ISyE 312 Final Report

Gabriela Barreiro Pujol, Linus Gustafsson, Anthony Hartono, Nandini Jagtiani

Problem statement

In this project, we have utilized the historical Olympic data of every Olympic Games from 1896 until 2012. Each row represents one athlete and their event and includes the following data points:

- ID - Unique number for each athlete
- Name - Athlete's name
- Sex - M or F
- Age - Integer
- Height - In centimeters
- Weight - In kilograms
- Team - Team name
- NOC - National Olympic Committee 3-letter code
- Games - Year and season
- Year - Integer
- Season - Summer or Winter
- City - Host city
- Sport - Sport
- Event - Event
- Medal - Gold, Silver, Bronze, or NA

The dataset is substantial with 271,116 unique rows.

Using this dataset we wanted to analyze the different sports within the Olympics to find some insights into the universal truth of what it takes to be a successful Olympian. Within the dataset, there are an infinite number of questions you could answer regarding this, but what we found most interesting to look into was if there is an age correlation to gathering medals as well as taking in the total number of athletes and the athlete's BMI in regards to medals. Age together with the vast number of sports is an interesting data point to look at because it generally handles the tradeoff between mental maturity against physical athleticism. Looking deeper into the dataset we got curious if we could find any signals to if there was a home team advantage for the host country. This led us to the following problem statement *"Use regression techniques to explore the relationship that physical and mental influences have on Olympic performance."*

From our curiosity and the mentioned problem statement, we aim to answer the following questions through our analysis:

- Does the age of an athlete relate to winning medals?
- Can the number of athletes in a team, the average age, and the average BMI of a country's Olympic team influence the total medal count?

- Is there an advantage for the host country?

Some key features we will use to answer these questions are Age, Weight, Height, Team, City, and Medal, as we as a few new variables that we will create using feature engineering.
By answering these questions, the results can indicate what variables will lead a country to perform better in the Olympics and give a better understanding of how different physical and mental influences can influence performance.

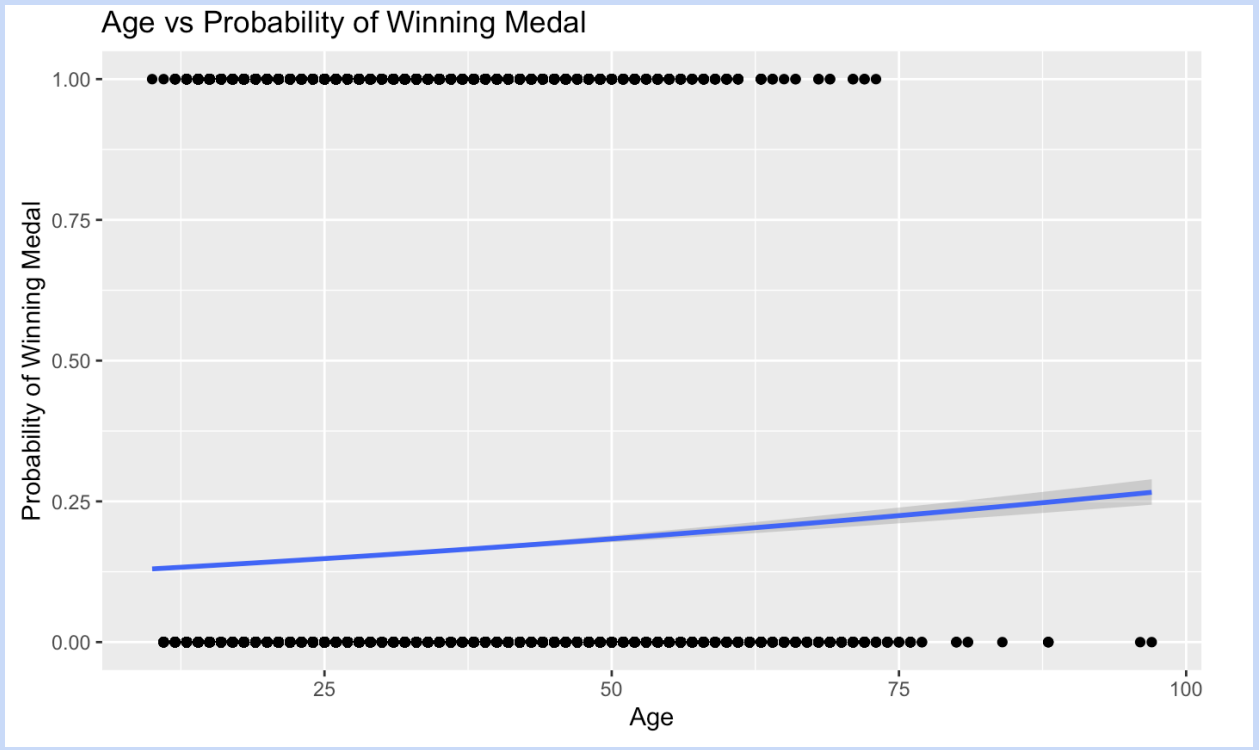
Simple Logistic Regression

"How does age relate to winning medals?"

To address this question, we did a simple logistic regression to see age vs the probability of winning medals. Since this is a logistic regression, we transformed the target variable into a binary variable. Originally, the target variable was Gold, Silver, Bronze if the athlete won a medal, and NA if the athlete did not win any medals. We transformed the data into 1 if the athlete won any medal at all, which is gold, silver, or bronze, and 0 if the athlete did not win any medals at all.

```
Call:
glm(formula = Medal ~ Age, family = "binomial", data = df)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.0021663  0.0220389  -90.85   <2e-16 ***
Age          0.0101811  0.0008256   12.33   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 220518  on 261641  degrees of freedom
Residual deviance: 220370  on 261640  degrees of freedom
  (9474 observations deleted due to missingness)
AIC: 220374
```

From the model adequacy test, we can see that the model fits the data reasonably well. The deviance is 220370, with a p-value of 1. This means that the model is adequate, with a p-value of more than 0.05. In terms of the deviance and degrees of freedom, we can see that if we divide the deviance by the degrees of freedom (220370/261640), the result is 0.84. Since the result is less than 1, it shows that the model has adequate fit. From the z-test, we can see that both the intercept and age variables have a very small p-value, which shows that both the explanatory and predictor variables are statistically significant.



From the plot, however, we can see that the best-fit line does not converge to 1. This is primarily down to the fact that the coefficient for the only predictor variable, age, is very small, at 0.01. This means that for every one-unit increase in age, the log odds of winning a medal increase by approximately 0.01, which is very small. The small coefficient can be attributed to the low number of features in this model, so a better plot might be to include more features to predict the probability of winning medals. From the plot, we can see that age alone is not enough to predict the probability of winning medals, so more features might be able to better predict the probability of winning medals. We also found that the correlation between age and winning medals is very low, at 0.02. From the findings, we can safely conclude that age does not necessarily relate all that much to winning medals, as there are a host of other factors that might come into play in predicting the probability of winning medals.

Multiple Linear Regression

“How do the number of athletes, average age, and average BMI of a country's Olympic team, influence its total medal count?”

To address this question, a multiple linear regression analysis was conducted, incorporating variables such as Total Athletes, Average Age, and Average BMI from both datasets. This analysis aims to uncover any statistically significant correlations between these factors and the team's performance, measured in terms of medal count.

```
Call:
lm(formula = Total_Medals ~ Total_Athletes + Average_Age + Average_BMI,
    data = country_summary)

Residuals:
    Min      1Q  Median      3Q     Max
-104.48   -4.20    2.62    4.59  324.50

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -13.436940   4.624326  -2.906  0.00369 **
Total_Athletes   0.308226   0.003535  87.204  < 2e-16 ***
Average_Age      0.031823   0.104877   0.303  0.76158
Average_BMI      0.301262   0.191894   1.570  0.11653
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19.06 on 3083 degrees of freedom
  (218 observations deleted due to missingness)
Multiple R-squared:  0.7133,    Adjusted R-squared:  0.7131
F-statistic:  2557 on 3 and 3083 DF,  p-value: < 2.2e-16
```

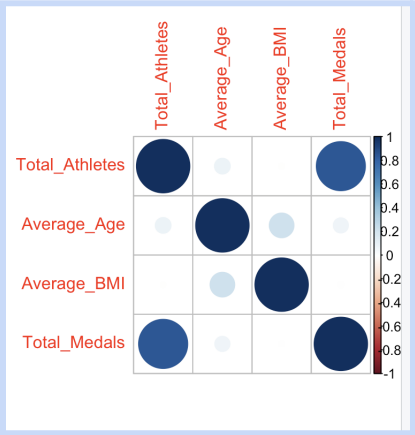
Initial analysis using multiple linear regression suggested that the chosen variables indeed influence Olympic medal outcomes, with an R-squared value of 0.7133, indicating a substantial effect. However, a subsequent review identified outliers and missing values within our dataset. To refine our insights, a revised model was constructed, excluding these anomalies.

```
Call:
lm(formula = Total_Medals ~ Total_Athletes + Average_Age + Average_BMI,
    data = country_summary_clean)

Residuals:
    Min      1Q  Median      3Q     Max
-32.157  -3.108   1.460   2.830  63.360

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -7.447242   2.203077  -3.380 0.000733 ***
Total_Athletes  0.257317   0.002173 118.417  < 2e-16 ***
Average_Age     0.036537   0.050294   0.726 0.467612
Average_BMI     0.124202   0.091214   1.362 0.173412
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.04 on 2960 degrees of freedom
  (218 observations deleted due to missingness)
Multiple R-squared:  0.827,     Adjusted R-squared:  0.8269
F-statistic:  4718 on 3 and 2960 DF,  p-value: < 2.2e-16
```

The new model's results were more robust, demonstrating an increased R-squared value of 0.827. This indicates that 82.7% of the variance in the total medal counts won by each country's Olympic team can be explained by the factors of Total Athletes, Average Age, and Average BMI. This high degree of correlation suggests a strong influence of these variables on a country's Olympic success, pointing to the importance of team composition in determining medal outcomes.

Multicollinearity

To check if this model is useful, we check for multicollinearity to make sure that none of the variables are highly correlated with each other



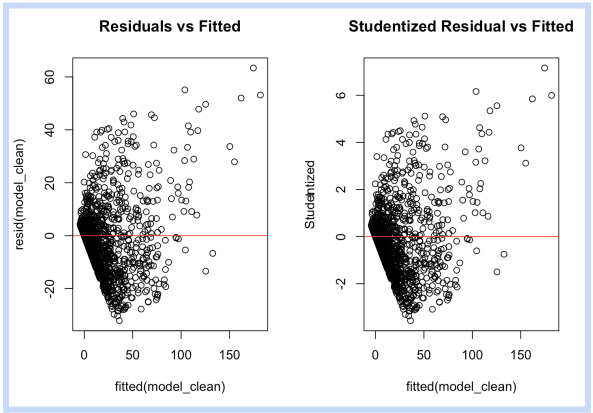
To ascertain the utility of our model, we conducted a thorough examination of multicollinearity, ensuring that the predictor variables are not inordinately correlated with each other. The analysis revealed low inter-variable correlations, suggesting that multicollinearity is not a concern.

```
> vif(model)
Total_Athletes    Average_Age    Average_BMI
      1.008197       1.055736       1.048016
```

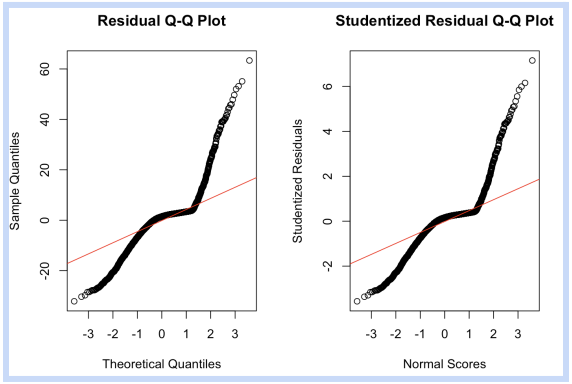
Additionally, we evaluated the Variance Inflation Factor (VIF) values, and as all were under the threshold of 5, it further confirms that multicollinearity does not significantly impact our model, affirming the reliability of our analysis.

Residual Plots and QQ plots - Multiple Linear Regression

Following the verification of VIF values for each variable, we proceeded to generate residual and studentized residual plots. These plots serve to evaluate the suitability of our multiple regression model, ensuring that it accurately reflects the data after adjustments made in the new model.



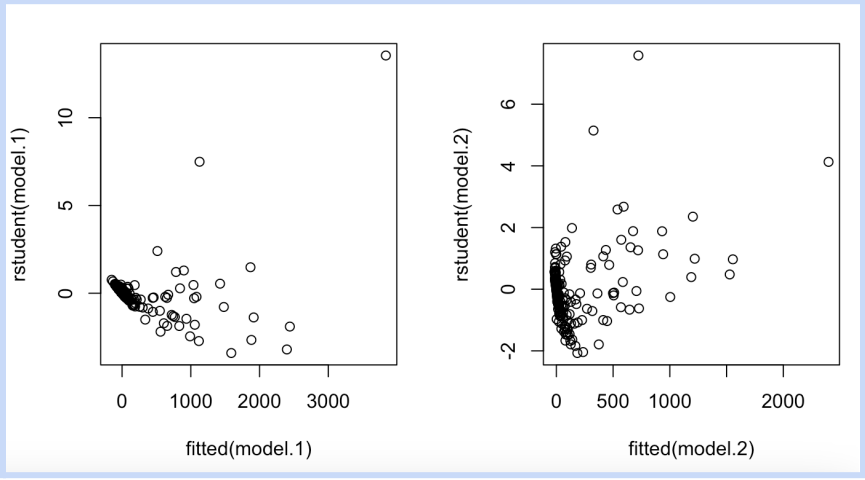
Furthermore, Q-Q plots were employed to assess the normality of the residual distribution. This analysis helps ensure that the residuals conform to a normal distribution without any significant skewness.



To address the issues identified in the residual and QQ plots, a different analytical approach is necessary. The residual plots displaying a funnel shape suggest heteroscedasticity, where the variance of the residuals is not constant across the range of values. This can lead to inefficient and biased estimates in ordinary least squares regression.

The QQ plots indicating skewness imply that the residuals do not follow a normal distribution, which is another assumption in OLS regression. Deviations from normality can affect the reliability of hypothesis tests.

To rectify these issues, a Weighted Least Squares analysis is performed



Anova - Multiple Linear Regression

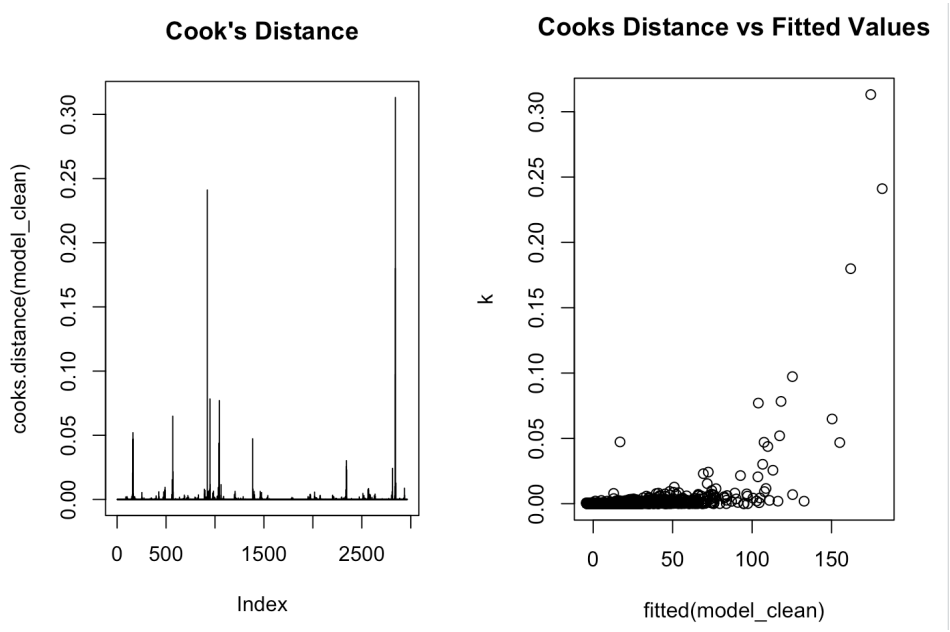
An additional ANOVA test was conducted for the multiple linear regression model. This highlighted view of the F-values from the test reveals that there is a large F-value and very small

p-value for the Total_Athletes variable in the model, further indicating only that variable is significant.

```
> anova(model_clean)
Analysis of Variance Table

Response: Total_Medals
                Df  Sum Sq Mean Sq    F value Pr(>F)
Total_Athletes   1 1156453 1156453 14150.6598 <2e-16 ***
Average_Age      1      88      88     1.0824 0.2983
Average_BMI      1     152     152     1.8541 0.1734
Residuals     2960  241904      82
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

Cook's Distance

Upon concluding the diagnostic tests for the multiple regression model, the last measure was to verify the absence of significant outliers that might undermine the model's performance. This was accomplished by evaluating Cook's distance for each observation



The guideline is that Cook's distance should not exceed 1 to avoid the influence of outliers. As indicated by the plot, all Cook's distances are well below this threshold, with the highest value not surpassing 0.04, confirming that outliers do not present a concern in our dataset.

Lastly, to address the question of how the number of athletes, average age, and average BMI of a country's Olympic team influence its total medal count: The data analysis reveals that the Total Athletes is a statistically significant predictor of Olympic success, while Average Age and Average BMI do not have a significant impact on the total number of medals won by a country's Olympic team.

Proportion Testing

“Is there an advantage for the host country?”

Our final question searched to answer if athletes competing in their home country had an advantage over athletes competing in a foreign country. In answering this question, we aim to find what mental influences an athlete faces and how that may impact their performance. To study this relationship, we created a binary variable in which “1” indicated that the athlete was competing for the country that was hosting that year’s Olympic games, and “0” if they were not. We then followed the same strategy used above in which athletes any athlete who received a medal had the value “1” for the new “placed” binary variable, and “0” if they did not. After conducting these transformations, we summarized the dataset to find the proportion of both home and away athletes who received a medal. The results of this summary are stated in the table below:

Home Team?	Athletes Placed	Total Athletes	P(Athletes Placed)
Yes	1059	6706	0.1579183
No	38715	264040	0.1466255

We can see that the proportion of athletes who competed in their home country that received a medal is greater than those who did not compete in their home country. To test if this was a statistically significant difference, we conducted a two-proportion test, with the following hypothesis statement:

$$H_o : P(Place)_{Home} = P(Place)_{Away} \quad H_a : P(Place)_{Home} > P(Place)_{Away}$$

To complete this test, we used the prop.test function in R and received the following results:

```
data:  c(1059, 38715) out of c(6706, 264040)
X-squared = 6.565, df = 1, p-value = 0.0052
alternative hypothesis: greater
95 percent confidence interval:
 0.00380463 1.00000000
sample estimates:
   prop 1    prop 2
0.1579183 0.1466255
```

Because the p-value is less than our decided alpha of 0.05, we reject the null hypothesis. With 95% confidence, there is sufficient evidence to support that the proportion of medalists is greater when the team is competing in their home country.

We took this conclusion a step further and attempted to use this knowledge to predict if an athlete would receive a medal based on whether or not they were competing in their home country. This problem pushes what we learned throughout this course, as we are mapping a binary feature to a binary target variable. To do this, we used a generalized linear regression model with the following output.

```
Call:
glm(formula = Place ~ Home, family = "binomial", data = places)

Coefficients:
             Estimate Std. Error  z value Pr(>|z|)
(Intercept) -1.761317   0.005502 -320.145  < 2e-16 ***
Home         0.087517   0.033936    2.579  0.00991 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 225966  on 270745  degrees of freedom
Residual deviance: 225960  on 270744  degrees of freedom
AIC: 225964
```

We conducted a few different tests to see if this was a good model. We started with Cook's Distance, in which we found that no individual point has $D_i > 1$. This means that no one point is overly influential to the model. However, this is expected because, as stated above, we are using binary variables as both the predictor and dependent variables. As a result, it is near impossible for one point to be more influential than another, as they are all valued at either 0 or 1.

We also looked at the deviance as stated in the results above. In dividing the residual deviance of 255960 by the degrees of freedom of 270744, we got a value of about 0.8346. Since this value is less than 1, it indicates an adequate fit for the model. For the same reason as Cook's Distance, we must also take this result lightly, as deviance relates to residuals.

Finally, we conducted a 95% confidence interval of the coefficients using the function `confint`. This gave us an interval for the intercept coefficient of [-1.77211341, -1.7505474] and the interval [0.02051883, 0.1535583] for the "Home" coefficient. Focusing on the latter, we see that 0 is not included in this range. This means that we can say with at least 95% that β_1 is greater than 0, and as such athletes are more likely to win a medal when competing in their home country.

All the test we conducted above supported the results that an athlete who competes in the Olympics in their home country have an advantage over athletes who are competing in a foreign country, and they are more likely to win a medal as a result.

Conclusion

In summary, our group answered the questions that we outlined at the beginning of the project with various regression techniques. To answer if the age of an athlete relates to winning medals we used simple logistic regression. We found that there is no significant relationship, which we can attribute to other factors that might contribute to the probability of winning medals, such as skill, experience, and ability to cope under pressure.

To answer if the number of athletes, average age, or average BMI can influence the total number of medal counts of a country, we used multiple linear regression. We found that the only variable that is statistically significant in influencing the total number of medal counts of a country is the number of athletes. This makes sense, as the more athletes a country has competing increases the chance of that country winning medals.

We addressed the final question of whether there is an advantage for athletes competing for the host country by using hypothesis testing. From this result, we found that there is sufficient evidence to support that the proportion of medalists is greater when the team is competing in their home country. Using the hypothesis test, a generalized linear model, and a confidence interval, we concluded that athletes competing for the host country have an advantage over those who are competing in a foreign country.