stroke_analysis

Iman Taheri

2025-11-24

Stroke Risk Analysis in R

Google Data Analytics Capstone • Kaggle Stroke Dataset

This analysis uses a real-world stroke dataset to explore patterns, risk indicators, and demographic trends associated with stroke. The workflow follows the full Google Data Analytics framework: *Ask* → *Prepare* → *Process* → *Analyze* → *Share* → *Act*, implemented entirely in R.

Project Objectives

```
* Identify factors most strongly associated with stroke.
* Examine demographic and lifestyle patterns.
* Build clear, reproducible visualizations.
* Develop insights that can support early risk detection.
```

Required Packages

```r
library("tidyverse")
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.6
## v forcats   1.0.1     v stringr   1.6.0
## v ggplot2   4.0.1     v tibble    3.3.0
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.2.0
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

```r
library("ggplot2")
library("janitor")
```

```
##
## Attaching package: 'janitor'
##
## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

```r
library("readr")
library("skimr")
```

Dataset Importing and Overview

```r
stroke_df <- read_csv("data/stroke_data.csv")
```

```
## Rows: 5110 Columns: 12
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (6): gender, ever_married, work_type, Residence_type, bmi, smoking_status
## dbl (6): id, age, hypertension, heart_disease, avg_glucose_level, stroke
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
skim_without_charts(stroke_df)
```

Table 1: Data summary

Name	stroke_df
Number of rows	5110
Number of columns	12
Column type frequency:	
character	6
numeric	6
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
gender	0	1	4	6	0	3	0
ever_married	0	1	2	3	0	2	0
work_type	0	1	7	13	0	5	0
Residence_type	0	1	5	5	0	2	0
bmi	0	1	2	4	0	419	0
smoking_status	0	1	6	15	0	4	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
id	0	1	36517.83	21161.72	67.00	17741.25	36932.00	54682.00	72940.00
age	0	1	43.23	22.61	0.08	25.00	45.00	61.00	82.00
hypertension	0	1	0.10	0.30	0.00	0.00	0.00	0.00	1.00

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
heart_disease	0	1	0.05	0.23	0.00	0.00	0.00	0.00	1.00
avg_glucose_level	0	1	106.15	45.28	55.12	77.24	91.88	114.09	271.74
stroke	0	1	0.05	0.22	0.00	0.00	0.00	0.00	1.00

Data cleaning

To ensure analytical reliability and prepare the dataset for modeling, several preprocessing steps were performed:

1.Standardized column names

To maintain consistency and avoid case-sensitivity issues, all variable names were converted to snake_case using clean_names() from the janitor package:

```
clean_names(stroke_df)
```

```
## # A tibble: 5,110 x 12
##       id gender   age hypertension heart_disease ever_married work_type
##    <dbl> <chr>  <dbl>        <dbl>         <dbl> <chr>        <chr>
##  1  9046 Male      67            0             1 Yes          Private
##  2 51676 Female    61            0             0 Yes          Self-employed
##  3 31112 Male      80            0             1 Yes          Private
##  4 60182 Female    49            0             0 Yes          Private
##  5  1665 Female    79            1             0 Yes          Self-employed
##  6 56669 Male      81            0             0 Yes          Private
##  7 53882 Male      74            1             1 Yes          Private
##  8 10434 Female    69            0             0 No           Private
##  9 27419 Female    59            0             0 Yes          Private
## 10 60491 Female    78            0             0 Yes          Private
## # i 5,100 more rows
## # i 5 more variables: residence_type <chr>, avg_glucose_level <dbl>, bmi <chr>,
## #   smoking_status <chr>, stroke <dbl>
```

2.Harmonized categorical labels

The dataset contained a categorical variable (residence_type) with inconsistent capitalization. It was renamed for clarity:

```
stroke_df <- stroke_df %>% rename(residence_type=Residence_type)
```

3.Removed rare and uninformative categories

The gender column included a category labeled “Other”, which contained only one observation. To avoid unstable statistical estimates, the analysis was restricted to male and female patients

```
stroke_df <- stroke_df %>% filter(gender %in% c("Male", "Female"))
```

4.Converted BMI to numeric and Removed incompelte BMI records

The bmi feature was stored as character. It was converted to numeric to support correlation analysis and modeling and rows with invalid or missing BMI values (introduced during conversion) were removed:

```
stroke_df$bmi <- as.numeric(stroke_df$bmi)
```

```
## Warning: NAs introduced by coercion
```

```
stroke_df <- subset(stroke_df, complete.cases(bmi))
```

Result

After cleaning, the dataset was fully structured, free of missing critical values, and ready for exploratory data analysis and predictive modeling.

```
colSums(is.na(stroke_df))
```

```
##                id            gender               age      hypertension
##                 0                 0                 0                 0
##     heart_disease      ever_married         work_type    residence_type
##                 0                 0                 0                 0
## avg_glucose_level               bmi    smoking_status            stroke
##                 0                 0                 0                 0
```

Analysis

This section examines how demographic, behavioral, and clinical factors relate to stroke occurrence. Both numerical summaries and visualizations are included to generate insights and identify meaningful patterns in the data.

1.Gender and Stroke

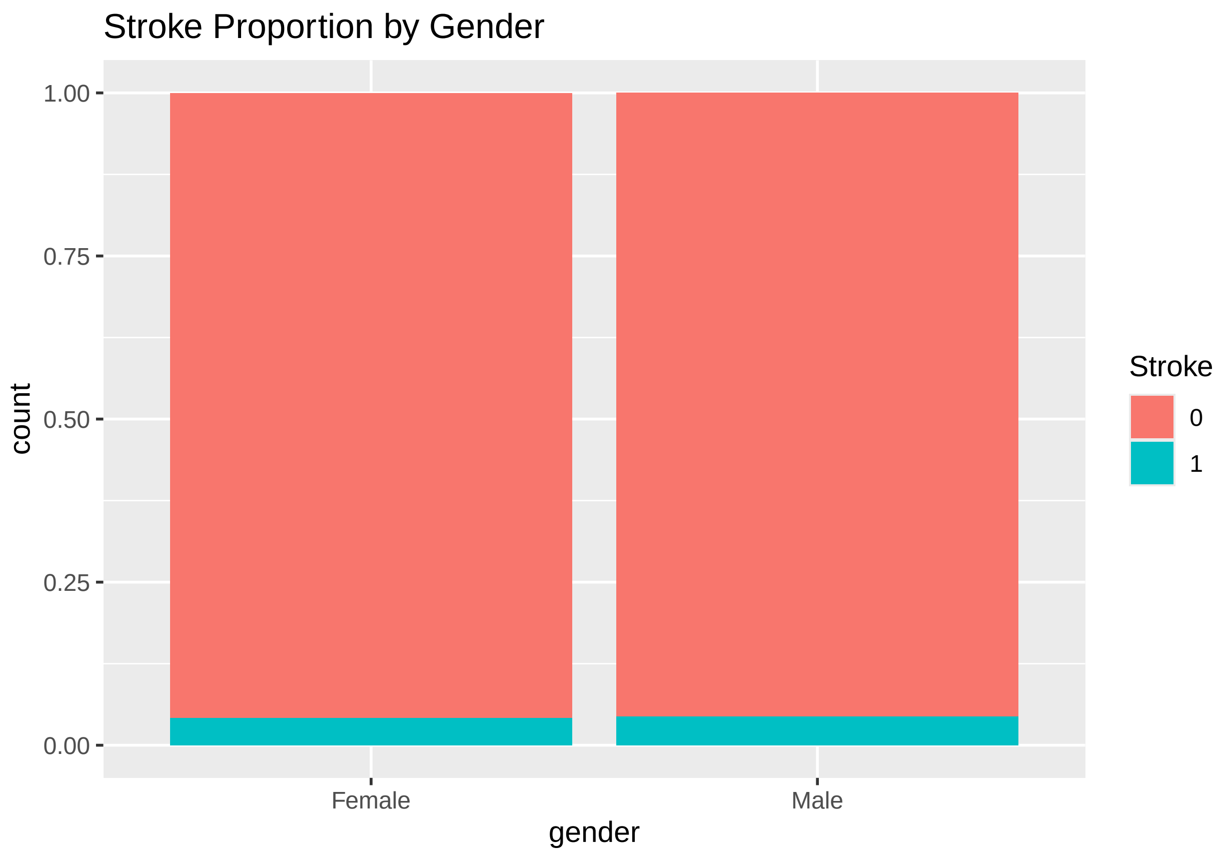
- Numerical Summary

```
stroke_df %>% group_by(gender) %>% summarize(sum(stroke == 1), sum(stroke == 0))
```

```
## # A tibble: 2 x 3
##   gender `sum(stroke == 1)` `sum(stroke == 0)`
##   <chr>               <int>              <int>
## 1 Female                120               2777
## 2 Male                   89               1922
```

- Visualization

```
ggplot(stroke_df, aes(gender, fill = factor(stroke))) +
  geom_bar(position = "fill") +
  labs(
    title = "Stroke Proportion by Gender",
    fill = "Stroke"
  )
```



- Correlation

```
table_gender <- table(stroke_df$gender, stroke_df$stroke)
chisq.test(table_gender)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table_gender
## X-squared = 0.16955, df = 1, p-value = 0.6805
```

- Observation: The proportion of stroke events is similar across males and females.
- Statistics: Chi-square test p = 0.68 → no significant association.
- Interpretation: Gender does not appear to influence stroke risk in this dataset.
- Practical Meaning: Gender-based differences in stroke risk may be negligible for this population.

2.Hypertension and Stroke

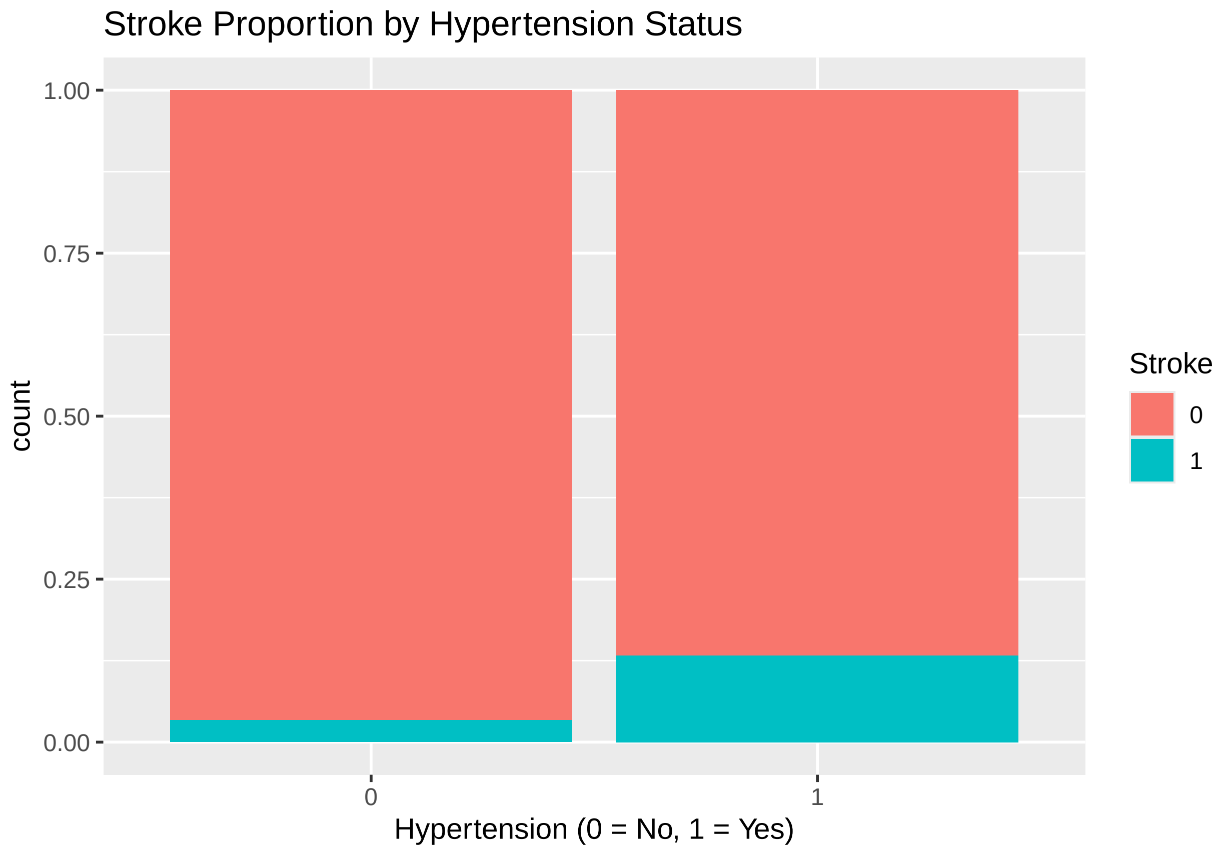
- Numerical Summary

```
stroke_df %>% group_by(hypertension) %>% summarize(sum(stroke == 1), sum(stroke == 0))
```

```
## # A tibble: 2 x 3
##   hypertension `sum(stroke == 1)` `sum(stroke == 0)`
##          <dbl>               <int>               <int>
## 1            0                 149                4308
## 2            1                  60                 391
```

- Visualization

```
ggplot(stroke_df, aes(factor(hypertension), fill = factor(stroke))) +
  geom_bar(position = "fill") +
  labs(
    title = "Stroke Proportion by Hypertension Status",
    x = "Hypertension (0 = No, 1 = Yes)",
    fill = "Stroke"
  )
```



- Correlation

```
table_hypertension <- table(stroke_df$hypertension, stroke_df$stroke)
chisq.test(table_hypertension)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table_hypertension
## X-squared = 97.239, df = 1, p-value < 2.2e-16
```

- Observation: Individuals with hypertension show a visibly higher stroke rate.
- Statistics: Chi-square test shows a significant association ($p < 0.001$).
- Interpretation: Hypertension is a major determinant of stroke occurrence.
- Practical Meaning: This reinforces established clinical evidence that high blood pressure is a leading modifiable risk factor.

3.Heart Disease and Stroke

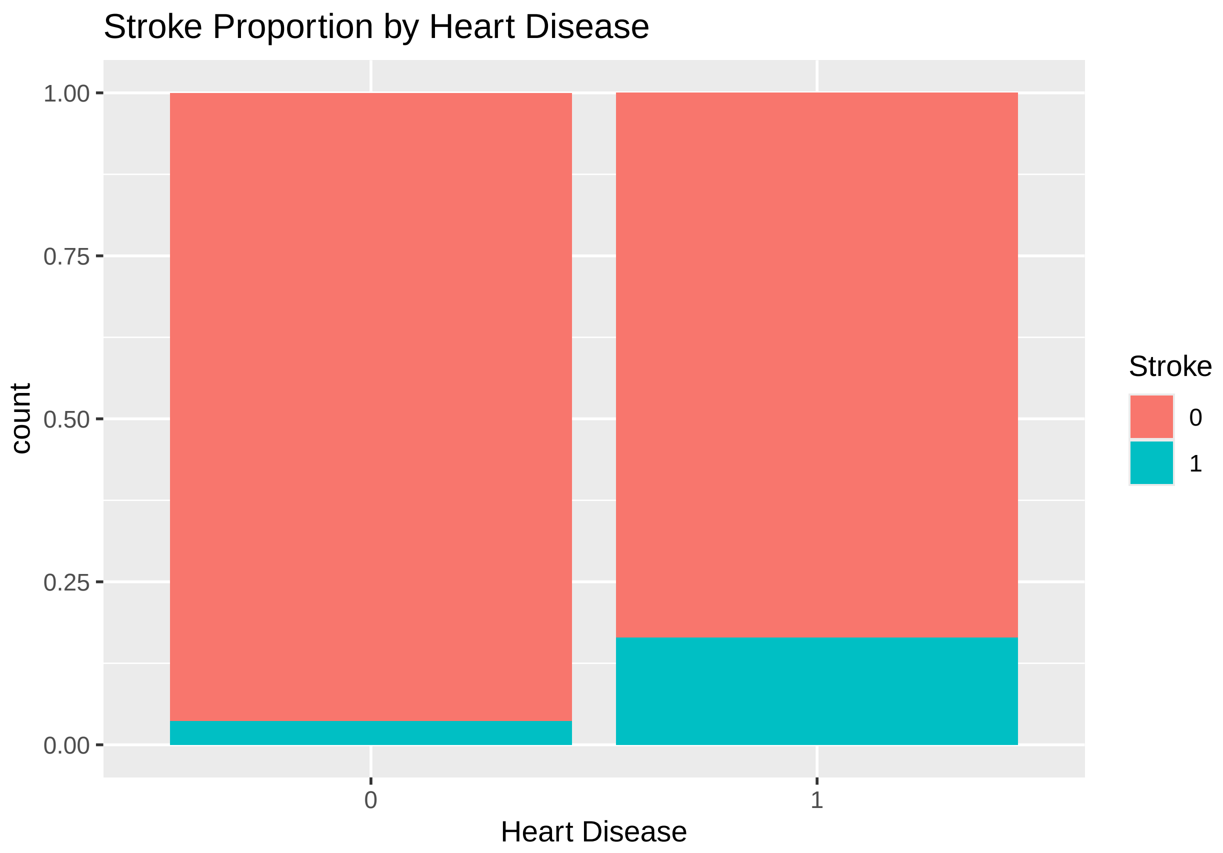
- Numerical Summary

```
stroke_df %>% group_by(heart_disease) %>% summarize(sum(stroke == 1), sum(stroke == 0))
```

```
## # A tibble: 2 x 3
##   heart_disease `sum(stroke == 1)` `sum(stroke == 0)`
##           <dbl>              <int>              <int>
## 1             0                169               4496
## 2             1                 40                203
```

- Visualization

```
ggplot(stroke_df, aes(factor(heart_disease), fill = factor(stroke))) +
  geom_bar(position = "fill") +
  labs(
    title = "Stroke Proportion by Heart Disease",
    x = "Heart Disease",
    fill = "Stroke"
  )
```



- Correlation

```
table_heartdx <- table(stroke_df$heart_disease, stroke_df$stroke)
chisq.test(table_heartdx)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table_heartdx
## X-squared = 90.25, df = 1, p-value < 2.2e-16
```

- Observation: Stroke is more common among individuals with existing heart disease.
- Statistics: Chi-square test confirms a significant relationship.
- Interpretation: Cardiovascular comorbidity contributes strongly to stroke risk.
- Practical Meaning: Patients with cardiac issues may require more aggressive stroke prevention strategies.

4.Smoking Status and Stroke

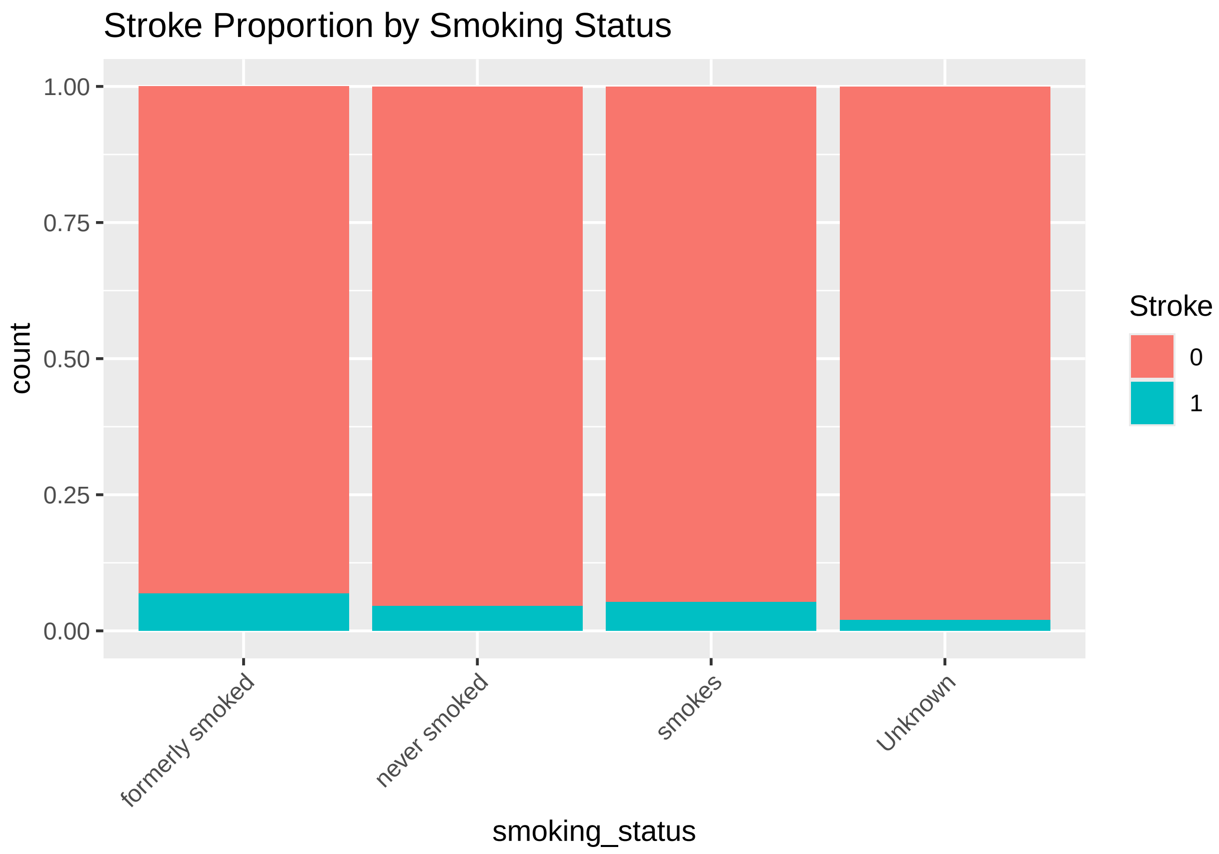
- Numerical Summary

```
stroke_df %>% group_by(smoking_status) %>% summarize(sum(stroke == 1), sum(stroke == 0))
```

```
## # A tibble: 4 x 3
##   smoking_status  `sum(stroke == 1)` `sum(stroke == 0)`
##   <chr>                        <int>              <int>
## 1 Unknown                         29               1454
## 2 formerly smoked                 57                779
## 3 never smoked                    84               1768
## 4 smokes                          39                698
```

- Visualization

```
ggplot(stroke_df, aes(smoking_status, fill = factor(stroke))) +
  geom_bar(position = "fill") +
  labs(
    title = "Stroke Proportion by Smoking Status",
    fill = "Stroke"
  ) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



- Correlation

```
table_smoking <- table(stroke_df$smoking_status, stroke_df$stroke)
chisq.test(table_smoking)
```

```
##
```

```
##  Pearson's Chi-squared test
##
## data:  table_smoking
## X-squared = 35.006, df = 3, p-value = 1.215e-07
```

- Observation: Stroke prevalence varies substantially across smoking categories.
- Statistics: Chi-square test: ^{2} = 35.0, p 1e-7 (highly significant).
- Interpretation: Smoking status is one of the strongest predictors in the dataset.
- Practical Meaning: Both current and former smokers exhibit greater stroke risk, aligned with known pathophysiology.

5.Residence Type and Stroke

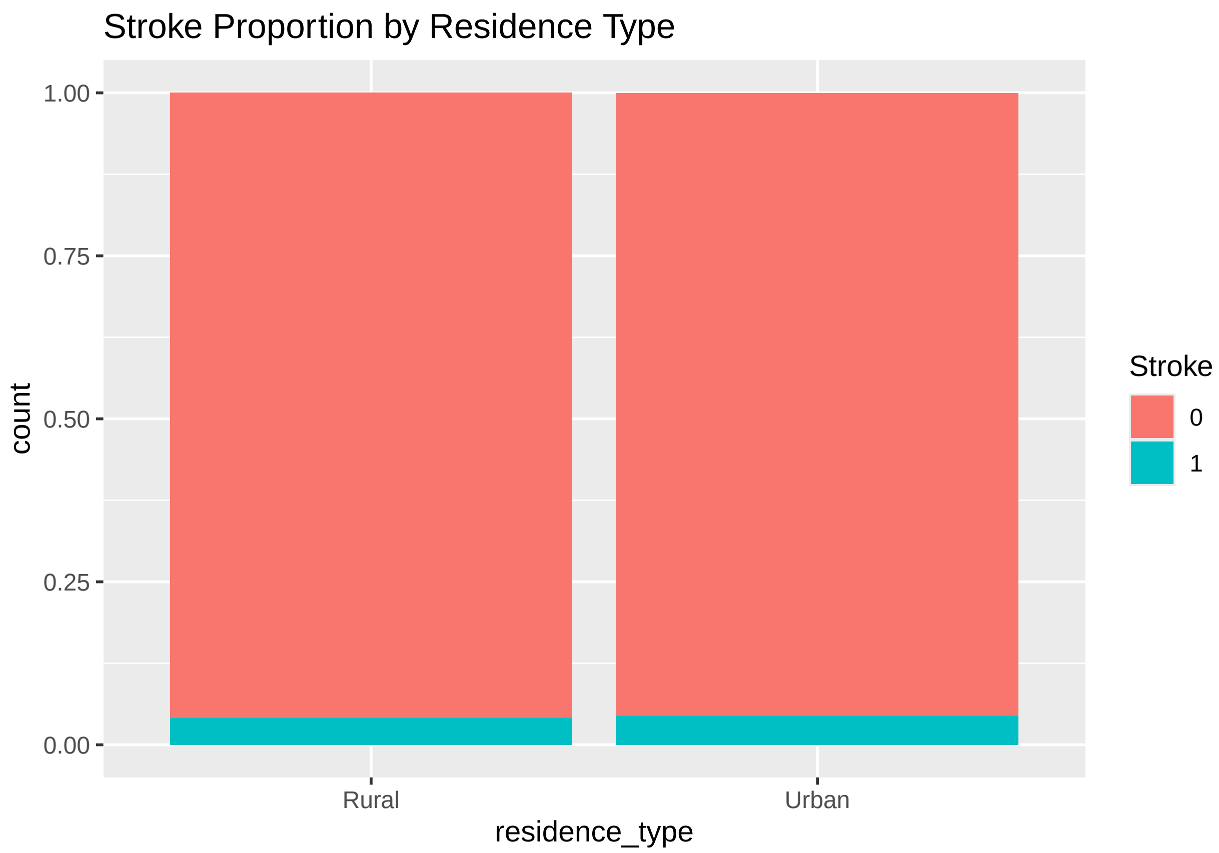
- Numerical Summary

```
stroke_df %>% group_by(residence_type) %>% summarize(sum(stroke == 1), sum(stroke == 0))
```

```
## # A tibble: 2 x 3
##   residence_type `sum(stroke == 1)` `sum(stroke == 0)`
##   <chr>                       <int>              <int>
## 1 Rural                         100               2318
## 2 Urban                         109               2381
```

- Visualization

```
ggplot(stroke_df, aes(residence_type, fill = factor(stroke))) +
  geom_bar(position = "fill") +
  labs(
    title = "Stroke Proportion by Residence Type",
    fill = "Stroke"
  )
```



- Correlation

```
table_residence <- table(stroke_df$residence_type, stroke_df$stroke)
chisq.test(table_residence)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table_residence
## X-squared = 0.12169, df = 1, p-value = 0.7272
```

- Observation: Stroke proportions are similar between urban and rural residents.
- Statistics: Chi-square suggests no significant relationship.
- Interpretation: Living environment does not appear to affect stroke risk within this dataset.
- Practical Meaning: Lifestyle or access to care differences may not be reflected here.

6.Work Type and Stroke

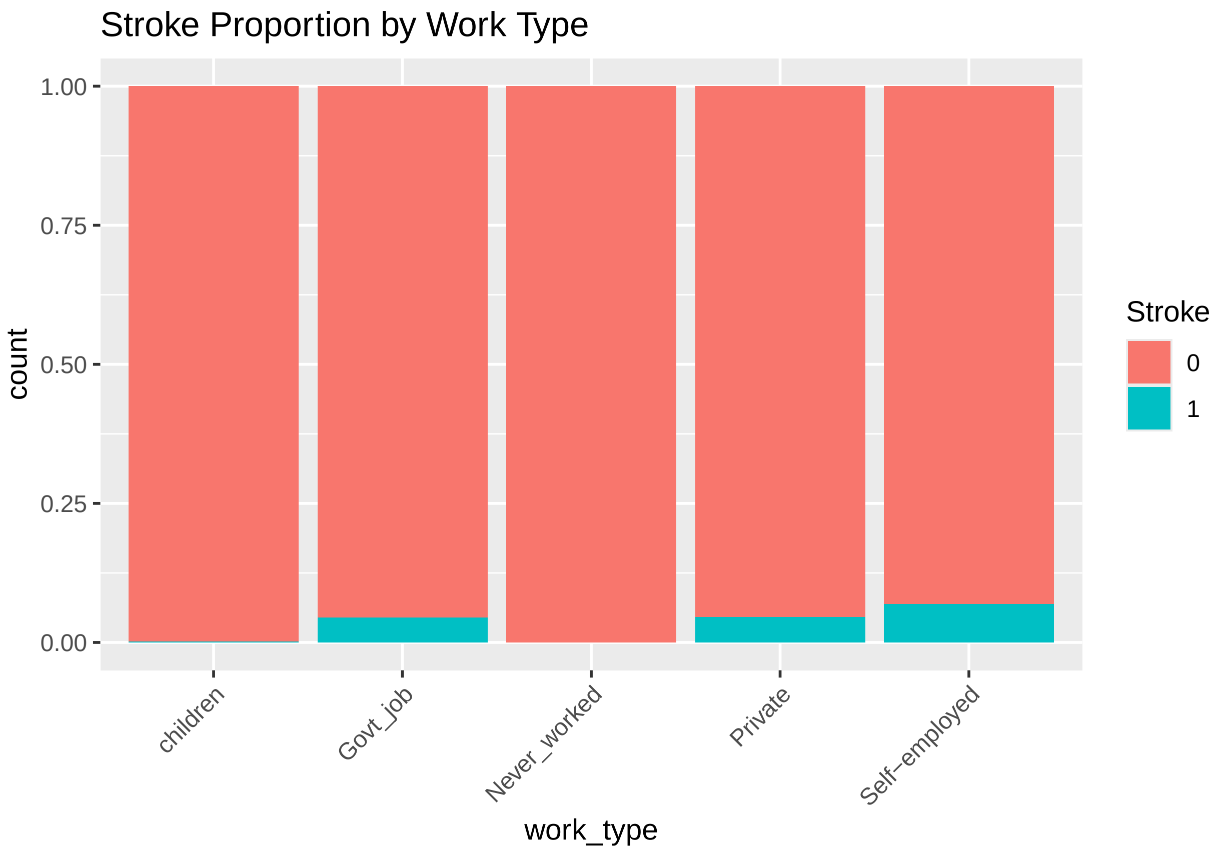
- Numerical Summary

```
stroke_df %>% group_by(work_type) %>% summarize(sum(stroke == 1), sum(stroke == 0))
```

```
## # A tibble: 5 x 3
##   work_type     `sum(stroke == 1)` `sum(stroke == 0)`
##   <chr>                      <int>              <int>
## 1 Govt_job                      28                602
## 2 Never_worked                   0                 22
## 3 Private                      127               2683
## 4 Self-employed                 53                722
## 5 children                       1                670
```

- Visualization

```
ggplot(stroke_df, aes(work_type, fill = factor(stroke))) +
  geom_bar(position = "fill") +
  labs(
    title = "Stroke Proportion by Work Type",
    fill = "Stroke"
  ) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



- Correlation

```
table_work <- table(stroke_df$work_type, stroke_df$stroke)
chisq.test(table_work)
```

```
## Warning in stats::chisq.test(x, y, ...): Chi-squared approximation may be
## incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  table_work
## X-squared = 41.951, df = 4, p-value = 1.708e-08
```

- Observation: Stroke rates vary across work categories, with self-employed and government workers showing slightly higher proportions.
- Statistics: Chi-square test shows a significant association ($p < 0.001$).
- Interpretation: Occupational lifestyle or socioeconomic differences may influence stroke risk.
- Practical Meaning: Work type may act as a proxy for stress, activity level, or chronic exposure factors.

7.Ever Married and Stroke

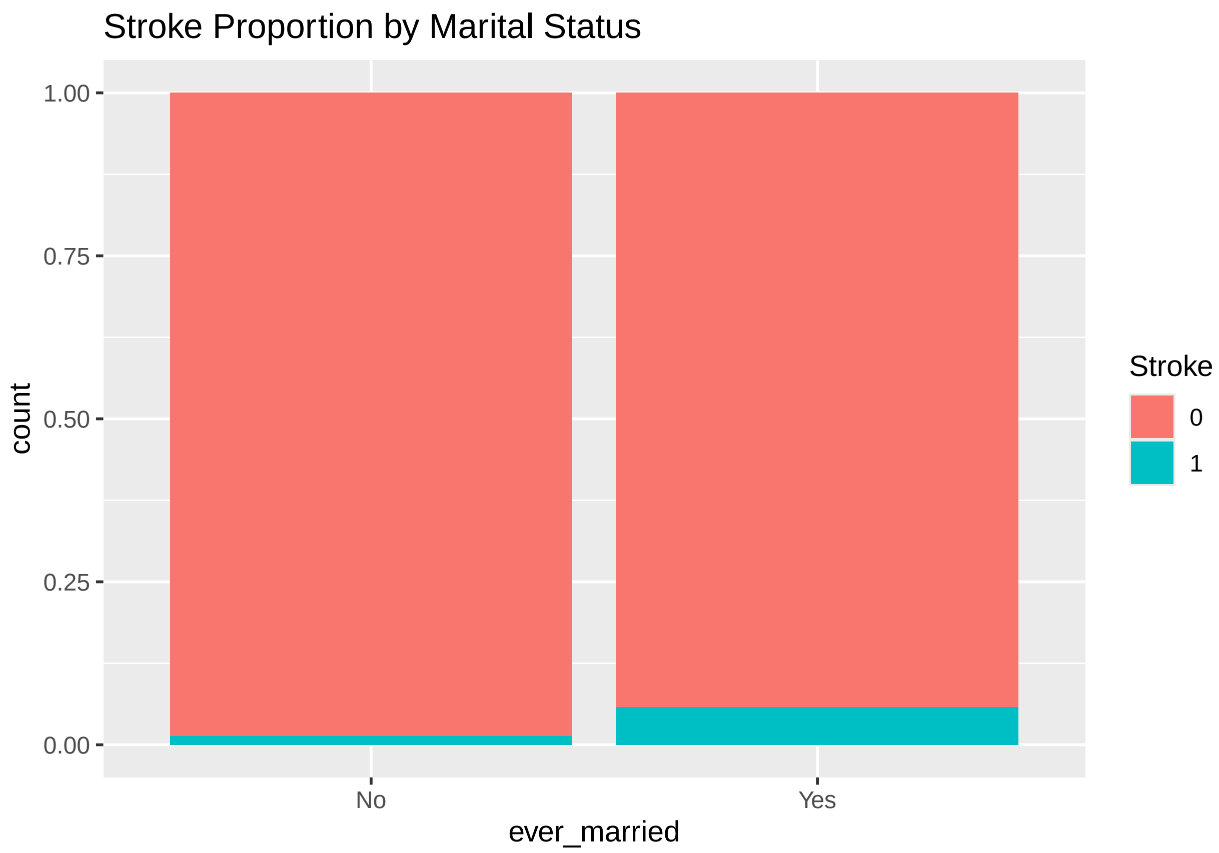
- Numerical Summary

```
stroke_df %>% group_by(ever_married) %>% summarize(sum(stroke == 1), sum(stroke == 0))
```

```
## # A tibble: 2 x 3
##   ever_married `sum(stroke == 1)` `sum(stroke == 0)`
##   <chr>                     <int>              <int>
## 1 No                           23               1681
## 2 Yes                         186               3018
```

- Visulization

```
ggplot(stroke_df, aes(ever_married, fill = factor(stroke))) +
  geom_bar(position = "fill") +
  labs(
    title = "Stroke Proportion by Marital Status",
    fill = "Stroke"
  )
```



- Correlation

```
table_marriage <- table(stroke_df$ever_married, stroke_df$stroke)
chisq.test(table_marriage)
```

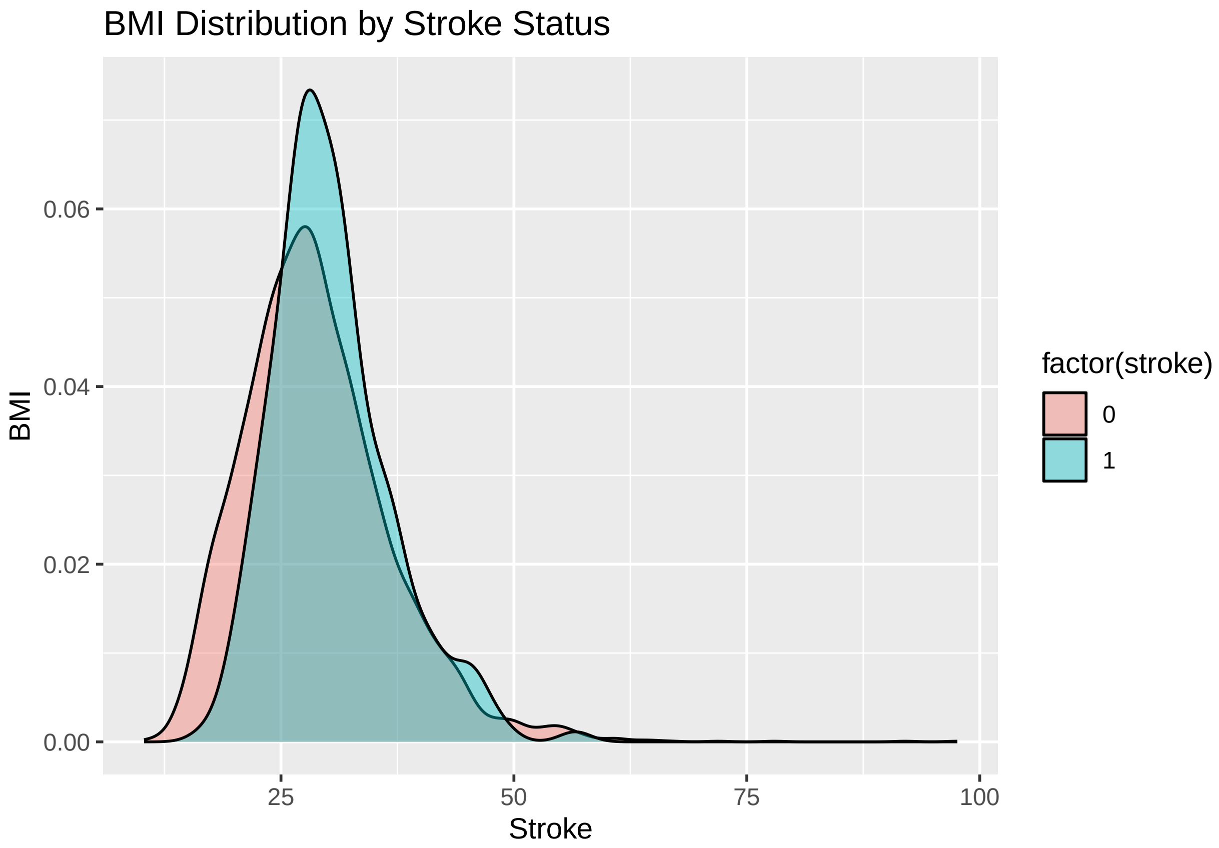
```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table_marriage
## X-squared = 53.076, df = 1, p-value = 3.209e-13
```

- Observation: Stroke proportions differ sharply between “Ever Married” vs “Never Married.”
- Statistics: ^{2} = 53.08, p = 3.2e–13 → highly significant association.
- Interpretation: Marital status is statistically linked with stroke risk.
- Practical Meaning: This may reflect confounding factors such as age — married individuals tend to be older, and age is a major driver of stroke risk. The association is strong but likely indirect.

8.BMI and Stroke

- Visualization

```
ggplot(stroke_df, aes(bmi, fill=factor(stroke))) +
  geom_density(alpha = 0.4) +
  labs(
    title = "BMI Distribution by Stroke Status",
    x = "Stroke",
    y = "BMI"
  )
```



- Correlation

```
cor(stroke_df$stroke, stroke_df$bmi)
```

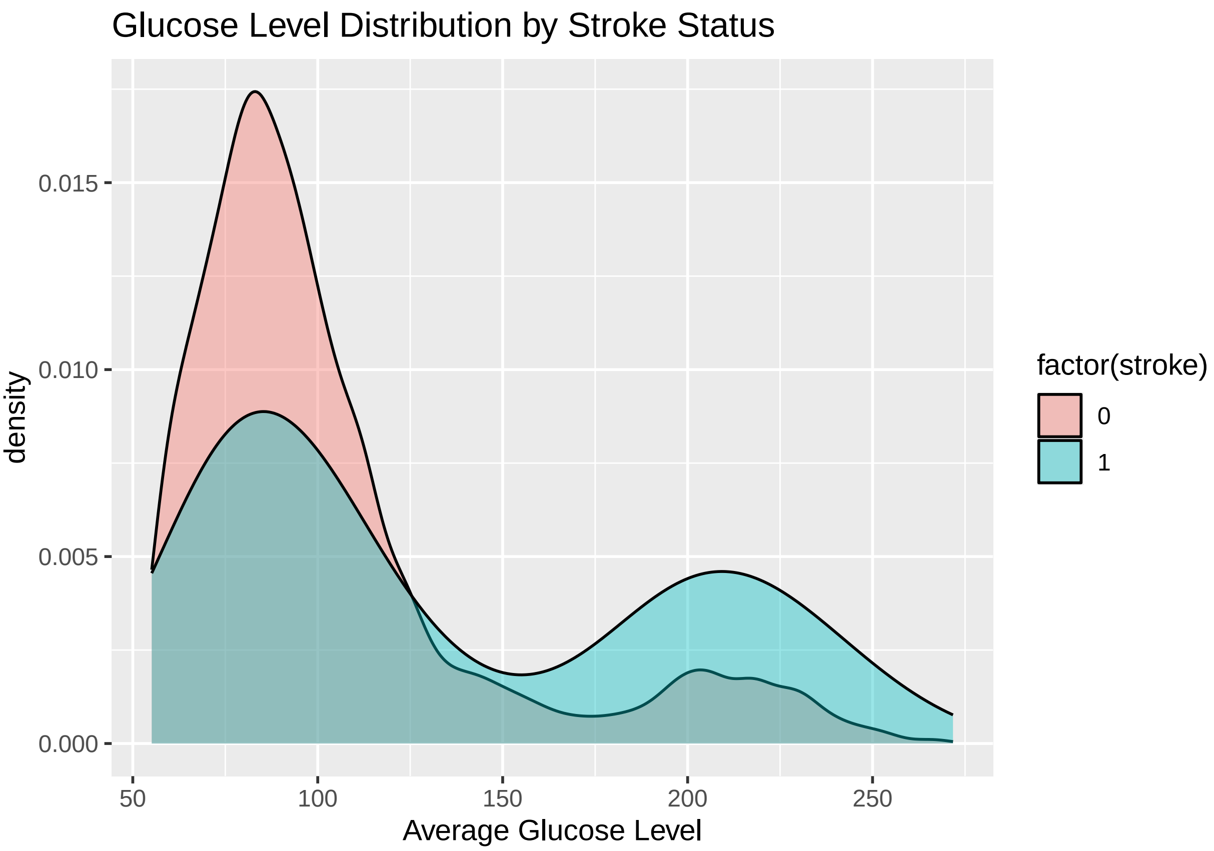
```
## [1] 0.04234128
```

- Observation: The correlation between BMI and stroke occurrence is extremely weak.
- Statistics: r = 0.042 → essentially no linear association.
- Interpretation: BMI does not meaningfully predict stroke risk in this dataset.
- Practical Meaning: Weight alone does not appear to indicate higher risk for stroke here; other factors may overshadow its impact.

9.Glucose Level and Stroke

- Visualization

```
ggplot(stroke_df, aes(avg_glucose_level, fill = factor(stroke))) +
  geom_density(alpha = 0.4) +
  labs(
    title = "Glucose Level Distribution by Stroke Status",
    x = "Average Glucose Level"
  )
```



- Correlation

```
cor(stroke_df$stroke, stroke_df$avg_glucose_level)
```

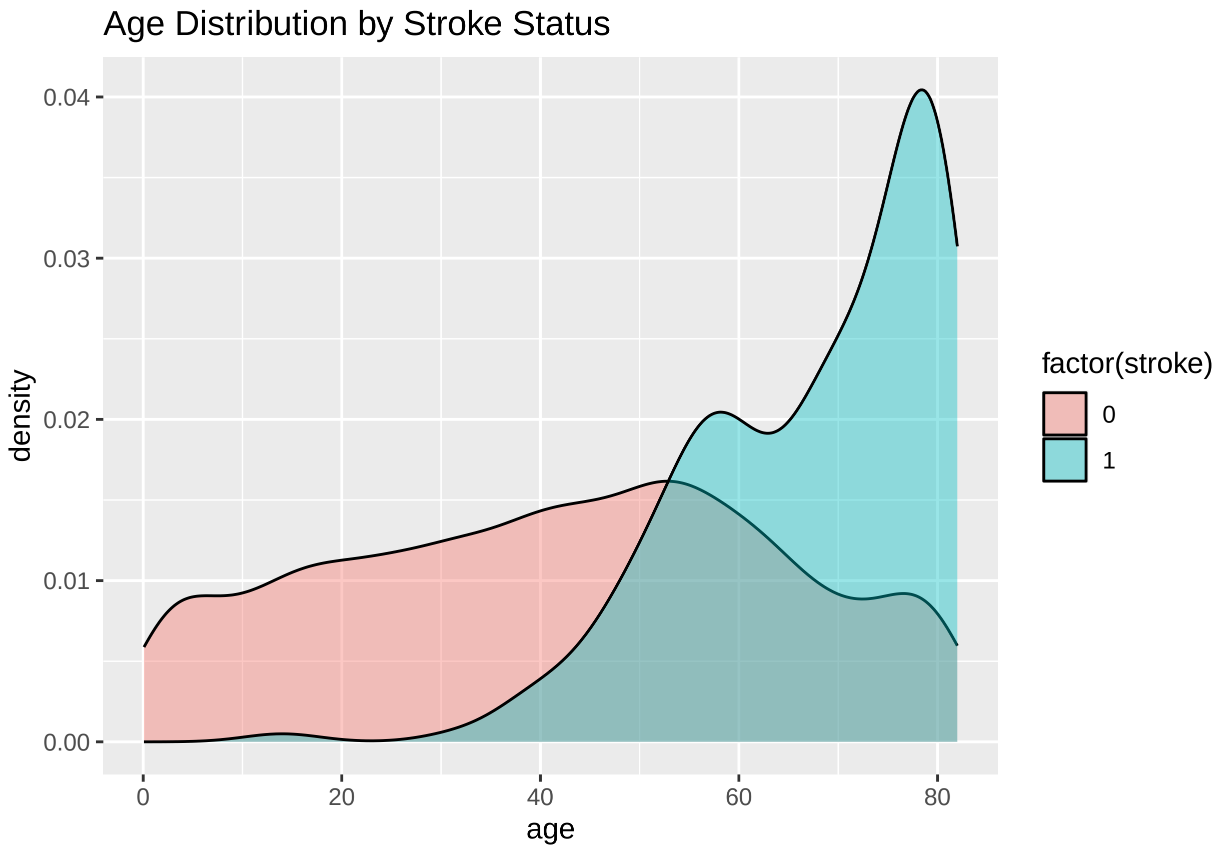
```
## [1] 0.1389836
```

- Observation: Stroke cases appear slightly more common among individuals with higher glucose levels.
- Statistics: r = 0.139 → weak positive correlation.
- Interpretation: Higher glucose may increase stroke risk, but the relationship is mild in this dataset.
- Practical Meaning: This points in the expected clinical direction (hyperglycemia → vascular risk), but the effect is small, likely needing more variables to strengthen prediction.

10.Age and Stroke

- Visualization

```r
ggplot(stroke_df, aes(age, fill = factor(stroke))) +
  geom_density(alpha = 0.4) +
  labs(title = "Age Distribution by Stroke Status")
```



- Correlation

```r
cor(stroke_df$stroke, stroke_df$age)
```

```
## [1] 0.232313
```

- Observation: Stroke cases increase noticeably with age.
- Statistics: r = 0.232 → moderate positive correlation.
- Interpretation: Age shows the strongest linear relationship among numeric predictors.
- Practical Meaning: Age is a clinically relevant risk factor and plays a meaningful role in stroke likelihood.

Overall Conclusion

This analysis demonstrates that stroke risk is strongly shaped by a combination of demographic, clinical, and lifestyle factors. Age emerged as the most influential continuous variable, while hypertension, heart disease, and smoking status showed clear categorical associations with stroke outcomes. These findings align with long-standing clinical evidence, reinforcing the critical role of vascular health and lifestyle behavior in cerebrovascular disease. Conversely, variables such as BMI, gender, and residence type contributed little explanatory value, suggesting they may play minor or context-dependent roles in this dataset. Overall, this project highlights how structured data analysis can convert raw patient information into meaningful clinical insights that support real-world decision-making.