MGSC 661 – Final Project

Rodrigo Castro

Clustering-Based Trend Analysis of Olympic Games

December 2023

1. Introduction

This project delves into the evolving landscape of the Olympic Games, employing data analytics to show how various facets of the games have transformed over time. The analysis extends beyond athlete profiles to encompass a broader spectrum of Olympic characteristics. Tools such as Principal Component Analysis (PCA) and K-Means clustering were utilized to dissect trends, encompassing shifts in sporting disciplines to the intricacies of athlete performance. The insights derived are not merely academic, they could hold substantial value for professionals across the sports industry. From event organizers and sports marketers to talent scouts and coaches, this analysis serves as a strategic asset, offering the foresight needed to navigate and excel in the dynamic and competitive world of international sports.

2. Data Description

The Olympics' dataset used encapsulates all the games spanning from Athens 1896 to Rio 2016. It profiles 98,546 athletes across 56 sports and 656 teams, representing the competitive spirit of 20 countries. Information about the athletes' characteristics and medals won in each game is also displayed. This extensive collection reflects the universal appeal and rich diversity of the Olympic movement.

Based on observations of athletes over recent years, the majority fall within the age range of 20 to 30, with some senior athletes aged up to 60 years. Height and weight data were used to calculate the Body Mass Index (BMI) for each player. The most common BMI range was found to be between 20 and 25. This indicates that the majority of participants in the game generally maintain an appropriate balance between their height and weight. For additional details, please refer to **Appendices I and II**.

Regarding the gender distribution, historically, only 32% of the players have been female, though there has been a notable increase in recent years. This trend may be attributed to the restrictions placed on female athletes in the early 20th century. Another significant observation is that among all athletes, only 15% have won a medal, with 34% of these medalists winning gold. This underscores the highly competitive nature of the Olympics. The performance score for each player, which considers the number of medals won, the type of medal, and the number of sports competed in, also reflects this competitive landscape. The distribution of performance scores is skewed towards players who have not won any medals. For additional information, please refer to **Appendix III.**

United States, Russia, and Germany were identified as the highest-performing teams (refer to **Appendices IV and V)**. However, this ranking correlates with the historical number of athletes each country has sent. For instance, athletes from the US represent 6.4% of the total, followed by Germany (6.1%), and Russia (4.5%). A larger number of athletes increases the likelihood of having individuals who excel in certain sports. Nevertheless, Australia and China, both in the top 5 for performance, have sent fewer athletes compared to other countries. This suggests that, in terms of individual performance, athletes from these countries tend to outperform their counterparts from the United States or Europe.

On the other hand, as indicated in **Appendix VI**, many variables in the dataset show no correlation, with a few exceptions. For instance, the season of the games is related to the type of sports played. This is because many winter sports, as well as water sports typically held in summer, are predominantly individual. Additionally, there is a correlation between an athlete's gender and their BMI, reflecting differing height and weight ratios between males and females. The data also reveals that there are fewer senior female athletes compared to male athletes, suggesting that female players face greater constraints in continuing competitive sports as they age. Lastly, there is a slight difference in performance between team and individual sports, with individual athletes generally showing better overall performance.

3. Model Selection & Methodology

Data Pre-processing:

The objective of this project is to cluster the dataset to uncover patterns among players and the Olympic Games over the years. To achieve this, new variables were introduced. For instance, as previously mentioned, the Body Mass Index (BMI) was calculated for each player, given the high correlation between height and weight. This index helps categorize an individual as underweight, overweight, or within a healthy weight range. The BMI is calculated using the following formula:

$$BMI = Weight\ (kg) / Height\ (meters)^2$$

Additionally, a performance score was calculated to quantitatively assess each athlete's achievements across different disciplines and competitions. This involved assigning a weight to each type of medal, with gold medals having a higher weight, and considering the number of events in which each athlete participated. The performance score was computed using the following formula:

$$Performance\_score = Sum\ of\ Medal\ Weights / Total\ Number\ of\ Events\ Participated$$

Continuing with the data preprocessing, all null values were removed. To facilitate the identification of patterns within the dataset, each game was categorized into specific subcategories based on both the year of the event and the type of sport.

- Periods: Early 20th Century, World War (WWI & WWII), Cold War, Actual Period.
- Sports: Winter Sports, Team Sports, Individual Sports, Water Sports, Others.

Finally, the variables of Age, BMI, Performance Score, and the dummified variables of the sports Subcategories were selected for the clustering.

Data Modeling:

Due to the numerous variables in the dataset, Principal Component Analysis (PCA) was employed to reduce the dimensionality of the input data, aiming to create an interpretable model. The historical data of the Olympic Games was divided into the previously mentioned time periods, allowing for clustering analysis to be conducted separately for each period. Additionally, K-means clustering was applied to complement the time series analysis within The Actual Period. To determine the optimal number of clusters, the Elbow Method was used. This method assesses the internal variance within each cluster, helping to identify the point where adding more clusters doesn't significantly improve the model's performance. The optimal number of clusters was identified as 3, as detailed in **Appendix VI**I.

4. Results

Below is a summary of the observations from each period, based on the visualizations of the Principal Component Analysis (PCA). These observations can be found detailed in **Appendices VIII to XI**. This summary aims to highlight key insights and trends identified through PCA for each distinct period.

Feature	Early 20th Century	World War	Cold War	Actual Period
Age	Avg. of **24** years.	Slight increase to **25** years.	Slight decrease to **24** years.	Slight increase to **25.6** years.
BMI	Avg. of **23.3.** Suggesting a relatively lean physique.	Slightly lower avg. of **22.9**. Less healthy physique, related to the conditions of the period.	Maintained around **22.9**. Not notable changes post-war.	Maintained around **22.9**. Probably due to the increase of senior athletes.

Performance	The avg. score was **0.265.**	Increased to **0.373**, showing better performance by perhaps less competition.	Significant to **0.145**, possibly indicating heightened competition.	Decrease to **0.126**, which could reflect an even more competitive environment and a larger pool of participants.
Gender	There were no female athletes during this period.	First inclusion of female athletes, **6.8%** on avg.	Noticeable increase in female participation to **23.5%.**	Significant presence of female athletes, accounting for **40%** of the participants
Season	Only summer sports as there were no winter sports included.	Mostly summer sports with a small representation of winter sports beginning to emerge.	Reduction in the predominance of summer sports, with an increased presence of winter sports.	Decline in the dominance of summer sports, with more athletes participating in winter sports.
Sports	Predominantly individual sports with very little representation in other categories. None winter sports.	Individual sports still dominated but with a noticeable increase in team and water sports. Introduction of winter sports, still by a small fraction.	Shift towards a more balanced distribution among individual, team, and water sports. A more substantial presence of winter sports at **14.5%.**	Diversified presence across individual, team, and water sports categories. Winter sports constituted a significant **23.2%** of the sports activities.

To enhance our analysis, we focused on information derived from K-means clustering, specifically applied to the 'Actual Period Sub-dataset'. This approach was taken to gain deeper insights into the current state of the Olympic Games. For more comprehensive and referential details, please refer to **Appendices XII and XIII**. This targeted analysis aims to provide a clearer understanding of recent trends and patterns in the Olympics.

Cluster	Age	BMI	Performance Score	% Female Athletes	% Summer Sports	% Winter Sports	Distribution in Sports
1	26.61	23.4	0.151	0.347	0.751	0.249	Balanced across individual, team, water
2	20.8	21.77	0.115	0.487	0.782	0.218	High in water sports

3	34.59	23.46	0.111	0.34	0.784	0.216	Notable in 'Others' category

The K-means analysis of the current Olympic landscape identifies three athlete profiles. Cluster 1 athletes are in their mid-20s, maintain a solid physique, and show a balanced involvement in both summer and winter sports, with competitive performance scores. Cluster 2 consists of younger, leaner athletes with nearly equal gender representation, a preference for summer and water sports, but lower performance scores, indicating potential room for growth. The oldest and most experienced athletes fall into Cluster 3, with a diverse participation in various sports, including non-traditional ones, and they maintain a steady level of performance. These profiles reflect the Olympics' progressive inclusivity and the broadening scope of competitions, pointing towards a future with a rich diversity of disciplines and participants.

5. Conclusions & Business Insights

The PCA and K-means analyses of Olympic data not only reflect the historical shifts in the Games but also offer a lens through which future trends and managerial strategies can be anticipated. The consistent increase in female participation and the rise of winter sports underscore a movement towards inclusivity and diversity. The stabilization of BMI and the slight rise in the average age of athletes may suggest an Olympic future that values sustained athlete health and longevity, indicating a need for management to invest in athlete development programs that cater to career sustainability.

The decrease in performance scores suggests that the Games are becoming more competitive. This competitive intensity implies that national committees and sponsors may need to focus on advanced training methods and analytics to stay ahead. The diversification of sports, with a more even distribution across individual, team, water, and winter sports, points to the potential for a broader commercial appeal and increased marketing opportunities. For example, brands could expand their sponsorship and target emerging sports with growing fanbases.

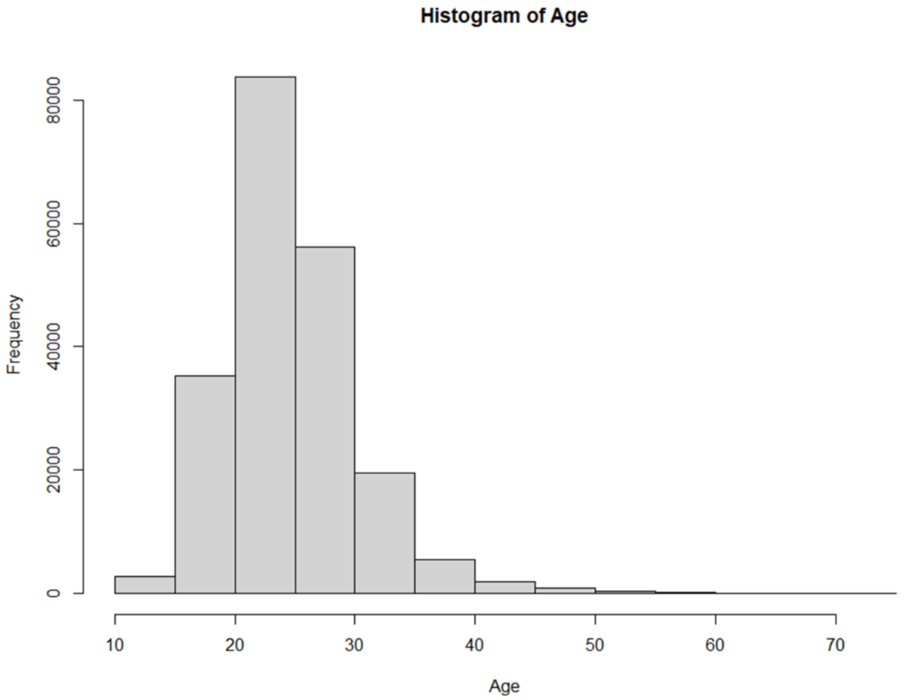
At the same time, the findings from these models suggest several strategic actions:

- **Talent Development:** Investment in long-term athlete development programs can help nations nurture younger talents (as seen in Cluster 2) and support seasoned athletes (as seen in Cluster 3) to maintain their competitive edge.
- **Gender Equality Initiatives**: With female participation nearing parity in some clusters, there is a clear opportunity for organizations to focus on gender equality both in sport and in related business initiatives, such as marketing and merchandise.
- **Seasonal Expansion:** The inclusion of winter sports and the rise in non-traditional sports categories suggest that the Olympic programming can be further diversified. This provides an opportunity for expanding the Games into new markets and seasons, which can attract new audiences and investors.
- **Technological Integration:** The future of the Games may involve greater integration of technology, not only in the sports themselves but also in how they are consumed by viewers. This opens doors for innovations in live streaming, virtual reality experiences, and interactive fan engagement.

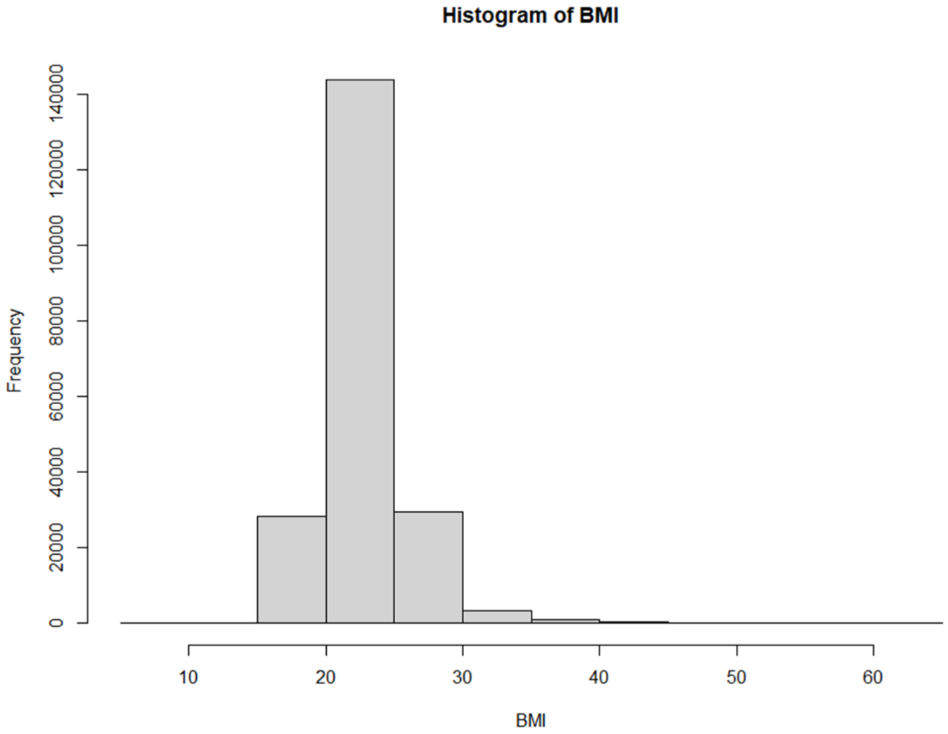
In conclusion, the patterns unearthed by these analytical models show an Olympics that is evolving with global societal trends. The managerial implications are vast, ranging from athlete management and training to marketing and global strategic positioning. Businesses and governing bodies alike can harness this data to inform future decisions, ensuring the continued growth and relevance of the Olympic Games.

6. Appendix

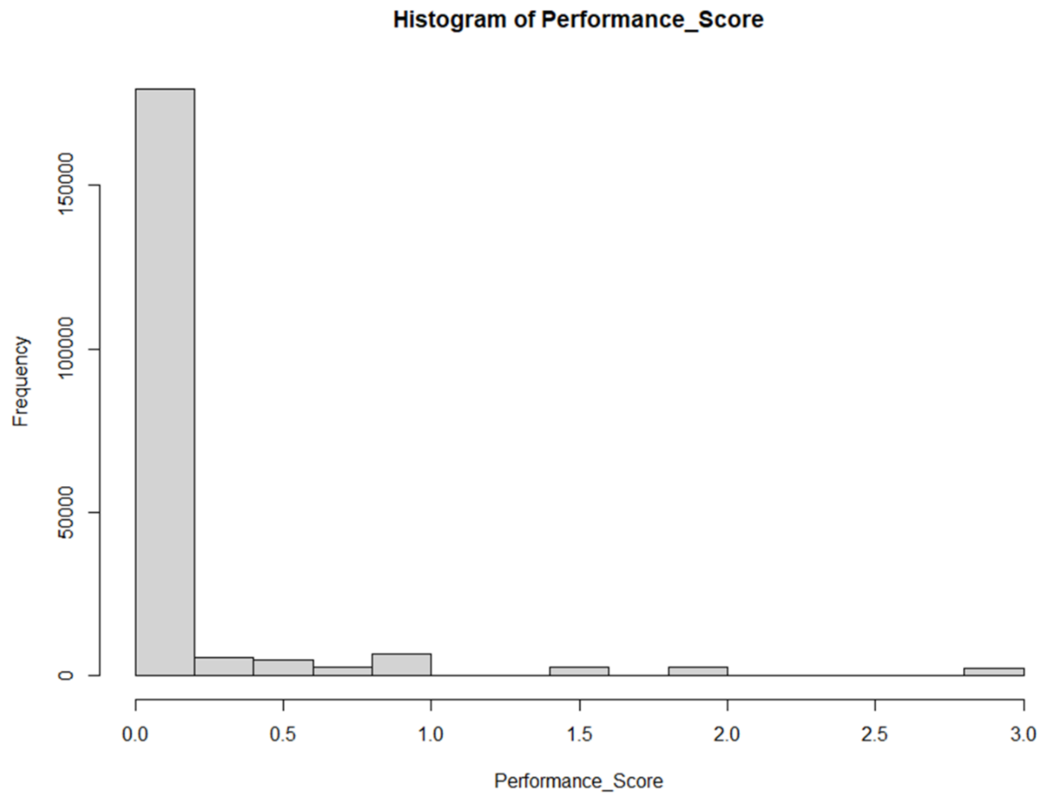
I. Historical Distribution (1896 to 2016) of Athletes' Age



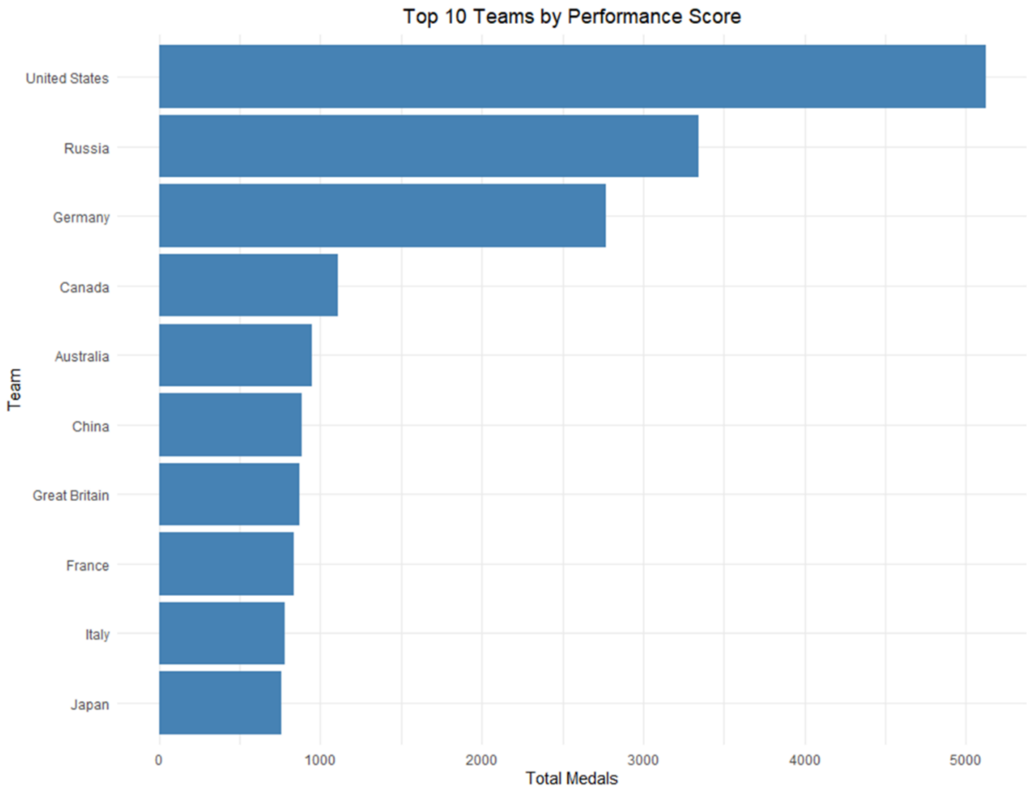
II. Historical Distribution (1896 to 2016) of Athletes' Body Mass Index (BMI)



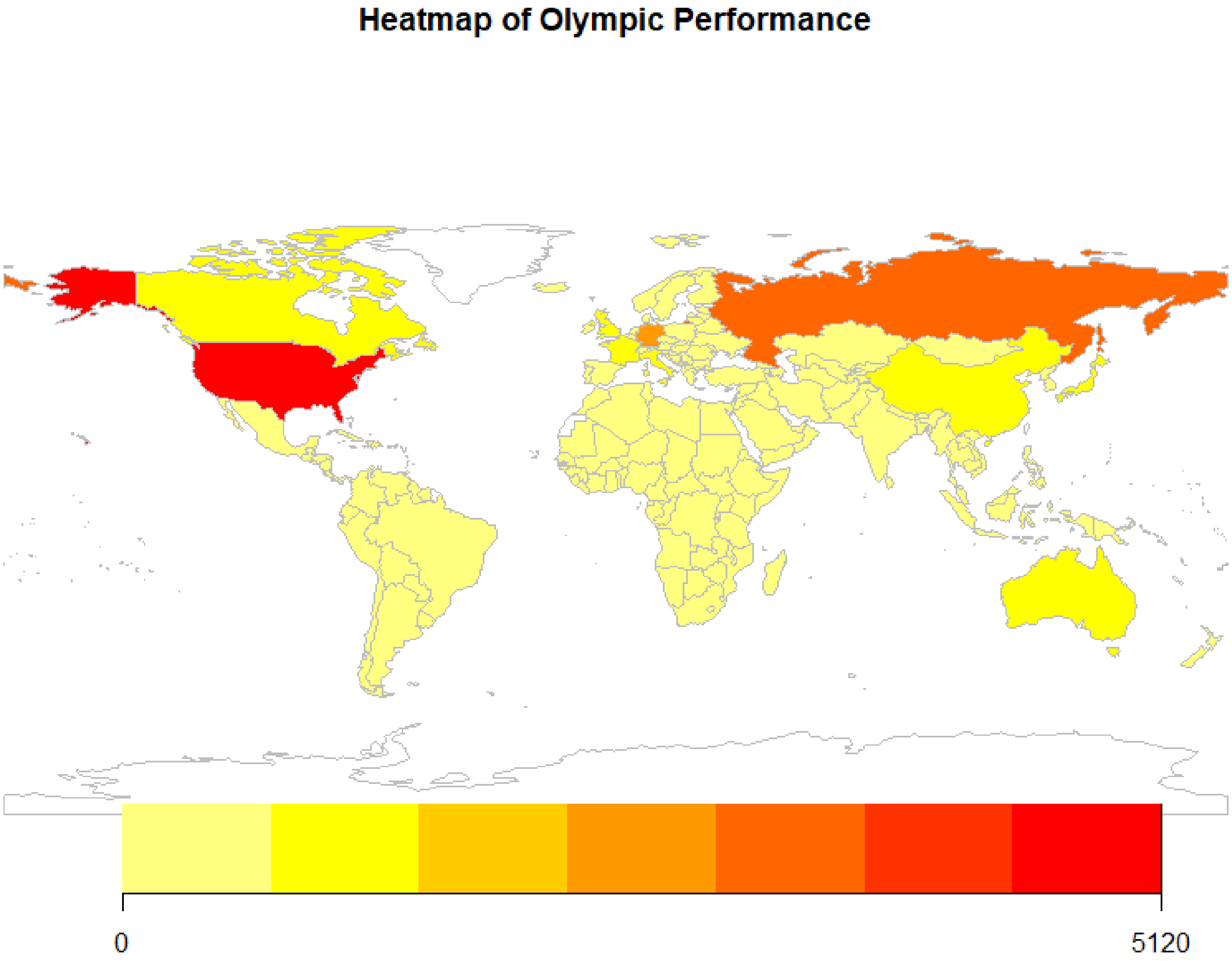
III. Historical Distribution (1896 to 2016) of Athletes' Performance Score



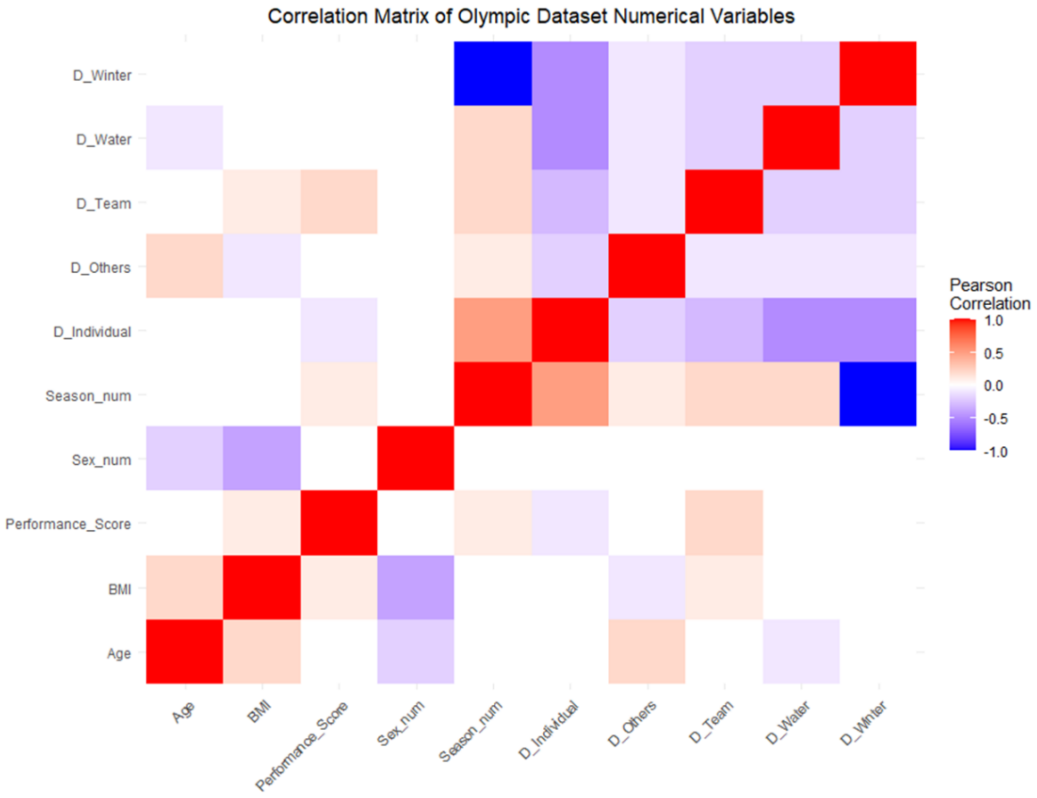
IV. Top 10 Olympic Teams by Historical Performance Score



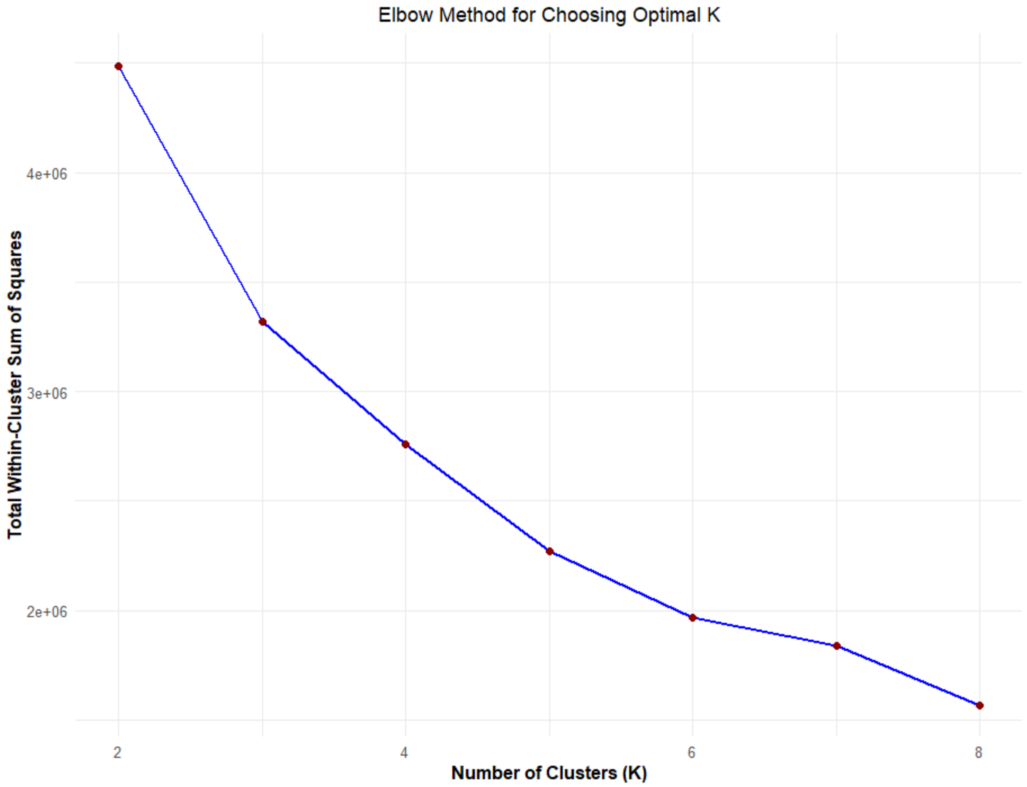
V. World Heatmap of Olympic Games Performance Score



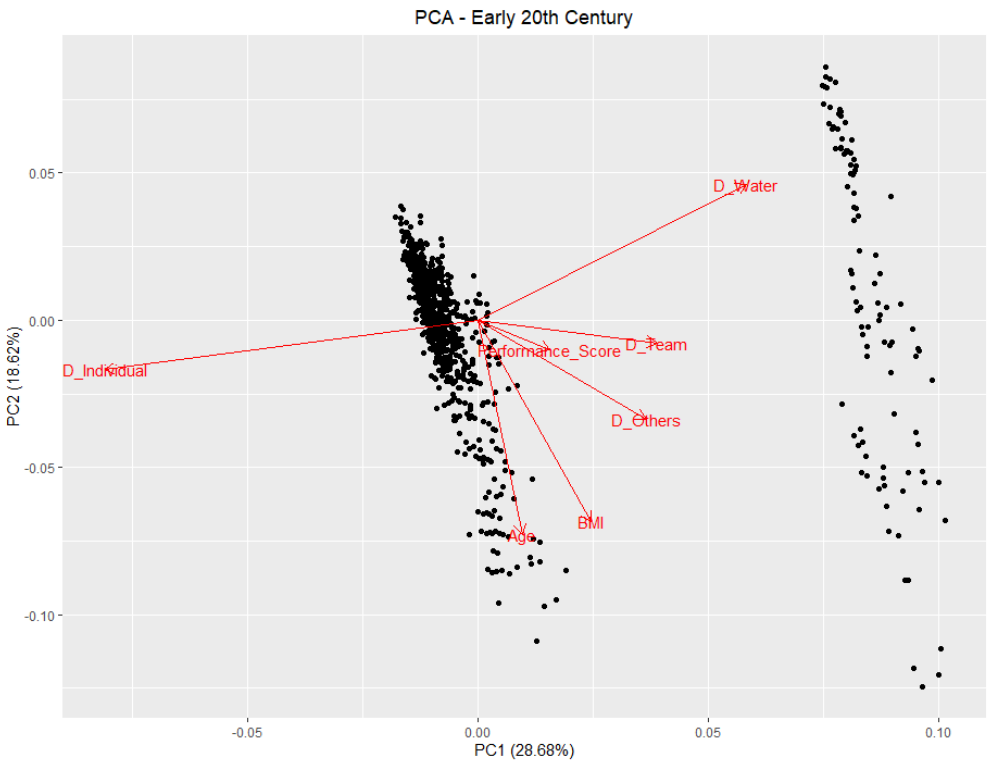
VI. Correlation Matrix of Numerical Variables



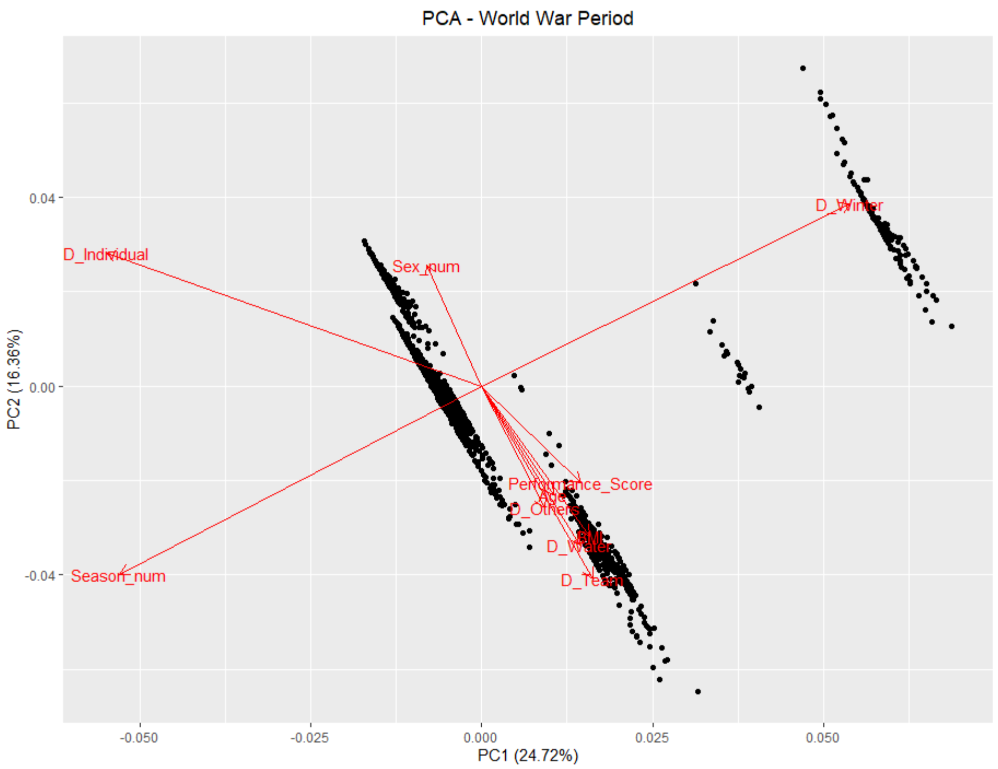
VII. Elbow Method Plot for Optimal K



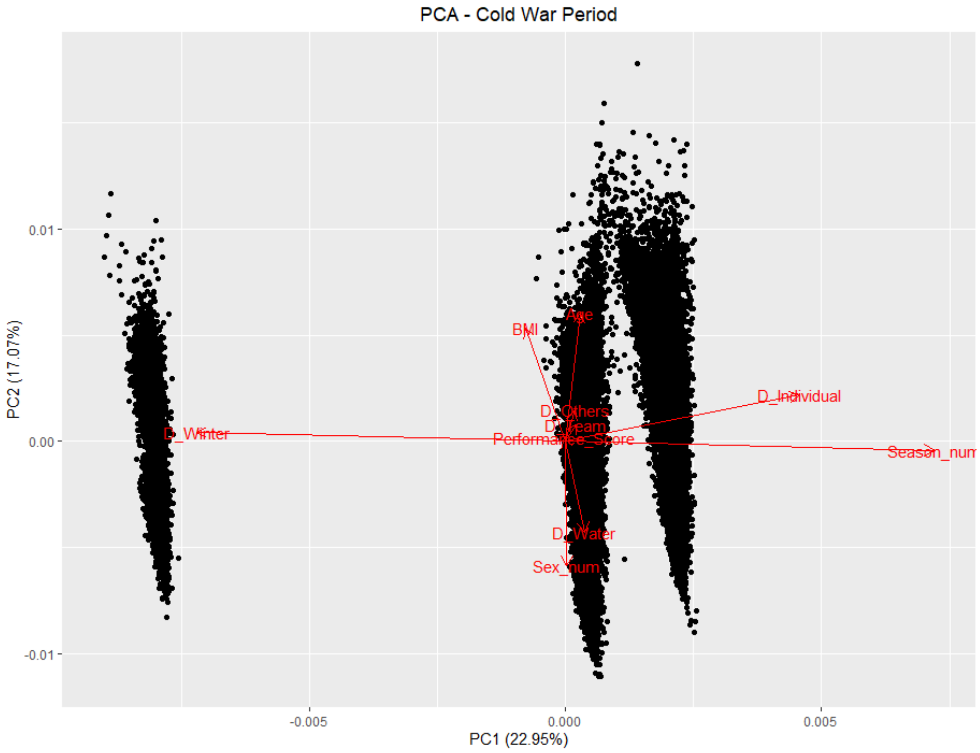
VIII. PCA Graph of Early 20th Century Period



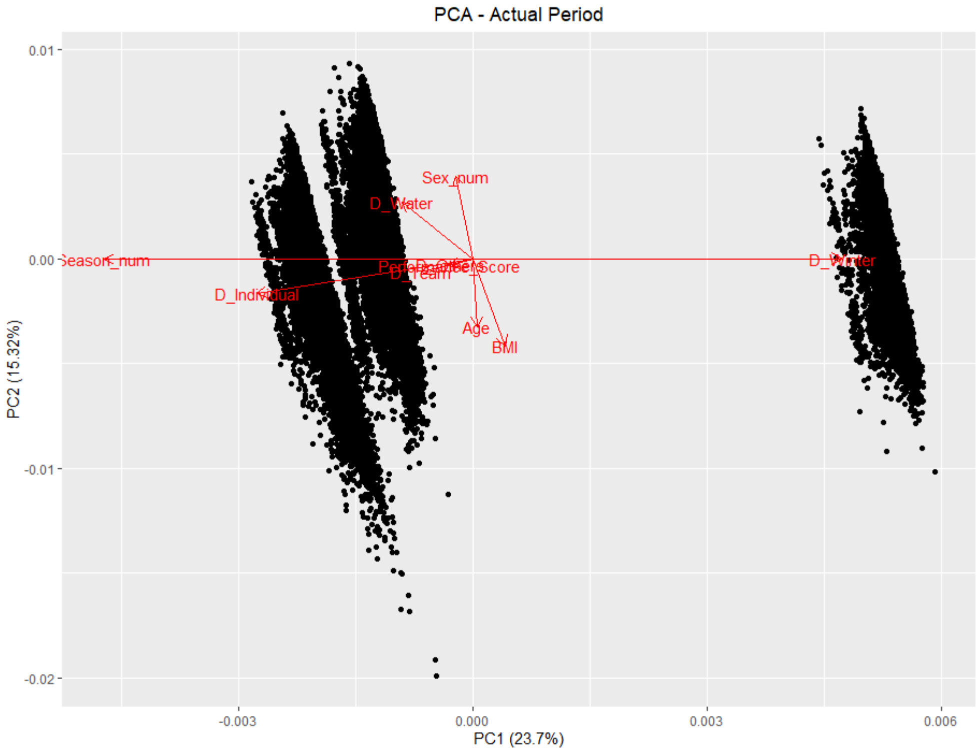
IX. PCA Graph of World War Period



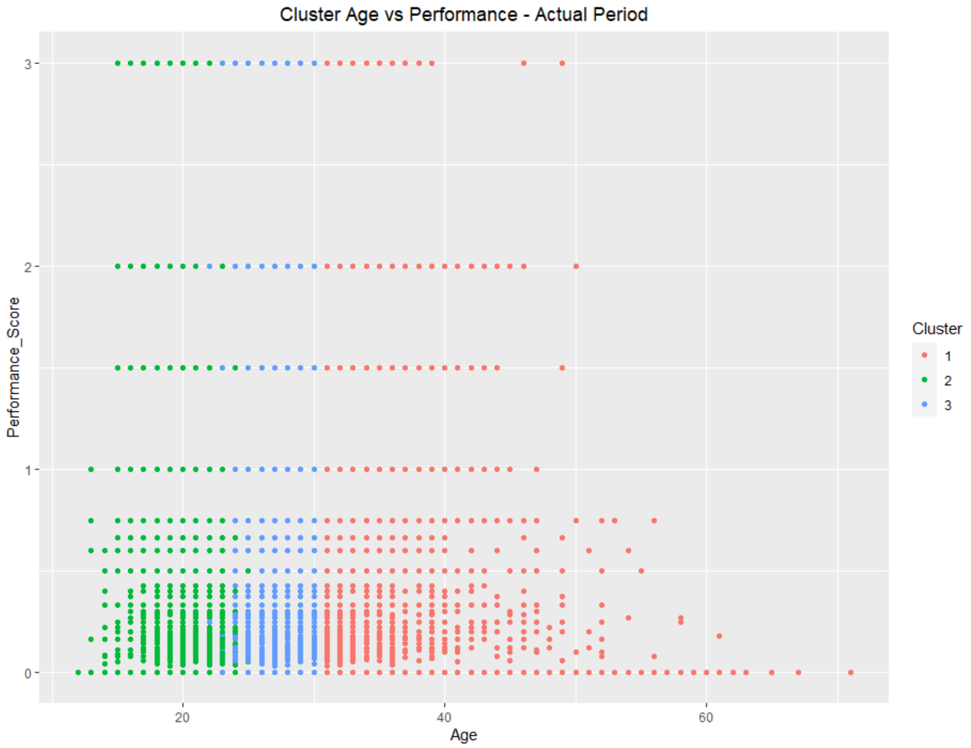
X. PCA Graph of Cold War Period



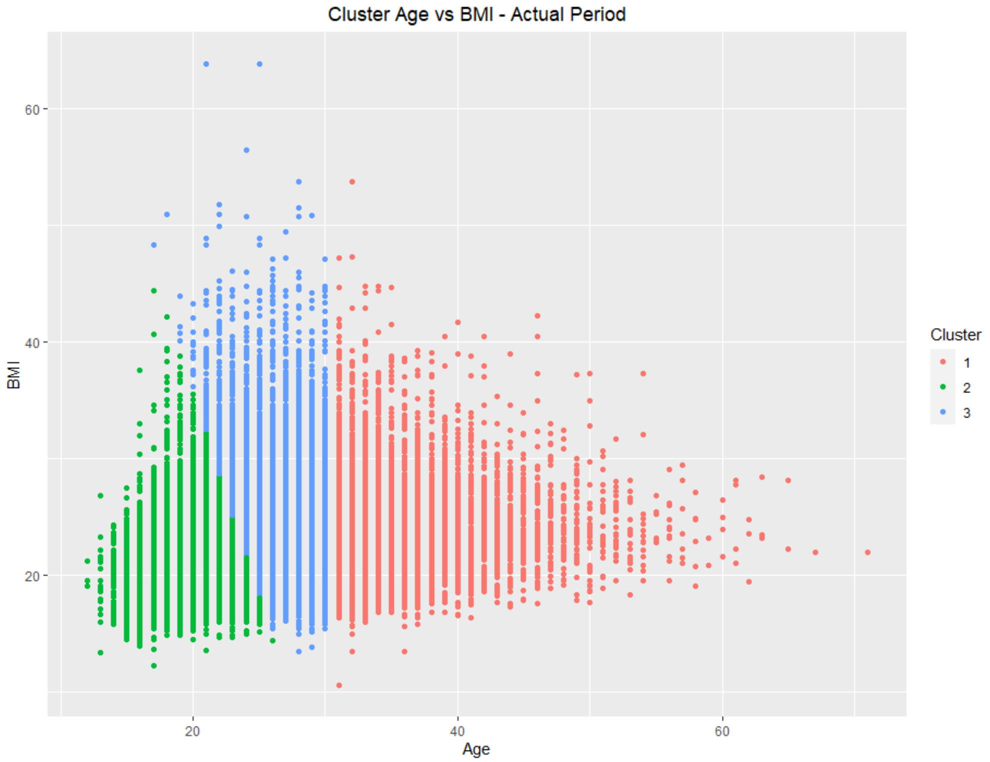
XI. PCA Graph of Actual Period



XII. K-Means Clusters Actual Period – Age & Performance Score



XIII. K-Means Clusters Actual Period – Age & BMI



7. Code

```
olympic = read.csv("C:/_McGill MMA 2023/2. Fall 2023/MGSC 661 Multivariate Statistical
Analysis/Final Project/Dataset 2 - Olympic events data.csv")

library(ggplot2)
library(reshape2)
#install.packages("rworldmap")
library(sp)
library(rworldmap)
library(MASS)
library(klaR)
require(psych)
library(ggfortify)
library(cluster)


# PRE-PROCESSING
################################

# Replace NA values in the Medal column
olympic$Medal[is.na(olympic$Medal)] = "No medal"

# Remove the missing values of Height and Weight
olympic = na.omit(olympic, cols = c("Height", "Weight"))

# Change some values from the Team column
olympic$Team = ifelse(olympic$Team == "East Germany", "Germany", olympic$Team)
olympic$Team = ifelse(olympic$Team == "West Germany", "Germany", olympic$Team)
olympic$Team = ifelse(olympic$Team == "Chinese Taipei", "China", olympic$Team)
olympic$Team = ifelse(olympic$Team == "Soviet Union", "Russia", olympic$Team)

# Create a column for Body Mass Index (BMI)
olympic$BMI = (olympic$Weight) / (olympic$Height/100)^2

# Create a column that categorizes the years of the games
olympic$Year_category = ifelse(olympic$Year < 1920, "Early 20th Century",
                        ifelse(olympic$Year < 1948, "World War Period",
                        ifelse(olympic$Year < 1992, "Cold War Period",
                        ifelse(olympic$Year >= 1992, "Actual Period", NA))))

# Create a column that categorizes the sports
winter_sports = c("Alpine Skiing", "Biathlon", "Bobsleigh", "Cross Country Skiing", "Curling", "Figure
Skating", "Freestyle Skiing", "Ice Hockey", "Luge", "Nordic Combined", "Short Track Speed Skating",
"Skeleton", "Ski Jumping", "Snowboarding", "Speed Skating")
```

```
team_sports = c("Baseball", "Basketball", "Beach Volleyball", "Football", "Handball", "Hockey", "Ice
Hockey", "Rugby", "Rugby Sevens", "Softball", "Volleyball", "Water Polo")
individual_sports = c("Archery", "Athletics", "Badminton", "Boxing", "Cycling", "Fencing", "Golf",
"Gymnastics", "Judo", "Modern Pentathlon", "Shooting", "Table Tennis", "Taekwondo", "Tennis",
"Trampolining", "Weightlifting", "Wrestling")
water_sports = c("Canoeing", "Diving", "Rowing", "Sailing", "Swimming", "Synchronized Swimming")
others = c("Aeronautics", "Alpinism", "Art Competitions", "Basque Pelota", "Cricket", "Croquet",
"Equestrianism", "Jeu De Paume", "Lacrosse", "Military Ski Patrol", "Motorboating", "Polo",
"Racquets", "Rhythmic Gymnastics", "Roque", "Tug-Of-War", "Triathlon")

olympic$Sport_category = ifelse(olympic$Sport %in% winter_sports, "Winter Sport",
                    ifelse(olympic$Sport %in% team_sports, "Team Sport",
                    ifelse(olympic$Sport %in% individual_sports, "Individual Sport",
                    ifelse(olympic$Sport %in% water_sports, "Water Sport",
                    ifelse(olympic$Sport %in% others, "Others", NA)))))

# Create a column that indicates the country where the game was held
olympic$Country = ifelse(olympic$Year == 1896, "Greece",
              ifelse(olympic$Year == 1900, "France",
              ifelse(olympic$Year == 1904, "United States",
              ifelse(olympic$Year == 1908, "United Kingdom",
              ifelse(olympic$Year == 1912, "Sweden",
              ifelse(olympic$Year == 1920, "Belgium",
              ifelse(olympic$Year == 1924, "France",
              ifelse(olympic$Year == 1928, "Netherlands",
              ifelse(olympic$Year == 1932, "United States",
              ifelse(olympic$Year == 1936, "Germany",
              ifelse(olympic$Year == 1948, "United Kingdom",
              ifelse(olympic$Year == 1952, "Finland",
              ifelse(olympic$Year == 1956, "Australia",
              ifelse(olympic$Year == 1960, "Italy",
              ifelse(olympic$Year == 1964, "Japan",
              ifelse(olympic$Year == 1968, "Mexico",
              ifelse(olympic$Year == 1972, "Germany",
              ifelse(olympic$Year == 1976, "Canada",
              ifelse(olympic$Year == 1980, "Russia",
              ifelse(olympic$Year == 1984, "United States",
              ifelse(olympic$Year == 1988, "South Korea",
              ifelse(olympic$Year == 1992, "Spain",
              ifelse(olympic$Year == 1996, "United States",
              ifelse(olympic$Year == 2000, "Australia",
              ifelse(olympic$Year == 2004, "Greece",
              ifelse(olympic$Year == 2008, "China",
              ifelse(olympic$Year == 2012, "United Kingdom",
              ifelse(olympic$Year == 2016, "Brazil",
              ifelse(olympic$Year == 1994, "Norway",
              ifelse(olympic$Year == 1998, "Japan",
              ifelse(olympic$Year == 2002, "United States",
               ifelse(olympic$Year == 2006, "Italy",
               ifelse(olympic$Year == 2010, "Canada",
               ifelse(olympic$Year == 2014, "Russia",
               ifelse(olympic$Year == 1906, "Greece", NA)))))))))))))))))))))))))))))))))))

# Create a new sex column that shows numerical values
olympic$Sex_num = ifelse(olympic$Sex == "F", 1, 0)

# Create a new season column that shows numerical values
olympic$Season_num = ifelse(olympic$Season == "Summer", 1, 0)

# Create a new medal column that shows numerical values
olympic$Medal_num = ifelse(olympic$Medal == "Gold", 3,
                ifelse(olympic$Medal == "Silver", 2,
                ifelse(olympic$Medal == "Bronze", 1, 0)))

# Create a performance score for each person
olympic$Total_Events_Participated = ave(rep(1, nrow(olympic)), olympic$ID, FUN = sum)
olympic$Performance_Score = olympic$Medal_num * (1 / olympic$Total_Events_Participated)

# Dummify the sport_category variables
sport_dummies = model.matrix(~olympic$Sport_category - 1, data=olympic)
colnames(sport_dummies) = paste("SportCat", colnames(sport_dummies), sep="_")
olympic = cbind(olympic, sport_dummies)
colnames(olympic)[(ncol(olympic)-4):ncol(olympic)] = c("D_Individual", "D_Others", "D_Team",
"D_Water", "D_Winter")


attach(olympic)

# Count unique values
length(unique(Country))
length(unique(Team))
length(unique(Sport))
length(unique(Year))
length(unique(Name))
table(Sex)
table(Medal)

US = olympic[olympic$Team == "United States", ]
Germany = olympic[olympic$Team == "Germany", ]
Russia = olympic[olympic$Team == "Russia", ]
Canada = olympic[olympic$Team == "Canada", ]
France = olympic[olympic$Team == "France", ]
Italy = olympic[olympic$Team == "Italy", ]
UK = olympic[olympic$Team == "Great Britain", ]
Japan = olympic[olympic$Team == "Japan", ]
Australia = olympic[olympic$Team == "Australia", ]
```

```
China = olympic[olympic$Team == "China", ]



# EXPLORATORY DATA ANALYSIS
##############################

# Correlation matrix for numerical values
numerical_var = olympic[c("Age", "BMI", "Performance_Score", "Sex_num", 'Season_num',
"D_Individual", "D_Others", "D_Team", "D_Water", "D_Winter")]
cor_matrix = round(cor(numerical_var), 1)

melted_cor_matrix = melt(cor_matrix)

ggplot(data = melted_cor_matrix, aes(Var1, Var2, fill = value)) +
  geom_tile() +
  scale_fill_gradient2(low = "blue", high = "red", mid = "white",
                       midpoint = 0, limit = c(-1,1), space = "Lab",
                       name="Pearson\nCorrelation") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        axis.title = element_blank(),
        plot.title = element_text(hjust = 0.5)) +
  ggtitle("Correlation Matrix of Olympic Dataset Numerical Variables")

# Check how numerical values are distributed
hist(Age)
hist(BMI)
hist(Performance_Score)

# Obtain the Top teams in terms of performance score
team_performance = aggregate(Performance_Score ~ Team, data = olympic, sum)
colnames(team_performance) = c("Team", "Performance_Score")
top_teams = head(team_performance[order(-team_performance$Performance_Score), ], 10)

ggplot(top_teams, aes(x = reorder(Team, Performance_Score), y = Performance_Score)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  theme_minimal() +
  labs(title = "Top 10 Teams by Performance Score", x = "Team", y = "Total Medals") +
  coord_flip() +
  theme(plot.title = element_text(hjust = 0.5))

map1 = joinCountryData2Map(team_performance, joinCode = "NAME", nameJoinColumn = "Team")
mapCountryData(map1, nameColumnToPlot="Performance_Score", catMethod="fixedWidth",
mapTitle="Heatmap of Olympic Performance",
               colourPalette="heat", addLegend=TRUE)
```

```
# CLUSTERING MODEL (PCA)
##############################

# Create a dataframe for each period of time
columns_1 = c("ID", "Age", "BMI", "Medal_num", "D_Individual", "D_Others", "D_Team",
"D_Water") # No woman in sports nor winter sports in the early 20th
early_20th = olympic[olympic$Year_category == "Early 20th Century", ]
early_20th = early_20th[columns_1]


columns_2 = c("ID", "Age", "BMI", "Medal_num", "Sex_num", 'Season_num', "D_Individual",
"D_Others", "D_Team", "D_Water", "D_Winter")
world_war = olympic[olympic$Year_category == "World War Period", ]
world_war = world_war[columns_2]

cold_war = olympic[olympic$Year_category == "Cold War Period", ]
cold_war = cold_war[columns_2]

actual_period = olympic[olympic$Year_category == "Actual Period", ]
actual_period = actual_period[columns_2]


#Recalculate the performance score for each period of time
early_20th$Total_Events_Participated = ave(rep(1, nrow(early_20th)), early_20th$ID, FUN = sum)
early_20th$Performance_Score = early_20th$Medal_num * (1 / early_20th$Total_Events_Participated)
early_20th = subset(early_20th, select = -c(ID, Medal_num, Total_Events_Participated))

world_war$Total_Events_Participated = ave(rep(1, nrow(world_war)), world_war$ID, FUN = sum)
world_war$Performance_Score = world_war$Medal_num * (1 / world_war$Total_Events_Participated)
world_war = subset(world_war, select = -c(ID, Medal_num, Total_Events_Participated))

cold_war$Total_Events_Participated = ave(rep(1, nrow(cold_war)), cold_war$ID, FUN = sum)
cold_war$Performance_Score = cold_war$Medal_num * (1 / cold_war$Total_Events_Participated)
cold_war = subset(cold_war, select = -c(ID, Medal_num, Total_Events_Participated))

actual_period$Total_Events_Participated = ave(rep(1, nrow(actual_period)), actual_period$ID, FUN =
sum)
actual_period$Performance_Score = actual_period$Medal_num * (1 /
actual_period$Total_Events_Participated)
actual_period = subset(actual_period, select = -c(ID, Medal_num, Total_Events_Participated))


# Run cluster for each period of time
pca_early_20th = prcomp(early_20th, scale. = TRUE)
autoplot(pca_early_20th, data = early_20th, loadings = TRUE, loadings.label = TRUE) + ggtitle("PCA
- Early 20th Century") + theme(plot.title = element_text(hjust = 0.5))
```

```
pca_world_war = prcomp(world_war, scale. = TRUE)
autoplot(pca_world_war, data = world_war, loadings = TRUE, loadings.label = TRUE) + ggtitle("PCA
- World War Period") + theme(plot.title = element_text(hjust = 0.5))

pca_cold_war = prcomp(cold_war, scale. = TRUE)
autoplot(pca_cold_war, data = cold_war, loadings = TRUE, loadings.label = TRUE) + ggtitle("PCA -
Cold War Period") + theme(plot.title = element_text(hjust = 0.5))

pca_actual_period = prcomp(actual_period, scale. = TRUE)
autoplot(pca_actual_period, data = actual_period, loadings = TRUE, loadings.label = TRUE) +
ggtitle("PCA - Actual Period") + theme(plot.title = element_text(hjust = 0.5))



# CLUSTERING MODEL (k-means)
##############################

# Obtain the optimal number of k with the elbow method
numerical_var = olympic[c("Age", "BMI", "Performance_Score", "Sex_num", 'Season_num',
"D_Individual", "D_Others", "D_Team", "D_Water", "D_Winter")]
wss = sapply(2:8, function(k){kmeans(numerical_var, centers=k)$tot.withinss})
elbow_df = data.frame(k = 2:8, wss = wss)

ggplot(elbow_df, aes(x = k, y = wss)) +
  geom_line(color = "blue", size = 1) +
  geom_point(color = "darkred", size = 2) +
  theme_minimal() +  # Minimal theme
  labs(
    title = "Elbow Method for Choosing Optimal K",
    x = "Number of Clusters (K)",
    y = "Total Within-Cluster Sum of Squares"
  ) +
  theme(
    plot.title = element_text(hjust = 0.5),
    axis.title = element_text(size = 12, face = "bold"),
    axis.text = element_text(size = 10))


# Run cluster for the actual period of time with optimal k=3
km_actual_period = kmeans(actual_period, 3)
km_actual_period
actual_period$Cluster = as.factor(km_actual_period$cluster)
ggplot(actual_period, aes(x = Age, y = Performance_Score, color = Cluster)) +
  geom_point() +  ggtitle("Cluster Age vs Performance - Actual Period") +
  theme(plot.title = element_text(hjust = 0.5))
ggplot(actual_period, aes(x = Age, y = BMI, color = Cluster)) +
  geom_point() +  ggtitle("Cluster Age vs BMI - Actual Period") +
  theme(plot.title = element_text(hjust = 0.5))
```