iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** XII **Month of publication:** December 2022

DOI: https://doi.org/10.22214/ijraset.2022.48046

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue XII Dec 2022- Available at www.ijraset.com

Olympic Data Analysis using Data Science

Nishant Kulkarni[1], Pratik Patil[2], Rugved Pande[3], Dhiraj Patil[4], Pranav Nair[5], Parth Prabhu[6], Pratyush Doshi[7], Pranav Bhosale[8]
Vishwakarma Institute of Technology, Pune

Abstract: ***The Olympic games are international sports events with more than 200 nations participating in various competitions. The Sportspersons from various countries participate in competitions and make their countries proud of their excellence in sports. The primary objective of this paper is to analyze the Olympic dataset using python to compare overall performance of countries and to evaluate the contribution of each country in the Olympics. These analyses will give deeper insight into the performance of countries in Olympics over the years and helps sportspersons to quickly analyze their own and the competitor's performance. In this paper, the exploratory data analysis techniques are used to provide comparison between performance of various countries and the contribution of each country in the Olympics. Visualization of Olympics dataset in many aspects provides the status of countries in Olympics and helps countries with poor performance to produce quality players and improve nation's performance in Olympics. Despite a lot of hard work, many countries or players are unable to perform well during the events and grab medals whereas there are many countries that perform very well in the event and secure many medals. An analysis needs to be done by each country to evaluate the previous statistics which will detect the mistakes which they have done previously and will also help them in future development. Visualization of the data over various factors will provide us with the statistical view of the various factors which lead to the evolution of the Olympic Games and Improvement in the performance of various Countries/Players over time. The primary objective of this Research paper is to analyze the large Olympic dataset using Exploratory Data Analysis to evaluate the evolution of the Olympic Games over the years.***
Keywords*: Olympic, Sports, Nations, Python, Dataset.*

I. INTRODUCTION

Olympics is considered as the most important event worldwide, which provides a common platform to players from various nations to show their talents. The Olympics started in 1896, which is conducted once every four years. The goal of this paper is to analyze performance and participation of nations in Olympics from 1896 to 2016.In addition, the field of sports of particular country in particular year, in which they have contributed the maximum can be identified. The comparison of the performance of each sport with another can be done. The field of sports that has to have more participation can be identified and necessary action can be taken by players and nations to enhance themselves in future contributions towards the Olympics. The modern Olympic Games or Olympics are leading international sporting events featuring summer and winter sports competitions in which thousands of athletes from around the world participate in a variety of competitions. The Olympic Games are considered the world's foremost sports competition with more than 200 nations participating. The Olympic Games are normally held every four years, alternating between the Summer and Winter Olympics every two years in the four years. Various scenarios come to our mind when we look into the Evolution of the Olympic Games over the years. These scenarios are: Increase in the number of participating nations, Increase in the number of participating Athletes, Increase/Decrease in the number of events, Increase in the expenditure cost of the event, improvement in the performance of the particular country, improvement in the performance of a particular player, Increase in women participation, Participation Ratio of Men to Women, improvement in medication facilities during competition, the effect of pandemic (if any) on the performance of the players. Analysis of these scenarios would depict the evolution of the Olympics over the years. This analysis would help in future prediction.

II. LITERATURE REVIEW

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 2016 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 is 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.

Machine learning techniques were used for heuristics prediction of Olympic medals of a country. Estimation of the success of a country can be done by efficiency analyses and importance of sports in society. When analyzing the sports categories are mainly being more representative towards viewpoint-based content rather than being a viewpoint that is spatiotemporal. The video content of the analysis has the significance of providing more interior information than structured collected data. In addition to these techniques, the exploratory data analysis uses visual methods to provide a deep understanding and statistical summary of the data.
Data interpretation and analysis is one of the primary tasks in the field of big data analytics. There has been a lot of analysis on the Olympic Games like statistics visualization, performance analysis of players, improvement in the performance of various countries, and many more. The type of analysis which is quite popular and suitable while analyzing the evolution of the Olympics is Exploratory Data Analysis. In Exploratory Data Analysis, we examine large data and elucidate its various characteristics basically in the visual format (Graphs, Charts, and many more). EDA is an approach that provides a deeper understanding of the dataset. There has been a research paper that analyses the outbreak of the Novel Coronavirus. The exploratory Data Analysis technique is used to analyze the data and find out the number of cases reported (positive, dead, discharged) inside China and Outside China. This paper took data from different datasets and apply the EDA technique to analyze various factors like the number of cases recovered during January and February inside and outside China, the number of cases confirmed in the different provinces of China, and outside China till 16 February 2020.The main aim of this analysis was to find out the growth in the performance of a country in the Olympics over the years. With the Help of such an Analysis, any player can check their progress record and can also have a look at their opponent's progress.

III.METHODOLOGY

An Approach is referred to 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 contribute to the evolution of the Olympics. To develop Olympics data analysis, we have followed methodology of:

1) *Data Collection:* The very first step of any type of Analysis, whether it is technical or non-technical, is Data Collection. In order to perform analysis on a certain problem, we require a large amount of Data on which we apply various techniques and algorithms to reach a particular conclusion and get our desired result. It is advised to take the data in abundance because larger the volume of data for analysis, the greater would be the accuracy in the result and also the greater would be the confidence in decision making based on these results. We have used data from various data sources for analysis on Evolution of the Olympics over the time. We have taken three datasets which provide us with a large volume and a large variety of data for Analysis. The 1st dataset consists of information about the players and their entire details like their Gender, Height, Weight, Country for which they play, Medals won (Gold, Silver and Bronze) and many more. This data can be used to analyze the performance of the particular player and can also help in the comparative study between two or more players. the 2nd dataset consists of the information of the countries which have participated in the Olympics so far and the list of the total number of medals (Gold, Silver and Bronze) won by them. This data can be used to perform a comparative study on the performance of the countries. the 3rd dataset consists of the list of countries along with their country code which is the identification of these countries. This data can be used to find out the total number of countries which have participated in the Olympics so far.
2) *Data Pre-Processing:* The next step after collecting Data is Data Processing. Data directly obtained from a data source such as dataset is known as Raw data. We can't apply various techniques or Machine Learning Algorithms like Linear Regression, Decision Tree, SVM etc directly to the Raw Data. This Data need to be processed and converted into useful data. Data Preprocessing is the process of translating the Raw data into Useful data by conscientiously checking for errors and eliminating redundant, incomplete, or incorrect data. The Dataset consists of various fields like Age, Gender, etc which consists of some null values which produces errors in the end result which is the Visualization of data in graphical format. These null values are needed to be omitted or replaced with some valid value which solves the error and generates accurate result. We have used a technique known as Deterministic Imputation to complete this task. Deterministic Imputation is a situation where the null values (NA or NaN) are determined with the help of the other values in the same column in the dataset. For this purpose, there are various models such as Basic Numeric Imputation Model in which the null value is replaced by Mean or Median of other values of the same column of the dataset. There is another model known as Hot Deck Imputation in which the null value is replaced by similar record in the dataset, i.e., some other value in the same column. Hot Deck Imputation can be applied to both Numerical as well as the Categorical value, but only if it contains enough values in the same column.

3) *Exploratory Data Analysis:* The next step after data pre-processing is data analysis. In this step, analysis is done on data using various Techniques like Text Analysis, Diagnostic Analysis, Exploratory Data Analysis, etc and Machine learning Algorithms like Linear Regression, Logistic Regression, SVM, Decision Tree etc to reach to a particular conclusion. As our field of Research is visualization and comparative study of various factors which leads to the Evolution of Olympic games over the time, we are using the Exploratory Data Analysis technique to complete this task. Exploratory Data Analysis (EDA) is an approach to analyze data thoroughly and encapsulate its primary attributes basically in visual format. Exploratory Data Analysis is mainly used to see what the data represents apart from applying various algorithms. With the help of EDA, we can understand the structure and content of the dataset by various types of graphs and plots which can be drawn with the help of EDA. We can View the data in the visual format and can explain the analysis on that basis and perform a Comparative Study between different plots. There are various types of plots which are used in EDA. Some of them are mentioned below:

a) Histogram
b) Bar Graph
c) Box Plot
d) Scatter Plot and many more.

IV. EXPERIMENTAL SETUP AND RESULT ANALYSIS

Experimental Setup Analysis on any type of data cannot be done without the help of Programming Language and a platform on which we can perform the Analysis with the help of Programming language. A Programming Language is a conventional language which consists of the various set of instructions by which one can produce a specific output by taking data from the system or ASCI- 2020 IOP Conf. Series: Materials Science and Engineering 1099 (2021) 012058 IOP Publishing doi:10.1088/1757-899X/1099/1/012058 6 by providing custom input. There are various programming languages which are used for the purpose of Data Analysis. Some of the widely used programming languages for Data Analysis are - Python, JavaScript, Scala, R, SQL, Julia and many more. With the help of any of these Programming, one can perform Analysis over data by applying various techniques. For our Project we have chosen R as the programming language and RStudio as the platform where we have analyzed the data using R language. R is a programming language used for Analytical Computation and Graphical Representations. It consists more than 10000 inbuilt packages.R provides various analytical techniques like linear and nonlinear modeling, Time Series Analysis, Clustering, Classification, etc and various Graphical Techniques, which makes it an ideal language for our task which is visualization of various factors which contributes in the evolution of Olympic Games over the years. Launched in 2011. RStudio is an open source IDE for R programming language written in C++, Java, JavaScript. It includes various R packages like Tidyverse, TensorFlow, Reticulate and many more. For our purpose, we have used various packages of R Language and imported them in RStudio to work with them. Some of them are : • Tidyverse - Tidyverse is a dictatorial collection of various R packages which are used inthe field of Data Analysis. The Packages included in Tidyverse are ggplot2, dplyr, tidyr, readr, purrr, tibble, stringr, and forcats.

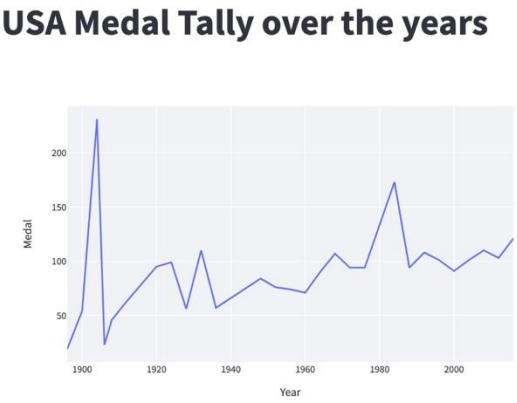
1) *ggplot2* - Released in 2005, ggplot2 is a package used to visualize almost all type of data in any format into graphs. It is one of the most popular packages of R Language.
2) *readr* - readr package is used to read Rectangular Data like csv (Comma Separated Values), tsv (Tab Separated Values), and fwf files.
3) *dplyr* - dplyr package is used to make Data Manipulation easier. It provides various methods which helps to manipulate data. Some functions in the dplyr package are mutate(), select(), filter(), summarize(), arrange() and many more.
4) *readxl* - readxl package is used to import data from the excel file. Data in xlsx, xls can be imported using readxl.

V. ANALYSIS AND VISUALIZATION

1) Identifying Contribution of Men And Women Participants In Olympics (1896-2016)

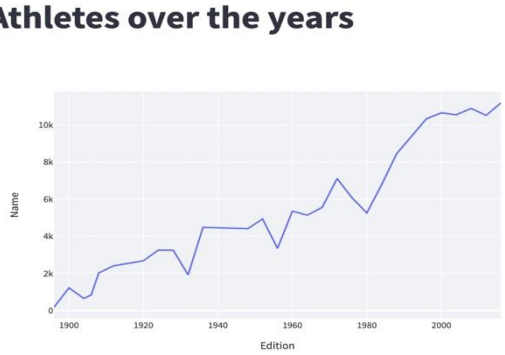
The total number of men and women participants in the Olympics from 1896-2016 is analyzed and the ratio between men and women participants can be obtained. The analysis shows that the contribution of men is higher than women all over the world. The figure 1 shows gender wise contribution of players in the Olympics.

2) Identifying total number of medals achieved by USA Country in Olympics(1900-2000)



In this analysis, the total number of gold, silver and bronze medals won by the participants from all countries in the Olympics from 1896 to 2016 can be identified. The count includes the number of individuals who contributed separately or as a team to receive medals for their nations. The following results are obtained in the analysis. (i) USA has won the highest number of gold medals when compared to other medals and almost equal percent of silver and bronze medals. (ii)Australia has received the least number of gold medals when compared to other medals and won the highest number of bronze medals. Japan has fewer gold medals than other medals. France has less gold and a high number of silver and bronze medals.

3) Identifying the performance of Particular Country in Olympics (1992-2016)



Excellence of a country in the Olympics can be viewed by the number of medals won by a country. This analysis identifies the performance of a particular country in Olympics from 1992 to 2016.This can be processed by calculating the total medals won by a particular country in a particular year from 1992 to 2016. Data visualization can be carried out to represent the result of a particular country. The results are (i)Performance of India was gradually increasing from 1992 with no medals,1996 with 1 medal and finally in 2016 with 6 medals.(ii)Performance of USA was found like zig-zag graph from 1992 with 220 medals,1996 with 260 medals, suddenly performance has decreased in 2000 with 240 medals, increased gradually from 2004,contributed best in 2008 with 350 medals.(iii)France‟s Performance was gradually increasing from 1996 to 2008 with medals within range of 40 and has performed well in 2016 with 80 medals.(iv)Performance of Australia was better during 1992 Olympics with 60 medals and there was a sudden increase in its performance with almost 200 medals over the period of 2000 and there has been gradual decrease in performance from 2004 to 2016.(v) Initially, performance of Japan was not so good ,but over the period of 2000 and 2004 there was a drastic increase in it and gained 100 medals which was higher than the rest.

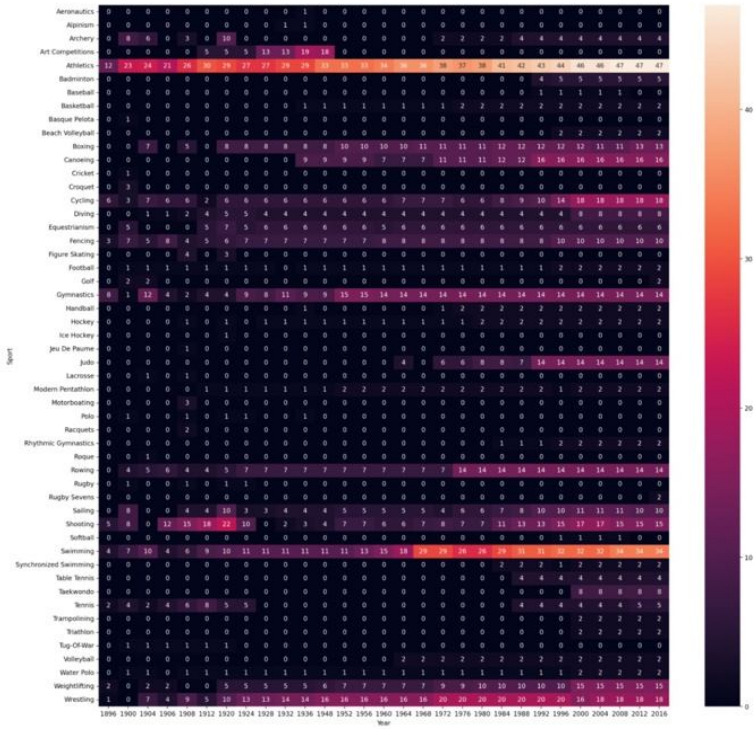
4) Comparing the performance between the countries in Olympics(1896-2016)

Overall Tally

	region	Gold	Silver	Bronze	total
0	USA	1035	802	708	2545
1	Russia	592	498	487	1577
2	Germany	444	457	491	1392
3	UK	278	317	300	895
4	France	234	256	287	777
5	China	228	163	154	545
6	Italy	219	191	198	608
7	Hungary	178	154	172	504
8	Sweden	150	175	188	513
9	Australia	150	171	197	518
10	Japan	142	134	161	437

The analysis compares the performance between the countries by medals won by the participants from selected countries in Olympics from 1896 to 2016.Countries such as USA, Hungary, France, Japan, Australia are selected for analysis .From this analysis, the following results has been inferred.(i)In the 1996 Olympics, among the five selected countries, USA is the leading country with a contribution of 7.53%,Australia is the second country with 2.25%,Japan with 1.61%,Hungary with 0.75% and France is the least country with 0.69% contribution among them.(ii)In 2000 Olympics, USA is the leading country with contribution of 6.55%,Australia is the second country with 4.019%,France with 1.58% and Hungary & Japan are the least country with 0.94% among them.(iii)In 2004 Olympics, USA is the leading country with contribution of 8.05%,Australia is the second country with 3.3%,Japan with 2.1%,France with 1.6% and Hungary is the least country with 0.9% among them.(iv)In 2008 Olympics, USA is the leading country with contribution of 8.52%,Australia is the second country with 3.72%,France with 1.71%,Japan with 1.42% and Hungary is the least country with 0.88% among them.(v)In 2016 Olympics, USA is the leading country with contribution of 8.5%,Australia is the second country with 3.01%,Japan with 1.8%,,France with 1.6% and Hungary is the least country with 0.9% among them.

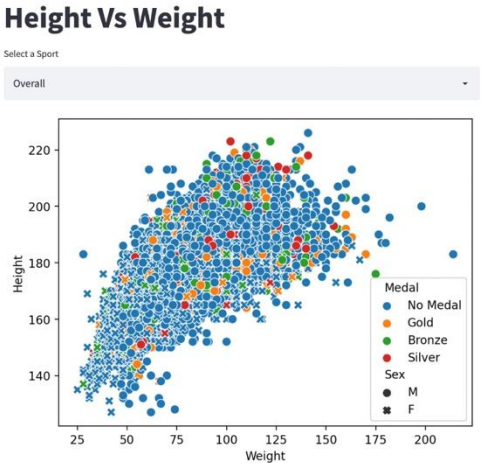
5) Identifying the Best Performed Field of Sports for Particular Country in Olympics (2000-2016)



The analysis represents the performance from participants of a particular country and their best performing field of sport in Olympics from 2000 to 2016.To identify the field of sports of a particular country in a particular year and to analyze which field of sport has to have more participation. This provides information to enhance themselves in future contributions towards Olympics.

a) In 2000, USA has performed best in the field of Aquatics and has performed least in the field of Weightlifting.
b) In 2000, Australia has performed best in the field of Aquatics and has performed least in the field of Gymnastics.
c) In 2000, France has performed best in the field of Fencing and has performed least in the field of Tennis.
d) In 2000, Australia has performed best in the field of Aquatics and has performed least in the field of Athletics.

6) Analyzing the height vs weight



From this analysis, the Most Females who have won the medal are between 160-180cm tall and their weight class wide ranging from 50-150kg. The Number of Gold winners has performed well irrespective of their weight but seen density high at range of 175-180cm height.

VI. RESULT AND DISCUSSION

This work highlights the broad range of diagnostic and therapeutic services available to athletes during the London 2016 Olympic Games. Peak usage of many of the facilities was seen around days 9 and 10 of the competition (5 and 6 August 2016). This is when there was the greatest number of event finals occurring11 and the athletes' village was at its busiest. As expected, most consultations were musculoskeletal in origin but a sizable proportion also related to dental and ophthalmic complaints. The demand for MRI was significant, reflecting the fact that this resource is considered not as freely available as others. wise as it is during Games time. Pathology investigations were performed steadily throughout the period of competition, but the demand for pharmacy services did spike considerably. It is interesting to note from the continent subanalysis that the greatest proportion of attendances was from athletes from African nations. This was for the gross number of overall attendances and also when corrected for multiple attendances by individual athletes. It is also interesting that although Oceania provided the smallest proportion of overall attendances (7%), this constituted the second largest fraction of visits by individual team members (30%). This reflects the fact that Oceania fielded the smallest number of athletes (670); therefore, individual attendances would constitute a greater proportion of the small Oceania cohort. Athletes were able to self-present to the Polyclinic and would often be accompanied by their NOC's medical or administrative staff. On arriving at the Polyclinic, they were quickly triaged to the appropriate department and rarely had a significant delay in being seen. Staffing levels appeared to meet the demands effectively; however, minimal waiting time was seen for some of the busier services such as physiotherapy, sports massage and radiology. Despite being serviced entirely by volunteers, staff had undergone a comprehensive recruitment and selection process involving an induction and orientation to the building and working environment prior to the start of the Games. This enabled an efficient working environment right from the start of the Games and limited any start-up issues. Daily work force meetings at the start and end of each shift further reinforced good communication and working relations among staff from different departments in the Polyclinic. Efficient assimilation and storage of medical encounter data were crucial throughout the Games. Workstations connected to the Games network were available in all medical venues including all fields of play to allow timely data input. This meant that records were kept contemporaneously and could be referred to during successive visits for the same individual. The Atos database provided an effective platform for these data to be securely stored and contained relevant data fields to be comprehensive and appropriate.

VII. FUTURE SCOPE

We all know that any Analysis is not perfect and it consists of some limitations which define the Future scope of the Research Work. This project work also contains some limitations which we are considering as the Future Scope of the Project. We can also describe the data in other formats like Geographical format where we can depict the countries on the World map. We can also apply various Machine Learning Algorithms to the data set after Analysis and can create a Predictive Model which can predict the statistics of the Future Olympic Games.

VIII. CONCLUSION

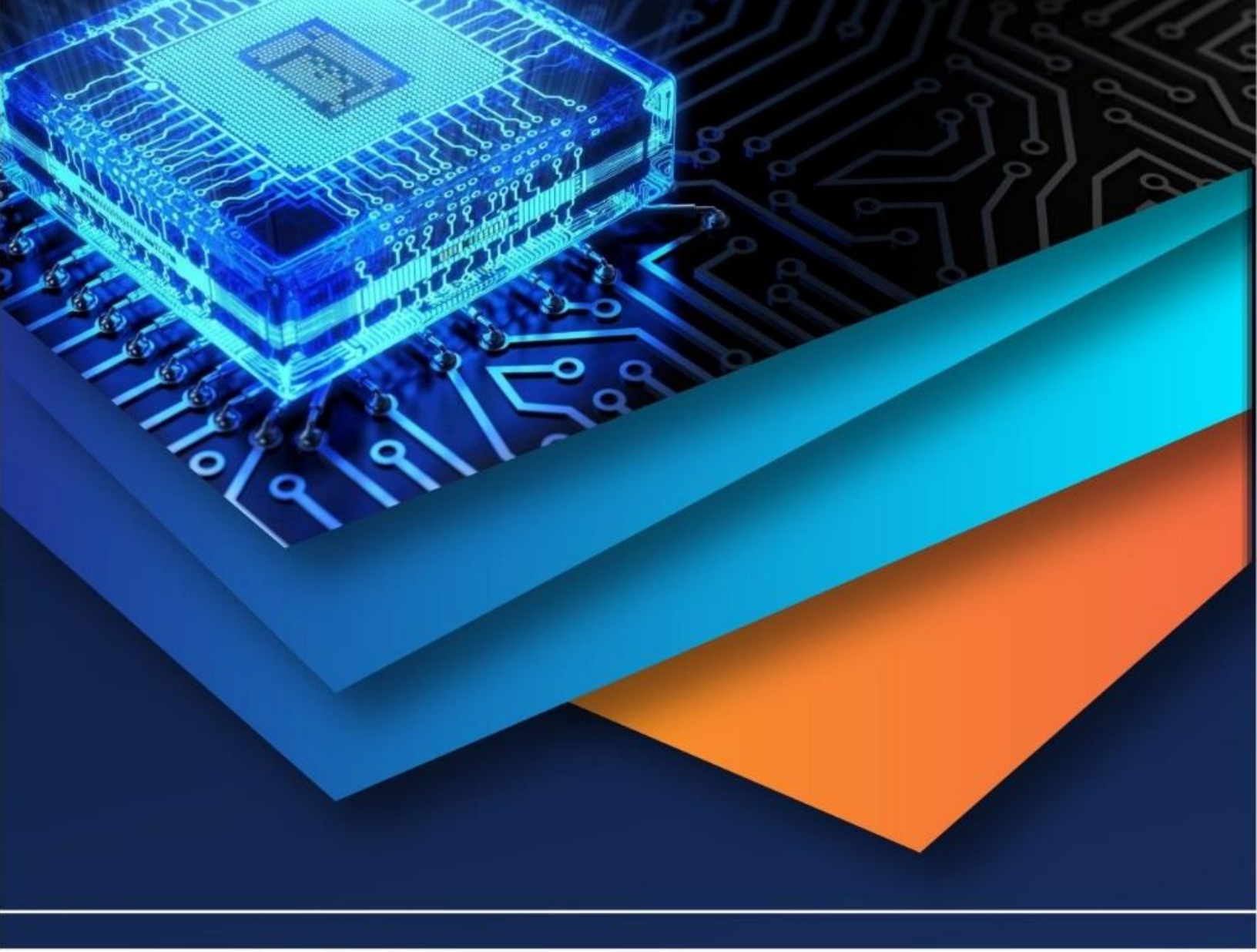
The main objective of this study was to analyze and visualize the various factors which have contributed to the Evolution of the Olympic Games over the years. This type of analysis is very helpful as this type of analysis can be performed by any Country or Player which can help them in analyzing their 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. Visualization 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.

IX. ACKNOWLEDGMENT

The group is feeling obliged in taking the opportunity to express our gratitude most sincerely to our guide Prof. Nishant Kulkarni who guided and motivated us in this course of time of understanding the concepts. We are grateful for the insightful comments offered by the anonymous peer reviewers. The generosity and expertise of one and all have improved this study in innumerable ways and saved us from many errors.

REFERENCES

[1] Wikipedia contributors: https://en.m.wikipedia.org/wiki/Olympic Games, last accessed 2020/11/02.
[2] Dey S K, Rahman M M, Siddiqi U R and Howlader A 2020 Analyzing the epidemiological outbreak of COVID-19: A visual exploratory data analysis approach J. Med. Virol. 92 632–8
[3] Bondu R, Cloutier V, Rosa E and Roy M 2020 An exploratory data analysis approach for assessing the sources and distribution of naturally occurring contaminants (F, Ba, Mn, As) in groundwater from southern Quebec (Canada) Appl. Geochem. 114 104500
[4] Cutait, M.: Management performance of the Rio 2016 Summer Olympic Games. Research Paper submitted and approved to obtain the Master's degree in Sports Administration at AISTS in Lausanne, Switzerland.
[5] Moreno A, Moragas M and Paningua R 1999 The evolution of volunteers at the Olympic Games Proceedings of Symposium on Volunteers (Lausanne, Switzerland: Global Society and the Olympic Movement) pp 1–18
[6] Abeza G, Braunstein-Minkove J R, S´eguin B, O'Reilly N, Kim A and Abdourazakou Y 2020 Ambush marketing via social media: The case of the three most recent Olympic Games Int. J. Sport Communication 1–25
[7] Yamunathangam D, Kirthicka G and Shahanas P 2018 Performance Analysis in Olympic Games using Exploratory Data Analysis Techniques International Journal of Recent Technology and Engineering (IJRTE) 7 251–3
[8] Wikipedia contributors: Exploratory data analysis, https://en.wikipedia.org/wiki/Exploratory data analysis, last accessed 2020/11/11.
[9] Ramachandran K. M. and Tsokos C P 2020 Mathematical statistics with applications in R (Academic Press)
[10] Lange D Summer Olympics: number of participating countries 1896-2016 Statista.com



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY