Data Set Title
Exploratory Analysis

Jaden Wilkins, jwilkins@bellarmine.edu
Alexander Kyrkjeeide, akyrkjeeide@bellarmine.edu

I. INTRODUCTION

We chose this dataset to explore the age, height, and weight of Olympic athletes. Our goal is to uncover patterns and relationships in the data, helping us understand the typical characteristics of these elite competitors. By examining histograms, scatterplots, and using statistical models, we aim to gain insights into how age, height, and weight interact among Olympic athletes. This analysis will provide valuable information about the physical profiles of individuals who excel in the Olympic arena. We found the link on Kaggle:
https://www.kaggle.com/datasets/samruddhim/olympics-althlete-events-analysis

II. DATA SET DESCRIPTION

Table 1: Data Types and Missing Data

Variable Name	*Data Type*	*Missing Data (%)*
ID	Int	0%
Name	Chr	0%
Sex	Chr	0%
Age	Int	0%
Height	Int	16%
Weight	Num	20%
Team	Chr	0%
NOC	Chr	0%
Games	Chr	0%
Year	Int	0%
Season	Chr	0%
City	Chr	0%
Sport	Chr	0%
Event	Chr	0%
Medal	chr	84%

```
> paste('Amount of missing data: ',sum(is.na(top_50)))
[1] "Amount of missing data:  60"
```

```
> str(top_50) #checking the data types
'data.frame':   50 obs. of  15 variables:
 $ ID    : int  1 2 3 4 5 5 5 5 5 5 ...
 $ Name  : chr  "A Dijiang" "A Lamusi" "Gunnar Nielsen Aaby" "Edgar Lindenau Aabye" ...
 $ Sex   : chr  "M" "M" "M" "M" ...
 $ Age   : int  24 23 24 34 21 21 25 25 27 27 ...
 $ Height: int  180 170 NA NA 185 185 185 185 185 185 ...
 $ Weight: num  80 60 NA NA 82 82 82 82 82 82 ...
 $ Team  : chr  "China" "China" "Denmark" "Denmark/Sweden" ...
 $ NOC   : chr  "CHN" "CHN" "DEN" "DEN" ...
 $ Games : chr  "1992 Summer" "2012 Summer" "1920 Summer" "1900 Summer" ...
 $ Year  : int  1992 2012 1920 1900 1988 1988 1992 1992 1994 1994 ...
 $ Season: chr  "Summer" "Summer" "Summer" "Summer" ...
 $ City  : chr  "Barcelona" "London" "Antwerpen" "Paris" ...
 $ Sport : chr  "Basketball" "Judo" "Football" "Tug-Of-War" ...
 $ Event : chr  "Basketball Men's Basketball" "Judo Men's Extra-Lightweight" "Football Men's Football" "Tug
-Of-War Men's Tug-Of-War" ...
 $ Medal : chr  NA NA NA "Gold" ...
```

III. Data Set Summary Statistics, Visualizations & Interesting Finds

a. **Summary**
b. **Histogram**
c. **Boxplots**
d. **Scatterplots**

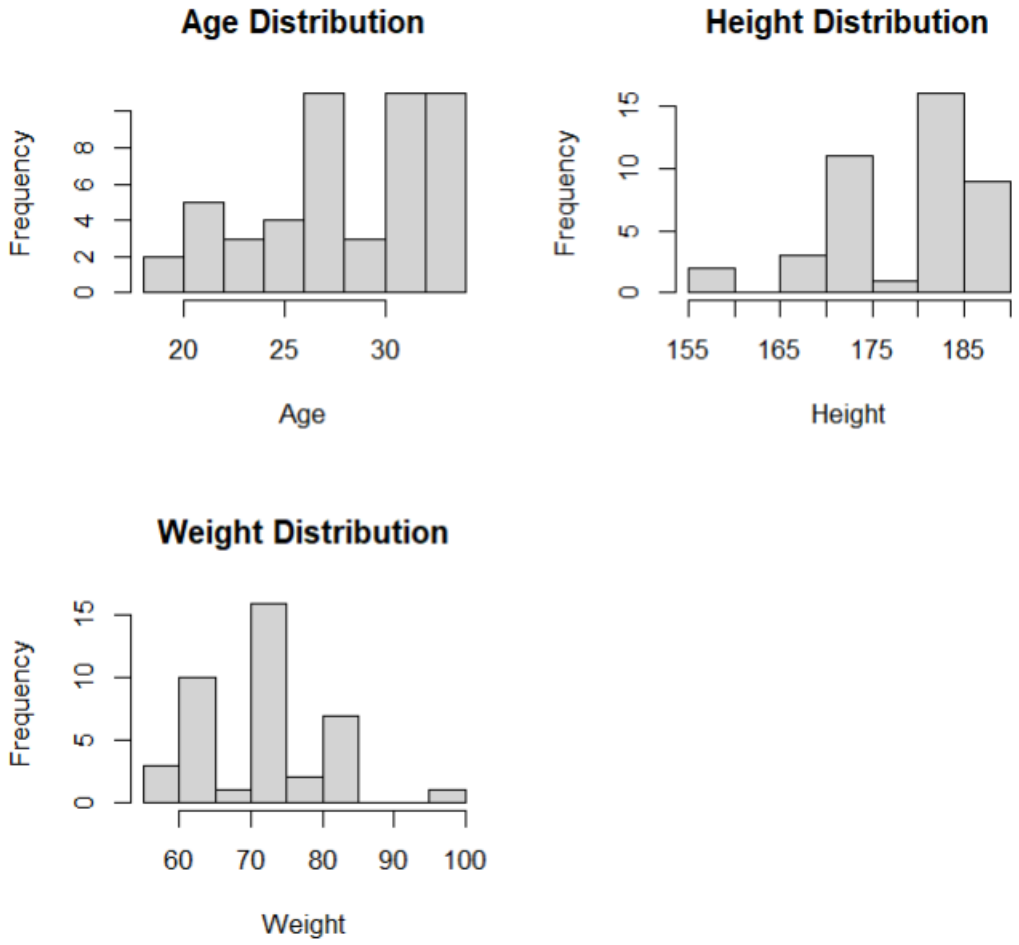
```
> summary(top_50)
       ID            Name               Sex                 Age            Height          Weight
 Min.   : 1.00   Length:50          Length:50          Min.   :18.00   Min.   :159.0   Min.   :55.50
 1st Qu.: 6.00   Class :character   Class :character   1st Qu.:26.00   1st Qu.:175.0   1st Qu.:64.00
 Median : 7.00   Mode  :character   Mode  :character   Median :29.00   Median :183.0   Median :72.00
 Mean   : 9.72                                         Mean   :28.44   Mean   :179.8   Mean   :72.19
 3rd Qu.:15.00                                         3rd Qu.:31.75   3rd Qu.:185.0   3rd Qu.:75.38
 Max.   :17.00                                         Max.   :34.00   Max.   :188.0   Max.   :96.00
                                                                       NA's   :8       NA's   :10
     Team               NOC               Games                Year         Season             City
 Length:50          Length:50          Length:50          Min.   :1900   Length:50          Length:50
 Class :character   Class :character   Class :character   1st Qu.:1948   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Median :1992   Mode  :character   Mode  :character
                                                          Mean   :1972
                                                          3rd Qu.:1994
                                                          Max.   :2014

    Sport              Event              Medal
 Length:50          Length:50          Length:50
 Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character
```

Histogram

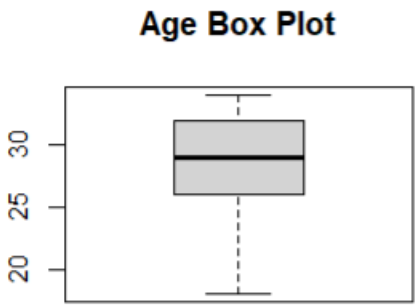
Through our exploration with histograms, we delved into the distribution of age, height, and weight, examining the frequency of occurrences. Our findings unveiled a central tendency in the dataset, indicating that the majority of athletes cluster around the age of 29 years, with an average height of 180 cm and a weight of approximately 70 kg. These key insights provide a succinct snapshot of the prevalent characteristics within the athlete population under scrutiny.

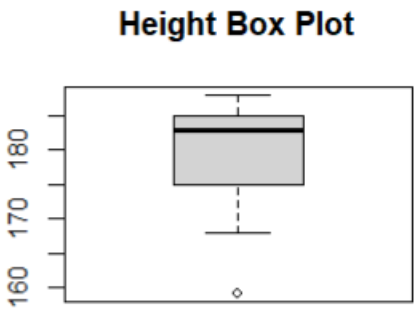
```r
par(mfrow=c(2, 2))
hist(top_50$Age, main = "Age Distribution", xlab = "Age")
hist(top_50$Height, main = "Height Distribution", xlab = "Height")
hist(top_50$Weight, main = "Weight Distribution", xlab = "Weight")
```

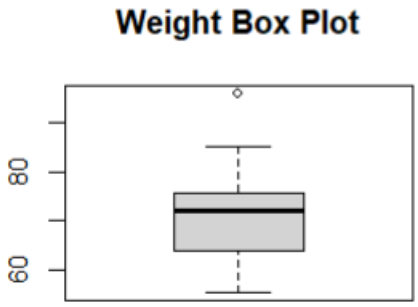


Box Plots

With out box plots, we confirmed what we could only eyeball from the histograms.

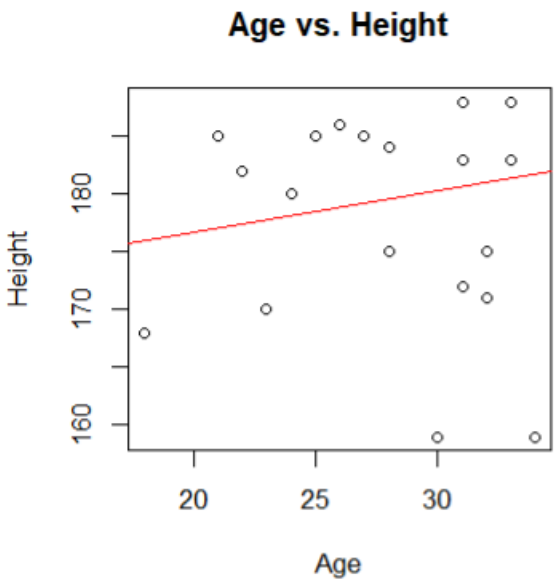


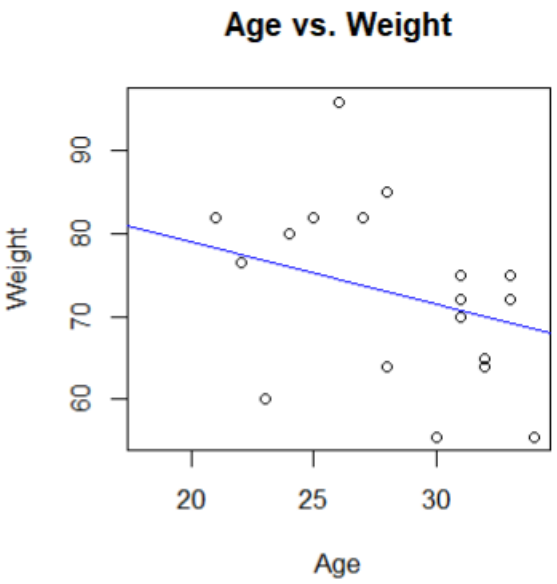


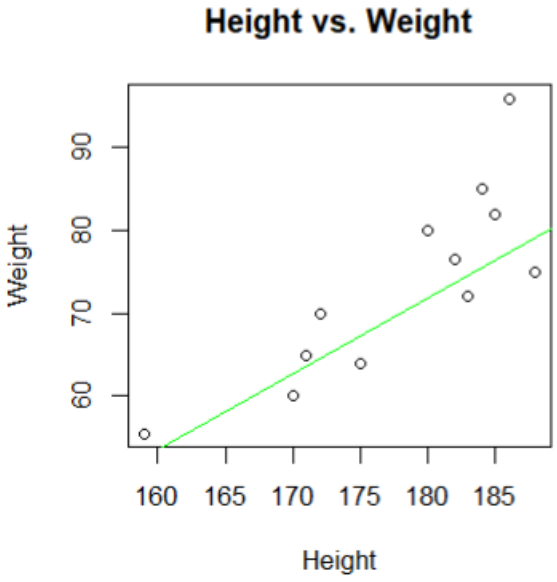


Scatterplots

By examining the scatterplots, we discerned the relationships between age and height, age and weight, as well as height and weight. Clear patterns emerged, revealing a positive linear trend in the association between Age and Height, and Height and Weight. However, a distinct negative linear trend was observed between Age and Weight. These trends, evident in our visual exploration, provide valuable insights into the interplay of these variables, highlighting their directional relationships within the dataset.







Interesting Finds

How many participants have been in each sport.

```
> table(top_50$Sport)

           Athletics            Badminton           Basketball             Biathlon
                   2                    1                    1                    1
Cross Country Skiing             Football             Gymnastics          Ice Hockey
                  17                    1                    9                    2
               Judo              Sailing       Speed Skating             Swimming
                   1                    2                    6                    6
         Tug-Of-War
                   1
```

How many participants were in each Olympic games

```
> table(top_50$Games)

1900 Summer 1912 Summer 1920 Summer 1924 Summer 1932 Summer 1948 Summer 1952 Summer 1980 Winter
          1           2           3           1           2           8           2           1
1988 Winter 1992 Summer 1992 Winter 1994 Winter 1996 Summer 2000 Summer 2002 Winter 2012 Summer
          2           1          10          11           1           2           1           1
2014 Winter
          1
```

How many people of each gender have participated

```
> table(top_50$Sex)

 F  M
11 39
```

How many participants in each EVENT total

```
> table(top_50$Event)

                      Athletics Women's 100 metres                 Athletics Women's 4 x 100 metres Relay
                                                 1                                                      1
                          Badminton Men's Singles                          Basketball Men's Basketball
                                                 1                                                      1
             Biathlon Women's 7.5 kilometres Sprint       Cross Country Skiing Men's 10 kilometres
                                                 1                                                      4
Cross Country Skiing Men's 10/15 kilometres Pursuit       Cross Country Skiing Men's 30 kilometres
                                                 4                                                      3
 Cross Country Skiing Men's 4 x 10 kilometres Relay       Cross Country Skiing Men's 50 kilometres
                                                 4                                                      2
                           Football Men's Football                Gymnastics Men's Floor Exercise
                                                 1                                                      1
                  Gymnastics Men's Horizontal Bar                   Gymnastics Men's Horse Vault
                                                 1                                                      1
             Gymnastics Men's Individual All-Around                Gymnastics Men's Parallel Bars
                                                 2                                                      1
                   Gymnastics Men's Pommelled Horse                         Gymnastics Men's Rings
                                                 1                                                      1
                  Gymnastics Men's Team All-Around                     Ice Hockey Men's Ice Hockey
                                                 1                                                      2
                      Judo Men's Extra-Lightweight                      Sailing Women's Windsurfer
                                                 1                                                      2
                Speed Skating Women's 1,000 metres                Speed Skating Women's 500 metres
                                                 3                                                      3
             Swimming Men's 200 metres Breaststroke       Swimming Men's 400 metres Breaststroke
                                                 3                                                      2
                Swimming Men's 400 metres Freestyle                      Tug-Of-War Men's Tug-Of-War
                                                 1                                                      1
>
```

```
> describe(top_50)
        vars  n    mean    sd median trimmed   mad    min  max range  skew kurtosis   se
ID         1 50    9.72  5.00    7.0    9.65  3.71    1.0   17  16.0  0.29    -1.44 0.71
Name*      2 50   10.32  4.89   10.0   10.53  7.41    1.0   17  16.0 -0.23    -1.47 0.69
Sex*       3 50    1.78  0.42    2.0    1.85  0.00    1.0    2   1.0 -1.31    -0.28 0.06
Age        4 50   28.44  4.32   29.0   28.85  4.45   18.0   34  16.0 -0.70    -0.47 0.61
Height     5 42  179.81  7.70  183.0  180.76  7.41  159.0  188  29.0 -0.97     0.26 1.19
Weight     6 40   72.19  8.50   72.0   72.11 11.86   55.5   96  40.5  0.23     0.01 1.34
Team*      7 50    4.62  1.24    4.0    4.75  1.48    1.0    6   5.0 -0.83     0.84 0.18
NOC*       8 50    3.68  1.10    3.0    3.75  1.48    1.0    5   4.0 -0.27    -0.61 0.16
Games*     9 50    9.14  3.94   11.0    9.25  3.71    1.0   17  16.0 -0.29    -0.90 0.56
Year      10 50 1971.72 31.33 1992.0 1974.85  8.90 1900.0 2014 114.0 -0.72    -0.94 4.43
Season*   11 50    1.52  0.50    2.0    1.52  0.00    1.0    2   1.0 -0.08    -2.03 0.07
City*     12 50    6.92  4.17    8.0    6.75  3.71    1.0   15  14.0 -0.03    -0.97 0.59
Sport*    13 50    7.22  3.18    7.0    7.22  2.97    1.0   13  12.0  0.19    -1.00 0.45
Event*    14 50   14.50  8.05   13.5   14.45 10.38    1.0   28  27.0  0.12    -1.48 1.14
Medal*    15  8    1.50  0.53    1.5    1.50  0.74    1.0    2   1.0  0.00    -2.23 0.19
```

```
> unique(top_50$Height)
 [1] 180 170  NA 185 188 183 168 186 182 172 159 171 184 175
> unique(top_50$NOC)
[1] "CHN" "DEN" "NED" "USA" "FIN"
> unique(top_50$City)
 [1] "Barcelona"      "London"         "Antwerpen"      "Paris"          "Calgary"
 [6] "Albertville"    "Lillehammer"    "Los Angeles"    "Salt Lake City" "Helsinki"
[11] "Lake Placid"    "Sydney"         "Atlanta"        "Stockholm"      "Sochi"
>
```

IV. Regression Analysis

a. **Weight & Age**
b. **Weight & Height**
c. **Age & Height**

```
> model <- lm(Weight ~ Age, data = top_50)
> summary(model)

Call:
lm(formula = Weight ~ Age, data = top_50)

Residuals:
    Min      1Q  Median      3Q     Max
-16.757  -9.011   2.734   4.610  21.490

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  93.9865    10.6692   8.809 1.02e-10 ***
Age          -0.7491     0.3639  -2.058   0.0465 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.167 on 38 degrees of freedom
  (10 observations deleted due to missingness)
Multiple R-squared:  0.1003,    Adjusted R-squared:  0.07663
F-statistic: 4.237 on 1 and 38 DF,  p-value: 0.04647
```

Coefficients:

- **Intercept (β0):** The intercept is 93.9865. This is the estimated value of "Weight" when "Age" is zero. However, the interpretation of the intercept might not be meaningful in this context because having an age of zero is not practically relevant.
- **Age (β1):** The coefficient for "Age" is −0.7491. This represents the estimated change in "Weight" for a one-unit increase in "Age." In this case, it suggests that, on average, the weight decreases by approximately 0.7491 units for each additional year of age.

P-values:

- The p-value associated with the "Age" coefficient is 0.0465. This p-value indicates the probability of observing a t-statistic as extreme as the one computed from the sample, assuming that there is no true effect. A p-value less than the significance level (commonly 0.05) suggests that the predictor variable ("Age") is statistically significant. In this case, the p-value is 0.0465, which is less than 0.05, so we consider the effect of "Age" to be statistically significant.

Residual Standard Error:

- The residual standard error is 8.167. It represents the standard deviation of the residuals, which are the differences between the observed and predicted values of "Weight." A lower residual standard error indicates a better fit of the model to the data.

R-squared and Adjusted R-squared:

- **Multiple R-squared:** The $R2$ value is 0.1003, indicating that approximately 10.03% of the variance in "Weight" is explained by the linear relationship with "Age." This value is relatively low, suggesting that the model explains only a small proportion of the variability in "Weight."
- **Adjusted R-squared:** The adjusted $R2$ adjusts the $R2$ value based on the number of predictors in the model. In this case, the adjusted $R2$ is 0.07663.

F-statistic:

- The F-statistic is 4.237 with 1 and 38 degrees of freedom. It tests the overall significance of the model. The associated p-value is 0.04647, which is less than 0.05. This suggests that the model as a whole is statistically significant.

```
> model_2 <- lm(Weight ~ Height, data = top_50)
> summary(model_2)

Call:
lm(formula = Weight ~ Height, data = top_50)

Residuals:
   Min     1Q Median     3Q    Max
-4.113 -3.267 -2.557  3.508 18.710

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -92.1965    20.4793  -4.502 6.21e-05 ***
Height        0.9112     0.1134   8.033 1.03e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.242 on 38 degrees of freedom
  (10 observations deleted due to missingness)
Multiple R-squared:  0.6294,    Adjusted R-squared:  0.6196
F-statistic: 64.54 on 1 and 38 DF,  p-value: 1.03e-09
```

Coefficients:

- **Intercept ($\beta 0$):** The intercept is −92.1965. This is the estimated value of "Weight" when "Height" is zero. However, the interpretation of the intercept might not be meaningful in this context because having a height of zero is not practically relevant.
- **Height ($\beta 1$):** The coefficient for "Height" is 0.91120. This represents the estimated change in "Weight" for a one-unit increase in "Height." In this case, it suggests that, on average, the weight increases by approximately 0.9112 units for each additional unit of height.

P-values:

- The p-value associated with the "Height" coefficient is 1.03×10^-9, which is extremely small. This indicates that the predictor variable ("Height") is highly statistically significant in predicting the target variable ("Weight").

Residual Standard Error:

- The residual standard error is 5.242. It represents the standard deviation of the residuals, which are the differences between the observed and predicted values of "Weight." A lower residual standard error indicates a better fit of the model to the data.

R-squared and Adjusted R-squared:

- **Multiple R-squared:** The $R2$ value is 0.6294, indicating that approximately 62.94% of the variance in "Weight" is explained by the linear relationship with "Height." This value

is relatively high, suggesting that the model explains a substantial portion of the variability in "Weight."

- **Adjusted R-squared:** The adjusted $R2$ is 0.6196.

F-statistic:

- The F-statistic is 64.54 with 1 and 38 degrees of freedom. The associated p-value is $1.03 \times 10-9$, which is extremely small. This suggests that the model as a whole is highly statistically significant.

```
> model_3<-lm(Age ~ Height, data = top_50)
> summary(model_3)

Call:
lm(formula = Age ~ Height, data = top_50)

Residuals:
    Min      1Q  Median      3Q     Max
-9.2848 -2.1369  0.7444  3.5363  7.6957

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.98200   15.38042   0.584    0.563
Height       0.10895    0.08546   1.275    0.210

Residual standard error: 4.211 on 40 degrees of freedom
  (8 observations deleted due to missingness)
Multiple R-squared:  0.03904,   Adjusted R-squared:  0.01502
F-statistic: 1.625 on 1 and 40 DF,  p-value: 0.2097
```

Coefficients:

- **Intercept ($\beta 0$):** The intercept is 8.982. This is the estimated value of "Age" when "Height" is zero. However, the interpretation of the intercept might not be meaningful in this context because having a height of zero is not practically relevant.
- **Height ($\beta 1$):** The coefficient for "Height" is 0.10895. This represents the estimated change in "Age" for a one-unit increase in "Height." In this case, it suggests that, on average, the age increases by approximately 0.10895 units for each additional unit of height.

P-values:

- The p-value associated with the "Height" coefficient is 0.210. This p-value indicates the probability of observing a t-statistic as extreme as the one computed from the sample,

assuming that there is no true effect. A p-value greater than the significance level (commonly 0.05) suggests that the predictor variable ("Height") is not statistically significant in predicting the target variable ("Age").

Residual Standard Error:

- The residual standard error is 4.211. It represents the standard deviation of the residuals, which are the differences between the observed and predicted values of "Age." A lower residual standard error indicates a better fit of the model to the data.

R-squared and Adjusted R-squared:

- **Multiple R-squared:** The $R2$ value is 0.03904, indicating that approximately 3.90% of the variance in "Age" is explained by the linear relationship with "Height." This value is relatively low, suggesting that the model explains only a small proportion of the variability in "Age."
- **Adjusted R-squared:** The adjusted $R2$ is 0.01502.

F-statistic:

- The F-statistic is 1.625 with 1 and 40 degrees of freedom. The associated p-value is 0.2097, which is greater than 0.05. This suggests that the model as a whole is not statistically significant.

V. SUMMARY OF FINDINGS

In our exploration of a dataset comprising 50 Olympians, we conducted a thorough analysis to uncover patterns and insights. Utilizing summary statistics, we gained a comprehensive understanding of the data, providing a foundation for accurate graph generation. Employing histograms, bar plots, and scatterplots, we visually represented the frequency distributions of age, height, and weight, as well as explored correlations between these variables.

Our scatterplots revealed interesting relationships. We identified a weak positive correlation between Age and Height, a strong positive correlation between Height and Weight, and a weak negative correlation between Age and Weight.

Moving beyond visualization, we applied simple linear regression models to delve deeper into the relationships within the data. The model examining the connection between Age and Weight indicated statistical significance, yet its explanatory power was constrained, as reflected in the

relatively low R-squared value. Notably, both the intercept and the coefficient for Age had p-values below 0.05, highlighting their statistical significance.

On the other hand, the linear regression model featuring Height as the predictor variable demonstrated a highly significant relationship with Weight. This model explained a substantial portion of the variability in Weight, supported by the relatively high R-squared value. The low Residual Standard Error of 5.242 suggested a favorable fit of the model to the data.

In contrast, the linear regression model exploring the relationship between Height and Age did not yield statistically significant results. The model's low R-squared value indicated limited explanatory capability, and the p-value for Height suggested it was not a significant predictor of Age in this context.

In summary, our analytical journey unveiled intriguing insights, showcasing both significant and non-significant relationships among the variables. The application of statistical models deepened our understanding of these relationships, laying the groundwork for informed interpretations and further investigation. Suggestions would be to analyse