Analysis of 120 Years of Olympic Results

Kassidy Chaikin, Christopher Egan, Emily Lepore, William Wenzel

Introduction:

For this project we have decided to look at the 120 years of Olympic history: athletes and results in R studio. It is a dataset on modern Olympics from 1896 to 2016. There are 271,116 rows and 15 columns with each row being an individual athlete. This dataset was posted by Rgriffin on Kaggle.com. The data was obtained by scraping and wrangling from a sport-statistic website. We decided to drop certain variables for different tests, to streamline the data more. The columns are ID, Name, Sex, Age, Height, Weight, Team, NOC, Games, Year, Season, City, Sport, Event, and Medal. We would like to research how over time gender inequality in participation in the Olympic Games has changed and see which countries have a higher percentage of women. We want to look specifically at four sports, two from Summer Olympic Games and two from Winter Olympic Games, to see how biological factors influence medal winnings for male and female athletes. We also want to look further into two specific athletes that have competed in multiple Olympic Games to see how age can have an impact on their performance and whether they are more or less likely to win a medal.

Dataset and Features:

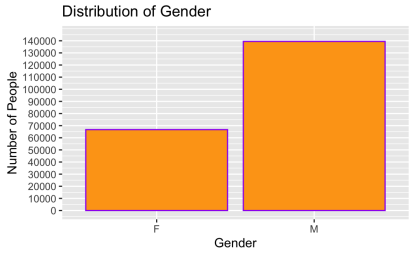

Fig. 1

In this plot, we can observe the sex distribution of the participations. We choose histogram because we want to see clearly the sex distribution from the plot by looking the columns. As shown in the plot, the male competitor ratio is twice the female competitor ratio.

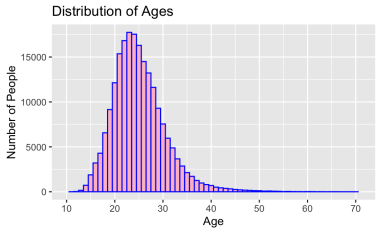

Fig. 2

In this plot, we want to see the distribution of ages by number of people. We choose histogram because we want to see the age distribution from the plot by looking at the columns. We can see that there is the highest number of athletes between the ages of 20-30. The 23-year-old athletes participate in the Olympics. There is minimal participation of athletes over the age of 56. In addition, the youngest athletes participating in the Olympics are 10 years old.

Summer: Winter:

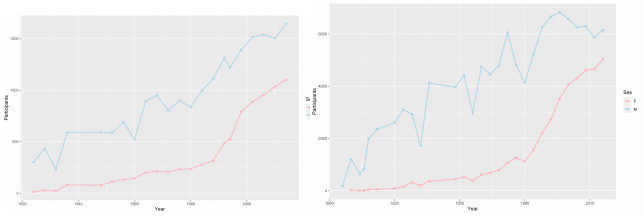

Fig. 3 Fig. 4

Overall, there has been a steady increase in both male and female athletes participating in the Olympic Games in both the summer and winter. In both summer and winter games there continuously has been a larger number of male participants to female participants.

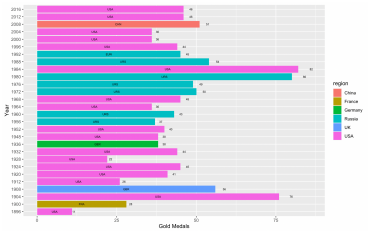

Fig. 5

The USA dominates the summer Olympic Games. Out of the 28 Summer Olympic Games in this dataset, the US has been in the top place 17 times.

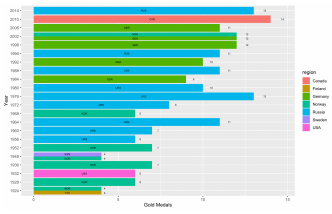

Fig. 6

In the Winter Olympic Games it seems that Russia does pretty well coming in top place 9 out of 22 times and Norway 7 out of 22 times. In the Winter games the US came in top place only once which is a significant change from their domination of the summer games.

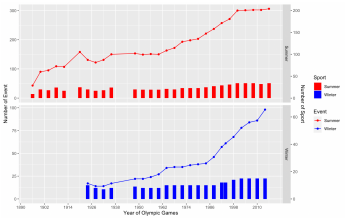

Fig. 7

Shows that the number of events increased significantly over the years, and that during the years of the World Wars there were no Olympic Games.

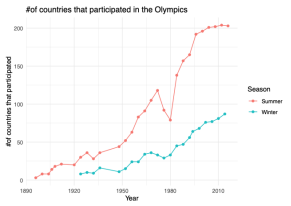

Fig. 8

Over the years, it seems that there has been an increase in the number of countries participating in the Olympic Games in both the summer and winter games, except in the Summer Olympic games between the 1972 and 1980 Olympics there was a sharp decrease in the number of countries participating. In the 1976 Olympic Games we see a drop due to the African Boycott involving 22 countries organized to protest New Zealands torturing of Africans during the Apartheid. The 1980 Olympics was met with a boycott of 65 countries led by the US due to the Soviet Union's invasion of Afghanistan.

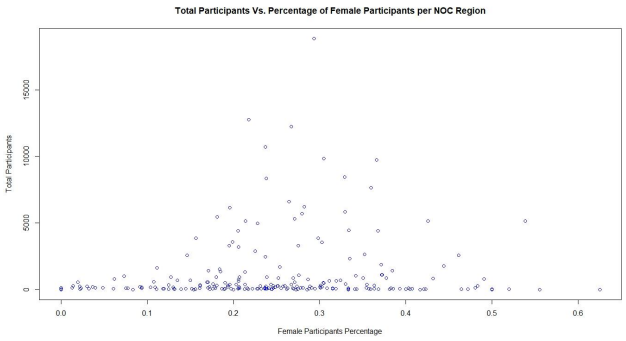

Fig. 9

The scatterplot above shows all 229 countries that participated in the Olympic Games over the last 120 years, since one NOC region did not send any participants. Each dot represents a country, and we can see this looks like a pretty normal distribution.

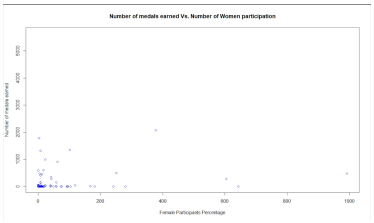

Fig. 10

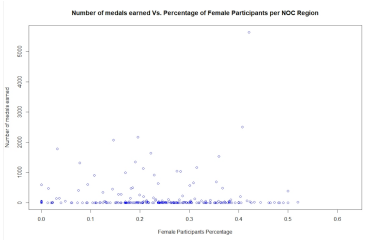

Fig. 11

Figures 10 and 11 we were hoping to look at the correlation between female participation in the Olympic Games and the amount of medals won by each country, but the graphs did not prove to show much of a correlation at all.

```
Call:
lm(formula = Medal ~ Age + Weight + Height + Year, data = athlete_events)

Residuals:
    Min      1Q  Median      3Q     Max
-0.6453 -0.3271 -0.2751 -0.2118  2.9683

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.726e+00  1.701e-01  10.145  < 2e-16 ***
Age          2.191e-03  3.182e-04   6.884 5.83e-12 ***
Weight       2.099e-03  1.986e-04  10.569  < 2e-16 ***
Height       3.358e-03  2.668e-04  12.585  < 2e-16 ***
Year        -1.118e-03  8.474e-05 -13.196  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7704 on 206160 degrees of freedom
Multiple R-squared:  0.007738,  Adjusted R-squared:  0.007719
F-statistic: 401.9 on 4 and 206160 DF,  p-value: < 2.2e-16
```

Fig. 12

```
Analysis of Variance Table

Response: Medal
              Df Sum Sq Mean Sq F value    Pr(>F)
Age            1     94   94.46  159.17 < 2.2e-16 ***
Weight         1    673  672.55 1133.28 < 2.2e-16 ***
Height         1     84   83.77  141.16 < 2.2e-16 ***
Year           1    103  103.34  174.14 < 2.2e-16 ***
Residuals 206160 122347    0.59
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Fig. 13

This regression analysis in Fig. 12 and Anova table in Fig. 13 shows that age, weight, height, and year are all significant predictors for winning medals in the Olympic Medal.

Snowboarding:

Method	Male Test Error	Female Error
LDA	0.08984375	0.1165049
QDA	0.1054688	0.1165049
OLS	0.09599739	0.104643
Lasso	0.06235766	0.07245025
Random Forest	0.3328304	0.3588307
Bootstrap	0.07907519	0.09839499

Fig. 14

Alpine Skiing:

Method	Male Test Error	Female Error
LDA	0.04841313	0.06912442
QDA	0.04142012	0.07834101
OLS	0.04800052	0.07173431
Lasso	0.04701825	0.06645356
Random Forest	0.2415144	0.28887
Bootstrap	0.04511856	0.06487793

Fig. 15

Swimming:

Method	Male Test Error	Female Error
LDA	0.1238132	0.1400189
QDA	0.1304011	0.1412015
OLS	0.1130407	0.1246211
Lasso	0.1090194	0.116051
Random Forest	0.3813993	0.4004171
Bootstrap	0.1140125	0.1223688

Fig. 16

Shooting:

Method	Male Test Error	Female Error
LDA	0.07768595	0.09014423
QDA	0.07290922	0.1009615
OLS	0.06808356	0.08011892
Lasso	0.06727093	0.06884449
Random Forest	0.2937732	0.3221107
Bootstrap	0.06629054	0.08004035

Fig. 17

Figures 14 through 17 all show a table for each sport we looked into deeper with columns for the test error we received for males and the test error we received for females.

Alpine Skiing:

```
Call:
lm(formula = o_train$Medal ~ o_train$Age, data = o_train)

Residuals:
     Min       1Q   Median       3Q      Max
-0.17393 -0.06627 -0.05336 -0.04044  0.97248

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.045686   0.024286  -1.881     0.06 .
o_train$Age  0.004306   0.001024   4.207 2.67e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2275 on 3159 degrees of freedom
Multiple R-squared:  0.00557,   Adjusted R-squared:  0.005255
F-statistic: 17.69 on 1 and 3159 DF,  p-value: 2.665e-05
```

Fig.18

Shooting:

```
Call:
lm(formula = o_train$Medal ~ o_train$Age, data = o_train)

Residuals:
     Min       1Q   Median       3Q      Max
-0.10970 -0.08811 -0.07633 -0.06259  0.96685

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.139143   0.017951   7.751 1.17e-14 ***
o_train$Age -0.001963   0.000546  -3.595 0.000329 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2655 on 3628 degrees of freedom
Multiple R-squared:  0.00355,   Adjusted R-squared:  0.003275
F-statistic: 12.92 on 1 and 3628 DF,  p-value: 0.0003287
```

Fig. 19

In figures 18 and 19 we ran a regression analysis on how Age affects the winning of a medal in Alpine Skiing and Shooting. In both cases our p-value was less than a significance level of 0.05, which means we would reject our null hypothesis that Age has no effect on medal winning.

Kjetil Andr Aamodt:

```
Call:
lm(formula = olympics$Medal ~ olympics$Age, data = olympics)

Residuals:
    Min      1Q  Median      3Q     Max
-0.4627 -0.3955 -0.3508  0.5597  0.6940

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.68657    0.64512   1.064    0.301
olympics$Age -0.01119    0.02480  -0.451    0.657

Residual standard error: 0.5135 on 18 degrees of freedom
Multiple R-squared:  0.01119,   Adjusted R-squared:  -0.04374
F-statistic: 0.2038 on 1 and 18 DF,  p-value: 0.6571
```

Fig. 20

Lars Jrgen Madsen:

```
Call:
lm(formula = olympics$Medal ~ olympics$Age, data = olympics)

Residuals:
    Min      1Q  Median      3Q     Max
-0.2465 -0.2006 -0.1699 -0.1546  0.8301

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.353760   0.417761   0.847    0.405
olympics$Age -0.003830   0.009726  -0.394    0.697

Residual standard error: 0.4089 on 24 degrees of freedom
Multiple R-squared:  0.00642,   Adjusted R-squared:  -0.03498
F-statistic: 0.1551 on 1 and 24 DF,  p-value: 0.6972
```

Fig. 21

Figures 20 and 21 show the regression analysis of two specific athletes Ages and how they affected their medal winnings. Kjetil was an Alpine Skier and Lars was a Shooter, both men competed in multiple Olympic Games in their sports. In both cases we received a p-value that was greater than a significance level of 0.05, which meant we failed to reject our null hypothesis that age had no effect on medal winning.

Methodology:

1. We removed certain columns because we did not find they would be useful to our research question, such as ID and Name.
2. We identified and resolved the null values by omitting them from the data.
3. We analyzed some aspects of the dataset to get more familiar with the data. For example, we looked at how the age and gender of participants has changed over the years, the number of events in the summer and winter games over the years, how many countries participated in the Olympic Games over the years, etc.
4. Next we looked at all the countries and the amount of participants they sent to the Olympic Games and looked at the ratio of male to female participants by country. We put this information into a scatter plot to understand the information better to be able to select the countries that have a higher percentage of women participants.
5. Then we viewed the amount of medals won by each country in the Olympic Games over the last 120 years and the percentage of female athletes that each country sent.
6. Switching gears, we turned the data under Medal and Sex into numerical data to better run regression analyses to see which variables had an effect on medal winnings.
7. We also split the data into male and female olympic athletes to be able to run regression analysis on each separately.
8. We ran a multiple linear regression to see which variables had a significant effect on medal winnings.
9. Next we chose to look further into Shooting and Swimming in the Summer Olympic Games, and Skiing and Snowboarding in the Winter Olympic Games.
10. Splitting these sports into male and female athletes we ran LDA, QDA, Lasso, Cross-Validation, Random Forest, and OLS regressions to examine the test error.
11. Lastly, we selected two athletes that have competed in multiple Olympic Games to see if they performed better, worse, or the same as they got older and had more Olympic experience.

Experiments/Results/Discussion:

When looking into how gender inequality has changed over the 120 years of Olympic Games in this data set, Fig. 1, Fig. 3, and Fig. 4 are a great place to start. These graphs all demonstrate that there is a significant difference in the number of male and female athletes that participate in the Summer and Winter Olympic games. Although there has been a steady increase in the number of female participants, the number of male participants still increases as well which causes there to be a constant gap in gender equality when it comes to the Olympic Games.

We were hoping when viewing the graphs we created for Fig. 10 and Fig. 11 that we would be able to draw a better conclusion on the effect of the percentage of female athletes sent by each country on the amount of medals each country won. Unfortunately, with the vast differences in how many participants are sent by each country this became extremely difficult. Looking at a country that sends 5,000 participants versus a country that sends 500 participants there is a huge difference there and it could be difficult to actually compare the percentages to one another.

After completing a linear regression using Medals as our response variable and Age, Weight, Height, and Year as our explanatory variables, we were able to come to the conclusion that all four of these explanatory variables are significant in predicting the response variable. This means that Age, Weight, Height and Year all play a vital role in predicting the outcome of whether or not an athlete will win a medal.

When looking at the specific sports we decided to exclude Year when running all the regression analyses. We did this because we wanted to look at the biological factors that affect how an athlete performs, and Year is something that no one can control. However, when looking specifically at the two athletes we chose we included Year as this did matter since the athletes were competing in multiple Olympic Games over the course of many years.

We originally tried to also run SVM for each gender and sport we choose, but due to the size of the datasets, SVM was too computationally extensive and our computers barely had enough computing power to run the code. We assume this is because the datasets were large and some of them are highly skewed, resulting in 30,000+ support vectors, making this not a good option for our project.

When looking at Snowboarding we discovered that Lasso regression proved to be more accurate in predicting medal winning for both Male and Female athletes. Snowboarding was added to the Olympics in 1998, making it a very new sport. This caused the dataset to be much smaller than the other sports.

When looking at Alpine Skiing we found QDA to be the most accurate in predicting medal winnings for male athletes and Cross-Validation to be most accurate in predicting winnings for female athletes. These were our overall best performing models. Alpine Skiing has many other factors that we could not include in our model, such as weather conditions and location of competition, which may be a strong determination of an individual's performance.We also could not take into account the advancements of skiing technology since 1936, when Alpine Skiing was introduced into the Olympics.

Swimming we also found Lasso to be the most accurate in predicting the medal winnings for both genders. Swimming produced our highest test errors. We believe this to be from the ranges of different events in the olympics, and how much the sport has progressed since 1896. For Shooting we found that Lasso is the most accurate for both sexes as well. Shooting has 9 different events, ranging from type of gun to different disciplines of the sport, we believe this is where some of the error comes from, along with the technology advancements of guns since 1896.

We chose two athletes to further investigate how the Age of a participant can affect their medal winning in their sport. Our athletes competed in Alpine Skiing and Shooting, so we first ran a regression on whether Age was a significant predictor for Medal winning for the entire sport. In both cases we received a p-value less than the significance level 0.05, which meant we reject our null hypothesis that age had no effect on medal winning.

We also ran a regression analysis to examine whether Age had an effect on the Medal winning of each individual athlete. We did this for an Alpine Skier and a Shooter so we could draw conclusions based on these regressions and the ones we did on the sports as a whole. After completing the individual regressions, we saw p-values in both cases that were greater than a significance level of 0.05, which meant we failed to reject our null hypothesis that age has no effect on medal winnings.

Conclusion:

For this project we found that the data set we chose, although extremely interesting, was actually a bit difficult to work with. Originally we planned to look into gender inequality and how the percentage of female participants for countries in the Olympic Games affects how many medals the country wins. We were able to conclude that there has continually been extensive inequality in the number of female participants to male participants. Even with the constant increase in female participation in Olympic Games, there has also been a constant increase in male participation as well, continuing to widen the gap.

We ran into some issues when it came time to look at the question of whether the percentage of female participation in the Olympic Games actually affected how many medals a country might win. We attempted to cluster the countries based on how high their percentage of female participation was into three groups: the top 33%, middle 33%, and bottom 33%. But this did not help as there are countries that send 5,000 athletes to the olympics and countries that send less than 500 as well, and this drastic difference made it impossible to compare these countries on an even scale.

We began to switch gears by running a multiple linear regression including Year, Age, Height, and Weight as our predictor variables and used Medal as our response and concluded that all four predictor variables were significant to our model. Then we decided to take things a step further and see how the biological factors affected medal winnings in four sports and divided them into male athletes and female athletes. We chose to use Age, Height and Weight for this since Year is not a biological factor and was something no athlete could control.

We ran six different methods of regression analysis- LDA, QDA, OLS, Lasso, Random Forest, and Cross-Validation - to look at the test errors each produced for four different sports. We selected two summer sports, Swimming and Shooting, and two winter sports, Alpine Skiing and Snowboarding. Relatively, the test errors were pretty minimal meaning most methods of regression were pretty accurate in predicting how the biological factors of Age, Weight, and Height affected the winning of a medal. We did notice, however, that Random Forest produced test errors that were much higher than the rest of the regression analyses that we ran, we believe that this could be due to the fact that Age, Weight, and Height are all so strongly correlated to the winning of a medal that it was not possible to de-correlate them enough for Random Forest to be accurate. We were able to conclude that Age, Weight, and Height were all strong predictors of medal winning.

We also were interested in how much age affected medal winnings for individual athletes. Many Olympians compete in multiple Olympic Games in their athletic careers and we were curious if they performed better, worse, or the same over time. In order to determine this we chose two athletes, one Alpine Skier and one Shooter.

We first did an OLS regression on Alpine Skiing and Shooting as a whole using Age as the predictor variable and Medal as the response. After conducting a hypothesis

test on each we were able to conclude that age was a strong predictor of medal winning in both sports. We determined that this definitely did make sense for the sports as a whole. Typically, one would associate someone that is much younger than their competitors with having more strength and more stamina to be able to handle the competitive nature of the Olympic Games. On the other end one might associate being much older with having more experience and perfecting their skill. Someone who is older also could have faced the pressures of the competition which could make them a bit more confident in their abilities allowing them to perform better and not succumb to the nerves.

Lastly, we completed an OLS regression analysis for each athlete we chose using Age as the predictor variable and Medal as the response. After conducting a hypothesis test on each we were able to conclude that age was not a strong predictor of medal winning in both sports. We believe that this does make sense for an individual athlete. Although one could make the same assumptions as above about being younger or older, we determined that this argument really holds true when it comes to sports as a whole. When thinking about a certain individual there is just so much more that goes into competing in the Olympics than just their age. The athlete could be facing a difficult time in their personal lives that affects their performance, or maybe the athlete could have gotten a bad night of sleep before the competition. Things like these could definitely affect the way an athlete competes, and would then affect their chances of winning a medal. We thought of Simone Biles as the perfect recent example for this. She is referenced as the best gymnast of all time and everyone always expects her to be in the Olympic Games and there is no question as to whether she will win the gold. However, in the 2020 Summer Olympic games, Simone performed pretty poorly compared to her usual self, and even ended up deciding not to compete in some of her events all due to a personal circumstance. If we were to examine Simone Biles performance using an OLS regression with Age as the predictor this past Olympic Games, when she was older than previous years, would negatively impact the way age affects medal winning, but from all the news surrounding the situation we know that this is not because of her age.

In conclusion we were able to find that there is significant gender inequality in the Olympic Games spanning all 120 years that we looked at. We were also able to determine that Age, Weight, and Height are all important biological factors in predicting the outcome of winning a medal. And finally, that Age does have an impact on sports overall when it comes to medal winners, but not for individual athletes that compete in multiple Olympic Games.

References:

Berkowitz, B., & Galocha, A. (2021, July 31). Olympians are probably older - and younger - than you think. The Washington Post. Retrieved December 9, 2021, from https://www.washingtonpost.com/sports/olympics/2021/07/31/oldest-youngest-olympians/.

Haas, J., & Herrera, M. (n.d.). (rep.). Olympic History: Athletes and Results Data Analysis (pp. 1–6).

Ioc. (2021, June 3). Montreal 1976 summer Olympics - athletes, medals & results. Olympics.com. Retrieved December 10, 2021, from https://olympics.com/en/olympic-games/montreal-1976.

Ramsay, G., Sinnott, J., & Wright, R. (2021, July 29). 'I have to focus on my mental health,' says Simone Biles after withdrawing from gold medal event. CNN. Retrieved December 9, 2021, from https://www.cnn.com/2021/07/27/sport/simone-biles-tokyo-2020-olympics/index.html.

Rgriffin. "120 Years of Olympic History: Athletes and Results." *Kaggle*, 15 June 2018, www.kaggle.com/heesoo37/120-years-of-olympic-history-athletes -and-results.

U.S. Department of State. (n.d.). U.S. Department of State. Retrieved December 10, 2021, from https://2001-2009.state.gov/r/pa/ho/time/qfp/104481.htm.

VanderPlas, Jake. "Visualization with Seaborn." *Visualization with Seaborn | Python Data Science Handbook*,jakevdp.github.io/PythonDataScienceHandbook/04.14-visualizati on-with-seaborn.html.