**Study of 120 years of Olympic history: Athletes and Results**

CA2 submitted in part fulfilment of the requirements for the degree of

MSc in Data Science (Computer Science)

at Technological University Dublin

Logo

Description automatically generated

Shubham Jangir

2020-2021 STUDENT NUMBER: D20124818

20th Dec’20 COURSE: TU059(FULL-TIME)

# Dataset:

# Data Source: This analysis is based on the dataset contains 120 years of Olympics History i.e., Athens 1896 to Rio 2016: Athletes and Result data

# There are two datasets:

* <https://www.kaggle.com/heesoo37/120-years-of-olympic-history-athletes-and-results?select=athlete_events.csv>
* <https://www.kaggle.com/heesoo37/120-years-of-olympic-history-athletes-and-results?select=noc_regions.csv>

The above datasets are taken from the Kaggle for 120 years of Olympics: Athletes

Results

* This is a historical dataset, covering all the Games from Athens 1896 to Rio 2016, about the current Olympic Games.
* Given the Historical Evidence for the Olympics , the odds for winning a medal (Gold, Silver, or Bronze) may be likely based on a few biological attributes of the athlete.
* There are two files one for the Athlete, Events, Medals, Biological Attributes, Season, Sport, Year, Name of the Athlete.

The other file contains the NOC (National Olympic Committee) and Regions.

# Data Import: Check the working directory and set up the working directory as per the location of the file and load the file using read\_csv.

# athlete\_events and noc\_regions will be loaded with 271116 obs. of 15 variables and 230 obs. of 3 variables respectively(1.1).

# Graphical user interface, text, application, chat or text message Description automatically generated

# Dataset Merging:

# Merge the two datasets to produce the final data frame using left join.

# Total number of records: There are 271116 data points (rows).

# Number of fields: 17 features (columns) describing mainly Athletes, Event, Medal, Biological Attributes, Sport, NOC, Regions(1.2).

# Graphical user interface, text, application Description automatically generated

# Describing and Exploring the Dataset

# Total number of records: There are 271116 data points (rows).

# Number of fields: 17 features (columns) describing mainly Athletes, Event, Medal, Biological Attributes, Sport, NOC, Regions.

# Head and tail command will show the top and bottom 5 rows of the dataset for each column so we can understand the dataset by looking at the few rows of the dataset.

# Using the summary command, we can explore the statistical summary of the numerical variable which will help us to find out the mean, median and max values in the dataset(1.3).

# Graphical user interface, text, application Description automatically generated

# Cleaning the Dataset:

# Duplicate rows/data:

# Checking the duplicate rows and removing them for our analysis(1.4).

# Graphical user interface, text, application, chat or text message Description automatically generated

# Missing/NA data:

# Checking the columns with NA or missing data and handle the missing data(1.5).

# A picture containing chart Description automatically generated

# There are missing values in multiple columns, let’s handle each of them separately.

# Medal Field: As most of the values are NA in this column because every athlete doesn’t win the medal, so we will assign ‘Did not win’ for NA values in the medal field.

# 

# Region Field: There are very fewer missing values in region field which can be filled with the Team field as both have the country name data.

# 

# Age, Height, Weight Fields: All these fields contains the missing values which can be filled with median of their respective fields group by Sports and Age as different sports have different Age, Height and Weight requirement.

# Graphical user interface, text, application Description automatically generated

# After managing most of the NA values we are still left with 0.5% missing values

# for Height and Weight which can be removed for our analysis.

# 

# Text Description automatically generated

|  |
| --- |
| Notes Field: This field contains 96-97% missing values, so we can go aheadand remove this column for our analysis because the values which are presentin this field also have the Team and Country details which is already present.After removing all the duplicate and missing values we are left with:Datatype Changing:Changing the datatype of the columns as per the analysis because most of columnswere considered as character instead of being the factor or categorical variable(1.6).Text  Description automatically generatedFinal Modified dataset/data frame after cleaning using R:After all the pre-processing and cleaning below is our final data frame for the analysis:Total number of records: There are 267,632 data points (rows).Number of fields: 16 features (columns) describing mainly Athletes, Event, Medal, Biological Attributes, Sport, NOC, Regions.A picture containing text  Description automatically generated\*Modified dataset attached in Brightspace |
|  |

# User Story

# A flag flying in the sky Description automatically generated with medium confidence

The Olympic Games are a well-known athletic platform, distinguished from the late 19th century, which is recognized all over the world. However, its roots can be traced back to the Greek empire, which consisted only of sprint race about 3,000 years ago and was kept only accessible to freeborn Greek citizens in Greece's city of Olympia. Since then, it has expanded and has now become a hub for all athletes worldwide to display their skill in more than 28 individual sporting competitions. It is usually held in various countries every 2 years, with Summer Olympics and Winter Olympics titles, each with their own collection of games. It has become a place that represents the strength of the individual athletes and has become a source of pride for the countries they represent.

At every Olympic Games, the medal count, Athlete performance(Men & Women), Top counties in regards with the highest number of medals, number of Athlete participation make big news before the start of the Olympic.

Various media channel, magazines and different website start showing the interactive charts and visualization for the increase in TRP and views.

As a data scientist, we are building this report and interactive charts for the BBC news channel and website before the Tokyo Olympic starts to get the understanding about all the previous Olympic data in form of interactive chart, visualizations and animations so that it can be shown over their website and can be accessible to all the users in the country.

# Data Wrangling and Visualizations /Insights:

# Process :

# Selecting the Topic 🡪 Data Collection 🡪 Data Cleaning 🡪 Defining User Story 🡪

# Data Wrangling 🡪 Visualization/Insights 🡪 Styling the Visualization

# In this section we will do Data Wrangling using Visualization in R we will answer our Question and Style our Visualization to give the clear picture to our User.

# 

# Insights from the data:

# Insight-1

# Question 1: Being called as “impractical, unaesthetic and incorrect” the idea of women participating in Olympic during the starting of Olympics. How are women performing in the Olympics in terms of:

# a)Athlete Numbers

# b)Total medal comparing 1936,1976 and 2016

# Note: There are two parts(a&b) for this Question which is shown differently in R-markdown notebook in the name of Insight-1a and Insight-2b(2.1 & 2.2) but for explanation purpose it’s been included in one part here.

# Data Wrangling:

# a) After 1992, Olympics were organized in every 2 year i.e., Winter and Summer for our analysis we are replacing Winter Olympic with its next Summer Olympic e.g., 1994Winter to 1996Summer.

# Counting Numbers of Male and Female Athletes over the years

# Graphical user interface, text Description automatically generated

# b) Calculating Total number of medals in Berlin 1936,1976 and 2016 for Female Athlete

# Graphical user interface, text, application Description automatically generated

# Visualization:

# a)Athlete number -As shown in the Figure1(a) in 1920 and before there was less or no participation of women athletes, but it has grown drastically after 1950 and 1980 and crossed 6000 in 2010 while the number of male athletes leveled off at 8000 but the female athlete continued to grow at higher rate. The highest participation was during the Rio 2016 Olympics this shows women are women are performing equivalent to men in terms of athlete numbers which shows active participation.

# Chart, line chart Description automatically generated

# Figure1(a)

# Extra Feature: Animation has been used with the increasing number of Years Athlete count for both male and female increases as shown in figure.

# b)Total Medal Won by women in 1936, 1976 and 2016 Olympics- As shown in the Figure1(b), (c), (d) the medal count for the female athletes have increased drastically before there was hardly 20 medals secured in 1936 but in recent Rio 2016 Olympics women have managed to secure more than 100+ medals by top nations which is good sign of active participation of the women athlete and medal count has increased over the period of time. The below graph shows the Countries with their Gold, Silver and Bronze medal count for the year 1936,1976 and 2016 if you hover over the graph.

# Chart, bar chart Description automatically generated

# Figure1(b)

# Chart, bar chart Description automatically generated

# Figure1(c)

# Chart, bar chart Description automatically generated

# Figure1(d)

# Extra Feature: Plotly's R graphing library makes interactive, publication-quality graphs. As shown in the figure1(b, c and d) you can hover over the chart and get all the Medal count for that particular country. There are lot of features including zoom in and out, reset axes, box select, auto scale and many more.

# Insight-2

# Question 2: What is the distribution of Total Medals all over the world using the World till 2016.

# Data Wrangling:

# Filtering out National Olympic Committee who won the Medals

# Taking the count of Medals by NOC and Event

# Taking the total of Medals as per Region/Countries and arranging by Descending order

# Loading the World Map

# Merging World Map and Total\_Medal\_Countries using left join

# Graphical user interface, text, application, email Description automatically generated

# Visualization:

# It is evident from the graph that North America particularly USA and Alaska (USA) have the highest number of medals after that its Russia. In European nations Germany have the highest number of medals followed by France, UK and Norway. In Asia, China has the highest number of medals if you observe closely these are the countries with highest GDP’s which contribute to better Sport facilities and participation.

# Map Description automatically generated

# Figure2(a)

# Extra Feature: World map has been used to show the countries with higher number of total medals.

# Plotly's R graphing library makes interactive, publication-quality graphs. As shown in the figure2(a) you can hover over the chart and get all the Medal count for that particular country. Total shows the number of medal and region shows the country.

# Insight-3

# Question 3: What are Top 5 countries which are performing best in terms of Highest Number of Gold, Silver and Bronze Medal.

# Data Wrangling:

# Gold, Silver and Bronze Medal group by Team, NOC, Medal, Event and Games.

# Gold, Silver and Bronze Medal Count by Team, NOC, Medal and Event.

# Merging Gold, Silver and Bronze Medal with Regions/ Countries.

# Taking the sum of Gold, Silver and Bronze Medal by Countries.

# Top 5 Countries with Gold, Silver and Bronze Medals.

# Merging Top 5 countries with Gold, Silver and Bronze Medal.

# The below screenshot in shown for Bronze medal for reference.

# Graphical user interface, text, application, email Description automatically generated

# Visualization:

# The below graph shows the top five nations with highest number of Gold, Silver and Bronze. USA has won the highest number of Gold medals and it is evident from the graph that these countries have a huge sports culture and when it is combined with money produces some of the best programs and in return higher Olympic Medals.

# Chart, bar chart Description automatically generated

# Figure3(a)

# Extra Feature: Animation has been used with the increasing number of Medal count for all the countries as shown in figure.

# Reorder has been used to order the medal count from descending order.

# Coordinate flip has been used to display the medal count.

# Insight-4

# Question 4: What is the variation of Height and Weight in Male and Female Athlete over the period of years.

# At what age does an athlete win the highest number of medals over the period of years

# Data Wrangling:

# Filter out year for Weight and Height category from 1906 to 2016.

# Run the sql file to the Variation in age for medals.

# 

# Visualization:

# It is evident from the below figure that Weight for the male athlete have increased and reduced for female athlete over the period of year(1906-2016).

# It is evident form the below figure that Height for the male athlete have increased and reduced for female athlete over the period of year(1906-2016).

# There is a relationship between height and weight, for male height and weight both increased and for female height and weight both decreased over the period of year(1906-2016).

# It is evident from the graph that there are more chances to win medal in the age group of 20-30yr.

# Chart Description automatically generated

# Figure4(a)

# A picture containing text, writing implement Description automatically generated

# Figure4(b)

# Chart, scatter chart Description automatically generated

# Figure4(c)

# Extra Feature: Plotly has been used for variation of age to find out the total number of medals won at that particular age.

# Boxplot have been introduced for Height and Weight to show the variation over the year and axis labels were made vertical using element text and set the angle to 90.

# Insight-5

# Question 5: What is the variation in total number of Athletes from 1936-1976-2016 using World Map.

# Data Wrangling:

# Calculating total number of athletes in 1936,1976 and 2016 Olympic Games and map it with the World map.

# Graphical user interface, text, application Description automatically generated

# Visualization:

# The below figures show the number of athletes for 1936,1976 and 2016. In 1936 there was not much participation from the Soviet Union, South America and Asian countries but over the year the number of athletes has drastically increased but in African and Middle east countries there is still less participation its due to the economic conditions and Political tension which is not letting Athletes to grow in numbers but let’s see in the upcoming Olympics if the situation changes and there is lot of participation from all over the world but overall participation have increased in the Olympics.

# Map Description automatically generated

# Figure5(a)

# Map Description automatically generated

# Figure5(b)

# Map Description automatically generated

# Figure5(c)

# Extra Feature: World map has been used to show the countries with number of Athletes if you hover over the country it will show the count and country; region signifies the location of the Olympic.

# Plotly R graphing library makes interactive, publication-quality graphs. As shown in the figure5(a, b &c) you can hover over the chart and get all the Athlete count with their respective countries.

# Insight-6(just an interesting insight while I was reading about Olympics)

# Question 6: What happened at the 1936 Olympics, that was the Nazi's era and Olympics held in Capital city Berlin.

# Data Wrangling:

# Calculating the number of athletes in 1936 Olympics using the world map and filtering the top ten nations.

# Text Description automatically generated

# Visualization:

# It is evident from the graph that Germany won the max number of medals, Hitler saw the 1936 Games as an opportunity to promote his promoted an image of a new, strong, and united Germany while masking the regime's targeting of Jews and Roma. USA boycotted the 1936 Olympics and most of the athletes haven’t participated in the game.

# Graphical user interface, application, table Description automatically generated

# Extra Feature: geom\_label is used to display the number of medals in front of their respective countries.

# Iterations:

# To reach the final plot we had go through various iterations and improvisation which is shown below:

# Iteration for Insight-1a:

# Chart, scatter chart Description automatically generated

# This iteration is for the Insight-1a before using geom\_label and geom\_point together and merging Olympics from 1992 because after1992, Olympics were organized in every 2 years i.e., Winter and Summer for our analysis we are replacing Winter Olympic with its next Summer Olympic e.g., 1994Winter to 1996Summer.

# Iteration for Insight-1b:

# Chart, bar chart Description automatically generated

# This iteration is for the Insight-1b before using the proper colour for the medal as they don’t signify anything to the user bar chart should be in the same colour of the real medals so that it can be easy to read for the user that’s why we used gold silver and bronze colour in our final analysis.

# The bar chart is not organized in the correct order that makes it difficult to read.

# Iteration for Insight-2:

# Map Description automatically generated

# This iteration is for the Insight-2 before using the Plotly it is difficult for the user to identify the name of the country if they aren’t aware of it so we used Plotly which is an interactive library and once you hover over the graph you can see the country and the total number of medals.

# Iteration for Insight-3:

# Chart, bar chart Description automatically generated

# This iteration is for the Insight-3 before using the proper colour for the medal as they don’t signify anything to the user bar chart should be in the same colour of the real medals so that it can be easy to read for the user that’s why we used gold silver and bronze colour in our final analysis.

# The bar chart is not organized in the correct order that makes it difficult to read.

# Iteration for Insight-4:

# Chart, box and whisker chart Description automatically generated

# Chart, box and whisker chart Description automatically generated

# This iteration is for the Insight-4 before using the as.factor for the year as it was aggerating all the data for the different year and showing only two box plot and if we use as.factor also year on the x-axis was not vising because it was displayed horizontally but with the help of element\_text and using angle 90, year axis label was displayed vertically which makes it easy to read.

# Iteration for Insight-5:

# Map Description automatically generated

# This iteration is for the Insight-5 before using the Plotly it is difficult for the user to identify the name of the country if they aren’t aware of it so we used Plotly which is an interactive library and once you hover over the graph you can see the country name and the total number of athletes we have changed the colour of the graph as well because using the default colour it was getting difficult to identify the countries because the colour is too dark and the light colour was signifying the more number of athletes which so we changed to a red and darker shade of red will display the more number of athletes.

# Technologies Used:

# We have created the visualization using:

# RStudio Version 1.3.1093: RStudio is an integrated development environment for R, a programming language for statistical computing and graphics.