Explanatory Data Analysis, Preparation of shot maps and Predictive model for top leagues in European football

Under guidance of

Dr. Bhalchandra Pujari

Presented By

Jayesh Patil (CMS1917)

Gaurav Tikhe (CMS1923)

Nirmitk Tripathi (CMS1924)

**Acknowledgement**

Firstly, we would like to thank our project guide Dr. Bhalchandra Pujari for his guidance and constant support throughout the duration of the project.

Secondly we would like to thank the organizations Statsbomb and Wyscout for making this data available publicly. Without this data, this project would have been a horrific nightmare and daunting challenge. Also we would like to thank Alin Secareanu for making the events data available on kaggle.

Finally, we would also like to thank our families, for their support, our friends, for their fun, support and discussions regarding the project.

Table of Contents

|  |  |  |
| --- | --- | --- |
| Chapter | Topic | Page number |
| 1 | [Introduction](#Introduction) | 5 |
| 2 | [EDA](#EDA) | 6 |
| **2.1** [EDA on off-the-ball events data](#EDA_off_the_ball) | 6 |
| **2.2** [EDA on on-the-ball events data](#EDA_on_the_ball) | 11 |
| 3 | [Valuing Player Actions (VAEP) & Player ratings](#VAEP) | 14 |
| 4 | [Predictive Model Development](#Pred_model) | 17 |
| **4.1** [Methodology](#Pred_model_methodology) | 17 |
| 5 | [Creation of GUI](#GUI) | 20 |
| 6 | [Results](#Results) | 25 |
| **6.1** [Results for off-the-ball events data](#Results_off_the_ball) | 25 |
| **6.2** [Analysis of on-the-ball EDA](#Results_on_the_ball) | 35 |
| **6.3** [Results for predictive model](#Results_pred_model) | 45 |
| 7 | [References](#References) | 48 |

**Project Summary**

In the era of big data and artificial intelligence, many different domains such as healthcare, finance, banking, retail, travel, social media and many more. Sports analytics is one such branch which has been attracting increasing interest due to availability advanced sensing technologies which make the data available. One such effort is being made in this project, where along with EDA, we are trying to develop a predictive model to predict the outcome of games as well as final points table with the use of expected goals parameter.

EDA is aimed at studying the football dataset, to analyse, extract information from it and make important conclusions based on the data. The main goal is to find the different styles of plays in different teams, finding weaknesses and strengths of the teams and assess the ways of measurement and improvement of the team performance. We got the most effective events and capitalised on their characteristics in order to achieve the set goal. For EDA we are working on Kaggle Football events dataset which contains the off-the -ball type events data (don’t include passing shots and location) of 4 years (season 2012-13 to season 2016-17) for 5 leagues (Laliga, Premier League, Bundesliga, League One, Serie A) as well as match events data made available for public access by Statsbomb and Wyscout.

For predictions we have used multiple output regressor with different regressors such as Random forest regressor (RF), Support Vector Regressor (SVR). In the points table, the main ranks to be predicted are top positions (vary as per league) and bottom 3 positions which is selected as accuracy measure. The prediction of season for 2019-20 season, top prediction accuracy is found to be 67 % whereas 33% for bottom positions.

Along with this, understanding attacking style of play of different teams using the shot maps, and of players using pass maps is also done. Shot maps of all Barcelona matches in the latest 19/20 season were created and analyzed. Similarly, Lionel Messi’s pass maps in the same matches were created and were used to extract inference on his style of play, positional preference and influence on the game. Along with that passing clusters were also created for Barcelona for the same set of matches to deeply understand the style of play.

Also along with this we have created a GUI wherein we display the goal plots in which all the previous actions leading to a goal are displayed and top 10 players, of each league, according to VAEP rankings are also displayed.

**Chapter 1: Introdu****ction**

In the era of big data, the amount of data getting generated every day is pretty huge. It is estimated that by the end of 2020 the digital universe will expand to 44 zettabytes. Thus data analysis helps us examining such large datasets in order to extract meaningful as well as useful conclusions. In recent years the analytics is being used in the field of sports for prediction and draw various insights.

Apart from some scattered attempts, there have been 3 major sources responsible for the availability of sports data in recent times due to availability of advanced sensing technologies providing ways to extract the high-quality data from football matches. The three of the major sources responsible for availability of the data are:

1. Football match logs describing the events in the match with use of a proprietary tagging software
2. Video-tracking data which describes trajectories of the players collected through video recordings
3. GPS data obtained during the training sessions.

Due to use of proprietary tagging software, not all the data obtained is publicly available. Thus for the current project we have used data from various sources. As mentioned above, the project comprises of three parts as:

1. Explanatory Data Analysis (EDA)

2. Preparation of shot maps, goal plots and passing networks, VAEP player rankings creating GUI.

3. Prediction model for predicting outcome of match and points table for the season

These activities are done for different top league in European football. For EDA and preparation of shot maps and passing networks we have used data from Kaggle titled “Football Events” and sites such as Wyscout and Statsbomb. For the development of the predictive model, we have used data for each season from a website “football-data.co.uk”. Further report is divided in 4 sections as EDA, shot maps and passing network preparation, predictive model development and results obtained so far.

**Chapter 2. ED****A**

The importance of analysing of football events has emerged. By analysing all events in game we can understand the strengths and weaknesses of each teams and players and by doing that we can get more insights about the game.

**2.1 EDA on Off-The-Ball even****ts data:**

**2.1.1) Data Selection and Import:**

Generally, EDA is carried out for better understanding of data for prediction models. But in this project we couldn’t find a specific data to work on all three parts of project. So we decided to find most suitable data for each part of the project. For EDA we are using Kaggle football event dataset which consist off-the-ball data. It includes events from more than 7,000 games from the top 5 European Leagues (Premier League, La Liga, Serie A, Bundesliga, and Ligue 1) from 2011 to 2017.In this data set, there are season, league, teams, match results, betting odds and events during the match.

In these events there are all events in the match: red card, attack, corner, foul, offside, player change, penalty etc. From them following events are used for analysis:

* **Examples of events used for EDA:**

1. Yellow and red cards served
2. Fouls against the team
3. Shot place
4. Too high
5. Goals against the team
6. Goals scored
7. Corners
8. Shoots on target
9. Shot-place
10. Location

**2.1.2 Pre-Processing:**

From Kaggle we received 3 file. In there were 2 csv files and 1 was text file.

CSV files: 1) events.csv, 2) ginf.csv

Text files: dictionary.txt

Events file consist the events data of all matches and ginfo (game information) file contain all details about the matches. And the last dictionary file was an information file which had all the full form for the events data. For better understanding in analysis we converted numerical data into categorical data with help of dictionary.

In pre-processing while looking for missing data we found that some of matches didn’t had all of events data. So by analysing data as per seasons we got to know that season (2012,2013,2014) were missing the most of data. So to move ahead we decided to neglect those seasons from EDA and continue on year 2015,2016,2017 data. And to perform EDA we were not going to need the odds data and some other features in data so we dropped them from the EDA.

* + 1. **EDA:**

We performed EDA in 3 steps:

* 1. **Analysis as per Leagues:**

First we find out total number of teams which play in all of leagues for the 3 seasons which we are using.

League One = 25

Premier League = 24

Bundesliga = 21

La Liga = 26

Serie A = 25

So now we can see that Bundesliga is 3-4 teams short with others. which is why it was important to calculate the ratios rather than dealing in absolute values. So now by that we get following data:

League shots and goals:



League fouls and cards served:



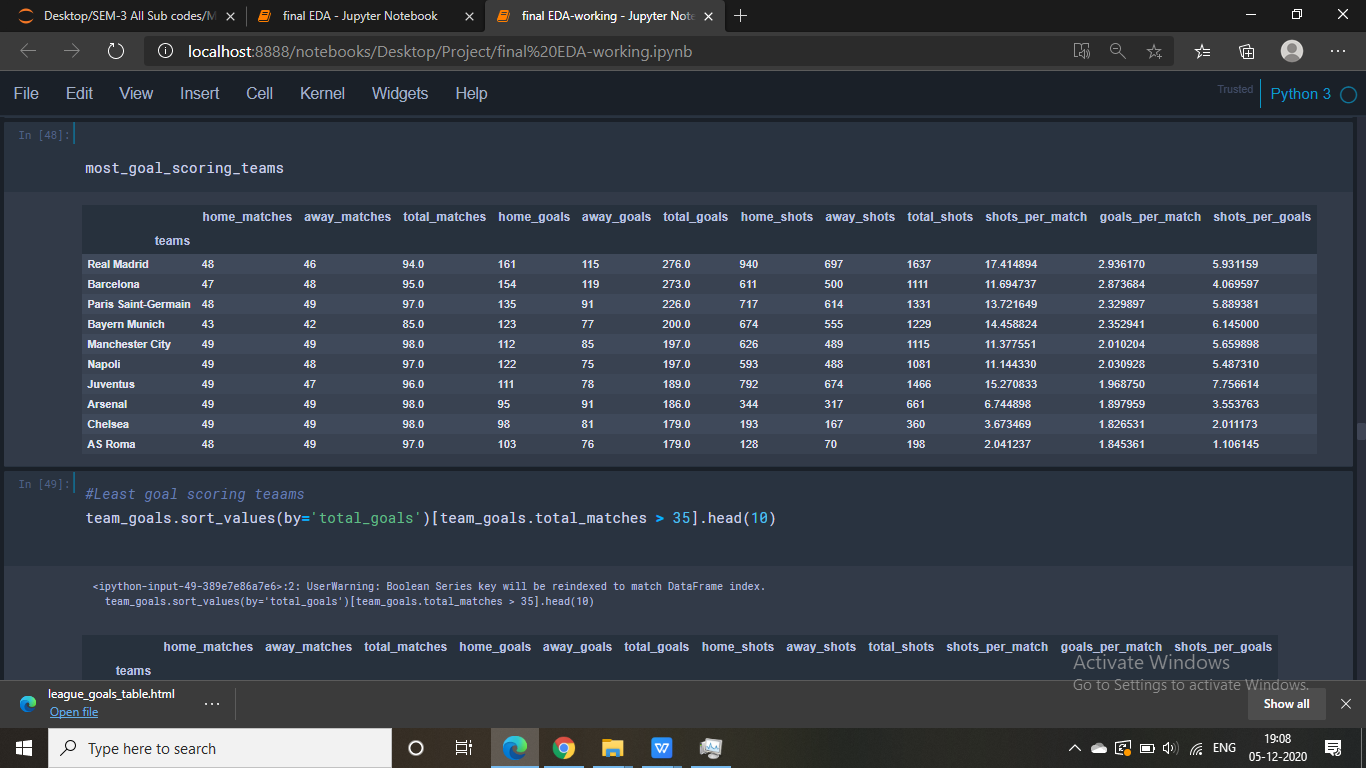
* 1. **Analysis per Teams:**

While doing analysis per team we got to know that some of teams played very few matches and some of teams scored very less goals. So they will act like outlier. Therefore, we decided to define some criteria for doing analysis per teams.

So now team must have played at-least 1 season (more than 35 matches) so they will have considerable amount of data to get accounted.

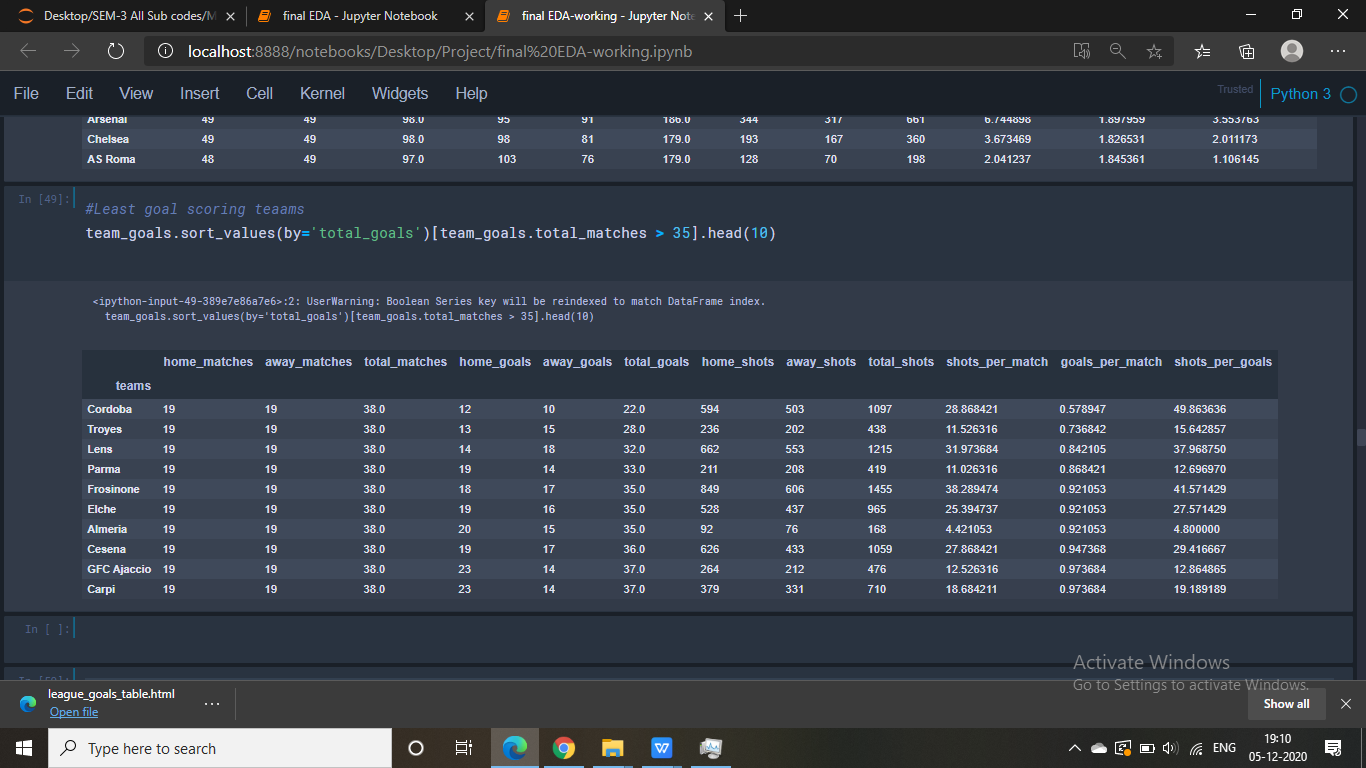
Stats of Top goal scoring teams:

Table 1. Table showing statistics of top goal scoring teams



Stats of Least goal scoring teams:

Table 2. Table showing statistics of least goal scoring teams



* 1. **Analysis as per Players:**

Same as teams here some players have played very less matches and also scored very less goals. So to avoid misunderstanding or outliers we decided to define some criteria.

So now player to be count he must have played at least 1 season (more than 35 matches) so now they will have considerable amount of data to be taken accounted for analysis.

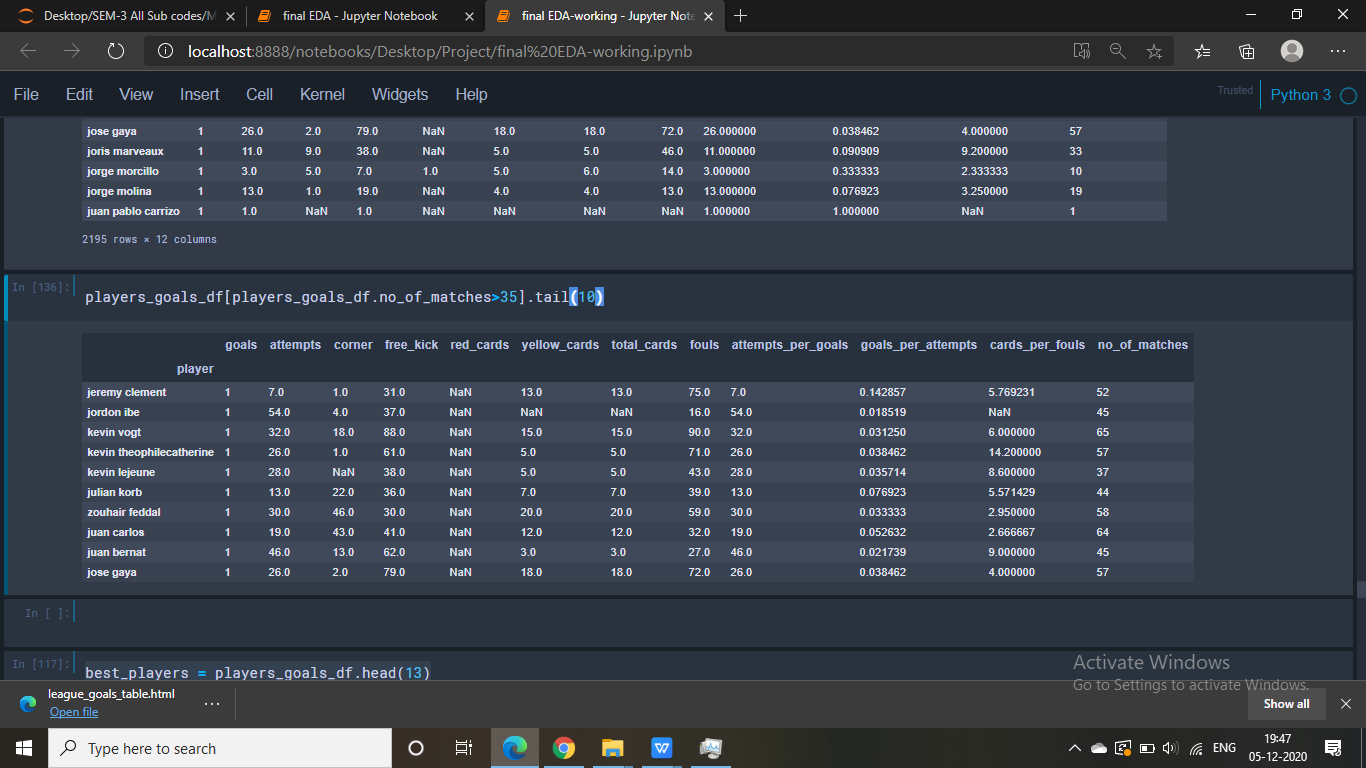
Stats of top goal scoring players:

Table 3. Table showing statistics of top goal scoring players



Stats of Least goal scoring players:

Table 4. Table showing statistics of least goal scoring players



**2.2 EDA on On-The-Ball e****vents data:**

**2.2.1 Data Selection and Import:**

Events data for various matches in various competitions for different seasons was made available by Statsbomb but not for all the matches in the season. For example, events data for only Barcelona’s matches for all the seasons from 2007/08 to present was available for public access by Statsbomb, and therefore, our EDA in La Liga was limited to Barcelona’s EDA in the latest season (2019/20).

Wyscout is another organization that keeps track of events data for every team and every match across various first tier leagues around the world for several seasons. They have made the complete set of events data across the Top 5 European Football Leagues (English Premier League, Spanish La Liga, Italian Serie A, German Bundesliga and French Ligue 1) for all matches for the entire 2017-18 season, freely available for public access to be utilized in football data analysis.   
  
Our EDA using Wyscout Data is on this available data. There is a popular research paper authored by **Pappalardo** et al titled “**A public data set of spatiotemporal match events in soccer competitions**” and our EDA is based on the aforementioned research paper.   
  
The given Wyscout (and Statsbomb) events data set has parameters that can be classified into 2 major categories:

**Spatial dimensions:** These parameters include all the data related to positions and locations of a player and /or the event. In Wyscout dictionary these parameters are placed under the heading **“positions**” which is defined as: “the origin and destination positions associated with the event. Each position is a pair of coordinates (x, y). The x and y coordinates are always in the range [0, 100] and indicate the percentage of the field from the perspective of the attacking team. In particular, the value of the x coordinate indicates the event's nearness (in percentage) to the opponent's goal, while the value of the y coordinates indicates the event's nearness (in percentage) to the right side of the field.”

**Temporal dimensions:** These parameters include all the data related to positions and locations of a player and /or the event. In Wyscout dictionary these parameters are placed under 2 headings – “matchPeriod” which is defined as: “the period of the match. It can be "1H" (first half of the match), "2H" (second half of the match), "E1" (first extra time), "E2" (second extra time) or "P" (penalties time)”; - “event Sec” which is defined as: “the time when the event occurs (in seconds since the beginning of the current half of the match)”. Using the above Statsbomb and Wyscout data, player analysis was done on Lionel Messi using shot maps and pass maps, team analysis on Barcelona was done using shot maps and passing clusters, and match analysis of Napoli vs Juventus match was done using passing networks.

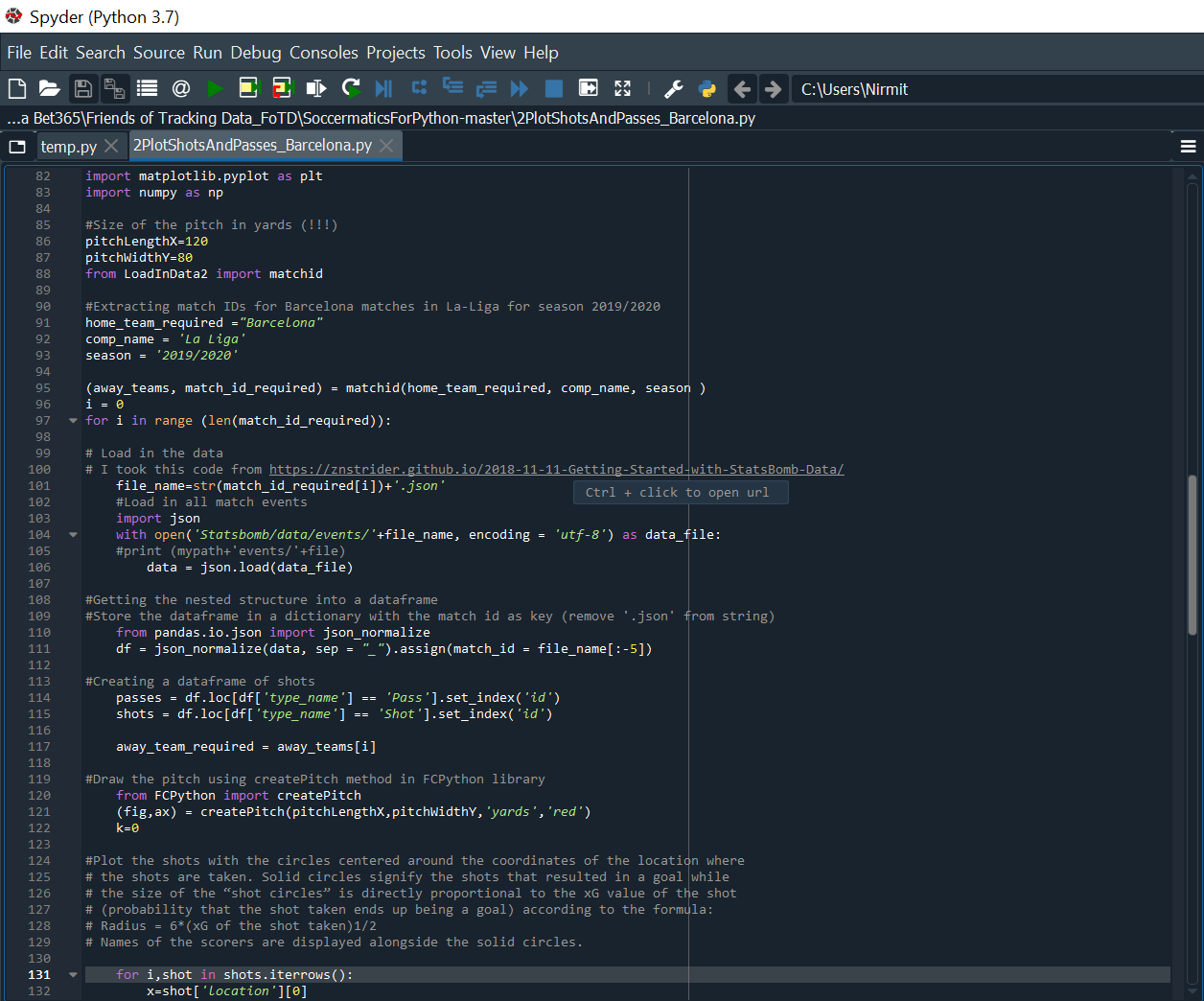
By looking at the position of the field *where* the events occur, we planned to investigate interesting aspects of a soccer match, such as the spatial distribution of players and events. One can expect that we observe differences in the spatial distribution of events when we select the players by their role: while the events of forwards are expected to be observed mainly in the opponent’s half of the field, the events of defenders are expected to be observed mostly in the own half and on the sides of the field. Similarly, there is also an expectation that the spatial distribution of events changes with their type: attacking events (e.g., shots) are mostly observed close to the opponent’s goal, while defensive events (e.g., clearances) are mostly observed close to the team’s own.

**2.2.2 Methodology:**

**Procedure to plot shot maps of Barcelona matches and pass maps and shot maps of Lionel Messi for the entire 2019/2020 La Liga season:**

Language used: Python

1. Extracting match Ids from Statsbomb data for Barcelona matches in La-Liga for season 2019/2020.
2. Load in the data.
3. Convert the data into a dataframe and store the dataframe in a dictionary with the match id as key (remove '.json' from string).
4. Creating a dataframe of shots.
5. Draw the pitch using createPitch method in FCPython library
6. Plot the shots with the circles centered around the coordinates of the location where the shots are taken. Solid circles signify the shots that resulted in a goal while the size of the “shot circles” is directly proportional to the xG value of the shot.
7. Plot pass maps and shot maps for Lionel Messi for all the matches during the season 2019/2020.



Code Snippet: Python code for extracting events data and plotting shots and passes

**Chapter 3. Valuing Player A****ctions (VAEP) and Player Rankings**

**3.1 Methodology:**

Language used: Python and Python Notebook

For this we utilised the available 2017/18 Wyscout data on Premier League (England), La Liga (Spain), Serie A (Italy), Bundesliga (Germany), and Ligue 1 (France) using a python package called socceraction and a research paper titled: **“Actions Speak Louder Than Goals: Valuing Player Actions in Soccer”** authored by **Decroos, Tom and Bransen, Lotte and van Haaren, Jan and Davis.**

**socceraction** is a Python package for objectively quantifying the impact of the individual actions performed by soccer players. It contains three components:

* **SPADL** (Soccer Player Action Description Language): a unified and expressive language for on-the-ball player actions.
* **VAEP** (Valuing Actions by Estimating Probabilities): a framework to value actions on their expected impact on the score line.

**socceraction** uses event stream data to value the individual actions performed by soccer players. Computing these action values requires the three steps described below.

**1. Conversion from event stream format to SPADL:** SPADL is a language for describing player actions, as opposed to the formats by commercial vendors that describe events. The distinction is that actions are a subset of events that require a player to perform the action. For example, a passing event is an action, whereas an event signifying the end of the game is not an action. SPADL was designed to be human-interpretable, simple and complete to accurately define and describe actions on the pitch. This package currently supports converters for Opta, Wyscout, and StatsBomb event stream data.

### 2.  Estimating scoring and conceding probabilities: The intuition is that all good actions should aim to

1. increase the *chance of scoring* a goal in the short-term future and/or,
2. decrease the *chance of conceding* a goal in the short-term future.

Valuing an action for a team then requires assessing the *change* in probability for both scoring and conceding as a result of action  moving the game from state to state. The change in probability for team**x** scoring, where **x** can be either the home team **h** or the visiting team **v**, can be computed as:

Suppose that for each game state **= [,…,],** we have access to the probabilities of scoring and conceding in the near future for the home team **h** and the visiting team **v**. Let  **and**denote the probability of the home team h respectively scoring and conceding in the near future. Similarly, let and  denote the probability of the visiting team **v** respectively scoring and conceding in the near future.

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |

This change will be positive if the action increased the probability that team **x** will score a goal. We call this change  the *offensive value* of an action  for team **x**. Similarly, the change in probability for team **x** conceding can be computed as:

|  |  |  |  |
| --- | --- | --- | --- |
| **(2)** |  |  |  |

This change will be positive if the action increased the probability that team **x** will concede a goal. However, all actions should always aim to decrease the probability of conceding. That is why we call the negation of this change**-** the *defensive value* of an action  for team **x**.

The VAEP framework provides a simple approach to valuing actions that is independent of the representation used to describe the actions. The framework’s strength is that it transforms the subjective task of valuing an action into the objective task of predicting the likelihood of a future event in a natural way. Valuing an action for a team then requires assessing the change in probability for both scoring and conceding as a result of an action. Therefore, **socceraction** converts each game state to a feature-vector format and trains a probabilistic classifier to estimate the probabilities of scoring and conceding in the near future for both teams.

### Compute VAEP values: An action moves the game state from one state to another. Using the probabilities computed in the previous step, we can define the offensive value of an action as the change in scoring probability before and after the action. This change will be positive if the action increased the probability that the team which performed the action will score (e.g., a successful tackle to recover the ball). Similarly, we define the defensive value of an action as the change in conceding probability. This change will be positive if the action increased the probability that the team will concede a goal (e.g., a failed pass). Finally, the total VAEP value of an action is the difference between that action's offensive value and defensive value.

We combine Equations [1](https://www.groundai.com/project/actions-speak-louder-than-goals-valuing-player-actions-in-soccer6805/2#S3.E1) and [2](https://www.groundai.com/project/actions-speak-louder-than-goals-valuing-player-actions-in-soccer6805/2#S3.E2) to derive an action’s total VAEP value.

###### **Definition (VAEP Value): The total VAEP value of an action is the sum of that action’s offensive value and defensive value.**

|  |  |  |  |
| --- | --- | --- | --- |
| (3) |  |  |  |

Given that we are usually interested in the value of an action for the team of the player performing the action, we use **)** to denote **)** where is the team of the player performing action .

**Chapter 4. Predictive mo****del development**

Since football is a team game, and there are 22 players involved on the field. Thus it is difficult to consider each and every individual player and actions he/she is involved in. Hence to make the model simpler, we have considered combined actions of each of the team as features in the model. For example, total shots taken by home or away team are considered instead of shots taken by individual player. This reduces complexity of the model as well computation load on the system. The methodology used for developing a model is as follows:

**4.1 Metho****dology**:

For the predictive model, we have used data from a website “football-data.co.uk”. This website contains data for many past seasons for variety of the football leagues in Europe. We have used expected goal approach to calculate expected goal for both the home and away for each of the match of the season along with some other parameters such as average home goals scored, average away goals scored, also replacement of relegated teams, home team attacking and defensive strength for every match, away team attacking and defensive strength for every match.

For calculation of the expected goals for both home and away team, we have used the data for the current season of premier league for which we have to predict result for each match, and data from past season of premier league and championship. From the last season data, we have taken goals scored by each team playing in current season when they played matches at home and away both in premier league and championship. For example, for prediction for season of 2017-18, we need list of teams played in 2017-18 season thus we use 2017-18 season data. There will be 3 teams promoted from championship of 2016-17 season and 3 teams will be relegated from premier league season of 2016-17. Number of goals scored and conceded while being at home and away by all teams playing 2017-18 season we obtained from this past season.

Expected goals calculations are done as follows:

Expected goals for home team = HTAS \* ATDS \* Avg. goals scored at home

Expected goals for away team = ATAS \* HTDS \* Avg. goals scored away from home

Where,

HTAS = Home team attacking strength

HTDS = Home team defensive strength

ATAS = Away team attacking strength

ATDS = Away team defensive strength

Calculation of terms HTAS, HTDS, ATAS, ATDS is as follows:

To accommodate promoted teams in premier league, the values of HTAS, HTDS, ATAS, ATDS are averaged with these values for relegated teams as the promoted teams replace relegated teams.

As an example, considering prediction for season 2017-18 of premier league following steps were taken:

1. Preprocessing of last season data for 2016-17 using data for 2015-16
2. Preprocessing of current season 2017-18 using data for 2016-17
3. Using Multioutputregressor with Random Forest or SVR as regressors
4. Preparation of final points tables based on the predictions
5. Finding accuracy using correct rank predictions for top 6 in table and bottom 3 in table.

The similar approach is taken for other leagues also. Different machine learning algorithms are tried for the prediction analysis. Further ahead as the performance of some of the algorithms was not so good, we decided to take the deep learning route to check its performance for the prediction purpose. Model used for deep learning is a sequential model with 1 hidden layer. Performances of different machine learning algorithms used is presented in the results chapter.

**Chapter 5. Creation of t****he GUI**

Data analysis is incomplete without some plots or values to show with it. To show what actions leads to goal, what particular pass, dribble, cross or save which are the most valuable actions we decided to create a GUI wherein it can be showed plots where it shows which actions previous to the goal made it possible or what players in a specific league which are involved in such valuable actions for their respective clubs. The GUI is designed for the season of 2017-18. The screenshot of the GUI is as follows:

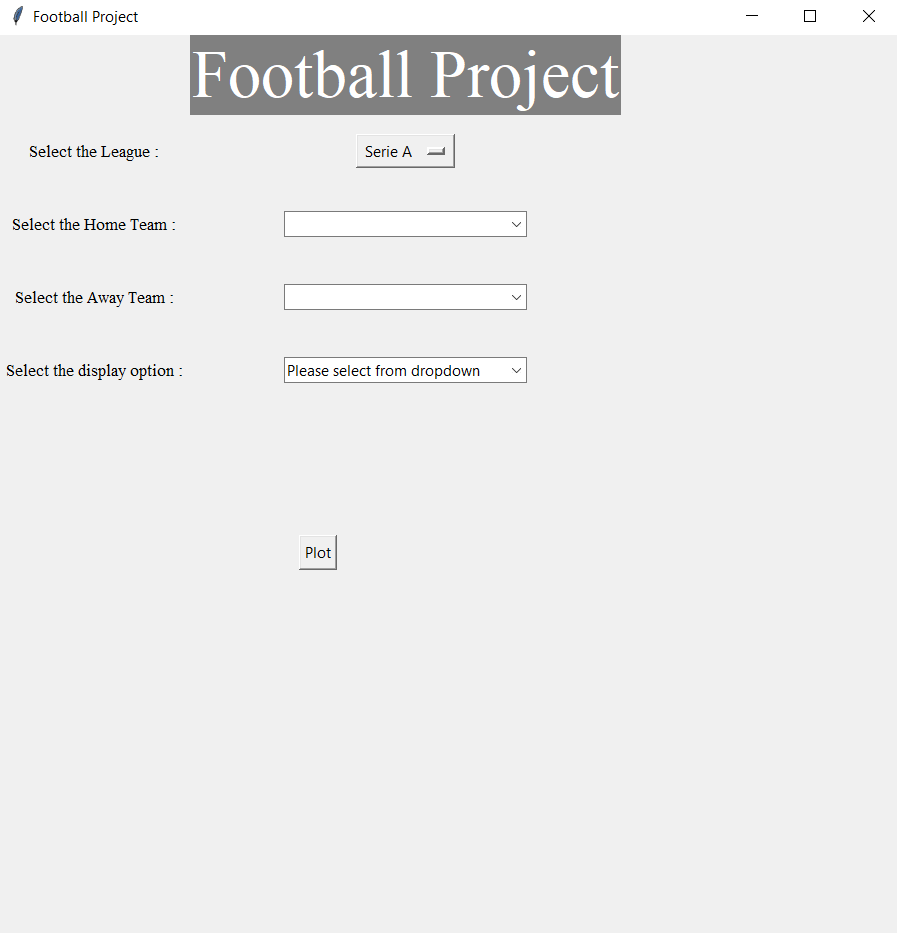


Fig 1. Figure showing GUI created

The structure of the GUI is as follows:

There are 4 inputs required from the user side:

1. Selecting a league from choices Serie A, La Liga, Premier League, Bundesliga and Ligue1.

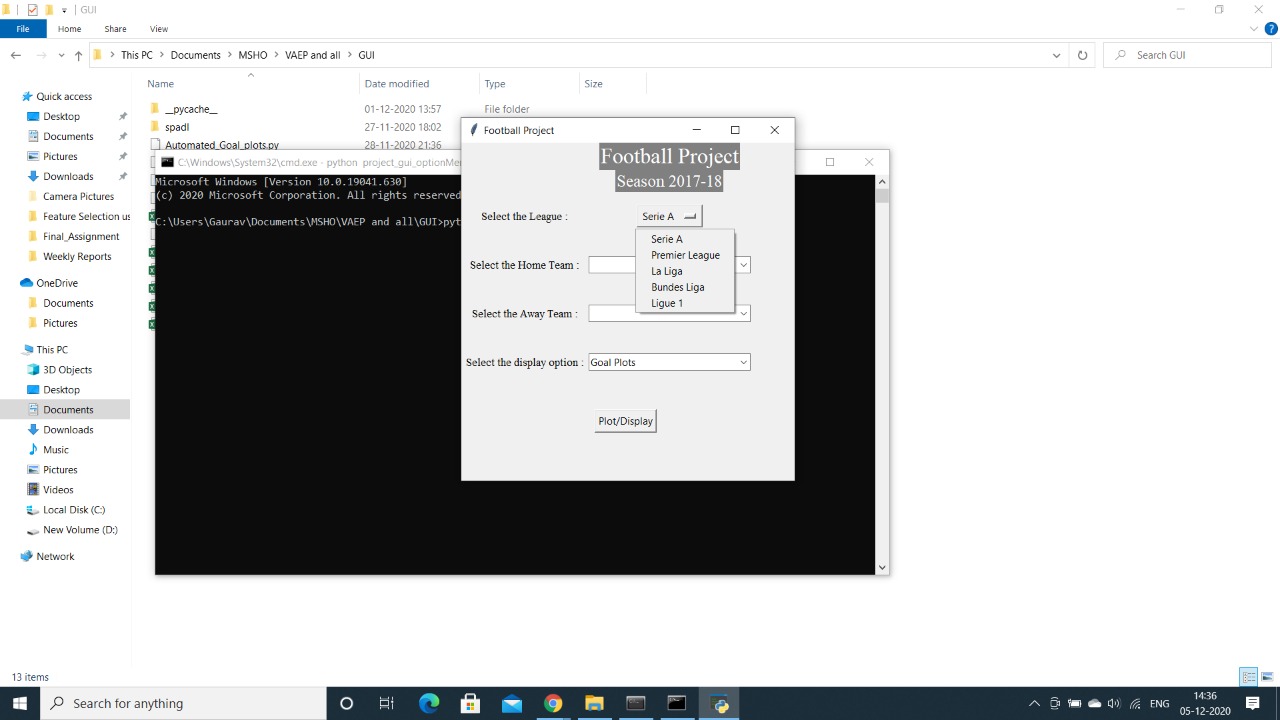


Fig 2. Selection of league from GUI

1. Selecting the home team.

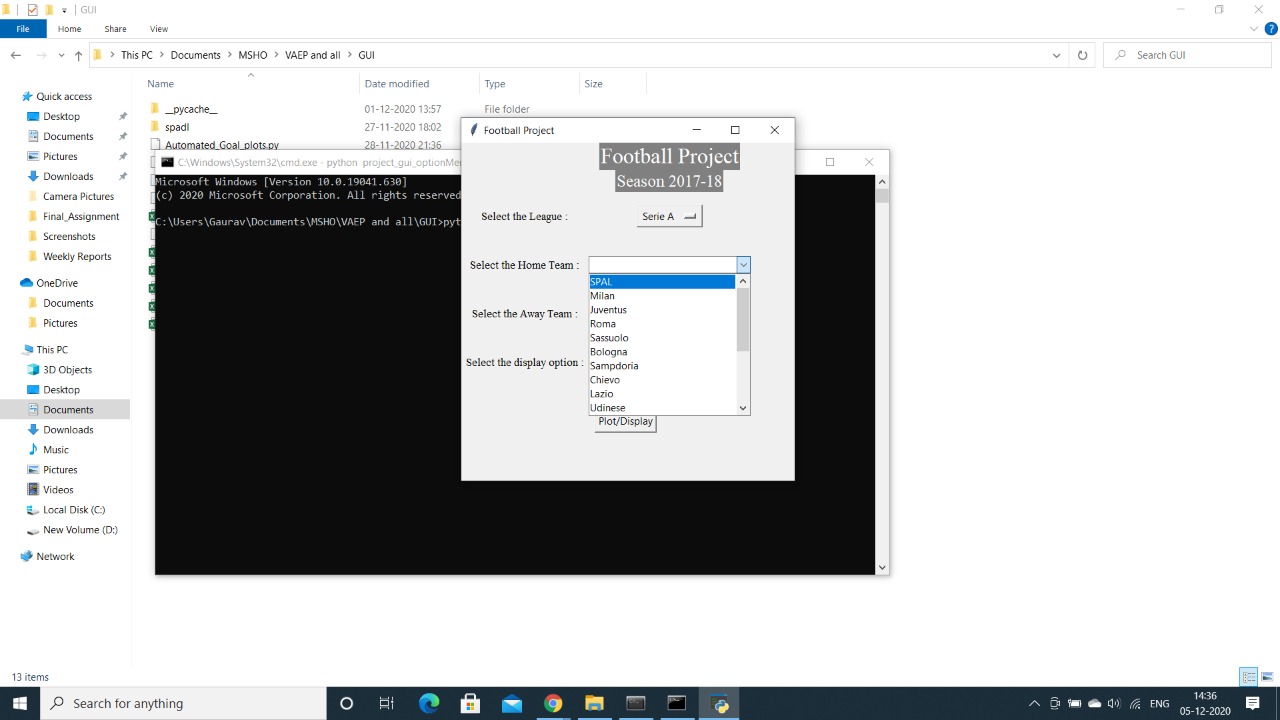


Fig 3. Selection of home team from GUI

1. Selecting the away team.

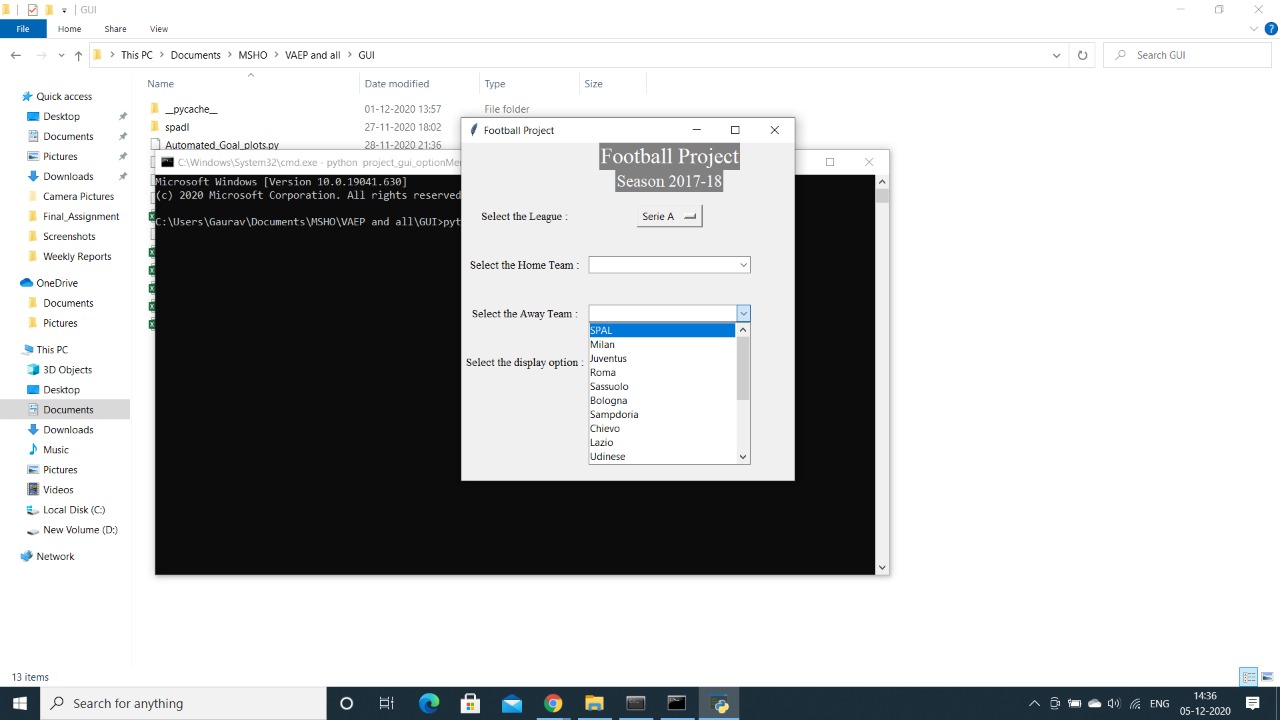


Fig 4. Selection of away team from GUI

1. Which thing user wishes to see Goal plots, VAEP ranking based on the league they have chosen.

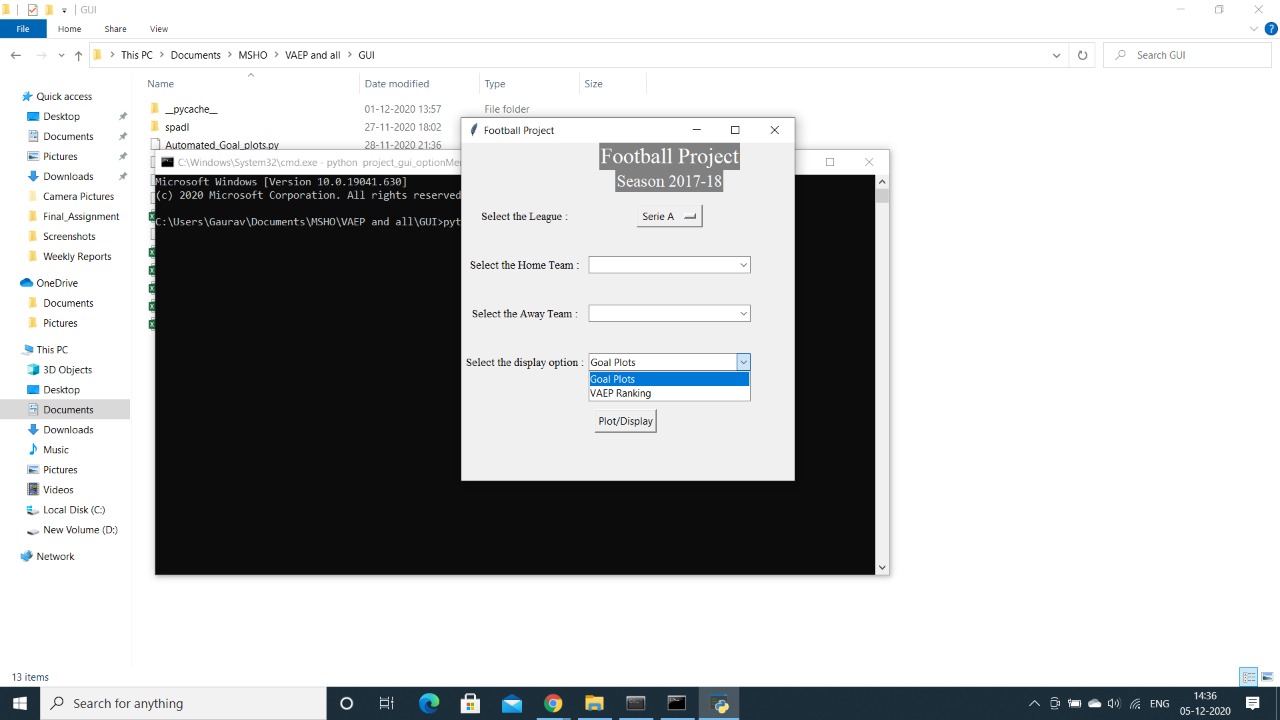


Fig 5. Selection of display method from GUI

Goal Plots:

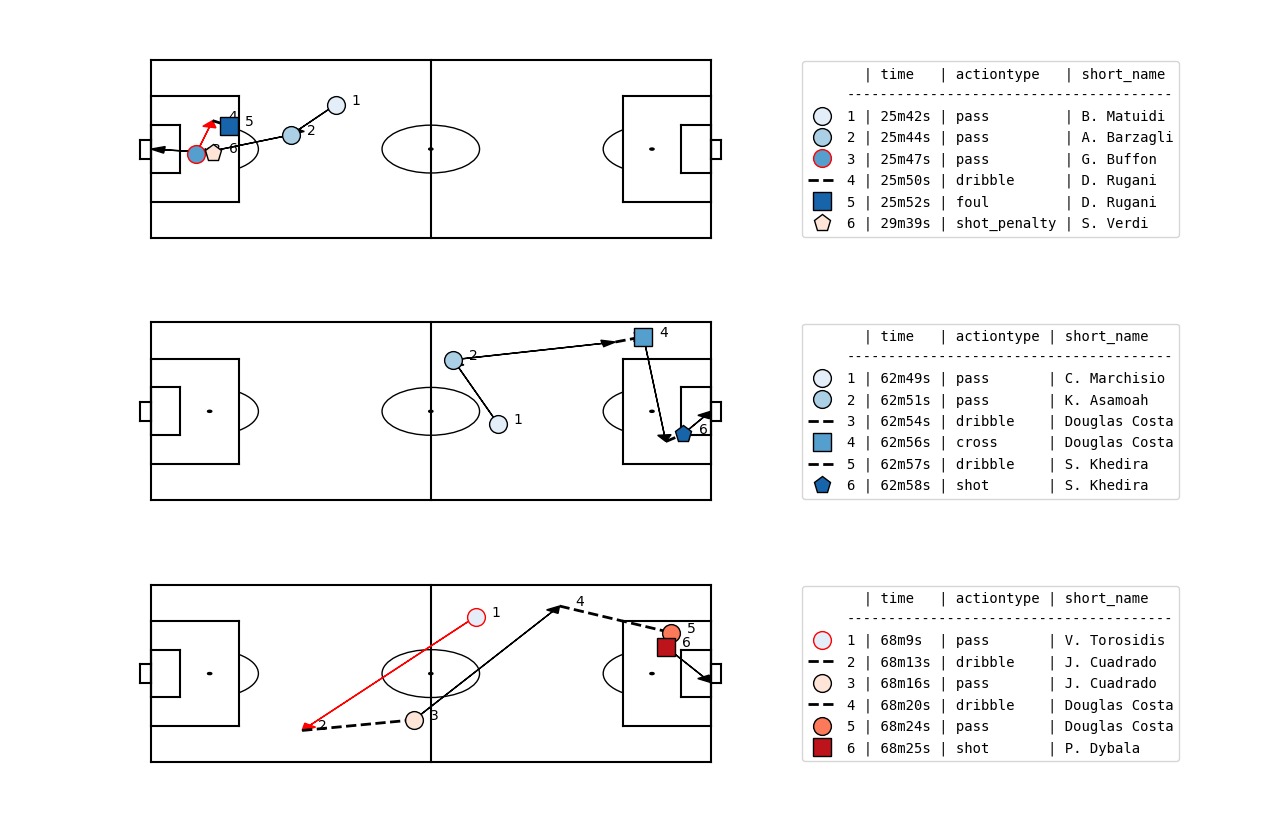


Fig 6. Goal plots shown in GUI

VAEP Ranking:

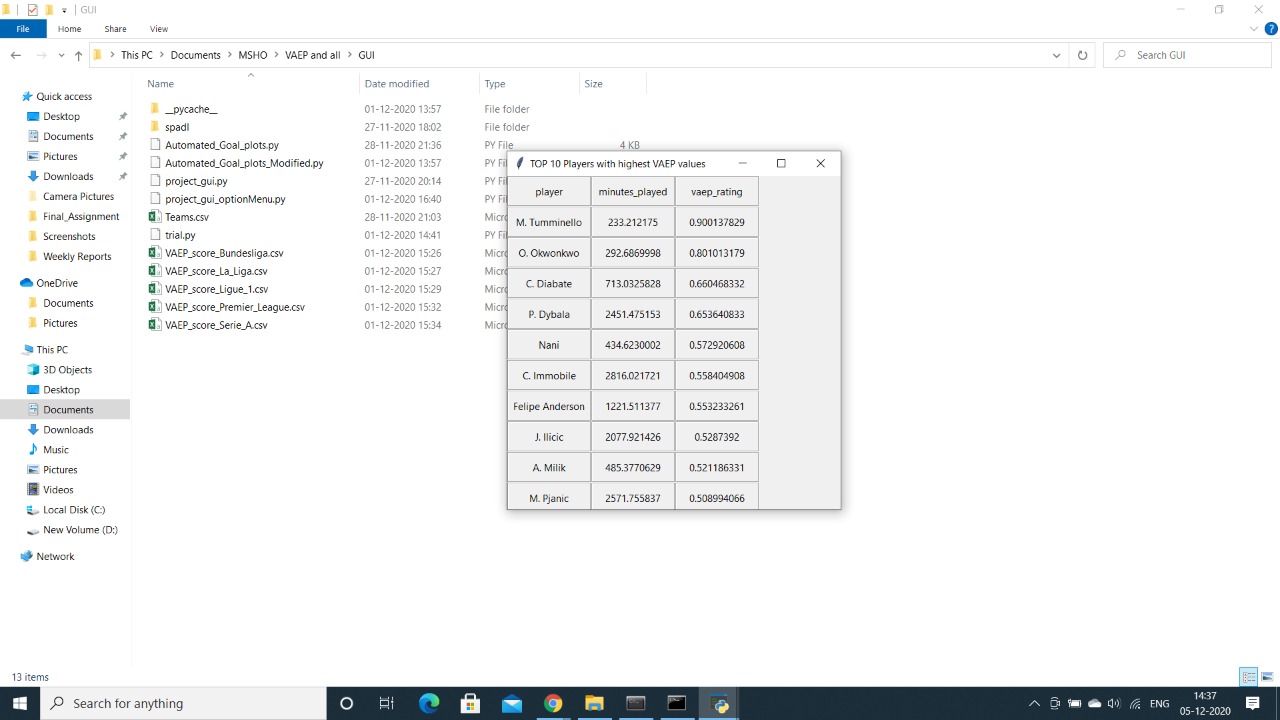


Fig 7. VAEP ranking shown in GUI

The goal plots are the ones which show the plots of actions leading to a goal and all such actions in a match played between the home and away team chosen by the user whereas the VAEP rankings show the top 10 ranked players from the league the user has chosen.

Once all the inputs are selected by the user just simply click on ‘plot/display’ as per choice of either goal plots or VAEP rankings, the plot or the rankings.

**Chapter 6. Re****sults**

**6.1 Results for Off-The-B****all events :**

1. **Analysis as per leagues:**
2. **Goals scored and No of matches in Leagues :**

Premier League was found to be the highest scoring league with 2631 goals in total, with La Liga a close second with around 2601 goals. Bundesliga was by far the lowest scoring league with 2122 goals scored. However, Bundesliga also had the least number of matches played indicating why it had the lowest goals scored tally.

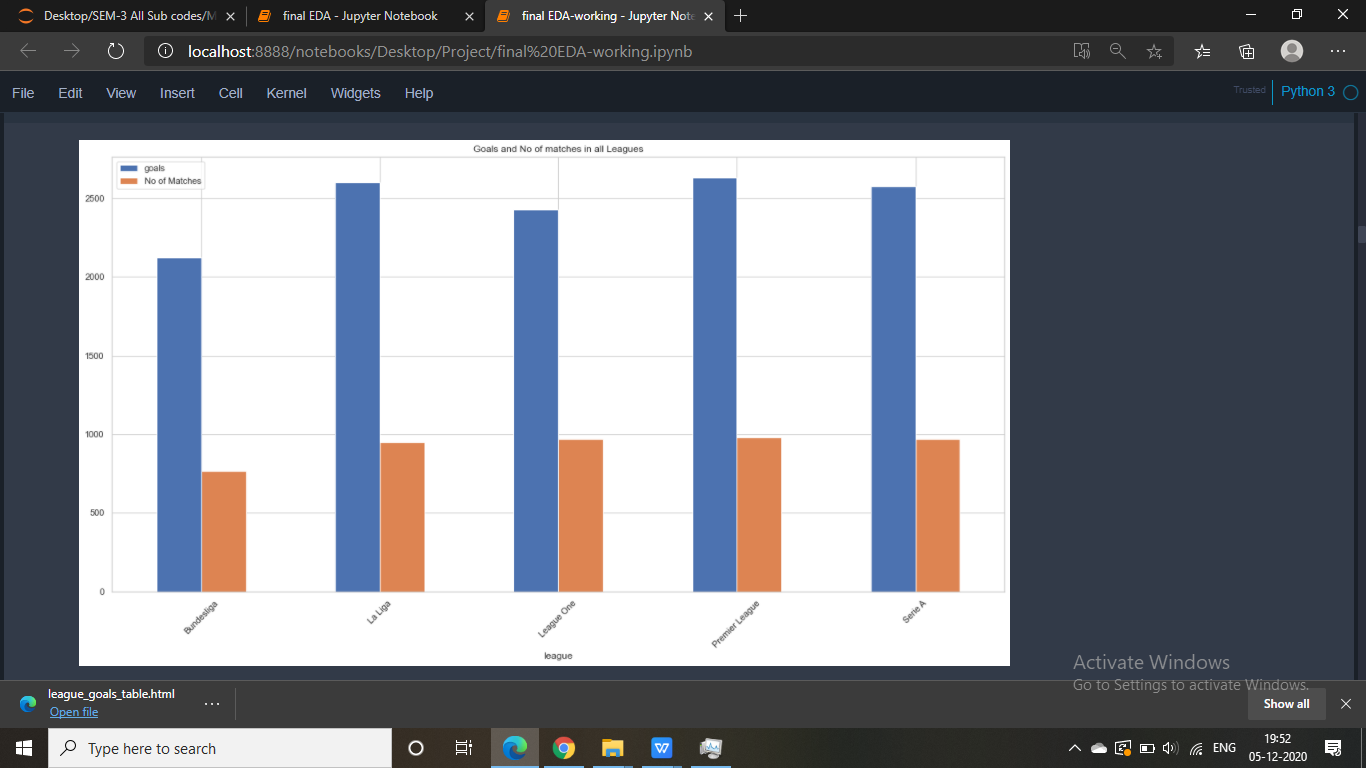


Fig.8 Number of goals and matches played as per different leagues

1. **Shots attempted per match in all leagues:**

Serie A had the most shots attempted per match with 26 shots being attempted in total by both the teams per match. On the other hand, La Liga had the least number of shots attempted per match.

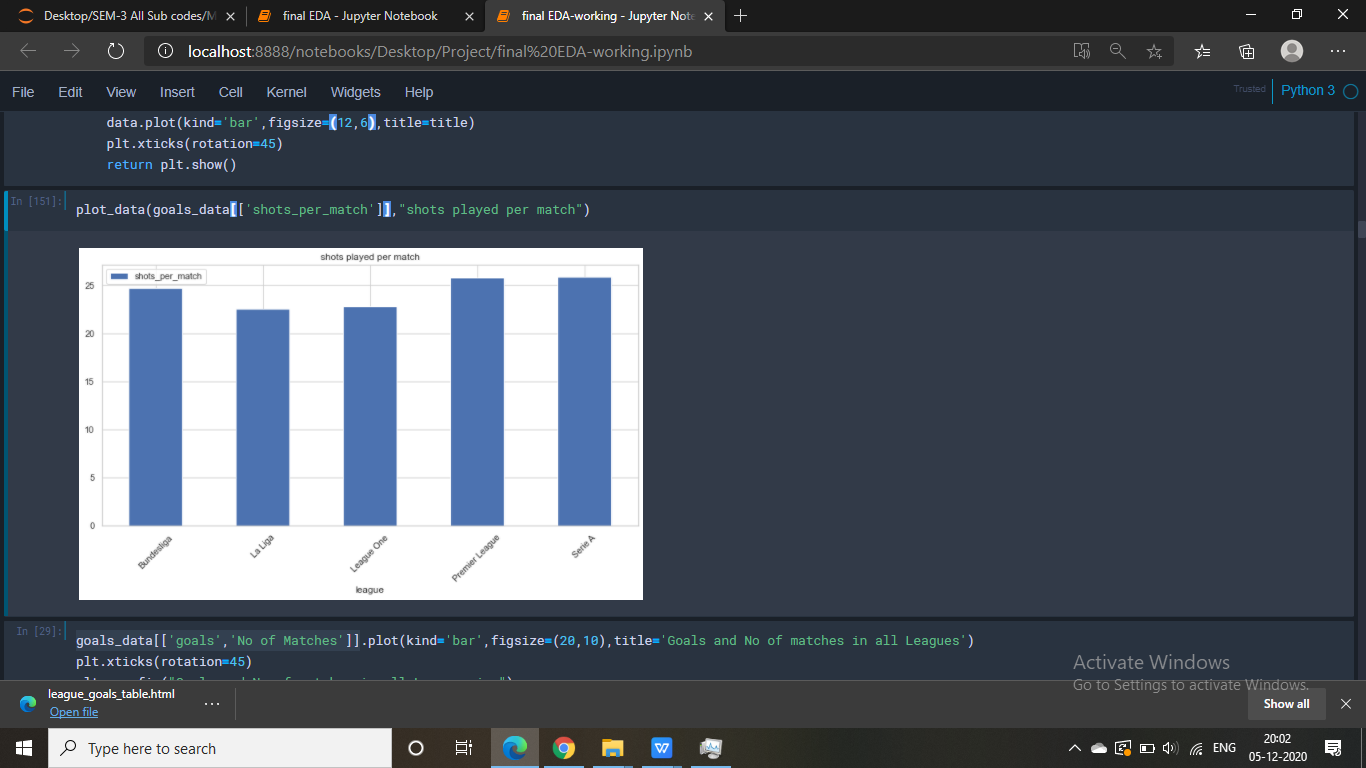


Fig.9 Number of shots per match in different leagues

1. **Goals Scored per match in all leagues:**

Bundesliga had the highest goals scored per match value with 2.76, even though it had lowest goals scored number. This is because of the fact that number of matches played in Bundesliga are the least among the 5 leagues.

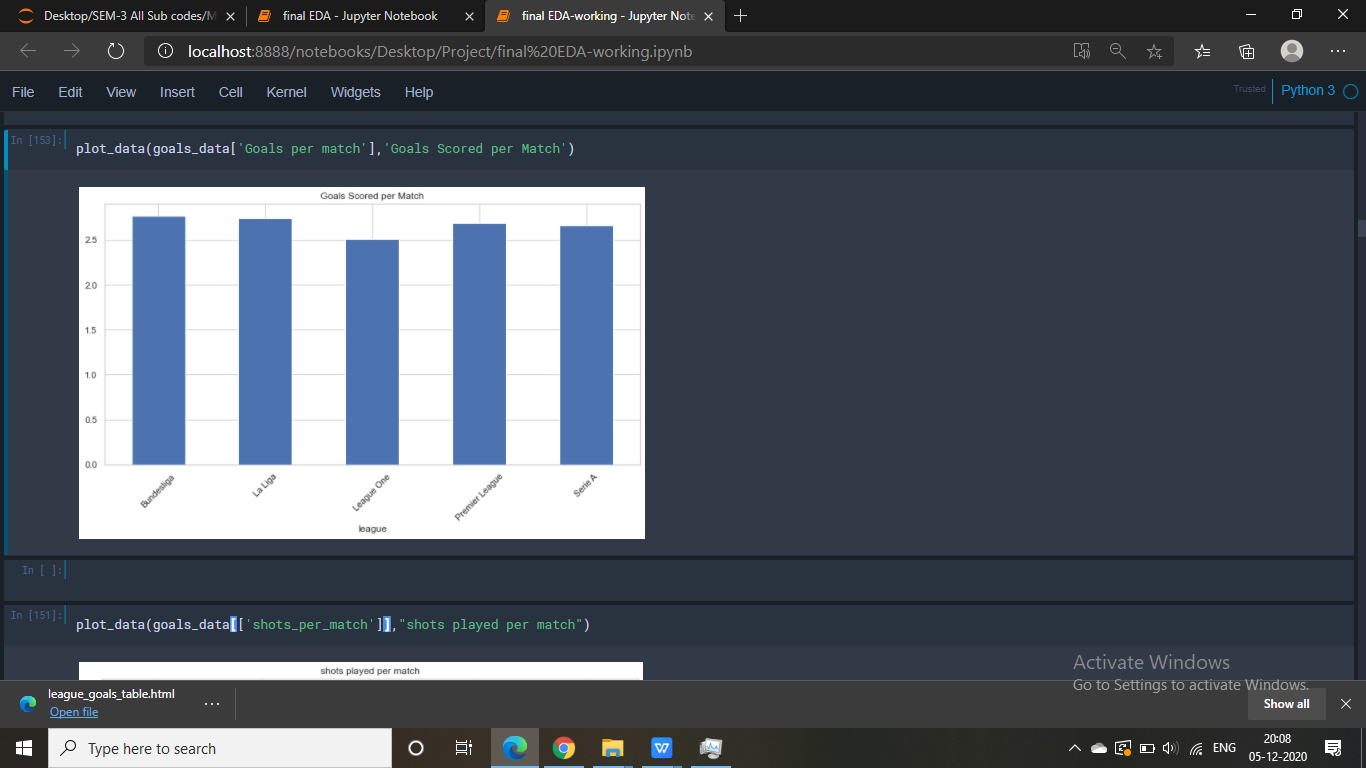


Fig.10 Number of goals scored per match in different leagues

**D.) Fouls commited per match in all leagues:**

Serie A has the highest fouls committed per match at 27.24 while Premier League has the lowest fouls committed per match at 20.80. This creates an impression that Serie A is one of the most aggressive and physically intensive league among the 5 while Premier League is the least aggressive and physically intensive.

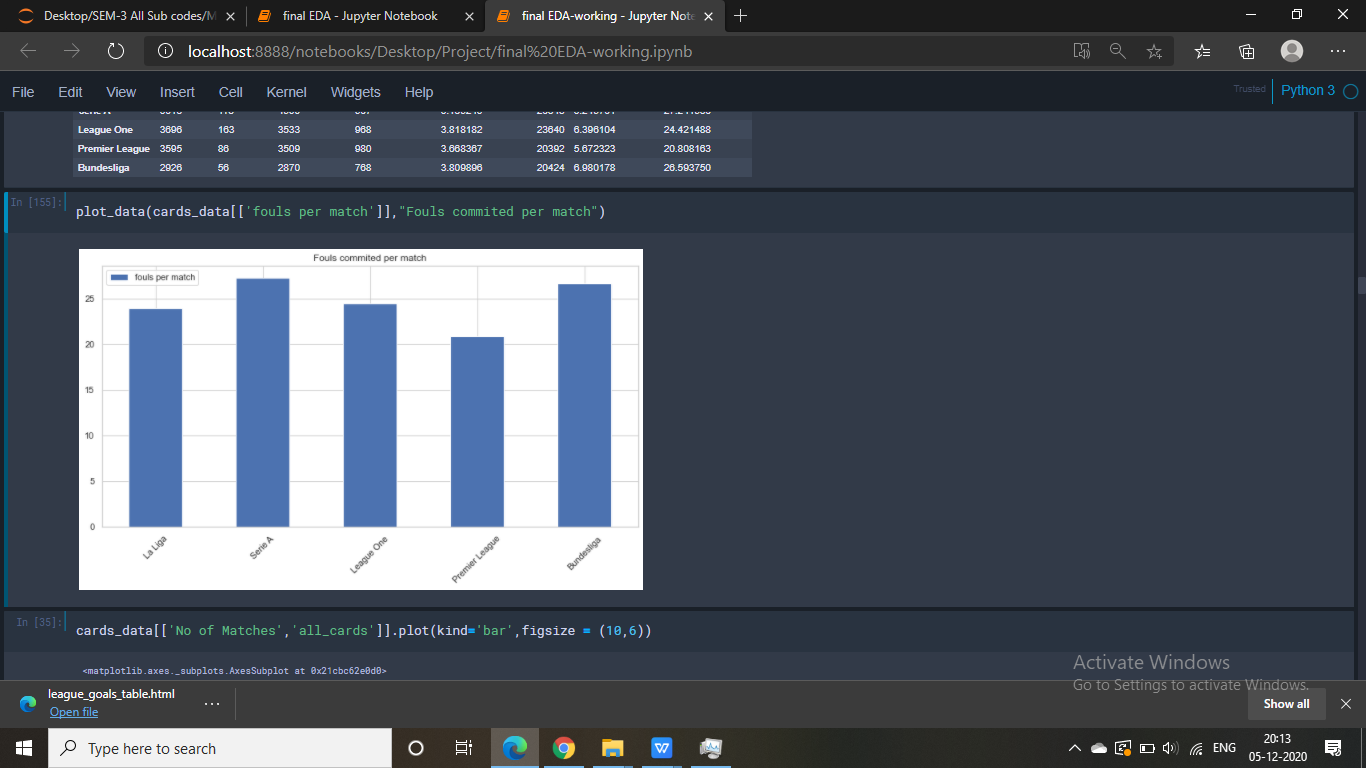


Fig.11 Number of fouls committed per match in different leagues

**E.) Fouls per cards and cards served per match in all leagues:**

Th number of cards issued per match in Serie A is 5.18, one of the highest, and that for Premier League is 3.66. This again reinforces our previous impression that Serie A is one of the most aggressive and physically intensive league among the 5 while Premier League is the least aggressive and physically intensive.

However, if we look at the fouls per card statistic, we can see that Premier League referees are far more lenient in booking for illegal plays and fouls than Serie A referees. The number of fouls per cards in Premier League is 5.67 while that in Serie A is 5.24, clearly showing that referees in Premier League allow more fouls to be committed before booking than Serie A referees. This can also be said about the previous 2 statistic (fouls committed per match and cards per match). Premier League have low values of these statistic because referees there do not deem minor foul plays as fouls committed and ignore them (or play advantage) as compared to other leagues (and especially Serie A and La Liga).

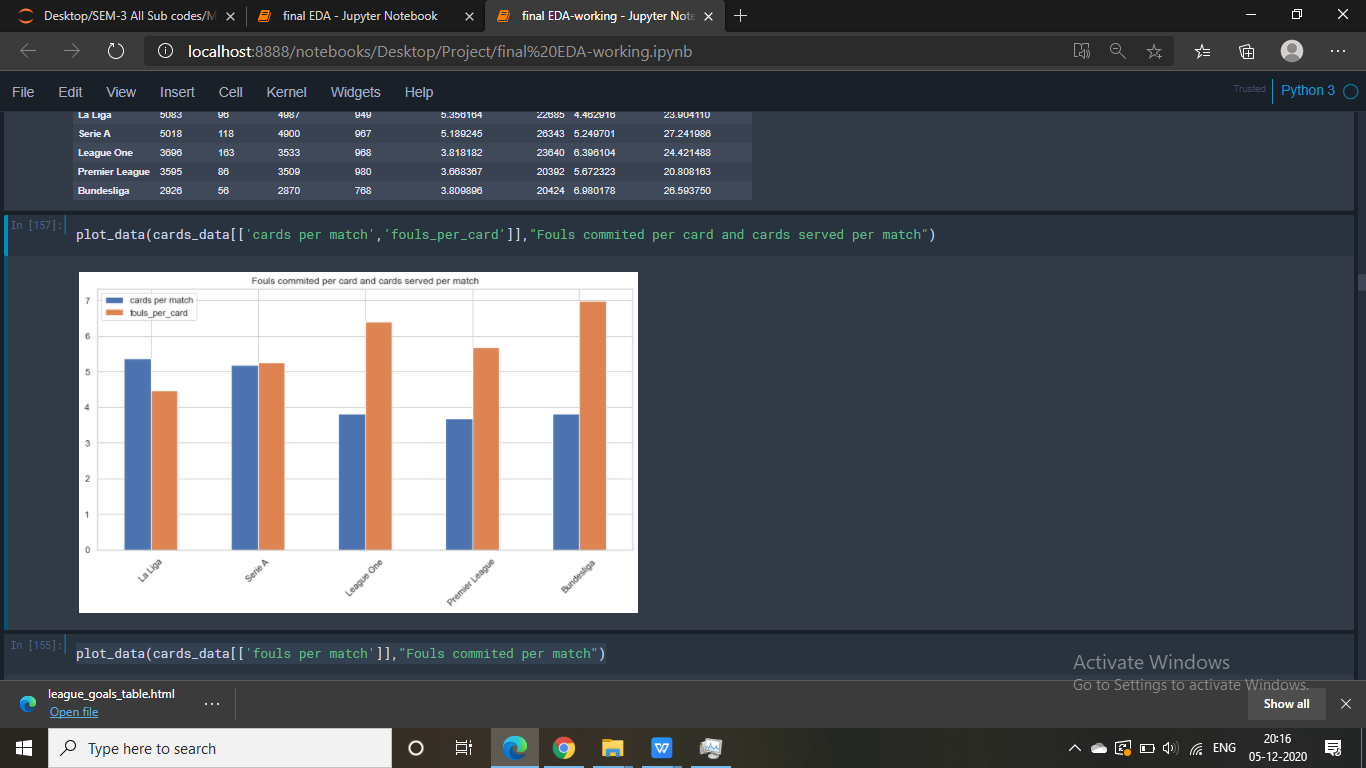


Fig.12 Cards per match and fouls per card in different leagues

1. **Analysis as per Teams:**

**# Analysis of Most Goal scoring teams**

**Goals and shots:**

Real Madrid with 276 goals in 1537 shots was the most offensive team amongst all the teams in the Top 5 leagues in Europe.

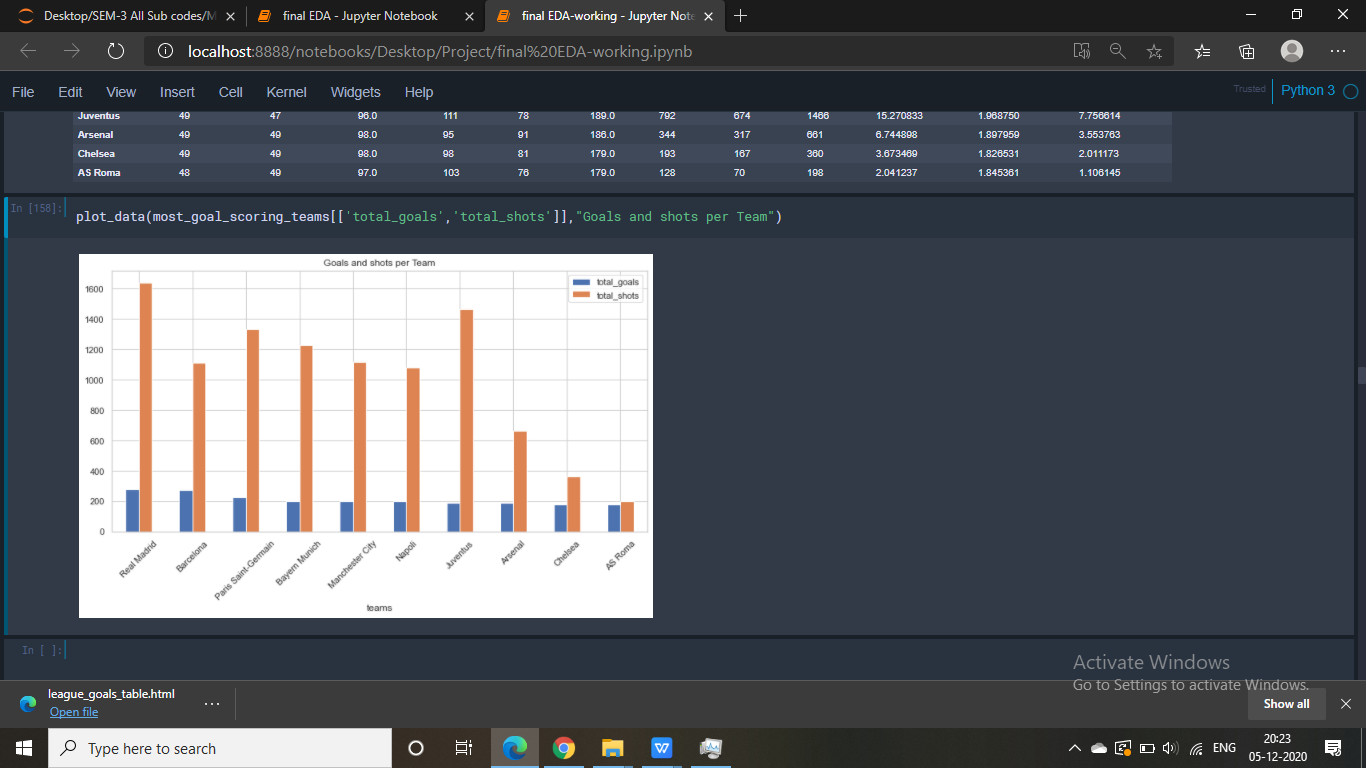


Fig.13 Most goal scoring teams

**Shots played per matches:**

Real Madrid have the most shots attempted per match with 17.41 with 2.93 goals per match. This shows that Real Madrid almost scored 3 goals per match. However, AS Roma was the most clinical team of all with 1.84 goals scored per match in 2.04 shots attempted per match.

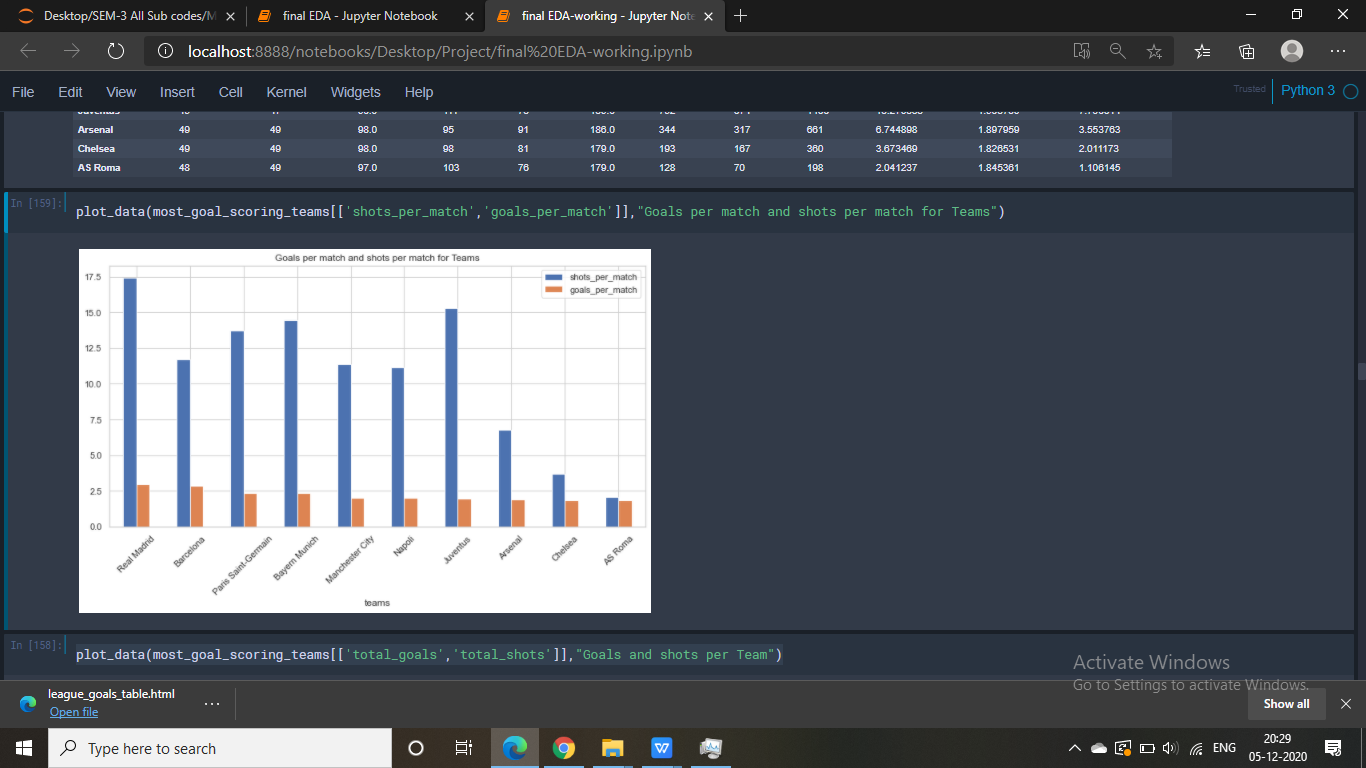


Fig.14 Shots played and goals scored per match

**# Analysis of Least Goal scoring teams:**

**Least Goals and shots:**

For the teams that played minimum 35 games, Cordoba had the least goals scored in the top 5 leagues with 22 goals.

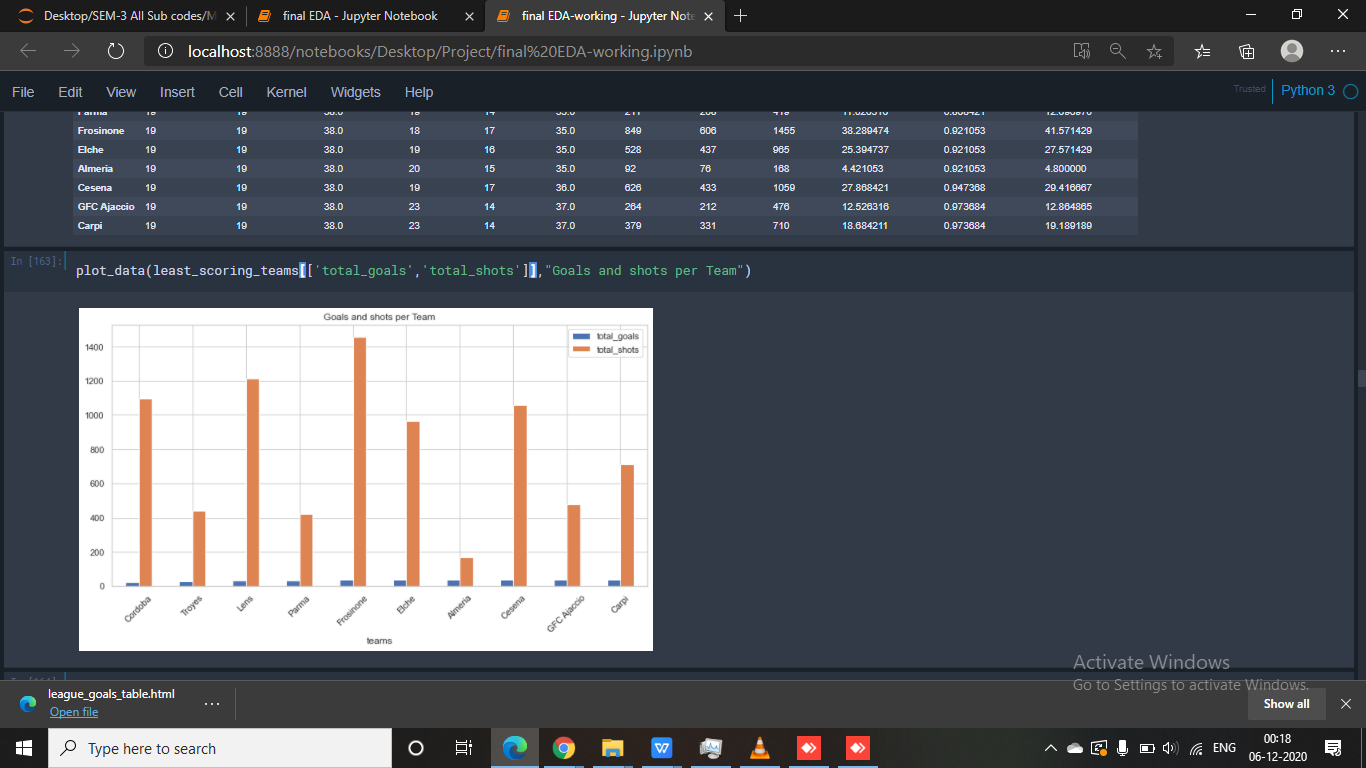


Fig.15 Least goal scoring and shot taking teams

**Shots played per matches:**

Team with least Shots per match was Almeria with 4.42 shots per match and Cordoba with 0.57 goals per match.

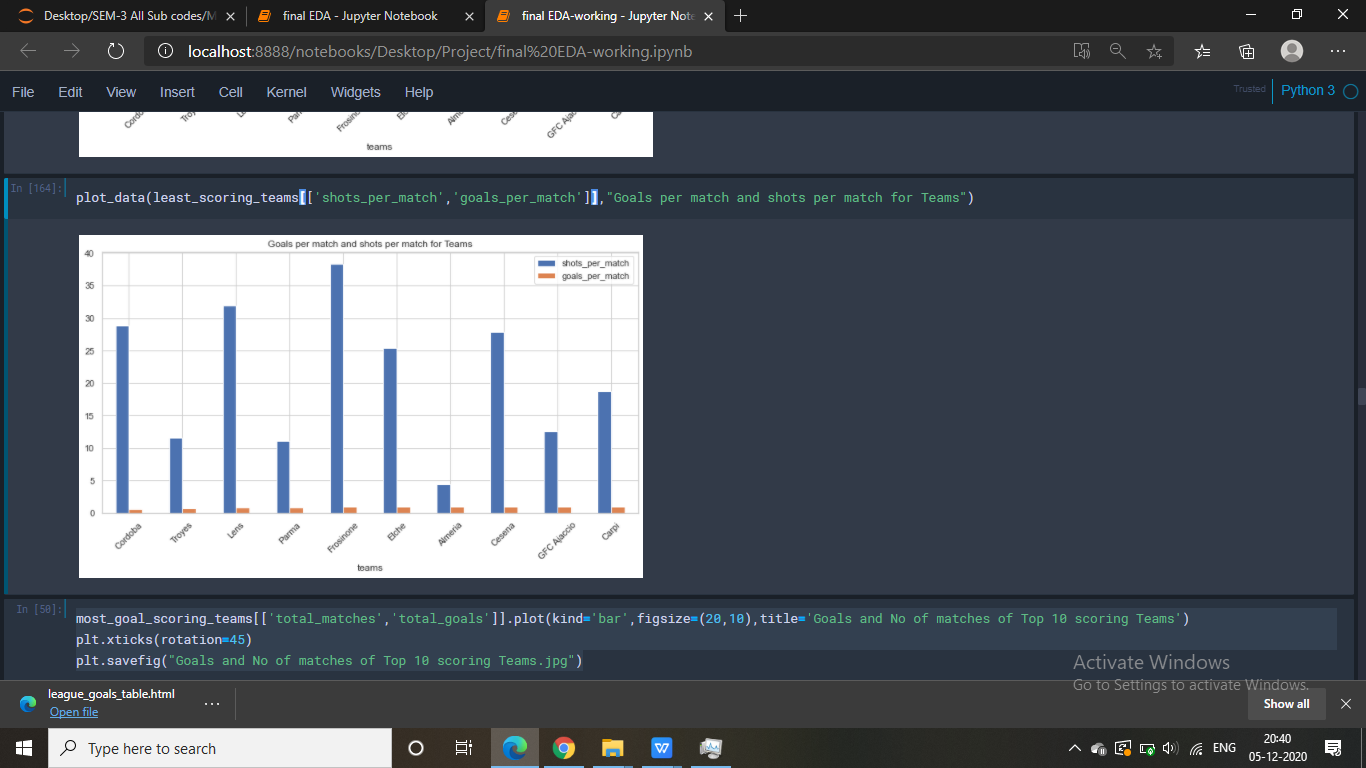


Fig.16 Shots and goals per match for least scoring teams

**Most Cards served:**

Granada had the highest number of cards issued to them with 300 cards.

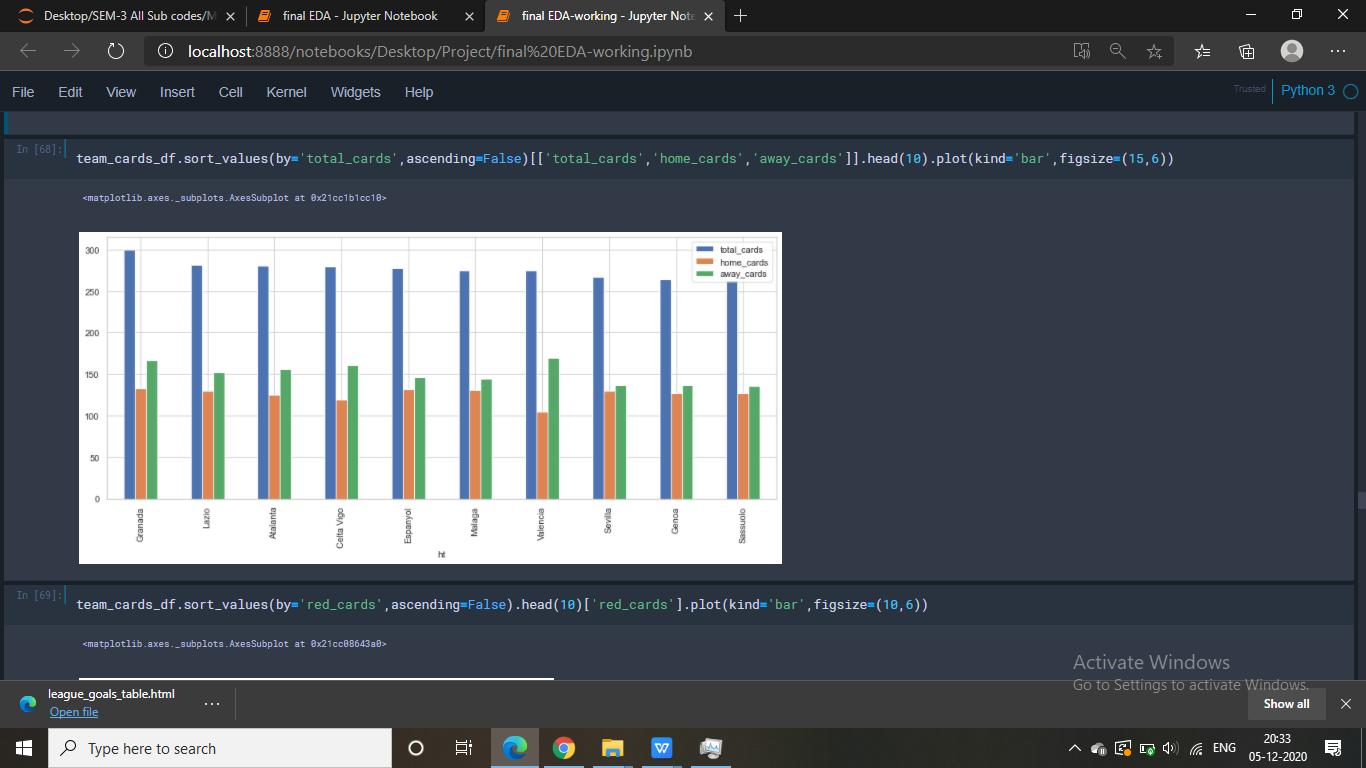


Fig.17 Cards served

1. **Analysis as per Players:**

**Goals scored and Matches played :**

Cristiano Ronaldo scored the most 89 goals in 80 matches, scoring more than a goal a game.

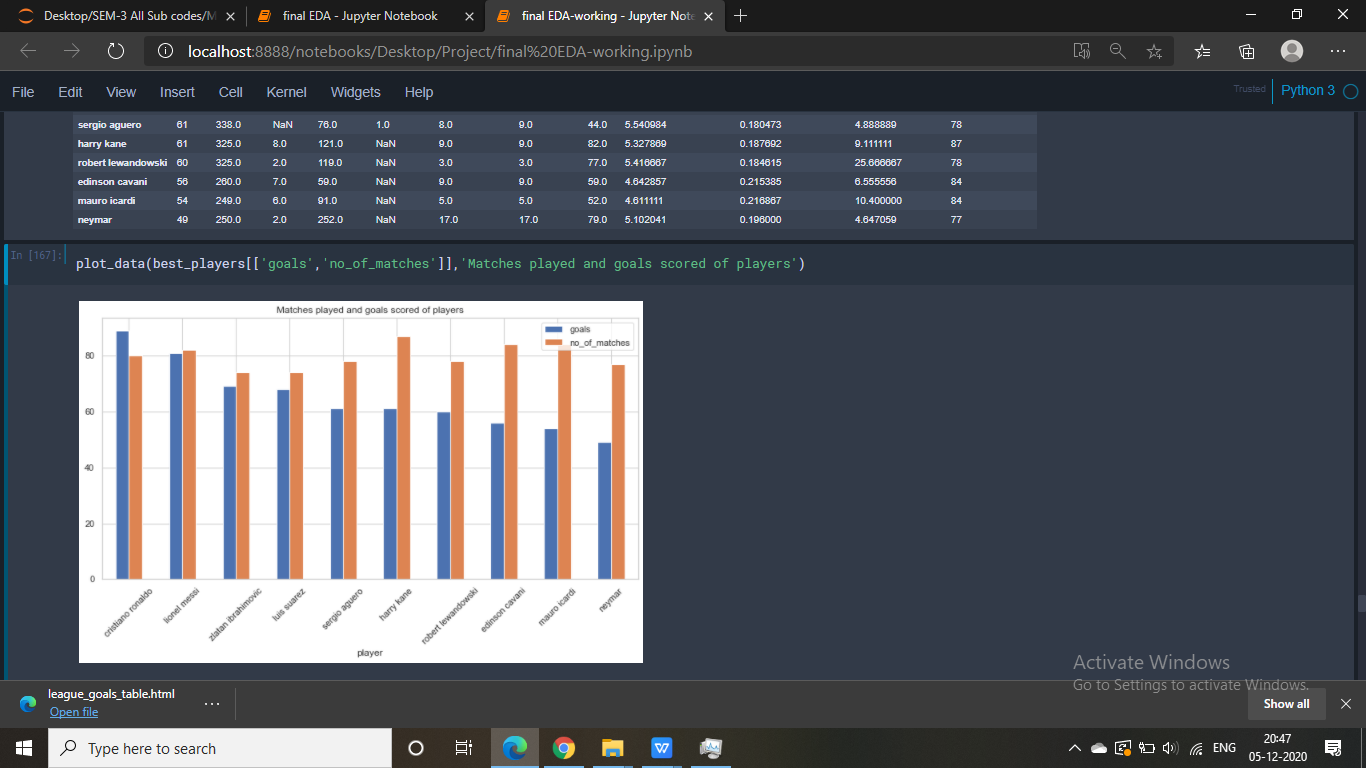


Fig.18 Most goal scoring players

**Attempts per Goal:**

Luis Suarez was the most clinical forward with 3.8 attempts needed per goal.

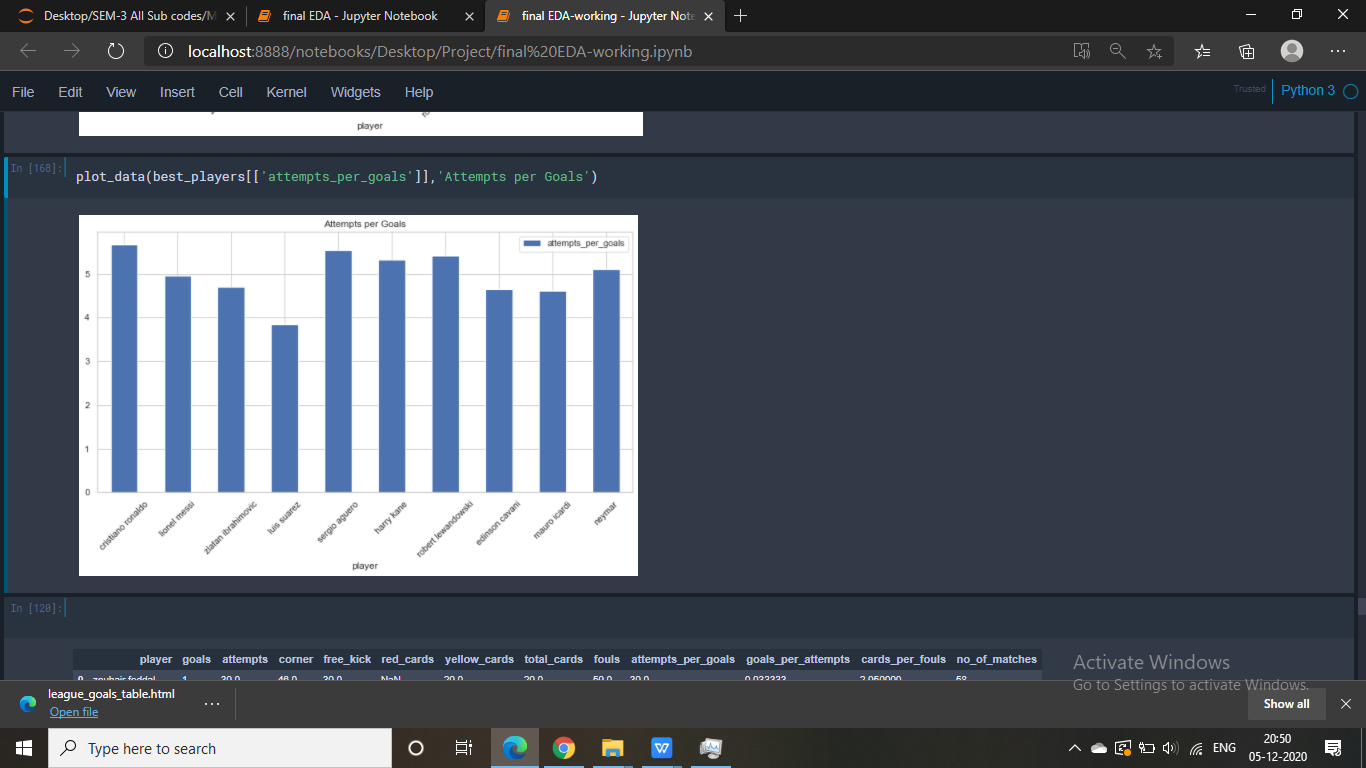


Fig.19 Attempts per goal for few top players

**6.2 Analysis of on-the****-ball EDA using Wyscout and Statsbomb matches events data set:**

In addition to above-mentioned various off-the-ball events, EDA on various on-the-ball events data made available to us by Statsbomb and Wyscout, was also done.

**Barcelona 2019/2020 season:**

The following figures show shot maps of both the El Clasicos (Barcelona vs Real Madrid) of 19/20 season:

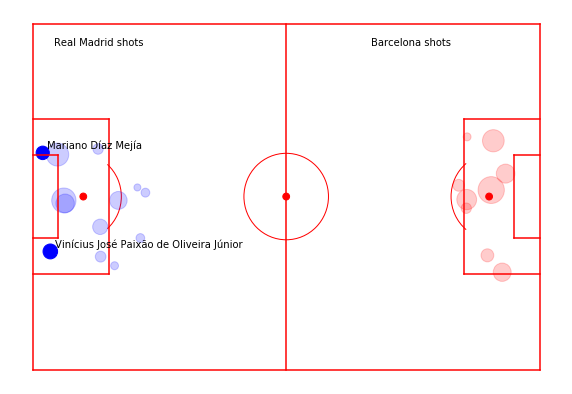


Fig. 20 El Classico Shot Maps: Real Madrid (H) vs Barcelona (A)

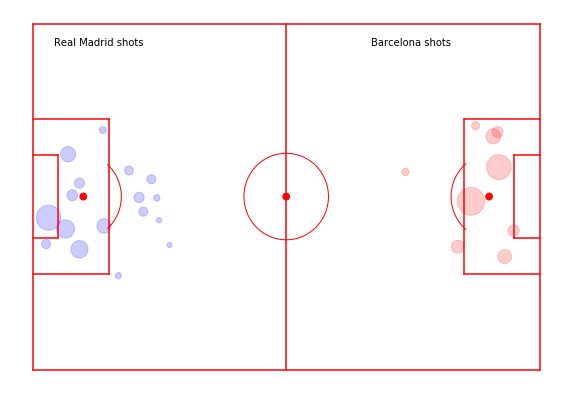


Fig. 21 El Classico Shot Maps: Real Madrid (A) vs Barcelona (H)

The circles are centered around the coordinates of the location where the shots are taken. Solid circles signify the shots that resulted in a goal while the size of the “shot circles” is directly proportional to the xG value of the shot (probability that the shot taken ends up being a goal) according to the formula:  
  
**Radius = 6\*(xG of the shot taken)1/2**

Names of the scorers are displayed alongside the solid circles.

The pass maps of Lionel Messi for the same matches are as follows:



Fig. 22 Lionel Messi pass maps vs Real Madrid (H)

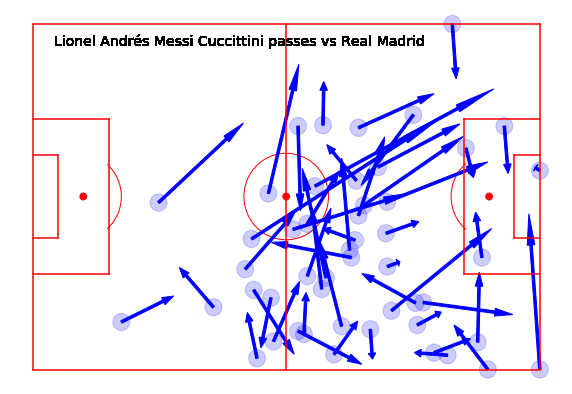


Fig. 23 Lionel Messi pass maps vs Real Madrid (A)

One can see from the above pass maps that although Messi is mostly deployed as a #10 player, in the hole just behind the striker (Suarez), he tends to drift all over the pitch, mostly to get on to the ball and control the game. However, the heavy concentration of passes on the right wing also suggests that he tends to drift towards the right wing and his passes and game involvements are predominantly on the right wing most probably because of his stronger left foot which will make him favour his left side and therefore making him prefer to cut inside on his stronger side from the right. These kinds of spatial dimensions of match events can provide us with information about a player’s behavior during a match, giving for example the possibility to determine a player’s profile from his average position during a match, or as seen above, the concentration of passes a player makes and influences a game from a given area of the pitch and his preference for that particular area.

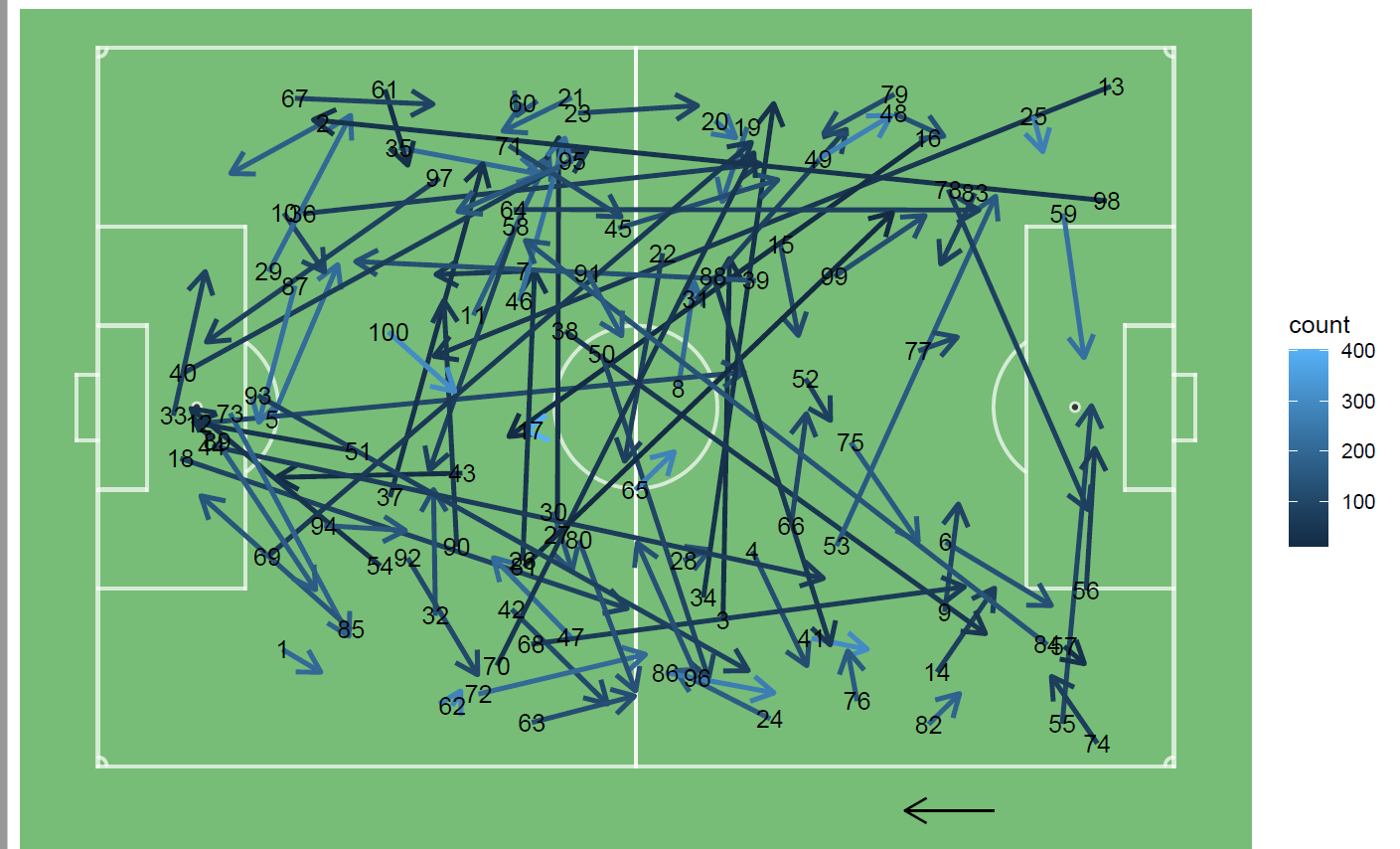


Fig. 24 Pass cluster of 100 most common passes of Barcelona 2019-2020 season

Along with this, the passing cluster of 100 most common passes of Barcelona is this season was also calculated. The clustering was done by K means cluster of 2,55,075 passes and 100 clusters with a restriction of a z –score of 2 w.r.t. that of number of passes in the league.

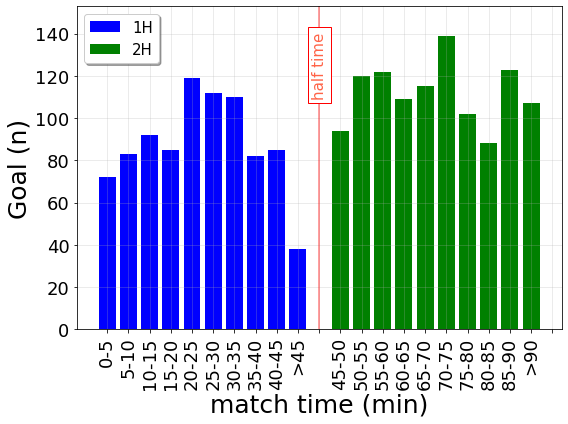


Fig. 25 Goals scored vs Time of the match



Fig 26. Red cards given vs Time of the match



Fig. 27 Yellow cards given vs Time of the match

By looking at *when* the events occur during a game, we can investigate interesting dynamics of teams and players. For example, **Fig. 25** above shows that goals are scored more frequently in the second half of the match, mirroring several of the possible factors that could affect scoring, such as a decrease of attention by the defenders towards the end of the match due to a loss of stamina and fatigue, or a more offensive attitude of the opponents who try to win or equalize the match, or the game gradually opening up leaving gaps in the defenses of both the teams, to be exploited by the respective opponents. Similarly, in **Figs 26** and **27**, we observe that the frequency of other rare events like yellow and red cards is the highest in the injury time. This aspect could highlight the presence of a bias by the referees who are less prone to award a card in the beginning of a match, a reduction of stamina or an increment of aggression of players at the end of the match.

**Passing Network analysis:**

By restructuring interlinks among the players of the same team, soccer-logs enable the analysis of the *interactions* between pairs of players through reconstructing the passing network of a particular team during a match. A passing network allows identifying the key players in the team, i.e., the ones having more connections to the teammates or a high passing activity or having a higher degree- anchoring and orchestrating the game-play.



Fig 28 Representation of the player passing networks of the match Napoli-Juventus. Nodes represent players,

edges represent passes between players. The size of the nodes reflects the number of ingoing and outgoing passes (i.e. node’s degree), while the size of the edges is proportional to the number of passes between the players.

**Figure 28** above shows two examples of a team passing network for the match Napoli - Juventus (Italian first division). Although Napoli engaged in more passes than Juventus (**666 vs. 332**), the two passing networks show similar average weighted out-degrees (**1.01 ± 0.93% and 1.10 ± 0.84%, respectively**). However, Juventus’ playing style resulted in a higher ***connectivity***(**19.89** as compared to Napoli’s **14.74**), defined as the network’s second smallest eigenvalue (i.e., a root of the characteristic equation of a matrix). This value indicates the robustness of a team, i.e., the strength of the links between its players. As a matter of fact, large values of connectivity between teammates are associated with a better overall team performance. The reconstruction of passing networks from soccer-logs enables several performance analyses. For example, by using the passing network and the players’ position during a pass it is possible to identify the most efficient tactical patterns across teams.

In addition to these we also analyzed player performances and evolution with time using Wyscout data.

**Player analysis using VAEP rankings:**

Using the method defined in previous sections on VAEP methodology, we calculated and plotted top 10 VAEP non shot actions (game plays) for Barcelona, Real Madrid, Juventus, Liverpool, Manchester United and Chelsea. Along with this, player rankings according to VAEP scores for La Liga, Serie A, Premier League, Bundesliga and Ligue 1 have been computed and tabulated in CSV files.

This was mainly done by first creating **.h5** files in **SPADL** format for each individual league and then creating required **“labels”** and **“features”** files from the **SPADL** files we created for the individual leagues. From these compiled files we created **“predictions”** file containing probability prediction of every action event occurring in a match and assigning the action a particular **“VAEP” value** according to its net probability score.   
With that done, we then generated a tabulated **“VAEP list”** csv file for the Top 5 European leagues.

Some of the Top VAEP non-shot action plots (non-shot game events with highest VAEP values) are:



Fig. 29 Non-shot VAEP events plots (Barcelona). VAEP rating: 0.535 (Zoomed-in)



Fig 30 Non-shot VAEP events plots (Juventus). VAEP rating: 0.454

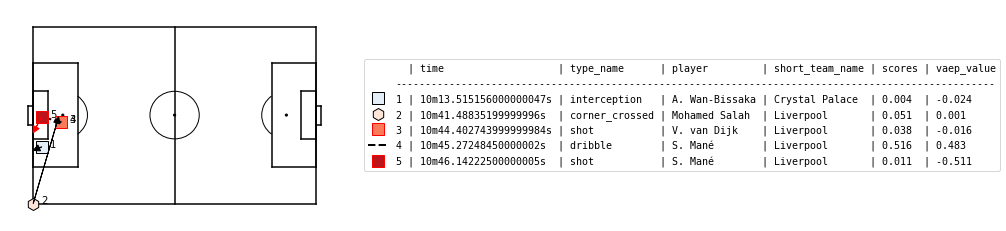


Fig 31 Non-shot VAEP events plots (Liverpool). VAEP rating: 0.483

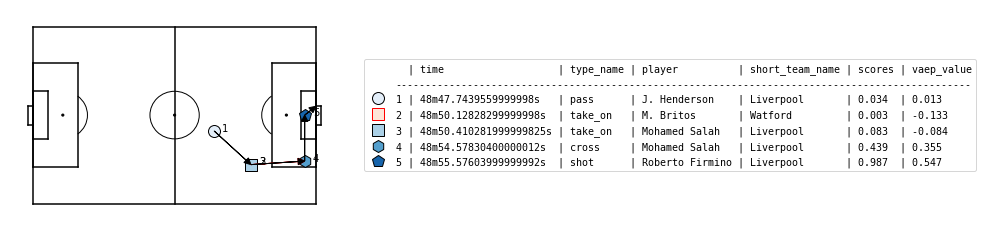


Fig 32 Non-shot VAEP events plots (Liverpool). VAEP rating: 0.547



Fig 33 Non-shot VAEP events plots (Manchester United). VAEP rating: 0.478

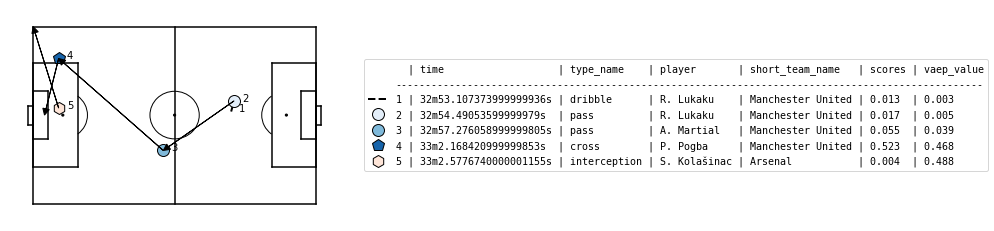


Fig 34 Non-shot VAEP events plots (Manchester United). VAEP rating: 0.468

**Top 10 Players according to VAEP values:**  
  
**1. Serie A:**

Table 5. Top 10 VAEP players in Serie A



**2. Premier League:**

Table 6. Top 10 VAEP players in Premier League



**3. La Liga:**

Table 7. Top 10 VAEP players in La Liga

****

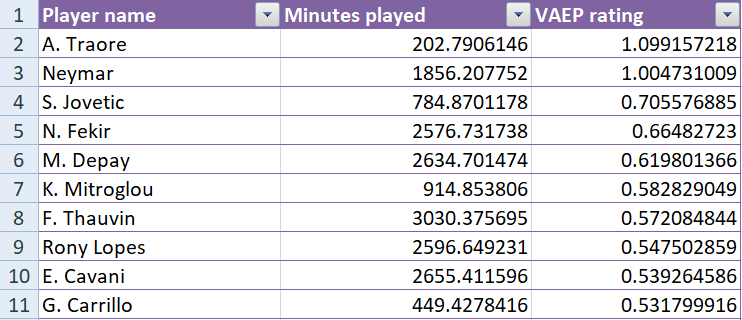
1. **Bundesliga:**

Table 8. Top 10 VAEP players in Bundesliga



1. **Ligue 1:**

Table 9. Top 10 VAEP players in Ligue 1

****

The top 10 VAEP scores of the above mentioned 5 leagues clearly show that in the season 2017-18, Lionel Messi and Neymar were amongst the top 3 players in those leagues (assuming the leagues are identical in terms of difficulty in playing). However, Adama Traore with a 1.099 VAEP score was clearly a revelation as an impact sub. His subsequent rise in footballing world now, as a marauding Right Winger at Wolverhampton Wanderers in the Premier League, is well known. Therefore, it is also clearly visible that the utility of VAEP ratings is not just in inferring an overall (over the course of the season) ranking system of top players, but it can also be utilised in scouting for exciting future prospects, which is also one of the greater strengths of EDA on-the-ball even data. Sergei Gnabry is another great example of this.

**6.3 Results for Predicti****ve Model**:

At current stage of the project we have used Random Forest regressor and SVR for predictions of matches. The results for the 2019 – 20 season real and predicted are as shown in following images.

Table 10. Results for Predictive model

|  |  |
| --- | --- |
|  |  |

Above tables show actual points table and predicted points table for season of 2019-20. In this model, we have obtained 67 % accuracy for top 6 team ranks and 33 % accuracy for bottom 3 teams. Similar to above seasons predictions we tried predicting for season 2018-19 season and 2017-18 season. For these seasons, the accuracy for top team was found similar to 2019-20 season but accuracy for bottom team was found to be lower. The different machine learning algorithms used in this project are Linear Regression, Random Forest Regression, Support Vector Regression and Deep learning. Similarly, for all the other leagues, these different machine learning approaches were taken to find out the performance. The performance tables for different leagues are as follows:

Table 11. Comparison table for different machine learning algorithms for Premier League

|  |  |  |  |
| --- | --- | --- | --- |
| Ranking Accuracy for Premier League | | | |
| Season Name | 2017-18 | 2018-19 | 2019-20 |
| Algorithm |
| Linear regression | 20 | 30 | 25 |
| Random Forest regression | 30 | 25 | 25 |
| Support Vector regression | 25 | 35 | 25 |
| Deep Learning | 30 | 30 | 25 |

Table 12. Comparison table for different machine learning algorithms for La Liga

|  |  |  |  |
| --- | --- | --- | --- |
| Ranking Accuracy for La Liga | | | |
| Season Name | 2017-18 | 2018-19 | 2019-20 |
| Algorithm |
| Linear regression | 20 | 10 | 35 |
| Random Forest regression | 25 | 20 | 25 |
| Support Vector regression | 20 | 25 | 35 |
| Deep Learning | 30 | 40 | 45 |

Table 13. Comparison table for different machine learning algorithms for Serie A

|  |  |  |  |
| --- | --- | --- | --- |
| Ranking Accuracy for Serie A | | | |
| Season Name | 2017-18 | 2018-19 | 2019-20 |
| Algorithm |
| Linear regression | 25 | 20 | 25 |
| Random Forest regression | 40 | 30 | 30 |
| Support Vector regression | 40 | 25 | 30 |
| Deep Learning | 45 | 25 | 30 |

Table 14. Comparison table for different machine learning algorithms for Bundes Liga

|  |  |  |  |
| --- | --- | --- | --- |
| Ranking Accuracy for Bundes Liga | | | |
| Season Name | 2017-18 | 2018-19 | 2019-20 |
| Algorithm |
| Linear regression | 38.88889 | 16.66667 | 16.66667 |
| Random Forest regression | 27.77778 | 27.77778 | 44.44444 |
| Support Vector regression | 27.77778 | 38.88889 | 38.88889 |
| Deep Learning | 33.33333 | 38.88889 | 27.77778 |

Table 15. Comparison table for different machine learning algorithms for Ligue 1

|  |  |  |  |
| --- | --- | --- | --- |
| Ranking Accuracy for Ligue 1 | | | |
| Season Name | 2017-18 | 2018-19 | 2019-20 |
| Algorithm |
| Linear regression | 10 | 10 | 15 |
| Random Forest regression | 40 | 20 | 20 |
| Support Vector regression | 20 | 20 | 35 |
| Deep Learning | 40 | 35 | 40 |

From above tables, it can be seen the accuracies for different algorithms are not very high. We tried to find the potential reasons behind such less accuracy and they are as follows:

1. The model does not take into consideration of the dynamic nature of the game.
2. The model is developed based on parameters such as shots taken by both teams, corners and free kicks taken by both teams, betting odds and expected goals for both the teams. The expected goals calculation is based on past performance of the club; that being said, it may or may not be possible for a club to reproduce its performance in consecutive seasons.
3. Maybe only last season data for training is not sufficient.

The potential remedies for the said reasons can be as follows:

1. Developing a model which takes in consideration team, player ratings, performance of teams in recent games.

2. Updating the expected goals, offensive and defensive rating values for the teams after every match the clubs plays, using them for the predictions.

3. Taking into consideration, data for past several seasons, giving a proper idea of performance structure of a club.

**Referen****ces**

1. Pappalardo, L., Cintia, P., Rossi, A. et al. A public data set of spatio-temporal match events in soccer competitions. Sci Data 6, 236 (2019). <https://doi.org/10.1038/s41597-019-0247-7>
2. Decroos T., Lotte B., Jan Van Haaren, Jesse D., Actions Speak Louder than Goals: Valuing Player Actions in Soccer, Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2019)
3. <https://dtai.cs.kuleuven.be/sports/blog/exploring-how-vaep-values-actions>
4. <https://znstrider.github.io/2018-11-11-Getting-Started-with-StatsBomb-Data/>
5. https://www.kaggle.com/secareanualin/football-events