**Project Title: Estimation of Obesity Levels Using Machine Learning**

**Data Analysis and Visualization  
Semester Project Proposal**

**Session *2022-2026***



|  |  |  |
| --- | --- | --- |
| 1 | *Aqeel Abbas Khan* | *NUM-BSCS-2022-38* |
| 2 | *Muhammad Asim Khan* | *NUM-BSCS-2022-14* |

Department of Computer Science

Namal University Mianwali

16 May 2025

**Introduction**

Obesity is a critical global health challenge, with the World Health Organization reporting that over 1 billion people were living with obesity in 2022 ([WHO Obesity Factsheet](https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight)). It is linked to chronic diseases like diabetes and heart disease, making it essential to understand its contributing factors and predict risk for early intervention. This project proposes using a publicly available dataset to analyze obesity levels through machine learning, aiming to identify key predictors and develop accurate predictive models. By exploring demographic, dietary, and lifestyle factors, the study seeks to provide insights that could inform public health strategies and personalized interventions.

**Dataset Source and Description**

The dataset, titled "Estimation of Obesity Levels Based On Eating Habits and Physical Condition," is sourced from the UCI Machine Learning Repository ([UCI Dataset](https://archive.ics.uci.edu/dataset/544/estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition)). It comprises 2111 records and 17 attributes, collected from individuals in Mexico, Peru, and Colombia. The attributes include:

| **Attribute** | **Type** | **Description** |
| --- | --- | --- |
| Gender | Categorical | Male or Female |
| Age | Continuous | Age in years |
| Height | Continuous | Height in meters |
| Weight | Continuous | Weight in kilograms |
| family\_history\_with\_overweight | Binary | Family history of overweight (yes/no) |
| FAVC | Binary | Frequent consumption of high caloric food (yes/no) |
| FCVC | Integer | Frequency of vegetable consumption in meals |
| NCP | Continuous | Number of main meals daily |
| CAEC | Categorical | Consumption of food between meals |
| SMOKE | Binary | Smoking status (yes/no) |
| CH2O | Continuous | Daily water intake in liters |
| SCC | Binary | Monitoring of calorie consumption (yes/no) |
| FAF | Continuous | Frequency of physical activity |
| TUE | Integer | Time using technological devices in hours |
| CALC | Categorical | Alcohol consumption frequency |
| MTRANS | Categorical | Mode of transportation (e.g., walking, public transport) |
| NObeyesdad | Categorical | Obesity level (target variable: Insufficient Weight, Normal Weight, Overweight Level I/II, Obesity Type I/II/III) |

The dataset includes no missing values, with 77% of the data synthetically generated using the Weka tool and SMOTE filter, and 23% collected directly from users via a web platform ([Dataset Paper](https://www.sciencedirect.com/science/article/pii/S2352340919306985)). This composition supports robust analysis but requires caution in generalizing findings.

**Research Questions**

Here are research topics for each feature in our dataset:

Here are research topics for each feature in your dataset:

**Gender**

Research Topic: "Gender-Based Differences in Obesity Risk Factors Among Young Adults."

**Age**

Research Topic: "Correlation Between Age and Obesity Onset: A Cross-Sectional Study."

**Height**

Research Topic: "Investigating the Influence of Height on Body Mass Index and Perceived Obesity Levels."

**Weight**

Research Topic: "Weight as a Predictor of Obesity Level: Machine Learning Versus Traditional Methods."

**Family History with Overweight**

Research Topic: "Genetic and Familial Predisposition to Obesity: Analyzing Family History as a Risk Factor."

**High Caloric Food Consumption**

Research Topic: "Impact of High-Calorie Diets on Body Weight and Obesity Prevalence."

**Use of Vegetables**

Research Topic: "Role of Vegetable Intake in Preventing and Managing Obesity."

**Number of Meals Daily**

Research Topic: "Effect of Meal Frequency on Metabolism and Weight Management."

**Food Between Meals**

Research Topic: "Snacking Habits and Their Association with Obesity Risk."

**Smoking**

Research Topic: "The Paradox of Smoking and Body Weight: Exploring Behavioral Trade-offs."

**CH2O (Water Intake)**

Research Topic: "Hydration and Obesity: Investigating the Link Between Water Consumption and Weight Control."

**Monitor Number of Calories**

Research Topic: "Effectiveness of Caloric Monitoring Apps in Reducing Obesity Rates."

**Physical Activity**

Research Topic: "The Role of Physical Activity Intensity and Frequency in Obesity Prevention."

**Use of Technological Devices**

Research Topic: "Screen Time and Sedentary Behavior: Technology Use as an Obesity Risk Factor."

**Alcohol**

Research Topic: "Alcohol Consumption Patterns and Their Contribution to Weight Gain and Obesity."

**Traveling (Transport Method)**

Research Topic: "Modes of Transportation and Their Impact on Physical Activity and Obesity."

**Obesity Level (Target Variable)**

Research Topic: "Predictive Modeling of Obesity Levels Using Lifestyle and Demographic Features."

These questions align with the dataset’s comprehensive feature set and support the goal of deriving actionable health insights.

**Proposed Methodology**

The project will follow a structured approach to analyze the dataset and answer the research questions:

* **Data Preprocessing:** Categorical variables will be encoded, and numerical features (e.g., Age, Height) will be standardized to ensure model compatibility.
* **Exploratory Data Analysis (EDA):** Initial analysis will examine feature distributions, correlations, and relationships with obesity levels to identify patterns and potential issues like multicollinearity.
* **Model Development:** Multiple machine learning models will be implemented for multi-class classification to choose the best fit model.

**Potential Challenges and Mitigation Strategies**

Several challenges may arise during the project, each with a planned mitigation strategy:

| **Challenge** | **Description** | **Mitigation Strategy** |
| --- | --- | --- |
| Synthetic Data | 77% of the data is synthetic, potentially limiting real-world applicability. | Focus on dataset-specific patterns and acknowledge generalizability limitations in findings. |
| Class Imbalance | Uneven distribution of obesity levels may skew model performance. | Apply class weighting to balance classes during model training. |
| Feature Correlation | High correlation between features may cause multicollinearity. | Conduct correlation analysis and use feature selection to reduce redundant features. |

Additional considerations include the dataset’s regional focus (Mexico, Peru, Colombia), which may limit applicability to other populations, and ethical concerns around obesity-related stigmatization, which will be addressed by framing results sensitively.

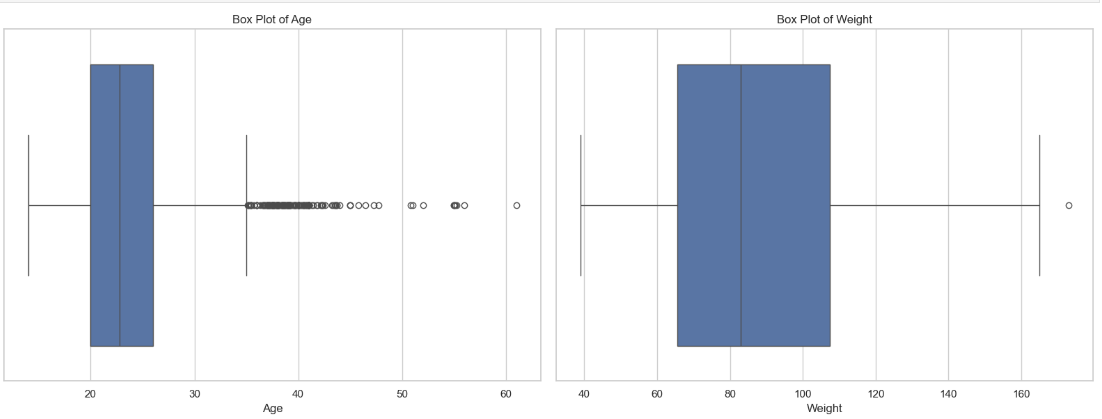
**Expected Outcomes**

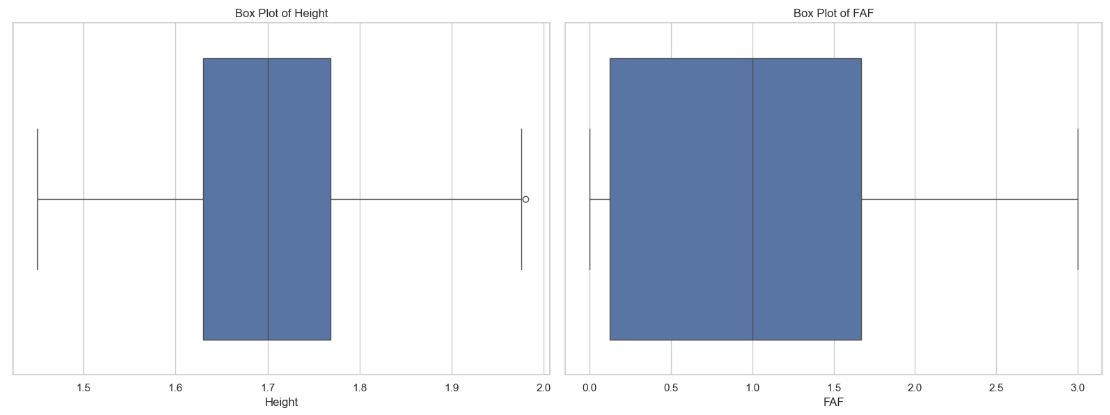
The project anticipates the following outcomes:

* **Identification of Key Factors:** Determining which demographic, dietary, or lifestyle factors most strongly predict obesity levels, potentially highlighting actionable areas like physical activity or dietary habits.
* **Accurate Predictive Models:** Developing machine learning models with high accuracy, comparable to studies achieving good accuracy.
* **Demographic Insights:** Uncovering variations in obesity levels by gender or age, informing targeted public health interventions.
* **Public Health Implications:** Providing insights that could guide personalized health recommendations and broader obesity prevention strategies.

These outcomes aim to contribute to the growing body of research on obesity prediction and support evidence-based health policies.

## **1.FINDING OUTLIERS:**

Let’s plot the box plot for finding outliers:  




* Box plot of a shows that there lies too much outlier in age column. So can not remove from this feature because this process can loss to much data from our data set.
* Weight feature contain just single outlier. So we have removed row related to that.
* Height also contain just single outlier. So we have also removed that one row.
* Physical activity have no outlier.

## **2. EDA:**

Let try to analyze the correlation of different other feature:

## Highly Correlated Feature Pairs:

1. **Height and Weight** (Correlation = **+0.46**)

Taller individuals tend to have higher body weight.

1. **Height and FAF (Physical Activity Frequency)** (Correlation = **+0.29**)

Taller individuals tend to engage more in physical activities.

1. **Height and NCP (Number of Main Meals Per Day)** (Correlation = **+0.24**)

Taller individuals may have slightly more main meals daily.

1. **Weight and FCVC (Vegetable Consumption Frequency)** (Correlation = **+0.21**)

Heavier individuals tend to eat vegetables slightly more frequently.

1. **CH2O (Water Intake) and FAF (Physical Activity Frequency)** (Correlation = **+0.16**)

Individuals who drink more water are somewhat more physically active.

## Most Negative (Inverse) Correlated Feature:

1. **Age and TUE (Technology Usage Time)** (Correlation = **-0.29**)

Older individuals spend less time using technological devices compared to younger individuals.

## Weakest (Almost No) Correlations:

1. **Age and Height** (Correlation = **-0.02**)

No meaningful relationship between a person's age and their height.

1. **FCVC (Vegetable Consumption) and FAF (Physical Activity)** (Correlation = **+0.02**)

No strong relation between vegetable eating habits and physical activity frequency.

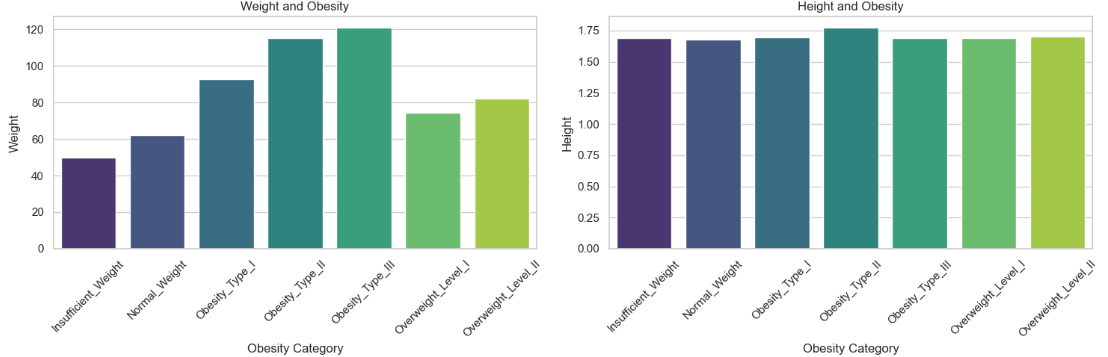
1. **NCP (Number of Meals) and TUE (Technology Usage)** (Correlation = **+0.04**)

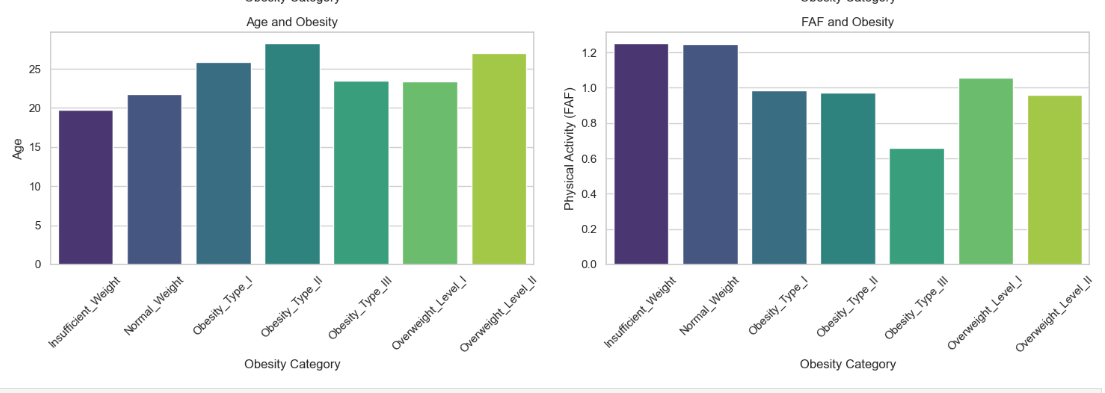
No strong relationship between how often someone eats and their technology usage time.

**Relationship Between Different Features And Obesity:**

In this section we will try to analyze the relation and effect of other features on target variable.

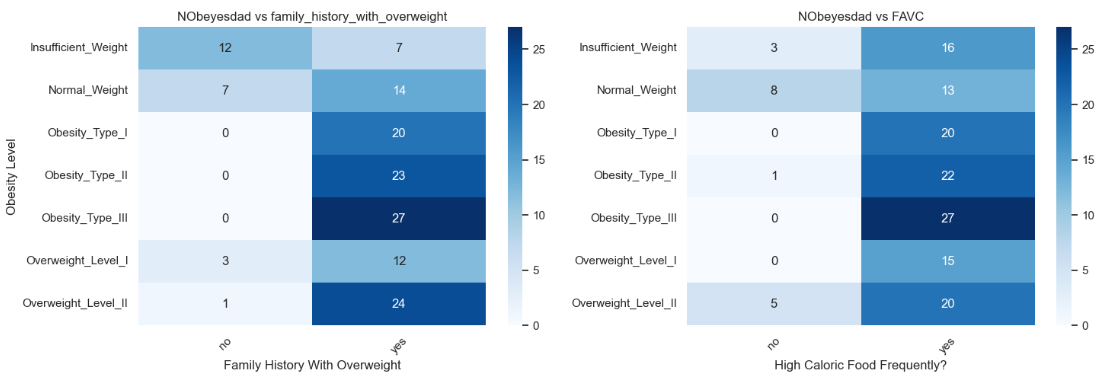
For that first we will plot bar graph on different features to analyze the impact of those features on obesity.

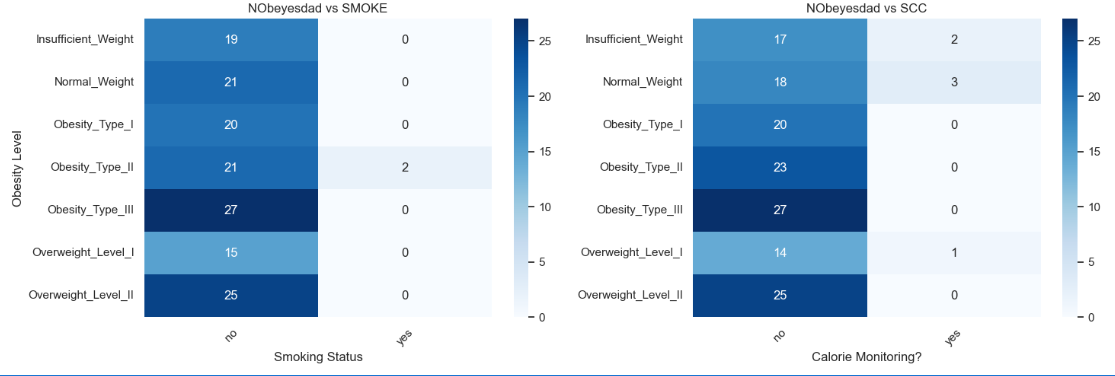




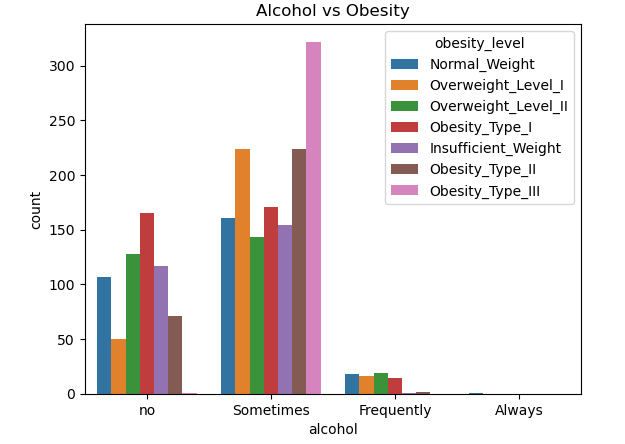
1. **Weight and Obesity**  
   As obesity category increases from **Insufficient Weight** to **Obesity\_Type\_III**, weight also increases significantly. Weight then decreases again slightly in **Overweight\_Level\_I** and **II**, but remains higher than normal weight categories.
2. **Height and Obesity**  
   Height remains relatively constant across all obesity categories with only minor fluctuations. There is no strong relationship between height and obesity in this dataset.
3. **Age and Obesity**  
   Age tends to increase with higher obesity categories up to **Obesity\_Type\_II**, but decreases slightly in **Obesity\_Type\_III** and **Overweight\_Level\_I** before rising again in **Overweight\_Level\_II**. Overall, higher obesity levels are associated with middle to older age groups.
4. **FAF (Physical Activity) and Obesity**  
   Higher physical activity levels (FAF) are associated with **Insufficient Weight** and **Normal Weight**. As obesity level increases, physical activity generally decreases, showing lower activity in **Obesity\_Type\_II** and **III**, but a slight increase again in **Overweight\_Level\_I**.

Now we will analyze the binary features to check their relation with obesity.





1. **NObeyesdad vs family\_history\_with\_overweight**  
   Individuals with a family history of overweight ("yes") have higher counts in obesity categories, especially **Obesity\_Type\_III** and **Overweight\_Level\_II**. Lack of family history shows higher counts in lower obesity levels or insufficient weight.
2. **NObeyesdad vs FAVC (High Caloric Food Frequently?)**  
   People who frequently consume high-calorie food ("yes") are concentrated more in higher obesity types (Obesity\_Type\_II, III). Those who do not consume high-calorie food regularly ("no") are more represented in Insufficient\_Weight and Obesity\_Type\_I.
3. **NObeyesdad vs SMOKE (Smoking Status)**  
   Both smokers and non-smokers show similar distribution across obesity levels, but non-smokers ("no") slightly dominate higher obesity categories (Obesity\_Type\_III, Overweight\_Level\_II). Smoking doesn't appear to drastically skew obesity levels.
4. **NObeyesdad vs SCC (Calorie Monitoring?)**  
   Individuals who monitor their calories ("yes") show slightly higher counts in lower obesity levels and insufficient weight. Those who do not monitor calories ("no") have greater numbers in higher obesity types, especially Obesity\_Type\_III and Overweight\_Level\_II.

Now analyzing obesity vs Alcohol  
**Graph:**  


**Description:**

 The grouped bar chart visualizes the distribution of **obesity levels** across different **alcohol consumption categories**.

 The **x-axis** represents alcohol consumption categories (e.g., *low*, *medium*, *high*).

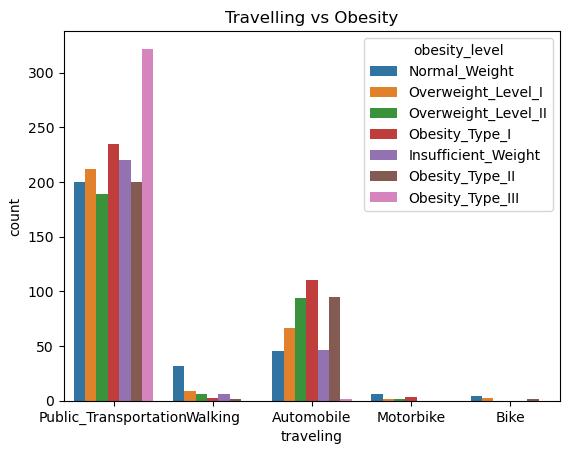
 The **y-axis** shows the count (frequency) of individuals within each alcohol category.

 Each bar group is divided by obesity categories (e.g., *non-obese*, *moderately obese*, *highly obese*), represented by different colors (legend).

 **Key observations**:

1. First it has been observed that most of the people are those who use alcohol sometime or do use it.
2. Those who not use lies in category of obesity i.e. insufficient, normal or obesity 1.
3. Those who use alcohol sometime lies in the category of obesity i.e. overweight, obesity level 2 and obesity level 3.

Now analyzing obesity vs Traveling  
**Graph:**



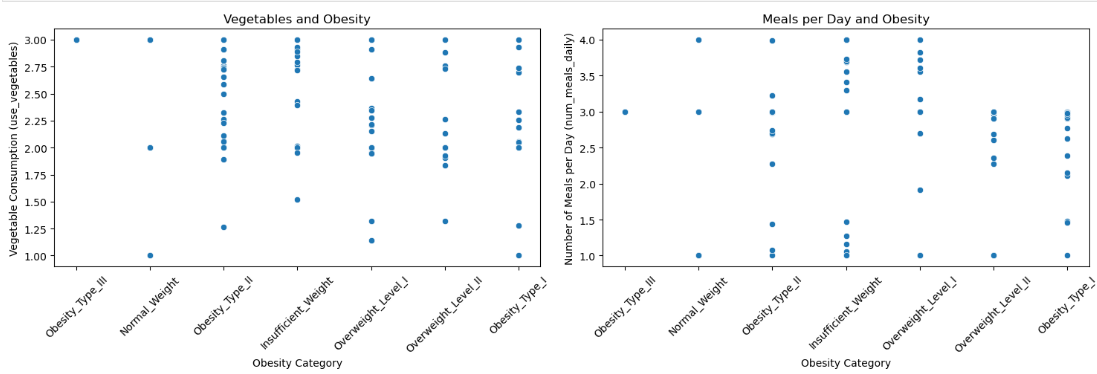
**Description:**

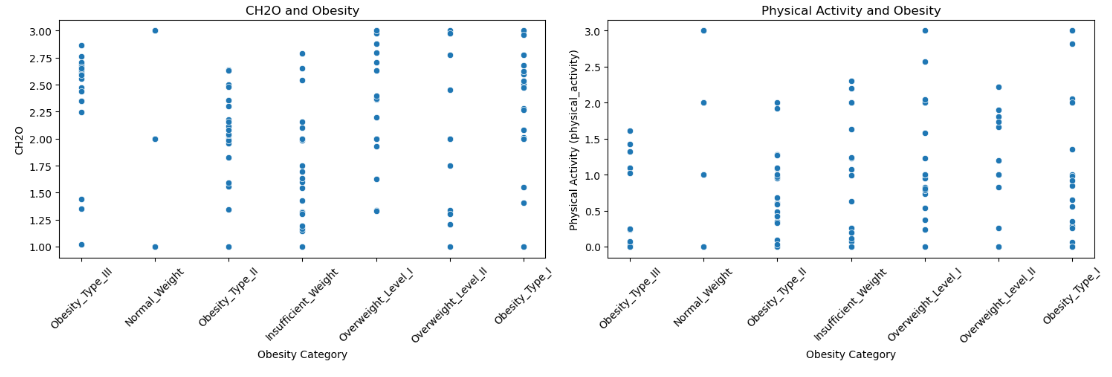
* The grouped bar chart visualizes the distribution of **obesity levels** across various **travel frequency categories**.
* The **x-axis** displays the categories of travel frequency (e.g., rarely travels, sometimes travels, frequently travels).
* The **y-axis** represents the count of individuals within each travel frequency category.
* Each bar group is subdivided into obesity categories (e.g., non-obese, moderately obese, highly obese), with different colors representing each obesity level as indicated in the legend.

**Observation**

* + First observation is that this dataset who mostly uses public transport. Then contain those people who uses automobile and other catagories have very low data.
  + Those people who uses public transport are effected by high obesity i.e. obesity 3 at large level.
  + Those who walks have normal weight.
  + Those who uses their auto mobile also lies between obesity level1 to overweight2. It means that automobile usage also effecting the obesity of people positively.

Now visualizing and observing features having float datatype;





**Description:**

The figure comprises **four scatter plots**, each displaying the distribution of different lifestyle factors across various **obesity categories**.

#### **1. Vegetables Consumption and Obesity**

* This plot illustrates the relationship between **vegetable consumption frequency** (on a 1–3 scale) and **obesity categories**.
* Vegetable consumption levels are fairly **evenly spread** across all obesity categories, indicating **no strong pattern** associating vegetable intake directly with obesity status.
* Individuals with **higher vegetable consumption (level 3)** appear across all obesity groups but people of obesity level 3 and overweight 1 are more then other categories.

#### **2. Meals per Day and Obesity**

* This plot shows the association between the **number of meals consumed per day** and **obesity categories**.
* Individuals consuming **3 or more meals per day** are present across overweight or insufficient weight.
* Conversely, individuals in the **normal weight** or **underweight** categories show **more variation**, with some consuming fewer meals (1–2 meals/day).

#### **3. CH2O (Carbohydrate Intake) and Obesity**

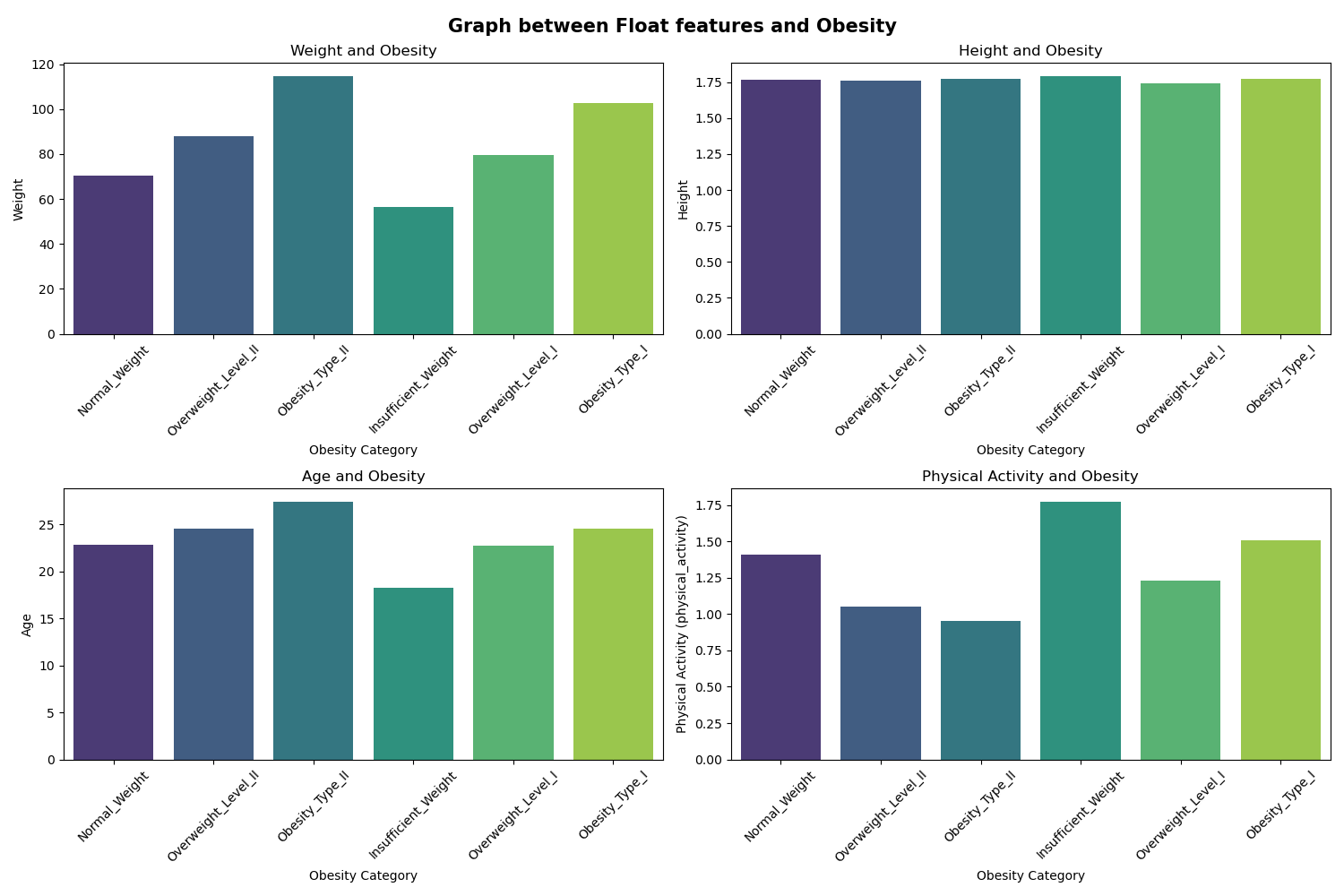
* This plot visualizes **carbohydrate consumption levels** (on a 1–3 scale) in relation to **obesity categories**.
* Carbohydrate intake appears **distributed across all obesity levels**, with **no clear concentration** of low or high carbohydrate intake in a specific obesity group.
* Similar to vegetable consumption, **carbohydrate intake** alone does not show an obvious pattern differentiating obesity categories.

#### **4. Physical Activity and Obesity**

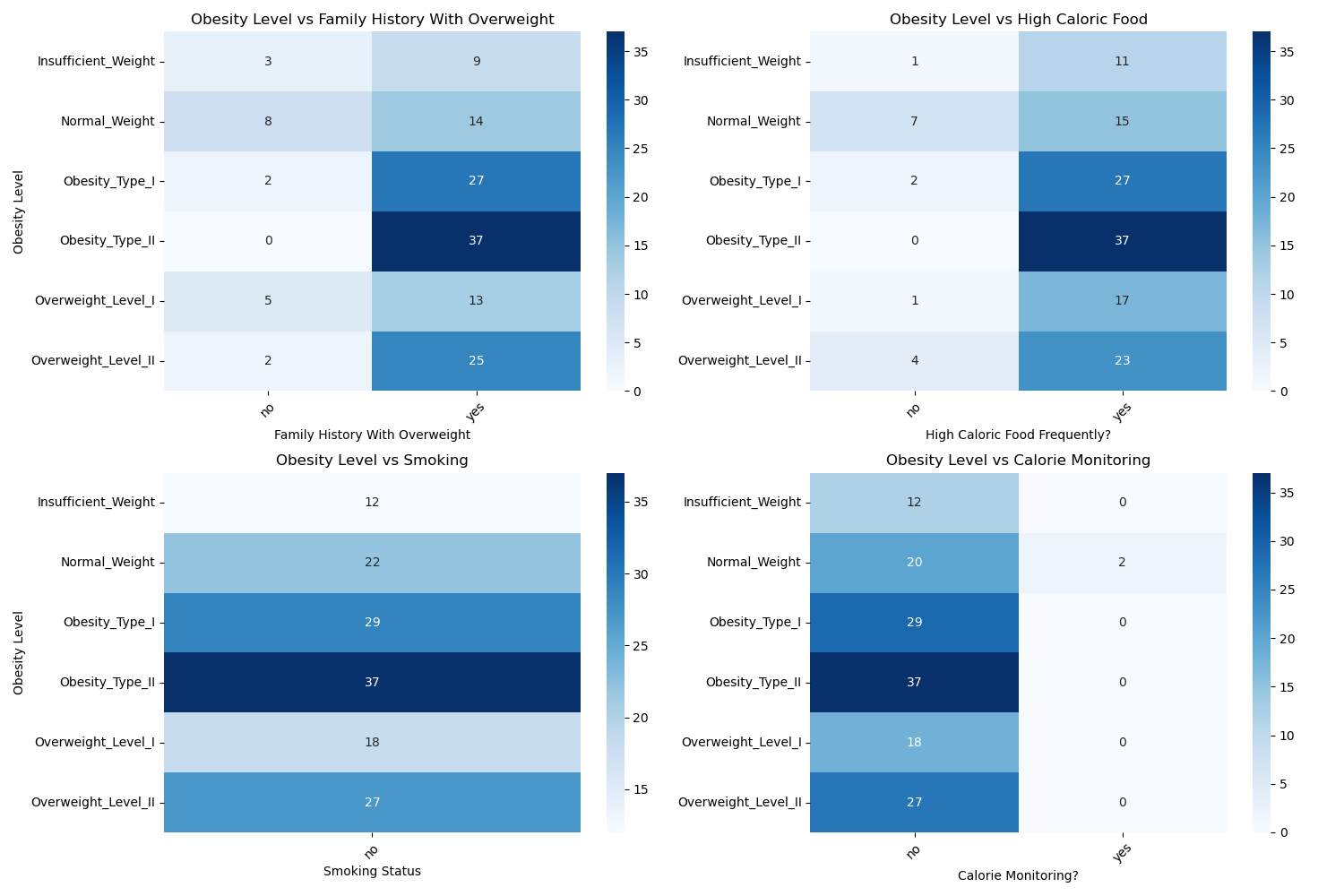
* This plot depicts the relationship between **physical activity frequency (days per week)** and **obesity categories**.
* Individuals engaging in **no or low physical activity (0–1 days/week)** are more commonly associated with **higher obesity levels** (Overweight and Obese categories).
* Conversely, individuals with **higher physical activity levels (2–3 days/week)** appear more frequently in the **normal weight** and **underweight** categories.
* This suggests a **negative association** between physical activity frequency and obesity — **higher physical activity** is linked with **lower obesity prevalence**.

**Male’s Data Visualization and Analysis:**

Bar Graph between some floating features and obesity levels:

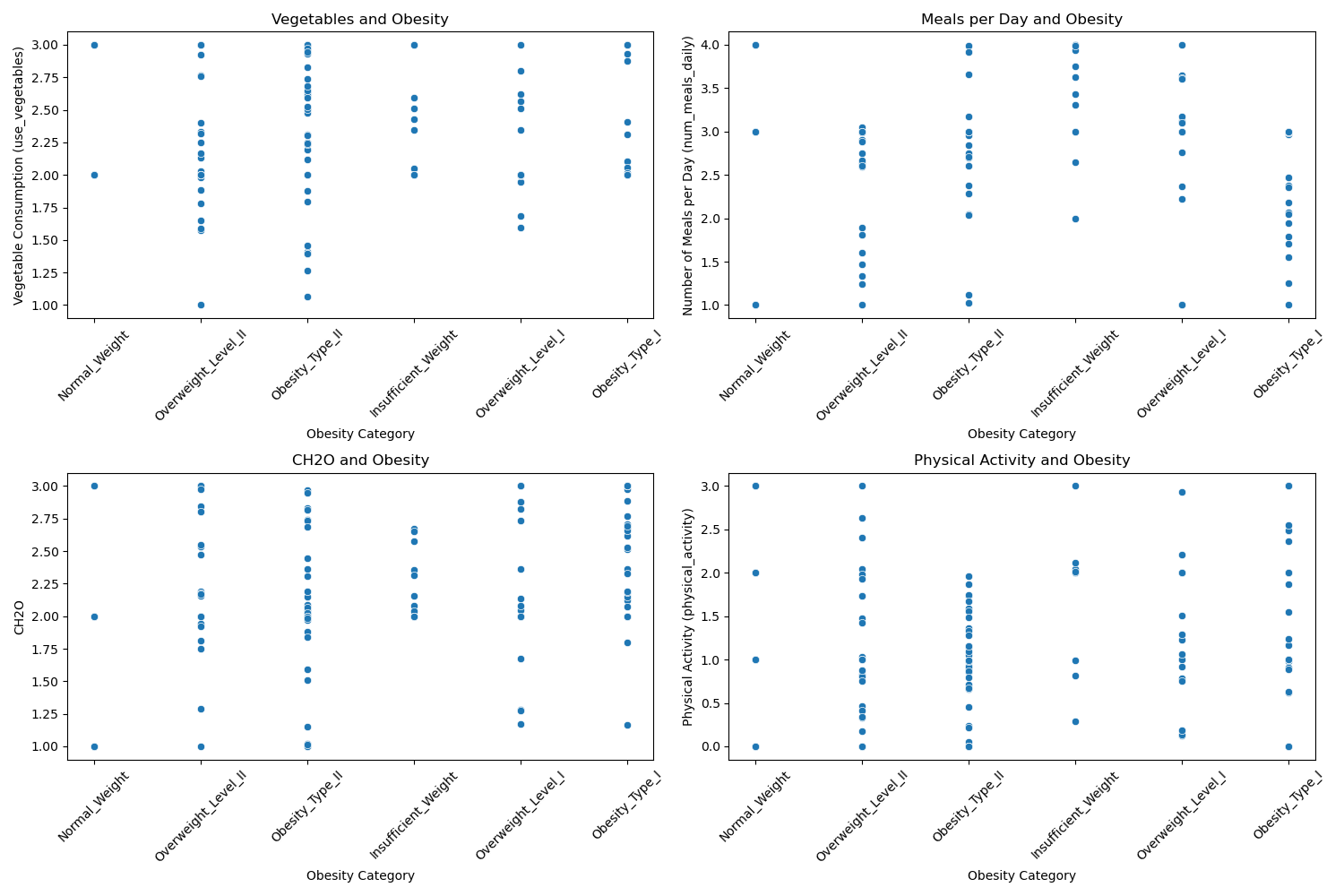
**Graph:**  
  
**Observations:**

* In weight and obesity graph it can be analyzed that people having weight more than 60 are the victim of the obesity or over weight.
* Height and obesity plot shows that there is no effective relation between height and obesity.
* Age and obesity graph clearly shows that people under 18 years have insufficient weight.
* Physical activity and obesity graph visualize that the people who do more physical activities are healthy i.e. insufficient weight or normal in weight.

**Heat plots between Binary Features and Obesity types feature:**  
  


**Observation:**

* The plot of **obesity vs family history of overweight** initially shows that most of the people in this data set are the people who family history shows that their family have some over weighted people. Secondly, plot shows that people who family member has family history ‘yes’ are also victim of obesity or overweight.
* **Obesity vs High chloric food graph** shows that people who user high chloric foods are over weighted people or the victim of obesity.
* **Obesity vs Smoking** plot shows that people of this data set does not smoke.
* **Obesity vs Calories graph** shows that people who do not monitor calories are over weighted people or the victim of obesity.

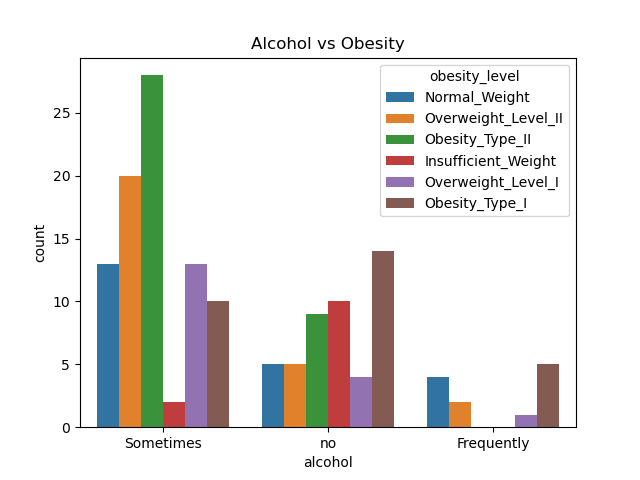


**Description:**  
 **Vegetables and Obesity**: People in the higher obesity categories tend to have more variation in vegetable consumption, indicating inconsistent healthy eating habits among obese individuals.

 **Meals per Day and Obesity**: Individuals in obese categories (especially Obesity\_Type\_III) tend to have more meals per day, suggesting a possible link between frequent eating and obesity.

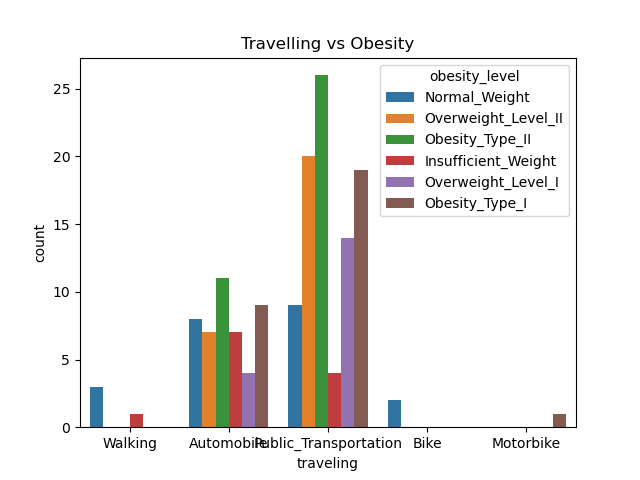
 **CH2O and Obesity**: The carbohydrate (CH2O) intake appears fairly distributed across all obesity levels, implying that CH2O intake alone may not directly correlate with obesity status in this dataset.

 **Physical Activity and Obesity**: Higher obesity levels are associated with a concentration of individuals having low physical activity, showing an inverse relationship between physical activity and obesity.



**Observations:**

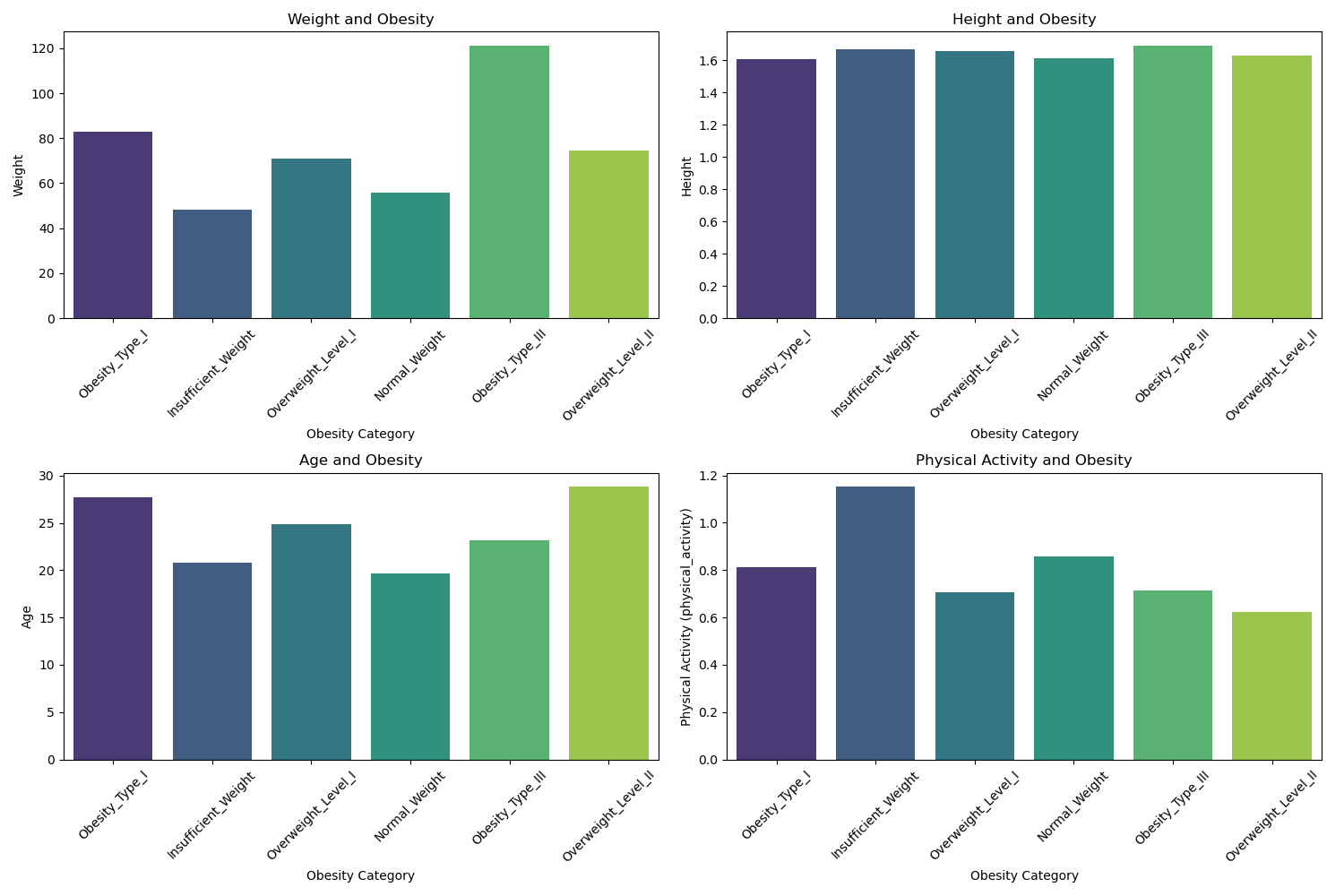
* This graph shows that majority of people drinks alcohol some time or never. Some people uses alcohol frequently.
* People who drinks alcohol sometime have obesity type 2 or overweight. But people with insufficient weight are too less than other categories.
* people who never uses alcohol have insufficient weight or lies in obesity level 1 or 2.
* Frequently use of alcohol have no greater impact on the obesity.



**Description:**

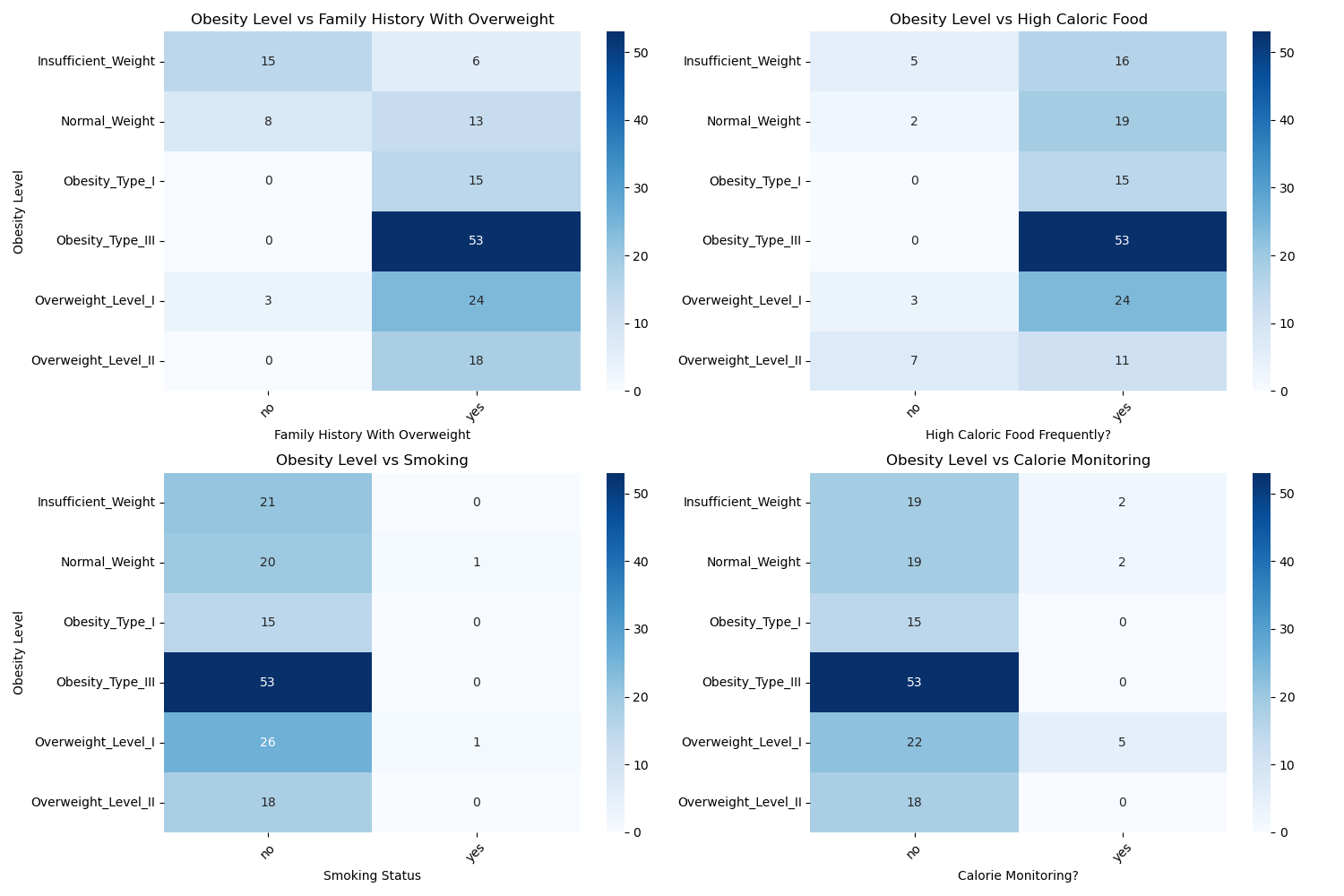
* People who walks are heathy means that they are normal in weight or insufficient weight.
* people who uses their automobiles are uses public transport are victim of obesity or overweight.
* People who uses their bikes are also normal in weight.

**Female’s Data Visualization and Analysis:**

****

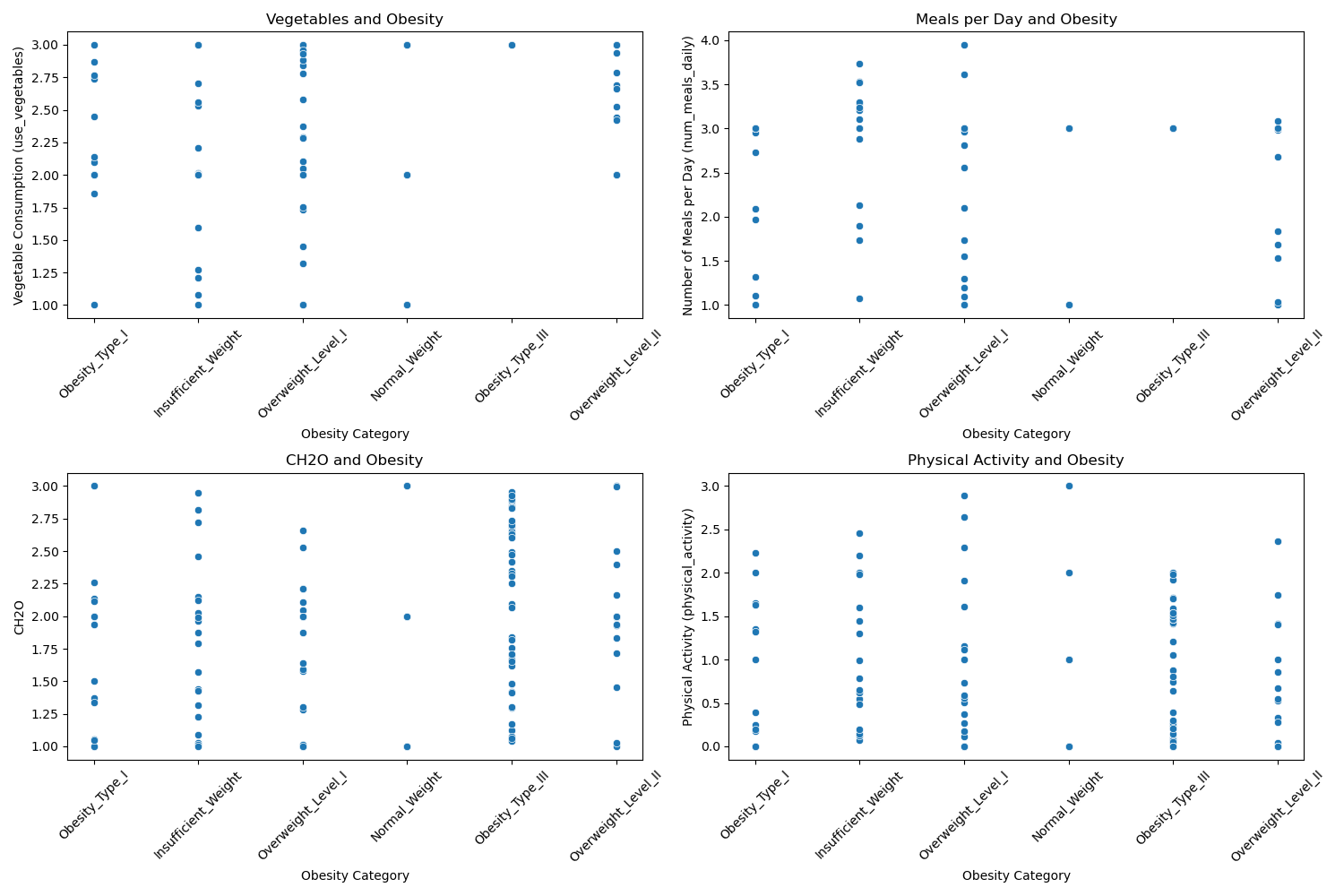
**Observations:**

* In **weight and obesity** graph it can be analyzed that females having weight more than 60 are the victim of the obesity or over weight.
* **Height and obesity** plot shows that there is no effective relation between height and obesity.
* **Age and obesity** graph clearly shows that females under 20 years have insufficient weight or normal weight.
* **Physical activity and obesity** graph visualize that the females who do more physical activities are healthy i.e. insufficient weight.

****

**Observations:**

* The plot of **obesity vs family history of overweight** initially shows that most of the females in this data set are the people who family history shows that their family have some over weighted females. Secondly, plot shows that females who family member has family history ‘yes’ are mostly victim of obesity level 3.
* **Obesity vs High chloric food graph** shows that females who uses high chloric foods are over weighted people or the victim of obesity level 3.
* **Obesity vs Smoking** plot shows that females who do not smoke mostly lies in the category of obesity level 3 and after that in overweight level 1.
* **Obesity vs Calories graph** shows that females who do not monitor calories are the victim of obesity level 3.

****

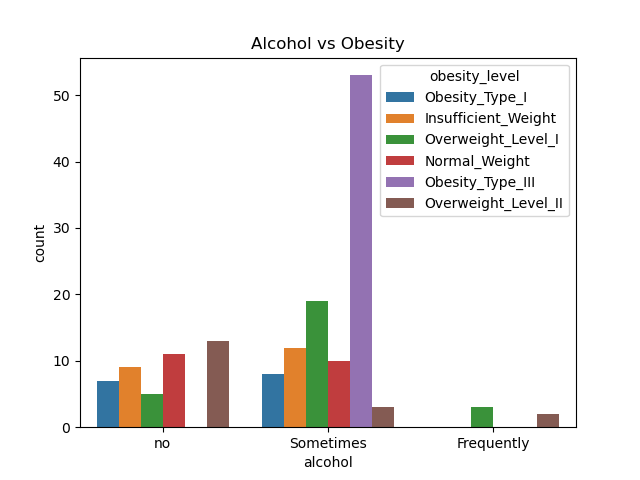
**Observations:**

 **Vegetables and Obesity**: There is no clear pattern indicating a strong relationship between the frequency of vegetable consumption and obesity category, as individuals across all obesity levels seem to consume vegetables at varying frequencies.

 **Meals per Day and Obesity**: People with higher obesity levels tend to have fewer meals per day, suggesting a possible negative correlation between number of meals and obesity severity.

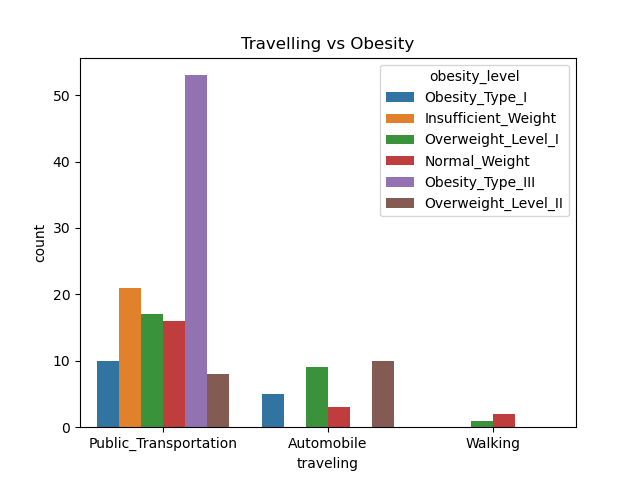
 **CH2O and Obesity**: The amount of carbohydrate consumption (CH2O) appears to be spread across all obesity categories, with no obvious trend, indicating that CH2O intake may not be a strong standalone predictor of obesity.

 **Physical Activity and Obesity**: Individuals with lower obesity levels seem to engage more frequently in physical activity, while higher obesity levels correlate with lower physical activity, suggesting a negative relationship between physical activity and obesity.

****

**Observations:**

* Female who do not uses alcohol have no strong relation with obesity.
* Females who uses alcohol sometimes are the victim of obesity i.e. they are over weighted.
* Other categories of alcohol usage have no effective impact on obesity.

****

**Observations:**

* Most of the females use public transport.
* Public transport has a greater impact on the obesity of females. Mostly female who usepublic transport lies in the category of obesity level 3.
* Plot shows that over weighted females uses automobiles more than others.