Code ▾

Olympics EDA

Varun Sukumar

August 6, 2018

1 Introduction to the Olympics:
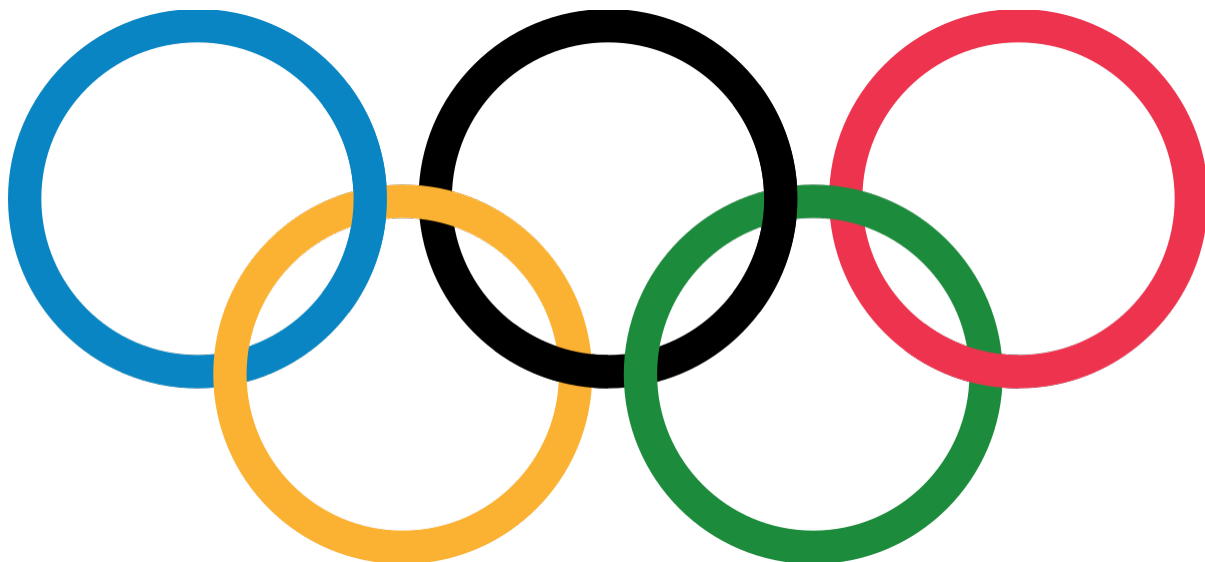
The modern Olympic Games or Olympics are leading international sporting events featuring summer and winter sports competitions in which thousands of athletes from around the world participate in a variety of competitions. The Olympic Games are considered the world's foremost sports competition with more than 200 nations participating. The Olympic Games are held every four years, with the Summer and Winter Games alternating by occurring every four years but two years apart.

The evolution of the Olympic Movement during the 20th and 21st centuries has resulted in several changes to the Olympic Games. Some of these adjustments include the creation of the Winter Olympic Games for snow and ice sports, the Paralympic Games for athletes with a disability, the Youth Olympic Games for athletes aged 14 to 18, the five Continental games (Pan American, African, Asian, European, and Pacific), and the World Games for sports that are not contested in the Olympic Games. The Deaflympics and Special Olympics are also endorsed by the IOC. The IOC has had to adapt to a variety of economic, political, and technological advancements. As a result, the Olympics has shifted away from pure amateurism, as envisioned by Coubertin, to allowing participation of professional athletes. The growing importance of mass media created the issue of corporate sponsorship and commercialisation of the Games. World wars led to the cancellation of the 1916, 1940, and 1944 Games. Large boycotts during the Cold War limited participation in the 1980 and 1984 Games. The latter, however, attracted 140 National Olympic Committees, which was a record at the time.

For more information refer Olympics' official page (https://www.olympic.org/)

2 Introduction about the Data:

This is a historical dataset on the modern Olympic Games, including all the Games from Athens 1896 to Rio 2016. I scraped this data from www.sports-reference.com in May 2018. The R code I used to scrape and wrangle the data is on GitHub. I recommend checking my kernel before starting your own analysis.

3 Content:

The file athlete_events.csv contains 271116 rows and 15 columns. Each row corresponds to an individual athlete competing in an individual Olympic event (athlete-events). The columns are:

1. ID - Unique number for each athlete
2. Name - Athlete's name
3. Sex - M or 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
11. Season - Summer or Winter
12. City - Host city
13. Sport - Sport
14. Event - Event
15. Medal - Gold, Silver, Bronze, or NA

4 Loading the dataset:

5 Loading the required libraries:

```
## Warning: package 'dplyr' was built under R version 3.5.1
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
## Warning: package 'ggplot2' was built under R version 3.5.1
```

6 Initial Exploration of the dataset:

```
## Observations: 271,116
## Variables: 15
## $ ID     <int> 1, 2, 3, 4, 5, 5, 5, 5, 5, 5, 6, 6, 6, 6, 6, 6, 6, 6, 7...
## $ Name   <chr> "A Dijiang", "A Lamusi", "Gunnar Nielsen Aaby", "Edgar ...
## $ Sex    <chr> "M", "M", "M", "M", "F", "F", "F", "F", "F", "F", "M", ...
## $ Age    <int> 24, 23, 24, 34, 21, 21, 25, 25, 27, 27, 31, 31, 31, 31,...
## $ Height <int> 180, 170, NA, NA, 185, 185, 185, 185, 185, 185, 188, 18...
## $ Weight <dbl> 80, 60, NA, NA, 82, 82, 82, 82, 82, 82, 75, 75, 75, 75,...
## $ Team   <chr> "China", "China", "Denmark", "Denmark/Sweden", "Netherl...
## $ NOC    <chr> "CHN", "CHN", "DEN", "DEN", "NED", "NED", "NED", "NED",...
## $ Games  <chr> "1992 Summer", "2012 Summer", "1920 Summer", "1900 Summ...
## $ Year   <int> 1992, 2012, 1920, 1900, 1988, 1988, 1992, 1992, 1994, 1...
## $ Season <chr> "Summer", "Summer", "Summer", "Summer", "Winter", "Wint...
## $ City   <chr> "Barcelona", "London", "Antwerpen", "Paris", "Calgary",...
## $ Sport  <chr> "Basketball", "Judo", "Football", "Tug-Of-War", "Speed ...
## $ Event  <chr> "Basketball Men's Basketball", "Judo Men's Extra-Lightw...
## $ Medal  <chr> NA, NA, NA, "Gold", NA, NA, NA, NA, NA, NA, NA, NA, NA,...
```

Code

```
## Observations: 230
## Variables: 3
## $ NOC    <chr> "AFG", "AHO", "ALB", "ALG", "AND", "ANG", "ANT", "ANZ",...
## $ region <chr> "Afghanistan", "Curacao", "Albania", "Algeria", "Andorr...
## $ notes  <chr> "", "Netherlands Antilles", "", "", "", "", "Antigua an...
```

6.1 Analysis by sex:

6.1.1 Rough Numbers:

The first part of our analysis will be helpful in understanding how the Olympics has evolved over the years. How have men and women been allowed to participate in the competitions overtime.
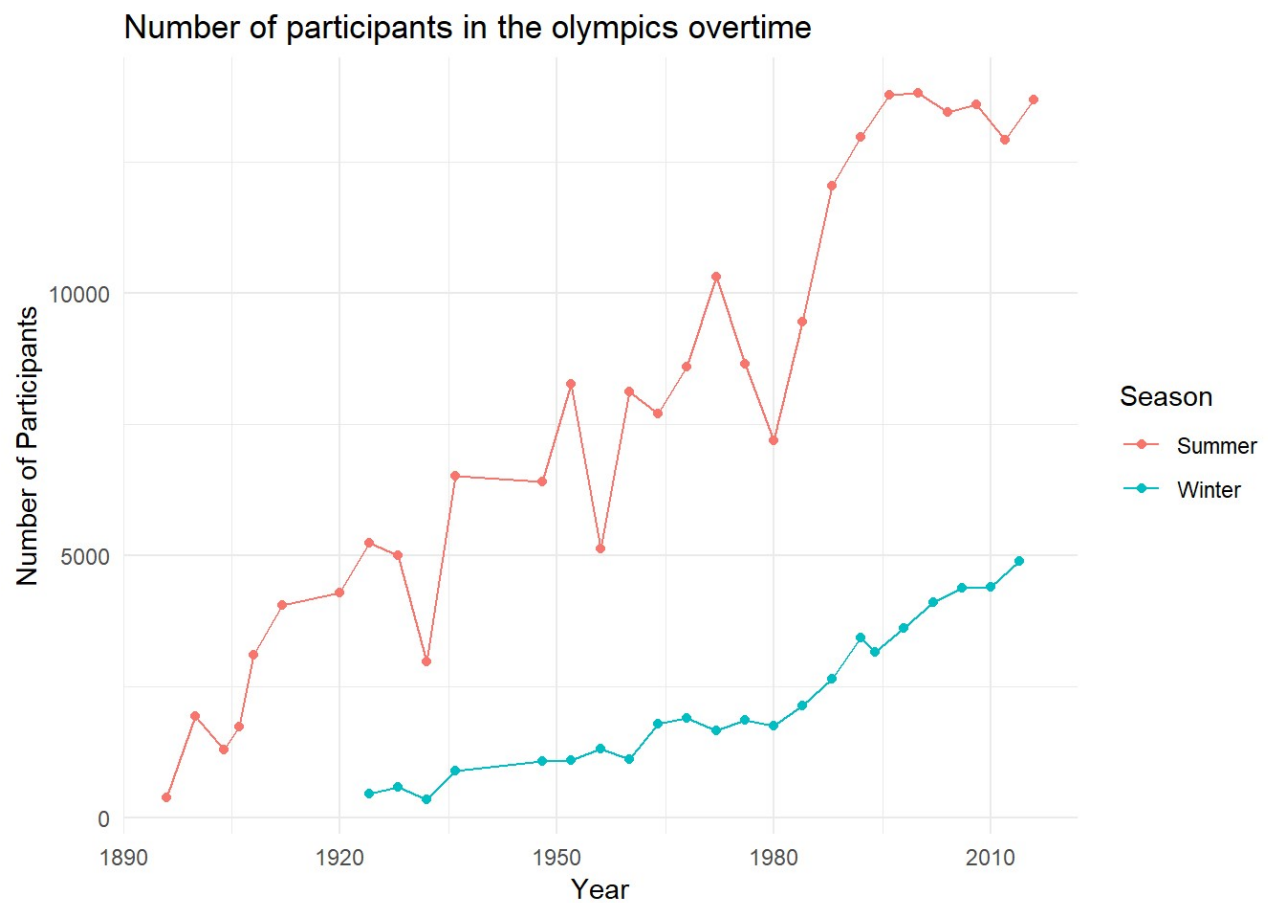
Code

Total number of transfers in a window(2007-2017)
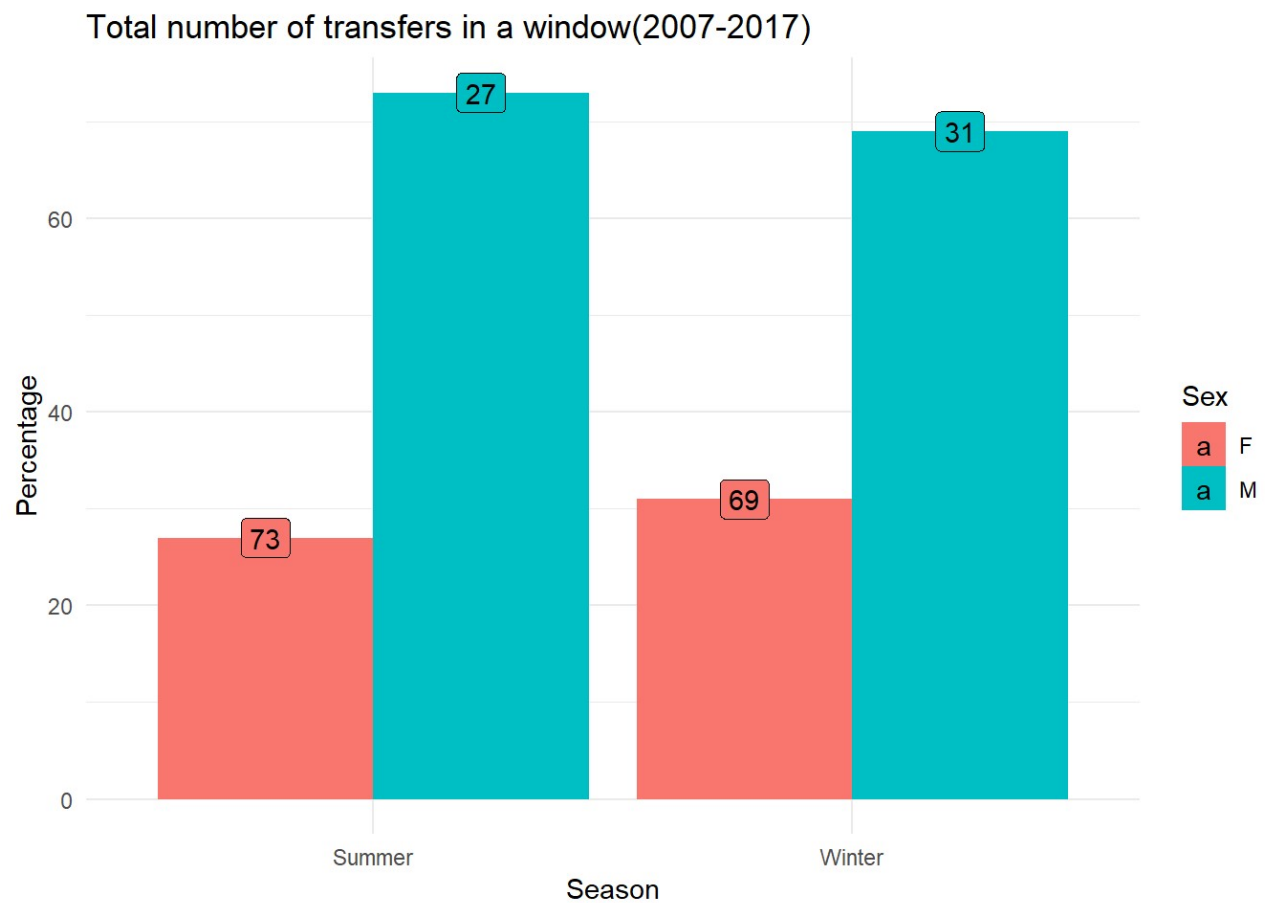
During the winter and the summer the percentage and women have remained the same.

6.1.2 By trend:

6.1.2.1 Overall trend:

As the olymics progressed through the ages the number of people participating in it would have increased, which meant that the number of men and women participating in the olympics grew.

Code

## Number of participants in the olympics overtime

The number of participants in the olympics have grown overtime. It is also obvious that the number of participants in the summer olympics are more than that of the winter olympics.

6.1.2.2 Trend of sex ratio:

Code

```
## Joining, by = c("Year", "Season")
```

Code

Sex Ratio in Olympics through the years

Sex Ratio in Olympics through the years after 1927
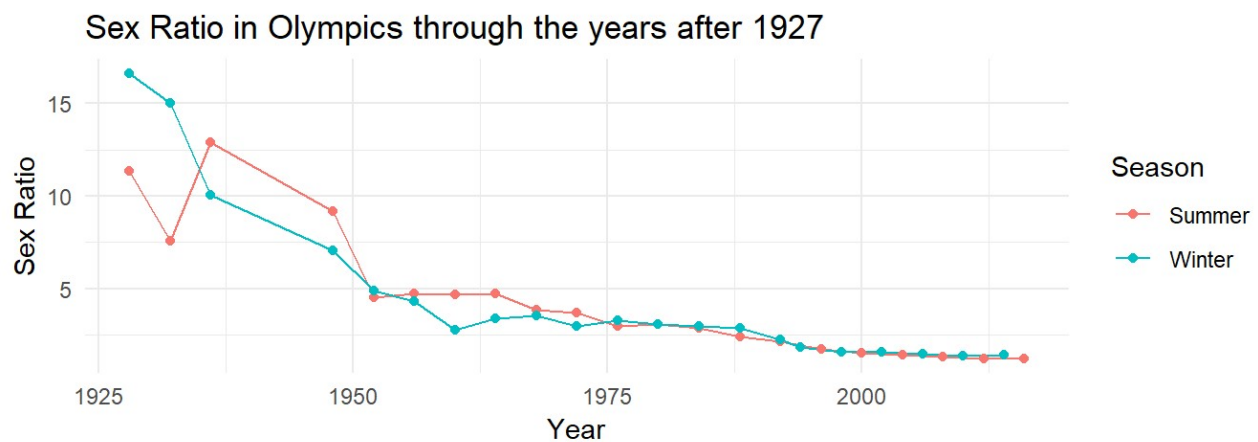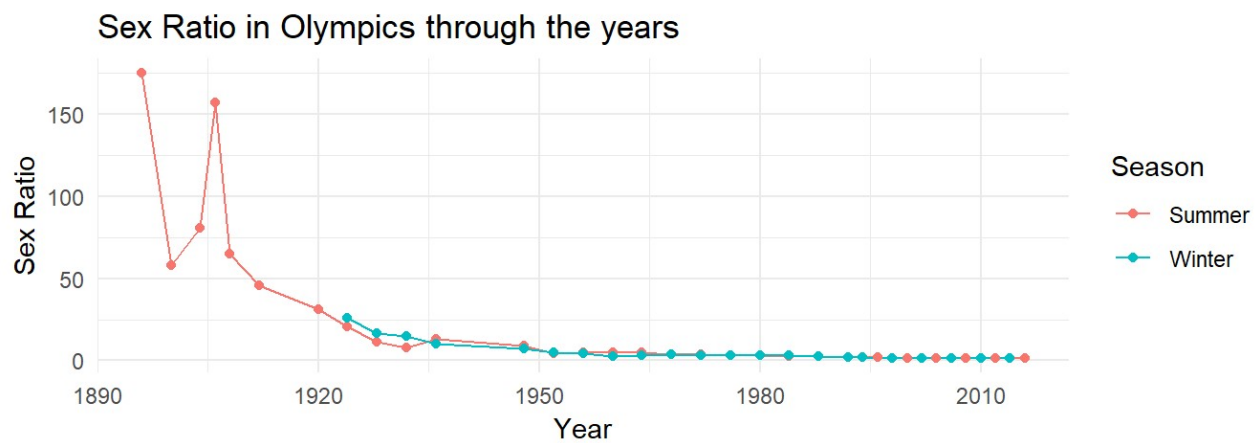
When the olympics started no women participated in the Olympics. In 1900 women started participating in the olympics. As years passed the sex ratio i.e. the ratio of men to women became smaller. After 2000 the ratio started to move towards 1, which means that the olympics now are more diverse than they used to be, which great.

6.2 Analysis by Age

Age is something might have changed from the olymipics started. Hypothesis: The participants during the 1900 of the olympics had a median age greater than the median age of the participants during the 2000's.

6.2.1 Age Density plots

Code

Code

```
## The median age of the athletes in the modern olympics is 24
```

Code

```
## The median age of the male athletes in the modern olympics is 25
```

```
## The median age of the female athletes in the modern olympics is 23
```

The median age for female athletes was lesser than that of median age of male athletes.

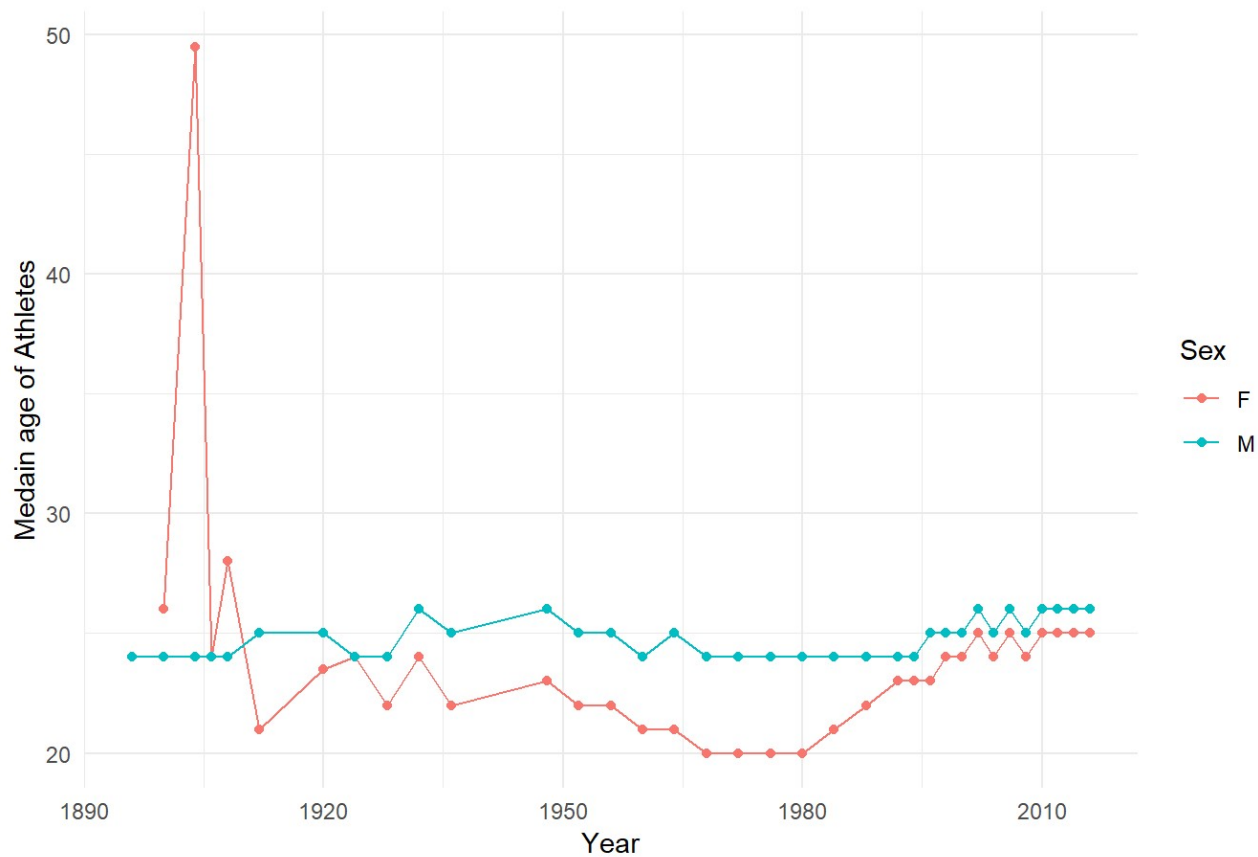## 6.2.2 Age of athletes over the years.

Median age of Male and Female athletes over the years

The median age of men and women participating in the olymipics has increased a bit after the 1980's.