



TEAM VISUAL SPORTS

CSE3020 – DATA VISUALISATION

PROJECT-BASED COMPONENT REPORT

By

Pratham Agarwal – 20BDS0142

Ronit Gehani – 20BCE2006

Dhairya Lunia – 20BCE2178

A Kajal Patro – 20BCE0736

School of Computer Science and Engineering

April 2022

DECLARATION

I hereby declare that the report entitle **“Visual Sports”** submitted by me, for the CSE3020 DATA VISUALISATION (EPJ) to VIT is a record of Bonafide work carried out by me under the supervision of Dr. S. VENGADESWARAN.

I further declare that the work reported in this report has not been submitted and will not be submitted, either in part or in full, for any other courses in this institute or any other institute or university.

Place : VIT Vellore

Date :29th April 2022

A handwritten signature in blue ink, appearing to read 'Dhairya', is positioned above the signature line.

Signature of the Candidate

CONTENTS

	P.No
1. ABSTRACT	1
2. INTRODUCTION TO THE PROJECT <ul style="list-style-type: none">• OBJECTIVE• PROBLEM STATEMENT• FUNCTIONAL REQUIREMENTS	2
3. DATA ABSTRACTION	3-5
4. TASK ABSTRACTION	6
5. DASHBOARD IMPLEMENTATION	7
6. RESULT ANALYSIS	8-12
7. CONCLUSION	13
8. APPENDIX <ul style="list-style-type: none">• SCREEN SHOTS• SAMPLE CODING	18-23

1. ABSTRACT

Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.

The 'modern Olympics' comprises all the Games from Athens 1986 to Tokyo 2020. The Olympics is more than just a quadrennial multi-sport world championship. It is a lens through which to understand global history, including shifting geopolitical power dynamics, women's empowerment, and the evolving values of society.

In this kernel, our 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? Which country wins the most medals? Which age group bags the most medals? What percentage of women win medals? What is the ratio of male over female in different years over summer and winter olympics ?

This project is aimed at analyzing Summer / Winter Olympics data. While a dataset was readily available for the initial 120 years of Olympics on Kaggle, no data was provided collectively for the 2020 Tokyo Olympics. So for that particular reason, We have created a dataset to incorporate all year ranges.

Future works include making a visualization dashboard and understanding data trends to a deeper level.

Data Set Used for Olympics from Athens 1986 to Rio 2016

<https://www.kaggle.com/heesoo37/120-years-of-olympic-history-athletes-and-results>

Data Set Used for Olympics Tokyo 2020

<https://www.kaggle.com/arjunprasadsarkhel/2021-olympics-in-tokyo>

Final Dataset:

After compiling and editing all the above dataset. The dataset contains 271116 rows and 15 columns. Each row corresponds to an individual athlete competing in an individual Olympic event (athlete-events).

<https://www.kaggle.com/datasets/fearsomejockey/olympics-dataset-2020-tokyo-dataset>

2. INTRODUCTION TO THE PROJECT

A. Objective:

Development of Visual Idioms to understand the Historical 120 years of Olympic sports and players

B. Problem Statement:

The Olympic Games, considered to be the world's foremost sports competition have more than 200 nations participating in the Summer and Winter Games alternating by occurring every four years but two years apart.

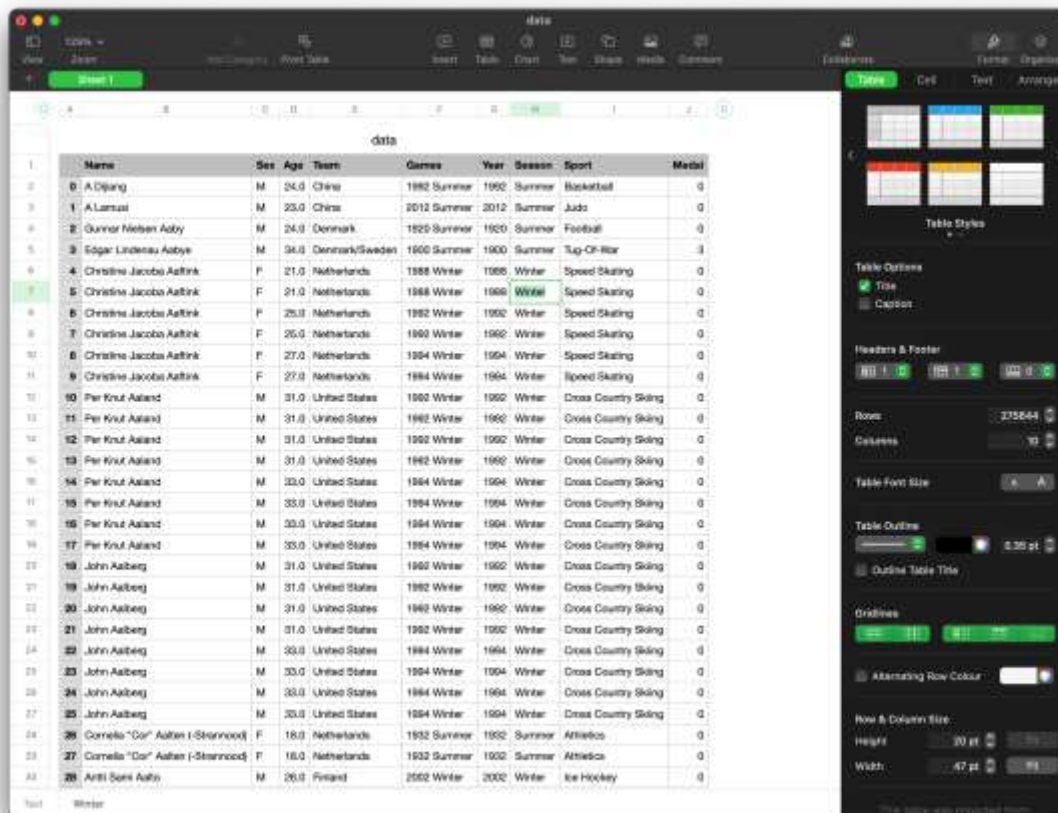
Throughout this project, we will explore the Olympics dataset consisting of sports, countries and athletes. We will look at some interesting visual idioms, and statistics and then try to compare the trends and countries who have dominated the Olympics, it will also focus on how countries have developed their participation and achieved medals with time and different seasons of the Olympics.

C. Functional Requirements:

The given project is required to provide ease in understanding the Olympics trends and comparisons between statistics of all Sports and nations by the use of a Dashboard Implementation made on the Python Jupyter Notebook. The use of appropriate Marks and Channels and a proper design is intended for an enjoyable and worthwhile experience for the user. The design idioms implementations would be made using Python language.

3. DATA ABSTRACTION

The chosen data set has the following attributes



The screenshot shows a data visualization tool interface. The main area displays a table with the following columns: ID, Name, Sex, Age, Team, Games, Year, Season, Sport, and Medal. The table contains 28 rows of data. The sidebar on the right includes options for Table Styles, Table Options (Title, Caption), Headers & Footer, Rows (275644), Columns (10), Table Font Size, Table Outline, Gridlines, Alternating Row Colour, and Row & Column Size (Height: 30 pt, Width: 47 pt).

ID	Name	Sex	Age	Team	Games	Year	Season	Sport	Medal
0	A. Djang	M	24.0	China	1992 Summer	1992	Summer	Basketball	0
1	A. Lamsal	M	23.0	China	2012 Summer	2012	Summer	Judo	0
2	Gunnar Nielsen Asby	M	24.0	Denmark	1920 Summer	1920	Summer	Football	0
3	Eggar Lindensau Asbye	M	24.0	Denmark/Sweden	1900 Summer	1900	Summer	Tug-Of-War	0
4	Christine Jacobsa Aalfink	F	21.0	Netherlands	1988 Winter	1988	Winter	Speed Skating	0
5	Christine Jacobsa Aalfink	F	21.0	Netherlands	1988 Winter	1988	Winter	Speed Skating	0
6	Christine Jacobsa Aalfink	F	25.0	Netherlands	1992 Winter	1992	Winter	Speed Skating	0
7	Christine Jacobsa Aalfink	F	25.0	Netherlands	1992 Winter	1992	Winter	Speed Skating	0
8	Christine Jacobsa Aalfink	F	27.0	Netherlands	1994 Winter	1994	Winter	Speed Skating	0
9	Christine Jacobsa Aalfink	F	27.0	Netherlands	1994 Winter	1994	Winter	Speed Skating	0
10	Per Knut Aaland	M	31.0	United States	1960 Winter	1960	Winter	Cross Country Skiing	0
11	Per Knut Aaland	M	31.0	United States	1962 Winter	1962	Winter	Cross Country Skiing	0
12	Per Knut Aaland	M	31.0	United States	1962 Winter	1962	Winter	Cross Country Skiing	0
13	Per Knut Aaland	M	31.0	United States	1962 Winter	1962	Winter	Cross Country Skiing	0
14	Per Knut Aaland	M	33.0	United States	1964 Winter	1964	Winter	Cross Country Skiing	0
15	Per Knut Aaland	M	33.0	United States	1964 Winter	1964	Winter	Cross Country Skiing	0
16	Per Knut Aaland	M	33.0	United States	1964 Winter	1964	Winter	Cross Country Skiing	0
17	Per Knut Aaland	M	33.0	United States	1964 Winter	1964	Winter	Cross Country Skiing	0
18	John Aalberg	M	31.0	United States	1960 Winter	1960	Winter	Cross Country Skiing	0
19	John Aalberg	M	31.0	United States	1962 Winter	1962	Winter	Cross Country Skiing	0
20	John Aalberg	M	31.0	United States	1962 Winter	1962	Winter	Cross Country Skiing	0
21	John Aalberg	M	31.0	United States	1962 Winter	1962	Winter	Cross Country Skiing	0
22	John Aalberg	M	33.0	United States	1964 Winter	1964	Winter	Cross Country Skiing	0
23	John Aalberg	M	33.0	United States	1964 Winter	1964	Winter	Cross Country Skiing	0
24	John Aalberg	M	33.0	United States	1964 Winter	1964	Winter	Cross Country Skiing	0
25	John Aalberg	M	33.0	United States	1964 Winter	1964	Winter	Cross Country Skiing	0
26	Cornelia "Cor" Aalton (Stenrood)	F	18.0	Netherlands	1932 Summer	1932	Summer	Artistic	0
27	Cornelia "Cor" Aalton (Stenrood)	F	18.0	Netherlands	1932 Summer	1932	Summer	Artistic	0
28	Ahti Sieni Aalto	M	26.0	Finland	2002 Winter	2002	Winter	Ice Hockey	0

All About Data

Attribute	Representing
ID	Specifies the Unique number for each athlete
Name	Specifies the Athlete's name
Sex	States the Sex either Male or Female
Age	Specifies the Age of the athlete in integer
Team	Describes the Team name
Games	Describes the Year and season
Year	Specifies the Year of olympics
Season	Specifies the season Summer or Winter
Sport	Specifies the sports participated in
Medal	Specifies the type of Medal won Gold, Silver, Bronze, or NA

- The used Dataset is of Tables type consisting of Items and Attributes (described above).
- There are 271116 items in the table.
- The Semantics for the dataset is described above.
- The used dataset is of Tables type consisting of Items and Attributes.

Classification of Attributes

Attributes	Attribute Type	Level of Measurement	Type of Quantitative data
ID	Quantitative	Interval	Discrete
Name	Qualitative	Nominal	-
Sex	Qualitative	Ordinal	Discrete
Age	Quantitative	Interval	-
Team	Qualitative	Nominal	-
Games	Qualitative	Ordinal	-
Year	Quantitative	Interval	Continuous
Season	Qualitative	Nominal	-
Sport	Quantitative	Nominal	-
Medal	Qualitative	Nominal	-

So all in all we see that there are 6 qualitative and 4 quantitative attribute types and the level of measurements of these attributes include 5 nominal, 2 ordinal, 3 Interval types of measurement where all the data is unordered

4. TASK ABSTRACTION

1.Actions

1.Analyze

1. Consume -> While discovering a relation between Gender, Age, Country, and medals won of an Athlete can be made, and also presented the data in an easy to visualize way.
2. Produce -> Records of new Olympics data like the Tokyo Olympics 2020 can be incorporated into the dataset and stored in an appropriate manner.
3. Enjoy -> Users/ Clients can use the dataset to understand and enjoy the different trends in the Olympics dataset.

2. Search

1. Perform various types of searches for athletes, countries in terms of how many medals they have won so far, or how many times have they hosted the Olympics so far.

3. Actions

1. Compare -> Compare 2 or more countries for their medals won, athletes participated, host countries, age comparison between winners, medals won over the years.

4. Target

1. All Data -> A relation trend can be observed between the number of medals won and the respective countries over the years. Several Outliers can be drawn for the scatter plot of the height and weight of athletes.
2. Attributes -> A positive correlation can be seen and observed between number of medals athletes won and small countries.

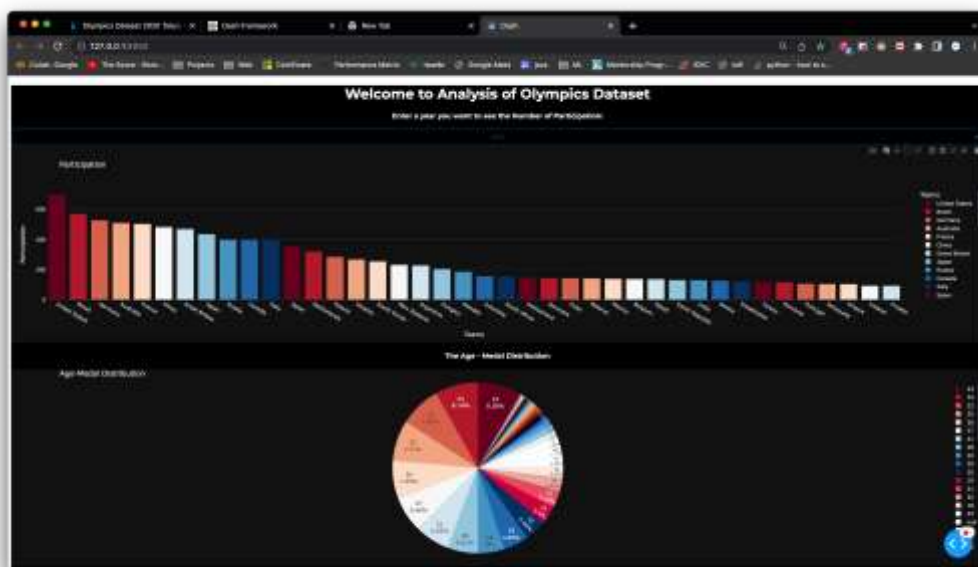
5. DASHBOARD IMPLEMENTATION

The Following dashboard has been created with the help of the python dashboard library “Dash”. Dash apps give a point-&-click interface to models written in Python, vastly expanding the notion of what's possible in a traditional “dashboard.”

Dash is a powerful library that simplifies the development of data-driven applications. It's especially useful for Python data scientists who aren't very familiar with web development. Users can create amazing dashboards in their browser using dash.

Built on top of Plotly.js, React, and Flask, Dash ties modern UI elements like dropdowns, sliders and graphs directly to your analytical Python code. Dash apps consist of a Flask server that communicates with front-end React components using JSON packets over HTTP requests.

A couple of Screenshots would show how the dashboard looks like after selecting some values for Interactive Graphs.

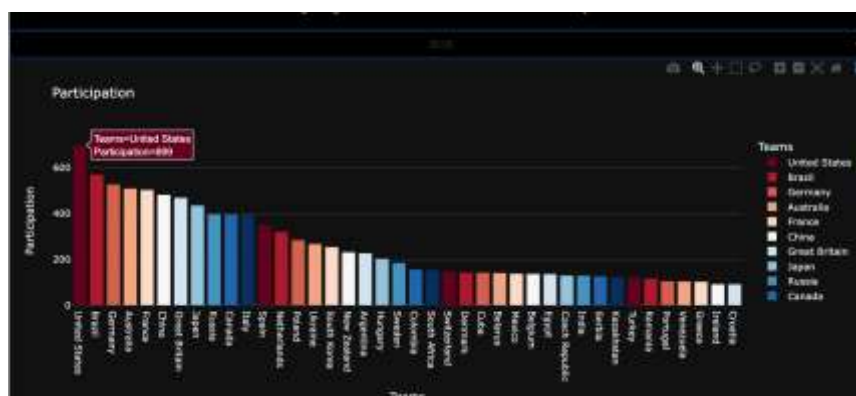
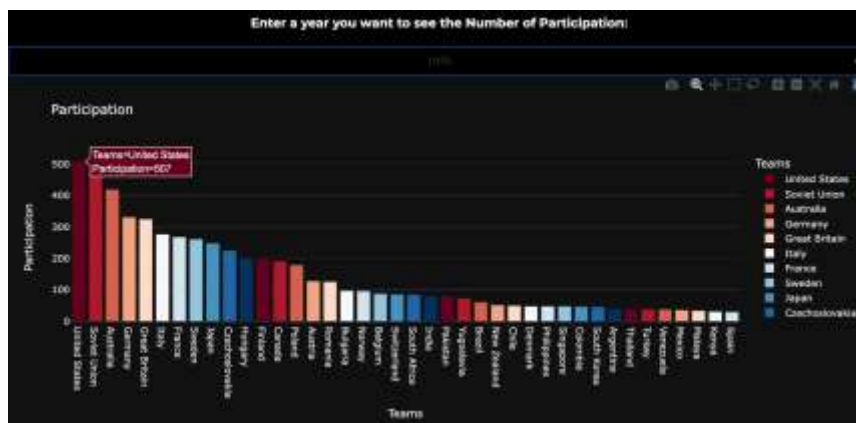


6. RESULT ANALYSIS

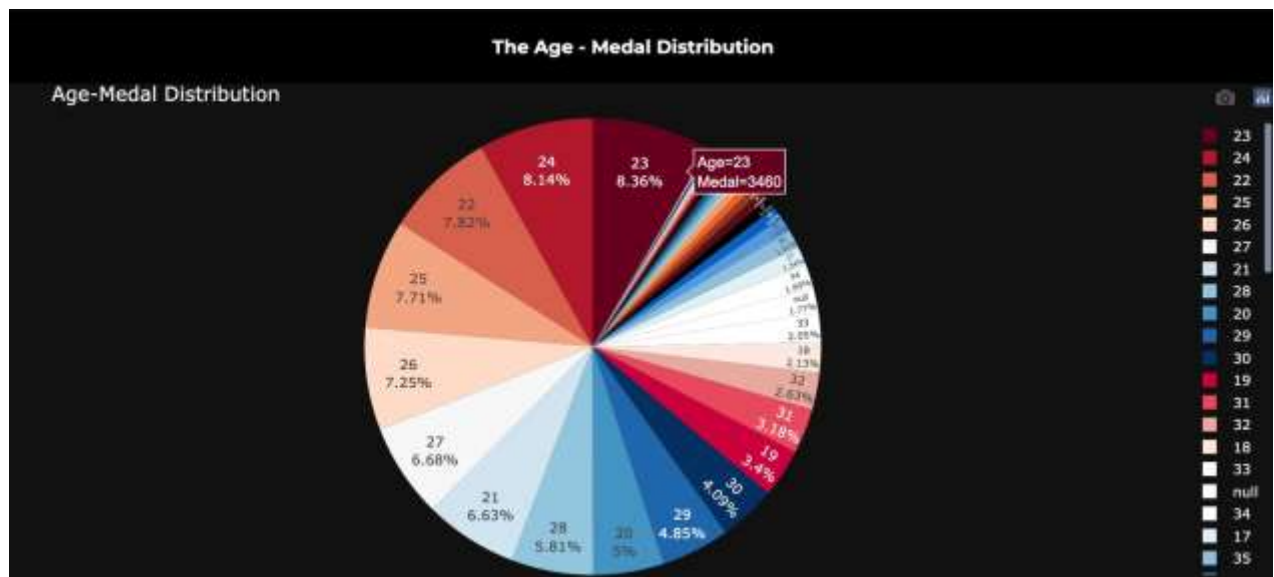
As the result of the Analysis, we can conclude that it is true that Olympic Games have evolved considerably over time from the 1896 Olympics Games till the 2020 Tokyo Olympics. There are various factors which provide the valid evidence that the Olympics have changed a lot some of these factors are the launch of the Winter Olympic Games apart from the Summer Olympic Games in 1924, the Increase in the number of participating countries in both Summer and Winter Olympics, 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 of Players who contributes to victory of Games in the event. Apart from these, there are many more factors which depict the Evolution of the Olympic Games over time. Visualization of these factors has been done to explain and validate the Analysis in various Graphical formats like Line graphs, Bar Graphs, Tree Maps etc.

Inferences from the Graphs->

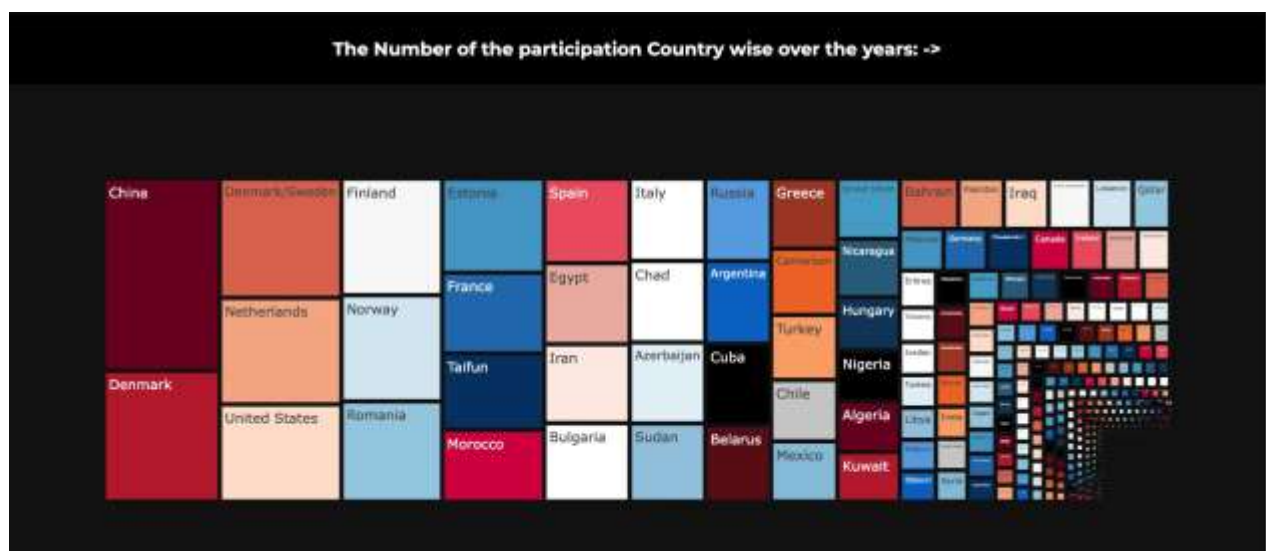
- From the first graph, we can infer that irrespective of the teams, the total participation for all participants have increased. Different teams have been dissolved and new teams have come up, old teams have increased their participation. We can infer using these graphs shown below.



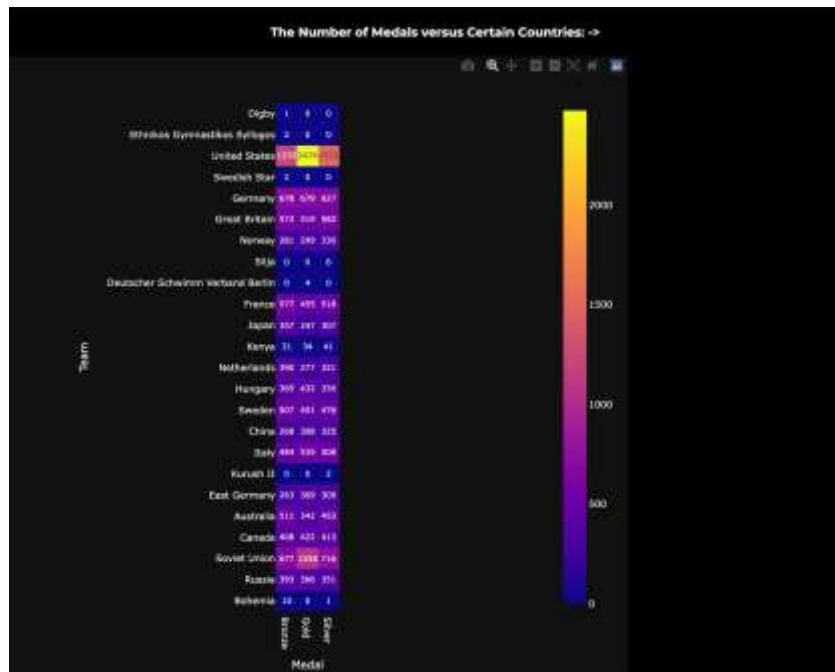
- The age medal distribution pie chart helps us understand which age group has won the most amount of medals. We can infer that the age of 23 has won the most amount of medal that is 3460.



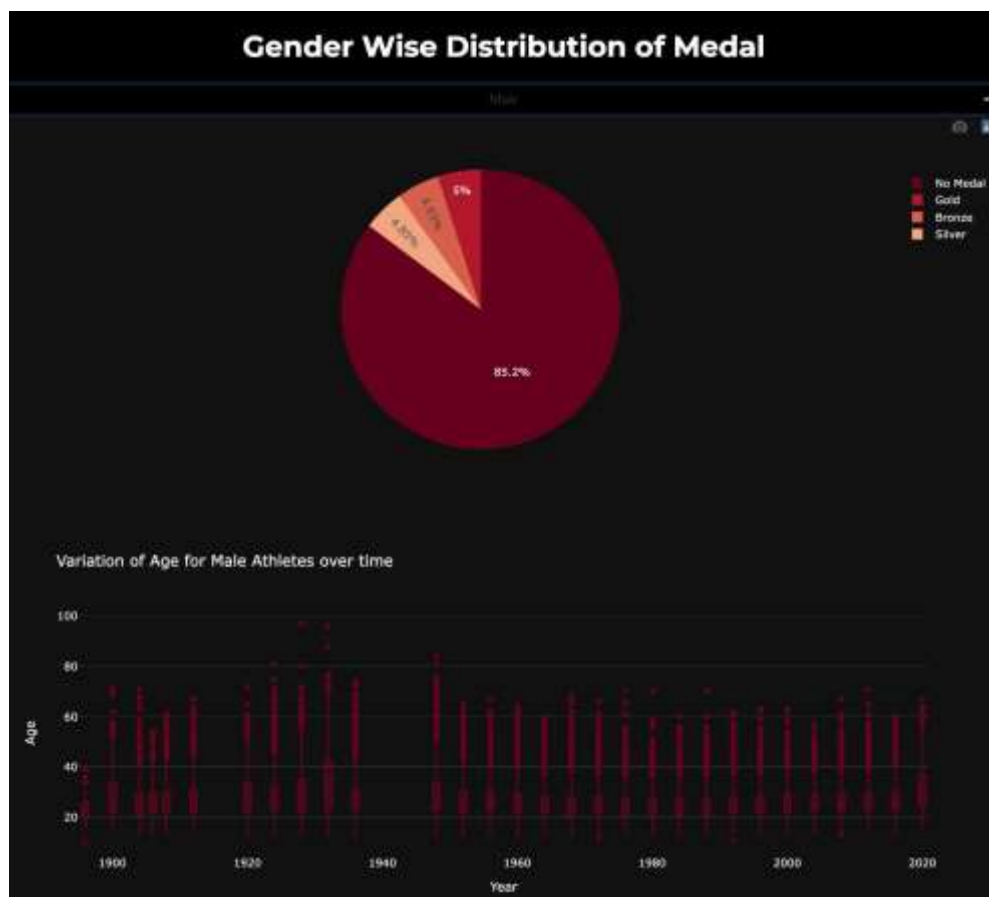
- The Treemap gives a conclusion about the Total participation by country over the years 1896 - 2020. The maximum participation has come from China.



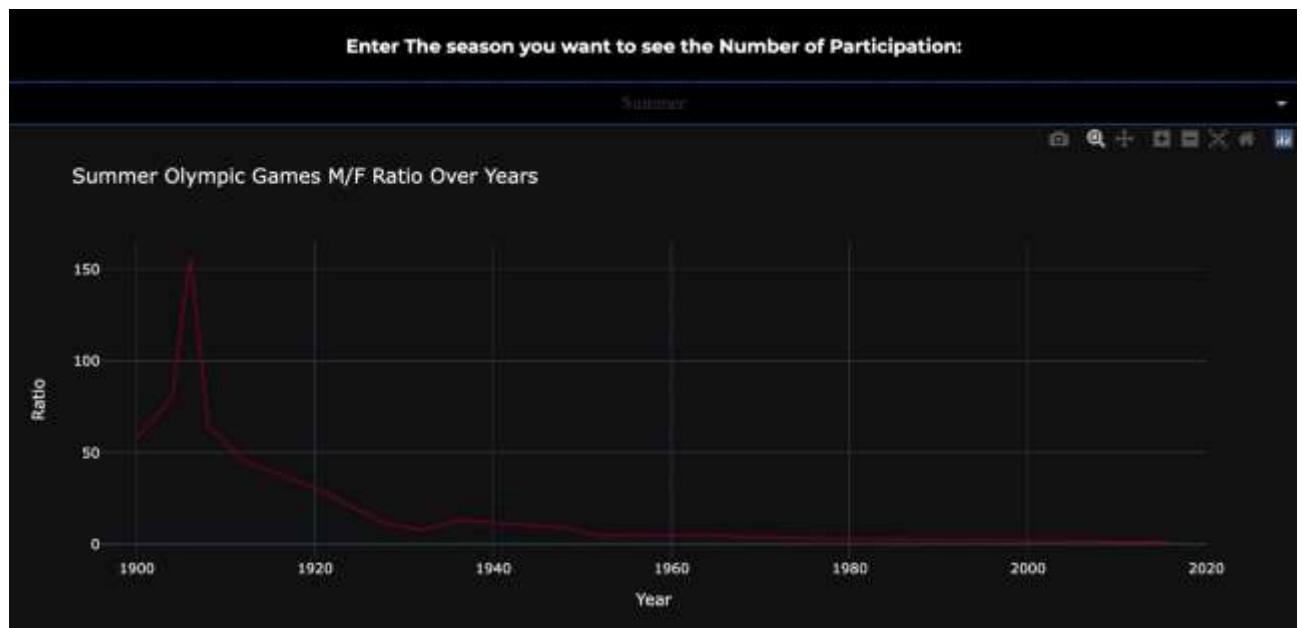
- The number of medals versus certain countries is represented by a heatmap.



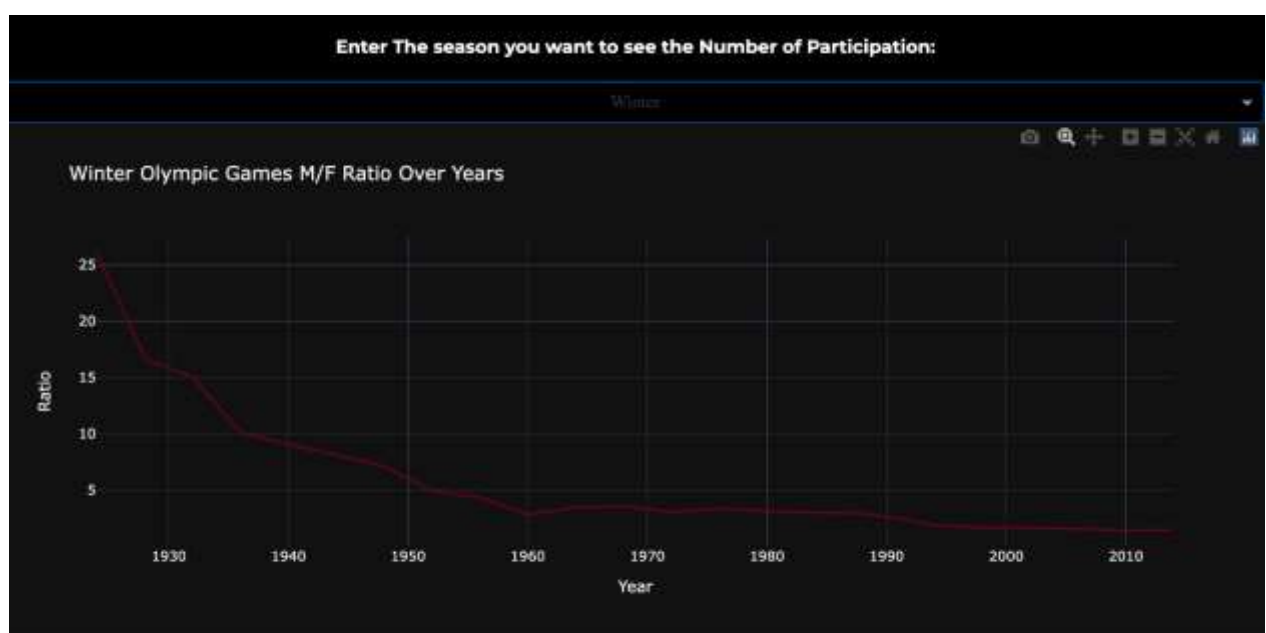
- The following 2 graphs are interrelated and change overusing one dropdown. The options are "Male" and "Female". The following charts show how each gender performed in terms of how many medals have they won, and which medals they win.



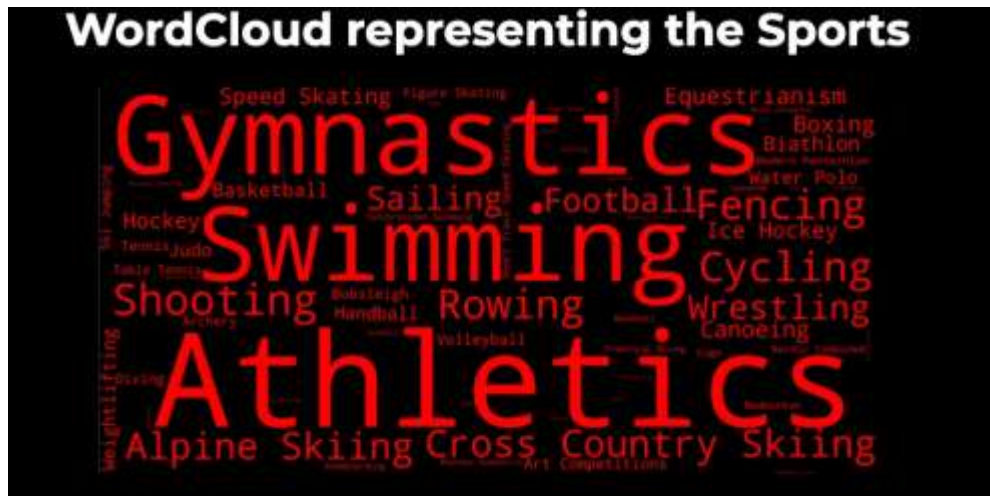
- Male dominance has always been an important factor all around the globe for various events and purposes. The next graph represents the Men/Women Ratio over either Summer/Winter Olympics over the years. We can conclude from this graph that there has been a gradual decrease in the M/F ratio in both the winter and summer Olympics. In 2020, the ratio for Summer was =1.19 which fell from 156% in the year 1906. Whereas for winter, the ratio fell from 26% in 1924 to 1.1 in 2018.
- Summer Graph



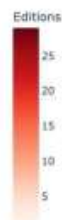
- Winter Olympics Graph



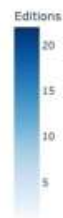
- The following choropleths signify the number of editions each country participated in 2 graphs (One for summer and one for Winter).



Olympic countries (Summer games)



Olympic countries (Winter games)



7. CONCLUSION

The main objective of this study was to analyze and visualize the various factors which have contributed to the Evolution of the Olympic Games over the years. These types of analyses 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.

Lastly, we would also like to mention how this project allowed us to experience team building and working together with peers as we not only coordinated amongst ourselves while working on the dataset but also learned so much about the Data Visualization functions, Dashboards, and the Statistics of Olympics together.

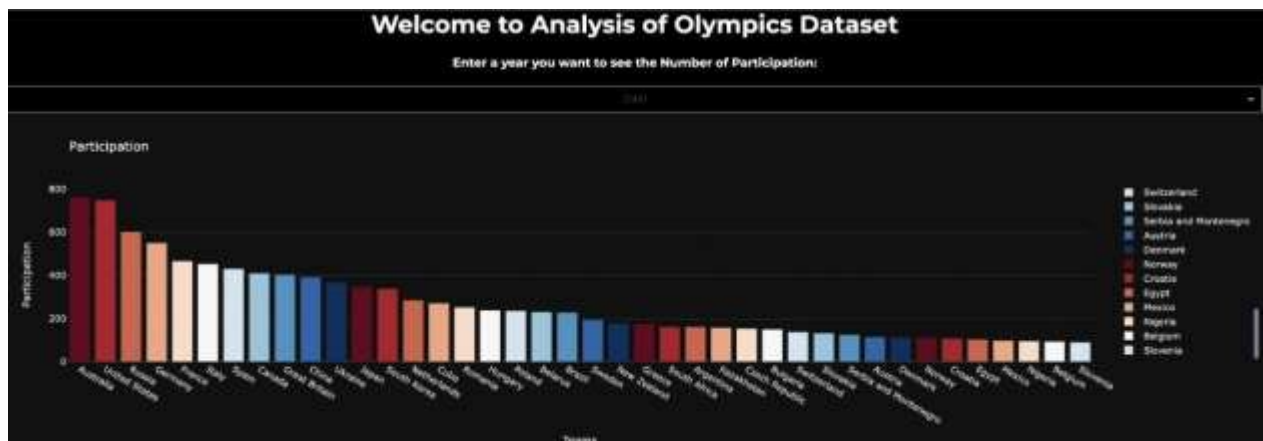
We would like to express our gratitude to Dr. Vengadeswaran S Sir for letting us take up this project and providing us with useful insights and recommendations all through the project without which this project would've never reached a level like this.

8. APPENDIX

1. Sample Output / Graph

Code ->

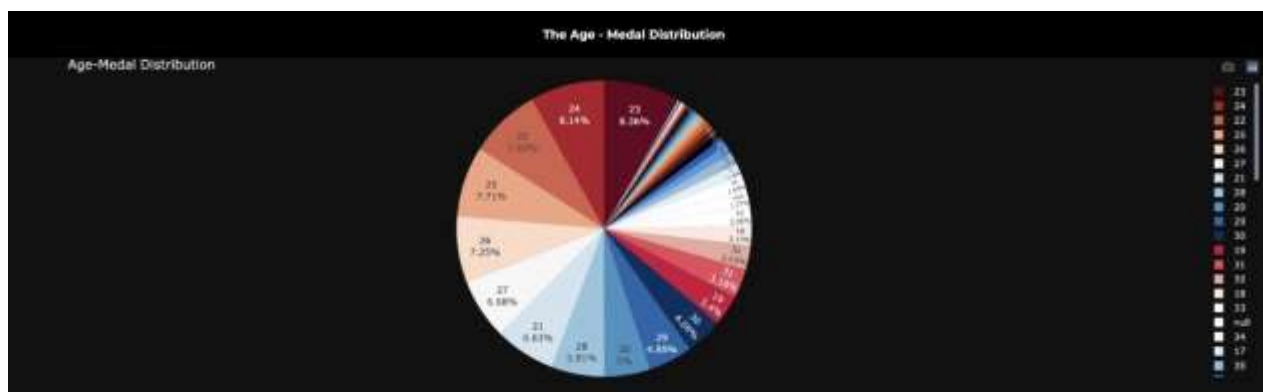
```
# Graph 1
@ app.callback(
    Output("graph1", "figure"),
    Input("dropdown", "value"))
def update_bar_chart(year):
    data_year = data.loc[data['Year'] == year]
    top_countries =
data_year.Team.value_counts().sort_values(ascending=False).head(40)
    indexes = list(top_countries.keys())
    values = top_countries.values
    graph1data = pd.DataFrame({"Teams": indexes,
"Participation": values})
    fig = px.bar(graph1data, x="Teams", y="Participation",
                  title='Participation',
template='plotly_dark', color="Teams",
color_discrete_sequence=px.colors.sequential.RdBu)
    return fig
```



Code ->

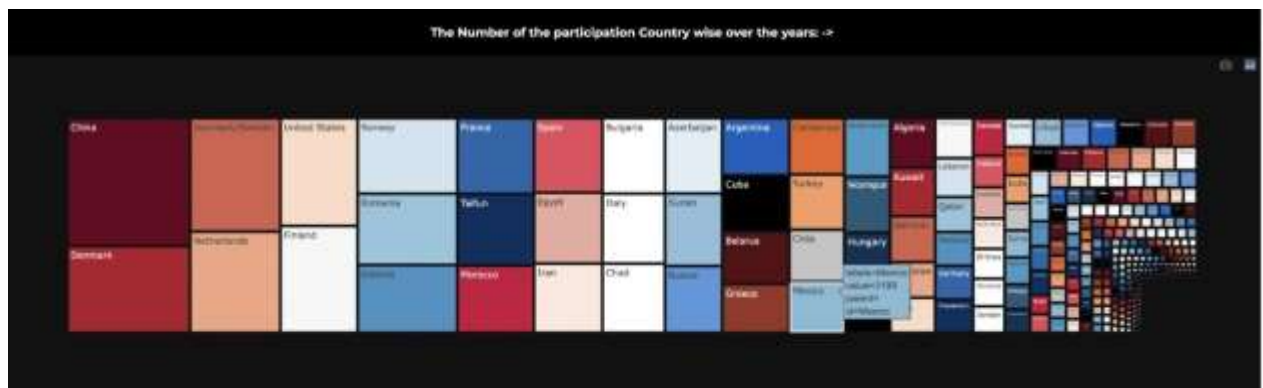
```
# Graph Age_Medal
data_age = list(data.Age)
data_medal = list(data.Medal)
for i in range(len(data_medal)):
    if(data_medal[i] > 0):
        data_medal[i] = 1

data_pie = pd.DataFrame({"Age": data_age, "Medal":
data_medal})
result = data_pie.sort_values('Age')
result.drop(result[result['Age'] > 70].index, inplace=True)
fig_age_medal = px.pie(result, values='Medal', names='Age',
                        title="Age-Medal Distribution",
                        color_discrete_sequence=px.colors.sequential.RdBu,
                        hover_data=['Age'],
                        template="plotly_dark",)
fig_age_medal.update_traces(textposition='inside',
textinfo='percent+label')
fig_age_medal.update_layout(margin=dict(t=30, b=30, l=30,
r=30))
```



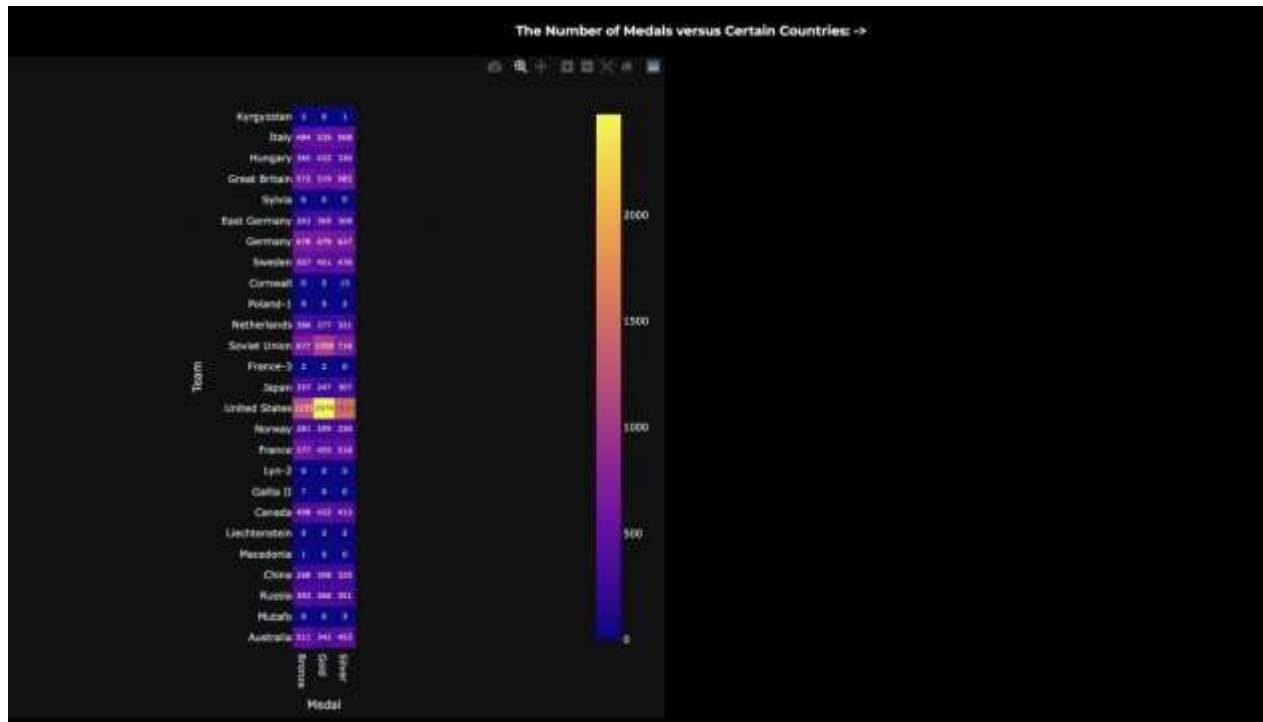
Code ->

```
df = pd.read_csv('athlete_events.csv', error_bad_lines=False,
engine="python")
newdf=df.groupby('Team').Medal.value_counts().unstack().fillna(
0)
topc=newdf.loc[newdf['Silver'] >= 300]
bottomc=newdf.loc[newdf['Silver'] <= 150 ]
bottomc=newdf.sample(n=10)
frames = [topc, bottomc]
result = pd.concat(frames)
result=result.sample(frac = 1)
fig_heatmap=px.imshow(result, text_auto=True, width=800, height=
400, template="plotly_dark")
fig_heatmap.layout.height = 800
fig_heatmap.layout.width = 800
```



Code ->

```
top_countries =  
data.Team.value_counts().sort_values(ascending=False).head(12  
03)  
u_team = data.Team.unique()  
fig_graph_2 = px.treemap(path=[u_team], values=top_countries,  
  
color_discrete_sequence=px.colors.sequential.RdBu,  
template="plotly_dark")
```



Code ->

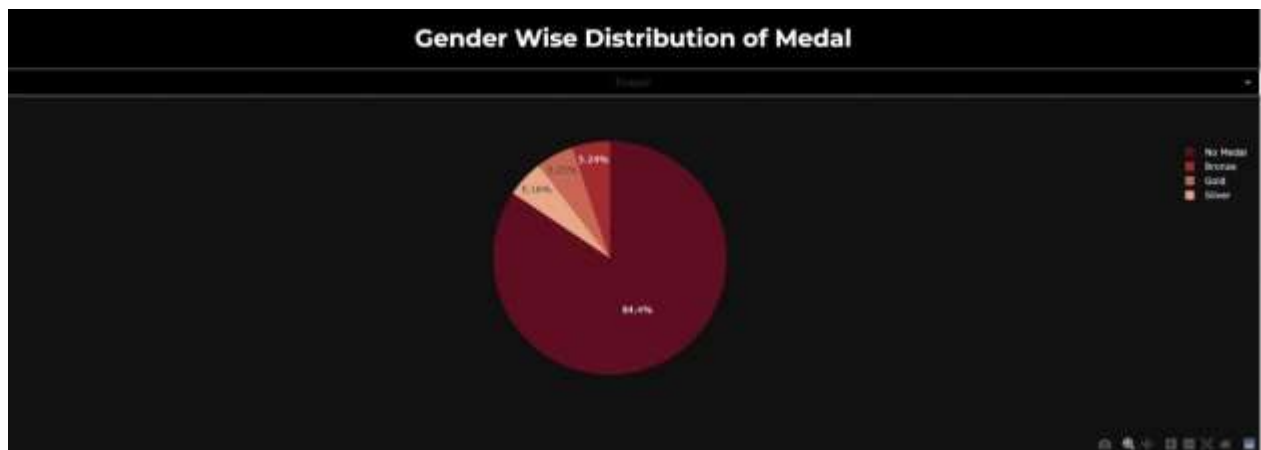
```
# Graph Male - Female Medal Distribution
@ app.callback(
    Output("graph_gender", "figure"),
    Input("gender_dropdown", "value"))
def update_gender(gender):
    data['Sex'] = data['Sex'].replace("M", "Male")
    data['Sex'] = data['Sex'].replace("F", "Female")
    data_male = data.loc[data['Sex'] == gender]

    medal_male = list(data_male['Medal'])
    for i in range(len(medal_male)):
        medal_male[i] = int(medal_male[i])

    medal_male_no = medal_male.count(0)
    medal_male_bronze = medal_male.count(1)
    medal_male_silver = medal_male.count(2)
    medal_male_gold = medal_male.count(3)
    medal_count = [medal_male_no, medal_male_bronze,
                    medal_male_silver, medal_male_gold]

    fig = px.pie(values=medal_count, names=["No Medal",
"Bronze", "Silver", "Gold"],

color_discrete_sequence=px.colors.sequential.RdBu,
template="plotly_dark")
    return fig
```

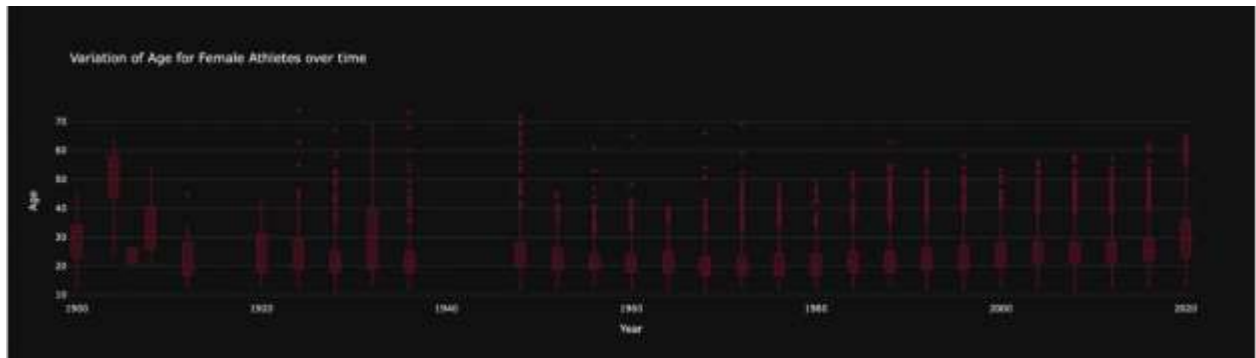


Code ->

```
# Graph MenOverTime
@app.callback(
    Output("graph_participation", "figure"),
    Input("gender_dropdown", "value"))
def update_box(gender):
    data['Sex'] = data['Sex'].replace("M", "Male")
    data['Sex'] = data['Sex'].replace("F", "Female")
    MenOverTime = data[(data.Sex == gender) & (data.Season ==
'Summer')]

    fig = px.box(MenOverTime, x='Year', y='Age',
hover_name='Year',

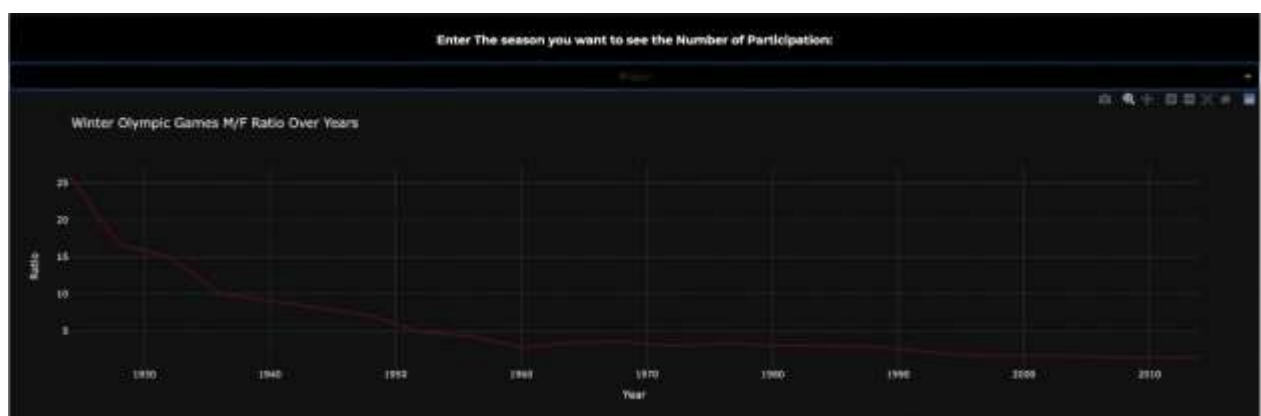
color_discrete_sequence=px.colors.sequential.RdBu,
template="plotly_dark",
                title="Variation of Age for "+str(gender)+"
Athletes over time")
    fig.update_traces(quartilemethod="inclusive")
    return fig
```



Code ->

```
# Graph SummerWinter Season
@app.callback(
    Output("graph_season", "figure"),
    Input("values_season", "value"))
def update_Season(season):
    data = pd.read_csv("data.csv")
    data.rename({'Unnamed: 0': 'ID'}, axis=1, inplace=True)
    summer_olympic = data.query("Season == 'Summer'")
    winter_olympic = data.query("Season == 'Winter'")
    games = {"Summer": summer_olympic, "Winter":
winter_olympic}
    sex_year_df = games[season].groupby(["Year",
"Sex"]).ID.count().rename("Count").reset_index()
    sex_year_df = sex_year_df.pivot(
        index="Year", columns="Sex",
values="Count").reset_index()
    sex_year_df["Ratio"] = sex_year_df["M"] /
sex_year_df["F"]
    fig = px.line(sex_year_df, x="Year", y="Ratio",
title=str(season)+" Olympic Games M/F Ratio Over Years",

color_discrete_sequence=px.colors.sequential.RdBu,
template="plotly_dark",)
    return fig
```



Code ->

```
values_year = data['Year'].unique()
values_year.sort()
values_year = list(values_year)
count_discipline = data.Sport.value_counts()
count_discipline
sns.set_style("ticks")
wordcloud = WordCloud(
    width=2000,
    height=1000,
    scale=1,
    normalize_plurals=False,
    repeat=False,
    random_state=42,
    background_color='black')
wordcloud.generate_from_frequencies(frequencies=count_discipline)
plt.figure(figsize=(17, 10))
plt.imshow(wordcloud, interpolation="bilinear")
plt.axis("off")
```



Code ->

```
dfS = df[df['Season']=='Summer'];
dfW = df[df['Season']=='Winter']

def draw_map(dataset, title, colorscale, reversescale=False):
    trace = go.Choropleth(
        locations = dataset['Country'],
        locationmode='country names',
        z = dataset['Editions'],
        text = dataset['Country'],
        autocolorscale =False,
        reversescale = reversescale,
        colorscale = colorscale,
        marker = dict(
            line = dict(
                color = 'rgb(0,0,0)',
                width = 0.5)
        ),
        colorbar = dict(
            title = 'Editions',
            tickprefix = '')
    )

    data = [trace]
    layout = go.Layout(
        title = title,
        geo = dict(
            showframe = True,
            showlakes = False,
            showcoastlines = True,
            projection = dict(
                type = 'orthographic'
            )
        )
    )
    fig = dict( data=data, layout=layout )
    iplot(fig)

draw_map(dfS, 'Olympic countries (Summer games)', "Reds")
```


Olympic countries (Summer games)



Olympic countries (Winter games)



The complete code and Jupyter notebooks can be found at this link:

<https://github.com/prathamagrawal/Analysing-Olympics-Dataset>

Thank you !