Available online at : http://www.advancedscientificjournal.com
http://www.krishmapublication.com
IJMASRI, Vol. 3, issue 1, pp.846- 852, Dec. -2023
https://doi.org/10.53633/ijmasri

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY ADVANCED SCIENTIFIC RESEARCH AND INNOVATION (IJMASRI)

ISSN: 2582-9130 **IBI IMPACT FACTOR 1.5** **DOI: 10.53633/IJMASRI**

RESEARCH ARTICLE

OLYMPIC DATA ANALYSIS

Yash Pathak[1] and Akash Kumar[2]

[1,2] Department of Information Technology, Maharaja Agrasen Institute of Technology affiliated to Guru Gobind Singh Indraprastha University, Rohini, Delhi
Email: yashpathak6302@gmail.com, Akashkumar200022@gmail.com

Abstract

This research paper's main goal is to analyse a sizable Olympic dataset using exploratory data analysis and assess how the Olympic Games have changed throughout time. Many data analyses on the Olympics have been conducted in the past using a variety of techniques, including artificial neural networks for predicting sports results, a novel sports prediction framework developed using machine learning, exploratory data analysis, and analytical and empirical techniques. Previous studies had a number of drawbacks, including the fact that data is redundant in many databases and they are unable to support the idea that a country's success in the Olympics is strongly correlated with the number of competitors they send. General Terms: Exploratory Data Analysis for Olympic data of years from 1896 to 2016.

Keywords: Metal Tally, Overall Analysis, Athlete Wise Analysis, Country Wise Analysis.

Introduction

The best athletic tournaments in the world are considered to be the Olympic Games. This event, which takes place once every four years, attracts more than 200 countries, which alternate the Summer and Winter Olympics every two years. The International Olympic Committee is the organisation in charge of organising the Olympic Games. An event that draws attention from all across the world and offers athletes the chance to make their country famous. In these competitions, athletes from various nations take part. It's seen as a symbol of fraternity. The main goal of the project is to create a web application that will allow users from various locations to track their nation's Olympic progress.

The research will analyse Olympic data in great detail and use machine learning methods to enhance search results. We shall conduct a thorough review of all nations' Olympic performances in this data analysis. The main goal of the project is to create a web application that will allow people in various locations to track their nation's Olympic progress.

Exploratory Data Analysis, the method we'll use, is well-liked and appropriate for examining how the Olympics have changed over time. In exploratory data analysis, we look at enormous data sets and primarily explain their many aspects visually (Graphs, Charts, and many more).

Literature Survey

Table 1 - Literature Survey table for Olympic Data Analysis.

S.No	Authors	Year	Technique	Subjects/findings
1.	Jiaxin Si [4]	2018	Exploratory Data Analysis is a popular and effective method for examining how the Olympics have changed throughout time. In exploratory data analysis, we look at enormous data sets and primarily explain their many aspects visually (Graphs, Charts, and many more).	Four major analysis sections make up this report. The basics of the Olympics are covered in the first section. We examine the fundamental study of the Summer and Winter Olympics in the second section. The third section is a combined study of the Winter and Summer Olympic Games.
2.	Yamunathanga m.D *et al* [10]	2018	In this paper, the exploratory data analysis techniques are used to provide comparison between performance of various countries and the contribution of each country.	The total number of men and women participants in Olympics from 1896-2012 is analyzed and the ratio between men and women participants can be obtained.
3.	Kabita Paul *et al*. [5]	2019	Analytical and empirical techniques utilized by Human Computer Interaction (HCI) interacts with users via computers.	Top countries in Olympic games were found to be USA, Soviet Union, Germany, Great Britain, France, Italy, Sweden, Australia, Canada, Hungary.
4.	Harshal S. Kudale *et al.* [3]	2019	The 120 years of Olympics and NOC regions data sets are Loaded in the background. The Spark-session library does ETL processes on data and provides the data frame, does Indexing, Joining Processes and provides Data frames (pandas	Apache Spark framework to do the data Analytics work, since it is about 100 times faster than the Apache Hadoop in in-memory data rocessing, and has wide variety of language API support like Java, Python, Scala, Ruby, R etc. whereas

			Compatible).	Hadoop mostly used Java as default language.
5.	Abhishek Gautam [1]	2020	Three datasets were used namely, "120 years of Olympic history: athletes and results [1]", "Gapminder GDP per capita, constant PPP dollars- v25[2]" and "Gapminder Total Population v6[3]"	The advantage of host country in any sporting activity is well known, as the participants will have familiarity of the field, and also there is a great support from the home crowd.
6.	Varagiri Shailaja *et al* [9]	2020	Proposed to use Artificial Neural Networks for sports result prediction. A novel sports prediction framework had been devised using Machine Learning.	Proposed to predict a nation in view of medals owned by 2012. To develop a novel technique, they used a combination of Pearson correlation and coefficient.
7.	Tanya Rieman *et al.* [8]	2020	This analysis was done using MySQL Workbench, Excel and Tableau. In MySQL Workbench an Olympics schema was created and the Table Data Import Wizard used the .csv file	There are a number of layers to this analysis, resulting in multiple queries and tabs which contribute to the ultimate set of data used to answer the question.
8.	Kassidy Chaikin *et al.* [6]	2020	When looking into how gender inequality has changed over the 120 years of Olympic Games. Although there has been a steady increase in the number of female participants, the number of male participants still increases as well which causes there to be a constant gap in gender equality when it comes to the Olympic Games.	They originally tried to also run SVM for each gender and sport we choose, but due to the size of the datasets, SVM was too computationally extensive and their computers barely had enough computing power to run the code.

9.	Rahul Pradhan *et al.* [7]	2020	Data interpretation and Analysis is one of the main and primary tasks in the field of big data analytics. This paper used EDA technique which is quite popular and suitable while analyzing the evolution of the Olympics is Exploratory Data Analysis.	The main objective of this study is to analyze the various factors mentioned above which plays a vital role in the evolution of Olympic Games over the years.
10.	Farkande Vaishnavi *et al.* [2]	2022	The type of analysis which is quite popular and suitable in Olympics is Exploratory Data Analysis.	Performance measures for a country in the Olympics can be predicted using their past performance.

Performance measures for a country in the Olympics can be predicted using their past performance. By predicting their win using the maximum value scored by them in previous participation, the chance of winning gold in Olympics has been identified. If a person wins a medal in an Olympics during a year, the chance of winning a medal in the upcoming Olympics was predicted. Having sports performance data, predicting one's future performance has been done.

Their performance can also be increased if they are not performing well in certain areas, and then placing them accordingly in the training program will provide considerable measure in their outcomes.

(Jiaxin Si 2018) This report has four main analysis parts. The first part introduces the base information about Olympics. In the second part, we explore the basic analysis about Summer and Winter Olympics. The third part consists of joint analysis of both the Summer and Winter the Olympic Games. The fourth part, involving the number of medals in the Summer Olympics and the Winter Olympics, will be explored.

(Yamunathangam *et al* 2018) used the exploratory data analysis techniques to provide comparison between performance of various countries and the contribution of each country.

(Harshal Kudale *et al.* 2019) Apache Spark framework to do the data Analytics work, since it is about 100 times faster than the Apache Hadoop in in- memory data processing, and has wide variety of language API support like Java, Python, Scala, Ruby, R etc. whereas Hadoop mostly used Java as default language.

(Kabita Paul *et al.* 2019) used Analytical and empirical techniques utilized by Human Computer Interaction (HCI) interacts. Age does have an impact on sports overall when it comes to medal winners, but not for individual athletes that compete in multiple Olympic Games.

(Riemann Nicol 2020) used MySQL Workbench, Excel, and Tableau, In MySQL Workbench an Olympics schema was created and the Table Data Import Wizard used the .csv file.

(Rahul Pradhan *et al.* 2020 used EDA technique which is quite popular and suitable while analyzing the evolution of the Olympics is Exploratory Data Analysis.

(Abhishek Gautam 2020) Three datasets were used namely, "120 years of Olympic history: athletes and results", "Gapminder GDP per capita, constant PPP dollars- v2" and "Gapminder Total Population v6". There is redundant data int the dataset which is creating ambiguity.

(Varagiri Shailaja *et al.*, 2020) Proposed to use Artificial Neural Networks for sports result

prediction. A novel sports prediction framework had been devised using Machine Learning. Proposed to predict a nation in view of medals owned by 2012. To develop a novel technique, they used a combination of Pearson correlation coefficient, Spearman correlation, coefficient and linear regression.

(Kassidy Chaikin *et al.* 2020) They originally tried to also run SVM for each gender and sport we choose, but due to the size of the datasets, SVM was too computationally extensive and their computers barely had enough computing power to run the code.

(Farkande Vaishnavi *et al.* 2022) used Exploratory Data Analysis, but the main drawback is that EDS is not effective when we deal with high-dimensional data.

Methodology

An Approach is referred as a systematic path to reach a solution. Every problem, whether technical or non-technical, requires a proper approach so that we can get a proper path on which we have to proceed to get the required result. This Research Paper aims to analyze the vast history of Olympic Games and determine the evolution of Olympic Games over the Time. There are various factors which contributes in the evolution of the Olympics. To determine these factors and perform a comparative study on these factors, we need to follow an Approach which will take us to our destination. Given below is the flowchart (See Figure. 2) of the proposed approach for the problem. We will discuss each step thoroughly.

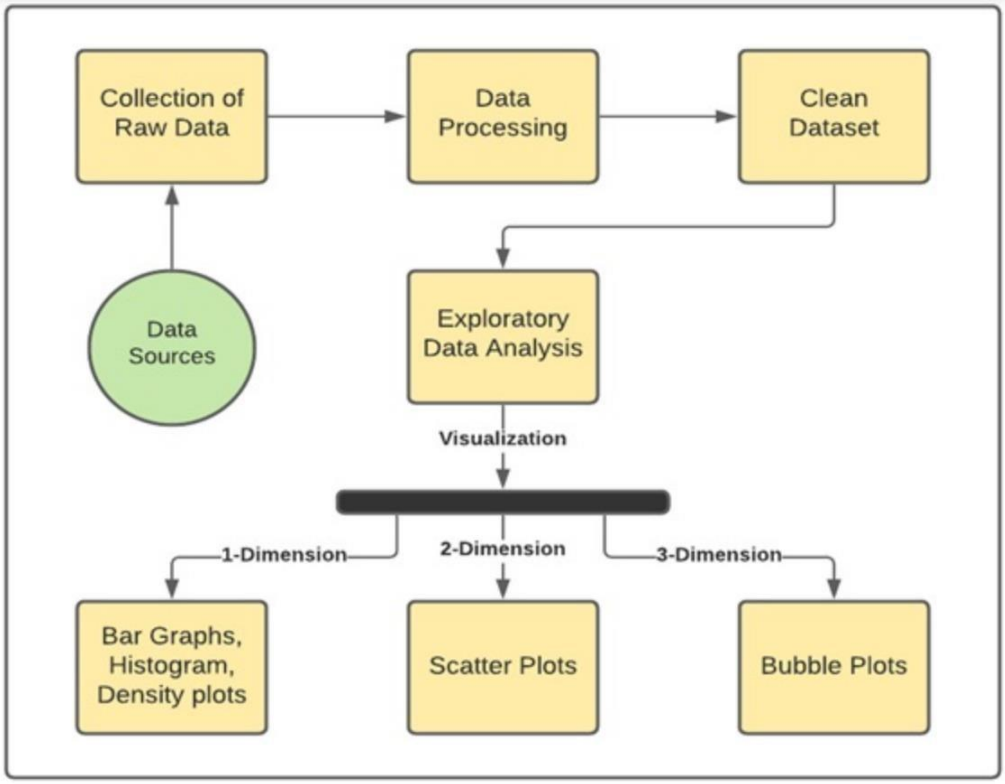

Fig. 1. This Figure depicts the steps of the Analysis. These are the steps in which the whole process has been executed

Data Collection

Data collection is the very first stage in any form of analysis, whether it be technical or not. We need a lot of data to undertake analysis on a given problem, and then we use a variety of approaches and algorithms to get our desired conclusions from the data. It is advisable to collect a lot of data since the more data that is analysed, the more accurate the results will be and the more confident you may be in the decisions you make based on those results. We have utilised information from a variety of data sources to analyse how the Olympics have changed over time. Three datasets that give us a lot of data in a wide range of types have been chosen by us for analysis. The first dataset contains facts about the players, including their gender, height, weight, country of origin, medals earned (gold, silver, and bronze), and many other things. This information can be used to evaluate a specific player's performance and to compare the performances of two or more players. The second dataset contains statistics on the nations that have so far competed in the Olympics, together with a list of the total amount of medals (Gold, Silver, and Bronze) they have earned. This information can be used to compare the performance of the various nations.The third dataset consists of a list of nations and their country codes, which serve as the nations' identifiers. The total number of nations that have competed in the Olympics to date can be determined using this data. These datasets give us access to a large number of precise data points in a variety of forms. Data processing comes next after data collection. Raw data is information that has been taken straight from a dataset or other data source. Different methods or machine learning algorithms, such as linear regression, decision trees, SVM, etc., cannot be applied directly to raw data.This data must be processed and transformed into usable information. By diligently screening for flaws and removing unnecessary, incomplete, or erroneous data, data pre-processing transforms raw data into useful data. The dataset includes a number of fields, including Age, Gender, and others, some of which include null values, which leads to problems in the final product, which is the graphical representation of the data.

The majority of a data analyst's time is spent preparing and cleaning the data. This often involves the following actions:

- Use Pandas to load the dataset into a data frame.
- Investigate the amount of rows and columns, value ranges, etc.
- Format data (convert text to numbers, remove extra formatting like "," from numbers, add a currency sign, format dates, etc.)
- Handle incomplete, inaccurate, and erroneous data (either remove it or fill it with appropriate data)
- Execute any extra actions (parsing dates, creating additional columns, merging multiple dataset etc.)

Exploratory Data Analysis

Data analysis comes after data pre-processing. In order to arrive at a specific conclusion, analysis of the data is done in this step using a variety of techniques, including text analysis, diagnostic analysis, exploratory data analysis, etc. and machine learning algorithms, such as linear regression, logistic regression, SVM, decision tree, etc. We are employing the exploratory data analysis technique to finish this work because our area of research is the visualisation and comparative analysis of many aspects that contribute to the Evolution of Olympic events over time. Exploratory Data Analysis (EDA) is a method for thoroughly analysing data and essentially visualising its key characteristics. [8] In addition to using different techniques, exploratory data analysis is largely utilised to understand what the data represents. By using various sorts of graphs and plots that may be created with EDA, we are able to grasp the structure and substance of the dataset. Plots come in many different varieties and are utilised in EDA.

The following list includes a few of them:

- Histograms
- Bar graphs
- Box plots
- Scatter plots

` We can view the data visually, explain the analysis based on that, and conduct a comparison study between several plots.

Conclusion

The main objective of this study was to analyse and visualise the various factors which have contributed to the Evolution of the Olympic Games over the years. This type of analysis is very helpful as this type of analysis can be performed by any Country or Player which can help them in analysing their performance so that they can improve their performance by changing their strategies.

We have used a technique named Exploratory Data Analysis which enables you to encapsulate the primary factors of a dataset into a visual format. We selected Python language to implement our work because It is one of the best languages suitable for Data Analysis and is the platform where we have performed this Analysis.

As a result of the Analysis, we can conclude that It is true that Olympic Games have evolved considerably over time from the 1896 Olympic Games till the 2016 Rio Olympics. Various factors provide valid evidence that the Olympics have changed a lot. some of these factors are the launch of the Winter Olympic Games apart from the Summer Olympic Games in 1924, an increase in the number of participating countries in both Summer and Winter Olympics, the Average age of players in the Olympic Games, the increase in the participation of the females in both Summer and Winter Olympics over the time, Total number of medals won by various participating countries over the years, Average height and the weight of Players who contributes to victory of Games in the event. Apart from these, there are many more factors that depict the Evolution of the Olympic Games over time. Visualisation of these factors has been done to explain and validate the Analysis in various Graphical formats like a Line graph, Scatter Plots, Bar, Graphs, Dist Plots, etc

Refrence

1. Abhisek Gautam (2020). Visual Analysis of Olympics Data. Jul 21, 2020.
2. Farkande Vaishnavi., Gurav Vaishnavi and Borse Tejas. (2022). Web Application of Olympic data analysis.July 2022.
3. Harshal,Kudale., Mihir, V. Phadnis. Pooja, J. Chittar, Kalpesh, P. Zarkar. Balaji, K. Bodhke. (2022). Data Analysis and Visualization Of Olympics Using Pyspark And Dash-Plotly. June-2022.
4. Jiaxin, S (2019). Olympic Games Medal Count Analysis Summer and Winter Olympic Games 2018.
5. Kabita Paul., Elif Demir. Anjali Bapat. (2019). Olympic Data Analysis - Final Project Report. May 2019.
6. Kassidy Chaikin. Christopher Egan. Emily Lepore. William Wenze. (2020). Analysis of 120 Years of Olympic Results. July 2020.
7. Rahul Pradhan., Kartik Agrawal1 and Anubhav Nag. (2021). Analyzing Evolution of the Olympics by Exploratory Data Analysis using R. 2021.
8. Tanya Riemann. Alyson Nicol. (2020). Olympic Data Analysis Project. December 2, 2020.
9. Varagiri Shailaja. (2020). Predictive Analytics of Performance of India in the Olympics using Machine Learning Algorithms. May 2020.
10. Yamunathangam, D., Kirthicka,. G. Shahanas Parveen. (2028). Performance Analysis in Olympic Games using Exploratory Data Analysis Techniques November 2018.

*****