Module Project 1

CECS 450 Data Visualisation

Team Member:
Ilmaan Zia - 030849843
Nikita Lalwani - 030861790
Ankit Raj - 030824350

I . Dataset Overview

Name : Estimation of obesity levels based on eating habits and physical condition
Description:
This dataset comprises information for assessing obesity levels among individuals in Mexico, Peru, and Colombia, determined by their dietary patterns and physical fitness. It consists of 17 attributes and 2111 entries, each labeled with the variable NObesity (Obesity Level), allowing classification into categories such as Insufficient Weight, Normal Weight, Overweight Level I, Overweight Level II, Obesity Type I, Obesity Type II, and Obesity Type III. Approximately 77% of the dataset was artificially generated using the Weka tool and the SMOTE filter, while the remaining 23% was gathered directly from users via a web platform.

Variable Table

Variable Name	Role	Type	Description	Missing Value
Gender	Feature	Categorical		no
Age	Feature	Continuous		no
Height	Feature	Continuous		no
Weight	Feature	Continuous		no
family_history_with_overweight	Feature	Binary	Has a family member suffered or suffers from being overweight?	no
FAVC	Feature	Binary	Do you eat high caloric food frequently?	no

FCVC	Feature	Integer	Do you usually eat vegetables in your meals?	no
NCP	Feature	Continuous	How many main meals do you have daily?	no
CAEC	Feature	Categorical	Do you eat any food between meals?	no
SMOKE	Feature	Binary	Do you smoke?	no

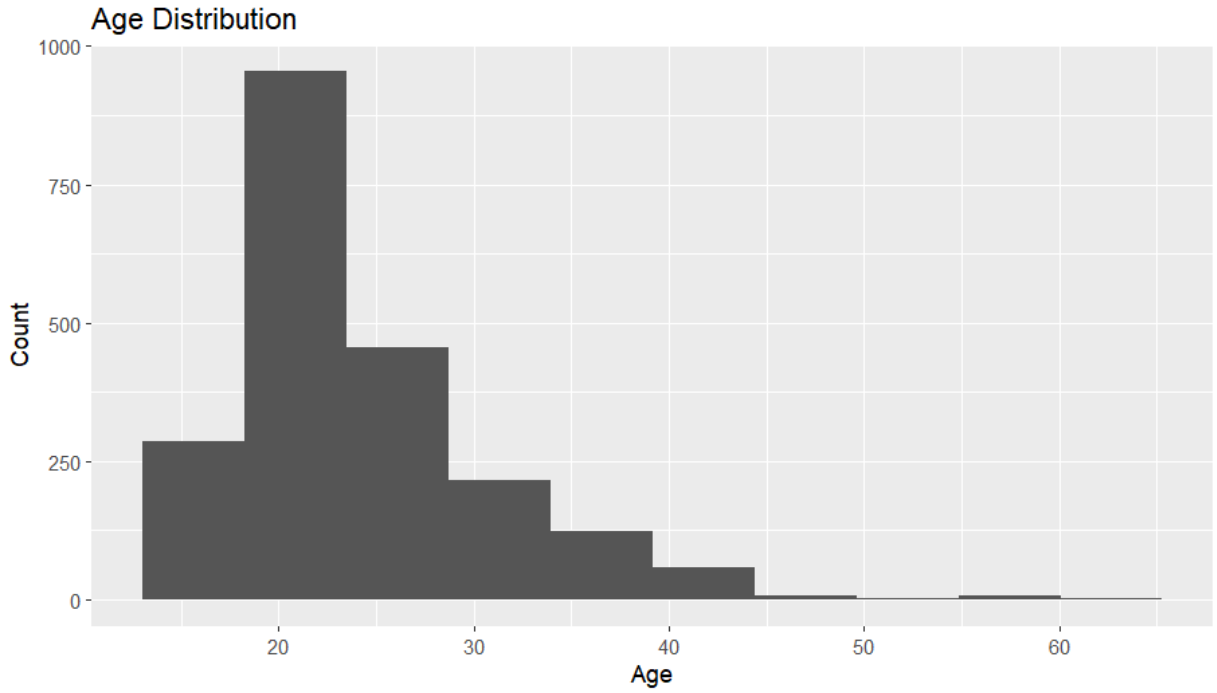
CH2O	Feature	Continuous	How much water do you drink daily?	no
SCC	Feature	Binary	Do you monitor the calories you eat daily?	no
FAF	Feature	Continuous	How often do you have physical activity?	no
TUE	Feature	Integer	How much time do you use technological devices such as cell phones, video games, television, computer and others?	no
CALC	Feature	Categorical	How often do you drink alcohol?	no
MTRANS	Feature	Categorical	Which transportation do you usually use?	no
NObeyesdad	Target	Categorical	Obesity level	No

Sample of Data

	Gender	Age	Height	Weight	family_history_with_overweight	FAVC	FCVC	NCP	CAEC	SMOKE	CH2O	SCC	FAF	TUE	CALC	MTRANS	NObeyesdad
1	Female	21	1.62	64.0	yes	no	2	3	Sometimes	no	2	no	0	1	no	Public_Transportation	Normal_Weight
2	Female	21	1.52	56.0	yes	no	3	3	Sometimes	yes	3	yes	3	0	Sometimes	Public_Transportation	Normal_Weight
3	Male	23	1.80	77.0	yes	no	2	3	Sometimes	no	2	no	2	1	Frequently	Public_Transportation	Normal_Weight
4	Male	27	1.80	87.0	no	no	3	3	Sometimes	no	2	no	2	0	Frequently	Walking	Overweight_Level_I
5	Male	22	1.78	89.8	no	no	2	1	Sometimes	no	2	no	0	0	Sometimes	Public_Transportation	Overweight_Level_II
6	Male	29	1.62	53.0	no	yes	2	3	Sometimes	no	2	no	0	0	Sometimes	Automobile	Normal_Weight

II. Visualizations:

1: Histogram for age distribution in dataset



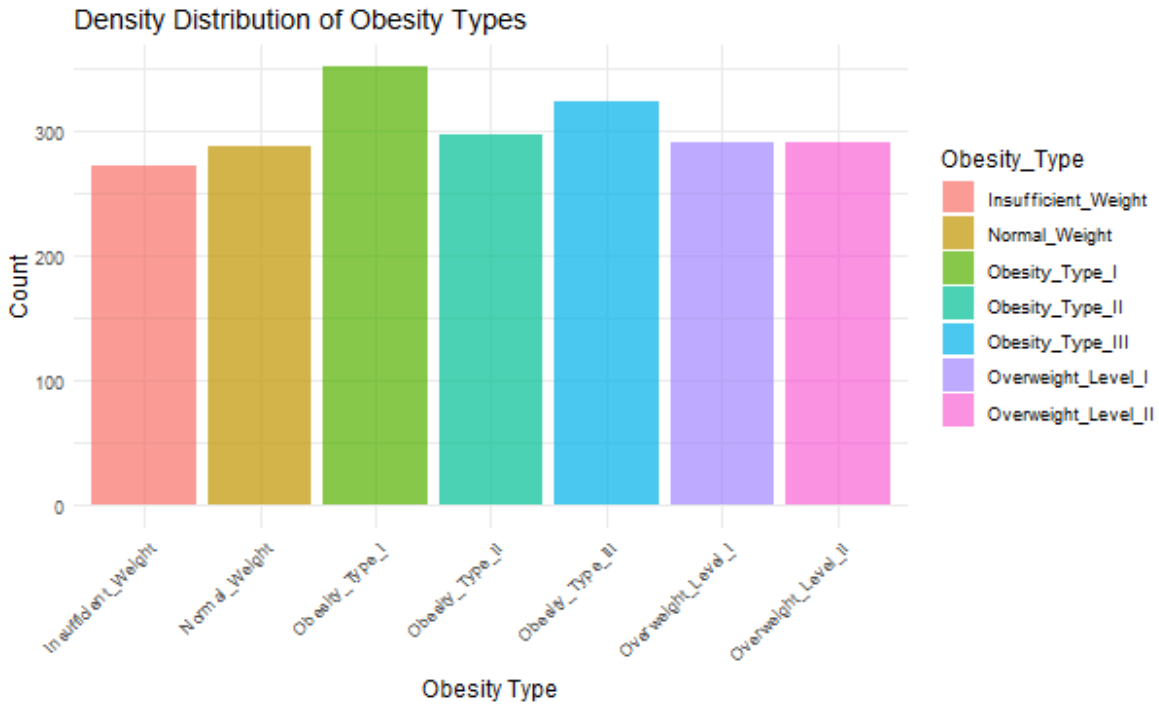
1.1 Inference:

1. The histogram depicts a skewed distribution toward younger ages.
2. The majority of respondents fall within the age range of 20 to 30 years.
3. Beyond the age of 30, there is a noticeable drop in the number of data points.
4. Understanding this age distribution is crucial for targeting a younger demographic.
5. The dataset is particularly relevant for insights and strategies focused on individuals aged 20 to 30, as they constitute the majority of the sampled population.

1.2 Code:

```
# Visualization 1: Histogram for age distribution in dataset

ggplot(data, aes(x=Age)) +
  geom_histogram(bins=10) +
  labs(title="Age Distribution",
       x="Age",
       y="Count")
```

2: Bar Chart: A Balanced Spectrum Across Obesity Types



2.1 Inference:

1. **Balanced Representation**: The roughly equal heights of the bars indicate a balanced representation of different obesity types. This suggests that individuals across various weight categories are well-represented in the dataset.

2. No Dominant Type: There is no clear dominance of any specific obesity type, indicating diversity in the dataset concerning weight-related conditions.

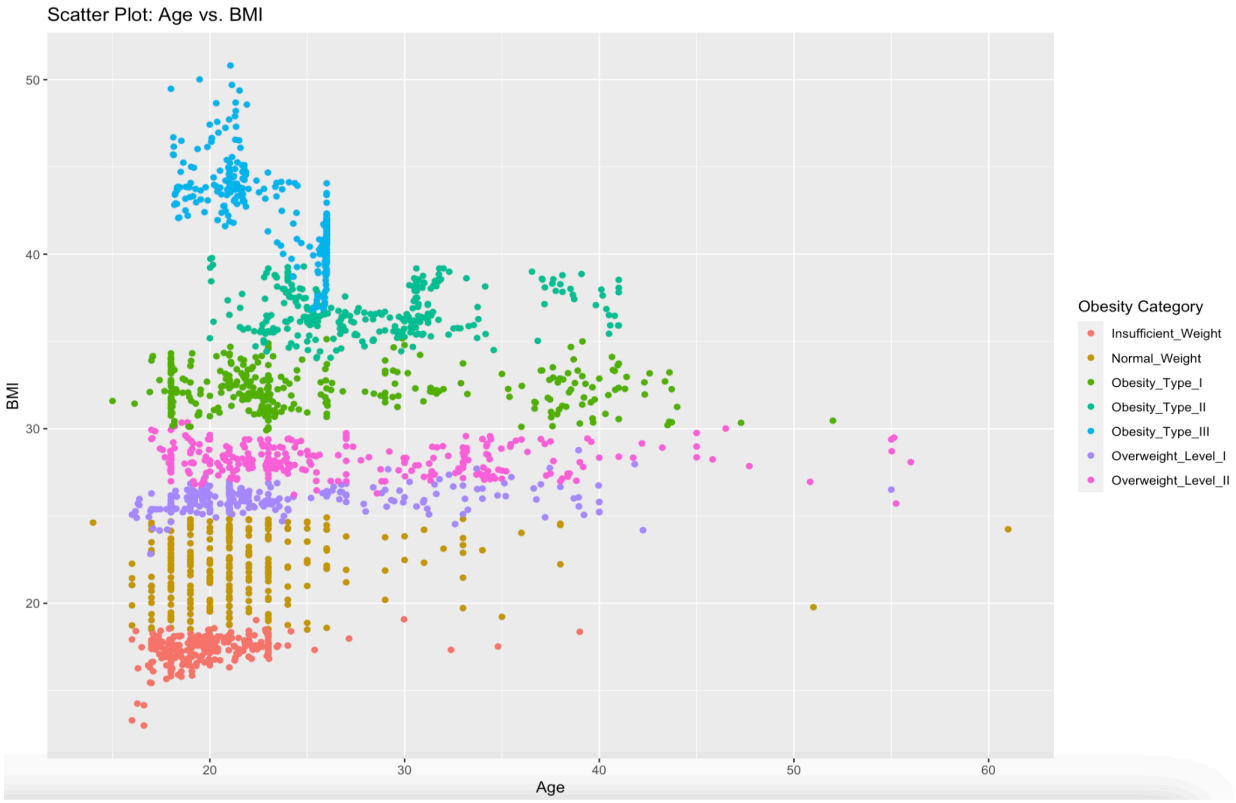
3. No Bias Towards Specific Conditions: The absence of a significant spike or drop in any particular obesity category suggests that there is no bias towards specific weight conditions in the dataset. This balance could be beneficial for conducting analyses that aim to capture a comprehensive understanding of obesity distribution.

2.2 Code:

```
#  Visualization 2: Count occurrences of each obesity type
obesity_counts <- table(data$NObeyesdad)

# Convert to data frame
obesity_counts_df <- as.data.frame(obesity_counts)
names(obesity_counts_df) <- c("Obesity_Type", "Count")
```

```
# Create bar plot
ggplot(data = obesity_counts_df, aes(x = Obesity_Type, y = Count, fill = Obesity_Type)) +
  geom_histogram(stat = "identity", alpha = 0.7) +  # Bar plot
  labs(title = "Density Distribution of Obesity Types",
       x = "Obesity Type",
       y = "Count") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))  # Rotate x-axis labels
```

3: Scatter plot for Age vs. BMI (Body Mass Index)



3.1 Inference:
The scatter plot illustrates the relationship between age and BMI, categorized by different levels of obesity. Here are some key insights:

1. **Insufficient Weight**: Individuals with insufficient weight are predominantly in the younger age group (10-20 years), indicating a lower BMI among younger individuals.
2. **Normal Weight and Overweight**: These categories are spread across all ages but are particularly concentrated in middle age. This suggests that as people age, they tend to move from normal weight to overweight.

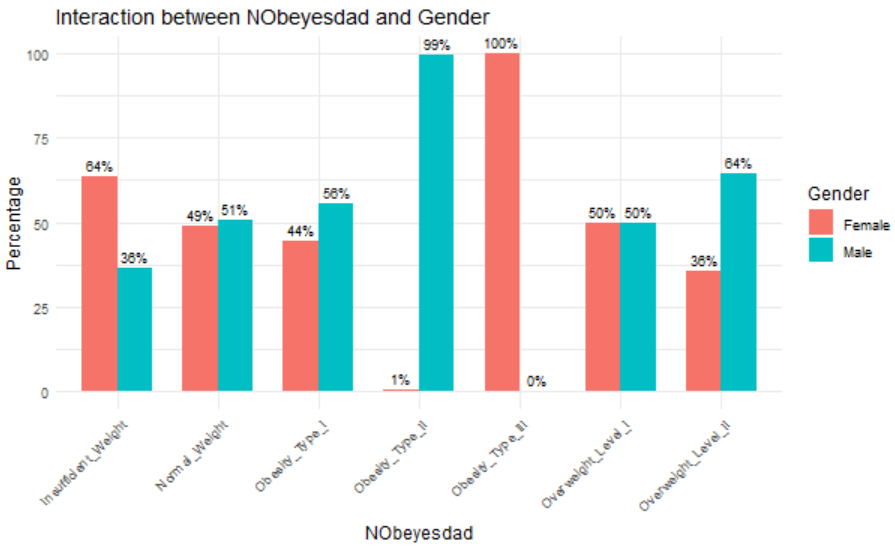
3. **Obesity**: There are fewer data points for Obesity Types I, II, III, indicating less prevalence or data availability. This could also suggest successful health interventions or lifestyle choices preventing obesity.

Overall, the plot suggests a trend where BMI increases with age, and the prevalence of obesity is less compared to other weight categories.

3.2 Code:

```
# Visualization 3: Scatter plot for Age vs. BMI (Body Mass Index)

ggplot(data, aes(x = Age, y = Weight / (Height^2), color = NObeyesdad)) +
  geom_point() +
  labs(title = "Scatter Plot: Age vs. BMI",
       x = "Age",
       y = "BMI",
       color = "Obesity Category")
```

4. Scattered bar plot : Disparities in Obesity Type vs Gender



4.1 Inference:

This is a clustered bar chart

1. Gender Disparities in Obesity Type:
    - Male consistently exhibit higher percentages across various “NObeysedad” categories compared to males.

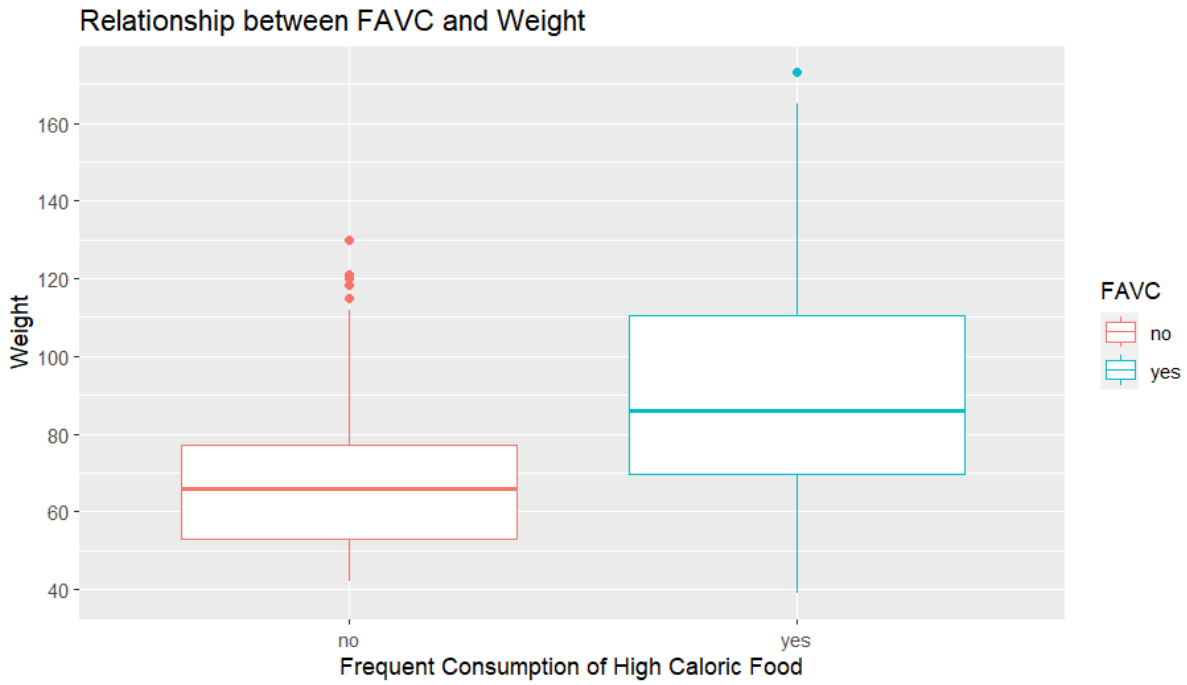
2. Behavioral Variances by Gender in the dataset:
    - Obesity Type 1:
        i. Among females, approximately 99% fall into this category.
        ii. There are no males in this group.
    - Obesity Type 2:
        i. This category is exclusively represented by females, with 100% prevalence.
        ii. No males are affected by Obesity Type 2.

4.2 Code:

```
# Visualization 4: Scattered bar plot : Disparities in Obesity Type vs Gender

cross_plot <- function(dataset, lead_category, sup_category, order = NULL) {
  # Calculate percentages
  percentage_data <- dataset %>%
    group_by(.data[[lead_category]], .data[[sup_category]]) %>%
    summarise(count = n()) %>%
    group_by(.data[[lead_category]]) %>%
    mutate(percentage = count / sum(count) * 100)

  # Create a cross plot
  p <- ggplot(data = percentage_data, aes(x = .data[[lead_category]], y = percentage, fill =
.data[[sup_category]])) +
    geom_bar(stat = "identity", position = "dodge", width = 0.7) +  # Set position to "dodge" and
adjust width of bars
    geom_text(aes(label = paste0(round(percentage), "%")), position = position_dodge(width =
0.7), vjust = -0.5, size = 3) +  # Add percentage labels
    labs(title = paste("Interaction between", lead_category, "and", sup_category),
         x = lead_category,
         y = "Percentage",
         fill = sup_category) +
    theme_minimal() +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))  # Rotate x-axis labels

  # Optionally, reorder the categories
  if (!is.null(order)) {
    p <- p + scale_x_discrete(limits = order)
  }

  print(p)
}
```

cross_plot(data, "NObeyesdad", "Gender")

5: Box plot for relationship between Weight vs. FAVC



5.1 Inference:

1. Non-Frequent Consumers of High Caloric Food:

- Median weight: 67 kg.
- Lower quartile range: 52 kg to 66 kg.
- Upper quartile range: 68 kg to 78 kg.
- Outliers observed with maximum weight: 150 kg.
- Conclusion: An outlier with a weight of 150 kg is considered an error.

2. Frequent Consumers of High Caloric Food:

- Median weight: 88 kg.
- Lower quartile range: Starts from 70 kg.
- Upper quartile range: 89 kg to 110 kg.
- Outlier observed with weight: 173 kg.
- Conclusion: An outlier with a weight of 173 kg is considered an error.

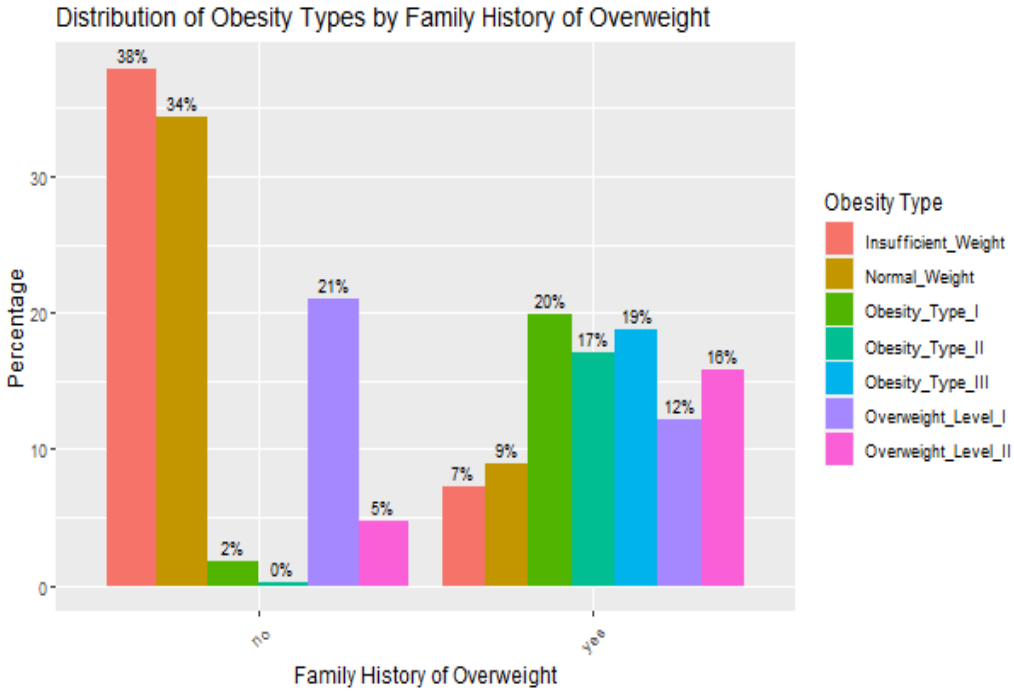
3. Overall Comparison:

- Substantial difference in weights between those who consume high caloric food frequently and those who don't.

5.2 Code:

```
# Visualization 5 : Box plot for relationship between Weight vs. FAVC

ggplot(data, aes(x = FAVC, y = Weight, color = FAVC)) +
  geom_boxplot() +
  labs(title = "Relationship between FAVC and Weight",
       x = "Frequent Consumption of High Caloric Food",
       y = "Weight") +
  scale_y_continuous(breaks = seq(0, max(data$Weight), by = 20))
```

6. Bar Chart + Histogram: Decoding Obesity - The Role of Family History and Genetic Predisposition



6.1 Inference:

Alongside histogram explains:

1. **Family History and Obesity**:
   - Individuals with a family history of overweight are more likely to experience various obesity types.
   - However, Obesity Type 2 does not exhibit significant differences based on family history.
2. **Genetic Predisposition**:Genetic factors play a crucial role in susceptibility to overweight or obesity.

- Shared risk factors may stem from familial lifestyle choices.

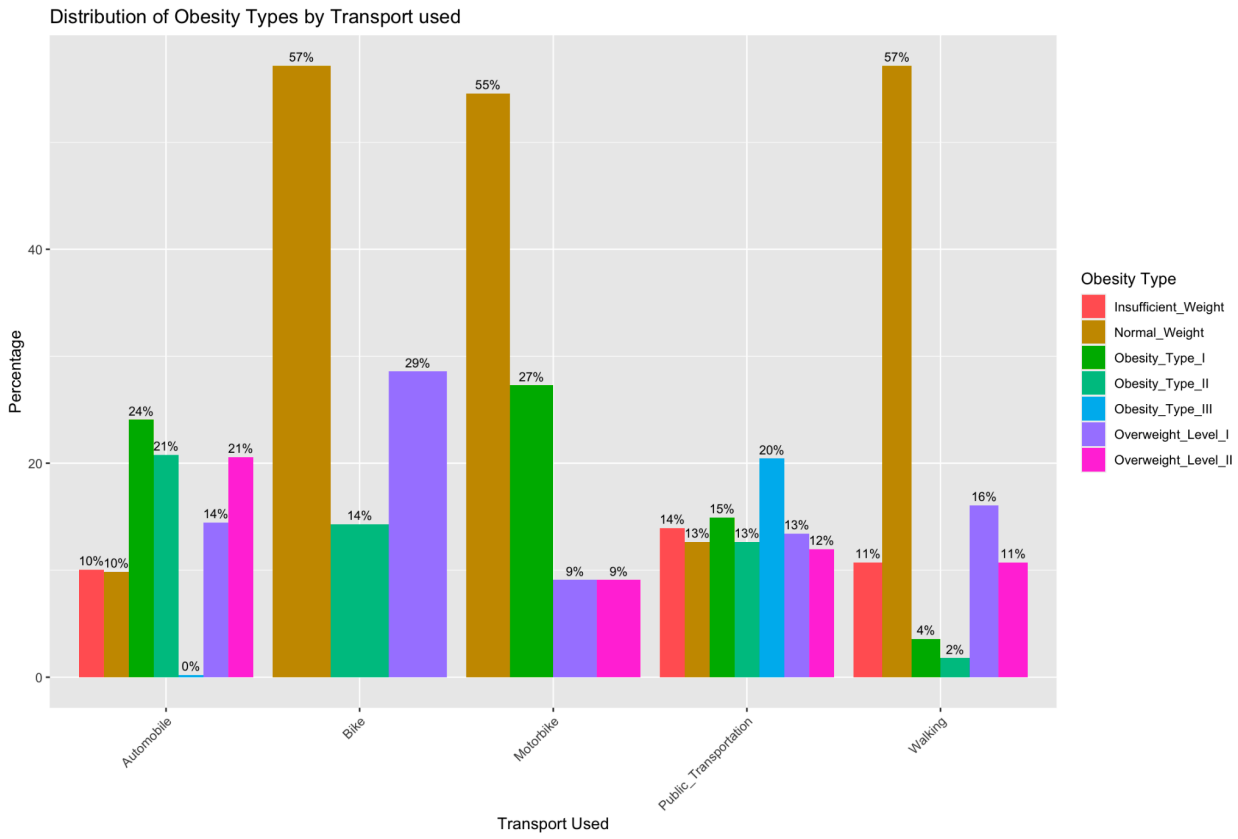
3. **Clinical Considerations**:
   - When designing obesity prevention strategies, incorporate family history as a predictor.

6.2 Code:

 # Visualization 6 :Bar Chart + Histogram: Decoding Obesity - The Role of Family History and Genetic Predisposition

```
# Calculate percentage levels
percentage_data <- data %>%
  group_by(family_history_with_overweight, NObeyesdad) %>%
  summarise(count = n()) %>%
  mutate(percentage = count / sum(count) * 100)

# Create histogram with percentage labels
ggplot(percentage_data, aes(x = family_history_with_overweight, y = percentage, fill = NObeyesdad)) +
  geom_bar(stat = "identity", position = position_dodge()) +
  geom_text(aes(label = paste0(round(percentage), "%")), position = position_dodge(width = 0.9), vjust = -0.5, size = 3) +
  labs(title = "Distribution of Obesity Types by Family History of Overweight",
       x = "Family History of Overweight",
       y = "Percentage",
       fill = "Obesity Type") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))  # Rotate x-axis labels by 45 degrees
```

7: Bar chart+histogram: Distribution of Obesity Types by Transport used



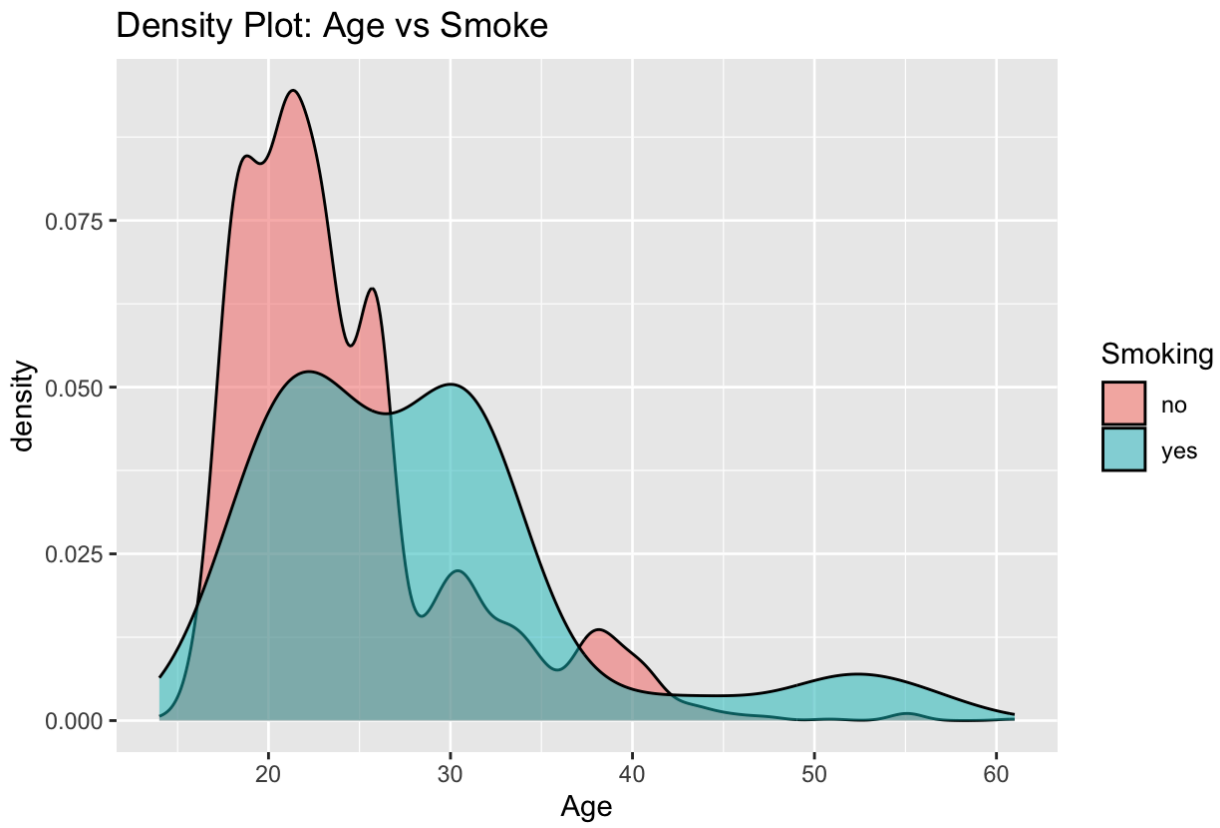
7.1 Inference:

1. **Walking Preference:**
   - Among those who prefer walking, the majority have normal weight.
2. **Public Transportation and Obesity:**
   - For individuals using public transportation, obesity types are equally distributed.
   - No specific trend is observed in obesity types among this group.
3. **Use of Automobile:**
   - Individuals using automobiles show an interesting pattern.
   - No person in this category has obesity type III.
4. **Bike as Transport:**
   - People using bikes as transportation predominantly fall into three categories: normal weight, obesity type I, or overweight level III.
5. **Motorbike Users:**
   - Those who travel by motorbike have no individuals classified as underweight.

7.2 Code:

```
# Visualization 7 : Distribution of Obesity Types by Transport used
# Calculate percentage levels
percentage_data <- data %>%
  group_by(MTRANS, NObeyesdad) %>%
  summarise(count = n()) %>%
  mutate(percentage = count / sum(count) * 100)

# Create histogram with percentage labels
ggplot(percentage_data, aes(x = MTRANS, y = percentage, fill = NObeyesdad)) +
  geom_bar(stat = "identity", position = position_dodge()) +
  geom_text(aes(label = paste0(round(percentage), "%")), position = position_dodge(width =
0.9), vjust = -0.5, size = 3) +
  labs(title = "Distribution of Obesity Types by Transport used",
       x = "Transport Used",
       y = "Percentage",
       fill = "Obesity Type") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

8: Density plot: Age vs Smoke



8.1 Inference:

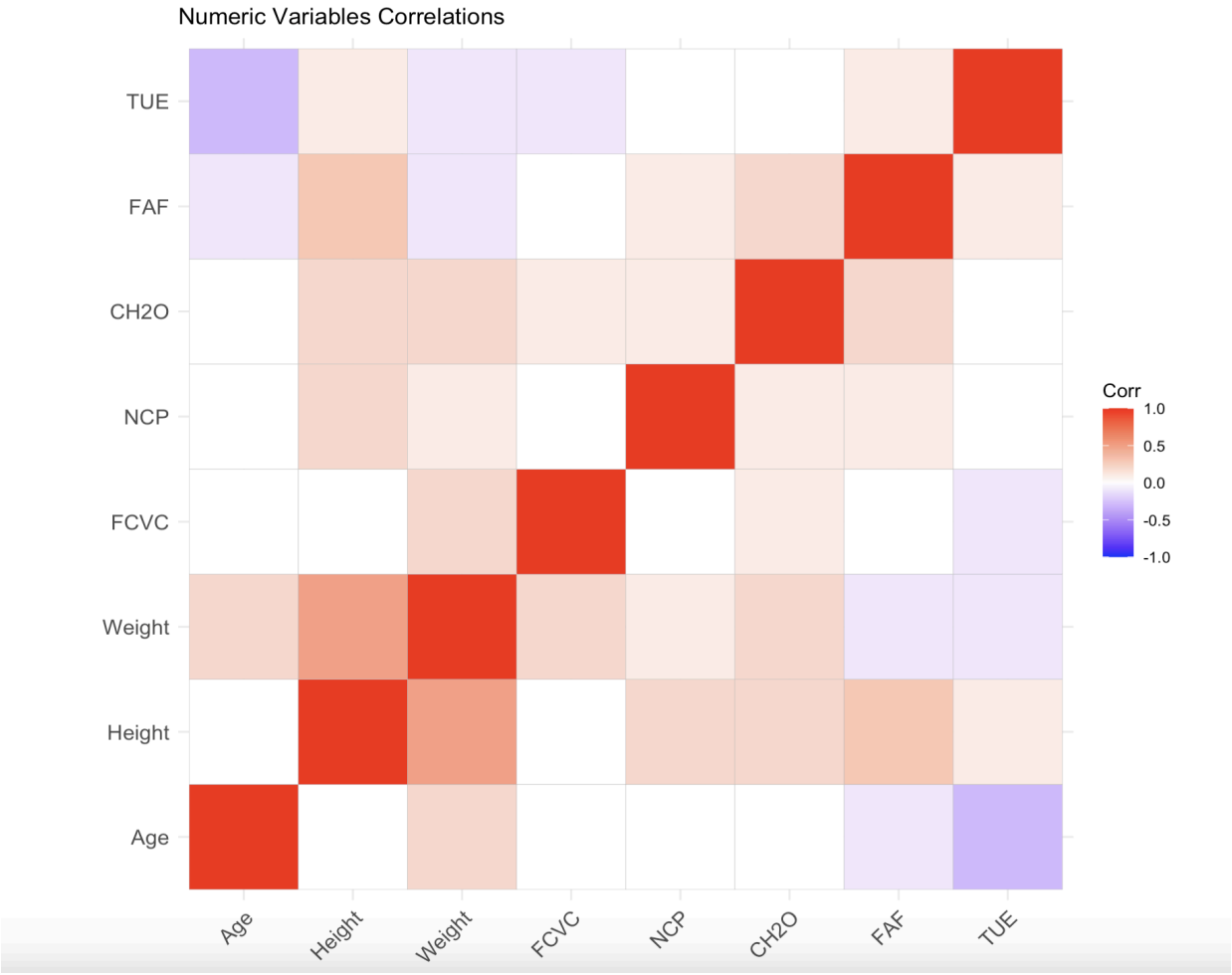
1. The density graph indicates that the highest number of non-smokers falls within the age group of 18-30.
2. Simultaneously, the maximum number of smokers also belongs to the 18-30 age category.
3. A noteworthy observation is that the highest number of individuals above 40 years of age are smokers.
4. Interestingly, the density graph doesn't reveal a clear connection between smoking and obesity.
5. The age group of 18-30, where smoking is prevalent, also contains individuals with obesity. However, the relationship isn't conclusive.

8.2 Code:

```
##Visualization 8: Density plot for Age vs smoking
ggplot(data, aes(x = Age, fill = SMOKE)) +
  geom_density(alpha = 0.5) +
  labs(title = "Density Plot: Age vs Smoke",
       x = "Age",
```

fill = "Smoking")

9: Correlation Map : Numerical variables in dataset



9.1 Inference:

1. TUE: Represents the amount of time you use technological devices such as cell phones, videogames, television, computers, and others.
2. FAF: Reflects how often you engage in physical activity.
3. CH2O: Quantifies how much water you drink daily.
4. NCP: Indicates the number of main meals you have daily.
5. FCVC: Measures how frequently you include vegetables in your meals.
6. Weight: Represents body weight.

7. Height: Represents height.
8. Age: Reflects your age.
9. Weight and Height:
    a. There is a strong positive correlation between weight and height (indicated by the dark red square).
    b. As weight increases, height tends to increase as well.
10. Age and FAF:
    a. There is a negative correlation between age and FAF (indicated by the purple square).
    b. As age increases, FAF (whatever it represents) tends to decrease.
11. Self-Correlation:
    a. The diagonal line from top left to bottom right consists of dark red squares.
    b. These squares represent the self-correlation of each variable (correlation coefficient = 1).

9.2 Code:

```
#Visualization 9:correlation map of variables
# Check the data types of each column
column_types <- sapply(data, class)

# Select columns with numerical values
numeric_columns <- names(column_types[column_types %in% c("numeric", "integer")])

# Display the selected numeric columns
selected_data <- data[, numeric_columns]

#compute correlations between the variables.
#The round function is then applied to round the correlation matrix to one decimal place.
corr <- round(cor(selected_data), 1)
ggcorrplot(corr) +labs(title = "Numeric Variables Correlations")
```