1 Organise Data

The dataset comprises 43,400 records and 12 columns, including Patient's ID, Gender, Age, Hypertension status, Heart Disease status, Marital status, Work Type, Residence Type, Average Glucose Level, BMI, Smoking status, and Brain Stroke occurrence. Each row represents an individual's demographic and health-related information, sourced from openly available health data.

Sr No	Name	Description	Data Type
1	ID	Patient's ID number	Nominal
2	Gender	Patient's gender	Nominal
3	Age	Age of patient	Numerical (Int)
4	Hypertension	Patient's high blood pressure status	Nominal
5	Heart Disease	Indication of patient's heart disease status	Nominal
6	Married	Indicates whether the patient is married	Nominal
7	Occupation	Patient's occupation	Nominal
8	Residence	Type of residence	Nominal
9	Average Glucose Level	Average blood glucose level of the patient	Numerical (Float)
10	BMI	Body Mass Index (BMI) of the patient	Numerical (Float)
11	Smoking	Indicates whether the patient is a smoker	Nominal
12	Brain Stroke	Indicates whether the patient has experienced a brain stroke	Nominal

Table 1: Data Dictionary

Preprocessing steps, such as renaming variables and addressing data quality issues (including handling missing values and detecting outliers), were undertaken to ensure the dataset's readiness for predictive modelling and analysis. These steps are described in the following section .

1.1 Data Preprocessing

The dataset is preprocessed using Python. We have used Google Colab for the implementation .

1.1.1 Data Exploration

- The dataset is imported using Pandas library as shown in Figure 1.

Figure 1: Loading the dataset

- Firstly, we will check the size of the dataframe as shown in Figure 2.

Figure 2: Size of the dataframe

The Dataset contains 43,400 row and 12 columns.

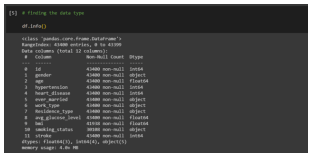


Figure 3: Data Type of each columns

- The type of variables can be seen using info() command. See Figure 3.

  Since, Stroke, heart disease or hypertension describes the plausibility of occurrence. These can be Nominal variables. So, the dataframe contains 4 Numerical fields and rest as Nominal fields.

- Then, we will find the summary of all columns.

  There is a possibility of outliers in the BMI and Average Glucose Level columns as shown in Figure 4(a).

  Figure 4(b) doesn't show any spelling mistakes. .

(a) (b)

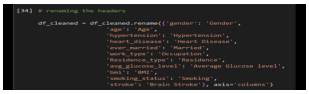
Figure 4: Summary of all columns

1.1.2 Data Cleaning

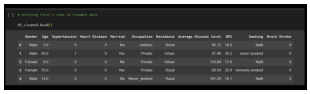
- Excluding the *ID* column, as it is not necessary for our analysis. See Figure 5.

Figure 5: Dropping the *ID* Column

- Renaming column headers to enhance understanding of all dataset columns. Printing and inspecting the first few columns of the dataset as shown in Figure 6.



(a)



(b)

Figure 6: Summary of all columns

- Finding NAs in the dataset. See Figure 7.

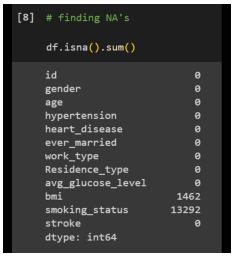


Figure 7: Finding NAs in the Dataset

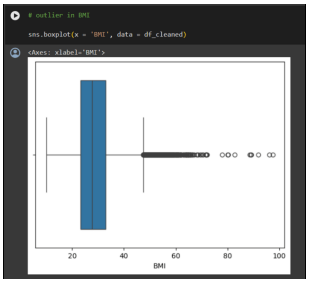
We have missing values (NA's) in the *BMI* and *Smoking Status* columns. We have decided to impute these missing values in the next step.

- Addressing missing values in the *Smoking* column by imputing NA values with an unknown category as shown in Figure 8.

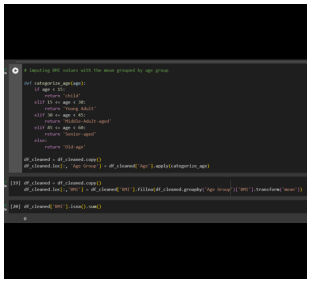
Figure 8: Solving in NAs in *Smoking*

- Checking Outliers in the *BMI* column. See Figure 9(a).

  We have observed outliers in the *BMI* column and have chosen to address them by imputing the BMI values with the mean, grouped by age group as shown in Figure 9(b). We are removing ages below 5, assuming that children under 5 years are not prone to brain stroke.



(a)



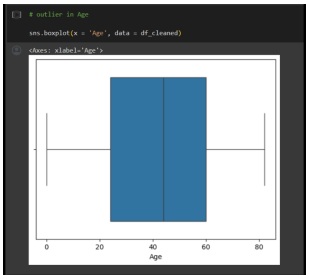
(b)

Figure 9: Solving NAs in *BMI*

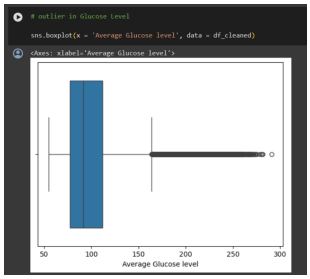
- Checking Outliers in the *Age* column

  There are no outliers present in the *Age* Columns seen in Figure 10(a).

- Checking Outliers in the *Glucose Level* column



(a)



(b)

Figure 10: NA's in Age and Glucose Level

We have identified outliers in the glucose level column as seen in the Figure 10(b), suggesting values beyond the typical range for most candidates in the dataset. To ensure data integrity, we have decided to remove all glucose level values exceeding 260.