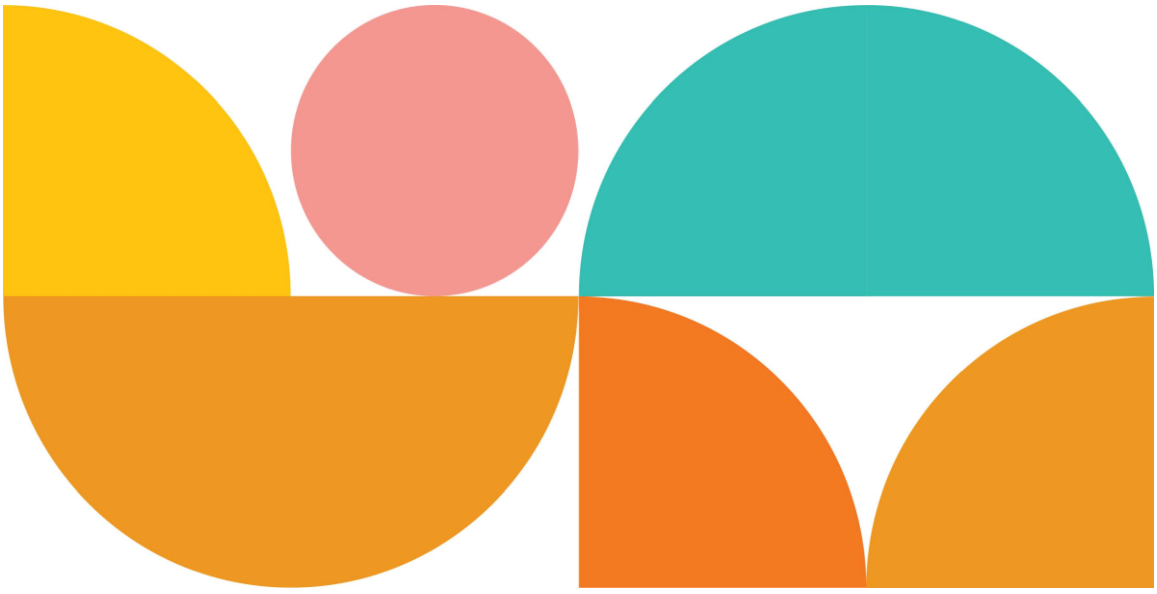
Report

Probability and Statistics

Dania Waseem 23i-2622
Munaza Tariq 23i-2545



Analysis of Factors Influencing Weight Using Multiple Linear Regression

Introduction

This project investigates the relationship between body weight and several other factors such as height, gender, exercise, consumption of sugary foods, meals per day, steps per day, screen time, deficiency / health issues. The dataset obtained was collected through a survey of students of FAST University where demographic and lifestyle variables were taken into consideration. We aim to identify key factors influencing weight using statistical techniques, visualizations and multiple linear regression.

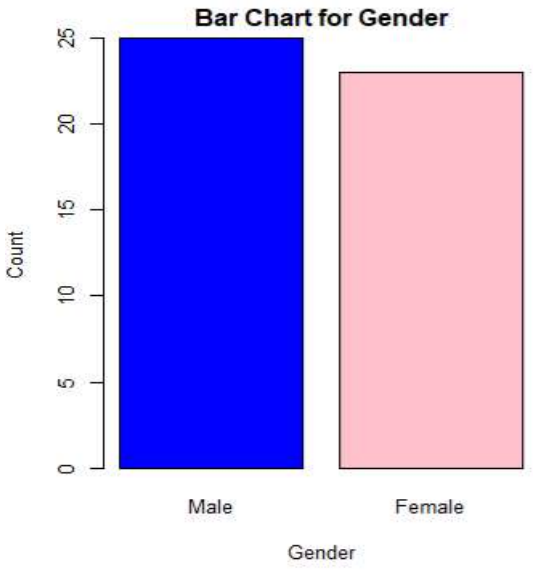
Initial Model

Dependent Variable

- Weight (kg)

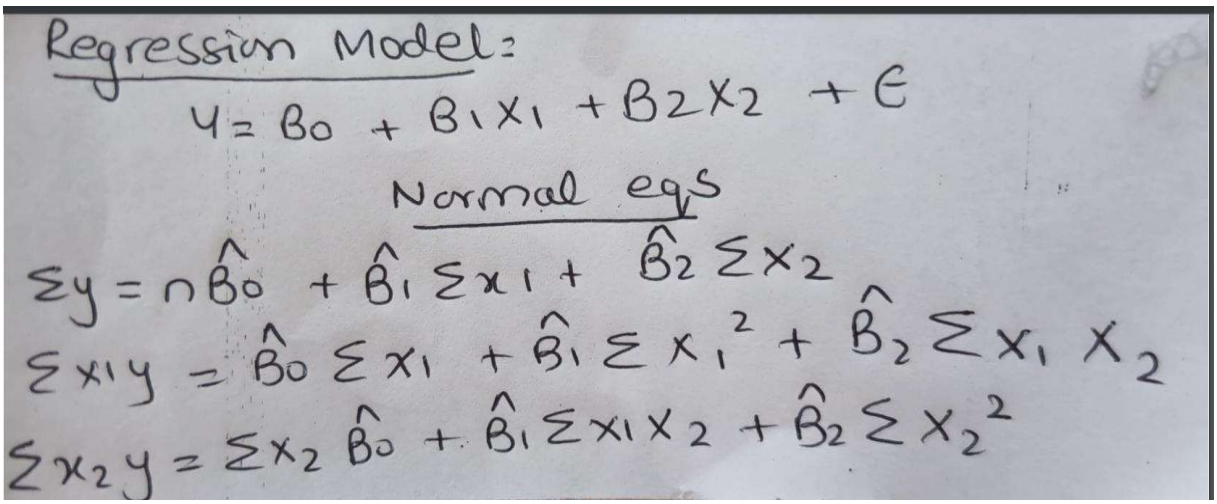
Independent variables

- Height
- Age
- Gender
- Exercise per week
- Deficiency / Health issues
- Consumption of sugar per week
- Meals per Day
- Screen Time
- Steps per Day



We included approximately equal numbers of males and females in this project to minimize potential bias and reduce error. This bar chart shows us the distribution of male and female amongst the 48 students who were surveyed. The graph shows that there were

Regression Model (using the significant variables)

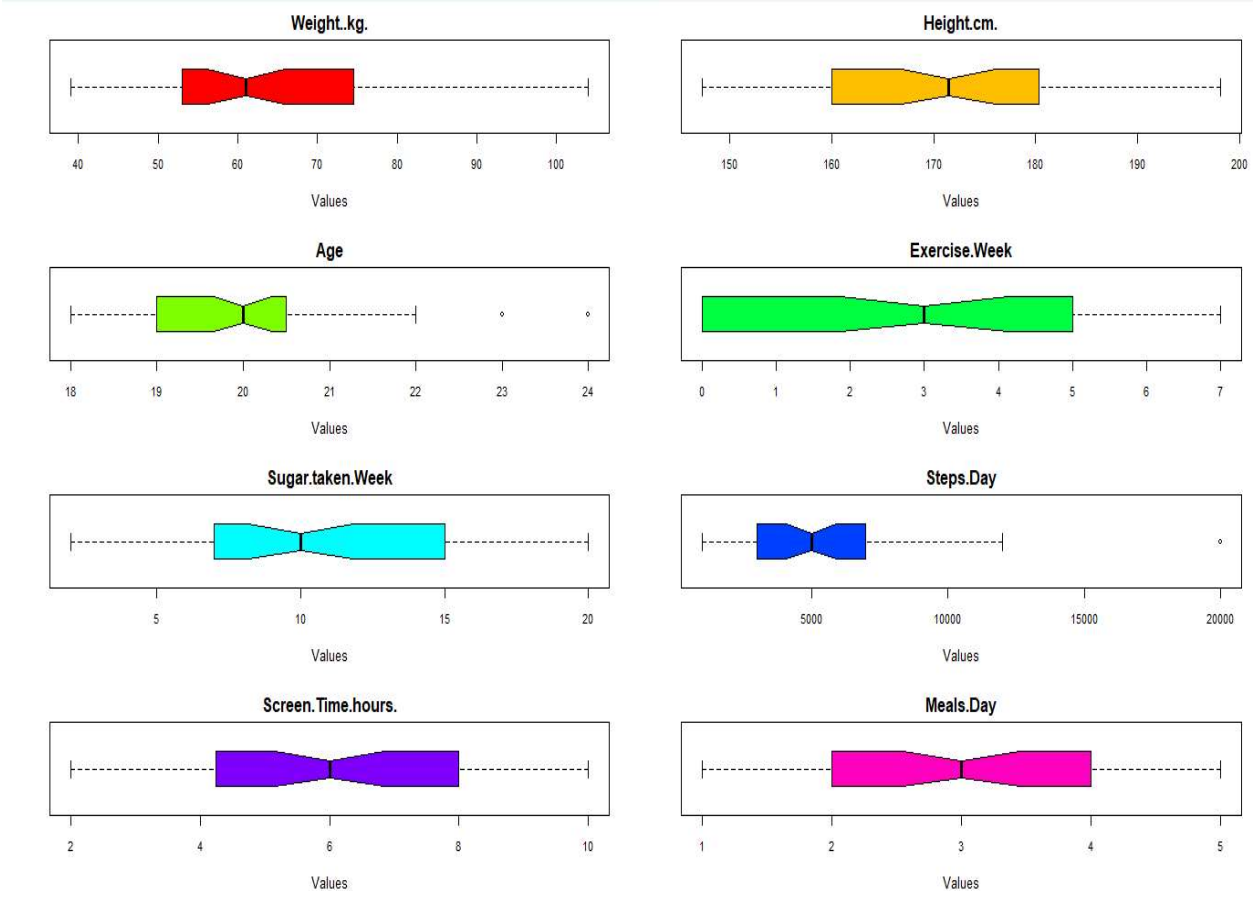


Box And Whisker Plots:

Interpretations:

- Weight: The values range from 39 kg to 104 kg with most values between 54 and 74 kg with median 61 kg.
- Height: The values range from 147 cm to 198 cm, with most values between 160 cm and 180 cm and median 171 cm.
- Age: The values range from 18 to 22 years, with most people around the age of 19 and 20. Median age is 20. The outliers are of age 23 and 24.

•Weekly Exercise: We can see that most exercise 0-5 times in a week however there are people who exercise 6 or 7 times. The median is 3.
•Sugar Consumption: The values range from 2 to 20 sugary meals per week, with most values between 7 and 15 and median 10.
•Steps/Day: The values range from 1000 to 12,000 steps, with most values between 3000 and 7000 steps and median 5000 steps. There is an outlier with 20,000 steps.
•Screen Time: The values range from 2 to 10 hours, with most values between 4 and 8 hours and median 6 hours.
•Meals/Day: The values range from 1 to 5 meals, with most values between 2 and 4 meals, and median 3 meals.

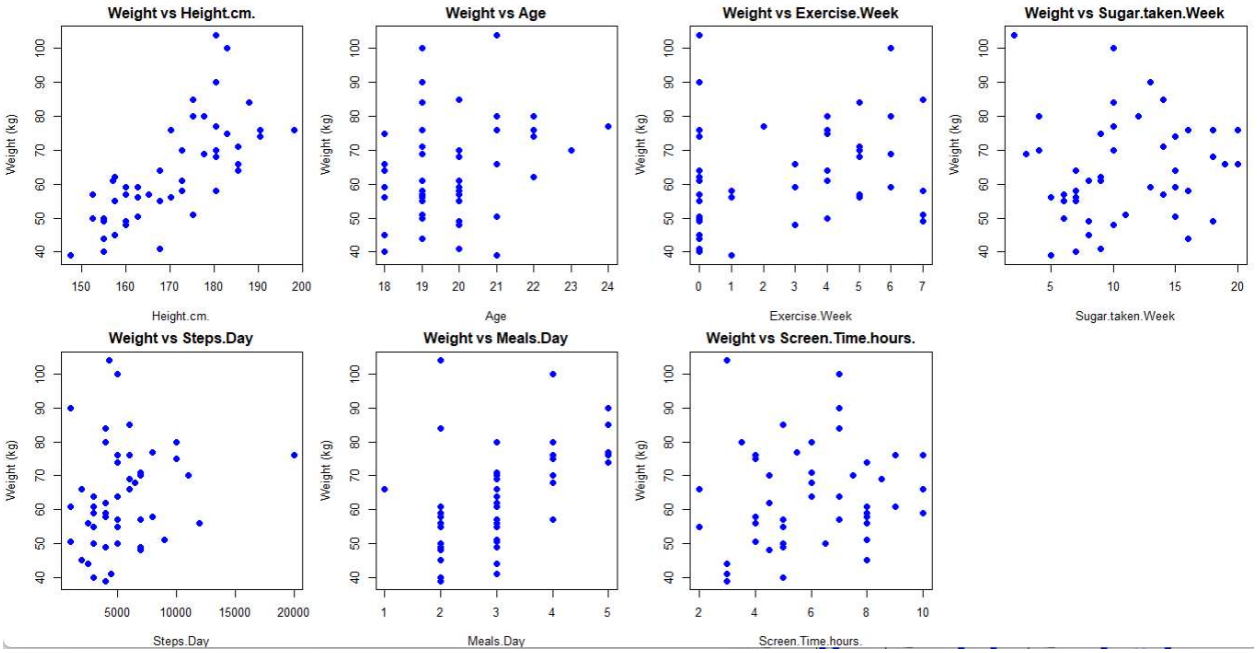


Scatter plots (all variables)

Interpretation:

The scatter plots above show a weak correlation and no consistent trend in variables age, exercise, sugar per week, steps per day, screen time and hours when compared with weight. However, we can observe a positive correlation between

weight and height. As the value of height increases, weight also increases. There is a weak positive correlation between meals per day and sugar taken per week.

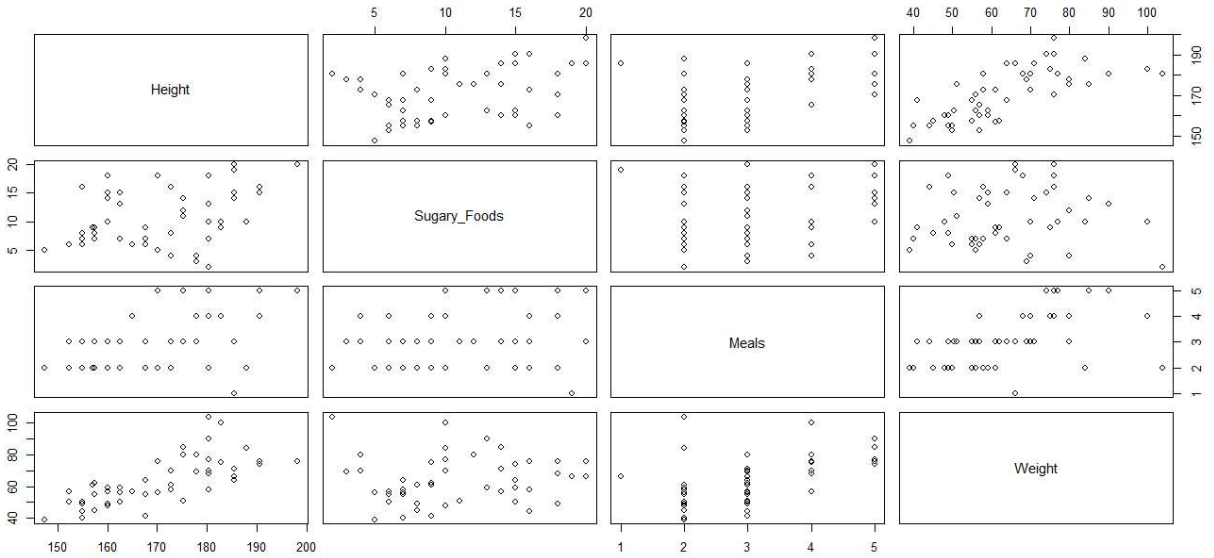


Summary Statistics

The summary statistics show the length, class and mode of string variables(eg. name, gender), and the numerical statistics like mean, median, minimum and maximum values and quartiles of numerical variables (eg. weight, height).

```
----------------------------------Summary Statistics----------------------------------------
     Name               Phone             Weight..kg.     Height.ft.in.        Height.cm.         Age        
 Length:48          Length:48          Min.   : 39.00   Length:48          Min.   :147.3   Min.   :18.00  
 Class :character   Class :character   1st Qu.: 54.00   Class :character   1st Qu.:160.0   1st Qu.:19.00  
 Mode  :character   Mode  :character   Median : 61.00   Mode  :character   Median :171.4   Median :20.00  
                                       Mean   : 63.59                      Mean   :170.7   Mean   :19.79  
                                       3rd Qu.: 74.25                      3rd Qu.:180.3   3rd Qu.:20.25  
                                       Max.   :104.00                      Max.   :198.1   Max.   :24.00  
    Gender          Exercise.Week   Deficiency..Health.Issue Sugar.taken.Week   Steps.Day     Screen.Time.hours.
 Length:48          Min.   :0.000   Length:48                Min.   : 2.00    Min.   : 1000   Min.   : 2.000    
 Class :character   1st Qu.:0.000   Class :character         1st Qu.: 7.00    1st Qu.: 3000   1st Qu.: 4.375    
 Mode  :character   Median :3.000   Mode  :character         Median :10.00    Median : 5000   Median : 6.000    
                    Mean   :2.667                            Mean   :10.62    Mean   : 5465   Mean   : 5.958    
                    3rd Qu.:5.000                            3rd Qu.:15.00    3rd Qu.: 7000   3rd Qu.: 8.000    
                    Max.   :7.000                            Max.   :20.00    Max.   :20000   Max.   :10.000    
   Meals.Day      
 Min.   :1.000  
 1st Qu.:2.000  
 Median :3.000  
 Mean   :3.021  
 3rd Qu.:4.000  
 Max.   :5.000  
```

Scatter plots separately for significant variables with weight



Weight vs. Height

- **Observation**: The scatterplot shows a positive trend, where as Height increases, Weight tends to increase as well.
- **Interpretation**: This strong, positive correlation indicates that Height is a significant predictor of Weight. Taller individuals tend to weigh more, which aligns with the significant p-value for Height in the regression output.

Weight vs. Sugary Foods

- **Observation**: The scatterplot shows a slight **positive trend**, where higher sugary food consumption is associated with higher weight. As individuals consume more sugary foods per week, their weight tends to increase.
- **Interpretation**: This makes sense because increased sugary food intake often contributes to weight gain. The regression output also shows that Sugary Foods is significant, reinforcing its role as a predictor of weight. The relationship is not as strong as that of Height.

. Weight vs. Meals

- **Observation**: The scatterplot shows a weak positive trend between the number of meals and Weight. Individuals who consume more meals per day appear to weigh slightly more.
- **Interpretation**: While the relationship is not as strong as with Height, the positive association supports the significance of Meals in the regression model.

Correlation Matrix:

```
--------------------------------------------------------------------------------
Correlation Matrix:
                Height Sugary_Foods     Meals     Weight
Height       1.0000000   0.36878670 0.4599900 0.71455472
Sugary_Foods 0.3687867   1.00000000 0.2580955 0.09751365
Meals        0.4599900   0.25809549 1.0000000 0.50933539
Weight       0.7145547   0.09751365 0.5093354 1.00000000
Multiple Linear Regression Model Summary:

Call:
lm(formula = Weight ~ Height + Sugary_Foods + Meals, data = LDF)

Residuals:
    Min      1Q  Median      3Q     Max
-21.136  -5.582  -2.069   4.558  30.560
```

Height and Weight: Strong positive correlation (0.7145), indicating height is a significant predictor of weight.

Meals and Weight: Moderate positive correlation (0.5093), suggesting more meals/day has some association with weight.

Sugary Foods and Weight: Weak correlation (0.0975), implying sugary food intake has negligible impact on weight.

Multiple Linear Regression Model

In this model we used 3 significant variables height, meals per day and sugar taken per week.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -78.0339    20.6926  -3.771 0.000481 ***
Height         0.8071     0.1348   5.988 3.52e-07 ***
Sugary_Foods  -0.6688     0.3193  -2.095 0.041989 *
Meals          3.6274     1.5591   2.327 0.024651 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.827 on 44 degrees of freedom
Multiple R-squared:  0.5926,    Adjusted R-squared:  0.5648
F-statistic: 21.33 on 3 and 44 DF,  p-value: 1.107e-08
```

Pr(>|t|) Values (P-values)

- **Definition**: P-values test the null hypothesis that the corresponding coefficient is zero (no effect). A small p-value (< 0.05) indicates strong evidence against the null hypothesis, meaning the variable significantly contributes to the model.

P-values in the output:

- **Height**: *3.52×10^(-7) :* This indicates that Height is a highly significant predictor.
- **Sugary_Foods**: 0.04189 :
  Sugary_Foods is significant at the 5% level (p < 0.05), meaning it has a small but meaningful impact on the dependent variable.
- **Meals**: 0.024651
  - Meals is also significant at the 5% level, showing it contributes meaningfully to explaining the dependent variable.

The **Intercept** (p = 0.000481) is also significant, meaning the baseline value when all predictors are 0 is not negligible.

Coefficient of Determination (R-squared)

- **R-squared**: 0.5926
- **Adjusted R-squared**: 0.5648

Interpretation:

- **R-squared** indicates that approximately 59.26% of the variation in the dependent variable (weight) is explained by the independent variables (Height, Sugary_Foods, Meals) in the model.
- **Adjusted R-squared** accounts for the number of predictors in the model and adjusts for the potential overfitting. At 0.5648, it means about 56.48% of the variance is explained by the predictors, even after penalizing for model complexity.

  The residual standard error is 9.827, indicating a reasonable prediction accuracy. Overall, the model is robust and identifies key predictors influencing the outcome.

Conclusion:

This study investigated the factors influencing weight by examining relationships between height, sugary food intake, meal frequency, and other variables. The analysis revealed that height and meal frequency were significant predictors of weight, as reflected in both the correlation matrix and regression results. Interestingly, sugary food consumption, though weakly correlated, demonstrated a minor influence in the regression model.

The analysis also highlighted certain limitations. Variables like and exercise habits, which are often considered critical, showed limited significance in this dataset. This could indicate the need for a broader dataset or the inclusion of additional variables, such as physical activity intensity, caloric intake, or metabolic factors, for a more comprehensive understanding of weight determinants.

These findings emphasize the importance of balanced dietary habits, particularly consistent meal patterns, in maintaining healthy weight levels

References

- Han, K., & Kim, M. K. (2023). Factors affecting high body weight variability. *Journal of Obesity & Metabolic Syndrome, 32*(2), 163–169. https://doi.org/10.7570/jomes22063
- Guinè, R. P. F., & Fernandes, S. R. (2016). Regression model of the factors that influence weight of young adolescents. *Journal of Food Science Research, 1*(1), 39–48. Retrieved from https://www.tsijournals.com/articles/regression-model-of-the-factors-that-influence-weight-of-young-adolescents.pdf
- ResearchGate. (n.d.). Multiple linear regression of weight gain. Retrieved from https://www.researchgate.net/figure/Multiple-Linear-regression-of-weight-gain_tbl1_343562364