Performance Analysis of Tokyo Olympics 2021 Using Exploratory Data Analysis Techniques

Dr. Pooja Bijlani1[1] Dr. Vikas Saxena[2]
[1]Department of Computer Science & Engineering [2]Department of Physical Education
[1]IES college of Technology, Bhopal, India [2]RNTU, Bhopal, India

Abstract— Python has rich libraries for data analysis and visualization. These libraries can be used for data analysis and visualization to find some in site of the data which is good to produce result. some leading companies use these results for improving customer satisfaction plays a great role to analyze factor , that affect the review rating. Exploratory data analysis (EDA) uses data interpretations which can be implemented in tabular format. This paper uses Python for data analysis. It is best for data analysis as because Python is object oriented and interactive programming language. It is an open source with rich sets of libraries like Pandas, Scikit, MATplotlib, seaborn etc. Different types of Graphs and various parameter are used to analyze Olympics 2021 review data sets which contains data three CSV files.

Keywords: Exploratory Data Analysis Techniques, Python

I. INTRODUCTION

Data science evolving very faster in today's world. Every minute millions of data is generated which is not so easy to process and stored the data manually. Data analysis and visualization projects used allow for reaching even deeper understanding. The programming language Python, with its easy commands and English words used syntax, offers wonderful open-source platform alternative to traditional techniques and applications. Data analytics help businesses with greater understandings with their efficiency and performance, and helps the companies take more informed decisions. Take a case, a company might be planning to analyze customer attributes in order to target customers and projecting ads for improving sales. Data analysis can be to almost any aspect of a business if one understands the tools available to process information. The reviews of customers are analyzed by using visualization methods. Exploratory Data Analysis (EDA) is an approachable tool for preparing the summarize data by taking their main characteristics and visual representations. EDA focuses more deeply on figuring assumptions required for model fitting and hypothesis testing, and handling missing values and making transformations of variables as needed. EDA encompasses IDA. EDA quickly describes the data sets number of rows/columns, missing data, data types and preview

Data Frame is Pandas structure, which is used to store data in row and column way. It is a 2D labeled array, which is actually an ordered collection of columns.

II. BACKGROUND

A. Python for Data Analytics:

Pandas's provide open source library for data analysis. Pandas has derived its name from "panel data system". Pandas make it a simple and easy process as compared to other available tools. Pandas have become a popular choice for data analysis which is a process of evaluating big data using analytical and statistical tools so as to discover useful information and conclusion to support business decision-making.

B. Data Structures

Data Structures is a particular way of storing and organizing data in a computer to suit a specific purpose so that it can be accessed and worked with in appropriate ways. Pandas offer many data structures to handle variety of data. Series and Dataframes are two popular data.

C. Data Visualization:

Refers to graphical or visual representation of statistics and data using visual elements like, bar graphs, pie charts and maps etc. Data visualization is immensely useful in decision making .Data visualization unveils patterns; trends outliers, correlations etc. in the data, and thereby helps decision makers understand the meaning of data to drive business decisions. For instance, a company has to decide about, 'which advertising solution it invest in to promote its new product?

Data visualization is the solution. For Data visualization in Python, the matplotlib library's Pyplot interface is used. The matplotlib is a high quality plotting library of Python that provides many interfaces and functionality for 2D- graphics similar to Matlab,s in various forms.

D. IDLE:

Integrated Development and Learning Environment. There are many development environments online or offline available that can be used to write and execute python code. Some of the examples are VSCode, Spyder that can be used offline, and there are some online platforms Jupyter notepad, Colaboratory.

E. Importing /Exporting File:

Dataframes are capable of storing data in 2D form .Most data files that we use to store data such as spreadsheet files or databases tables, also store the data in 2D tabular formats. Since dataframes can also hold data in similar way, we can transfer data from dataframes to such data files or from files into dataframes. Python Pandas provides modules that can help importing and exporting such files .There are 14 such formats available that we can import/export Data sets data. We can find the missing value of data. We can find the possible relationship among data. Data visualization is done by the use of tabular data and understanding the characteristics validating the data. III. Model Building We use the statistical model or machine learning model to describe the variable and working of the variable. Model can be supervised or unsupervised model. We can use classification, regression model to get the output. We can visualize the result by the use of model. After that we have to evaluate the model.

All rights reserved by www.ijsrd.com

III. LITERATURE SURVEY

Performance measure for a country in Olympics can be predicted using their past performance. By predicting their revised version manuscript received on 25 November, 2018. Yamunathangam.D, Assistant Professor, Kumaraguru College of Technology, Coimbatore (Tamil Nadu), India.

Kirthicka.G, UG scholars, Kumaraguru College of Technology, Coimbatore (Tamil Nadu), India. Shahanas Parveen.I, UG Scholars, Kumaraguru College of Technology, Coimbatore (Tamil Nadu), India

K. Ulaga Priya1 et al. [6] has done exploratory analysis on prediction of loan privilege for customers using random forest. They have used R programming for exploratory data analysis.

Bogumil M. Konopka et al. [7] has done exploratory data analysis of a clinical study group. Development of a procedure for exploring multi dimensional data.

IV. EXPLORATORY DATA ANALYSIS (EDA):

Primarily, exploratory data analysis is an approach to see what the data can communicate us away from the formal modeling or hypothesis testing task.

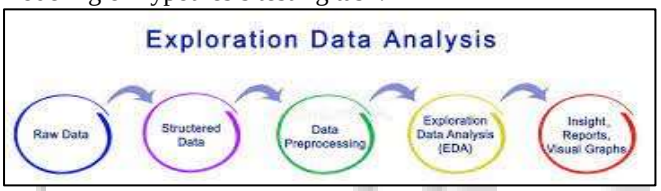


Fig. 1: Exploratory Data Analysis

It is the first stage of data analysis. Here we can know about the content of the data set and characteristic of data set. It tells about the size of the

A. Working with the Data Sets:

It's time to explore the data and find about it. The data we are using belongs to Olympics 2021 data set. Examaplory data analyse will be performed on this data set.

1) In the first step we have imported the Pandas libraries.

```
Import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
```

2) After that we have imported fairly large Olympics2021 CSV files as a dataframe df. It gives the datasets in the form of rows and columns. In our CSV file 93 rows and 7 columns are there. This is shown in figure. Here we are using pandas read_csv() method to read the data .the data so read is in form of DataFrame. Figure also shows the DataFrame.

Fig. 2: Importing pandas library and reading data set

3) Finding Dataset details:

Next task is to analyze the data set. Finding out what type of data is stored. Using functions like Medal.info( ) and Medal.shape( )

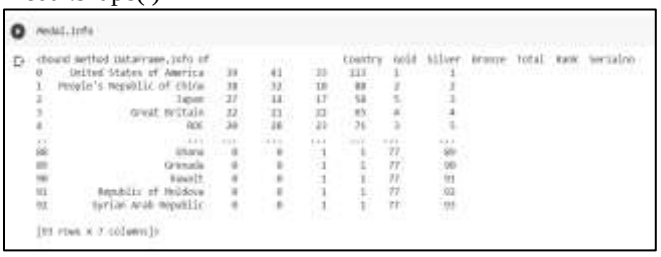


Fig. 3: Showing details of the data set

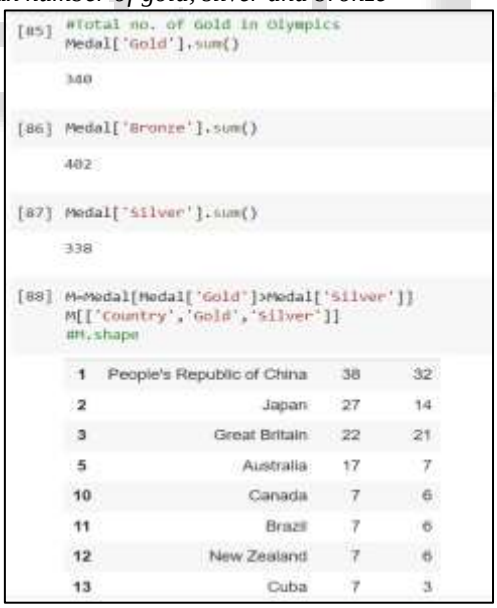
4) Data description:

describe() function is used to get descriptive statistics from Dataset like count, mean, minimum, maximum, quartiles etc.

	S.no	Gold	Silver	Bronze	Total	Rank
count	93.000000	93.000000	93.000000	93.000000	93.000000	93.000000
mean	46.333333	3.655914	3.634409	4.322581	11.612903	43.494624
std	26.219116	7.022471	6.626339	6.210372	19.091332	24.171769
min	1.000000	0.000000	0.000000	0.000000	1.000000	1.000000
25%	24.000000	0.000000	0.000000	1.000000	2.000000	23.000000
50%	46.000000	1.000000	1.000000	2.000000	4.000000	47.000000
75%	70.000000	3.000000	4.000000	5.000000	11.000000	66.000000
max	86.000000	39.000000	41.000000	33.000000	113.000000	77.000000

Fig. 4: Describe function to display statistics

5) Finding total number of gold, silver and bronze medals and Max number of gold, silver and bronze



1	People's Republic of China	38	32
2	Japan	27	14
3	Great Britain	22	21
5	Australia	17	7
10	Canada	7	6
11	Brazil	7	6
12	New Zealand	7	6
13	Cuba	7	3

Fig. 5: Using sum function to show statistics

6) Data visualization essentially means graphical representation of compiled data. Thus, graphs and charts are very effective tools for data visualization. To create different type of graphs and charts using Pyplot. Some commonly used chart types are:

1) Line Chart: A line chart or line graph is a type of chart which displays information as a series of data points called 'markers' connected by straight line

All rights reserved by *www.ijsrd.com*

segments. With PyPlot, a line chart is created using plot() function.

2) Bar Chart: A bar chart is a graph that represents categorical data with rectangular bars with heights or lengths proportional to the values that they represent .The bars can be plotted vertically or horizontally. With PyPlot, a bar chart is created using bar() and barh() functions.
3) Pie Chart: A pie chart is a circular statistical graphic, which is divided into slices to illustrate numerical proportion. With PyPlot, a pie chart is created using scatter() function .

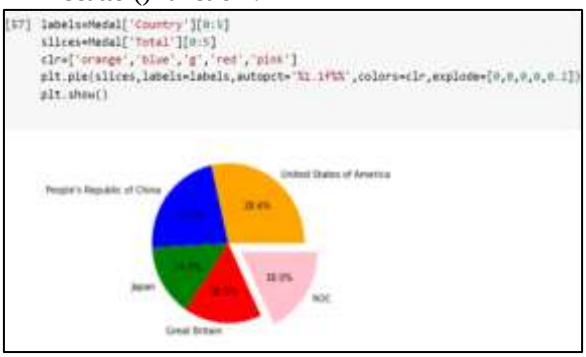


Fig. 6: Pie Chart to show top 5 Countries.

7) Based on dataset a detailed visualization of total number of medals count Country wise Chart sorted

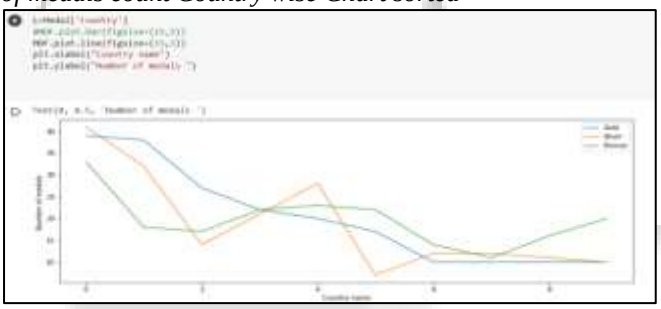


Fig. 7: Bar chart to show comparative analysis of medals

8) In the next round of commands Excel data is read using pandas function Athletes data is in Excel. Further an analysis is done with column "NOC" number of athletes belong to various countries is generated and first 10 listed records are plotted. The comparative study of participant's number among countries can be easily done using matplotlib library's functions.

In this data analysis according to column discipline certain statistics are drown and further plotted.

plt.grid() function is used to show the graph in grid format . The Data shows the names of different sports and the number of participants in that.

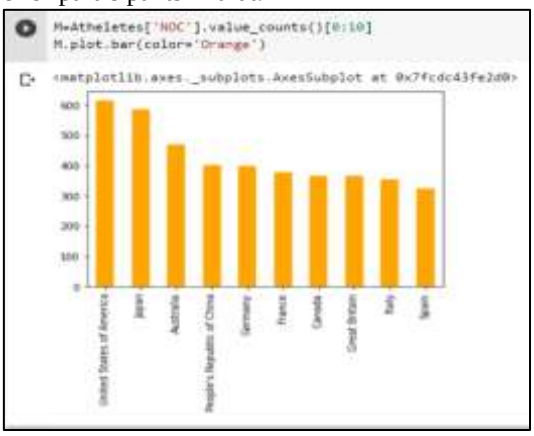


Fig. 8: Top 10 Athletes with their countries

9) In third read process Gender description Data set is in Excel. Function read_excel() is used . All the data analysis tools are used like info(), sum() ,sort_values() . Total number of Females and total number of males displayed.

Fig. 9: Showing sum of male and female candidates

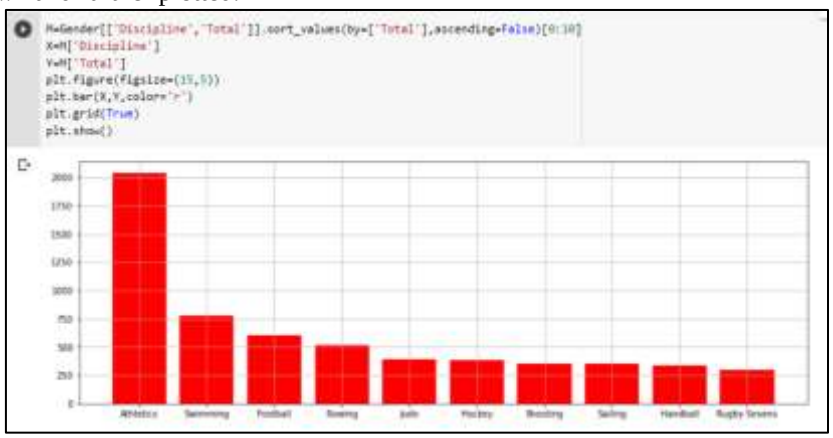


Fig. 10: Further the Dataset is analyzed

All rights reserved by *www.ijsrd.com*

V. CONCLUSION:

This article helps reader understand tools of Pandas and Matplotlib library for Data Visualization. Also the import/export of CSV and excel files are shown; the main aim of this Data Analysis is to find incite of data using online platform. We got the required result taking different parameter. In future we will use more data sets and other functions to get the clear idea related to exploratory data analysis.

So, in the analyses shown, it is useful to identify the country that needs more skills, the field of sports in which players are performing well and players who need practice to enhance themselves in upcoming Olympics.
Countries have performed well so far.

REFERENCES

[1] Antarlina Sen and Gaurang Margaj,"A prediction model for which country will win highest number of „Gold" , 2016
[2] Leonardo De Marchi,"Data mining of Sports performance data", 2011.
[3] Huang-Chiashih,"Survey on content-aware Video Analysis for Sports" , IEEE Transactions on Circuits and Systems for Video Technology, Vol. 99, No. 9, January 2017.
[4] Chandra Segar Thirumalai and Monica Sankar,"Heuristic Prediction of Olympics using Machine Learning" , International Conference on Electronics, Communication and Aerospace Technology ,April 2017.
[5] Alexander Rathke and Ulrich Woitek,"Economics and Olympics: An Efficiency analysis" , January 2007. 6. Olympics Dataset, Available: https://www.kaggle.com.
[6] Analysis and Comparison with Modern Software Development international journal Approaches Nat. Nat. Volatile s& Essent. Oils, 2021;8(5):7553-7559

All rights reserved by *www.ijsrd.com*