WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Data Analytics on Olympics Datasets

[1]Surya Sena Reddy, [2]Suraj Kumar

[1,2]Department of Information Technology, Maturi Venkata Subba Rao Engineering College

ABSTRACT:

Today, Olympics is one of the most famous sporting events across the world, with almost all major countries taking a part in it. Over the time, Olympics has marketed itself into the countries and general public in such a way that, winning a medal in this competition has become a prestigious issue. Moreover, these games contribute to the world peace and coordination among the world countries. The primary purpose of this paper is to publish the results of performing careful data analytical operations on the data collected from the Olympics 1896 to 2016. This analysis aids in developing a resourceful knowledge from the data, about the athletes and countries performance. For this purpose, two datasets are considered, namely Athlete Events and Athlete BMI. This paper finds it's base in Descriptive and Predictive form of Analytics. Descriptive Analytics, helps in knowing what has happened in a clear way such that one can look for reasons to substantiate the findings. Whereas, Predictive analytics casts light on what would happen or what should happen scenarios.

Keywords: Descriptive Analytics and Predictive Analytics.

1. INTRODUCTION

Modern Olympics are leading sporting events across the world. They were first conducted in Athens in the year of 1896. In order to draw useful insights from the Olympic events conducted across the years, till 2016, two datasets have been considered. The first dataset is Athlete Events, which contains approximately 271116 data rows distributed over 15 attributes, which are ID, Name, Age, Sex, Height, Weight, Team, NOC, Games, Year, Season, City, Sport, Event, and Medal. These all attributes correspond to an athlete basic information, type of Sport event he or she participated in, the year and season of participation, and whether that athlete has won any medal or not. This data is used to establish relationships among the attributes involved, like the relationship between height and weight parameters, women participation status over the years, countries and athletes with most medals, countries with most gold medals in a specific Olympic year, whether a country has won atleast one Olympic medal, and the participation trend in Summer and Winter seasons. This concludes the Descriptive Analytics part.

For the Predictive Analytics part, Athlete BMI dataset has been considered which contains three attributes, Athlete Name, Sport and BMI (Body Mass Index). The primary purpose of this dataset is to train a model on the Sport and BMI parameters, and to employ it to forecast a suitable sport based on the BMI values of an individual. Also, this part includes predicting weight of an athlete based on the height.

Before the data analytical operations are performed, the datasets are transformed into DataFrames, and thorough data cleaning routines are performed to remove any Null values. Thereafter, data analytical operations are performed using Python and libraries such as NumPy, Pandas, Sklearn, Plotly, Streamlit and Matplotlib are used. Visualization such as bar charts, pie charts have been used to showcase the results.

2. LITERATURE SURVEY

Here is an overview of previous publications that have been referred during the process of research.

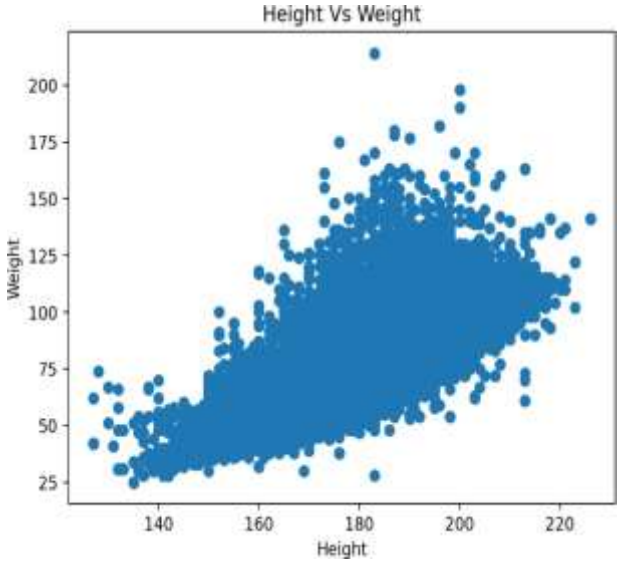
1. Performance Analysis in Olympic Games using Exploratory Data Analysis Techniques by **Yamunathangam.D, Kirthicka.G, Shahanas Parveen.**
2. 120 years of Olympic Games — How to analyze and visualize the history with R by **Saul Buentello.**
3. Analyzing Evolution of the Olympics by Exploratory Data Analysis using R by **Rahul Pradhan, Karthik Agrawal, Anubhav Bag.**

3. DESCRIPTIVE ANALYTICS

This type of analytics is done in order to know the accurate information on what has been happened. In this context, here some questions have been answered which are given below.

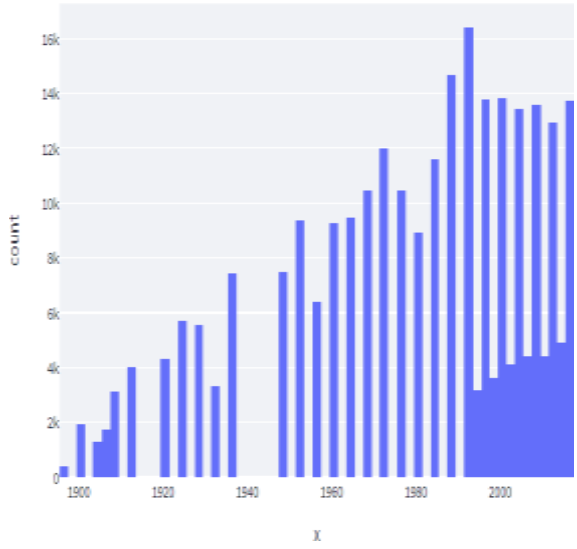
A. Analyse the change in Weight of an athlete with respect to the Height.

The relationship between the Heights and Weights of the athletes is illustrated in the form of a scatterplot in the figure below. This plot clearly shows that as height increases, weight of an athlete also changes correspondingly. Therefore, one can deduce that both the parameters are positively correlated.



B. Determine women participation over the years

From the given histogram, it can be inferred that there has been a gradual increase in the number of women athletes over the years, from 1896 to 2000's. Initially there has been fewer participation of women, even less than 1000 in 1896. However, from 1900 the graph of women involvement has picked up pace, reaching maximum in 1990's at a number slightly greater than 16k. In 2000's however there has been a slight decline, but the number has been levelled at an average of 14k. Also, the years 1904, 1932, 1956 and 1976- 80 have witnessed fewer participation than the preceding years.



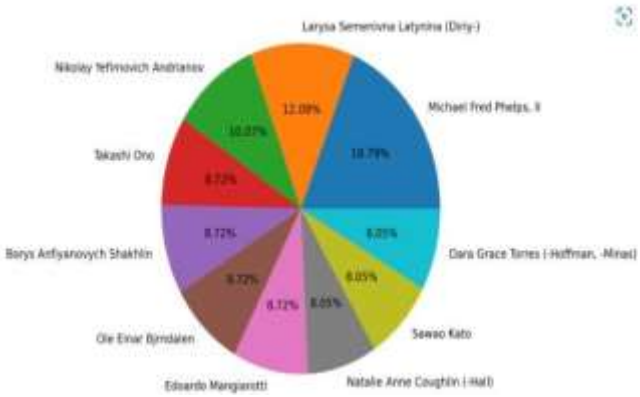
C. Consider Athletes with most medals

Given DataFrame, contains ten athletes ranked according to the number of individual Olympic medals won by them. From the data, it can be known that Michael Fred Phelps has won the most medals so far in history of the Olympics. With 28, he leads the way, followed by Larysa and Nikolay with

18 and 15 individual Olympic medals respectively. Thereafter, the competition seems stiff as many athletes (Takashi, Borys, Ole and Edoardo) are tied at same spot with equal performance, judging by the 13 medals won by them. The DataFrame concludes with Natalie, Dara, and Sawao winning 12 Olympic medals each.

The pie chart illustrates the share of each athlete against the total sum of Olympic medals won by all of them. Here, the total sum amounts to 149, of which 18.79 percent are won by Michael. The percentage share of each athlete in the list can be seen in the pie chart, although with a different shade of colour to identify the athletes uniquely. A similar analysis can be made out from the pie chart, like the DataFrame list.

	Total
Michael Fred Phelps, II	28
Larysa Semenivna Latynina (Diriy-)	18
Nikolay Yefimovich Andrianov	15
Takashi Ono	13
Borys Anfiyanovych Shakhlin	13
Ole Einar Bjrndalen	13
Edoardo Mangiarotti	13
Natalie Anne Coughlin (-Hall)	12
Sawao Kato	12
Dara Grace Torres (-Hoffman, -Minas)	12



D. Determine Countries with Most medals

In this, we try to estimate the countries with the most medals in the Olympics history. The countries are grouped against the medals they have won, and the list is provided in the form of a DataFrame. This gives the top ten countries with most medals, with the United States leading the way, followed by the Soviet Union, Germany and so on. Note that data doesn't consider the dissolution of the Soviet Union.

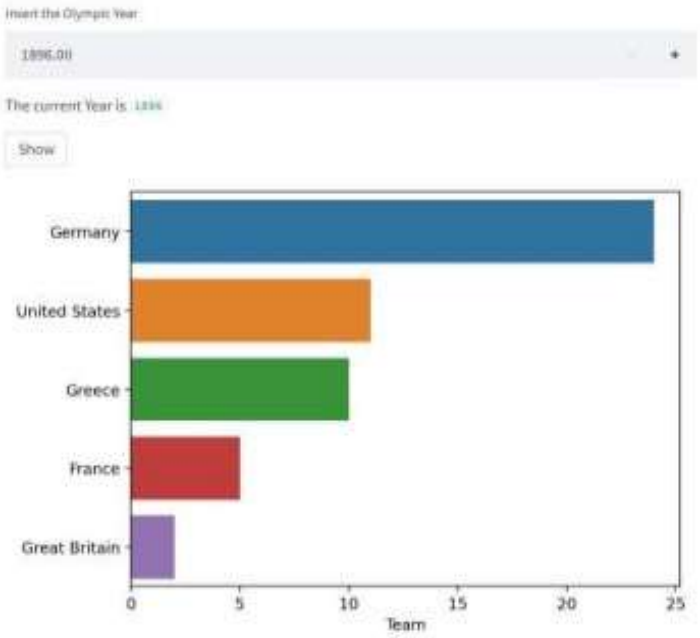
	0
0	United States
1	Soviet Union
2	Germany
3	Great Britain
4	France
5	Italy
6	Sweden
7	Australia
8	Canada
9	Hungary

E. Find if your country is in the Zero-Medal list

Zero-Medal list is a list of countries that haven't won any medals so far in the history of Olympics, that is till 2016. One can check if a particular country is in this list, by just providing the country name as an input. Now there can be three possible outputs depending on the performance of the input country. The first scenario outputs the phrase "Your country has won at least 1 Medal, so chill" if the given country has won 1 or more medals. Or the second scenario, if the country has won no medals, outputs the phrase, "Sorry to break it to you, your country is in the Zero-Medal list". Lastly, there can be a situation where the given country isn't present in the dataset. In such cases the output would be given as "The country isn't listed in the dataset''. Given below is a scenario where a country isn't present in the Zero-Medal list. This concludes this topic.

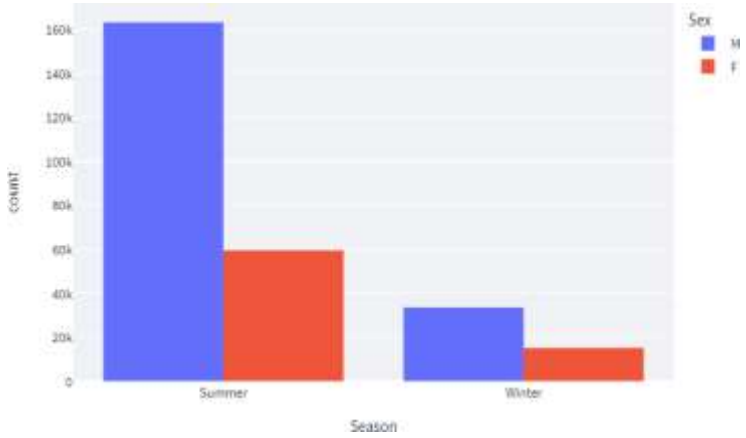
F. Countries winning the most Gold in a specific year

Here, an Olympic year should be provided as an input which produces a bar chart of countries that won the most gold in that specific year. An Olympic year is a year in which Olympics have been conducted, either Summer or Winter. Given below is a bar chart that corresponds to the year 1896, the first Olympic year. In this year, Germany has won the most gold, about twenty-four. It's followed by the highest medal winning country in the history, the United States. Then comes another European country Greece with ten gold medals, then France with five and lastly the Great Britain with two gold medals. Like this one can view the share of countries in the gold medals every Olympic year.



G. F) Summer and Winter Olympics participation

Olympics is conducted in two seasons, summer and winter. Usually, Summer Olympics are conducted every leap year, while Winter Olympics are conducted every two years after the Summer Olympics. Given below is a histogram showing the participation trend in both the seasons. From this, it can be drawn that Summer Olympics participation is far more superior when compared to the Winter Olympics participation. A total of 163k men and about 60k women participated in the Summer, while only 35k men and 20k women participated in the Winter. The probable reasons need to be ascertained.



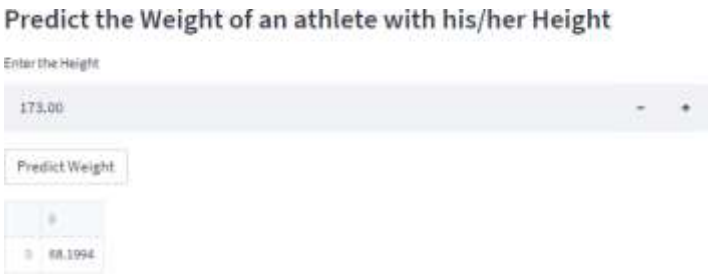
4. PREDICTIVE ANALYTICS

Involves forecasting a value based on an independent parameter.

A. Predicting Weight using height

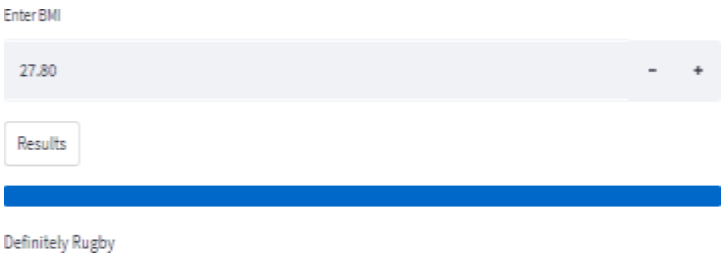
Linear Regression algorithm has been used to forecast weight of an athlete with the corresponding height. This model estimates the value of dependent variable based on independent variable. Here, in this study Height is an independent entity on which Weight depends. For this Athlete Events dataset has been considered, which contains over 270k data tuples spread over fifteen attributes, of which Height and Weight are the required part for this prediction. The choice of the algorithm is done after evaluating the scatter plot between these parameters, which shows a linear relationship.

Whenever height is provided as an input, then the predicted weight can be seen as output. However, the height shouldn't be less than 130 cm or greater than 220 cm. Given below is a case, where the height is 173cm, and the predicted weight is given below. Note that the accuracy of prediction model is approximately 63 percent.



B. Predicting an Apt Sport

In this part, an apt sport for an individual is predicted based on his or her BMI values. For this, an Athlete BMI dataset has been considered that contains twenty data tuples distributed over three attributes (Name, Sport and BMI). Here, we focus only on Sport and BMI parameters. Linear Regression algorithm has been used to forecast the suitable Sport (dependent entity) using the BMI (independent entity) of a person. Note that the BMI values should lie between 15 and 40. Given below is a test case, consider the input as 27.8 BMI, and the predicted sport is Rugby. Note that the accuracy of this prediction is approximately 90%.



5. CONCLUSION

Descriptive and predictive form of analytics have been performed on the considered datasets, containing information from 1896 to 2016 Olympics. Countries with the most medals are found, and needs to maintain their excellent performance. Women participation has increased over the years, which stands as a testimony to the women empowerment through the Olympics. While accuracy of the predictions can be increased.

References

1. https://www.researchgate.net/publication/330847008_Performance_analysis_in_olympic_games_using_exploratory_data_analysis_techniques
2. https://www.researchgate.net/publication/265033380_Da ta_mining_of_sports_performance_data
3. https://www.researchgate.net/publication/23756788_Eco nomics_and_Olympics_An_Efficiency_Analysis
4. https://docs.streamlit.io
5. Sources: Athlete Events: https://www.kaggle.com/datasets/heesoo37/120-years-of- olympic-history-athletes-and-results

Athlete BMI:

Joe Kovacs,40,4

Patty Mills,24,2

Ryan Crouser,35.9,4

Richie Mccaw,30.6,3

Goran Dragic,23.8,2

Brigid Kosgei,17.3,1

Seth Curry,23.8,2

Heather Moyce,22.6,3

Zerseney Tadese,21.1,1

Fernando Portugal,28.1,3

Kyrie Irving,24.9,2

Eliud Kipchoge,18.6,1

Valerie Adams,32.2,4

David Harvey,27.8,3

Tom Walsh,35.1,4

Lelisa Desisa,20.1,1

Ivanka Khristovia,30.4,4

Santiago Gomez,25.2,3

Abdi Nageeye,19.8,1

Bruce Brown,24.7,2

Note: 1 – Marathon, 2- Basketball, 3- Rugby, 4- Shotput