Football Data Analysis

As a huge football (or soccer) fan and at the time of the world's greatest sporting event, FIFA World Cup 2022, I decided to take football data analysis and visualization in R as my final project for the STA 518. Having to complete such a project during the last month really helps grasp and use all the concepts and techniques learned throughout the semester in a fun way.

For this project I wanted to look at many aspects of football matches from the stages or game weeks in a season, each of the goals scored in those games and the players who scored them to who the manager was and what was the formation the team adopted for the game. I knew looking at things in these minute details would take a lot of time and I could not finish the project on time. Although time constraint was one of the challenges I faced, I still wanted to start and work on this project even after the semester to see what I can achieve.

Data

Another challenge for me would be to get the data. Although we can find a lot of football data in Kaggle or other data hosting platforms, I needed data that were very detailed to each aspect of the game. I was able to find some and the ones I did not find, I scraped the web using Python. After a lot of searches, I found a ".sqlite" file on Kaggle that had data sets on the matches and players with acceptable details. The matches dataset includes information on each game played from the season 2008/2009 to season 2015/2016 for all the major football leagues in Europe. I decided to limit my project on those seasons as the data was readily available and there were only a few things that I needed to scrape off the web.

Here is a table showing the data I used and the source of those data:

database.sqlite	https://www.kaggle.com/datasets/hugomathien/soccer
Country	Table in the database
League	Table in the database
Matches	Table in the database
Players	Table in the database
Teams	Table in the database
EPL Manager	Scraped Wikipedia
EPL Referee	Scraped (https://www.bdfutbol.com/)
LaLiga Manager	Scraped Wikipedia
LaLiga Referee	Scraped (https://www.bdfutbol.com/)

The database.sqlite file contained many tables. Matches was the most important table for me which gave me a detailed match details for all the games played in 11 leagues in their respective country. Country table gave the information on country id, League table gave information on the league id, Players table gave information on different attributes of players and their id and Teams table gave information on the team id and their attributes. All these details were important to my analysis and needed to be merged with the Matches table.

Information on the managers of the teams playing those games and match referees was not available in the file and I could not find a dataset online meeting my criteria. Therefore, I decided to scrape the web for the missing information. I found all the manager's data on Wikipedia and used python's beautifulsoup library to extract all the information. I needed the season in which the manager was in office for each team and the name of team, which I planned on using to merge the manager's data with the matches data. The most difficult was to find the data on who was the referee on all those games. With a lot of hard work, I got to the BDFootball's website, and it had all the necessary information I needed. I scraped the name of those referees, home team and away team for the match and the date of the match. The date was an important detail as the same teams will have played against one another in multiple occasions under different referees.

There are not much missing data in the data apart from the player data. Unsubstantial amount of these player data is missing which I replaced with some fake data to calculate age. There are a lot of features I have used throughout this project, and I have listed all of them in the data dictionary below:

Variable	Description	Type	Class
country_id	ID for the Country (Found in Country Table)	Integer	Numeric
league_id	ID for the League (Found in League Table)	Integer	Numeric
team_api_id	ID for the Teams (Found in Teams Table)	Integer	Numeric
player_api_id	ID for the Players (Found in Players Table)	Integer	Numeric
season	Season of the game	Character	Character
stage	Stage of the game	Integer	Numeric
date	Date the game was played	Double	Date
home_team_goal	Number of Goals Home Team Scored	Integer	Numeric
away_team_goal	Number of Goals Away Team Scored	Integer	Numeric
on_target_shot_home_team	Number of Shots on Target for Home Team	Integer	Numeric
on_target_shot_away_team	Number of Shots on Target for Away Team	Integer	Numeric
off_target_shot_home_team	Number of Shots off Target for Home Team	Integer	Numeric
off_target_shot_away_team	Number of Shots off Target for Away Team	Integer	Numeric
foul_home_team	Number of Fouls by Home Team	Integer	Numeric
foul_away_team	Number of Fouls by Away Team	Integer	Numeric
yellow_card_home_team	Number of Yellow Cards Home Team	Integer	Numeric
yellow_card_away_team	Number of Yellow Cards Away Team	Integer	Numeric
red_card_home_team	Number of Red Cards Home Team	Integer	Numeric
red_card_away_team	Number of Red Cards Away Team	Integer	Numeric
crosses_home_team	Number of Cross Home Team	Integer	Numeric
crosses_away_team	Number of Cross Away Team	Integer	Numeric
corner_home_team	Number of Corner Home Team	Integer	Numeric
corner_away_team	Number of Corner Away Team	Integer	Numeric
possession_home_team	Possession for Home Team	Integer	Numeric
possession_away_team	Possession for Away Team	Integer	Numeric
H_Age	Home Team Players Average Age	Double	Numeric
A_Age	Away Team Players Average Age	Double	Numeric
HomeTeam	Home Team	Character	Character
AwayTeam	Away Team	Character	Character
home_player_1 to home_player_11	Names of 11 Starting Home Players	Character	Character
away_player_1 to away_player_11	Names of 11 Starting Away Players	Character	Character
HomeManager	Manager of Home Team	Character	Character
AwayManager	Manager of Away Team	Character	Character
Ref	Name of Referee who officiated	Character	Character
FullTimeResult	Result in Full Time	Character	Character
HomePoints	Total Points Home Team Scored	Integer	Numeric
AwayPoints	Total Points Away Team Scored	Integer	Numeric
HomePosition	Position of Home Team on Table	Integer	Numeric
AwayPosition	Position of Away Team on Table	Integer	Numeric

HomeGoalsFor	Total Home Goals Scored by Home Team	Integer	Numeric
AwayGoalsFor	Total Away Goals Scored by Away Team	Integer	Numeric
HomeGoalsAgainst	Total Home Goals Scored by Away Team	Integer	Numeric
AwayGoalsAgainst	Total Away Goals Scored by Home Team	Integer	Numeric

Methods

After getting all the necessary data I started working on cleaning them first. The matches table had a lot of missing values and most of the information in the table were IDs instead of names. For example, there were IDs instead of names for Country, League, all the players, Teams and many more. I started by reading in the data in R and cleaning it. I had decided to use only two leagues in Europe, English Premier League (EPL) and LIGA BBVA (LaLiga) as they are the top two leagues in the continent. I merged both these tables together to get all the matches from these leagues from 2009/10 – 2015/16 seasons.

The next task for me was to replace all the ids with their respective name and I started with replacing the country id and league id in the matches table with the name of the country and league the match was played on. After this, I replaced the name of home team id and away team id with the actual name of the team. This gave me an idea of who the teams are in the dataset. To replace these ids, I went through each row of the dataset (each match played) and matched the id with the names in other datasets.

There were a lot of missing information in the dataset. Luckily, most of them were in the columns we would not be looking at in our analysis like the sport betting odds for each match. A lot of IDs were missing from the player's columns but not for the two leagues we were looking at. And for the player's IDs that were missing, I gave a fake id to the player and updated that fake id in the Players dataset too. This way we would have all the data regarding the players to get to the average age of the team that played the match. I then replaced all the player's ids in the matches data with their respective names by performing a left join and merging the matches dataset and the players dataset. Although I did not look at the average age of the players and the players themselves in this phase of the project, I am planning on working with them to get new information in the future.

Next, I needed to merge the matches dataset with the managers and referees' datasets which I had scraped off the web. While scraping I kept in mind the columns, I needed to perform the merge but quite unluckily the names on the matches datasets for the teams that played in those matches were different. For example, in the EPL Referee's dataset, a team's name was "Tottenham" while the name on the matches dataset for the same team was "Tottenham Hotspur" which was the official name. and this was the case with LaLiga Referee's dataset, and LaLiga manager's dataset. I then decided to rename each of the teams on other datasets to match the ones in the matches dataset for which I renamed the teams individually using the stringR

package. I also needed the dates the games were played to match on formatting on all the datasets so that I could use the name of home team, away team and dates to merge the matches table with the referee's datasets which I did using the lubridate package. I then merged the matches datasets to the manager's datasets giving the home and away team's manager information on each row and the referee who officiated each of the games.

I also created a column called "FullTimeResult" which included information on who won the game, home team or the away team, or was it a draw by comparing the goals that the home team and away team scored. With all the columns I had on my matches dataset, I started looking at some summary statistics of groups of data by grouping the data based on seasons and leagues and others. For example, I looked at how many goals were scored in each season and grouped them based on the two leagues I was looking at. These, however, were statistics for all the seasons in the dataset. There was some other interesting information hidden in the individual seasons and leagues.

So, my next task was to create a function that would take two parameters, season, and league, and would ultimately filter the matches dataset to those criteria and return two datasets, one that gave the season's overall picture and the other which had information on all the games played in the give season for the given league. The dataset with the season's overall picture would include information on teams and their points, how many goals they scored, how many did they concede, what the goal difference was and their ranks at the end of the season. However, the other dataset the function returned would include all these information on a game-by-game basis. So, it would include the points and the goals the team scored and conceded at the end of each game.

And since we now have a season specific data, we can look at more things like how many games in the season were won by home teams, how many by away teams and how many ended in a draw. We can also group referees based on the number of games they officiated and the number of cards they gave out (red and yellow) throughout the season. Using the seasonal data, I also looked at how many goals each team scored throughout the season and split them based on home goals and away goals. To visualize this data with a stacked bar chart, I had to pivot it in a longer format. And finally, to get the information on individual teams, not divided as home and away teams, I created another function called "stagePointsGoals" which would take a variable number of teams, one or more, and would return a dataset with each of these teams individually listed for each game they played in a season with updated information as they progressed through the season. This would be a basis of one of the most important visualizations, the line chart, which would show how the teams that we selected as an argument in the "stagePointsGoals" function performed in the season and league we selected in the first function.

The dashboard I created in another .rmd file with knit parameters, so we need to knit this file using the "knit with parameters" option, which would prompt the user to provide the season, league and team/s as input and then knit the document based on those parameters. This is a

much better approach because we get the option of choosing our parameters in a designated spot rather than having to find the function in the file and update the parameters.

Results

After cleaning and performing analysis on the data, I was able to figure out a lot of things. I performed exploratory data analysis and looked at various descriptive statistics like mean, median, mode, minimum and maximum value, quartiles and many more. The minimum and maximum value played an important role in the analysis. For example, seeing more than 10 in the goals column would be a something that needed further examination. I also looked various statistical summaries of total and grouped data. This gave valuable insights to the overall seasons that we looked at and while we looked at the data for individual seasons. Let's look at some of the tables and graphs we prepared that helps us to get valuable insights.

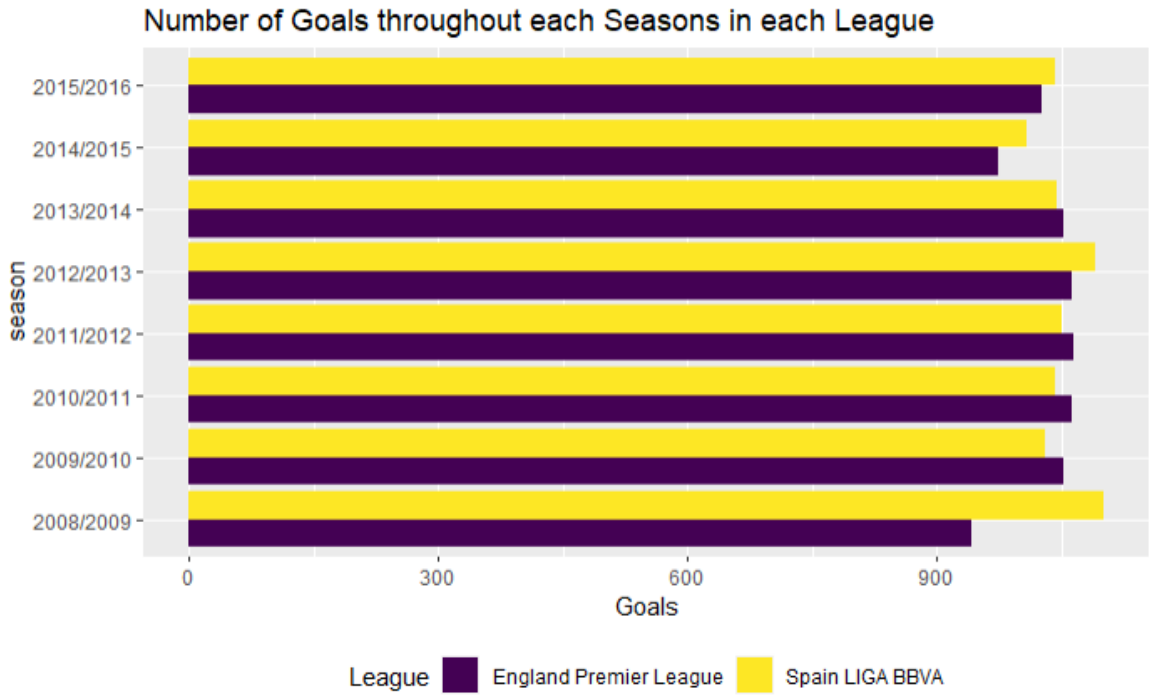

Figure 1 Number of goals throughout each season in each league

Figure 1 shows a grouped bar chart that shows how many goals were scored in each of the seasons in both EPL and LaLiga. Here we can see that most of the seasons have similar number of goals and only the season 2008/2009 having a substantial difference. EPL being the most competitive football league in the world may be one of the reasons for this difference where scoring goals were thought more difficult. But the later seasons show that there is no difference between scoring in EPL and LaLiga.

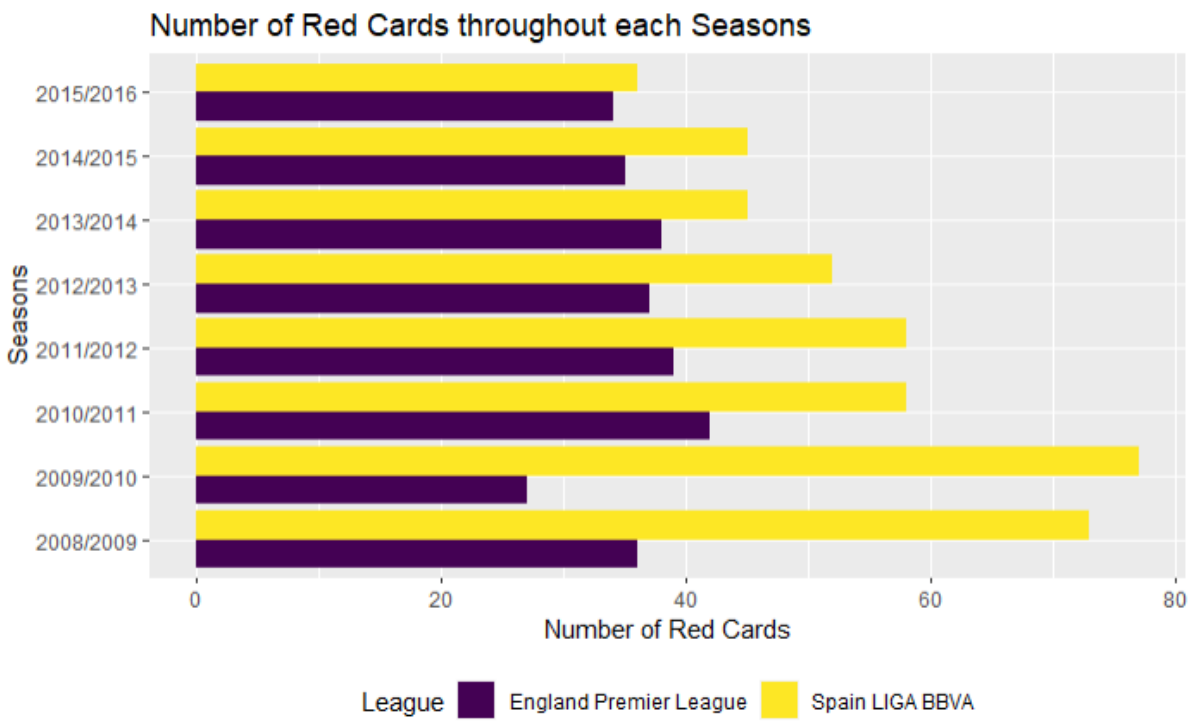

Figure 2 Number of red cards throughout each season in each league

Figure 2 shows how the two leagues differ in number of red cards issued in each season of the leagues. We see an important trend here, the number of red cards in EPL have fairly remained constant while the number of red cards in LaLiga has decreased every season. In the past, the number of red cards in the LaLiga games were more than double what were in the EPL.

Rank <dbl>	Teams <chr>	Played <dbl>	GF <dbl>	GA <dbl>	GD <dbl>	Points <dbl>
1	FC Barcelona	38	95	21	74	96
2	Real Madrid CF	38	102	33	69	92
3	Valencia CF	38	64	44	20	71
4	Villarreal CF	38	54	44	10	62
5	Atlético Madrid	38	62	53	9	58
6	Athletic Club de Bilbao	38	59	55	4	58
7	Sevilla FC	38	62	61	1	58
8	RCD Espanyol	38	46	55	-9	49
9	CA Osasuna	38	45	46	-1	47
10	Real Sporting de Gijón	38	35	42	-7	47

Figure 3 Table showing the season level statistic of the teams

Figure 3 is a partial screenshot of the table that shows the season level description of how each team in the given league performed in the given season. It includes information about the team, it's rank, number of games played, goals scored, goals conceded and points, all at the end

of the season. This is an important statistic in any game where we would want to know where each team stands. All the figures now, including 3, are based on season 2010/2011 and the league is LIGA BBVA.

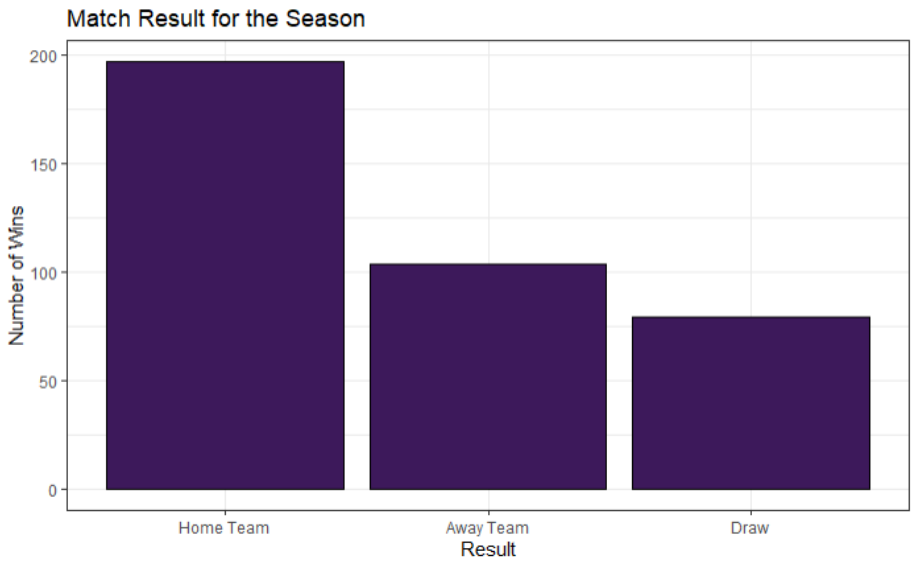

Figure 4 Result for the season

Figure 4 is a bar chart that shows the number of home and away wins and the number of draws. Home teams have a competitive advantage as per our result. This is true because home teams are more comfortable and have the large number of home fans who push them to perform better.

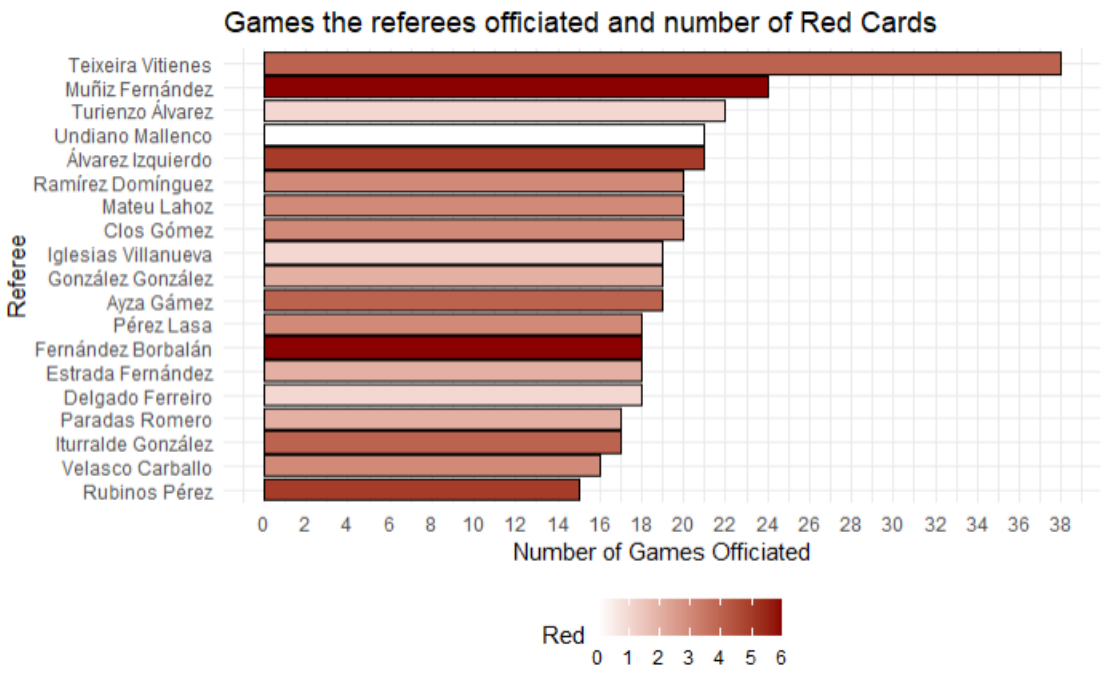

Figure 5 Number of games for each referee and number of red cards issued

Figure 5 is yet another bar chart that shows the names of the referees and the numbers of games each officiated. Teixeira Vitienes seems to have officiated in most of the games, 38, and the most red cards were given by Fernández Borbalán, 4. I also created another bar chart that shows how many yellow cards these referees gave in the season.

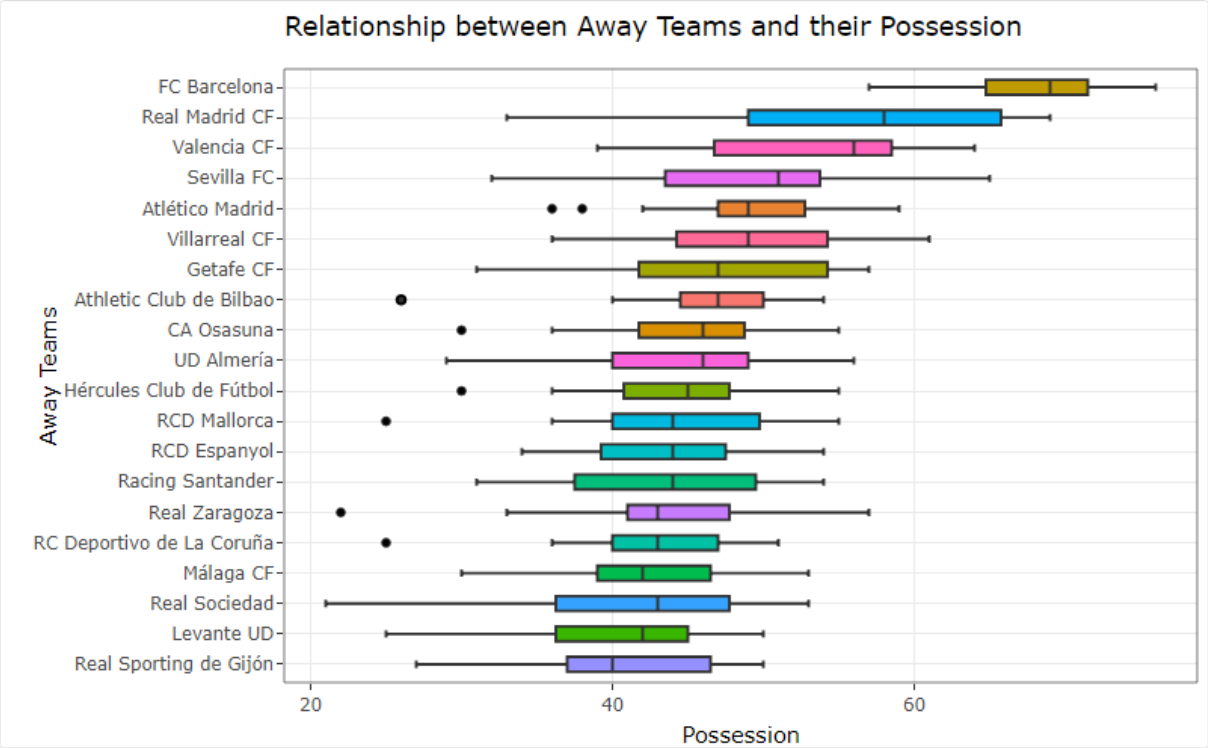

Figure 6 Possession of Away Teams

Figure 6 is box plot that shows the distribution of possession each teams had when they played away from their home. I also have another box plot that shows the distribution of possession each teams had when they played in their home. In both plots we can see FC Barcelona on top. This is the period when the manager of FC Barcelona was Pep Guardiola, who is famous for his “Tika Taka” style of play. This resulted Barcelona in complete control of the ball always.

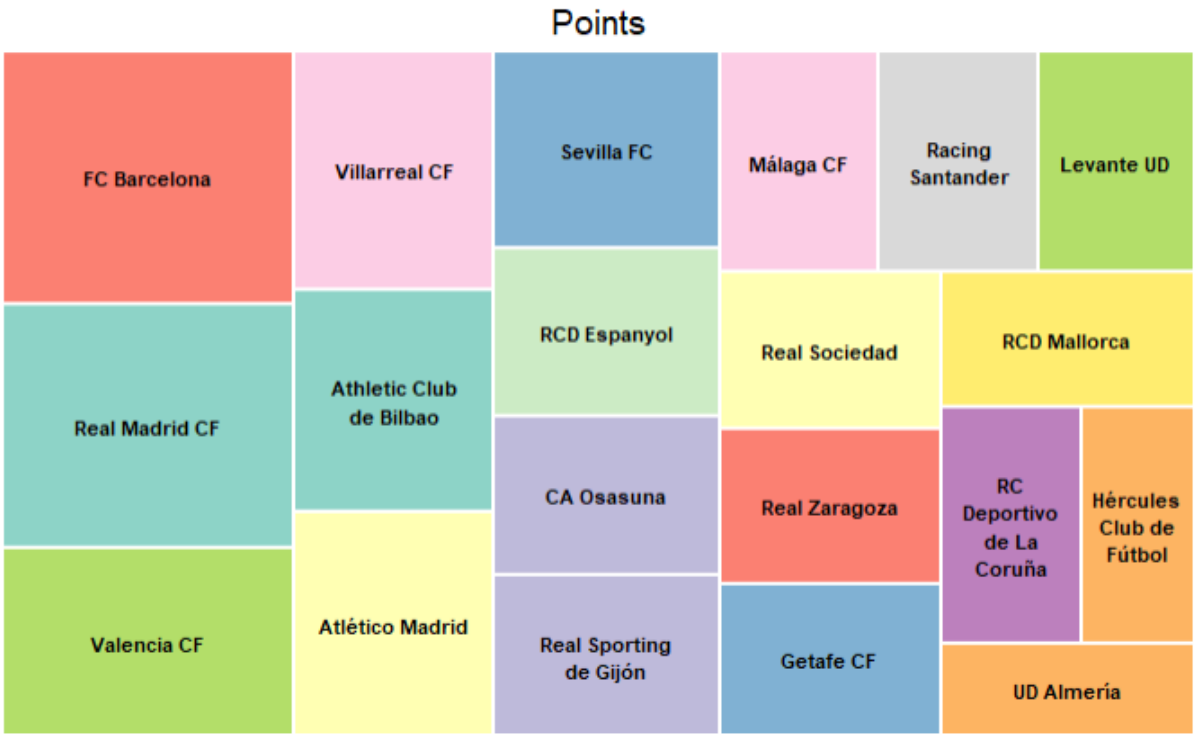

Figure 7 Points distribution of Teams

Figure 7 is a Tree graph which shows the proportion of points each team won. From the table we can clearly see that FC Barcelona was the top team as it has the largest area of box which means it scored the highest point. On the opposite spectrum is UD Almeria, who was at the bottom of the league and was relegated.

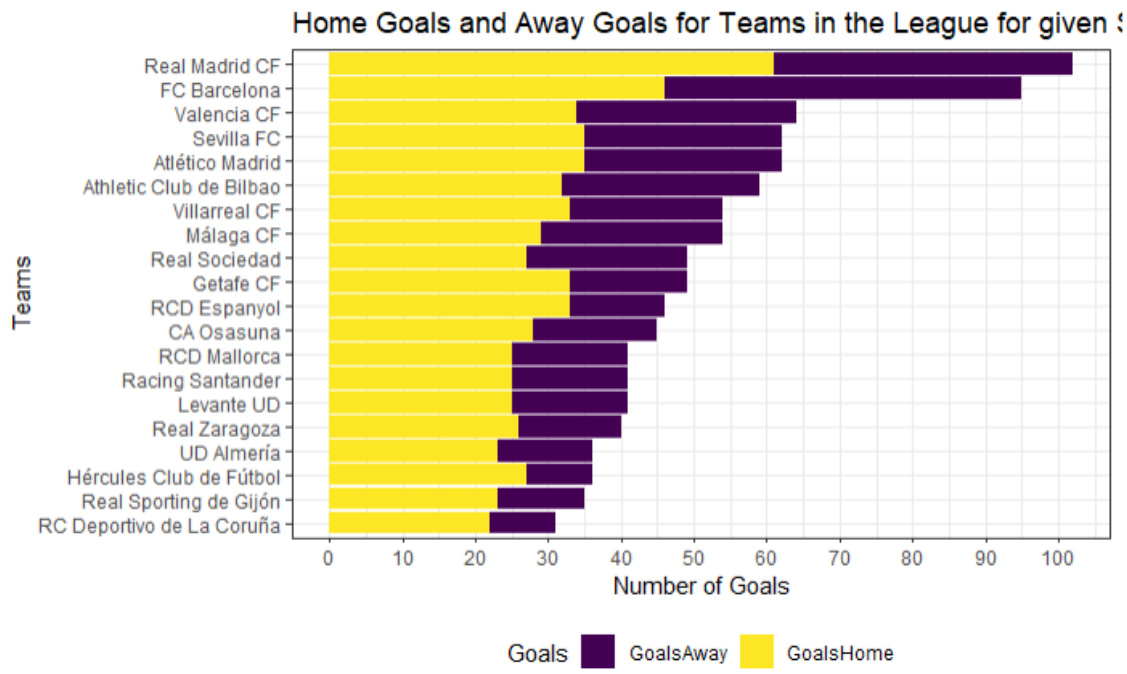

Figure 8 Home and Away goals for teams in the league

Figure 8 is the stacked bar chart which shows how many goals each team scored in home, while away and in total. The purple bar represents away goals and yellow bar represents home goals. FC Barcelona, who won the league title that season, seems to be beaten by their rivals Real Madrid who were the runners-up in the league in terms of number of goals scored.

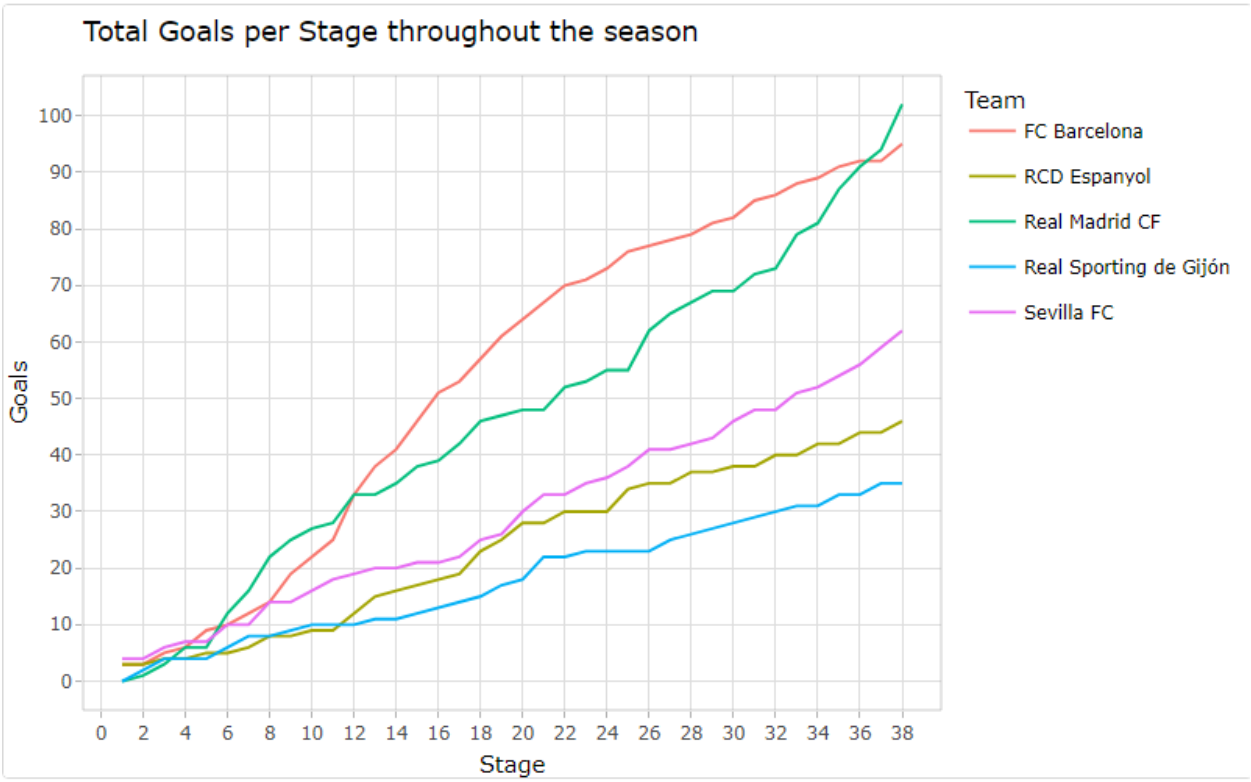

Figure 9 Total goals per stage in the season

Figure 9 is a line chart that shows how the team performs on the goals front in each of the 38 game weeks or stages. I had selected a list of the shown five teams in the "stagePointsGoals" function. If we select any other teams in that function, we can change this graph and show information for those teams. As the teams started the season Real Madrid had more goals than others and were surpassed by FC Barcelona from 12th stage. At the end of the season, Real Madrid, again seems to have secured the win in terms of goals. These plots are very good at visualizing time series data like this one. I also have another line chart that shows how total points change throughout the season.

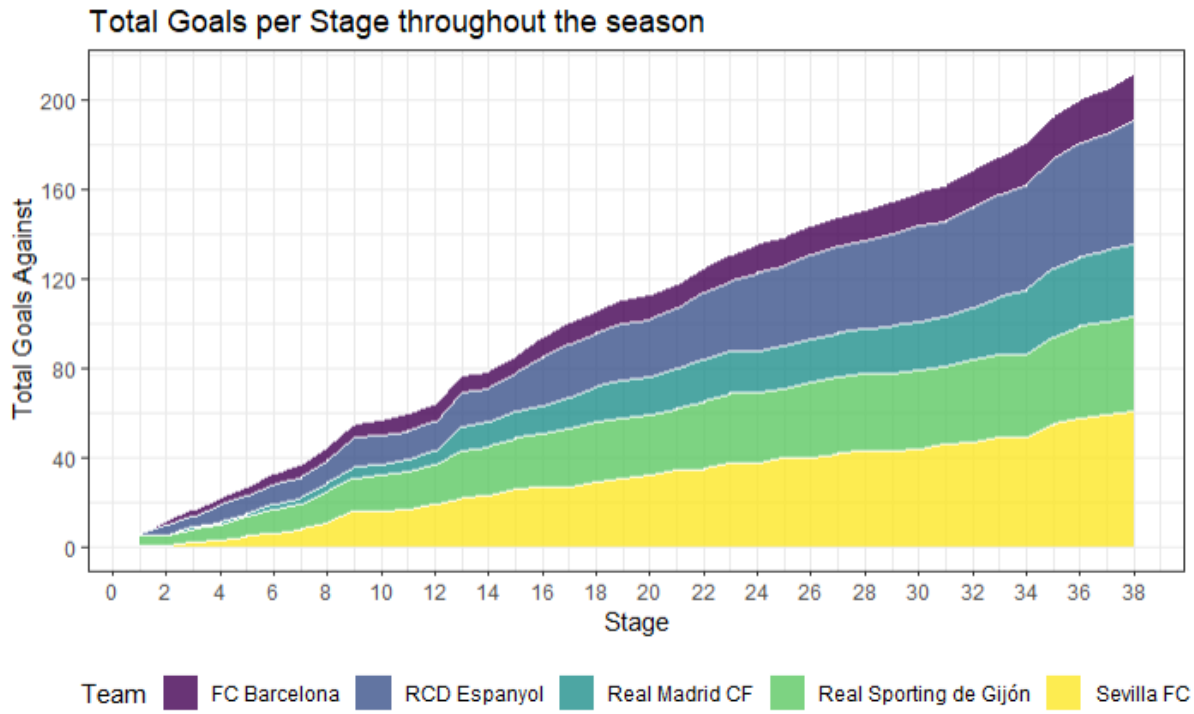

Figure 10 Total Goals per Stage throughout the season

Figure 10 is a stacked area chart that shows how the teams progressed through the season in terms of the number of goals they conceded. From the graph we can see that FC Barcelona are the one team that conceded very few goals when compared to others and this graph is also based on the teams, we passed in our second function. If we want data for more teams, we just need to pass those teams in the function.

Conclusion

Working on a project like this one, where I was doing everything, I liked, from football to writing codes to visualizing the data was very fun and interesting. I am happy with how the project turned out and about the fact that I met all the requirements for the project. All the graphs and tables in the project give some interesting insights to how the world of football is and how it might be changing right underneath our nose and how we are still unaware of it. There are still a lot of analyses we can do under this project from analyzing player statistics in each of the game and ultimately determining winning characteristics of a team. These sorts of analysis are increasing in the field of sports and sport analytics is a field. So, this project has certainly helped me in increasing my skills in R and in analytics and am planning on continuing to work in this project and making it better in many ways. Next I want to focus on creating a dashboard for getting all these insights using Shiny package.