Project Report Group- 06

Estimation of obesity levels based on eating habits and physical condition.

Lahari Kaja

Sivaramakrishna Pallabothula

Sudharani Yeruva

Uma Buchammagari

Ushaswini Sunkara

Introduction

The global obesity surge poses significant challenges across health, economics, and societal well-being, with obesity defined by WHO as excessive adipose tissue accumulation. Despite efforts to promote healthy habits, understanding the complex interplay between diet, exercise, and obesity requires innovative approaches (Bag et al., 2023).

Research Question

This study investigates the factors influencing obesity levels based on individuals' eating habits and physical condition. By analyzing the relationships between lifestyle factors and obesity, the research aims to identify contributors to the obesity epidemic, informing prevention and management strategies

Dataset and variables

The dataset, "ObesityDataSet.csv," comprises 2111 observations and 17 variables related to demographics, physical measurements, eating habits, and lifestyle factors. With a robust sample size, it enables a comprehensive examination of potential influences on obesity levels.

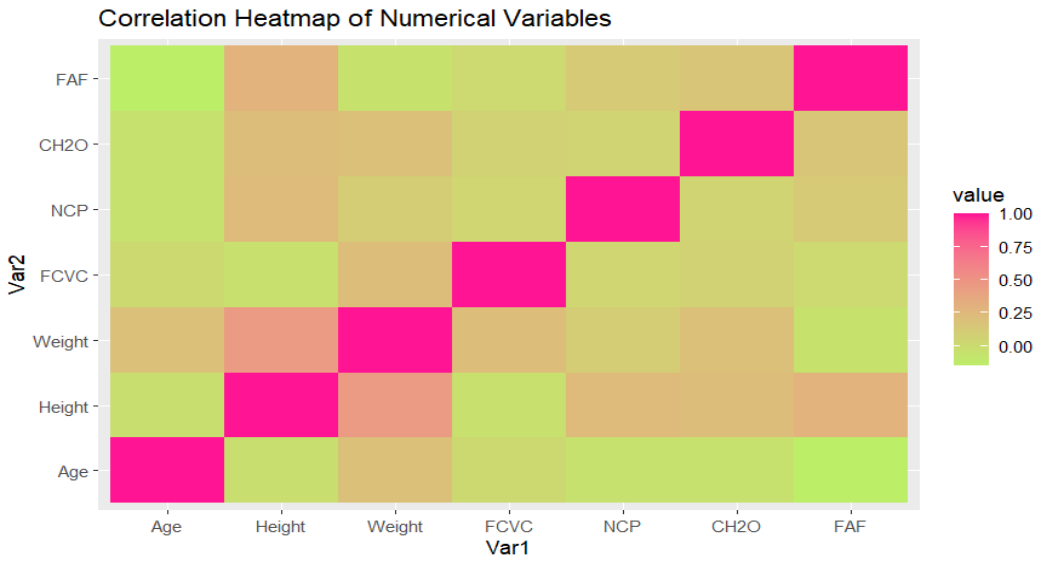
Statistical Analysis:

Descriptive Statistics:

Descriptive statistics provide insights into the dataset's characteristics, such as central tendency and dispersion, revealing patterns in variables like age, weight, and dietary habits. These analyses offer valuable context for understanding the factors influencing obesity levels, emphasizing the significance of healthy behaviors and the need for targeted interventions to address obesity risk factors effectively.

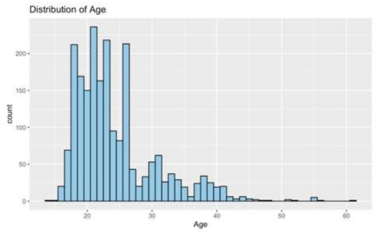
Correlation Analysis:

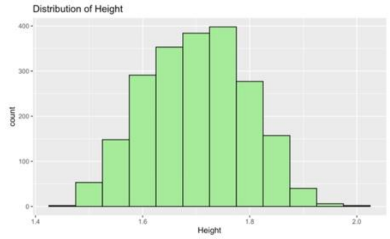
Correlation analysis identifies relationships between variables, such as age, weight, and dietary habits, revealing patterns of association that inform our understanding of obesity determinants. These findings highlight potential factors influencing obesity levels, guiding the development of targeted interventions to address modifiable risk factors and mitigate the global obesity epidemic.In the analysis, the correlation matrix showed the correlation coefficients between pairs of numerical variables such as Age, Height, Weight, FCVC (Frequency of Consumption of Vegetables), NCP (Number of Main Meals), and CH2O (Water Consumption).

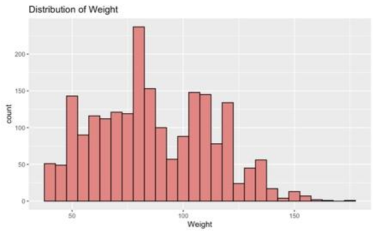


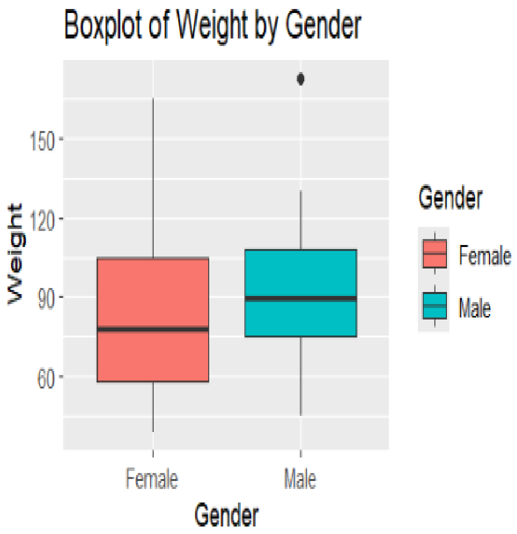
Data Visualization:

Data visualization is a powerful tool that we have used for exploring and understanding the patterns and relationships within the data. Histograms are used to display the distribution of continuous variables, such as Age, Height, and Weight. They provided insights into the shape of the distribution, the presence of outliers, and the concentration of values.





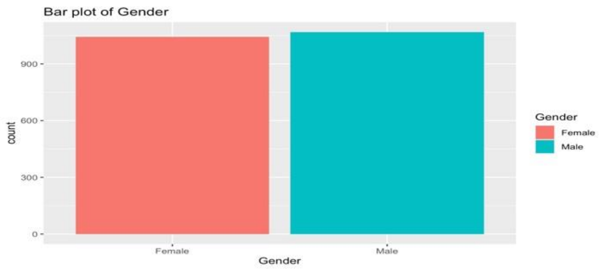


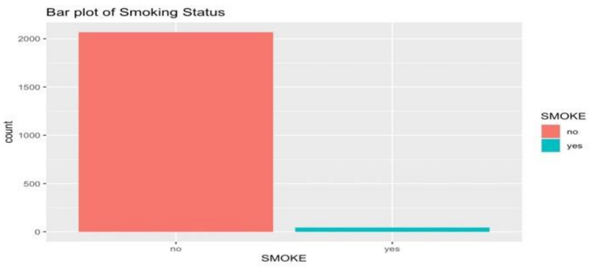


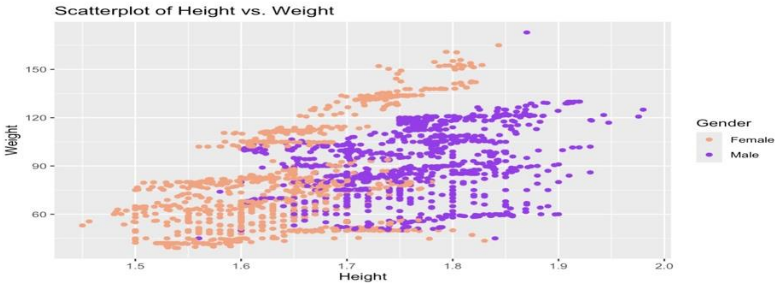
Boxplots are used to compare the distribution of a continuous variable across different categories. In the analysis, boxplots are used to compare the distribution of Weight by Gender and the distribution of Age by Obesity Level.

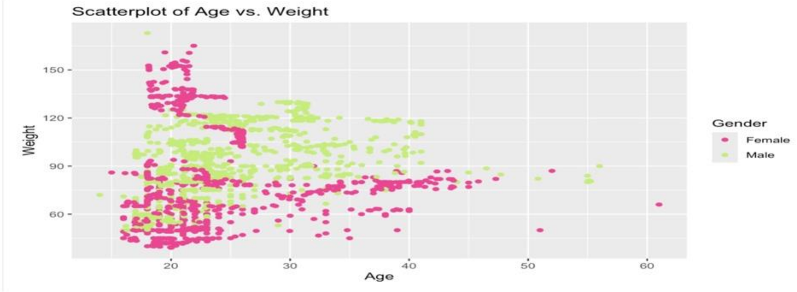
Bar plots are employed to visualize the frequency or proportion of categorical variables. The analysis included bar plots for Gender, Smoking Status, and the distribution of Obesity Levels by Gender and Family History with Overweight.

Scatterplots are used to explore the relationship between two continuous variables. The analysis included scatterplots of Age vs. Height and Height vs. Weight, with the points colored by Gender. These plots help identify any patterns, trends, or clustering of data points based on the variables being analyzed.









Inferential Statistics:

Inferential statistics, specifically chi-square tests, were utilized to assess relationships between categorical variables and obesity levels, comparing observed frequencies with expected frequencies under the null hypothesis. A significance level of $p < 0.05$ indicated strong evidence against the null hypothesis, suggesting significant associations. Variables such as Gender, Family History with Overweight, FAVC, and CAEC were subjected to chi-square tests, with results showing p-values < 0.05, indicating significant relationships with obesity levels. For instance, the chi-square test for Gender and Obesity Level demonstrated a highly significant association ($p < 2.2e-16$), highlighting gender's impact on obesity levels.

Regression Analysis:

Regression analysis models relationships between predictor variables and an outcome, here, BMI or obesity levels. Linear regression assessed BMI's relationship with Weight Category, revealing a significant positive association. Multiple linear regression considered various factors simultaneously, finding significant associations between BMI and variables like high-calorie food consumption, family history of overweight, and not counting calories. Regression coefficients indicated the change in BMI associated with predictor variables, providing insights into their effects and significance. This analysis identified key factors influencing BMI, informing strategies for obesity prevention and management.

Limitations and Appropriateness:

Acknowledging limitations, the study's observational nature hinders causal inference, despite identifying associations and correlations. Linear regression assumptions may oversimplify relationships, urging future exploration of non-linear modeling techniques. The cross-sectional dataset limits understanding of long-term effects, emphasizing the need for longitudinal data to elucidate temporal dynamics in obesity research. Nonetheless, the study's comprehensive statistical approach offers valuable insights into obesity determinants given the available data.

Interpretation and Conclusion:

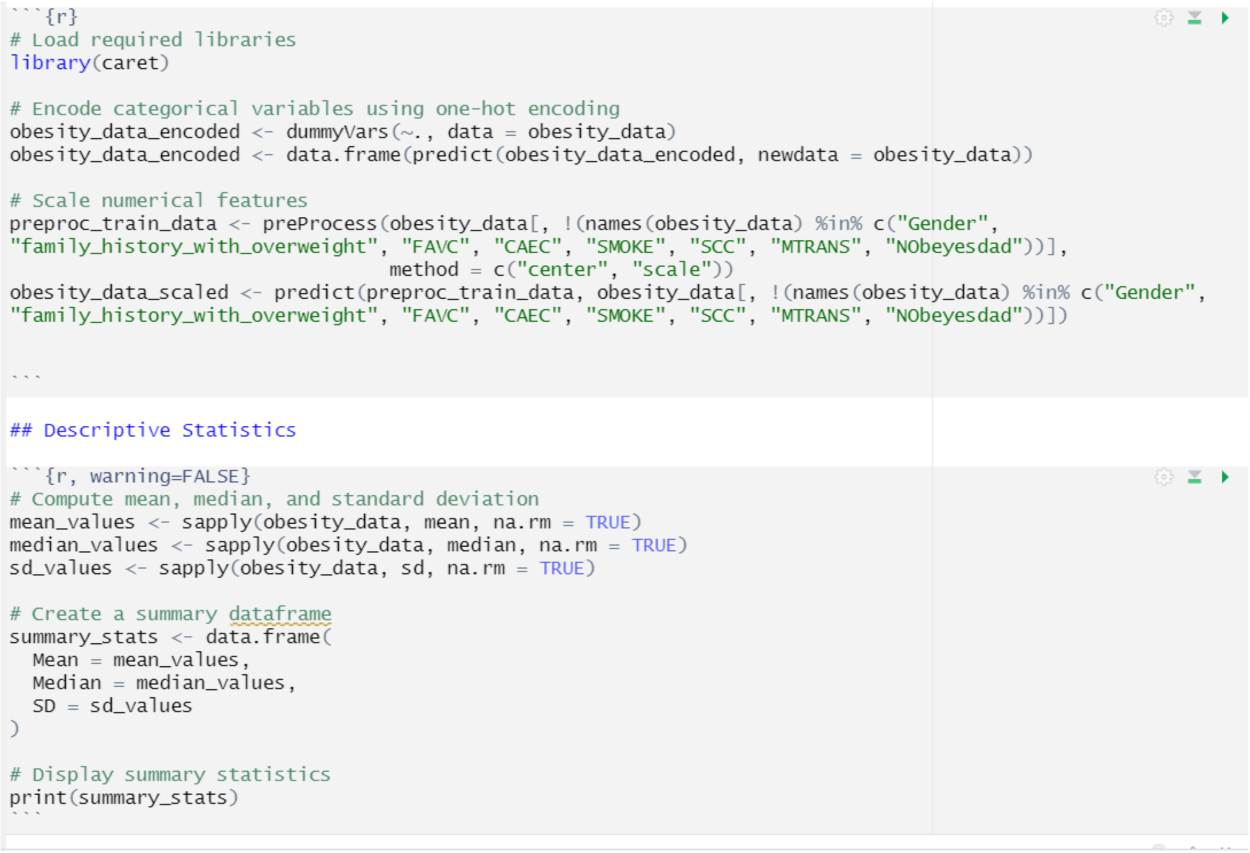
This study identifies dietary habits, notably high-calorie food consumption and neglecting calorie counting, as significant predictors of higher BMI and obesity levels, advocating for promoting healthier eating habits and calorie awareness. Family history of overweight emerges as influential, emphasizing the necessity for targeted interventions to support individuals with such predispositions. Gender differences in obesity levels highlight the need for gender-specific strategies, while the roles of smoking and physical activity warrant further investigation for comprehensive obesity prevention and management. Overall, this study underscores the importance of understanding and addressing multifactorial influences on obesity and suggests avenues for future research to enhance intervention effectiveness.

References

Bag, H. G., Yagin, F. H., Görmez, Y., Prieto-Gonzalez, P., Colak, C., Gulu, M., Badicu, G., & Ardigo, L. P. (2023). Estimation of Obesity Levels through the Proposed Predictive Approach Based on Physical Activity and Nutritional Habits. *Diagnostics*, *13*(18), 2949. https://doi.org/10.3390/diagnostics13182949

Appendix

Descriptive statistics:

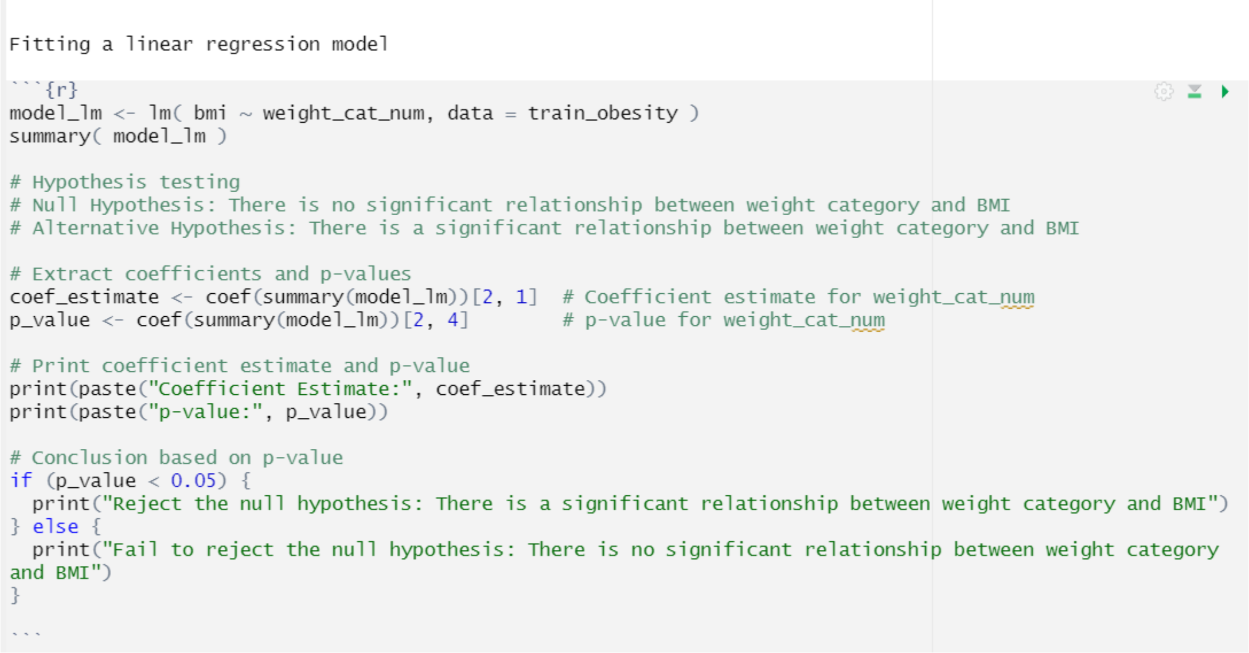


~~~
```{r}
# Load required libraries
library(caret)

# Encode categorical variables using one-hot encoding
obesity_data_encoded <- dummyVars(~., data = obesity_data)
obesity_data_encoded <- data.frame(predict(obesity_data_encoded, newdata = obesity_data))

# Scale numerical features
preproc_train_data <- preProcess(obesity_data[, !(names(obesity_data) %in% c("Gender",
"family_history_with_overweight", "FAVC", "CAEC", "SMOKE", "SCC", "MTRANS", "NObeyesdad"))],
                                 method = c("center", "scale"))
obesity_data_scaled <- predict(preproc_train_data, obesity_data[, !(names(obesity_data) %in% c("Gender",
"family_history_with_overweight", "FAVC", "CAEC", "SMOKE", "SCC", "MTRANS", "NObeyesdad"))])


~~~

Descriptive Statistics

```{r, warning=FALSE}
# Compute mean, median, and standard deviation
mean_values <- sapply(obesity_data, mean, na.rm = TRUE)
median_values <- sapply(obesity_data, median, na.rm = TRUE)
sd_values <- sapply(obesity_data, sd, na.rm = TRUE)

# Create a summary dataframe
summary_stats <- data.frame(
  Mean = mean_values,
  Median = median_values,
  SD = sd_values
)

# Display summary statistics
print(summary_stats)
```
```
```

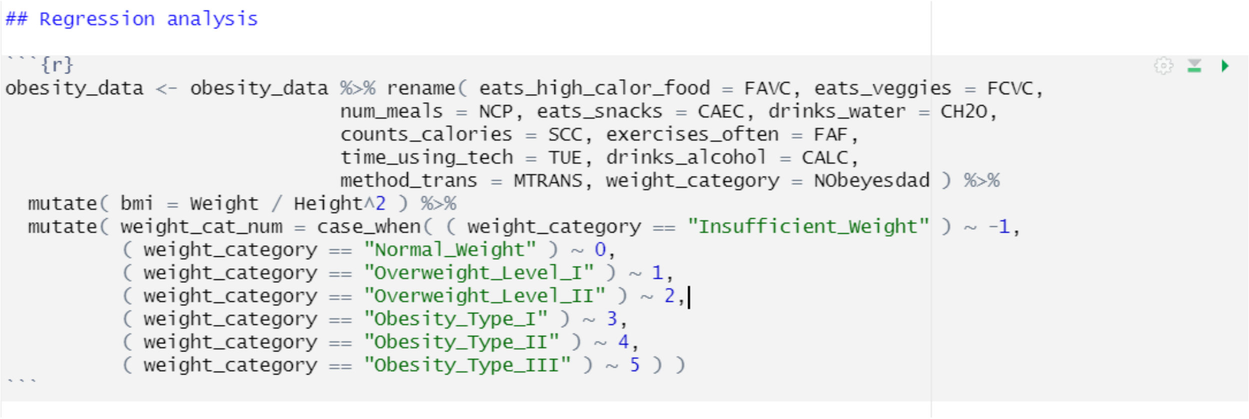
Linear Regression Model:



Fitting a linear regression model

~~~
```{r}
model_lm <- lm( bmi ~ weight_cat_num, data = train_obesity )
summary( model_lm )

# Hypothesis testing
# Null Hypothesis: There is no significant relationship between weight category and BMI
# Alternative Hypothesis: There is a significant relationship between weight category and BMI

# Extract coefficients and p-values
coef_estimate <- coef(summary(model_lm))[2, 1]  # Coefficient estimate for weight_cat_num
p_value <- coef(summary(model_lm))[2, 4]        # p-value for weight_cat_num

# Print coefficient estimate and p-value
print(paste("Coefficient Estimate:", coef_estimate))
print(paste("p-value:", p_value))

# Conclusion based on p-value
if (p_value < 0.05) {
  print("Reject the null hypothesis: There is a significant relationship between weight category and BMI")
} else {
  print("Fail to reject the null hypothesis: There is no significant relationship between weight category 
and BMI")
}

```
~~~

Regression Analysis:



Regression analysis

~~~
```{r}
obesity_data <- obesity_data %>% rename( eats_high_calor_food = FAVC, eats_veggies = FCVC,
                             num_meals = NCP, eats_snacks = CAEC, drinks_water = CH2O,
                             counts_calories = SCC, exercises_often = FAF,
                             time_using_tech = TUE, drinks_alcohol = CALC,
                             method_trans = MTRANS, weight_category = NObeyesdad ) %>%
  mutate( bmi = Weight / Height^2 ) %>%
  mutate( weight_cat_num = case_when( ( weight_category == "Insufficient_Weight" ) ~ -1,
          ( weight_category == "Normal_Weight" ) ~ 0,
          ( weight_category == "Overweight_Level_I" ) ~ 1,
          ( weight_category == "Overweight_Level_II" ) ~ 2,
          ( weight_category == "Obesity_Type_I" ) ~ 3,
          ( weight_category == "Obesity_Type_II" ) ~ 4,
          ( weight_category == "Obesity_Type_III" ) ~ 5 ) )
```
~~~