# Assignment II

***1.*** The following is the report:

* 1. The Diabetes Dataset contains information about individuals diagnosed with diabetes, including demographic attributes, medical history, and clinical measurements. This dataset serves as a valuable resource for studying diabetes management, risk factors, and predictive modelling for disease outcomes.
  2. To approach the analysis of the Diabetes Dataset, consider the following methodology:

1.Descriptive Statistics:

Calculate and analyze key descriptive statistics (mean, median, standard deviation) for relevant variables.Data Exploration

2. Data Exploration:

Understand the structure of the dataset.

Explore summary statistics, distributions and relationships between variables.

3. Data Visualization:

Use ggplot or other visualization tools to create plots (scatter plots, histogram) for insights into data patterns.

Correlation Analysis:

Explore correlations between variables to identify potential relationships.

4. Predictive Modelling:

Consider Using Machine Learning algorithms (eg. logistics regression, decision trees) to predict diabetes outcomes based on relevant features.

5. Interpretation:

Summarize findings and interpret the results in the context of diabetes management and risk factors.

c.

# Install and load necessary packages

install.packages(c("tidyverse", "ggplot2"))

library(tidyverse)

library(ggplot2)

# Read the dataset (replace path/to/your/diabetes\_dataset.csv with the actual path)

diabetes\_data <- read.csv("/Users/anamay/Desktop/Thanmayee/archive-2/Training.csv")

# Explore the structure of the dataset

str(diabetes\_data)

# Summary statistics

summary(diabetes\_data)

# Data preprocessing (if needed)

# For example, handling missing values

# diabetes\_data <- na.omit(diabetes\_data)

# Explicitly load the dplyr package for the pipe operator

library(dplyr)

# Create scatter plots

scatter\_plots <- diabetes\_data %>%

ggplot(aes(x = Glucose, y = BMI, color = factor(Outcome))) +

geom\_point() +

labs(title = "Scatter Plot: Glucose vs BMI by Outcome",

x = "Glucose",

y = "BMI",

color = "Outcome") +

theme\_minimal()

# Create histograms

histograms <- diabetes\_data %>%

ggplot(aes(x = Age, fill = factor(Outcome))) +

geom\_histogram(binwidth = 5, position = "identity", alpha = 0.7) +

labs(title = "Histogram: Age by Outcome",

x = "Age",

y = "Frequency",

fill = "Outcome") +

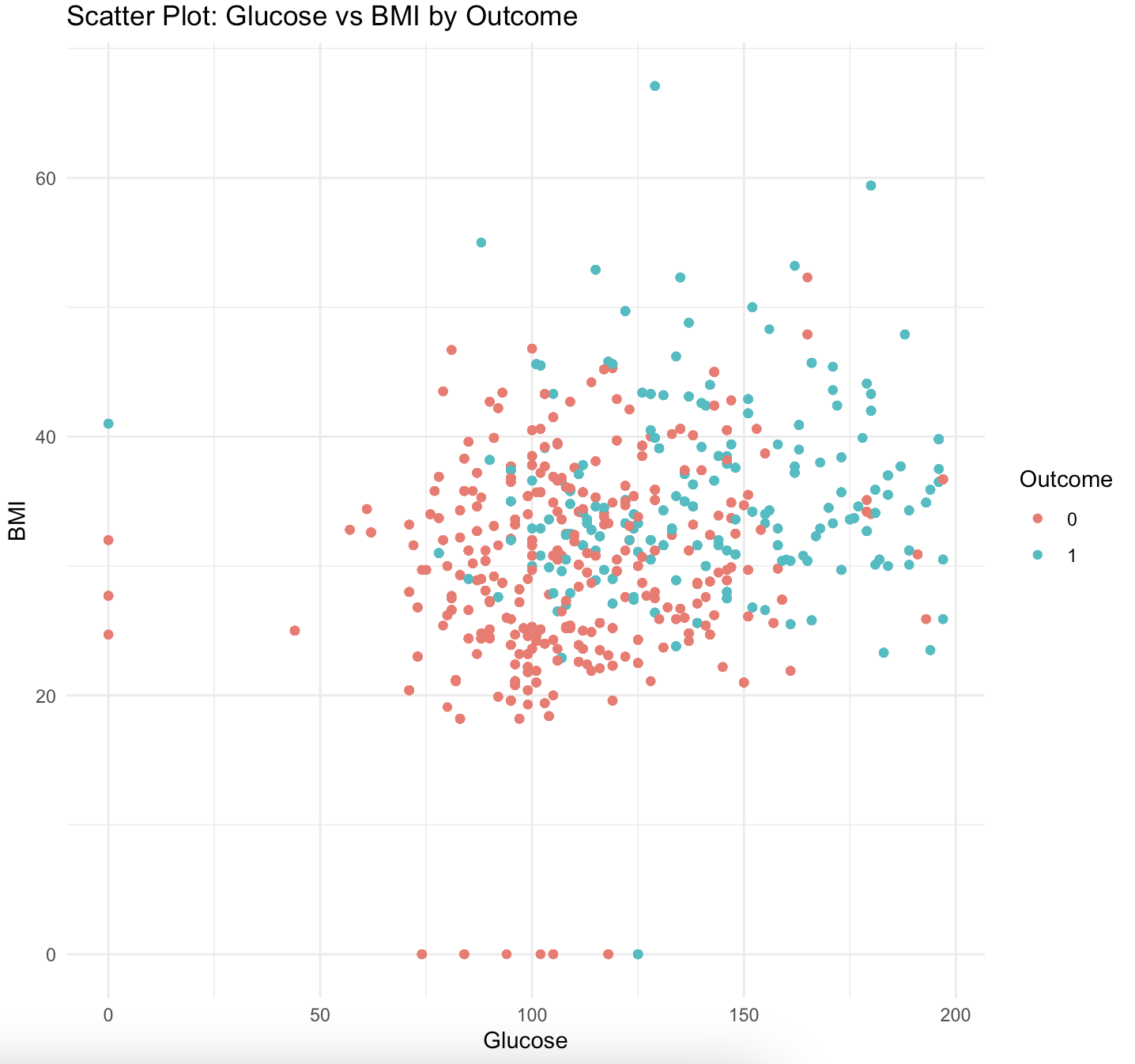
theme\_minimal()

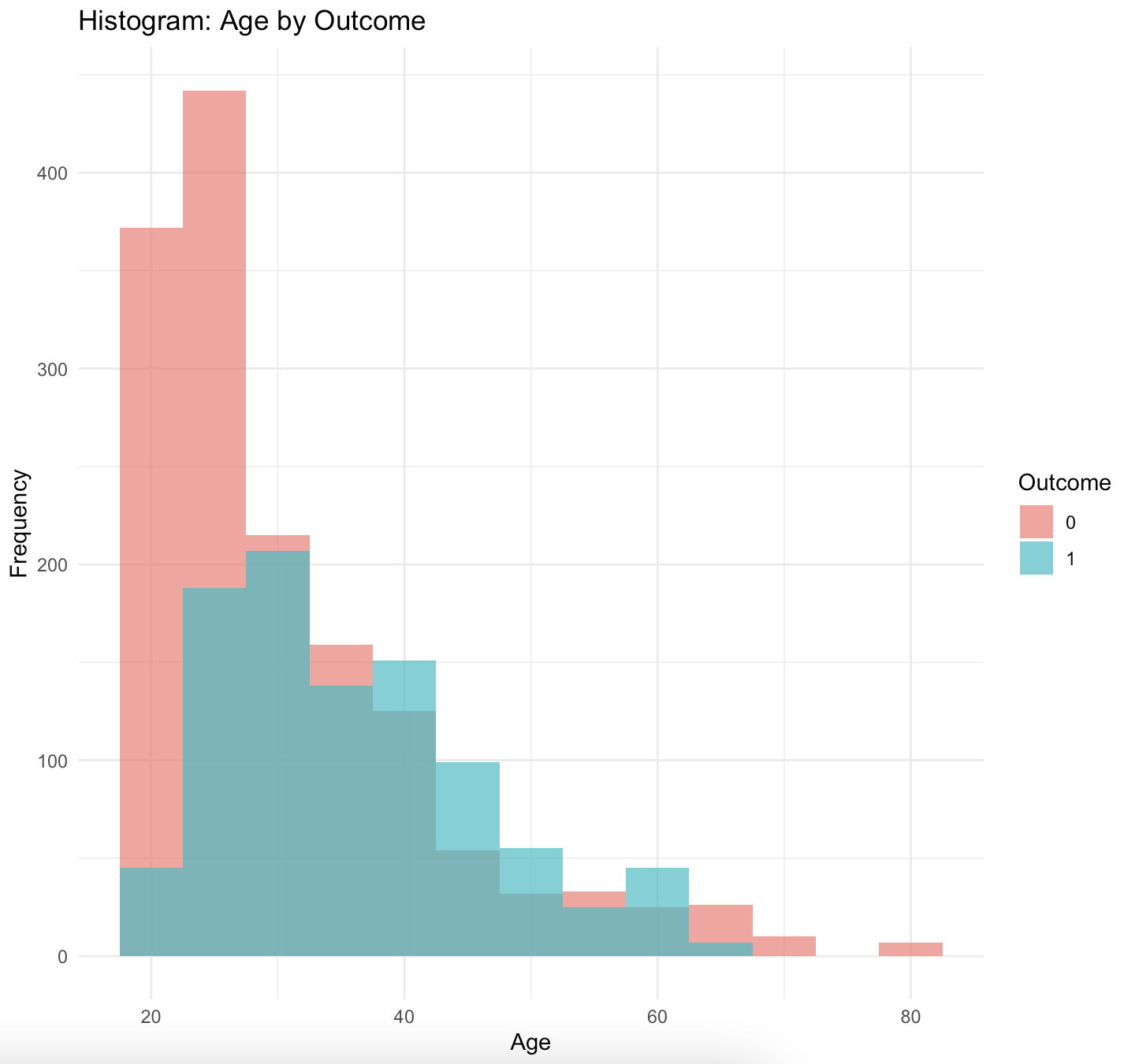
# Display the plots

print(scatter\_plots)

print(histograms)

1. Results including some plots and histogram created using

ggplot:



e. Data:

1. Description of the Dataset:

The dataset used for this analysis contains information about individuals diagnosed with diabetes, including demographic attributes, medical history, and clinical measurements.

2. Source or Citation:

The dataset was obtained from [https://www.kaggle.com/datasets/ehababoelnaga/diabetes-dataset/data].

3. Format:

The data is structured in CSV (Comma-Separated Values) format and consists of rows and columns.

4. Variables:

- Pregnancies:- Number of times pregnant.

- Glucose:- Plasma glucose concentration.

- BloodPressure:- Diastolic blood pressure.

- SkinThickness- Triceps skinfold thickness.

- Insulin:- 2-Hour serum insulin.

- BMI:- Body mass index.

- DiabetesPedigreeFunction:- Diabetes pedigree function,

providing a measure of diabetes hereditary risk.

- Age:- Age in years.

- Outcome:- Binary variable indicating if the individual has

diabetes (1) or not (0).

5. Size:

The dataset comprises 768 observations (rows) and 9 variables (columns).

***2.***

a.i) Formulating the hypothesis:

Hypothesis 1:- Problem: Explore whether there is a significant difference in the average glucose levels between individuals with and without diabetes.

b.i) Hypotheses:

* + - * Null Hypothesis (H₀): The average glucose levels are the same for individuals with and without diabetes.
      * Alternative Hypothesis (H₁): There is a significant difference in average glucose levels between individuals with and without diabetes.

c.i) Methodology:

* + - * Two-sample t-test will be used to compare the means of glucose levels in the two groups.

d.i) *R code:*

*# Assuming diabetes\_data is your dataset*

*# H₀: The average glucose levels are the same for individuals with and without diabetes.*

*# H₁: There is a significant difference in average glucose levels between individuals with and without diabetes.*

*t\_test\_result <- t.test(Glucose ~ Outcome, data = diabetes\_data)*

*t\_test\_result*

*e.i)* Results:

The t-test result will include the t-statistic, degrees of freedom, p-value, and confidence interval. If the p-value is less than the significance level (e.g., 0.05), you may reject the null hypothesis and conclude that there is a significant difference in average glucose levels.

a.ii) Hypothesis 2:- Problem: Investigate whether there is an association between the number of pregnancies and the likelihood of diabetes.

b.ii) Hypotheses:

* + - * Null Hypothesis (H₀): The number of pregnancies is not associated with the likelihood of diabetes.
      * Alternative Hypothesis (H₁): There is a significant association between the number of pregnancies and the likelihood of diabetes.

c.ii) Methodology:

* + - * Chi-square test of independence will be used to assess the association between the two categorical variables.

*d.ii) R code:*

# Assuming diabetes\_data is your dataset

# H₀: The number of pregnancies is not associated with the likelihood of diabetes.

# H₁: There is a significant association between the number of pregnancies and the likelihood of diabetes.

chi\_square\_result <- chisq.test(table(diabetes\_data$Pregnancies, diabetes\_data$Outcome))

chi\_square\_result

e.ii) Results:

The chi-square test result will include the chi-square statistic, degrees of freedom, and p-value. If the p-value is less than the significance level, you may reject the null hypothesis and conclude that there is a significant association between the number of pregnancies and the likelihood of diabetes.