Project2: Soccer database analysis

Created By	Fares Lassoued
Last Edited	@Apr 14, 2020 9:48 AM
Property	Report
Tags	Data Analyst Nanodegree

1.Introduction:

For this project, I decided to use the soccer database for European teams provided on Kaggle which contains data for soccer matches, players, and teams from several European countries from 2008 to 2016. I want to explore and discover the important factors for Elite players and most winning teams, as well as players and team progressions over the years.

Note: All the code for this project will be provided seperately

Motivation:

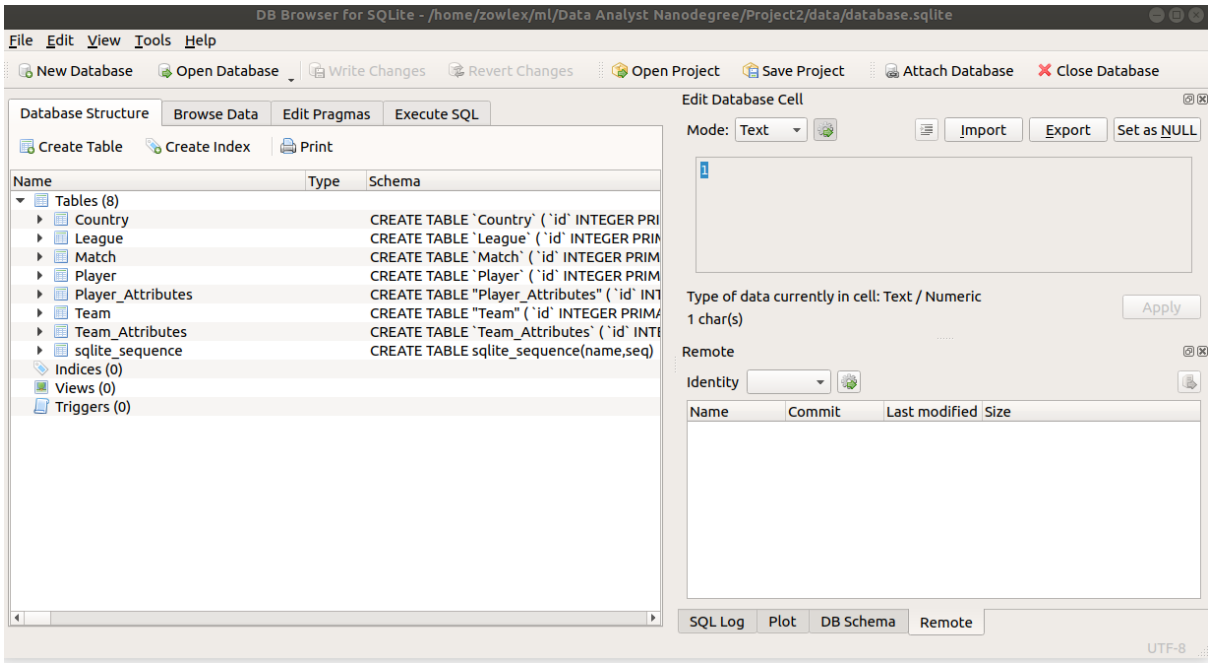
- I wanted to experience what it is like to work with data from its original source (SQLite DB in this case).

- I chose to work with this dataset because I am interested in soccer and this is the best chance to practice what I learned.

2. Data Wrangling

2.1 Gathering data

DB was downloaded from this link and examined using DB Browser.



Convert Tables to CSV files

I used a custom method to go through all tables in db and convert them to csv files using pandas and sqlite3 modules

Note: Code is in notebook file

2.2 Assess Data

DFs summary after converting tables to csv files

Aa df	shape
Country	(11,2)
League	(11,3)
Match	(25979, 115)

Aa df	shape
Player	(11060,7)
Player_Attributes	(183978, 42)
Team_Attributes	(1458, 25)

So far we know that the data covers:

- 11 Countries and 11 leagues
- 25979 played matches and stats
- 11060 player
- Player attributes are taken multiple times throughout years
- 1458 Teams

We'll end up working with team_df, team_att_df, and player_df, player_att_df, and league_df, match_df. So I thought of merging between each pair and work on 3 dataframes in total only.

Aa df	shape
team_merged_df	(1458, 29)
player_merged_df	(183978, 48)
match_merged_df	(25979, 118)

2.3 Clean Data

Steps taken for each of 3 dataframes:

▼ Drop unnecessary columns

▼ Check for missing values and drop

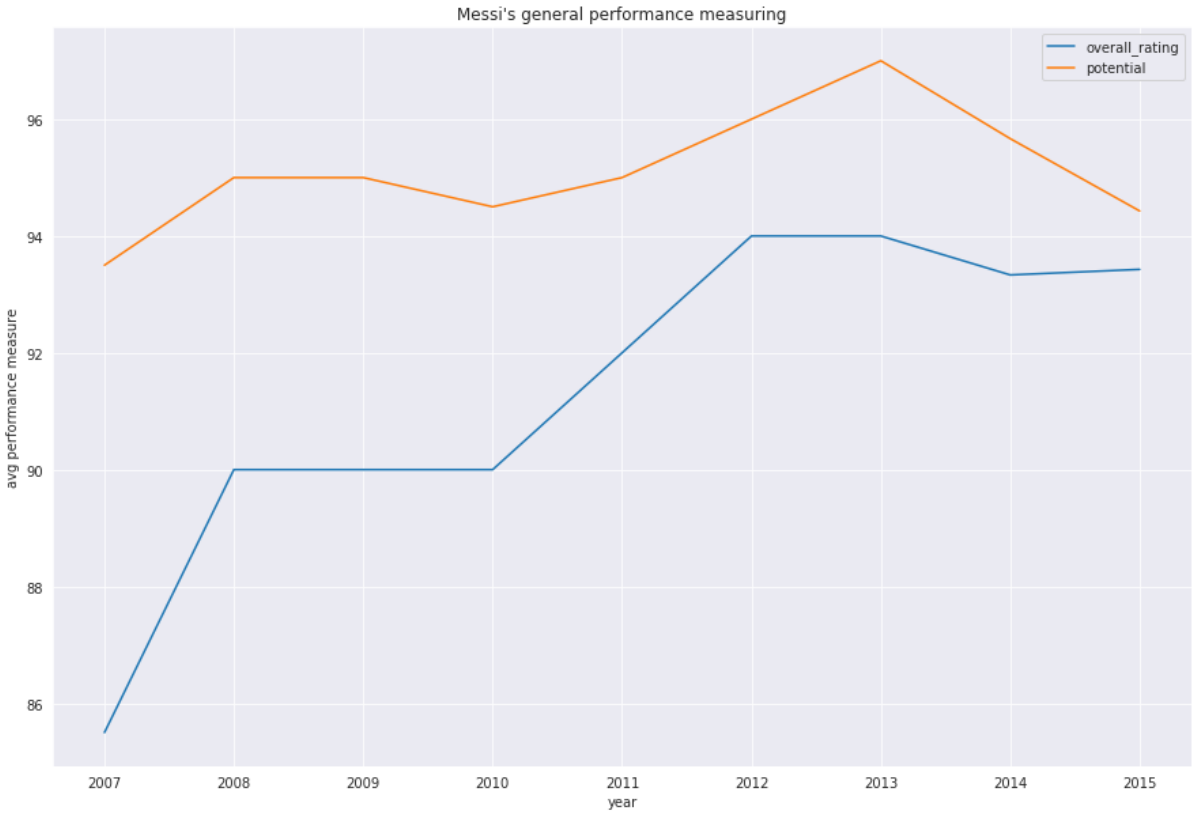
▼ Check for duplicates and dedup

▼ Check and convert datatypes (convert dates from object to datetime)

3. EDA

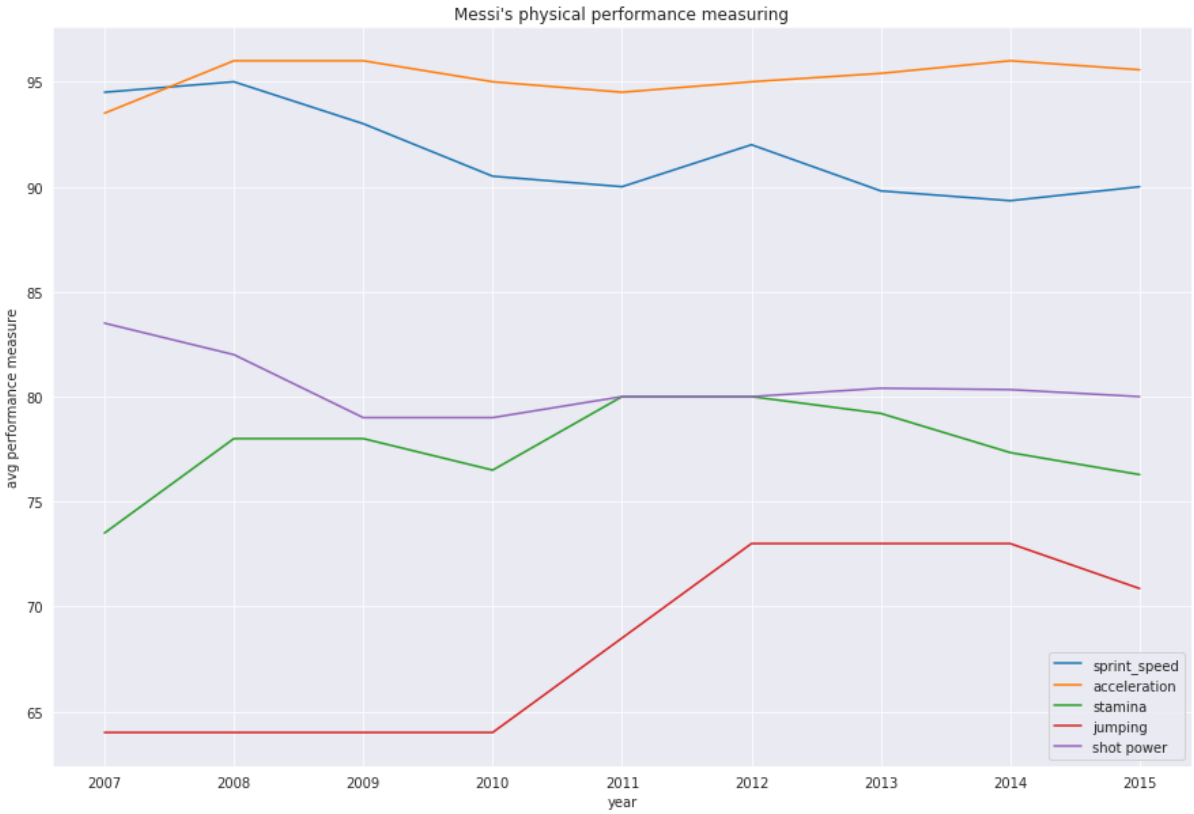
Question1: How is Lionel Messi's performance over the years?

First we compare his overall performance metric between 2007 and 2015



As expected the average overall_rating per year of Lionel Messi has improved by 9.2%

Second We compare his physical performance over the years (like jumping, sprinting,...)



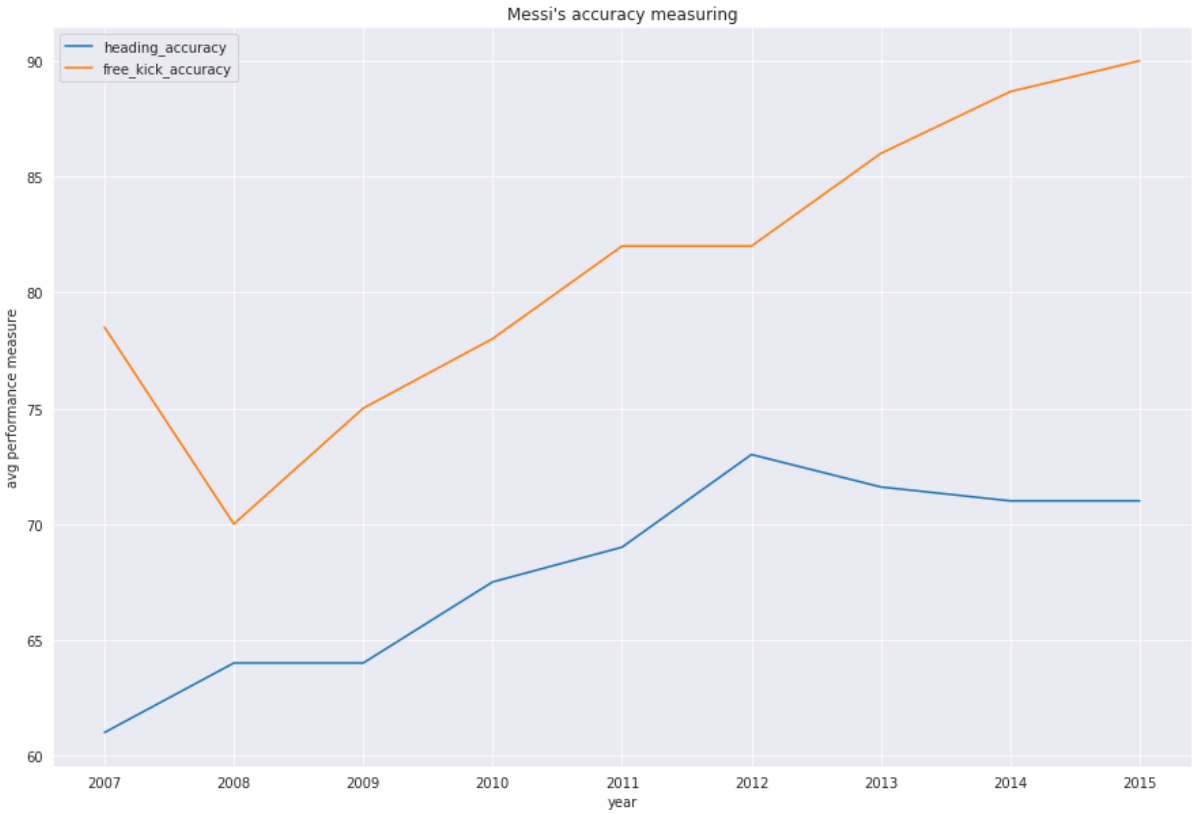
Acceleration and sprint_speed: Since Messi is relatively short and fit, his acceleration is an important key factor which depends on small periods of top speed, unlike sprint speed which decreases with aging.

Stamina: This metric is related to the physical condition and it has seen the most peek between 2011-2012.

Jumping: Jumping has tremendously improved since 2010 and starts decreasing since 2014

Shot power: Shot power has decreased and was stable from 2009 since his playstyle doesn't depend on the shooting

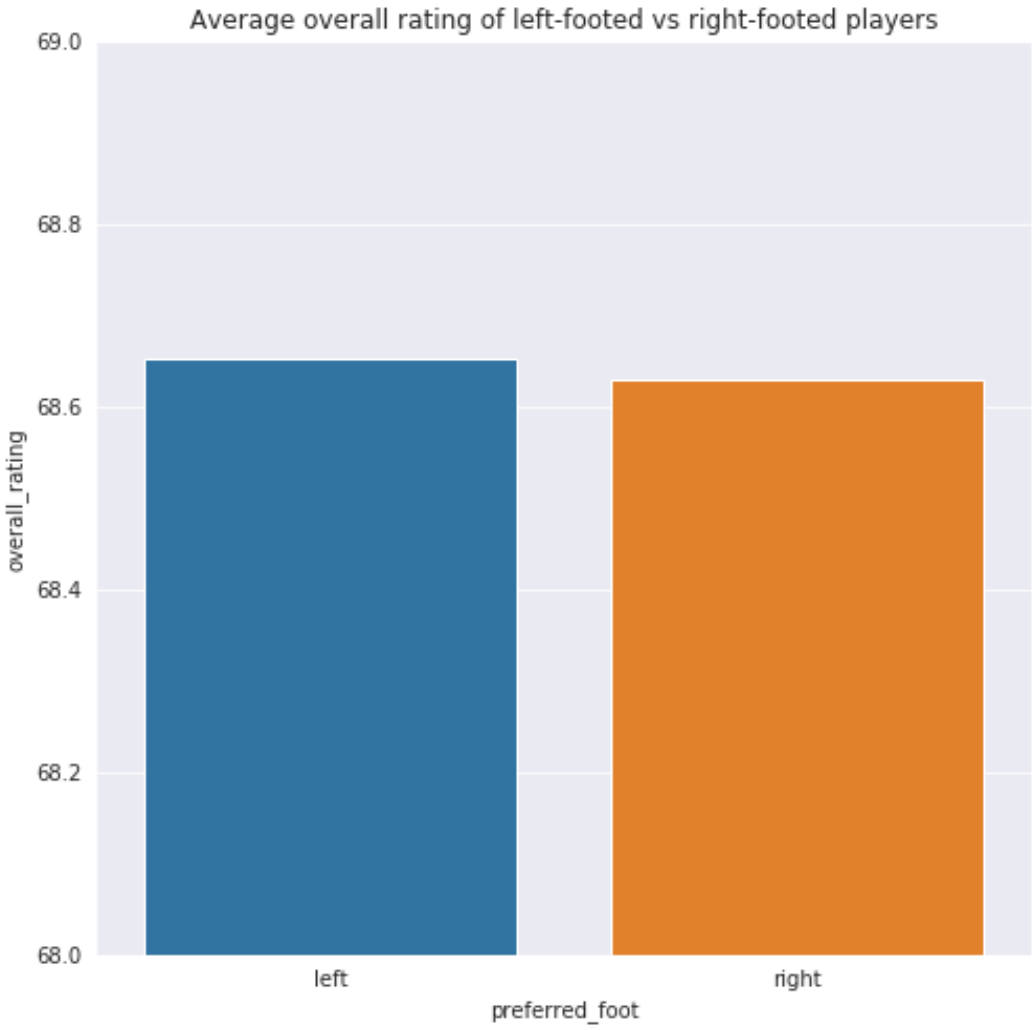
Accuracy measuring



Free kick accuracy: Free kick accuracy improved over the years, reaching its highest in 2015

Heading accuracy: Like overall performance, heading accuracy improved to reach its peek in 2012 and then starts decreasing.

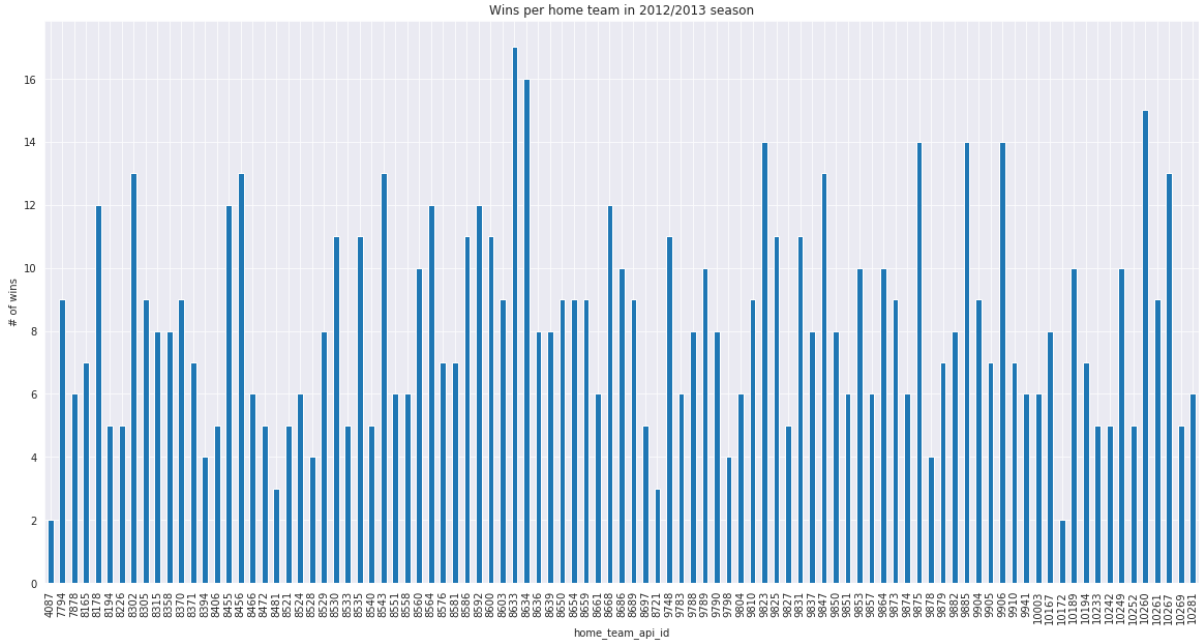
Question 2: Who is better, left-footed or right-footed players?



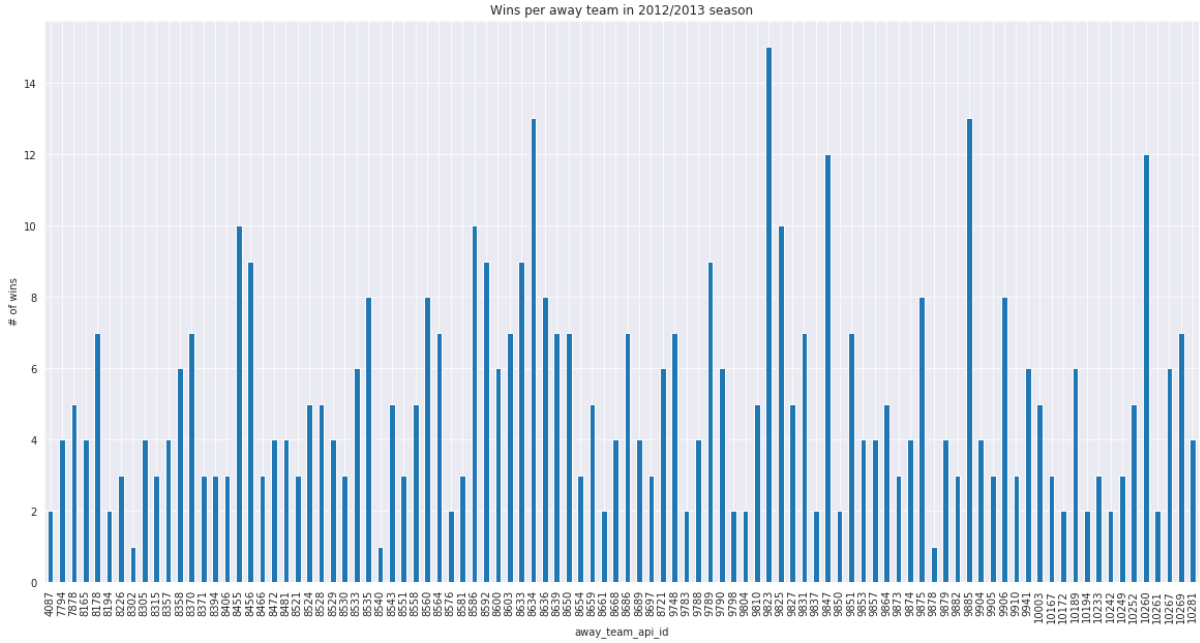
According to the previous plot, the left-footed average players overall rating between 2007 and 2015 is slightly better (by ~0.03).

Question 3: Which is the most winning team in 2012/2013 season (home and away)?

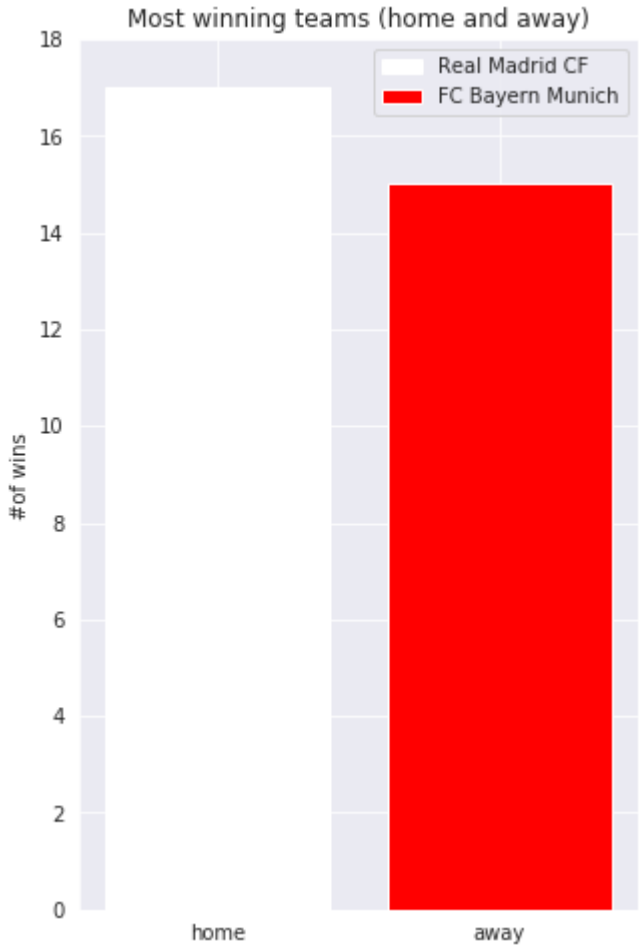
First we plot the number of wins per home team:



then we plot the number of wins per away team:

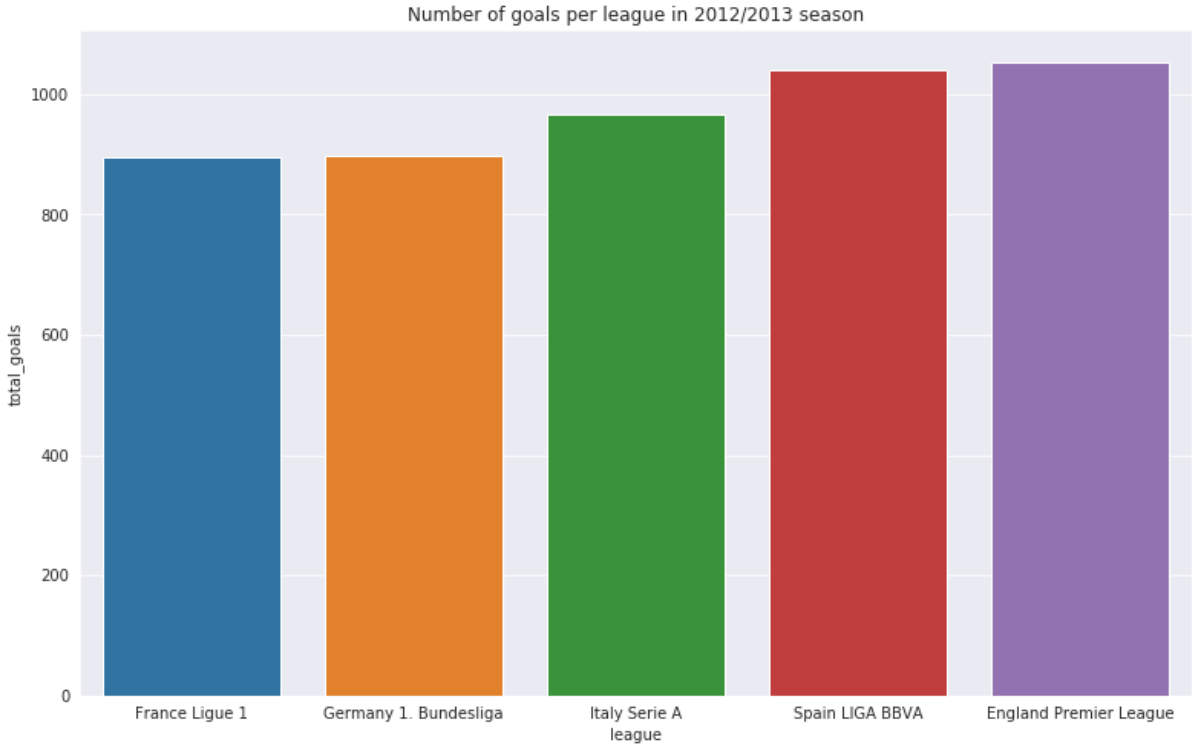


After we determine the most winners we plot them:



The best results in the 2012/2013 season occur in the premier league where the home team with the most wins is **Real Madrid CF** with 17 wins and the away team with the most wins is **FC Bayern Munich** with 15 wins.

Question 4: Which league has the most goals in the 2012/2013 season ?



After plotting match goals grouped by league we can see that England's **Premier League** has the most goals (1054 goals)

4. Conclusions

Question 1: How is Lionel Messi's performance over the years?

By plotting different performance measures for **Lionel Messi** between the years 2007 and 2015 we can conclude that he kept improving to reach his peak performance in 2012-2013, after that most performance measures start decreasing with aging except for free-kick accuracy which does not require high physical performance.

Question 2: Who is better, left-footed or right-footed players?

By plotting the average overall performance of left-footed players and right-footed players we can see that **left-footed** players are slightly better

Question 3: Which is the most winning team in 2012/2013 season (home and away)?

After plotting the number of wins per home team and the number of wins of away team we can see that **Real Madrid** has won the most (17) when playing

as a home team while **FC Bayern Munich** has won the most games (15) when playing as away team

Question 4: Which league has the most goals in the 2012/2013 season ?

The league with the most goals in the 2012/2013 season is **England Premier League** with a total number of **1054** scored goals.

Limitations

- For analysing matches from match dataset we were only limited to the season of 2012/2013, whereas we can get more insights from different seasons.
- Databese was normalized, so I had to merge tables and execute sql-like query using pandas to get information about certain relations.