33Text

Description automatically generated

|  |  |
| --- | --- |
| **Student Name** | **SUBODH SATYANARAYAN UGILE** |
| **SRN No** | **202201987** |
| **Roll No** | **04** |
| **Program** | **Computer Engg.** |
| **Year** | **Second Year (SY)** |
| **Division** | **F** |
| **Subject** | **Applied Statistical Analysis lab** |
| **CIE NO** | **CIE** |

**MINI PROJECT**

**TOPIC :- Healthcare Dataset ( Stroke Data )**

Use **healthcare** dataset to implement Chi-Square test and visualise the patterns of data, data behaviour, analysis of data and create standardized (Z-) scores for several variables, run Frequencies to explore the distributions of several variables, obtain summary statistics for scale variables using Box-plot, visualize the relationship between two scale variables creating scatter plots to comment your findings.

**Implimantion :-**

|  |
| --- |
| # Load necessary libraries if not already loaded  library(ggplot2)  data <- read.csv("subodh.CSV")  # 1. Chi-Square Test  chi\_square\_result <- chisq.test(data$hypertension, data$stroke)  print(chi\_square\_result)  # 2. Visualizing Data Patterns and Behavior  ggplot(data, aes(x = age)) +  geom\_histogram(binwidth = 5, fill = "green", color = "red") +  labs(x = "Age", y = "Frequency")  # 3. Z-Scores for Several Variable  # Calculate Z-scores for age, avg\_glucose\_level, and bmi  data$age\_z <- scale(data$age)  data$avg\_glucose\_level\_z <- scale(data$avg\_glucose\_level)  data$bmi\_z <- scale(data$bmi)  # 4. Frequency Distribution for Categorical Variables  ggplot(data, aes(x = smoking\_status)) +  geom\_bar(fill = "blue") +  labs(x = "Smoking Status", y = "Frequency") +  theme\_minimal()  ggplot(data, aes(x = work\_type)) +  geom\_bar(fill = "pink") +  labs(x = "work\_type", y = "Frequency") +  theme\_minimal()  ggplot(data, aes(x = gender)) +  geom\_bar(fill = "purple") +  labs(x = "gender", y = "Frequency") +  theme\_minimal()  ggplot(data, aes(x = ever\_married)) +  geom\_bar(fill = "orange") +  labs(x = "ever\_married", y = "Frequency") +  theme\_minimal()  table(data$work\_type)  table(data$gender)  table(data$ever\_married)  # 5. Summary Statistics with Box Plots for Numerical Variables  boxplot(data$age, main = "Age Distribution", ylab = "Age", col = "blue")  boxplot(data$avg\_glucose\_level, main = "Avg. Glucose Level Distribution", ylab = "Avg. Glucose Level", col = "green")  # 6. Scatter Plots to Visualize the Relationship between Two Scale Variables  ggplot(data, aes(x = avg\_glucose\_level, y = bmi, color = Residence\_type)) +  geom\_point() +  labs(x = "Avg. Glucose Level", y = "BMI") +  theme\_minimal()  ggplot(data, aes(x = age, y = bmi, color = work\_type)) +  geom\_point() +  labs(x = "age", y = "bmi") +  theme\_minimal() |

**Explaination :-**

**1. Load necessary libraries:**

- The line `library(ggplot2)` loads the "ggplot2" library, a popular R package for data visualization.

**2. Read Data:**

- `data <- read.csv("subodh.CSV")` reads a CSV file named "subodh.CSV" into a data frame named "data." It assumes that the CSV file contains the dataset you want to work with.

**3. Chi-Square Test:**

- `chi\_square\_result <- chisq.test(data$hypertension, data$stroke)` performs a chi-square test to assess the association between the variables "hypertension" and "stroke."

- `print(chi\_square\_result)` prints the result of the chi-square test to the console.

4. **Visualizing Data Patterns and Behavior:**

- `ggplot(data, aes(x = age))` sets up a ggplot object, specifying that the "age" variable should be on the x-axis.

- `geom\_histogram(binwidth = 5, fill = "green", color = "red")` creates a histogram, specifying the bin width, fill color, and border color.

- `labs(x = "Age", y = "Frequency")` adds labels to the x and y axes.

**5. Z-Scores for Several Variables:**

- `data$age\_z <- scale(data$age)` calculates Z-scores for the "age" variable and stores them in a new column called "age\_z."

- `data$avg\_glucose\_level\_z <- scale(data$avg\_glucose\_level)` calculates Z-scores for "avg\_glucose\_level" and stores them in a new column.

- `data$bmi\_z <- scale(data$bmi)` calculates Z-scores for "bmi" and stores them in a new column.

**6. Frequency Distribution for Categorical Variables:**

- The code creates bar charts to visualize the frequency distribution of categorical variables ("smoking\_status," "work\_type," "gender," and "ever\_married").

- It uses ggplot2 to create bar charts, specifying the fill color and axis labels.

- `table(data$work\_type)`, `table(data$gender)`, and `table(data$ever\_married)` display frequency tables for these variables.

**7. Summary Statistics with Box Plots for Numerical Variables:**

- `boxplot(data$age, main = "Age Distribution", ylab = "Age", col = "blue")` creates a boxplot for the "age" variable, with a title, y-label, and boxplot color.

- A similar boxplot is created for the "avg\_glucose\_level" variable.

8. Scatter Plots to Visualize Relationships:

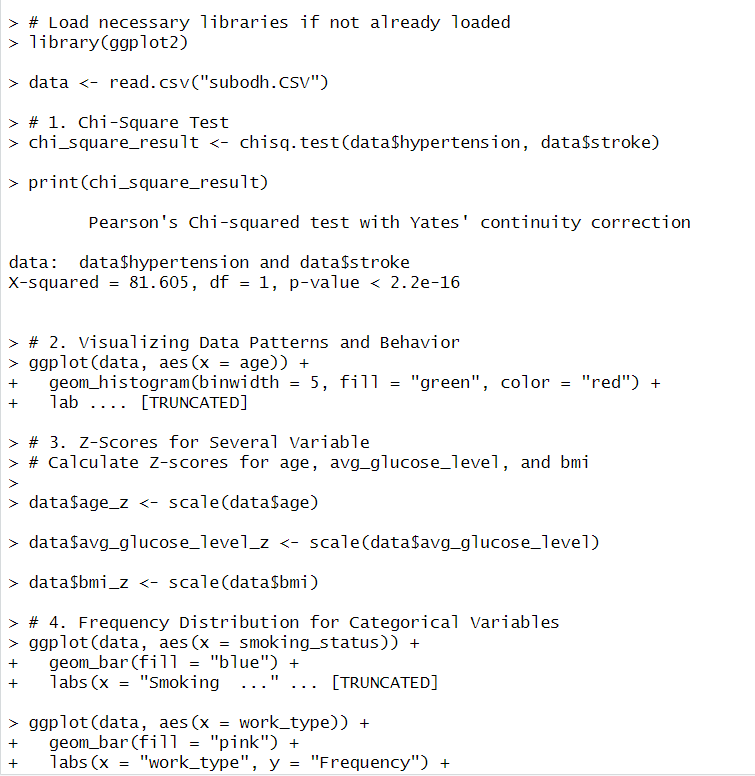
- The code creates scatter plots to visualize the relationship between two scale variables.

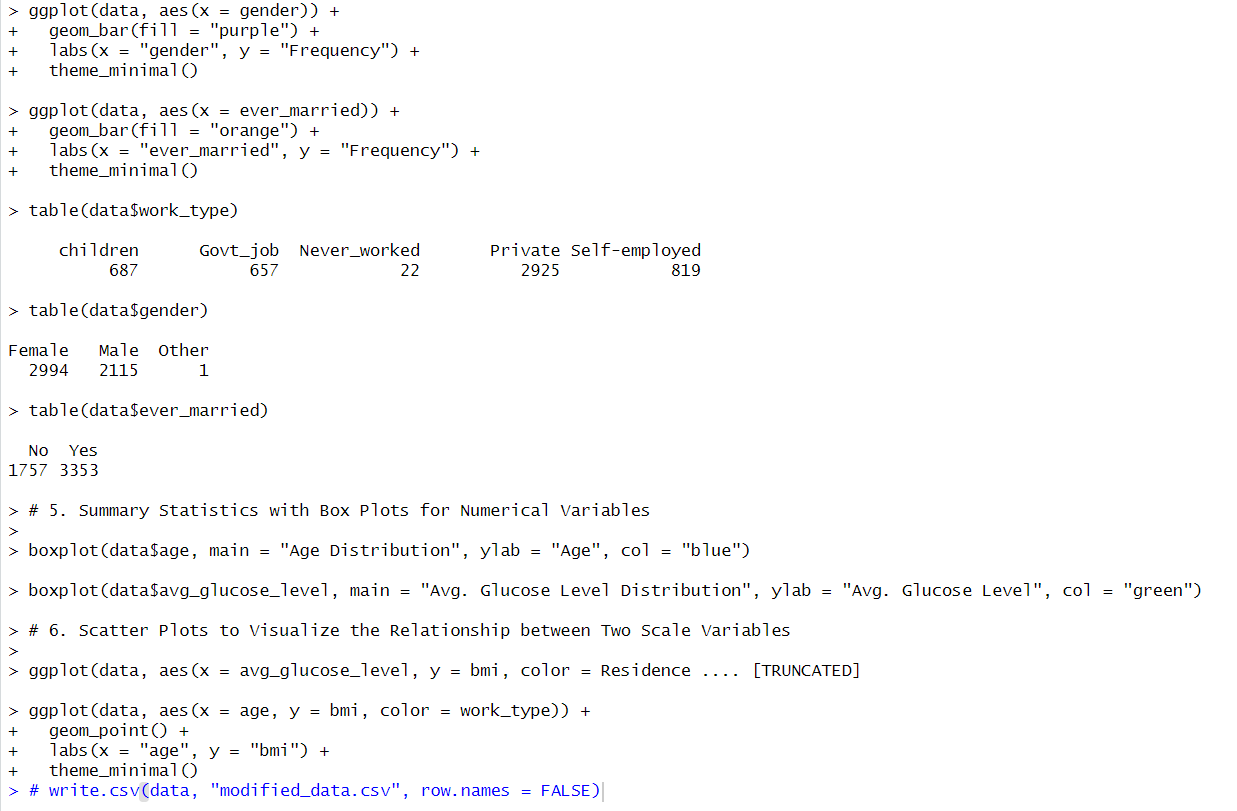
- For example, `ggplot(data, aes(x = avg\_glucose\_level, y = bmi, color = Residence\_type))` sets up a scatter plot, specifying the x and y variables and color-coding by "Residence\_type."

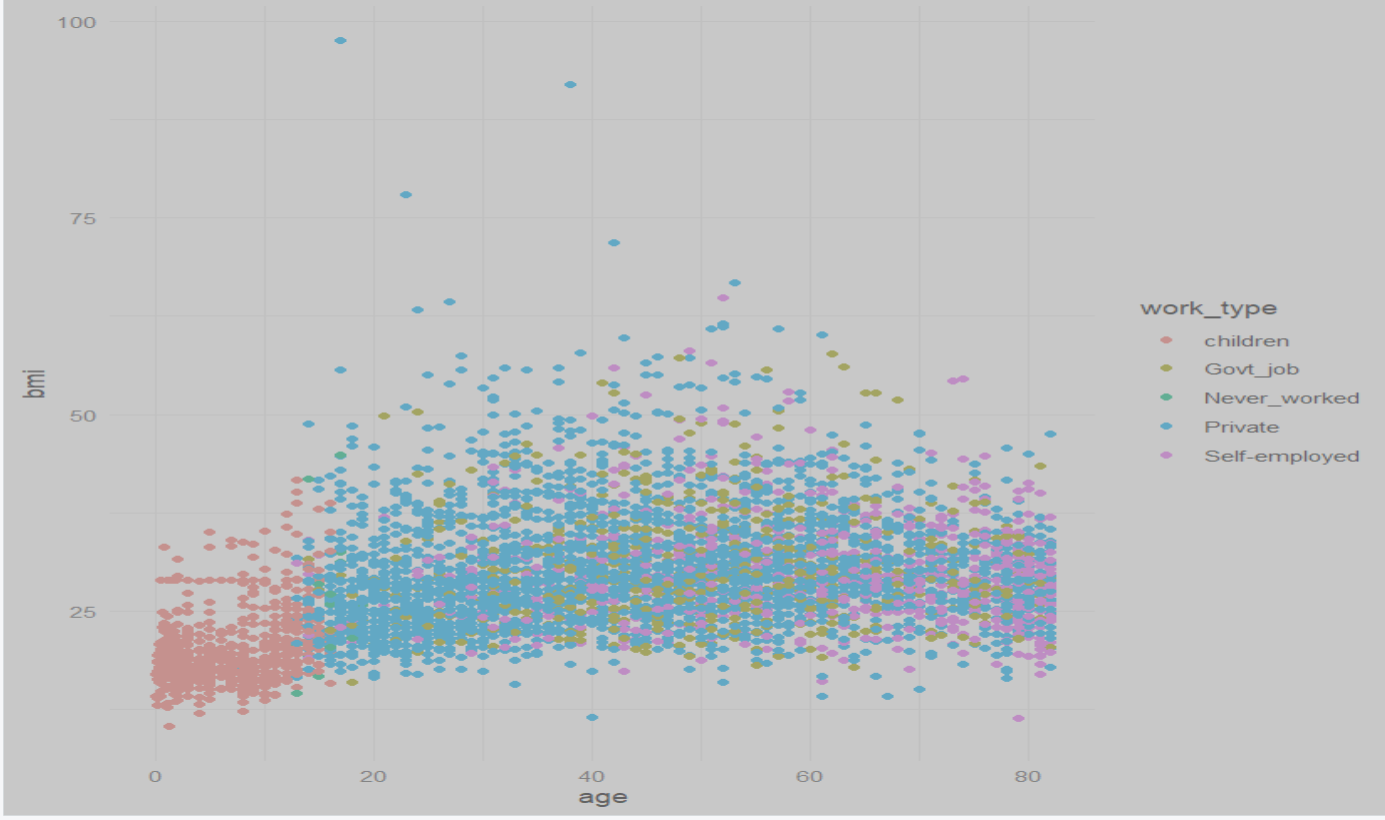
- `geom\_point()` adds points to the plot, and `labs()` adds axis labels.

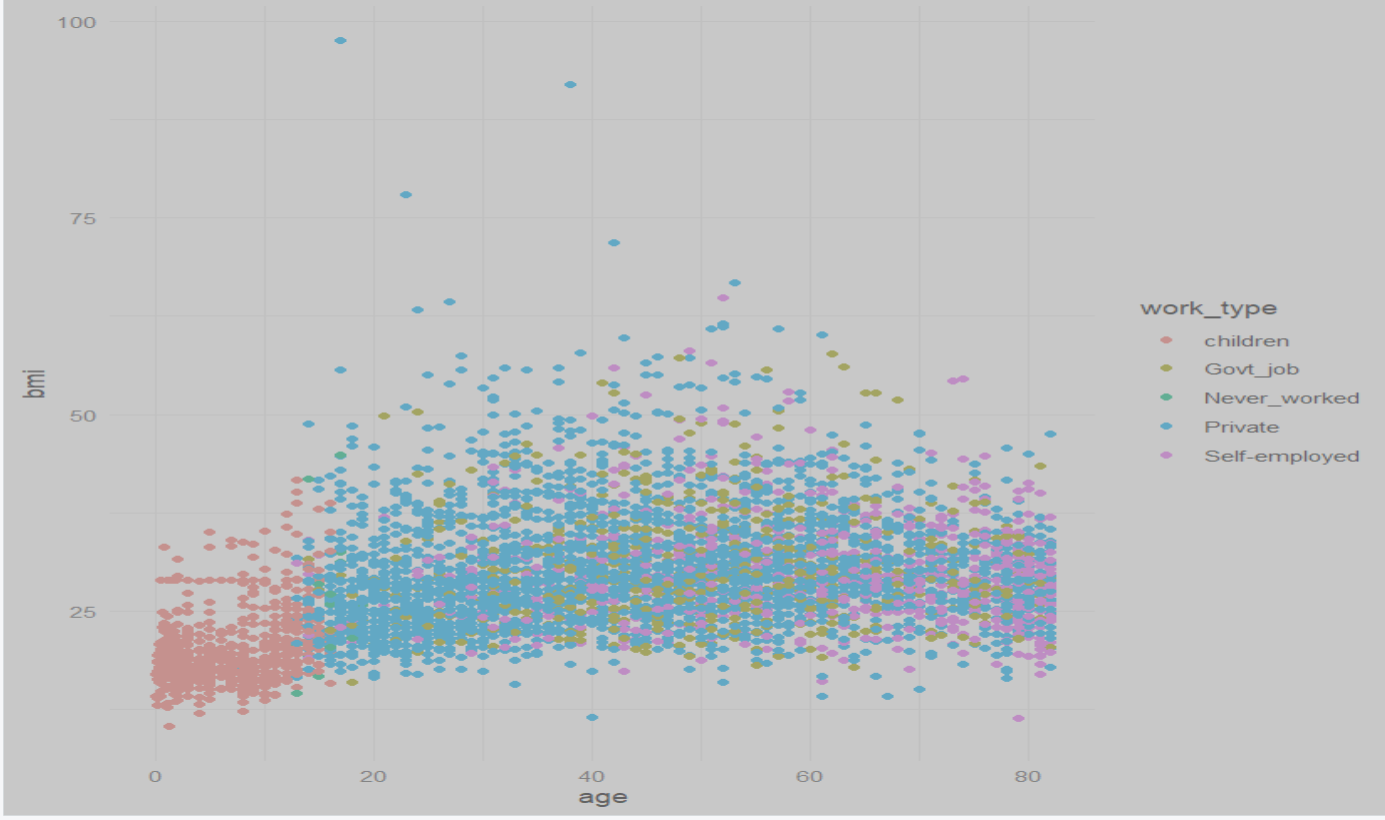
- A similar scatter plot is created for "age" and "bmi," with points color-coded by "work\_type."

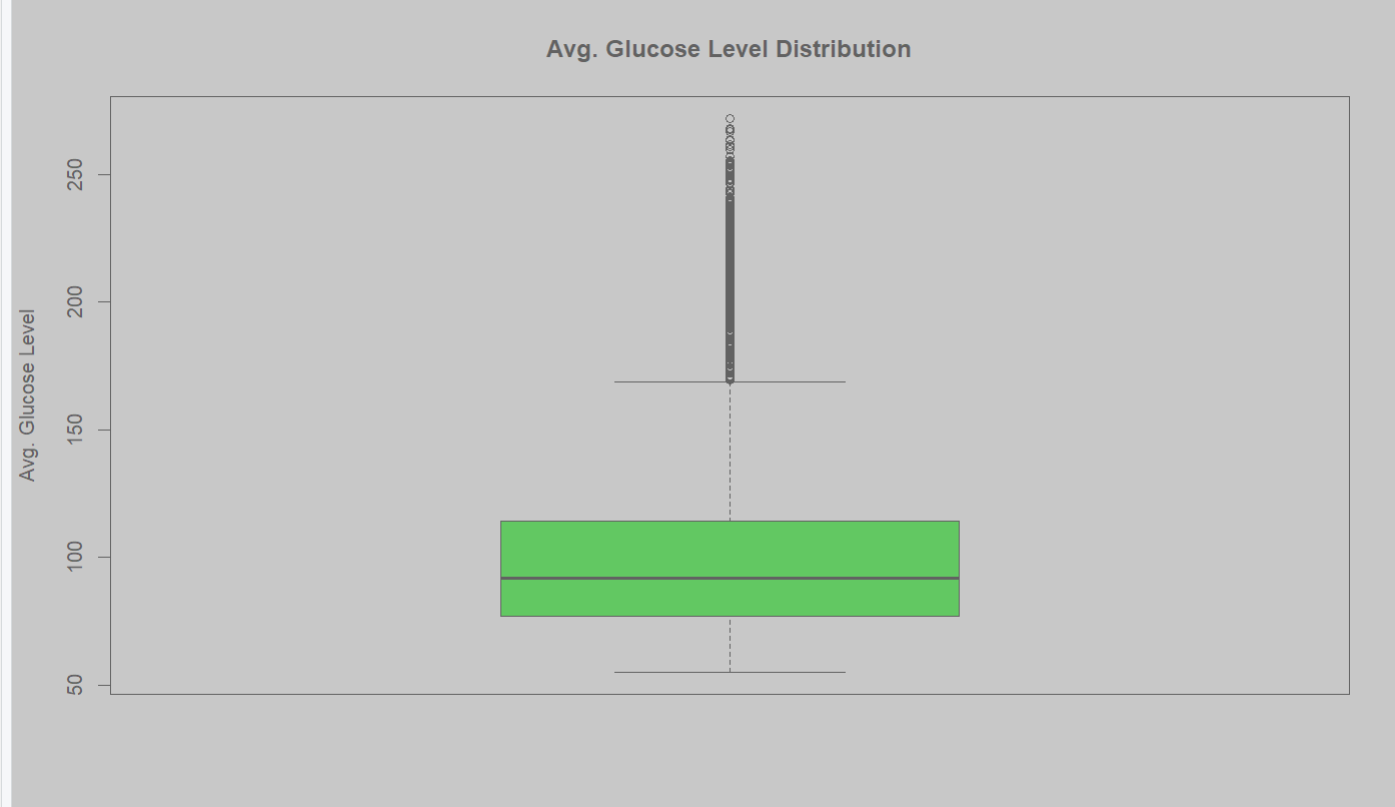
**Output :-**

****

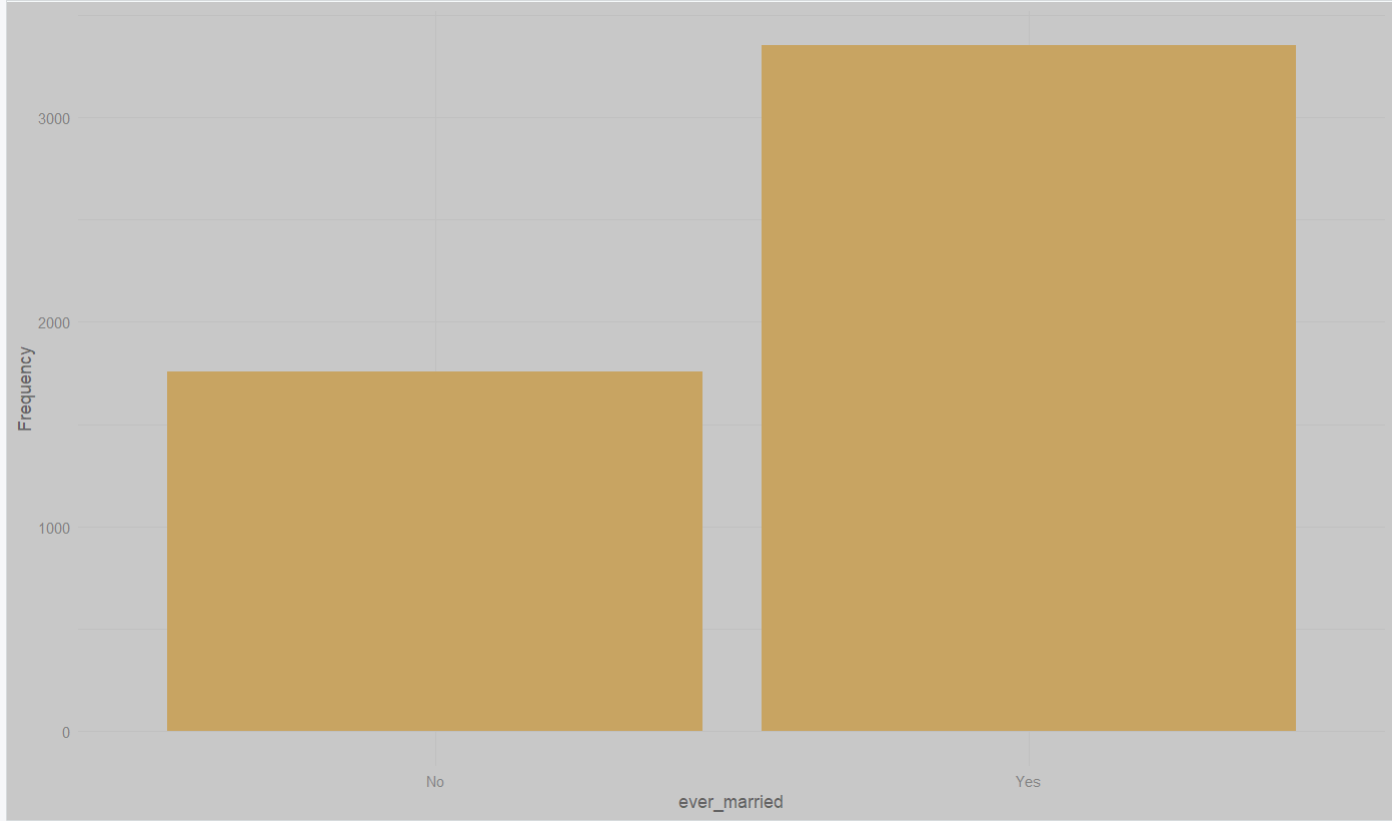


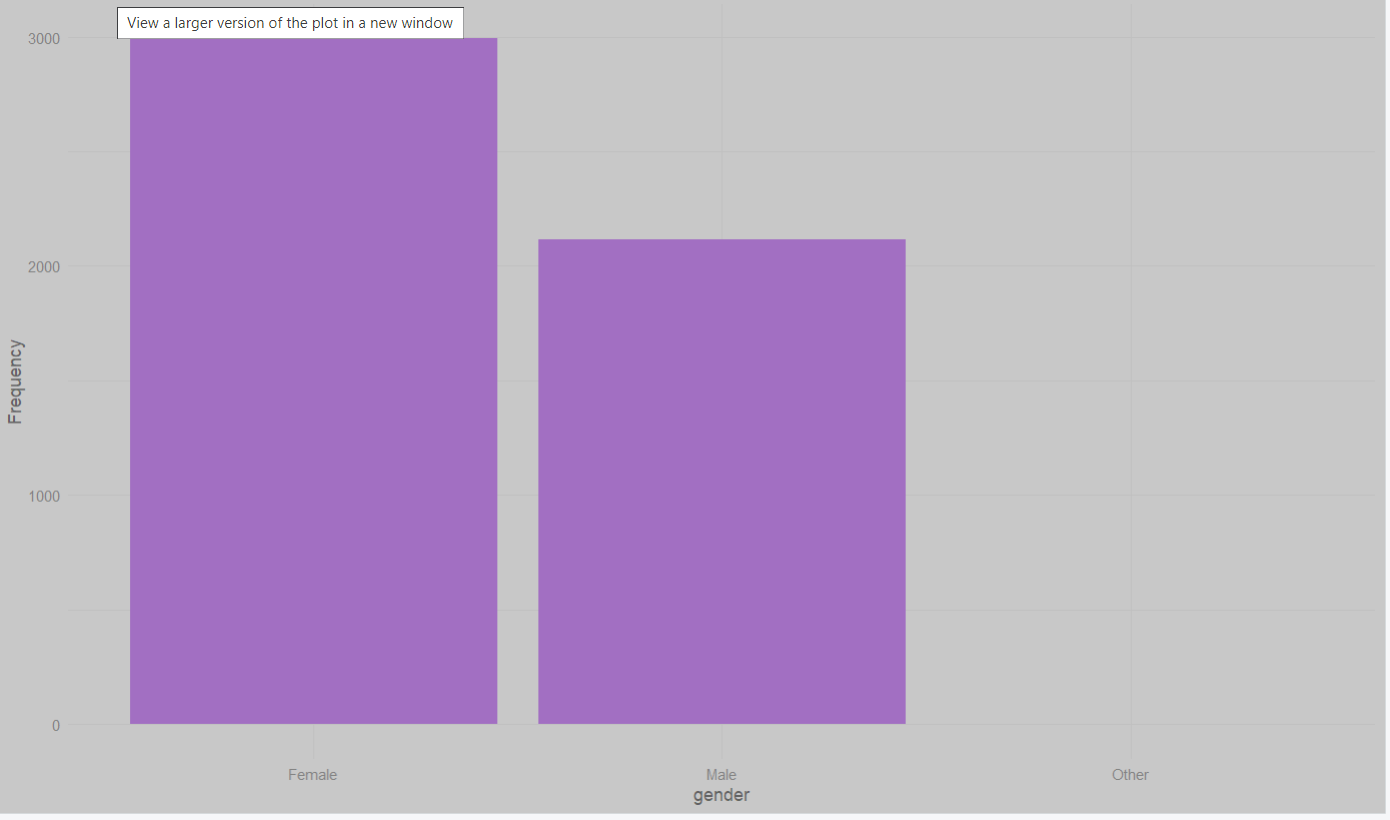
-

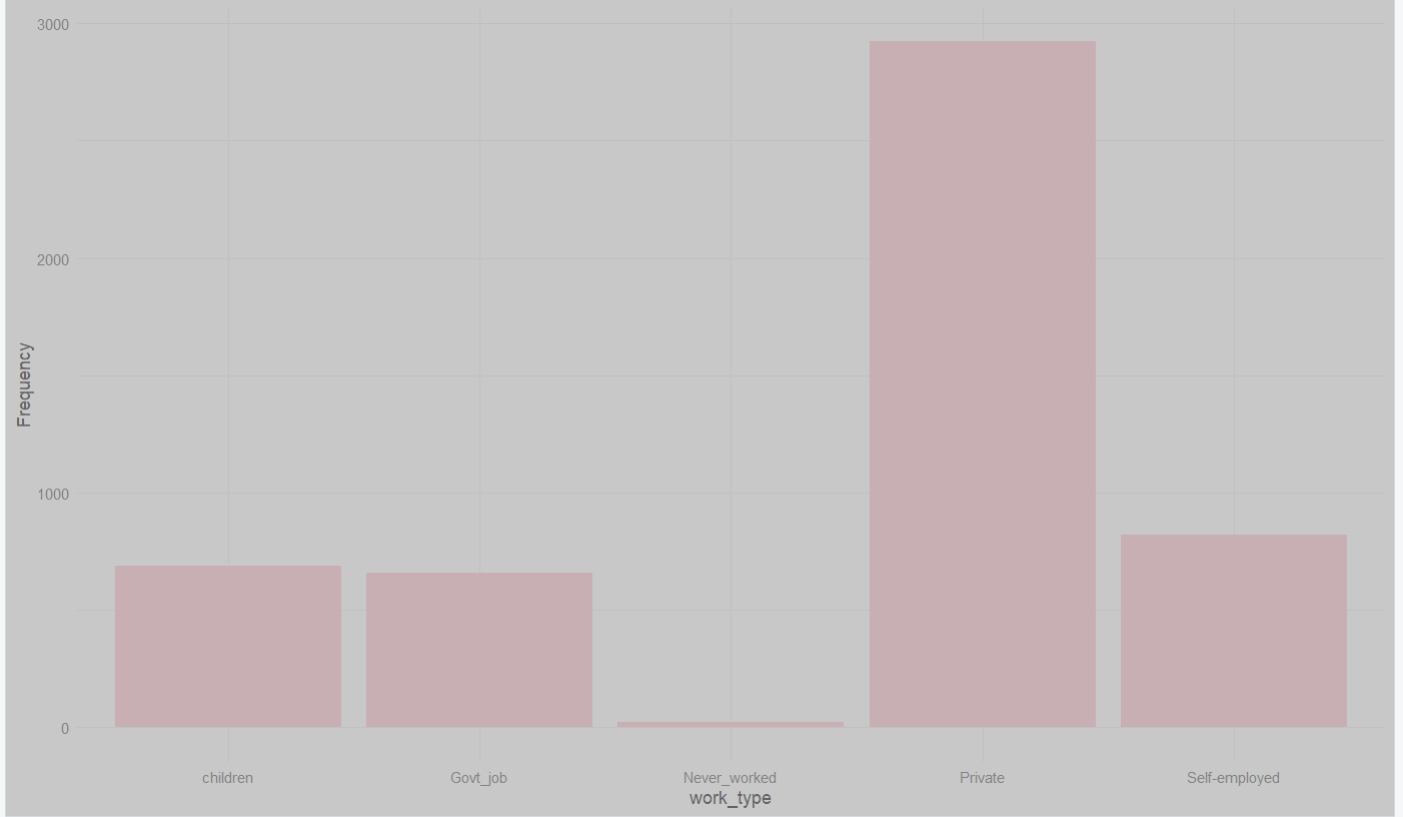


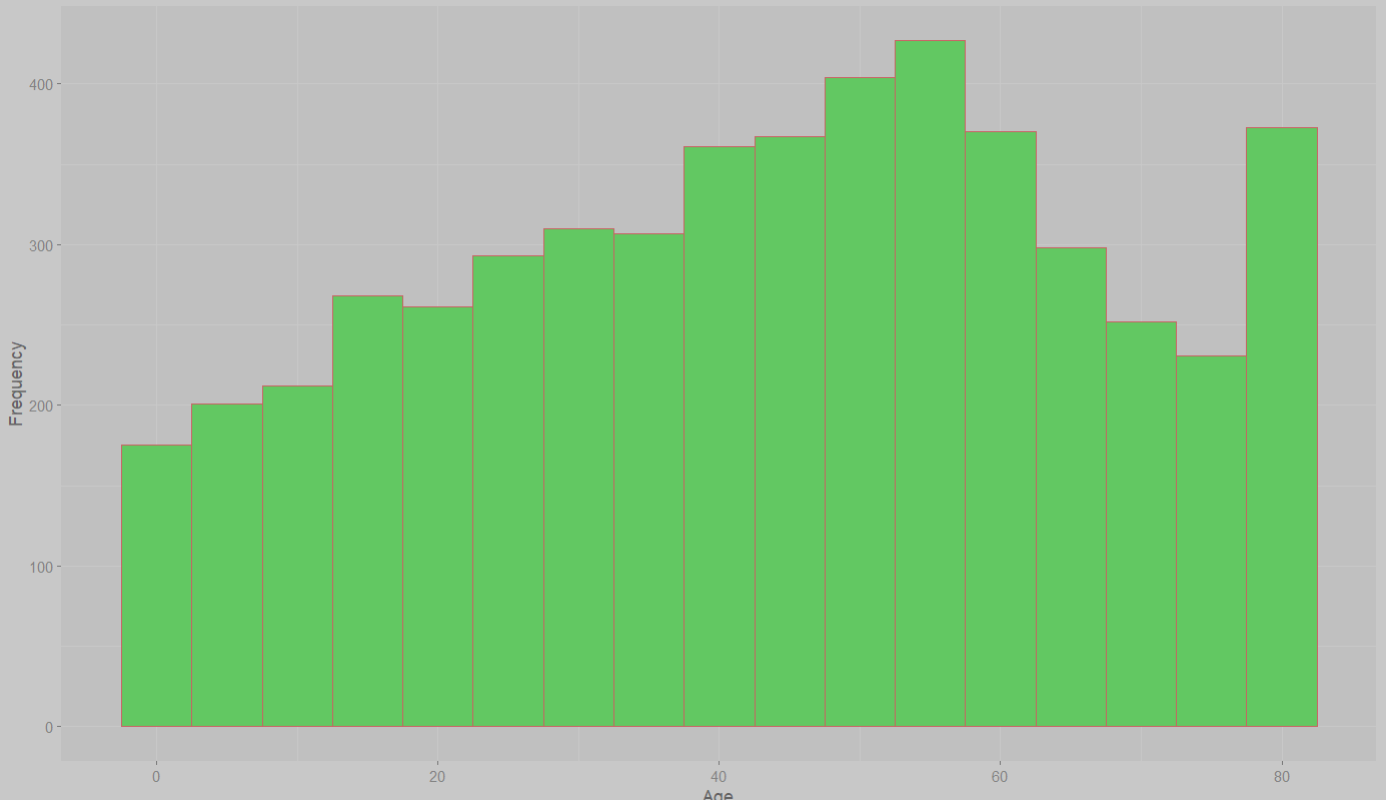


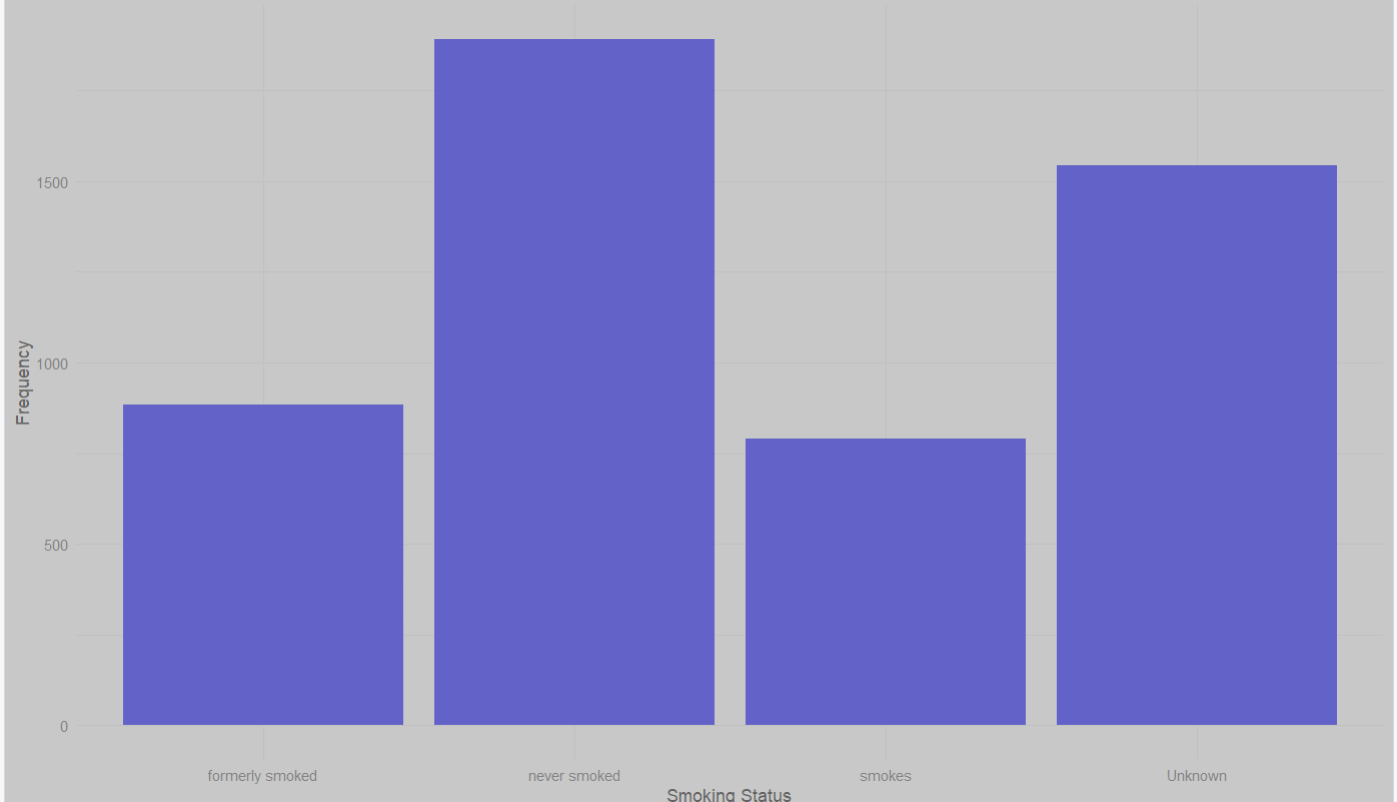












In summary, this code is a comprehensive data analysis script that covers data loading, statistical tests, data visualization, Z-score calculation, frequency distribution analysis, summary statistics, and scatter plot creation. It uses the ggplot2 package for data visualization and various R functions for data manipulation and analysis.