Obesity Analysis Using Lifestyle and Physical Condition Factors

Rekha, Reshma Vangala, Kavya | Statistics for Data Science | November 2025

1. Abstract:

This study investigates how lifestyle habits, eating behaviors, and physical conditions relate to obesity levels using the UCI Obesity Dataset. Statistical analysis, including t-tests, Chi-square tests, and ANOVA, identified significant differences across groups, such as higher average weights in males, greater weight among individuals with a family history of overweight, and strong impacts of snacking frequency on weight. A binary logistic regression model further revealed that BMI is the strongest predictors of obesity, while lifestyle factors such as water intake, exercise frequency, and meal patterns contribute smaller but meaningful effects. Overall, the findings show that physical measures are the primary determinants of obesity, with lifestyle behaviors providing supportive insights. This analysis emphasizes the value of data-driven approaches for improving obesity assessment and risk prediction.

2. Introduction:

Obesity is a growing global health concern linked to lifestyle habits, physical activity, and genetics. Understanding the factors that contribute to obesity is essential for prevention and early intervention. This project uses the “Estimation of Obesity Levels Based on Eating Habits and Physical Condition” dataset from the UCI Machine Learning Repository, which includes information on eating patterns, physical activity, and demographic factors of individuals from Mexico, Peru, and Colombia. The goal is to analyze the relationship between lifestyle behaviors and obesity levels and to develop predictive models that classify individuals’ obesity risk. The study highlights key factors influencing obesity and provides insights for data-driven health interventions.

2.1 Data Source and Scope:

The dataset used in this project comes from the UCI Machine Learning Repository, specifically the file **ObesityDataSet_raw_and_data_sinthetic.csv** (UCI Obesity Dataset). It contains **2,111 records** of individuals from Mexico, Peru, and Colombia, with **17 attributes** covering demographics, eating habits, physical activity, and family history related to being overweight.

The unit of analysis is an **individual person**, with each row representing a unique participant. The dataset includes:

- **Demographics:** Gender, Age, Height, Weight, Family history of overweight.
- **Eating habits:** HighCalFreq, VegMealFreq, MealsPerDay, Snacking, WaterIntake, TrackCalories.
- **Lifestyle habits:** ExerciseFreq, ScreenTime, AlcoholFreq, SMOKE, TransportMode.
- **Target variable:** ObesityLevel, classified into “Insufficient Weight,” “Normal Weight,” “Overweight Level I & II,” and “Obesity Type I–III.”

Notably, **77% of the data was generated synthetically using the Weka tool and SMOTE filter**, while **23% was collected directly from users through a web platform**. This combination ensures a robust dataset for analysis while maintaining representation of real-world behavior.

This project focuses on **analyzing relationships between lifestyle habits and obesity using hypothesis testing and regression analysis**, providing insights into key factors influencing obesity risk.

3. Data Acquisition and Cleaning/Preprocessing :

3.1 Data Acquisition and Initial Load :

The dataset used in this project was obtained from the UCI Machine Learning Repository. Specifically, the file ObesityDataSet_raw_and_data_sinthetic.csv was downloaded in CSV format.

For the analysis, Python libraries such as pandas, numpy, matplotlib, and seaborn were used. The dataset was loaded into a Pandas DataFrame directly from the GitHub repository:

```
url =
"https://raw.githubusercontent.com/RohitG57/Estimation-of-obesity-levels-based-on-eating
-habits-and-physical-condition/main/ObesityDataSet_raw_and_data_sinthetic.csv"
df= pd.read_csv(url)
```

The dataset was loaded into a Pandas DataFrame for inspection and further analysis using pd.read_csv().

3.2 Data Cleaning and Preprocessing:

Step 1. Renaming Columns: The original dataset contained long and complex column names. To simplify analysis and avoid syntax issues in Python, the columns were renamed with shorter, standardized names.

Step 2. Converting Columns to Appropriate Data Types: To optimize memory usage and improve performance, we converted certain columns to more suitable data types.

● **Categorical Conversion -** Columns with repeated text values, such as Gender, Snacking, AlcoholFreq, TransportMode, ObesityLevel, were converted to the category data type. This reduces memory usage and improves performance when analyzing repeated text values.

Step 3. Checking for missing value : The dataset was checked for missing or invalid values. No missing values were found, so no imputation or cleaning was necessary for this step.

Step 4. Checking for Duplicate Records : The dataset was checked for duplicate rows, and a few duplicate observations were found. These duplicates were removed to ensure data accuracy and prevent repeated information from affecting the analysis. The final dataset contained **2,087 rows and 17 columns**, ensuring cleaner and more reliable data for analysis.

Step 5. Created New Features: BMI

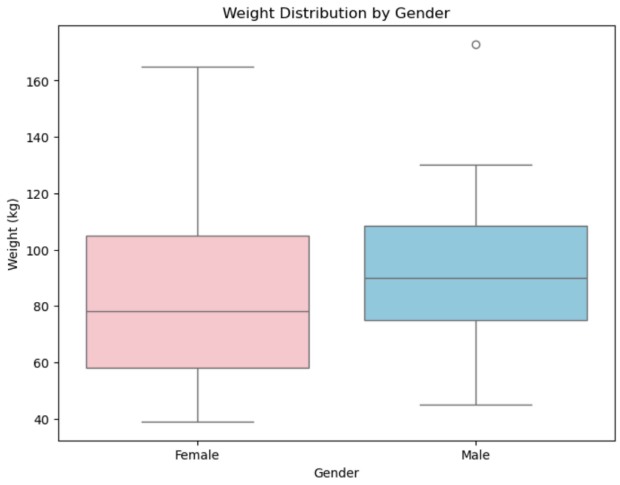
```
df['BMI'] = df['Weight'] / (df['Height']**2)
```

4. Analysis

4.1 Exploratory Analysis Using Boxplot

Question : Is there a significant difference in average weight between males and females?

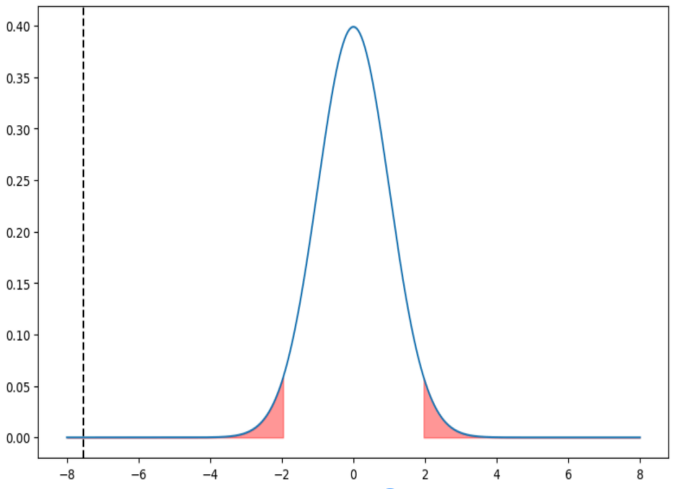
To examine how weight is distributed across genders, we first created a boxplot comparing Weight vs. Gender.



- **Higher Average Weight in Males**: Males weigh more on average (90.8 kg) than females (82.3 kg), with a higher median as well.
- **More Variability in Females**: Female weights are more spread out (std=29.7) compared to males (std=21.4).
- **Tighter Male Grouping**: The middle 50% of male weights (IQR=33.5 kg) is more clustered than females (IQR=47.0 kg).
- **Outliers Present**: There is one high-weight outlier among males (173 kg).

4.1.1 T-Test: Weight vs Gender

To determine if the difference in average weight between males and females is statistically significant, an independent two-sample t-test was performed.



Results:

- **T-statistic:** -7.53
- **P-value:** 7.68×10^{-14}
- **Critical t-value (α = 0.05)** ±1.96

Interpretation:

- The t-statistic falls far outside the acceptance range (|t| > t_critical), and the p-value is extremely small.
- This indicates that the difference in mean weights between males and females is statistically significant.

Conclusion:

- Males have a significantly higher average weight than females.
- The result supports the observation from the boxplot that weight distributions differ by gender.

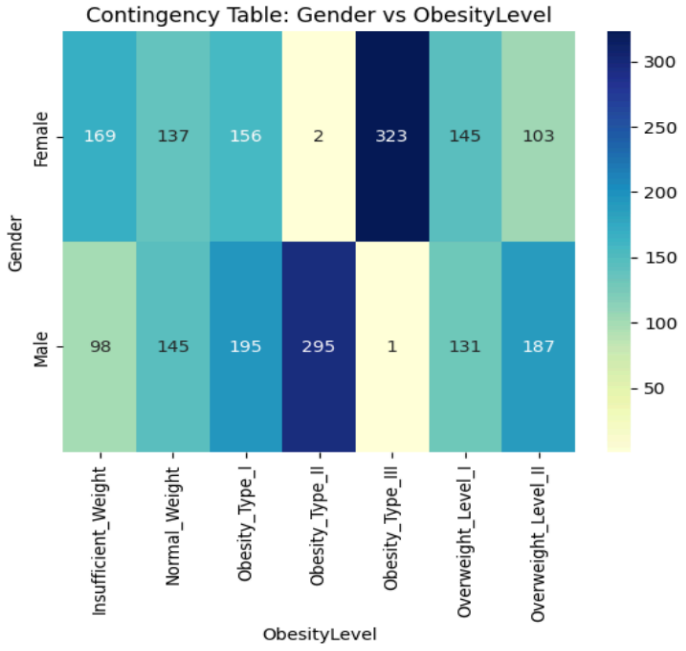
4.2 Chi-Square Test: Gender and Obesity Level

Question : Is obesity level associated with gender in this dataset?

Hypothesis

H_0: Gender and Obesity Level are independent (no association).
H_1: Gender and Obesity Level are not independent (significant association).

A Chi-square test of independence was conducted to determine whether there is a significant relationship between **Gender** (Male, Female) and **Obesity Level** (7 categories). A contingency table was created to compare how obesity levels are distributed across genders.



Interpretation:

- The distribution of obesity levels **differs between males and females**.

- For example:
  - More **males** fall into **Obesity_Type_II** (295 cases).
  - More **females** fall into **Obesity_Type_III** (323 cases).
- This suggests that **gender influences the pattern of obesity classification** in this dataset.

4.2.1 Chi-Square Test Procedure

A Chi-square test was performed on the contingency table for Gender and Obesity Level to compare the observed counts with the expected counts and determine whether the two variables are related.

Chi-Square Test Results:
- Chi-square Statistic: 657.45
- Degrees of Freedom: 6
- P-value: 0.0000
- Expected Counts:
  [[132.41 139.85 174.07 147.29 160.68 136.88 143.82]
  [134.59 142.15 176.93 149.71 163.32 139.12 146.18]]

Conclusion:
Since the p-value is 0.0000, there is a **significant association** between Gender and Obesity Level. This means obesity levels differ meaningfully between males and females.

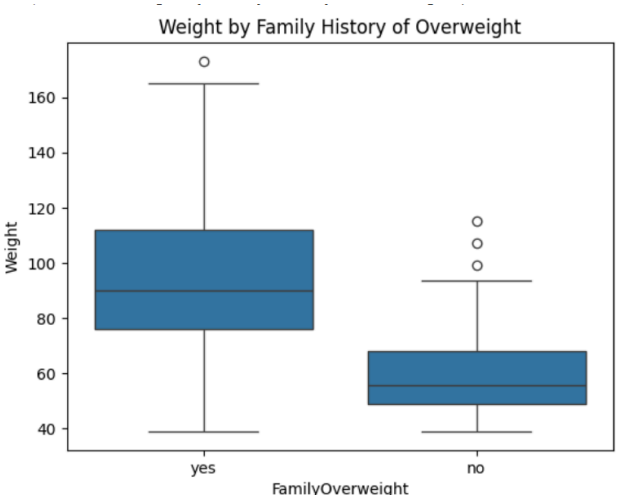
4.3 T-Test: Weight and Family History of Overweight

Question : Does weight differ significantly between individuals with and without a family history of being overweight?

Hypothesis:
- **Null Hypothesis (H_0):** There is **no difference** in average weight between the two groups.
- **Alternative Hypothesis (H_1):** There **is a significant difference** in average weight between the two groups.

Method: An independent two-sample t-test (with unequal variances) was performed using weight data from:
- Individuals with family history (“yes”)
- Individuals without family history (“no”)

A boxplot was also used to visually compare weight distributions.



Results:
- **T-statistic:** 35.8177
- **P-value:** 1.4275×10^{-173}

Conclusion:

- The p-value is extremely small, far below 0.05. Therefore, we reject the null hypothesis.
- There is a highly significant difference in weight between people with and without a family history of being overweight. The boxplot supports this result, showing that individuals with a family history tend to have higher weights, on average.

4.4 T-Test: Exercise Frequency vs Obesity Status

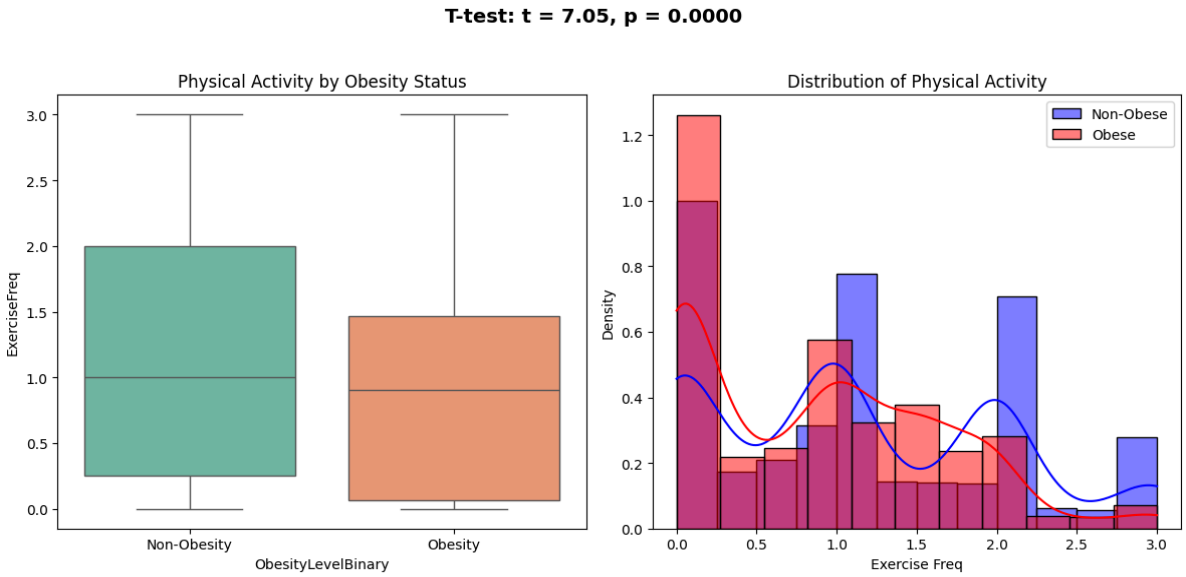
Question : Do non-obese individuals exercise more frequently than obese individuals on average?

- **H_0 (Null):** There is **no difference** in mean exercise frequency between obese and non-obese individuals.
- **H_1 (Alternative):** Non-obese individuals **have a higher** mean exercise frequency than obese individuals.

Method:
An independent two-sample t-test (with unequal variances) was performed using Obesity data with Exercise Frequency:

- Sample 1 : Individuals classified as Obese((Obesity_type_I, Obesity_type_II, Obesity_type III)
- Sample 2 : Individuals classified as Non - Obese(Insufficient_Weight, Normal_weight, Overweight_I, Overweight_II)

A boxplot & Histogram was also used to visually compare weight distributions.



Results:

- **T-statistic:** 7.05
- **P-value:** approximately 0.00000000000246

Conclusion:

- As the p-value is zero. Therefore, we reject the null hypothesis. This indicates the exercise frequency affects the obesity level. So we can reject the Null hypothesis that exercise does not affect obesity.

Boxplot: The median exercise frequency appears higher for the non obese group compared to the obese group. The spread of data looks somewhat similar but with the higher range of non-obese groups.

Histogram/KDE(with density curves) :

Non-obese(blue) : The distribution shows higher exercise frequencies are more common. **Obese(red)** : This distribution is skewed towards lower exercise frequencies.

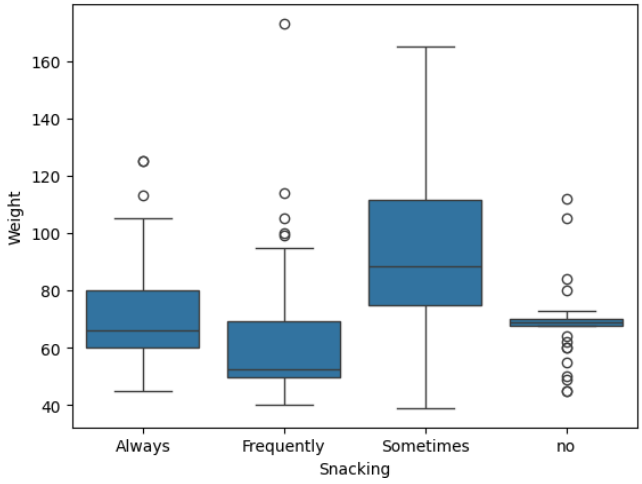
Overall the result suggests that there is a significant difference in exercise frequency between non-obese & obese groups. The non-obese group tends to have a higher exercise frequency compared to the obese group.

4.5 ANOVA: Snacking vs Weight

Question : To test whether the average weight differs among people with different snacking frequencies : Always, Frequently, Sometimes, and No.

4.5.1 Descriptive Statistics - Snacking vs Weight

Before running the ANOVA test, descriptive statistics were analyzed for each snacking group:



- **Sometimes:** Highest average weight (91.4 kg) with the widest variation (std = 25).
- **Always:** Moderate mean weight (71.1 kg).
- **No:** Mean weight (68.5 kg) with the least variation (std = 12.8).
- **Frequently:** Lowest mean weight (59.1 kg).

These results suggest that individuals who snack *sometimes* tend to have higher and more varied weights than other groups, supporting further testing with ANOVA.

4.5.2 ANOVA Test - Snacking vs Weight

- **Null Hypothesis (H_0):** Mean weight is the same across all snacking groups.
- **Alternative Hypothesis (H_1):** At least one group mean is different.

Results:

- F-statistic = 142.85
- p-value = 3.58×10^{-84}

Conclusion:
Since the p-value is far below 0.05, we reject the null hypothesis.
This means weight differs significantly among people with different snacking habits.

4.5.3 Post-hoc Analysis : Snacking vs Weight

After finding a significant ANOVA result, a **Bonferroni-corrected pairwise t-test** was used to see which snacking groups differ in weight.

Test Multiple Comparison ttest_ind FWER=0.05
method=bonf alphacSidak=0.01, alphacBonf=0.008

group1	group2	stat	pval	pval_corr	reject
Always	Frequently	4.6444	0.0	0.0	True
Always	Sometimes	-5.8805	0.0	0.0	True
Always	no	0.7432	0.4594	1.0	False
Frequently	Sometimes	-19.3069	0.0	0.0	True
Frequently	no	-3.2798	0.0012	0.007	True
Sometimes	no	5.5716	0.0	0.0	True

Observation:
- Significant weight differences were found between most pairs ($p < 0.05$).
- No significant difference between **“Always”** and **“No”** snacking groups ($p = 0.46$).
- The largest differences occurred between **“Frequently”** and **“Sometimes”** snackers.

Conclusion:
Snacking frequency has a strong impact on weight, people who snack sometimes tend to have **much higher average weights** than those who snack frequently or always.

Logistic Regression Model for Obesity Prediction

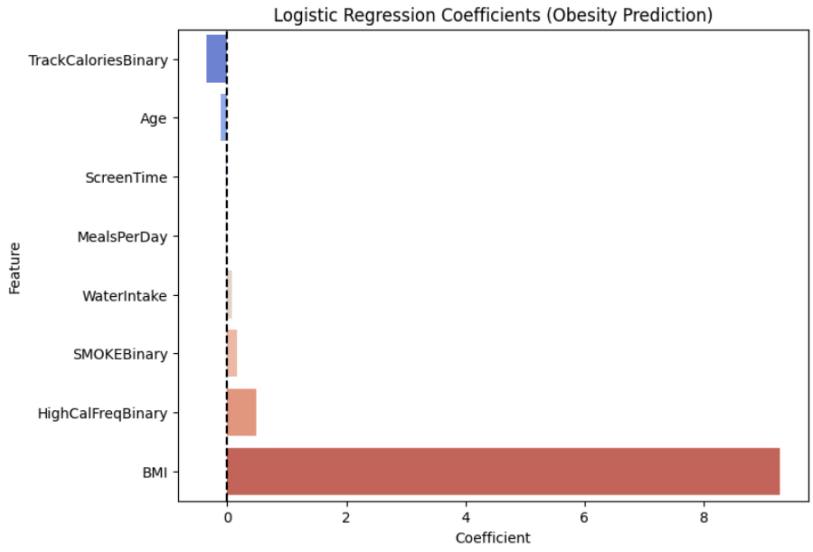
Question : To understand which factors (BMI, lifestyle habits, etc.) are most important in predicting a person’s obesity level.

Objective
The purpose of this analysis is to understand which factors, such as height, weight, BMI, and lifestyle habits, are most important in predicting whether a person is obese.

We used a logistic regression model to classify each individual as either:
- **Obese** (Obesity class I, II, III), or
- **Non-Obese** (Insufficient weight, Normal, Overweight I/II).

How to Interpret Coefficients
- **Positive coefficient** : increases the probability of being in that obesity group
- **Negative coefficient** : decreases the probability
- **Coefficient near zero** : little or no influence

- **Large absolute value** : strong impact



Observation:

1. BMI is the strongest predictor

- BMI has the **largest positive coefficient**, meaning people with a higher BMI are far more likely to be classified as obese.
- This strongly affects the model's performance because obesity categories themselves are based on BMI.
- The model becomes extremely accurate when BMI is included.

2. Lifestyle factors have much smaller effects

- Variables like water intake, screen time, meals per day, and exercise habits have coefficients close to zero.
- This means they contribute very little to predicting obesity in our dataset.

3. Some lifestyle factors still show patterns

- **High-calorie food frequency** slightly increases the chance of obesity.
- **Smoking** also has a small positive effect.
- **Tracking calories** has the strongest negative coefficient — people who track their food are less likely to be obese.

These effects are small compared to BMI.

Model Accuracy Summary:

To understand how each feature affects obesity prediction, we tested three different logistic regression models:

Model 1 (With BMI):

- Accuracy was **99.5%** – the highest.
- BMI is the best predictor of obesity.

Model 2 (Without BMI):
- Accuracy dropped to **68.3%**.
- Lifestyle habits alone don't predict obesity very well.

Model 3 (With Weight Only):
- Accuracy improved to **88%** when weight was added (but no BMI).
- Weight is a strong predictor, but still not as strong as BMI.

Model Version	Features Used	Accuracy
Model 1: With BMI	Age, HighCalFreqBinary, SMOKEBinary, TrackCaloriesBinary, WaterIntake, ScreenTime, MealsPerDay, **BMI**	**0.995**
Model 2: Without BMI	Age, HighCalFreqBinary, SMOKEBinary, TrackCaloriesBinary, WaterIntake, ScreenTime, MealsPerDay	**0.683**
Model 3: With Weight Only (no BMI)	Age, HighCalFreqBinary, SMOKEBinary, TrackCaloriesBinary, WaterIntake, ScreenTime, MealsPerDay, **Weight**	**0.88**

Key Findings

1. Gender Differences in Weight

Males weigh significantly more than females, confirmed by the t-test and supported by the boxplot.

2. Gender and Obesity Level Are Connected

The Chi-square test shows a clear association between gender and obesity level. Obesity categories differ meaningfully between males and females.

3. Family History Strongly Affects Weight

People with a family history of being overweight tend to weigh much more. This was confirmed by both statistical results and visual graphs.

4. Exercise Frequency Influences Obesity

Non-obese individuals exercise more often compared to obese individuals. Exercise frequency shows a strong link with obesity status.

5. Snacking Habits Impact Weight

Weight varies significantly across snacking groups. Those who snack "Sometimes" have the highest average weight. Post-hoc tests show meaningful differences between several snacking habits.

6. BMI Is the Best Predictor of Obesity

Logistic regression results show:
- With BMI: 99.5% accuracy
- Without BMI: Accuracy drops to 68.3%
- With Weight Only: Accuracy = 88%

BMI explains obesity far better than any lifestyle factor.

7. Lifestyle Factors Have Small Predictive Power

Habits such as water intake, smoking, exercise, and calorie tracking have very small effects individually compared to BMI.