Exploratory And Predictive Analysis of Olympic Data Using Web Scraping

Sushila Palwe [1], *Nikhil Mundey*[1], *Prasad Khandke*[1], *Masoom Raza*[1]

[1]*School of Computer Engineering and Technology, MITWPU, Pune, India.*

Corresponding Author: sushila.palwe@mitwpu.edu.in

Abstract: - The Olympics games are the most popular and crucial sports event that happens every 4 years. The Olympic games is the world's most important sports competition in which around 200 nations participate. Thousands of athletes participate in 35 different sports and over 400 events. It is crucial to analyze the data related to Olympic games as it takes in account the data related to thousands of athletes belonging to around 200 countries. This paper represents the extraction of this data through web scraping by targeting several websites and performing an analysis study on these collected data. This analysis represents the role of economic factors like country's GDP, GDP per capita, hosting country, population versus the medal count that each country is obtaining, so that the obtained analysis measures can be used to enhance the performance of the athletes belonging to a particular country.

Key Words*: — Web Scraping, Data Visualization, Data Analysis, Linear Regression.*

I. INTRODUCTION

The Olympic Games are a world-famous sporting event, and have flourished since 1896. The history of the Olympic games can be found thousands of years ago where it was started by the Greek, which consisted only of normal running races, and was used to be hosted in Greece's city Olympia and only the people of Greece can participate in it. It has developed very much from then and now has become a center for a large number of athletes from all over the world to showcase their abilities and skills in more than 30 solo sports events. Currently the Olympic games happen every four-year alternating between Winter and Summer Olympic. Both of them have their respective set of games. The Olympic games has become a way where the thousands of contestants can showcase their hard trained skills, and become a symbol of pride by representing their country.

The data related to these Olympic games are a bit difficult to get and are mostly present in text form in several websites. These data are embedded in the HTML tags that represent the webpages.

The extraction of this kind of data is done by the technique called web scraping. This automates the manual tasks of data extraction with much efficiency and with less chance of error. Web scraping enables us to convert the unstructured data in form of simple text and non-tabular data to structured form represented in csv format. Web scraping is very much similar to web indexing, that is used by most of the search engines like google, yahoo etc. When compared, web indexing parses the entire data of the web page in order to make it searchable while web scraping looks for the targeted information available on the web page as per the requirement. For example, several e-commerce companies often search and analyze the available data on the web of their rivals, they subsequently change their item prices, item features, etc and then use this information to increase their productivity by adjusting the parameter of their items and services. Another example is "contact scraping" in which the individual's information at personal level like phone numbers or email ids are collected for marketing purposes. Through web scraping particular web sites can be analyzed to get the unstructured data and convert them in a structured manner to carry out beneficial activities.

The paper proposes the usage of linear regression algorithm on the extracted data to gain the insights and statistical relationship between the factors important to Olympic games. The main idea is to get the line that fits most of the data points. The aim of the best-fit regression line is to keep the overall estimated error (far data points) to the least. Where error counts for the distance between the data point and the regression line.

Manuscript revised December 23, 2022; accepted December 24, 2022. Date of publication December 26, 2022.

This paper available online at www.ijprse.com

ISSN (Online): 2582-7898; SJIF: 5.59

II. Literature Review

It is well known that the participants have an advantage if their country is hosting the Olympic games as they have prior knowledge about this field, environment, and also great support that they get from the audience. Countries hosting the event are expected to get three times the number of medals that they achieve normally (Clarke, 2000) [2]. Country that organizes the event are familiar to the environment and this makes a positive effect on medal count. (Bian 2005, 2005) [4].

It is also said that social-economic factors such as a country's GDP plays an important role in the performance of participants due to advance training. A country's GDP and population are greatly responsible for the winning medal count. (Bian 2005, 2005) [4].

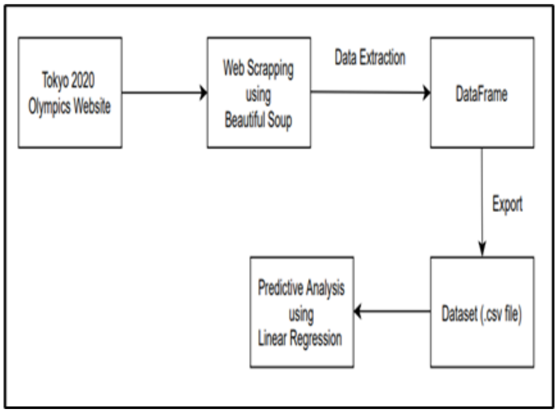

Tokyo 2020 Olympics Website
Web Scrapping using Beautiful Soup
Data Extraction
DataFrame
Export
Dataset (.csv file)
Predictive Analysis using Linear Regression

Fig.1. Process flow

Participant's age always plays an important role in the sports events, and athletes of particularly the same age range, the relative age effect (RAE) is used to determine who wins (Fletcher and Sarkar, 2012) [3]. The RAE states that one athlete may have an advantage over another in terms of maturity, experience and prior skill (Neil, Cotton, Cudros & Connor, 2016).

The Olympics are a part of history, and have influenced and great history. The Olympic games have seen several major political and social implications in its history. Like the entry of womens in the sport events, the stance against racial issues, the promotion of civil rights, the unification of nations, and the exercise of power by nations (O'Connell). [5]. Similarly, the politics of nations such as racism, World War I, terrorism and the Cold War have greatly affected the Olympics games several times in the past (Dwyer & McMaster, 2018) [6].

(David Mathew Thomas, Sandeep Mathur, 2019) Shows the process of collection of data using web scraping and python language for analytical process for, collecting, organizing, data cleaning, analysis, creating models and algorithms and at last outputting the required results [8].

III. Motivation

The Olympics have a long history, starting from 1896 to till date and marks the importance of sport from a long time. Therefore, it is useful to take in account how these crucial sports events and activities have affected the subtleties of the Olympics and how it has developed since then. Therefore, it is appropriate to address the following questions regarding historical events:

- What is the impact on the number of medals won if the country is itself hosting the Olympic event?
- How does the performance change for countries participating in Olympic Games in comparison to economic factors like GDP (GDP per capita)?
- Does country's population important in the Olympic Games?

IV. Approach

Collecting the data by using Beautiful Soup, the several web pages containing Olympic related data are scrapped and then combined to get an overall collective data. Data are visualized to get the comparative analysis of several participating countries.

Also, the related countries data are extracted with their economic factors, to make the analysis of the Olympic games at the economic level of countries. Linear regression enables us to make statistical comparisons such as countries vs medal count, medal count vs country's GDP per capita, GDP vs population, etc. Below is the process flow diagram:

4.1 Terminologies Used

4.1.1 Web scraping

Web scraping also known as web data extraction or web harvesting is a data scraping technique which is used to retrieve data and information from the target websites. World Wide Web pages can be directly accessed through web scraping techniques using a HTTP request through a web browser.

4.1.2 Beautiful Soup

The Python package that is used widely to parse the contents represented in the XML and HTML documents and formats which also includes the badly structured sequence of HTML and XML tags. The parsed pages are represented in the form of a parsed tree through which we get data from HTML tags containing information, this helps in web scraping.

4.1.3 Python

Open source, high level language creates a simpler way to read, analyze and visualize the collected data.

4.1.4 Linear Regression

According to wikipedia, linear regression is a supervised approach that is used for the formation of relationships between the dependent and independent variables. A linear line is formed that tries to cover most of the data points to show trends in the dataset.

V. Finding And Discussion

- Analysing countries participating in the Olympic games with the most population: China has the highest population and countries like India, USA, Indonesia is trailing behind. It's clear from the graph that the population of a country doesn't play an important role in the Olympics, as countries like India, Indonesia, Brazil are in the list of most populous countries but their medal count is low. On the other hand, countries like China, USA having massive population have most of the medal count

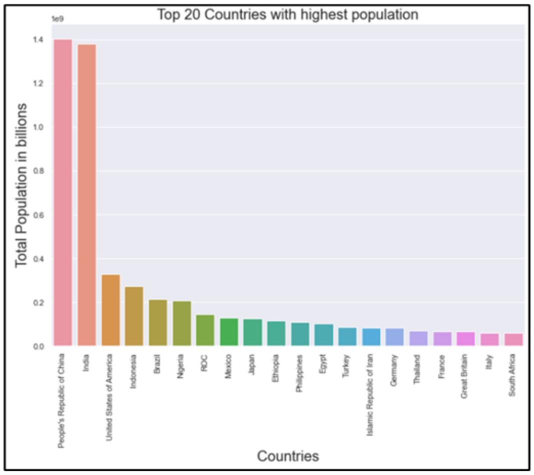


Fig.2. Countries vs population graph.

- Olympic medal share per country: We see that the USA has the highest Olympic medal share, then China, Britain and Japan are in the second, third and fourth position respectively. These countries in the top list have hosted the greatest number of Olympic games. These top countries are France, Japan, USA, Britain and so on [13]. As the countries in the top list have hosted the Olympic events in their country, we can infer that the hosting country mostly gets more medal share due to the familiar environment. Consider the latest example, the current Olympic games Tokyo 2020, Japan secured 40% more medals than the previous Olympic medal count. Even the gold medal count got raised by more than 100%, from 12 to 27. So, hosting the Olympic event for a country does play a great role in the medal count.
- Comparing the country's GDP per capita and the number of medals that respective country has won. It is observed that GDP (Gross domestic product) plays a very important role in these kinds of sports events and is considered for the estimation of a country's performance in an Olympics games (Bernard & Bus, 2000). One of the subsets of GDP is GDP per capita, which also looks at per capita earnings in a country (individual earning). We can infer that the countries with greater GDP per capita are getting more medals than the countries with less GDP per capita. Below graph depicts the points representing medals more than 20 are lying in high GDP per capita region.

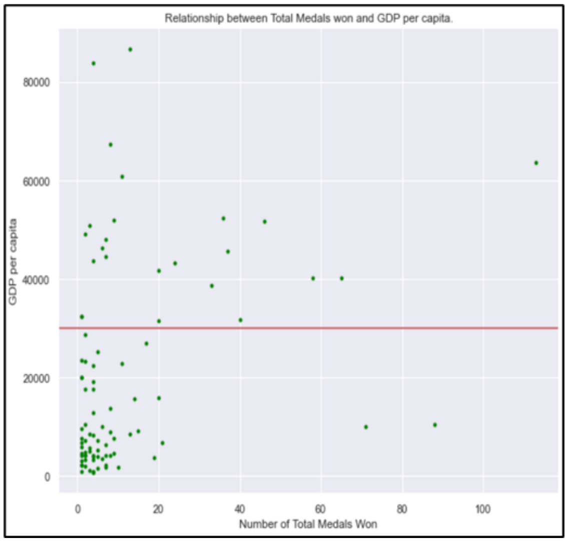


Fig.3. GDP per capita vs Medal tally.

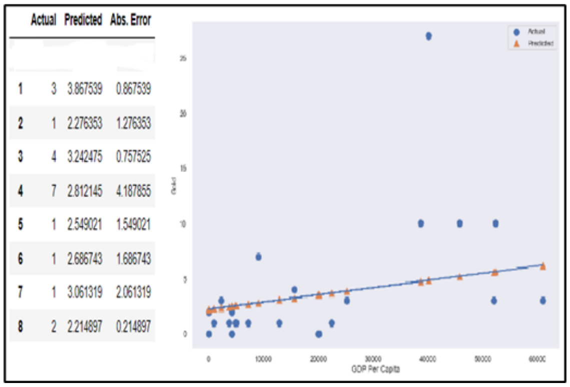
VI. Results

6.1 Correlation

To analyze the correlations of total population with GDP and several other factors, we get to see how they affect the number of medals won, GDP is always counted as a better parameter for when considering a correlation with the number of medals won. But it is not always correct to say that just because of a better economy and GDP factors, the country will win a good number of medals. Large population also contributes to selecting the best participants, such as selecting from a large sample space. From the below matrix, there is a positive correlation of around 0.3 between the GDP and the medals. Though this is an average correlation due to less data point, it does infer the existence of the relation.

6.2 Linear Regression

In this, the relationships are built between data features (Gold medal count vs GDP per capita) using linear functions which predict the unknown parameters (dependent variable) from the known data point (Independent variable). Below is the output of the regression model for GDP per capita vs gold medals won. The accuracy obtained is 42% percent which is quite remarkable for the web scraped data.



	Actual	Predicted	Abs. Error
1	3	3.867539	0.867539
2	1	2.276353	1.276353
3	4	3.242475	0.757525
4	7	2.812145	4.187855
5	1	2.549021	1.549021
6	1	2.686743	1.686743
7	1	3.061319	2.061319
8	2	2.214897	0.214897

Fig.4. Linear regression model output.

VII. Conclusion

The host nations always have a good chance of winning a medal at the Olympics. They can win at least 20-30 percent more medals. In terms of economic impact, although a country's population and per capita GDP plays an important role in analysis of the number of medals won, a country's overall GDP has been very important in determining successes in recent years. Therefore, an athlete belonging from the country hosting an Olympic event having high GDP, tends to win more medals at the Olympics games.

References

[1]. Mahtani, K.R., Protheroe, J., Slight, S.P., Demarzo, M.M.P., Blakeman, T., Barton, C. A., Brijnath, B., Roberts, N. (2012).

[2]. Clarke, S. R. (2000, June). Home advantage in the Olympic Games. In Proceedings of the 5th Australian Conference on Mathematics and Computers in Sport. University of Technology Sydney, Australia (pp. 76-85).

[3]. Fletcher, D., & Sarkar, M. (2012). A grounded theory of psychological resilience in Olympic champions. Psychology of sport and exercise, 13(5), 669-678.

[4]. Bian, X. (2005). Predicting Olympic medal counts: The effects of economic development on Olympic performance. The park place economist, 13(1), 37-44.

[5]. O'Connel, C. 13 Olympic Moments that Changed History.

[6]. Dwyer, B. B., & McMaster, A. (2018). 18 Times Politics Trumped Sport in Olympic Games' History.

[7]. Bernard, A. B., & Busse, M. R. (2000). Who wins the Olympic Games: Economic development and medal totals (No. w7998). National Bureau of Economic Research. O'Neill, K. S., Cotton, W. G., Cuadros, J. P., & O'Connor, D. (2016). An investigation of the relative age effect amongst Olympic athletes. Talent Development & Excellence, 8(1), 27-39.

[8]. D. M. Thomas and S. Mathur, "Data Analysis by Web Scraping using Python," 2019 3rd International conference on Electronics, Communication and Aerospace Technology (ICECA), 2019, pp. 450-454.

[9]. Ghazvinian, Holbert, Viswanathan. "Simple Web Scraping.

[10]. Population data: The World Bank Data.

[11]. GDP data: The World Bank Data.