

**Final Project**

**Data Mining Applications**

**Olympics Data Analysis**

**PROJECT REPORT**

**Team Members: Arvind Pawar, Nidhi Shah, Viraj Shukla**

**Assignment Title: Week 3 – Assignment 3**

**Course Number: ALY6040 - 21289**

**Title: Data Mining Applications**

**Academic Term: Winter 2019**

**Instructor’s Name: Amin Karimpour**

**Assignment completion date: 02/16/19**

**Project Description**

The Olympic Games are a global games celebration, held at regular intervals of four years. The games are held independently as Summer and Winter Games where competitors from all over the globe participate. The event goal is to grow individuals, through a game, and add to world harmony. It is exciting to know that more than 200 nations take part in this event. Innovation in data analytics has become increasingly important in producing marginal gains in any area of excellence. Using data analysis teams can perform better and take decisions based on insights from historical data and certain pieces of evidence. Technology and innovations have offered more options for analysis, so this area has grown in importance in high-performance sport. Well, the sport has funny things of twisting and turning the expected into something that no one could ever imagine, however, advanced analytics technologies can tell us the correlations of various factors and nearly accurate prediction that can help to develop game strategies for better performance.

We worked on historical data of Olympic amusements from Athens 1986 to Rio 2016 and applied analytical techniques, determined relationship between various columns and changes made every four years since 1986. The goal of this project is to implement the knowledge of data mining techniques and perform analysis on the historical dataset of Olympics games. The importance of sports games has grown so much that nearly every nation is now representing their teams in the Olympics. This growth has created many challenges, and every country wants to perform the best and win medals. Each country and its teams need to understand the game trends and strategies of their competitor. Insights from the available historical dataset would help to understand the game strategies of winning teams and their performance trends over the years. This project gave us a chance to demonstrate:

• The Olympics development over time.

• The changing trends of participation

• Winning attributes of the teams and gender

• Performance of women over the years and so on.

It is also essential to know the effects of season on events and performances of athletes and their winnings. The analysis model used can help to predict the effect of games concerning the season in which they are going to conduct. We first went through the data to see how it looks and if there are any missing values. We used “tidyverse” package that includes various other required packages such as ggplot, dplyr, tidyr, readr and so on. To predict the effects of season on games, we used the Decision Tree algorithm for this dataset. Decision tree belongs to the family of supervised learning algorithms. This algorithm can be used for solving regression and classification problems. Decision Tree algorithm is easy compared with other classification algorithms, and it helps to create a training model which can use to predict a value of target variables by learning decision rules derived from training data and by using tree representation it tries to solve the problem. Insights from this project will envision the performance trends in graphical presentation, and further can help to predict, plan and prepare for future endeavors accordingly.

**Dataset Used**

* The Olympics historic dataset is collected from www.sports-reference.com in May 2018.
* The files are “athlete\_events.csv” and “noc\_regions.csv”.
* The file “athlete\_events.csv” contains 271116 rows and 15 columns. It also includes the data related to medal won by the athletes.

**Data Attributes:**

A screenshot of a cell phone

Description automatically generated

The structure of the data set looks like:

A screenshot of a cell phone

Description automatically generated

**Data Exploration**

Analysis of the data is meaningful only if the data quality is good. This step is performed to clean and manipulate the data in order to extract the valuable information that can be used further for analysis and predictions.

Here, we perform

* Loading of the data, installation of the required packages and libraries.
* Cleaning of the data by determining missing values, filtering out the groups based on requirements and combining the data frames.
* Identifying the valuable records, correlation and extract them to interpret the results and predictions.

We Installed the tidyverse package by running below code and then loaded the basic tidyverse for that we run the library(tidyverse) as shown below. We then installed the other packages that are required for the project.

#Olympics Games Analysis

#------- Installing Required Packages --------

>install.packages("tidyverse")

>install.packages("dplyr")

>install.packages("ggmap")

>install.packages("tseries")

>install.packages("rpart")

>install.packages("rpart.plot")

>install.packages("ggplot2")

>library(stats)

>library(dplyr)

>library(ggplot2)

>library(tidyverse)

>library(rvest)

>library(magrittr)

>library(ggmap)

>library(stringr)

>library(forecast)

>library(tseries)

The tidyverse is a set of packages that work together as they share common API design and data representations. This package eases the process of installation and loading of core packages from tidyverse in a single command.

This package includes:

* ggplot2: used for creating the graphics by mapping variables to the aesthetics.
* dplyr: used for data manipulation
* tidyr: used to get the clean data
* readr: used to read the csv files
* stringr: provides set of functions that makes working with strings as easy as possible

After this, we read the data and got the necessary details:

#Read the data

>athlete\_events<-read\_csv("~/Downloads/120-years-of-olympic-history-athletes-and-results/athlete\_events.csv")

>noc\_regions <- read\_csv("~/Downloads/120-years-of-olympic-history-athletes-and-results/noc\_regions.csv")

We updated the attributes by replacing the NA values with the mean average of that attribute. To update the NA values with mean we use if else statement, ave and mean function.

#Updating the attributes to remove NA values

>athlete\_events$Weight <- ifelse(is.na(athlete\_events$Weight),

ave(athlete\_events$Weight,FUN = function(x) mean(x,na.rm = TRUE)),

athlete\_events$Weight)

>athlete\_events$Height <- ifelse(is.na(athlete\_events$Height),

ave(athlete\_events$Height,FUN = function(x) mean(x,na.rm = TRUE)),

athlete\_events$Height)

>athlete\_events$Age <- ifelse(is.na(athlete\_events$Age),

ave(athlete\_events$Age,FUN = function(x) mean(x,na.rm = TRUE)),

athlete\_events$Age)

To see the dimension of dataset, we used dim() function.

#Dimensions of data

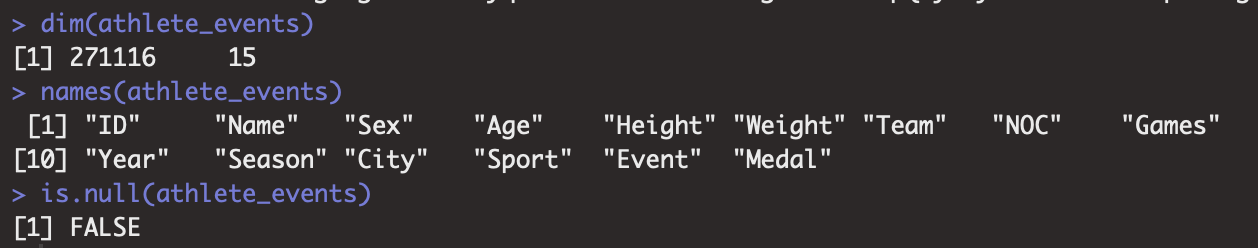
dim(athlete\_events)

#Name of the columns

names(athlete\_events)

#Checking for null values

is.null(athlete\_events)



There are total 15 columns and 271116 rows of data in the given dataset. And now there are no null values in athlete data.

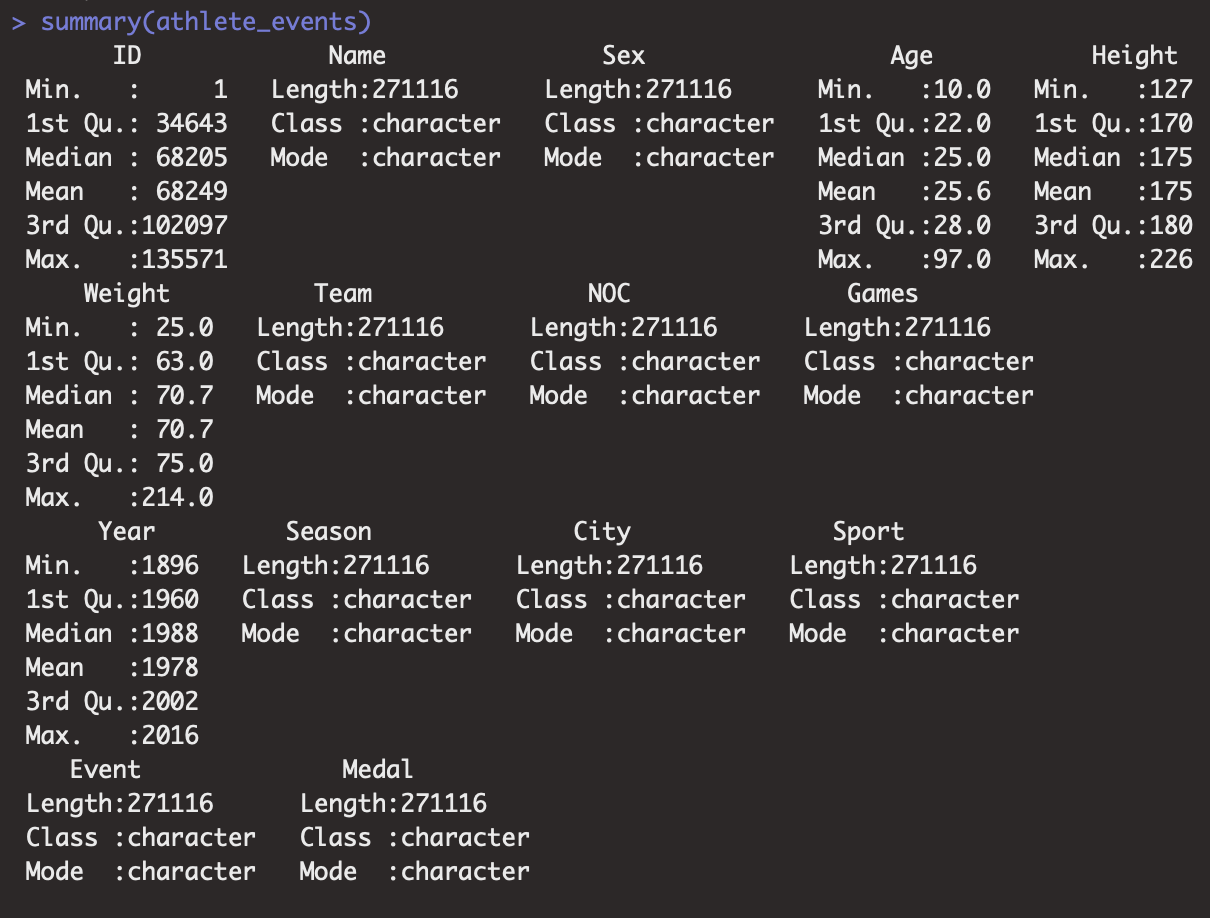
#Taking as Data Frame

>athlete\_events = as.data.frame(athlete\_events)

>str(athlete\_events)

#Summary of data

summary(athlete\_events)



#Data Exploration-(EDA) with plots

Participation Statistics:

Using group\_by() and summarize() functions:

Group by() function is similar to SQL group by call. Many data analysis tasks are performed using the “split-apply-combine”. Split the data into groups, apply some analysis to each group, and then combine the results., dplyr package makes this very easy.

#Getting the count of athletes, nations and events

>counts <- athlete\_events %>%

group\_by(Year, Season) %>%

summarize(

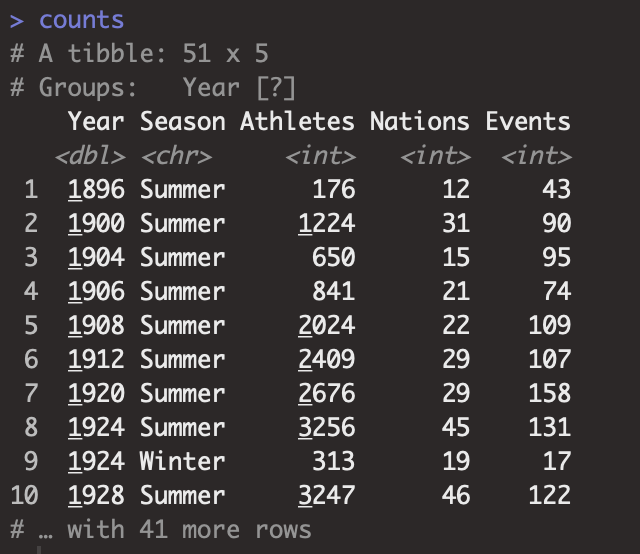
Athletes = length(unique(ID)),

Nations = length(unique(NOC)),

Events = length(unique(Event))

)

>counts



#Plotting the participation statistics

#Athletes

>athletes\_plot <- ggplot(counts, aes(x=Year, y=Athletes, group=Season, color=Season)) +

geom\_point(size=2) + ggtitle("Number of Athletes participated over the Years") +

geom\_line()

>athletes\_plot

A close up of a map

Description automatically generated

Here we can see the growth level of athletes is starting from approximately handful athletes to more than 11000 after 2010. We can also observe that participants are not much for the winter games as ice and snow sports are not that popular.

Then, we moved further to see participation of countries trend in both the seasons.

#Nations

>nations\_plot <- ggplot(counts, aes(x=Year, y=Nations, group=Season, color=Season)) +

geom\_point(size=2) + ggtitle("Number of Countries participated over the Years") +

geom\_line()

>nations\_plot

A close up of a map

Description automatically generated

Here, we can observe that number of nations increased over the time. We can see a slight dip in the nations during 1980 in the summer games as 66 nations boycotted including USA.

#No. of Events

>events\_plot <- ggplot(counts, aes(x=Year, y=Events, group=Season, color=Season)) +

geom\_point(size=2) + ggtitle("Number of Events occurred over the Years") +

geom\_line()

>events\_plot

A close up of a map

Description automatically generated

We can observe that the list of events grew indeterminately. The summer Games events crossed 300 around the year 2000 whereas the winter games just reached near 100 events.

Let’s see the participation of males and females in Olympics.

#Male and Female over the years

>mf\_counts<- athlete\_events %>% filter(Sport != "Art Competitions")

>original <- c(1994,1998,2002,2006,2010,2014)

>new <- c(1996,2000,2004,2008,2012,2016)

>for (i in 1:length(original)) {

mf\_counts$Year <- gsub(original[i], new[i], mf\_counts$Year)

}

>mf\_counts$Year <- as.integer(mf\_counts$Year)

>counts\_sex <- mf\_counts%>% group\_by(Year, Sex) %>%

summarize(Athletes = length(unique(ID)))

>counts\_sex$Year <- as.integer(counts\_sex$Year)

>gender\_plot <- ggplot(counts\_sex, aes(x=Year, y=Athletes, group=Sex, color=Sex)) +

geom\_point(size=2) +

geom\_line() +

labs(title = "Number of Male and Female participation over the Years") +

theme(plot.title = element\_text(hjust = 0.5))

>gender\_plot

A close up of a map

Description automatically generated

We can observe that participation of male athletes was higher in number than the participation female athletes.

However, the number women in the Olympics has been increasing since their first participation in 1900

More data analysis performed to look the winning statistics.

#Winning trends

#Total number of medals for top 10 nations

>medal\_counts <- athlete\_events %>%

filter(Medal != "<NA>") %>%

group\_by(Team) %>%

summarise(Total = length(Medal))%>%

arrange(desc(Total)) %>%

ungroup() %>%

mutate(country = reorder(Team,Total)) %>%

top\_n(10)

>total\_medals\_plot <- ggplot(medal\_counts, aes(x = country,y = Total)) +

geom\_bar(stat='identity',colour="white", fill = "lightgreen") +

geom\_text(aes(x = country, y = .1, label = paste0("(",round(Total,2),")",sep="")),

hjust=0, vjust=.5, size = 4, colour = 'black',

fontface = 'bold') +

labs(title = "Top 10 Nations - Total Number of Medals") +

theme(plot.title = element\_text(size=10),

axis.title = element\_text(size=10),

axis.text = element\_text(size=10)) +

labs(x = 'Country',

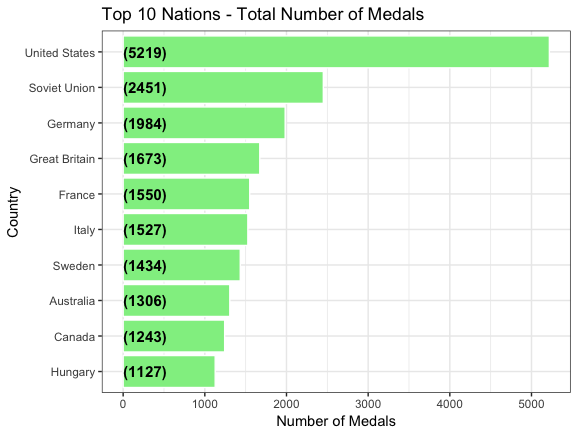
y = 'Number of Medals'

) +

coord\_flip() +

theme\_bw()

>total\_medals\_plot



United States have won the highest medals crossing 5000 in count over the time and then followed by Soviet Union, Germany and so on.

We have also determined the total medals for male and female as well.

# Count number of medals awarded to females at Olympics

>counts\_female <- athlete\_events %>% filter(Medal != "<NA>") %>% filter(Sex=="F") %>%

group\_by(NOC, Medal) %>%

summarize(Count=length(Medal))

# Order NOC by total medal count

NOC is a column in dataset that has region names.

>l\_female <- counts\_female %>%

group\_by(NOC) %>%

summarize(Total=sum(Count)) %>%

arrange(Total) %>%

top\_n(10) %>%

select(NOC)

>counts\_female$NOC<- factor(counts\_female$NOC, levels=l\_female$NOC)

# Plot female medals

>female\_medal\_plot <- ggplot(counts\_female, aes(x=NOC, y=Count, fill=Medal)) +

geom\_col() +

coord\_flip() +

scale\_fill\_manual(values=c("gold4","gold1", "gray70")) +

ggtitle("Medal counts for women at the 2016 Olympics") +

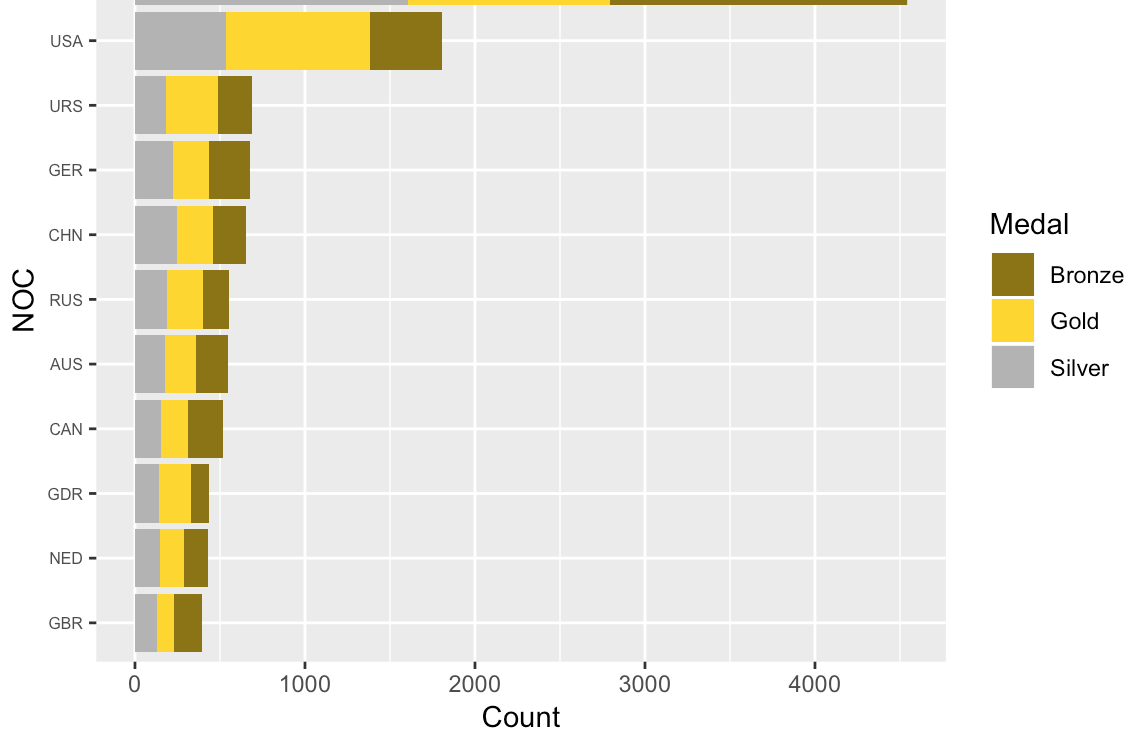
theme(plot.title = element\_text(hjust = 0.5),

axis.text.y = element\_text(size=6))

>female\_medal\_plot

We also categorized the medals to better understand the count of gold, silver and bronze medals.





We can see that overall medals earned by women are near to 2000 and USA is at the top to grab the highest medals.

# Count number of medals awarded to males at Olympics

>counts\_male <- athlete\_events %>% filter(Medal != "<NA>") %>% filter(Sex=="M") %>%

group\_by(NOC, Medal) %>%

summarize(Count=length(Medal))

# Order NOC by total medal count

>l\_male <- counts\_male %>%

group\_by(NOC) %>%

summarize(Total=sum(Count)) %>%

arrange(Total) %>%

top\_n(10) %>%

select(NOC)

>counts\_male$NOC<- factor(counts\_male$NOC, levels=l\_male$NOC)

# Plot male medals

>male\_medal\_plot <- ggplot(counts\_male, aes(x=NOC, y=Count, fill=Medal)) +

geom\_col() +

coord\_flip() +

scale\_fill\_manual(values=c("gold4","gold1", "gray70")) +

ggtitle("Medal counts for men at Olympics") +

theme(plot.title = element\_text(hjust = 0.5),

axis.text.y = element\_text(size=6))

>male\_medal\_plot



A screenshot of a cell phone

Description automatically generated

We can see that overall medals earned by men are around 2000.

USA is at the top to grab the highest medals overall.

We can also calculate number of medals for different sports. Let’s see number of medals for Basketball game.

#Analysis for Basketball

>sport\_bb<- athlete\_events %>%

filter(Sport == "Basketball")

#count number of medals awarded to each Team (Basketball)

>medal\_counts <- sport\_bb %>% filter(Medal != "<NA>") %>%

group\_by(Team, Medal) %>%

summarize(Count=length(Medal))

>medal\_counts

#order Team by total medal counts

>arrange\_levs <- medal\_counts %>%

group\_by(Team) %>%

summarize(Total=sum(Count)) %>%

arrange(Total) %>%

select(Team) %>%

top\_n(10)

>arrange\_levs

>medal\_counts$Team <- factor(medal\_counts$Team, levels=arrange\_levs$Team)

>bb\_plot <- ggplot(medal\_counts, aes(x=Team, y=Count, fill=Medal)) +

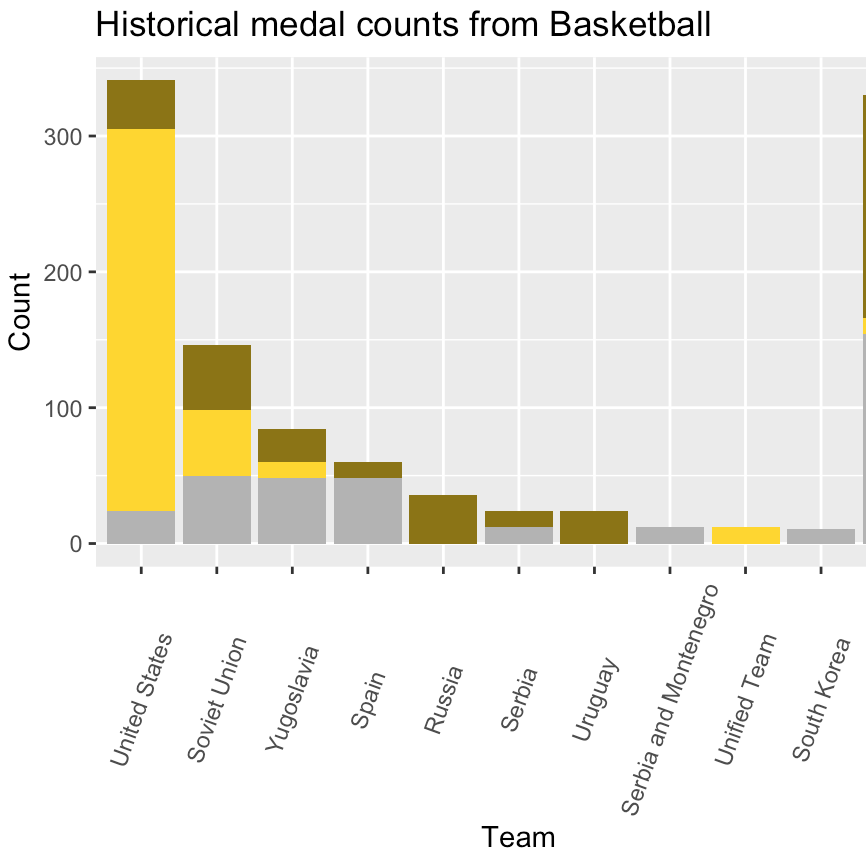
geom\_bar(stat = 'identity') +

scale\_fill\_manual(values=c("gold4","gold1","gray70")) +

ggtitle("Historical medal counts from Basketball")+

theme(axis.text.x = element\_text(angle=70, vjust=0.5))

>bb\_plot





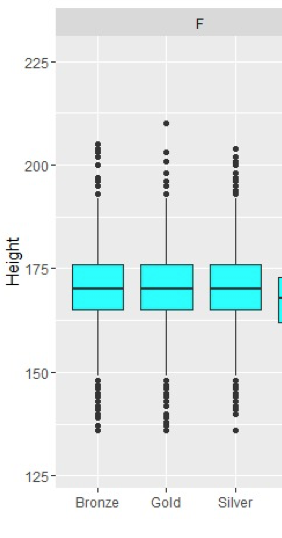
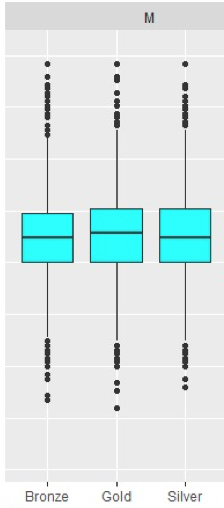
We can see that more than 350 medals are earned by USA.

Similarly, we can find the medal counts for other sports also that would be helpful to see in which sport a particular country is a champion.

We then tried to find how height and weight of athletes are related to winning the medals.

> ggplot(athletes\_events1, aes(x = Medal, y = Height)) +

geom\_boxplot(fill="cyan") + facet\_grid(.~Sex)

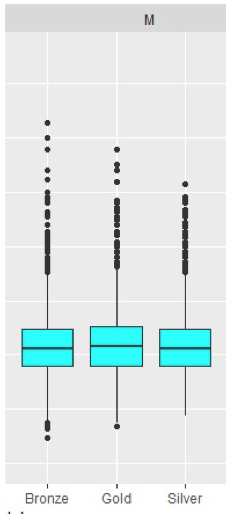
 

# the median height of athletes was around 170 cm who won the all kind of medals.

>ggplot(athletes\_events1, aes(x = Medal, y = Weight)) +

geom\_boxplot(fill="cyan") + facet\_grid(.~Sex)

A screenshot of a cell phone

Description automatically generated 

#Similarly, the median weight of athletes was around 75 kg who won the all kind of medals.

We can also consider the age factor of athletes. Differences in the age of medalists in different sports? Some sports seem to favor younger athletes more for example gymnastics and skating, while in other sports we see more veterans are success because of their experience.

**Algorithm Used**

**Decision tree:**

Decision is recursive subdividing fundamental tool in data mining. It encourages us investigate the structure of a lot of information, while growing simple to envision choice principles for foreseeing a clear cut (order tree) or ceaseless (relapse tree) result. The tree helps us to visualize the data much easier and even gives the probability of an event to occur based on a certain outcome at each and every branch.

R-code:

We are using the dataset columns, which are being assigned a NULL value as they don’t have much correlation in the classification Decision tree

> athlete\_events$Age <- as.numeric(athlete\_events$Age)

> athlete\_events$Height <- as.numeric(athlete\_events$Height)

> athlete\_events$Year <- as.numeric(athlete\_events$Year)

In the above R code, we are using the dataset columns, which are being converted into numeric values as they required before a conflict between data columns take place. Here we are converting Age, Height, Year of the dataset for classification using decision tree.

> library(rpart)

> library(rpart.plot)

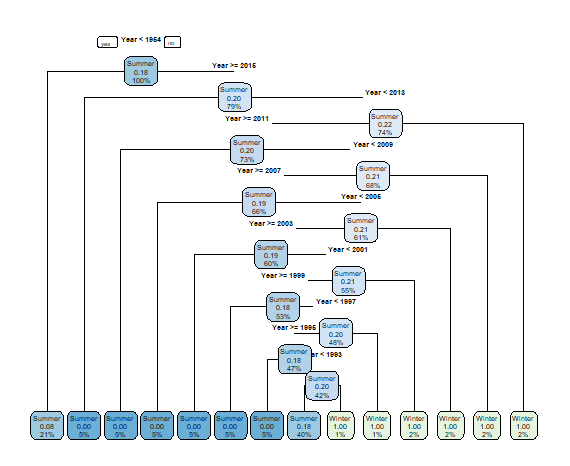
> dt\_model<-rpart(Season~.,data = athlete\_events)

>

> rpart.plot(dt\_model,type = 1)

We will be using rpart library as it contains the decision tree algorithm and a decision tree model can be build using the Rpart function. We have used the Season Attribute (column) of the data with every other columns of the athlete event dataset. dt\_model model is generated using the rpart which is the decision tree model. The given decision tree model is provided as input to rpart.plot. rpart.plot function is used to plot the decision tree graphically on the console, so that user like us can view the plot and understand the working of the decision tree.

Output:



The above is the plot for the decision tree model built using the Season attribute of Olympic dataset. Each node represents a feature and link represents a condition. Similarly, in the first condition of the above graph is year less than 1954 then it is Summer, else it is Winter. The recursive pattern of decision tree goes on till the classification is complete and results is obtained by the decision tree.

> index\_dt\_1 <- createDataPartition(athlete\_events$Season, p = 0.75, list = FALSE)

> train\_dt\_1 <- athlete\_events[index\_dt\_1,]

> test\_dt\_1 <- athlete\_events[-index\_dt\_1,]

In the above code we have divide the dataset into two parts. Train and test to predict the accuracy of the model that we have build and check how good our model works. Train contains 75% of the data and test contain remaining 25% of the data that is needed for prediction.

Also in the code below we built model for train data.

> dt\_model\_train <- rpart(Season~., data = train\_dt\_1)#train Model

> summary(dt\_model\_train)

> #predicting Values using test data

> pred\_class <- predict(dt\_model\_train, test\_dt\_1, type = "class")

> pred\_class[1:10]

3 7 13 17 31 39 45 57 64 65

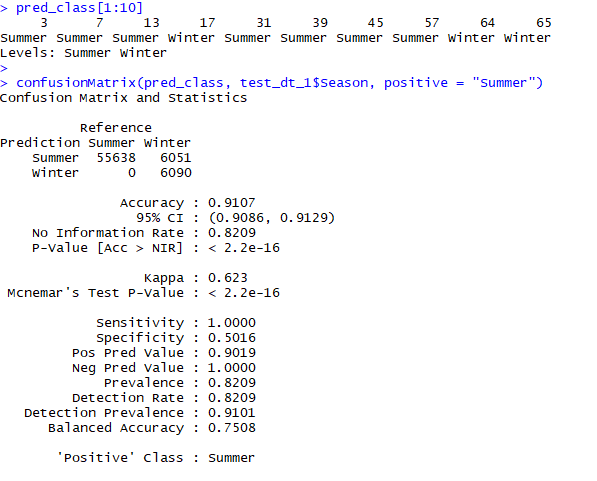
Summer Summer Summer Winter Summer Summer Summer Summer Winter Winter

Levels: Summer Winter

>

> confusionMatrix(pred\_class, test\_dt\_1$Season, positive = "Summer")

Output:



In the above code have performed prediction using the predict function on the train data decision tree model. Which is following up with a prediction. The confusion matrix gives the accuracy of 91% as it is given in the above output.

**Conclusion**

We could successfully clean, visualize and analyze the patterns from Olympic historical dataset that gave us the opportunity to demonstrate how the Olympic Games have developed over the years? How males and females have performed during summer and winter, the participation of various countries, growth or decline in their number of medals, aggregated value comparison of the age of males and females and so on. Using the train, test data and confusion matrix we could successfully predict the effect of season and our accuracy of the model is 91.2%. Also, this project gave us a chance to learn & practice data mining strategies to derive significant learning from the data. It provided us bits of knowledge related to the Olympics development timeline, participation, performance and participation of women, diverse nations, unique games and events.

**References**

#### Information about Olympic Games｜2020 Games Preparation | Bureau of Tokyo 2020 Olympic and Paralympic Games Preparation. 2020games.metro.tokyo.jp. Retrieved from <https://www.2020games.metro.tokyo.jp/eng/taikaijyunbi/olympic/index.html>

* 120 years of Olympic history: athletes and results. Kaggle.com. Retrieved from <https://www.kaggle.com/heesoo37/120-years-of-olympic-history-athletes-and-results/home>
* Olympics In-depth EDA | Kaggle. (2019). Kaggle.com. Retrieved from <https://www.kaggle.com/varunsukumar27/olympics-in-depth-eda>
* Holding an Olympic Games means evoking history | Kaggle. (2019). Kaggle.com. Retrieved from <https://www.kaggle.com/josephgpinto/holding-an-olympic-games-means-evoking-history>
* Easily Install and Load the 'Tidyverse'. (2019). Tidyverse.tidyverse.org. Retrieved 17 February 2019, from <https://tidyverse.tidyverse.org/>
* contributors, D. (2019). Aggregating and analyzing data with dplyr. Datacarpentry.org. Retrieved 17 February 2019, from <https://datacarpentry.org/R-genomics/04-dplyr.html>