**An In-Depth Statistical Analysis of Olympic Swimming**

**MGSC 401**

**Professor Juan Serpa**

Group 12

Sam Greene

Jamie Park

Nicholas Toronga

Yuan Tian

**I. Introduction**

The Olympics is an international sporting event that garners the attention of billions across the globe. The athletes competing in the multitude of events in both the summer and winter editions of the games are top-shape citizens whose lives have been shaped around performing their sport of choice. However, the unfortunate fact is that not everybody starts on a level playing field. With the diversity of countries participating in the Olympics, we expect to see a divergence in economic and developmental stages amongst them. We wanted to see what extrinsic factors top athletes of the world were stacked up against. Maybe being tall, strong, or nimble are not the only factors at play in determining the success of an Olympian.

The base of our analysis started out with 120 years of Olympic data. The bulk of the dataset was comprised of various athletes’ statistics as well as specifics on the location and events from the 1896 Athens Olympics to the 2016 Rio de Janeiro Olympics, with each row corresponding to an athlete’s performance in one event. Thus, there are multiple rows for the same athlete if they competed in more than one event or more than one Olympics. Given the diversity of sports we have at hand, we believe it would be in the best interest of thorough and insightful analysis to focus on a single sport. We chose swimming, which has been a part of the Summer Olympics since 1896 for men and 1912 for women. It is comprised of a variety of events including different swimming styles such as backstroke, butterfly, and varying distances.

The central goal of this analysis is to observe the impact of variables beyond the athlete’s control that determine his or her success. Observing whether uncontrollable factors play into an athlete’s likelihood of performing well in the Olympics could lead to the discovery of ways in which an athlete’s background can be accounted for in his or her training. In a nutshell: if an athlete is at an inherent disadvantage, is there anything that can be done to correct for it?

**II. Data Description**

For our predictive model, we chose to analyze data from the 2000 Summer Sydney Olympics to the 2016 Rio de Janeiro Olympics. This decision was made based on the assumption that time-related factors such as changes in rules and regulations, technological developments in sports equipment, and advancements in athlete training methods have vastly changed since 1896.

A subset including Olympic events from 2000-onwards leaves us with 85,258 observations, representing 31.45% of our original 271,116 observations. Our further and final subset of just athletes that competed in swimming events leaves us with 6718 observations.

To begin the investigation of the significance of external factors as well as individual athlete characteristics, we combined a few datasets to include more variables related to the characteristics of each athlete’s home country. In the original athlete dataset, we already had variables for the athletes’ personal characteristics such as height, weight, age, as well as their medal result in the event (None, Bronze, Silver, or Gold). To further explore the relationship between athlete success and external factors we added variables such as the GDP and population of the athlete’s home country in the year in which they competed. We also added the country’s area, number of infant deaths per 1000 people, literacy rate, number of phones per 1000 people, and population density. These factors reflect the size, level of economic activity, and stage of development of a country. The GDP per capita column was added based on the existing variables, GDP and population.

Our dataset was then composed of two categories of variables: athlete characteristics and home country characteristics. We made the decision to focus our analysis on variables describing an athlete’s home country with the hopes of determining whether external factors heavily influenced their performance. This left us with eight variables to make our final model: GDP, GDP per capita, population, population density, area, phones per 1000 people, infant deaths per 1000 people, and literacy rate. To gain a primary level of insights into these external factors’ relationship with athlete success, we plotted histograms and boxplots of each variable. Also, since each athlete in the dataset is assigned their country’s characteristics, we created histograms that show both the distribution for all participating athletes, as well as the distribution by individual country. For example, all American swimmers from 2016 have the same GDP, thus the breakdown by athlete, and by individual country will look slightly different.

We first looked at GDP per capita. From the scatterplot (Appendix 1), we can see that in general, the chance of winning a medal increases as the GDP per capita increases. This shows that richer countries have had more success in recent Olympic in recent history. The mean GDP per capita also increased from $14,778 to $32,220 from 2000 to 2012, but then dropped to $28,739 dollars in 2016. We can see that GDP per capita is distributed asymmetrically and skewed right, indicating that most of the athletes come from a country with a GDP per capita below $50,000 (Appendix 2). Then, we looked at the area of the country, the infant mortality rate, and population density and found that all three variables are also skewed right (Appendix 3). Different from the previous few variables, the literacy rate is skewed left showing that most participating athletes come from a country with a literacy rate between 75% and 100% (Appendix 4). The number of phones per 1000 people is the only variable that is distributed normally. Phone usage achieves a peak at around 480 phone users per 1000 people (Appendix 5). We then checked for collinearity, and from the correlation matrix (Appendix 6), we can see that all the results are below 0.80, therefore there is no collinearity between two variables that needs to be accounted for.

The dataset also included the gender of each athlete. Overall, there were 3853 swimming events for women and 4361 for men. In our dataset, there are 1844 male swimmers and 1466 women. Originally, we thought that we should further subset the data based on gender because of the difference in physical characteristics. However, since we are not accounting for physical characteristics, the overall impact of external factors on success for both male and female athletes should hold the same.

Finally, we wanted to see which countries have had the most success in Olympic swimming events since 2016. Among all the represented countries that appear in the dataset, USA and Australia are the clear dominant nations when it comes to swimming (Appendix 8). We also found that only a few countries have enjoyed great success. In fact, from 2000 to 2016, only 20.32% of the countries that participated won a medal (38 out of 187). These findings lead us to our first main insight which is that wealthier countries are more likely to be successful because they could provide better training and equipment for athletes. We first started to see this relationship between success and economic development when we plotted GDP per capita against total medals earned by country. We will further explore this relationship in the following sections, to see if there truly is a great advantage for swimmers from wealthier countries, and if so what is the magnitude of this advantage.

**III. Model Selection**

Our team wanted to create a model from which we could draw insights that could lead to recommendations and applications for athletes and countries in future Olympics. Thus, we decided to focus our model solely on how the characteristics of a country that the athlete is representing influences whether they win a medal, and which medal they are most likely to win in a swimming event.

Our dataset contained variables that represented a country’s development such as the nation’s GDP per capita, literacy rate, and number of phones per 1000 people. We also had variables related to a country’s size such as area, population, and population density to work with. From this, we wanted to predict the probability that an athlete from a given country has of winning a gold, silver, bronze, or no medal, without knowing anything about the athlete, to show how unlikely it truly is for athletes from certain countries to win a medal in Olympic swimming events. Since we are trying to predict a categorical variable with more than two outcomes, we decided to use linear discriminant analysis or quadratic discriminant analysis as a final model (we get to the decision between the two later). We also decided to use discriminant analysis because we wanted to see the probabilities of winning each medal based on the predictors. However, we used tools such as decision trees and random forests to determine which predictors to include in the final model.

**III.A Visualizing Relationships**

In order to gather our first insights into how a country’s characteristics effected its athletes’ success, we visualized the relationships between the different types of medals, and the variables using boxplots. Immediately we could see that there were some significant relationships, as well as some insignificant relationships. In Appendix 9 you can see that the number of phones used per 1000 people in a country that an athlete is from, and the medal which they win is positively correlated as the median rate increases as the medal goes up in value from no medal to a gold medal. Other variables that seemed to have a significant and positive relationship with the medal won included area, population, and GDP. Another variable that was found to have a significant relationship with the medal earned was infant mortality rate. However, this turned out to be a negative relationship as we can see through the boxplot (Appendix 10). On the other hand, a country’s literacy rate was found to have almost no relationship with the success of its athletes in swimming events (Appendix 11). Another variable that was taken out of the running to be included in our model due to its lack of a clear relationship with medal earned was the country’s population density (Appendix 12).

**III.B Using Classification Trees and Random Forests**

After visualizing the data and getting a general idea of what variables should be included in the final model, we further explored the importance of each variable in predicting an athlete’s success using classification trees and random forests. We did this to gain more concrete and quantitative reasons to include certain variables in our model. We first ran a decision tree with a CP of 0.00001, with the variables that had not been ruled out based on the boxplot visualizations, to see which variables were the most important (Appendix 13). We can immediately see that GDP, area, and population are the most important country-related factors in determining which medal an athlete is most likely to win according to the tree.

In order to quantitatively analyze the importance of each variable, we ran a random forest model with the same variables. This then allowed us to use the ‘importance’ function to see the mean decrease in accuracy with the exclusion of each variable. It also allows us to see the mean decrease in the Gini coefficient with the exclusion of each variable. The Gini coefficient measures node purity in a decision tree, so the higher mean decrease in the Gini coefficient, the more important the variable is for making accurate predictions. We can see from the table (Appendix 14) that the three most important country variables for predicting medals earned at the Olympics in swimming events are the country’s number of phones used per 1000 people, the country’s GDP, and the country’s population. This test also allowed us to confirm that the rest of the variables that we included in this test are still relatively important and should be included in the model.

**III. C Final Model**

Our previous analysis showed us that the best variables for predicting the likelihood that an athlete will win a given medal based on their country’s characteristics are the country’s population, GDP, and number of phones used per 1000 people. These variables each represent an important aspect of a country; population is a good measure of a country’s size, GDP is a prominent measure of economic activity, and phone usage shows us how developed a country is. The other variables that we decided to include in our final model were area, and number of infant deaths per 1000 people, as these factors both proved to have a relationship with Olympic success in swimming events.

Once we had our variables, we were faced with the decision between linear discriminant analysis, and quadratic discriminant analysis. Since, our previous analysis of the data showed that each variable had a different distribution and standard deviation, we decided to use quadratic discriminant analysis to predict the likelihood that an athlete from a certain country has of winning either no medal, a bronze medal, a silver medal, or a gold medal.

**IV. Results**

The main goal of our analysis was to find out if athletes from certain types of countries are at a significant advantage or disadvantage when it comes to Olympic swimming success. With our data from the 2000 Olympic onwards we built a quadratic discriminant analysis with the following predictors; the number of phones per 1000 people in the athlete’s home country as well as the country’s GDP, population, area, and infant mortality rates. This model accounts for the athlete’s home country’s size, economic activity, and development. Our model produced the following results (Appendix 15), and an error rate of 12.5%. This rate shows that 87.5% of the time our model can successfully predict which medal an athlete competing in a swimming event at the Olympics will win only based off what country they are from.

By looking at the group means of each category for each predictor we can see what type of countries are more likely to be successful (Appendix 15), and it coincides with our original hypothesis. The group mean of phones per 1000 people increases drastically (from 393 to 714) from the no medal category to the gold medal category. Meanwhile, the group mean of infant deaths per 1000 people decreases significantly from the no medal category to the gold medal category. However, the variable with the mean that changes the most from the no medal category to the gold medal category is GDP. We can see that the group mean of GDP in the gold category is almost 7 times greater than the group mean in the no medal category. The group means of Area and Population also increase significantly as the medal categories increase in value.

The analysis of the group means produced from our QDA model shows us that athletes from countries that are larger in size are more likely to win a gold medal in swimming. This makes sense as Olympic pools, and training centers take up a lot of space, so in order to build a lot of these types of training facilities for their athletes, countries need to be relatively large in size. Secondly, we found that athletes from countries with greater economic activity are more likely to have success in Olympic swimming events. The main reason for this is that along with being large, swimming training facilities, and nice indoor pools in general are expensive, so countries with a lower GDP have more trouble affording them, and thus their athletes have trouble competing with athletes from countries that have more resources to allocate towards athletic training. Lastly, we can conclude that athletes from less developed countries are at a disadvantage when it comes to competing for Olympic medals in swimming compared to athletes from more developed countries based on the group means analysis of infant mortality per 1000 people and phones per 1000 people.

After analyzing the group means, we wanted to see how accurate our model was at making predictions, so we found our error rate of 12.5%. We did this by making a confusion matrix with the predicted medal based off our model and the actual medal that the athlete won (Appendix 16). This matrix shows that our model is extremely good at classifying athletes on both sides of the success spectrum (no medal, and gold medal). However, it struggled when it came to predicting when athletes would win silver and bronze medals, as it did not correctly classify any of the athletes that won silver or bronze medals.

Our model is certainly not perfect. Accurately predicting an athlete’s success without looking at individual athlete characteristics is nearly impossible. However, our model and analysis did show us that athletes from smaller countries, that have less economic activity, and are less developed are very unlikely to win a gold medal in a swimming event at the upcoming Olympics.

**V. Conclusion/Predictions**

After our model showed us that athletes from large countries that are developed and have a large amount of economic output are much more likely to be successful in Olympic swimming events, we wanted to see the magnitude of this advantage. We first decided to predict the likelihood of an athlete from a hypothetical country that has the median value for each of the characteristics that we included in our model. This hypothetical country, let us call it ‘Medianland’ has 176 phones per 1000 people, an area of 86,600 square miles, 21 infant deaths per 1000 people, a GDP of 36,037,519,521, and a population of 10,038,188 people. Our model predicted that an athlete from Medianland has a 98% chance of winning no model, 0.6% chance of winning a bronze medal, a 0.9% chance of winning a silver medal, and a 0.5% chance of winning a gold medal.

Next, to see how much of an advantage that athletes from large, developed countries with high economic activity have at these events, we used our model to predict the likelihood of winning each medal for Canadian swimmer. Based on the number of phones per 1000 people, GDP, population, area, and infant mortality per 1000 people, any Canadian Olympic swimmer has a 56% chance of winning no medal at all, a 13% chance of winning a bronze medal, a 23% chance of winning a silver medal, and an 8% chance of winning a gold medal.

As we can see a Canadian Olympic swimmer is much more likely to be successful than a swimmer from Medianland, a country of mediocre size, development, and economic activity.

These predictions and results show us that swimming is a sport that gives a great advantage to athletes from larger, more developed countries. We were not surprised by these results as you often see swimmers from the United States, and Australia on the podium at every Olympics, and rarely, if ever do we see swimmers from countries such as Algeria achieve great success. We had originally chosen to narrow our analysis down to swimming because of the vast differences in training and technology required across different sports. So, our results led us to wonder if this advantage for larger, developed countries held true in other sports.

To explore this, we ran the exact same quadratic discriminant analysis model on a subset of data that included only athletes that participated in track and field events from the 2000 Olympics onwards. We chose track and field because it is also one of the main Olympic sports that gathers a lot of attention. A model with the exact same predictors as the one we ran before gave us the following results (Appendix 17). We can see that the group means of each variable are far less spread out across the different medal categories than they were for the swimming model. This immediately shows us that country characteristics seem to be less important for determining success in track and field events as they are in swimming events.

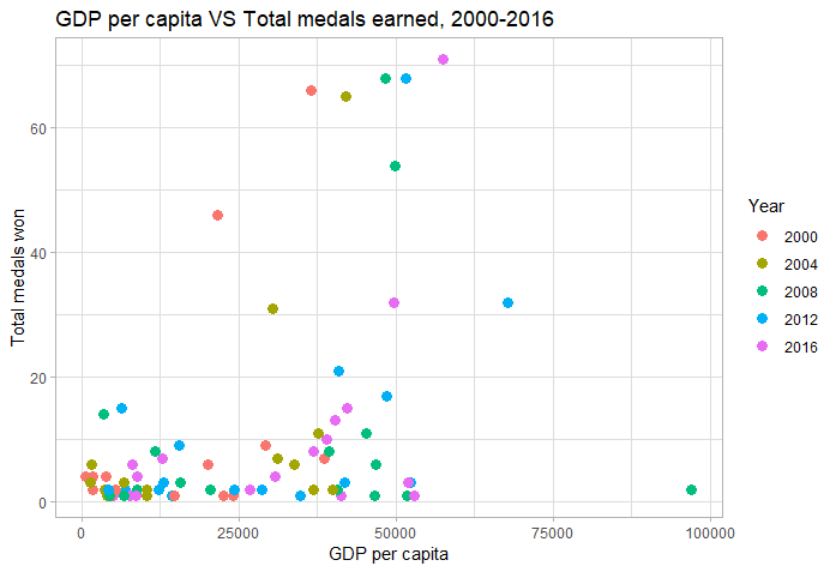
To further test this hypothesis, we predicted the likelihood of winning each medal for an athlete from the same two countries we had made predictions with using the swimming model; a Medianland, and Canada. Based on our model, an athlete from the Medianland has a 5% chance of winning a gold medal in track and field, compared to the 0.5% chance he/she had in swimming. Meanwhile, the Canadian athlete’s likelihood of success decreased significantly. The results from these predictions showed that in fact, athletes from smaller, less developed countries have a significantly higher chance of winning a gold medal in a track and field event, than in swimming.

In conclusion, we can see that the characteristics of a country in which an athlete is from has a large effect on the success of the athlete in Olympic swimming events. It also showed that bigger and richer countries are the most successful. This is most likely due to swimming being a sport in which success relies heavily on what type of facilities you train in. Larger countries with more resources are much more likely to have a higher number of Olympic sized pools for their athletes to train in. However, it turns out to be quite the opposite for track and field events. Countries that aren’t as wealthy or large have much more success in these events. This is likely due to the fact that training for these events does not require nearly the amount of resources that training for swimming does. You can run anywhere.

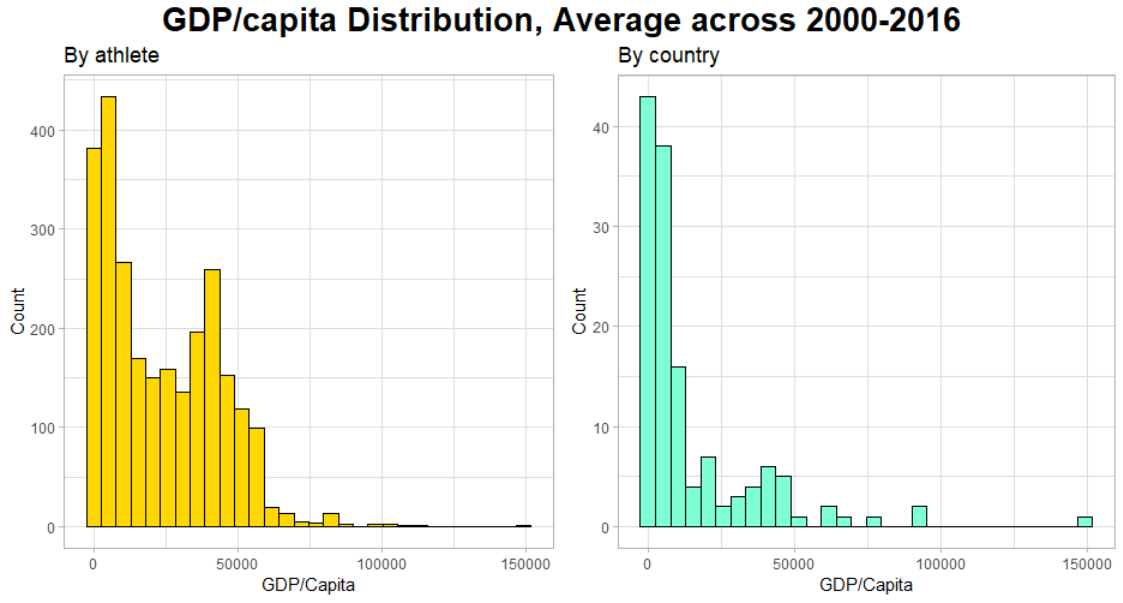
Our in-depth analysis showed that swimming is not a sport that athletes from developing countries should focus on. Thus, we recommend that these smaller, developing countries invest more in other sports that require less resources, as greater success in important sports such as track and field on the world stage at the Olympics will aid these countries in their development.

**VI. Appendices**

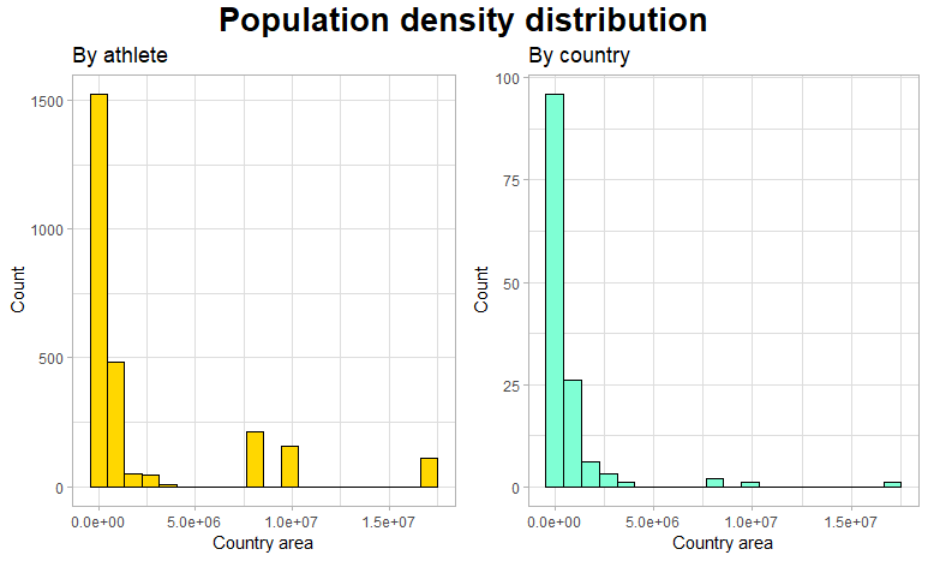
Appendix 1: Scatterplot of GDP per Capita vs Total Medals Earned

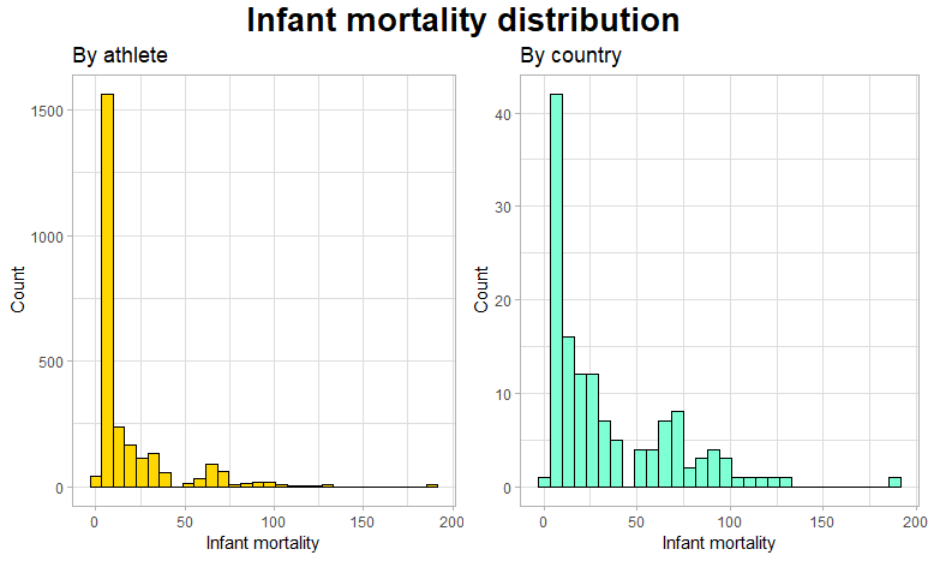


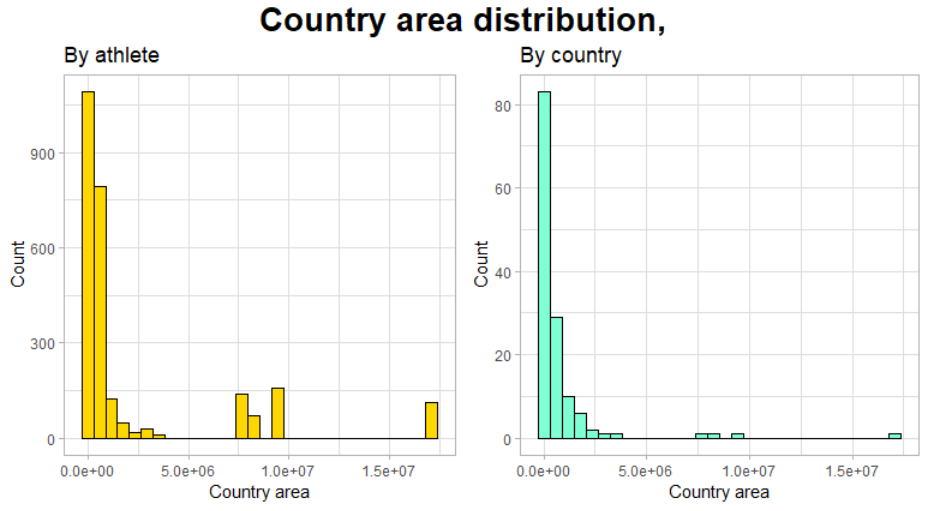
Appendix 2: Distribution of GDP per Capita by Athlete and by Country



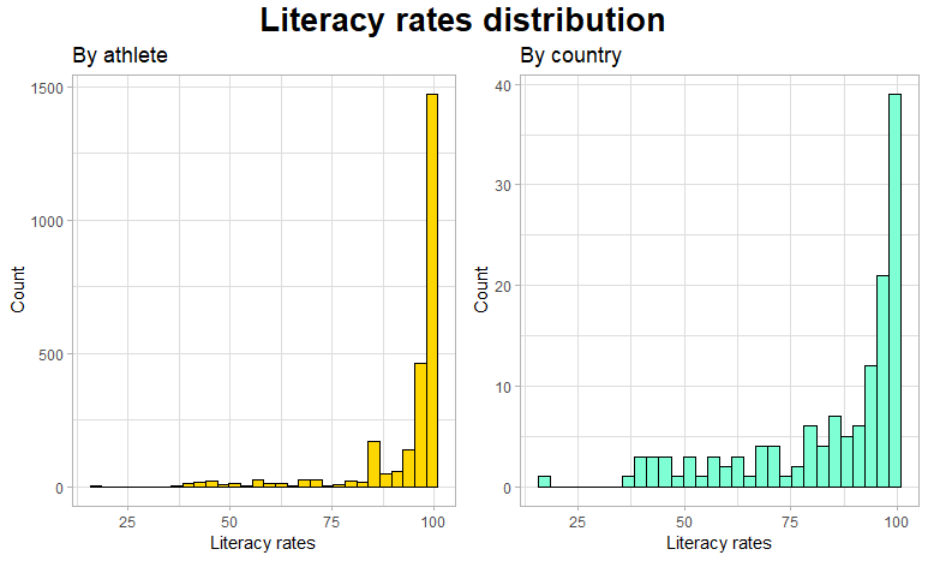
Appendix 3: Distribution of Population Density, Infant Mortality, and Area by Athlete and by Country



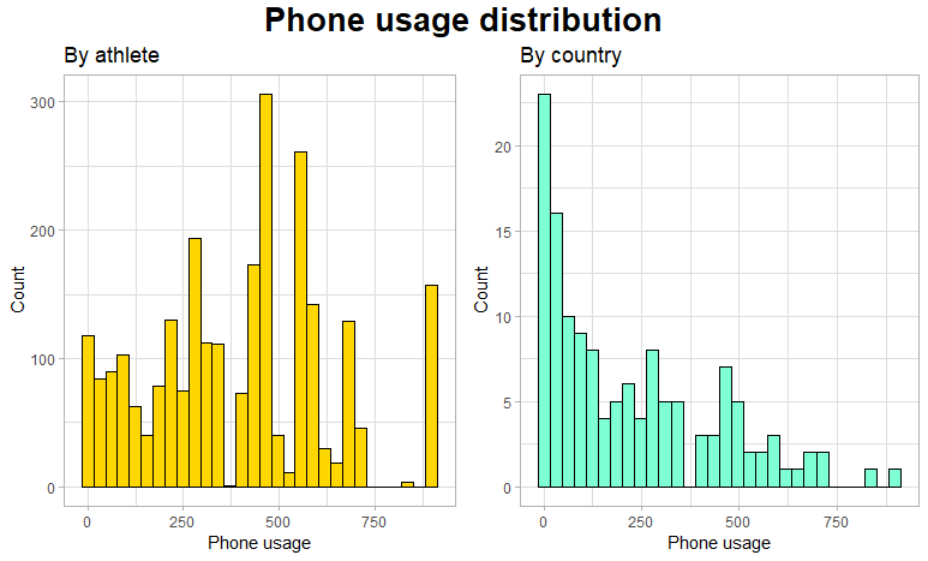




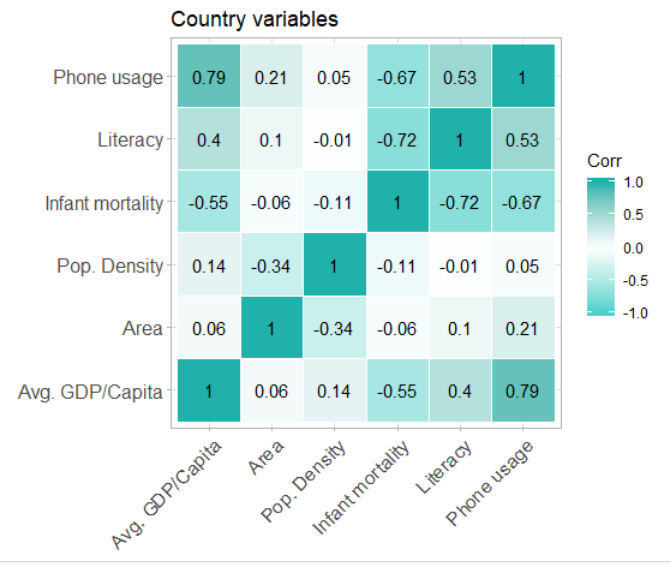
Appendix 4: Distribution of Literacy Rate by Athlete and Country



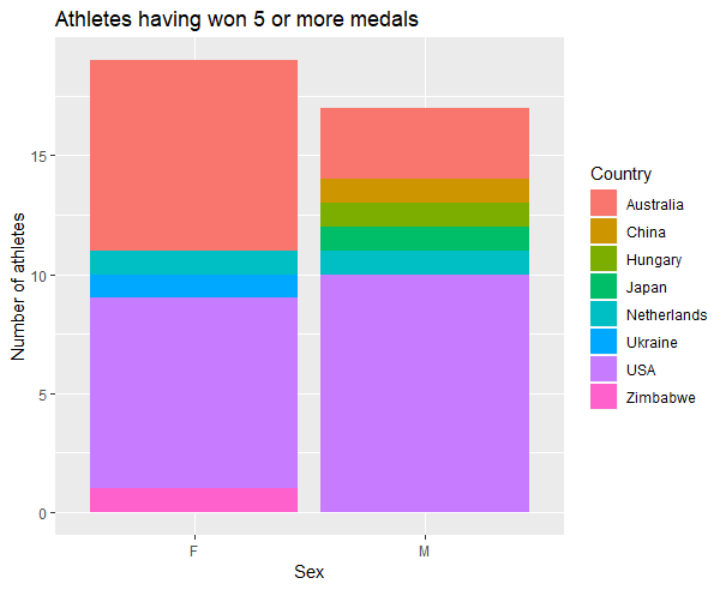
Appendix 5: Distribution of Phones per 1000 People by Athlete and Country

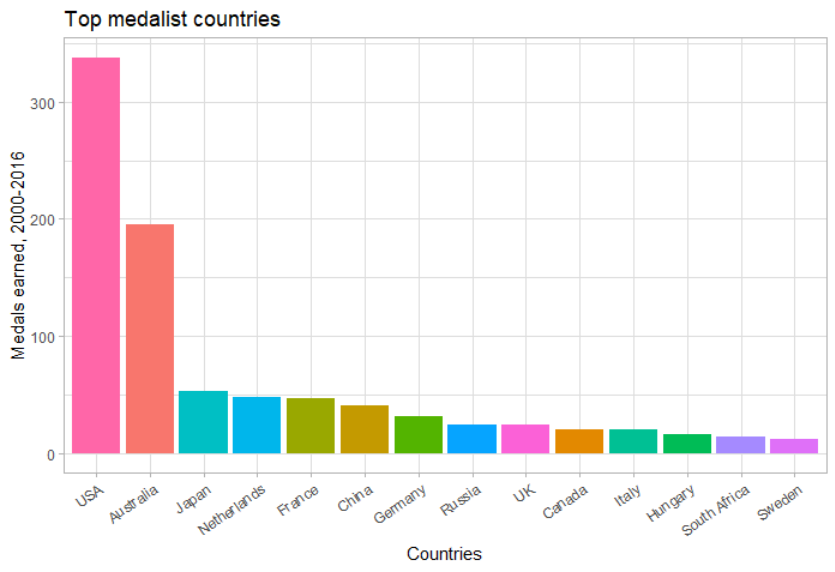


Appendix 6: Correlation Matrix of all Country Variables

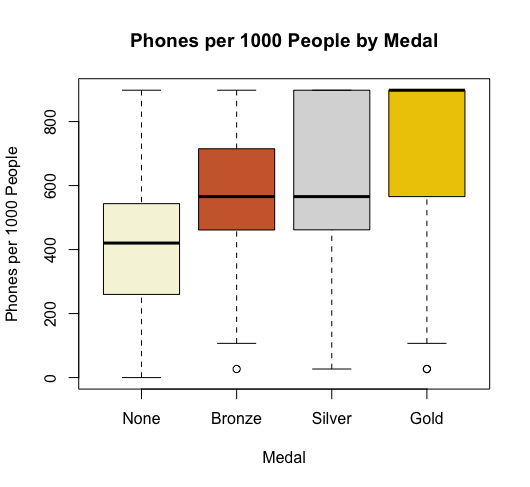


Appendix 8: Graphs Showing Country Success in Swimming Events

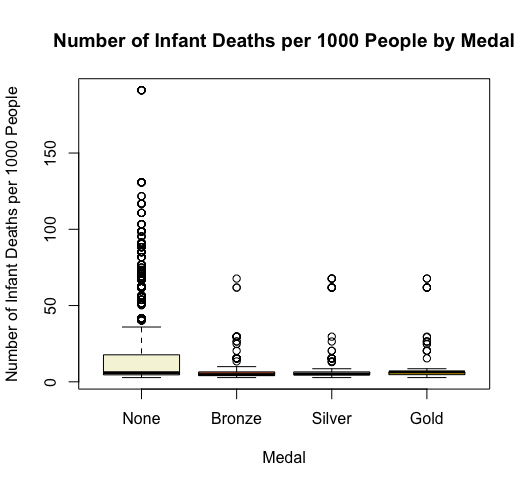




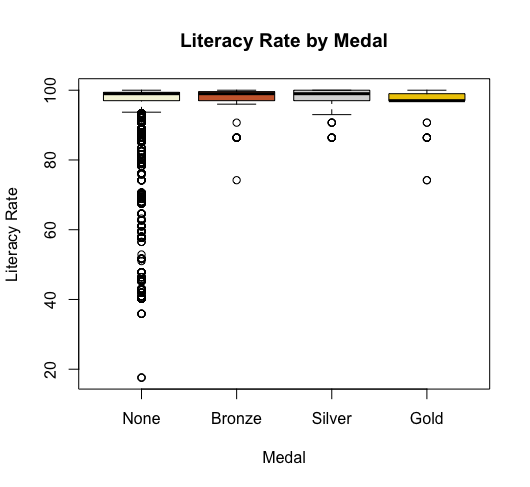
Appendix 9: Boxplot of Phones per 1000 People by Medal



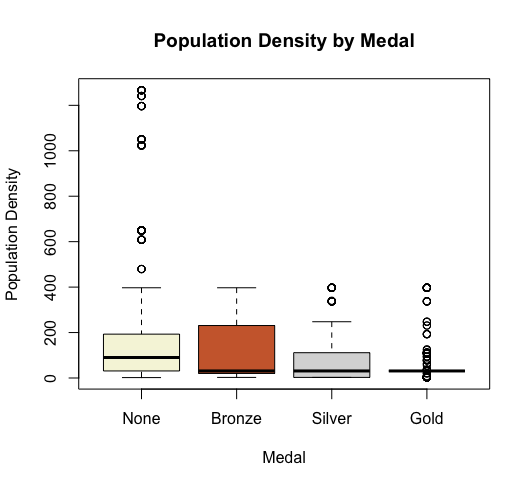
Appendix 10: Boxplot of Infant Deaths per 1000 People by Medal



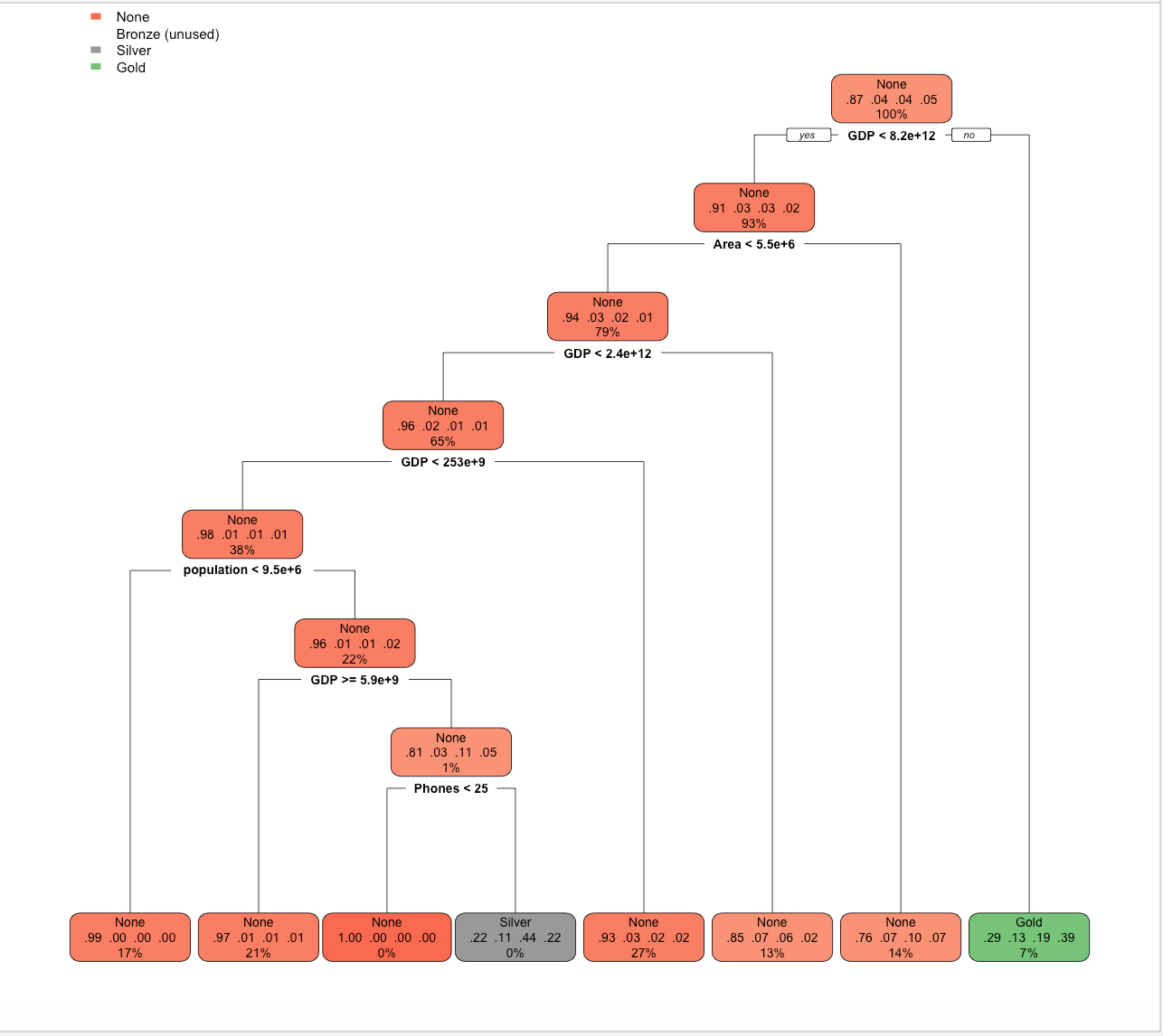
Appendix 11: Boxplot of Literacy Rate by Medal



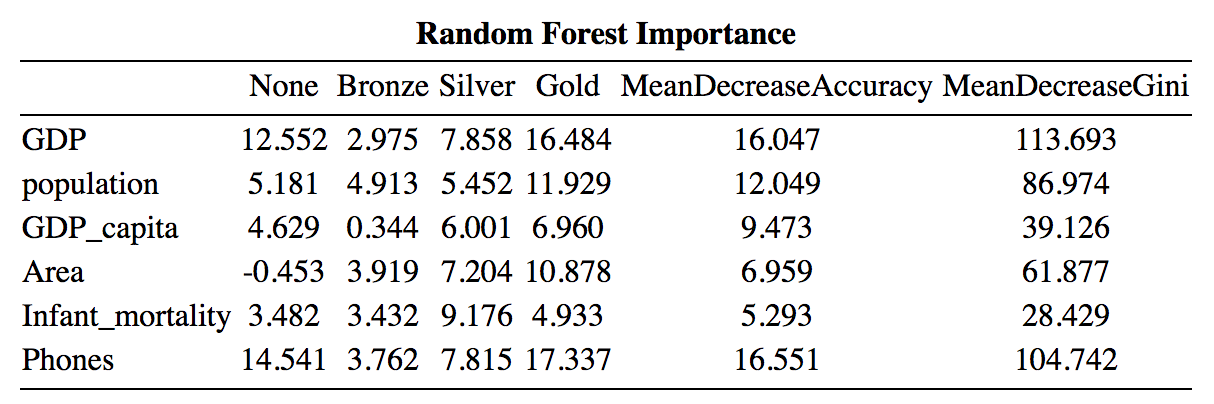
Appendix 12: Boxplot of Population Density by Medal



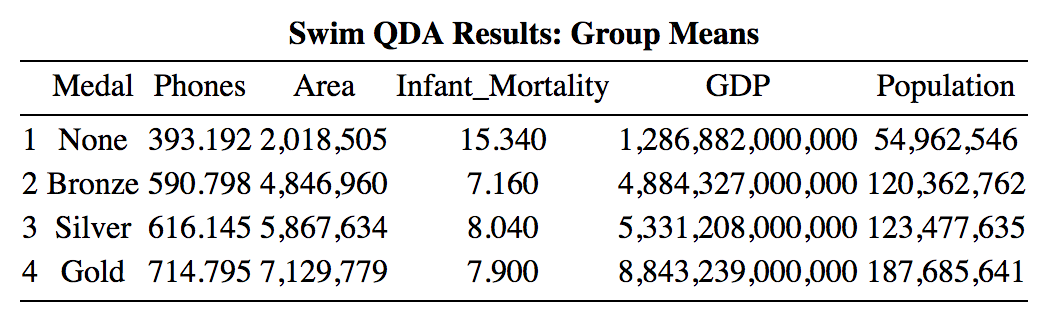
Appendix 13: Decision Tree with CP 0.0001

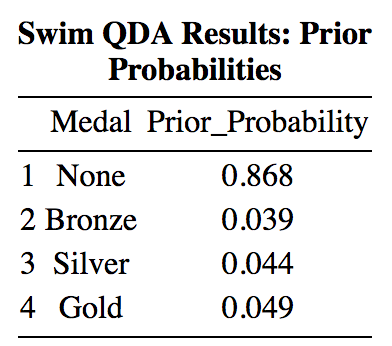


Appendix 14: Importance of Each Variable from the Random Forest Model



Appendix 15: Results from the Quadratic Discriminant Model

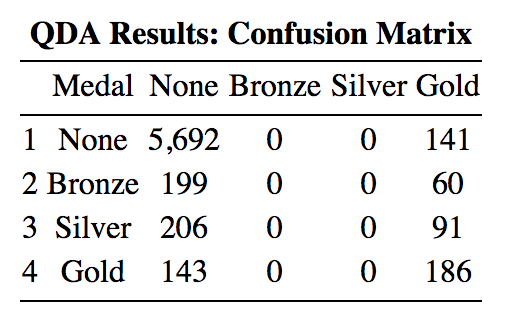




Appendix 16: Confusion Matrix of Actual Medal versus Predicted Medal based on the Model

Rows: Actual Medal

Columns: Predicted Medal



Appendix 17: Group Means of the Model ran on Track and Field Athletes

