Brain Stroke Data Analysis Report

Team Members:

1. Mostafa Fathy Mahmoud Abdel Rahim
2. Abanoub Fahem Fawzy Fahem
3. Michael Mamdouh Sedrak
4. Mohamed Ayman Sobhi

Supervisor: Abdullah Kamal

1. Introduction

This report presents the findings from a brain stroke data analysis project. The project aimed to explore the factors contributing to strokes and to identify significant trends through data analysis and visualization. The initial data was sourced from multiple formats: CSV files, Excel spreadsheets (XLSX), and a database. These different data sources were consolidated into a single dataset for analysis. Additionally, new data from another database was incorporated, adding two important columns: death and hospital visits.

Various tools were used, including Excel for preliminary handling, Python for cleaning and manipulation, and Tableau for visualization.

2. Dataset Overview

The final dataset contains 5,233 records with 13 variables, including the newly added columns:

Demographics: Gender, age, and marital status.

Health Indicators: Hypertension, heart disease, BMI (Body Mass Index), and average glucose levels.

Lifestyle Factors: Smoking status and work type , Residence type .

Outcome Variables: Stroke occurrence, hospital visits, and death.

This dataset enables a comprehensive analysis of stroke risk factors and outcomes, such as hospital visits and mortality due to strokes.

3. Data Cleaning Process

The raw dataset exhibited various data quality issues that were addressed before analysis:

Missing Values: Several columns, including hypertension, heart_disease, and stroke, contained missing values. These were imputed with appropriate methods—median values for numerical columns and the most frequent value for categorical columns.

Data Entry Errors: Typographical errors, especially in fields like gender and smoking_status, were corrected. For example, "mle" was corrected to "Male."

Outliers: Unrealistic values, such as an age of 149, were removed.

Inconsistent Formatting: Variations in text case (e.g., "PRIVATE" vs. "private") were standardized.

Non-Numeric Values in Numerical Fields: Columns such as avg_glucose_level and bmi had nonnumeric values like "high" or "NAN," which were replaced with valid numeric data where appropriate.

The additional columns, death and hospital visits, were added after cleaning, ensuring consistency with the rest of the dataset. These new variables allowed us to further explore the consequences of strokes.

4. Tools and Methodology

To process, analyze, and visualize the data, we used the following tools:

Excel: For initial inspection and simple calculations.

Python: For data cleaning, manipulation, and preliminary visualizations. Libraries like pandas and numpy were used for handling the dataset, while Matplotlib was used for early-stage visualizations.

Tableau: After the data was cleaned and updated with the new columns, Tableau was used to create interactive dashboards, helping visualize the impact of strokes and hospital visits or deaths.

Methodological Steps:

Descriptive Analysis: We computed summary statistics for key variables such as age, BMI, and average glucose level, including the newly added hospital visits and deaths.

Correlation Analysis: We explored relationships between variables such as hypertension and stroke, as well as stroke outcomes like hospital visits and death.

Visualization: Dashboards were created in Tableau to visually represent key insights about stroke risk factors, mortality rates, and hospitalization trends.

5. Key Insights

5.1. Age and Stroke Risk

The analysis showed a strong correlation between age and stroke occurrence. Older individuals had a higher stroke incidence, which aligns with medical research indicating age as a key risk factor.

5.2. Hypertension's Impact

Hypertension was identified as a significant risk factor for strokes. Individuals with high blood pressure were far more likely to experience a stroke than those without.

5.3. BMI and Stroke Risk

A higher Body Mass Index (BMI) was associated with an increased stroke risk, especially in those with preexisting heart conditions.

5.4. Smoking and Stroke

Former and current smokers were at a higher risk of stroke compared to non-smokers. The data showed that former smokers had a higher stroke incidence, suggesting that the negative effects of smoking may persist even after quitting.

5.5. Hospital Visits and Mortality

With the addition of the new columns, we found:

Hospital Visits: Many stroke victims required hospitalization, particularly those with hypertension or high BMI.

Death: There was a noticeable correlation between age, stroke occurrence, and death. Older individuals who had strokes were more likely to pass away, especially if they also had conditions like heart disease or hypertension.

5.6. Other Observations

Work Type: Private-sector employees had a higher rate of strokes compared to other occupational categories.

Residence Type: There were no significant differences between urban and rural dwellers in terms of stroke rates or outcomes.

6. Visualization in Tableau

After integrating the new data, we visualized the stroke outcomes, hospital visits, and death rates using Tableau. The dashboards provided insights into:

Stroke risk by age, gender, and health conditions.

The relationship between stroke occurrences and hospital visits.

Stroke-induced mortality rates based on age and comorbidities like hypertension and heart disease.

7. Conclusion

The analysis of brain stroke data provided valuable insights into the factors contributing to strokes, including age, hypertension, BMI, and smoking status. The integration of new data, such as hospital visits and death, added depth to the analysis, highlighting the severe outcomes of strokes. By using tools like Python for data cleaning and Tableau for visualization, we were able to uncover trends and correlations that can guide future research and public health policies for stroke prevention.