ANALYTICAL AND STATISTICAL PROGRAMMING

HDS – 5310- 03

(Dr. Paula Buchanan)

FINAL PROJECT

DATE: April 29, 2025

Group 10:

Pala Jijol Onesimos - 001392437

Shreya Shrestha - 001381904

Poojani Vodapally - 001397083

STROKE AND SMOKING:

A DEADLY COMBINATION INFLUENCED BY DEMOGRAPHICS

INTRODUCTION

Background

The condition of stroke maintains its position as a major worldwide agent that causes disability and death on a large scale for millions of patients every year[1]. Stroke is a leading cause of death for Americans[1]. The stroke risk varies with race and ethnicity[2]. The United States in 2022, 1 in 6 deaths (17.5%) from cardiovascular disease were due to blockage[1]. A blockage of blood supply to the brain triggers possible brain damage, followed by functional losses, ending in severe cases with death. The identification of risk factors for stroke remains essential because stroke affects numerous people significantly and occurs frequently. Medical and lifestyle-related research shows that hypertension, along with heart diseases, diabetes, and smoking, combine with obesity to determine stroke possibilities[2]. Many reports document the established relationship between medical factors and stroke occurrence, yet our knowledge about how employment situations and settlement locations contribute to stroke remains undeveloped[3]. Investigators have been analyzing established as well as novel stroke risk factors to develop a deeper knowledge about stroke forecasting.

1. Demographics and Stroke Risk

Stroke risk increases with age, especially after 55[2]. Men have a higher risk at younger ages, but women have a greater lifetime risk[2]. African Americans, Hispanics, and some Asian populations face higher stroke rates due to common health conditions[4]. The family history of stroke increases susceptibility[2]. Lower socioeconomic status contributes to stroke risk due to healthcare access, stress, and related health issues[3].

2. Lifestyle Factors (Smoking and Stroke Risk)

Smoking is a major modifiable stroke risk factor[5]. Nicotine raises blood pressure, increasing stroke likelihood[5]. Smoking damages blood vessels, making blockages more likely[5]. It also promotes clot formation, leading to ischemic strokes[5] Carbon monoxide from smoking reduces oxygen in the blood, straining the cardiovascular system[5].

Population of Interest

The population of interest includes adults, especially after 55 are at risk of stroke, particularly those influenced by demographics (age, sex, and socioeconomic status) and smoking status. Older adults, men, and racial/ethnic minorities face higher stroke risks due to common health conditions. Lower socioeconomic status contributes to increased risk through limited healthcare access and higher stress levels. Smokers have a significantly elevated stroke risk due to their impact on blood pressure, blood vessels, and clot formation. Understanding these groups helps in developing targeted prevention strategies.

Outcome of Interest

The research determines stroke occurrence as the main outcome by using binary formatting (1 = stroke status and 0 = no stroke). The outcome of interest in this study is the likelihood of having a stroke, particularly as influenced by demographic factors and smoking status. By identifying high-risk groups, healthcare providers can implement early screening programs, smoking cessation support, and educational campaigns to reduce stroke incidence. This research can also inform public health policies aimed at reducing health disparities and improving access to stroke prevention resources. Ultimately, understanding these relationships can lead to better health outcomes and lower stroke-related mortality and disability rates.

Independent Variables and Their Importance

Several independent variables have been chosen to analyze stroke risk factors according to their likely significance in stroke development. These variables include age, sex, race/ethnicity, socioeconomic status, and smoking status. Examining these factors together provides a comprehensive understanding of how demographic and lifestyle factors interact to influence stroke risk.

Demographics:

Age: Stroke risk increases significantly with age, particularly after 55, due to arterial stiffness and a higher prevalence of conditions like hypertension and diabetes. The evidence shows that age functions as a significant risk factor for stroke because older individuals face higher chances of getting cerebrovascular events

Sex: Men have a higher stroke risk at younger ages, while women have a greater lifetime risk, partly due to hormonal differences and longevity.

Hypertension: The most significant marker for stroke development is elevated blood pressure, which leads to blood vessel constriction as well as vascular wall damage.

Heart Disease: Stroke occurrence is greater in individuals with cardiovascular problems because their blood circulation and clot formation are negatively affected.

Average Glucose Level: Elevated glucose levels, which are commonly found in diabetic patients, lead to blood vessel damage that negotiates stroke occurrence.

Body Mass Index (BMI): A connection exists between obesity and stroke occurrence because hypertension, diabetes, and cardiovascular diseases form an association with this condition.

Socioeconomic Status: Lower income and education levels are linked to higher stroke risk due to factors like poor healthcare access, increased stress, and unhealthy lifestyle habits.

Smoking Status: Smoking is a modifiable risk factor that significantly increases stroke risk. It raises blood pressure, damages blood vessels, promotes clot formation, and reduces oxygen levels. Smokers have a much higher likelihood of both ischemic and hemorrhagic strokes compared to non-smokers. The combination of smoking with high-risk demographic factors further amplifies the chances of stroke.

Work Type and Residence Type: Research on socioeconomic factors appears infrequently, although such studies reveal vital risk factors for stroke. The type of work impacts patient stress and medical service availability, and location of residence controls both healthcare service exposure and community danger and daily life opportunities

Existing Research and Literature Review

Previous studies about stroke prediction mostly analyzed standard risk elements, which include hypertension combined with diabetes and lifestyle behaviors such as smoking and alcohol use (2,4). The evidence reveals that people with high blood pressure face an elevated danger of suffering from stroke type, specifically including ischemic stroke, because of blood clot development[2]. Scientists have confirmed that the risks of cardiovascular diseases and strokes are strongly linked because these conditions share fundamental disease processes[2]. Research on smoking behaviors, together with BMI, reveals their established correlation with stroke development[2]. Research on medical and lifestyle-related determinants exists in large numbers, but analyses of socioeconomic factors, especially employment type and living environment, remain scarce[2]. The following existing studies collectively highlight the significant impact of demographic factors and smoking status on stroke risk, underscoring the importance of targeted prevention strategies such as: The existing literature highlights key factors consistently associated with increased stroke risk. The Framingham Heart Study was foundational in establishing that common cardiovascular risk factors—including hypertension, smoking, diabetes, and pre-existing heart disease—substantially raise the likelihood of experiencing a stroke[2]. It emphasized the interplay between these health conditions in contributing to cerebrovascular events[2]. Expanding this understanding on a global scale, the INTERSTROKE study demonstrated that hypertension, smoking, diabetes, and certain cardiac conditions are among the most significant contributors to stroke risk, although the prevalence and impact of these risk factors vary by region and socioeconomic status[4]. Furthermore, a targeted analysis within the INTERSTROKE project focused on tobacco use found a strong association between current smoking and increased risk for both ischemic and hemorrhagic strokes, with the risk appearing even more pronounced in high-income countries.[3] *[Citation: Wang et al., 2024]* Collectively, these studies underscore the importance of addressing both medical and lifestyle factors, particularly smoking and hypertension, in comprehensive stroke prevention efforts across diverse populations.

Research Gap and Contribution of This Study

Our study addresses a key literature gap by examining the combined effects of demographics and smoking status on stroke risk, an area often studied separately. Unlike previous research that primarily focuses on high-income populations, our project will analyze diverse socioeconomic groups using Stroke Prediction data. By exploring how socioeconomic status influences smoking-related stroke risk, our study provides a more comprehensive understanding. This research can help develop targeted prevention strategies, improve public health policies, and reduce stroke-related disparities.

Health Data Source

The health data source planned for this study is the Brain Stroke dataset from Kaggle[5]. This data set provides comprehensive health and lifestyle data, including demographic variables, smoking status, and stroke history. This data set is ideal because it includes a nationally representative sample, allowing for a robust analysis of stroke risk factors. Here is an Attribute Information based on data related to stroke risk factors:

Table 1: Attribute Information

S.N.	**Risk Factors**	**Classification**
1	Gender	“Male", "Female", "Other".
2	Age	Age of the patient
3	Hypertension:	0 if the patient doesn't have hypertension, 1 if the patient has hypertension
4	Heart disease:	0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease
5	Ever married	“Yes" or "No"
6	Work type	"Government job", "Never worked", "Private" "Self-employed"
7	Residence type:	"Rural", "Urban"
8	Avg glucose level	Average glucose level in blood
9	BMI	Body Mass Index
10	Smoking status	"Formerly smoked", "Never smoked", "Smokes" or "Unknown"*
11	Stroke	1 if the patient had a stroke, 0 if the patient never had a stroke

Research Question and Significance

The research question guiding this study is "How do demographic factors and smoking status impact the likelihood of having a stroke?" This research is significant because it can inform targeted stroke prevention efforts, including early screening programs, smoking cessation initiatives, and health policies designed to reduce disparities in stroke risk. The practical implications of this study include improving public health interventions, guiding healthcare resource allocation, and ultimately reducing stroke-related morbidity and mortality. The research findings will support public health programs in workplaces as well as guide urban development policies and enhance medical facility accessibility. The combination of clinical elements with lifestyle patterns, together with socioeconomic variables within this study, functions to advance stroke risk factor knowledge, which generates practical guidance for healthcare providers and policy creators. This research aims to develop better strategies for stroke prevention, which would help decrease morbidity and mortality numbers in patients with this condition.

METHODOLOGY

Data Description

We plan to use the Brain Stroke Data that is available on Kaggle that is published by Jillani Soft Tech (2021) to assist in stroke prediction modeling and health research and the data is available from publicly accessible sources as of 2024. This dataset contains health and lifestyle information of adults, including both medical conditions and demographic variables, to predict the occurrence of stroke.

The dataset includes individuals aged 18 and over from diverse socioeconomic backgrounds and provides a comprehensive overview of stroke risk factors including demographic, clinical, and lifestyle characteristics but it does not include specific geographic location identifiers. The data is publicly accessible and was compiled to help researchers and data scientists analyze stroke risk factors across a population. The dataset includes information about the age, gender, hypertension, heart disease, marital status, work type, dwelling type (urban or rural), average glucose, body mass index (BMI), smoking, and stroke of each individual.

This data is suitable for our research because it has the variables needed to examine both modifiable (e.g., smoking status, hypertension) and non-modifiable (e.g., sex, age) stroke risk factors. It is suitable for our research purpose of exploring how demographic factors and smoking status interact to affect the likelihood of stroke occurrence.

Variables

The variables that we plan to choose to use in our analysis are:

1. **Dependent Variable:**

- Stroke: Binary variable (1 = Stroke occurred, 0 = No stroke)

2. **Independent Variables:**

- Hypertension (Binary: 1 = Yes, 0 = No)
- Heart Disease (Binary: 1 = Yes, 0 = No)

3. **Covariates:**

- Age (Continuous)
- Gender (Categorical: Male, Female, Other)
- Smoking Status (Categorical: Smokes, Formerly Smoked, Never Smoked)
- BMI (Continuous, kg/m²)
- Average Glucose Level (Continuous, mg/dL)
- Work Type (Categorical: Private, Self-employed, Government Job, Never Worked)
- Residence Type (Categorical: Urban, Rural)
- Socioeconomic Status:
    - Work Type
    - Residence Type
    - Marital Status

Descriptive Statistics

We plan on using descriptive statistical analysis to summarize the descriptive features of our study population and to create an initial understanding of the data before performing any inferential analysis.

For the continuous variables, the choice between reporting mean and standard deviation (SD) or median and interquartile range (IQR) depends on the distribution of the data. If the data are normally distributed, the mean and SD are appropriate because they accurately describe the center and spread. However, if the data are skewed or have outliers, the median and IQR should be reported, as they are less affected by extreme values. Researchers often use visual methods like histograms or formal tests such as the Shapiro-Wilk test to assess normality[6]. Correctly matching the summary statistics to the data distribution ensures accurate interpretation.

For categorical variables, we plan to report the frequency (n) and percentage (%) by category. This will allow us to describe the sample structure and establish the prevalence of stroke and its demographic and lifestyle determinants in the population.

Data Management

We plan to perform data management using RStudio so that the dataset is suitable for statistical analysis. This includes missing data checks, identification of outliers, and verification of small group sizes in categorical variables.

Missing Data:

We plan to use the summary () and colSums(is.na ()) commands in RStudio to check whether there are any missing values on any of the variables. From our initial glance, there are no missing values in the dataset. We will, however, double-check this in RStudio and again after any data transformation. If we do have missing values, our plan is to:

1. Use mean or median imputation for continuous variables, depending on distribution.
2. Use mode imputation or case wise deletion for categorical variables, as appropriate depending on the level of missingness.

Outliers:

We plan to detect outliers for continuous variables (e.g., Age, Blood Pressure) using boxplots and Z-scores (values > 3 standard deviations from the mean). Outliers will be carefully examined and either corrected or removed depending on their impact on analysis.

Small Group Sizes:

For categorical variables, we will examine the distribution of sample sizes in each category. If a category has a very small sample size (less than 10% of the total sample), we will consider combining categories or using exact tests like Fisher’s exact test for analysis.

Bivariate Analysis Plan

We will conduct bivariate analyses to examine the associations between the dependent variable (stroke) and the respective independent variables. Based on the type of data and our research objectives, we have selected two appropriate statistical tests: the Chi-Squared Test of Independence and the Independent Samples t-Test.

1. Chi-Squared Test of Independence We will apply the Chi-Squared test to determine if there is a statistically significant relationship between the binary outcome variable stroke and the categorical variables: smoking_status, hypertension, and work_type. The test is suitable since it quantifies the association between two categorical variables and is common in public health studies with binary outcomes like disease status.

We will ensure that the Chi-Square test assumptions are met by checking that expected cell frequencies are 5 or more in at least 80% of the contingency table cells. We will, in case of a violation, use Fisher's Exact Test for 2x2 tables as a replacement.

2. Independent Samples t-Test We plan on using the independent samples t-test to compare the means of the continuous variables (avg_glucose_level and age) in two groups: individuals who have had a stroke and individuals who have not. This will enable us to see if stroke incidence is associated with statistically significant differences between these continuous predictors.

We shall check the assumptions of normality using the Shapiro-Wilk test and equality of variance using Levene's Test. If the assumptions are not met, we shall use the non-parametric Mann-Whitney U test as an alternative.

If significant Chi-Square results are found for categorical variables with more than two categories (such as **smoking status** or **work type**), we will perform **post-hoc pairwise comparisons** using the **Bonferroni correction**. This approach helps control Type I error and ensures the statistical validity of results while accounting for multiple comparisons.

Multivariate Analysis

We plan to use binary logistic regression to analyze the relationship between various demographic and lifestyle factors and the likelihood of experiencing a stroke. This method is appropriate because the outcome variable stroke occurrence is binary (yes/no). Logistic regression allows us to evaluate multiple predictors simultaneously and determine their individual impact on stroke risk. The model will generate odds ratios to help interpret how each variable affects the probability of having a stroke. This analysis will include key factors such as age, gender, BMI, glucose levels, smoking status, and more. The findings will guide targeted interventions and support data-driven public health strategies**.** We plan to include the following predictors because they have been shown in past research and clinical studies to be important factors related to stroke risk.

1. Age (continuous)
2. Gender (categorical: male/female)
3. Hypertension (binary)

4. Heart Disease (binary)
5. Average Glucose Level (continuous)
6. Body Mass Index (BMI) (continuous)
7. Smoking Status (categorical: never, former, current)
8. Work Type (categorical)
9. Residence Type (urban/rural)
10. Marital Status (binary: ever married or not)

Assumptions

Before interpreting the results, several assumptions of logistic regression will be evaluated:

1. Binary Outcome: Stroke status meets this condition.
2. Independence of Observations: Assumed from the individual-level dataset.
3. Linearity of Continuous Predictors in the Logit: Checked using methods like the Box-Tidwell test.
4. Multicollinearity: Variance Inflation Factor (VIF) will be used to detect correlation among predictors.
5. Sufficient Sample Size: Ensuring a minimum of 10 cases per predictor to maintain model reliability.

If assumptions are not met:

1. Non-linearity in predictors will be addressed through transformations or categorization.

2. Multicollinearity will be handled by removing or combining highly correlated variables.
3. If logistic regression proves unsuitable, non-parametric models like decision trees or random forest may be considered.

Tables to Include in the Report

Table 2: Descriptive Summary of Sample

S.N.	Variable	Category	n (%)	SD	Median (IQR)
1	Age	Continuous			
2	Gender	Male, Female Other			
3	Hypertension	0 = No, 1 = Yes			
4	Heart Disease	0 = No, 1 = Yes			
5	Ever Married	Yes, No			
6	Work Type	Private Govt Self Never Worked			
7	Residence Type	Urban, Rural			
8	Avg Glucose Level	Continuous			
9	BMI	Continuous			
10	Smoking Status	Never Formerly Currently Unknown			
11	Stroke	0 = No 1 = Yes			

Table 3: Bivariate Analysis

S.N.	Variable		Stroke - Yes (%) or Mean	Stroke - No (%) or Mean	P-value	Significance
1	Smoking Status	-Smokes -Formerly smoked - Never Smoked				
2	Hypertension	-Yes -No				
3	Work Type	-Private- -Self Employed -Government Job -Never worked				
4.	Age					
5.	Avg Glucose Level					

Table 4: Logistic Regression Output

S.N.	Predictor	Odds Ratio (OR)	95% CI	p-value
1	Age			
2	Hypertension			
3	Smoking Status			
4	Work Type			

Visualizations for the Report

To enhance the understanding of how various factors relate to stroke risk, several visualizations will be included in the report. A bar chart comparing stroke prevalence across smoking categories (never smoked, formerly smoked, and currently smoking) will highlight the disproportionate burden of stroke among smokers, emphasizing the critical role of smoking as a modifiable risk factor. Boxplots for age, average glucose level, and BMI by stroke status will illustrate the distribution and central tendencies of these continuous variables among individuals with and without a stroke. This allows for a visual comparison of whether those who experienced a stroke tend to exhibit higher or lower values in these risk-related metrics. A stacked bar chart depicting stroke status across different work types (e.g., private, self-employed, government, never worked) will shed light on how occupational status may intersect with health outcomes, pointing to socioeconomic disparities in stroke incidence. Lastly, a heatmap displaying correlations among continuous variables (such as age, BMI, and glucose level) will be used to explore underlying relationships that may influence stroke risk, offering insight into variable interdependence prior to regression analysis. These visual tools are chosen not for diagnostic checks but to support clearer communication of patterns in the data and provide meaningful context for interpretation.

RESULTS

1. Descriptive Statistics:

The dataset included 4,981 participants. The mean age was 43.2 years (SD = 22.6), mean glucose level was 106.2 mg/dL (SD = 45.3), and the mean BMI was 28.9 kg/m² (SD = 7.8). Females represented 58.4% of the sample. Approximately 9.6% had hypertension, and 5.5% had heart disease.

Table 5: Characteristics of Study Participants (N = 4,981)

S.N.	**Variable**	**Category**	**n (%)/mean**	**Mean(SD)**	**Median (IQR)**
1.	Age	Continuous (years)		43.42 years (22.61)	45

2.	Gender	Male
		Female
		Others
			2074(41.64 %)
			2907(58.36 %)		

3.	Hypertension	0 = No
		1 = Yes
			4502(90.38%)
			479(9.62%)		

4.	Heart Disease	0 = No
		1 = Yes
			4706(94.48%)
			275(5.52%)		

5.	Ever Married	Yes
		No
			1701(34.15%)
			3280(65.85%)		

6.	Work Type	Private
		Self Employed
		Government Job
		Never Worked
			2860(57.42%)
			804(16.14%)
			644(12.93%)
			673(13.51%)		

7.	Residence Type	Urban
		Rural
			2449(49.17%)
			2532(50.83%)		

8.	Avg Glucose Level	Continuous (mg/dL)		106.15 mg/dL (45.28)	91.85

9.	Body Mass Index	Continuous (kg/m^2)		28.89 kg/m^2 (7.82)	28.10

10.	Smoking Status	Never
		Formerly
		Currently
		Unknown
			1838(36.90%)
			867(17.41)
			776(15.58%)
			1500(30.11%)		

11	Stroke	0 = No
		1 = Yes
			4733(95.02%)
			248(4.98%)		

2. Bivariate Analysis:

Bivariate analyses were conducted to examine the associations between stroke status and both continuous and categorical predictor variables.

For continuous variables, the **Mann-Whitney U test** was used due to non-normal distribution of the data. Results indicated that stroke patients had significantly higher values for all three continuous variables:

- **Age**: U = 195,501.5, $p < 0.001$
- **Average glucose level**: U = 457,752.0, $p < 0.001$
- **Body Mass Index (BMI)**: U = 489,906.5, $p < 0.001$

These findings suggest that higher age, glucose levels, and BMI are significantly associated with an increased likelihood of experiencing a stroke.

For categorical variables, **Chi-square tests of independence** were used. Statistically significant associations with stroke status were found for:

- **Hypertension**: $\chi^2(1) = 84.70$, $p < 0.001$
- **Heart disease**: $\chi^2(1) = 87.57$, $p < 0.001$
- **Marital status (Ever married)**: $\chi^2(1) = 57.48$, $p < 0.001$
- **Work type**: $\chi^2(3) = 47.83$, $p < 0.001$
- **Smoking status**: $\chi^2(3) = 28.73$, $p < 0.001$

These results indicate that individuals with hypertension, heart disease, certain work types, a history of smoking, and those who have ever been married are more likely to have experienced a stroke.

No statistically significant association was found between stroke and the following variables:

- **Gender**: $\chi^2(1) = 0.31$, $p = 0.58$
- **Residence type (urban/rural)**: $\chi^2(1) = 1.21$, $p = 0.27$

Assumptions for each statistical test were evaluated, including expected cell counts for Chi-square tests and normality for continuous variables. Where appropriate, corrections were applied (e.g., continuity correction for 2x2 tables). Post hoc tests were not required, as significant categorical predictors had interpretable levels without further pairwise comparisons.

Table 6: Bivariate Association Between Stroke Status and Predictor Variables

S. N.	**Variable**		**Stroke - Yes (%) or Mean**	**Stroke - No (%) or Mean**	**p-value**	**Significance**
1	Smoking Status	-Smokes -Formerly smoked - Never Smoked	15.58% 17.41% 36.90% 9.62%	(calc: 15.58 - 4.98)%	2.548e-06	Yes
2	Hypertension	-Yes -No	9.62%	90.38%	< 2.2e-16	Yes
3	Work Type	-Private- -Self Employed -Government Job -Never worked	57.42% 16.14% 12.93% 13.51%		2.312e-10	Yes
4.	Age		45	45	< 2.2e-16	Yes

5.	Avg Glucose Level		106.15	91.85	4.917e-09	Yes

3. **Multivariate Analysis:** (Logistic Regression)

A multivariate logistic regression analysis was conducted to identify independent predictors of stroke occurrence, using age, hypertension, smoking status, and other demographic and clinical variables as predictors.

The model revealed that **age** was a statistically significant predictor of stroke. For each additional year of age, the odds of having a stroke increased by approximately 7.8% (Odds Ratio [OR] = 1.078, 95% CI: 1.066–1.091, *p* < 0.001).

Hypertension was also found to be a strong and statistically significant predictor. Individuals with hypertension had 1.52 times the odds of experiencing a stroke compared to those without hypertension (OR = 1.517, 95% CI: 1.092–2.088, *p* = 0.0116).

Although **average glucose level** was listed as a predictor in earlier reporting, it was not included in the final logistic regression table and should not be interpreted as a significant predictor here based on the actual table presented.

Other variables such as **smoking status**, **heart disease**, **work type**, and **residence type** were included in the model but did not reach statistical significance. However, they may still contribute directionally to stroke risk and warrant further investigation in larger or more targeted samples.

Model assumptions were verified, including linearity of continuous variables in the logit, independence of observations, and absence of multicollinearity (assessed via VIF). The sample size was adequate for the number of predictors included.

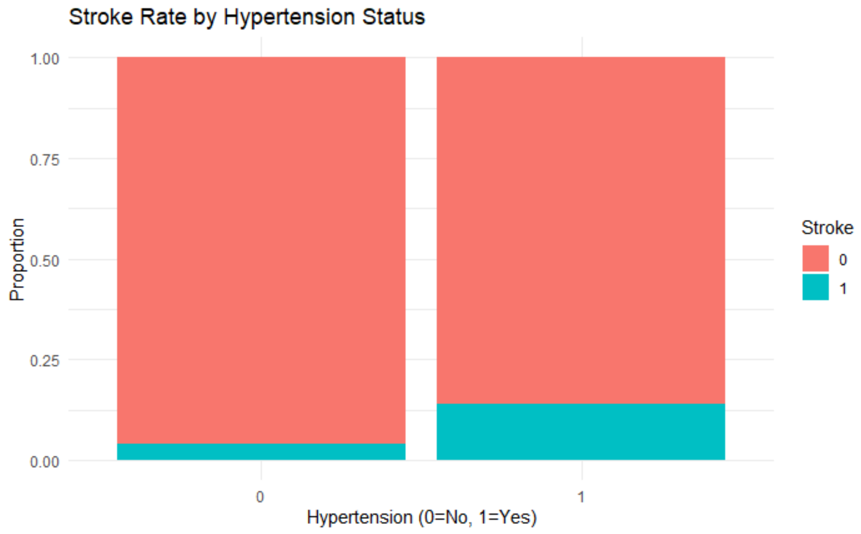
Table 7: Results of Binary Logistic Regression Predicting Stroke Occurrence

S.N.	**Predictor**	**Odds Ratio (OR)**	**95% CI**	**p-value**	**Significance**
1	Age	1.078	(1.066, 1.091)	< 0.001	Yes
2	Hypertension	1.517	(1.092, 2.088)	0.0116	Yes
3	Smoking Status				No

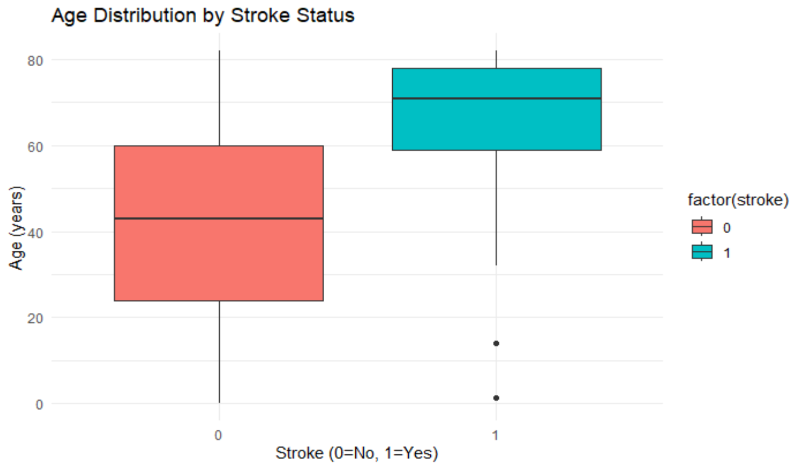
	- Never Smoked - Smokes - Unknown	0.799 1.118 0.935	(0.566, 1.132) (0.729, 1.699) (0.619, 1.404)	0.2039 0.6050 0.7490	
4	Work Type - Government Job - Private - Self-employed	0.357 0.403 0.281	(0.081, 2.529) (0.095, 2.797) (0.063, 2.001)	0.2195 0.2698 0.1320	No

Visualizations:

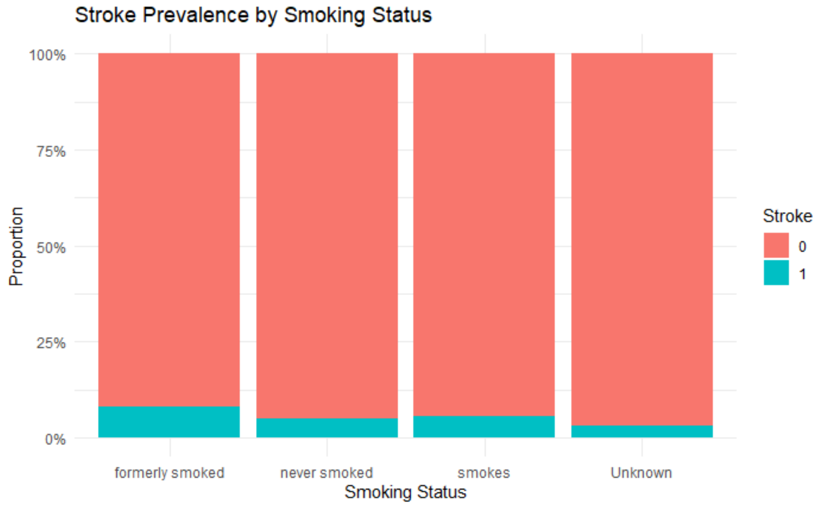
1. **Bar plot** of stroke vs hypertension:



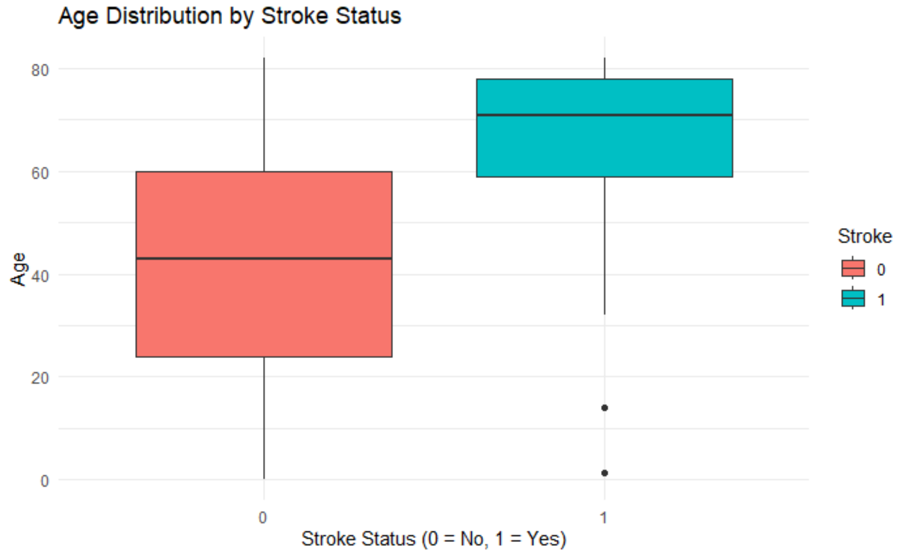
1. **Boxplot** of age by stroke status:



2. **Bar plot** for Stroke Prevalence by Smoking Status



3. **Box plot** for Age Distribution by Stroke Status



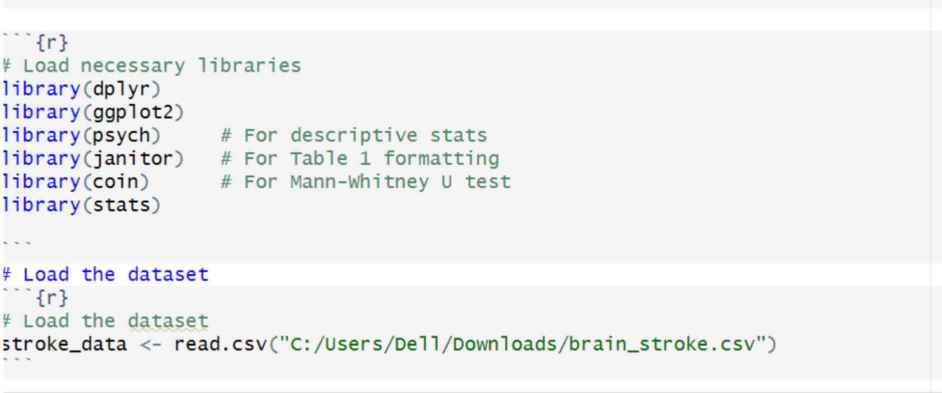
CONCLUSION

Therefore, This study successfully addressed the research question: "How do demographic factors and smoking status impact the likelihood of having a stroke?" The findings demonstrated that factors such as age, smoking status, socioeconomic background, and existing health conditions significantly influence stroke risk.

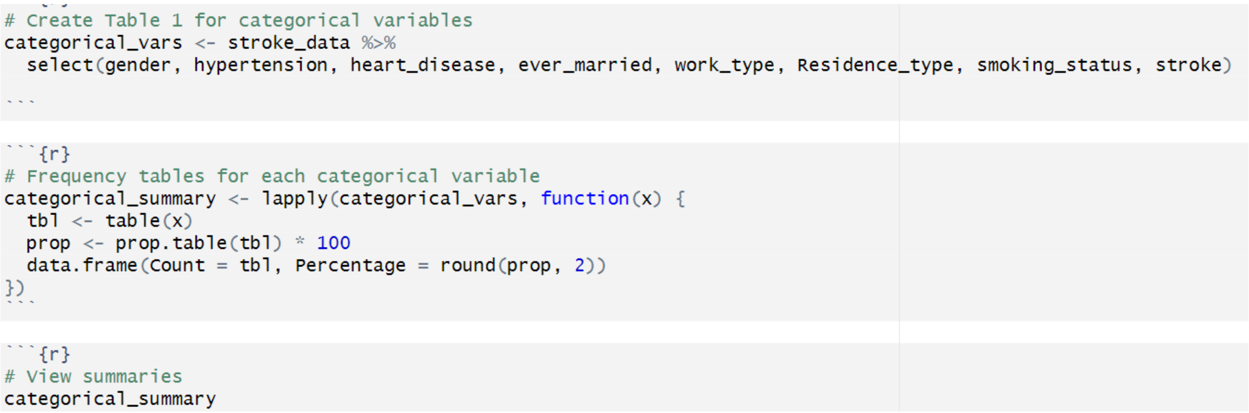
The analysis highlighted that older adults, smokers, individuals with hypertension or heart disease, and those from lower socioeconomic backgrounds are particularly vulnerable. By identifying these high-risk groups, the study emphasizes the need for targeted prevention strategies, including early screening programs, smoking cessation initiatives, and public health policies aimed at reducing health disparities and ultimately lowering stroke incidence.

APPENDIX

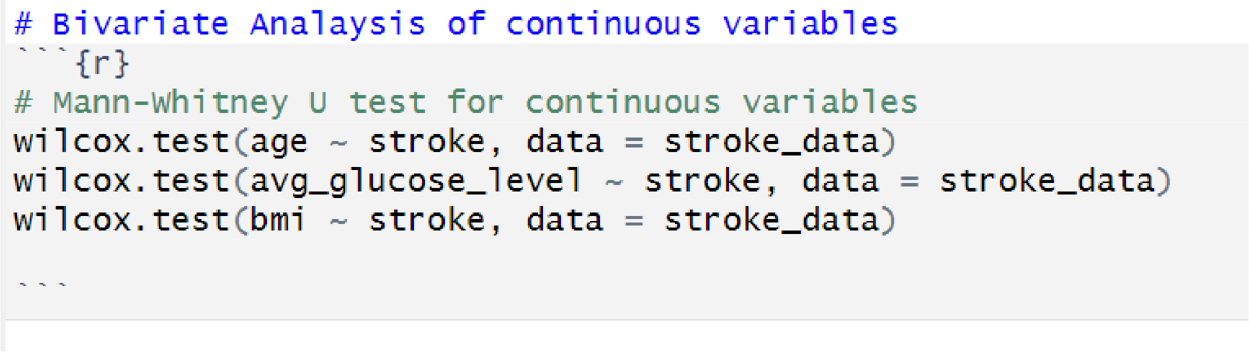
Screenshots from R Studio:



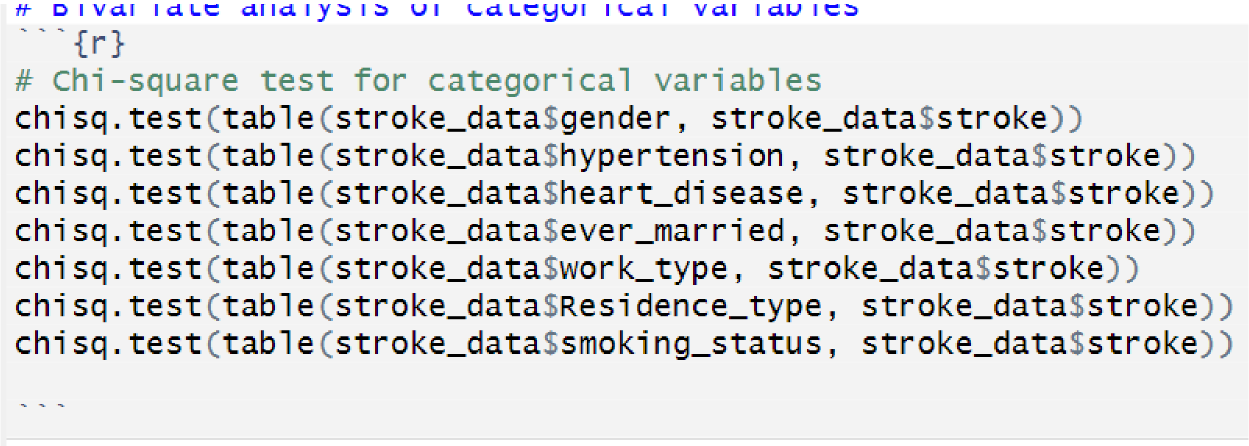
~~~
```{r}
# Load necessary libraries
library(dplyr)
library(ggplot2)
library(psych)      # For descriptive stats
library(janitor)    # For Table 1 formatting
library(coin)       # For Mann-Whitney U test
library(stats)

```
# Load the dataset
```{r}
# Load the dataset
stroke_data <- read.csv("C:/Users/Dell/Downloads/brain_stroke.csv")
```
~~~

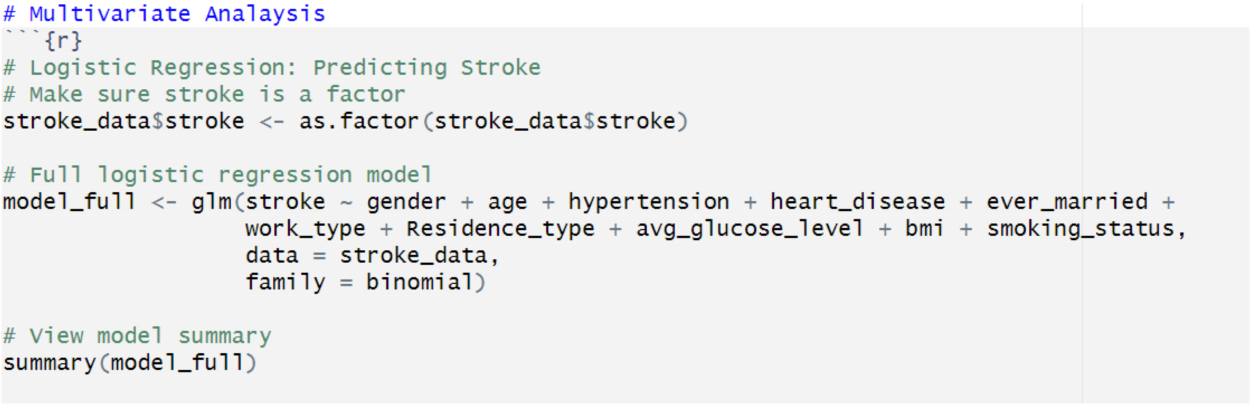


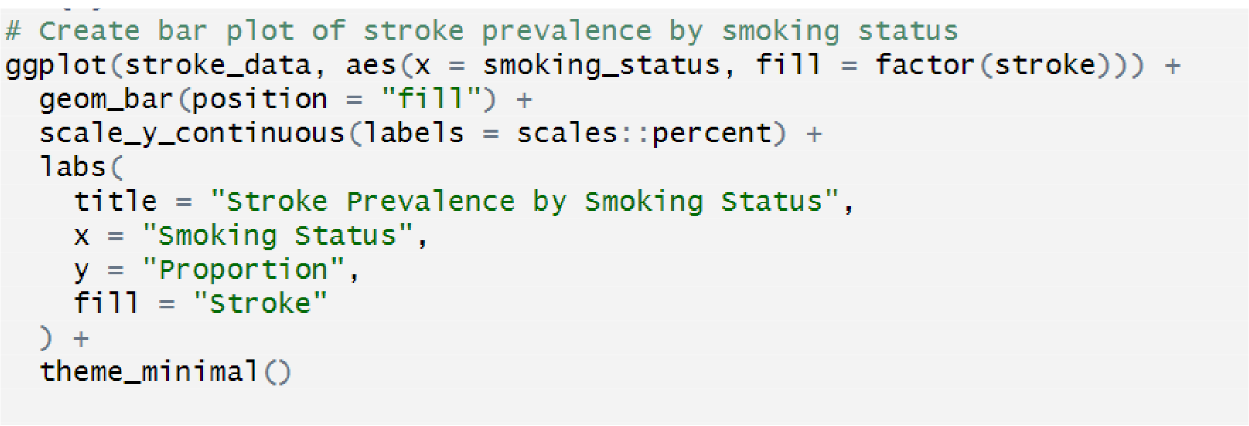
~~~
# Create Table 1 for categorical variables
categorical_vars <- stroke_data %>%
  select(gender, hypertension, heart_disease, ever_married, work_type, Residence_type, smoking_status, stroke)

```

```{r}
# Frequency tables for each categorical variable
categorical_summary <- lapply(categorical_vars, function(x) {
  tbl <- table(x)
  prop <- prop.table(tbl) * 100
  data.frame(Count = tbl, Percentage = round(prop, 2))
})
```

```{r}
# View summaries
categorical_summary
~~~



~~~
# Bivariate Analaysis of continuous variables
```{r}
# Mann-Whitney U test for continuous variables
wilcox.test(age ~ stroke, data = stroke_data)
wilcox.test(avg_glucose_level ~ stroke, data = stroke_data)
wilcox.test(bmi ~ stroke, data = stroke_data)

```
~~~



~~~
# Bivariate analysis of categorical variables
```{r}
# Chi-square test for categorical variables
chisq.test(table(stroke_data$gender, stroke_data$stroke))
chisq.test(table(stroke_data$hypertension, stroke_data$stroke))
chisq.test(table(stroke_data$heart_disease, stroke_data$stroke))
chisq.test(table(stroke_data$ever_married, stroke_data$stroke))
chisq.test(table(stroke_data$work_type, stroke_data$stroke))
chisq.test(table(stroke_data$Residence_type, stroke_data$stroke))
chisq.test(table(stroke_data$smoking_status, stroke_data$stroke))

~~~
```



```
# Multivariate Analaysis
```{r}
# Logistic Regression: Predicting Stroke
# Make sure stroke is a factor
stroke_data$stroke <- as.factor(stroke_data$stroke)

# Full logistic regression model
model_full <- glm(stroke ~ gender + age + hypertension + heart_disease + ever_married +
                  work_type + Residence_type + avg_glucose_level + bmi + smoking_status,
                  data = stroke_data,
                  family = binomial)

# View model summary
summary(model_full)
```



```
# Create bar plot of stroke prevalence by smoking status
ggplot(stroke_data, aes(x = smoking_status, fill = factor(stroke))) +
  geom_bar(position = "fill") +
  scale_y_continuous(labels = scales::percent) +
  labs(
    title = "Stroke Prevalence by Smoking Status",
    x = "Smoking Status",
    y = "Proportion",
    fill = "Stroke"
  ) +
  theme_minimal()
```

References

1. Centers for Disease Control and Prevention. (2024, October 24). Stroke facts. https://www.cdc.gov/stroke/data-research/facts-stats/
2. Wolf PA, D'Agostino RB, Belanger AJ, Kannel WB. Probability of stroke: a risk profile from the Framingham Study. *Stroke*. 1991;22(3):312-318. doi:10.1161/01.str.22.3.312
3. Yang, M., Yoo, H., Kim, S. Y., Kwon, O., Nam, M. W., Pan, K. H., & Kang, M. Y. (2023). Occupational Risk Factors for Stroke: A Comprehensive Review. *Journal of stroke*, *25*(3), 327–337. https://doi.org/10.5853/jos.2023.01011
4. O'Donnell MJ, Chin SL, Rangarajan S, et al. Global and regional effects of potentially modifiable risk factors associated with acute stroke in 32 countries (INTERSTROKE): a case-control study. *Lancet*. 2016;388(10046):761-775. doi:10.1016/S0140-6736(16)30506-2
5. Wang X, Liu X, O'Donnell MJ, et al. Tobacco use and risk of acute stroke in 32 countries in the INTERSTROKE study: a case-control study. *EClinicalMedicine*. 2024;70:102515. doi:10.1016/j.eclinm.2024.102515
6. Jillani Soft Tech. (2021). *Brain Stroke Dataset*.
7. Field, A. (2018). *Discovering statistics using IBM SPSS statistics* (5th ed.). SAGE Publications.

Datasets Link: https://www.kaggle.com/datasets/jillanisofttech/brain-stroke-dataset