**UNIT 5: Regression Analysis**

**1. Box Plot of Student Scores**

library(ggplot2)

set.seed(123) # For reproducibility

student\_scores <- data.frame(

Class = rep(c("Class A", "Class B", "Class C", "Class D", "Class E"), each = 20),

Score = c(

rnorm(20, mean = 75, sd = 10),

rnorm(20, mean = 80, sd = 12),

rnorm(20, mean = 70, sd = 8),

rnorm(20, mean = 85, sd = 9),

rnorm(20, mean = 78, sd = 11)

)

)

ggplot(student\_scores, aes(x = Class, y = Score, fill = Class)) +

geom\_boxplot(outlier.color = "red", outlier.shape = 16, outlier.size = 3) +

scale\_fill\_brewer(palette = "Set2") +

labs(title = "Comparison of Student Scores Across Classes",

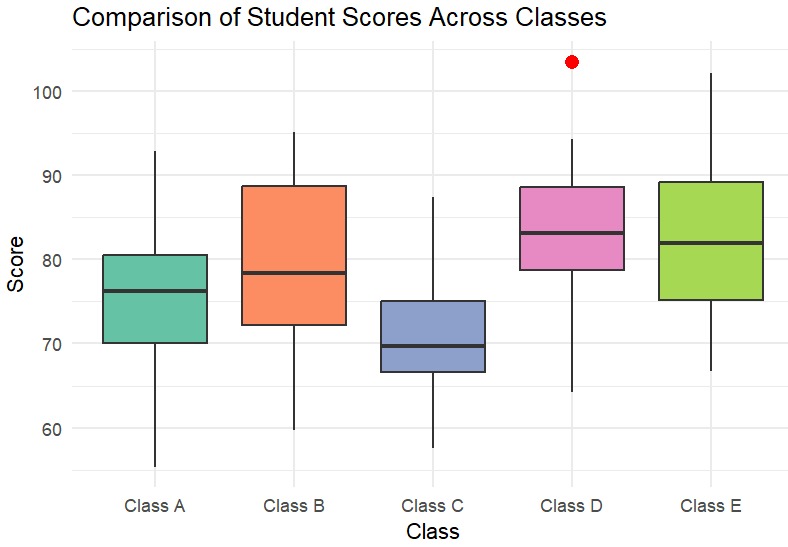
x = "Class",

y = "Score") +

theme\_minimal() +

theme(legend.position = "none")

**Output:**

****

**2. Linear Regression Visualization**

library(ggplot2)

set.seed(123)

data <- data.frame(

Experience = seq(1, 20, by = 1),

Salary = 30000 + 2500 \* seq(1, 20, by = 1) + rnorm(20, mean = 0, sd = 5000)

)

ggplot(data, aes(x = Experience, y = Salary)) +

geom\_point(color = "deepskyblue", size = 3, alpha = 0.7) +

geom\_smooth(method = "lm", color = "darkmagenta", se = TRUE, fill = "pink") +

labs(title = "Years of Experience vs. Salary",

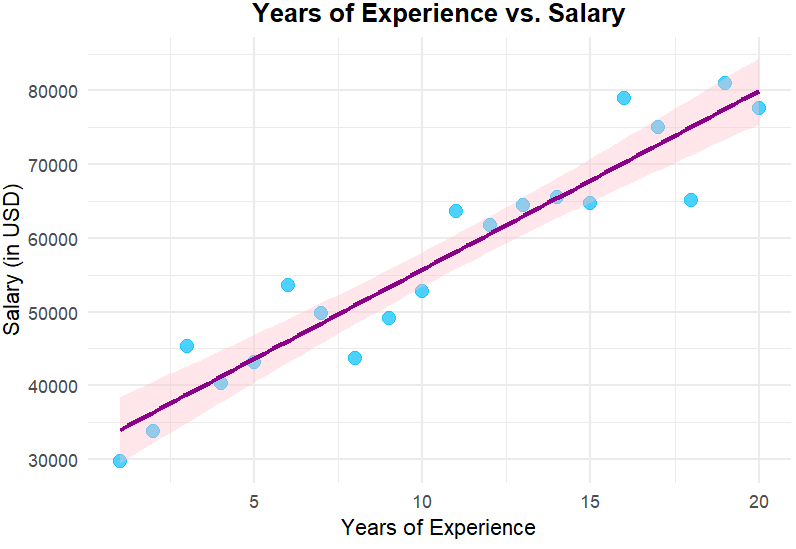
x = "Years of Experience",

y = "Salary (in USD)") +

theme\_minimal() + # Clean theme

theme(plot.title = element\_text(hjust = 0.5, face = "bold"))

**Output:**

****

**3. Histogram Generation**

**set.seed(123)**

**scores <- rnorm(100, mean = 75, sd = 10)**

**hist(scores,**

**main = "Histogram of Student Scores",**

**xlab = "Scores",**

**ylab = "Frequency",**

**col = "skyblue",**

**border = "black")**

**hist(scores,**

**breaks = 15,**

**main = "Histogram with 15 Bins",**

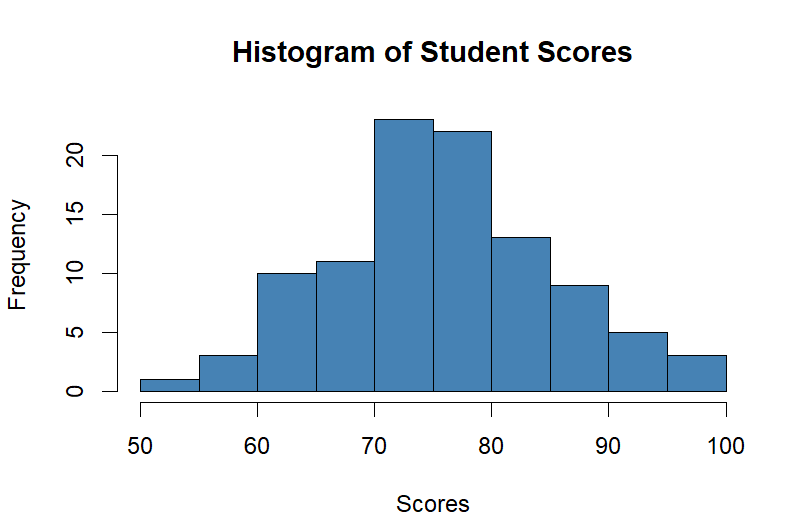
**xlab = "Scores",**

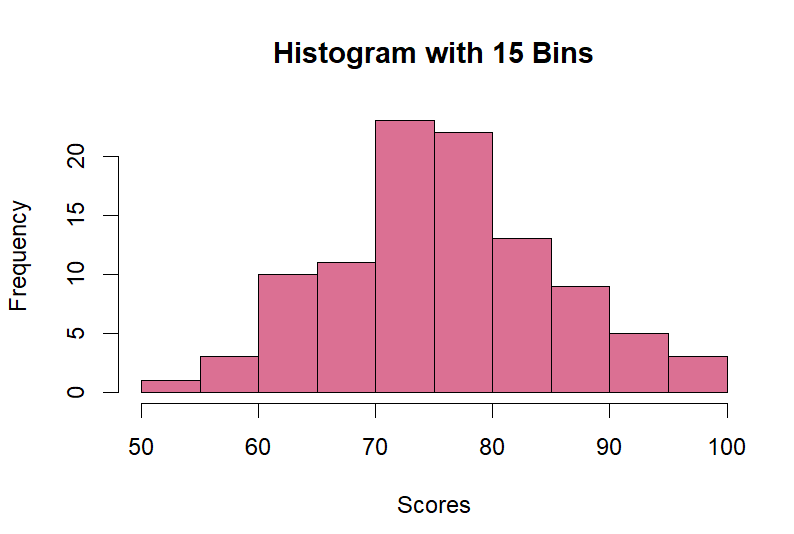
**ylab = "Frequency",**

**col = "lightgreen",**

**border = "black")**

**Output:**

****

****

**4. Add a Legend to a Plot using legend()**

set.seed(123)

x <- 1:10

y1 <- x \* 2 + rnorm(10, sd = 2)

y2 <- x \* 3 + rnorm(10, sd = 2)

plot(x, y1, type = "o", col = "mediumorchid", pch = 16, lty = 1, ylim = c(min(y1, y2), max(y1, y2)),

xlab = "X Values", ylab = "Y Values", main = "Plot with Legend")

lines(x, y2, type = "o", col = "darkred", pch = 17, lty = 2)

legend("topleft",

legend = c("Line 1 (y = 2x)", "Line 2 (y = 3x)"),

col = c("mediumorchid", "darkred"),

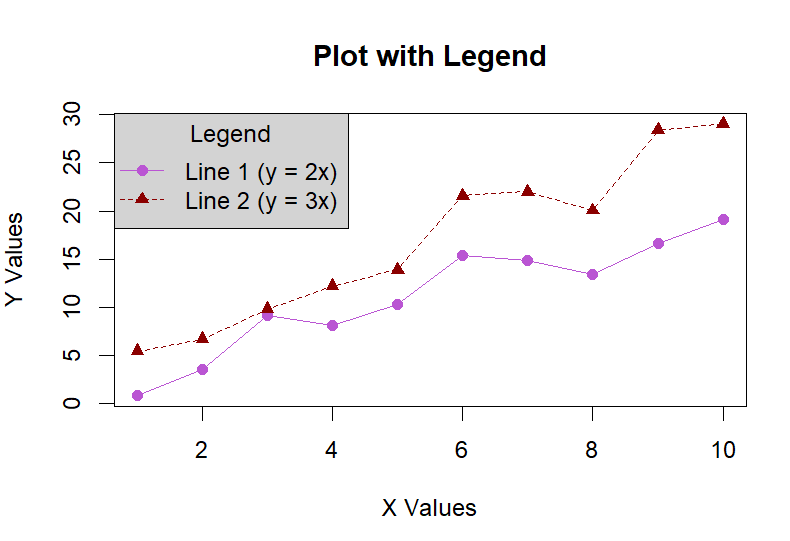
pch = c(16, 17),

lty = c(1, 2),

title = "Legend",

bg = "lightgray")

**Output:**

****

**5. Generate Grouped Histograms using lattice::histogram()**

library(lattice)

set.seed(123)

data <- data.frame(

Group = rep(c("Male", "Female"), each = 50),

Score = c(rnorm(50, mean = 75, sd = 10),

rnorm(50, mean = 80, sd = 12))

)

histogram(~Score | Group,

data = data,

col = "mediumpurple",

border = "black",

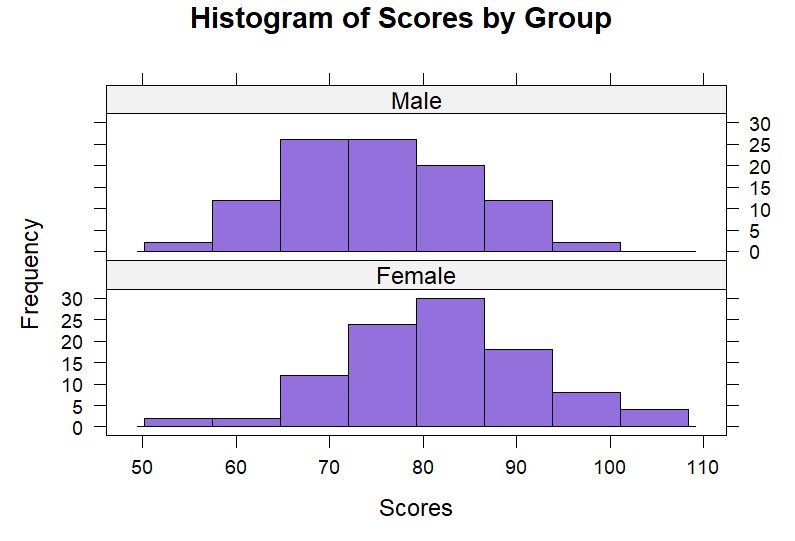
main = "Histogram of Scores by Group",

xlab = "Scores",

ylab = "Frequency",

layout = c(1, 2))

**Output:**

****

**6. Visualize Data Distribution Using Density Plots**

library(ggplot2)

set.seed(123)

data <- data.frame(

Score = rnorm(200, mean = 75, sd = 10)

)

ggplot(data, aes(x = Score)) +

geom\_density(fill = "thistle", alpha = 0.6, color = "indianred") + # Density plot

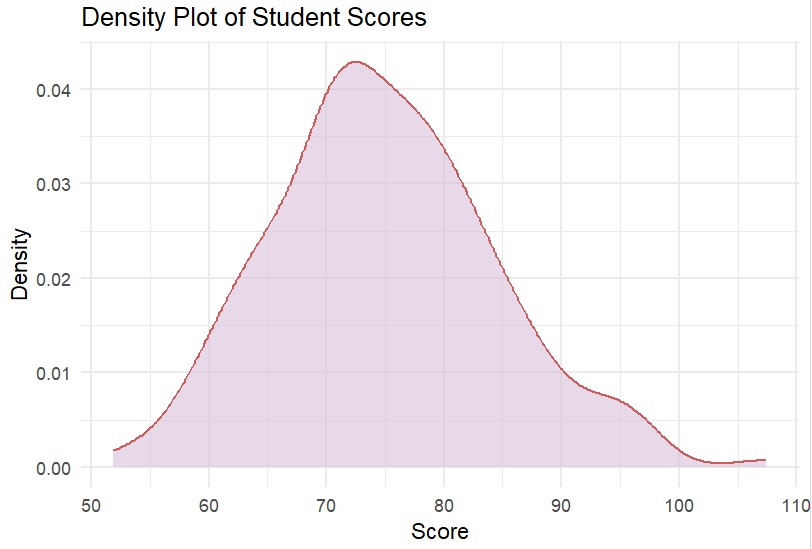
labs(title = "Density Plot of Student Scores",

x = "Score",

y = "Density") +

theme\_minimal()

**Output:**

****

**7. Interactive Bar Plot for Cylinder Frequency**

library(plotly)

library(dplyr)

data(mtcars)

cyl\_freq <- mtcars %>%

count(cyl)

plot <- plot\_ly(cyl\_freq,

x = ~cyl,

y = ~n,

type = "bar",

marker = list(color = "plum")) %>%

layout(title = "Frequency of Cylinders in mtcars Dataset",

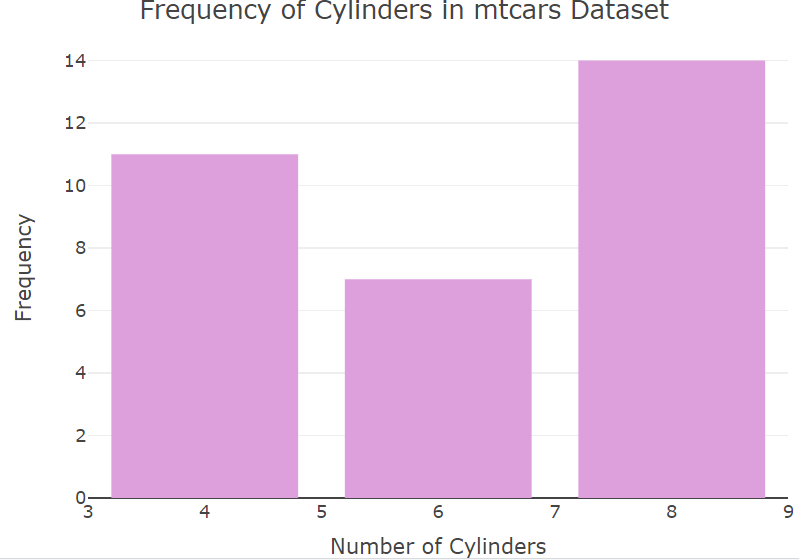
xaxis = list(title = "Number of Cylinders"),

yaxis = list(title = "Frequency"),

bargap = 0.2)

plot

**Output:**

****

**8. Simple Linear Regression Analysis**

library(ggplot2)

data(mtcars)

model <- lm(mpg ~ wt, data = mtcars)

summary(model)

ggplot(mtcars, aes(x = wt, y = mpg)) +

geom\_point(color = "midnightblue", size = 3, alpha = 0.6) +

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

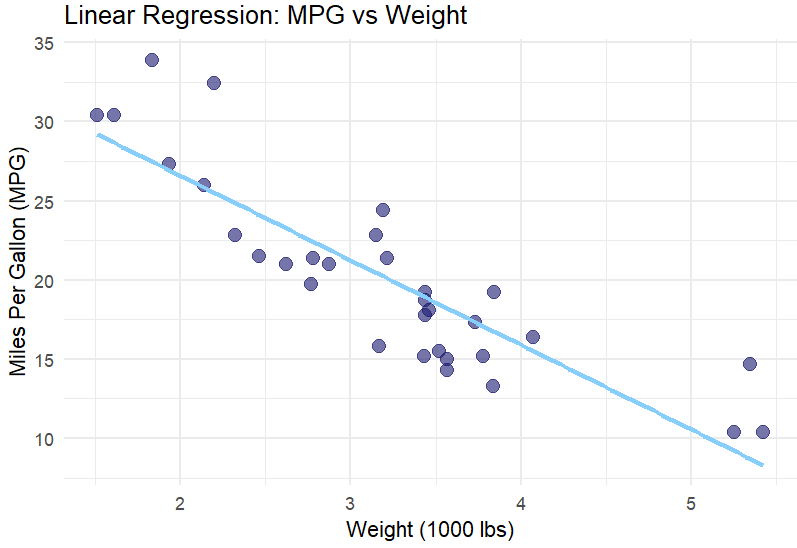
labs(title = "Linear Regression: MPG vs Weight",

x = "Weight (1000 lbs)",

y = "Miles Per Gallon (MPG)") +

theme\_minimal()

**Output:**

****

**9. Scatter Plot for Sepal Dimensions**

library(ggplot2)

data(iris)

ggplot(iris, aes(x = Sepal.Length, y = Sepal.Width, color = Species)) +

geom\_point(size = 3, alpha = 0.7) +

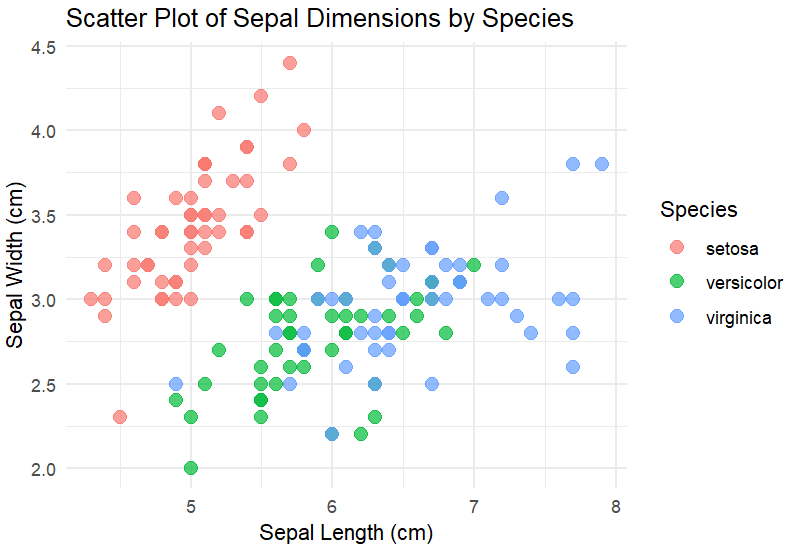
labs(title = "Scatter Plot of Sepal Dimensions by Species",

x = "Sepal Length (cm)",

y = "Sepal Width (cm)") +

theme\_minimal()

**Output:**

****

**10. Multiple Linear Regression to Predict House Prices**

library(ggplot2)

house\_data <- data.frame(

Area = c(1500, 1800, 2400, 3000, 3500, 4000, 1200, 2500, 2700, 3200),

Bedrooms = c(3, 4, 3, 5, 4, 6, 2, 3, 4, 5),

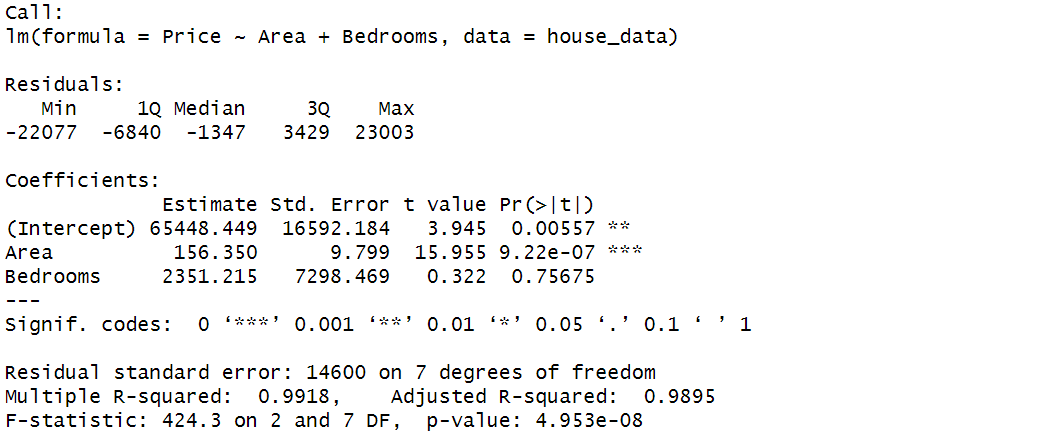
Price = c(300000, 350000, 450000, 550000, 600000, 700000, 250000, 480000, 520000, 580000)

)

model <- lm(Price ~ Area + Bedrooms, data = house\_data)

summary(model)

**Output:**

****

**11. Histogram of Survey Participants' Ages**

library(ggplot2)

survey\_data <- data.frame(

Age = c(25, 30, 35, 40, 45, 50, 55, 30, 25, 35, 45, 60, 40, 50, 55, 35, 25, 30, 45, 50)

)

hist\_plot <- ggplot(survey\_data, aes(x = Age)) +

geom\_histogram(binwidth = 5, fill = "maroon", color = "black", alpha = 0.7) +

labs(title = "Histogram of Ages in the Survey",

x = "Age Groups",

y = "Frequency") +

theme\_minimal()

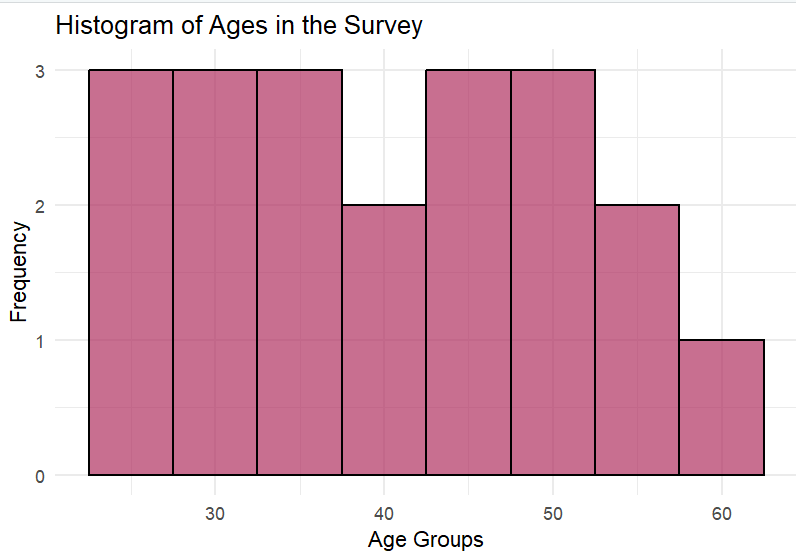
print(hist\_plot)

age\_freq <- table(cut(survey\_data$Age, breaks = seq(20, 65, by = 5))) # Categorize ages in 5-year bins

most\_frequent\_group <- names(age\_freq[which.max(age\_freq)])

cat("The age group with the highest frequency is:", most\_frequent\_group, "\n")

**Output:**

****

**12. Pie Chart of Species Proportions in the iris Dataset**

library(ggplot2)

data(iris)

species\_count <- table(iris$Species)

species\_df <- as.data.frame(species\_count)

colnames(species\_df) <- c("Species", "Count")

ggplot(species\_df, aes(x = "", y = Count, fill = Species)) +

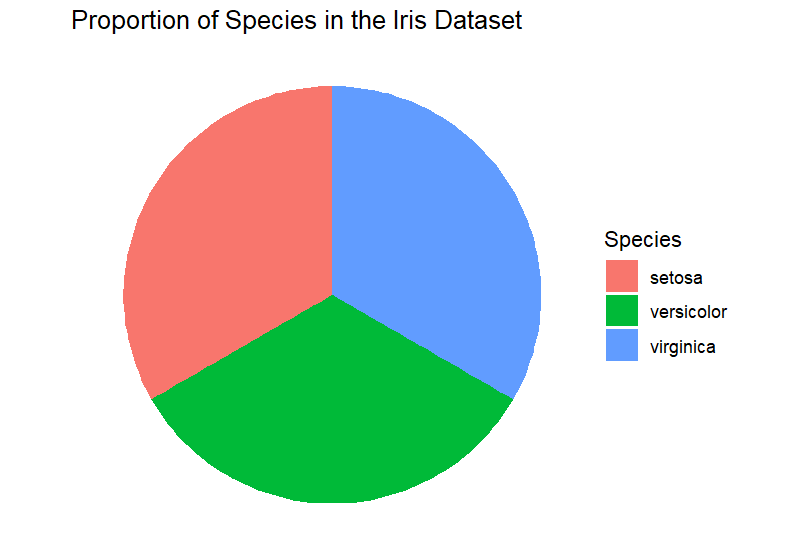
geom\_bar(stat = "identity", width = 1) +

coord\_polar(theta = "y") +

labs(title = "Proportion of Species in the Iris Dataset") +

theme\_void()

**Output:**

****

**13. Analyzing the ChickWeight Dataset**

data(ChickWeight)

library(ggplot2)

ggplot(ChickWeight, aes(x = weight)) +

geom\_histogram(binwidth = 25, fill = "skyblue", color = "black", alpha = 0.7) +

labs(title = "Distribution of Chick Weights", x = "Weight (grams)", y = "Frequency") +

theme\_minimal()

missing\_values <- sum(is.na(ChickWeight$weight))

cat("Number of missing values in weight:", missing\_values, "\n")

median\_weight <- median(ChickWeight$weight, na.rm = TRUE)

cat("Median weight of chickens:", median\_weight, "\n")

ggplot(ChickWeight, aes(y = weight)) +

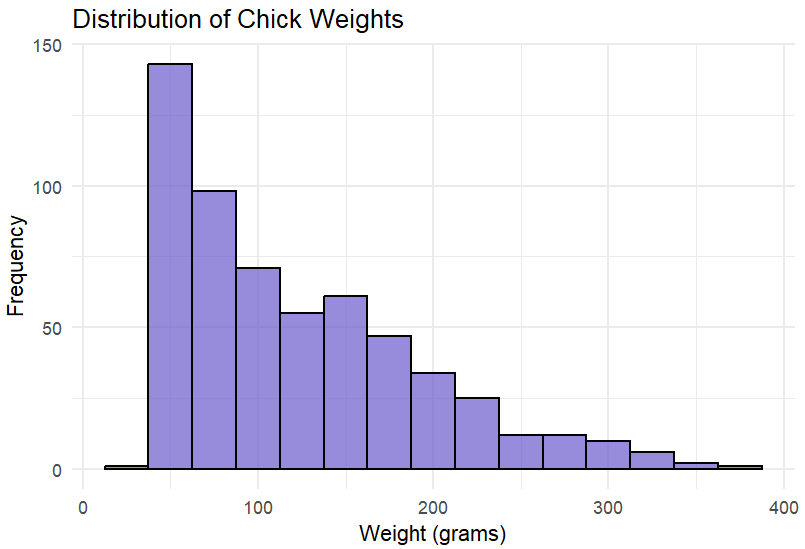
geom\_boxplot(fill = "orange", color = "black", alpha = 0.7) +

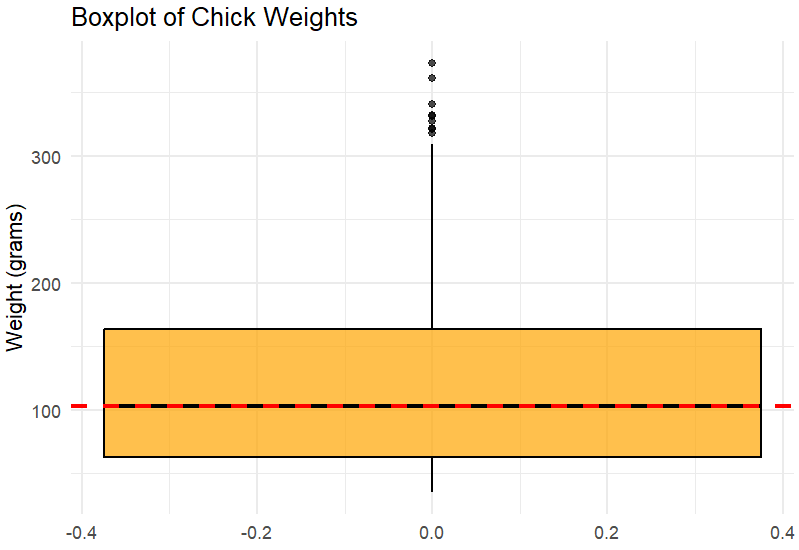
labs(title = "Boxplot of Chick Weights", y = "Weight (grams)") +

geom\_hline(yintercept = median\_weight, linetype = "dashed", color = "red", size = 1) +

theme\_minimal()

**Output:**

****

****

**14. EDA on a Wine Dataset**

library(ggplot2)

library(dplyr)

library(corrplot)

library(ggcorrplot)

library(tidyr)

wine\_data <- read.csv("winequality-red.csv", sep = ";") # Update the path if needed

cat("Summary of the dataset:\n")

summary(wine\_data)

cat("\nStructure of the dataset:\n")

str(wine\_data)

missing\_values <- colSums(is.na(wine\_data))

cat("\nMissing values in each column:\n")

print(missing\_values)

wine\_data <- wine\_data %>% drop\_na()

ggplot(wine\_data, aes(x = alcohol)) +

geom\_histogram(binwidth = 0.5, fill = "blue", color = "black", alpha = 0.7) +

labs(title = "Distribution of Alcohol Content", x = "Alcohol", y = "Frequency") +

theme\_minimal()

ggplot(wine\_data, aes(x = quality)) +

geom\_bar(fill = "red", color = "black", alpha = 0.7) +

labs(title = "Distribution of Wine Quality", x = "Quality", y = "Count") +

theme\_minimal()

cor\_matrix <- cor(wine\_data)

ggcorrplot(cor\_matrix, lab = TRUE, title = "Correlation Heatmap of Wine Attributes")

pairs(wine\_data[, c("alcohol", "citric.acid", "density", "quality")],

main = "Pair Plot of Selected Wine Features", col = "blue")

ggplot(wine\_data, aes(x = factor(quality), y = alcohol, fill = factor(quality))) +

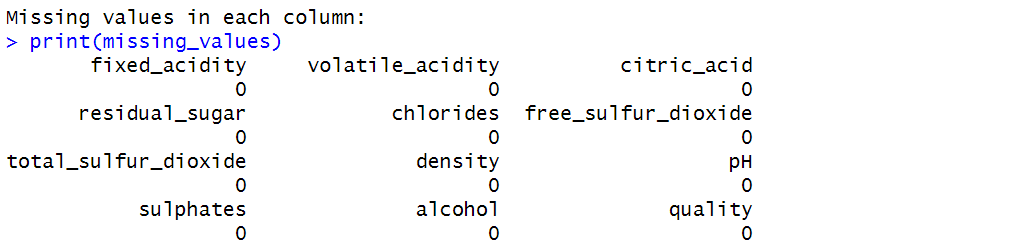
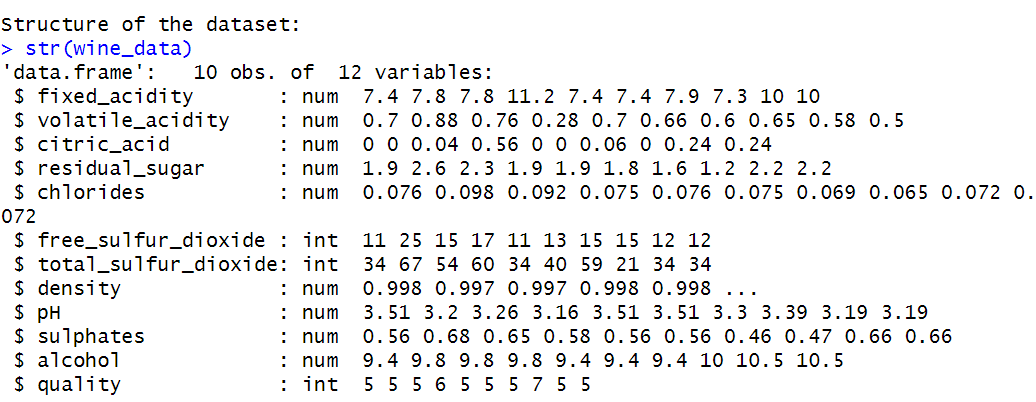
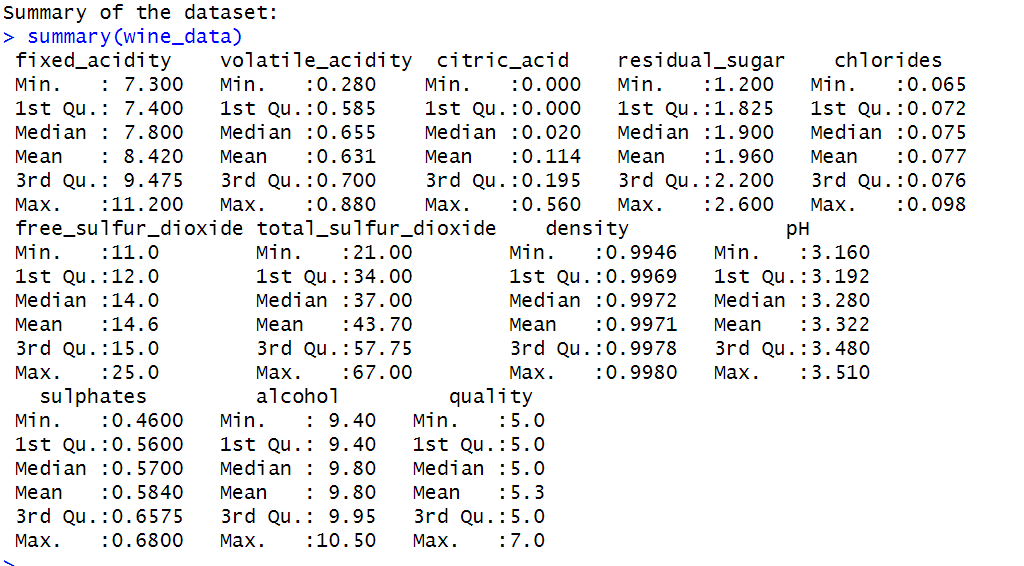
geom\_boxplot() +

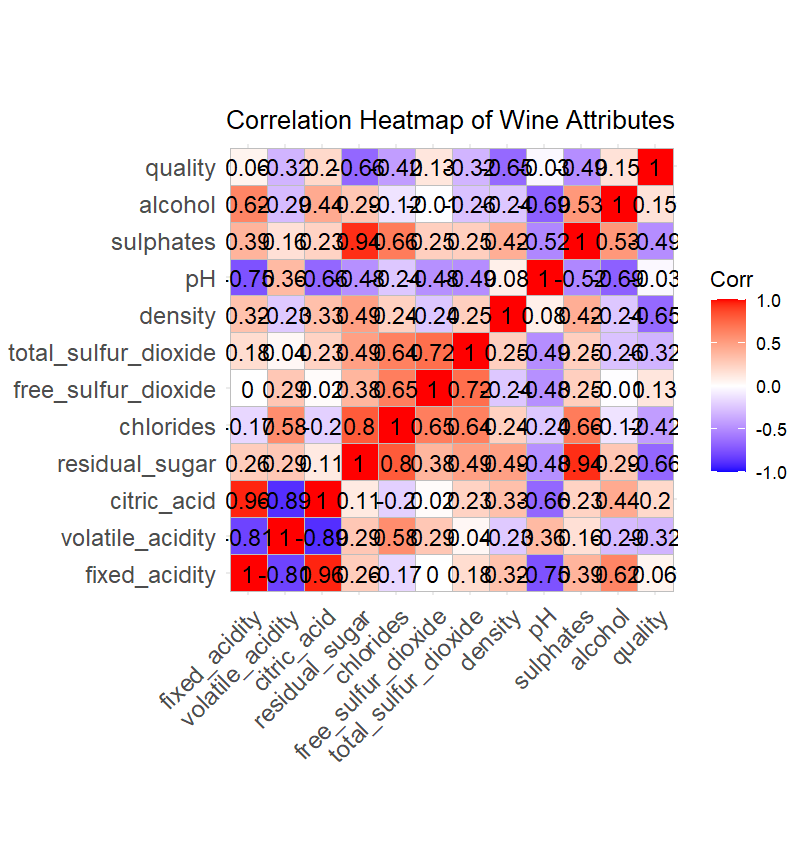
labs(title = "Alcohol Content by Wine Quality", x = "Quality", y = "Alcohol Content") +

theme\_minimal()

cat("\nExploratory Data Analysis (EDA) completed successfully!\n")

**Output:**

****

****

**15. Air Quality dataset**

data("airquality")

if (is.data.frame(airquality)) {

cat("airquality is a data frame.\n")

} else {

cat("airquality is NOT a data frame.\n")

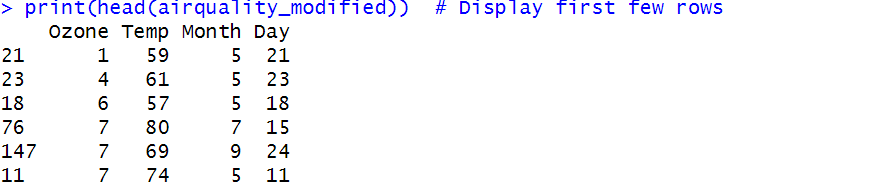
}

airquality\_ordered <- airquality[order(airquality$Ozone, airquality$Solar.R, na.last = TRUE), ]

airquality\_modified <- subset(airquality\_ordered, select = -c(Solar.R, Wind))

print(head(airquality\_modified))

**Output:**

****

**16. Titanic Data Analysis**

if (!require(titanic)) install.packages("titanic", dependencies=TRUE)

if (!require(ggplot2)) install.packages("ggplot2", dependencies=TRUE)

if (!require(dplyr)) install.packages("dplyr", dependencies=TRUE)

library(ggplot2)

library(dplyr)

library(titanic)

titanic\_data <- titanic::titanic\_train

titanic\_data$Survived <- as.factor(titanic\_data$Survived)

titanic\_data$Pclass <- as.factor(titanic\_data$Pclass)

titanic\_data$Sex <- as.factor(titanic\_data$Sex)

ggplot(titanic\_data, aes(x = Pclass, fill = Survived)) +

geom\_bar(position = "dodge") +

labs(title = "Survival Count by Passenger Class",

x = "Passenger Class", y = "Count") +

theme\_minimal()

ggplot(titanic\_data, aes(x = Pclass, fill = Survived)) +

geom\_bar(position = "dodge") +

facet\_wrap(~Sex) +

labs(title = "Survival Count by Passenger Class and Gender",

x = "Passenger Class", y = "Count") +

theme\_minimal()

ggplot(titanic\_data, aes(x = Age)) +

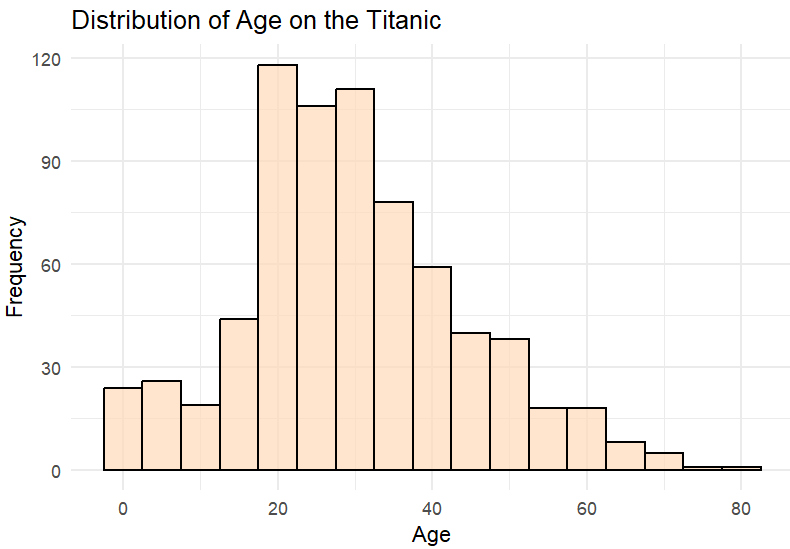
geom\_histogram(binwidth = 5, fill = "peachpuff", color = "black", alpha = 0.7, na.rm = TRUE) +

labs(title = "Distribution of Age on the Titanic",

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

theme\_minimal()

**Output:**

****

**17. Salary Distribution Analysis**

if (!require(dplyr)) install.packages("dplyr", dependencies=TRUE)

library(dplyr)

salary\_data <- data.frame(

Employee\_ID = 1:15, # Employee IDs

Salary = c(50000, 52000, 48000, 55000, 60000, 58000, 62000, 49000,

51000, 57000, 54000, 59000, 56000, 53000, 60000) # Salaries

)

mean\_salary <- mean(salary\_data$Salary)

median\_salary <- median(salary\_data$Salary)

get\_mode <- function(x) {

unique\_x <- unique(x)

unique\_x[which.max(tabulate(match(x, unique\_x)))]

}

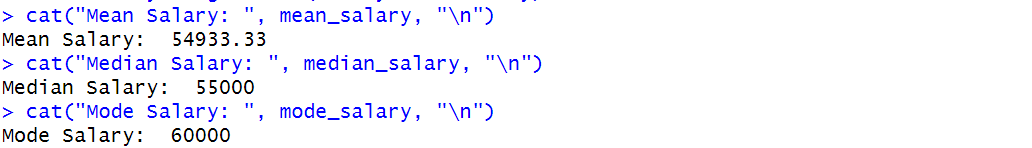
mode\_salary <- get\_mode(salary\_data$Salary)

cat("Mean Salary: ", mean\_salary, "\n")

cat("Median Salary: ", median\_salary, "\n")

cat("Mode Salary: ", mode\_salary, "\n")

**Output:**

****

**18. Air Quality Dataset Manipulation**

**data("airquality")**

**if (is.data.frame(airquality)) {**

**cat("The airquality dataset is a data frame.\n")**

**} else {**

**cat("The airquality dataset is NOT a data frame.\n")**

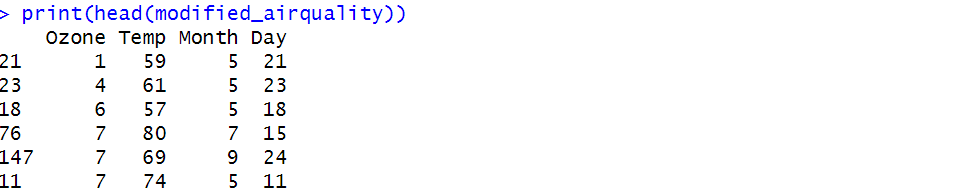
**}**

**ordered\_airquality <- airquality[order(airquality$Ozone, airquality$Solar.R, na.last = TRUE), ]**

**modified\_airquality <- subset(ordered\_airquality, select = -c(Solar.R, Wind))**

**print(head(modified\_airquality))**

**Output:**

****

**19. Line Chart & Bar Chart**

library(ggplot2)

data("pressure")

ggplot(data = pressure, aes(x = temperature, y = pressure)) +

geom\_line(color = "steelblue", size = 1) +

geom\_point(color = "deeppink", size = 2) +

ggtitle("Line Chart of Pressure vs Temperature") +

xlab("Temperature (°C)") +

ylab("Pressure (mm of Mercury)") +

theme\_minimal()

ggplot(data = pressure, aes(x = factor(temperature), y = pressure)) +

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

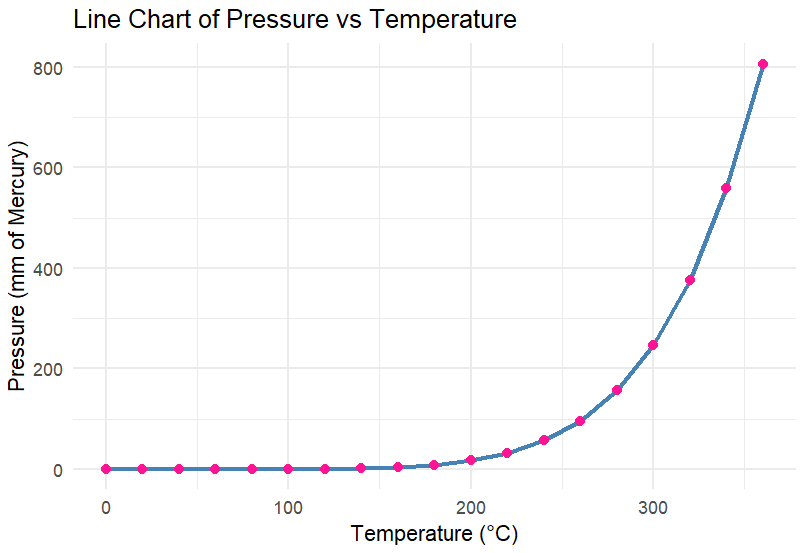
ggtitle("Bar Chart of Pressure vs Temperature") +

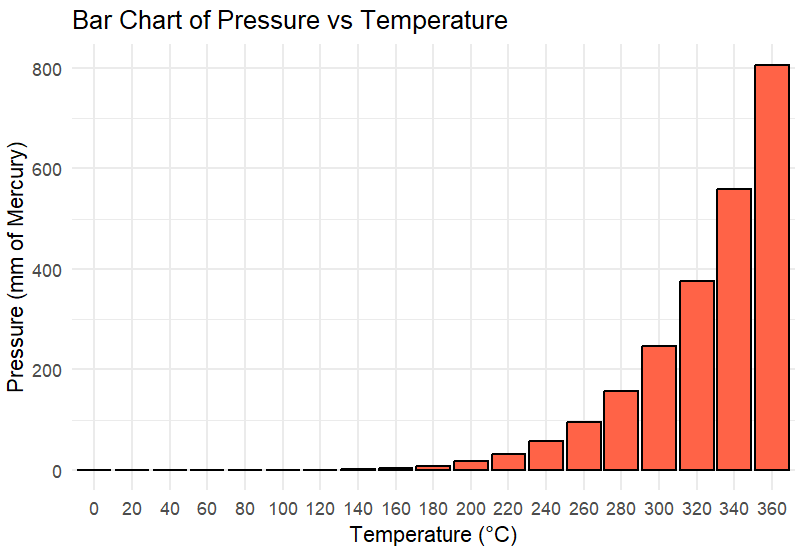
xlab("Temperature (°C)") +

ylab("Pressure (mm of Mercury)") +

theme\_minimal()

**Output:**

****

****

**20. Multiple Regression Analysis**

data("mtcars")

model <- lm(mpg ~ wt + hp, data = mtcars)

summary(model)

library(ggplot2)

ggplot(mtcars, aes(x = wt, y = mpg, color = hp)) +

geom\_point(size = 3) +

geom\_smooth(method = "lm", formula = y ~ x, se = FALSE, color = "blue") +

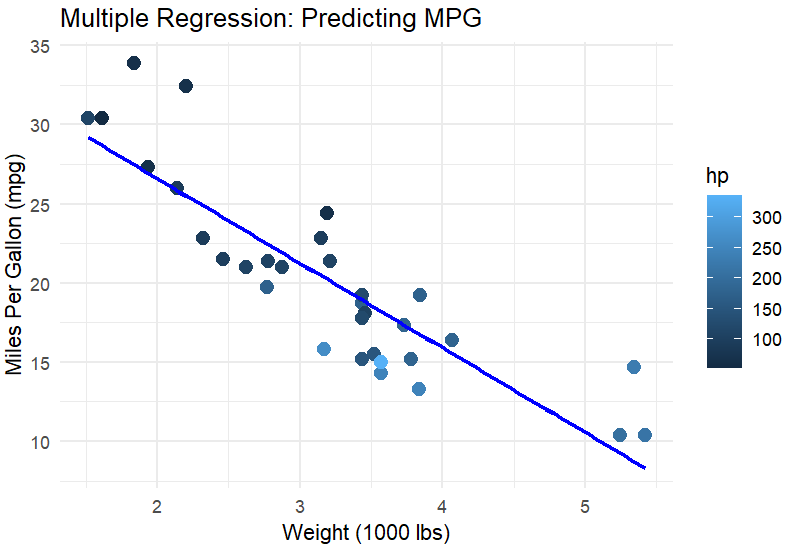
ggtitle("Multiple Regression: Predicting MPG") +

xlab("Weight (1000 lbs)") +

ylab("Miles Per Gallon (mpg)") +

theme\_minimal()

**Output:**

****