Data analysis of number of medalists in Summer Olympic

Edward Kim

December 2019

1 Code

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

def men(a):
    male = a['Sex'] == 'M'
    return a[male]

def women(a):
    female = a['Sex'] == 'F'
    return a[female]

def winter(a):
    winter = a['Season'] == 'Winter'
    return a[winter]

def summer(a):
    summer = a['Season'] == 'Summer'
    return a[summer]

def medal_only(a):
    return pd.notnull(a["Medal"])

def grp_yr_cnt(a):
    return a.groupby("Year").count()

def no_of_sport(a):
    a = a[["Year","Sport"]]
    x = a.drop_duplicates(["Year","Sport"])
    x1 = x.groupby("Year").size()
    return x1

def graph_plot1(x):
    for i in range(0,len(x)):
        plt.subplot(2,2,i+1)

        if (i+1) == 1:
            plt.title('Figure 1(a). Number of medalists in Summer')
            plt.ylabel('No. of medals')
        elif (i+1) == 2:
            plt.title('Figure 1(b). Number of medals in Winter')
        elif (i+1) == 3:
            plt.ylabel('No. of olympian')
            plt.xlabel('Year')
            plt.title('Figure 2(a). Number of olympian in Summer')
        elif (i+1) == 4:
            plt.xlabel('Year')
            plt.title('Figure 2(b). Number of olympian in Winter')
        for j in range(0,len(x[0])):
            plt.plot(x[i][j])
        plt.legend(['Male','Female'],loc="upper right")
    plt.savefig('graph1')

def graph_plot2(x):
    for i in range(0,len(x)):
        plt.subplot(2,1,i+1)
```

```
        if (i+1) == 1:
            plt.title('Figure 3(a). Number of different type of sports played in Winter
        else:
            plt.title('Figure 3(b). Number of different type of sports played in Summer
            plt.xlabel('Year')
        plt.ylabel('No. of sports')
        for j in range(0,len(x[0])):
            plt.plot(x[i][j])
        plt.legend(['Male','Female'],loc="upper left")
    plt.savefig('graph2')

def cnt_no_olympian(x):
    x = x.drop_duplicates(subset=['ID','Year'])
    x = grp_yr_cnt(x)
    return x['ID']

def cnt_no_medal(x):
    x = x.drop_duplicates(subset=["Year","Medal","Event"]) #remove duplicates
    return x[medal_only(x)]

if __name__ == '__main__':

    #read excel file
    data = pd.read_csv("C:\\shkim\\Internship\\nPlan\\data\\athlete_events.csv")

    #Data extraction and simplification

    #Display coloumns that are only necessary
    gb = data["NOC"]=="GBR" #Extract players who are assigned to GBR only
    data_gb = data[gb]

    #remove players playing more than once in same olympic
    data_gb_wt = winter(data_gb) #Seperate it by Season
    data_gb_sm = summer(data_gb)

    data_gb_wt_m = men(data_gb_wt) #Seperate it by Gender
    data_gb_wt_f = women(data_gb_wt)
    data_gb_sm_m = men(data_gb_sm)
    data_gb_sm_f = women(data_gb_sm)

    #Calculate a number of different type of sports played and assign a variable
    data_gb_wt_m_sport = no_of_sport(data_gb_wt_m)
    data_gb_wt_f_sport = no_of_sport(data_gb_wt_f)
    data_gb_sm_m_sport = no_of_sport(data_gb_sm_m)
    data_gb_sm_f_sport = no_of_sport(data_gb_sm_f)

    #Narrow the data to medalists only
    data_gb_wt_m_medal = cnt_no_medal(data_gb_wt_m)
    data_gb_wt_f_medal = cnt_no_medal(data_gb_wt_f)
    data_gb_sm_m_medal = cnt_no_medal(data_gb_sm_m)
    data_gb_sm_f_medal = cnt_no_medal(data_gb_sm_f)

    m_medal_summer = grp_yr_cnt(data_gb_sm_m_medal) #Group the data by Year
    m_medal_winter = grp_yr_cnt(data_gb_wt_m_medal)
    f_medal_summer = grp_yr_cnt(data_gb_sm_f_medal)
    f_medal_winter = grp_yr_cnt(data_gb_wt_f_medal)

    m_summer = cnt_no_olympian(data_gb_sm_m)#Count the number of olympian
```

```
m_winter = cnt_no_olympian(data_gb_wt_m)
f_summer = cnt_no_olympian(data_gb_sm_f)
f_winter = cnt_no_olympian(data_gb_wt_f)

#graph plotting
#Take out only necessary column for the calculation
a1 = m_medal_summer['Medal']
a2 = m_medal_winter['Medal']
b1 = f_medal_summer["Medal"]
b2 = f_medal_winter["Medal"]

z1 = data_gb_wt_m_sport
z2 = data_gb_wt_f_sport
z3 = data_gb_sm_m_sport
z4 = data_gb_sm_f_sport

x = [[a1,b1],[a2,b2],[m_summer,f_summer],[m_winter,f_winter]]
y = [[z1,z2],[z3,z4]]
plt.figure(figsize = (12,16))
plt.figure(1)
graph_plot1(x)
plt.figure(figsize = (12,16))
plt.figure(2)
graph_plot2(y)

plt.show()
```

2 Figures

Figure 1. Number of medalists in Summer

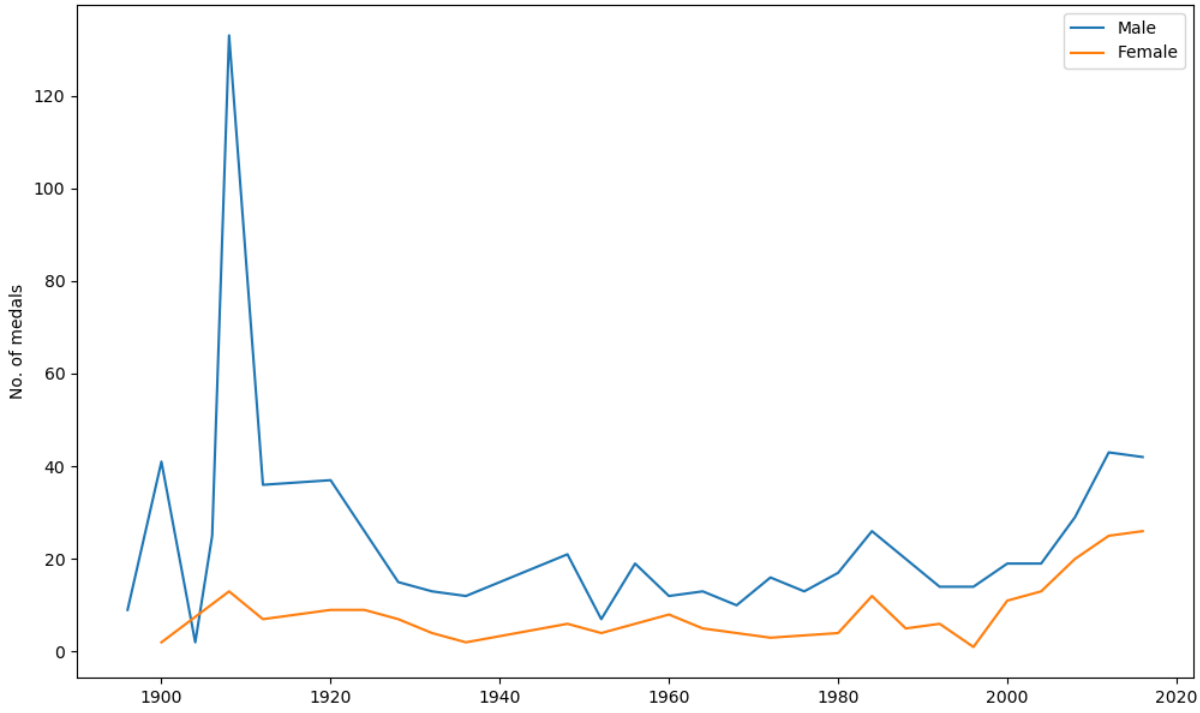


Figure 2. Number of olympian in Summer

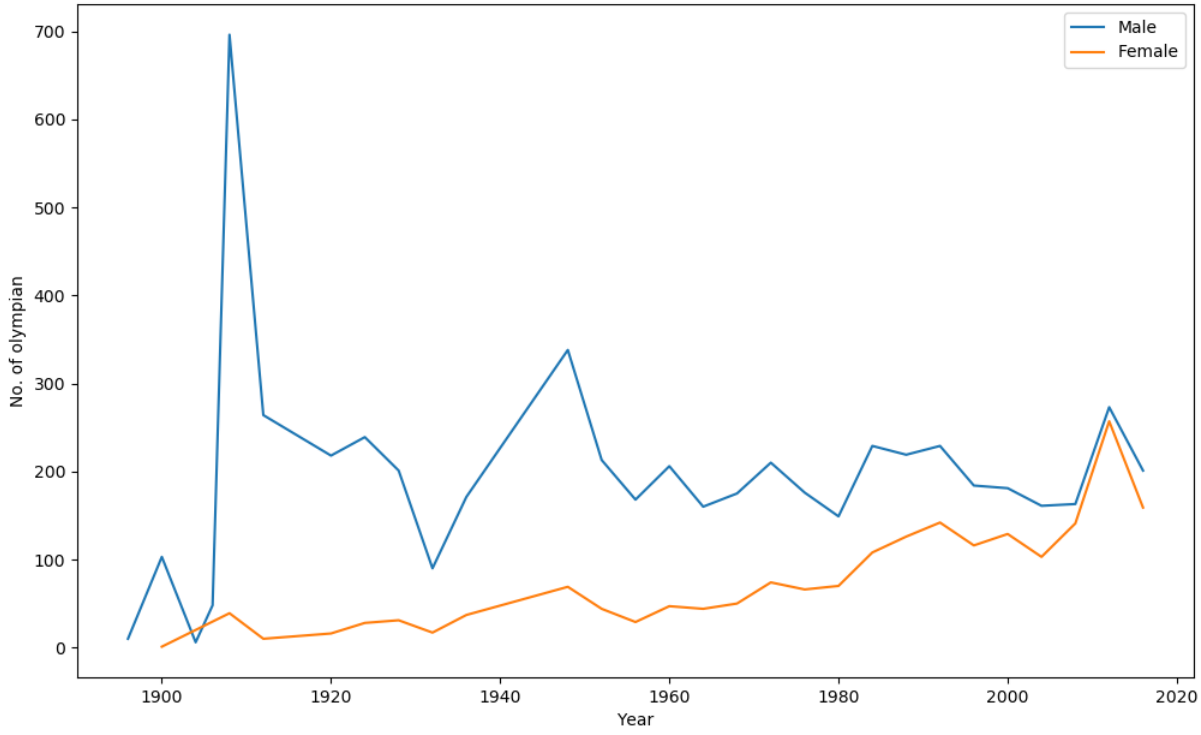
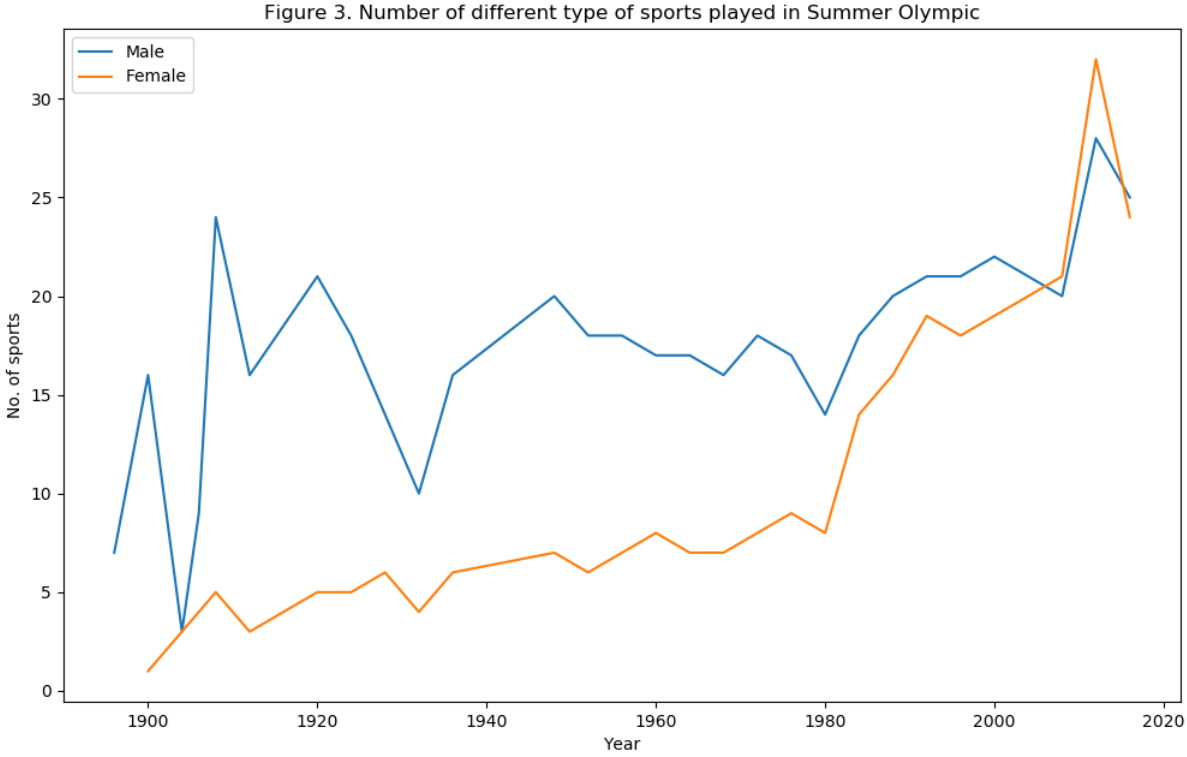


Figure 3. Number of different type of sports played in Summer Olympic



3 Introduction

Using the data of the Olympic history, I compared and analysed the data between the number of female and male medalists throughout each Olympic. Since, the original data is very large and I wanted to increase the relevance of the results for readers, I decided to extract the data for British players only. Then, I classified the data by 'Gender' and 'Season' to generate graphs. The reason why I did not include the data for Winter Olympic is that the number of medals and Olympians are too few that the data fluctuates abruptly which makes difficult to analyse the data properly.

4 Data Analysis

4.1 Main

The graph in fig1 shows the number of medals achieved by British Olympians throughout each Olympic. As fig1 indicates, the number of men medalists are generally higher than the women. The index for men has a peak in 1908 followed by a sudden drop and then a slow recovery. On the other hand, the index for women is initially very low compare to the men's index, but it gradually increases throughout the year. Also, throughout the each Olympic, the gap between the number of men and women medalists tend to decreases. To comprehend this trend of the graph further, I have also come up with other figures.

4.2 Others

Fig2 shows the number of Olympian for British players participated. The graph generally looks very similar to the fig1. For example, the peak and a huge drop followed by a slow recovery are similar to the men's index in fig1. Furthermore, the index for women is small initially but gradually increases which are also akin to the trend of women's index in fig1. It shows that the number of medalists depend on the number of Olympians. Hence, this factor may have contributed to the increase in the number of women's medalist and the reduction in the gap between the men and women's index in fig1 as well. Fig3 shows the number of different type of sports played by British players in each Olympic. The women initially has much lower index compare to the men. However, it constantly increases and catches up the men's index in the end which explains the decrease in the gap between the men and women in fig1. The trend of women's index in fig3 is also similar to the women's index in fig1 which indicates that the number of different type of sports is also a factor of the number of medals.

5 Evaluation

There is a slight difference between the values of fig1 and the actual values. In order to measure the number of medalists, I only extracted data with the value of 'GBR' in NOC column. This was done to include medals for British players playing in a mixed team. For example, in 1896, a British Olympian, 'John Pius Boland' won a 1st place in 'Tennis Men's Doubles', but it was not recorded on the British team since he played in a mixed team with 'Fritz Traun' who is a German Tennis player. Hence, to include all these occurrences, I extracted data by selecting the data with 'GBR' for NOC. Furthermore,

6 Conclusion

This research carries some limitations too. Throughout the Olympic, there are many sports that have been excluded and included, so this also have contributed to the result of figure1. For example, in the early 1900's there were sports such as 'tug-of-war', 'Racquets' and 'Polo' which British players are very strong at, but since these sports were excluded, it led to the decline as fig1 shows. To conclude, the number of medals achieved by female player is increasing and the gap between the men are reducing Which are affected by the factors I have explained such as the number of Olympians and the number of different type of sports played.