FWD Udacity Professional Data Analysis Nanodegree

Project 2

Investigate a dataset

European Soccer Database

By: Karim Taher Ibrahim



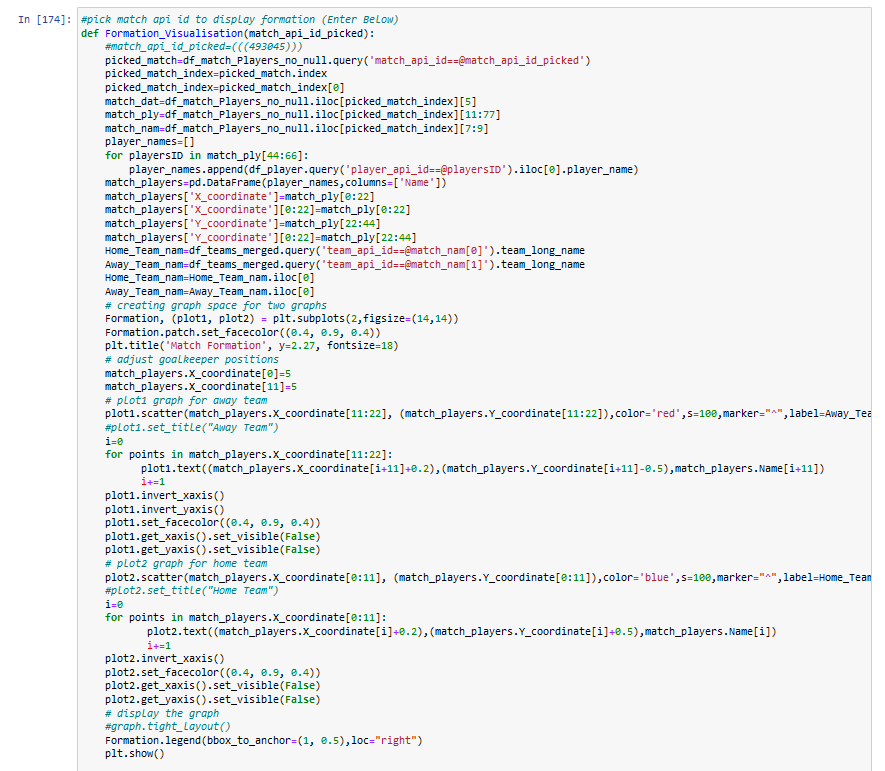
1. Questions asked:
2. What is the formation of a certain team match   
   in a certain match?
3. Which formation has the highest winning chances?
4. Does higher possession affect winning chances?
5. Which bidding agency has the most accurate biddings?
6. Which team has improved the most between   
   2008/2009 season and 2015/2016 season?
7. Which was most consistent team across the 8 seasons period?
8. Which player scored the most goals of a certain subtype (for example bicycle kicks) across the 8 seasons period?
9. Which players, teams and leagues had the   
   highest committed fouls per game?
10. Which players were fouled the most per game?
11. Question (a) (formation visualization):

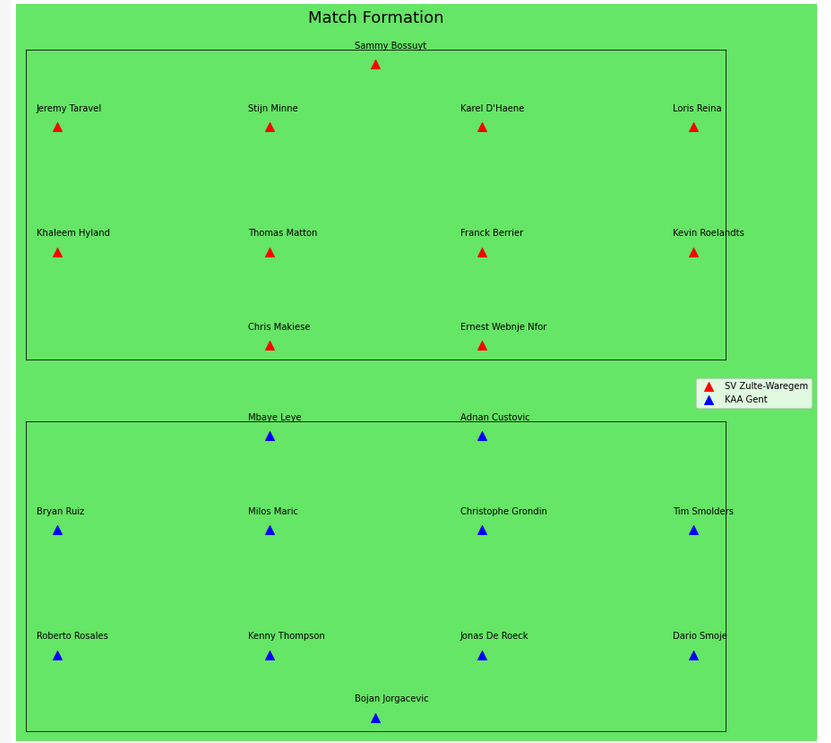
-First, matches with null values in both player coordinates and   
 player api id were removed from the dataframe using the code below:

-Next, using the match api id as input, a new dataframe was created for   
 the names and x and y coordinates of each player in this match:

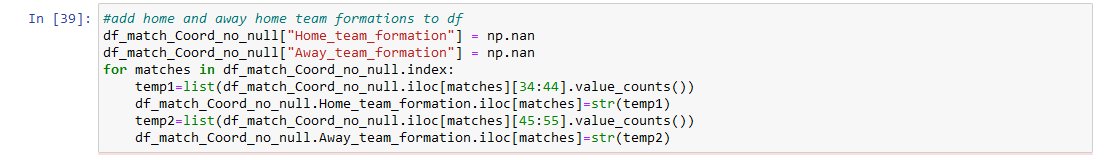
-Finally, a scatter plot was created showing the formation of the   
 teams in this match:

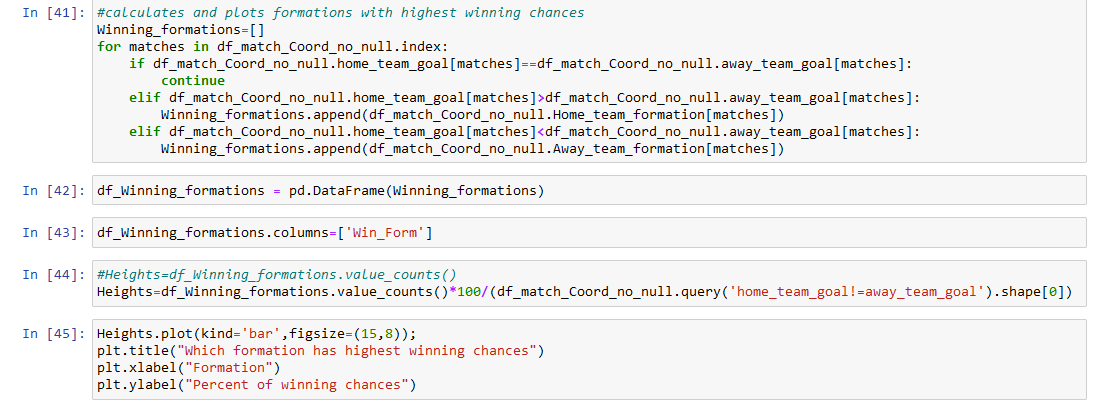


 -The code parts were then defined into one function named   
 Formation\_Visualisation:

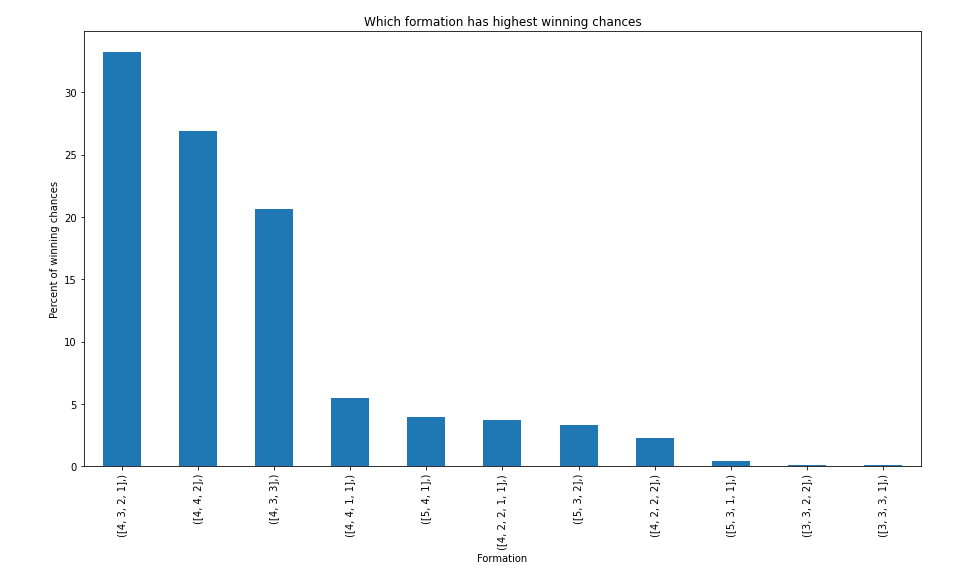
-Here is the final result:

1. Question (b) (Which formation has the highest   
   winning chances?):

-First, two new columns were added to the dataframe named   
 df\_match\_coord\_no\_null, each contains the formation of the home   
 and away team.

-Then, winning formations were appended to a list and a bar chart   
 was plotted with the results.

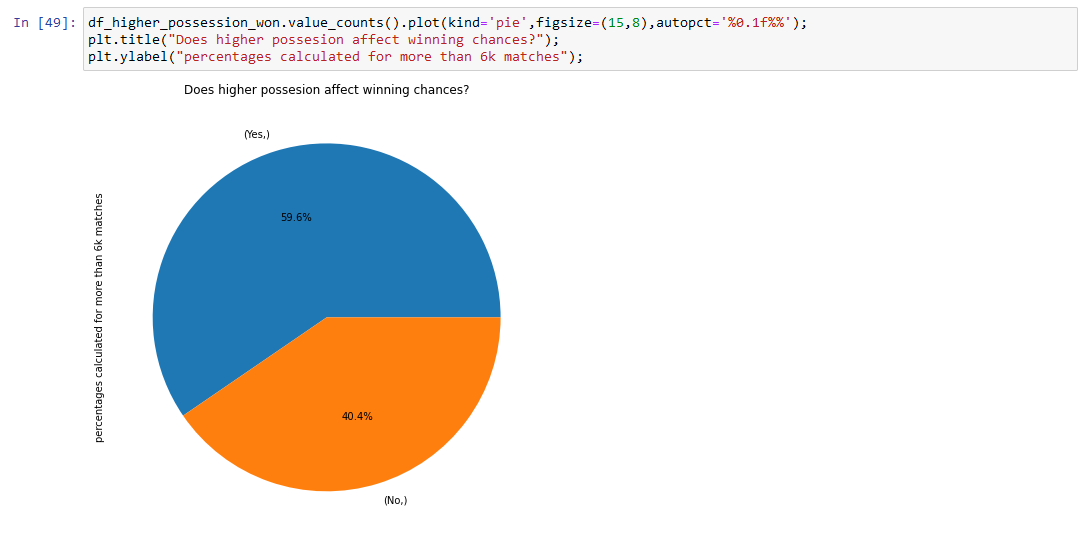
-The bar plot is displayed below:



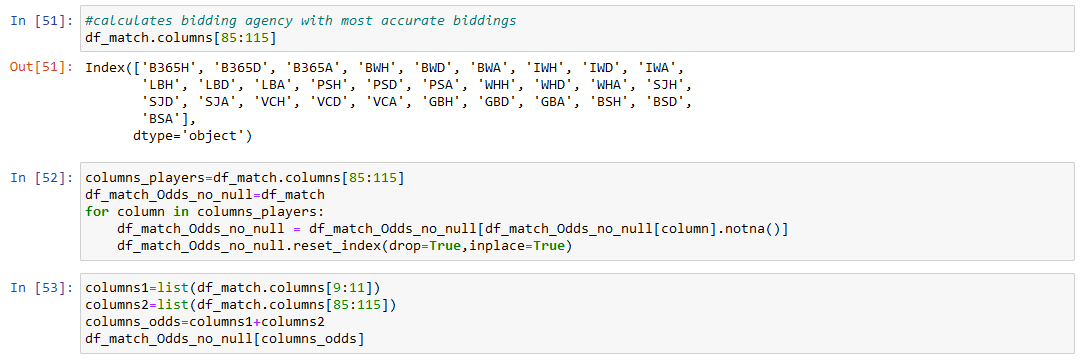
1. Question (c) (Does higher possession affect winning chances?):

-The code below uses beautiful soup to extract possession rating at the end of the match for both home and away team then if the team with the higher possession won yes is appended to a list:

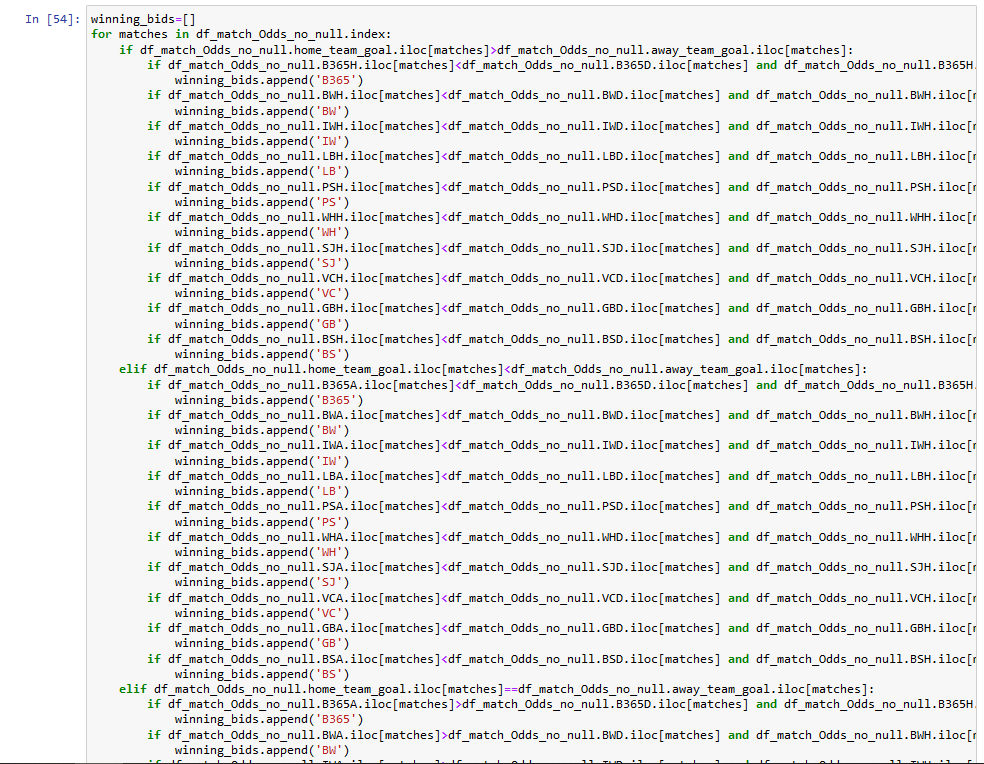
-A pie chart was plotted with the results:

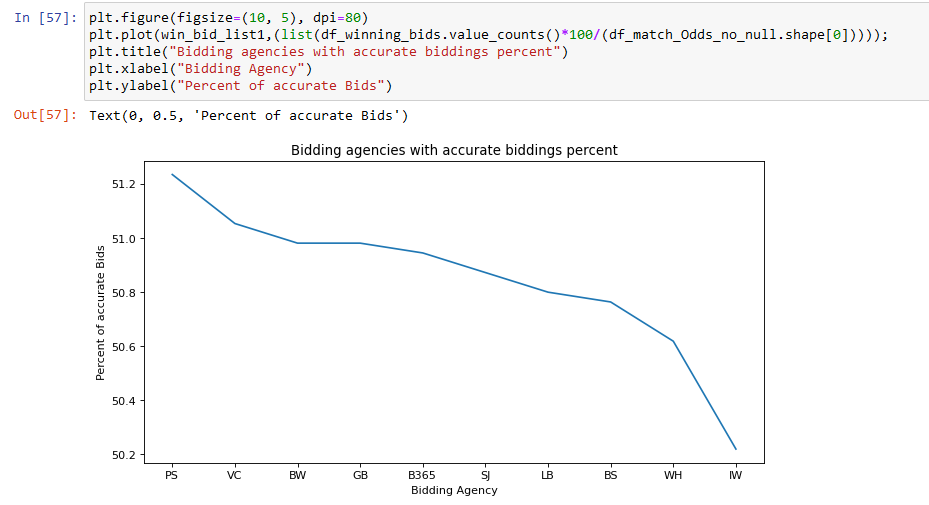


1. Question (d) (Which bidding agency has the most accurate biddings?):

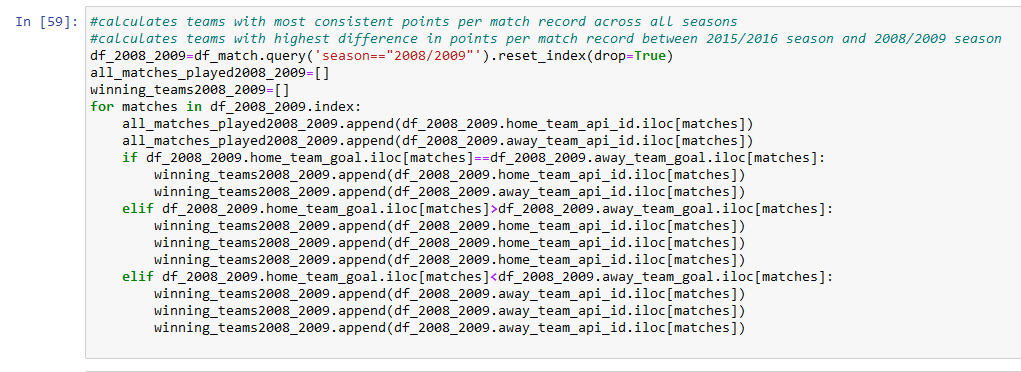
-A New dataframe was created with rows that have no null values in all   
 bidding columns:

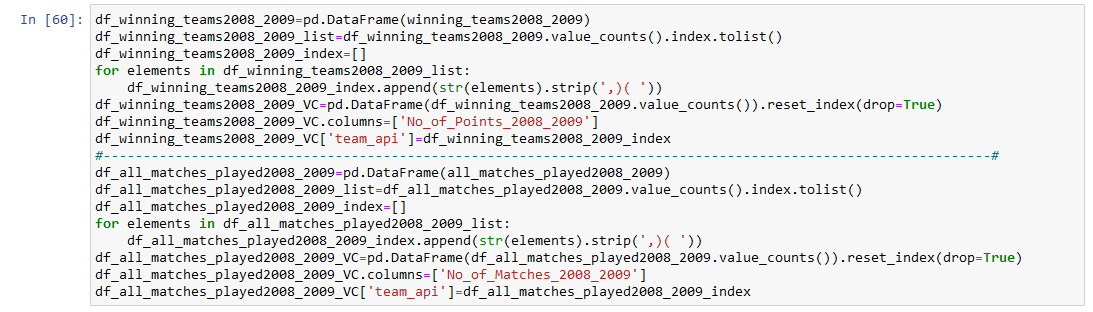
- Next, if the lowest bid (most probable) is equal to the real match   
 result, the bidding agency is appended to a list:



- A line plot is used to show the differences between the bidding   
 agencies:

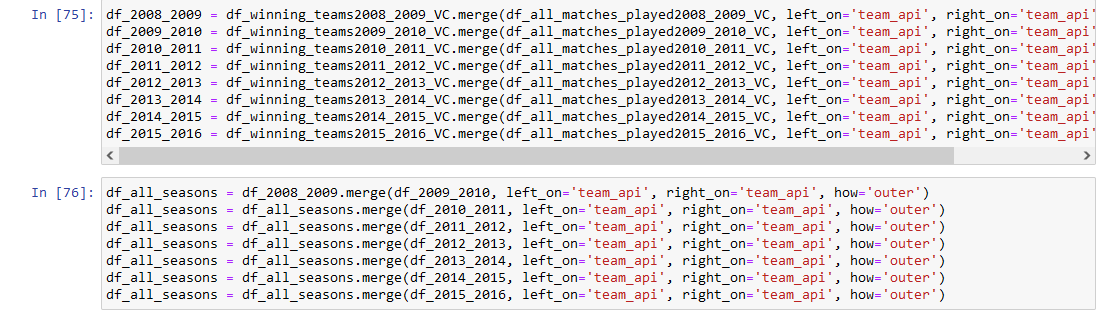
1. Question (e & f) (Which team has improved the most between 2008/2009 season and 2015/2016 season?) (Which was most consistent team across the   
   8 seasons period?):

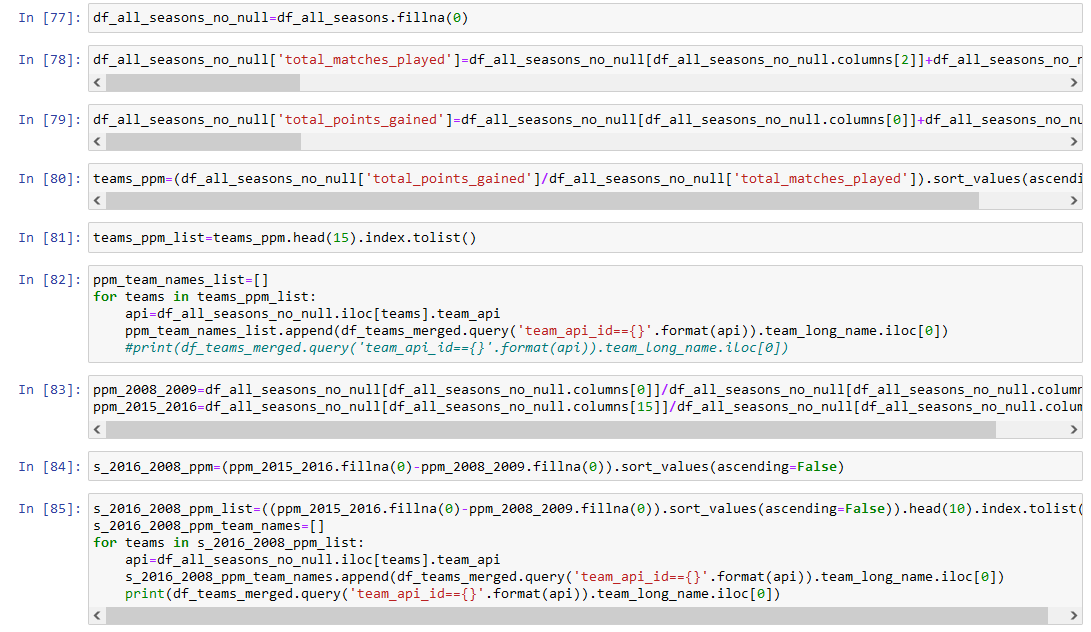
- Number of points for each team during 2008/2009 season was   
 calculated below:

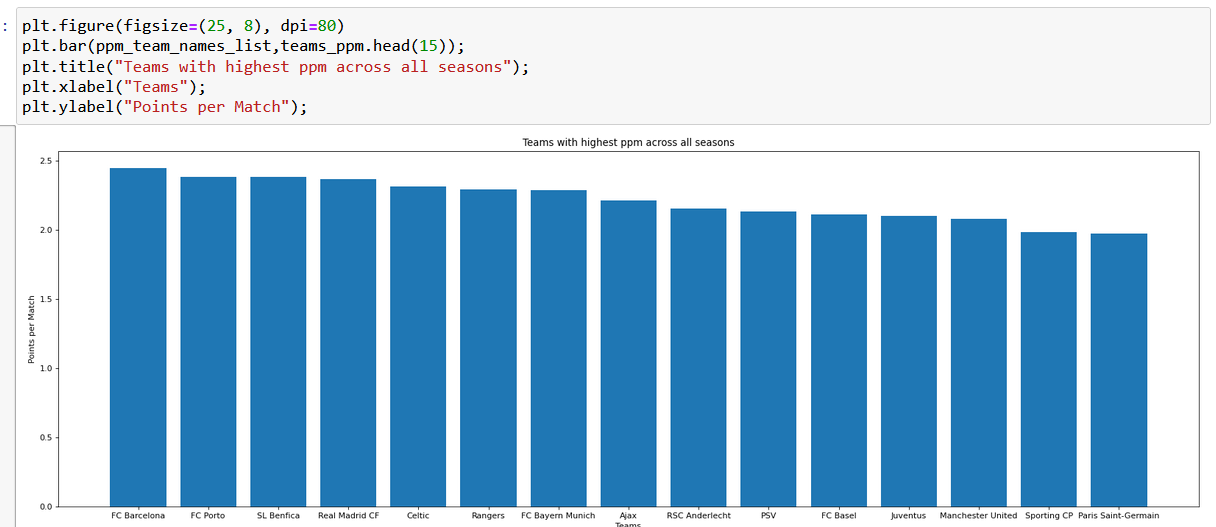
- Two new dataframes were created containing the number of points for   
 each team this season and number of matches played:

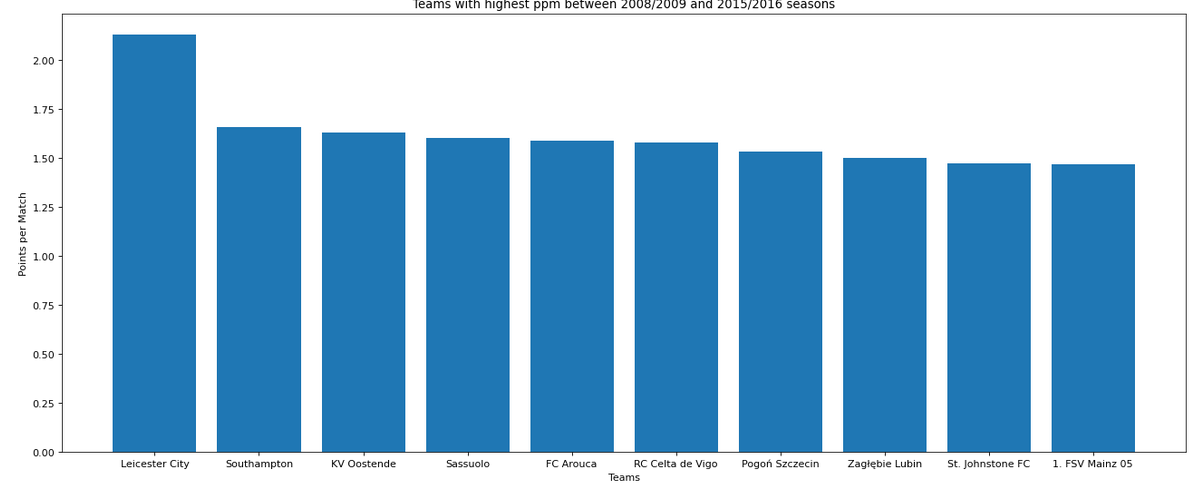
- The process was repeated for all other seasons:



-The dataframes for total points and total matches values were   
 counted then merged to one dataframe for each season and all the   
 seasons were used to create a dataframe containing   
 all seasons combined:

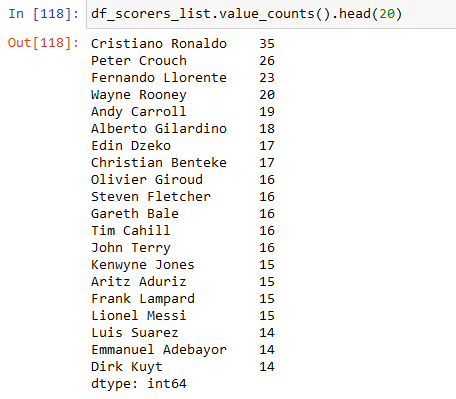
-Null was converted to 0 points or matches and points per match   
 were calculated for all seasons case and only 2008/2009 and   
 2015/2016 only case. Also, team api id was changed to team name:

- A bar graph was plotted for both cases, here is the all seasons   
 case (1) (most consistent teams across the period):

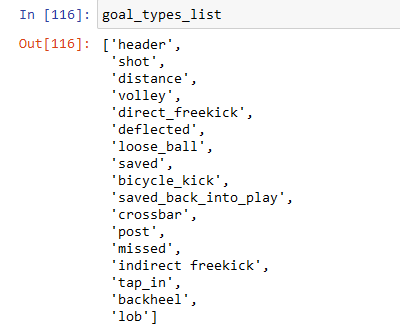
- case (2) (team that improved the most between 2008 and 2016)

1. Question (g) (Which player scored the most goals of a certain subtype (for example bicycle kicks) across the 8 seasons period?):

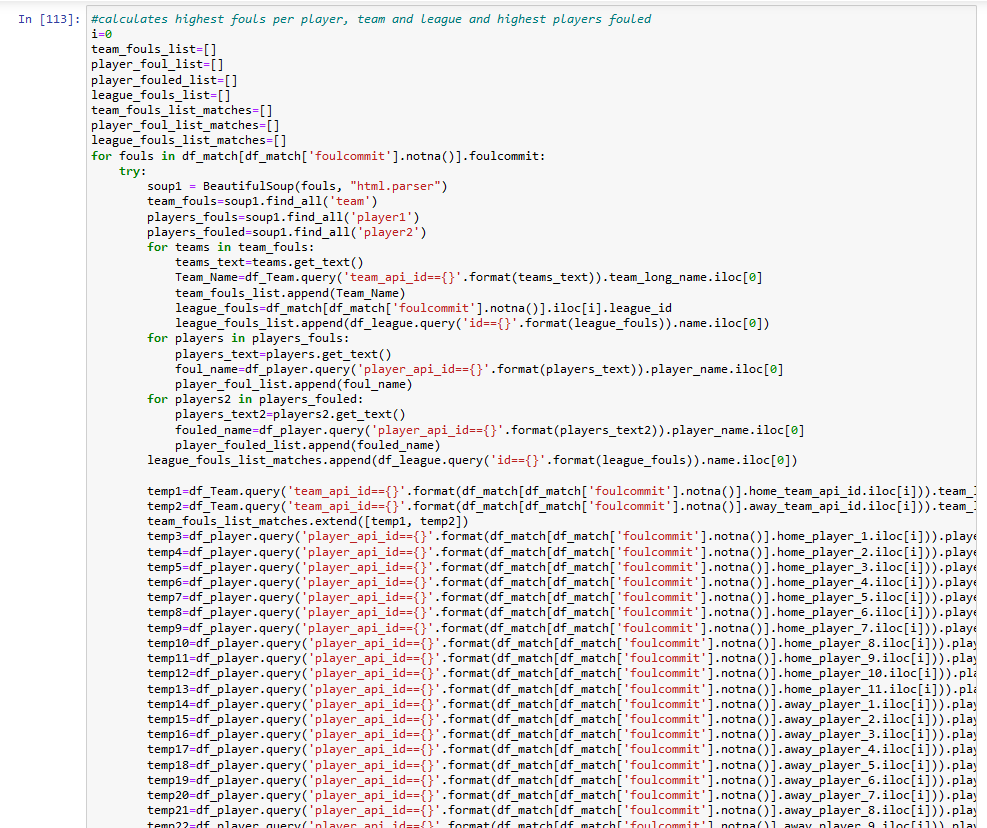
- The code below was used to calculate the frequency a player scored   
 a certain subtype of goals. Beautiful soup was used to extract the goal   
 subtype and scorer’s name. The subtype investigated was chosen   
 and the player’s name was appended to a list each time he scored   
 this subtype of goal:

 - Below is a list of the top 20 header scorers across all leagues and all   
 seasons:

-Here are all the goal subtypes available:



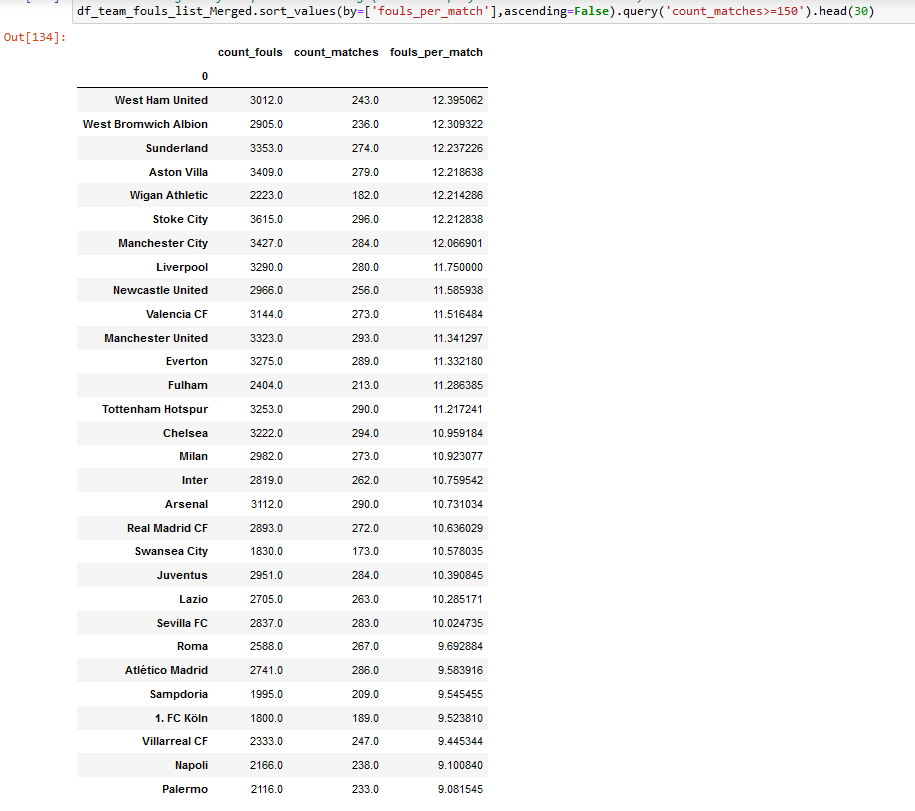
1. Question (h&I) (Which players, teams and leagues had the highest committed fouls per game?)   
   (Which players were fouled the most per game?):

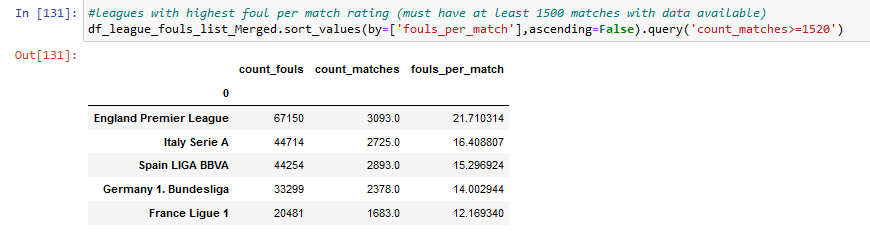
 -This code uses beautiful soup to extract the player who commited the   
 foul, his team and league. It also extracts the player fouled. For each   
 extracted api id, it is converted to a name (whether the player, team or   
 league) and appended to a list. The number of appearances of each   
 player is calculated using the starting formation in the match   
 dataframe. The number of matches a team played and number of   
 matches in each league are also calculated using the  
 match dataframe.

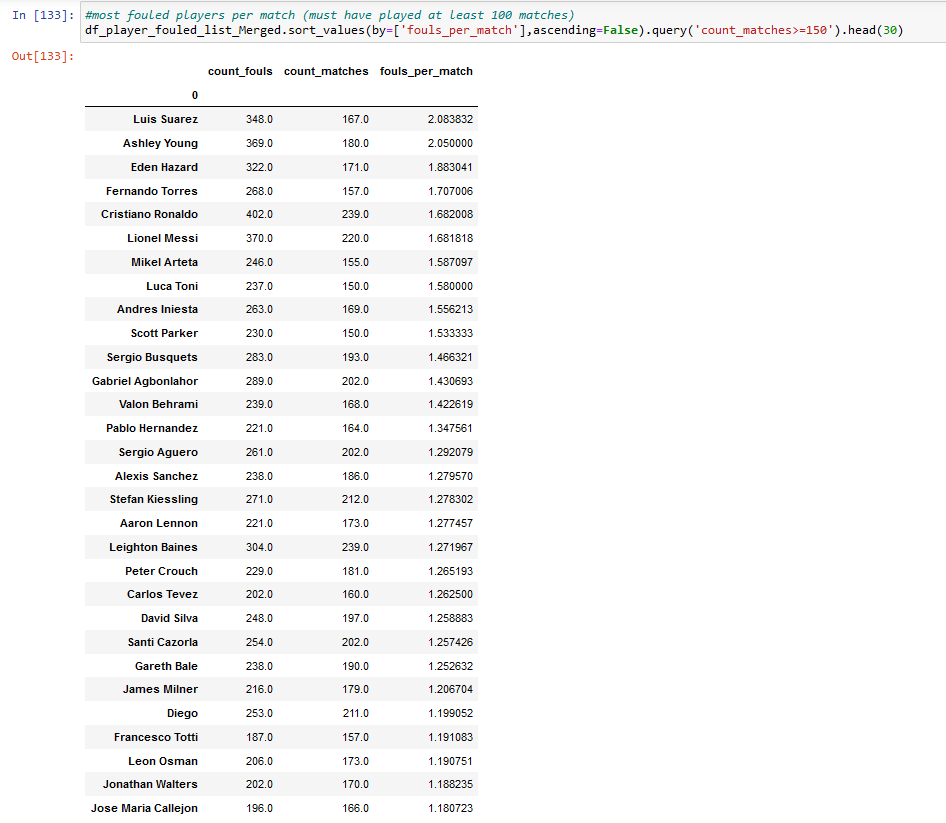
-Lists are converted to dataframes, value counts are calculated and   
 finally fouls per match are calculated for each case:

-The results are shown below:

-Case (1) players with most fouls committed (player must have at least   
 played 150 matches (about 50%):

-Case (2) teams with most fouls committed (team must have at least   
 played 150 matches (about 50%):

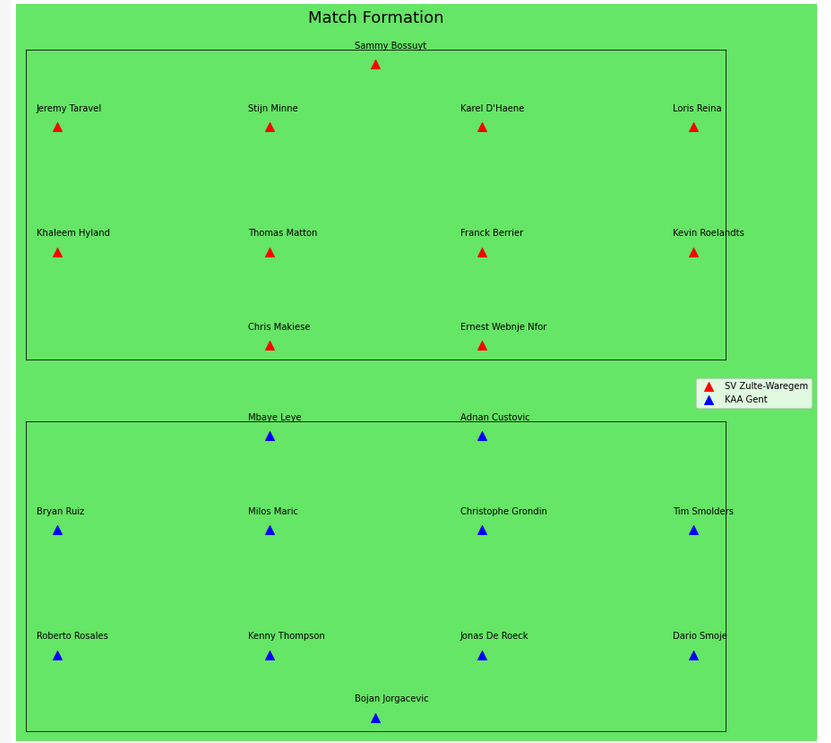
-Case (3) leagues with most fouls committed (data must be available   
 for at least 1520 matches) (about 50%):

-Case (4) players fouled the most (player must have at least played   
 150 matches (about 50%):

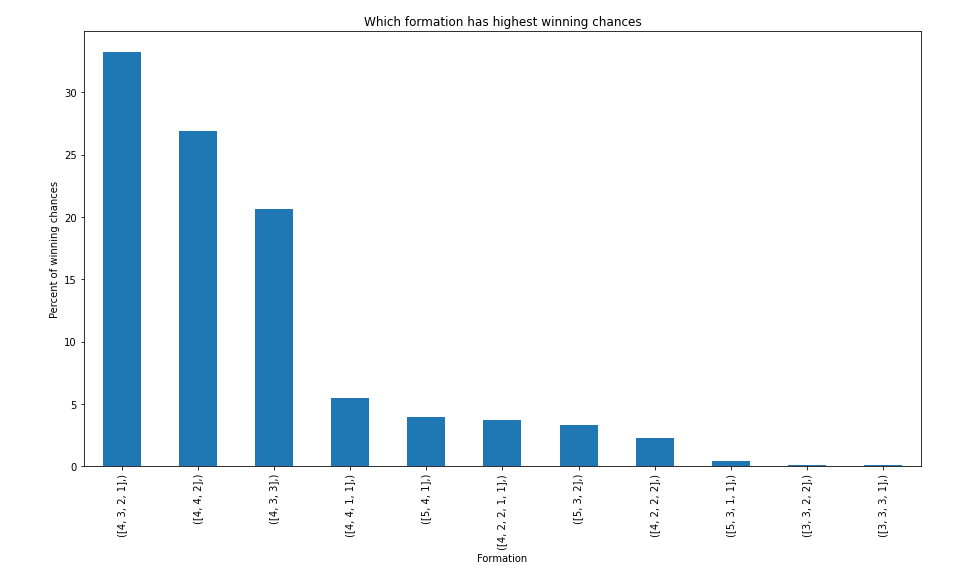
1. Conclusion:

Question (a) What is the formation of a certain team match   
in a certain match?

It was answered by plotting each player coordinates in a scatter diagram.



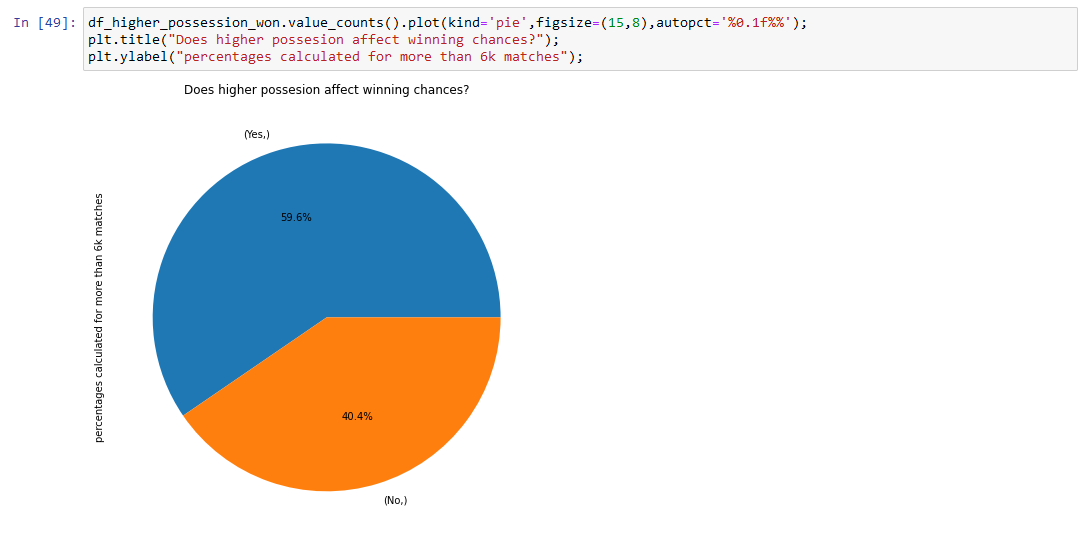
Question (b) Which formation has the highest winning chances?

Every match result was analyzed and the winner’s formation was recorded. A bar chart was plotted with the results.

The plot shows that 4-3-2-1 is the most successful formation in football matches with a success rate of 33.2%

Question (c) Does higher possession affect winning chances?

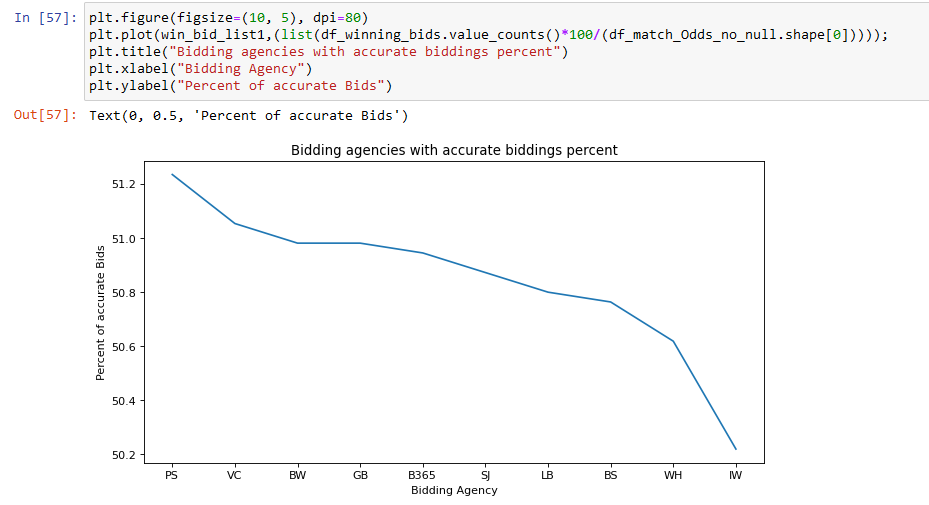
Every match result and possession numbers were analyzed and the winner was recorded if they were more possessive than their opponent. A pie chart was plotted with the results.



The plot shows that possession affects winning chances as 59.6% of teams with higher possession won.

Question (d) Which bidding agency has the most accurate biddings?

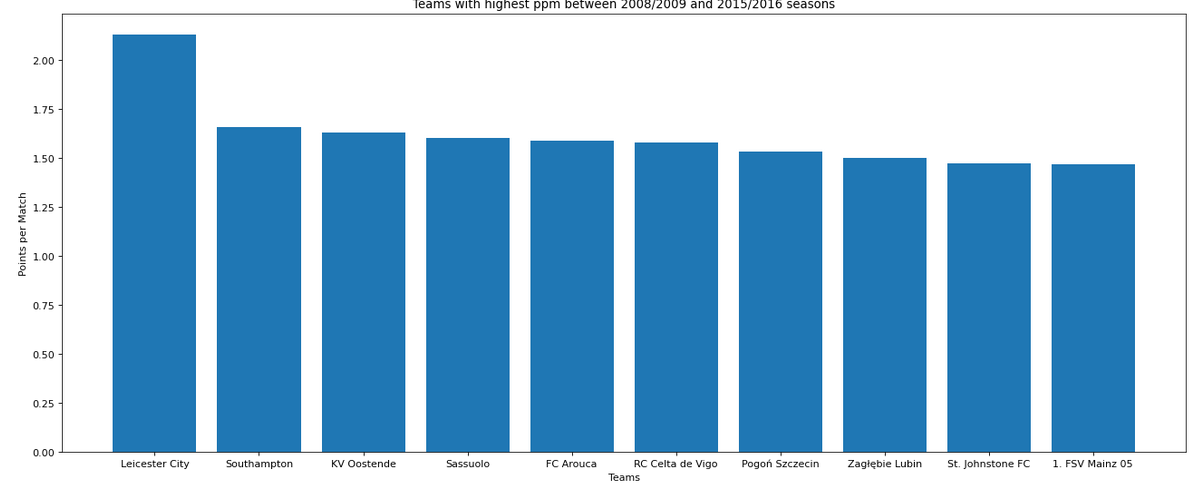
Every match result and bidding odds were analyzed and the bidding agency was recorded if the lowest bid matched the real match result. A line plot was used to graph the results.



The plot shows that PokerStars bidding agency has the most accurate bids between al the agencies by a 51.2% accuracy.

Question (e) Which team has improved the most between 2008/2009 season and 2015/2016 season?

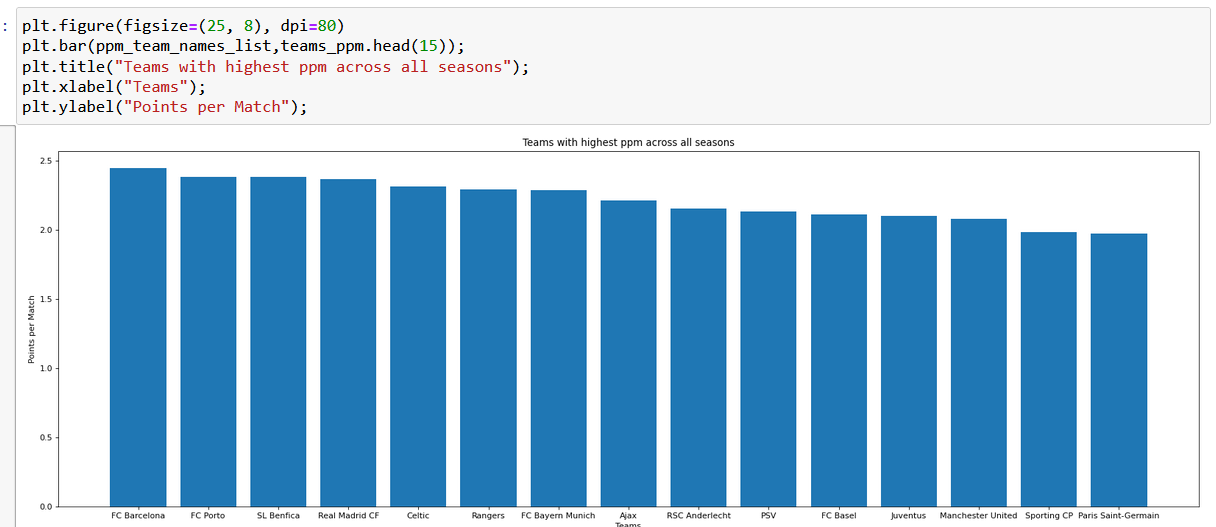
Every team points and number of matches were calculated each season. A bar plot illustrates the results found.



The plot shows that Leicester City has improved the most during the time period with an impressive difference of 2.13 points per match between 2008/2009 and 2015/2016 seasons. It makes sense since Leicester City was not in the premier league during the 2008/2009 season and it won the Title in the 2015/2016 season.

Question (f) Which was most consistent team across the 8 seasons period?

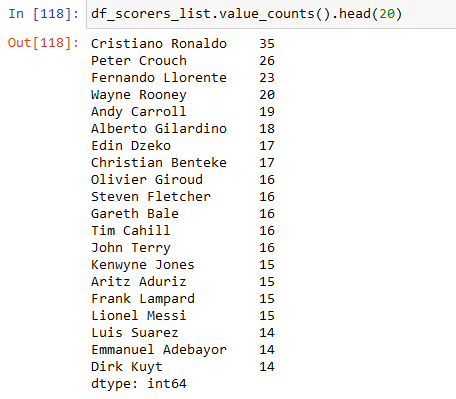
Every team points and number of matches were calculated each season. A bar plot was used to show the results found.



The plot shows that FC Barcelona was the most consistent team across the time period with an incredible average of 2.45 points per match between all seasons from 2008/2009 to 2015/2016. This time period is considered as the golden years of Barcelona under the management of the masterclass tactician Pep Guardiola.

Question (g) Which player scored the most goals of a certain subtype (for example bicycle kicks) across the 8 seasons period?

Every goal type scored and its scorer were extracted and recorded.



This List shows the top headers scorers during the time period with Cristiano Ronaldo scoring a whooping 35 headers during the 8 seasons period.

Question (h) Which players, teams and leagues had the

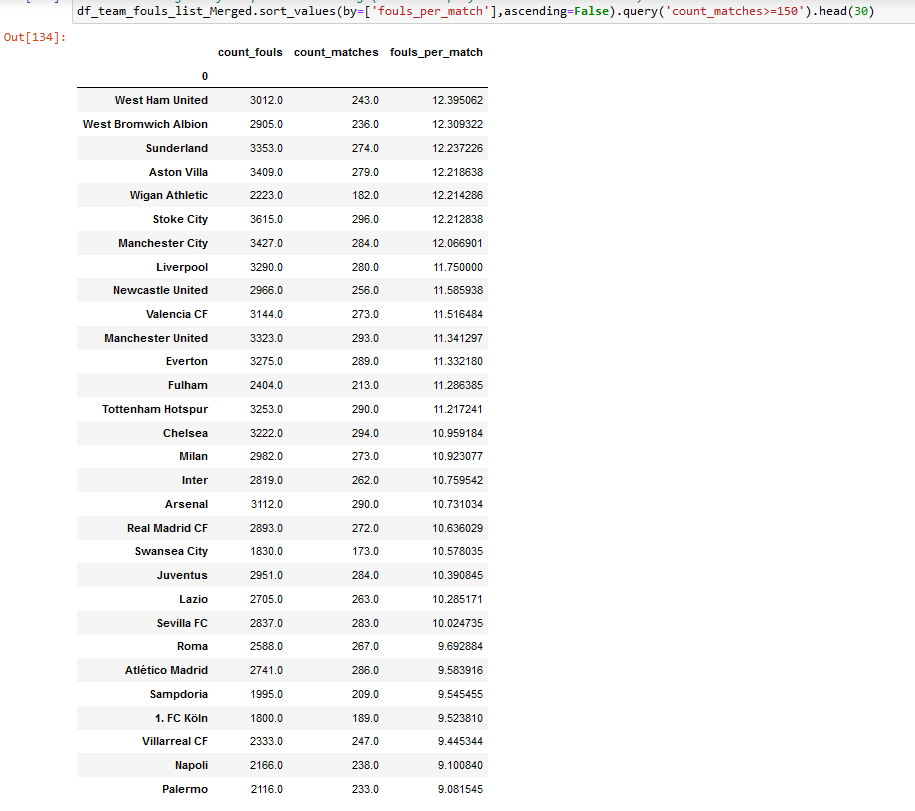
highest committed fouls per game?

Info about foul committed including league, team and player were extracted and recorded. The number of fouls was divided by the number of matches for each player, team or league.



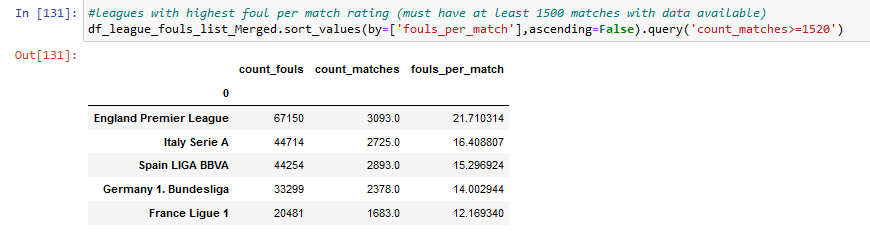
This List shows the players with the highest fouls committed per game. Marouane Fellaini tops the list with 2.95 fouls per match.

Next List shows the teams with the highest fouls committed per game.



It appears that many English teams are ranked higher on the list than other countries. West Ham United is towering the list with 12.4 fouls per game.

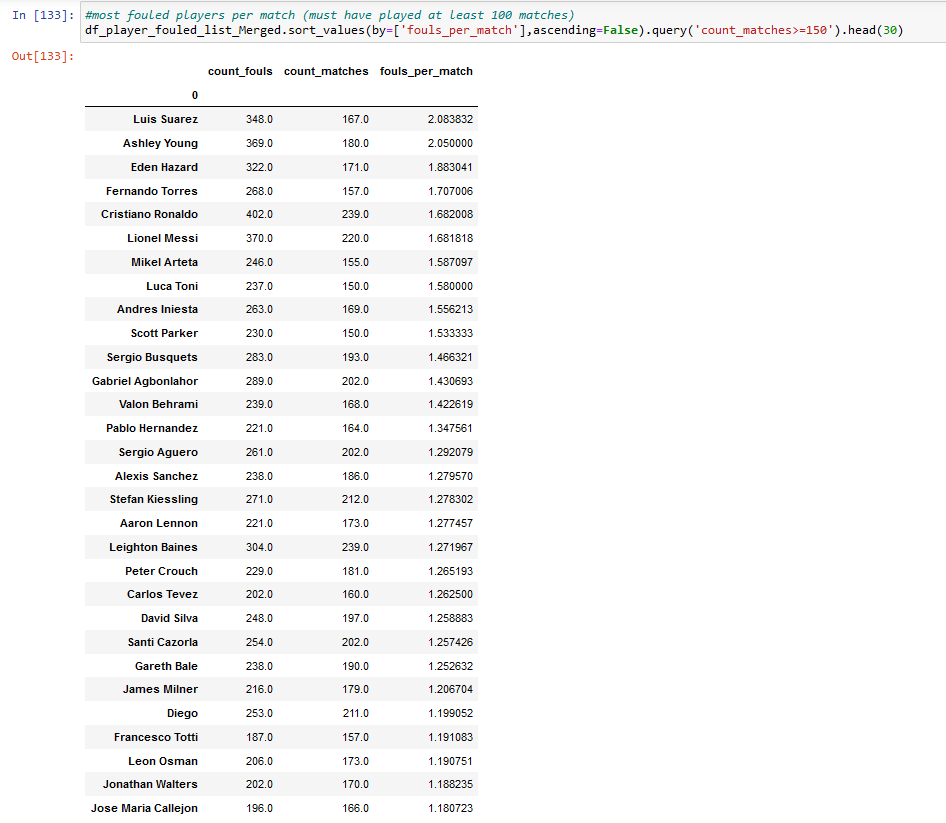
The Final list similarly calculates fouls committed per game but for leagues instead of teams.



The English Premier League has the highest number of fouls per game which is 21.7. This is consistent with the other results since both the top player and top team before play on the English Premier League.

Question (i) Which players were fouled the most per game?

The same method was used as before to calculate players fouled each match then results were recorded.



The list displays players who were fouled the most per game. It appears that most of them are famous players who play on attacking roles, which makes sense since attackers have a high chance to be fouled by defenders in a match. Luis Suarez tops the list with 2.08 fouls won per game.

Limitations

The Soccer Database is an exceptionally valuable treasure for any one who loves the beautiful game, but of course it has some drawbacks.

The main problem with the data set is that it has many missing data especially in leagues other than The Top Five (England, Spain, Italy, Germany and France). The English Premier League has the most complete data probably because it is considered the best football league in the world, so its data is easier to obtain. The dataset only includes the highest division in each league, which causes accuracy problems when calculating stats per match because some players and teams played relatively few matches as they were relegated or transferred during the time period. That is why when calculating stats like fouls per match, a minimum number of matches (about 50%) was used to filter out players or teams who did not make enough appearances in the dataset.

Another problem with this dataset is that it is considered outdated today. The last season 2015/2016 was finished six years ago. Football is everyday changing with new tactics, managers and players. The results collected from this dataset may not correctly apply with today’s football, for example winning formations or possession may not have the same results if the analysis was repeated for newer seasons. That is not to say the analysis is wrong by today’s standards, it is very likely to be close, but using a more updated database would give us more accurate results.