Course Number: DATA 301 *Instructor: Dr. Osei Tweneboah*

Stroke Data Analysis and Visualization

Anish Bhurtyal[1], Bijay Adhikari [2], Sadikshya Shrestha [3,] and Saurav Dahal[4]

[1] Affiliation 1;abhurtya@ramapo.edu
[2] Affiliation 2; badhika2@gmail.com
[3] Affiliation 2; sshres24@ramapo.edu
* Correspondence: sdahal4@ramapo.edu

Abstract: Strokes pose significant health risks globally, emphasizing the importance of early intervention and preventive strategies. This study utilizes the "Stroke Prediction Dataset" from Kaggle to analyze and predict stroke occurrences. The dataset encompasses various demographic and health-related factors, including age, hypertension, heart disease, marital status, occupation, residence type, average glucose level, BMI, smoking status, and stroke occurrence. Through exploratory data analysis, missing value imputation, and statistical modeling, we uncover insights into stroke risk factors. Key findings reveal significant associations between stroke and hypertension, heart disease, age, average glucose level, BMI, marital status, work type, and smoking status. Logistic regression analysis highlights age, heart disease, average glucose level, BMI, hypertension, and gender as significant predictors of stroke occurrence. Visual representations aid in understanding the impact of these factors on stroke risk. This study underscores the importance of comprehensive risk assessment and emphasizes potential avenues for preventive interventions.

Keywords: Stroke prediction; Logistic regression; Exploratory data analysis; Demographic factors; Hypothesis testing

---

1. Introduction

Stroke, a critical medical emergency, occurs when blood flow to the brain is interrupted or when there is sudden bleeding in the brain. This interruption deprives brain cells of oxygen and nutrients, leading to potentially severe consequences such as lasting brain damage, long-term disability, or even death. According to the Centers for Disease Control and Prevention (CDC), strokes accounted for 1 in 6 deaths from cardiovascular disease in 2021, highlighting their significant impact on public health [1].

The prevalence of strokes is substantial, with more than 795,000 people in the United States experiencing a stroke annually, a considerable portion of which are first or new occurrences. The incidence of stroke is influenced by various factors, including lifestyle choices and underlying health conditions. High blood pressure, high cholesterol, smoking, obesity, and diabetes stand as leading contributors to stroke risk, with approximately one-third of U.S. adults having at least one of these risk factors [1].

Understanding the factors contributing to stroke risk is crucial. This study aims at examining demographic and health-related variables, we seek to uncover patterns and predictors of stroke risk, thereby contributing to the development of targeted preventive measures and interventions. In this paper, we present our analysis of the "Stroke Prediction Dataset" with exploratory data analysis, statistical modeling, and visualization techniques. We highlight significant associations between various factors and stroke occurrence. Ultimately, our work aims to enhance understanding of stroke risk factors and inform strategies for stroke prevention and management.

Research Questions

In pursuit of our objective, we address the following research questions:

- **Research Question 1:** What is the distribution of the predictive physical health factors and demographic factors such as hypertension, health stroke, BMI, and glucose levels, etc., among individuals in the dataset?

- **Research Question 2:** How do different factors interact to influence the risk of stroke in a population?

- **Research Question 3:** Does higher mean age, BMI, and glucose level correlate with an increased risk of stroke compared to individuals with lower values in these parameters?

- **Research Question 4:** Based on the previous visual and statistical analyses, how do the different factors affect the chance of having a stroke, and how can these be visually represented to aid understanding?

2. Materials and Methods

The Materials and Methods section outlines the procedures and techniques employed in the analysis of stroke occurrences using the "Stroke Prediction Dataset" obtained from Kaggle. This section provides detailed information to facilitate replication and extension of the study's findings.

2.1 Data Collection and Preparation

The primary dataset utilized in this study is the "Stroke Prediction Dataset," sourced from Kaggle. The dataset contains a comprehensive collection of health-related variables, demographic information, and stroke occurrences. Prior to analysis, the dataset underwent preprocessing steps to ensure data quality and suitability for statistical modeling. Missing values in the dataset, particularly in the 'bmi' column, were addressed through imputation using the median value. Additionally, categorical variables were encoded appropriately for subsequent analysis.

The dataset features:

- Age: Continuous variable indicating the age of the patients.
- Hypertension: Binary variable indicating if the patient has hypertension.
- Heart Disease: Binary variable showing if the patient has any heart-related diseases. Marital Status: Categorical variable indicating marital status.
- Work Type: Categorical variable indicating the type of occupation.
- Residence Type: Categorical variable differentiating between rural and urban residence.
- Average Glucose Level: Continuous variable showing the average glucose level in blood.
- BMI: Continuous variable indicating the Body Mass Index.
- Smoking Status: Categorical variable indicating the smoking status.
- Stroke: Binary target variable indicating if the patient had a stroke.

This dataset is ideal for uncovering patterns and predictors of stroke risk through detailed data analysis and visualization.

2.2 Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) was conducted to gain insights into the distribution and relationships among variables. Descriptive statistics, visualizations, and correlation analyses were employed to elucidate patterns and trends within the dataset. EDA techniques included histogram plots, count plots, and correlation matrices, providing a comprehensive overview of the dataset's characteristics.

3. Results

The Results section outlines the findings obtained from the analysis "Stroke Prediction Dataset". It offers a precise description of the experimental results, their interpretation, and the conclusions drawn from the analysis. The section provides insights into the distribution of predictive factors, the influence of different variables on stroke risk, correlations with stroke occurrence, and factors affecting the chance of having a stroke. These results are presented following each of the four research questions for this study.

3.1. Research Question 1: Distribution of Predictive Factors

To address Research Question 1, an in-depth analysis of the distribution of predictive physical health factors and demographic factors such as hypertension, heart stroke, BMI, and glucose levels among individuals in the dataset was conducted. This involved generating density plots and count plots to visualize the distribution of these variables and identify any notable patterns or trends.

3.1.1 Distribution of Numerical Predictors

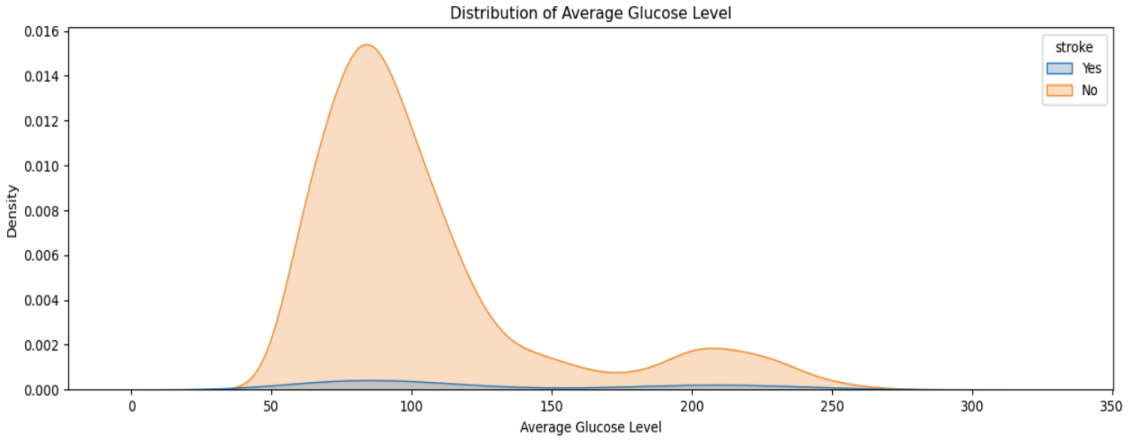
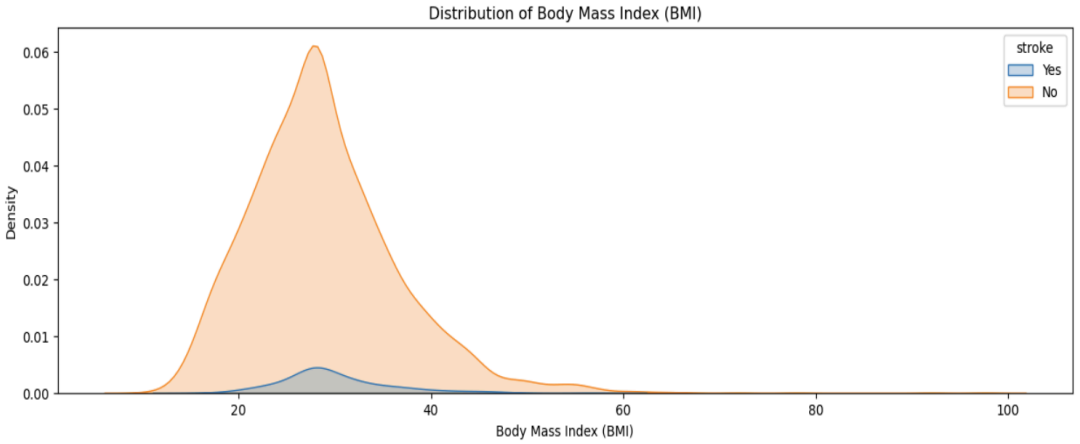


Figure 1. (a) Distribution of Average Glucose Level for Stroke and Non-Stroke Patients.

This density plot shows the distribution of average glucose levels among patients, comparing those who have had a stroke (Yes) to those who have not (No). The majority of patients have glucose levels around 100 mg/dL, with non-stroke patients having a higher density in this range. The distribution indicates higher glucose levels are less common, with a small increase in density for stroke patients at extreme low and extreme high glucose levels.



(b) Figure 1(b): Distribution of Body Mass Index (BMI) for Stroke and Non-Stroke Patients.

This density plot illustrates the BMI distribution among patients, comparing those who have had a stroke (Yes) to those who have not (No). The majority of patients have a BMI around 25. Non-stroke patients show a higher density across most BMI values, indicating they are more prevalent in the dataset.

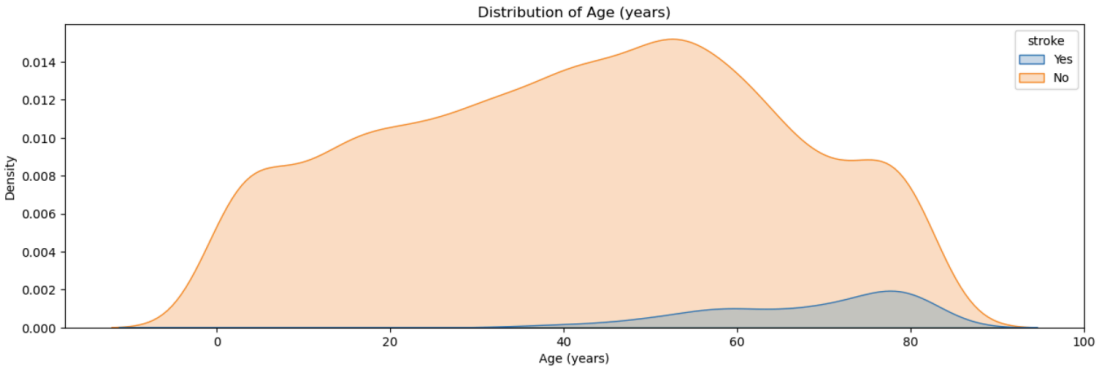


Figure 1(c): Distribution of Age for Stroke and Non-Stroke Patients.

This density plot shows the age distribution among patients, comparing those who have had a stroke (Yes) to those who have not (No). The plot indicates that the density of stroke patients increases in older age groups, particularly around 60 years and above. Non-stroke patients are more prevalent across all age ranges, with a higher density observed up to around 60 years. The distribution suggests that older age is associated with a higher likelihood of having a stroke.

3.1.2 Proportion of Categorical Variables

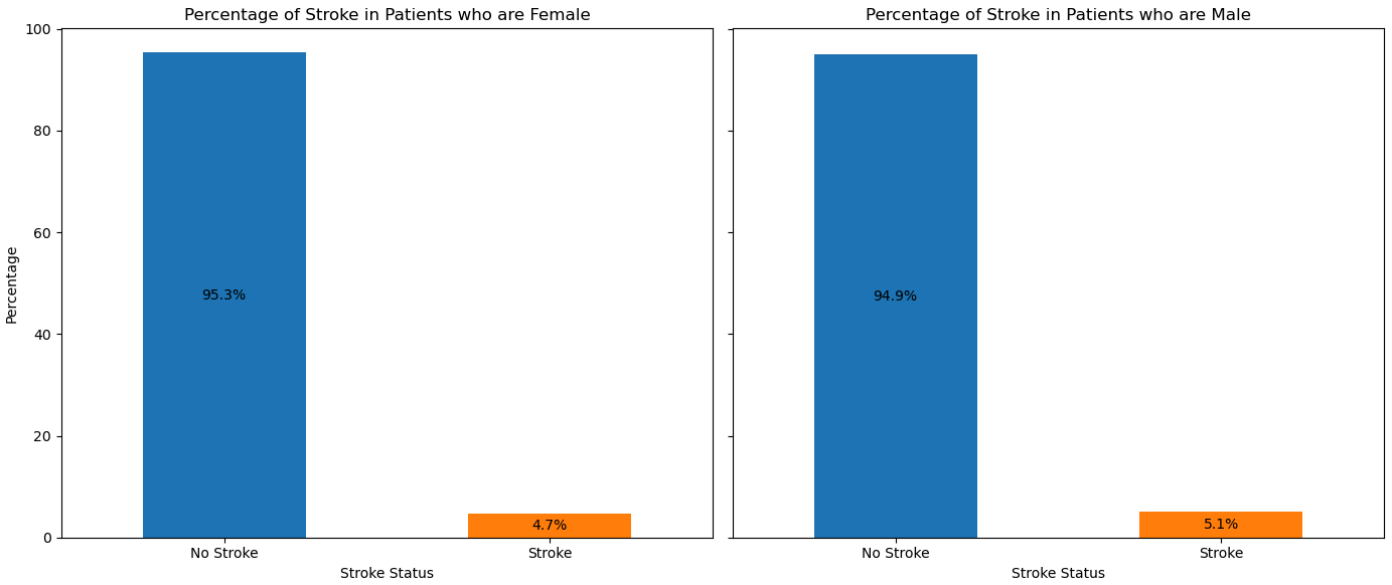


Figure 2(a): Percentage of Stroke in Patients who are Female and Male.

This bar chart compares the percentage of stroke occurrences between female and male patients. The chart on the left shows that 4.7% of female patients have had a stroke, while 95.3% have not. The chart on the right indicates that 5.1% of male patients have had a stroke, whereas 94.9% have not. The data suggests a slightly higher percentage of stroke occurrences in male patients compared to female patients.

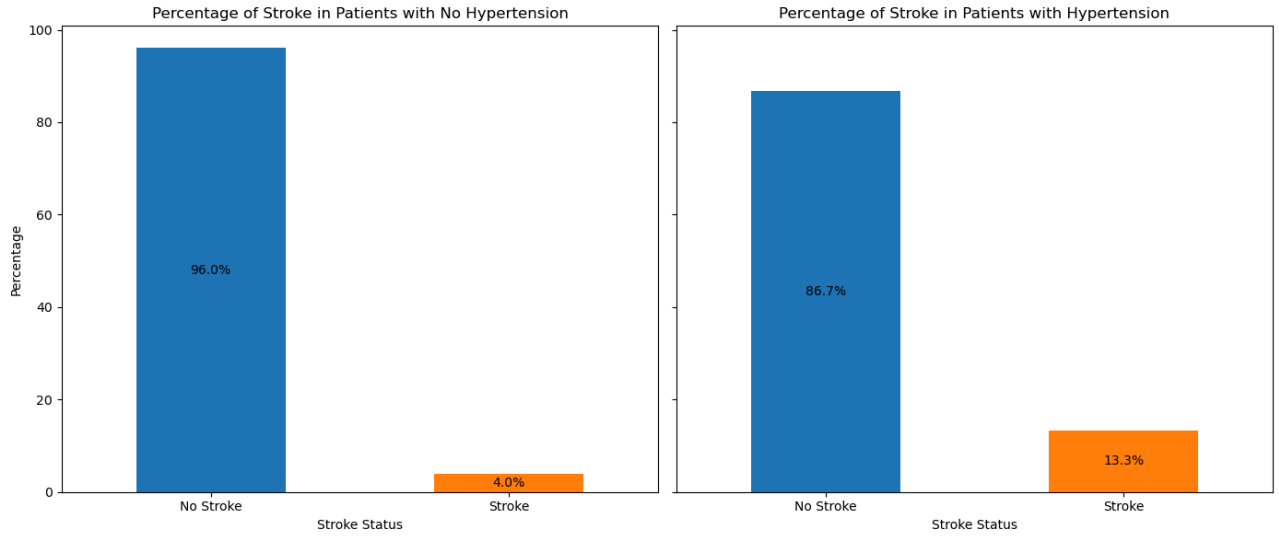


Figure 2(b): Percentage of Stroke in Patients with No Hypertension and Hypertension.

This bar chart compares the percentage of stroke occurrences between patients with and without hypertension. The chart on the left shows that 4.0% of patients without hypertension have had a stroke, while 96.0% have not. The chart on the right indicates that 13.3% of patients with hypertension have had a stroke, whereas 86.7% have not. The data suggests that hypertension is associated with a higher percentage of stroke occurrences.

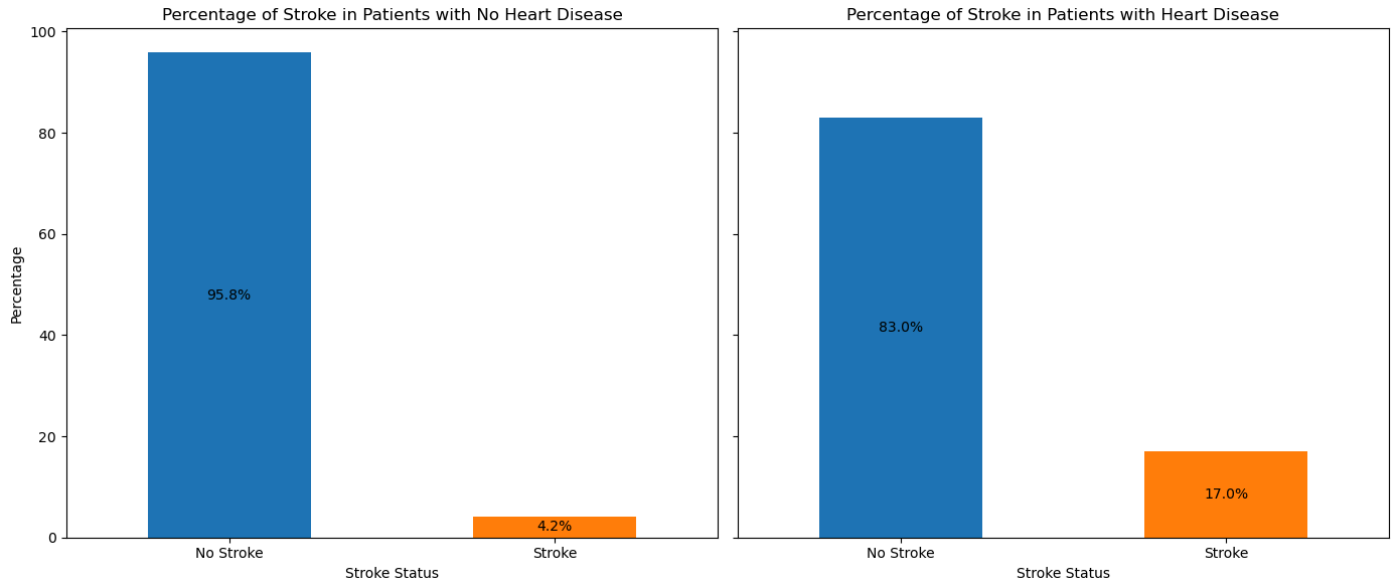


Figure 2(c): Percentage of Stroke in Patients with No Heart Disease and Heart Disease.

This bar chart compares the percentage of stroke occurrences between patients with and without heart disease. The chart on the left shows that 4.2% of patients without heart disease have had a stroke, while 95.8% have not. The chart on the right indicates that 17.0% of patients with heart disease have had a stroke, whereas 83.0% have not. The data suggests that heart disease is associated with a higher percentage of stroke occurrences.

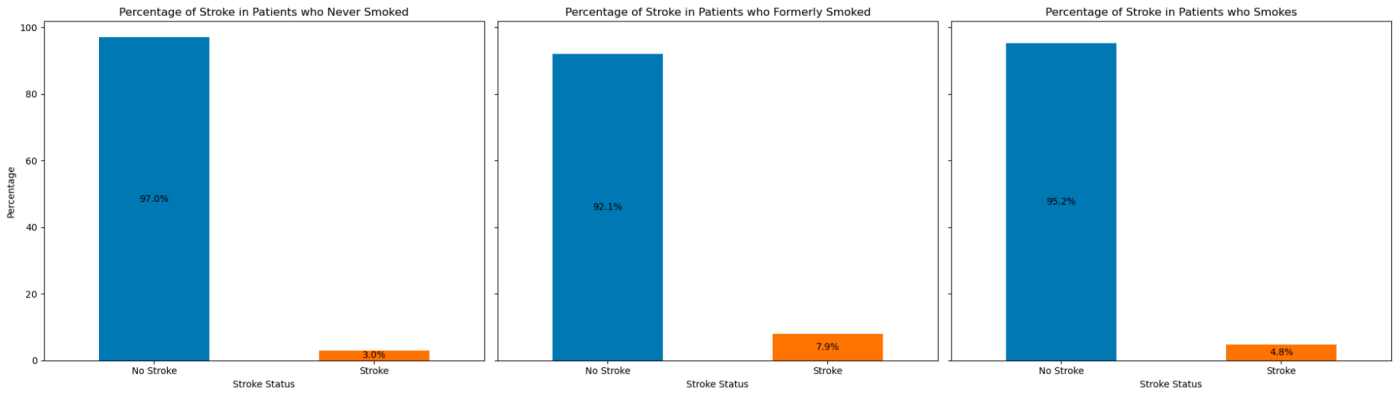


Figure 2(d): Percentage of Stroke in Patients who Never Smoked, Formerly Smoked, and Currently Smoke.

This bar chart compares the percentage of stroke occurrences among patients based on their smoking status. The chart on the left shows that 3.0% of patients who never smoked have had a stroke, while 97.0% have not. The middle chart indicates that 7.9% of patients who formerly smoked have had a stroke, whereas 92.1% have not. The chart on the right shows that 4.8% of patients who currently smoke have had a stroke, while 95.2% have not. The data suggests that former smokers have a higher percentage of stroke occurrences compared to never smokers and current smokers.

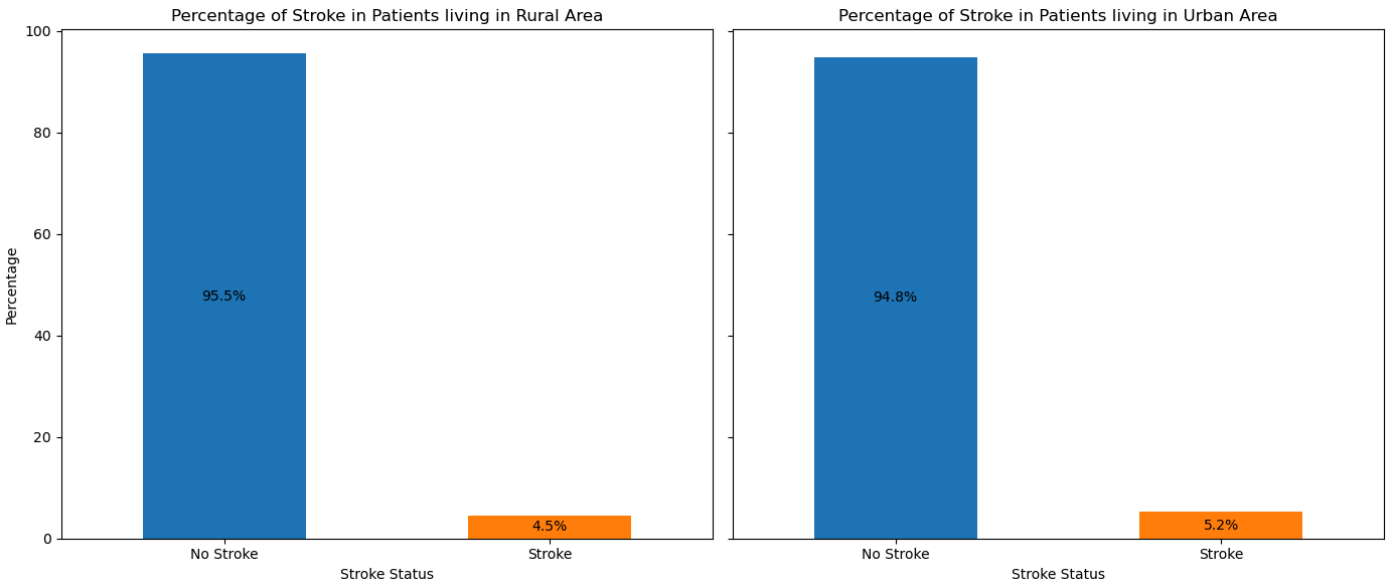


Figure 2(e): Percentage of Stroke in Patients living in Rural and Urban Areas.

This bar chart compares the percentage of stroke occurrences between patients living in rural and urban areas. The chart on the left shows that 4.5% of patients living in rural areas have had a stroke, while 95.5% have not. The chart on the right indicates that 5.2% of patients living in urban areas have had a stroke, whereas 94.8% have not. The data suggests that the percentage of stroke occurrences is slightly higher in urban areas compared to rural areas.

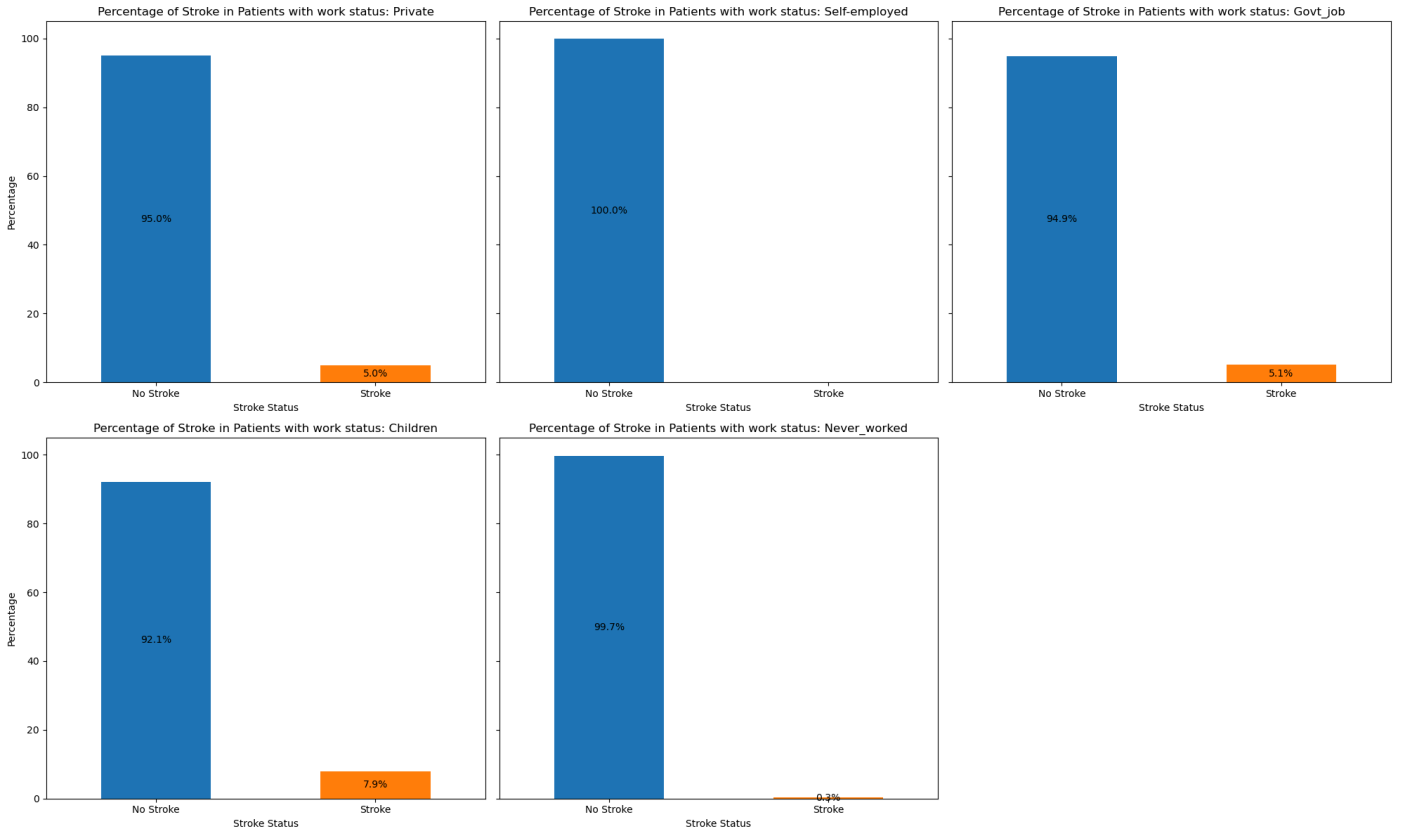


Figure 2(f): Percentage of Stroke in Patients with Different Work Statuses.

This series of bar charts compares the percentage of stroke occurrences among patients based on their work status.

- Private: 5.0% of patients in private jobs have had a stroke, while 95.0% have not.
- Self-employed: 100% of self-employed patients have not had a stroke.
- Govt Job: 5.1% of government job holders have had a stroke, while 94.9% have not.
- Children: 7.9% of children have had a stroke, while 92.1% have not.
- Never worked: 0.3% of patients who never worked have had a stroke, while 99.7% have not.

The data suggests that the highest percentage of stroke occurrences is found among children and government job holders, whereas self-employed individuals show no stroke occurrences in this dataset.

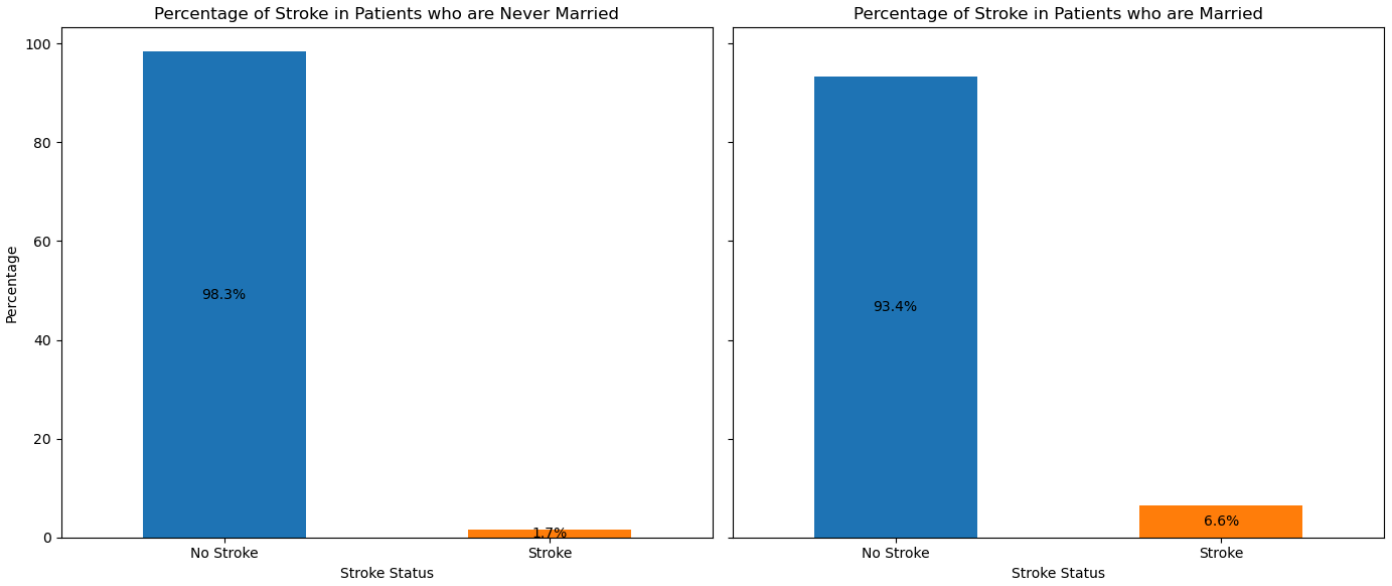


Figure 2(g): Percentage of Stroke in Patients who are Never Married and Married.

This bar chart compares the percentage of stroke occurrences between patients who are never married and those who are married. The chart on the left shows that 1.7% of patients who are never married have had a stroke, while 98.3% have not. The chart on the right indicates that 6.6% of married patients have had a stroke, whereas 93.4% have not. The data suggests that married patients have a higher percentage of stroke occurrences compared to those who are never married.

3.2.Research Question 2: Factors Influencing Stroke Risk

Research Question 2 focused on understanding how different factors interact to influence the risk of stroke in the population. Exploratory data analysis techniques, including correlation analysis and scatter plots, were utilized to examine the relationships between predictor variables and stroke occurrence. This enabled the identification of potential predictors and their relative importance in predicting stroke risk.

3.2.1 Correlation between the Numerical and Binary Predictors

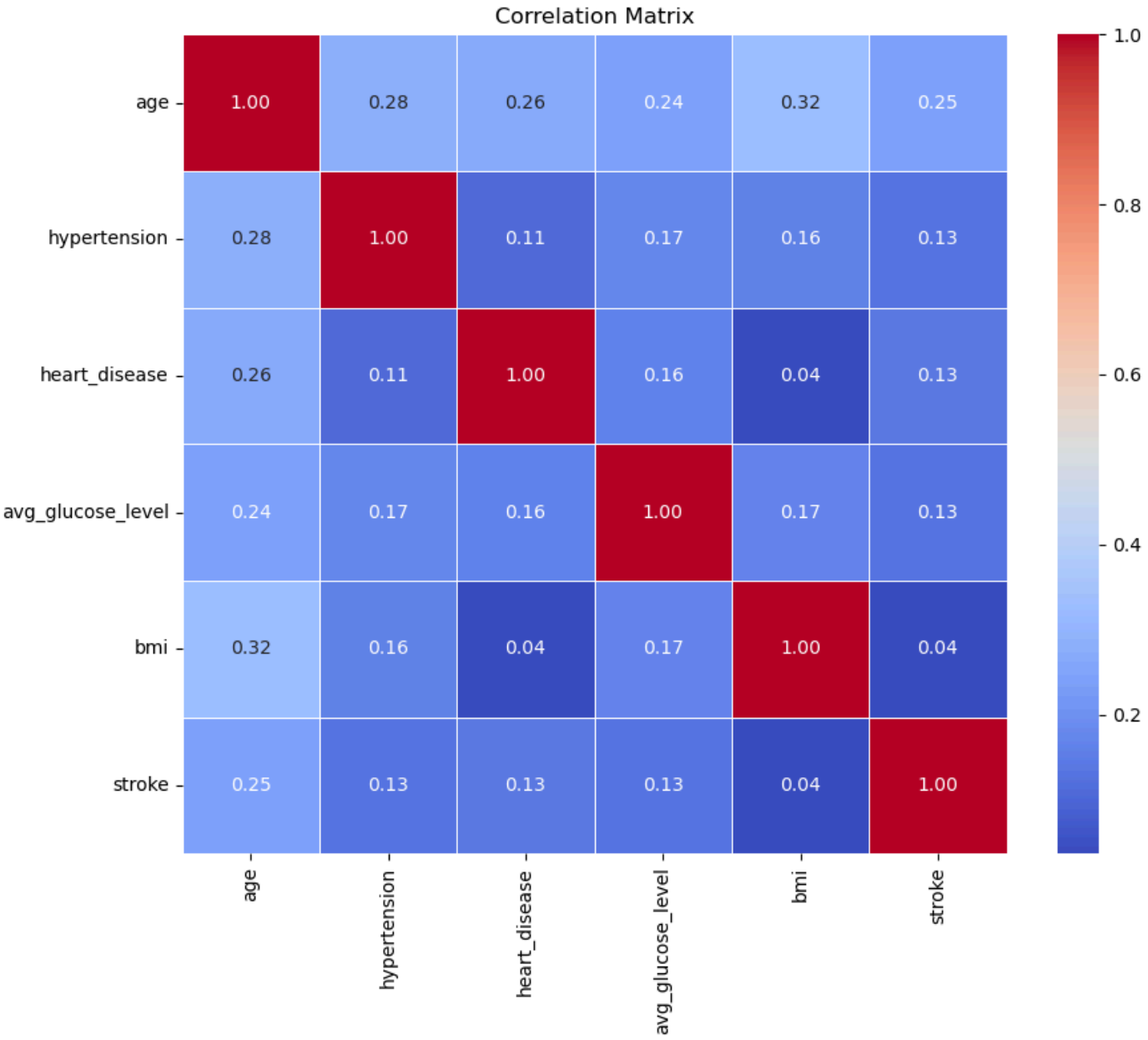


Figure 3(a): Correlation Matrix of Key Variables.

This heatmap displays the correlation coefficients between key variables in the stroke dataset, including age, hypertension, heart disease, average glucose level, BMI, and stroke.

Key Observations:

- Age has a moderate positive correlation with hypertension (0.28), heart disease (0.26), BMI (0.32), and stroke (0.25).
- Hypertension shows a weak positive correlation with stroke (0.13).
- Heart Disease also has a weak positive correlation with stroke (0.13).
- Average Glucose Level is weakly correlated with stroke (0.13) and other variables.
- BMI has a very weak positive correlation with stroke (0.04).

The matrix highlights that while there are some correlations between the predictors and stroke, none of them are very strong, suggesting that stroke occurrence is influenced by a combination of these factors rather than any single one.

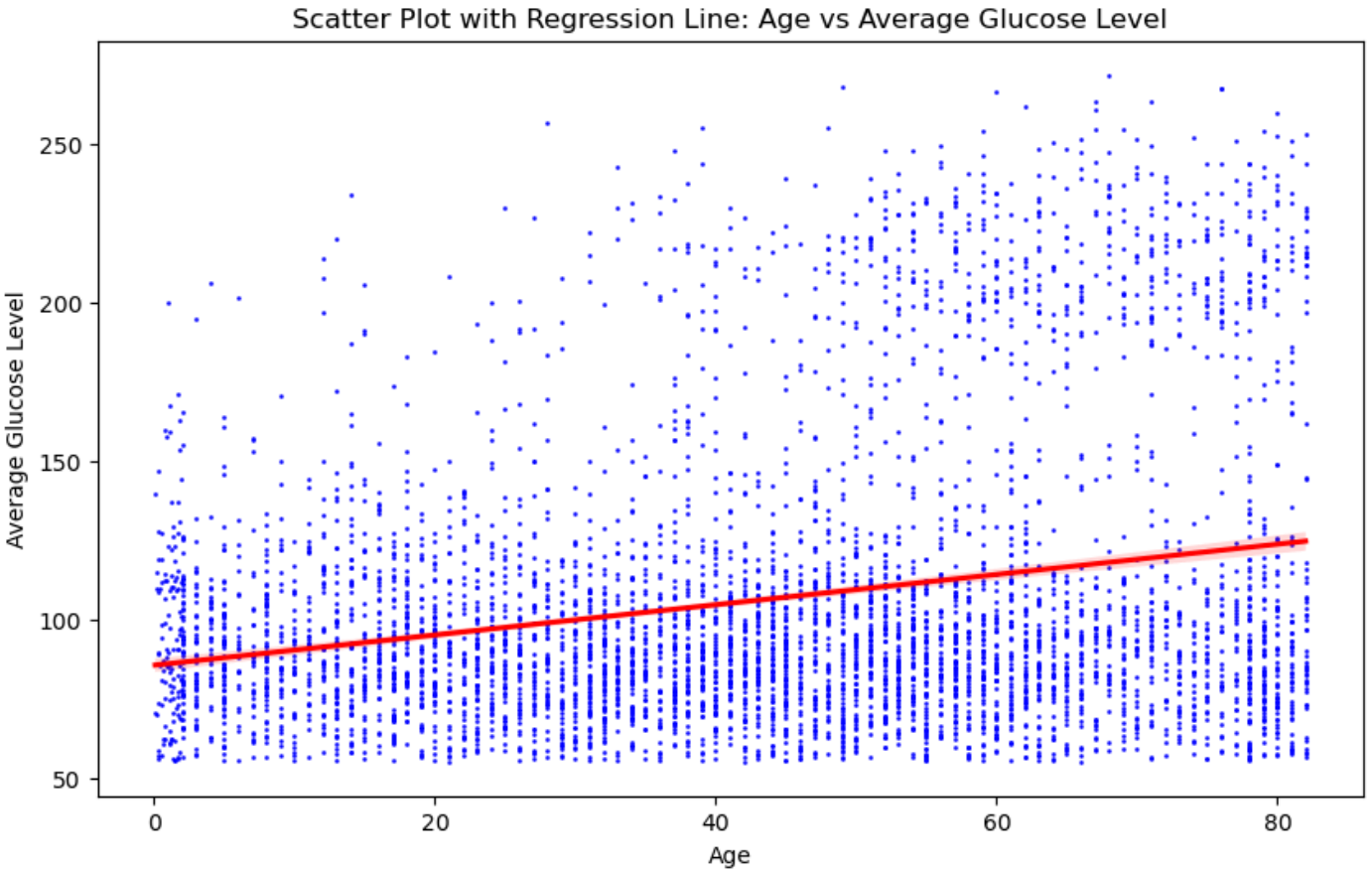


Figure 3(b): Scatter Plot with Regression Line: Age vs Average Glucose Level.

Following the correlation matrix, Figure 3(a), which highlighted the relationships between various predictors, Figure 3(b) provides a more detailed look at the correlation between age and average glucose level. This scatter plot shows the relationship between age and average glucose level, with a regression line indicating the trend. As age increases, there is a slight upward trend in average glucose levels, suggesting a positive correlation between these two variables. This aligns with the moderate correlation coefficient observed in the correlation matrix.

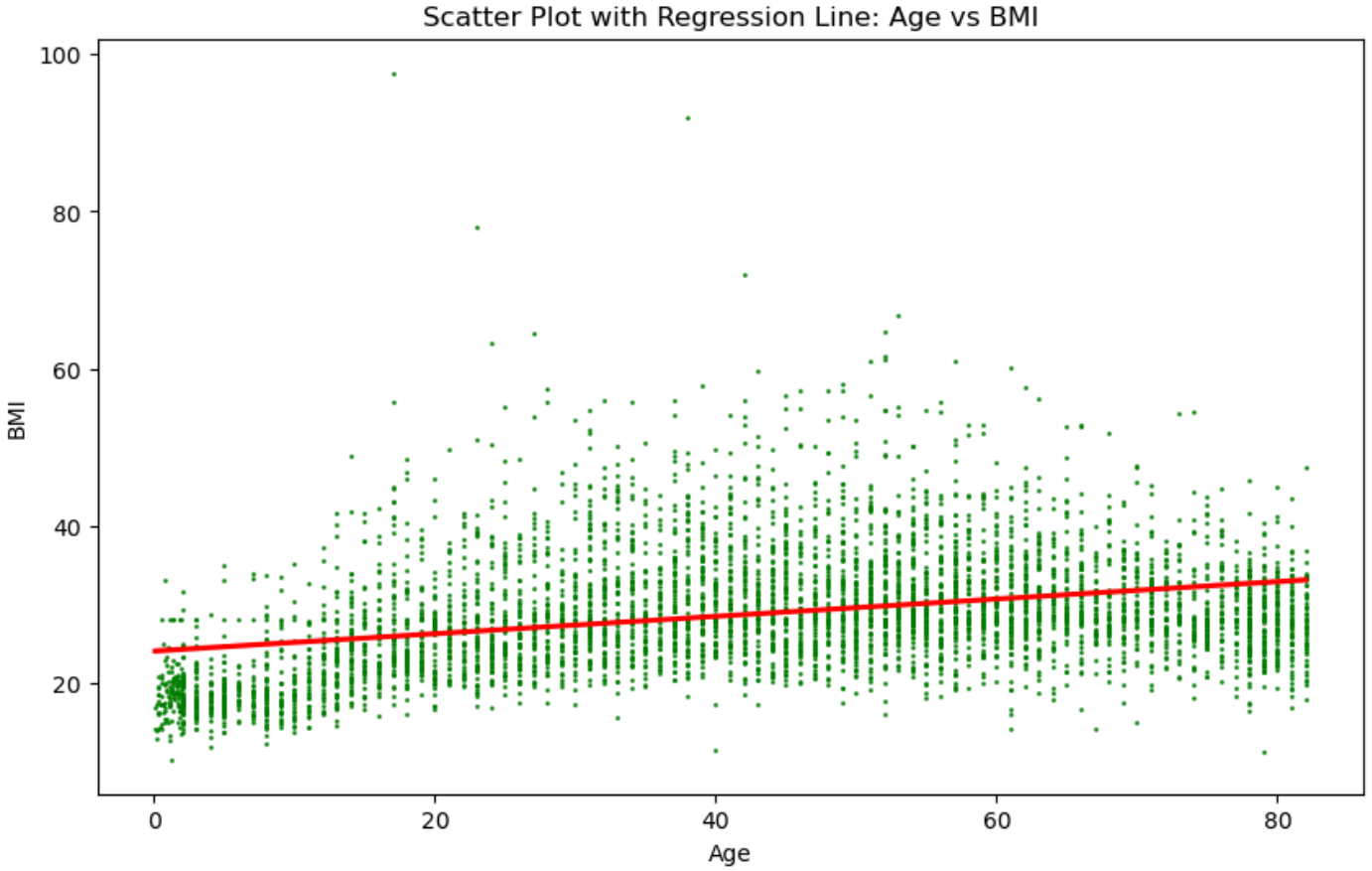


Figure 3(c): Scatter Plot with Regression Line: Age vs BMI.

Building on the previous scatter plot showing the relationship between age and average glucose level, we now examine the correlation between age and BMI. This scatter plot illustrates the relationship between age and BMI, with a regression line showing the trend. The plot reveals a slight upward trend, indicating that as age increases, BMI also tends to increase slightly. This is consistent with the moderate positive correlation between age and BMI observed in the correlation matrix.

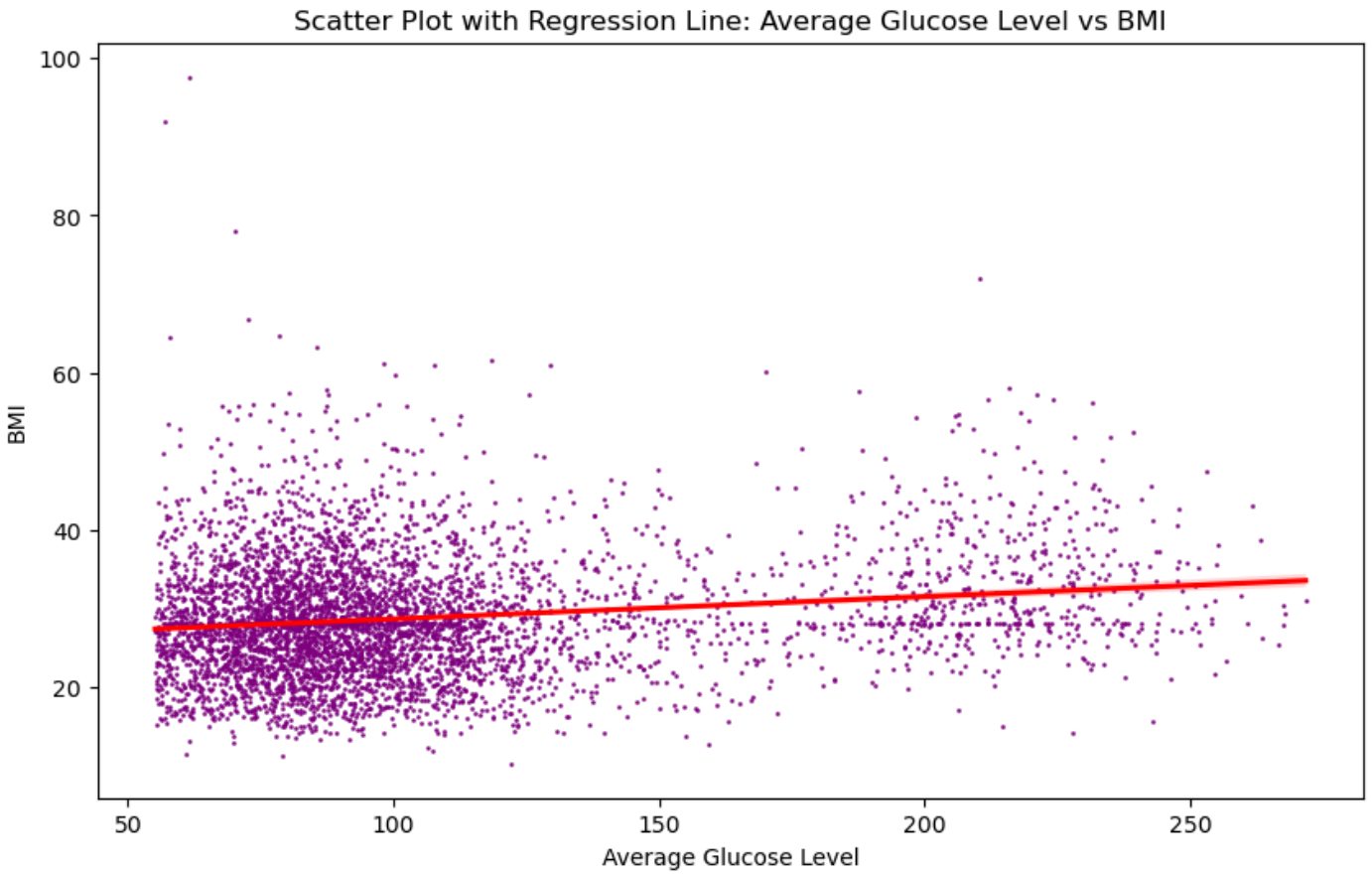


Figure 3(d): Scatter Plot with Regression Line: Average Glucose Level vs BMI.

Continuing the analysis of correlations between variables, we now explore the relationship between average glucose level and BMI. This scatter plot displays the relationship between average glucose level and BMI, with a regression line indicating the trend. The plot shows a slight upward trend, suggesting a weak positive correlation between average glucose level and BMI. This finding is consistent with the weak correlation observed between these two variables in the correlation matrix.

3.2.2 Analysis of Factors Associated with Stroke Using Chi-Square Test

Predictor	**Chi-square**	**P-value**
Hypertension	81.57	0.0000
Heart Disease	90.23	0.0000
Ever Married	58.87	0.0000
Work Type	49.16	0.0000
Residence Type	1.07	0.2998
Smoking Status	29.23	0.0000

The Chi-Square test was employed to examine the association between various categorical variables and the occurrence of stroke. Following hypothesis was used:

- **Ho: There is no association between the categorical factor and the occurrence of stroke**
- **Ha: There is an association between the categorical factor and the occurrence of stroke**

The results indicate significant associations for several variables, except for the type of residence. Here is a detailed analysis:

Hypertension

Chi-Square Statistic: 81.57

p-value: 0.0000

Interpretation: The high Chi-Square statistic and the p-value of 0.0000 indicate a statistically significant association between hypertension and stroke. Patients with hypertension are significantly more likely to have had a stroke compared to those without hypertension.

Heart Disease

Chi-Square Statistic: 90.23

p-value: 0.0000

Interpretation: Similar to hypertension, the Chi-Square statistic and the p-value suggest a strong association between heart disease and stroke. Patients with heart disease have a significantly increased risk of stroke.

Ever Married

Chi-Square Statistic: 58.87

p-value: 0.0000

Interpretation: There is a significant association between marital status and stroke occurrence. Being married or having been married is associated with a higher likelihood of stroke, potentially reflecting lifestyle or social support factors.

Work Type

Chi-Square Statistic: 49.16

p-value: 0.0000

Interpretation: The type of work individuals engage in is significantly associated with stroke risk. This could be related to stress levels, physical activity, and other job-related factors that influence health.

Residence Type

Chi-Square Statistic: 1.07

p-value: 0.2998

Interpretation: The Chi-Square statistic and p-value indicate no significant association between the type of residence (urban or rural) and the occurrence of stroke. This suggests that living in an urban or rural area does not significantly impact the likelihood of having a stroke.

Smoking Status

Chi-Square Statistic: 29.23

p-value: 0.0000

Interpretation: There is a significant association between smoking status and stroke. Smoking is identified as a risk factor for stroke, with smokers having a higher likelihood of stroke compared to non-smokers.

These results highlight the importance of managing hypertension, heart disease, and smoking to reduce the risk of stroke. Additionally, the findings suggest that marital status and work type are associated with stroke risk, which may reflect broader lifestyle and socioeconomic factors. Conversely, residence type does not appear to significantly influence stroke occurrence.

3.3.Research Question 2: Factors Influencing Stroke Risk

Research Question 2 focused on understanding how different factors interact to influence the risk of stroke in the population. Exploratory data analysis techniques, including correlation analysis and scatter plots, were utilized to examine the relationships between predictor variables and stroke occurrence. This enabled the identification of potential predictors and their relative importance in predicting stroke risk.

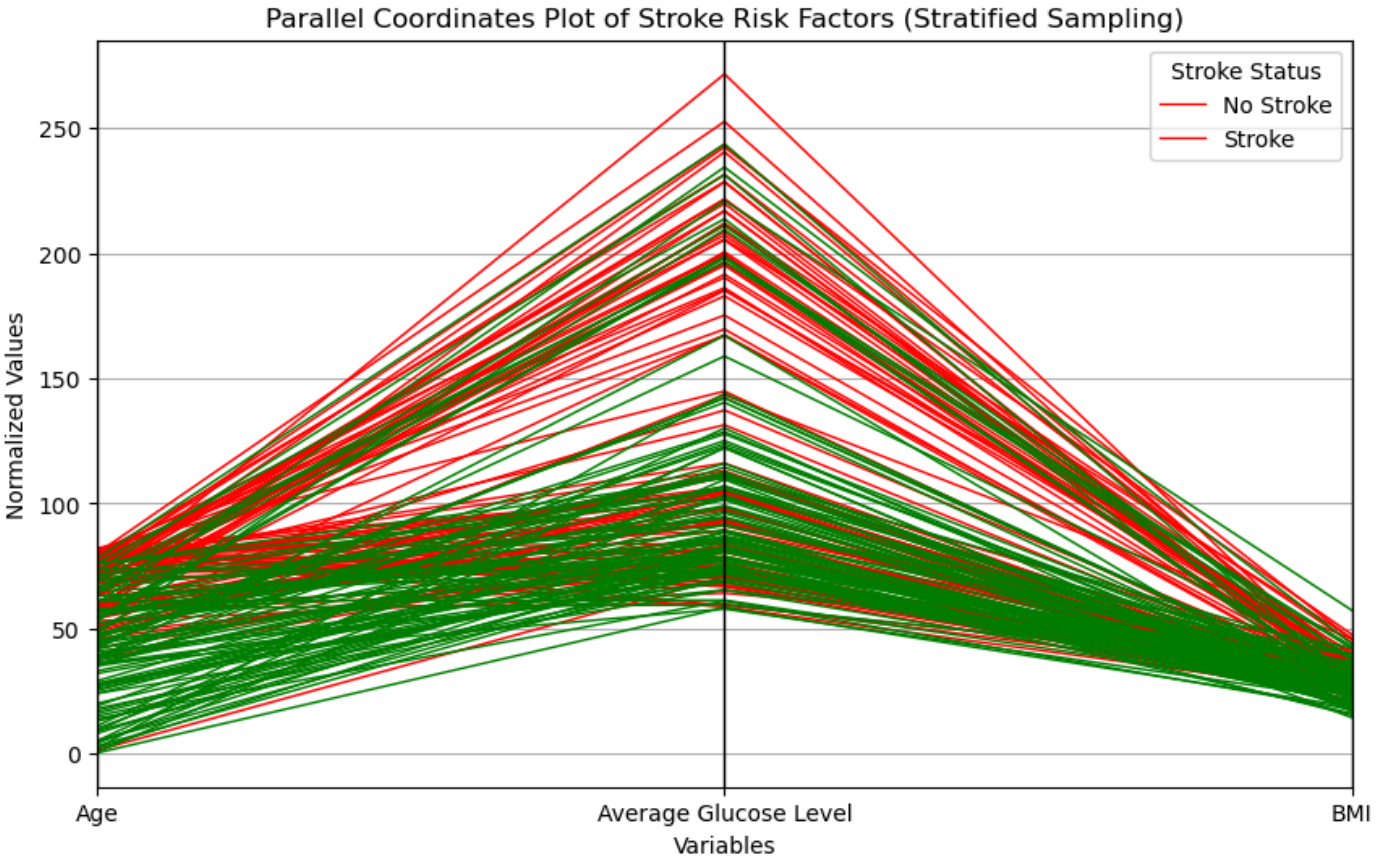


Figure 4(a): Parallel Coordinates Plot of Stroke Risk Factors (Stratified Sampling).

This plot visualizes the normalized values of age, average glucose level, and BMI for patients with and without stroke. The lines represent individual patients, with red lines indicating stroke patients and green lines indicating non-stroke patients. The plot highlights differences in these variables between the two groups.

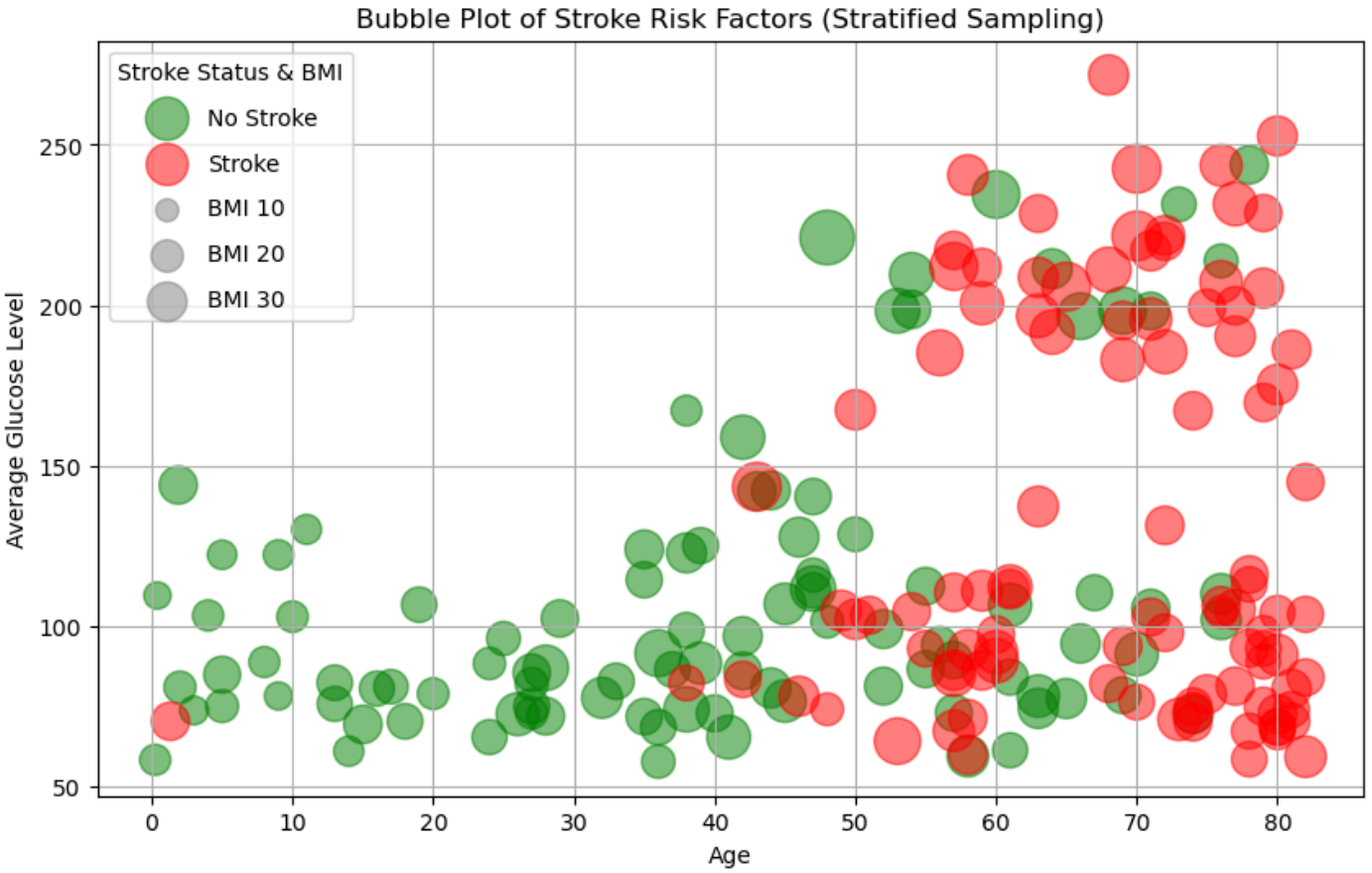


Figure 4(b): Bubble Plot of Stroke Risk Factors (Stratified Sampling).

This bubble plot shows the relationship between age, average glucose level, and BMI, with bubble size representing BMI. Red bubbles indicate stroke patients, while green bubbles indicate non-stroke patients. The plot demonstrates that stroke patients tend to have higher average glucose levels and are generally older compared to non-stroke patients.

Notably, the stroke group exhibits higher lines correlated with age, average glucose level, and BMI, indicating potential associations with increased stroke risk.

3.3.1 Hypothesis Test for Age, BMI, and Average Glucose Level

3.3.1.1 Hypothesis Test for Age Predictor:

We conducted a one-sided t-test for the 'age' predictor, with the alternative hypothesis suggesting that the mean age of the stroke group is greater than that of the non-stroke group.

- **Ho (Null Hypothesis): The mean age of individuals in the stroke group is less than or equal to the mean age of individuals in the non-stroke group.**
- **Ha (Alternative Hypothesis): The mean age of individuals in the stroke group is greater than the mean age of individuals in the non-stroke group.**

p-value: 3.71 * 10^-71

Test Stat: 18.0776

Hence, we reject Ho and conclude the following: The mean age is significantly higher for individuals who have experienced a stroke compared to those who have not.

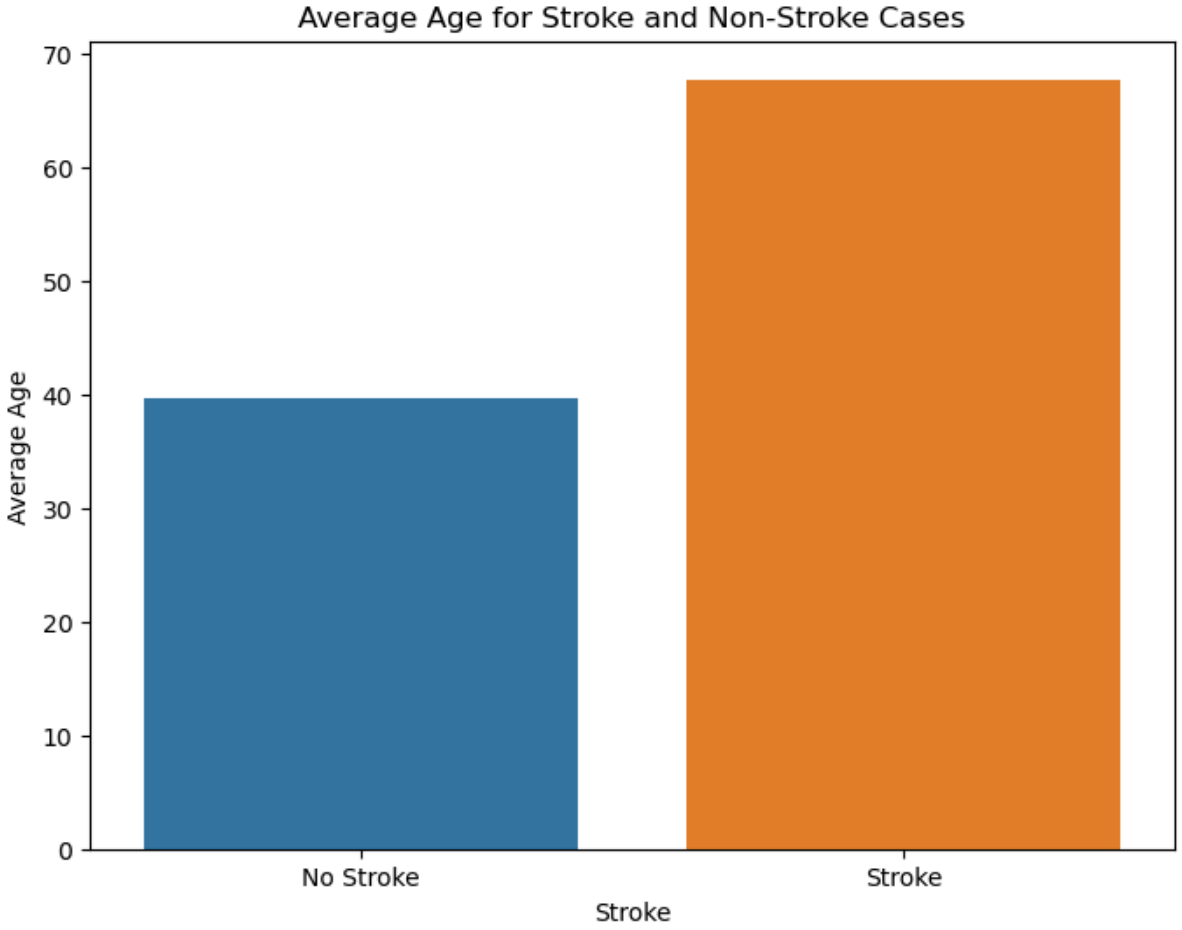


Figure 5(a): The average age of stroke patients is significantly higher than that of non-stroke patients.

3.3.1.2 Hypothesis Test for Glucose Level Predictor:

A one-sided t-test was performed for the glucose level predictor, with the alternative hypothesis proposing that the mean glucose level of the stroke group is greater than that of the non-stroke group.

- **Ho (Null Hypothesis): The mean glucose level of individuals in the stroke group is less than or equal to the mean glucose level of individuals in the non-stroke group.**
- **Ha (Alternative Hypothesis): The mean glucose level of individuals in the stroke group is greater than the mean glucose level of individuals in the non-stroke group.**

p-value: 1.35 * 10^-21

Test Stat: 9.515

Hence, we reject Ho and conclude the following: The mean glucose level is significantly higher for individuals who have experienced a stroke compared to those who have not.

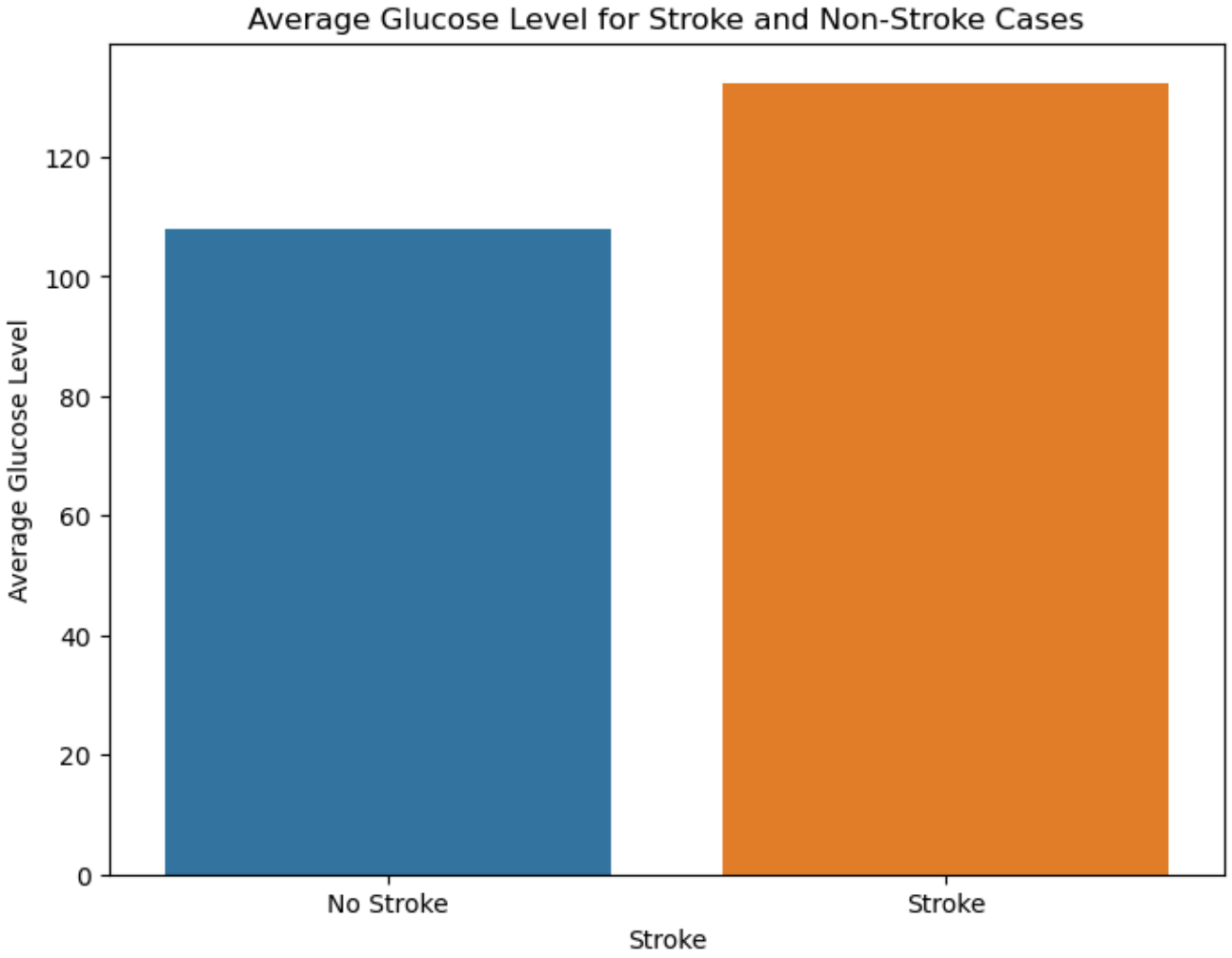


Figure 5(b): The average glucose level is higher in stroke patients compared to non-stroke patients.

3.3.1.3 Hypothesis Test for BMI Predictor:

We conducted a one-sided t-test for the BMI predictor, with the alternative hypothesis suggesting that the mean BMI of the stroke group is greater than that of the non-stroke group.

- Ho (Null Hypothesis): The mean BMI of individuals in the stroke group is less than or equal to the mean BMI of individuals in the non-stroke group.
- Ha (Alternative Hypothesis): The mean BMI of individuals in the stroke group is greater than the mean BMI of individuals in the non-stroke group.

p-value: 0.0049

Test Stat: 2.579

Hence, we reject Ho and conclude the following: The mean BMI is significantly higher for individuals who have experienced a stroke compared to those who have not.

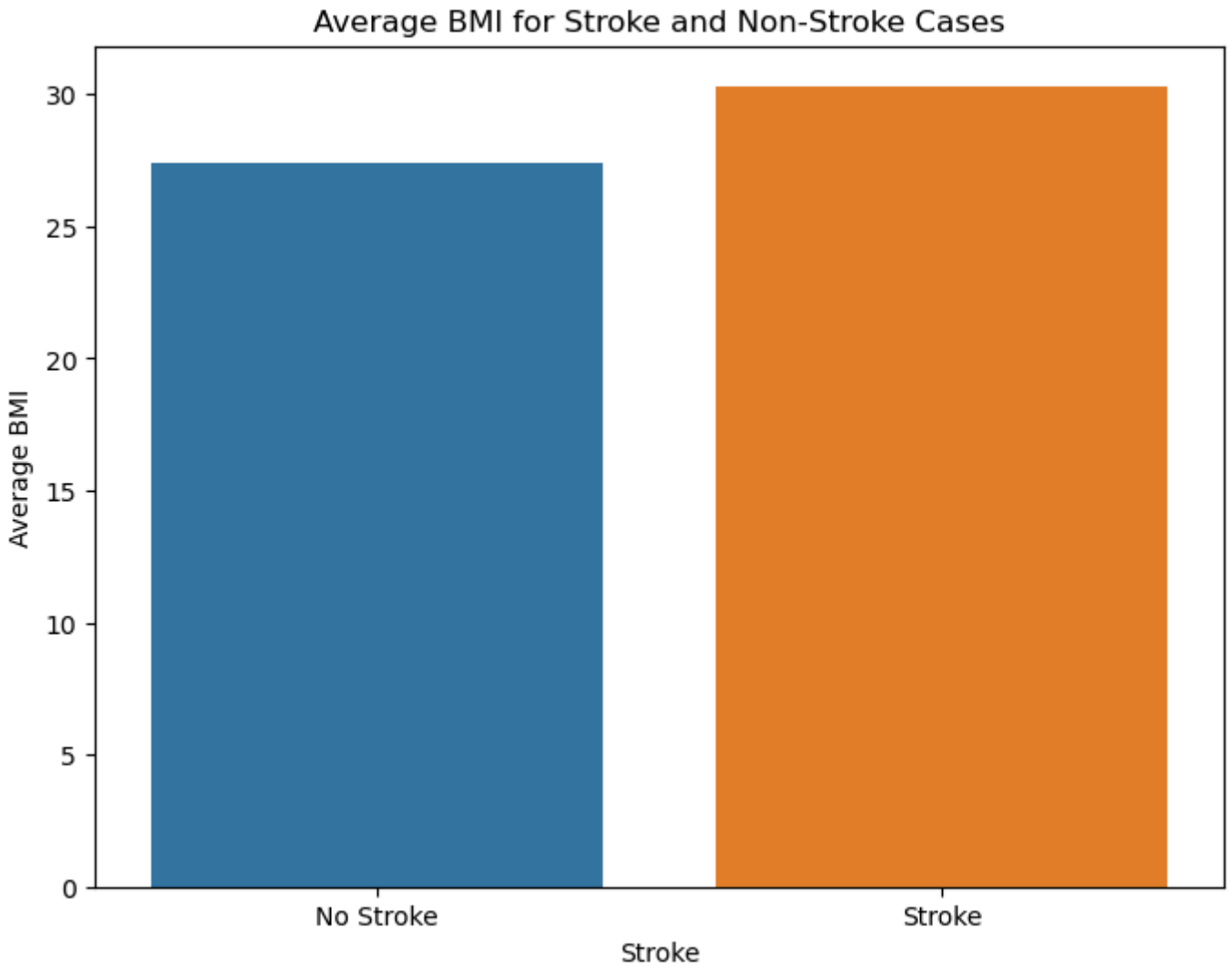


Figure 5(c): Stroke patients have a higher average BMI than non-stroke patients.

3.4.Research Question 4: Factors Affecting Stroke Chance

In Research Question 4, we aimed to understand how different factors affect the likelihood of having a stroke and how these relationships can be visually represented to aid understanding. By employing various visualization techniques and logistic regression analysis, we sought to identify and illustrate the complex interactions between predictor variables and stroke occurrence.

3.4.1 Logistic Regression

A logistic regression model was fitted with the predictors:
'Age', 'Heart Disease', 'Avg Glucose Level', 'BMI', 'Hypertension', 'Gender (Male)'

These were the results:

Predictors	**Coefficient**
Intercept	-4.1493
Age	1.7234
Average Glucose Level	0.1177
BMI	0.0153
Hypertension	0.0144
Heart Disease	0.1680
Gender (Male)	0.0123

Intercept

Coefficient: -4.1493

Interpretation: The baseline log-odds of having a stroke are very low when all predictors are zero.

Age

Coefficient: 1.7234

Interpretation: Older age significantly increases the log-odds of having a stroke.

Average Glucose Level

Coefficient: 0.1177

Interpretation: Higher glucose levels slightly increase the log-odds of having a stroke.

Body Mass Index (BMI)

Coefficient: 0.0153

Interpretation: Higher BMI has a minimal effect on the log-odds of having a stroke.

Hypertension

Coefficient: 0.0144

Interpretation: Having hypertension slightly increases the log-odds of having a stroke.

Heart Disease

Coefficient: 0.1680

Interpretation: Having heart disease increases the log-odds of having a stroke.

Gender

Coefficient: 0.0123

Interpretation: Being male has a minimal effect on the log-odds of having a stroke.

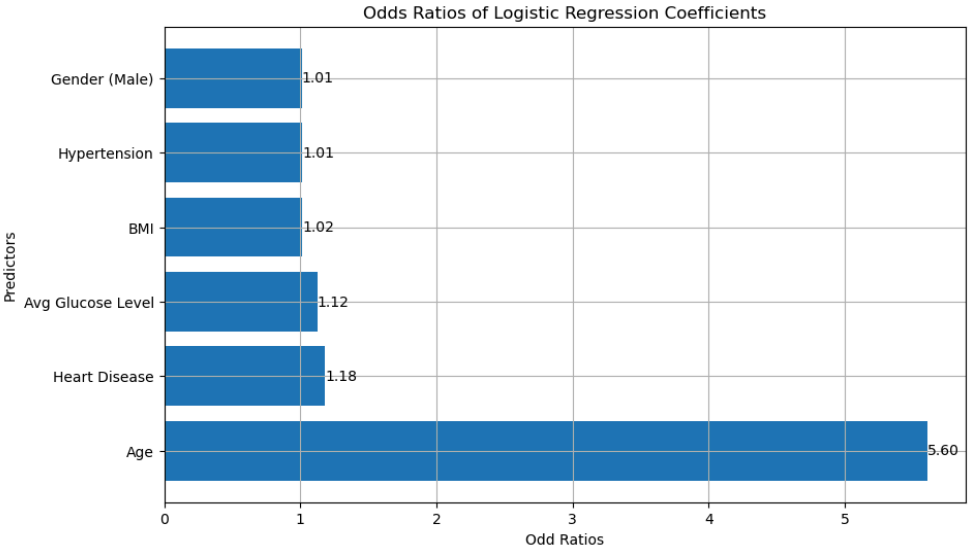


Figure 6: Odds Ratios of Logistic Regression Coefficients.

This bar chart illustrates the odds ratios for various predictors of stroke. Age has the highest odds ratio at 5.60, indicating a strong association with stroke. Heart disease and average glucose level also show significant associations with odds ratios of 1.18 and 1.12, respectively. Other factors, such as BMI, hypertension, and gender (male), have odds ratios close to 1, suggesting a weaker association with stroke.

4. Conclusions

This study aimed to uncover the factors contributing to stroke risk by analyzing the "Stroke Prediction Dataset" from Kaggle. Through exploratory data analysis, statistical modeling, and visualization techniques, we addressed four key research questions, each shedding light on different aspects of stroke risk.

We analyzed the distribution of physical health and demographic factors among individuals in the dataset. Density plots revealed that higher age, BMI, and average glucose levels were more prevalent among stroke patients. The count plots showed a higher percentage of stroke occurrences in males, individuals with hypertension or heart disease, smokers, and those living in urban areas. These findings highlight the importance of these factors in predicting stroke risk.

We used correlation analysis and scatter plots to examine how different factors interact to influence stroke risk. The correlation matrix indicated moderate correlations between age, hypertension, heart disease, BMI, and stroke. Scatter plots confirmed positive correlations between age and average glucose level, age and BMI, and average glucose level and BMI. These interactions suggest that stroke risk is influenced by a combination of these factors.

Hypothesis tests confirmed that stroke patients tend to have higher mean age, BMI, and average glucose levels compared to non-stroke patients. This supports the idea that older age, higher BMI, and elevated glucose levels are significant predictors of stroke.

Logistic regression analysis identified age, heart disease, average glucose level, BMI, hypertension, and gender as significant predictors of stroke. The odds ratios indicated that older age had the strongest association with stroke, followed by heart disease and average glucose level. Visualizations, such as the parallel coordinates plot and bubble plot, helped illustrate these relationships, making it easier to understand how these factors interact to influence stroke risk.

In summary, this study highlights the multifaceted nature of stroke risk, emphasizing the significant roles of age, heart disease, average glucose level, BMI, hypertension, and gender. By understanding these factors, we can better identify individuals at higher risk and develop targeted preventive measures. The findings underscore the importance of comprehensive risk assessment in stroke prevention and management, paving the way for more effective interventions and strategies to reduce the incidence of stroke.

Author Contributions: Sadikshya Shresha led the investigation into Research Question 1, examining the distribution of predictive physical health factors and demographic factors among individuals in the dataset. Bijay Adhikari contributed to the exploration of Research Question 2, investigating how different factors interact to influence stroke risk in a population. Anish Bhurtyal conceptualized and conducted research Question 3, focusing on the correlation between higher mean age, BMI, and glucose level with an increased risk of stroke. Saurav Dahal spearheaded Research Question 4, analyzing the impact of different factors on stroke occurrence and visualizing the findings to aid understanding.

Data Availability Statement: The dataset used in this study, titled "Stroke Prediction Dataset," is publicly available on Kaggle at the following link: Stroke Prediction Dataset

References

1. [1] Centers for Disease Control and Prevention. (n.d.). Stroke Facts. Retrieved from https://www.cdc.gov/stroke/facts.htm