Report: Data Analysis on Olympic Winter Games

Report Prepared By:

Mouli Banerjee

Project Engineer, Wipro Limited, India

Email: banerjeemouli932@gmail.com

Table of Content

1. Introduction 1

2. Proposed Approach 2

2.1 Data Source 2

2.2 Collection of Data 3

2.3 Processing of Data 3

2.4 Cleaning of Dataset 4

2.5 Exploratory Data Analysis 4

3. Analysis Setup and Result 5

3.1 Setup 5

3.1.1 Pandas 5

3.1.2 NumPy 5

3.1.3 Matplotlib 6

3.1.4 Seaborn 6

3.2 Result Analysis 6

3.2.1 Descriptive Analysis of the Data 6

3.3 Age difference of the entire Dataset 10

3.4 PDF & CDF of the entire Dataset 11

4. Hypothesis Testing on female Dataset 12

5. Summary 13

6. Bibliography 13

1. Introduction

The Olympics is considered the prime international event where each country tries to give their best performance to achieve glory. Modern Olympics has been featured in the Summer and Winter Olympics. Winter Olympics features snow sports which are not possible during the summer. Participants performing in the Olympics are not only able men and women, but the impaired (physically disabled) talented competitors who are given chance in the Paralympics. Given the dataset, it is categorized under two games Alpine Skiing and Snowboard, where we can see both men and women athletes participate in it and of a wide range of ages. It has been said that there is an expectation for snowboard athletes to be younger than Skiing. An analysis of the above-mentioned winter Olympic dataset has been done on whether there is an age difference between the two games. Therefore, the primary object of this analysis report is to analyze the age of the participants using exploratory data analysis to evaluate the dataset and perform descriptive data analysis using statistical methods for all the three Olympic years. Visualization of the dataset over various variables will show the statistical view of how the games are being held in the years. This report not only provides whether there is an age difference but also proves statistical hypotheses on the given analysis. This helps the IOC governing body of the Olympics to supervise the age of the Snowboarding participants with respect to the other winter games which take place. This analysis will not enhance the performance of a particular sport, but also set a clear requirement on the participant list.

In an overview, the report on the Winter Olympics dataset has been categorized under statistical measures, which shows the distribution of the participants for both the games and also the distribution of the gender in the games. A detailed graphical measure has been taken on the basis of the variables provided in the dataset. Followed by a hypothesis which supports the statistics observed in the dataset.

2. Proposed Approach

A proposed procedure is a systematic path to reach an outcome. Each and every problem needs an approach to obtain its correct decision or result. Similarly, this analysis has been followed with a fixed procedure, which includes various factors that help us approach the problem. As given below in Figure. 1, we can see what is the proposed approach in this analysis. Each step has been discussed in detail accordingly.

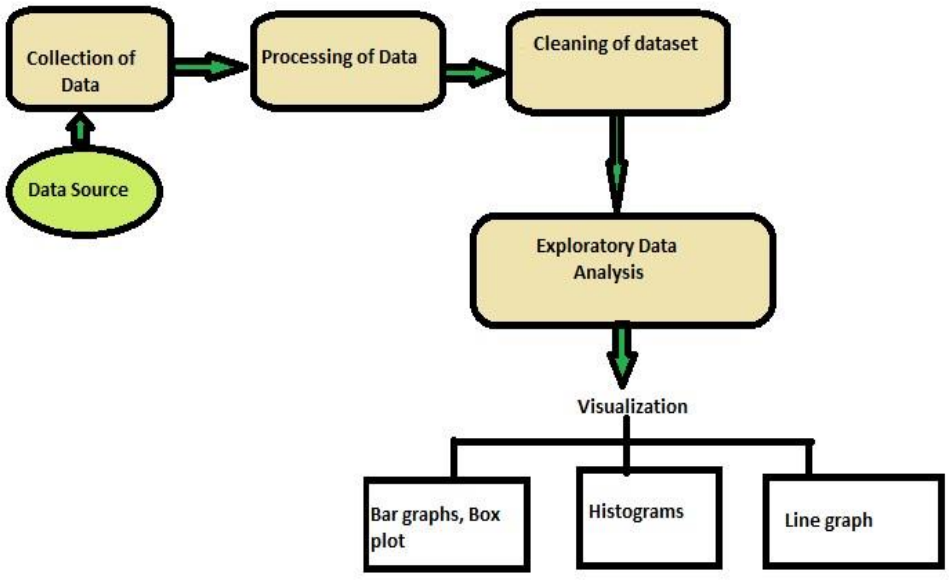

Figure 1: The diagram shows the steps of Data Analysis performed in the report.

2.1 Data Source

The provided data on the Olympic Winter games have been scraped from the recent Olympics which were held. The data is assumed to be retrieved from the data world dataset. When dealing with Olympic data we have a CSV file, which holds all the variables required for the data exploratory. Since the dataset holds a multiple dimension record, we try to follow the steps of data collection and processing the data for further analysis in the upcoming topics.

2.2 Collection of Data

Collecting data from a source is the process of gathering, extracting and storing the data which may be present in a structured or unstructured form like null values, missing values, non-text formats, etc. Since data collection is the very first step which is taken into account during data analysis, it should be meaningful data from which at least an interpretation can be derived by analyzing the variables. Usually, during the process, data is collected raw which is later cleaned and shaped into information. We usually tend to ask a certain number of questions when reading data. We often try to categorize them to read them easily and hope to gather at least a certain analysis as to what step can be taken further. The data has been used to perform a comparative study between two Olympic games. This has not only helped to analyze my report, but also can help the Olympic committee to make a fair participation list. The data has been used to find the total number of participants in each game, which has been divided into two different games as well as gender. With the high volume of data provided, we can perform descriptive analysis which includes standard deviation and variance, skewness, PDF and CDF on the entire data set (EDA) and hypothesis testing to reach a final conclusion.

2.3 Processing of Data

The step which follows data collection in Figure 1, is data processing. The data which has been collected from a particular source needs to be processed as it might be raw data which holds various null and missing values. Applying direct methods to raw data would lead to unnecessary wrong values and errors in analysis. This can be executed by checking and reducing errors, rectifying the incomplete and wrong data, reduce redundancy. Data processing helps in converting correct data so that it would not lead to negative results or output. The converted data can result in texts, numbers, and graphs so that the script being developed can easily be read as we move on with the analysis. The data can also be processed via algorithms or commands. Without data processing, the analysis would lead to a more complicated structure. The dataset consists of age, gender, game and year. Any missing or incorrect value which is not processed might lead to error results and visualization leading to a wrong hypothesis.

2.4 Cleaning of Dataset

Data cleaning is the process of fixing and correcting error details, example: incorrect format, corrupted data, missing values. When we work with various data sets, there might be possibilities that we end up with duplicate data. Algorithms are run for it so that we can eliminate the duplicate data and process an error free result. If the above step is not performed before exploratory as shown in Figure 1, then it might lead to unreliable or incorrect outcomes leading to incorrect data categorized or inaccurate graphs, which a person might think looks correct. While fixing a dataset, we start with correcting the inaccurate data, eliminating duplicate and null values, following with validating the dataset with a set of questions prepared for the particular data which will be the key element of the dataset. Benefit of this process leads to minimal errors, easy to relocate further errors, making it easier to take decisions on analysis as well as business.

2.5 Exploratory Data Analysis

In this step, analysis is been done on data using various methods. As this report contains visualization and hypothesis, I am using Exploratory Data Analysis to achieve the task. This step not only analyzes data minutely but also, summarizes the primary variables or attributes present. Applying this step besides various algorithms makes us understand the structure and the meaning of the dataset. The various plots used in this analysis are mentioned below and shown in the figure 1:

- Box plot
- Histogram
- Line plot

We can view the data in a visual format and can explain the analysis with a comparative study taking the variables into consideration. In this analysis Python programming language has been used to perform the EDA processes.

3. Analysis Setup and Result

3.1 Setup

We have always seen; that an analysis has always taken a helping hand of a programming language and a platform where the analysis will be performed. There are various programming languages which can be used to perform data analysis like Python, R, JavaScript, SQL and many more. With the help of Python, I have successfully completed the data analysis and the hypothesis. Python is a readable language. Its large library helps in performing many python design decisions. In this report, I have used various libraries to work on the descriptive analysis as well as the hypothesis. The IDE which has been used to perform the analysis is Jupyter Notebook, it is an open web-based interactive platform, which gets connected to various Kernels to allow development codes. Since it is a web interface it helps to deliver output easily. For the data analysis I have used various packages which are listed below:

3.1.1 Pandas

Pandas is a python library which helps to analyze data [code: import pandas as pd] Pandas basically helps in importing data from various sources and file formats, it is used for data manipulation and analysis. While working I have used the Dataframes which uses tabular data formats such as the Winter Olympics CSV file, which holds the data in rows and columns.

3.1.2 NumPy

NumPy is being used to support the large dimensional data which we use as our source data for the analysis [code: import numpy as np]. NumPy being an open-source universal data structure supports simplifying the programming codes and its workflow. With the help of NumPy many linear algebra operations can be executed, which helps in the descriptive analysis of the data.

3.1.3 Matplotlib

Matplotlib is used as a plotting library in Python, it is an open-source and free library. In this data analysis report, I have used the Seaborn toolkit, which helps in making statistical graphs. Matplotlib is one of the interactive visualizations which helps in working with huge data. [code: import matplotlib.pyplot as plt %matplotlib inline]

3.1.4 Seaborn

Seaborn has been built with the use of matplotlib with the help of pandas data structure. In this analysis Seaborn has played a major role in exploratory data analysis. It has helped in working with the dataframes used while coding. [code: import seaborn as sns].

3.2 Result Analysis

I have mainly tried to explain the analysis with the plots and graphs and hypothesis. The language being used in the analysis is Python which consists of packages, libraries, it has helped in the visualization and analyzing of the data provided. Variables like age, sex, year, and game has been taken into account to work on the analysis and hypothesis.

3.2.1 Descriptive Analysis of the data

1. Measure of Central Tendency (mean, median, mode)
2. Measure of Spread (Range, Quartile, Percentiles, absolute deviation, variance and std. deviation)
3. Measure of symmetry (Skewness)

1. Measure of Central Tendency (mean, median, mode):

Mean is defined as the following:

$$Mean(or \bar{X}) = \sum(X_i)/n = (X_1 + X_2 + ... + X_n)/n$$

Where $\bar{X}$ = The symbol we use for *"Mean"* (is pronounced as X bar)

$\sum$ = Symbol of summation

X_i = Value of the *i*th item X, i = 1,2,3…, n

n = total number of items

Median: Middle value of an ordered sample of numerical values.
Mode: Value that occurs most frequently.

The above formula and description have been taken from the Wikipedia source: https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/samplemean/

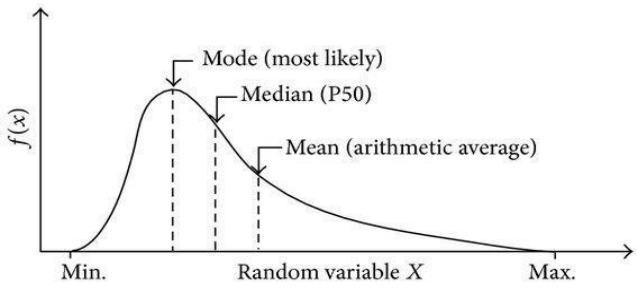

Figure 2: Image showing Mean, Median, Mode for random variable 'X' Mean, median and mode has been calculated over the entire data.

In analysis I have found out the central tendency on the variable age, listed below:

Obtained value of Age for the central tendency are as follows:

1. Mean: 25.8825 years: [code: np.mean(df.age)]
2. Median: 26 years: [code: np.median(df.age)]
3. Mode: 26 years: [code: df.age.mode()]
4. [code: plt.hist(df.age)] : a histogram has been used which represents the frequencies and helps in visualizing the data distribution. With the shape of the histogram, it is seen below that the data is right skewed with a tail. Below is the result of the histogram function.

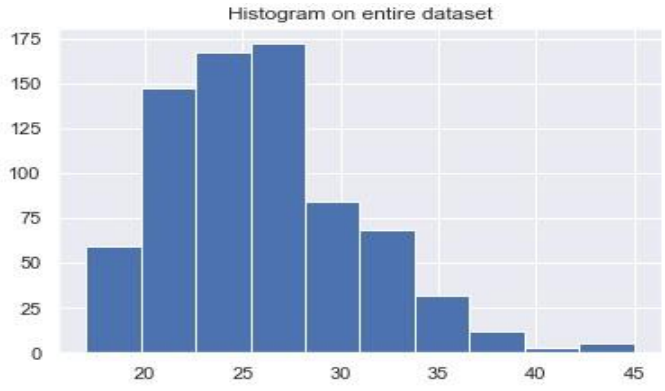

Figure 3: Histogram for the entire dataset

2. Measure of Spread (Range, Quartile, Percentiles, absolute deviation, variance and std. deviation):

- [code: sns.boxplot(df.age)]: A boxplot always shows the distribution of data which clearly shows the comparison between categorical set. It represents the quartiles of the distribution. ***As per the boxplot below we have the following value of the spread for 'age':***

- **Min**: 17 years; **Max**: 39 years; **1st Quartile (Q1)**: 22 years;

 2nd Quartile (Q2)/Median: 26 years; **3rd Quartile (Q3)**: 29 years

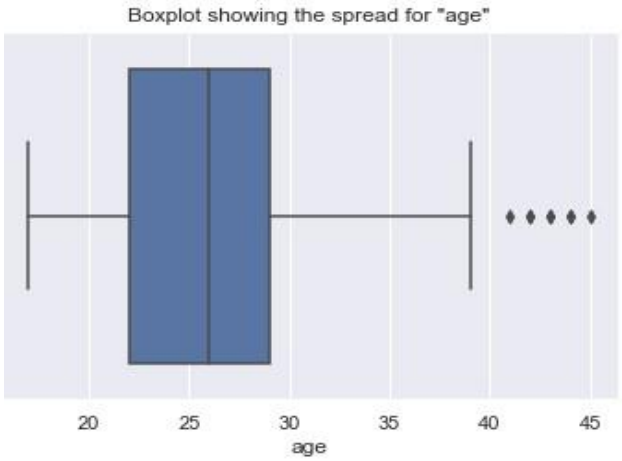

Figure 4: Boxplot showing the spread for 'age'

Interquartile range has been calculated using the formula [code: np.percentile], that provides with the result:

Q1: 22.0,

Q3: 29.0,

IQR: 7.0,

Therefore, the Interquartile Range (IQR) is 7 years.

- ***Standard Deviation and Variance***

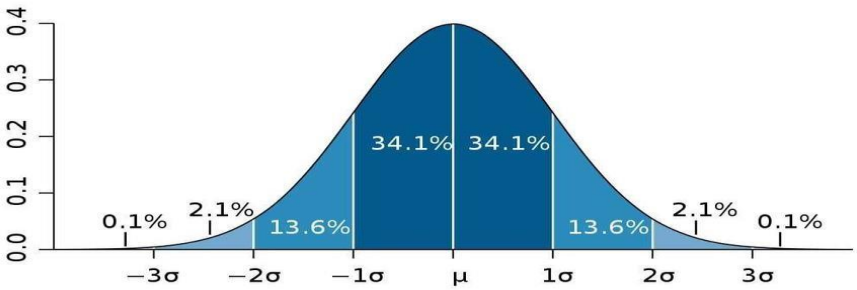

Figure 5: Image showing the Empirical Rule: 68:95:99

The above figure 5 is a Wikipedia source, the first step is the variance, which measures how far is the data from the mean. The second step is the standard deviation, which is the square root of the variance and calculates the amount of variation of the data being used. Here, the Mean being: 25.88

- and -σ to σ: 1st Standard Deviation would be ~1.86 (i.e., 68.2% of the data)
- the -σ2 to σ2: 2nd Standard Deviation would be ~3.72 (i.e., 95% of the data)
- and -σ3 to σ3: 3rd Standard Deviation would be ~5.58 (i.e., 99.7% of the data)

Below is the distplot of the data having the age as the argument:

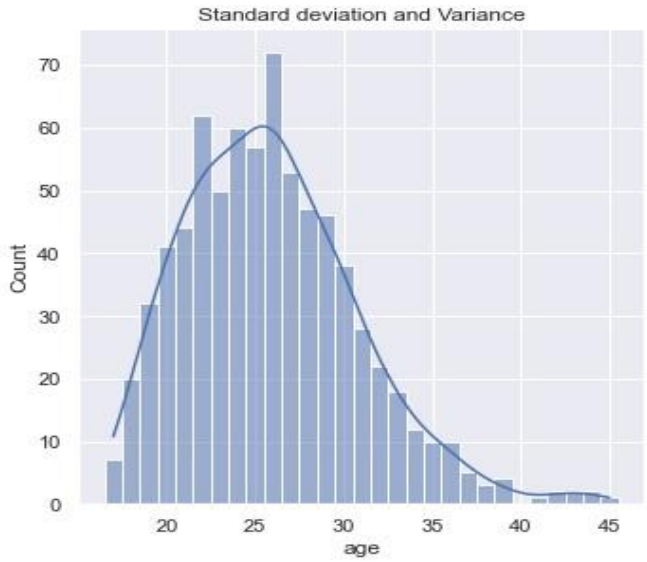

Figure 6: Standard deviation Variance plot using seaborn

3. Skewness:

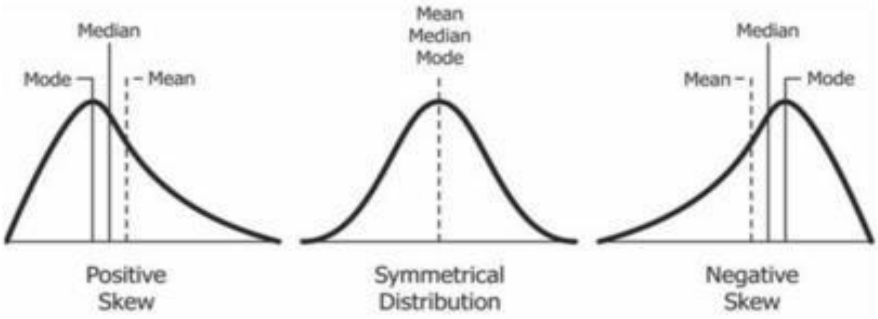

Figure 7: Image showing Positive Skew, Symmetrical Distribution, & Negative Skew

Therefore, the Skewness of the above data is right /positive skewed.

3.3 Age Difference of the entire dataset

The age difference for both the games as well as the count has been shown below with the figure. It is visible that there is a clear difference of age in both the games. The boxplot as well as the distplot has been executed to show the overall comparison and the distribution of the continuous data considering age as the primary variable and hue as discipline.

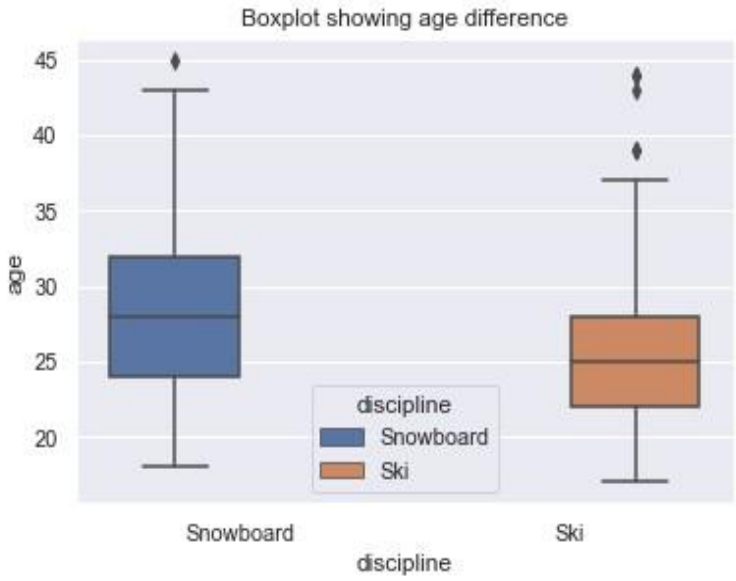

Figure 8: Boxplot showing age difference

3.4 PDF and CDF of the entire Dataset

PDF is defined as the Probability Density Function that is used for both discrete and continuous variables, it in turn helps in extracting the metadata. CDF is defined as Cumulative Density Function that is used for continuous data. It provides a probability for both the games in this dataset. Below figures describes PDF and CDF for the entire dataset as well as for both the Olympic games Snowboard and Alpine skiing played by the female participants.

Figure 9: PDF & CDF of entire dataset

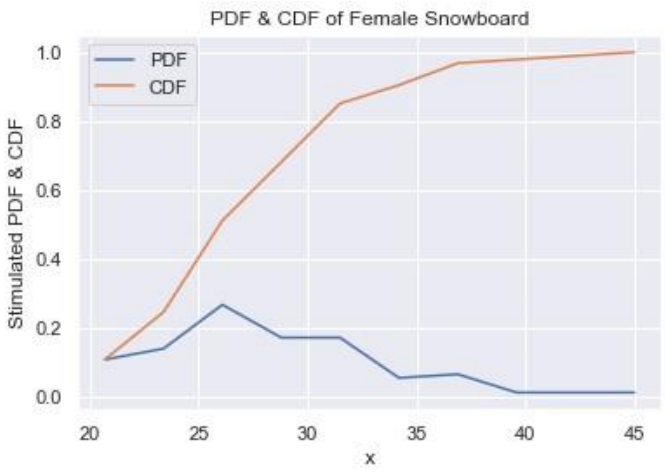

Figure 10: Female Snowboard

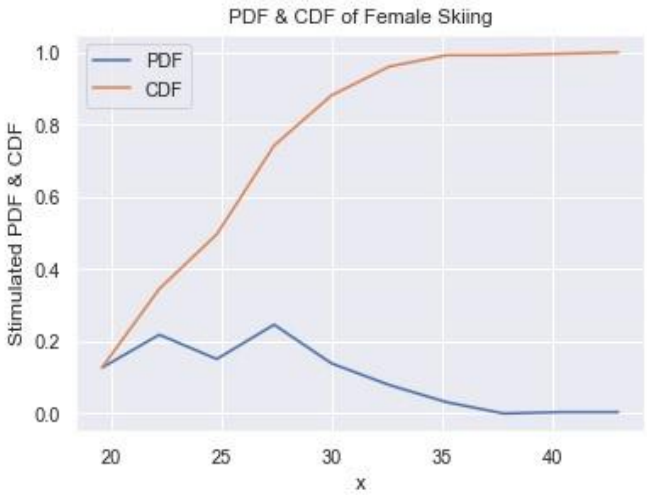

Figure 11: Female Skiing

4. Hypothesis Testing on Female Dataset

A Hypothesis testing always structures a dataset so that we can use the statistical solutions to check the assumptions. A hypothesis is always defined under two parts null hypothesis (Ho) and alternative hypothesis (Ha).

[code :
mu_1 = np.mean(df_snowboard.age),
mu_2 = np.mean(df_ski.age)
np.abs(mu_2 - mu_1)
which gives a result of 2.2651975683890555]

Assumption Check

- Ho: there is age difference, Ha: there is no age difference. This is to check the hypothesis being used whether null hypothesis can be accepted or rejected. The four steps to check the hypothesis are combining datasets, shuffle and random sample, calculating the test statistics, comparing the datasets making a decision.

Combing datasets: The two datasets which has been created on snowboard and skiing data with the variable age are combined :[code: np.concatenate([dataset1, dataset2]].

Shuffle and Random Sample: The dataset which has been concatenated is permuted in the next step.
[code: np.array([np.random.permutation(len(dataset1) + len(dataset2)) for i in range (10000)])]

Calculating test statistics: A test statistic is a random value that is calculated from the permuted data.

Compare datasets for P value: A p value for a statistical analysis is a probability that when a null value is accepted the alternative hypothesis is rejected. That is when the p value is less than 0.05, it indicates that null hypothesis should be rejected. If p value is greater than 0.05 then we accept the null hypothesis, rejecting the alternative hypothesis.
[code: p_value = 2*np.sum(testdata >= np.abs(test_sat))/10000]

5. Summary

The main objective of this report is to analyze and visualize the various variables which leads to the successful winter Olympic games. This analysis doesn't only help in determining the probability and prove hypothesis, it also helps the Olympic association to keep a fair selection of participants for the game. We have clearly executed the age difference of the two games in which females has also participated beside male. Following, the hypothesis test has been done to prove the assumption from the beginning that the participation is correctly handled. In this report we have Descriptive Analysis and Exploratory Data Analysis which has not only provided us the correct data but also helped us visualizing over graphs and plots. I have selected the Python language as it is easy to handle, open source and versatile in nature. As a result, we can conclude with the hypothesis testing that there is a significant age difference in between the female participants of Snowboard and Alpine Skiing game.

6. Bibliography

[1] Wikipedia: https://en.wikipedia.org/wiki/Olympic_Games

[2] Wikipedia: https://en.wikipedia.org/wiki/Exploratory_data_analysis

[3] Seaborn: https://seaborn.pydata.org/generated/seaborn.distplot.html

[4] Exploratory Data Analysis using Python: International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8, Issue-12, October2019 by Jitendra Pramanik, Kabita Sahoo, Abhaya Kumar Samal, Dr. Subhendu Kumar Pani

[5] Data Visualization: https://towardsdatascience.com/visualizing-your-exploratory-data-analysisd2d6c2e3b30e by Thomas Plapinger

[6] IDE: Jupyter Notebook

[7] Figure 5: http://www.differencebetween.net/science/mathematics-statistics/difference-betweenvariance-and-standard-deviation/

[8] Figure 7: https://en.wikipedia.org/wiki/Skewness.

[9] Libraries used: NumPy, Pandas, Matplotlib, Seaborn Wikipedia Source