Week 1 SportStats Analysis

June 25, 2023

[1]:
```
import pandas as pd
import pandasql as ps
import matplotlib.pyplot as plt
```

[2]:
```
noc_regions = pd.read_csv('Data/noc_regions.csv')
```

[3]:
```
athlete_events = pd.read_csv('Data/athlete_events.csv')
```

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

[4]:

	ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season	City	Sport	Event	Medal
0	1	A Dijiang	M	24.0	180.0	80.0	China	CHN	1992 Summer	1992	Summer	Barcelona	Basketball	Basketball Men's Basketball	NaN
1	2	A Lamusi	M	23.0	170.0	60.0	China	CHN	2012 Summer	2012	Summer	London	Judo	Judo Men's Extra-Lightweight	NaN
2	3	Gunnar Nielsen Aaby	M	24.0	NaN	NaN	Denmark	DEN	1920 Summer	1920	Summer	Antwerpen	Football	Football Men's Football	NaN
3	4	Edgar Lindenau Aabye	M	34.0	NaN	NaN	Denmark/Sweden	DEN	1900 Summer	1900	Summer	Paris	Tug-Of-War	Tug-Of-War Men's Tug-Of-War	Gold
4	5	Christine Jacoba Aaftink	F	21.0	185.0	82.0	Netherlands	NED	1988 Winter	1988	Winter	Calgary	Speed Skating	Speed Skating Women's 500 metres	NaN

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

[5]:

	NOC	region	notes
0	AFG	Afghanistan	NaN
1	AHO	Curacao	Netherlands Antilles
2	ALB	Albania	NaN

```
3  ALG      Algeria                   NaN
4  AND      Andorra                   NaN
```

[6]:
```
query1 = "SELECT Sport, COUNT(DISTINCT Name) AS 'Athlete Count' FROM␣
↪athlete_events GROUP BY Sport ORDER BY COUNT(DISTINCT Name) DESC LIMIT 5"
```

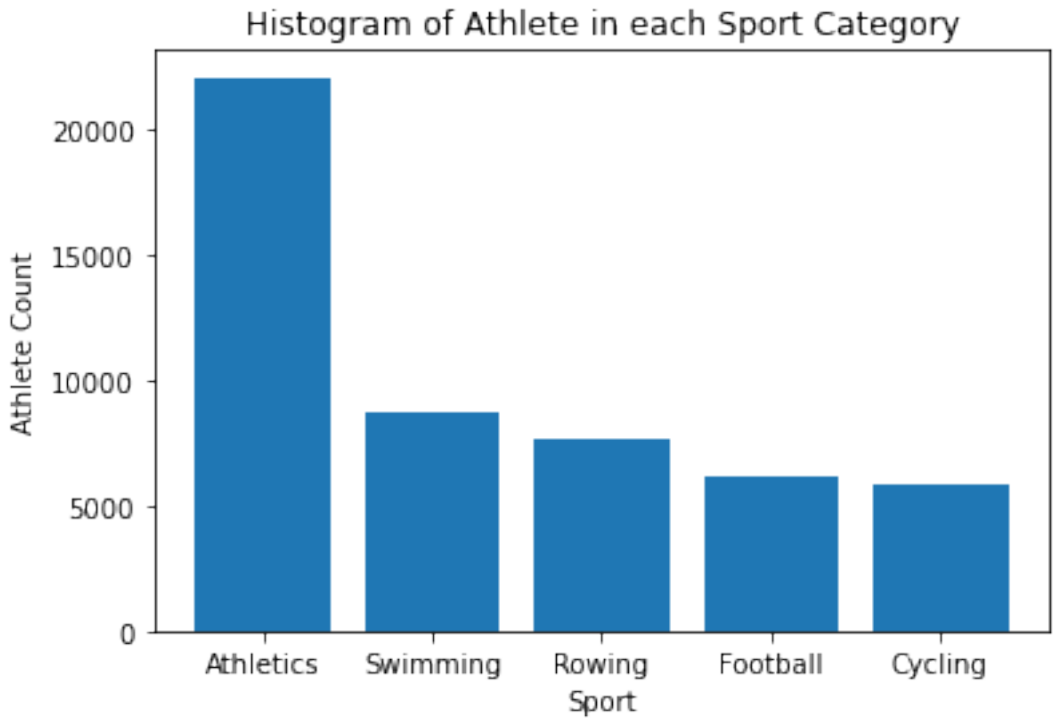
[7]:
```
athlete_count = ps.sqldf(query1, locals())
athlete_count
```

[7]:

	Sport	Athlete Count
0	Athletics	22053
1	Swimming	8761
2	Rowing	7684
3	Football	6161
4	Cycling	5819

[8]:
```
plt.bar(athlete_count['Sport'],athlete_count['Athlete Count'])
plt.xlabel('Sport')
plt.ylabel('Athlete Count')
plt.title('Histogram of Athlete in each Sport Category')
```

[8]:
```
Text(0.5, 1.0, 'Histogram of Athlete in each Sport Category')
```



[9]:
```
query2 = "SELECT Sex, COUNT(Sex) AS Count FROM athlete_events GROUP BY Sex"
```

[10]:
```
sex_count = ps.sqldf(query2,locals())
sex_count
```

[10]:

	Sex	Count
0	F	74522
1	M	196594

[11]:
```
plt.bar(sex_count['Sex'], sex_count['Count'])
plt.xlabel('Sex')
plt.ylabel('Count')
plt.title('Histogram of Sex Distribution')
```

[11]: Text(0.5, 1.0, 'Histogram of Sex Distribution')