Lance Hinds

@03045386

**The Gold of Athletes**

**Objective**

The objective of this project was to perform an analysis on a dataset from a set of possible dataset topics. The dataset that was chosen for this project included statistics of past Olympic athletes. The statistics included categories such as measurements of the athletes (age, height, weight), their country, and if they have won a medal. These items of this dataset were determined to be the most interesting to utilize.

To analyze this dataset, dataset sorting and observing methods were used from Python and Pandas coding. The methods were used in order to find unique values from the dataset that could be used for ideas of implementations of the dataset. After observing common values and categories that appeared in the dataset, visualizations were created to reflect the ideas gained from the experience. The visualizations consisted of many plotting functions that projected the desired data in a visual manner.

From the visualizations and observations taken from the testing process, generalizations could be made about the dataset. The generalizations reflected results seen that could be proved to be true and trustworthy. The generalizations could also be applied to determine essential ideas of the dataset and as a way of determining future data.

An important goal of the project was to enhance the understanding of how to sort and manipulate a dataset. The assignment reflected this by providing critical thinking and requiring creativity within the field of data science.

**Introduction**

The project began with the requirement of thinking of a topic or trend to observe from the dataset. After performing a few hours of analysis on the dataset, some of the categories appeared to be great topics of discussion. Specifically, the wide range of height and weight measurements were intriguing in addition to the years that were observed in the dataset. The common topic that created the most interest was the medal status of the athletes. Achieving a gold medal in the Olympics is an important event and must be attached to some sort of trend. Furthermore, is the earning of a gold medal related to some of the other categories that exist in the dataset? Could being of a certain height or specific year be related to the winning of a gold medal?

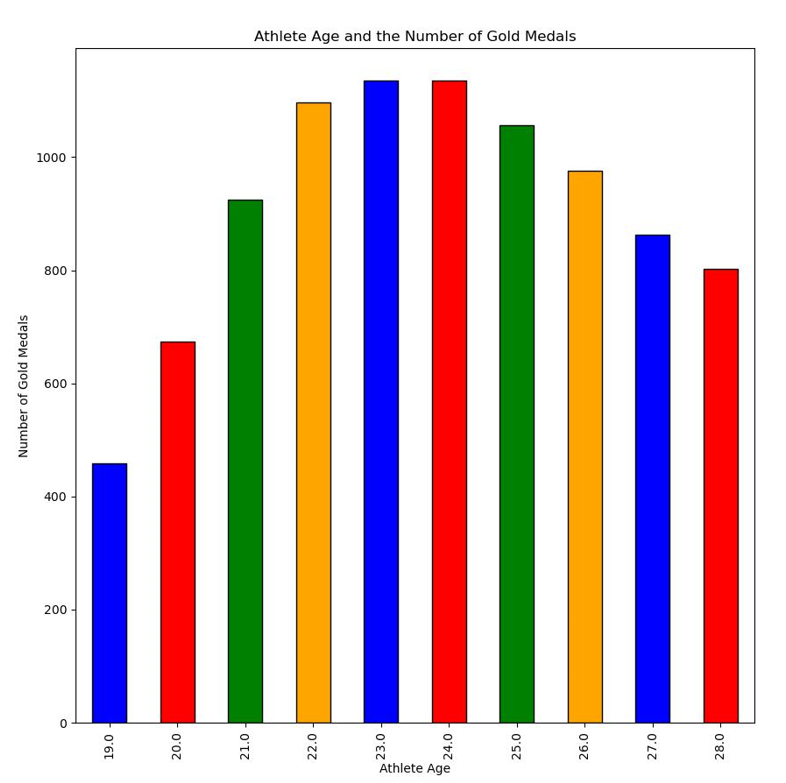
The idea of determining the criteria of winning a gold medal is the main theme of this project. The visualizations and commentary provided in this report will include the discussion of gold medal earnings. A generalization of who is qualified to win a gold medal based on their characteristics is an essential part of this project.

**Method**

Before creating visualizations or implementations using the dataset, handling missing values and data wrangling was required to be completed. Missing values were seen in different categories of the dataset. The categories of height, weight, age, and medal status had missing information for some of the athletes in the dataset. Missing values can cause misinterpretations and errors to result when analyzing the dataset using them. To get rid of the missing values, the mean of the existing values in a given column was used to replace the missing values in that column. For example, a missing value of height was replaced by the mean of all of the known height values. Even though this method got rid of the missing values, misrepresentations of the data of measurements resulted. The missing values were then handled instead by deleting rows that were missing information.

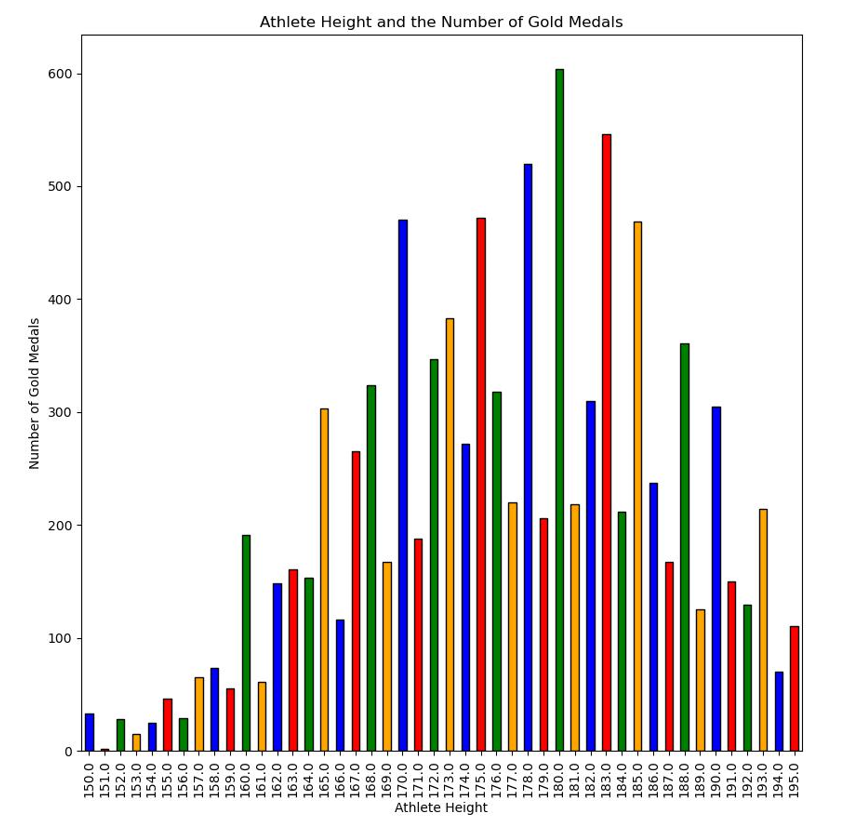
Data wrangling followed the missing value resolution. Firstly, a new dataset was created without the missing values. This new dataset will be used and altered to create visualizations in the project. Qualities of the dataset were observed further before creating implementations. The total number of rows and columns were determined, values such as mean and maximum were calculated, and sorting was utilized to determine categories of the dataset to study. The handling of the missing values and data wrangling prepared the dataset to be studied without difficulty.

**Storytelling**

The visualizations that were created in this project all involve a comparison of the number of gold medals achieved. The first visualization that was created was based on a comparison between two categories within the dataset. The categories implemented were the athlete’s age and the number of gold medals. The purpose of this comparison was to see if an athlete’s age had an influence on their performance in winning a medal. Using Python and Pandas coding, a bar plot was created as the form of the visualization. The first visualization of the project can be seen below.

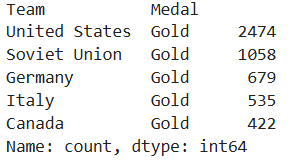
The athlete age is on the horizontal axis and the number of gold medals earned is on the vertical axis. Observing the plot, we can see that the majority of gold medals were earned by athletes around the age of 23. It is also seen that athletes who are too older or younger than 23 have earned fewer gold medals. We can conclude from this visualization that being of a young age in the Olympics is a strong factor in earning a gold medal.

The second visualization utilizes the same bar plotting format as the first visualization, but uses a different category to understand the gold medal earning. The second visualization makes a comparison between height and the number of medals. Due to its similarity to age, height was thought to have a strong correlation with the number of gold medals. The second visualization is shown below.

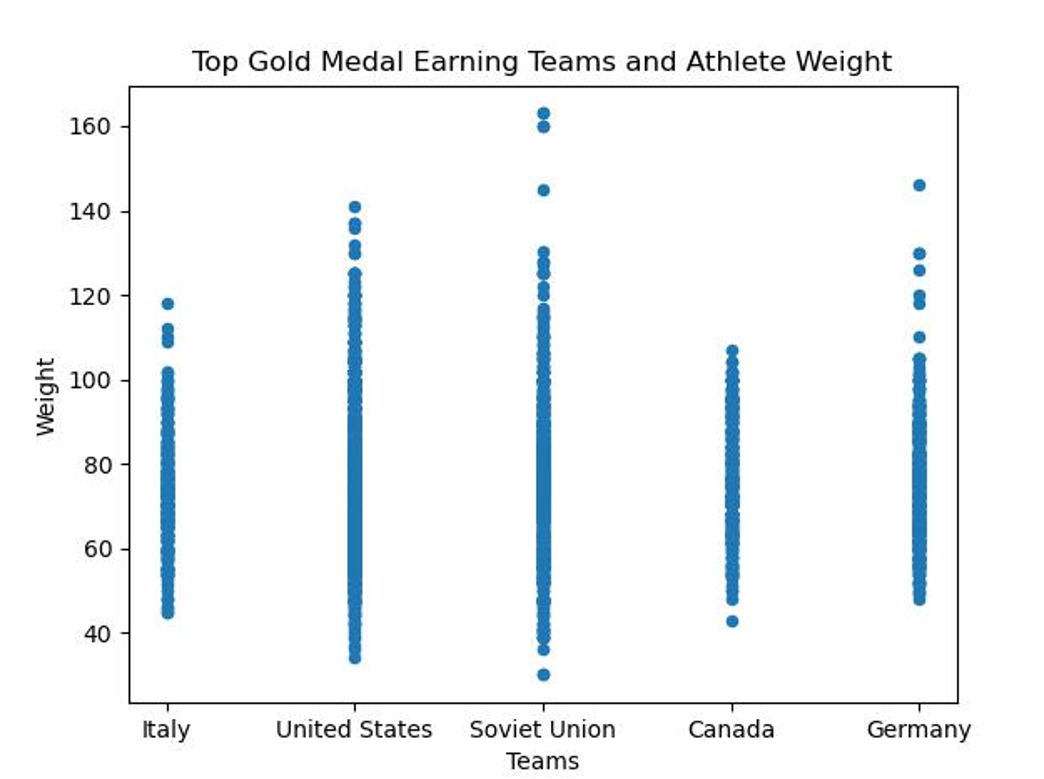


When observing the display, the trend in the number of gold medals is more difficult to see compared to the first visualization. Height values that have a large medal count are sometimes surrounded by adjacent height values that correspond with significantly lower medal counts. Overall, the most gold medals are received by athletes who are around 180 cm tall. It can be concluded then that athletes who are around 180 cm tall will have the best chance of receiving the most gold medals.

The third visualization compared the weight to the number of gold medals earned. A different approach was taken for this implementation. Firstly, the teams that participated in the Olympics and earned the most medals was determined from the dataset. The result of the top 5 in this category is listed below.

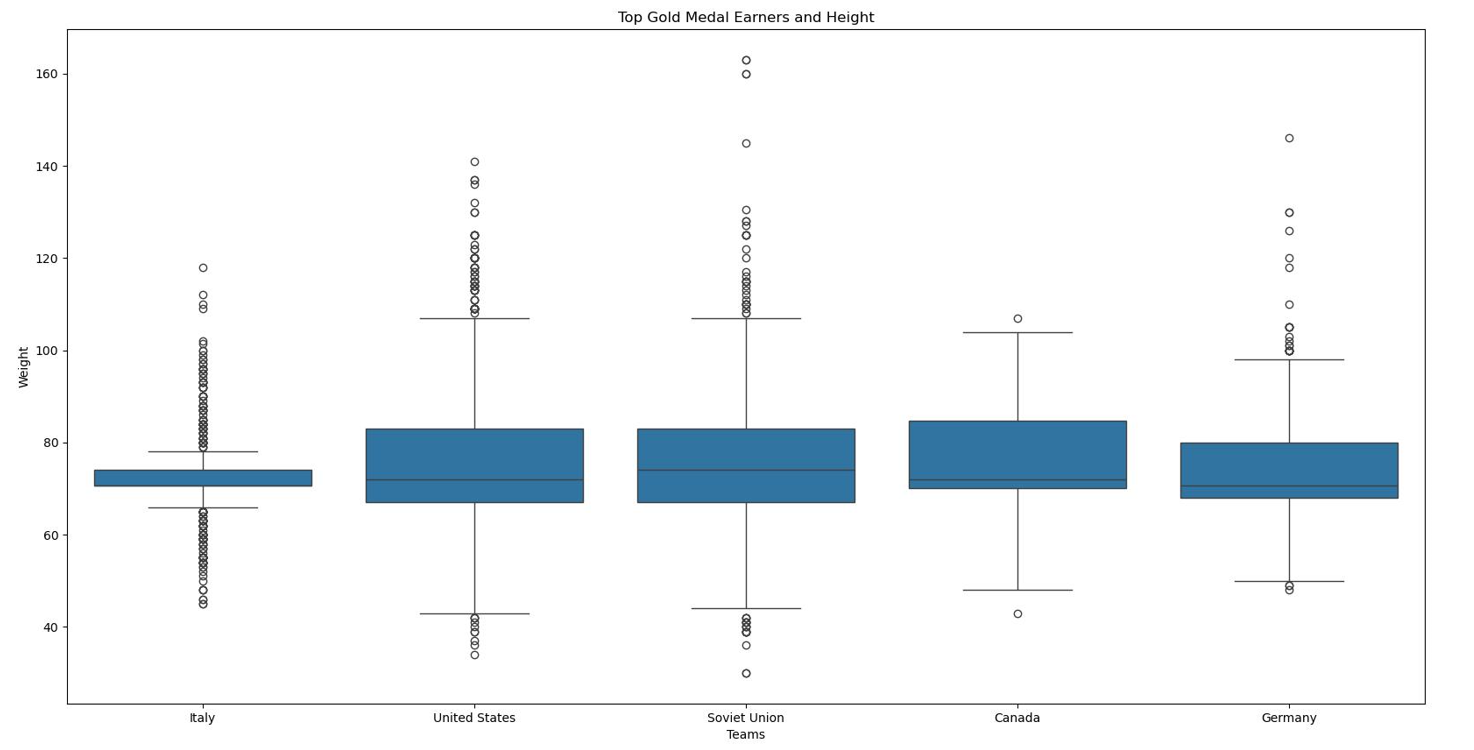


Based on this figure, the United States has earned the most gold medals in the dataset and the Soviet Union has earned the second most. The dataset was modified to include only the athletes of these 5 teams in order to more closely study the weight measurement. Visualization 3 shows the weight range of athletes from the top gold medal earning teams. The visualization is seen below.



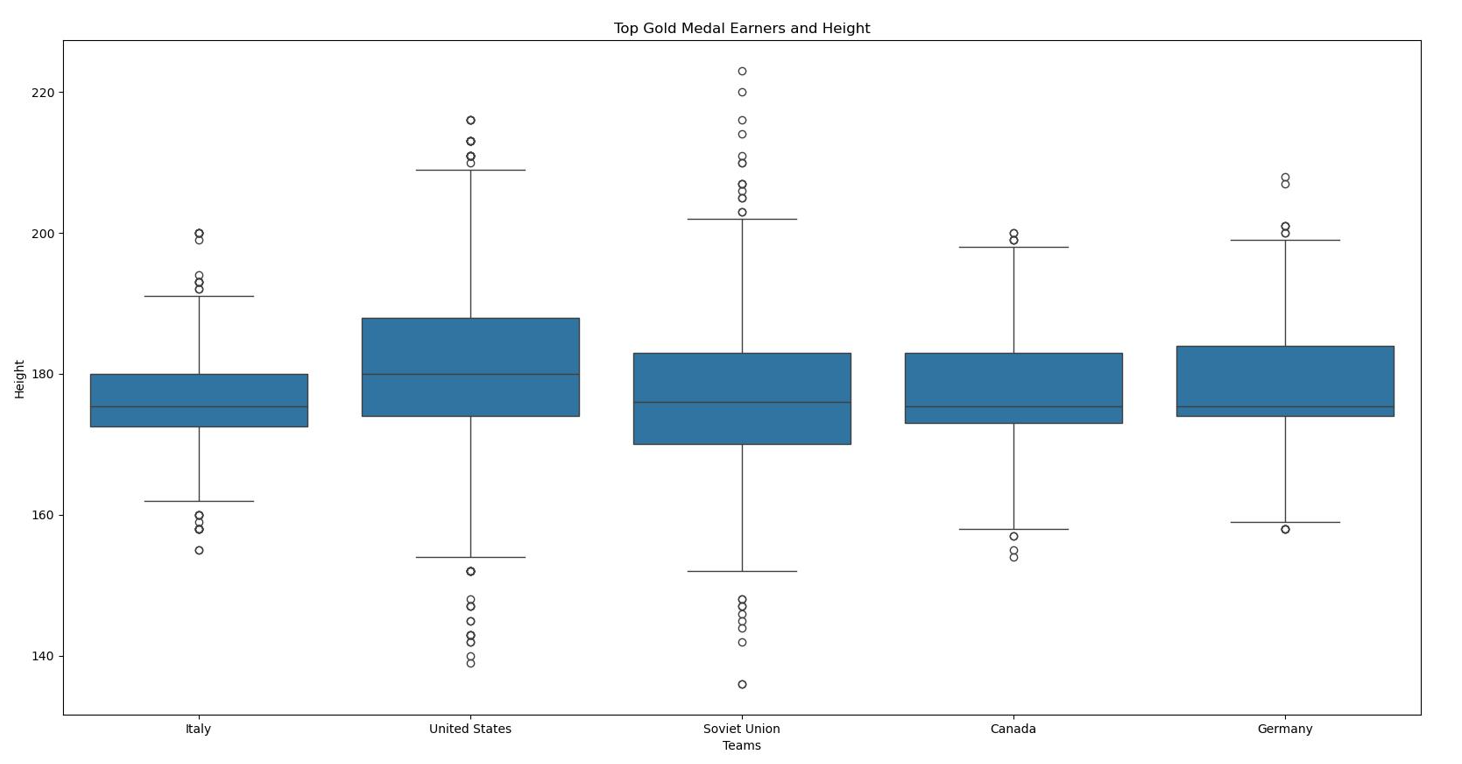
It can be seen that despite the United States earning the most gold medals, the Soviet Union and Germany have heavier athletes. It is also noted that the Soviet Union has the most outliers compared to the other countries.

To thoroughly see the weight comparison, a second visualization was made using s different graphing format. The second visualization is shown below.



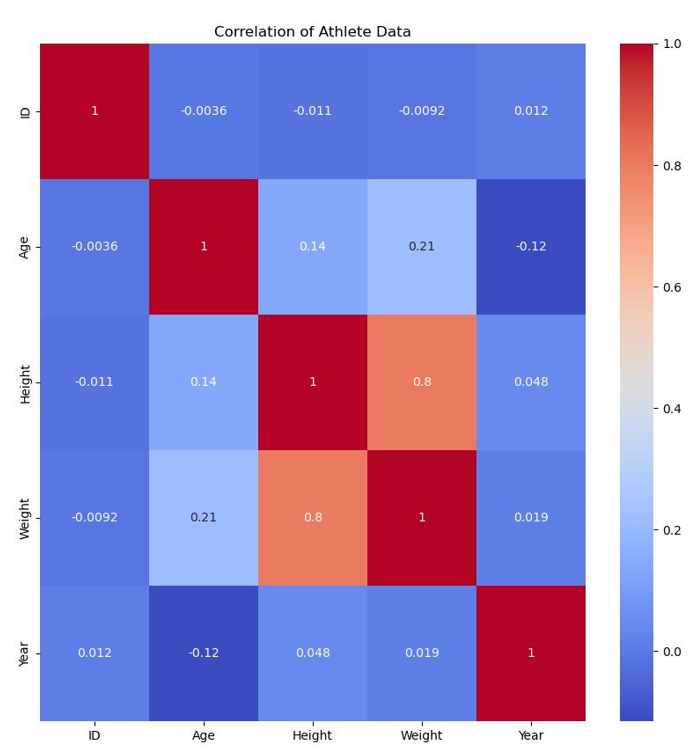
The second weight visualization displays the average of each countries weight values. The average of each countries weight can be seen to be between 60 and 80 kg. This means that an athlete weighing about 70 kg has a high chance of earning a gold medal.

The fourth visualization of this project follows the same process as the previous one. The implementation compares the top 5 earners of gold medals with their height values. The same graph type is used and the average value (this time height) is observed. The fourth visualization is included below.



Based on the display, the average height is seen to be about 180 cm. The height data corresponds accordingly to the number of gold medals. The United States (having the most gold medals) has the largest mean and is second in terms of high height outliers. On the other hand, Canada (having the least medals) has the smallest mean and the smallest range of height values. These noticed trends make the height a very reliable correlator to the number of gold medals. In addition, athletes with a height of about 180 cm will have a great chance of receiving a gold medal.

The final visualization of this project is the implementation of a heatmap. In the heatmap, a correlation is desired to be determined between numerical factors of height, weight, and age. Knowing these correlations will strengthen the idea of the factors relating to the probability of earning a gold medal. The heatmap for the overall dataset is shown below.



Age is seen to have a strong positive correlation with weight and a slighter weaker correlation with height. Height and weight also have a very strong positive correlation. The dataset has shown that many athletes included have a high height and a large weight. It can then be determined that athletes will also have a relatively high age due to the positive correlation.

**Conclusion**

The analysis of the Olympic athlete dataset proved to be beneficial to the understanding of manipulating datasets. Form the dataset analysis a generalization was determined in relation to the winning of a gold medal. It can be said that athletes around the age of 23, weighing about 70 kg, and being around 180 cm tall are the best at winning gold medals. This generalization can now be used to predict the future winners of Olympic events. Because of the project, further study and implementation of the Olympic dataset can be done to achieve other generalizations of the athletes.

**References**

[120 years of Olympic history: athletes and results](https://www.kaggle.com/datasets/heesoo37/120-years-of-olympic-history-athletes-and-results)

**Acknowledgements**

Professor Nerolu

Rgriffin from Kaggle

Classmates of the data science course