# Implementation of R Language in Analyzing the Influence of Smoking on the Risk of Patients Getting Stroke


**Oleh:**

**Nama: Darrien Rafael Wijaya**

**NIM: 2602064241**

**Kelas: LA06**

**Bina Nusantara University**

**2023/2024**

# Table of Contents

# Chapter 1 – INTRODUCTION

## 1.1. Introduction

This study delves into the utilization of RStudio for analyzing the correlation between smoking habits and the risk of stroke in patients, employing the "Stroke Prediction Dataset" as the primary data source. Employing exploratory data analysis methods in RStudio, the study models statistics, visualizes data, and tests hypotheses to identify relevant variables for analyzing the impact of smoking on stroke risk.

Smoking, being a prevalent global habit, poses adverse effects on health, with cigarettes containing toxins such as nicotine, carbon monoxide, and tar that can damage blood vessels and trigger plaque formation in arteries. As a result, smoking can elevate the risk of cardiovascular diseases, including stroke, accounting for 1 in 10 global stroke cases (Imanda et al., 2019). Therefore, in-depth analyses regarding the influence of smoking on the risk of stroke are crucial.

In this study, the R programming language is implemented to analyze the influence of smoking on the likelihood of contracting stroke. R was chosen due to its popularity as a programming language equipped with numerous libraries and tools for data analysis, coupled with its easily understandable syntax, making it a preferred choice for researchers conducting exploratory data analysis. Through exploratory data analysis conducted using the R programming language, it is anticipated that this observation can provide a deeper understanding of the relationship between smoking and the risk of developing stroke. The findings of this research can contribute to efforts in prevention, management, and a better understanding of factors contributing to stroke.

Key attributes considered in our predictive model include the presence of hypertension, heart disease, marital status, occupation type, residence type, average glucose level, body mass index (BMI), and smoking status. The inclusion of these factors is designed to capture the intricate interplay between individual health, lifestyle, and demographic characteristics that contribute to the overall stroke risk. Adhering to ethical standards and privacy considerations, it is imperative to note that the dataset used in this research is sourced from a confidential database intended exclusively for educational purposes. Researchers are reminded to credit the author if utilizing this dataset for further investigations. Through this research, we aim to contribute valuable insights to the field of preventive healthcare by developing a predictive model that can aid healthcare professionals in identifying individuals at higher risk of stroke. By understanding the nuanced relationships between various factors, our study strives to pave the way for more targeted interventions and personalized healthcare approaches, ultimately reducing the global burden of stroke-related morbidity and mortality.

## 1.2. Significance of the Study

### 1.2.1. Tailoring Interventions Across Age Groups:

Understanding how stroke risk varies across age groups is crucial for tailoring effective interventions. By categorizing ages into distinct ranges, from early adulthood to advanced years, we can pinpoint when individuals might be more susceptible. This insight enables

healthcare practitioners to develop age-specific preventive measures and allocate resources efficiently.

1.2.2. Smoking Status as a Modifier:

Smoking is known to influence cardiovascular health. This study aims to shed light on how smoking status interacts with age in shaping the risk of stroke. Identifying age groups where smoking has a more pronounced impact provides a foundation for targeted anti-smoking campaigns and smoking cessation programs.

1.2.3. Informing Public Health Policies:

Findings from this study contribute valuable insights to inform public health policies. By pinpointing age brackets with heightened stroke risk, policymakers can allocate resources for awareness campaigns and healthcare infrastructure where they are needed most. Moreover, identifying high-risk groups helps in shaping policies that address lifestyle factors such as smoking within these demographics.

1.2.4. Holistic Approach to Stroke Prevention:

Stroke prevention is not a one-size-fits-all endeavor. This study advocates for a holistic approach by considering age and smoking status as intertwined factors. It prompts a reevaluation of prevention strategies, emphasizing personalized healthcare that recognizes the unique risks associated with different age groups and smoking behaviors.

In conclusion, this study embarks on a comprehensive exploration of stroke risk factors, unraveling the intricate connections between age, smoking status, and the incidence of strokes. The significance of this research lies in its potential to guide targeted interventions, inform public health policies, and ultimately contribute to a holistic approach to stroke prevention.

# Chapter 2 – DATASET DESCRIPTION

This dataset contains the relation between smoker, stroke, hypertension, and hearth disease. Its attributes have a boolean value (‘0’ or ‘1’) to indicates if they have the disease or not. This dataset is used to predict whether a patient is likely to get stroke. Each row in the data provides relavant information about the patient.

This dataset contains 14 variables and 3426 rows. That includes:

1) id: unique identifier
2) gender: "Male", "Female" or "Other"
3) age: age of the patient
4) hypertension: 0 if the patient doesn't have hypertension, 1 if the patient has hypertension
5) heart_disease: 0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease
6) ever_married: "No" or "Yes"
7) work_type: "children", "Govt_jov", "Never_worked", "Private" or "Self-employed"
8) Residence_type: "Rural" or "Urban"

9) avg_glucose_level: average glucose level in blood
10) bmi: body mass index
11) smoking_status: "formerly smoked", "never smoked", "smokes" or "Unknown"
12) stroke: 1 if the patient had a stroke or 0 if not

The dataset under consideration exhibits instances of outliers and missing data; however, it is noteworthy that there is an absence of duplicate data, a verification that has been duly undertaken and confirmed.

## 2.1. Finding if there's outliers

To ascertain the presence of outliers within the dataset, a methodical approach was employed, specifically utilizing scatterplots as a visual tool for the purpose of outlier detection. This systematic procedure allowed for a comprehensive examination of the data, aiding in the identification and assessment of any data points that deviate significantly from the overall trend or distribution.

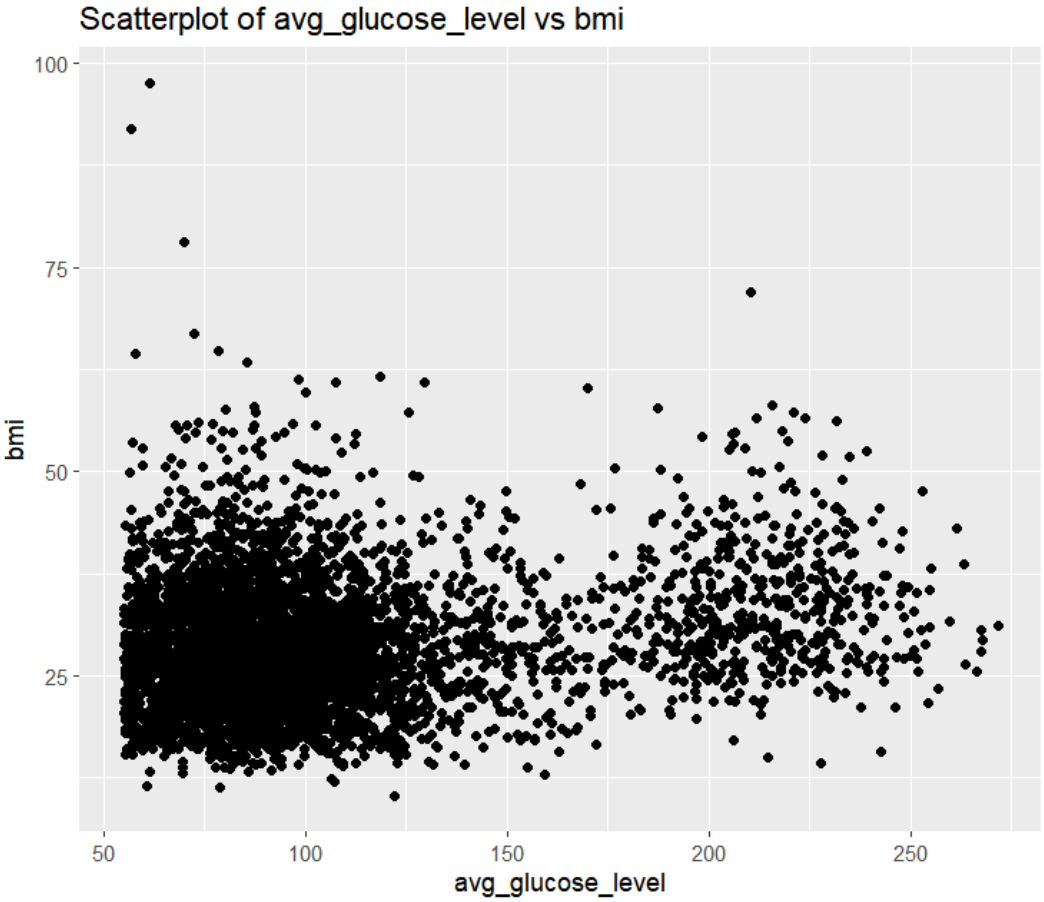



**Image 2.1**
**Scatterplot of Glucose Level Average and BMI**
(code is provided in the attached R file and this report last page)

From the scatterplot above, it has been discerned that the dataset does, indeed, contain certain data points that exhibit characteristics indicative of outliers. These outliers, identified through rigorous examination, warrant further investigation and consideration in the subsequent stages of data interpretation and analysis.

## 2.2. Finding and removing the missing value

Within the dataset, there are instances where data is missing its value, denoted by the placeholder "N/A." Recognizing the imperative need for a pristine dataset conducive to rigorous analysis, a judicious data cleaning process was meticulously executed. This systematic procedure involved the identification and rectification of missing data, ensuring the integrity and reliability of the dataset for subsequent phases of analytical exploration.

| work_type | Residence_type | avg_glucose_level | bmi |
|---|---|---|---|
| Private | Urban | 228.69 | 36.6 |
| Self-employed | Rural | 202.21 | *NA* |
| Private | Rural | 105.92 | 32.5 |

**Image 2.2**
**One of the Missing Value in the dataset**

The main goal here was to carefully find and get rid of the rows or entries that had "N/A" values in the dataset. This way, we wanted to make sure our dataset is stronger and more complete for the next steps in analysis and modeling. So, in simple terms, we did some cleaning to kick out those "N/A" values.

```
unique_values <- unique(df$bmi)
print(unique_values)
df[df == "N/A"] <- NA
data <- na.omit(df)
```

## 2.3. Detecting if there are duplicated data

To find out if there were any duplicated data in the dataset, researcher wrote a code to take a closer look. The code was designed to carefully check the dataset and identify any entries that were duplicated. This thorough examination was done to make sure the dataset is accurate and free from unnecessary repetitions, setting the stage for further analysis without any redundant information.

```
duplicate_rows <- duplicated(data)
duplicate_rows_data <- data[duplicate_rows, ]
print(duplicate_rows_data)
```

After carefully looking into it, we found out that there aren't any repeated values in the dataset. This tells us that the dataset is solid and accurate, showing that our efforts to check for duplicates worked well in keeping the data clean and free from unnecessary repetition.

```
 [1] id                gender            age               hypertension      heart_disease
 [6] ever_married      work_type         Residence_type    avg_glucose_level bmi
[11] smoking_status    stroke            age_skewed
<0 rows> (or 0-length row.names)
```

**Image 2.3**
**Scatterplot of Glucose Level Average and BMI**
(code is provided in the attached R file and this report last page)

Subsequently, the research delved deeply into the dataset to unravel the intricate connections among numeric variables. This meticulous exploration involved the application of rigorous statistical methods aimed at discerning discernible patterns and relationships inherent within these numerical parameters. By employing various statistical techniques, the researcher sought to illuminate the underlying structure and interdependencies that characterize the dataset.

One of the key methodologies employed was the investigation of correlations between numeric variables. This analytical approach allowed for the identification of statistically significant relationships, shedding light on how changes in one variable may correspond to changes in another. Through correlation analysis, the researcher aimed to capture the inherent dynamics and dependencies that might exist among the numeric components of the dataset.

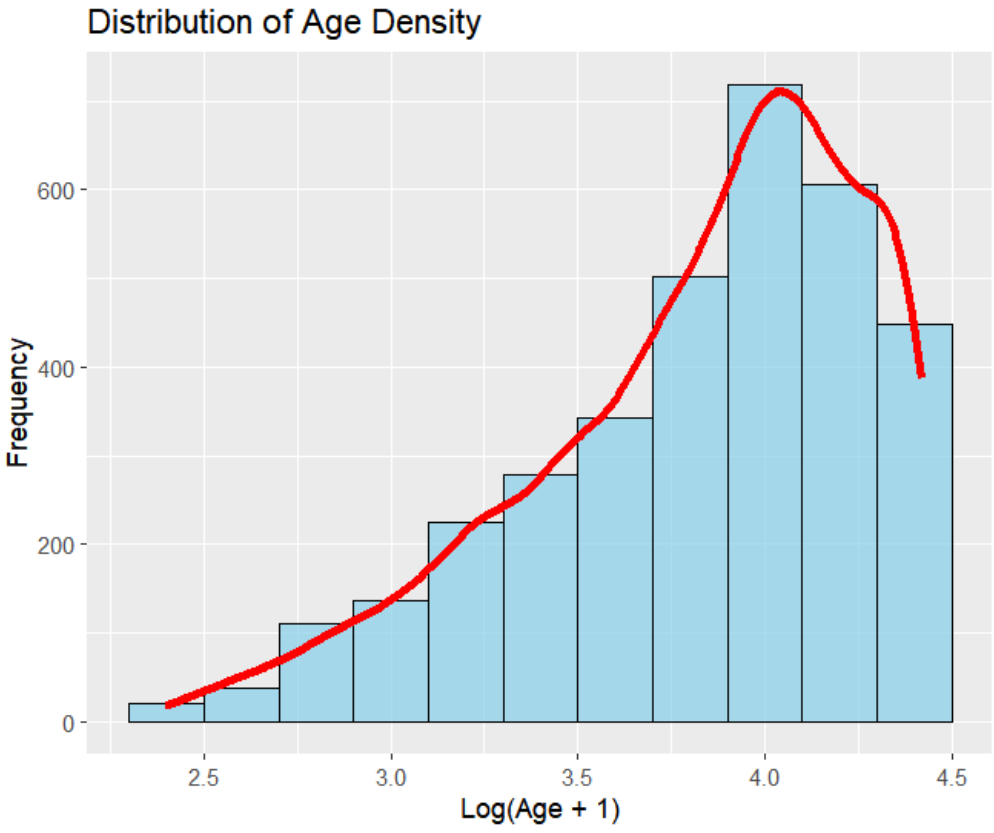



**Image 2.4**
**Distribution of Age Density**
(code is provided in the attached R file and this report last page)

The graph provides insight into the distribution of ages within the dataset, emphasizing the right-skewed nature of the distribution. The blue bars in the histogram showcase the frequency of occurrences for different log-transformed age ranges. Meanwhile, the red density line offers a smoothed estimate of the probability density function, highlighting the underlying shape of the age distribution. The x-axis reflects the log-transformed age values, and the y-axis indicates the frequency of occurrences. This visualization allows for a comprehensive understanding of the distributional characteristics, particularly the skewness, providing a nuanced perspective on the dataset's age distribution.

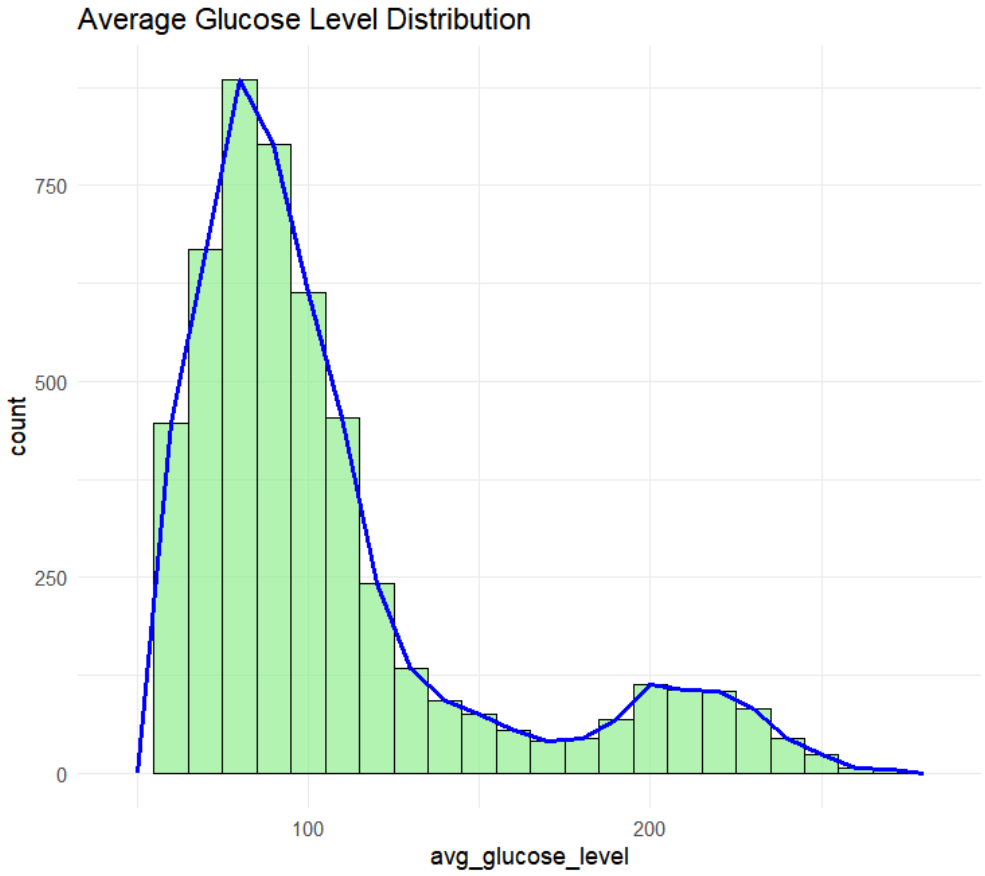



**Image 2.5**
**Scatterplot of Glucose Level Average and BMI**
(code is provided in the attached R file and this report last page)

The image presents a histogram portraying the distribution of average glucose levels within the dataset. This histogram serves as a visual representation of the frequency of individuals across various glucose levels, essentially illustrating how many people fall into specific glucose level ranges. For instance, the most prevalent glucose level appears to be around 100, with approximately 500 individuals registering this particular value.

The observation derived from the histogram analysis is the non-normal distribution of glucose levels. Rather than conforming to a symmetrical bell curve, the distribution exhibits a skewness towards the right. This skewness indicates that a larger proportion of individuals in the dataset possess lower glucose levels, while fewer individuals exhibit higher glucose levels.

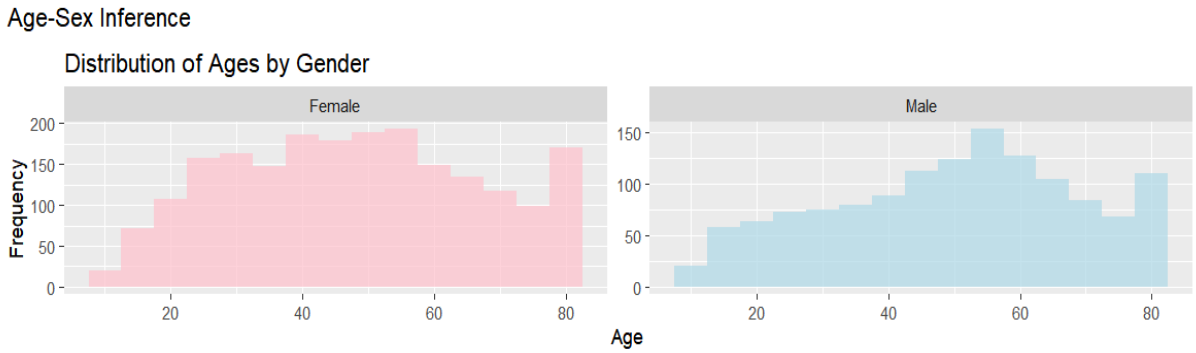



**Image 2.6**
**Distribution of Age with Male and Female**
(code is provided in the attached R file and this report last page)

Looking at the distribution, we can see that the ages of men are spread out, but there's a noticeable group of older individuals. On the other hand, for women, there's a larger cluster of people aged between 40 and 60.

This suggests that more men in the data are older, while women are more evenly distributed, with a concentration in the middle-age range. It might be interesting to explore why there's this difference in age distribution between men and women. This finding could give us insights into the demographics of our study participants.

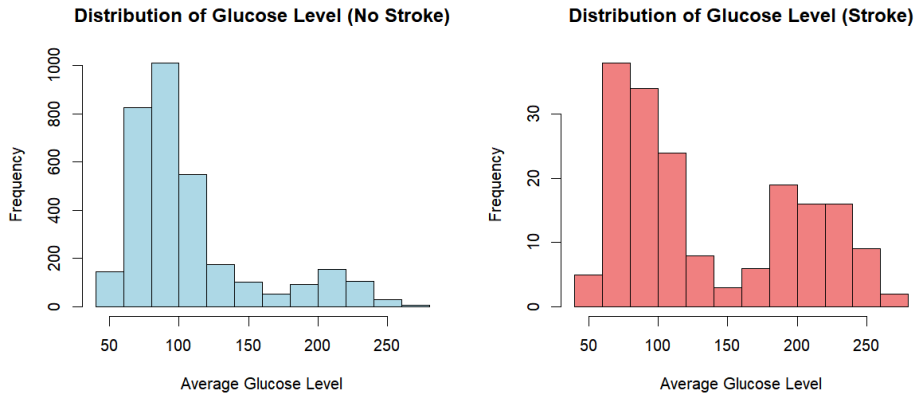
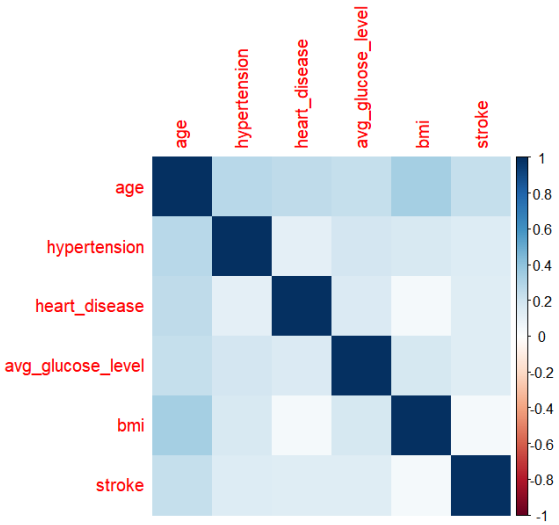



**Image 2.7**
**Distribution of Glucose Level with Stroke and No Stroke**
(code is provided in the attached R file and this report last page)

The presented histograms illuminate noteworthy disparities in the distribution of glucose levels between individuals with and without a history of stroke. Specifically, the histogram for individuals with a history of stroke exhibits a rightward shift, indicating a prevailing trend of higher glucose levels compared to their non-stroke counterparts. This aligns seamlessly with existing research underlining the significance of elevated blood sugar levels as a prominent risk factor for stroke.

Furthermore, a distinctive feature of the histogram for those with a history of stroke is its broader spread, signifying increased variability in glucose levels within this cohort. This expanded range may be attributed to several contributing factors, including variations in the

type and severity of strokes experienced by individuals, as well as the presence of underlying health conditions such as hypertension and heart disease.



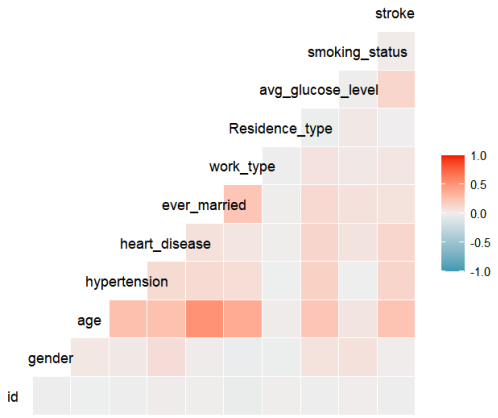




**Image 2.8 – Image 2.9**
**Correlation between variables in the dataset**
(code is provided in the attached R file and this report last page)

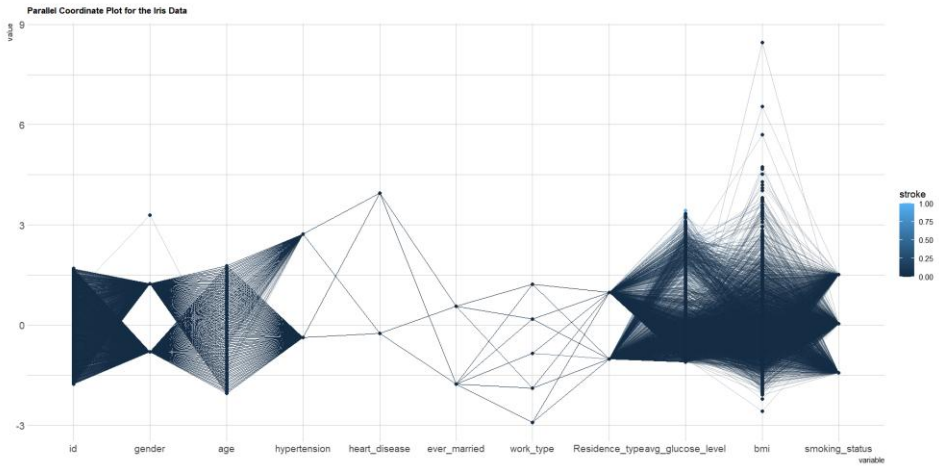



**Image 2.10**
**Correlation between variables in the dataset**
(code is provided in the attached R file and this report last page)

The dataset under consideration comprises comprehensive health-related information on individuals, encompassing demographic details such as age, gender, marital and work status, residence type, and crucial health indicators including the presence of hypertension, heart disease, average glucose level, BMI, and smoking habits.

A notable observation within the dataset reveals a significant incidence of stroke, particularly prevalent among the older demographic characterized by a higher prevalence of hypertension or heart disease. However, it is imperative to acknowledge the existence of missing BMI data, which warrants attention and consideration in subsequent analyses.

In summation, this dataset offers valuable insights into the health characteristics influencing the occurrence of strokes. However, to unveil more intricate patterns and nuances, a more in-depth and detailed analysis is requisite, ensuring a comprehensive understanding of the interplay between various health factors and stroke incidence.

## Chapter 3 – EXPLORATION AND VISUALIZATION

```
Call:
glm(formula = stroke ~ age + avg_glucose_level + heart_disease +
    hypertension + bmi + ever_married + bmi, family = binomial(link = "logit"),
    data = data)

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       -7.753005   0.644218 -12.035  < 2e-16 ***
age                0.068910   0.006537  10.541  < 2e-16 ***
avg_glucose_level  0.004570   0.001371   3.334 0.000856 ***
heart_disease      0.453786   0.216948   2.092 0.036467 *
hypertension       0.556688   0.182468   3.051 0.002282 **
bmi                0.006714   0.012846   0.523 0.601202
ever_marriedYes   -0.157026   0.261124  -0.601 0.547609
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1411.0  on 3425  degrees of freedom
Residual deviance: 1149.3  on 3419  degrees of freedom
AIC: 1163.3

Number of Fisher Scoring iterations: 7
```

**Image 3.1**
**Generalized Linear Model**
(code is provided in the attached R file and this report last page)

**Age:**
Interpretation: For every one-unit increase in age, the estimated likelihood (log odds) of having a stroke increases by 0.06.
Significance: The p-value associated with age is very small, suggesting that age is a statistically significant predictor of stroke. In other words, age is an important factor in determining the likelihood of having a stroke.

**Average Glucose Level:**
Interpretation: A one-unit increase in average glucose level corresponds to a 0.005 increase in the log odds of having a stroke.
Significance: The low p-value for average glucose level indicates that it is a statistically significant predictor. This implies that variations in average glucose level are associated with meaningful changes in the likelihood of having a stroke.

**Heart Disease:**
Interpretation: Individuals with heart disease are expected to have 0.404 higher log odds of having a stroke compared to those without heart disease.

Significance: The p-value for heart disease is below the conventional significance level of 0.05, suggesting that the presence of heart disease significantly influences the likelihood of having a stroke.

**Hypertension:**
Interpretation: Individuals with hypertension are anticipated to have 0.534 higher log odds of having a stroke compared to those without hypertension.
Significance: The low p-value associated with hypertension indicates that it is a statistically significant predictor. Thus, hypertension plays a meaningful role in predicting the likelihood of having a stroke.

**BMI and Ever Married:**
Interpretation: Both BMI and marital status (ever_married) do not have a statistically significant impact on the likelihood of having a stroke, as their p-values are above 0.05.
Significance: The lack of significance implies that, in this particular dataset, variations in BMI and marital status are not associated with meaningful changes in the likelihood of having a stroke.

In essence, the analysis suggests that age, average glucose level, heart disease, and hypertension are crucial factors in predicting the likelihood of having a stroke. On the other hand, BMI and marital status do not appear to be significant predictors in this context. This information provides valuable insights into the specific variables that contribute significantly to stroke prediction based on the given dataset.

## 3.1. Age and Smoking Status on Stroke Risk

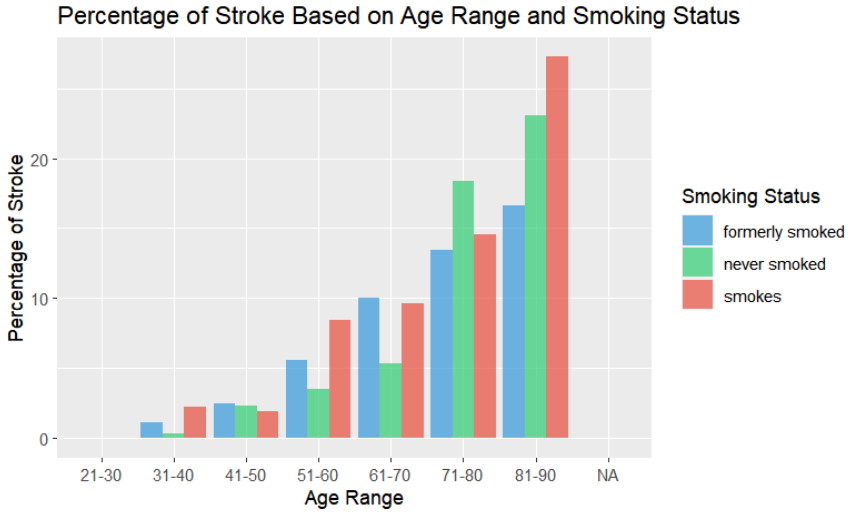



**Image 3.2**
**Histogram of Stroke percentage based on Age Range and Smoking Status**
(code is provided in the attached R file and this report last page)

So, when it comes to smoking and having a stroke, it's a bit of a mixed bag depending on how old you are. If you're between 31 and 70 or 81 to 90, there's a higher chance of having a stroke

if you're a smoker or used to be one. But, in the 71-80 age range, it's the non-smokers who seem to be topping the charts in terms of stroke cases.

So, it looks like age and smoking are teaming up in some age groups to increase the chances of a stroke. Even though there are these ups and downs in different age gangs, the overall trend is that getting older seems to go hand in hand with a higher likelihood of having a stroke.

To sum it up, the relationship between age, smoking, and strokes is a bit like a rollercoaster – it has its twists and turns. But when we step back and look at the big picture, age seems to play a role in cranking up the chances of a stroke, especially when there's smoking in the mix. Understanding these connections can help us figure out how to tackle the risks of strokes in different groups of people.

## 3.2. Gender and Smoking Status on Stroke Risk

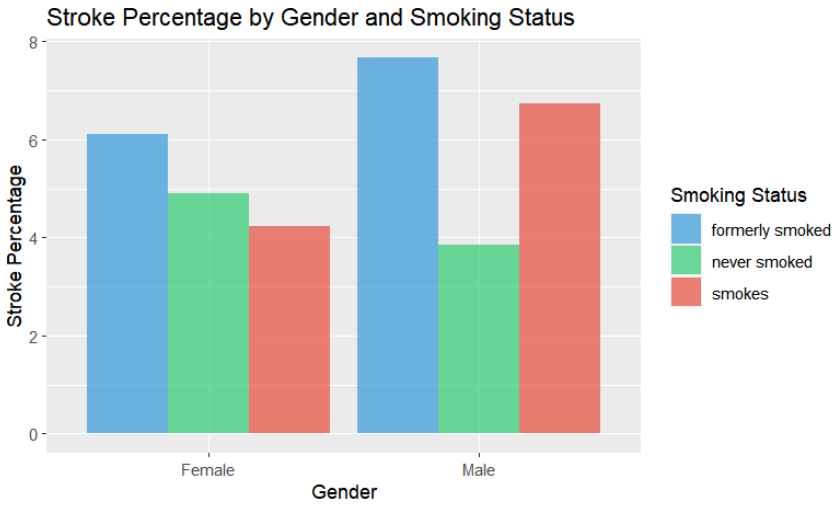



**Image 3.3**
**Histogram of Stroke Percentage by Gender and Smoking Status**
(code is provided in the attached R file and this report last page)

From the diagram above, we noticed something interesting, the likelihood of having a stroke is higher for men compared to women. Looking at the graph, it's clear that people who have smoked before or are still smoking (whether they quit or not) have a higher chance of having a stroke compared to those who never smoked. So, men who smoked in the past or still smoke are more likely to have a stroke compared to women with similar smoking histories.

This finding tells us that there's a connection between being a guy, smoking, and the risk of strokes. The graph paints a clear picture of this relationship, showing that smoking, especially for men, can be a significant factor in having a stroke.

In wrapping it up, this study not only highlights that men might be more prone to strokes but also emphasizes how smoking plays a big role in increasing that risk. By digging into these connections, our research adds valuable info about how gender and smoking habits interact when it comes to the likelihood of having a stroke. This insight can help in planning targeted strategies to prevent strokes and give better healthcare advice.

## 3.3. Percentage of Stroke and No Stroke looking from Gender

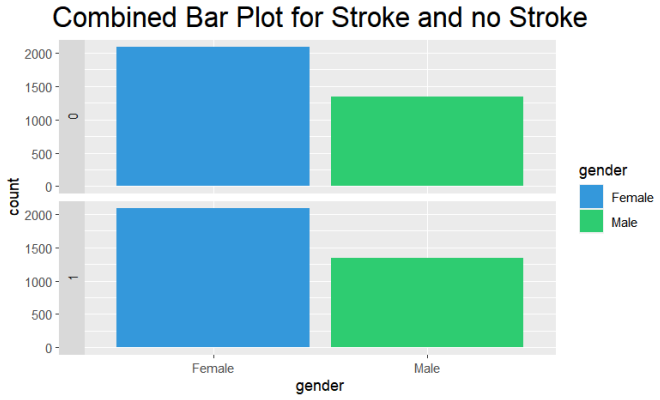



**Image 3.4**
**Bar plot of Stroke and No Stroke Prevalence by Gender**
(code is provided in the attached R file and this report last page)

The graphical diagram represent noteworthy gender disparity in stroke prevalence, revealing that a higher number of females experience strokes compared to males. Exploring and comprehending these distinctions is crucial for developing healthcare strategies that are not only effective but also tailored to the specific needs of different gender groups. By delving into the nuances of these differences, healthcare professionals can refine their interventions, ensuring they are more targeted and impactful.

This nuanced understanding forms the basis for more personalized and gender-sensitive approaches to stroke prevention, diagnosis, and treatment, ultimately contributing to improved health outcomes for individuals across diverse demographic groups.

## 3.4. Age and BMI looking from Stroke or No Stroke

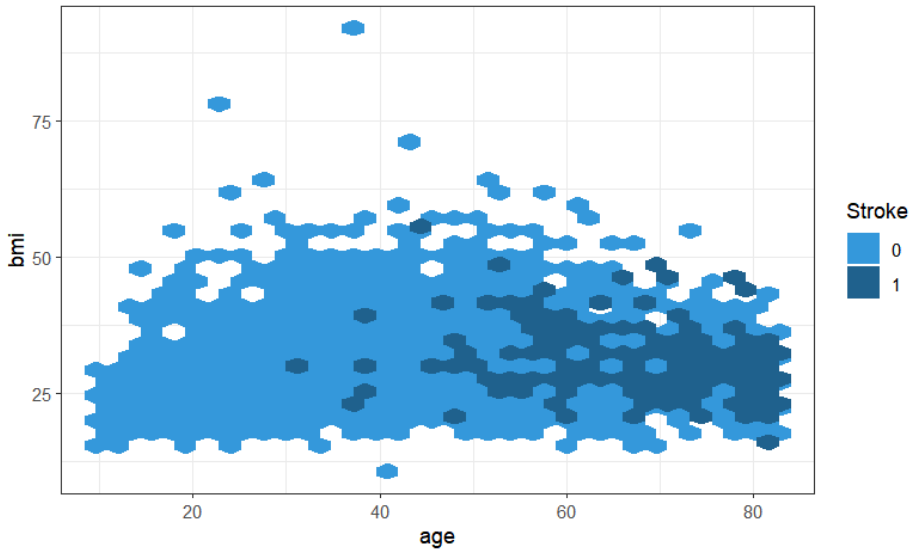



**Image 3.5**
**Hex Diagram of Age and BMI for Stroke and No Stroke**
(code is provided in the attached R file and this report last page)

Examining the hexagonal diagram, it's clear that as people get older and maintain a decent BMI, their chances of experiencing a stroke increase. This simple observation provides a valuable understanding of how age and BMI are connected to the likelihood of having a stroke. The hexagonal visualization helps us see that, as age goes up and BMI stays moderate, the risk of strokes also goes up. This insight encourages us to dig deeper into the reasons behind this connection, considering various factors like genetics, lifestyle choices, and existing health conditions that might contribute to this trend.

By starting to understand this link between age, BMI, and strokes, we set the stage for future research. Exploring these connections further can not only enhance our scientific knowledge about stroke risk factors but also guide the development of specific actions to prevent strokes, especially for people in certain age and BMI groups.

## 3.5. Smoking contributes to increasing Strokes Risk

Through a comparative analysis, researchers have identified a considerable distinction in the susceptibility to stroke among active smokers, former smokers, and individuals who do not smoke. This disparity in stroke occurrence is notably significant. The prevalence of stroke cases is notably higher in the group of active smokers compared to non-smokers. This observation underscores the influential impact of smoking on elevating the risk of developing a stroke, emphasizing the need for heightened awareness and preventive measures in individuals with a history of smoking.

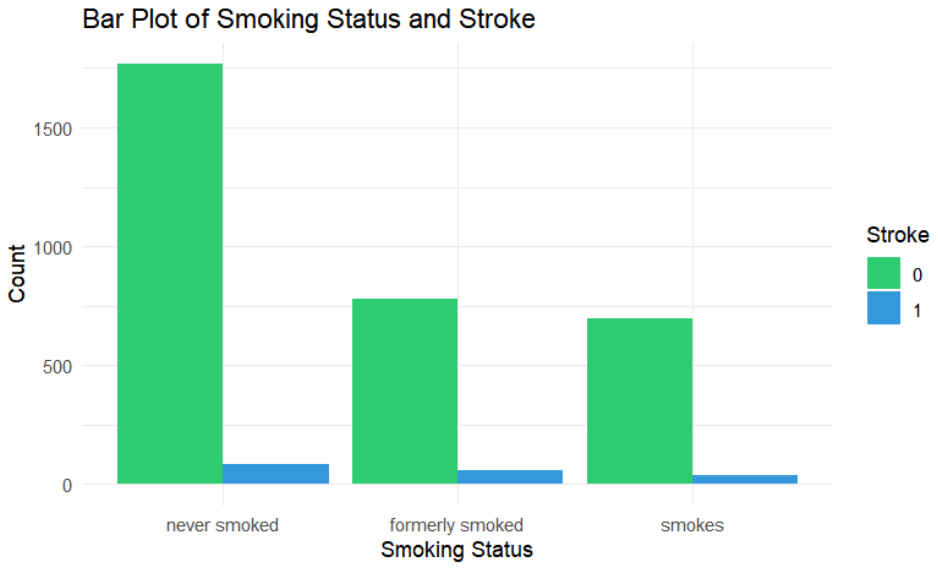



**Image 3.6**
**Bar Plot of Smoking Status with Stroke and No Stroke**
(code is provided in the attached R file and this report last page)

Looking at the graph, it's clear that there are big differences in the number of active smokers, former smokers, and non-smokers. The data shows that non-smokers have a much lower risk of having a stroke compared to both active smokers and people who used to smoke.

This information is important for our health because it tells us that not smoking is associated with a lower risk of stroke. It suggests that encouraging a smoke-free lifestyle can be a key factor in preventing strokes. So, based on the graph, it's pretty clear that avoiding smoking is a good idea for staying healthier and reducing the risk of having a stroke.

## 3.6. Relation of Other Diseases and Smoking Status to Stroke Risk

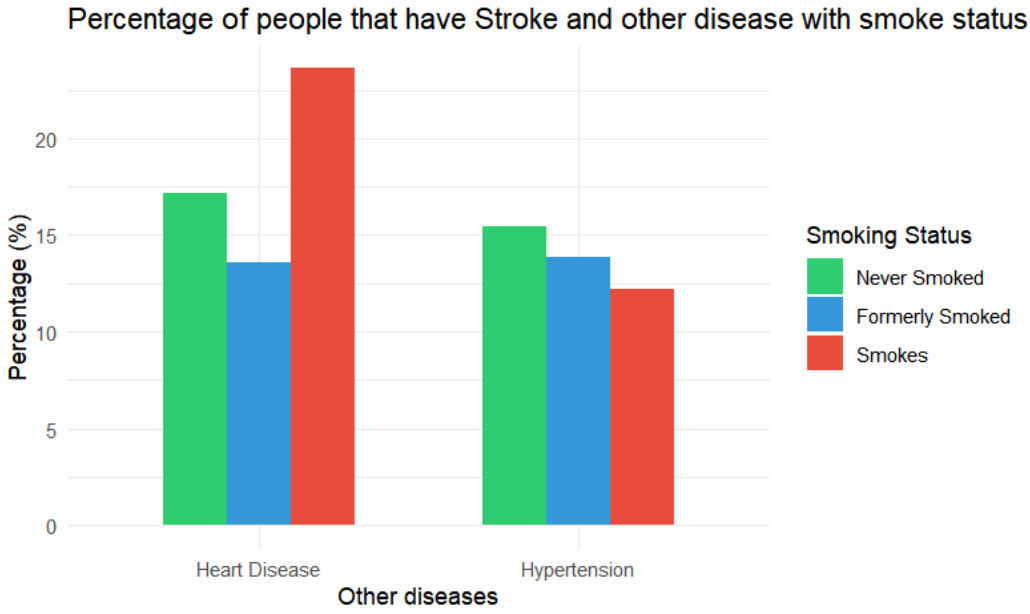



**Image 3.7**
**Bar Plot of Percentage of people that have Stroke and Other Disease**
(code is provided in the attached R file and this report last page)

From the graph above, we know that when people have other health issues, like high blood pressure (hypertension) or heart disease, it can make them more likely to have a stroke. Looking at the pictures of this information, we can see that both hypertension and heart disease make the percentage of strokes go up. But, if someone has high blood pressure and smokes, it doesn't seem to directly make the chance of having a stroke much higher. On the other hand, people with heart disease who smoke have a much bigger increase in the chance of having a stroke. This shows that different health problems and lifestyle choices can affect the chances of having a stroke in different ways.

## 3.7. Stroke Cases Percentage by Work Type

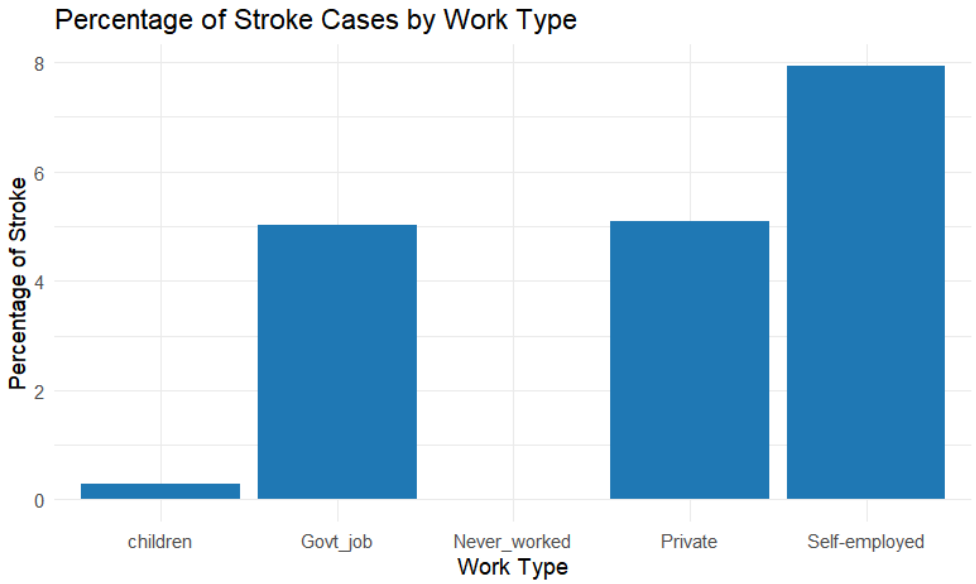
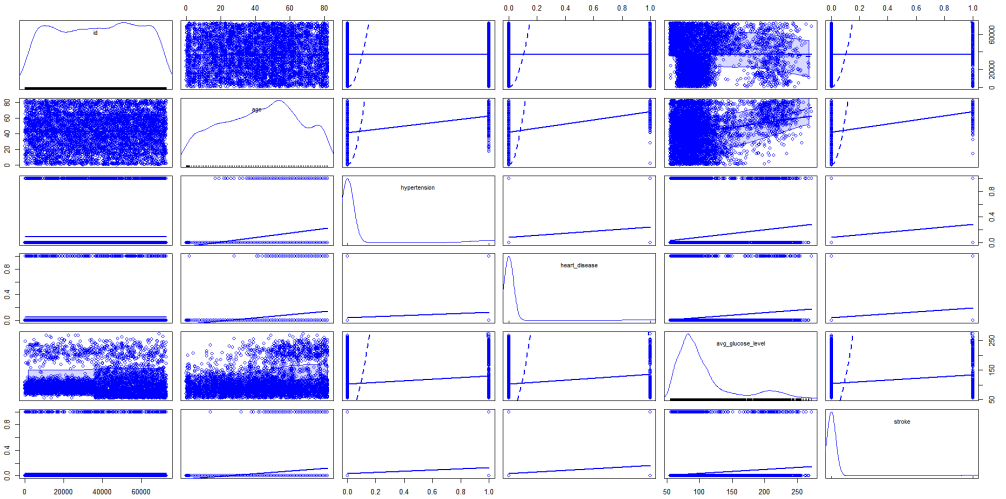



**Image 3.8**
**Bar Plot of Stroke Cases Percentage by Work Type**
(code is provided in the attached R file and this report last page)

Looking at the diagram, it's clear that people who are self-employed have a higher percentage of strokes compared to those in other job types. Interestingly, individuals working for the government show a slightly lower percentage of strokes than those who are self-employed. This observation sparks curiosity about why there might be these differences. Factors like stress at work, lifestyle choices, and access to healthcare could be playing a role. Exploring these aspects further could help us understand why certain jobs seem to be linked to a higher or lower risk of strokes.

In a simpler sense, this information isn't just about strokes; it's also about how the type of job someone has might relate to their health. Understanding these patterns can lead to better ways of promoting health in the workplace and tailoring interventions to reduce the risk of strokes for different job types.

### 3.8. Overall Plot




## Chapter 4 – DISCUSSION AND CONCLUSION

### 4.1. Summary

This dataset encapsulates information related to individuals' health and demographics. Each entry delineates a unique individual, providing details such as gender, age, marital status, work type, residence type, average glucose level, BMI, smoking status, and a binary indicator for the occurrence of strokes. Notably, the dataset denotes whether individuals have hypertension or heart disease through binary indicators. While some entries lack BMI information, the dataset's primary focus lies in understanding the factors associated with stroke occurrences.

The analysis of various factors influencing stroke risk in our dataset has yielded valuable insights. Age, average glucose level, heart disease, and hypertension emerged as significant predictors of stroke likelihood, providing a foundation for targeted preventive measures. The relationship between age, smoking, and strokes revealed a nuanced interplay, with age playing a role in elevating stroke chances, particularly in certain age groups. Gender and smoking status demonstrated a noteworthy connection, indicating that men, especially smokers, face a higher risk of strokes. The prevalence of strokes displayed a gender disparity, with a higher incidence among females. The interaction of age and BMI highlighted an increased risk of strokes as individuals age, particularly with a moderate BMI. Smoking emerged as a significant contributor to stroke risk, emphasizing the importance of promoting a smoke-free lifestyle. Additionally, the impact of other diseases, such as hypertension and heart disease, on stroke risk was explored. Finally, the examination of stroke cases by work type provided insights into potential occupational influences on health. These findings collectively contribute to a comprehensive understanding of stroke risk factors, informing targeted interventions, public health strategies, and further research endeavors.

## 4.2. Conclusion

The comprehensive exploration and visualization of the dataset have provided valuable insights into the complex dynamics influencing stroke risk. The following key conclusions can be drawn from the detailed analyses:

### 4.2.1. The Variables are Significant Predictors:

- Age emerges as a crucial factor, with a consistent increase in the likelihood of stroke with each unit increase.
- Average glucose level, heart disease, and hypertension also exhibit significant impacts on stroke risk, with higher levels or the presence of these conditions associated with elevated odds of stroke.

### 4.2.2. BMI and Marital Status Lack Significance:

In contrast, BMI and marital status (ever_married) do not demonstrate a statistically significant influence on stroke likelihood in this dataset.

### 4.2.3. Age and Smoking Interaction:

The relationship between age, smoking, and stroke risk is intricate, with varying trends across different age groups. While smoking appears to increase the chances of stroke in specific age ranges, the pattern fluctuates, emphasizing the need for a nuanced understanding of age-related risk factors.

### 4.2.4. Gender Disparity and Smoking Impact:

Men exhibit a higher likelihood of having a stroke compared to women, and this gender difference is more pronounced among individuals with a history of smoking. Smoking, especially in men, emerges as a significant contributing factor to stroke risk.

### 4.2.5. Age and BMI Interplay:

The hexagonal diagram illustrates that as individuals age and maintain a moderate BMI, their risk of experiencing a stroke increases. This observation highlights the interconnectedness of age, BMI, and stroke risk, paving the way for further exploration into the underlying factors driving this association.

### 4.2.6. Smoking as a Major Contributor to Stroke Risk:

Active smokers and former smokers demonstrate a notably higher prevalence of stroke compared to non-smokers. This underscores the critical role of smoking in elevating the risk of stroke and emphasizes the importance of promoting a smoke-free lifestyle for stroke prevention.

### 4.2.7. Impact of Other Diseases and Lifestyle Choices:

The analysis reveals that individuals with heart disease and smokers with heart disease face a significantly higher risk of stroke. Understanding these nuanced connections between health conditions and lifestyle choices provides valuable information for targeted interventions.

#### 4.2.8. Occupational Influence on Stroke Risk:

Certain occupational groups, particularly the self-employed, exhibit a higher percentage of strokes. The reasons behind these differences warrant further investigation, considering factors such as job-related stress, lifestyle, and access to healthcare.

In conclusion, this comprehensive exploration contributes to our understanding of the multifaceted factors influencing stroke risk. The findings underscore the importance of personalized approaches to stroke prevention, considering age, gender, lifestyle choices, and health conditions. Further research into the underlying mechanisms driving these associations can inform targeted interventions, ultimately contributing to improved public health outcomes.

## 4.3. Benefit

Understanding the various factors contributing to stroke risk, such as age, smoking status, and gender, offers several significant benefits:

1. **Personalized Healthcare Interventions**: The insights gained from the analysis allow for the development of more personalized healthcare interventions. Healthcare professionals can tailor preventive measures, screenings, and advice based on an individual's specific risk factors, improving the effectiveness of healthcare strategies.
2. **Public Health Planning:** Public health initiatives can be better designed and targeted to address specific risk factors prevalent in different demographic groups. This knowledge is crucial for shaping policies aimed at reducing the overall burden of strokes on society.
3. **Clinical Decision-Making**: Healthcare providers can use this information to make more informed clinical decisions. For example, understanding that certain age groups are more susceptible to strokes or that smoking significantly contributes to risk allows for better risk assessment and management.
4. **Educational Campaigns**: The findings can be used in educational campaigns to raise awareness about stroke risk factors. Public awareness can lead to lifestyle modifications, such as smoking cessation or healthier living, which can contribute to stroke prevention.
5. **Research Guidance**: The analysis provides a foundation for further research. Researchers can delve deeper into specific areas identified in the analysis, exploring additional factors or refining understanding in order to continually improve stroke prediction and prevention strategies.
6. **Resource Allocation**: Healthcare resources can be allocated more efficiently based on the identified risk factors. This includes directing resources towards groups or communities with higher vulnerability, optimizing the impact of healthcare interventions.
7. **Improved Health Outcomes**: Ultimately, the goal is to enhance health outcomes by identifying and addressing key determinants of stroke risk. This can lead to a reduction in the incidence of strokes, better management of stroke cases, and improved overall public health.

## Chapter 5 – SOURCE CODE

```
library(ggplot2)
library(corrplot)
library(dplyr)
library(tidyverse)
library(corrplot)
library(car)
library(ggcorrplot)
library(GGally)
library(hrbrthemes)
library(patchwork)
library(gridExtra)
library(RColorBrewer)

dev.off()

# Import dataset
df <- read.csv("StrokeData.csv")

# Data cleaning
unique_values <- unique(df$bmi)
print(unique_values)
df[df == "N/A"] <- NA
data <- na.omit(df)

# Checking Outlier
numeric_vars <- c("age", "avg_glucose_level", "bmi")
scatterplotMatrix(data[, sapply(data, is.numeric)])
ggplot(data, aes(x = avg_glucose_level, y = bmi)) +
  geom_point() +
  labs(title = "Scatterplot of avg_glucose_level vs BMI",
       x = "Average Glucose Level",
       y = "BMI")

# Checking Duplicate data
duplicate_rows <- duplicated(data)
duplicate_rows_data <- data[duplicate_rows, ]
print(duplicate_rows_data)

# Data Summary
summary(data)
table(data$smoking_status)
plot(data)

# Image 2.1. Scatter Plot of Avg_glucose_level vs BMI
ggplot(data, aes(x = avg_glucose_level, y = bmi)) +
  geom_point() +
```

```
  labs(title = "Scatter Plot: avg_glucose_level vs bmi",
       x = "Average Glucose Level",
       y = "BMI")

# Image 2.4. Histogram distribution of age distribution
set.seed(123)
data$age_skewed <- log(data$age + 1)
ggplot(data, aes(x = age_skewed)) +
  geom_histogram(binwidth = 0.2, fill = "skyblue", color = "black", alpha
= 0.7) +
  geom_density(aes(y = ..count.. * 0.2), color = "red", size = 1.5) +
  labs(title = "Distribution of Age Density", x = "Log(Age + 1)", y =
"Frequency")

# Image 2.5. histogram with a frequency polygon
glucose_distribution <- ggplot(data, aes(x = avg_glucose_level, y =
..count..)) +
  geom_histogram(binwidth = 10, fill = "lightgreen", color = "black",
alpha = 0.7) +
  geom_freqpoly(binwidth = 10, color = "blue", size = 1) +
  labs(title = "Average Glucose Level Distribution") +
  theme_minimal()
print(glucose_distribution)

# Image 2.6. Distribution of Gender
hist_combined <- ggplot(data %>% filter(gender %in% c("Female", "Male")),
aes(x = age, fill = gender)) +
  geom_histogram(binwidth = 5, position = "identity", alpha = 0.7,
show.legend = FALSE) +
  labs(title = "Distribution of Ages by Gender", x = "Age", y =
"Frequency") +
  scale_fill_manual(values = c("Female" = "pink", "Male" = "lightblue")) +
  facet_wrap(~gender, scales = "free_y", ncol = 2)
final_plot <- hist_combined +
  plot_layout(guides = "collect") +
  plot_annotation(title = "Age-Sex Inference")

final_plot

# Scatterplot glucose and age
scatter_matrix <- ggplot(data, aes(x = age, y = avg_glucose_level)) +
  geom_point() +
  labs(title = "Scatterplot glucose and age")
data$avg_glucose_level <- as.numeric(data$avg_glucose_level)
print(scatter_matrix)

# Image 2.7. distribution of glucose level by stroke
par(mfrow=c(1,2))
hist(data$avg_glucose_level[data$stroke == 0],
     main = "Distribution of Glucose Level (No Stroke)",
```

```
     xlab = "Average Glucose Level",
     col = "lightblue")

hist(data$avg_glucose_level[data$stroke == 1],
     main = "Distribution of Glucose Level (Stroke)",
     xlab = "Average Glucose Level",
     col = "lightcoral")


# Image 2.8. Correlation
str(data[, c("age", "hypertension", "heart_disease", "avg_glucose_level",
"bmi", "stroke")])
data$avg_glucose_level <- as.numeric(data$avg_glucose_level)
data$bmi <- as.numeric(data$bmi)
correlation_matrix <- cor(data[, c("age", "hypertension", "heart_disease",
"avg_glucose_level", "bmi", "stroke")])
corrplot(correlation_matrix, method = "color")
print(correlation_matrix)

# Image 2.9. Correlation 2
ggcorr(data, method = c("everything", "pearson"))

# Image 2.10. Parallel Coordinat Plot for the Iris Data
fig(20,20)
ggparcoord(data,
           columns = 1:11, groupColumn = 12,
           showPoints = TRUE,
           title = "Parallel Coordinate Plot for the Iris Data",
           alphaLines = 0.3
) +
  theme_ipsum()+
  theme(
    plot.title = element_text(size=10)
  )

# Image 3.1. Generalized Linear Model
log  <- glm(stroke ~ age + avg_glucose_level + heart_disease +
hypertension + bmi + ever_married + bmi, family=binomial (link="logit"),
data=data)
summary(log)

ggplot(data, aes(x = avg_glucose_level, y = bmi, color = smoking_status))
+
  geom_point() +
  facet_wrap(~ smoking_status, scales = "free") +
  labs(title = "Scatterplot of BMI vs Avg Glucose Level Faceted by Smoking
Status",
       x = "Average Glucose Level",
       y = "BMI")
```

```r
# Image 3.2. Percentage of Age Range by Smoking
data <- data %>%
  filter(smoking_status != "Unknown" & !is.na(age))

data$age_range <- cut(data$age, breaks = c(20, 30, 40, 50, 60, 70, 80,
90),
                      labels = c("21-30", "31-40", "41-50", "51-60", "61-
70", "71-80", "81-90"),
                      include.lowest = TRUE)

stroke_percentage <- data %>%
  group_by(age_range, smoking_status) %>%
  summarize(stroke_percentage = mean(stroke == 1, na.rm = TRUE) * 100)

ggplot(stroke_percentage, aes(x = age_range, y = stroke_percentage, fill =
smoking_status)) +
  geom_bar(stat = "identity", position = "dodge", alpha = 0.7) +
  labs(title = "Percentage of Stroke Based on Age Range and Smoking
Status",
       x = "Age Range",
       y = "Percentage of Stroke",
       fill = "Smoking Status") +
  scale_fill_manual(values = c("never smoked" = "#2ecc71", "formerly
smoked" = "#3498db", "smokes" = "#e74c3c"))

# Image 3.3. Gender and Smoking Status on Stroke Risk
data_filtered <- data[data$gender != "Other", ]

stroke_percentage <- data_filtered %>%
  group_by(gender, smoking_status) %>%
  summarize(stroke_percentage = mean(stroke == 1, na.rm = TRUE) * 100)

ggplot(stroke_percentage, aes(x = gender, y = stroke_percentage, fill =
smoking_status)) +
  geom_bar(stat = "identity", position = "dodge", alpha = 0.7) +
  labs(title = "Stroke Percentage by Gender and Smoking Status",
       x = "Gender",
       y = "Stroke Percentage",
       fill = "Smoking Status") +
  scale_fill_manual(values = c("never smoked" = "#2ecc71", "formerly
smoked" = "#3498db", "smokes" = "#e74c3c"))

# Image 3.4. Stroke by gender
combined_plot <- ggplot(data = combined_data, aes(x = gender, fill =
gender)) +
  geom_bar(stat = "count") +
  facet_grid(stroke ~ ., scales = "free_y", switch = "y") +
  ggtitle("Combined Bar Plot for Stroke and no Stroke") +
```

```
  theme(plot.title = element_text(size = 20, hjust = 0.5)) +
  scale_fill_manual(values = c("Male" = "#2ecc71", "Female" = "#3498db"))

options(repr.plot.width = 15, repr.plot.height = 8)
print(combined_plot)

# Image 3.5. Hex Diagram of Age and BMI for Stroke and No Stroke
ggplot(data, aes(x = age, y = bmi, fill = factor(stroke))) +
  geom_hex() +
  scale_fill_manual(values = c("0" = "#3498db", "1" = "#1f618d"), name =
"Stroke") +
  theme_bw()

# Image 3.6. Bar Plot of Smoking Status with Stroke and No Stroke
data$smoking_status <- factor(data$smoking_status, levels = c("never
smoked", "formerly smoked", "smokes"))

ggplot(data, aes(x = smoking_status, fill = factor(stroke))) +
  geom_bar(stat = "count", position = "dodge") +
  scale_fill_manual(values = c("0" = "#2ecc71", "1" = "#3498db"), name =
"Stroke") +
  labs(title = "Bar Plot of Smoking Status and Stroke",
       x = "Smoking Status",
       y = "Count") +
  theme_minimal()

# Image 3.7. Relation of Other Diseases and Smoking Status to Stroke Risk
never_smoked <- c(13.89, 13.58)
formerly_smoked <- c(15.45, 17.14)
smokes <- c(12.20, 23.64)
disease <- c('Hypertension', 'Heart Disease')

data <- data.frame(disease, never_smoked, formerly_smoked, smokes)

colors <- c("#2ecc71", "#3498db", "#e74c3c")

data_long <- tidyr::gather(data, key = "smoke_status", value =
"percentage", -disease)

ggplot(data_long, aes(x = disease, y = percentage, fill = smoke_status)) +
  geom_bar(position = "dodge", stat = "identity", width = 0.6) +
  labs(x = 'Other diseases',
       y = 'Percentage (%)',
       title = 'Percentage of people that have Stroke and other disease
with smoke status') +
  scale_fill_manual(values = colors, name = "Smoking Status",
                    labels = c("Never Smoked", "Formerly Smoked",
"Smokes")) +
  theme_minimal()
```

```r
# Image 3.8. Bar Plot of Stroke Cases Percentage by Work Type
options(repr.plot.width = 5, repr.plot.height = 5)

percentage_data <- df %>%
  group_by(work_type) %>%
  summarize(percentage_stroke = sum(stroke == 1) / n() * 100)

ggplot(percentage_data, aes(x = work_type, y = percentage_stroke)) +
  geom_bar(stat = "identity", fill = "#1f78b4") +
  labs(x = 'Work Type', y = 'Percentage of Stroke', title = 'Percentage of
Stroke Cases by Work Type') +
  theme_minimal()
```