maths

by Muskan Yadav

Submission date: 14-May-2023 11:39PM (UTC+0530)
Submission ID: 2092112681
File name: Statistical_Report_on_the_Olympic_Games.docx (521.59K)
Word count: 702
Character count: 30697

MATHEMATICS FOR ENGINEER-II

Statistical Report on the Olympic Games: Exploring the Numbers and Trends

Project Report

SUBMITTED IN PARTIAL FULFILLMENT REQUIREMENT FOR THE AWARD OF DEGREE OF

BACHELOR OF TECHNOLOGY

SUBMITTED BY

Lalit Kumar	220334
Muskan Yadav	220343
Chirag Sharma	220351
Kashish	220637

UNDER THE SUPERVISION OF

RANJIB BANERJEE

SCHOOL OF ENGINEERING AND TECHNOLOGY

BML MUNJAL UNIVERSITY™
FROM HERE TO THE WORLD

BML MUNJAL UNIVERSITY Gurugram, Haryana - 122413

May 2023

CANDIDATE'S DECLARATION

We hereby declare that the work on the project entitled, "statistical Report on the Olympic Games: Exploring the Numbers and Trends", in partial fulfilment of requirements for the award of Degree of Bachelor of Technology in School of Engineering and Technology at BML Munjal University, having University Roll No.1232434, is an authentic record of my own work carried out during a period from February 2023 to May 2023 under the supervision of SUPERVISOR NAME.

Lalit Kumar

Muskan Yadav

Chirag Sharma

Kashish

SUPERVISOR'S DECLARATION

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

Faculty Supervisor Name: Ranjib Banerjee

Signature:

ABSTRACT

Since 1896, the Olympics have handed out more than 35,000 medals. Athletes were retroactively given gold, silver, and bronze medals based on their rankings by the IOC (International Olympic Committee). The dataset we utilised comprises a row for each Olympian who has won a medal since the first games for each of the Summer Olympics (1896–2012) and the Winter Olympics (1924–2014). Additionally, this information includes 2012–2014 GDP and demographic data for each IOC nation. Four major analysis sections make up this report. The first section provides background information on the Olympics. We examine the fundamental study of the Summer and Winter Olympics in the second section. The joint study of the Summer and Winter Olympic Games is presented in the third section. We'll look at the fourth section, which compares the amount of medals won at the Summer and Winter Olympics. At the same time, in 2012 and 2014, the average high temperature in winter and GDP per capita were incorporated in order to show the relationship between the number of medals and the fundamental characteristics of each country. The amount of medals a nation wins depends on the correlation between the average high temperature in winter and GDP per capita.

ACKNOWLEDGEMENT

We would like to express our gratitude to our course faculty Dr.Ranjib Banerjee for his valuable guidance and support throughout the course. I couldn't have finished my thesis without his enlightened guidance, outstanding kindness, and patience.His attentive and active academic observation informs not only this thesis but also my upcoming research.

Statistical Report on the Olympic Games: Exploring the Numbers and Trends

INTRODUCTION

The Olympic Games are a massive global sports event that happens every four years. It brings together top athletes from around the world to compete and showcase their skills. Behind the amazing performances and memorable moments, there is a wealth of statistical information that tells us more about the Games.

In this report, we will take a closer look at the numbers behind the Olympics. We will analyze data from previous editions of the Games to discover interesting patterns, trends, and changes that have shaped this extraordinary event.

Our goal is to provide valuable insights into different aspects of the Olympics. We will explore things like the ages and nationalities of the athletes, which countries have won the most medals, the popularity of different sports, and how host nations impact the Games.

This statistical report, fortified with R programming, serves as a valuable resource for sports enthusiasts, statisticians, policymakers, athletes, and researchers. By leveraging the analytical power of R, we can gain a comprehensive understanding of the Olympic Games' broader significance. The insights derived from this report can aid in strategic decision-making, sports policy formulation, and provide a foundation for further research in the field.

Join us on this statistical journey as we harness the potential of R programming to unravel the hidden stories within the Olympic data. Together, we will explore the triumphs, trends, and transformations that have shaped the Olympic Games, gaining valuable insights into this extraordinary event through the lens of statistical analysis with R programming.

This project will focus on analyzing a dataset encompassing 120 years of Olympic history. The dataset comprises 270897 row and 14 column, providing information about the athletes who participated in both the Winter and Summer Olympic Game from 1900 to 2016.

Summary of the data:-

```
> olympic = read.csv("athlete_events.csv")
> summary(olympic)
```

```
       ID              Name               Sex                 Age              Height      
 Min.   :     1   Length:270897      Length:270897      Min.   :10.00   Min.   :127.0  
 1st Qu.: 34619   Class :character   Class :character   1st Qu.:21.00   1st Qu.:168.0  
 Median : 68156   Mode  :character   Mode  :character   Median :24.00   Median :175.0  
 Mean   : 68195                                        Mean   :25.56   Mean   :175.3  
 3rd Qu.:102031                                        3rd Qu.:28.00   3rd Qu.:183.0  
 Max.   :135474                                        Max.   :97.00   Max.   :226.0  
                                                       NA's   :9474    NA's   :60128  
     Weight            Team               NOC               Games                Year      
 Min.   : 25.0   Length:270897      Length:270897      Length:270897      Min.   :1896  
 1st Qu.: 60.0   Class :character   Class :character   Class :character   1st Qu.:1960  
 Median : 70.0   Mode  :character   Mode  :character   Mode  :character   Median :1988  
 Mean   : 70.7                                                            Mean   :1978  
 3rd Qu.: 79.0                                                            3rd Qu.:2002  
 Max.   :214.0                                                            Max.   :2016  
 NA's   :62833                                                                          
    Season              City              Sport              Event          
 Length:270897      Length:270897      Length:270897      Length:270897      
 Class :character   Class :character   Class :character   Class :character   
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   
                                                                             
                                                                             
                                                                             
    Medal          
 Length:270897      
 Class :character   
 Mode  :character   
```

Conclusion:-

This is the overview of the dataset, showcasing the minimum, maximum, median, mean, a nd quartile values for each numerical column. Additionally, it highlights the presence of mis sing values (NA) in the 'Age', 'Height', and 'Weight' columns.

Following dataset contains total 5 numerical variable

>Each athlete is given a specific id in which athlete's name , athlete's age, weight(kg), heig ht(cm), and year is given.

Other variables are character type.

- Name (Athlete's)
- Sex (M or F)
- Team (team name, mostly it's a country name)
- NOC (National Olympic Committee 3-
  letter code, more uniform way of referring to a team)
- Games (Year and season)
- Season (winter or summer)
- City (place where game was hosted)
- Sport (discipline)

- Event (subcategory – as running will be categorised by distance)
- Medal (Gold medal , Silver medal , Bronze medal, or NA)

Analysing Data:-

```
> # Checking the missing values(NA)
```

```
> count_NA <-
sapply(olympic, function(olympic) sum(length(which(is.na(olympic)))))
> count_NA
```

```
   ID        Name       Sex       Age    Height    Weight      Team       NOC     Games      Year    Season      City
Sport
    0           0         0      9474     60128     62833         0         0         0         0         0         0
    0
 Event     Medal
     0    231145
```

Upon analyzing the dataset, we found that several variables contain missing values (NA). The 'Age' variable has 9,474 missing values, indicating that the age informatio n is unavailable for some athletes. Similarly, the 'Height' variable has 60,128 missing values, suggesting that the height data is not recorded for a significant number of at hletes. Additionally, the 'Weight' variable has 62,833 missing values, indicating that t he weight information is absent for many athletes.

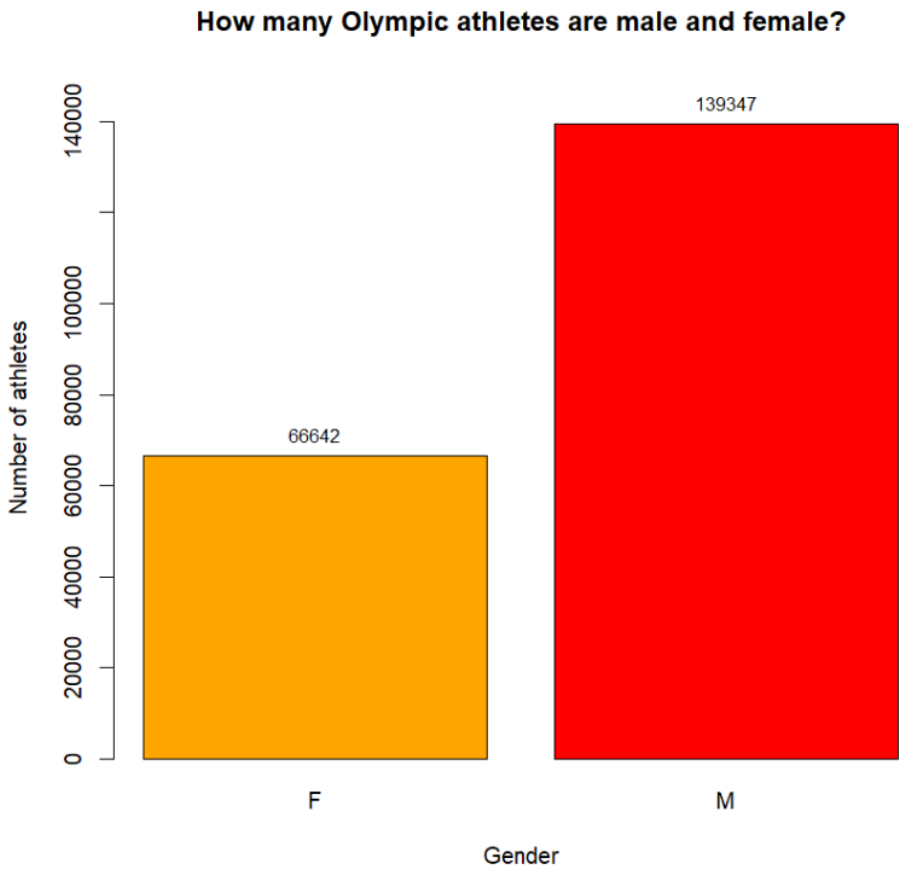
Now, let's remove the outliers or the rows which have null (NA) numerical values

```
#removing all rows with null values
df = olympic %>% drop_na(Height, Weight, Age)
```

Now, let's start with the distribution of athletes between males and females:

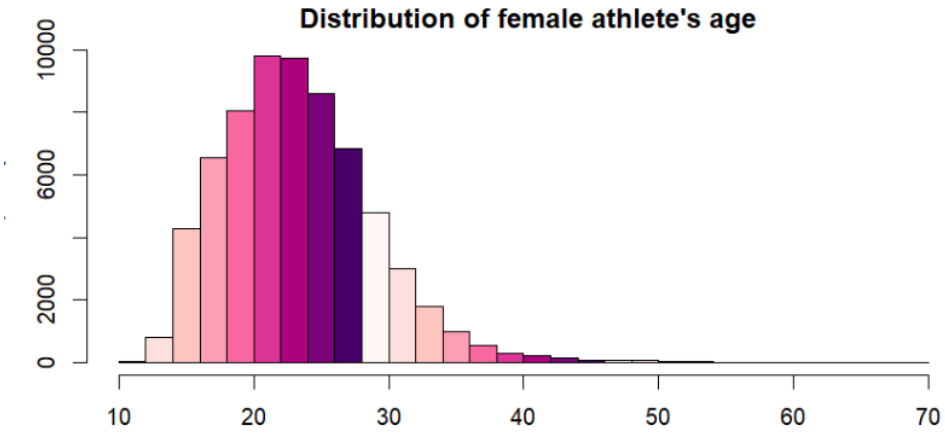
```
a = barplot(table(df1$Sex),
            main = "How many Olympic athletes are male and female?",
            xlab = "Gender",
            ylab = "Number of athletes",
            ylim = c(0,150000),
```

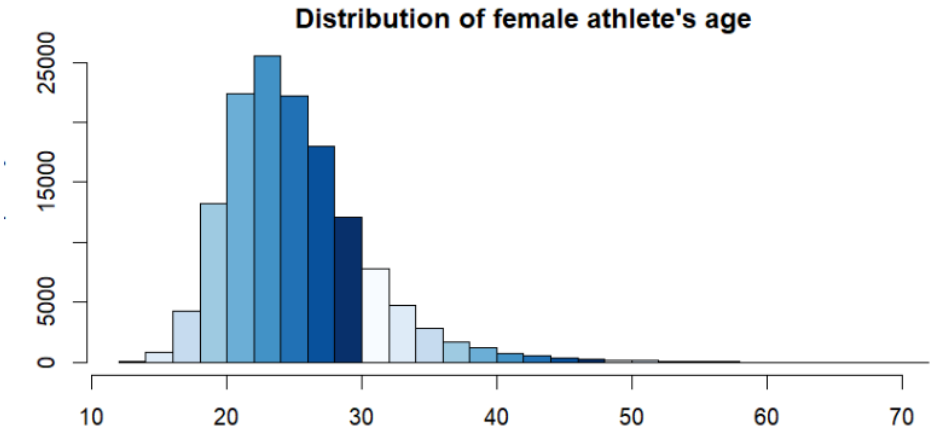
```
                col = c("orange", "red"))

text(y = table(df$Sex),
     a,
     table(df$Sex),
     cex=0.8,
     pos = 3)
```



Now that we know there are twice as many male athletes as female athletes, it's important to examine how numerical variables are distributed based on gender. To begin our analysis , we'll use diagnostic charts like histograms and box plots.

```
# creating subset to have a look separately on both the genders
women = dplyr::filter(df1, Sex == "F")
men = dplyr::filter(df1, Sex == "M")

# Let's represent it on histograms and boxplots:
par(mfrow=c(2,1))
par(mar=c(3,3,2,1)) #  margins changes

hist(women$Age,
     breaks = 30,
     col = brewer.pal(9, "RdPu"),
     main = " Distribution of female athlete's age")

hist(men$Age,
     breaks = 30,
     col = brewer.pal(9, "Blues"),
     main = "Distribution of female athlete's age")
```





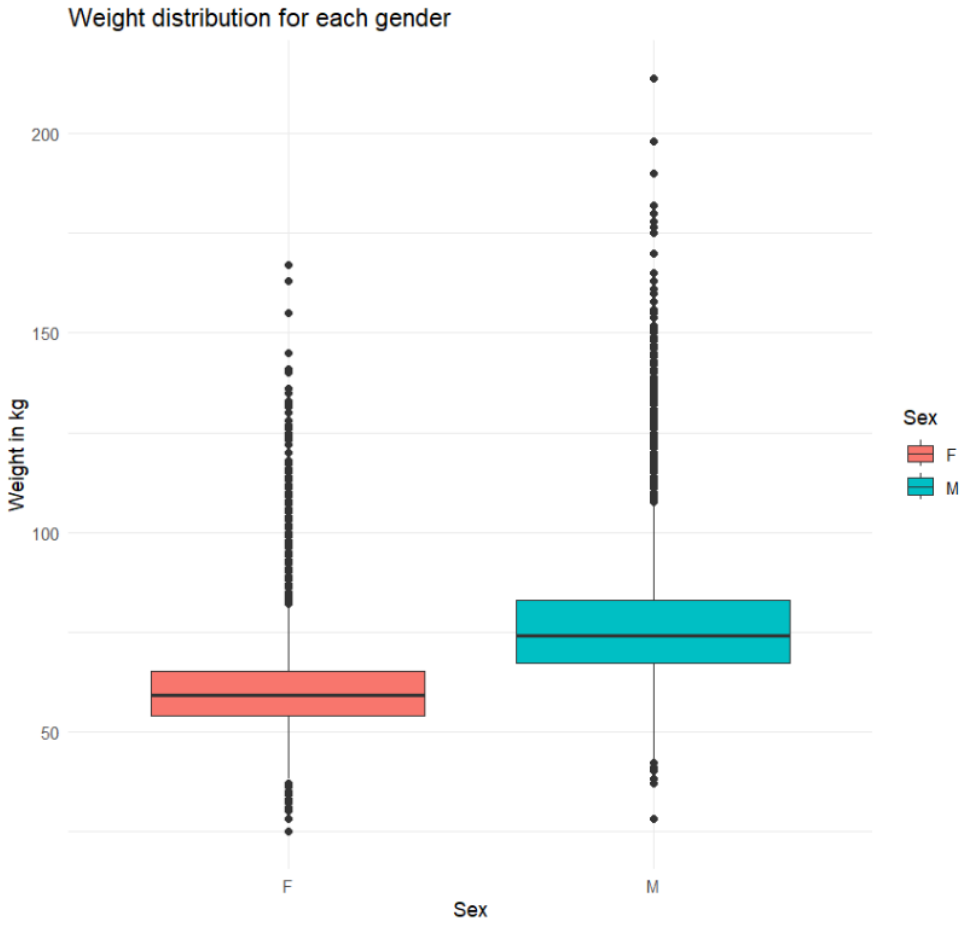
The histograms show that the distribution of ages for male and female at hletes at the Olympic Games is right-
skewed, with a higher frequency of females under the age of 20 compare d to males. Both distributions have a peak between the ages of 20 and 3 0, but the mode (most frequent age) for female athletes is lower than for male athletes.

Let’s explore the other numerical variables:

Weight

```
# Box plots for Men's and Women's weight
 ggplot(df1, aes(x = Sex, y = Weight, fill = Sex)) +
```

```
geom_boxplot() +
scale_color_manual(values=c("lightcoral", "cornflowerblue")) +
labs(title = "Weight distribution for each gender", y = "Weight(Kg)") +
theme_minimal()
```

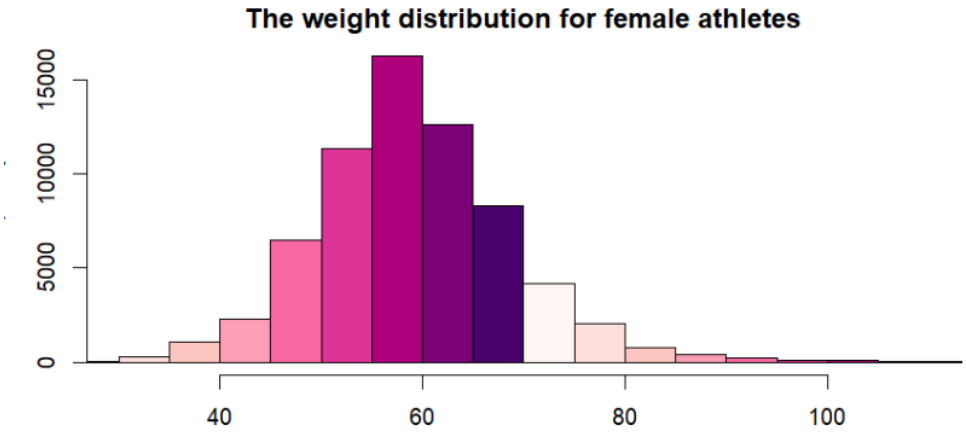


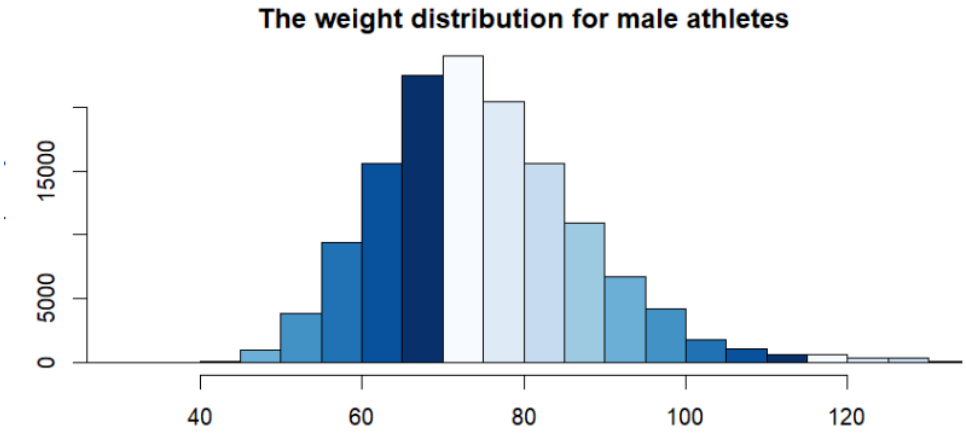
Conclusion:

- Box plots show a five-
  number summary: minimum value, first quartile, median, third quart
  ile, and maximum value.
- The figure displays the weight distribution for both genders.

- The median weight is higher for male athletes compared to females , as indicated by the thick line in the middle of the box plots.
- The "boxes" represent the interquartile range, showing the distributi on of weights.
- The distribution of weights is more dispersed for male athletes than for females.
- The dots above and below the plots represent outliers, which are d ata points outside the range of 1.5 times the interquartile range.
- Both genders have outliers that are higher and lower in weight than the rest of the data.

```
#Now let's have a look on  how it looks  on histograms:
par(mfrow=c(2,1))
par(mar=c(3,3,2,1))

hist(women$Weight,
     xlim = c(30, 110),
     breaks = 30,
     col = brewer.pal(9, "RdPu"),
     main = "The weight distribution for female athletes")

hist(men$Weight,
     xlim = c(30, 130),
     breaks = 30,
     col = brewer.pal(9, "Blues"),
     main = "The weight distribution for male athletes")
```





From the histogram, it appears that the variable representing weight follo ws an approximately normal distribution, with a slight right skew. Addition ally, the mode (most frequent value) for the weight of male athletes is hig her than for female athletes.

To determine the exact percentage of female athletes weighing 60 kilogra ms or less, we would need the specific counts or frequencies for each we ight interval in the histogram. Without this information, it is challenging to provide an exact percentage. However, by examining the histogram, we c an estimate that a significant portion of female athletes falls within the 60 kilograms or less range.

```
pnorm(60, mean = mean(women$Weight), sd = sd(women$Weight))
```

Output:-

```
[1] 0.4993863
```

Examining the outliers plotted above the box plot, we will focus on two sp ecific groups:

1. Female athletes who weigh more than the third quartile plus 1.5 tim es the interquartile range. These athletes have weights that fall out side the upper limit of the typical range for females.
2. Male athletes who weigh less than the value of the first quartile min us 1.5 times the interquartile range. These athletes have weights th at fall below the lower limit of the typical range for males.

Analyzing these outliers allows us to identify female athletes with higher-than-average weights and male athletes with lower-than-average weights. By examining these individuals, we can gain insights int o potential exceptional cases or factors that contribute to weight differenc es among athletes of both genders.

```
# First, let's find the first and third quartiles
q1 <- quantile(women$Weight, 0.25)
q3 <- quantile(women$Weight, 0.75)
iqr <- q3 - q1

# Now, calculate the upper and lower limit
lower <- q1 - 1.5 * iqr
upper <- q3 + 1.5 * iqr

# Lastly, find outliers
out_upper = women %>%
  select(c(1:7)) %>%
  filter(Weight > upper)
head(out_upper)
```

```
  ID                Name Sex Age Height Weight      Team
1  5 Christine Jacoba Aaftink   F  21    185     82 Netherlands
2  5 Christine Jacoba Aaftink   F  21    185     82 Netherlands
```

3	5	Christine Jacoba	Aaftink	F	25	185	82	Netherlands
4	5	Christine Jacoba	Aaftink	F	25	185	82	Netherlands
5	5	Christine Jacoba	Aaftink	F	27	185	82	Netherlands
6	5	Christine Jacoba	Aaftink	F	27	185	82	Netherlands

Upon closer examination, it is observed that there are a substantial number of observations (1668) classified as outliers with weights higher than the third quartile plus 1.5 times the interquartile range. However, it is important to note that this count includes duplicate entries for one athlete who participated in multiple events, resulting in her being mentioned multiple times in the dataset. To gain a clearer understanding of the actual number of male athletes weighing less than the first quartile minus 1.5 times the interquartile range, let's explore this subset.

By focusing on males who weigh less than the value of the first quartile minus 1.5 times the interquartile range, we can expect to find a smaller number of athletes in this category. This subset may help us identify male athletes who have weights significantly lower than the typical range for males.

```
# Find the first and third quartile (overwriting the previous values )
 q1 <- quantile(men$Weight, 0.25)
 q3 <- quantile(men$Weight, 0.75)
 iqr <- q3 - q1

 # Calculate  lower and upper limits
 lower <- q1 - 1.5 * iqr
 upper <- q3 + 1.5 * iqr

 #   find outliers
 men %>%
   select(c(1:6)) %>%
   filter(Weight < lower)
```

Output:-

	ID	Name	Sex	Age	Height	Weight
1	18005	Cao Yuan	M	17	160	42
2	18005	Cao Yuan	M	21	160	42
3	18005	Cao Yuan	M	21	160	42
4	24814	Barry Edward Dagger	M	39	147	41
5	24814	Barry Edward Dagger	M	47	147	41
6	38251	Wayne Bruce Gammon	M	14	155	38
7	46484	Ali Ismael Hassan	M	21	157	42
8	55906	Daniel Jorge	M	13	152	41
9	73023	Thomas P. Mack, Jr.	M	14	147	41
10	73023	Thomas P. Mack, Jr.	M	18	147	41
11	87343	Miguel Nez Lima	M	20	170	42
12	96860	Robert Archibald "Bob" Prentice	M	35	156	42
13	99902	Erik Remmerswaal	M	31	147	38
14	100743	Michel Riendeau	M	21	167	41
15	101936	Mircea Roger	M	13	146	40
16	101936	Mircea Roger	M	13	146	40
17	113413	Alaeddin Soueidan	M	13	148	37
18	115166	Orlando Maurits Stewart	M	34	153	40
19	134370	Albert Ferdinand "Al" Zerhusen	M	24	183	28

Conclusion:
19 men are outliers in the lower end distribution of weight.

In statistical hypothesis testing, one can assess whether a sample statistic deviates significantly from the expected value. In this case, we will conduct a one-tailed test t o compare the means of male and female athletes' weights. There are two common statistical tests used for hypotheses related to means: the t-test and the z-test.
For this analysis, we will begin by employing a t-test. The hypotheses are formulated as follows:

- Null Hypothesis (H0): There is no significant difference in the average weight between male and female athletes.
- Alternative Hypothesis (H1): Male athletes, on average, have a higher weight compared to female athletes.

By conducting the t-test, we aim to examine whether the sample data provides enou gh evidence to reject the null hypothesis in favor of the alternative hypothesis.

```
t.test(men$Weight, mu = mean(women$Weight), alternative = 'greater')
```

```
##
##
##
##          One Sample t-test
```

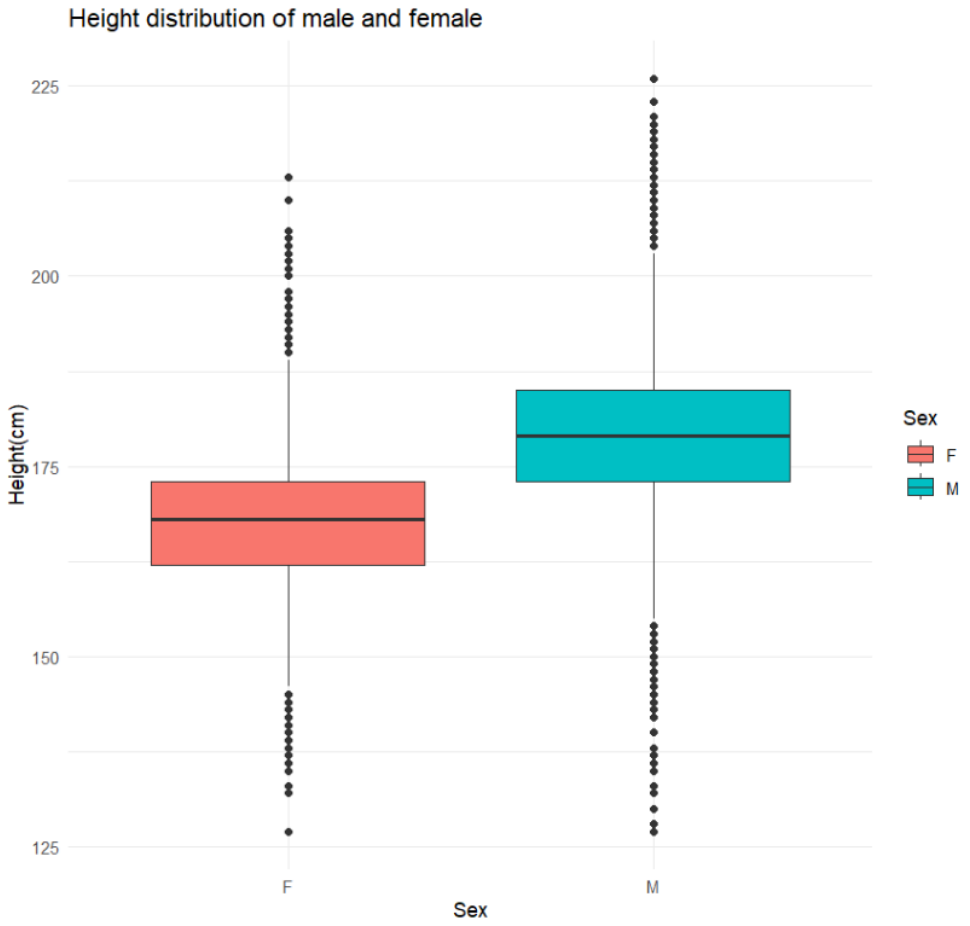
```
##   data:  men$Weight
##   t = 446.61, df = 139346, p-value < 2.2e-16
##   alternative hypothesis: true mean is greater than 60.0157
##   95 percent confidence interval:
##   75.72893      Inf
##   sample estimates:
##   mean of x
##   75.78702
```

In hypothesis testing, the p-value represents the probability of obtaining results as extreme as, or even more extreme than, the observed data, assuming that the null hypothesis is true. A small p-value indicates that the observed results are unlikely to have occurred by chance alone.

In this case, since the calculated p-value is very close to zero, it provides strong evidence against the null hypothesis. This implies that we can reject the null hypothesis and conclude that there is a significant difference in the average weight between male and female athletes. This result holds true for a chosen confidence interval of 95% and would also hold for a higher confidence level, such as 99%, as the p-value is lower than the significance level (alpha) in both cases.

Height

```
# Box plots
 ggplot(df, aes(x = Sex, y = Height, fill = Sex)) +
   geom_boxplot() +
   scale_color_manual(values=c("lightcoral", "cornflowerblue")) +
   labs(title = "Height distribution of male and female ", y = "Height(cm)") +
   theme_minimal()
```

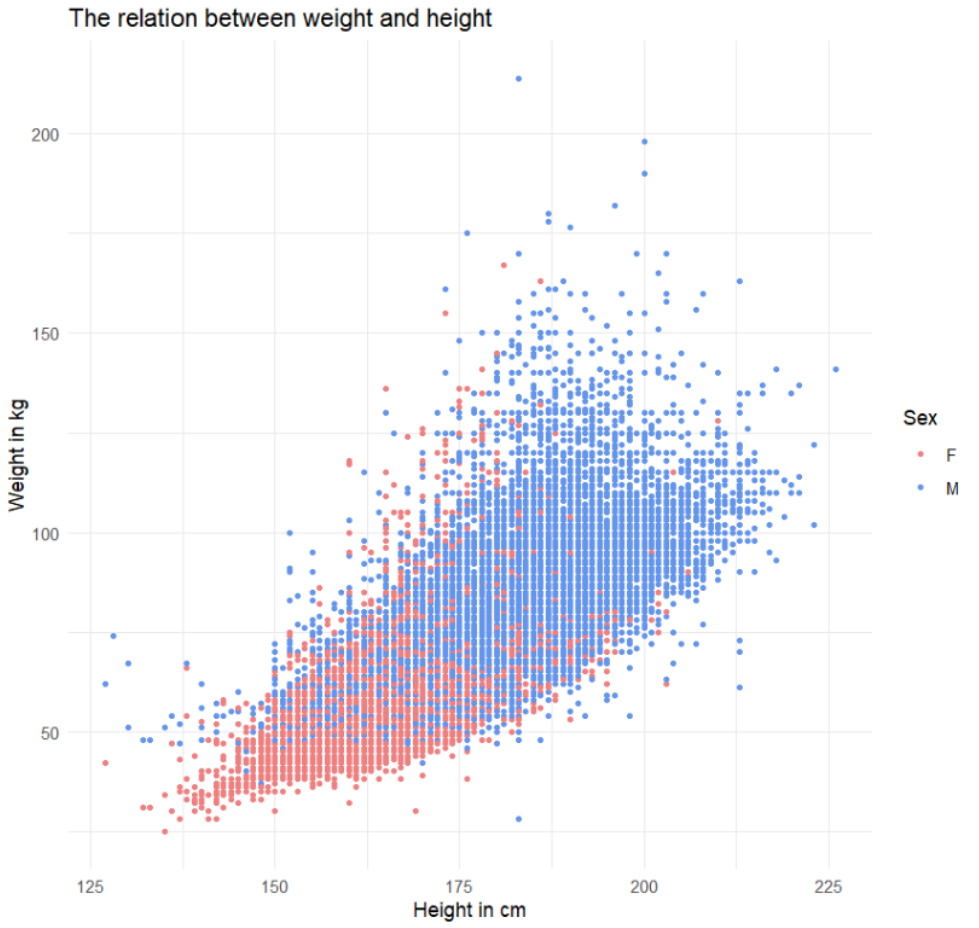


It is observed that, similar to weight, the height of male athletes tends to be higher than tha t of female athletes. While there are athletes of both genders who have heights just above 125 cm, it is worth noting that there are no female athletes taller than 215 cm.

In addition to examining the height and weight individually, exploring the relationship betwe en these two variables is indeed intriguing. To illustrate this relationship, a scatter plot will b e used. A scatter plot visually represents the data points as individual dots, with one variabl e plotted on the x-axis (in this case, height) and the other variable on the y-axis (weight). This plot will help us visualize any patterns or trends between height and wei ght among athletes.

```
ggplot(df, aes(x = Height, y = Weight, color = Sex)) +
  geom_point(size = 1) +
  scale_color_manual(values=c("lightcoral", "cornflowerblue")) +
  theme_minimal() +
```

```
  labs(title = "The relation between weight and height", y = "Weight in kg", x = "Height in c
m")
```



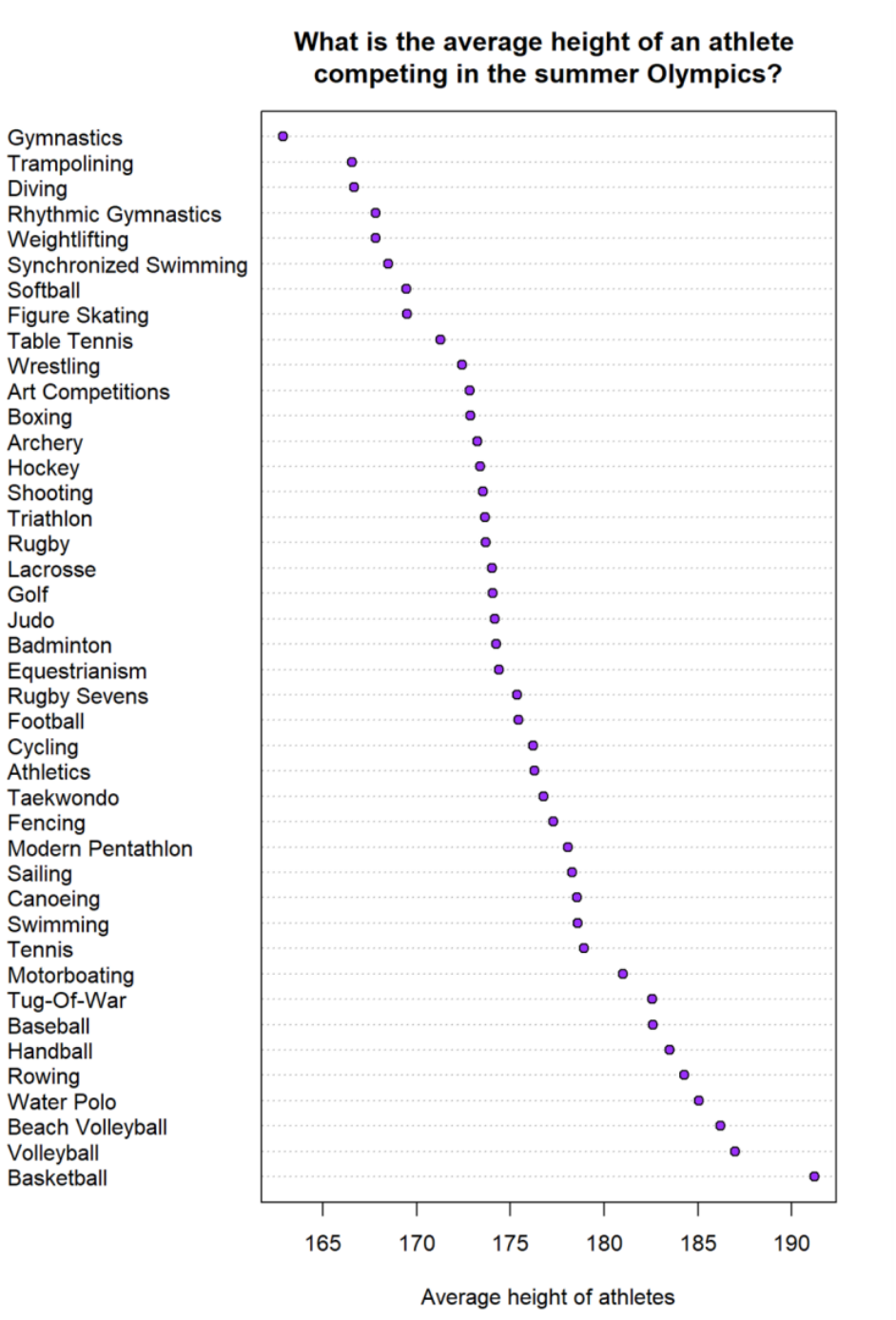
A scatter plot is a graphical representation of ordered pairs of numbers, with one variable (height in this case) as the independent variable (x-axis) and another variable (weight) as the dependent variable (y-axis). It visually depicts tnature of relationship between the two variables. In this analysis, t he scatter plot demonstrates a positive linear relationship between height and weight, indic ating that taller individuals tend to weigh more. The plot also incorporates color coding to re present gender, revealing that, on average, men are taller and weigh more, consistent with the findings from previous box plots. However, there are exceptions within the dataset due t o variations in sports disciplines. For example, gymnasts are typically shorter, weightlifters t end to have higher weights regardless of gender, and basketball players are generally taller .

To further analyze the numerical variables, subsets were created for winter and summer sp orts. However, it was noticed that ice hockey was included in the summer Olympics only on ce (in 1920), so it was removed from the summer sports subset and included solely in the winter sports subset. It's important to consider that height and weight distributions for specif ic sports disciplines may vary, and analyzing data from a single year may not fully represen t the trends in athlete height and weight over time.

```
# for clear visualizations create subsets separately
summer = dplyr::filter(df, Season == "Summer")
summer = summer[!grepl("Ice", summer$Sport),] #  in 1920 ice hockey was clasified as summer )
winter = dplyr::filter(df, Season == "Winter")
```

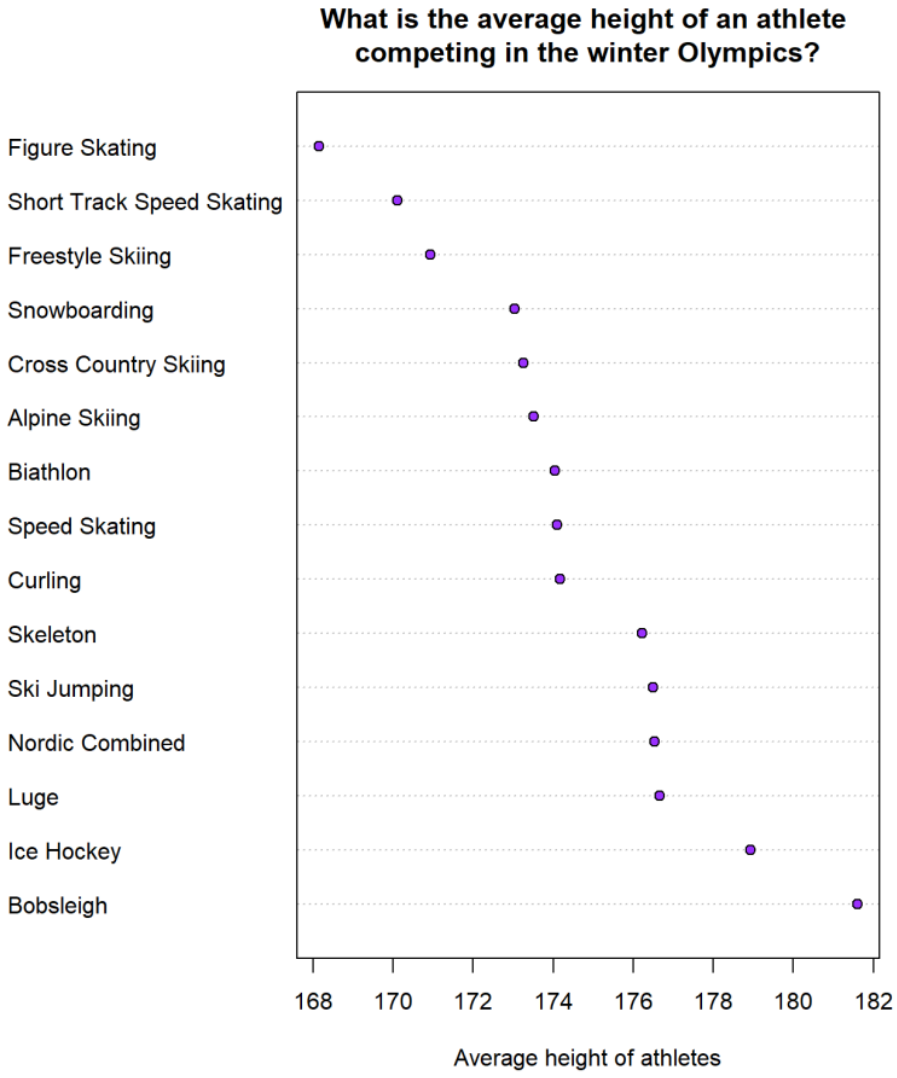
Dot plots are a type of chart where each value is represented as a dot placed above a hori zontal axis. In your analysis, you plan to utilize dot plots to visualize the average height of athletes participating in different sports disciplines. Instead of displaying individual data poin ts, you will summarize the data by calculating the average height for each sport. This appro ach allows you to focus on the average values and eliminates clutter from the chart, makin g it easier to compare the average heights across different sports. By examining the averag e height values, you can gain insights into the typical height range for athletes in each spec ific sport discipline.

```
# Let's have a look on  average height of thr  summer sports
Summe.heights = tapply(sumer$Height,  summer$Sport, mean)
Summer.heights = sort(sumer_heights, decreasing = TRUE)


dotchart(Summer.heights, pch = 21, bg = "purple1",
         xlab="Average height of athlete",
         main = "What is the average height of an athlete \ncompeting in the summer Olympics?")
```


What is the average height of an athlete competing in the summer Olympics?
Gymnastics
Trampolining
Diving
Rhythmic Gymnastics
Weightlifting
Synchronized Swimming
Softball
Figure Skating
Table Tennis
Wrestling
Art Competitions
Boxing
Archery
Hockey
Shooting
Triathlon
Rugby
Lacrosse
Golf
Judo
Badminton
Equestrianism
Rugby Sevens
Football
Cycling
Athletics
Taekwondo
Fencing
Modern Pentathlon
Sailing
Canoeing
Swimming
Tennis
Motorboating
Tug-Of-War
Baseball
Handball
Rowing
Water Polo
Beach Volleyball
Volleyball
Basketball
165
170
175
180
185
190
Average height of athletes

The dot plot clearly shows that there are significant differences in height between different sports. Gymnasts tend to be the shortest, while basketball players are typically the tallest. This visual representation highlights the varying heights required for different sports and co nfirms the common understanding that athletes in each sport have specific physical charact eristics.

```
# Let's have a look on averagee heights of winter sport
winter.heights = tapply(winter$Height, winter$Sport, mean)
winter.heights = sort(winter_heights, decreasing = TRUE)


dotchart(winter.heights, pch = 21, bg = "purple1",
         xlab="Averagee height of the athlete",
         mai n = "What is the average height of an athlete \ncompeting in the winter Olympics?")
```



The distribution of heights for athletes in the winter games appears to be less spread out compared to sports in the summer Olympics. This could be attributed to the presence of fewer disciplines in the winter games. Additionally, certain winter sports may require specific physical attributes from athletes, leading to a narrower range of heights among participants.

One surprising observation is that bobsleigh, on average, has the highest placement in terms of height among the winter sports. This finding challenges initial expectations and suggests that bobsleigh athletes tend to have relatively taller heights compared to athletes in other winter sports. It highlights the diverse physical requirements and characteristics associated with different sports, even within the context of the winter Olympics.

Conclusion

In analyzing the data from 120 years of Olympic history, several interesting insights have b
een revealed. Firstly, it became apparent that the distribution of weight and height is not sol
ely determined by gender. Instead, these attributes are heavily influenced by the specific sp
orts discipline. For instance, a female volleyball player is likely to be taller than a male gym
nast, highlighting the importance of considering the sport when examining numerical attribut
es.

Additionally, a significant gender disparity was observed in Olympic participation, with only
half of the athletes in the dataset being female. This aligns with the information provided by
the International Olympic Committee (IOC), which states that even in recent Olympics, suc
h as Tokyo, female athletes constituted only 49% of participants. Notably, Tokyo had the hi
ghest level of gender equality in Olympic history. Conducting a more comprehensive analys
is of the dataset would be valuable, particularly by examining trends over time using line ch
arts, such as the proportion of female athletes.

Furthermore, hypothesis testing was conducted to validate assumptions regarding the weig
ht and height of Olympic athletes. The results confirmed that there is no significant differen
ce in average weight between athletes in winter and summer sports. Additionally, it was est
ablished that male athletes generally weigh more than their female counterparts, and baske
tball players tend to be significantly taller than gymnasts.

By uncovering these findings and utilizing visualizations, a deeper understanding of the attri
butes and patterns within Olympic athletes' data has been obtained.

Link for the code:-

https://docs.google.com/document/d/1mZ95bZPgWLk0uXcbXgZ5xSkCyZWS3CGXJ3WTXHb7KV0/edit?usp=sharing

maths

ORIGINALITY REPORT

9%	6%	1%	5%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.coursehero.com Internet Source	4%
2	Submitted to University of Warwick Student Paper	2%
3	Submitted to Aligarh Muslim University, Aligarh Student Paper	1%
4	Sandra M. Mathioni, André Beló, Jeffrey P. Townsend, Nicole M. Donofrio. "Chapter 5 Getting the Most Out of Your Fungal Microarray Data: Two Cost- and Time-Effective Methods", Springer Science and Business Media LLC, 2011 Publication	1%
5	how2electronics.com Internet Source	1%
6	grietinfo.in Internet Source	1%

Exclude quotes	On	Exclude matches	< 6 words
Exclude bibliography	On		