TCET
DEPARTMENT OF COMPUTER ENGINEERING (COMP)
(Accredited by NBA for 3 years, 3rd Cycle Accreditation w.e.f. 1st July 2019)
Choice Based Credit Grading Scheme (CBCGS)
Under TCET Autonomy

Analysing Evolution of the Olympics by Exploratory Data Analysis using Python

By

Group No.

(Sonali Rasal, BECOMP-C, 02)

(Pranjal Singh, BECOMP-C, 24)

(Aroma Sinha, BECOMP-C, 30)

Under the Guidance of

Mr. Aniket Mishra
Assistant Professor

for the subject
Data Analytics

In

B.E. COMPUTER ENGINEERING

(Academic Year: 2022-23)

TCET

DEPARTMENT OF COMPUTER ENGINEERING (COMP)

(Accredited by NBA for 3 years, 3rd Cycle Accreditation w.e.f. 1st July 2019)

Choice Based Credit Grading Scheme (CBCGS)

Under TCET Autonomy

CERTIFICATE

This is to certify that

Authors

Ms. Sonali Rasal, Mr. Pranjal Singh, Ms. Aroma Sinha

Have satisfactorily completed the requirements of the B.E Capstone Project Report

On

"Analysing Evolution of the Olympics by Exploratory Data Analysis using Python"

Mr. Aniket Mishra	Dr. Harshali Patil
Subject In-charge	**HOD COMP**

Examiners

1. Signature: …….
   Name:

2. Signature: …….
   Name:

Date:

Place: Mumbai

CONTENTS

List of Figures

1. INTRODUCTION

Olympics games are considered as one of the most prime event which provides a valid and common platform for players across different countries to show their talent and skills. Modern Olympic Games were originated by taking inspiration from Ancient Olympic Games held in Olympia, Greece from the 8th Century BC to the 4th Century AD [1]. The following timeline outlines the main events in the history of Modern Olympic Games. The Olympics consists of various games (Approximately 45) in which players from various countries (Approx 205) participate to win a medal for their country. Olympics has a great history of evolution. From 13 participating nations in 1st Olympics (1896) to 207 participating nations in 2016 Rio Olympics, the Olympics have come across a long way. There are various scenarios which comes in our mind when we look into Evolution of Olympic Games over the years.

These scenarios are: Increase in number of participating nations, Increase in number of participating Athletes, Increase/Decrease in number of events, Increase in the expenditure cost of the event, improvement in performance of particular country, improvement in performance of a particular player, Increase in women participation, Participation Ratio of Men to Women, improvement in medication facilities during competition, the effect of pandemic (if any) on the performance of the players. Analysis over these scenarios would depict the evolution of the Olympics over the years. This analysis would help in the future prediction of the number of participating countries, players; winners of various games; Women participation and many more. These type of Analysis can also serve as a performance indicator of a particular country or Player. The main objective of this study is to analyze the various factors mentioned above which plays a vital role in the evolution of Olympic Games over the years. The Analysis will include the visualization and explanation of the change in trends of the various factors over the years which will help to predict the information of future Olympic Games. As Olympic Games are one of the most important sporting event across the world, each country and each player tries to give their best performance in the event. In order to improve their performance, every country should perform such Analysis which would help them in the improvement of their policies and strategies by providing current statistics to them.

2. PROBLEM DEFINITION

The main objective of this study is to analyse the various factors mentioned above which plays a vital role in the evolution of the Olympic Games over the years. The Analysis will include the visualisation and explanation of the change in trends of the various factors over the years which will help to predict the information of future Olympic Games. As the Olympic Games are one of the most important sporting events across the world, each country and each player tries to give their best performance in the event. To improve their performance, every country should perform such an Analysis which would help them in the improvement of their policies and strategies by providing current statistics to them. Building Olympic Data Analysis Web Application for analyzing data over the years which helps athletes in widening their scope of winning a medal and might be useful for further predictions.

3. TECHNOLOGY USED

Language : Python

Libraries : NumPy, Pandas, Seaborn, Matplotlib

IDE : Google Collab

Other Software: Microsoft Excel

Data : Kaggle/Book-Crossing Dataset

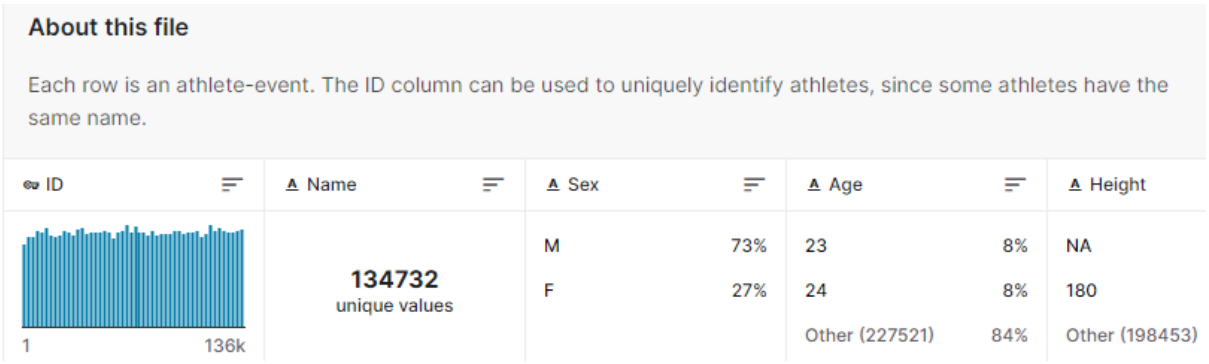
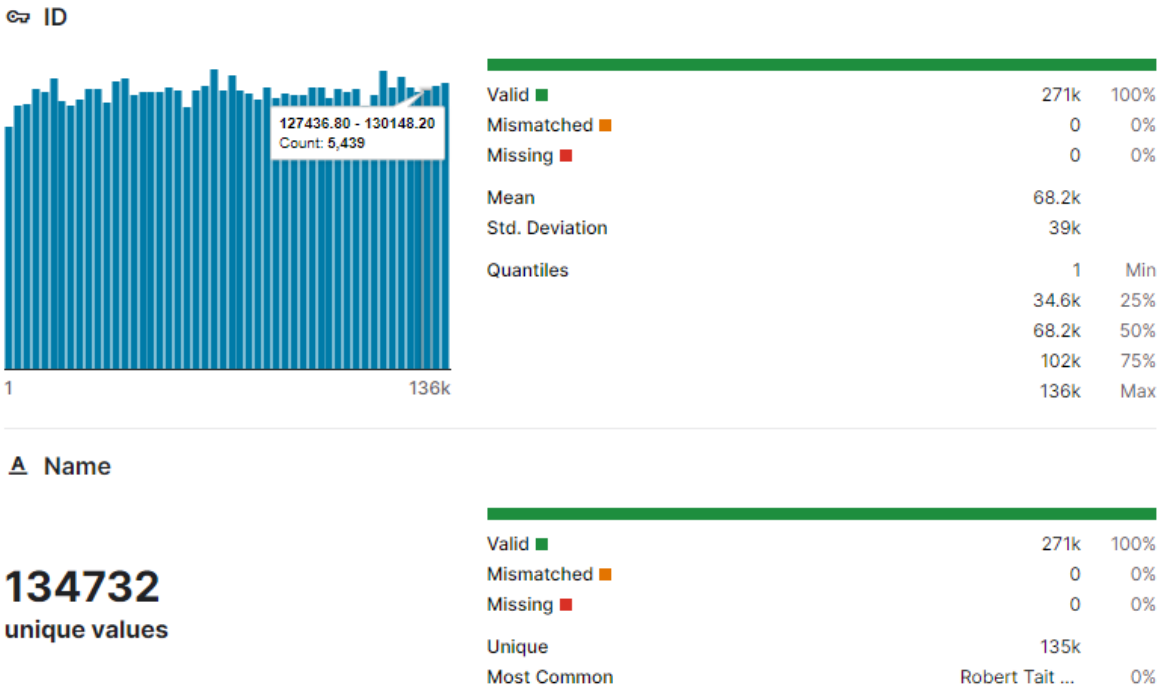


Fig. 1 - Statistics



3. IMPLEMENTATION

Data Collection The very first step of any type of Analysis, whether it is technical or non-technical, is Data Collection. In order to perform analysis on a certain problem, we require a large amount of Data on which we apply various techniques and algorithms to reach to a particular conclusion and get our desired result. It is advised to take the data in abundance because larger the volume of data for analysis, the greater would be the accuracy in the result and also the greater would be the confidence in decision making based on these results.

We have used data from various data sources for analysis on Evolution of the Olympics over the time. We have taken two datasets which provide us with large volume and a large variety of data for Analysis. 1st dataset consists the information about the players and their entire details like their Gender, Height, Weight, Country for which they play, Medals won (Gold, Silver and Bronze) and many more. This data can be used to analyze the performance of the particular player and can also help in the comparative study between two or more players. 2nd dataset consists the information of the countries and their NOC .

The Dataset consists of various fields like Age, Gender, etc which consists of some null values which produces errors in the end result which is the Visualization of data in graphical format. These null values are needed to be omitted or replaced with some valid value which solves the error and generates accurate result. We have used a technique known as Deterministic Imputation to complete this task. Deterministic Imputation is a situation where the null values (NA or NaN) are determined with the help of the other values in the same column in the dataset. For this purpose, there are various models such as Basic Numeric Imputation Model in which the null value is replaced by Mean or Median of other values of the same column of the dataset.

Exploratory Data Analysis The next step after data pre-processing is data analysis. In this step, analysis is done on data using various Techniques like Text Analysis, Diagnostic Analysis, Exploratory Data Analysis, etc and Machine learning Algorithms like Linear Regression, Logistic Regression, SVM, Decision Tree etc to reach to a particular conclusion.We are using the Exploratory Data Analysis technique to complete this task. Exploratory Data Analysis (EDA) is an approach to analyze data thoroughly and encapsulate its primary attributes basically in visual format.

Code:

```
import pandas as pd
import numpy as np
import plotly.express as px
import plotly.graph_objects as go
from plotly.subplots import make_subplots
```

```
url1 = 'https://raw.githubusercontent.com/AromaSinha/Olympics/master/athlete_events.csv'
url2='https://raw.githubusercontent.com/AromaSinha/Olympics/master/noc_regions.csv'
df = pd.read_csv(url1)
#df = pd.read_csv('athlete_events.csv')
region=pd.read_csv(url2)
df.head()
```

Fig. 2 - Imports

```
df["Age"].fillna((df["Age"].mean()), inplace = True)
df["Height"].fillna((df["Height"].mean()), inplace = True)
df["Weight"].fillna((df["Weight"].mean()), inplace = True)
```

```
df['Medal'].fillna('No Medal', inplace = True)
df.head()
```

```
df.isnull().sum()
```

```
#Finding out the cities that have hosted Games
#grouping by the cities based on unique year values and later we will be sorting the data based on
#no. of occurance of each cities which hosted based on occurance

City = df.groupby('City').apply(lambda x:x['Year'].unique()).to_frame().reset_index()
City.columns=['City','Years']
City['Occurence']=[len(c) for c in City['Years']]
City.sort_values('Occurence',ascending=False)
```

```
#Finding out participation of Men and Women at the Olympic games

print('Total number of athletes in Olymics:',len(df.ID.unique()))
print('Number of female participants in 120 years:',len(df[df.Sex=='F']))
print('Number of male participants in 120 years:',len(df[df.Sex=='M']))
```

```
df_summer = df[df["Season"] == "Summer"]
df_summer.head()
```

```
x = Trend["Year"]
y1 = Trend["F"]
y2 = Trend["M"]

fig = make_subplots(specs=[[{"secondary_y": True}]])
fig.add_trace(go.Scatter(x = x, y = y1, mode = "lines+markers", name = "Female",
                         line=dict(color='Blue', width=2)), secondary_y=False,)
fig.add_trace(go.Scatter(x = x, y = y2, mode = "lines+markers", name = "Male",
                         line=dict(color='Orange', width=2)), secondary_y=True,)

# Add figure title
fig.update_layout(
    title_text="Number of men and women athelete over time"
)

# Set x-axis title
fig.update_layout(title="Variation in count of male and female players",
                  xaxis_title = "Year")

# Set y-axes titles
fig.update_yaxes(title_text="Female", secondary_y=False)
fig.update_yaxes(title_text="Male", secondary_y=True)
fig.show()
```

```
fig = px.histogram(df, x="Season", color="Sex", barmode = "group",
             color_discrete_map= {'M': 'Orange', 'F': 'Blue'},
            )
fig.update_layout(
            title = "Participation of male and female athelete in both season",
            yaxis_title = "Athelete count")
```

```
import seaborn as sns
import matplotlib.pyplot as plt

x=sns.distplot(df['Age'].dropna(),color='Green')
x.set_title('Age Distribution of Athletes',fontsize=16,fontweight=200)
```

```
h=sns.distplot(df['Height'].dropna(),color='Yellow')
h.set_title('Height Distribution of Athletes',fontsize=16,fontweight=200)
```

```
w=sns.distplot(df['Weight'].dropna(),color='Blue')
w.set_title('Weight Distribution of Athletes',fontsize=16,fontweight=200)
```

```
noc = pd.read_csv('noc_regions.csv')
noc.head(5)
```

```
data = pd.merge(df, region, on='NOC', how='left')
data.head(5)
```

```
plt.figure(figsize=(15, 10))
topc=data.groupby('region')['Medal'].count().nlargest(10).reset_index()
sns.barplot('region','Medal',data=topc)
plt.title('Top Countries in Olympic Medals')
plt.show()
```

```
India = data[(data['region']=='India')]
medals = India['Medal'].value_counts()
medals
```

```
plt.figure(figsize=(10,9), facecolor='lavender')

sns.scatterplot(df['Weight'], df['Height'],hue=df['Medal'])
plt.title("Weight vs Height plot")
```

```
fig = px.scatter(df, x = "Weight", y = "Height", color = "Sport")

fig.update_layout(title = "Distribution of height and weight according to sport")
fig.show()
```

```
#country wise medals list
df=df.merge(region,on='NOC',how='left')
temp_df=df.dropna(subset=['Medal'])
temp_df.drop_duplicates(subset=['Team','NOC','Games','Year','City','Sport','Event','Medal'],inplace=True)
```

```
new_df=temp_df[temp_df['region']=='India']
final_df=new_df.groupby('Year').count()['Medal'].reset_index()
final_df
```

```
fig=px.line(final_df,x="Year",y="Medal")
fig.show()
```

```
#heatmap : which counry is better in which sport
new_df=temp_df[temp_df['region']=='UK']
plt.figure(figsize=(20,20))
sns.heatmap(new_df.pivot_table(index='Sport',columns='Year',values='Medal',aggfunc='count').fillna(0))
```

Fig. 4 – Code(Data Analysis)

5. RESULT AND ANALYSIS

With the help of EDA, we can understand the structure and content of the dataset by various types of graphs and plots which can be drawn with the help of EDA. There are various types of plots which used in EDA. Some of them are mentioned below: • Histogram • Bar Graph • Box Plot • Scatter Plot and many more. We can View the data in the visual format and can explain the analysis on that basis and also perform a Comparative Study between different plots.

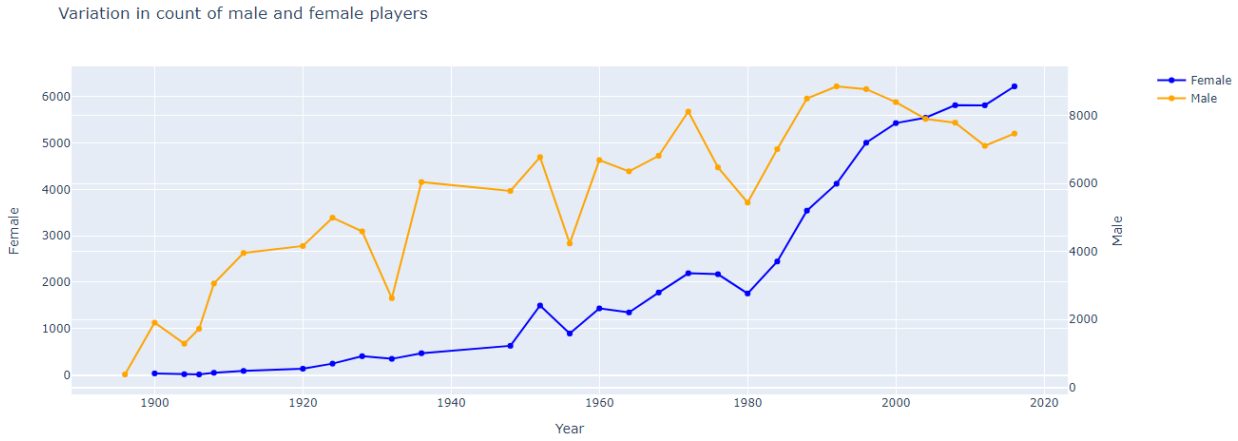


Fig 5. - Comparison

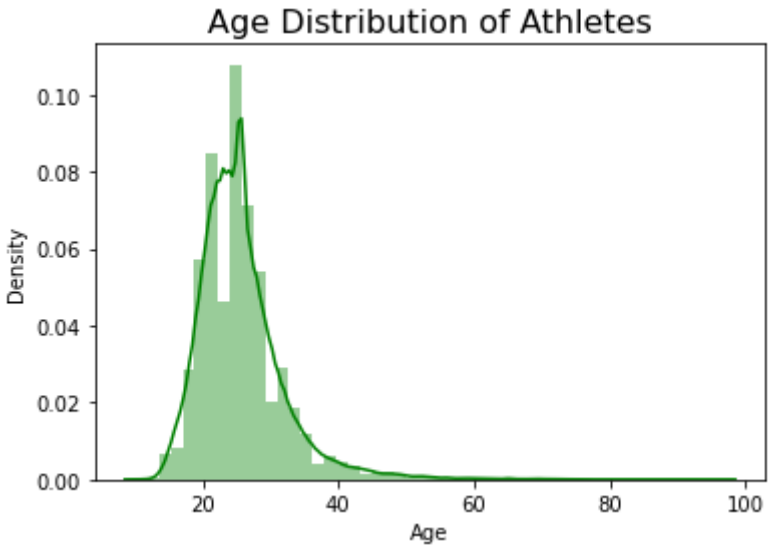


Fig 6. – Age Distribution

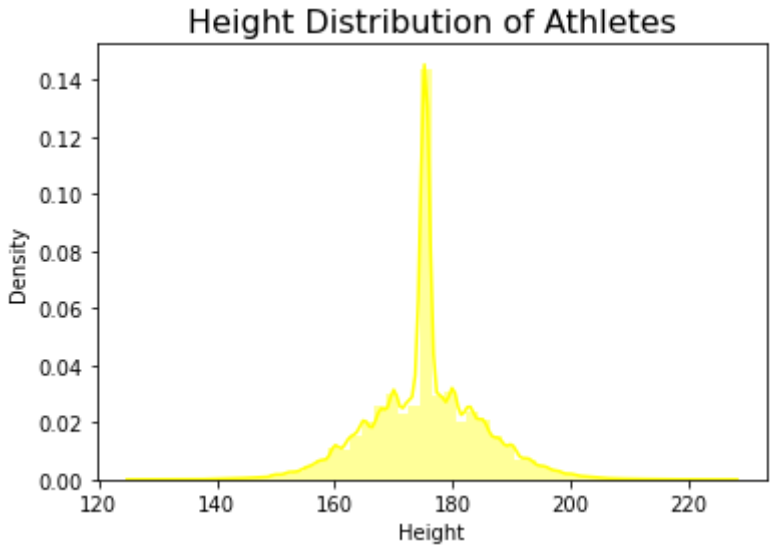


Fig 7. – Height Distribution

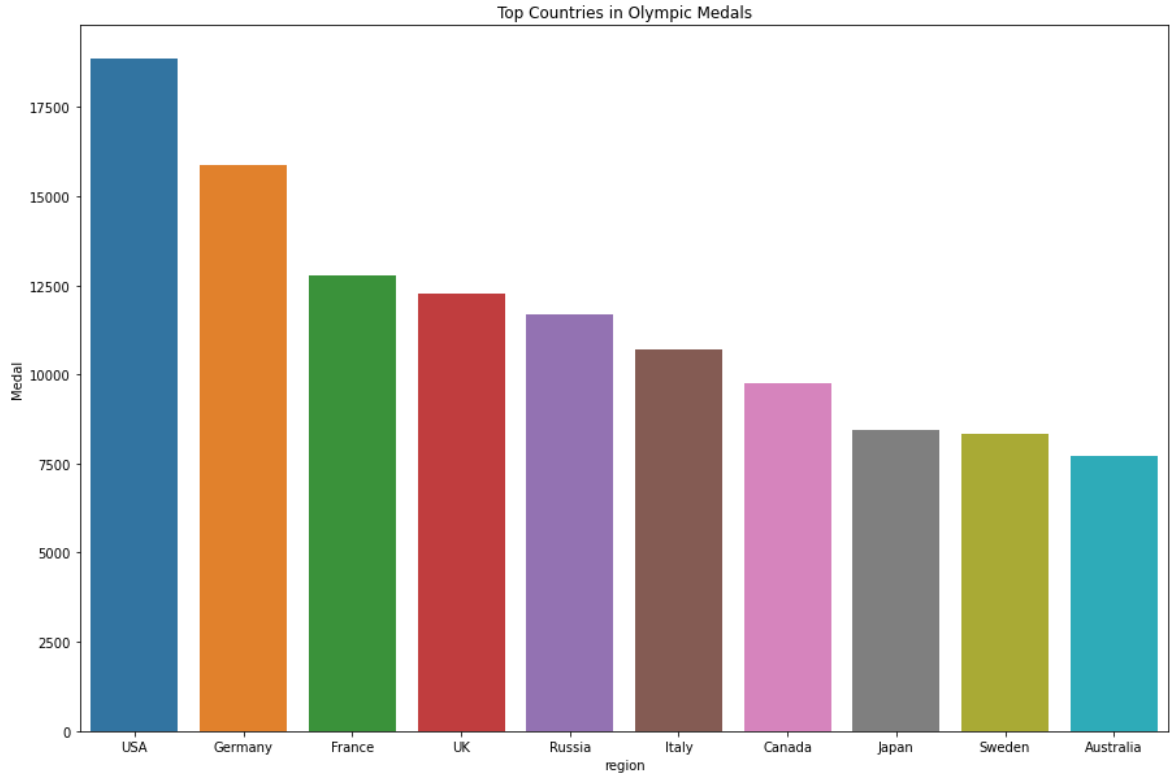


Fig. 8 – Countries Performance.

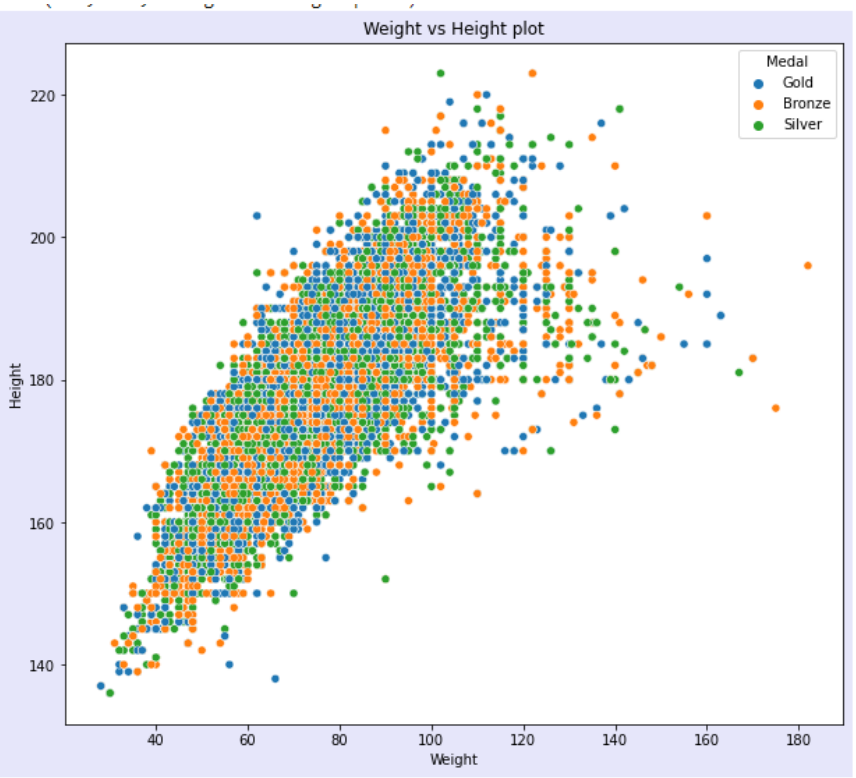


Fig. 8 –Weight VS Height

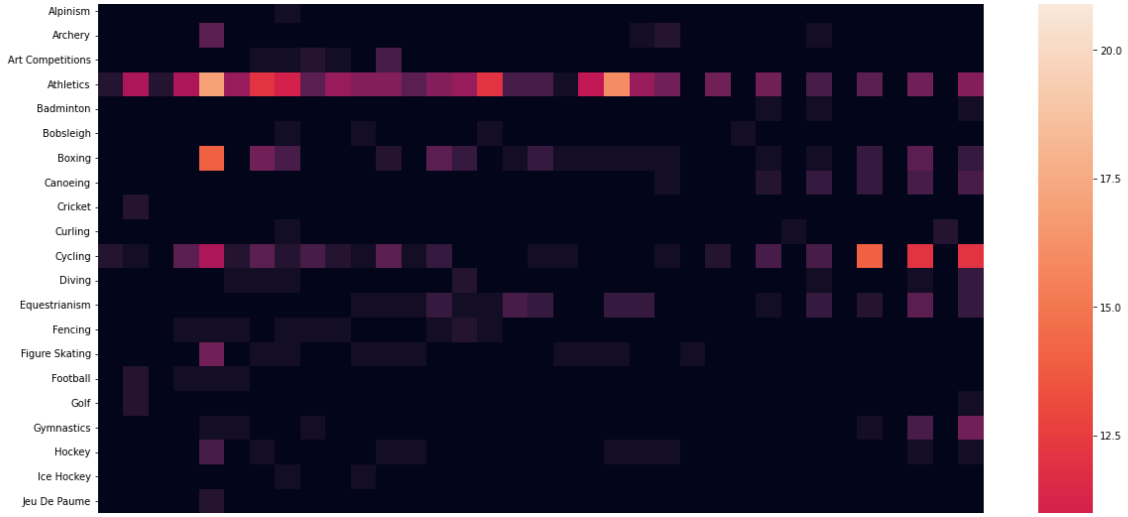


Fig. 9 – Heatmap.

6. CONCLUSION

The main Objective of this study was to Analyze and visualize the various factors which have contributed in the Evolution of Olympic Games over the years. These type of Analysis are very helpful as this type of Analysis can be performed by any Country or Player which can help them in analyzing their performance so that they can improve their performance by changing their strategies. We have used a technique named as Exploratory Data Analysis which enables you to encapsulate the primary factors of a dataset into visual format. We selected Python language to implement our work because It is one of the best language suitable for Data Analysis and Jupyer notebook as the platform where we have preformed this Analysis. As the result of Analysis, we can conclude that It is true that Olympic Games have evolved considerably over the time since 1896 Olympic Games till 2016 Rio Olympics. There are various factors which provides the valid evidence that the Olympics have changed a lot.the Average age of players in Olympic Games, the increase in the participation of the females in both Summer and Winter Olympics over the time, the Total number of medals won by various participating countries over the years, Average height and the weight of Players who contributes to victory of Games in the event

We all know that any Analysis is not perfect and it consists of some limitations which defines the Future scope of the Research Work. This project work also contains some limitations which we are considering as Future Scope of the Project.

These are:.

● We have visualized our data only in Graphical format. We can also describe the data in other formats like Geographical format where we can depict the countries on the World map.

● Till now we have only performed Data Analysis using Exploratory Data Analysis. We can also apply various Machine Learning Algorithms on the data set after Analysis and can create a Predictive Model which can predict the statistics of Future Olympic Games.

● We can also perform Correlation Analysis on the data set and analyze the relation between two continuous variables.

REFERENCES

[1] Wikipedia contributors: https://en.m.wikipedia.org/wiki/Olympic Games, last accessed 2020/11/02.

[2] Dey S K, Rahman M M, Siddiqi U R and Howlader A 2020 Analyzing the epidemiological outbreak of COVID-19: A visual exploratory data analysis approach J. Med. Virol. 92 632–8

[3] Bondu R, Cloutier V, Rosa E and Roy M 2020 An exploratory data analysis approach for assessing the sources and distribution of naturally occurring contaminants (F, Ba, Mn, As) in groundwater from southern Quebec (Canada) Appl. Geochem. 114 104500

[4] Cutait, M.: Management performance of the Rio 2016 Summer Olympic Games. Research Paper submitted and approved to obtain the Master's degree in Sports Administration at AISTS in Lausanne, Switzerland.

[5] Moreno A, Moragas M and Paningua R 1999 The evolution of volunteers at the Olympic Games Proceedings of Symposium on Volunteers (Lausanne, Switzerland: Global Society and the Olympic Movement) pp 1–18

[6] Abeza G, Braunstein-Minkove J R, S´eguin B, O'Reilly N, Kim A and Abdourazakou Y 2020 Ambush marketing via social media: The case of the three most recent Olympic Games Int. J. Sport Communication 1–25

[7] Yamunathangam D, Kirthicka G and Shahanas P 2018 Performance Analysis in Olympic Games using Exploratory Data Analysis Techniques International Journal of Recent Technology and Engineering (IJRTE) 7 251–3

[8] "The Modern Olympic Games" (PDF). The Olympic Museum. Archived from the original (PDF) on 6 September 2008. Retrieved 29 August 2008.

[9] Antarlina Sen and Gaurang Margaj, "A prediction model for which country will win the highest number of Gold", 2016.

[10] Leonardo De Marchi, "Data mining of Sports performance data", 2011.

[11] Huang-Chih Shih," Survey on content-aware Video Analysis for Sports", IEEE Transactions on Circuits and Systems for Video Technology, Vol. 99, No. 9, January 2017.

Analyzing Evolution of the Olympics by Exploratory Data Analysis using Python

Sonali Rasal
Thakur College Of Engineering and Technology
(Student)

Pranjal Singh
Thakur College Of Engineering and Technology
(Student)

Aroma Sinha
Thakur College Of Engineering and Technology
(Student)

Mr. Aniket Mishra
Thakur College Of Engineering and Technology
(Assistant Professor)

Abstract: Olympic Games are one of the main international event and also a matter of prestige for countries and therefore each country tries to give their best performance during the event.An Analysis need to be done by each country to evaluate the previous statistics which will detect the mistakes which they have done previously and will also help them in future development. The primary objective of this Research paper is to analyze the large Olympic dataset using Exploratory Data Analysis to evaluate the evolution of Olympic Games over the years. This analysis will provide detailed and accurate information regarding various factors which leads to the evolution of Olympic Games and improvement of Countries/Players over the time in visual format. Visualization of the data over various factors will provide us with the statistical view of the various factors which leads to the evolution of the Olympic Games and Improvement in performance of various Countries/Players over the time.

1. Introduction

Olympics games are considered as one of the most prime event which provides a valid and common platform for players across different countries to show their talent and skills. Modern Olympic Games were originated by taking inspiration from Ancient Olympic Games held in Olympia, Greece from the 8th Century BC to the 4th Century AD [1]. The following timeline outlines the main events in the history of Modern Olympic

Games. The Olympics consists of various games (Approximately 45) in which players from various countries (Approx 205) participate to win a medal for their country. Olympics has a great history of evolution. From 13 participating nations in 1st Olympics (1896) to 207 participating nations in 2016 Rio Olympics, the Olympics have come across a long way. There are various scenarios which comes in our mind when we look into Evolution of Olympic Games over the years. These scenarios are: Increase in number of participating nations, Increase in number of participating Athletes, Increase/Decrease in number of events, Increase in the expenditure cost of the event, improvement in performance of particular country, improvement in performance of a particular player, Increase in women participation, Participation Ratio of Men to Women, improvement in medication facilities during competition, the effect of pandemic (if any) on the performance of the players. Analysis over these scenarios would depict the evolution of the Olympics over the years. This analysis would help in the future prediction of the number of participating countries, players; winners of various games; Women participation and many more. These type of Analysis can also serve as a performance indicator of a particular country or Player. The main objective of this study is to analyze the various factors mentioned above which plays a vital role in the evolution of Olympic Games over the years. The Analysis will include the visualization and explanation of the change in trends of the various factors over the years which will help to predict the information of future Olympic Games. As Olympic Games are one of the most important sporting event across the world, each country and each player tries to give their best performance in the event. In order to improve their performance, every country should perform such Analysis which would help them in the improvement of their policies and strategies by providing current statistics to them.

2. Literature Survey

Data interpretation and Analysis is one of the main and primary task in the field of big data analytics . There has been a lot of analysis on the Olympic Games like statistics visualization, performance analysis of players, improvement in the performance of various countries and many more. The type of analysis which is quite popular and suitable while analyzing the evolution of the Olympics is Exploratory Data Analysis. In Exploratory Data Analysis, we examine large data and elucidate its various characteristics basically in the visual format(Graphs, Charts, and many more). EDA is an approach which provides deeper

understanding of the dataset.There has been a research paper which analyze 2016 Rio Olympics to find out the various legacies on which these games depends and which are the main reason to explain the hosting of Olympic Games. [4] . This paper used a methodology which uses a performance indicator used in public sector assessment and with the help of it, they find out approximately 32 legacies which plays a major role in the smooth functioning of Olympic Games. [4] There is another research paper which analyze the evolution of volunteering activities in Olympic Games. [5] Volunteers are the ones who offers to take part in the event, arrangement, social activities or works for an organization without being paid. This paper detailed analyzed official reports of each Olympic Games (Winter as well as Summer) and a survey of Olympic Bibliography [5]. They also made an effort and tried to achieve the direct corroboration from the volunteers who had participated in the respective Olympic Game [5].

3. Methodology

3.1. Data Collection

The very first step of any type of Analysis, whether it is technical or non-technical, is Data Collection. In order to perform analysis on a certain problem, we require a large amount of Data on which we apply various techniques and algorithms to reach to a particular conclusion and get our desired result. It is advised to take the data in abundance because larger the volume of data for analysis, the greater would be the accuracy in the result and also the greater would be the confidence in decision making based on these results. We have used data from various data sources for analysis on Evolution of the Olympics over the time. We have taken two datasets which provide us with large volume and a large variety of data for Analysis. 1st dataset consists the information about the players and their entire details like their Gender, Height, Weight, Country for which they play, Medals won (Gold, Silver and Bronze) and many more. This data can be used to analyze the performance of the particular player and can also help in the comparative study between two or more players. 2nd dataset consists the information of the countries and their NOC .

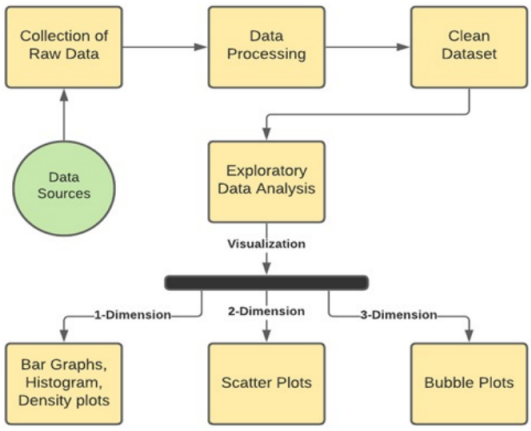


Fig 1 EDA Process

3.2. Data preprocessing

The Dataset consists of various fields like Age, Gender, etc which consists of some null values which produces errors in the end result which is the Visualization of data in graphical format. These null values are needed to be omitted or replaced with some valid value which solves the error and generates accurate result. We have used a technique known as Deterministic Imputation to complete this task. Deterministic Imputation is a situation where the null values (NA or NaN) are determined with the help of the other values in the same column in the dataset. For this purpose, there are various models such as Basic Numeric Imputation Model in which the null value is replaced by Mean or Median of other values of the same column of the dataset.

3.3. Exploratory Data Analysis

The next step after data pre-processing is data analysis. In this step, analysis is done on data using various Techniques like Text Analysis, Diagnostic Analysis, Exploratory Data Analysis, etc and Machine learning Algorithms like Linear Regression, Logistic Regression, SVM, Decision Tree etc to reach to a particular conclusion.We are using the Exploratory Data Analysis technique to complete this task. Exploratory Data Analysis (EDA) is an approach to analyze data thoroughly and encapsulate its primary attributes basically in visual format. [8] Exploratory Data Analysis is mainly used to see what the data represents apart from applying various algorithms. [8] With the help of EDA, we can understand the structure and content of the dataset by various types of graphs and plots which can be drawn with the help of EDA. There are various types of plots which used in EDA. Some of them are mentioned below:

- Histogram
- Bar Graph
- Box Plot
- Scatter Plot and many more.

We can View the data in the visual format and can explain the analysis on that basis and also perform a Comparative Study between different plots.

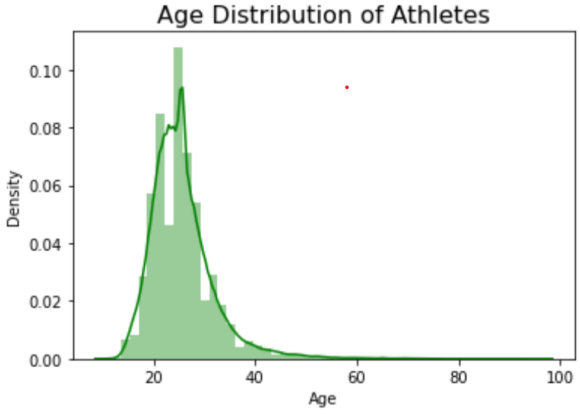


Fig 2. Age distribution of athletes

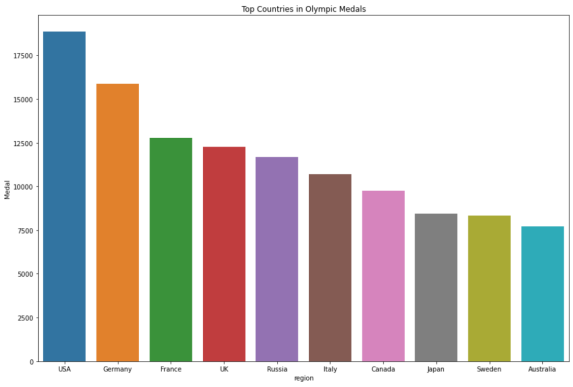


Fig 3. Top countries for medals

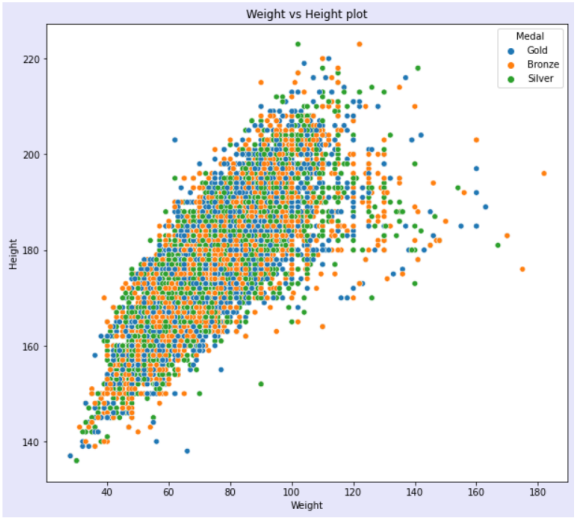


Fig 4. Weight vs height plot

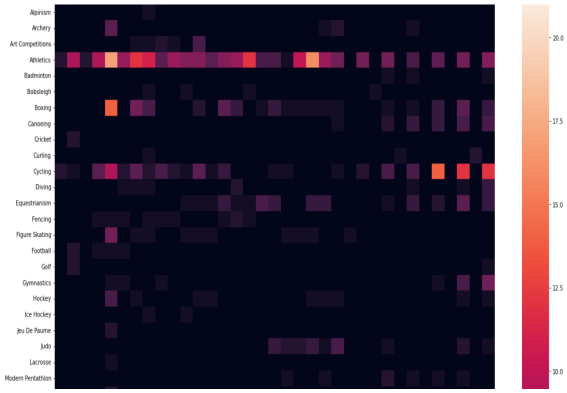


Fig 5. Heatmap

4. Conclusion

The main Objective of this study was to Analyze and visualize the various factors which have contributed in the Evolution of Olympic Games over the years. These type of Analysis are very helpful as this type of Analysis can be performed by any Country or Player which can help them in analyzing their performance so that they can improve their performance by changing their strategies. We have used a technique named as Exploratory Data Analysis which enables you to encapsulate the primary factors of a dataset into visual format.

We selected Python language to implement our work because It is one of the best language suitable for Data Analysis and Jupyer notebook as the platform where we have preformed this Analysis. As the result of Analysis, we can conclude that It is true that Olympic Games have evolved considerably over the time since 1896 Olympic Games till 2016 Rio Olympics. There are various factors which provides the valid evidence that the Olympics have changed a lot.the Average age of players in Olympic Games, the increase in the participation of the females in both Summer and Winter Olympics over the time, the Total number of medals won by various participating countries over the years, Average height and the weight of

Players who contributes to victory of Games in the event.

5. Future Scope

We all know that any Analysis is not perfect and it consists of some limitations which defines the Future scope of the Research Work. This project work also contains some limitations which we are considering as Future Scope of the Project. These are:.

- We have visualized our data only in Graphical format. We can also describe the data in other formats like Geographical format where we can depict the countries on the World map.
- Till now we have only performed Data Analysis using Exploratory Data Analysis. We can also apply various Machine Learning Algorithms on the data set after Analysis and can create a Predictive Model which can predict the statistics of Future Olympic Games.
- We can also perform Correlation Analysis on the data set and analyze the relation between two continuous variables.

References

[1] Wikipedia contributors: https://en.m.wikipedia.org/wiki/Olympic Games, last accessed 2020/11/02.

[2] Dey S K, Rahman M M, Siddiqi U R and Howlader A 2020 Analyzing the epidemiological outbreak of COVID-19: A visual exploratory data analysis approach J. Med. Virol. 92 632–8

[3] Bondu R, Cloutier V, Rosa E and Roy M 2020 An exploratory data analysis approach for assessing the sources and distribution of naturally occurring contaminants (F, Ba, Mn, As) in groundwater from southern Quebec (Canada) Appl. Geochem. 114 104500

[4] Cutait, M.: Management performance of the Rio 2016 Summer Olympic Games. Research Paper submitted and approved to obtain the Master's degree in Sports Administration at AISTS in Lausanne, Switzerland.

[5] Moreno A, Moragas M and Paningua R 1999 The evolution of volunteers at the Olympic Games Proceedings of Symposium on Volunteers (Lausanne, Switzerland: Global Society and the Olympic Movement) pp 1–18

[6] Abeza G, Braunstein-Minkove J R, S´eguin B, O'Reilly N, Kim A and Abdourazakou Y 2020 Ambush marketing via social media: The case of the three most recent Olympic Games Int. J. Sport Communication 1–25

[7] Yamunathangam D, Kirthicka G and Shahanas P 2018 Performance Analysis in Olympic Games using Exploratory Data Analysis Techniques International Journal of Recent Technology and Engineering (IJRTE) 7 251–3