PROJECT REPORT

Uncovering Critical Factors Linked to the Risk of Developing Stroke

Lakshmi Aparna Valiveti, Navya Kamepalli, Sreevidhya P V, Timothy Gudisa, Surya R Sripathi

Introduction:

Stroke is a leading cause of death worldwide, leaving the majority of survivors with long-term impairment. According to recent projections, the annual cost of stroke care in the United States alone would be $46 billion, or roughly $140 per person in the country. The complex pathophysiology of stroke involves multiple causative factors, the most prevalent of which are embolus, thrombus, and atherosclerosis. (Shehjar et al., 2023). Stroke is the nation's second major risk of death globally, accounting for about 11% of all fatalities, according to WHO data.

Stroke is still the second significant cause of death globally and the third noticeable cause of death and disability combined, according to disability-adjusted life years lost (DALYs). These findings were reported by Feigin et al. (2022). Elderly people are more vulnerable, and they are more prevalent in wealthy nations. A worldwide survey conducted between 1990 and 2010 revealed a significant rise in overall strokes and related deaths, as well as a 25% increase in strokes among people aged 20 to 64 and a 113% increase in stroke survivors.

Problem statement:

The aim of this project is to pinpoint the key factors that impact the likelihood of an individual suffering a stroke and to evaluate the probability of stroke occurrence based on their health and demographic attributes.

Research questions:

1. Are there significant differences in BMI between stroke-affected and non-affected individuals?
2. Are lifestyle factors (smoking) correlated with stroke risk?
3. Which demographic elements (age, gender) are associated with higher stroke risk?

DATA

Link for data set:
https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

Data set and variables

The stroke dataset, comprising 5110 records and 12 variables, provides a comprehensive snapshot of various health-related aspects. Among the numerical variables, 'age,' 'avg_glucose_level,' and 'bmi' exhibit considerable variability. The 'age' variable spans from 0.08 to 82, with an average age of 43.23, indicating a diverse age distribution. Similarly, 'avg_glucose_level' ranges from around 55.12 to 271.74, with an average of 106.15, highlighting

a broad spectrum of glucose levels. The 'bmi' values range from 10.30 to 97.60, with an average BMI of 28.86, highlighting diversity in body mass indices.

Categorical variables such as 'gender,' 'ever_married,' and 'work_type' offer insights into demographic and lifestyle factors. 'Gender' reveals a distribution of 2994 females, 2115 males, and one individual listed under 'Other.' Marital status indicates that 1757 individuals have not been married, while 3353 are married. Work types encompass diverse categories, including 'children,' 'Govt_job,' 'Never_worked,' 'Private,' and 'Self-employed.' The 'Residence_type' variable reflects that 2514 people live in rural areas, while 2596 reside in urban areas.

Furthermore, health-related variables like 'hypertension' and 'heart_disease' shed light on the prevalence of these conditions. The 'stroke' variable, indicating the presence or absence of stroke, has a mean of 0.04873, suggesting a low prevalence of stroke in the dataset. However, it is crucial to note that the 'smoking_status' variable contains missing values (NA's), indicating instances where smoking status is not recorded. 'Smoking_status' has three categories: '0' for 'never smoked,' '1' for 'smokes,' and 'Unknown.' On average, 4.87% of individuals in the dataset have experienced a stroke.

METHODS

Exploratory Data analysis: To gain an understanding of the connections between these predictor and predictive variables, we have examined the data and produced visualizations. We produced histogram of the BMI column to understand the distribution of the data. We infer from the plot that most of the sample fell in a BMI range between 20 to 3030. Later to dig deep into the BMI influence on the stroke we plotted a histogram between stroke-positive patients and BMI. We were able to clearly visualize the peak of the histogram at BMI level 28 with highest frequency (Figure.1., Figure.2.).

To visualize the smoking status column, we excluded the sample size having an age less than 20 to avoid bias as the sample included minors. By using this pattern, we inferred that half of the sample size were smokers and the other half were non-smokers (Figure.3.). To delve deeper we plotted a bar graph including just the stroke-positive patients where we got results like the prior analysis (Figure.4.).

To visualize the age data, we opted for box plots as they produce a clear picture and stand of age influence on stroke. The plot showed that the people with stroke were between 60 to 80 years old (Figure.5.). This provided a lead making us estimate that age would have an association with stroke occurrence. To view our other research demographic feature gender, we employed bar graphs for the visualization. We inferred from the bar plot that the female patients are comparatively higher than the males (Figure.6.,Figure.7.).

STATSTICAL ANALYSIS

Q 1) Are there significant differences in BMI between stroke-affected and non-affected individuals?

HYPOTHESIS:

Null hypothesis: There is no significant BMI difference between stroke and non-stroke individuals.

Alternative hypothesis: There is a Significant BMI difference between stroke and non-stroke individuals.

Rationale

The Shapiro-Wilk test was employed to assess the 'bmi' column from the 'stroke.data' dataset for normality. This test is crucial as it evaluates whether the BMI values adhere to a normal distribution, a necessary assumption before performing statistical analysis. Ensuring normality supports the validity of applying these parametric tests in statistical analyses.To ensure consistency, a seed of 123 was set for reproducibility, enabling consistent results across multiple code executions. This practice enhances result stability and is pivotal for validation and comparability across various research analyses.

The R code executes a Wilcoxon signed rank test on the 'BMI' values, aiming to compare two distinct groups represented by the 'stroke' column in the 'stroke.data' dataframe. The rationale for conducting the Wilcoxon test lies in its ability to assess differences in median BMI between two groups, specifically individuals affected by stroke and those not affected. This test serves as a robust alternative to parametric tests when assumptions of normality are violated or when dealing with non-normally distributed data. It operates without requiring strict assumptions about the data distribution, making it suitable for scenarios where data might not meet the criteria for parametric analyses.

FINDINGS:

The provided R code conducts a Shapiro-Wilk test on a sample of BMI (Body Mass Index) data extracted from the 'stroke.data' dataset. This statistical test evaluates whether the distribution of BMI values conforms to a normal distribution. Upon executing the test, the obtained results exhibit a W statistic of 0.9502 and an extremely low p-value of less than 2.2e-16. These results strongly indicate that the BMI data, when sampled from the 'stroke.data' dataset, significantly deviates from a normal distribution.

The Wilcoxon rank sum test results show a statistically significant difference in BMI between those who have had a stroke and those who have not. Test Statistic (W): The test statistic is 522643. To find out if there is a difference in the median BMI between the two groups, the Wilcoxon test uses this value as the calculated test statistic. The calculated p-value is 0.0002769, which is below the accepted significance level of 0.05. The observed difference in median BMI between stroke-affected and non-affected individuals is unlikely to have happened by accident, as suggested by this small p-value, which provides strong evidence against the null hypothesis. These findings suggest that the BMI values of the two groups under investigation differ significantly. There may be a relationship between having a stroke and BMI levels because the statistically significant difference shows that there is more than just random variability between those who have had a stroke and those who have not. (Figure.8.).

Limitations

Large sample sizes can highlight minor deviations as significant, while smaller ones may limit the test's ability to detect deviations, requiring consideration in interpretation. Although it identifies deviations broadly, the test cannot precisely specify the nature or degree of deviation, like skewness or kurtosis. Extreme values can disproportionately impact the test, causing rejection of normality even if most data follow a normal distribution. The Wilcoxon test's limitations for comparing BMI between stroke-affected and non-affected groups include sensitivity to sample size, reliance on ranking rather than actual values, lack of direct quantification of effect size, assumptions about variance homogeneity, and the absence of consideration for covariates or confounding factors.

Q2) Are lifestyle factors (smoking) correlated with stroke risk?

Hypothesis:

Null Hypothesis (H0): There is no significant association between smoking status and stroke risk.

Alternative Hypothesis (H1): There is a significant association between smoking status and stroke risk.

Rationale:

Our choice of the Chi-squared test stems from its suitability for analyzing categorical data. This test is ideal for evaluating the independence of two variables - in our case, smoking status, and stroke risk. We employed random sampling to ensure our sample was representative of the larger dataset, thereby aiming to enhance the generalizability of our findings. Additionally, setting an example for reproducibility ensures that our analysis can be independently verified, a crucial aspect of scientific research.

Findings:

The R code evaluates the relationship between a dataset's smoking status and stroke risk using a Pearson's Chi-squared test. Contingency table analysis results indicate that the assessed statistic (X)-squared is 3.7743 combined with 2 degrees of freedom and a significance level (p-value) of 0.1515 when a randomly sampled subset with a seed set for reproducibility is used, and samples under the age of 20 are excluded. Since the p-value is higher than the 0.05 threshold, the test's conclusion indicates that there is no meaningful correlation between smoking status and stroke risk across the examined dataset and age range. As a result, the null hypothesis is not rejected, suggesting that, in this particular situation, smoking status may not be a significant predictor of stroke risk (Figure.9.).

Limitations:

Our study's limitations include the potential for reporting bias due to reliance on self-reported smoking status. We acknowledge that our approach does not account for other variables that might influence stroke risk, such as diet or genetic factors. Our data are cross-sectional, which restricts our ability to prove causation and leaves us only with associations. These limitations highlight the need for cautious interpretation of our findings and suggest areas for future research.

Q3) Which demographic elements (age, gender) are associated with higher stroke risk?

Hypothesis:

Null (H0): There is no significant association between age and stroke occurrence.

Alternative (H3): There is a significant association between age and stroke occurrence.

Rationale for age

The Shapiro-Wilk test assesses normality, a crucial assumption for many statistical analyses. In this case, it verifies whether the age distribution meets the normality assumption. Deviations from normality might impact subsequent parametric analyses. The Wilcoxon-signed rank test for independent samples looks at whether there is a significant difference in mean age between stroke victims and survivors. A major age difference between the stroke-positive and stroke-negative groups is implied by a significant p-value, which indicates that the observed difference in mean age is unlikely to have happened by coincidence.

Findings

The Shapiro-Wilk test was conducted on the age data. The test result indicated that the age distribution was not normally distributed (p-value = 2.2e-16). This suggests that the age data deviates significantly from a normal distribution. The R code performs a Wilcoxon-signed rank test on age data from stroke dataset. Results show a highly significant age difference between stroke-positive and stroke-negative groups ($t = -23.293$, $p < 2.2e-16$). The mean age for stroke-positive individuals (68.21) is notably higher than stroke-negative (49.96). Findings imply age

significantly correlates with stroke occurrence, with stroke-affected individuals older than non-affected individuals (Figure.10.).

Limitations

The results of the age test may have limitations because of sample representation. Results may not be generalizable if the dataset is not a complete representation of the population. Furthermore, if samples under the age of 20 are excluded, it may be possible to miss stroke cases in younger people, which could affect the analysis's comprehensiveness. Furthermore, although the test indicates a noteworthy age gap between the groups afflicted by stroke and those who are not, it does not prove a cause-and-effect relationship; the incidence of stroke may be influenced by other unidentified factors.

Rationale for gender

With the help of a chi-squared test, the provided R code seeks to examine any possible correlation between gender and stroke risk within the stroke.data dataset. First, the code guarantees reproducibility of the results by first setting a seed. After that, a contingency table is created and displayed, classifying and arranging the frequency of stroke incidents by gender. Using this contingency table, the chi-squared test is run to see if there are any statistically significant gender-based variations in the incidence of stroke. The test determines degrees of freedom, a corresponding p-value, and a chi-squared statistic. The test results are accompanied by a warning message, though, indicating possible doubts regarding the chi-squared approximation's accuracy for this dataset.

Findings

Examining the test results, the computed p-value of 0.7743 is greater than the standard significance level of 0.05. Given the contingency table's indication of an observed relationship between gender and stroke risk, the high p-value implies that the relationship is not statistically significant. Accordingly, this analysis indicates that there is insufficient data to substantiate a relationship between gender and the chance of having a stroke in this dataset (Figure.11).

Limitations

Representing gender in binary categories in the gender analysis may restrict the test's applicability by ignoring gender variability outside of the binary classification. A lack of representation for some genders in the dataset could skew the results and make it more difficult to comprehend the full relationship between gender and stroke risk. If the effect size is tiny or the sample size is insufficient, the chi-squared test may also have poor statistical power and overlook subtle relationships.

References:

Shehjar, F., Maktabi, B., Rahman, Z., Bahader, G. A., James, A. W., Naqvi, A., Mahajan, R., & Shah, Z. A. (2023). Stroke: Molecular mechanisms and therapies: Update on recent developments. *Neurochemistry International*, *162*, 105458. https://doi.org/10.1016/j.neuint.2022.105458

Feigin, V. L., Brainin, M., Norrving, B., Martins, S. C. O., Sacco, R. L., Hacke, W., Fisher, M., Pandian, J. D., & Lindsay, P. (2022). World Stroke Organization (WSO): Global stroke fact sheet 2022. *International Journal of Stroke*, *17*(1), 18–29. https://doi.org/10.1177/17474930211065917

APPENDIX

Figure.1.

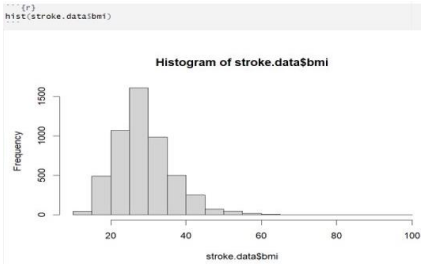


Figure.2.

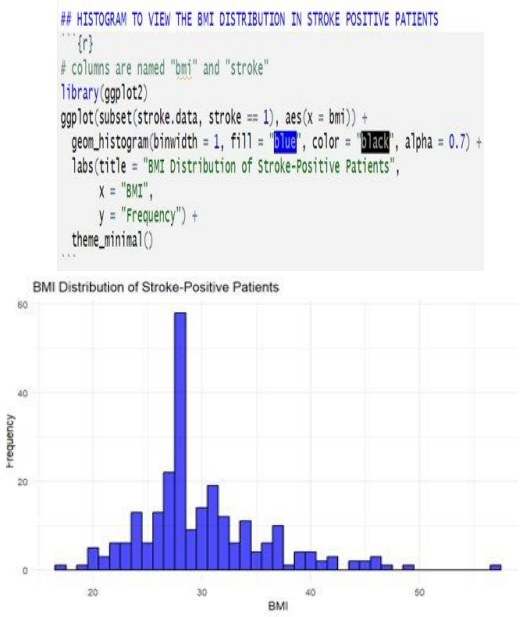


Figure.3.

```
## BAR GRAPH TO VIEW THE SMOKING STATUS DISTRIBUTION
```

```{r}
# Assuming your column is named "smoking_status"
# Assuming your age column is named "age"

library(ggplot2)

# Filter data for age greater than or equal to 20
filtered_data <- stroke.data[stroke.data$age >= 20,]

ggplot(filtered_data, aes(x = factor(smoking_status))) +
  geom_bar(fill = "skyblue", color = "black") +
  labs(title = "Smoking Status Distribution (Age >= 20)",
       x = "Smoking Status",
       y = "Frequency") +
  theme_minimal()
```

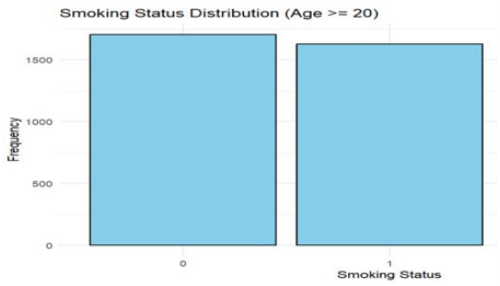


Figure.4.

```
## BAR GRAPH TO VIEW THE SMOKING STATUS DISTRIBUTION IN STROKE  POSITIVE PATIENTS
```

```{r}
# Assuming your columns are named "smoking_status" and "stroke"
library(ggplot2)

ggplot(subset(stroke.data, stroke == 1), aes(x = factor(smoking_status))) +
  geom_bar(fill = "coral", color = "black") +
  labs(title = "Smoking Status for Stroke-Positive Patients",
       x = "Smoking Status",
       y = "Frequency") +
  theme_minimal()
```

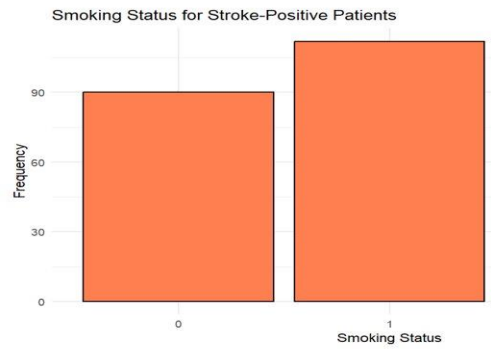
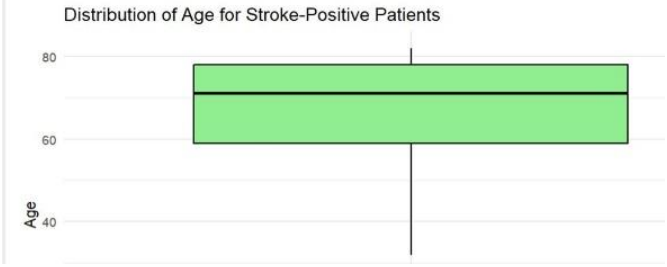


Figure.5.

```
## BOXPLOT TO VIEW THE AGE DISTRIBUTION IN STROKE POSITIVE PATIENTS
```

```{r}
# Assuming your columns are named "age" and "stroke"
library(ggplot2)

ggplot(subset(stroke.data, stroke == 1), aes(x = factor(stroke), y = age)) +
  geom_boxplot(fill = "lightgreen", color = "black") +
  labs(title = "Distribution of Age for Stroke-Positive Patients",
       x = "Stroke",
       y = "Age") +
  theme_minimal()
```



```r
library(ggplot2)
ggplot(stroke.data, aes(x = factor(gender))) +
  geom_bar(fill = "skyblue", color = "black") +
  labs(title = "Distribution of Gender", x = "Gender", y = "Frequency") +
  theme_minimal()
```

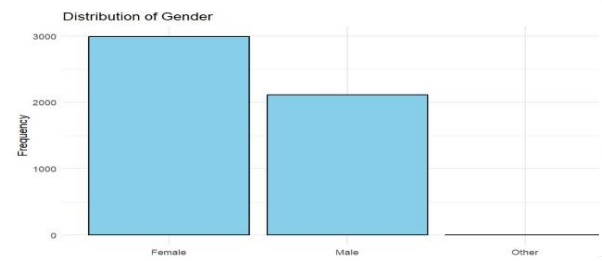


Figure.6.

```r
library(ggplot2)
ggplot(subset(stroke.data, stroke == 1), aes(x = factor(stroke), fill = gender)) +
  geom_bar(position = "dodge", color = "black") +
  labs(title = "Distribution of Gender for Stroke-Positive Patients",
       x = "Stroke",
       y = "Frequency",
       fill = "Gender") +
  theme_minimal()
```

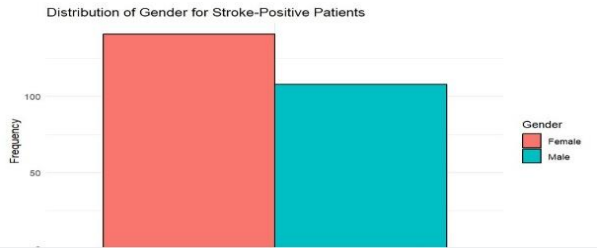


Figure.7.

Figure.8.

```r
set.seed(123)  # Setting a seed for reproducibility
sample_size <- min(5000, nrow(stroke.data))
sample_data <- stroke.data[sample(1:nrow(stroke.data), sample_size), ]
# Performing Shapiro-Wilk test
shapiro_test_result <- shapiro.test(sample_data$bmi)
# Printing the result
print(shapiro_test_result)
# Checking if the p-value is less than 0.05
if (shapiro_test_result$p.value < 0.05) {
  cat("The BMI data is not normally distributed based on the Shapiro-Wilk test.\n")
} else {
  cat("The BMI data appears to be normally distributed based on the Shapiro-Wilk test.\n")
}
```

```
        Shapiro-Wilk normality test

data:  sample_data$bmi
W = 0.9502, p-value < 2.2e-16

The BMI data is not normally distributed based on the Shapiro-Wilk test.
```

```r
# Assuming your data frame is named stroke.data and "BMI" and "stroke" are the columns of interest
wilcoxon_test_result <- wilcox.test(bmi ~ stroke, data = stroke.data)

# Print the result
print(wilcoxon_test_result)

# Check if the p-value is less than 0.05
if (wilcoxon_test_result$p.value < 0.05) {
  cat("There is a significant difference in BMI between stroke-affected and non-affected individuals.\n")
} else {
  cat("There is no significant difference in BMI between stroke-affected and non-affected individuals.\n")
}
```

```
        Wilcoxon rank sum test with continuity correction

data:  bmi by stroke
W = 522643, p-value = 0.0002769
alternative hypothesis: true location shift is not equal to 0

There is a significant difference in BMI between stroke-affected and non-affected individuals.
```

Figure.9.

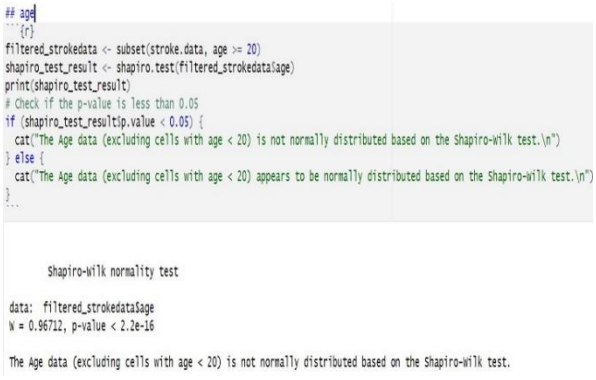
```r
# Random Sampling
set.seed(123)  # Setting a seed for reproducibility
sample_size <- min(5000, nrow(stroke.data))
sample_data <- stroke.data[sample(1:nrow(stroke.data), sample_size), ]
# Excluding samples with age less than 20
sample_data_filtered <- sample_data[sample_data$age >= 20, ]
# Creating a contingency table
cross_table <- table(sample_data_filtered$smoking_status,
sample_data_filtered$stroke)
# Performing Chi-squared test
chi_square_test <- chisq.test(cross_table)
# Printing the result
#print(cross_table)
print(chi_square_test)
# Make an inference based on the p-value
if (chi_square_test$p.value < 0.05) {
  cat("There is a significant association between smoking status and stroke
risk.\n")
} else {
  cat("There is no significant association between smoking status and
stroke risk.\n")
}
```

```
        Pearson's Chi-squared test with Yates' continuity correction

data:  cross_table
X-squared = 3.1085, df = 1, p-value = 0.07788

There is no significant association between smoking status and stroke
risk.
```

Figure.10

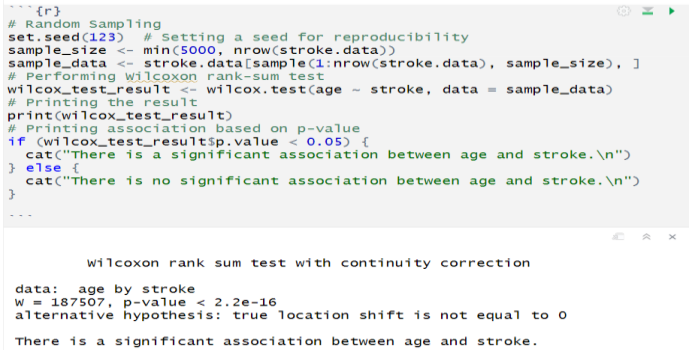


```r
## age
```{r}
filtered_strokedata <- subset(stroke.data, age >= 20)
shapiro_test_result <- shapiro.test(filtered_strokedata$age)
print(shapiro_test_result)
# Check if the p-value is less than 0.05
if (shapiro_test_result$p.value < 0.05) {
  cat("The Age data (excluding cells with age < 20) is not normally distributed based on the Shapiro-Wilk test.\n")
} else {
  cat("The Age data (excluding cells with age < 20) appears to be normally distributed based on the Shapiro-Wilk test.\n")
}
```
```

```
        Shapiro-Wilk normality test

data:  filtered_strokedata$age
W = 0.96712, p-value < 2.2e-16

The Age data (excluding cells with age < 20) is not normally distributed based on the Shapiro-Wilk test.
```



```r
```{r}
# Random Sampling
set.seed(123)  # Setting a seed for reproducibility
sample_size <- min(5000, nrow(stroke.data))
sample_data <- stroke.data[sample(1:nrow(stroke.data), sample_size), ]
# Performing Wilcoxon rank-sum test
wilcox_test_result <- wilcox.test(age ~ stroke, data = sample_data)
# Printing the result
print(wilcox_test_result)
# Printing association based on p-value
if (wilcox_test_result$p.value < 0.05) {
  cat("There is a significant association between age and stroke.\n")
} else {
  cat("There is no significant association between age and stroke.\n")
}
```
```

```
        Wilcoxon rank sum test with continuity correction

data:  age by stroke
W = 187507, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0

There is a significant association between age and stroke.
```

Figure.11.

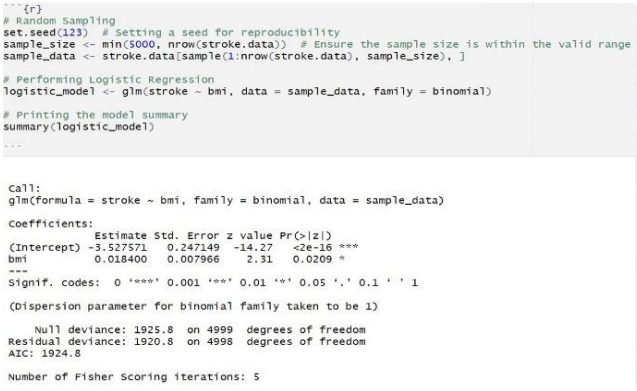
```r
## gender
```{r}
set.seed(123)  # Setting a seed for reproducibility
sample_size <- min(5000, nrow(stroke.data))
sample_data <- stroke.data[sample(1:nrow(stroke.data), sample_size), ]
contingency_table <- table(sample_data$gender, sample_data$stroke)
chi_square_test <- chisq.test(contingency_table)
print(chi_square_test)
if (chi_square_test$p.value < 0.05) {
  cat("There is a significant association between gender and stroke risk.\n")
} else {
  cat("There is no significant association between gender and stroke risk.\n")
}
```
```

```
Warning: Chi-squared approximation may be incorrect
        Pearson's Chi-squared test

data:  contingency_table
X-squared = 0.51158, df = 2, p-value = 0.7743

There is no significant association between gender and stroke risk.
```

Figure.12. (Logistic Regression)



```r
```{r}
# Random Sampling
set.seed(123)  # Setting a seed for reproducibility
sample_size <- min(5000, nrow(stroke.data))  # Ensure the sample size is within the valid range
sample_data <- stroke.data[sample(1:nrow(stroke.data), sample_size), ]

# Performing Logistic Regression
logistic_model <- glm(stroke ~ bmi, data = sample_data, family = binomial)

# Printing the model summary
summary(logistic_model)
```
```

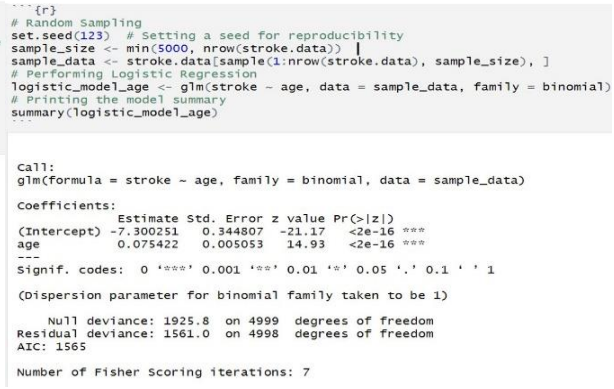
```
Call:
glm(formula = stroke ~ bmi, family = binomial, data = sample_data)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.527571   0.247149  -14.27   <2e-16 ***
bmi          0.018400   0.007966    2.31   0.0209 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1925.8  on 4999  degrees of freedom
Residual deviance: 1920.8  on 4998  degrees of freedom
AIC: 1924.8

Number of Fisher Scoring iterations: 5
```



```r
```{r}
# Random Sampling
set.seed(123)  # Setting a seed for reproducibility
sample_size <- min(5000, nrow(stroke.data))
sample_data <- stroke.data[sample(1:nrow(stroke.data), sample_size), ]
# Performing Logistic Regression
logistic_model_age <- glm(stroke ~ age, data = sample_data, family = binomial)
# Printing the model summary
summary(logistic_model_age)
```
```

```
Call:
glm(formula = stroke ~ age, family = binomial, data = sample_data)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -7.300251   0.344807  -21.17   <2e-16 ***
age          0.075422   0.005053   14.93   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1925.8  on 4999  degrees of freedom
Residual deviance: 1561.0  on 4998  degrees of freedom
AIC: 1565

Number of Fisher Scoring iterations: 7
```

```{r}
# Random Sampling
set.seed(123)  # Setting a seed for reproducibility
sample_size <- min(5000, nrow(stroke.data))
sample_data <- stroke.data[sample(1:nrow(stroke.data), sample_size), ]
# Performing Logistic Regression
logistic_model_gender <- glm(stroke ~ gender, data = sample_data, family = binomial)
# Printing the model summary
summary(logistic_model_gender)
```

```
Call:
glm(formula = stroke ~ gender, family = binomial, data = sample_data)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.02543    0.08781 -34.456   <2e-16 ***
genderMale    0.09066    0.13355   0.679    0.497
genderOther  -9.54063  324.74371  -0.029    0.977
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1925.8  on 4999  degrees of freedom
Residual deviance: 1925.3  on 4997  degrees of freedom
AIC: 1931.3

Number of Fisher Scoring iterations: 11
```

```{r}
# Random Sampling
set.seed(123)  # Setting a seed for reproducibility
sample_size <- min(5000, nrow(stroke.data))
sample_data <- stroke.data[sample(1:nrow(stroke.data), sample_size), ]
# Performing Logistic Regression
logistic_model <- glm(stroke ~ smoking_status, data = sample_data, family = binomial)
# Printing the model summary
summary(logistic_model)
```

```
Call:
glm(formula = stroke ~ smoking_status, family = binomial, data = sample_data)

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     -3.0048     0.1092 -27.516   <2e-16 ***
smoking_status   0.3498     0.1481   2.362   0.0182 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1504.3  on 3492  degrees of freedom
Residual deviance: 1498.6  on 3491  degrees of freedom
  (1507 observations deleted due to missingness)
AIC: 1502.6

Number of Fisher Scoring iterations: 5
```