**UNIT 4: EDA Analysis**

**1. mtcars**

data(mtcars)

mean\_mpg <- mean(mtcars$mpg)

median\_mpg <- median(mtcars$mpg)

get\_mode <- function(v) {

uniq\_vals <- unique(v)

freq <- table(v)

mode\_vals <- uniq\_vals[freq == max(freq)]

return(mode\_vals)

}

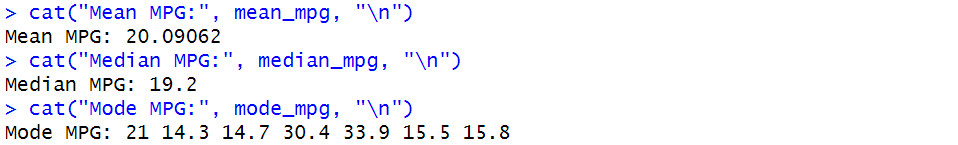
mode\_mpg <- get\_mode(mtcars$mpg)

cat("Mean MPG:", mean\_mpg, "\n")

cat("Median MPG:", median\_mpg, "\n")

cat("Mode MPG:", mode\_mpg, "\n")

**Output:**



**2. Chick Weight**

data(ChickWeight)

sd\_weight <- sd(ChickWeight$weight)

cat("Standard Deviation of Chickens' Weights:", sd\_weight, "\n")

cat("The ChickWeight dataset does not contain gender information.\n")

heaviest\_chicken <- ChickWeight[which.max(ChickWeight$weight), ]

cat("Heaviest Chicken:\n")

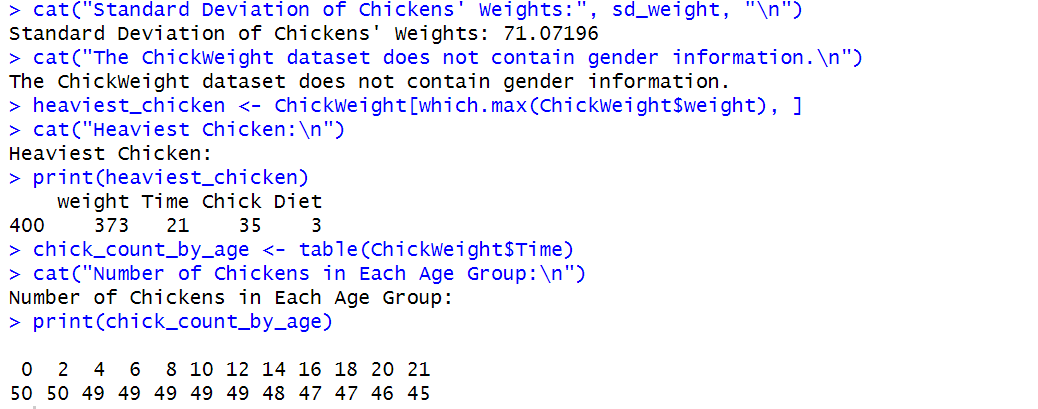
print(heaviest\_chicken)

chick\_count\_by\_age <- table(ChickWeight$Time)

cat("Number of Chickens in Each Age Group:\n")

print(chick\_count\_by\_age)

**Output:**

****

**3. Sepal Width**

data(iris)

quartiles <- quantile(iris$Sepal.Width, probs = c(0.25, 0.75))

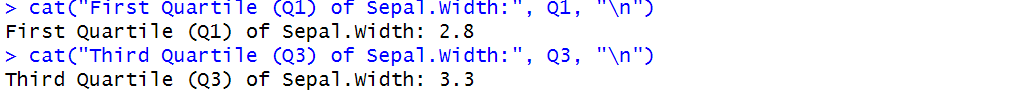
Q1 <- quartiles[1] # First quartile (25th percentile)

Q3 <- quartiles[2] # Third quartile (75th percentile)

cat("First Quartile (Q1) of Sepal.Width:", Q1, "\n")

cat("Third Quartile (Q3) of Sepal.Width:", Q3, "\n")

**Output:**

****

**4. mtcars mean**

data(mtcars)

mean\_mpg <- mean(mtcars$mpg)

cat("Mean of mpg in mtcars dataset:", mean\_mpg, "\n")

**Output:**

****

**5. Read csv**

data <- data.frame(

ID = c(1, 2, 3, 4),

Name = c("Alice", "Bob", "Charlie", "David"),

Age = c(25, 30, 35, 40),

Score = c(90, 85, 88, 92)

)

write.csv(data, "sample\_data.csv", row.names = FALSE)

read\_data <- read.csv("sample\_data.csv")

print(read\_data)

**Output:**

****

**6. Sepal Width**

data(iris)

quartiles <- quantile(iris$Sepal.Width, probs = c(0.25, 0.75))

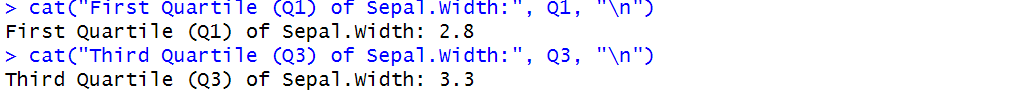
Q1 <- quartiles[1] # First quartile (25th percentile)

Q3 <- quartiles[2] # Third quartile (75th percentile)

cat("First Quartile (Q1) of Sepal.Width:", Q1, "\n")

cat("Third Quartile (Q3) of Sepal.Width:", Q3, "\n")

**Output:**

****

**7. Width and length of rectangle**

calculate\_area <- function(length, width) {

return(length \* width)

}

rectangles <- data.frame(

length = c(5, 8, NA, 6, 10),

width = c(3, NA, 4, 7, 2)

)

rectangles$length[is.na(rectangles$length)] <- 1

rectangles$width[is.na(rectangles$width)] <- 1

rectangles$area <- mapply(calculate\_area, rectangles$length, rectangles$width)

print(rectangles)

**Output:**

****

**8. Rectangle areas**

library(ggplot2)

calculate\_area <- function(length, width) {

return(length \* width)

}

lengths <- rep(1:10, each = 10)

widths <- rep(1:10, times = 10)

areas <- mapply(calculate\_area, lengths, widths) # Compute area

rectangles <- data.frame(Length = lengths, Width = widths, Area = areas)

ggplot(rectangles, aes(x = Length, y = Width, fill = Area)) +

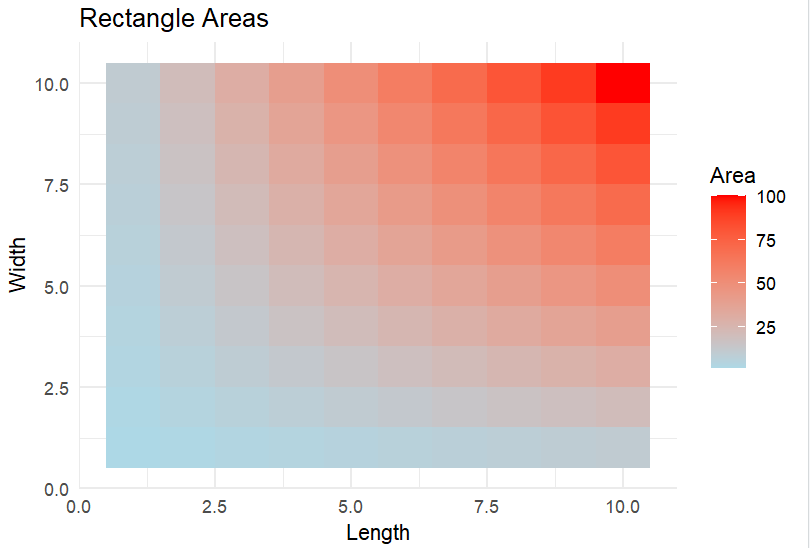
geom\_tile() +

scale\_fill\_gradient(low = "lightblue", high = "red") +

labs(title = "Rectangle Areas", x = "Length", y = "Width", fill = "Area") +

theme\_minimal()

**Output:**

****

**9. minimum, maximum, range, variance, and standard deviation of the test scores**

scores <- c(85, 90, 78, 92, 88, 76, 95, 89, 84, 91,

87, 82, 90, 93, 88, 85, 77, 94, 80, 79)

min\_score <- min(scores)

max\_score <- max(scores)

range\_score <- max\_score - min\_score

variance\_score <- var(scores)

sd\_score <- sd(scores)

cat("Minimum Score:", min\_score, "\n")

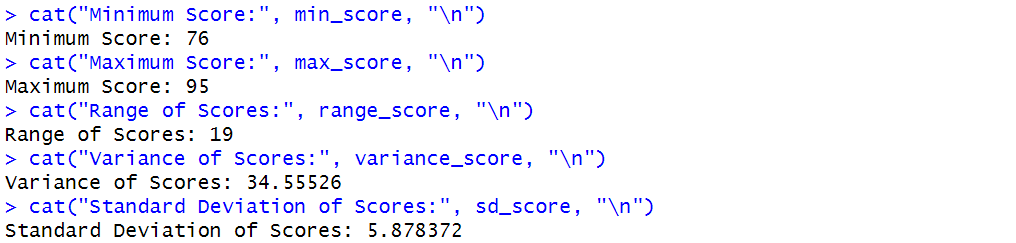
cat("Maximum Score:", max\_score, "\n")

cat("Range of Scores:", range\_score, "\n")

cat("Variance of Scores:", variance\_score, "\n")

cat("Standard Deviation of Scores:", sd\_score, "\n")

**Output:**

****

**10. covariance and correlation between advertising expenditure and sales revenue**

advertising <- c(1000, 1500, 2000, 2500, 3000, 3500, 4000, 4500, 5000, 5500)

sales <- c(15000, 18000, 21000, 24000, 27000, 30000, 33000, 36000, 39000, 42000)

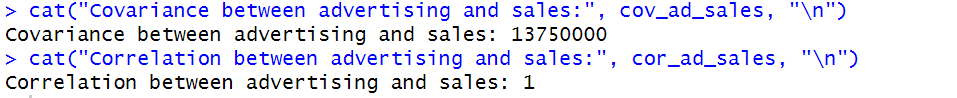
cov\_ad\_sales <- cov(advertising, sales)

cor\_ad\_sales <- cor(advertising, sales)

cat("Covariance between advertising and sales:", cov\_ad\_sales, "\n")

cat("Correlation between advertising and sales:", cor\_ad\_sales, "\n")

**Output:**

****

**11. Dataframe Stock X Stock Y**

stock\_data <- data.frame(

Stock\_X = c(0.02, 0.05, -0.01, 0.04, 0.03, 0.06, -0.02, 0.07, 0.01, 0.05),

Stock\_Y = c(0.03, 0.04, 0.00, 0.05, 0.02, 0.07, -0.01, 0.06, 0.02, 0.04)

)

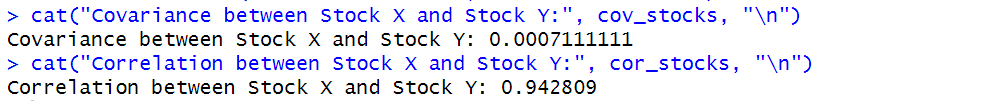
cov\_stocks <- cov(stock\_data$Stock\_X, stock\_data$Stock\_Y)

cor\_stocks <- cor(stock\_data$Stock\_X, stock\_data$Stock\_Y)

cat("Covariance between Stock X and Stock Y:", cov\_stocks, "\n")

cat("Correlation between Stock X and Stock Y:", cor\_stocks, "\n")

**Output:**

****

**12. Statistical measures**

library(dplyr)

performance\_scores <- c(85, 90, 78, 92, 88, 76, 95, 89, 84, 91,

87, 82, 90, 93, 88, 85, 77, 94, 80, 79)

get\_mode <- function(v) {

uniq\_vals <- unique(v)

freq <- table(v)

mode\_vals <- uniq\_vals[freq == max(freq)]

return(mode\_vals)

}

data <- data.frame(Performance\_Scores = performance\_scores)

summary\_stats <- data %>%

summarise(

Mean = mean(Performance\_Scores),

Median = median(Performance\_Scores),

Mode = toString(get\_mode(Performance\_Scores)),

Minimum = min(Performance\_Scores),

Maximum = max(Performance\_Scores),

Range = max(Performance\_Scores) - min(Performance\_Scores),

Q1 = quantile(Performance\_Scores, 0.25),

Q3 = quantile(Performance\_Scores, 0.75),

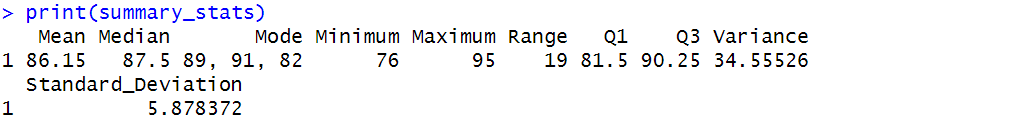
Variance = var(Performance\_Scores),

Standard\_Deviation = sd(Performance\_Scores)

)

print(summary\_stats)

**Output:**

****

**13. Exam score box plot**

library(ggplot2)

exam\_scores <- data.frame(

Class = rep(c("Class A", "Class B", "Class C", "Class D"), each = 10),

Score = c(85, 90, 78, 92, 88, 76, 95, 89, 84, 91,

87, 82, 90, 93, 88, 85, 77, 94, 80, 79,

65, 70, 68, 72, 74, 66, 75, 69, 73, 71,

55, 60, 58, 62, 64, 56, 65, 59, 63, 61)

)

quartiles <- quantile(exam\_scores$Score, probs = c(0.25, 0.5, 0.75))

cat("Q1 (25th percentile):", quartiles[1], "\n")

cat("Median (50th percentile):", quartiles[2], "\n")

cat("Q3 (75th percentile):", quartiles[3], "\n")

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

geom\_boxplot() +

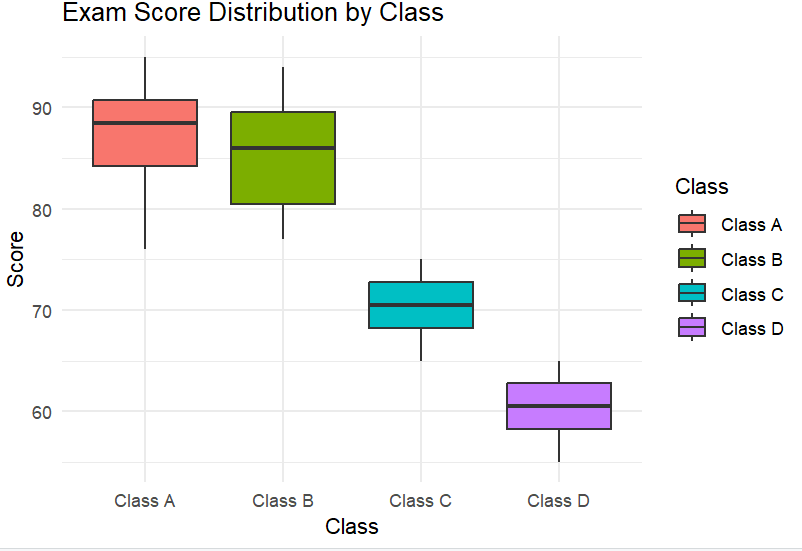
labs(title = "Exam Score Distribution by Class",

x = "Class",

y = "Score") +

theme\_minimal()

**Output:**



**14. EDA of patients**

library(dplyr)

library(moments)

patients\_data <- data.frame(

Age = c(25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95),

Blood\_Pressure = c(120, 125, 130, 135, 140, 145, 150, 155, 160, 165, 170, 175, 180, 185, 190)

)

get\_mode <- function(v) {

uniq\_vals <- unique(v)

freq <- table(v)

mode\_vals <- uniq\_vals[freq == max(freq)]

return(mode\_vals)

}

eda\_summary <- patients\_data %>%

summarise(

Mean\_Age = mean(Age),

Median\_Age = median(Age),

Mode\_Age = toString(get\_mode(Age)),

SD\_Age = sd(Age),

Skewness\_Age = skewness(Age),

Kurtosis\_Age = kurtosis(Age),

Mean\_BP = mean(Blood\_Pressure),

Median\_BP = median(Blood\_Pressure),

Mode\_BP = toString(get\_mode(Blood\_Pressure)),

SD\_BP = sd(Blood\_Pressure),

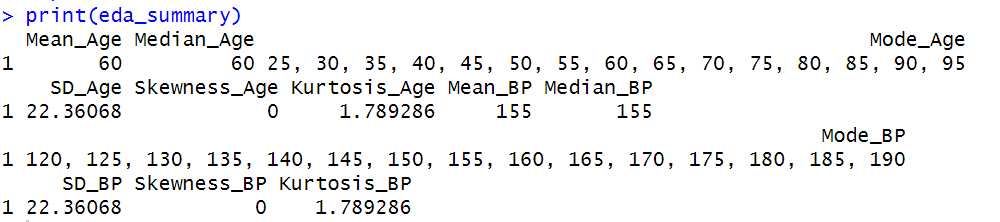
Skewness\_BP = skewness(Blood\_Pressure),

Kurtosis\_BP = kurtosis(Blood\_Pressure)

)

print(eda\_summary)

**Output:**

****

**15. Covariance and Correlation**

sales\_data <- data.frame(

Products\_Sold = c(50, 60, 45, 70, 90, 30, 80, 100, 55, 75),

Price = c(20, 18, 22, 15, 12, 25, 14, 10, 19, 13)

)

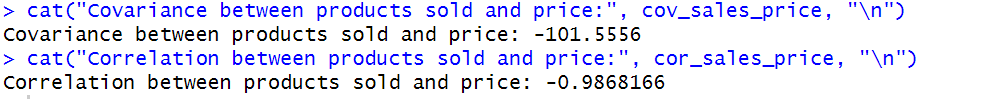
cov\_sales\_price <- cov(sales\_data$Products\_Sold, sales\_data$Price)

cor\_sales\_price <- cor(sales\_data$Products\_Sold, sales\_data$Price)

cat("Covariance between products sold and price:", cov\_sales\_price, "\n")

cat("Correlation between products sold and price:", cor\_sales\_price, "\n")

**Output:**

****

**16. Series of plots**

library(ggplot2)

library(gridExtra)

sales\_data <- data.frame(

Products\_Sold = c(50, 60, 45, 70, 90, 30, 80, 100, 55, 75),

Price = c(20, 18, 22, 15, 12, 25, 14, 10, 19, 13) # Product price per unit

)

hist\_products <- ggplot(sales\_data, aes(x = Products\_Sold)) +

geom\_histogram(binwidth = 10, fill = "blue", color = "black", alpha = 0.7) +

labs(title = "Distribution of Products Sold", x = "Products Sold", y = "Count") +

theme\_minimal()

hist\_prices <- ggplot(sales\_data, aes(x = Price)) +

geom\_histogram(binwidth = 2, fill = "green", color = "black", alpha = 0.7) +

labs(title = "Distribution of Product Prices", x = "Price per Unit", y = "Count") +

theme\_minimal()

boxplot\_products <- ggplot(sales\_data, aes(y = Products\_Sold)) +

geom\_boxplot(fill = "blue", alpha = 0.5) +

labs(title = "Boxplot of Products Sold", y = "Products Sold") +

theme\_minimal()

boxplot\_prices <- ggplot(sales\_data, aes(y = Price)) +

geom\_boxplot(fill = "green", alpha = 0.5) +

labs(title = "Boxplot of Product Prices", y = "Price per Unit") +

theme\_minimal()

scatter\_plot <- ggplot(sales\_data, aes(x = Price, y = Products\_Sold)) +

geom\_point(color = "red", size = 3) +

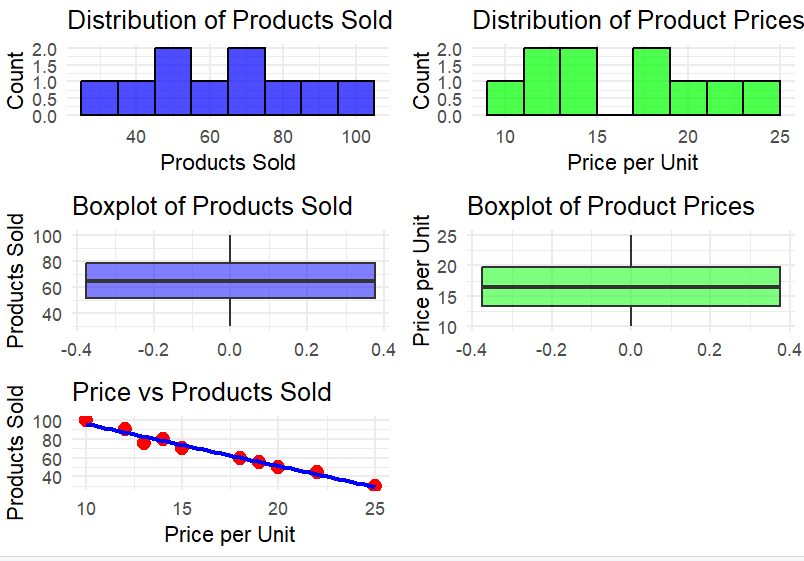
geom\_smooth(method = "lm", se = FALSE, color = "blue") + # Trend line

labs(title = "Price vs Products Sold", x = "Price per Unit", y = "Products Sold") +

theme\_minimal()

grid.arrange(hist\_products, hist\_prices, boxplot\_products, boxplot\_prices, scatter\_plot, ncol = 2)

**Output:**

****

**17. Covariance, Correlation & Visualization**

library(ggplot2)

set.seed(123)

data <- data.frame(

Variable\_X = c(10, 15, 20, 25, 30, 35, 40, 45, 50, 55),

Variable\_Y = c(22, 29, 35, 45, 50, 60, 65, 72, 78, 85)

)

cov\_xy <- cov(data$Variable\_X, data$Variable\_Y)

cor\_xy <- cor(data$Variable\_X, data$Variable\_Y)

cat("Covariance between Variable\_X and Variable\_Y:", cov\_xy, "\n")

cat("Correlation between Variable\_X and Variable\_Y:", cor\_xy, "\n")

ggplot(data, aes(x = Variable\_X, y = Variable\_Y)) +

geom\_point(color = "lightblue", size = 3) +

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

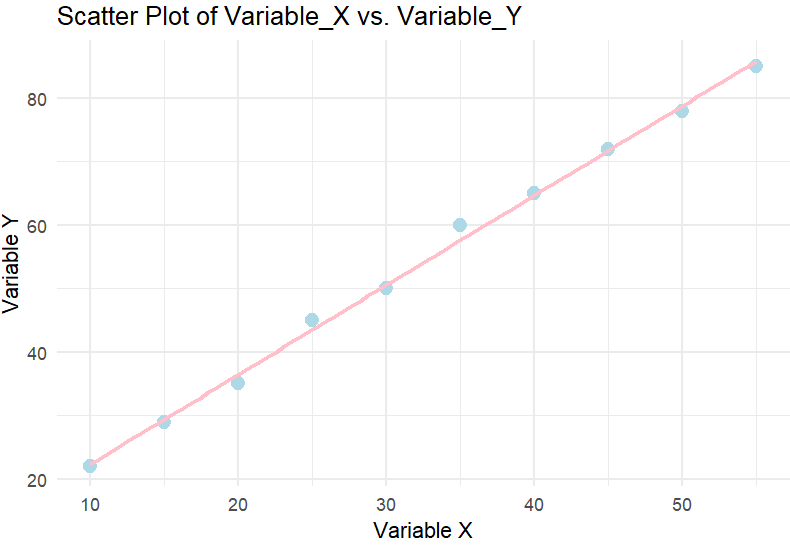
labs(title = "Scatter Plot of Variable\_X vs. Variable\_Y",

x = "Variable X",

y = "Variable Y") +

theme\_minimal()

**Output:**

****

**18. Scatter plot with line trend**

library(ggplot2)

set.seed(123)

data <- data.frame(

X\_Variable = c(10, 15, 20, 25, 30, 35, 40, 45, 50, 55),

Y\_Variable = c(22, 29, 35, 45, 50, 60, 65, 72, 78, 85)

)

ggplot(data, aes(x = X\_Variable, y = Y\_Variable)) +

geom\_point(color = "mediumturquoise", size = 3) +

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

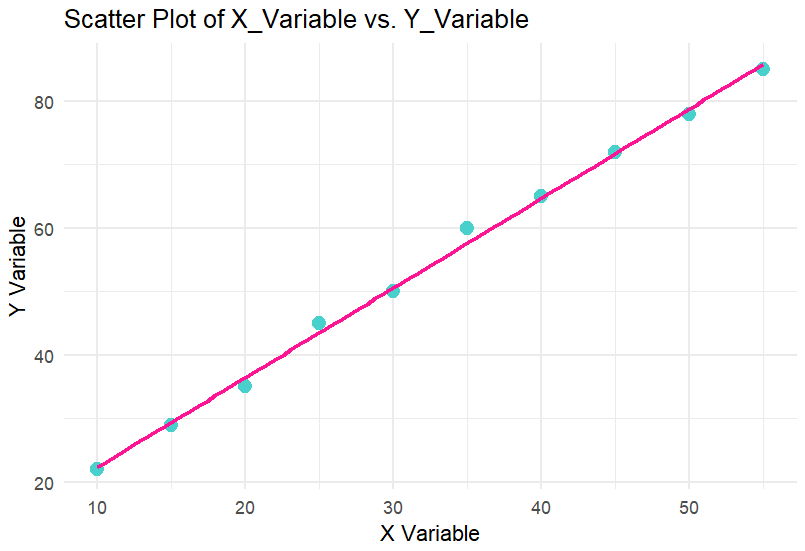
labs(title = "Scatter Plot of X\_Variable vs. Y\_Variable",

x = "X Variable",

y = "Y Variable") +

theme\_minimal()

**Output:**

****

**19. Heatmap**

library(ggcorrplot)

set.seed(123)

data <- data.frame(

Variable\_A = rnorm(100, mean = 50, sd = 10),

Variable\_B = rnorm(100, mean = 30, sd = 5),

Variable\_C = rnorm(100, mean = 75, sd = 15),

Variable\_D = rnorm(100, mean = 60, sd = 8)

)

cor\_matrix <- cor(data)

ggcorrplot(cor\_matrix, method = "square", type = "lower", # "square" or "circle"

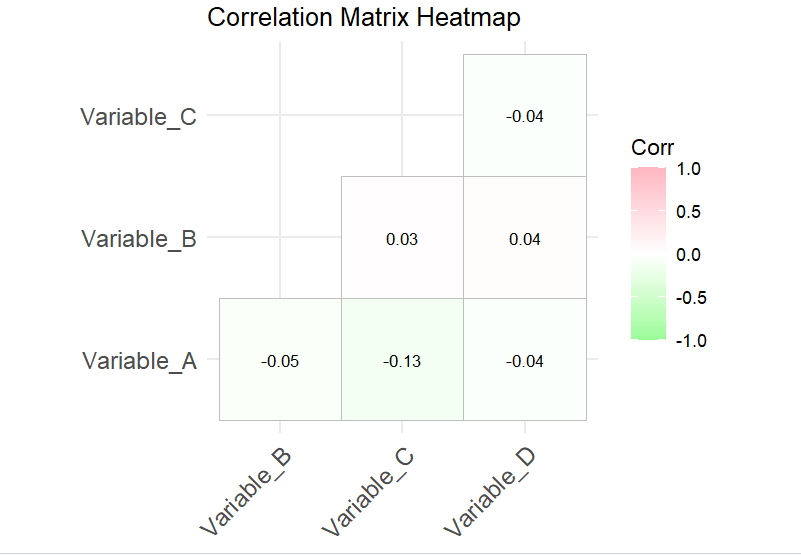
lab = TRUE, lab\_size = 3,

colors = c("palegreen", "white", "lightpink"),

title = "Correlation Matrix Heatmap",

ggtheme = theme\_minimal())

**Output:**

****

**20. Pair plot**

library(GGally)

library(ggplot2)

set.seed(123)

data <- data.frame(

Height = rnorm(100, mean = 170, sd = 10),

Weight = rnorm(100, mean = 70, sd = 15),

Age = rnorm(100, mean = 35, sd = 10),

Income = rnorm(100, mean = 50000, sd = 15000)

)

custom\_theme <- theme(

panel.background = element\_rect(fill = "gray95"),

panel.grid.major = element\_line(color = "white"),

panel.border = element\_rect(color = "black", fill = NA),

strip.background = element\_rect(fill = "skyblue"),

axis.text = element\_text(color = "black")

)

ggpairs(data,

upper = list(continuous = wrap("cor", size = 4, color = "red")),

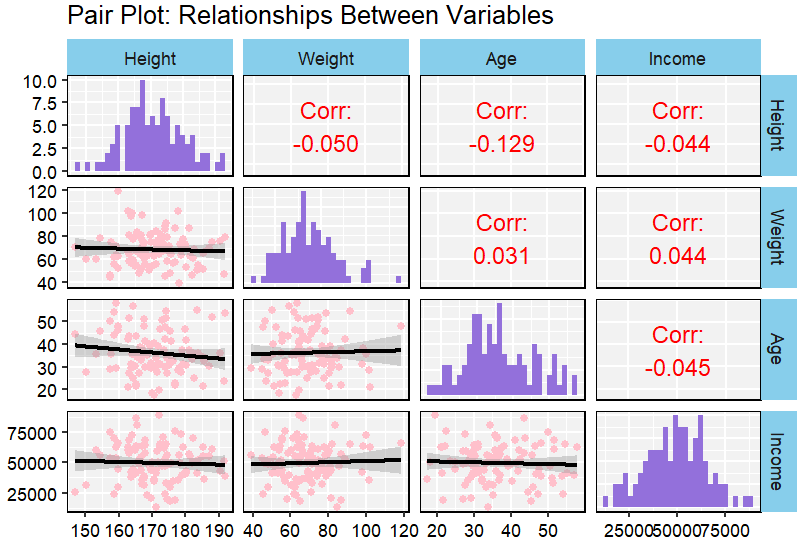
lower = list(continuous = wrap("smooth", color = "pink")),

diag = list(continuous = wrap("barDiag", fill = "mediumpurple")),

title = "Pair Plot: Relationships Between Variables") +

custom\_theme

**Output:**

****