Investigate a Dataset

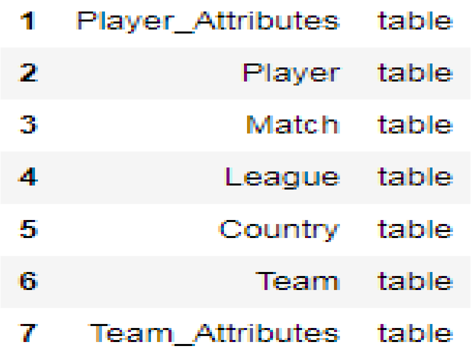
Data analyst nanodegree first project

Introduction:

I have chosen Soccer dataset to analyze it. I chose it specially to develop my skills in sql using python

Datasets:

Here we have only 1 dataset which is stored in a sql database file in this database we have 6 tables which are:



1	Player_Attributes	table
2	Player	table
3	Match	table
4	League	table
5	Country	table
6	Team	table
7	Team_Attributes	table

Software used:

1- SQL to store data in a database

2- Python (Jupyter notebook) to analyze data and visualize it

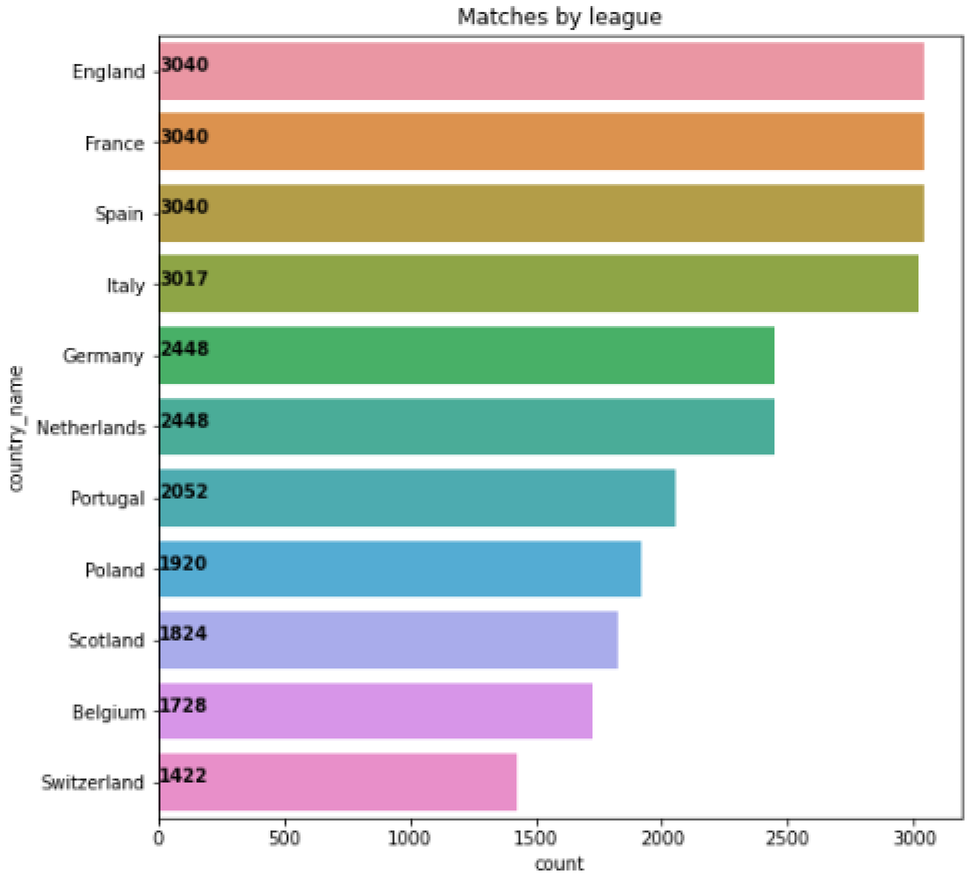
Questions to be answered:

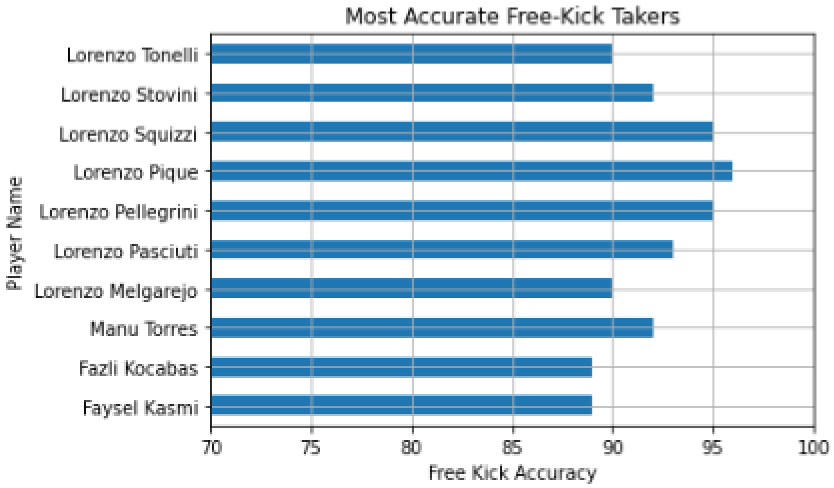
1- Who are most accurate free kicks taker?
2- Which is more home or away teams' goals in each league?
3- How many matches played by Season in each league?
4- Which teams are the top by home and away goals number in all seasons?
5- Is there any advantage for home team over away team?
6- Which teams are the best winners in each season?
7- Which teams are the worst loser in each season?
8- Which teams are the best all over the time?

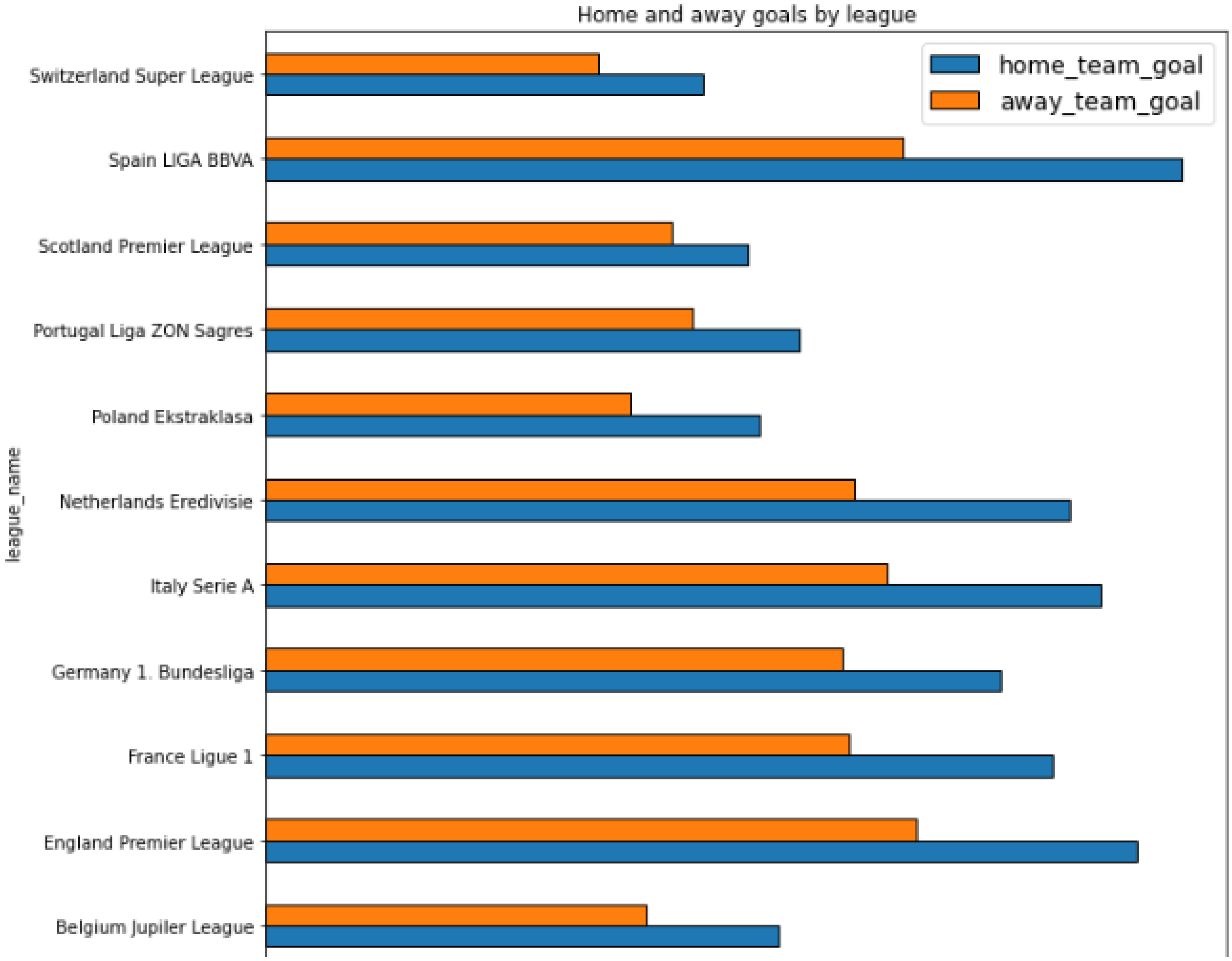
Dealing with nulls

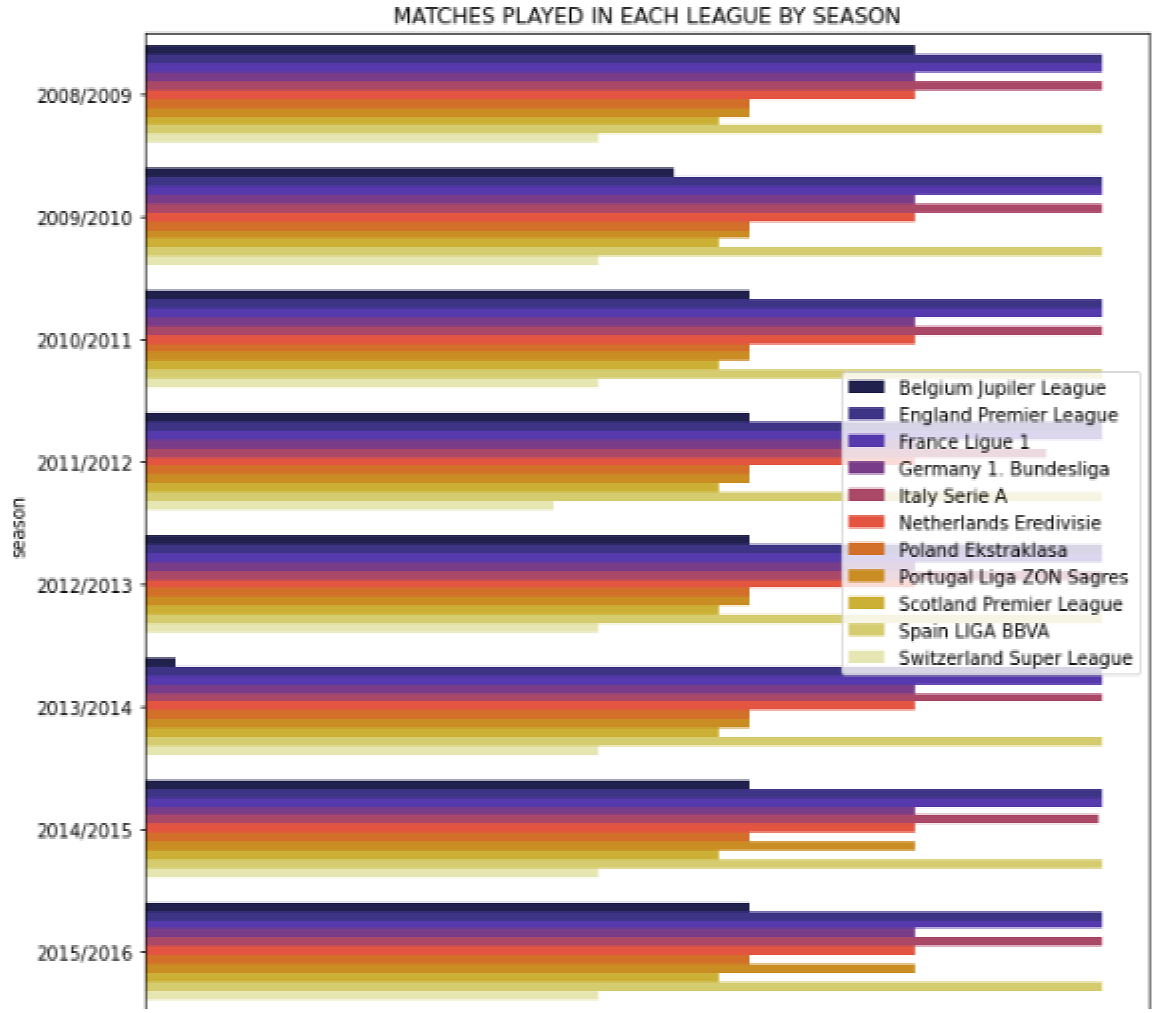
In many cases I drop nulls in my analysis except for **buildUpPlayDribbling** has null values which are 969 but we have another column **buildUpPlayDribblingClass** which can provide info for null values to impute it

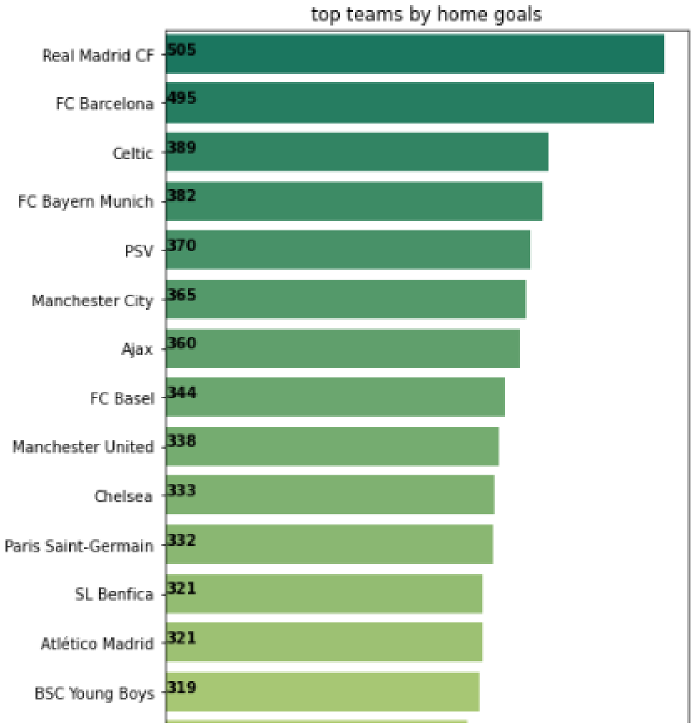
And after investigation i figure out that number from 41 are classified to normal so i choose to impute missing values with an average value of **35 but I didn't use it in my investigation**

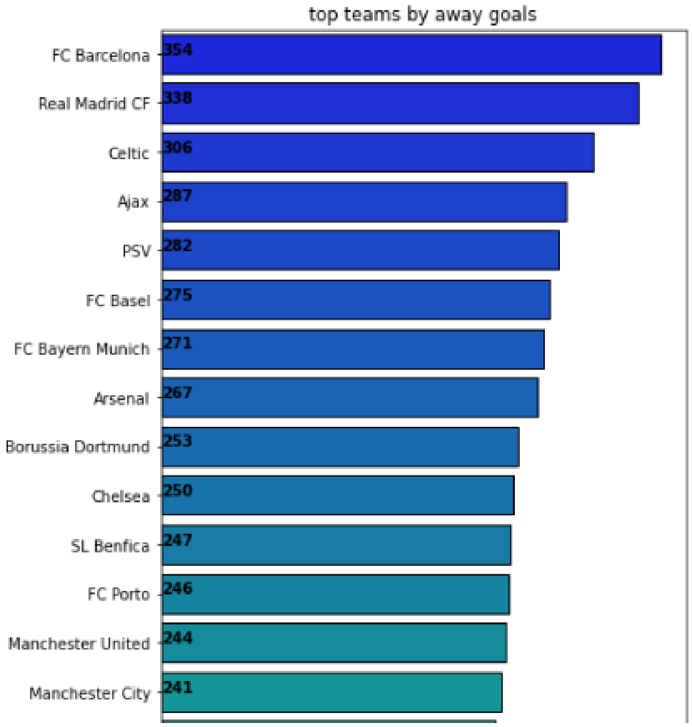


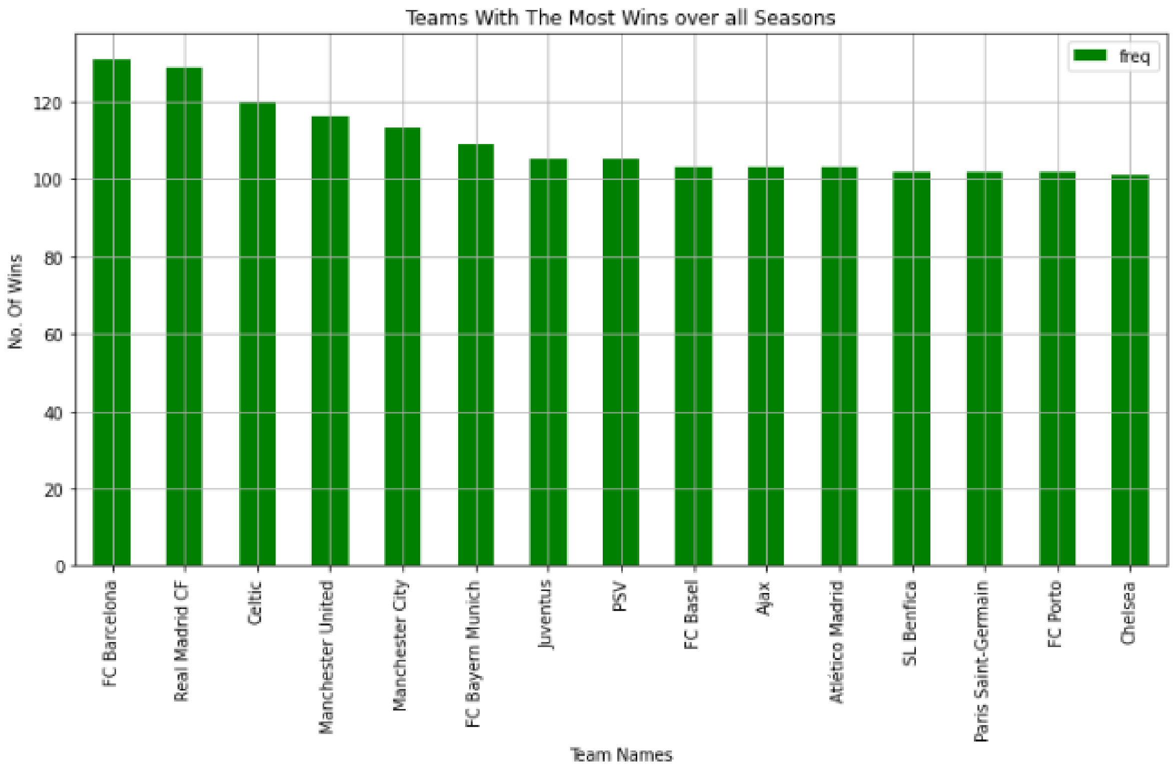

Most Accurate Free-Kick Takers
Player Name
Lorenzo Tonelli
Lorenzo Stovini
Lorenzo Squizzi
Lorenzo Pique
Lorenzo Pellegrini
Lorenzo Pasciuti
Lorenzo Melgarejo
Manu Torres
Fazli Kocabas
Faysel Kasmi
70
75
80
85
90
95
100
Free Kick Accuracy


Home and away goals by league
home_team_goal
away_team_goal
league_name
Switzerland Super League
Spain LIGA BBVA
Scotland Premier League
Portugal Liga ZON Sagres
Poland Ekstraklasa
Netherlands Eredivisie
Italy Serie A
Germany 1. Bundesliga
France Ligue 1
England Premier League
Belgium Jupiler League


MATCHES PLAYED IN EACH LEAGUE BY SEASON
season
2008/2009
2009/2010
2010/2011
2011/2012
2012/2013
2013/2014
2014/2015
2015/2016
Belgium Jupiler League
England Premier League
France Ligue 1
Germany 1. Bundesliga
Italy Serie A
Netherlands Eredivisie
Poland Ekstraklasa
Portugal Liga ZON Sagres
Scotland Premier League
Spain LIGA BBVA
Switzerland Super League


top teams by home goals
Real Madrid CF 505
FC Barcelona 495
Celtic 389
FC Bayern Munich 382
PSV 370
Manchester City 365
Ajax 360
FC Basel 344
Manchester United 338
Chelsea 333
Paris Saint-Germain 332
SL Benfica 321
Atlético Madrid 321
BSC Young Boys 319


top teams by away goals
FC Barcelona 354
Real Madrid CF 338
Celtic 306
Ajax 287
PSV 282
FC Basel 275
FC Bayern Munich 271
Arsenal 267
Borussia Dortmund 253
Chelsea 250
SL Benfica 247
FC Porto 246
Manchester United 244
Manchester City 241



Conclusion:

No doubts here, home advantage is a very big part of football. Very interestingly even though La Liga takes the cake in terms of most number of goals score EPL still has more away goals scored. Also important to note here that even though the German League has significantly lower number of matches (2448 for 8 seasons whereas others have 3040) it still rakes in a lot of goals and has the highest average goals per game amongst all.

I believe I could have expressed the visualizations in a clearer manner with different kind of plots which I have yet to learn, and draw correlations more easily & quickly.
I have concluded that home team stadium and audience didn't contribute a lot in matches result as many leagues away teams win the matches

Real Madrid and Barcelona both are top teams in all leagues in away and home goals they have incredible attack line and striker like **Cristiano Ronaldo and Messi**

References:

I use some of kaggle kernel to help my in my works