Women at the Olympics

Pratul Singh Raghava

University of Sydney

Contents

Executive Summary

- The aim of this report is to analyze the trends in female participation and results in the Summer Olympics from 1896-2016.
- The main discoveries are :
    1) There has been a steady increase in the number of female participants in the Summer Olympics, with a 500-fold increase in the number of female Olympians and a 50-fold increase in the percentage of female Olympians.
    2) USA has shown the highest numbers of female medal winners over a century of the Olympic games, with nearly 3-times more female Olympic medals than its closest competitor.
- Soviet Union still holds the 2nd place for all-time number of female medals won in Olympics history, even though it hasn't participated since 1988.
- Developed countries have more successful Olympic programs and result in more number of female medals won than developing or third world countries.

Full Report

Initial Data Analysis (IDA)

- The data came from R. H. Griffin's data repository on Kaggle, for which they had originally scraped the data from www.sports-reference.com/olympics.

- The data is valid because the facts and numbers mentioned in the data are easily verifiable on the internet. Furthermore, R. H. Griffin is a reputed Data Scientist from Harvard and this supports the quality and integrity of the data compiled by them.
- Possible issues include :
    1) Incomplete data from early 1900s due to lack of technology and means of data collection. This may have have resulted in under reporting of numbers and maybe even false reporting since there are no means or records to confirm the data from that period.
    2) Duplicate rows for the same participant in the same edition of the Games, if the said participant participated in more than one events. However, this issue doesn't affect the quality of the analysis being done since this report aims to measure the degree of female participation in the Olympics and therefore, participation in multiple events by even the same athlete is included in the domain of the objective.
- Potential stakeholders include aspiring future female Olympians around the world, the International Olympic Committee, the Olympic committees of developed nations, former female Olympians.
- Each row represents an entry in the one of the editions of the Summer or Winter Olympics. This entry row contains all possible information about the player and the event they are participating in.
- Each column represents the participant's Name, Gender, Age, Height, Weight, Country, Edition of Olympics, Year of Olympics, Season, Host City, Sport, Event, Medal won(if any).
- The key variables are Sex, Team, NOC, Season, Medal, Year.
    1) Sex, a character variable, holds the value of either 'M' or 'F'. it represents the gender of the participant and helps us to classify them as Male or Female so that they be distinctly categorised for further analysis. Team, a character variable, holds the value of name of the participant's country and helps us to categorise the participants by their countries.
    2) NOC, a character variable, holds the value of a 3-letter abbreviation of the participant's country and helps us to categorise the participants by their countries with much more ease. Categorisation using this is easier to do than by Team, since the name of various countries are often long and prone to be misspelled.
    3) Season, a character variable, holds the value of either "Summer" or "Winter". This variable helps us to filter out data for the Summer editions of the Olympics since that is the domain of this project.
    4) Medal, a character variable, holds the value of either "Gold", "silver" or "Bronze". While the absolute value of the variable is not analysed, it's the count of these medals that is important to our analysis.
    5) Year, an integer variable, holds the value of the year of occurrence of each edition of the Olympics and helps us to see data and their analyses for each edition quite easily.

```
## read in data
athlete=read.csv("athlete_events.csv")

##Load R packages
library("ggplot2")
library("dplyr")
library("tidyverse")

## show classification of variables
str(athlete)
```

```
## 'data.frame':    271116 obs. of  15 variables:
##  $ ID    : int  1 2 3 4 5 5 5 5 5 5 ...
##  $ Name  : chr  "A Dijiang" "A Lamusi" "Gunnar Nielsen Aaby" "Edgar Lindenau Aabye" ...
##  $ Sex   : chr  "M" "M" "M" "M" ...
##  $ 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  : chr  "China" "China" "Denmark" "Denmark/Sweden" ...
##  $ NOC   : chr  "CHN" "CHN" "DEN" "DEN" ...
##  $ Games : chr  "1992 Summer" "2012 Summer" "1920 Summer" "1900 Summer" ...
##  $ Year  : int  1992 2012 1920 1900 1988 1988 1992 1992 1994 1994 ...
##  $ Season: chr  "Summer" "Summer" "Summer" "Summer" ...
##  $ City  : chr  "Barcelona" "London" "Antwerpen" "Paris" ...
##  $ Sport : chr  "Basketball" "Judo" "Football" "Tug-Of-War" ...
##  $ Event : chr  "Basketball Men's Basketball" "Judo Men's Extra-Lightweight" "Football Men's Footbal
##  $ Medal : chr  NA NA NA "Gold" ...
```

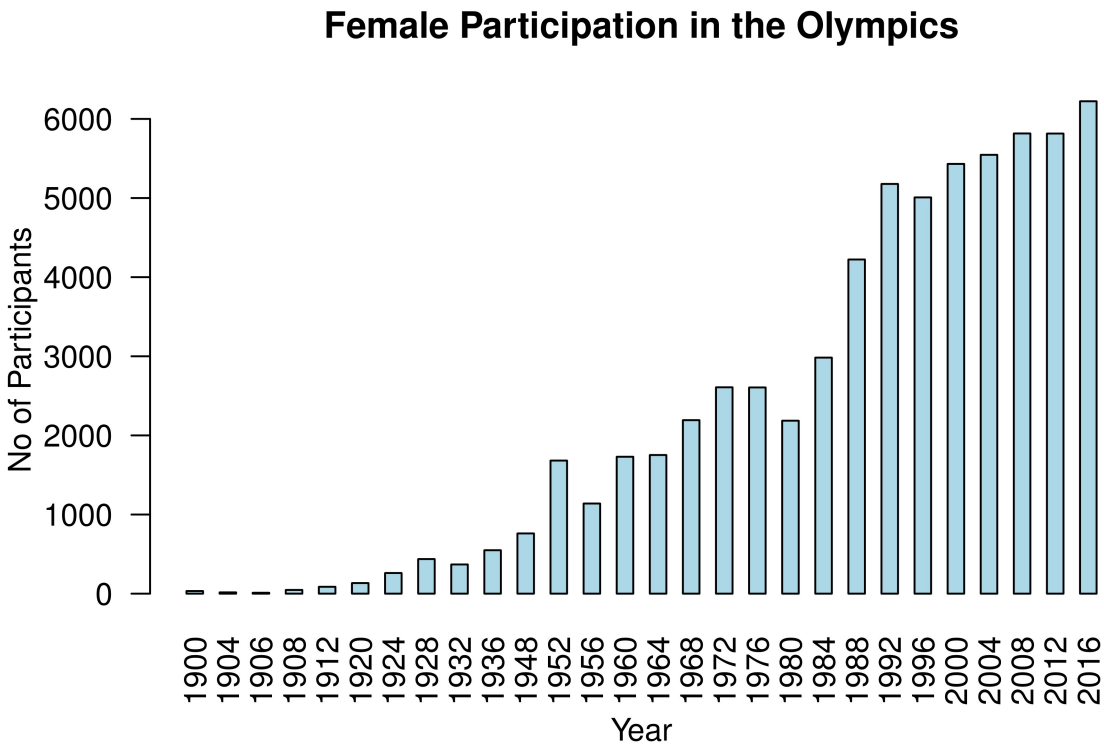
```r
##Store attributes in variables
year=athlete$Year
gender=athlete$Sex
```

Research Question 1 : How has female participation in the Olympics evolved since the 1890's ?

-> We aim to analyse the trends of female participation in the Summer Olympics from the 1890s up until 2016. We achieve this by producing multiple charts that look at both the absolute number of female participants and their ratio among all participants in each edition of the Summer Olympics.

First, we produce a bar plot to analyse the trends observed in the absolute number of female participants in each edition of the Summer Olympics. We represent Year on the X-axis and Number of participants on the Y-Axis such that we get vertical bars for each edition with the height of the bars signifying the number of female participants in that edition.

```r
#Excluding the Winter Olympics
gender[year==2014]<-0
gender[year==2010]<-0
gender[year==2006]<-0
gender[year==2002]<-0
gender[year==1998]<-0
gender[year==1994]<-0

#Including only Female Participants
g2=gender[gender=="F"]
y2=year[gender=="F"]

x2=table(g2,y2)

#Producing the Bar Plot
barplot(x2,main = "Female Participation in the Olympics", xlab = "Year", ylab="No of Participants",
    col = c("lightblue"),las=2,beside=TRUE)
```



Summary : On producing this bar plot, we observe that there has been a steady increase in female participation numbers over the years with little to none instances where numbers have gone down in consecutive editions of the Olympics. We see that the bar plot shows female numbers only from 1900 onwards even though the data supplied was from 1896 onwards, implying that at the 1896 Games, there were no female participants at all. Even in the early 1900s, female participation was minimal, with only 11 females in the 1906 edition to 1952 being the first edition with over 1000 females participating. Since then, the numbers have only increased with 6000 females participating in the latest edition in 2016, indicating a near 500x increase over a 110-year period.

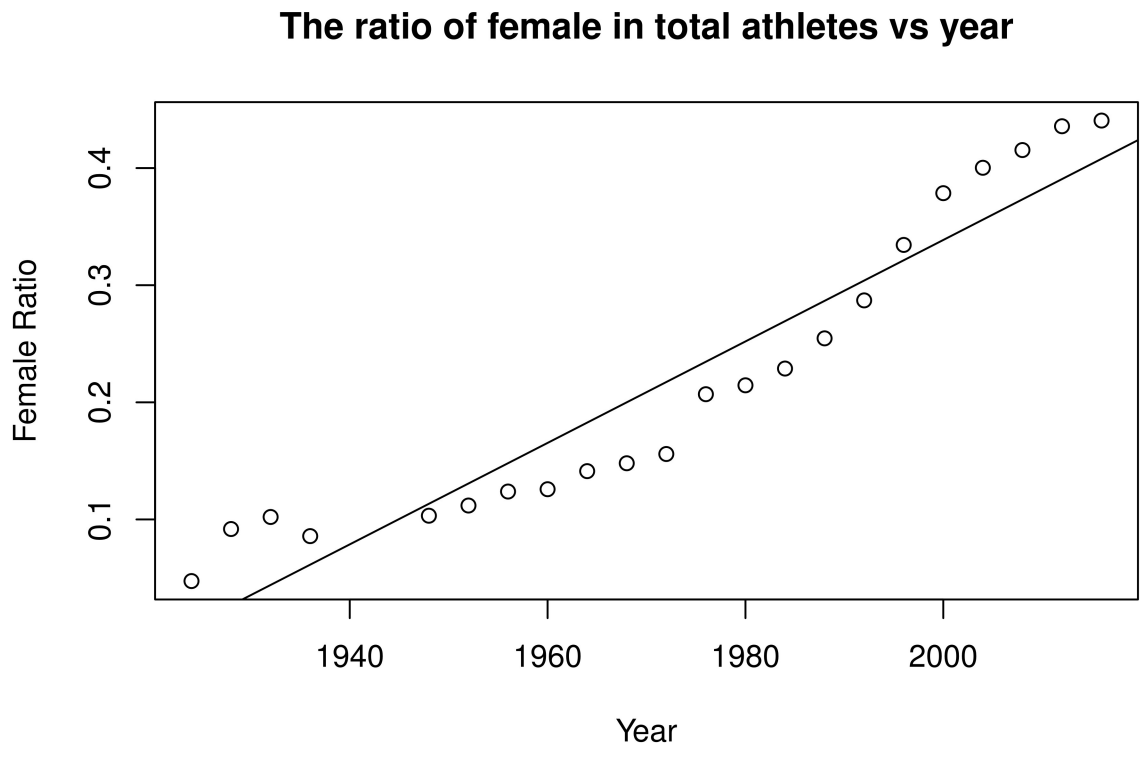
Second, we produce a linear regression model to observe what percentage of total participants have the females constituted in the Olympics over the years. We do so through a scatter plot produced by plotting the ratio of female participants to the total participants on the Y-Axis against the X-Axis representing the Year for each edition. We then plot a regression line which best represents the data plotted on the scatter plot. The correlation coefficient for the data points is :

```
## To match the year of summer games, the year of winter games after 1992 is re-coded
## Thus, each Olympics occurred
initial <- c(1994,1998,2002,2006,2010,2014)
 new <- c(1996,2000,2004,2008,2012,2016)
for (z in 1:length(initial)) {athlete$Year <- gsub(initial[z], new[z], athlete$Year)}
athlete$Year <- as.integer(athlete$Year)

## Group and count the number of athletes according to year and sex
count_sex <- athlete %>% filter(Year>1920) %>% group_by(Year, Sex) %>%
summarize(Number = length(unique(ID)))

## Filter female athletes from data group
```

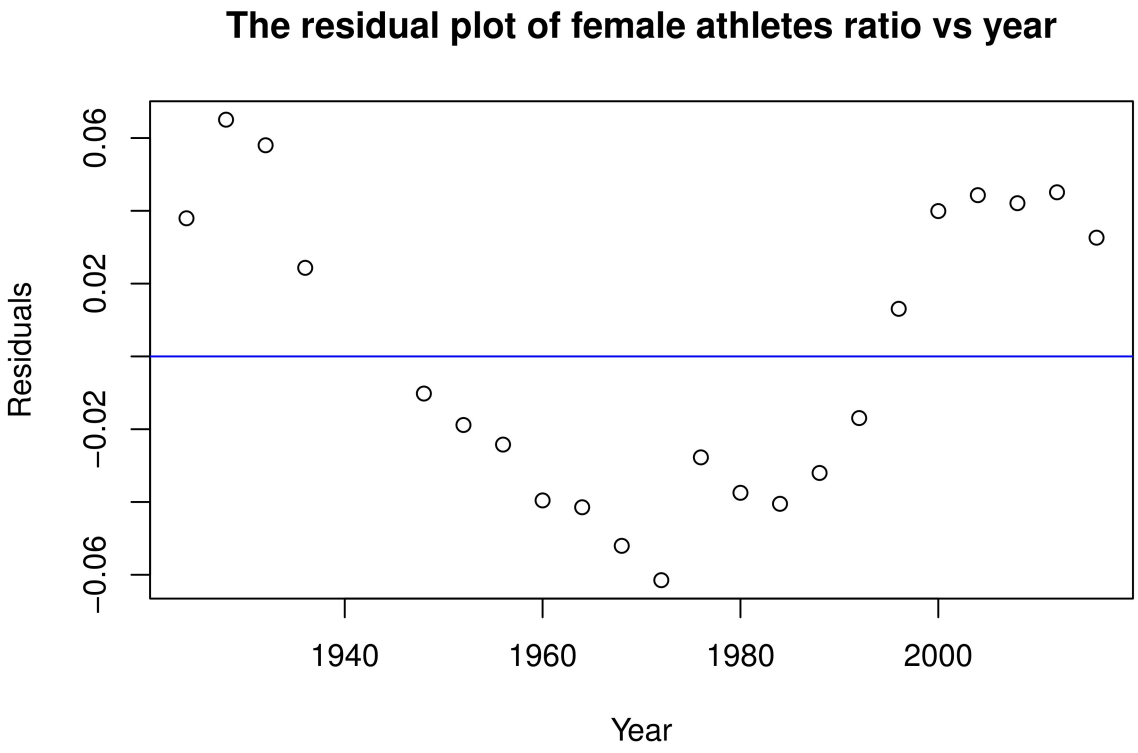
```r
Female <- count_sex %>% filter(Sex == "F")
## Count the total Participants by Year
Total_athletes <- athlete %>% filter(Year>1920) %>% group_by(Year) %>% summarize(Athletes = length(uniq
## Find the female ratio in athletes
Female_Ratio <- Female %>% left_join(Total_athletes) %>% mutate(female_ratio = Number/Athletes)

## Plot female ratio vs year Basic
cor(Female_Ratio$Year, Female_Ratio$female_ratio)
```

```
## [1] 0.9495026
```

```r
L = lm(Female_Ratio$female_ratio ~ Female_Ratio$Year)
plot(Female_Ratio$Year, Female_Ratio$female_ratio, xlab = "Year", ylab = "Female Ratio", main = "The ra
abline(L)
```




```r
## Residual plot for female athletes ratio vs year
plot(Female_Ratio$Year, L$residuals, xlab = "Year", ylab = "Residuals", main = "The residual plot of fe
abline(h = 0, col = "blue")
```



The correlation coefficient for female athletes ratio and Year is close to 1, which indicate a strong positive linear relationship between female athletes ratio and year.This is verified by the Gender Equality Report by the International Olympic Committee themselves. To demonstrate this relationship, a linear regression graph has been established and the regression line shown that the female athletes ratio increase over years. However, the distance between actual points and the regression trend is distinct. There are less clustering points. Thus, a residual plot has been created to detect the undiscovered patterns.As a result, a random scatter is shown in the residual plot which represents that the linear model is likely appropriate.

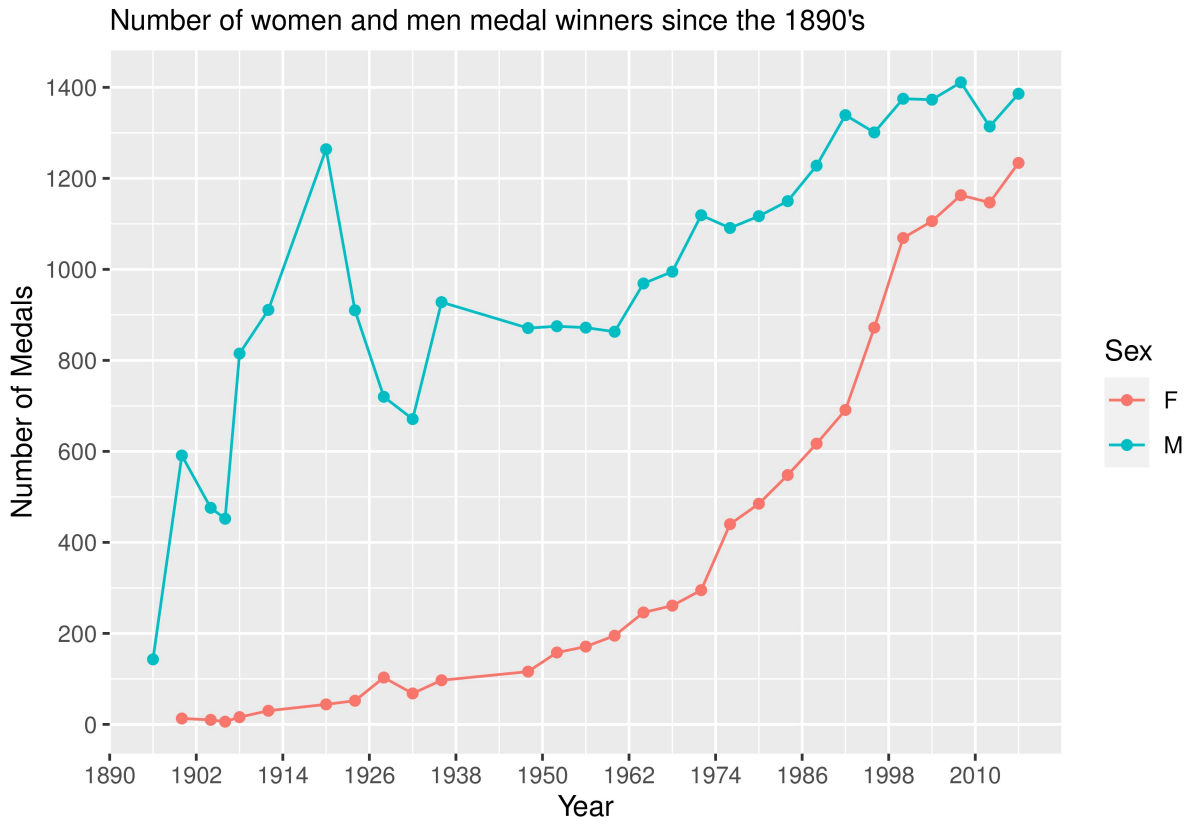
Research Question 2 : Which countries have produced the highest number of medal winning female Olympians?

-> We aim to find out the most successful countries in terms of medals won by female participants throughout Olympic history. We achieve this by producing multiple charts to compare every edition's medals by gender and then to see the top 10 countries with the highest number of female Olympic medals ever.

First, we produce a line chart to show a comparison of total medals won by females compared to those won by males in each edition of Olympics. We represent Year on the X-axis and Number of participants on the Y-Axis such that we get two different lines on the graph representing the number of medals won by each gender over the years in the Summer Olympics. A legend on the side assigns colours to each line to represent different genders.

```
## Filter and Count the number of medals by year and sex
medal_sex <- athlete %>% filter(!is.na(Medal)) %>% group_by(Year,Sex) %>% summarize(Medals = length(Medal

## Plot the number of female/male medal winners since the 1890's
ggplot(medal_sex, aes(x = Year , y = Medals, colour = Sex ,
```

```
group = Sex)) + geom_line(aes(color = Sex)) +
geom_point(aes(color = Sex)) +
scale_y_continuous(breaks = seq(0, 1500, by = 200)) + scale_x_continuous(breaks = seq(1890, 2016, by =
labs(x = "Year", y = "Number of Medals", title = "Number of women and men medal winners since the 1890'
theme(plot.title = element_text(size = 11))
```

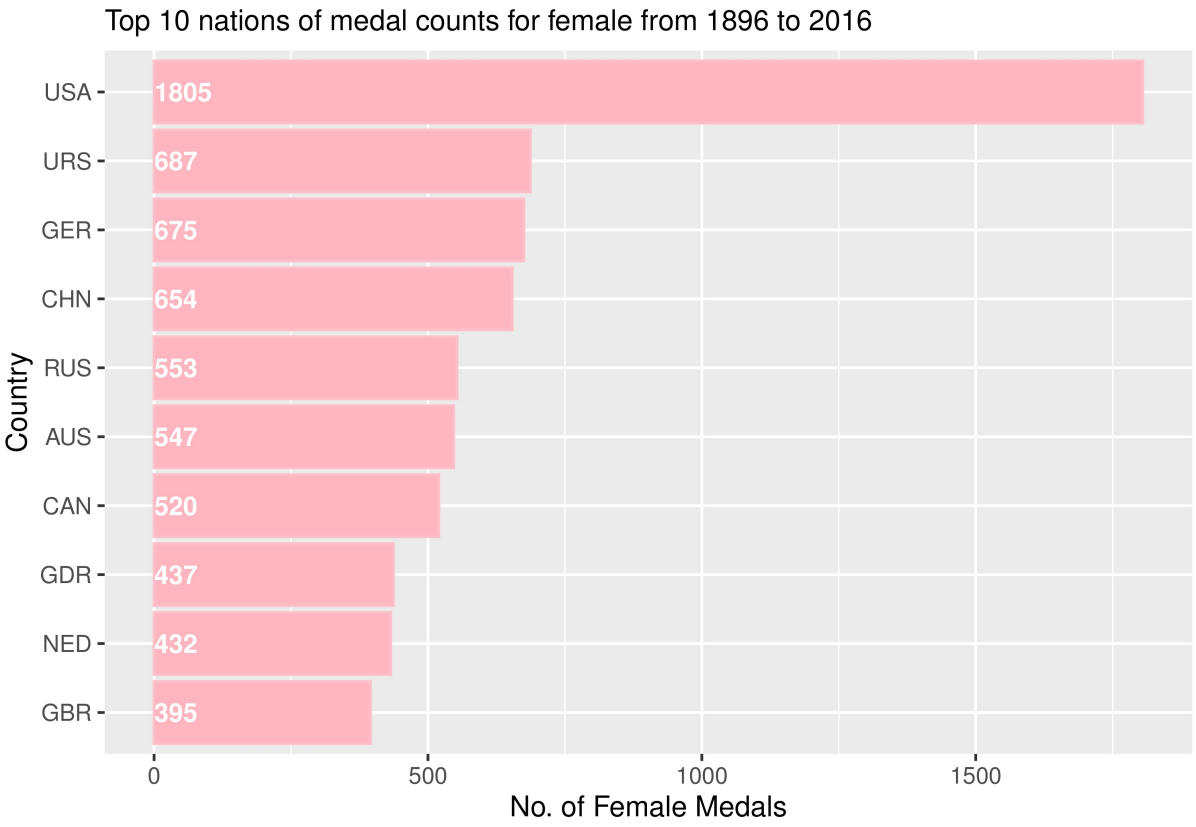


Summary: On producing this line chart, we observe that although in the early 1900s there was a huge disparity between the number of medals won my males and females, over the years, this gap has reduced to a very small number and over the past decade the trends shown by both the lines are also the same.

Second, we produce a horizontal bar plot to list the top 10 nations with the highest number of medals won by female Olympians. We represent the number of medals on the X-Axis and the individual top 10 countries on the Y-Axis to produce horizontal bars with their lengths proportional to the number of female medals they won.

```
## Count the top 10 nations with greater female medal counts
female_medal<- athlete %>%filter(!is.na(Medal), Sex == "F") %>%
group_by(NOC) %>% summarise(Medal_count = length(Medal)) %>% arrange(desc(Medal_count)) %>% ungroup() %
mutate(NOC_new = reorder(NOC, Medal_count)) %>% slice(10:1)

## Plot the Top 10 nations of medal counts for female
ggplot(female_medal, aes(x = NOC_new ,y = Medal_count)) +
coord_flip() +
geom_bar(stat = "identity",colour = "pink", fill = "lightpink") +
geom_text(aes(x = NOC_new, y = 0.2, label = paste(Medal_count)),
hjust = 0, size = 3.5, colour = "white", fontface = "bold") +
```

```
labs(x = "Country",  y = "No. of Female Medals", title = "Top 10 nations of medal counts for female fro
theme(plot.title = element_text(size = 11))
```



On producing this bar plot, we see that the United States with 1805 medals takes top position with a huge gap to the others, having won nearly 3-times more female medals than its closest competitor, the Soviet Union, which although hasn't participated since 1988, still holds second place on the list with 687 medals. Other countries that feature on the list have a record like that of Germany with not man differences among them. One interesting feature that emerges in this list is that the countries listed are all developed ones: be it European countries like Germany, former East Germany and Netherlands or global superpowers like USA, Russia, China, Australia, former Soviet Union. This indicates that as expected, developed countries have the most successful female Olympic campaigns.

References

Griffin, R.H. (2018, June 15). 120 years of Olympic history: athletes and results. https://www.kaggle.com/heesoo37/120-years-of-olympic-history-athletes-and-results

(2021, July 23). Women in the Olympic Movement. https://stillmed.olympics.com/media/Documents/Olympic-Movement/Factsheets/Women-in-the-Olympic-Movement.pdf

(2016, August 22). Here's some proof that women have always crushed the Olympics. https://pri.org/stories/2016-08-22/how-women-olympians-help-some-nations-stand-out-crowd

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2021). dplyr: A Grammar of Data Manipulation. R package version 1.0.7. https://CRAN.R-project.org/package=dplyr

H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686