Week 2 SportStats Analysis

June 25, 2023

[25]:
```
import pandas as pd
import pandasql as ps
import matplotlib.pyplot as plt
```

[26]:
```
noc_regions = pd.read_csv('Data/noc_regions.csv')
```

[27]:
```
athlete_events = pd.read_csv('Data/athlete_events.csv')
```

[28]:
```
athlete_events.head()
```

[28]:
```
   ID                      Name Sex   Age  Height  Weight            Team  \
0   1                 A Dijiang   M  24.0   180.0    80.0           China
1   2                  A Lamusi   M  23.0   170.0    60.0           China
2   3       Gunnar Nielsen Aaby   M  24.0     NaN     NaN         Denmark
3   4      Edgar Lindenau Aabye   M  34.0     NaN     NaN  Denmark/Sweden
4   5  Christine Jacoba Aaftink   F  21.0   185.0    82.0     Netherlands

   NOC        Games  Year  Season       City          Sport  \
0  CHN  1992 Summer  1992  Summer  Barcelona     Basketball
1  CHN  2012 Summer  2012  Summer     London           Judo
2  DEN  1920 Summer  1920  Summer  Antwerpen       Football
3  DEN  1900 Summer  1900  Summer      Paris     Tug-Of-War
4  NED  1988 Winter  1988  Winter    Calgary  Speed Skating

                              Event Medal
0       Basketball Men's Basketball   NaN
1      Judo Men's Extra-Lightweight   NaN
2           Football Men's Football   NaN
3       Tug-Of-War Men's Tug-Of-War  Gold
4  Speed Skating Women's 500 metres   NaN
```

[29]:
```
noc_regions.head()
```

[29]:
```
   NOC       region                 notes
0  AFG  Afghanistan                   NaN
1  AHO      Curacao  Netherlands Antilles
2  ALB      Albania                   NaN
```

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

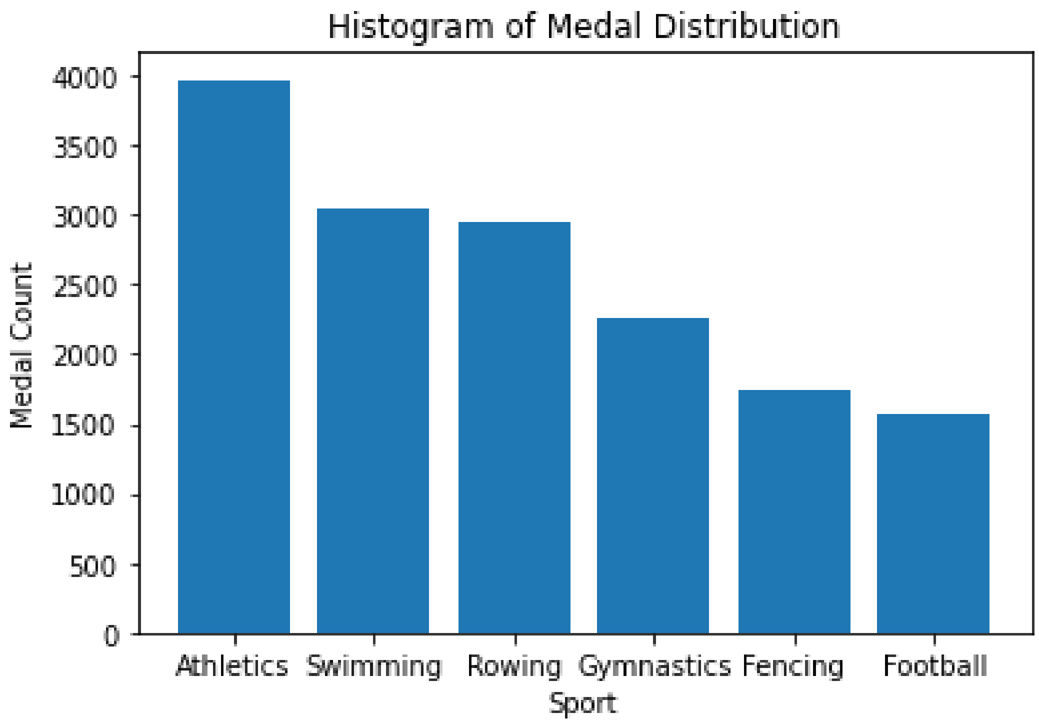
[30]:
```
query1 = "SELECT Sport, SUM(CASE WHEN Medal != 'NaN' THEN 1 ELSE 0 END) AS␣
 ↪'Medal Count' FROM athlete_events GROUP BY Sport ORDER BY COUNT(Medal) DESC␣
 ↪LIMIT 6"
```

[31]:
```
sport_medal_count = ps.sqldf(query1, locals())
sport_medal_count
```

[31]:
```
        Sport  Medal Count
0   Athletics         3969
1    Swimming         3048
2      Rowing         2945
3  Gymnastics         2256
4     Fencing         1743
5    Football         1571
```

[32]:
```
plt.bar(sport_medal_count['Sport'],sport_medal_count['Medal Count'])
plt.xlabel('Sport')
plt.ylabel('Medal Count')
plt.title('Histogram of Medal Distribution')
```

[32]:
```
Text(0.5, 1.0, 'Histogram of Medal Distribution')
```



```
[33]: query2 = "SELECT Name, Sport, MAX(Medal) AS 'Medal Count' FROM (SELECT DISTINCT␣
       ↪Name, Sport, COUNT('Medal') AS Medal FROM athlete_events WHERE Medal !=␣
       ↪'NaN' GROUP BY Name) WHERE Sport = 'Football' OR Sport = 'Swimming' OR Sport␣
       ↪= 'Basketball' GROUP BY Sport ORDER BY MAX(Medal) DESC "
```

```
[34]: player_medal_count = ps.sqldf(query2, locals())
      player_medal_count
```

```
[34]:                               Name       Sport  Medal Count
      0          Michael Fred Phelps, II    Swimming           28
      1                   Teresa Edwards  Basketball            5
      2  Christie Patricia Pearce-Rampone    Football            4
```

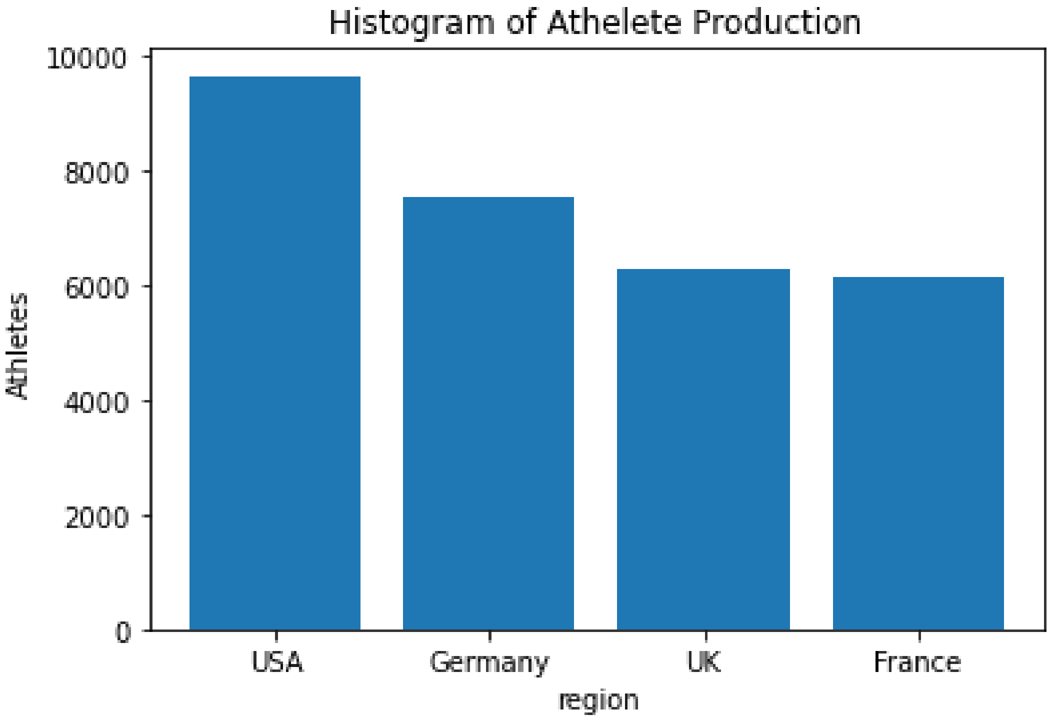
```
[45]: query3 = "SELECT Region, Team, COUNT(DISTINCT Name) AS Athletes FROM␣
       ↪athlete_events INNER JOIN noc_regions ON athlete_events.NOC = noc_regions.
       ↪NOC GROUP BY Region ORDER BY COUNT(DISTINCT Name) DESC lIMIT 4"
```

```
[46]: country_athlete_production = ps.sqldf(query3,locals())
      country_athlete_production
```

```
[46]:     region           Team  Athletes
      0      USA  United States      9652
      1  Germany        Germany      7541
      2       UK  Great Britain      6273
      3   France         France      6161
```

```
[47]: plt.
       ↪bar(country_athlete_production['region'],country_athlete_production['Athletes'])
      plt.xlabel('region')
      plt.ylabel('Athletes')
      plt.title('Histogram of Athelete Production')
```

```
[47]: Text(0.5, 1.0, 'Histogram of Athelete Production')
```


Histogram of Athelete Production
10000
8000
6000
4000
2000
0
Athletes
USA
Germany
UK
France
region

[ ]: