pr11-fifa

February 5, 2024

1 Problem Statement

1.1 With FIFA is in the blood of many people of the world. You are tasked to tell the story of unsung analysts who put great efforts to provide accurate data to answer every question of fans. The FIFA World Cup is a global football competition contested by the various football-playing nations of the world. It is contested every four years and is the most prestigious and important trophy in the sport of football.

1.2 The World Cups dataset shows all information about all the World Cups in history, while the World Cup Matches dataset shows all the results from the matches contested as part of the cups. Find key metrics and factors that influence the World Cup win. Do your own research and come up with your findings

1.3 Importing the Libraries

[1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
import plotly as py
import cufflinks as cf
```

1.4 Getting the Datasets

[2]:
```
players = pd.read_csv("WorldCupPlayers.csv")
matches = pd.read_csv("WorldCupMatches.csv")
world_cup = pd.read_csv("WorldCups.csv")
```

[3]:
```
players.head()
```

[3]:

	RoundID	MatchID	Team Initials	Coach Name	Line-up	Shirt Number	\
0	201	1096	FRA	CAUDRON Raoul (FRA)	S	0	
1	201	1096	MEX	LUQUE Juan (MEX)	S	0	
2	201	1096	FRA	CAUDRON Raoul (FRA)	S	0	
3	201	1096	MEX	LUQUE Juan (MEX)	S	0	

```
4      201     1096           FRA  CAUDRON Raoul (FRA)       S             0

        Player Name Position Event
0       Alex THEPOT       GK   NaN
1   Oscar BONFIGLIO       GK   NaN
2  Marcel LANGILLER      NaN  G40'
3      Juan CARRENO      NaN  G70'
4   Ernest LIBERATI      NaN   NaN
```

[4]:
```
matches.head()
```

[4]:
```
      Year              Datetime    Stage         Stadium         City  \
0  1930.0  13 Jul 1930 - 15:00   Group 1         Pocitos  Montevideo
1  1930.0  13 Jul 1930 - 15:00   Group 4  Parque Central  Montevideo
2  1930.0  14 Jul 1930 - 12:45   Group 2  Parque Central  Montevideo
3  1930.0  14 Jul 1930 - 14:50   Group 3         Pocitos  Montevideo
4  1930.0  15 Jul 1930 - 16:00   Group 1  Parque Central  Montevideo

  Home Team Name  Home Team Goals  Away Team Goals Away Team Name  \
0         France              4.0              1.0         Mexico
1            USA              3.0              0.0        Belgium
2     Yugoslavia              2.0              1.0         Brazil
3        Romania              3.0              1.0           Peru
4      Argentina              1.0              0.0         France

  Win conditions  Attendance  Half-time Home Goals  Half-time Away Goals  \
0                     4444.0                   3.0                   0.0
1                    18346.0                   2.0                   0.0
2                    24059.0                   2.0                   0.0
3                     2549.0                   1.0                   0.0
4                    23409.0                   0.0                   0.0

                  Referee               Assistant 1  \
0  LOMBARDI Domingo (URU)     CRISTOPHE Henry (BEL)
1       MACIAS Jose (ARG)  MATEUCCI Francisco (URU)
2     TEJADA Anibal (URU)   VALLARINO Ricardo (URU)
3   WARNKEN Alberto (CHI)       LANGENUS Jean (BEL)
4     REGO Gilberto (BRA)      SAUCEDO Ulises (BOL)

                  Assistant 2  RoundID  MatchID Home Team Initials  \
0         REGO Gilberto (BRA)    201.0   1096.0                FRA
1       WARNKEN Alberto (CHI)    201.0   1090.0                USA
2         BALWAY Thomas (FRA)    201.0   1093.0                YUG
3    MATEUCCI Francisco (URU)    201.0   1098.0                ROU
4  RADULESCU Constantin (ROU)    201.0   1085.0                ARG

  Away Team Initials
```

```
0                MEX
1                BEL
2                BRA
3                PER
4                FRA
```

[5]:
```
world_cup.head()
```

[5]:
```
    Year      Country      Winner      Runners-Up    Third      Fourth  \
0  1930      Uruguay     Uruguay       Argentina      USA  Yugoslavia
1  1934        Italy       Italy  Czechoslovakia  Germany     Austria
2  1938       France       Italy         Hungary   Brazil      Sweden
3  1950       Brazil     Uruguay          Brazil   Sweden       Spain
4  1954  Switzerland  Germany FR         Hungary  Austria     Uruguay

   GoalsScored  QualifiedTeams  MatchesPlayed Attendance
0           70              13             18    590.549
1           70              16             17    363.000
2           84              15             18    375.700
3           88              13             22  1.045.246
4          140              16             26    768.607
```

1.5 Data Cleaning

[6]:
```
matches.dropna(subset=['Year'], inplace=True)
```

[7]:
```
matches.tail()
```

[7]:
```
       Year              Datetime                     Stage  \
847  2014.0  05 Jul 2014 - 17:00            Quarter-finals
848  2014.0  08 Jul 2014 - 17:00               Semi-finals
849  2014.0  09 Jul 2014 - 17:00               Semi-finals
850  2014.0  12 Jul 2014 - 17:00   Play-off for third place
851  2014.0  13 Jul 2014 - 16:00                     Final

                 Stadium            City Home Team Name  Home Team Goals  \
847     Arena Fonte Nova        Salvador     Netherlands              0.0
848     Estadio Mineirao  Belo Horizonte          Brazil              1.0
849   Arena de Sao Paulo       Sao Paulo     Netherlands              0.0
850     Estadio Nacional        Brasilia          Brazil              0.0
851  Estadio do Maracana  Rio De Janeiro         Germany              1.0

     Away Team Goals Away Team Name                        Win conditions  \
847              0.0     Costa Rica  Netherlands win on penalties (4 - 3)
848              7.0        Germany
849              0.0      Argentina    Argentina win on penalties (2 - 4)
850              3.0    Netherlands
```

```
851              0.0      Argentina          Germany win after extra time

     Attendance  Half-time Home Goals  Half-time Away Goals  \
847     51179.0                   0.0                   0.0
848     58141.0                   0.0                   5.0
849     63267.0                   0.0                   0.0
850     68034.0                   0.0                   2.0
851     74738.0                   0.0                   0.0

                   Referee                   Assistant 1  \
847  Ravshan IRMATOV (UZB)  RASULOV Abduxamidullo (UZB)
848  RODRIGUEZ Marco (MEX)      TORRENTERA Marvin (MEX)
849     C neyt  AKIR (TUR)         DURAN Bahattin (TUR)
850  HAIMOUDI Djamel (ALG)          ACHIK Redouane (MAR)
851   Nicola RIZZOLI (ITA)        Renato FAVERANI (ITA)

                   Assistant 2   RoundID      MatchID Home Team Initials  \
847  KOCHKAROV Bakhadyr (KGZ)  255953.0  300186488.0                 NED
848     QUINTERO Marcos (MEX)  255955.0  300186474.0                 BRA
849         ONGUN Tarik (TUR)  255955.0  300186490.0                 NED
850   ETCHIALI Abdelhak (ALG)  255957.0  300186502.0                 BRA
851      Andrea STEFANI (ITA)  255959.0  300186501.0                 GER

    Away Team Initials
847                CRC
848                GER
849                ARG
850                NED
851                ARG
```

[8]:
```
matches['Home Team Name'].value_counts()
```

[8]:
```
Home Team Name
Brazil                          82
Italy                           57
Argentina                       54
Germany FR                      43
England                         35
                                ..
Wales                            1
Norway                           1
rn">United Arab Emirates         1
Haiti                            1
rn">Bosnia and Herzegovina       1
Name: count, Length: 78, dtype: int64
```

[9]:
```
names = matches[matches['Home Team Name'].str.contains('rn">')]['Home Team␣
↪Name'].value_counts()
names
```

[9]:
```
Home Team Name
rn">Republic of Ireland       5
rn">United Arab Emirates      1
rn">Trinidad and Tobago       1
rn">Serbia and Montenegro     1
rn">Bosnia and Herzegovina    1
Name: count, dtype: int64
```

[10]:
```
wrong = list(names.index)
wrong
```

[10]:
```
['rn">Republic of Ireland',
 'rn">United Arab Emirates',
 'rn">Trinidad and Tobago',
 'rn">Serbia and Montenegro',
 'rn">Bosnia and Herzegovina']
```

[11]:
```
correct = [name.split('>')[1] for name in wrong]
correct
```

[11]:
```
['Republic of Ireland',
 'United Arab Emirates',
 'Trinidad and Tobago',
 'Serbia and Montenegro',
 'Bosnia and Herzegovina']
```

[12]:
```
old_name = ['Germany FR', 'Maracan  - Est dio Jornalista M rio Filho', 'Estadio␣
↪do Maracana']
new_name = ['Germany', 'Maracan Stadium', 'Maracan Stadium']
```

[13]:
```
wrong = wrong + old_name
correct = correct + new_name
```

[14]:
```
wrong, correct
```

[14]:
```
(['rn">Republic of Ireland',
  'rn">United Arab Emirates',
  'rn">Trinidad and Tobago',
  'rn">Serbia and Montenegro',
  'rn">Bosnia and Herzegovina',
  'Germany FR',
  'Maracan  - Est dio Jornalista M rio Filho',
  'Estadio do Maracana'],
```

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

[15]:
```
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):
    players = players.replace(wrong[index], correct[index])
```

[16]:
```
names = matches[matches['Home Team Name'].str.contains('rn">')]['Home Team␣
 ↪Name'].value_counts()
names.name = 'Home Team Name'
names
```

[16]:
```
Series([], Name: Home Team Name, dtype: int64)
```

1.6 Most Number of World Cup Winning Title

[17]:
```
winner = world_cup['Winner'].value_counts()
winner.name = 'Winner'
winner
```

[17]:
```
Winner
Brazil       5
Italy        4
Germany      4
Uruguay      2
Argentina    2
England      1
France       1
Spain        1
Name: Winner, dtype: int64
```

[18]:
```
runnerup = world_cup['Runners-Up'].value_counts()
runnerup.name='Runners-Up'
runnerup
```

[18]:
```
Runners-Up
Germany           4
Argentina         3
Netherlands       3
Czechoslovakia    2
Hungary           2
Brazil            2
Italy             2
Sweden            1
France            1
Name: Runners-Up, dtype: int64
```

[19]:
```
third = world_cup['Third'].value_counts()
third.name='Third'
third
```

[19]:
```
Third
Germany        4
Brazil         2
Sweden         2
France         2
Poland         2
USA            1
Austria        1
Chile          1
Portugal       1
Italy          1
Croatia        1
Turkey         1
Netherlands    1
Name: Third, dtype: int64
```

[20]:
```
teams = pd.concat([winner, runnerup, third], axis=1)
teams.fillna(0, inplace=True)
teams = teams.astype(int)
teams.sort_index(inplace=True)
teams
```

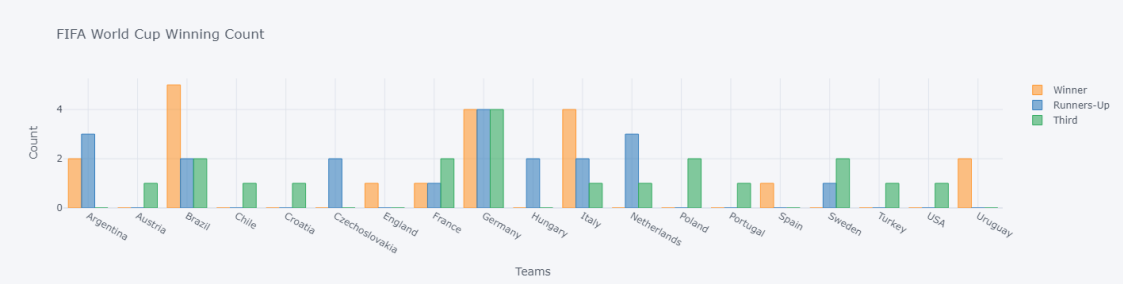
[20]:

	Winner	Runners-Up	Third
Argentina	2	3	0
Austria	0	0	1
Brazil	5	2	2
Chile	0	0	1
Croatia	0	0	1
Czechoslovakia	0	2	0
England	1	0	0
France	1	1	2

```
Germany          4          4       4
Hungary          0          2       0
Italy            4          2       1
Netherlands      0          3       1
Poland           0          0       2
Portugal         0          0       1
Spain            1          0       0
Sweden           0          1       2
Turkey           0          0       1
USA              0          0       1
Uruguay          2          0       0
```

[21]:
```
from plotly.offline import iplot
py.offline.init_notebook_mode(connected=True)
cf.go_offline()
```

[22]:
```
teams.iplot(kind = 'bar', xTitle='Teams', yTitle='Count', title='FIFA World Cup␣
↪Winning Count')
```



1.7 Number of Goals per country

[23]:
```
matches.head()
```

[23]:
```
     Year              Datetime    Stage         Stadium         City  \
0  1930.0  13 Jul 1930 - 15:00   Group 1         Pocitos  Montevideo
1  1930.0  13 Jul 1930 - 15:00   Group 4  Parque Central  Montevideo
2  1930.0  14 Jul 1930 - 12:45   Group 2  Parque Central  Montevideo
3  1930.0  14 Jul 1930 - 14:50   Group 3         Pocitos  Montevideo
4  1930.0  15 Jul 1930 - 16:00   Group 1  Parque Central  Montevideo

  Home Team Name  Home Team Goals  Away Team Goals Away Team Name  \
0         France              4.0              1.0         Mexico
1            USA              3.0              0.0        Belgium
2     Yugoslavia              2.0              1.0         Brazil
3        Romania              3.0              1.0           Peru
4      Argentina              1.0              0.0         France
```

```
  Win conditions  Attendance  Half-time Home Goals  Half-time Away Goals  \
0                      4444.0                   3.0                   0.0
1                     18346.0                   2.0                   0.0
2                     24059.0                   2.0                   0.0
3                      2549.0                   1.0                   0.0
4                     23409.0                   0.0                   0.0

                  Referee               Assistant 1  \
0  LOMBARDI Domingo (URU)     CRISTOPHE Henry (BEL)
1       MACIAS Jose (ARG)  MATEUCCI Francisco (URU)
2     TEJADA Anibal (URU)   VALLARINO Ricardo (URU)
3   WARNKEN Alberto (CHI)       LANGENUS Jean (BEL)
4     REGO Gilberto (BRA)      SAUCEDO Ulises (BOL)

                  Assistant 2  RoundID  MatchID Home Team Initials  \
0         REGO Gilberto (BRA)    201.0   1096.0                 FRA
1       WARNKEN Alberto (CHI)    201.0   1090.0                 USA
2         BALWAY Thomas (FRA)    201.0   1093.0                 YUG
3    MATEUCCI Francisco (URU)    201.0   1098.0                 ROU
4  RADULESCU Constantin (ROU)    201.0   1085.0                 ARG

  Away Team Initials
0                MEX
1                BEL
2                BRA
3                PER
4                FRA
```

[24]:
```
home = matches[['Home Team Name', 'Home Team Goals']].dropna()
away = matches[['Away Team Name', 'Away Team Goals']].dropna()
```

[25]:
```
home.columns = ['Countries', 'Goals']
away.columns = home.columns
```

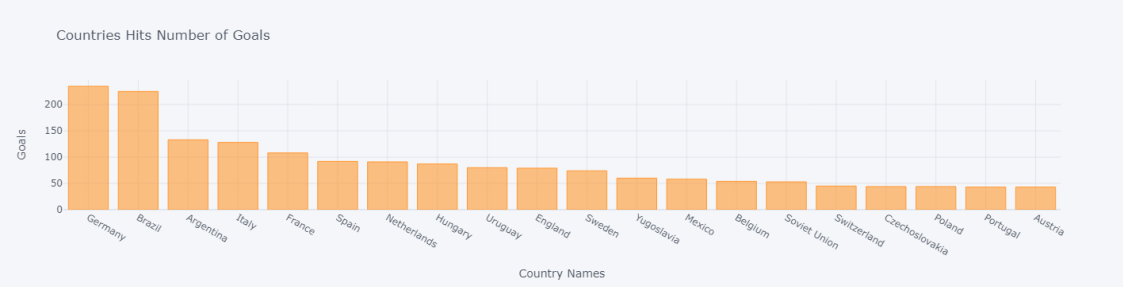
[26]:
```
goals = pd.concat([home, away], ignore_index=True)
```

[27]:
```
goals = goals.groupby('Countries').sum()
goals = goals.sort_values(by = 'Goals', ascending=False)
goals
```

[27]:
```
                Goals
Countries
Germany         235.0
Brazil          225.0
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 x 1 columns]
```

[28]:
```
goals[:20].iplot(kind='bar', xTitle = 'Country Names', yTitle = 'Goals', title
↪= 'Countries Hits Number of Goals')
```



1.8 Attendance per Year

[29]:
```
world_cup['Attendance'] = world_cup['Attendance'].str.replace(".", "")
```

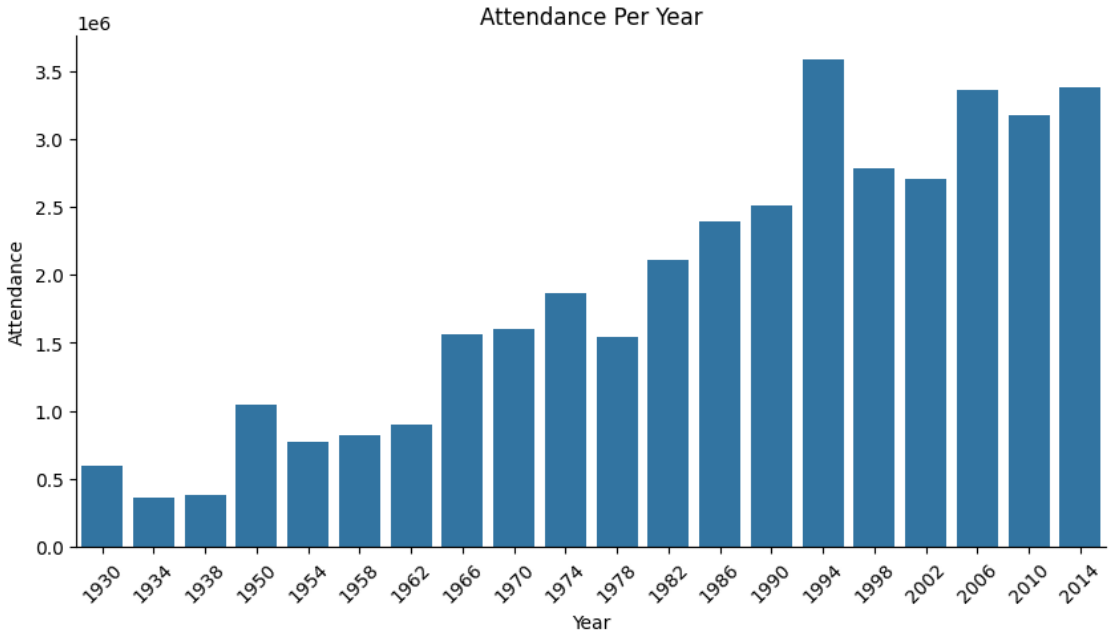
[30]:
```
world_cup['Attendance'] = world_cup['Attendance'].astype(int)
world_cup.head()
```

[30]:

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

	GoalsScored	QualifiedTeams	MatchesPlayed	Attendance
0	70	13	18	590549
1	70	16	17	363000
2	84	15	18	375700
3	88	13	22	1045246
4	140	16	26	768607

[31]:
```
fig, ax = plt.subplots(figsize = (10,5))
sns.despine(right = True)
g = sns.barplot(x="Year", y="Attendance", data=world_cup)

ax.tick_params(axis="x", labelrotation=45)
g.set_title("Attendance Per Year")

# Show the plot
plt.show()
```

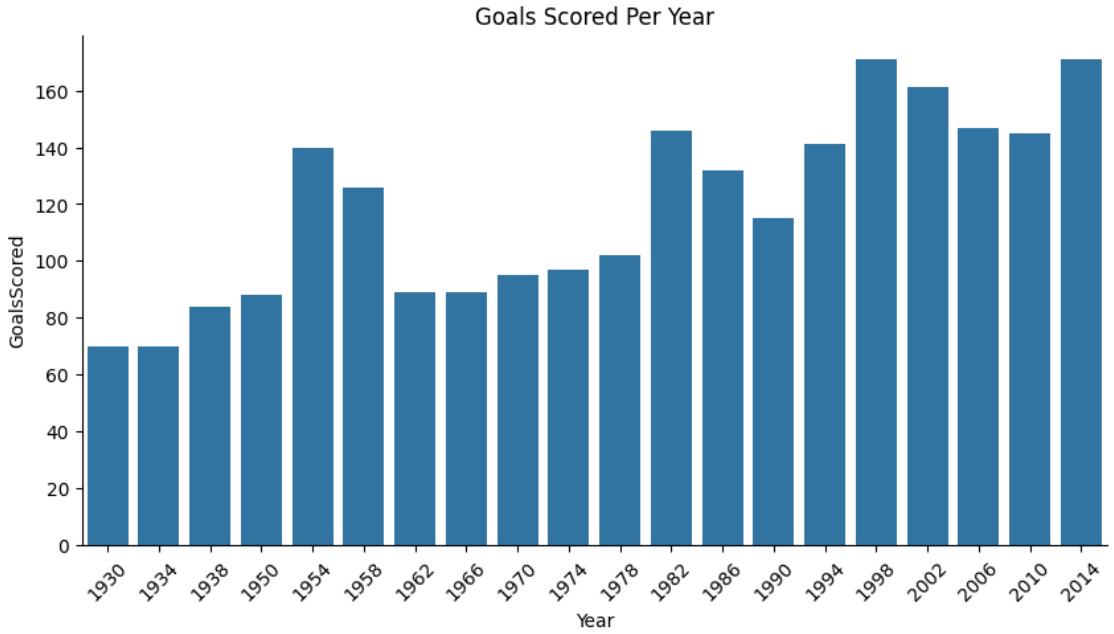


1.9 Qualified Teams per Year

[32]:
```
fig, ax = plt.subplots(figsize = (10,5))
sns.despine(right = True)
g = sns.barplot(x = 'Year', y = 'QualifiedTeams', data = world_cup)

ax.tick_params(axis="x", labelrotation=45)
g.set_title('Qualified Teams Per Year')
plt.show()
```

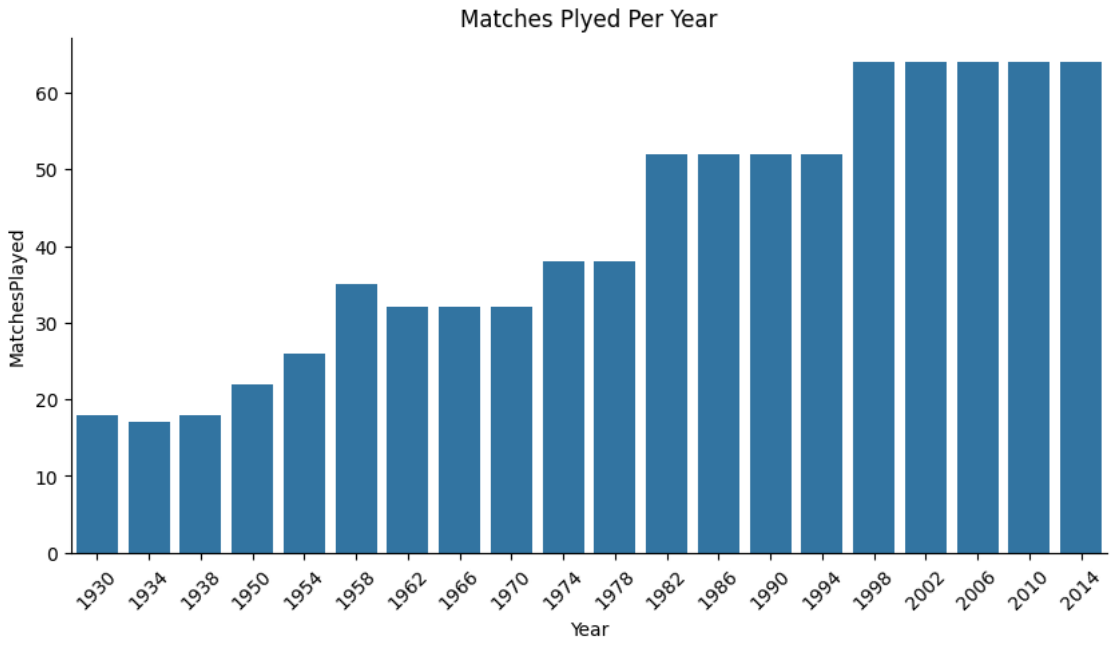
1.10 Total Goals Scored per Year

[33]:
```
fig, ax = plt.subplots(figsize = (10,5))
sns.despine(right = True)
g = sns.barplot(x = 'Year', y = 'GoalsScored', data = world_cup)

ax.tick_params(axis="x", labelrotation=45)
g.set_title('Goals Scored Per Year')
plt.show()
```



1.11 Total Matches Played per Year

[34]:
```
fig, ax = plt.subplots(figsize = (10,5))
sns.despine(right = True)
g = sns.barplot(x = 'Year', y = 'MatchesPlayed', data = world_cup)

ax.tick_params(axis="x", labelrotation=45)
g.set_title('Matches Plyed Per Year')
plt.show()
```



1.12 Goals per Team (Year-Wise) - Top N

[35]:
```
matches.head()
```

[35]:

	Year	Datetime	Stage	Stadium	City	Home Team Name	Home Team Goals	Away Team Goals	Away Team Name	Win conditions	Attendance	Half-time Home Goals	Half-time Away Goals
0	1930.0	13 Jul 1930 - 15:00	Group 1	Pocitos	Montevideo	France	4.0	1.0	Mexico		4444.0	3.0	0.0
1	1930.0	13 Jul 1930 - 15:00	Group 4	Parque Central	Montevideo	USA	3.0	0.0	Belgium		18346.0	2.0	0.0
2	1930.0	14 Jul 1930 - 12:45	Group 2	Parque Central	Montevideo	Yugoslavia	2.0	1.0	Brazil		24059.0	2.0	0.0
3	1930.0	14 Jul 1930 - 14:50	Group 3	Pocitos	Montevideo	Romania	3.0	1.0	Peru		2549.0	1.0	0.0
4	1930.0	15 Jul 1930 - 16:00	Group 1	Parque Central	Montevideo	Argentina	1.0	0.0	France		23409.0	0.0	0.0

```
                  Referee               Assistant 1  \
0  LOMBARDI Domingo (URU)     CRISTOPHE Henry (BEL)
1       MACIAS Jose (ARG)  MATEUCCI Francisco (URU)
2     TEJADA Anibal (URU)   VALLARINO Ricardo (URU)
3   WARNKEN Alberto (CHI)       LANGENUS Jean (BEL)
4     REGO Gilberto (BRA)      SAUCEDO Ulises (BOL)

                  Assistant 2  RoundID  MatchID Home Team Initials  \
0         REGO Gilberto (BRA)    201.0   1096.0                 FRA
1       WARNKEN Alberto (CHI)    201.0   1090.0                 USA
2         BALWAY Thomas (FRA)    201.0   1093.0                 YUG
3    MATEUCCI Francisco (URU)    201.0   1098.0                 ROU
4  RADULESCU Constantin (ROU)    201.0   1085.0                 ARG

  Away Team Initials
0                MEX
1                BEL
2                BRA
3                PER
4                FRA
```

[36]:
```
home = matches.groupby(['Year', 'Home Team Name'])['Home Team Goals'].sum()
home
```

[36]:
```
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
```

[37]:
```
away = matches.groupby(['Year', 'Away Team Name'])['Away Team Goals'].sum()
away
```

[37]:
```
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
```

[38]:
```
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
```

[38]:
```
                  Goals
Year
1930.0 Argentina   18.0
       Brazil       5.0
       Chile        5.0
       France       4.0
       Paraguay     1.0
…                   …
1998.0 Iran         2.0
       Mexico       8.0
       Norway       5.0
       Tunisia      1.0
2006.0 IR Iran      0.0

[427 rows x 1 columns]
```

[39]:
```
goals = goals.reset_index()
goals.columns = ['Year', 'Country', 'Goals']
goals = goals.sort_values(by = ['Year', 'Goals'], ascending = [True, False])
goals
```

[39]:
```
        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
..       …          …     …
355   2014.0       Japan    2.0
361   2014.0      Russia    2.0
340   2014.0    Cameroon    1.0
352   2014.0    Honduras    1.0
353   2014.0     IR Iran    1.0
```

```
[427 rows x 3 columns]
```

[40]:
```
top5 = goals.groupby('Year').head()
top5.head(10)
```

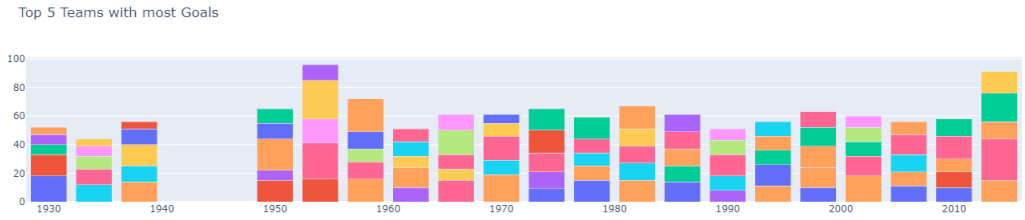
[40]:

	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

[41]:
```
import plotly.graph_objects as go
```

[42]:
```
x, y = goals['Year'].values, goals['Goals'].values
```

[43]:
```
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 = False)

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



1.13 Matches with Highest Number of Attendance

[44]:
```
matches.head()
```

[44]:

	Year	Datetime	Stage	Stadium	City	Home Team Name	Home Team Goals	Away Team Goals	Away Team Name	Win conditions	Attendance	Half-time Home Goals	Half-time Away Goals	Referee	Assistant 1	Assistant 2	RoundID	MatchID	Home Team Initials	Away Team Initials
0	1930.0	13 Jul 1930 - 15:00	Group 1	Pocitos	Montevideo	France	4.0	1.0	Mexico		4444.0	3.0	0.0	LOMBARDI Domingo (URU)	CRISTOPHE Henry (BEL)	REGO Gilberto (BRA)	201.0	1096.0	FRA	MEX
1	1930.0	13 Jul 1930 - 15:00	Group 4	Parque Central	Montevideo	USA	3.0	0.0	Belgium		18346.0	2.0	0.0	MACIAS Jose (ARG)	MATEUCCI Francisco (URU)	WARNKEN Alberto (CHI)	201.0	1090.0	USA	BEL
2	1930.0	14 Jul 1930 - 12:45	Group 2	Parque Central	Montevideo	Yugoslavia	2.0	1.0	Brazil		24059.0	2.0	0.0	TEJADA Anibal (URU)	VALLARINO Ricardo (URU)	BALWAY Thomas (FRA)	201.0	1093.0	YUG	BRA
3	1930.0	14 Jul 1930 - 14:50	Group 3	Pocitos	Montevideo	Romania	3.0	1.0	Peru		2549.0	1.0	0.0	WARNKEN Alberto (CHI)	LANGENUS Jean (BEL)	MATEUCCI Francisco (URU)	201.0	1098.0	ROU	PER
4	1930.0	15 Jul 1930 - 16:00	Group 1	Parque Central	Montevideo	Argentina	1.0	0.0	France		23409.0	0.0	0.0	REGO Gilberto (BRA)	SAUCEDO Ulises (BOL)	RADULESCU Constantin (ROU)	201.0	1085.0	ARG	FRA

1.13.1 Cleaning the datetime column

[45]:
```
f_matches=matches.copy()
```

[46]:
```
f_matches['Datetime'] = pd.to_datetime(f_matches['Datetime'], format="mixed",␣
↪errors='coerce')
```

[47]:
```
f_matches.head()
```

[47]:
```
     Year            Datetime    Stage         Stadium         City  \
0  1930.0 1930-07-13 15:00:00  Group 1         Pocitos  Montevideo
1  1930.0 1930-07-13 15:00:00  Group 4  Parque Central  Montevideo
2  1930.0 1930-07-14 12:45:00  Group 2  Parque Central  Montevideo
3  1930.0 1930-07-14 14:50:00  Group 3         Pocitos  Montevideo
4  1930.0 1930-07-15 16:00:00  Group 1  Parque Central  Montevideo

  Home Team Name  Home Team Goals  Away Team Goals Away Team Name  \
0         France              4.0              1.0         Mexico
1            USA              3.0              0.0        Belgium
2     Yugoslavia              2.0              1.0         Brazil
3        Romania              3.0              1.0           Peru
4      Argentina              1.0              0.0         France

  Win conditions  Attendance  Half-time Home Goals  Half-time Away Goals  \
0                     4444.0                   3.0                   0.0
1                    18346.0                   2.0                   0.0
2                    24059.0                   2.0                   0.0
3                     2549.0                   1.0                   0.0
4                    23409.0                   0.0                   0.0

                  Referee               Assistant 1  \
0  LOMBARDI Domingo (URU)     CRISTOPHE Henry (BEL)
1       MACIAS Jose (ARG)  MATEUCCI Francisco (URU)
2     TEJADA Anibal (URU)   VALLARINO Ricardo (URU)
3   WARNKEN Alberto (CHI)       LANGENUS Jean (BEL)
4     REGO Gilberto (BRA)      SAUCEDO Ulises (BOL)

                  Assistant 2  RoundID  MatchID Home Team Initials  \
0         REGO Gilberto (BRA)    201.0   1096.0                 FRA
1       WARNKEN Alberto (CHI)    201.0   1090.0                 USA
2         BALWAY Thomas (FRA)    201.0   1093.0                 YUG
3    MATEUCCI Francisco (URU)    201.0   1098.0                 ROU
4  RADULESCU Constantin (ROU)    201.0   1085.0                 ARG

  Away Team Initials
0                MEX
1                BEL
```

```
2                BRA
3                PER
4                FRA
```

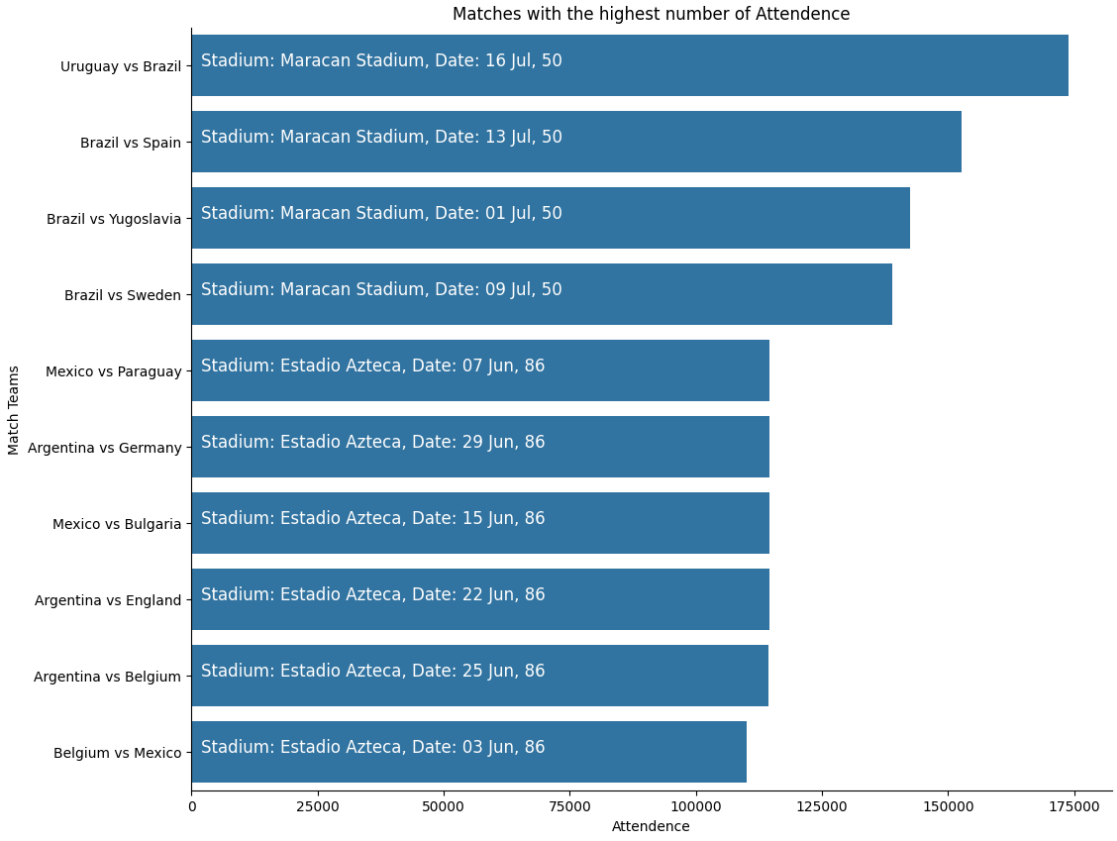
[48]:
```
f_matches = f_matches.dropna(subset=['Datetime'])
```

[49]:
```
f_matches['Datetime'] = f_matches['Datetime'].apply(lambda x: x.strftime('%d␣
↪%b, %y'))
```

[50]:
```
f_matches.head()
```

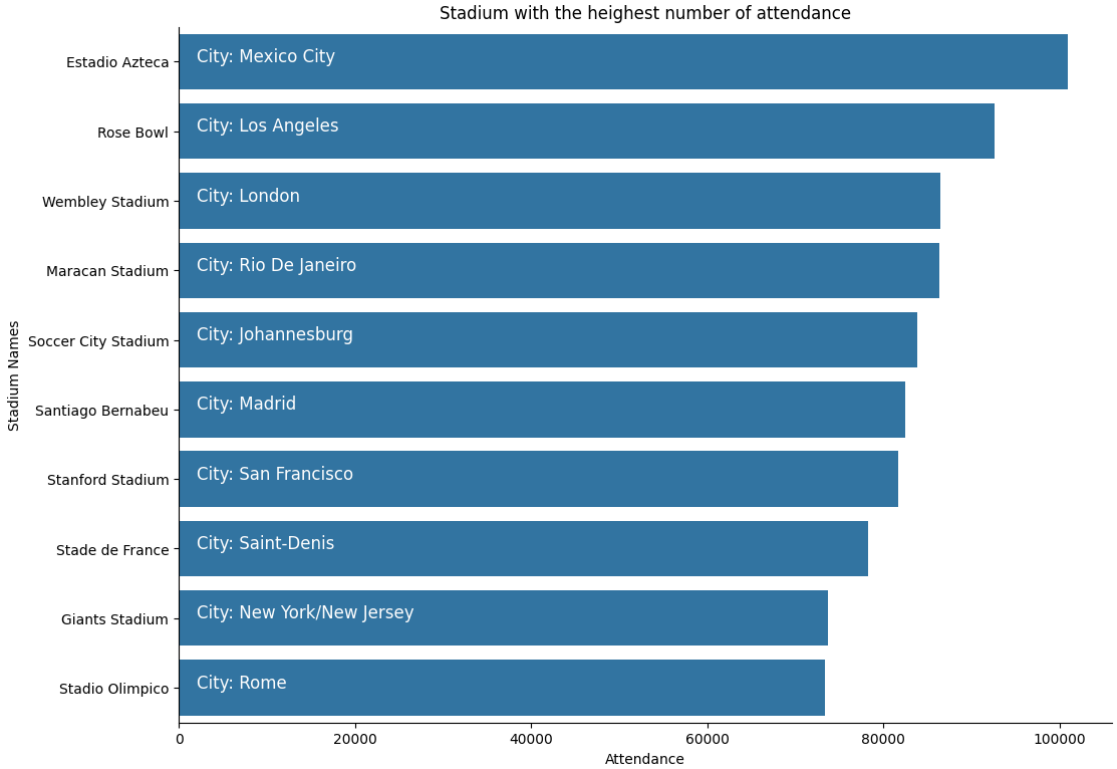
[50]:
```
     Year    Datetime    Stage         Stadium         City Home Team Name  \
0  1930.0  13 Jul, 30  Group 1         Pocitos  Montevideo          France
1  1930.0  13 Jul, 30  Group 4  Parque Central  Montevideo             USA
2  1930.0  14 Jul, 30  Group 2  Parque Central  Montevideo      Yugoslavia
3  1930.0  14 Jul, 30  Group 3         Pocitos  Montevideo         Romania
4  1930.0  15 Jul, 30  Group 1  Parque Central  Montevideo       Argentina

   Home Team Goals  Away Team Goals Away Team Name Win conditions  Attendance  \
0              4.0              1.0         Mexico                     4444.0
1              3.0              0.0        Belgium                    18346.0
2              2.0              1.0         Brazil                    24059.0
3              3.0              1.0           Peru                     2549.0
4              1.0              0.0         France                    23409.0

   Half-time Home Goals  Half-time Away Goals                Referee  \
0                   3.0                   0.0  LOMBARDI Domingo (URU)
1                   2.0                   0.0       MACIAS Jose (ARG)
2                   2.0                   0.0     TEJADA Anibal (URU)
3                   1.0                   0.0   WARNKEN Alberto (CHI)
4                   0.0                   0.0     REGO Gilberto (BRA)

                Assistant 1                 Assistant 2  RoundID  MatchID  \
0     CRISTOPHE Henry (BEL)         REGO Gilberto (BRA)    201.0   1096.0
1  MATEUCCI Francisco (URU)       WARNKEN Alberto (CHI)    201.0   1090.0
2   VALLARINO Ricardo (URU)         BALWAY Thomas (FRA)    201.0   1093.0
3       LANGENUS Jean (BEL)    MATEUCCI Francisco (URU)    201.0   1098.0
4      SAUCEDO Ulises (BOL)  RADULESCU Constantin (ROU)    201.0   1085.0

  Home Team Initials Away Team Initials
0                FRA                MEX
1                USA                BEL
2                YUG                BRA
3                ROU                PER
4                ARG                FRA
```

[51]:
```
top10 = f_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()
```



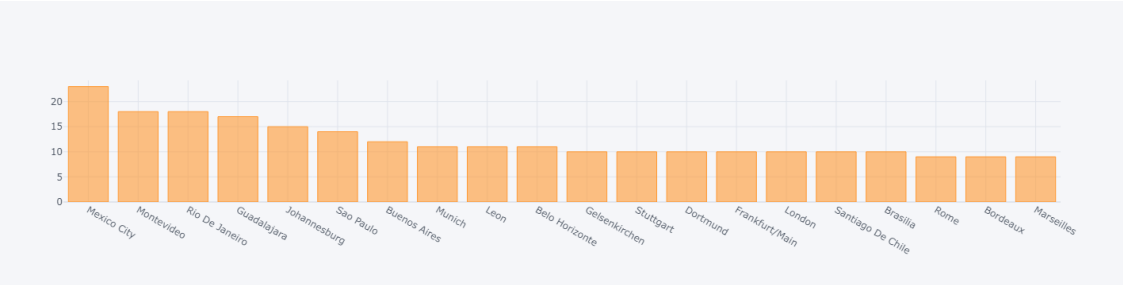
1.14 Stadium with Highest Average Attendance

[52]:
```
f_matches['Year'] = f_matches['Year'].astype(int)

std = f_matches.groupby(['Stadium', 'City'])['Attendance'].mean().reset_index().
 ↪sort_values(by = 'Attendance', ascending =False)

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 = 'white')

plt.show()
```



1.15 Matches played per City

[53]:
```
matches['City'].value_counts()[:20].iplot(kind = 'bar')
```



1.16 Match Outcome by home and away teams

[54]:
```
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'
```

[55]:
```
matches['outcome'] = matches.apply(lambda x: get_labels(x), axis=1)
```

[56]:
```
matches.head()
```

[56]:

	Year	Datetime	Stage	Stadium	City	\
0	1930.0	13 Jul 1930 - 15:00	Group 1	Pocitos	Montevideo	
1	1930.0	13 Jul 1930 - 15:00	Group 4	Parque Central	Montevideo	
2	1930.0	14 Jul 1930 - 12:45	Group 2	Parque Central	Montevideo	
3	1930.0	14 Jul 1930 - 14:50	Group 3	Pocitos	Montevideo	
4	1930.0	15 Jul 1930 - 16:00	Group 1	Parque Central	Montevideo	

	Home Team Name	Home Team Goals	Away Team Goals	Away Team Name	\
0	France	4.0	1.0	Mexico	
1	USA	3.0	0.0	Belgium	
2	Yugoslavia	2.0	1.0	Brazil	
3	Romania	3.0	1.0	Peru	
4	Argentina	1.0	0.0	France	

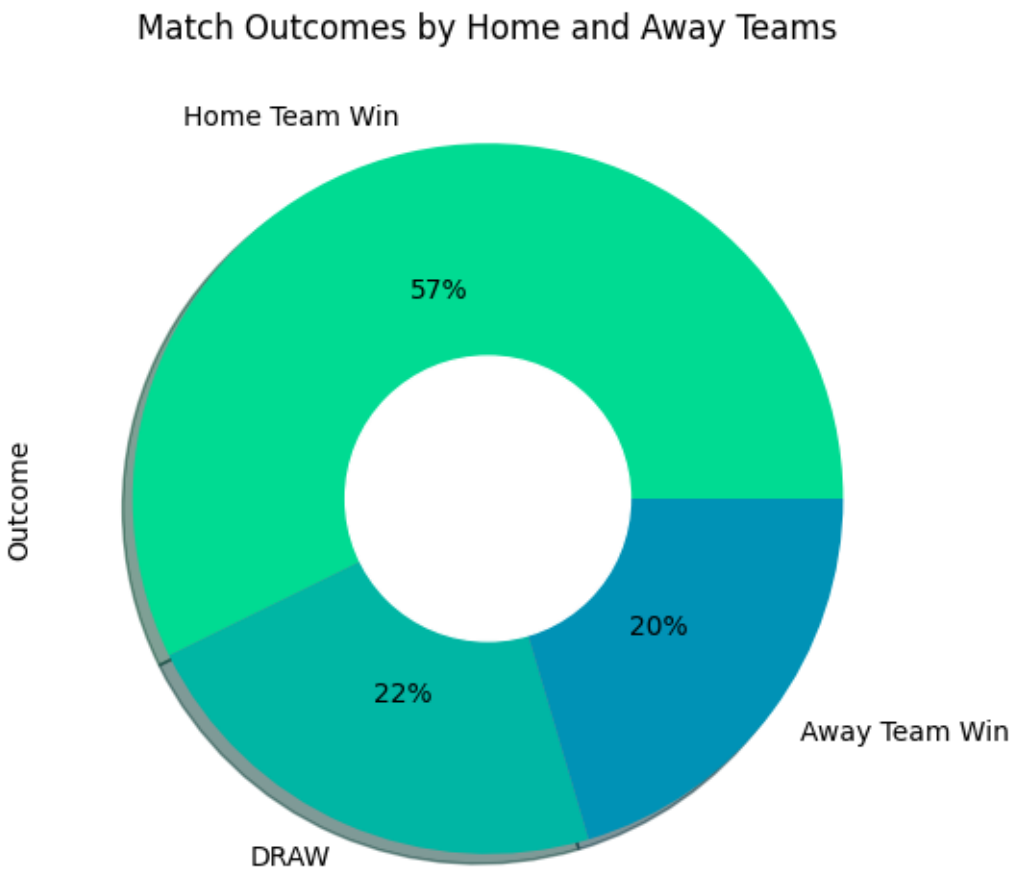
	Win conditions	…	Half-time Home Goals	Half-time Away Goals	\
0		…	3.0	0.0	
1		…	2.0	0.0	
2		…	2.0	0.0	
3		…	1.0	0.0	

```
4                 …                   0.0                   0.0

                  Referee               Assistant 1  \
0  LOMBARDI Domingo (URU)     CRISTOPHE Henry (BEL)
1       MACIAS Jose (ARG)  MATEUCCI Francisco (URU)
2     TEJADA Anibal (URU)   VALLARINO Ricardo (URU)
3   WARNKEN Alberto (CHI)       LANGENUS Jean (BEL)
4     REGO Gilberto (BRA)      SAUCEDO Ulises (BOL)

                  Assistant 2 RoundID  MatchID  Home Team Initials  \
0         REGO Gilberto (BRA)   201.0   1096.0                 FRA
1       WARNKEN Alberto (CHI)   201.0   1090.0                 USA
2         BALWAY Thomas (FRA)   201.0   1093.0                 YUG
3    MATEUCCI Francisco (URU)   201.0   1098.0                 ROU
4  RADULESCU Constantin (ROU)   201.0   1085.0                 ARG

  Away Team Initials        outcome
0                MEX  Home Team Win
1                BEL  Home Team Win
2                BRA  Home Team Win
3                PER  Home Team Win
4                FRA  Home Team Win

[5 rows x 21 columns]
```

[57]:
```
final = matches['outcome'].value_counts()
final.name='Outcome'
final
```

[57]:
```
outcome
Home Team Win    488
DRAW             190
Away Team Win    174
Name: Outcome, dtype: int64
```

[58]:
```
plt.figure(figsize = (6,6))

final.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()
```


Match Outcomes by Home and Away Teams
Home Team Win
57%
Outcome
20%
22%
Away Team Win
DRAW