Report: European Soccer Database Investigation

Table of Contents

Introduction

- This soccer database comes from Kaggle. It contains data for soccer matches, players, and teams from several European countries from 2008 to 2016.
- The database is stored in a SQLite database. You can access database files using software like DB Browser.
- After exploring the data in DB Browser, we are going to assess and clean the data so that we can answer these research questions:

1. **What player attributes is linked with the high potential & overall rating?**
2. **What teams attributes lead to most goals scoring & most defeated teams?**

Data Wrangling

Data Gathering & Loading

- Data is downloaded from kaggle through this link.
- Data is loaded using pandas & SQL select queries.

Data Assessing

1) Quality

1. 'player_id', 'team_id' & 'match_id' dtypes are inconsistent
2. 'birthday' & 'date' (in the three datasets) dtypes are inconsistent
3. 'birthday' is less meaningful than age
4. Missing data of 'buildUpPlayDribbling' in teams
5. Missing data in players
6. Duplicate data in players and teams

Data Cleaning

> Creating copies of the data in order to start cleaning them

1) Quality

1. 'player_id', 'team_id' & 'match_id' dtypes are inconsistent

- Convert all 'ids' dtypes to strings

2. 'birthday' & 'date' (in the three datasets) dtypes are inconsistent

- Convert all dates dtypes to datetime

3. 'birthday' is less meaningful than age

- Replace 'birthday' column with 'age' column after calculating it using 'date'

4. Missing data of 'buildUpPlayDribbling' in teams

1. Group data by 'buildUpPlayDribblingClass' to find the mean of each class
2. Fill null values with the mean of each class

5. Missing data in players

- Remove rows containing null values from 'players' data

6. Duplicate data in players & teams

- Remove duplicates from both datsets

Exploratory Data Analysis

Research Question 1: What player attributes is linked with the high potential & overall rating?

- 1) First, we will explore all player attributes to see patterns
- 2) Then, we will find the top rating players
- 3) Finally, we will compare the average attributes of the top 5 rating players & the top 5 potential players to the average attributes of other players

Research Question 2: What teams attributes lead to most goals scoring & most defeated teams?

- 1) First, we will modify 'match_df' dataframe to obtain points, goals, victories and losses for each team
- 2) Then, we will find the teams matching our criteria
- 3) Finally, we will find the attributes of those teams from 'teams_df' dataframe

Conclusions

1) Limitations in player attributes data:

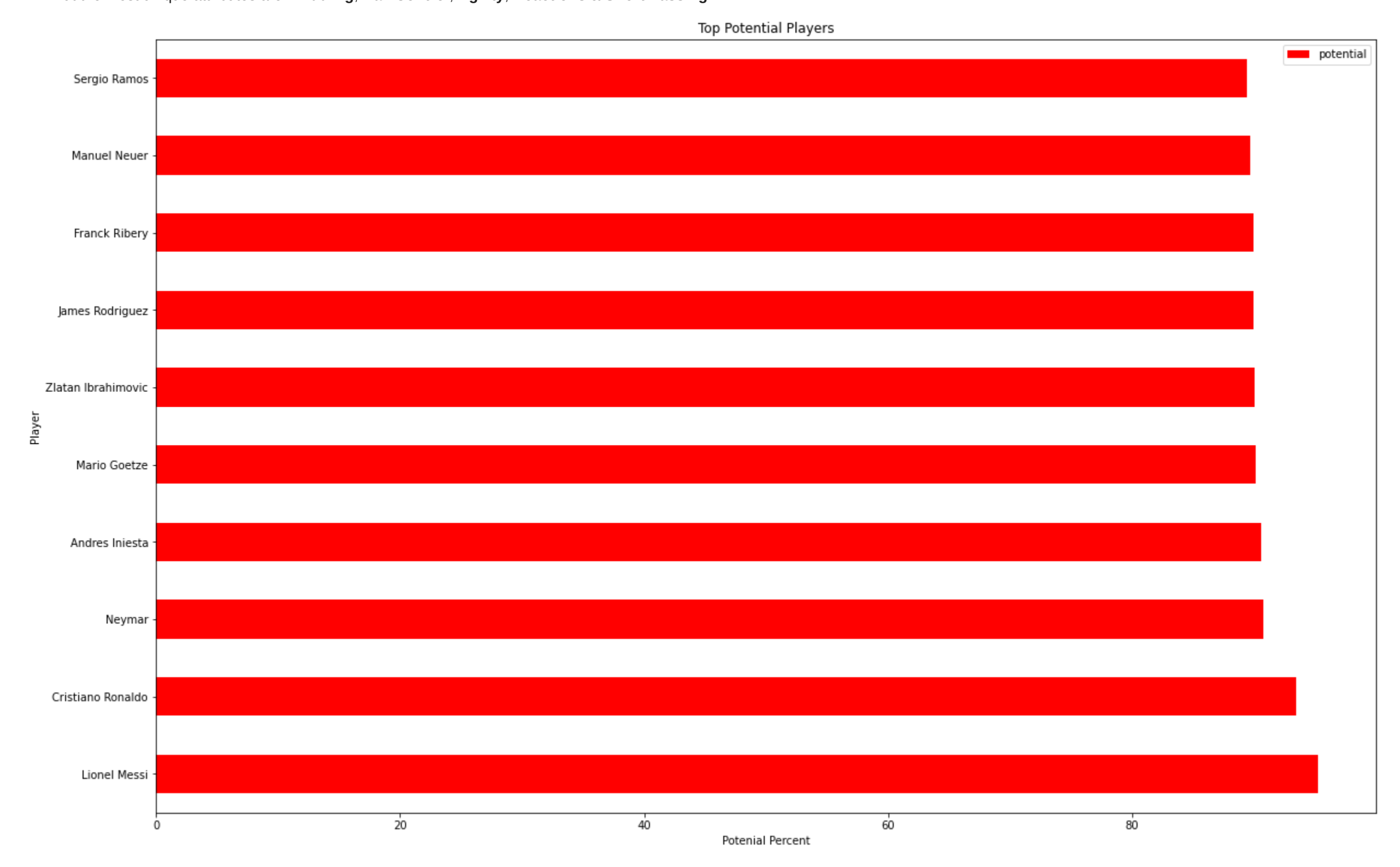
- It seems to be there are a lot of outliers in almost all attributes, but actually, they are no outliers.
- The thing is that there is a very important missing factor in these comparisons which is the position of the players or in other way the role of its player in the game.
- As the attributes of defenders will not be like the forwards, and attributes of gool keepers will differ from those of midfielders. Of course, they might share some like ('agility', 'reactions', 'balance', 'jumping', 'stamina', 'strength', 'aggression') but still there are big gaps between them.
- The funny thing is that the outliers in all gool keeping attributes (gk_diving to gk_reflexes) are the data of the actual gool keepers which we should explore and study, not get rid of. Also, they are shown as outliers because the ratio of goal keepers is about (3/23) of each team.
- So, this is a big limitation in this dataset, not mentioning the player role in the field beside his attributes to enable us to filter by roles and get accurate and efficient analysis.

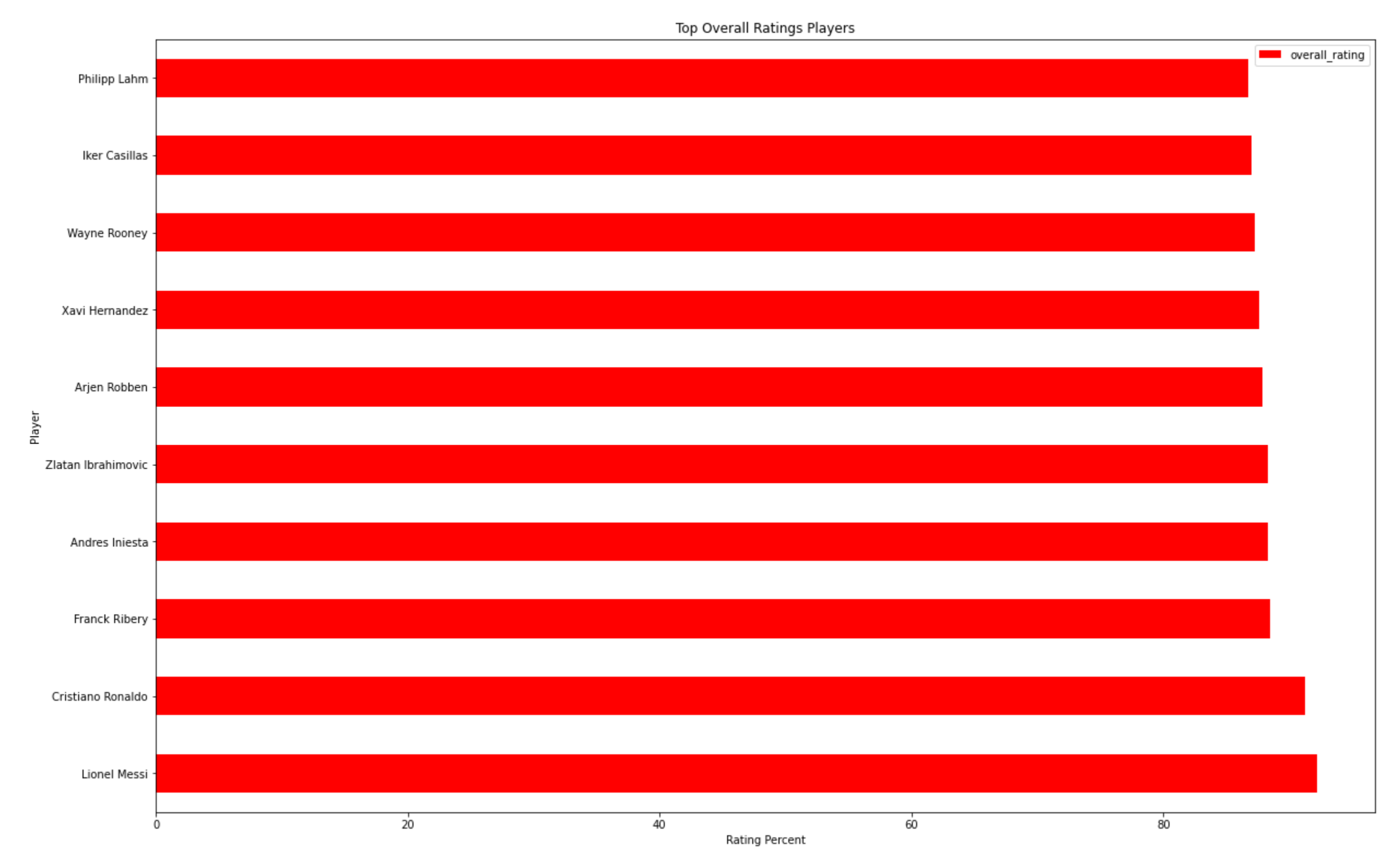
2) Limitations in teams & players data:

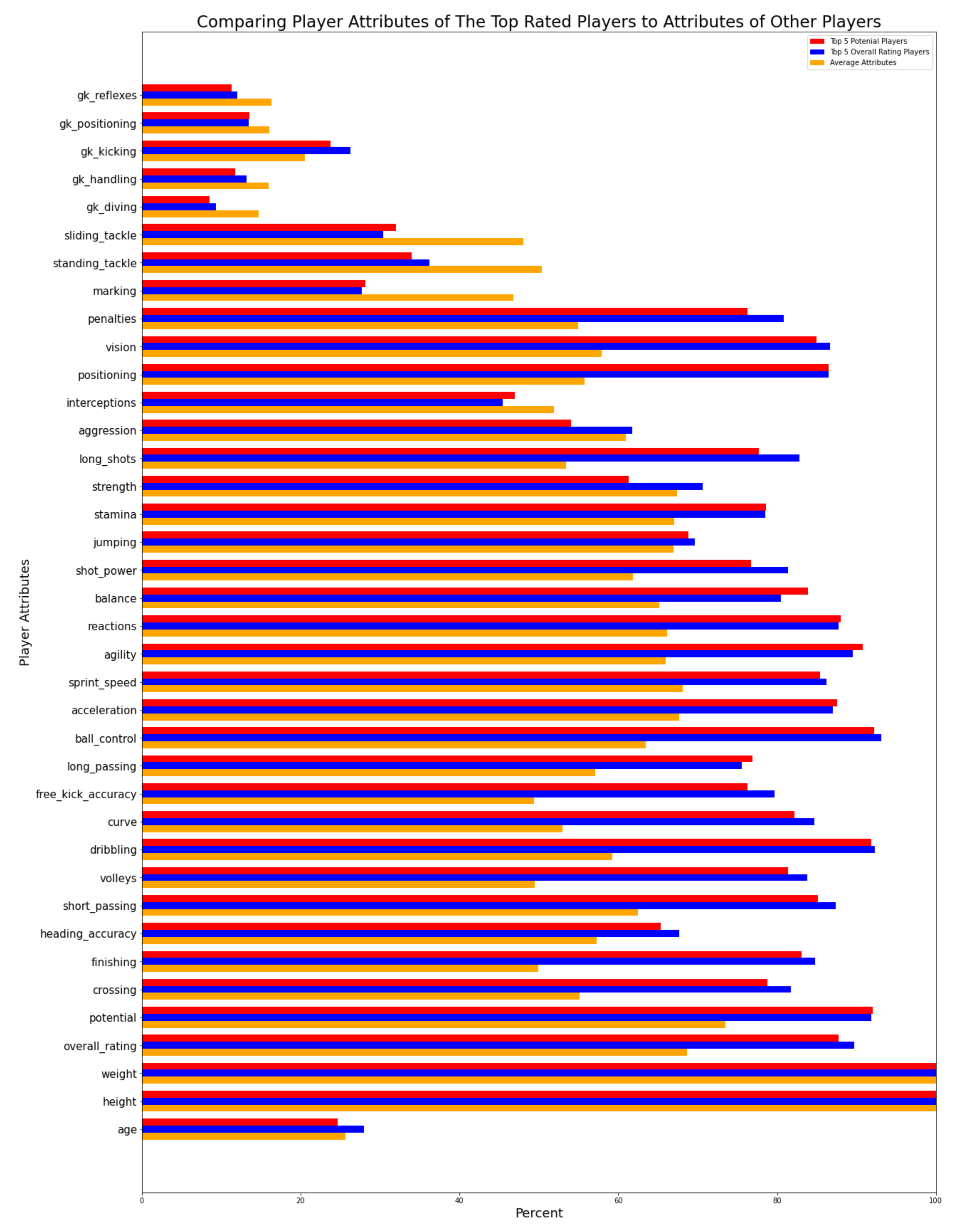
- Not knowing whom player belongs to which team is another big limitation in this dataset.
- Even though, if we tried to get players' teams from "match" data starting players, it will take forever and it will not be accurate.
- Imagine a team with three high potential players compared to another team with 7 high potential players compared to a team with no high potential players, would these conditions affect team results, wins and goals scored through the season?
- I think it will have a big influence, but unfortunately, we cannot make sure of it because lack of data.

3) The top potential & overall rating players:

- "Lionel Messi" & "Cristiano Ronaldo" are sitting on top of both lists as shown in the below figure.
- Top rated players are higher than other players in almost all attributes except the gool keeping ones.
- But the most unique attributes are **Dribbling**, **Ball Control**, **Agility**, **Reactions** & **Short Passing**.

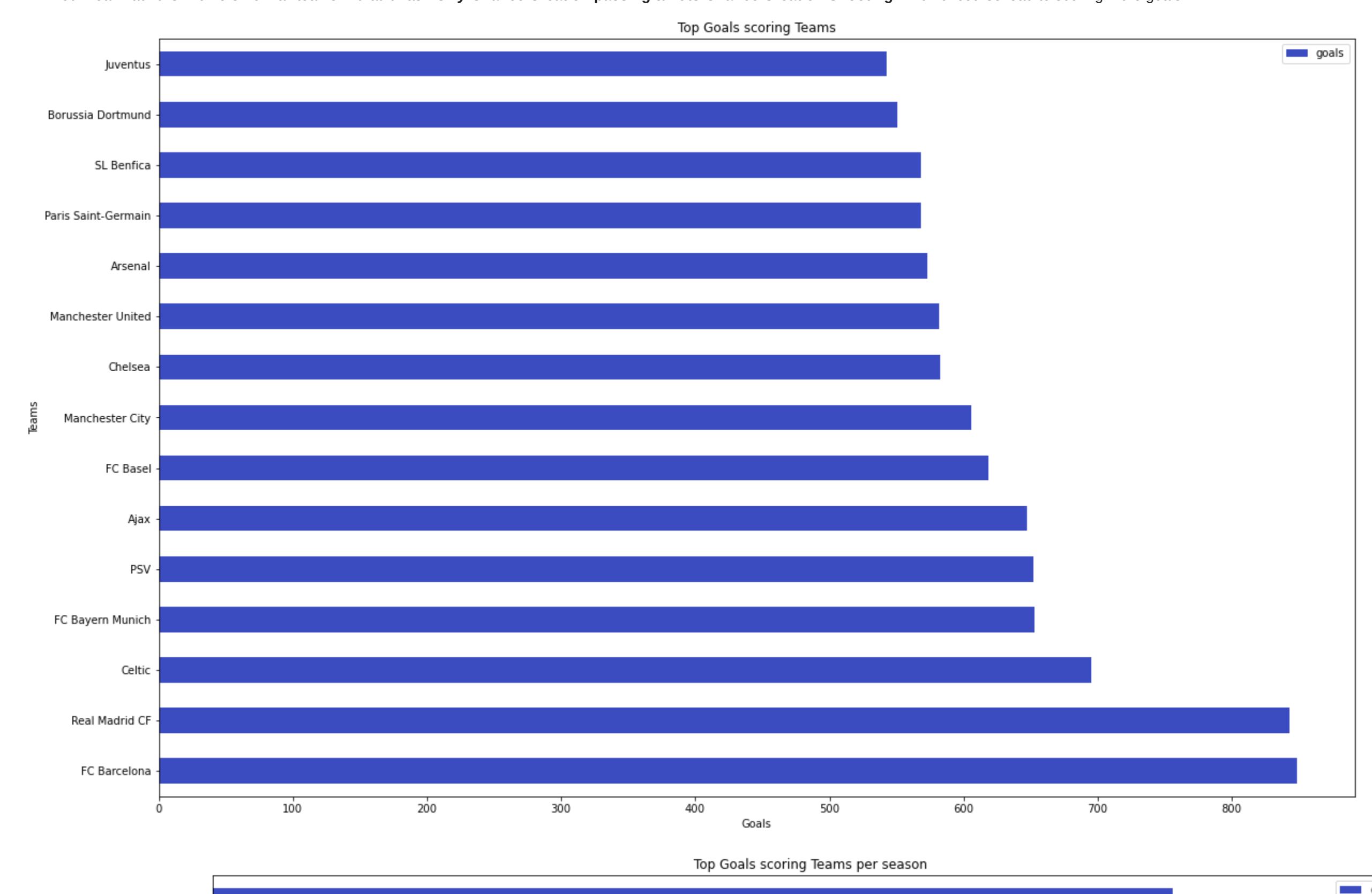


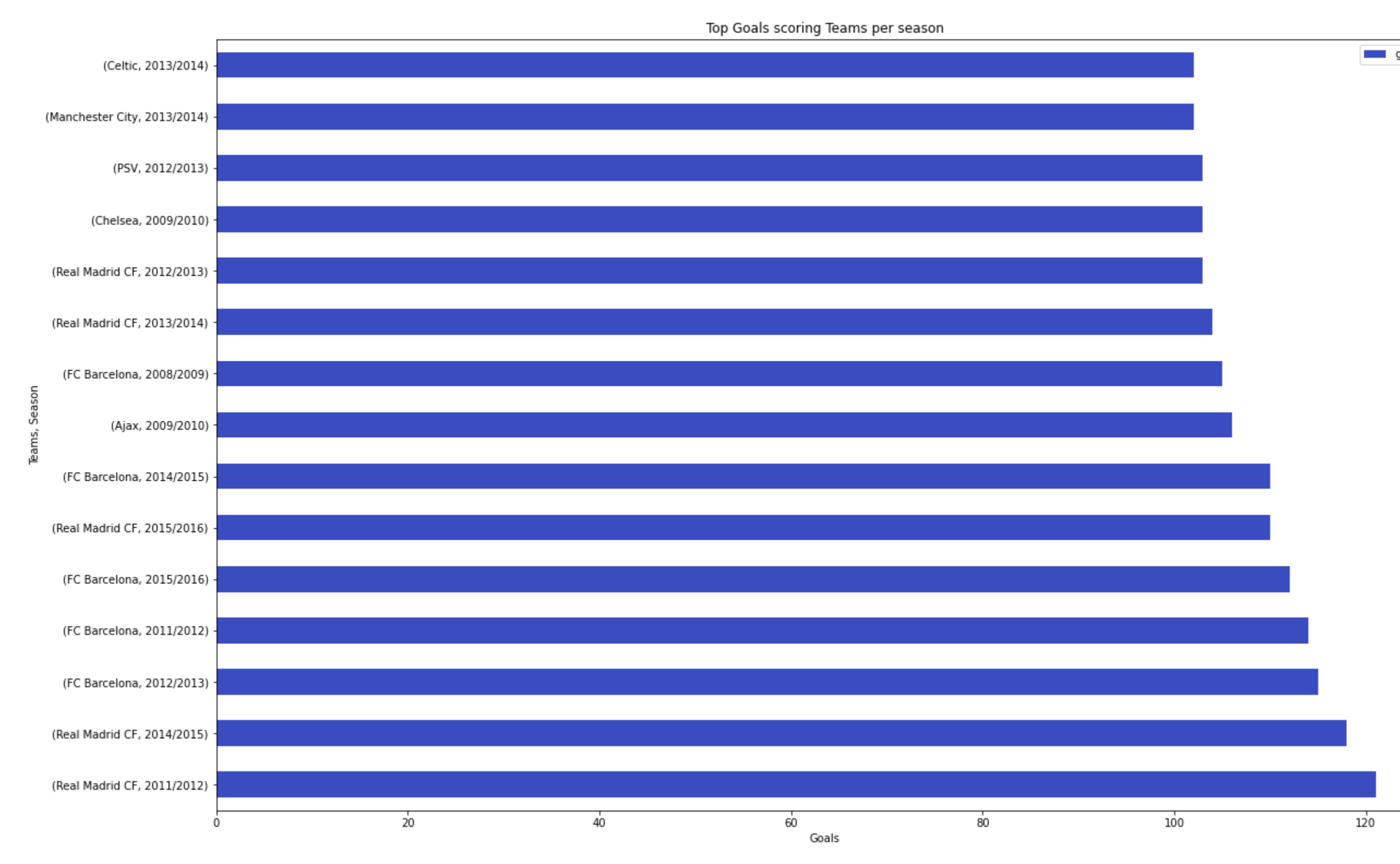


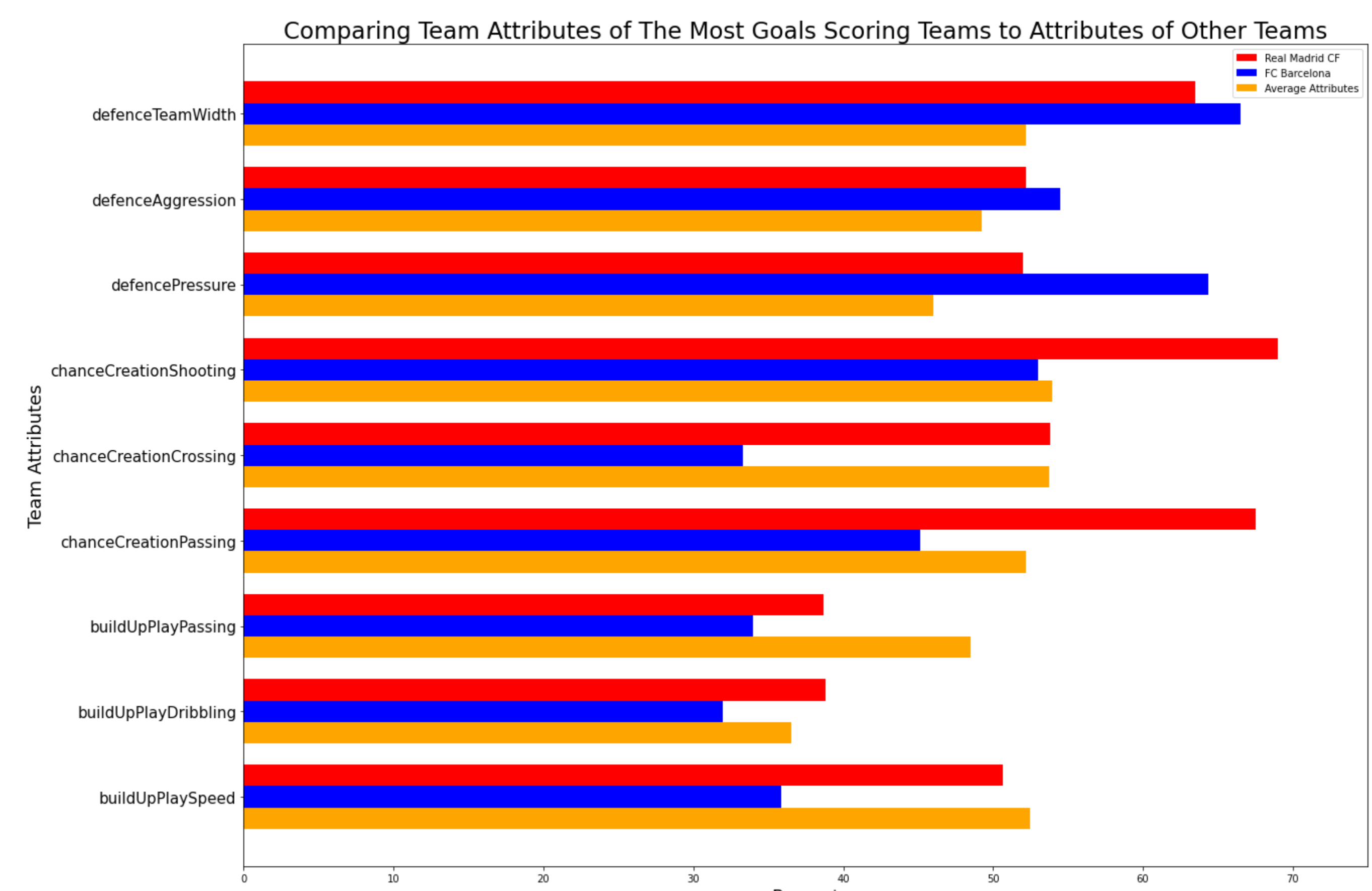


4) The most goals scoring teams:

- "FC Barcelona" & "Real Madrid CF" are in the top as in the below figure.
- They both have two common attributes that differ from other teams: ***Mixed*** **Build Up Play Passing** & ***almost Wide*** **high Defence Team Width**.
- But "Real Madrid CF" differs from all teams in that it has ***Rsiky*** **Chance Creation passing** & ***Lots*** **Chance Creation Shooting** which of course lead to scoring more goals.







5) The most defeated teams:

- "Kilmarnock", "Getafe CF" & "Sunderland" are the top three defeated teams as in the below figure.
- "Kilmarnock" & "Sunderland" attributes do not differ much from other teams attributes.
- But "Getafe CF" does differ in that it has ***Very slow*** **Build Up Play Speed** ***little*** **Chance Creation Crossing** & ***little*** **Defence Aggression** which of course lead to receiving more goals resulting in more defeats.