Data Analysis of olympic Games(120 Years)

Parshwa Shah
Department of Computer Science
University of Windsor
Windsor, Canada
Student id:110021970
shah52@uwindsor.ca

Jaimin Patel
Department of Computer Science
University of Windsor
Windsor, Canada
Student id:110017550
patel2go@uwindsor.ca

Soham Patel
Department of Computer Science
University of Windsor
Windsor, Canada
Student id:110017511
patel2gh@uwindsor.ca

***Abstract*—The Olympic games have been popular all over the world for decades. Every year we get new champions and participants. As per our research, we found an article on Kaggle and found a historical dataset of 2.8 million records including all the Games from Athens 1896 to Rio 2016. So, we have decided to analyze the data using MongoDB. In every Olympic year, there are new participants and new champions in different games and countries. We know the benefits and functionality of MongoDB like it varies, is easy to scale, supports dynamic queries and easy to install and setup. Based on data available of Olympics predictions can also be done about the response of new games, new participants and many more. We have not worked on MongoDB before; that is why we selected it because it will be a new skill to develop and experience for us.**

I. Introduction

In our project, we have analyzed the 120 years of Olympic data using MongoDB, and that is visualized using Jupyter Notebook and different python libraries. With the help of data analysis and visualization, we can analyze how the Olympics has been evaluated in a span of 120 years. It would also give the information about how the number of participants changed over the years and how the number of male and female participants increased and many more cases have been analyzed for that we have divided our project into two phases: Data Analysis and Data Visualization. In the data analysis, we imported 2.8 million of data into MongoDB compass. Which was built to address three primary goals like schema discovery, data discovery and visual construction of queries. All the data is displayed with datatypes in compass using schema discovery. With the help of a data discovery compass is able and displays the histogram to represent the data distribution and frequency within the data collections. Using different queries, the collected data is stored in text files and that is used in data visualization. In the data visualization, we have imported the data files and the different libraries like Numpy, pandas, Seaborn and Matplotlib. First, the data is cleaned and all the null values are removed before the data visualization. Different graphs are created using the above-mentioned libraries. This is how our project has developed the best way to analyze the Olympic evolution over the 120 years.

II. Objective

The main objective of this project is to provide information about the development of different Olympic games over the years and how the new games have become successful and famous. Also, it will provide an overview to the introduction of a new game if it will become successful or not.

III. Related Work

They have characterized data into events, sports level data and athletes level data. They have also excluded art competitions. Their application is a concrete analysis example. They have implemented the project using various packages of the R language, including tidyverse, gridExtra, knitr and gganimate. They have detailed observations using a name, sex, age, weight, game type, location. However, they did not execute with all of these.

IV. Tools and Techniques

A. Data analysis

For data analysis tools and languages like MongoDB Compass, PyCharm, python and pymongo are used. MongoDB Compass is considered as a Graphical User Interface for MongoDB. It allows users to do the analysis of the data in a more straightforward way. Aggregation operations on MongoDB can be executed in a quick and easy way by using aggregation in compass. It uses the concept of a pipeline to process the given data, so documents in a collection pass through different stages and at the end the final result is displayed. In short, the output of one stage goes to the input of another stage and so on. In order to do the analysis, we imported all records in MongoDB then created various cases to do analysis. Cases such as a list of gold medalists greater than 10, female winner by year and many more. All cases were implemented in MongoDB Compass. In order to execute all cases to gather, we used python language and pymongo library. This library acts as a bridge between Python and MongoDB. As a result output of all cases can be seen when this python program is executed.

B. Data Visualization

Data visualization: Data visualization means to take out information from large amount of data and put it on the visual context. Because human brain to understand the data easier to detect trends from charts and graphs. It makes complex datasets into more meaningful and clearer into message. In out project for Advanced database system we are using Jupyter

notebook as a platform to perform data visualization. Python libraries are being more popular to create graphs and charts for more complex datasets and makes it better to visualize the data. Here, we are using NumPy, Pandas, Seaborn, Matplotlib libraries to create different types of graphs and charts. NumPy helps to work with arrays, and array object is faster than the traditional python lists. Pandas is a library written for python that helps to manipulate data which contains data structures and many operations to analyze numerical tables and time series. To create high level statistical and informative graphs seaborn is the best option. Because seaborn is based on matplotlib. Matplotlib uses the numerical extensions of NumPy and it is a plotting library. It helps to create animated and interactive visualizations in Python.

V. Result

A. Data Analysis

In data analysis, when all the cases were executed, we got various statistics of the Olympic games. In the form of a result, one can see an increase in different sports year-wise, maximum gold winner by country and many more. Below is the screenshot generated as an output from the MongoDB Compass and PyCharm.

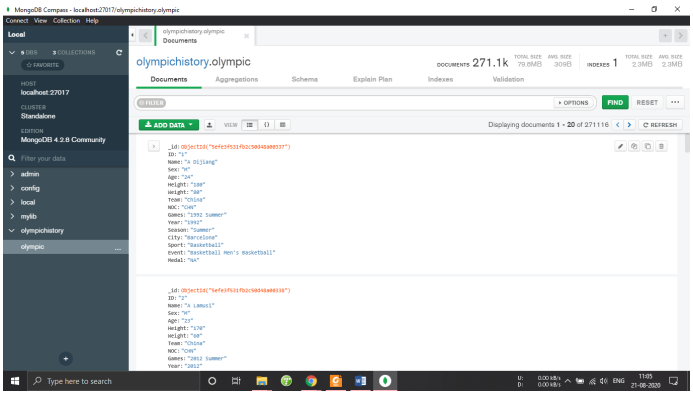

Fig. 1. Query Data

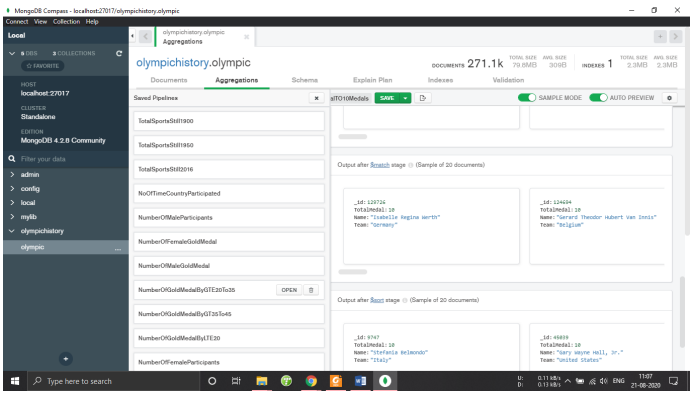

Fig. 2. Number of Gold medal by age Group

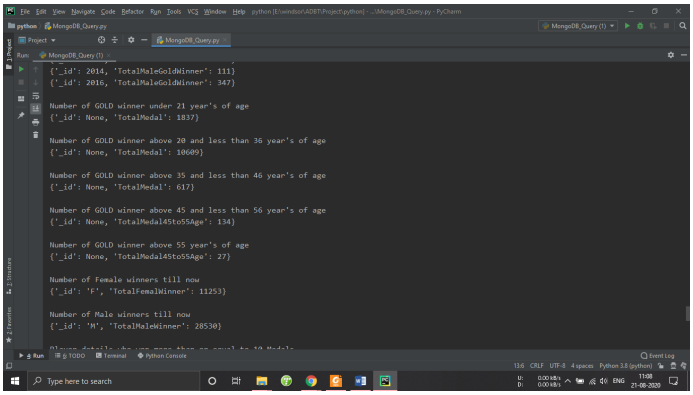

Fig. 3. Total Number of Winners

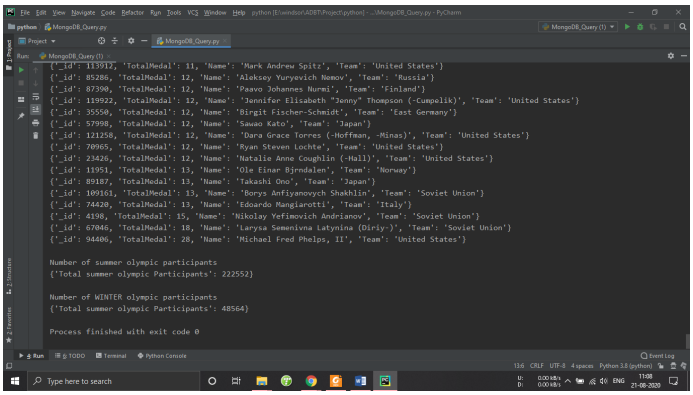

Fig. 4. Total Number of Olympic Participants

B. Data Visualization

Data visualization on the 120 years of data gives us lot of details and anyone can get to understand the trend and variation happen in between the years 1896 to 2016. As you can see that we have performed data visualization for summer and winter both the categories. By creating the different types of graphs and charts we can give the answer of the following questions.

1. How the distribution of medals (Gold, Silver) occurred according to age.

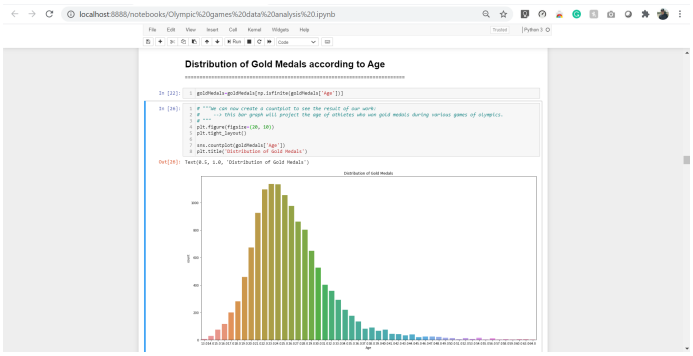

Fig. 5. Number of Gold medal disribution

2. What are the categories of sports in which athletes above the age 50, 60, 70 won the medals?

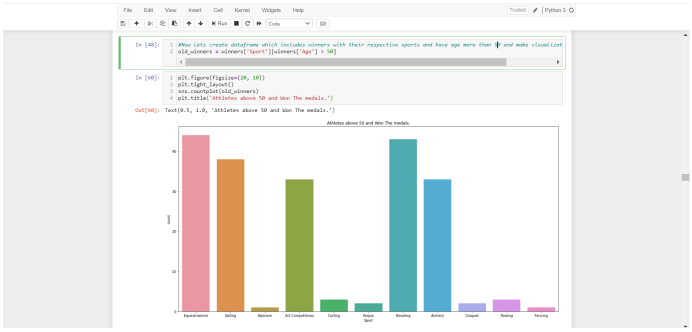

Fig. 6. Categories of sports With winners age more than 50, 60,70.

3. Who are the youngest and oldest medalist in Olympic till now?

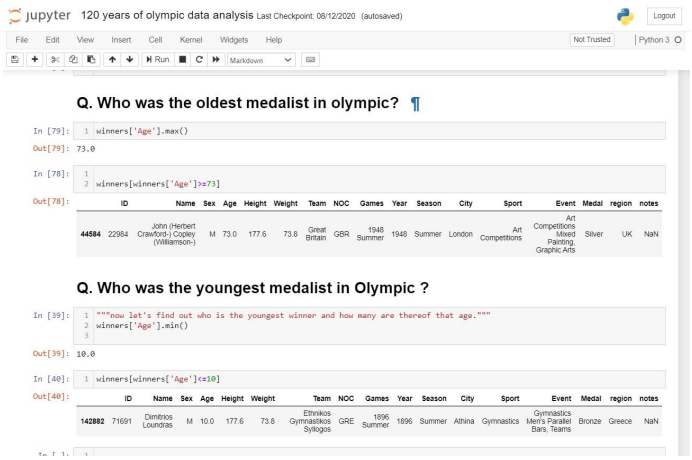

Fig. 7. Youngest and Oldest Winner

4. How the participation of women evolved in every Olympic year?

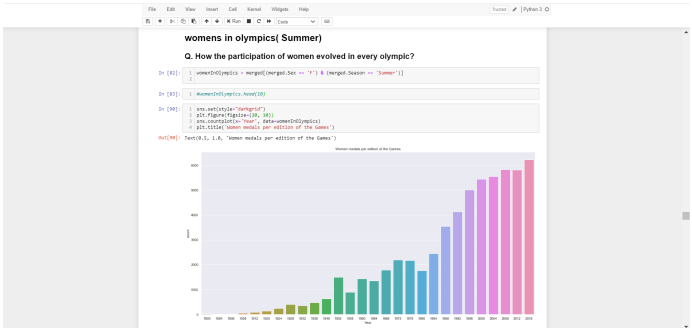

Fig. 8. Women Evolution

5. What are the countries that won highest number of Gold, Silver, Bronze medals?

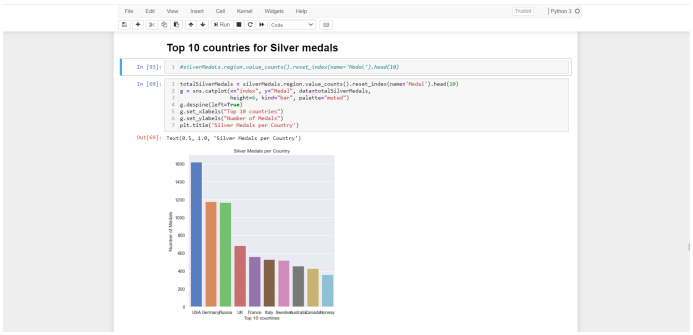

Fig. 9. Highest number of medalist Country

6. How does the athlete participation increased every Olympic year?

Fig. 10. Number of Participants increased.

7. How the gender distribution occurred in both Olympic category?

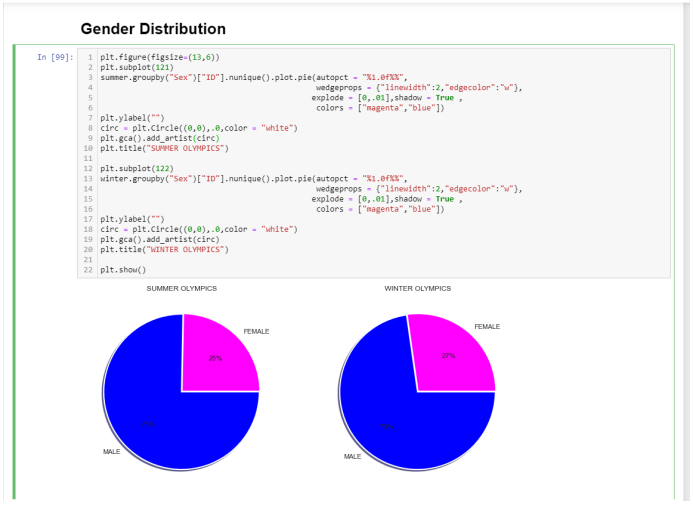

Fig. 11. Gender Distribution.

8. How many countries participated in Olympic according to every year?

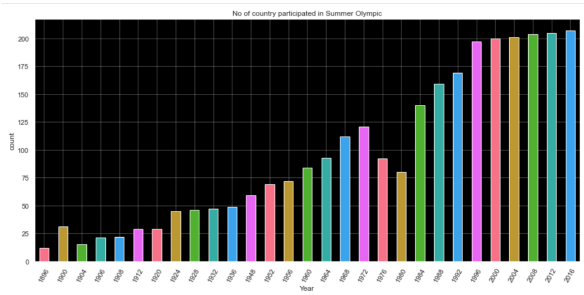

Fig. 12. Number of countries participated over the years

9. Which country hosted the Olympic and in which year?

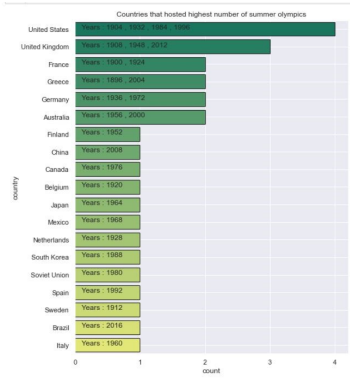

Fig. 13. Host Country.

10. Number of sports increased every Olympic year.

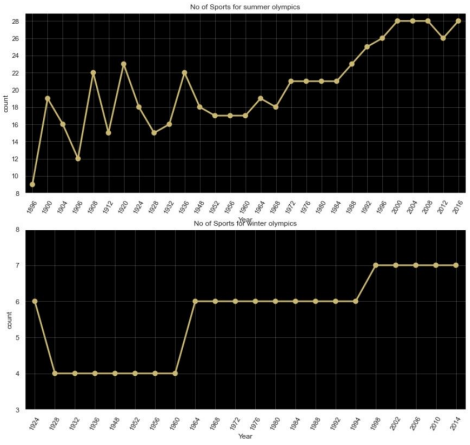

Fig. 14. Number of sports each year

VI. CONCLUSION

We have used 2.8 million data to analyze. After the analysis of these data we have found how the olympic games are evaluated over the 120 years and how the number of participants and games are increased over the years. We have used these data analysis to make visualizations. This data is imported to jupyter notebook and uses different python libraries to make proper visual analysis. These python libraries are used to easily create graphs. This visualization also helps us to make predictions about if the new game is introduced then it will become successful or not.

REFERENCES

[1] Docs.mongodb.com. 2020. Mongodb Compass — Mongodb Compass Stable. [online] Available: https://docs.mongodb.com/compass/master/

[2] Kb.objectrocket.com. 2020. How To Perform Aggregation In Mongodb Compass Community — Objectrocket. [online] Available : https://kb.objectrocket.com/mongo-db/how-to-perform-aggregation-in-mongodb-compass-community-398

[3] Api.mongodb.com. 2020. Tutorial — Pymongo 3.9.0 Documentation. [online] Available: https://api.mongodb.com/python/current/tutorial.html

[4] Tutorialspoint.com. 2020. Mongodb - Query Document - Tutorialspoint. [online] Available: https://www.tutorialspoint.com/mongodb/mongodb$_{q}uery_{d}ocument.htm$

[5] Medium. 2020. An Introduction To Mongodb Query For Beginners. [online] Available: https://blog.exploratory.io/an-introduction-to-mongodb-query-for-beginners-bd463319aa4c

[6] Rpubs.com. 2020. Rpubs - 120 Years Of Olympics Analysis - Report. [online] Available: https://rpubs.com/elifdemir/olympics$_{a}nalysis_{r}eport$