```
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
%matplotlib inline
import plotly as py
import cufflinks as cf
```

```
/kaggle/input/fifa-world-cup/WorldCups.csv
/kaggle/input/fifa-world-cup/WorldCupMatches.csv
/kaggle/input/fifa-world-cup/WorldCupPlayers.csv
/kaggle/input/fifaworld/1_0A8eTfcCEI4vQdErHdrwEQ.jpeg
```

```
players = pd.read_csv("WorldCupPlayers.csv")
matches = pd.read_csv("WorldCupMatches.csv")
world_cup = pd.read_csv("WorldCups.csv")
```

```
players.head()
```

| | 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 |
| | | | | CAUDRON | | | | | |

```
matches.head()
```

| | Year | Datetime | Stage | Stadium | City | Home Team Name | Home Team Goals | Away Team Goals | Away Team Name | cond |
|---|---|---|---|---|---|---|---|---|---|---|
| **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 |

```
matches.tail()
```

| | Year | Datetime | Stage | Stadium | City | Home Team Name | Home Team Goals | Away Team Goals | Away Team Name | Win conditions | Atte |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **4567** | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | |
| **4568** | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | |
| **4569** | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | |
| **4570** | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | |
| **4571** | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | |

```
world_cup.head()
```

| | Year | Country | Winner | Runners-Up | Third | Fourth | GoalsScored | Qualifi |
|---|---|---|---|---|---|---|---|---|
| **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 | |

# Data Cleaning

```
matches.dropna(subset=['Year'], inplace=True)
```

```
matches.tail()
```

| | Year | Datetime | Stage | Stadium | City | Home Team Name | Home Team Goals | Away Team Goals | Away Tea Nam |
|---|---|---|---|---|---|---|---|---|---|
| **847** | 2014.0 | 05 Jul 2014 - 17:00 | Quarter-finals | Arena Fonte Nova | Salvador | Netherlands | 0.0 | 0.0 | Costa Ric |
| **848** | 2014.0 | 08 Jul 2014 - 17:00 | Semi-finals | Estadio Mineirao | Belo Horizonte | Brazil | 1.0 | 7.0 | Germar |
| **849** | 2014.0 | 09 Jul 2014 - 17:00 | Semi-finals | Arena de Sao Paulo | Sao Paulo | Netherlands | 0.0 | 0.0 | Argentir |
| **850** | 2014.0 | 12 Jul 2014 - 17:00 | Play-off for third place | Estadio Nacional | Brasilia | Brazil | 0.0 | 3.0 | Netherlanc |
| **851** | 2014.0 | 13 Jul 2014 - 16:00 | Final | Estadio do Maracana | Rio De Janeiro | Germany | 1.0 | 0.0 | Argentir |

```
matches['Home Team Name'].value_counts()
```

```
Brazil                        82
Italy                         57
Argentina                     54
Germany FR                    43
England                       35
                              ..
rn">Bosnia and Herzegovina     1
Angola                         1
Bolivia                        1
Haiti                          1
Wales                          1
Name: Home Team Name, Length: 78, dtype: int64
```

```
names = matches[matches['Home Team Name'].str.contains('rn">')]['Home Team Name'].value_c
names
```

```
rn">Republic of Ireland       5
rn">Trinidad and Tobago       1
rn">Serbia and Montenegro     1
rn">Bosnia and Herzegovina    1
rn">United Arab Emirates      1
Name: Home Team Name, dtype: int64
```

```
wrong = list(names.index)
wrong
```

```
['rn">Republic of Ireland',
 'rn">Trinidad and Tobago',
 'rn">Serbia and Montenegro',
 'rn">Bosnia and Herzegovina',
 'rn">United Arab Emirates']
```

```
correct = [name.split('>')[1] for name in wrong]
correct
```

```
['Republic of Ireland',
 'Trinidad and Tobago',
 'Serbia and Montenegro',
 'Bosnia and Herzegovina',
 'United Arab Emirates']
```

```
old_name = ['Germany FR', 'Maracan\ufffd - Est\ufffddio Jornalista M\ufffdrio Filho', 'Estadio do Mara
new_name = ['Germany', 'Maracan Stadium', 'Maracan Stadium']
```

```
wrong = wrong + old_name
correct = correct + new_name
```

```
wrong, correct
```

```
(['rn">Republic of Ireland',
  'rn">Trinidad and Tobago',
  'rn">Serbia and Montenegro',
  'rn">Bosnia and Herzegovina',
  'rn">United Arab Emirates',
  'Germany FR',
  'Maracan\ufffd - Est\ufffddio Jornalista M\ufffdrio Filho',
  'Estadio do Maracana'],
 ['Republic of Ireland',
  'Trinidad and Tobago',
  'Serbia and Montenegro',
  'Bosnia and Herzegovina',
  'United Arab Emirates',
  'Germany',
  'Maracan Stadium',
  'Maracan Stadium'])
```

```
for index, wr in enumerate(wrong):
    world_cup = world_cup.replace(wrong[index], correct[index])

for index, wr in enumerate(wrong):
    matches = matches.replace(wrong[index], correct[index])

for index, wr in enumerate(wrong):
```

```
for index, wr in enumerate(wrong):
    players = players.replace(wrong[index], correct[index])


names = matches[matches['Home Team Name'].str.contains('rn">')]['Home Team Name'].value_c
names
```

```
Series([], Name: Home Team Name, dtype: int64)
```

# Most Number of World Cup Winning Title

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

```
Brazil       5
Italy        4
Germany      4
Uruguay      2
Argentina    2
France       1
Spain        1
England      1
Name: Winner, dtype: int64
```

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

```
Germany           4
Argentina         3
Netherlands       3
Hungary           2
Czechoslovakia    2
Brazil            2
Italy             2
Sweden            1
France            1
Name: Runners-Up, dtype: int64
```

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

```
Germany        4
Brazil         2
Poland         2
Sweden         2
France         2
Italy          1
Chile          1
Croatia        1
Portugal       1
```

```
Portugal       1
Austria        1
Turkey         1
Netherlands    1
USA            1
Name: Third, dtype: int64
```

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

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

```
from plotly.offline import iplot
py.offline.init_notebook_mode(connected=True)
cf.go_offline()
```

```
teams.iplot(kind = 'bar', xTitle='Teams', yTitle='Count', title='FIFA World Cup Winning C
```

## Number of Goal Per Countary

```
matches.head(2)
```

| | Year | Datetime | Stage | Stadium | City | Home Team Name | Home Team Goals | Away Team Goals | Away Team Name | conditi |
|---|---|---|---|---|---|---|---|---|---|---|
| **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 | |

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

home.columns = ['Countries', 'Goals']
away.columns = home.columns

goals = home.append(away, ignore_index = True)

goals = goals.groupby('Countries').sum()
goals
```

| Countries | Goals |
|---|---|
| **Algeria** | 14.0 |
| **Angola** | 1.0 |
| **Argentina** | 133.0 |
| **Australia** | 11.0 |
| **Austria** | 43.0 |
| **...** | ... |
| **United Arab Emirates** | 2.0 |
| **Uruguay** | 80.0 |
| **Wales** | 4.0 |
| **Yugoslavia** | 60.0 |
| **Zaire** | 0.0 |

82 rows × 1 columns

```
goals = goals.sort_values(by = 'Goals', ascending=False)
goals
```

| Countries | Goals |
|---|---|
| **Germany** | 235.0 |
| **Brazil** | 225.0 |
| **Argentina** | 133.0 |

| | |
|---|---|
| **Argentina** | 133.0 |
| **Italy** | 128.0 |
| **France** | 108.0 |
| **...** | ... |
| **Trinidad and Tobago** | 0.0 |
| **Canada** | 0.0 |
| **China PR** | 0.0 |
| **Dutch East Indies** | 0.0 |
| **Zaire** | 0.0 |

82 rows × 1 columns

```
goals[:20].iplot(kind='bar', xTitle = 'Country Names', yTitle = 'Goals', title = 'Countri
```

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

# Cup

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

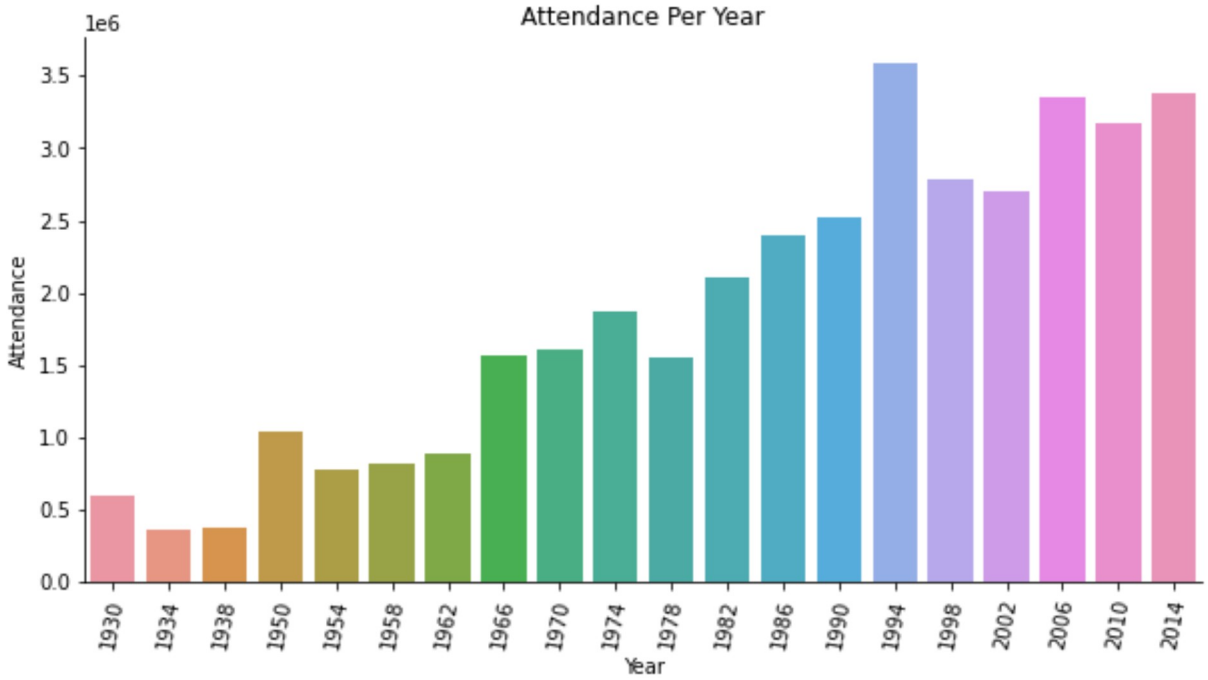
```
world_cup.head()
```

| | Year | Country | Winner | Runners-Up | Third | Fourth | GoalsScored | Qualifi |
|---|---|---|---|---|---|---|---|---|
| **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 | Hungary | Austria | Uruguay | 140 | |

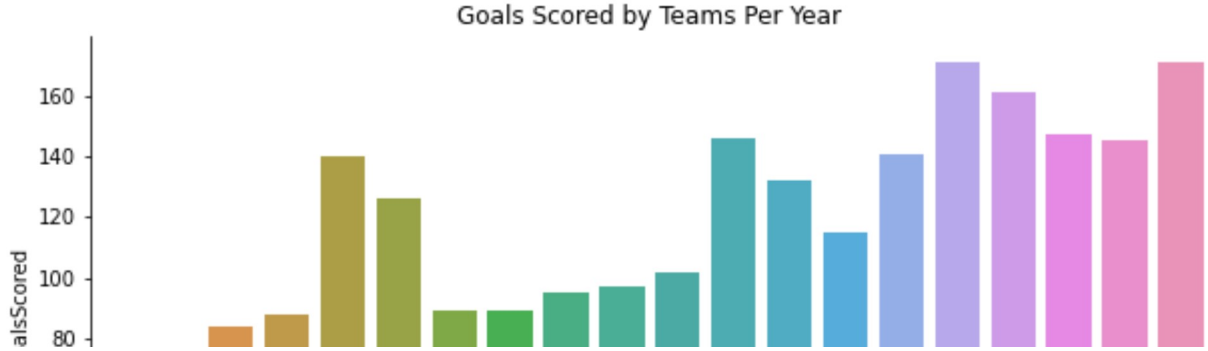
```
fig, ax = plt.subplots(figsize = (10,5))
sns.despine(right = True)
g = sns.barplot(x = 'Year', y = 'Attendance', data = world_cup)
g.set_xticklabels(g.get_xticklabels(), rotation = 80)
g.set_title('Attendance Per Year')

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

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)
```

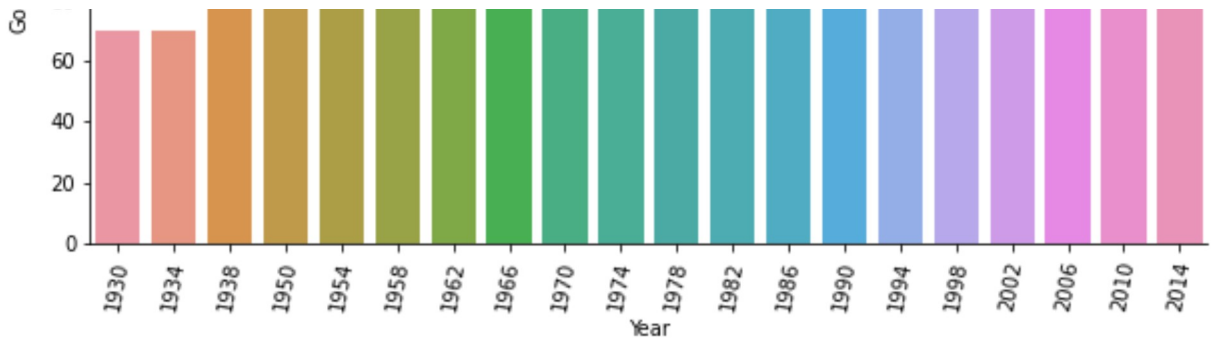
```
g.set_xticklabels(g.get_xticklabels(), rotation = 80)
g.set_title('Matches Plyed Scored by Teams Per Year')
```

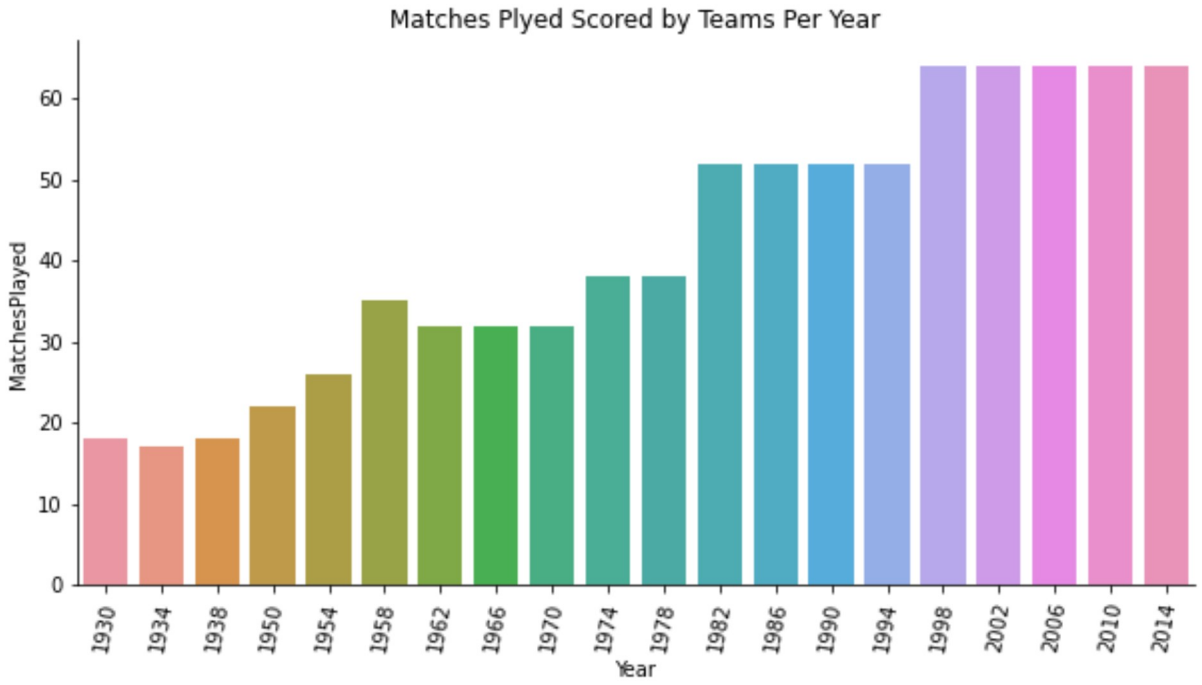
```
Text(0.5, 1.0, 'Matches Plyed Scored by Teams Per Year')
```










# Goals Per Team Per World Cup

```
matches.head(2)
```

| | Year | Datetime | Stage | Stadium | City | Home Team Name | Home Team Goals | Away Team Goals | Away Team Name | conditi |
|---|---|---|---|---|---|---|---|---|---|---|
| **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 | |

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

```
home
```

```
Year    Home Team Name
1930.0  Argentina         16.0
        Brazil             4.0
        Chile              4.0
        France             4.0
        Paraguay           1.0
                          ...
2014.0  Russia             1.0
        Spain              1.0
        Switzerland        4.0
        USA                2.0
        Uruguay            3.0
Name: Home Team Goals, Length: 366, dtype: float64
```

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

```
Year    Away Team Name
1930.0  Argentina         2.0
        Belgium           0.0
        Bolivia           0.0
        Brazil            1.0
        Chile             1.0
                         ...
2014.0  Russia            1.0
        Spain             3.0
        Switzerland       3.0
        USA               4.0
        Uruguay           1.0
Name: Away Team Goals, Length: 411, dtype: float64
```

```
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 = 1)
goals
```

| | | Goals |
|---|---|---|
| **1930.0** | **Argentina** | 18.0 |
| | **Belgium** | 0.0 |
| | **Bolivia** | 0.0 |
| | **Brazil** | 5.0 |
| | **Chile** | 5.0 |
| **...** | **...** | ... |
| **2014.0** | **Russia** | 2.0 |

| | | |
|---|---|---|
| **2014.0** | **Russia** | 2.0 |
| | **Spain** | 4.0 |
| | **Switzerland** | 7.0 |
| | **USA** | 6.0 |
| | **Uruguay** | 4.0 |

427 rows × 1 columns

```
goals = goals.reset_index()
```

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

| | Year | Country | Goals |
|---|---|---|---|
| **0** | 1930.0 | Argentina | 18.0 |
| **11** | 1930.0 | Uruguay | 15.0 |
| **10** | 1930.0 | USA | 7.0 |
| **12** | 1930.0 | Yugoslavia | 7.0 |
| **3** | 1930.0 | Brazil | 5.0 |
| **...** | ... | ... | ... |
| **416** | 2014.0 | Japan | 2.0 |
| **422** | 2014.0 | Russia | 2.0 |
| **401** | 2014.0 | Cameroon | 1.0 |
| **413** | 2014.0 | Honduras | 1.0 |
| **414** | 2014.0 | IR Iran | 1.0 |

427 rows × 3 columns

```
top5 = goals.groupby('Year').head()
top5.head(10)
```

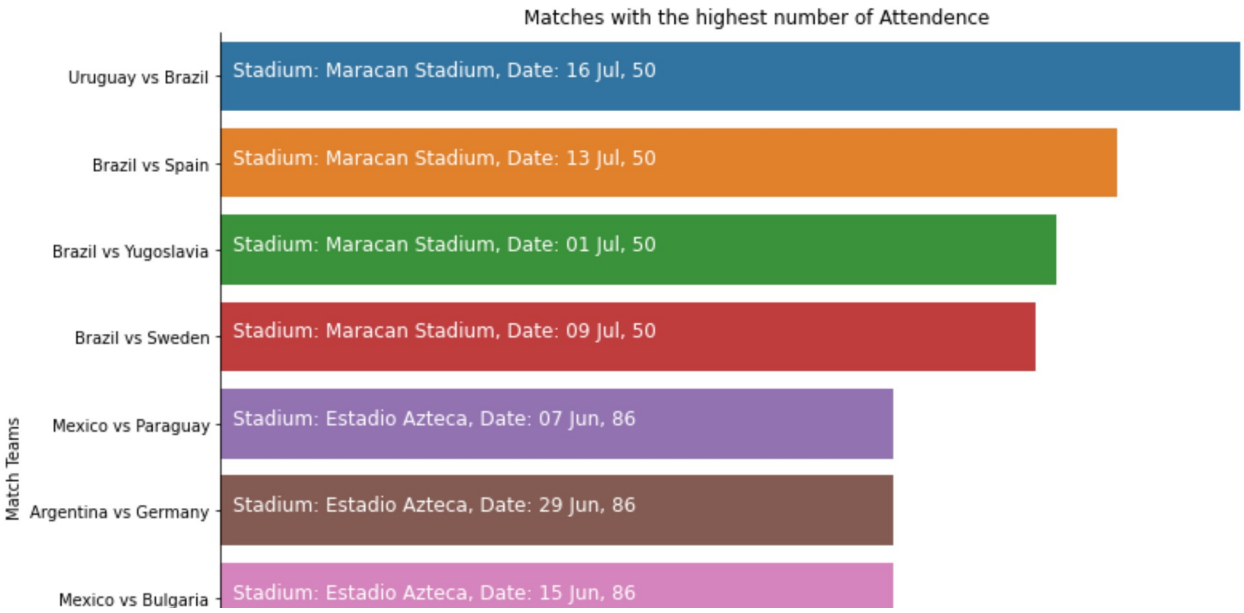
| | Year | Country | Goals |
|---|---|---|---|
| **0** | 1930.0 | Argentina | 18.0 |
| **11** | 1930.0 | Uruguay | 15.0 |
| **10** | 1930.0 | USA | 7.0 |
| **12** | 1930.0 | Yugoslavia | 7.0 |

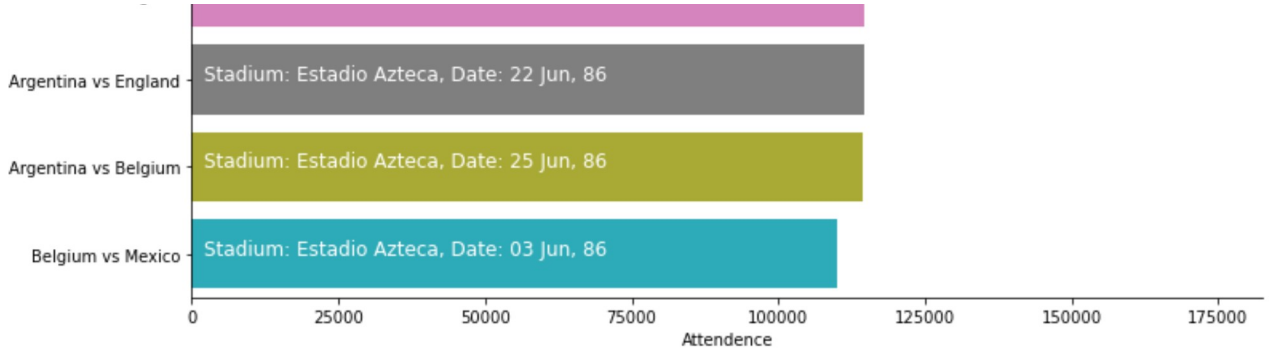
| | | | |
|---|---|---|---|
| **12** | 1930.0 | Yugoslavia | 7.0 |
| **3** | 1930.0 | Brazil | 5.0 |
| **22** | 1934.0 | Italy | 12.0 |
| **20** | 1934.0 | Germany | 11.0 |
| **17** | 1934.0 | Czechoslovakia | 9.0 |
| **14** | 1934.0 | Austria | 7.0 |
| **21** | 1934.0 | Hungary | 5.0 |

```
import plotly.graph_objects as go



x, y = goals['Year'].values, goals['Goals'].values


data = []
for team in top5['Country'].drop_duplicates().values:
    year = top5[top5['Country'] == team]['Year']
    goal = top5[top5['Country'] == team]['Goals']

    data.append(go.Bar(x = year, y = goal, name = team))
layout = go.Layout(barmode = 'stack', title = 'Top 5 Teams with most Goals', showlegend =

fig = go.Figure(data = data, layout = layout)
fig.show()
```

## Matches With Heihest Number Of Attendance

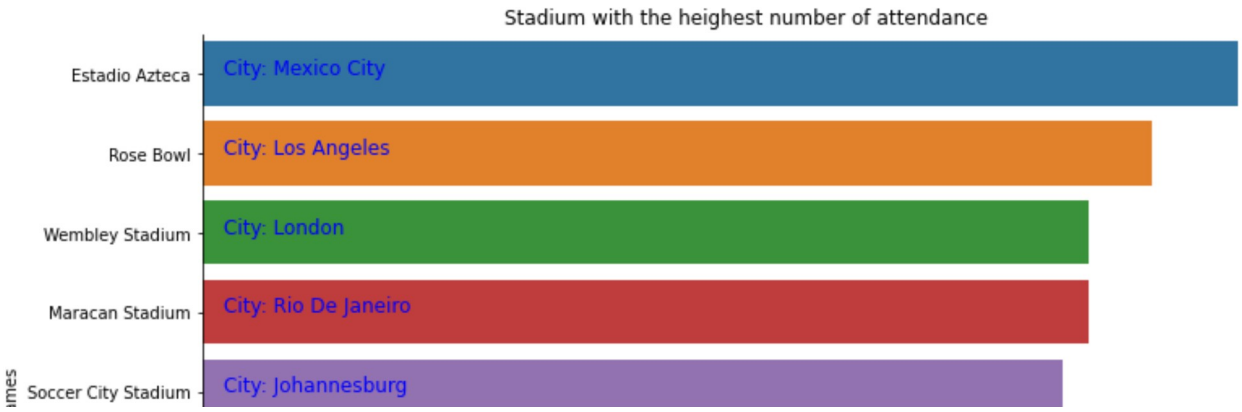
```
matches['Datetime'] = pd.to_datetime(matches['Datetime'])


matches['Datetime'] = matches['Datetime'].apply(lambda x: x.strftime('%d %b, %y'))


top10 = matches.sort_values(by = 'Attendance', ascending = False)[:10]
top10['vs'] = top10['Home Team Name'] + " vs " + top10['Away Team Name']

plt.figure(figsize = (12,10))

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')

for i, s in enumerate("Stadium: " + top10['Stadium'] +", Date: " + top10['Datetime']):
    ax.text(2000, i, s, fontsize = 12, color = 'white')
plt.show()
```

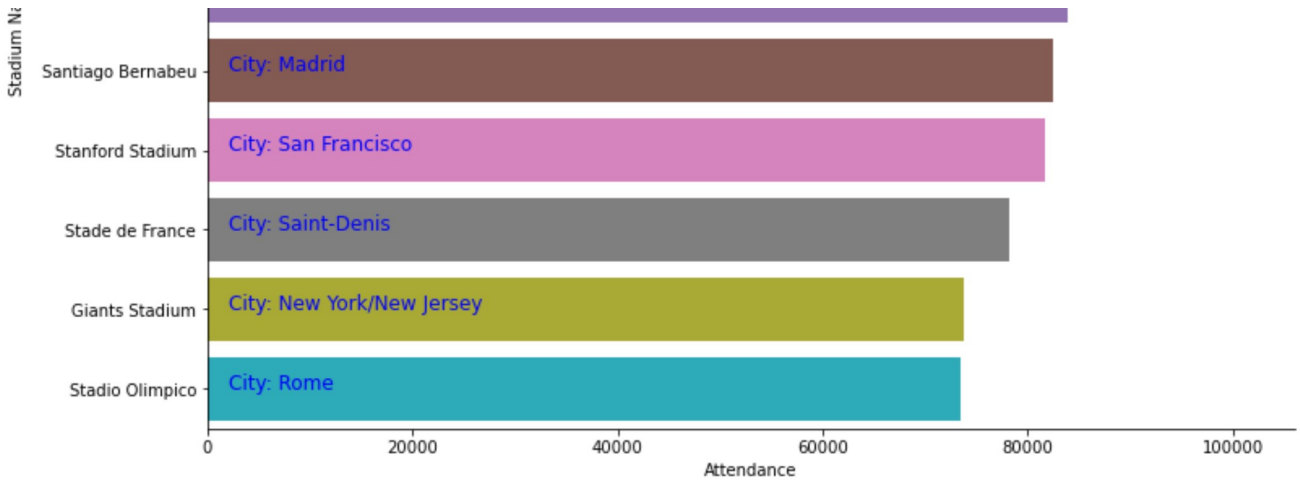





## Stadium with Highest Average Attendance

```
matches['Year'] = matches['Year'].astype(int)

std = matches.groupby(['Stadium', 'City'])['Attendance'].mean().reset_index().sort_values

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()
```



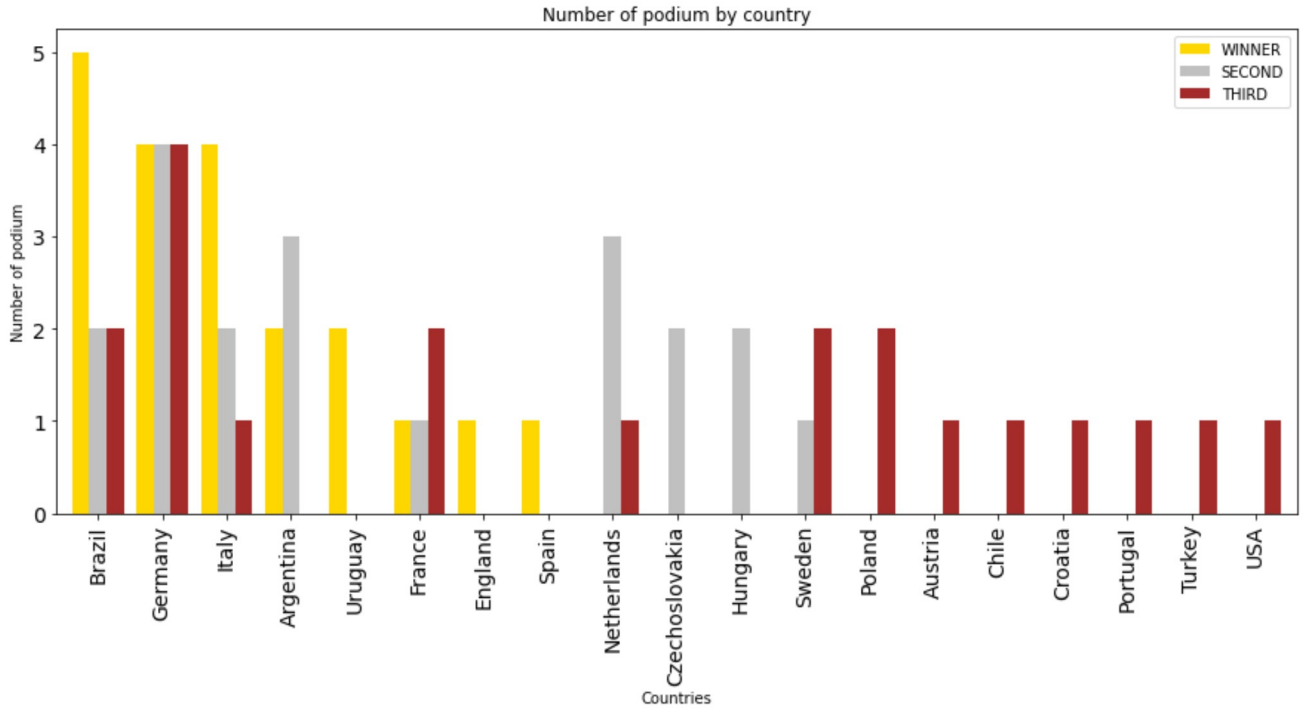



```
matches['City'].value_counts()[:20].iplot(kind = 'bar')
```

# Which countries had won the cup ?

```
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, how='outer')
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'], ascending=False

podium_count.plot(y=['WINNER', 'SECOND', 'THIRD'], kind="bar",
                  color =['gold','silver','brown'], figsize=(15, 6), fontsize=14,
                 width=0.8, align='center')
plt.xlabel('Countries')
plt.ylabel('Number of podium')
plt.title('Number of podium by country')
```

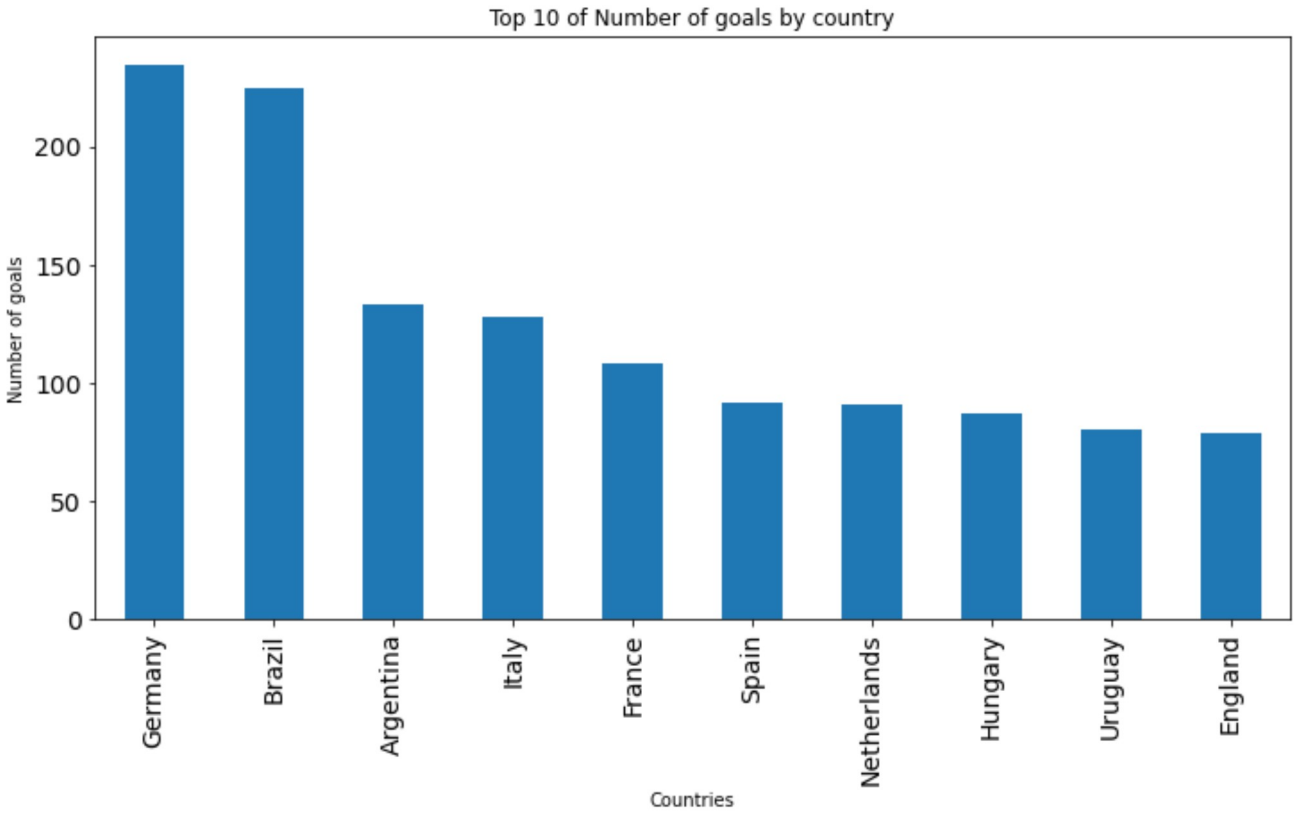
```
Text(0.5, 1.0, 'Number of podium by country')
```



# Number of goal per country

```
#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={'Home Team Nam
goal_per_country = goal_per_country.append(away.rename(index=str, columns={'Away Team Nam

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

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

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

```
Text(0.5, 1.0, 'Top 10 of Number of goals by country')
```



# Match outcome by home and away temas

```
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'


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


matches.head()
```

| | Year | Datetime | Stage | Stadium | City | Home Team Name | Home Team Goals | Away Team Goals | Away Team Name | condit |
|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 1930 | 13 Jul, 30 | Group 1 | Pocitos | Montevideo | France | 4.0 | 1.0 | Mexico | |
| **1** | 1930 | 13 Jul, 30 | Group 4 | Parque Central | Montevideo | USA | 3.0 | 0.0 | Belgium | |
| **2** | 1930 | 14 Jul, 30 | Group 2 | Parque Central | Montevideo | Yugoslavia | 2.0 | 1.0 | Brazil | |
| **3** | 1930 | 14 Jul, 30 | Group 3 | Pocitos | Montevideo | Romania | 3.0 | 1.0 | Peru | |
| **4** | 1930 | 15 Jul, 30 | Group 1 | Parque Central | Montevideo | Argentina | 1.0 | 0.0 | France | |

5 rows × 21 columns

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

```
Home Team Win    488
DRAW             190
Away Team Win    174
Name: outcome, dtype: int64
```

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

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

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()
```