STA 9750 Final Project

Kimberly Yee Tan, Juan Rodriguez, Eric Sedaghat, Kimberly Choi

12/18/2020

Olympics: Event Winning Predictors and Performance by Gender over time

Introduction

It has been 124 years since the first modern Olympics took place in 1896. For our project, our group will be analyzing 50 years of Olympic history to:

- Determine if there are certain factors that can predict the chances of an athlete winning an event
- Compare male versus female performance by countries over the years

The main data set we will be using for our analyses will be "120 years of Olympic history: athletes and results" from Kaggle.com (dataset). This data set contains the biographical data of athletes and the medal results from the 1896 to 2016 Olympics games.

We have obtained other data sets to add a "GDP" column to our main data set to aid further analysis later. As there are many countries participating in the Olympics, we have decided to group countries into developed and developing countries based on the countries' GDP for our analyses. We have obtained another data set, "API-GDP1", under the subheading "Per Capita GDP in US Dollars" from the United Nations Statistics Division. "noc_regions" was a data set provided with our main data set and "host_cities" was obtained from The World Bank.

Summary of Data Set

Our main data set, "athlete_events", originally had 271,116 rows of data. However, perhaps the data dates too far back in time and there wasn't as much of an importance in recording the biographical information of athletes in the old days, there are a lot of NAs in certain variables of interest for our analyses (Age, Weight, and Height).

To obtain meaningful data by reducing the number of NAs in the columns with data of said variables of interest, we have filtered the data to observations from the past 50 years (1966-2016). We have also filtered out remaining observations with NAs in the Age, Weight, Height, and Sex columns.

We are now left with 175,350 rows for our analyses. The current data set has 18 variables: ID, Name, Sex, Age, Height, Weight, Team, NOC, Team_Country, GDP, Games, Year, Season, Host_City, Host_Country, Sport, Event, Medal. Main subgroups of potential relevance present in the data are: Sex (Male and Female), Team Countries, Year, Winter and Summer season, Sport, Host Countries, Medal, and development state of countries by GDP.

Hypothesis

We are looking to explore two questions with our data set:

- What is the importance of an athletes' age, height, weight, BMI, and home country's GDP in their Olympic performance?
- How does the gender affect the athlete's performance over time with regards to their country's GDP and stage of development?

Our hypothesis is as follows:

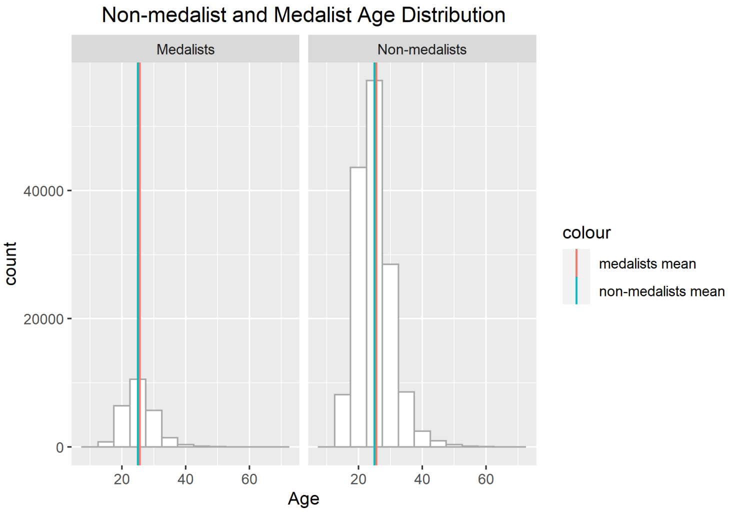
Athletes of age 18 to 25 are more likely to win. Athletes from developed countries will also perform better than those from developing countries due to having more training equipment, space, and overall resources available for them to prepare for the Olympics. Females overall will have a higher rate of performance improvement over time due to improved gender equality in recent years.

Before we start analyzing, below are some limitations we have identified about our data:

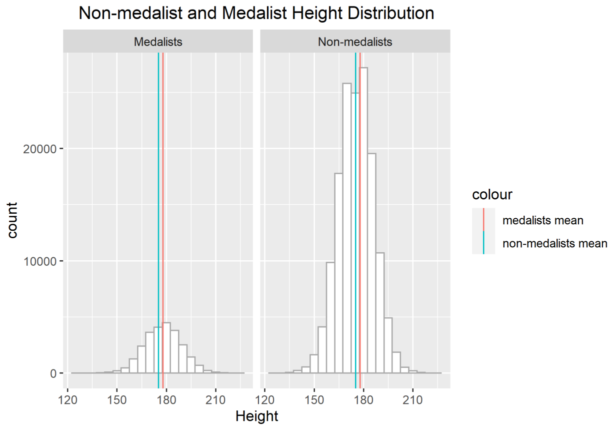
- Limited amount of variables we can use for data analysis (many more factors go into determining performance of athletes)
- Loss of data due to missing data in one or more variables of interest (have to be filtered out)

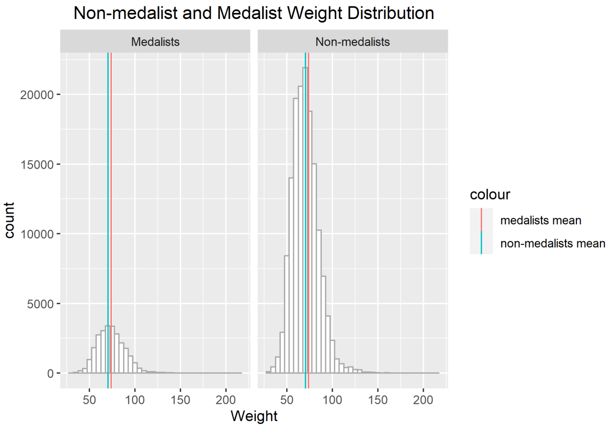
Exploratory Analysis

Potential Differentiating Factors for Performance



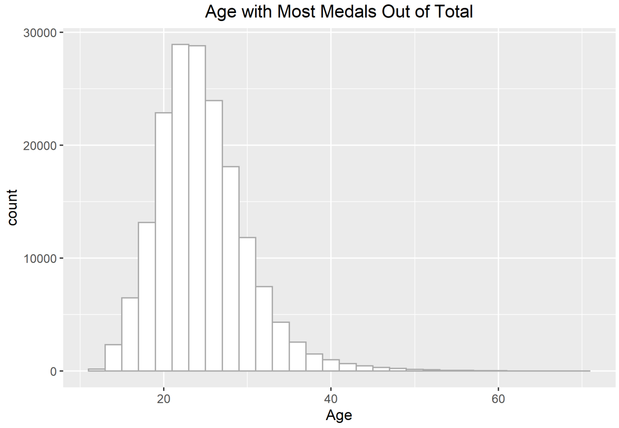
To see whether age, height, and weight affects performance, the distribution of age, height, and weight of non-medalists and medalists are compared side-by-side to see if there are any striking differences. As we can see from the figure above, the age distribution between non-medalists and medalists are similar. When we compare the mean age of both, it is noticed that the average age of medal winners are slightly higher than non-medalists. However, with such small differences, there is little to suggest age plays a significant role in predicting the performance of athletes and whether they will win.





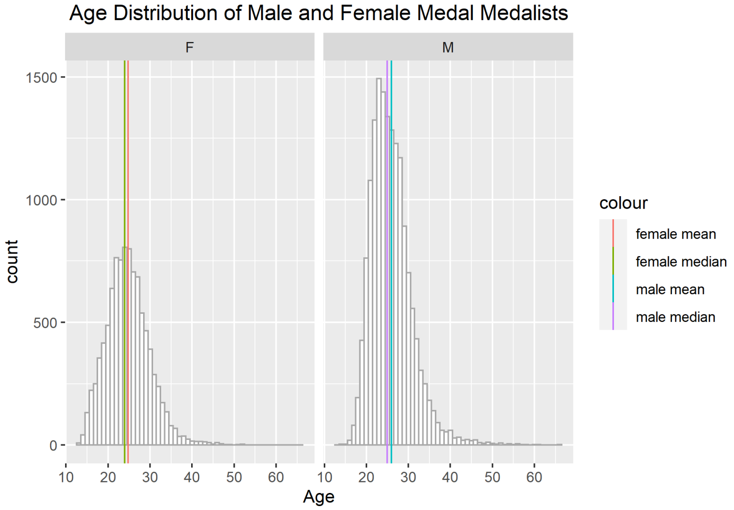
Similar to age, the height and weight distribution between non-medalists and medalists are similar too. However, a few things are observed. The height distribution of medalists is a smooth bell-shaped curve with the majority of medalists having a height of 180cm. However, the height distribution of non-medalist shows that heights slightly lower or higher than 180cm are more common among non-medalists. This could suggest that there is an optimal height that could predict performance.

We also observe from the figure on weight distribution that the weight range for medalists is slightly smaller than non-medalists, with the left tail of the medalists weight distribution shorter than that of non-medalists. It is also noted that the mean height and weight of medalists are also higher than non-medalists. This could indicate that there is a certain range of weight and height that medalists tend to have, predicting performance.

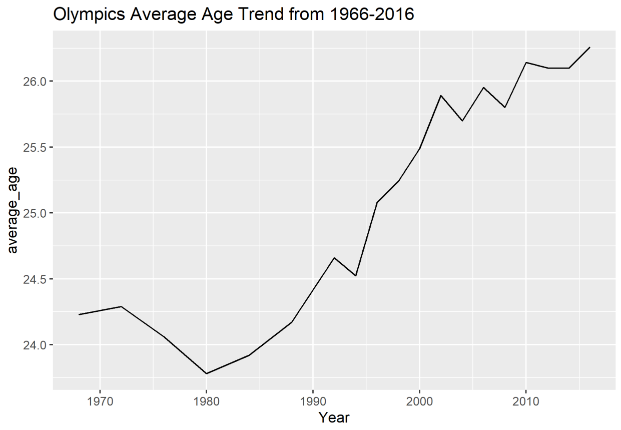


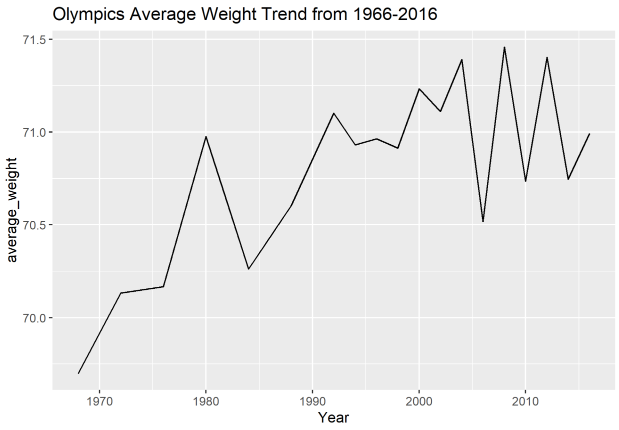
The above figure shows the amount of medals won out of the total number of medals awarded in the past 50 years by age. We aim to see if there is a certain age range that wins more medals than others. From the figure above, we can see that indeed there is a difference. Those who are 22 to 24 years old earns the most medals compared to other ages. This is perhaps because athletes at these ages tend to have more experience and their bodies are in prime condition (not too young, not too old). This implies that age does have a certain effect on performance and that it is an important variable in a model predicting winning an Olympic event.

Having identified the most common age among medalists in general, we use histograms next, to see if there are differences in the age distributions among male and female medalists to gain insight on what ages perform the best for each.



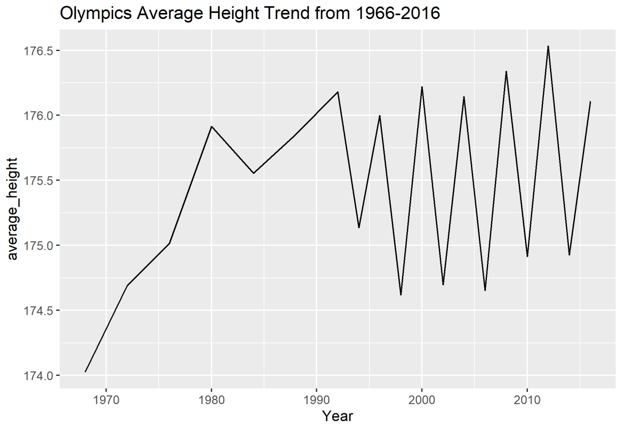
From the age distributions of medalists we can see that the male distribution is more left skewed with a higher mean and median than the female age distribution of medalists. The peak of the female age distribution is at age 24 and 25 while the peak for the male age distribution is 23 and 24, slightly younger. The male distribution had a mean age of 25.9669055 while the female distribution had a mean age of 24.7355509.



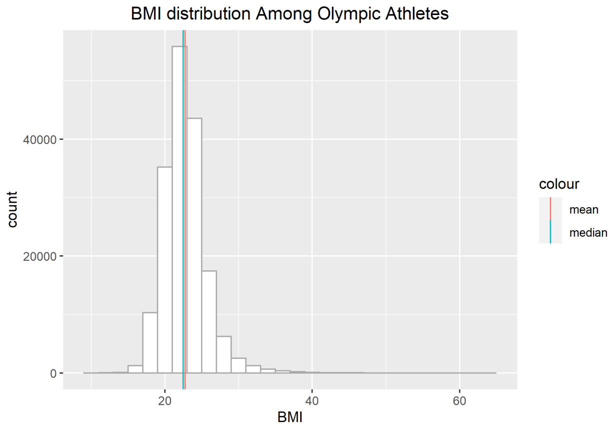


As the Olympics gets increasingly more competitive over the years, instead of only comparing the difference in age, height, and weight between medalists and non-medalists, we decided to see if there are any trends in age, height, and weight over the years that could indicate its importance for Olympics-eligible performance. From the above line graph, we can see that the age for Olympians has been increasing since 1980. This could indicate that experience and a grown body is related to better performance. It also may be due to advancements in certain fields and technologies, life better rehabilitation for atheletes, that have allowed for athletes to continue performing for longer moving up the average age of competitors.

We also observe that the average weight for Olympians has also been increasing over the years. This is in alignment with our previous findings where medalists have a heavier weight on average than non-medalists. This indicates a potential relationship between weight and performance.

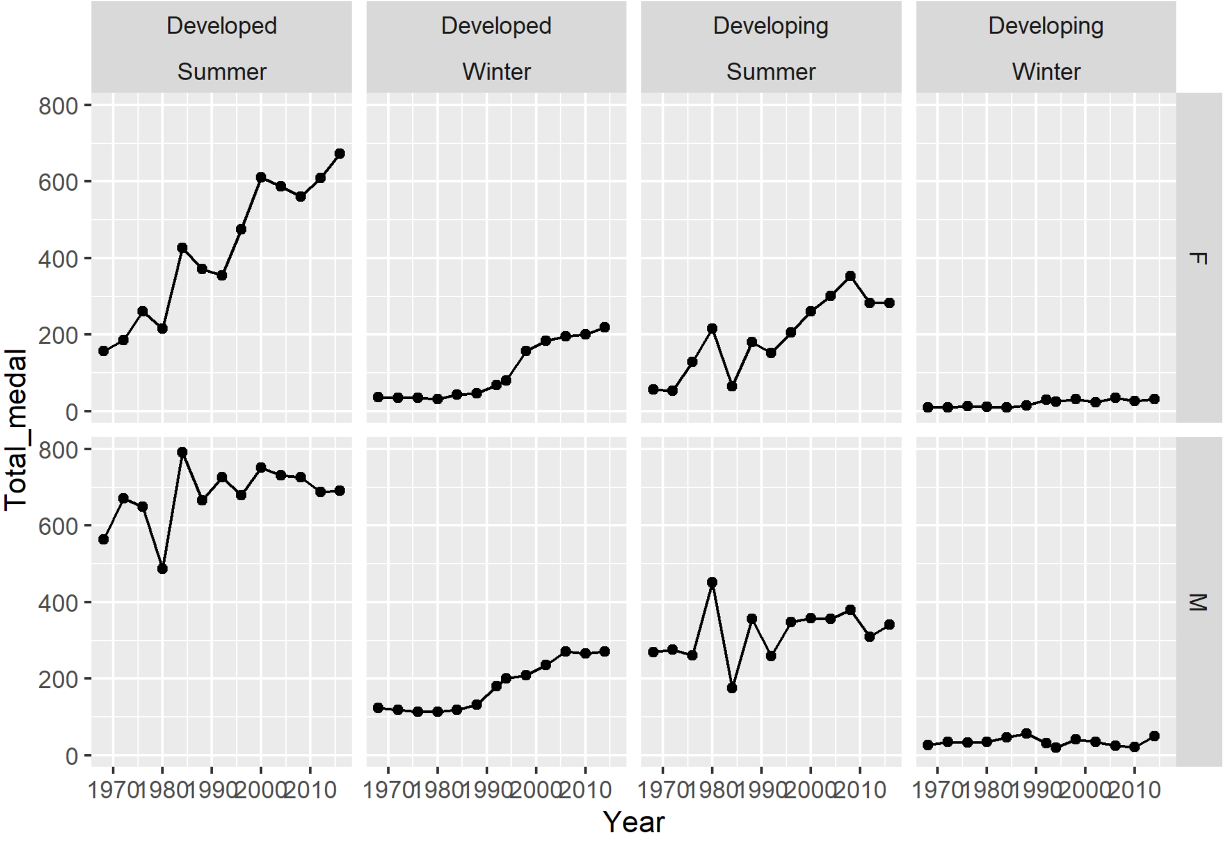


We also observe that there have been an increase in average height over the years for Olympians. However, it seems to be showing signs of slowing down (fluctuations of similar scale since 1994). This could be because there is a limit as to how tall a person can grow. The overall trend observed in the line graph shows that height could be a factor for great performance (qualify to participate in the Olympics), but its impact could be limited/ smaller than other factors for winning considered. We should also take into account that the increase in the average height of athletes could also be attributed to how the average height of humans is increasing with each generation.



Above is a histogram showing the BMI distributions of all athletes. Our mean BMI is 22.7470217 with a standard deviation of 2.9535037. The BMI follows a relatively normal distribution with mean and median values found grouped together near the peak. The most common BMI among athletes is 22.

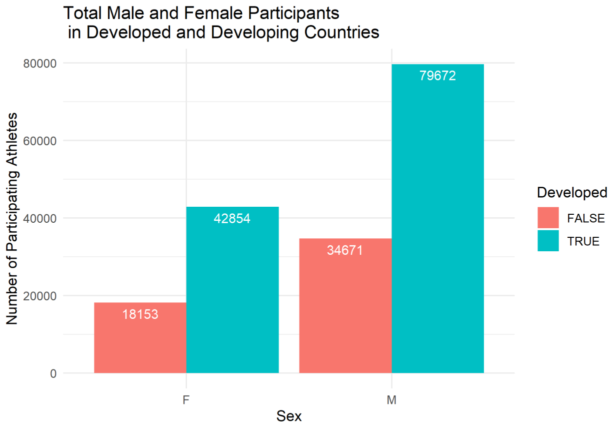
Male Versus Female Performance Over the Years

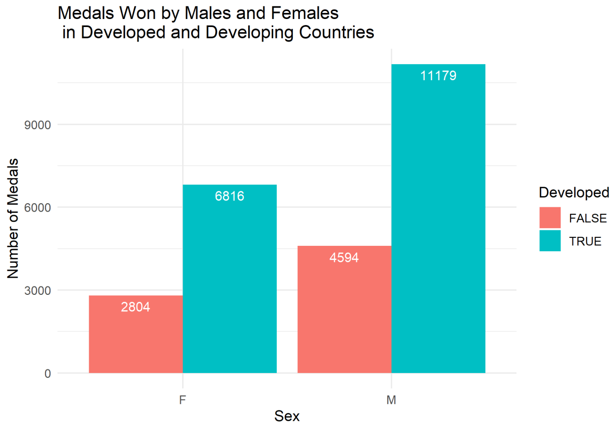


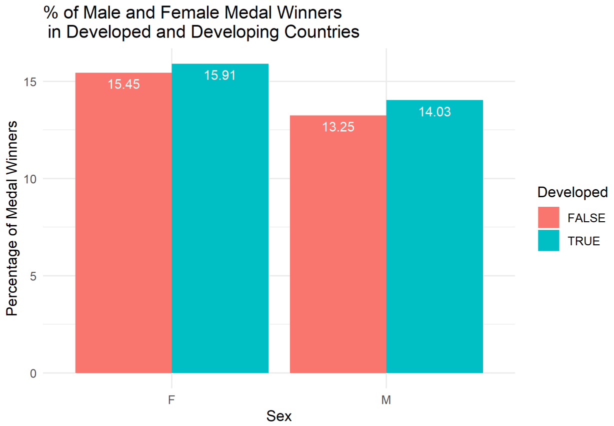
To compare the performance of male and female athletes from and developing countries, the total number of gold medals won per year among the sub groups of male from developing country, female from developing country, male from countries, and male from developing countries was used to measure performance. We further split it by season to account for any potential season-related differences such as sports and countries participating in each season.

From the line graphs above, a few interesting observations are derived. Over the years, the performance of female athletes from countries in the Summer season have been improving exponentially. On the other hand, the performance of male athletes in the Summer season seems to be stagnating and hasn't shown much improvement over the years, especially when compared to the performance of females.

For the Winter season, however, we see that the performance of both male and female athletes from countries have been improving at a similar rate over the years. The same can be said for performance of both male and female athletes from developing countries during the Winter season. During the Summer season however, the rate of improvement for female athletes from developing countries (by seeing the steepness of the slope) seems to be higher than that of male athletes from developing countries.







Countries were sorted into two categories, developing and developed using the 2018 GDP of 12,000 or less for developing and greater than 12,000 for developed. We next calculated the total number of athletes who participated in Olympic games, the number of medal winners and the percentage of medal winners for males and females. Our first bar graph shows that there are approximate half the number of participants from developing countries than developed countries for both males and females. Our next bar graph shows that both the male and female athletes from developing countries won less medals than those of developed countries. The number of medals won by athletes from developing countries is about 41% of those from developed countries for both males and females. The percentage of women vs men winning medals in developing and developed countries both average at about 61%. Therefore, although the total number of medals for developing countries is lower, the ratio of female to male medals is comparable for developing and developed countries. The last bar graph shows the percentage of medal winners by comparing them to the total number of participants in the same group. This graph depicts that there is little statistical difference in the probability of winning a medal based on whether an athlete is from a developing or developed country.

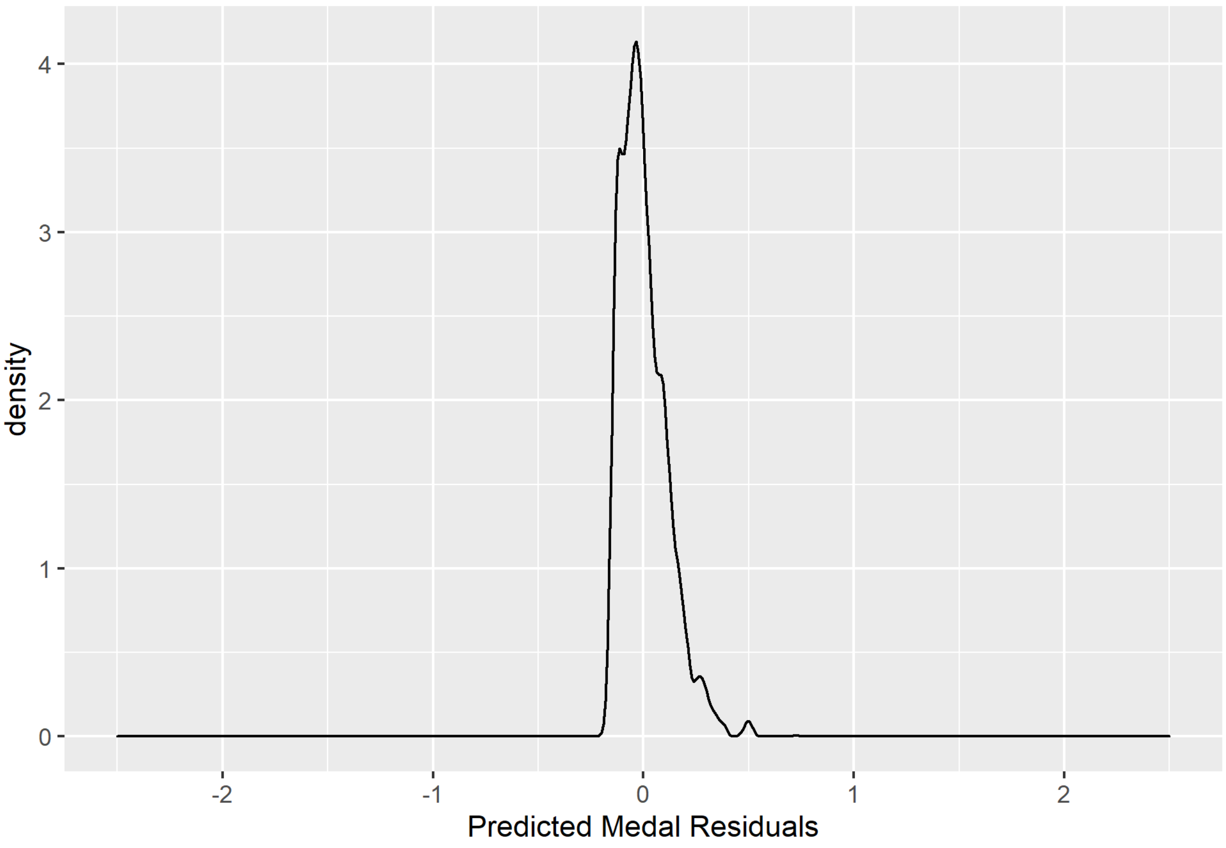
Advanced Models and Analysis

Model: How Well Does Age, Height, Weight, BMI, and GDP Predict Olympic Winnings

Linear Model We will be using all of the following variables: Age, Weight, Height, BMI, Developmental status of home country, and Sex in a linear model to see how well it can predict the average medals won per country.

```
Call:
lm(formula = Medal_countryavg ~ Age + Weight + Height + BMI +
    Developed + Sex, data = olympics_train)
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.23965 -0.08543 -0.02036  0.06509  0.77911

Coefficients:
                 Estimate  Std. Error t value             Pr(>|t|)
(Intercept)    0.30618518  0.03182089   9.622 < 0.0000000000000002 ***
Age            0.00021627  0.00005713   3.785             0.000154 ***
Weight         0.00288825  0.00022248  12.982 < 0.0000000000000002 ***
Height        -0.00100567  0.00018130  -5.547         0.0000000291 ***
BMI           -0.00778342  0.00071100 -10.947 < 0.0000000000000002 ***
DevelopedTRUE  0.00547736  0.00065536   8.358 < 0.0000000000000002 ***
SexM          -0.03364325  0.00076480 -43.990 < 0.0000000000000002 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 0.1128 on 141456 degrees of freedom
Multiple R-squared:  0.02071,   Adjusted R-squared:  0.02067
F-statistic: 498.5 on 6 and 141456 DF,  p-value: < 0.00000000000000022

[1] 0.1124627
```



With this linear model, we obtained a rˆ2 value of 0.02023 and rmse value of 0.1133922. Although these values are not ideal (do not indicate high predictive power), this is the best fit we could obtain using the variables from the data we are currently using.

According to the summary results we obtained from this model, all variables included have a low p-value (below the 5% threshold for significance). Looking at these p-values, there is no one variable that stands out in showing high predictive ability.

Above is a plot of the residuals from our model. The majority of the values are around 0 which shows that our regression model is pretty accurate at predicting the data. There seems to be a higher density of residuals that are negative (below the regression line on the left side) than positive ones. This means that the majority of this time, our model will over predict the average medals won per country.

We have run other linear models with different combinations of predictor variables. However, the rˆ2 and rmse values for those models suggest poorer predictive ability (rˆ2 value was lower and rmse value was higher) than the model we currently have.

Model: Does Sex and Year Predict Number of Wins Per Team?

Linear Model 2

```
Call:
lm(formula = win ~ Sex * Year + Season, data = olympics_train)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.1649 -0.1432 -0.1429 -0.1145  0.8860 

Coefficients:
                Estimate  Std. Error t value            Pr(>|t|)    
(Intercept)   0.22495538  0.23320376   0.965               0.335    
SexM         -0.10143105  0.28138453  -0.360               0.718    
Year         -0.00003053  0.00011670  -0.262               0.794    
SeasonWinter -0.02868949  0.00230561 -12.443 <0.0000000000000002 ***
SexM:Year     0.00004029  0.00014092   0.286               0.775    
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 0.3511 on 141458 degrees of freedom
Multiple R-squared:  0.001908,  Adjusted R-squared:  0.00188 
F-statistic: 67.61 on 4 and 141458 DF,  p-value: < 0.00000000000000022
```

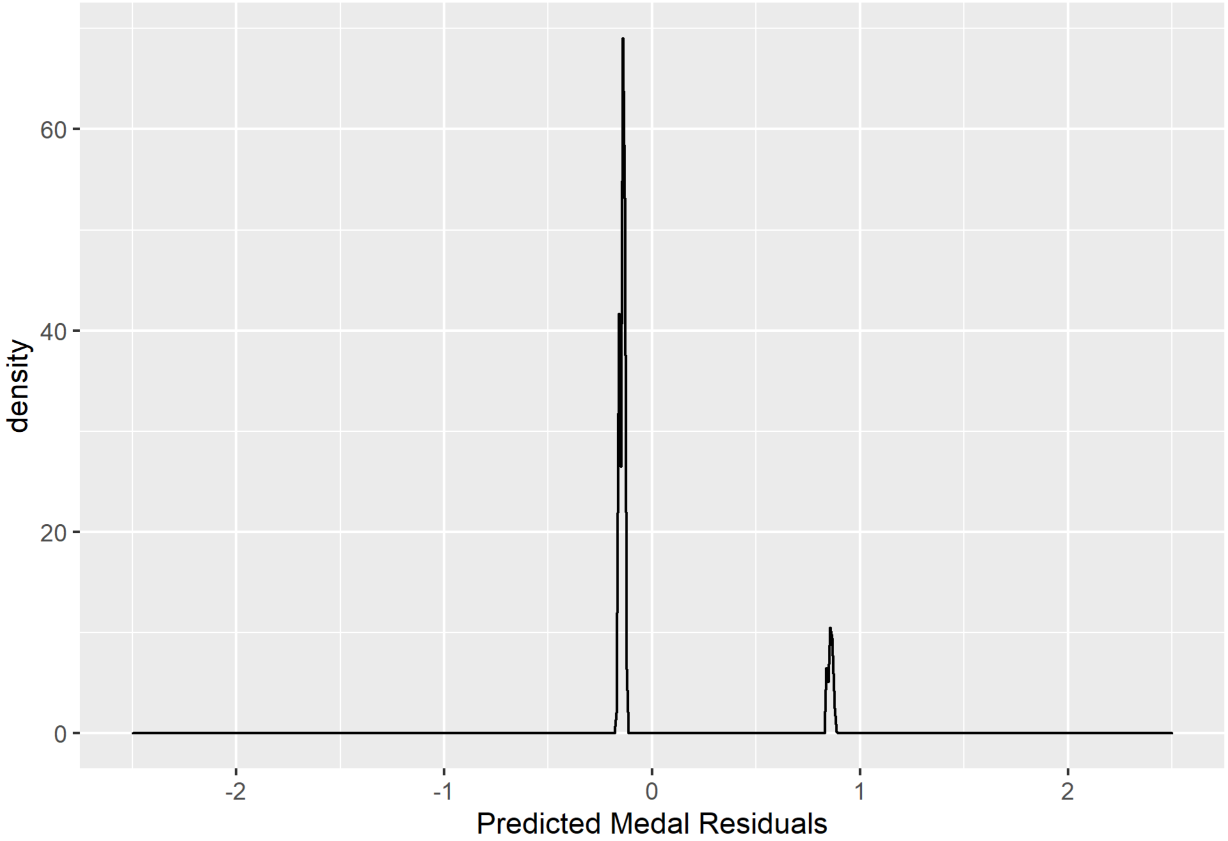
```
[1] 0.3536402
```

```
Call:
lm(formula = win ~ Sex * Year + Developed, data = olympics_train)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.1616 -0.1516 -0.1396 -0.1308  0.8700 

Coefficients:
                 Estimate   Std. Error t value  Pr(>|t|)
(Intercept)    0.209377589  0.233812448   0.895     0.371
SexM          -0.006067689  0.281589302  -0.022     0.983
Year          -0.000028844  0.000116952  -0.247     0.805
```

```
DevelopedTRUE  0.008947230  0.002036940   4.392 0.0000112 ***
SexM:Year     -0.000007534  0.000141018  -0.053     0.957
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3513 on 141458 degrees of freedom
Multiple R-squared:  0.0009519, Adjusted R-squared:  0.0009236
F-statistic: 33.69 on 4 and 141458 DF,  p-value: < 0.00000000000000022
```

```
[1] 0.3539753
```

By comparing the r^2 and rmse value of the two linear models, mod_olympics_sex2 and mod_olympics_fandm, we see that the first model performs better in predicting whether an athlete wins or not. mod_olympics_sex2 has an r^2 value of 0.00188 and rmse value of 0.3536402 while mod_olympics_fandm has a r^2 and rmse value of 0.0009236 and 0.3539753 respectively. The higher the r^2 value and lower the rmse, the better the fit of the model. From this, we see that season a slightly better predictor variable than the development of a country. The poor predictive ability of the model shows that gender and year does not play a big role in determining whether an ahtlete wins.



From this residual plot of our linear model for predicting whether an athlete will win based on the interaction variable of Sex and Competing year, and developmental status of their home country, we can see that the majority of the residuals are close to 0, on the negative side. This means that our model will over predict if someone is a medalist. Because our dependent variable, win, is binary, we see two peaks in the residual plot.

Random Forest We then ran a random forest model to see how well the same variables would be at predicting whether an athlete would win at the Olympics.

```
Call:
 randomForest(formula = as.factor(win) ~ Age + Weight + Height +      BMI + Developed + Sex, data = oly
               Type of random forest: classification
                     Number of trees: 200
No. of variables tried at each split: 2

        OOB estimate of  error rate: 14.52%
Confusion matrix:
       0   1 class.error
0 120594 453 0.003742348
1  20092 324 0.984130094
```

```
          MeanDecreaseAccuracy
Age                   54.41108
Weight                32.60656
Height                41.68960
BMI                   41.63883
Developed             45.26780
Sex                   69.10010
```

```
Warning in Ops.factor(response(model, data), stats::predict(model, data)): '-'
not meaningful for factors
```

```
[1] NaN
```

The most important variables to predict whether someone will be a medalist are in the following order: Sex, Age, Developmental Status of home country, Height, BMI, and Weight. The OOB error is really high and stays around 14%-15% regardless of how many trees we have on the model. We would need additional variables to better predict whether an individual would be a medalist. There is a much higher error rate on class 2 than class 1. From this information, we have concluded that we cannot accurately predict a medalist from a non-medalist with our current data and variables.

When calculating the rmse for forest_olympic_data_medal, we were met with the message "not meaningful for factors". This further supports our results from our linear model analysis, showing that the current variables investigated are insufficient for predicting whether an athlete wins at the Olympics.

Next we will create another random forest model to see the predictive ability of the variables and how they predict the average medal count per country per year.

```
Call:
 randomForest(formula = Medal_countryavg ~ Age + Weight + Height +      BMI + Developed + Sex, data = ol
               Type of random forest: regression
                     Number of trees: 200
No. of variables tried at each split: 2

          Mean of squared residuals: 0.01144075
                    % Var explained: 11.91
```

```
           %IncMSE
Age       76.48066
Weight    45.39110
```

```
Height    47.31625
BMI       56.21708
Developed 85.71497
Sex       93.29617
```

```
[1] 0.1061367
```

We can see that the most predictive variable is still Sex, followed by the developmental status of a country, Age, BMI, Height and Weight. The % Var explained for this randomForest model is around 12%, which shows that this is a poorly predicative model. The variables chosen do not predict either dependent variables well.

However, by calculating the rmse value for forest_olympic_medal, we see that our current variables are better at predicting the average number of medals won per team per year.

Conclusion:

Predicting whether an athlete will win the Olympics is a very intricate task. After all, there are many factors that can contribute to this, from an individual's preparation time to their genetics, and much more. Our data set had limited variables for us to consider and thus even after running multiple models, we still feel like more variables would have to be considered for one to better predict winning attributes of an athlete.

With the linear models we ran, based on the r-square values of each variable, we saw that the top two most important variables were Weight and Height. But when we ran the ran the Random Forest model, we saw that Age and Sex were the most important. While these results are not mutually exclusive, we thought it was interesting that they did not match up. We also did not have a high RMSE value when we ran different versions of the random forest models. The only time we did see above 50% for the RMSE was when we included the variable of NOC or Home Team's Country which we later decided to take out since it would create some bias in the model. For example, if certain countries were more susceptible to winning within the 50-year period we looked at, then the athlete representatives from those countries would have a "boost" simply from being from their home country and not the individual's statistics. Additionally, Team Country were included in the calculation of the dependent variable, which could be the main reason why there was such a high boost in the predictive power of the model.

Other factors and further comments:

Limitation to our data/bias:

We should note that a country's GDP changes year over year, but we are only using their 2018 GDP throughout our analysis. Countries could have changed from a "developing" to a "developed" country within the 50 years we analyzed and this categorization could have caused a bias in our models.

We also based performance on whether an athlete would win any medal. Therefore, our two performance buckets were winners and participants. A bronze, silver or gold medal all counted the same in our analysis. We did this because picking values for each medal would be strictly arbitrary. For example, if we assigned values: 1,2,3, respectively, we would be assuming (with no proven data) that a gold medal is 3 times harder to win than a bronze one.

Additionally, we looked at winners at the Olympics as a whole. There are many sports and events in the Olympics that require different weights, heights, ages for optimal performance. We tried to do exploratory analysis but realized the data varied too much.

Overall Analysis

We believe that the predictive ability of our models needs to be improved upon before predicting the next Olympic winners. We believe that more information about a myriad of other factors may help us to improve our predictive abilities in Olympic Medal winners. To conclude, we have found that the variables we used do not provide enough predictive ability to accurately and consistently predict a a medal winner and more data would be needed to improve upon that.