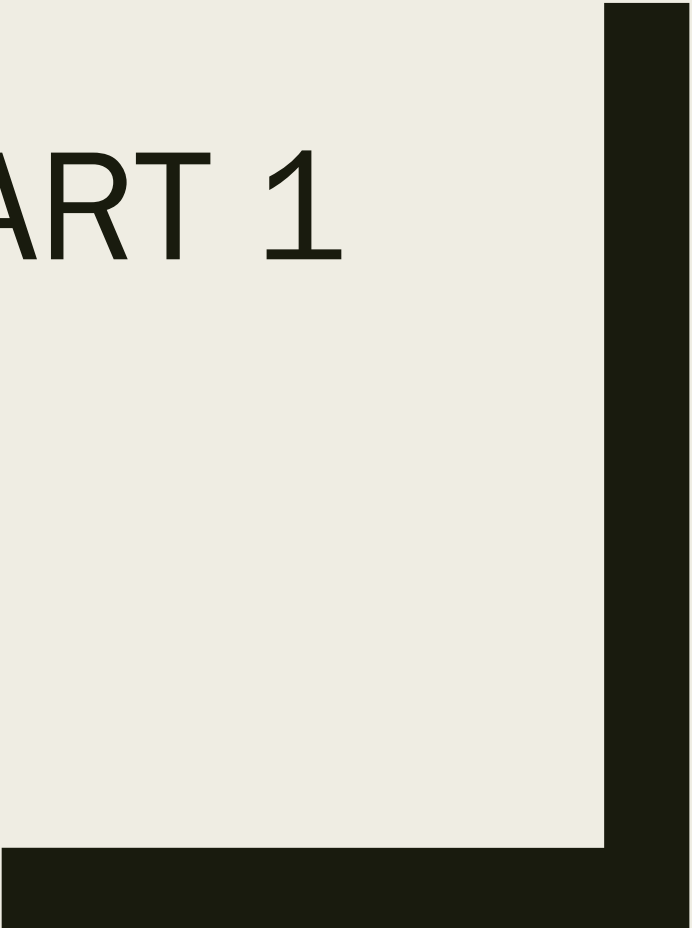




FINAL PROJECT PART 1

Guangyu Song 1006707459



Introduction

- **Research Question:**

Exploring the association between whether **Stroke** and these various factors, which include **Age, BMI, Average glucose level, Gender, Smoking status** and **Hypertension**.

- **Response:** Stroke

- **Predictor:** Age, BMI, Avg glucose level, Gender, Smoking status, Hypertension

- **potential confounders:** Heart disease(Heart disease is not only related to whether or not you have had a stroke, but heart disease may indirectly lead to Hypertension)

- **Regression Model:** Generalized Linear Model (GLM)

Introduction

■ Relevance or Importance of Research Question:

Stroke is a main indicator of neurological disability and controlling risk factors can significantly reduce the severity of stroke (Kelly-Hayes, 2010). It is commonly considered that stroke is highly linked to age, health status, and lifestyle. Understanding how these factors correlate with stroke incidence helps identify at-risk populations, helps prevent stroke, and manages risk factors among different populations. To be more specific, factors including age, BMI, glucose level, gender, smoking status, and hypertension are potential indicators of interest.

Supporting Literature

- **Title:**

Influence of Age and Health Behaviors on Stroke Risk: Lessons from Longitudinal Studies

- **Description:**

The study seeks to answer how frequently stroke occurs over a person's lifetime based on gender and age (research question/variables) (Kelly-Hayes, 2010). Using data from the Framingham Heart Study (FHS) (dataset), the researchers analyze trends across different populations over time (method) (Kelly-Hayes, 2010). The study concludes that risk increases significantly with age. Also, men have a higher stroke risk at younger ages, while women are more likely to experience stroke at older ages (conclusion) (Kelly-Hayes, 2010).

- **Similarity:** Both include age, gender, hypertension, and smoking as indicators. The study uses poor diet, and we use BMI, which is a similar aspect regarding weight.

- **Difference:** The study conducts its study based on longitude studies, meta-analyses, and epidemiological evidence that link various risk factors to the incidence and outcomes of stroke. The method we are implementing is GLM, which takes all potential indicators in a logistic prediction model.

Supporting Literature

- **Title:**

Body Mass Index Measured Repeatedly over 42 Years as a Risk Factor for Ischemic Stroke: The HUNT Study

- **Description:**

The study aims analyze the associations of weight and stroke risks (research question /variables) (Horn et al., 2023). Using data from the HUNT study (dataset), researchers calculated average BMI values and group-based trajectory models and related these to the risk of stroke (method) (Horn et al., 2023). The study observed a higher risk for stroke among participants with obesity (BMI >30 kg/m²) or overweight (BMI 25–30 kg/m²) over adulthood (conclusion) (Horn et al., 2023).

- **Similarity:** Both analyze the association between stroke and weight, especially excess weight. Both use BMI as the indicator.

- **Difference:** The study is using the Trøndelag Health Study (HUNT), which is an ongoing longitude population-based study. Analyses are conducted over 20 years. We are using one-time data and a GLM method considering factors besides weight for analyses.

Supporting Literature

- **Title:**

Glucose and Acute Stroke : Evidence for an Interlude

- **Description:**

The study investigates how glucose levels affect stroke risk (research question/variables) (Helgason, 1988). By analyzing clinical data can categorizing them by blood glucose levels (dataset), the researchers used risk assessments and correlation analyses to identify patterns related to stroke severity (method) (Helgason, 1988). The study concludes that both hyperglycemia and hypoglycemia significantly affect stroke risk (conclusion) (Helgason, 1988).

- **Similarity:** Both analyze the association between blood glucose level and stroke.

- **Difference:** The study uses comparative analyses and correlation analyses where it compares the outcomes of strokes in individuals with normal blood glucose levels, hyperglycemia, and hypoglycemia, and finds the strength of the relationship between glucose levels and stroke. We are using a GLM method where we take factors including not only blood glucose level to predict stroke risk.

Data Description

- Title: Stroke Prediction Dataset
- Number Of Observation: 5111
- The website from which I chosen data was obtained:
<https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset/data>

Variable	Description
ID	unique identifier
Gender	"Male", "Female" or "Other"
Age	age of the patient
Hypertension	If patients have hypertension(0, 1)
Heart disease	If patients have heart disease(0, 1)
Ever married	"No" or "Yes"
Work type	"children", "Govt job", "Never worked", "Private" or "Self-employed"
Residence type	"Rural" or "Urban"
Avg glucose level	average glucose level in blood
BMI	body mass index
Smoking status	"formerly smoked", "never smoked", "smokes" or "Unknown"
Stroke	1 if the patient had a stroke or 0 if not

All variables are in the data set

Data Description – Variables I choosed

- Response: **Stroke** (*Binary Categorical Variable*) stroke ~ 1, not stroke ~ 0
- Predictor: (Numerical Variable) :
 - **Age:** This represents how old the person is.
 - **BMI:** A score that use patients' height and weight to work out if their weight is healthy
 - **Avg glucose level:** This number shows the typical amount of sugar (glucose) in the person's blood over a period.
- Predictor: (Categorical Variable) :
 - **Gender:** This indicates the gender with which the patient identifies. The options are "Male," "Female," or "Other,"
 - **Smoking status:** This tells us about the person's smoking habits. It categorizes individuals into four groups: "formerly smoked", "never smoked", "smokes" or "Unknown"
 - **Hypertension:** This is a simple yes-or-no indicator (represented by 1 for 'yes' and 0 for 'no') that signifies whether the patient has hypertension

Data Description - EDA

■ Numerical Variable:

	Age	BMI	Avg glucose level
min	0.08	10.30	55.12
1 st Qu	25.00	23.50	77.07
median	44.00	28.10	91.68
mean	42.87	28.89	105.31
3rd Qu	60.00	33.10	113.57
max	82.00	97.60	271.74

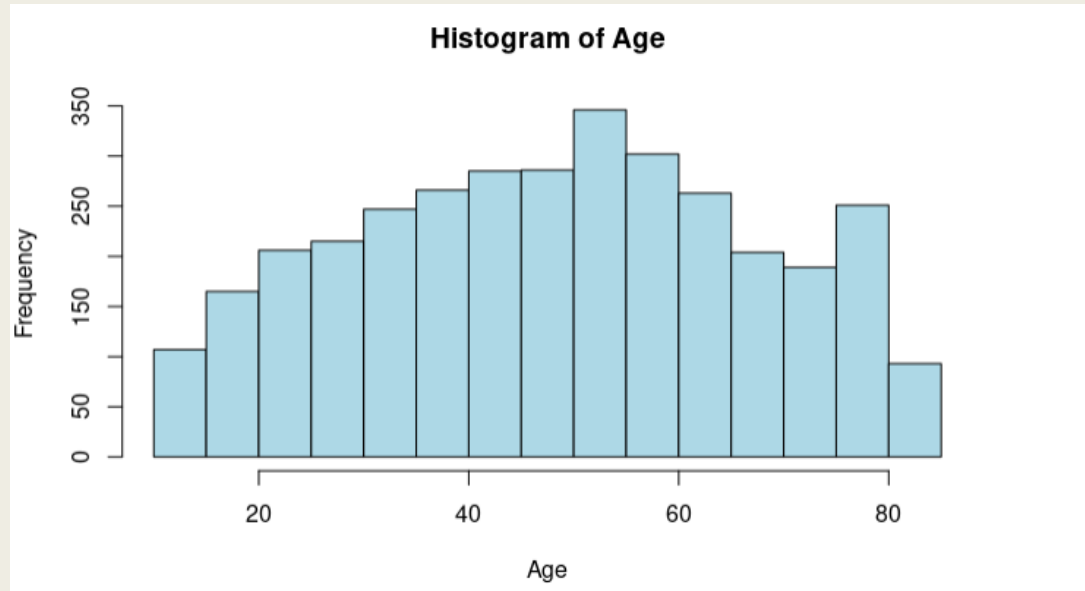
■ Categorical Variable:

	Gender			Smoking status				Hypertension	
Classification	Male	Female	Other	Never Smoked	Formerly Smoked	Smokes	Unknown	Yes(1)	No(0)
Proportion%	0.409	0.590	0.0002	0.377	0.171	0.150	0.302	0.0918	0.9081

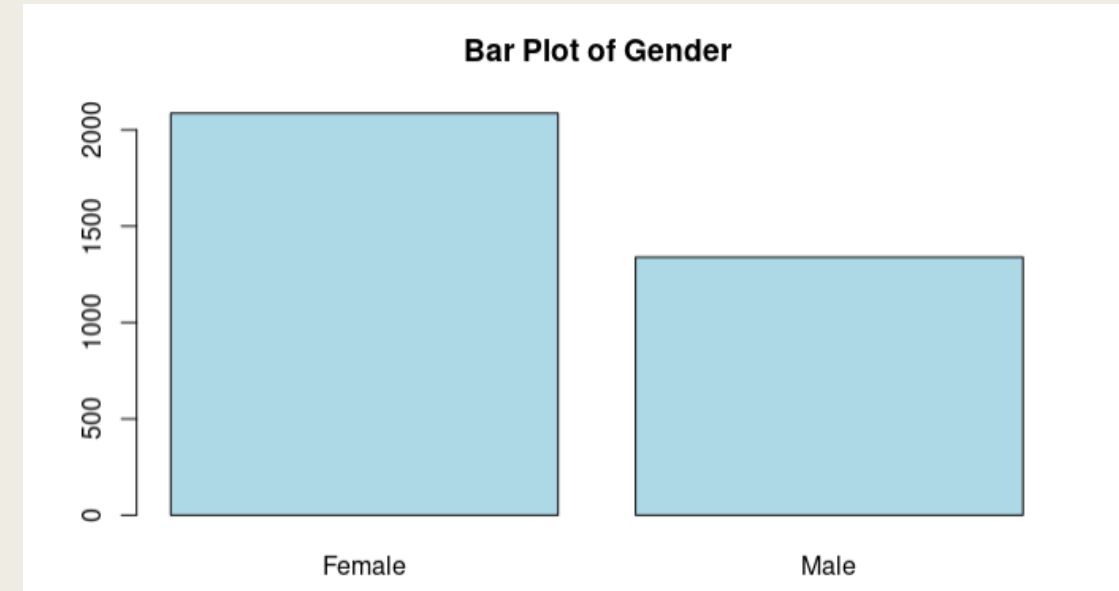
Data Description - EDA

- For the Gender variable, Through the analysis in R, we find that gender = "Other" has only one sample. For the convenience of later analysis, we can choose to delete the line where gender = "Other", because a single line of data has little impact on the analysis result.
- For the Smoking status = Unknown. Because we wanted to study the relationship between the frequency of smoking and whether or not we had a stroke, Unknown did not contribute anything to our study and could be considered Missing data. So we remove rows for Smoking status = Unknown

Data Description - EDA

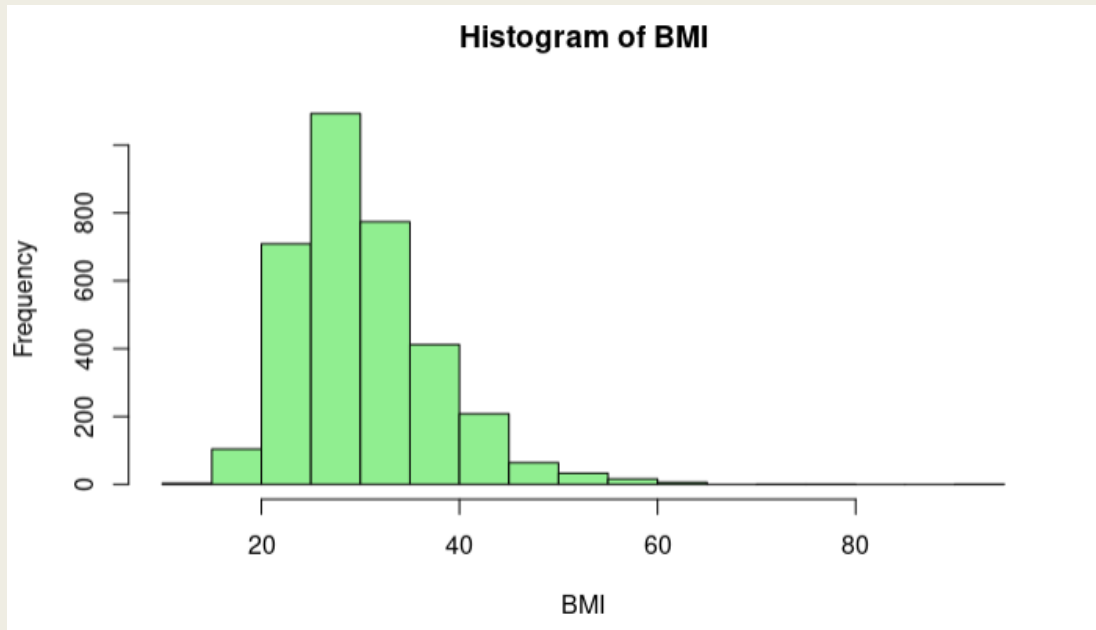


histogram of the frequency distribution of age in the dataset. Most common age group is between 50-60 years old, and there are also relatively large numbers of people around 80 years old.

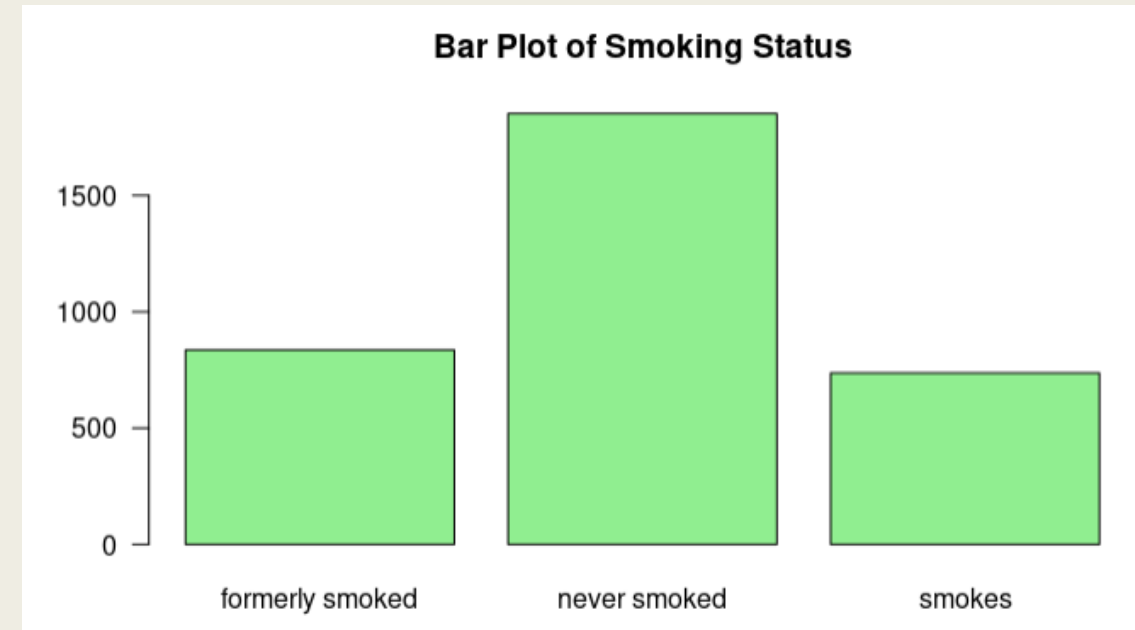


A bar plot of gender that illustrates the count of patients by gender. The proportion of women in the sample is greater than that of men.

Data Description - EDA

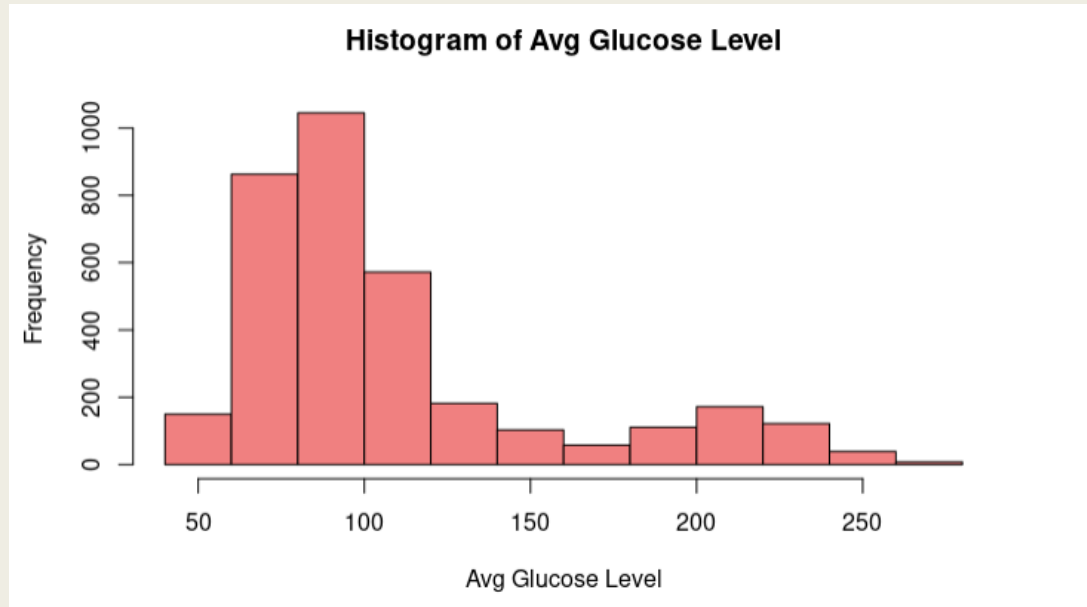


Histogram of BMI, showing a distribution that peaks between approximately 20 to 30.

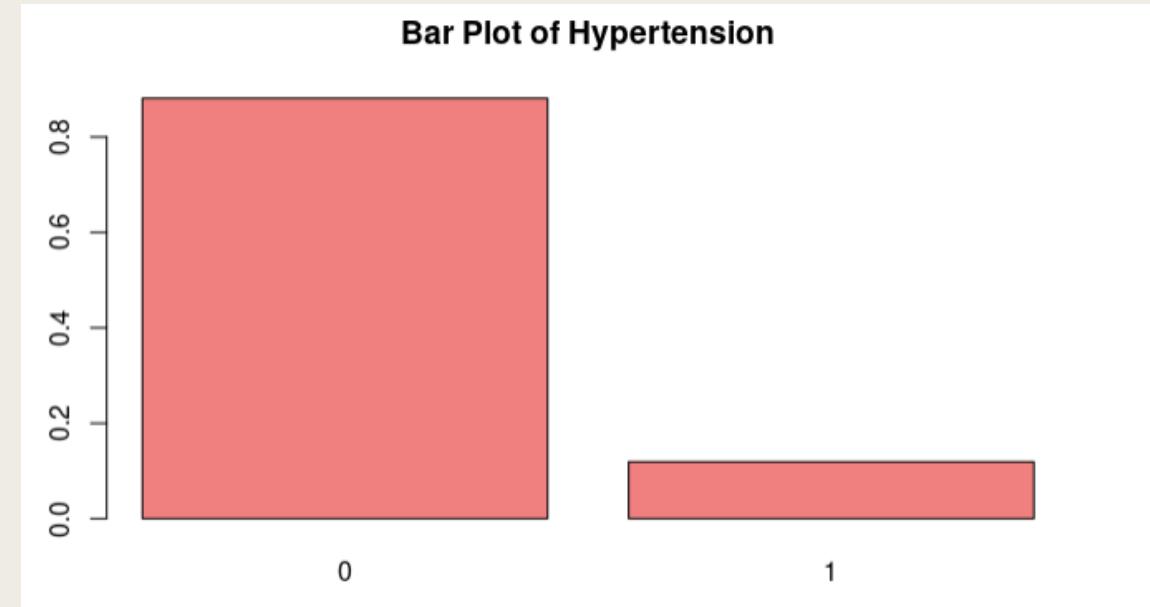


Bar plot of smoking status, indicating that the largest category consists of patients who have never smoked, followed by those who formerly smoked, with the fewest people being current smokers.

Data Description - EDA

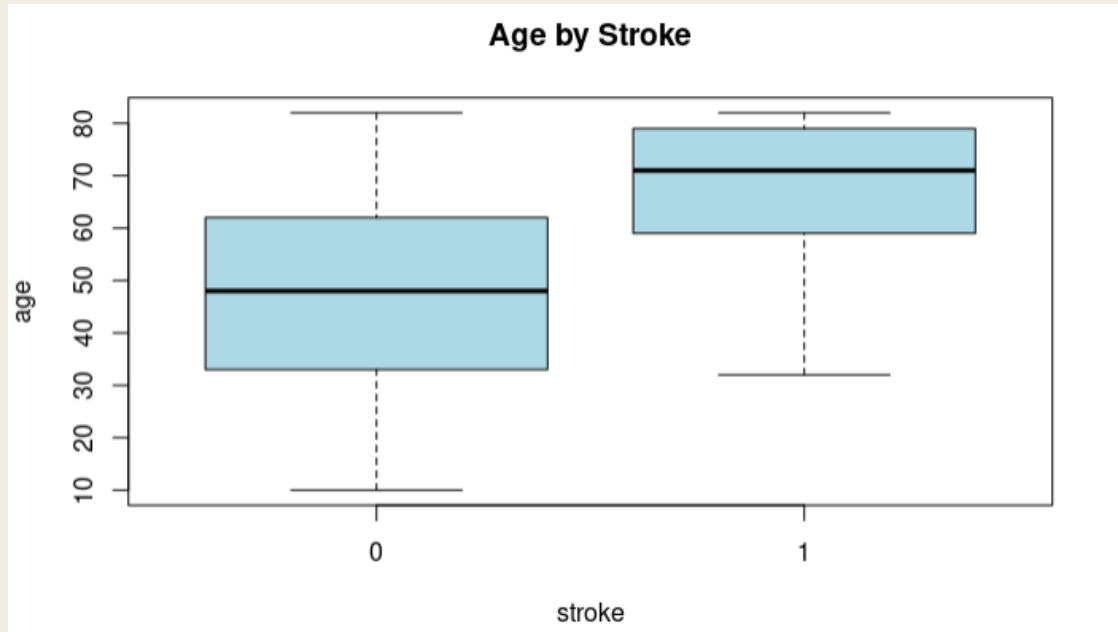


Histogram of average glucose level shows that the peak frequency between approximately 50 to 100 mg/dL, which is an important consideration when assessing risk factors for stroke.

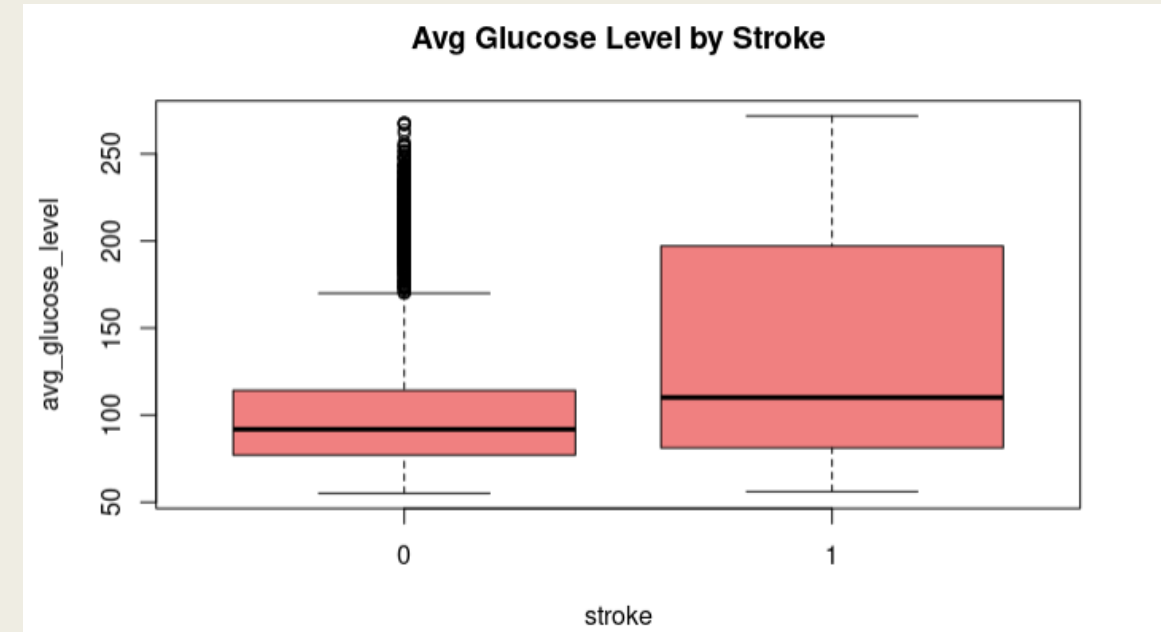


Bar plot of hypertension indicates that the proportion of individuals without hypertension ('0') is larger compared to those with hypertension ('1'). The difference can help us find the potential impact of hypertension on stroke occurrence.

Data Description - EDA

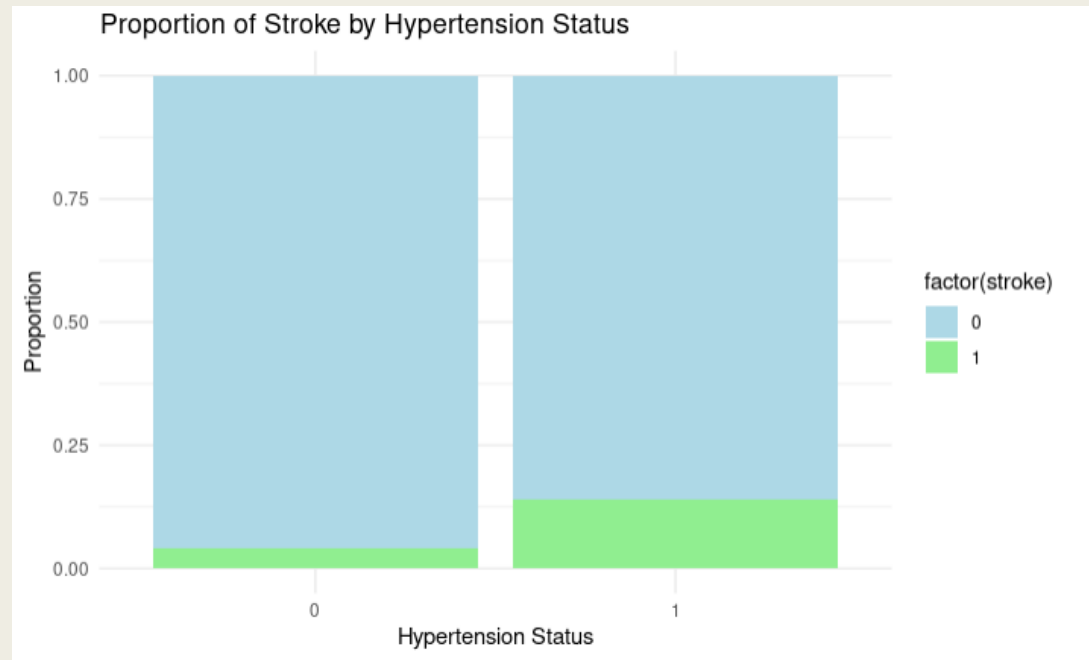


Boxplot comparing the distribution of ages between patients who have had a stroke and those who have not. The boxplot shows a higher median age for patients who have had a stroke

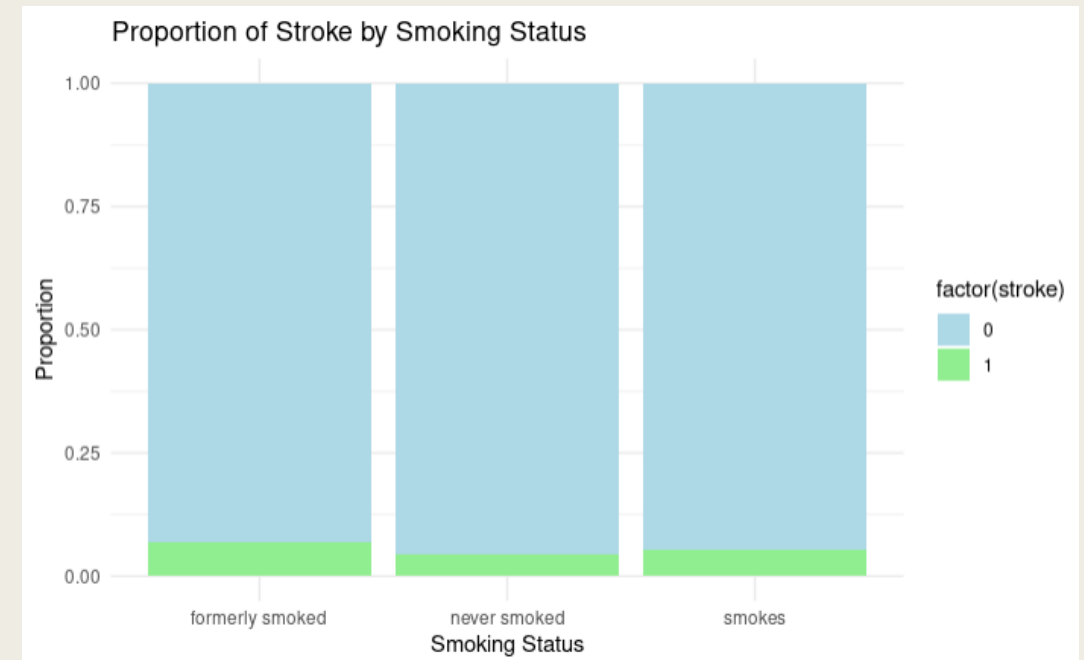


Boxplot of average glucose level by stroke status. It shows a higher median and wider range of average glucose levels in patients who have had a stroke compared to those who have not.

Data Description - EDA



Bar plot shows that the proportion of patients who had a stroke and those who did not into two groups: with hypertension and without hypertension. We can find that the proportion of stroke with hypertension is higher than without hypertension.



Bar plot shows that the proportion of patients who had a stroke and those who did not into three groups of smoking status. We can find that the proportion with never smoked is lowest.

Discussion

- How each predictor works for the model and research question?
 - **Age:** Age is a key predictor as the risk of stroke increases with age. This variable can help establish the trend or pattern in stroke incidence as age increases.
 - **BMI:** BMI is used as an indicator of healthy weight. Since overweight are known risk factors for stroke, this variable could provide insights of how body weight relates to stroke risk.
 - **Average Glucose Level:** *The increase of* blood glucose levels can indicate a risk of diabetes, which is also a risk factor for stroke.
 - **Gender:** Gender might influence the risk of stroke due to the difference of biological factors. Including gender in the model can help understand if there's a significant difference in stroke incidence between males, females
 - **Smoking Status:** Smoking status is a well-known risk factor for stroke. This variable can help measure the extent to which smoking contributes to stroke risk.
 - **Hypertension:** As a major risk factor for stroke, hypertension is a crucial binary predictor. The model can assess the strength of the association between hypertension and stroke occurrence.

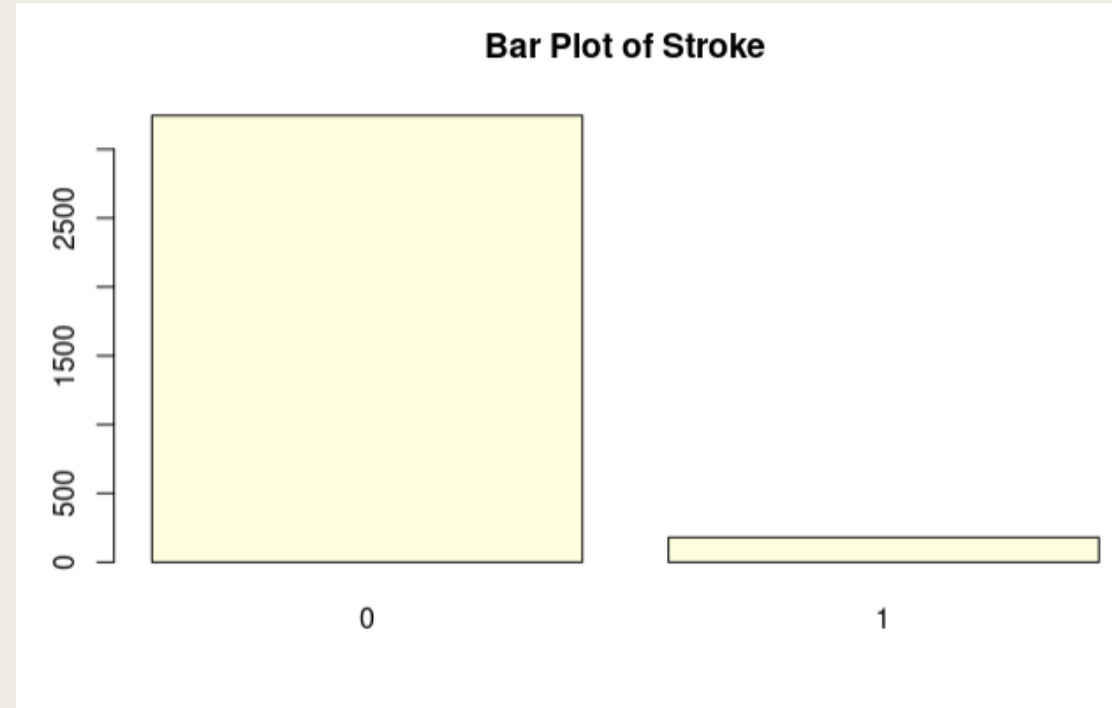
Discussion

- Why we could use GLM model to answer our research question?
 - *Our response Stroke is a binary categorical variables* (stroke = 1, no stroke = 0), which follows Bernoulli distribution, GLM is relatively suitable for this kind of data.
 - GLMs can handle both numerical and categorical variables, making them ideal for analyzing the diverse set of predictors we have: age, BMI, average glucose level, gender, smoking status, and hypertension.
 - GLM can produce various types of graphs to visualize the relationships between predictors and the binary response Stroke.
 - Since we've been studying GLMs in class, applying them to a real-world scenario will deepen our understanding.

Discussion

■ Assumptions of GLM

- ***Binomial distributed for outcome:***
Based on the bar plot of Stroke, our response Stroke is a binary categorical variable.
- ***Independence for outcome:***
For our dataset, observations of patients are independent of each other. So the stroke incidence of each patient is independent.
- ***Linearity:***
Our GLM model satisfies linearity assumption



Reference

1. Helgason, C. M. (1988). Blood glucose and stroke. *Stroke*, 19(8), 1049.
<https://doi.org/10.1161/01.STR.19.8.1049>
2. Horn, J. W., Feng, T., Mørkedal, B., Aune, D., Strand, L. B., Horn, J., Mukamal, K. J., & Janszky, I. (2023). Body Mass Index Measured Repeatedly over 42 Years as a Risk Factor for Ischemic Stroke: The HUNT Study. *Nutrients*, 15(5), 1232.
<https://doi.org/10.3390/nu15051232>
3. Kelly-Hayes M. (2010). Influence of age and health behaviors on stroke risk: lessons from longitudinal studies. *Journal of the American Geriatrics Society*, 58 Suppl 2(Suppl 2), S325–S328. <https://doi.org/10.1111/j.1532-5415.2010.02915.x>