**Project Report**

# On

#### **“Olympic Data Analysis”**

#### **Submitted to**

**Rashtrasant Tukadoji Maharaj Nagpur University Nagpur**

In the partial fulfillment for the Award Degree of

**MASTER OF COMPUTER MANAGEMENT**

#### **Submitted By**

**“RAJJAT KUMARE1”**

**“KRISTINA ANNAPAREDDY2”**

**“LOKESH SONARE3”**

##### Under the Guidance of

**Prof. S. S. Anwaney**



**Department of Computer Education**

**Dhanwate National College**

**Congress Nagar, Nagpur - 440 012.**

**2021 – 2022**

*ShriShivaji Education Society, Amravati’s*

**Dhanwate National College**

**Department of Computer Education**

**CERTIFICATE**

This is to certify that the Project entitled **“Olympic Data Analysis”** submitted by Rajjat Kumare, Kristina Annapareddy, Lokesh Sonare in partial fulfillment of the requirements for the award of the **MASTER of Computer MANAGEMENT** from Rashtrasant Tukdoji Maharaj Nagpur University and is a bonafide record of the work done by him at Dhanwate National College, Nagpur during the academic year 2020 - 21.

Guide

**(Prof. S. S. Anwaney)**

Place: Nagpur

Date:

**Principal**

Dhanwate National College

Nagpur

**Internal Examiner External Examiner**

**DECLARATION**

To,

**The Principal,**

Dhanwate National College,

Congress Nagar,

Nagpur – 440 012.

Respected Sir,

We the undersigned, hereby declare that the work entitled **“Olympic Data Analysis”** developed and submitted by us is ours original work. The system presented here is developed by us independently and has not been duplicated from any other source.

We understand that any such copying is liable to be punished in any way the University deem fit.

Place: Nagpur

**“RAJJAT KUMARE1”**

**“KRISTINA ANNAPAREDDY2”**

**“LOKESH SONARE3**

**ACKNOWLEDGEMENT**

#### We wish to express my sincere thanks to the honorable **Dr. J. D. Wadte,** Principal, Dhanwate National College, Congress Nagar, Nagpur, for providing me varieties of opportunities, infrastructural facilities and inspiration to gather professional knowledge and material without which it would have been impossible to complete this hard task.

#### We take this opportunity to express our deep gratitude and whole hearted thanks to us **Project guide Guide (Prof. S. S. Anwaney)** for his guidance throughout this work. Weare very much thankful to him for his kindness, encouragement and the valuable time, which he has devoted to me.

We wish to thanks all those, who have helped me in us way or the others in bringing out this project successful.

**“RAJJAT KUMARE1”**

**“KRISTINA ANNAPAREDDY2”**

**LOKESH SONARE3**

**INDEX**

1. Brief Review of Project
   1. Introduction
   2. Aims and Objectives
2. Preliminary System Analysis
   1. Identification of needs
   2. Tools, Components and Platform
   3. Study of Existing System
3. Project Category Software/Hardware Requirements
   1. Hardware Specification
   2. Software Specification
   3. Language Used
4. Detail Design System
   1. Data Flow Diagram
5. Datastructures and Table

5.1 Athletes Data

5.2 Region Data

6. Exploratory Data Analysis

6.1 Data preparaton and Data Cleaning

6.2 Data Manipulation

6.3 Data Visualisaton

7. Source Code for Deployment

8. Screen

8.1 Overall Tally

8.2 Medal Tally

8.3 Overall Analysis

8.4 Country-wise Analysis

8.5 Athlete wise Analysis

9. Data Report

10. Maintenance

11. Fututre Scope of Project

12. Conclusion

13. Limitations

14. BiblioGraphy

**INTRODUCTION**

Present system of ”OLYMPIC DATA ANALYSIS” is handled manually which is a tedious job & time consuming job, this is so because everyday many analysis take place and each time we have to done to do complete analysis in respective analyze and registers in which new records are maintained , this analysis report shows how sports analytics crucial, ultimately resulting in chaos.

the web app dashboard developed eases the job of data analysis by making the maintenance of registers computerized and producing the required information quickly by withdrawing few outputs.

with the help of this project the job of the become easy by making maintenance of register computerized. A very adequate data visualization will be established.

In medal tally form we can display all information about overall Olympic medals such as medal tally. We can draw insight of data report using EDA(Exploratory Data Analysis). In the medal tally the module is used to show the tally analysis of overall medal counting which show when and which year it had won it

In this dashboard, my goal is to shed light on major patterns in Olympic history. How many athletes, sports, and nations are there? Where do most athletes come from? Who wins medals? What are the characteristic of the athletes (e.g., gender and physical size)?

In country-wise analysis, we can display data visualization graph how to country were performing in history in Olympic games. Also shows the top Athletes of country.

**AIM & OBJECTIVE**

* This data analysis dashboard main objective is to draw insight of the sports industry and help sports person to make decision about the game career.
* This dashboard helps sports industry to make better decision that which age group men and women are participating in sports faculty and which country paying more attention to their sport players.
* It also helps in sports analytics to achieve their objectives and customizes the all record and maintain it to show the visual graph to understand the data
* Using the data visualization plot to define the data pattern. To clarify purpose of sports industry.
* This dashboard shows the distribution of players age according to their country and what ideal age is achieve more gold medal in games.

**IDENTIFICATION OF NEED**

One must know what the problem is before it can be solved. It helps to find out real problem of the existing system without clear understanding of the problem the further work will be wastage of time and energy. So it is necessary to identify the actual problem. The initial investigation must satisfy the following conditions:

1. What effect does the host country have in the medals won at the olympic?
2. Is the age of winning Olympics changing?
3. How many countries were participated in these?
4. Overall Medal Tally

The success of the system depends largely on how accurately a problem is defining thoroughly investigation`& properly carried out through the choice of the solution here mostly concern with the user need rather that what he or she wants.

**PRELIMINARY ANALYSIS**

Preliminary Analysis deals with investigations on the basic of initial stage of the project. In this stage of project development we have required various information related to data.

In this project “Olympic Data Analysis” we have collect information form the required website called ‘kaggle’.

**PROJECT CATEGORY**

We have selected the project category for this project is python as Front End and MS Excel as database , for deployment streamlit and for data .

.

**About pyhon:**

Data Analysis is the technique to collect, transform, and organize data to make future predictions, and make informed data-driven decisions. It also helps to find possible solutions for a business problem. There are six steps for Data Analysis. They are:

* Ask or Specify Data Requirements
* Prepare or Collect Data
* Clean and Process
* Analyze
* Share
* Act or Report

Each step has its own process and tools to make overall conclusions based on the data. Also uses following process to analyze the data

* **Data Cleaning**: It involves finding and correcting any inaccuracies or ambiguity present in the stored data.
* **Data Preprocessing**: It is the process of modifying data into formats that are more suitable for performing Data Analytics with Python.
* **Data Manipulation**: It is the process of implementing Machine Learning models on data to obtain desired results. Tasks like Clustering, Classification, Regression, etc. fall under Data Manipulation as shown in the below image.

**About Ms-Excel:**

Excel spreadsheets have been around for more than 30 years and they’re still valuable. The original concept isn’t much different than what we use today, it just looks better and has a lot of new capabilities.

It’s true that larger corporations have moved away from spreadsheets on a big data scale, however spreadsheets are still used for everyday items. In its most basic form, Excels holds data points in each cell. Anything like raw data exports, date of sales, SKU, or units sold are entered (or imported) into a spreadsheet for easier viewing and organization purposes. A successful Excel spreadsheet will organize raw data into a readable format that makes it easier to extract actionable insights. With more complex data, Excel allows you to customize fields and functions that make calculations for you. Even with larger data sets, segmented data can be studied more carefully and visualized without using other software. Determine hypothetical profit margins or department budgets. While it won’t build a full-scale data product alone, it can present easy-to-read visualizations and accurate calculations.

As mention as above about

**SOFTWARE REQUIREMENT**

When you initiate any Project then there is need of some software requirement. Here we have also need of some software requirement some of them are listed below:

* Windows 10
* Visual studio Code
* Microsoft excel
* Jupyter notebook(for Data manipulation)
* Streamlit (Web Deployment)
* Pycharm

Packages for Running:

* Jupyter notebook
* Matplotlib
* Scikit-learn
* Numpy
* Pandas
* Streamlit
* Plotly express
* Plotly figure\_factory
* Seaborn

Cloud for Deployment:

* Streamlit hosting
* Streamli cloud
* Git and Github
* Heroku

Note:-

It is also mention that all software as mention as above is properly install in your computer, then only your project work properly.

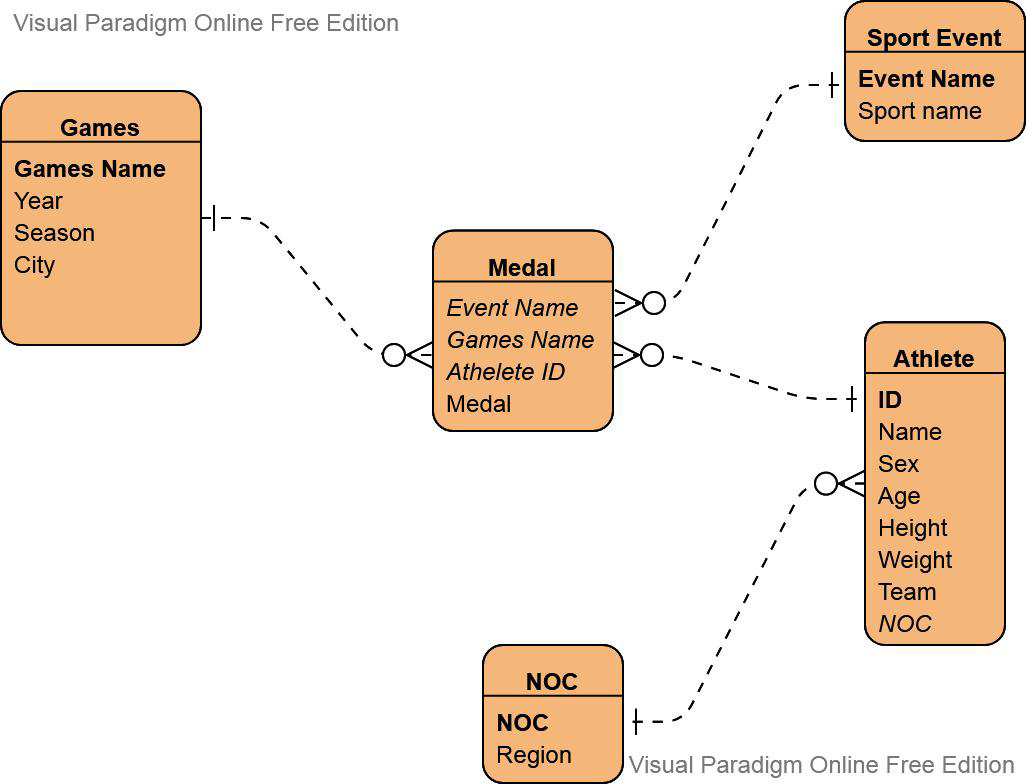
**HARDWARE REQUIREMENT**

Any physical component of the computer system is known as “Hardware”.

Our project is basically Windows based. We have used many devices of computer system which are as follow:

* Motherboard: i5 gen 11
* RAM: 32 GB, 16 GB, 8 GB
* Processor: 2.30 GHz
* Hard Disk: 230 GB

**DATA FLOW DIAGRAM**

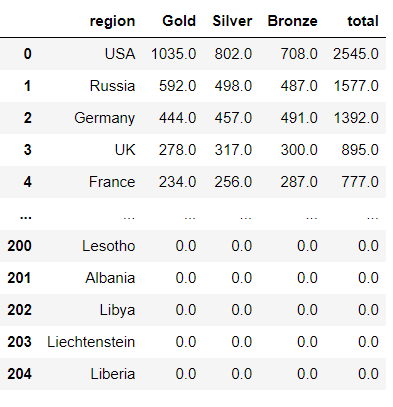
****

**DATA STRUCTURE AND TABLE**

**Athlete Data:**



**Region Data:**



**EXPLORATORY DATA ANALYSIS**

**Data preparation and Data cleaning:**

As the first step of EDA, we have to understand what the whole dataframe is like. Generally in this step, we’re going to do following things:

Explore the number of rows and columns, range of values etc.

Handle missing, incorrect, invalid, and duplicated data.

df.isnull().sum()

# Sum of all duplicate value

df.duplicated().sum()

# Drop all duplicate value

df.drop\_duplicates(inplace=True)

# Count Medal

df['Medal'].value\_counts()

# Merge with region\_df

df = df.merge(region\_df, on='NOC', how='left')

# One hot encoding medals

df = pd.concat([df,pd.get\_dummies(df['Medal'])],axis=1)

# Groupby

df.groupby('NOC').sum()[['Gold','Silver','Bronze']].sort\_values('Gold',ascending=False).reset\_index().head(25)

**Source Code Deployment**

**App.py**

# Import all the packages

import streamlit as st

import pandas as pd

import preprocessor,helper

import plotly.express as px

import matplotlib.pyplot as plt

import seaborn as sns

import plotly.figure\_factory as ff

# Load The Datasets

df = pd.read\_csv('athlete\_events.csv')

region\_df = pd.read\_csv('noc\_regions.csv')

df = preprocessor.preprocess(df,region\_df)

# Set the Title

st.title("Olympic Data Analysis Dashboard")

st.sidebar.title("Olympics Analysis")

st.sidebar.image('https://e7.pngegg.com/pngimages/1020/402/png-clipart-2024-summer-olympics-brand-circle-area-olympic-rings-olympics-logo-text-sport.png')

user\_menu = st.sidebar.radio(

'Select an Option',

('Description','Medal Tally','Overall Analysis','Country-wise Analysis','Athlete wise Analysis')

**)**

# Olympics Analysis

st.sidebar.title("Olympics Analysis")

if user\_menu == 'Description':

with st.expander("What is this app about?"):

st.markdown("""

###### The Olympic Games is a well-known sporting platform which is recognized all over the world, has been distinguished from the late 19th century. Its origin however can be traced back to the Greek empire, at around 3,000 years ago, which consisted only of sprint race, and was held in Greece’s city Olympia only accessible to freeborn Greek people (Young & Abraham, 2020). It has grown since then and now has become a hub for all the athletes worldwide to demonstrate their abilities in more than 28 individual sporting contests. Currently, it is held every 2 years in different countries, with names Summer Olympics and Winter Olympics, both having their own set of games (Young & Abraham, 2020). It has become a place which reflects the power of the individual contestants and has become a source of pride for the countries they represent.

### Motivation and Research Question

###### Olympics has a rich history, spanning from 1896 till 2018, and has been a part of history. So, it is an interesting topic to see how the historical events have affected the specifics of Olympics and how it has been changing till date. Hence, this report attempts to build around the following questions, with some connection to historical happenings where suitable:

- 1. What effect does the host country have in the medals won at the Olympics?

- 2. Is the age of winning Olympics changing?

- 3. How many countries were participated in this event?

- 4. Overall Medal Tally

""")

with st.expander("Literature View: "):

st.markdown("""

The advantage of host country in any sporting activity is well known, as the participants will have familiarity of the field, and also there is a great support from the home crowd. Host countries are expected to win 3 times the medals that they were winning while playing as away (Clarke, 2000). Being a host nation and also having a communist background is also going to have a positive effect in the number of medals won (Bian, 2005).

It is found that a country’s socio-economic variables, such as GDP affects the country’s performance in the Olympics by a great factor. Country’s population and its GDP is seen to have a correlation with the number of medals won in Olympics (Bian, 2005).

Age factor is also one of the important ones when it comes to sports and even among the athletes of the same age, relative age effect (RAE) comes into factor which determines who triumphs (Fletcher & Sarkar, 2012). RAE states that an athlete can have more advantage as compared to another who is younger by almost a year with respect to maturity, experience and early specialization (Neill, Cotton, Cuadros & Connor, 2016).

Olympics has been part of the history, and has affected, and also has been affected by the history.

Olympic Games has made major social and political impacts throughout the history like including women in sports, has taken a stand against racial matters, promoted civil rights, has unified countries and even has been a tool to demonstrate power by different countries (O’Connell). Similarly, the politics of the countries such as racial separation, terrorism, World Wars and the Cold War have also affected Olympics at different times in the history (Dwyer & McMaster, 2018).

""")

with st.expander("Approach: "):

st.markdown("""

To answer the questions, three datasets were used namely, “120 years of Olympic history: athletes and results[1]”, “Gapminder GDP per capita, constant PPP dollars- v25[2]” and “Gapminder Total Population v6[3]”.

The Olympics dataset had the names of participants, their demographics, which sport they participated in and on which Olympic games. Gapminder’s GDP per capita dataset consisted of GDP of countries from 1960 forecasted till 2040, and their Population dataset had populations of all the countries of the world from 1800 till 2019 and forecasted from there onwards till 2100.

A custom dataset was also created that maps the cities mentioned in the Olympics dataset to country names.

Visualizations were created in Tableau and Python. Initial data cleaning was done on Excel, and visualization specific data manipulation were carried out as needed on Python.

""")

with st.expander("Conclusion: "):

st.markdown("""

It is clear that the host countries have always a better chance of winning medals in the Olympics;

they can win at least 10–20 percent more medals.

Looking at the economic effect, even though country’s population and per capita GDP affected the number of medals won in the past, the total GDP of the country is more significant to determine the winnings in the recent years.

With the age factor, the age range of players winning medals has decreased over the years, and an optimal age for each sport can be identified in the recent years.

Thus, there is a high chance for an athlete from a host country with high GDP, whose age range falls in the optimum age range for the sport to win a medal in the Olympics.

""")

st.markdown(

"""

<style>

[data-testid='stSidebar"][aria=expanded="true"] > div:first-child{

width: 350px

}

[data-testid='stSidebar"][aria=expanded="false"] > div:first-child{

width: 350px

margin-left: -350px

}

</style>

""",

unsafe\_allow\_html=True,

)

# Medal Tally

if user\_menu == 'Medal Tally':

st.sidebar.header("Medal Tally")

years,country = helper.country\_year\_list(df)

selected\_year = st.sidebar.selectbox("Select Year", years)

selected\_country = st.sidebar.selectbox("Select Country", country)

medal\_tally = helper.fetch\_medal\_tally(df,selected\_year,selected\_country)

if selected\_year == 'Overall' and selected\_country == 'Overall':

st.title("Overall Tally")

if selected\_year != 'Overall' and selected\_country == 'Overall':

st.title("Medal Tally in" + str(selected\_year) + " Olympics")

if selected\_year == 'Overall' and selected\_country != 'Overall':

st.title(selected\_country + " Overall performance")

if selected\_year != 'Overall' and selected\_country != 'Overall':

st.title(selected\_country + " performance is " + str(selected\_year) + " Olympics")

st.table(medal\_tally)

if user\_menu == 'Overall Analysis':

editions = df['Year'].unique().shape[0] - 1

cities = df['City'].unique().shape[0]

sports = df['Sport'].unique().shape[0]

events = df['Event'].unique().shape[0]

athletes = df['Name'].unique().shape[0]

nations = df['region'].unique().shape[0]

with st.expander("More info about Overall Analysis:"):

st.markdown("""

Overall Analysis of the olympic medal System to show you the how many

Edition was organised in the past 120 years. It will show you the which

country had hosted the event and played all around the world.

""")

st.title("Top Statistics")

with open('style.css') as f:

st.markdown(f'<style>{f.read()}</style>', unsafe\_allow\_html=True)

col1,col2,col3 = st.columns(3)

with col1:

st.header("Editions")

st.subheader(editions)

with col2:

st.header("Hosts")

st.subheader(cities)

with col3:

st.header("Sports")

st.subheader(sports)

col1, col2, col3 = st.columns(3)

with col1:

st.header("Events")

st.subheader(events)

with col2:

st.header("Nations")

st.subheader(nations)

with col3:

st.header("Athletes")

st.subheader(athletes)

nations\_over\_time = helper.data\_over\_time(df,'region')

fig = px.line(nations\_over\_time, x="Edition", y="region")

st.title("Participating Nations over the years")

st.plotly\_chart(fig)

events\_over\_time = helper.data\_over\_time(df, 'Event')

fig = px.line(events\_over\_time, x="Edition", y="Event")

st.title("Events over the years")

with st.expander("Notes: "):

st.markdown("""

Text

.""")

st.plotly\_chart(fig)

athlete\_over\_time = helper.data\_over\_time(df, 'Name')

fig = px.line(athlete\_over\_time, x="Edition", y="Name")

st.title("Athletes over the years")

st.plotly\_chart(fig)

st.title("No. of Events over time(Every Sport)")

fig,ax = plt.subplots(figsize=(20,20))

x = df.drop\_duplicates(['Year', 'Sport', 'Event'])

ax = sns.heatmap(x.pivot\_table(index='Sport', columns='Year', values='Event', aggfunc='count').fillna(0).astype('int'),annot=True)

st.pyplot(fig)

st.title("Most successful Athletes")

sport\_list = df['Sport'].unique().tolist()

sport\_list.sort()

sport\_list.insert(0,'Overall')

selected\_sport = st.selectbox('Select a Sport',sport\_list)

x = helper.most\_successful(df,selected\_sport)

st.table(x)

with st.expander("Notes: "):

st.markdown("""

This statistic shows the athletes with the number of Summer Olympic victories since

the start of the Olympic Games from 1896 to 2021. Michael Phelps, the legendary American swimmer,

won 23 gold medals across his participation in the games.

""")

# Country wise analysis

if user\_menu == 'Country-wise Analysis':

st.sidebar.title('Country-wise Analysis')

country\_list = df['region'].dropna().unique().tolist()

country\_list.sort()

selected\_country = st.sidebar.selectbox('Select a Country',country\_list)

country\_df = helper.yearwise\_medal\_tally(df,selected\_country)

fig = px.line(country\_df, x="Year", y="Medal")

st.title(selected\_country + " Medal Tally over the years")

over\_the\_year = st.checkbox('Load Graph')

if over\_the\_year:

st.plotly\_chart(fig)

st.title(selected\_country + " excels in the following sport")

pt = helper.country\_event\_heatmap(df,selected\_country)

fig,ax = plt.subplots(figsize=(20,20))

ax = sns.heatmap(pt, annot=True)

st.pyplot(fig)

st.title("Top 10 athletes of " + selected\_country)

top10\_df = helper.most\_successful\_countrywise(df,selected\_country)

st.table(top10\_df)

# Athlete wise Analysis

if user\_menu == 'Athlete wise Analysis':

athlete\_df = df.drop\_duplicates(subset=['Name', 'region'])

x1 = athlete\_df['Age'].dropna()

x2 = athlete\_df[athlete\_df['Medal'] == 'Gold']['Age'].dropna()

x3 = athlete\_df[athlete\_df['Medal'] == 'Silver']['Age'].dropna()

x4 = athlete\_df[athlete\_df['Medal'] == 'Bronze']['Age'].dropna()

fig = ff.create\_distplot([x1, x2, x3, x4], ['Overall Age', 'Gold Medalist', 'Silver Medalist', 'Bronze Medalist'],

show\_hist=False, show\_rug=False)

fig.update\_layout(autosize=False, width=1000,height=600)

st.title(" 1. Distribution of Age")

Dist\_age= st.checkbox('Load Graph')

if Dist\_age :

st.plotly\_chart(fig)

x = []

name = []

famous\_sports = ['Basketball', 'Judo', 'Football', 'Tug-Of-War', 'Athletics',

'Swimming', 'Badminton', 'Sailing', 'Gymnastics',

'Art Competitions', 'Handball', 'Weightlifting', 'Wrestling',

'Water Polo', 'Hockey', 'Rowing', 'Fencing',

'Shooting', 'Boxing', 'Taekwondo', 'Cycling', 'Diving', 'Canoeing',

'Tennis', 'Golf', 'Softball', 'Archery',

'Volleyball', 'Synchronized Swimming', 'Table Tennis', 'Baseball',

'Rhythmic Gymnastics', 'Rugby Sevens',

'Beach Volleyball', 'Triathlon', 'Rugby', 'Polo', 'Ice Hockey']

for sport in famous\_sports:

temp\_df = athlete\_df[athlete\_df['Sport'] == sport]

x.append(temp\_df[temp\_df['Medal'] == 'Gold']['Age'].dropna())

name.append(sport)

fig = ff.create\_distplot(x, name, show\_hist=False, show\_rug=False)

fig.update\_layout(autosize=False, width=1000, height=600)

st.title(" 2. Distribution of Age wrt Sports(Gold Medalist)")

wrt\_sport= st.checkbox('Load Graph', key=int)

if wrt\_sport:

st.plotly\_chart(fig)

sport\_list = df['Sport'].unique().tolist()

sport\_list.sort()

sport\_list.insert(0, 'Overall')

st.title(" 3. Height vs Weight")

selected\_sport = st.selectbox('Select a Sport', sport\_list)

temp\_df = helper.weight\_v\_height(df,selected\_sport)

fig,ax = plt.subplots()

ax = sns.scatterplot(temp\_df['Weight'],temp\_df['Height'],hue=temp\_df['Medal'],style=temp\_df['Sex'],s=60)

st.pyplot(fig)

st.title(" 4. Men Vs Women Participation Over The Years")

with st.expander("Notes: "):

st.markdown("""

The share of male athletes at the Summer Olympics has always been greater than the share of female athletes.

The first modern Olympic Games in Athens in 1896 was exclusively for male competitors, and although some female events were introduced in Paris in 1900, the share of events was just 2.2 percent.

Over the next century, the ratio of female to male events has gradually narrowed, and at a faster rate in recent decades, reaching almost 49 percent in Tokyo 2020.

The Tokyo Games in particular saw the introduction of several mixed events, in an attempt to increase this participation further; these included mixed shooting events and both sprinting and swimming mixed relays.

In the history of the Summer Olympics, the share of events for women athletes has grown gradually; from no events at all in 1896, to a 51.3 percent share in 2020. Growth was gradual and slow, crossing the 25 percent mark in 1984, and exceeding 50 percent in 2020.

It is important to note, however, that a number of these events are mixed, for both male and female competitors, and the total number of single-gender events remains slightly higher for men. The majority of sports have mirrored events for male and female competitors,

and there had been an increased number of mixed events in recent years, with a particularly high number of mixed events added in Tokyo 2020.

""")

final = helper.men\_vs\_women(df)

fig = px.line(final, x="Year", y=["Male","Female"])

fig.update\_layout(autosize=False, width=1000, height=600)

men\_vs\_women = st.checkbox('Load Graph', key=str)

if men\_vs\_women:

st.plotly\_chart(fig)

**helper.py**

import numpy as np

# Fetch the medal from the given datasets

def fetch\_medal\_tally(df, year, country):

medal\_df = df.drop\_duplicates(subset=['Team', 'NOC', 'Games', 'Year', 'City', 'Sport', 'Event', 'Medal'])

flag = 0

if year == 'Overall' and country == 'Overall':

temp\_df = medal\_df

if year == 'Overall' and country != 'Overall':

flag = 1

temp\_df = medal\_df[medal\_df['region'] == country]

if year != 'Overall' and country == 'Overall':

temp\_df = medal\_df[medal\_df['Year'] == int(year)]

if year != 'Overall' and country != 'Overall':

temp\_df = medal\_df[(medal\_df['Year'] == year) & (medal\_df['region'] == country)]

if flag == 1:

x = temp\_df.groupby('Year').sum()[['Gold', 'Silver', 'Bronze']].sort\_values('Year').reset\_index()

else:

x = temp\_df.groupby('region').sum()[['Gold', 'Silver', 'Bronze']].sort\_values('Gold',

ascending=False).reset\_index()

x['total'] = x['Gold'] + x['Silver'] + x['Bronze']

x['Gold'] = x['Gold'].astype('int')

x['Silver'] = x['Silver'].astype('int')

x['Bronze'] = x['Bronze'].astype('int')

x['total'] = x['total'].astype('int')

return x

def medal\_tally(df):

medal\_tally = df.drop\_duplicates(subset=['Team', 'NOC', 'Games', 'Year', 'City', 'Sport', 'Event', 'Medal'])

medal\_tally = medal\_tally.groupby('region').sum()[['Gold', 'Silver', 'Bronze']].sort\_values('Gold',ascending=False).reset\_index()

medal\_tally['total'] = medal\_tally['Gold'] + medal\_tally['Silver'] + medal\_tally['Bronze']

medal\_tally['Gold'] = medal\_tally['Gold'].astype('int')

medal\_tally['Silver'] = medal\_tally['Silver'].astype('int')

medal\_tally['Bronze'] = medal\_tally['Bronze'].astype('int')

medal\_tally['total'] = medal\_tally['total'].astype('int')

return medal\_tally

def country\_year\_list(df):

years = df['Year'].unique().tolist()

years.sort()

years.insert(0, 'Overall')

country = np.unique(df['region'].dropna().values).tolist()

country.sort()

country.insert(0, 'Overall')

return years,country

def data\_over\_time(df,col):

nations\_over\_time = (df.drop\_duplicates(['Year', col])['Year']).value\_counts().reset\_index().sort\_values('index')

nations\_over\_time.rename(columns={'index':'Edition','Year': col},inplace=True)

return nations\_over\_time

def most\_successful(df, sport):

temp\_df = df.dropna(subset=['Medal'])

if sport != 'Overall':

temp\_df = temp\_df[temp\_df['Sport'] == sport]

x = temp\_df['Name'].value\_counts().reset\_index().head(15).merge(df, left\_on='index', right\_on='Name', how='left')[

['index', 'Name\_x', 'Sport', 'region']].drop\_duplicates('index')

x.rename(columns = {'index': 'Name', 'Name\_x': 'Medals'}, inplace=True)

return x

def yearwise\_medal\_tally(df,country):

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'] == country]

final\_df = new\_df.groupby('Year').count()['Medal'].reset\_index()

return final\_df

def country\_event\_heatmap(df,country):

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'] == country]

pt = new\_df.pivot\_table(index='Sport', columns='Year', values='Medal', aggfunc='count').fillna(0)

return pt

def most\_successful\_countrywise(df, country):

temp\_df = df.dropna(subset=['Medal'])

temp\_df = temp\_df[temp\_df['region'] == country]

x = temp\_df['Name'].value\_counts().reset\_index().head(10).merge(df, left\_on='index', right\_on='Name', how='left')[

['index', 'Name\_x', 'Sport']].drop\_duplicates('index')

x.rename(columns={'index': 'Name', 'Name\_x': 'Medals'}, inplace=True)

return x

def weight\_v\_height(df,sport):

athlete\_df = df.drop\_duplicates(subset=['Name', 'region'])

athlete\_df['Medal'].fillna('No Medal', inplace=True)

if sport != 'Overall':

temp\_df = athlete\_df[athlete\_df['Sport'] == sport]

return temp\_df

else:

return athlete\_df

def men\_vs\_women(df):

athlete\_df = df.drop\_duplicates(subset=['Name', 'region'])

men = athlete\_df[athlete\_df['Sex'] == 'M'].groupby('Year').count()['Name'].reset\_index()

women = athlete\_df[athlete\_df['Sex'] == 'F'].groupby('Year').count()['Name'].reset\_index()

final = men.merge(women, on='Year', how='left')

final.rename(columns={'Name\_x': 'Male', 'Name\_y': 'Female'}, inplace=True)

final.fillna(0, inplace=True)

return final

**Data manipulation**

**Preprocessor.py**

import pandas as pd

# Preprocessing the Dataset

# Also use summer olympic Datasets

def preprocess(df,region\_df):

# filtering for summer olympics

df = df[df['Season'] == 'Summer']

# merge with region\_df

df = df.merge(region\_df, on='NOC', how='left')

# dropping duplicate

df.drop\_duplicates(inplace=True)

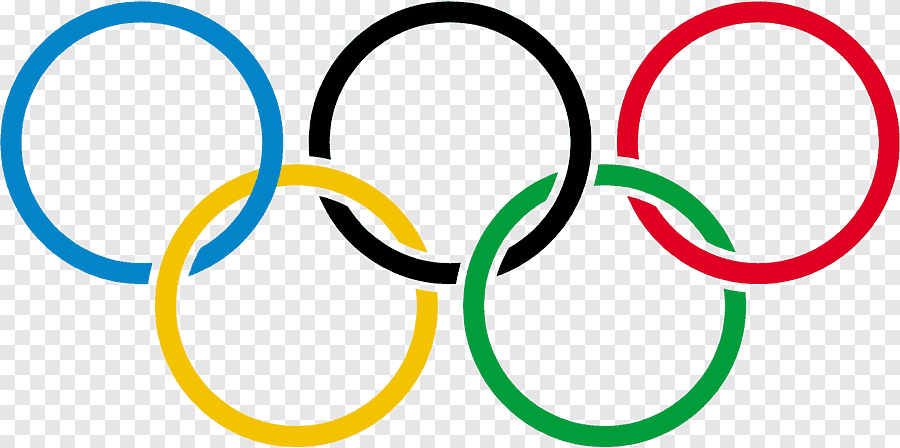
# one hot encoding medals

df = pd.concat([df, pd.get\_dummies(df['Medal'])], axis=1)

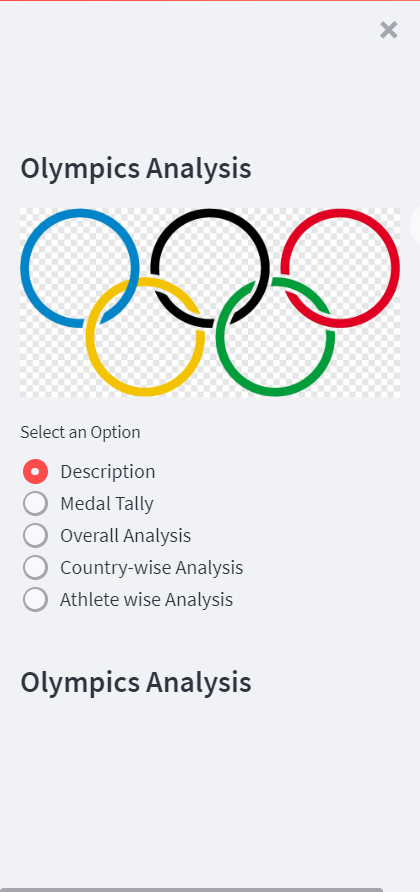
return df

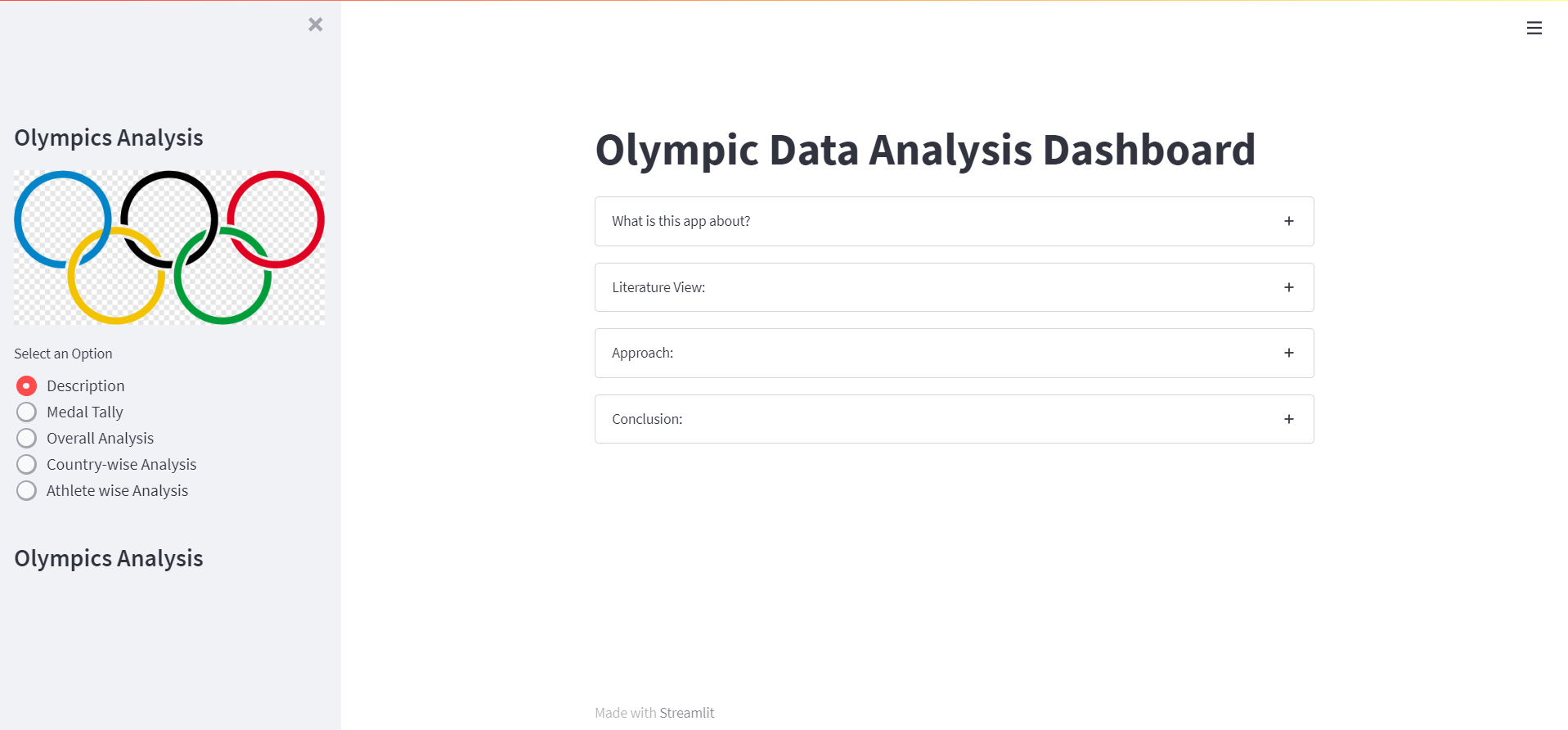
**SCREEN**

**LOGO:**

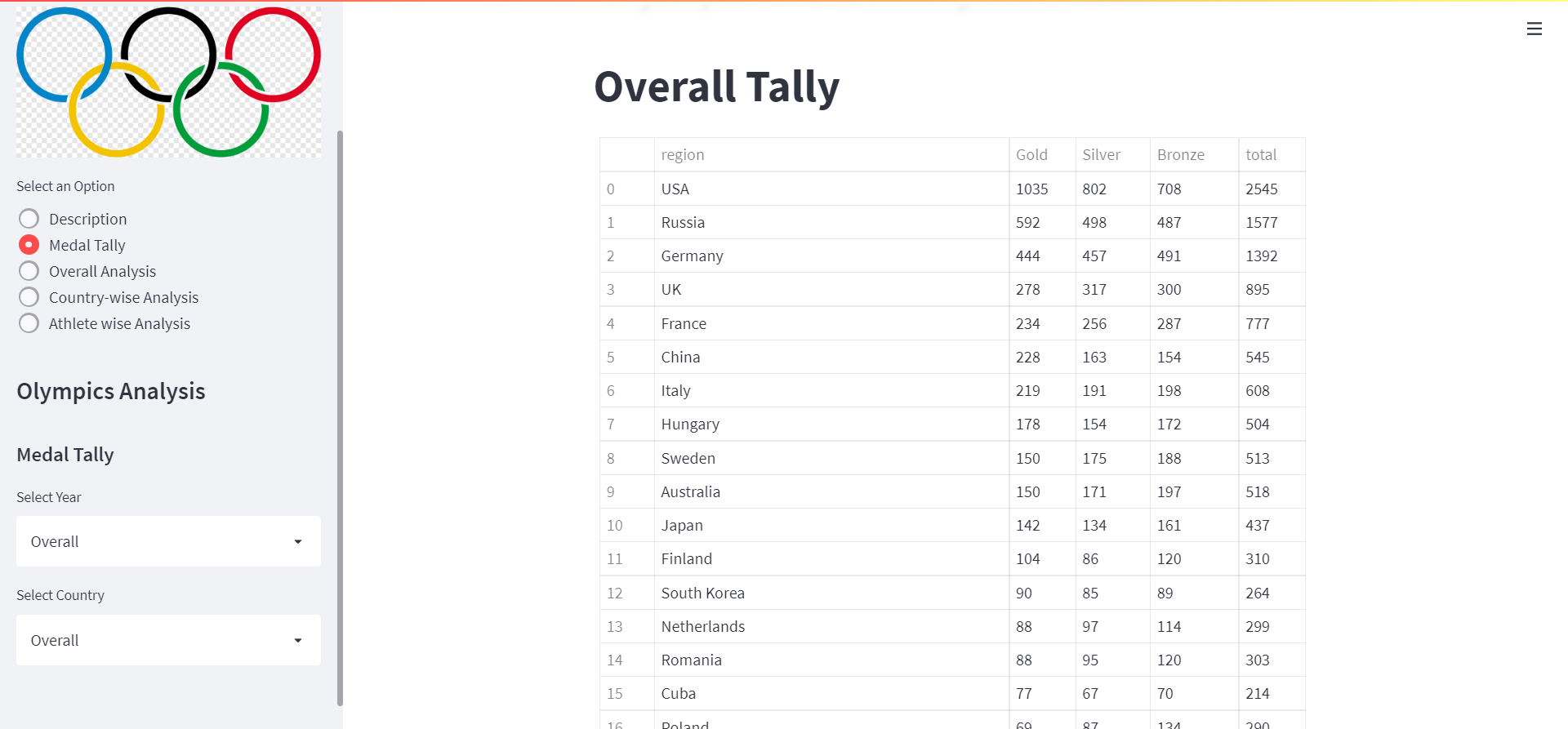


**SELECT AN OPTION:**

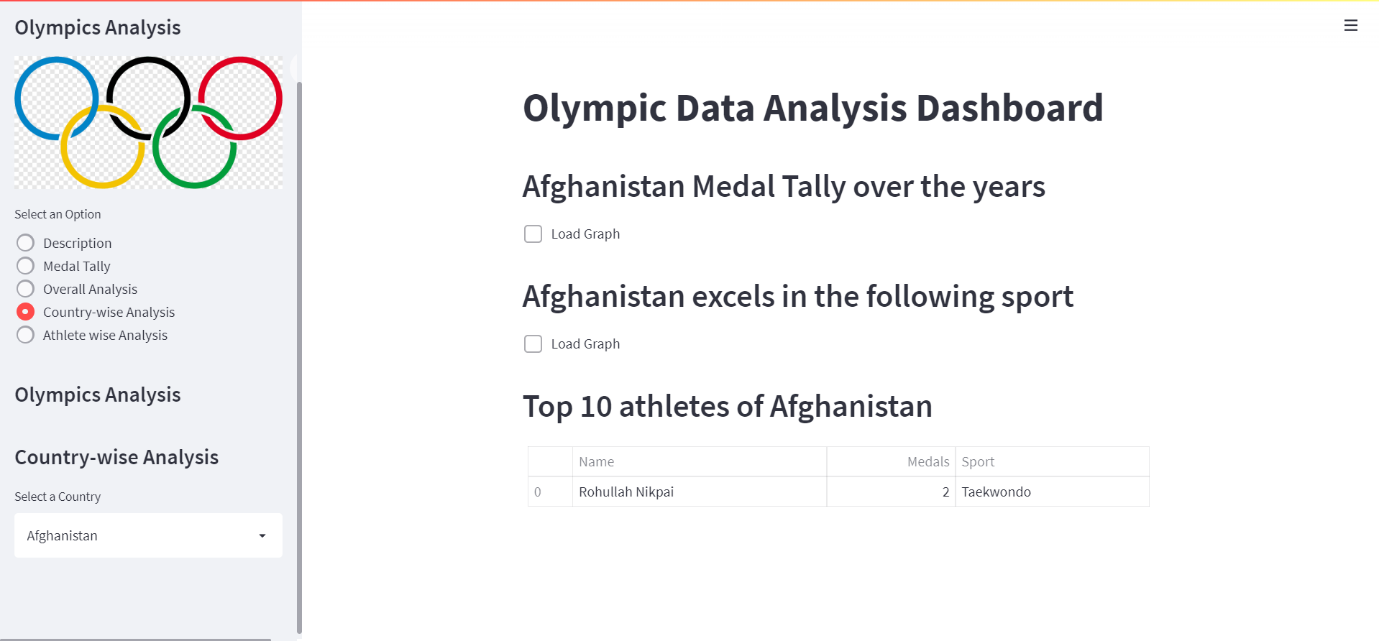
****

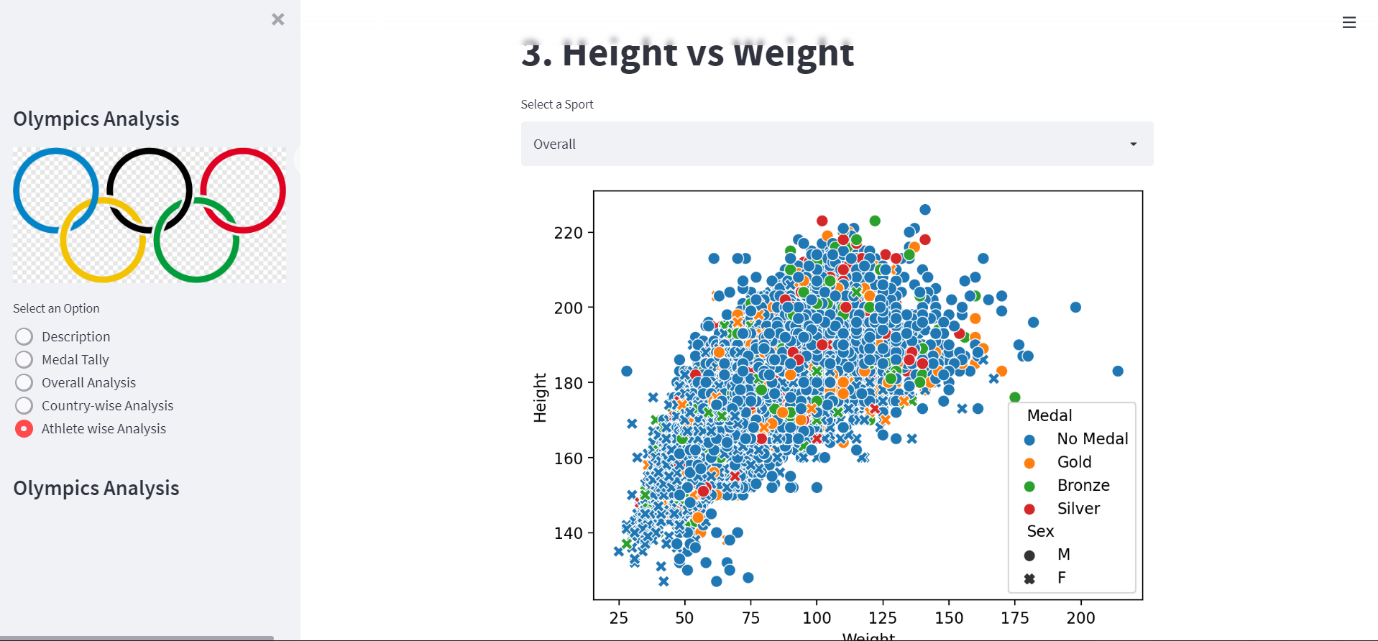
**DESCRIPTION:**

**MEDAL TALLY**:

****

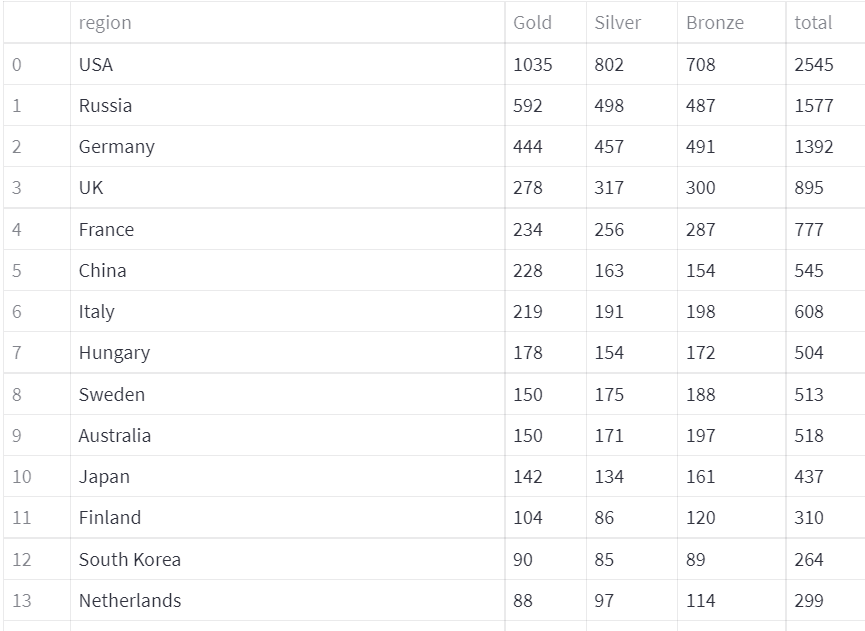
**COUNTRY-WISE:**

****

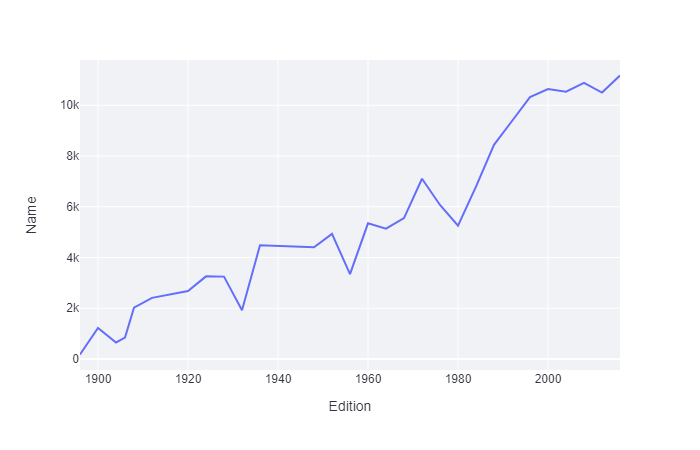
**ATHLETE-WISE:**

**DATA VISUALISATION**

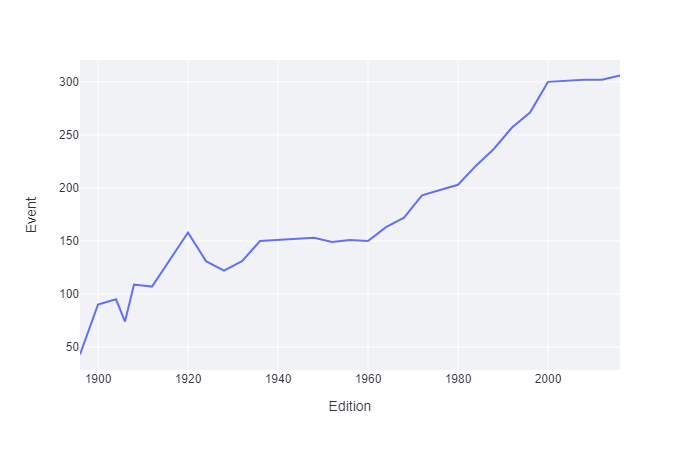
**OVERALL TALLY:**

****

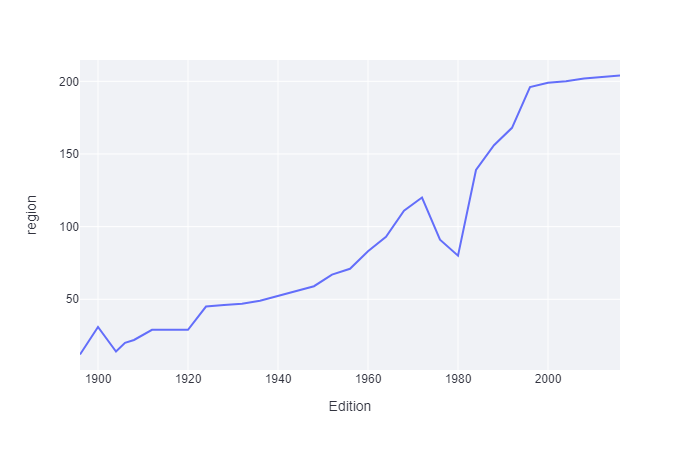
1. **Participated Nations Over the years:**

****

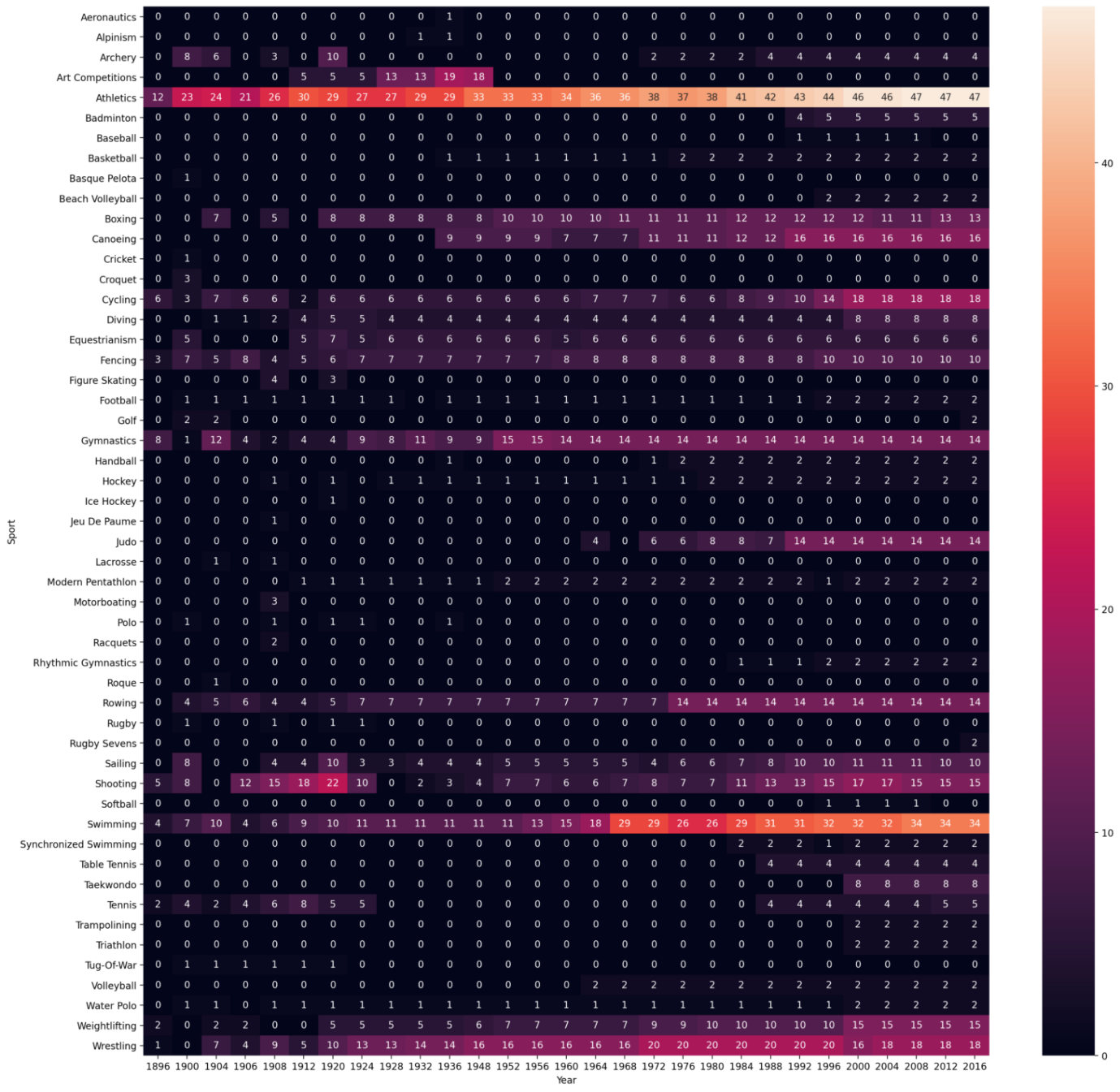
1. **Events Over the years:**

****

1. **Athletes Over the years:**

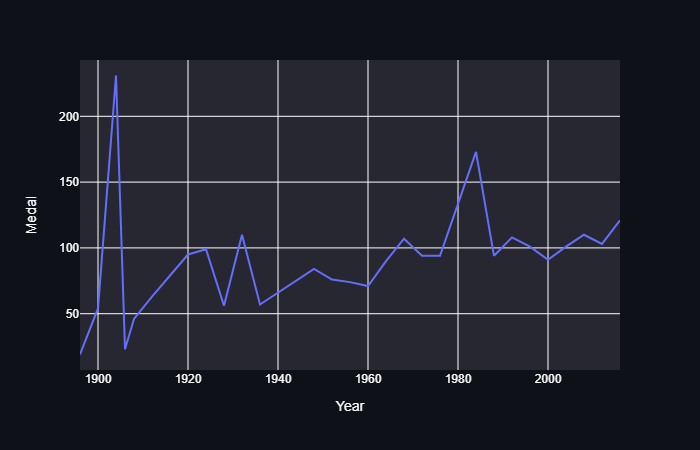
****

1. **No. of Events over time ( Every Sport):**

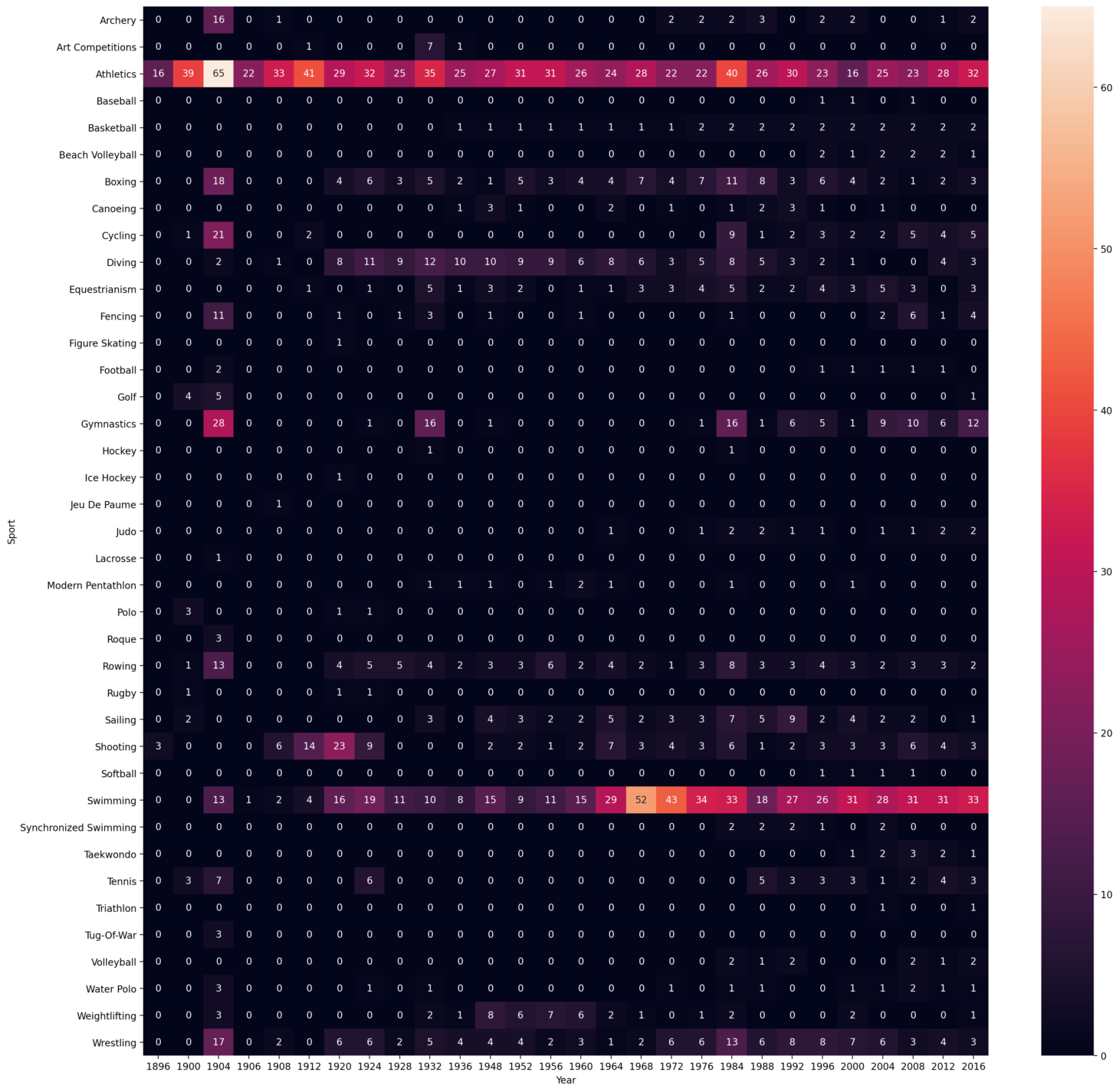
****

**COUNTRY WISE ANALYSIS**

1. **Country Medal Tally Over the years:**

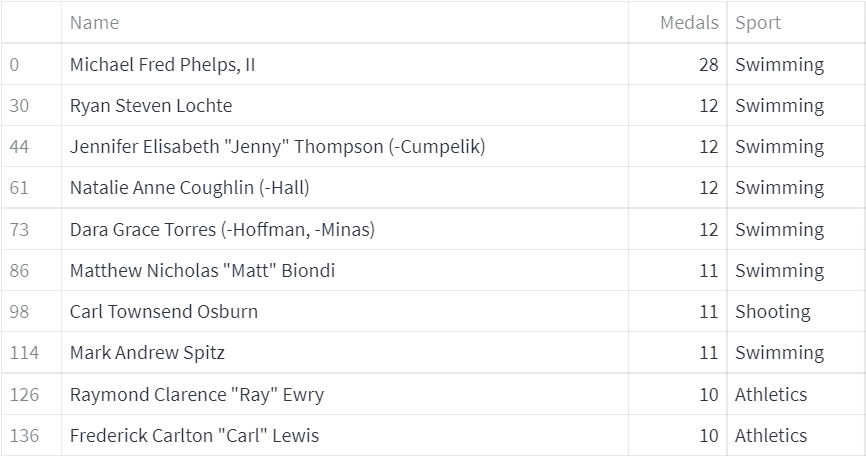
**** Note: Country’s Graph changes according to country

1. **Country Excel in the following sport:**

Note: Country’s Graph changes according to ****sport

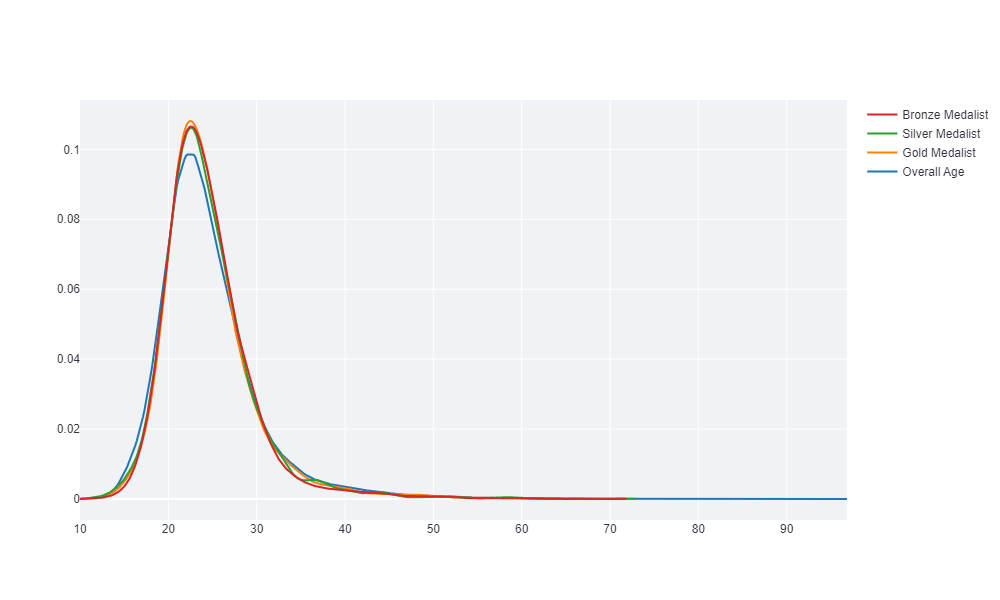
1. **Top 10 athletes of country:**

Note: Country’s Graph changes according to athlete

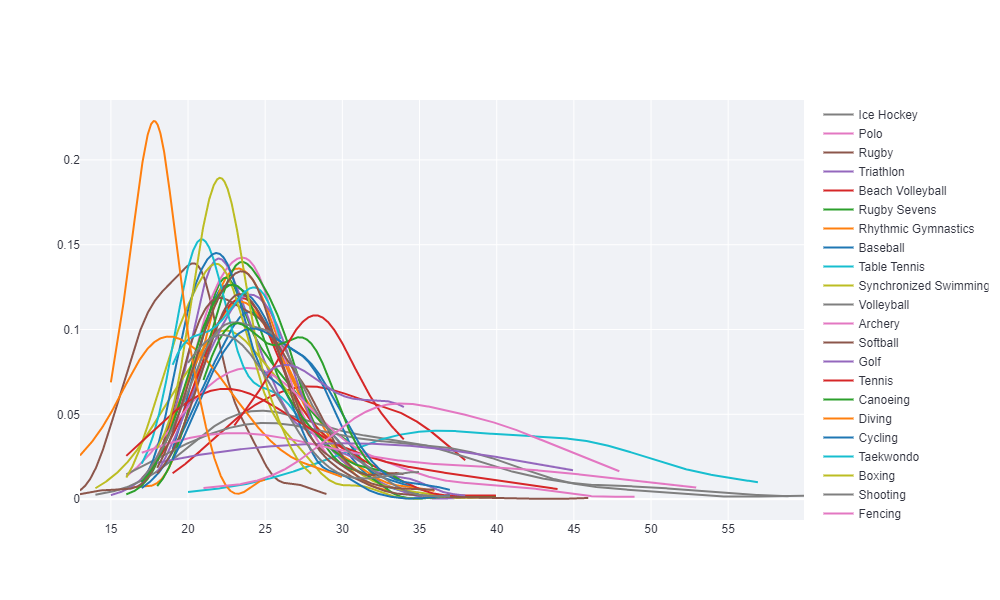
****

**ATHLETE WISE ANALYSIS**

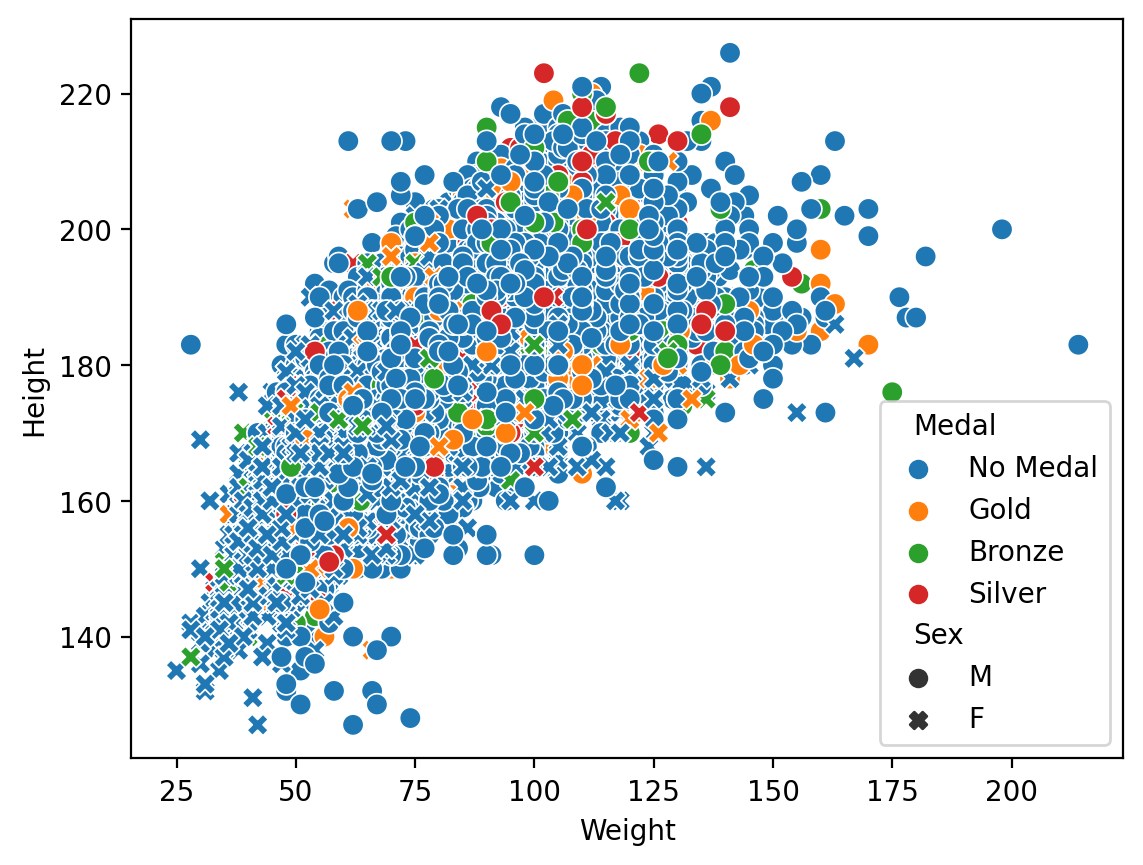
1. **Distribution of Age:**

****

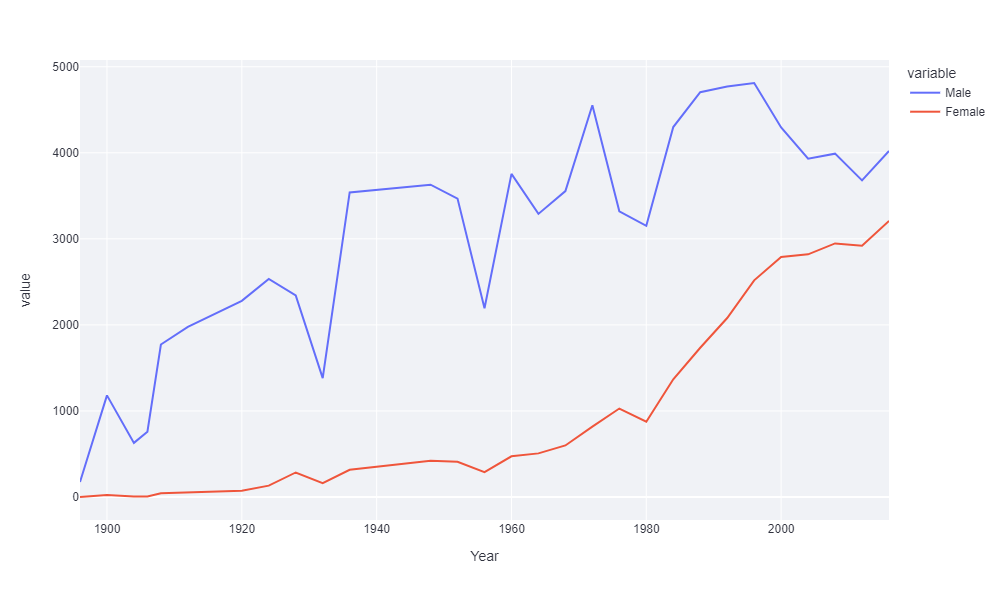
1. **Distribution of Age wrt Sports(Gold Medalist):**

****

1. **Height vs Weight:**

****

1. **Men vs Women Participation Over the Years:**

****

**MAINTENANCE**

Maintenance is the enigma of the system development analyst’s and programmer’s spent far more time on maintaining the program than any do writing them.

1. Maintenance over the wide range of activities, including correcting coding and design errors updating documentation rest data and updating user support.
2. Web app can be stored to it’s original position by software maintenance
3. Maintenance, which is also referred to as web maintenance, means adding, modifying or developing the code to support changes in the specification.
4. The dashboard developed will need proper monitoring so that any or

unexpected result does not lead to the system crash when the system will be implemented there may be condition which might have been overlooked by the user during testing. For the system to run efficiently weekly evaluation of result set will be done and the user will be consulted for the system performance.

If there is any rectification or enhancement posted by the user, the software will be modified and the desired modules if added will be implemented.

**FUTURE SCOPE OF THE PROJECT**

In order to get rid of the shortcomings in the developed system, future modification with regard to the program and management requirement can be made.

To increase the efficiency & functionality of the developed system, the earlier discussed drawbacks are to be eradicated by including relevant programmed module.

The system can be modified to include interactive reports, which would have been used by the user. The interactive report which would return the count of copies of the same title.

**CONCLUSION**

It clear that the host countries have alwaya a better chance of winning medal in the Olympics; they can win at least 10-20 percent more medals. Looking at the economic effect, even though country’s population and per capita GDP the number of medals won in past,

the total GDP of the country is more significant to de significant to determine the winnings in the recent years.

With the age factor, the age range of players winning medals has decreased over the years, and an optimal age for each sport can be identified in the recent years.

Thus, there is a high chance for an athlete from a host country with high GDP, whose age range falls in the optimum age range for the sport to win a medal in the Olympics.

Also, the most of the gold medalist are mostly higher than the other athletes and theire weight is also in less than the other competitor.

In The Olympic games womens participation increases over the year it is good sign for women percepective as well.

**LIMITATIONS**

The developed application though satisfies the desired objective but still has few limitations, which one cannot ignore. These limitations could have been eliminated, but due to time constraint and the restricted scope of this project still persists. The limitations of this application are discussed as under:

* We must take notice that this system deals with only stationary products
* Anyone Can use it.

**BIBLIOGRAPHY**

* **The self – taught Programmer**

**Cory Althoff**

* **Data smart\_using data science to transform information into sight**

**Foreman, John W**

* **Python for Data Analysis(2017, O’Reilly Media Inc)**

**Mckenney, Wes**

* **Python Automation Cookbook\_75 Python automation ideas for Web scraping, Data wrangling**

**Jaima buelta**