*Olympic Data Analysis in R*

Project Description

In the final project, our team decided to use the Olympic Games dataset, which comes from Kaggle.com. Olympic Games are a world known sports game, and it is held in a interval of every four years. There are Summer Games, and Winter Games held independently, and there are over 200 nations participating in the events. The main goal of organizing Olympic games is to bring the countries together, grow as an individual through competitive games, and add to world harmony and educating the youth by participating in sports without any discrimination, solidarity and fair play. Olympic Games are too famous for leading nations to want to gain more medals to add light to their country. With the new innovations and improvements in the field of data analytics, it provides minute and important information of the teams and individuals to gain advantage over their rivals. Using data analysis teams can perform better and take decisions based on insights from historical data and certain pieces of evidence. With the growing technology and innovations, more options for detailed analysis are made possible, Hence, analysis has a high importance in sports. Sports has funny and nail-biting things of twisting and turning the expected into something that no one could ever imagine,

We will work on and analyze the historical Olympics data from Athens 1986 to Rio 2016 to determine the relationship between some valuable variables and the changes that happened in the past Olympics Games. We will use R Studio as the analytical tool to analyze our Olympics data. We will analyze data by the next few steps:

• Clean Olympics dataset by deleting all the NA data.

• Do basic data statistical analysis.

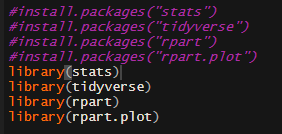
• Do data visualization on athletes participated number, top 10 nations ranking by winning medals, females and males’ medals winning counts, top 10 nations medals winning on basketball, and the athletes' physical conditions.

• Use the decision tree to forecast which season will gain a greater number of medals, summer games or winter games?

**Data Cleaning**

120 years of Olympic history from 1896 to 2016 is collected from [www.Kaggle.com](http://www.kaggle.com). Two datasets are used, they are “athlete\_events.csv” and “noc\_regions.csv”. We have used “noc\_regions” as it contains full forms of the abbreviations in the “athlete\_events.csv”.

We loaded the datasets and installed the required packages and libraries for this analysis. The installed the following packages and libraries:



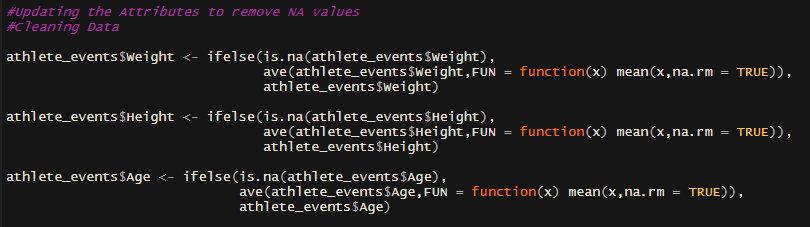
As we can see in the above image, we have installed a package called as “tidyverse”, this package consists of various core packages such as:

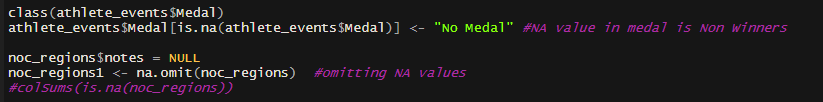
* Ggplot2: Used to create the graphics by mapping variables to the aesthetics.
* Dplyr: Used in manipulating data.
* Tidyr: Used in cleaning the data.
* Readr: Used to read the csv files
* Stringr: provides set of functions that makes working with strings as easy as possible.

Now we will start the analysis, first we will load the data.



After loading the data, we will clean the data by replacing the NA values with the mean average of that particular attribute or by deleting the column containing the NA values. We found out NA values in 3 columns they are Weightm, Height and Age column. We carried out the cleaning as followed:





We will check out the dimensions of the dataset by using dim() funcion:



Output:



From the above image we observed that the dataset contains 15 columns and 271116 rows.

Now we will find out the names of the 15 columns by:



Output:



After cleaning the data we will check if there are any other Null values left in the dataset:



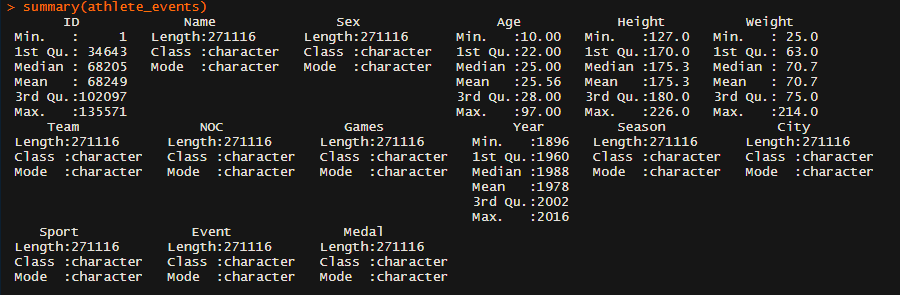
Output:



Now we will have a look at the summary of the dataset:

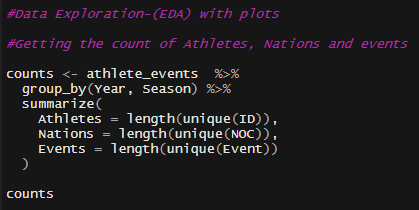


Output:

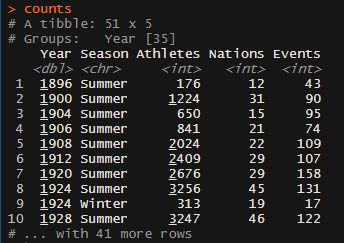


**Exploratory Data Analysis**

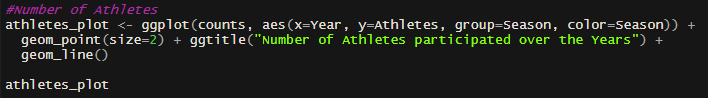
We will perform EDA (Exploratory Data Analysis), we will split the dataset into groups, and analyze the groups. Grouping can be done by using the group by() function:



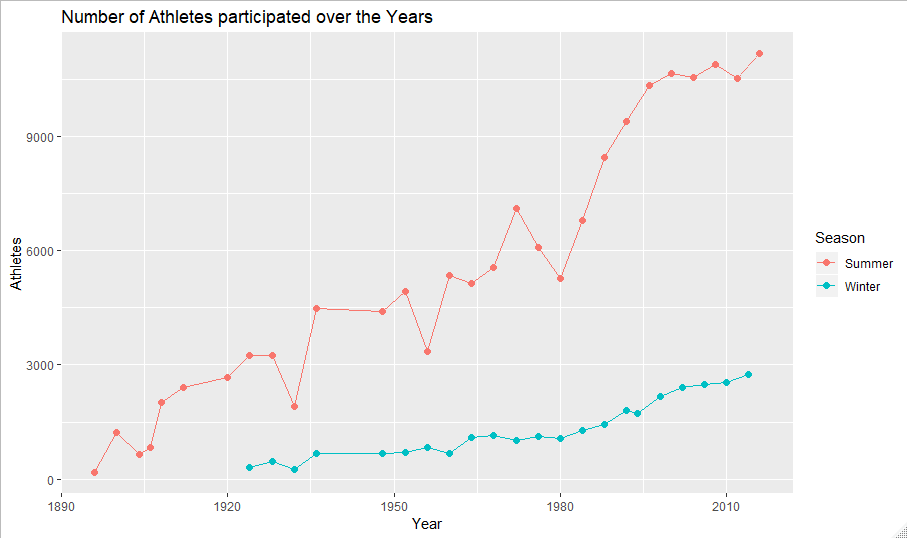
Output:



Now we will plot how many athletes participated in the olympics over the years:

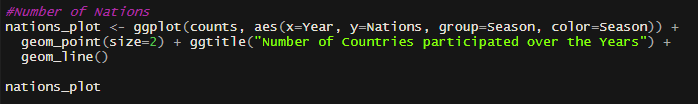


Output:

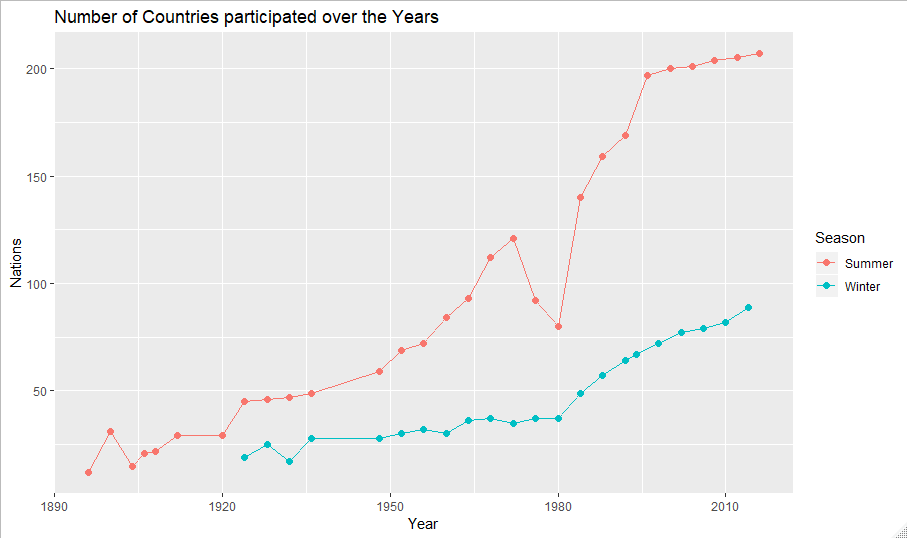


From the above line graph, we observed that, there has been a gradual increase in the participation of athletes from the year 1980 for the summer olympics. Whereas winter olympics has slight growth in the participation number as winter olympics are not popular in many countries.

Next, we tried to visualize if there was an increase in the number of countries participating in the olympics over the years:

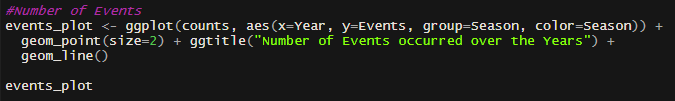


Output:

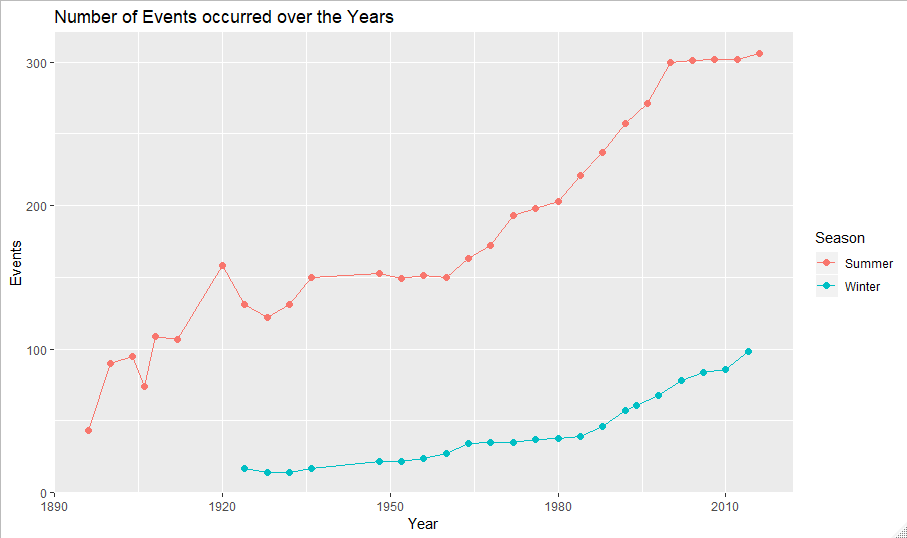


From the above image we observed that there was a dip in the participation in 1980, there was a fall in number of countries participated because around 66 countries had boycotted the tournaments including America.

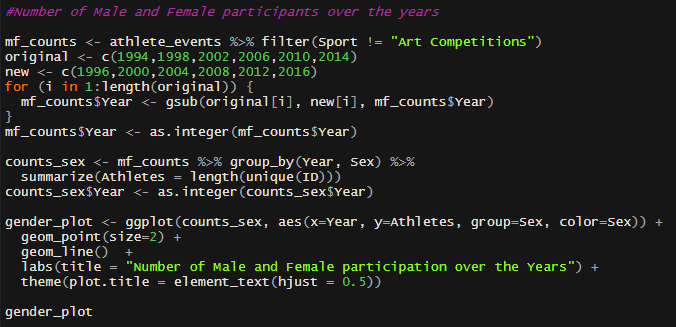
Further, we will see whether there was any increase in the number of events that occured over the years:



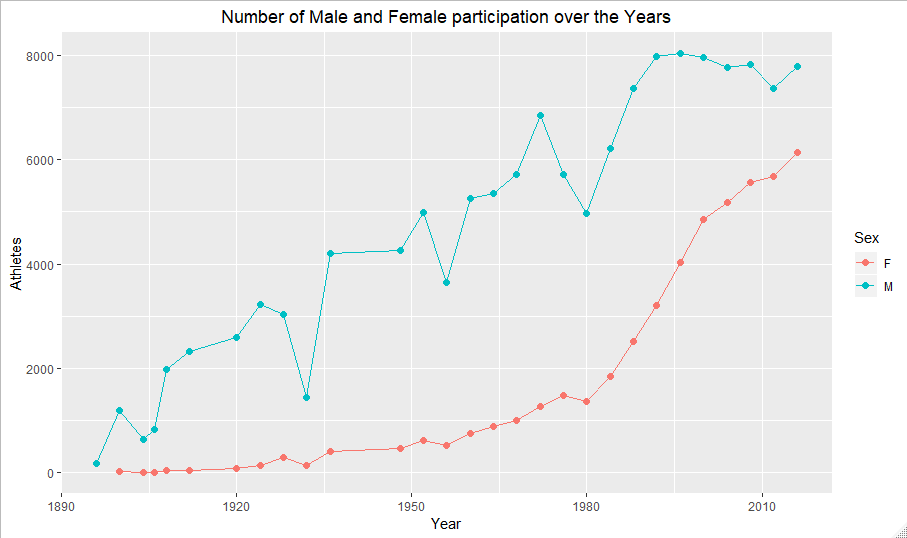
Output:



From the above image we observed that there has been steady increase in the number of sports events in the olympics over the year for the summers season as well as winter season. But Summer season leads the list as it has over 300 events by the year 2010.

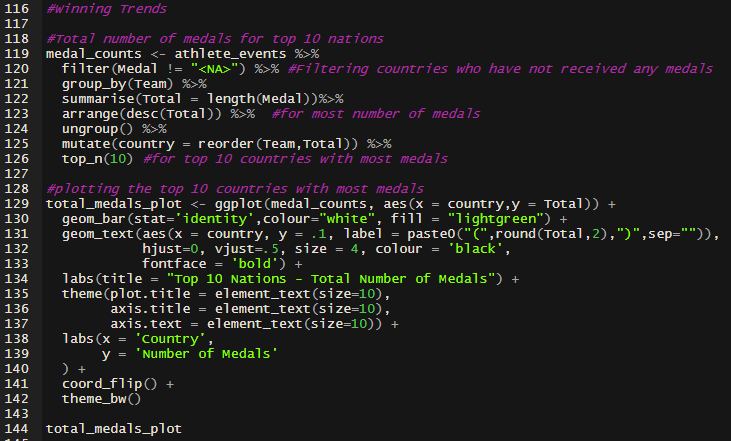
Next, we plot a graph comparing the number of male and female participants over the years and see the growth or fall in the participation with respect to the gender.

Output:

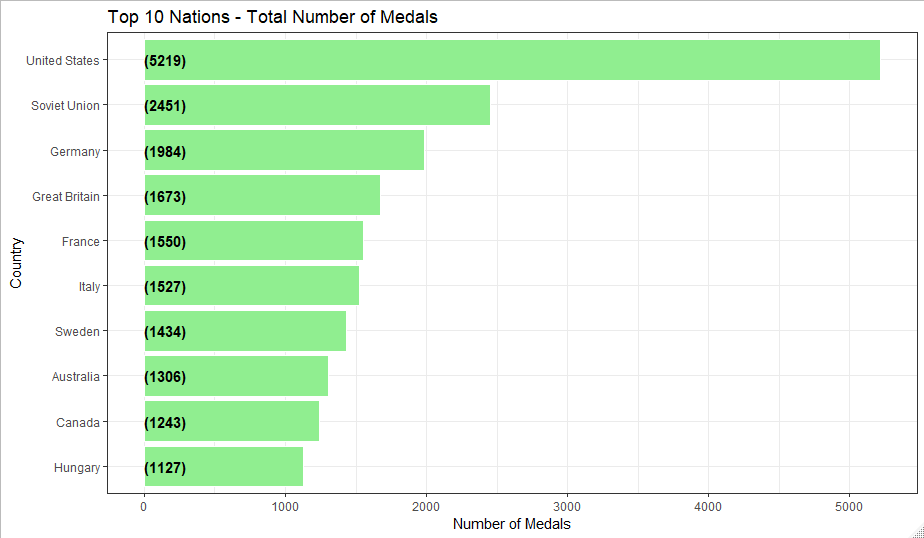


From the above graph, we can say that the male participation has always been higher compared to the females, but the female participation has an exponential growth and it is slowly catching up to the male participants.

Next, we find the nations with the most medals. As the plot would be very conjusted we would be focusing on the top 10 countries with most medals.

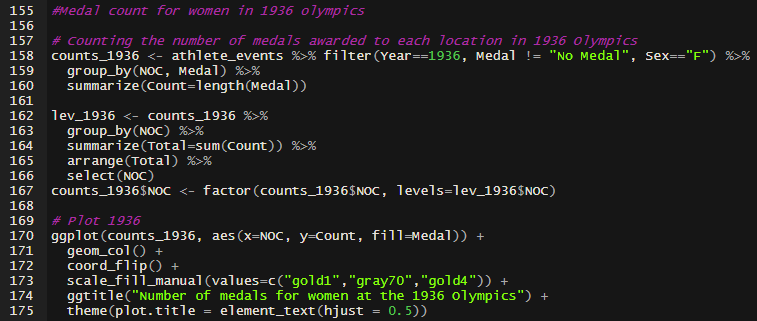


Output:

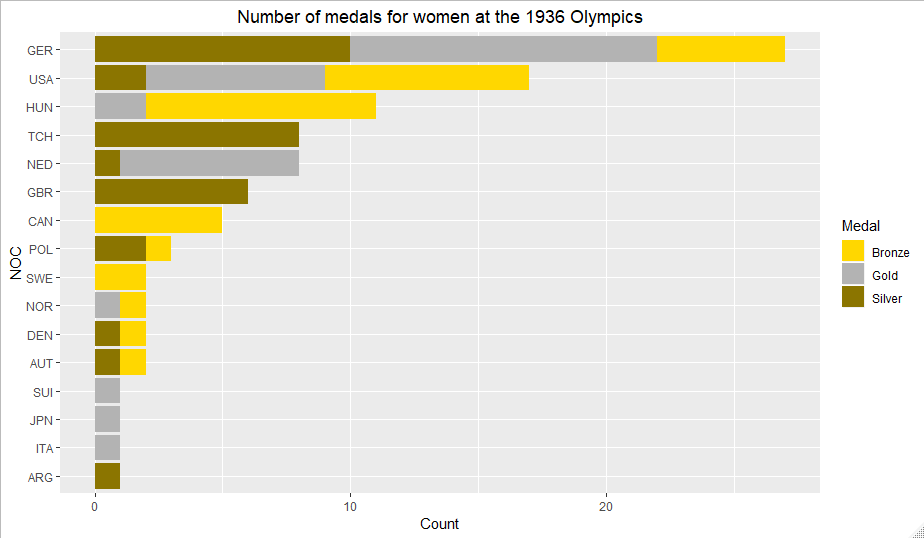


By looking We can say that United States has won the most medals with more than 5000 medals with Soviet Union at second and Germany at third and so on.

Now we would be comparing women medal winners in 1936 olympics to 2016 olympics winners along with their respective countries to determine women's participation in 1936 and in 2016.

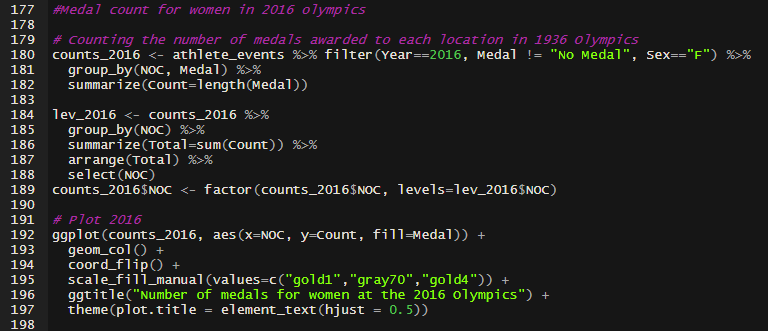


Output:

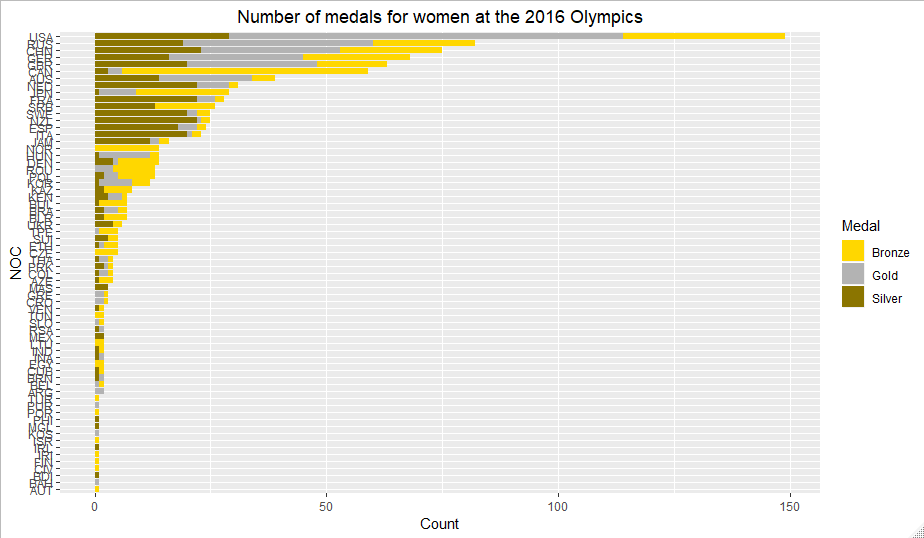


From the above plot we can tell that Germany had the greatest number of medal winners in women's category.

Now we compare the above results to the 2016 Olympics to see the changes in women winners and their respective representing nation.

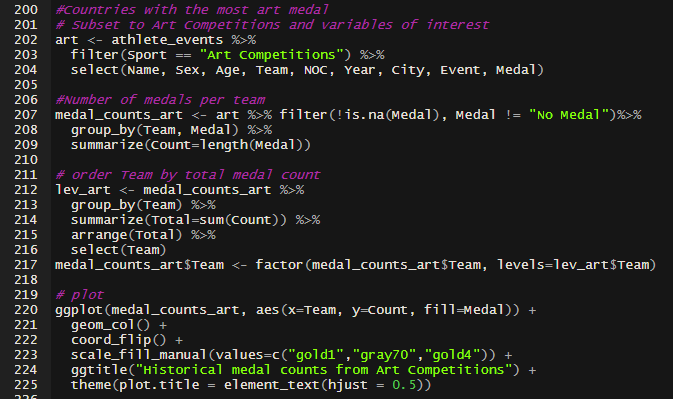


Output:

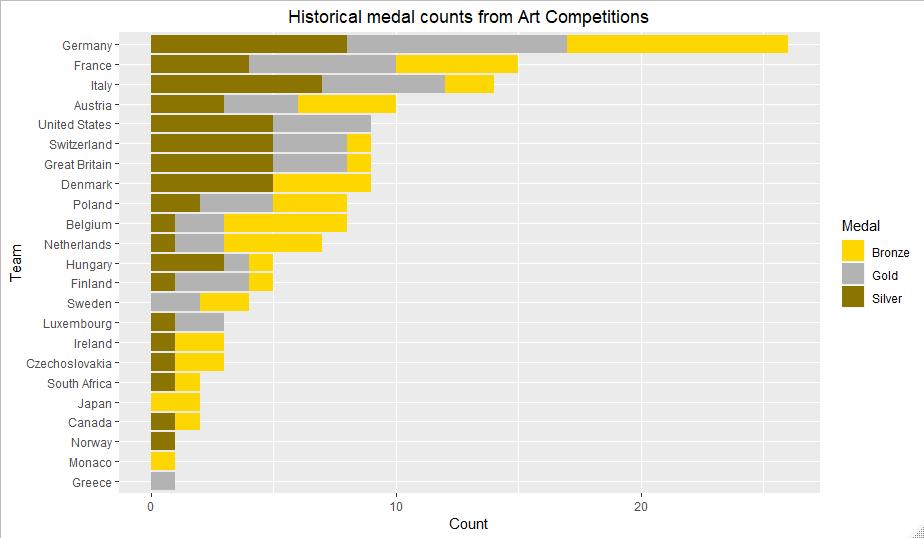


As seen in the above plot we can tell that women winners have increased drastically along with the increase in participants from more nations and we can also tell that USA took a major leap in women's category compared to Germany in 1936 Olympics.

Now we try to check the countries with the most art medal winners.

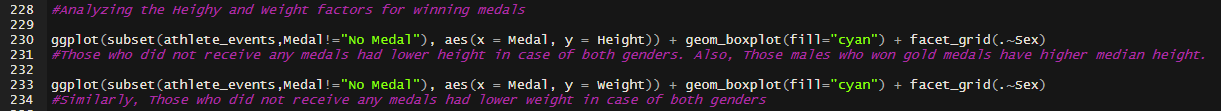


Output:

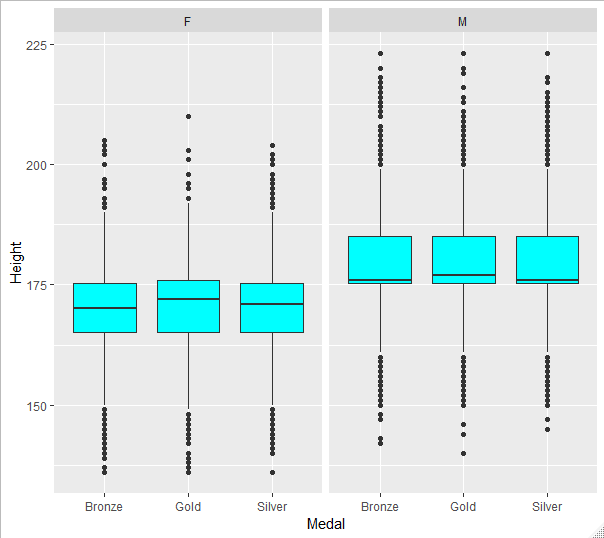


As seen from the above plot we can say that winners from Germany are leap ahead compared to other nations in Works of Art category.

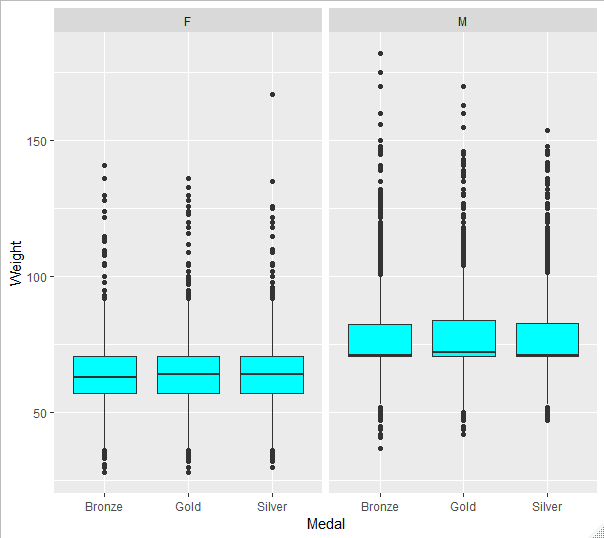
Now we try to compare winning characteristics of winners by looking at their body characteristics like height and weight, to be noted weight may play major role in discrepancies as some sports may require higher or lower weight depending on the game.



Output:



As seen from the above box plot we see that the gold medal winners in male category have slight higher height than compared to silver and bronze winners and looking at the female category we say the same that gold medal winners tend to have slightly higher height than compared to silver and bronze medal winners.



As mentioned before, weight characteristics of an athlete may vary in respect to specific game requirement so we can see there is miniscule difference between the gold, silver and bronze winners for both women and men.

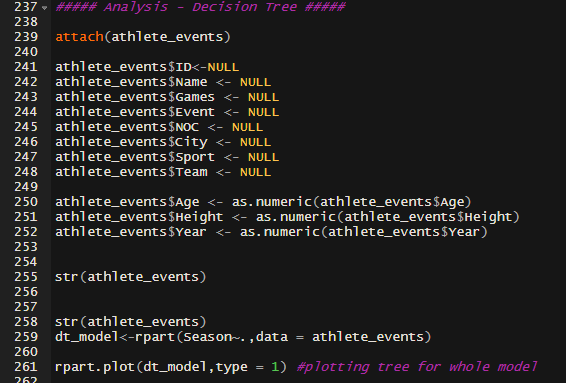
**Decision Tree Analysis:**

A decision tree is decision-based support model which use tree like model for decisions and its possible consequences. This model structure can help us in envisioning the forecast in an ordered tree fashion. This tree can assist us in visualizing the data and help us in checking the probability of an event occurrence based on the outcome’s outputs in every branch.

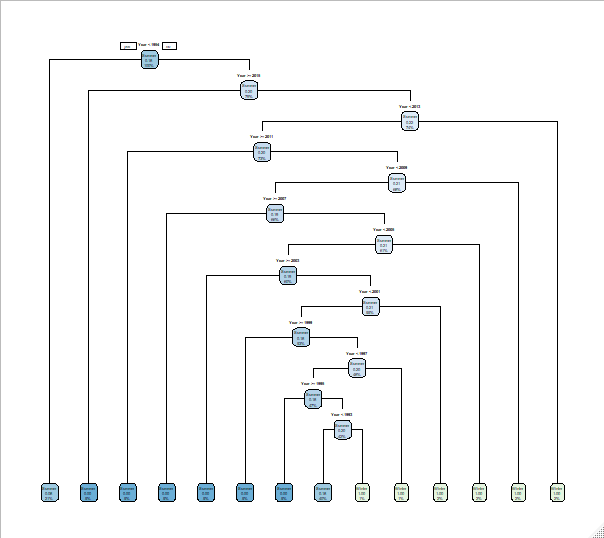
R- code:

We will be using the datasets columns and will be assigning them a NULL value as they don’t have a requirement for the classification in the decision tree.

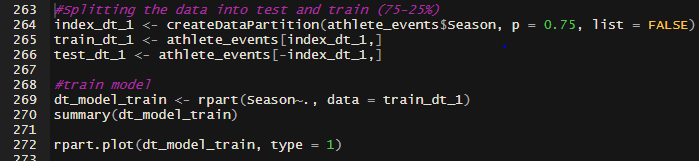
And we will be converting the required variable to numeric to simplify the values to fit in the model.



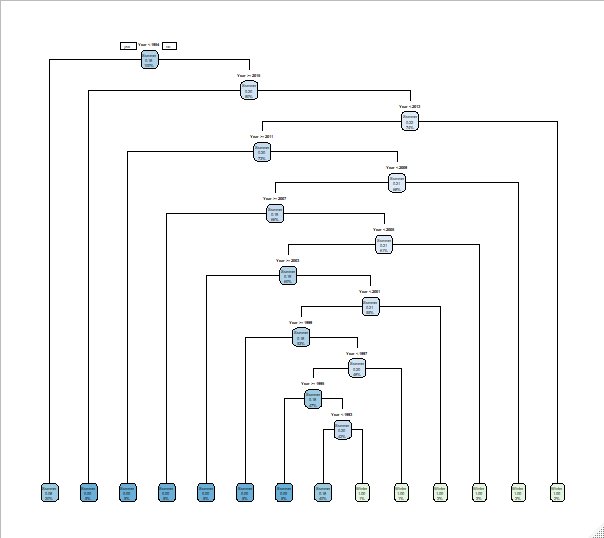
Output:



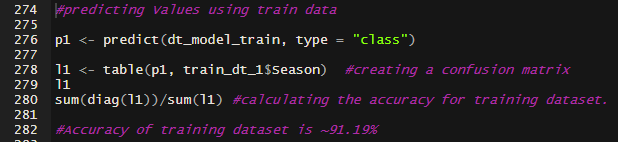
The above plot of decision tree model is built with season attribute from Olympic dataset. Each node represents a link with a condition. As seen above till year 1954 the winter games get less chosen game compared to summer from later on.

After this we created the train and test dataset for our season prediction.

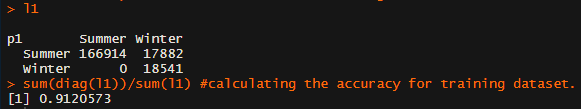
Output:



Now making the confusion matrix and calculating the accuracy of training dataset.

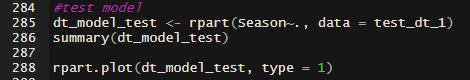


Output:

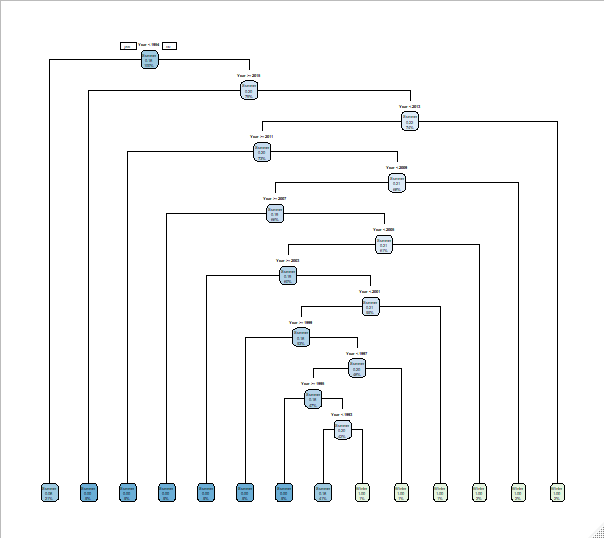


The accuracy of training model is nearly 91%

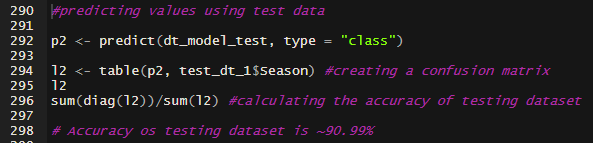
Doing the same for the test dataset, we get



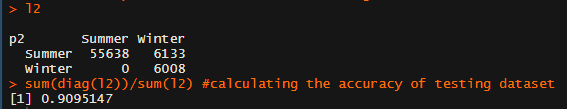
Output:



Now we create a similar confusion matrix and calculating the accuracy of test dataset.



Output:



The accuracy of test dataset is also close to 91%.

**Conclusion:**

We could effectively clean, visualize and analyze the patterns from Olympics Historical dataset that allowed us the chance to show how the Olympic Games have developed throughout the years? How males and females have performed during summer and winter, the participation increase of different nations, increase or decrease in their number of winners, totaled value correlation of the age of males and females, and so on. Utilizing the train, test information and confusion matrix we could effectively anticipate the impact of season and our precision of the model is close to 91%. Likewise, this project allowed us to learn and implement data mining techniques to get critical insights from the dataset. It gave us bits of information identified with the Olympics advancement course of events, participation, performance and increase in participation of females, increase in participation of various countries and overall participation over diverse games.

References:

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>

Easily Install and Load the 'Tidyverse'. (2019). Tidyverse.tidyverse.org. Retrieved 17 February 2019, from <https://tidyverse.tidyverse.org/>

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>