Stroke Dataset Analysis

Helly Thakar

June 10, 2021

Student ID Number: 19155625

Total word count : **2365 words**

Data Visualisation: CMP5352

Birmingham City University
Faculty of Computing Engineering and the Built Environment
School of Computing and Digital Technology

Contents

1 Abstract **3**

2 Section 1: Introduction **4**

3 Section 2: Motivation and Objectives **5**

4 Section 3: Experimental Results **13**

5 Section 4: Summary **19**

6 Reference **20**

1 Abstract

With the use of data visualisation, this report provides information on stroke. The R programming language is used to do data analysis and build web applications in RStudio. The major goal of this report is to enable a random individual to become aware of a stroke. Identified some intriguing non-trivial dataset problems. To solve the stated research questions, I used data visualisation ideas. The web application and data analysis components are designed to operate with patient data. When constructing data visualisations using ggplot2 and the shiny package, categorical, binary numerical, and continuous numerical variables from the dataset are used. The key findings of the report show that age is the most important predictor of having a stroke, followed by hypertension, heart disease, and average glucose level.

2 Section 1: Introduction

The central and peripheral nervous systems are both affected by neurological illnesses. Some of these illnesses are treatable, while others are not. Neurological issues, such as Alzheimer's disease and Parkinson's disease, mostly affect people over the age of 60, making ageing a crucial role in the development of these diseases. Genetic diseases, infections, and lifestyle choices are among the reasons, as are other health issues that may damage the brain. There are over 600 nervous system illnesses, including stroke, brain tumours, epilepsy, and many others. Every year, around 15 million people suffer from a stroke.

To better understand data by mapping from data space to visual space, I'm using data visualisation concepts like data type that can be visualised, chart type to show the visualisation, and graphics grammar, which is for better data communication using a set of rules, to create a better visualisation to gain insight into the data. The data type is structural data and the chart type I used to display the comparison, distribution, and connection between the variables in the data in the bar chart.

With the use of data visualisation, the report attempts to understand which variables influence stroke and how many individuals suffer a stroke vs how many don't, so we can raise awareness about it.

The report's accomplishments include determining the number of persons who have had a stroke, the gender and age groups that have had more strokes, and characteristics such as work type, hypertension, heart disease, habitation type, marital status, and others that influence the likelihood of having a stroke. And I produced visualisations using the ggplot2 package in R studio, as well as a web app using the shiny package, which explains the entire tale using data visualisation simply and interactively.

The report is divided into four sections: Introduction, Motivation, and Objectives, which includes a description of the chosen data set and a detailed description of the potential interesting non-trivial questions; Experimental Results, which includes detailed results in high-quality graphs which were created using the R programming language in R studio to answer those nontrivial questions, and lastly, a summary of the selected questions, conclusions, and major results in this report.

3 Section 2: Motivation and Objectives

Stroke is the second biggest cause of mortality worldwide, accounting for over 11% of all fatalities, according to the World Health Organization (WHO).

The Stroke Dataset was taken from the Kaggle website and has 5000 records with 12 variables. However, I used 1000 records and 11 variables in this analysis. I eliminated the column id because it was ineffective, and I didn't include the entries that had the smoking status variable set to 'Unknown.' The Unknown category denoting that we have no way of knowing whether or not an individual smoked. So I eliminated the unknown.

Based on input criteria such as gender, age, numerous illnesses, and smoking status, this data set is used to forecast if a patient is likely to have a stroke. Each row of data in the table contains pertinent information about the patient. There are both categorical and numerical variables.

The data-set consists of the following:
1) Categorical Features: gender, ever_married, work_type, Residence_type, smoking_status
2) Binary Numerical Features: hypertension, heart_disease, stroke
3) Continuous Numerical Features: age, avg_glucose_level, bmi

The following are the information of attributes:
2) gender: "Male", "Female"
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_job", "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"
12) stroke: 1 if the patient had a stroke or 0 if not

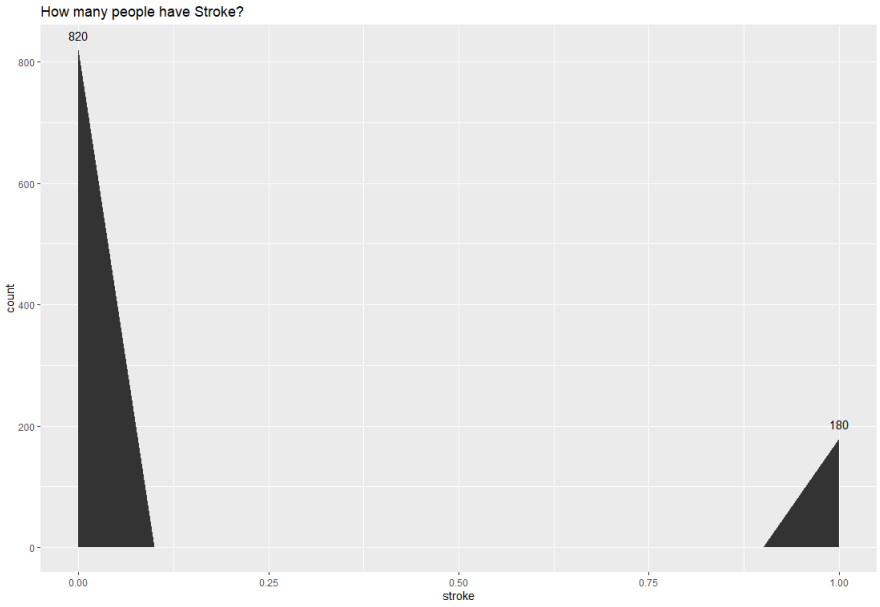

Figure 1: Count of people with stroke and no stroke in data set

Moreover, 800 people have had a stroke in our stroke data collection, but less than 200 people haven't, as seen in the graph above. I used the stroke variable to produce this visualisation in R. According to the graph above, 180 persons out of 1000 experience a stroke and 820 do not.

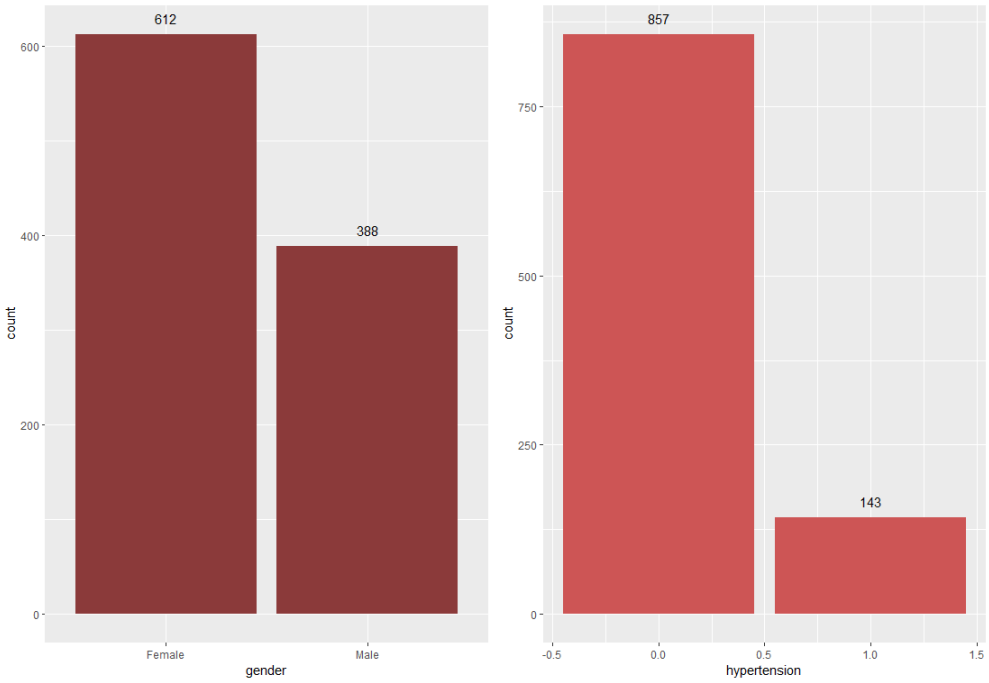

Figure 2: Categorical Variables Analysis
In our data, females outnumber males. There are 612 ladies and 388 men out of a total of 1000. People who do not have hypertension are far fewer than those who do. In our dataset, 857 persons do not have hypertension and 143 have hypertension out of 1000.

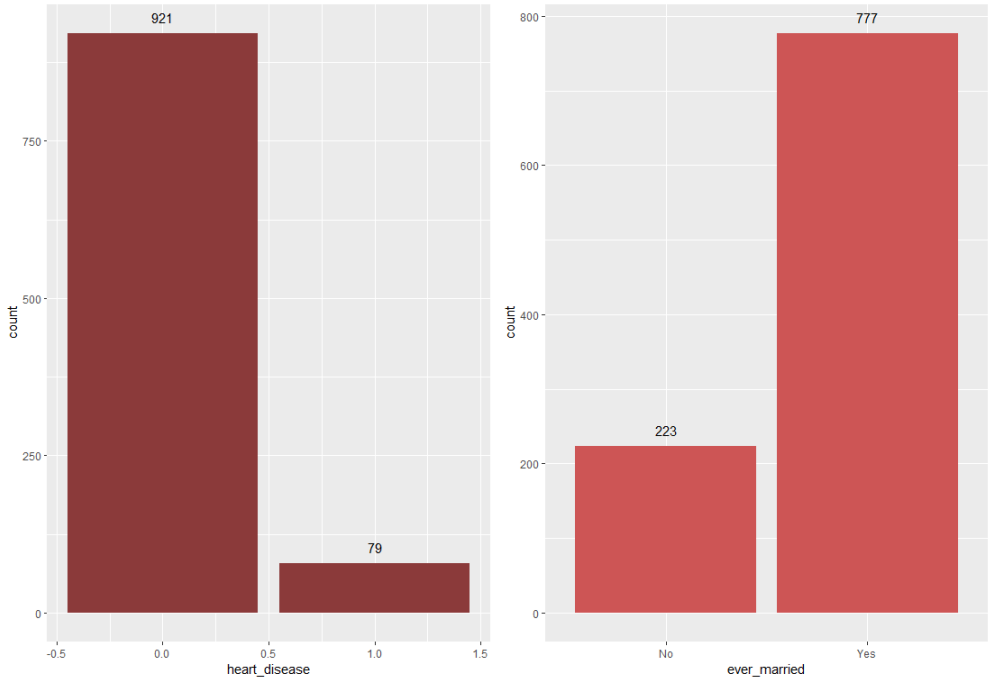

Figure 3: Categorical Variables Analysis
Heart disease affects a small percentage of the population. 921 persons out of 1000 have no heart disease and 79 have heart disease. The number of married persons much outnumbers the number of single persons (makes sense as the distribution is between 0 and 60). 223 persons out of 1000 are unmarried, while 777 are married.

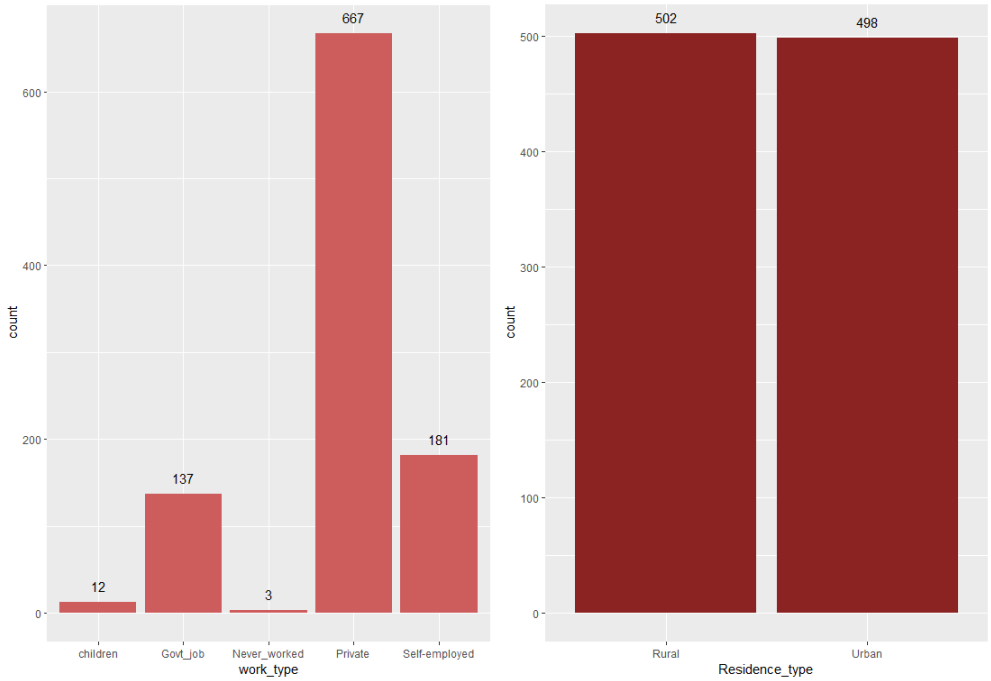

Figure 4: Categorical Variables Analysis
People seem to prefer working for private enterprises, despite the fact that the number of self-employed/government employees and children appears to be equal. The number of unemployed persons is quite low. Out of 1000 persons, 12 have children, 137 work for the government, 3 have never worked, 667 work in the private sector, and 181 are self-employed. There is not much of a distinction between the urban and rural populations. Out of 1000 people, 502 live in rural areas and 498 live in urban areas.

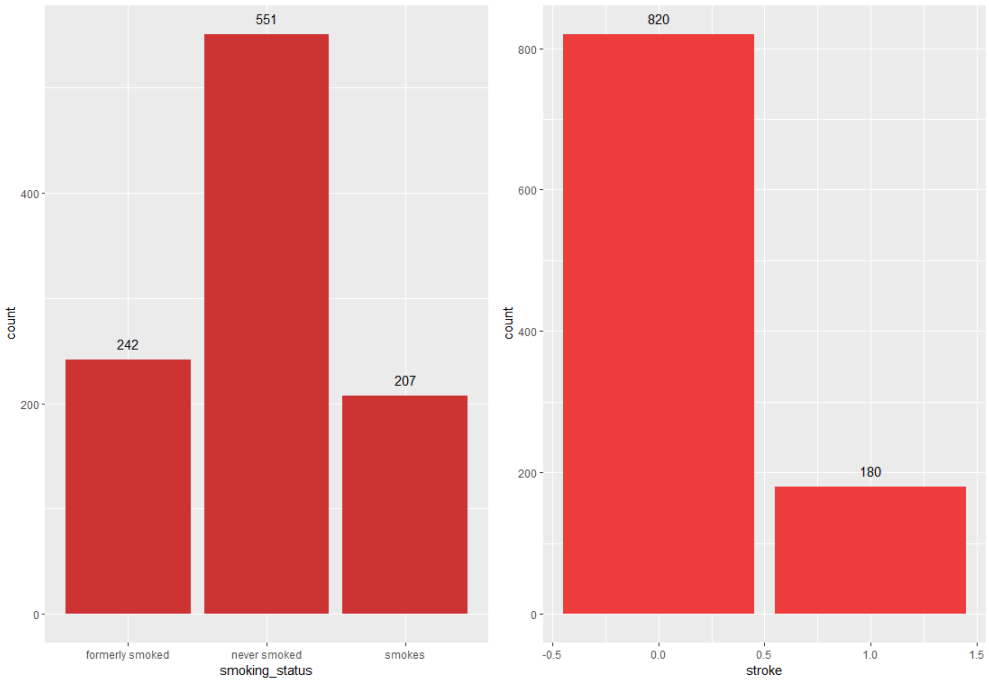

Figure 5: Categorical Variables Analysis
Nonsmokers outnumber smokers/ex-smokers by a large margin, which is a positive thing. There were 1000 records in which 242 former smokers, 551 never smokers, and 207 smokes. The number of persons who have had a stroke is 820 out of 1000.

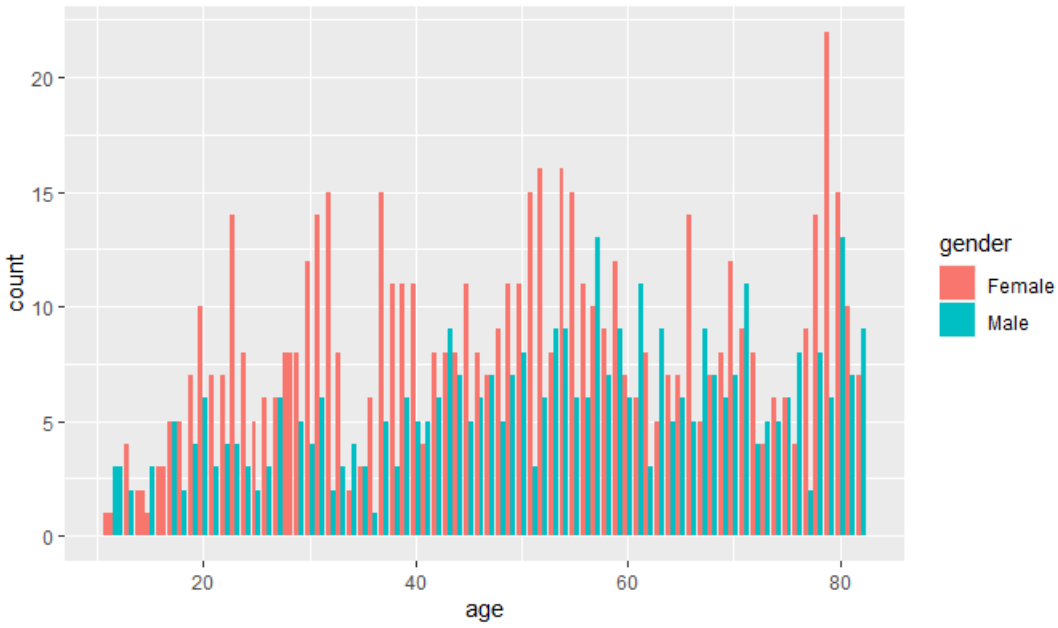

Figure 6: Count of age by gender
Females outnumber males in our data set, and the majority of persons are between the ages of 40 and 80, as seen in the graph above. As a result, elderly adults have a higher risk of stroke than younger individuals.

"A recent study led by researchers from the University of California- Los Angeles Health Sciences (UCLA) determined the running fluid spiked with a Covid-19-like protein through a 3D-printed model to explain how the virus increases the risk of stroke," I chose this data set after reading recent news. As a result, I had to first take its data set and extract more information on stroke.

There are four non-trivial questions.

The first non-trivial question is whether age, BMI, and glucose level have an impact on the stroke variable. For starters, because the frequency of cardiovascular and metabolic diseases increases with age, older persons are more prone to have strokes. The most major risk factor for stroke is age. Second, being overweight increases our chances of getting a stroke significantly. Inflammation is likely to arise as a result of excess fat in the body, resulting in poor blood flow and possible blockages—two key causes of stroke. Finally, high blood sugar levels can harm blood vessels and neurons, increasing the risk of a stroke. When the blood supply to the brain is cut off, a stroke develops. As a result of the above, I'd want to observe how all of the elements listed in the question affect stroke, which we'll observe through data visualisation.

The second non-trivial question that may be of interest is: Do work type, marital status, dwelling type, smoking status, and gender all have an impact on stroke? What impact does that have? To begin, work-related mental stress may raise the risk of stroke, therefore it's important to determine which types of jobs, such as private, government, never worked, and self-employed, are affected and by how much. Second, I want to see who has more stroke, married or unmarried, by examining marital status. Finally, I'd like to examine if it makes a difference where individuals reside when they have a stroke. Then you'll want to see if the persons with stroke smoke, and you'll also want to see if males or females are more likely to have a stroke.

The third non-trivial question is perhaps the most intriguing: how does the stroke distribution differ by gender? This question will explain the gender distribution of women and men who have had a stroke and those who have not had a stroke. The fourth non-trivial question, How do heart disease and hypertension affect the stroke variable, has the potential to be intriguing? I had heard that coronary artery disease, which is connected to the heart, increases our risk of stroke because plaque builds up in the arteries and restricts the transport of oxygen-rich blood to the brain.

The above mentioned four questions are non-trivial because we are searching for precise and more explainable answers (solutions) to that questions (problems).

4 Section 3: Experimental Results

The following are the results (answers) of each identified non-trivial questions with the high-quality graphs:

Question 1: Do factors such as age, BMI, glucose level affect the stroke variable?

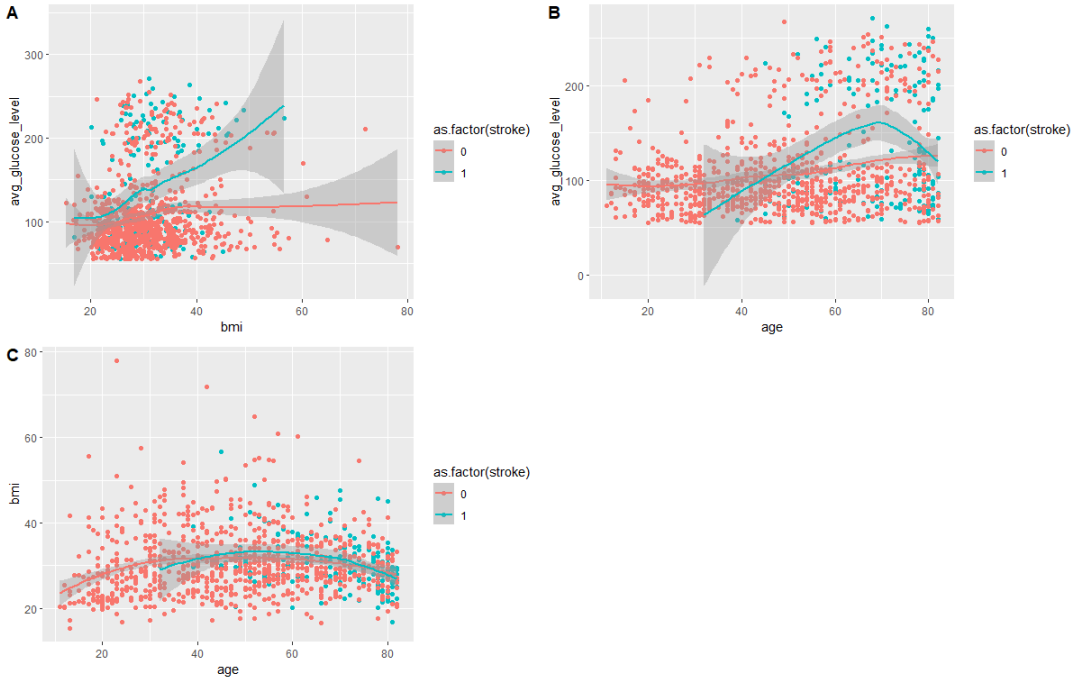

Figure 7: Analysis of numerical variables

The data appears to be highly imbalanced as only a few points where stroke = 1. In A graph i.e. average glucose level vs BMI, it seems that most of the people having a stroke have glucose level between 50 to 270 and BMI between 20 to 55. In the B graph i.e. average glucose level vs age, it seems that people with age between 35 to 85 and having average glucose level between 40 to 260 are more prone to having Stroke. In the C graph i.e BMI VS age it seems that people with mage more than 40 and BMI between 20-60 are more prone to Stroke. We can see from graph B and C that most people in the 40 to 90 age group are having a stroke.

Question 2: Work type, married status, residence type, smoking status, gender does these factors also affect the stroke?

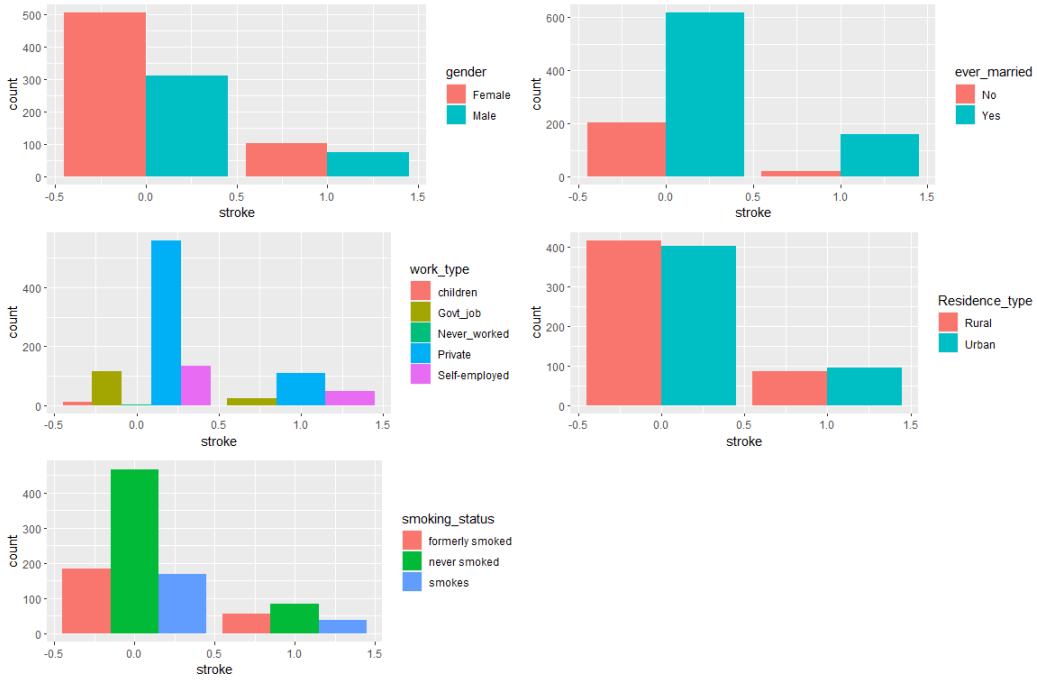

Figure 8: Analysing Categorical Variables with Stroke

Seems that number of male and female who has stroke are equal in number. The people who got married show signs of stroke way more than people who are unmarried (expected I guess). Private employees seem to experience stroke more than other work types(may be due to work pressure). Self-employed people do show signs of stroke(maybe due to reasons like heart disease, tension etc). Children can be ignored. Almost no difference between people living in urban and rural areas in terms of stroke occurrence. People who formerly smoked and who smoke (combined) are showing signs of stroke way more than people who never smoked (considering the sample size of people who never smoked and people who used to smoke and smoke now).

Question 3: How is the stroke distribution compared to gender?

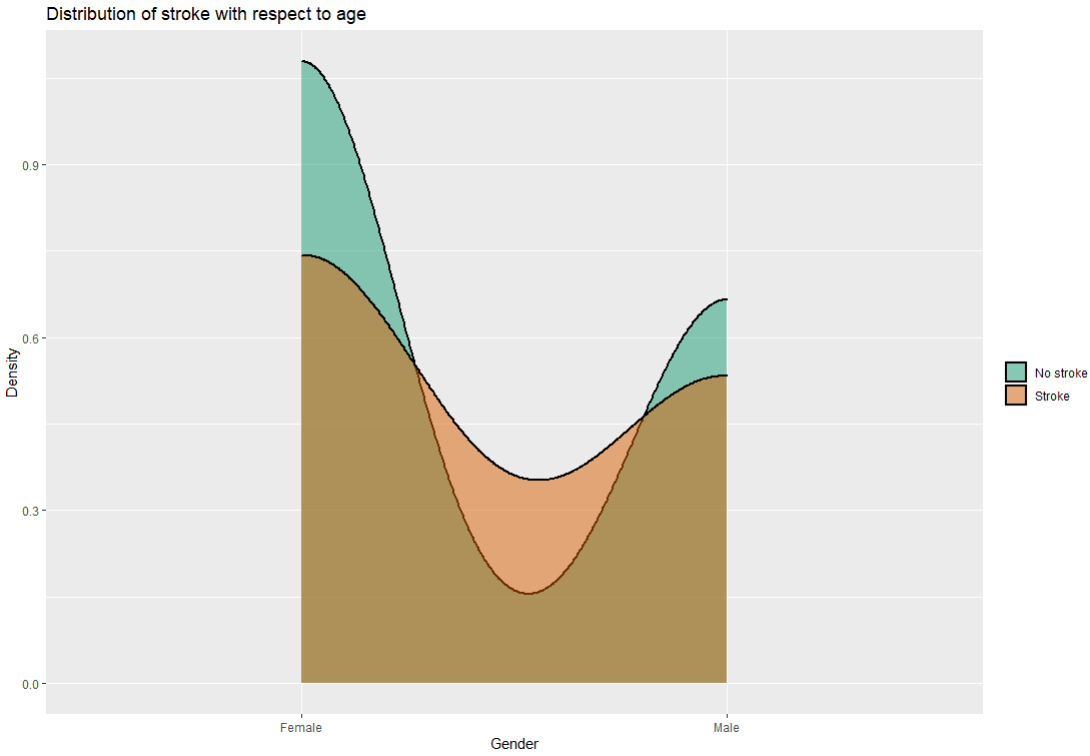

Figure 9: Distribution of gender with respect to stroke
Female and male are more with no stroke but on the other hand, female and male seems equal to female with stroke but the curve is slightly bending higher towards female so female have more stroke them, male.

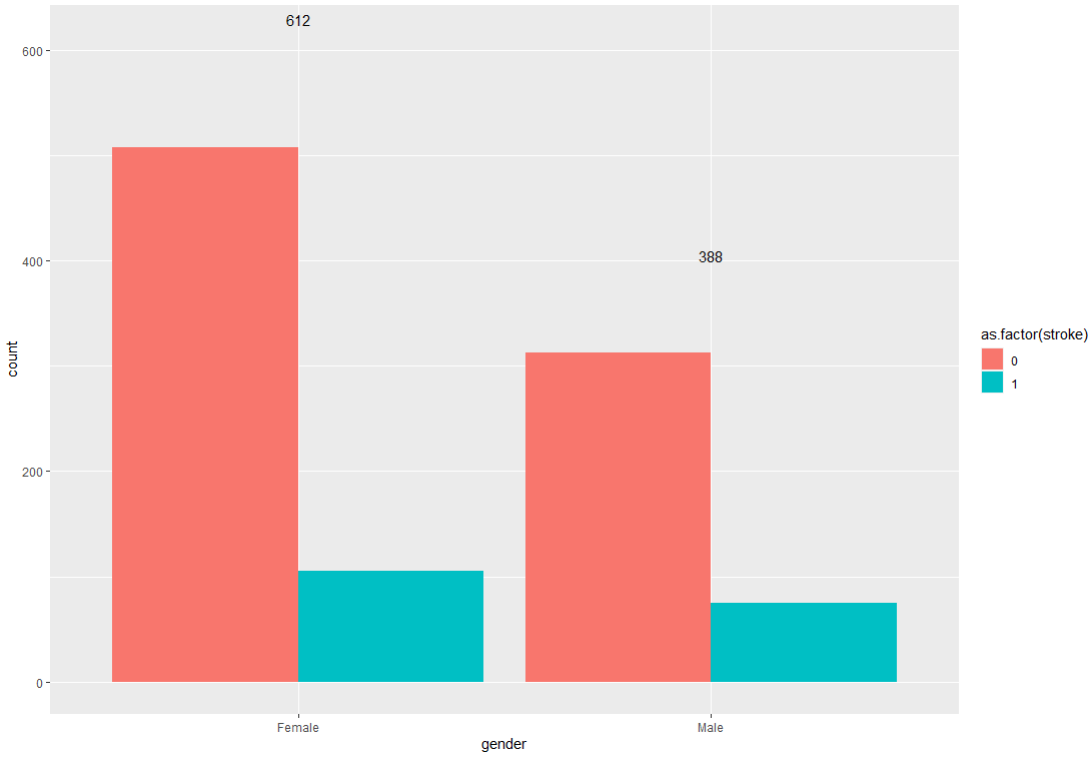

Figure 10: Distribution of gender with respect to stroke
The above graph explains that 612 are females and 388 are males. Out of 613 females, more than 500 females doesn't have a stroke and more than 100 females have a stroke as well as out of 388 males, more than 300 males don't have a stroke and less than 100 males have a stroke. More people don't have a stroke in our data set and female are more in our data set compared to male as well as males are less having a stroke than female.

Question 4: Does Heart disease and hypertension affect the stroke variable?

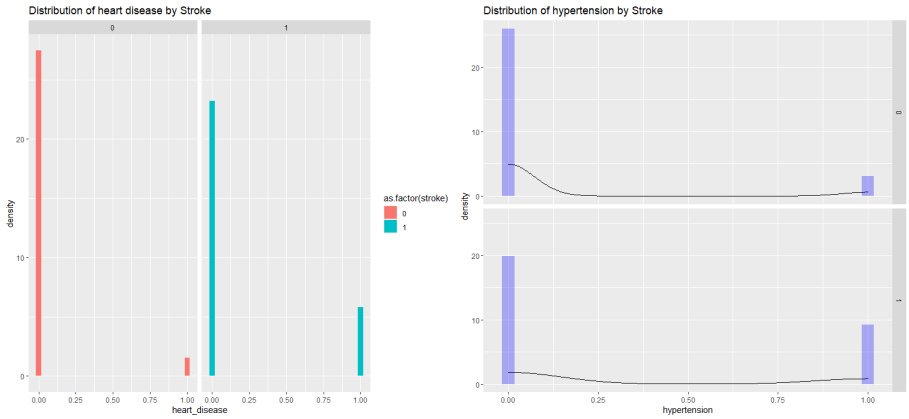

Figure 11: Distribution of heart disease and hypertension with respect stroke

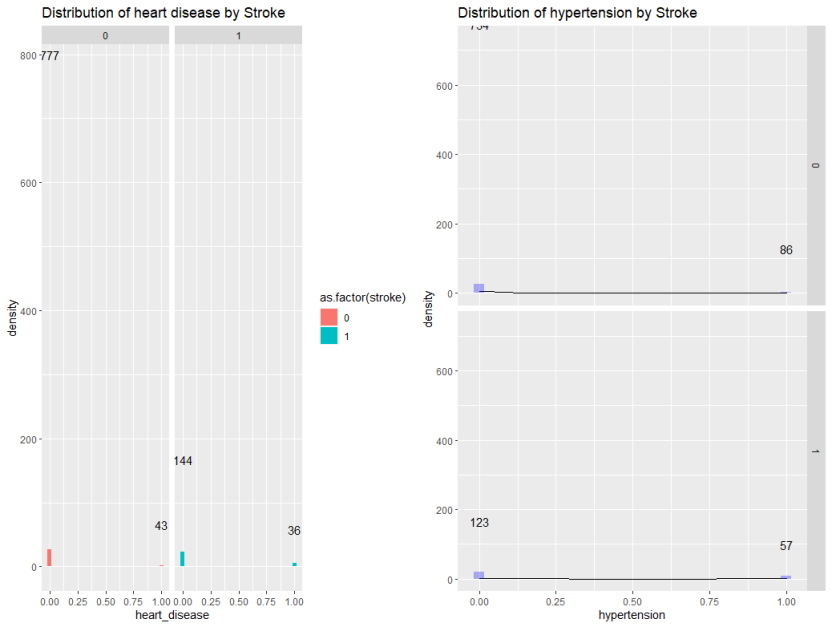

Figure 12: Distribution of heart disease and hypertension with respect stroke People with heart disease show signs of stroke. Out of 180 people with stroke, 79 people have heart disease which shows that heart disease is an important factor in people having a stroke. The number of people who do not have hypertension also shows signs of no stroke. And people with hypertension show signs of stroke. Out of 180 people with stroke, 143 people have hypertension which shows that hypertension is an important factor in people having a stroke. So we see both the values of people having hypertension and heart disease which is 143 and 79 out of 180 people having a stroke then it seems that some people have both hypertension and heart disease because we add 79 with 180 then it gives 222 but here only 180 people have a stroke so it proves that some people having a stroke also have both hypertension and heart disease. Therefore, hypertension and heart disease will be considerable factors while seeing the stroke.

5 Section 4: Summary

The inclusion of data analysis through developing visualisations utilising data and using a web application is effectively done with this study, according to the report's conclusion. The report and web application demonstrate how the findings of data analysis may be made available to users. This form of integration aids the user in comprehending plots and statistical work in data analysis via data visualisation.

The report's primary results include that the most important predictor of having a stroke is age, followed by hypertension, heart disease, and average glucose level. It gives healthcare practitioners the knowledge they need to focus on the causes of stroke and then give appropriate preventative medical recommendations. Females have more strokes than men, individuals who live in cities have more strokes than those who live in rural areas, and individuals who work in the private sector have the most strokes. The most critical factors that influence a stroke are age, hypertension, and heart disease. As age is the most important factor affecting stroke so the web application depicts that so user can easily see by comparing the age variable with other variables in the data set.

6 Reference

1] Ashokan, S., Narayanan, S. and Anand, P., 2021. [online] Irjet.net. Available at: ¡https://www.irjet.net/archives/V7/i3/IRJET-V7I3799.pdf¿ [Accessed 3 June 2021].

2]www.ETHealthworld.com. 2021. Diagnostics, Latest Diagnostics News, Health News - ET HealthWorld. [ONLINE] Available at: https://health.economictimes.indiatimes.com /news/diagnostics/ [Accessed 20 May 2021].

3]March — 2015 — Medication Junction. 2021. March — 2015 — Medication Junction. [ONLINE] Available at: https://www.medicationjunction.com/2015/03/[Accessed 20 May 2021].

4]PubMed. 2021. Stroke prevention and management in older adults - PubMed. [ONLINE] Available at: https://pubmed.ncbi.nlm.nih.gov/16966926/ [Accessed 20 May 2021].

5]PubMed. 2021. Stroke prevention and management in older adults - PubMed. [ONLINE] Available at: https://pubmed.ncbi.nlm.nih.gov/16966926/ [Accessed 20 May 2021].

6]Howley, Elaine K. "A Patient's Guide to Coronary Artery Disease." US News & World Report, U.S. News & World Report, 2019, [ONLINE] Available at: health.usnews.com/conditions/heart-disease/coronary-artery-disease[Accessed 20 May 2021].

7]"Stroke Prediction Dataset." Kaggle.com,[ONLINE] Available at: www.kaggle.com/fedesoriano/stroke-prediction-dataset[Accessed 25 March 2021].