DATA ANALYTICS ON OLYMPICS DATASET

BY:

C.SURYA SENA REDDY(2451-19-737-121)

SURAJ KUMAR JANA(2451-19-737-123)

UNDER THE GUIDANCE OF

DR.A.V. KRISHNA PRASAD

ASSOCIATE PROFESSOR,IT.

CONTENTS

- Abstract
- Statement of Problem
- Literature Survey
- Scope of Work
  - Exploratory Data Analysis
  - Data processing
  - Descriptive Analysis
  - Predictive Analysis
- Languages used
- Hardware and Software Requirements
- Existing System
- Proposed System
- Architecture Overview
  - Home
  - Exploratory Data Analysis
  - Data preprocessing
  - Trends
  - Linear Regression
  - Prediction
- Applications/Usage
- Conclusion
- Future Scope
- References

ABSTRACT

- Olympics is one of the leading sporting events and this project revolves around performing careful data analytics operations on the data collected from it. For this objective, two datasets that contain information about the various events and the participated athletes has been analyzed. This project finds its base in Descriptive and Predictive forms of Analytics

STATEMENT OF PROBLEM

- The problem statement revolves around knowing the trends and relationship between attributes of the participated athletes. For this purpose, the Athlete events dataset containing a total of 15 attributes, describing about the Athlete object has been considered.
- It also includes a basic predictive analysis on BMI values of athletes and their concerned sport. For this purpose Athlete BMI dataset has been imported. The results must be embedded into an application (or) interface.

LITERATURE SURVEY

1)

- Title : Performance Analysis in Olympic Games using Exploratory DataAnalysis Techniques
- Authors: Yamunathangam.D, Kirthicka.G, Shahanas Parveen
- Year : Jan 2019
- Limitations : Lack of Interface

2)

- Title : 120 years of Olympic Games– How to analyze and visualize thehistory with R
- Author : Saul Buentello
- Year: Aug 1 2021
- Limitations: No interface, data preprocessing, or documentation.

3)

- Title : Analyzing Evolution of the Olympics by Exploratory Data Analysisusing R
- Author(s): Rahul Pradhan, Karthik Agrawal, Anubhav Bag
- Year: March 2021
- Limitations: Lack of Interface, unappealing Visualization.

SCOPE OF WORK

The work includes four major activities after importing the few data sets:

- **<u>Exploratory Data Analysis:</u>** Includes a careful examination of datasets, getting a aquatinted with the attribute column properties, estimating the NAN values etc.
- **<u>Data Preprocessing:</u>** Includes cleaning the dataset to bring out a more polished one such that it would make the analytics part easier.

- **Descriptive Analysis:** It is performed for knowing the trends or for answering the question “**what happened** “. This is done using various visualisation tools such as matplot lib, plotly.libraries
- **Predictive Analysis:** It involves making basic predictions based on learnt data.

The outputs are shown using an interface based on streamlit(python).

LANGUAGES USED

- **Python**(Libraries such as Numpy, Pandas, sklearn , plotly etc.)
- **Anaconda environment** for easy package/modules management.

HARDWARE AND SOFTWARE REQUIREMENTS

- **Hardware**

  Processor: Pentium V (or) higher.

  RAM: 1GB

  Space on Hard disk: minimum 512MB

- **Software**

  Web browser/engine: Google chrome (or) IE

  Python libraries (matplotlib, plotly, numpy, pandas, sklearn, seaborn)

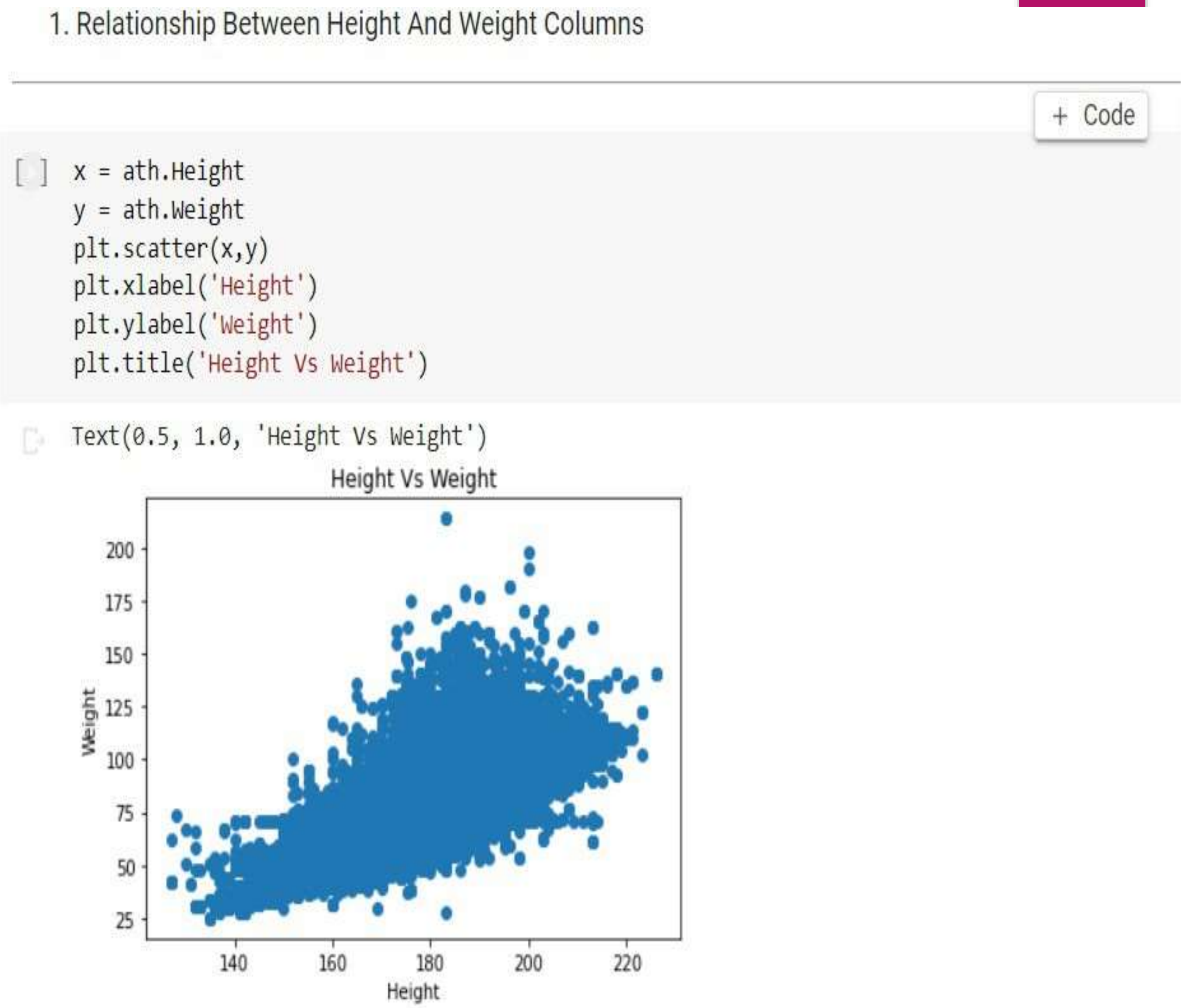
  Anaconda environment

  PC running with windows 7 (or) more

  Streamlit framework

EXISTING SYSTEM

Usually,the analytics part is done using Jupyter Notebook, Google colab and other tools. In those environments, cells are present which contains the code that produces output on run command. A serious drawback is the lack of a proper interface, that would make results even more appealing to look at.



1. Relationship Between Height And Weight Columns

```
x = ath.Height
y = ath.Weight
plt.scatter(x,y)
plt.xlabel('Height')
plt.ylabel('Weight')
plt.title('Height Vs Weight')
```

```
Text(0.5, 1.0, 'Height Vs Weight')
```

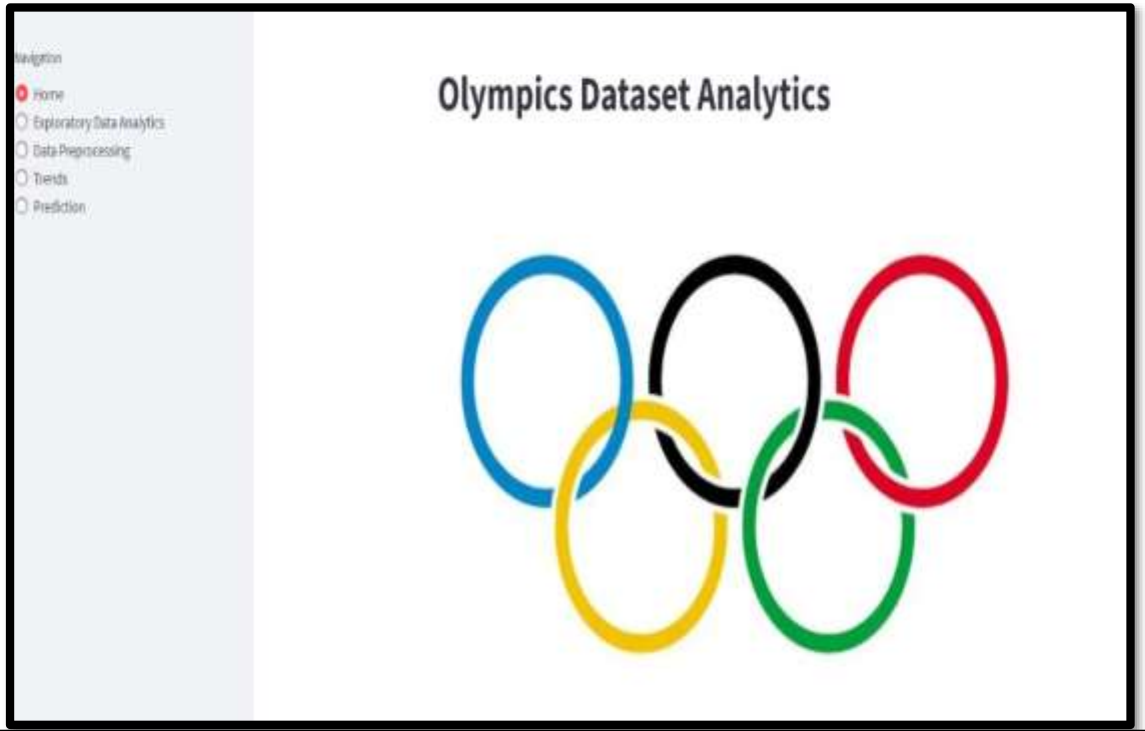
PROPOSED SYSTEM

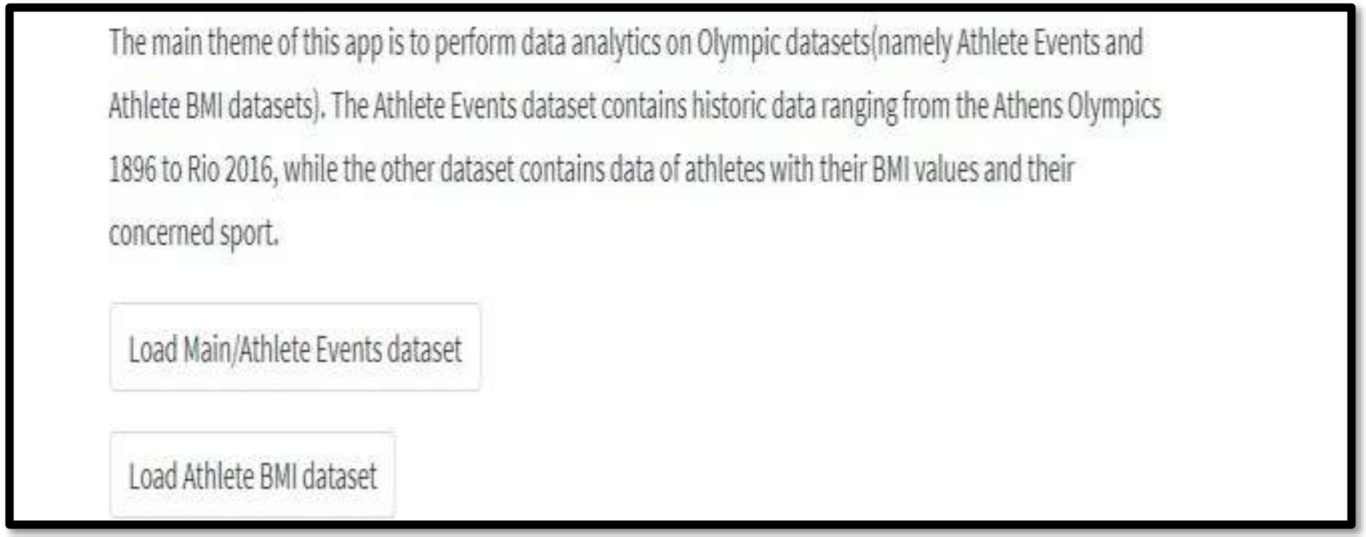
A simple application (or) Graphical User Interface that would act as an intermediate between the results and the users. The coding part is done using Python, for building the application streamlit(python) has been used and for analytics part libraries such as Numpy, pandas, matplotlib etc(python) have been implemented accordingly.

ARCHITECTURE OVERVIEW

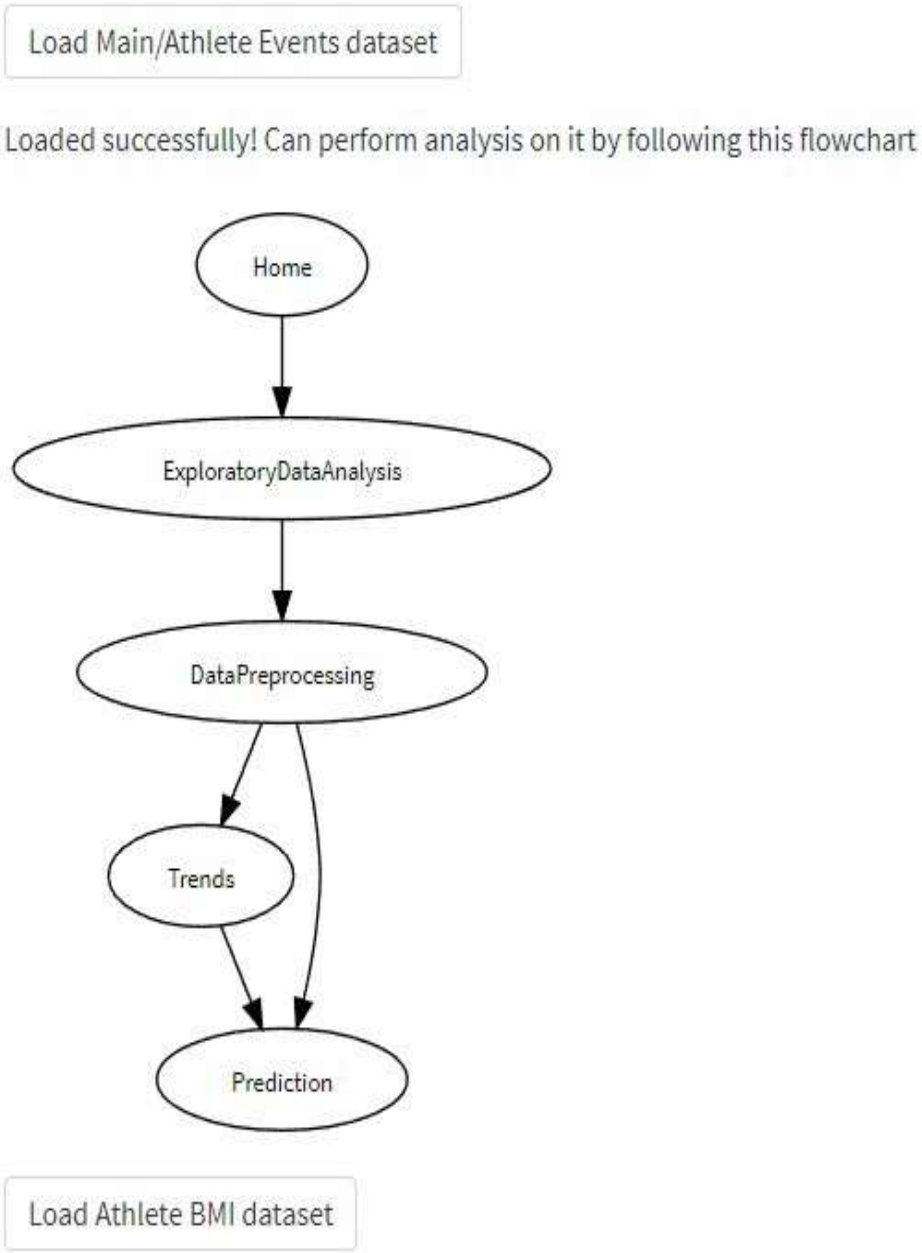
- **Home:**

Contains basic information on olympics and LOAD buttons to import the dataset which on load produces a map to be followed.





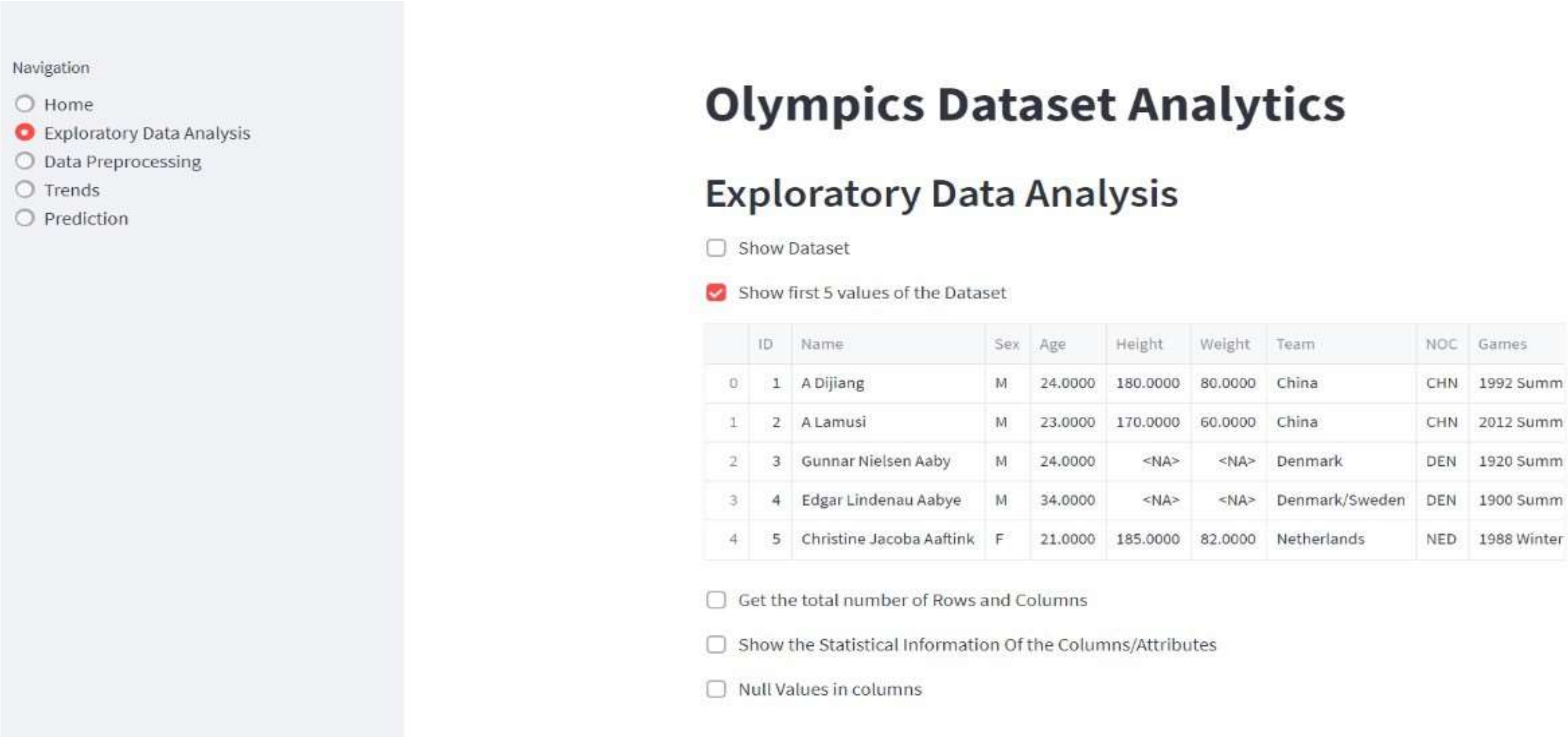
The main theme of this app is to perform data analytics on Olympic datasets(namely Athlete Events and Athlete BMI datasets). The Athlete Events dataset contains historic data ranging from the Athens Olympics 1896 to Rio 2016, while the other dataset contains data of athletes with their BMI values and their concerned sport.



Loaded successfully! Can perform analysis on it by following this flowchart

- **Exploratory Data Analysis:**

Contains the options for exploration of the datasets.



Navigation

○ Home

● Exploratory Data Analysis

○ Data Preprocessing

○ Trends

○ Prediction

Olympics Dataset Analytics

Exploratory Data Analysis

□ Show Dataset

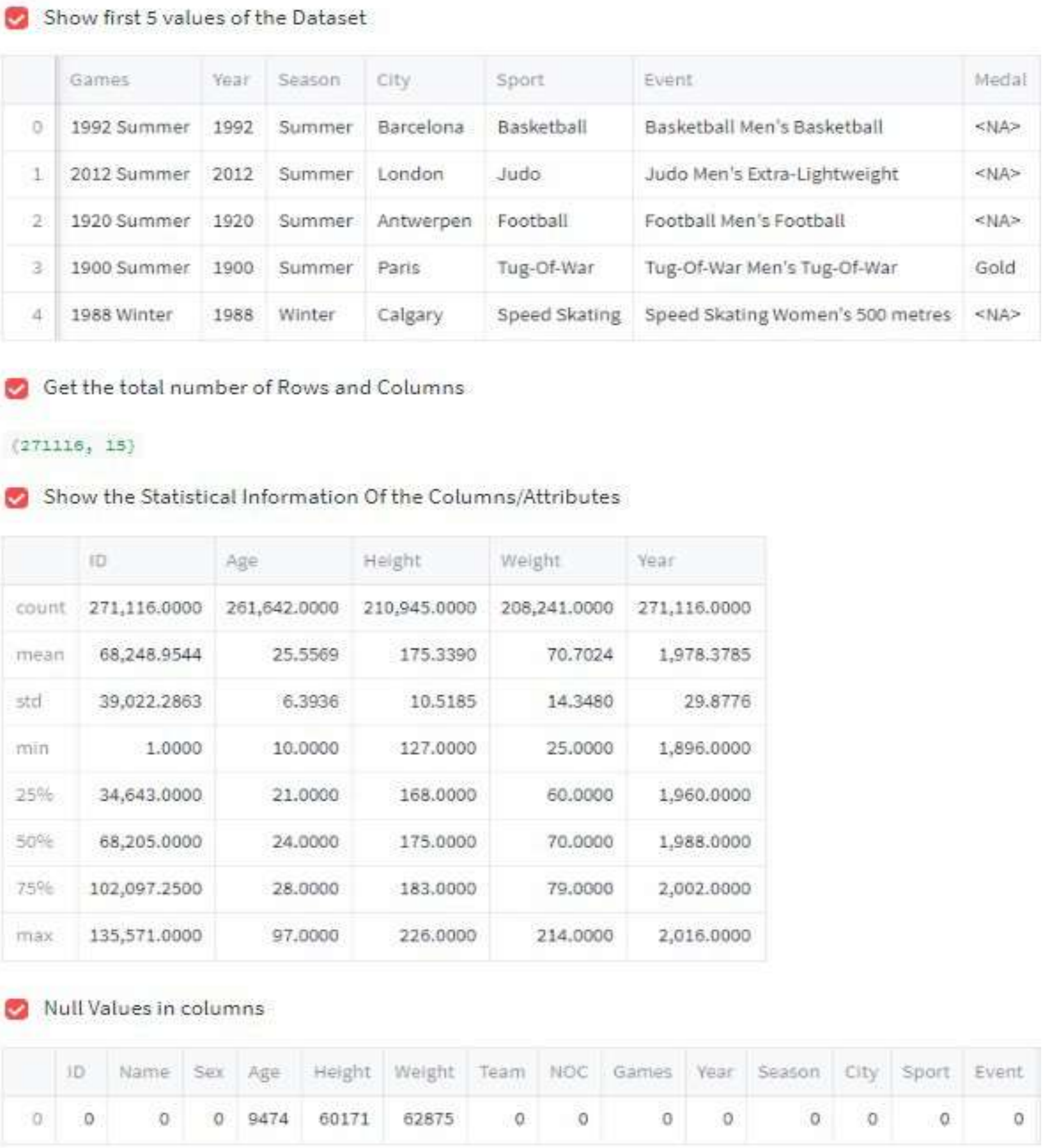
☑ Show first 5 values of the Dataset

	ID	Name	Sex	Age	Height	Weight	Team	NOC	Games
0	1	A Dijiang	M	24.0000	180.0000	80.0000	China	CHN	1992 Summ
1	2	A Lamusi	M	23.0000	170.0000	60.0000	China	CHN	2012 Summ
2	3	Gunnar Nielsen Aaby	M	24.0000	<NA>	<NA>	Denmark	DEN	1920 Summ
3	4	Edgar Lindenau Aabye	M	34.0000	<NA>	<NA>	Denmark/Sweden	DEN	1900 Summ
4	5	Christine Jacoba Aaftink	F	21.0000	185.0000	82.0000	Netherlands	NED	1988 Winter

□ Get the total number of Rows and Columns

□ Show the Statistical Information Of the Columns/Attributes

□ Null Values in columns



☑ Show first 5 values of the Dataset

	Games	Year	Season	City	Sport	Event	Medal
0	1992 Summer	1992	Summer	Barcelona	Basketball	Basketball Men's Basketball	<NA>
1	2012 Summer	2012	Summer	London	Judo	Judo Men's Extra-Lightweight	<NA>
2	1920 Summer	1920	Summer	Antwerpen	Football	Football Men's Football	<NA>
3	1900 Summer	1900	Summer	Paris	Tug-Of-War	Tug-Of-War Men's Tug-Of-War	Gold
4	1988 Winter	1988	Winter	Calgary	Speed Skating	Speed Skating Women's 500 metres	<NA>

☑ Get the total number of Rows and Columns

```
(271116, 15)
```

☑ Show the Statistical Information Of the Columns/Attributes

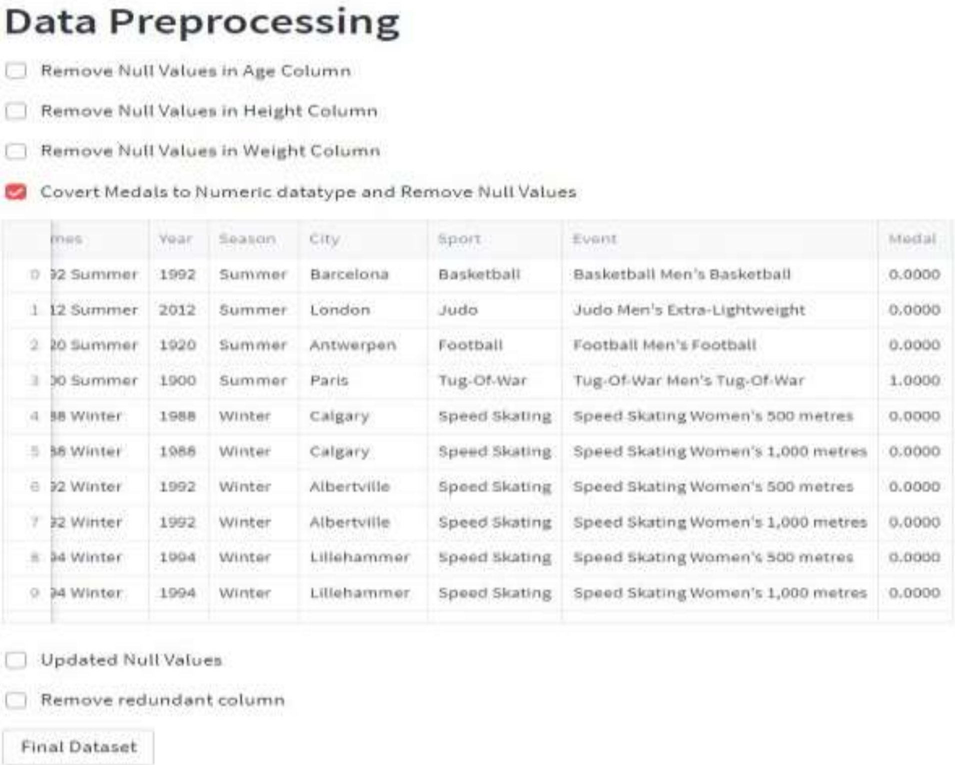
	ID	Age	Height	Weight	Year
count	271,116.0000	261,642.0000	210,945.0000	208,241.0000	271,116.0000
mean	68,248.9544	25.5569	175.3390	70.7024	1,978.3785
std	39,022.2863	6.3936	10.5185	14.3480	29.8776
min	1.0000	10.0000	127.0000	25.0000	1,896.0000
25%	34,643.0000	21.0000	168.0000	60.0000	1,960.0000
50%	68,205.0000	24.0000	175.0000	70.0000	1,988.0000
75%	102,097.2500	28.0000	183.0000	79.0000	2,002.0000
max	135,571.0000	97.0000	226.0000	214.0000	2,016.0000

☑ Null Values in columns

	ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season	City	Sport	Event
0	0	0	0	9474	60171	62875	0	0	0	0	0	0	0	0

- **Data preprocessing:**

Contains options to polish the dataset.



Data Preprocessing

☐ Remove Null Values in Age Column

☐ Remove Null Values in Height Column

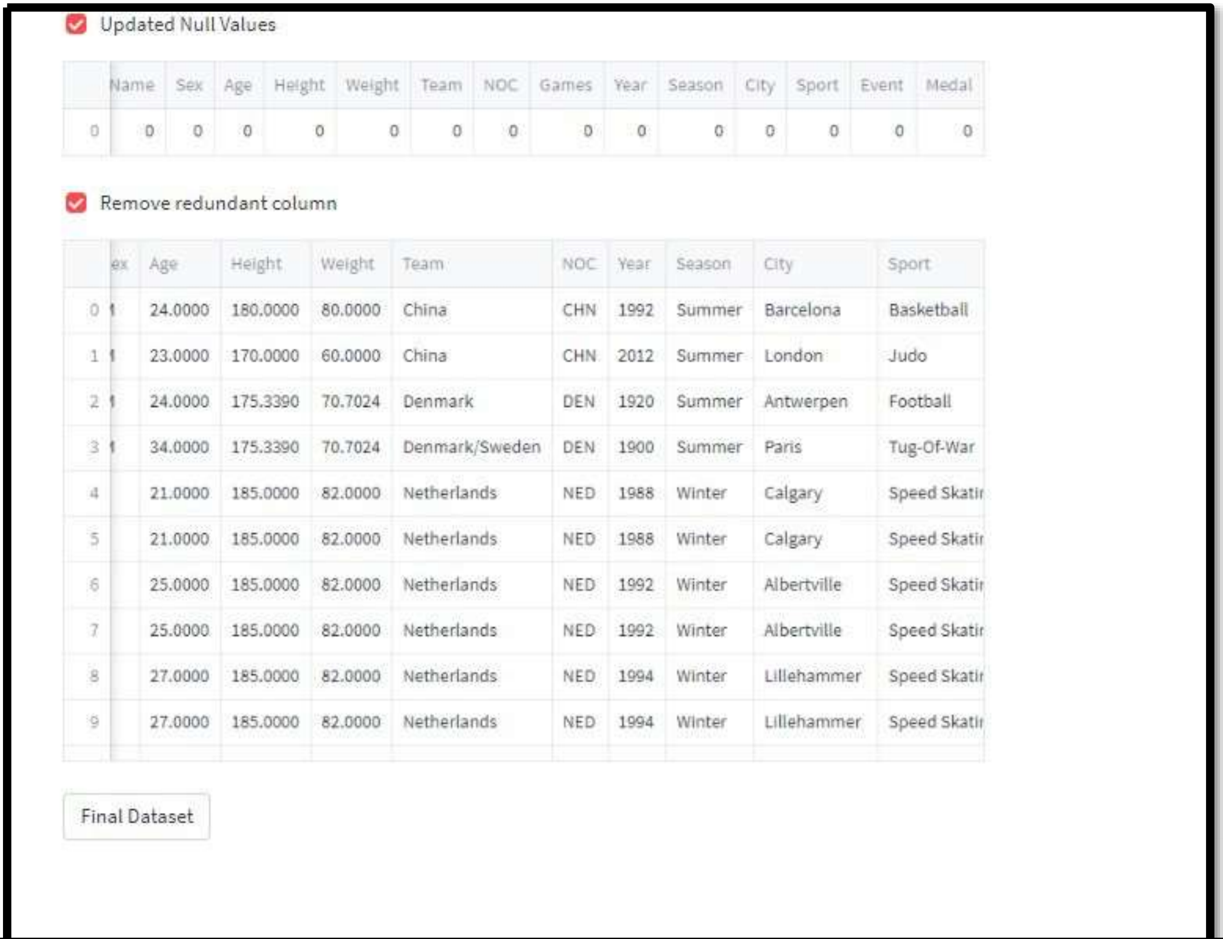
☐ Remove Null Values in Weight Column

☑ Covert Medals to Numeric datatype and Remove Null Values

	mes	Year	Season	City	Sport	Event	Medal
0	92 Summer	1992	Summer	Barcelona	Basketball	Basketball Men's Basketball	0.0000
1	12 Summer	2012	Summer	London	Judo	Judo Men's Extra-Lightweight	0.0000
2	20 Summer	1920	Summer	Antwerpen	Football	Football Men's Football	0.0000
3	00 Summer	1900	Summer	Paris	Tug-Of-War	Tug-Of-War Men's Tug-Of-War	1.0000
4	88 Winter	1988	Winter	Calgary	Speed Skating	Speed Skating Women's 500 metres	0.0000
5	88 Winter	1988	Winter	Calgary	Speed Skating	Speed Skating Women's 1,000 metres	0.0000
6	92 Winter	1992	Winter	Albertville	Speed Skating	Speed Skating Women's 500 metres	0.0000
7	92 Winter	1992	Winter	Albertville	Speed Skating	Speed Skating Women's 1,000 metres	0.0000
8	94 Winter	1994	Winter	Lillehammer	Speed Skating	Speed Skating Women's 500 metres	0.0000
9	94 Winter	1994	Winter	Lillehammer	Speed Skating	Speed Skating Women's 1,000 metres	0.0000

☐ Updated Null Values

☐ Remove redundant column



☑ Updated Null Values

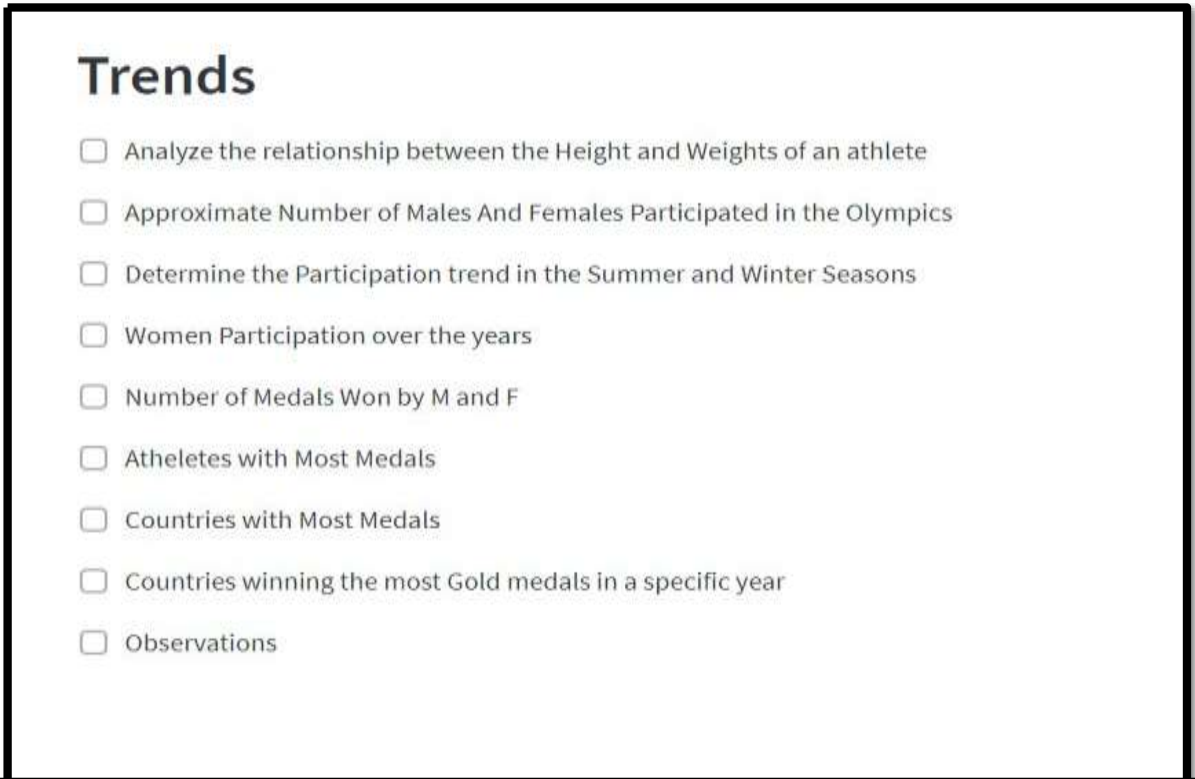
	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season	City	Sport	Event	Medal
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

☑ Remove redundant column

	ex	Age	Height	Weight	Team	NOC	Year	Season	City	Sport
0	1	24.0000	180.0000	80.0000	China	CHN	1992	Summer	Barcelona	Basketball
1	1	23.0000	170.0000	60.0000	China	CHN	2012	Summer	London	Judo
2	1	24.0000	175.3390	70.7024	Denmark	DEN	1920	Summer	Antwerpen	Football
3	1	34.0000	175.3390	70.7024	Denmark/Sweden	DEN	1900	Summer	Paris	Tug-Of-War
4		21.0000	185.0000	82.0000	Netherlands	NED	1988	Winter	Calgary	Speed Skatir
5		21.0000	185.0000	82.0000	Netherlands	NED	1988	Winter	Calgary	Speed Skatir
6		25.0000	185.0000	82.0000	Netherlands	NED	1992	Winter	Albertville	Speed Skatir
7		25.0000	185.0000	82.0000	Netherlands	NED	1992	Winter	Albertville	Speed Skatir
8		27.0000	185.0000	82.0000	Netherlands	NED	1994	Winter	Lillehammer	Speed Skatir
9		27.0000	185.0000	82.0000	Netherlands	NED	1994	Winter	Lillehammer	Speed Skatir

- **Trends:**

Contains list of options to show the relationship between attributes.



Trends

□ Analyze the relationship between the Height and Weights of an athlete

□ Approximate Number of Males And Females Participated in the Olympics

□ Determine the Participation trend in the Summer and Winter Seasons

□ Women Participation over the years

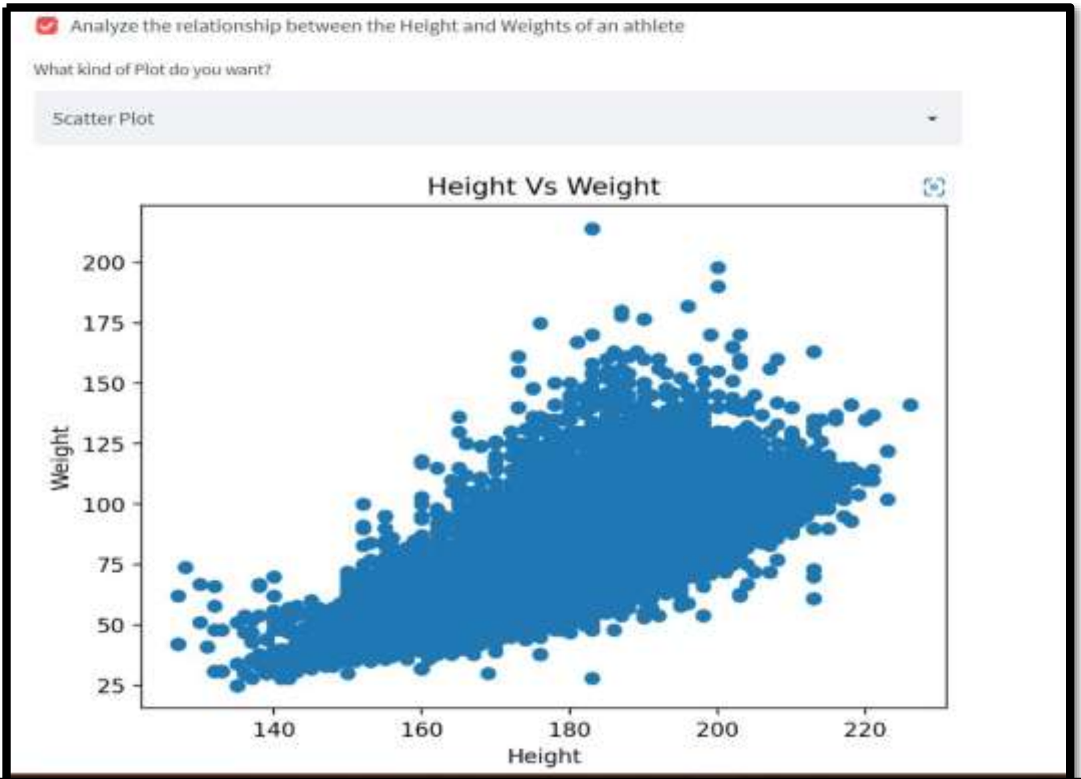
□ Number of Medals Won by M and F

□ Atheletes with Most Medals

□ Countries with Most Medals

□ Countries winning the most Gold medals in a specific year

□ Observations

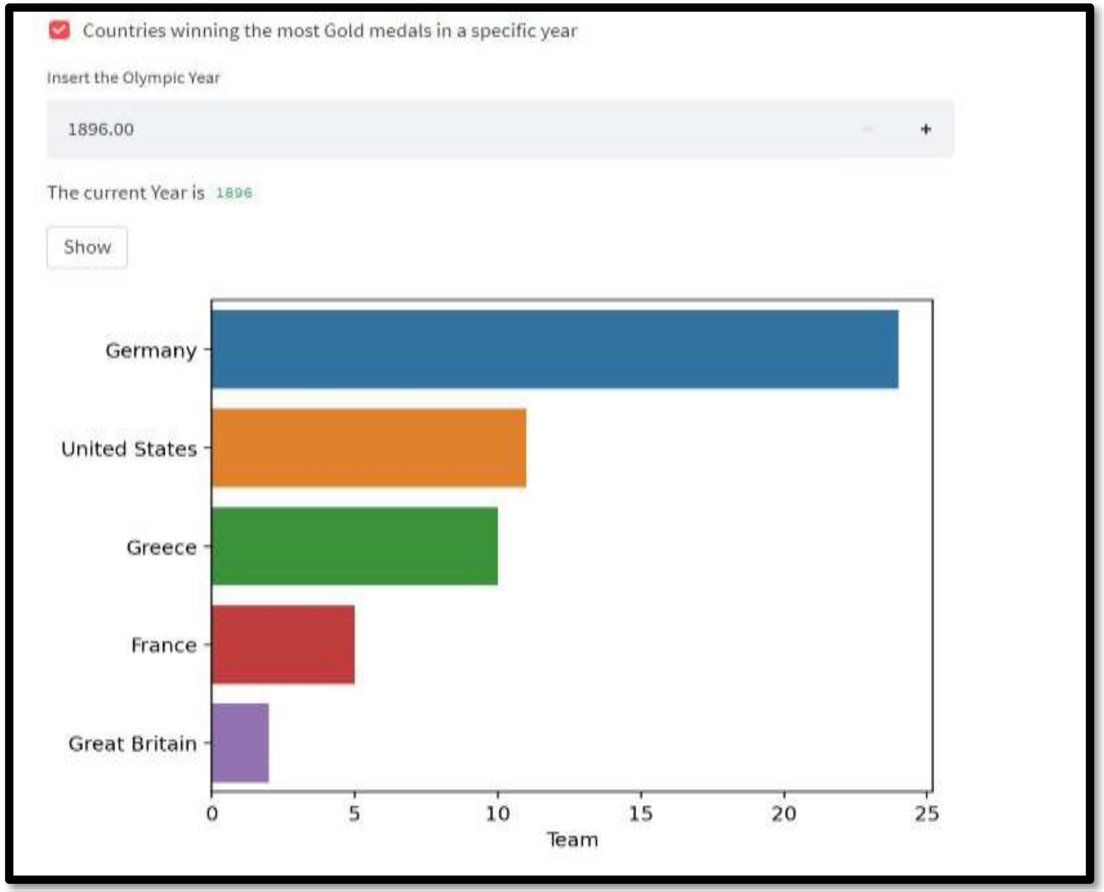


☑ Analyze the relationship between the Height and Weights of an athlete

What kind of Plot do you want?

Scatter Plot

Height Vs Weight



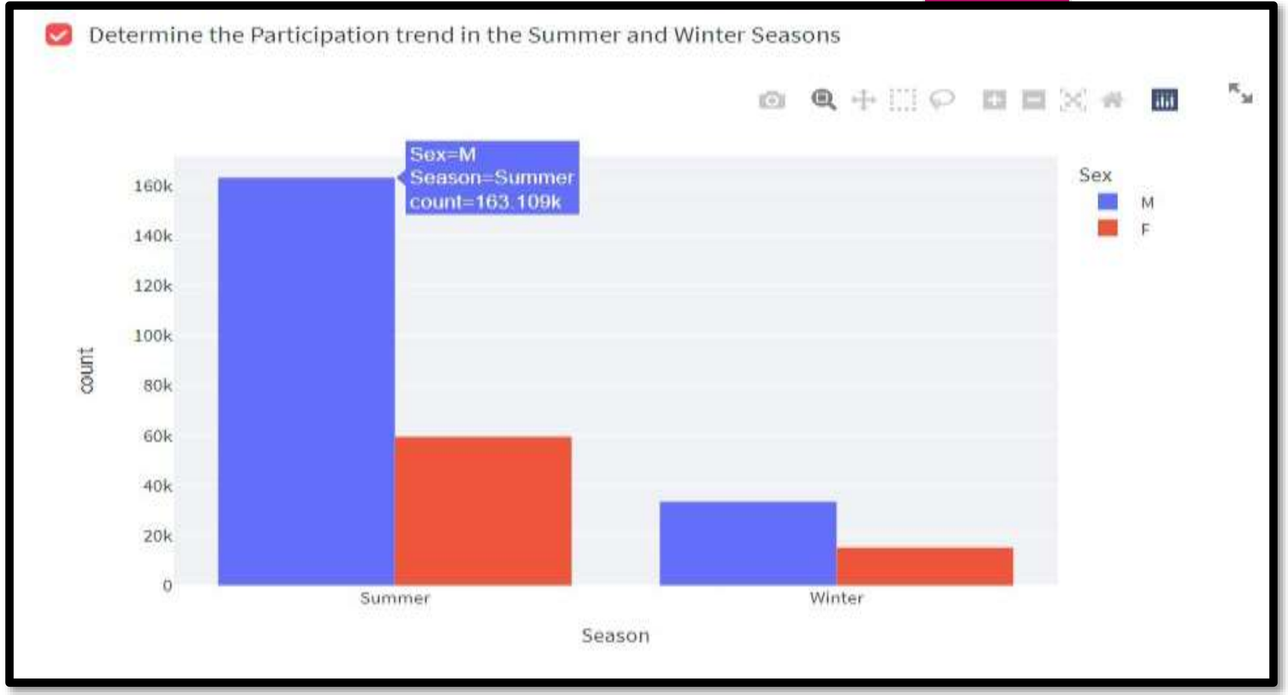
☑ Countries winning the most Gold medals in a specific year

Insert the Olympic Year

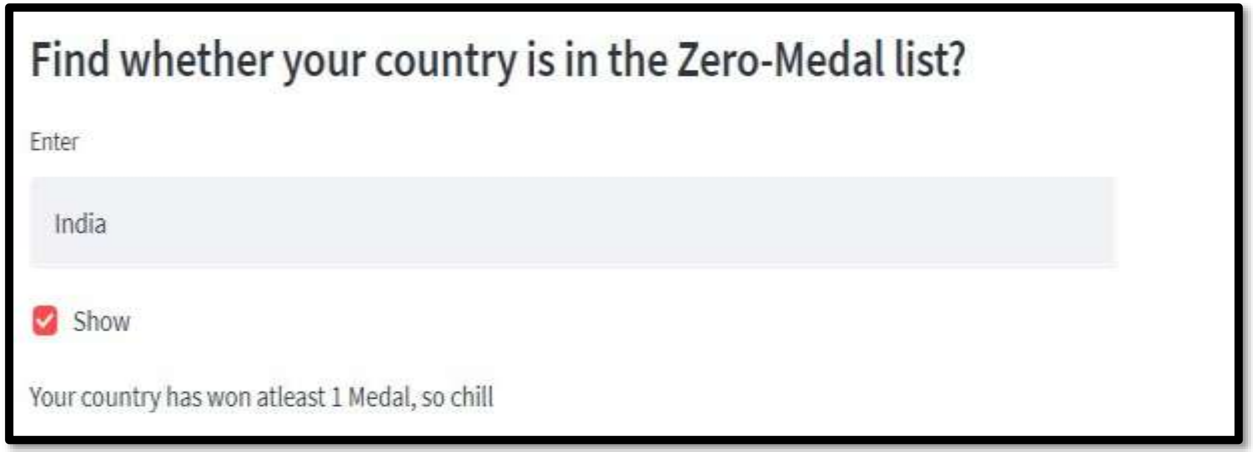
1896.00 − +

The current Year is `1896`

Show



☑ Determine the Participation trend in the Summer and Winter Seasons



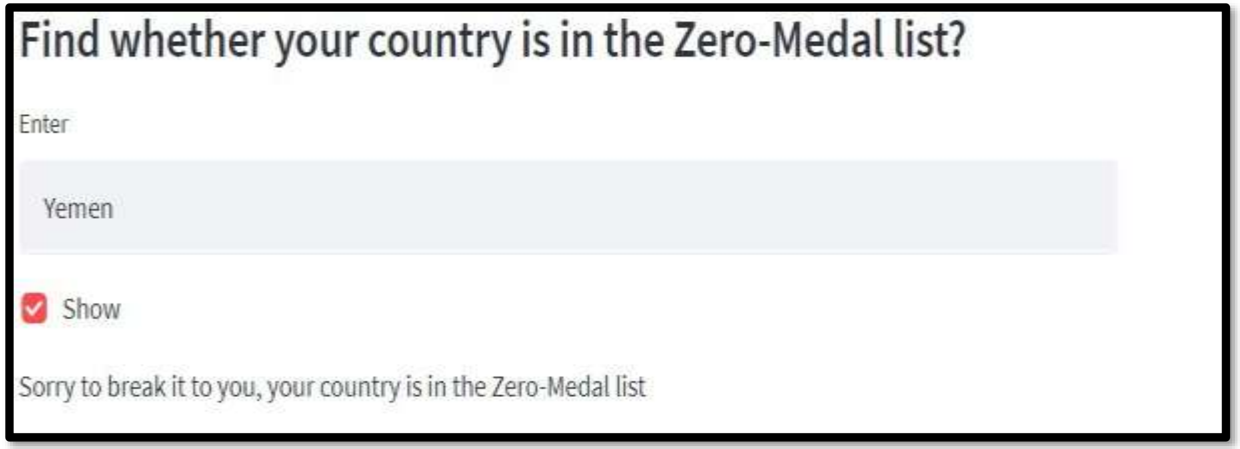
Find whether your country is in the Zero-Medal list?

Enter

India

☑ Show

Your country has won atleast 1 Medal, so chill



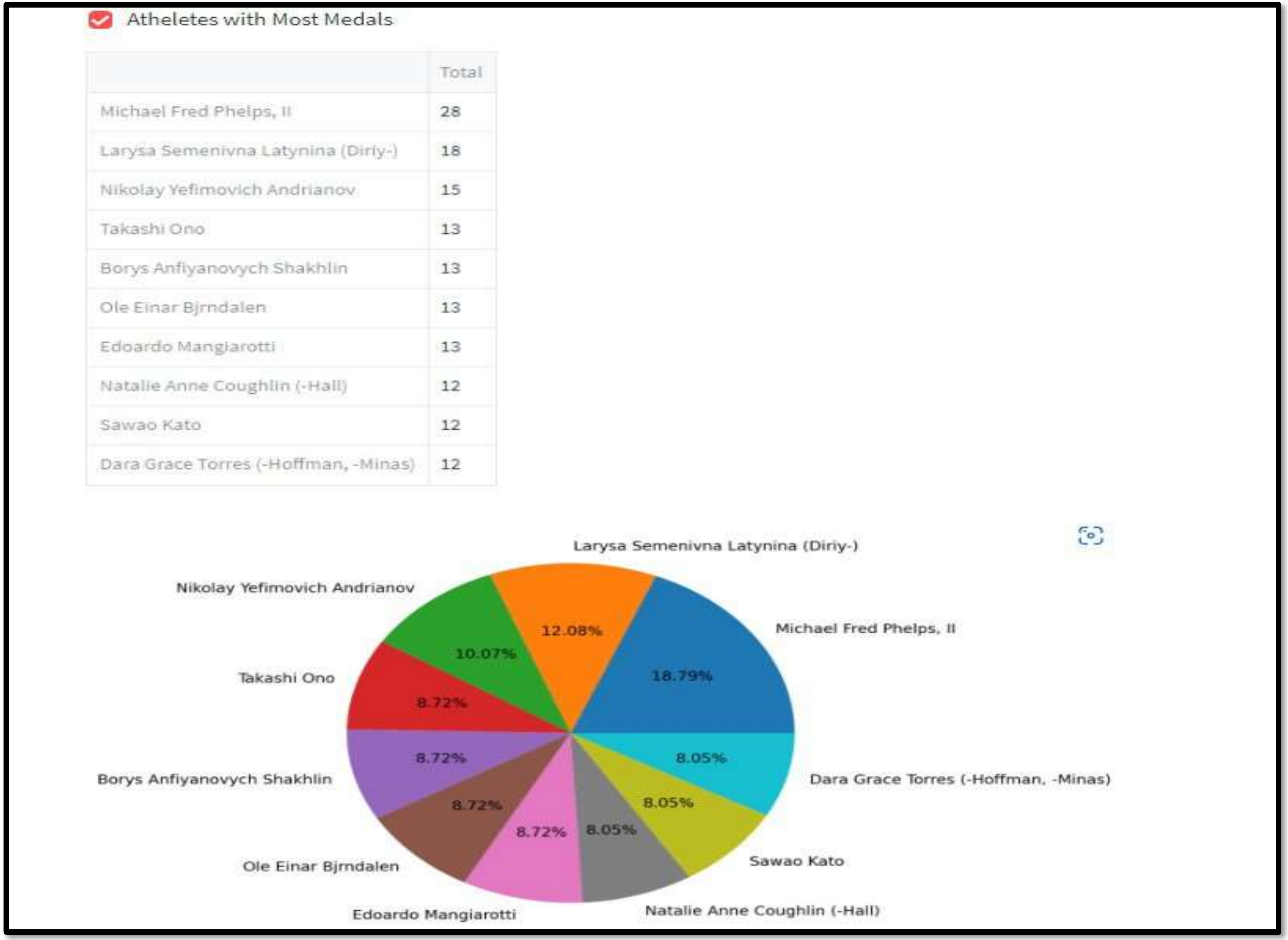
Find whether your country is in the Zero-Medal list?

Enter

Yemen

☑ Show

Sorry to break it to you, your country is in the Zero-Medal list



☑ Atheletes with Most Medals

	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

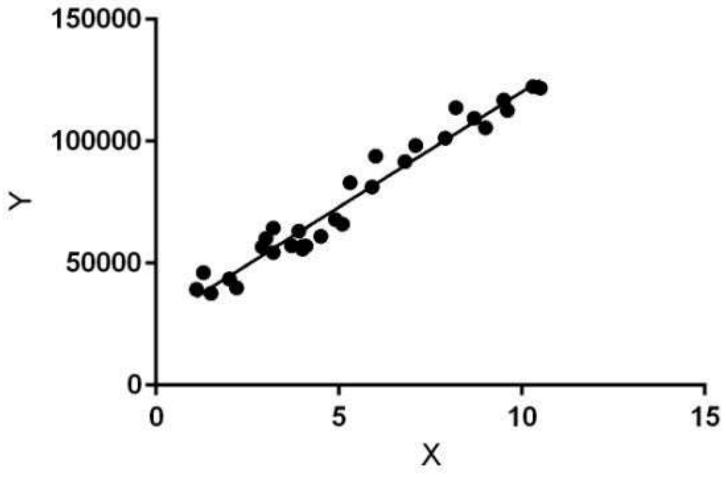
Larysa Semenivna Latynina (Diriy-)
12.08%
Nikolay Yefimovich Andrianov
10.07%
Michael Fred Phelps, II
18.79%
Takashi Ono
8.72%
Borys Anfiyanovych Shakhlin
8.72%
Dara Grace Torres (-Hoffman, -Minas)
8.05%
8.72%
8.05%
8.72%
8.05%
Sawao Kato
Ole Einar Bjrndalen
Edoardo Mangiarotti
Natalie Anne Coughlin (-Hall)

LINEAR REGRESSION

- Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task.
- Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x).
- The regression line is the best fit line for our model.
- Hypothesis function for Linear Regression :

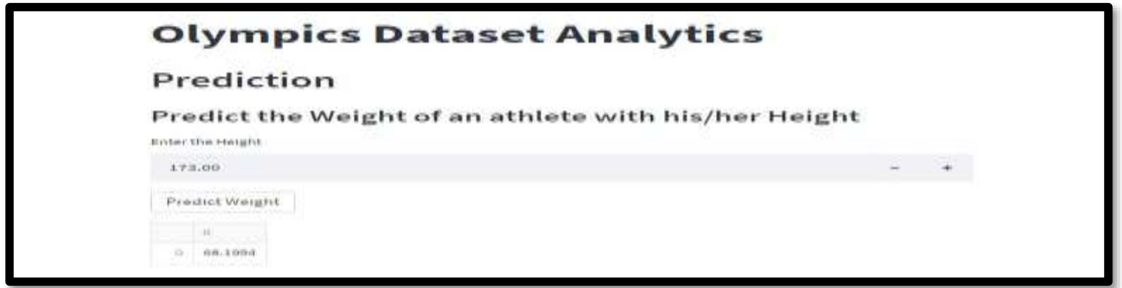
$$y = \theta_1 + \theta_2.x$$

- x: input training data (univariate – one input variable(parameter))
- y: labels to data (supervised learning)
- θ1: intercept
- θ2: coefficient of x



- **Prediction:**

- Contains two prediction options: one for weight and other for sport.
- Accuracy in predicting weight is 62%.
- Accuracy in predicting sport is 89.8%.



Olympics Dataset Analytics

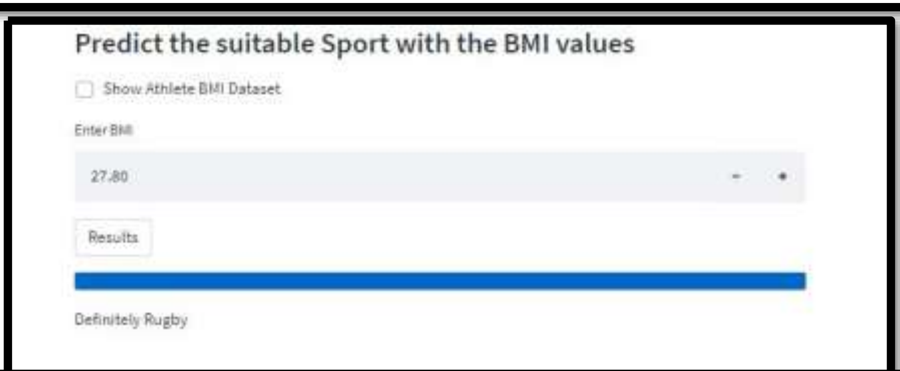
Prediction

Predict the Weight of an athlete with his/her Height

Enter the Height

173.00

	0
0	68.1094



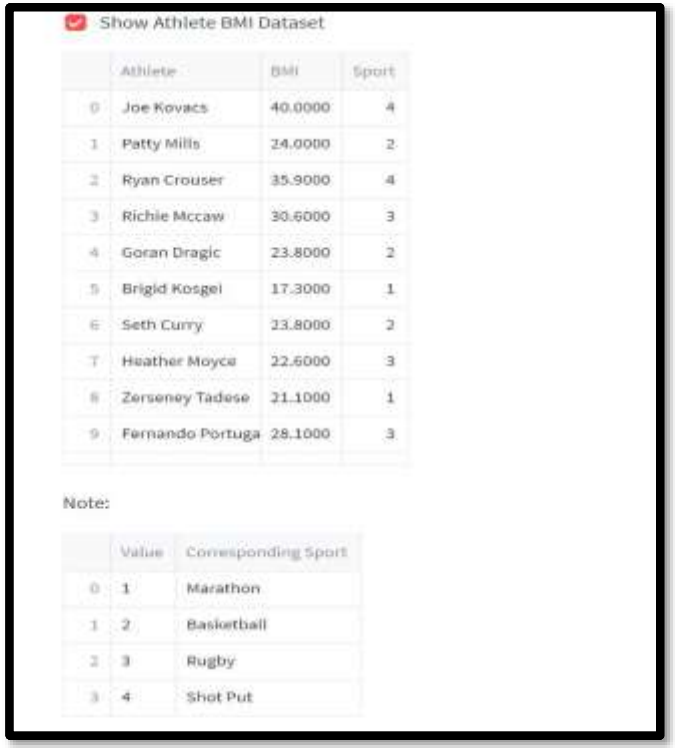
Predict the suitable Sport with the BMI values

□ Show Athlete BMI Dataset

Enter BMI

27.80

Definitely Rugby



☑ Show Athlete BMI Dataset

	Athlete	BMI	Sport
0	Joe Kovacs	40.0000	4
1	Patty Mills	24.0000	2
2	Ryan Crouser	35.9000	4
3	Richie Mccaw	30.6000	3
4	Goran Dragic	23.8000	2
5	Brigid Kosgei	17.3000	1
6	Seth Curry	23.8000	2
7	Heather Moyce	22.6000	3
8	Zerseney Tadese	21.1000	1
9	Fernando Portuga	28.1000	3

Note:

	Value	Corresponding Sport
0	1	Marathon
1	2	Basketball
2	3	Rugby
3	4	Shot Put

APPLICATIONS / USAGE

The olympics dataset analytics application/ interface can be helpful for:

- The managing authorities of olympics.
- To those who are interested in knowing about olympics.
- Can be useful as an easy interface.
- Can lay foundation to build much more interactive Data Analytics applications.

CONCLUSION

We’re able to build an interactive application that’d perform analytical operations on the olympic datasets, and fetch results in an easy and appealing manner. Once the user clicks on load datasets option, he/she presented with a flowchart that guides them in a wonderful data analytic journey. Therefore, we conclude that the interface/application has been built successfully.

FUTURE SCOPE

The scope of project can be extend to, but not limited to:

- Options to load more than two datasets.
- Options to contribute to the datasets.
- More options at each stage of analytics.
- Improving the accuracy of Prediction

REFERENCES

1. http://www.researchgate.net/publication/330847008_Performance_analysis_in_olympic_games_using_exploratory_data_analysis_techniques
2. https://www.researchgate.net/publication/265033380_Data_mining_of_sports_performance_data
3. https://www.researchgate.net/publication/23756788_Economics_and_Olympics_An_Efficiency_Analysis
4. https://ieeexplore.ieee.org/abstract/document/9725496
5. https://docs.streamlit.io/

THANK YOU