**EXPLORATORY DATA ANALYSIS**

**Introduction:**

This report explores the cardiovascular dataset, which contains information on patients’ medical history and physical measurements. The goal of this analysis is to identify patterns and relationships that may be useful for predicting the risk of cardiovascular disease.

**Dataset:** Cardiovascular\_train\_new dataset.

**Total variables:** 18

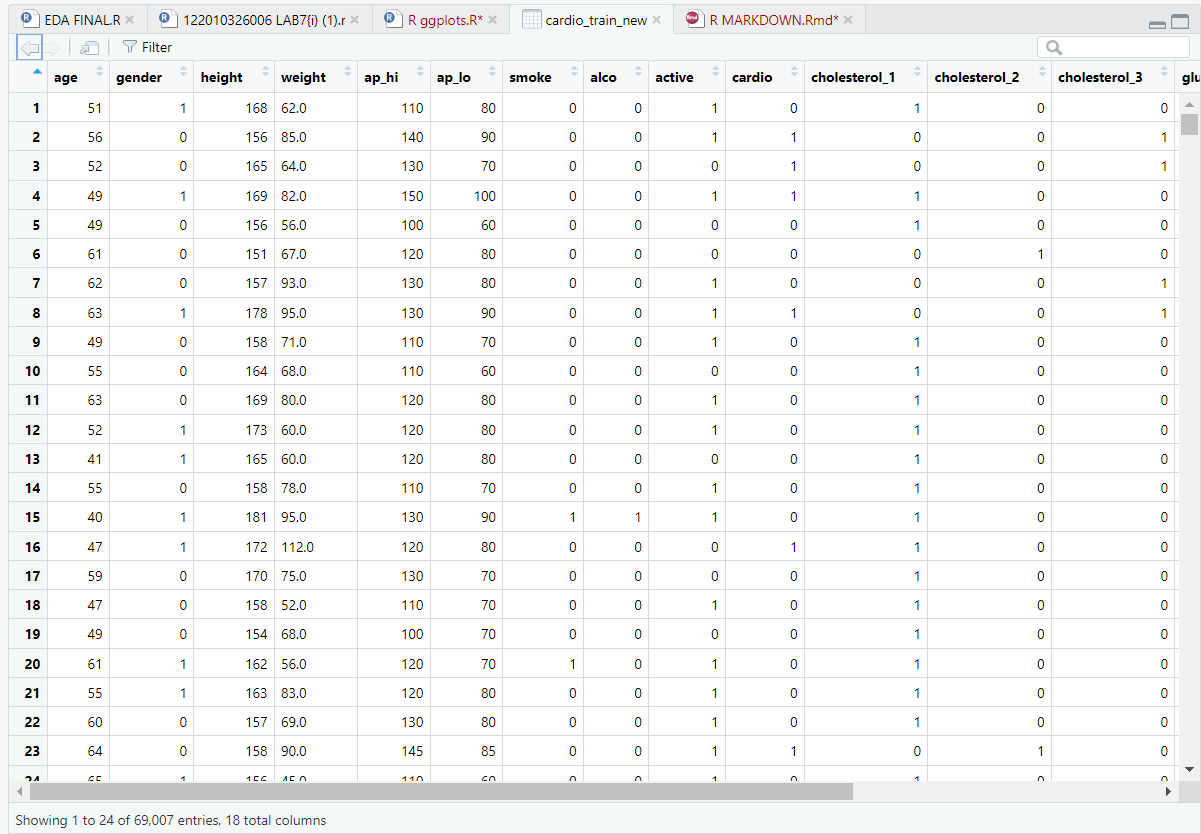
**IMPORTING AND READING DATASET:**

# IMPORTING AND READING DATASET

library(readxl)

d <- read\_excel("C:\\Users\\ramri\\Desktop\\cardio\_train new.xlsx")

**OUTPUT:**

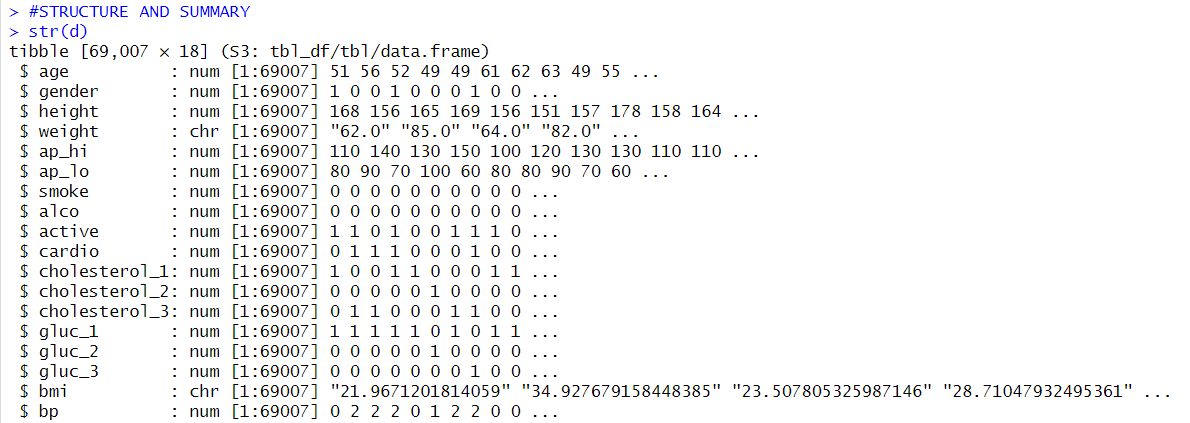


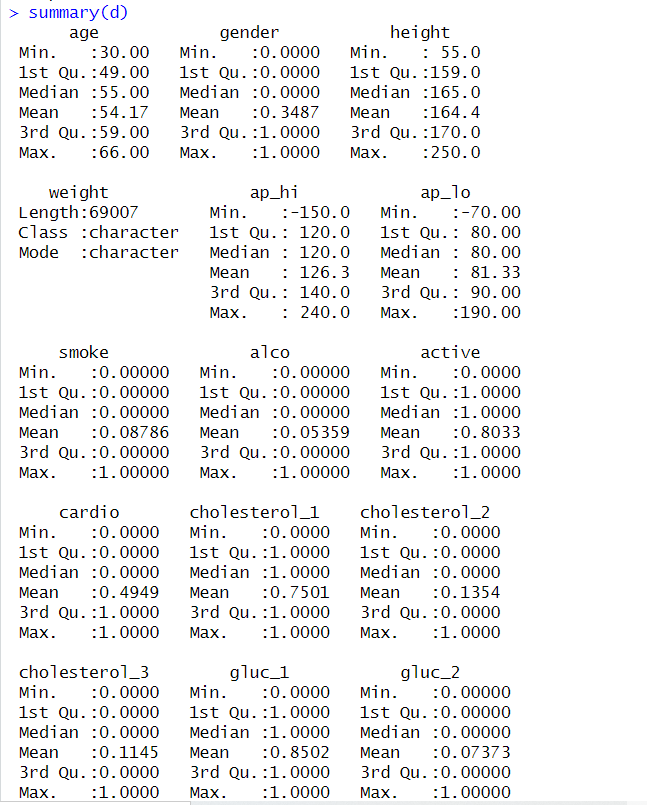
**STRUCTURE AND SUMMARY OF DATASET:**

str(d)

summary(d)

**OUTPUT**:





**LIBRARY:**

library(ggplot2)

library(dplyr)

**BAR CHART OF CATEGORICAL VARIABLE**

**Code:**

#BAR CHART FOR DISTRIBUTION OF CATEGORICAL VARIABLE.

# Create a table of the counts of each category of the variable of interest

variable\_counts <- d %>%

count(active)

# Create bar chart of smoking status

ggplot(variable\_counts, aes(x = active, y = n, fill = gender)) +

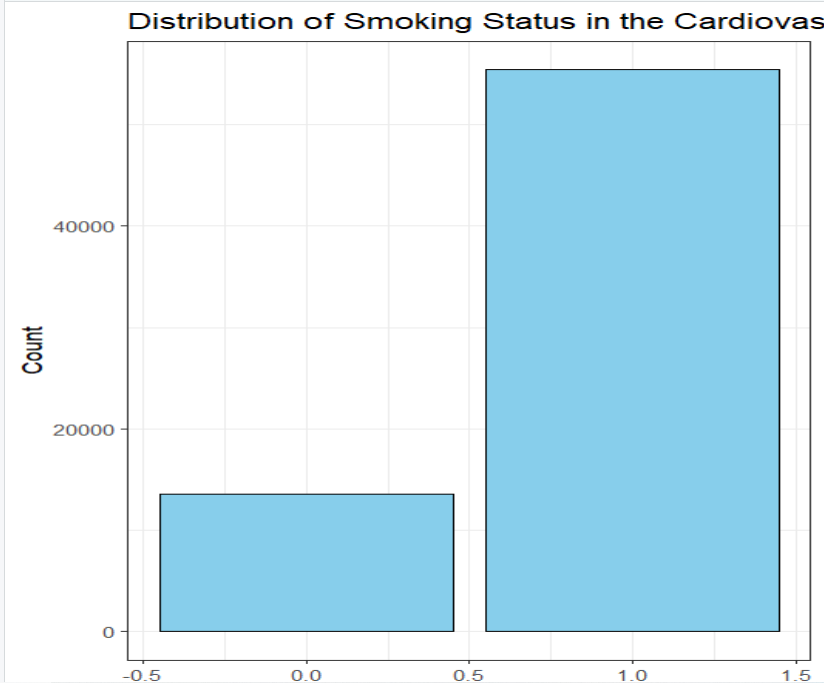
geom\_bar(stat = "identity", fill = "skyblue", color = "black") +

labs(x = "Smoking Status", y = "Count",

title = "Distribution of Smoking Status in the Cardiovascular Dataset")+

theme\_bw()

**Output:**

****

**FREQUENCY POLYGON**

**Code:**

#FREQUENCY POLYGON

ggplot(d, aes(x=age, color=gender)) +

geom\_freqpoly(alpha = 2, color = "red") +

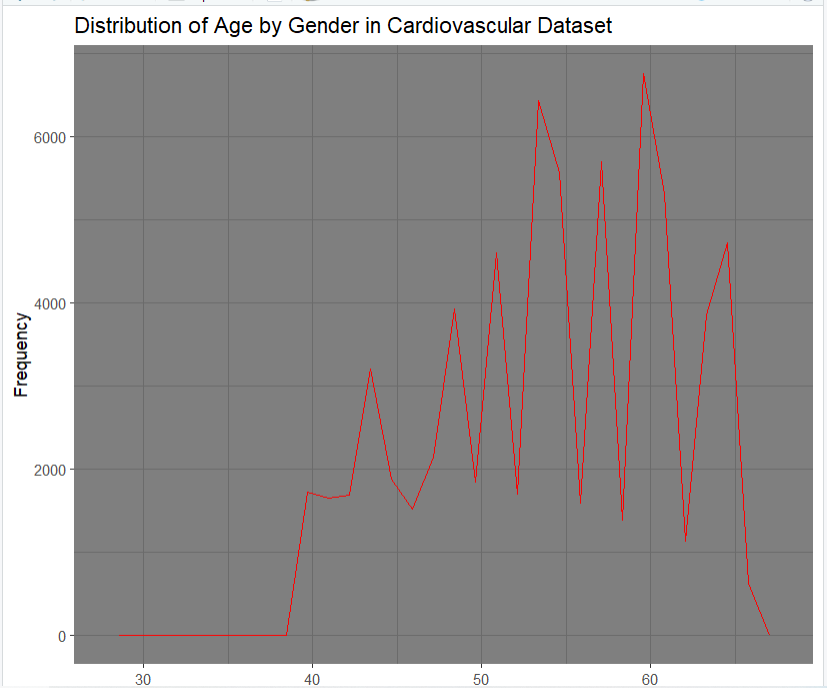
labs(title="Distribution of Age by Gender in Cardiovascular Dataset",

x="Age", y="Frequency") +

scale\_fill\_manual(values=c("#E69F00", "#0072B2")) +

theme\_dark()

**Output:**



**COUNTING**

**Code:**

# Compute basic statistics of the age variable

mean\_age <- mean(d$age)

sd\_age <- sd(d$age)

median\_age <- median(d$age)

q1\_age <- quantile(d$age, 0.25)

q3\_age <- quantile(d$age, 0.75)

**# Print the results**

cat("Mean age: ", mean\_age, "\n")

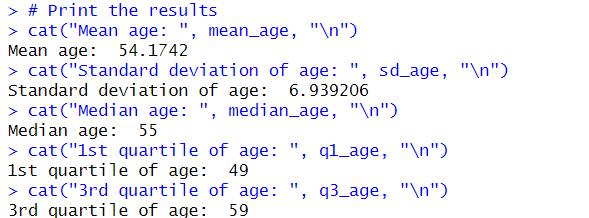
cat("Standard deviation of age: ", sd\_age, "\n")

cat("Median age: ", median\_age, "\n")

cat("1st quartile of age: ", q1\_age, "\n")

cat("3rd quartile of age: ", q3\_age, "\n")

**Output:**



**Scatterplot of age vs. height**

**Code:**

# Create a scatterplot of age vs. height with point size based on count

ggplot(d, aes(x=age, y=height)) +

geom\_count() +

labs(title="Scatterplot of Age vs. Height", x="Age", y="Height")

**# Analyze Exponential Relationship between the Weight and Age**

ggplot(d, aes(x=age, y=gender)) +

geom\_point() +

labs(title="Scatterplot of Age vs. Gender", x="Age", y="Gender")

**# Fit a linear regression model**

linear\_model <- lm(cardio ~ age, data=d)

# Add the linear regression line to the scatterplot

ggplot(d, aes(x=age, y=gender)) +

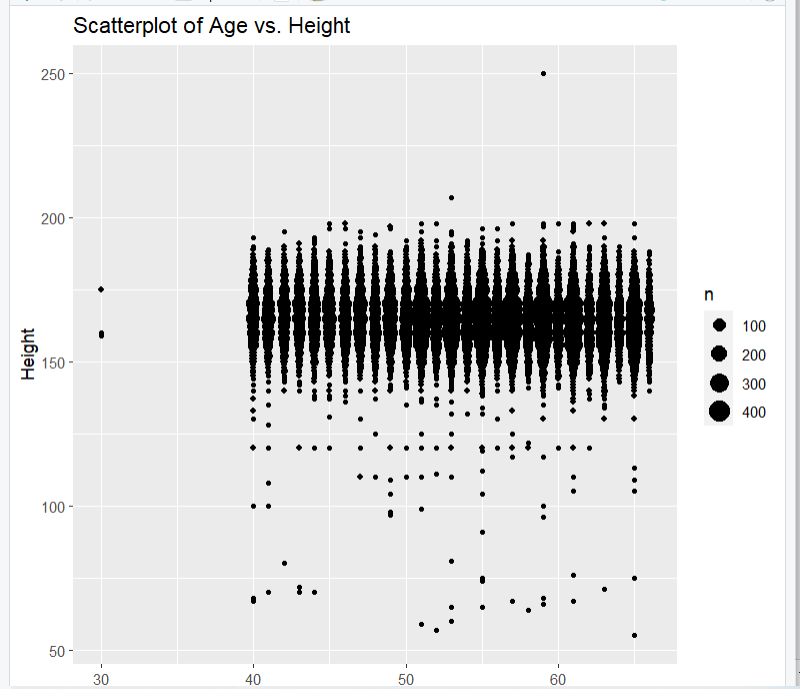
geom\_point() +

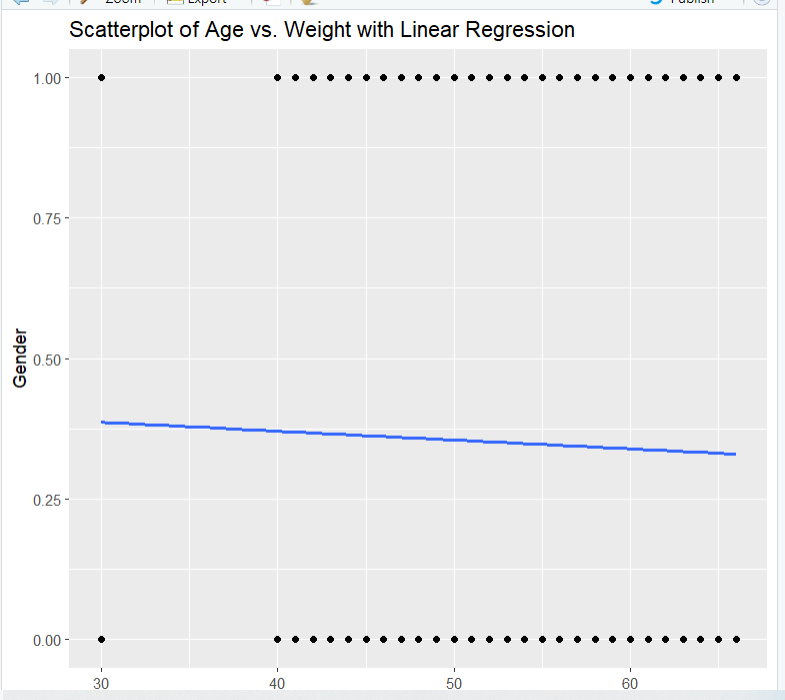
geom\_smooth(method="lm", se=FALSE) +

labs(title="Scatterplot of Age vs. Weight with Linear Regression",

x="Age", y="Gender")

**Output:**





**TILE PLOT**

**Code:**

#TILE PLOT

library(ggcorrplot)

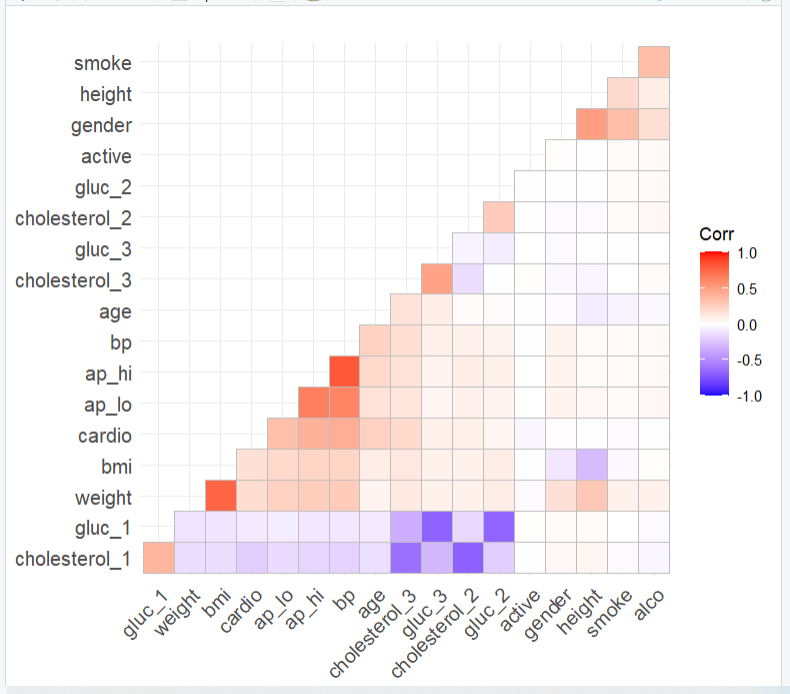
d[] <- lapply(d, as.numeric)

d <- na.omit(d)

corr\_matrix <- cor(d)

ggcorrplot(corr\_matrix, hc.order = TRUE, type = "lower", lab = FALSE)

**Output:**



**HEXBIN**

**Code:**

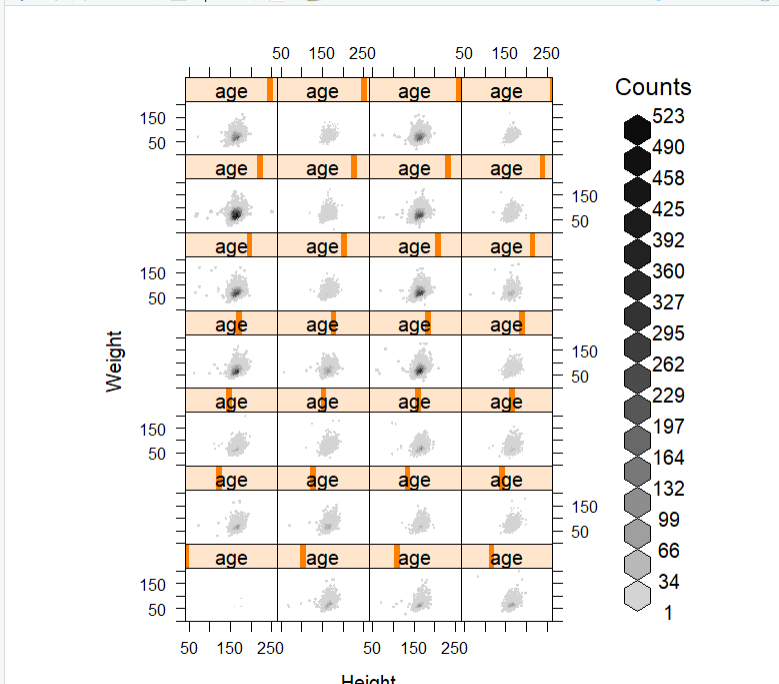
#HEXBIN

library(hexbin)

hexbinplot(weight ~ height | age,

data = d, xlab = "Height", ylab = "Weight")

**Output:**



**Visualizing the combination of a categorical and a continuous variable.**

**Code:**

#Visualizing the combination of a categorical and a continuous variable

my\_colors <- c("#E69F00", "#56B4E9")

boxplot(age ~ gender, data = d,

xlab = "Gender Class", ylab = "Age",

main = "Age Distribution by Cardio Class",

col = my\_colors,

border = "gray",

notch = TRUE,

notchwidth = 0.5,

boxwex = 0.6,

whisklty = 2,

whisklwd = 1.5,

outcex = 0.5,

axes = FALSE)

axis(side = 1, at = 1:2, labels = levels(d$gender))

axis(side = 2, las = 1)

legend("topright", legend = c("Male", "Female"), fill = my\_colors)

#Display the number of points on same bin

boxplot\_data <- d %>%

group\_by(gender, cholesterol\_2) %>%

summarize(count = n())

box\_plot <- ggplot(boxplot\_data, aes(x = cholesterol\_2, y = count,

fill = gender)) +

geom\_boxplot() +

geom\_text(aes(label = count), color = "white", size = 6, fontface = "bold",

position = position\_dodge(width = 0.8))

**# Customize the plot**

box\_plot +

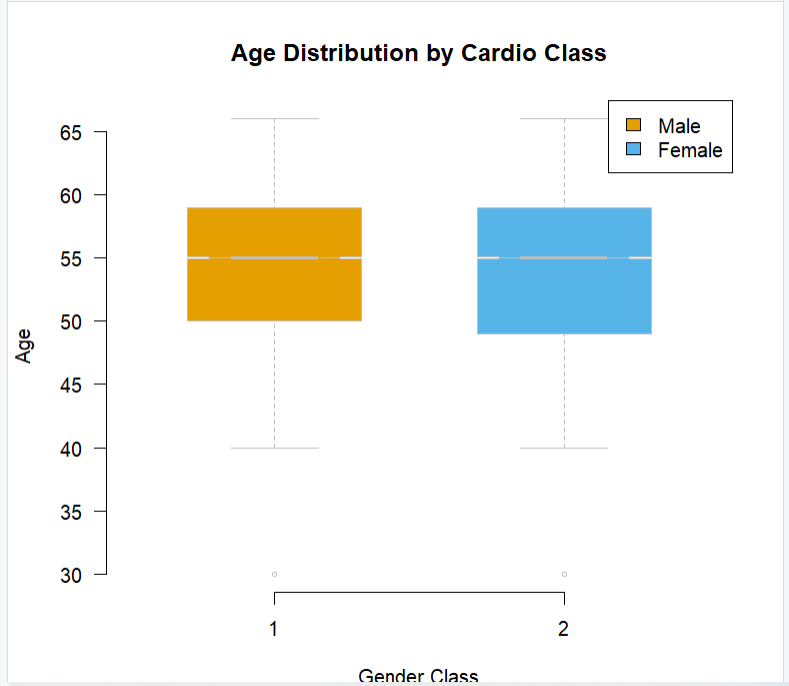
labs(x = "Cholesterol level", y = "Count",

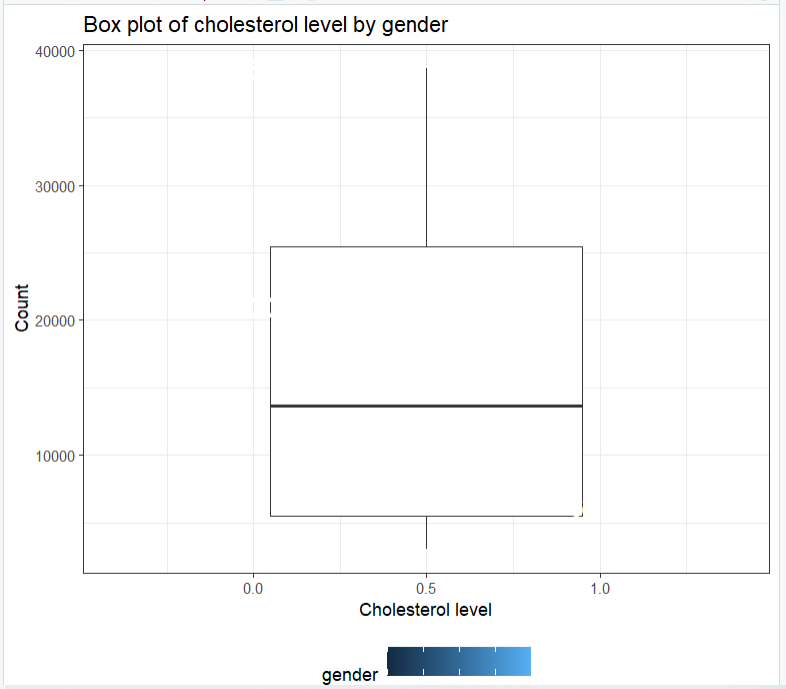
title = "Box plot of cholesterol level by gender") +

theme\_bw() +

theme(legend.position = "bottom")

**Output:**

****

****