POLITECNICO DI TORINO

Master in Engineering and Management

Master Thesis

How big data are changing football: analytic method for evaluating players' economic value

Tutor
Prof. Roberto Fontana

Candidate
Marco Bungaro

December 2019

Abstract

For football clubs the most valuable assets are the contracts of the players. The transfer value of football players has increased dramatically in the last years as more money is involved in the football industry, but the estimation of such value is still a matter of subjective opinions of professionals and public. The sport industry is generating high quantity of data and the problem of building a data-driven model for the analytical evaluation of transfer fees is still open. The first part of the research is dedicated at first, defining which are the variables that have higher impact on the performances of a player, according to his role in the team, using as pilot dataset the Serie A 2017/2018 data, secondly building a model for evaluating the market value of all the strikers and comparing predicted values with the actual transfer fees of the summer transfers and with the estimated market value proposed by the reliable source Transfermarkt.

The second part of the research focuses on automating this process to extend the dataset to other seasons and leagues, creating a scalable workflow that scrapes the data from internet, cleans the row data and prepares the final dataset. As last step, a similar approach as before is applied to a bigger dataset to understand if the precedent model would still be valid and if there are any significant differences between leagues and seasons.

Summary

1 Introduction

1.1 Background

This research starts as a combination of my two greatest interests: a fanatic passion for football and a great curiosity for data analytics. Football is the most popular sport in the world and it is growing its fan base year-on-year (Frick, 2007). Football clubs in Europe compete to maintain and improve their competitive position at the highest rankings, in order to have more visibility and win the highest prizes offered by the competitions. To be the best in Europe teams are willing to invest increasing amount of money in their most valuable assets: football players.

During the last years transfer fees of football players have reached shocking levels creating surprise in the crowd and rising several questions. Is it worth to invest 228M€ in signing Neymar JR? Does Higuain really value 90M€?

It is very difficult to answer these questions as the entertainment industry has a variety of factor to keep into account. The objective of this research is to evaluate football players based on their performances to have an analytical and reliable support in decision-making and negotiation of new transfers.

1.2 Research questions

The main question this research wants to solve is: *"How can football players' market value be evaluated through their performance?"*. There is a subset of questions to be answered to have a clear understanding of the topic.

- "How is evaluated today the market value of a player?"
- "How can performances be analysed and how can they influence the economic value?"
- "Who are the most undervalued or overvalued players and why?"

1.3 Research Scope

The research focused on the Italian league Serie A through the whole 2017/2018 season as it was the latest season with fully seasonal data. Market data are defined at June 2018 and performances are individuated across all the matches of a singular season.

The development of the research started with the problem definition, followed by the literature review, data collection, model building and conclusion.

1.4 Research Method

The research has been built on defining a problem and investigating the literature review. The tools used for researching the literature review are both open and licensed by the universities Politecnico di Torino and ESCP Europe.

The open tool I used the most for having a wide resume of the literature review is Google Scholar. Through the license of the university I was also able to access some online dataset that collects many articles such as Scopus and Statista for statistics.

The literature review has been continuously applied during the whole research process to highlight the scientific knowledge of every subtopic. The literature research was not narrowed only to football and Serie A but was widely conducted through different sports and different approaches to data mining for athletic performances and economic evaluation. The most inspiring academic paper on which I referred to find the datasets and the information is the 2015 research of Miao He at Leiden University, called "Exploring the Relationship between Football Players' Performance and Their Market Value" that will be cited in the next phases of the study.

2 Literature review

2.1 Big Data in Football

The usage of big data is changing the world we live in and the football industry is part of this change. Football has been a laggard in introducing data analytics, respect to the USA sports as Basket, Baseball and American football (Kidd, 2018).

There are several reasons explaining this phenomenon. The first is that football has always been thrived by passion and has a more intrinsic relationship with its fan. Football teams generally represent cities and their history, thus the cold-blooded analysis have lower weight than the emotions. The second reason is that football generates less and lower quality data than other sports such as basket.

As of today, the players' movements are all tracked on the field and are able to generate a high density of data. All the main stadiums are filled with 8-10 cameras dedicated to collect players' movements and during training the footballer generally wears GPS tracker to monitor their fitness (Marr, 2015).

Football is not an individual sport and the outcome of the matches is highly determined by the interaction between players. It is defined tactics and has a highly level of complexity to be analysed but with the big data technologies might be in a short term future be a solution for tactics simulations (Rein and Memmert, 2016). Moreover, there are some researches that are moving towards the use of big data with the aim of identifying the key players through network analysis (McHale and Relton, 2018).

Football players are the most valuable asset in football companies (Lozano and Carrasco Gallego, 2011), thus I will focus on using data to understand the valuation of such assets based on their performances.

2.2 Market value estimation

The aim of this research is to expose how is evaluated the market value of a football player and how can this process be improved with new technological trends. The first important aspect to clarify is the definition of transfer fee and market value.

Transfer fees are the representation of the actual price paid by the acquiring club to the selling club and comes out after negotiations, while market value is the estimation of a potential fair value of the transfer fee for that player in that moment (Herm, Callsen-Bracker and Kreis, 2014).

This value has been generally estimated by football experts but during the last years crowdsourcing methods have come out and gradually substituted it (Müller, Simons and Weinmann, 2017). As of today, the most trusted institution in football players' market value estimation is the online website Transfermarkt.com, that has achieved a respected credibility through the years. Transfermarkt offers a wide variety of football data including goals, assists, number of matches played and others as transfer fees and market value estimation for teams and individual players.

The market value is based on crowd evaluation. Every registered user can participate to thread discussion related to a single player proposing market values and arguing the reasons behind (Müller, Simons and Weinmann, 2017).

The users have different weights in the decision-making process depending on their historical trustworthy suggestions. There are some users who are in charge of taking the final decision who are called the "judges" (Herm, Callsen-Bracker and Kreis, 2014).

Today Transfermarkt is still the most used and reliable source for estimating the market value even if many researchers have tried to create algorithms capable of outsmarting human decision making and defining less biased market values.

2.3 Analytical evaluation

The literature review highlighted different techniques for evaluating the impact of football players' characteristics on their market value. First of all, qualitative researches focused on the importance of identifying and classifying the skills needed to be a high performer player in every specific role (Hughes *et al.*, 2012). The identification of such factors is influenced by several information, which need both statistical and professional knowledge and the building of the model should be performed by a multi-disciplinary team (Memmert and Rein, 2018). I will rely on my personal experience in football to compensate the lack of professionals in football performance analysis. Majewski proposed an econometric model taking into account the impact of 14 variables subdivided in seasonal performance counting the number of appearances, substitutes and age, field performance counting goals, cards and assists and team performance counting the level of the player's current club to estimate the market value (Majewski, 2016). The relationship between seasonal records of La Liga football players and their Transfermarkt estimated market value has been statistically analysed to understand which factors have higher impacts (He, 2013). This research will follow a similar approach to exploit online accessible data with a different dataset, focusing on the Serie A.

3 Data

3.1 Data Selection

The main reason for choosing the Italian league is that I am a Serie A fan and I have a considerable personal knowledge of players' performances and valuations across the last years. Moreover, while Serie A was recognized as the best league, today is ranked number 4 in the UEFA season country coefficient (UEFA, 2019) but is gaining international importance again. Indeed, in the last three year, from 2016 to 2018, the overall expenditure of Serie A clubs in the summer registration period has increased from 477 M€ to 1000 M€, which is an increase of 110%. Among the most important leagues (Italian, Spanish, English, German and French leagues) it is the second highest increase in percentage after France and it moved from being the third country with highest expenditure in 2016 to be the second highest in 2018 (KPMG, 2018). The Italian movement is highly growing and it is few explored in the literature making it an interesting field to be studied.

3.2 Data Collection

The dataset has been collected first analysing different potential sources and then choosing the most accessible and useful ones for this research topic. There are several open databases that collect several data on football matches with different aims. As the scope of the project is to individuate the seasonal performances of every football player

and individuate their market value at the end of the year I decided to focus on two reliable and well-known open datasets.

The first is Whoscored.com, which is a UK based football website that collects all the seasonal data on every player of the major European and South American leagues. The data are wide and cover almost all the events that can happen in a single match such as goals scored, assists, key passes, tackles and even more detailed ones such as number of shots in the penalty area and aerial tackles won.

It has been already validated by previous research as a reliable website for accuracy in data and for the variety of variables collected (He, 2013). The scraping technique was in this case quite difficult as the website displays all the players in several pages. This means that I could not identify a single table from which I could extract all the data.

The data were extracted following a repetitive process of copying the information on all the different pages of football players for every specific attribute and then repeated again across all the different attributes. Moreover, as the full database included more than 500 players, I decided to narrow the selection as most of these did not play a sufficient number of games to be statistically significant for analysing the performance. Thus, the final dataset includes only the players who played during the season 2017/2018 at least the same number of matches as the average player in that season, which was 19 games comprehensive of starting and substitute appearances. The dataset was thus restricted to 284 players.

The second database I scraped from the Internet is related to data transfers and market value estimation and it is the German website Transfermarket.com, which is the most

used and most reliable source in terms of identification of market value for specific football players.

This time I used the free version of an online data scraping tool called WebHarvy. This application made easy to collect all the data from a single football team just by selecting all the columns of the table and labelling them with the correspondent data.

Unfortunately, as the data collected refer to the season 17/18 the market value is related to June 2017. For the purposes of this research it is not an acceptable proxy value for the market value of 2018, as it will be affected by the football performances during the seasons and by other external factors. Thus, I had to collect all the updated market value to the date of June 2018. I only looked at those players who had enough data regarding the seasonal performances coming from the Whoscored.com database which were around 280. This reduced the amount of values to collect but it was not possible to scrape automatically this data from Transfermarkt.com as the website updates the market value every 3 months so the value related to the 18/19 season was not of June 18 but of January 19, which was biased by the performances of the beginning of season 18/19 that are not taken into account in this research. To solve this situation, I had to check the market values in June 18 of every player one by one.

The use of this database for estimating market value is widely recognised in the literature as an efficient and reliable method (Herm, Callsen-Bracker and Kreis, 2014). Moreover, it has become through years a reference point for the football insiders such as journalists, administrative directors of football clubs and market agents who sponsor their football player.

Merging these two datasets I was able to identify the data in three different categories:

Players' general information – all the information that identify the football player but are not related to any aspect of the game (name, age, nationality, preferred foot, height etc…)

Players' market information – all the information related to the market value of the player (team, contract length, position, market value etc…)

Players' performance information – all the information related to the seasonal performance on the field (goals, assists, tackles, rating etc…)

3.3 Data Preparation

Once the dataset has been completed, the delicate procedure of preparing the data for analysing the relationship among them was carried out. This step is of fundamental importance for providing a coherent data mining procedure (Zhang, Zhang and Yang, 2003). There are 3 main aspects that are affected by the data preparation: (1) the impurity of data in real word, (2) the quality of the data and (3) the quality of the patterns (He, 2014).

In this case data come from two different data sources, thus I had to merge the two of them into a single dataset before focusing on the preparation of specific data.

Encryption

As the SAS English version was having troubles with the codification of accented letters in the dataset, I saved all the excel files to be imported in the UTF-8 format and finally used the Italian version of SAS.

3.4 Data Merge

The two datasets have few attributes in common such as the nationality, the age, the number of seasonal appearances and the player's first and last name. This last attribute is the key for merging the two datasets as there were no homonymous in the 2017-2018 Italian league.

However, the two datasets refer to the player's name in different ways. First of all while Transfermarkt shows only the players' name, Whoscored at the voice "name" inserts "FirstName LastName Team, Age, Position,". For example if you look for Dybala in Transfermarkt the record would be Paulo Dybala while on Whoscored the record would be "Paulo DybalaJuventus, 25, AM(CR),FW".

The attributes are all divided by a comma except the team, which was the most difficult part to deal with. As the research was not intended to focus on how to deal in data preparation and the current team was a duplicated information as it was already present in the Transfermarkt database the quickest and most efficient solution was to remove in a scalable way all the 20 Serie A team names from the column "players". I had to deal separately with two specific players that contain the name of a team inside their Last Name, which are Alessio Romagnoli and Romagna (Appendix 2*)*.

Once the team name was removed all the variables were divided by a comma, which made it easy to separate in three different columns related to “Player name”, “Age” and “Position”.

The name was then used as the key attribute to merge the two datasets. Initially 14 data were lost during the merge because of incongruency in naming. Mostly for Argentinian and Brazilian players the First name could be substituted in some cases by the nickname. For instance, while Whoscored uses the full name for “Alejandro Gomez” and “Jose Reina”, Transfermarkt names them as “Papu Gomez” and “Pepe Reina” referring to how they are generally called in the football industry.

As the focus of this research is not the name-matching algorithms and the errors were less than 5%, I decided to solve the issue by manually changing the nicknames into the official name directly on SAS (Appendix 2*)*.

Finally, the last check was to remove all those players who retired at the end of the season and who had a market value equals to 0 as this would have created some noise in the market value estimation.

3.5 Data conversion

Before starting analysing the data, an ultimate step has to be taken into account. Some attributes were not ready to be read in the proper way by the machine. The duration of the contract and the time spent in the club were both inserted as dates. It is impossible

for the basics SAS programs to understand the real correlation between dates, as they are considered as a string, and market value.

The solution was to convert dates into number of years. I have considered as actual date June 2018 and then subtracted the contract ending date and the actual date. The same was done for the year spent in clubs, subtracting the starting date in the club and the actual date (ex. 06/2021 → 06/2021 – 06/2018 = 3).

For the age was not necessary as it was considered as the birth date in the Transfermarkt database but in the Whoscored database there was the same information in numeric format. As the Whoscored age was referring to June 17 I had to increase all the values by one to have a consistent dataset (ex. 06/2012 → 06/2018 – 06/2012 = 6).

Market value was initially saved as number of millions and the character "M" or "Th" in case the value was in thousands of euros. It was then converted into a simple number in the scale of millions (ex. 65 M. → 65, 350 Th. → 0.35).

3.6 Final Dataset

The final dataset contains 280 rows and 94 columns. It means that there are 280 football players with the minimum conditions required for the study in the Serie A 2017/2018 and that I have identified 94 attributes that can potentially have an impact in evaluating the market value. The variables are explained in appendix 1.

4 Dataset Analysis

4.1 Market value distribution

The market value estimated by Transfermarkt has been used as the initial proxy value for transfer fees. In the following paragraph I will discuss the distribution of this value across the Serie A players to understand whether it is normally distributed or not and if there are any variables that can have an impact on the valuation of the player.

This procedure has been largely done in the literature review as one of the first steps once the full dataset was collected (Müller, Simons and Weinmann, 2017). It could be helpful to identify where is the mean and the median and if there are any tails very far from the mean region. As the football evolves every year and gather more money in the main leagues, the superstars caught the major percentage of this increasing market, creating a greater separation in absolute terms with the average football player. The superstar is defined as a person who can earn an unproportioned salary due to his status in comparison with the other people involved in the same sector (Rosen, 1981).

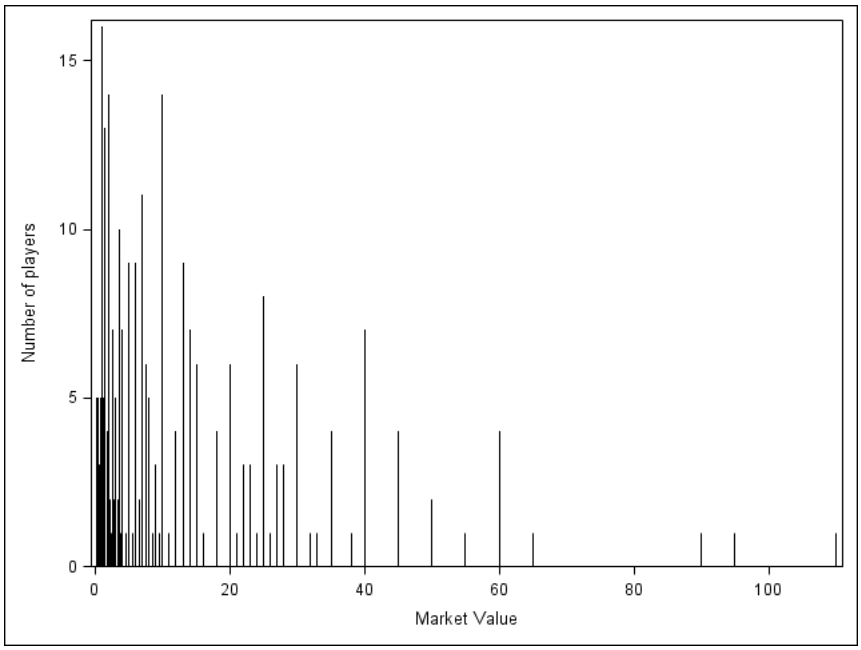

Figure 1- Distribution of market value by number of players

Figure 1 shows the market value distribution. The left region, the closest to the zero, is the most crowded of players as it can be seen by the overlapping of discrete bars and the high volumes on the y axis. It is indicative to look at the mean and median of the market value. The mean is 12.38 Mln € which is somehow quite high, but if we look at the median it falls dramatically to 6 Mln €. It indicates how the average value is dragged to the right by the top players high values.

The more we go to right the less players we find, and we can see a right tail with very few players whose market value is almost 10 times higher than the average. Thus, the market value is not distributed following a normal distribution.

4.2 Market Value and Position on the field

In the literature review we can easily find the differentiation of players by role and position on the football field, as Hughes suggested a classification in 7 different main roles of football players on the field: Goalkeepers, full backs, centre backs, holding midfield, attacking midfield, wide midfield and strikers (Hughes *et al.*, 2012). This is not the only differentiation for roles that can be done in football going more in details of every position specifying the side of the field on which the player majorly acts (left/right) or individuating fluid position as the football is moving more and more towards a more dynamic statement of positions than the past.
As I merged two different datasets some incongruency might emerge on the definition of the position on the field of a specific football player. It generally happens for those players who play in the middle between two roles, playing both as a full back or a side midfielder, rather than playing as a striker or a wing. Indeed, while Transfermarkt only identifies one position for every player, Whoscored allocates from one to a maximum of three different roles for every player, depending on their flexibility.

For every player in the Whoscored dataset the order of the positions is not based on the relevance but on the physical location on the field, from the goalkeeper to the centre forward. Thus, it might happen that a player who generally plays as a right wing, but occasionally can play as a right midfielder, will have in the first position the "Right Midfielder" and in the second one the "Right Wing".

I have preferred to use the Transfermarkt definition for the position statement, as it identifies only the most relevant position for every player. This differentiation proposes 13 subcategories defined as "Position":

- *Goalkeeper*
- *Right-Back*
- *Left-Back*
- *Centre-Back*
- *Right Midfield*
- *Left Midfield*
- *Defensive Midfield*
- *Central Midfield*
- *Attacking Midfield*
- *Left Winger*
- *Right Winger*
- *Second Striker*
- *Centre-Forward*

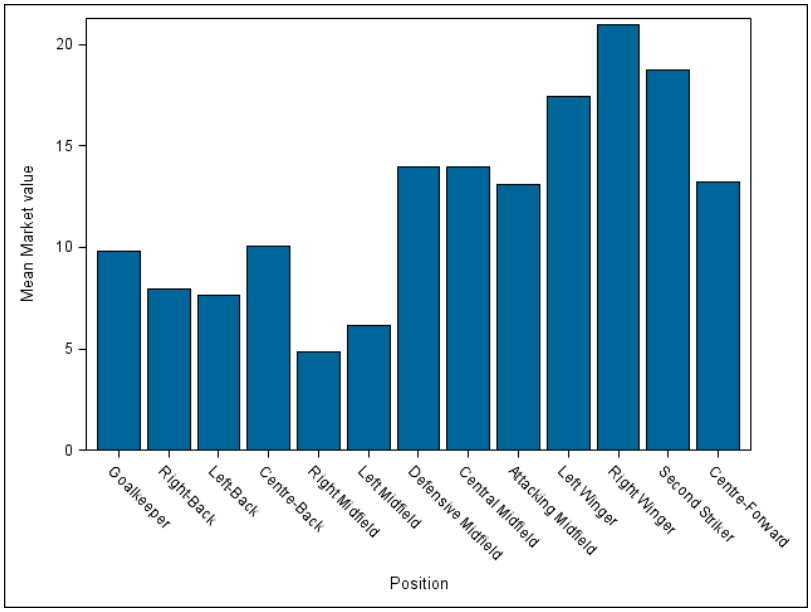

Figure 2 - Histogram of market values by position on the field

In *figure 2* the average market value of football players belonging to each position is shown. The positions are sorted on the x axis from defence to attack, starting from the

goalkeeper and finishing with the centre-forward. The more the player plays closer to the opponents' goal, the higher is generally his market value. As the overall dataset contains 280 players, dividing them in 13 categories generates groups with a low frequency and so inconsistent significance. This is the reason for which right winger have very high average and right midfielder have very low ones.

It would be better to identify high level categories, in order to have a significant number of players per each category. The classic definition of "role" in football divides the players in 4 level: Goalkeeper, Defender, Midfielder and Forward.

Every position has been related to a role, according this scheme:

Position	**Role**
Goalkeeper	*Goalkeeper*
Right-Back	*Defender*
Left-Back	*Defender*
Centre-Back	*Defender*
Right Midfield	*Midfielder*
Left Midfield	*Midfielder*
Defensive Midfield	*Midfielder*
Central Midfield	*Midfielder*
Attacking Midfield	*Midfielder*
Left Winger	*Forward*
Right Winger	*Forward*
Second Striker	*Forward*
Centre-Forward	*Forward*

The categories are then reduced to just 4. The group of Goalkeepers only counts 18 players, while the others count on average 90 players per group, which is a reasonable numerosity for analysis purposes. I took into consideration the same metrics, comparing the average market value for each role, sorting the x axis based on roles from goalkeeper to forward.

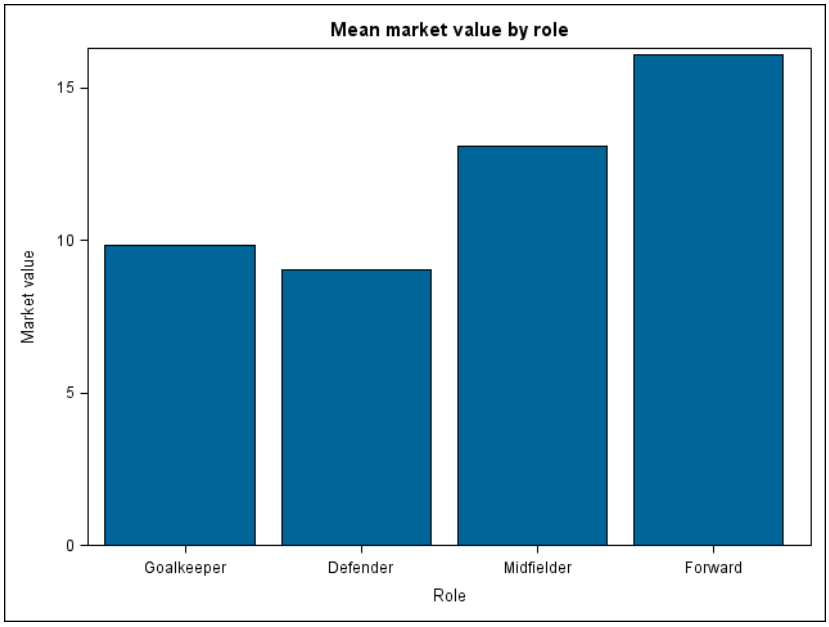

Figure 3 - Histogram of market values by roles

Figure 3 shows the differences in evaluating football players based on the roles they play. It is visible the increasing in market value by moving the position towards the goal. Even if football is a team sport and every player is fundamental to the final win, the status of the man who is signed on the final table is always recognized as more important. It is fundamental to score goals for winning and clubs are willing to pay more for those people who will take the responsibility of scoring. On the other hand also defending the goal from opponents is considered as fundamental and the goalkeepers are achieving the role of superstars, as Alisson in June 2018 was evaluated 60 M€ by Transfermarkt and was transferred from Rome to Liverpool on a 62.5 M€ deal, the highest in history for a Goalkeeper.

Thus, the strikers are the players that catch more the attention, both for being the constant presence in highlights of matches and for making the difference in the final result, but also for being considered economically more valuable than the rest of the players.

4.3 Market Value and Age

The career duration in the same league of a European football player, defined as the number of years without interruption, neither by moving in a lower league neither moving abroad, is rather short as the average is around 4 years (Frick, 2007). Frick studied the Bundesliga, which is the major league in Germany and that could be comparable to the league of this study, Serie A, as both are the major leagues of their respective countries and they are respectively 3rd and 4th in the UEFA country coefficient.

The UEFA country coefficient is a classification of every European country based on the performances of the teams of each nation in the UEFA Champions League and UEFA Europa League. It is a good proxy of valuation of competitiveness of the league and it is also used to determine the number of teams allowed to participate every year to the UEFA Champions League and UEFA Europa League to represent their country. Italy being 4th has the maximum number of teams accepted, since the first four teams of the championship are invited to the UEFA Champions League and 3 more teams are invited to the Europa League, as it happens for the Premier League which is 1st in the ranking.

As Frick points out in his research, *"age has a statistically positive influence on the probability of being eliminated by Bundesliga"* which means the older you are the more probable is that you will not play in Bundesliga next season.

It might then be significant to understand how the market value changes in relationship to the age, as it is done in the literature (He, 2014), and how age is distributed across the dataset.

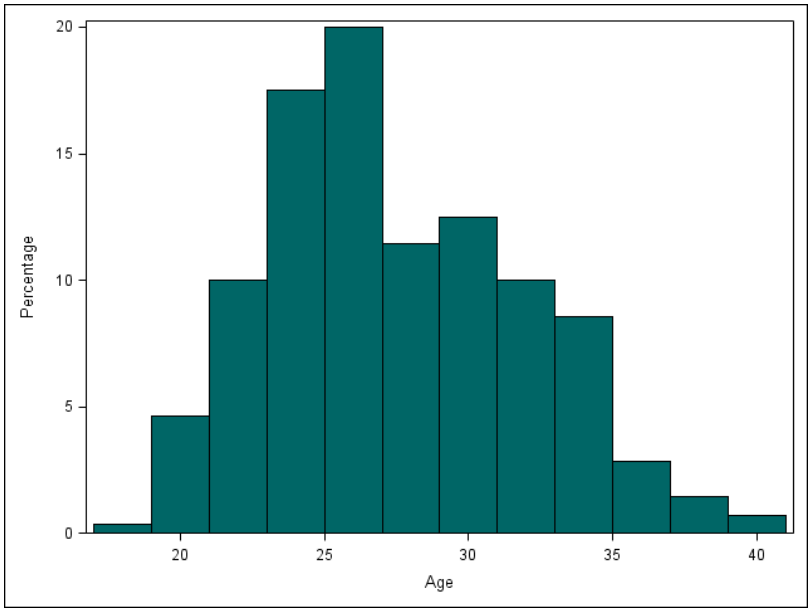

Figure 4 - Histogram of market values by age

In *figure 4* we can see that, as expected, the age is distributed following approximately a normal distribution with a peak around the 26/27 years old. There are almost no players younger than 20 years old or older than 40. The percentage decreases dramatically after the 35 years old threshold.

In *figure 5* the market value has been shown depending on the age of the players in a simple scatterplot. The graph shows that the football players with highest market values are all clustered in the range 24-30 years old, peaking at 25.

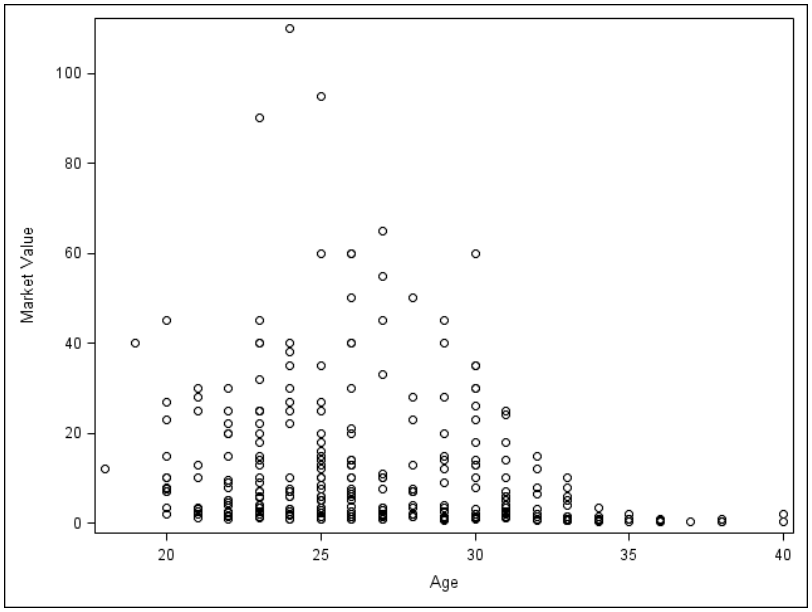

Figure 5 - Scatterplot of age and market value

It can be seen that there is not a linear relationship between the age and the market value of a football player as it will be increasing from the beginning of the career until it will reach a peak around 25 to 30 years old, probably depending on the maturity of performance of the player, and then it will start to decrease due to the short perspective of the career as the probability of retirement will increase dramatically (Frick, 2007).

4.4 Market Value and Rating

One of the most interesting variables in the Whoscored database is the "rating". It is a valuation of players' performances during the year. The ratings are calculated live during the game by a unique algorithm, based on the events that happen in the match. These events may have several outcomes that will impact positively or negatively the players' vote. The algorithm uses more than 200 raw statistics and each of them is weighted according to their specific influence on the game outcome (Whoscored.com, no date).

For example, an attempted dribble may have two outcomes: successful or unsuccessful. In the first case the attacking player will benefit of an increase in the rating weighted based on the position in the field, the closer to the opponents' goal the higher will be the increase. The defender on the other way will suffer of a decrease in the rating. In the second case the increase and decrease will be vice versa.

It is thus interesting to look at the relationship between the Whoscored ratings and the Transfermarkt market value estimations, to check whether there is a correlation between them. In *figure 6* the scatterplot of market value and rating is shown. Although we can see that there is a positive linear a trend for which players with higher ratings generally have higher market value, in the lower right zone we can find some outliers. These are players who, even if they had better performances than many other colleagues, are valued less. One of the reasons that may explain this phenomenon is the fact that players with different roles have differences in their economic value estimation.

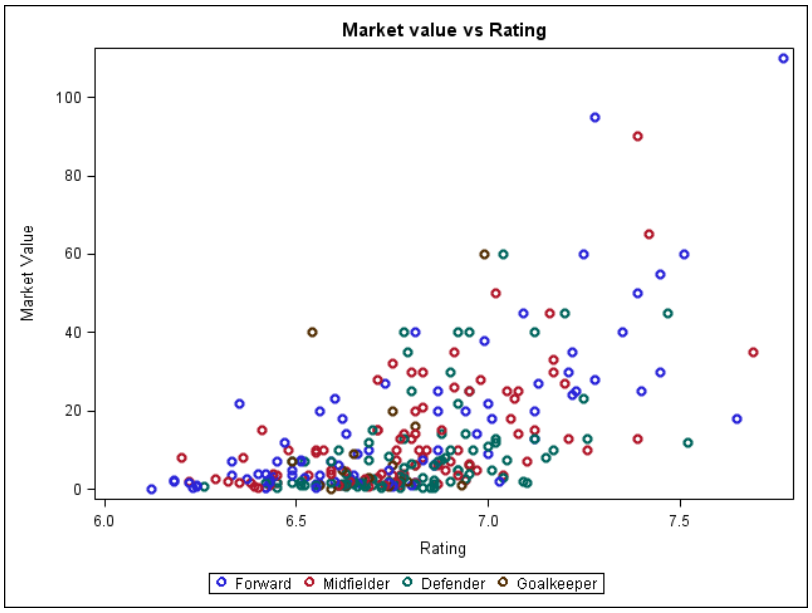

Figure 6 - Scatterplot of market value and rating grouped by role

Indeed, in the graph most of the players with good performances and low value belong to the categories of midfielder and defender, while the blue ones who are the forwards fit better the trend. The second reason is that, according to what was exposed before, there are other factors that impact on the market value such as the age.

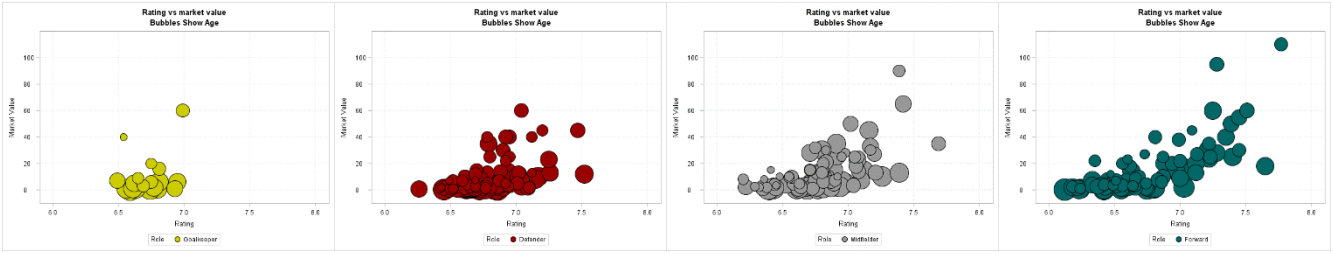

Figure 7 - Bubbleplot of rating and market value divided by role

The difference among roles is exploited better in *figure 7.* On the horizontal axis the rating is defined in a range between 6.0 and 8.0 as the minimum rating average is 6.12, scored by Sergio Pellissier and the maximum average rating is 7.77 scored by Paulo Dybala, while the market value on vertical axis ranges from 0 to 110 Mln €. The

goalkeeper table, the yellow one, shows the low numerosity of the group and few differentiations among the players. All the goalkeepers had a rating between 6.5 and 7.0 and almost all of them have a market value lower than 20 Mln €. The two exceptions who are valued 40 and 60 Mln € are respectively Donnarumma and Alisson, whose high value is also influenced by the young age represented by the small bubble.

For the field players, defender in red, midfielder in grey and forward in blue-green, we can see that the positive linear trend is maintained in all of them, but as mentioned before the slope of this trend looks different. The more we move to the right (from defender to forward) the steeper is the trend. Moreover, in all of the graph the higher valued players have generally low ages as showed by the small dimensions of the bubbles.

4.5 Defenders analysis

I will now dive deep into every specific role to better understand the main attributes that drive market values according to the position on the field.

The "Goalkeeper" category has a low numerosity as it is composed by 18 members, so I will no longer focus on this subset. Starting with the defender I decided to split the group in two: lateral and central defenders. All the centre-back have been considered as central while the left and right back have been considered as lateral defender. The *figure 8* shows that even in the same role different categories might be taken into account in judging a player. On the horizontal axis there is the number of total clearances done every 90 minutes. We can see that all the central players in red are positioned on the right part

of the graph, while all the lateral defenders are on the left part. It means that they cannot be evaluated both on this attribute.

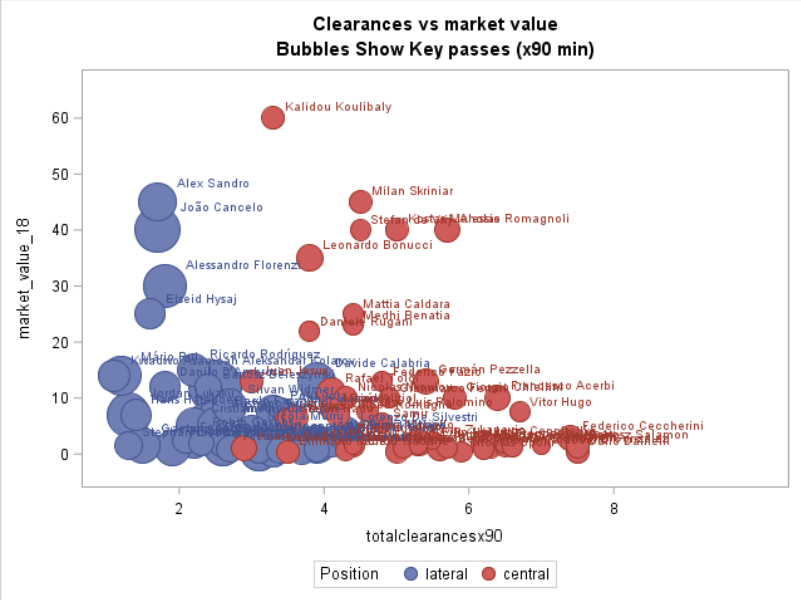

Figure 8 - Bubbleplot of defenders' clearances and market value divided by position

Thus, the dimension of the bubble has been defined as the total number of key passes every 90 minutes. This is because generally the lateral defenders play also an important in the offensive phase, creating chances for their teammates. The graph shows that there is a tendency to high evaluate lateral defenders with a high number of key passes, as the highest market values are attributed to the players with bigger bubbles on the left side.

Regarding the central defenders the market value does not seem having a great relationship with the total number of clearances, as the players who play in low ranking teams generally have to defend more and consequently have more chances to clear the ball.

4.6 Midfielders analysis

For the midfielder analysis, as the numerosity of the sample was very high and graphically showing the tendency with the bubble plot would have created a chaotic image, only the players with 2000 seasonal minutes played have been taken into consideration.

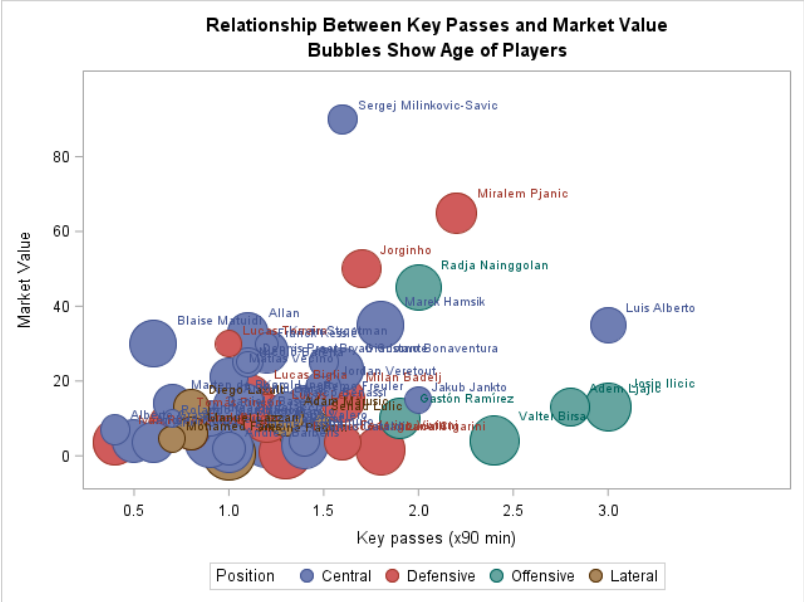

Figure 9 - Bubbleplot of midfielders number of key passes and market value grouped by position

The total number of key passes is the attribute positioned on the horizontal axis, as the main variable to analyse the potential of a midfielder. The bubbles dimension shows the age and the grouping has been made through the position occupied on the field: defensive, central, offensive and lateral. Although, as we could imagine, the players with higher number of key passes are the attacking midfielder on the right, they are not the most valuable players. It means that as the position on the field allows them to be more influential in the game, the number of chances created is weighted in a different way.

The lateral midfielder in the dataset are very few and are set in the low-left corner of the graphs.

The central and defensive midfielders are distributed close to a linear tendency that confirms the relationship between the number of key passes and the market value, on top of which we find Miralem Pjanic, the playmaker of Juventus. The player with the highest market value is Sergej Milinkovic Savic, the youngster of Lazio is an outlier of the tendency probably due to the high number of goals scored in the season (12) and to his young age (23).

4.7 Forwards analysis

Following the same reasoning applied to the midfielders, as the number of forwards is too high, I selected only the players who scored at least 5 goals during the season, to show only those players with significant skills. The right and left wingers have been grouped into the category “Winger” while the second striker and the centre-forward have been grouped into the “Striker” category.

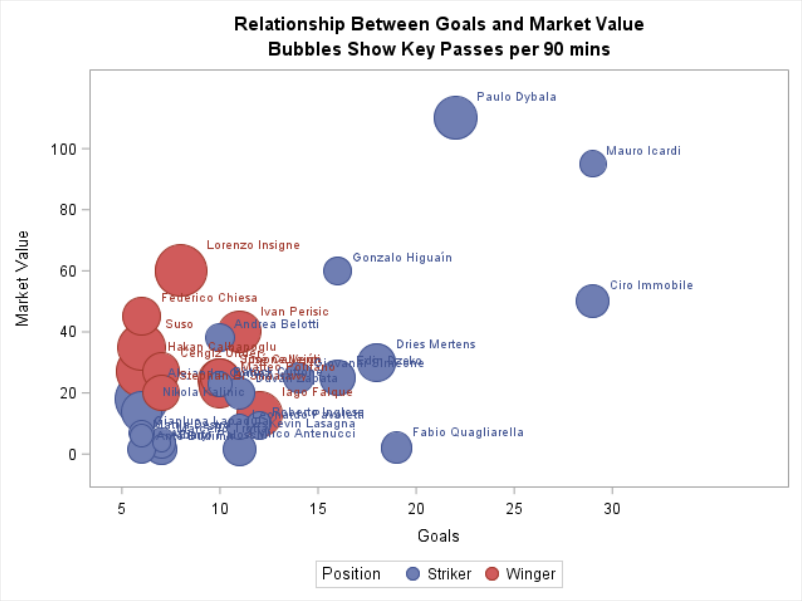

Figure 10 - Bubbleplot of number of goals by forwards and market value grouped by position

The number of goals is significant for evaluating the players who play in a central zone of the attack. The Striker category fits a linear tendency with some outliers Paulo Dybala and Mauro Icardi who are respectively 24 and 25 years old, being in their prime years of productivity. Regarding the wingers a different attribute must be taken into account. As they are players who generally play on the sides of the pitch, an important factor might be the number of chances they create for their strikers' teammates. The bubbles thus indicate the number of key passes in the 90 minutes, as it was used for the midfielders. We can note that generally the bigger the bubble the higher is the market value. In the following chapter the forwards will be studied deeper to understand if there is the possibility to create a linear model for estimating the market value.

5 Linear Model

After having analysed the interesting relationship between some specific football performances of every role and the market value, I will try to create a linear model for evaluating the players. The process will start with the analysis of the correlation matrix between all the variables and to understand which have to be deleted and which have to be considered in the final model.

5.1 Correlation analysis

In this first step the objective will be to understand the relationship between the numeric variables on the field and the market value. For this reason, all the non-numeric variables have not been taken into account. *Figure 11* shows the correlation matrix between the variables left. All the blue-green squares represent a highly positive correlation between the variables (>0.8) while all the gold squares represent a highly negative correlation (<-0.8).

Many of the variables highly correlated are actually a duplicable representation of the same attribute, as there are some performances that are collected whether per game played whether per 90 minutes. For example, the total tackles are collected over the season, then the same value is divided by the overall appearances to create the “tackles per game” indicator and by the minutes played and then multiplied by 90 to create the “tackles per 90 mins” indicator.

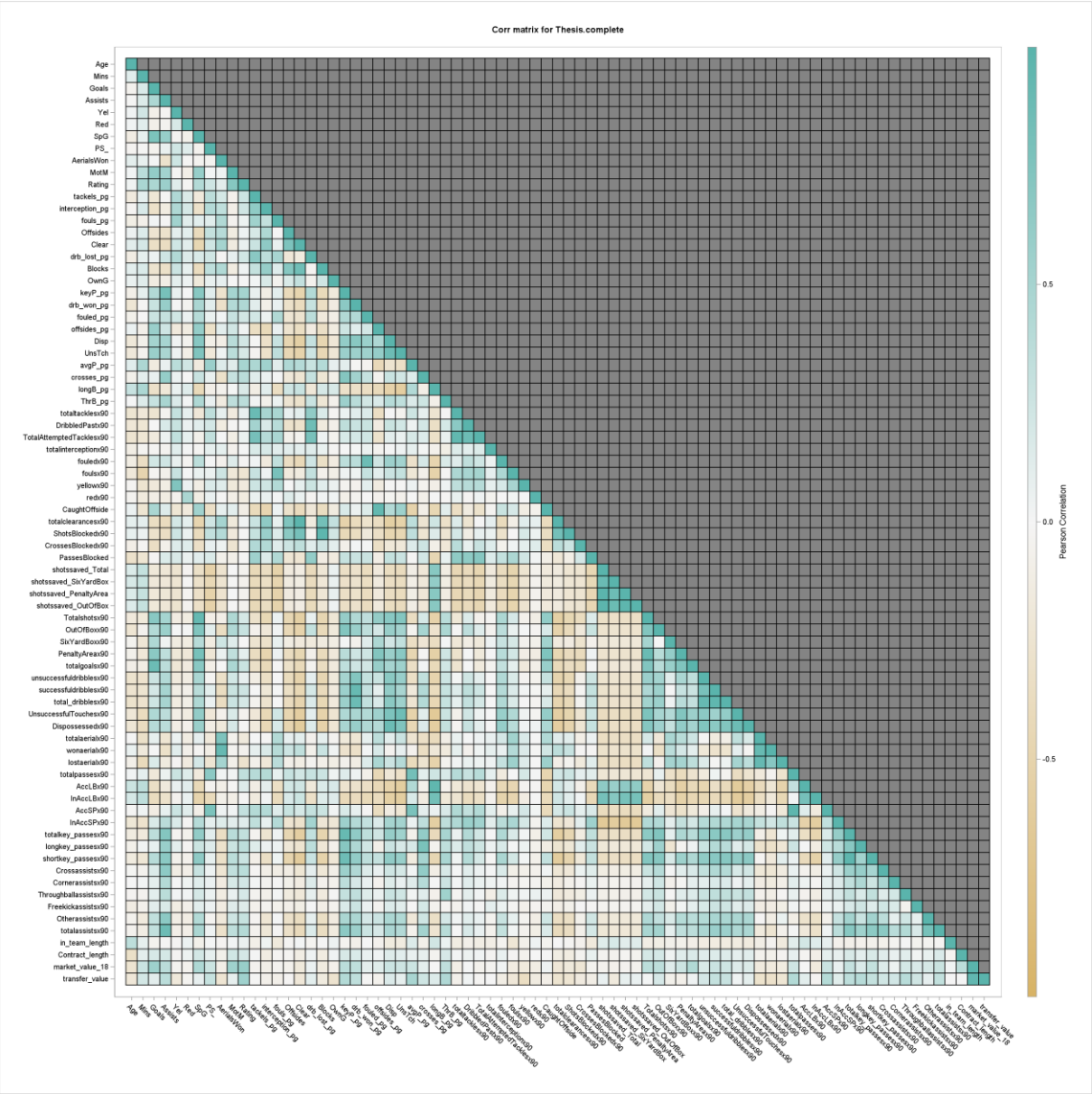

Figure 11 - Correlation matrix of all numeric variables

5.1.2 Variables drop

After having analysed all the correlations I decided to drop variables that were overrepresented and to keep only the necessary ones to prepare the model. Regarding all the variables that were represented both as "per game" and "per 90 mins" the general rule I followed was to keep the "per 90 mins", in order to consider the minutes played rather than the effective appearances as it would have created inequality in the judgment. The dataset has already been created by players who had a minimum number of

appearances, to avoid statistically incongruency. Thus, considering that not all the players play the whole match, it is more precise to consider the variables per 90 mins played except for the goals that will be explained in a few rows. In the following paragraph I will explain all the variables dropped and the reason behind it. The variables considered have all correlation higher than 0.8 or lower than -0.8.

Data dropped:

- *Goals and totalgoalsx90:* this is the only exception to the 90 mins rule. I decided to keep the overall number of goals as some players who play the final minutes of lot of matches might actually have higher totalgoalsx90 than the best scorer of Serie A.
- *keyP_pg totalkey_passesx90 assists and totalassistx90:* I took into consideration only the "x90" variables and I also decided to drop the total assist as a key passage is an assist but not vice versa. As a direct consequence I dropped all the variables related to the field position where the assist was made (Cornerassistsx90 Throughballassistsx90 Freekickassistsx90 Otherassistsx90 Crossassistsx90)
- *Yel and Yellowx90/ Red and Redx90:* Dropped Yel and Red
- *SPG and totalshotx90:* Dropped SPG (Shots per game)
- *Aerialswon and wonaerialsx90:* Dropped aerialswon
- *Motm and rating:* As the Motm (Man of the match) is chosen as the highest rating out of all the players' ratings of the match, I dropped the Motm.
- *Tackles_pg and totalatacklesx90:* Dropped Tackles_pg
- *Fouls_pg and foulsx90:* Dropped Foul_pg
- *Clear and totalclearancesx90 blocks and shotsblockedx90:* Dropped Clear and blocks and shotsblockedx90 as a shot blocked is a specific type of clearances
- *drb_lost_pg, drb_won_pg, successfuldribblesx90, dribbledpastx90, disp, dispossedx90 and total_dribblesx90:* I dropped the drb_lost_pg and the drb_won_pg, moreover as the dribble can only have two outcomes, either is

successful either is unsuccessful I decided to drop the total dribbles as it is a combination of the successfuldribblesx90 and dispossesedx90.

- *Totalattemptedtackles and totaltacklesx90:* Dropped Totalattemptedtackles
- *Fouled_pg vs fouledx90:* Dropped fouled_pg
- *offsides_pg vs caughtoffside(x90):* Dropped offsides_pg
- *dispossedx90 unsuccessfultouchesx90 and unstch:* I dropped the two variables related to the unsuccessful touches as they were a correlation of dribbles lost and wrong passes.
- *avgP_pg vs totalpassesx90:* Dropped avgP_pg
- *crosses_pg vs longkeypassesx90:* Dropped crosses_pg
- *longB_pg vs AccLbx90:* Dropped longB_pg
- *totalaerialx90 wonaerialx90 and lostaerialx90* -> Dropped totalaerialx90 as a combination of won aerial and lost aerial
- *Totalpassesx90 and AccSPx90:* Dropped AccSPx90
- *shortkey_passesx90 and total key passes:* Dropped shortkey_passesx90
- *Interception_pg and totalinterceptionx90:* Dropped Interception_pg

Once all the exceeded variables were removed the database counted of less than 50 variables, which is easier to manage and to understand which are the most important attributes to create a linear model. As I previously divided the dataset in four different categories, I will keep working with these roles and for every role I will now dive deep in their respective correlation matrix to individuate whether there are some variables to drop. For example, while for a goalkeeper the number of shots saved is a fundamental performance indicator, for the rest of the players that value is null.

5.1.2 Drop by roles

Goalkeeper

The goalkeeper matrix *(figure 13)* shows grey zones for Goals redx90 CaughtOffside ShotsBlockedx90 CrossesBlockedx90 Totalshotsx90 unsuccessfuldribblesx90 Dispossessedx90 lostaerialx90 shortkey_passesx90, which means that there are no values for any goalkeeper in these categories and thus have been eliminated. The final matrix is in appendix 3 *(figure 14).*

Defender

The defender matrix *(figure 15)* shows grey zones for Shotssaved_Total Shotssaved_SixYardBox Shotssaved_PenaltyArea Shotssaved_OutOfBox and thus they have been deleted. The final matrix is in appendix 3 *(figure 16).*

Midfielder

The midfielder matrix *(figure 17)* shows grey zones for the same categories as the defender so they have been deleted from this dataset as well. The final matrix is in appendix 3 *(figure 18).*

Forward

The forward matrix shows grey zones for the same as defender and midfielder and also for redx90. The final matrix is shown in *figure 12.*

As the forwards are the most valuable players as showed in the initial phases of this research, I will develop a linear model only for this category, leaving the others for future potential studies.

5.2 Forward model

5.2.1 Multiple regression linear model

The final step of the research is the analysis of the final correlation matrix for the category "Forward" to identify the variables that will be part of the regression model to estimate the market value. In *figure 12* the market value row has been highlighted in red and from there I selected all the variables with a significant correlation (higher than 0.3 or lower than -0.3).

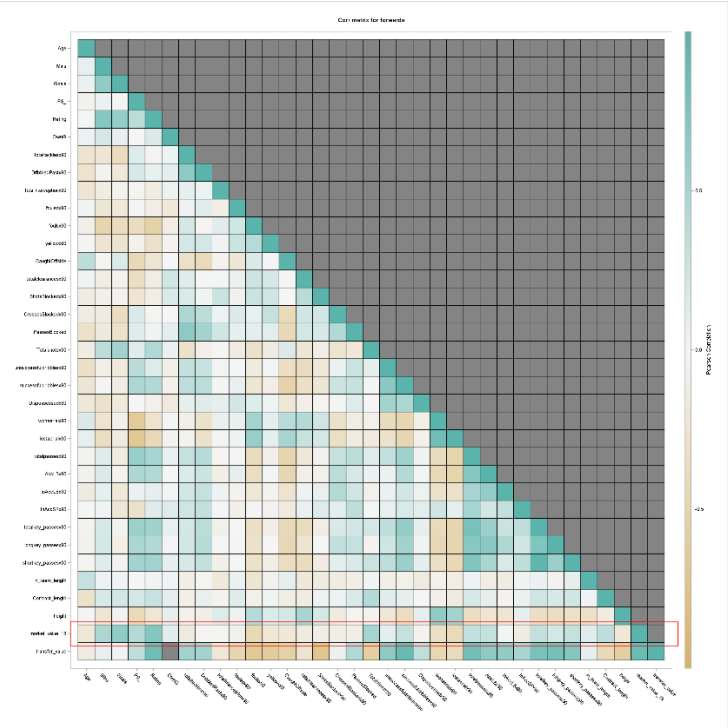

Figure 12 - Correlation matrix of final variables for forwards. Market value is highlited in red

I have selected 7 variables, 6 with a positive correlation and 1 with a negative correlation.

Positive

- Goals (0.65)
- Rating (0.73)
- Totalshotsx90 (0.51)
- Successfuldribblesx90(0.32)
- Totalkey_Passes (0.33)
- Contract_length(0.37)

Negative

- Lostaerialx90 (-0.39)

The rating is a value calculated by an algorithm based on the performances present in the dataset and therefore it has been removed from the model, in order to have a cleaner understanding of the data. In table 1 is presented the linear multiple regression output. In table 1 the star denotes a significant variable when α=5%, where α is the probability of type-1 error.

Parameter	**Estimate**	**Pr > |t|**	**α=5%**	**α=10%**
Intercept	-15.53696490	0.0304	*	*
Goals	2.02462982	<.0001	*	*
Totalshotsx90	2.93175610	0.2190		
successfuldribblesx9	5.47946662	0.0034	*	*
totalkey_passesx90	1.59128609	0.5928		
Contract_length	1.84934746	0.0967		*
lostaerialx90	-1.46286548	0.1756		

Table 1 - Multiple linear regression model factors

Thus, considering only significant variables, a striker will be positively evaluated if he scores more goals and completes successfully more dribbles. This model can explain 64% of the total variance (R= 0.64) which is sufficiently high.

5.2.2 Stepwise model

The stepwise method is a sequential procedure that aims at maximizing the line fit with the data by selecting automatically the variables. In every step of the algorithm is chosen whether every single variable has to enter or exit the model. It is a combination of backward and forward selection techniques. “Stepwise regression is a modification of the

forward selection so that after each step in which a variable was added, all candidate variables in the model are checked to see if their significance has been reduced below the specified tolerance level. If a nonsignificant variable is found, it is removed from the model" (Statistical and Ncss, no date).

Parameter	**Estimate**	**Pr > |t|**
Intercept	-8.45262	0.1068
Goals	2.19091	<.0001
successfuldribblesx9	6.01138	0.0002
Contract_length	2.49902	0.0157
lostaerialx90	-1.69358	0.1043

Table 2 - First Stepwise output

In table 2 the final outcome of the stepwise method is exposed. The variables are less than the previous technique, the R-square is 0.63 and the adj R-square is 0.61.

At the beginning of the chapter I excluded the variable "rating" as it was calculated through the other variables included in the dataset. In order to have an interesting comparison I executed a Stepwise procedure with the rating, to see if it would have impacted the selection of the variables. The following table (table 3*)* shows the outcome of the stepwise regression with the rating included.

.

Parameter	Estimate	Pr > |t|
Intercept	-115.29	0.0044
Goals	1.58	<.0001
Rating	16.28	0.0122
Contract_length	2.16	0.0177

successfuldribblesx9	4.27	0.0340

Table 3 - Second Stepwise output

The negative impact of the variable lostaerialx90 has been removed by the algorithm that has inserted the rating on the contrary. The R-square is 0.64 and the adj R-square is 0.63, slightly higher than before.

In the next chapter I will discuss the differences in the outcome of the three procedures.

6 Model Interpretation

6.1 Transfer value comparison

The final estimated market value of every player has been compared with the transfer value. Unfortunately, only 11 players remained in our dataset have been transferred in the summer of 2018, therefore I used only the estimation of these values to understand if the regression generated a better or worse estimation than Transfermarkt procedure. I thus calculated the square difference by the values, summed and then applied the square root to understand the average distance from every model by the actual transfer value.

	Transfermarkt	Regression	Stepwise	Stepwise rtg
Avg distance	6.47288	7.59869	7.098217	8.232085

Table 4 - Comparison of average distance from evaluation and real market value

The lowest value is the Transfermarkt one, which is the most accurate according to this comparison. As expected, the stepwise is more precise than the linear regression but if the rating is included it biases the calculation of the fit creating a worst model. The difference between the Stepwise model and the Transfermarkt estimation is quite low anyway, differing by only 0.5 M€ on average.

Most overvalued and undervalued players

Using the Transfermarkt value estimation as a proxy for the market value, I compared the estimated value with the Stepwise method to understand which are the most overvalued and undervalued players.

The most overvalued player is Paulo Dybala. Indeed, while Transfermarkt.com valued him at 110 Mln € in June 2018, the regression would state a value at 71 mln €, generating a difference of almost 40 Mln €.

On the other side the most undervalued player is Fabio Quagliarella. The 35 years old Sampdoria striker had a golden season in 2017/2018 and his performances granted him a 40 Mln € valuation, against the 2 Mln € stated by the German website Transfermarkt.

The main reason that could explain such a difference in both the evaluations is that I did not take into account the age in the determinants factor as it did not have a linear correlation with the market value. Dybala's value is indeed increased by the fact that he is just 24, in his prime of his years, while Quagliarella is getting close to the end of his career as he was 35 in the time period taken into consideration. In the future developments one of the objectives will be to introduce the age as a factor.

6.2 Model limitations

The collected data are not sufficient to evaluate all the roles and different positions on the field, as they might be affected by different factors. Collecting only data from the Serie A are a limitation for the dimension of the dataset. Moreover, I collected only data related to one season, while it could be helpful collecting several following football seasons to understand the potential trends. On the other hand, the availability of such data is not always free and easy to access.

Football is nonetheless a very interactive and team-based game. The evaluation of the market player based only on his performances considered in absolute terms, without taking into consideration the impact on the final result or on his teammates' performances, will always be biased and create inaccuracy in final evaluation.

7 Data automation

The proposed model limitations can be partially overcome by enriching the dataset with more data. Two data dimensions can be improved: temporal, including more seasons, and geographical, including more leagues.

The aim of this second part of the work is to find an easy and scalable way to first of all access and download the data and secondly to revise and clean these data to create a proper dataset. The objective is to work with the 5 main European leagues over the last 5 years, from season 13/14 to season 18/19.

7.1 Data scraping automation

I focused the attention of the second part of this work in trying to create a scalable algorithm that could fetch data from Whoscored.com and Transfermarkt.com. The choice was consistent with the idea of creating an end to end analysis, from dataset creation through data cleaning and finally data mining and analytics.

The tool used to create the RPA algorithm to automate the data extraction is UiPath Studio, that is free to use and of ease interpretation and learning. It uses "activities" that enable specific code to complete different actions.

The difficulties met during this phase were mostly related to the complicated definition of the websites.

7.2 Whoscored scraping

Whoscored contains all the seasonal player and team statics, divided in 5 macro areas:

- Summary *(generic statistics of the season)*
- Offensive *(statistics regarding attack phase)*
- Defensive *(statistics regarding defensive phase)*
- Passing *(statistics regarding passing skills)*

- Detailed *(divided in 16 tables of specific data, i.e. passes inside penalty area)*

The first step was to create a loop that could iterate on all the leagues needed. Leagues are defined on top bar navigation of the website and are inside html tags named upon the league represented.

The “For Each” activity was defined looping on an array of strings, containing the names of the leagues: “Serie A, Bundesliga, Ligue 1, Premier League and La Liga”. This process is very scalable as all the code within the loop can be applied to any league selected. In order to add a specific league to the dataset it would be sufficient manual update of the array, inserting the new league name and then waiting for the algorithm to complete all the steps.

Once we enter the loop the first activity is the “open browser”, which will redirect to the homepage of Whoscored. Here the selector will take the variable assigned to the iteration, hence the league name and click on it.

Inside the league page, the algorithm scrapes the years drop menu selector, and saves all the seasons founded in a specific array. This means that if we would want to use the same algorithm next year, it would take into account even the new year without any modification. Now a new “For Each” node creates a loop on every year, starting from “year-1” thus avoiding season 2019/2020 as it is just started and would be an outlier in the analysis.

Then I inserted a break after the season “2014/2015” because I did not want to scrape more than 5 years backward, but it would be simple to gather data until year 2000, when the Whoscored.com dataset starts.

Within the selection of the year, the excel to export all the data for that specific season and league is created, concatenating the league name and the first year of the season (i.e. Premier League 2018/2019 becomes Premier league_2018.xlsx).

The excel name will be useful during the data cleaning and preparation since in the dataset is not present neither the league neither the season, I had to implement the columns by using the reference excel name.

After creating the excel, the algorithm will click on the “player statistics” tab, opening the final web page to scrape the data. Here the automation is split into two paths.

The first one iterates on the first 4 statistics tabs as described before (Summary, Offensive, Defensive and Passing), scrape the 10 rows of data, clicking on next and iterating it for all the players in any of the categories. Then will add an excel sheet to the excel opened before with the name of the specific tab scraped.

The second one enters the detailed page that has a different layout then previous one. It selects the option “per 90 mins” showing all the statistics divided by the minutes played by the player and multiplied times 90 mins, to normalize the numbers across a standard match.

Then selects and saves into an array all the tabs in the detailed drop-down menu and scrapes all of them adding as many sheets as the variables. Overall the final excel file will contain 16 tabs, each with almost 500 rows. Everything multiplied times 5 years and 5 leagues.

The Whoscored scraping is now complete and the row statistics dataset is pulled, with a scalable process both in terms of temporal horizon and geographical diversity.

7.3 Transfermarkt scraping

The next step is to pull the market value data from Transfermarkt.com. I used UiPath Studio again, reapplying partially the activities used and customizing them for the new website. The data used from this data set to include in the Whoscored are the market value, the height, preferred foot, the team, the age and the contract length.

The data are nested into a higher level as all the players are shown in the “team” page. On the other side there is only one table for every team, which means the final excel composition will be a single excel named upon the league and the season as per previous procedure but with each sheet inside associated to a specific team.

Transfermarkt.com has a different html base, making the reapplication of same procedure impossible and prone to errors. Thus, I decided to work with the URLs, as in this case they are created dynamically with the same path. I stored the URL for the 5 leagues on which I am interested in an array and looped through them.

Then I stored the years of the seasons I am interested in and looped on them as well. Once inside the loop, I created the final URL and the excel sheet, concatenating general location for the league plus the year at the end.

As mentioned before the players are stored inside the team but every year teams can be relegated to inferior leagues or promoted to superior leagues. Thus, the composition of a league cannot be the same across the time. Therefore, I decided to scrape the name of the teams from the league and store them into an array.

After this I implemented a new loop on every team and within the team, I scraped the data needed and stored into a sheet named upon the team. When all team are analysed the algorithm steps into a new year and then into a new league.

Dataset is now row and ready to be cleaned and prepared for the analytical part.

7.4 Data preparation

Now all the data have been stored into 50 excel files each one containing 15 tabs. The next step is to group all of these into a single ordered database. To do so I

used another free and open source tool called Knime. It is single and easy environment to crunch, clean, prepare and analyse data.

The interface is quite intuitive with pre-coded nodes that enables activities or java snippets for customized necessities. Firstly, I had to extract all the excel files and sheets, therefore I used the List Files node, which takes all the files' location name inside a specific folder. Then the "Table Row to Loop Variable" allowed the transformation of the location names from table into a specific variable, over which I iterated. The variable is passed every iteration to the "Read Excel Sheet" node that receives as input the location name and returns a table with all the sheets name inside. Again a "Table Row to Loop Variable" node transforms the table into a variable, creating a loop inside the loop. Now we are reading the data inside the excel sheet and the "Excel Reader Node" can be used to import them into a Knime table.

The duplicate row filter removes any potential row duplicated during the data extraction, then the proper cleaning data phase begins. Since the excel sheets inside the same excel file contain the same players, the loop inside has to append the data to previous iteration, joining based on the same player name. On the second loop instead, the data will have to be added at the end of the dataset as every iteration will have the same columns.

Inside a "Column List Loop" node is used to iterate on every column and add the name of the sheet at the end in order to understand every statistic to which macro type is referred. For instance, we will have the statistic "penalty_area" both in shots and block saved thus the final outcome will be penalty_area_shots and penalty_area_blocksaved.

Finally, I had to define the "Row ID" on which the "loop end column appends" node will join the several tables. As key joiner I decided to use the player name, which is composed by the name and surname of the player, the team where he played during that season, the age and the role. All of these generate a quite robust specific definition of the observation and it can be used as our primary key.

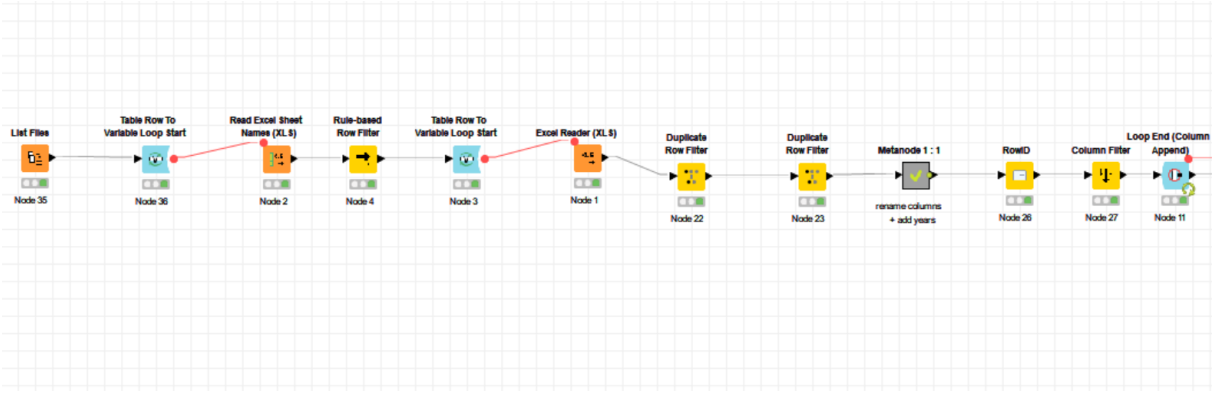

Figure 133 – Transfermarkt workflow

When the inner loop is concluded, I use the "Constant Column" node to add the season column, taking the number from the excel name and I filter out all the columns that refers to the same value. For instance, in every excel sheet is present the "rating" of the player, repeating this data 16 times for every player.

Now the loop is finished, and we can work on the whole Whoscored dataset to clean some remaining dirty data. Specifically, the "player name" as mentioned before includes data that we do not need any longer. Here an example of a data before cleaning it: *"Arjen RobbenBayern Munich, 35, M(CLR),FW"* as said before the name and surname precedes the team name, separated then by the age and the roles by commas.

It was quite easy to separate name, age and roles through a "cell splitter" node, using the comma as string delimiter but, as the name of the player can be composed by one, two or more names, using the uppercase as a character delimiter to separate player and team name was not an option. Therefore I had to create a workaround, I used again UiPath to quickly scrape all the teams name from whoscored and I uploaded the list in the Knime workflow.

Then I used a "Recursive" loop to iterate on every team and on every player and use a string replacer to remove the team name from the player name. The outcome will be

only the player name, which is not a loss of data as age, team and position will be taken from the Transfermakrt dataset: *"Arjen Robben"*.

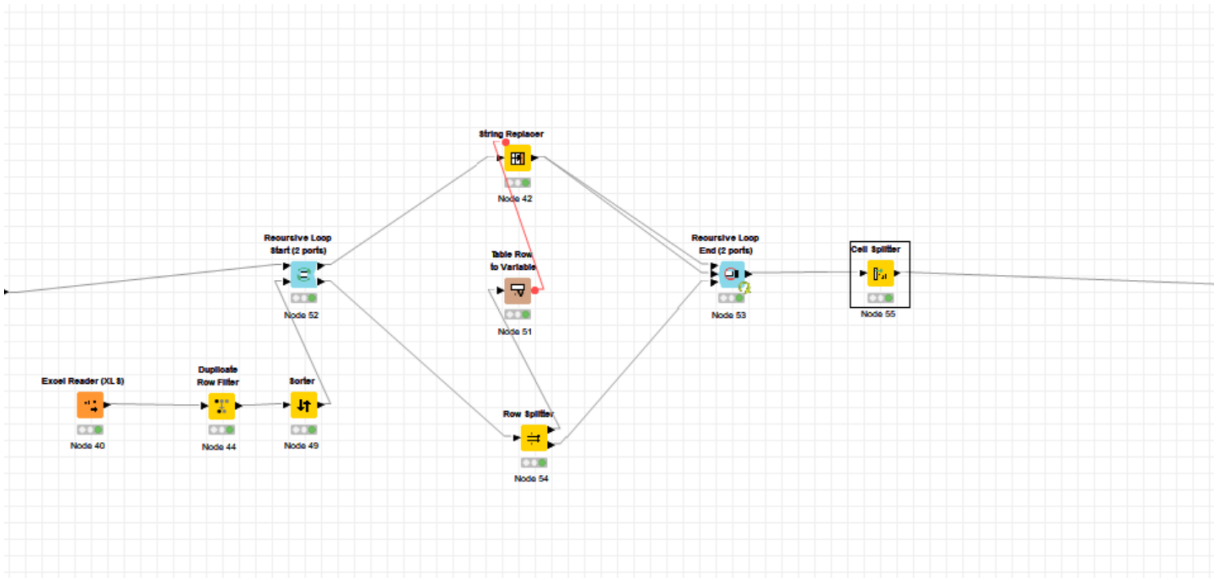

Figure 144 – Player name cleaning

The Whoscored dataset is ready, in parallel I had to apply similar rules to the Transfermarkt dataset. The first part has the same routine, reading the locations inside the folder and then the excel sheet names. Looping on those and creating a unified dataset.

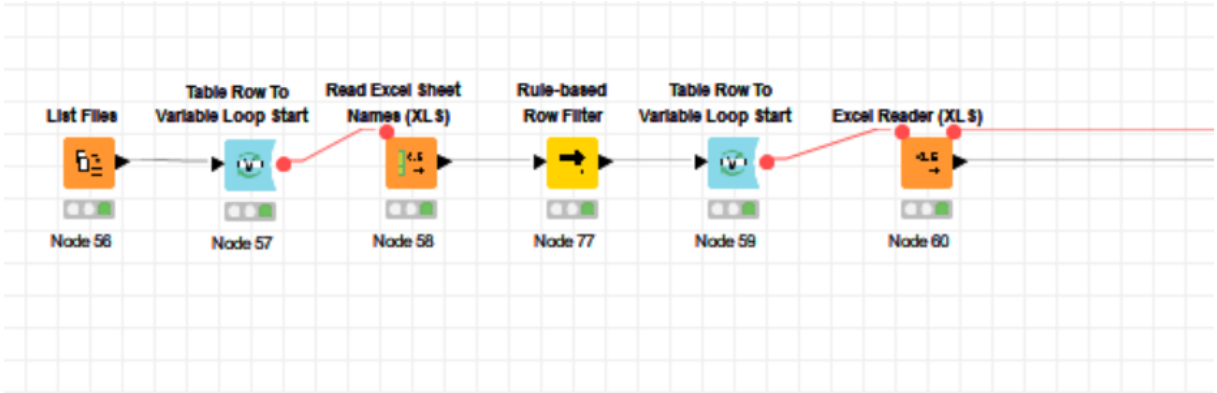

Figure 155 – Whoscored workflow

The second part has to deal with the most important data, which is the market value. It is stored as a string, and defined of two types, Millions of euros or Thousands of euros.

In the first case the last characters will be the “mil. €” in the second case will be the “k €”. Thus a “java snippet” node was necessary to modify the string into a specific number and then convert it into a double, based on Millions of euros. First the value would have been “16,00 mil. €” or “500 K €” and the output would be “16” or “0.5”.

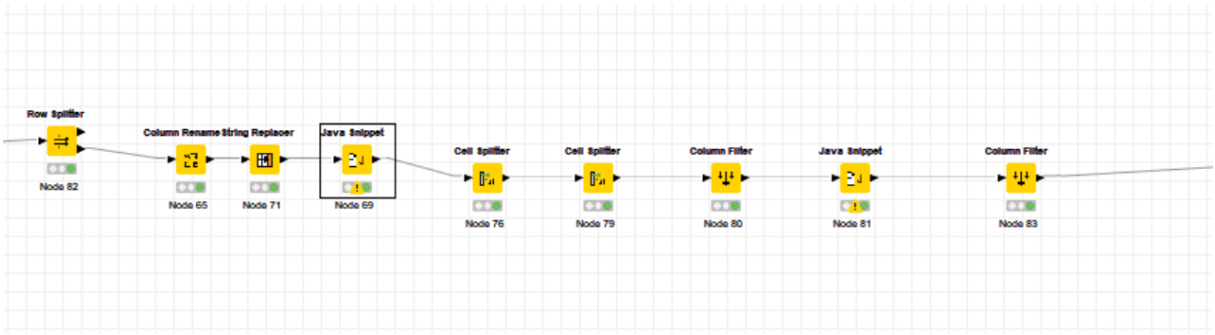

Figure 166 – Market Value extraction

Now both datasets are ready and is time to merge them, the “joiner” node operates on two key variables, the player name and the season. Then all columns that are not needed are removed and the final dataset is composed by more than 7000 rows and 54 columns.

The dataset composition is similar to the one created for the first analysis, plus the league name and the season.

8 Final model and conclusions

8.1 Final model

The new dataset includes more observations from different seasons and different leagues, which is a good way to increase the dimension but at the same time data will have higher diversity.

In order to have a comparable outcome the model will be specific to forwards, as already defined in previous chapter.

To create the new model, I took inspiration from the specific case that was presented before on Serie A 2017/2018. Nonetheless I applied the "linear correlation" node to calculate correlation within variables and especially between variables and the market value. Some of the data collected are not exactly the same. The contract length indeed could not be calculated as we did in the specific case because Transfermarkt only stores actual contract expiration date and not the one related to the season in the past.

Market Value vs Variable	**Correlation**
Goals_Summary	0.56
Total_Shots	0.34
Total_Key passes	0.23
Age	-0.11
UnsuccessfulTouches_Possession loss	-0.17

Table 5

The choice of including the age is due to the outcome of previous regression, whereas the UnsuccessfhoulTouches replaces a less complete variable as the "lost aerial". Contract length as explained before was impossible to calculate. The other variables were already included in previous model.

I started the analysis by checking if these variables were solid both in terms of season diversity and league diversity.

8.1.1 League Analysis

Across the same year 2017-2018, I checked whether the linear regression with these variables would have been a good fit for data for every specific league.

Ligue 1:

Parameter	**Estimate**	**Pr > |t|**

Age	-1.55	0.05
Goals_Summary	1.32	0.04
Total_Key Passes	11.20	0.00
Total_Shots	10.05	0.03
UnsuccessfulTouches_Possession loss	-6.47	0.04
Intercept	29.56	0.21

Bundesliga:

Parameter	**Estimate**	**Pr > |t|**
Age	-1.77	0.00
Goals_Summary	1.36	0.00
Total_Key Passes	13.18	0.00
Total_Shots	-0.29	0.84
UnsuccessfulTouches_Possession loss	2.26	0.18
Intercept	28.10	0.00

La Liga:

Parameter	**Estimate**	**Pr > |t|**
Age	-2.15	0.00
Goals_Summary	3.83	0.00
Total_Key Passes	15.41	0.00
Total_Shots	6.01	0.02
UnsuccessfulTouches_Possession loss	-1.45	0.28
Intercept	24.21	0.08

Premier League:

Parameter	Estimate	Pr > |t|
Age	-1.84	0.00
Goals_Summary	1.53	0.00
Total_Key Passes	10.29	0.01
Total_Shots	11.15	0.01
UnsuccessfulTouches_Possession loss	-9.91	0.00
Intercept	39.26	0.05

Serie A:

Parameter	Estimate	Pr > |t|
Age	-0.99	0.00
Goals_Summary	1.55	0.00
Total_Key Passes	2.03	0.43
Total_Shots	5.02	0.03
UnsuccessfulTouches_Possession loss	-4.34	0.00
Intercept	24.58	0.03

All these data are limited to the season 2017/2018. We can see that the model is consistent in every league, with few outlier exceptions for the t-value as the total shots in Bundesliga and the key passes in Serie A, which are probably related to data misses.

Generally, the trend is always the same, the intercept is set around 25 and age and unsuccessful touches affect the value negatively, which is logical thinking about a player that loses several balls will have lower value and at the same time an old player will lose his market value due to the loss of potential future returns.

On the other side the number of goals, the total shots and the key passes increase the value of the forward.

Interesting to notice that the intercept for "La Liga" is set to a higher base, close to 40 Mln €. This might be driven by the presence of Messi and Ronaldo during most of the seasons that increased the average value of La Liga players.

8.1.2 Season Analysis

I repeated the analysis using only the Serie A league and iterating the correlation on all the seasons.

2014/2015:

Parameter	**Estimate**	**Pr > |t|**
Age	-0.76	0.00
Goals_Summary	0.50	0.02
Total_Key Passes	1.83	0.36
Total_Shots	1.92	0.17
UnsuccessfulTouches_Possession loss	-2.27	0.21
Intercept	23.16	0.01

2015/2016:

Parameter	**Estimate**	**Pr > |t|**
Age	-0.67	0.00
Goals_Summary	1.33	0.00
Total_Key Passes	3.13	0.06
Total_Shots	0.98	0.30
UnsuccessfulTouches_Possession loss	-2.05	0.08
Intercept	16.39	0.03

2016/2017:

Parameter	**Estimate**	**Pr > |t|**
Age	-0.84	0.02
Goals_Summary	1.32	0.00
Total_Key Passes	6.12	0.00
Total_Shots	1.33	0.53
UnsuccessfulTouches_Possession loss	-3.26	0.12
Intercept	19.67	0.13

2017/2018:

Parameter	**Estimate**	**Pr > |t|**
Age	-0.99	0.00
Goals_Summary	1.55	0.00
Total_Key Passes	2.03	0.43
Total_Shots	5.02	0.03
UnsuccessfulTouches_Possession loss	-4.34	0.00
Intercept	24.58	0.03

2018/2019:

Parameter	**Estimate**	**Pr > |t|**
Age	-1.39	0.00
Goals_Summary	0.74	0.18
Total_Key Passes	7.74	0.04
Total_Shots	3.83	0.21
UnsuccessfulTouches_Possession loss	-3.79	0.17
Intercept	38.60	0.02

Season diversity is a little bit more emphasised, as several values go above the 0.1 threshold in different seasons. I will try to create boolean variables to be used in the

linear correlation both for seasons and leagues to control whether they would result in being significant for model explanation.

8.1.3 Including Seasons and Leagues

I created 4 variables (n-1) which will be set to 0 or 1 in case the player is playing in the respective league. The comparison is chosen with Serie A, therefore in case all of the variables would be set to 0, it would mean that the player was currently playing in the Italian league.

Correlation with Leagues:

Parameter	**Estimate**	**Pr > |t|**
age	-1.20	0.00
Goals_Summary	1.90	0.00
Total_Key passes	7.46	0.00
Total_Shots	1.42	0.01
UnsuccessfulTouches_Possession loss	-2.64	0.00
ligue1	1.58	0.28
la_liga	3.21	0.01
premier	6.84	0.00
bundesliga	-3.45	0.00
Intercept	23.21	0.00

Almost all of the added variables have very high p-value, except for the French league which has 0.28. The estimate shows that they are all positive respect to the Italian league, with high difference for premier league, almost 7 times Italian value. The only negative related variable is Bundesliga, implying a lower value of players in the league compared to Italian ones.

After comparing leagues, I wanted to check whether there is a positive trend across season in player evaluation, as we have seen in the introduction.

Correlation with Seasons:

Parameter	Estimate	Pr > |t|
age	-1.25	0.00
Goals_Summary	1.95	0.00
Total_Key passes	7.85	0.00
Total_Shots	1.21	0.02
UnsuccessfulTouches_Possession loss	-3.26	0.00
2014	-9.05	0.00
2015	-8.68	0.00
2016	-5.82	0.00
2018	2.60	0.03
Intercept	32.03	0.00

The season set as reference is the used during the first analysis, the 2017/2018. We can see that the precedent seasons are all negatively correlated while the following season is positively correlated. Moreover, they are all highly significant with p-value lower than 5%. This confirms the initial hypothesis that the market value is increasing year over year, with an average increase of almost 3 M€.

Both of the variables seem to be significant, therefore I added them to the final linear model.

8.2 Conclusions

I finally wrote the code to extract the final model, including all the significant variables.

Correlation with Seasons:

Parameter	**Estimate**	**Pr > |t|**
age	-1.24	0.00
Goals_Summary	1.89	0.00
Total_Key passes	7.34	0.00
Total_Shots	1.65	0.00
UnsuccessfulTouches_Possession loss	-3.57	0.00
ligue1	0.81	0.56
la_liga	3.52	0.00
premier	7.23	0.00
bundesliga	-3.95	0.00
2014	-9.13	0.00
2015	-8.66	0.00
2016	-5.74	0.00
2018	3.18	0.01
Intercept	29.93	0.00

This is the final model, all the variables have p-values lower than 5% and most of them close to 0 except for Ligue 1. The negative correlated variables are the age, possession loss, Bundesliga and the antecedent seasons. Goals scored, key passes, future seasons and the other leagues are positively correlated, while the intercept is equal to 29.93.

Metric	**Value**
R^2	0.51
mean absolute error	9.74
mean squared error	207.66
root mean squared deviation	14.41
mean signed difference	0.00

The R-squared shows that the model explains 51% of dataset variability, which is a good result but leaves room for improvement, while the mean absolut error is 9.74, almost 2 points higher than the results from the single season analysis.

The model shows that there is difference of evaluation among leagues, players with similar statistics have higher value in Premier League rather than Italy. The value is therefore not only influenced by performances on the field, as the goals scored or the key passes, it also includes players' personal data as the age or external information as the team in which he plays and the year of evaluation.

8.3 Future studies

The model can evaluate forwards who played during last 5 years in one of the 5 best European leagues. Even if 5 years might look as a sufficient amount of observation, player career can last up to 15-20 years. The dataset could be extended both on temporal dimension and league horizon. Moreover, the model could be enlarged affecting players of different roles, highlighting statistic that were not taken into consideration using forwards as group.

Now that the effort of data extraction and data cleaning have been completed, it could be very interesting also focusing on different statistical approach to predict the market value. For instance, using a regression tree might optimize the estimation of the variables used.

The dataset used during this thesis will be open and accessible to researchers who might be interested in further development of the topic.

Bibliography

Frick, B. (2007) 'The football players' labor market: empirical evidance from the major european leagues', 54(3), pp. 422–446.

He, M. (2013) 'Football Player ' s Performance and Market Value', pp. 1–82.

He, Y. (2014) 'Predicting Market Value of Soccer Players Using Linear Modeling Techniques', *Stat.Berkeley.Edu*, pp. 1–15. Available at: http://www.stat.berkeley.edu/~aldous/Research/Ugrad/Yuan_He.pdf.

Herm, S., Callsen-Bracker, H. M. and Kreis, H. (2014) 'When the crowd evaluates soccer players' market values: Accuracy and evaluation attributes of an online community', *Sport Management Review*. Sport Management Association of Australia and New Zealand, 17(4), pp. 484–492. doi: 10.1016/j.smr.2013.12.006.

Hughes, M. *et al.* (2012) 'Moneyball and soccer - an analysis of the key performance indicators of elite male soccer players by position', 7(2), pp. 402–412. doi: 10.4100/jhse.2012.72.06.

Kidd, R. (2018) *Soccer's Moneyball Moment: How Enhanced Analytics Are Changing The Game*, *Forbes*. Available at: https://www.forbes.com/sites/robertkidd/2018/11/19/soccers-moneyball-moment-how-

enhanced-analytics-are-changing-the-game/#4fccdd0576b2 (Accessed: 5 April 2019).

KPMG (2018) *Player Valuation: Putting data to work on transfer market analysis*, *KPMG Football Benchmark*. Available at: https://www.footballbenchmark.com/library/player_valuation_putting_data_to_work_on_transfer_market_analysis (Accessed: 2 April 2019).

Lozano, F. J. M. and Carrasco Gallego, A. (2011) 'Deficits of accounting in the valuation of rights to exploit the performance of professional players in football clubs. A case study', *Journal of Management Control*, 22(3), pp. 335–357. doi: 10.1007/s00187-011-0135-6.

Majewski, S. (2016) 'Identification of Factors Determining Market Value of the Most Valuable Football Players', *Journal of Management and Business Administration. Central Europe*, 24(3), pp. 91–104. doi: 10.7206/jmba.ce.2450-7814.177.

Marr, B. (2015) *How Big Data and Analytics are Changing Soccer*, *Linkedin*. Available at: https://www.linkedin.com/pulse/how-big-data-analytics-changing-soccer-bernard-marr/ (Accessed: 6 April 2019).

McHale, I. G. and Relton, S. D. (2018) 'Identifying key players in soccer teams using network analysis and pass difficulty', *European Journal of Operational Research*. Elsevier B.V., 268(1), pp. 339–347. doi: 10.1016/j.ejor.2018.01.018.

Memmert, D. and Rein, R. (2018) 'Match analysis, Big Data and tactics: current trends in elite soccer', *Deutsche Zeitschrift für Sportmedizin*, 2018(03), pp. 65–72. doi:

10.5960/dzsm.2018.322.

Müller, O., Simons, A. and Weinmann, M. (2017) 'Beyond crowd judgments: Data-driven estimation of market value in association football', *European Journal of Operational Research*, 263(2), pp. 611–624. doi: 10.1016/j.ejor.2017.05.005.

Rein, R. and Memmert, D. (2016) 'Big data and tactical analysis in elite soccer: future challenges and opportunities for sports science', *SpringerPlus*. Springer International Publishing, 5(1). doi: 10.1186/s40064-016-3108-2.

Rosen, S. (1981) 'The Economics of Superstars', *American Economic Review*, 52(4), pp. 845–858.

Statistical, N. and Ncss, S. (no date) 'Stepwise Regression', pp. 1–9.

UEFA (2019) *Uefa Club Coefficients*. Available at: https://www.uefa.com/memberassociations/uefarankings/index.html (Accessed: 5 April 2019).

Whoscored.com (no date) *WhoScored Ratings Explained*. Available at: https://www.whoscored.com/Explanations (Accessed: 9 April 2019).

Zhang, S., Zhang, C. and Yang, Q. (2003) 'Data preparation for data mining', *Applied Artificial Intelligence*. Taylor & Francis, 17(5–6), pp. 375–381. doi: 10.1080/713827180.

Table of figures

List of tables

Appendix 1

Variable	Explanation
Player	Name and surname of the player
Age	Age of the player
Position1	First position on the field according to Whoscored
Position2	Second position on the field according to Whoscored
Position3	Third position on the field according to Whoscored
Apps	Total number of appearances
Mins	Total number of minutes played
Goals	Total number of goals
Assists	Total number of assists
Yel	Total number of yellow cards
Red	Total number of red cards
SpG	Number of shots per game
PS_	Percentage of successful passes
AerialsWon	Number of aerial tackles won per game
MotM	Total number of “man of the match” award won
Rating	Average rating based on Whoscored algorithm
tackels_pg	Number of tackles per game
interception_pg	Number of interceptions per game
fouls_pg	Number of fouls per game
Offsides	Number of offsides per game
Clear	Number of clearances per game
drb_lost_pg	Number of dribbles lost per game
Blocks	Number of shots blocked per game
OwnG	Total number of own goals
keyP_pg	Number of key passes per game
drb_won_pg	Number of dribbles won per game
fouled_pg	Number of times fouled per game
Disp	Number of times dispossessed per game
UnsTch	Number of unsuccessful touches per game
avgP_pg	Average number of passes per game
crosses_pg	Number of crosses per game
longB_pg	Number of long balls per game

ThrB_pg	Number of through balls per game
totaltacklesx90	Number of tackles won per 90 mins
DribbledPastx90	Number of dribbles past per 90 mins
TotalAttemptedTacklesx90	Number of attempted tackles per 90 mins
totalinterceptionx90	Number of interceptions per 90 mins
fouledx90	Number of times fouled per 90 mins
foulsx90	Number of fouls committed per 90 mins
yellowx90	Number of yellow cards per 90 mins
redx90	Number of red cards per 90 mins
CaughtOffside	Number of offsides per 90 mins
totalclearancesx90	Number of clearances per 90 mins
ShotsBlockedx90	Number of shots blocked per 90 mins
CrossesBlockedx90	Number of crosses blocked per 90 mins
PassesBlocked	Number of passes per 90 mins
shotssaved_Total	Number of shots saved per 90 mins (GK)
shotssaved_SixYardBox	Shots saved from six yard per 90 mins (GK)
shotssaved_PenaltyArea	Shots saved from penalty areaper 90 mins (GK)
shotssaved_OutOfBox	Shots saved from out of box per 90 mins (GK)
Totalshotsx90	Number of shots per 90 mins
OutOfBoxx90	Number of shots from out of box per 90 mins
SixYardBoxx90	Number of shots from six yard per 90 mins
PenaltyAreax90	Number of shots from penalty area per 90 mins
totalgoalsx90	Number of goals per 90 mins
unsuccessfuldribblesx90	Number of unsuccessful dribbles per 90 mins
successfuldribblesx90	Number of successful dribbles per 90 mins
total_dribblesx90	Number of total attempted dribbles per 90 mins
UnsuccessfulTouchesx90	Number of unsuccessful touches per 90 mins
Dispossessedx90	Number of dispossessed times per 90 mins
totalaerialx90	Number of aerial tackles per 90 mins
wonaerialx90	Number of aerial tackles won per 90 mins
lostaerialx90	Number of aerial tackles lost per 90 mins
totalpassesx90	Number of passes per 90 mins
AccLBx90	Number of successful long passes per 90 mins
InAccLBx90	Number of unsuccessful long passes per 90 mins
AccSPx90	Number of successful short passes per 90 mins
InAccSPx90	Number of unsuccessful short passes per 90 mins
totalkey_passesx90	Number of key passes per 90 mins
longkey_passesx90	Number of long key passes per 90 mins
shortkey_passesx90	Number of short key per 90 mins
Crossassistsx90	Number of assists by cross per 90 mins

Cornerassistsx90	Number of assists by corner per 90 mins
Throughballassistsx90	Number of assists by through balls per 90 mins
Freekickassistsx90	Number of assists by free kick per 90 mins
Otherassistsx90	Number of assists by other ways per 90 mins
totalassistsx90	Number of assists per 90 mins
#	Shirt number of player
Date_of_Birth___Age_	Date of birth and age
Nat_	Nationality
Current_club	Current club
Foot	Preferred foot
in_team_length	Time spent in the current club (years)
Contract_length	Time to the end of the contract (years)
Market_value_17	Market value in June 2017 (Transfermarkt)
position	Position on the field (Transfermarkt)
role	Role on the field
height	Height
market_value_18	Market value in June 2018 (Transfermarkt)
transfer_value	Transfer value if transferred in summer 2018

Appendix 2

```
/*import excel*/
%macro import_playerstatistics(sheet_name);
proc                    import                    datafile                    =
"C:\Users\Marco\Desktop\Marco\Thesis\Data\Whoscored_17_18_final.xlsx"
replace dbms=xlsx out= Thesis.Seriea_&sheet_name;
sheet = "&sheet_name";
run;
%mend;

%import_playerstatistics(Summary);
%import_playerstatistics(Defensive);
%import_playerstatistics(Offensive);
%import_playerstatistics(Passing);
%import_playerstatistics(Tackles);
%import_playerstatistics(Interception);
%import_playerstatistics(Fouls);
%import_playerstatistics(Cards);
%import_playerstatistics(Offside);
%import_playerstatistics(Clearences);
%import_playerstatistics(Blocks);
%import_playerstatistics(Saves);
%import_playerstatistics(Shots);
%import_playerstatistics(Goals);
%import_playerstatistics(Dribbles);
%import_playerstatistics(Possession_losses);
%import_playerstatistics(Aerial);
%import_playerstatistics(Passes);
%import_playerstatistics(Key_Passes);
%import_playerstatistics(Assists);


%macro import_stat_transfermarkt(sheet_name);
proc                    import                    datafile                    =
"C:\Users\Marco\Desktop\Marco\Thesis\Data\Transfermarkt\Serie_A.xlsx"
      /*importo file trsfmkt*/
replace dbms=xlsx out= Thesis.Trsfmkt_&sheet_name;
sheet = "&sheet_name";
```

```
run;
%mend;


%import_stat_transfermarkt(TOTAL);

%macro xset(category, old_name, new_name);

data Thesis.Seriea_&category;
set Thesis.Seriea_&category;
rename &old_name = &new_name;
label &old_name = &new_name;
run;
%mend;

%xset (defensive , tackles, tackels_pg);
%xset (defensive , inter, interception_pg);
%xset (defensive , drb, drb_lost_pg);
%xset (defensive , fouls, fouls_pg);
%xset (offensive , keyP, keyP_pg);
%xset (offensive , drb, drb_won_pg);
%xset (offensive , fouled, fouled_pg);
%xset (offensive , off, offsides_pg);
%xset (passing , keyP, keyP_pg);
%xset (passing , avgP, avgP_pg);
%xset (passing , crosses, crosses_pg);
%xset (passing , longB, longB_pg);
%xset (passing , ThrB, ThrB_pg);


%macro xsetcategory90(category, old_name,new_name);

data Thesis.Seriea_&category;
set Thesis.Seriea_&category;
rename &old_name = &old_name&new_name;
label &old_name = &old_name&new_name;
run;
%mend;

%xsetcategory90 (tackles , totaltackles, x90);
%xsetcategory90 (tackles , DribbledPast, x90);
```

```
%xsetcategory90 (tackles , TotalAttemptedTackles, x90);
%xsetcategory90 (interception , total, interception);
%xsetcategory90 (interception , totalinterception, x90);
%xsetcategory90 (fouls , fouled, x90);
%xsetcategory90 (fouls , fouls, x90);
%xsetcategory90 (cards , yellow, x90);
%xsetcategory90 (cards , red, x90);
%xsetcategory90 (offisde , CaughtOffside, x90);
%xsetcategory90 (clearences , total, clearances);
%xsetcategory90 (clearences , totalclearances, x90);
%xsetcategory90 (blocks , ShotsBlocked, x90);
%xsetcategory90 (blocks , CrossesBlocked, x90);
%xsetcategory90 (saves , total, saves);
%xsetcategory90 (saves , totalsaves, x90);
%xsetcategory90 (saves , SixYardBox, x90);
%xsetcategory90 (saves , PenaltyArea, x90);
%xsetcategory90 (saves , OutOfBox, x90);
%xsetcategory90 (shots , Total, shots);
%xsetcategory90 (shots , Totalshots, x90);
%xsetcategory90 (shots , OutOfBox, x90);
%xsetcategory90 (shots , SixYardBox, x90);
%xsetcategory90 (shots , PenaltyArea, x90);
%xsetcategory90 (goals , total, goals);
%xsetcategory90 (goals , totalgoals, x90);
%xsetcategory90 (goals , SixYardBox, x90);
%xsetcategory90 (goals , PenaltyArea, x90);
%xsetcategory90 (goals , OutOfBox, x90);
%xsetcategory90 (dribbles , unsuccessful, dribbles);
%xsetcategory90 (dribbles , unsuccessfuldribbles, x90);
%xsetcategory90 (dribbles , successful, dribbles);
%xsetcategory90 (dribbles , successfuldribbles, x90);
%xsetcategory90 (dribbles , total_dribbles, x90);
%xsetcategory90 (possession_losses , UnsuccessfulTouches, x90);
%xsetcategory90 (possession_losses , Dispossessed, x90);
%xsetcategory90 (aerial , total, aerial);
%xsetcategory90 (aerial , won, aerial);
%xsetcategory90 (aerial , lost, aerial);
%xsetcategory90 (aerial , totalaerial, x90);
%xsetcategory90 (aerial , wonaerial, x90);
%xsetcategory90 (aerial , lostaerial, x90);
%xsetcategory90 (passes , total, passes);
```

```
%xsetcategory90 (passes , totalpasses, x90);
%xsetcategory90 (passes , AccLB, x90);
%xsetcategory90 (passes , InAccLB, x90);
%xsetcategory90 (passes , AccSP, x90);
%xsetcategory90 (passes , InAccSP, x90);
%xsetcategory90 (key_passes , total, key_passes);
%xsetcategory90 (key_passes , totalkey_passes, x90);
%xsetcategory90 (key_passes , long, key_passes);
%xsetcategory90 (key_passes , longkey_passes, x90);
%xsetcategory90 (key_passes , short, key_passes);
%xsetcategory90 (key_passes , shortkey_passes, x90);
%xsetcategory90 (assists , total, assists);
%xsetcategory90 (assists , totalassists, x90);
%xsetcategory90 (assists , Cross, assists);
%xsetcategory90 (assists , Crossassists, x90);
%xsetcategory90 (assists , Corner, assists);
%xsetcategory90 (assists , Cornerassists, x90);
%xsetcategory90 (assists , Throughball, assists);
%xsetcategory90 (assists , Throughballassists, x90);
%xsetcategory90 (assists , Freekick, assists);
%xsetcategory90 (assists , Freekickassists, x90);
%xsetcategory90 (assists , Throwin, assists);
%xsetcategory90 (assists , Throwinassists, x90);
%xsetcategory90 (assists , Other, assists);
%xsetcategory90 (assists , Otherassists, x90);


data Thesis.Seriea_summary;
set Thesis.Seriea_summary;
if Player = "Sergej Milinkovic0Savic" then Player = "Sergej Milinkovic-
Savic";
run;

data Thesis.Seriea_defensive;
set Thesis.Seriea_defensive;
if Player = "Sergej Milinkovic0Savic" then Player = "Sergej Milinkovic-
Savic";
run;

data Thesis.Seriea_offensive;
set Thesis.Seriea_offensive;
```

```
if Player = "Sergej Milinkovic0Savic" then Player = "Sergej Milinkovic-
Savic";
run;

data Thesis.Seriea_passing;
set Thesis.Seriea_passing;
if Player = "Sergej Milinkovic0Savic" then Player = "Sergej Milinkovic-
Savic";
run;

proc sql;
create table Thesis.WSC_final as
select A.*, B.*,C.*,D.*,E.*, F.*,G.*,H.*,I.*, J.*,K.*,
M.*,N.*, O.*,P.*,Q.*,R.*, S.*,T.*
from

Thesis.Seriea_summary as A, Thesis.Seriea_defensive as B,
Thesis.Seriea_offensive as C, Thesis.Seriea_passing as D,
Thesis.Seriea_tackles as E, Thesis.Seriea_interception as F,
Thesis.Seriea_fouls as G, Thesis.Seriea_cards as H,
Thesis.Seriea_offside as I, Thesis.Seriea_clearences as J,
Thesis.Seriea_blocks as K,
Thesis.Seriea_shots as M, Thesis.Seriea_goals as N, Thesis.Seriea_dribbles
as O, Thesis.Seriea_possession_losses as P,
Thesis.Seriea_aerial as Q, Thesis.Seriea_passes as R,
Thesis.Seriea_key_passes as S, Thesis.Seriea_assists as T

where

A.Player = B.Player and B.Player = C.Player and C.Player = D.Player and
D.Player = E.Player and E.Player = F.Player and F.Player = G.Player and
G.Player = H.Player and H.Player = I.Player and I.Player = J.Player and
J.Player = K.Player and K.Player = M.Player and
M.Player = N.Player and N.Player = O.Player and O.Player = P.Player and
P.Player = Q.Player and Q.Player = R.Player and R.Player = S.Player and
S.Player = T.Player;

quit;

run;
```

```
data Thesis.Wsc_final;
set Thesis.Wsc_final;
if Player = "Alessio gnoli" then Player = "Alessio Romagnoli";
if Player = "Filippo gna" then Player = "Filippo Romagna";
if Player = "M'Baye Niang" then Player = "MBaye Niang";
drop B;
drop Throwinassistsx90;
run;



data Thesis.Trsfmkt_total;
set Thesis.Trsfmkt_total;
if Player = "Papu Gómez" then Player = "Alejandro Gómez";
if Player = "Álex Berenguer" then Player = "Alex Berenguer";
if Player = "Ali Adnan" then Player = "Ali Adnan Kadhim";
if Player = "Édgar Barreto" then Player = "Edgar Barreto";
if Player = "Gian Marco Ferrari" then Player = "Gianmarco Ferrari";
if Player = "Pepe Reina" then Player = "Jos□é Reina";
if Player = "Konstantinos Manolas" then Player = "Kostas Manolas";
if Player = "Davide Faraoni" then Player = "Marco Faraoni";
if Player = "Maxi López" then Player = "Maximiliano López";
if Player = "Nicolás Burdisso" then Player = "Nicolas Burdisso";
if Player = "Nicolas N'Koulou" then Player = "Nicolas Nkoulou";
if Player = "M'Baye Niang" then Player = "MBaye Niang";
drop Signed_from;
run;

%macro deleteplayer(player, team);

data Thesis.trsfmkt_total deleted;
set Thesis.trsfmkt_total;

if Player = "&player" and current_club = "&team" then output deleted;
else
output Thesis.trsfmkt_total;

run;
%mend;
```

```
%deleteplayer (Boukary Dram□E atalanta);            /*cancello i giocatori
doppi*/
%deleteplayer (Bruno Petkovic, verona);
%deleteplayer (Christian Puggioni, sampdoria);
%deleteplayer (Cristiano Lombardi, lazio);
%deleteplayer (Cyril Théréau, udinese);
%deleteplayer (Daniel Bessa, genoa);
%deleteplayer (Daniel Pavlovic, sampdoria);
%deleteplayer (Diego Falcinelli ,fiorentina );
%deleteplayer (Duván Zapata, napoli );
%deleteplayer (Emanuele Giaccherini, napoli);
%deleteplayer (Federico Bonazzoli , sampdoria);
%deleteplayer (Federico Ricci, genoa );
%deleteplayer (Federico Ricci, sassuolo );
%deleteplayer (Filip Djuricic, sampdoria);
%deleteplayer (Giovanni Simeone, genoa);
%deleteplayer (Ivan strinic, napoli );
%deleteplayer (jasmin kurtic, atalanta );
%deleteplayer (khouma babacar, fiorentina);
%deleteplayer (leonardo pavoletti, napoli);
%deleteplayer (luca antei, sassuolo);
%deleteplayer (luca rizzo, atalanta);
%deleteplayer (luca rossettini, torino);
%deleteplayer (Marco Borriello, cagliari);
%deleteplayer (Marco Capuano,cagliari );
%deleteplayer (Martú‹ Cáceres , lazio);
%deleteplayer (Maxi López, torino);
%deleteplayer (Niccol□EZanellato , milan );
%deleteplayer (pietro iemmello, sassuolo);
%deleteplayer (Rafael ,napoli );
%deleteplayer (Ryder Matos, udinese);
%deleteplayer (Vittorio Parigini, torino);
%deleteplayer (MBaye Niang, milan);
%deleteplayer ( Cristian Ansaldi, inter);
%deleteplayer ( Cyril Théréau, udinese);
%deleteplayer ( Diego Falcinelli, fiorentina);
%deleteplayer ( Giovanni Simeone, genoa);
%deleteplayer ( Leonardo Pavoletti,napoli );
%deleteplayer ( Luca Rossettini, torino);
%deleteplayer ( Maximiliano López, torino);
%deleteplayer ( Moise Kean, juventus);
```

```
%deleteplayer ( Nikola Kalinic, fiorentina);
%deleteplayer ( Patrik Schick, sampdoria);


proc sql;
create table Thesis.Complete as

select  distinct A.*,B.*

from

Thesis.Wsc_final as A, Thesis.Trsfmkt_total as B

where

A.Player = B.Player;

quit;

run;


data Thesis.Complete elimina;
set Thesis.Complete;
drop R;
if market_value_18 = 0 then
output elimina;
else output Thesis.complete;
run;

%macro changet(oldvar);
data thesis.complete;
set thesis.complete;
   new = input(&oldvar,comma10.);
   drop &oldvar;
   rename new=&oldvar;
run;
%mend;

%changet(height);
%changet(market_value_18);
```

```
%changet(transfer_value);



/*remove variables*/
data Thesis.Final;
set Thesis.Complete;
drop A totalgoalsx90    assists     Yel   Red   SpG   Aerialswon  Offsides
MotM tackels_pg interception_pg     Fouls_pg    clear drb_lost_pg
      totalattemptedtacklesx90      blocks      keyP_pg     drb_won_pg
      total_dribblesx90 fouled_pg   offsides_pg disp  unsuccessfultouchesx90
      Unstch      avgP_pg     crosses_pg  longB_pg    PenaltyAreax90
      SixYardBoxx90     OutOfBoxx90 totalaerialx90    AccSPx90
      totalassistsx90     _     Market_value_17    Prestito_riscattato_nel_18
Prestito_riscattato_nel_19  apps  Date_of_Birth__Age_  Position1  Position2
Position3  ThrB_pg  Crossassistsx90  Cornerassistsx90  Throughballassistsx90
Freekickassistsx90 Otherassistsx90;
run;

/*create role tables*/


proc sql;
create table Thesis.forward as

select A.*

from

Thesis.Final as A

where A.role = "Forward";


quit;

proc sql;
create table Thesis.midfielder as

select A.*
```

```
from

Thesis.Final as A

where A.role = "Midfielder";


quit;

proc sql;
create table Thesis.defender as

select A.*

from

Thesis.Final as A

where A.role = "Defender";


quit;

proc sql;
create table Thesis.goalkeeper as

select A.*

from

Thesis.Final as A

where A.role = "Goalkeeper";


quit;


/*remove variables from each role table*/
data Thesis.Goalkeeper;
set Thesis.Goalkeeper;
```

```
drop Goals redx90 CaughtOffside ShotsBlockedx90 CrossesBlockedx90
Totalshotsx90 unsuccessfuldribblesx90 Dispossessedx90 lostaerialx90
shortkey_passesx90;
run;

/*correlation matrix*/

ods pdf file
="C:\Users\Marco\Desktop\Marco\Thesis\PDF\corr_goalkeeper2.pdf";
ods graphics /height=1200px width=1200px imagemap tipmax=4000;

%prepCorrData(
  in = Thesis.goalkeeper(drop= Player),
  out=goalkeeper_r);
proc sgrender data=goalkeeper_r template=corrHeatmap;
   dynamic _title="Corr matrix for goalkeeper";
run;
ods pdf close;
ods graphics off;


data Thesis.Defender;
set Thesis.Defender;
drop Shotssaved_Total Shotssaved_SixYardBox Shotssaved_PenaltyArea
Shotssaved_OutOfBox;
run;

ods pdf file ="C:\Users\Marco\Desktop\Marco\Thesis\PDF\corr_defender2.pdf";
ods graphics /height=1200px width=1200px imagemap tipmax=4000;

%prepCorrData(
  in = Thesis.defender(drop= Player),
  out=defender_r);
proc sgrender data=defender_r template=corrHeatmap;
   dynamic _title="Corr matrix for defender";
run;
ods pdf close;
ods graphics off;
```

```
data Thesis.Midfielder;
set Thesis.Midfielder;
drop    Shotssaved_Total    Shotssaved_SixYardBox    Shotssaved_PenaltyArea
Shotssaved_OutOfBox;
run;

ods                                pdf                                file
="C:\Users\Marco\Desktop\Marco\Thesis\PDF\corr_midfielder2.pdf";
ods graphics /height=1200px width=1200px imagemap tipmax=4000;

%prepCorrData(
  in = Thesis.Midfielder(drop= Player),
  out=Midfielder_r);
proc sgrender data=Midfielder_r template=corrHeatmap;
   dynamic _title="Corr matrix for Midfielder";
run;
ods pdf close;
ods graphics off;



data Thesis.Forward;
set Thesis.Forward;
drop    Shotssaved_Total    Shotssaved_SixYardBox    Shotssaved_PenaltyArea
Shotssaved_OutOfBox redx90;
run;

ods pdf file ="C:\Users\Marco\Desktop\Marco\Thesis\PDF\corr_forward2.pdf";
ods graphics /height=1200px width=1200px imagemap tipmax=4000;

%prepCorrData(
  in = Thesis.Forward(drop= Player),
  out=Forward_r);
proc sgrender data=Forward_r template=corrHeatmap;
   dynamic _title="Corr matrix for Forward";
run;
ods pdf close;
ods graphics off;
```

```
/* data vs market value*/


ods pdf file ="C:\Users\Marco\Desktop\Marco\Thesis\PDF\scatterdataMV.pdf";

%macro scatterplotmv(xvar);


ods listing style= statistical sge = on;
ods graphics on;
proc sgplot data=Thesis.complete;
   scatter x=&xvar y=market_value_18;
   run;

ods graphics off;
ods listing sge = off;

%mend;

%scatterplotmv(Age);
%scatterplotmv(Mins);
%scatterplotmv(Goals);
%scatterplotmv(Assists);
%scatterplotmv(Yel);
%scatterplotmv(Red);
%scatterplotmv(SpG);
%scatterplotmv(PS_);
%scatterplotmv(AerialsWon);
%scatterplotmv(MotM);
%scatterplotmv(Rating);
%scatterplotmv(tackels_pg);
%scatterplotmv(interception_pg);
%scatterplotmv(fouls_pg);
%scatterplotmv(Offsides);
%scatterplotmv(Clear);
%scatterplotmv(drb_lost_pg);
%scatterplotmv(Blocks);
%scatterplotmv(OwnG);
%scatterplotmv(keyP_pg);
%scatterplotmv(drb_won_pg);
%scatterplotmv(fouled_pg);
```

```
%scatterplotmv(offsides_pg);
%scatterplotmv(Disp);
%scatterplotmv(UnsTch);
%scatterplotmv(avgP_pg);
%scatterplotmv(crosses_pg);
%scatterplotmv(longB_pg);
%scatterplotmv(ThrB_pg);
%scatterplotmv(totaltacklesx90);
%scatterplotmv(DribbledPastx90);
%scatterplotmv(TotalAttemptedTacklesx90);
%scatterplotmv(totalinterceptionx90);
%scatterplotmv(fouledx90);
%scatterplotmv(foulsx90);
%scatterplotmv(yellowx90);
%scatterplotmv(redx90);
%scatterplotmv(CaughtOffside);
%scatterplotmv(totalclearancesx90);
%scatterplotmv(ShotsBlockedx90);
%scatterplotmv(CrossesBlockedx90);
%scatterplotmv(PassesBlocked);
%scatterplotmv(shotssaved_Total);
%scatterplotmv(shotssaved_SixYardBox);
%scatterplotmv(shotssaved_PenaltyArea);
%scatterplotmv(shotssaved_OutOfBox);
%scatterplotmv(Totalshotsx90);
%scatterplotmv(OutOfBoxx90);
%scatterplotmv(SixYardBoxx90);
%scatterplotmv(PenaltyAreax90);
%scatterplotmv(totalgoalsx90);
%scatterplotmv(unsuccessfuldribblesx90);
%scatterplotmv(successfuldribblesx90);
%scatterplotmv(total_dribblesx90);
%scatterplotmv(UnsuccessfulTouchesx90);
%scatterplotmv(Dispossessedx90);
%scatterplotmv(totalaerialx90);
%scatterplotmv(wonaerialx90);
%scatterplotmv(lostaerialx90);
%scatterplotmv(totalpassesx90);
%scatterplotmv(AccLBx90);
%scatterplotmv(InAccLBx90);
%scatterplotmv(AccSPx90);
```

```
%scatterplotmv(InAccSPx90);
%scatterplotmv(totalkey_passesx90);
%scatterplotmv(longkey_passesx90);
%scatterplotmv(shortkey_passesx90);
%scatterplotmv(Crossassistsx90);
%scatterplotmv(Cornerassistsx90);
%scatterplotmv(Throughballassistsx90);
%scatterplotmv(Freekickassistsx90);
%scatterplotmv(Throwinassistsx90);
%scatterplotmv(Otherassistsx90);
%scatterplotmv(totalassistsx90);
%scatterplotmv(in_team_length);
%scatterplotmv(Contract_length);
%scatterplotmv(height);
%scatterplotmv(transfer_value);


proc sql;
create table forwardgoals as

select A.*

from

Thesis.forward as A

where A.goals > 5;


quit;

proc template;
  define statgraph bubbles;
    begingraph;
      entrytitle 'Radius of Influence';
      entrytitle 'Bubbles Show Distance Covered by Observation';
      layout overlay;
        bubbleplot x=age y=market_value_18
            size=goals / datalabel=player;
      endlayout;
    endgraph;
```

```
  end;

proc sgrender data=forwardgoals template=bubbles;
run;


ods listing style= statistical sge = on;
ods graphics on;
proc bubbleplot data=Thesis.forward;
   scatter x=&xvar y=market_value_18;
   run;

ods graphics off;
ods listing sge = off;


ods pdf close;



ods listing style= statistical sge = on;
ods graphics on;
proc template;
  define statgraph bubbles;
    begingraph;
      entrytitle 'Relationship Between Goals and Market Value';
      entrytitle 'Bubbles Show Key Passes per 90 mins';
      layout overlay;
        bubbleplot x=goals y=market_value_18
            size=totalkey_passesx90/    name="sp"    group    =    position
datalabel=player ;
            discretelegend "sp" / title="Position";
      endlayout;
    endgraph;
  end;

proc sgrender data=forwardgoals template=bubbles;
run;

ods graphics off;
ods listing sge = off;
```

```
ods pdf close;


proc sql;
create table midfieldertmp as

select A.*

from

Thesis.midfielder as A

where A.mins > 2000;


quit;

ods                                    pdf                                    file
="C:\Users\Marco\Desktop\Marco\Thesis\PDF\bubblesmidfielder.pdf";
proc template;
  define statgraph bubbles;
    begingraph;
      entrytitle 'Relationship Between Key Passes and Market Value';
      entrytitle 'Bubbles Show Age of Players';
      layout overlay;
        bubbleplot x=totalkey_passesx90 y=market_value_18
            size= age/ group = position datalabel=player;
      endlayout;
    endgraph;
  end;

proc sgrender data=midfieldertmp template=bubbles;
run;
ods pdf close;



proc template;
  define statgraph bubbles;
```

```
    begingraph;
      entrytitle 'Radius of Influence';
      entrytitle 'Bubbles Show Distance Covered by Observation';
      layout overlay;
        bubbleplot x=shotssaved_Total y=market_value_18
            size= age/ name="sp"  group = contract_length datalabel=player;
            discretelegend "sp" / title="Position";
      endlayout;
    endgraph;
  end;

proc sgrender data=Thesis.goalkeeper template=bubbles;
run;



proc template;
  define statgraph bubbles;
    begingraph;
      entrytitle 'Clearances vs market value';
      entrytitle 'Bubbles Show Age';
      layout overlay;
        bubbleplot x=totalclearancesx90 y=market_value_18
            size= age/  name="sp" group = position datalabel=player;
            discretelegend "sp" / title="Position";
      endlayout;
    endgraph;
  end;

proc sgrender data=Thesis.defender template=bubbles;
run;



proc template;
  define statgraph bubbles;
    begingraph;
      entrytitle 'Rating vs market value';
      entrytitle 'Bubbles Show Age';
      layout overlay;
        bubbleplot x=rating y=market_value_18
```

```
            size= age/ name="sp" group = role;
                  discretelegend "sp" / title="Role";
      endlayout;
    endgraph;
  end;

proc sgrender data=Thesis.final template=bubbles;
run;




proc sql;
create table finaltmp as

select COUNT(A.market_value_18),A.market_value_18

from

Thesis.final as A

group by market_value_18;


quit;



/* final model*/

%macro linearmodel(var);
title "&var";
ods graphics on;
ods pdf file ="C:\Users\Marco\Desktop\Marco\Thesis\PDF\linearmodel&var.pdf";
ods listing style = statistical sge = on;
proc glm data = Thesis.forward;
model market_value_18 = &var;
run;
ods listing sge = off;
ods graphics off;
%mend;
```

```
%linearmodel(goals);
%linearmodel(PS_);
%linearmodel(Mins);
%linearmodel(Rating);
%linearmodel(Totalshotsx90 );
%linearmodel(Successfuldribblesx90);
%linearmodel(totalkey_passesx90);
%linearmodel(Contract_length );
%linearmodel(Foulsx90);
%linearmodel(Lostaerialx90);


proc glm data= Thesis.forward;
ods graphics on;
ods pdf file ="C:\Users\Marco\Desktop\Marco\Thesis\PDF\final_model.pdf";
ods listing style = statistical sge = on;
title = "final_model";
model market_value_18 = goals Totalshotsx90 Successfuldribblesx90
totalkey_passesx90 Contract_length Lostaerialx90;
ods graphics off;
ods listing sge = off;
run;


proc reg data= Thesis.forward;
ods graphics on;
ods pdf file
="C:\Users\Marco\Desktop\Marco\Thesis\PDF\final_model_stepwise.pdf";
ods listing style = statistical sge = on;
model market_value_18 = goals Totalshotsx90 Successfuldribblesx90
totalkey_passesx90 Contract_length Lostaerialx90
/ vif selection= stepwise;
ods graphics off;
ods listing sge = off;
run;
```

-

Appendix 3

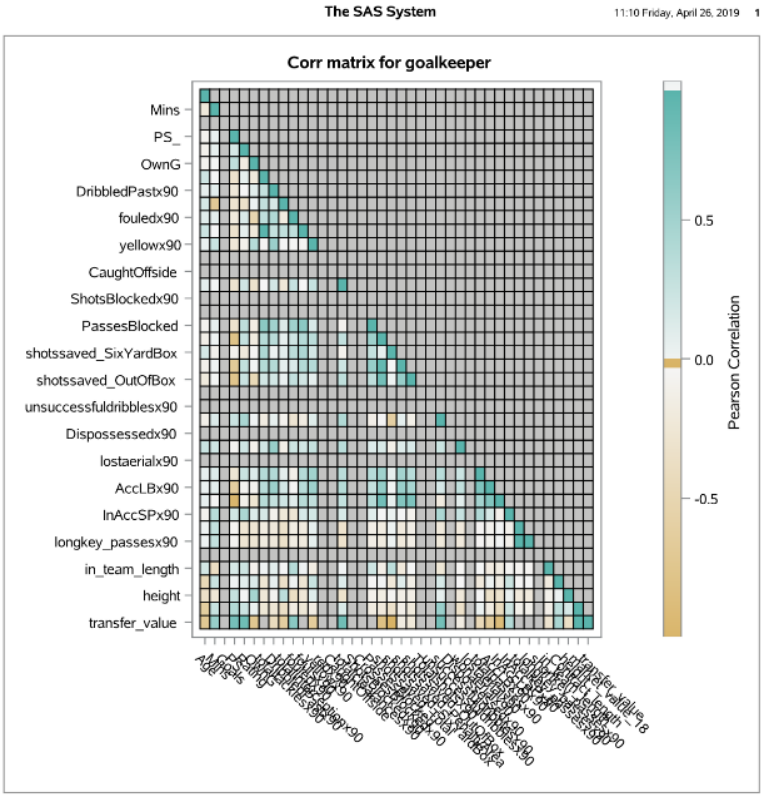

Figure 17 - Correlation matrix for goalkeepers

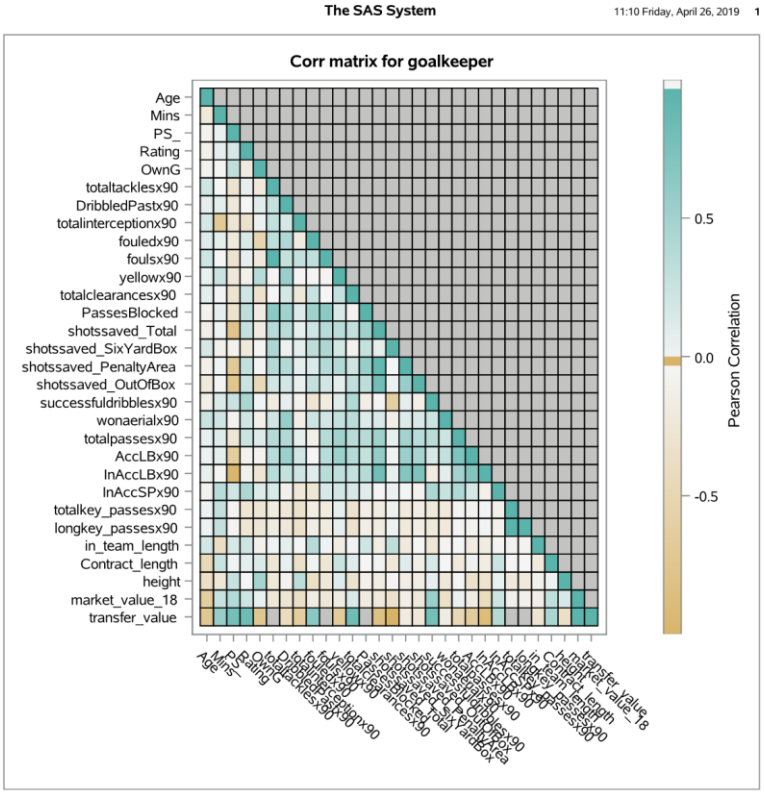

Figure 18 - Final correlation matrix goalkeepers

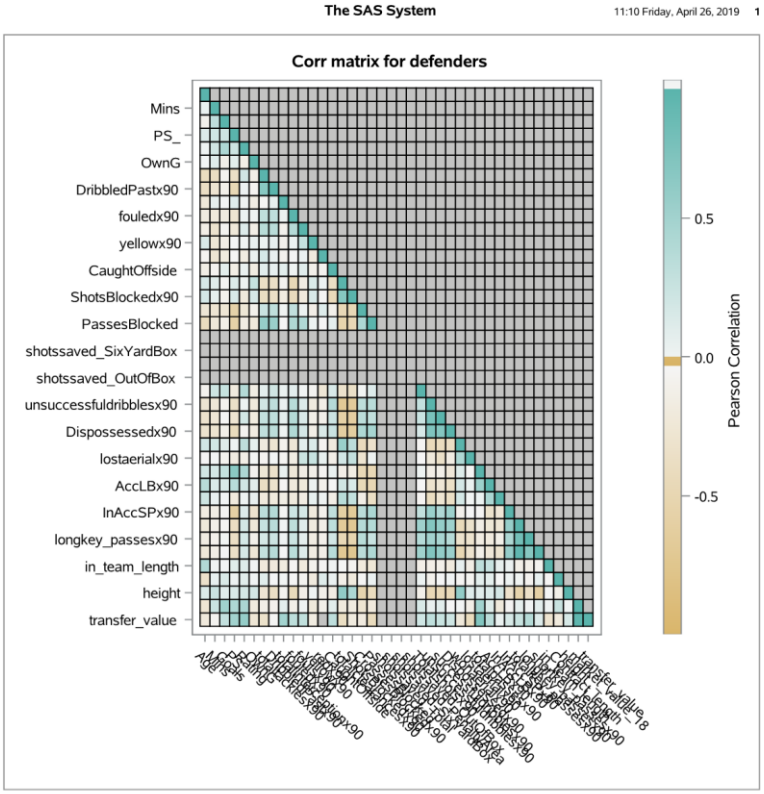

Figure 19 - Correlation matrix for defenders

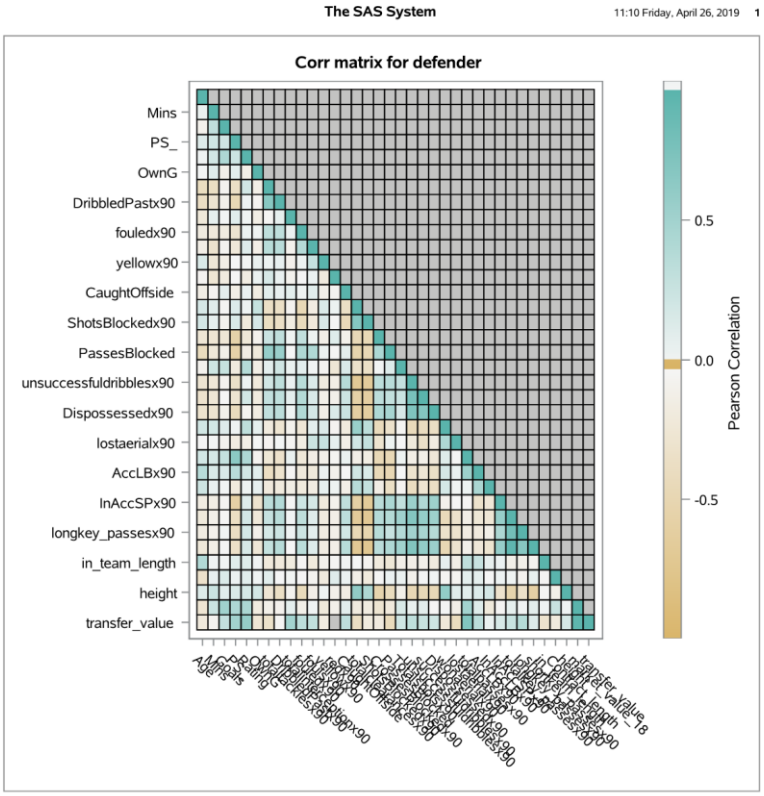

Figure 20 – Final correlation matrix for defenders

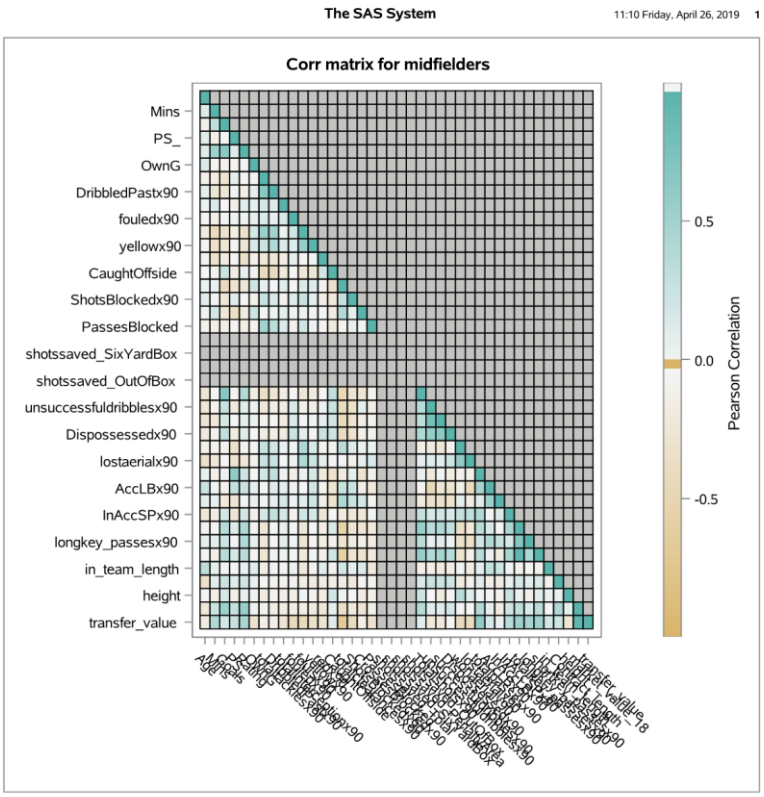

Figure 21 - Correlation matrix for midfielders

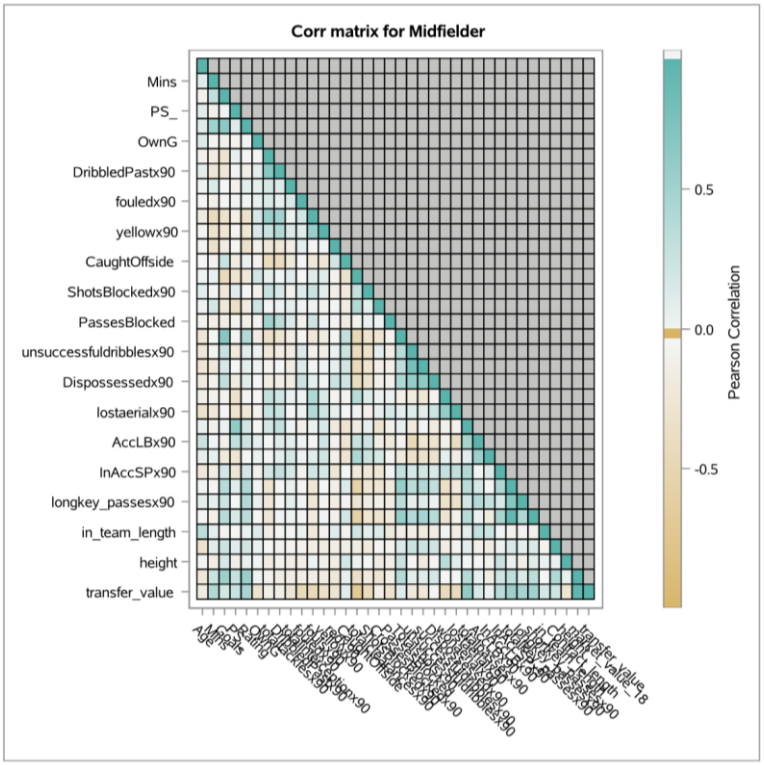

Figure 22 – Final correlation matrix for midfielders