Analysis of Home Advantage Bias in the Summer Olympics

Group 2
Aluya Omofuma, Elie Tetteh-Wayoe, Jessica Fogerty, Mihir Gadgil and Pierre Bamba

Presentation Objectives

- Background
- SMART Questions
- Methodology
  - Exploratory Data Analysis
  - Data Visualizations
- Communicate Results
  - What do the results mean
- Summary

Background

- The Olympics games are a prestigious international sports event that feature winter and summer sports.
- Our report focuses on the summer olympic games
- We imported the Olympic medal winners dataset from kaggle
- Background research helped us arrive at our SMART questions.

Background

- We chose to investigate the performance of host nations to see if they had a considerable advantage when hosting.
- In a study conducted by Wachtel and Medvedkov, they reported on the pride Russians took in their great performance in the Olympics.
    - This led us to investigate the performance of Russia during the cold war and compare it to the performance of the USA in that period.
- In the first Olympic games, the IOC did not allow female competitors, so we analyzed the data to study the increased female participation in the Olympics over the years

SMART Qs

Does a country have advantage when hosting Olympics compared to when it isn’t?

- **Specific:** Specifically examining host country wins when hosting and not hosting
- **Measure:** Mean of the proportions of medals won as host vs mean of the proportions of medals won as a non host
- **Answerable:** By a paired t-test
- **Relevant:** For training plans and for sponsors to decide which teams to sponsor
- **Time bound:** Data is already available publicly

SMART Qs

When will men and women participation in the Olympic games be equal?

- **Specific:** Specifically examining when women will have equal participation with men
- **Measure:** Projecting when women will have 50% participation
- **Answerable:** Linear Regression
- **Relevant:** For studying progression and equality
- **Time bound:** Data is already available publicly

Questions of Interest

Did the Cold War have an affect on the participation and performance of the US and Russian teams?

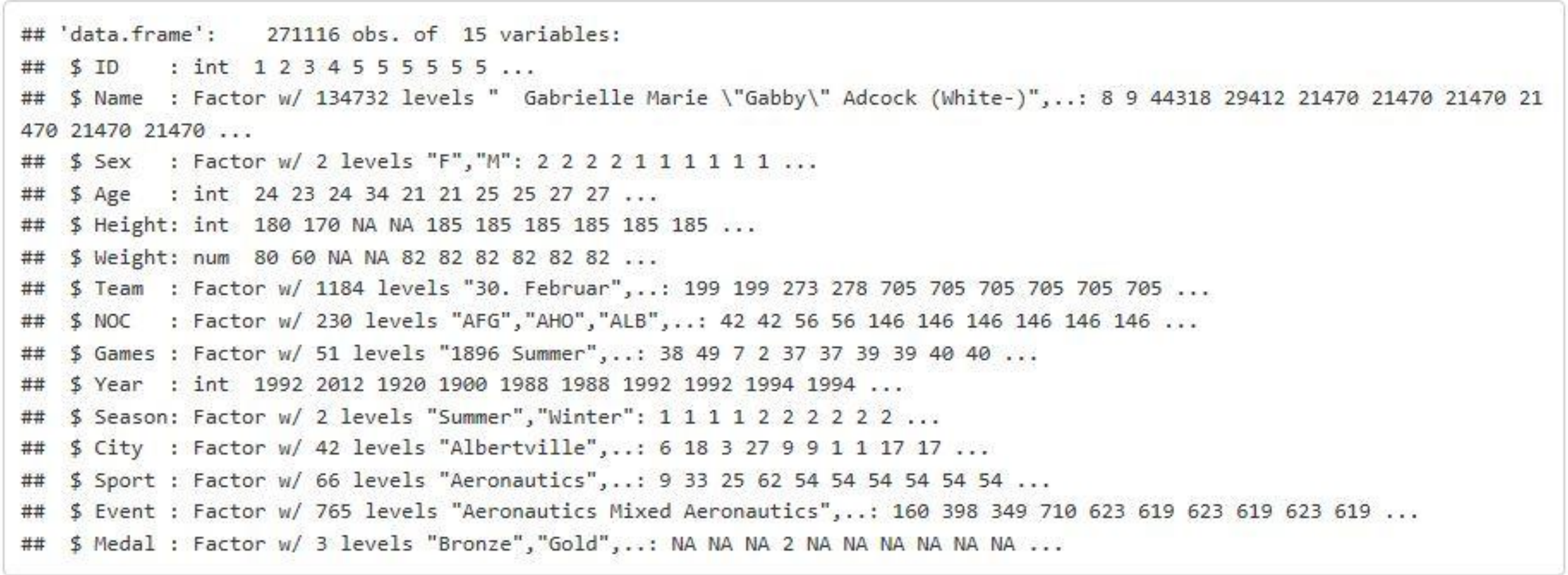
Who are the top 5 medal winning countries?

What was the number of medals earned by each country per discipline?

Methodology

- Load data from: https://www.kaggle.com/heesoo37/120-years-of-olympic-history-athletes-and-results .
- Cleanup data
- Insert host country column into the dataset
- Change some country codes to simplify analysis
- Use the dplyr package to group and summarize the data
- Create plots with ggplot2
- Fit to the data where necessary
- Perform a statistical test where appropriate

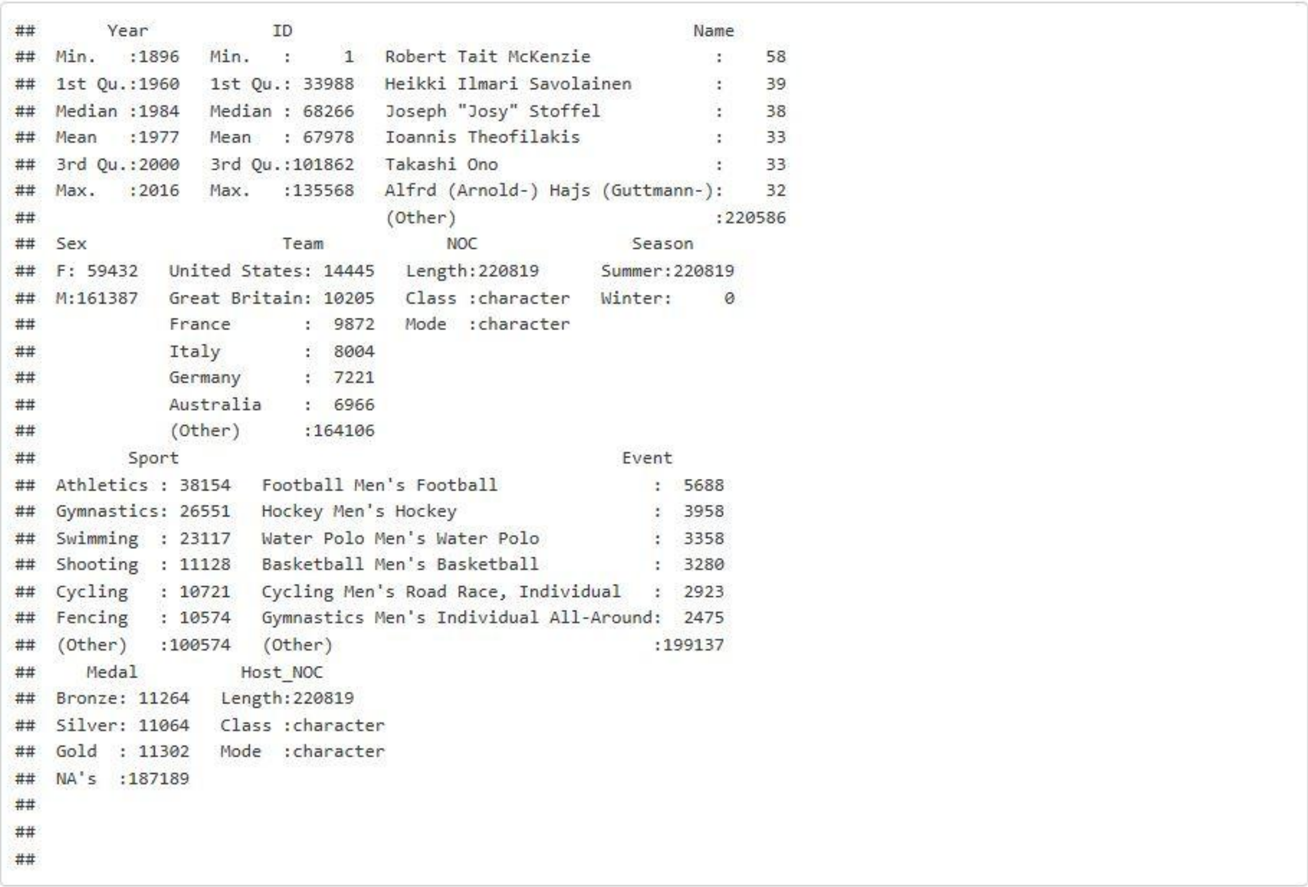
Examine structure of dataset :



```
## '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 21
470 21470 21470 ...
##  $ Sex   : Factor w/ 2 levels "F","M": 2 2 2 2 1 1 1 1 1 1 ...
##  $ Age   : int  24 23 24 34 21 21 25 25 27 27 ...
##  $ Height: int  180 170 NA NA 185 185 185 185 185 185 ...
##  $ Weight: num  80 60 NA NA 82 82 82 82 82 82 ...
##  $ 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 ...
##  $ Games : Factor w/ 51 levels "1896 Summer",..: 38 49 7 2 37 37 39 39 40 40 ...
##  $ 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 ...
##  $ City  : Factor w/ 42 levels "Albertville",..: 6 18 3 27 9 9 1 1 17 17 ...
##  $ 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 ...
```

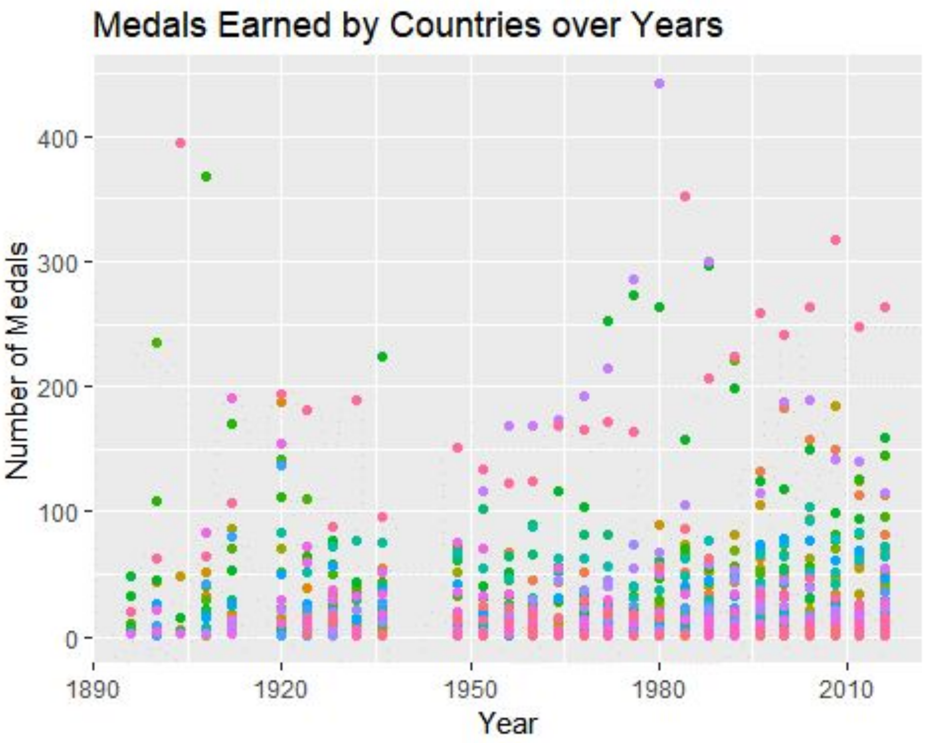
1. We changed the structure of some of the columns such as medal, changed to ordered factor. We dropped Age, Height, Weight, City and Games. Select only the Summer Olympics data.
2. Insert host country column and data into the data frame.
3. Change some country codes to simplify analysis.

Five Summary Statistics

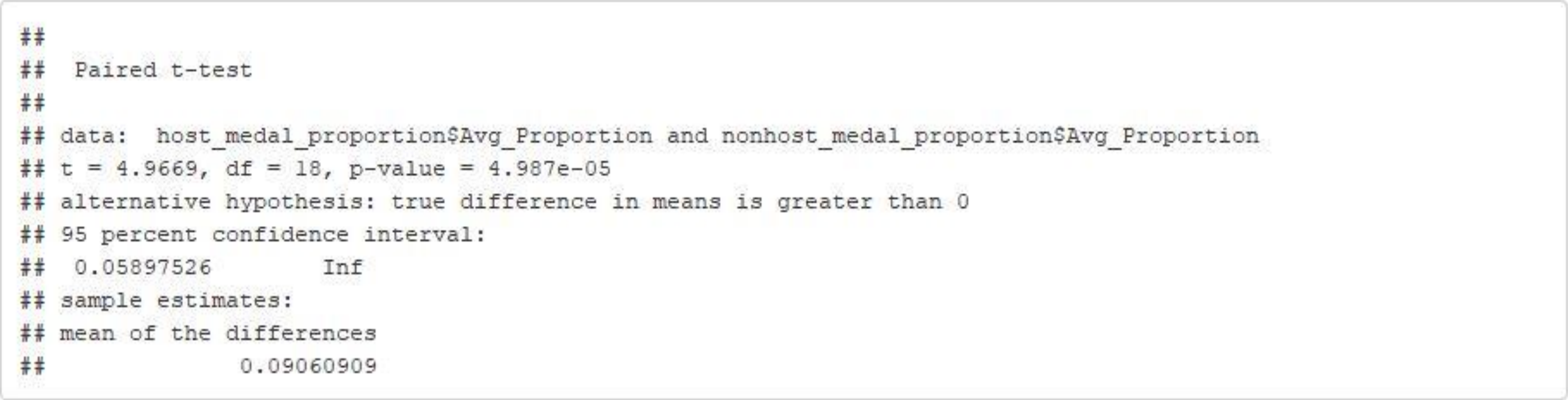


```
##       Year            ID                                       Name
##  Min.   :1896   Min.   :     1   Robert Tait McKenzie            :    58
##  1st Qu.:1960   1st Qu.: 33988   Heikki Ilmari Savolainen        :    39
##  Median :1984   Median : 68266   Joseph "Josy" Stoffel           :    38
##  Mean   :1977   Mean   : 67978   Ioannis Theofilakis             :    33
##  3rd Qu.:2000   3rd Qu.:101862   Takashi Ono                     :    33
##  Max.   :2016   Max.   :135568   Alfrd (Arnold-) Hajs (Guttmann-):    32
##                                  (Other)                         :220586
##  Sex                   Team            NOC               Season
##  F: 59432   United States: 14445   Length:220819      Summer:220819
##  M:161387   Great Britain: 10205   Class :character   Winter:     0
##             France       :  9872   Mode  :character
##             Italy        :  8004
##             Germany      :  7221
##             Australia    :  6966
##             (Other)      :164106
##         Sport                                           Event
##  Athletics : 38154   Football Men's Football               :  5688
##  Gymnastics: 26551   Hockey Men's Hockey                   :  3958
##  Swimming  : 23117   Water Polo Men's Water Polo           :  3358
##  Shooting  : 11128   Basketball Men's Basketball           :  3280
##  Cycling   : 10721   Cycling Men's Road Race, Individual   :  2923
##  Fencing   : 10574   Gymnastics Men's Individual All-Around:  2475
##  (Other)   :100574   (Other)                               :199137
##     Medal          Host_NOC
##  Bronze: 11264   Length:220819
##  Silver: 11064   Class :character
##  Gold  : 11302   Mode  :character
##  NA's  :187189
##
##
##
```

Exploratory Data Analysis: Host Country Advantage



Paired T-Test



```
##
##  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
```

Hypothesis:

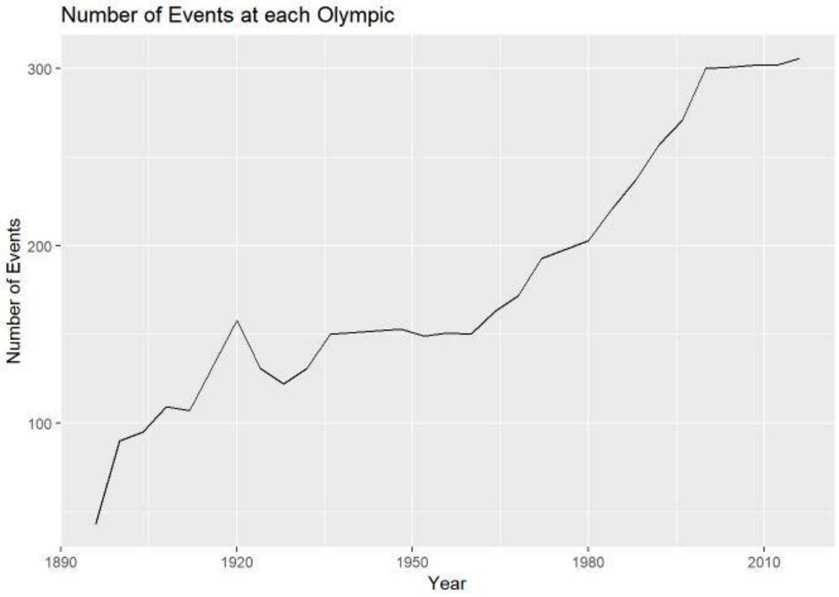
$$H_0\colon \mu_D = 0$$
$$H_1\colon \mu_D > 0$$

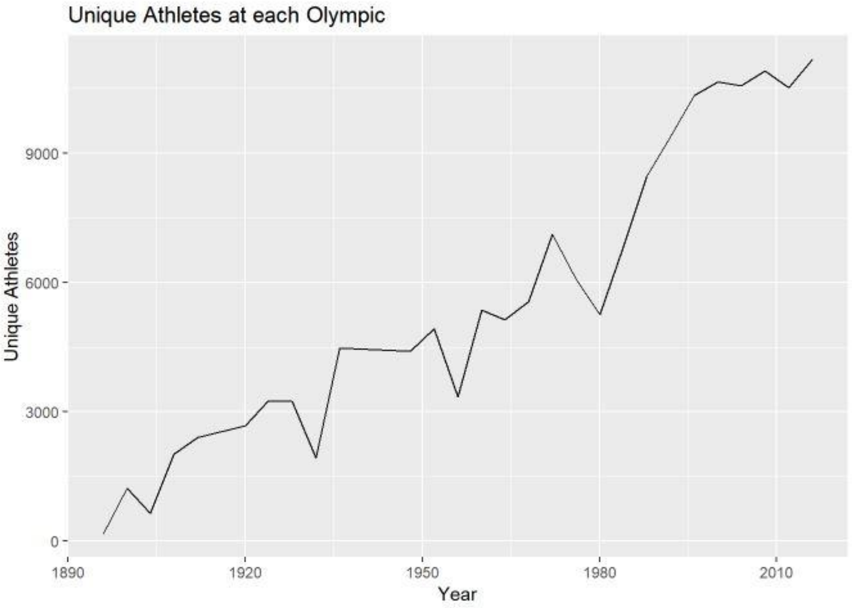
The p-value (4.987×10^{-5}) is smaller than the significance level (0.05)

T-test result interpretation

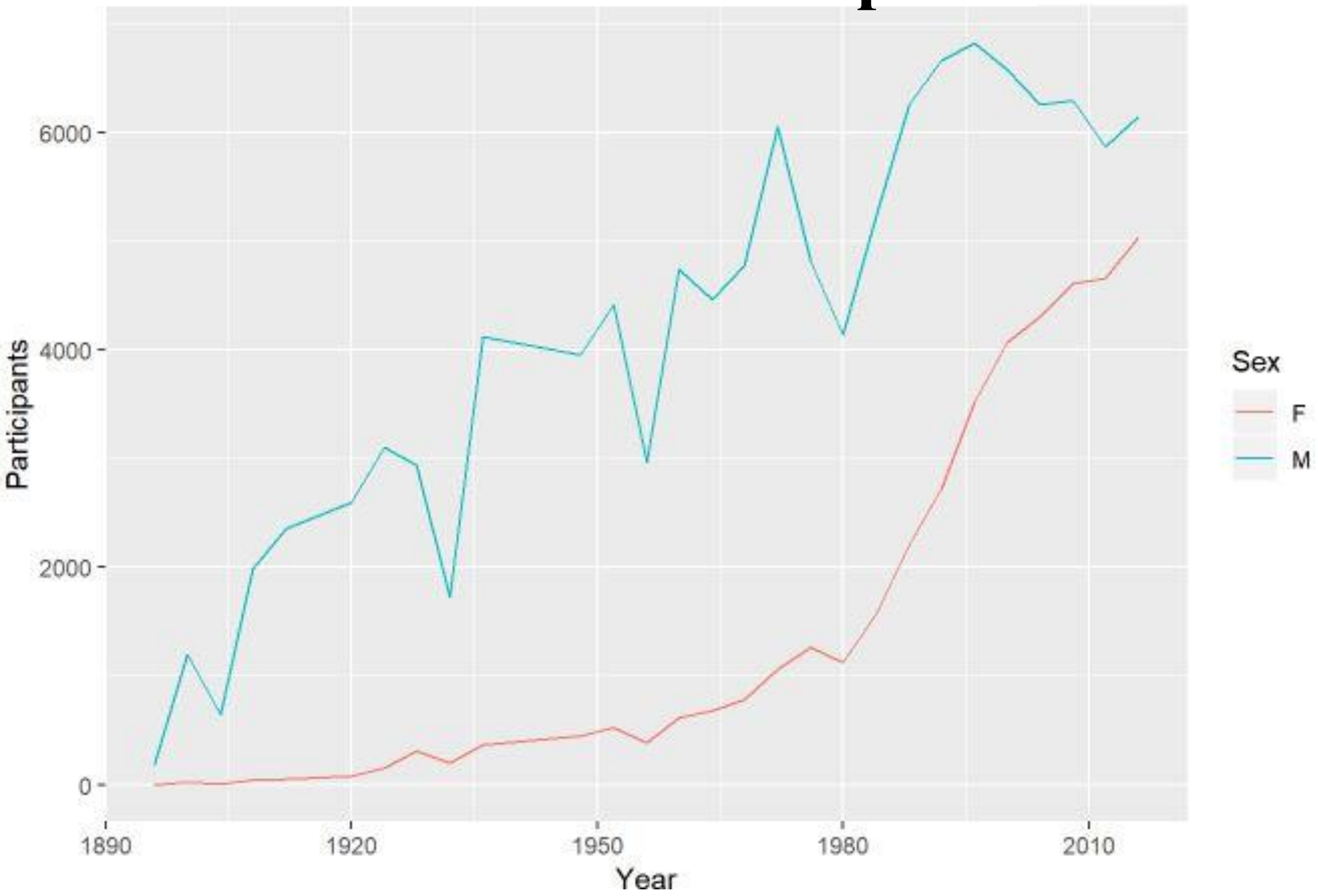
- μ_D is the difference between the average proportion of medals won as a host country and as a non host country
- We Conducted a right tailed t-test to check whether a host country gets any advantage or not. Significance level α=0.05
- Since the p-value (4.987×10^{-5}) is less than the significance level, we rejected the null hypothesis. This is evidence for our alternative hypothesis, that there is home advantage in Olympics

Evolution of Number of Events and Participants

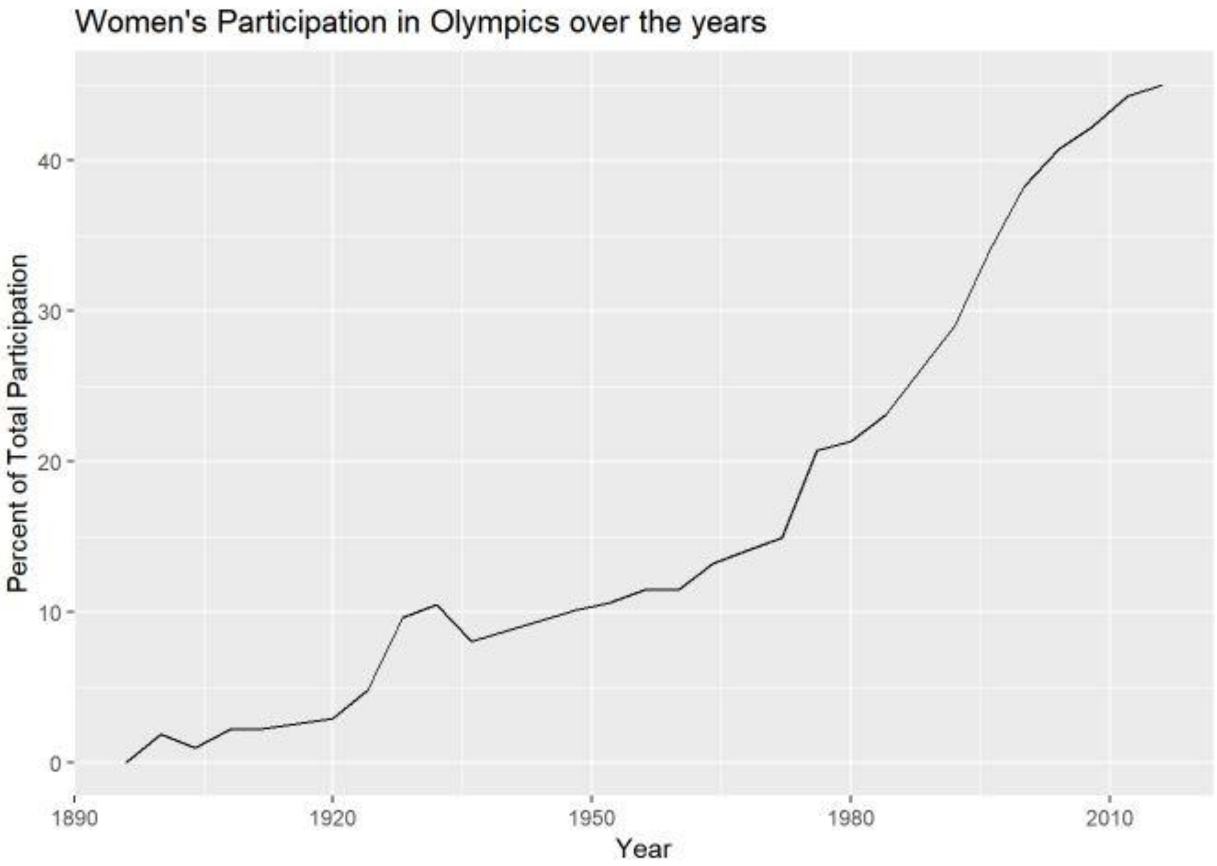




Men vs. Women Participation

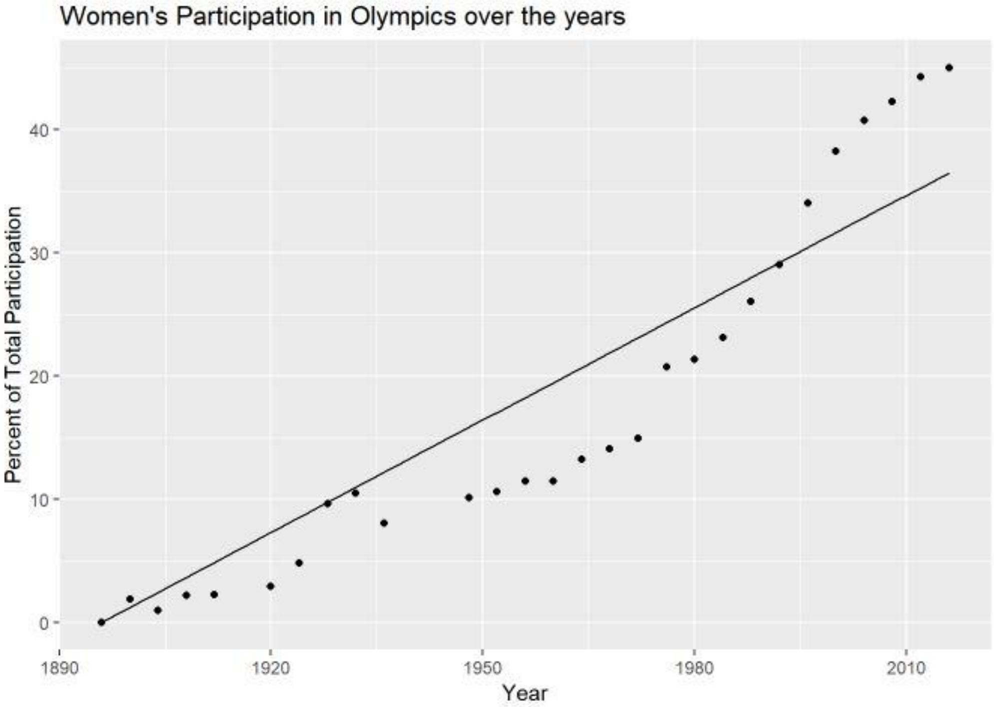


2016: Women participation was at 45%



Linear Model for Women’s Participation

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

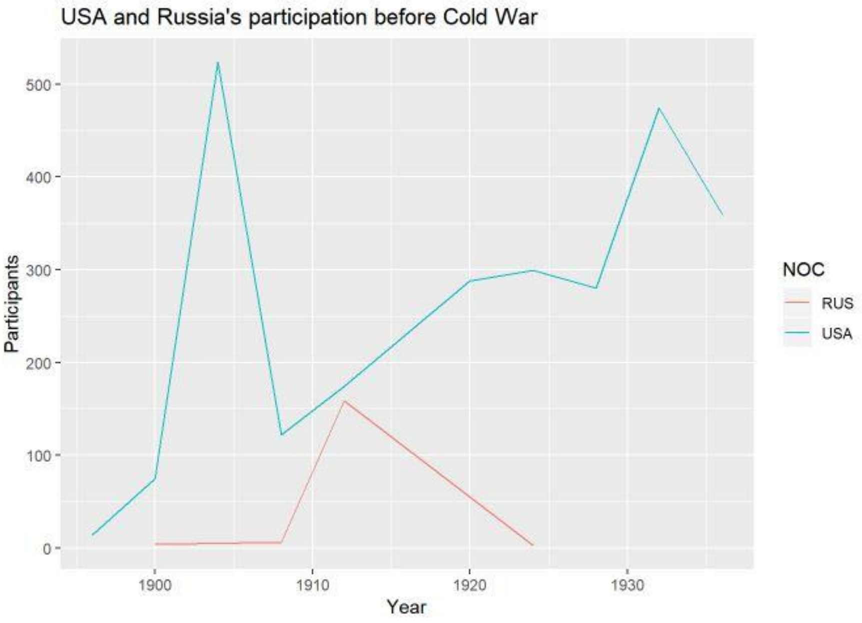


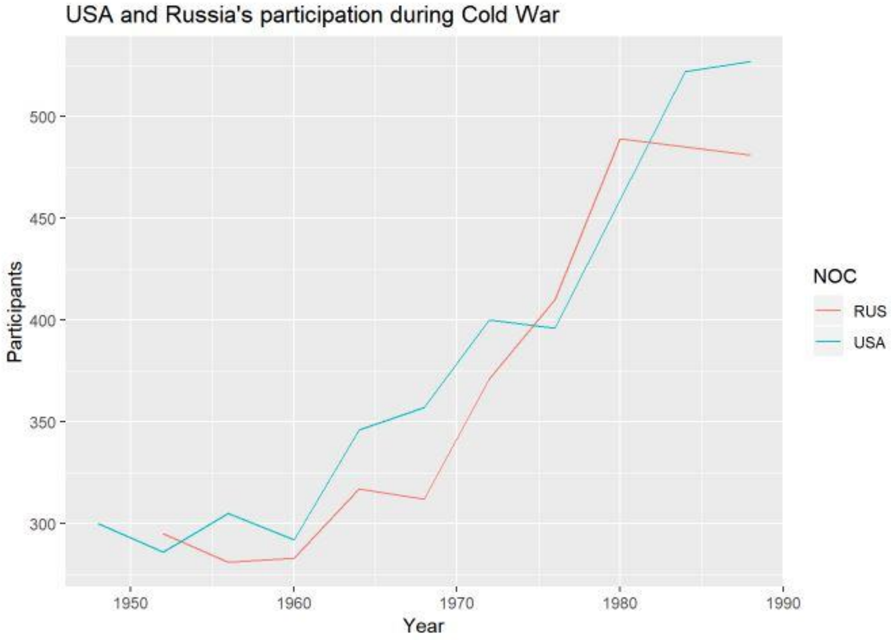
Slope: 0.304
Intercept: (1896, 0)

Multiple R-squared: 0.9431

Expected 50% year: 2060

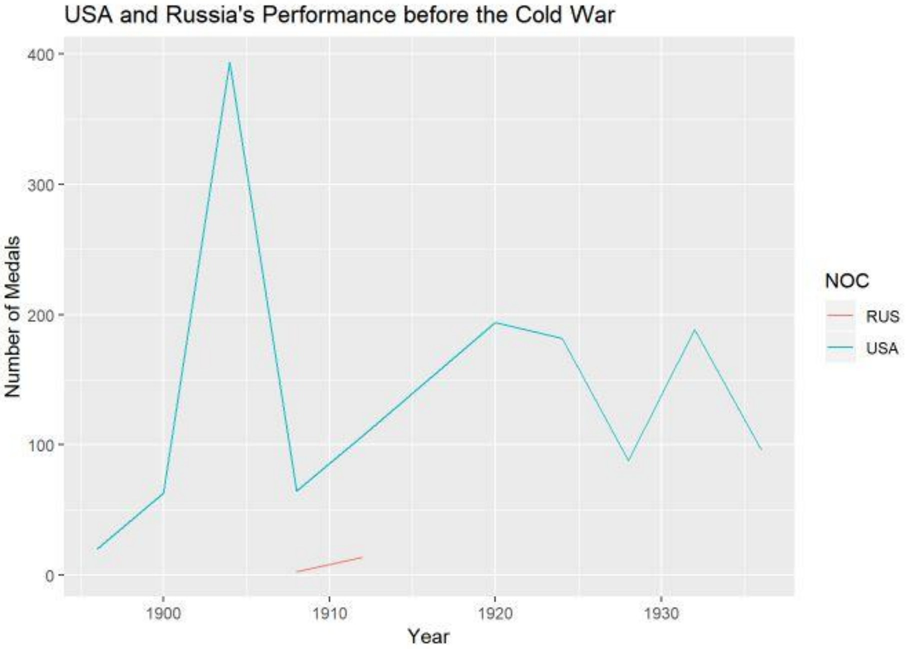
Cold War's Effect on US and Russian Participation

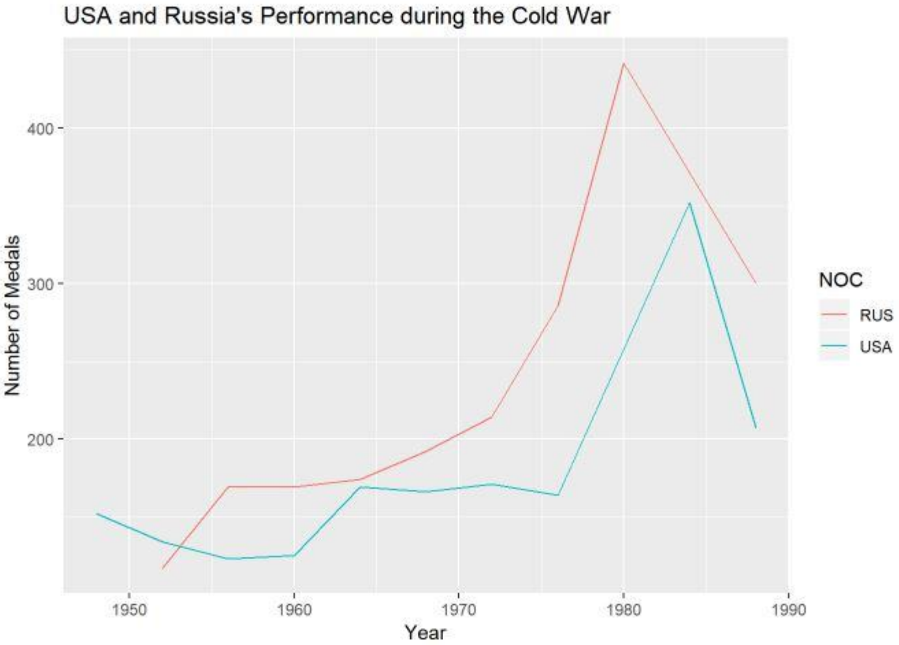




- Number of participants by Russia before the cold war: 172
- Number of participants by USA before the cold war: 2609
- Total Number of participants by Russia: 3239
- Total Number of participants by USA: 3731

Cold War’s Effect on US and Russian Performance



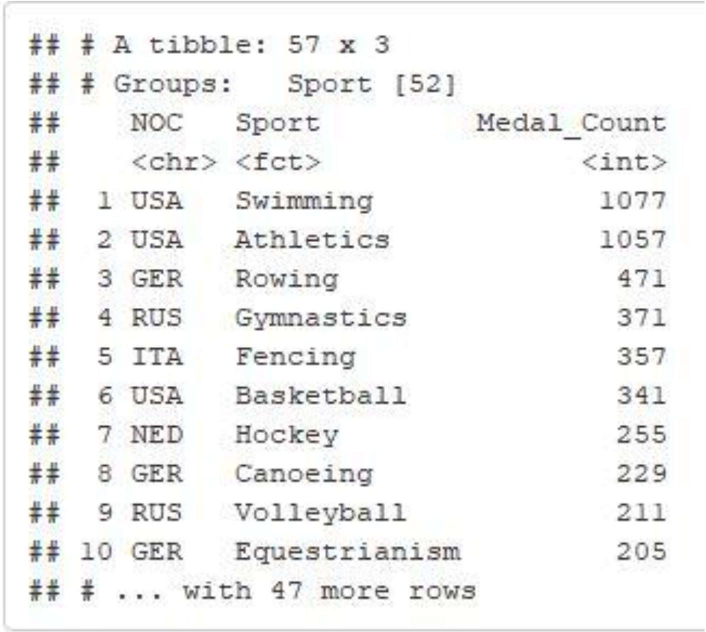


- Number of medals for Russia before the cold war: 17
- Number of medals for USA before the cold war: 1398
- Number of medals for Russia during the cold war: 2063
- Number of medals for USA during the cold war: 1763

Top 5 Medal Winners per Country

1.	USA	4978
2.	GER	3096
3.	RUS	2968
4.	GBR	1946
5.	FRA	1563

Which discipline has the most events?



```
## # 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
```

Summary

- Major highlights from the olympic games:
    - Performance of countries as host and when not as host
    - Women participation throughout the olympics from its inception
    - Cold War effect on Russia and U.S.A
    - Top medal winners(Discipline and Country)
- The focus of this research paper was to determine whether or not a host country had an advantage in the number of medals won other than when not hosting.
- Also, the results from the test performed affirmed the thought had in mind.
- Russia dominated the USA throughout the cold war.
- Again it is quite interesting to note that USA has the highest medal count across all the olympics hosted from its inception.With swimming being the discipline that counts highest to this success.
- In a nutshell, I think there should be a regulation to restrict the number of events a country can partake in so to bring about equal likelihood for each country in chase for a medal.

Questions?