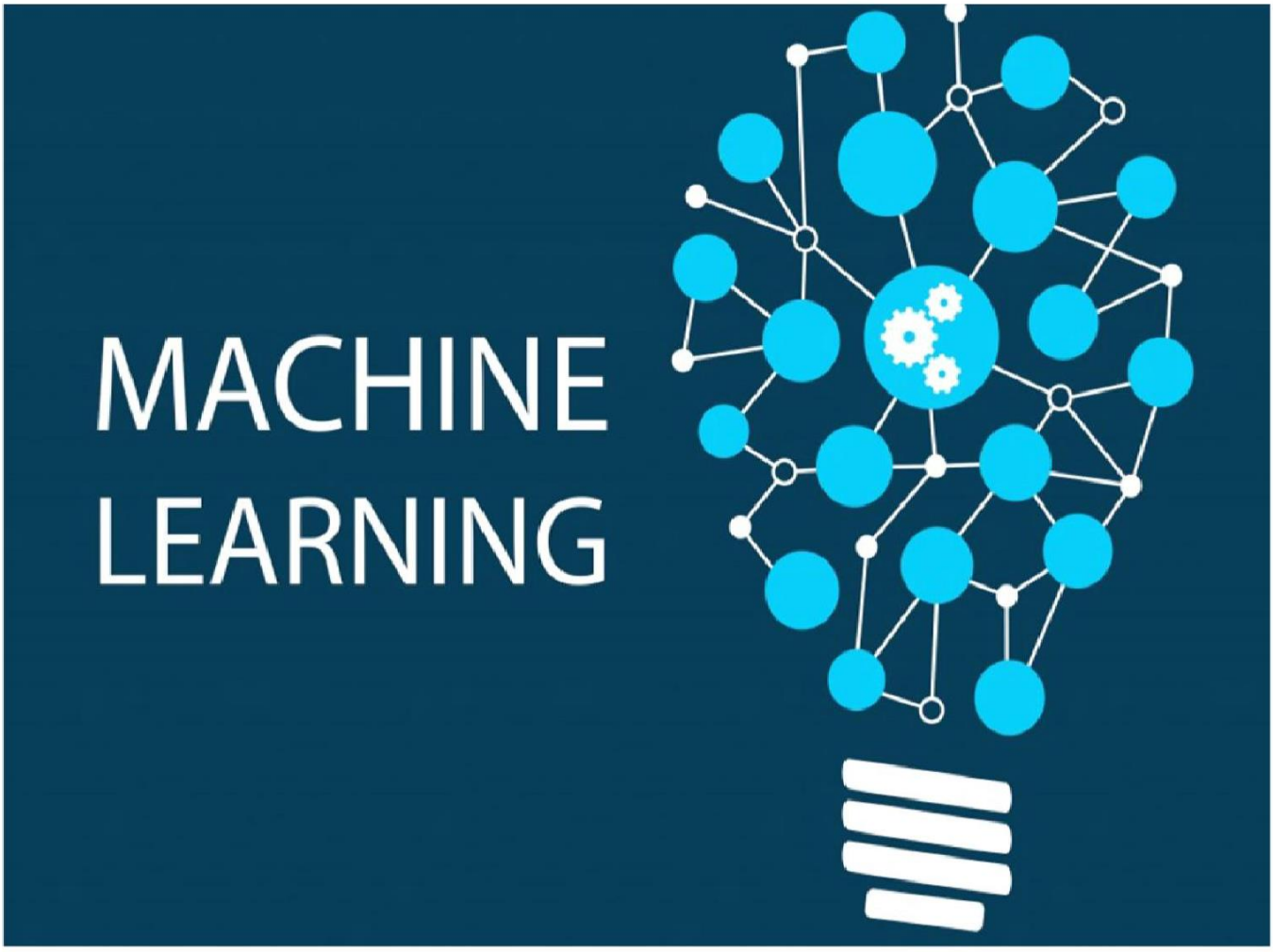



# Olympics Data Mining

Submitted to: Yogendra Sir

Submitted by: Gaurav Rajpurohit,
Gunjan Khandelwal, Kashish Mathur

# TABLE OF CONTENT


- Certificate
- Acknowledgement
- Abstract
- List of figures
- Introduction
- Theory
- Methodology
- Data Pre-processing
- Libraries
- Regression Method Used
- Classification Method used
- Data visualization
- Result Analysis
- Annexure

# Certificate

Date: 27/06/19

This is to certify that Mr. **Gaurav Rajpurohit**, student of 3rd Year from Department of Computer Science, Global Institute Of Technology, Sitapura, has undergone a Project work from May 1, 2019 to June 30, 2019 in **Data Science & Machine Learning** titled "**Olympics Data Mining**".

_____ _____

Project Incharge Seal

# Certificate

Date: 27/06/19

This is to certify that Ms. **Gunjan Khandelwal**, student of 3rd Year from Department of Computer Science, Global Institute Of Technology, Sitapura, has undergone a Project work from May 1, 2019 to June 30, 2019 in **Data Science & Machine Learning** titled "**Olympics Data Mining**".

_____ _____

Project Incharge Seal

# Certificate

Date: 27/06/19

This is to certify that Ms. **Kashish Mathur**, student of 3rd Year from Department of Computer Science, Global Institute Of Technology, Sitapura, has undergone a Project work from May 1, 2019 to June 30, 2019 in **Data Science & Machine Learning** titled "**Olympics Data Mining**".

_____ _____

Project Incharge Seal

# Acknowledgement

We would like to express our special thanks of gratitude to our course instructor Mr. Yogendera Singh and Dr. Sylvester Fernandes who gave us the golden opportunity to do this wonderful project on the topic Olympic Data Mining, which also helped us in getting to know about Olympics events. We would also like to show our gratitude towards Rohit sir and Kunal sir, who helped us throughout to develop our project. Last but not least we would like to thank our friends who helped us to improve our project.

# Abstract

Rapid growth of data comes with a challenge of sorting and analysing them, where raw data exists in graphical form, textual form or in images. Data science and machine learning finds its application in various fields like stock market, recommendation systems, image processing, aerial photography, military, weather forecasting etc.

This report is about our project on “Olympics Data Mining” that addresses about data pre-processing which includes cleaning data and plotting various queries regarding data and the ability of machine learning algorithms to deal with different set of data. In this project, we have tackled three different datasets which are combined into one dataset and an unsupervised machine learning problem. We have used the KMeans algorithm for dividing the height and weight dataset into clusters. In addition to this, we have also made use of several libraries to plot the data points on different types of graph.

# List of Figures


1. Data Science
2. Decision Tree
3. Logistic Regression
4. Linear Regression
5. KNN Classification
6. K-means clustering
7. Athlete participation count over years.
8. Athlete Participation by gender over years.
9. Gender Distribution in the games.
10. Number of countries participated in the games.
11. Highest number of participation nation wise.
12. Countries that hosted the games for the highest number of times.
13. Cities that hosted the games for the highest number of times.
14. Average age, height and weight of the athletes for various sports categories. (3 separate representation)
15. Total unique sports activities over years in Olympics.
16. Event ratio by gender.
17. Visualize the events by genders over years.
18. Visualize the sports ratio in each revenue category.
19. Visualize the medals won in each revenue category.

20. Visualize above results

21. Visualize the top 100 athletes with the highest total medal first separate for both seasons and then in combined form.

# Introduction

According to historical records, the first ancient Olympic Games can be traced back to 776 BC. They were dedicated to the Olympian gods and were staged on the ancient plains of Olympia. They continued for nearly 12 centuries, until Emperor Theodosius decreed in 393 A.D. that all such "pagan cults" be banned. This is a historical dataset on the modern Olympic Games, including all the Games from Athens 1896 to Rio 2016. The target is to analyse the data properly, visualizing various queries and using the unsupervised machine learning algorithm to divide the data into clusters.

Our approach to the problem is very simple, we first retrieve the dataset that consists of all the athlete details participated in the Olympics till Rio 2016, dataset of all the host countries and the NOC codes assigned to all the countries participating. Then, we will check the dataset for some missing data i.e. the data pre-processing part. Then, we move towards visualizing the data using different methods like bar graph, pie chart, donut chart etc.

Finally, we compare the WCSS score and silhouette score to decide the number of cluster the data can be divided and choose the appropriate number of clusters.

# Theory

Data science is a "concept to unify statistics, data analysis and their related methods" in order to "understand and analyse actual phenomena" with data. It employs techniques and theories drawn from many fields within the broad areas of mathematics, statistics, information science, and computer science, in particular from the subdomains of machine learning, classification, cluster analysis, data mining, databases, and visualization.

Data science – discovery of data insight

This aspect of data science is all about uncovering findings from data. Diving in at a granular level to mine and understand complex behaviours, trends, and inferences. It's about surfacing hidden insight that can help enable companies to make smarter business decisions.

For example:

Netflix data mines movie-viewing patterns to understand what drives user interest, and uses that to make decisions on which Netflix original series to produce

Data science – development of data product

A "data product" is a technical asset that: (1) utilizes data as input, and (2) processes that data to return algorithmically generated results. The classic example of a data product is a recommendation engine, which ingests user data, and makes personalized recommendations based on that data.

For example:

Amazon's recommendation engines suggest items for you to buy, determined by their algorithms. Netflix recommends movies to you. Spotify recommends music to you.

Machine learning and statistics are part of data science. The word learning in machine learning means that the algorithms depend on some data, used as a training set, to fine-tune some model or algorithm parameters. This encompasses many techniques such as regression, naive Bayes or supervised clustering.

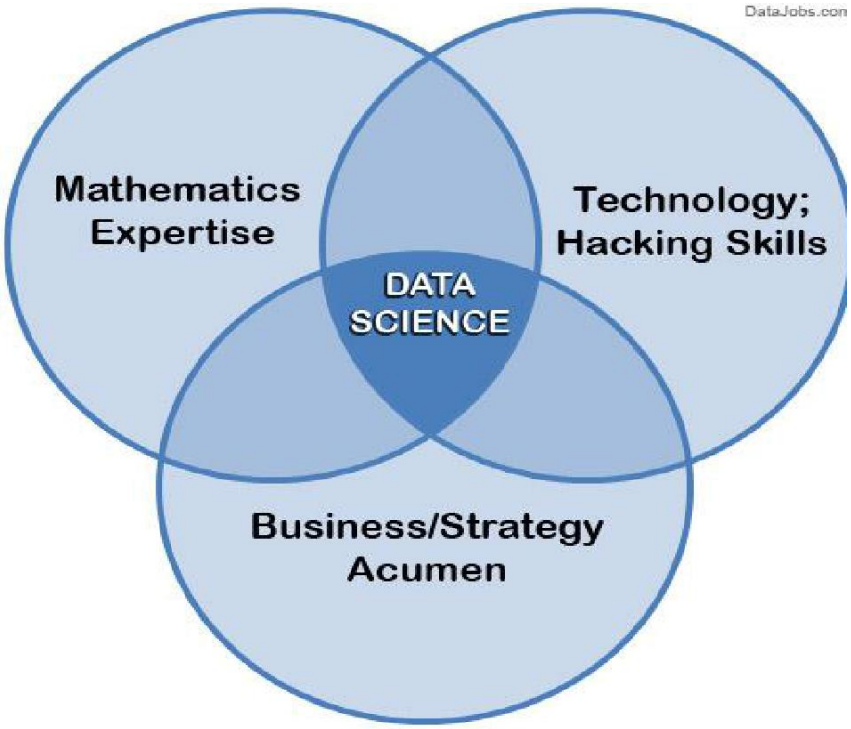



**Fig. 1 Data Science**

Supervised and unsupervised learning describe two ways in which machines algorithms can be set loose on a data set and expected to learn something useful from it.

**Supervised:**

If we are training our machine-learning task for every input with corresponding target, it is called supervised learning, which will be able to provide target for any new input after sufficient training. Our learning algorithm seeks a function from inputs to the respective targets. If the targets are expressed in some classes, it is called *classification* problem. Alternatively, if the target space is continuous, it is called *regression* problem.

- **Regression** analysis is widely used for prediction and forecasting, where its use has substantial overlap with the field of machine learning. Regression analysis is also used to understand which among the independent variables are related to the dependent variable, and to explore the forms of these relationships. In restricted circumstances, regression analysis can be used to infer causal relationships between the independent and dependent variables.

- **Classification** model attempts to draw some conclusion from observed values. Given one or more inputs a classification model will try to predict the value of one or more outcomes. Outcomes are labels that can be applied to a dataset.

**Unsupervised:** If we are training our machine-learning task only with a set of inputs, it is called unsupervised learning, which will be able to find the structure or relationships between different inputs. Most important unsupervised learning is *clustering*, which will create different cluster of inputs and will be able to put any new input in appropriate cluster.

- **Cluster** analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, bioinformatics, data compression, and computer graphics.

1. **Decision Trees:** A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance-event outcomes, resource costs, and utility.

   From a business decision point of view, a decision tree is the minimum number of yes/no questions that one has to ask, to assess the probability of making a correct decision, most of the time. As a method, it allows you to approach the problem in a structured and systematic way to arrive at a logical conclusion.

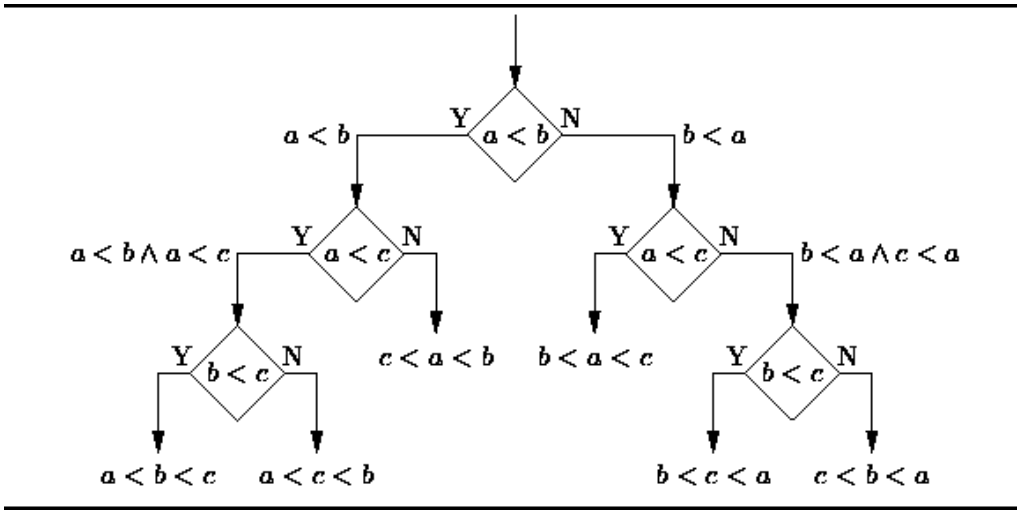



**Fig. 2 Decision Tree**

2. **Logistic Regression:** Logistic regression is a powerful statistical way of modelling a binomial outcome with one or more explanatory variables. It measures the relationship between the categorical dependent variable and one or more independent variables by estimating probabilities using a logistic function, which is the cumulative logistic distribution.

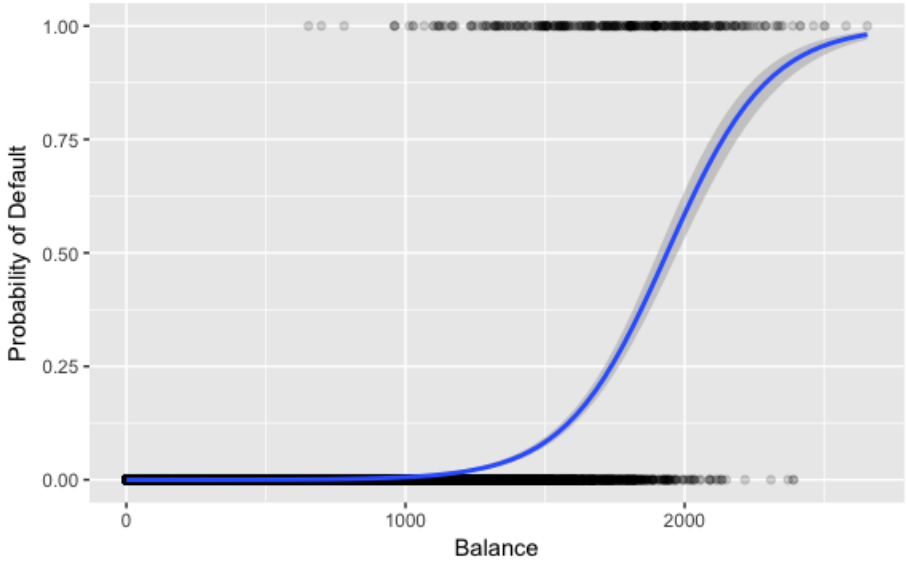



**Fig. 3 Logistic Regression**

## 3. Linear Regression

It is used to estimate real values (cost of houses, number of calls, total sales etc.) based on continuous variable(s). Here, we establish relationship between independent and dependent variables by fitting a best line. This best fit line is known as regression line and represented by a linear equation Y= a *X + b.

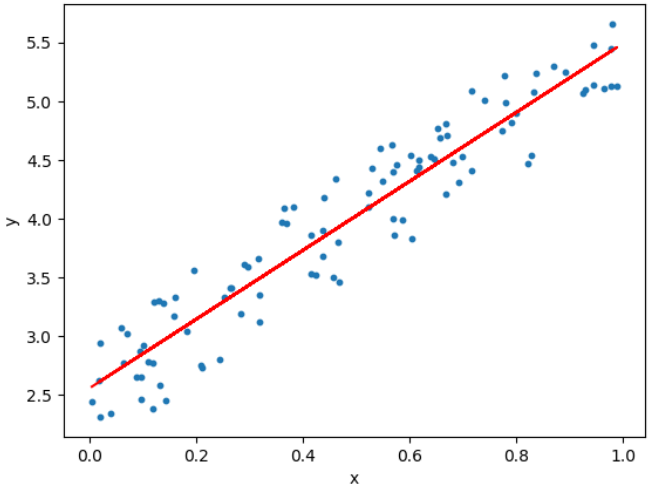



**Fig. 4 Linear Regression**

### 4. KNN (K- Nearest Neighbours)

It is also a lazy algorithm. What this means is that it does not use the training data points to do any generalization. K nearest neighbours is a simple algorithm that stores all available cases and classifies new cases by a majority vote of its k neighbours. The case being assigned to the class is most common amongst its K nearest neighbours measured by a distance function. These distance functions can be Euclidean, Manhattan, Minkowski and Hamming distance. First three functions are used for continuous function and fourth one (Hamming) for categorical variables. If K = 1, then the case is simply assigned to the class of its nearest neighbour. At times, choosing K turns out to be a challenge while performing KNN modelling.

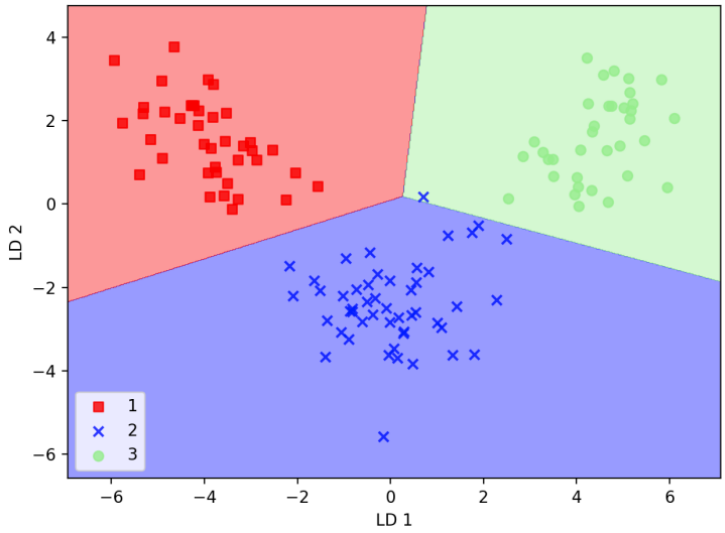



**Fig. 5 K-Nearest Neighbours Classification**

## 5. CLUSTERING (K-means)

K-means clustering is a clustering algorithm that aims to partition n observations into k clusters.

There are 3 steps:

- Initialization – K initial “means” (centroids) are generated at random
- Assignment – K clusters are created by associating each observation with the nearest centroid.
- Update – The centroid of the clusters becomes the new mean

Assignment and Update are repeated iteratively until convergence

The end result is that the sum of squared errors is minimized between points and their respective centroids.

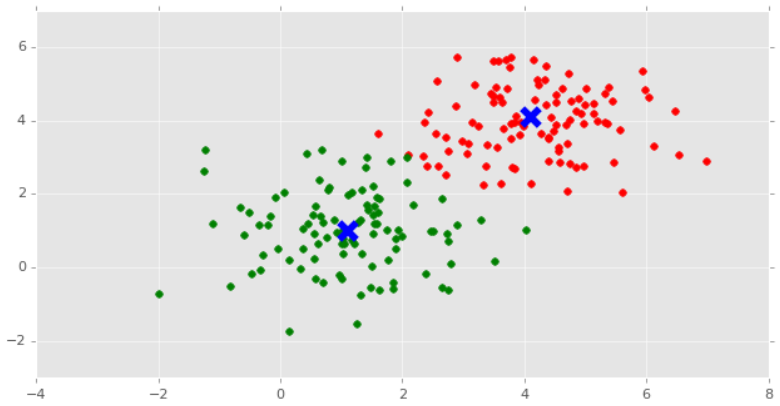



**Fig. 6 Clustering with K-means Classification**

# Methodology

To understand the methodology adopted, we first understand our dataset and the variables.

**Dataset**

Our dataset consists of roughly 2,69,369 athlete information having their name, age, height, weight, year, season, medal, host city and country, latitude, longitude etc. The target is to visualize this data and perform clustering using KMeans method.

**Features:**

- ID: Unique id provided to the athlete
- Name: Athlete's name
- Sex: Gender (F/M)
- Age: Athlete's age
- Height: Athlete's height
- Weight: Athlete's weight
- Team: Name of the team
- NOC: National Olympic Committee 3-letter code
- Year: Year of participation
- Season: Season of participation
- City: Host city
- Sport: Sport of the athlete
- Event: Name of the event
- Medal: Medal won (Gold, Silver, Bronze)
- Region: Region of the athlete
- Country: Host country
- Latitude: Coordinate of the host city
- Longitude: Coordinate of the host city

# Data Pre-processing

For pre-processing, we considered Python as our options for the project. After some experimentation, we found that while R was easier for statistics and analysis of the data, the lack of uniformity among the various ML packages made Python our preference. The ML algorithms provided by the scikit-learn package do not function if the input data has missing values. Hence we either had to impute data at the missing slots or remove the instances that had these missing fields. Upon combining the three datasets into a single dataset, we found that our data consisted information of cancelled Olympics, there were various missing values in the age, height and weight field. To fill these values we used another dataset retrieved from the Google and tried to fill the missing values and the dropped the data of the cancelled Olympics.

Additionally, we combine several columns into one columns and dropped the unnecessary columns from the dataset.

# Libraries used

- numpy
- pandas
- matplotlib.pyplot
- plotly.plotly
- plotly.offline
- plotly.graph_objs
  - download_plotlyjs
  - init_notebook_mode
  - plot
- plotly.io
- folium
- scikitlearn
  - sklearn.cross_validation
  - sklearn.cluster
  - sklearn.neighbors
  - sklearn.ensemble
  - sklearn.metrics

# Classification Method

Here for making classification of the athlete on the basis of height and weight we used K-means clustering. We made height and weight as features. Then using WCSS elbow method and silhouette method we found out that number of clusters are 3. We classified those clusters into 3 categories: Low static sports, Medium static sports and High static sports.

Then we plotted a graph through matplotlib.pyplot library and visualized it.

# Data visualization

For doing visualization of the data we had used libraries like plotly. Through plotly we were able to visualize our queries through different graphs. Like, number of medals won by top athlete were visualize using the marker attribute of the scatter function of the plotly, we used bar graphs to show the average age, height and weight required by each sport, bubble chart was used to visualize the ratio of male and female in the sports. We used the geo attribute of the Bar graph to show the world map with host city being marked and lots more.

# Result Analysis

We plotted following graphs:

1. The participant who won the most medals with their sports and country


**Fig. 6**

2. Athlete participation count over years

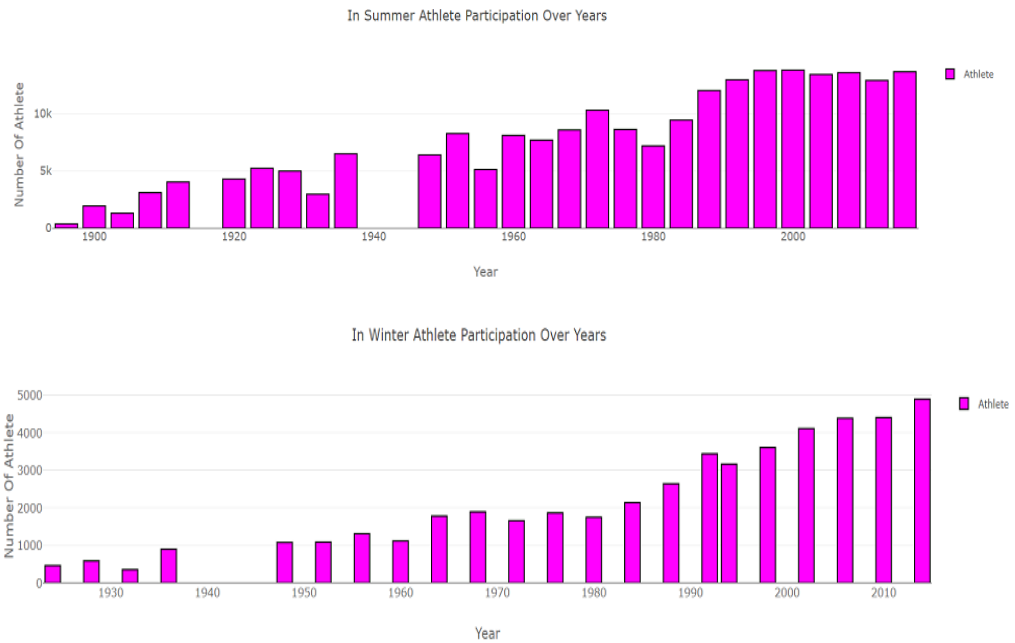



**Fig. 7**

3. Athlete Participation by gender over years

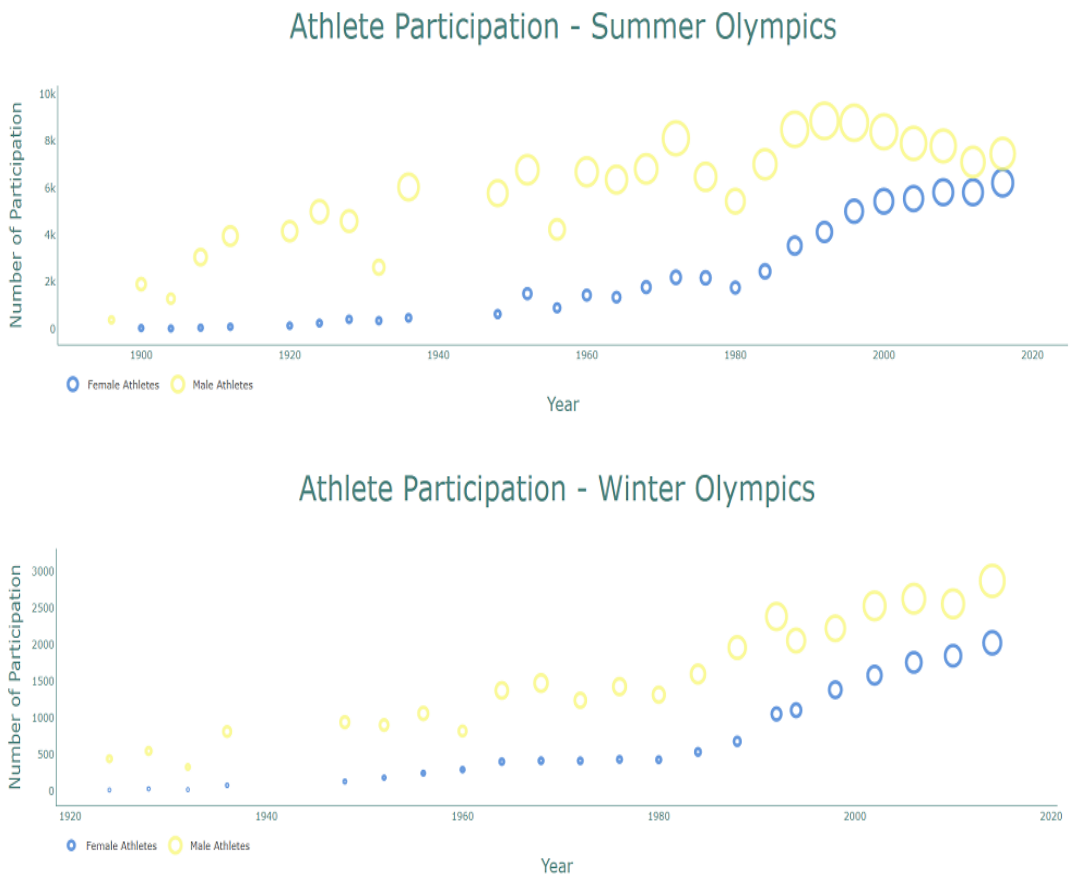
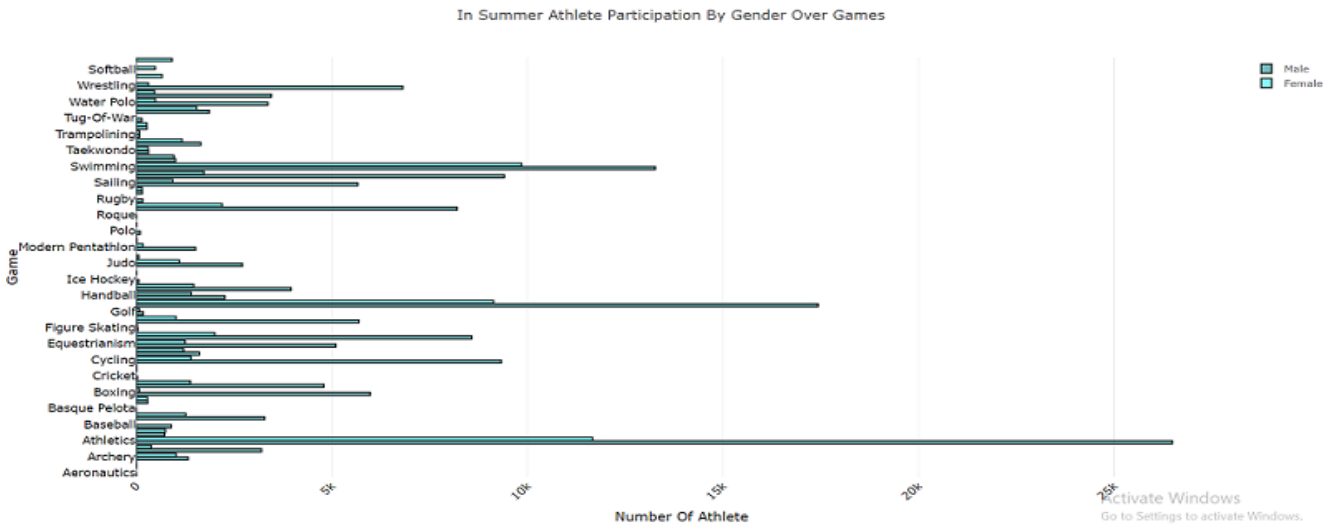



**Fig. 8**

4. Gender Distribution in the games



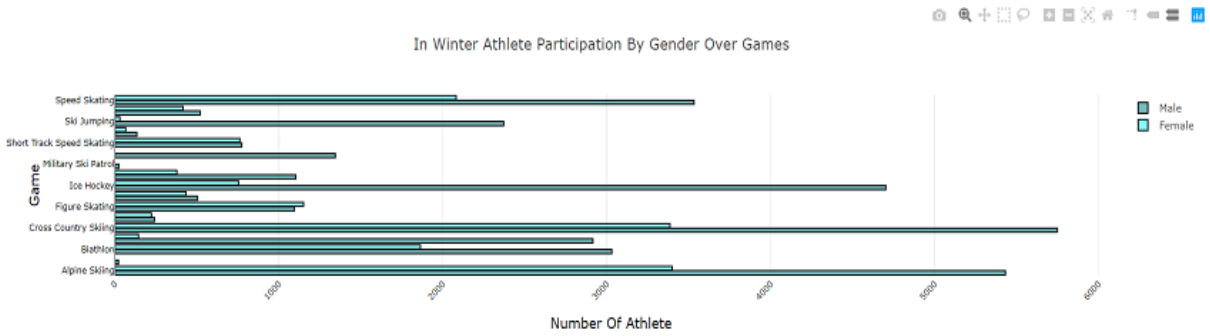
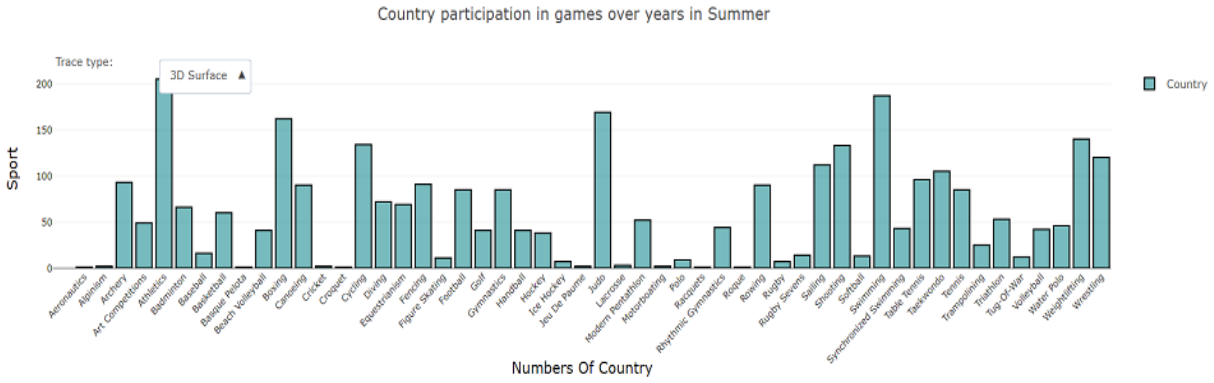



**Fig. 9**

5. Number of countries participated in the games



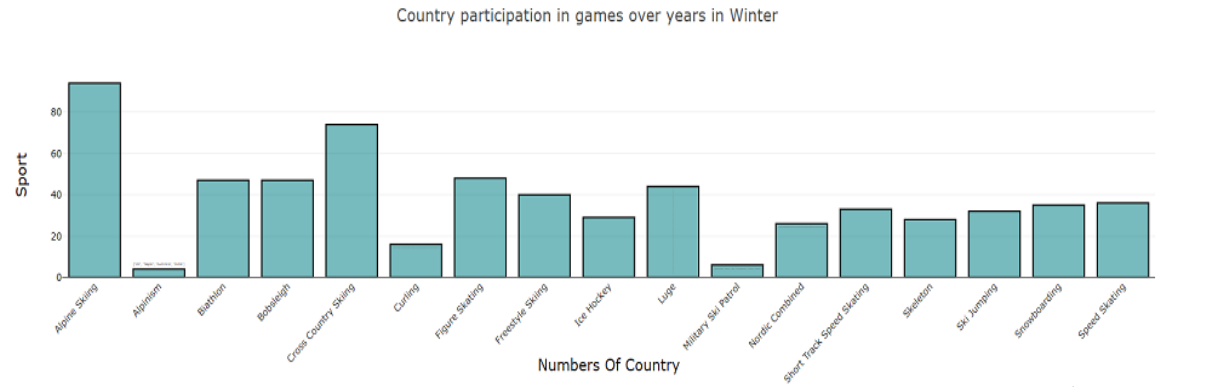




**Fig. 10**

6. Highest number of participation nation wise

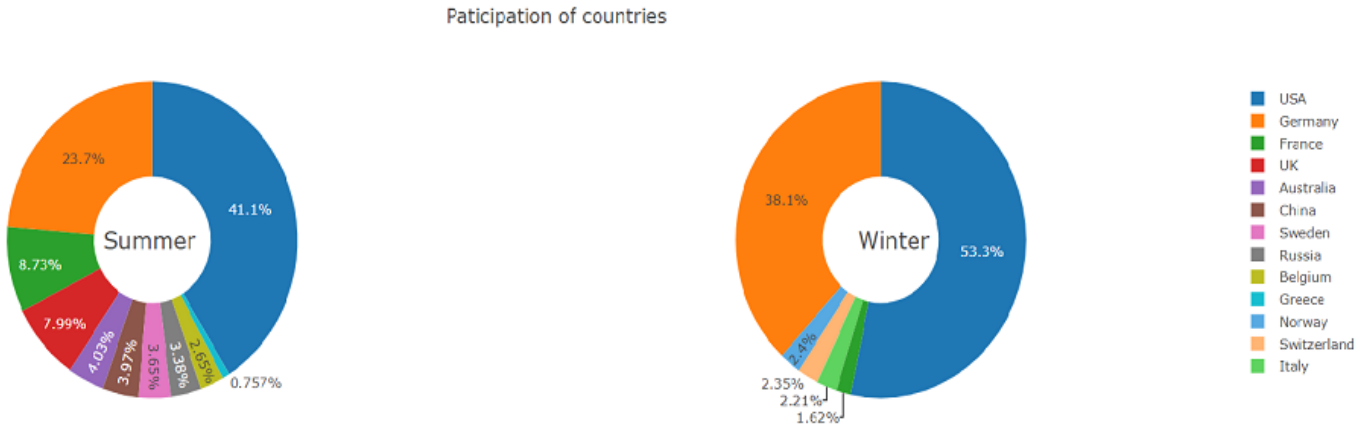



**Fig. 11**

7. Countries that Hosted the games for the highest number of times

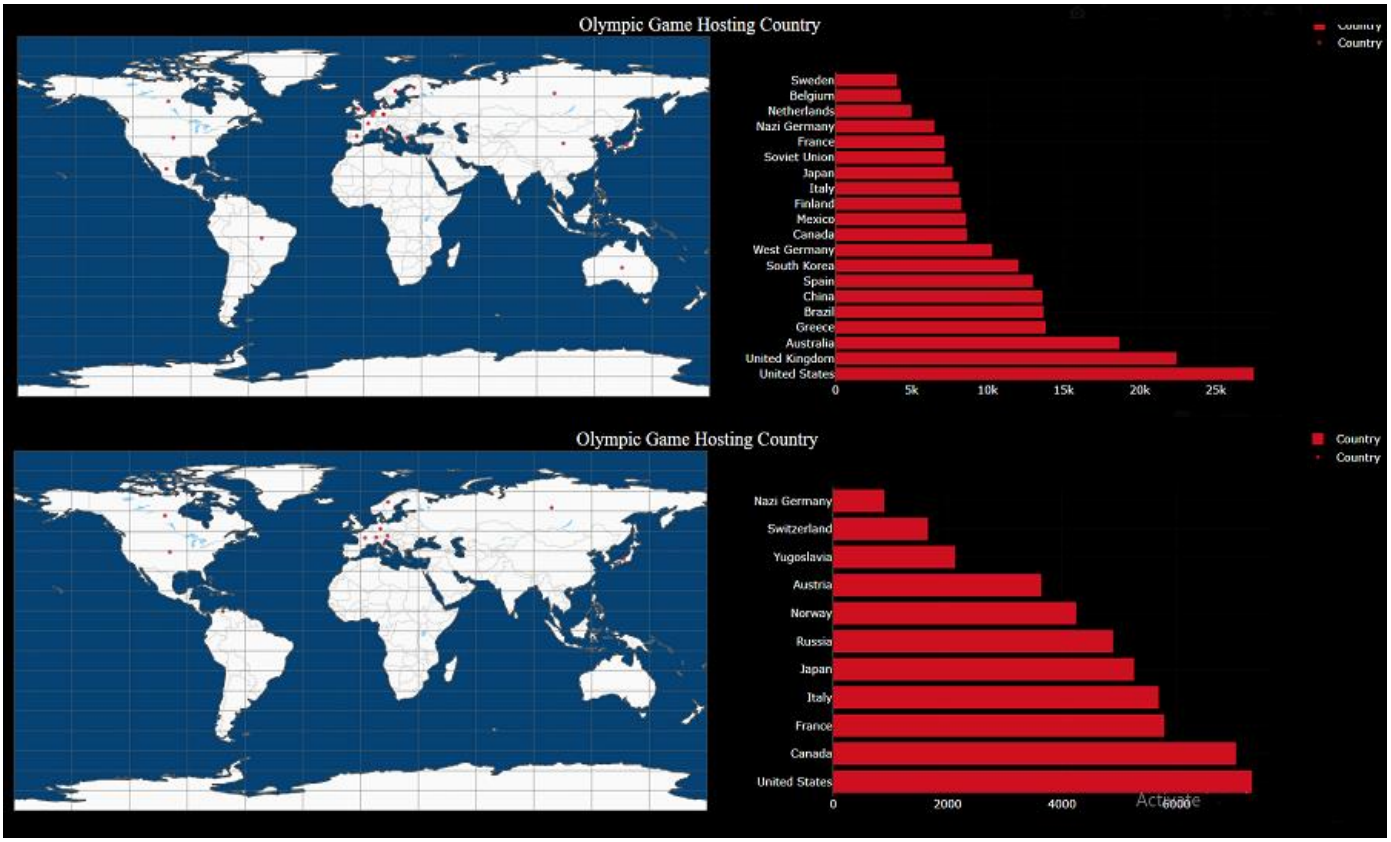



**Fig. 12**

8. Cities that hosted the games for the highest number of times

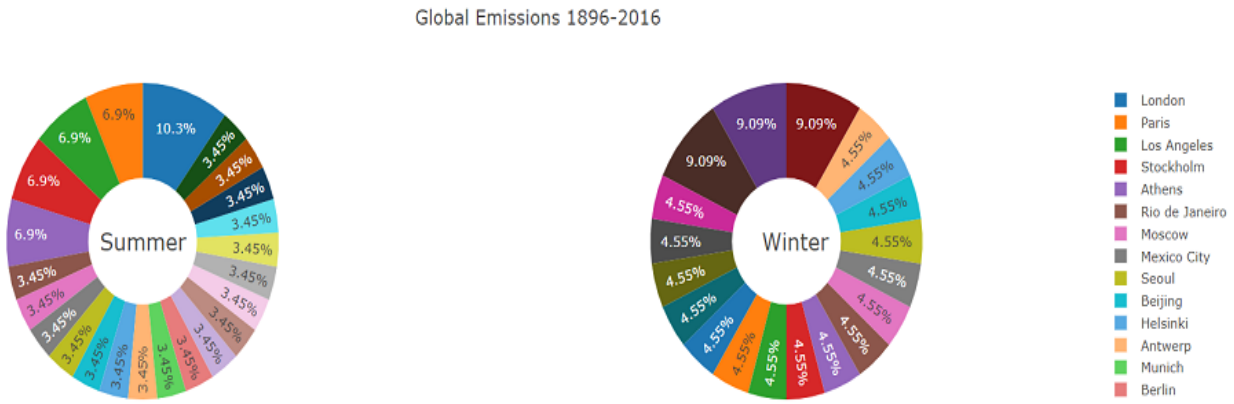



**Fig. 13**

9. The average age, height and weight of the athletes for various sports categories

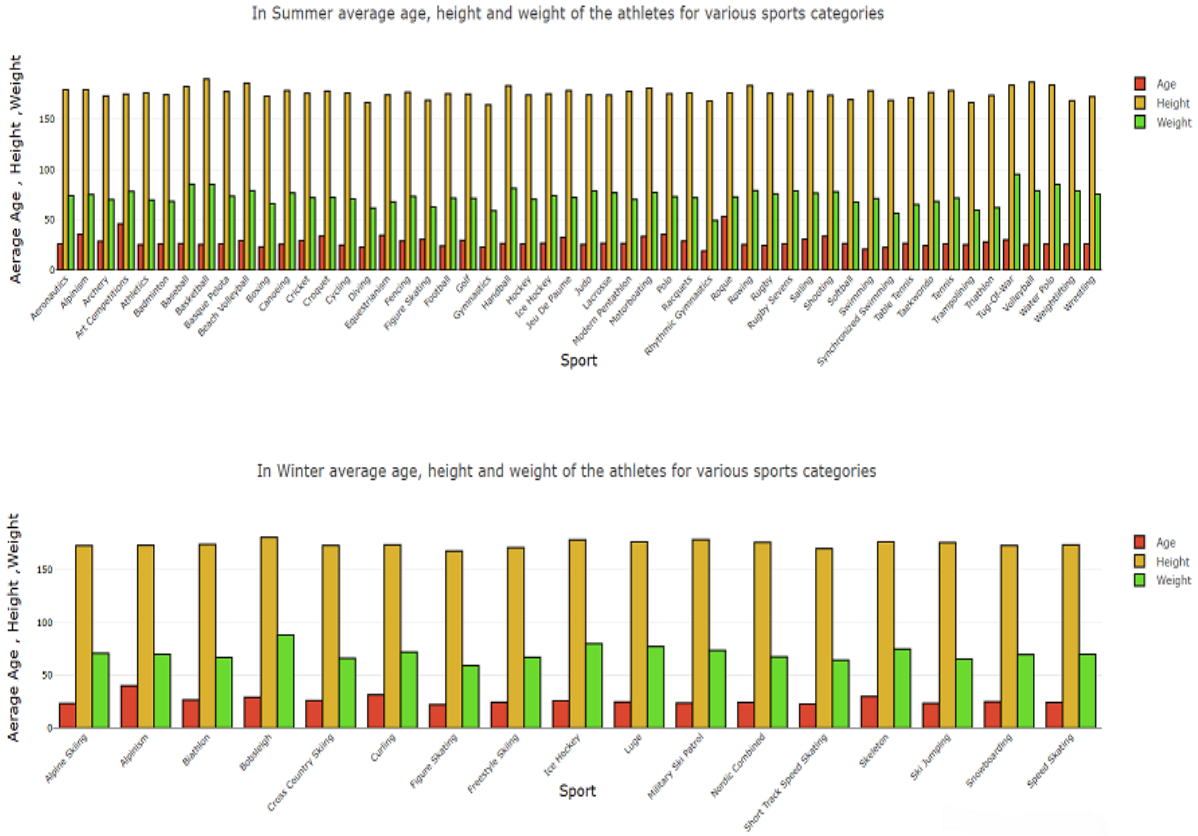



**Fig. 14**

10. Total unique sports activities over years in Olympics

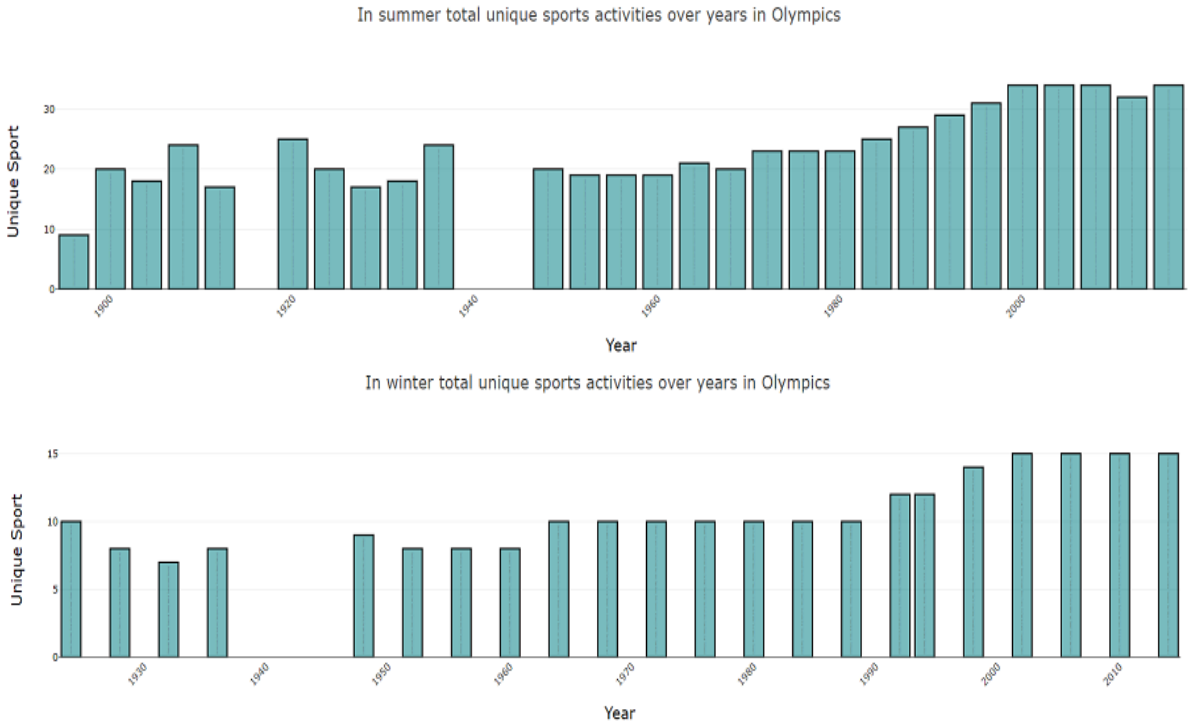



**Fig. 15**

## 11. Event ratio by gender

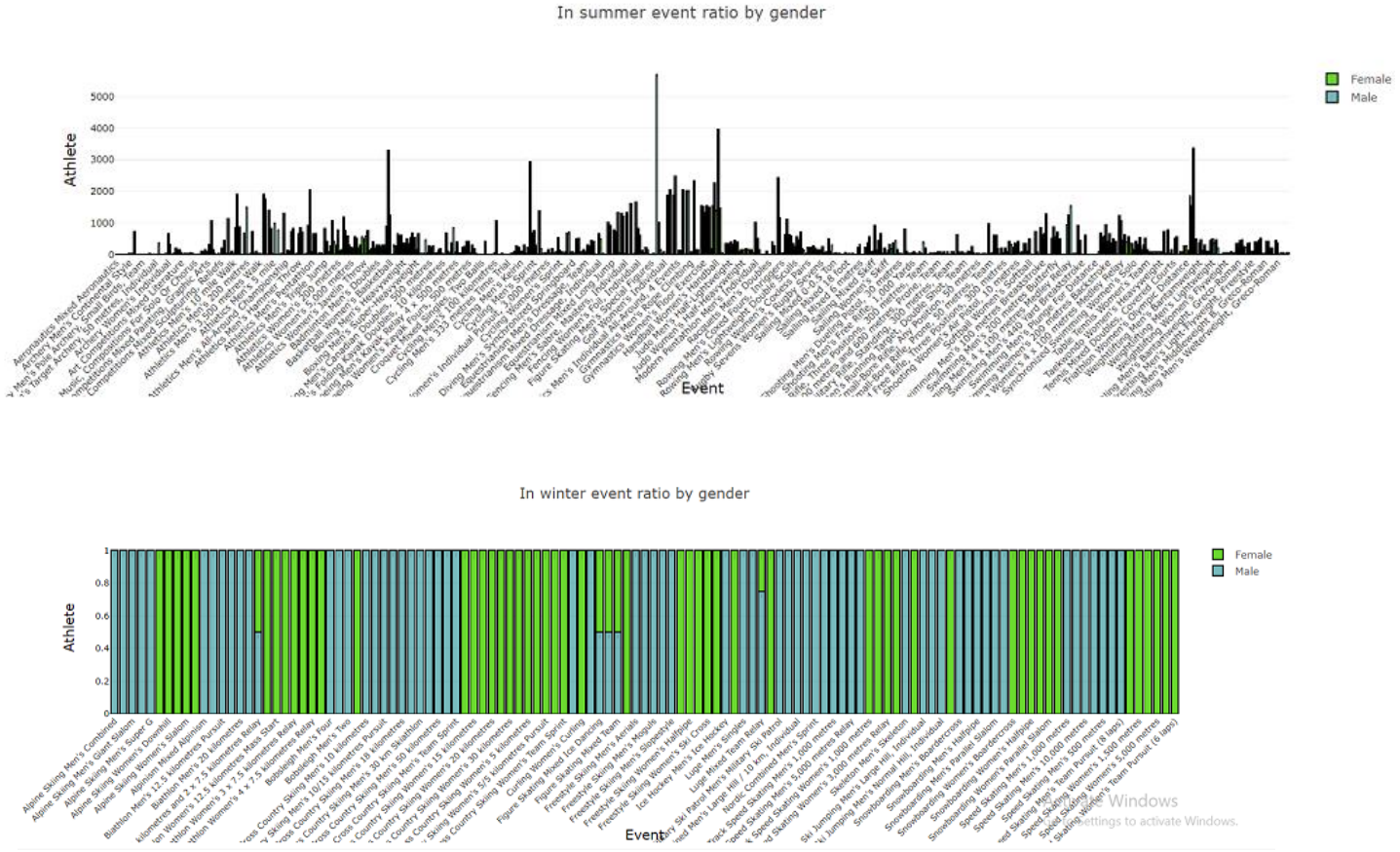



**Fig. 16**

## 12. Events by genders over years

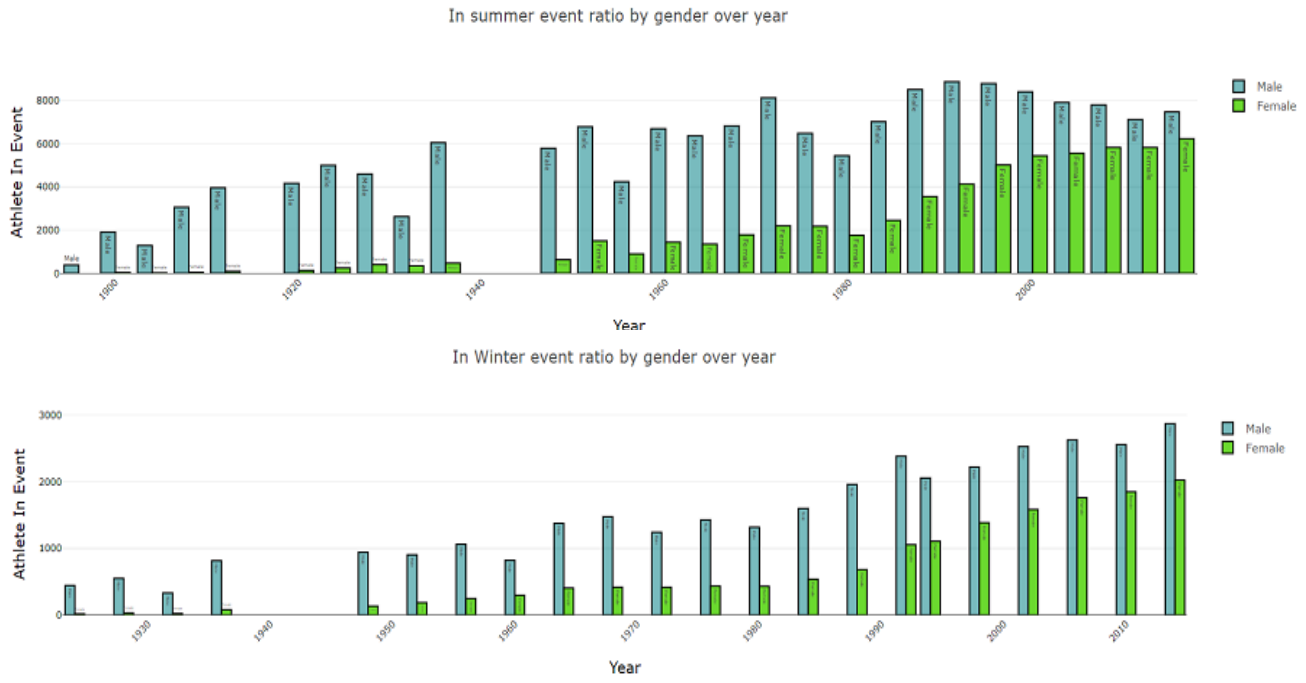



**Fig. 17**

13. Sports ratio in each revenue category

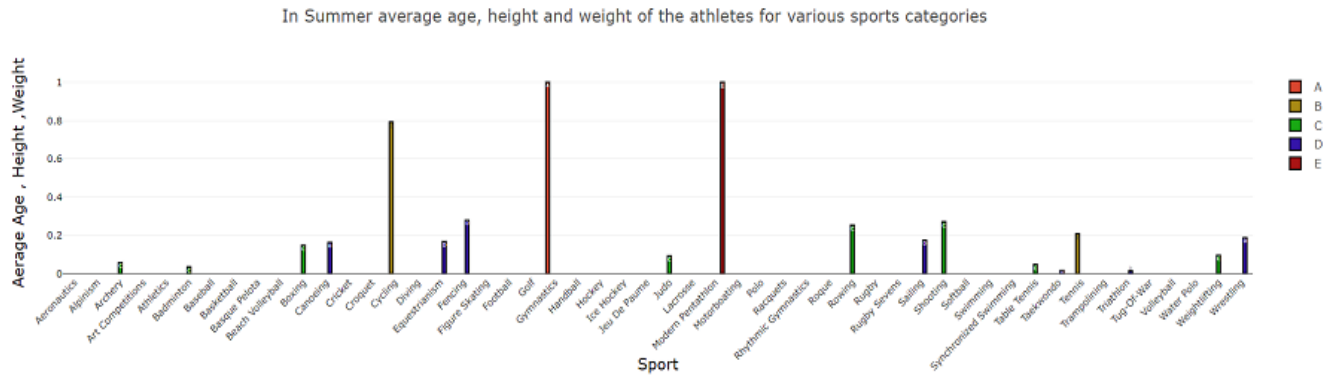



**Fig. 18**

14. Medals won in each revenue category

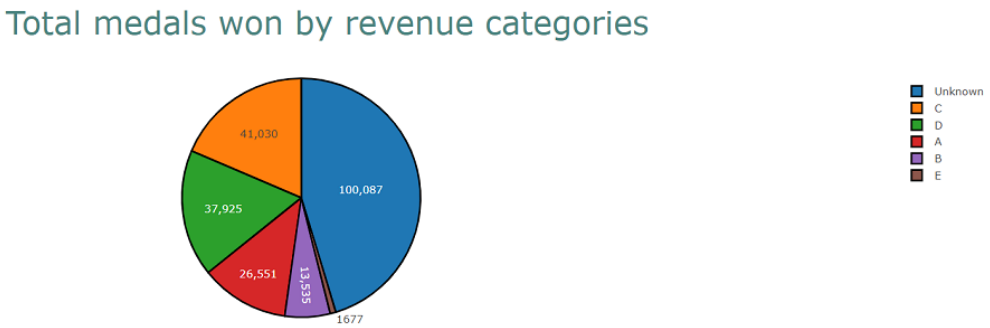



**Fig. 19**

15. Visualize above results

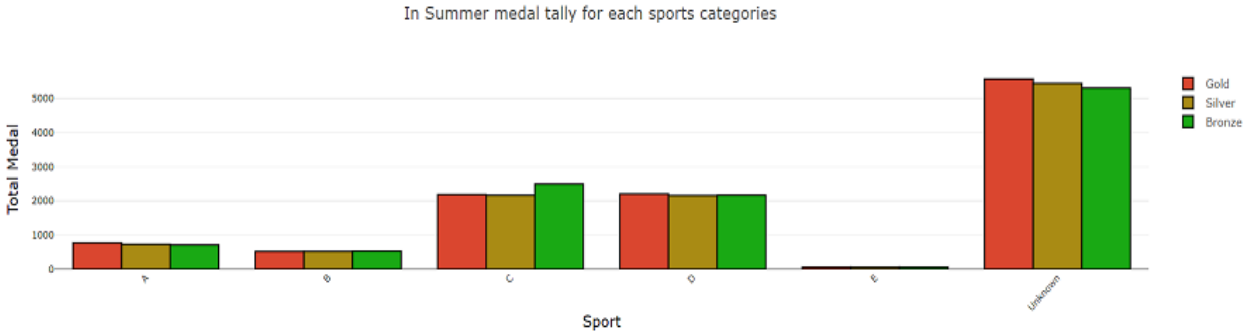
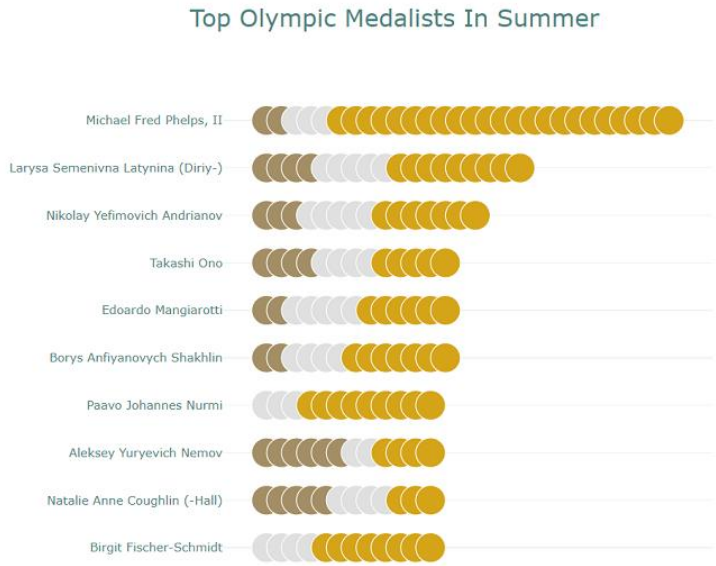
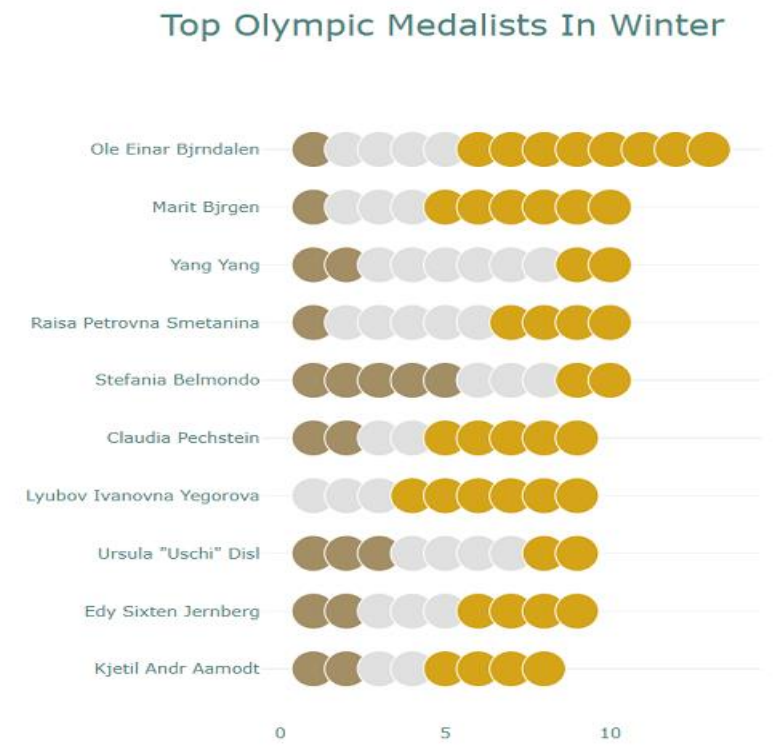



**Fig. 20**

16. Top 100 athletes with the highest total medal first separate for both seasons and then in combined form.




Top Olympic Medalists In Winter
Ole Einar Bjrndalen
Marit Bjrgen
Yang Yang
Raisa Petrovna Smetanina
Stefania Belmondo
Claudia Pechstein
Lyubov Ivanovna Yegorova
Ursula "Uschi" Disl
Edy Sixten Jernberg
Kjetil Andr Aamodt
0
5
10

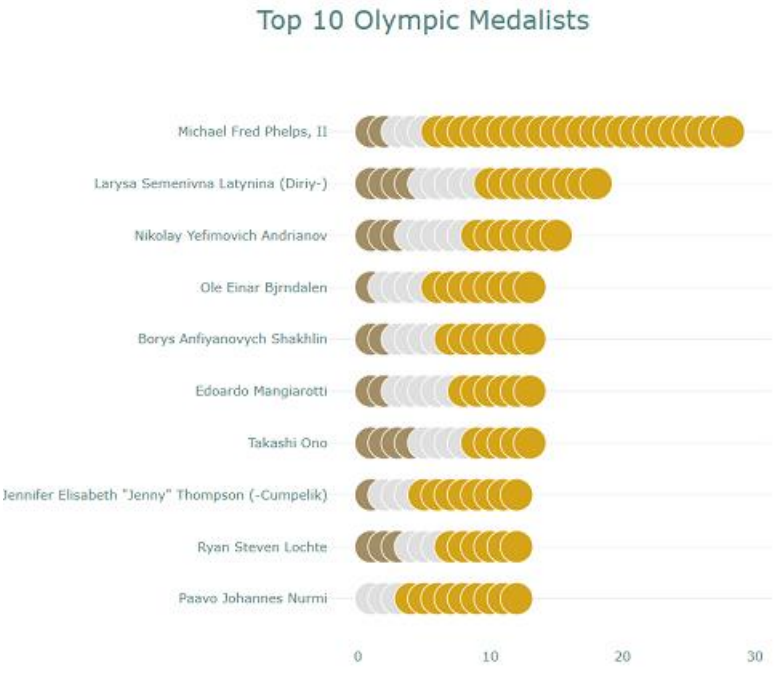

Top 10 Olympic Medalists
Michael Fred Phelps, II
Larysa Semenivna Latynina (Diriy-)
Nikolay Yefimovich Andrianov
Ole Einar Bjrndalen
Borys Anfiyanovych Shakhlin
Edoardo Mangiarotti
Takashi Ono
Jennifer Elisabeth "Jenny" Thompson (-Cumpelik)
Ryan Steven Lochte
Paavo Johannes Nurmi
0
10
20
30

**Fig. 21**

# Conclusion

Here in this project we visualized the data of Olympics events, sports, season and athletes. We precisely classified the Olympics athletes data based on their height and weight which represents what type of games they play. We mapped the cities and countries where Olympics events took place. We also made the depiction of the upcoming athletes which type of sport they can play.