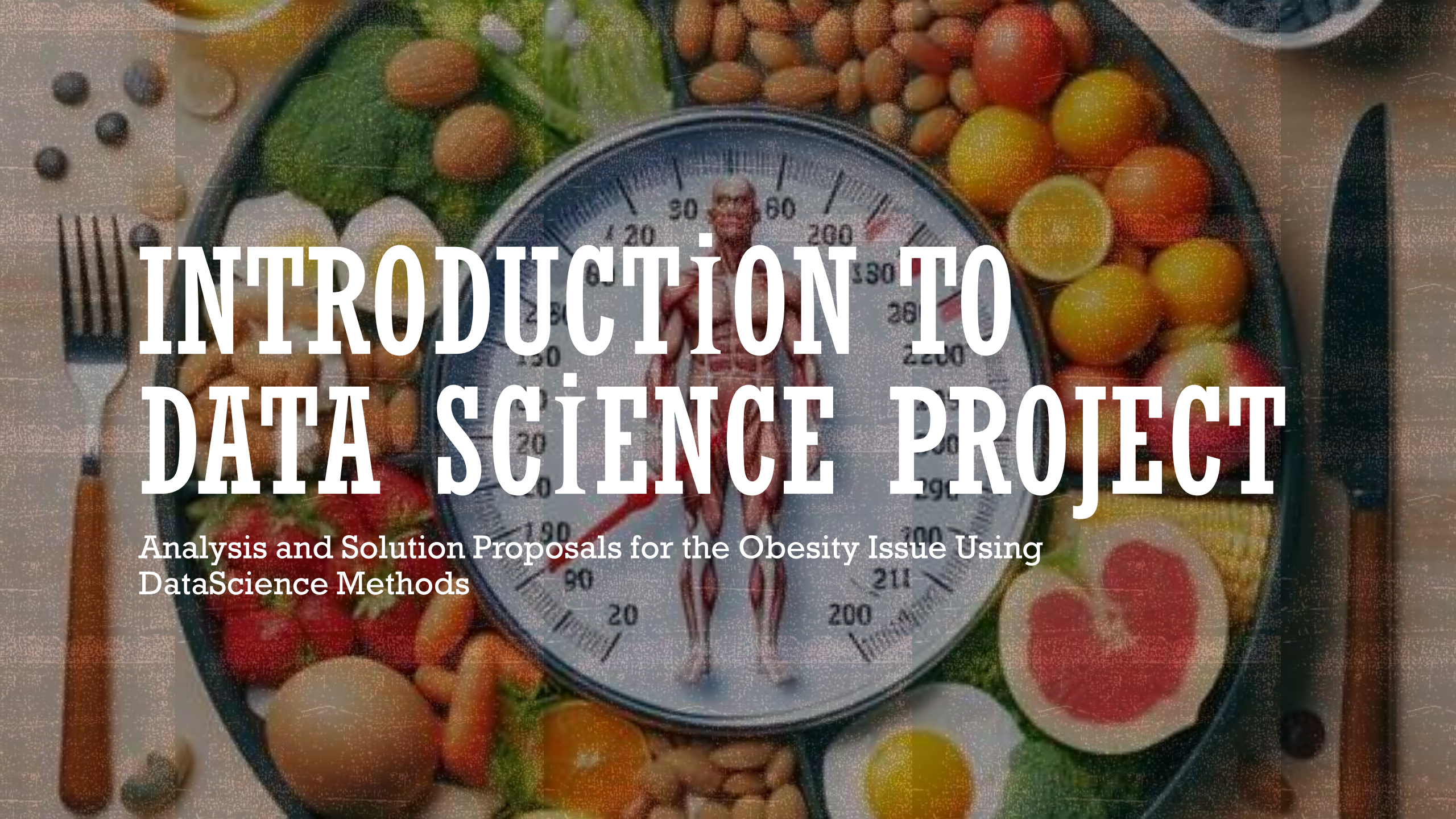


INTRODUCTİON TO DATA SCİENCE PROJECT

Analysis and Solution Proposals for the Obesity Issue Using DataScience Methods

STEPS

1. Problem Definition
2. Data Collection
3. Exploratory Data Analysis and Data Visualization
4. Model
5. Model Testing and Results

GANTT

No	Work Packages	MONTHS							
		1	2	3	4	5	6	7	8
1	Problem Detection								
2	Dataset Collection								
3	EDA and Data Visualizations								
4	Model Creation								
5	Model Testing and Interpretation								

1.PROBLEM DEFINITON

Obesity is a serious health problem linked to heart diseases, diabetes, and chronic illnesses, reducing quality of life. Data science methods help analyze the factors causing obesity and create personalized health solutions.

2.DATA COLLECTION

A suitable dataset has been found on Kaggle for conducting a data science project related to the topic. You can access the dataset through the following link: kaggle.com/datasets/lesumitkumarroy/obesity-data-set

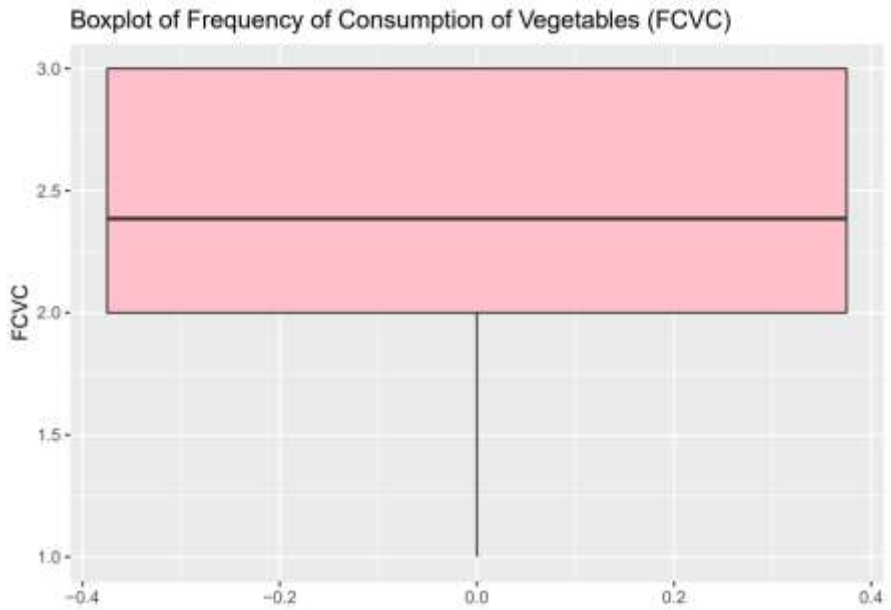
3. EXPLORATORY DATA ANALYSIS AND DATA VISUALIZATION

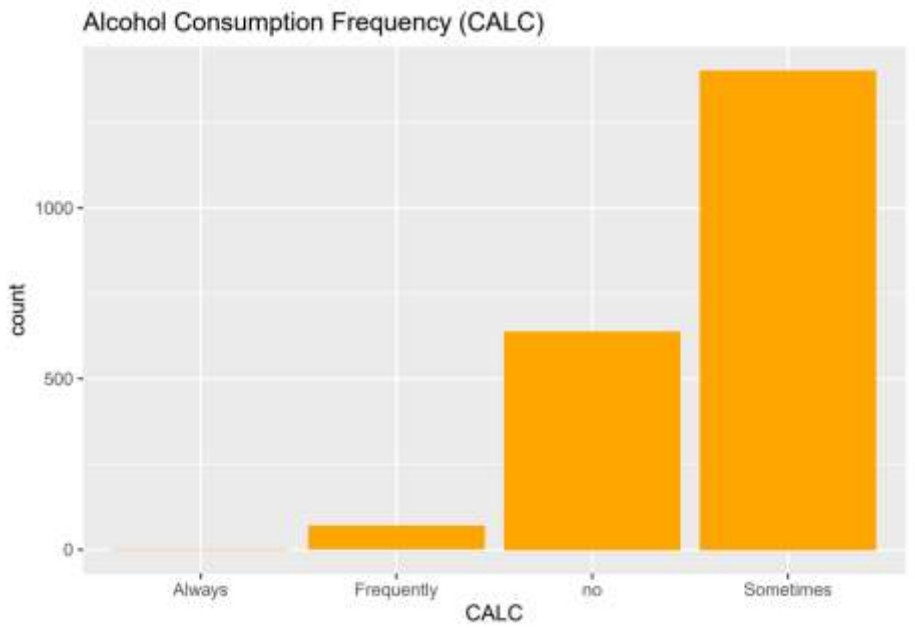
DEFINITION OF VARIABLES

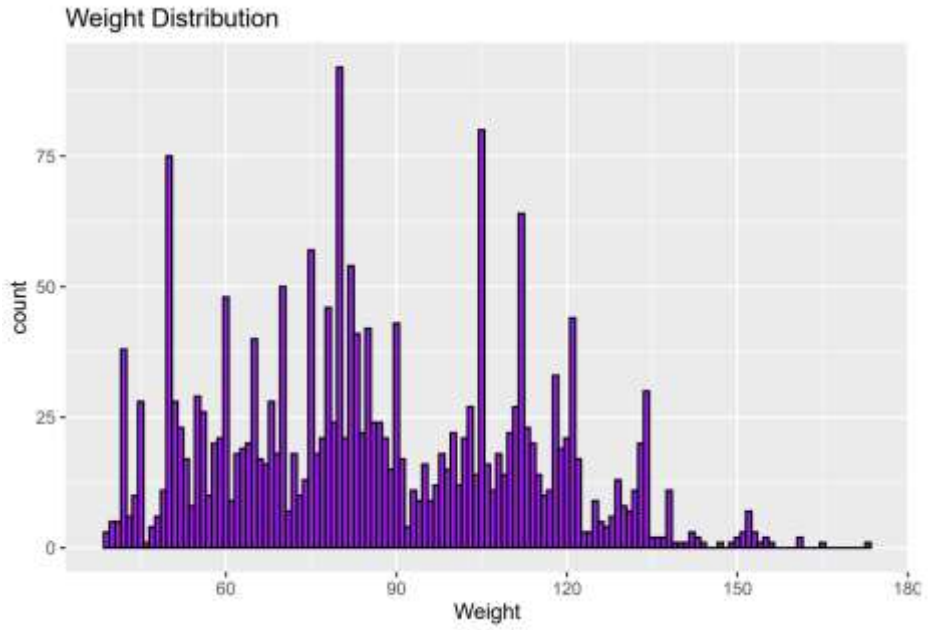
- **Gender**: Gender (Male/Female).
- **Age**: Age (numeric value).
- **Height**: Height (in meters).
- **Weight**: Weight (in kilograms).
- **family_history_with_overweight**: Family history of obesity (Yes/No).
- **FAVC**: Habit of consuming high-calorie food (Yes/No).
- **FCVC**: Frequency of vegetable consumption (a value between 1-3).
- **NCP**: Number of main meals per day (numeric value).
- **SMOKE**: Smoking habit (Yes/No).
- **CH2O**: Daily water consumption (in liters).
- **SCC**: Habit of controlling calories (Yes/No).
- **FAF**: Weekly physical activity time (in hours).
- **TUE**: Daily technology usage time (in hours).
- **CALC**: Alcohol consumption frequency (None/Weekly/Daily/Special Occasions).
- **MTRANS**: Mode of daily transportation (Car/Motorbike/Bicycle/Walking/Public transport).
- **NObeyesdad**: Target variable. Obesity level (Normal weight, Overweight, Slightly obese, Severely obese, etc.).

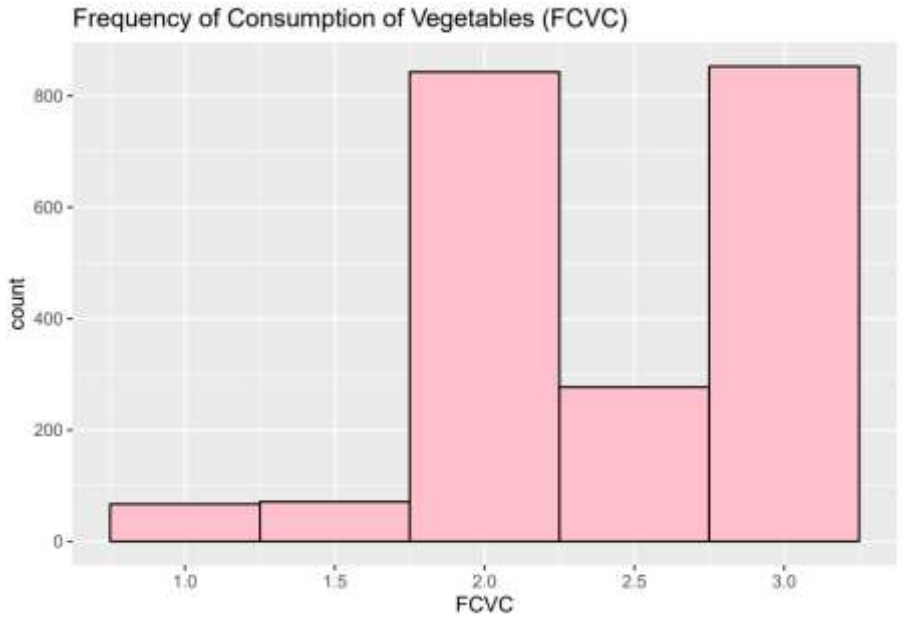
```r
# Check the number of rows and columns
str(data)
```

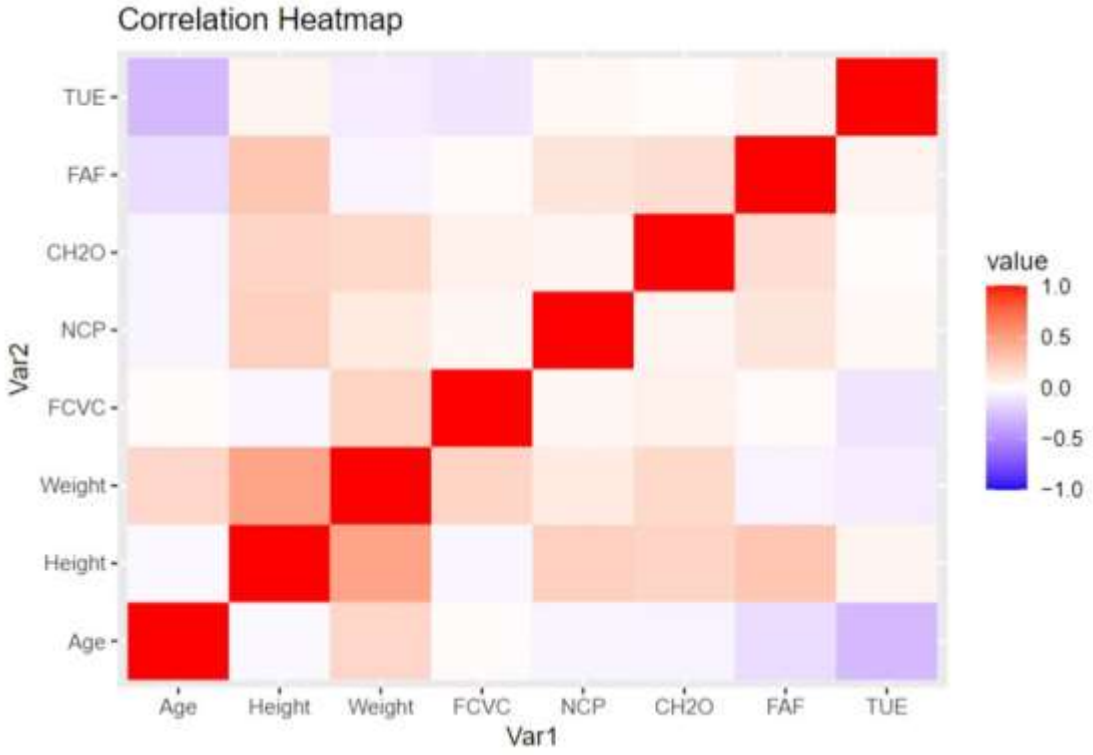
```
## spc_tbl_ [2,111 x 17] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ Gender                         : chr [1:2111] "Female" "Female" "Male" "Male" ...
##  $ Age                            : num [1:2111] 21 21 23 27 22 29 23 22 24 22 ...
##  $ Height                         : num [1:2111] 1.62 1.52 1.8 1.8 1.78 1.62 1.5 1.64 1.78 1.72 ...
##  $ Weight                         : num [1:2111] 64 56 77 87 89.8 53 55 53 64 68 ...
##  $ family_history_with_overweight: chr [1:2111] "yes" "yes" "yes" "no" ...
##  $ FAVC                           : chr [1:2111] "no" "no" "no" "no" ...
##  $ FCVC                           : num [1:2111] 2 3 2 3 2 2 3 2 3 2 ...
##  $ NCP                            : num [1:2111] 3 3 3 3 1 3 3 3 3 3 ...
##  $ CAEC                           : chr [1:2111] "Sometimes" "Sometimes" "Sometimes" "Sometimes" ...
##  $ SMOKE                          : chr [1:2111] "no" "yes" "no" "no" ...
##  $ CH2O                           : num [1:2111] 2 3 2 2 2 2 2 2 2 2 ...
##  $ SCC                            : chr [1:2111] "no" "yes" "no" "no" ...
##  $ FAF                            : num [1:2111] 0 3 2 2 0 0 1 3 1 1 ...
##  $ TUE                            : num [1:2111] 1 0 1 0 0 0 0 0 1 1 ...
##  $ CALC                           : chr [1:2111] "no" "Sometimes" "Frequently" "Frequently" ...
##  $ MTRANS                         : chr [1:2111] "Public_Transportation" "Public_Transportation" "Pub
##  $ NObeyesdad                     : chr [1:2111] "Normal_Weight" "Normal_Weight" "Normal_Weight" "Ove
##  - attr(*, "spec")=
##   .. cols(
##   ..   Gender = col_character(),
##   ..   Age = col_double(),
##   ..   Height = col_double(),
##   ..   Weight = col_double(),
##   ..   family_history_with_overweight = col_character(),
##   ..   FAVC = col_character(),
##   ..   FCVC = col_double(),
##   ..   NCP = col_double(),
##   ..   CAEC = col_character(),
##   ..   SMOKE = col_character(),
##   ..   CH2O = col_double(),
##   ..   SCC = col_character(),
##   ..   FAF = col_double(),
##   ..   TUE = col_double(),
##   ..   CALC = col_character(),
##   ..   MTRANS = col_character(),
##   ..   NObeyesdad = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```


Boxplot of Frequency of Consumption of Vegetables (FCVC)
FCVC
3.0
2.5
2.0
1.5
1.0
−0.4
−0.2
0.0
0.2
0.4


Alcohol Consumption Frequency (CALC)
count
1000
500
0
Always
Frequently
no
Sometimes
CALC


Weight Distribution
count
75
50
25
0
60
90
120
150
Weight


Frequency of Consumption of Vegetables (FCVC)
count
800
600
400
200
0
1.0
1.5
2.0
2.5
3.0
FCVC



As a result of examining and visualizing the data, the variables that are considered to directly affect obesity and are relatively easier to address with solutions have been identified as follows;

- FCVC (Frequency of Vegetable Consumption): Increase vegetable consumption.
- FAVC (Consumption of High-Calorie Food): Reduce the consumption of high-calorie foods.
- FAF (Physical Activity Time): Increase physical activity.
- CH2O (Water Consumption): Increase daily water consumption.
- TUE (Technology Usage Time): Reduce technology usage time to create more time for physical activity.

In this context, we can begin exploring the obesity situation by focusing on the positive changes in these variables.

4.MODEL

Obesity status (NObeyesdad) is usually determined using a metric like Body Mass Index (BMI). BMI is a parameter that helps estimate a person's body fat percentage based on their height and weight.

BMI is calculated using the following formula:

$$BMI = \text{Weight (kg)} / \text{Height (m)}2$$

```
data$BMI <- data$Weight / (data$Height^2)  # BMI hesaplama

# Select overweight and obese individuals (BMI >= 25)
overweight_obese_data <- data[data$BMI >= 27.45, ]

# Display the number of overweight or obese individuals
cat("Number of overweight or obese individuals: ",
    nrow(overweight_obese_data), "\n")

## Number of overweight or obese individuals:  1210
```

```r
# Standardization
overweight_obese_data_scaled <- overweight_obese_data
overweight_obese_data_scaled[, c("FCVC", "FAVC", "FAF", "CH2O", "TUE")] <-
  scale(overweight_obese_data[, c("FCVC", "FAVC", "FAF", "CH2O", "TUE")])

# Recreate model using the standardized data
model_scaled <- lm(Weight ~ FCVC + FAVC + FAF + CH2O + TUE,
                   data = overweight_obese_data_scaled)
```

```r
# New data for prediction based on mean values with specific adjustments
new_data <- data.frame(
  FCVC = mean(overweight_obese_data$FCVC) * 1.02,
  # 2% increase in vegetable consumption
  FAVC = mean(overweight_obese_data$FAVC) * 0.98,
  # 2% decrease in high-calorie food consumption
  FAF = mean(overweight_obese_data$FAF) * 1.05,
  # 5% increase in physical activity
  CH2O = mean(overweight_obese_data$CH2O) * 1.02,
  # 2% increase in water consumption
  TUE = mean(overweight_obese_data$TUE) * 0.95
  # 5% decrease in technology use
)
```


Original vs Adjusted Weights
Visualizing the Difference Between Original and Adjusted Weights
Adjusted Weight (kg)
150
125
100
75
50
60
90
120
150
Original Weight (kg)
Weight_Difference
10

```
##
## Call:
## lm(formula = Weight ~ FCVC + FAVC + FAF + CH2O + TUE, data = train_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -41.990 -11.569  -1.485  13.150  53.117
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   6.7254     5.0468   1.333    0.183
## FCVC         14.0693     1.0108  13.919  < 2e-16 ***
## FAVC         25.6381     2.0906  12.263  < 2e-16 ***
## FAF           3.4866     0.6673   5.225 2.14e-07 ***
## CH2O          4.7735     0.8750   5.455 6.22e-08 ***
## TUE          -0.2230     0.9223  -0.242    0.809
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.09 on 964 degrees of freedom
## Multiple R-squared:  0.3049, Adjusted R-squared:  0.3013
## F-statistic: 84.55 on 5 and 964 DF,  p-value: < 2.2e-16
```

```r
# Limit the weight change predictions between -10 and 10
predicted_weight_changes_scaled <-
  pmin(pmax(predicted_weight_changes_scaled, -10), 10)
```

5. MODEL TESTING AND RESULTS

Good food
is *good mood.*

THANK YOU FOR LISTENING!