Project 1: Does height correlate with athletic success in the Olympics?

Rikki Chiba, George Jiang, Emma Wang

Our group decided to explore the question of whether heights influence the ability to win medals in the Olympics. We used the data set of the 2016 Olympics in Rio de Janeiro, which consists of the official statistics on the 11,538 athletes and 306 events at the Olympics game. The source data came from Kaggle who sourced it from Rio Olympics 2016 website, and it includes 3 files, athletes, countries, and events. The athletes' file was the most relevant to our project, so we utilized and modified the dataset based on our needs.

Variables:

There are 10 variables in the dataset: name, nationality, the number of bronze, silver, and gold medals won, date of birth, height in meters, sex, sports category, id, and weight.

- Name: Name of athletes
- Nationality: Nationality of athletes
- Bronze: Number of bronze medals won
- Silver: Number of silver medals won
- Gold: Number of gold medals won
- Dob: Date of birth of athletes
- Height: Height athletes in meters
- Sex: Sex of athletes
- Sport: Sports category the athletes is in
- Id: The Identification number of athletes
- Weight: The weight of athletes in kilograms

Data Filtering and Processing:

- Filtered out athletes without height - The first step of data filtering was to filter out any data points that don't have any height to avoid any error in plotting data.
- Separated by non-medalist and medalists - Since we want to examine the relationship between heights and medalists, we separated the non-medalists and medalists in the dataset. We used a for each loop and collected any athletes that had a sum of zero in the number of gold, silver, and bronze medals for non-medalist, and the remaining data is the medalist.
- Filtered by sports: Since there are 33 sports in an Olympic Game, we wanted to visualize the data better without crowding out the canvas and overloading our audience. Therefore, we chose sports where height could be a factor in winning medals, such as basketball, gymnastics, and weight lifting, sports where height is neutral in winning medals, such as shooting, badminton, and sailing, and also excluded the sports that were too broad, such as aquatics as it encompasses diving and swimming, two very different sports.
- Calculation of median height of each medal category for each sport - We decided to calculate the median height of the winners for each medal category for each sport because the median is not as easily influenced by outliers as the mean is. We also thought this would be the easiest for someone to interpret rather than looking at a bunch

of scattered data points, so it would be a good introductory graph that summarizes the data. First, we used a for each loop to separate the data by sport and then further separated the data by the non-medalist and medalist (as explained above). Then, for each medal category within the medalist data, we calculated the median height using d3.median.

- Separated by female and male - We also separated the separated dataset by sex because we thought there are statistical differences in height between females and males. This would help us to better visualize the relationship between height and sports. However, in the end, we decided to remove the graph visualizing the male and female heights, since it did not add any additional information to the graph that didn't separate male and female athletes.

Design Rationale

Scatterplot design rationale (1st and 3rd chart)

- Scatterplot - We decided to use a scatterplot because we wanted to explore the correlation between the number of medals won and athletes' height. Since height is a continuous variable, a scatterplot is the best way to show if there is a correlation between number of medals won and height of the athlete.
- Scales - We used the extent function to find out the minimum and maximum for price and height. We wanted to ensure that the data has enough padding, therefore, we added margins to the canvas. We also adjusted the domain of the extents slightly (make the min a little lower and the max a little higher) in order to make the value of some of the points more clear. That way every point falls between two labeled tick marks or near a one labeled tick mark so that the value of those points is more clear.
- Jitters - Since we were dealing with discrete data, we added some horizontal jitter to avoid data clustering. Specifically, we only added jitter for athletes that won 1 to 2 medals, because there are many overlapping data points. This is a trade-off because it is impossible for an athlete to win 1.4 medals; however, without the jitter, the graph is extremely cluttered to the point where no relationship could be inferred. It further emphasizes the number of athletes who won 1 medal during the Olympics and shows the structure of the game. On the other hand, we minimized this trade-off by not adding any jitter to athletes who won more than 2 medals, where the data points are sparse. Even though this may create confusion among the users as 2 columns had a cluster of points and the rest had clean vertical points, the users can see the data in the most accurate way possible. This design rationale illustrates that there is no perfect way to visualize the data.
- Legend on the side: We added a legend to the side of the graphs to make it more clear what each color represented. For the last chart where we look at the distribution of the height of medalists and non-medalists, we also made clear that larger circles meant that more medals were won in that particular medal category (e.g. multiple gold medals won would mean a bigger circle).
- Color: We set corresponding colors to gold, silver, and bronze medals won. This way it is intuitive to the users regarding the quality of the medal the athlete earned.

Non-medalists were black as it was distinct from the other three medal colors and blended in with the color scheme of our charts. We also lowered the opacity to 0.75 because that way it was easier to see any points that were potentially overlapping

- Shape: For both scatterplots, we used circles to represent each data point because it was the most intuitive as medals are round. For the third graph, we chose to incorporate the area as a means to express that more medals were won in a particular medal category for a sport. While the exact area is hard to differentiate, we felt that this was not an important detail as we had only wanted to express that certain people with a certain height won more medals in one category, but we don't really care exactly how many more. Using area also allowed us to utilize the axis for information that we felt was more important (height and sport).

Bar chart design rationale (2nd chart)

- Bar Chart - We chose a bar chart because it is a lot easier to see the difference in median height between bars than it is to see the area difference between circles. It is also fitting since height is used to measure how tall something is and we are using the height of the bar to represent the median height. It effectively shows the median heights of the athletes in relation to gold, silver, bronze, and non-medals in each sport category.
- Scales - Different from the scatterplot, we are using the median height instead of the exact height of each athlete who participated in the sports category. Therefore, we filtered our processed dataset further by selecting sports and calculating the median height of medal-winning and non-medal athletes to see how height influences their ability to win medals.
- Bar width: After several rounds of experimentation, we chose skinnier bars and incorporated spaces to make the graph cleaner and less crowded.
- Colors: Similar to the scatterplot, we used the same color scheme to keep our visualization theme consistent and intuitive to the users. We chose black as the non-medalist color since it is different from all the other colors we used and it fits in with our color scheme
- Legend on the side: We added a legend to the side of the graphs to make it more clear what each color represented.
- Axis Scale - We first started out from 0 to max height in meters. However, this was not intuitive because it is impossible for anyone to be 0.1 meters. Since there is a minimum and average height, there is no point in showing the entire spectrum of height, so we decided to use 1.4 meters as the minimum height on the y-axis rather than 0. We made this trade-off to magnify the difference. Because there is a significant difference between 1.6 meters and 1.9 meters, we decided to zoom in to visualize the data better.

The Story

Our visualization provides a story that height does not necessarily mean success in sports categories. We were surprised at how the scatterplot turned out at first because none of us thought that we were dealing with a discrete dataset. However, the data visualization helped us dissect the problem layer later.

1. Scatterplot gave us a general idea of how height could relate to medal-winning probability. Our data points were extremely concentrated and seemed to distribute equally across the spectrum of heights. As most Olympic athletes participate in 1 sports category, there was more data in the lower number of medals won.
2. That scatterplot led us to think about how to look at a problem from a different perspective. A bar chart of selected sports displaying the median height was clear to see whether height played a factor for the athlete to win a medal in the corresponding sports category. From the chart, we can see that there doesn't seem to be a big difference in height between medalist categories within each sport.
3. Since we only visualized the median of each sport category, we wanted to see a more complete view of the data. Thus, we made a vertical scatter plot separated by medalist and non-medalists. It appears that the distribution of individual athletes with or without medals is quite similar. However, we can see that there is a general height difference among sports.

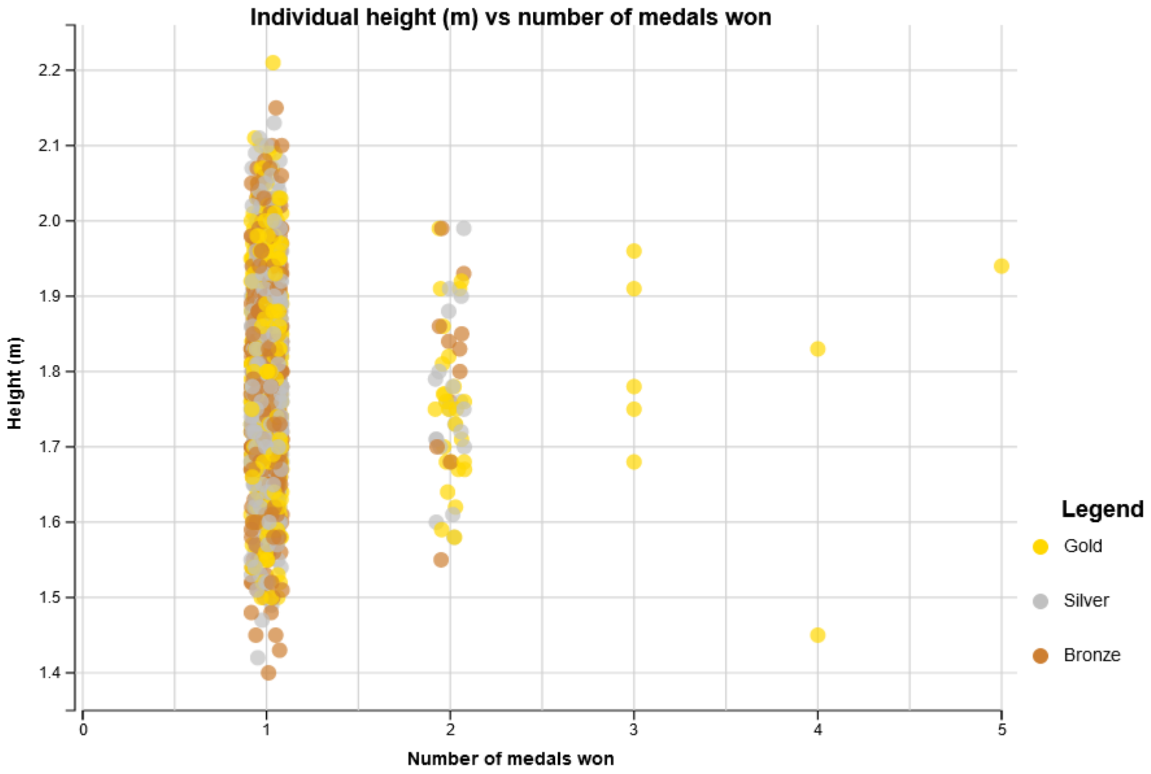
Team Contribution

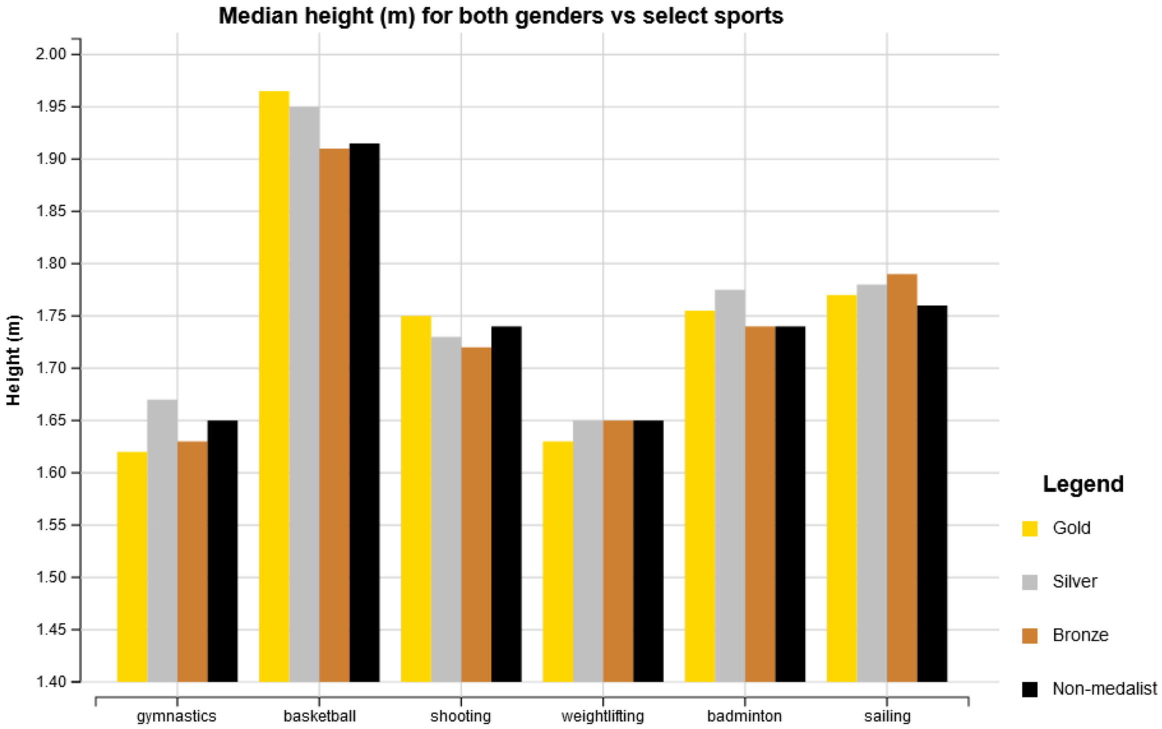
All three of us worked together to come up with the design of the visualizations.

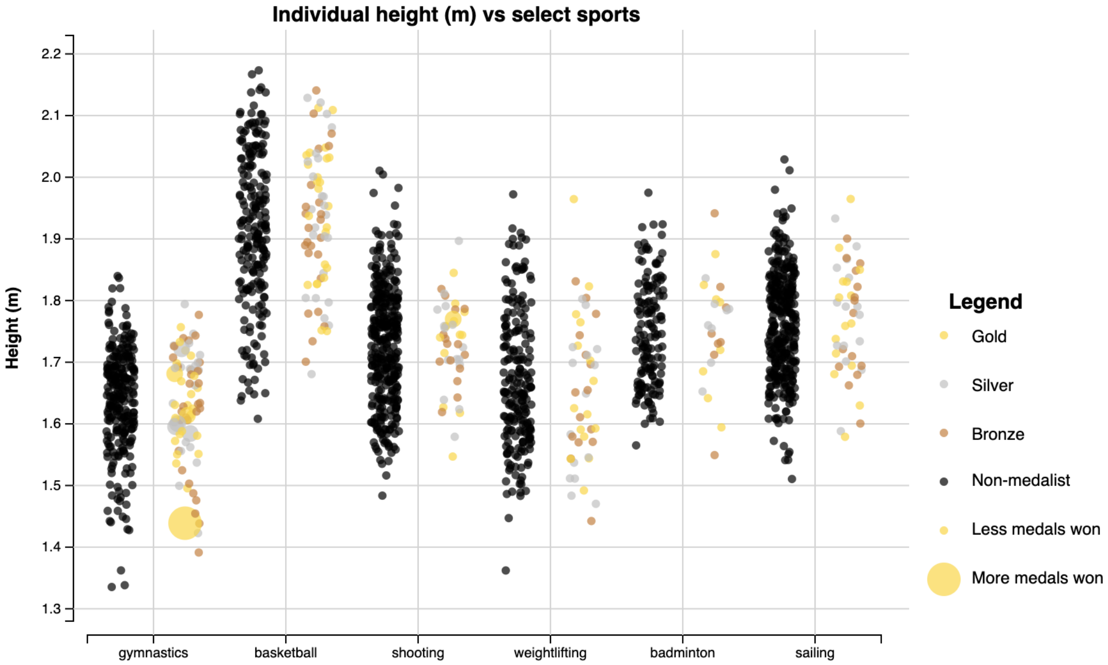
- Emma: Helped code the second and third charts and the legends, with the data processing, and fixing bugs
- Rikki: Helped with ideating through data filtering and processing, drafting up the written report
- George: Helped code the first chart, and fixing bugs, and revising data visualizations for legends, axis titles

Time
~2 hours spent ideating and figuring out the design of our visualizations
~10 hours spent coding the visualizations
~4 hours spend on drafting and finalizing the written report

Appendix


Individual height (m) vs number of medals won
Height (m)
2.2
2.1
2.0
1.9
1.8
1.7
1.6
1.5
1.4
0
1
2
3
4
5
Number of medals won
Legend
Gold
Silver
Bronze


Median height (m) for both genders vs select sports
Height (m)
2.00
1.95
1.90
1.85
1.80
1.75
1.70
1.65
1.60
1.55
1.50
1.45
1.40
gymnastics
basketball
shooting
weightlifting
badminton
sailing
Legend
Gold
Silver
Bronze
Non-medalist


Individual height (m) vs select sports
Height (m)
2.2
2.1
2.0
1.9
1.8
1.7
1.6
1.5
1.4
1.3
gymnastics
basketball
shooting
weightlifting
badminton
sailing
Legend
Gold
Silver
Bronze
Non-medalist
Less medals won
More medals won