Final Project

Max Liu (maxliu), Steven Tang (sjtang), Uma Pradeepan (upradeep)

Contents

Overview of Data

The dataset our group has chosen to explore is the 'Rio 2016' dataset. This dataset, available publicly on Kaggle, contains the official measurement of over 10,000 athletes in over 300 events in the 2016 Olympics. The dataset consists of 3 csv files.

The first file, athletes.csv, has a row for every athlete that participated in the 2016 Olympics. Each athlete's name, nationality, gender, date of birth, height, weight, sport, and quantity of gold, silver, and bronze medals won is recorded.

The second file, countries.csv, has a row for every country that participated. The information recorded is not directly related to the Olympics, but is nonethless interesting to perform data analysis with in conjunction with athletes.csv. For each country, the country name, country code, population, and GDP per capita is recorded.

The third and final file, events.csv, has a row for every event that the athletes participated in. For each event, the name, sport, discipline, gender, and list of venues is recorded.

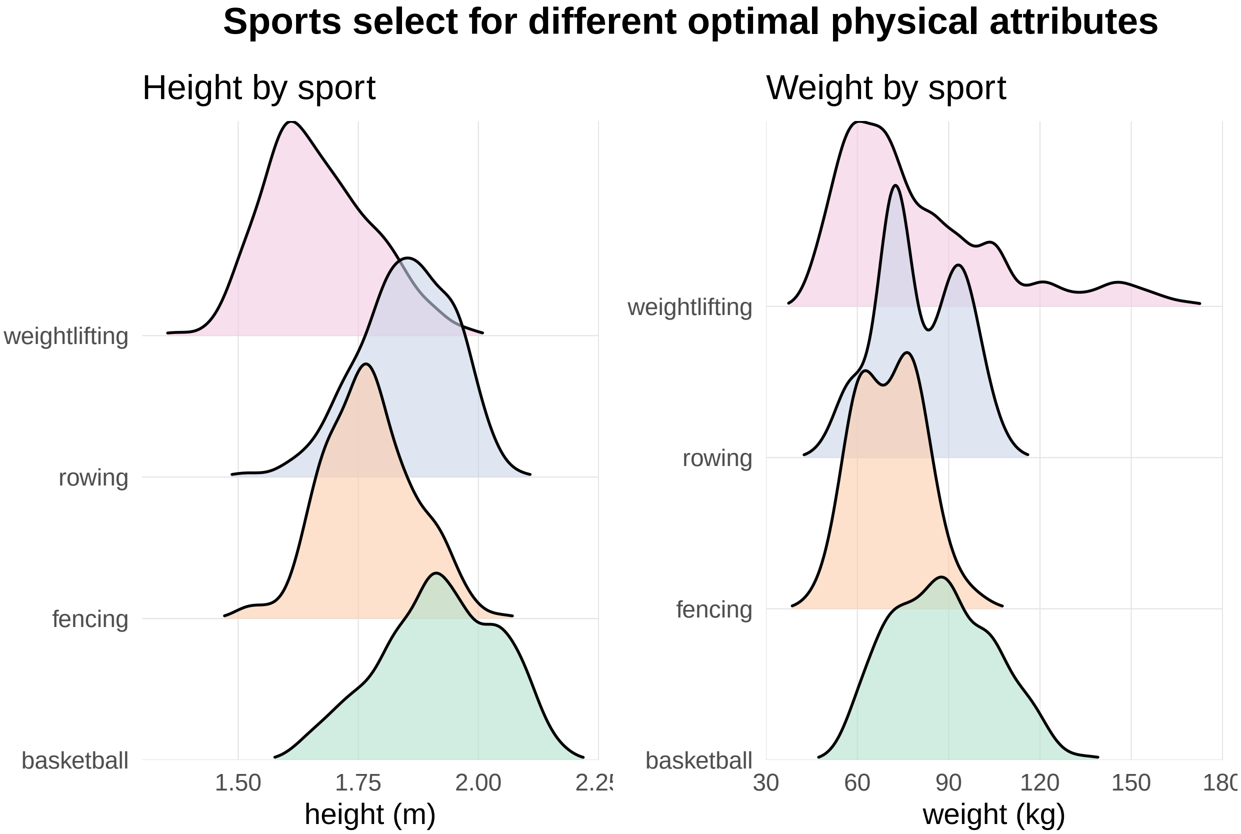
Although there are these 3 files available, in our analysis we focus on the first two files: athletes.csv and countries.csv. We ask questions about how bodily proportions vary between sports, how winning of medals varies across different countries with differeing GDP per capita's, and how performance measured in medals won varies between male and female athletes in different countries. We ask these three seemingly distinct questions in an effort to uncover interesting patterns in how different qualities physical and social qualities affect performance in, arguably, the most high stakes athletic competition in the world.

Question 1: How do Olympic Athletes' Height and Weight Vary by Sport?

Motivation

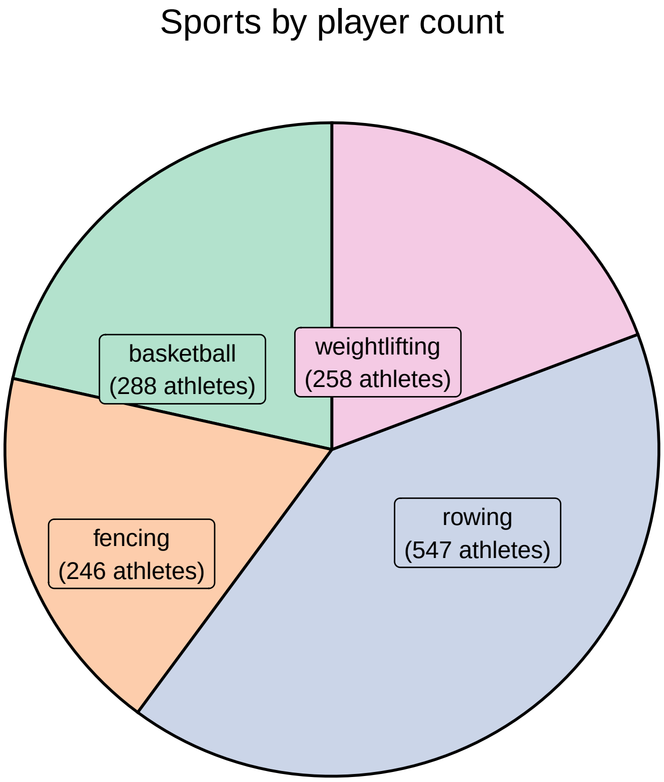
In some sports, it is clear that certain body proportions yield a definite advantage. An obvious example is that basketball selects for height in athletes. In addition to basketball, we want to see whether other high level sports naturally select for different body structures which are optimal for that specific sport. Here we will do a visual analysis of the differences in height and weight distributions of athletes among some selected popular Olympic sports. The sports selected for analysis here are weightlifting, basketball, rowing, and fencing. We select these sports for some variety in the type of movements demanded by the sport.

1.1: What are the Height Distributions and Weight Distributions for each Sport?



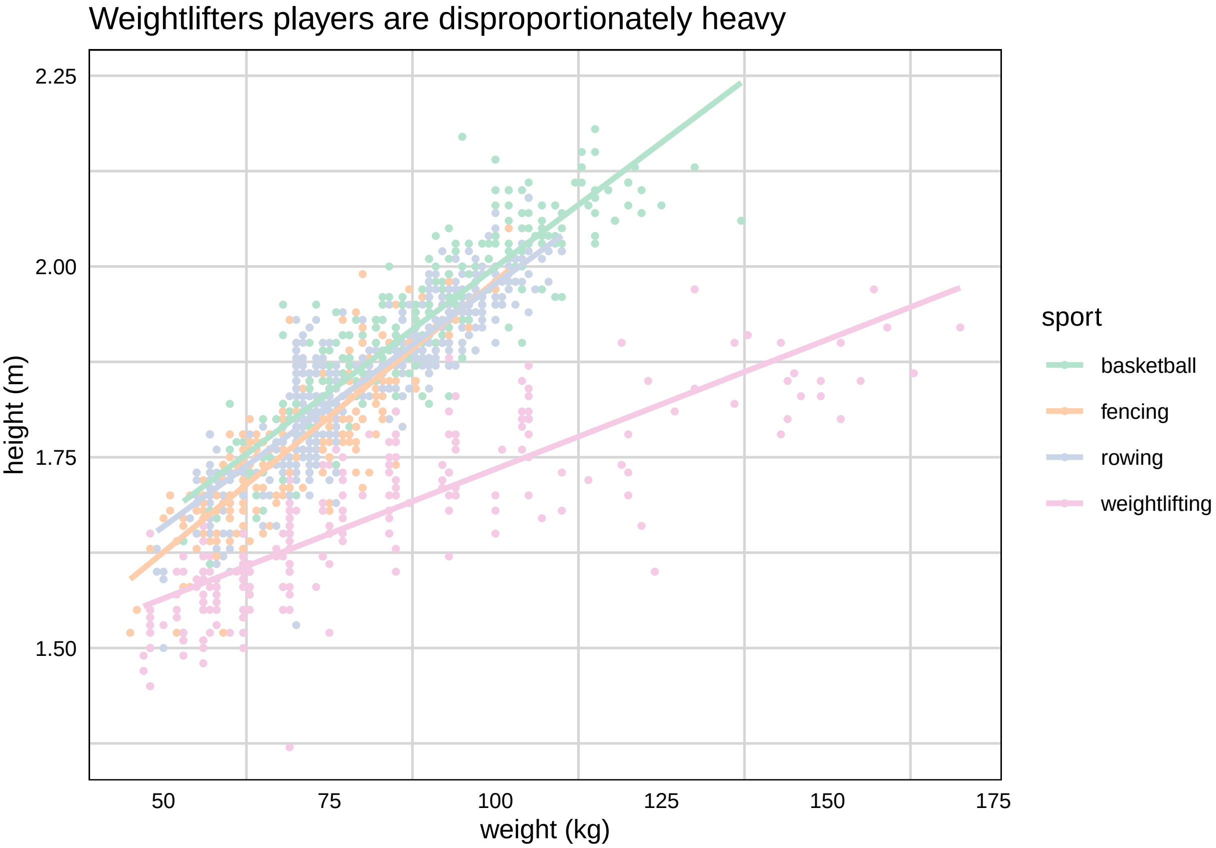
These two plots show the relative distributions of heights and weights of athletes grouped by the sport that they play. The most notable feature is that weightlifters appear to be much shorter with a wider distribution of weights than athletes in the other sports. As expected, basketball players are taller than in other sports. Despite the data containing both male and female athletes, there does not appear to be significant bimodality in most of the distributions, which could indicate optimal height and weight are selected for regardless of sex. However, the weight distribution among weightlifters is super spread out with multiple humps, perhaps due to differing weight classes.

1.2: How does Number of Participants in the Sport Relate to the Observed Height and Weight Distributions?



We expected rowing would have a greater range in height and weight distribution since the sport requiring a greater number of participants. This is because in comparison to the other sports, it has a greater number of participants, which may result in the sport having to be more accepting of a wider variety of body types. However, looking at the graph in 1.1, this did not appear to be the case as rowing exhibits a slightly more limited range than a couple of the other sports in both the height and weight fields, suggesting this hypothesis does not hold. We conclude there is no noticeable correlation between height and/or weight distribution and number of participants in the sport.

1.3: How do the Height-Weight Combinations Vary by Sport?



Despite the differences in distributions of height and weight that we saw in the previous chart, most of the sports we are looking at appear to have similar proportions of height and weight. Even basketball players appear to have similar proportions to other sports (despite being taller). The exception is weightlifters, who are much heavier relative to their height.

Question 2: How does a Country's GDP per Capita Relate to the Number of Medals a Country Wins?

Motivation

An interesting feature of this dataset is that it includes information about the countries participating in the 2016 Rio de Janeiro Summer Olympics, including GDP per capita and population. This is in addition to the data on the number of medals an athlete of a particular nationality won. To answer the question posed in the heading, we will present several plots to answer some smaller subquestions.

These questions include:

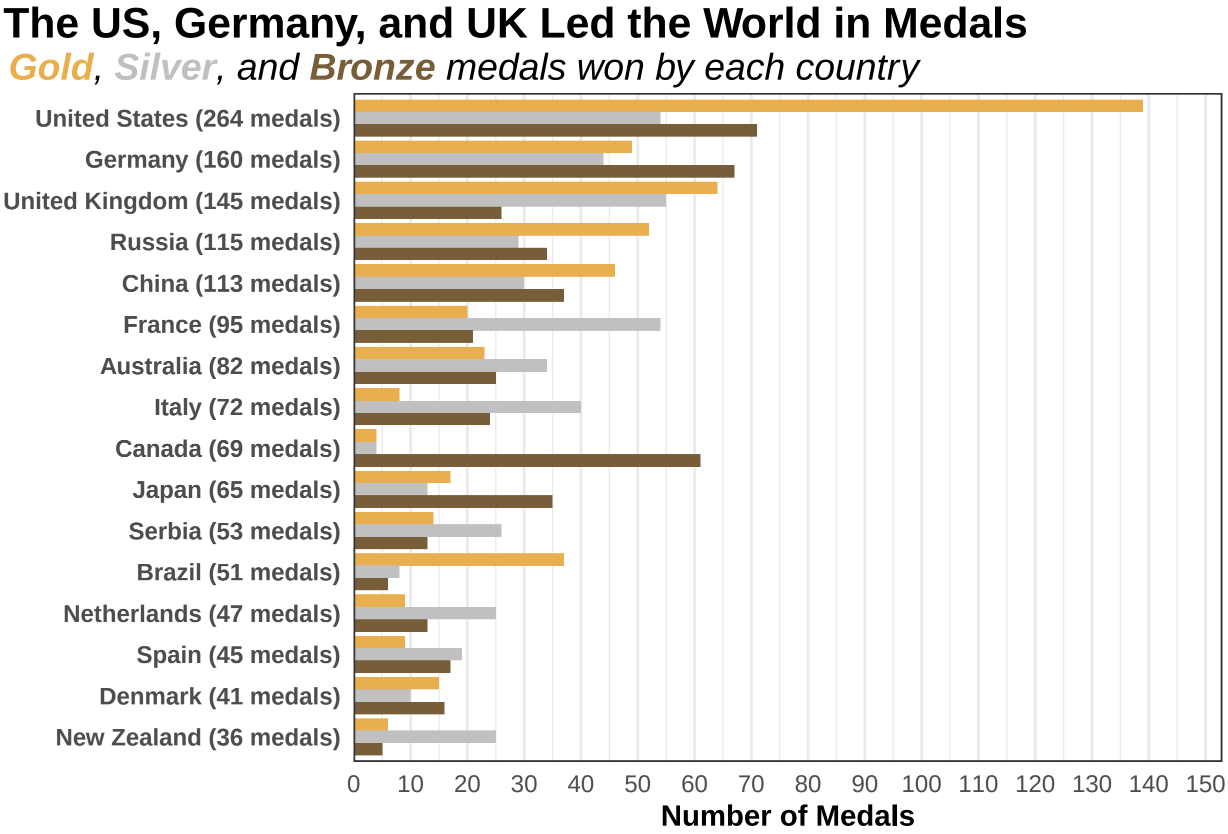
- Which countries won the most medals, and of what types (gold, silver, or bronze)?
- Of the countries that won at least one medal, what is the proportion of countries having a particular medal type as their most abundent winning? Do they have more gold medals, silver medals, or bronze medals?
- What is the distribution of the total number of medals won by a country?
- What is the distribution of countries' GDP per capita?
- Most importantly, what is the relationship between the GDP per capita of country and the number of medals a country wins?

We first explore the amount and distribution of medal types by country.

2.1: Which countries won the most medals, and of what types (gold, silver, or bronze)?

Since the data on athletes includes how many gold, silver, and bronze medals each athlete won, most of the work was done on that data set. We aggregated the number of gold, silver, and bronze medals won by athletes of each nationality, in order to determine the total amount of gold, silver, and bronze medals won by the country. Because nationality was written as a three-letter code like SRB for Serbia or AUS for Australia, and the country data included the full names of the country, we had to join the data together by nationality and country. There was a discrepancy in the three-letter codes denoting nationalities, which allowed the datasets to combine. In particular, we had to manually change the nationality of SCB into SCB so that the country can be matched correctly. Since the medal types were out of order as expected, we reordered the levels of the medals, with “Gold” being the highest distinction, “Silver” being the next highest distinction, and “Bronze” being the lowest distinction, so the coloring can be done correctly at a later time. Since gold, silver, and bronze were among variables of a single observation, because we want a side-by-side bar chart to display this information, we pivoted the data, by duplicating the data for each medal for each country.

After performing these data manipulation, we proceed to plot a barplot of the 16 countries with the most total medals. Countries are labeled with the total number of medals won by the country, which determines the order the countries are listed. Each country has a breakdown of how many gold, silver, and bronze medals won by the country.

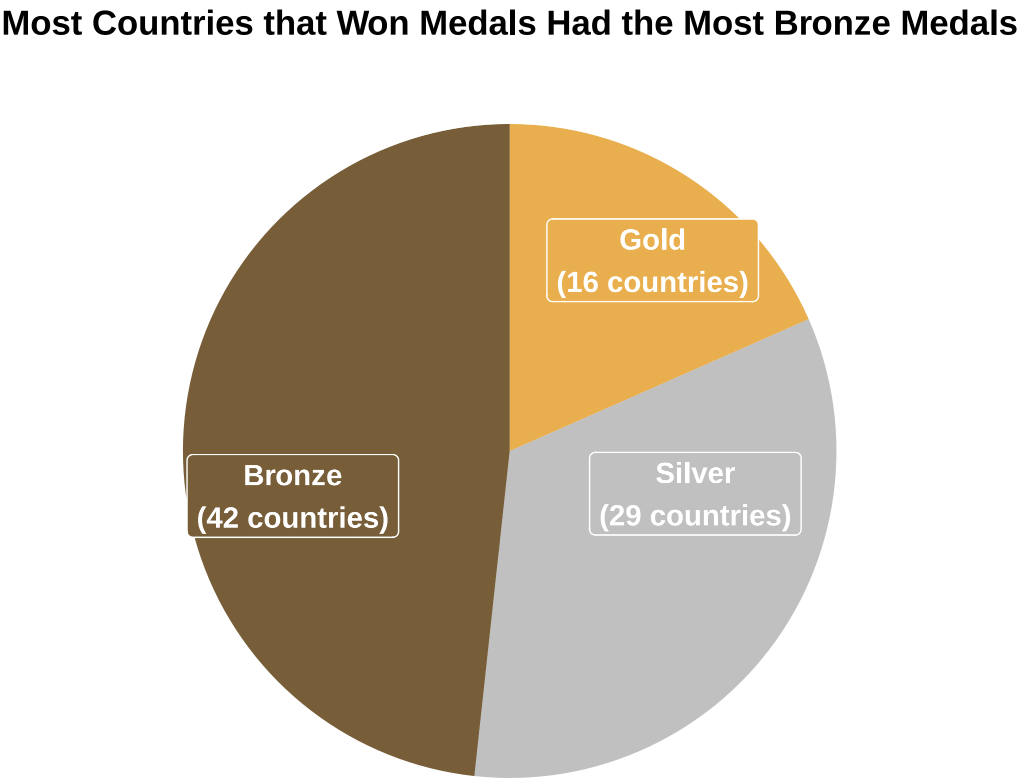


We can see from the side-by-side bar plot above, where each medal in each country is colored by the medal color, that the United States, Germany, and the United Kingdom led the world in total medals (disregarding team sports which may result in the double-counting of certain medals). The United States, in particular, has overwhelmingly the most Gold medals in comparison to other countries. Note that the medal counts encompass the absolute total amount of medals won by all athletes of a country, not by event as medal results are usually reported, so it may contradict what is reported here.

2.2: What share of the country has gold, silver, or bronze medals as its most abundant winning?

We then decided to make a pie chart displaying the proportion of all countries that won a medal that, of their total medal count, had the most gold, silver, and bronze medals.

To construct such a plot, we counted the total number of gold, silver, and bronze medals won by athletes of a certain nationality, expanding the dataset to have each medal as its own observation. We then filtered the dataset to include the medals of each nationality with the most medals. If there is a tie, the greatest prize would be filtered in. For example, if a country has 0 Gold medals, 2 Silver medals, and 2 Bronze medals, then the observation containing the silver medal will be kept, indicating that that country has the most silver Medals in comparison to the other medals they have.

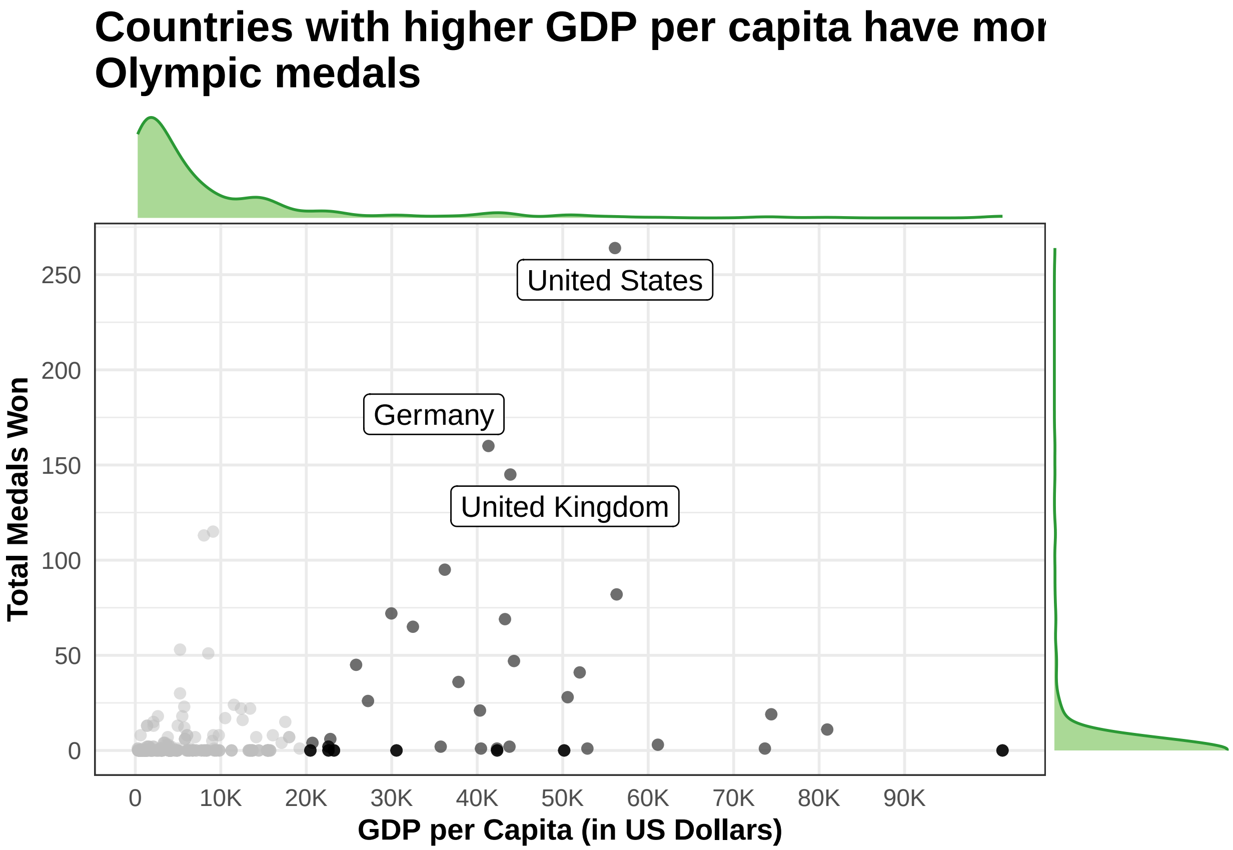


From the pie chart, we can see that the most medals that most countries had are bronze medals, followed by silver, and gold medals. More specifically, most of the medals that the 42 countries had are bronze, most of the medals that the 29 countries had are silver, and most of the medals that the 16 countries had are gold. This suggests that the higher value medals are concentrated amongst fewer countries which are able to specialize in them.

2.3: What is the distribution of and the relationship between the total number of medals country have and the country's GDP per capita?

The last three questions posed in the beginning of the subsection will be answered with this final plot. Using the manipulated data set used for the first plot, which because we combined the athletes and country data, and therefore the dataset contains information about the total number of medals countries won and the country's GDP per capita, we decided to utilize it in the construction of the plot.

To answer the question, the scatter plot plots GDP per Capita (in US dollars) against the total medals won by the country, with marginal plots displaying the distribution of those variables. We highlighted the datapoints whose GDP per capita forms a cluster of similarly behaving countries. These cluster indicate countries with particular high GDP per capita and how prosperous the country's population is and therefore how prosperous the country is.



We can see that both GDP per capita and the total number of medals won is skewed to the right, with more nations having lower GDP per capita and less medals. There are, in other words, more countries with low GDP per capita and less countries with high GDP per capita, and similarly for the total number of medals won.

Also, there is an interesting pattern, wherein the points between a GDP per capita 0 and 20000 is a miniature, similar version of points laying between a GDP per capita of 20k and 90k. More specifically, in both of these prosperity groups (which are highlighted), there are some countries around the groups median which won the most medals. More visibly though is countries with higher GDP per capita have more Olympic medals. This may point to the country's priorities, particularly for countries with low GDP per capita, where in compared to other countries in its group, China and Russia tend to historically perform really well, perhaps due to drive to perform well in the world state and demonstrate patriotism and their competitive spirit. A similar dynamic occurs in countries—US, UK, and Germany—whose GDP per capita is greater than 20k.

As previously stated, countries around the median of the prosperity groups (above and below a GDP per Capita of 20k) tend to gain the most medals, but this is more pronounced for countries with a GDP per capita greater than 20k because those countries, even though there are less countries composing the group, tend to have greater variance and win more total medals.

Question 3: How does the Difference in Performance Between Male and Female athletes Vary by GDP of the Country?

Motivation

The Olympics are expensive for a country to partake in. From the cost of facilities, to training athletes, to funding coaches and equipment, these costs can add up. Historically, Men's athletics have been more popular and better funded than Women's athletics. So, for countries with a bigger budget crunch, we explore whether this translates into unevenness in the performance of men and women by country.

Here we decide to examine the top ten countries by total number of medals won, filtering out all other countries. We do this so as to avoid overcrowding of data in visualizations.

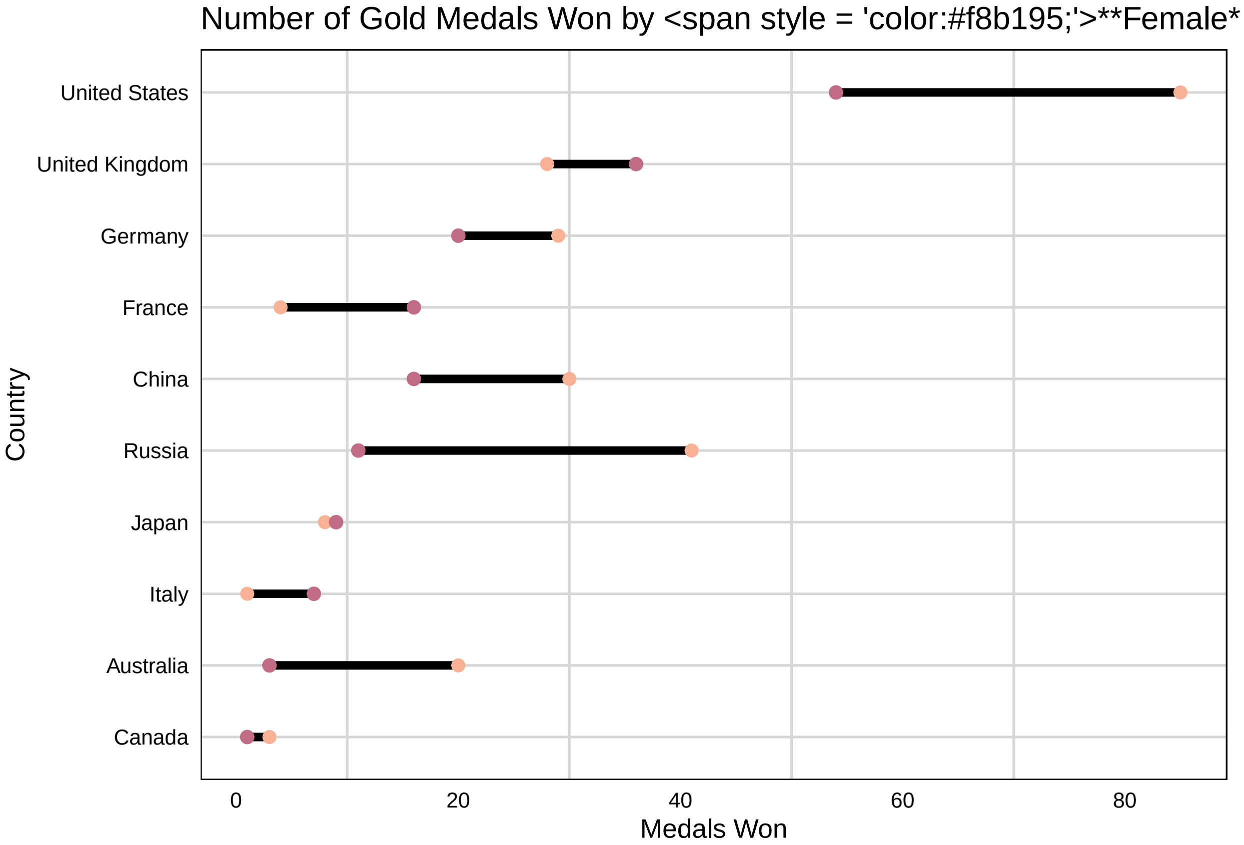
3.1: What are the GDP's per Capita for the Top 10 Countries by Total Number of Medals Won?

GDP per Capita by Country

Country	GDP per Capita (2016 USD)
Australia	56310.963
United States	56115.718
United Kingdom	43875.970
Canada	43248.530
Germany	41313.314
France	36205.568
Japan	32477.215
Italy	29957.804
Russia	9092.581
China	8027.684

Now that we have a table of GDP per Capita by Country. The countries with the highest GDP per Capita are Australia and United States. Meanwhile, the countries with the lowest GDP per Capita are China and Russia. Next we will examine the actual difference in Men's and Women's performance, keeping in mind the GDP's per capita of the country they are from.

3.2: Do Men or Women Perform Better in the Top Ten Gold Medal Earning Countries?

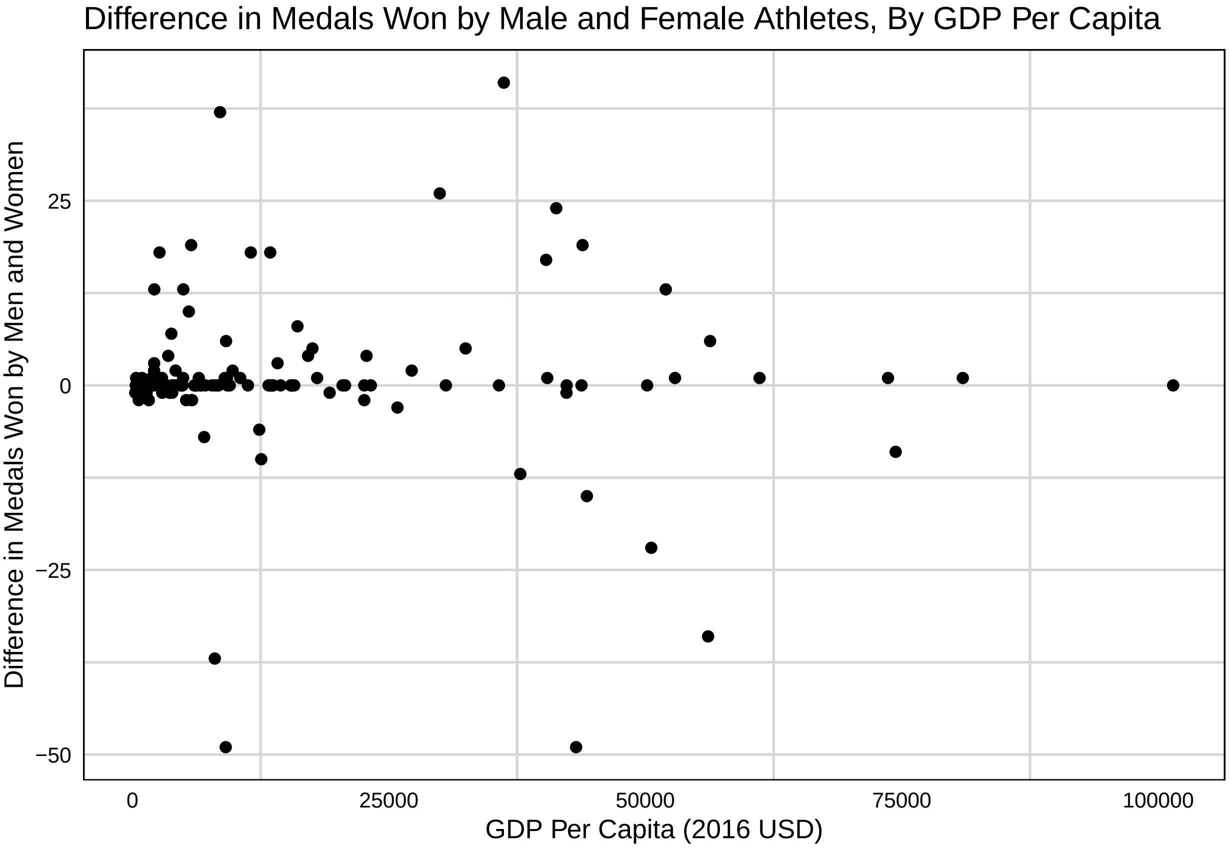


Looking at this dumbbell, plot it appears that GDP per capita does not affect the difference in performace (measured in gold medals earned) in male and female athletes. In fact, it's interesting to note the the two countries with the largest difference in number of gold medals earned by male and female athletes are USA and Russia, which glancing back at the table mapping countries to the GDP per capita's we see that USA has on of the highest GDP per capita's while Russia has one of the lower GDP per capita's.

Another fact that it interesting to note is that the countries with larger differences; that is USA, Russia, Australia, and China, women are earning more gold medals than men.

3.3: What are the GDP's per Capita for the Top 10 Countries by Total Number of Medals Won?

We have seen the top ten gold medal earning countries, but for a more comprehensive perspective, it'd be beneficial to see how the difference in male to female medal earning varies by GDP per Capita.



Looking at the plot above, again it appears that there is no strong trend between difference in medals won by female and male athletes and GDP per capita. Perhaps the costs associated with the Olympics, although they are high, are not a sizable enough ratio of the GDP of the country to affect the level of investment in the Olympics. On the other hand, it could be possible that countries are investing significantly different amounts in the Olympics based on their GDP per capita, but that the amount of investment does not translate very strongly to performance in the athletes. Whichever of these two listed reasons, or some other reason, for why the GDP per capita is not strongly correlated with the difference in medals won by athletes, it would be interesting to do additional exploration on how GDP per capita affects level of investment in the Olympics, and how level of investment in the Olympics affects performance. Additionally, coming back to the theme of prioritization of male or female athletics programs, it would be interesting to see if there are difference in performance of men's and women's teams in leagues lower than the Olympics where budgeting can be expected to be tighter and the amount of budget available has greater potential to affect the split of investment in male and female athletes.

Conclusion

In conclusion, we have shown using our visualization that there is a correlation between country GDP and the number of medals won. However, there isn't much relation between GDP and whether male or female athletes win more medals. We also showed that different sports have different optimal height/weight distributions which are optimal. Some limitations on our data include only having summer Olympic events and only having events from 2016. If we had data from other yers, for example, we could look at trends over time.