Stroke Analysis

Vram Papyan, Leonid Sarkisyan, Elina Davtyan, Elvina Nosrati

2023-12-03

```
library(readxl)
library(plotly)
```

```
## Loading required package: ggplot2
```

```
##
## Attaching package: 'plotly'
```

```
## The following object is masked from 'package:ggplot2':
##
##     last_plot
```

```
## The following object is masked from 'package:stats':
##
##     filter
```

```
## The following object is masked from 'package:graphics':
##
##     layout
```

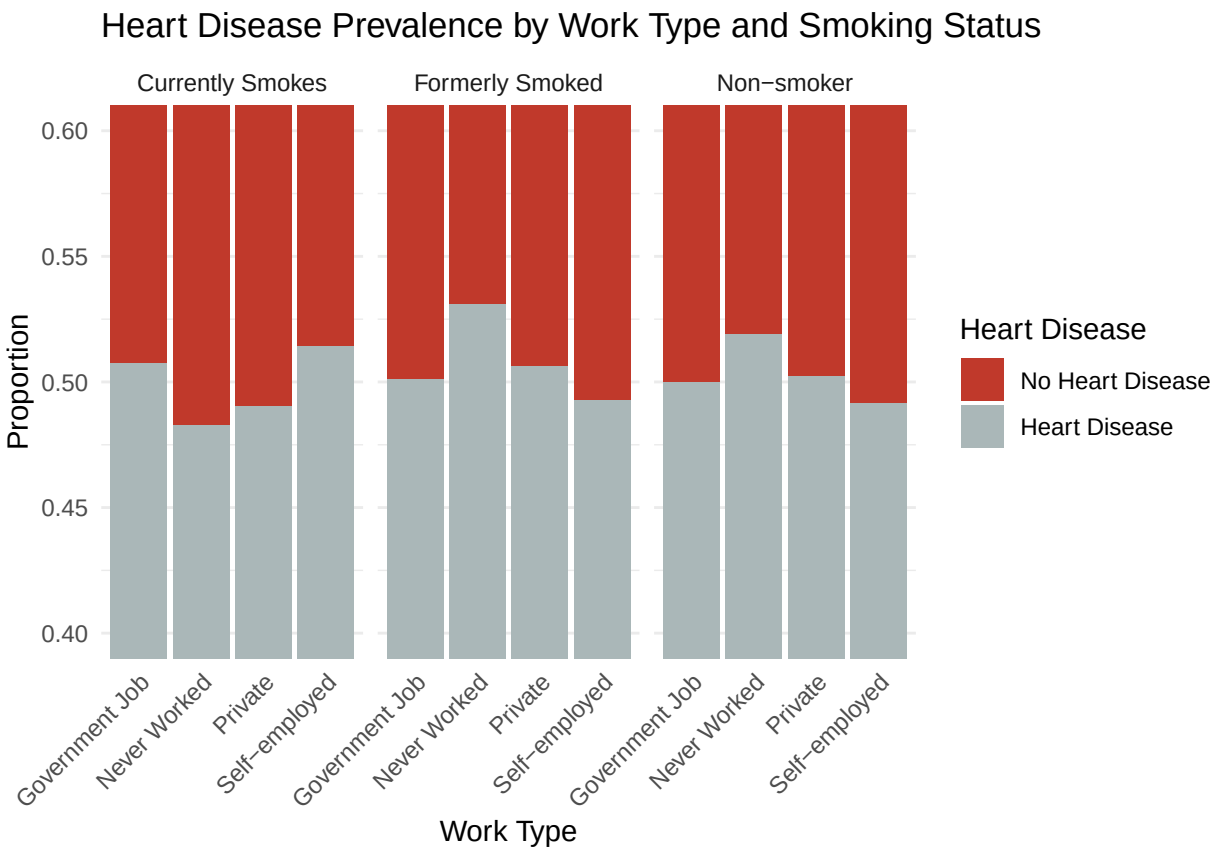
```
library(stringr)
library(viridis)
```

```
## Loading required package: viridisLite
```

```
data <- read_excel("/Users/vrampapyan/Desktop/Datavis project/dv-p-ls-v.1.1.0/dv-p-ls/heart_stroke_data
data <- na.omit(data)
```

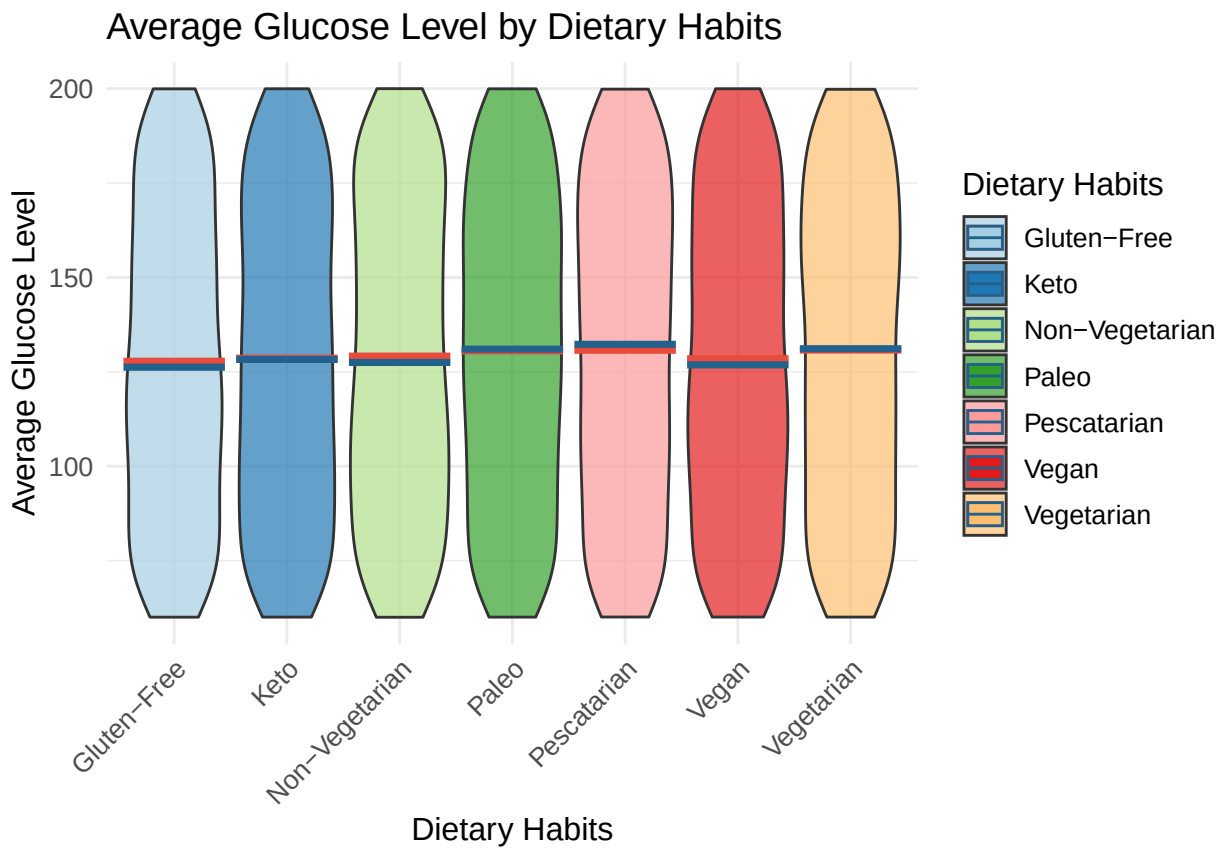
Hypothesis 1 The prevalence of heart disease varies significantly between different work types and is influenced by smoking status.

```
ggplot(data, aes(x = `Work Type`, fill = as.factor(`Heart Disease`))) +
      geom_bar(position = "fill") +
      facet_wrap(~`Smoking Status`) +
      labs(title = "Heart Disease Prevalence by Work Type and Smoking Status",
           x = "Work Type", y = "Proportion", fill = "Heart Disease") +
      theme_minimal() + scale_fill_manual(values = c("0" = "#C0392B", "1" = "#AAB7B8"),
                    labels = c("No Heart Disease", "Heart Disease")) +
      coord_cartesian(ylim = c(0.4, 0.6)) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



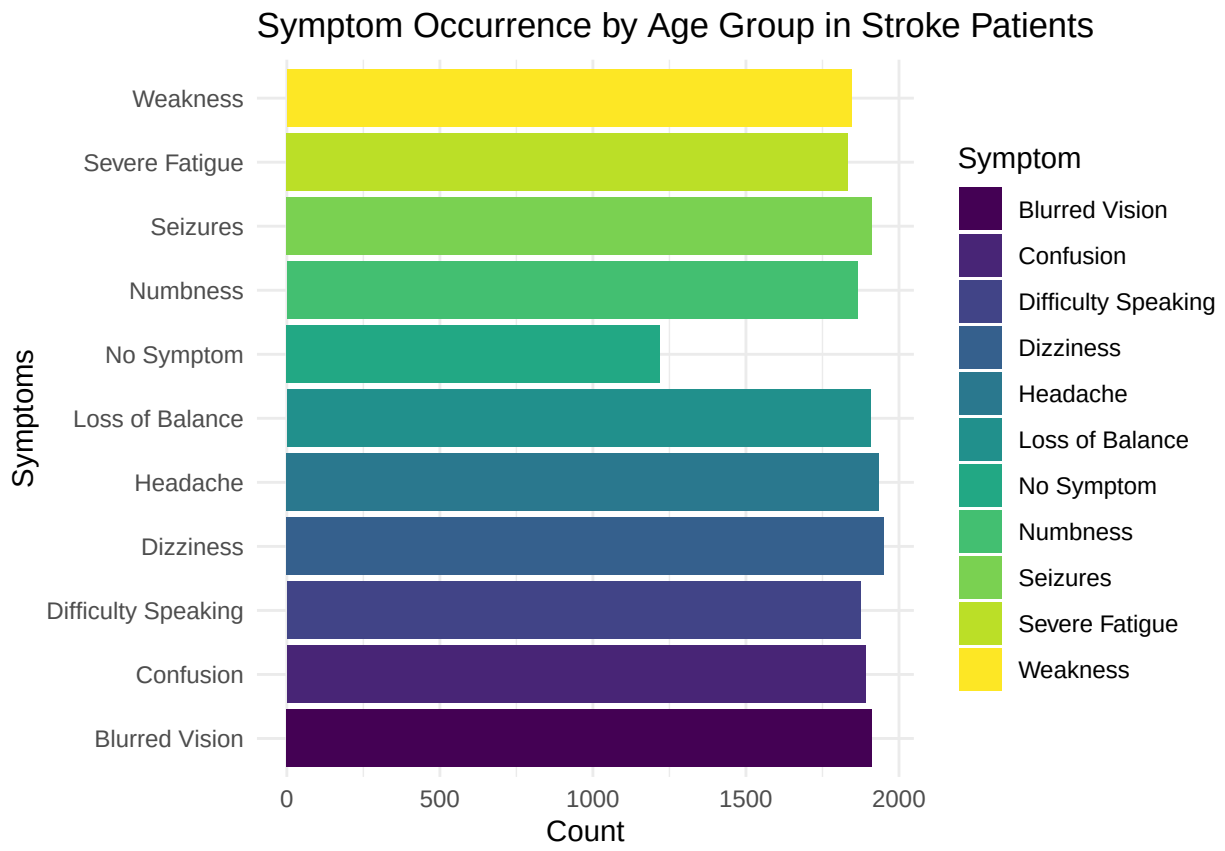
Hypothesis 2 Average glucose levels are significantly different among patients with different dietary habits.

```
ggplot(data, aes(x = `Dietary Habits`, y = `Average Glucose Level`, fill = `Dietary Habits`)) +
  geom_violin(scale = "count", adjust = 1.5, alpha = 0.7) +
  scale_fill_brewer(palette = "Paired") +
  stat_summary(fun = "mean",
               geom = "crossbar",
               color = "#E74C3C") +
  stat_summary(fun = "median",
               geom = "crossbar",
               color = "#21618C") +
  labs(title = "Average Glucose Level by Dietary Habits",
       x = "Dietary Habits", y = "Average Glucose Level") +
  theme_minimal() +
  theme(text = element_text(size = 12),
        axis.text.x = element_text(angle = 45, hjust = 1),
        legend.position = "right")
```



Hypothesis 3 The occurrence of different symptoms varies across age groups for stroke patients.

```
stroke_patients <- data[data$Diagnosis == "Stroke",]
    symptoms_list <- str_split(stroke_patients$Symptoms, ", ", simplify = FALSE)
    symptoms_unlisted <- unlist(symptoms_list)
    symptoms_df <- data.frame(Symptom = symptoms_unlisted)
    ggplot(symptoms_df, aes(x = Symptom, fill = Symptom)) +
      geom_bar() +
      coord_flip() +
      labs(title = "Symptom Occurrence by Age Group in Stroke Patients",
           x = "Symptoms", y = "Count") +
      theme_minimal() + scale_fill_viridis_d()
```



Hypothesis 4 Different work types have distinct distributions of stress levels.

```
ggplot(data, aes(x = `Stress Levels`)) +
      geom_density(fill = "blue", alpha = 0.5) +
      facet_wrap(~`Work Type`) +
      labs(title = "Stress Levels Distribution by Work Type",
           x = "Stress Levels", y = "Density") +
      theme_minimal()
```

Hypothesis 5 The interaction of average glucose level and BMI impacts the stroke risk differently for various age groups.

```
plot_ly(data, x = ~`Average Glucose Level`, y = ~`Body Mass Index (BMI)`, z = ~Age,
        color = ~as.factor(`Stroke History`),
        text = ~ifelse(`Stroke History` == 0, "No Stroke", "Stroke"),
        hoverinfo = "text+x+y+z+color",
        symbols = ~ifelse(`Stroke History`== 0, "No Stroke", "Stroke"),
        type = "scatter3d", mode = "markers") %>%
  layout(title = "3D Scatter Plot: Stroke Risk by Glucose Level, BMI, and Age",
         scene = list(xaxis = list(title = 'Average Glucose Level'),
                      yaxis = list(title = 'BMI'),
                      zaxis = list(title = 'Age')))
```

```
## Warning in RColorBrewer::brewer.pal(N, "Set2"): minimal value for n is 3, returning requested palette

## Warning in RColorBrewer::brewer.pal(N, "Set2"): minimal value for n is 3, returning requested palette
```

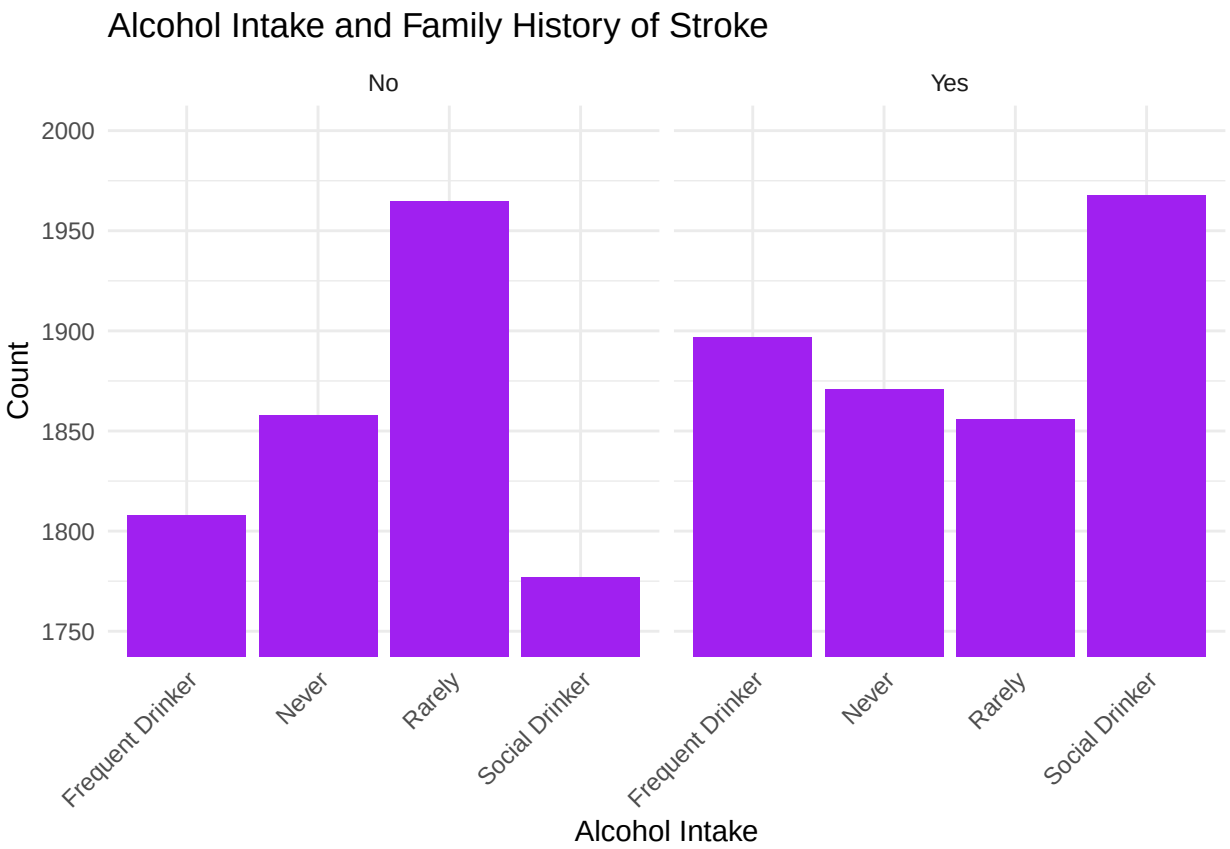
Hypothesis 6 The distribution of stress levels across different age groups follows distinct patterns.

```
data$AgeGroup <- cut(data$Age, breaks = seq(min(data$Age), max(data$Age), by = 20), include.lowest = TRU
    ggplot(data, aes(x = `Stress Levels`, fill = AgeGroup)) +
      geom_density(alpha = 0.6) +
      labs(title = "Stress Level Distribution Across Age Groups",
           x = "Stress Levels", y = "Density") +
```

```
    theme_minimal() +
    scale_fill_viridis(discrete = TRUE)
```

Hypothesis 7 Alcohol intake patterns differ among patients with and without a family history of stroke.

```
ggplot(data, aes(x = `Alcohol Intake`)) +
  geom_bar(fill = "purple") +
  facet_wrap(~`Family History of Stroke`) +
  labs(title = "Alcohol Intake and Family History of Stroke",
       x = "Alcohol Intake", y = "Count") +
  coord_cartesian(ylim = c(1750, 2000)) +
  theme_minimal() + theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Hypothesis 8 This hypothesis suggests that the relationship between a patient's Age and their Average Glucose Level may vary depending on their Gender.

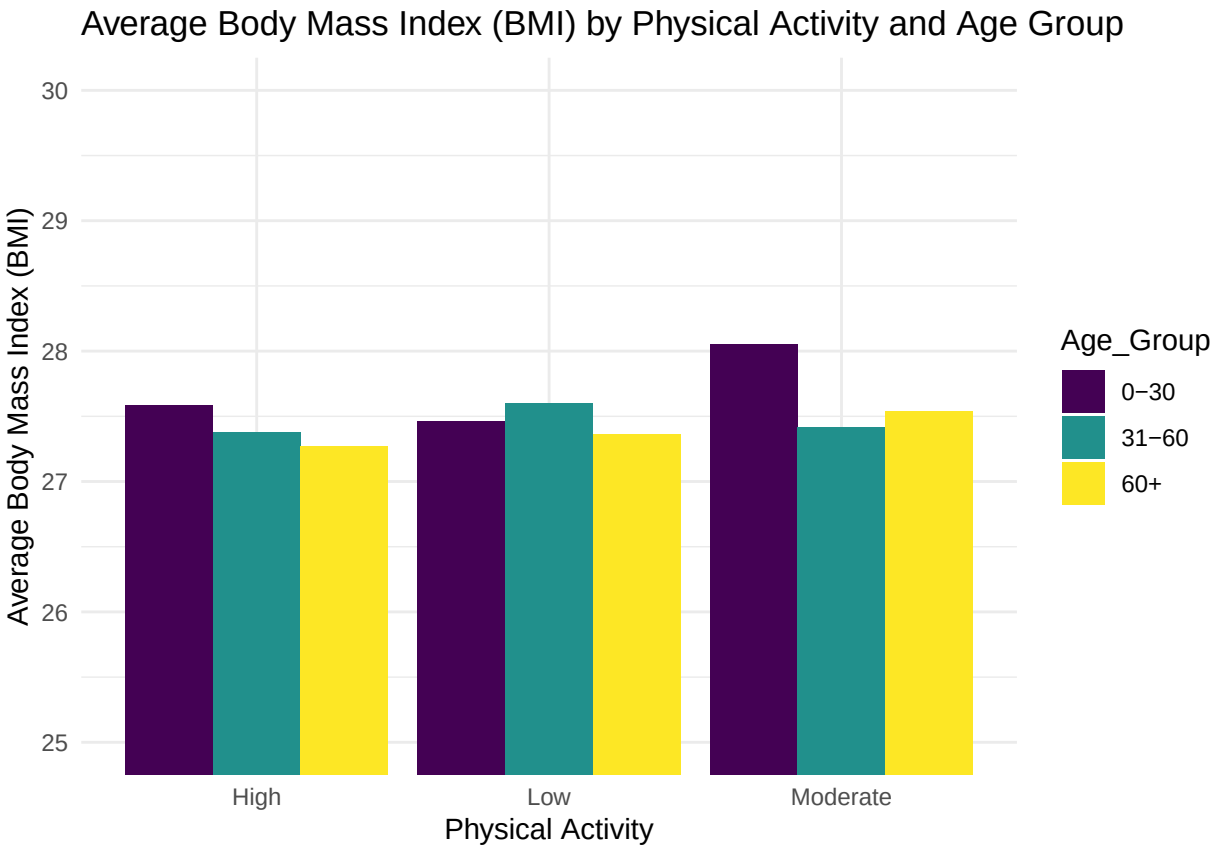
```
data$Age_Group <- cut(data$Age, breaks = c(0, 30, 60, Inf),
                      labels = c("0-30", "31-60", "60+"))

ggplot(data, aes(x = `Average Glucose Level`, fill = Gender)) +
  geom_histogram(binwidth = 10, position = "identity", alpha = 0.7, color = "white") +
  labs(title = "Distribution of Average Glucose Level by Gender and Age Group",
       x = "Average Glucose Level", y = "Frequency") +
  facet_grid(Gender ~ Age_Group) +
  theme_minimal()
```

Hypothesis 9 Average BMI varies significantly between age groups and is affected by physical activity levels. Younger individuals (aged 0-30) are expected to have lower average BMI than older groups (31-60, 60+), and within each group, more physically active individuals will have lower BMI than less active ones.

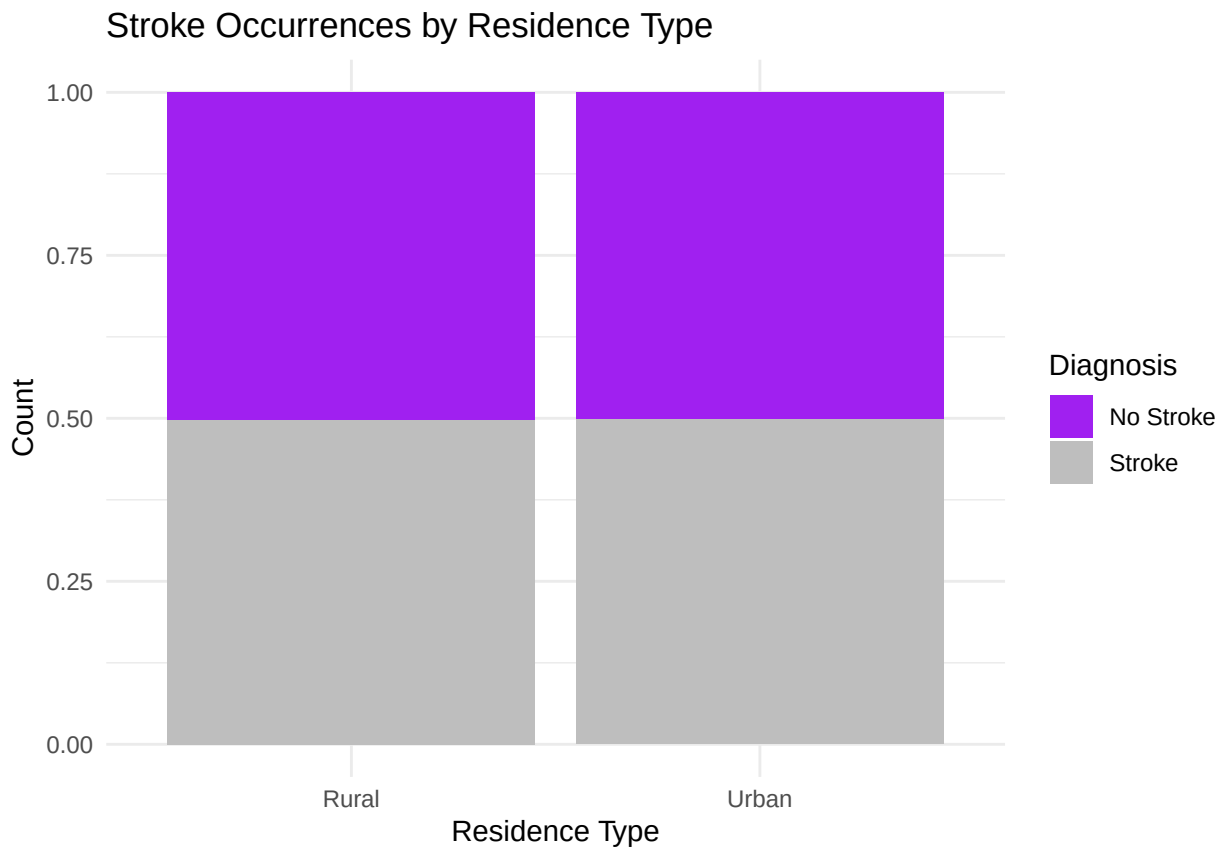
```
data$Age_Group <- cut(data$Age, breaks = c(0, 30, 60, Inf),
                      labels = c("0-30", "31-60", "60+"))

ggplot(data, aes(x = `Physical Activity`, y = `Body Mass Index (BMI)`, fill = Age_Group)) +
  stat_summary(fun = mean, geom = "bar", position = "dodge") +
  labs(title = "Average Body Mass Index (BMI) by Physical Activity and Age Group",
       x = "Physical Activity", y = "Average Body Mass Index (BMI)") +
  scale_fill_viridis_d() +
  coord_cartesian(ylim = c(25, 30)) + theme_minimal()
```



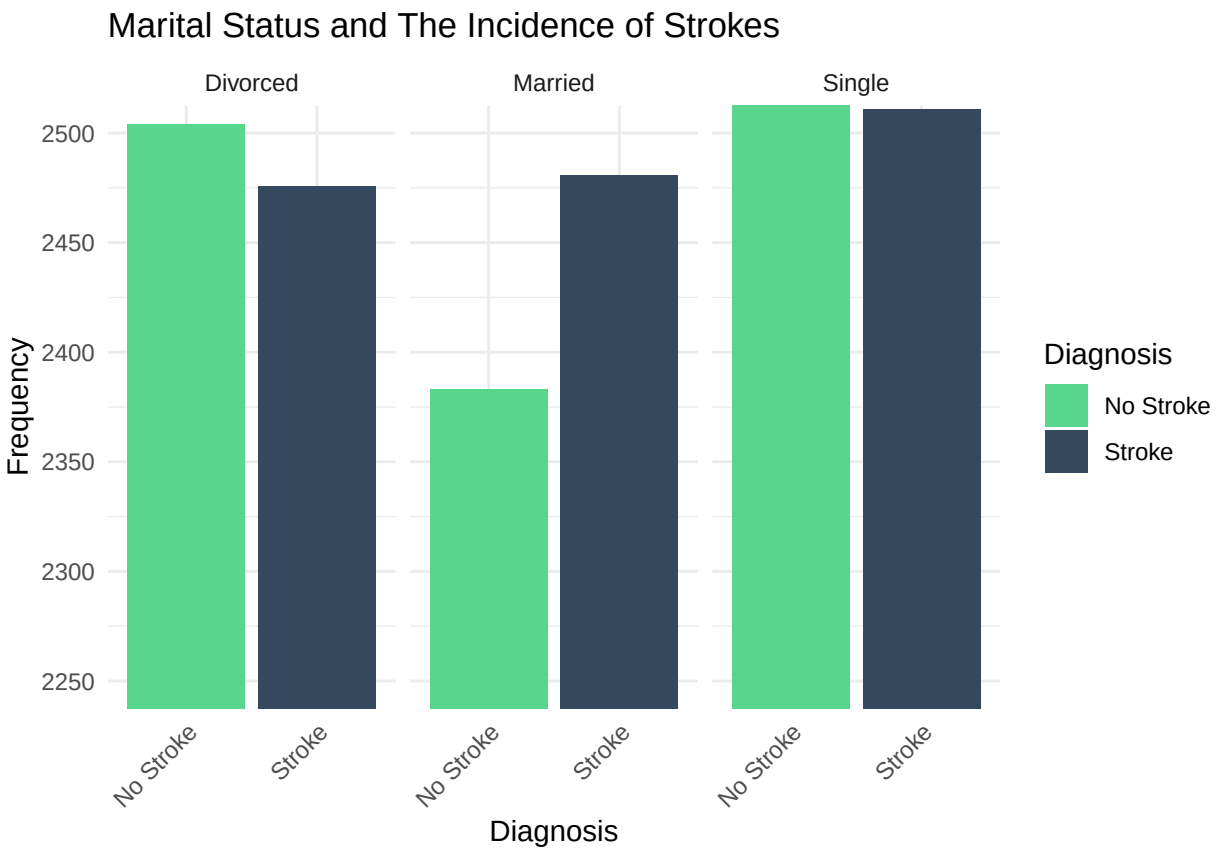
Hypothesis 10 There is a differing prevalence of strokes between individuals residing in urban and rural areas.

```
ggplot(data, aes(x = `Residence Type`, fill = Diagnosis)) +
  geom_bar(position = "fill", stat = "count") +
  labs(title = "Stroke Occurrences by Residence Type",
       x = "Residence Type",
       y = "Count",
       fill = "Diagnosis") +
  scale_fill_manual(values = c("purple", "grey"), labels = c("No Stroke", "Stroke")) +
  theme_minimal()
```



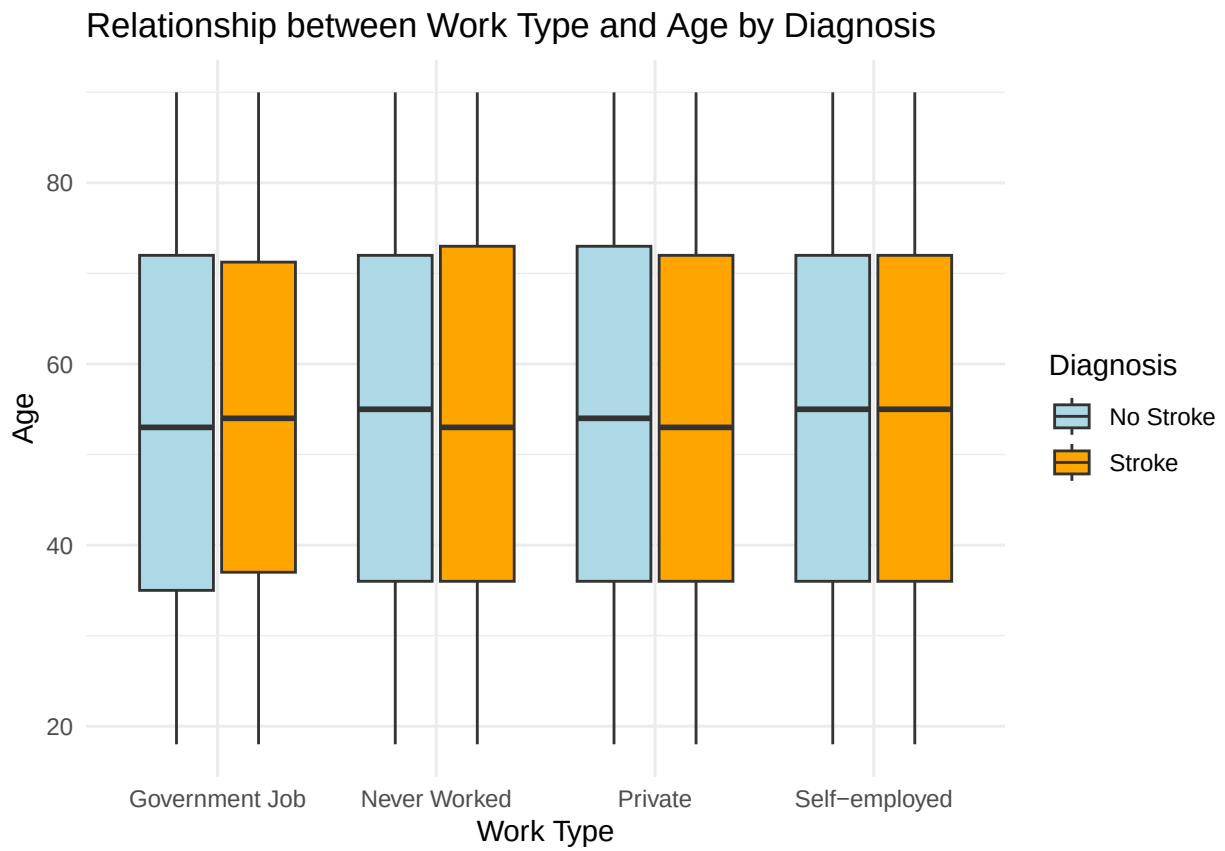
Hypothesis 11 There is a correlation between marital status and the incidence of strokes.

```
ggplot(data, aes(x = Diagnosis, fill = Diagnosis)) +
  geom_bar() +
  facet_wrap(~ `Marital Status`) +
  labs(title = "Marital Status and The Incidence of Strokes",
       x = "Diagnosis",
       y = "Frequency") +
  theme_minimal() +
  scale_fill_manual(values = c("Stroke" = "#34495E", "No Stroke" = "#58D68D")) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))+
  coord_cartesian(ylim = c(2250, 2500))
```



Hypothesis 12 Sedentary work environments, such as government jobs have higher risk of stroke than other work types.

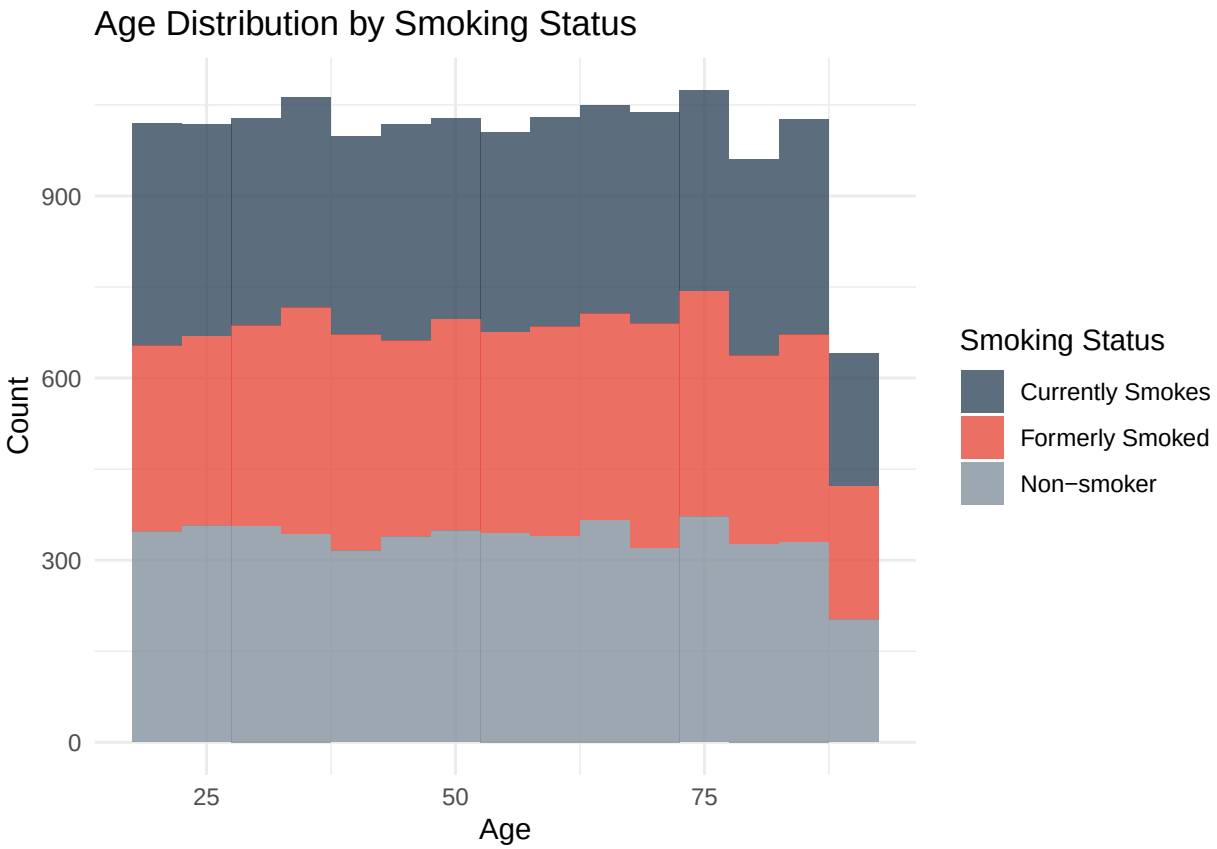
```
ggplot(data, aes(x = `Work Type`, y = Age, fill = Diagnosis)) +
  geom_boxplot() +
  labs(title = "Relationship between Work Type and Age by Diagnosis",
       x = "Work Type", y = "Age") +
  scale_fill_manual(values = c("No Stroke" = "Light Blue", "Stroke" = "Orange")) +
  theme_minimal()
```



Hypothesis 13 The older you are, the more likely you are to have a stroke.

```
ggplot(data, aes(x = Age, fill = `Smoking Status`)) +
geom_histogram(binwidth = 5, alpha = 0.8) +
labs(title = "Age Distribution by Smoking Status", x = "Age", y = "Count") + scale_fill_brewer(palette =
  scale_fill_manual(values = c("Non-smoker" = "#85929E",
                               "Formerly Smoked" = "#E74C3C",
                               "Currently Smokes" = "#34495E")) +
theme_minimal()
```

```
## Scale for fill is already present.
## Adding another scale for fill, which will replace the existing scale.
```


Age Distribution by Smoking Status
Count
900
600
300
0
25
50
75
Age
Smoking Status
Currently Smokes
Formerly Smoked
Non−smoker