10/16/2018

Olympic Sports Exploratory Data Analysis

DATS6101 Section: 11

**ALUYA OMOFUMA, ELIE TETTEH-WAYOE, JESSICA FOGERTY, MIHIR GADGIL, PIERRE BAMBA**

## Introduction

## The Olympic games is international sporting events featuring winter and summer sporting competitions for men and women. The Olympics games is arguably the most prestigious sporting event in the world and its popularity is on the increase. In the 2016 edition held in Rio-Brazil, there were 306 events compared to the very first Olympic games in 1896 held in Athens-Greece which only had 43 events. The growth of the Olympics has been a trend from the early years of the event. In the 1950s, after the second world war and the cusp of the cold war the number of events had tripled to 150 from the very first Olympics. The growth continued till present time although it slowed down a bit. There might be an increase in the number of unique events over the next few years, but it would most likely be minute. The focus of our study would be on the performance of the host nations compared to their performance when they are not hosting. There has been 29 unique host cities over the years in 19 countries with some being joint hosts. We would be investigating if host nations have an advantage in their editions by comparing their performance in other all events to their performance when hosting. To do this, we examine the proportion of medals won by a nation in an edition of the Olympics i.e. what percentage of available medals did the country win and use this metric to measure their performance in the Olympics. We used a paired t-test to compare their performance when hosting and when not hosting. This method as used in a study on fivethrityeight by Pettigrew and Reiche, gave us a solid means of measuring host nation performance. In addition to studying this phenomenon, we studied other trends in the dataset, like male and female participation in the games, the cold war effects on the USA and Russia’s participation in the games. We also tackled the performance of countries based on their economic strength. For instance, the USA is the richest country and thus invest most in their sports. Our trend analysis shows that they have won the most medals over the years and we infer that this is a direct relation to their investment in sports.

**Methodology**

## Objectives

### First we imported the data from: <https://www.kaggle.com/heesoo37/120-years-of-olympic-history-athletes-and-results> .

### Added Host\_Country Column.

### We noticed that each athlete was awarded a medal including each team member, we were able to subset the data so that each team was awarded a medal so that there was no discrepancy.

### Called the necessary packages for analysis. We used: ggplot2, dplyr.

### Examined Structure of the data frame and then examined the five statistic summary.

### Examined the formulated questions

### Created visualizations for analysis

### Ran appropriate tests

## EDA

Loaded the dataset and reviewed the structure

## 'data.frame': 271116 obs. of 15 variables:  
## $ ID : int 1 2 3 4 5 5 5 5 5 5 ...  
## $ Name : Factor w/ 134732 levels " Gabrielle Marie \"Gabby\" Adcock (White-)",..: 8 9 44318 29412 21470 21470 21470 21470 21470 21470 ...  
## $ Sex : Factor w/ 2 levels "F","M": 2 2 2 2 1 1 1 1 1 1 ...  
## $ Team : Factor w/ 1184 levels "30. Februar",..: 199 199 273 278 705 705 705 705 705 705 ...  
## $ NOC : Factor w/ 230 levels "AFG","AHO","ALB",..: 42 42 56 56 146 146 146 146 146 146 ...

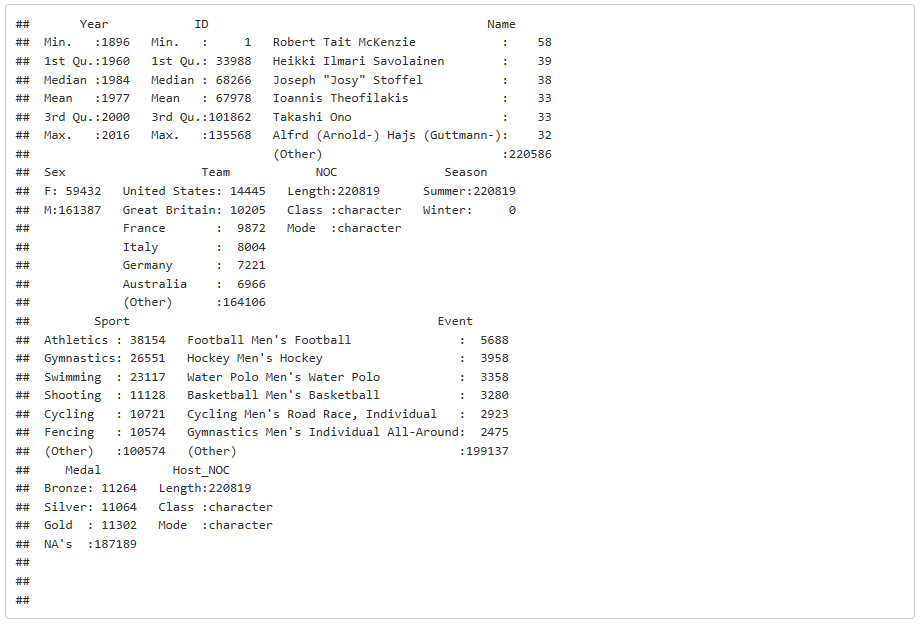
## $ Year : int 1992 2012 1920 1900 1988 1988 1992 1992 1994 1994 ...  
## $ Season: Factor w/ 2 levels "Summer","Winter": 1 1 1 1 2 2 2 2 2 2 ...  
## $ Sport : Factor w/ 66 levels "Aeronautics",..: 9 33 25 62 54 54 54 54 54 54 ...  
## $ Event : Factor w/ 765 levels "Aeronautics Mixed Aeronautics",..: 160 398 349 710 623 619 623 619 623 619 ...  
## $ Medal : Factor w/ 3 levels "Bronze","Gold",..: NA NA NA 2 NA NA NA NA NA NA ...

We changed the structure of some of the columns such as medal, changed to ordered factor. We dropped Age, Height, Weight, City and Games. Select only the Summer Olympics data.

Inserted host country data into the data frame. We then changed some country codes to simplify analysis.

Interesting fact: 1906 had Olympic games, too. They were officially recognized as such at that time, but aren’t considered Olympic games now.

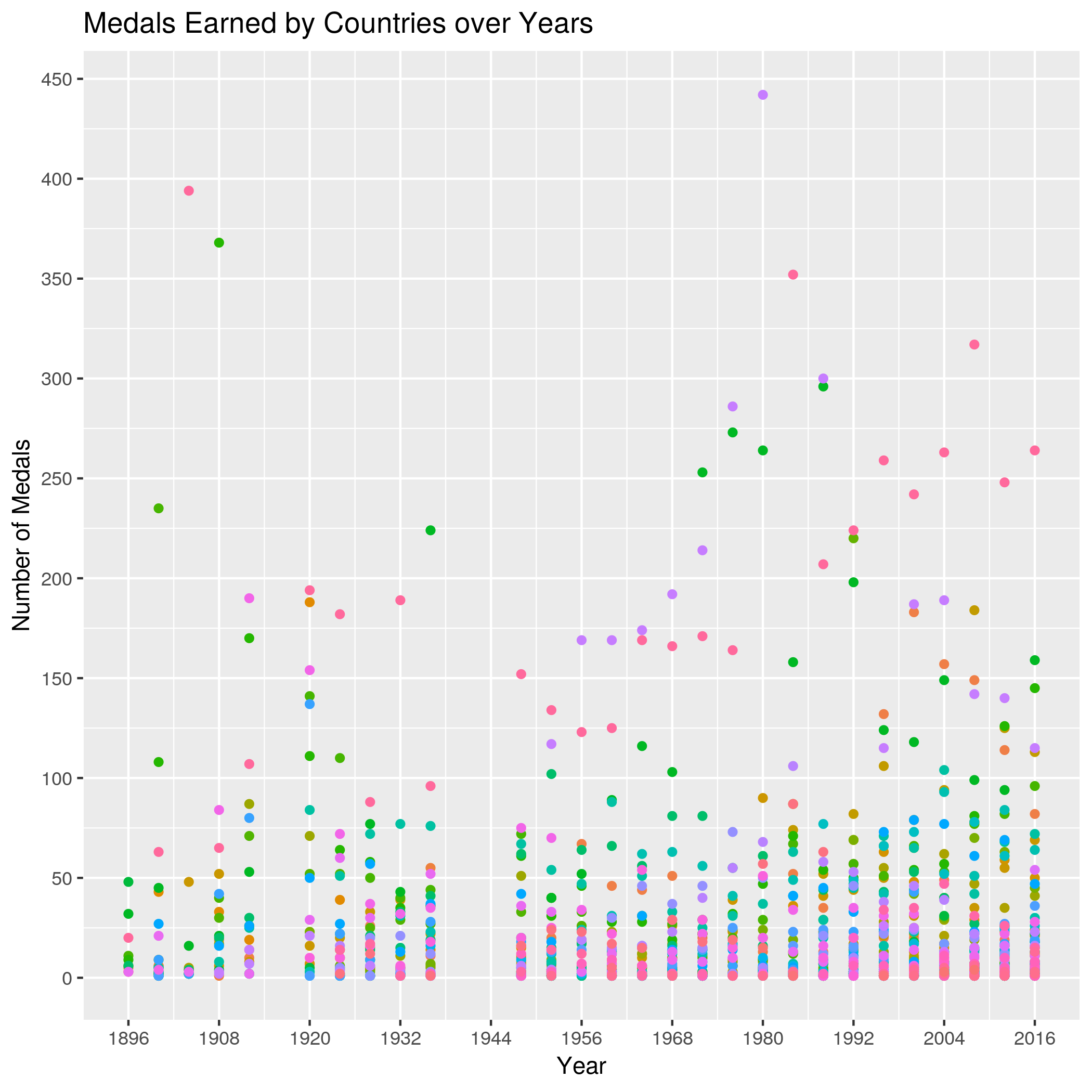
Five summary statistics of the dataset



The summary showed us that the United States, Great Britain, and France are the teams that participated the most frequently. The sports that had the highest number of participants were: Athletics, Gymnastics, Swimming. It is important to note that the three types of medals are not equal in numbers. This is because of ties.

**Does the Host Country have an advantage in Olympics?**

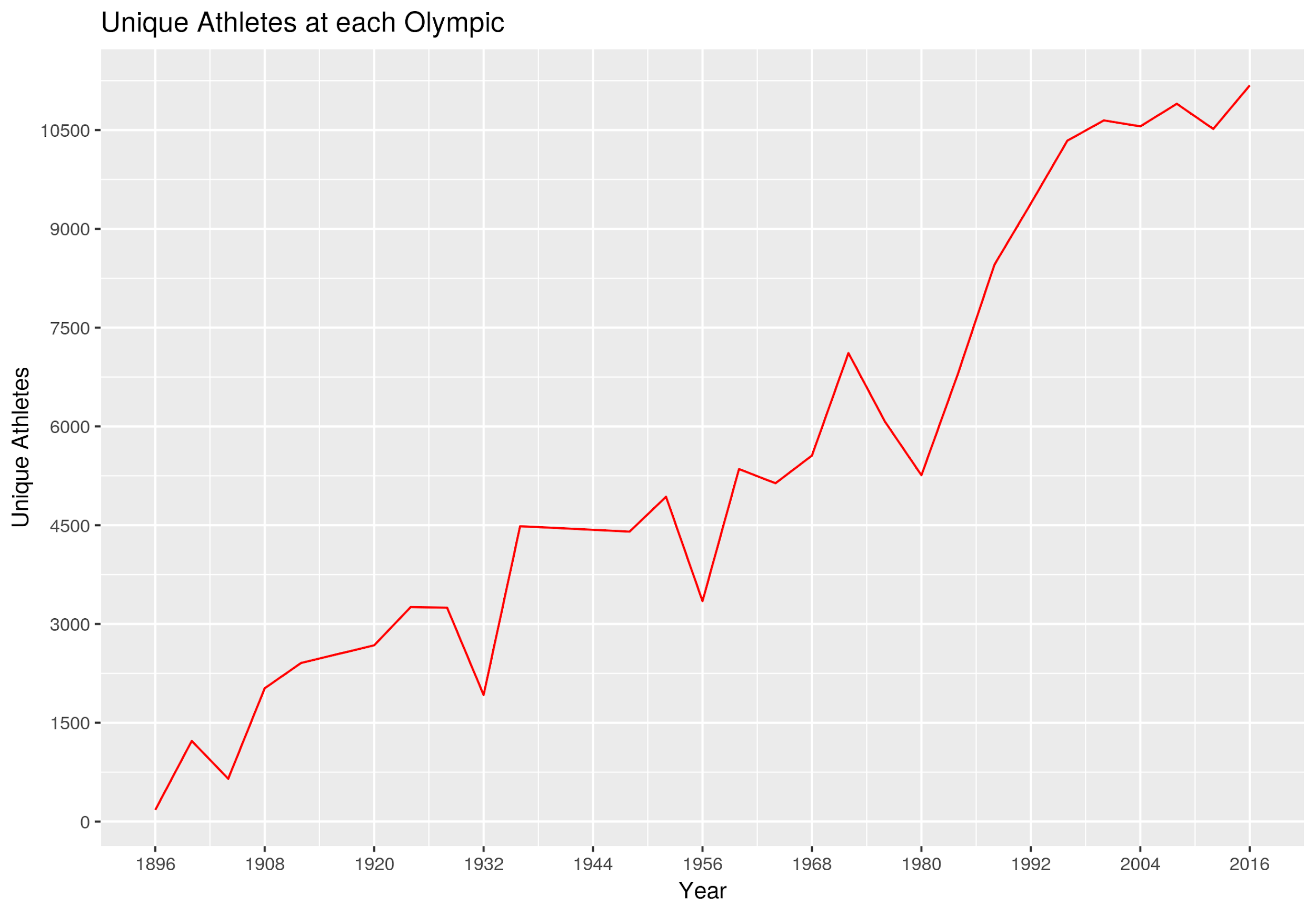
First we examined how many medals per country are won per Olympic games



Based on the visual inspection we would like to conduct a hypothesis test with the following hypothesis:  
 is the difference between the average proportion of medals won as a host country and as a non-host country

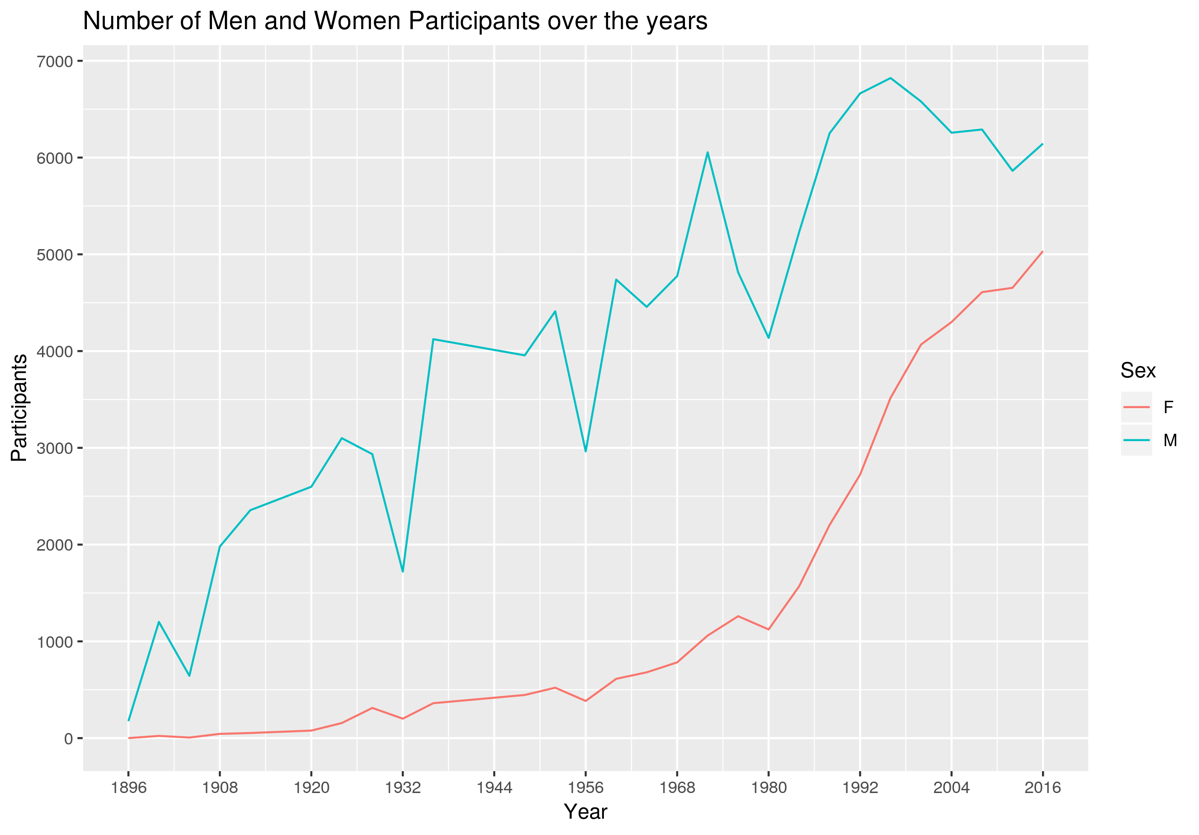
We decided to use the paired t-test because we are comparing two means () from the same sample to determine whether the means are different. We conducted a right tailed t-test to see if the mean () is positive or not, a positive mean will show that the host country wins more when they are hosting the Olympics. Significance level, , we chose this significance level before we ran the test.

The p-value is smaller than the significance level .  
 We reject the null hypothesis.

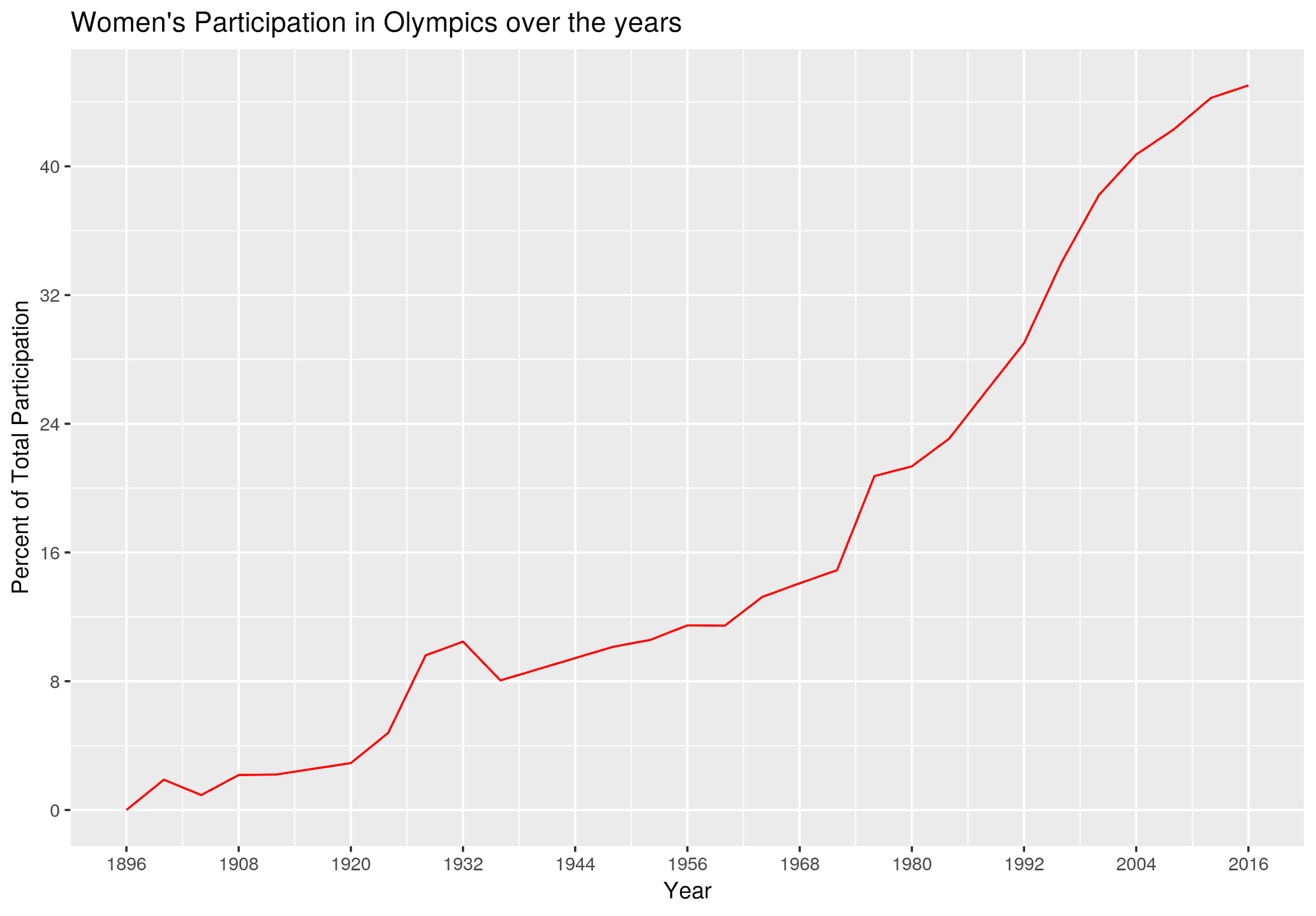
**Participation of Men vs. Women**

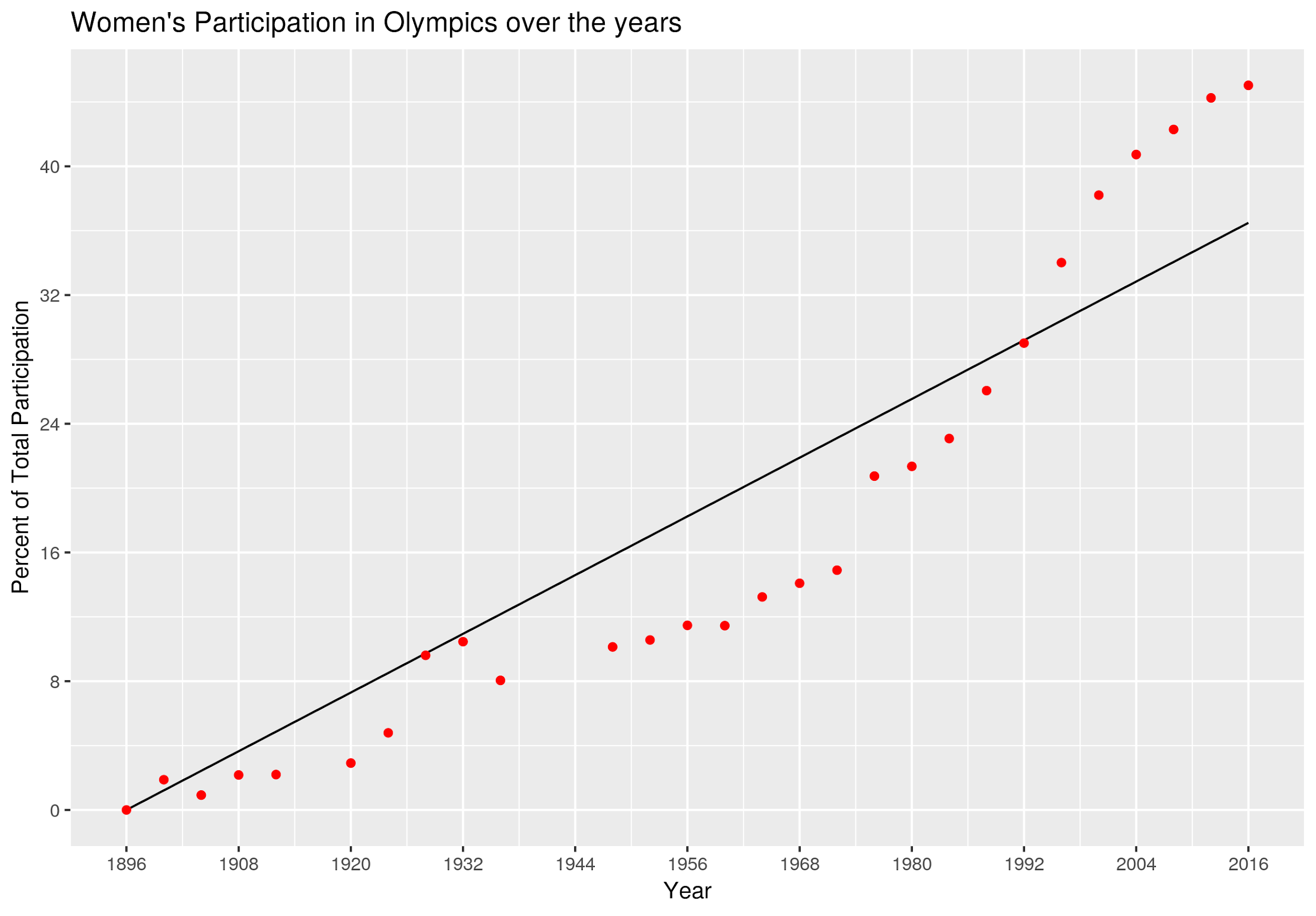
We took a look at the number of unique athletes through the years.

Men/Women events counted separately. This shows that the women participation is increasing and closing the gap (difference) between men and women participation.



In fact, the previous graph (Number of Men and Women Participants over the years) doesn’t do the situation justice. The proportion (total woman participants divided by total number of participants) of women participating each year gives a better picture. As of 2016 45% of the total participants were women.



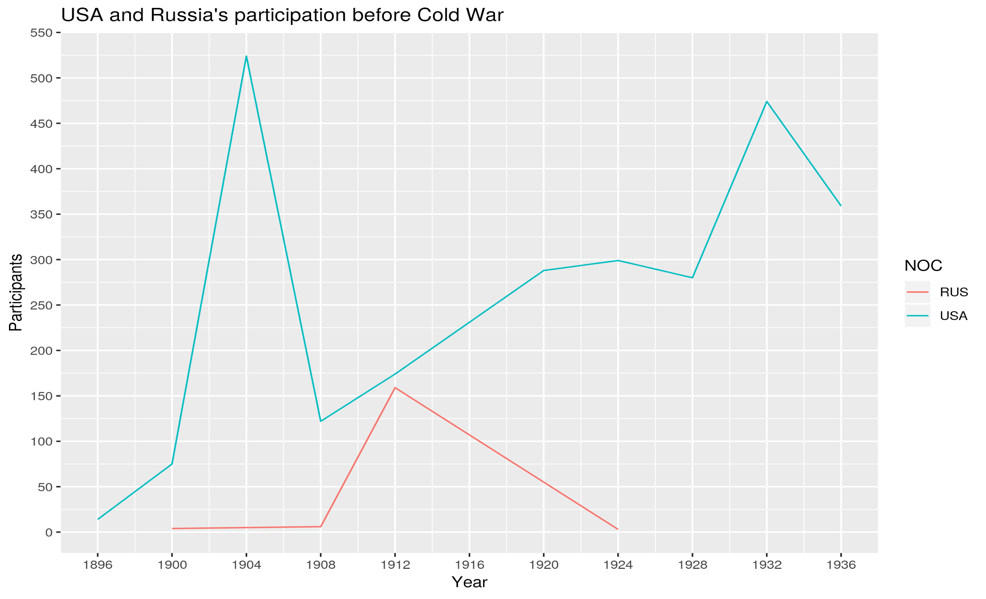


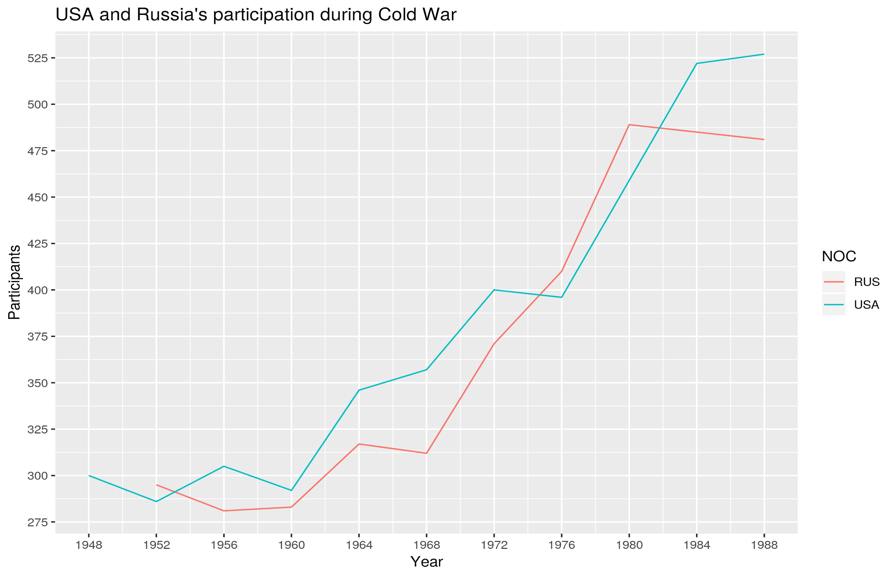
We then created a linear model to predict when women and men participation would be at an equal 50%.

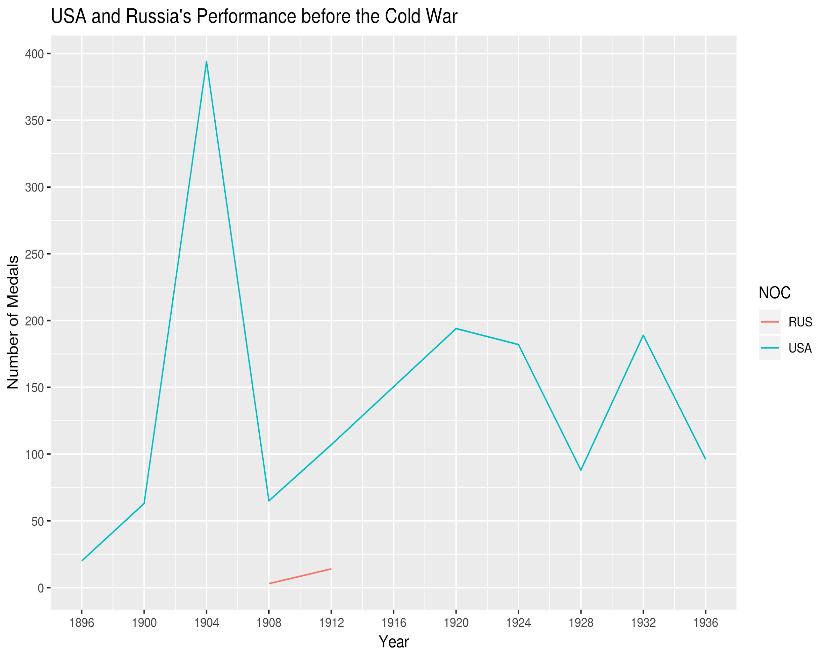
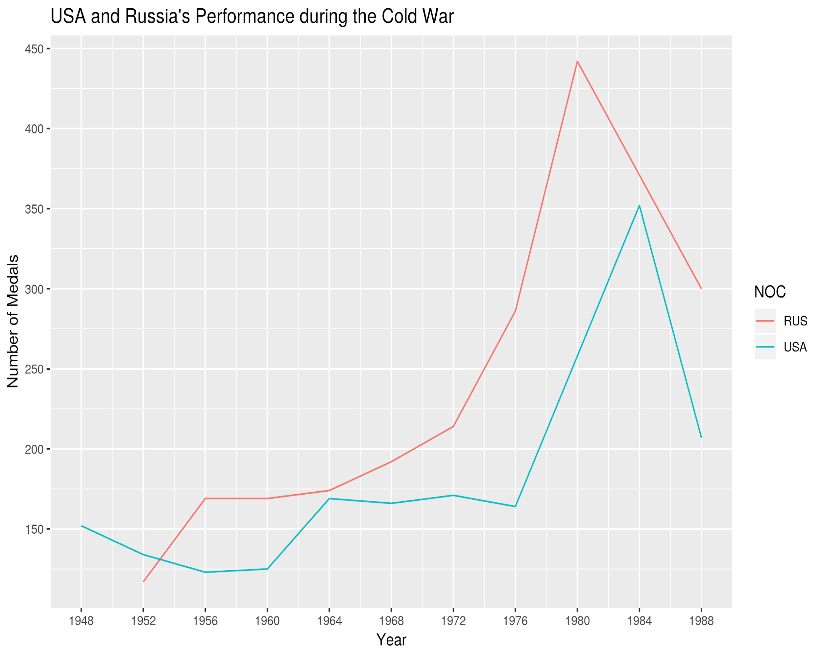
The linear model is:  
Proportion = Intercept + Slope (×) Year

The slope indicates the rate of increase of women’s participation in percentage. Each year we see .3% more women (1.2% at each Olympic games). By the year 2060 Men and Women participation will be equal.

**Cold war’s effect on USA and Russia’s participation**

****Selected data from cold war years and only for USA and Russia. First we analyzed overall participation over the years. We then analyzed the two countries performance. Russia won a total of 2063 medals, the USA won a total of 1763. The plots below show the participation between the US and Russia before and after the Cold War. The bottom two show the performance of the two countries before and after the war. It is clear that Russian participation and performance has increased after the start of the war. Please see the figures below that reference these findings.

****

****

**Top 5 medal winners over all Olympics**

We found the top 5 performing countries, this trend/result indicates that these countries place the most emphasis on athletics, if we were to investigate the economic endowment of the sport ministries of these countries we would also see that they have the biggest budget compared to other countries. This analysis could indicate that these countries have the strongest Olympic teams. The USA leads, then Germany then Russia. This analysis is taking into consideration all of the Olympic games over the years.

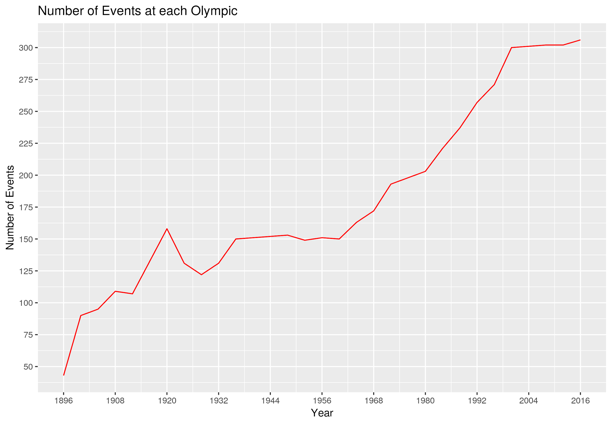
## NOC Medal\_Count  
## 1 USA 4978  
## 2 GER 3096  
## 3 RUS 2968  
## 4 GBR 1946  
## 5 FRA 1563

**Maximum medals earned by each sport**

This tells us which sport is awarded the most medals by country. And by translation, they have the most participants. We can also infer that these are the sports with the most variation in their evens, e.g. athletics have 100m sprint and 110m hurdles. The USA is top in Swimming and Athletics, Germany in Rowing and Russia in Gymnastics.

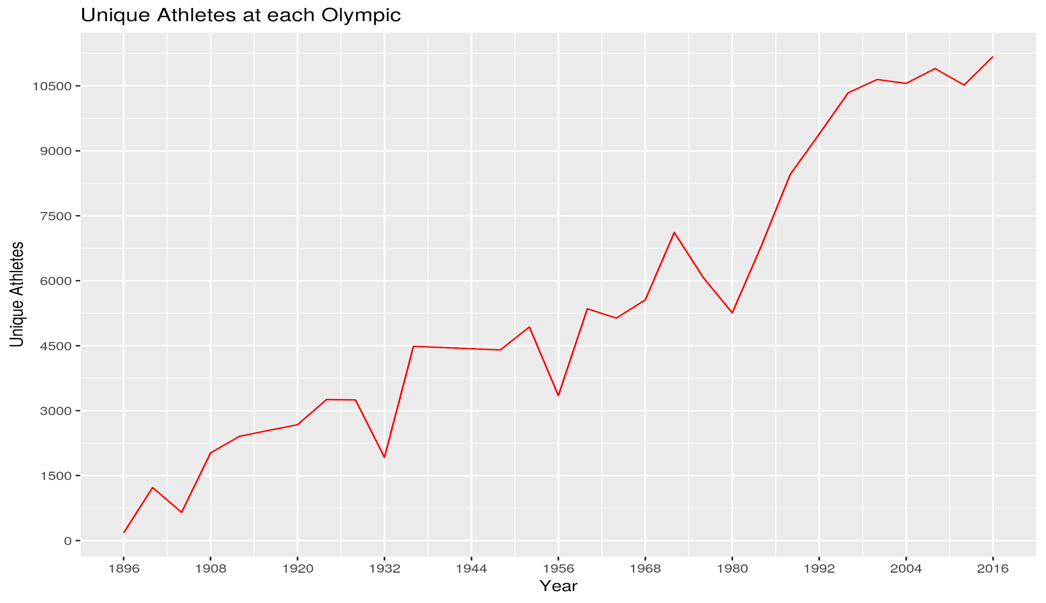
## # A tibble: 57 x 3  
## # Groups: Sport [52]  
## NOC Sport Medal\_Count  
## <chr> <fct> <int>  
## 1 USA Swimming 1077  
## 2 USA Athletics 1057  
## 3 GER Rowing 471  
## 4 RUS Gymnastics 371  
## 5 ITA Fencing 357  
## 6 USA Basketball 341  
## 7 NED Hockey 255  
## 8 GER Canoeing 229  
## 9 RUS Volleyball 211  
## 10 GER Equestrianism 205  
## # ... with 47 more rows

**INTERPRETATION**



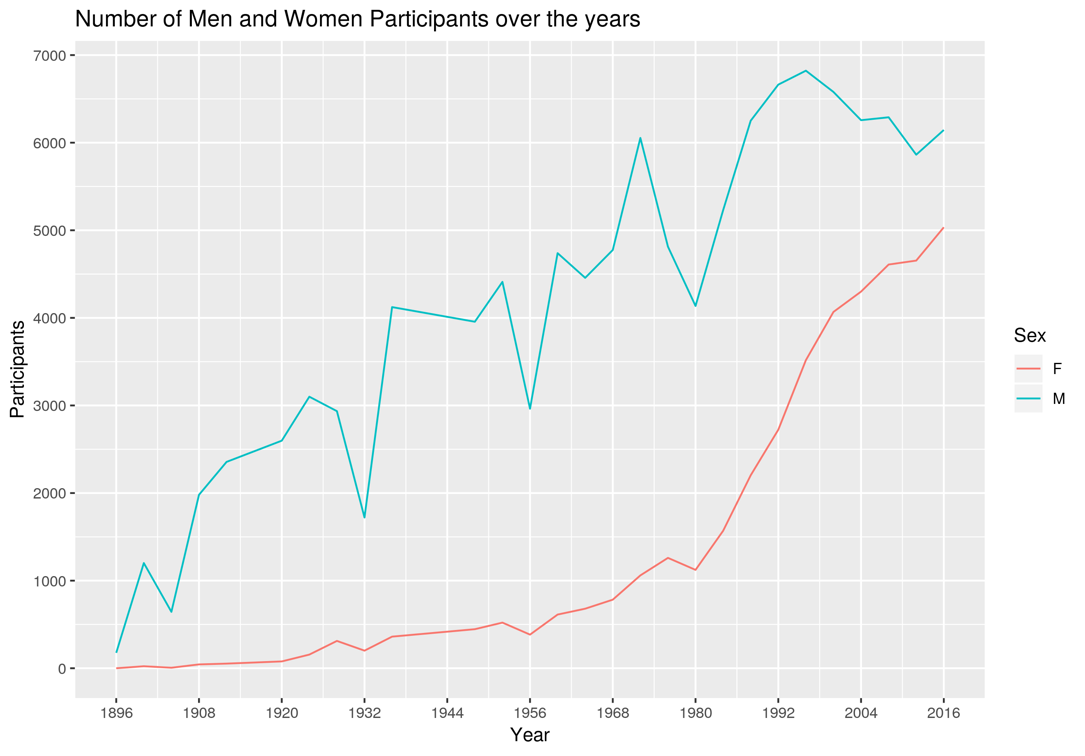
In this line graph, we are taking a look at the number of events conducted through years in the Olympic. Here Men and Women events are counted separately. The count started in 1896 to the most recent Olympics in 2016. The graph records the number of events that happen in an Olympic event.

The graph tells the story of the progress of the number of events in the Olympics. The first section of the graph indicates a positive and ascending progress in accruing events between Olympics from 1900 until 1920 where it gets hit its peak. The second section of the graph indicates a drastic decrease in events accruing and this section is automatically follow by a third plat section where the number of events stay constant. These last two sections happen to be in the world war and cold war period from 1920-1975’s. From 1975’s -1995’s we experience a continued increase of number of even and a flat increase from 1995’s till today. These changes are due to the peace and stability in the world, increase in number of countries participating to the Olympics and also, the increase in the number of discipline.



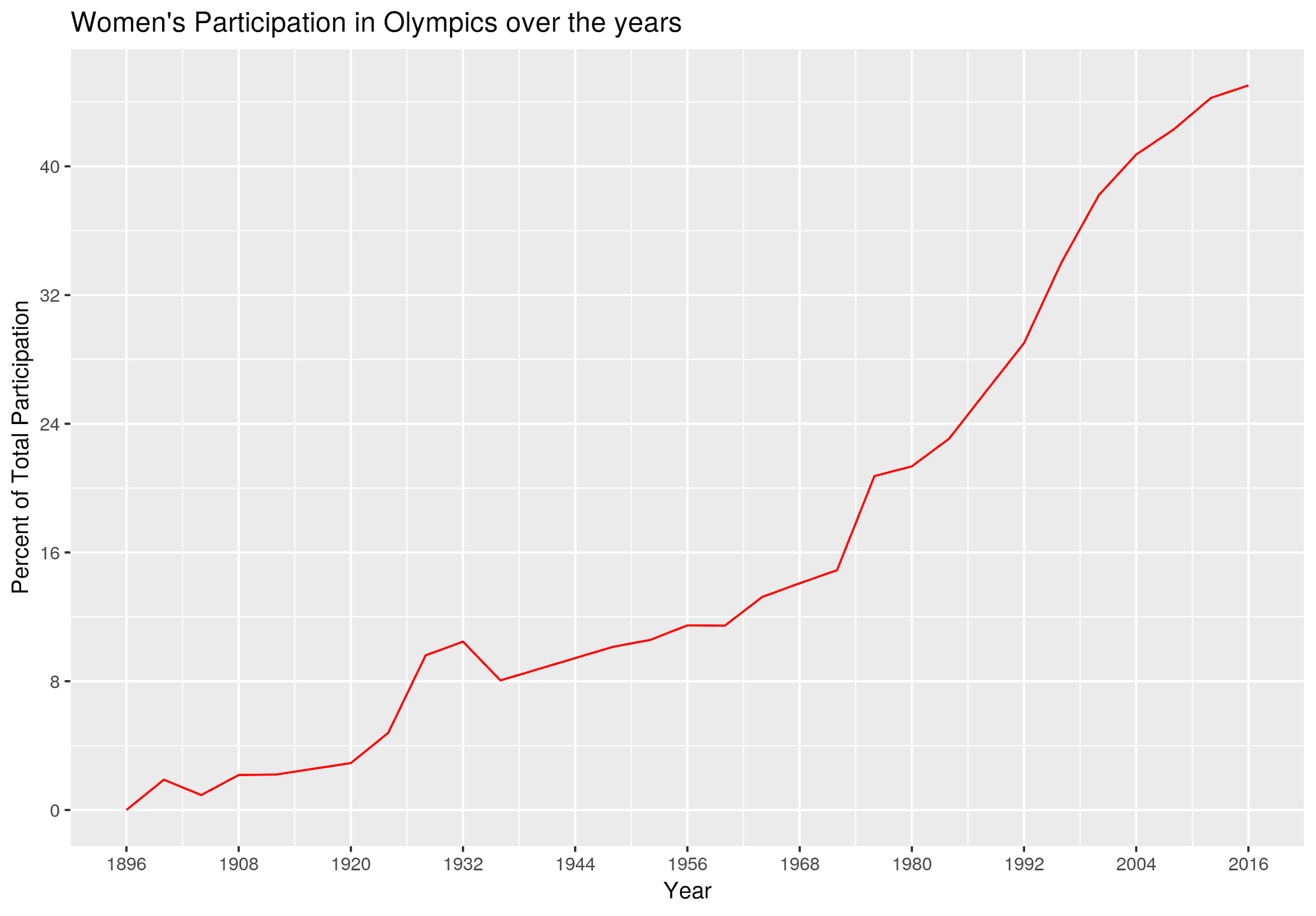
In this line graph, we are looking the Olympics between 1896 and 2016. The graph records the number of athletes participating in an Olympic event at each year.

From the beginning of the Olympics the graph presents a saw-tooth evolution until the year 1980. This period was one of the unstable periods in the history of the world. From 1980 we experience a continue increase in the number on athlete participation at the Olympics and This is due to the peace and stability  in the world, increase in number of countries participating to the Olympics and also, the increase in the number of discipline.

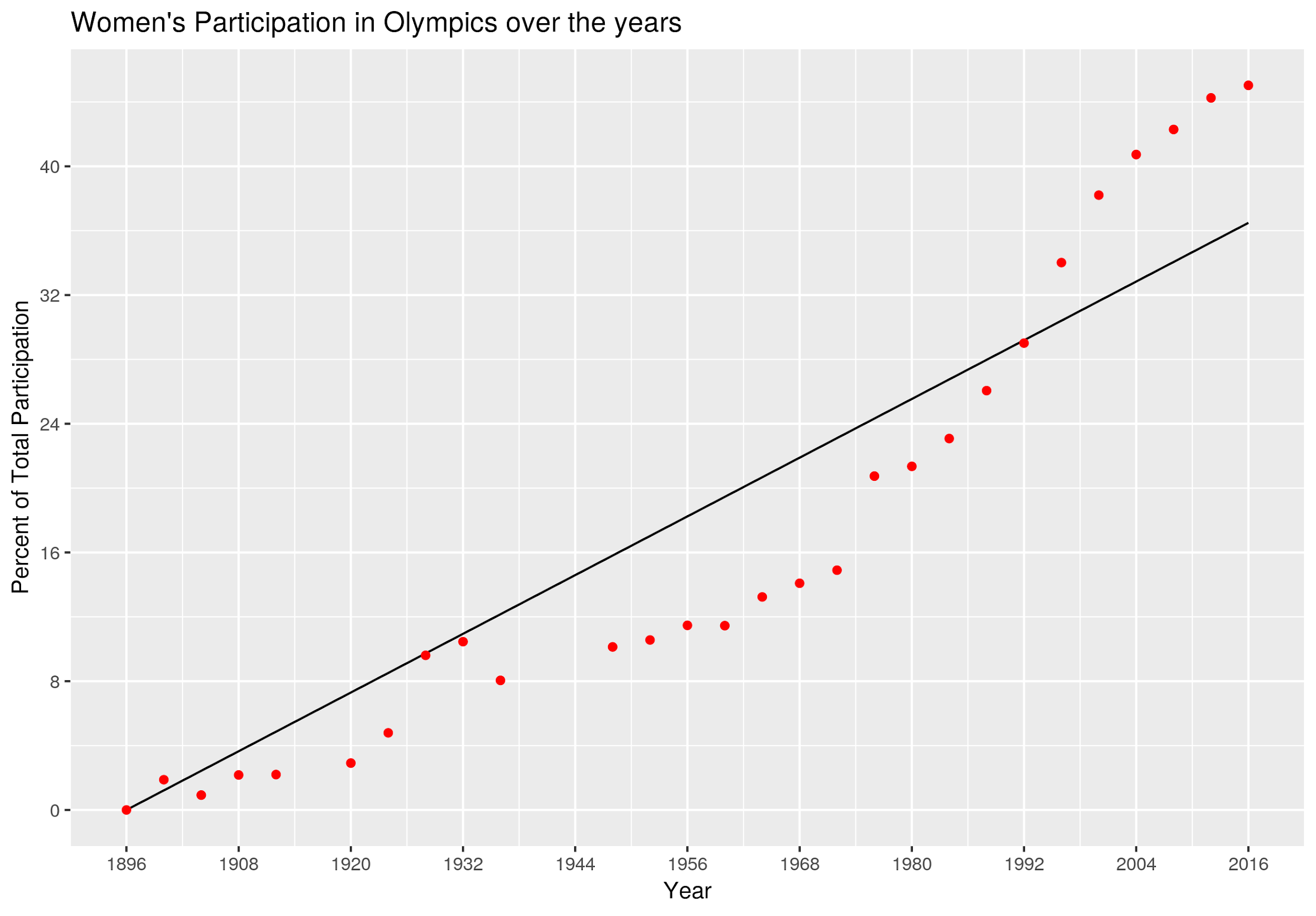


In this first line graph, we are looking at raw data as total number of men in comparison to women. The count started in 1896 to the most recent Olympic games. The graph records a total number of men and women participation in an Olympic at Olympic year intervals.

Even though the graph shows a saw-tooth evolution for men participation, women participation line always falls under men participation line. This means that men have always recorded higher participation at Olympics than women. However, form the year 2010’s it looks like women and men participation converge toward the same amount in number of participation.



This graph above is showing women’s participation as a percent of participation as a whole. As in the previous graph, the percentage of women participation have been growing over the Olympics years until it reaches 50% of the number of participation in the year 2060. This means from 2060 women and men are presenting the same number of participants at the Olympics where men have always recorded the higher percentage in participation over the past Olympics years before 2060. You can see in the graph below the best fit linear model for predicting equal participation.



* Least square line:  
  Proportion = Intercept + Slope \* Year   
  Participation =   1896 + 0.3040617\*year
* Slope = 0.3040617  
  On average, the change in Olympics years is associated with an increase in percentage of women participation until the next Olympic by 1.2%
* Intercept = (1896, 0). The intercept tells us that women’s participation was at 0% in the first Olympic Games. We made it a fixed point for the linear regression.

With regards to main topic at hand, a paired t-test was conducted in order to ascertain whether or not a country won more medals as a host other than not being a host.

The hypothesis:  
 is the difference between the average proportion of medals won as a host country and average proportion of medals won when not hosting Olympics.

Conduct a right tailed t-test to check whether a host country gets any advantage or not. Significance level

Results:

Paired t-test

data:  host\_medal\_proportion$Avg\_Proportion and nonhost\_medal\_proportion$Avg\_Proportion

t = 4.9669, df = 18, p-value = 4.987e-05

alternative hypothesis: true difference in means is greater than 0

95 percent confidence interval:

0.05897526        Inf

sample estimates:

mean of the differences = 0.09060909

From the above results we can see that the p-value is 0.00004987 and at the current significance level we reject the null hypothesis. This is strong evidence for our alternative hypothesis that host country advantage actually exists.

**CONCLUSION AND RECOMMENDATION**

The focus of this research paper is to ascertain whether a host nation for an Olympic Game had an advantage in winning more medals as compared to that country not being a host in a particular Olympic game. This is very important so as to clear the air about bias in the dominance of certain countries over the past years in the number of medals won.

The results from the paired T-test affirmed the hypothesis that there is a host advantage in the number medals won by a country as host as compared to not being a host. Again, certain discoveries were made with respect to the Olympic Games as a whole. It is quite interesting to note that Russia won more medals than the United States of America in the Olympic Games held during the cold war period. This is very essential to note since Russian athletes were less in number as compared to their American counterparts.

Another highlight worth noting is the fact that female athletes have made tremendous improvements in terms of the numbers that show up in the Olympic Games over the past 20 years. There is an estimated catch up to men in the year 2026 Olympic Games. The Country with the most medals won since the Olympic Games begun in 1896 is the U.S.A with a total medal count of 4978 with swimming being the discipline with most medals.

In order for countries to improve upon their medal count there should be a regulation on the number of events in total a country can partake in by gender in the Olympic Games. This is due to the fact that the country with the highest medal count contests in almost every event held in the Olympic Games. This creates a higher likelihood for winning and invariably the total number of medals won.

**REFERENCES**

1. Lieven, Dominic, and Martin McCauley. “Russia.” *Encyclopædia Britannica*, Encyclopædia Britannica, Inc., 20 Oct. 2018, [www.britannica.com/place/Russia/Sports-and-recreation](http://www.britannica.com/place/Russia/Sports-and-recreation).
2. rink\_stats. “Is There Home-Field Advantage At The Olympics?” *FiveThirtyEight*, FiveThirtyEight, 10 Aug. 2016, fivethirtyeight.com/features/is-there-home-field-advantage-at-the-olympics/.
3. “Statistics.” *International Olympic Committee*, 5 Mar. 2018, www.olympic.org/women-in-sport/background/statistics.