MTH 765P Mini-project

Temporal Analysis of Olympics

Mohit Raju Burkule 210504927

1 Introduction

The Olympic Games are a well-known sporting event. The Olympics have a long and illustrious history that dates back to the Ancient Greek Empire, some 3,000 years ago. It has evolved since then and are now a centre for all athletes from across the world to exhibit their talents in more than 28 different athletic competitions. It is currently hosted every two years in various nations under the labels Summer Olympics and Winter Olympics.

In this report, the the temporal aspects of history of the Olympics is explored.

This includes visualizations that show the change in average age and BMI (Body Mass Index) of participants along the course of the Olympics. Special emphasis is made on senior participants (age more than 50) as it is intriguing to explore if age is just a number in Olympics.

Lets Start with the Olympic Flag 1.

Figure 1: Olympic Flag.

2 Data

2.1 Obtaining Data

Data is obtained from kaggle. `www.kaggle.com/heesoo37/120-years-of-olympic-history-athletes-and-results` It contains 120 years of Olympic history, from Athens 1896 to Rio 2016. It contained both the summer and winter Olympics, but only the summer Olympics are explored here , to limit the scope. Data contained 271116 rows and 15 columns Each row corresponds to an individual athlete competing in an individual Olympic event. There is also a second csv 'regions.csv' which was merged to to get actual country/region names from NOC column (National Olympic Committees). kaggle user heesoo37 scraped this data from a `www.sports-reference.com`. This data is interesting because it allows you to ask questions about how the Olympics have changed over time, such as how women's participation and performance have changed over time, as well as queries regarding different nations, sports, and events.

2.2 Missing Values

The columns of height, weight and medals contained missing values. It is assumed to be due to athlete not receiving any medal during that event . Missing height and weight values showed a negative trend with years (older years had more missing data). Since they were later grouped by year averaged by mean, the missing values in height and weight were ignored.(The outcome wouldn't change if we impute it with mean)

2.3 Data Validation

Some rows were randomly selected and manually googled to validate them.Also, the total medals for a few countries were found to be nearly the same. The deviations are due to the Olympic committee changing the medals due to factors such as cheating and official reporting errors.

2.4 Feature Generation

Body Mass Index(BMI) was generated from height and weight columns. It is used as a proxy for health in subsequent visualizations.

To understand the trend of the number of senior participants(age more than 50) in the Olympics to the young(age less than 50) participants, a ratio (Number of Senior /Number of Young)*100 was calculated. This is because the ratio highlights an important fact in the visualization, and also because the absolute number of senior participants is very low on average.

The inverse Sex ratio was calculated to understand the ratio of females to males . The inverse was used as initially there were very few female participants so the standard sex ratio(number of males/number of females) was very high 70

Getting the number of medals per country required a 6 column groupby ('Year','region','Team','Sport','Event','Medal') to get the usual number of medals reported in media. Usually, when a team wins in a game like rugby all players get medals (13 medals for 13 players) but the number reported is 1 gold as only 1 team won that particular event.

3 Description of data set

Columns in the data set are

1. ID - Unique number for each athlete
2. Name - Athlete's name
3. Sex - Male(M) or Female(F)
4. Age - Integer
5. Height - In centimeters
6. Weight - In kilograms
7. Team - Team name
8. NOC - National Olympic Committee 3-letter code
9. Games - Year and season
10. Year - Integer (1896 to 2016)
11. Season - Summer or Winter
12. City - Host city
13. Sport - Sport
14. Event - Event
15. Medal - Gold, Silver, Bronze, or NA

4 Analysis

Python along with Pandas, Seaborn and Matplotlib was used for visualization

4.1 Medals and Countries

Olympic is an international sport. Hence everyone is interested in knowing how their country performed compared to other countries. To understand this, we can see the cumulative trend in total Medals earned by a few top countries, in Figure 2 . USA leads the race and its cumulative performance is almost linear to number of years, while other countries are seen catching up

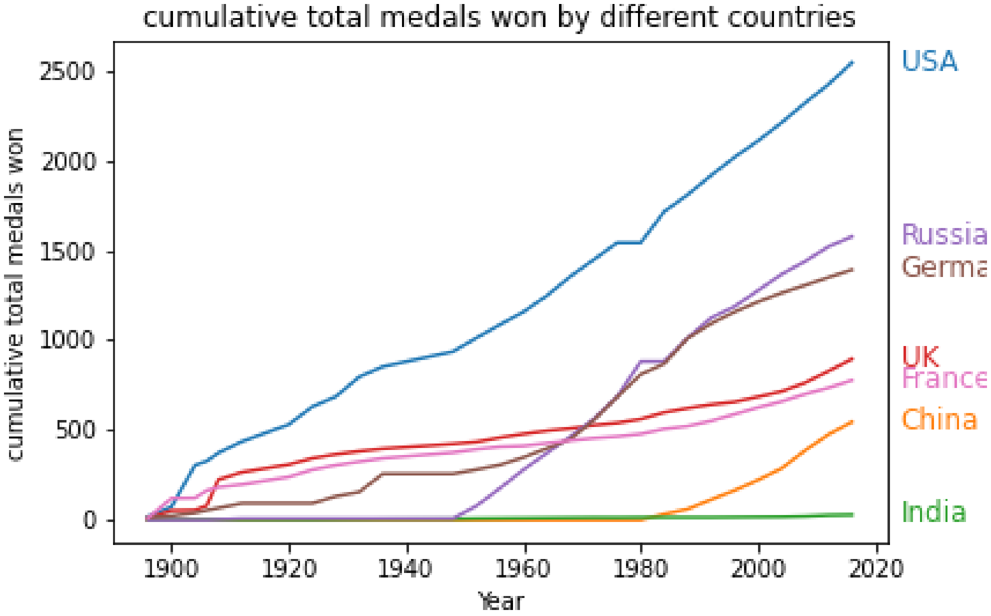

Figure 2:

4.2 BMI and Medals

The importance of temporal analysis is demonstrated here(Figure 3 right), BMI is converging to a particular value(23 in case of Atheletes with 1 or more medals) and its Confidence Intervals are decreasing . We can also see the difference in BMI between the two groups.

This information would not have been available if only a single Olympic was analysed Figure 3 left.

Another point to be noted is that might not be causation and just correlation.

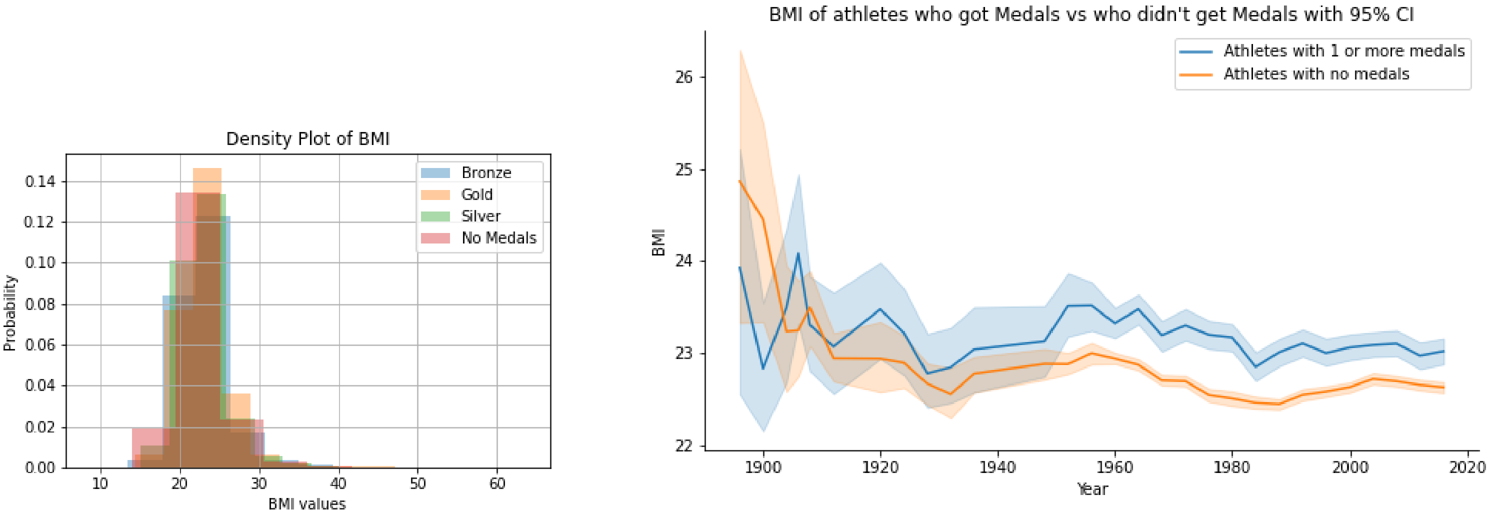

Figure 3:

4.3 BMI and countries

Till now we understand that the USA has most medals and BMI is different for medal winners . Figure 4 shows the change in BMI over time for USA and UK. We see a similar trend as they tend towards BMI 23.

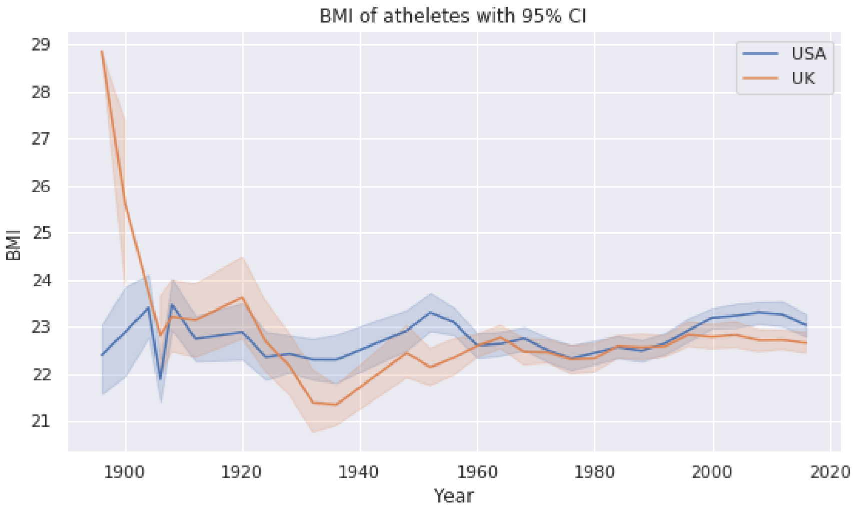

Figure 4:

4.4 Females to Male Ratio

The trend in the number of Females to Males is shown by Figure 5 . We see that earlier there were very few females to males but recently this ratio is tending to 1. In recent years,the ratio for participants is lesser than ratio for winners. This means that even though lesser women are participating , they are winning more.

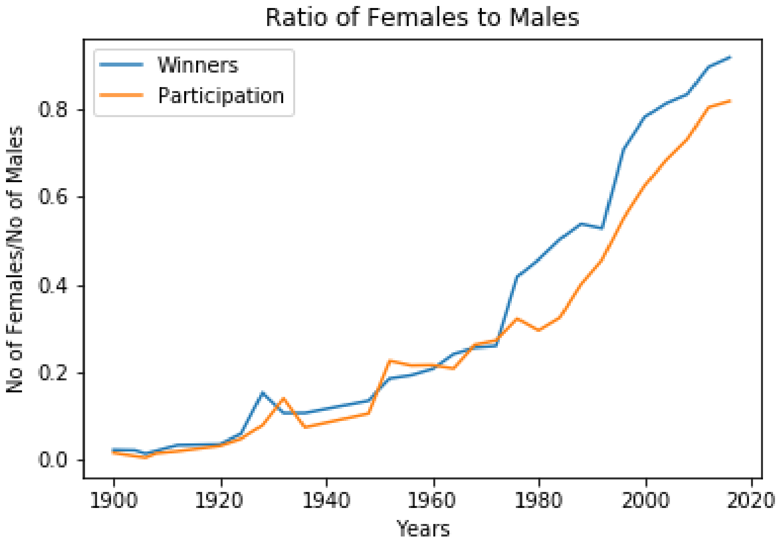

Figure 5:

4.5 Is Age just a Number for Olympics

Let us see the histogram of Age Figure 6 Left We see that most of the participants are in age 20 to 40 , but we also see a few who are over the age of 50! , lets zoom on it Figure 8 Right.

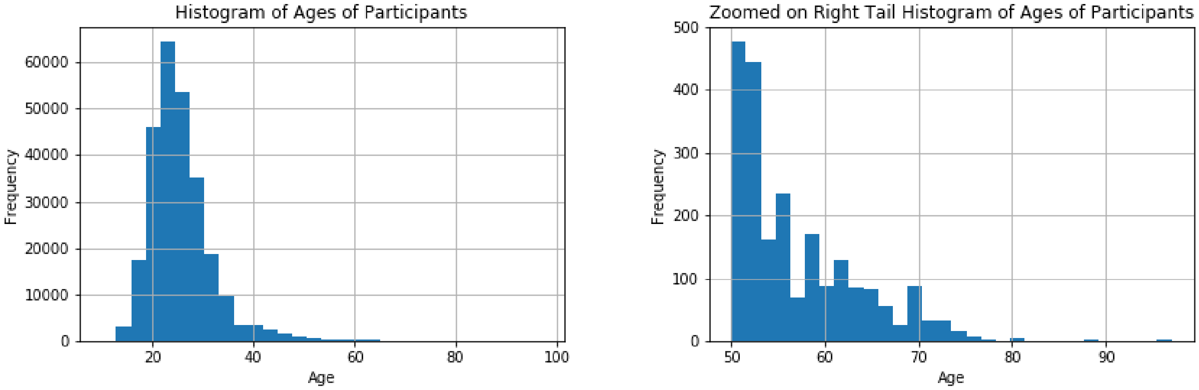

Figure 6:

Lets also plot a temporal view of Age Figure 7

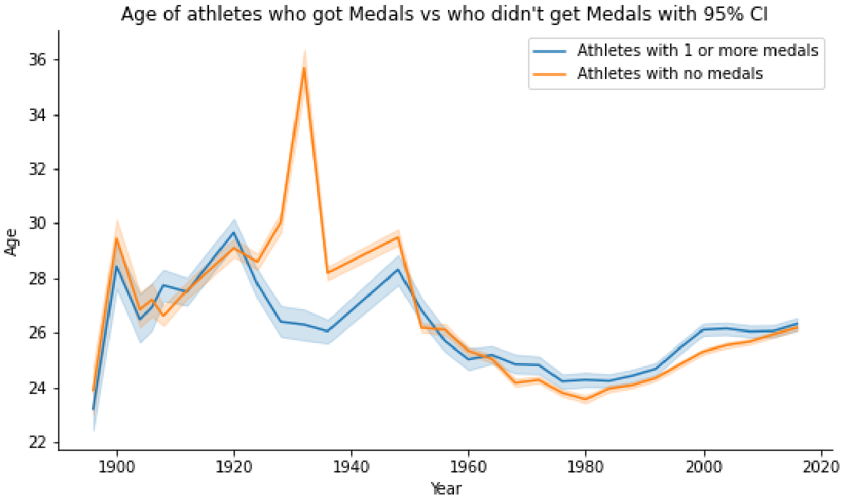

Figure 7:

It looks like average age of athletes who didn't get medal peaked during 1930s , Let's see if it is related to the senior participants in any way.

4.5.1 What happened in 1932?

As we have seen in BMI, a histogram in Figure 5 gives us little information compared to a temporal chart.

Let us try to plot the number of senior participants and number of young participants in the time series diagram . Oops , we soon realize that the senior participants are very less compared to the youngsters, which would not allow us to plot them correctly(without axis scaling). To circumvent this problem we can plot the number of senior contestants per 100 young contestants -Figure 8

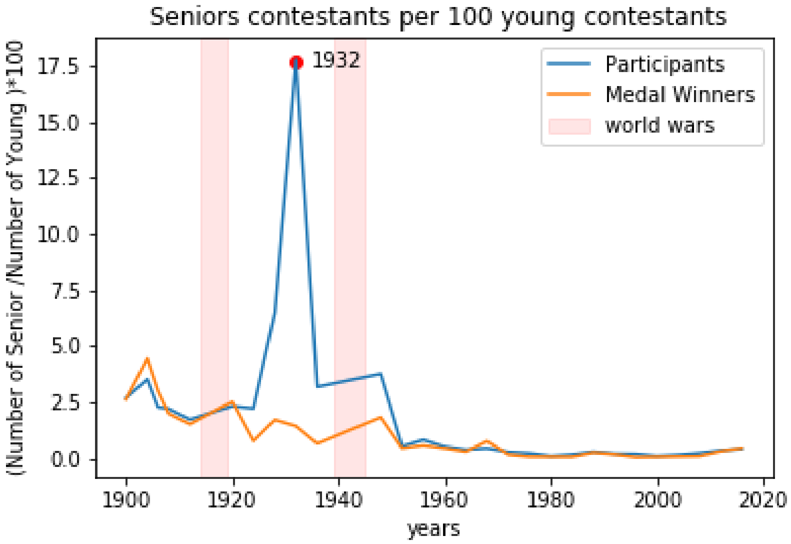

Figure 8:

We see that this participation ratio spiked in 1932 and then fell back . As this was between the two world wars , one could speculate that the elite young population were involved in the war

But This may not be the complete story...

4.5.2 Which games did the seniors participate in?

Lets see which game was most played by senior competitors in Figure 9

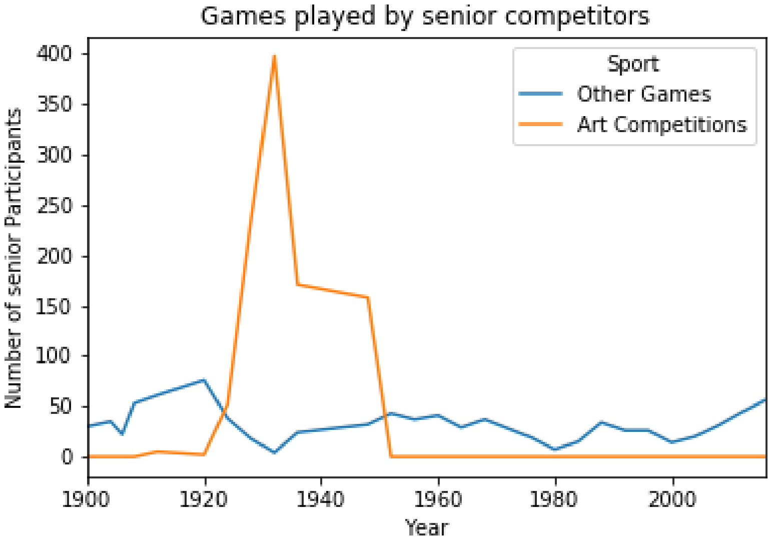

Figure 9:

We see that Art Competitors was most participated in by the senior participants , and the other games were less played in the year 1932 (dip in blue line) This is because Art Competitions were recognized then as a valid sport but it was later discontinued. (Hence there is no spike after 1932) (Another reason might be that the host city was Los Angeles which was then a small town)

5 Questions

In this report, the focus was more on the temporal aspects of the Olympic games as a whole, because most of the people only focus on comparing countries, One could compare countries and their medals and answer questions such as countries with most or least medals / never won gold/silver/bronze , countries with a hat trick gold/silver/bronze in a particular sport

One could also compare individual athletes and answer questions like athletes with most medals/most participation/most years participated

One could merge this data with GDP/population and analyse if GDP/population and number of medals are correlated some other directions can be does the host country get more medals than otherwise.