# A Data-Driven Analysis of Factors Influencing Obesity

By Evalyn Njagi

# Table of Contents

## INTRODUCTION:

Obesity has emerged as a significant health challenge, impacting a large portion of the population in Mexico, Peru, and Colombia. Various factors, such as genetic predisposition, the frequency of consuming junk food, regularity of physical exercise, and water intake, contribute to the prevalence of obesity. To better understand these contributing factors, an analysis was conducted on a dataset comprising health and dietary information from individuals in these countries. The primary objective of this study is to estimate obesity levels based on physical conditions and eating habits, utilizing Python for Data Science to uncover actionable insights.

## EXECUTIVE SUMMARY:

This report outlines the findings from an exploratory data analysis (EDA) and machine learning model implementation aimed at understanding and predicting obesity levels based on various lifestyle, demographic, and health factors. Key insights, data preparation steps, and predictive model performance are highlighted to inform stakeholders of actionable outcomes and strategic recommendations.

## DATA DESCRIPTION:

This dataset includes data to estimate obesity levels, with 17 attributes and 2111 records.

### Key Features:

1. **Gender** (Categorical)
2. **NObeyesdad** (Categorical): Insufficient Weight, Normal Weight, Over Weight Type1, Over Weight Type2, Obesity Type 1, Obesity Type 2, Obesity Type 3.
3. **Age** (Continuous): Age of the individual.
4. **Height** (Continuous): Height in meters.
5. **Weight** (Continuous): Weight in kilograms.
6. **family_history_with_overweight** (Binary): Whether the individual has a family member suffering from overweight.
7. **FAVC** (Binary): Whether the individual eats high-calorie food frequently.
8. **FCVC** (Continuous): Frequency of vegetable consumption in meals.
9. **NCP** (Continuous): Number of main meals consumed daily.
10. **CAEC** (Categorical): Food consumption between meal.
11. **SMOKE** (Binary): Whether the individual smokes.
12. **CH2O** (Continuous): Amount of water consumed daily.
13. **SCC** (Binary): Whether the individual monitors calorie intake
14. **FAF** (Continuous): Frequency of physical activity.
15. **TUE** (Continuous): Time spent using technological devices
16. **CALC** (Categorical): Alcohol consumption frequency.
17. **MTRANS** (Categorical): Mode of transportation used

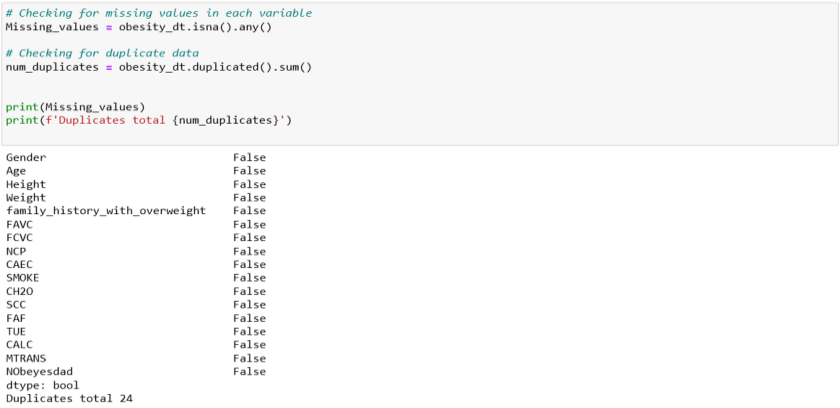
## CODE LINK:

https://github.com/EvalynTheAnalyst/Obesity-Level-Analysis/blob/main/Obesity%20Python%20Project%20code1.pdf

# DATA CLEANING:

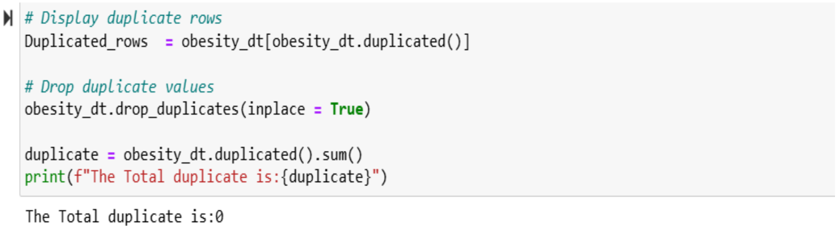
## 1) Data Integrity:

- No missing values were identified.
- Duplicate rows were removed to ensure data accuracy.



```
# Checking for missing values in each variable
Missing_values = obesity_dt.isna().any()

# Checking for duplicate data
num_duplicates = obesity_dt.duplicated().sum()


print(Missing_values)
print(f'Duplicates total {num_duplicates}')
```

```
Gender                            False
Age                               False
Height                            False
Weight                            False
family_history_with_overweight    False
FAVC                              False
FCVC                              False
NCP                               False
CAEC                              False
SMOKE                             False
CH2O                              False
SCC                               False
FAF                               False
TUE                               False
CALC                              False
MTRANS                            False
NObeyesdad                        False
dtype: bool
Duplicates total 24
```

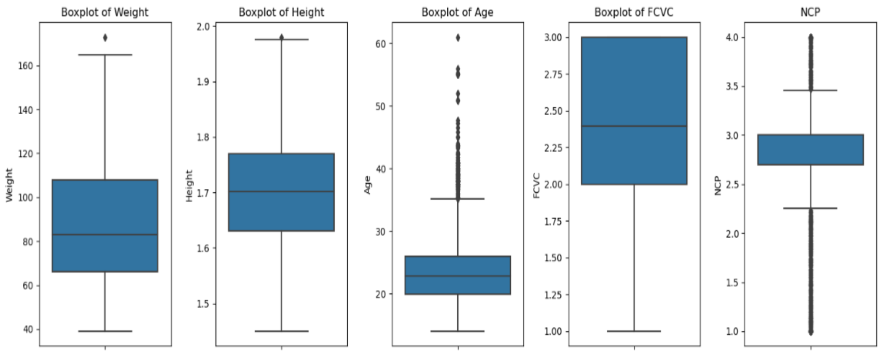


```
# Display duplicate rows
Duplicated_rows  = obesity_dt[obesity_dt.duplicated()]

# Drop duplicate values
obesity_dt.drop_duplicates(inplace = True)

duplicate = obesity_dt.duplicated().sum()
print(f"The Total duplicate is:{duplicate}")
```

```
The Total duplicate is:0
```

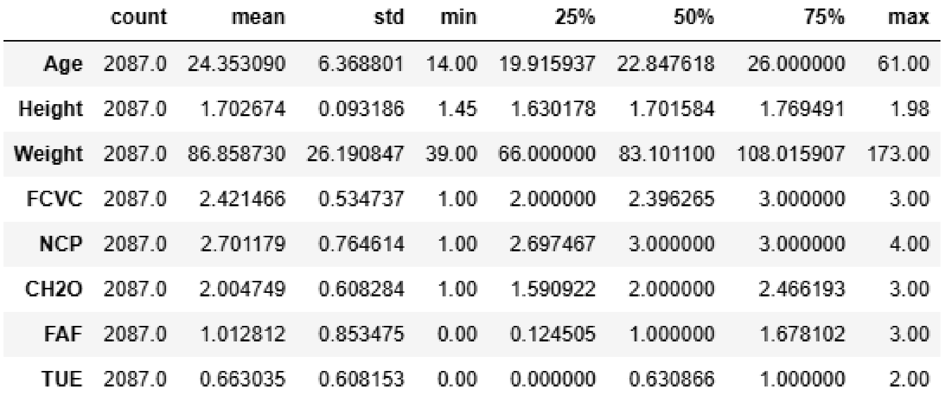
## 2) Outlier Detection and Handling

- Boxplots revealed outliers in features like Weight and Height.
- The outliers for age and NCP would not affect the analysis thus were not altered.
- Outliers were capped using the 1st and 99th percentiles.



# EXPLARATORY DATA ANALYSIS(EDA)

## Summary statistics

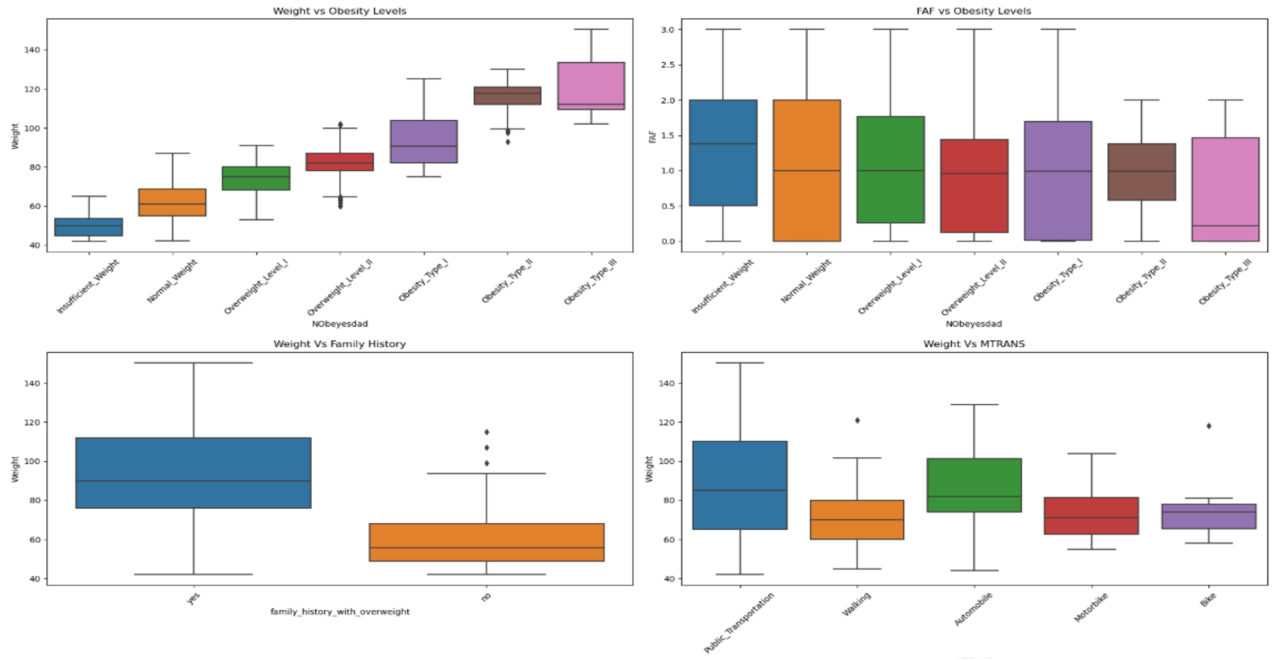


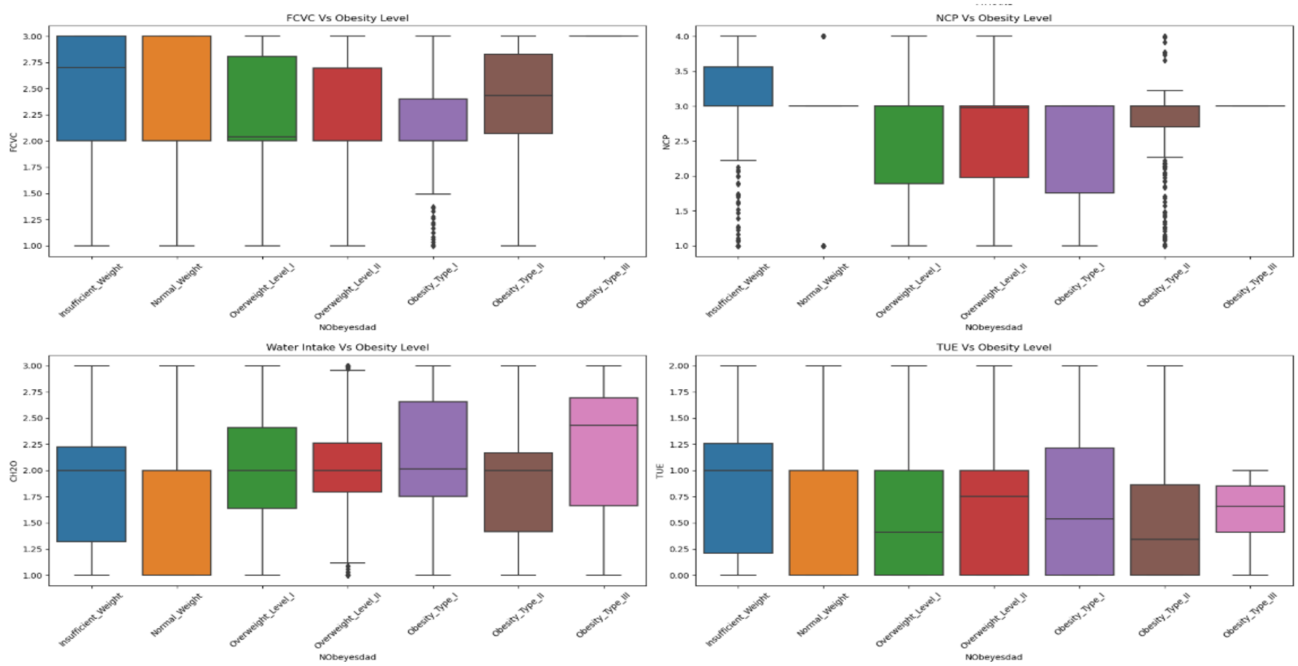
|  | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| **Age** | 2087.0 | 24.353090 | 6.368801 | 14.00 | 19.915937 | 22.847618 | 26.000000 | 61.00 |
| **Height** | 2087.0 | 1.702674 | 0.093186 | 1.45 | 1.630178 | 1.701584 | 1.769491 | 1.98 |
| **Weight** | 2087.0 | 86.858730 | 26.190847 | 39.00 | 66.000000 | 83.101100 | 108.015907 | 173.00 |
| **FCVC** | 2087.0 | 2.421466 | 0.534737 | 1.00 | 2.000000 | 2.396265 | 3.000000 | 3.00 |
| **NCP** | 2087.0 | 2.701179 | 0.764614 | 1.00 | 2.697467 | 3.000000 | 3.000000 | 4.00 |
| **CH2O** | 2087.0 | 2.004749 | 0.608284 | 1.00 | 1.590922 | 2.000000 | 2.466193 | 3.00 |
| **FAF** | 2087.0 | 1.012812 | 0.853475 | 0.00 | 0.124505 | 1.000000 | 1.678102 | 3.00 |
| **TUE** | 2087.0 | 0.663035 | 0.608153 | 0.00 | 0.000000 | 0.630866 | 1.000000 | 2.00 |

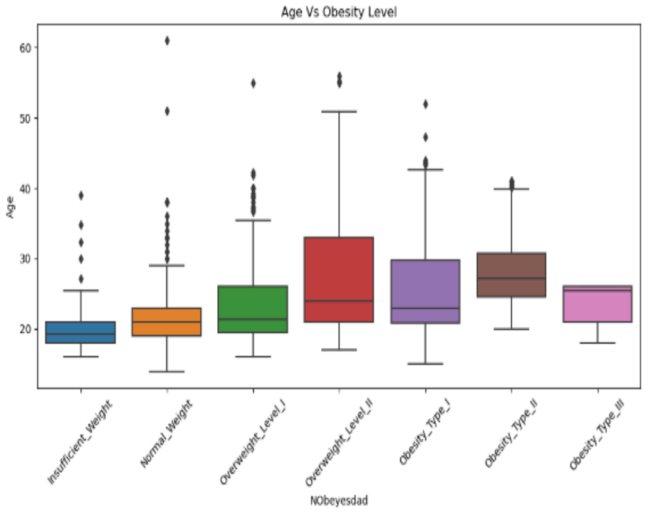
- Age - Mean: 24 years, Youngest: 14 years and Oldest: 61 years
- Height - Mean: 1.7meters
- Weight - Mean: 86 kg

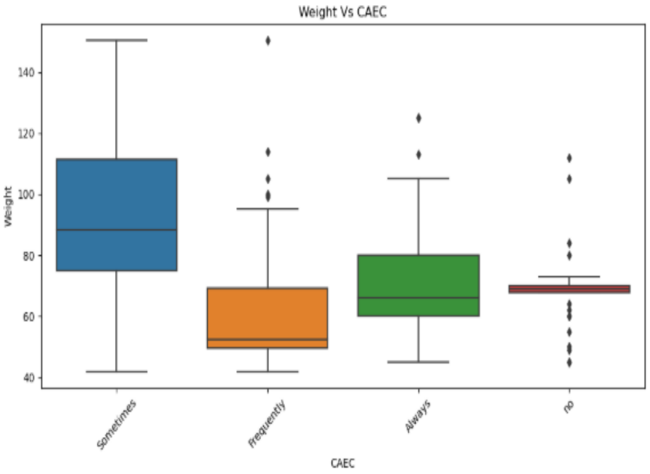
## Interactions between variables influencing obesity levels.

- Boxplots revealed significant variations in obesity levels based on family history, eating habits, and transportation modes.
- A clear upward trend in weight is associated with higher obesity levels.
- Those using public transport are at a risk of developing obesity than those walking or cycling.
- Lower FAF (exercise frequency) is slightly associated with higher obesity levels.
- Increased screen time correlates mildly with higher obesity levels and CAEC frequency, suggesting a sedentary lifestyle link.
- Food consumption between Meals (CAEC): Higher CAEC are linked to greater weight and possibly lower physical activity levels.
- Majority of those with high weight have association with family history of overweight.
- Water intake has no correlation with the obesity level.


Weight vs Obesity Levels
Weight
40
60
80
100
120
140
Insufficient_Weight
Normal_Weight
Overweight_Level_I
Overweight_Level_II
Obesity_Type_I
Obesity_Type_II
Obesity_Type_III
NObeyesdad
FAF vs Obesity Levels
FAF
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Insufficient_Weight
Normal_Weight
Overweight_Level_I
Overweight_Level_II
Obesity_Type_I
Obesity_Type_II
Obesity_Type_III
NObeyesdad
Weight Vs Family History
Weight
40
60
80
100
120
140
yes
no
family_history_with_overweight
Weight Vs MTRANS
Weight
40
60
80
100
120
140
Public_Transportation
Walking
Automobile
Motorbike
Bike


FCVC Vs Obesity Level
FCVC
1.00
1.25
1.50
1.75
2.00
2.25
2.50
2.75
3.00
Insufficient_Weight
Normal_Weight
Overweight_Level_I
Overweight_Level_II
Obesity_Type_I
Obesity_Type_II
Obesity_Type_III
NObeyesdad
NCP Vs Obesity Level
NCP
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Insufficient_Weight
Normal_Weight
Overweight_Level_I
Overweight_Level_II
Obesity_Type_I
Obesity_Type_II
Obesity_Type_III
NObeyesdad
Water Intake Vs Obesity Level
CH2O
1.00
1.25
1.50
1.75
2.00
2.25
2.50
2.75
3.00
Insufficient_Weight
Normal_Weight
Overweight_Level_I
Overweight_Level_II
Obesity_Type_I
Obesity_Type_II
Obesity_Type_III
NObeyesdad
TUE Vs Obesity Level
TUE
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
Insufficient_Weight
Normal_Weight
Overweight_Level_I
Overweight_Level_II
Obesity_Type_I
Obesity_Type_II
Obesity_Type_III
NObeyesdad

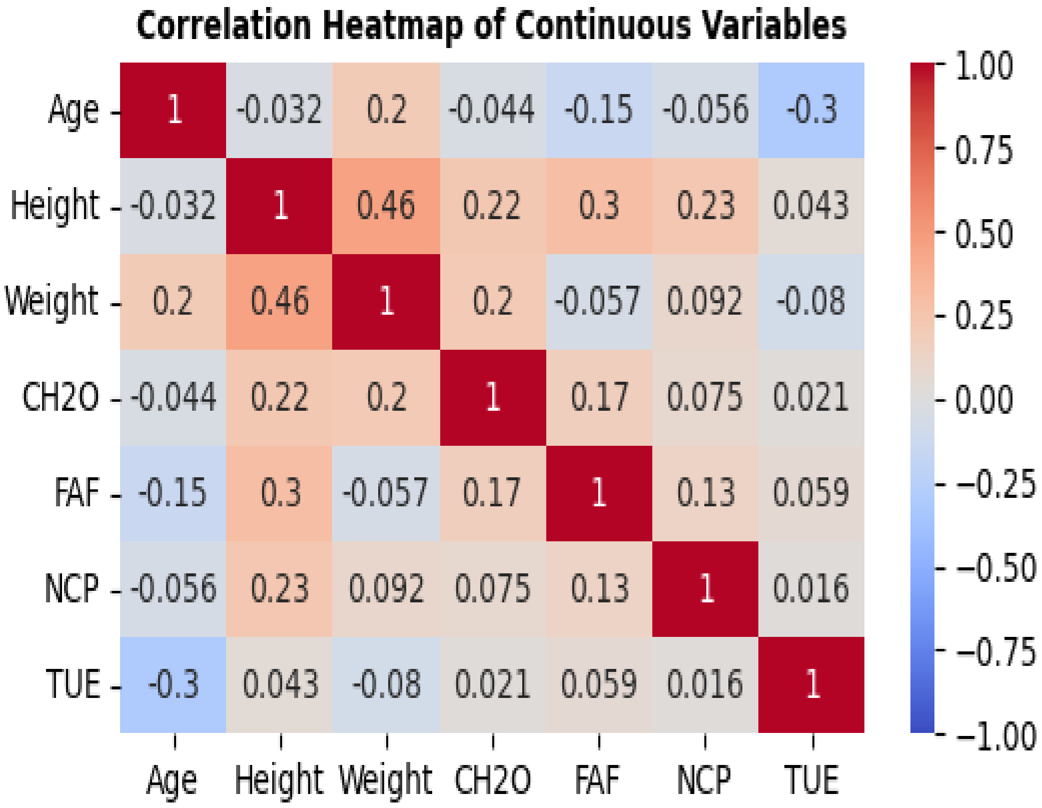




## Relationships between continuous variables:

1. **Diagonal Elements (Perfect Correlation)**:
    - The diagonal values are all **1** because each variable is perfectly correlated with itself.
2. **Positive Correlations**:
    - **Height and Weight (0.46)**:
        - There is a moderate positive correlation between height and weight, indicating that taller individuals tend to weigh more.
    - **Weight and Age (0.20)**:
        - A weak positive correlation suggests that weight slightly increases with age.
    - **Height and Age (0.22)**:
        - A weak positive correlation indicates that taller individuals might be slightly older, but this could be attributed to natural growth in younger populations.
    - **Height and CH2O (0.22)**:
        - A weak positive relationship between height and water consumption suggests that taller individuals might consume slightly more water.
    - **FAF and CH2O (0.17)**:
        - A positive correlation suggests that individuals engaging in frequent physical activity might also consume more water.
3. **Negative Correlations**:
    - **Age and TUE (-0.30)**:
        - A moderate negative correlation suggests that older individuals spend less time using electronic devices compared to younger people, likely due to lifestyle or generational differences.
    - **Age and FAF (-0.15)**:
        - A weak negative correlation indicates that older individuals tend to engage in less frequent physical activity.
    - **TUE and Weight (-0.08)**:
        - A very weak negative correlation suggests that time spent on electronic devices has a negligible association with weight.
    - **TUE and FAF (0.059)**:
        - An extremely weak positive correlation between screen time and physical activity frequency.

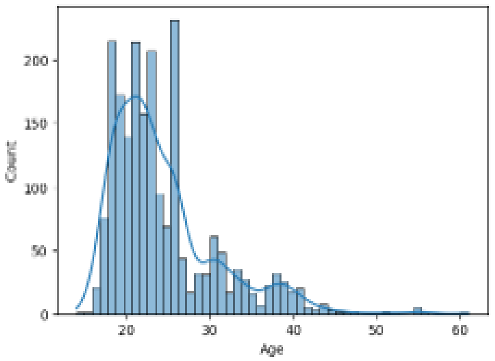
4. **Minimal Correlations (Near Zero)**:
   - Variables like **CH2O and Weight (0.20)** or **NCP and TUE (0.016)** have almost no meaningful correlation.
   - Many of these relationships suggest weak or no linear association between these variables.

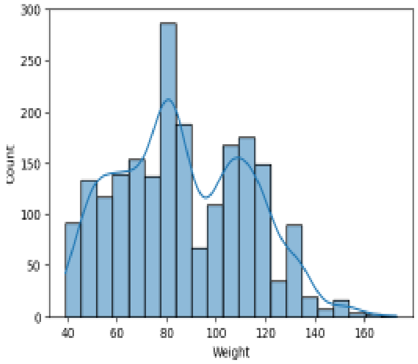


## Data Distribution:

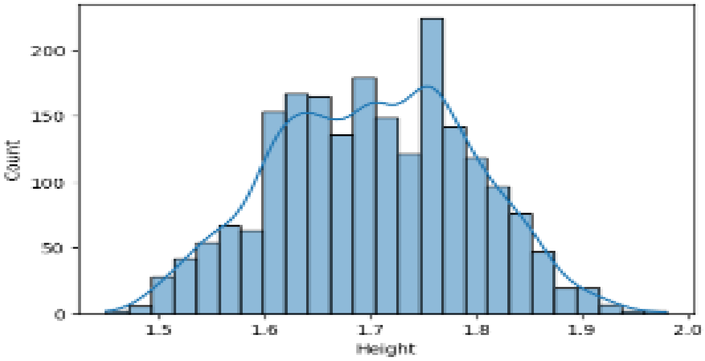
**Age**: The data is right skewed means that the distribution of ages has a longer tail on the right side. This indicates that the majority of the individuals in the dataset are younger, with fewer individuals falling into older age groups.

**Weight:** Most people weight range from 40 to 120 kg. Fewer have weight past 120kg thus the data is slightly right skewed.

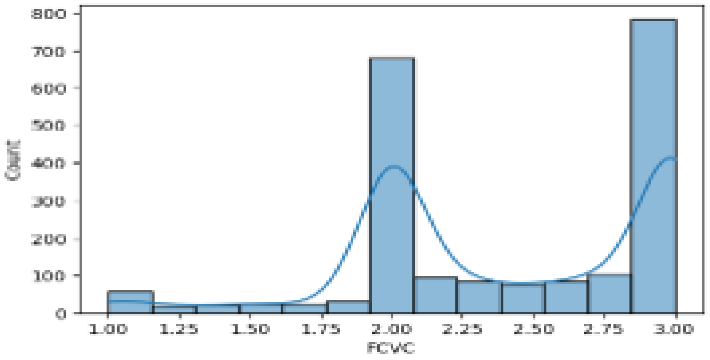




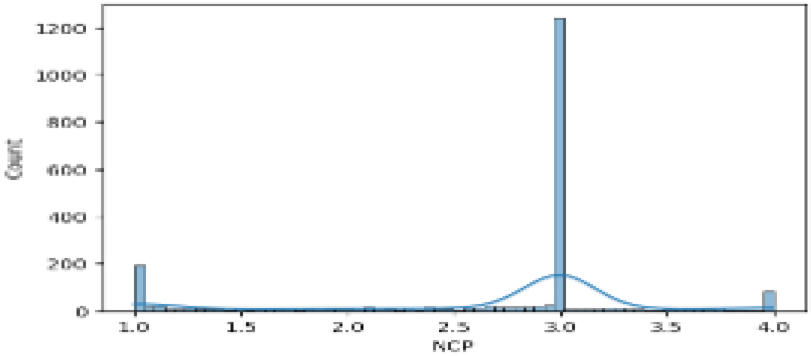
**Height:** The data is close to a normal distribution. It is slightly symmetric. About as many people are shorter than 1.7m (between 1.6m–1.7m) as there are taller (between 1.7m–1.8m).



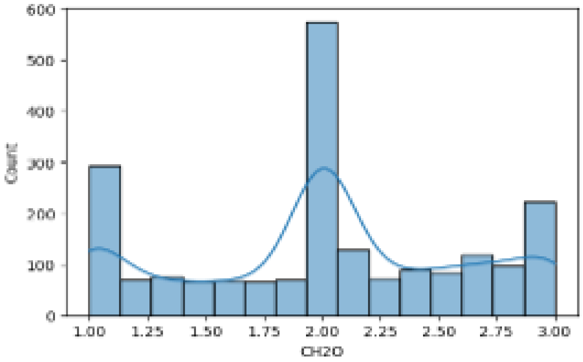
**FCV:** The data shows a multimodal distribution suggesting that individuals in the dataset tend to consume vegetables at specific frequencies values close to 2 and 3 times per meal.



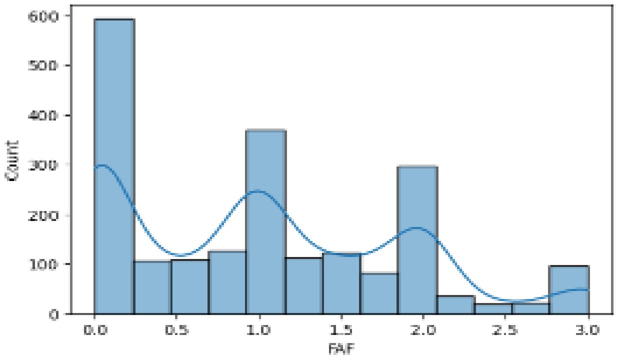
**NCP:** The data is Sparse distributed. Showing Most individual sticks to consuming the standard 3 days per meal. Only a few consumes 1,2 or more meals per day.



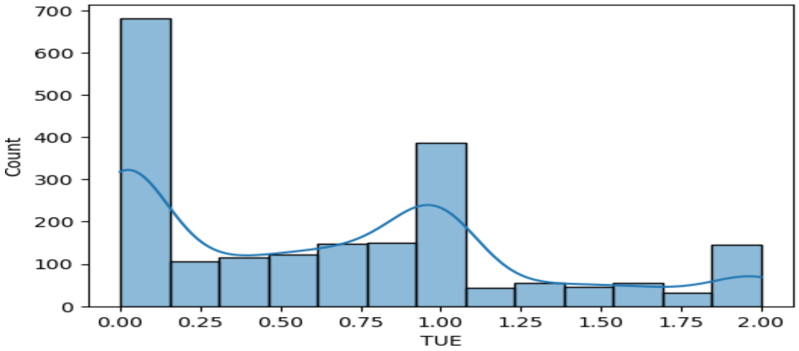
**CH2O:** The distribution is unimodal. Has one peak. Which means most people drunk 2litres of water per day. The rest was distributed between 1liter to 3 liters per day.



**FAF:** The distribution is multimodal and slightly right skewed as majority do exercise less than one hour per week.

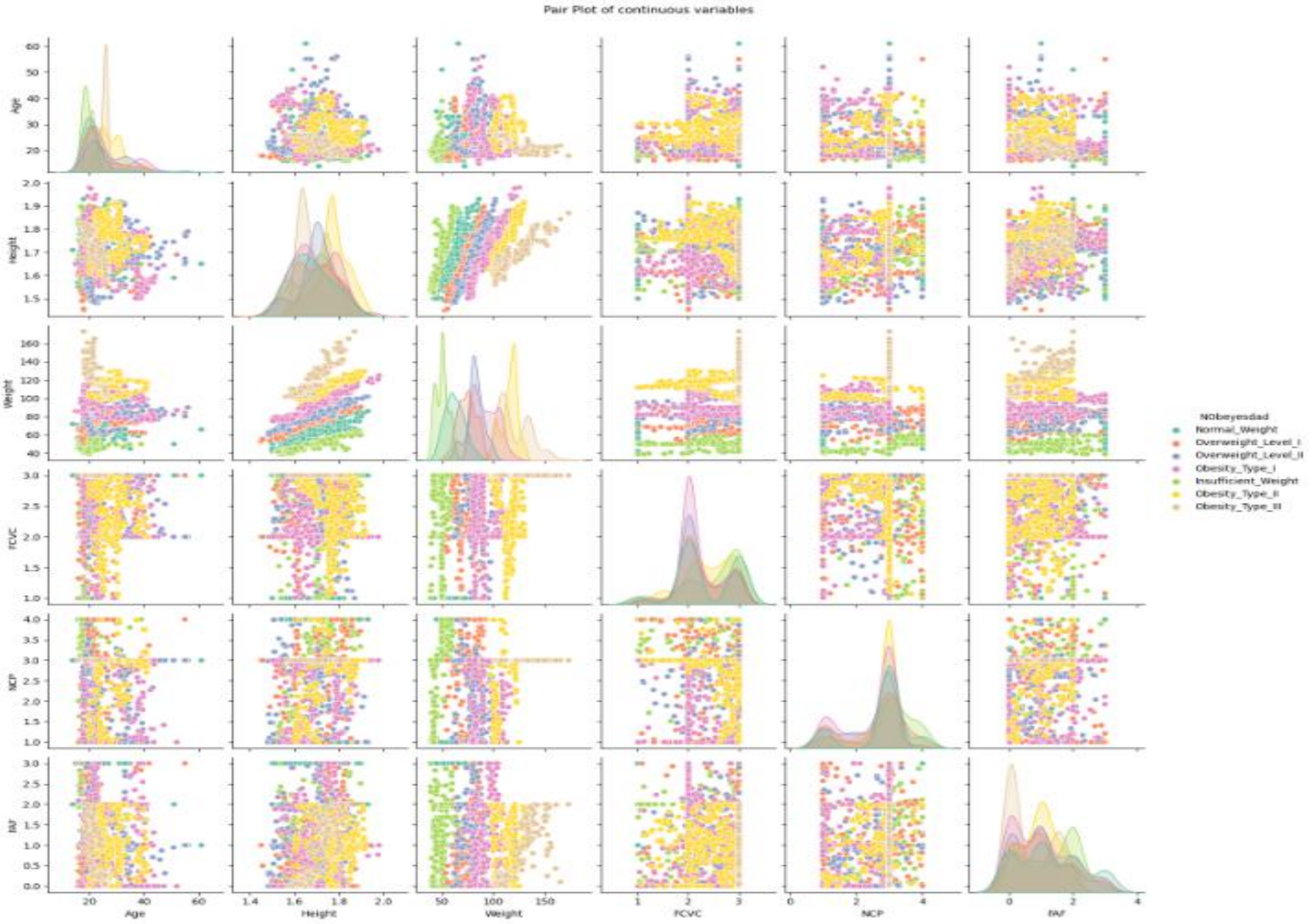


**TUE:** The distribution is unimodal and slightly right skewed. This means most people spent less time on technology devices as it is less than an hour. Only a few spend 2 to 3hours on their electronic gadget.



## Pair plot:

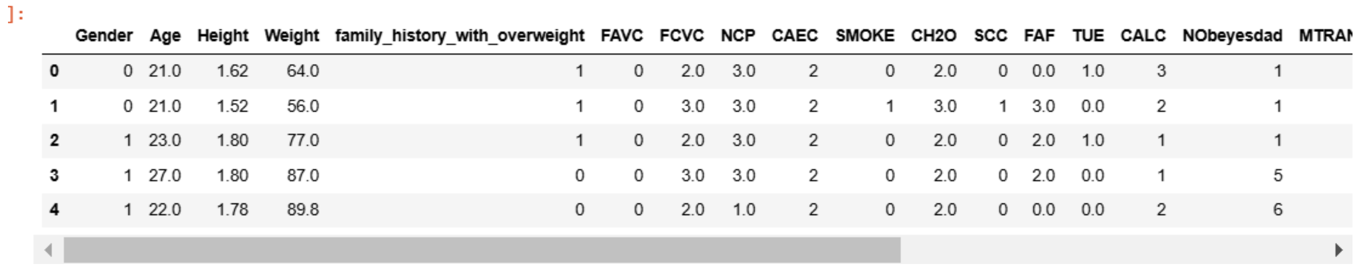
The pair plot reveals several insights about the dataset. Firstly, it demonstrates the distribution of variables like age and height, which appear somewhat normally distributed. Secondly, it highlights relationships between variables, such as the positive correlation between age and weight. Thirdly, the pair plot shows clustering of data points based on obesity categories, suggesting potential associations between specific obesity levels and variable ranges.



# MACHINE LEARNING ANALYSIS

## Data Preparation:

1. **Encoding**: Categorical variables were encoded using label and one-hot encoding techniques.
2. **Normalization**: Continuous features like Age, Weight, and Height were scaled using Min-Max normalization for consistent model input.
3. **Feature Engineering**: Polynomial features were generated to capture interactions, and feature selection was performed using chi-squared tests.



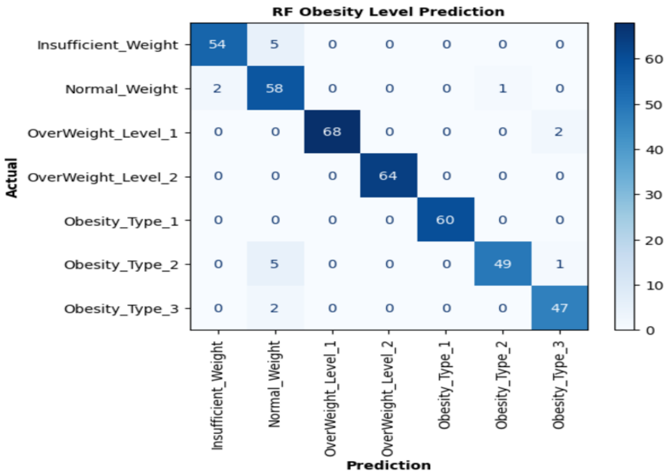
]:

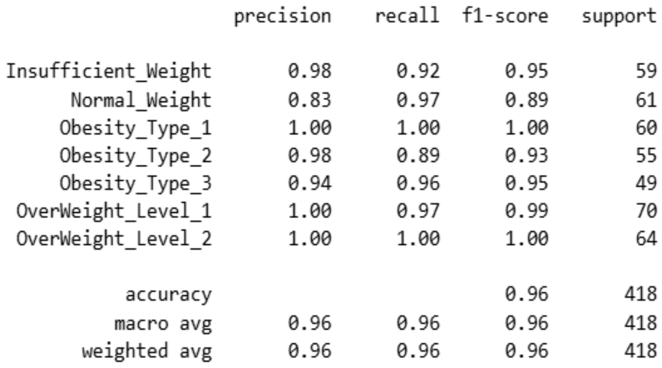
| | Gender | Age | Height | Weight | family_history_with_overweight | FAVC | FCVC | NCP | CAEC | SMOKE | CH2O | SCC | FAF | TUE | CALC | NObeyesdad | MTRA |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 0 | 21.0 | 1.62 | 64.0 | 1 | 0 | 2.0 | 3.0 | 2 | 0 | 2.0 | 0 | 0.0 | 1.0 | 3 | 1 | |
| **1** | 0 | 21.0 | 1.52 | 56.0 | 1 | 0 | 3.0 | 3.0 | 2 | 1 | 3.0 | 1 | 3.0 | 0.0 | 2 | 1 | |
| **2** | 1 | 23.0 | 1.80 | 77.0 | 1 | 0 | 2.0 | 3.0 | 2 | 0 | 2.0 | 0 | 2.0 | 1.0 | 1 | 1 | |
| **3** | 1 | 27.0 | 1.80 | 87.0 | 0 | 0 | 3.0 | 3.0 | 2 | 0 | 2.0 | 0 | 2.0 | 0.0 | 1 | 5 | |
| **4** | 1 | 22.0 | 1.78 | 89.8 | 0 | 0 | 2.0 | 1.0 | 2 | 0 | 2.0 | 0 | 0.0 | 0.0 | 2 | 6 | |

## Modelling Approach

### 1 Random Forest Classifier

- **Accuracy**: 95% on test data.
- **Recall**: 97% on test data.
- **F1 score: 95%** on test data.
- **Strengths**: Robust to overfitting, handled feature importance effectively.
- **Confusion Matrix Insights**: High precision and recall for predicting severe obesity levels (Type 2 and Type 3).

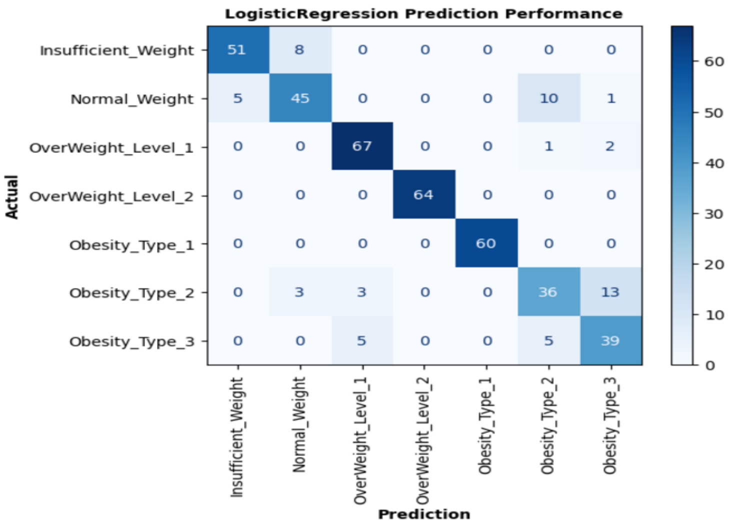


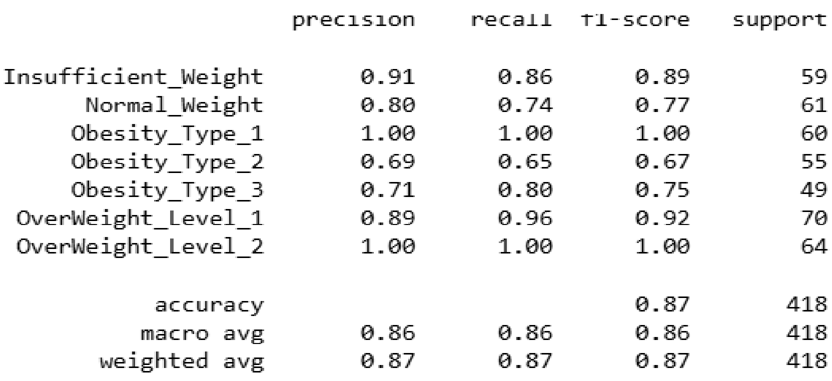


```
                    precision    recall  f1-score   support

 Insufficient_Weight       0.98      0.92      0.95        59
       Normal_Weight       0.83      0.97      0.89        61
      Obesity_Type_1       1.00      1.00      1.00        60
      Obesity_Type_2       0.98      0.89      0.93        55
      Obesity_Type_3       0.94      0.96      0.95        49
 OverWeight_Level_1       1.00      0.97      0.99        70
 OverWeight_Level_2       1.00      1.00      1.00        64

            accuracy                           0.96       418
           macro avg       0.96      0.96      0.96       418
        weighted avg       0.96      0.96      0.96       418
```

## 2. Logistic Regression

- **Accuracy**: 87% on test data.
- **Optimization**: Feature interactions improved accuracy to 87%.
- **Confusion Matrix Insights**: Struggled with intermediate obesity levels but showed high accuracy for normal weight predictions.





```
                    precision    recall  f1-score   support

 Insufficient_Weight       0.91      0.86      0.89        59
       Normal_Weight       0.80      0.74      0.77        61
      Obesity_Type_1       1.00      1.00      1.00        60
      Obesity_Type_2       0.69      0.65      0.67        55
      Obesity_Type_3       0.71      0.80      0.75        49
 OverWeight_Level_1       0.89      0.96      0.92        70
 OverWeight_Level_2       1.00      1.00      1.00        64

            accuracy                           0.87       418
           macro avg       0.86      0.86      0.86       418
        weighted avg       0.87      0.87      0.87       418
```

## 3. Comparative Performance

- Random Forest outperformed Logistic Regression in overall classification accuracy and handling of complex feature interactions.

## KEY FINDING

1. **Determinants of Obesity**
   - Sedentary lifestyles (low FAF) and poor dietary habits (high NCP, low FCVC) significantly increase obesity risk.
   - Individuals with a family history of obesity are more prone to higher obesity levels.
2. **Model Insights**
   - Random Forest’s high accuracy and detailed predictions make it a preferred choice for operational use.
   - Logistic Regression, when optimized, serves as a simpler, interpretable alternative for certain scenarios.

## RECOMMENDATIONS

1. **Data-Driven Interventions**
   - Targeted programs encouraging physical activity and healthy eating habits for high-risk groups.
   - Promote walking or cycling for short commutes to work instead of relying on public transportation or vehicles.
2. **Policy Formulation**
   - Design public health policies emphasizing early detection and intervention strategies for individuals with obesity risk factors.
3. **Future Work**
   - Incorporate additional features like genetic markers or socioeconomic indicators for improved predictive accuracy.

## CONCLUSION:

This analysis highlights the effectiveness of data-driven approaches in understanding and combating obesity. The Random Forest model proves highly accurate in predicting obesity levels, making it a valuable asset for designing personalized health interventions. The findings also emphasize the importance of implementing targeted programs that promote physical activity, such as walking or cycling for short commutes, and encourage healthy eating and hydration habits. These initiatives can serve as practical, preventative measures to improve health outcomes for high-risk groups. Public health policies should prioritize early detection and tailored intervention strategies to address obesity risk factors effectively. Incorporating additional features like genetic markers and socioeconomic indicators into predictive models will further improve their accuracy and enable more comprehensive solutions for tackling obesity.