PMP LAB PROJECT

NAME- ASAD JAWAID

SID- 2246063

Submitted to :- Mr.VIVEK PANDEY Sir

Content

- Visualization
  - Most Number of World Cup Winning Title
  - Number of Goal Per Country
  - Attendance, Number of Teams, Goals, and Matches per Cup
  - Goals Per Team Per World Cup
  - Matches With Heighest Number Of Attendance
  - Stadium with Highest Average Attendance
  - Which countries had won the cup ?
  - Number of goal per country
  - Match outcome by home and away temas

In [1]:
```
#importing libraries

import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
import matplotlib.pyplot as plt
import seaborn as sns
```

In [3]:
```
#Readind the dataset

matches=pd.read_csv("C:\\Users\\farid\\Downloads\\archive\\WorldCupMatches.
players=pd.read_csv("C:\\Users\\farid\\Downloads\\archive\\WorldCupPlayers.
world_cup=pd.read_csv("C:\\Users\\farid\\Downloads\\archive\\WorldCups.csv"
```

In [5]:
```
players.head()
```

Out[5]:

	RoundID	MatchID	Team Initials	Coach Name	Line-up	Shirt Number	Player Name	Position	Event
0	201	1096	FRA	CAUDRON Raoul (FRA)	S	0	Alex THEPOT	GK	NaN
1	201	1096	MEX	LUQUE Juan (MEX)	S	0	Oscar BONFIGLIO	GK	NaN
2	201	1096	FRA	CAUDRON Raoul (FRA)	S	0	Marcel LANGILLER	NaN	G40'
3	201	1096	MEX	LUQUE Juan (MEX)	S	0	Juan CARRENO	NaN	G70'
4	201	1096	FRA	CAUDRON Raoul (FRA)	S	0	Ernest LIBERATI	NaN	NaN

In [6]:
```
matches.head()
```

Out[6]:

	Year	Datetime	Stage	Stadium	City	Home Team Name	Home Team Goals	Away Team Goals	Away Team Name	W conditic
0	1930.0	13 Jul 1930 - 15:00	Group 1	Pocitos	Montevideo	France	4.0	1.0	Mexico	
1	1930.0	13 Jul 1930 - 15:00	Group 4	Parque Central	Montevideo	USA	3.0	0.0	Belgium	
2	1930.0	14 Jul 1930 - 12:45	Group 2	Parque Central	Montevideo	Yugoslavia	2.0	1.0	Brazil	
3	1930.0	14 Jul 1930 - 14:50	Group 3	Pocitos	Montevideo	Romania	3.0	1.0	Peru	
4	1930.0	15 Jul 1930 - 16:00	Group 1	Parque Central	Montevideo	Argentina	1.0	0.0	France	

In [7]:
```
world_cup.head()
```

Out[7]:

	Year	Country	Winner	Runners-Up	Third	Fourth	GoalsScored	QualifiedTe
0	1930	Uruguay	Uruguay	Argentina	USA	Yugoslavia	70	
1	1934	Italy	Italy	Czechoslovakia	Germany	Austria	70	
2	1938	France	Italy	Hungary	Brazil	Sweden	84	
3	1950	Brazil	Uruguay	Brazil	Sweden	Spain	88	
4	1954	Switzerland	Germany FR	Hungary	Austria	Uruguay	140	

In [10]:
```
#finding information of the Dataset

players.info
```

Out[10]:
```
<bound method DataFrame.info of        RoundID    MatchID Team Initials
Coach Name Line-up  \
0          201       1096           FRA      CAUDRON Raoul (FRA)       S
1          201       1096           MEX         LUQUE Juan (MEX)       S
2          201       1096           FRA      CAUDRON Raoul (FRA)       S
3          201       1096           MEX         LUQUE Juan (MEX)       S
4          201       1096           FRA      CAUDRON Raoul (FRA)       S
...        ...        ...           ...                       ...     ...
37779   255959  300186501           ARG  SABELLA Alejandro (ARG)       N
37780   255959  300186501           GER       LOEW Joachim (GER)       N
37781   255959  300186501           ARG  SABELLA Alejandro (ARG)       N
37782   255959  300186501           GER       LOEW Joachim (GER)       N
37783   255959  300186501           ARG  SABELLA Alejandro (ARG)       N

       Shirt Number       Player Name Position       Event
0                 0       Alex THEPOT       GK         NaN
1                 0   Oscar BONFIGLIO       GK         NaN
2                 0  Marcel LANGILLER      NaN        G40'
3                 0      Juan CARRENO      NaN        G70'
4                 0   Ernest LIBERATI      NaN         NaN
...             ...               ...      ...         ...
37779            19           ALVAREZ      NaN         NaN
37780             6           KHEDIRA      NaN         NaN
37781            20            AGUERO      NaN  IH46' Y65'
37782            21           MUSTAFI      NaN         NaN
37783            23           BASANTA      NaN         NaN

[37784 rows x 9 columns]>
```

In [11]:
```
#Checking the shape of dataset

players.shape
```

Out[11]:
```
(37784, 9)
```

In [13]:
```
#Describing the dataset

players.describe()
```

Out[13]:

	RoundID	MatchID	Shirt Number
count	3.778400e+04	3.778400e+04	37784.000000
mean	1.105647e+07	6.362233e+07	10.726022
std	2.770144e+07	1.123916e+08	6.960138
min	2.010000e+02	2.500000e+01	0.000000
25%	2.630000e+02	1.199000e+03	5.000000
50%	3.370000e+02	2.216000e+03	11.000000
75%	2.559310e+05	9.741000e+07	17.000000
max	9.741060e+07	3.001865e+08	23.000000

In [14]:
```
#Checking the NULL values

players.isnull()
```

Out[14]:

	RoundID	MatchID	Team Initials	Coach Name	Line-up	Shirt Number	Player Name	Position	Event
0	False	False	False	False	False	False	False	False	True
1	False	False	False	False	False	False	False	False	True
2	False	False	False	False	False	False	False	True	False
3	False	False	False	False	False	False	False	True	False
4	False	False	False	False	False	False	False	True	True
...	...	...	...	...	...	...	...	...	...
37779	False	False	False	False	False	False	False	True	True
37780	False	False	False	False	False	False	False	True	True
37781	False	False	False	False	False	False	False	True	False
37782	False	False	False	False	False	False	False	True	True
37783	False	False	False	False	False	False	False	True	True

37784 rows × 9 columns

In [16]:
```
#Checking the data types in datasets

players.dtypes
```

Out[16]:
```
RoundID           int64
MatchID           int64
Team Initials    object
Coach Name       object
Line-up          object
Shirt Number      int64
Player Name      object
Position         object
Event            object
dtype: object
```

In [17]:
```
matches.dtypes
```

Out[17]:
```
Year                    float64
Datetime                 object
Stage                    object
Stadium                  object
City                     object
Home Team Name           object
Home Team Goals         float64
Away Team Goals         float64
Away Team Name           object
Win conditions           object
Attendance              float64
Half-time Home Goals    float64
Half-time Away Goals    float64
Referee                  object
Assistant 1              object
Assistant 2              object
RoundID                 float64
MatchID                 float64
Home Team Initials       object
Away Team Initials       object
dtype: object
```

In [18]:
```
world_cup.dtypes
```

Out[18]:
```
Year               int64
Country           object
Winner            object
Runners-Up        object
Third             object
Fourth            object
GoalsScored        int64
QualifiedTeams     int64
MatchesPlayed      int64
Attendance        object
dtype: object
```

Most Number of World Cup Winning Title

In [15]:
```
winner = world_cup['Winner'].value_counts()
winner
```

Out[15]:
```
Brazil        5
Italy         4
Germany FR    3
Uruguay       2
Argentina     2
England       1
France        1
Spain         1
Germany       1
Name: Winner, dtype: int64
```

In [21]:
```
runnerup = world_cup['Runners-Up'].value_counts()
runnerup
```

Out[21]:
```
Argentina         3
Germany FR        3
Netherlands       3
Czechoslovakia    2
Hungary           2
Brazil            2
Italy             2
Sweden            1
Germany           1
France            1
Name: Runners-Up, dtype: int64
```

In [22]:
```
third = world_cup['Third'].value_counts()
third
```

Out[22]:
```
Germany        3
Brazil         2
Sweden         2
France         2
Poland         2
USA            1
Austria        1
Chile          1
Portugal       1
Germany FR     1
Italy          1
Croatia        1
Turkey         1
Netherlands    1
Name: Third, dtype: int64
```

In [23]:
```
teams = pd.concat([winner, runnerup, third], axis=1)
teams.fillna(0, inplace=True)
teams = teams.astype(int)
teams
```

Out[23]:

	Winner	Runners-Up	Third
Brazil	5	2	2
Italy	4	2	1
Germany FR	3	3	1
Uruguay	2	0	0
Argentina	2	3	0
England	1	0	0
France	1	1	2
Spain	1	0	0
Germany	1	1	3
Netherlands	0	3	1
Czechoslovakia	0	2	0
Hungary	0	2	0
Sweden	0	1	2
Poland	0	0	2
USA	0	0	1
Austria	0	0	1
Chile	0	0	1
Portugal	0	0	1
Croatia	0	0	1
Turkey	0	0	1

Number of Goal Per Country

In [25]:
```
matches.head(2)
```

Out[25]:

	Year	Datetime	Stage	Stadium	City	Home Team Name	Home Team Goals	Away Team Goals	Away Team Name	Win conditions
0	1930.0	13 Jul 1930 - 15:00	Group 1	Pocitos	Montevideo	France	4.0	1.0	Mexico	
1	1930.0	13 Jul 1930 - 15:00	Group 4	Parque Central	Montevideo	USA	3.0	0.0	Belgium	

In [27]:
```
home = matches[['Home Team Name', 'Home Team Goals']].dropna()
away = matches[['Away Team Name', 'Away Team Goals']].dropna()
```

In [28]:
```
home.columns = ['Countries', 'Goals']
away.columns = home.columns
```

In [29]:
```
goals = home.append(away, ignore_index = True)
```

```
C:\Users\farid\AppData\Local\Temp\ipykernel_16332\2748964524.py:1: Future
Warning: The frame.append method is deprecated and will be removed from p
andas in a future version. Use pandas.concat instead.
  goals = home.append(away, ignore_index = True)
```

In [30]:
```
goals = goals.groupby('Countries').sum()
goals
```

Out[30]:

Countries	Goals
Algeria	14.0
Angola	1.0
Argentina	133.0
Australia	11.0
Austria	43.0
...	...
rn">Bosnia and Herzegovina	4.0
rn">Republic of Ireland	10.0
rn">Serbia and Montenegro	2.0
rn">Trinidad and Tobago	0.0
rn">United Arab Emirates	2.0

83 rows × 1 columns

In [31]:
```
goals = goals.sort_values(by = 'Goals', ascending=False)
goals
```

Out[31]:

Countries	Goals
Brazil	225.0
Argentina	133.0
Germany FR	131.0
Italy	128.0
France	108.0
...	...
Dutch East Indies	0.0
China PR	0.0
Canada	0.0
Zaire	0.0
rn">Trinidad and Tobago	0.0

83 rows × 1 columns

Attendance, Number of Teams, Goals, and Matches per Cup

In [32]:
```
world_cup['Attendance'] = world_cup['Attendance'].str.replace(".", "")
```

```
C:\Users\farid\AppData\Local\Temp\ipykernel_16332\902531040.py:1: FutureW
arning: The default value of regex will change from True to False in a fu
ture version. In addition, single character regular expressions will *not
* be treated as literal strings when regex=True.
  world_cup['Attendance'] = world_cup['Attendance'].str.replace(".", "")
```

In [33]:
```
world_cup.head()
```

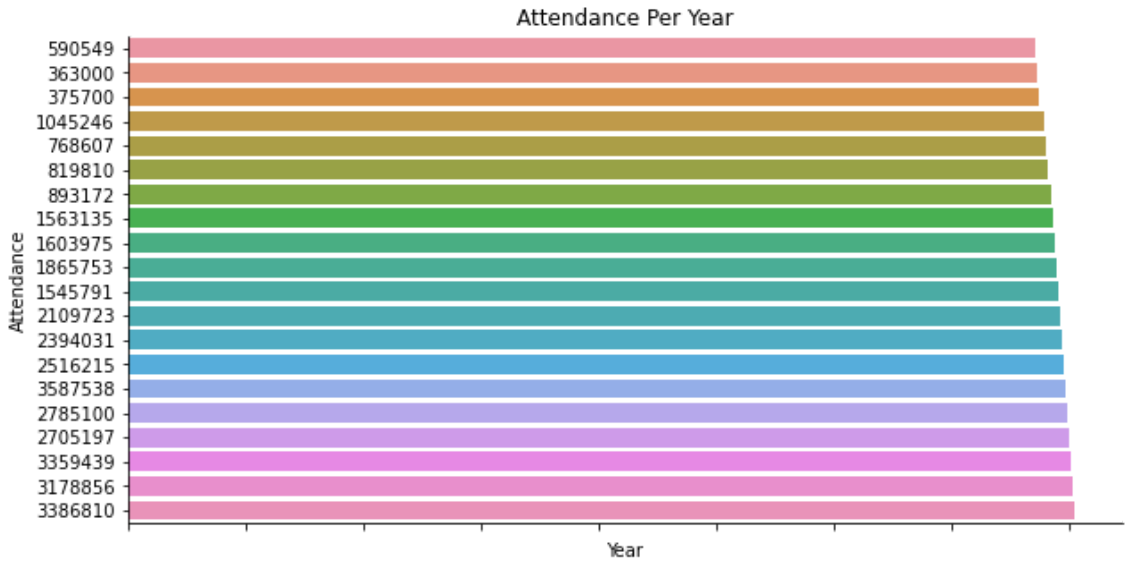
Out[33]:

	Year	Country	Winner	Runners-Up	Third	Fourth	GoalsScored	QualifiedTe
0	1930	Uruguay	Uruguay	Argentina	USA	Yugoslavia	70	
1	1934	Italy	Italy	Czechoslovakia	Germany	Austria	70	
2	1938	France	Italy	Hungary	Brazil	Sweden	84	
3	1950	Brazil	Uruguay	Brazil	Sweden	Spain	88	
4	1954	Switzerland	Germany FR	Hungary	Austria	Uruguay	140	

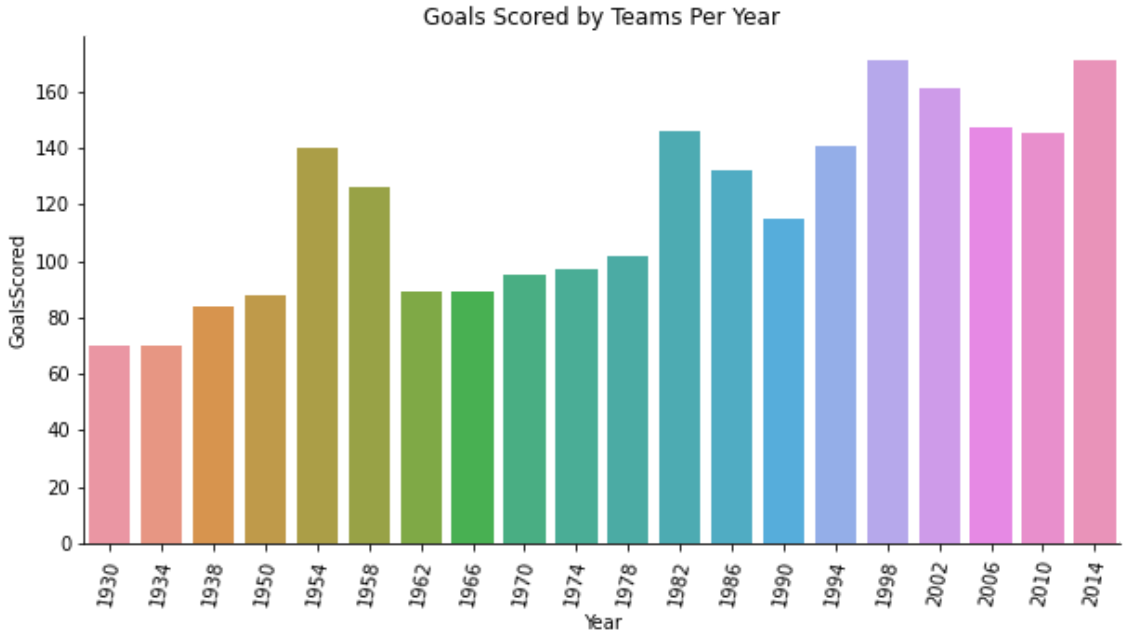
In [34]:
```
fig, ax = plt.subplots(figsize = (10,5))
sns.despine(right = True)
g = sns.barplot(x = 'Year', y = 'Attendance', data = world_cup)
g.set_xticklabels(g.get_xticklabels(), rotation = 80)
g.set_title('Attendance Per Year')

#=========================================

fig, ax = plt.subplots(figsize = (10,5))
sns.despine(right = True)
g = sns.barplot(x = 'Year', y = 'QualifiedTeams', data = world_cup)
g.set_xticklabels(g.get_xticklabels(), rotation = 80)
g.set_title('Qualified Teams Per Year')

#=========================================

fig, ax = plt.subplots(figsize = (10,5))
sns.despine(right = True)
g = sns.barplot(x = 'Year', y = 'GoalsScored', data = world_cup)
g.set_xticklabels(g.get_xticklabels(), rotation = 80)
g.set_title('Goals Scored by Teams Per Year')


#=========================================


fig, ax = plt.subplots(figsize = (10,5))
sns.despine(right = True)
g = sns.barplot(x = 'Year', y = 'MatchesPlayed', data = world_cup)
g.set_xticklabels(g.get_xticklabels(), rotation = 80)
g.set_title('Matches Plyed Scored by Teams Per Year')
```

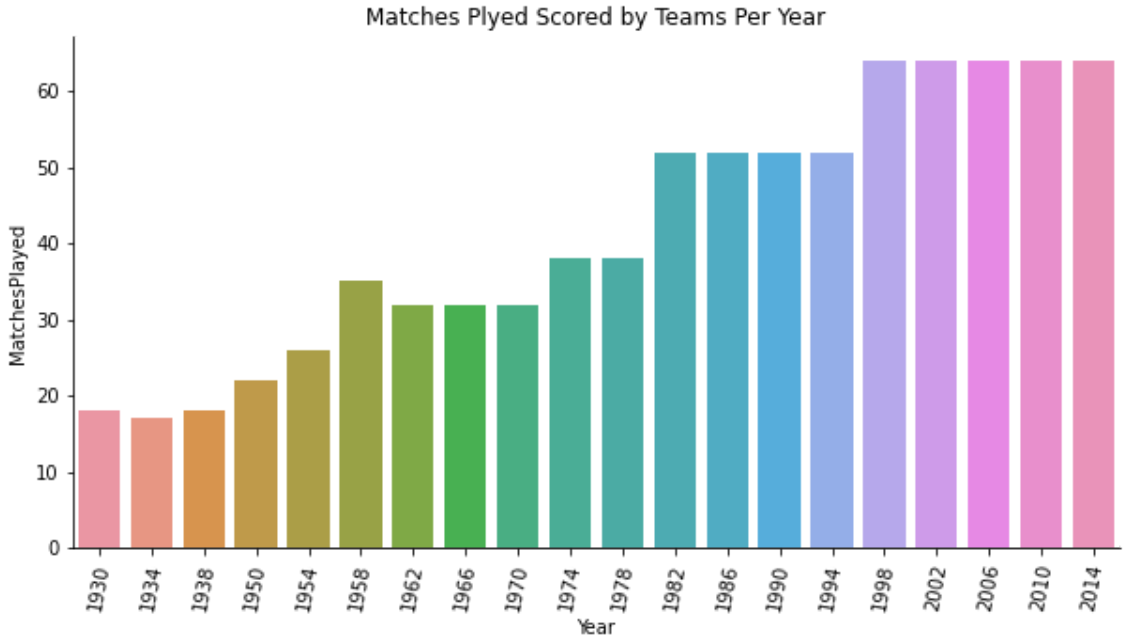
```
C:\Users\farid\AppData\Local\Temp\ipykernel_16332\882942790.py:4: UserWar
ning: FixedFormatter should only be used together with FixedLocator
  g.set_xticklabels(g.get_xticklabels(), rotation = 80)
```

Out[34]:
```
Text(0.5, 1.0, 'Matches Plyed Scored by Teams Per Year')
```




Qualified Teams Per Year
QualifiedTeams
0
5
10
15
20
25
30
1930
1934
1938
1950
1954
1958
1962
1966
1970
1974
1978
1982
1986
1990
1994
1998
2002
2006
2010
2014
Year


Goals Scored by Teams Per Year
GoalsScored
0
20
40
60
80
100
120
140
160
1930
1934
1938
1950
1954
1958
1962
1966
1970
1974
1978
1982
1986
1990
1994
1998
2002
2006
2010
2014
Year


Matches Plyed Scored by Teams Per Year
MatchesPlayed
0
10
20
30
40
50
60
1930
1934
1938
1950
1954
1958
1962
1966
1970
1974
1978
1982
1986
1990
1994
1998
2002
2006
2010
2014
Year

Goals Per Team Per World Cup

In [36]:
```
matches.head(2)
```

Out[36]:

	Year	Datetime	Stage	Stadium	City	Home Team Name	Home Team Goals	Away Team Goals	Away Team Name	Win conditions
0	1930.0	13 Jul 1930 - 15:00	Group 1	Pocitos	Montevideo	France	4.0	1.0	Mexico	
1	1930.0	13 Jul 1930 - 15:00	Group 4	Parque Central	Montevideo	USA	3.0	0.0	Belgium	

In [37]:
```
home = matches.groupby(['Year', 'Home Team Name'])['Home Team Goals'].sum()
home
```

Out[37]:
```
Year    Home Team Name
1930.0  Argentina                     16.0
        Brazil                         4.0
        Chile                          4.0
        France                         4.0
        Paraguay                       1.0
                                      ...
2014.0  Spain                          1.0
        Switzerland                    4.0
        USA                            2.0
        Uruguay                        3.0
        rn">Bosnia and Herzegovina     3.0
Name: Home Team Goals, Length: 366, dtype: float64
```

In [38]:
```
away = matches.groupby(['Year', 'Away Team Name'])['Away Team Goals'].sum()
away
```

Out[38]:
```
Year    Away Team Name
1930.0  Argentina                     2.0
        Belgium                       0.0
        Bolivia                       0.0
        Brazil                        1.0
        Chile                         1.0
                                     ...
2014.0  Spain                         3.0
        Switzerland                   3.0
        USA                           4.0
        Uruguay                       1.0
        rn">Bosnia and Herzegovina    1.0
Name: Away Team Goals, Length: 411, dtype: float64
```

In [39]:
```
goals = pd.concat([home, away], axis=1)
goals.fillna(0, inplace=True)
goals['Goals'] = goals['Home Team Goals'] + goals['Away Team Goals']
goals = goals.drop(labels = ['Home Team Goals', 'Away Team Goals'], axis =
goals
```

Out[39]:

		Goals
Year		
	Argentina	18.0
	Brazil	5.0
1930.0	**Chile**	5.0
	France	4.0
	Paraguay	1.0
...	**...**	...
	Iran	2.0
	Mexico	8.0
1998.0		
	Norway	5.0
	Tunisia	1.0
2006.0	**IR Iran**	0.0

427 rows × 1 columns

In [40]:
```
goals = goals.reset_index()
```

In [41]:
```
goals.columns = ['Year', 'Country', 'Goals']
goals = goals.sort_values(by = ['Year', 'Goals'], ascending = [True, False]
goals
```

Out[41]:

	Year	Country	Goals
0	1930.0	Argentina	18.0
7	1930.0	Uruguay	15.0
6	1930.0	USA	7.0
8	1930.0	Yugoslavia	7.0
1	1930.0	Brazil	5.0
...	...	...	...
354	2014.0	Japan	2.0
360	2014.0	Russia	2.0
339	2014.0	Cameroon	1.0
351	2014.0	Honduras	1.0
352	2014.0	IR Iran	1.0

427 rows × 3 columns

In [42]:
```
top5 = goals.groupby('Year').head()
top5.head(10)
```

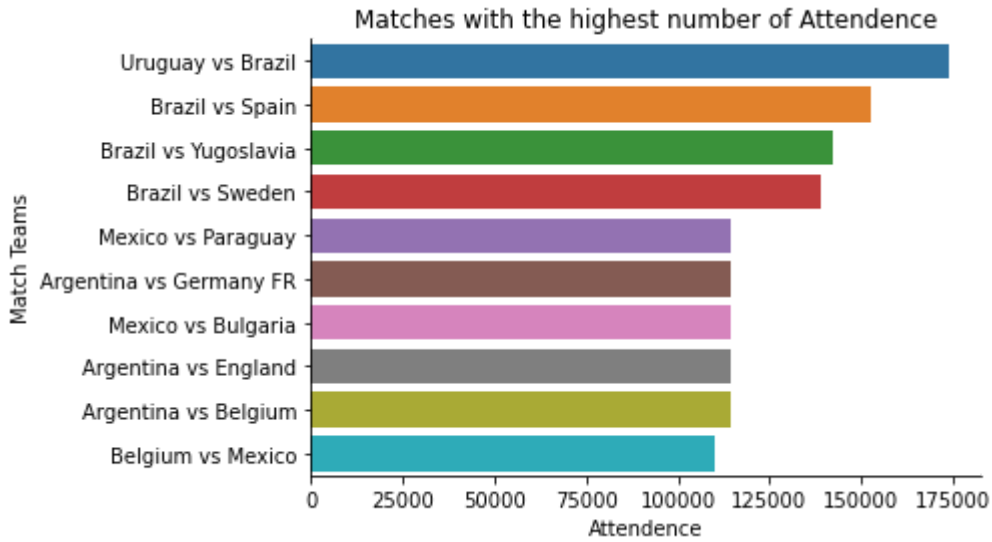
Out[42]:

	Year	Country	Goals
0	1930.0	Argentina	18.0
7	1930.0	Uruguay	15.0
6	1930.0	USA	7.0
8	1930.0	Yugoslavia	7.0
1	1930.0	Brazil	5.0
13	1934.0	Italy	12.0
11	1934.0	Germany	11.0
10	1934.0	Czechoslovakia	9.0
9	1934.0	Austria	7.0
12	1934.0	Hungary	5.0

Matches With Heighest Number Of Attendance

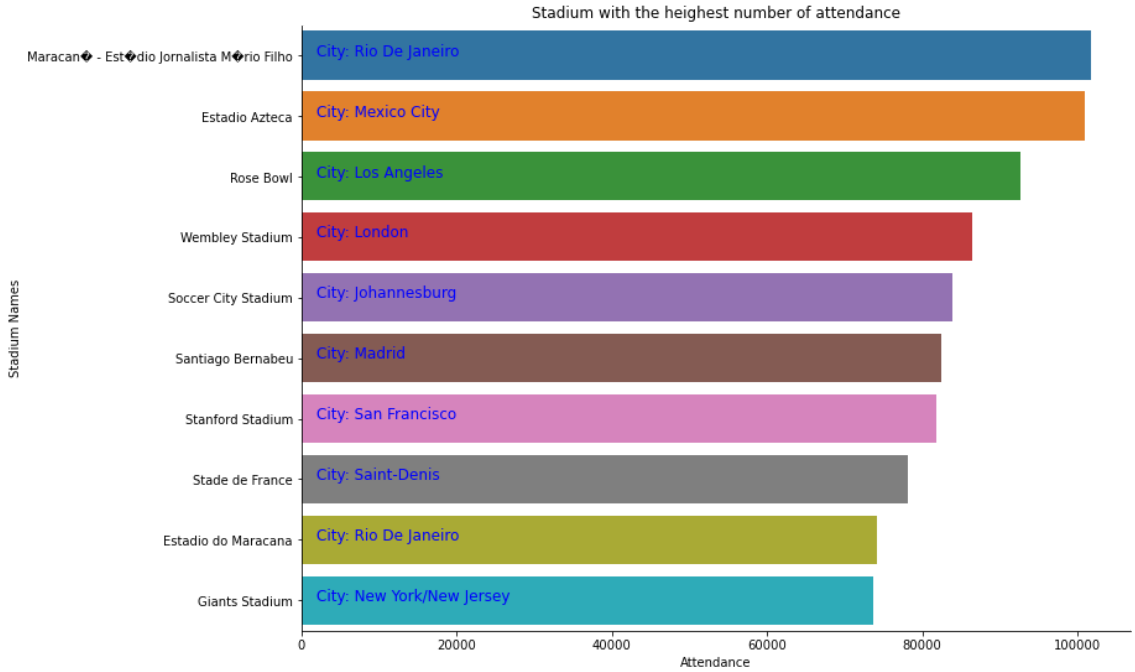
In [43]:
```
matches['Datetime'] = pd.to_datetime(matches['Datetime'])
```

In [49]:
```
ax = sns.barplot(y = top10['vs'], x = top10['Attendance'])
sns.despine(right = True)

plt.ylabel('Match Teams')
plt.xlabel('Attendence')
plt.title('Matches with the highest number of Attendence')

plt.show()
```



Stadium with Highest Average Attendance

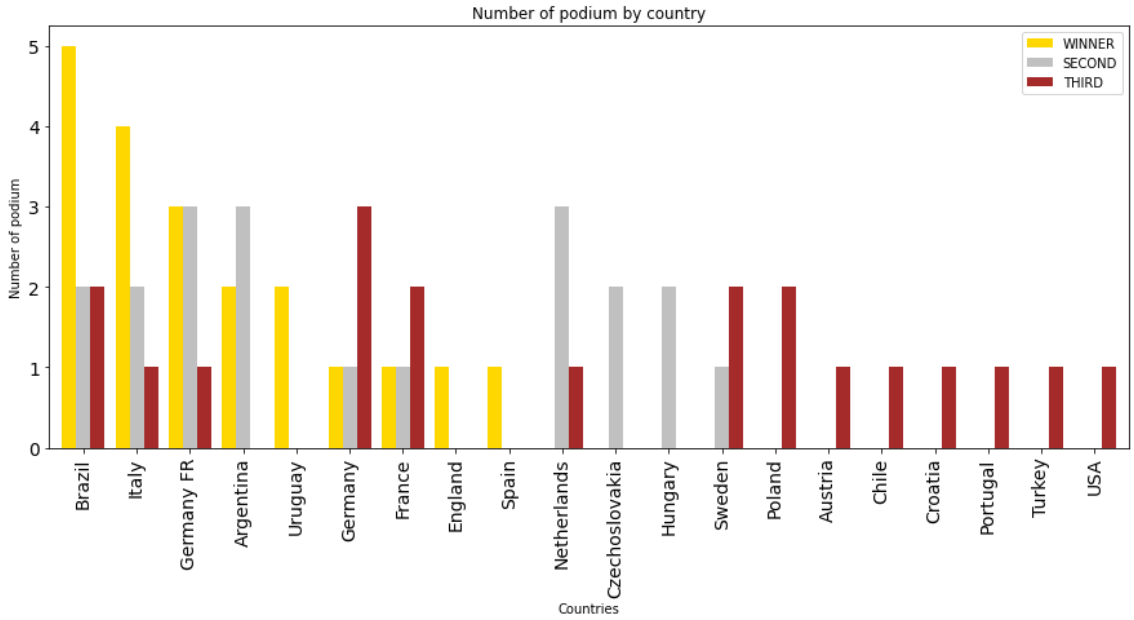
In [53]:
```
std = matches.groupby(['Stadium', 'City'])['Attendance'].mean().reset_index

top10 = std[:10]

plt.figure(figsize = (12,9))
ax = sns.barplot(y = top10['Stadium'], x = top10['Attendance'])
sns.despine(right = True)

plt.ylabel('Stadium Names')
plt.xlabel('Attendance')
plt.title('Stadium with the heighest number of attendance')
for i, s in enumerate("City: " + top10['City']):
        ax.text(2000, i, s, fontsize = 12, color = 'b')

plt.show()
```



Which countries had won the cup ?

In [54]:
```
gold = world_cup["Winner"]
silver = world_cup["Runners-Up"]
bronze = world_cup["Third"]

gold_count = pd.DataFrame.from_dict(gold.value_counts())
silver_count = pd.DataFrame.from_dict(silver.value_counts())
bronze_count = pd.DataFrame.from_dict(bronze.value_counts())
podium_count = gold_count.join(silver_count, how='outer').join(bronze_count
podium_count = podium_count.fillna(0)
podium_count.columns = ['WINNER', 'SECOND', 'THIRD']
podium_count = podium_count.astype('int64')
podium_count = podium_count.sort_values(by=['WINNER', 'SECOND', 'THIRD'], a

podium_count.plot(y=['WINNER', 'SECOND', 'THIRD'], kind="bar",
                  color =['gold','silver','brown'], figsize=(15, 6), fontsi
                 width=0.8, align='center')
plt.xlabel('Countries')
plt.ylabel('Number of podium')
plt.title('Number of podium by country')
```

Out[54]: Text(0.5, 1.0, 'Number of podium by country')

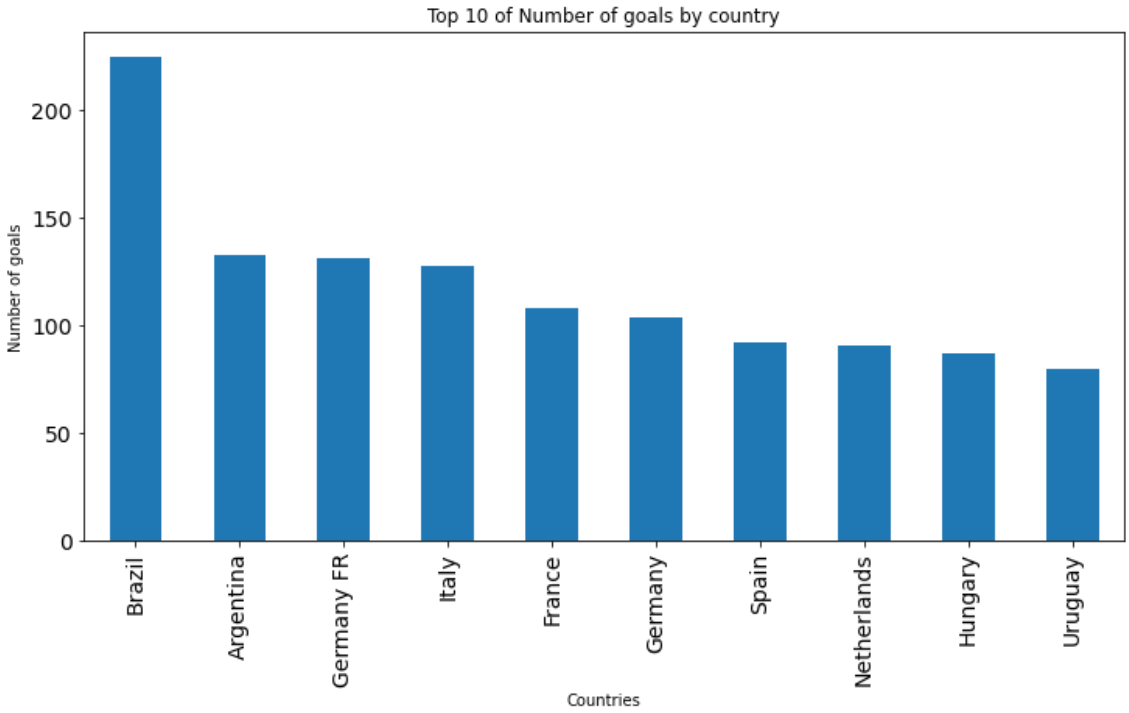


Number of goal per country

In [55]:
```
#world_cups_matches['Win conditions'].value_counts()
home = matches[['Home Team Name', 'Home Team Goals']].dropna()
away = matches[['Away Team Name', 'Away Team Goals']].dropna()

goal_per_country = pd.DataFrame(columns=['countries', 'goals'])
goal_per_country = goal_per_country.append(home.rename(index=str, columns={
goal_per_country = goal_per_country.append(away.rename(index=str, columns={

goal_per_country['goals'] = goal_per_country['goals'].astype('int64')

goal_per_country = goal_per_country.groupby(['countries'])['goals'].sum().s

goal_per_country[:10].plot(x=goal_per_country.index, y=goal_per_country.val
plt.xlabel('Countries')
plt.ylabel('Number of goals')
plt.title('Top 10 of Number of goals by country')
```

```
C:\Users\farid\AppData\Local\Temp\ipykernel_16332\2805712483.py:6: Future
Warning: The frame.append method is deprecated and will be removed from p
andas in a future version. Use pandas.concat instead.
  goal_per_country = goal_per_country.append(home.rename(index=str, colum
ns={'Home Team Name': 'countries', 'Home Team Goals': 'goals'}))
C:\Users\farid\AppData\Local\Temp\ipykernel_16332\2805712483.py:7: Future
Warning: The frame.append method is deprecated and will be removed from p
andas in a future version. Use pandas.concat instead.
  goal_per_country = goal_per_country.append(away.rename(index=str, colum
ns={'Away Team Name': 'countries', 'Away Team Goals': 'goals'}))
```

Out[55]:
```
Text(0.5, 1.0, 'Top 10 of Number of goals by country')
```



Match outcome by home and away teams

In [61]:
```
def get_labels(matches):
    if matches['Home Team Goals'] > matches['Away Team Goals']:
        return 'Home Team Win'
    if matches['Home Team Goals'] < matches['Away Team Goals']:
        return 'Away Team Win'
    return 'DRAW'
```

In [62]:
```
matches['outcome'] = matches.apply(lambda x: get_labels(x), axis=1)
```

In [63]:
```
matches.head()
```

Out[63]:

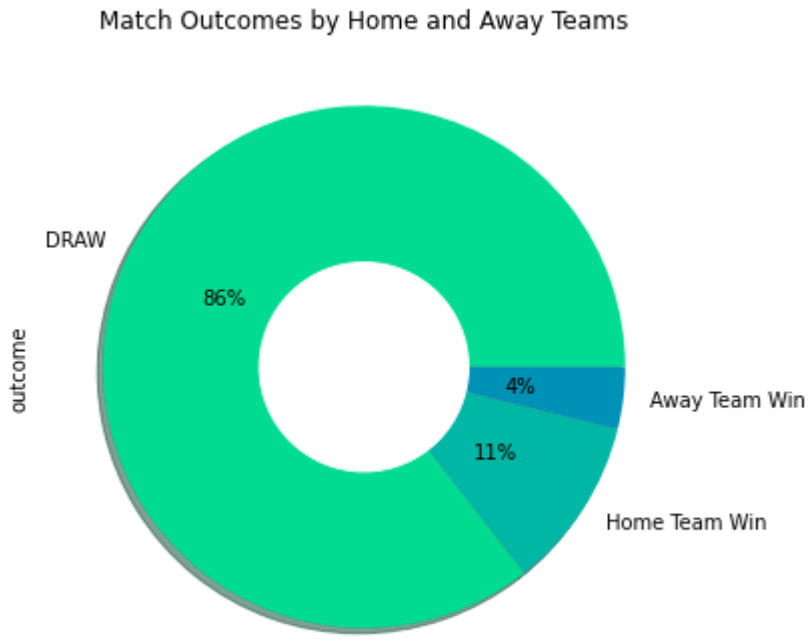
	Year	Datetime	Stage	Stadium	City	Home Team Name	Home Team Goals	Away Team Goals	Away Team Name	W conditic
0	1930.0	1930-07-13 15:00:00	Group 1	Pocitos	Montevideo	France	4.0	1.0	Mexico	
1	1930.0	1930-07-13 15:00:00	Group 4	Parque Central	Montevideo	USA	3.0	0.0	Belgium	
2	1930.0	1930-07-14 12:45:00	Group 2	Parque Central	Montevideo	Yugoslavia	2.0	1.0	Brazil	
3	1930.0	1930-07-14 14:50:00	Group 3	Pocitos	Montevideo	Romania	3.0	1.0	Peru	
4	1930.0	1930-07-15 16:00:00	Group 1	Parque Central	Montevideo	Argentina	1.0	0.0	France	

5 rows × 21 columns

In [64]:
```
mt = matches['outcome'].value_counts()
mt
```

Out[64]:
```
DRAW             3910
Home Team Win     488
Away Team Win     174
Name: outcome, dtype: int64
```

In [65]:
```
plt.figure(figsize = (6,6))

mt.plot.pie(autopct = "%1.0f%%", colors = sns.color_palette('winter_r'), sh

c = plt.Circle((0,0), 0.4, color =  'white')
plt.gca().add_artist(c)
plt.title('Match Outcomes by Home and Away Teams')
plt.show()
```



THANK YOU...!

In [ ]:
```
```