**SDS PROJECT REPORT**

**Title:** Performance of Women in Olympics

**Members’ Details:**

1. Gaurav J ‘C’ 2. Sai Manoj ‘C’

PES1UG19CS599 PES1UG19CS155

3. Gitika Jain ‘C’ 4. Guruprasad N Bamane ‘C’

PES1UG19CS169 PES1UG19CS174

* **Abstract:**

Our goal as a team was to perform some analysis on the female participants in sports. How many athletes, sports, and nations are there? Where do most athletes come from? Who wins medals? What are the characteristics of the athletes who win medals? And so on.

So, we decided to look at the Olympic games data as it has been held for over 100 years and is well recorded. This allowed us to shed light on major progress in the participation of women insights in Olympics through various analysis on the data such as correlations, numerical patterns, verification of hypotheses etc.

* **Introduction:**

The rate of participation of women in the Olympics has been increasing since their first participation in 1900. Some sports are uniquely for women, others are contested by both sexes, while some older sports remain for men only. Studies of media coverage of the Olympics consistently show differences in the ways in which women and men are described and the ways in which their performances are discussed. So here we are to analyse the participation of women in Olympics in a detailed manner and visualize using graphs.

* We need to analyse and visualise how each characteristic of an athlete affects the athlete’s performance. We also determine the correlation of different characteristics of an athlete. The normalisation of data is an important aspect. Normal Probability plot is plotted for each numeric data column by considering the sample dataset which contains details of women from a country with highest women participation so that plotting would be more efficient and simpler.
* **Dataset:**

The dataset chosen is a historical dataset on the modern Olympic Games, including all the Games from Athens 1896 to Rio 2016. The dataset was taken from [www.kaggle.com](http://www.kaggle.com/) which was originally scraped from [www.sports-reference.com](http://www.sports-reference.com/). We have to note that the Winter and Summer Games were held in the same year up until 1992. After that, they staggered them such that Winter Games occur on a four year cycle starting with 1994, then Summer in 1996, then Winter in 1998, and so on. A common mistake we as people might make when analysing this data is to assume that the Summer and Winter Games have always been staggered.

**Content:**

The dataset contains 271116 rows and 15 columns. Each row corresponds to an individual athlete competing in an individual Olympic event (athlete-events). The columns are:

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

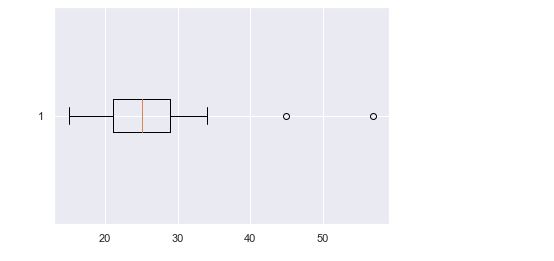
**Inspiration:**

This dataset provides an opportunity to ask questions about how the Olympics have evolved over time, including questions about the participation and performance of women, different nations, and different sports and events.

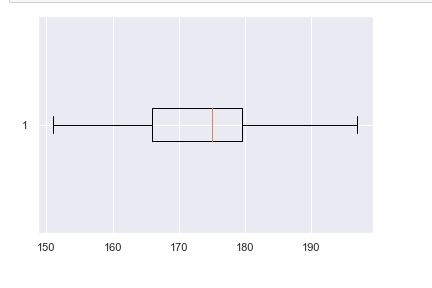
* **Pre-processing or Data Cleaning:**

Data cleaning or cleansing is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database. The main columns which we try to handle here are ‘Age’, ‘Height’ and ‘Weight’. Since they are all numeric columns, we use a very efficient technique, box plot to determine any outliers. First, we use the describe function for each column to know the mean, median, etc. Then, with the help of box plot we replace the missing values from a column accordingly with mean or median.

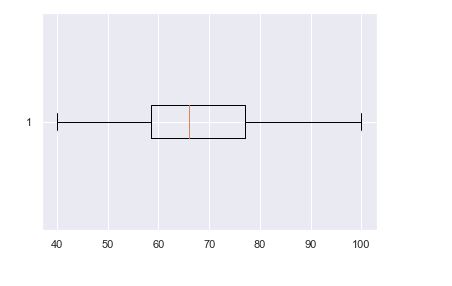
**Box plot for Age:** As we can see from the below boxplot, the plot is not symmetric and has outliers, so we replace the missing values with median of the age column.



**Box plot for Height:** As we can see from the below boxplot, the plot is not symmetric, so we replace the missing values with median of the height column.



**Box plot for Weight:** As we can see from the below boxplot, the plot is symmetric, so we replace the missing values with mean of the weight column.



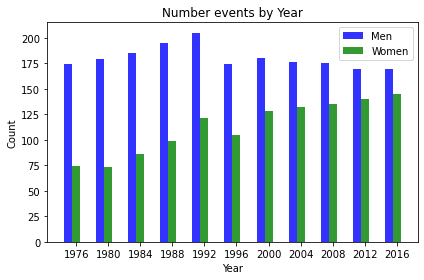
Note: Cleaning of categorical data (Medals) is done in the later part.

* **Exploratory Data Analysis:**

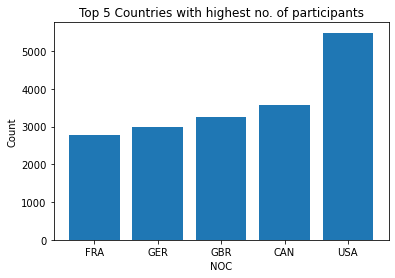
Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations. Our main aim is to analyse the participation of women in sports events.

We create a sub dataset which contains the details of female participants alone. Now, the analysis of this dataset should be done. We have to create another sub dataset which contains the details of male participants. Now, we have to compare

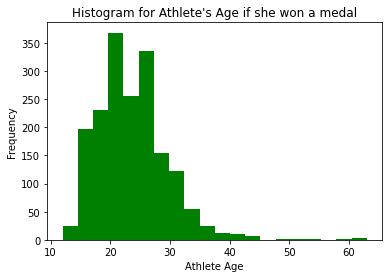
the gender subsets to compare the number of events hosted for male and female participants respectively. The below graph is plotted:



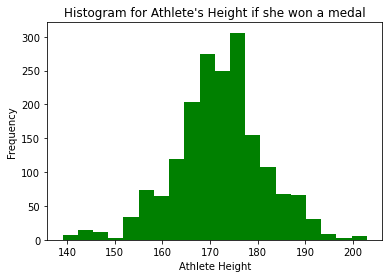
Next, we plot a graph to depict the top 5 countries with highest number of women participants:



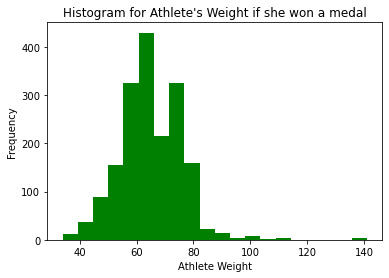
From the above graph we can infer that the USA has the highest number of women participants. Next, we analyse the number of medals won by participants with age, height and weight as parameters. We consider the female participants from USA alone due to the simplicity factor because most women participants are from The USA. The below graphs are plotted:



Since, the above graph is right skewed, we can infer that the most winning athletes are quite young.



Since, the above graph is almost symmetric, we can infer that the ideal height for winning is around 165-180 centimetres.

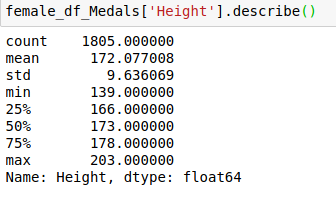


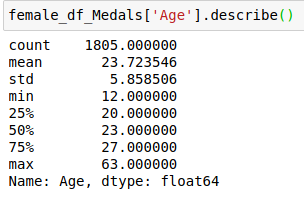
The above graph is right skewed. Therefore, the ideal weight for winning medals is around 55-75 kgs.

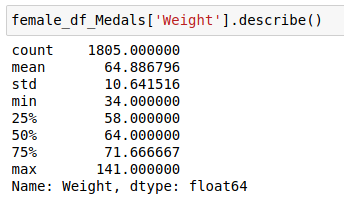
* **Normalization**

“Normalizing” a vector most often means dividing by a norm of the vector. It also often refers to rescaling by the minimum and range of the vector, to make all the elements lie between 0 and 1 thus bringing all the values of numeric columns in the dataset to a common scale.

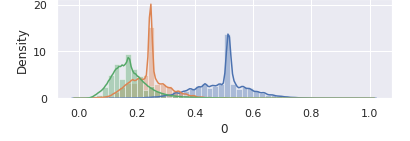
Before normalization

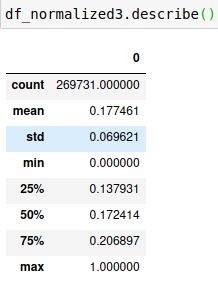
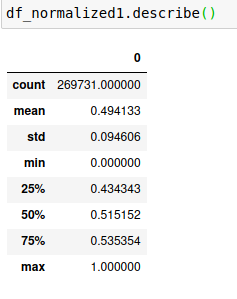
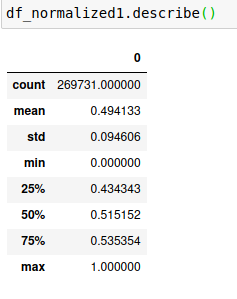






After normalization:





* Standardization

A vector most often means subtracting a measure of location and dividing by a measure of scale. For example, if the vector contains random values with a Gaussian distribution, you might subtract the mean and divide by the standard deviation, thereby obtaining a “standard normal” random variable with mean 0 and standard deviation 1.

Why do we need to standardize?

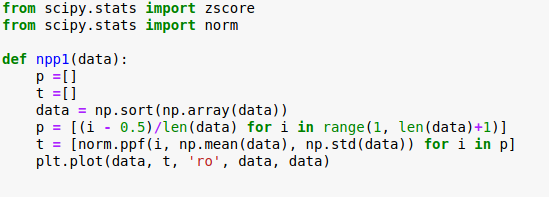
We need to standardize our data, so it is easy to calculate the probabilities, from the standard normal probability table.

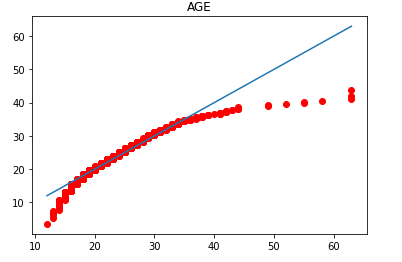
And we don’t have to integrate every time.

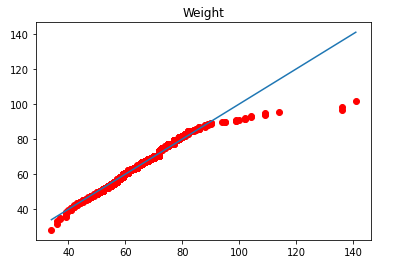
****

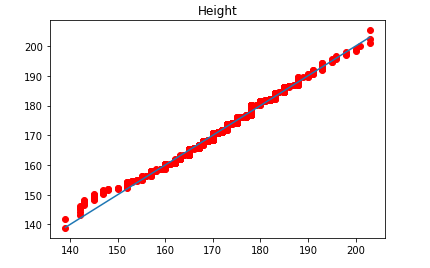
**To check whether the data is normal:**

We use normal probability plot









So we see that only Height distribution of the Female participants is approximately normally distributed.

* **Hypothesis Testing**

We have assumed our hypothesis test from something like a case study,

Many times, in the news or media we have seen reports like, for example:

“the Russian female team have prepared strong for upcoming olympic games The mean age of the squad is just 21,

they claim that as that average age of Olympic medallists in every Olympic event over the years have been less than 23. So, the Russians are the favourites”

So now we test this claim:

Ho: u<=23

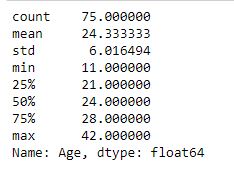
Ha: u>23

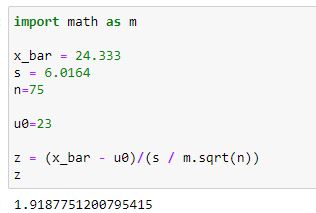
here u0=23

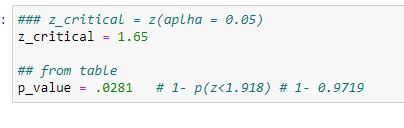
We take a random sample of size 500 from our data frame of female\_df (female participants over the years)

Now from this sample we select only those women who have gotten a medal









Let us assume 5% significance rule

z\_critical = 1.65

Therefore, as P-value<= α

### So, we reject Ho,

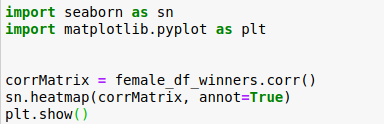
We conclude that we have evidence against u<=23 (for winning female participants),

therefore u>23.

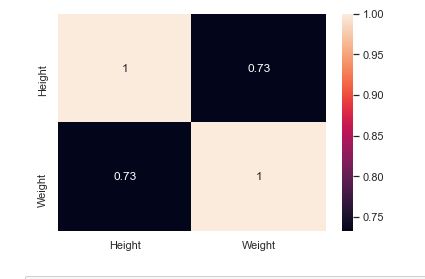
* **Correlation**

We are going to see how correlated the Height and Weight variables are

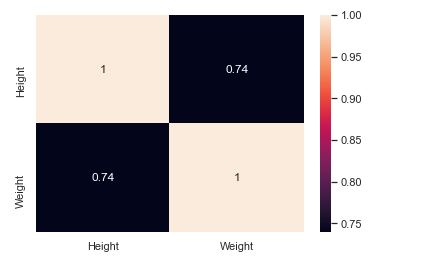
1. in case of male participants who have won a medal
2. in case of female participants who have won a medal



1.in case of women winners, we see that height and weight are positively correlated



2.in case of men winners, we see that height and weight are positively correlated



* **Result:**

1. We conclude the number of events being held for women have been increasing gradually over the years.
2. We found out the top 5 countries with highest number of female participants.
3. Irrespective of the events, we have found out the ideal age, height and weight of a woman to win a medal.
4. We have conducted a hypothesis regarding the mean age of a woman to win a medal.
5. We also found out that the height and weight of an athlete are positively correlated.

\* \* \*