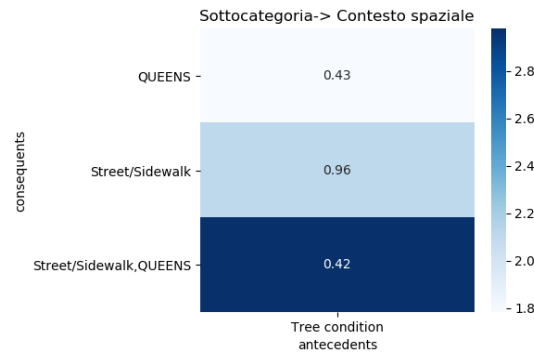


Figura 5.43: Regole associative sottocategoria Park & Trees → contesto spaziale

La figura 5.44 si riferisce al solo contesto temporale. Riusciamo a dedurre che segnalazioni sulla cattiva condizione di alberi si registrano con una certa frequenza durante le giornate feriali del periodo estivo.

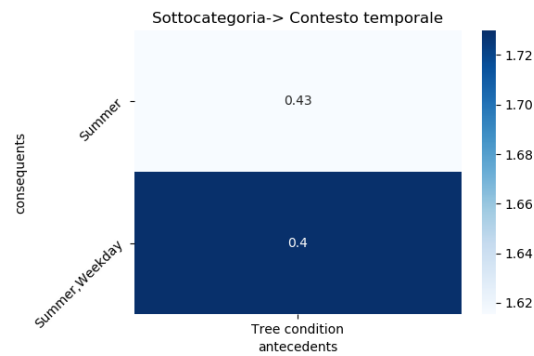


Figura 5.44: Regole associative sottocategoria Park & Trees → contesto temporale

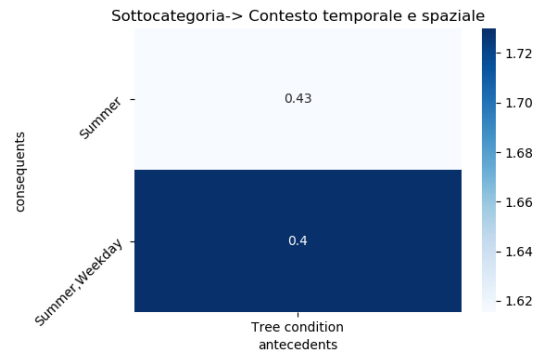


Figura 5.45: Regole associative sottocategoria Park & Trees → contesto temporale e spaziale

Mentre la mappa 5.45 mette in relazione questa sottocategoria sia al contesto spaziale che temporale, e dalla regole si rimarkano alcune informazioni precedenti, in aggiunta si associano i problemi legati al distretto di Queens principalmente nei giorni feriali, stessa cosa la correlazione con il periodo estivo. Le regole estratte rimangono comunque molto valide e robuste.

Risultano molto interessanti le indicazioni fornite dalle regole in figura 5.46 sui sottotipi di Streets & Sidewalks relativamente al contesto geografico della città di New York. Compaiono 4 sottocategorie: Street/Road/Sidewalk obstruction, Street/Sidewalk repair, Street light repair e Street/Road sign repair. Il 40% di segnalazioni di ostruzione stradale si registrano nelle strade del distretto di Queens, con lift 2.86. Stesse segnalazioni con un livello di confidence e lift leggermente minore lungo le strade del distretto di Brooklyn. Dalle regole emerge anche nelle strade del distretto di Queens e Brooklyn si potrebbero verificare segnalazioni di riparazione stradale. In linea generale le regole confermano che segnalazioni per riparazioni di segnali e luci stradali si verificano lungo le strade.

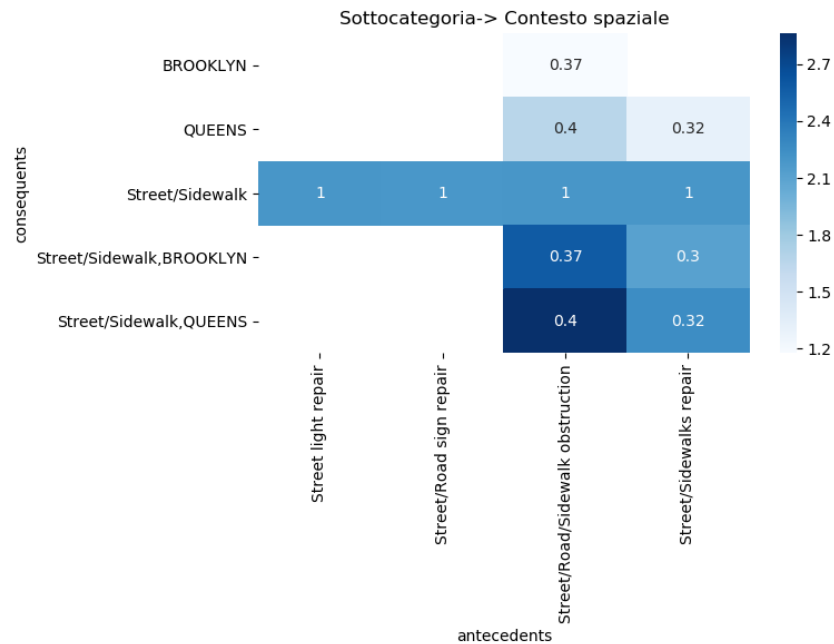


Figura 5.46: Regole associative sottocategoria Streets & Sidewalks → contesto spaziale

Per quanto riguarda la sola associazione con il contesto temporale emergono particolari che potrebbero rivelarsi utili, si associano richieste di riparazioni stradali nei giorni feriali d'Estate e Primavera. Inoltre si evince che nel 95% di segnalazioni relative a problemi legati all'illuminazione stradale sono effettuate nei giorni feriali. Troviamo il tutto riportato nella figura 5.47.

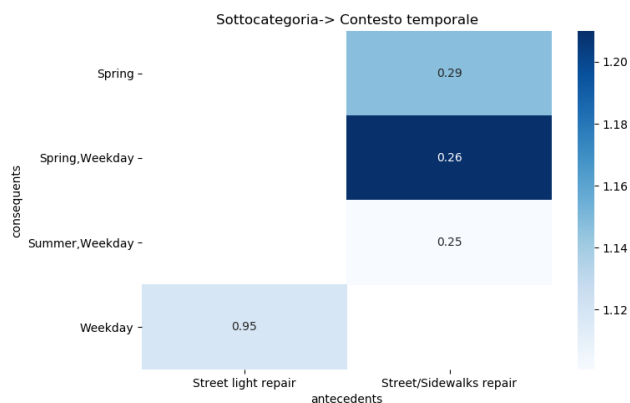


Figura 5.47: Regole associative sottocategoria Streets & Sidewalks → contesto temporale

Nella tabella 5.48 troviamo le correlazioni se si valuta insieme il contesto temporale e geografico per le sottocategorie associate a Streets & Sidewalks. Molto particolare il fatto che si hanno correlazioni abbastanza forti tra luoghi e stagioni che nell'analisi solo temporale non erano presenti. È il caso delle regole che associano problemi di ostruzione stradale lungo le strade nel periodo Invernale e Autunnale, lift per entrambe indicativamente 2.5. Invece la stessa correlazione con le stagioni Estate e Primavera presenta metriche inferiori a livello di lift. La regola più forte è quella che evidenzia la possibilità di segnalazioni relativi a possibili ostacoli stradali lungo le strade di Queens nei giorni feriali. Stessa relazione si ha con il distretto di Brooklyn. Alla stagione Inverno si associano anche segnalazioni di riparazione delle strade/marciapiedi, ma correlazione piuttosto forte è legata anche alla stagione Primavera. Leggermente più basso è il valore di lift che mette in relazione segnalazioni di riparazioni stradali lungo le strade nel periodo estivo e autunnale. Inoltre dalle regole si apprende che potrebbero verificarsi segnalazioni di riparazioni su strada principalmente nei distretti Queens e Brooklyn nei giorni settimanali.

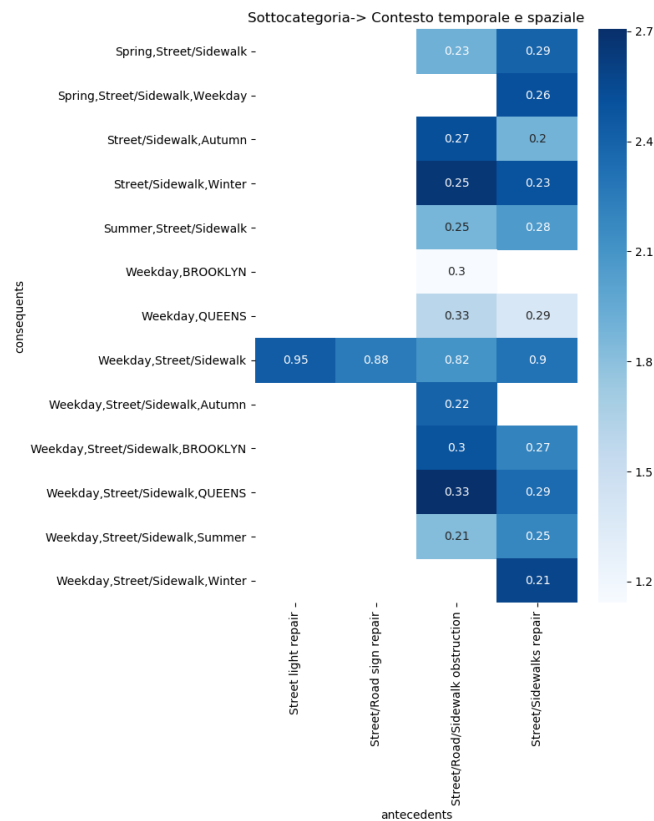


Figura 5.48: Regole associative sottocategoria Streets & Sidewalks → contesto temporale e spaziale

Per ultimo si valutano le sottocategorie associate a Transportation. Se si considera la tabella 5.49 che mette in relazione le sottocategorie agli attributi geografici del dataset riusciamo a dedurre alcune informazioni. Segnalazioni di parcheggi poco consoni si verificano con confidence di circa il 36% lungo le strade di Brooklyn, e con confidence del 31% nel distretto di Queens. Le rispettive regole hanno lift > 2 . Inoltre si prende atto che segnalazioni di veicoli abbandonati si registrano principalmente lungo le strade.

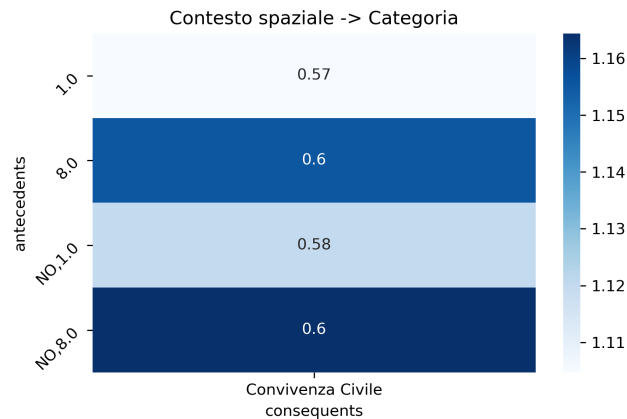


Figura 5.49: Regole associative sottocategoria Transportation $\rightarrow$ contesto spaziale

Per quanto riguarda l'associazione con gli attributi temporali, figura 5.50, si deduce che invece il 2021 è risultato un anno parecchio problematico da questo punto di vista, a quanto pare ci sono stati parecchi disagi legati ai parcheggi illegali.

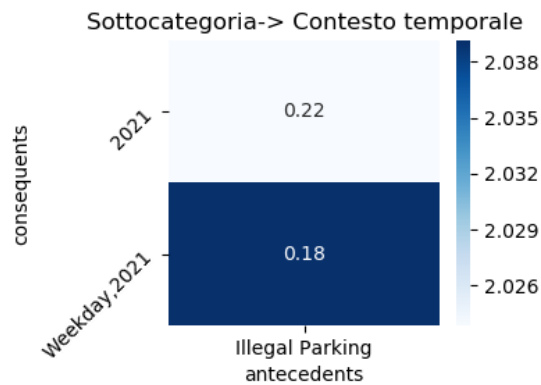


Figura 5.50: Regole associative sottocategoria Transportation $\rightarrow$ contesto temporale

Le regole in cui si prendono in esame sia il contesto temporale che spaziale possono aiutarci ad avere informazioni aggiuntive. Le regole così strutturate sono contenute nella mappa in figura 5.51. Principalmente le informazioni riguardano i parcheggi illegali, riusciamo ad aggiungere nuove informazioni rispetto a quanto detto in precedenza. Troviamo regole robuste con $\text{lift} > 2$. Questa sottocategoria di segnalazione ha avuto una certa incidenza lungo le strade nei giorni feriali del 2021. Inoltre si potrebbero verificare segnalazioni di parcheggi illegali tutto l'anno, in quanto nelle regole ci sono associazioni con tutte le stagioni. Leggermente più correlate le regole di questo tipo con Autunno e Inverno. Il distretto maggiormente colpito da questo disagio sembra essere quello di Brooklyn. Leggermente minori i valori delle metriche per il distretto Queens. Dalle regole si deduce anche la relazione di parcheggi illegali in tali distretti con una certa frequenza nei giorni feriali. L'unica sottocategoria differente è legata ai veicoli abbandonati, viene aggiunta l'informazione che ci indica che le segnalazioni relative a questo disagio sono registrate su strada e nei giorni feriali.

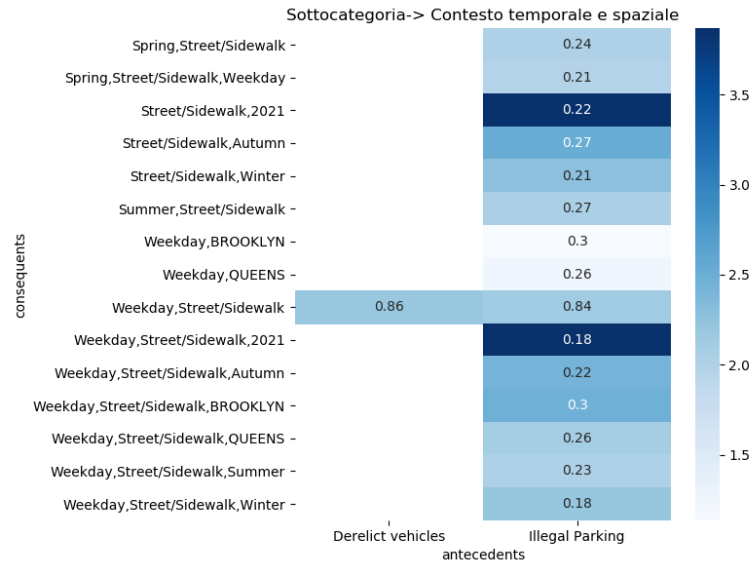


Figura 5.51: Regole associative sottocategoria Transportation → contesto temporale e spaziale

Quanto emerso da questa prima analisi ci permette di delineare delle linee guida per identificare delle situazioni da monitorare in città per determinate tipologie di segnalazioni. Abbiamo osservato che ci sono delle zone della città in cui si verificano determinate richieste e inoltre si potrebbero identificare degli archi temporali precisi in cui si verificano con maggiore frequenza. Oltre a fornirci alcune considerazioni interessanti, questa analisi ci permette di capire il potenziale informativo di questi dati urbani per adottare delle misure per risolvere problematiche all'interno della città.

5.5 Flipping correlation

Nonostante la conoscenza estratta dalle regole di associazione è stato deciso di continuare l'analisi cercando di trovare ulteriori correlazioni particolari tra i dati. In particolar modo si è scelto un algoritmo molto interessante, il Flipper, che permette di estrarre correlazioni flipping (contrastanti) tra livelli gerarchici. Questo algoritmo si potrebbe adattare bene ai nostri dati sia perché tra gli attributi si possono individuare determinate gerarchie sia perché l'algoritmo si adatta bene agli itemset con supporto medio-basso. Per la corretta esecuzione di questo algoritmo il dataset è stato strutturato in tre diversi file da dare in input all'algoritmo. Si è realizzato un file di tipo dizionario, con almeno 2 livelli di gerarchia per elemento. Formato: ID elemento, seguito dal livello di astrazione più elevato ed eventualmente da livelli più specifici. Il successivo file da dare in input è strutturato in questo modo: righe per ogni transazione, ogni elemento della transazione è indicato con il rispettivo ID ed è separato da un delimitatore. L'ultimo file è relativo ai supporti, per ogni livello della gerarchia si specifica il suo supporto minimo. Inoltre si danno in input all'algoritmo Flipper le due soglie di supporto minimo che ci aiuteranno a definire se una correlazione è positiva o negativa. Si sceglie una soglia minima positiva così che tutte le correlazioni al di sopra di essa sono considerate positive. Viene scelta anche una soglia massima negativa che indica che tutte le correlazioni al di sotto di essa sono considerate negative. L'intervallo di valori per le soglie è compreso tra 0 e 1. L'output del programma permetterà di visualizzare diverse righe che includono gli elementi a diversi livelli di astrazione e la rispettiva correlazione. Viene visualizzato anche il valore della rispettiva metrica di *Kulczynsky* usata per definire la correlazione.

Sono state effettuati diversi tentativi e quindi sono stati dati in input diverse strutture dati. Principalmente sono stati formati gerarchie a 2 livelli, si è tenuto fissa la parte di attributo temporale o spaziale, ed è stata associata ai 2 livelli, categoria e sottocategoria. Successivamente il dataset è stato sempre organizzato in 2 livelli gerarchici, mantenendo fissa la parte categoria o sottocategoria e facendo variare gli attributi temporali o spaziali. Per ultimo queste due tipologie di strutture

sono state allargate a 3 livelli gerarchici.

5.6 Flipping correlation sulle segnalazioni di Torino

In questo paragrafo presenteremo alcune correlazioni di tipo flipping estratte dal dataset sulle segnalazioni urbane della città di Torino. Per primo vedremo i risultati ottenuti dai vari dataset organizzati in due livelli gerarchici, e analizzeremo opportuni spunti interessanti. Per questo tipo di struttura ad albero sono stati forniti in input come valori di supporto per i due livelli rispettivamente 0.01 e 0.001. Il valore di soglia minima positiva 0.12 mentre 0.1 come soglia massima negativa. Se il numero di regole estratte era particolarmente alto per rendere il risultato più preciso viene impostata la soglia positiva a 0.2. Si è prima analizzata la struttura in cui rimane fisso l'attributo relativo all'anno e varia categoria e sottocategoria. Possiamo apprezzare che grazie a questo algoritmo e questa struttura troviamo nuove informazioni che prima non si erano generate. La figura 5.52 ne è l'esempio. Si prende nota che nell'anno 2012 nonostante la categoria Allarme sociale fosse negativamente correlata a tale anno, la rispettiva sottocategoria Atti di vandalismo presentava una certa correlazione durante l'anno. Vengono anche estratte altre 2 relazioni flipping, in cui l'anno 2014 e 2015 sono positivamente correlati alla categoria Allarme sociale, mentre la sottocategoria Atti di vandalismo non trova una particolare relazione con questi anni.

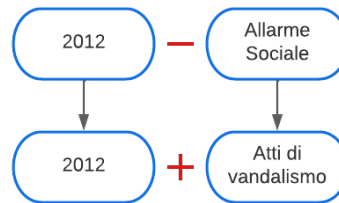


Figura 5.52: Correlazione Flipping Allarme Sociale e sottocategoria con anno

Molte regole di associazione flipping si trovano con lo stesso pattern ma prendendo in esame la categoria Convivenza Civile. Interessante come si trovino solo correlazioni con pattern positivo al primo livello di astrazione e negativo al secondo. Ci sono correlazioni positive tra la categoria e tutti gli anni, quindi di conseguenza delle correlazioni negative con alcune sottocategorie, tra questi disturbi da locali dall'anno 2016 al 2019, Rumori molesti con diversi anni ma non il 2015, molto utile questa informazione in quanto questo anno era correlato con tale sottocategoria se osserviamo le regole dell'altro metodo, quindi anche con questo algoritmo

implicitamente la cosa viene confermata. L'attributo Altro(C) non compare mai nelle regole ed inoltre troviamo correlazioni negative con sottocategorie che non comparivano nel precedente metodo, è questo conferma l'informazione estratta in precedenza. La figura 5.53 è l'esempio tra i tanti.

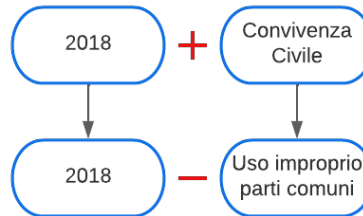


Figura 5.53: Correlazione Flipping Convivenza Civile e sottocategoria con anno

Per quanto riguarda la sottocategoria Qualità urbana si estraggono 6 flipping correlation, anche qui troviamo solo correlazioni positive al primo livello e negative al secondo. Appare in una regola la sottocategoria Altro(Q) e nelle altre Veicoli abbandonati. Si trovano correlati positivamente a questa categoria gli anni 2013, 2015, 2016, 2017, 2018, 2019. Rispetto al precedente metodo risulta interessante la non presenza degli anni 2012 e 2014 che invece risultavano correlati in un certo modo. Mentre il 2015 è un anno che compare in entrambi i metodi. Il 2013 invece è correlato qui positivamente. Mentre per la sottocategoria vengono confermate alcune deduzioni, gli anni che qui sono correlati negativamente non rientravano tra gli anni correlati con il metodo precedente.

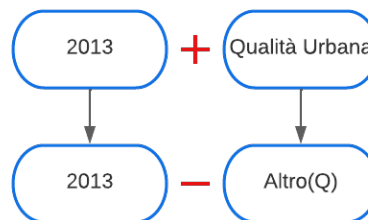


Figura 5.54: Correlazione Flipping Qualità Urbana e sottocategoria con anno

Se prendiamo in esame l'attributo temporale stagione si estraggono correlazioni flipping solo per la categoria Convivenza Civile. Non ci sono correlazioni positive nel secondo livello gerarchico. Troviamo nei risultati descritti da questo algoritmo correlazioni positive con le stagioni Inverno e Autunno. Nel precedente metodo

comparivano nelle regole correlazioni tra la categoria e la stagione Estate e Primavera. Si deduce che abbiamo estratto una nuova conoscenza che non eravamo riusciti a cogliere precedentemente. Per le sottocategorie presenti nelle regole precedenti si ottenevano correlazioni tra tutte le stagioni, qui invece alcune sottocategorie non presenti in precedenza sono correlate negativamente, confermandone l'affidabilità. Le sottocategorie negativamente correlate alla stagione Inverno sono: Uso improprio parti comuni, Comportamenti molesti e disturbi altri animali. Aggregazioni giovanili e comportamenti molesti con Autunno. Diventano interessanti le due regole presentate sotto.

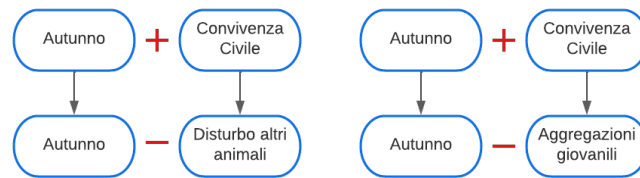
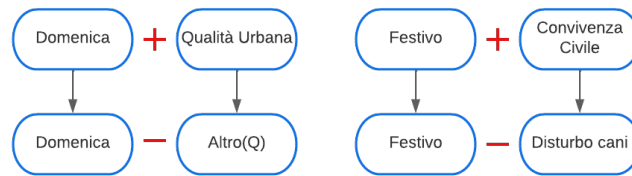
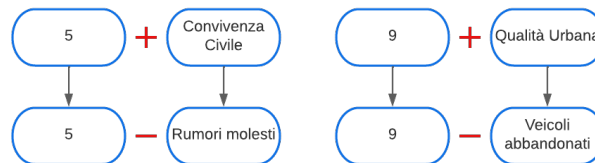


Figura 5.55: Correlazione Flipping Convivenza Civile e sottocategoria con Stagione

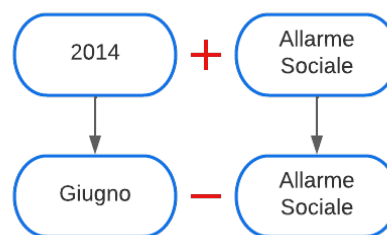
Per quanto riguarda l'attributo mese si trovano correlazioni flipping con tutti i mesi. Si ottengono quindi relazioni nuove rispetto al precedente metodo. Sono tutte correlate negativamente con le sottocategorie e positivamente con le categorie. Si trovano diverse correlazioni che confermano le precedenti informazioni estratte. In questo caso non si trova nessuna correlazione per la categoria e le sottocategorie di Allarme Sociale. Molto più interessanti le correlazioni di tipo flipping in cui si prendono in considerazione tra i due livelli i giorni settimanali. Non si trovano livelli correlati negativamente ma l'informazione immediata che salta fuori è che ci sono correlazioni positive generalmente con giorni come Sabato e Domenica, e di conseguenza correlazioni negative con determinati sottotipi. Se prima non avevamo ottenuto grandi informazioni dal punto di vista di giorni festivi qui invece possiamo osservare che invece esiste qualche correlazione con la categoria generale e questi giorni. La cosa viene anche confermata se prendiamo l'attributo specifico che differenzia i giorni feriali dai festivi, anche se si trovano correlazioni solo con la categoria Convivenza Civile.

Successivamente si è valutata la presenza di flipping correlation per l'attributo spaziale Circostrizione. Si sono trovate associazione positive al primo livello con tutte le categorie e correlazione negativa con le sottocategorie. Si nota la non presenza di Allarme Sociale.

Successivamente si è deciso di strutturare la categoria nei 2 livelli e si sono variati attributi spaziali e temporali. Spaziando nei 2 livelli anno e stagione si ottengono

**Figura 5.56:** Correlazioni Flipping**Figura 5.57:** Correlazioni Flipping Categoria e sottocategoria con Circonscrizione

correlazioni flipper solo con la categoria Allarme sociale che risulta correlata positivamente all'anno 2014 ma negativamente per tutte le quattro stagioni, fig. 5.58. Creando due livelli Stagione-Mese e Mese-Giorno settimanale non otteniamo correlazioni flipping. Dopo si è creato un livello gerarchico per i valori temporali Anno-Mese, trovando nuovamente correlazioni positive nell'anno 2014 con la categoria Allarme Sociale, ma questa risulta correlata negativamente a tutti i mesi dell'anno. Questo ci porta a pensare che probabilmente si evidenziano particolari correlazioni in particolare con l'anno 2014 per questa categoria nonostante sia poco rappresentata nel dataset.

**Figura 5.58:** Correlazioni Flipping Categoria con Anno e mese

Prossimo passo è stato quello di strutturare il dataset tenendo fissa la parte relativa alla sottocategoria e variando gli attributi temporali su due livelli gerarchici.

Utilizzando gli attributi Anno e Mese, si sono trovate correlazioni positive per il primo livello e negative con il mese, secondo livello. Sono tutte correlazioni in cui sono presenti due sottocategorie relative alla categoria Qualità Urbana: Aggregazioni giovanili e Veicoli abbandonati. Per la prima sottocategoria non era mai stata trovata alcuna correlazione con attributi temporali, quindi otteniamo una particolare nuova informazione. Troviamo correlazioni positive per questi attributi durante gli anni che vanno dal 2012 al 2014. Mentre per il primo attributo questa informazione estratta è nuova, invece per l'attributo Veicoli abbandonati era già stata trovata una correlazione proprio con i corrispondenti anni, attraverso questo algoritmo riusciamo a ben identificare determinati mesi che sono correlati negativamente durante un certo anno. Informazioni simili sono ottenute se si sceglie come primo livello l'attributo anno e l'attributo stagione nel secondo. Si ottengono 3 regole in cui compaiono solo correlazioni positive tra gli anni 2012 e 2014 e segnalazioni di Veicoli abbandonati, mentre si trovano correlazioni negative durante le stagioni Inverno e Autunno. La figura 5.59 mostra queste correlazioni flipping.

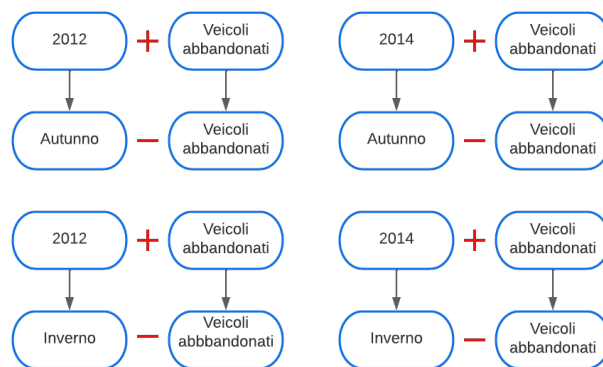


Figura 5.59: Correlazioni Flipping Sottocategoria con Anno e stagione

Se passiamo alla formazione di una gerarchia tra stagione e mese per trovare correlazioni flipper con la sottocategoria otteniamo diverse regole interessanti, in cui si trovano correlazioni positive tra diversi mesi ed alcune sottocategorie di tipo Convivenza Civile. Riusciamo per esempio a vedere che esiste una correlazione positiva tra segnalazioni di Comportamenti molesti e il periodo estivo ma si nota una correlazione negativa durante il mese di Luglio. Molto particolare la regola che associa una correlazione positiva tra i disturbi da locali e il periodo estivo ma contrariamente è correlata negativamente con i mesi di Settembre. Questo potrebbe coincidere con il fatto che molte segnalazioni legate a disturbi provenienti da locali si potrebbero verificare durante l'Estate ma nella prime due settimane estive di Settembre si potrebbe assistere ad una minor frequenza. Inoltre compaiono anche correlazioni con sottocategorie a cui precedentemente non era stato possibile trovare

una correlazione. Troviamo Uso improprio parti comuni associato positivamente alla stagione Primavera ma negativamente al mese di Maggio. Invece Altro(Q) risulta correlato positivamente nel primo livello con l'attributo Autunno e negativamente nel secondo livello con il mese di Settembre. Non emergono correlazioni per questa struttura dati con le sottocategorie di Allarme Sociale. Nella figura 5.60 possiamo apprezzare alcune di queste relazioni estratte.

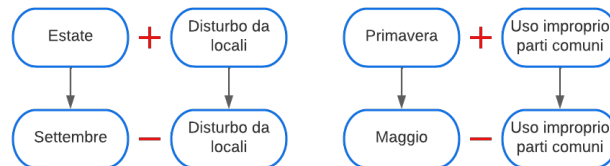


Figura 5.60: Correlazioni Flipping Sottocategoria con Stagione e Mese

Se ci spostiamo nel livello temporale Mese-Giorno settimanale riusciamo ad identificare una regola parecchio interessante. Tale associazione flipping presenta una correlazione negativa al primo livello con un determinato mese, di conseguenza positiva con un determinato giorno. Nel dettaglio si trova una associazione negativa tra il mese di Agosto e la sottocategoria Veicoli abbandonati ma scendendo di livello vi è una correlazione positiva tra sabato e questa tipologia di sottocategoria. Questa particolare associazione contrastante tra livelli potrebbe essere utile per individuare in un dato mese il giorno in cui potrebbero verificarsi determinati eventi. Altre regole sono relative anche alla categoria Decoro e degrado urbano, come mostrato in figura 5.61.

Purtroppo per quanto riguarda associazioni tra sottocategorie e il contesto spaziale non si riescono ad estrarre valide regole contrastanti.

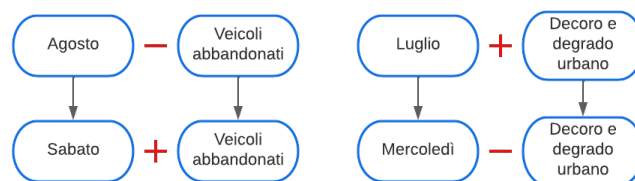


Figura 5.61: Correlazioni Flipping, Sottocategoria con Mese e Giorno

Successivamente si è deciso di allargare la struttura ad albero su 3 livelli. Variando sulla categoria e sottocategoria non è stato possibile ottenere nulla. Dopo si è cercato di variare l'attributo temporale, impostando la struttura Anno-Stagione-Mese e mantenendo fisso il livello sulla categoria e sottocategoria.

Per la categoria non siamo riusciti ad estrarre alcuna informazione, contrariamente per la sottocategoria si sono generate 3 associazioni che vengono mostrate nella figura 5.62. Si è portata su due livelli la categoria e nell'ultimo livello è stata riportata la sottocategoria di riferimento. Le associazioni estratte sono tutte relative alla convivenza civile, il pattern ottenuto è $+ - +$. Per quanto riguarda queste regole per esempio si può dedurre che l'anno 2014 è positivamente correlato alla categoria Convivenza Civile, si scopre una correlazione negativa durante la stagione Primavera con il sottotipo disturbo da locali, ma nel mese di Giugno relativo alla stagione primaverile vi è una correlazione positiva con questo tipo di segnalazioni. Potrebbe essere interpretato come indice di un aumento di segnalazioni di questo tipo con l'avvicinarsi del periodo estivo.

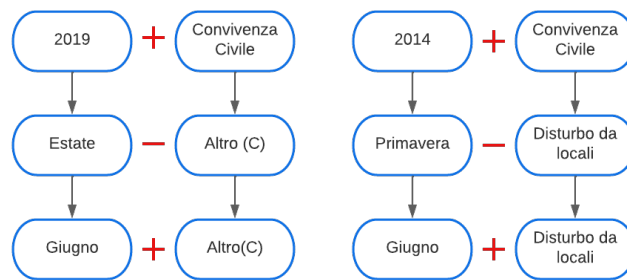


Figura 5.62: Correlazioni Flipping su 3 livelli

5.7 Flipping correlation sulle segnalazioni di New York

La stessa procedura è stata applicata al dataset sulle segnalazioni urbane della città di New York. In primo luogo si sono strutturati i dataset su due livelli di gerarchia. Prima è stata mantenuta fissa la parte relativa al contesto temporale o spaziale. Prendendo in esame l'anno si sono confermati alcuni aspetti estratti precedentemente. In linea generale diversi anni sono correlati positivamente a delle categorie, ma per molte sottocategorie poi non viene trovata una relazione valida. Le categorie che presentano queste associazioni positive sono relative alle categorie Complaint e Maintenance. Si è osservata una regola che presenta un pattern $(-)(+)$, in cui si evince che la categoria Information è correlata negativamente all'anno 2020, ma la sua sottocategoria relativa alle segnalazioni di tipo License ha una correlazione positiva con questo anno. La figura 5.63 mostra quanto appena descritto.

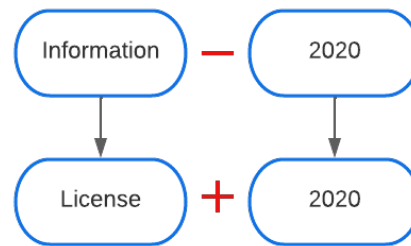


Figura 5.63: Correlazioni Flipping Categoria e sottocategoria con Anno

Proseguendo con questa struttura a 2 livelli, proviamo ad individuare alcune correlazioni di tipo flipping con la stagione. Si individuano 6 regole flipping, in cui troviamo solamente correlazioni positive al primo livello e negative al secondo. Si generano 4 regole flipping per la categoria Maintenance e 2 per la categoria Complaint. La prima si trova correlata positivamente alle quattro differenti stagioni, da questo punto di vista non otteniamo nessuna novità rilevante, ma scendendo di livello vediamo che per esempio la sottocategoria Heat and Hot water è correlata negativamente alla stagione Primavera ed Estate. Da questa informazione rafforziamo le informazioni delle regole associative precedenti in cui si evinceva che tale sottocategoria raggiunge correlazioni forti durante i mesi freddi dell'anno. Mentre troviamo correlazioni negative anche per la sottocategoria Vacant Property e le stagioni autunnali e invernali. Questa sottocategoria non era presente tra le precedenti regole. Mentre per quanto riguarda la categoria Complaint è correlata positivamente per entrambe le regole alla stagione Autunno ma negativamente alle sottocategorie Drinking e Nuisance Complaint.

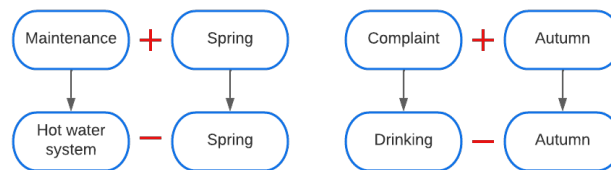


Figura 5.64: Correlazioni Flipping, Categoria e sottocategoria con Stagione

Per quanto riguarda invece le correlazioni flipping tra Categoria e Sottocategoria da una parte e il mese dall'altra, si individuano correlazioni positive al primo livello e negative al secondo. Le categorie e sottocategorie presenti in questo tipo di correlazione sono Complaint e Maintenance. Le informazioni che possiamo dedurre sono legate a delle sottocategorie poco presenti e per cui non erano state rilevate alcun tipo di correlazioni interessanti quando si sono generate le regole associative. Tra queste vi sono Elevator, Drinking, Smoking, Illegal posting. A livello informativo queste associazioni flipping ci permettono di individuare per alcune sottocategorie non molto rilevanti i mesi esatti in cui vi è una particolare correlazione negativa. Nei grafici in figura 5.65 troviamo due possibili associazioni.

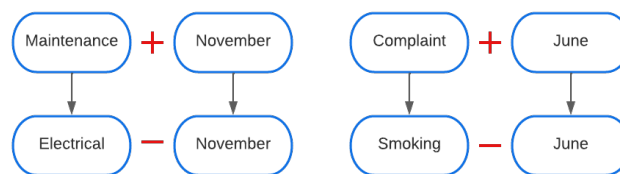


Figura 5.65: Correlazioni Flipping, son Categoria, sottocategoria e Mese

Stessa cosa si può dire se si prende in considerazione il giorno settimanale. Anche qui troviamo correlazioni flipping solo per le categorie Complaint e Maintenance. Troviamo anche qui solo correlazioni positive per il primo livello e negative per il secondo. L'informazione più interessante è data da alcuni giorni che si confermano avere una certa correlazione con queste categorie. Per le sottocategorie non riusciamo ad evidenziare particolarità in quanto già precedentemente non presentavano alcuna correlazione.

Successivamente sono state prese in considerazione i contesti geografici della città di New York. La struttura scelta è ancora di tipo Categoria-Sottocategoria relazionata all'attributo Borough relativo ai distretti. Alcune correlazioni contrastanti che sono state estratte da questo algoritmo evidenziano alcune deduzioni interessanti. In linea generale alcune regole confermano la correlazione positiva di alcune categorie con dei distretti. È il caso di categorie come Complaint, Maintenance e Cleaning & Garbage. Inoltre otteniamo informazioni su sottotipi di cui prima non sapevamo nulla che sono correlate negativamente in certi distretti. Una delle correlazioni flipping fornisce qualche novità a livello di informazione in quanto è stata trovata una conoscenza non identificata prima, questa è legata alla relazione positiva tra il distretto di Manhattan e la categoria Poverty. Inoltre tra queste viene trovata una sola correlazione negativa al primo livello gerarchico e positiva al secondo, questa riguarda ancora il distretto di Manhattan correlato negativamente con la

categoria Transportation, contrariamente è correlata positivamente alla sottocategoria Derelict Bicycle. Molto interessante sapere quindi che si potrebbero verificare nel distretto di Manhattan segnalazioni di bici abbandonate. Nella figura 5.66 troviamo rappresentato quanto detto.

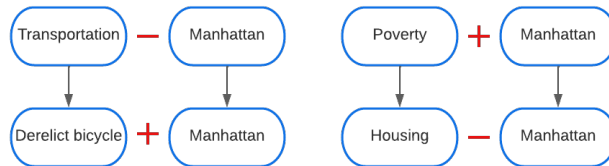


Figura 5.66: Correlazioni Flipping, Categoria e sottocategoria con Borough

Se invece si ispezionano eventuali correlazioni flipping per l'attributo Community Board si riescono ad individuare precise zone all'interno dei distretti che sono correlate positivamente con le categorie Complaint o Maintenance. Dai risultati riusciamo ad individuare alcune community board dei distretti di Manhattan, Brooklyn, Bronx e Queens che presentano correlazioni flipping. Nell'immagine sotto alcune informazioni.

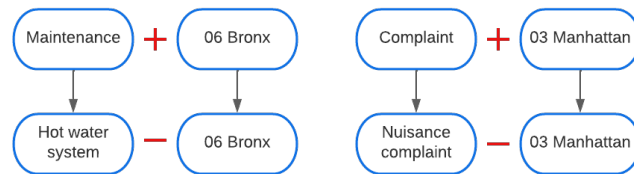


Figura 5.67: Correlazioni Flipping, Categoria e sottocategoria con Community Board

Molto interessanti le associazioni flipping strutturati con l'attributo Location type che indica il luogo in cui è avvenuta la segnalazione. I risultati più interessanti mostrano diversi tipi di luogo e diverse categorie per cui non era stata trovata alcuna associazione nell'analisi precedente. Tra queste l'associazione tra la categoria Information correlata negativamente a luoghi commerciali e la sottocategoria License positivamente correlata con tale luogo. Questo ci porta a concludere che nei negozi si verificano segnalazioni legate a problematiche su eventuali licenze. Altra informazione è legata a segnalazioni sulla categoria animali che avvengono principalmente su strade ma negativa invece è la correlazione della sottocategoria Animal control care con questo luogo. Inoltre riusciamo ad associare i luoghi commerciali e luoghi di ristorazione e bar positivamente alla categoria Complaint,

con rispettive sottocategorie Mold e Nuisance Complaint negativamente correlate. Si raffigurano queste correlazioni flipping nel grafico 5.68.

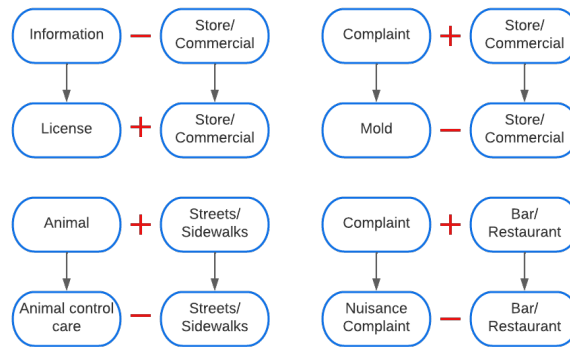


Figura 5.68: Correlazioni Flipping, Categoria e sottocategoria con Luogo

Dopo si è passati ad analizzare sempre una struttura su due livelli ma mantenendo fissa la parte sulla categoria e sottocategoria e variando nei due livelli gli attributi spaziali e temporali. Diverse correlazioni sono state trovate con le categorie usando nei due livelli gerarchici la stagione e il mese. Molte associazioni confermavano le informazioni già in nostro possesso dalla precedente analisi, diversamente per quelle categorie da cui non si riusciva ad ottenere alcuna conoscenza. Sono state trovate solo correlazioni positive nel primo livello e negative nel secondo, tra le associazioni positive troviamo categorie Poverty e il mese primaverile o Public and Mobility Services con la categoria Estate.

Mentre per quanto riguarda l'associazione tra la categoria e mese e giorno della settimana si trovano correlazioni ridondanti, troviamo infatti correlazioni di categorie come Complaint, Maintenance e Streets & Sidewalks positivamente con alcuni mesi mentre negativamente con dei giorni settimanali specifici. Nella figura 5.69 troviamo alcune correlazioni tra la categoria e il contesto temporale stagione e mese.

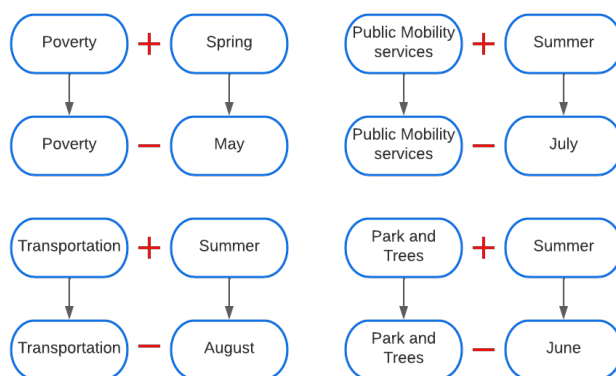


Figura 5.69: Correlazioni Flipping, Categoria con Stagione e Mese

Dopo sono stati creati livelli gerarchici tra l'attributo Borough e Community Board e si è cercato di estrarre relazioni flipping con la categoria. Sono state trovate 2 correlazioni che presentano correlazione negativa al primo livello e correlazione positiva al secondo livello. Queste riguardano il distretto di Queens e la categoria Public Mobility Services, mentre le community board n.83 e 80 di tale distretto mostrano una correlazione con questo tipo di segnalazione. Altre associazioni riguardano categorie come Streets & Sidewalks, Maintenance, Cleaning & Garbage, Complaint, Transportation e Parks & Trees che confermano la correlazione esistente con alcuni distretti come precedentemente già ricavato.

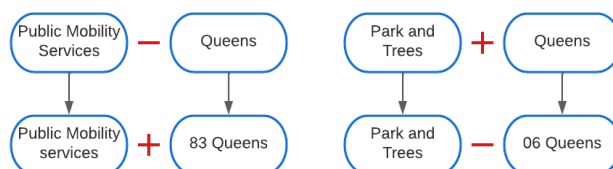


Figura 5.70: Correlazioni Flipping, Categoria con Borough e Community Board

Poi si è deciso di legare gerarchicamente l'attributo Location Type a Borough. Alcune associazioni flipping trovate riguardano il distretto di Staten Island, questi risultanti acquistano molta valenza perchè con questo distretto non erano state trovate molte correlazioni. Infatti tale distretto è correlato negativamente al primo livello di astrazione che riguarda la categoria di appartenenza della segnalazione, mentre positivamente con un determinato luogo. Troviamo il distretto di Staten Island correlato negativamente alla categoria Complaint in due flipping correlation, mentre Bar/Restaurant e Store/Commercial positivamente correlati alla categoria Complaint. Questo ci porta a dedurre che nonostante non si trovi una certa

associazione per le segnalazioni di categoria Complaint nel distretto di Staten Island si potrebbero comunque verificare segnalazioni di questo tipo in luoghi come bar, ristoranti e negozi. Altra correlazione è con la categoria Maintenance che risulta correlata negativamente al distretto di Staten Island ma positivamente con luoghi etichettati come Unspecified. Altra informazione rilevante che tra queste associazioni troviamo nuovamente una correlazione positiva tra il distretto di Manhattan e la categoria Poverty. Troviamo nella figura 5.71 queste correlazioni flipping.

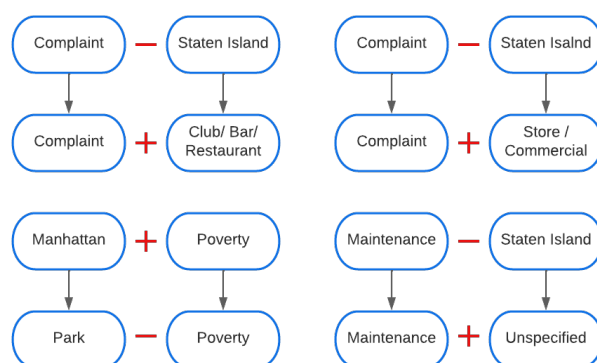


Figura 5.71: Correlazioni Flipping, Categoria con Borough e Location Type

Successivamente è stata analizzata la struttura gerarchica che associa la stagione e il mese da una parte e la sottocategoria dall'altra. Interessanti questi risultati in quanto permettono di apprezzare correlazioni con sottocategorie per cui non erano state estratte valide informazioni. Una tra le più rilevanti è quella che mostra una correlazione positiva tra la sottocategoria Snow/Ice removal e la stagione Inverno, mentre i mesi della stagione invernale relativi a Marzo mostrano una correlazione negativa con questa tipologia di segnalazione. Questa associazione mostra come si verificano segnalazioni per problematiche legate alla neve durante l'Inverno ma nelle prime settimane di Marzo la situazione urbana legata a questa problematica non risulterebbe così allarmante. Cosa simile possiamo dire per le segnalazioni legate a Tree Condition che mostrano una correlazione positiva con l'Estate ma risultano negativamente correlate con il mese di Luglio. Troviamo anche la sottocategoria Nuisance Complaint correlata positivamente alla stagione Primavera ma negativamente al mese di Marzo. Si individuano anche segnalazioni di animali morti con una certa correlazione positiva durante il periodo estivo ma con correlazione negativa nel mese di Giugno. Altre correlazioni flipping confermano correlazioni già precedentemente ottenute con alcune sottocategorie, è l'esempio la sottocategoria Street Noise che viene associata positivamente alla

stagione Primavera ma negativamente al mese di Marzo.

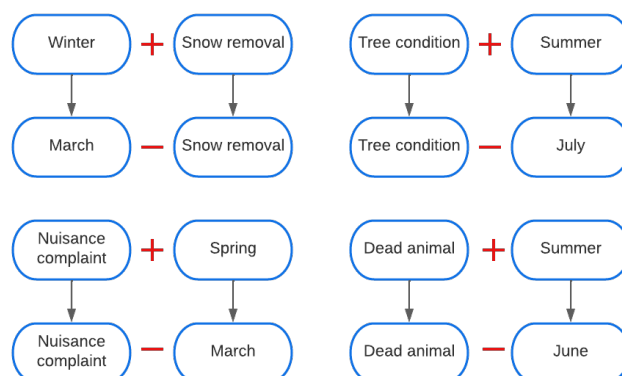


Figura 5.72: Correlazioni Flipping, Sottocategoria con stagione e mese

Quando invece passiamo ad analizzare i risultati prendendo in esame gli attributi legati al mese e al giorno settimanale si ottengono associazioni flipping in cui si confermano alcune informazioni già in nostro possesso. Troviamo infatti correlazioni positive tra sottocategorie come Heat and Hot Water system con i mesi di Gennaio, Febbraio, Novembre e Dicembre, mentre si associano determinati giorni ad essi correlati negativamente. Stessa cosa per quanto riguarda la sottocategoria Street Noise che è positivamente correlata al mese di Giugno ma negativamente ad alcuni giorni del mese. Le novità provengono dalle correlazioni flipping positive tra i primi livelli gerarchici tra i mesi di Gennaio e Febbraio e la sottocategoria Snow/Ice removal, che invece è negativamente correlata ad alcuni giorni della settimana. Se prima era già stata individuata una correlazione positiva tra questa segnalazione urbana e il mese inverno e negativa nel mese di Marzo, con questa struttura abbiamo confermato la cosa ed inoltre abbiamo aggiunto al nostro bagaglio informazioni che effettivamente esiste una correlazione tra queste richieste e determinati mesi invernali, Gennaio e Febbraio in questo caso. Analizzando le correlazioni flipping tra le sottocategorie e gli attributi geografici Borough e Community Board otteniamo associazioni che confermano la correlazione tra alcune sottocategorie e determinati distretti, anche se con questo tipo di correlazione riusciamo ad identificare in quali zone del distretto non esiste alcuna correlazione. In aggiunta si ottengono nuove informazioni su sottotipi di segnalazione che potrebbero verificarsi in certi distretti ed inoltre abbiamo l'informazione su quali Community Board non vi è correlazione. Nuove informazioni sono date dalla sottocategoria Graffiti che ha una certa correlazione positiva con il distretto di Brooklyn anche se non in tutte le aree. In questo distretto si trovano anche correlazioni con segnalazioni su Illegal Posting. Questa segnalazione ha una certa correlazione anche con il distretto di Queens. Dai

risultati osserviamo che in tale distretto vi sono correlazioni anche con segnalazioni di veicoli abbandonati, ed inoltre si trovano due correlazioni negativamente correlate con il primo livello di astrazione e positivamente con il secondo, sono informazioni che avevamo già precedentemente ottenuto da altre regole flipping per la categoria di riferimento, e che adesso puntualizzano il tipo di problema riscontrato. Queste riguardano problematiche legate a reclami sui taxi in determinate community board del distretto di Queens, infatti troviamo che questa sottocategoria è negativamente correlata con tale distretto ma esiste una certa correlazione con alcune zone al suo interno, 80 Queens e 83 Queens. Problematiche sui taxi emergono anche nel distretto di Manhattan, come anche segnalazioni sulle auto a noleggio e rumori causati da elicotteri o aerei e segnalazioni urbane legate ai senzatetto, ovviamente si identificano i rispettivi community board in cui non esiste tale correlazione.

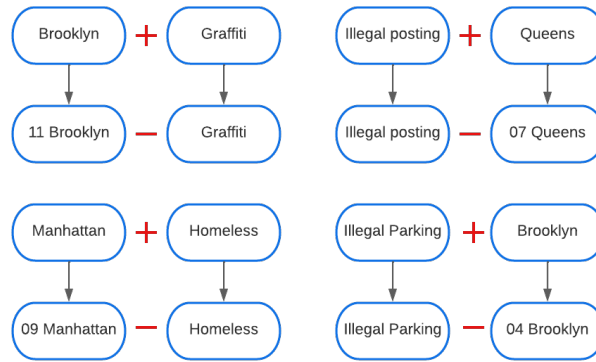


Figura 5.73: Correlazioni Flipping, Sottocategoria con Borough e Community Board

Sono state analizzate anche le correlazioni flipping tra la sottocategoria e i distretti e la tipologia di luogo. Da queste emergono correlazioni positive al secondo livello di astrazione tra alcuni luoghi e alcune sottocategorie. Troviamo reclami da parte dei consumatori nelle attività commerciali del distretto di Queens, nonostante questo distretto non abbia una correlazione positiva con questa tipologia di segnalazioni. Riusciamo ad identificare luoghi in cui si verificano segnalazioni su problematiche legate ai roditori in edifici vuoti nell'area di Queens. Inoltre otteniamo nuove informazioni sul distretto di Staten Island che registra problematiche di rumori e condizioni poco sanitarie nei bar o ristoranti. Altre associazioni flipping con correlazioni positive al primo livello di astrazione denotano una certa relazione tra il distretto di Brooklyn e segnalazioni su graffiti e il distretto di Queens su problematiche di fognatura e acqua piovana. Troviamo nel grafico sotto alcune di queste correlazioni.

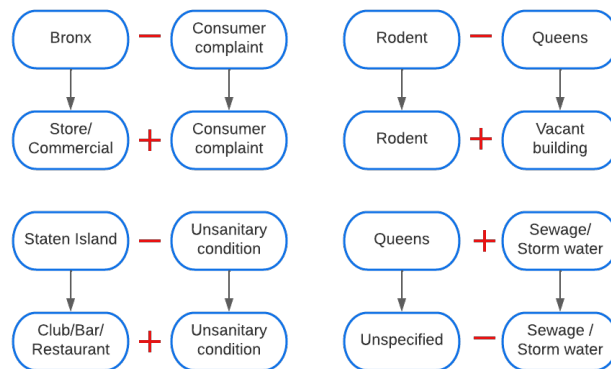


Figura 5.74: Correlazioni Flipping, Sottocategoria con Borough e Location Type

Le strutture su 3 livelli gerarchici generano correlazioni flipping solo con i contesti spaziali. I 3 livelli sono composti da Borough, Community Board e Location Type, e si generano correlazioni con le categorie o sottocategorie. Si trovano solo correlazioni flipping + - +. Per la categoria troviamo correlazioni che confermano relazioni tra alcuni distretti e alcuni particolari luoghi. La cosa interessante è che scendendo di supporto si riescono a trovare alcune correlazioni peculiari tra categorie e luoghi che non erano state riscontrate nella precedente analisi. Per esempio la categoria Cleaning & Garbage si trova correlata positivamente al distretto di Queens, poi negativamente in una zona all'interno del distretto ma trova correlazione positiva in luoghi di ristorazione. Da ciò deduciamo segnalazioni di poca pulizia in dei locali all'interno del distretto di Queens. Altra correlazione flipping conferma la relazione emersa anche in altre regole tra la categoria Poverty e il distretto di Manhattan, qui diventa interessante il fatto che nel terzo livello di astrazione

troviamo che questa categoria è correlata positivamente a luoghi di culto, come le chiese. Altra informazione interessante è estratta dalla regola che mostra una correlazione positiva tra il distretto di Bronx e problematiche di manutenzione che evidenziano poi nell'ultimo livello gerarchico segnalazioni di questa tipologia nelle scuole.

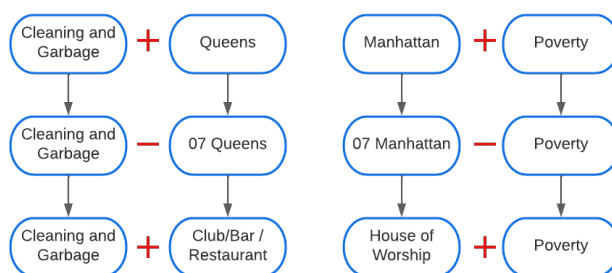


Figura 5.75: Correlazioni Flipping, Categoria con Borough, Community Board e Location Type

Se ci spostiamo nelle sottocategorie troviamo molte correlazioni particolari, tra queste per esempio la correlazione tra il distretto di Brooklyn con la sottocategoria Rodent al primo livello di astrazione, mentre troviamo correlazioni negative con alcune Community Board, successivamente al terzo livello emerge la correlazione positiva tra questa segnalazione e luoghi identificati come edifici vuoti. Quanto emerge è molto plausibile e permette di avere una informazione nuova. Da questi risultati riusciamo anche ad identificare i luoghi in cui si verificano graffiti nel distretto di Brooklyn, da queste regole emerge che sono da monitorare per questo tipo di segnalazioni i luoghi come negozi. Riusciamo ad ottenere conferma da alcune segnalazioni che si verificano in alcuni distretti ed in particolari luoghi, è il caso di segnalazioni sui rumori stradali che da queste associazioni si deduce potrebbero verificarsi in luoghi come bar, ristoranti e club. Se dalle correlazioni flipping di prima avevamo dedotto che ci sono correlazioni tra la categoria Poverty e il distretto di Manhattan, qui riusciamo ad identificare anche quale sottocategoria mostra tale correlazione in questo determinato distretto, infatti si evincono segnalazioni di senza tetto e principalmente in luoghi come chiese. Nelle vicinanze di chiese si potrebbero verificare anche segnalazioni di rumori lungo le strade di Brooklyn e Manhattan. Emergono anche segnalazioni su riparazione di edifici nelle scuole nel distretto di Manhattan. Mentre nel distretto di Brooklyn si potrebbero verificare segnalazioni di richieste o informazioni sulle scuole. Queste correlazioni flipping su 3 livelli risultano particolarmente interessanti per entrare nel dettaglio anche di correlazioni trovate in precedenza, permettendo di delineare qualche informazione interessante. I grafici in figura 5.76 riportano alcuni esempi.

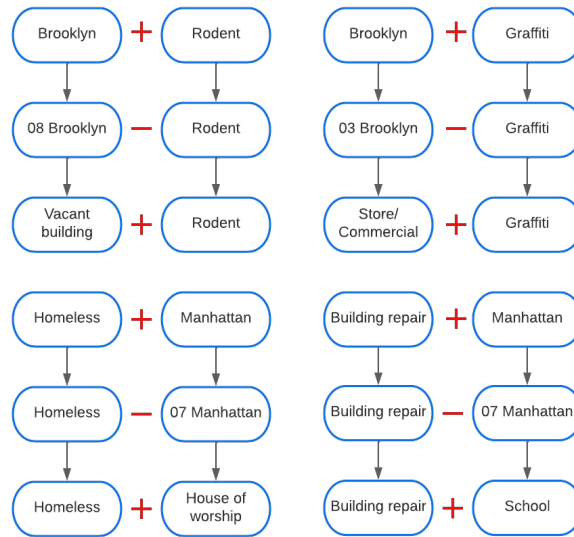


Figura 5.76: Correlazioni Flipping, Sottocategoria con Borough, Community Board e Location Type

I risultati ottenuti da questo algoritmo permettono di individuare diverse caratteristiche interessanti per tipologie di segnalazioni urbane poco frequenti ma peculiari. Questa tipologia di correlazioni flipping permette di individuare correlazioni specifiche per un dato sottogruppo in un determinato livello di astrazione. Infatti grazie alla struttura di questo algoritmo si possono identificare delle sottoaree di una città e di analizzare le condizioni temporali in cui determinate tipologie di segnalazioni hanno una certa rilevanza.

Capitolo 6

Conclusioni e sviluppi futuri

In questo elaborato si è descritto il lavoro di tesi relativo all'analisi dei dati in ambito urbano utilizzando due diverse tecniche di data mining legate al contesto delle regole di associazione. Come si è visto la metodologia presentata è stata validata su due casi d'uso reali appartenenti a due contesti urbani differenti e con tipologie di segnalazioni in parte diverse tra di loro. L'obiettivo principale è stato quello di analizzare i dati relativi a segnalazioni di non emergenza legati al contesto urbano con lo scopo di trasformarli in informazione dopo aver opportunamente strutturato i dataset per ottenere risultati il più possibile accurati. I diversi metodi utilizzati hanno permesso di estrapolare conoscenza nascosta all'interno di particolari dataset urbani e di valutarne l'utilità. Si sono confrontate le due tecniche e dopo aver ottenuto i risultati, questi sono stati visualizzati con degli opportuni pattern di facile interpretazione. È stato possibile delineare delle zone cittadine e degli archi temporali in cui determinate richieste di segnalazioni si verificano con una certa frequenza. Le analisi effettuate e i risultati presentati ci permettono di abbozzare delle linee guida per i governatori locali che sono interessati a integrare un sistema di previsione del livello di sicurezza nella loro città. Possono rappresentare una prima base di conoscenza per costruire un futuro modello di previsione in modo da migliorare la qualità e la percezione di sicurezza in città. Informazioni di questo genere potrebbero essere utilizzate per provare ad identificare delle situazioni allarmanti di sicurezza e disturbo in determinate aree della città, ma anche per osservare come determinate problematiche sono cambiate nel tempo. Di fatto i risultati ottenuti in questa analisi dei dati urbani non solo possono fornire informazioni utili da utilizzare nel contesto di una smart city per migliorare la qualità di vita dei cittadini e la percezione di sicurezza in città, ma anche contribuire a rendere il cittadino sempre più partecipe del contesto urbano in cui vive. Come lavoro futuro, si potrebbe estendere lo studio a diverse città e paesi e arricchire il quadro attuale con ulteriori metadati contestuali, come per esempio informazioni socio-demografiche. Si potrebbe anche pensare alla creazione di uno

standard di riferimento mondiale o nazionale per questa tipologia di segnalazioni urbane in modo da poter sviluppare sistemi interoperabili e strumenti per creare applicazioni e piattaforme innovative per la comunità.

Bibliografia

- [1] AperTO. *Riepilogo segnalazioni Contact Center Polizia Municipale*. URL: <https://aperto.comune.torino.it/dataset>.
- [2] M. Barsky, K. Sangkyum, T. Weninger e J. Han. *Mining Flipping Correlations from Large Datasets with Taxonomies*. 2011. arXiv: 1201.0233 [cs.DB].
- [3] L. Cagliero, T. Cerquitelli, S. Chiusano, P. Garino, M. Nardone, B. Pralio e L. Venturini. *Monitoring the citizens' perception on urban security in Smart City environments*. Dipartimento di Automatica e Informatica Politecnico di Torino.
- [4] Datetime. URL: <https://docs.python.org/3/library/datetime.html>.
- [5] FLIPPER. URL: <https://github.com/mgbarsky/flipper>.
- [6] Mahdi Hashemi. «Studying and Clustering Cities Based on Their Non-Emergency Service Requests». In: *Information* 12.8 (2021). ISSN: 2078-2489. URL: <https://www.mdpi.com/2078-2489/12/8/332>.
- [7] Itertools. URL: <https://docs.python.org/3/library/itertools.html>.
- [8] Matplotlib. URL: <https://matplotlib.org/>.
- [9] MLxtend. URL: <https://rasbt.github.io/mlxtend/>.
- [10] NumPy. URL: <https://numpy.org/>.
- [11] NYCOpenData. *311 Service Requests from 2010 to Present*. URL: <https://opendata.cityofnewyork.us/>.
- [12] Grace Paet. *An analysis of New York City's 311 Service, 2010 to 2014*. URL: <https://public.tableau.com/app/profile/grace.paet/v%20iz/shared/JPCSBP2G7>.
- [13] Pandas. URL: <https://pandas.pydata.org/>.
- [14] *Public Holidays USA*. URL: <https://www.public-holidays.us/>.
- [15] Seaborn. URL: <https://seaborn.pydata.org/>.
- [16] P. Tan, M. Steinbach, A. Karpatne e V. Kumar. «Association Analysis: Basic Concepts and Algorithms». In: *Introduction to Data Mining*.

- [17] L. Wang, C. Qian, P. Kats, C. Kontokosta e S. Sobolevsky. «Structure of 311 service requests as a signature of urban location». In: *PloS one* 12(10):e0186314 (2017). URL: <https://doi.org/10.1371/journal.pone.0186314>.