

In []:

S06 T01: Tasca dades, probabilitats i estadístiques

In []:

#Nivell 1

In [124]:

```
#Exercici 1
#Agafa un conjunt de dades de tema esportiu que t'agradi i selecciona un atribut del co
njunt de dades.
#Calcula la moda, la mediana, la desviació estàndard i la mitjana aritmètica.

#Llibreries
import numpy as np
import pandas as pd
import random
import seaborn as sns
import warnings
import statistics

#Prenem el conjunt de dades de tema esportiu de la pàgina web
#https://www.kaggle.com/abecklas/fifa-world-cup/version/5

#Dataset FIFA World Cup
#Fitxer: Python/WorldCupPlayers.csv
#Fitxer: Python/WorldCups.csv
#Fitxer: Python/WorldCupMatches.csv

wcupplay_df = pd.read_csv('Python/WorldCupPlayers.csv', engine="python", error_bad_line
s=False, warn_bad_lines=False, sep=',')
wcup_df = pd.read_csv('Python/WorldCups.csv', engine="python", error_bad_lines=False, w
arn_bad_lines=False, sep=',')
wcupmatch_df = pd.read_csv('Python/WorldCupMatches.csv', engine="python", error_bad_lin
es=False, warn_bad_lines=False, sep=',')
```

In [125]:

```
#Calcula la moda, la mediana, la desviació estàndard i la mitjana aritmètica.
#Escollim l'atribut 'Attendance' - Població que va assistir al World Cup
wcup_df['Attendance'] = wcup_df['Attendance'].astype(str).str.replace('.', '').astype(in
t)
```

In [126]:

```
#Mitja aritmètica d'assistència als World Cup de 1930 a 1978
print('La mitjana aritmètica és:', round(statistics.mean(wcup_df['Attendance'])))
```

La mitjana aritmètica és: 1872882

In [127]:

```
#Mediana d'assistència World Cup 1930-1978
print('La mediana és:', round(statistics.median(wcup_df['Attendance'])))
```

La mediana és: 1734864

In [128]:

```
#Moda d'assistència World Cup 1930-1978
print('La moda és:', statistics.mode(wcup_df['Attendance']))
```

La moda és: 590549

In [129]:

```
#Desviació standard d'assistència World Cup 1930-1978
print('La desviació standard és:', round(wcup_df['Attendance'].std()))
```

La desviació standard és: 1071842

In [130]:

```
#Exercici 2
#Continuant amb les dades de tema esportiu, selecciona dos atributs i calcula'n la seva
correlació.
#Escollim els atributs wcupmatch_df['Home Team Goals'] i wcupmatch_df['Away Team Goals']

print('La correlació dels atributs Home Team Goals i Away Team Goals és',
      round(np.corrcoef(wcupmatch_df['Home Team Goals'], wcupmatch_df['Away Team Goals']
                        ))[0,1],8))
```

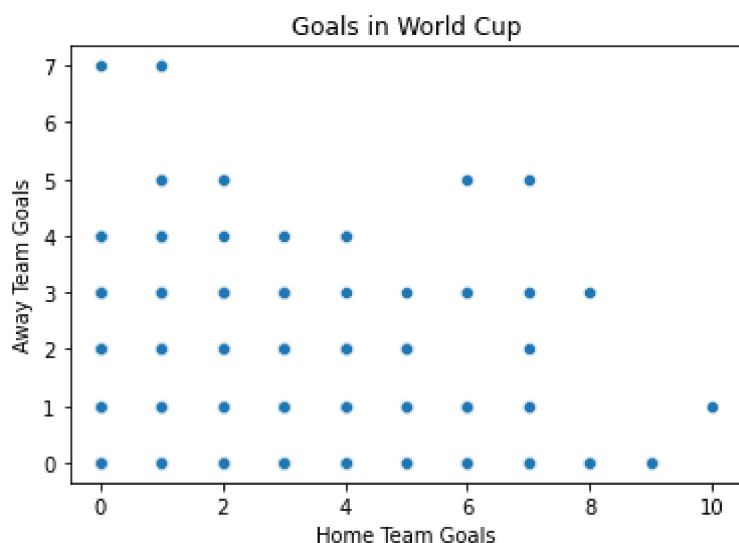
La correlació dels atributs Home Team Goals i Away Team Goals és 0.0124739

9

In [131]:

```
#Graphic World Cup 1 - Home Team Goals vs Away Team Goals
import matplotlib.pyplot as plt
import seaborn as sns

sns.scatterplot(data=wcupmatch_df, x="Home Team Goals", y="Away Team Goals", cmap='Greens')
plt.title('Goals in World Cup')
plt.show()
```



In [132]:

```
from scipy import stats
stats.pearsonr(wcupmatch_df['Home Team Goals'], wcupmatch_df['Away Team Goals'])
```

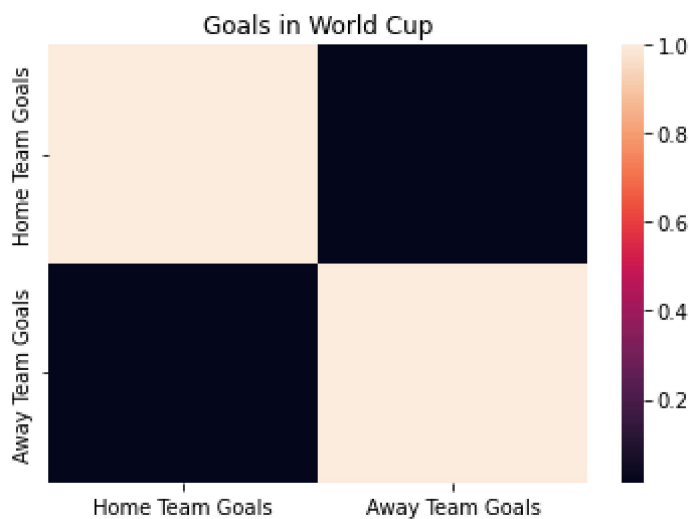
Out[132]:

```
(0.012473988573227651, 0.7161691736497401)
```

In [133]:

```
##Graphic World Cup 2 - Home Team Goals vs Away Team Goals
import matplotlib.pyplot as plt
import seaborn as sns

cormat=wcupmatch_df[['Home Team Goals','Away Team Goals']].corr()
sns.heatmap(cormat);
plt.title('Goals in World Cup')
plt.show()
```



In [134]:

```
#NivelL 2
```

In [135]:

```
#Exercici 3
#Continuant amb les dades de tema esportiu, calcula la correlació de tots els atributs
#entre si i representa'ls en una matriu amb diferents colors d'intensitat.
```

```
#Escollim els atributs de wcupplay_df
```

```
corr = wcupmatch_df.corr() #Prenem només els atributs numèrics de wcupplay_df
```

```
fig = plt.figure(figsize=(10,10), dpi=80)
```

```
ax = fig.add_subplot(111)
```

```
cax = ax.matshow(corr, cmap='coolwarm', vmin=-1, vmax=1) # plot de tipus matshow
```

```
fig.colorbar(cax)
```

```
ticks = np.arange(0, len(corr.columns), 1) #nombre de columnes
```

```
ax.set_xticks(ticks)
```

```
plt.xticks(rotation=90)
```

```
ax.set_yticks(ticks)
```

```
ax.set_xticklabels(corr.columns)
```

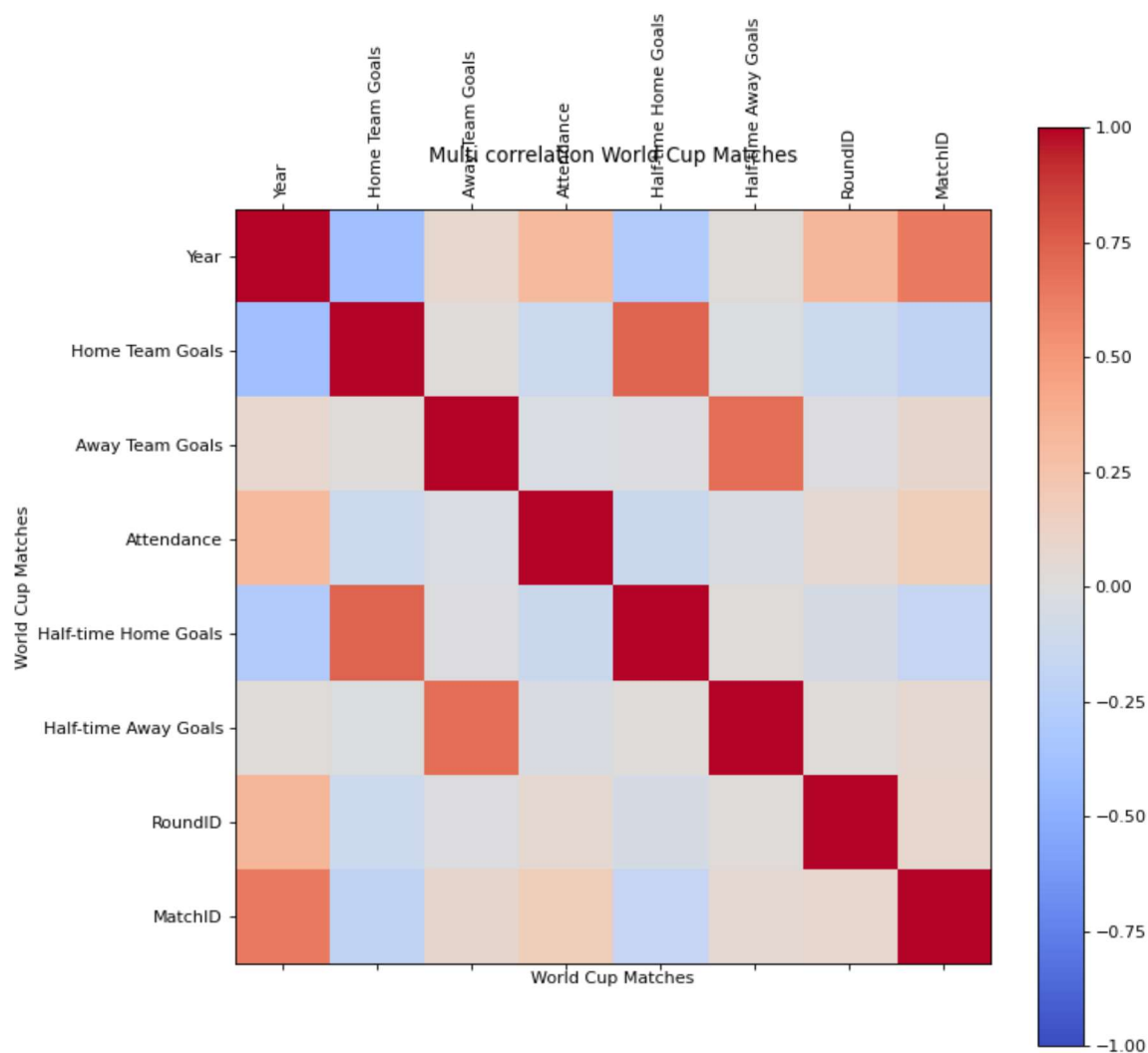
```
ax.set_yticklabels(corr.columns)
```

```
ax.set_title('Multi correlation World Cup Matches')
```

```
ax.set_xlabel('World Cup Matches')
```

```
ax.set_ylabel('World Cup Matches')
```

```
plt.show()
```



In [136]:

```
#Nivell 3
```

In [137]:

```
#Exercici 4
#Continuant amb les dades de tema esportiu, selecciona un atribut i calcula la mitjana
geomètrica i la mitjana harmònica.
```

In [138]:

```
from scipy.stats.mstats import gmean
#Escolliu l'atribut 'Attendance' - Població que va assistir al World Cup
wcup_df['Attendance'] = wcup_df['Attendance'].astype(str).str.replace('.', '').astype(int)
```

In [139]:

```
#Mitja geomètrica d'assistència als World Cup de 1930 a 1978
print('La mitjana geomètrica és:', round(gmean(wcup_df['Attendance'])))
```

La mitjana geomètrica és: 1518630.0

In [141]:

```
#Mitjana harmònica d'assistència als World Cup de 1930 a 1978
print('La mitjana harmònica és:', round(statistics.harmonic_mean(wcup_df['Attendance'])))
```

La mitjana harmònica és: 1152259

In []: