# Identifying Risk Factors for Stroke Disease

**Tam Nguyen**

**Uyen Tran**

## 1. Introduction

According to the World Health Organisation (WHO) (n.d), stroke is a leading cause of death and disability worldwide, with almost 15 million people worldwide suffering from the disease every year, among whom 5.5 million died and another 5 million are left permanently disabled. A stroke is a medical emergency that occurs when blood flow to the brain is disrupted, leading to brain damage and neurological deficits. Stroke is a major cause of disability and death worldwide. Several factors affect Stroke Disease including age, gender, hypertension, smoking, and cardiovascular disease. Besides, there is growing evidence that socioeconomic factors, such as income, education, and occupation, may also play a role in stroke risk. The proposed study is important because it could help identify patients at high risk of stroke and allow for targeted interventions to prevent stroke and improve patient outcomes. The study's findings could also inform the development of more effective risk prediction models and preventive strategies for stroke. Additionally, understanding the role of socioeconomic factors in stroke risk can help address health disparities and ensure that preventive interventions reach those who need them most. The project aims to investigate the following questions:

- How do demographic factors affect the risk of developing Stroke disease?
- Are hypertension, heart disease, and smoking habits related to Stroke risk?
- Is there any interaction between socio-economic factors (such as work type, place of residence) and other biological risks of Stroke (such as BMI)?

This can be solved by creating several different types of visualizations on data on stroke risk factors and socioeconomic factors to identify patterns and relationships between these variables. This analysis can inform the development of more accurate stroke risk and targeted preventive interventions.

## 2. Data descriptions

The dataset being used in this study is obtained from Kaggle, titled “Stroke Prediction Dataset” by Fedesoriano. It includes information on key parameters that are relevant to stroke development such as age, gender, heart disease, married status, work type, BMI, and smoking habits of over five thousand individuals. Specific details about the variables’ description and data type are demonstrated in the below table:

| Data type | Variables | Description |
|---|---|---|
| Numerical | -age | age of the patient |
| | -avg_glucose_level | average glucose level in blood |
| | -bmi | body mass index |
| Categorical | -id | unique identifier of patient |
| | -gender | "Male", "Female" or "Other" |
| | -hypertension | “0” if the patient doesn't have hypertension, “1” if the patient has hypertension |
| | -heart_disease | “0” if the patient doesn't have any heart diseases, 1 if the patient has a heart disease |
| | -ever_married | "No" if the patient is not married,<br>"Yes"  if the patient is married |
| | -work_type | "children": patient is children<br>"Govt_jov": those working in government sector |

| | | |
|---|---|---|
| | | "Never_worked": those out of the workforce<br>"Private": those working in the private sector<br>"Self-employed": business owners. |
| | -Residence_type | "Rural": those living in rural areas,<br>"Urban": those living in urban areas. |
| | -smoking_status | "formerly smoked", "never smoked", "smokes" or "Unknown" |
| | -stroke | “1” if the patient had a stroke or “0” if not |

Before conducting any analysis, we pre-processed the data by checking for missing values and inconsistencies in the data. We found that there were missing values in the BMI variable, and so we used imputation techniques to fill in the missing values. We also converted categorical variables, such as gender, work type, residence type, and smoking status, into binary variables. After preprocessing the data, we conducted exploratory data analysis, particularly using descriptive statistics and visualizations to summarize the data and identify any potential outliers. For the visualizations, we use Tableau as it allows us to explore different approaches and aid decision-making when choosing the most appropriate types of graphs for different combinations of data types.

## 3. Data exploratory analysis

### 3.1. Summary of key descriptive statistics

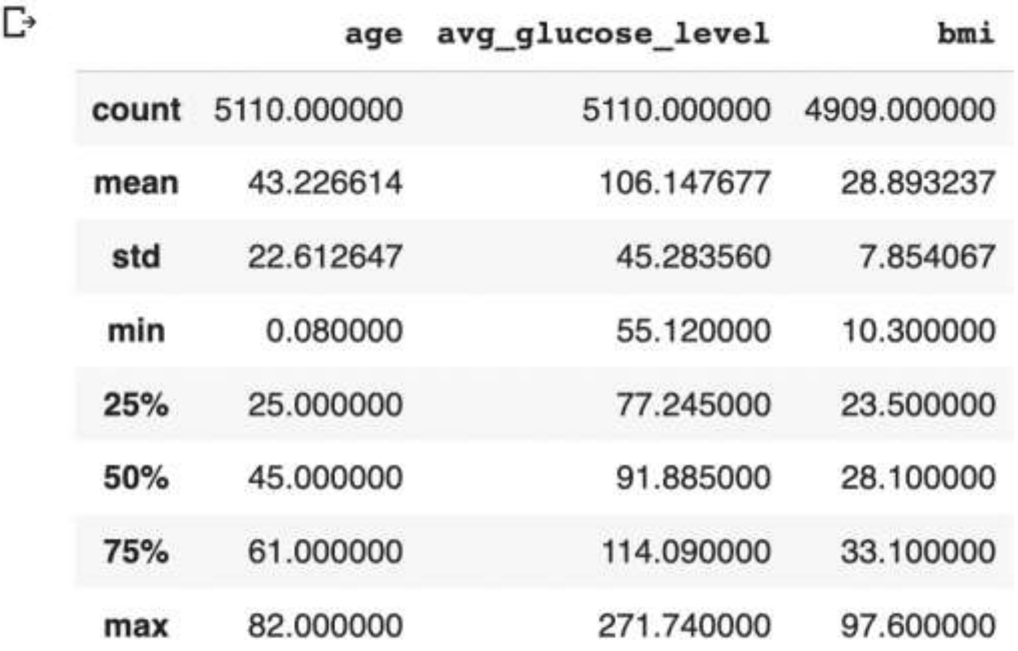
The summary of descriptive statistics in this study was performed using the pandas library in Python with the function panda.DataFrame.describe() function (Appendix 1 &

Appendix 2). Figure 1 demonstrates key statistics for the three important numeric variables in the dataset. It can be observed from the *age* column that the average age of 5110 individuals in the dataset is approximately 43.23 years (mean value), the median value is 45, the youngest individuals are newborns (minimum value of 0.08) and the oldest individuals are 82 years old (maximum value of 82). The age values have a spread of around 22.61 years from the average age.

Regarding the *average glucose level*, the average glucose level of all individuals in the sample is approximately 106.15 mg/dL, with a spread of approximately 45.28 mg/dL. Based on the indication of Centers for Diseases control and Prevention (*Diabetes Testing* 2023), the average glucose level in the sample falls within the prediabetes range (100 - 125 mg/dL). While the minimum recorded average glucose level is 55.12 mg/dL, the maximum level is 271.74 mg/dL, suggesting high levels of variation in data.

As for *BMI*, the average BMI of individuals in the dataset is approximately 28.89, which is relatively higher than the standard BMI for healthy people according to Centers for Diseases control and Prevention (*Assessing Your Weight* 2022). The BMI values have a spread of approximately 7.85 units around the mean, with a minimum record of 10.3, and a maximum record of 97.6. This suggests the presence of potential outliers in the dataset.

In general, the data suggests that the majority of individuals in the sample are middle-aged (with a median age of 45) and have glucose levels and BMI within a slightly overweight range. The wide range of values for both glucose levels and BMI indicates that there is likely a high level of variability in the population. Additionally, the presence of missing values for BMI (491 missing values out of 5110) should be taken into consideration when interpreting the results to avoid potential bias.



|  | age | avg_glucose_level | bmi |
|---|---|---|---|
| count | 5110.000000 | 5110.000000 | 4909.000000 |
| mean | 43.226614 | 106.147677 | 28.893237 |
| std | 22.612647 | 45.283560 | 7.854067 |
| min | 0.080000 | 55.120000 | 10.300000 |
| 25% | 25.000000 | 77.245000 | 23.500000 |
| 50% | 45.000000 | 91.885000 | 28.100000 |
| 75% | 61.000000 | 114.090000 | 33.100000 |
| max | 82.000000 | 271.740000 | 97.600000 |

Figure 1. Summary Statistics of Age, Average Glucose Level and BMI

### 3.2. Exploratory Data Analysis

### 3.2.1. Stroke Patient Demographic Analysis

First, we explore the proportion of stroke patients among the total examined population using a bar chart, which is ideal for demonstrating the part-to whole relationship. As can be seen from Figure 2, in a database of over 5000 observations, 4.87% of people are stroke patients.

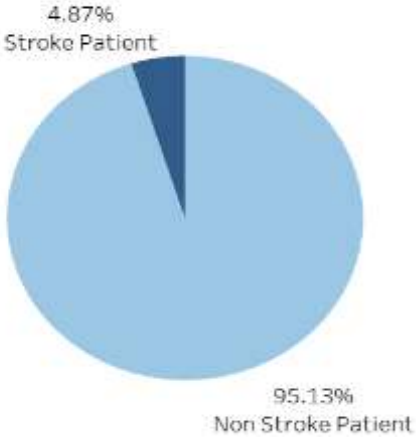


Figure 2. Percentage of Stroke Patients

Next, a stacked bar chart is used to compare the number of stroke incidents among different age ranges (Figure 3). In this scenario, since the age range of participants in the sample is wide (from 0 to 82 years old), the bar chart can function as a histogram which shows the relative distribution of stroke count across  different age groups. This provides an informative overview on the scope of the problem, particularly, the higher the age, the higher the chance of a person to suffer a stroke. Additionally, it is easy to point out that elders within the age range from 78 to 81 are at highest risk, with the count of stroke incidents being twofold compared to those from 50 to 77 and almost threefold  compared to those below 50.

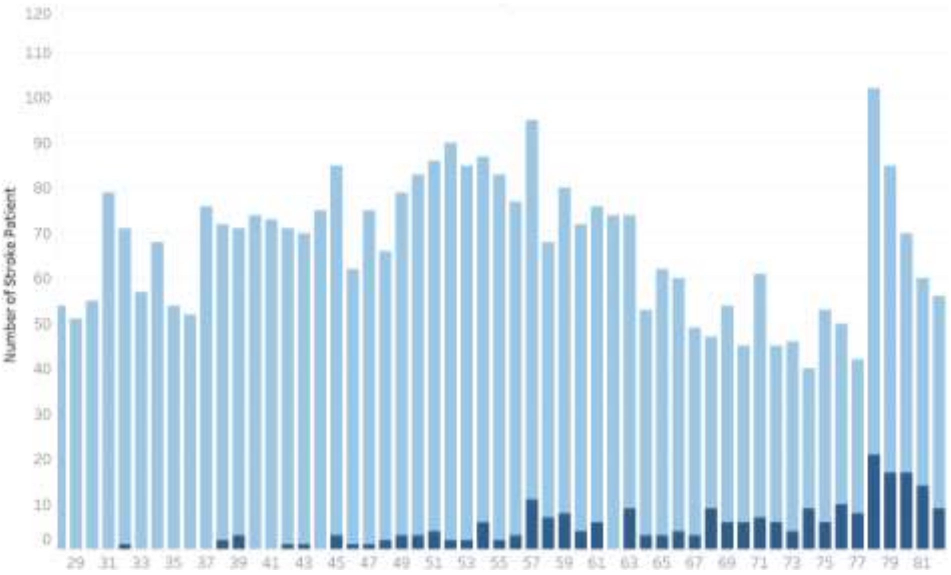


Figure 3. Number of Stroke Patients by Age

Similarly, bar charts are also used to provide an overview of the distribution of place of residence, gender, and marital status related to stroke patients. The data is evenly distributed between urban and rural areas, with 50.81% of patients residing in urban areas and 49.18% in rural areas (Figure 4). On the other hand, stroke distribution by gender shows a clear difference, indicating that females are more likely to suffer stroke than males (Figure 5). Similarly, more married people suffer from stroke than those who are not married (Figure 6). These findings make logical sense, given that strokes are in direct proportion to age as indicated in the above finding, and women generally live longer than men (Rexrode et al., 2022). Besides, there are several sex-specific risk factors that place women in a disproportionate burden such as the ability to be pregnant (which associates with higher risks of high blood pressure, depression and hormonal fluctuations) (*Women and Stroke* 2023)

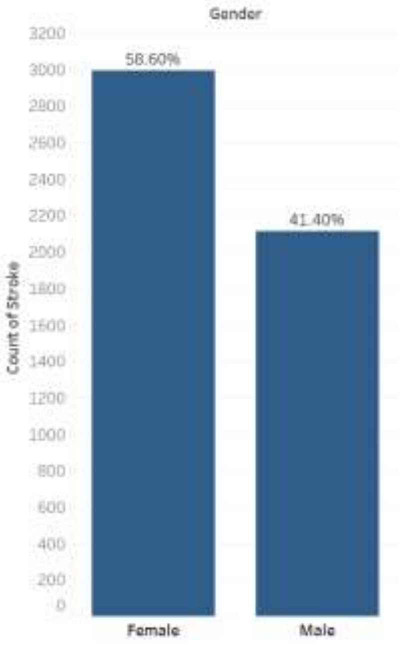


Figure 4. Percent of Patients by Gender

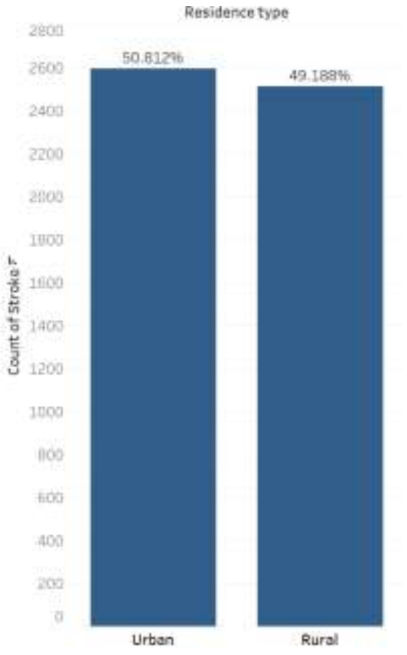


Figure 5. Percent of Patients by Place of Residence

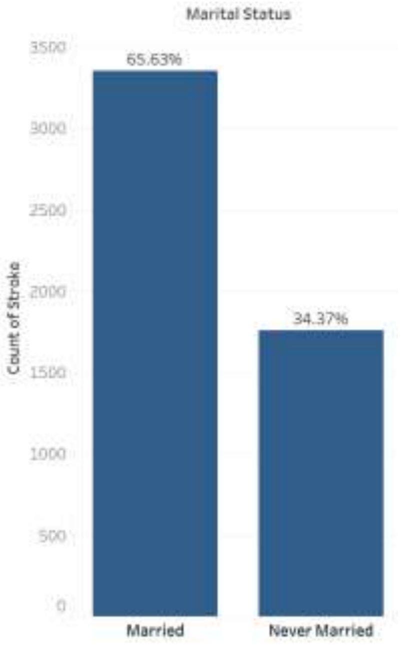


Figure 6. Percent of Patient by Marital Status

### 3.2.2. Stoke patient analysis - Hypertension, Heart Disease, and Smoking Habit

Smoking, hypertension and heart disease have been identified as major risk factors of stroke in many previous studies. In this study, we used a bar chart to discover the proportion of people associated with each of those factors and their likelihood of developing stroke. The plots suggest that among stroke patients, 43.75% were smokers (Figure 7), 18.88% had heart disease (Figure 8), 26.51% had hypertension (Figure 9).

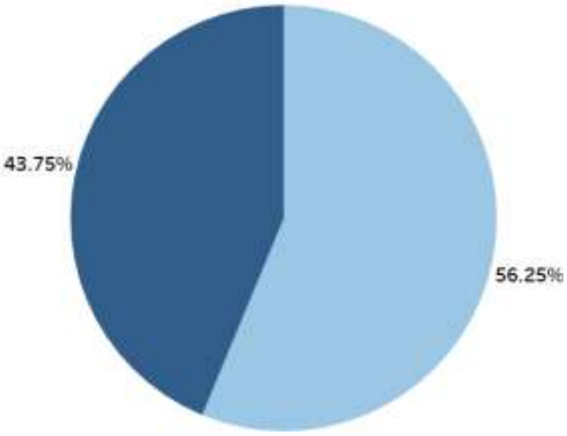


Figure 7. Relation between Stroke Patient and Smoke

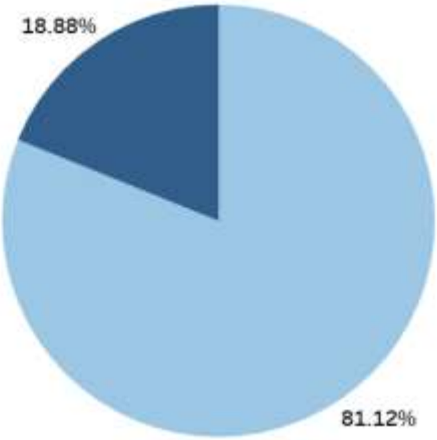


Figure 8. Relation between Stroke Patient and Heart Disease

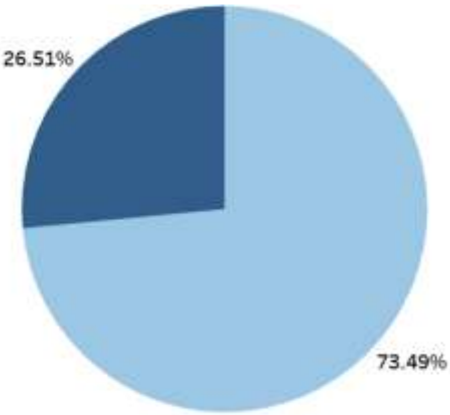


Figure 9. Relation between Stroke Patient and Hypertension

### 3.2.3. Stoke patient analysis - Relationship between biological factors and Age

In the summary statistics, we discovered that the average BMI and glucose level of the participants are slightly high. BMI and glucose level are associated with heart disease, and while heart disease is associated with stroke risks in previous papers, we did not observe such pattern through pie charts. Thus, we employed the use of scatter plots to verify the relationship between BMI and glucose level and age and use that as the bridge to indicate the relationship between Heart Disease and stroke risk. Figure 10 evidence that both the average BMI and glucose level are higher in older individuals compared to middle-aged and young adults. Besides, while the group 70-80 years old has the highest occurrence of strokes, the rate of stroke patients in groups 50-60 and 60-70 are fairly the same.

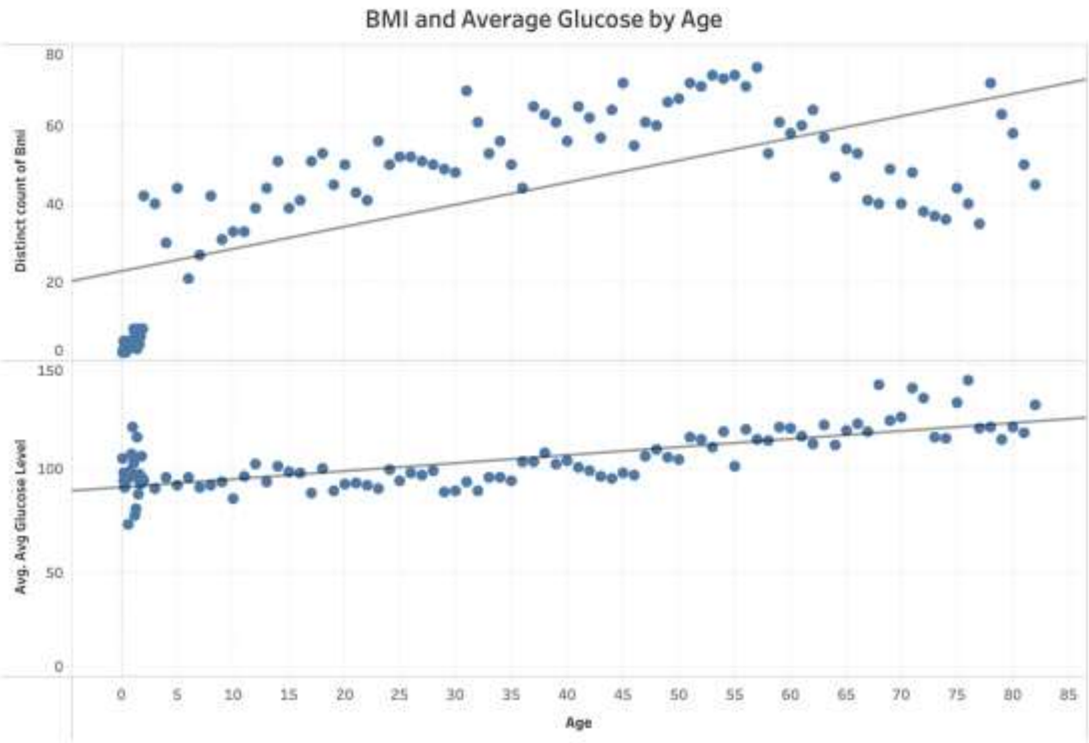


Figure 10. Relation between BMI and Glucose Level by Age

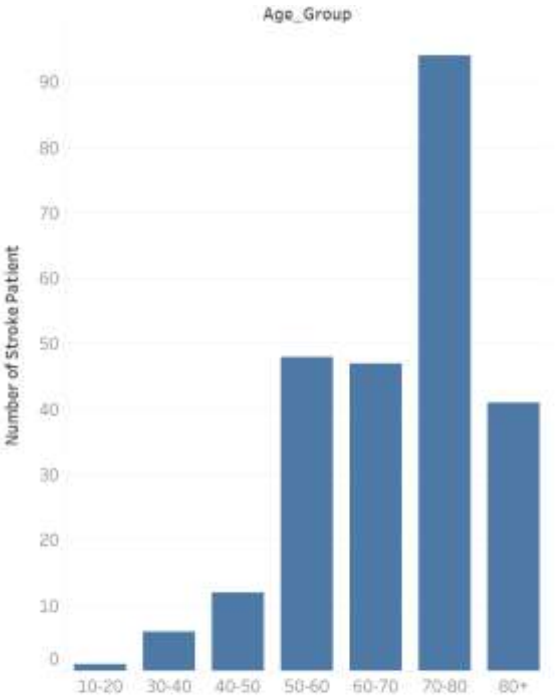


Figure 11. Stroke Patient by Age Group

### 3.2.4. Stoke patient analysis - Relationship between socio-economic factor and BMI

The Figure 12 indicates that the median BMI is higher for the private work type compared to the other work types, and the distribution of BMI values is wider for the self-employed work type. This means that individuals in the private work type may have a higher risk of being overweight or obese compared to other work types. Additionally, as we can see in Figure 13, individuals living in urban areas tend to have slightly higher BMI values than those in rural areas.

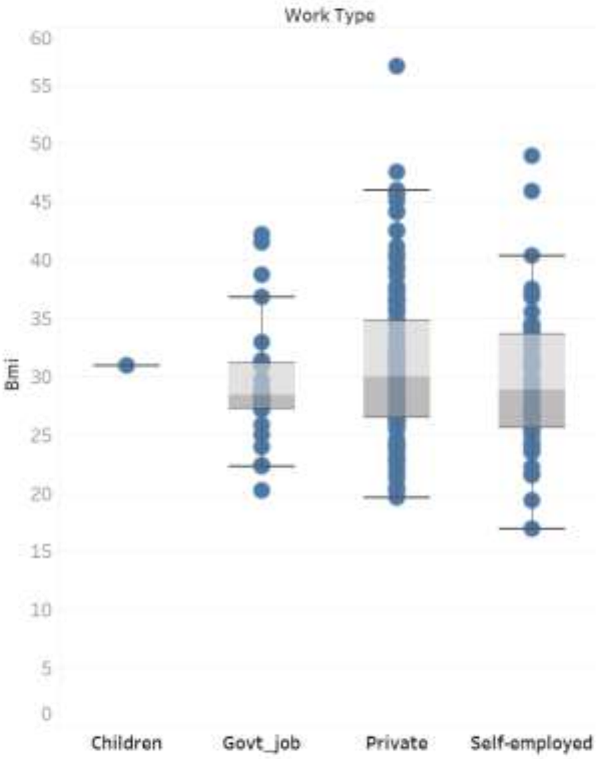


Figure 12. Relation between Work Type and BMI

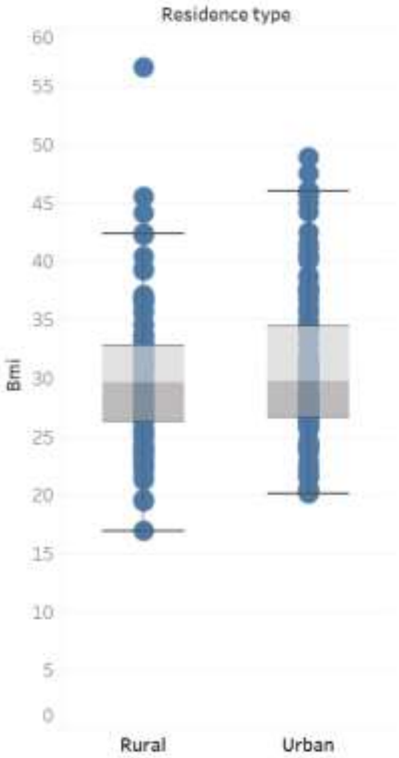


Figure 13. Relation between Residence Type and BMI

## 4. Data visualization design

When creating visualizations, we followed several design principles and best practices to ensure that the visualizations effectively conveyed the key findings and insights:

- Choosing appropriate chart types: we chose chart types that best suited the data and the message we wanted to convey. For example, scatter plots were used to show the correlation between age and average glucose level or BMI as both variables on the x-axis and y-axis are numerical. On the other hand, bar charts were used to compare the percentage of stroke patients by sex or by places of residence for its inherent simplicity in making comparisons. Particularly, bar charts' preattentive attribute is length, which also aligns with our intuition to draw the focus on the end points when comparing two or more variables.
- Emphasizing key findings: key findings were emphasized by using color to draw the viewer's attention. For example, we used two colors of dark blue and light blue to distinguish the stroke status of participants, with dark blue being used for those

suffering from stroke  to draw attention and emphasize the impact of higher risk of stroke for patients.

- Using appropriate color schemes: First, we used color schemes that are easy on the eyes and clearly distinguishable, such as light blue for stroke-free and dark blue for stroke occurrence.We avoided using too many colors or using bright colors that could be distracting. Second, we use color as an evocation of emotions. The blue color is typically used in the medical field as it evokes the feelings of calm, professionalism, trust, power and credibility, all of which are appreciated in the healthcare community.

For a better interpretation of the data, the following types of visualization are employed based on the characteristic of variables and variable combinations:

- Pie chart (figure 2, figure 7, figure 8, figure 9)
- Bar chart (figure 3, figure 4, figure 5, figure 6, figure 11)
- Box plot (figure 12, figure 13)
- Scatterplot (figure 10)

First of all, we used pie charts in figure 1, figure 6, figure 7 and figure 8 to compare two categorical variables. It can be helpful for comparing the prevalence of hypertension, heart disease, and smoking habits among stroke patients. Besides, we used different colors to distinguish between the two variables in the pie chart. In figure 3, figure 4 and figure 5, we used bar charts to compare two categories within a single group, such as male and female. The bar chart was also used in figure 10 to compare stroke patients in different age groups. This can be useful for analyzing the percentage of stroke patients based on demographic factors such as place of residence, gender, and marital status.

We used box plots in Figure 11 and Figure 12 to display the distribution of a continuous variable by a categorical variable. Specifically, we used box plots to identify any differences in BMI based on work type and residence type. In Figure 9, the use of scatter plot for the relation between BMI and glucose level by age is appropriate for visualizing the correlation between two variables. We also used a regression line to show the trend of these points on a scatter plot, which aids in deriving accurate insights and patterns towards the readers.

In terms of interactive features, we used the hover-over tooltips, which can allow users to explore and interact with the data in more detail. For example, hovering over a data point in a scatter plot can display additional information about that point, such as its value and label.

## 5. Results

The pie charts showed that hypertension, heart disease, and smoking habits are all significantly less common in stroke patients than in non-stroke patients. Even though the findings are contrary to our natural assumptions, they align with some previous studies from Pan et al. (2019), which indicates that stroke occurrence has a dose-dependent correlation to smoking (meaning the effect is trivial with smaller doses) and there is no association between the incidents of stroke and former smokers. The bar charts revealed that the majority of stroke patients live in urban areas, are female, and are married. This is because life expectancy of women is generally longer than men, hence compared to men, more women have strokes over their lifetimes. Women also have unique risk factors for stroke, such as having high blood pressure during pregnancy and hormonal fluctuations. Besides, married patients experienced strokes compared to those who were never married, which can be explained by participants with marital transitions and living with children who had increased stroke risk. This information may be useful for public health interventions targeting stroke prevention in different demographic groups.

The box plots showed that individuals who work in the private sector or are self-employed tend to have higher BMI values than those who work in the government or children. Self-employed individuals may have irregular working hours and may not have a fixed schedule for meals and exercise, leading to weight gain. Additionally, individuals living in urban areas tend to have slightly higher BMI values than those in rural areas. This could be due to the availability of highly processed and high-calorie food options in urban areas and a more sedentary lifestyle. These findings suggest that occupation and place of residence may be important factors to consider in stroke prevention strategies.

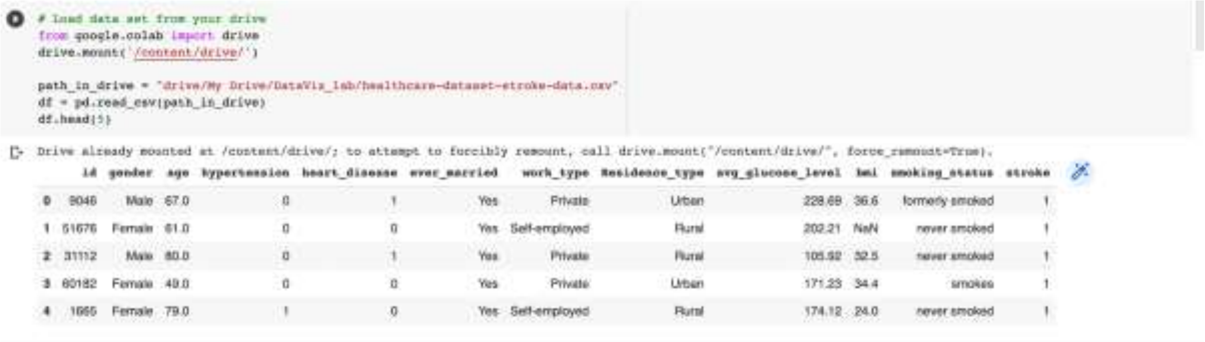
The scatter plot revealed a positive correlation between BMI and glucose level, particularly in older age groups. This suggests that maintaining a healthy weight and managing glucose levels may be particularly important for older individuals at risk for stroke.

## 6. Discussion and conclusion

The findings of this project are significant in highlighting the important demographic and lifestyle factors associated with stroke risk. The analysis shows that hypertension, heart disease, and smoking habits are little related to stroke risk. The little correlation implies that these risk factors are modifiable (for instance, former smokers have no correlation to stroke occurrence, meaning that quitting cigarettes can positively reduce stroke risk). Thus individuals with these conditions can reduce their risk of developing stroke by making positive changes in their lifestyle and habits. Additionally, the project reveals that the majority of stroke patients live in urban areas, are female, and are married, and older people are more likely to suffer from stroke, emphasizing the importance of taking a holistic approach to stroke prevention that considers all relevant demographic factors. Moreover, the findings show a significant association between work type and BMI, and the relation between residence type and BMI shows that patients living in urban areas had a higher median BMI compared to those living in rural areas. Furthermore, there is a positive correlation between BMI and glucose level. suggesting that higher BMI is associated with higher glucose levels in the blood, which is a risk factor for stroke. These insights can have significant implications for public health policies and initiatives aimed at reducing the burden of stroke on individuals and society. By identifying the most important risk factors for stroke, policymakers and healthcare providers can better target their efforts towards the populations most at risk, providing targeted education, screening, and prevention strategies.

However, there are several limitations to this study that should be addressed. Firstly, the dataset used in this study only includes information from a single source, which may affect the accuracy and representativeness of the data. Another thing is that the study was limited to exploring the relationship between demographic, lifestyle, and health-related factors and stroke risk. Other factors, such as genetic predisposition or environmental exposures, were not explored in this study but could potentially impact stroke risk. Based on these limitations and findings of the study, suggestions are to collect a more diverse and representative sample of participants to avoid potential bias in the data. Besides, consider incorporating more variables into the analysis, such as exercise habits or family medical history, to gain a more comprehensive understanding of the risk factors for stroke.

## Appendix

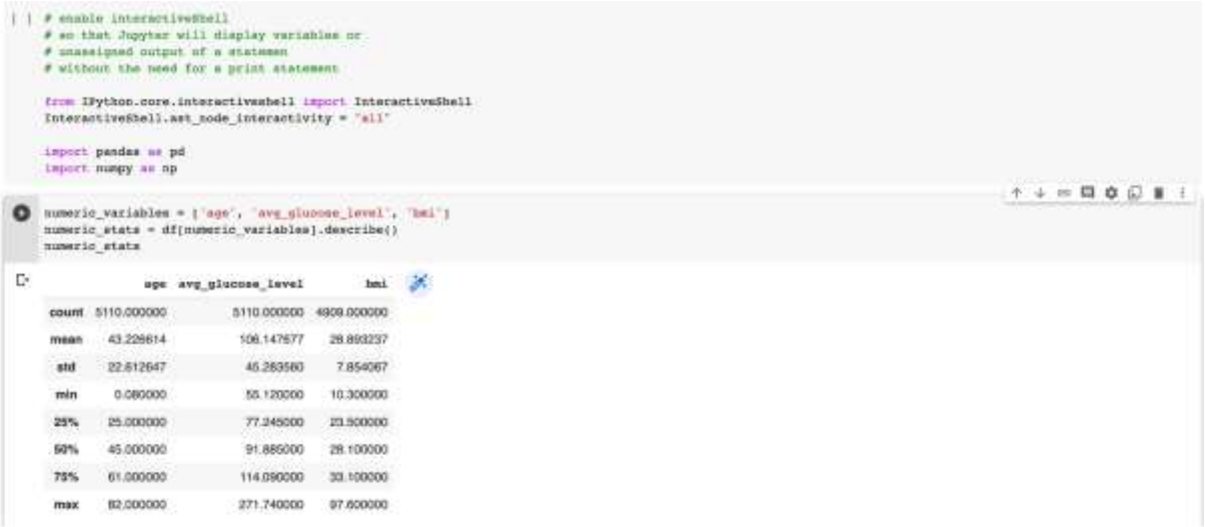


```python
# load data set from your drive
from google.colab import drive
drive.mount('/content/drive/')

path_in_drive = "drive/My Drive/DataViz_lab/healthcare-dataset-stroke-data.csv"
df = pd.read_csv(path_in_drive)
df.head(5)
```

```
Drive already mounted at /content/drive/; to attempt to forcibly remount, call drive.mount("/content/drive/", force_remount=True).
      id  gender   age  hypertension  heart_disease  ever_married      work_type  Residence_type  avg_glucose_level   bmi   smoking_status  stroke
0   9046    Male  67.0             0              1           Yes        Private           Urban             228.69  36.6  formerly smoked       1
1  51676  Female  61.0             0              0           Yes  Self-employed           Rural             202.21   NaN     never smoked       1
2  31112    Male  80.0             0              1           Yes        Private           Rural             105.92  32.5     never smoked       1
3  60182  Female  49.0             0              0           Yes        Private           Urban             171.23  34.4           smokes       1
4   1665  Female  79.0             1              0           Yes  Self-employed           Rural             174.12  24.0     never smoked       1
```

Appendix 1: Loading the dataset from the drive and read csv file



```python
# enable interactiveShell
# so that Jupytar will display variables or
# unassigned output of a statemen
# without the need for a print statement

from IPython.core.interactiveshell import InteractiveShell
InteractiveShell.ast_node_interactivity = "all"

import pandas as pd
import numpy as np
```

```python
numeric_variables = ['age', 'avg_glucose_level', 'bmi']
numeric_stats = df[numeric_variables].describe()
numeric_stats
```

```
               age  avg_glucose_level          bmi
count  5110.000000        5110.000000  4909.000000
mean     43.226614         106.147677    28.893237
std      22.612647          45.283560     7.854067
min       0.080000          55.120000    10.300000
25%      25.000000          77.245000    23.500000
50%      45.000000          91.885000    28.100000
75%      61.000000         114.090000    33.100000
max      82.000000         271.740000    97.600000
```

Appendix 2. Creating summary statistics in pandas library

## References

*Assessing Your Weight*. (2022, December 8). Centers for Disease Control and Prevention. https://www.cdc.gov/healthyweight/assessing/index.html

*Diabetes Testing*. (2023, February 28). Centers for Disease Control and Prevention. https://www.cdc.gov/diabetes/basics/getting-tested.html#:~:text=A%20fasting%20blood%20sugar%20level,higher%20indicates%20you%20have%20diabetes.

Pan, B., Jin, X., Jun, L., Qiu, S., Zheng, Q., & Pan, M. (2019b). The relationship between smoking and stroke. *Medicine*, *98*(12), e14872. https://doi.org/10.1097/md.0000000000014872

Rexrode, K. M., Madsen, T. E., Yu, A. Y., Carcel, C., Lichtman, J. H., & Miller, E. C. (2022). The Impact of Sex and Gender on Stroke. *Circulation Research*, *130*(4), 512–528. https://doi.org/10.1161/circresaha.121.319915

*WHO | Stroke, Cerebrovascular accident | Health topics*. (n.d.). World Health Organization - Regional Office for the Eastern Mediterranean. https://www.emro.who.int/health-topics/stroke-cerebrovascular-accident/index.html

*Women and Stroke | cdc.gov*. (2023, May 4). Centers for Disease Control and Prevention. https://www.cdc.gov/stroke/women.htm#:~:text=Because%20women%20generally%20live%20longer,especially%20if%20they%20also%20smoke.