Homework3

October 7, 2024

1 Homework 3 -OLYMPICS

2 Student name: Ramya Chowdary Patchala

3 Assignment submisision date: 9/27/2024

[2]:
```
#importing necessary libraries
import pandas as pd
import numpy as np
import matplotlib
import matplotlib.pyplot as plt
```

[3]:
```
# NOTE: Please make sure that the olympics1992_2008.zip file has been uploaded
  ↪to the same folder where you have this
#       notebook file
odata = pd.read_csv('olympics1992_2008.zip',skiprows=4)
```

[4]:
```
# Start exploratory data analysis
odata.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9619 entries, 0 to 9618
Data columns (total 10 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   City          9619 non-null   object
 1   Edition       9619 non-null   int64
 2   Sport         9619 non-null   object
 3   Discipline    9619 non-null   object
 4   Athlete       9619 non-null   object
 5   NOC           9619 non-null   object
 6   Gender        9619 non-null   object
 7   Event         9619 non-null   object
 8   Event_gender  9619 non-null   object
 9   Medal         9619 non-null   object
dtypes: int64(1), object(9)
memory usage: 751.6+ KB
```

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

[5]:
```
        City  Edition     Sport Discipline            Athlete  NOC Gender  \
0  Barcelona     1992  Aquatics     Diving          XIONG, Ni  CHN    Men
1  Barcelona     1992  Aquatics     Diving        SUN, Shuwei  CHN    Men
2  Barcelona     1992  Aquatics     Diving    DONIE, Scott R.  USA    Men
3  Barcelona     1992  Aquatics     Diving  CLARK, Mary Ellen  USA  Women
4  Barcelona     1992  Aquatics     Diving        FU, Mingxia  CHN  Women

           Event Event_gender   Medal
0  10m platform            M  Bronze
1  10m platform            M    Gold
2  10m platform            M  Silver
3  10m platform            W  Bronze
4  10m platform            W    Gold
```

[ ]:
```
# Add cells with any additional exploratory data analysis commands/functions␣
 ↪that you think are necessary. This will
# not be graded but will help you in solving this homework's tasks
# Hint.. get the unique entries for columns of interest
```

[11]:
```
# Different sports and no.of entries
odata.value_counts("Sport")
```

[11]:
```
Sport
Aquatics            1498
Athletics            902
Rowing               732
Hockey               481
Gymnastics           478
Football             455
Handball             443
Canoe / Kayak        420
Volleyball           407
Basketball           358
Baseball             335
Cycling              324
Fencing              321
Wrestling            293
Judo                 280
Sailing              261
Boxing               232
Shooting             231
Equestrian           227
Weightlifting        194
Softball             180
Archery              120
```

```
Badminton             120
Table Tennis          102
Tennis                 94
Taekwondo              80
Modern Pentathlon      33
Triathlon              18
Name: count, dtype: int64
```

[10]:
```
# Different Gender
odata.value_counts("Gender")
```

[10]:
```
Gender
Men      5522
Women    4097
Name: count, dtype: int64
```

[12]:
```
# Medal count values
odata.value_counts("Medal")
```

[12]:
```
Medal
Bronze    3304
Gold      3164
Silver    3151
Name: count, dtype: int64
```

[14]:
```
# Different countries
odata.value_counts("NOC")
```

[14]:
```
NOC
USA    1311
GER     691
AUS     678
RUS     638
CHN     550
       …
SRI       1
KUW       1
MKD       1
MRI       1
AFG       1
Name: count, Length: 116, dtype: int64
```

Solve the following tasks. You can add as many additional cells as you need to solve each one of them.

3.1 Task #1

a) List the 5 countries that accumulated the most medals across all the olympic game editions covered in the dataset

b) List the 5 countries that accumulated the most GOLD medals across all the olympic game editions covered in the dataset

[16]:
```
# Finding 5 countries that accumulated most medals

# Checking for na values in medals
na_count_medal = odata['Medal'].isna().sum()

# Display the count of NaNs
print(f'Number of NaNs in medal: {na_count_medal}')

# Since there are no na's in medals, we can find top 5 countries from␣
 ↪value_counts
odata.value_counts("NOC").head(5)
```

```
Number of NaNs in medal: 0
```

[16]:
```
NOC
USA    1311
GER     691
AUS     678
RUS     638
CHN     550
Name: count, dtype: int64
```

[17]:
```
# Finding countries that accumulated most Gold Medals

# Filtering based on Gold Medals
gold_df = odata[odata['Medal'] == 'Gold']
gold_df.value_counts("NOC").head(5)
```

[17]:
```
NOC
USA    620
GER    237
CHN    202
RUS    192
AUS    186
Name: count, dtype: int64
```

3.2 Task #2

List the number of Gold, Silver and Bronze medals obtained by Women and Men across all the olympic game editions covered in the dataset

[29]:
```
# Grouping data based on Medals and Gender
grp_df = odata.groupby(['Medal', 'Gender'])

# We can no.of medals obtained by that gender
count = grp_df.size()
print(count)
```

```
Medal   Gender
Bronze  Men       1918
        Women     1386
Gold    Men       1807
        Women     1357
Silver  Men       1797
        Women     1354
dtype: int64
```

3.3 Task #3

List the names of the 5 male athletes and 5 female athletes that obtained the most medals across all the olympic game editions covered in the dataset

[39]:
```
# Obtaining male athletes data
male = odata[odata['Gender']=='Men']

# Obtaining female athletes data
female = odata[odata['Gender']=='Women']

# Obtaining top five male athletes names
print('Male : ', male['Athlete'].value_counts().head(5))

# Obtaining top five female athletes names
print('\n Female : ', female['Athlete'].value_counts().head(5))
```

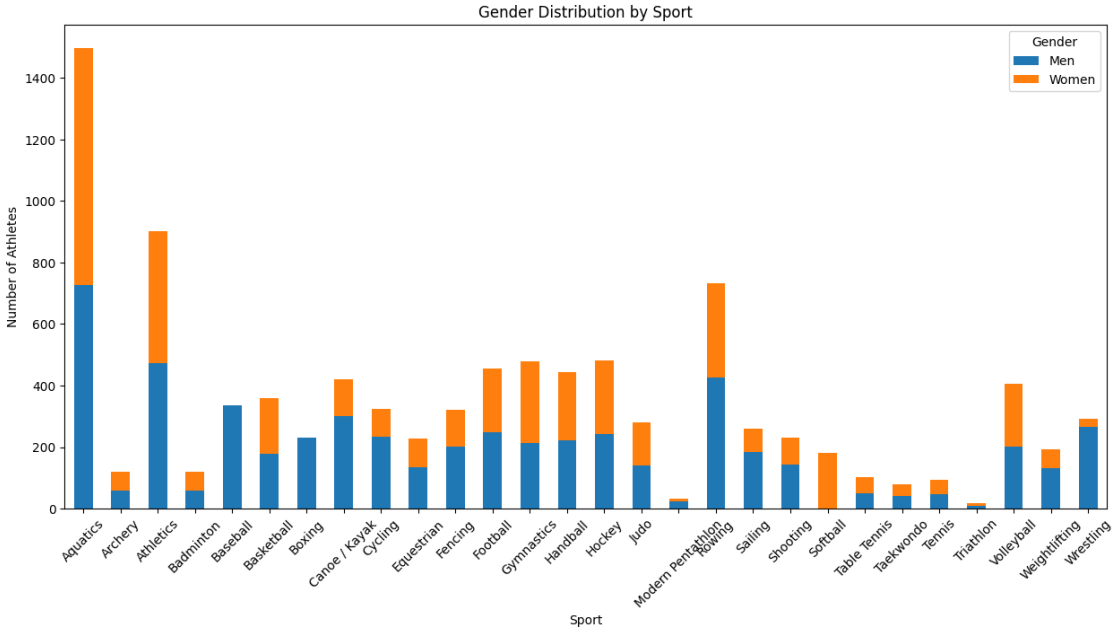
```
Male :  Athlete
PHELPS, Michael     16
NEMOV, Alexei       12
SCHERBO, Vitaly     10
HALL, Gary Jr.      10
POPOV, Alexander     9
Name: count, dtype: int64

 Female :  Athlete
THOMPSON, Jenny           12
COUGHLIN, Natalie         11
VAN ALMSICK, Franziska    10
TORRES, Dara               9
THOMAS, Petria             8
Name: count, dtype: int64
```

3.4 Task #4

Provide two additional analysis results that you can derive from the dataset (they must be different than those obtained in tasks 1 to 3). The results can include graphs (but it is not required). Describe the results obtained in the cell provided for that purpose

[52]:
```
#Correlation Between Sports and Gender
gender_sport_counts = odata.groupby(['Sport', 'Gender']).size().
 ↪unstack(fill_value=0)
gender_sport_counts.plot(kind='bar', stacked=True, figsize=(15, 7))
plt.title('Gender Distribution by Sport')
plt.xlabel('Sport')
plt.ylabel('Number of Athletes')
plt.xticks(rotation=45)
plt.legend(title='Gender')
plt.show()
```



[59]:
```
# Best-Performing Countries by Event Based on Medal Count

# Group by NOC (country) and Event, then count the number of medals
medals_by_event_country = odata.groupby(['NOC', 'Event']).size().
 ↪reset_index(name='Medal Count')

# Find the country with the maximum medals for each event
best_performing_countries = medals_by_event_country.loc[medals_by_event_country.
 ↪groupby('Event')['Medal Count'].idxmax()]
```

```
# Sort by Medal Count in descending order
best_performing_countries = best_performing_countries.sort_values(by='Medal␣
 ↪Count', ascending=False)

# Display the results
display(best_performing_countries)
```

```
      NOC                              Event  Medal Count
2286  USA                         basketball          120
93    AUS                             hockey          113
547   CUB                           baseball          111
2234  USA                 4x100m medley relay           79
240   BRA                           volleyball           72
…     …                                    …          …
420   CHN  Laser Radial - One Person Dinghy             1
403   CHN  59 - 64kg, total (featherweight)            1
272   BUL   70 - 76kg, total (middleweight)            1
834   FRA     71 - 78kg (half-middleweight)            1
985   GEO                           85 - 97kg           1

[277 rows x 3 columns]
```

RESULTS ANALYSIS Summarize your findings here.

3.4.1 In the first analysis of correleation between the sports and gender, we can see some games have very high male participation than the female participation like Baseball and Boxing etc.

3.4.2 In the second analysis which is identifying countries played best in that events. We can see that USA has 120 Medals in basketball, and Australia has 113 Medals in Hockey etc.

4 PART 2

Explore the dataset of Paris Olympics medallists. See https://www.kaggle.com/datasets/piterfm/paris-2024-olympic-summer-games?select=medallists.csv for more information. Describe your exploration and observations of **TWO** unique findings.

[63]:
```
paris_df = pd.read_csv('Parismedallists.csv')
paris_df.head()
```

[63]:

	medal_date	medal_type	medal_code	name	gender	country_code	\
0	2024-07-27	Gold Medal	1.0	EVENEPOEL Remco	Male	BEL	
1	2024-07-27	Silver Medal	2.0	GANNA Filippo	Male	ITA	
2	2024-07-27	Bronze Medal	3.0	van AERT Wout	Male	BEL	
3	2024-07-27	Gold Medal	1.0	BROWN Grace	Female	AUS	
4	2024-07-27	Silver Medal	2.0	HENDERSON Anna	Female	GBR	

```
         country   country_long nationality_code    nationality  … team  \
0        Belgium        Belgium              BEL        Belgium  …  NaN
1          Italy          Italy              ITA          Italy  …  NaN
2        Belgium        Belgium              BEL        Belgium  …  NaN
3      Australia      Australia              AUS      Australia  …  NaN
4  Great Britain  Great Britain              GBR  Great Britain  …  NaN

  team_gender    discipline                          event event_type  \
0         NaN  Cycling Road    Men's Individual Time Trial        ATH
1         NaN  Cycling Road    Men's Individual Time Trial        ATH
2         NaN  Cycling Road    Men's Individual Time Trial        ATH
3         NaN  Cycling Road  Women's Individual Time Trial        ATH
4         NaN  Cycling Road  Women's Individual Time Trial        ATH

                                           url_event  birth_date code_athlete  \
0  /en/paris-2024/results/cycling-road/men-s-indi…  2000-01-25      1903136
1  /en/paris-2024/results/cycling-road/men-s-indi…  1996-07-25      1923520
2  /en/paris-2024/results/cycling-road/men-s-indi…  1994-09-15      1903147
3  /en/paris-2024/results/cycling-road/women-s-in…  1992-07-07      1940173
4  /en/paris-2024/results/cycling-road/women-s-in…  1998-11-14      1912525

   code_team is_medallist
0        NaN         True
1        NaN         True
2        NaN         True
3        NaN         True
4        NaN         True

[5 rows x 21 columns]
```

[64]:
```
paris_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2315 entries, 0 to 2314
Data columns (total 21 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   medal_date        2315 non-null   object
 1   medal_type        2315 non-null   object
 2   medal_code        2314 non-null   float64
 3   name              2315 non-null   object
 4   gender            2315 non-null   object
 5   country_code      2315 non-null   object
 6   country           2315 non-null   object
 7   country_long      2315 non-null   object
 8   nationality_code  2314 non-null   object
```

```
 9   nationality       2314 non-null   object
 10  nationality_long  2314 non-null   object
 11  team              1555 non-null   object
 12  team_gender       1555 non-null   object
 13  discipline        2315 non-null   object
 14  event             2315 non-null   object
 15  event_type        2315 non-null   object
 16  url_event         2294 non-null   object
 17  birth_date        2315 non-null   object
 18  code_athlete      2315 non-null   int64
 19  code_team         1555 non-null   object
 20  is_medallist      2315 non-null   bool
dtypes: bool(1), float64(1), int64(1), object(18)
memory usage: 364.1+ KB
```

[75]:
```
paris_df.describe()
```

[75]:
```
        medal_code  code_athlete
count  2314.000000  2.315000e+03
mean      2.023336  1.893321e+06
std       0.820390  2.628276e+05
min       1.000000  1.532872e+06
25%       1.000000  1.896552e+06
50%       2.000000  1.924464e+06
75%       3.000000  1.950498e+06
max       3.000000  4.980004e+06
```

[76]:
```
paris_df['event_type'].value_counts()
```

[76]:
```
event_type
HTEAM    843
TEAM     638
ATH      494
HATH     266
HCOUP     42
COUP      32
Name: count, dtype: int64
```

[79]:
```
paris_df['is_medallist'].value_counts()
```

[79]:
```
is_medallist
True     2268
False      47
Name: count, dtype: int64
```

[80]:
```
# Top 5 Countries Winning Most Gold Medals in Couple Events
```

```
# Filter the dataset for 'COUP' events (you can replace 'COUP' with your␣
 ↪specific couple event type)
couple_events = paris_df[paris_df['event_type'] == 'COUP']

# Further filter for Gold medals
gold_couples = couple_events[couple_events['medal_type'] == 'Gold Medal']

# Count the number of Gold medals by country
top_countries = gold_couples['country'].value_counts().head(5)

# Display the top 5 countries
print("Top 5 countries with the most Gold medals in couple events:")
print(top_countries)
```

```
Top 5 countries with the most Gold medals in couple events:
country
China          6
Germany        2
New Zealand    2
Name: count, dtype: int64
```

[81]:
```
# Group by country and count the number of each type of medal
medal_counts = paris_df.groupby('country')['medal_type'].value_counts().
 ↪unstack(fill_value=0)

# Filter countries based on the condition: bronze > silver > gold
result = medal_counts[(medal_counts['Bronze Medal'] > medal_counts['Silver␣
 ↪Medal']) &
                      (medal_counts['Silver Medal'] > medal_counts['Gold␣
 ↪Medal'])]

# Display the result
print("Countries that won more Bronze than Silver and more Silver than Gold:")
print(result)
```

```
Countries that won more Bronze than Silver and more Silver than Gold:
medal_type           Bronze Medal  Gold Medal  Silver Medal
country
Argentina                      16           1             2
Brazil                         35           4            28
Great Britain                  80          40            42
India                          21           0             1
Kazakhstan                      4           1             3
Kyrgyzstan                      4           0             2
Lithuania                       5           0             2
Republic of Moldova             3           0             1
South Africa                   14           1             7
```

Switzerland	7	1	2
Türkiye	7	0	4

Sure! Here's the analysis written in a more conversational tone:

4.0.1 Analysis of Medal Winners in Couple Events

Top 5 Countries Winning the Most Gold Medals in Couple Events In our look at gold medals awarded in couple events, the following countries topped the list:

1. **China**: 6 Gold Medals
2. **Germany**: 2 Gold Medals
3. **New Zealand**: 2 Gold Medals

Observations: - **China** clearly stands out as the top performer in couple events, with a total of 6 gold medals, which is significantly higher than any other country. - **Germany** and **New Zealand** both earned 2 gold medals each, showing that they are competitive in this area, but they still have a long way to go to catch up with China.

This shows that China has a strong focus and investment in couple events, which likely contributes to their athletes' success.

4.0.2 Countries with More Bronze Medals than Silver and More Silver than Gold

We also found some interesting trends among countries that have won more bronze medals than silver medals and more silver medals than gold medals. Here are those countries:

- **Argentina**: 16 Bronze, 2 Silver, 1 Gold
- **Brazil**: 35 Bronze, 28 Silver, 4 Gold
- **Great Britain**: 80 Bronze, 42 Silver, 40 Gold
- **India**: 21 Bronze, 1 Silver, 0 Gold
- **Kazakhstan**: 4 Bronze, 3 Silver, 1 Gold
- **Kyrgyzstan**: 4 Bronze, 2 Silver, 0 Gold
- **Lithuania**: 5 Bronze, 2 Silver, 0 Gold
- **Republic of Moldova**: 3 Bronze, 1 Silver, 0 Gold
- **South Africa**: 14 Bronze, 7 Silver, 1 Gold
- **Switzerland**: 7 Bronze, 2 Silver, 1 Gold
- **Türkiye**: 7 Bronze, 4 Silver, 0 Gold

Observations: - These countries show an interesting pattern where they have more bronze medals than silver and more silver than gold. - For example, **Argentina** and **Brazil** have a lot of bronze medals, which suggests they are regularly competitive, even if they're not winning as many gold medals. - **Great Britain** has a really high count of bronze medals but also performs well with silver and gold medals, indicating they have a balanced approach across different events. - Countries like **India**, **Kyrgyzstan**, and **Lithuania** show that they might be strong in certain events where they are getting bronze but aren't securing many gold medals.

Overall, this analysis gives us a clearer picture of how different countries perform in couple events, highlighting unique strengths and patterns that could help them improve in future competitions.