SI618 Project 2

Analysis of Olympic History Data

Joshua Chang, joshch@umich.edu December 7, 2021

Motivation

There have always been stereotypes such as African athletes tend to win more medals in track & field events due to their athleticism or European and North American countries have better resources to train athletes, so these countries tend to win more medals than others.

In part one, I intended to find out the correlation between a country's GDP and its chances of winning medals. Here, in part two, my goal is to gain further insights from the dataset by using exploratory data analysis, and hopefully find out whether the stereotypes I mentioned above are true or not.

To be more specific, these are the following questions I would like to answer:

1. What is the distribution of the total medals won by different countries? Is it true that countries from Europe or North America tend to win more medals?
2. Does the stereotype that Asian athletes are less likely to win a medal in athletics (track & field) hold true? What about African athletes? Do athletes from Africa have a higher chance of winning medals in athletics? What about other sports?
3. In the past, females were not encouraged to do sports due to many different historical and stereotypical factors. However, were there more and more women athletes rising and shining in the Olympics games? Moreover, do countries with better economic levels tend to send more female athletes to compete?

Datasets

1. 120 Years of Olympic History

This is a historical dataset about the modern Olympic Games from 1896 to 2016, scraped by Randi H Griffen, and can be downloaded as a CSV file, which contains different athletes' personal information, such as their NOC code(National Olympic Committee 3 letter code), gender, year, season, the sport they play, and their outcomes. The raw dataset has 271116 rows and 15 columns, and the variables of interest are Name, Sex, NOC, Year, and Medal. Here I did not use the Team column for two reasons. First, during 1970-2016, there were unified teams that were mostly joined by the Soviet Union. Hence, using the country code would be better to classify where these athletes are really from. I accomplished this by joining the dataset with *"**NOC Region Dataset**"* on the NOC column. For more information about the Unified Team at the Olympics, go check out https://en.wikipedia.org/wiki/Unified_Team_at_the_Olympics.

The URL link of this dataset can be found here:
https://www.kaggle.com/heesoo37/120-years-of-olympic-history-athletes-and-results?select=athlete_events.csv

2. **NOC Region Dataset**

This dataset is about the NOC code that matches the full name of each country. This allows us to add a full country name column to the "***120 Years of Olympic History***" dataset by matching the NOC column.

The raw dataset has 230 rows and 3 columns, and the variables of interest are NOC and Region. One more thing to notice here is that the full country names of this dataset do not completely match the ones in the "***Countries of the World***" dataset, additional data cleaning and manipulation would be covered in the Methods section.

The URL link of this dataset can be found here:
https://www.kaggle.com/heesoo37/120-years-of-olympic-history-athletes-and-results?select=noc_regions.csv

3. **Countries of the World**

The data is compiled from the CIA website by the US government. It shows world facts on the average of the population, region, area size, infant mortality, GDP, and literacy rate, etc for different countries. In this project, I would use this as an indicator of different countries' profiles.

The raw dataset has 227 rows and 21 columns, and the variables of interest are Country, Population, GDP.

The URL link of this dataset can be found here:
https://www.kaggle.com/fernandol/countries-of-the-world?select=countries+of+the+world.cs

Methods

I did two joins to gain the dataframe ***join2*** for further analysis for the three main questions. I used a left join to combine ***athleteEvents*** with ***nocRegions*** on the "NOC" column, and gained a joined dataset called ***join1***. Next, I used a left join again to join the countries dataframe with ***join1*** on the column "Country". However, during this process, I found out that there were white spaces in the values of "Country" column in both ***countries*** and ***nocRegions*** dataframes. I used .str.strip() to remove the white spaces on both ends. Besides, I also found out that although most of the values in the

two “Country” columns were the same name, there were some names that needed to be revised to match. After cleaning up all the messy data and conducting two joins, I formed the dataframe ***join2*** that has 271116 rows and 39 columns, which allows me to begin my exploratory data analysis.

Question1: What is the distribution of the total medals won by different countries? Is it true that countries from Europe or North America won more medals in the past?

First, I selected the columns "Country", "Region", "Season", "Medal", and created another “Total Medals” column with 1 that represents an athlete has won a medal, 0 that means an athlete has not won a medal. Next, I dropped all the rows that contain NaN values in the “Medal” column, so I could focus on the data that won medals. At last, I found out there were whitespaces in the “Region” column, hence I used str.strip() to clean them up. The biggest challenge that I encountered was trying to deal with white spaces for making the last join that I mentioned above, other than that there were no big challenges faced here.

Question 2: Does the stereotype that Asian athletes are less likely to win a medal in athletics (track & field) hold true? What about African athletes? Do they have higher chances of winning medals in athletics? What about other sports?

First, I selected the columns "Country", "Region", "Medal" that I want for this question, and replaced all the NaN values in column “Medal” with “DidNotWinMedals”. Also, there were whitespaces in the “Region” column, hence I used str.strip() to clean them up. After doing this, the data were prepared for analysis.

Question3: Were there more and more female athletes participating in the Olympics games? Moreover, do countries with better economic levels tend to send more female athletes?

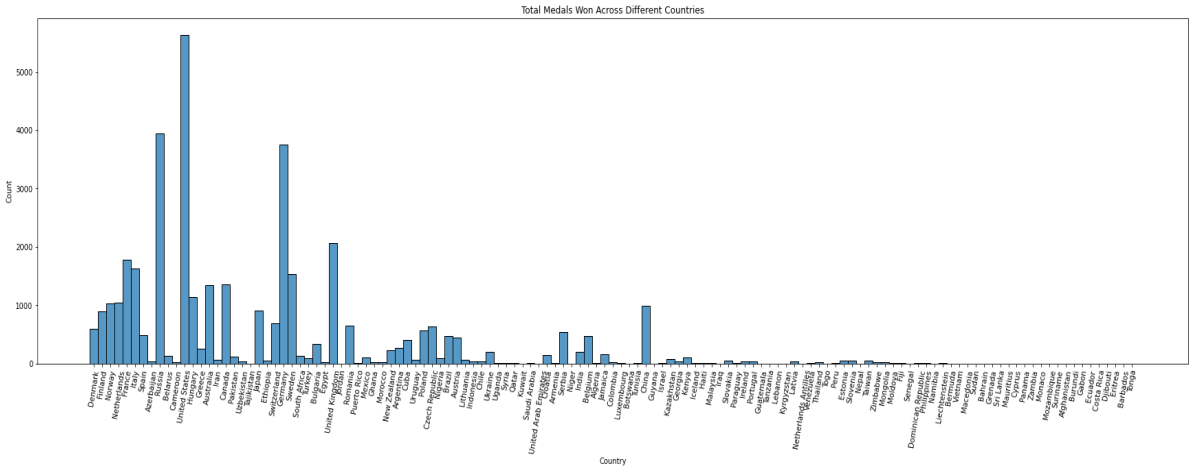
As above, I selected "Country","Region","Sex","Season", "Medal", 'GDP ($ per capita)', 'Year' for the columns of dataframe ***Q3***. After doing this, the data were prepared for analysis. There were no big issues handling data here.

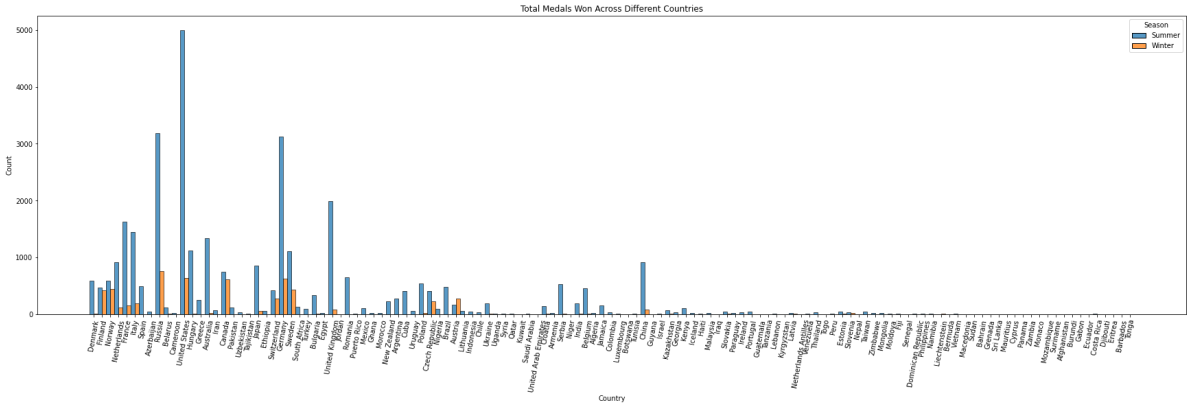
Analysis and Results

Question1: What is the distribution of the total medals won by different countries? Is it true that countries from Europe or North America tend to win more medals?

From part 1, the data shows that European countries and North American countries tend to have higher Total GDP with more winning medals as well. Here, the purpose of this question is that I would like to use exploratory data analysis to find out whether these countries do win more medals, and try to visualize them.

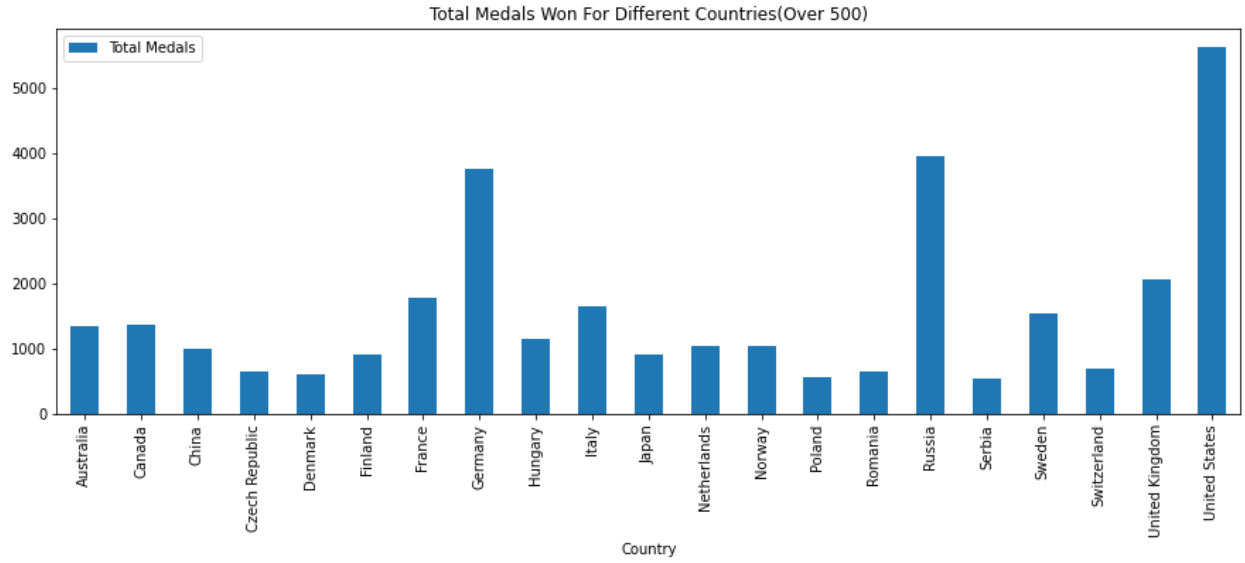
First, in addition to extracting the columns that I want from ***join2***, I created a new column “Total Medals” to show whether an athlete wins gold medals or not. Next, I dropped all the rows that contain “NaN” in the medal column and used histplot to count the total medals won by each country. After using plt.figure() and plt.xticks(), I gained two graphs with one containing the medals altogether, the other shows summer and winter medals respectively:



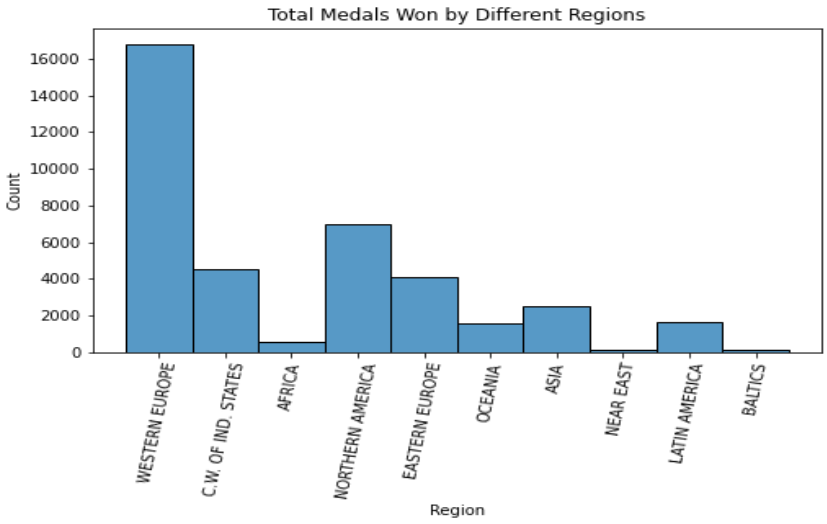


We can see that the top 5 countries are Russia, the United States, Germany, the United Kingdom, and France in both conditions (containing winter-season medals or not). However, the visualization was hard to look at, so I tried to narrow it down to the

countries that have won over 500 medals to give a clearer graph down below:

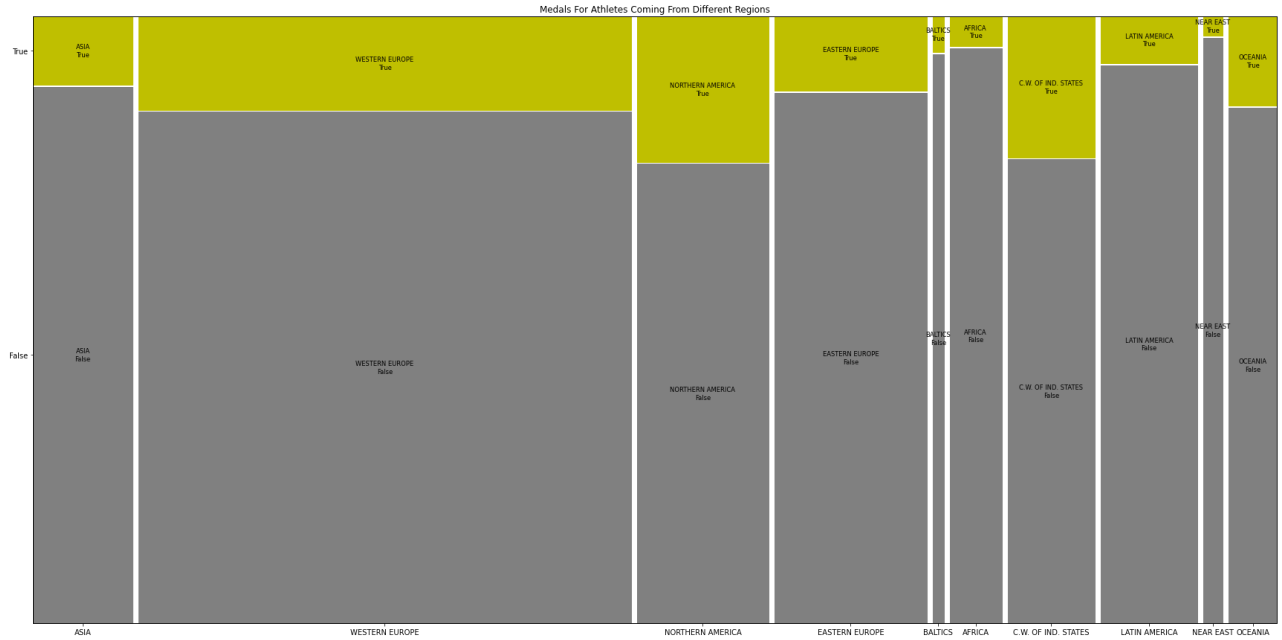


As far as the distribution of total medals for different regions is concerned, I created another WinMedalsOrNot column that shows True to represent an athlete who has won a medal, and false representing an athlete who did not. Also, the region names in the original dataset were too long, I revised it and then used replace() to combine 'NORTHERN AFRICA' and 'SUB-SAHARAN AFRICA' to ‘Africa’. Then repeated the steps I used in histplot to gain a graph down below:



We can only see the medals won by different regions over the years, they are highly centered in North American and European countries. However, this does not mean these regions are “more likely” to win medals because we did not take how many countries or how many athletes into account. Hence, I decided to use a **mosaic plot** to

see what the proportion of winning and not winning medals looks like for different regions, and gained a graph like this:

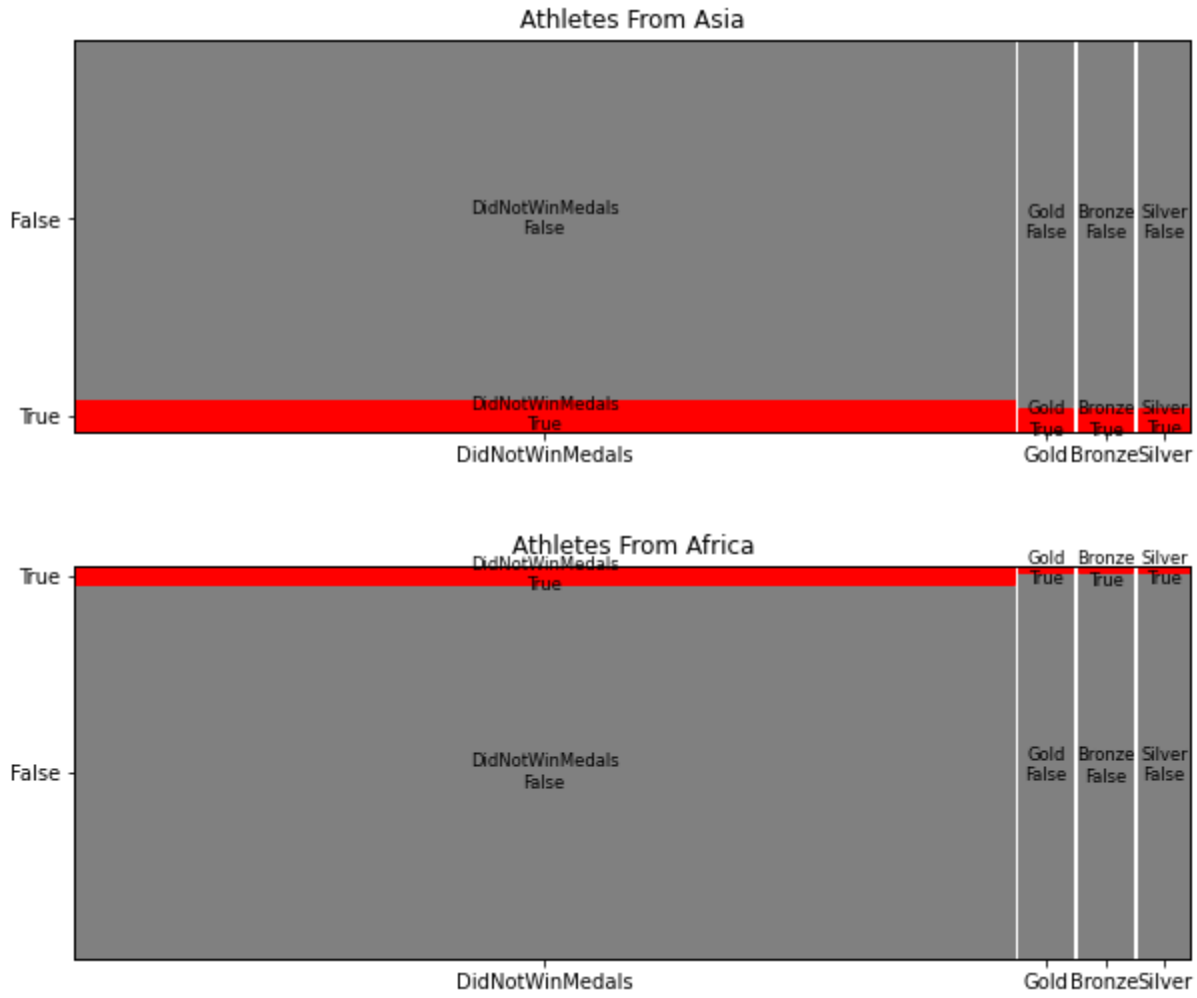


Here, looking at the proportion for different regions, Western Europe dropped significantly, and the differences between regions were not that significant.

Hence, ***Is it true that countries from Europe or North America won more medals in the past?*** The answer is Yes, but when taking the proportion of winning into account, the difference is slight for different regions.

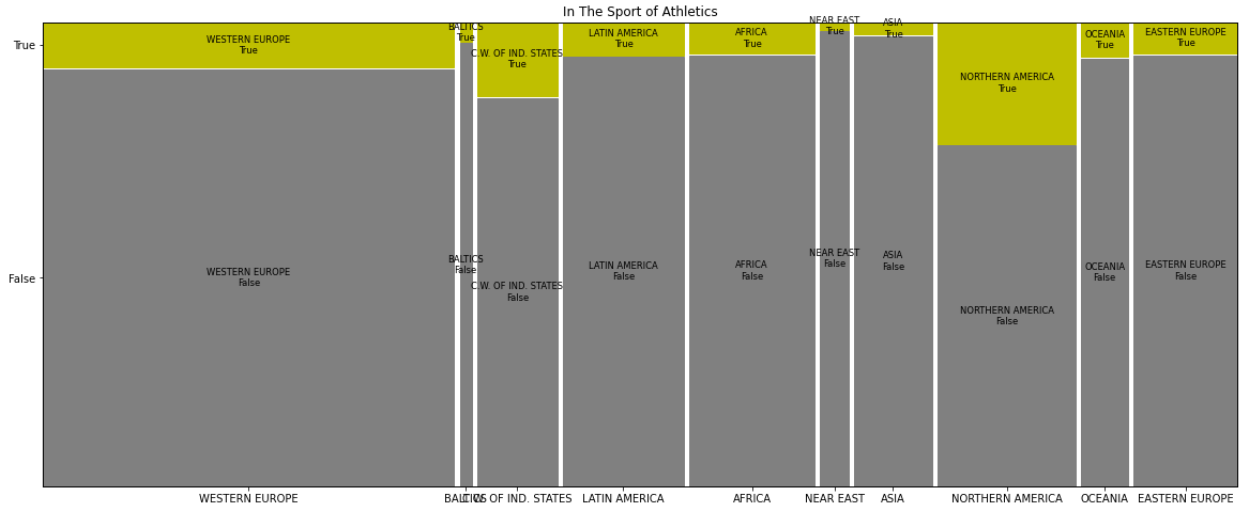
Question 2: Does the stereotype that Asian athletes are less likely to win a medal in athletics (track & field) hold true? What about African athletes? Do they have a higher chance of winning medals in athletics? What about other sports?

Originally, the question was not narrowed down to the sports ‘’athletics”, but for every sport, which was not suitable for analysis because some sports do not require high athleticism. Moreover, another reason for its unsuitability is that the analysis method I did was wrong. These are the mosaic plot I did for the original question:



We can see, whichever region the athlete is from(Asian or not, African or not), the chances of not winning medals will always be higher than the chances of winning medals. I shouldn't analyze them respectively, because I would always gain the same outcome that their chances of not winning medals are lower than winning medals, which is an inevitable outcome.

Therefore, I revised my question to narrow it down to the sport "athletics" and tried to filter out the rows that are in this domain. Next, instead of creating another two columns representing an athlete is from Asia or not and Africa or not. I let the column "region" do its job to show whether the athlete is from that region or not. Then, I created another column and used .replace() to change all "Gold", "Silver", and "Bronze" to "True", and "DidNotWinMedals" to false in this column, denoting whether an athlete won a medal or not. Finally, I built a mosaic plot that shows the proportion of winning medals for different regions in the sports "athletics":

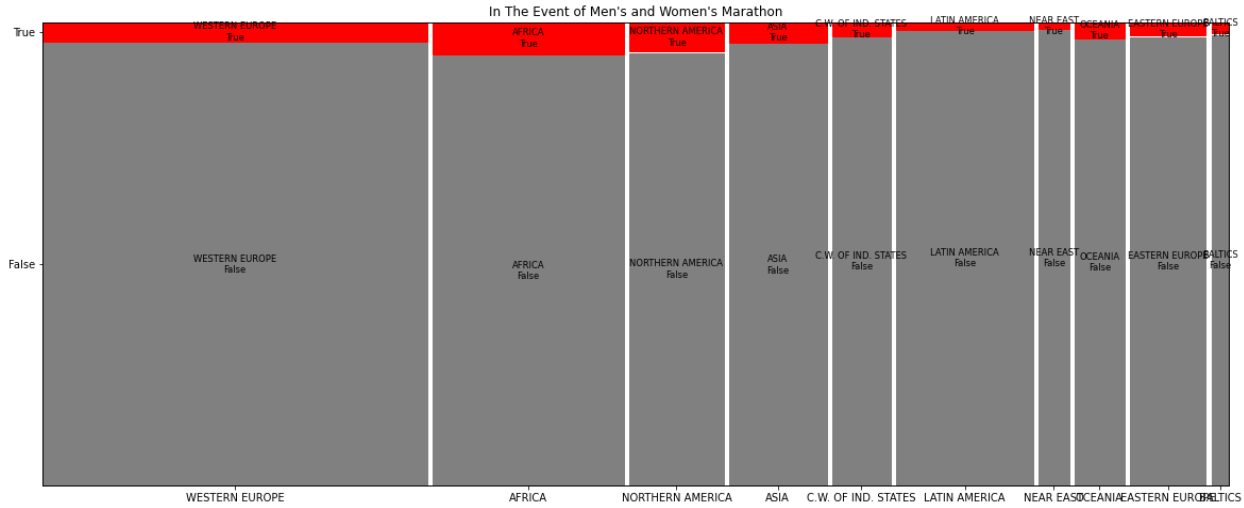


We can see, the proportion of Asian athletes winning medals in “Athletics” is relatively smaller than other regions, whereas the proportion of athletes coming from Africa is not as dominating as I thought. In “athletics”, athletes from North America have the highest proportion of winning medals.

What about other sports?

1. Marathon

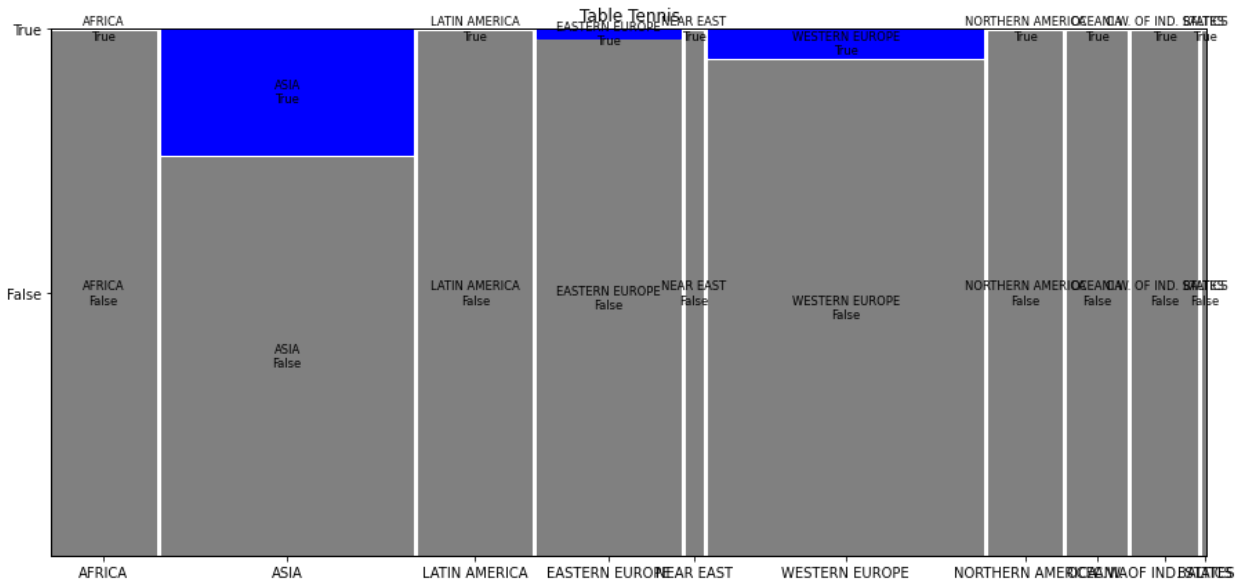
Marathon is the most sacred sport in the Olympics and takes place on the last day as the grand finale. I assumed that this event will be more likely won by African athletes. To prove my assumption, I used the where() to narrow down to the event of "Athletics Women's Marathon" and "Athletics Men's Marathon". Next, repeating the steps of creating the mosaic plot, I gained the graph down below:



Here, we can see in the marathon events, African athletes do account for a higher proportion of winning medals.

2. Table Tennis

Repeating the steps again by narrowing down to table tennis events, I built another mosaic plot down below:



In table tennis, the proportion of Asian athletes winning medals is significantly high compared with other regions.

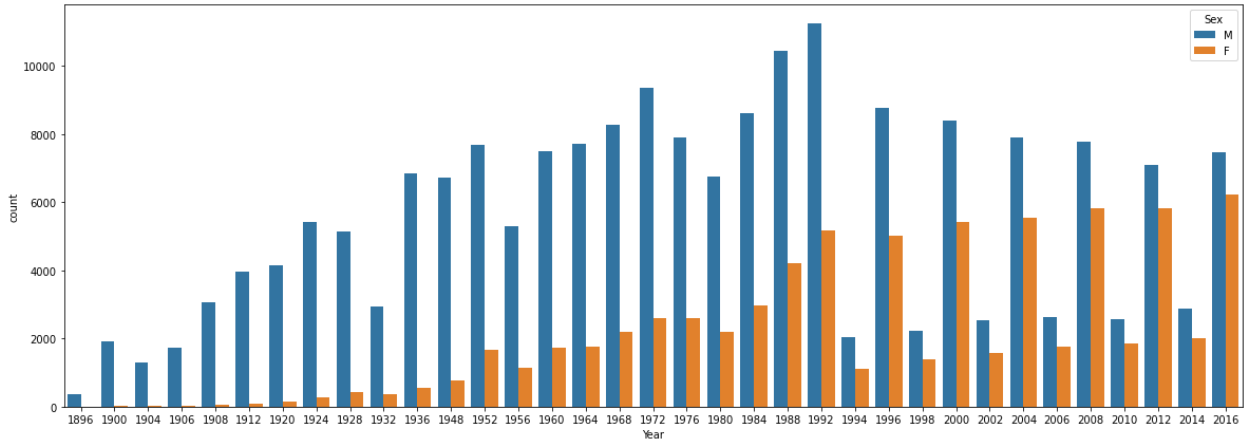
Conclusion:

Yes, the stereotype of Asain athletes is less likely to win medals in athletics holds true. Asain athletes did account less for the total winning medals in the past. However, Athletes from Africa did not show a higher proportion of winning in athletics either. But for the marathon event, the outcome shows that African athletes had a higher proportion of winning in the past. As for table tennis events, this sport is highly dominated by Asian athletes.

Question3: Were there more and more women athletes participating in the Olympics games? Moreover, do countries with better economic levels tend to send more female athletes?

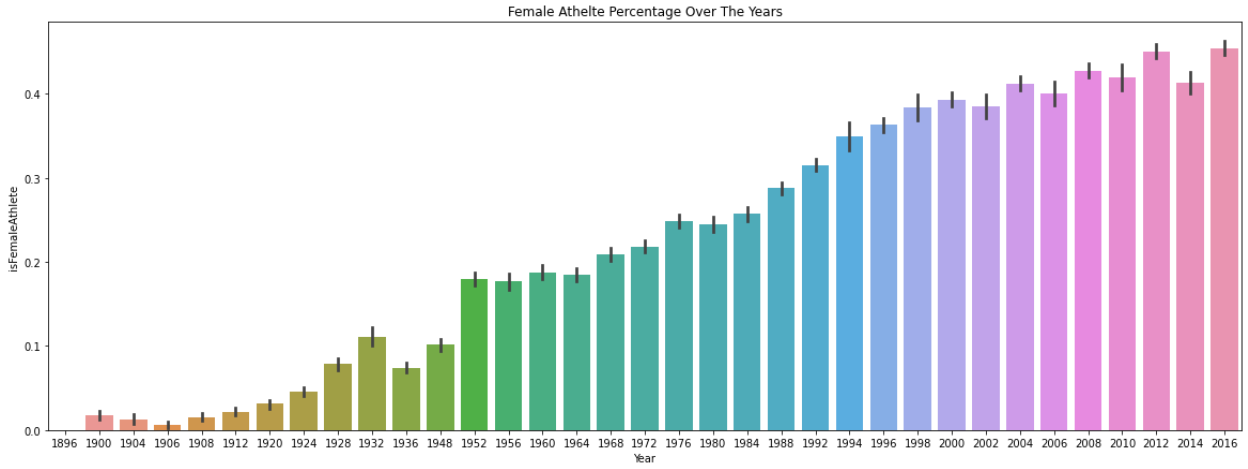
I created a dataframe ***Q3*** that contains the columns of "Country", "Region", "Sex", "Season", "Medal", 'GDP ($ per capita)' and 'Year' from ***join2,*** and also cleared up the white spaces in the column “Region” and renamed the values. Next, I created a

countplot with x denoting different years and y representing gender and gained the graph down below:



It shows that there is an upward trend for the total female athletes participating in the Olympics over the years.

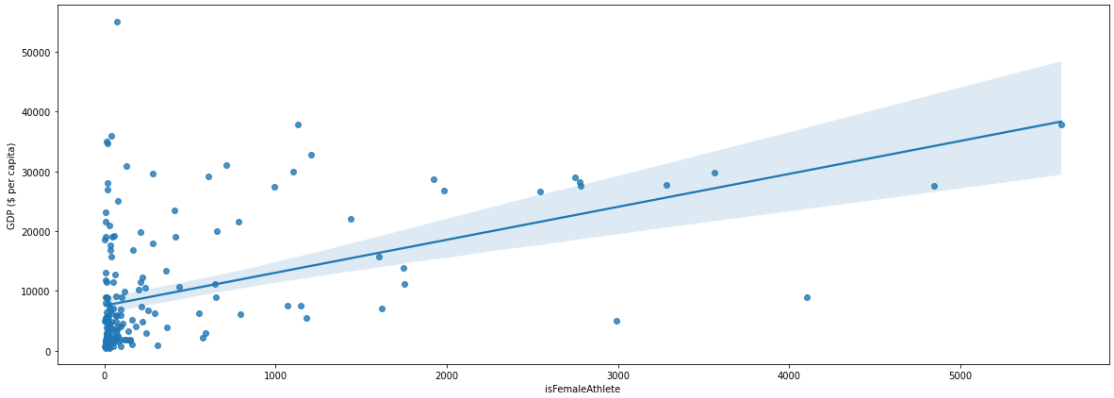
Next, I created another column “isFemaleAthlete” with 1 denoting “is female”, 0 denoting “is not female”. And used a barplot to see the proportion of women’s participation over the years:



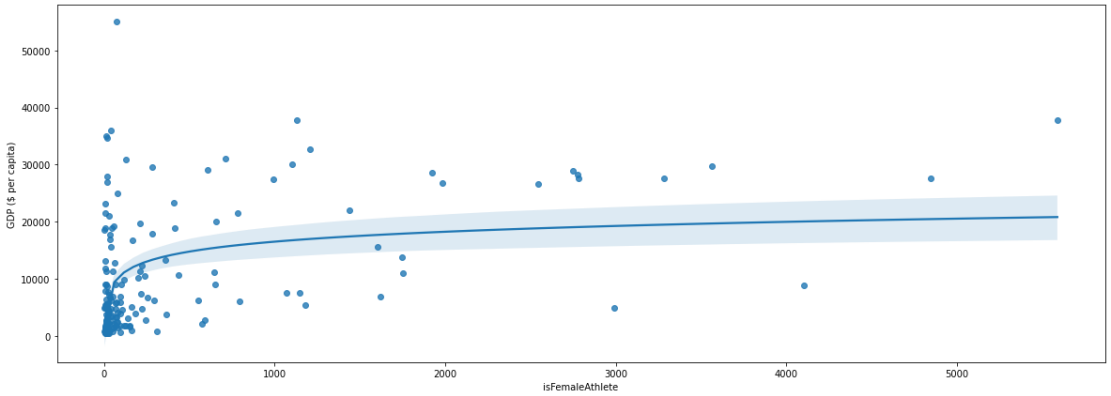
Here, we can see the rising participation of women athletes over the period of 1896 to 2016.

At last, ***do countries with better economic levels tend to send more female athletes?***

I used grouby() then reset_index() to gain the dataframe, after sorting the dataframe in descending order it looks like there is a relationship aligned with the question. Then I used regplot to plot the relationship between GDP per capita and the column “isFemaleAthlete” and gained the regression graph down below:



I also tried to fit the regression model using log(x):



Finally, I created an OLS Summary down below:

OLS Regression Results

Dep. Variable:	GDP	**R-squared:**	0.217
Model:	OLS	**Adj. R-squared:**	0.213
Method:	Least Squares	**F-statistic:**	49.75
Date:	Tue, 07 Dec 2021	**Prob (F-statistic):**	3.68e-11
Time:	22:27:01	**Log-Likelihood:**	-1909.3
No. Observations:	181	**AIC:**	3823.
Df Residuals:	179	**BIC:**	3829.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	7531.7913	756.091	9.961	0.000	6039.793	9023.789
isFemaleAthlete	5.5078	0.781	7.053	0.000	3.967	7.049

Omnibus:	72.324	**Durbin-Watson:**	2.074
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	205.359
Skew:	1.691	**Prob(JB):**	2.55e-45
Kurtosis:	6.973	**Cond. No.**	1.06e+03

The p-value shows that the coefficient was significant, and the coefficient term, 5.5078, tells us the change in Y for a unit change in X. Hence if 'isFemaleAthlete' rises by one unit then 'GDP' rises by 5.5078.

Conclusion:

Yes, there were more and more female athletes participating in the Olympic games, and the participation was significantly low in the early stage of the Olympics. Using different plot methods we can see a slightly different upward trend for the female athletes, in the barplot, it shows a more reasonable upward trend of female athletes instead of just taking only the total number of them into account.

For the regression analysis, we can see that there is a positive correlation between economic levels and the participation of female athletes.