Analyzing the Olympic Results

Stephanie Cao, Brandon Huang, Hannah Koschmeder
Economics and Data Science (73-265)
Professor Gasper

Table of Contents

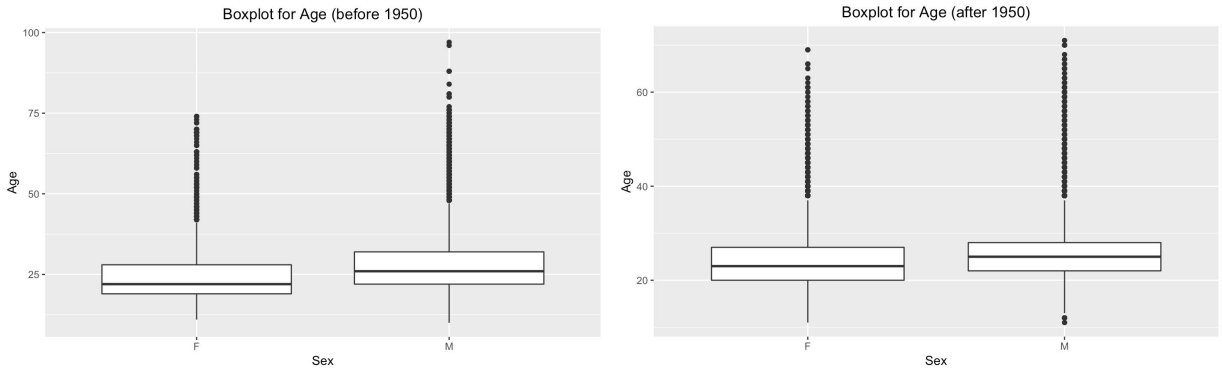
Motivation of Topic

The Olympic Games is an international sports competition featuring athletes from all over the globe. Occuring every two years, the Summer and Winter Olympic games have gained huge popularity amongst people from different nations. People not only want to watch the best athletes in the world come together and compete, but also want to cheer for their own country. The main motivation for our study stems from this worldwide Olympic spirit. Furthermore, this topic focuses on what Olympic fans are interested in as well - what makes certain countries strong in certain sports; what makes athletes successful; and predicting the next medal winners.

Description of Data and Exploratory Data Analysis

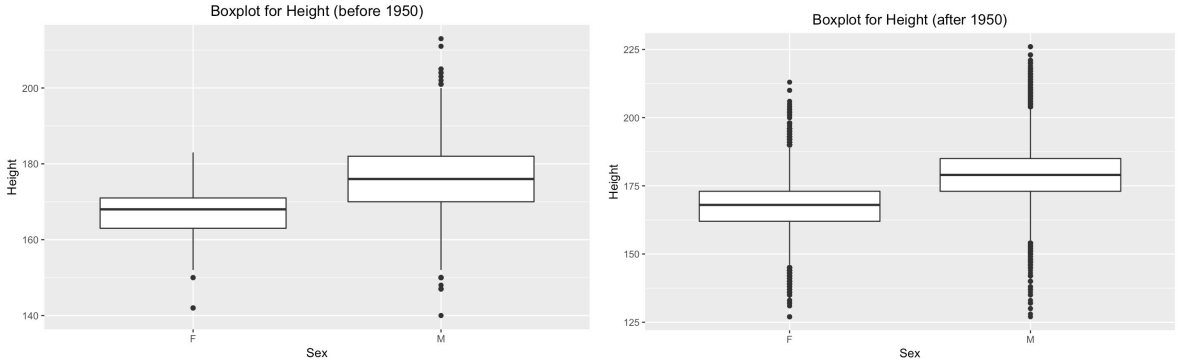
Our Olympics dataset includes 120 years of data spanning from Athens 1896 to Rio 2016. It contains 271,116 instances, one for an athlete per event. Each instance contains the following variables: Name, Sex, Age, Height, Weight, Country, Olympic season and year, City, Sport, Event, and whether or not the athlete won a medal. Our data was collected from Kaggle, which gathered data from sports-reference.com.

For our initial exploratory analysis, we first plotted boxplots to investigate into the distributions of age, height, and weight in the Olympics. We split the box plot distributions for data before 1950 and after 1950 to account for potential differences in athlete characteristics over time. The following is the distribution for Age of athletes split between male and female for both results before and after 1950.



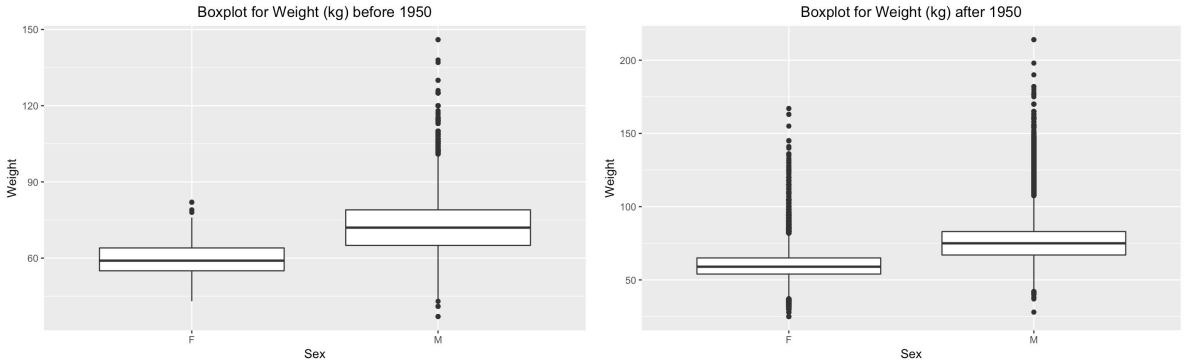
According to the boxplots, the average age for athletes before 1950 are around 23 for females and 25 for males, and for athletes after 1950 are also similarly around 23 for females and 25 for males. Thus, the average age for athletes seem to stay relatively similar over time. However the range of age for athletes before 1950 is much larger than athletes after 1950. The box plot for before 1950 shows that female athletes used to compete as young as 12 and would range up to 75 years old. As for males, athletes also would compete as young as 10 and would range up to around 90 years old. On the other hand, the age range for female athletes after 1950 range from 12 to 70, and for male athletes it ranges from 12 to 72. This may indicate that over time the Olympics has become increasingly more competitive, thus calling for better conditions to compete, and oftentimes age is major factor to the performance of an athlete.

Next, we looked into the distributions for height.



While the average heights for athletes appear relatively similar before and after 1950, the range of heights after 1950 is much larger than its counterpart athletes before 1950. Female athletes before 1950 had ranges from 142 to 182cm and male athletes had ranges from 140 to 213cm. As for after 1950, female athletes have ranges from 126 to 210cm and male athletes have ranges from 127 to 225cm. The larger range of heights may point to the increased introduction of more sports after 1950 such as archery, boxing, and cycling.

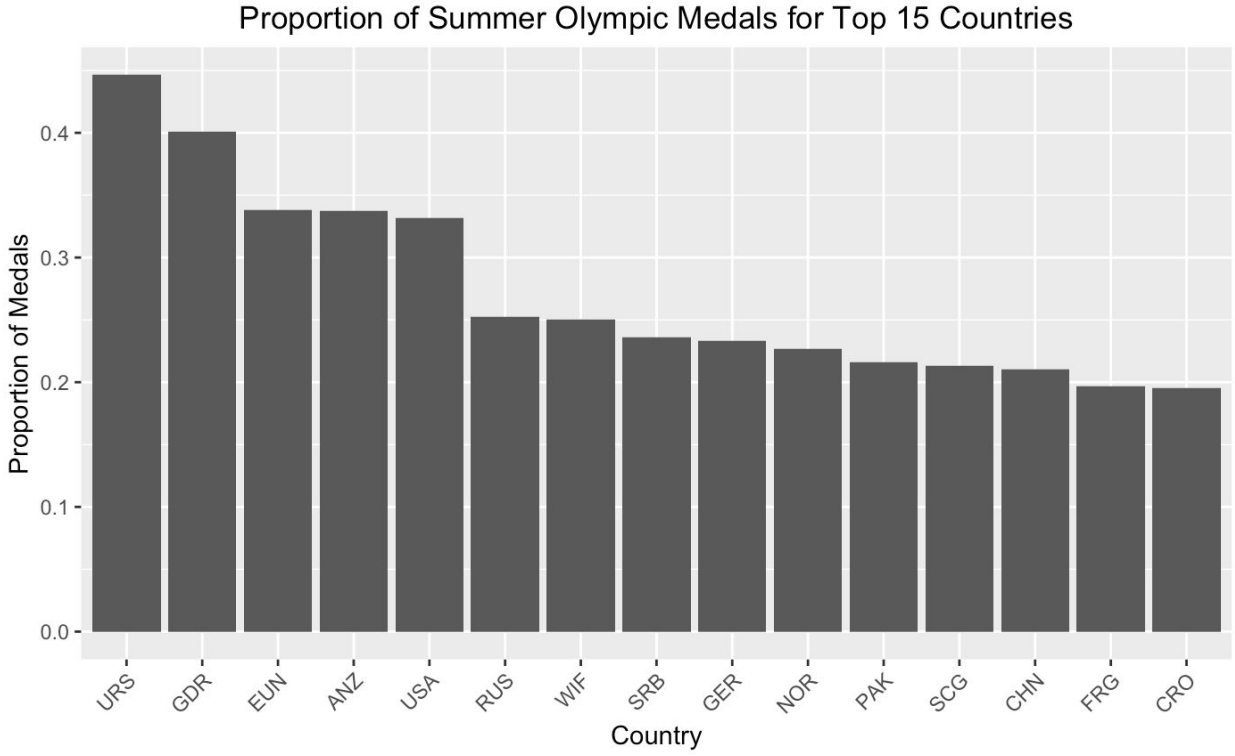
Finally, the following are the distributions for weight.



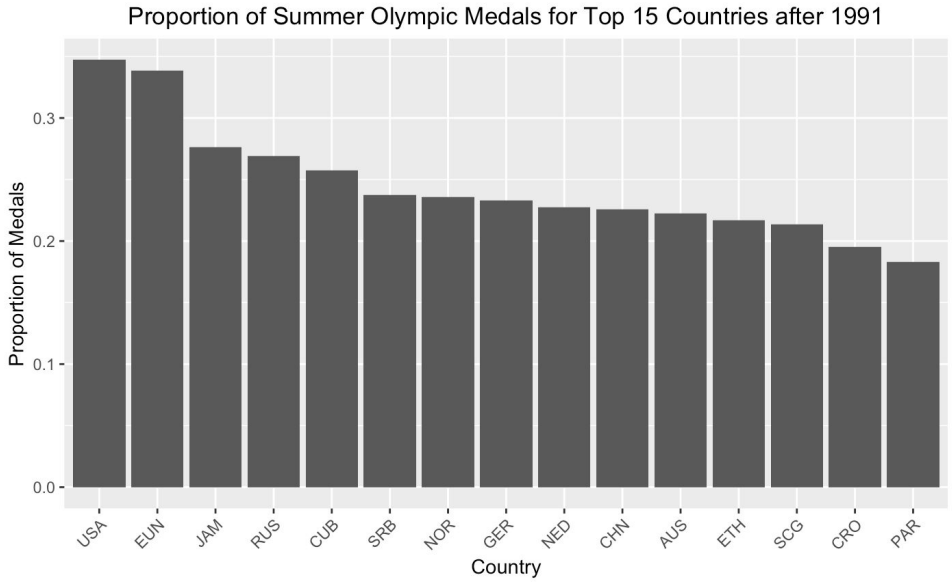
Similarly to height, the weight distributions for athletes before 1950 and after 1950 are relatively similar, but the ranges are much larger for both female and male after 1950. The range for female athletes before 1950 is around 45 to 80 kg and for male athletes it is around 40 to 150 kg. For athletes after 1950, the female weight range lies as low as 30 to 170 kg, whereas the male weight range lies from 40 to as high as 215 kg. The larger weight range may also indicate the larger availability of sports after 1950 that offers a greater variety of athletes. A larger spectrum of sports offered in the olympics may demand different types of athletes that typically lie in a certain weight class.

For our next exploratory data analysis, we investigated into the top 15 countries with the highest proportion of medals for both summer and winter olympics. We grouped the countries together and calculated the proportion of medals by dividing total event medals by total events attended for. The following graph is the top 15 countries for proportion of summer olympic medals. Here is the key for each country (URS = Soviet Union, GDR = East Germany, EUN = Unified Team, ANZ = Australasia, USA = United States, RUS = Russia, WIF = West Indies Federation, SRB =

Serbia, GER = Germany, NOR = Norway, PAK = Pakistan, SCG = Serbia and Montenegro, CHN = China, FRG = West Germany, CRO = Croatia).

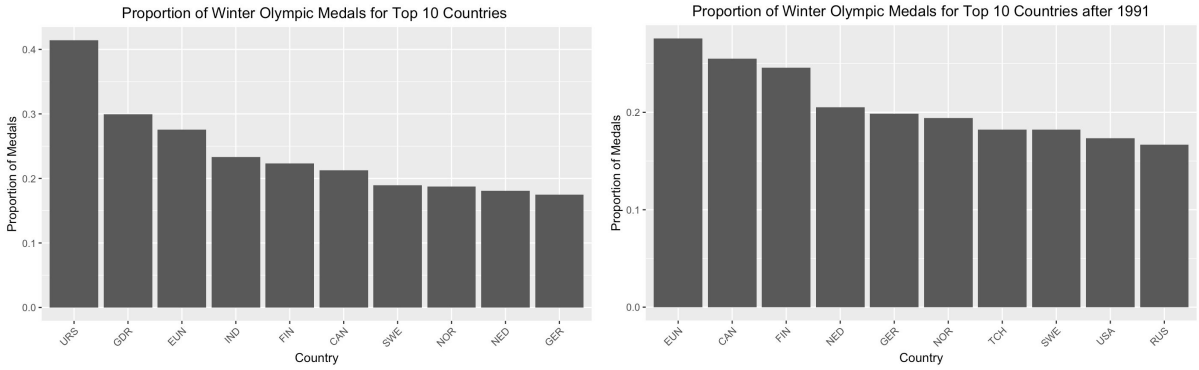


According to the bar chart, a lot of leading countries in the summer olympics include pre cold war countries such as Soviet Union, East Germany, Unified Team, and the West Indies Federation. Although most top countries shown in the chart do not exist today, it is interesting to see that the countries with the greatest proportion of olympic medals were also countries that were among the most powerful at the time. Since the bar chart includes many countries that do not participate in the olympics today, we looked into summer olympic results after the end of the cold war, which is 1991. Below is a bar chart of the top 15 countries for proportion of summer olympic medals after 1991. Here is the key for each country (USA = United States, EUN = Unified Team, JAM = Jamaica, RUS = Russia, CUB = Cuba, SRB = Serbia, NOR = Norway, GER = Germany, NED = Netherlands, CHN = China, AUS = Australia, ETH = Ethiopia, SCG = Serbia and Montenegro, CRO = Croatia, PAR = Paraguay).



After filtering to countries after 1991, the bar chart looks much more different with USA leading and countries that did not make the top 15 on the previous chart such as Jamaica, Cuba, Netherlands, Ethiopia, and Paraguay.

As for the Winter Olympics, below is the bar chart for top 10 countries with greatest proportion of medals for the whole olympics history and after 1991.

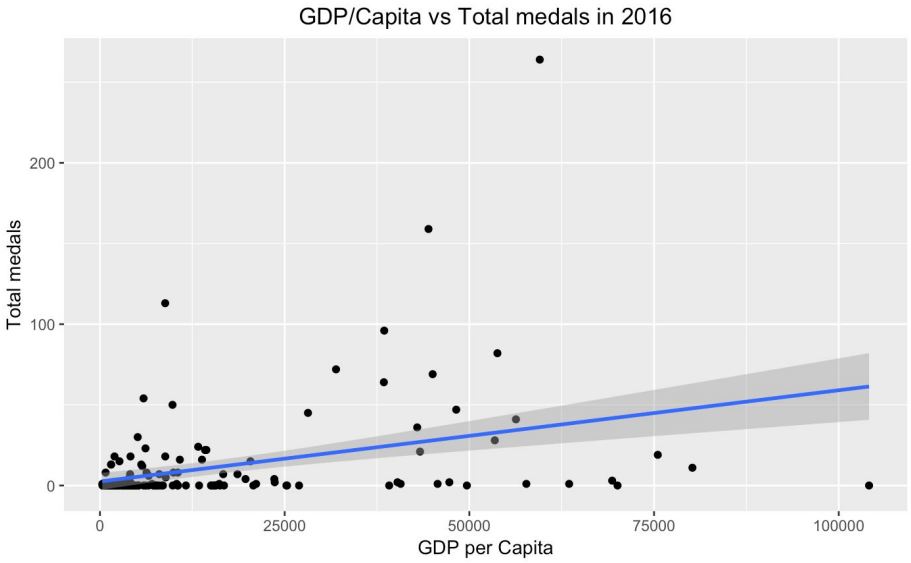


Keys (IND = India, FIN = Finland, SWE = Sweden, TCH = Czechoslovakia)
Similar to the summer olympics, pre cold war countries such as the Soviet Union and East Germany lead in terms of medal performance in the olympics history, but after the cold war, countries such as Canada and the United States emerged to the top 10. Another thing to note is that since it is the winter olympics, both bar chart also shows that most top countries happen to be in areas with colder climates.

Predictability of GDP per Capita on Winning a Medal

The next part of our analysis is on finding out the effects of GDP per capita on winning a medal. The purpose of this study is to see if the average wealth of a country affects the ability of an athlete to win. We focused on the Rio Summer 2016 Olympics results since our GDP data is for

the year of 2017. We ran a linear regression model between GDP per Capita and Total Medals won for Rio Summer 2016. Below is a plot of the relationship between the two.



Although there appears to be a positive correlation between the two variables, it seems to be very weak. The regression results are the following: R-squared of 0.134, and coefficient of 0.0005655 for GDP per Capita with a statistically significant estimate. Although the coefficient estimate is statistically significant, GDP per Capita alone does not explain a lot of the variation in the dataset. Also the residual plot shows that the residuals are heteroscedastic, so the linear regression model does not appear to be an appropriate model for predicting total medals. We also ran a logistic regression predicting whether an athlete wins a medal by GDP per Capita, and although the accuracy is 85%, the kappa was 0. Thus, it was predicting not winning a medal for every athlete and shows that it has little predictability of a country winning a medal.

Predictability of Country on Attaining a Medal

Our next variable analysis is finding out the predictability of country alone on an athlete or team attaining a medal. The purpose of this analysis is to see whether athletes from certain countries are in really more competitive in some sports than athletes from other countries. For this study, we analyzed for two very different sports, Basketball and Sailing. We ran a logistic regression predicting medal on just the Country variable for both men's and women's basketball with thresholds of 0.5. For men's basketball, the logistic regression model had an accuracy of 0.9027, Kappa of 0.6616, and AUC of 0.9414. For women's basketball, the logistic regression model had an accuracy of 0.8686, Kappa of 0.6842, and AUC of 0.9381. Both logistic regression models had high Kappa and accuracies, indicating that the model is not just predicting one value for all instances and has very high predictability with just the country variable alone. Next, we ran a logistic regression model for sailing just to see if the predictability of country is also effective for other sports. For men's sailing, the logistic regression model had an accuracy of 0.9025, Kappa of 0.6328, and AUC of 0.9262. For women's sailing, the model had an accuracy of 0.8683, Kappa of 0.1847, and AUC of 0.8027. Both models also had high

accuracies and greater than 0 Kappa. However, women's sailing had a much lower Kappa than for men's sailing. Although our analysis is limited to just basketball and sailing, our results indicate that country is only effective in predicting athletes or teams winning medals for some but not all sports. This is understandable since some countries are known to have won a lot of medals in some sports whereas some sports, usually not as popular ones, have more mixed results.

Effects of Country and Population on Total Medals

Attending the Olympics with competitive athletes, is prestigious in itself, but often only the final total of medals is important to the powerhouse countries who consistently perform well. What are the most important factors to predict the final medal count for each country?

Some important factors may include: population, GDP per capita, government funding and corruption, and availability of facilities. Population, and the composition of the population, seem directly correlated to success. The larger a population is, the larger the pool of athletes is to choose from, and thus more chances of those athletes being successful. A younger population, or a population with large, well-spaced generations will likely also contain successful athletes. A higher GDP per capita, or average wealth per person, allows individuals to have access to more sports because some sports are more expensive than others. However, GDP per capita doesn't always tell the full story if the distribution of wealth is concentrated among a small group. Olympians can be required to be amateurs - unpaid for their sport, fully-funded by the government, or funded by sponsorships. A more financially secure athlete can fully focus on their sport, compared to an athlete that might have to seek regular employment.

Although all of these factors might be relevant to predict the final medal count, they aren't all easily accessible. The two sources that were most accessible were population estimates from the UN since 1950, and GDP per capita data from The World Bank. One issue with this data is how different it is from Olympic data. Olympic data has team codes that haven't changed throughout its history, only new ones have been added, but countries have changed immensely over time, and UN estimates reflect this. Therefore, the population estimates are as true as possible to the countries that existed at the time.

The predictor with a large R^2 is Country, but it's messy to include each country as a factor, and not all of the country's predictors are significant because there is such a large difference in medals each year. The range of total medals is 0 to 442, and the most common number of medals is 0. This means that the majority of the coefficients are very close to 0, which may make sense for the majority of countries that consistently don't win many medals. However, negative coefficients are a definite extrapolation because a country cannot have a negative medal count. A flaw with including country as a factor is that it says more about the fact that a country's advantage, or disadvantage will carry on through each Olympic Games, than what factors made that country successful. This simple linear regression is promising, and it's almost all one needs.

After seeing the power of country history, it's easy to understand that using the total number of medals won at the previous Olympics would also be a powerful predictor. In fact, lag from the two previous Olympics have extremely high correlations with the total number of medals. Surprisingly, the variable of interest, population, has a low correlation with total number of medals, and doesn't add much to the final model that isn't captured by the country or the previous number of medals won.

Similar to the common log transformation for population, the other variables' residual plots and distribution plots reveal that the model might benefit from their transformation. A Box-Cox function can be used to see an applicable transformation, but all data points must be greater than 0, even if just by a small decimal. The Box-Cox suggests that all variables of interest, medals from the prior Olympics, medals from 2 prior Olympics, total number of medals, and population should be log transformed. The positive impact of these transformations, or at least of the transformation on total medals, can be seen in the extreme reduction of the Akaike information criterion from around 17,000 from the model with country and log(population) to 10,000. While Box-Cox, and the distribution of Total medals suggested a transformation was necessary, the best transformation came from trial and error, and ended up being the fourth root.

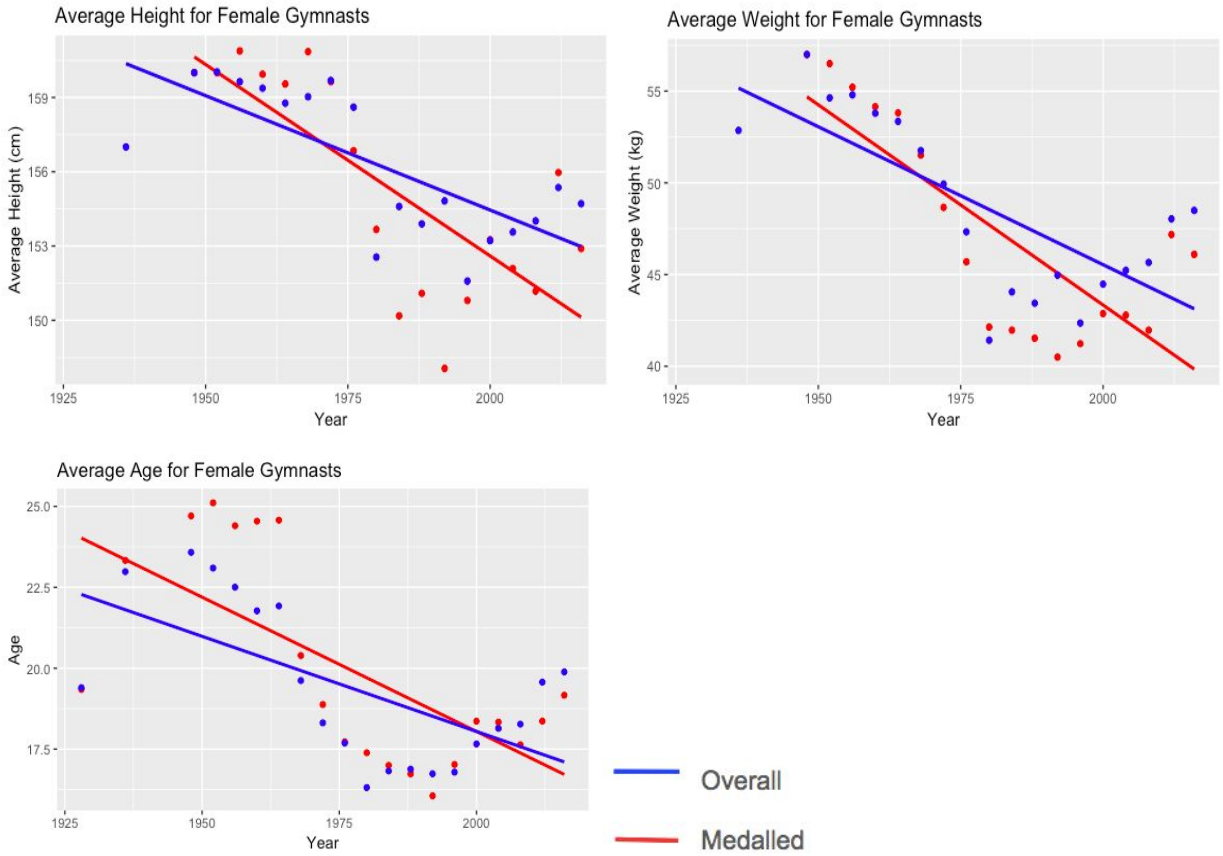
Analysis of 3 Popular Olympic Sports

Our study analyzed Gymnastics, Swimming, and Beach Volleyball as they are, according to Statista, "fan favorites". For each, we studied: 1) variables: age, height, weight, and country and how these variables changed overtime and affect the chances of an athlete winning a medal; 2) logistic models to predict athletes' success at the Olympics; and 3) the effect of Olympic experience. Our analysis on all three sports indicate that Country has the greatest effect on whether an athlete wins a medal at the Olympics or not.

Women's Gymnastics

Analysis on Variables

Looking at just the scatter points, we observe an U-shaped curve and that utilizing a linear model may not be reasonable for the entire data. Since we are more interested in recent data as they are more reflective of current gymnasts and gymnastics rules, we decided to only study the data after 1980s, which also, is a more reasonable for a linear model.



Predicting Athletes chances of winning a Medal

To predict an athletes chances of winning a medal, we utilized a logistic regression, in which the independent variables were age, height, weight, country, and event. The AUC of our model was 0.9446. We analyzed two Olympians: Simone Biles and Oksana Chusovitina.

We compared Biles, who won 5 medals in Rio 2016, to the average features of other gymnasts who have medalled. Surprisingly, none of her features are “average.” We conclude that while these features do contribute to the success of a gymnast, there are other qualities as well.

Next, we predicted Oksana Chusovitina’s chances of winning a medal on the vault in 2020. Chusovitina is the only female gymnast who has competed in seven Olympic games. She competed in the 2016 Rio games and placed 7th on vault. Our logistic regression predicted that she had a 4.85e-09 chance of winning a medal as she will be 44.

The Effect of Olympic Experience

The Olympic Number and its corresponding average number of medals are in the table below. We also included the number of gymnasts to show relative population size.

Olympic Number	Average Number of Medals at that Olympics	Number of Gymnasts
1	0.496	113
2	0.389	113

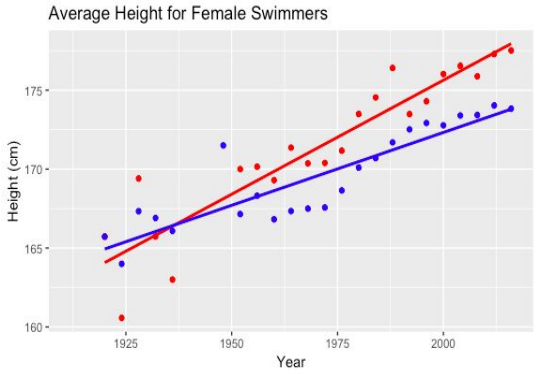
3	0.267	15
4	0	2
5	0.5	2
6	0	1
7	0	1

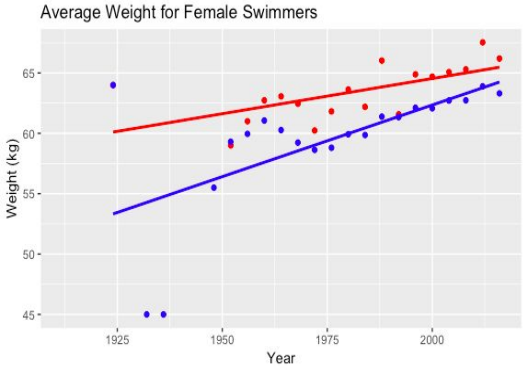
We see a general decline in the average number of medals. The best comparison is between the first and second Olympic where the number of gymnasts are the same. Although the 5th Olympic does have the highest average, there is only two data points, and one of the gymnasts (Chusovitina) during their 5th Olympics medalled. Essentially, the best indication is between the first two points, in which we are comparing the same athletes.

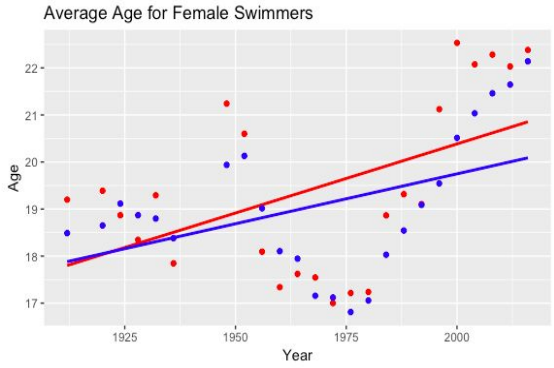
Swimming

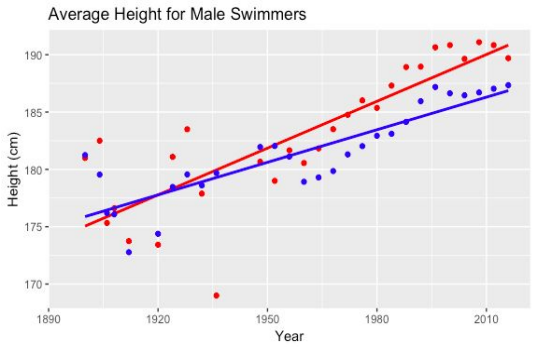
Analysis on Variables

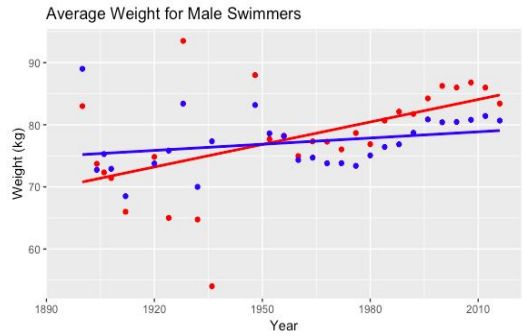
The average height and weight for overall and for those swimmers who won medal(s) over the years has increased. Most of the swimmers who medalled are, on average, taller and heavier. Unlike average height and weight, average age has not had a positive increase every Olympic year. From the years 1948 to 1976, there was decline in average age over the years, and after 1976, there was a steady increase in average age. Between 1948 and 1976, the average age for swimmers who medalled was less than the overall average; on the other hand, for the years after 1976, the average age for for swimmers who medalled is greater than the overall average.

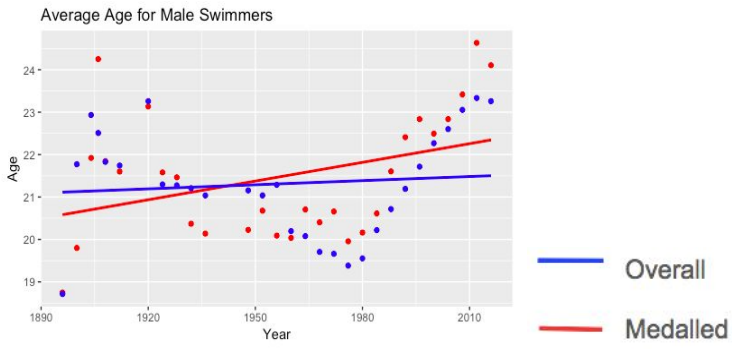












Athlete Analysis

For our athlete analysis, we studied two male swimmers - Michael Phelps and Joseph Schooling. We utilized a logistic regression, which had an AUC of 0.9, on Age, Height, Weight, Country, Sex, and Event.
We compared Michael Phelps (the most decorated Olympian in history) and Joseph Schooling (the “underdog” who beat Phelps in the 100 meter butterfly in 2016). For 2016, our model predicted that Phelps had a 70% chance of making the podium, and Schooling had a 1% chance. Essentially, Team has a significant effect in our model. We tested that if Joseph Schooling was on team USA, his chances would increase to 57% in 2016.

The Effect of Olympic Experience

For women’s swimming, returning Olympians for the second and third time have a higher percentage of winning a medal. Comparing the first and second Olympics, we notice that athletes on average perform slightly better than their first. Third-time Olympians, on average, have the highest medal count; however, it is difficult to draw conclusions because the number of athletes have declined. The average number of medals for the first four Olympic Games is relatively the same for men’s swimming, with the second Olympics having the highest at 0.34 medals per game.

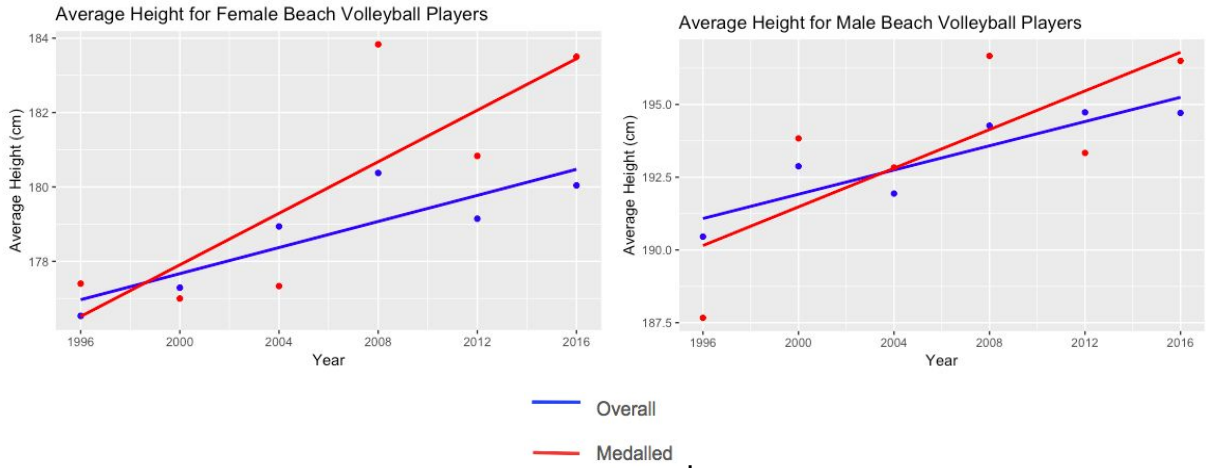
Average Number of Medals at that Olympics (Red for Female. Blue for Male)

Olympic Number	Average Number of Medals	Number of Swimmers	Average Number of Medals	Number of Swimmers
1	0.326	973	0.29	1448
2	0.367	973	0.336	1448
3	0.370	238	0.333	369
4	0.300	50	0.254	71
5	0.333	9	0.700	10
6	0	1	0	2

Beach Volleyball

Heights of Beach Volleyball Players

Beach volleyball became an official sport in 1996. The average height beach volleyball players has strictly increased over the years. For most of the Olympic years, medalled volleyball players are, on average, taller than the overall population.



Effect of Olympic Experience

For women's, the average number of medals has decreased. However, the number has increased for the third of fourth time. For mens, their second Olympics, on average, is better than their first. The fifth Olympics has the highest average number of medals because there was only one athlete who ended up medalling. This increase is because there are fewer athletes that have participated in that many Olympics, and those returning athletes have done very well at prior Olympics to continue to play.

Average Number of Medals at that Olympics (Red for women's. Blue for men's)

Olympic Number	Average Number of Medals at that Olympics	Number of Volleyball Players	Average Number of Medals at that Olympics	Number of Volleyball Players
1	0.17	66	0.15	67
2	0.15	66	0.16	67
3	0.29	17	0.14	22
4	0.67	3	0.25	4
5	0	1	1	1

Conclusion

This analysis has provided greater insight into not only the Olympic dataset, but also into the politics and policies of the Olympic Games and the Olympians that achieve glory. It will be a testament to these models to see if our predictions for individual athletes in 2020 will be correct, and it will certainly provide more viable data to continue our tests.

In the end, the best model to predict total medal count is from a country's past performance, but there is still a missing factor, given that there is such a clear divide between the countries that consistently perform poorly and the ones that perform well. The country's success is a shallow method of prediction because it doesn't accurately address what made this country successful. Through our analysis, we have determined that the average wealth of a country does not correlate to greater success in winning medals. Additionally, some countries are historically better at some sports than others, so it is easier to predict the winners of some sports with reasonably high accuracy. Overall, the country is the top factor in our predictions, and more data analysis from different sources, as well as historical research could point to interesting reasons why, or how to build a great Olympic team from the ground up.

Sources:

Statista:
Techbargains.com. (n.d.). What is your favorite Olympic event?. In Statista - The Statistics Portal. Retrieved December 12, 2018, from https://www-statista-com.proxy.library.cmu.edu/statistics/237282/favorite-olympic-events-among-americans-in-2012/.

Olympic Data:
https://www.kaggle.com/heesoo37/120-years-of-olympic-history-athletes-and-results

UN Population Data:
https://population.un.org/wpp/Download/Standard/Population/

UN Country Codes:
https://unstats.un.org/unsd/tradekb/Knowledgebase/50377/Comtrade-Country-Code-and-Name

World Bank:
https://data.worldbank.org/indicator/NY.GDP.PCAP.CD?view=map