Winter Olympics Analysis: Pre and Post 21st Century

Rijo Kuruvilla (14171011)

1. Abstract

a. **Introductory Statement:** The Winter Olympic Games are a major international multi-sport event held every four years for snow and ice sports. The first Winter Olympics was held in 1924 in Chamonix, France [1]. Although not as well-known as the Summer Olympics, the Winter Olympics has evolved over time with the addition of new sports and a greater number of participating nations, making it one of the most popular sporting events after the FIFA Football World Cup and the Summer Olympics.

b. **Purpose of the report:** The objective of this report is as follows:
- To analyse country performance pre and post 21st century.
- To analyse the overall athlete participation (Male and Female) in the winter Olympics.
- To determine the sport popularity by athlete participation in the 20th and the 21st century.

c. **Methodological approach**: For data processing, exploratory data analysis, and visualisation, R programming and its libraries such as dplyr, ggplot2, tidyverse, and patchwork were used. The original Olympics data was filtered by the winter season after which data tidying and data wrangling procedures were carried out to fulfill the purposes of the report as mentioned in section 1.b.

d. **Findings:**
- In the twentieth and twenty-first centuries, the Soviet Union (URS) and Canada (CAN) were the best performing countries.
- The number of female athletes has increased significantly since the commencement of the winter Olympics.
- Ice Hockey had the most participants both before and after the twenty-first century, while Skeleton had the fewest.

e. **Conclusion:** According to the findings, the popularity of the winter Olympics has grown over time. The number of countries and athletes participating in various sports increases with each season. This report can also be used to analyse individual athlete performance by sport/event, as well as individual national performance.

2. Introduction

The Winter Olympics have grown in popularity since their inception in 1924. After starting with only five sports, several new sports have been added to the Olympic programme since 1992, broadening the appeal of the Winter Olympics beyond Europe and North America [1]. From 1924 to 1936, the games were held every four years, but were interrupted by World War II in 1940 and 1944 and were later resumed in 1948. Initially, since 1992 the summer and the winter Olympics were held in the same year. But in 1986, The International Olympic Committee (IOC) ruled that Summer and Winter Olympics will be held on separate four-year cycles in alternating even-numbered years [1].
As an avid sports fan, I had been following the Summer Olympics since I was a child, but I had little knowledge of the Winter Olympics. This inspired me to investigate and analyse the Winter Olympics, the participation and performances of various athletes and countries in each season, and the various sports played during each event.

3. Data

The original dataset ‘athlete-events.csv’ is a comma-separated values(CSV) file, which consisted of 271116 observations and 15 variables, was originally scraped by Randi H. Griffin in May 2018 [6], from the sports-reference.com website which is now closed for the Olympics data [2]. This data has now been uploaded to Kaggle and is available as a public data [7]. The information was obtained as part of an observational study that depicts various athletes and sport/event scenarios held during the Olympics between 1924 and 2014. Because this study only looks at the Winter Olympics, the original data was filtered by the winter season, and after applying the processing and wrangling techniques, the final sample size of the data contains 5695 observations and 13 variables.

Variable Name	Data Type	Description
ID	integer	Unique number for each athlete
Name	character	Name of each Athlete
Sex	character	Gender of the athlete (Male, Female)
Age	Integer	Age of the athletes
Height	integer	Height of the athlete in centimetres
Weight	number	Weight of the athlete in kilograms
Team	character	Country name
NOC	character	National Olympic Committee – 3 letter code
Year	integer	Year when the Olympics was held.
Season	character	Summer or Winter
Sport	character	Type of Sport.
Event	character	Type of Event under a certain sport.
Medal	character	Type of Medal – Gold, Silver, Bronze, NA

FIGURE 1: Data Summary table

4. Methods

The analysis was carried out with R programming in RStudio studio version '2022.2.3.492.' (release name – Prairie Trillium).

Data representation: The original data is CSV (comma-separated values) file which is loaded into the Rstudio using **read.csv()** command. **View()** command was used to check if the data has been loaded correctly. The **class()** function was used to find the class of the data. The data is a tabular data and is represented in R programming as a data frame. Basic exploratory data analysis (EDA) was conducted using the **str()** and **summary()** commands which displayed the structure of the data and the summary of the variables in the data.

Unstructured to Structured data: Transformation from unstructured to structured data was not required as the imported data was already in a structured format with all the relevant column names matching the row values.

Data subset selection and/or subsampling: Since only the Winter Olympics are being analysed, a subset of the original data was taken and filtered by the winter season using the **filter()** command. The filter command was also used to split the data between the 20th and the 21st century.

Data cleaning and Missing value imputation: The command **colSums(is.na()**) was used to find the missing values (NA) in each column of the data set. One of the dataset's limitations was the large number of missing values. After filtering the data by winter season, the 'Medal' column was missing nearly 88.27% of the data, so it was removed and only the rows containing medal values (Gold, Silver, Bronze) were considered for analysis. Univariate imputation was used to fill in the missing values in the other columns (Age, Height, and Weight). The missing values were imputed using the mean values of the non-missing values from the same column. When dealing with a symmetric distribution, it is best to use the mean as a measure of central tendency [3]. The variables were tested for normal distribution and were found to be approximately normally distributed, before proceeding with missing value imputation using mean values. To test the variables for normality, the **summary()** function and a **histogram** created with the ggplot2 library were used.

Type conversion: To convert the columns 'Sex' and 'Medal' from characters to factors, the **factor()** function was used. The 'Sex' column was converted into a nominal factor with levels 'M' and 'F,' while the 'Medal' column was transformed into an ordinal factor with levels 'Gold,' 'Silver,' and 'Bronze.'

Group-based data summarisation and Variable Selection/ Transformation: After cleaning the data, imputing missing values, and type conversion, the data was arranged in ascending order by the 'Year' column using the **dplyr** package's **arrange()** function. The **pipe operator (%>%)**, in conjunction with the **summarize()** and **filter()** functions, was used to group, summarise, and select variables at the same time, fulfilling the purpose of the report and the findings mentioned in sections 1.b and 1.c. The pipe operator in R aids in the replacement of a sequence of functions by making coding simpler and more readable. The dplyr package includes all the aforementioned commands.

Exploratory visualisation using ggplot2: The commands **geom_bar(),geom_point()**, and **geom_treemap()** were used to generate bar charts, scatter plots, and tree maps. The bar chart was used to examine the countries with the most medals (performance), the scatter plot to determine the relationship between the number of athletes and the year, and the tree map to determine the popularity of the sport based on athlete participation.

5. Results and Discussion

Overall Country Performance: Figures 2 and 3 show the top five performing countries in the twentieth and twenty-first centuries. With 440 and 319 medals, respectively, the Soviet Union (URS) and Canada (CAN) are the best performing countries in the twentieth and twenty-first centuries. Austria (AUT) is the country with the fewest medals, with 177.

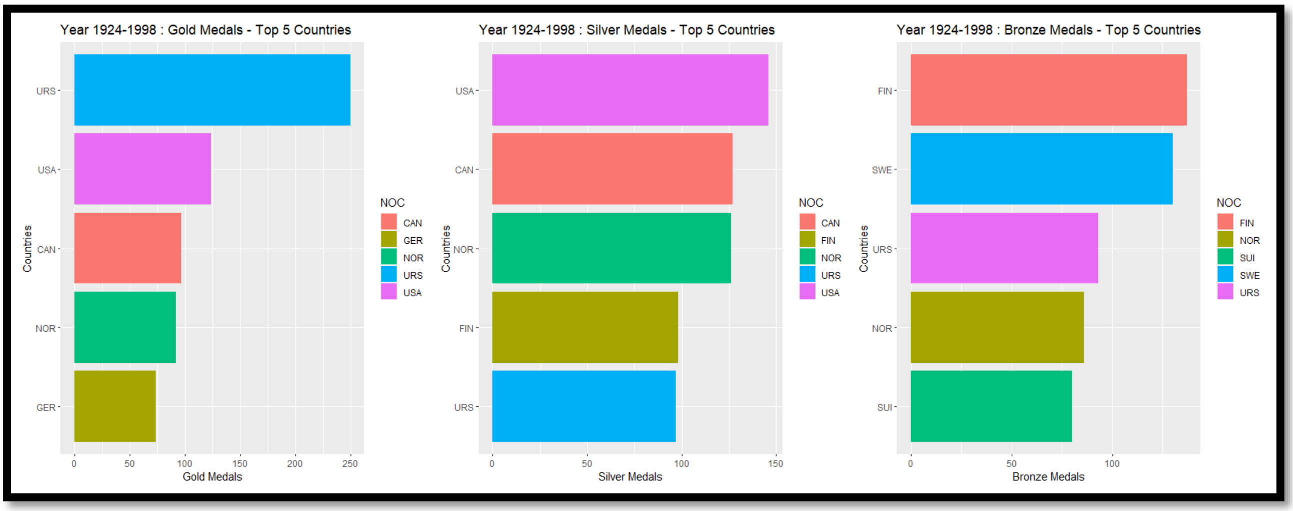

FIGURE 2: Top performing countries 20th century

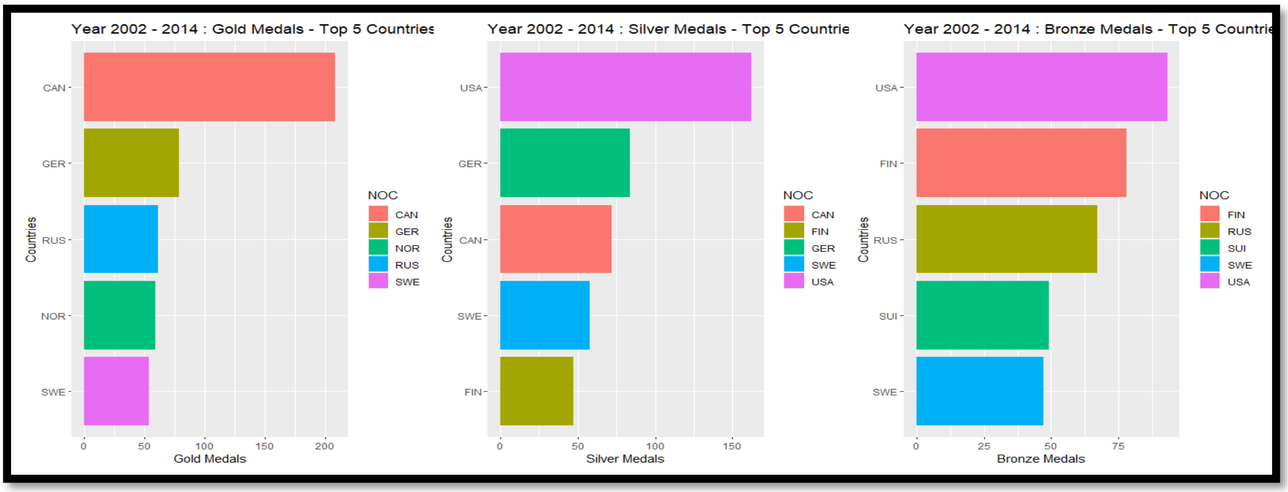

FIGURE 3: Top performing countries 21st century

Athlete participation over the years: The scatter plot (FIGURE 4) depicts the relationship between the number of male and female athletes from 1924 to 2014. The number of participants in both genders increases over the course of each winter Olympic games. The number of male athletes has always been higher than the number of female athletes. It is interesting to note that only 27 female athletes competed in the first four winter Olympic games (1924-1936), but this gradually changed, with the number of female athletes almost matching the number of male participants by 2010.
In recent years, the IOC's progressive initiative to make the Olympic Games the largest, gender-equal sporting event in the world has also contributed to an increase in the number of female athlete participants [4].

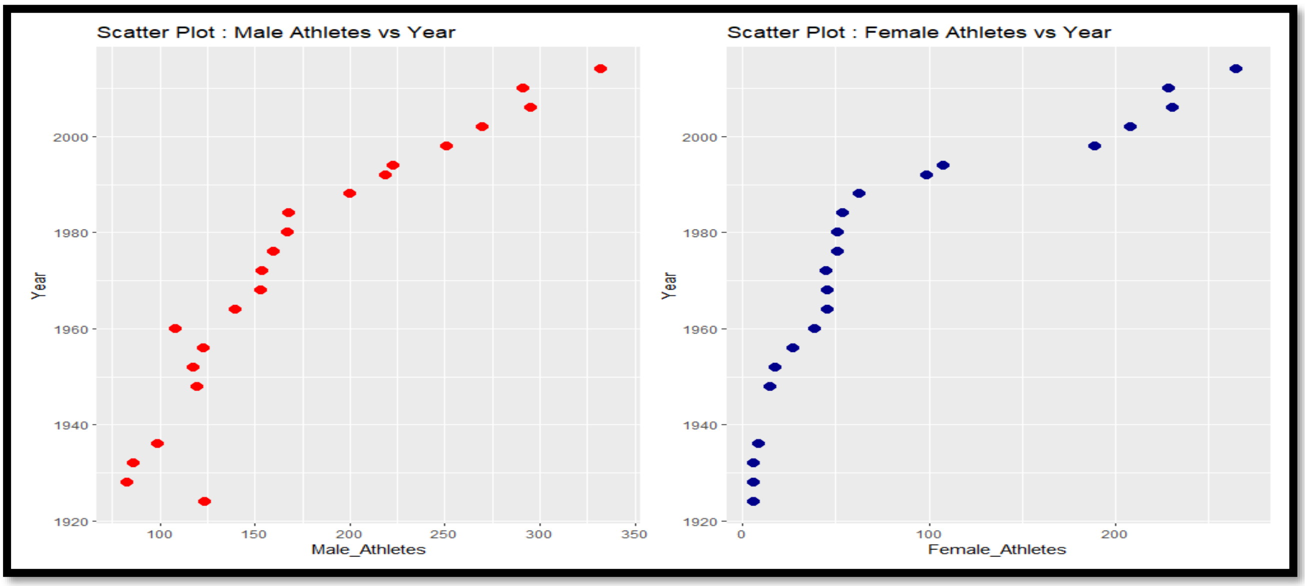

FIGURE 4: Athlete participation vs Year

Sport Popularity by Athlete Participation

Tree Maps were used to illustrate the popularity of a sport based on athlete participation as shown in FIGURE 5. Ice hockey (1235) and cross-country skiing (371) have consistently been the most popular sports among athletes in both the twentieth and twenty-first centuries. Because ice hockey is a team sport, the number of athletes is higher than in other sports. Skeleton remains the least popular sport, with a total of 26 participants in both the twentieth and twenty-first centuries, though the overall number of participants in Skeleton in each century has increased.

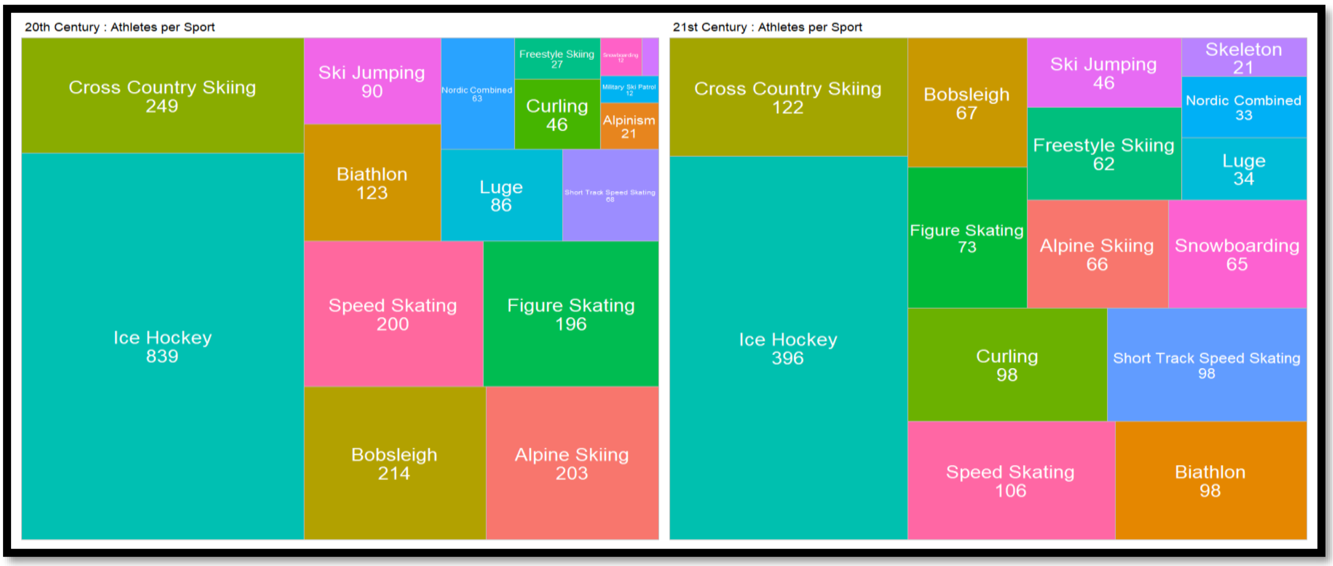

FIGURE 5: Athlete participation per Sport

6. Conclusions

Based on the increase in participation of countries with each participating year and the increase in participation of female athletes, the analysis concludes that the winter Olympics continue to grow in popularity. Although the Soviet Union and Canada had the most gold medals, the United States and Finland came in second in the total medal count in the twentieth and twenty-first centuries, respectively. Alpinism and Military Ski Patrol were both medalled events in the winter Olympics but were dropped after the 1948 Sankt Moritz games.

The structured and cleaned data can be used to build a foundation for statistical analysis. Linear Regression analysis can be used to determine whether there is a link between an athlete's age and his or her ability to win a medal.

Linear Regression can also be used to determine whether an athlete's physical characteristics (height and weight) over the course of each competing year influence the performance of athletes to win a medal. A limitation of this data would be the huge number of Null values which should be handled strategically before proceeding with the analysis.

References

[1] Wikipedia Contributors. (2019, November 4). Winter Olympic Games. Retrieved from Wikipedia website: https://en.wikipedia.org/wiki/Winter_Olympic_Games

[2] *Olympics Site Closed | Olympics at Sports-Reference.com*. (n.d.). Www.sports-Reference.com. https://www.sports-reference.com/olympics.html

[3] Frost, J. (2019, February 11). Measures of Central Tendency: Mean, Median, and Mode. Retrieved from Statistics By Jim website: https://statisticsbyjim.com/basics/measures-central-tendency-mean-median-mode/

[4] Gender equality through time: at the Olympic Games. (2021, April 29). Retrieved from International Olympic Committee website: https://olympics.com/ioc/gender-equality/gender-equality-through-time/at-the-olympic-games

[5] Treemaps in ggplot2 with treemapify. (2021, April 13). Retrieved from R CHARTS | A collection of charts and graphs made with the R programming language website: https://r-charts.com/part-whole/treemapify/

[6] (2018). Randigriffin.com. https://www.randigriffin.com/2018/05/27/olympic-history-1-web-scraping.html

[7] 120 years of Olympic history: athletes and results. (n.d.). Kaggle.com. https://www.kaggle.com/heesoo37/120-years-of-olympic-history-athletes-and-results

7. Appendix

```
# Load the required libraries

#install.packages("tidyverse")
#install.packages("dplyr")
#install.packages("ggplot2")
#install.packages("Hmisc")
#install.packages("tm")
#install.packages("treemapify")
#install.packages("patchwork")
library(dplyr)
library(ggplot2)
library(cluster)
library(Hmisc)
library(ggplot2)
library(tidyverse)
library(treemapify)
library(patchwork)

## WINTER OLYMPICS

# Load the dataolympics_data = read.csv('athlete_events.csv')
View((olympics_data))

# Filter by Winter season
winter_subset = filter(olympics_data, Season == 'Winter')
View(winter_subset)   # View the data
str(winter_subset)    # structure of the data
class(winter_subset)  # class of the data
is.na(winter_subset)  # Check for missing values
colSums(is.na(winter_subset))  # Count the number of missing values per column
which(colSums(is.na(winter_subset))>0) # Identify the position of the columns with atleast one missing value
names(which(colSums(is.na(olympics_data))>0))  # Return the column names with missing values
```

# LIMITATIONS - Winter Subset:

The original dataset consists of 48564 obs and 15 variables.

The dataset was filtered considering the question in hand. The Winter subset had the following number missing data

```
"Age"  "Height" "Weight" "Medal"

 285   8314      9021    42869
```

# Since 88.27% (42869)of the values are missing in the Medal columns, we filter the data further by Medals

```
# Filtering the Winter subset with Medal

winter_subset_2 = filter(selected_winter_variables, (Season == 'Winter' & Medal %in% c('Gold','Silver','Bronze')))
str(winter_subset_2)

# Count the number of missing values per column
colSums(is.na(winter_subset_2))
summary(winter_subset_2)
```

# IMPUTING MISSING VALUES USING UNIVARIATE IMPUTATION

```
# Before Imputation

# Check if the variables are normally distributed or not, before proceeding with mean or median

summary(winter_subset_2$Age)
summary(winter_subset_2$Height)
summary(winter_subset_2$Weight)

# GGPLOT RESULTS - Before Imputation

ggplot(data = winter_subset_2, mapping = aes(Age)) + geom_histogram(bins = 40, colour = "black") + ggtitle('Age - Before Imputation')
ggplot(data = winter_subset_2, mapping = aes(Height)) + geom_histogram(bins = 40, colour = "black") + ggtitle('Height - Before Imputation')
ggplot(data = winter_subset_2, mapping = aes(Weight)) + geom_histogram(bins = 40, colour = "black")+ ggtitle('Weight - Before Imputation')

# Apply Univariate Imputation on AGE, HEIGHT, and WEIGHT

winter_subset_2$Age = impute(winter_subset_2$Age , fun = mean)
winter_subset_2$Height = impute(winter_subset_2$Height , fun = mean)
winter_subset_2$Weight = impute(winter_subset_2$Weight , fun = mean)

### After Imputation results

summary(winter_subset_2$Age)
summary(winter_subset_2$Height)
summary(winter_subset_2$Weight)

# GGPLOT RESULTS - After Imputation

ggplot(data = winter_subset_2, mapping = aes(Age)) + geom_histogram(bins = 40, colour = "green") + ggtitle('Age - After Imputation')
ggplot(data = winter_subset_2, mapping = aes(Height)) + geom_histogram(bins = 40, colour = "red") + ggtitle('Height - After Imputation')
ggplot(data = winter_subset_2, mapping = aes(Weight)) + geom_histogram(bins = 40, colour = "blue") + ggtitle('Weight - After Imputation')

# The data is now cleaned and has no missing values.
```

# Variable Type Conversion

```
# 1. Medal: Character -> Oridnal Factor ( Gold, Silver, Bronze)

winter_subset_2$Medal = factor(winter_subset_2$Medal, levels = c('Bronze','Silver','Gold'), ordered = TRUE)
levels(winter_subset_2$Medal)

# 2. Sex : Character -> Nominal Factor

winter_subset_2$Sex = factor(winter_subset_2$Sex , levels = c("M","F"))
levels(winter_subset_2$Sex)

winter_subset_2 = arrange(winter_subset_2, Year)  # sort the data by Year
```

# COUNTRY PERFORMANCE ANALYSIS BETWEEN 20TH CENTURY AND 21st CENTURY

```
summarize = dplyr::summarize  # Initialize the summarize function from the dplyr package to be used for summarizing the data.

# Total Medal Count
```

```
tapply(winter_subset_2$Medal, winter_subset_2$Medal , length)
```

# 20th Century results

```
twentieth_century = winter_subset_2 %>% filter(Year > 1900 & Year < 2000)  # Filter the winter season by the 20th century

# Distinct medals in the 20th century by NOC(Country)

gold_20th_century = twentieth_century[twentieth_century$Medal == 'Gold' , c(1:13)] # Gold
silver_20th_century = twentieth_century[twentieth_century$Medal == 'Silver' , c(1:13)]  # Silver
bronze_20th_century = twentieth_century[twentieth_century$Medal == 'Bronze' , c(1:13)] # Bronze

## Gold Medal by countries

# 1.Medal Count by Countries - 20th century

gold_medal_by_countries_20th = gold_20th_century %>% group_by(NOC) %>% summarize(Country = n()) %>% arrange(desc(Country))
gold_medal_by_countries_20th

silver_medal_by_countries_20th = silver_20th_century %>% group_by(NOC) %>% summarize(Country = n()) %>% arrange(desc(Country))
silver_medal_by_countries_20th

bronze_medal_by_countries_20th = bronze_20th_century %>% group_by(NOC) %>% summarize(Country = n()) %>% arrange(desc(Country))
bronze_medal_by_countries_20th
```

## Bar Plot for Gold , Silver, and Bronze Medals – 20th century

```
fig_1 = ggplot(data = head(gold_medal_by_countries_20th,5)) + geom_bar(mapping = aes(x = Country, y = reorder(NOC,Country), fill= NOC), stat = 'identity') + ggtitle("Year 1924-1998 : Gold Medals - Top 5 Countries") + labs( x = 'Gold Medals' ,y = 'Countries')     # Gold

fig_2 = ggplot(data = head(silver_medal_by_countries_20th,5)) + geom_bar(mapping = aes(x = Country, y = reorder(NOC,Country), fill= NOC), stat = 'identity') + ggtitle("Year 1924-1998 : Silver Medals - Top 5 Countries") + labs( x = 'Silver Medals' ,y = 'Countries')  # Silver

fig_3 = ggplot(data = head(bronze_medal_by_countries_20th,5)) + geom_bar(mapping = aes(x = Country, y = reorder(NOC,Country), fill= NOC), stat = 'identity') + ggtitle("Year 1924-1998 : Bronze Medals - Top 5 Countries") + labs( x = 'Bronze Medals' ,y = 'Countries')  # Bronze

fig_1 | fig_2 | fig_3   # Combine the plots using patchwork()
```

# 21st Century results

```
twenty_first_century  = winter_subset_2 %>% filter(Year >= 2000)  # Filter the winter season by the 21st century

# Distinct medals in the 21st century by NOC(Country)

gold_21st_century = twenty_first_century[twenty_first_century$Medal == 'Gold' , c(1:13)] # Gold
silver_21st_century = twenty_first_century[twenty_first_century$Medal == 'Silver' , c(1:13)]  # Silver
bronze_21st_century = twenty_first_century[twenty_first_century$Medal == 'Bronze' , c(1:13)] # Bronze

# Medal Count by Countries – 21st century

gold_medal_by_countries_21st = gold_21st_century %>% group_by(NOC) %>% summarize(Country = n()) %>% arrange(desc(Country))  # Gold
```

```
silver_medal_by_countries_21st = silver_21st_century %>% group_by(NOC) %>% summarize(Country = n()) %>% arrange(desc(Country))  # Silver

bronze_medal_by_countries_21st = bronze_21st_century %>% group_by(NOC) %>% summarize(Country = n()) %>% arrange(desc(Country))  # Bronze
```

## Bar Plot for Gold, Silver , Bronze – 21st Century

```
fig_4 = ggplot(data = head(gold_medal_by_countries_21st,5)) + geom_bar(mapping = aes(x = Country, y = reorder(NOC,Country), fill= NOC), stat = 'identity') + ggtitle("Year 2002 - 2014 : Gold Medals - Top 5 Countries") + labs( x = 'Gold Medals' ,y = 'Countries')  # Gold

fig_5 = ggplot(data = head(silver_medal_by_countries_21st,5)) + geom_bar(mapping = aes(x = Country, y = reorder(NOC,Country), fill= NOC), stat = 'identity') + ggtitle("Year 2002 - 2014 : Silver Medals - Top 5 Countries") + labs( x = 'Silver Medals' ,y = 'Countries')  #Silver

fig_6 = ggplot(data = head(bronze_medal_by_countries_21st,5)) + geom_bar(mapping = aes(x = Country, y = reorder(NOC,Country), fill= NOC), stat = 'identity') + ggtitle("Year 2002 - 2014 : Bronze Medals - Top 5 Countries") + labs( x = 'Bronze Medals' ,y = 'Countries') # Bronze

fig_4 | fig_5 | fig_6   # Combine the plots using patchwork()
```

# ATHLETE ANALYSIS - OVERALL

```
# Athletes participation(M/F) each year - Grouped by Year

athlete_participation = winter_subset_2 %>% group_by(Year) %>%

  summarize('Male_Athletes' = sum(Sex == 'M'), 'Female_Athletes' = sum(Sex == 'F'))
View(athlete_participation)
```

# Scatter Plot : Year- Athlete Participation

```
#?geom_point

# MALE ATHLETES vs YEAR

SP_1 = ggplot(athlete_participation) +

geom_point(mapping = aes(x = Male_Athletes , y = Year),  size = 5 , pch = 20, color = 'red') + ggtitle("Scatter Plot : Male Athletes vs Year") + xlim(80,340)

# FEMALE ATHLETES vs YEAR

SP_2 = ggplot(athlete_participation) +

geom_point(mapping = aes(x = Female_Athletes , y = Year),   size = 5 , pch = 20, color = 'darkblue') + ggtitle("Scatter Plot : Female Athletes vs Year") + xlim(5,270)


SP_1 | SP_2    # Combining plots using patchwork()
```

# ATHLETE - SPORT PARTICIPATION ANALYSIS - 20TH and 21st Century - To determine Sport Popularity

```
athlete_sport_participation_20th = twentieth_century %>% group_by(Sport) %>%

  summarize(Number_of_Athletes = n_distinct(ID)) %>% arrange(desc(Number_of_Athletes))  # athlete sport participation - 20th century

View(athlete_sport_participation_20th)

athlete_sport_participation_21st = twenty_first_century %>% group_by(Sport) %>%
  summarize(Number_of_Athletes = n_distinct(ID)) %>% arrange(desc(Number_of_Athletes))
```

```
View(athlete_sport_participation_21st)   #athlete sport participation - 21st century
```

## Tree Map

```
# Heat Map - Athlete per Sport - 20th Century
TM_1 = ggplot(athlete_sport_participation_20th, aes(area = Number_of_Athletes, fill = Sport, label = paste(Sport,Number_of_Athletes, sep = '\n')))
+ geom_treemap() + geom_treemap_text(colour = "white",place = "centre", size = 20) + theme(legend.position = "none") + labs(title=" 20th Century : Athletes per Sport")
```

```
# Heat Map - Athlete per Sport - 2stth Century
```

```
TM_2 = ggplot(athlete_sport_participation_21st, aes(area = Number_of_Athletes, fill = Sport, label = paste(Sport,Number_of_Athletes, sep = '\n')))
+ geom_treemap() + geom_treemap_text(colour = "white", place = "centre", size = 20) + theme(legend.position = "none") + labs(title=" 21st Century : Athletes per Sport")
```

```
TM_1 | TM_2  # Combining the plots using patchwork()
```