## **The Impact of Hosting and Physical Attributes on Olympic Performance and Participation**

### **Introduction**

This project explores the impact of hosting the Olympics and the physical attributes of athletes on their performance and participation utilizing a comprehensive dataset spanning 120 years of Olympic history, from 1896 to 2016. This dataset includes detailed information on all participating athletes with data points spanning from height and weight to team and placing.

### **Limitations**

Before diving into our research questions and analysis, we identified and addressed several limitations inherent in this dataset. Our primary goal was to transform the data into a more manageable form for analysis. To begin with, we acknowledged that there were missing data points for key attributes such as height, weight, and age. These missing values, if left unaddressed, could potentially introduce biases into our findings. Additionally, we considered the evolution of data collection methods over the years, recognizing that older data might be less precise due to advances in data collection techniques through modern times.

We also took into account the changes in Olympic scheduling. Until 1992, both the Summer and Winter Olympics were held in the same year every four years. From 1994 onwards, the scheduling was altered, with the Games alternating every two years. This change affects the continuity and comparability of data over time. Moreover, the introduction and removal of events over the years impact the number of medals available and the distribution of athlete attributes, leading to fluctuations in participation and performance metrics. The eligibility criteria for athletes have also evolved, allowing professional athletes to compete in many sports that previously only permitted amateurs. This shift could affect the level of competition and performance outcomes.

### **Data Cleaning**

To facilitate easier analysis, before we even started our data cleaning, we condensed our dataset to the last 15 occurrences of both the Summer and Winter Olympics or all Olympic occurrences since 1960. After condensing the dataset, we started our data cleaning process by addressing the null values in the medal column. It was easily understood that these null data points were for athletes who did not medal in their respective event. This led us to simply replace the nulls with "DNM", standing for “did not medal”. After cleaning up our medal column, we focused on rows with null values in the age, weight, and height columns. While contemplating what to do with these rows, we found out they only made up 6% of the total 210,000 remaining rows. Due to this small percentage, we decided to simply drop these rows. While we acknowledged that this might introduce some bias, the size of this percentage made it a reasonable decision that the integrity of our data and analysis would not be greatly impacted.

With a cleaner, condensed dataset, we were able to focus on our research questions and subsequent analyses for the overarching question of the impact of hosting and athlete attributes on participation and performance.

### **Research Questions**

The following research questions, t-test, and regression will be answered and addressed in the analysis below:

1. Is there one country who has hosted more Olympics since 1960? Does hosting impact the percentage of medals won by the home country?
2. What is the age distribution for Summer and Winter Olympics? Which sport(s) have the highest average age?
3. Does height or weight impact the likelihood someone is competing in a certain event? Has the average height and weight changed since 1960?
   * **T-Test Analysis:** Is there a significant difference in the height and weight distributions from Summer to Winter Olympics?
4. *Regression*: Does age, height, or weight impact the likelihood of medaling at the Olympics?

### **Analysis and Findings**

#### ***Question 1: The Impact of Hosting***

To answer this first research question we matched and appended country to city data so we could count the number of times each country has hosted the Olympics since 1960. There were a few snags that needed individually ironed out, for example the Soviet Union was dissolved and needed to be changed to the host country of “Russia '' and Bosnia and Herzegovina were formerly Yugoslavia.

Once the data was cleaned up and organized we were able to count the number of times each country hosted the Olympics. As can be seen on chart number one in our appendix, the United States has hosted the most Olympics since 1960, hosting one three separate occasions. Following that, Canada and Japan have both hosted three times while most other host countries have only hosted once. A conclusion that large potential Olympic host countries tend to only host one or twice over a thirty Olympic occurrence period.

After looking at hosting frequency we pivoted to the percentage of medals won by these hosting countries in both hosting and non-hosting years. To start this process we added up the total possible medals available for each Olympic occurrence and how many medals the hosting team won. This was placed into a new dataframe for ease of calculations. Then a new column was added to this dataframe to find the percentage of medals won by the host country. We also added up the total number of medals won by each host country for Summer and Winter Olympics in the years they did not host. Once we had these numbers we were able to divide by the average number of medals available for the Summer and Winter Olympics respectively to get the average percentage of medals won in non-hosting years. This data was appended to the new dataframe for comparison.

Looking at chart number two in the appendix, which displays the percentage of medals won by the host country in both hosting and non-hosting years, it is easily seen that hosting the Olympics increased the percentage of medals won. This means there is a positive effect on the number of medals won by the host country. To investigate why this might be, we did some further research and decided the following three factors are what we believe most impact this “home-country advantage”. One, host country athletes do not need to travel long distances or acclimate to new climates, time zones, or cultures. Two, there is strong local support potentially boosting athlete morale. And three, a host country gains automatic qualification for certain events, increasing the number of medals they could potentially win in a year.

#### ***Question 2: Age Distribution***

We wanted to compare the overall participants from each season, evaluating if the average age differed depending on the seasonal sport. Our hypothesis was that the average age would not vary significantly, as most people achieve their peak physical state in their early to mid-20s. Creating the data frames and extracting the information from our larger data sets was not very complicated. It really only required me to pull from the necessary columns in order to access and visualize the information accurately.

The results were not surprising. While the number of participants differed vastly between the two seasons, the age distribution was fairly similar, with the average age being in the early 20s for the Summer Olympics and the mid-20s for the Winter Olympics. A closer look at the data plots reveals that the Summer Olympics experience a steeper decline in athlete participation after the peak average age compared to the Winter Olympics, which show a steadier decline after the mid-20s. This could be due to several factors for which I couldn't acquire concrete data, though I have some theories. I believe this is because many summer sports exert more wear and tear on the body, whereas winter sports often involve tools like skis or snowboards, which may reduce strain and allow for greater longevity in athletes' careers, especially for Olympic-level athletes with access to top-tier facilities and equipment.

The second part of this question was: Which sport across all games has the highest average age? Locating this information was not difficult; We separated the specific columns related to sport and age from the overall data set. The most challenging part was creating a function to evaluate the average age of athletes in each sport and compile that information into a single row. Once visualized, the results were unsurprising. Equestrianism had the highest average age for athletes, at 34, followed by sports like shooting, cycling, and golf. These sports likely have higher average ages because they allow for greater physical longevity, with less strain on the body. Additionally, these sports are highly technical, and athletes with more experience often excel due to their familiarity with the equipment and their bodies.

There are many factors that contribute to being an exceptional athlete, and age plays a significant role. The convergence of physical abilities with experience and technical expertise can determine whether an athlete has a long career in their designated sport.

#### ***Question 3: Height and Weight Analysis***

For our third and final research question we looked at height and weight of Olympic athletes. To find this out we grouped our data by sport. Once grouped by sport we were able to find an average height and weight and created a new dataframe to hold this information. Once this new dataframe was fully created we plotted these average heights and weights by sport and sorted from lowest to highest. These two charts can be seen in the appendix, charts five and six.

What these two charts display is that while there is a large central tendency for height and weight, there is an upward trend when the charts are sorted. This means there are sports where height and weight factor into the likelihood of an athlete participating. For example, the lighter and shorter athletes offer greater agility and maneuverability and are often seen in sports like gymnastics while taller and heavier athletes may be seen in events like basketball and volleyball due to their longer reach and strength.

This analysis gave us an easy transition into looking at height and weight of Olympic athletes over time. For this analysis we averaged height and weight by year and separated them by Summer and Winter Olympics. We then took these plot points and made two line charts that can be seen in the appendix, charts seven and eight.

Taking a closer look at these charts, we can see an overall upward trend for both height and weight as well as Summer and Winter Olympics with a few peaks and valleys. These peaks and valleys led to some further investigation that let us know that in 1988 a large volume of new events were added to the Winter Olympics. This change had a few years long impact on both the average height and weight for the Winter Olympics.

Overall, we should note that while there has been an increase in both average height and weight for the Olympics since 1960, there is a larger change in average weight than average height. The biggest change in average height is from around one hundred seventy centimeters to one hundred seventy six centimeters which is about a inch to inch and a half difference. While the biggest change in average weight is from sixty eight kilograms to seventy two kilograms, which is the equivalent of about ten pounds.

***T-Test****:*

This final research question prompted us to conduct a t-test during our research. While looking at the heights and weights over the years we also looked at the height and weight distributions for the Summer and Winter Olympics. These distributions can be seen in the appendix in the form of violin plots, charts nine and ten. Looking at these violin plots, we are able to see that both height and weight plots for Summer and Winter are similar but not identical. Some easily identifiable differences include the tail on the Summer weight plot as well as the elongated shape of the summer height plot.

To conduct these two t-tests we first identified our significance level and our hypotheses. We chose a significance level of .05 with our null hypotheses being that there is no significant difference between the height and weight plots for the Summer and Winter Olympics and our alternative hypotheses being that there is a significant difference between the height and weight plots for the Summer and Winter Olympics. The outcome for our t-tests were a p-value of 1.32 e-48 for height and .084 for weight.

For our height t-test, a p-value of 1.32e-48 was less than our significance level of .05. This means we should reject our null hypothesis and understand that there is a significant difference in the height distribution for the Summer and the Winter Olympics. To finalize our analysis here we looked at the means and medians of height for both as well. While these numbers were nearly identical, and can be seen in the appendix as image one, we are confident in rejecting the null due to the large volume of our sample size.

For our weight t-test, a p-value of .084 was greater than our significance level of .05. This means we can accept the null hypothesis and understand that there is not a significant difference in weight distribution for the Summer and Winter Olympics.

### **Regression Analysis**

Our regression analysis aimed to identify any linear relationships between the physical attributes of athletes (age, height, weight) and their success over their career in the Olympics quantified as a medal count. The results showed extremely low R-squared values, indicating no linear relationship (Charts 11-16):

* **Age of Winter athletes**: R2=0.0028R^2 = 0.0028R2=0.0028
* **Age of Summer athletes**: R2=0.0080R^2 = 0.0080R2=0.0080
* **Height of Winter athletes**: R2=0.0007R^2 = 0.0007R2=0.0007
* **Height of Summer athletes**: R2=0.0034R^2 = 0.0034R2=0.0034
* **Weight of Winter athletes**: R2=0.0002R^2 = 0.0002R2=0.0002
* **Weight of Summer athletes**: R2=0.0031R^2 = 0.0031R2=0.0031

These findings suggest that while age, height, and weight do not have a strong linear correlation with medal counts, factors such as survivorship bias and versatility in physical build might play roles in athletic success. For instance:

* **Age**: Older athletes with higher medal counts may reflect survivorship bias, where only the most skilled and experienced athletes continue competing at older ages, which can explain the difference between what most people expect where athletic ability declines with age once adulthood is reached.
* **Height and Weight**: Athletes with a more average build might be more versatile and able to compete and excel in multiple events, leading to higher medal counts than an athlete that has a super specialized body type for a specific event. For example, Michael Phelps, with a relatively average height and weight for an Olympian, excelled in multiple events.

### **Conclusion**

This exploratory data analysis delved into the multifaceted impact of hosting and the physical attributes of athletes on Olympic performance and participation. The exploration of data spanning 30 Olympic occurrences revealed intriguing insights into the dynamics shaping athletic success at the highest level of competition.

#### 

#### ***Hosting Impact:***

The analysis of hosting frequency and medal performance highlighted that hosting the Olympics significantly enhances the medal count for the host nation. Factors such as home advantage, local support, and automatic qualifications contribute to the success of host countries. The findings underscore the strategic advantage of hosting the Olympics in boosting national performance and fostering a sense of pride and achievement.

#### 

#### ***Physical Attributes and Performance:***

The examination of age, height, and weight distributions among Olympic athletes revealed interesting trends over time. While there was an overall upward trend in height and weight since 1960, the regression analysis showed a lack of strong linear correlations between physical attributes and medal counts. Factors like survivorship bias and versatility in physical build were identified as potential influencers of athletic success beyond mere physical attributes.

#### 

#### ***Future Research and Implications:***

Future research endeavors could delve deeper into exploring non-linear relationships between physical attributes and Olympic success, as well as investigating the impact of psychological factors, training methods, and competition strategies on athletic performance. Understanding these dynamics can provide valuable insights for athletes, coaches, and sports organizations aiming to optimize performance and participation in the Olympics.

In conclusion, this analysis sheds light on the intricate interplay between hosting, physical attributes, and athletic success in the Olympic arena. By leveraging data-driven approaches and continuous exploration, the sports community can further enhance training programs, talent identification, and performance optimization strategies for Olympic athletes, ultimately elevating the standard of competition and achievement in the world's most prestigious sporting event.

### 

### ***Recommendations:***

* Continuously monitor and analyze hosting patterns and their impact on medal performance to inform future hosting decisions.
* Conduct further research on non-linear relationships between physical attributes and Olympic success to uncover hidden factors influencing athletic performance.
* Implement data-driven strategies in athlete training programs and talent identification processes to optimize performance and maximize medal potential in the Olympics.

This analysis serves as a foundation for ongoing research and exploration in the realm of Olympic sports analytics, offering valuable insights for athletes, coaches, and sports organizations striving for excellence on the global stage.

### **Appendix:**

### Image One

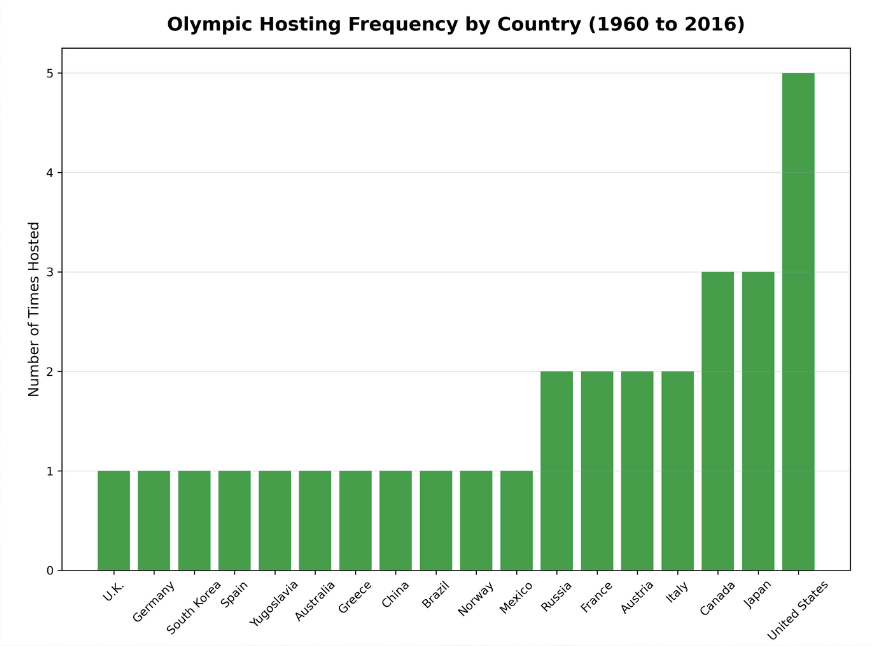
Chart One

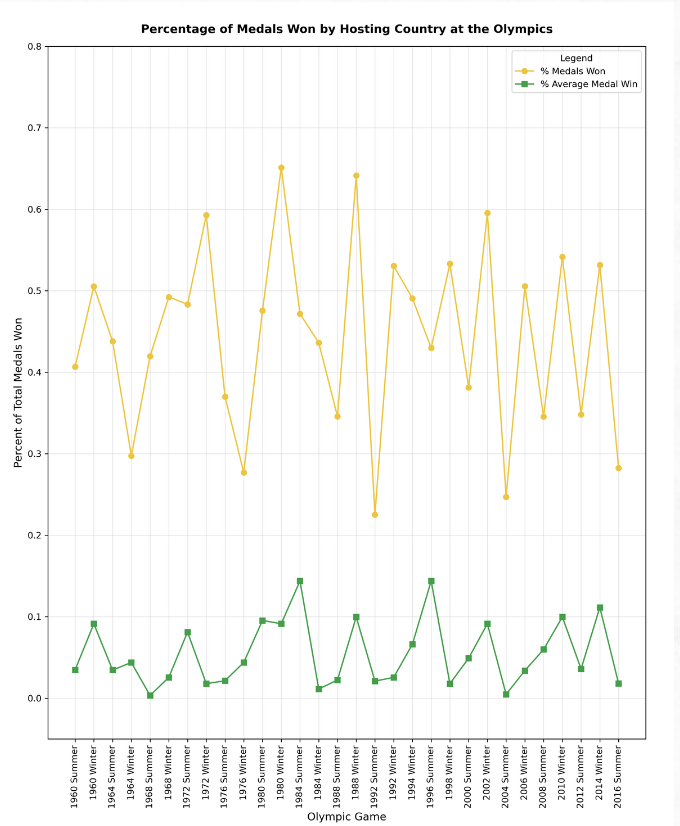
Chart Two

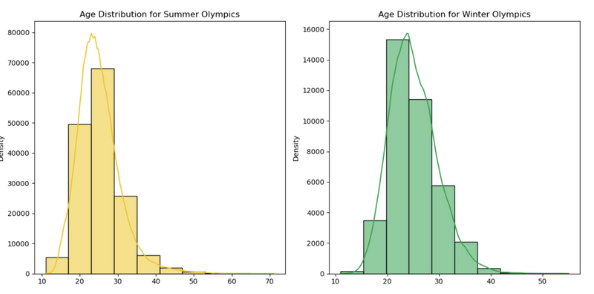
Chart Three

Chart Four

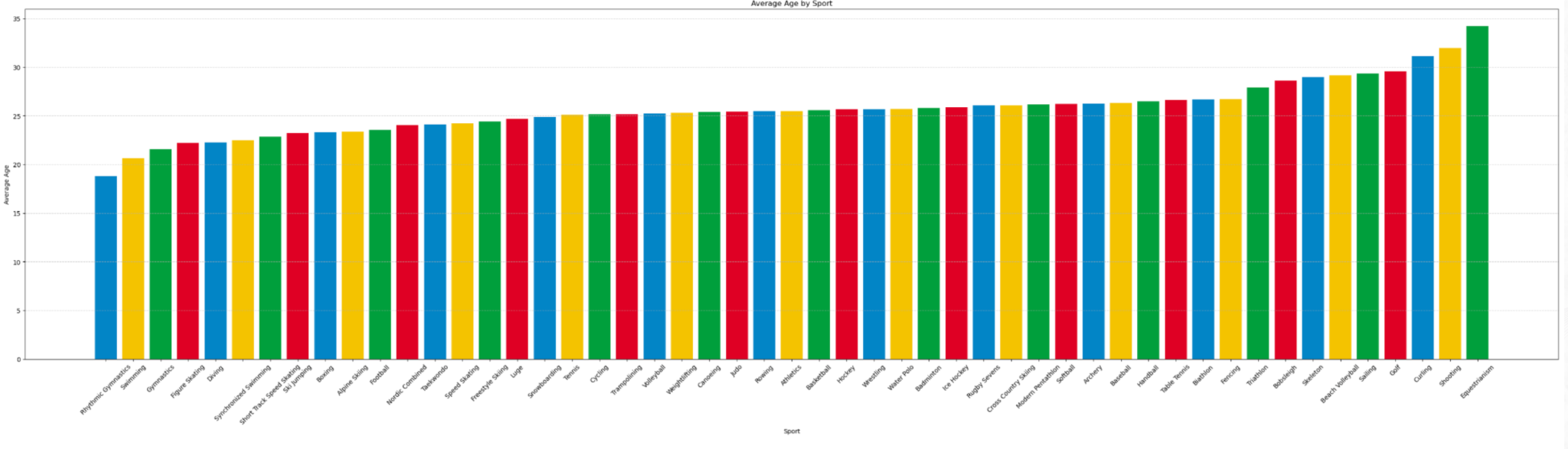


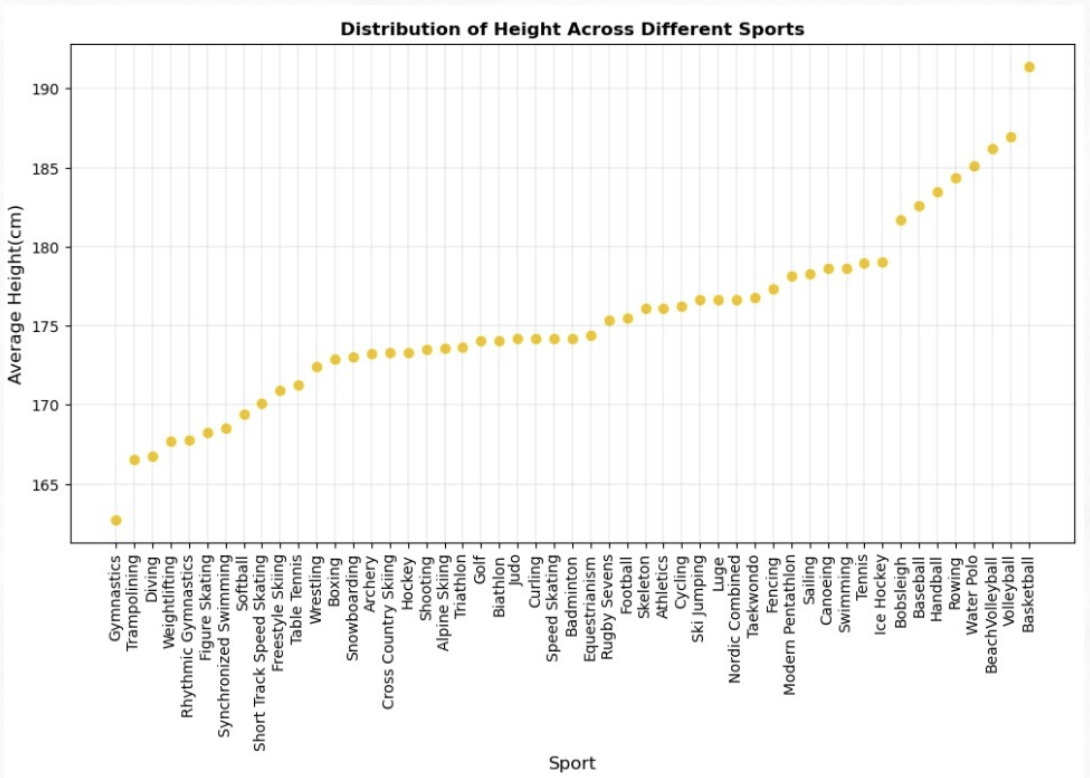
Chart Five

Chart Six

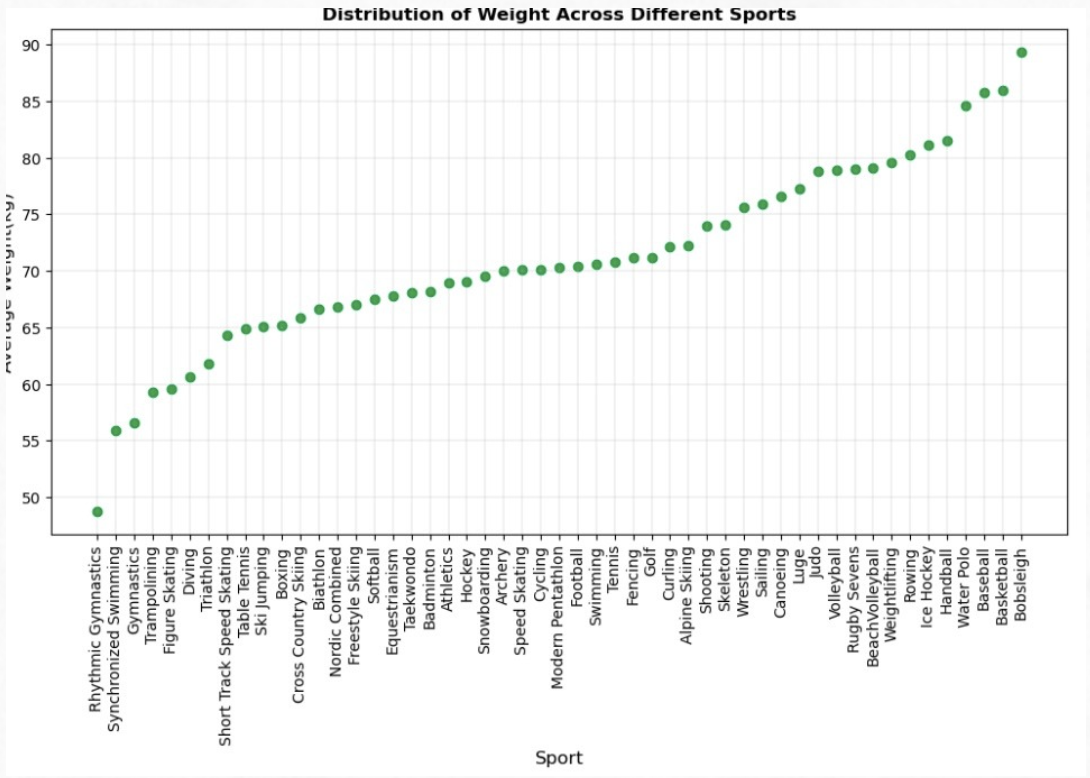


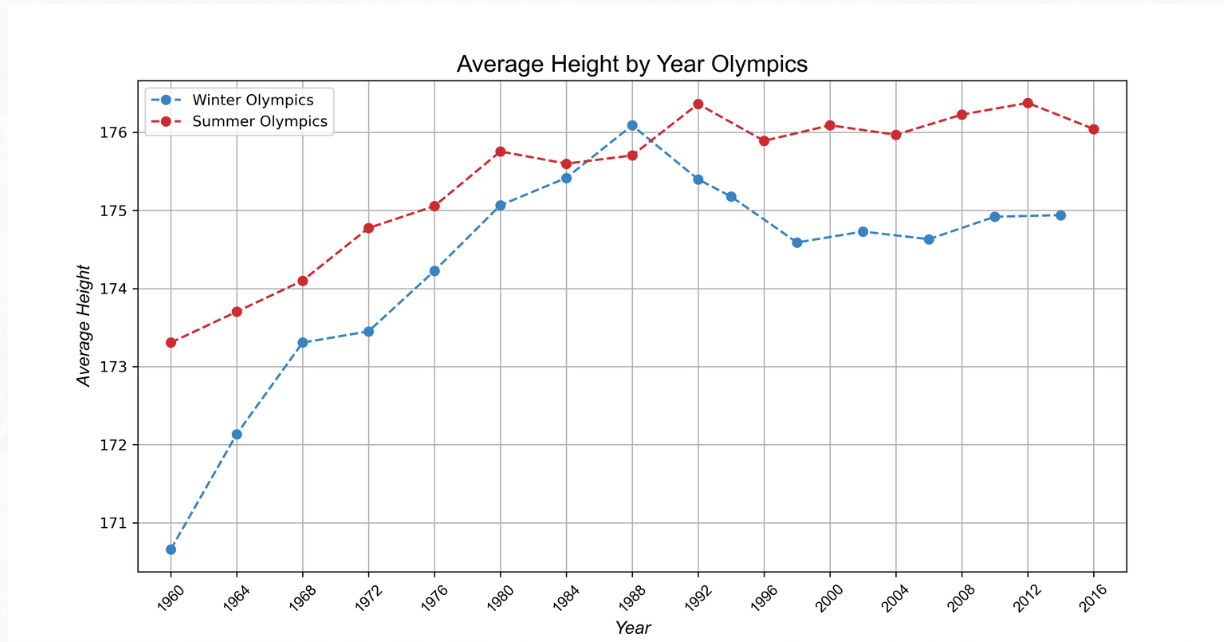
Chart Seven

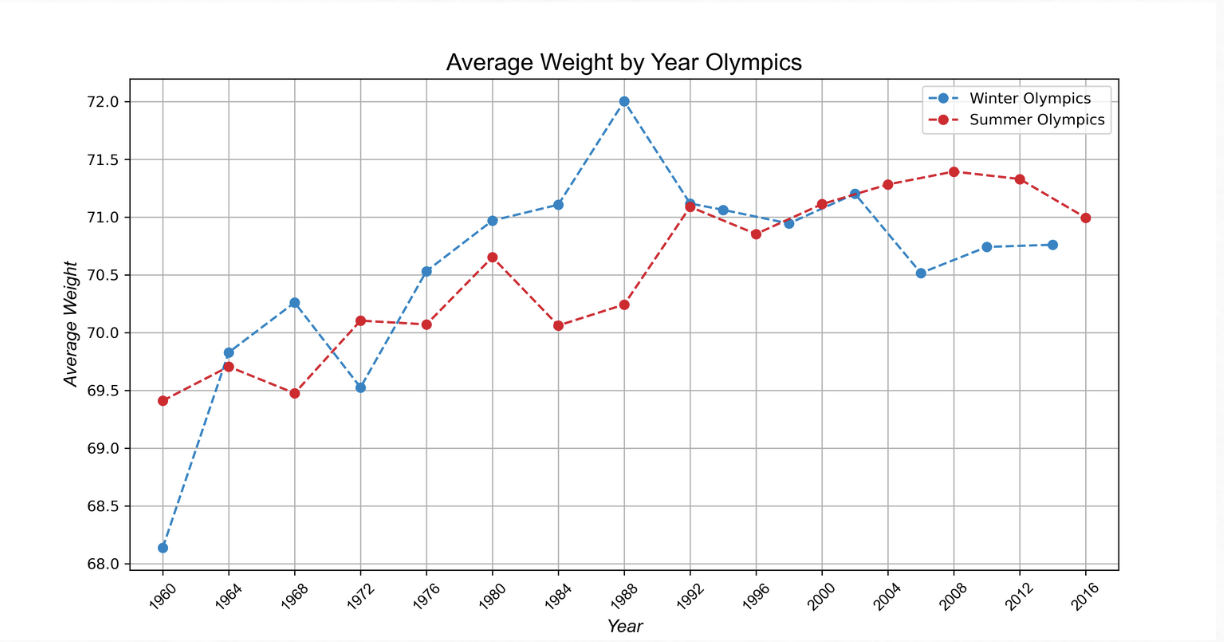
Chart Eight

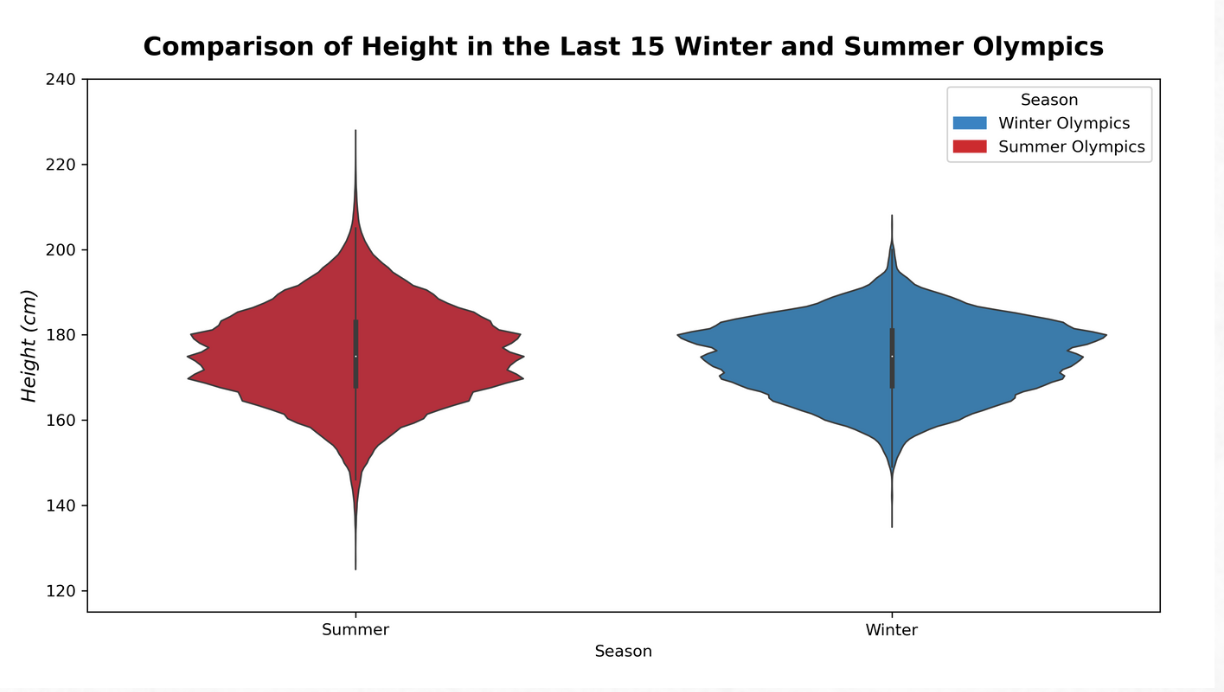
Chart Nine

Chart Ten

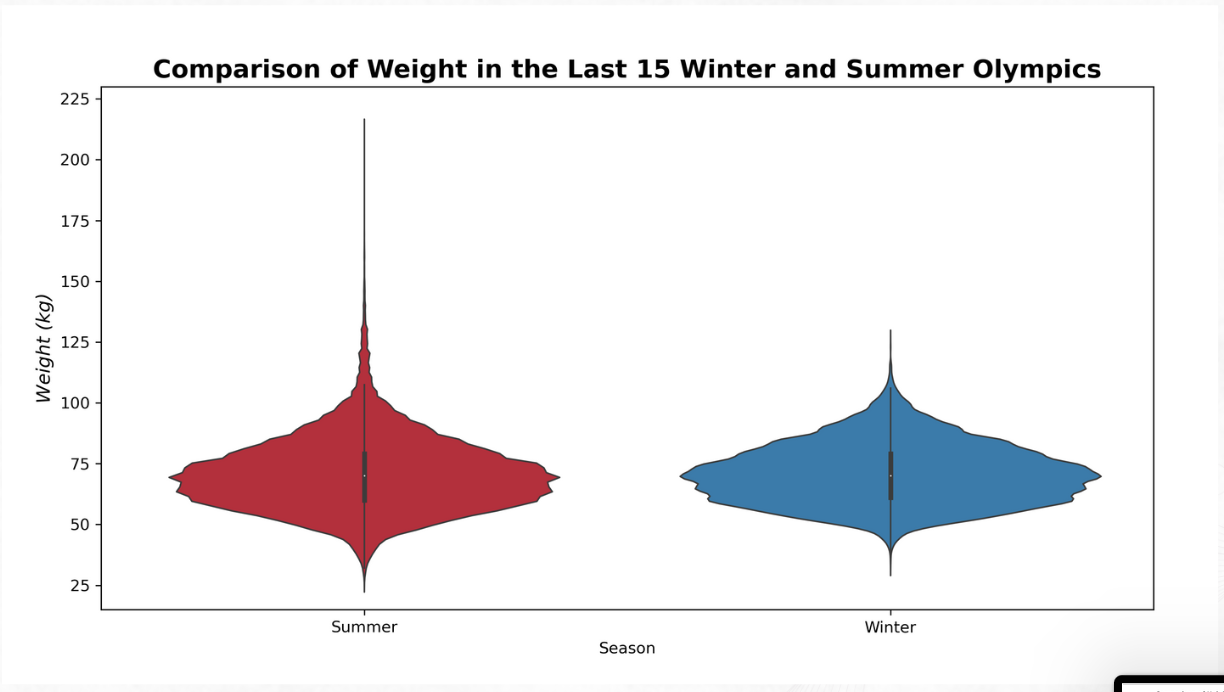


Chart Eleven

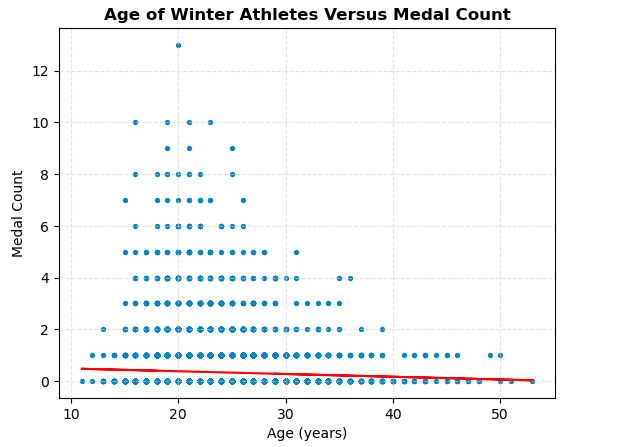


Chart Twelve

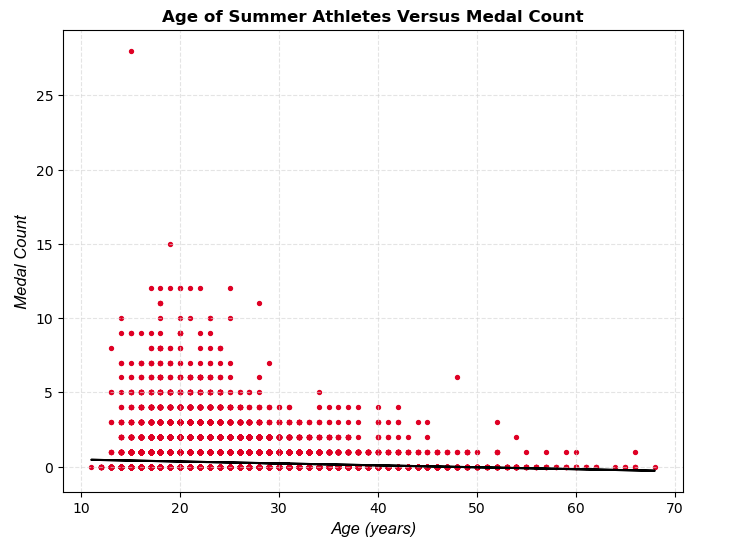


Chart Thirteen

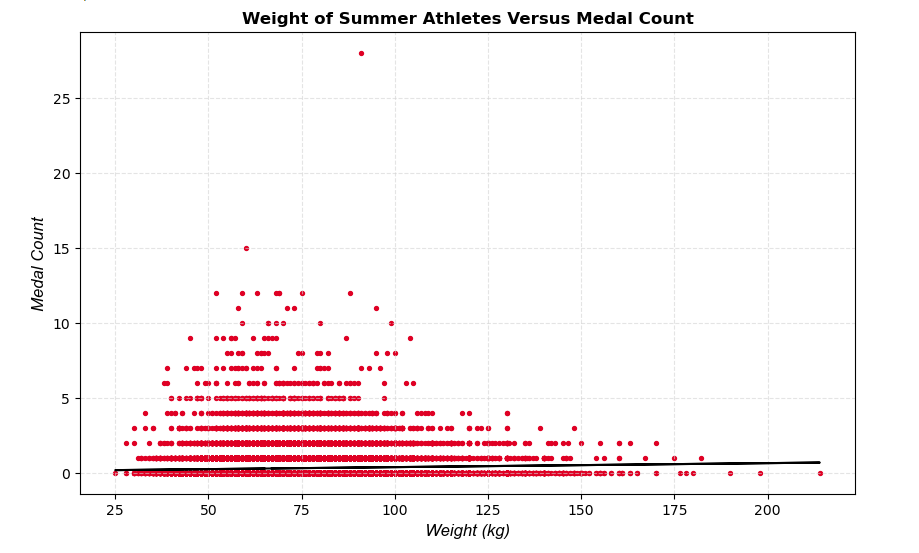


Chart Fourteen

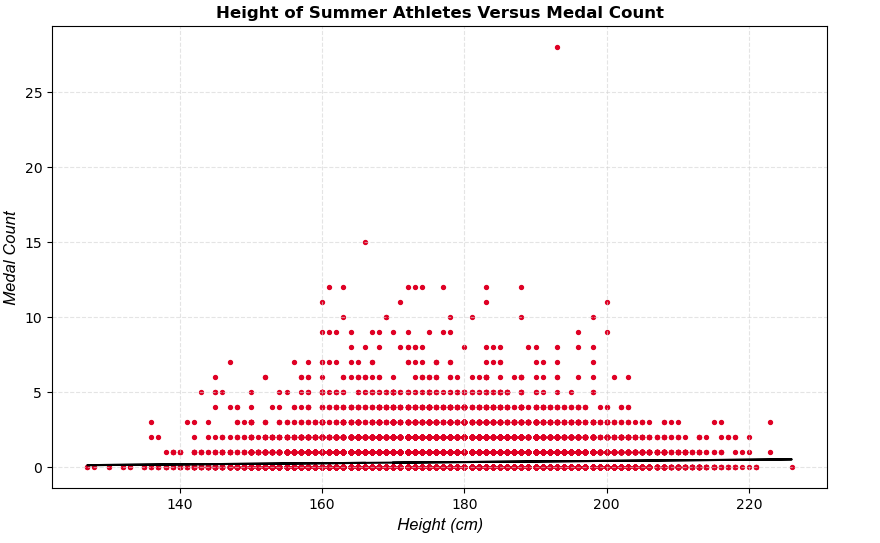


Chart Fifteen

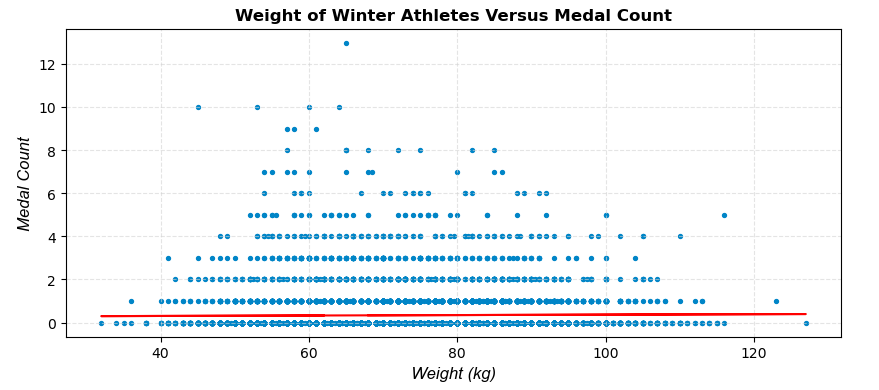


Chart Sixteen

