**INTRODUCTION:**

I have performed EDA and data Cleaning on Stroke Prediction Dataset available on Kaggle : <https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset>

This Report consists of the Summary of the tasks that I have performed and the results that I obtained .

### **Dataset Description**

The dataset consists of the following key attributes:

* **id**: Unique identifier for each patient.
* **gender**: Male, Female, or Other.
* **age**: Patient's age in years.
* **hypertension**: Binary indicator (0: No, 1: Yes).
* **heart\_disease**: Binary indicator (0: No, 1: Yes).
* **ever\_married**: Whether the patient was ever married.
* **work\_type**: Type of employment (Private, Govt, etc.).
* **Residence\_type**: Urban or Rural.
* **avg\_glucose\_level**: Average blood glucose level.
* **bmi**: Body Mass Index.
* **smoking\_status**: Current, Former, or Never smoked.
* **stroke**: Target variable (0: No Stroke, 1: Stroke).

The dataset comprises **5,110** records with some missing values in the BMI column and outliers which required handling during analysis. However there were no duplicates present.

**Data Cleaning & Preprocessing**

## **Data Cleaning & Preprocessing**

### **1. Handling Missing Values**

* Identified missing values in the BMI column.
* Used median imputation to fill missing BMI values.
* Checked for missing values in categorical variables and filled them with the mode.

### **2. Handling Duplicates**

* Verified if duplicate records existed and removed them where necessary.

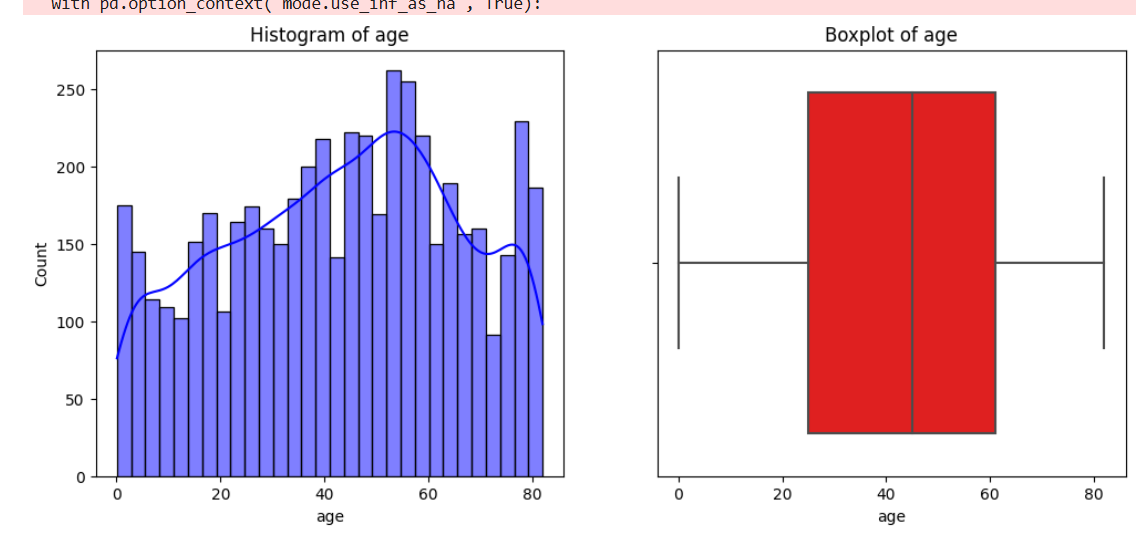
### **3. Handling Outliers**

* Used Box Plots and IQR Method to detect and handle outliers.
* Outliers in BMI and Glucose Level were replaced using the median.

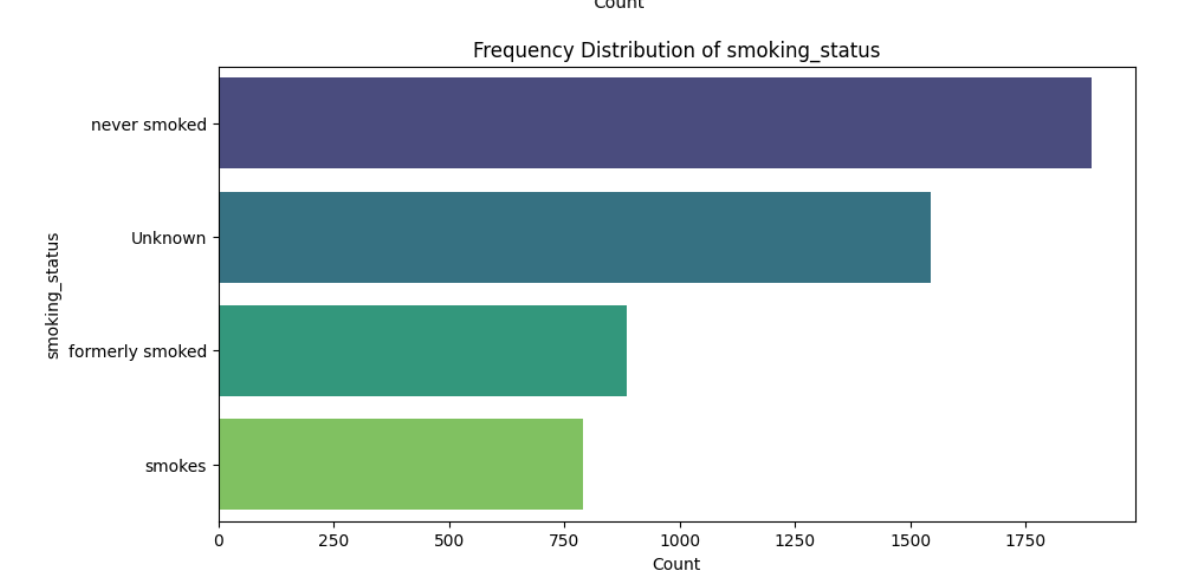
## **Exploratory Data Analysis (EDA)**

### **1. Univariate Analysis**

To analyze individual variables, we performed the following steps:

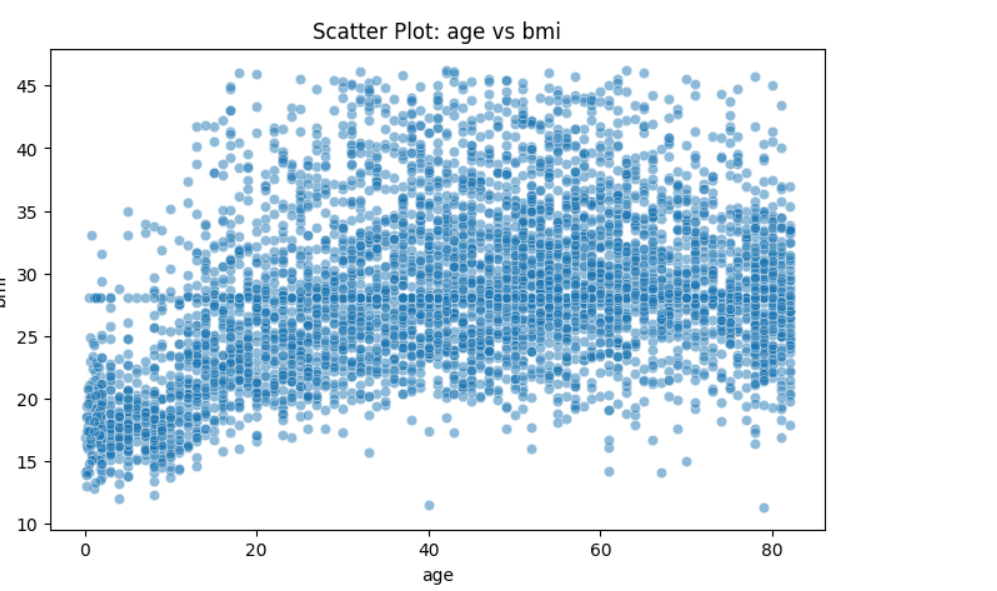
* **Histograms :** Visualized distributions of age, bmi, and avg\_glucose\_level.
* **Bar Charts:** Analyzed categorical variable distributions.
* **Box Plots:** Identified outliers in numerical variables.

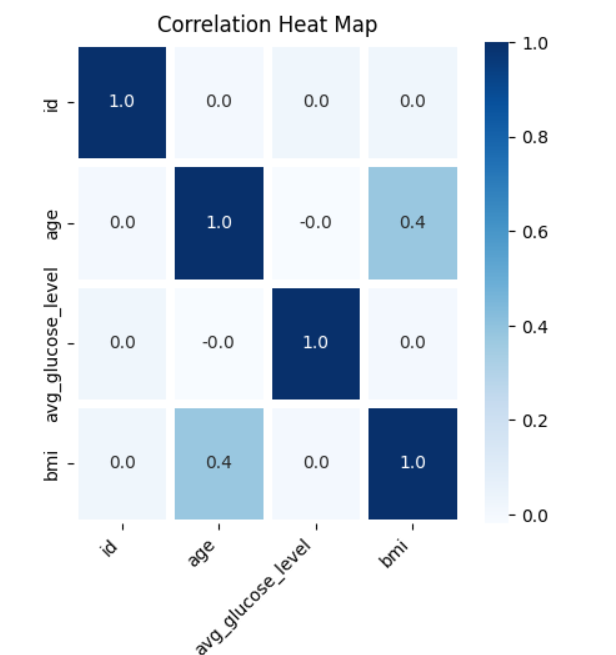
**Histogram and BoxPlot**

****

### **2. Bivariate Analysis**

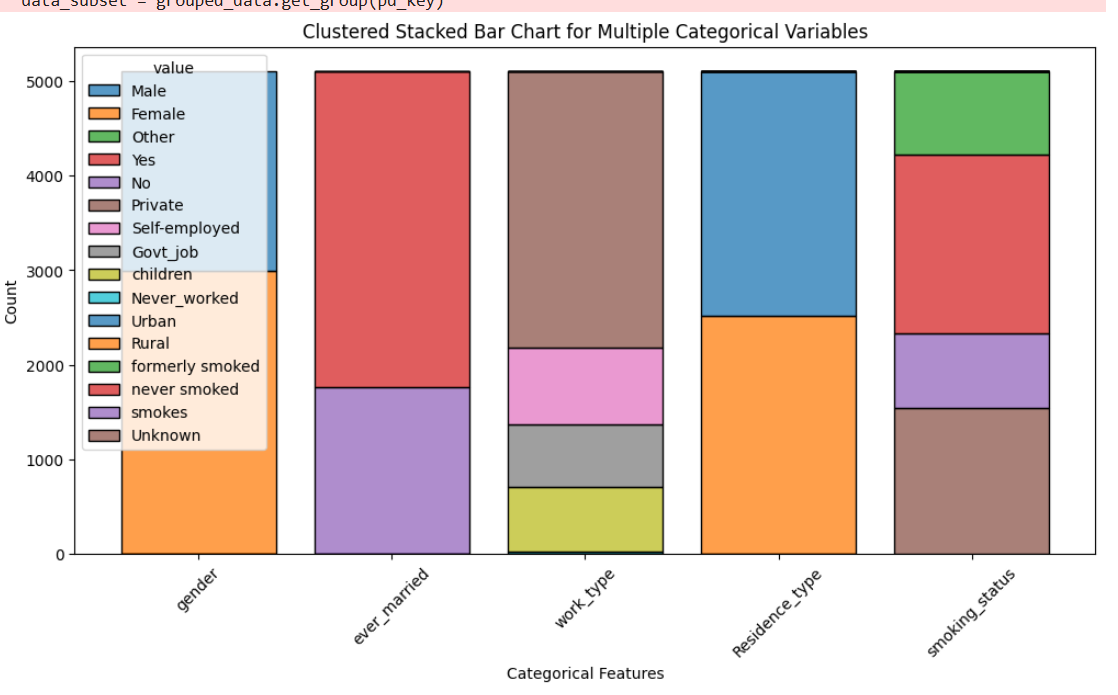
* **Violin Plots:** Examined the relationship between categorical variables (hypertension, smoking\_status) and numerical variables (bmi, age).(Numerical v/s Categorical)
* **Scatter Plots:** Showed interactions between continuous variables like avg\_glucose\_level and bmi. (Numerical v/s Categorical)
* **Correlation Matrix:** Used seaborn.heatmap to display relationships among numerical variables.

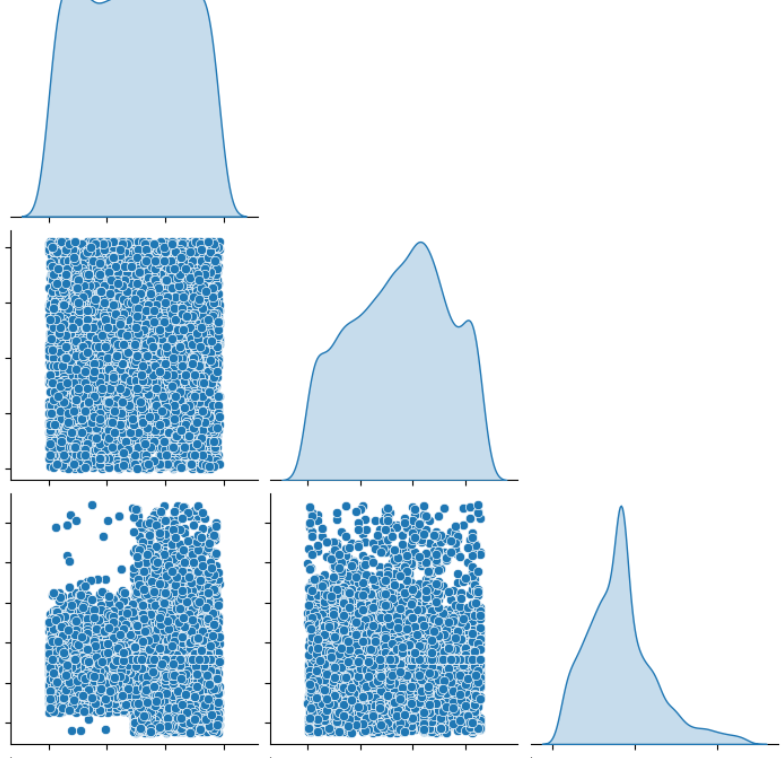


* 

### **Multivariate Analysis**

* **Pair Plots:** Identified patterns across multiple numerical variables.
* **Stacked Bar Charts:** Examined patterns across various categorical variables





### **Conclusion:**

1. **Age and BMI have a moderate positive correlation (0.4)**
   * This suggests that as **age increases, BMI tends to increase moderately**.
   * This makes sense as older individuals may have higher BMI due to lifestyle or metabolic changes.
2. **No significant correlation between avg\_glucose\_level and other variables**
   * The correlation between avg\_glucose\_level and age **(≈ 0.0)** and BMI **(≈ 0.0)** is **very weak or negligible**.
   * This indicates that glucose levels do not strongly depend on age or BMI in this dataset.
3. **ID has no correlation with any feature**
   * As expected, the ID column has a **correlation of 0 with all variables**, indicating it is just an index.