POLITECNICO DI TORINO

Master's Degree in Engineering And Management

Master's Degree Thesis

EVALUATION OF DATA QUALITY TOOLS

Supervisor:
Prof. Torchiano Marco

Candidate:
Mba chizaramekpere .c

Academic Year 2020/2021

Abstract

Data plays an important role in our day-to-day activities and its importance cannot be over emphasized. For organizations, it has become a valuable asset that drives strategy and informed decision making but the benefits of data are compromised if the data is of bad quality. With the large amount of data generated daily, it is common within datasets to find anomalies such as inconsistency, outdated values, missing values, duplicate values, wrong formats or representations of data, etc. Such anomalies negatively impact the quality of data significantly and degrade the quality of information, insights or decisions derived from such datasets. For this reason, it is necessary to assess and ensure the quality of data is reliable and up to standards. The goal of this thesis is to provide an understanding of data quality, the dimensions of data quality and the standards in place to assess and measure data quality. Exploration of available tools which can be used to assess and improve data quality, and finally a comparative analysis is then carried out on these tools to understand their capabilities.

Table of Contents

List of figures

List of tables

Preface

Providing high quality data has become of great importance to both the government and business organizations. Information has always played an important role to making decisions. From simple daily decisions such as carrying an umbrella, to more critical decisions made by managers and the government, having well rounded accurate information can improve the quality decisions made. In the same way, making decisions based on bad quality data could have negative effects and consequences. For this reason, attention is being drawn to the importance of data quality management which ensures the quality of data is assessed, improved and maintained. But what exactly is bad data quality? What causes bad data quality? How can data quality be qualified or quantified? The first section (chapter 1 and chapter 2) of this thesis aims to answer these questions. In the first chapter, the concepts of data quality, the benefits of good data quality and implications of bad data quality are described. The second chapter explains the standards of data quality according to the ISO/ IEC 25000. These standards are important because the concept of data quality can be relative. Experts have slightly different views on what should be considered good quality data and several methods have been proposed for detecting and measuring data quality problems. The ISO standard presents 15 measurable categories or dimensions for evaluating data quality. These dimensions are discussed extensively in chapter 2.

Data quality tools are emerging in the market, to automate the remediation of data quality issues. These tools support the identification and improvement of data quality issues with features such as data profiling, management of metadata, record matching, monitoring, privacy and security, etc. But what can be expected from such tools? How efficiently and effectively are they able to assess and improve data quality? How do they compare with each other? how well do the tools cover the quality characteristics of the ISO 25012 and the measures of ISO 25024? The second section of this thesis (chapter 3-4) aims to answer these questions. A comparison matrix which compares the functionalities of 21 tools are presented. The matrix is normalized to give allocate a superficially grade or rating to each of the tools based on the functionalities of the tools on paper. We then associated the features and functionalities of the tools to the dimensions presented in the ISO 25012 standard, this served as a basis for the testing of some of the tools which is done in chapter 4. In order to observe how the features on paper apply in real life instances, OpenRefine and Trifacta, were tested using 3 datasets. The analysis and results of these tools are documented in chapter 4. Chapter 5 provides a detailed conclusion.

Chapter 1

1. Introduction to Data Quality

1.1 Data quality definition

Data quality has been defined in several ways in literature but in general terms, it is referred to as "fitness for use," which means the ability of data to meet the user's requirement. This definition implies that the concept of data quality is relative because data with quality considered appropriate, for one use may be considered insufficient quality for another purpose. For example, a processed dataset of sales may be of high quality for predicting future sales, despite not representing the sizes of sold items and therefore not fit for the purpose of predicting the number of items to be kept in inventory based on their sizes. Likewise, a good quality dataset of customer information will be quite irrelevant to predicting the weather. The degree by which data meets the expectations of data consumers, based on their intended uses of the data, is represented by the level of data quality [1]. Data quality is therefore directly related to the perceived or established purpose of the data and can also be defined as a measure of the reliability and application efficiency of data.

This shows that in order to fully characterize the quality of data, it is important to consider multiple dimensions such as accuracy, consistency, completeness, timeliness, credibility, accessibility, compliance, confidentiality, efficiency, precision, traceability, understandability, availability, portability and recoverability. These dimensions measure data quality from different angles and will be discussed in detail in subsequent chapters. For the purpose of this thesis, Data quality is studied as independent of a particular purpose, because it is assessed with respect to all possible purposes.

The relevance of data quality in organizations has increased over the years, as data processing has become more strongly associated with business operations, and data analytics increasingly used by organizations to help drive business decisions. With the emergence of big data, data quality management has become more important than ever, especially to organizations who seek to attain business value through data.

Data quality problems can lead to liability consequences, even errors considered as minor, can result in lost revenue, process or business inefficiencies, missed opportunities and risks of paying huge fines due to failures to comply with industry and government regulations. With adequate knowledge about data, data quality can be improved right from the beginning of the production process [2]. Moreover, rather than simply changing data values manually, a variety of data quality tools are currently available. These tools can be adopted to automate some processes, in a bid to increase quality.

1.1.1 Potential benefits of good data quality to organizations

1. More Informed Decision-Making: The idea of using data is often to analyse patterns and facts and use the insights to make decisions, develop strategies and activities that benefit the business in a number of areas. Using bad quality data could result in a counterproductive effect. The quality of data determines how good the decisions made based on the data will be. Improved data quality leads to better and more confident decision-making, reduces risk and can result in consistent improvements in results across an organization.

2. Better Audience Targeting: Without quality data, marketers are forced to reach a broad audience or try to guess at who their target audience should be, which is very inefficient. Using quality data helps to accurately determine who the target audience should be. This helps with more personalized product/content development that appeals to the right people and better advertising campaigns. While customer data can increase the effectiveness of advertisement, it is important to note that the issue of using customer data for targeted advertisement is not only a very controversial and sensitive topic, but it also requires compliance to the laws set in place (e.g., GDPR). Figure 1.1 shows how data on Facebook users is used for targeted advertisement.

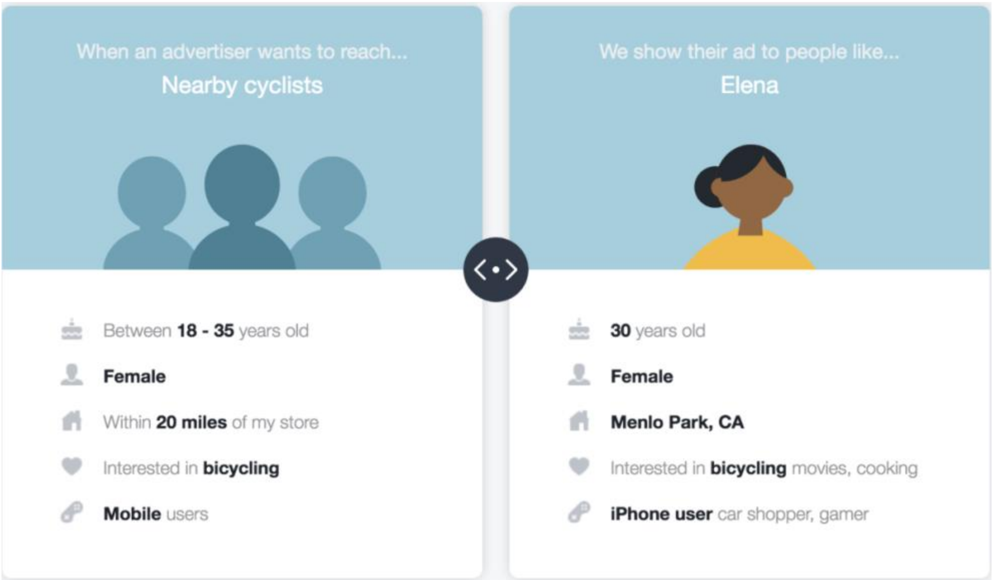

Figure 1-1-Targeting options on Facebook

https://www.digitalmarketing.org/blog/how-do-facebook-ads-work

3. Improved Relationships with Customers: High-quality data is important to boost Customer Relationship Management (CRM), which is crucial for success in any industry. CRM is a combination of strategies, technologies and practices that are used by companies to manage and analyse their customers' data and interactions throughout the whole customer lifecycle. The customer lifecycle involves the process of considering a product or service, actually purchasing this product or service, using and maintaining loyalty to the product or service. Collecting data about your customers helps you to know them better, identify trends and insights about them and offer better products and services to them. For example, having easy access to data of past purchases and history of previous interactions, can help customer support representatives provide better and faster customer service.

4. Easier Implementation of Data: Acquiring high quality data saves the company time of cleaning and processing data to make it usable. It also reduces the risk of having conclusions and decisions derived from poor quality data. Such conclusions could have errors that will be expensive and time consuming to fix. This time takes away from other activities and decreases the efficiency of the company or team. Consider an example of an order sent to the wrong address, due to an error in the dataset. In order to fix this, the item has to be returned to the warehouse and a new item has to be sent, incurring additional shipping cost and delaying the time the customer gets their correct item. This could further lead to an unhappy and frustrated customer, bad reviews, cancelation of the order etc.

5. Competitive Advantage and increased profitability: High quality data is a very valuable resource that can give competitive advantage to a company. If a company has higher data quality than its competitors, or if they are able to use their data more efficiently than their competitors, they are able to discover opportunities before their competitors and take advantage of these insights and opportunities to improve their business. Taking all these into effect, high quality data can lead to increased profitability.

1.2 Implications of data quality

The quality of data can have significant business consequences for companies. Poor-quality data is often pegged as the source of operational disarray, inaccurate analytics, ill-conceived business strategies and dissatisfied customers. If not identified and corrected early, the negative effect of poor data quality, can lead to further contamination of information assets and downstream systems and servers. The direct economic damage caused by poor data quality problems could be in the form of added expenses due to shipping products to the wrong address, lost sales opportunities due to incorrect or incomplete customer records, fines due to data quality breaches for incorrect risk assessment, improper financial or regulatory compliance reporting.

According to a research carried out by Gartner, it was found that organizations believe poor data quality to be responsible for an average of $15 million per year in losses [3]. An estimate by IBM brings the total annual cost of poor-quality data in the U.S, to $3.1 trillion in 2016 [4]. Thomas C. Redman, the president of Data Quality Solutions in an article he wrote for MIT Sloan Management Review in 2017, points out "*we estimate the cost of bad data to be 15% to 25% of revenue for most companies. These costs come as people accommodate bad data by correcting errors, seeking confirmation in other sources, and dealing with the inevitable mistakes that follow.*" [5] This financial impact contributes to the increasing trends of the quest of data quality solutions by organizations. A study carried out by RingLead [6], shows that there is a huge payoff in spending on cleaning bad records instead of bearing the cost of the impact of the bad records. For this study, the cost of preventing bad data was set at $1, while the cost of correcting the impact of bad quality data and the cost of doing nothing was set to $10 and $100 respectively. The results of this study are shown in Figure *1.2.*

Figure 1-2- The Cost of Bad Data

http://ww2.ringlead.com/rs/leadmdringlead/images/The-Cost-of-Bad-Data-Infographic.pdf

1.3 Challenges associated with the quality of data:

The diversity of the available data sources allows for the acquisition of data from organizations of which the quality level of data management or production process is unknown or weak. The many different types of data structures and types also make it difficult for data integration. In recent times, the data collected and analysed by organizations has surpassed the scope of just data generated from within their own business systems. Data could be sourced from the internet, compilations of data from various industries, scientific experimental and observational data. These sources could produce different data types which include:

- Structured data: Data which has been formatted, transformed and organized into a well-defined data model. The elements of structured data are usually contained in rows and columns and each data element has an associated fixed structure. This makes structural data easy to extract, search and organize. The most common type of structured data are relational databases.

- Semi-structured data: This type of data lies between structured and unstructured data. They have some consistent and definite characteristics, but it does not conform to the rigid structure that is expected in relational databases. They have a high degree of flexibility in that with some processing, they can be formatted and stored in a relational database. An example of semi-structured data is email messages. While the actual content of the email is unstructured, structured data such as the name of the sender and recipient, their email address, the time and date sent, etc are contained in the email message. Other examples of semi-structured data include CSV, XML and JSON documents, NoSQL databases, HTML, etc

- Unstructured data: Data with no specific structure, format or pre-defined organization. Majority of the data that exists today is unstructured such as videos, audios, texts, social media content, etc. the lack of structure, makes this type of data difficult to search, manage or analyse. Machine learning algorithms and artificial intelligence can be used to process unstructured data and make sense out of it. For example, with sentiment analysis, a machine learning model can be carried out on social media content to figure out what is trendier to consumers or to determine how effective a marketing campaign is.

When organizations acquire data from different sources and with complex structures, integrating them effectively becomes difficult especially in the cases of big voluminous data. In addition to this, data generated from within the organization could be marred due to data entry errors by the employees or errors by customers, when they are allowed to enter data about themselves directly to the operational systems. Examples of such errors include misspellings, placing data in the wrong fields, missing or incorrect codes, etc.

The manner in which data changes over time is a challenge for maintaining the quality of data. Experts say about 2% of records in a customer file become outdated in a month due to factors like death, divorce, marriage and movement [7]. If organisations are unable to acquire required data in real time or work to constantly update processed data, they could be working with obsolete or invalid data. Analysis carried out based on such data will produce misleading results and lead to decision making mistakes.

The massive volume of data makes it difficult to ascertain the quality of data within a reasonable amount of time. The volume and rapid growth of data is increasing every day and the majority of this data is unstructured. While organizations can easily acquire large amounts of various data types and structures, sufficient processing abilities and technological infrastructure is required to clean, integrate and manage such data. Due to the very high proportion of unstructured data in big data, more time and skill is required to transform these unstructured data types into structured and further process the data to obtain the necessary high-quality data.

Chapter 2

2. The Standards for data quality

The term ‘quality’ is very common and well known. Although the word has been used for so many years, it is quite ambiguous and very often misunderstood. This ambiguity is partly due to the fact that quality is not just a single idea but a multidimensional concept, where dimensions are practical yardsticks for data quality that need to be defined and measured. For example, I could argue that tap water is bad quality, but what makes it bad quality? Does it have a bad taste? Is it an issue with how clean or how old the pipes are? Is this a general problem or country specific situation? Is there some coloration in the water? Does it have to do with the PH levels of the water, the chemicals or impurities inside the water? All of these attributes can be used to measure the quality of water by different individuals which could lead to several different opinions on the quality of tap water. The problem of the conflicting ideas when it comes to quality can be solved when standards are set.

With regards to quality, the concept of ‘dimension’ classifies aspects of data quality expectations and provides measures to evaluate conformance to these measures.[8] This implies that dimensions of data quality describes a context for data quality attributes, a frame of reference to have these attributes measured as well as suggested units of measurements. These metrics make it possible for the levels of data quality to be measured. They are also used to identify the gaps and opportunities for improvement of data quality across an information flow.

Some national bodies have come together and agreed upon uniform standards for quality requirements and evaluation to which companies, developers and vendors should align their data quality management mechanisms with. These standards include the ISO 8000 and the ISO/IEC 25000. In this chapter, the divisions of the ISO/IEC 25000 series will be briefly described but we will be focusing particularly on the data quality models in the ISO/IEC 25012 and their measurements as stated in the ISO/IEC 25024.

2.1 The ISO / IEC 25000 Standard

The International Organization for Standardization - ISO and The International Electrotechnical Commission- IEC make up a specialized system for worldwide standardization. The members of the ISO or IEC, usually national bodies participate in developing international standards. This is achieved through technical committees established to focus on particular technical activities.[9] One of the technical committees is the ISO/IEC JTC 1 for the field of information technology.

The ISO/IEC 25000 consists of a series of standards that outline guidelines for quality requirements and their evaluation. SQuaRE (System and Software Quality Requirements and Evaluation) is the latest framework that supports the ISO/IEC 25000. It is made up of fragments of standards based on directives present in ISO/IEC 9126 and ISO/IEC 14598 as shown in Figure 2.1. The objective of the SQuaRE series is to support the specification of software quality requirements and the evaluation of software quality, through defined and standardized criteria for measurement and evaluation. The standards assist with development and acquisition processes of system and software products.

The standards related to the "SQuaRE" series are divided into the following 5 divisions:

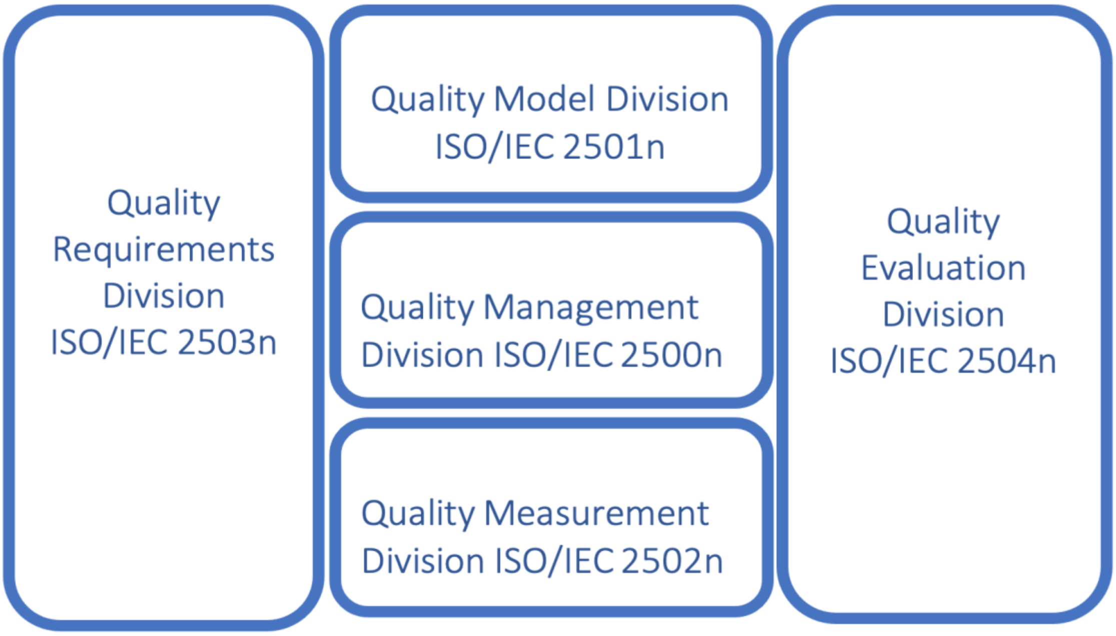

Figure 2-1 - Organization of SQuaRE series of standards

- **Quality Management Division (ISO/IEC 2500n)**: The standards that form this group give an overview of the models, terms and definitions used by all the other standards in the square series. It also provides guidance to manage technologies required for the use of SQuaRE. This division currently includes:

    - ISO/IEC 25000 - Guide to SQuaRE: Provides a general overview of the contents of SQuaRE, the referenced models, terminology, documents overview, as well as specification of the intended users.
    - ISO/IEC 25001 - Planning and Management: Provides support in the form of recommendations, technology, tools, management skills and experiences to organizations involved in the management and planning of systems and software product quality requirements specification and evaluation process.

- **Quality Model Division (ISO/IEC 2501n)**: The standards that make up this division provide detailed quality models for data, systems and software products and quality in use. It also presents practical guidelines on the use of the quality models. This division currently includes:

    - ISO/IEC 25010 - System and software quality models: It provides characteristics and sub characteristics of product quality model and quality in use model. Which are applicable to both software products and computer systems.
    - ISO/IEC 25012 - Data Quality model: Provides a quality model for data which is maintained in a structured format within a computer system. It describes 15 quality dimensions, applicable to data used by humans and systems.

- **Quality Measurement Division (ISO/IEC 2502n)**: The standards that make up this group provide a referenced model for software quality measurement including metrics for quantifying the quality measurement and practical guidelines for their applications. This division currently includes the following standards:

    - ISO/IEC 25020 - Measurement reference model and guide: It provides guidance for the selection and customization of software quality measure, according to the use case.
    - ISO/IEC 25021 - Quality measure elements: Provides measures to be used throughout the whole life cycle of software development. In addition, it presents guidelines for designing quality measure elements or verifying the design of already existing quality measure elements.

    - ISO/IEC 25022 - Measurement of quality in use: Provides metrics and guidelines for measuring quality in use.
    - ISO/IEC 25023 - Measurement of system and software product quality: Provides metrics and guidelines for measuring system and software product quality.
    - ISO/IEC 25024 - Measurement of data quality: Provides quantity measures useful for the quantitative assessment of the data quality characteristics described in ISO/IEC 25012.

- **Quality Requirements Division (ISO/IEC 2503n):** This division is made up of only one standard, which supports the specification of quality requirements for software products. The quality requirements can be used in the software product development or as an input for an evaluation process. It consists of:

    - ISO/IEC 25030 - Quality requirements: This standard presents requirements and recommendations for both quality requirements and the process used in the development of quality requirements.

- **Quality Evaluation Division (ISO/IEC 2504n):** The standards that form this group present software product evaluation requirements, guidelines and recommendations. This division currently includes the following standards:

    - ISO/IEC 25040 - Evaluation reference model and guide: Provides and evaluation framework for software product quality. It also specifies the requirements for the methods of measuring and evaluating software products.
    - ISO/IEC 25041 - Evaluation guide for developers, acquirers and independent evaluators: It presents guidelines and recommendations and for quality evaluation to be used by acquirers, developers and independent evaluators.
    - ISO/IEC 25042 - Evaluation modules: Provides a description of the structure and content of the documentation made for the purpose of describing an evaluation module. The evaluation modules consist of the specification of the quality model, the associated data and information about its application.
    - ISO/IEC 25045 - Evaluation module for recoverability: Presents the specification for the assessment of recoverability which is a sub characteristic defined under reliability quality model.

2.2 Data Quality Dimensions

The ISO/IEC 25012 presents a quality model that organizes data quality into 15 characteristics or dimensions based on the inherent and system dependent point of view.

The inherent data quality indicates the degree to which the characteristics of data quality have intrinsic potential to satisfy needs when data is used under certain conditions. From this point of view, data quality refers to data itself, in particular to a) Data domain values and possible restrictions b) Relationships of data values (e.g., consistency) c) Metadata [9]

System dependent data quality indicates to the degree to which data quality is attained and preserved within a computer system when data is used under specified conditions. From this point of view, data quality is dependent on the technological domain in which data is used. It depends on the capability of the computer systems (the hardware and software). For example, the capability to make data available or to obtain precision is achieved by the hardware, while migration tools or backup tools to achieve recoverability is achieved by software systems.[9]

Table 2.1. summarizes data quality model characteristics, classifying them based on their relevance to the inherent and system dependent point of views. As seen on Table 1, some characteristics are relevant to both sides. These characteristics are defined in detail in this section.

2.2.1 **Accuracy**

The degree to which data value correctly corresponds to the intended actual real-world values in a specific use case. In other words, it asks the question of how much or to what extent the recorded data represents or conforms to the true value which was intended. It can also be defined as the proximity between a value v and a value v^1, where v^1 represents the real-life phenomenon that v aims to portray [10]. It can be classified into syntactic accuracy and sematic accuracy.

Table 2-1-Data quality model characteristics

Characteristics	**Data Quality**	
	Inherent	**System dependent**
Accuracy	X	
Completeness	X	
Consistency	X	
Credibility	X	
Currentness	X	
Accessibility	X	X
Compliance	X	X
Confidentiality	X	X
Efficiency	X	X
Precision	X	X
Traceability	X	X
Understandability	X	X
Availability		X
Portability		X
Recoverability		X

<u>Syntactic accuracy:</u> The proximity of a value v in our dataset to the elements of a corresponding domain constraint D. That is, the level of closeness of the values in our dataset to a set of defined values which are considered syntactically correct in a domain. So, with syntactic accuracy, it does not matter whether the value v actually corresponds to the true value v^1, as long as the value v is considered correct in the domain. For example, consider a section of the Kaggle IMDb dataset[11], on the fourth row of Table 2.2, the language specified for the movie '18 regali' is English (v). While the correct language (v^1) should be Italian, this is not syntactically inaccurate because English is an acceptable value in the domain for the language column. On the other hand, 'Englh' is syntactically inaccurate because it does not correspond to any existing language. It is most likely a misspelling of the word 'English'. Syntactic accuracy can be measured by comparison functions such as the edit distance. The edit distance quantifies the dissimilarity between two strings[12]. In the example of 'Englh' to 'English', the edit distance is 2 because a minimum of 2 edits (inserting 'i' and 's') is required to change the string Englh' to 'English'. These edits could be a deletion, insertion or replacement of a character.

Sematic accuracy: The proximity of the value v to the true value v^1, which v intends to represent. For example, a sematic error can be seen in tuple 2 and 3 of the directors' column in Table 2.2, where the directors have been switched. While 'Prince Bagdasarian' and 'Nicolas Pesce' are names of directors making them admissible to the directors' column and therefore syntactically correct, 'Nicolas Pesce' is not the director of the movie 'Diverted Eden' and 'Prince Bagdasarian' is not the director of the movie 'The grudge'. This switch has created a sematic error because in both cases, the v does not correspond with the true value v^1 which it intends to represent.

The grudge Prince Bagdasarian
Diverted Eden Nicolas Pesce

Measuring the semantic accuracy of a value v requires that the corresponding true value v^1 should be known or there should be a possibility with additional knowledge to deduce whether the value v is or is not the true value v^1. Taking this into consideration, it is clear that sematic accuracy is more complex to calculate than syntactic accuracy. Semantic accuracy can be measured with a <yes, no> or a <correct, not correct> domain.

Table 2-2- IMDb movies dataset

	imdb_title_id	title	date_published	duration	avg_vote	country	language	director	Year	genre
1	tt8810394	The Point of No Return	6/16/2020	110	3	UK	NaN	Rick Roberts	2020	War
2	tt3612126	The Grudge	3/5/2020	94	4.2	USA, Canada	English	Prince Bagdasarian	2020	Horror, Mystery
3	tt3580692	Diverted Eden	3/1/2020	110	4.2	USA	English	Nicolas Pesce	2020	Action, Crime, Drama
4	tt10816484	18 regali	1/2/2020	115	6.7	Italy	Italian	Francesco Amato	2020	Drama
5	tt10814876	7 ore per farti innamorare	4/20/2020	93	5.9	Italy	Italian	Giampaolo Morelli	2020	Comedy
6	tt5747714	Unbound	2/7/2020	97	4.5	USA	Englh	Steve Rahaman	2021	Action, Crime, Drama
7	tt10806028	Agir Romantik	2/14/2020	97	8	Turkey	Turkish	Deniz Denizciler	2020	Comedy, Drama, Romance
8	tt8675288	Il mio nome è Imp@vido_	8/14/2020	89	4.8	Canada	English	Cory Edwards	2019	Animation, Comedy, Family
9	tt4789618	Still Here	8/28/2020	99	7	USA	English	Vlad Feier	2020	Crime, Drama, Thriller
10	tt10801196	Pressure Cooker	2/21/2020	135	6.4	India	Telugu	Sujoi, Sushil	2020	Comedy, Drama, Family

2.2.2 Completeness

The degree to which all data necessary to represent an entity are available and recorded in the system. This implies that there is a recorded value for all the expected attributes and related instances, which are considered fundamental for a specific context of use or corresponds to the real-world system. For example, a data set of students living in a hostel is considered to have a completeness issue if some students' records do not contain the data regarding the phone number of their next of kin, who can be contacted in case of an emergency. Completeness problems are usually identified by the presence of missing or null value i.e., a value that should exist in the real world but is not available in the data set. Consider the language column of Table 2.2., there is a missing value 'NaN' associated with the movie title 'The point of no return'. This is a completeness issue because the language for this movie indeed exists in the real world and the value should be 'English'.

2.2.3 Consistency

The degree to which the attributes of data do not have discrepancies and are coherent with and verifiable by other data in a specific context use. Consistency can be verifiable within the same dataset, for example, a data set which contains an entity with the marital status filled as 'married' and the age filled as '3', is clearly an inconsistency problem because it is well known to everybody that a three-year-old cannot be married. In this case, from the attribute 'married' in the dataset, we are able to identify a discrepancy in the attribute 'age'. Consistency can also be verifiable also across similar data that are comparable. An example could be seen in the case, where the sum of the students in each department, is not equal to the generally known and accepted total population of students in a university. From both examples, there is a rule or constraint that must not be violated in order for the data to be consistent.

For a clearer picture of the idea of constraints, consider the 'date_published' and the 'Year' columns on Table 2.3. It is clear that year on both of these columns must be the same for the data to be considered the same. In order to check if there is indeed consistency among the columns, a python code can be written to create a new column, where 'True' is 'False' is returned depending on whether the year is coherent on both columns or not. The result of the code below identifies two inconsistent values highlighted on Table 2.3.

```
df6['comparison_column'] = np.where(df6['Year'] !=
pd.DatetimeIndex(df6['date_published']).year, 'False','True')
```

2.2.4 Credibility

The extent to which the attributes of data are considered to be trusted or believable in terms of their source or content, for a specific context of use. Data can be considered credible, if it has been certified from an independent and trusted organization [9]. For example, credit risk information that has been certified by internal audit is considered credible and can be used by banks for evaluating credit risk.

Table 2-3- Result of the code to check for consistency

	imdb_title_id	title	date_published	duration	avg_vote	country	language	director	Year	genre	comparison_column
1	tt8810394	The Point of No Return	6/16/2020	110	3	UK	NaN	Rick Roberts	2020	War	TRUE
2	tt3612126	The Grudge	3/5/2020	94	4.2	USA, Canada	English	Prince Bagdasarian	2020	Horror, Mystery	TRUE
3	tt3580692	Diverted Eden	3/1/2020	110	4.2	USA	English	Nicolas Pesce	2020	Action, Crime, Drama	TRUE
4	tt10816484	18 regali	1/2/2020	115	6.7	Italy	Italian	Francesco Amato	2020	Drama	TRUE
5	tt10814876	7 ore per farti innamorare	4/20/2020	93	5.9	Italy	Italian	Giampaolo Morelli	2020	Comedy	TRUE
6	tt5747714	Unbound	2/7/2020	97	4.5	USA	Englh	Steve Rahaman	2021	Action, Crime, Drama	FALSE
7	tt10806028	Agir Romantik	2/14/2020	97	8	Turkey	Turkish	Deniz Denizciler	2020	Comedy, Drama, Romance	TRUE
8	tt8675288	Il mio nome è Imp@vido_	8/14/2020	89	4.8	Canada	English	Cory Edwards	2019	Animation, Comedy, Family	FALSE
9	tt4789618	Still Here	8/28/2020	99	7	USA	English	Vlad Feier	2020	Crime, Drama, Thriller	TRUE
10	tt10801196	Pressure Cooker	2/21/2020	135	6.4	India	Telugu	Sujoi, Sushil	2020	Comedy, Drama, Family	TRUE

2.2.5 Currentness

The degree to which the attributes of data are up to date with the facts in the real world. As explained in Chapter 1, data needs to be updated in order to avoid the risk of working with obsolete data. For example, the flight itinerary must be updated with the frequency required to allow passengers to catch a flight even if the scheduled time or gates change. According to its change frequency, Carlo Batini[10] classifies data into:

- Stable data: which is unlikely to change, such as scientific publications. While new publications can be added to the source, the older publications remain the same.
- Long-term changing data: which has a very low tendency to change, such as currency, addresses.
- Frequently changing data: which is very change intensive, such as the real time traffic information, stock prices, etc.

The measurement of currentness has to consider not just how quickly data is updated, but also if the updated data is available before the time of its intended use.

2.2.6 Accessibility

The degree to which data can be made available or retrievable with ease, particularly to users who due to some disability, require supporting technology or special configuration. The technology could be in form of screen readers, for visually impaired people to access text or text alternatives for people with hearing impairments to access audio or video content. An example of accessibility issue could be storing data as an image, when it is intended to be managed by a screen reader.

2.2.7 Compliance

The extent to which the attributes of data adhere to the rules, standards, conventions or regulations in force. For example, data managed by all credit card companies, must be PCI compliant. The PCI DSS ensures that credit card transactions in the payments industry are secure. Also credit risk data managed by banks must be compliant to the specific standards and regulations.

2.2.8 Confidentiality

The extent to which access to data is appropriately restricted and protected, only to be made available or interpretable to authorized users. This implies that, nonpublic personal data or confidential information such as patient's health records, must be protected and only accessible by authorized users or be written in secret code which only authorized users can interpret. In order to achieve this, several methods can be applied such as data swapping. Data swapping is a technique for statistical disclosure limitation (SDL) used to modify characteristics in a database, by exchanging a subset of attributes between selected pairs of records, making it impossible for an intruder to identify the individual entities in the database [13].

2.2.9 Efficiency

The degree to which the attributes of data can be processed in such a manner that the expected levels of performance are provided using the appropriate types and number of resources. For example, storing data in such a way that more space than necessary is used, can result to a waste of memory, time, storage, money and reduced efficiency of processing. Deduplication (i.e., eliminating copies of the same data stored in a dataset) and compression (i.e., reducing the number of bits required to represent data) are methods that can be used to reduce the total storage space and increase efficiency. The format in which the data is stored, is another aspect to be considered. Data is best stored in open formats such as csv, xml, JSON, etc., to ensure easier processing, interoperability and reduce the risks of mistakes or data loss during conversion between formats.

2.2.10 Precision

The degree to which the attributes of data provide distinguishable characteristics or the exact amount of information, required in the context of which it is used. For example, in numeric data, the number of decimal places to the right of the decimal point, can specify the level of precision. Consider the following numbers below, the number ‘5.3842’ with the highest level of precision, allows for more functionalities than the number ‘5’ with the lowest level of precision.

5 → 5.3 → 5.38 → 5.384 → 5.3842

Lowest level of precision *Highest level of precision*

2.2.11 Traceability

The extent to which the attributes of data provide information, that identifies the sources of any new or updated data attribute. In other words, the extent to which an audit trail is provided for any access or changes to the data. An audit trail can be defined as data stored in a record, which identifies how, when, and by whom data was created, accessed or modified [14]. For example, public administrations keep information about the access executed by users, this is helpful for investigating who read or wrote confidential data [9]. Figure 2.2 shows an example of a reporting table entry which includes a reference to source data element.

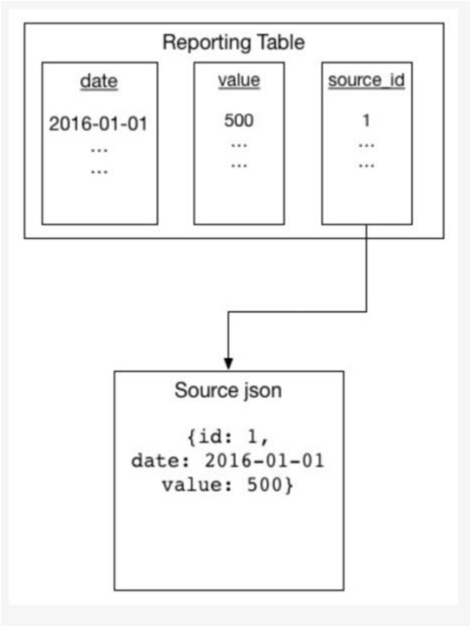

Figure 2-2- A reporting table entry which includes a reference to source data element

https://engineering.squarespace.com/blog/2016/date-traceability-and-lineage

2.2.12 Understandability

The extent to which data attributes are presented in such a manner that they are easily to read and interpret in appropriate languages, symbols and units. In other words, they are free from ambiguity and easily understandable. For example, a dataset which contains the regions in Italy, it is more understandable if the regions are represented with the standard acronyms rather than numeric code. Understandability can be facilitated by either linked or existing meta data.

2.2.13 Availability

The extent to which data attributes are retrievable for use, by authorized users or applications, in the specified context and time frame in which they are expected. Availability considers two aspects [9];

- Availability as a form of concurrent access, that allows multiple users or applications to read or modify data. For example, during intensive managing operations such as backup, data should also be available.
- Availability as a subsection of currentness, that allows data to be retrievable within a specific time frame. For example, a weather forecast application should be able make current data available.

2.2.14 Portability

The degree to which data has attributes that allow for it to be applied in as many set of situations as possible, while maintaining its existing quality [10]. In other words, data is stored in a format that allows for installing, replacing or transferring the data from one platform to another, maintaining the existing quality.

2.2.15 Recoverability

The ability of data to maintain and preserve a specific level of operation and integrity, both physically and logically, even in the event of failures. Failures could include accidental deletion, data loss due to power outages, equipment malfunction, etc. Recoverability can be achieved through backup recovery features such as commit (a feature that guarantees updates to a database are written to disk at a point in time), rollback (a feature that returns the dataset to a previous state) or using cloud storage solutions where data has a higher rate of retrievability, amazon web services replicates object data stored across 3 availability zones.

In subsequent chapters, data quality tools will be studied and these characteristics, will serve as a guide to assess and evaluate the functionality of these tools.

Chapter 3

3 Data Quality Tools

Data quality tools are used for assessing data, with the aim of detecting and fixing the data problems that influence the overall quality of data. They include technologies and processes used to identify, understand and correct flaws in data, review the data source, transform data so it aligns with the generally accepted standards and business rules. In recent times, several tools have been developed by different vendors to solve or assist in solving the data quality problems. In this chapter, we will look at some of these tools and their features.

3.1 Data Quality Processes

Data quality tools can be grouped according to the data quality processes they support. These processes include data profiling, data cleaning, data integration, data monitoring, data enrichment, data governance. In this section, we will be defining these processes and some of their features which are found on the comparison matrix in Table 3-2.

3.1.1 Data cleaning

Data cleaning is a process of detecting and removing invalid, incorrect, irrelevant, outdated, redundant, inconsistent, poorly formatted or inaccurate records from a dataset.

- Deduplication: A technique for eliminating redundant or excessive copies of data.
- Data transformation: A technique for converting a data values, structures or format, from the data format of a source data system into the data format of a destination data system. Transformation changes the representation of a value without changing the content [15]. For example, gender maybe represented as 1 and 2 in the source data system, a transformation can translate 1s to M and 2s to F if it is required in the destination data system.

- Data parsing and standardization: A technique for formatting data values to a consistent and uniform pattern, based on, local or user-defined standards. For example, changing ‘Avenue’ in address column to ‘Ave’. Regular expressions are often used to achieve this.
- Identity resolution and Record matching: Techniques used to recognize variations that suggest whether two records refer to the same entity or determines that they truly represent distinct entities [8]. Data faceting is an example of such techniques.
- Data imputation: Techniques used to assign to missing data with a plausible estimated value (e.g., mean) based on available information.

3.1.2 Data Profiling

Data profiling is a process of examining a dataset and retrieving statistical and technical information about that data. Data profiling creates informative summaries of a database which give insights to the structure, content, quality of data and relationships among values in the dataset.

- Uniqueness analysis: A techniques used to Identify duplicate records and determine whether there are unique values in the key columns [16].
- Missing/ Null values identification: Identifies the occurrence of missing or incomplete records in the dataset.
- Pattern detection: Is used to identify patterns within the data and facilitate error detection by identifying pattern violations.
- Column property analysis: Are used to obtain details about the columns in a table such as the data types (string, int, float, date, etc), frequency or value distribution of patterns, the mean, median, max and min values, outliers, etc.
- Cross-column profiling: It is used to identify dependencies across columns within the same dataset. It consists of key analysis, which identifies primary keys across collections of attribute values, and dependency analysis, which identifies relationships between attributes in the same table.
- Value distribution: It shows the relative frequency (count and percentage) of the assignment of distinct values, missing values, etc.
- Cross-table analysis: Compares data between tables and indicates foreign keys. By identifying overlapping or identical set of values each column, it determines relationships that exists across every table loaded into a project.
- Clustering: Used to identify groups/clusters of data with the same actual value but different representations.

3.1.3 Data Integration

Data integration is a process of combining data from different sources and presenting the data in a unified view that facilitates analysis [17]. some of the features include:

- Data extraction, Transformation and consolidation is a process used to blend data from several sources. It involves taking data from a source system, converting the data to a format that can be analyzed and stored or loaded into a target database. It is usually used to build a data warehouse.
- Metadata management: techniques to capture and document metadata (i.e., data about other data). This is important because information contained in metadata provides understanding to both humans and machines, facilitating interoperability and integration.

3.1.4 Data monitoring

Data monitoring is a process used to enforce data quality standards and rules. In other words, it is used to maintain regulatory or best practices compliance [18]. With proper data monitoring, when there are issues with the data, users can identify and address such issues before there is a decline in the quality of data. Some features that support this process include:

- Data lineage tracking: used to track the origin, transference and transformations of data over time. It requires reporting the details of how data is manipulated, where it is used and who has access to it at every layer of action [19].
- Modification history tracking: used to identify changes or modifications that has been done to a database and when they were done.

3.1.5 Data Governance

Data governance a combination of policies, procedures, technology and tools, which are necessary to maintain control and effective operation of data quality [19].

- Data privacy and security: used to provide controlled access to data on computer system.
- Access controls: they are security restrictions used to control who can create, update, or delete data based on an identifying value or user id on a data object or a range of data within an object [19].

3.1.6 Data Enrichment

Data enrichment a process of adding or updating information to existing databases to improve accuracy. We consider email address validation, phone number validation and address validation as features of this process.

3.2 Exploring the data quality tools on the matrix

- **OpenRefine**: OpenRefine is a powerful Java-based tool designed to work with messy data and improve it. With this tool, it is possible to load, understand, clean, format, transform, reconcile, and augment data with web services and external data, for analytics and other purposes[20].

- **Datacleaner:** Data cleaner is a profiling and wrangling tool which supports data quality analysis. It can be used for data cleansing, transformations, enrichment, deduplication, matching and merging[21].

- **SQL Power Dqguru:** SQL Power DQguru is a cleaning tool from the SQL power group. It ca be used for cleaning data, validating and correcting addresses, identifying duplicates, performing deduplication, and building cross-references between source and target tables[22].

- **SQL Power Architect:** SQL Power Architect is a data modelling and profiling tool from the SQL power group, used for facilitating warehouse design. With this tool, it is possible to reverse-engineer existing databases, perform data profiling on source databases, and auto-generate ETL metadata[23].

- **CSVkit**: Csvkit is a set of command line tools that allow supports converting different formats such as excel and JSON to CSV. It also supports data extraction from PostgreSQL in CSV format and can export csv formated data directly into PostgreSQL database tables. While working on formatted data, the user is able to view, select and reorder columns, and also filter rows and records based on the data they contain[24].

- **Trifacta:** Trifacta is data analysis software that includes features such as data discovery, data visualization, high volume processing, predictive analytics, regression analysis, sentiment analysis, statistical modeling, and text analytics. It also facilitates building, deploying and managing self-service data pipelines. Trifacta is available on cloud[25].

- **Cloudingo:** Cloudingo is data cleansing software provided by SaaS. It can be used for data deduplication, data migration, data profiling, master data management and match & merge. It also helps to identify human errors and other flaws or inconsistencies[26].

- **Microsoft DQS-data quality services:** Data Quality Services (DQS) is a knowledge-driven data quality tool that provides ways to manage the integrity and quality of your data. It provides a medium to discover, build and manage knowledge about your data. That knowledge base can then be used to perform several data quality tasks such as data matching, profiling, correction, enrichment, standardization, and deduplication. It also provides features which enables you to ensure the quality of your data by comparing it with data guaranteed by a third-party company[27].

- **Talend Open Studio:** Talend open studio is a tool which supports data quality analysis of different types of fields, databases and file types. It can be used for data deduplication, validation, standardization and includes pre-built connectors and monitoring tools[28].

- **Data ladder:** Data ladder is a data quality tool that supports cleaning, matching and deduplication of any type of data. It also features address cleansing, verification and geocoding and Includes more than 300,000 prebuilt rules, templates and connectors for most major applications. Data ladder is available on premise and on cloud[29].

- **TIBCO Clarity:** TIBCO Clarity is a tool used for discovering, profiling, cleansing, validating and standardizing raw data collected from different sources, and providing good quality data for accurate analysis and intelligent decision-making[30].

- **Validity DemandTools:** Validity DemandTools is a data quality tool that can be used to control, standardize, deduplicate, import and generally manipulate Salesforce data. The modules of this tool can be divided into 3 sections and they include:
    1. Cleaning tools which provide solutions for identifying, preventing and merging duplicates, and flexible lead conversion.
    2. Maintenance tools which provide solutions for clean data loading, on-demand data backups, and data manipulation.

3. Discovery Tools which provide solutions for comparing external data to existing Salesforce data before import and verifying email addresses on demand[31].

- **Ataccama DQ Analyzer:** DQ Analyzer is a sub section of Ataccama Data Quality Center that focuses on data analysis. The tool allows for execution of complex transformations, supports data profiling and reveals relevant information which could be hidden within the data[32].

- **Datameer:** Datameer is a cloud-native platform that allows users to integrate, transform, discover, and operationalize datasets to their projects without any code. It includes features such as collaboration, data blends, data cleansing, data mining, data visualization, data warehousing, high volume processing, No-Code sandbox, and templates[33].

- **Informatica Data Explorer:** Data Explorer is a tool by informatica that provides data profiling and data quality solutions which enables developers to carry out a faster and thorough analysis of data in the repository. It can identify anomalies and hidden relationships by scanning all data records from any source. It features pre-built rules which can be applied to data for profiling[34].

- **SAS Data Management**: SAS Data Management is a cloud-based master data management software. The tool allows users to improve, integrate, manage and govern data. It includes features such as data capture, data integration, data migration, data quality control, and master data management. SAS Data Management helps you access the data you need, create rules, collaborate with other teams and manage metadata so you're prepared to run analytics for better decision making [35]. It works well with the data profiling tool, DataFlux which is also offered by SAS [36].

- **Pentaho**: Pentaho is a data integration and analytics tool that allows users to access, manage, cleanse and prepare diverse data from different sources. Although Pentaho is mostly a data integration tool, it works well with data profiling and cleaning tools such as data cleaner [37].

- **WINpure:** Winpure is a data cleaning and matching tool designed to increase the accuracy of customer data. The features of this tool includes Profiling, Cleansing, Matching, Deduplication, Global Address Verification, phone and email verification and developer API toolkit [38].

- **Experian data quality (Aperture Data Studio):** Aperture Data Studio is a data quality management platform that allows users to understand their data and make it fit for business use. Some of its features include data transformation, name, address, and email validation, consumer data enrichment, and data profiling [39].

- **Aggregate Profiler/osDQ**: Aggregate Profiler is a data profiling and quality tool which can be used for quality assessment and correction, profiling of data, both statistical analysis of data and visualization in form of charts. Some other features of the tool includes anomality detection, random data generation, populating database values, looking into database metadata and fetching and storing data from/to databases cardinality checks between different tables within one data source [40].

- **Semarchy xDM**: xDM is an all-in-one platform for master data reference data, application data, data quality, and data governance. The tool allows users to:
    - Discover Access any source, profile data, discover critical assets, and build data catalogs.
    - Integrate Connect applications and external data in real-time or batch with REST APIs.
    - Manage Deploy apps for data champions & users with built-in data quality, match/merge & more
    - Govern Build business glossaries, define & enforce policies with rules and business processes
    - Measure Analyze metrics on any data, define ad-hoc KPIs, and take actions with Dashboards [41].

3.3 Comparison Matrix

The comparison matrix for the 21 data quality tools described in section 3.2 rates the tools in absolute terms against the specified criteria using the symbols;

+ Feature is supported

± Feature is somewhat supported with additional extensions

“Blank” - Not available or not supported

The criteria considered in order of appearance on the tables, include the following:

Table 3-1

- The operation environments or operating systems which are supported by the tools.
- The license type of each tool, assigning symbols to the tools which are open source.
- The pricing models of each tool, classified as free, paid and free trial.
- The supported data sources, including file formats and databases which are supported by each tool.
- The reporting format, which could be graphical or tabular.

Table 3-2

- The data quality processes described in section 3.1

Table 3-1-Data quality tools comparison matrix

TOOL			Open Refine	DataCleaner	SQL Power Dqguru	SQL Power Architect	CSVkit	Trifacta	Cloudingo	Microsoft DQS-data quality	Talend Open Studio	Data Ladder	TIBCO Clarity	Validity DemandTools	Ataccama DQ ANALYZER	DATAMEER	Informatica Data Explorer	Sas data management	Pentaho Data Integration	WINpure	Experian data quality	Aggregate Profiler / osDQ	Semarchy xdm
SUPPORTED OPERATING SYSTEM		Windows	+	+	+	+	+	+		+	+	+	+	+	+		+	+	+	+	+	+	+
		Mac os	+	+	+	+	+	+			+		+			+			+				+
		Linux	+	+	+	+	+	+		+	+		+			+	+	+	+			+	+
		Others		+	+	+	+	+	+		+			±	±			+			+	+	+
LICENSE TYPE		Open source	+	+	+	+	+				+								+			+	
PRICING		Free	+	+	+	+	+				+				+				+	+		+	
		Paid		+	+	+		+	+	+	+	+	+	+		+	+	+	+	+	+		+
		Free trial		+				+	+		+	+	+	+		+	+	+	+	+	+		+
SUPPORTED DATA SOURCE	FILE FORMAT	CSV, TSV	+	+			+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
		JSON	+	+			+	+			+		+			+	+						+
		MS Excel (.xls or .xlsx)	+	+			+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
		XML and other RDF	+	+				+			+	+	+	±			+	+	+	+		+	+
		Support for other formats	+	+				+			+		+		+	+	+	+	+	+	+		+
	DATA BASES	MySQL	+	+	+	+	+	+			+	+	+		+	+	+		+	+	+	+	
		SQL Server		+	+	+		+		+	+	+	+		+		+	+	+	+	+	+	+
		Oracle		+	+	+		+			+	+	+		+	+	+	+	+	+	+	+	+
		PostgreSQL	+	+	+	+	+	+			+	+	+		+		+		+	+	+	+	+
		DB2				+		+			+	+	+		+	+	+	+	+		+	+	
		Redis											+										
		NoSQL(MongoDB, Hbase,etc)		+		+		+			+		±				+	+	+		+		
		Others	+	+				+			+		+		+	+	+	+		+	+		+
REPORTING		Tabular	+	+				+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
		Graphical	+	+	+	+		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+

Table 3-2- Data quality tools comparison matrix II

TOOL			Open Refine	DataCleaner	SQL Power Dqguru	SQL Power Architect	CSVkit	Trifacta	Cloudingo	Microsoft DQS-data quality services	Talend Open Studio	Data Ladder	TIBCO Clarity	Validity DemandTools	Ataccama DQ ANALYZER	DATAMEER	Informatica Data Explorer	Sas data management	Pentaho Data Integration	WINpure	Experian data quality	Aggregate Profiler / osDQ	Semarchy xdm
DATA QUALITY PROCESSES	DATA PROFILING	Uniqueness analysis	+	+		+		+	+	+	+	+	+	+	+	+	+	+	±	+	+	+	+
		Missing values identification	+	+		+	+	+	+	+	+	+	+	+	+	+	+	+	±	+	+	+	+
		Pattern detection	+	+	±	+	+	+		+	+	+	+	+		+	+	+		+		+	+
		Column property analysis	+	+		+	+	+	+	+	+	+	+	+	+	+	+	+		+	+	+	+
		Value distribution (frequency analysis)	+	+		+	+	+		+	+	+	+		+	+	+	+		+	+	+	+
		Cross-column Profiling (functional dependencies)	+	+		+	±	+		+	+	+	+	+	+	+	+	+	±	+	+	+	+
		clustering	+			+		+	+	+	+		+			+	+	+	±	+		+	+
		Outliers	+					+		+	+		+			+	+	+					
		Precision			+	+		+		+	+					+	+		+				+
		Cross-Table analysis	±	+		+	±	+		+	+						+	+				+	±
	DATA CLEANING	Deduplication	+	+	+			+	+	+	+	+	+	+	+	+	+	+		+	+	+	+
		Parsing and Standardization	+	+	+		±	+		+	+	+	+	+	±	+	+	+	+	+	±	+	+
		Record matching and Identity resolution(Faceting, mispelled value correction)	+	+	±	±		+		+	+	+	+	+	+	+	+	+	±	+	±	+	+
		Data Transformation	+	+	+		±	+		+	+	+	+	+	+	+	+	+	+	+	+		+
		Data Imputatiion (for missing values)	±	+				+		+	+	+	+	+			+	+	+			+	±
	DATA INTEGRATION	Data Extraction, Transformation and Consolidation	+			+		+			+		+	±	+	+	+	+	+				+
		Metadata Management	+	+	+	+		+		+	+			+	+	+	+	+	+		+	+	+
		Others		+		+		+			+		+			+	+	+	+				
	DATA MONITORING	Metadata respository	+	+		+		+		+	+		+	+	+	+	+	+	+		+	+	+
		Data lineage tracking			+	+		+			+			+		+	+	+	+		+	±	+
		Modification history tracking	+	+	+	+		+	+	+	+	+	+	+	+	+	+	+	+		+		+
		Others		+				+	+		+			+		+	+	+	+				+
	DATA GOVERNANCE	Data Privacy & Security	+					+	+	+	+	+		+	+	+	+	+	+		+		+
		Access controls						+	+	+	+	+	+	+	+	+	+	+	+		+	±	+
	DATA ENRICHMENT	Email validation		±						±			+	+						+	+		±
		Phone number Validation		±						±		+	+	±				±		+	+		±
		Address validation		±	±				+	±		+	+	±			+	+	+	+	+	+	±
		Others	+	+				+	+	+	+	+	+								+		+

Table 3-3-Associating the dimensions to the features of the tools

	TOOLS	ABSOLUTE WEIGHT	Open Refine	DataCleaner	SQL Power Dqguru	SQL Power Architect	CSVkit	Trifacta	Cloudingo	Microsoft DQS-data quality services	Talend Open Studio	Data Ladder	TIBCO Clarity	Validity DemandTools	Ataccama DQ ANALYZER	DATAMEER	Informatica Data Explorer	Sas data management	Pentaho Data Integration	WINpure	Experian data quality	Aggregate Profiler / osDQ	Semarchy xdm
Accuracy	Pattern detection	13	1	1	0.5	1	1	1	0	1	1	1	1	1	0	1	1	1	0	1	0	1	1
	Column property analysis		1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1
	Cross-column Profiling (functional dependencies)		1	1	0	1	0.5	1	0	1	1	1	1	1	1	1	1	1	0.5	1	1	1	1
	clustering		1	0	0	1	0	1	1	1	1	0	1	0	0	1	1	1	0.5	1	0	1	1
	Outliers		1	0	0	0	0	1	0	1	1	0	1	0	0	1	1	1	0	0	0	0	0
	Precision		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Cross-Table analysis		0.5	1	0	1	0.5	1	0	1	1	0	0	0	0	0	1	1	0	0	0	1	0.5
	Deduplication		1	1	1	0	0	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1
	Parsing and Standardization		1	1	1	0	0.5	1	0	1	1	1	1	1	0.5	1	1	1	1	1	0.5	1	1
	Record matching and Identity resolution(Faceting, mispelled value correction)		1	1	0.5	0.5	0	1	0	1	1	1	1	1	1	1	1	1	0.5	1	0.5	1	1
	Data Transformation		1	1	1	0	0.5	1	0	1	1	1	1	1	1	1	1	1	1	1	1	0	1
	Data Imputatiion (for missing values)		0.5	1	0	0	0	1	0	1	1	1	1	1	0	0	1	1	1	0	0	1	0.5
	Metadata respository		1	1	0	1	0	1	0	1	1	0	1	1	1	1	1	1	1	0	1	1	1
	TOTAL		84.62	76.92	30.77	50	30.77	92.31	23.08	92.31	92.31	61.54	84.62	69.23	50	76.92	92.31	92.31	42.31	61.54	46.15	76.92	76.92
Completeness	Missing values identification	3	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0.5	1	1	1	1
	Data Imputatiion (for missing values)		0.5	1	0	0	0	1	0	1	1	1	1	1	0	0	1	1	1	0	0	1	0.5
	Metadata respository		1	1	0	1	0	1	0	1	1	0	1	1	1	1	1	1	1	0	1	1	1
	TOTAL		83.33	100	0	66.67	33.33	100	33.33	100	100	66.67	100	100	66.67	66.67	100	100	83.33	33.33	66.67	100	83.33
Consistency	Uniqueness analysis	11	1	1	0	1	0	1	1	1	1	1	1	1	1	1	1	1	0.5	1	1	1	1
	Column property analysis		1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Value distribution (frequency analysis)		1	1	0	1	1	1	0	1	1	1	1	0	1	1	1	1	0	1	1	1	1
	Cross-column Profiling (functional dependencies)		1	1	0	1	0.5	1	0	1	1	1	1	1	1	1	1	1	0.5	1	1	1	1
	clustering		1	0	0	1	0	1	1	1	1	0	1	0	0	1	1	1	0.5	1	0	1	1
	Outliers		1	0	0	0	0	1	0	1	1	0	1	0	0	1	1	1	0	0	0	0	0
	Cross-Table analysis		0.5	1	0	1	0.5	1	0	1	1	0	0	0	0	0	1	1	0	0	0	1	0.5
	Deduplication		1	1	1	0	0	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1
	Parsing and Standardization		1	1	1	0	0.5	1	0	1	1	1	1	1	0.5	1	1	1	1	1	0.5	1	1
	Record matching and Identity resolution (Faceting, mispelled value correction)		1	1	0.5	0.5	0	1	0	1	1	1	1	1	1	1	1	1	0.5	1	0.5	1	1
	Data Transformation		1	1	1	0	0.5	1	0	1	1	1	1	1	1	1	1	1	1	1	1	0	1
	TOTAL		95.45	81.82	31.82	50	27.27	90.91	27.27	90.91	90.91	63.64	81.82	54.55	59.09	81.82	90.91	90.91	36.36	72.73	54.55	72.73	77.27
Credibility	Email validation	3	0	0	0	0	0	0	0	0.5	0	0	1	1	0	0	0	0	0	1	1	0	0.5
	Phone number Validation		0	0	0	0	0	0	0	0.5	0	1	1	0.5	0	0	0	0.5	0	1	1	0	0.5
	Address validation		0	0.5	0.5	0	0	0	1	0.5	0	1	1	0.5	0	0	1	1	1	1	1	1	0.5
	TOTAL		0	16.67	16.67	0	0	0	33.33	50	0	66.67	100	66.67	0	0	33.33	50	33.33	100	100	33.33	50
Currentness	Cross-Table analysis	2	0.5	1	0	1	0.5	1	0	1	1	0	0	0	0	0	1	1	0	0	0	1	0.5
	Record matching and Identity resolution(Faceting, mispelled value correction)*		1	1	0.5	0.5	0	1	0	1	1	1	1	1	1	1	1	1	0.5	1	0.5	1	1
	TOTAL		75	100	25	75	25	100	0	100	100	50	50	50	50	50	100	100	25	50	25	100	75
Accessibility	Access controls	1	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	0	1	0.5	1
	TOTAL		0	0	0	0	0	100	100	100	100	100	100	100	100	100	100	100	100	0	100	50	100
Compliance	Data Privacy & Security	1	1	0	0	0	0	1	1	1	1	1	0	1	1	1	1	1	1	0	1	0	1
	TOTAL		100	0	0	0	0	100	100	100	100	100	0	100	100	100	100	100	100	0	100	0	100
Confidentiality	Data Privacy & Security	2	1	0	0	0	0	1	1	1	1	1	0	1	1	1	1	1	1	0	1	0	1
	Access controls		0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	0	1	0.5	1
	TOTAL		50	0	0	0	0	100	100	100	100	100	50	100	100	100	100	100	100	0	100	25	100
Efficiency	Outliers	4	1	0	0	0	0	1	0	1	1	0	1	0	0	1	1	1	0	0	0	0	0
	Deduplication		1	1	1	0	0	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1
	Parsing and Standardization		1	1	1	0	0.5	1	0	1	1	1	1	1	0.5	1	1	1	1	1	0.5	1	1
	Data Transformation		1	1	1	0	0.5	1	0	1	1	1	1	1	1	1	1	1	1	1	1	0	1
	TOTAL		100	75	75	0	25	100	25	100	100	75	100	75	62.5	100	100	100	50	75	62.5	50	75
Precision	Precision	1	0	0	1	1	0	1	0	1	1	0	0	0	0	1	1	0	1	0	0	0	1
	TOTAL		0	0	100	100	0	100	0	100	100	0	0	0	0	100	100	0	100	0	0	0	100
Traceability	Metadata respository	3	1	1	0	1	0	1	0	1	1	0	1	1	1	1	1	1	1	0	1	1	1
	Data lineage tracking		0	0	1	1	0	1	0	0	1	0	0	1	0	1	1	1	1	0	1	0.5	1
	Modification history tracking		1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	0	1	0	1
	TOTAL		66.67	66.67	66.67	100	0	100	33.33	66.67	100	33.33	66.67	100	66.67	100	100	100	100	0	100	50	100
Understandability	Metadata Management	2	1	1	1	1	0	1	0	1	1	0	0	1	1	1	1	1	1	0	1	1	1
	Metadata respository		1	1	0	1	0	1	0	1	1	0	1	1	1	1	1	1	1	0	1	1	1
	TOTAL		100	100	50	100	0	100	0	100	100	0	50	100	100	100	100	100	100	0	100	100	100
Availability	Access controls	1	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	0	1	0.5	1
	TOTAL		0	0	0	0	0	100	100	100	100	100	100	100	100	100	100	100	100	0	100	50	100
Portability	Data Extraction, Transformation and Consolidation	12	1	0	0	1	0	1	0	0	1	0	1	0.5	1	1	1	1	1	0	0	0	1
	CSV, TSV		1	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	JSON		1	1	0	0	1	1	0	0	1	0	1	0	0	1	1	0	0	0	0	0	1
	MS Excel (.xls or .xlsx)		1	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	XML and other RDF		1	1	0	0	0	1	0	0	1	1	1	0.5	0	0	1	1	1	1	0	1	1
	MySQL		1	1	1	1	1	1	0	0	1	1	1	0	1	1	1	0	1	1	1	1	0
	SQL Server		0	1	1	1	0	1	0	1	1	1	1	0	1	0	1	1	1	1	1	1	1
	Oracle		1	1	1	1	0	1	0	0	1	1	1	0	1	1	1	1	1	1	1	1	1
	PostgreSQL		1	1	1	1	1	1	0	0	1	1	1	0	1	0	1	0	1	1	1	1	1
	DB2		0	0	0	1	0	1	0	0	1	1	1	0	1	1	1	1	1	0	1	1	0
	Redis		0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
	NoSQL(MongoDB, Hbase,etc)		0	1	0	1	0	1	0	0	1	0	0.5	0	0	0	1	1	1	0	1	0	0
	TOTAL		66.67	75	33.33	58.33	41.67	91.67	16.67	25	91.67	66.67	95.83	25	66.67	58.33	91.67	66.67	83.33	58.33	66.67	66.67	66.67
Recoverability	Modification history tracking	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	0	1	0	1
	TOTAL		100	100	100	100	0	100	100	100	100	100	100	100	100	100	100	100	100	0	100	0	100

Table 3-3 presents a quantitative analysis of the relationship between the criterion used in the comparison analysis and the dimensions of data quality. Each dimension is matched to the criterion/feature that can be used to identify and fix data quality issues related to the dimension. Score points were allocated to the + , ± , blank (1, 0.5, 0 respectively), then the average score points (%) for each dimension is calculated for each tool. The result and ranking of the tools are summarized in Table 3-4 and Figure 3-1 It is important to note that this matrix relies solely on information found on the websites and documentation of the tools (on paper) and may not reflect the actual functionality of the tools in real life instances. In chapter 4, some tools are tested with datasets to prove their functionality.

Table 3-4- Normalizing the matrix and ranking the tools

	Open Refine	DataCleaner	SQL Power Dqguru	SQL Power Architect	CSVkit	Trifacta	Cloudingo	Microsoft DQS-data quality services	Talend Open Studio	Data Ladder	TIBCO Clarity	Validity DemandTools	Ataccama DQ ANALYZER	DATAMEER	Informatica Data Explorer	Sas data management	Pentaho Data Integration	WINpure	Experian data quality	Aggregate Profiler / osDQ	Semarchy xdm
Accuracy	84.615	76.923	30.769	50	30.769	92.308	23.077	92.308	92.308	61.538	84.615	69.231	50	76.923	92.308	92.308	42.3	61.5	46.2	76.9	76.9
Completeness	83.333	100	0	66.667	33.333	100	33.333	100	100	66.667	100	100	66.667	66.667	100	100	83.3	33.3	66.7	100	83.3
Consistency	95.455	81.818	31.818	50	27.273	90.909	27.273	90.909	90.909	63.636	81.818	54.545	59.091	81.818	90.909	90.909	36.4	72.7	54.5	72.7	77.3
Credibility	0	16.667	16.667	0	0	0	33.333	50	0	66.667	100	66.667	0	0	33.333	50	33.3	100	100	33.3	50
Currentness	75	100	25	75	25	100	0	100	100	50	50	50	50	50	100	100	25	50	25	100	75
Accessibility	0	0	0	0	0	100	100	100	100	100	100	100	100	100	100	100	100	0	100	50	100
Compliance	100	0	0	0	0	100	100	100	100	100	0	100	100	100	100	100	100	0	100	0	100
Confidentiality	50	0	0	0	0	100	100	100	100	100	50	100	100	100	100	100	100	0	100	25	100
Efficiency	100	75	75	0	25	100	25	100	100	75	100	75	62.5	100	100	100	50	75	62.5	50	75
Precision	0	0	100	100	0	100	0	100	100	0	0	0	0	100	100	0	100	0	0	0	100
Traceability	66.667	66.667	66.667	100	0	100	33.333	66.667	100	33.333	66.667	100	66.667	100	100	100	100	0	100	50	100
Understandability	100	100	50	100	0	100	0	100	100	0	50	100	100	100	100	100	100	0	100	100	100
Availability	0	0	0	0	0	100	100	100	100	100	100	100	100	100	100	100	100	0	100	50	100
Portability	66.667	75	33.333	58.333	41.667	91.667	16.667	25	91.667	66.667	95.833	25	66.667	58.333	91.667	66.667	83.3	58.3	66.7	66.7	66.7
Recoverability	100	100	100	100	0	100	100	100	100	100	100	100	100	100	100	100	100	0	100	0	100
Total average	**61.449**	**52.805**	**35.284**	**46.667**	**12.203**	**91.659**	**46.134**	**88.326**	**91.659**	**65.567**	**71.929**	**76.03**	**68.106**	**82.249**	**93.881**	**86.659**	**76.9**	**30.1**	**74.8**	**51.6**	**86.9**
	14	15	19	17	21	2	18	5	2	13	11	9	12	7	1	6	8	20	10	16	4

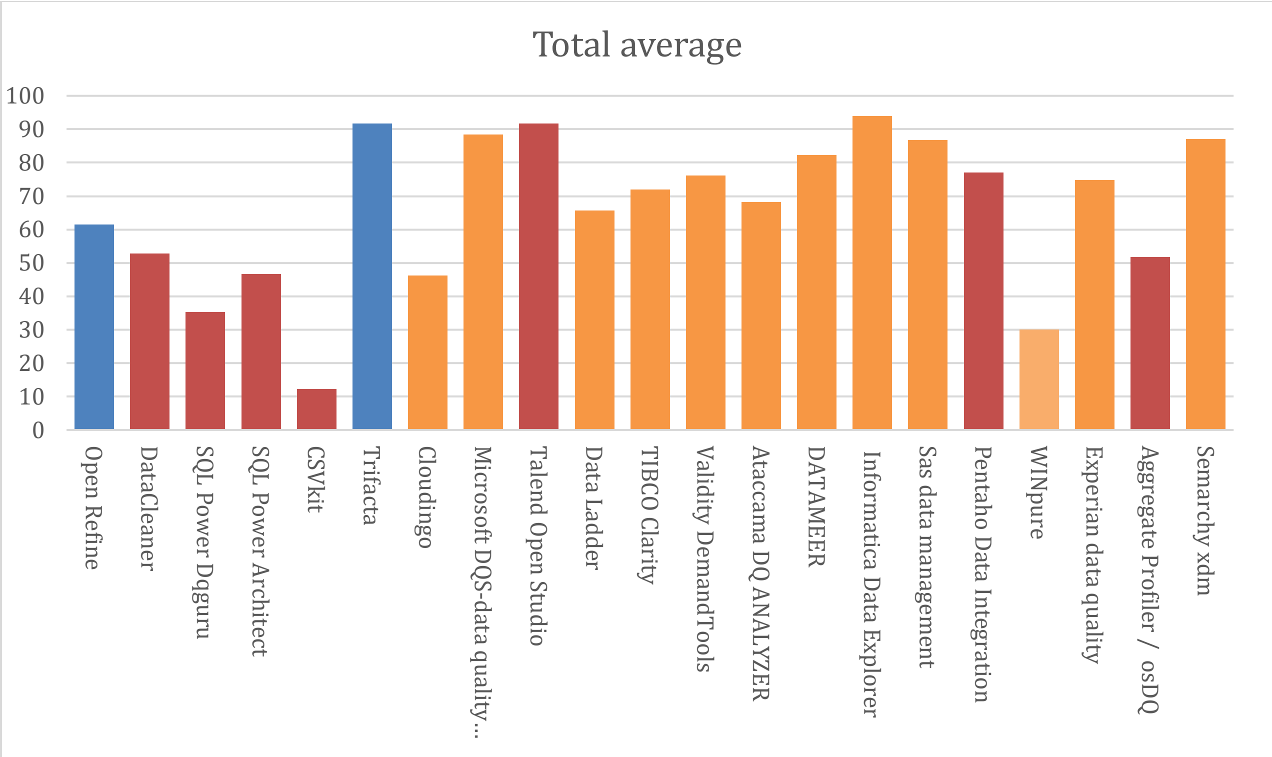

Figure 3-1-Graphical representation of the result of the comparison matrix

Selected tools for testing
Open-source tools
Closed-source tools

Chapter 4

4 Testing the data quality Tools

4.1 Method

- Three datasets (universitydata, wikidata and hosteldata) with significant number of errors were selected for the test.
- The Openrefine and Trifacta tool were selected for the test using the following criteria:
    - Tool should be free or have a free trial or demo version
    - There should be sufficient learning resources, documentation or tutorials for the tool.
- The datasets were first analysed with ad-hoc python code, to correctly identify the existing data quality problems in them. The number of errors identified are highlighted (green for duplicates and yellow highlight for other errors) in Table 4-2 for hostel dataset, Table 4-3 for wiki dataset and Table 4-1 for university dataset.
- Each data set is explored with the two tools, specifically to identify and fix the errors identified with the ad-hoc python code and the results are recorded.
- Using the applicable guidelines for measurement in the ISO 25024 [42] and considering only inherent data quality dimensions applicable to the dataset (accuracy, completeness, consistency and efficiency), the tools were evaluated.
- For the accuracy, consistency and efficiency dimensions, the number of errors remaining after we clean the dataset with the tools, are recorded in Tables 4-4, 4-5 and 4-6. On the other hand, for the completeness dimension, the number of existing records and the total number of records, before and after deduplication is recorded.

4.2 Dataset Description

For testing the tools, we will be working with open-sourced data which is unrefined and uncleaned. Open-source data is the type of data which is available for anyone to access, modify, reuse and share. They are usually derived from open-source science, hardware, government materials which are not licensed and are free to access. Three datasets were selected for the comparison of the tools. In this section, the datasets and the errors found in them are described.

4.2.1 UniversityData

The original dataset consists of 75043 rows and 10 columns (a subsection of 55 rows were used to test) of university information extracted from Wikipedia [43] with the aim of comparing the relationship between the number of students at a university and the size of the university's endowment. Upon observation of this dataset, some problems can be detected.

- Duplicate records: The dataset is heavily duplicated, having some records such as 'Washington State University' appear 320 times and 'California Institute of Technology' appear 1080 times.
- Inconsistent representation: There are some inconsistent records in the country column for example 'United States' is represented as 'USA', 'U.S.A', 'US', 'U.S.'.
- Embedded data errors: For example, in the country column, "Canada B1P 6L2", "Canada C1A 4P3 Telephone: 902-566-0439 Fax: 902-566-0795" some embedded errors can also be found in the established column "1793 as Hamilton-Oneida Academy, 1812 as Hamilton College1", etc.
- Wrong data formats: consider the established column represents dates, but they are in a string format and without a consistent datetime standard. For example, some dates are represented in yyyy-mm-dd ("1890-03-28"), some others are in the format yyyy ("1876"). The "numPostgrad", "numUndergrad", "numStudents", "numStaff", "numFaculty", "numDoctoral" and "endowment" columns are also wrongly represented as strings.
- Wrongly filled data: For example, "Some postdoctoral students and visiting scholars" in the "numGrad" column, "Day Course and Evening Course" in the numFaculty column, etc.
- Syntax errors: In the university column for example "Lumi%C3%A8re University Lyon 2", "California State University%2C Los Angeles" etc.
- The endowment column is very inconsistent. It consists of several currencies (USD, CAD$, etc). the numbers are also represented wrongly

as strings in several ways e.g. “5.00E+07”, “US $239 million”,” $44 million USD”, etc. there are also some embedded texts in the column e.g. “CHF 183 million annual budget”.

Table 4-1-Universitydata

university	endowment	numFaculty	numDoctoral	country	numStaff	established	numPostgrad	numUndergrad	numStudents
%C3%89cole Polytechnique de Montr%C3%A9al	$CAD145 million		NA	Canada	220	1873	1615	3929	
%C3%89cole Polytechnique de Montr%C3%A9al	$CAD145 million		NA	Canada	220	1873	1615	3929	
Acadia University	4.00E+07			Canada	211	1838 Queen's Co	76	2760	3485
Bowdoin College	9.04E+08	217	NA	USA	NA	1794-06-24	Some postdoc	1777	
California State University%2C Los Angeles	$19.2 million 2011	1031		United States		1947	4611	16008	
Cape Breton University	3400		not available	Canada B1P 6L2		1951	181	2987	3168
Confederation College	4700000			Canada		1967	not available	pre-university s	21160
Defiance College	$12.5 million.	86		U.S.A.		1850	100	900	1000
Durham University	¬£61.3M		NA	England		1832	4521	11278	16355
Durham University	¬£61.3M		NA	England		1832	4521	11278	One MEELL
East Carolina University	USD$130.0 millior	1804	NA	United States	5354	03/8/07	6417	20974	27816
Hamilton College	7.02E+08	219		USA		1793 as Hamilton-Oneida Acad		1812	
Idaho State University	40200750	838		United States	1269	1901	2661	12892	15553
Idaho State University	40200750	838		USA	1269	1901	2661	12892	15553
Idaho State University	40200750	838		United States	1269	1947	2661	12892	15553
Idaho State University	40200750	838		United States	1269	1901	2661	12892	15553
Lancaster University	5950000	1490	NA	England	3025	1964	3346	8780	12125
Lancaster University	5950000	1490	NA	England, UK	3025	1964	3346	8780	12125
Lumi%C3%A8re University Lyon 2	121		1355	France		1835	7046	14851	27393
Osaka University of Foreign Studies	US$ billion	Day Course and Evening (		Japan		Founded Mar. 19	N/A	N/A	
Otterbein University	US$70,025,283			United States		1847	400	2700	
Otterbein University	US$70,025,283			United States		1847	400	2700	
Paris Universitas	15	5500	8000	France		2005		25000	70000
Rocky Mountain College	16586100			United States		1878	66	878	894
Rocky Mountain College	16586100			USA		1878	66	878	894
Santa Clara University College of Arts & Sciences	$603.6 million pare	239		United States		1851	1047	2786	8846
Santa Clara University College of Arts & Sciences	$603.6 million pare	239 179		United States		1851	1047	2786	8846
Savonia University of Applied Sciences	approx. $100 million			Finland	600	provisional 1992	100	6400	4500
Savonia University of Applied Sciences	approx. $100 million			Finland	600	provisional 1992	100	6400	4500
SCU Leavey School of Business	$603.6 million pare	488268		United States		1923	1047	1491	8846
Smith College	1.43E+09	285		US		Chartered in 1871; opened its d		2600	
St. Mary's College of Maryland	U.S. $30.3 million	231		United States		1840	40	2035	
University of Central Oklahoma	1.70E+07	834		United States)		1890	1850	15251	17101
University of Delaware	$1.008 billion USD			USA	4004	1743	3634	15757	19391
University of Delaware	$1.008 billion USD			USA	4004	1743	3634	15757	19391
University of Delaware	$1.008 billion USD			USA	4004	1743	3634	15757	19391
University of Liverpool	1.21E+08			England, UK		1881 - University	3860	16805	20655
University of Michigan	US $6.56 billion	6238	NA	US	18426	1817	15309	26208	41674
University of Milan	562000000	4210		Italy	2455	1924	4354	49476	62801
University of Minnesota	US$2.224 billion ir	3374		United States		1851	16948	30375	51611
University of Minnesota	US$2.224 billion ir	3374		United States		1851	16948	30375	51611
University of Minnesota	US$2.224 billion ir	3391		United States		1851	16948	30375	51611
University of North Carolina at Charlotte	US$105.9 million	1280	817	U.S.		1961	4994	20283	25063
University of North Carolina at Charlotte	US$140.9 million	1280	817	U.S.		1961	5308	19755	25277
University of Northern Iowa	$65.8 M http://www	800	NA	United States		1876	1933	11147	
University of Prince Edward Island	2.00E+07	250		Canada C1A 4P3	Telephon	1969	324	4276	4600
University of St. Gallen	CHF 183 million ai	1325	782	Switzerland	285	1898-05-25	3043	3656	6726
University of the Philippines Los Ba%C3%B1os	‚Ç±4.46 billion	933		Philippines		03/6/09	1305	10756	10688
University of the Philippines Los Ba%C3%B1os	‚Ç±4.46 billion	817		Philippines		03/6/09	1305	10756	12557
University of the Philippines Los Ba%C3%B1os	‚Ç±4.46 billion	933		Philippines		03/6/09	1305	10756	12557
University of Toronto	C$1.518 billion	2551		Canada	4795	1827-03-15	12732	43141	
University of Utah	US$513.4 million	2687		United States	14362	1850-02-28	7448	23371	30819
Washington State University	6.20E+08	1304	611	U.S.		1890-03-28	2241	15380	-18,234
Washington State University	6.20E+08	1304	611	U.S.		1890-03-28	2241	15380	21016

4.2.2 Hostel data

The dataset describes hostel type accommodation in Torino [44]. It provides information on their locations, contact details, proximity to places such as metro stations, prices, etc. It consists of 221 columns and 51 rows (a subsection of 13 columns were used to test). Some errors identified in the dataset include:

- Inaccurate data: on the ‘cap’ column, all records have the zip code ‘10100’. Considering different addresses in different areas, the zip codes should be different for some records. Because the dataset is relatively small, the correct data was obtained from google maps and reserved in another file, for cross table analysis, record matching or reconciliation.

- Inconsistency: for example, on the 'DistanzeNomeStazioneFerroviaria' column, "PORTA SUSA", "fs porta susa", "FS Porta Susa" can be identified to all represent the porta susa station. The unit to measure distance on the 'DistanzeParcheggioEsternoM' column and the 'DistanzeStazioneFerroviariaKm' column are not in a uniform standard. For example, '700 m' instead of '0.7 km' or 'km 3,7' instead of '3.7 km'. Moreover, the presence of the units embedded within the column is going to be problematic for any data analysis. Ideally, all records should be converted to a single unit which should be indicated at the top of the column.

Table 4-2- Hostel data

Provincia	Comune	Cap	DenominazioneStruttura	Indirizzo	NroCivico	Telefono	RecapitiFax	EMail	SitoWeb	DistanzeParcheggioEsternoM	DistanzeStazioneFerroviariaKm	DistanzeNomeStazioneFerroviaria
TORINO	TORINO	10100	BUENA VISTA	Via Giordano Bruno	191	3914089452-0112386330		info@buenavista.torino.it	www.acmos.net	0 m		
TORINO	TORINO	10100	CASA IN CENTRO	San Domenico	13/I	3290552565	114319268	casaincentro@coopaccomazzi.it	www.coopaccomazzi.it	15 mt	1 Km	PORTA SUSA
TORINO	TORINO	10100	CASA OASI	Via Capriolo Luigi	18	0113835245-3371320952	113802905	casa.oasi@gruppoarco.org	www.gruppoarco.org/casaoasi		2 km	PORTA SUSA
TORINO	TORINO	10100	CASA SANT'ANNA	Via Massena Andrea	36	0115166532-3317049877	115166599	casasantanna.to@istituto-santanna.it		500 mt	0,6 km	PORTA NUOVA
TORINO	TORINO	10100	CIVIVO 15	Via Cottolengo	15	3429924123		civivo15@providencehouse.it		50	2,4	fs porta susa
TORINO	TORINO	10100	COLLEGIO UNIVERSITARIO R. EINAUDI - SE	VIA DELLE ROSINE	3	118126856	118171008	concorsi@collegioeinaudi.it	www.collegioeinaudi.com		1,9 km	Porta Nuova
TORINO	TORINO	10100	COLLEGIO UNIVERSITARIO R. EINAUDI - SE	Via Maria Vittoria	39	118126853	118171008	concorsi@collegioeinaudi.it	www.collegioeinaudi.it		1,9 km	Porta Nuova
TORINO	TORINO	10100	COLLEGIO UNIVERSITARIO R. EINAUDI - SE	Via Bobbio	3	113851944	118171008	concorsi@collegioeinaudi.it	www.collegioeinaudi.it		3,3 km	FS porta nuova
TORINO	TORINO	10100	DON BOSCO YOUTH HOUSE	Corso Unione Sovietica	312	116198311	116198421	economo@agnelli.it	www.agnelli.it	10	4	Lingotto
TORINO	TORINO	10100	FOYER - YWCA UCDG	Via San Secondo	70	0115683369-0115819571	115131427	segreteria@ywcaitalia.it			1 Km	PORTA NUOVA
TORINO	TORINO	10100	FRATERNITA'	Via Lanfranchi Francesco	16	0118192658-3358091345		fratuniversitarie@gmail.com		0 m	1,5 Km	PORTA NUOVA / SUSA
TORINO	TORINO	10100	ISTITUTO ALFIERI - CARRU'	Via Accademia Albertina	14	118395391	118395391	amministrazione@istitutoalfiericarru	www.istitutoalfiericarru.it	100 mt		
TORINO	TORINO	10100	ISTITUTO SUORE SAN GIUSEPPE	Via Giolitti Giovanni	29	118177874	118123466	istitsg_ferie@yahoo.it		200 mt	0,8 Km	PORTA NUOVA
TORINO	TORINO	10100	OASI MARIA CONSOLATA	Via Santa Lucia	89/97	116612300	116612300	oasicavoretto@gruppoabele.org	www.oasicavoretto.it	300 mt	6 Km	PORTA NUOVA
TORINO	TORINO	10100	OPEN 011 - CASA DELLA MOBILITA' GIOVA	Corso Venezia	11	11250535	112215919	info@open011.it	www.open011.it	0 m	4 km	FS Porta Susa
TORINO	TORINO	10100	OSTELLO DELL'ANTICA ABBADIA	Strada Comunale Cascinotto	59	112730972		ostello.abbadia@virgilio.it	www.ostelloanticaabbadia.it	10		
TORINO	TORINO	10100	PENSIONATO LAVORATORI LA SALETTE	Via Maddona Della Salette	20	3663573585	11710753	torinolasalette@gmail.com		10	3,5 Km	PORTA SUSA
TORINO	TORINO	10100	PENSIONATO REBADUE	Piazza Rebaudengo Conti D	22	112429711	112429799	economato@rebanet.it	www.rebanet.it	20 mt	6 km	PORTA SUSA
TORINO	TORINO	10100	PENSIONATO ROSA GOVONE	Via Delle Rosine	7	3420184774	114319268	casagovone@coopaccomazzi.it	www.coopaccomazzi.it	50 mt	2,5 km	PORTA NUOVA / SUSA
TORINO	TORINO	10100	PENSIONATO UNIVERSITARIO ARTIGIANEL	Corso Palestro	14	3484008019	115625824	universitari@educarecoop.org		0	0,8	Fs Porta Susa
TORINO	TORINO	10100	PENSIONATO UNIVERSITARIO SALESIANO	Via Maria Ausiliatrice	36	115224822	115224395	cus.valdocco@31gennaio.net		30 mt	2 Km	PORTA SUSA
TORINO	TORINO	10100	ATTIC HOSTEL TORINO	PIAZZA PALEOCAPA	2	1119704651	1119704651	info@attichostel.it	www.attichostel.it	10 MT	0,2 KM	FS PORTA NUOVA
TORINO	TORINO	10100	BAMBOO ECO HOSTEL	CORSO PALERMO	90/D	11235084		info@bambooecohostel.it	www.bambooecohostel.it		4 km	PORTA NUOVA / SUSA
TORINO	TORINO	10100	CAMPLUS LINGOTTO	Via Nizza	230	116939393	116939350	lingotto.guest@camplus.it	www.camplusguest.it	10 mt	5 km	PORTA NUOVA
TORINO	TORINO	10100	CAMPUS SAN PAOLO	Via Caraglio	97	113828416	115175486		www.campussanpaolo.it	15 mt	3,2 km	FS Porta Susa
TORINO	TORINO	10100	CASA DELLA GIOVANE	Via C. I. Giulio	8	114362681	114390169	toconsolata@fma-ipi.it		50 mt	2 Km	PORTA NUOVA
TORINO	TORINO	10100	CASA ENRICHETTA DOMINICI	Via Massena Andrea	34	0115166532-3317049877	115166599	casasantanna.to@istituto-santanna.it		500 mt	0,6 km	PORTA NUOVA
TORINO	TORINO	10100	CASA FEMMINILE VALDESE	Via San Pio V	15	116692838	1119834888	csd.casa.femminile@tiscali.it	www.torinovaldese.org		0,2 km	PORTA NUOVA
TORINO	TORINO	10100	CASA MAMMA MARGHERITA	Via Maria Ausilliatrice	9	115224201	115224680	accoglienza@valdocco.it	www.accoglienza.valdocco.i	50 mt	1 Km	PORTA SUSA
TORINO	TORINO	10100	CENTRO FORMATIVO ONAOSI	Via Della Basilica	4	115290500	115290510	cftorino@onaosi.it				
TORINO	TORINO	10100	CENTRO PUZZLE	VIA CIMABUE	2	113119900	113010078	info@centropuzzle.it	www.centropuzzle.org	10 mt	6,5 Km	PORTA NUOVA
TORINO	TORINO	10100	COLLEGIO UNIVERSITARIO R. EINAUDI - SE	CORSO LIONE	24	113851922	118171008	concorsi@collegioeinaudi.it	www.collegioeinaudi.it		3,2 km	Porta Nuova
TORINO	TORINO	10100	COLLEGIO UNIVERSITARIO R. EINAUDI-SEZ	VIA GALLIARI BERNAR	30	114222505	118171008	concorsi@collegioeinaudi.it	www.collegioeinaudi.it		0,85 km	FS Porta Nuova
TORINO	TORINO	10100	COLLEGIO UNIVERSITARIO SAN GIOVANNI	Via Madama Cristina	1	1119839492	110703992	cus.sangiovanni@31gennaio.net		100	0,5 km	PORTA NUOVA
TORINO	TORINO	10100	COLLEGIUM TRINITATIS	VICOLO CROCETTA	5/A	110810354		info@collegiumtrinitatis.it	www.trini.to.it	30 MT	1,7 KM	FS PORTA NUOVA
TORINO	TORINO	10100	CONVITTO PER STUDENTI E LAVORATORI (	VIA SPOLETO	9	0115690832-0115212812	117935870	orionecooperativa@libero.it	www.camereaffitto.it		3 Km	PORTA SUSA
TORINO	TORINO	10100	CONVITTO SAN SALVARIO	Via Saluzzo	58	116694728		contatti@sport-residence.it	www.sportresidence.com	100 mt	1 km	FS
TORINO	TORINO	10100	DELLA BARCA	Strada Comunale Cascinotto	59	112730972		ostello.abbadia@virgilio.it	www.ostelloanticaabbadia.it	10		
TORINO	TORINO	10100	D'ORHO-DON ORIONE HOUSING	Corso Principe Oddone	22	3883254331		dorho.torino@gmail.com			1 km	PORTA SUSA
TORINO	TORINO	10100	GIOVANNA D'ARCO VARSITY HOUSE	Via Pomba Giuseppe	21	110208430		segreteria@residenzagiovannadarcc	www.residenzagiovannadarco.it		700 m	FS Porta Nuova
TORINO	TORINO	10100	LINK HOUSE	STATI UNITI	11/H	115631562	110700808	segreteria@cooperativaparadigma.it	www.cooperativaparadigma.	1	0,5	FS Porta Nuova
TORINO	TORINO	10100	MICHELE MAGONE	VIA SALERNO	12	115224279		casamichelemagone@gmail.com		30 m	2 km	PORTA SUSA
TORINO	TORINO	10100	OSPITERIA DELL'ARSENALE DELLA PACE	Via Andreis Vittorio	18/27	114368566	115215571	ospiteria@sermig.org	www.sermig.org/ospiteria	20 mt	3 Km	PORTA SUSA
TORINO	TORINO	10100	PENSIONATO MADRE CABRINI	Via Tarino Luigi	1	11835858	11835859	msccabrini.to@libero.it		10 mt	2 km	PORTA NUOVA
TORINO	TORINO	10100	PENSIONATO REBAUNO	Piazza Rebaudengo Conti D	22	112429711	112429799	economato@rebanet.it	www.rebanet.it	20 mt	6 km	PORTA SUSA
TORINO	TORINO	10100	PENSIONATO UNIVERSITARIO VALDOCCO	Via Maria Ausiliatrice	36	115224822	115224395	cus.valdocco@31gennaio.net		30 m	2 km	PORTA SUSA
TORINO	TORINO	10100	RESIDENZA UNIVERSITARIA CARLO MOLLI	Corso Peschiera	90	01119752000-346076321	119752171	residenza.mollino@camplus.it	www.camplusapartments.it	5 MT	2 KM	PORTA SUSA
TORINO	TORINO	10100	RESIDENZA VALPIANA	Strada Valpiana	31	118998555	118998555	mail@fondazione-df.com	www.fondazione-df.com		km 3,7	FS Porta Nuova
TORINO	TORINO	10100	SGUARDO SU TORINO	Via Capriolo	18	0113835245-3371320952	113802905	sguardosutorino@gruppoarco.org	www.gruppoarco.org/sguardosutorino		2 km	PORTA SUSA
TORINO	TORINO	10100	WINS BOARDING	VIA TRAVES	28	111972111	111972150	info@worldinternationalschool.com	www.worldinternationalscho	250 MT	5,5 KM	GTT DORA

WikiDataset: This dataset was scrapped from Wikipedia by [45]. It consists of a list of countries, their population, % of world population, Total Area, Percentage Water, Total Nominal GDP and Per Capita GDP. The dataset has 7 rows and 197 columns (a sub section of 30 rows were used to test). The dataset is very messy with a lot of unnecessary embedded data as shown in Table 4-3.

Table 4-3- WikiDataset.csv

Country(or dependent terr	Population	% of worldpopulation	Total Area	Percentage Water	Total Nominal GDP	Per Capita GDP
China[Note 2]	1,394,350,000	18.20%	9,596,961¬†km2 (3,705,407¬†sq¬†mi)[g] (3rd/4th)	2.8%[h]	$14.092 trillion[16] (2nd)	$10,087[16] (71st)
India[Note 3]	1,337,630,000	17.50%	3,287,263[5]¬†km2 (1,269,219¬†sq¬†mi)[d] (7th)	9.6	$2.848 trillion[16] (6th)	$2,134[16] (133rd)
United States[Note 4]	327,918,000	4.28%	3,796,742¬†sq¬†mi (9,833,520¬†km2)[8] (3rd/4th)	6.97	$19.390 trillion[11] (1st)	$59,501[11] (7th)
Brazil	209,650,000	2.74%	8,515,767¬†km2 (3,287,956¬†sq¬†mi) (5th)	0.65	$2.139 trillion[7] (9th)	$10,224[7] (65th)
Pakistan	202,169,000	2.64%	881,913¬†km2 (340,509¬†sq¬†mi)[a][18] (33rd)	2.86	$304.4 billion[21] (42nd)	$1,629[22] (145th)
Nigeria	193,392,517	2.53%	923,768¬†km2 (356,669¬†sq¬†mi) (32nd)	1.4	$376.28 billion[3] (31st)	$1,994[3] (137th)
Bangladesh	165,278,000	2.16%	147,570[5]¬†km2 (56,980¬†sq¬†mi) (92nd)	6.4	$285.817 billion[8] (43rd)	$1,754[8] (148th)
Russia[Note 5]	146,877,088	1.92%	17,098,246¬†km2 (6,601,670¬†sq¬†mi)[5] (without Crimea)[n	13[7]¬†(including swamps)	$1.719 trillion[9] (12th)	$11,946[9] (67th)
Japan	126,420,000	1.65%	377,973.89[9]¬†km2 (145,936.53¬†sq¬†mi)[10] (61st)	0.8	$5.167 trillion[12] (3rd)	$40,849[12] (20th)
Mexico	124,737,789	1.63%	1,972,550¬†km2 (761,610¬†sq¬†mi) (13th)	2.5	$1.250 trillion[6] (16th)	$10,021[6] (69th)
Ethiopia	107,534,882	1.40%	1,104,300¬†km2 (426,400¬†sq¬†mi) (26th)	0.7	$85.664 billion[5]	$910[5]
Philippines	106,540,000	1.39%	300,000[4][5]¬†km2 (120,000¬†sq¬†mi) (63rd)	0.61[6]¬†(inland waters)	$371.8 billion[8]	$3,541[8]
Egypt	97,639,400	1.28%	1,010,408[2]¬†km2 (390,121¬†sq¬†mi) (29th)	0.632	$237.073 billion[4] (49th)	$2,501[4] (113th)
Vietnam	94,660,000	1.24%	331,698[4]¬†km2 (128,069¬†sq¬†mi) (65th)	6.4[5]	$240.779 billion[7] (47th)	$2,546[7] (129th)
DR Congo	84,004,989	1.10%	2,345,409¬†km2 (905,567¬†sq¬†mi) (11th)	3.32	$40.415 billion[3]	$446[3]
Germany	82,792,400	1.08%	357,386¬†km2 (137,988¬†sq¬†mi)[4] (62nd)	82,800,000[5] (16th)	$3.685 trillion[6] (5th)	$44,550[6] (17th)
Iran	81,830,600	1.07%	1,648,195¬†km2 (636,372¬†sq¬†mi) (17th)	7.07	$438.3 billion[8] (27th)	$5,383[8]
Turkey	80,810,525	1.06%	783,356¬†km2 (302,455¬†sq¬†mi) (36th)	1.3	$909 billion[4] (17th)	$11,114[4] (60th)
Thailand	69,183,173	0.90%	513,120¬†km2 (198,120¬†sq¬†mi) (50th)	0.4 (2,230 km2)	$514.700 billion[11]	$7,588[11]
France[Note 6]	67,323,000	0.88%	640,679¬†km2 (247,368¬†sq¬†mi)[3] (42nd)	551,695¬†km2 (213,011¬†sq	$2.583 trillion[7] (7th)	$39,869[7] (22nd)
United Kingdom[Note 7]	66,040,229	0.86%	242,495¬†km2 (93,628¬†sq¬†mi)[7] (78th)	1.34	$2.624¬†trillion[10] (5th)	$39,734[10] (19th)
Italy	60,421,460	0.79%	301,340¬†km2 (116,350¬†sq¬†mi) (71st)	2.4	$2.181 trillion[5] (8th)	$35,913[4] (25th)
South Africa	57,725,600	0.75%	1,221,037¬†km2 (471,445¬†sq¬†mi) (24th)	0.38	$371 billion[6] (35th)	$6,459[6] (88th)
Tanzania[Note 8]	54,199,163	0.71%	947,303¬†km2 (365,756¬†sq¬†mi) (31st)	6.4[6]	$55.666 billion[9]	$1,100[9]
Myanmar	53,862,731	0.70%	676,578¬†km2 (261,228¬†sq¬†mi) (39th)	3.06	$69,322 billion[5] (70th)	$1,299[5] (152nd)
Georgia[Note 15]	3,729,600	0.05%	69,700¬†km2 (26,900¬†sq¬†mi) (119th)	3,718,200[a][5] (131st)	$15.23 billion[7] (116th)	$4,370[8] (112th)
Slovenia	2,066,880	0.03%	20,273¬†km2 (7,827¬†sq¬†mi) (151st)	0.7[6]	$56.933¬†billion[9]	$27,535[9] (32nd)
Latvia	1,923,400	0.03%	64,589¬†km2 (24,938¬†sq¬†mi) (122nd)	1.57% (1,014 km2)	$30.176 billion[6]	$18,472[6]
Kosovo[Note 17]	1,798,506	0.02%	10,908¬†km2 (4,212¬†sq¬†mi)	1.0[2]	$7.73 billion[4]	$4,140[5]
Guinea-Bissau	1,584,763	0.02%	36,125¬†km2 (13,948¬†sq¬†mi) (134th)	22.4	$1.295 billion[3]	$761[3]

4.3 Working with OpenRefine

4.3.1 UniversityData

To identify the missing values, the ‘facet blank value per column’ features was applied, resulting 102 blank records in total (Fig 4-1). After deduplication, the number of blank records is a total of 80. This method was unable to identify “NA” values as missing values.

To identify the duplicate data, the ‘university’ column was reordered in alphabetical order, to have rows with similar text clustered together. Then the blank down feature is applied identifying 18 duplicate data which are all completely removed with the ‘remove matching rows’ feature (Fig 4-2).

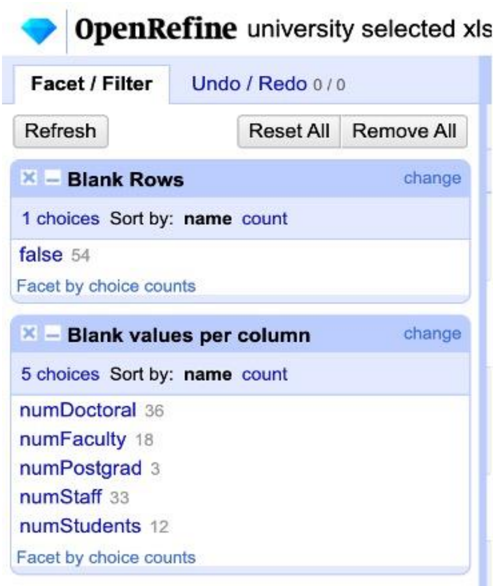

Figure 4-2-Identifying missing values in the university dataset

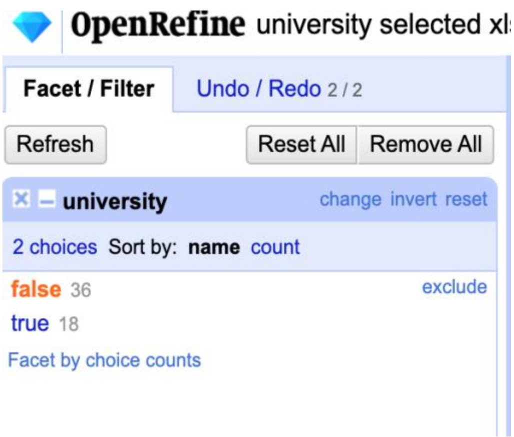

Figure 4-1-identifying duplicate data in university dataset

To fix the accuracy problems in the university column the unescaped('url') was used to remove illegal and reserved characters within the text (Fig 4-3). Excluding duplicate data, 4 records were fixed.

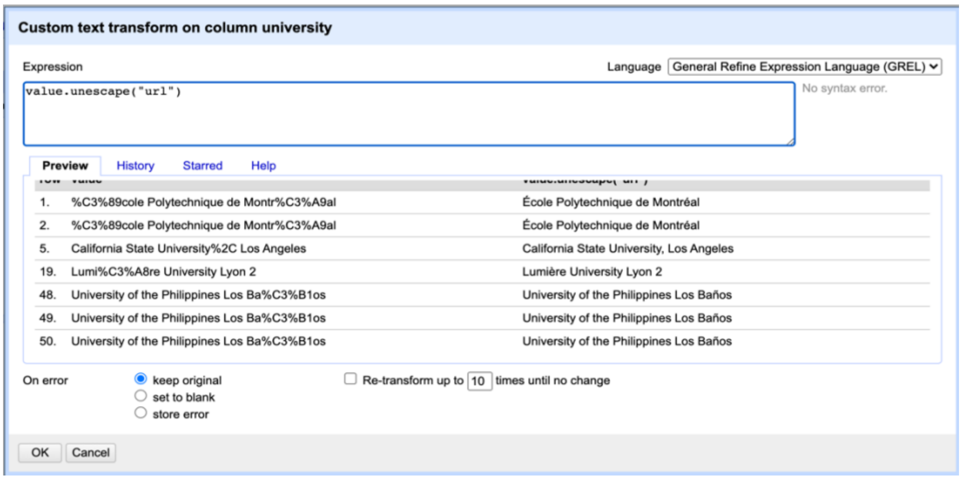

Figure 4-3-Removing syntax error in the university dataset

The non-uniform representation of some records in the 'country' column was identified, and mass edited with the text facet feature as shown in Fig 4-4. The text facet allows for strings with similar values to be clustered together and mass edited. It has 6 algorithms for clustering: fingerprint, ngram-fingerprint, metaphone3, cologne-phonetic, levenshtein and PPM.

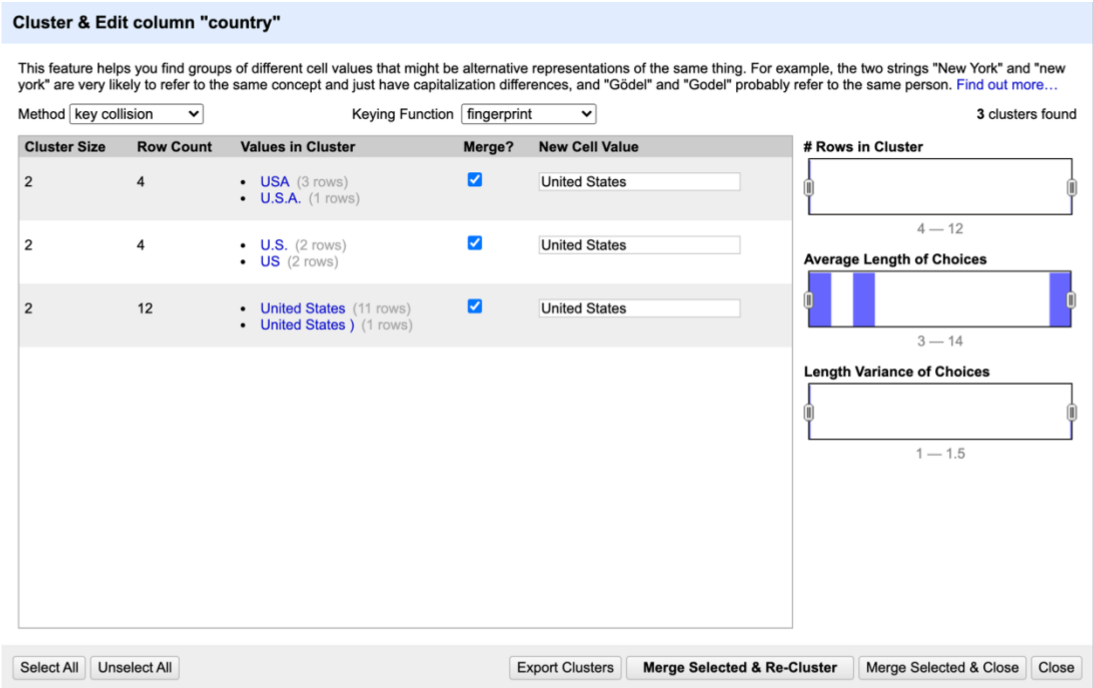

Figure 4-4- Identifying the different representations of the United states

Fixing the endowment column required using the 'Transform' function and some coding to eliminate non numerical data Fig 4-5. The 'million' and 'billion' text were converted to numerical form by multiplying the values by 10^6 and 10^9 respectively Fig 4-5. Then the whole column was then transformed into number format using "common transforms". Similar functions were applied to the 'establishment' column, then the column was converted to date format.

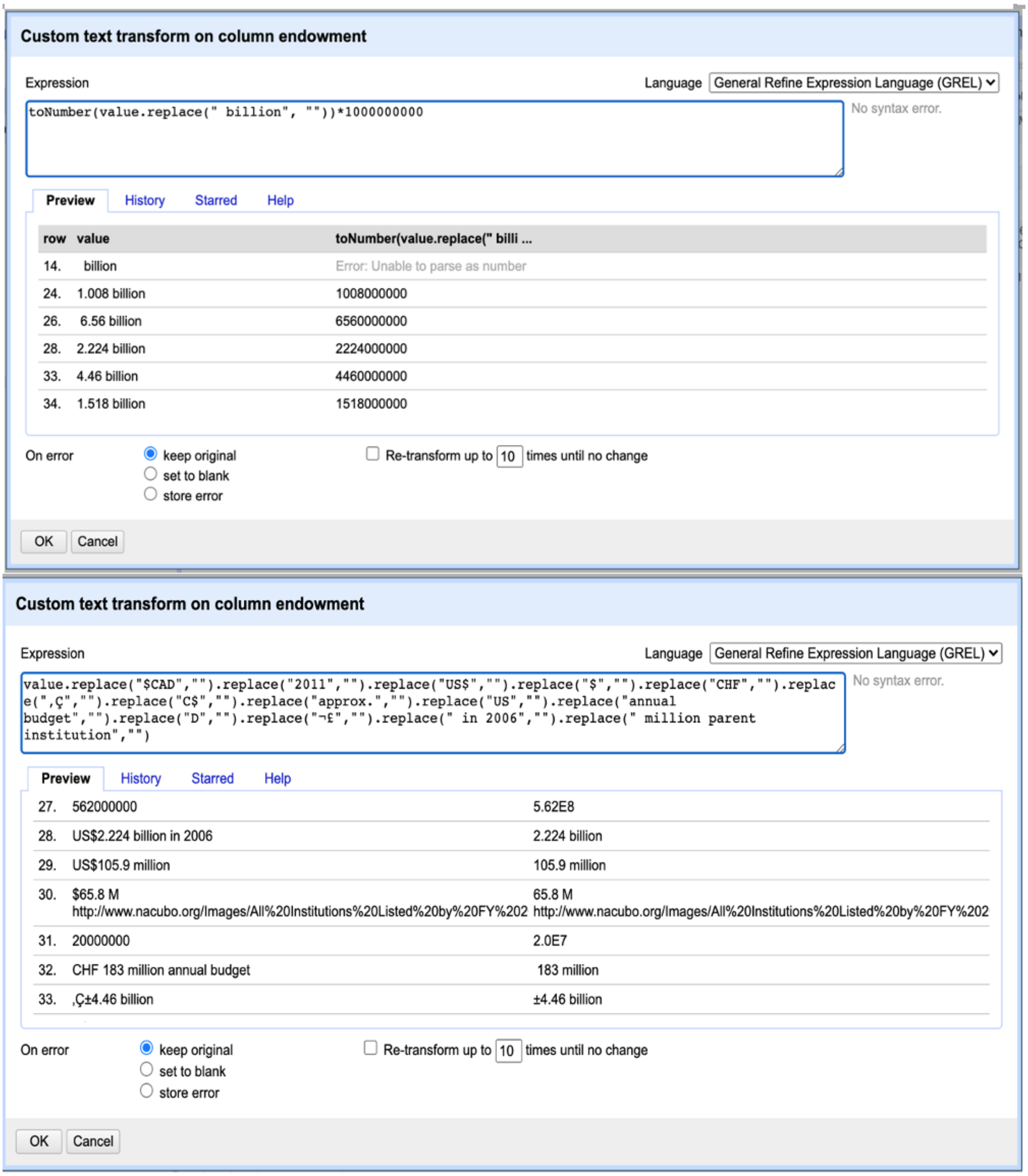

Figure 4-5- Formatting the endowment column

The ‘numPostgrad’, ‘numUndergrad’ and ‘numStudents’ columns were transformed with the ‘common transforms’ function, into numbers and the non-numeric records were set to blank Fig 4-6.

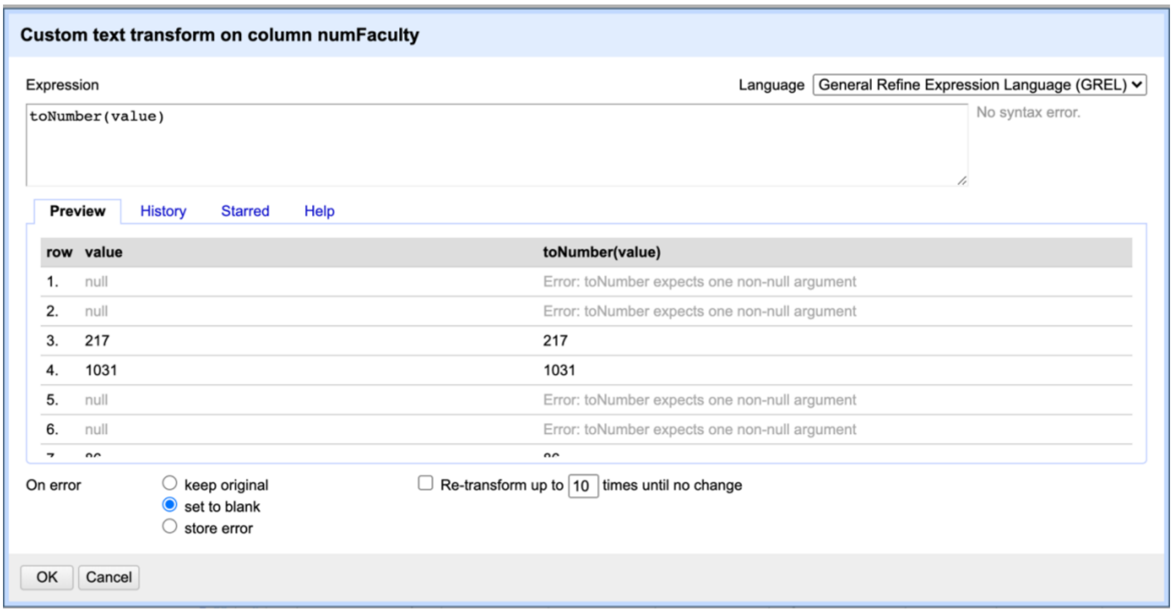

Figure 4-6-Transforming columns to number format

4.3.2 Hostel data

Using the same methods as in the previous datasets, the null values were identified as 53 with 0 completely empty rows (Fig 4-7)

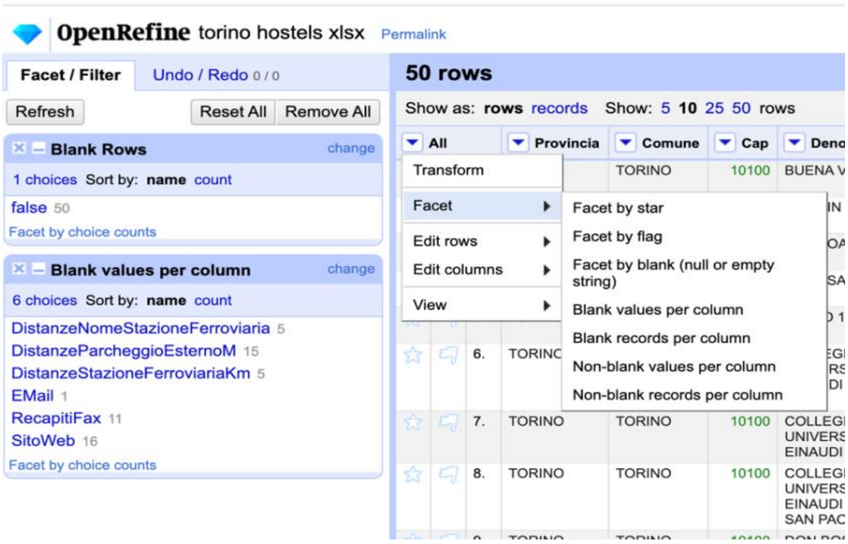

Figure 4-7 Identifying missing values in the hostel data

To correct the inaccurate “cap” column, both the inbuilt reconciling feature and the external source reconciling feature was used but none were effective. The external source reconciling feature could not accurately match one zip code to one address. Fig 4-8. Finally, the cross-table join was used to match the similar columns and extracting the relevant column from the local dataset with accurate values. Based on the result, a new column was created Fig 4-9.

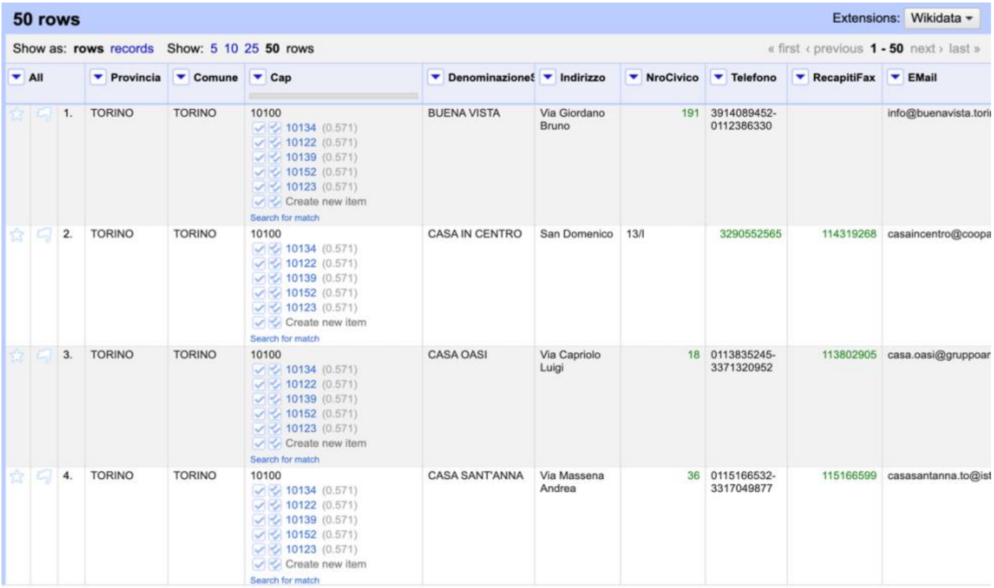

Figure 4-8-Reconciling the Cap column against a local dataset

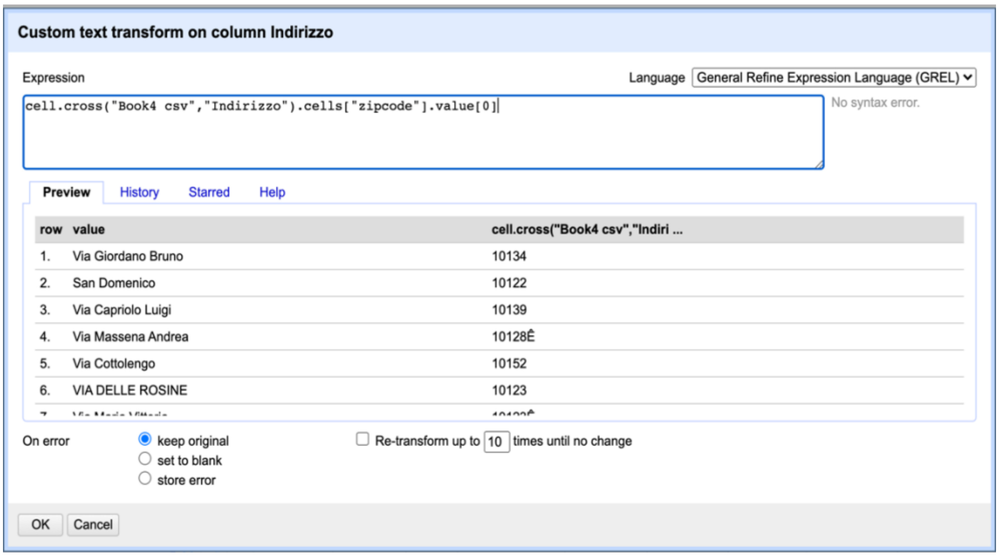

Figure 4-9-Matching the cap column with the join feature

A new column was created based on the 'Telefono' column to extract multiple phone number entries leaving only a single phone number entry in each column. As shown below:

"Create new column phone number 2 based on column Telefono by filling 9 rows with grel:value.split("-")[1]"

The text facet feature was applied to the column "DistanzeNomeStazioneFerroviaria" to mass edit the different representations of 'Porta susa' and 'Porta Nuova' to a uniform format. (25 records were affected) Shown in Fig 4-10.

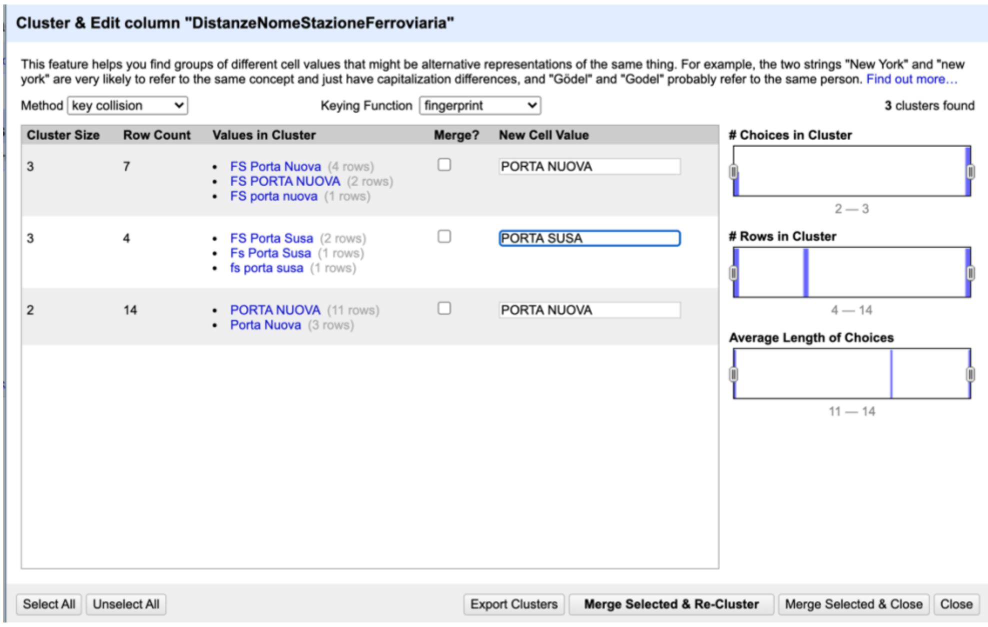

Figure 4-10-Formatting the DistanzeNomeStazioneFerroviaria column

The 'DistanzeParcheggioEsternoM' and the 'DistanzeStazioneFerroviaria' columns were formatted with similar methods used for the numeric data in the University dataset (Fig 4-6).

4.3.3 WikiDataset.csv

No blank, null or empty strings were found in this dataset. No duplicate records were found as well.

To format and process this dataset, regular expressions were applied using the GREL coding feature of open refine across each column. This process is not automatic and heavily relies on the skills of the person working with the data (Fig 4-11).

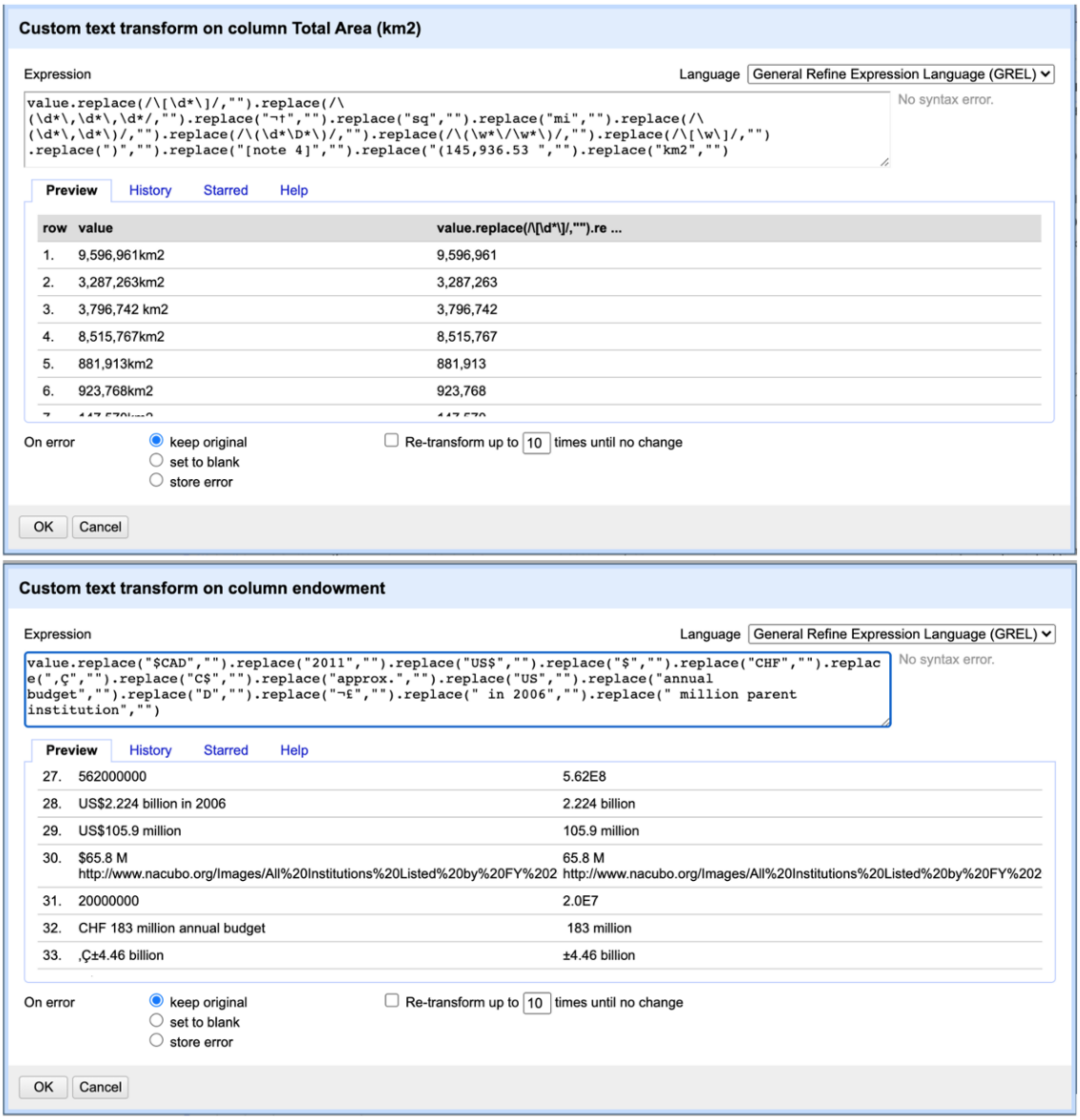

Figure 4-11-Formatting the total area Coolum with regular expressions

Similar methods (as seen in fig4-5) were used to format the "Total Nominal GDP($)" and other columns with numeric data including "population", "% of world population", "Total Area", "Percentage Water", and "Per Capita GDP" columns.

4.4 Working with Trifacta

4.4.1 University.csv

To identify and delete the duplicate data with Trifacta, the inbuilt "Remove duplicate rows" transformation was applied but it was not effective in identifying any of the duplicate records in the data. A different method was used [46];

- Creating a new primary key column (merging the university and established columns).
- Ordering the dataset by the new primary key column.
- Creating a new window to compare the records in the primary key column.
- Creating a new column 'isdupe', which represents match or not matched by true or false.

  IF((window==PrimaryKey), true, false)
- Deleting the "true" rows.

With this method, 17 duplicates were deleted (Fig 4-12).

Trifacta has a data quality bar and a column view that can be used to profile data effectively. They are able to identify category counts, unique values, missing values, etc. From the column view, 70 missing values and 21 mismatched values (Fig 4-13).

There are no direct methods to remove the syntax errors in the university column.

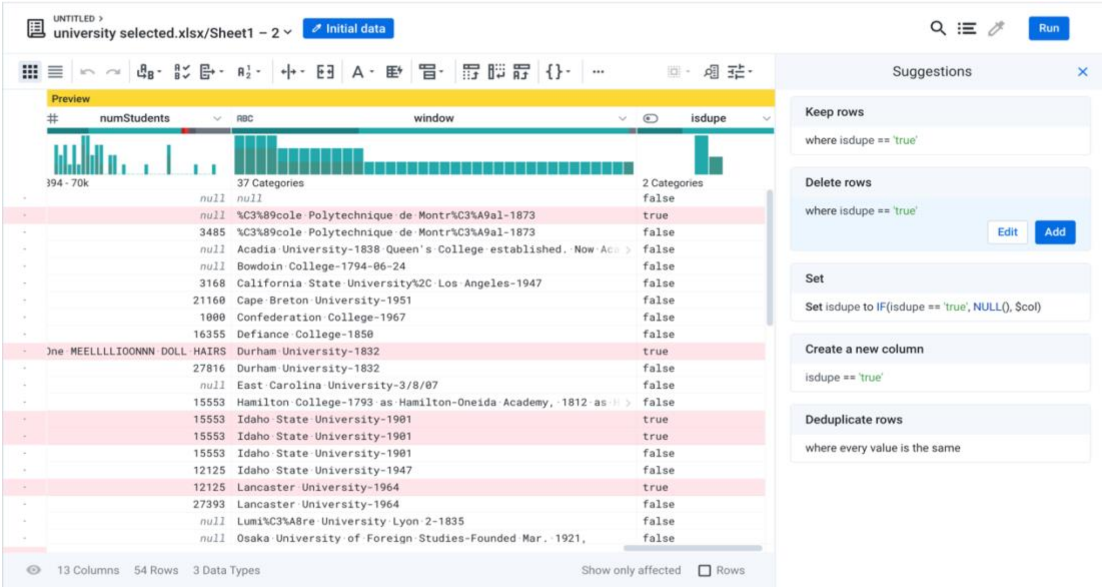

Figure 4-11-Removing duplicate data with Trifacta

To format the endowment column, all non-numerical values (except ‘million’ and ‘billion’) were replaced with “”. The column was then split with a space delimiter, separating the numeric values and the non-numeric values into different columns fig 4-14. In the new ‘endowment 2’ column, ‘million’ is replaced with 1000000 and ‘billion’ is replaced with 1000000000. The two endowment columns are multiplied, and the results are inserted into a new “multiplied” column. The null values in the ‘multiplied’ column are set to the corresponding values in the “endowment 1” column. It is important to note here that there were no direct features to achieve this formatting and the coding flexibility is strictly limited to the functions offered by Trifacta. This makes the cleaning of similar dirty data types highly reliant on the skills of the person working with the tool. Similar methods were applied to the numeric data in exp format (e.g., 1.43E+09) Fig 4-14.

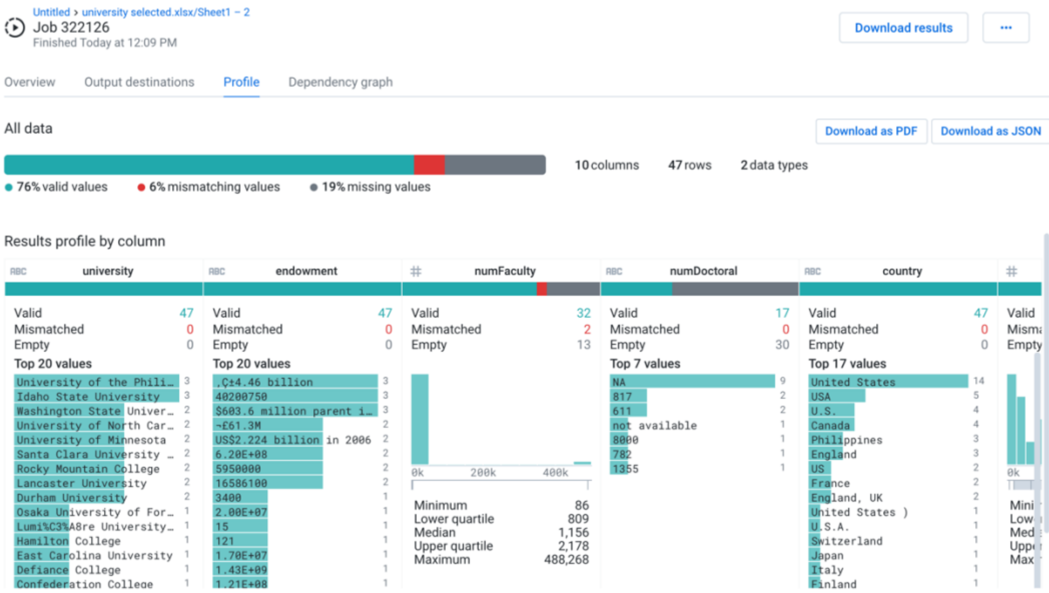

Figure 4-12-The profiling feature of Trifacta

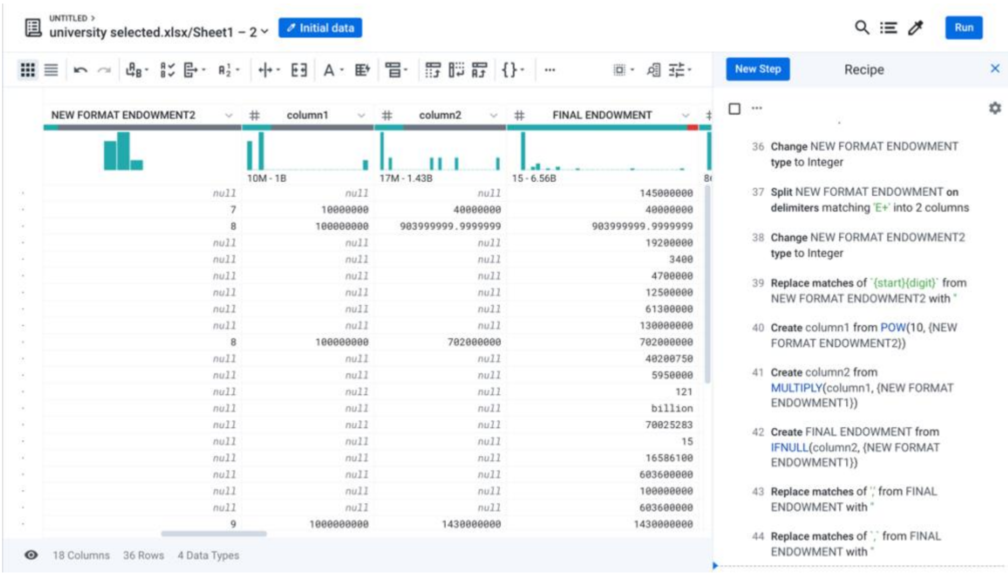

Figure 4-13-Formatting the endowment column with Trifacta

The country column was formatted using the pattern recognition of Trifacta. It identifies all the patterns within the column, then mass edits (usually suggested by Trifacta) can be carried out on the clusters of the patterns (Fig 4-15).

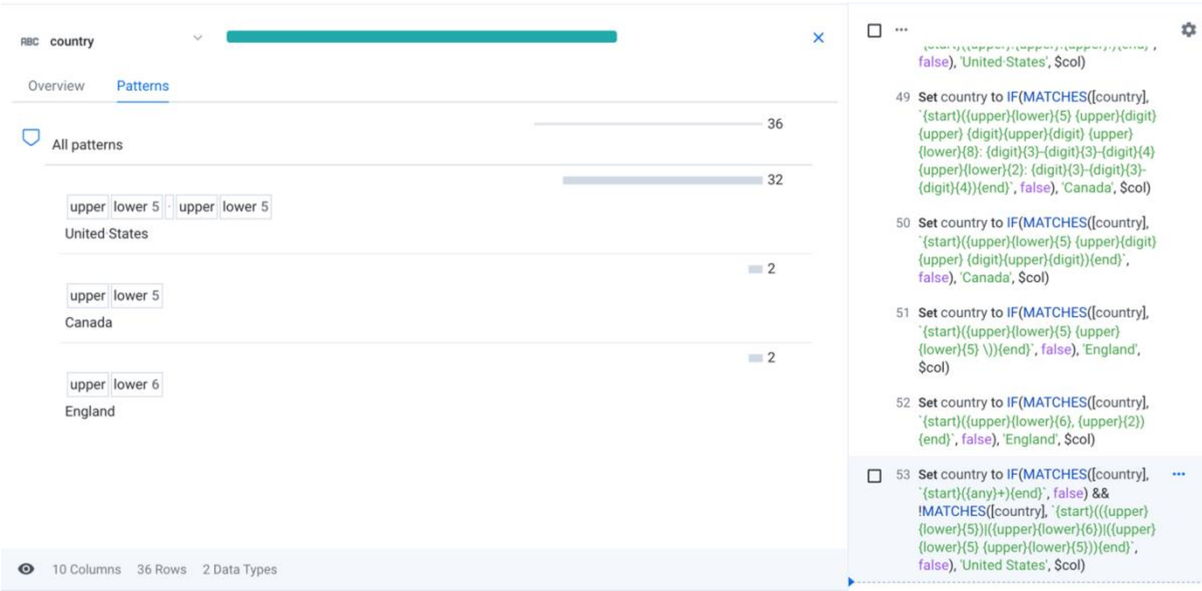

Figure 4-14- Formatting the country column with Trifacta pattern recognition

The established column was first converted to string, then the not "yyyy" format was selected in one record. From the Trifacta suggestions tab, a replace match function is used to mass format similar patterned records (Fig 4-16).

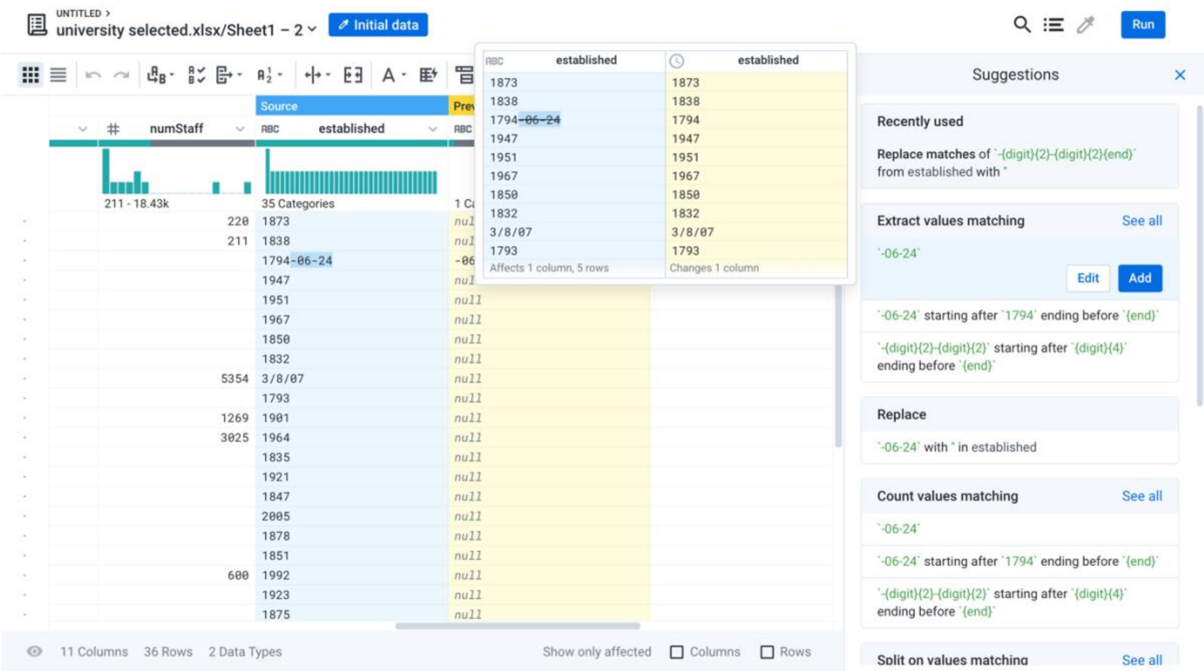
Figure 4-15-Formatting the established column with Trifacta from the suggestions tab

The 'numPostgrad', 'numUndergrad' and 'numStudents' columns were formatted by clicking on the identified mismatched values, and editing for each case (e.g., converting N/A to blank)

4.4.2 Hostel data

The column view of Trifacta indicates 599 valid data, 15 mismatched values (6 in "NroCivico" and 9 in 'telefono') and 53 missing values. The embedded data in all the columns were easier to clean on Trifacta because of the automatic pattern recognition and the suggestions when any record on the dataset is highlighted (Fig 4-16). Trifacta is also able to recognize and profile the data in more formats. For example, the "SitoWeb" and "Email" columns were automatically identified as URL and email address respectively.

The incorrect "cap" column was matched by using the join recipe from the transform builder (Fig 4-17).

Formatting the 'DistanzeNomeStazioneFerroviaria' was mainly done manually by selecting the records and using the replace function to fill in the correct data. This is because the cluster clean feature that allows for standardization of values in a column by clustering similar values, is not available on the demo version of Trifacta.

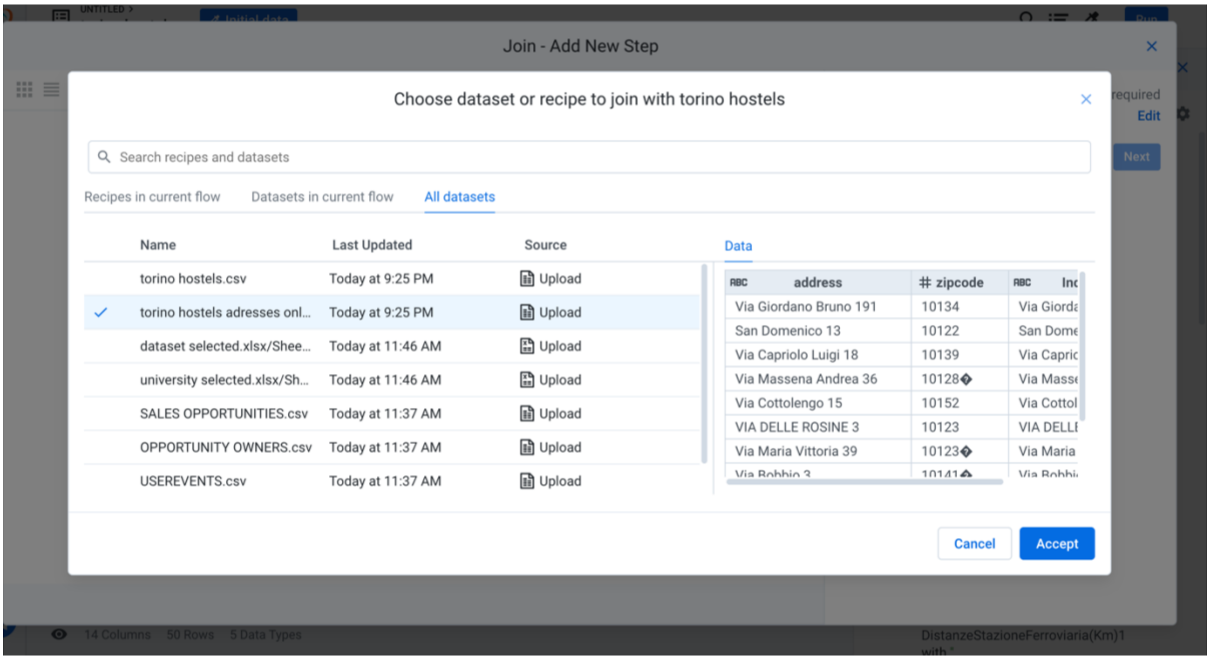

Figure 4-16-Join recipe in Trifacta

4.4.3 WikiDataset

The Trifacta tool rightfully did not identify any missing values or duplicate values in this data set. Although 3 mismatched values in the "Percentage Water" column were automatically identified. Just like in the hostel dataset, the embedded data in all the columns were easy to clean due to the automatic pattern recognition and the suggestions when any record on the dataset is highlighted. A summary of some of the transforms carried out on this dataset is shown in Fig 4-18.

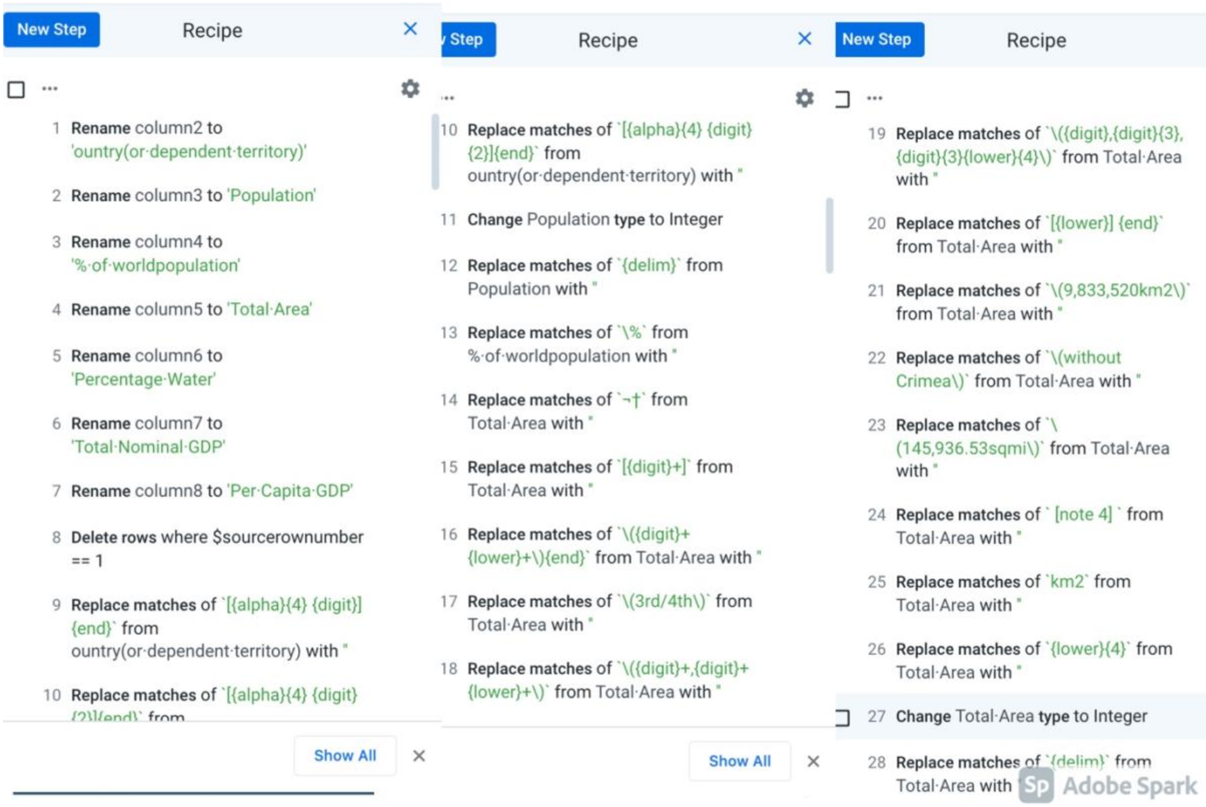

Figure 4-17-Summary of some transforms carried out with Trifacta on the wikidataset

4.5 Result and Observations

The tools were used to the best of their capabilities, based on the knowledge acquired from their documentation and tutorials. The demo version of Trifacta was used for this test and does not have all the features available in the paid versions. Each tool has its own strengths and weaknesses for example, OpenRefine has a very flexible coding functionality with GREL, Clojure and Python but it is lacking in terms of graphical representations for profiling. On the other hand, Trifacta does not have a flexible coding functionality and only relies on the inbuilt transformation features. While Trifacta is lacking in coding flexibility, it has a very good graphical representation for profiling. With the data quality bar, discrepancies are easily identified, the automatic pattern recognition helps to extract the regex of any record and the suggestions bar displays transformations that could be carried out on these discrepancies. This is particularly helpful for users who are not proficient with coding. Both tools have history tracking that allows a return to a previous point in the work progress. In addition, Trifacta has a preview feature that allows a view of any attempted changes, showing the exact way the change will affect the dataset before it is made.

Both tools were unable to match the address in the hostel dataset and extract the zip codes. While OpenRefine has a reconciling feature that supports reconciliation against a local dataset or other services such as wikidata, the wikidata reconciliation service was unable to match the address column to anything in its database and the manually made local dataset was matched wrongly with the reconciliation feature (Fig 4-8). As an alternative, the join table features of both tools were used and produced similar results (Table 4-1), with the exception of 9 embedded errors encountered in the new zip codes column, after the join in Trifacta. The errors could not be removed by any functions in the suggestions bar (Fig 4-19).

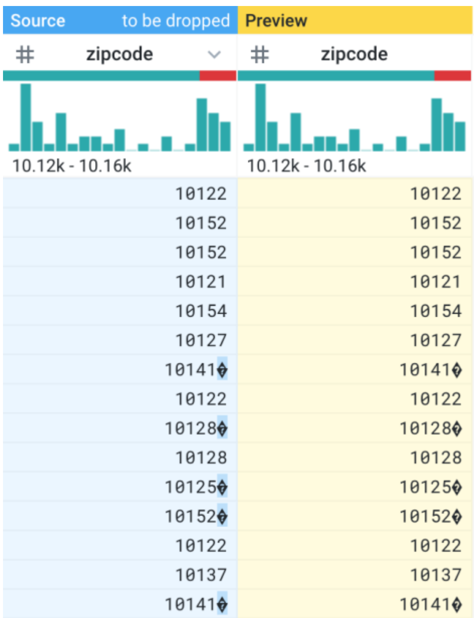

Figure 4-18-Embedded errors after the table join in Trifacta

While Trifacta has a built-in deduplication feature, it was ineffective on the University dataset and could not detect any duplicate records. For both tools, other multiple step methods were used which were explained in the previous section. The method used on OpenRefine identified and removed all the duplicate data while the method on Trifacta could detect and remove all but one duplicate records. This was because the primary key, on which duplicate records were checked in Trifacta, was a merge of the "university" and the "establish column", and the unidentified duplicate record was matched in the university column but had different value in the establish column.

The different results in accuracy in the university column (Table 4-3) is due to Trifacta tool's inability to remove the url encoded characters in the column. A feature which was achieved with 'unescaped url' in OpenRefine. There were also 2 records in the establish column which could not be reformatted by the tool.

The faceting feature in OpenRefine was easier to use for mass editing the country column of the university dataset in comparison to Trifacta. Trifacta has a similar feature called 'cluster clean', only available in the paid versions. Notwithstanding, similar results were achieved with the automatic pattern detection and the replace feature.

Table 4-4-Comparing Test Results Hostel dataset

HOSTEL DATASET						
		Columns	Description	Origina l	Openrefine	Trifacta
COMPLETEN	Existing missing values before deduplication			53	53	53
	Total records in dataset before deduplication			650	650	650
	Existing missing values after deduplication			53	53	53
	Total records in dataset after deduplication			650	650	650
EFFICIENCY	Number of numeric data stored as strings	Cap		0	0	0
		NroCivico		50	6	6
		RecapitiFax		0	0	0
		DistanzeParcheggioEsternoM		35	0	0
		DistanzeStazioneFerroviariaKm		45	0	0
				130	6	6
ACCURACY	Inaccurate records in data set	Cap	50 same zipcode was recorded for all addreses	50	0	9
		DistanzeParcheggioEsternoM	27 non uniform units (assuming the column should just have integers with no units within) of measurements	27	0	0
		DistanzeStazioneFerroviariaKm	Embedded units km, m, KM	41	0	0
		DistanzeNomeStazioneFerroviaria	1 'FS' entered on the 36th row, 39 noN uniform stan	40	1	1
				158	1	10
CONSISTENC	Inconsistent records in dataset	DistanzeParcheggioEsternoM	27 non uniform units (assuming the column should just have integers with no units within) of measurements	27	0	0
		DistanzeStazioneFerroviariaKm	Inconsistent records '700 m' , 'km 3,7'	2	0	0
		DistanzeNomeStazioneFerroviaria	39 no uniform standard	39	0	0
				68	0	0

Table 4-4-Comparing Test Results Wiki dataset

WIKI DATASET						
		Columns	Description	Original	Openrefine	Trifacta
OMPLETENES	Existing missing values before deduplication			0	0	0
	Total records in dataset before deduplication			270	270	270
	Existing missing values after deduplication			0	0	0
	Total records in dataset after deduplication			270	270	270
EFFICIENCY	Number of numeric data stored as strings	Population		0	0	0
		% of worldpopulation		0	0	0
		Total Area		30	0	0
		Percentage Water		12	0	0
		Total Nominal GDP		30	0	0
		Per Capita GDP		30	0	0
				102	0	0
ACCURACY	Inaccurate records in data set	Country(or dependent territory)	9 embedded records in Country(or dependent territoı	9	0	0
		Population		0	0	0
		% of worldpopulation		0	0	0
		Total Area	30 embedded records	30	0	0
		Percentage Water	12 embedded records	12	0	0
		Total Nominal GDP	30 embedded records	30	0	0
		Per Capita GDP	30 embedded records	30	0	0
				111	0	0
CONSISTENCY	Inconsistent records in dataset	Percentage Water	3 outliers (3,718,200), (551,695), (82,800,000)	3	0	0
		Total Nominal GDP	1 inconsistent format record (69,322 billion comma instead of dot)	1	0	0
				4	0	0

Table 4-5-Comparing Test Results University dataset

UNIVERSITY DATASET						
		Columns	Description	Original	Openrefine	Trifacta
COMPLETENESS	Existing missing values before deduplication			115	102	102
	Total records in dataset before deduplication			540	540	540
	Existing missing values after deduplication			80	70	71
	Total records in dataset after deduplication			360	360	360
EFFICIENCY	Number of numeric data stored as strings	endowment		20	1	1
		numFaculty		1	0	1
		numDoctoral		1	0	0
		numStaff		0	0	0
		established		11	0	2
		numPostgrad		2	0	1
		numUndergrad		1	0	1
		numStudents		0	0	0
				36	1	6
ACCURACY	Inaccurate records in data set	university	4 syntax errors	4	0	4
		endowment	18 embeded data, 1 incorrect entry ('US$ billion') and 3 outliers(3400, 15, 121)	19	4	4
		numFaculty	1 incorrect entry (Day Course and Evening Course)	1	0	0
		numDoctoral	1 non uniform standard for missing value (not available	1	0	0
		country	4 embedded data ('Canada B1P 6L2', 'Canada C1A 4P3 Telephone: 902-566-0439 Fax: 902-566-0795', 'England, UK' , 'United States)'), 12 non uniform standard representations (USA , u.s.a, US , etc)	16	0	0
		numStaff		0	0	0
		established	6 embedded data	6	0	2
		numPostgrad	2 incorrect data ('Some postdoctoral students and visiting scholars ', "not available ")	2	0	0
		numUndergrad	I incorrect record ('pre-university students; technical')	1	0	0
		numStudents	1 embedded data in numStudents(-18234).	1	0	0
				51	4	10
CONSISTENCY	Inconsistent records in dataset		18 duplicate data	18	18	17
		endowment	3 outliers, 19 inconsistent data in the endowment column both in terms of currency and format	22	3	4
		country	12 non uniform standard representations	12	0	0
		established	7 inconsistent date time format in the established column (yyyy and yyyy-mm-dd and yyyy-mm-dd hh:min:sec)	7	0	2
		numPostgrad	2 incorrect data ('Some postdoctoral students and visiting scholars ', "not available ")	2	0	1
		numUndergrad	I incorrect record ('pre-university students; technical')	1	0	1
		numStudents	1 embedded data in numStudents(-18234).	1	0	0
				63	21	25

Table 4-6-Results using the ISO 25024 Metrics

TOOLS			OPENREFINE			TRIFACTA			ORIGINAL		
	Data Quality Measure N	Measurement Function	UNIVERSITY	HOSTEL	WIKIDATASET	UNIVERSITY	HOSTEL	WIKIDATASET	UNIVERSITY	HOSTEL	WIKIDATASET
Accuracy	Record's field accuracy	A=number of records with the specified field accurate	286	596	270	279	587	270	229	439	159
		B=number of records	360	650	270	360	650	270	360	650	270
		A/B	79.4	91.7	100.0	77.5	90.3	100.0	63.6	67.5	58.9
Completeness	Completeness of data within a file	A= number of records with associated values not null for a specific data item	290	597	270	289	597	270	280	597	270
		B= number of records counted	360	650	270	360	650	270	360	650	270
		A/B	80.6	91.8	100.0	80.3	91.8	100.0	77.8	91.8	100.0
Consistency	Consistency of a data file	A=number of data consistent in the file	269	597	270	264	597	270	217	529	266
		B=number of data recorded in file	360	650	270	360	650	270	360	650	270
		A/B	74.7	91.8	100.0	73.3	91.8	100.0	60.3	81.4	98.5
Efficiency	Numbers stored as strings	Numbers stored as strings	4	6	0	6	6	0	36	130	102
		A= number of data items that are stored in a format that are qualified stored in a format that are qualified as efficient (for this case number of data stored as strings)	194	213	180	192	213	180	162	89	78
		B= number of data items for which format is tested for efficient operation (total number of records expected to be numbers i.e int or float or date)	198	219	180	198	219	180	198	219	180
		A/B	98.0	97.3	100.0	97.0	97.3	100.0	81.8	40.6	43.3

Table 4-7 shows a summary of the results of the evaluation of the Openrefine and Trifacta tool in terms of the ISO- 25024 [42]. The 'Measurement Function' column contains the components of the ratios recommended for measuring the dimensions of data quality in the ISO- 25024. The ratios (A/B) are expressed in percentages. Comparing the outcomes of the tools with the original dataset, clearly there is an improvement in the quality of the datasets on applications of the tools.

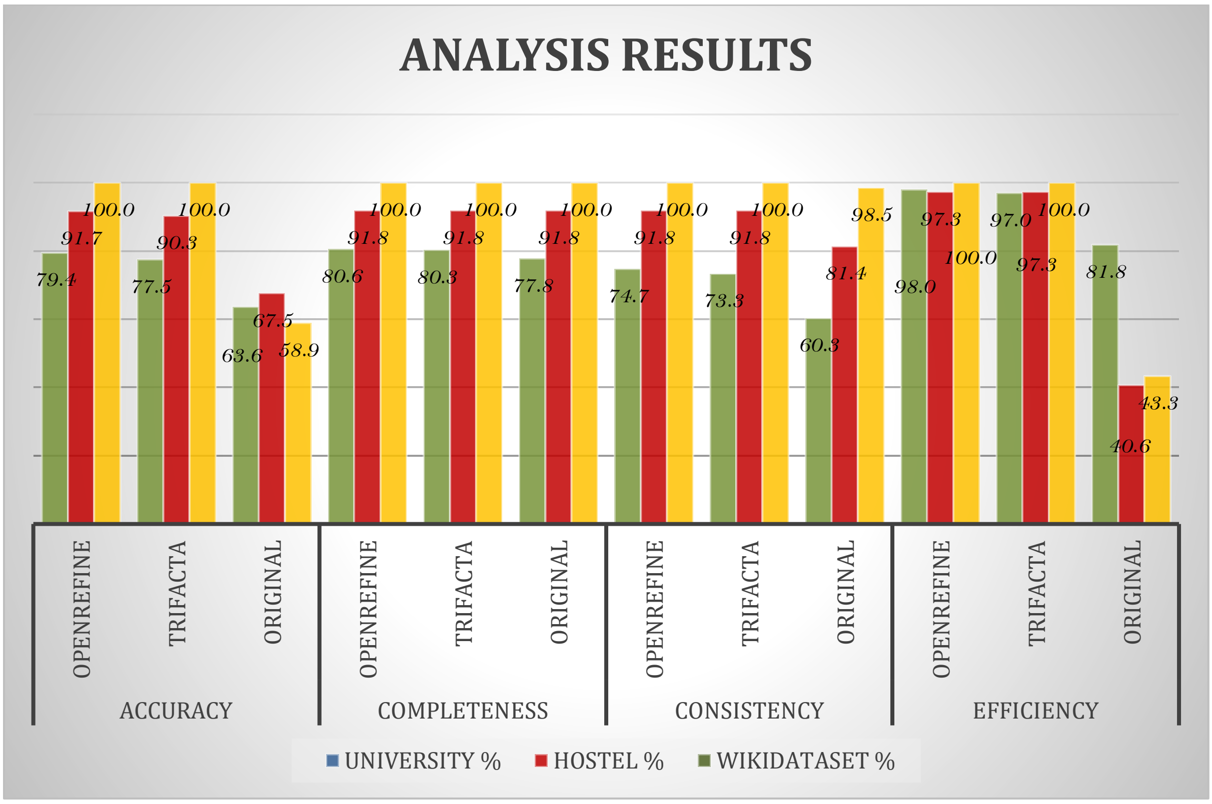

Figure 4-19-Graphical representation of the final results

The two tools had similar method applied to achieve better data quality of the three datasets. The results in Figure 4-19 show both tools are effective for cleaning data and improving data quality. The results also show OpenRefine has a slight edge over the trial version of Trifacta for the university and hostel dataset. Although both tools were 100% effective in fixing quality problems in the hostel dataset.

Chapter 5

5 Conclusion

Data quality is a complex concept with varying perspectives, but its importance is evident both in business and government organizations. In this thesis, we have defined and explained the concepts of data quality. Some benefits of good data quality to organizations were presented, and the risks and implications associated with bad data quality were discussed. The challenges that bring about bad quality data were also described, highlighting the many different types of data structures and types, making it difficult for data integration and the data animalities caused by human errors within the organizations either by customers or employees. The standards of data quality as stated in the ISO 25000 series were further explained. These standards aim to provide a uniform framework to support the specification of software quality requirements and the evaluation of software quality, through defined and standardized criteria for measurement and evaluation. The 15 dimensions of data quality; accuracy, consistency, completeness, timeliness, credibility, accessibility, compliance, confidentiality, efficiency, precision, traceability, understandability, availability, portability and recoverability, stated in the ISO25012 were discussed. The dimensions describe a context for data quality attributes and a frame of reference to have these attributes measured. Practical examples of each of the dimensions were also explained.

Chapters 3 and 4, focused on the study of data quality tools. Data quality tools automate the process of assessing and enhancing the quality of data, by detecting and fixing the data problems that influence the overall data quality. The common data quality processes (data profiling, data cleaning, data integration, data monitoring, data enrichment, data governance, etc.) were used to create a comparison matrix for some data quality tools. The comparison matrix identified several features of the different tools that serve as support to the data quality processes. The comparison matrix was normalized to provide a ranking of the tools according to how their features cover the 15 dimensions of data quality.

Finally, to get better understanding of the functionality of these tools, two of the tools (OpenRefine and Trifacta) were tested with three real life open-sourced datasets. Exploratory analysis on the datasets was initially carried out, identifying and recording the errors that exists within each dataset, then more work was done to improve the quality of these data sets, using the selected tools. For testing the tools, only the inherent data quality dimensions which were applicable to the datasets was considered. The testing of the datasets with the two tools, showed that while the tools assist in identifying and solving the data quality problems, the level of automation still needs to be developed further as some of the features of the tools were dependent to an extent, on the skills of the user. The result of our analysis placed Openrefine at a slight edge over the demo version of Trifacta.

References

[1] L. Sebastian-Coleman, *Measuring data quality for ongoing improvement: a data quality assessment framework*. Amsterdam: Elsevier/MK, Morgan Kaufmann, 2013.

[2] P. Glowalla, P. Balazy, D. Basten, and A. Sunyaev, “Process-Driven Data Quality Management -- An Application of the Combined Conceptual Life Cycle Model,” in *2014 47th Hawaii International Conference on System Sciences*, Waikoloa, HI, Jan. 2014, pp. 4700–4709, doi: 10.1109/HICSS.2014.575.

[3] “How to Create a Business Case for Data Quality Improvement.” //www.gartner.com/smarterwithgartner/how-to-create-a-business-case-for-data-quality-improvement/ (accessed Dec. 30, 2020).

[4] “The Four V's of Big Data,” *IBM Big Data & Analytics Hub*. https://www.ibmbigdatahub.com/infographic/four-vs-big-data (accessed Dec. 30, 2020).

[5] T. C. Redman, “Seizing Opportunity in Data Quality,” *MIT Sloan Management Review*. https://sloanreview.mit.edu/article/seizing-opportunity-in-data-quality/ (accessed Dec. 30, 2020).

[6] “How to quantify Data Quality?. From individual data quality metrics to… | by Yannick Saillet | Towards Data Science.” https://towardsdatascience.com/how-to-quantify-data-quality-743721bdba03 (accessed Jan. 04, 2021).

[7] B. W. W. Eckerson and 05/01/2002, “Data Warehousing Special Report: Data quality and the bottom line -,” *ADTmag*. https://adtmag.com/articles/2002/05/01/data-warehousing-special-report-data-quality-and-the-bottom-line.aspx (accessed Dec. 30, 2020).

[8] D. Loshin, “The Organizational Data Quality Program,” in *The Practitioner’s Guide to Data Quality Improvement*, Elsevier, 2011, pp. 17–34.

[9] “ISO/IEC Guide 59:2019(en), ISO and IEC recommended practices for standardization by national bodies.” https://www.iso.org/obp/ui/#iso:std:iso-iec:guide:59:ed-2:v1:en (accessed Jan. 05, 2021).

[10] C. Batini and M. Scannapieco, *Data quality: concepts, methodologies and techniques*. Berlin: Springer, 2006.

[11] “IMDb movies extensive dataset.” https://kaggle.com/stefanoleone992/imdb-extensive-dataset (accessed Jan. 05, 2021).

[12] “Edit distance,” *Wikipedia*. Jan. 03, 2021, Accessed: Jan. 09, 2021. [Online]. Available:https://en.wikipedia.org/w/index.php?title=Edit_distance&oldid=998049392.

[13] A. F. Karr and A. P. Sanil, “Data Quality and Data Confidentiality for Microdata: Implica- tions and Strategies,” p. 5.

[14] D. McGilvray, *Executing Data Quality Projects: Ten Steps to Quality Data and Trusted Information^{TM}*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2008.

[15] J. E. Olson, *Data quality: the accuracy dimension*, Nachdr. Amsterdam: Morgan Kaufmann, 2008.

[16] T. Kusumasari and Fitria, *Data profiling for data quality improvement with OpenRefine*. 2016, p. 6.

[17] D. McGilvray, *Executing Data Quality Projects: Ten Steps to Quality Data and Trusted Information™*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2008.
[18] S. Latifi, Ed., *Information Technolog: New Generations*, vol. 448. Cham: Springer International Publishing, 2016.
[19] J. M. Barker, *Data Governance: The Missing Approach to Improving Data Quality*. University of Phoenix, 2016.
[20] "OpenRefine user manual | OpenRefine." https://docs.OpenRefine.org// (accessed Feb. 02, 2021).
[21] "The premier open source Data Quality solution | DataCleaner." https://datacleaner.github.io/ (accessed Feb. 02, 2021).
[22] "Data Cleansing & Address Correction: SQL Power DQguru | SQL Power Software." http://www.bestofbi.com/page/dqguru (accessed Feb. 02, 2021).
[23] "Data Modeling & Profiling Tool: SQL Power Architect | SQL Power Software." http://www.bestofbi.com/page/architect (accessed Feb. 02, 2021).
[24] "Tutorial — csvkit 1.0.5 documentation." https://csvkit.readthedocs.io/en/latest/tutorial.html (accessed Feb. 02, 2021).
[25] "Documentation," *Trifacta Documentation*. https://docs.Trifacta.com/display/r076/Documentation (accessed Feb. 02, 2021).
[26] "Cloudingo - Salesforce Data Cleansing and Management Tool." https://cloudingo.com/ (accessed Feb. 02, 2021).
[27] swinarko, "Data Quality Services - Data Quality Services (DQS)." https://docs.microsoft.com/en-us/sql/data-quality-services/data-quality-services (accessed Feb. 02, 2021).
[28] "Talend - A Cloud Data Integration Leader (modern ETL)," *Talend Real-Time Open Source Data Integration Software*. https://www.talend.com/ (accessed Feb. 02, 2021).
[29] "Data Ladder: Enterprise Data Profiling, Cleansing, and Matching," *Data Ladder*. https://dataladder.com/ (accessed Feb. 02, 2021).
[30] "TIBCO Clarity Cloud - Data Cleansing and Transformation Software." https://clarity.cloud.tibco.com/landing/index.html (accessed Feb. 03, 2021).
[31] "DemandTools: #1 CRM Data Quality Tool," *Validity*. https://www.validity.com/products/demandtools/ (accessed Feb. 03, 2021).
[32] Ataccama, "Self-Driving Data Management & Governance." https://www.ataccama.com/ (accessed Feb. 03, 2021).
[33] "The Simple Way to Unlock Your Raw Data," *Datameer*. https://www.datameer.com/ (accessed Feb. 03, 2021).
[34] "Enterprise Cloud Data Management | Informatica." https://www.informatica.com/ (accessed Feb. 03, 2021).
[35] "Data Management Software." https://www.sas.com/en_us/solutions/data-management.html (accessed Feb. 03, 2021).
[36] "DataFlux Data Management Server." https://support.sas.com/en/software/dataflux-data-management-server-support.html (accessed Feb. 03, 2021).
[37] "Data Management and Analytics." https://www.hitachivantara.com/en-us/products/data-management-analytics.html (accessed Feb. 03, 2021).

[38] “#1 Data Cleansing Tool & Data Matching Software ▷▷ WinPure,” *WinPure*. https://winpure.com/ (accessed Feb. 03, 2021).
[39] “Data Quality Management Solutions & Services | Experian,” *Experian Data Quality*, Nov. 07, 2014. https://www.edq.com/ (accessed Feb. 03, 2021).
[40] “osDQ Documentation,” *osDQ Documentation*. www.arrahtech.com/docs/profiler_user_guide.html (accessed Feb. 03, 2021).
[41] “Meet xDM,” *semarchy.com*. https://www.semarchy.com/xdm/ (accessed Feb. 03, 2021).
[42] ISO/IEC 25024:2015, Systems and Software engineering Measurement of data quality.
[43] “2018 UUtah Reproducibility Short Course,” Jun. 2018, https://osf.io/39fus/
[44] “Hostels 2017 | data.gov.it.” https://dati.gov.it/view-dataset/dataset?id=e2634520-0fb0-4b4d-b882-b9c9402807e8 (accessed Feb. 24, 2021).
[45] K. Bhanot, *kb22/Web-Scraping-using-Python*. 2021. https://github.com/kb22/Web-Scraping-using-Python/blob/master/Dataset.csv
[46] “Deduplicate Data,” *Trifacta Documentation*. https://docs.Trifacta.com/display/AWS/Deduplicate+Data (accessed Feb. 25, 2021).

Appendix

Table I - Summary of list of data sets, tables, python notebooks and graphs which can be found in the GitHub repository https://github.com/chizzymara/thesis

File Name	Description
Dataset.csv	The original wikidataset
IMDb ratings.csv	Original imbd dataset
Comparison matrix.xlsx	Excel file with all the tables found in the thesis
Dataset selected.xlsx	A subsection of the wiki dataset which was used for the testing of the tools.
Dataset selected.xlsx_Sheet cleaned with trifacta.csv	Final version of the wiki dataset processed with Trifacta
Imbd code1.ipynb	Here all tables and code found in chapter 3 of the thesis are found.
Imbd subsection.xlsx	Subsection of the imbd dataset, originally from kaggle https://www.kaggle.com/stefanoleone992/imdb-extensive-dataset/download

python analysis.ipynb	Notebook with the python analysis used to identify problems in the dataset.
reg_ostelli_2017.csv	Original torino hostel dataset.
torino hostels adresses only.csv	Dataset manually created with accurate zip codes for the addresses found in the torino hostels data. For matching or reconciliation.
torino hostels cleaned with trifacta.csv	Final version of torino hostels dataset cleaned with Trifacta.
torino-hostels-xlsx cleaned with open refine.xls	Results of torino hostels dataset cleaned with OpenRefine.
university selected.xlsx	Subsection of the university dataset used to test the tools.
university selected.xlsx cleaned with trifacta.csv	Result of processing the University dataset with Trifacta
university-selected-xlsx cleaned with open refine.xls	Result of processing the University dataset with OpenRefine
universityData.csv	Original university dataset
wikidataset cleaned with openrefine.xls	Result of processing the wikidataset dataset with OpenRefine