# Olympics Web Scraping Report
Yuqian(Tracy) Ran

**Introduction:**

From judging to timing to preliminary rounds to finals to the various Olympic records, there is data for every sport and country at the Olympics Games.
We analysis the Summer Olympic Games from three aspects:

1. location: the US, UK and China will top the medals tables when it is all over, why these countries are so successful.
2. history:
   Analyzing the number of Olympic medals won by geographic region by geographic region year reveals the true impact and extent of medal diversification. For example, whilst more countries are winning Olympic medals, how many medals are they capturing compared to traditionally strong Olympic nations? Is their success fairly minor or more pronounced? What are the possible contributing factors to their success in the Olympics? I listed the history of total medals won by the top 13 leading countries.
3. different sports: Which countries dominate the 2 traditional sports: athletics and swimming?
4. gender and age: Have the Olympic games achieved gender equality in competitors? Does age impact the number of medals the athletes can get?

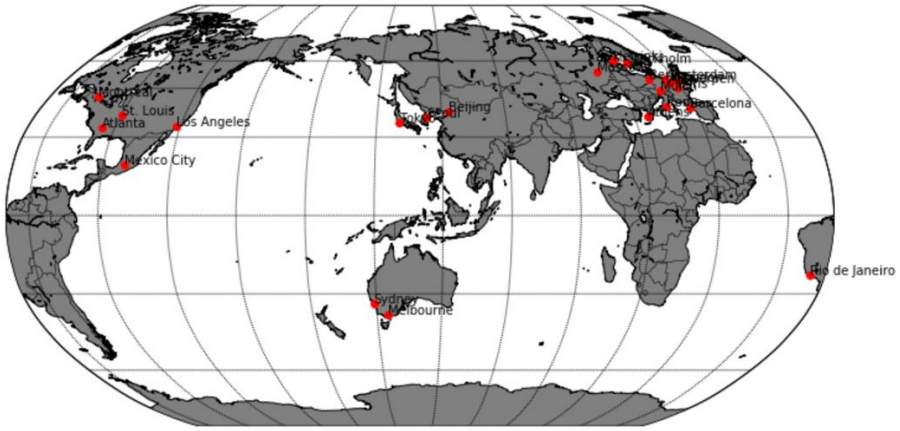
Method:
→ web scraping:
the overall website is organized and structured very well. I wrote a web scraper in Python using the package "Beautiful Soup" and extracted the following data for the future analysis

1. events of country medal leaders among 29 summer Olympics
2. Events of country medal leaders of 2 sports among 29 summer Olympics
3. Events of sports, athletes and medalists of 84 countries during 2016 Olympics

→ K-means clustering:
K-means clustering is an unsupervised learning algorithm that tries to cluster data based on their similarity. Unsupervised learning means that there is no outcome to be predicted, and the algorithm just tries to find pattern in the data. The key of k-means clustering is to determine the number of clusters, K. I used elbow graph to determine K.

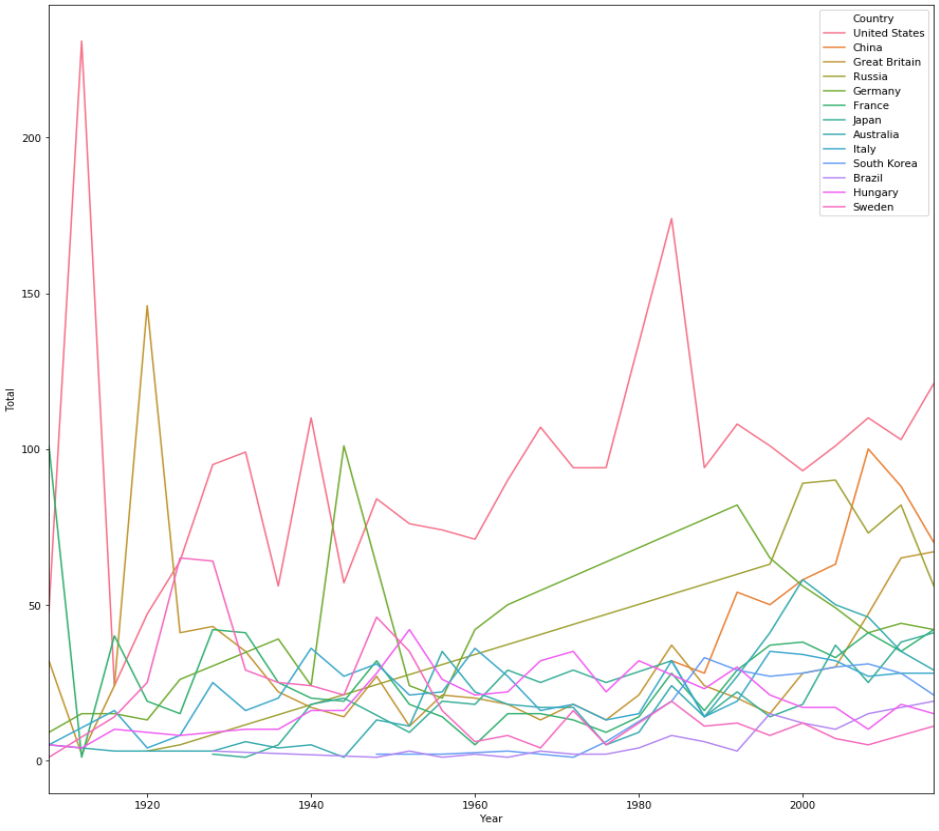
## Data Exploration and Modeling

1. which country host the most summer Olympics?




From the map visualization, we can see that USA host the most summer Olympics in four cities

2. What affects the total medals won by different countries?
    I. overall years/time

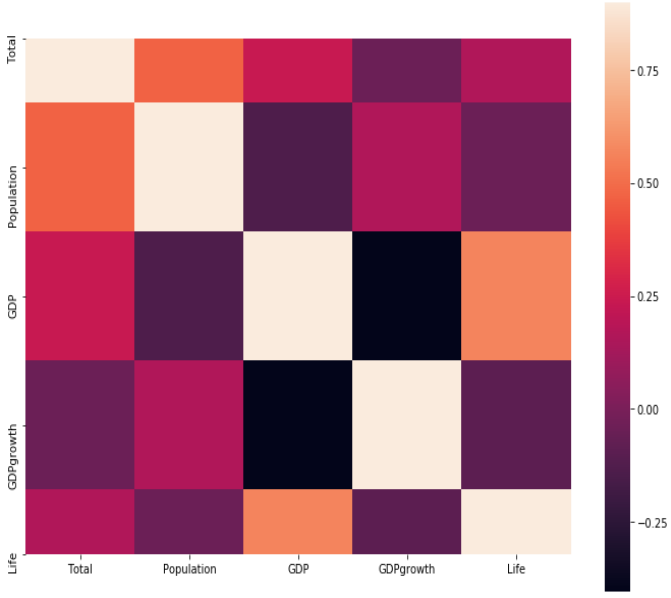


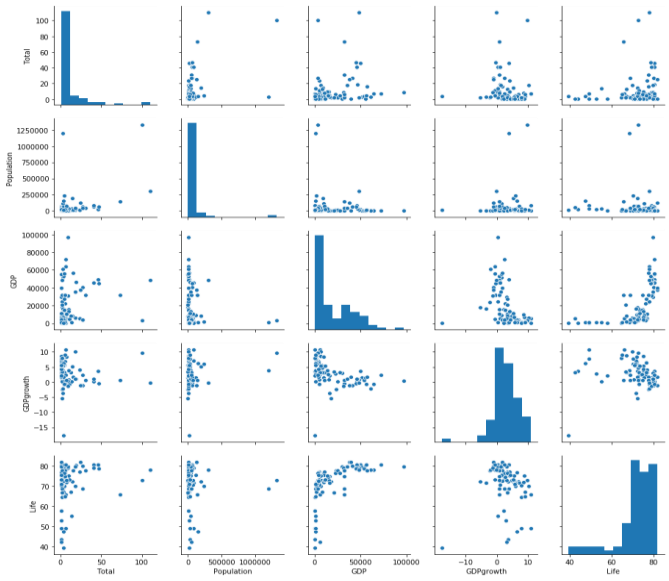
This graph shows that since the first modern Olympic games, the landscape of medal-winning nations has markedly changed. Before World War II, Olympic success was dominated by the US and Europe. Afterwards, more African and Asian countries begin to participate in the Olympics and the medal standings are marked by the arrival and growth of many regions including Japan, South Korea, China and Hungary. Here are three factors that affect medal standing revealed by analysis:

1) Past Olympic success: Medals won in the past can be seen as an indicator of a "sports culture". The US, for example, always perform quite well. Sporting prowess is important to them so many people take part.
2) Host-country effect: The US hosted the 1904 Olympics and won 231 medals compared to 48 at the previous games. The phenomena occur again and again. For example, China hosted the 2008 Olympics and collected 100 medals compared to 63 at the previous Olympics. This is a recognized pattern. Performing in front of a home crowd combined with extra investment in sport gives the host country a medals boost.
3) Future-host effect: Australia won 27 medals in 1992 followed by 41 medals four years later. This was probably due to increased investment in sport in the run-up to the 2000 Sydney Games. The UK, as another example, increased its medal haul from 30 to 47 between 2004 and 2008, prior to hosting the 2012 Games

II. particular one year

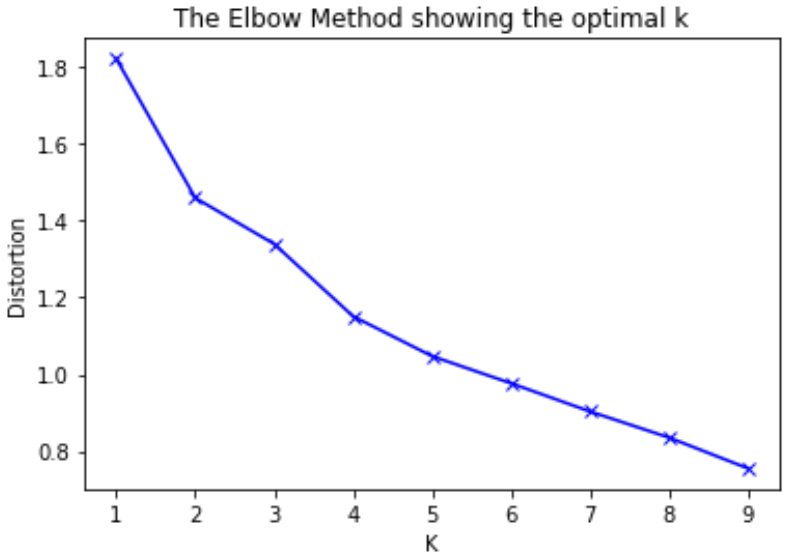
Analyze the detailed year, I extracted the data from 2012 Olympics – including the medal winning record of all the participating countries – and analyzed the relationship between total winning-medals won by each country and its population, GDP, GDP growth and life expectancy



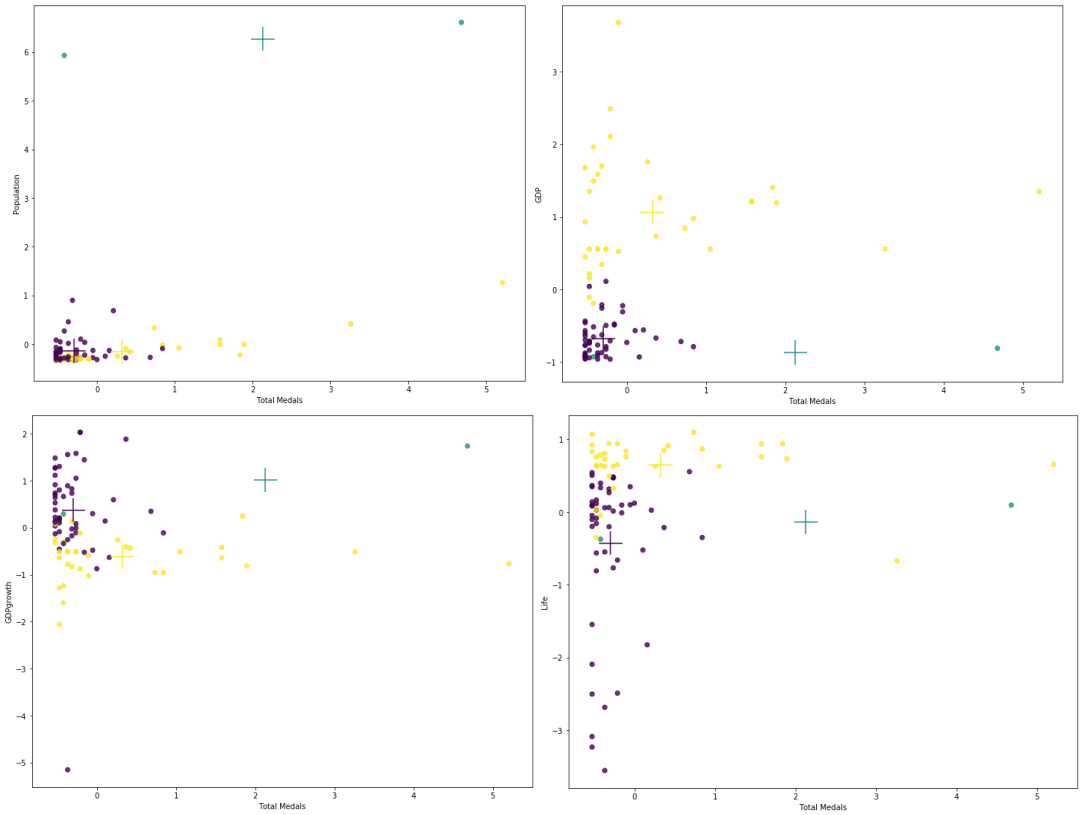



This correlation graph demonstrates the total winning-medals is primarily correlated with population and also other variables such as GDP and life expectancy affect the total winning-medals. And the scatterplots illustrate cluster pattern between the five variables. Thus, a K-means clustering model is applied to find the underlying pattern

- K-means




When K increases, the centroids are closer to the cluster's centroids. The improvements will decline, at some point rapidly, creating the shape. The point is the optimal value for K. In the image above, k = 3.
The three clusters belong to: Indian and china, developed countries such as USA, Great Britain, developing countries such a Hungary, South Africa



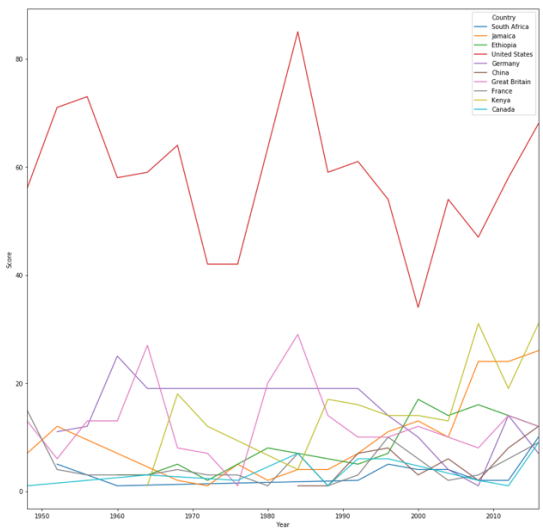

This graph illustrates some national indicators such as GDP population, GDP growth and life expectancy also possibly play roles in the medal-winning battle of the Olympics:

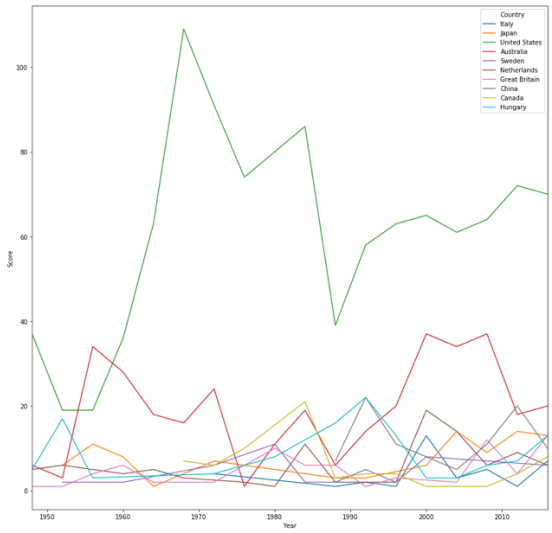
1) Wealth: Countries with a high GDP, like Germany or the USA, can afford to invest in sports facilities and their populations have enough leisure time and money to take part in sports. This may not be the case in poorer countries.
2) Population: A big population means a big talent pool to choose athletes from – in China's case, 1.36 billion people
3) Planned economies: These countries with a high GDP growth tend to invest more in sport because they value the prestige that sporting success brings.
4) Health: Countries with a high life expectancy have a big healthy pool to choose athletes from such as Japan

3. Which countries dominate the 2 traditional sports: athletics and swimming?

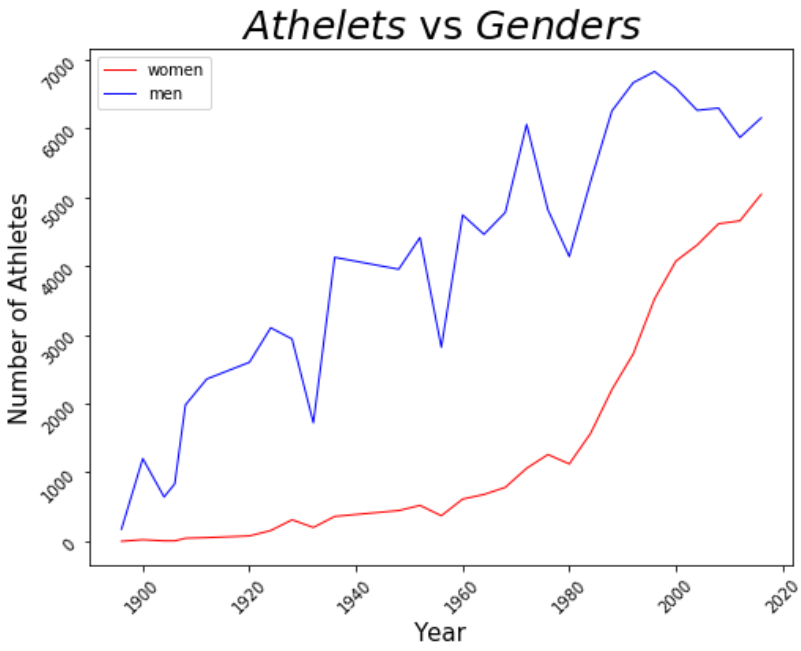
   I created a breakdown of which countries the data shows will excel in which sports by investigating its past performance in 29 summer Olympics, given the total medal values- where a gold medal is worth three points, sliver two and bronze one.

   The two graphs show that US, UK and Jamaica are likely to dominate the athletics events, the US, Australia and China excel in swimming

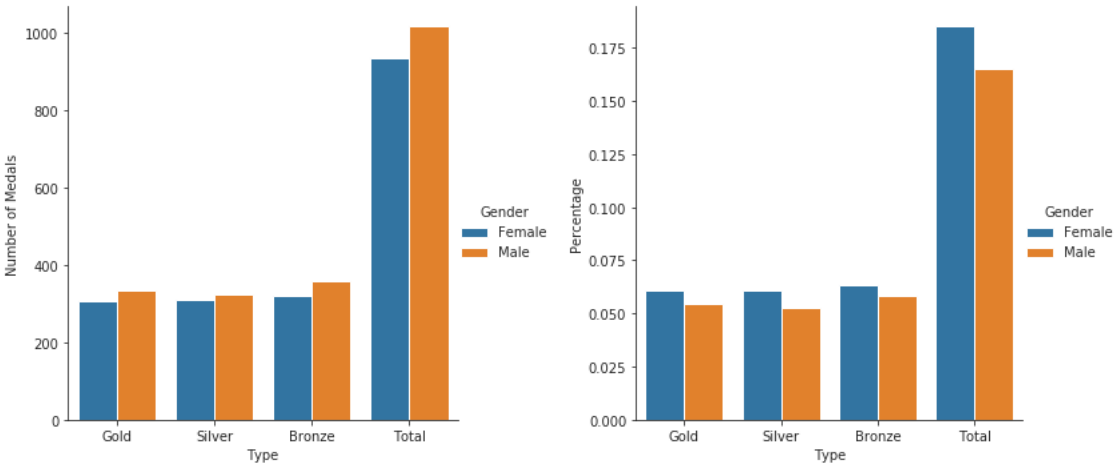


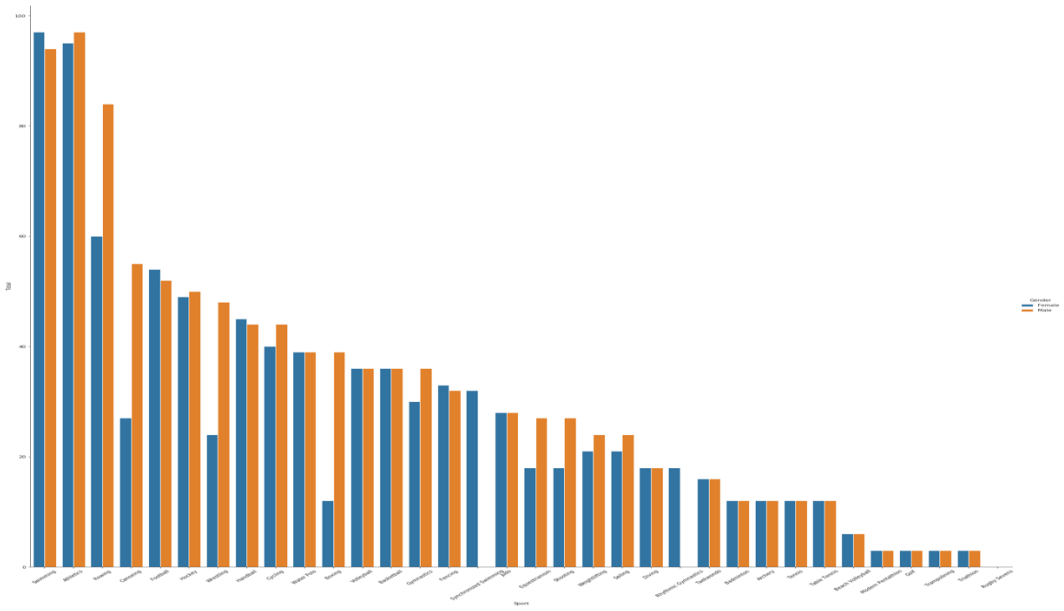

4. Does the Olympics show gender equality?



This graph shows that more women participate in the Olympics. What differences remain between the ways that male and female athletes are involved in Olympic competitions? I analyze all of the men's and women's events at the 2016 Olympics to identify gender differences in the structure and rules of the sports, and in the opportunities for male and female athletes.

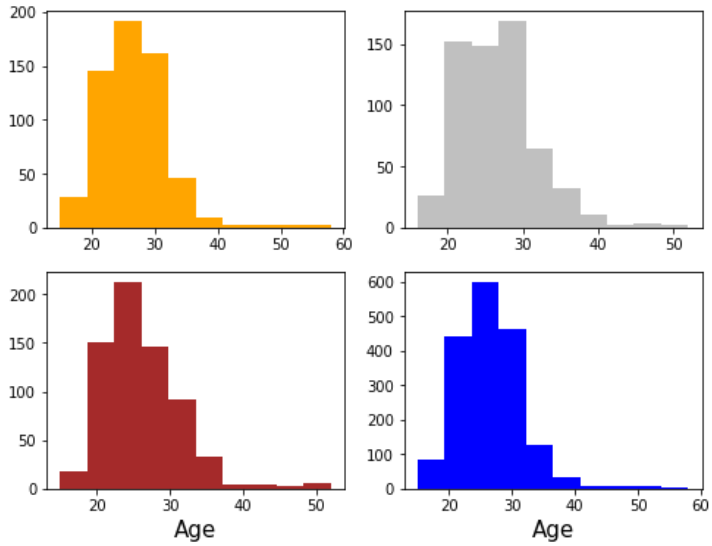



This graph shows that more men won medals while the percentage of women who won medals in higher than that of men.

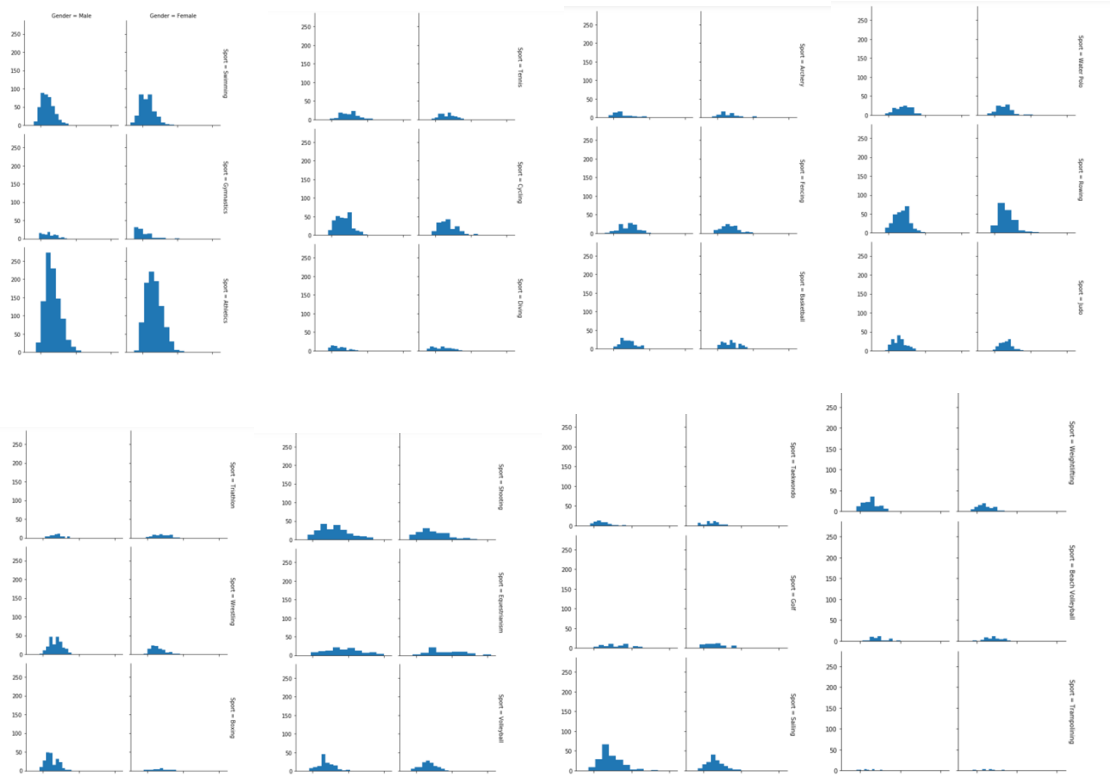


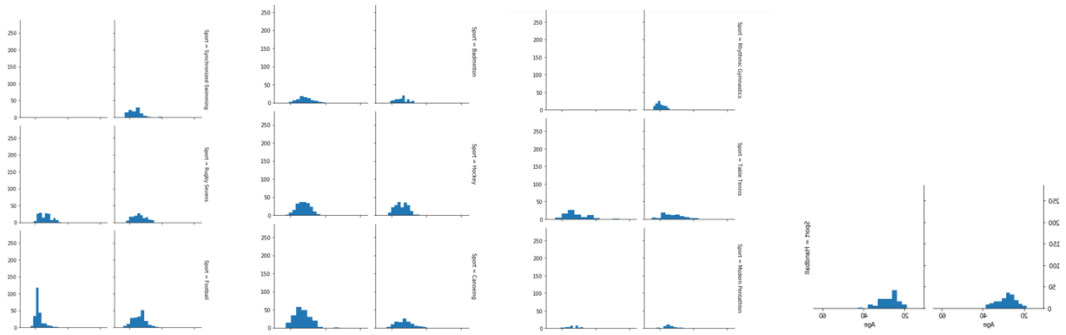
This graph shows distribution of medals winners for different events. We can find that athletics, swimming and rowing produce the most top 3 medals. There are more men events than women events in several field, such as rowing, canoeing, wresting, boxing. In the opposite, synchronized swimming only has women attendants

5. Does age impact the number of medals?




I created histogram plots of the distribution of athletes' age by 4 types: all athletes, athletes who won gold medals, athletes who won silver medals, and athletes who won bronze medals. From the plots, we can see that very young medalists (early teens) and older medalists (late 30s-over 40) tend to win fewer medals than medalists within the "sweet sport" of late teens-early 30s





Some sport, like equestrianism, have older athletes winning medals, whereas a sport like gymnastics has a peak age-range of early-to-late teens and early 20s.