Project Draft

Group: Andres Quintana

Researchers at CSU Fullerton's College of Health and Human Development have been conducting a long-running study to understand the factors affecting the physical health of college students. The goal of this project is to come up with an approximation of the Total Fitness Factor Score for a subject using only the available variables of the dataset given by the Department of Kinesiology.

1. Data Exploration and Cleaning

The dataset initially contained approximately 35 variables. Columns with more than 6000 non-empty values were retained to streamline the analysis.

2. Exploratory Data Analysis

In this phase of the project, we aimed to understand the relationship between different variables and the Total Fitness Factor Score (FFTotal). The primary tool used for this exploration was the correlation matrix. Once the correlation matrix was obtained, selected variables exhibited a notable correlation with the Total Fitness Factor Score (FFTotal). This was crucial for building a predictive model, as variables with stronger correlations are more likely to contribute meaningfully to the model.

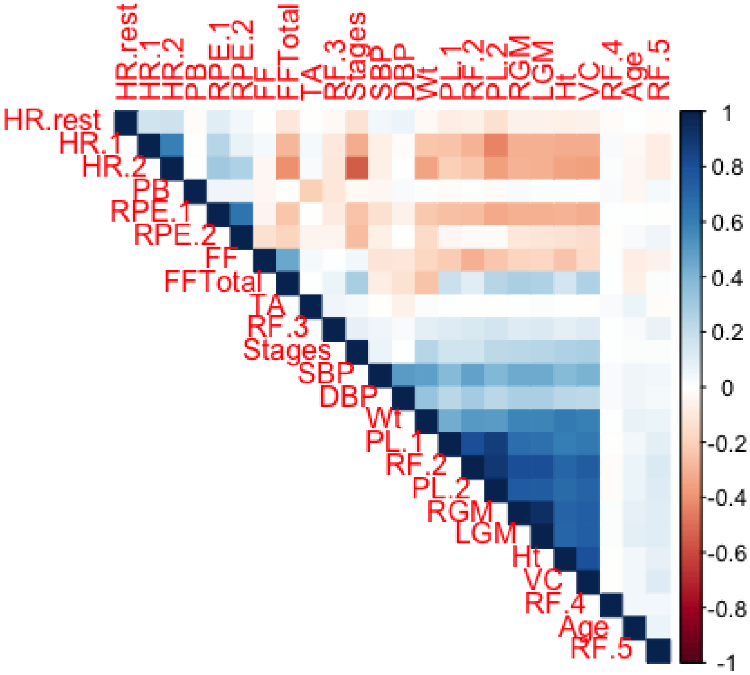


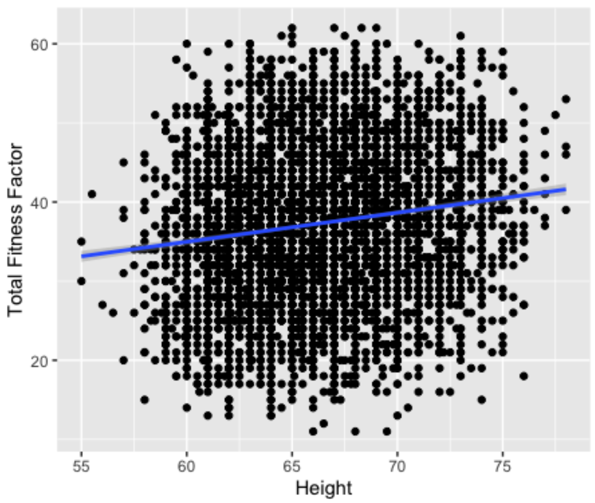
Figure 1: Correlation Matrix

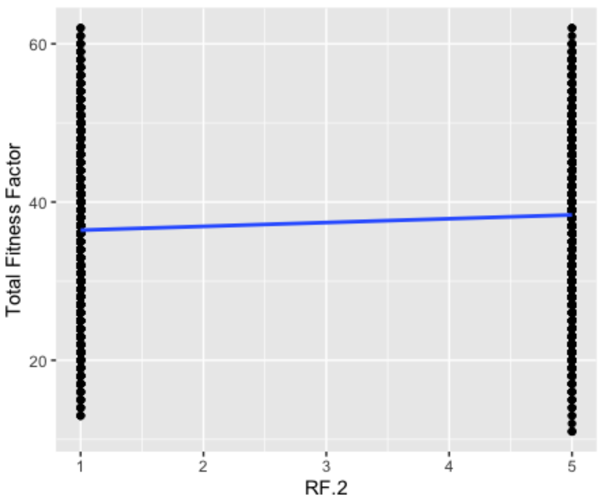
3. Scatter Plots and Linear Regression

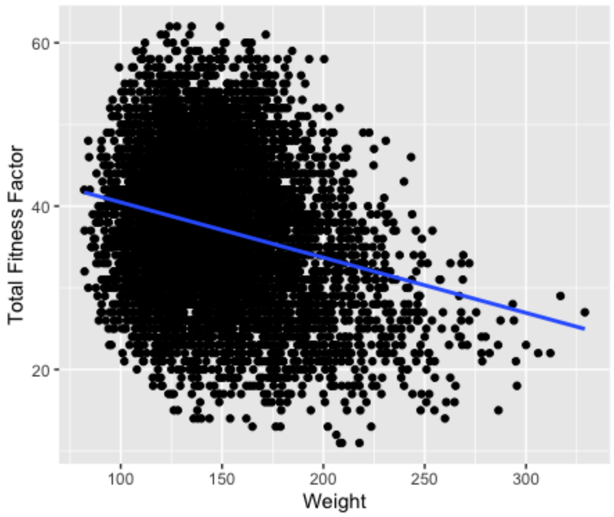
A correlation matrix was computed to explore relationships between variables. The following variables were selected based on their correlation with the Total Fitness Factor Score (FFTotal):

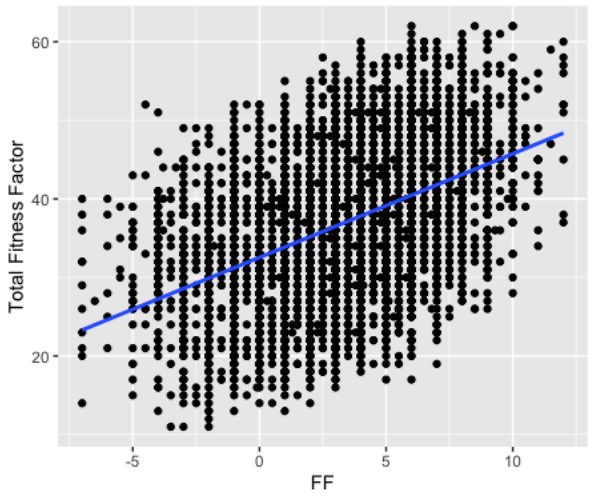
- **Height (Ht):** Taller individuals may have different fitness expectations than shorter individuals. The correlation suggests a meaningful relationship between height and overall fitness.
- **Weight (Wt):** Weight is a crucial factor in assessing overall fitness and health. The correlation indicates a relationship between weight and the Total Fitness Factor Score.
- **Forward Flexion (FF):** A component of the Total Fitness Factor Score (FFTotal), FF represents flexibility, a key indicator of musculoskeletal health. The positive correlation highlights the importance of flexibility in overall fitness.
- **Vital Capacity (VC):** Reflecting respiratory health, VC is closely tied to cardiovascular fitness. The positive correlation suggests a relationship between respiratory health and overall fitness.
- **Resting Heart Rate (HR rest):** A marker of cardiovascular fitness, a lower resting heart rate is associated with better cardiovascular health. The negative correlation indicates the potential impact of cardiovascular health on overall fitness.
- **Right and Left Grip Max (RGM, LGM):** Grip strength provides insights into overall muscular strength and physical capability. The positive correlation emphasizes the role of muscular strength in fitness.
- **Systolic Blood Pressure (SBP):** An important cardiovascular health indicator, SBP's correlation suggests a connection between cardiovascular health and overall fitness.

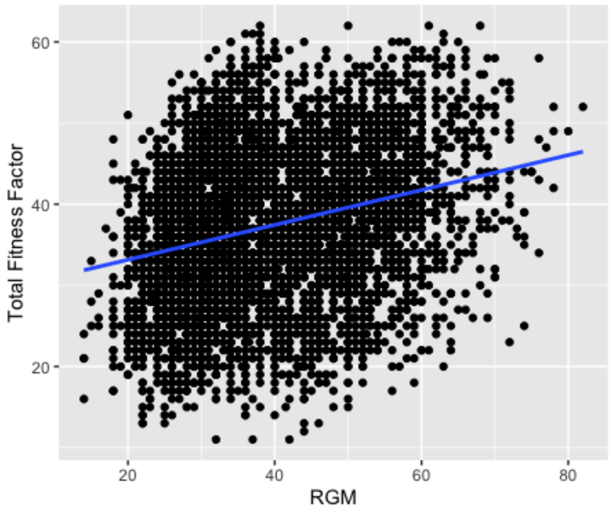
Individual scatter plots with regression lines were created to visualize the relationship between selected variables and FFTotal.

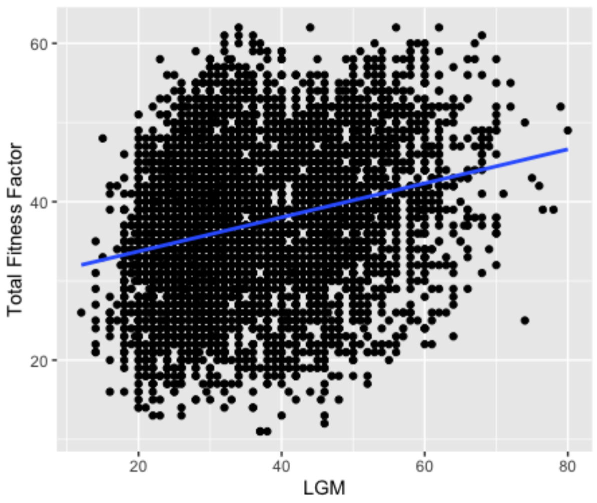

Total Fitness Factor
60
40
20
55
60
65
70
75
Height

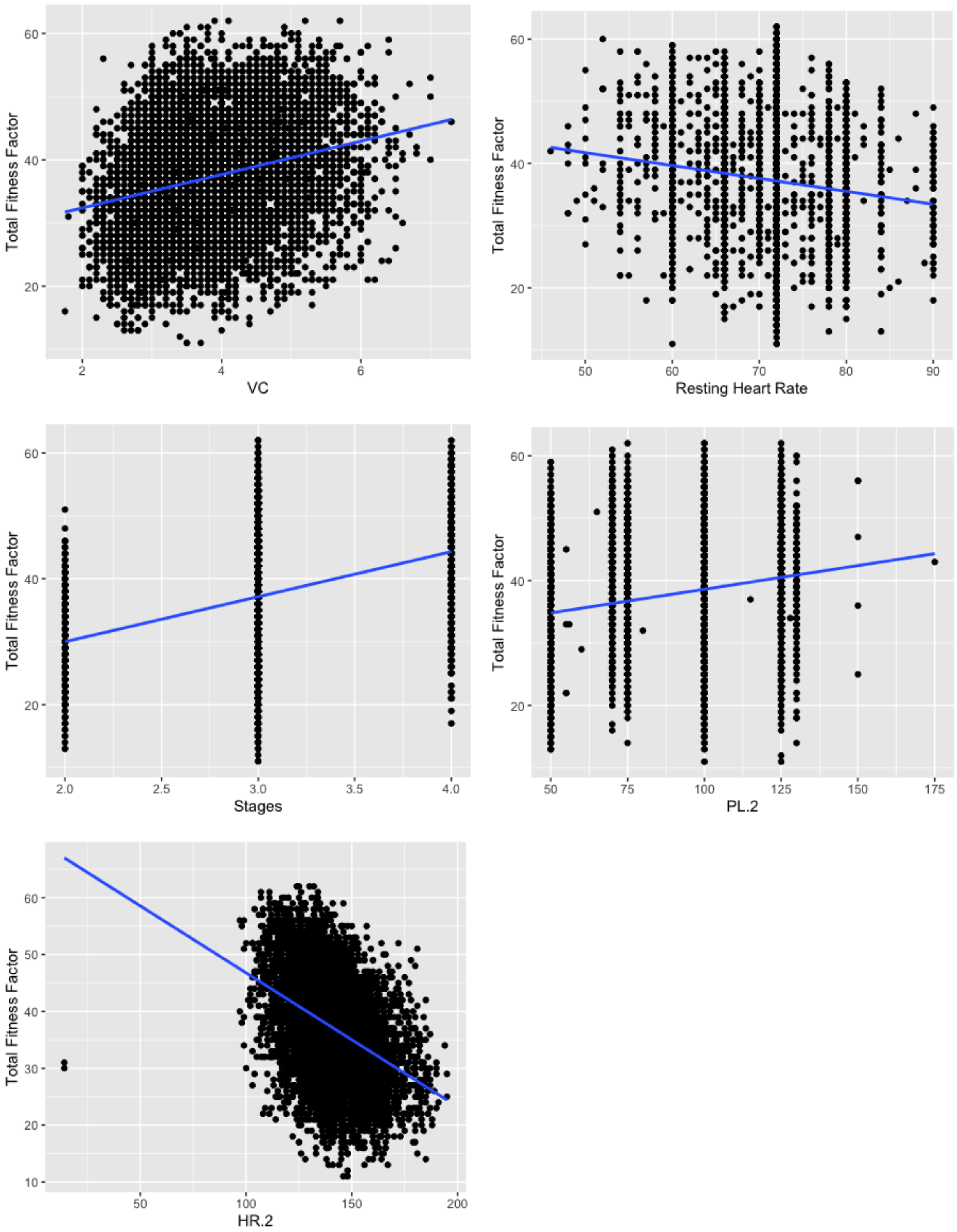

Total Fitness Factor
60
40
20
1
2
3
4
5
RF.2


Total Fitness Factor
60
40
20
100
150
200
250
300
Weight


Total Fitness Factor
60
40
20
-5
0
5
10
FF

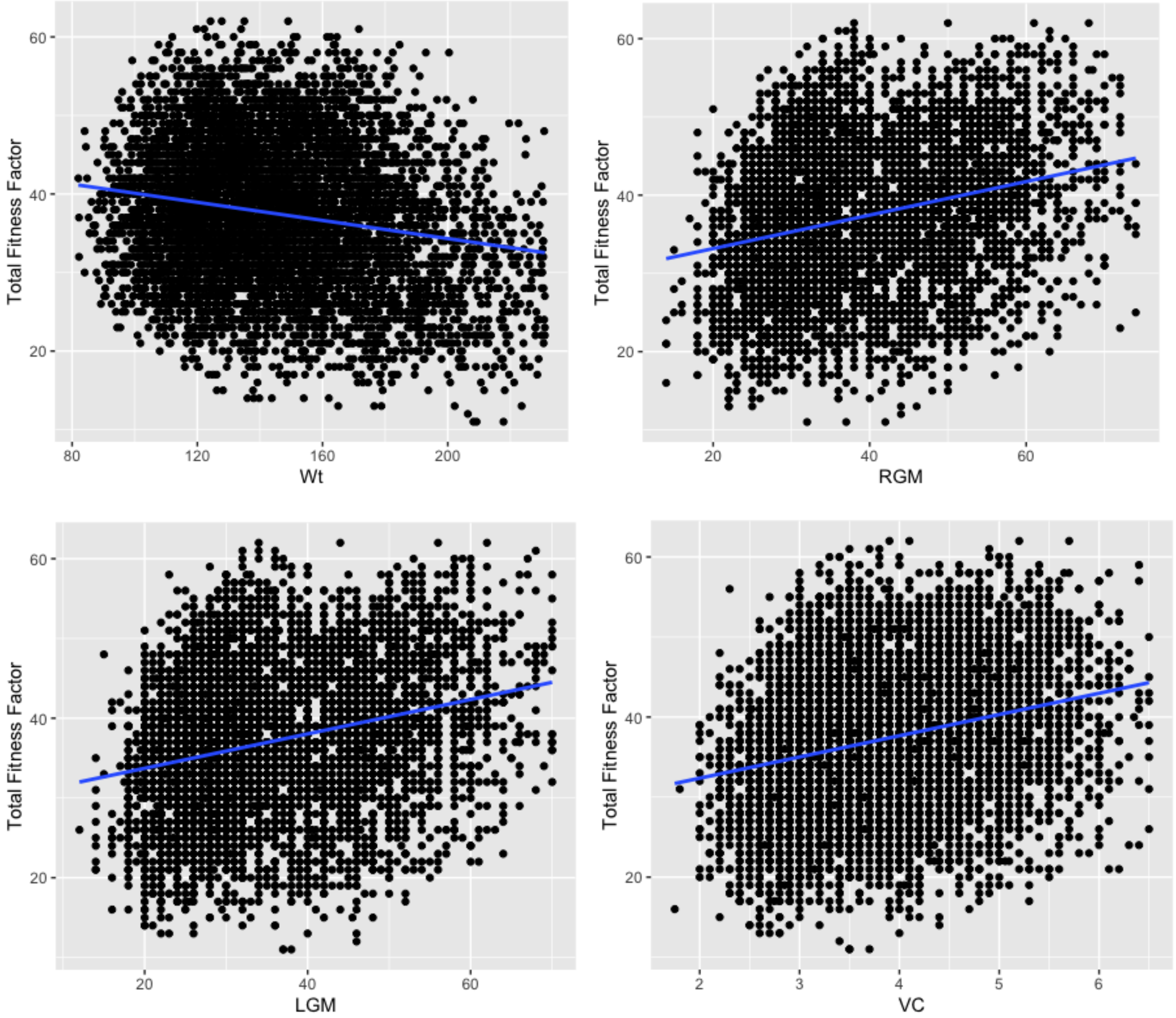

Total Fitness Factor
60
40
20
20
40
60
80
RGM

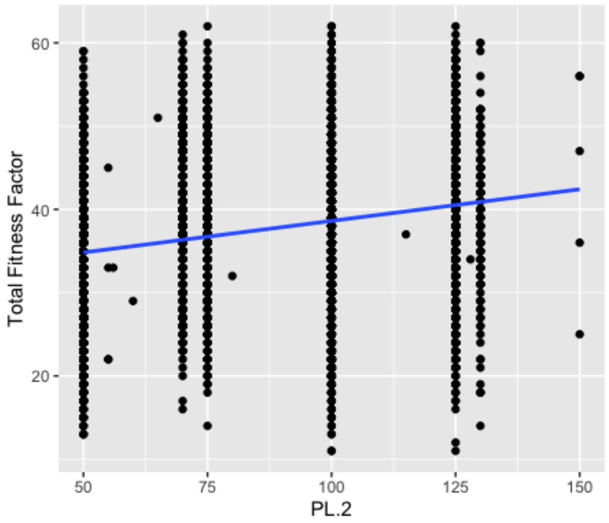

Total Fitness Factor
60
40
20
20
40
60
80
LGM

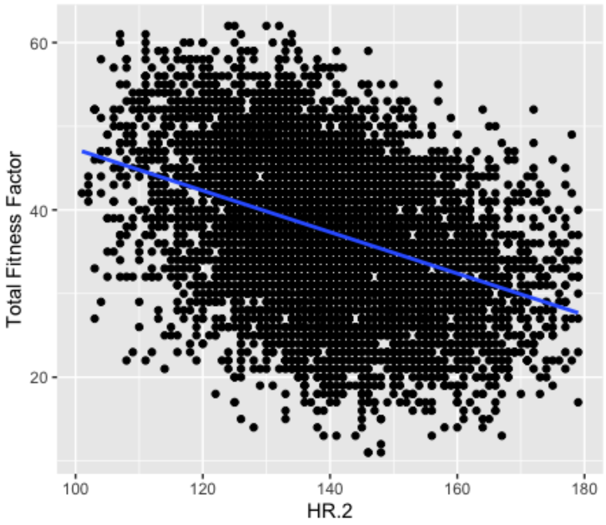

Total Fitness Factor
60
40
20
2
4
6
VC
Total Fitness Factor
60
40
20
50
60
70
80
90
Resting Heart Rate
Total Fitness Factor
60
40
20
2.0
2.5
3.0
3.5
4.0
Stages
Total Fitness Factor
60
40
20
50
75
100
125
150
175
PL.2
Total Fitness Factor
60
50
40
30
20
10
50
100
150
200
HR.2

4. Outlier Removal

Outliers were removed for each variable using the interquartile range (IQR) method.This process is applied to each selected variable, resulting in filtered datasets for each variable. The removal of outliers ensures that the subsequent analysis and model building are more robust and not unduly affected by extreme values.







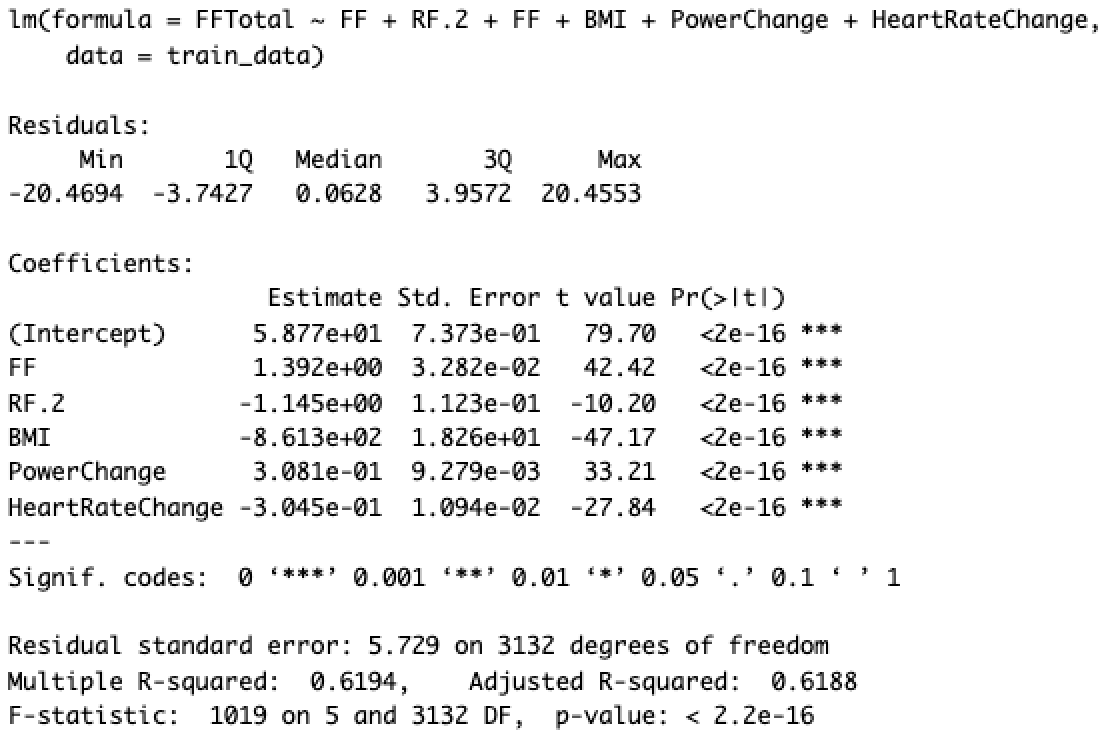
5. Variable Transformations:

- **Body Mass Index (BMI):** Standardized measure of body composition. This simplifies complex relationships between height and weight into a single index. Creating composite scores simplifies the interpretation of multiple variables, providing a standardized measure of overall fitness.
- **Power at Different Stages (PL 1, PL 2):** Indicates sustained effort and adaptation to intensity between stages, revealing how individuals respond to increasing physical demand and offering insights into endurance.
- **Heart Rate Change (HR 1, HR 2):** Provides insights into cardiovascular adaptation between stages reveals how individuals respond to increasing physical demand, offering insights into Heart rate changes and cardiovascular adaptation

6. Modeling:

We developed several linear regression models using transformed variables to predict Total Fitness Factor Scores (FFTotal). The models were trained on the training subset and evaluated on the test subset.

- Model 1:



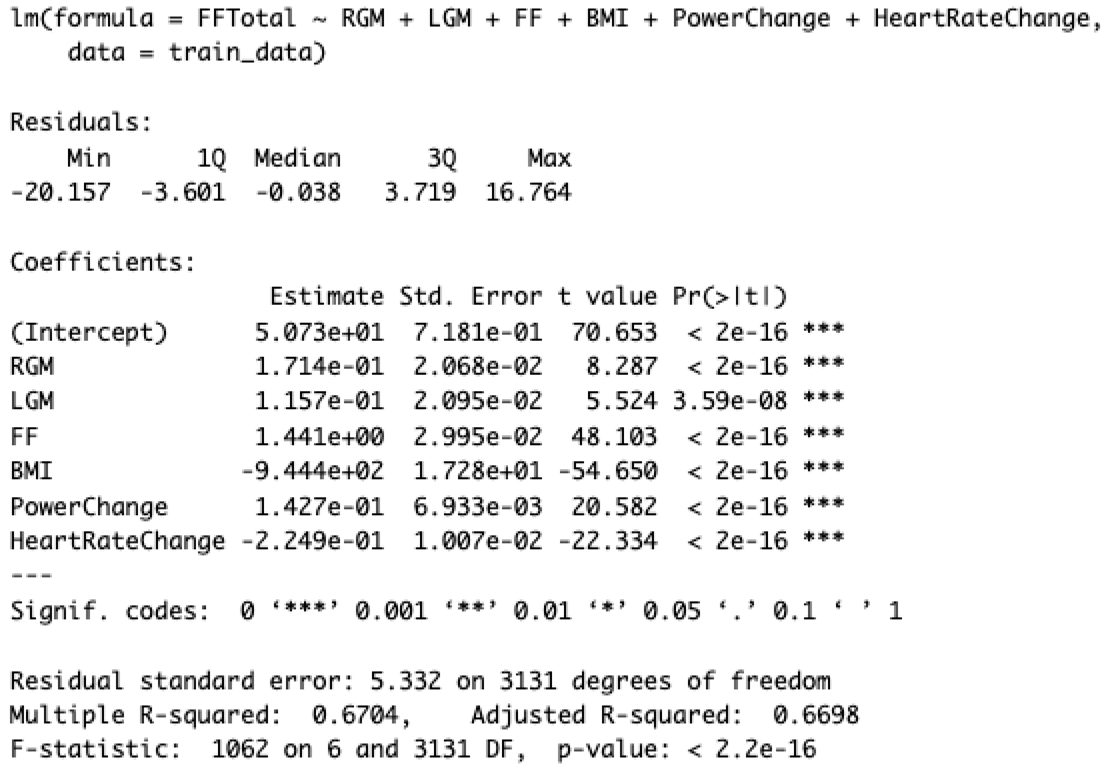
```
lm(formula = FFTotal ~ FF + RF.2 + FF + BMI + PowerChange + HeartRateChange,
    data = train_data)

Residuals:
     Min       1Q   Median       3Q      Max 
-20.4694  -3.7427   0.0628   3.9572  20.4553 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)      5.877e+01  7.373e-01   79.70   <2e-16 ***
FF               1.392e+00  3.282e-02   42.42   <2e-16 ***
RF.2            -1.145e+00  1.123e-01  -10.20   <2e-16 ***
BMI             -8.613e+02  1.826e+01  -47.17   <2e-16 ***
PowerChange      3.081e-01  9.279e-03   33.21   <2e-16 ***
HeartRateChange -3.045e-01  1.094e-02  -27.84   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.729 on 3132 degrees of freedom
Multiple R-squared:  0.6194,    Adjusted R-squared:  0.6188 
F-statistic:  1019 on 5 and 3132 DF,  p-value: < 2.2e-16
```

- Model 2:



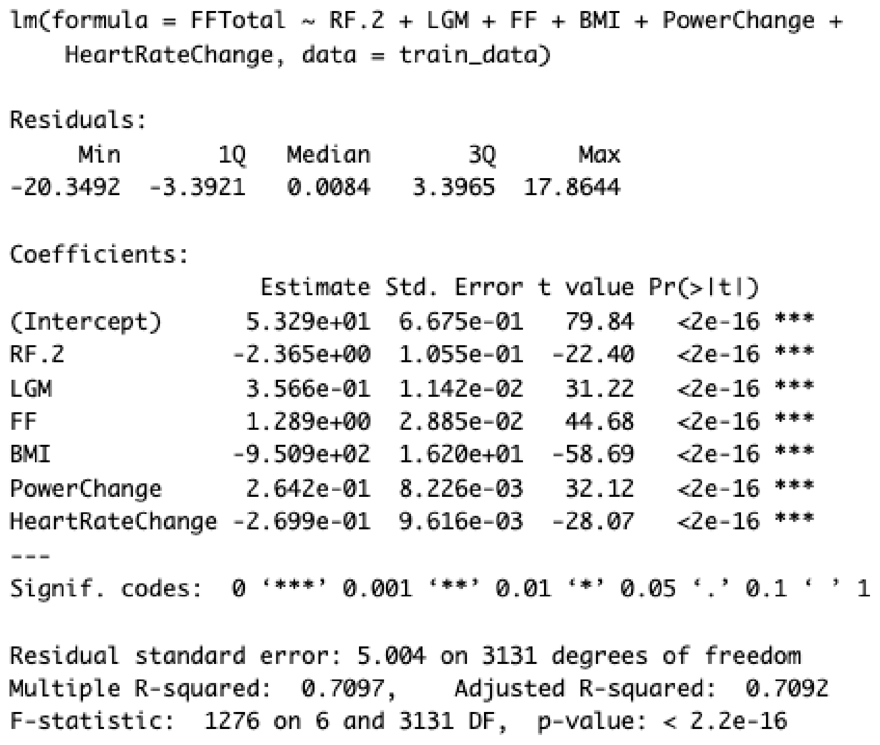
```
lm(formula = FFTotal ~ RGM + LGM + FF + BMI + PowerChange + HeartRateChange,
    data = train_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-20.157  -3.601  -0.038   3.719  16.764 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)      5.073e+01  7.181e-01  70.653  < 2e-16 ***
RGM              1.714e-01  2.068e-02   8.287  < 2e-16 ***
LGM              1.157e-01  2.095e-02   5.524 3.59e-08 ***
FF               1.441e+00  2.995e-02  48.103  < 2e-16 ***
BMI             -9.444e+02  1.728e+01 -54.650  < 2e-16 ***
PowerChange      1.427e-01  6.933e-03  20.582  < 2e-16 ***
HeartRateChange -2.249e-01  1.007e-02 -22.334  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.332 on 3131 degrees of freedom
Multiple R-squared:  0.6704,    Adjusted R-squared:  0.6698 
F-statistic:  1062 on 6 and 3131 DF,  p-value: < 2.2e-16
```

- Model 3:



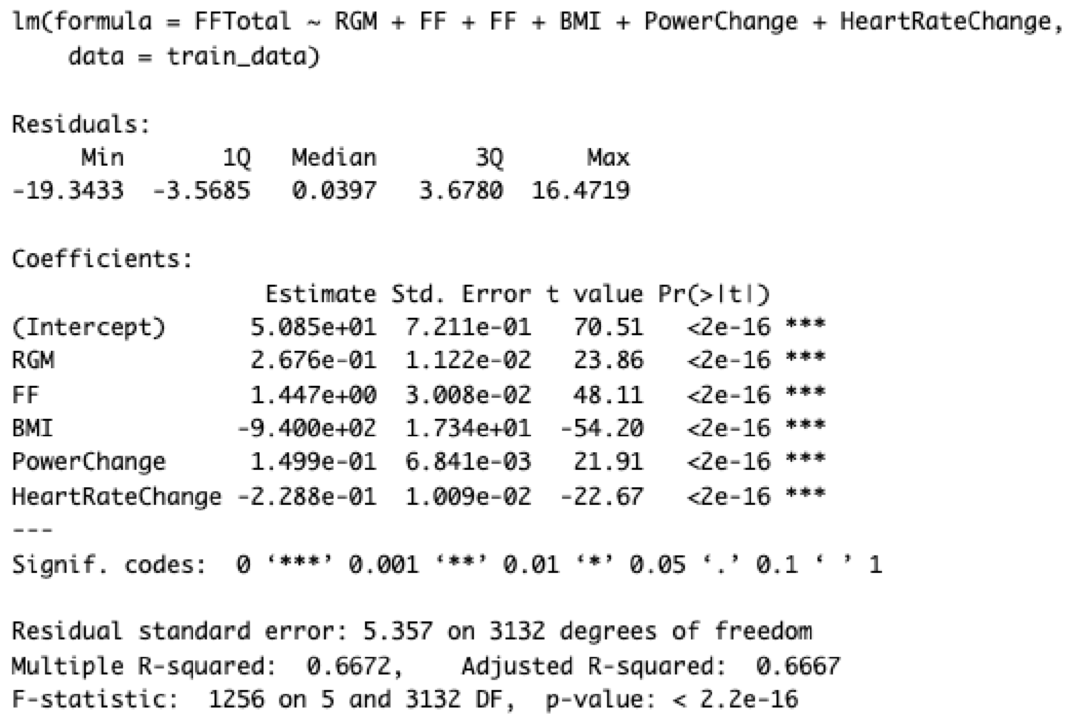
```
lm(formula = FFTotal ~ RF.2 + LGM + FF + BMI + PowerChange +
    HeartRateChange, data = train_data)

Residuals:
     Min       1Q   Median       3Q      Max
-20.3492  -3.3921   0.0084   3.3965  17.8644

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      5.329e+01  6.675e-01   79.84   <2e-16 ***
RF.2            -2.365e+00  1.055e-01  -22.40   <2e-16 ***
LGM              3.566e-01  1.142e-02   31.22   <2e-16 ***
FF               1.289e+00  2.885e-02   44.68   <2e-16 ***
BMI             -9.509e+02  1.620e+01  -58.69   <2e-16 ***
PowerChange      2.642e-01  8.226e-03   32.12   <2e-16 ***
HeartRateChange -2.699e-01  9.616e-03  -28.07   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.004 on 3131 degrees of freedom
Multiple R-squared:  0.7097,	Adjusted R-squared:  0.7092
F-statistic:  1276 on 6 and 3131 DF,  p-value: < 2.2e-16
```

- Model 4:



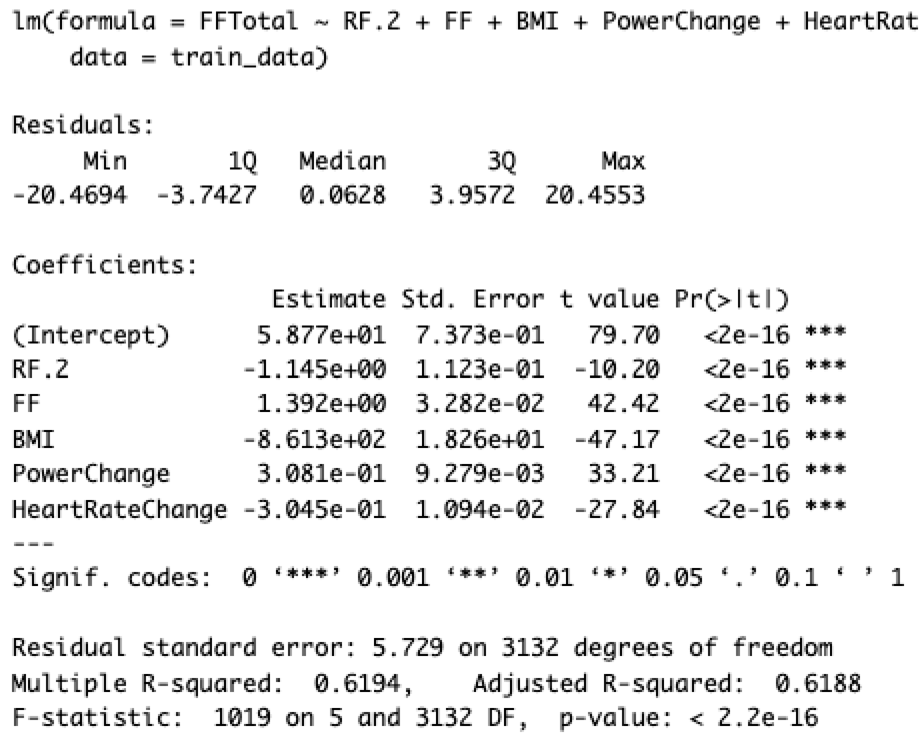
```
lm(formula = FFTotal ~ RGM + FF + FF + BMI + PowerChange + HeartRateChange,
    data = train_data)

Residuals:
     Min       1Q   Median       3Q      Max
-19.3433  -3.5685   0.0397   3.6780  16.4719

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      5.085e+01  7.211e-01   70.51   <2e-16 ***
RGM              2.676e-01  1.122e-02   23.86   <2e-16 ***
FF               1.447e+00  3.008e-02   48.11   <2e-16 ***
BMI             -9.400e+02  1.734e+01  -54.20   <2e-16 ***
PowerChange      1.499e-01  6.841e-03   21.91   <2e-16 ***
HeartRateChange -2.288e-01  1.009e-02  -22.67   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.357 on 3132 degrees of freedom
Multiple R-squared:  0.6672,	Adjusted R-squared:  0.6667
F-statistic:  1256 on 5 and 3132 DF,  p-value: < 2.2e-16
```

- Model 5:



```
lm(formula = FFTotal ~ RF.2 + FF + BMI + PowerChange + HeartRateChange,
    data = train_data)

Residuals:
     Min       1Q   Median       3Q      Max 
-20.4694  -3.7427   0.0628   3.9572  20.4553 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)      5.877e+01  7.373e-01   79.70   <2e-16 ***
RF.2            -1.145e+00  1.123e-01  -10.20   <2e-16 ***
FF               1.392e+00  3.282e-02   42.42   <2e-16 ***
BMI             -8.613e+02  1.826e+01  -47.17   <2e-16 ***
PowerChange      3.081e-01  9.279e-03   33.21   <2e-16 ***
HeartRateChange -3.045e-01  1.094e-02  -27.84   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.729 on 3132 degrees of freedom
Multiple R-squared:  0.6194,	Adjusted R-squared:  0.6188 
F-statistic:  1019 on 5 and 3132 DF,  p-value: < 2.2e-16
```

Model 3 is the best model based on its highest R-squared value (0.7097). To evaluate its performance on the test subset, we calculated the Root Mean Squared Error

RMSE for the Best Model on Test Data: 8.222061 The RMSE provides a measure of the difference between the predicted and actual FFTotal values.

7. Conclusion

In this comprehensive analysis of the dataset, we aimed to understand the factors influencing Total Fitness Factor Scores (FFTotal) through exploratory data analysis, variable transformations, and linear regression modeling. The dataset, encompassing approximately 35 variables related to physical fitness, prompted a focused investigation into five key variables.

The initial exploratory data analysis revealed insights into the relationships between FFTotal and selected variables. The visualization of variables such as Height (Ht), Power at Different Stages (PL.1, PL.2), Resting Heart Rate (HR.rest), and others provided a foundation for variable selection in subsequent modeling.

We developed five distinct linear regression models, each incorporating a combination of predictor variables and transformations. Notably, Model 3 emerged as the most robust, achieving an R-squared value of 0.7097. This model included the variables RF.2, LGM, FF, BMI, PowerChange, and HeartRateChange.

To assess the performance of our models, we split the dataset into training and test subsets. The evaluation of Model 3 on the test subset demonstrated a Root Mean Squared Error (RMSE) of 8.222061. While this value provides a measure of the difference between predicted and actual FFTotal values, it's crucial to consider the context of the dataset and the nature of fitness factors.

The high R-squared value of Model 3 indicates a substantial portion of the variability in FFTotal can be explained by the selected variables and transformations. The negative impact of RF.2, the positive contributions of LGM, FF, and BMI, as well as the effects of PowerChange and HeartRateChange, offer valuable insights into the dynamics of physical fitness.

In conclusion, our analysis underscores the importance of specific variables and transformations in predicting Total Fitness Factor Scores. Further research and validation may refine the models and contribute to a deeper understanding of the complex interplay between various fitness-related factors.