STATISTICS FOR DATA SCIENCE (UE19CS203)

PROJECT REPORT

Title: 120 years of Olympic history

Team details:

Sl No	Name	SRN	Section
1	Sanjana S	PES2UG19CS363	F
2	Sanjana S Murthy	PES2UG19CS364	F
3	Sohan Naidu	PES2UG19CS398	F
4	Soundarya K	PES2UG19CS402	F

1. Abstract

The Olympics is an international sporting event. Participation in the event has expanded from 241 athletes to 11,500 since the last Olympics. Given the historical data throughout the Olympics, the odds of winning a medal (gold, silver, or bronze) could perhaps be given based on a few biological attributes of the athletes.

2. Introduction

The Olympic Games have been expanding every year which can be seen by the records of the nations participating. The number has grown from 14 nations in 1896 in Athens to 207 nations in 2016 at the Rio Olympics. This international sporting event where thousands of athletes from various countries compete in various sports every four years, has experienced enough growth in which we can begin to ask questions on the evolution of the Olympics based on gender participation or their performance and results based on basic biological information.

Therefore, we decided to do exploratory data analysis so we may visualize patterns within the dataset. Furthermore, we wanted to predict if an athlete would win a medal based on those few attributes given.

3. Dataset

The dataset was taken from https://www.kaggle.com/heesoo37/120-years-of-olympic-history-athletes-and-results. The dataset provided consists of 271,116 unique athletes with 15 attributes.

1. ID - Unique number for each athlete
2. Name - Athlete's name
3. Sex - M or F
4. Age - Integer
5. Height - In centimeters
6. Weight - In kilograms
7. Team - Team name
8. NOC - National Olympic Committee 3-letter code
9. Games - Year and season
10. Year - Integer
11. Season - Summer or Winter
12. City - Host city
13. Sport - Sport
14. Event - Event
15. Medal - Gold, Silver, Bronze, or NA

The collection includes all games from Athens 1896 to Rio 2016. Another file called "noc_regions.csv" was provided as well, however, we made the decision to drop the file.

4. Preprocessing or Data Cleaning

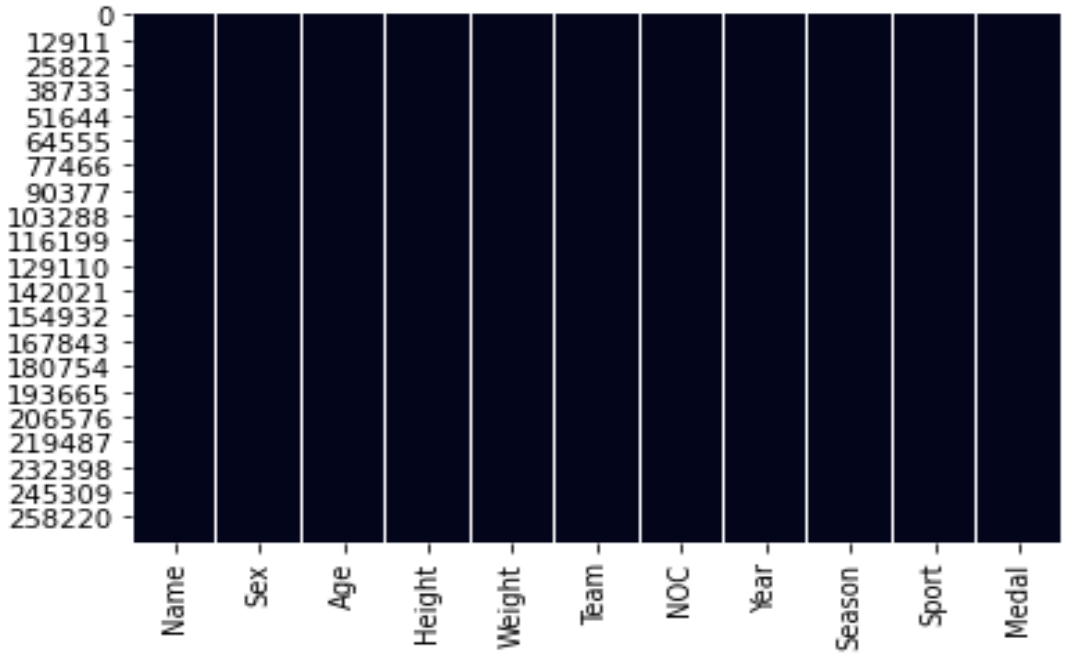
We did preprocessing of the data by selecting attributes we deemed relevant such as: Name, Sex, Age, Weight, Height, Team, NOC, Year, Season, Sport and Medal. We made the decision to remove ID, Games, City and Event. These were removed based on the idea that personal identifying information would not be useful in many predictions or data analysis.

```
#Dropping the passed columns from the data frame
data.drop(["ID","Games","City","Event"],axis=1,inplace=True)
#Displaying the Data frame
data
```

	Name	Sex	Age	Height	Weight	Team	NOC	Year	Season	Sport	Medal
0	A Dijiang	M	24.0	180.0	80.0	China	CHN	1992	Summer	Basketball	NaN
1	A Lamusi	M	23.0	170.0	60.0	China	CHN	2012	Summer	Judo	NaN
2	Gunnar Nielsen Aaby	M	24.0	NaN	NaN	Denmark	DEN	1920	Summer	Football	NaN
3	Edgar Lindenau Aabye	M	34.0	NaN	NaN	Denmark/Sweden	DEN	1900	Summer	Tug-Of-War	Gold
4	Christine Jacoba Aaftink	F	21.0	185.0	82.0	Netherlands	NED	1988	Winter	Speed Skating	NaN
...	...	...	...	...	...	...	...	...	...	...	...
271111	Andrzej ya	M	29.0	179.0	89.0	Poland-1	POL	1976	Winter	Luge	NaN
271112	Piotr ya	M	27.0	176.0	59.0	Poland	POL	2014	Winter	Ski Jumping	NaN
271113	Piotr ya	M	27.0	176.0	59.0	Poland	POL	2014	Winter	Ski Jumping	NaN
271114	Tomasz Ireneusz ya	M	30.0	185.0	96.0	Poland	POL	1998	Winter	Bobsleigh	NaN
271115	Tomasz Ireneusz ya	M	34.0	185.0	96.0	Poland	POL	2002	Winter	Bobsleigh	NaN

The dataset came with null values that had to be resolved. We identified them by checking existing null values for each column within the dataset. Our results were as shown in the table.

The reason for medal column returned so many null values was because of the dataset had the tags gold, silver and bronze medalists and null tag for non-medalists. The decision was made to give non-medalists the “NOMEDAL” string value to make further data analysis easier.
This picture below depicts visualization of null values after imputation.



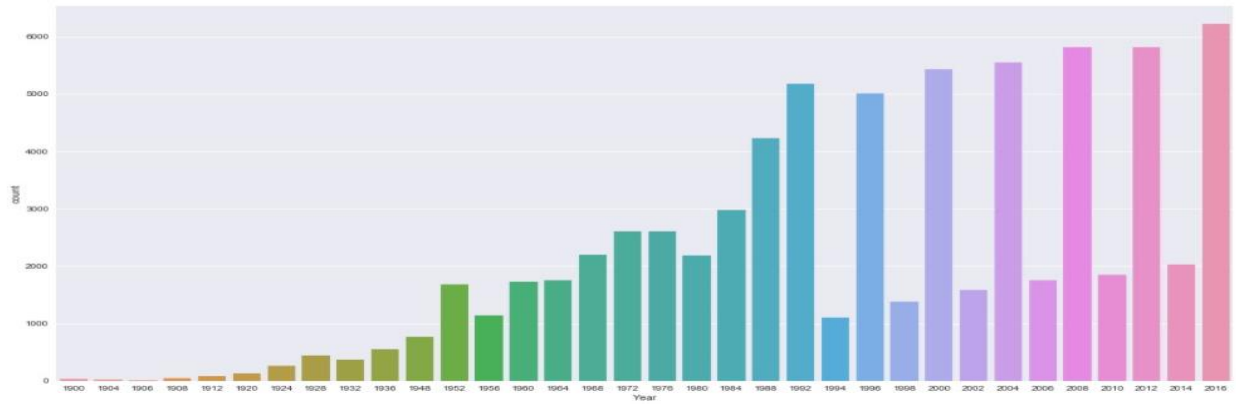
5. Exploratory Data Analysis

We explored the data and wanted to find the participation of women, men, India in Olympics over the years. We plotted the data in histogram.

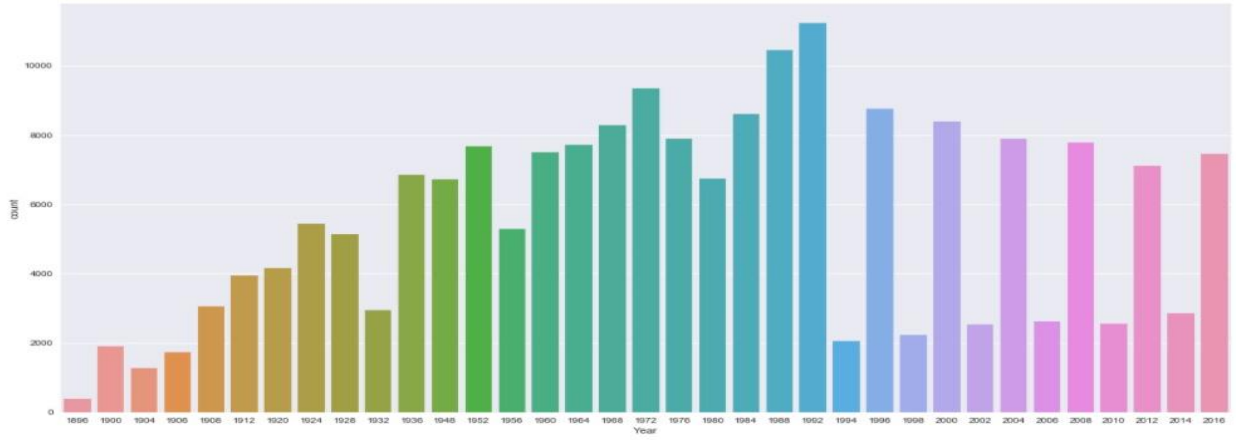
We also standardized and normalized the data related to height, weight and age of the athletes.

```
Average height of participants in the Olympics is :  175.33896987366376
Average weight of participants in the Olympics is :  70.70239290053351
Average age of participants in the Olympics is :  25.556898357297374

Standard Deviation in height of participants in the Olympics is :  10.518462222679224
Standard Deviation in weight of participants in the Olympics is :  14.348019999019392
Standard Deviation in age of participants in the Olympics is :  6.393560847035813

Variance in height of participants in the Olympics is :  175.33896987366376
Variance in weight of participants in the Olympics is :  205.86567789226046
Variance in age of participants in the Olympics is :  40.87762030474931
```

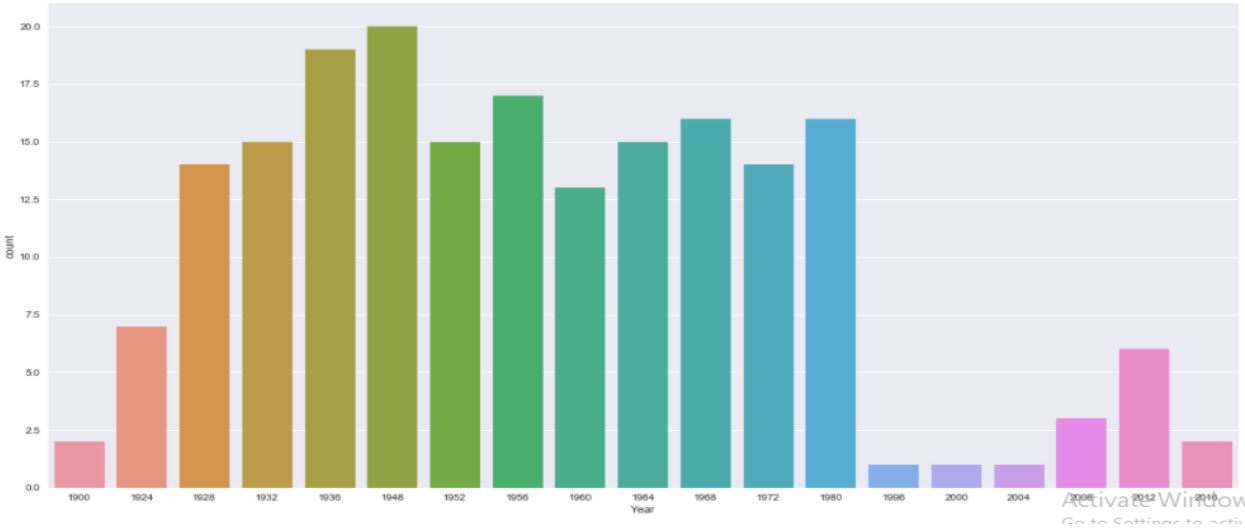
Participation of women



Participation of men



Medals secured by Indian Athletes

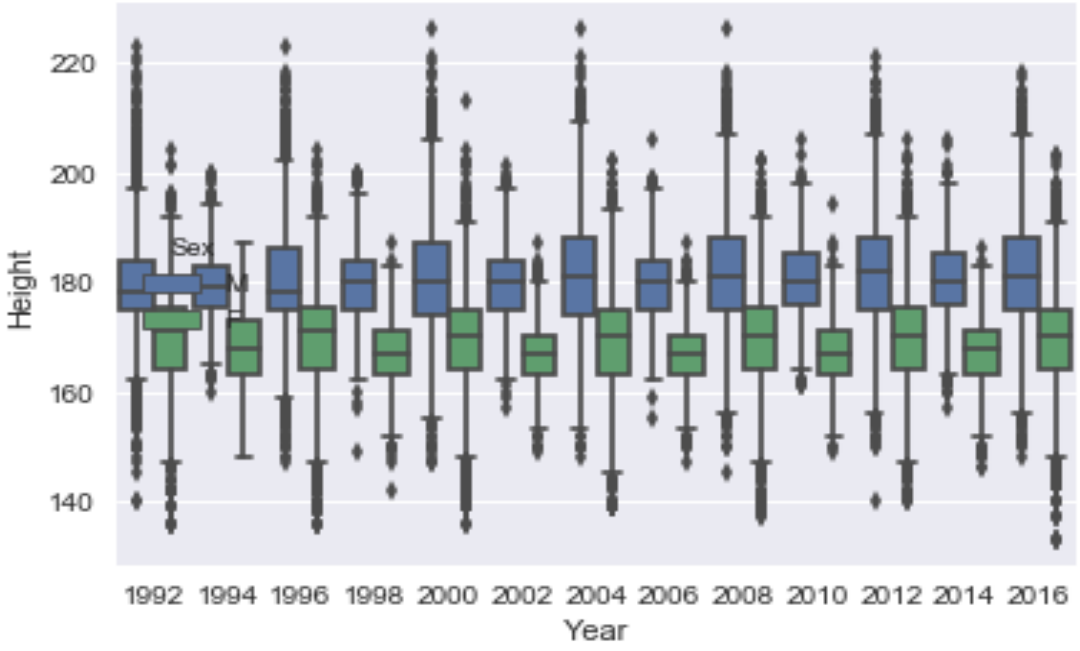


We wanted to view the highest number of medals secured by the countries that participated in the Olympics held during the different seasons.

```
Country that has won the highest number of gold medals in Summer Olympics: Zimbabwe
Country that has won the highest number of gold medals in Winter Olympics: West Germany-2
Country that has won the highest number of silver medals in Summer Olympics: Zut
Country that has won the highest number of silver medals in Winter Olympics: Yugoslavia
Country that has won the highest number of bronze medals in Summer Olympics: Zimbabwe
Country that has won the highest number of bronze medals in Winter Olympics: Yugoslavia
```

We also wanted to view the distribution through their quartiles with the height attribute between the men and women. As seen on the box and whisker plot, the average height is focused starting from the year 1992.

We viewed the distribution through their quartiles with the weight attribute between the men and women. As seen on the box and whisker plot, the average height is focused starting from the year 1992



.

6. Hypothesis Testing

In our hypothesis testing, we took random samples such as percentage of ‘male basketball players’, ‘female cyclists’ among all participants, fewest and most number of participants, age of youngest male and female who participated in Olympics of 2000. Here are the results we obtained

```
Percentage of male basketball players among all male participants in 2000 was: 2.2%
Percentage of female cyclists in 2000 was: 3.0%
In 2000 Olympics, age of youngest male was 14.0 and age of youngest female was 13.0
```

From these results, we tried to derive conclusions about the population. We understood that the percentage of male basketball players among all male participants remained consistent as we got a result of 2.5% from the population.

The average percentage of female cyclists, however, was 2.1% for the overall population.

Similarly, the youngest male and female who ever took part in the Olympics was 10 and 11 respectively which is around the value we got from the sample.

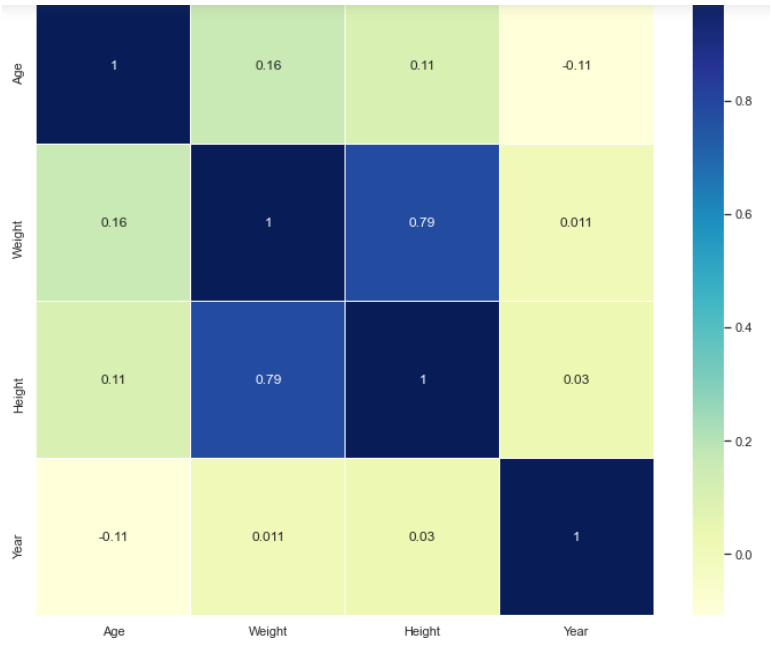
```
Percentage of male basketball players among all male participants in all of olymics were: 2.5%

Percentage of female cyclists among all female participants in all of olymics were: 2.1%

In the entire 120 yrs of olympics, age of youngest male was 10.0 and age of youngest female was 11.0
```

So, we can conclude that the participation of men and woman in respective sports and the age category are consistent over the years. That is, not too much of a rise or drop is observed.

7. Results and Discussion

We believe Height and Weight played a vital role as well so we wanted to see if there was a trend that existed within our data.

Correlation matrix is a table showing correlation coefficients between variables. In our respective dataset, we have taken age, year, height and year as the variables. We wanted to use correlation matrix to summarize the large amount of data in pattern. The observable pattern in our dataset is that all the variables highly correlate with each other.



Various useful data visualization and machine learning libraries were used. The knowledge learned during this project will be incredibly useful later on. There can be many different combinations of features to be used in this project that may give better predictions.