DATA ANALYSIS OF EUROPEAN SOCCER DATABASE

ABSTRACT

In Today's world the Ad agencies are at top positions in marketing the products for reaching to people choices. The perspective of Ad agencies is changing from the past decades to the current date. Most of them choose the celebrities with high fame to shoot an Ad so that their fans have a zeal to try the product as their favorite person is using them. There is no denying that professional athletes have an impact on their supporters' shopping choices. Fans of a team will purchase merchandise featuring their favorite players, including tickets, jerseys, T-shirts, and other items. This effect may even extend to goods like food, automobiles, and clothing that are produced outside of a team environment. A powerful, swift, athletic player can become a good target market for an advertiser that utilizes that player to support a product since the fan wants to emulate those qualities.

To determine how well the players influence Ad Agencies is the main goal of the analysis of the "European Soccer Database." In this project, we typically examine the gaming trends from a database that contains information on soccer matches from 2008 to 2016 played in various nations. This pertains to data on athletes, games, leagues, nations, etc. In order for the advertising agencies to use the data, we will analyze it to derive relevant insights and determine which player and country has a bigger impact on individuals. At the end, we can analyze a person's fan base with maximum games won and their rating which can be helpful for the Ad agencies to shoot some Ad's with the person so that they can help the corporate sector in reaching their targets. Moreover, identified the parameters that can influence the game or player over time.

INTRODUCTION:

We all know how sports have an impact on people these days. Earlier this is just a purpose of entertainment but today it really got changed the way people look at it. It is more beyond entertainment as people showing more interest into it. One of such famous sport is Soccer which is the favorite game for most of the people across the globe. As FIFA is world's one of the best sport draws attention of millions, simply it consists of two teams each team has 11 players and uses only body except their hands and arms to pass the ball, and their will be a goalkeeper to safeguard the goal from opponent team. So, they will be so many popular or big companies interested in promoting their bands on this platform by advertising or by supporting the football club. Therefore, we are working on the European soccer database

The primary motive for analyzing the "European Soccer Database" is to see how well the players influence Ad Agencies. In this project we usually analyze the gaming patterns from the database with data about soccer games from 2008 to 2016 across different countries. This involves the information about players, matches, leagues, countries, etc. We will analyze the data to extract the meaningful insights and understanding which player has a greater impact on people so that the Ad agencies can make use of it.

LITERATURE REVIEW:

There are many analysis that were made on European soccer game. Most of them speaks the same way that covers the key points of total matches country wide. One of the interesting analysis that is being carried out is the goal comparison with league which helps to depict the high chances of attaching ability and winning. Besides that, comparing the total goals per league helps to know year wide also shows the improvement of game across the league. The Overall rating of the player helps to evaluate the fame of the player along with their physical traits and gaming strategies.

"*I would like to know if a player's potential for assisting a goal is high, his potential for making a goal is also high*"[3]. The statement made speaks the reality that the goals have an higher impact on a player to win. Besides that, the player physical strategies also help to achieve the goal in a better way.

Ahmed Mohamed AbdElHameed AbdElHameed research says that "*There is a positive correlation between attack features and a negative between attack and defense features*"[4]. The interesting point is that the correlation speaks the relation between the data and it can help us to draw the better insights. These important attributes and correlations among them may affect the overall rating of a player. With this inference, we have considered highly correlated attributes like agility, sprintspeed, and free kick accuracy to determine the overall rating of the player.

Most of the analysis till today was more focused on analysis the game and highlighting the key take aways like the country with highest matches, leagues with highest goals, players that were born in a particular month. From all this analysis, one of the major insight that can also be extracted is the popularity of the player from the above analysis[2]. We all know that the Ad agencies looking for celebrities to increase their sales in today's market. This dataset can help us to analyze the game and wining team players by which we can list out some of the players that can be referred to the Ad agencies to increase their sales or to franchise their brand across the globe.

DESCRIPTION OF DATASET:

The Dataset we used is collected from Kaggle[1]. The Dataset had seven different csv files from which we created seven different tables. The csv files we used to convert into tables are Country, League, Match, Player, Player_Attributes, Team, Team_Attributes.

Country Table:

Column Name	**Description**	**Type**
id	Unique id number for country	Number
name	Name of the Country	Plain Text

League Table:

Column Name	**Description**	**Type**
id	Unique id number for League	Number
Country_id	Unique id number for country	Number
name	Name of the League	Plain Text

Match Table:

Column Name	**Description**	**Type**
id	Unique id number for Match	Number
Country_id	Unique id number for country	Number
League_id	Unique id number for League	Number
Season	Match Year	Year
date	Date of Match	Date
match_api_id	Unique id number for match	Number
home_team_api_id	Home Team API Id	Number
away_team_api_id	Away Team API Id	Number
Home_team_goal	Home Team Goal	Number

Player Table:

Column Name	**Description**	**Type**
id	Unique id number for Player	Number
Player_api_id	Name of the Player	Plain Text
Player_Name	Unique id number for League	Number
player_fifa_api_id	Match Year	Year
birthday	Date of Match	Date
height	Height of the Player	Number
Weight	Weight of the Player	Number

Player_Attributes Table:

Column Name	Description	Type
Player_api_id	Name of the Player	Plain Text
player_fifa_api_id	Match Year	Year
Overall_rating	Player rating	Number
potential	Height of the Player	Number
Preferred_foot	Preferred Player foot	Plain Text
defensive_work_rate	The defensive work Rate	Plain Text
Crossing	Crossing	Number

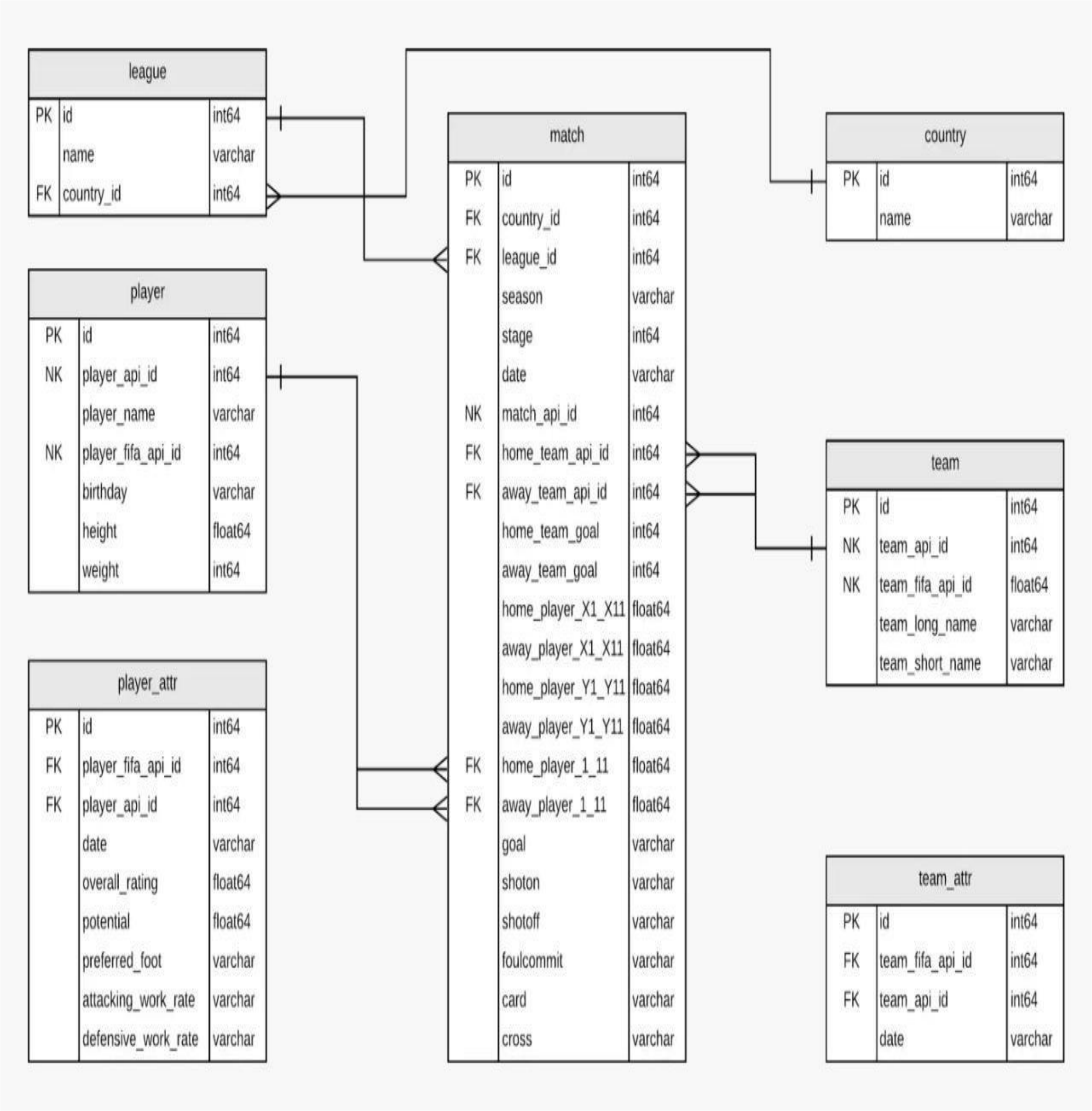
Team Table:

Column Name	Description	Type
team_api_id	Unique id number for team	Number
team_fifa_api_id	Match Year	Number
team_long_name	Team long name	Plain Text
team_short_name	Team short name	Plain Text

Team_Attributes Table:

Column Name	Description	Type
team_api_id	Unique id number for team	Number
team_fifa_api_id	Match Year	Number
team_long_name	Team long name	Plain Text
team_short_name	Team short name	Plain Text
date	Date	Date
buildUpPlaySpeed	Speed of play	Number
buildUpPlaySpeedClass	Class of play speed	Plain Text
buildUpPlayDribbling	Dribbling Speed	Number
buildUpPlayDribblingClass	Class of Dribbling speed	Plain Text
buildUpPlayPassing	Play pass	Number
buildUpPlayPassingClass	Class of Play Pass	Number

ER DIAGRAM:



league		
PK	id	int64
	name	varchar
FK	country_id	int64

player		
PK	id	int64
NK	player_api_id	int64
	player_name	varchar
NK	player_fifa_api_id	int64
	birthday	varchar
	height	float64
	weight	int64

player_attr		
PK	id	int64
FK	player_fifa_api_id	int64
FK	player_api_id	int64
	date	varchar
	overall_rating	float64
	potential	float64
	preferred_foot	varchar
	attacking_work_rate	varchar
	defensive_work_rate	varchar

match		
PK	id	int64
FK	country_id	int64
FK	league_id	int64
	season	varchar
	stage	int64
	date	varchar
NK	match_api_id	int64
FK	home_team_api_id	int64
FK	away_team_api_id	int64
	home_team_goal	int64
	away_team_goal	int64
	home_player_X1_X11	float64
	away_player_X1_X11	float64
	home_player_Y1_Y11	float64
	away_player_Y1_Y11	float64
FK	home_player_1_11	float64
FK	away_player_1_11	float64
	goal	varchar
	shoton	varchar
	shotoff	varchar
	foulcommit	varchar
	card	varchar
	cross	varchar

country		
PK	id	int64
	name	varchar

team		
PK	id	int64
NK	team_api_id	int64
NK	team_fifa_api_id	float64
	team_long_name	varchar
	team_short_name	varchar

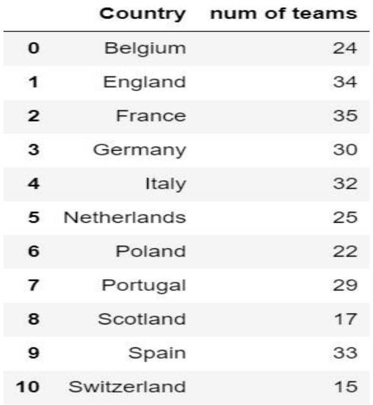
team_attr		
PK	id	int64
FK	team_fifa_api_id	int64
FK	team_api_id	int64
	date	varchar

METHODOLOGY

After clear observation of the data, we observed that the data is clean and the data is in integerformat. The following analysis is done by SQL queries:

Number Teams Each Country Has:

```
q1 = pd.read_sql_query(
    '''
        SELECT
              ct.name AS Country,
              COUNT(DISTINCT(team_long_name)) AS 'num of teams'
              FROM Match AS ma
              LEFT JOIN Country AS ct
              ON ma.country_id = ct.id
              LEFT JOIN Team AS t
              ON ma.home_team_api_id = t.team_api_id
              GROUP BY Country
    ''', conn
)
```

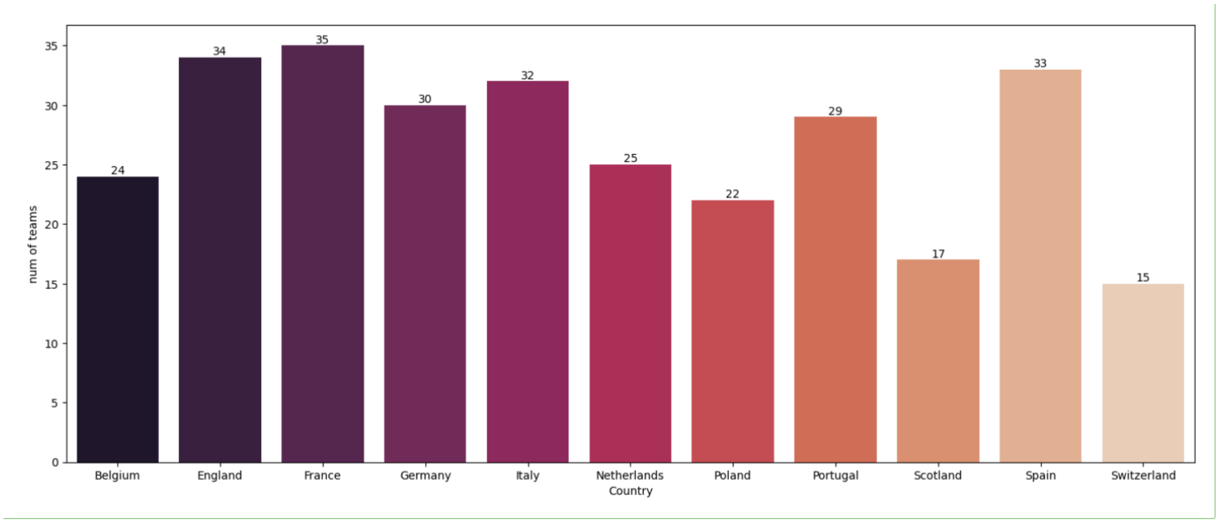


	Country	num of teams
0	Belgium	24
1	England	34
2	France	35
3	Germany	30
4	Italy	32
5	Netherlands	25
6	Poland	22
7	Portugal	29
8	Scotland	17
9	Spain	33
10	Switzerland	15

plt.figure(figsize = (18,7))

ax=sns.barplot(x = 'Country', y = 'num of teams', data = q1, palette='rocket');

ax.bar_label(ax.containers[0])



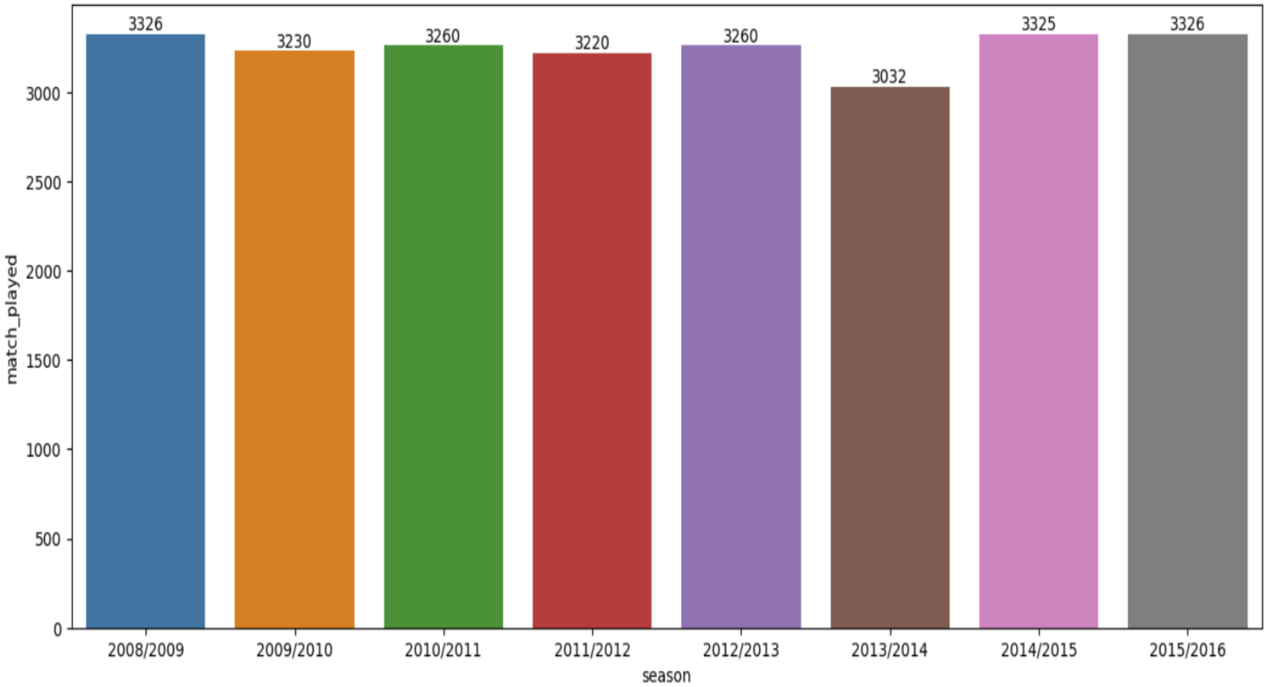
Matches Played Per Season:

matchperseason = pd.read_sql_query("SELECT season, count(*) match_played from match group by season", conn)

plt.figure(figsize = (15,6))

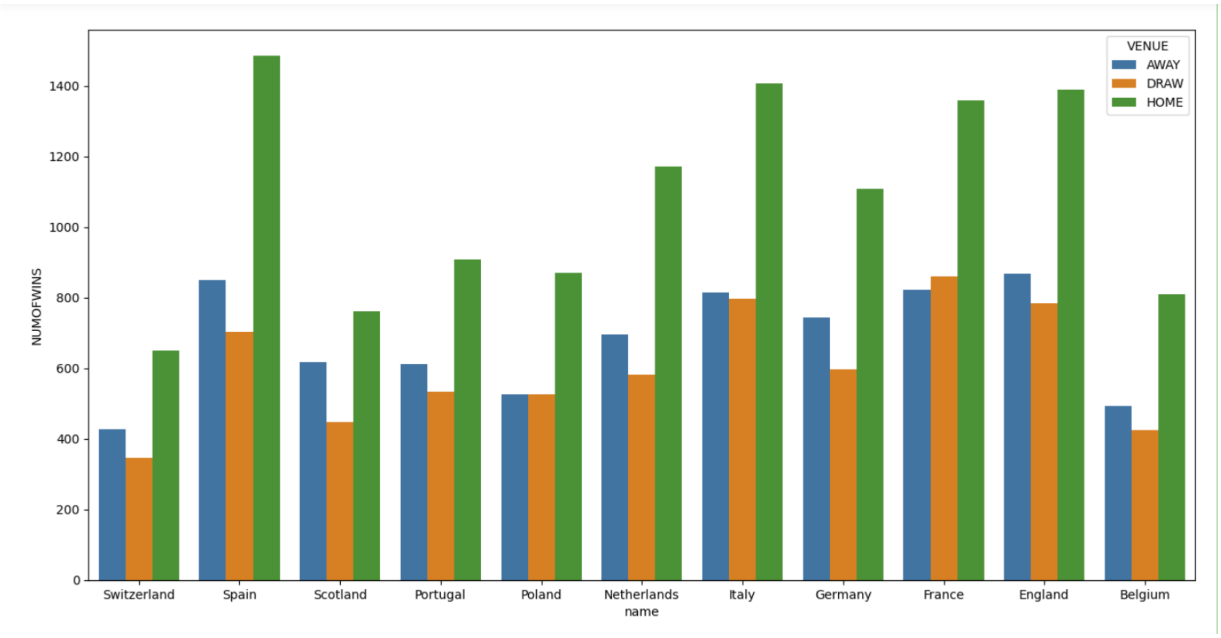
ax = sns.barplot(x = 'season', y = 'match_played', data = matchperseason)

ax.bar_label(ax.containers[0])



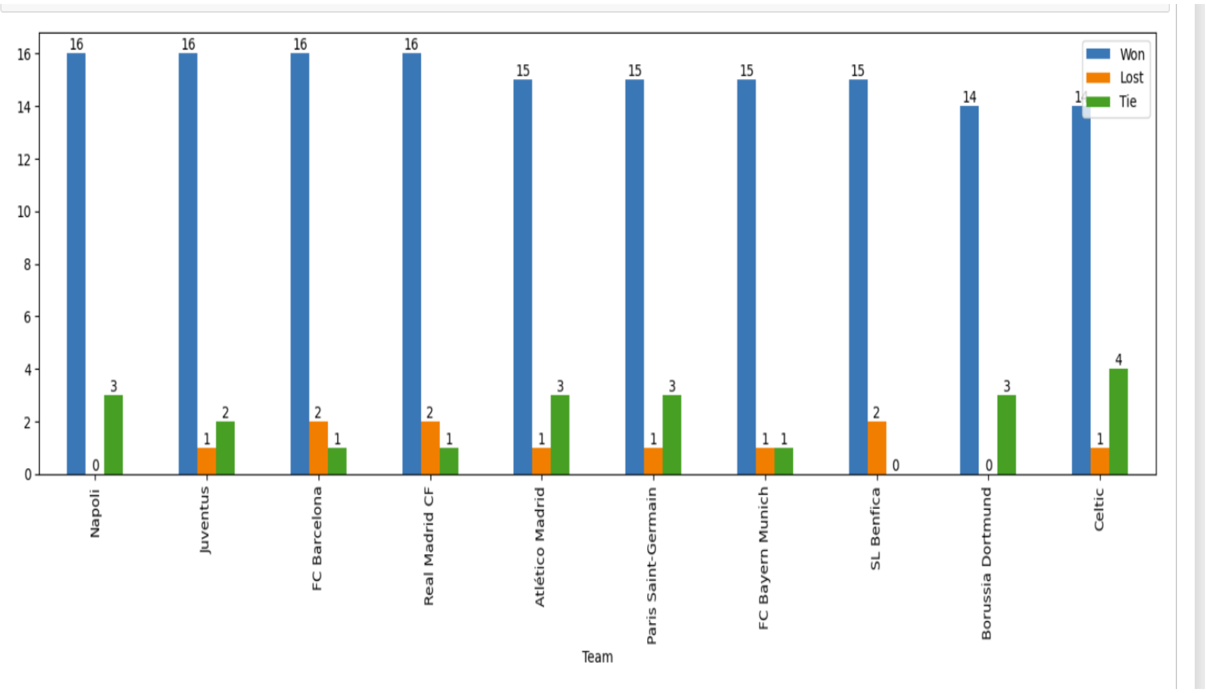
Performance Of The Teams On Home Ground vs Away From Home:

```
q_hwd = pd.read_sql_query("SELECT ct.name, case when ma.home_team_goal > ma.away_team_goal Then'HOME'\
                    when ma.home_team_goal < ma.away_team_goal Then 'AWAY'\
               else 'DRAW' end as VENUE,\
               count(*) as NUMOFWINS\
            FROM\
               Match as ma\
               join Country as ct\
               on ct.id = ma.country_id\
            group by ct.name, VENUE\
            order by ct.name desc", conn)


plt.figure(figsize = (16,8))
sns.barplot(x = 'name', y = 'NUMOFWINS', data = q_hwd, hue = 'VENUE' )
```



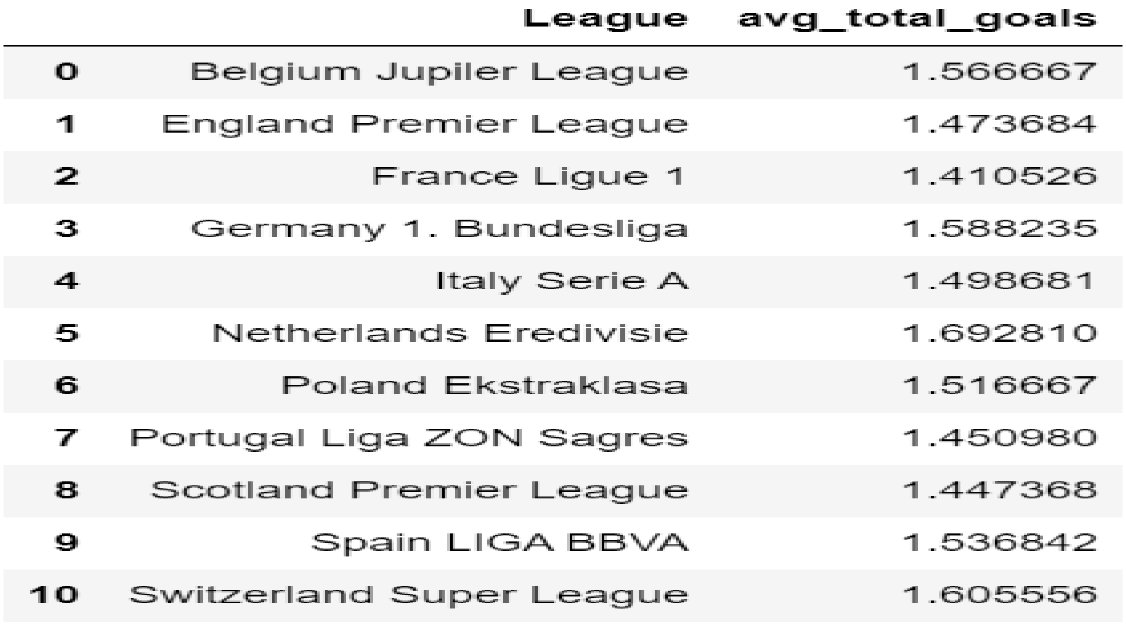
Count Of Matches Won, Lost, And Tie:

```
ax = q8.plot.bar(x= 'Team', y = ['Won','Lost','Tie'],figsize = (17,5))

for container in ax.containers:

    ax.bar_label(container)
```



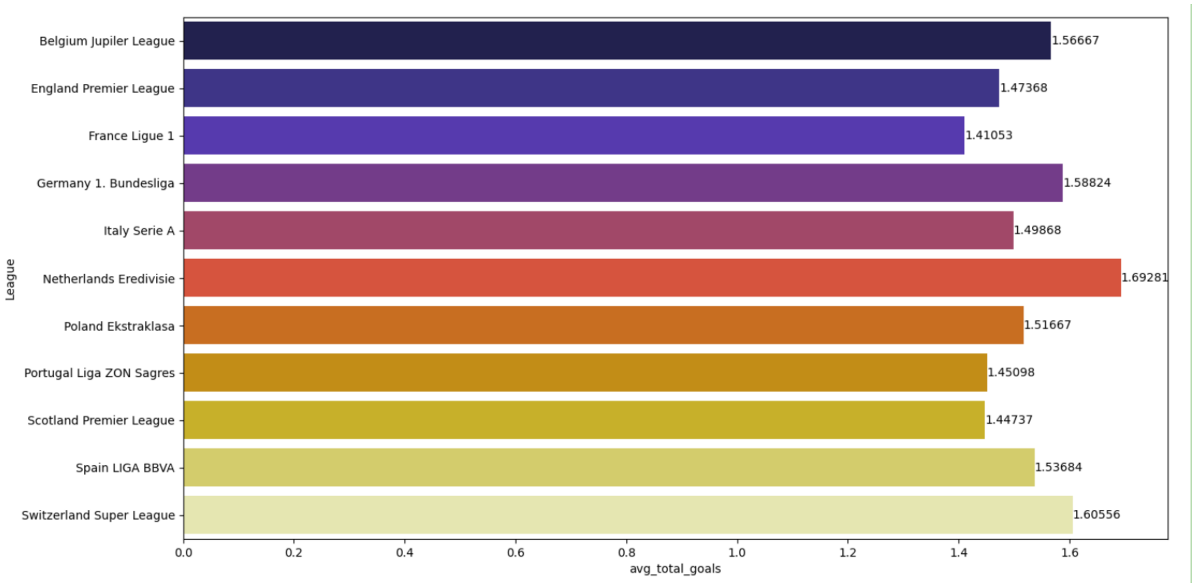
Finding Average Total Goals For Each League:

```
q10 = pd.read_sql_query(

    '''

              SELECT

                   l.name AS League,

                   avg(ma.home_team_goal) AS avg_total_goals

         FROM Match AS ma

         LEFT JOIN League AS l

         ON ma.country_id = l.id

         where ma.season = '2014/2015'

         GROUP BY League

         ORDER BY League
''', conn
   )
q10
```



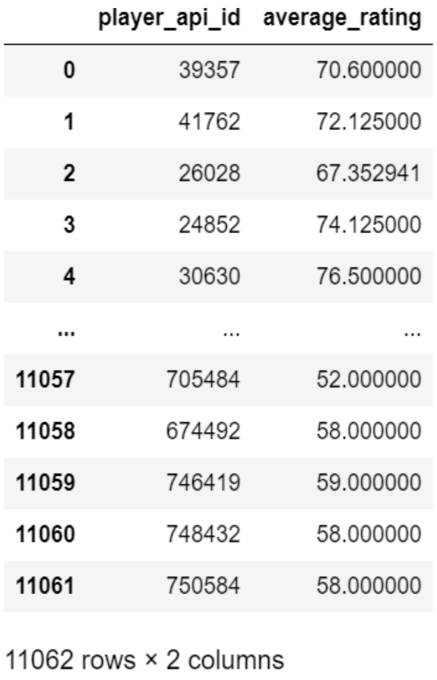
	League	avg_total_goals
0	Belgium Jupiler League	1.566667
1	England Premier League	1.473684
2	France Ligue 1	1.410526
3	Germany 1. Bundesliga	1.588235
4	Italy Serie A	1.498681
5	Netherlands Eredivisie	1.692810
6	Poland Ekstraklasa	1.516667
7	Portugal Liga ZON Sagres	1.450980
8	Scotland Premier League	1.447368
9	Spain LIGA BBVA	1.536842
10	Switzerland Super League	1.605556

```
plt.figure(figsize = (15,8))
ax=sns.barplot(x = 'avg_total_goals', y = 'League', data = q10, palette='CMRmap');
ax.bar_label(ax.containers[0])
```



Average Rating Of Each Player:

```
query11  = pd.read_sql_query(

    ''' SELECT player_api_id, avg(overall_rating) as average_rating from player_attributes group by player_fifa_api_id
    ''', conn)

query11
```

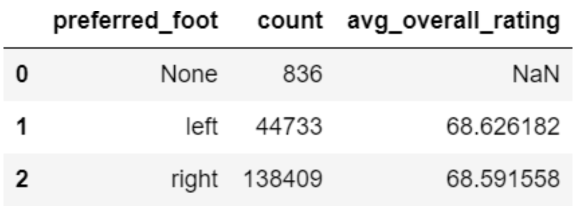


	player_api_id	average_rating
0	39357	70.600000
1	41762	72.125000
2	26028	67.352941
3	24852	74.125000
4	30630	76.500000
...	...	...
11057	705484	52.000000
11058	674492	58.000000
11059	746419	59.000000
11060	748432	58.000000
11061	750584	58.000000

11062 rows × 2 columns

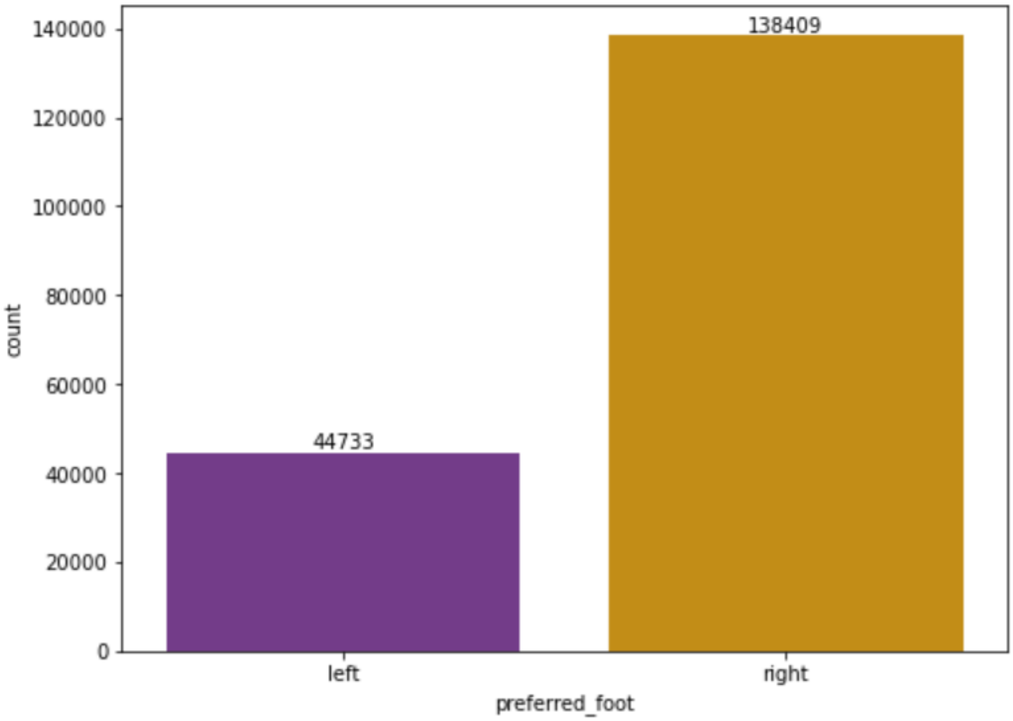
Finding Count Of Players With Different Preferred Foot And Average Of Overall Rating Of Those Players:

```
query_11  = pd.read_sql_query(

    ''' SELECT preferred_foot,count(player_api_id) as count , avg(overall_rating) as avg_overall_rating from player_attributes group by preferred_foot

    ''', conn)

query_11
```



	preferred_foot	count	avg_overall_rating
0	None	836	NaN
1	left	44733	68.626182
2	right	138409	68.591558

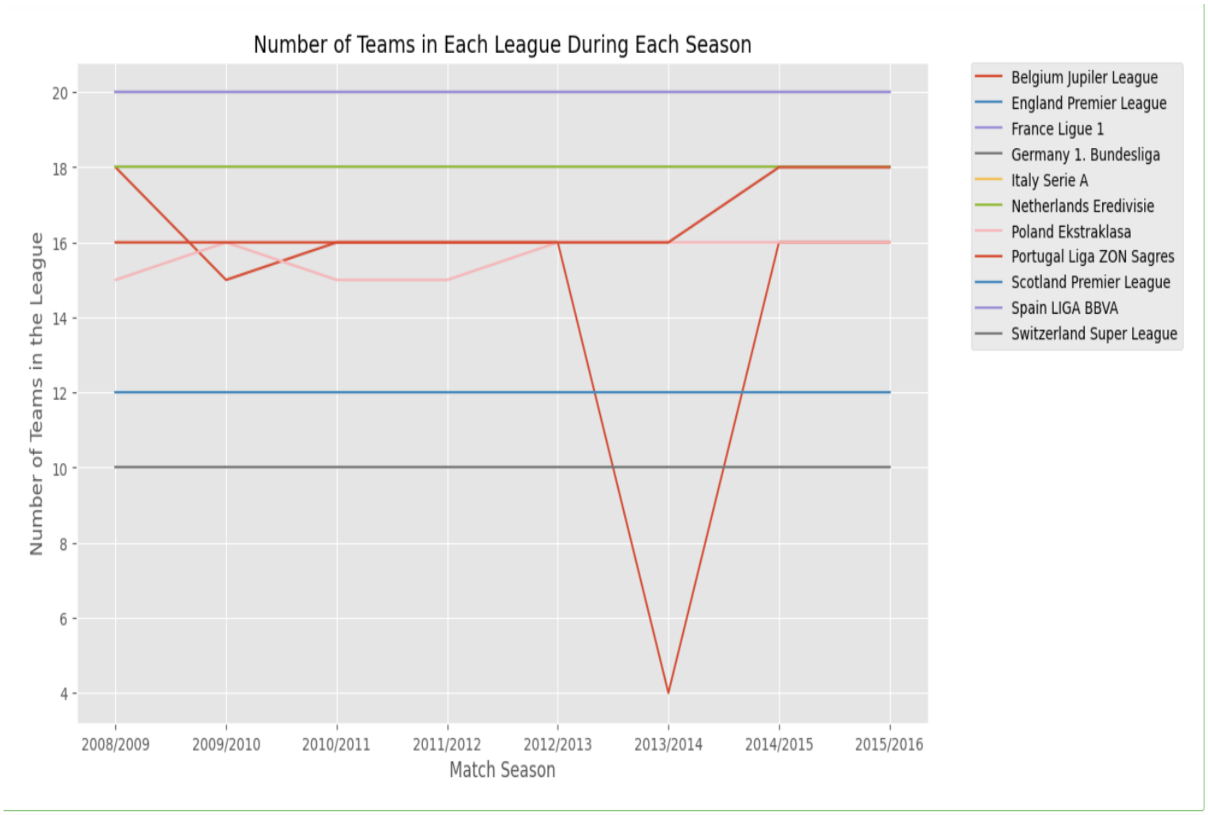
```
plt.figure(figsize = (8,6))ax=sns.barplot(x = 'preferred_foot', y = 'count', data = query10, palette='CMRmap');ax.bar_label(ax.containers[0])
```



Plotting How Number Of Home Teams Varies For Each Leaue As Season Changes:

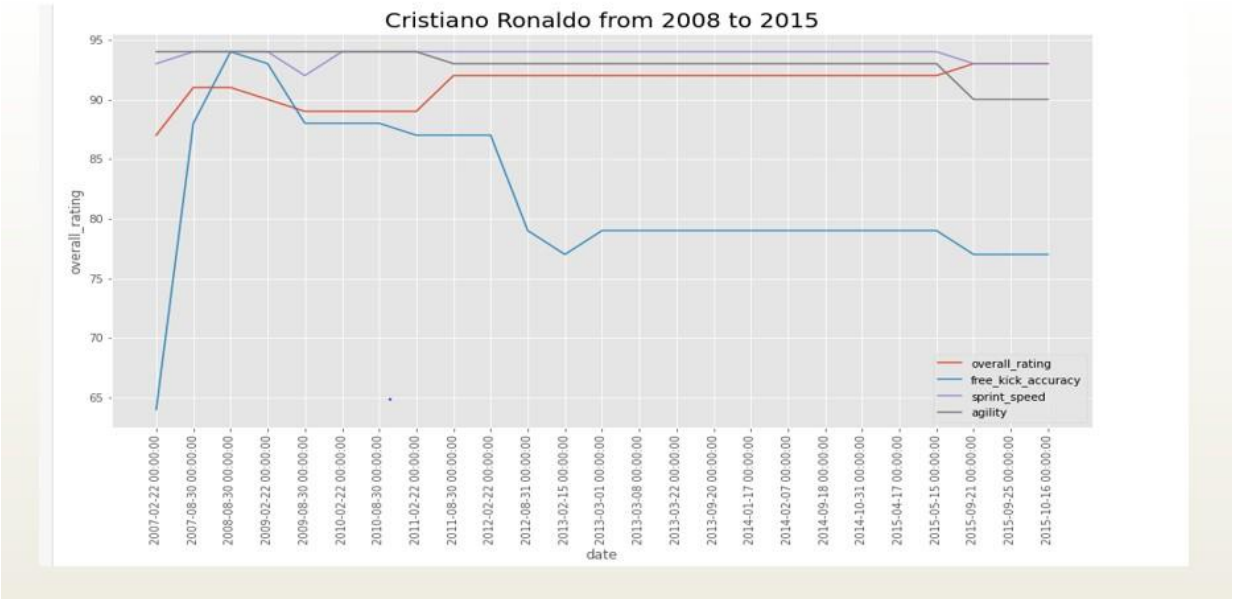
```
import numpy as np

seasons = np.array(['2008/2009', '2009/2010', '2010/2011', '2011/2012', '2012/2013',
       '2013/2014', '2014/2015', '2015/2016'], dtype=object)
seasons
import matplotlib.pyplot as plt
%matplotlib inline
#set ggplot style
plt.style.use('ggplot')
# plot data
fig, ax = plt.subplots(figsize=(12,7))
# use unstack()
Count_League_Team_Season.groupby(['match_Season','League_Name']).sum()['Number_of_Home_Teams'].unstack().plot(ax=ax)
ax.set_xlabel('Match Season')
ax.set_ylabel('Number of Teams in the League')
```

```
plt.xticks(range(len(seasons)),seasons)

plt.legend(bbox_to_anchor=(1.05, 1), loc=2, borderaxespad=0.)

plt.title('Number of Teams in Each League During Each Season')
```



To Vizualize The Player Attribute Of One Of Best Players Cristiano_Ronaldo:

```
plt.figure(figsize=(15, 7))

sns.lineplot( x = Cristiano_Ronaldo['date'], y = Cristiano_Ronaldo["overall_rating"], palette = 'Wistia', label="overall_rating")

sns.lineplot(x = Cristiano_Ronaldo['date'], y = Cristiano_Ronaldo["free_kick_accuracy"], palette = 'Wistia', label="free_kick_accuracy")

sns.lineplot(x = Cristiano_Ronaldo['date'], y = Cristiano_Ronaldo["sprint_speed"], palette = 'Wistia', label="sprint_speed")

sns.lineplot(x = Cristiano_Ronaldo['date'], y = Cristiano_Ronaldo["agility"], palette = 'Wistia', label="agility")

plt.tick_params(axis='x', rotation=90)

plt.title("Cristiano Ronaldo from 2008 to 2015", fontsize=20)
```



Top 5 Players And Their Rating:

top_players = pd.read_sql("""SELECT play.player_name, pl.player_api_id, ROUND(AVG(pl.overall_rating),2) AS avgRating

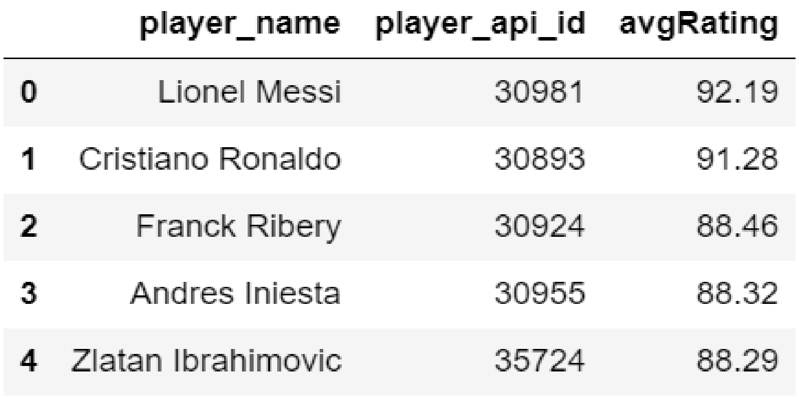
FROM Player play, Player_Attributes pl

WHERE play.player_api_id = pl.player_api_id

GROUP BY play.player_api_id

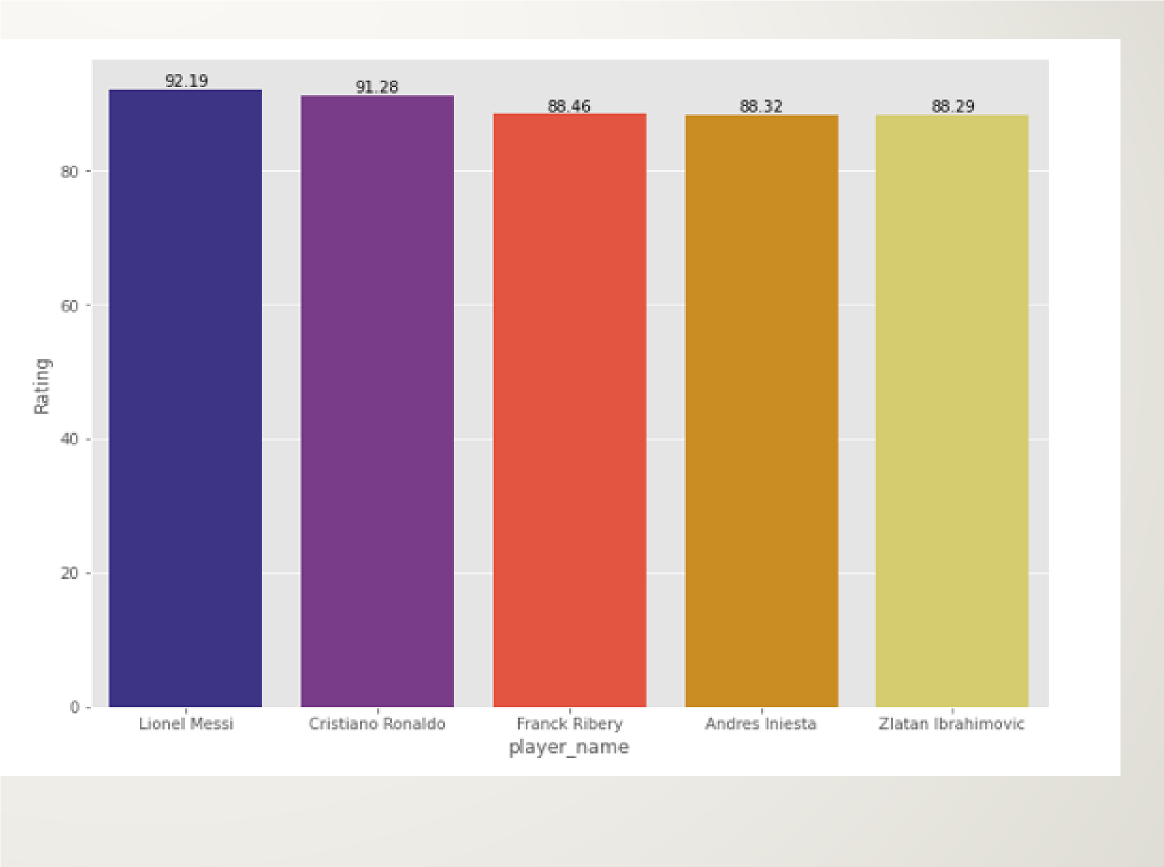
ORDER BY 3 DESC LIMIT 5;""",conn)

top_players



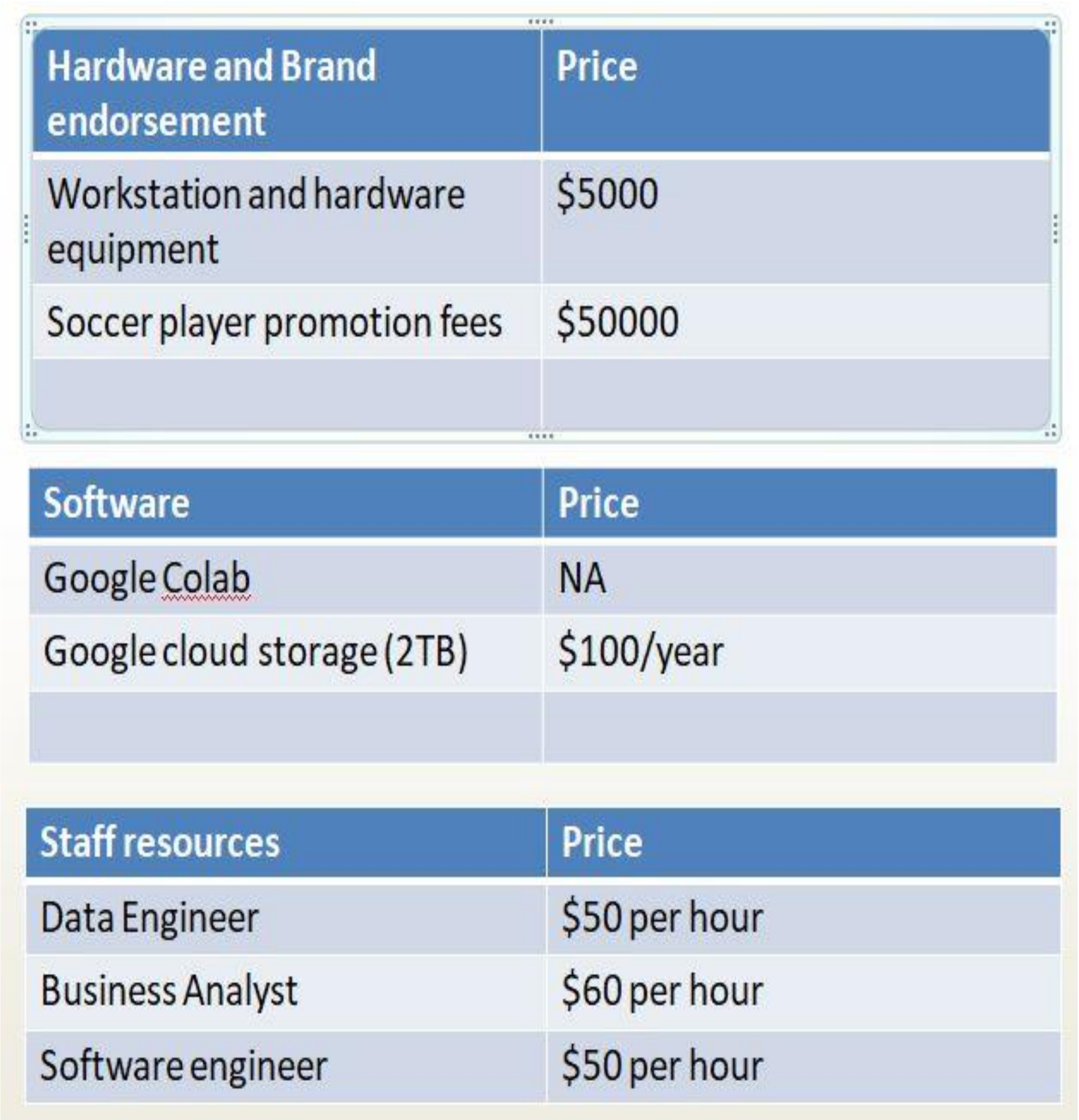
	player_name	player_api_id	avgRating
0	Lionel Messi	30981	92.19
1	Cristiano Ronaldo	30893	91.28
2	Franck Ribery	30924	88.46
3	Andres Iniesta	30955	88.32
4	Zlatan Ibrahimovic	35724	88.29

```
plt.figure(figsize = (11,8))

ax=sns.barplot(x = 'player_name', y = 'avgRating', data = top_players, palette='CMRmap');

ax.bar_label(ax.containers[0])
```



EXPENSES ESTIMATE

We have estimated the price to install the equipment for analyzing the data to be used for target marketing to generate profits. The estimates are listed as follows.



Hardware and Brand endorsement	Price
Workstation and hardware equipment	$5000
Soccer player promotion fees	$50000

Software	Price
Google Colab	NA
Google cloud storage (2TB)	$100/year

Staff resources	Price
Data Engineer	$50 per hour
Business Analyst	$60 per hour
Software engineer	$50 per hour

s

RESULTS:

- Data analysis of the soccer database helped us find different solutions toimprove ad revenue.
- Most no.of goals i.e,262 are scored by Belgium in 2008.
- According to the analysis right footed players show better performance.
- Real Madrid CF stands as the best team.
- Most matches were conducted in the 2015-2016 season.
- Teams won more no of matches at home country fields.
- Lionel Messi stands as the top player.
- Another interesting insight is that number of teams in all leagues are constant during different seasons except for “Belgium Jupiler League” which seems to have fluctuating teams during different seasons.
- Players with height between 165 and 170 have got higher rating when compared to players out of this range

DISCUSSION AND FUTURE WORK:

From the detailed analysis of the data, we found a few points that are interesting. The interesting part is that the performance of the team depends on many factors like where the match is being conducted, the height and other attributes of the player, etc.

Using the following analysis we can choose the most popular or well-performing player to target the audience for the required brands. For ex: As the agility of the player is one of the main factors that influence the performance of the player, shoes can be advertised using the player with higher agility. This data not only can be used for ad revenue but also for the improvement of teams to them towards winning.

In future, we can derive the important features by applying correlation process and doing feature selection and elimination process , and then feeding them to linear regression model to determine the overall rating of the playe

REFERENCES:

[1] Abdelrhman Raga, 2021: European Soccer Database
https://www.kaggle.com/datasets/abdelrhmanragab/european-soccer-database

[2] Yiou Wang , 2020: Data Analyses of European Soccer
https://www.researchgate.net/publication/343947808_Data_Analyses_of_European_Soccer

[3] Seanyeon, 2021: Soccer Data analysis

https://www.seanyeon.com/post/soccer-data-analysis

[4] Ahmed Mohamed AbdElHameed, 2021: Soccer Data analysis
https://am2958.medium.com/soccer-data-analysis-e911eccc8369