1.Covariance and correlation

Children of three ages are asked to indicate their preference for three photographs of adults. Do the data suggest that there is a significant relationship between age and photograph preference? What is wrong with this study?

**Photograph:**

**Age of child** A B C

5-6 years: 18 22 20

7-8 years: 2 28 40

9-10 years: 20 10 40

1. Use cov() to calculate the sample covariance between B and C.
2. Use another call to cov() to calculate the sample covariance matrix for the preferences.
3. Use cor() to calculate the sample correlation between B and C.
4. Use another call to cor() to calculate the sample correlation matrix for the preferences.

**# Creating a matrix with the provided preference data**

**preferences <- matrix(c(22, 28, 10, 20, 40, 40), ncol = 3, byrow = TRUE)**

**rownames(preferences) <- c("5-6 years", "7-8 years", "9-10 years")**

**colnames(preferences) <- c("A", "B", "C")**

**# Calculate the sample covariance between B and C**

**cov\_bc <- cov(preferences[, "B"], preferences[, "C"])**

**print("Sample Covariance between B and C:")**

**print(cov\_bc)**

**# Calculate the sample covariance matrix for the preferences**

**cov\_matrix <- cov(preferences)**

**print("Sample Covariance Matrix for Preferences:")**

**print(cov\_matrix)**

**# Calculate the sample correlation between B and C**

**cor\_bc <- cor(preferences[, "B"], preferences[, "C"])**

**print("Sample Correlation between B and C:")**

**print(cor\_bc)**

**# Calculate the sample correlation matrix for the preferences**

**cor\_matrix <- cor(preferences)**

**print("Sample Correlation Matrix for Preferences:")**

**print(cor\_matrix)**

|  |
| --- |
| 18, 18, 18, 20, 20, 20, 20, 20, 20, 20, 21, 21, 21, 21, 25, 25, 25, 25, 25, 28, 28, 30, |
| 30, 30.  (i) Partition the dataset using an equal-frequency partitioning method with bin equal to 3 (ii) apply data smoothing using bin means and bin boundary. (iii) Plot Histogram for the above frequency division |

2.Imagine that you have selected data from the All Electronics data warehouse for analysis. The data set will be huge! The following data are a list of All Electronics prices for commonly sold items (rounded to the nearest dollar). The numbers have been sorted: 1, 1, 5, 5, 5, 5, 5, 8, 8, 10, 10, 10, 10, 12, 14, 14, 14, 15, 15, 15, 15, 15, 15, 18, 18, 18, 18, 18,

3.Two Maths teachers are comparing how their Year 9 classes performed in the end of year exams. Their results are as follows:  
Class A: 76, 35, 47, 64, 95, 66, 89, 36, 84, 76, 35,47,64,95,66,89,36,84

Class B: 51, 56, 84, 60, 59, 70, 63, 66, 50, 51,56,84,60,59,70,63,66,50

(i) Find which class had scored higher mean, median and range.  
(ii) Plot above in boxplot and give the inferences

Class B: 51, 56, 84, 60, 59, 70, 63, 66, 5051,56,84,60,59,70,63,66,50

**# Data for Class A and Class B**

**classA <- c(76, 35, 47, 64, 95, 66, 89, 36, 84)**

**classB <- c(51, 56, 84, 60, 59, 70, 63, 66, 50)**

**# Calculate mean, median, and range for Class A**

**mean\_A <- mean(classA)**

**median\_A <- median(classA)**

**range\_A <- max(classA) - min(classA)**

**# Calculate mean, median, and range for Class B**

**mean\_B <- mean(classB)**

**median\_B <- median(classB)**

**range\_B <- max(classB) - min(classB)**

**# Compare means**

**if (mean\_A > mean\_B) {**

**mean\_higher <- "Class A"**

**} else if (mean\_A < mean\_B) {**

**mean\_higher <- "Class B"**

**} else {**

**mean\_higher <- "Both classes have the same mean"**

**}**

**# Compare medians**

**if (median\_A > median\_B) {**

**median\_higher <- "Class A"**

**} else if (median\_A < median\_B) {**

**median\_higher <- "Class B"**

**} else {**

**median\_higher <- "Both classes have the same median"**

**}**

**# Compare ranges**

**if (range\_A > range\_B) {**

**range\_higher <- "Class A"**

**} else if (range\_A < range\_B) {**

**range\_higher <- "Class B"**

**} else {**

**range\_higher <- "Both classes have the same range"**

**}**

**# Print results**

**cat("Mean: Class", mean\_higher, "had a higher mean.\n")**

**cat("Median: Class", median\_higher, "had a higher median.\n")**

**cat("Range: Class", range\_higher, "had a higher range.\n")**

**# Combine data for both classes**

**combined\_data <- data.frame(Class = c(rep("A", length(classA)), rep("B", length(classB))),**

**Score = c(classA, classB))**

**# Create a boxplot**

**boxplot(Score ~ Class, data = combined\_data,**

**main = "Boxplot of Exam Scores by Class",**

**xlab = "Class", ylab = "Scores")**

**# Provide inferences**

**cat("\nInferences based on the boxplot:\n")**

**cat("- Class A has a wider range of scores compared to Class B.\n")**

**cat("- Class A's distribution appears to be skewed towards higher scores, while Class B's distribution is more symmetric.\n")**

**cat("- Class A has a higher median and a more spread-out interquartile range compared to Class B.\n")**

4.Let us consider one example to make the calculation method clear. Assume that the minimum and maximum values for the feature F are $50,000 and $100,000 correspondingly. It needs to range F from 0 to 1. In accordance with min-max normalization, v = $80,

b) Use the two methods below to normalize the following group of data: 200, 300, 400, 600, 1000

(a) min-max normalization by setting min = 0 and max = 1

(b) z-score normalization

**data <- c(200, 300, 400, 600, 1000)**

**v <- 80**

**min\_value <- 50000**

**max\_value <- 100000**

**# Min-Max Normalization formula: (x - min) / (max - min)**

**min\_max\_normalized\_a <- (data - min(data)) / (max(data) - min(data))**

**cat("Min-Max Normalization (Method A):\n")**

**print(min\_max\_normalized\_a)**

**# Z-Score Normalization formula: (x - mean) / standard deviation**

**z\_score\_normalized <- (data - mean(data)) / sd(data)**

**cat("\nZ-Score Normalization:\n")**

**print(z\_score\_normalized)**

1. Make a histogram for the “AirPassengers “dataset, start at 100 on the x-axis, and from values 200 to 700, make the bins 150 wide

**# Load the necessary library**

**library(datasets)**

**# Load the AirPassengers dataset**

**data("AirPassengers")**

**# Set up the histogram parameters**

**start\_value <- 100**

**end\_value <- 700**

**bin\_width <- 150**

**# Create the histogram**

**hist(AirPassengers, breaks = seq(start\_value, end\_value, by = bin\_width),**

**xlab = "Passenger Count", ylab = "Frequency",**

**main = "Histogram of AirPassengers",**

**col = "skyblue", border = "black")**

1. Obtain Multiple Lines in Line Chart using a single Plot Function in R.Use attributes“mpg”and“qsec”of the dataset “mtcars”

**# Load the necessary dataset**

**data("mtcars")**

**# Create a line chart with the first line (mpg)**

**plot(mtcars$qsec, mtcars$mpg, type = "l", col = "blue", xlab = "qsec", ylab = "mpg", main = "Line Chart with Multiple Lines")**

**# Add the second line (qsec) using the lines() function**

**lines(mtcars$qsec, mtcars$qsec, col = "red")**

1. Download the Dataset "water" From R dataset Link.Find out whether there is a linear relation between attributes"mortality" and"hardness" by plot function.Fit the Data into the Linear Regression model.Predict the mortality for the hardness=88.

**# Load the dataset**

**water\_data <- read.csv("water.csv")**

**# Display the structure of the dataset**

**str(water\_data)**

**# Create a scatter plot**

**plot(water\_data$hardness, water\_data$mortality,**

**xlab = "Hardness", ylab = "Mortality",**

**main = "Scatter Plot of Mortality vs. Hardness")**

**# Fit a linear regression model**

**linear\_model <- lm(mortality ~ hardness, data = water\_data)**

**# Display the model summary**

**summary(linear\_model)**

**# Predict mortality for hardness = 88**

**new\_data <- data.frame(hardness = 88)**

**predicted\_mortality <- predict(linear\_model, newdata = new\_data)**

**cat("Predicted Mortality for Hardness = 88:", predicted\_mortality)**

1. Create a Boxplot graph for the relation between "mpg"(miles per galloon) and "cyl"(number of Cylinders) for the dataset "mtcars" available in R Environment.

**# Load the necessary dataset**

**data("mtcars")**

**# Create a boxplot graph**

**boxplot(mpg ~ cyl, data = mtcars,**

**xlab = "Number of Cylinders", ylab = "Miles per Gallon",**

**main = "Boxplot of MPG by Number of Cylinders",**

**col = "skyblue", border = "black")**

9. Assume the Tennis coach wants to determine if any of his team players are scoring

outliers. To visualize the distribution of points scored by his players, then how can he

decide to develop the box plot? Give suitable example using Boxplot visualization

technique.

**# Example dataset of points scored by tennis players**

**points\_scored <- c(10, 15, 20, 25, 30, 40, 50, 60, 70, 80, 100, 120, 150)**

**# Create a boxplot to visualize points scored**

**boxplot(points\_scored,**

**main = "Boxplot of Points Scored by Tennis Players",**

**ylab = "Points Scored")**

**# Add a title and labels**

10. Implement using R language in which age group of people are affected byblood pressure based on the diabetes dataset show it using scatterplot and bar chart (that is BloodPressure vs Age using dataset “diabetes.csv”)

**# Load the necessary library**

**library(ggplot2)**

**# Load the dataset**

**diabetes\_data <- read.csv("diabetes.csv")**

**# Display the structure of the dataset**

**str(diabetes\_data)**

**# Create a scatterplot**

**scatter\_plot <- ggplot(data = diabetes\_data, aes(x = Age, y = BloodPressure)) +**

**geom\_point() +**

**labs(title = "Scatterplot of BloodPressure vs Age",**

**x = "Age", y = "BloodPressure")**

**print(scatter\_plot)**

**# Create age groups**

**age\_group\_data <- diabetes\_data %>%**

**mutate(AgeGroup = cut(Age, breaks = c(20, 30, 40, 50, 60, 70, Inf),**

**labels = c("20-29", "30-39", "40-49", "50-59", "60-69", "70+")))**

**# Calculate average blood pressure by age group**

**avg\_bp\_by\_age\_group <- age\_group\_data %>%**

**group\_by(AgeGroup) %>%**

**summarise(AvgBloodPressure = mean(BloodPressure))**

**# Create a bar chart**

**bar\_chart <- ggplot(data = avg\_bp\_by\_age\_group, aes(x = AgeGroup, y = AvgBloodPressure)) +**

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

**labs(title = "Average Blood Pressure by Age Group",**

**x = "Age Group", y = "Average Blood Pressure")**

**print(bar\_chart)**