Analysis and Visualization of Olympics Games

Vinupriya Sanjay Kumar , SID - 7719
Data Visualization, Department of Data Analytics
San Jose State University
San Jose, US

Greeshma Venkatesh , SID - 7041
Data Visualization, Department of Data Analytics
San Jose State University
San Jose, US

Saroj Saran , SID - 6014
Data Visualization, Department of Data Analytics
San Jose State University
San Jose, US

***Abstract*— The Olympics, the most significant event in the lives of athletes and prospective athletes worldwide, is revered and appreciated. There are a lot of aspiring athletes all over the world who wants to participate in the Olympic Games. This project focuses on analyzing the Olympic Games by visualization through various types of graphs and charts which have been used to understand the patterns of the Olympics games, athletes, and countries. Further visualizations are done between the impact of GDP and Population with the Olympic Games. This can help both the athletes and the countries over time. The results are useful for aspiring athletes, viewers of Olympics, economists and working professionals who are interested in knowing the various factors that can help in the improvement of the performance and evolution of the Olympics games as a whole.**

Keywords— visualization, Olympics, GDP, Population, Tableau

I. Introduction

The Olympic Games is one of the most prestigious competitions held in the world. Athletes are always honored to participate in Olympics because it is a pinnacle in the career of an athlete. There is no other big international event in the world where all the topmost athletes come and participate in one city. The audience is excited to watch the topmost athletes compete against each other and people who share common values are brought together from all over the world to watch the Olympics Games. The host city is selected by a committee after careful discussion and consideration and the selected host city becomes the main center of attraction for a while and can attract a lot of tourists which in turn generates revenue for that host country. All of this shows the importance of the Olympic Games in our lives and for the economy[1].

However, there has always been a need to understand the patterns that are involved in the Olympic Games.Since it has both social and economic impact, it is extremely useful in understanding the various trends that are involved in Olympics.. In this study, we have used the Olympics Athlete Events dataset from Kaggle which contains about 2.75 lakhs records of the athletes, age, gender, height, weight events, medals, team, city and country over the years from 1896-2016. We have also used the GDP and Population dataset from world bank by merging it with the Athletes events dataset to understand the economic effect of Olympics in countries. Along with that we have taken the Tokyo Olympics 2020 dataset from Kaggle where we created an interactive Olympics event schedule dashboard of Tokyo 2020 Olympics.

In this project, we have explored the factors that affect the Olympics Games by analyzing the whole dataset and to provide related graphs by using Tableau as the visualization tool. Maps have been used to find the medal tally of countries in Olympics. Bar charts have been used to find average age peak performance, average height and weight of the athletes, the number of participation of athletes, the number of Olympic medals for the host country . Line graphs have been used to show trends over time between the number of events and the number of nations. Scatter plots have been used to show relations between the effect of GDP with the number of medals won and the medals of winter and summer Olympics respectively. Finally, we have created a tableau story with all visualizations and the interactive dashboards with all the results combined.

II. Related Work

In a paper by Rahul Pradhan et al. [4], they have evaluated the evolution of Olympics over the years through an exploratory data analysis. They provide analysis in visual format which help in the countries and players improve their skills. In exploratory data analysis, large data and its different characteristics is examined using charts, graphs, etc. The analysis is done using R language and using RStudio as a platform. Their main goal was to help an athlete and country analyze the performances and trend over the years so that they can perform better and change their strategies.

The idea that visual aids help to quickly perceive the data compared to traditional tabular methods is shown using the winter Olympics data [5]. Data can be better understood when there are innovative forms to view the data and make it interactive. To convey any information efficiently and clearly, the use of data visualization tools like infographics, charts, graphs, etc. are useful [6].

Collecting data and reports has been a tradition for long time for all sports. In the paper by Perin et al. [7] they have written about the challenges of visualizing sports data. The extensive research being made in this area indicates that timeliness of sports reports is important. Some of the tasks in this research area include new visualization techniques are explored, existing visualizations are adapted to new domains, design studies are conducted and evaluated. In this report three categories of sports data were analyzed – box score data which contains summary of a sport, tracking data which has the data about trajectories and in-game actions, and meta data about the participants and sport. Critical research gaps are identified, and their contribution of sports visualization classification helps provide actionable insights where data visualization is better combined with sports data.

It is well proven that team games have a home advantage. A study was made to assess [8] the impact of home advantage for the Olympics summer games held between 1896 and 1996. The sports were categorized into five event groups: weightlifting, athletics, gymnastics, boxing, and team games. Athletics and weightlifting came under predominantly objectively judged, gymnastics and boxing under subjectively judged, and team games which involved subjective decisions. It was concluded that for the events which are based on subjective decisions and judgement had highly significant home advantage. Also, it was observed that no home advantage was found in the events where objective judgement was made.

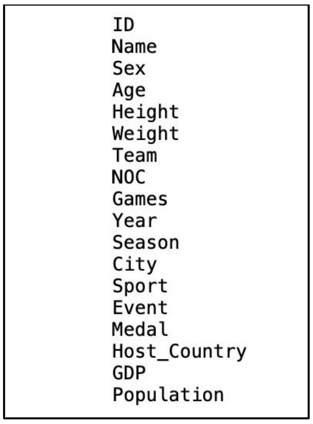
Olympics games played an important role to convey a message to society that is striving hard for gender equality [9]. The evolution of the female athletes' participation was analyzed [10] in the Olympics games using data from various sources reports from Olympic Committee. It has been observed that from the beginning of 20th century there is a significant shift in the women participation which shows a positive trend towards full equality.

We have referred multiple papers mentioned above about Olympics and different visualization techniques of how to view sports data for our project. We then have carefully chosen the type of graph to infer key points from the Olympics data.

III. THE DATA SET

The essential part of any good analysis is finding the correct data. The Olympics athlete events dataset used here in our project has been taken from Kaggle [2] which contains around 2.7 lakhs records from 1896 to December 2016. We have also used the Olympic event schedule dataset [3] of Tokyo 2021 for creating an interactive event scheduling page of the Olympics games. For our analysis we have also used datasets [12,13] which contain GDP and population of all the countries in the world.

Our final athlete events dataset after combining the Olympics events, GDP and Population has the following columns as shown in Fig. 1. There are few null values present in the dataset.



```
ID
Name
Sex
Age
Height
Weight
Team
NOC
Games
Year
Season
City
Sport
Event
Medal
Host_Country
GDP
Population
```

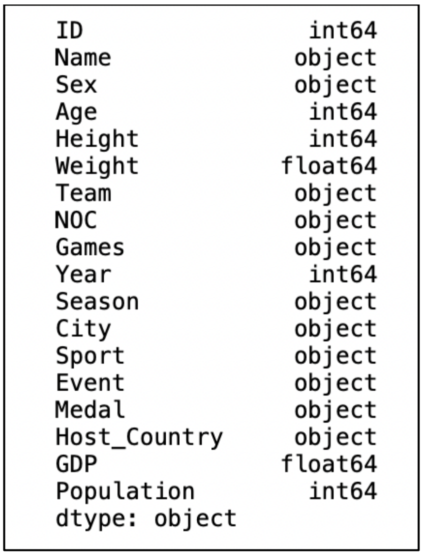
Fig. 1. Column Names in Olympics Athlete Events Dataset

TABLE I below gives the details of the different attributes representing our data. It has three main columns giving details about the name of attribute, what the attribute means and if it can hold a null value or not.

TABLE I. OLYMPICS EVENTS ANALYSIS DATASET

Sl.No	Attribute	Details	Nullable
1	ID	Unique identifier of the athlete participating in Olympics	No
2	Name	Name of the athlete	No
3	Sex	Gender of the athlete – M : Male , F : Female	Yes
4	Age	Age of the athlete	Yes
5	Height	Height of the athlete in cm	No
6	Weight	Weight of the athlete in kg	No
7	Team	The country which has been represented by the participating athletes	No
8	NOC	The name of the country in code as used by Olympics committee	No
9	Games	Name of the games based on the season in which it was held i.e, Summer or Winter	No
10	Year	Year in which the Olympics were held	No
11	Season	Season in which the Olympics were held	No
12	City	City in which the Olympics were held	No
13	Sport	Name of the Olympics sport	No
14	Event	Name of the Olympics sport event	No
15	Medal	Medal won by the athlete: Gold, Silver or Bronze	No
16	Host Country	Country which hosted the Olympics	No
17	GDP	GDP of the athlete's country in the year when Olympics was held	Yes
18	Population	Population of the athlete's country in the year when Olympics was held	Yes

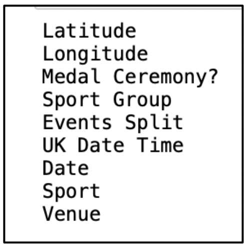
The datatype of these attributes can be seen in the Fig. 2.



```
ID                 int64
Name              object
Sex               object
Age                int64
Height             int64
Weight           float64
Team              object
NOC               object
Games             object
Year               int64
Season            object
City              object
Sport             object
Event             object
Medal             object
Host_Country      object
GDP              float64
Population         int64
dtype: object
```

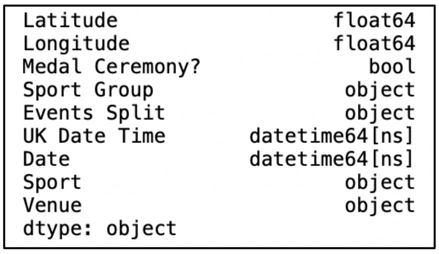
Fig. 2. Datatype of attributes

The Tokyo 2021 Olympic schedule dataset contains around 1900 rows with the following 9 columns as shown in Fig 3. The datatype of these attributes can be seen in Fig. 4.



```
Latitude
Longitude
Medal Ceremony?
Sport Group
Events Split
UK Date Time
Date
Sport
Venue
```

Fig. 3. Column Names in Tokyo 2021 Event Scheduling dataset



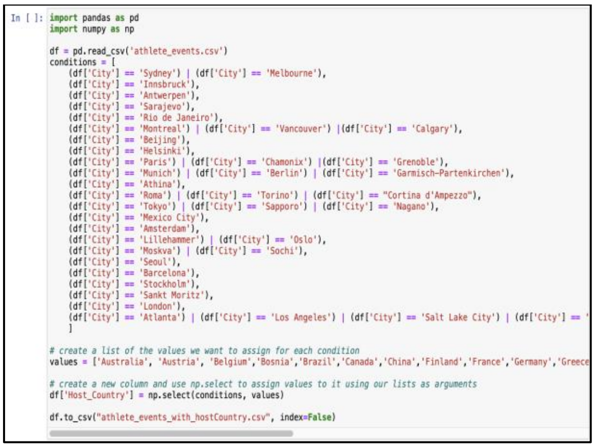
```
Latitude                  float64
Longitude                 float64
Medal Ceremony?              bool
Sport Group                object
Events Split               object
UK Date Time       datetime64[ns]
Date               datetime64[ns]
Sport                      object
Venue                      object
dtype: object
```

Fig. 4. Data of scheduling events dataset

IV. Methods

A. Data Merging and Cleaning

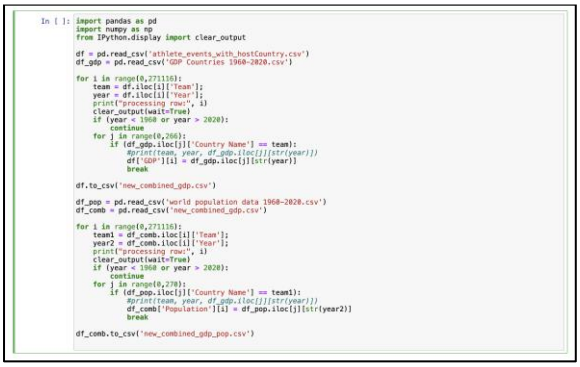
In our athlete events dataset, we first added a host country column which takes the value of the country name where the Olympics was held based on the 'City' column which represented the city where Olympics was hosted. We have used Python to add a new column of the host country to the dataset we obtained. A snippet of the python code written to add the host country is shown in Fig. 5.



```
In [ ]: import pandas as pd
        import numpy as np

        df = pd.read_csv('athlete_events.csv')
        conditions = [
            (df['City'] == 'Sydney') | (df['City'] == 'Melbourne'),
            (df['City'] == 'Innsbruck'),
            (df['City'] == 'Antwerpen'),
            (df['City'] == 'Sarajevo'),
            (df['City'] == 'Rio de Janeiro'),
            (df['City'] == 'Montreal') | (df['City'] == 'Vancouver') |(df['City'] == 'Calgary'),
            (df['City'] == 'Beijing'),
            (df['City'] == 'Helsinki'),
            (df['City'] == 'Paris') | (df['City'] == 'Chamonix') |(df['City'] == 'Grenoble'),
            (df['City'] == 'Munich') | (df['City'] == 'Berlin') | (df['City'] == 'Garmisch-Partenkirchen'),
            (df['City'] == 'Athina'),
            (df['City'] == 'Roma') | (df['City'] == 'Torino') | (df['City'] == "Cortina d'Ampezzo"),
            (df['City'] == 'Tokyo') | (df['City'] == 'Sapporo') | (df['City'] == 'Nagano'),
            (df['City'] == 'Mexico City'),
            (df['City'] == 'Amsterdam'),
            (df['City'] == 'Lillehammer') | (df['City'] == 'Oslo'),
            (df['City'] == 'Moskva') | (df['City'] == 'Sochi'),
            (df['City'] == 'Seoul'),
            (df['City'] == 'Barcelona'),
            (df['City'] == 'Stockholm'),
            (df['City'] == 'Sankt Moritz'),
            (df['City'] == 'London'),
            (df['City'] == 'Atlanta') | (df['City'] == 'Los Angeles') | (df['City'] == 'Salt Lake City') | (df['City'] == '
            ]

        # create a list of the values we want to assign for each condition
        values = ['Australia', 'Austria', 'Belgium','Bosnia','Brazil','Canada','China','Finland','France','Germany','Greece

        # create a new column and use np.select to assign values to it using our lists as arguments
        df['Host_Country'] = np.select(conditions, values)

        df.to_csv("athlete_events_with_hostCountry.csv", index=False)
```

Fig. 5. Python code to add a host country column to our dataset

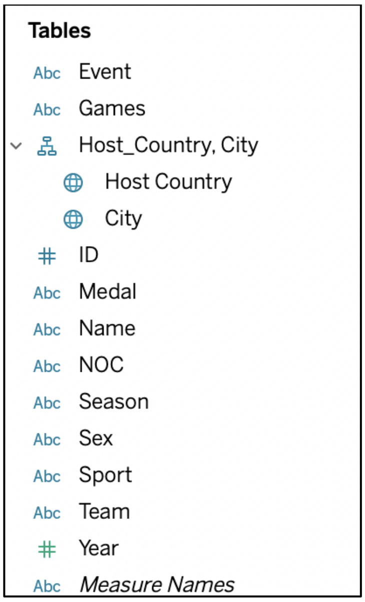
We wanted to visualize the effect of GDP and Population of a country on the Olympics games and so we needed 2 more columns, the GDP and Population. Python code was used to pick GDP and population values of each country year wise from 2 datasets and was added to our previous dataset. The code snippet of the python code written is as shown in Fig. 6.



```
In [ ]: import pandas as pd
        import numpy as np
        from IPython.display import clear_output

        df = pd.read_csv('athlete_events_with_hostCountry.csv')
        df_gdp = pd.read_csv('GDP Countries 1960-2020.csv')

        for i in range(0,271116):
            team = df.iloc[i]['Team'];
            year = df.iloc[i]['Year'];
            print("processing row:", i)
            clear_output(wait=True)
            if (year < 1960 or year > 2020):
                continue
            for j in range(0,266):
                if (df_gdp.iloc[j]['Country Name'] == team):
                    #print(team, year, df_gdp.iloc[j][str(year)])
                    df['GDP'][i] = df_gdp.iloc[j][str(year)]
                    break

        df.to_csv('new_combined_gdp.csv')

        df_pop = pd.read_csv('world population data 1960-2020.csv')
        df_comb = pd.read_csv('new_combined_gdp.csv')

        for i in range(0,271116):
            team1 = df_comb.iloc[i]['Team'];
            year2 = df_comb.iloc[i]['Year'];
            print("processing row:", i)
            clear_output(wait=True)
            if (year < 1960 or year > 2020):
                continue
            for j in range(0,270):
                if (df_pop.iloc[j]['Country Name'] == team1):
                    #print(team, year, df_gdp.iloc[j][str(year)])
                    df_comb['Population'][i] = df_pop.iloc[j][str(year2)]
                    break

        df_comb.to_csv('new_combined_gdp_pop.csv')
```

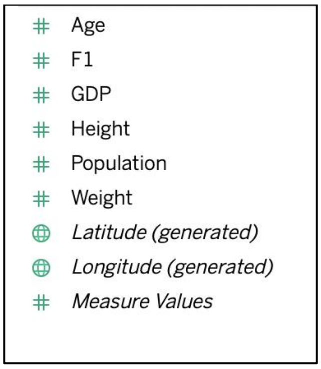
Fig. 6. Python code to add GDP and Population to our dataset

Our final dataset was uploaded in Tableau to look for the dimensions Fig. 7 and measures Fig. 8.



Tables
Abc Event
Abc Games
Host_Country, City
Host Country
City
ID
Abc Medal
Abc Name
Abc NOC
Abc Season
Abc Sex
Abc Sport
Abc Team
Year
Abc *Measure Names*

Fig. 7. Dimensions



Age
F1
GDP
Height
Population
Weight
Latitude (generated)
Longitude (generated)
Measure Values

Fig. 8. Measures

We used Tableau Prep to remove few duplicate rows after merging the data. Also, the rows that had null values were removed using the tableau 'Clean' step. Duplicates were removed using the 'Aggregate' function. Fig 9 shows the flow of the data cleaning process.

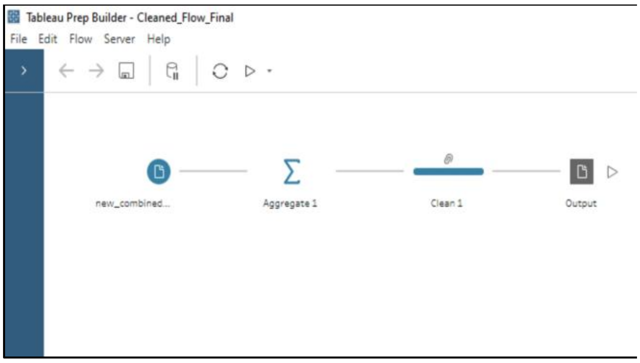


Tableau Prep Builder - Cleaned_Flow_Final
File Edit Flow Server Help
new_combined...
Aggregate 1
Clean 1
Output

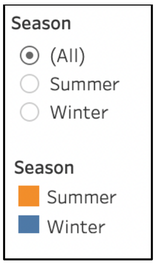
Fig. 9. Data Cleaning done used Tableau Prep

The columns of the scheduling event dataset were straightforward and well prepared, so we used them as it is for creating a scheduler dashboard.

B. Visualization Design

It is complicated to analyze the different attributes of the Olympics over the years to get meaningful insights. We have tried to categorize different aspects while visualizing the data to make it simple and easy to understand.

Firstly, we try to find the difference between Summer and Winter Olympics. For this the Tableau feature of filters was very useful where for the same attributes we can see the change in trend for different seasons.



Season
● (All)
○ Summer
○ Winter

Season
🟧 Summer
🟦 Winter

Fig. 10. Filter Option in Tableau

For other analysis like effect of GDP, Population, and home country advantage we use the count measure of the medal column instead of using it as a dimension to get the number of medals.

The geographic spread of the number of medals won by each country over the history of Olympics is shown using the maps of tableau. Heatmap is used for each country to know the rank of countries based on total medals won.

Based on the type of analysis we have carefully selected the type of graph to clearly present the data. We have used filters and tooltips in all the graphs to give more details about the attribute being analyzed.

C. Tableau Operation

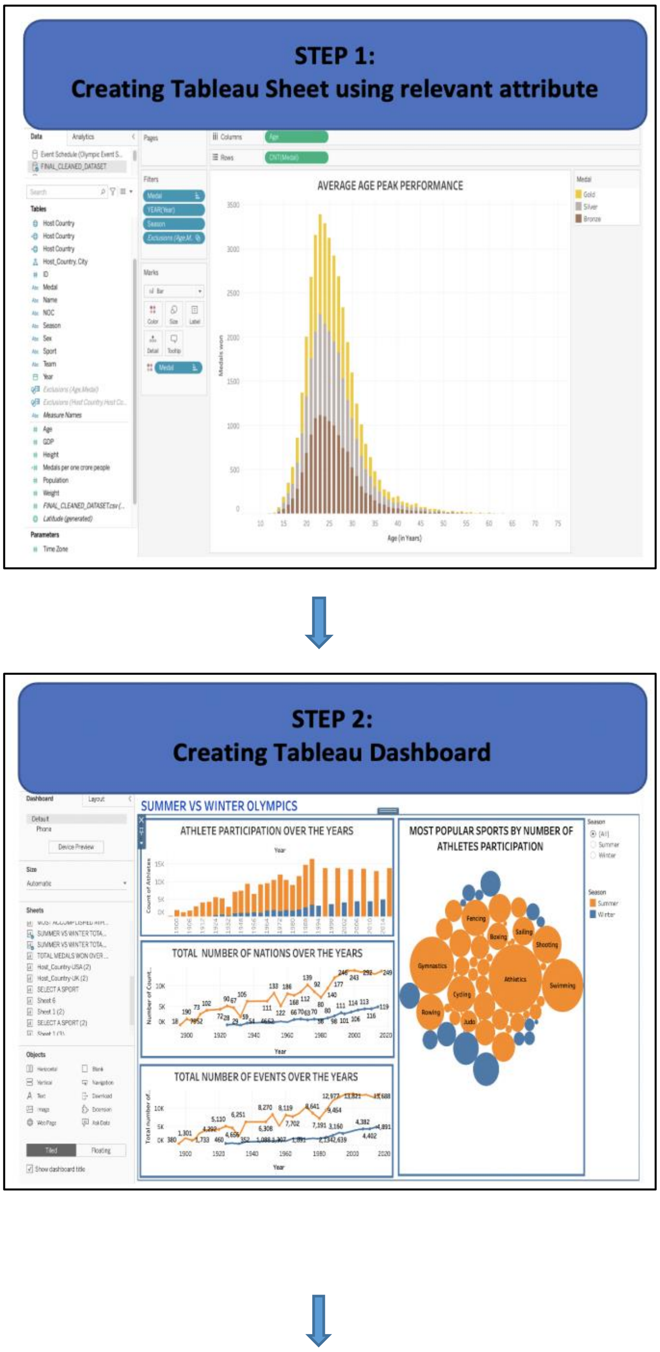
In this project we have used Tableau's Story feature to demonstrate the analysis of various attributes in Olympics dataset. We have created multiple interactive dashboards which clearly show the correlations of the data to the users.

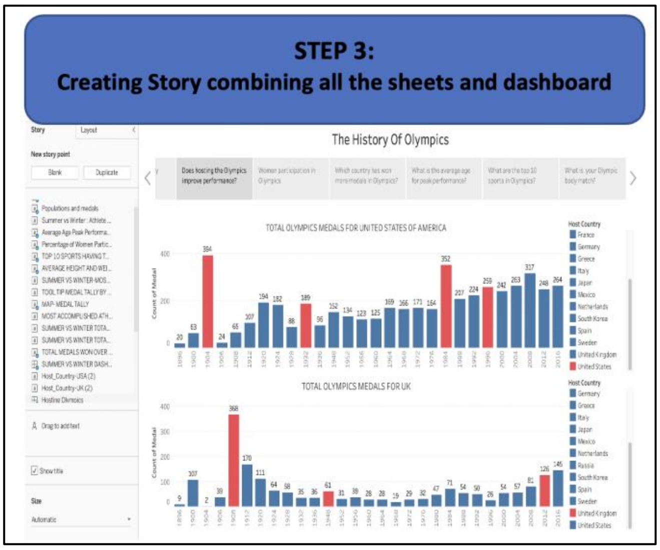
We first created individual sheets which analyzed the different attributes using the different graphs of Tableau. Some part of data cleaning was done here to remove the null values from graph. Also, various filters were included to visualize the relevant data.

Next, the similar sheets were grouped together in a dashboard to combine similar information. Dashboard is designed and arranged carefully for better visualization and adding common filters.

Finally, a storyline is created to elate all our visualizations. The story contains a combination of interactive dashboards and individual sheets. Each story point includes a feature which can lead users to better analyze and get insights from the dataset. This gives a deeper understanding of the various factors that are contribute to our data.

In Fig. 10 the steps used for creating a Tableau story are shown.


STEP 1:
Creating Tableau Sheet using relevant attribute
AVERAGE AGE PEAK PERFORMANCE
Medals won
Age (in Years)
Medal
Gold
Silver
Bronze
STEP 2:
Creating Tableau Dashboard
SUMMER VS WINTER OLYMPICS
ATHLETE PARTICIPATION OVER THE YEARS
TOTAL NUMBER OF NATIONS OVER THE YEARS
TOTAL NUMBER OF EVENTS OVER THE YEARS
MOST POPULAR SPORTS BY NUMBER OF ATHLETES PARTICIPATION
Fencing
Rowing
Sailing
Shooting
Gymnastics
Athletics
Swimming
Cycling
Judo
Year
Season
Summer
Winter


STEP 3:
Creating Story combining all the sheets and dashboard

The History Of Olympics
Does hosting the Olympics improve performance?
Women participation in Olympics
Which country has won more medals in Olympics?
What is the average age for peak performance?
What are the top 10 sports in Olympics?
What is your Olympic body match?
TOTAL OLYMPICS MEDALS FOR UNITED STATES OF AMERICA
TOTAL OLYMPICS MEDALS FOR UK
Count of Medal
Host Country
Story
Layout
New story point
Blank
Duplicate
Drag to add text
Show title
Size
Automatic

Fig. 11. Tableau Operations

V. Results

We have combined all our visualizations and interactive dashboards in the form of tableau story. The tableau story will be explained from top to bottom, and they are:

- Page 1: Summer vs Winter Olympics Interactive Dashboard
- Page 2: Effect of GDP on the number of medals won
- Page 3: Medals and Population in 2016 Summer Olympics
- Page 4: Analysis of Host Country - USA & UK
- Page 5: Participation of Women in Olympics
- Page 6: Map View of the Medal Tally of the Countries
- Page 7: Average Age Peak Performance
- Page 8: Top 10 Sports with Highest Number of Medals
- Page 9: Average Height and Weight of Each Sport by Gender
- Page 10: Most Accomplished Athletes by the Number of Medals Won
- Page 11: Total Medals Won Over the Years by Countries
- Page 12: Tokyo 2021 Olympics Events Schedule Interactive Dashboard

A. *Summer Vs Winter Olympics Interactive Dashboard*

Fig.12 shows the first page in the story, which is a summer vs winter Olympics comparison represented in an interactive dashboard. The first visualization represents the athlete participation over the years in Summer and Winter Olympics respectively. In Summer Olympics the highest number of participants was in the year 2000 with a count of 13,821 athletes, and the highest number in Winter Olympics was in the year 2014 with a count of 4,891 athletes. The second visualization represents total number of events over the years where the number of events has continuously increased over the years with the current standing of 306 events in the Summer Olympics and a current standing of 98 events in the Winter Olympics. The Third visualization in the dashboard represents the number of nations over the years and the highest number of nations that have participated in the Summer Olympics is 292 nations in the year 2008, and the highest number of nations that have participated in the Winter Olympics is 119 nations in the year 2014. The fourth visualization represents the most popular sport in the Summer and Winter Olympics by the number of athlete participation. The most popular sport in Summer is athletics with 38,624 participants and the most popular sport in Winter is cross country skiing with 9,133 participants.

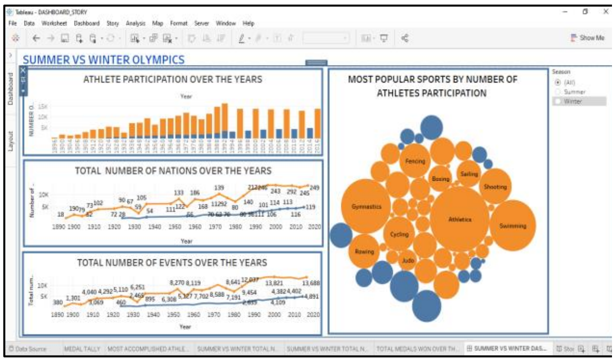


Tableau - DASHBOARD_STORY
SUMMER VS WINTER OLYMPICS
ATHLETE PARTICIPATION OVER THE YEARS
MOST POPULAR SPORTS BY NUMBER OF ATHLETES PARTICIPATION
TOTAL NUMBER OF NATIONS OVER THE YEARS
TOTAL NUMBER OF EVENTS OVER THE YEARS
Year
Season
(All)
Summer
Winter
Fencing
Boxing
Sailing
Shooting
Gymnastics
Athletics
Swimming
Cycling
Rowing
Judo

Fig. 12. The Summer vs Winter Olympics Interactive Dashboard

B. *Effect of GDP on the number of medals won*

In the second page in the story, we have tried to visualize the effect of economy of the country on the performance by their athletes in Olympics. We find that there seems to be a positive correlation between the GDP of the country and medal tally. We have created a scatter plot graph that plots the total number of medals won a country in a particular Olympics games as seen in Fig. 13. There is a filter to select between the seasons of Olympics and the games relevant to that season shows up. The countries are plotted based on the NOC and trend line is added. The trendline clearly shows a strong positive correlation among the GDP and number of medals variables.

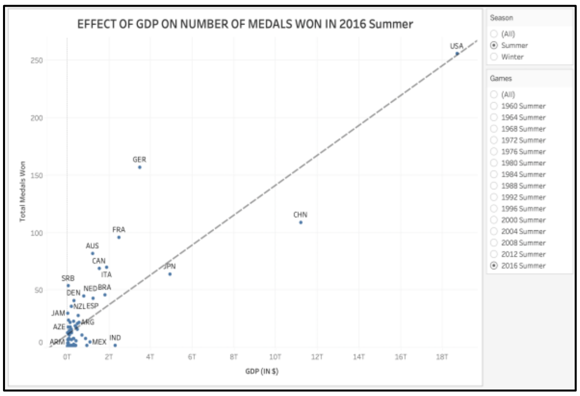


Fig. 13. GDP and medal tally analysis page

C. *Medals and Population in 2016 Summer Olympics*

To analyze the effect of population of a country on the number of medals won we have used the 2016 Summer Olympics data. A chart is plotted with total medals won by a country against the number of medals won per crore population as shown in Fig. 14. We find that there is no correlation between medals won and the population of a country. For example, US which has won the highest medals and is among a most populated countries, it stands quite below when we consider the medals won per population. On the other hand we that Fiji which is sparsely populated turns out to have come in top though it has won only 13 medals.

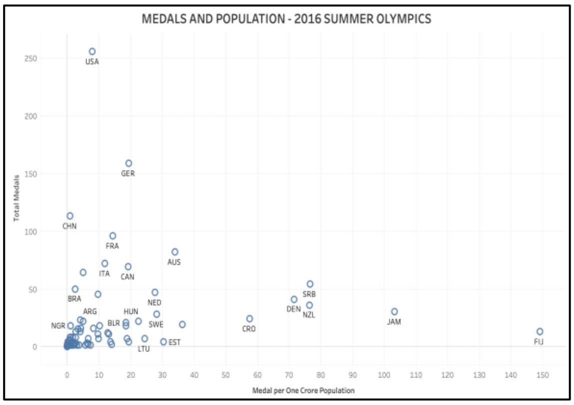


Fig. 14. Medals and population analysis page

D. *Analysis of Host Country advantage*

In the fourth page, we try to find out if the host country has any advantage in winning medals in the Olympic games. We have plotted the total medals won over the years in Summer Olympics for US and UK in two separate charts. In the first chart which represents the medals won by US, the year in which US themselves hosted the Summer Olympics is highlighted with a different color. It is clearly seen that the average of the number of medals is less when US did not host the Olympics whereas the medal count is high when they hosted. A similar trend can be seen for UK as well. The dashboard is as shown in Fig. 15

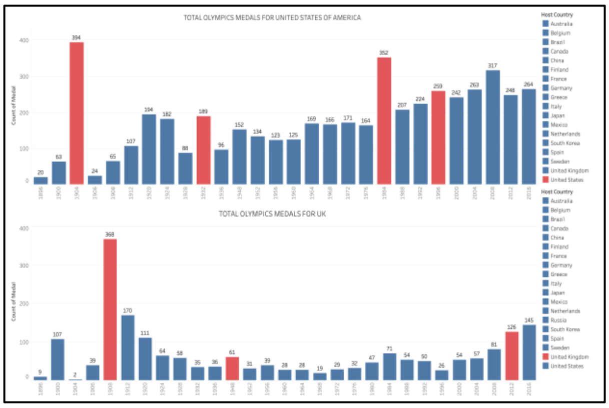


Fig. 15. Host Country advantage analysis

E. *Participation of Women in Olympics*

In this page we show analyze the female participation in Olympics over the years. We have used the filter to see the trend in both summer and winter Olympics. We clearly see that the number of women participants have increased over the years from a mere count of 33 when Olympics began in 1896 to over 6300 in 2016. But this increase has been gradual in the initial years and after 1980's we see a big increase of numbers as seen in Fig. 16.

We have added a filter to select both male and female participants to check the ratio of the male female participation. As seen in Fig 17 the ratio is quite less till 1990. After 1990 we see the ratio increasing and in 2016 we can say the ratio is nearly 50:50 showing equal participation of men and women.

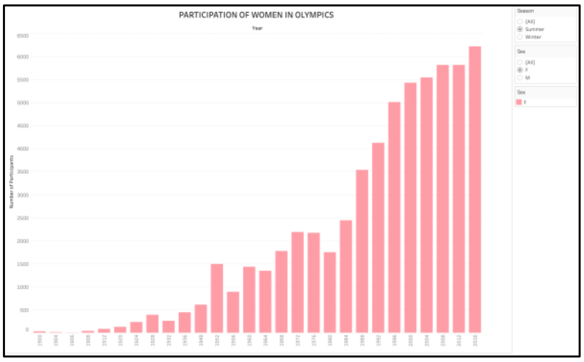


Fig. 16. Women participation trend page

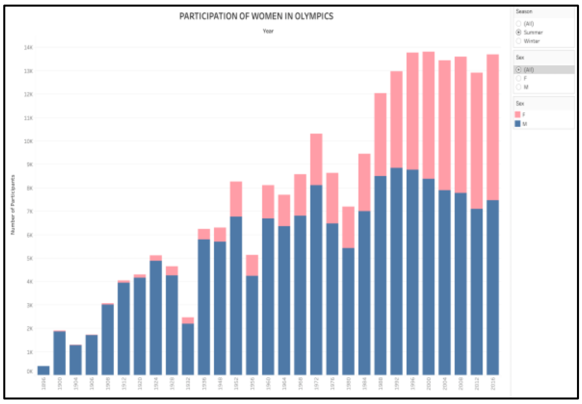


Fig. 17. Men vs Women participation trend over the years

F. *Map view of the Medal Tally of the Countries*

Fig.18 shows the Map View of the Medal Tally of the Countries. United States is the country with the highest number of medals in Olympics. United States have won an overall of 2,474 gold medals, 1,512 silver medals 1,233 bronze medals. There are a total of 453 countries, and they are ranked based on the total number of medal counts. They have been represented in the shades of blue where the darkest blue shade represents the country with the highest number of medals and the lightest shade represents the country with the lowest number of medals. The total medal counts of the country have been also labelled in the maps.

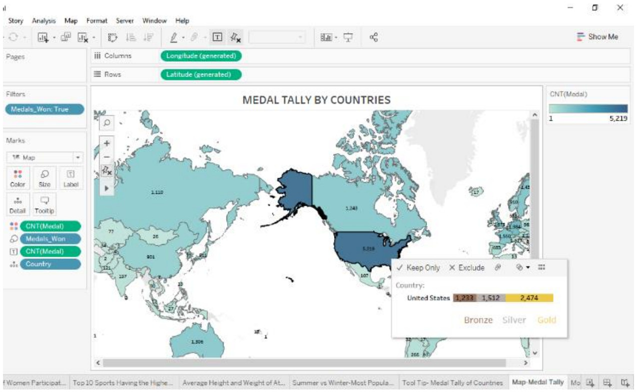


Fig. 18. Map View of the Medal Tally of the Countries

G. *Average Age Peak Performance*

Fig. 19 shows the total medals won by the athletes vs their age considering all the Olympics games held over the years. We see that most of medals are won by athletes who are in the age of 21 to 27. The average age of the athletes who give their best performance can be said to between 20-30 considering the previous history of data. We have differentiated between the three medals bronze, silver and gold for each age group.

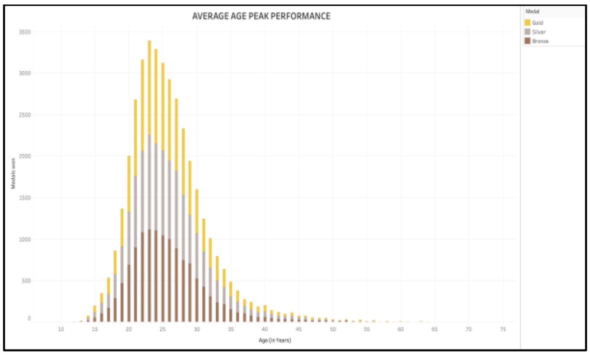


Fig. 19. Average age peak performace analysis

H. *Top 10 Sports with the Highest Number of Medals*

Fig. 20 shows The Top 10 Sports with the Highest Number of Medals are Athletics, Swimming, Rowing, Gymnastics, Fencing, Wrestling, Cycling, Shooting, Cross Country Skiing and Alpine Skiing. Athletics stands as the Top 1 with the highest number of medals consisting of 1,339 Gold medals, 1,334 Silver medals and 1,296 Bronze medals.

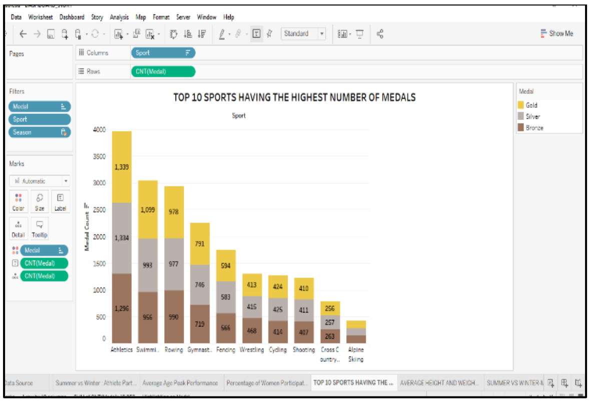


Fig. 20. Top 10 Sports with the Highest Number of Medals

I. *Average Height and Weight of Athletes by Each Sport and Gender*

Fig. 21 shows the Average Height, and Weight of Athletes by Each Sport and Gender that varies accordingly to the different sports. The highest average height for female is the sport beach volleyball with 175.00 cm. The highest average height for male is the sport volleyball with 185.7 cm. The highest average weight for female is the sport bobsleigh with 72.80 kg. The highest average weight for male is the sport Rugby sevens with 91.01 kg.

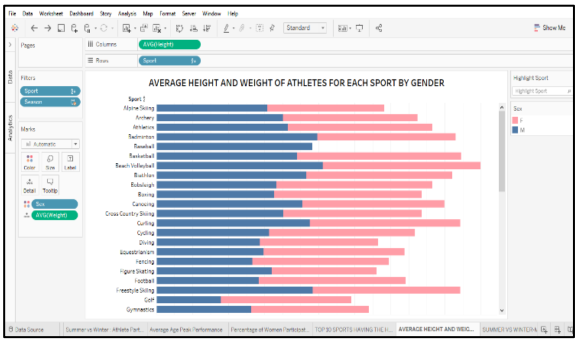


Fig. 21. Average Height and Weight of Athletes by Each Sport and Gender

J. *Total Medals Won Over the Years by Countries*

Fig. 22 shows the Total Medals Won Over the Years. The Medals won by countries varies according to the season a i.e., according to the Summer and Winter Olympics. From the visualization we can infer that there is a fewer number of medals won in Winter Olympics compared to Summer Olympics. This is mainly because the countries participating in Winter Olympics are fewer compared to the Summer Olympics.

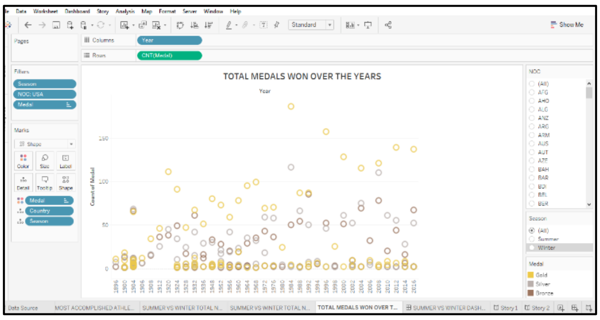


Fig. 22. Total Medals Won Over the Years

K. *Tokyo 2021 Olympics Events Schedule Dashboard*

Fig. 23 shows the interactive dashboard which we have built for Olympics. This dashboard allows user to customize their dashboard based on the Olympic sport events of their choice. They can either add one sport at a time or add multiple sports one below the other as per their choice. The dashboard contains a list of all the Olympic sport events happening from where the user can select the sport of their choice. It also has a feature to remove a particular sport event if the user wants to make edits to his/her dashboard. Once the user customizes the dashboard based on their choice, they can also view if the respective sport is participating for Qualifications or Finals. It also provides details of how many numbers of sports in each category is scheduled. Further the dashboard also provides details on date and month on which a particular sport is scheduled for. It also shows the categorization of a particular sport and the sport group it belongs to.

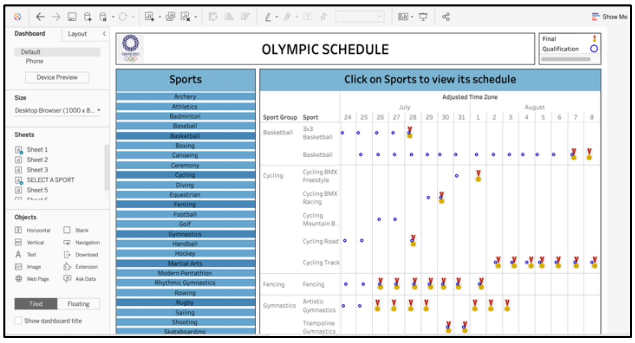


Fig. 23. Olympics events schedule dashboard

VI. Discussions

Our expectation for the audience is to understand the patterns and trends of the Olympics Games. Since Olympics is such a huge international platform for sports, there will be a lot of audience, athlete aspirers, working professionals, previous and future Olympic participants who wants to know how Olympics Games have a huge social and economic impact in the world.

From Fig. 12, the audience will get an overview of the Summer and Winter Olympics Games. The athlete participation for Summer and Winter Olympics varies. The audience can see that athlete participation in Summer Games is more than the athlete participation in the Winter Games. One of the main reasons could be the different sports happening in both games and thus the athletes are restricted to participate in all the events that happens especially Winter Games. This is mainly because there are countries which don’t have winters and it becomes difficult for them to practice for winter sports in their respective countries. The total number of events have increased over the years for both summer and winter. This shows the progress of the Olympics Games where they have been adding new events each year, and thus on a current standing for summer Olympics there are 306 events as of the year 2016 events from the year 1890 which started of which started off with 48 events. The winters were first held in the year 1920 and similarly, the events have increased from 17 events to 98 events. The total number of nations that have participated in the Olympics have been increasing over the years in both Summer and Winter Olympics Games. This shows how people are interested in sports and that there have been an increasing number of Olympic participants from all over the world. The most popular sport in Summer Olympics is athletics which has the highest number of participants and the most popular sport in Winter Olympics is Cross Country Skiing which has the highest number of participants.

Fig. 13 shows the effect of GDP on medals in summer 2016. From this the audience can understand that the size of the economy seems to have a positive correlation to the number of podium finishes. This means that as the economy increases the medals are also increasing in a positive correlation. If we look in the chart the countries that have less GDP and fewer number of medals are the bottom and countries that have more GDP and higher number of medals are the top. This relates that the total medal count is related to the total GDP of a country.

Fig. 14 shows the medals and population The figure compares a country's overall number of medals won at the Tokyo Olympics to the number of medals won per crore people. The trendline indicates that there is no relationship between a country's population size and medals earned. For example, the United States, which is one of the most populated countries, earned the most medals but ranks worse when medals won per crore people are considered. On the other side, despite winning only a few medals, a thinly populated region like Fiji came out on top. This shows that there is no effect of population on medals.

In Fig. 15, which shows the analysis of host countries United States & UK. From these two charts the audience can understand that when United States had been hosting for the Olympics, they have won the greatest number of medals compared to other countries. United states have hosted in 1904, 1932, 1984 and 1996. Similarly, when UK had been hosting Olympics, they have won the highest number of medals. UK had hosted Olympics 1908, 1948, 2012. Thus, from this the audience can understand that host country does have an impact on the number of medal winnings.

Fig. 16 shows the participation of women in Olympics, where the participation has increased widely over the years. From this chart the audience can understand that women have been giving a lot of importance to sports and are interested in excelling such fields.

Fig. 18 shows the Map View of the Medal Tally of the Countries. From this visual the audience can understand and identify those countries that have won the highest number of medals in Olympics based on the total number of medal counts. The ranking has been done on color scheme where the lightest shade of the color represents the country with the lowest number of medals and the darkest shade of the color represents the country with the highest number of medals. The audience can also identify the number of golds, silver and bronze medal won by the countries respectively.

Demographic analysis is done by Fig. 19, Fig. 20, Fig. 21, and Fig. 22. Fig. 19 represents the average age peak performance of athlete. Fig. 20 represents the Top 10 Sports with the highest number of medals. Fig. 21 represents the average height and average weight of each sport by gender. Fig. 22 represents the total medals won over the years by countries in summer and winter. All these visuals can be easily understood by the audience since most the charts are bar charts with properly named labels and axis.

Fig. 23 shows the Tokyo 2021 Olympics Events Schedule Dashboard. This Dashboard represents the sports and their events in a schedule calendar form. It also shows the number of events that have qualified from the sport. This dashboard is widely useful for the Olympic audience, Olympic participants and even for working professionals. Audience can also understand the different events that are happening on the different days by just having a look at the Olympics Schedule. It will also give an idea to the working professionals on creating Olympic schedules in an interactive manner so that it will be easier for the readers and audience to understand.

To summarize, we want the audience to understand the different patterns, trends, and outliers in the Olympic Games. In our visualization, we have tried our best to cover all the important factors in the Olympics Games. Our visuals are made in a way that it can be easily understood by any audience. The visualization is mainly focused for the aspiring athletes and the audience who are interested Olympics. Thus, our visuals have covered the social, economic, and demographic analysis of Olympics.

VII. Future Work

In this project, we have created a brief visualize for the Olympic Schedule, which represents the date and month of a particular sport upon selection. It contains the details of the classification which represents whether the respective sports event is scheduled for qualifications or finals. The total number of sports present under each category is also visualized. We can indeed include more further insights and features with deeper information to this interactive dashboard with other analysis which would excite many of the Olympic enthusiasts.

For Example, additional details like the status of each of the sports mentioning if the sport is finished or scheduled or In-Progress can be added. We can include live updates about countries leading position numbers for all the sports. This would give an idea for the Olympic enthusiasts who are willing to see timely updates of the country's winning chance. This can be achieved with the help of collecting more data on the above fields and merging all the required data using python packages. To support the integration of live updates on the tableau dashboard we can use a database like Oracle database which is updated with live data on a timely basis, and we can establish a live connection on tableau which sends queries to the database and updates the visuals based on the changed results.

To make these insights more easily accessible to people, mobile applications with necessary analysis and visualization can be built. We can also enable notification feature to remind people who are eagerly waiting for a particular event to begin and would want to follow and visualize timely updates of the sport events.

Also, as another additional feature, since know that there has been an exponential increase in the number of different sport events that is being added in the Olympics every year. Upon more research and study of different Olympic factors and by choosing an appropriate machine learning model based on the study results, we can predict the chances of a particular sport being added or not for the year. This would be helpful for the preparation of an athlete who is planning to participate and win in Olympics.

VIII. What Did We Learn

Working on this project together as a group gave us a great learning experience in every stage of the project. Since Olympics is one of the most momentous events and brings new hope for collective unity amongst people, we decided to derive insights on Olympic dataset which people will be curious to know about. As a first step we researched several research papers and learnt several points which we need to cover in our project which has not yet been touched upon until now. Then we started with reviewing multiple datasets on various data sources to obtain all the features which we would require for visualization. During this process we understood that we will have to work on merging several datasets based on our needs and perform data cleaning to obtain accurate results in our visualization.

For merging which is the required and the most important step in our project we used a data manipulation tool which is Pandas and we also used NumPy as our dataset contained a lot of numerical operations to be worked upon like GDP and Population. We learnt new concepts of iloc and dictionary in python for obtaining the required merged dataset with all the necessary columns with required details. A lot of time and learning was catered to this step as it is the baseline for the entire project to build upon. To make sure the data merged is correct we had to perform a lot of trial runs in a set of small batches. This was done because there was a total of 266 country names with 271116 row items in the dataset which took almost 3 hours to run every line item to perform matching based on the country name, population, and GDP in3 different datasets. We made sure that accurate results are obtained upon merging.

Then, as a next step we had to perform cleaning for our data. Since we had several columns which contained null values and repeated variables. We had to remove these anomalies in our merged dataset as it could lead to incorrect results during the process of analysis and visualization. We learnt how to perform these cleaning procedures using Tableau Prep which was easy and effective to perform. We proceed further with the visualizations of the Olympic attributes we had decided on, the visualization step consumed a lot of time and gave a lot of learning experience in terms of using Tableau Operations and tools. We learnt different techniques of visualization and used various chart techniques which provides easy understanding of the Olympic insights for the viewers using our Dashboard. We learnt various options in Tableau like creating measures, calculated fields, importing images from other sources and incorporating it in shapes, creating interactive dashboards etc.

In the process of project development, we learned how to create several sheets with required features and then how to merge them into one single dashboard which gives an overview of summary of results. Then, we learned how to merge all dashboards into one story and create a story line. We tried to implement all the concepts learnt in class using various visualization principles and color combinations keeping in mind of color blinded people. All these concepts helped us a lot in developing the dashboards and building a storyline for our project.

IX. NOVELTY OF THE IDEA

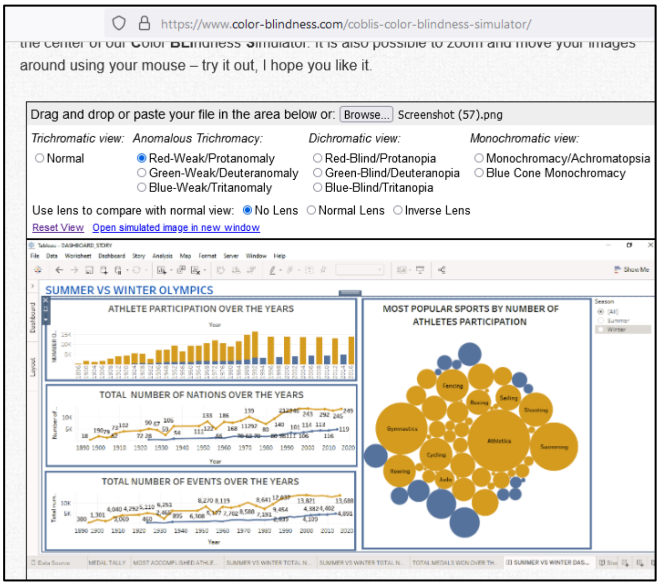
We have found that there is no research paper that has covered a full analysis on both GDP and Population effect on Olympics as well, as summer vs winter Olympics. We created an interactive dashboard which gives an overview of the summer and winter Olympics. We have also created an interactive dashboard of the Tokyo 2020 Olympics Event Schedule which has not been found in any other literature. This interactive event schedule dashboard can help working professionals, the audience and the Olympics participants understand the different events that are being held in different days of the Olympics calendar. We have also shown the effect of GDP and population in the Olympics Host Country. There has been no other literature who has widely covered the aspects of Olympics.

X. CHARTING

We have tried to implement the Gestalts Law of Perceptual Grouping in our visualizations, and they are:

- **Proximity:** In our visualizations we have made sure that it is divided into subsets and by using grids so thus it is easier for the audience to understand the relationship between the design elements.
- **Closure:** In the visualizations we have marked all the details such as the name and marks of the axis, legends and filters, and the color of each visualization.
- **Similarity:** We have tried to design the visuals in a way that they look like a part of similar group and not in a different way.
- **Good Continuation:** The visuals are place in an order in both the story as well as the dashboard. The pages in the story are followed by another as well as in dashboard, one can easily differentiate the visuals in the order.
- **Symmetry:** There is a symmetry that is maintained in all our charts. That is visuals that are similar such as the total number of nations over the years and the total number of events over the years are symmetrical.
- **Figure and Ground:** In our visuals we have made sure that there is a proper Figure and Ground. The text and the visuals are easily distinguishable, and we have also given proper space and have separated our visuals in the dashboard equally.

We have made sure that all our visualizations can be easily understood by people with color blindness. To check that we have checked using the color blindness simulator [14]. Fig. 24 shows the snapshot of how our dashboard looks on the color blindness simulator.



https://www.color-blindness.com/coblis-color-blindness-simulator/

the center of our Color BLindness Simulator. It is also possible to zoom and move your images around using your mouse – try it out, I hope you like it.

Drag and drop or paste your file in the area below or: Browse... Screenshot (57).png

Trichromatic view: ○ Normal

Anomalous Trichromacy: ● Red-Weak/Protanomaly ○ Green-Weak/Deuteranomaly ○ Blue-Weak/Tritanomaly

Dichromatic view: ○ Red-Blind/Protanopia ○ Green-Blind/Deuteranopia ○ Blue-Blind/Tritanopia

Monochromatic view: ○ Monochromacy/Achromatopsia ○ Blue Cone Monochromacy

Use lens to compare with normal view: ● No Lens ○ Normal Lens ○ Inverse Lens

Reset View Open simulated image in new window

Tableau - DASHBOARD_STORY
File Data Worksheet Dashboard Story Analysis Map Format Server Window Help
Show Me
SUMMER VS WINTER OLYMPICS
ATHLETE PARTICIPATION OVER THE YEARS
Year
TOTAL NUMBER OF NATIONS OVER THE YEARS
Year
TOTAL NUMBER OF EVENTS OVER THE YEARS
Year
MOST POPULAR SPORTS BY NUMBER OF ATHLETES PARTICIPATION
Fencing
Sailing
Boxing
Shooting
Gymnastics
Athletics
Swimming
Cycling
Rowing
Judo
Season
(All)
Summer
Winter
Dashboard
Layout

Fig. 24 Olympics Dashboard on the Color blindness simulator

Since we had a lot of important visualizations, we have didn't want to fill the dashboard with a lot of information and make it look messy, so we have created two main dashboards with similar filters and legends. Thus, since we wanted to portray all our information, we have combined our visuals and dashboards in a story form.

We have tried to choose our colors for visuals in such a way that it's not noisy and doesn't affect the eyes. We have made sure that our visualization can be easily understood by a worldwide audience in different categories We have portrayed our visualization in a story form so that it is easier to capture the details in the visualization and will stay in the memory of a person for a long time.

References

[1] Nansystar, "Why Are The Olympic Games So Important To Many of Us?," *Nansy Damianova*, 02-May-2013. [Online]. Available: https://nansydamianova.wordpress.com/2013/05/02/why-are-the-olympic-games-so-important-to-many-of-us/. [Accessed: 06-Oct-2021].

[2] Samruddhi Mhatre. (2020, September).Olympics Athlete Events Analysis, Version 1. Retrieved Sep 4, 2021 from https://www.kaggle.com/samruddhim/olympics-althlete-events-analysis

[3] Google. (n.d.). *Olympic event schedule.xlsx*. Google Drive. Retrieved September 5, 2021, from https://drive.google.com/file/d/1NF46NSX3lme_43e2QQaYRxKdf1hs1An-/view.

[4] Pradhan, R., Agrawal, K., & Nag, A. (2021). Analyzing Evolution of the Olympics by Exploratory Data Analysis using R. *IOP Conference Series. Materials Science and Engineering*, *1099*(1), 12058–. https://doi.org/10.1088/1757-899X/1099/1/012058

[5] J. Wen and X. Wang, "Study of the visualization and Interaction of data : Take the Historical Data of the Winter Olympics as an Example," *2020 International Conference on Innovation Design and Digital Technology (ICIDDT)*, 2020, pp. 78-82, doi: 10.1109/ICIDDT52279.2020.00022.

[6] M. Friendly, "A brief history of data visualization" in Handbook of data visualization, Berlin, Heidelberg:Springer, pp. 15-56, 2008.

[7] Perin, C., Vuillemot, R., Stolper, C. D., Stasko, J. T., Wood, J., & Carpendale, S. (2018, June). State of the art of sports data visualization. In Computer Graphics Forum (Vol. 37, No. 3, pp. 663-686).

[8] Balmer, Nigel J., Alan M. Nevill, and A. Mark Williams. "Modelling home advantage in the Summer Olympic Games." *Journal of sports sciences* 21.6 (2003): 469-478.

[9] DeFrantz, Anita. "The changing role of women in the Olympic Games." *Olympic Review* 26.15 (1997): 18-21.

[10] Nunes, Rita Amaral. "Women athletes in the Olympic Games." (2019).

[11] "2021: Week 29 - PD x WOW - Tokyo 2020 Calendar." *2021: Week 29 - PD x WOW - Tokyo 2020 Calendar*, 20 July 2021, https://preppindata.blogspot.com/2021/07/2021-week-29-pd-x-wow-tokyo-2020.html.

[12] Population, total. (n.d.). Retrieved September 5, 2021, from https://data.worldbank.org/indicator/SP.POP.TOTL

[13] GDP (current US$). (n.d.). Retrieved September 5, 2021, from https://data.worldbank.org/indicator/NY.GDP.MKTP.CD

[14] "Coblis - color blindness simulator," *Colblindor*. [Online]. Available: https://www.color-blindness.com/coblis-color-blindness-simulator/. [Accessed: 10-Oct-2021].