Analyzing the Rio 2016 Olympics

Jack Doyle
Jack.doyle@marquette.edu

Josie Zucca
Josephine.zucca@marquette.edu

Tyler Hackel
Tyler.hackel@marquette.edu

Amy Geraghty
Amy.geraghty@marquette.edu

Abstract
This paper analyzes the data from the athletes of the Rio 2016 Olympics. By cleaning, grouping, and visualizing the data, patterns and trends can be made about Olympic athlete bodies and medals won. The primary questions being investigated focus on determining the average athlete's height, weight, and age for each sport and the standard deviation. The optimal athlete is determined by medal winners in each category. By analyzing these factors, a logistic regression model has been created that can predict, to a certain degree of accuracy, whether or not an athlete will win a medal in his or her individual sport. Additionally, we investigated how significant the differences between the bodies of men and women competing in the same sport were.

Author Keywords
Olympics; Rio 2016; height and weight analysis; map of athletes; medalists; logistic regression model; confusion matrix

Introduction
The Olympic Games occur every two years, alternating between Summer and Winter. Athletes competing in the games spend innumerable hours training and preparing for their events. Determining the ideal body type for an event is important to prospective athletes when determining how best to train for their sport, or which sport to compete in. Countries also strive to compete in number of medals won by their athletes and knowing how best to maximize their chances is essential.

Data Set

The athletes.csv [1] file contains the information of the 11,000+ athletes who competed in the 2016 Olympics held in Rio de Janeiro, Brazil. This data set was uploaded by the International Olympic Committee (IOC) to the official Rio Olympic Games website, rio2016.com. Matt Riggott, a programmer based in Iceland, web scraped the Rio site, structuring and uploading the data to GitHub. Notable fields of the data set include ID, name, sex, nationality, date of birth, height, weight, and gold, silver, and bronze medals won. There are 28 different events for athletes to compete in. Descriptive statistics from the data set are displayed in Figure 1.

	id	height	weight	gold	silver	bronze	medals won	age
count	1.153800e+04	11208.000000	10879.000000	11538.000000	11538.000000	11538.000000	11538.000000	11538.000000
mean	4.999885e+08	5.794889	158.883006	0.057722	0.056769	0.061016	0.175507	26.697435
std	2.908648e+08	0.369812	35.664874	0.255910	0.239147	0.243320	0.428628	5.378624
min	1.834700e+04	3.969816	68.343220	0.000000	0.000000	0.000000	0.000000	14.000000
25%	2.450997e+08	5.544620	132.277200	0.000000	0.000000	0.000000	0.000000	23.000000
50%	5.002011e+08	5.774278	154.323400	0.000000	0.000000	0.000000	0.000000	26.000000
75%	7.539874e+08	6.036746	178.574220	0.000000	0.000000	0.000000	0.000000	30.000000
max	9.999878e+08	7.250656	374.785400	5.000000	2.000000	2.000000	6.000000	62.000000

Figure 1: Descriptive statistics of variables from the Rio 2016 athlete data set.

Cleaning the Data

In order to use the data for calculations and observations, several adjustments had to be made to the original data set. The athletes' height, originally listed in meters, was converted to inches. Similarly, height was converted from kilograms to pounds. This was to increase readability for our American target audience. The date of birth column was used to calculate a new ag e field, to compare athletes more easily. Additionally, a column for total medals won, a summation of the gold, silver, and bronze columns, was created. Athlete weights for boxing were missing a majority of the time, so the mean weight was found on the Rio website and inserted for the purpose of analysis.

Height vs Weight

An Olympic athlete's body is extremely important. Height and weight are primary factors when predicting an individual's potential success in a particular sport. Research has previously been done on these factors. For instance, Andrew Moony compiled the height and weight of medal-winning gymnasts through 2010 [5]. The average medal-winner was 5 feet 1 inch tall and weighed 103 pounds. In contrast, as found by Robert Wood, the average Rio basketball player measured 6 feet 5 inches and weighed 192 pounds [10].

Figure 2 displays height versus weight distribution graphs for 3 of the 28 Rio Olympic events. Many observations can be made from these graphs. As mentioned before, it can be observed that the Olympic gymnasts have a noticeably lower average height (5 feet 4 inches) and weight (119 pounds) when compared to all the other events. For instance, the volleyball graph demonstrates the average height of athletes is 6 feet 2 inches, and the average weight is 177 pounds. Events not pictured, such as weightlifting, wrestling, and judo, have noticeable vertical clusters because these events are split into distinguishable weight categories. Aquatics is one of the events with the largest number of participants (1445 total). This event is split into numerous smaller events. This explains the dense cluster of points on the graph, as well as the large range of heights and weights observed.

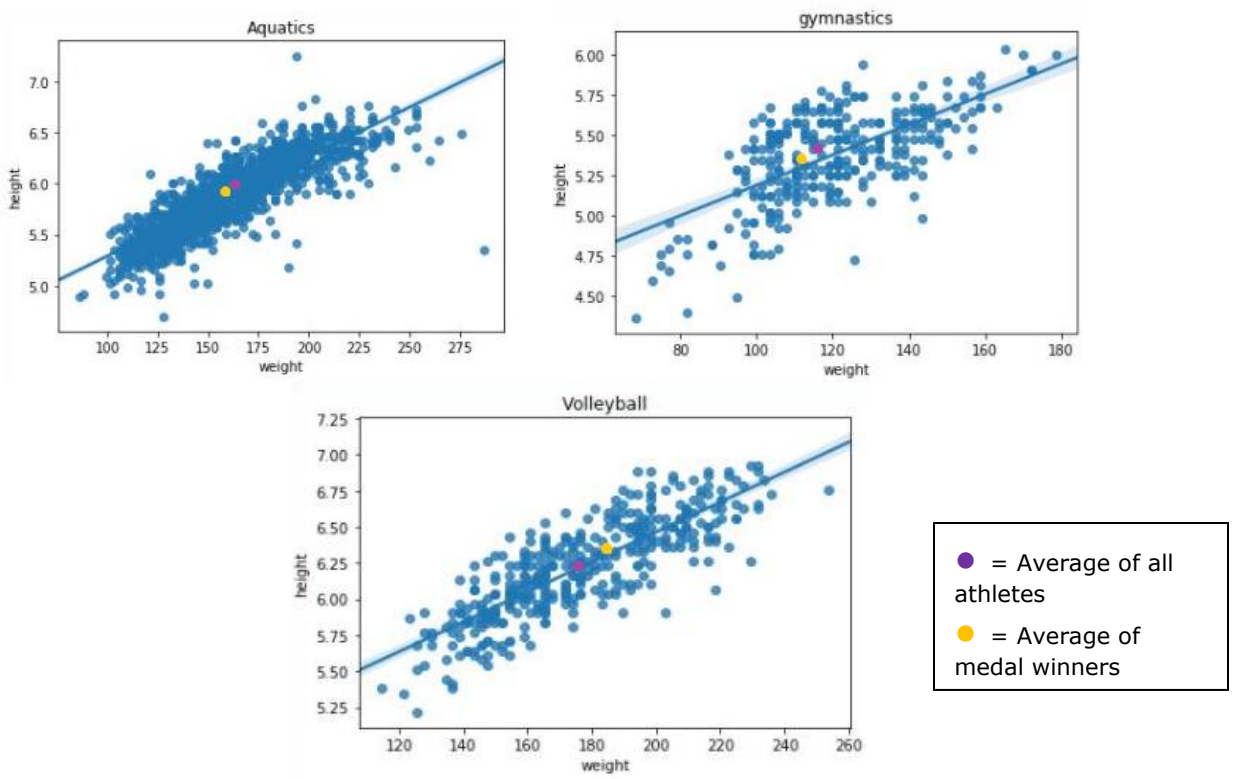

Figure 2: Selection of height vs weight graphs for individual sports, marking the average and optimal height and weights.

The purple dot represents the average height and weight for all athletes competing in the event, while the golden dot represents the statistics for the medal winners. It is interesting to note that in each graph, both averages lie nearly right on top of the line of regression, implying that there is an ideal height to weight ratio for each sport. When considering whether the optimal athlete has a height and weight above or below the average, the answer depends on the sport. For gymnastics, the optimal lies slightly below the average, implying that a smaller build is beneficial. In contrast, the volleyball optimal is slightly above the average, implying that a taller build is preferred. These observations make sense when considering how height and weight contribute to

these individual sports. Similar research by Alvin Chang resulted in the comparison of an individual's height with the past 10 gold medalists in each of the events [2]. This analysis describes the perfect height for each event, the reasons a height has certain advantages, and indicates whether the individual could realistically compete. Similarly, Nassos Stylianou takes height, weight, date of birth, and sex into account to search the dataset and identifies the 3 Rio athletes with the closest body type [9]. The site also displays the average height, weight, and age of all Rio 2016 athletes. Finally, Simon Rogers provides numerous filters from the 2010 Olympics held in London, England to analyze and visualize the same data fields of the athletes [7]

Aquatics Logistic Regression Modeling

The primary goal of this analysis was to create a model to predict if a particular athlete would win a medal. When attempting to determine medal winners, height weight and age are important factors to consider. Models were created for each individual sport. This provides more helpful information than creating a model to include all athletes, because the optimal body type is very different for each sport, as indicated by the graphs in Figure 2.

Aquatics was selected as the primary sport to do in depth research on. Figure 3 in the side bar contains the descriptive statistics for all athletes in the aquatics category. Early research done in 1988 found that medal winners in swimming averaged 18 years old for women and 21 years old for men. However, recent results from 2014 shows that the age of medal winners has changed over time [4]. Figure 3 shows that the mean age of swimmers competing in the Rio Olympics is 24, which supports the idea that the ideal age to win a swimming medal has increased over time.

	height	weight	age
count	1396.000000	1396.000000	1396.000000
mean	5.867251	159.343806	23.906160
std	0.361777	31.601357	4.341903
min	4.691601	85.980180	14.000000
25%	5.577428	134.481820	21.000000
50%	5.839895	154.323400	23.000000
75%	6.135171	180.778840	27.000000
max	7.250656	286.600600	41.000000

Figure 3: Descriptive statistics for height, weight, and age of all swimmers competing in aquatics.

The logistic regression model determined for swimmers is pictured in Figure 4.

```
Medal_Winner = 1.2072(height) + .0016(weight) +
.0444(age) - .9443(sex_male)
```

Figure 4: Logistic regression model for athletes in aquatics.

We predicted that the body composition, particularly height and weight, would be crucial to determining medal winners for swimming. Research done by Reel and Gill asserted that a lighter weigh results in faster times in races, possibly because a slimmer frame may be more hydrodynamic [6]. However, research by Sinning contradicts these findings, stating that a slighter larger frame is also beneficial because it improves buoyancy. Therefore, both leanness and fatness have benefits for swimmers [8]. This finding is supported by our model, because the p-value and coefficient for the weight variable was very small, implying that it is not statistically significant. Thus, it can be inferred that a swimmer's height is important, but his or her weight is less so.

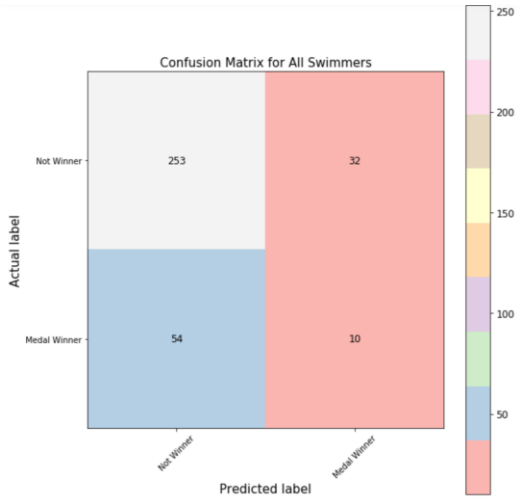

Figure 5: Confusion matrix for the logistic regression model to determine the aquatics medal winners.

The confusion matrix for the aquatics logistic regression model is show in Figure 5. The model correctly predicted whether or not an athlete won a medal 81.37% of the time. Achieving a much higher score is unlikely, because although height, weight, and age impact an individuals aptitude for a sport, overall ability and human behavior are the ultimate deciders of an individual's success. That being said, the model provides good insight into an individual's potential in aquatics. By using similar models for the other 13 sports, an athlete can determine which sport his or her body type has the best chance for winning an Olympic medal.

Sex Logistic Regression Model

In addition to predicting an individual's chance to win a medal, logistic regression was also used to examine the degree of significance between male and female body types. Figure 6 displays the confusion matrix for classification of all athletes, not separated by sport.

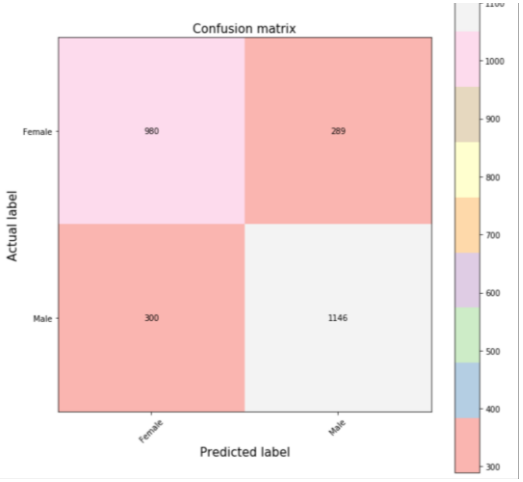

Figure 6: Confusion matrix for logistic regression of males and females from all sports.

We anticipated that this model would not be very accurate, because there is such a wide range in the body types of all athletes. For instance, judged solely on height and weight, male gymnasts and female basketball players could easily be misclassified. However, this model was correct 78.3% of the time, much higher than we had predicted.

Figure 7 represents the confusion matrix for the classification of males and females within the gymnastics events. Because gymnasts have similar body types, this model was predictably more accurate, with a score of 87.5%. These models demonstrate that even within the same sport, male and female bodies are considerably

different; enough so that a model and distinguish between the two the majority of the time.

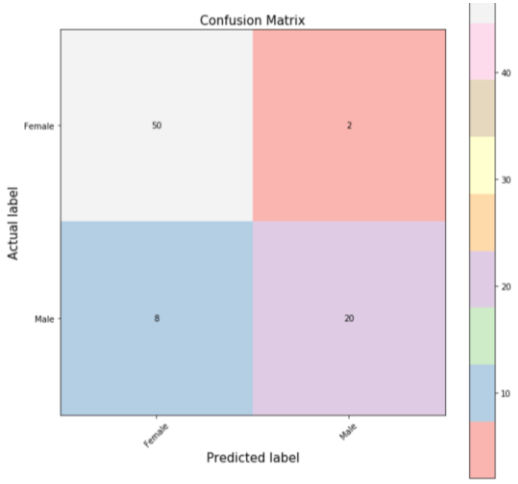

Figure 7: Confusion matrix for logistic regression of males and females from gymnastics events.

Age of Athletes

In addition to height and weight is an important factor when considering an athlete's potential in a sport.

The minimum age to compete in the Olympics varies by sport. The youngest competitor, swimmer Gaurika Singh of Nepal, was 14 years old and the oldest competitor, equestrian Julie Brougham of New Zealand, was 62. Age can be an advantage in sports that require more mental strength, but can be a disadvantage in more physical sports when competing against younger Olympians [3].

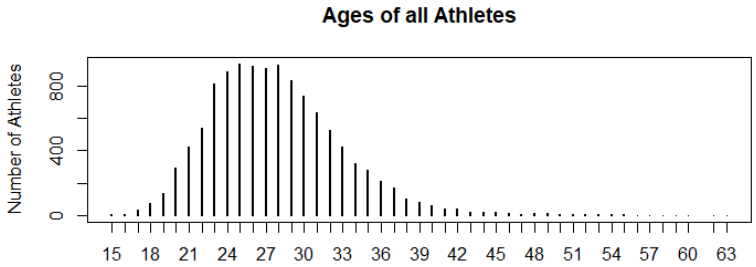

Figure 8: Shows the distribution of all athletes' ages. The mean age is 27.

Final Inferences

Data from the Rio 2016 Olympics concerning height, weight, age, and sex can be used to create a model to predict the likelihood of a particular athlete winning a medal in his or her event. While the perfect average height and weight for a particular sport may not always lead to a medal, there is a strong positive correlation between these variables and a resulting victory. Additionally, the observed difference between male and female body types of athletes in the same sport is significant enough to create a model that is accurate the majority of the time.

References

1. 2017. Data for the 2016 Olympic Games in Rio de Janeiro. (April 12, 2017). Retrieved Februrary 26, 2018 from https://github.com/flother/rio2016.
2. Alvin Chang. 2016. Want to win Olympic gold? (August 9, 2016). Retrieved March 25, 2018 from

https://www.vox.com/2016/8/9/12387684/olympic-heights

3. Janissa Delzo. 2016. The oldest and youngest Olympic Athletes. (August 11, 2016). Retrieved March 27, 2018 from https://www.cnn.com/2016/08/11/health/youngest-and-oldest-olympic-athletes/index.html
4. Stefan König, Fabio Valeri, Stefanie Wild, Thomas Rosemann, Christoph Alexander Rüst, and Beat Knechtle. 2014. Change of the age and performance of swimmers. (2014). Retrieved May 6, 2018 from https://springerplus.springeropen.com/track/pdf/10.1186/2193-1801-3-652
5. Andrew Moony. 2012. The bodies of champion gymnasts. (August 3, 2012). Retrieved March 25, 2018 from http://archive.boston.com/sports/blogs/statsdriven/2012/08/the_bodies_of_champion_gymnast.html
6. Justine J. Reel and Diane L. Gill. 2001. Slim Enough to Swim? (April 2001). Retrieved May 6, 2018 from http://thesportjournal.org/article/slim-enough-to-swim-weight-pressures-for-competitive-swimmers-and-coaching-implications/
7. Simon Rogers. 2012. Olympic athletes by age, weight and height visualized. (August 7, 2012). Retrieved March 25, 2018 from https://www.theguardian.com/sport/datablog/interactive/2012/aug/07/olympic-athletes-age-weight-height
8. Wayne E. Sinning. Body Composition and Athletic Performance. Retrieved May 6, 2018 from http://www.nationalacademyofkinesiology.org/AcuCustom/Sitename/DAM/129/TAP_18_LimitsofHumanPerformance_06.pdf
9. Nassos Stylianou, John Walton, Nathan Mercer. 2018. Who is your Olympic body match? Retrieved March 25, 2018 from http://www.bbc.com/sport/olympics/36984887
10. Robert Wood. 2008. Body Size & Basketball. Retrieved March 25, 2018 from https://www.topendsports.com/sport/basketball/body-size.htm