Soccer Database Report

Introduction

I selected this data because I love football and the most reason for selecting these is I knew in football very much which make it easy for me to analyze these data.

I used all the techniques that I learned from Udacity and other places to make my analysis easy and more readable from all genres of people.

The questions I want to answer are:

1- Who is the best player in the world?
2- How many goals for each season and which season has scored the most goals?
3- Who are the best two players in penalties for each season with his score?
4- Who are the best two players for each season with their score?
5- Who are the best two players in finishing for each season with their score?
6- Who are the best two teams in defence pressure for each season?
7- What is the relationship between ball control, crossing, finishing and short passing?
8- Who is the best team in buildup play speed ?

Data Wrangling Process

This process consists of three steps:

1- Data Gathering
2- Data Assessing
3- Data Cleaning

Data Gathering

This is the first step in Data Wrangling Process.

I gathered these data from database.

Data Assessing

This step is useful when the data is dirty but fortunately the data is cleaning

So we didn't need this step.

Data Cleaning

As we said the data is cleaning, It only needs some small operations such as converting the data in the column from one type to another, adding some important columns that are calculated by statistics and so on...

It is just very simple operations that do not require the previous step.

In these dataset we added league_name column to df_match data frame because I need these column, I also convert birthday column from string data type to datetime because I needed these column to calculate the age and add this new column to the data frame.

Data Exploratory Process

The purpose of this process is to show your ability for asking good questions about data and to answer these questions accurately, it is more interesting process.

The first question crossed my mind is (Who is the best player in the world?) what I did in this question is I grouped the data by player name and took the average rating for each player and put all in the dictionary and iterated through the dictionary and took the max rating then I printed the player name which is Lionel Messi.

The question number 2 is (How many goals for each season and which season have scored the most goals?), what I do is we first create a figure then we group the data by seasons and calculate the number of goals for each season. We make the plot more readable by putting x and y label and title. The last line of code is interesting for me, we put above each column the number of goals so there is no misleading.

The question number 3 also is more interesting (Who are the best two players in penalties for each season with his score?) in this question we needed to know who are the best two players for each season in penalties. We grouped the data by two indexes (season, player name) and calculate the average of the penalties. We created seasons variables which contain a list of all seasons, we needed it in iteration. We iterated over seasons then we create a new figure for each season. We

create a X data frame which sort data in descending order and select only the first two rows. Then we plotted a point plot which the x axis is the player name and the y axis is the average rating in penalties. We putted the labels and title. The last line of the code made some space between each plot.

The question number 4 is (Who are the best two players for each season with their score?) in this question we needed to know who are the best two players in each season with their score. We grouped data by two indexes too (season, player name) and calculate the average of overall rating. We iterated over seasons then we create a new figure for each season. We created a X data frame which sort data in descending order and select only the first two rows. Then we plotted a point plot which the x axis is the player name and the y axis is the average of overall rating. We putted the labels and title.

The question number 5 is (Who are the best two players in finishing for each season with their score?) in this question we needed to know who are the best two players in each season in finishing with their score. We grouped data by two indexes too (season, player name) and calculate the average of finishing. We iterated over seasons then we create a new figure for each season. We created a X data frame which sort data in descending order and select only the first two rows. Then we plotted a point plot which the x axis is the player name and the y axis is the average of finishing. We putted the labels and title.

The question number 6 is (Who are the best two teams in defence pressure for each season?) in this question we needed to know who are the best two teams in each season in defence pressure with their score. We grouped data by two index too (season, team name) and calculate the average of defence pressure. When I tried to plot all seasons, there were errors because the data in some seasons missing. We iterated over seasons then we create a new figure for each season. We created a X data frame which sort data in descending order and select only the first two rows. Then we plotted a point plot which the x axis is the team name and the y axis is the average of defence pressure. We putted the labels and title.

The question number 7 is (What is the relationship between ball control, crossing, finishing and short passing?) what I did in this question is I plotted a scatter plot then made the x axis is ball control and the y axis is crossing and the size of points short passing and the color of the points is finishing, the results, there are a positive relationships among these attributes.

The question number 8 is (Who is the best team in buildup play speed?) and the answer is first I sorted the values by the buildup play speed in descending order and selected the first seven teams then I plotted a bar chart which the x axis is the team name and the y axis is the score of the buildup play speed, the result was the best team in buildup play speed is carpi.

Conclusion

Results

After analyzing this data, I found that the data is clean enough to work with it without any effort, and sufficient to answer the questions I putted. The data also doesn't have duplicates rows.

Our data suggests that:

1- The best player in the world is Messi.
2- The shot power for most players is between 60 and 80.
3- The season which scored the most goals is 2015/2016.
4- The best team in buildup play speed is Carpi.
5- There is a positive relationship among ball control, crossing, finishing and short passing.
6- The team that has the best score in defence pressure is FC Bayern Munich.
7- The player that has the best score in finishing is Messi and in penalties is lambert.

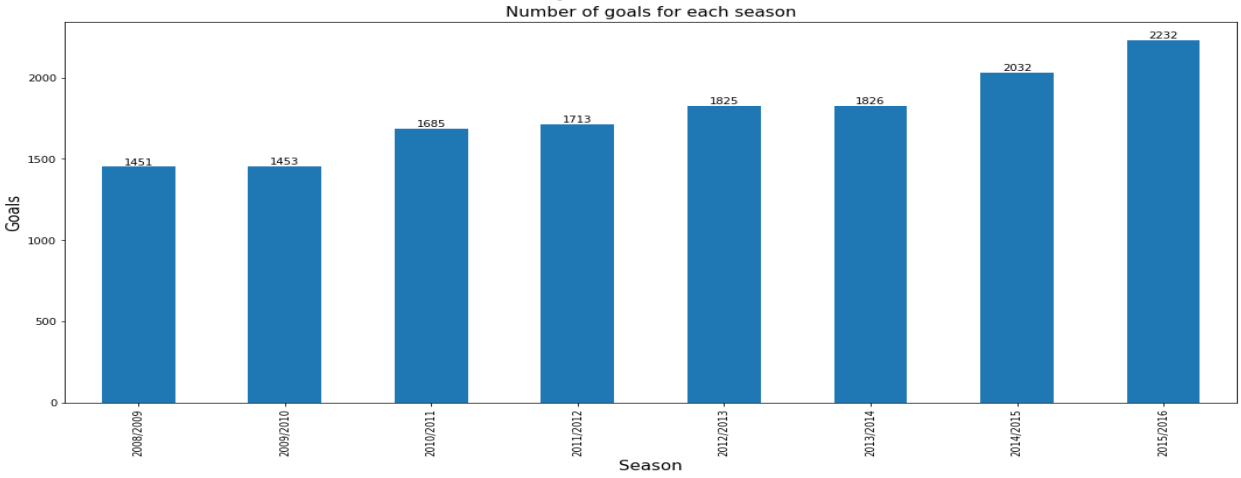
Limitations: There are some issues with this data:

1- There are missing values in our data which is not good for our analysis.
2- There are too many columns that don't useful.
3- There are some data that doesn't in appropriate type.
4- There are some columns that really important and don't exist.

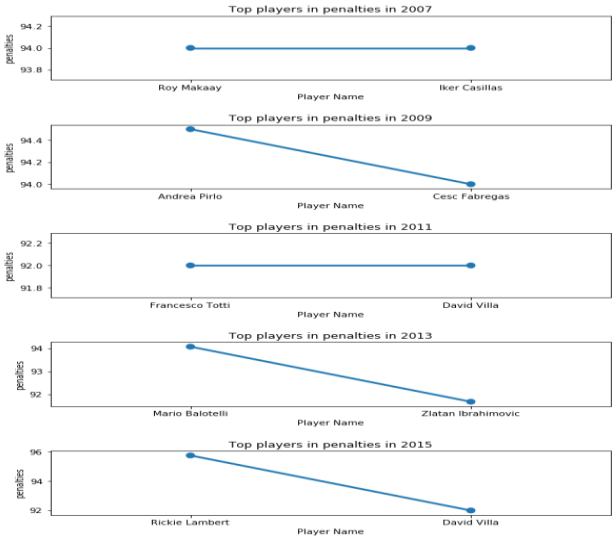
Communication results

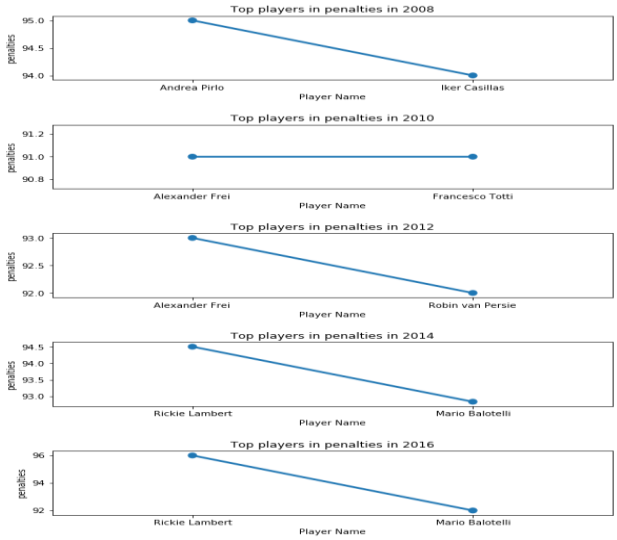
I will share my results with you. After analyzing these data, I discover that the best player in the world is Lionel Meesi.

The season which is scored the most goals is 2015/2016.

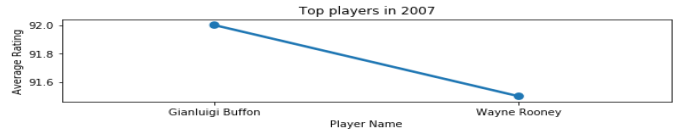


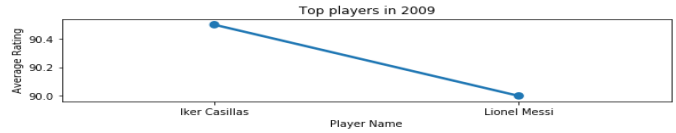
I also discover the best player in penalties in 2007 is Roy Makkay, in 2008 is Pirlo, in 2009 is Pirlo too, in 2010 is Alexander Frei, in 2011 is Totti, in 2012 is Alexander Frei, in 2013 is Balotelli, in 2014 is Lambert, in 2015 is Lambert too, and in 2016 is Lambert.

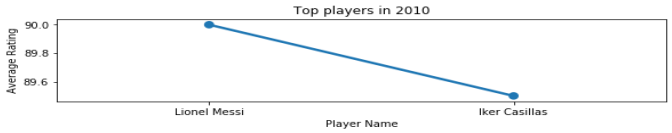


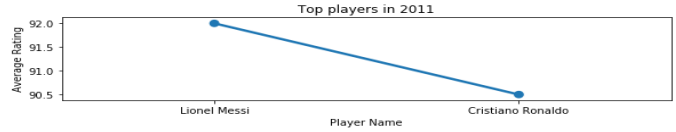


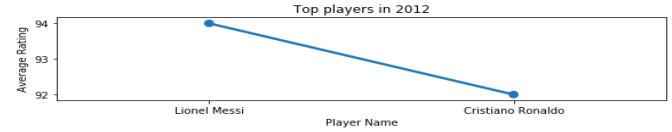
And the best two players for each season in the average rating and finishing are

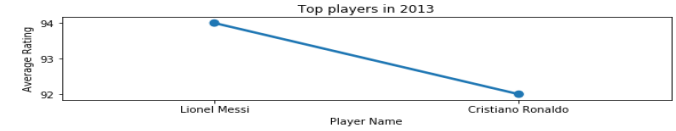


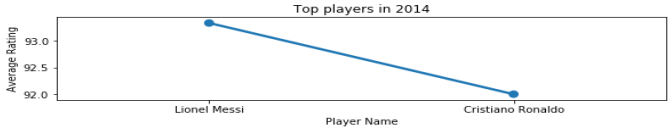


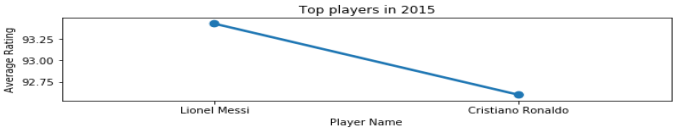


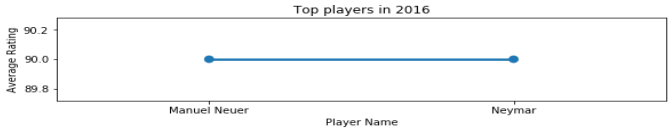


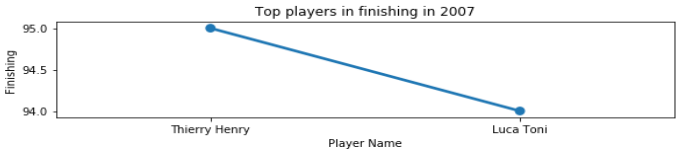


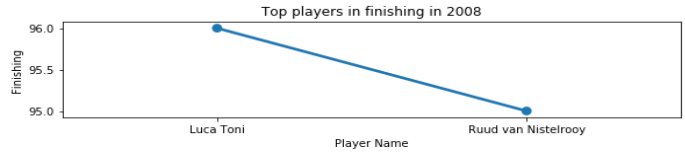


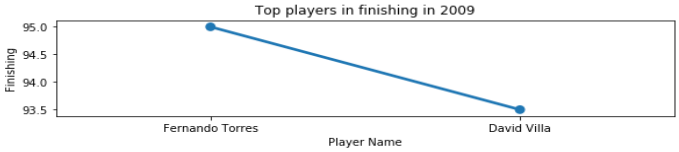


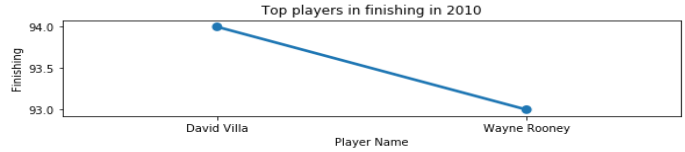


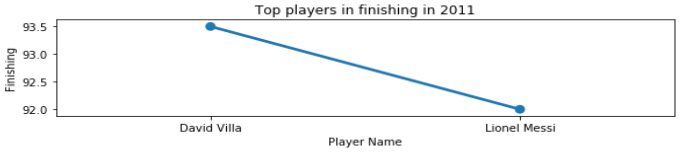


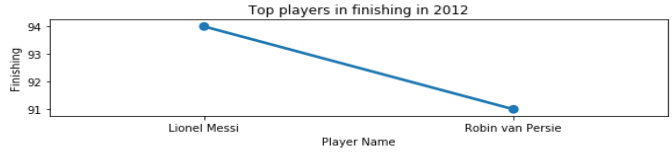


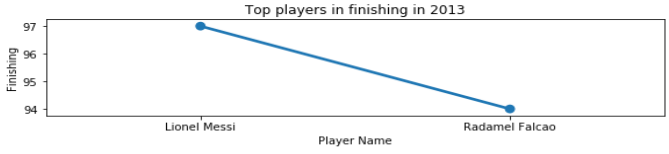


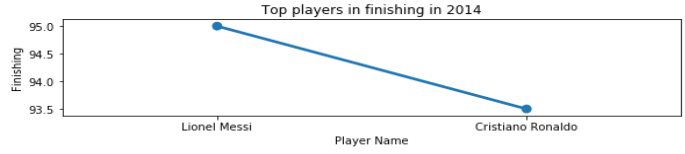


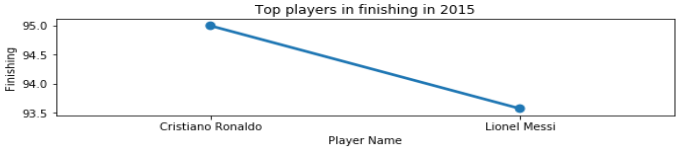


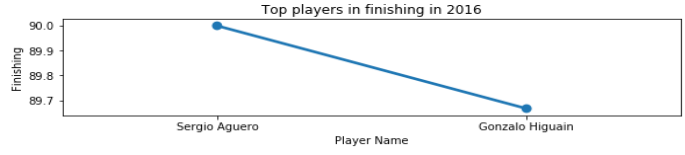




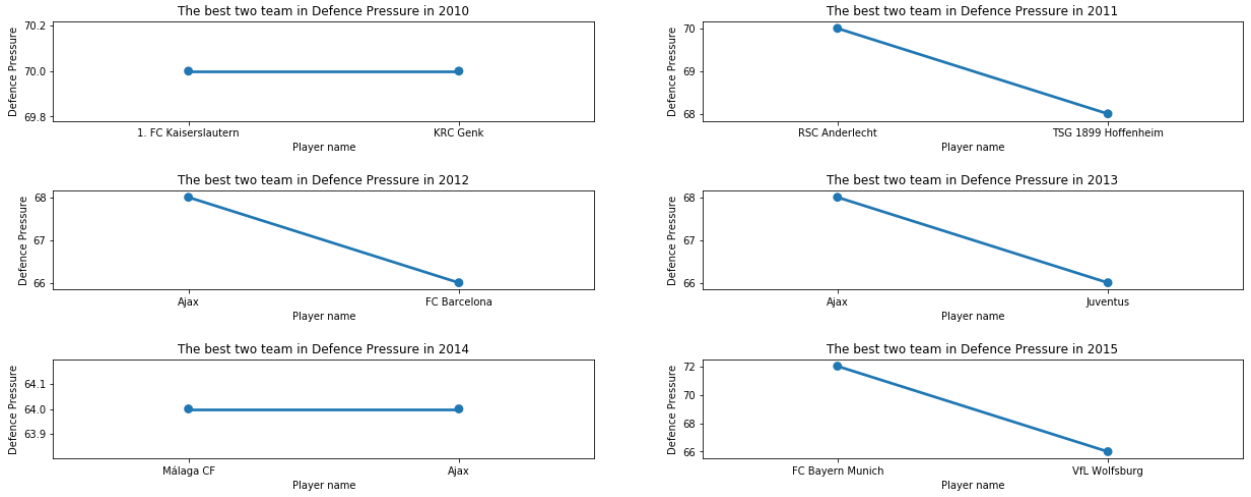




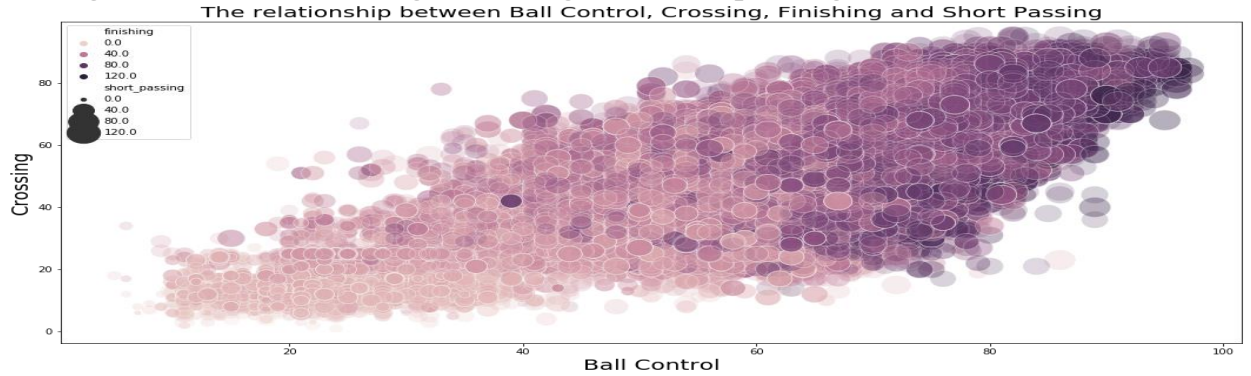




And the best two team for each season in defence pressure are



When I did summary statistics, I discovered that there are a positive relationship among ball control, crossing, finishing and short passing.



I also found the best team in buildup play speed which is Carpi.