Introduction

In this project, I will use a dataset from a survey held by NHANES in 2015-16. I have collected this dataset from Kaggle. I will perform descriptive and diagnostic analysis with the aim to collect and compare data of BMI with age and gender.
This dataset contains 28 columns and 5735 rows. I have isolated the relevant columns for my analysis.

Brief description of the columns:

- **SEQN:** Respondent Sequence Number
- **SMQ020:** Smoking
- **RIAGENDR:** Gender
- **RIDAGEYR:** Age (years)
- **DMDEDUC2:** Education level
- **BMXWT:** Weight(kg)
- **BMXHT:** Height(cm)
- **BMXBMI:** BMI

Data cleaning

Analysis is focused on 7 columns - 3 with categorical variables and 4 with numerical variables.

- Categorical - Smoking, gender, education level
- Numerical - Age, weight, height, BMI

Steps for data cleaning:

1. Checking duplicate or unnecessary values
2. Checking null or missing values
3. Checking outliers

No duplicate rows were found so there was no need for sequence column.

For missing values

I have missing values in education, weight, height and BMI. Addressing the missing values in education, we have remove them as filling the missed value can lead to inaccuracies in analysis. For weight, height and BMI, I will drop them also as the number is small compared to the dataset (>5000 rows).

Handling the outliers

Utilizing the histogram and boxplot,
A histogram shows frequency vs value ranges.
It is found that height shows normal distribution curve while weight and BMI has right right-skewed distribution.
Histograms only suggest outliers — they don't confirm them. To confirm them, boxplot are plotted.

For box plot

IQR = Q3 − Q1
Lower bound = Q1 − 1.5 × IQR
Upper bound = Q3 + 1.5 × IQR
Q3 = 75 percentile
Q1 = 25 percentile

Through boxplot, we are ensured that age has no outliers while height has normal distribution with few outliers. Weight and bmi has more outliers in upper values
Based on observation from the histogram and boxplot, I have considered value lower than minimum and larger than the maximum as outliers. For weight and BMI, values above maximum are considered are outliers

After removing outliers, the result is data with normal distribution.

Feature Engineering

4.1 Encoding Categorical Variables

In this step, categorical variables were converted into a numerical format to make them suitable for machine learning algorithms.

- **Method Used:** One-Hot Encoding (using Pandas `get_dummies`).
- **Variables Processed:** The categorical features identified were `smoking`, `gender`, and `education`.
- **Implementation Details:** The parameter `drop_first=True` was utilized. This technique creates k-1 binary columns for a categorical feature with k levels. This is a standard practice to avoid multicollinearity (also known as the "dummy variable trap"), where one variable can be predicted from the others.
- **Result:** The dataset was expanded with binary variables representing the presence or absence of specific categories (e.g., specific education levels or smoking status).

4.2 Handling Skewness in Numerical Variables

An analysis was conducted to check if the numerical variables required transformation (such as Log Transformation) to address skewness, which can affect the performance of certain statistical models.

- **Variables Analyzed:** `age`, `weight`, `height`, and `bmi`.
- **Threshold Set:** A skewness limit of **0.75** was established. Variables exceeding this threshold would be candidates for transformation.
- **Findings:**
    1. **Age:** ~0.11
    2. **Weight:** ~0.65
    3. **Height:** ~0.08
    4. **BMI:** ~0.52
- **Conclusion:** All numerical variables exhibited skewness values below the 0.75 limit. Consequently, **no log transformations were applied**, as the data was determined to be sufficiently symmetric for the intended analysis.

4.3 Feature Correlation Analysis

- **Method:** A Pairplot (using the Seaborn library) was generated.
- **Objective:** To visually inspect the pairwise relationships between numerical variables (`age`, `weight`, `height`, `bmi`) and identify potential correlations or patterns that could inform future modeling or hypothesis testing.

Hypothesis Testing

This section investigates three specific questions about the population using statistical inference. For each question, a null hypothesis (H_0) and an alternative hypothesis (H_1) were formulated and tested using a significance level (α) of 0.05.

5.1 Hypothesis 1: Obesity in Females Aged 40-50

Objective: To determine if the mean BMI of females aged 40-50 is significantly greater than the obesity threshold of 30.

- **Hypotheses:**
  - **Null Hypothesis (H_0):** The mean BMI of females aged 40-50 is < 30 (Not obese).
  - **Alternative Hypothesis (H_1):** The mean BMI of females aged 40-50 is > 30 (Obese).
- **Methodology:**
  - **Test:** One-sample Z-test/T-test approach.
  - **Sample Statistics:** Mean BMI approx 30.17, Sample Size n = 476.
- **Results:**
  - **P-value:** 0.45
- **Conclusion:**
  - Since p-value(0.45) > α(0.05), we **fail to reject the null hypothesis**.
  - **Interpretation:** While the sample mean (30.17) is slightly above 30, there is insufficient statistical evidence to conclude that the population of females in this age group is significantly "obese."

5.2 Hypothesis 2: Gender and Smoking Habits

Objective: To investigate if there is a dependency between gender and smoking status (i.e., do men smoke more or less than women?).

- **Hypotheses:**
  - **Null Hypothesis (H_0):** Gender and smoking status are **independent** (Proportion of smokers is equal across genders).
  - **Alternative Hypothesis (H_1):** Gender and smoking status are **dependent** (Proportion of smokers differs by gender).
- **Methodology:**
  - **Test:** Chi-Square Test of Independence (chi2_contingency).
  - **Input:** Contingency table of Gender vs. Smoking.
- **Results:**
  - **P-value:** 8.66*10^(-57)(< 0.001)
- **Conclusion:**
  - Since p-value < α(0.05), we **reject the null hypothesis**.
  - **Interpretation:** There is extremely strong evidence that smoking habits are statistically dependent on gender.

5.3 Hypothesis 3: BMI Comparison by Gender

Objective: To compare the Body Mass Index (BMI) between males and females to see if one group has a significantly different average BMI.

- **Hypotheses:**
  - **Null Hypothesis (H_0):** The mean BMI of males equals the mean BMI of females μ(male) = μ(female).
  - **Alternative Hypothesis (H_1):** The mean BMI of males does not equal the mean BMI of females μ(male) != μ(female).
- **Methodology:**
  - **Test:** Welch's Two-Sample t-test (ttest_ind with equal_var=False).
  - **Justification:** Welch's t-test was chosen to account for potential unequal variances between the male and female groups.
- **Results:**
  - **T-statistic:** -4.33
  - **P-value:** 1.54*10^(-5) (< 0.001)
- **Conclusion:**
  - Since p-value < α(0.05), we **reject the null hypothesis**.
  - **Interpretation:** There is a statistically significant difference in the average BMI between males and females in this dataset.