Stroke Data Analysis - Group Assignment

## Objective

The groups' objective of this task is to perform data cleaning and exploratory data analysis (EDA) on the given dataset related to strokes.

Specific Objectives:

1. **Identify Demographic Trends**

Explore and analyze demographic factors such as age, gender, marital status, and residence type to identify any trends or patterns related to strokes.

1. **Assess Health Metrics Impact**

Investigate the impact of health metrics, including hypertension, heart disease, average glucose level, and BMI, on the occurrence of strokes.

1. **Evaluate Lifestyle Factors**

Examine lifestyle factors, specifically smoking status and work type, to understand their association with strokes.

1. **Explore Correlations**

Conduct correlation analysis to identify relationships between different health conditions (hypertension, heart disease) and understand their potential collective influence on strokes.

## Dataset:

You are provided with a dataset containing information about individuals and their health status, with a focus on strokes. The dataset is contained [here](https://drive.google.com/file/d/1JcqDBxq0D9a_hR1pP36SF159DmJvFU_k/view?usp=sharing). The dataset is in tabular form and includes the following columns:

## Data Dictionary:

1. **id**: Unique identifier for each individual.
2. **gender**: Gender of the individual (Male/Female).
3. **age**: Age of the individual.
4. **hypertension**: Indicates if the individual has hypertension (1: Yes, 0: No).
5. **heart\_disease**: Indicates if the individual has heart disease (1: Yes, 0: No).
6. **ever\_married**: Marital status.
7. **work\_type**: Type of employment.
8. **Residence\_type**: Type of residence
9. **avg\_glucose\_level**: Average glucose level in the blood.
10. **bmi:** Body Mass Index.
11. **smoking\_status**: Smoking status.
12. **stroke**: Indicates if the individual had a stroke (1: Yes, 0: No).

## Tasks

### Data Loading

### Data Inspection

* Preview rows of the dataset to get a sense of its shape and structure.
* Check for the data types of each column and handle any inconsistencies.
* Check for missing values in each column.

### Data Cleaning:

* Handle missing values
* Check for and handle any duplicate rows.
* Handle any outliers
* Convert data types as needed.

### Exploratory Data Analysis (EDA):

* Generate descriptive statistics for numerical columns (mean, median, std, etc.).
* Explore the distribution of age, avg\_glucose\_level, and bmi.
* Visualize the distribution of strokes in the dataset.
* Analyze the distribution of gender, work\_type, and smoking\_status with respect to strokes.
* Explore the correlation between hypertension, heart\_disease, and strokes.
* Identify and analyze any patterns or trends that might be interesting.
* Use any tool i.e. groupby, pivot tables and crosstab to perform any additional analyses that you find relevant.

### Summary and Insights

* Summarize your findings and insights from the analysis.
* Provide recommendations or further steps based on your analysis.

## Deliverables

GitHub Repository

* Python script or Jupyter Notebook with the code for data loading, cleaning, and analysis and contain visualizations (plots, charts) generated during EDA.  ***Ensure that your code is well-commented and easy to understand.***
* Powerpoint slides.
* ReadMe

Submit the [github repository link](https://docs.google.com/spreadsheets/d/1X-A29vWAodTZNUIyI-c7F46obSaKtWkfKZ10nKLRBEs/edit#gid=0) here.