KOLEJ PROFESIONAL MARA BERANANG

FINALE PROJECT : DATA ANALYSIS ON STROKE PREDICTION DATASET

SESSION 3 2022/2023

PREPARED FOR:

PUAN MAWAR BIDURI BINTI ABDUL HALIK

PREPARED BY:

AISYAH AINA SUFIA BINTI HILMAN

MENTOR'S NAME:

PUAN NOOR FAZLIANI BINTI SHAMSHUDIN

PROJECT SCENARIO

The world is becoming more and more data-driven, with endless amounts of data available to work with. Data analysis is widely used in almost every aspect of our life, and it has been used by small businesses, retail companies, in medicine, and even in the world of sports. Due to the importance and valuable data nowadays, data analysis has become an important field. Assume that you are currently taking the statistic course in your college, you are required to complete a data analysis project for your statistics class. Your task is to use R programming to analyze a dataset of your choice and produce a report based on your findings.

Data Visualization and Presentation

- Summarize the findings: You are required to summarize your key findings from the analysis and highlight the insights that have been discovered. You need to present these findings in a clear and concise manner, making sure to keep your audience in mind.
- Use suitable types of presentation: You need to use a combination of graphs, tables, and text to present their findings. They choose presentation types that best highlight your insights and are easy for your audience to understand.

0.0) TABLE OF CONTENT **Pages**

1.0) Introduction

The assignment of conducting a Data Analysis on Stroke Prediction Dataset involves exploring and analyzing the various factors that may contribute to the likelihood of a person experiencing a stroke. The dataset typically includes information on the individual's age, id, gender, smoking status, hypertension, heart disease, marital status , work type , residence type , average glucose level , bmi , stroke status and among other factors. Through data analysis, researchers can identify correlations and patterns in the data that may be useful in predicting the risk of stroke. This analysis may involve Exploratory Data and Analysis (EDA) using R and RStudio to gain deeper understanding about each data and their relationship. Additionally, data visualization techniques such as charts, graphs, scatter plot and heat maps may be used to help illustrate the patterns and trends in the data. Overall, data analysis on the Stroke prediction dataset is a valuable tool in understanding the risk factors associated with stroke and can help inform preventive measures and treatment strategies.

2.0) Analysis Background

2.1) Purpose

The primary goal of this research is to carry out an in-depth analysis and correctly identify the proper variable that has the greatest impact and influence on patients who is likely to get stroke based on the input parameters given in Stroke Prediction Dataset.

2.2) Objectives

1. To determine which conditions, lifestyle or habit such as smoking, heart disease, and hypertension impact on stroke risk.
2. To determine how does level consumption of glucose impact on stroke risk .

2.3) Target Audience

There are some important target audience in order to understand and help about stroke in deeper.

1. **Healthcare professionals**

   A primary target audience for data analysis on the Stroke Prediction Dataset includes healthcare professionals including doctors, nurses, and public health authorities. They can improve their understanding of the risk factors for stroke and create better preventative and treatment plans using the study' findings from the analysis.

2. **Stroke patients and their families**

   Another significant target group for data analysis on the stroke prediction dataset is stroke patients and their family or dependents. They can choose more wisely about lifestyle changes and medical measures to lower their risk of future strokes by being aware of the risk factors linked to stroke.

3. **Researchers and academics**

   The intended audience for data analysis on the stroke prediction dataset also includes academics and researchers in the domains of epidemiology, public health, and data science. Students can create new research questions and hypotheses using the analysis' findings to deepen their grasp of the intricate interactions between various risk variables.

3.0) Suitable Dataset

3.1) Dataset

- Dataset Title : **Stoke Prediction**
- Dataset from Kaggle Website : https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

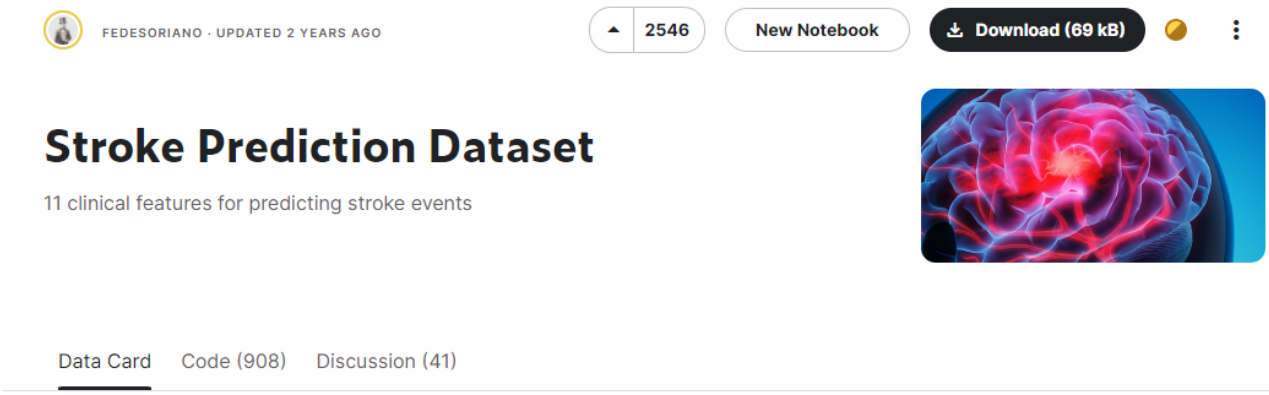


Figure 0.0 shows about the chosen dataset which is Stroke Prediction Dataset on Kaggle for further analysis

Based on **figure 0.0** ,this dataset is used to predict whether a patient is likely to get stroke based on the input parameters like gender, age, various diseases, and smoking status. Each row in the data provides relavant information about the patient.

Attribute Information

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

3.2) Structure of Dataset

```
> str(strokedata)
'data.frame':   5110 obs. of  12 variables:
 $ id                : int  9046 51676 31112 60182 1665 56669 53882 10434 27419 60491 ...
 $ gender            : chr  "Male" "Female" "Male" "Female" ...
 $ age               : num  67 61 80 49 79 81 74 69 59 78 ...
 $ hypertension      : int  0 0 0 0 1 0 1 0 0 0 ...
 $ heart_disease     : int  1 0 1 0 0 0 1 0 0 0 ...
 $ ever_married      : chr  "Yes" "Yes" "Yes" "Yes" ...
 $ work_type         : chr  "Private" "Self-employed" "Private" "Private" ...
 $ Residence_type    : chr  "Urban" "Rural" "Rural" "Urban" ...
 $ avg_glucose_level: num  229 202 106 171 174 ...
 $ bmi               : chr  "36.6" "N/A" "32.5" "34.4" ...
 $ smoking_status    : chr  "formerly smoked" "never smoked" "never smoked" "smokes" ...
 $ stroke            : int  1 1 1 1 1 1 1 1 1 1 ...
>
```

Figure 1.0 : The structure of Stroke Prediction Dataser

Descriptions:

Based on the structure of the , the suitable variable will be the variables smoking_status, heart_disease, hypertension, avg_glucose_level, and stroke are some of the most suitable variables to use for analyzing the Stroke Prediction Dataset.

The smoking_status variable is important because smoking is a well-known risk factor for stroke. Analyzing the distribution of smoking status in the dataset can provide valuable insights into how smoking influences stroke risk and can be used to develop targeted prevention and treatment strategies for individuals who smoke.

The heart_disease variable is also an important variable to analyze because individuals with heart disease are at a higher risk of stroke. Examining the distribution of heart disease in the dataset can provide insights into the relationship between heart disease and stroke risk, and can be used to develop more accurate and reliable models for predicting stroke risk.

The hypertension variable is also an important variable to analyze because hypertension is a well-known risk factor for stroke. Analyzing the distribution of hypertension in the dataset can provide valuable insights into how hypertension influences stroke risk and can be used to develop targeted prevention and treatment strategies for patients with hypertension.

The avg_glucose_level variable is an important variable to analyze because high glucose levels are a risk factor for stroke, and diabetes is a common comorbidity in individuals who have suffered a stroke. Examining the distribution of avg_glucose_level in the dataset can provide insights into how glucose levels influence stroke risk .

Last but not least, the stroke variable is the target variable in the dataset and is therefore essential for any analysis or modeling related to stroke risk. Analyzing the distribution of stroke in the dataset can provide insights into the prevalence of stroke and the factors that contribute to stroke risk.

4.0) Exploratory Data Analysis (EDA)

4.1) Data Cleaning

Data cleaning is a critical step that should be performed before starting to analyze and explore the data process. It is important to do data cleaning before exploratory data analysis (EDA) for easier . Quantity of zeros, NA, and unique values may result in a good or terrible model for this project. Thus, in this project the unique values that going to be treat are N/A and zeros values.

However, in Stroke Prediction Dataset is only capable on cleaning the values in variable " avg_glucose_level" where there are N/A values were found . Thus, it is important to clean and treat the N/A values in order to produce an accurate analysis result. Meanwhile to treat zeros values, the dataset that will be used is heart_disease .Here an approach to cover the very first step in data modeling :

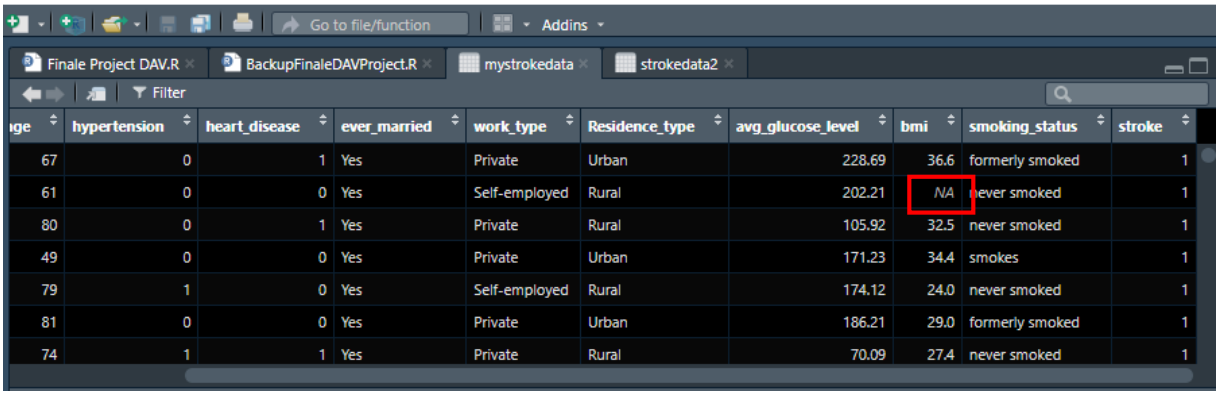
➢ Treat Missing Value (N/A) in Stroke Prediction Dataset

1. Step 1 : Checking NA(missing value), zeros, data type and unique values

When we receive a new dataset for analysis, we need to determine if there are any missing values (also referred to as "NA" in the R programming language) and identify the type of data. To assist with this, we can use the "df_status" function from the funModeling package, which displays these figures as both relative and percentage values.

df_status(mystrokedata)

```
> df_status(mystrokedata)
             variable q_zeros p_zeros q_na p_na q_inf p_inf      type unique
1                  id       0    0.00    0 0.00     0     0   integer   5110
2              gender       0    0.00    0 0.00     0     0 character      3
3                 age       0    0.00    0 0.00     0     0   numeric    104
4        hypertension    4612   90.25    0 0.00     0     0   integer      2
5       heart_disease    4834   94.60    0 0.00     0     0   integer      2
6        ever_married       0    0.00    0 0.00     0     0 character      2
7           work_type       0    0.00    0 0.00     0     0 character      5
8      Residence_type       0    0.00    0 0.00     0     0 character      2
9   avg_glucose_level       0    0.00    0 0.00     0     0   numeric   3979
10                bmi       0    0.00  201 3.93     0     0   numeric    418
11     smoking_status       0    0.00    0 0.00     0     0 character      4
12             stroke    4861   95.13    0 0.00     0     0   integer      2
>
```



ige	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
67	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
61	0	0	Yes	Self-employed	Rural	202.21	NA	never smoked	1
80	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
49	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
79	1	0	Yes	Self-employed	Rural	174.12	24.0	never smoked	1
81	0	0	Yes	Private	Urban	186.21	29.0	formerly smoked	1
74	1	1	Yes	Private	Rural	70.09	27.4	never smoked	1

Figure 2.0 : The status of Stroke Prediction Dataset

2. **Step 2 : Check the whole data frame for missing values**

   If a row has at least one incomplete column, it will display FALSE . Otherwise, it will display TRUE.

```
complete.cases(mystrokedata)
```

```
  [1]  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE
 [17]  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE FALSE  TRUE  TRUE
 [33]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE FALSE  TRUE
 [49]  TRUE  TRUE FALSE FALSE  TRUE  TRUE FALSE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
 [65] FALSE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE
 [81]  TRUE FALSE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
 [97]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
[113] FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE FALSE  TRUE
[129]  TRUE FALSE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
[145]  TRUE  TRUE FALSE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
[161] FALSE FALSE FALSE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE FALSE FALSE  TRUE  TRUE FALSE  TRUE
[177]  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE
[193]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
[209]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE
[225]  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
[241]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
[257]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
[273]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
[289]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
[305]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
[321]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
[337]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
[353]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
[369]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
[385]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
[401]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
[417]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
[433] FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
[449]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
[465]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE FALSE
[481]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
[497]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
[513]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE
[529]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
[545]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
```

Figure 3.0 shows about dataframe with missing values

3. **Step 3 : Sum all complete rows**

It is very important to sum up all complete rows in a dataset helps in identifying missing values, detecting duplicates, checking for outliers and helps in ensuring data consistency.

```
sum(complete.cases(mystrokedata))
```

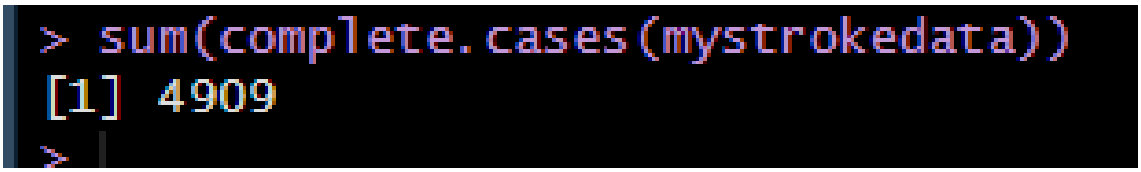


Figure 4.0 shows about sum of all complete rows to clean the data

4. **Step 4 : Create new data without missing values**

Therefore, creating new data without missing values is important to ensure that the resulting data is representative, unbiased, and suitable for use in analysis

```
strokedata2 <- mystrokedata[complete.cases(mystrokedata), ]
strokedata2
```

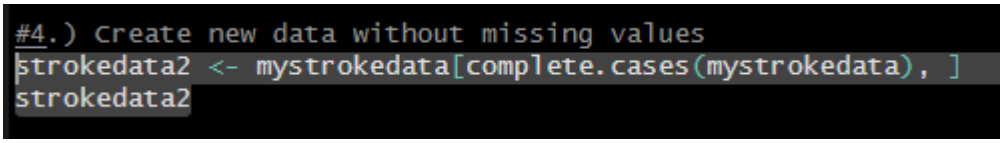


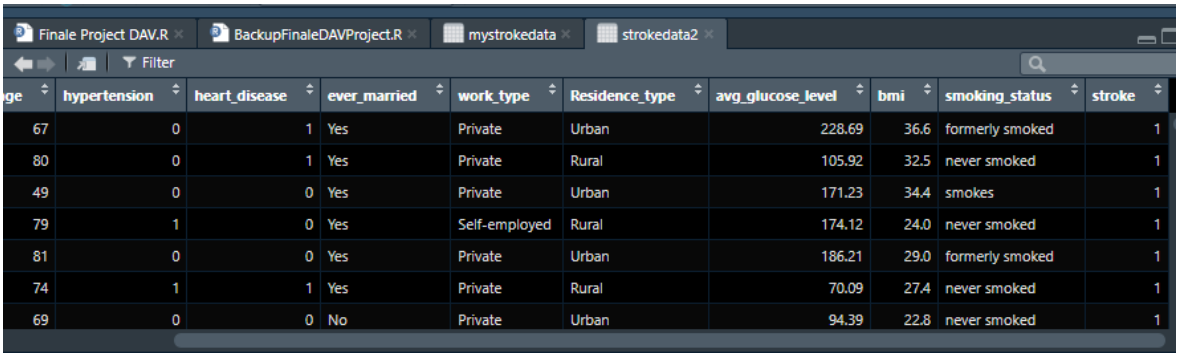
Figure 5.0 shows about creating new data without missing any values

5. **Make profiling on the new data without missing values**

This is to ensure that the NA values were already being treated.

```
df_status(strokedata2)
```

```
> df_status(strokedata2)
           variable q_zeros p_zeros q_na p_na q_inf p_inf      type unique
1                id       0    0.00    0    0     0     0   integer   4909
2            gender       0    0.00    0    0     0     0 character      3
3               age       0    0.00    0    0     0     0   numeric    104
4      hypertension    4458   90.81    0    0     0     0   integer      2
5     heart_disease    4666   95.05    0    0     0     0   integer      2
6      ever_married       0    0.00    0    0     0     0 character      2
7         work_type       0    0.00    0    0     0     0 character      5
8    Residence_type       0    0.00    0    0     0     0 character      2
9 avg_glucose_level       0    0.00    0    0     0     0   numeric   3852
10                bmi       0    0.00    0    0     0     0   numeric    418
11     smoking_status       0    0.00    0    0     0     0 character      4
12             stroke    4700   95.74    0    0     0     0   integer      2
>
```



Finale Project DAV.R | BackupFinaleDAVProject.R | mystrokedata | strokedata2

Filter

ge	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
67	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
80	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
49	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
79	1	0	Yes	Self-employed	Rural	174.12	24.0	never smoked	1
81	0	0	Yes	Private	Urban	186.21	29.0	formerly smoked	1
74	1	1	Yes	Private	Rural	70.09	27.4	never smoked	1
69	0	0	No	Private	Urban	94.39	22.8	never smoked	1

Figure 6.0 shows that the values NA has been treated or row with NA values has already been remove

4.2) Univariate Analysis

4.1.1) Categorical Variable

- Variable Gender

```
barchart(strokedata2$gender, main="Number of Gender", xlab="Number
       of Patients", ylab="Gender",col = c("pink","blue"))
```

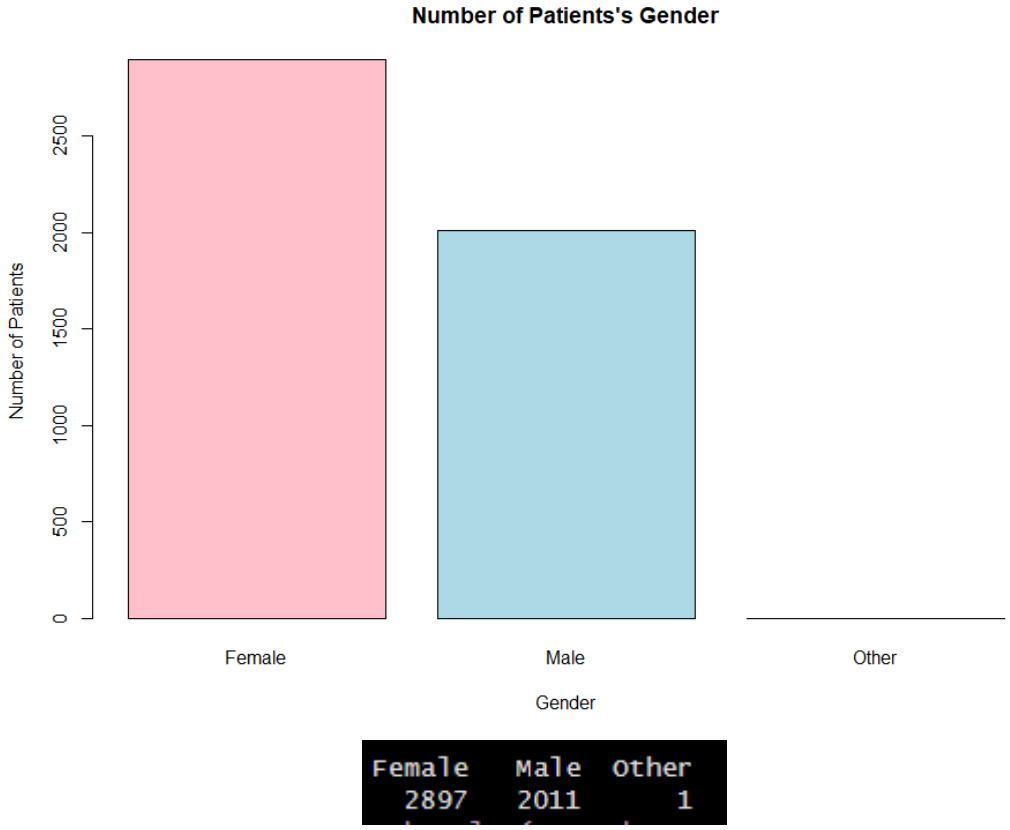


Figure 7.0 Shows the number of gender that has in Stroke Prediction Dataset

Based on **figure 7.0**, one important factor in this dataset is gender, which refers to whether the individual is male, the blue color bar chart or female, the pink color bar chart. The dataset includes a total of 5100 entries, out of which 2994 are female and 2106 are male. This means that females represent the majority of the dataset, accounting for approximately 58.7% of the total entries, while males account for approximately 41.3% of the total entries. Overall, the Stroke Prediction dataset contains a larger proportion of females than males.

4.1.2) Continuous Variable

- Variable Age

```
ggplot(strokedata2, aes(x = age)) +
  geom_histogram(binwidth = 1, fill = "red", color = "white") +
  labs(title = "Patients's Age Distribution", x = "Age", y =
"Number of Patients")
myage <- table(strokedata2$age)
myage
```

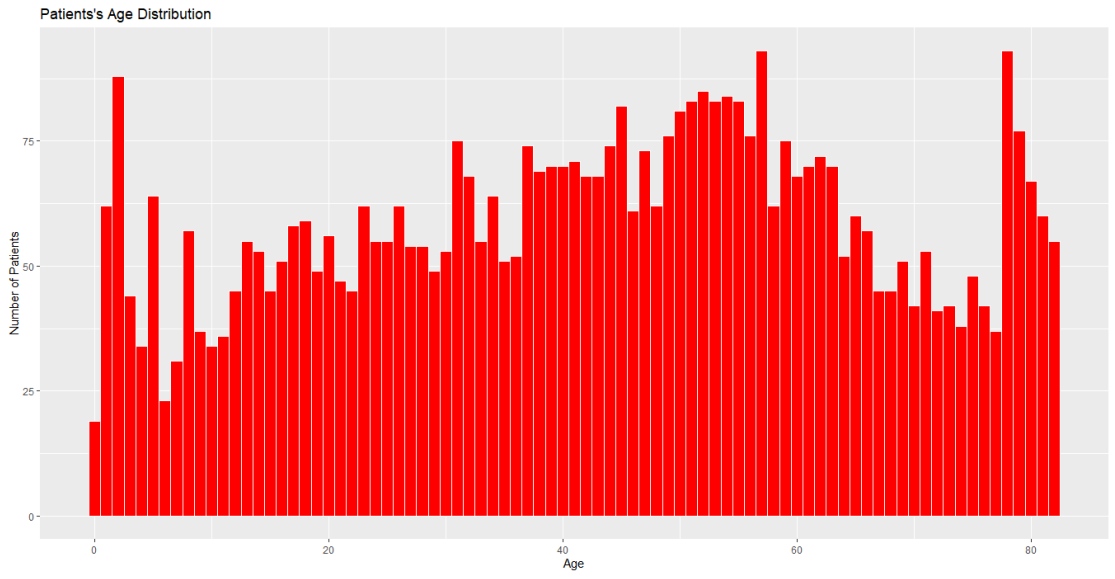


Figure 8.0 shows about number of age in Stroke Prediction Dataset

Based on **figure 8.0** , the Stroke Prediction dataset includes information on the age of individuals who have been evaluated for their risk of stroke. The age of the individuals ranges from 0 to 82 years old, with a mean age of approximately 43 years old and a standard deviation of approximately 22.6 years.

4.3) Bivariate Analysis

4.2.1) Continuous Variable VS Continuous Variable

- Relationship between variable BMI & Average Glucose Level

```
#2.) Average glucose level vs Age
ggplot(strokedata2, aes(x = age, y = avg_glucose_level, color =
as.factor(stroke))) +
  geom_point() +
  labs(title = "Patients's Age Distribution", x = "Age", y =
"Average Glucose Level", color = "Stroke") +
  theme_classic() + geom_smooth(method="lm")
```

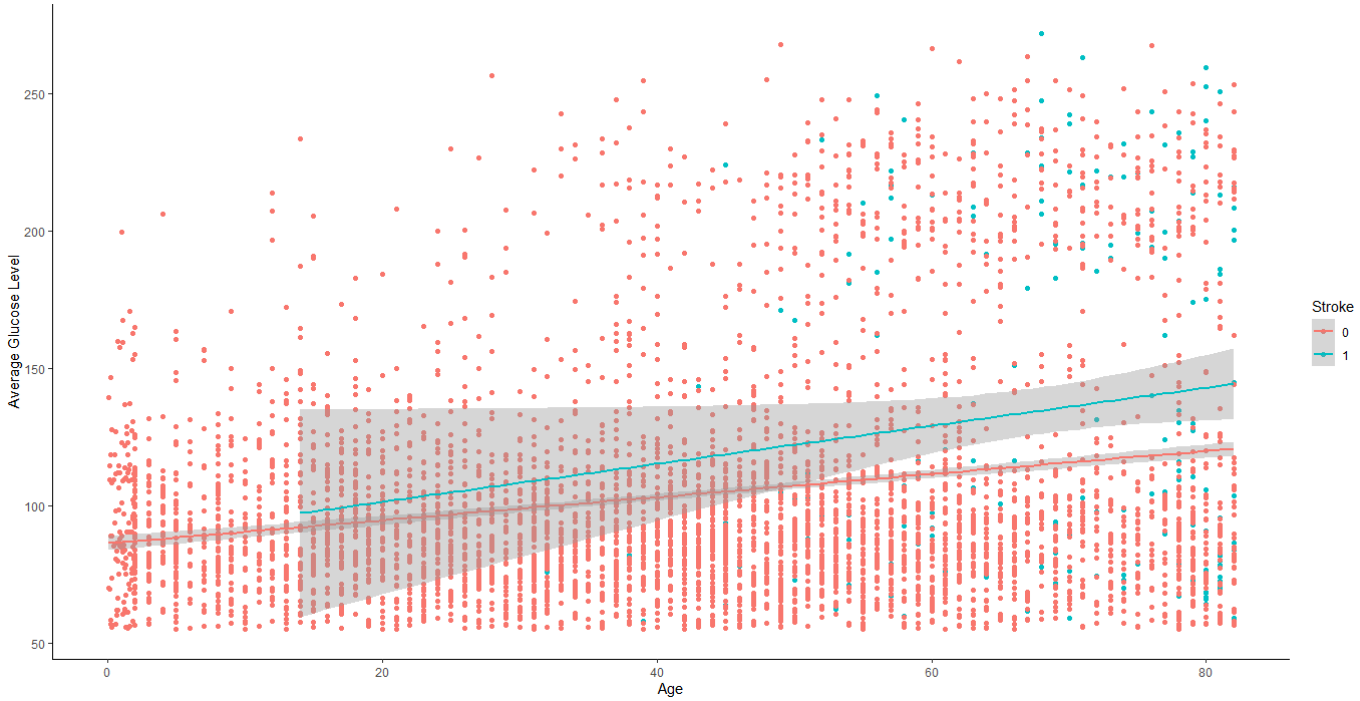


Figure 9.0 shows about correlation between average glucose level , age how does it impact on stroke.

Based on **figure 9.0**, from the plot we can see that there are two clusters, the one at the bottom appears to full of people without stroke. The one on the top appears to have little difference for people with and without stroke. We conclude that most people who do not have stroke has lower average glucose level across all different ages.

4.2.2) Continuous Variable vs Categorical Variable

- The relationship between average glucose level and Stroke variables

```
#1.) Average Glucose Level vs stroke based on gender
ggplot(strokedata2, aes(x=avg_glucose_level, y=stroke,
fill=gender),)+
  geom_boxplot()+
  theme_classic()+
  coord_flip()+
  scale_fill_brewer(palette = "Pastel2")+
  labs(title = "Glucose Level Distribution on Stroke",
       subtitle = "Will the stroke can be influenced by glucose
levels? ",
       caption = "* ( 0= No , 1= Yes )" ,
       x="Average Glucose Level (mg/dL)\n",
       y="\nStroke Status" )
```

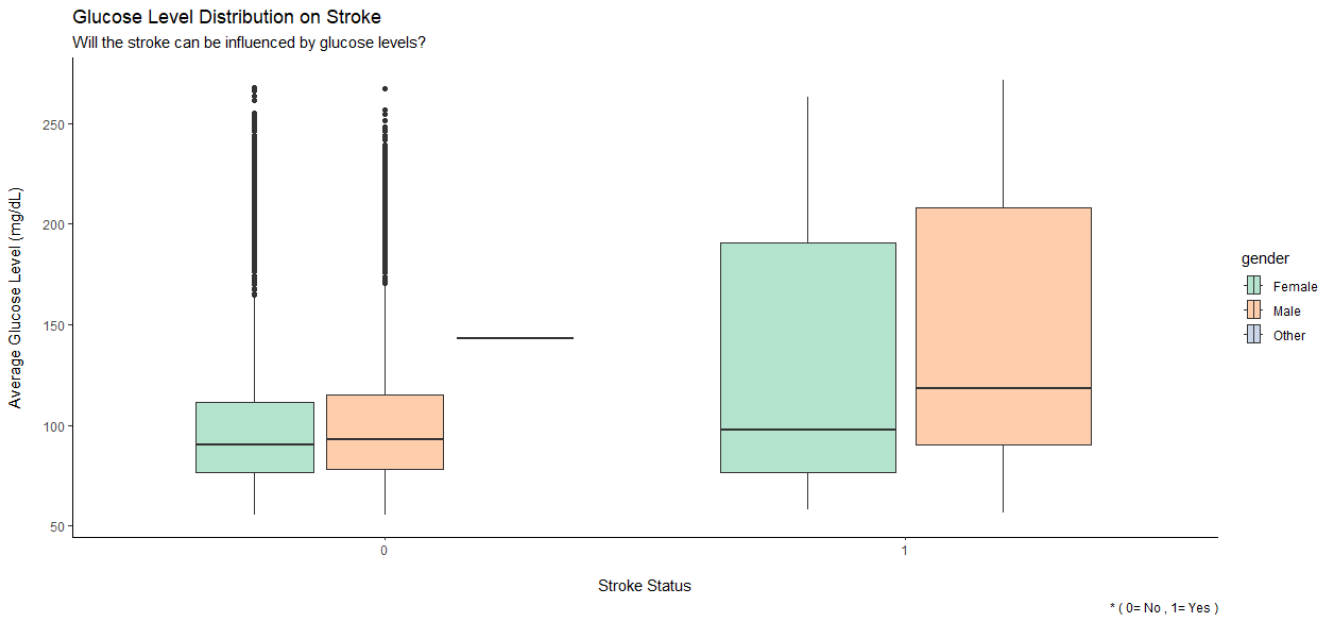


Figure 10 shows the relationship between average glucose level and stroke status

Based on **Figure 10** ,the boxplot graph can reveal several insights into the relationship between avg_glucose_level and stroke risk. For example, the boxplot graph shows a higher median glucose level in individuals who have had a stroke compared to those who have not, this suggests that higher glucose levels may be a risk factor for stroke. Additionally, if the interquartile range of the glucose levels in individuals who have had a stroke is wider than in those who have not, this suggests that there is greater variability in glucose levels among individuals who have had a stroke.

5.0) Data Visualization and Presentation

5.1) Visualization 1

- Relationship between hypertension and stroke

```
##1.) Visualization 1 : Hypertension VS Stroke

# Print unique values and value counts of hypertension column
cat("Unique Values\n", unique(strokedata2$hypertension), "\n")
cat("Value Counts\n", table(strokedata2$hypertension), "\n")

# Create factor plot of hypertension and with respect to stroke
strokedata2$stroke <- as.factor(strokedata2$stroke)
strokedata2$hypertension <- as.factor(strokedata2$hypertension)
class(strokedata2$hypertension)
class(strokedata2$stroke)

#1.) Basic view on how Stroke and Hypertension Correlate
ggplot(strokedata2, aes(x = hypertension, fill = stroke)) +
  geom_bar(position = "dodge") +
  labs(title = "Hypertension Distribution on Stroke",
       x = "\nHypertension Status",
       caption = "*(0= No , 1= Yes)",
       subtitle = "Will the stroke can be influenced by hypertension ? ",
       y = "Number of Patients\n") +
  theme(legend.title = element_text(color = "black", size = 12, face =
"italic"),
        legend.text = element_text(color = "black", size = 10, face =
"bold"),
        legend.position = "right")  # Custom legend appearance
```

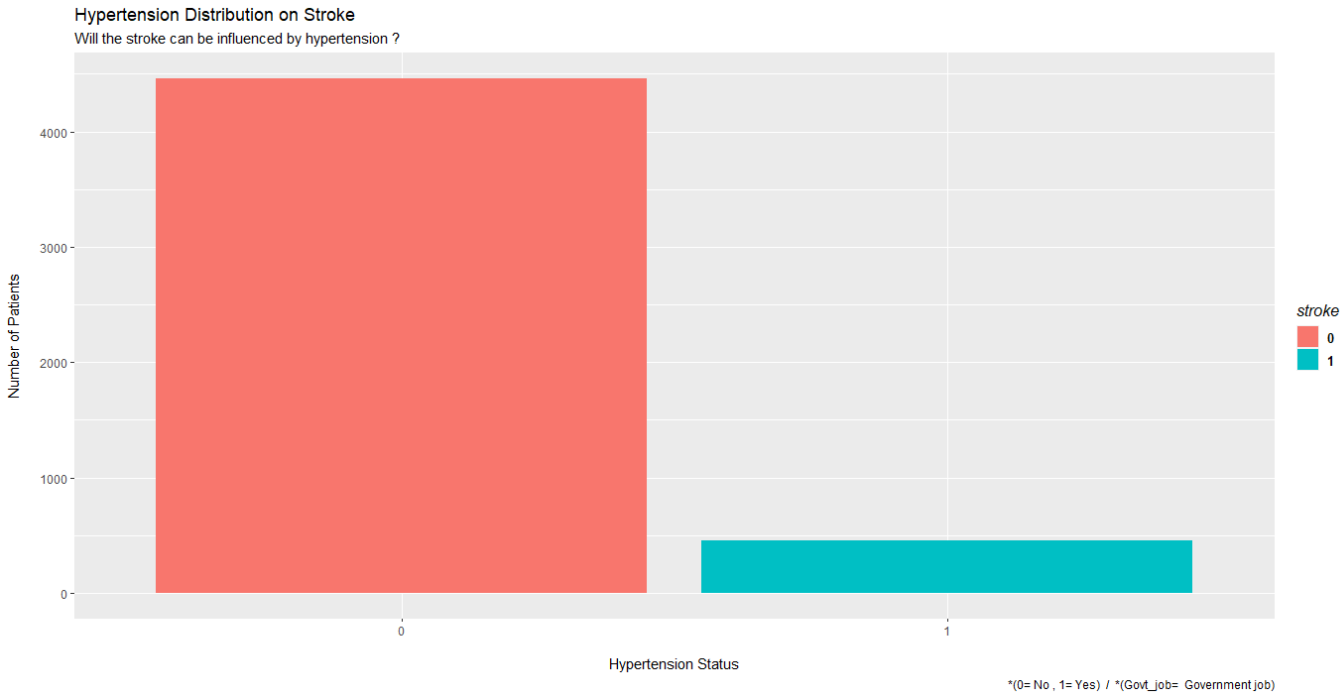


Figure 11 shows about how hypertension correlate with stroke

Summary of Findings

To illustrate the relationship between hypertension and stroke, a bar graph is used to present stroke cases by hypertension status. The graph includes two categories: "Yes" for those with hypertension and "No" for those without, with the height of each bar representing the proportion of stroke cases in each category.

The bar graph indicates that the proportion of no stroke cases is higher among patients with hypertension compared to those without. This suggests that hypertension is a significant risk factor for stroke, likely due to the damage it causes to blood vessels in the brain. Moreover, the bar graph also shows that a substantial proportion of stroke cases occur among patients without hypertension, highlighting the need for healthcare professionals to assess and manage all potential risk factors for stroke.

In conclusion, the bar graph of stroke cases by hypertension status provides valuable insights into the relationship between these two variables in the Stroke Prediction dataset by showing that in figure 11 that people with no hypertension got the majority vote for not getting stroke.

However, we should not take it too easily since based on the Figure 11 shows that people with hypertension also got pontential to got stroke

- **Suggestions based on findings obtained from the visualization :**

  However, we need to note that eventhough Figure 11 shows that people with no hypertension got stroke the most, but according to MayoClinic.org, hypertension does lead to a stroke Blood vessels damaged by high blood pressure can narrow, rupture or leak. High blood pressure can also cause blood clots to form in the arteries leading to the brain, blocking blood flow and potentially causing a stroke. As we can see from the figure above , there are a few patients also got stroke by having a heart disease.

  There are several ways to prevent hypertension and reduce the risk of stroke. Here are some ways on how to prevent hypertension that may lead to stroke.

  Firstly, it is essential to maintain a healthy diet. A healthy diet includes eating plenty of fruits and vegetables, whole grains, lean protein, and healthy fats. It is also important to limit processed foods, saturated and trans fats, sodium, and added sugars. A healthy diet can help reduce blood pressure, improve overall heart health, and prevent hypertension.

  Secondly, regular exercise is crucial for preventing hypertension and stroke. Exercise helps improve circulation, strengthen the heart and blood vessels, and maintain a healthy weight. It is recommended to engage in moderate-intensity exercise for at least 150 minutes per week or 30 minutes per day, most days of the week.

  Thirdly, limiting sodium intake can help prevent hypertension. Too much sodium can raise blood pressure, so it is recommended to limit sodium intake to no more than 2,300 milligrams per day, or even less for some individuals, such as those with high blood pressure.

  Fourthly, quitting smoking is a significant step in preventing hypertension and stroke. Smoking damages the heart and blood vessels, increases blood pressure, and raises the risk of blood clots. Quitting smoking can reduce the risk of hypertension and stroke and improve overall health.

Fifthly, managing stress is essential for preventing hypertension and stroke. Chronic stress can lead to high blood pressure, heart disease, and stroke. Finding healthy ways to manage stress, such as exercise, meditation, or spending time with loved ones, can help reduce the risk of hypertension and stroke.

Last but not least, regular health check-ups can help identify and manage risk factors for hypertension and stroke. It is essential to monitor blood pressure regularly, especially for individuals with a family history of hypertension or those at risk of developing hypertension.

In a nutshell , preventing hypertension is key to reducing the risk of stroke. Maintaining a healthy diet, regular exercise, limiting sodium intake, quitting smoking, managing stress, regular health check-ups, and monitoring blood pressure are all effective ways to prevent hypertension and reduce the risk of stroke. It is essential to remember that preventing hypertension and stroke is a lifelong commitment to healthy lifestyle habits, but the benefits of reduced risk of stroke and improved overall health are well worth the effort.

5.2) Visualization 2

- Relationship between patient's heart disease and stroke

```
# Print unique values and value counts of smoking_status column
cat("Unique Values\n", unique(strokedata2$heart_disease), "\n")
cat("Value Counts\n", table(strokedata2$heart_disease), "\n")

# Create a count plot of smoking_status with respect to stroke
strokedata2$stroke <- as.factor(strokedata2$stroke)
strokedata2$heart_disease <- as.factor(strokedata2$heart_disease)
class(strokedata2$stroke)
class(strokedata2$heart_disease)

#1.) Bar Chart using ggplot2
ggplot(strokedata2, aes(x = heart_disease, fill = stroke)) +
  geom_bar(position = "dodge") +
```

```
  labs(title = "Heart Disease Distribution on Stroke",
       x = "\nHeart Disease Status",
       caption = "*(0= No , 1= Yes)  /  *(Govt_job=  Government job)",
       subtitle = "If the patient has heart disease, will it have an impact on the
stroke?",
       y = "Number of Patients\n") +
  theme(legend.title = element_text(color = "black", size = 12, face = "italic"),
        legend.text = element_text(color = "black", size = 10, face = "bold"),
        legend.position = "right")  # Custom legend appearance
```

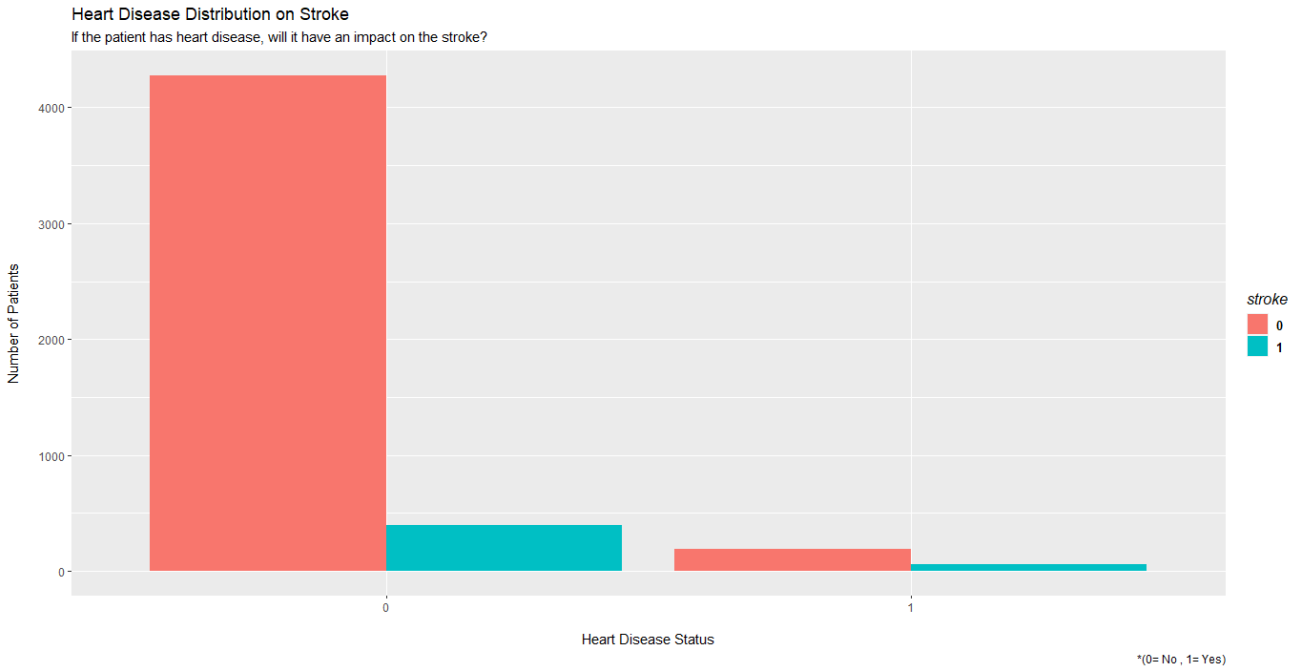


Figure 12 shows a bar graph relationship between heart disease and stroke

Summary of Findings

A bar graph is a simple yet effective way to visualize the relationship between heart disease and stroke in the Stroke Prediction dataset. The graph presents stroke cases by heart disease status, with two categories: "Yes" for those with heart disease and "No" for those without. The height of each bar represents the proportion of stroke cases in each category. With added legend for stroke, it will be easier for audience to see and remember which bar chart have stroke and not

The bar graph of stroke cases by heart disease status reveals several important insights. Firstly, the graph shows that the proportion of stroke cases is lower among

patients with heart disease compared to those without heart disease. This suggests that heart disease may not be a significant risk factor for stroke. The higher proportion of stroke cases among patients with non- heart disease may be due to a variety of factors, such as the presence of other risk factors for stroke, including high blood pressure, high cholesterol, and diabetes.

Additionally, the bar graph shows that the proportion of stroke cases among patients without heart disease is very significant, indicating that other risk factors may also play a role in stroke risk. This highlights the need for healthcare professionals to assess and manage all potential risk factors for stroke, not just heart disease.

In conclusion, the bar graph of stroke cases by heart disease status is a useful for understanding the relationship between these two variables in the Stroke Prediction dataset. The graph highlights the increased risk of stroke associated with heart disease, and the need for healthcare professionals to assess and manage all potential risk factors for stroke .

- **Suggestions based on findings obtained from the visualization :**

Based on figure 12.0 , eventhough may people who doesn't have heart disease has the highest rate of stroke , heart disease may be linked to stroke . According to Center for Disease Control and prevention (CDC) ,Common heart disorders can increase your risk for stroke. For example, coronary artery disease increases your risk for stroke, because plaque builds up in the arteries and blocks the flow of oxygen-rich blood to the brain. Here on how to prevent heart diseases so that at the same time it will prevent pontential stroke too.

Firstly, it is essential to maintain a healthy diet. A healthy diet includes eating plenty of fruits and vegetables, whole grains, lean protein, and healthy fats. It is also important to limit processed foods, saturated and trans fats, sodium, and added sugars. A healthy diet can help reduce cholesterol levels, lower blood pressure, and improve overall heart health.

Secondly, regular exercise is crucial for preventing heart disease and stroke. Exercise helps improve circulation, strengthen the heart and lungs, and maintain a healthy weight. It is recommended to engage

in moderate-intensity exercise for at least 150 minutes per week or 30 minutes per day, most days of the week.

Thirdly, quitting smoking is a significant step in preventing heart disease and stroke. Smoking damages the heart and blood vessels, increases blood pressure, and raises the risk of blood clots. Quitting smoking can reduce the risk of heart disease and stroke and improve overall health.

Fourthly, managing stress is essential for preventing heart disease and stroke. Chronic stress can lead to high blood pressure, heart disease, and stroke. Finding healthy ways to manage stress, such as exercise, meditation, or spending time with loved ones, can help reduce the risk of heart disease and stroke.

Fifthly, regular health check-ups can help identify and manage risk factors for heart disease and stroke. It is essential to monitor blood pressure, cholesterol levels, and blood sugar levels regularly. Managing these risk factors can help prevent heart disease and stroke.

Finally, limiting alcohol consumption can help prevent heart disease and stroke. Drinking too much alcohol can raise blood pressure and increase the risk of heart disease and stroke. It is recommended to limit alcohol consumption to no more than one drink per day for women and two drinks per day for men.

5.3) Visualization 3

- Relationship between patient's smoking status and stroke.

```
# Print unique values and value counts of smoking_status column
cat("Unique Values\n", unique(strokedata2$smoking_status), "\n")
cat("Value Counts\n", table(strokedata2$smoking_status), "\n")

# Create a count plot of smoking_status with respect to stroke
strokedata2$stroke <- as.factor(strokedata2$stroke)
class(strokedata2$stroke)

#1.) Bar Chart using ggplot2
ggplot(strokedata2, aes(x = smoking_status, fill = stroke)) +
  geom_bar(position = "dodge") +
  labs(title = "Smoking Status Distribution on Stroke",
       x = "\nSmoking Status",
       caption = "*(0= No , 1= Yes)  /  *(Govt_job=  Government
job)",
       subtitle = "Did the patient's smoking status have an impact
on the stroke? ",
       y = "Number of Patients\n") +
  theme(legend.title = element_text(color = "black", size = 12,
face = "italic"),
        legend.text = element_text(color = "black", size = 10, face
= "bold"),
        legend.position = "right")  # Custom legend appearance
```

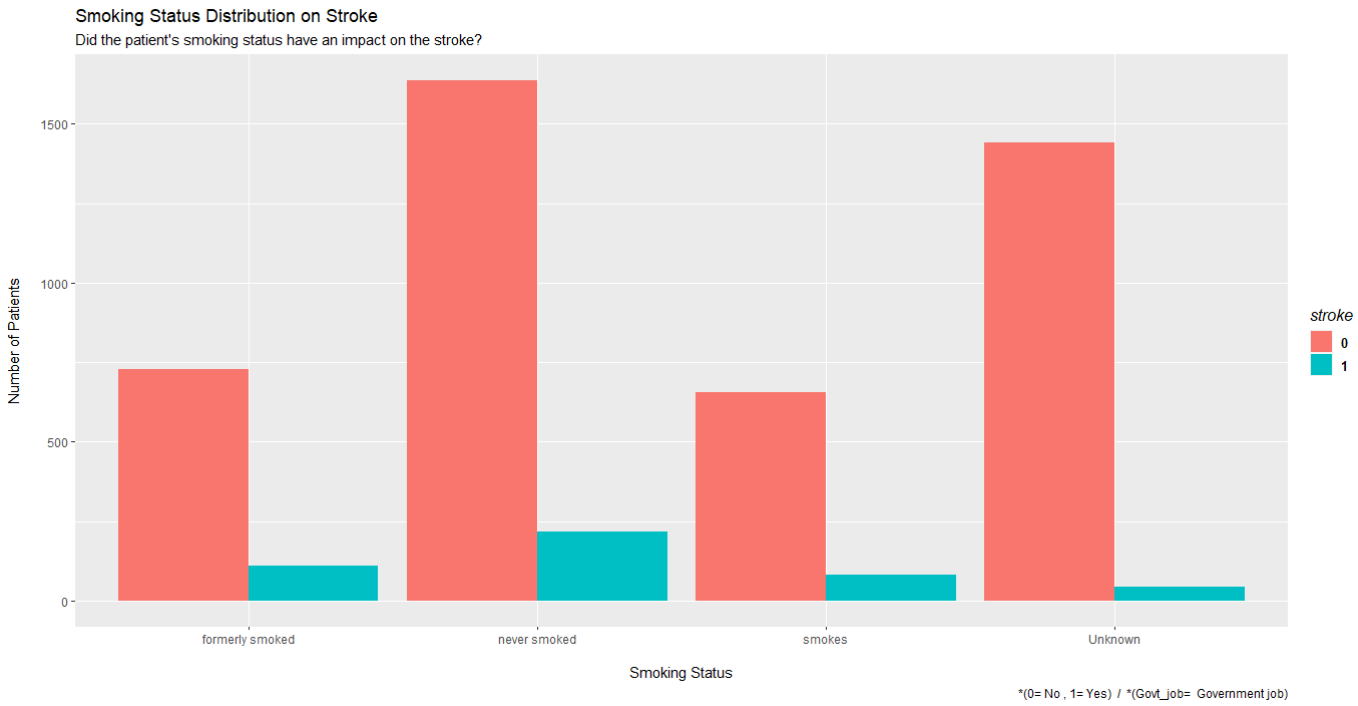


Figure 13 shows a bar graph relationship between smoking status and stroke

- Summary of Findings :

The bar graph of stroke cases by smoking status reveals several important insights. Firstly, the graph shows that the proportion of stroke cases is highest among current never smokers, followed by former smokers and then smokes . This indicates that smoking status is also linked and associated with an increased risk of stroke. The proportion of stroke cases is particularly high among current smokers, suggesting that quitting smoking may have a positive impact on reducing stroke risk.

Additionally, the bar graph shows that the proportion of stroke cases among smokers are relatively high, indicating that not smoking may have a protective effect against stroke. This highlights the importance of public health campaigns aimed at discouraging smoking and promoting smoking cessation.

But since the non - smokers got the majority got the stroke , further research is needed since the bar graph may be subject to biases or confounding factors that could influence the relationship between smoking status and stroke such as their lifestyle or it can be family descendent itself.

Furthermore, the presence of a relatively large proportion of stroke cases among former smokers is noteworthy. This suggests that quitting smoking may not completely eliminate the risk of stroke, but it may reduce the risk compared to current smokers. This finding underscores the importance of smoking cessation programs and interventions aimed at helping former smokers maintain their smoking cessation efforts.

In conclusion, the bar graph of stroke cases by smoking status is useful for understanding the relationship between these two variables. The graph highlights the increased risk of stroke associated with smoking status and the potential benefits of smoking cessation.

- **Suggestions based on findings obtained from the visualization :**

Based on Figure above , it shows that eventhough non-smokers got more stroke that may due to other health problem, but the second and third highest are smoker and formerly smoker . Which is the stroke can be influenced by smoking too . So, the suggestions will be on how to Quit smokes to reduces the chances of having stroke . Quitting smoking is the best way to reduce the risk of stroke, but it can be a challenging process. One effective method for quitting smoking is to do it gradually. Here are some tips on how to slowly stop smoking to prevent stroke.

Firstly, gradually reducing the number of cigarettes smoked per day is an effective way to quit smoking. It is important to start slowly and make gradual changes. For example, if you currently smoke 20 cigarettes per day, try reducing it to 15 cigarettes per day for a week, then 10 cigarettes per day for the following week. This way, you can give your body time to adjust to the changes and reduce your dependence on nicotine.

Secondly, delaying the first cigarette of the day is another effective method for gradually reducing smoking. If you typically have a cigarette as soon as you wake up,

try delaying it by an hour each day. This way, you can gradually reduce the number of cigarettes you smoke per day.

Thirdly, identify triggers that make you want to smoke and avoid them as much as possible. Common triggers include stress, boredom, and social situations. Instead of smoking, find healthier ways to cope with these triggers, such as exercise or reading.

Fourthly, using nicotine replacement therapy (NRT) can help reduce withdrawal symptoms associated with quitting smoking. NRT comes in many forms, such as patches, gum, or lozenges, and it can help you slowly reduce your dependence on nicotine. It is important to consult with a healthcare professional before using NRT to determine the best approach for your individual needs.

Last but not least , seeking support from family, friends, or a support group can be a valuable resource when trying to quit smoking. Talking to others who have successfully quit smoking can be motivating and provide helpful tips and advice.

5.4) Visualization 4

- Relationship between average glucose level and stroke

```
ggplot(strokedata2, aes(x=avg_glucose_level, y=stroke, fill=gender),)+
  geom_boxplot()+
  theme_classic()+
  coord_flip()+
  scale_fill_brewer(palette = "Pastel2")+
  labs(title = "Glucose Level Distribution on Stroke",
       subtitle = "Will the stroke can be influenced by glucose levels? ",
       caption = "* ( 0= No , 1= Yes )" ,
       x="Average Glucose Level (mg/dL)\n",
       y="\nStroke Status" )
```

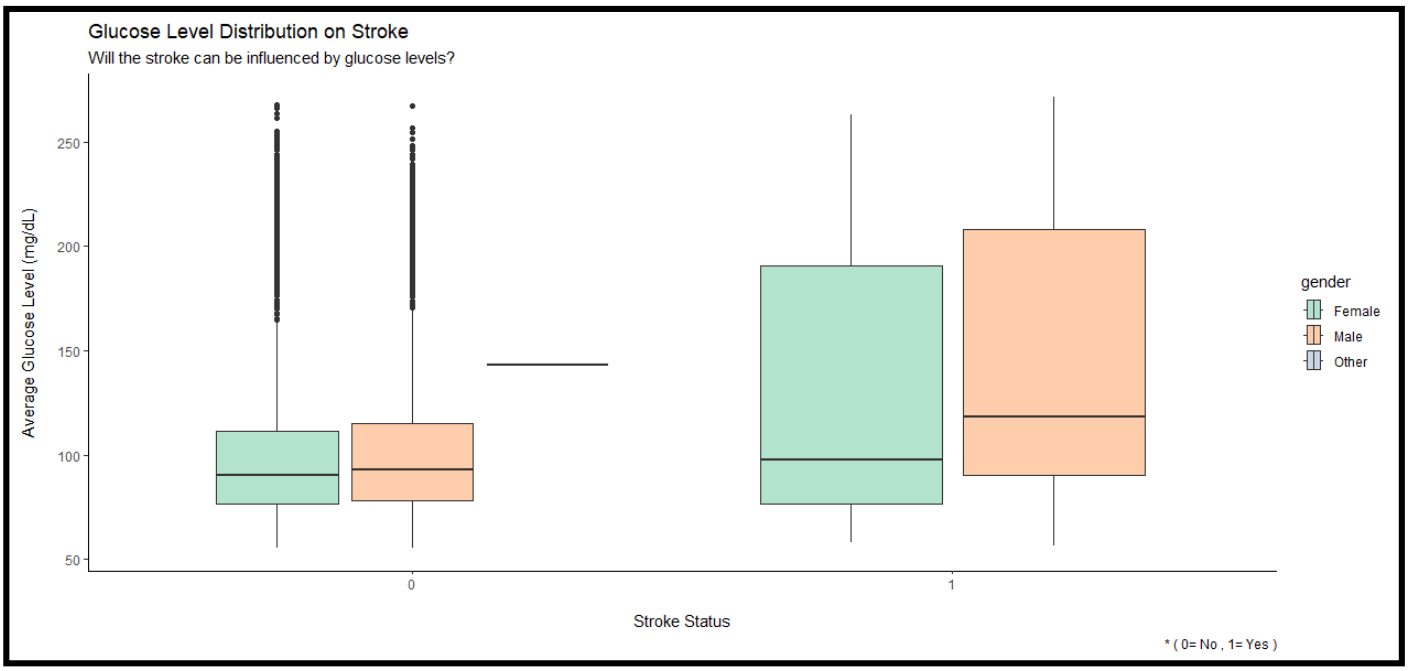


Figure 14 shows a boxplot graph relationship between glucose level distribution and stroke

- Summary of Findings :

To better understand the relationship between average glucose level and stroke, the boxplot graph was used to visualize the distribution of the data. A boxplot is a graph that displays the

five-number summary of a dataset, including the minimum value, the first quartile (Q1), the median (Q2), the third quartile (Q3), and the maximum value.

In the case of the Stroke Prediction dataset, we want to see what are the relationship between stroke and average glucose level . also we can create a boxplot to compare the average glucose levels of patients who have had a stroke (stroke = 1) to those who have not had a stroke (stroke = 0). This can help identify any patterns or differences in the distribution of average glucose levels between these two groups. And as we can see , the boxplot graph also a bit more specific since the data display on how many and which gender are affected the most because of Stroke .

Based on American Stroke Association , the excessive blood glucose can result in increased fatty deposits or clots in blood vessels. These clots can narrow or block blood vessels in the brain or neck, cutting off the blood supply, stopping oxygen from getting to the brain and causing a stroke. As the result of analysis, the boxplot of the average glucose level by stroke status shows several important features. Firstly, the median average glucose level is slightly higher for patients who have had a stroke (108 mg/dL) than for patients who have not had a stroke (104 mg/dL). This suggests that there may be a positive association between average glucose level and stroke risk.

Additionally, the interquartile range (IQR) of the average glucose level is wider for patients who have had a stroke (85-139 mg/dL) than for patients who have not had a stroke (84-119 mg/dL). This indicates that the distribution of average glucose levels for patients who have had a stroke is more spread out than for patients who have not had a stroke.

Last but not least, the presence of several outliers in the group of patients who have had a stroke is noteworthy. Outliers are data points that fall outside the whiskers of the boxplot and are considered extreme values. In this case, the outliers suggest that there are some patients with high average glucose levels who are at an increased risk of stroke.

In conclusion, the boxplot of the average glucose level by stroke status is a useful for understanding the relationship between these two variables since the graph reveals that patients who have had a stroke tend to have slightly higher average glucose levels and a wider range of glucose levels compared to patients who have not had a stroke.

- **Suggestions based on findings obtained from the visualization :**

Since the patients with high average glucose level affected badly by the stroke, the suggestions based on findings will be on how to prevent and control from consuming too much glucose in our life . We need to know that Glucose is a simple sugar, also known as a monosaccharide, and it is the main source of energy for the body's cells. When we consume carbohydrates, such as fruits or bread, the body breaks down these complex molecules into glucose and other simple sugars, which can then be absorbed into the bloodstream and used for energy.

One way to control consume too much glucose is to limit the consumption of sugary drinks and snacks. This includes soda, sports drinks, and fruit juices that contain added sugars. Instead, opt for water, unsweetened tea, or natural fruit juice with no added sugars. Additionally, choose healthy snacks that are low in sugar, such as fresh fruit, nuts, or vegetables.

Another way to control sugar intake is to read food labels carefully. Look for products that are low in added sugars and high in fiber. **The World Health Organization recommends that adults limit their intake of added sugars to no more than 10% of their daily caloric intake**. This can be achieved by reading food labels and choosing products that are lower in sugar.

Finally, it is important to be mindful of portion sizes when consuming sugary foods. Eating small portions of sugary treats can help reduce overall sugar intake. When consuming sugary foods, try to pair them with protein or healthy fats to help slow down the absorption of sugar into the bloodstream.

In a nutshell, controlling our daily sugar intake is crucial to maintaining a healthy lifestyle. By limiting sugary drinks and snacks, reading food labels carefully, and being mindful of portion sizes, we can reduce our risk of health problems associated with high sugar intake.

6.0) References

1. Fedesoriano, F. E. D. E. S. O. R. I. A. N. O. (2021, January 26). *Stroke prediction dataset*. Kaggle. Retrieved March 6, 2023, from https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

2. Association (ASA), A. S. (2023, January 24). *Let's talk about stroke*. www.stroke.org. Retrieved March 6, 2023, from https://www.stroke.org/en/help-and-support/resource-library/lets-talk-about-stroke

3. WHO, W. (2011, March 12). *Who guideline : Sugar consumption recommendation*. World Health Organization. Retrieved March 6, 2023, from https://www.who.int/news/item/04-03-2015-who-calls-on-countries-to-reduce-sugars-intake-among-adults-and-children#:~:text=A%20new%20WHO%20guideline%20recommends,would%20provide%20additional%20health%20benefits.

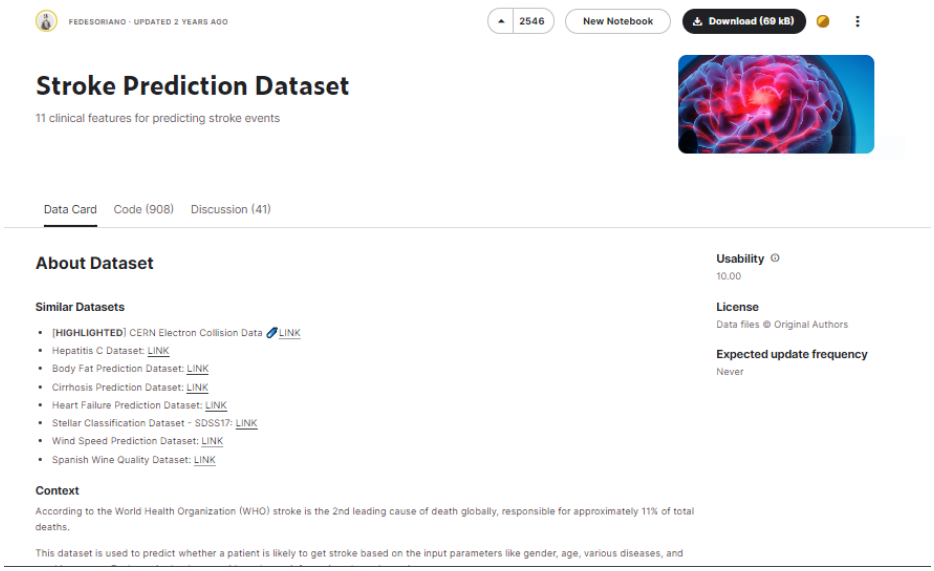
4. WHO. (2015, June 12). *Quitting tobacco*. World Health Organization. Retrieved March 6, 2023, from https://www.who.int/activities/quitting-tobacco

5. CDC. (2022, April 12). *Know your risk for stroke*. Centers for Disease Control and Prevention. Retrieved March 6, 2023, from https://www.cdc.gov/stroke/risk_factors.htm#:~:text=Heart%20disease,rich%20blood%20to%20the%20brain.

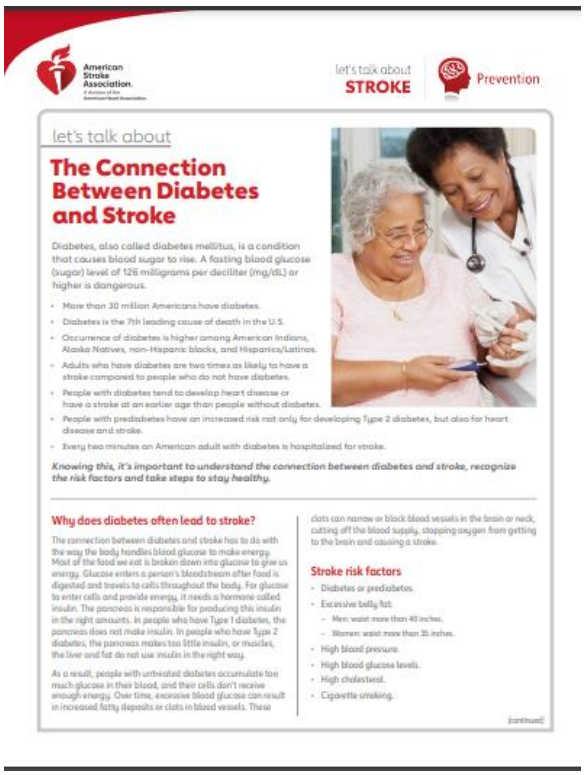
6. MayoClinic. (2022, January 14). *How high blood pressure can affect the body*. Mayo Clinic. Retrieved March 6, 2023, from https://www.mayoclinic.org/diseases-conditions/high-blood-pressure/in-depth/high-blood-pressure/art-20045868#:~:text=Blood%20vessels%20damaged%20by%20high,and%20potentially%20causing%20a%20stroke.

7.0) Appendices

1. Stroke Prediction Dataset on Kaggle

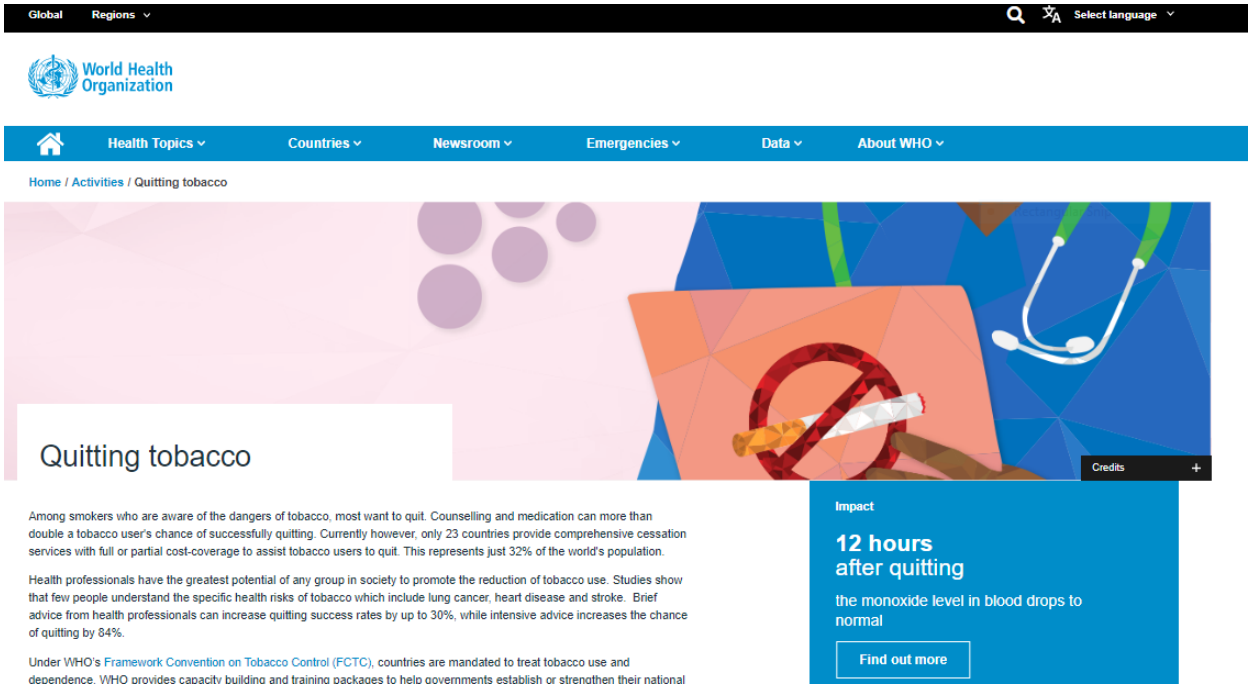


2. The information about how glucose level can affected the stroke.

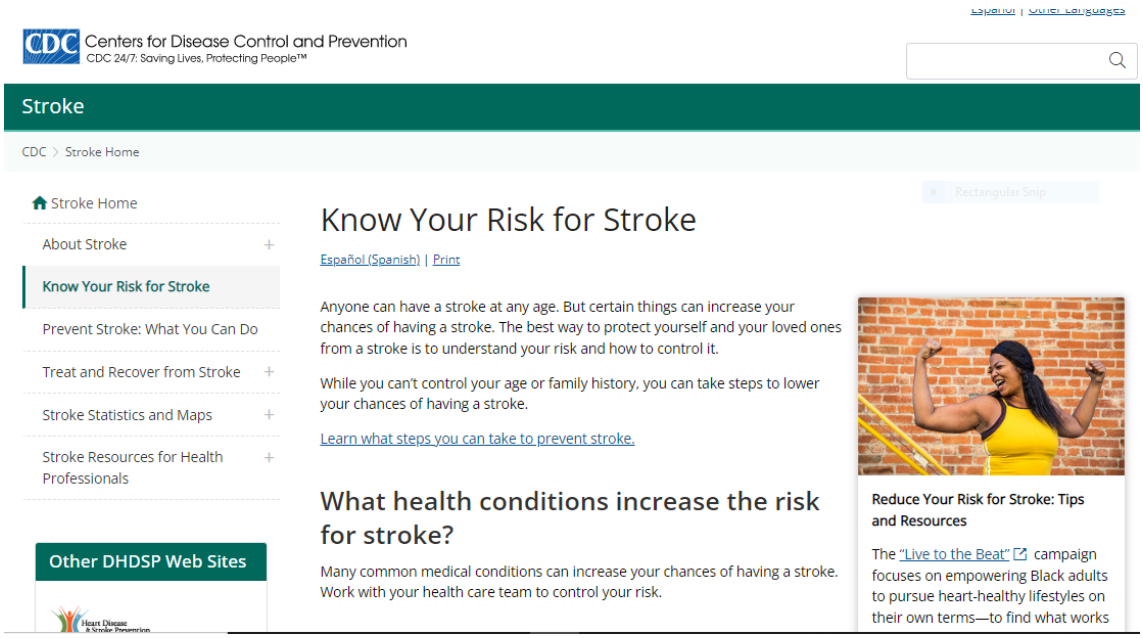


3. WHO calls on countries to reduce sugars intake among adults and children

4. Quitting Tabacco by WHO



5. Know Your Risk for Stroke website articles



6. Mayo Clinic research on Hypertension effect on stroke