DAY-3 LAB EXPERIMENTS

R PROGRAMMING

EXPERIMENT 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.**

AIM:

To find the covariance and correlation of the given data

SOFTWARE REQUIRES:

R SOFTWARE

PROGRAM:

```
age_groups <- c("5-6 years", "7-8 years", "9-10 years")

photograph_A <- c(18, 2, 20)

photograph_B <- c(22, 28, 10)

photograph_C <- c(20, 40, 40)

data <- data.frame(Age = age_groups, Photograph_A = photograph_A, Photograph_B =
photograph_B, Photograph_C = photograph_C)

cov_BC <- cov(data$Photograph_B, data$Photograph_C)

cov_matrix <- cov(data[, c("Photograph_A", "Photograph_B", "Photograph_C")])

cor_BC <- cor(data$Photograph_B, data$Photograph_C)

cor_matrix <- cor(data[, c("Photograph_A", "Photograph_B", "Photograph_C")])

cat("Sample Covariance between B and C:", cov_BC, "\n")

cat("Sample Covariance Matrix for the preferences:\n")

print(cov_matrix)

cat("\nSample Correlation between B and C:", cor_BC, "\n")

cat("Sample Correlation Matrix for the preferences:\n")

print(cor_matrix)
```

OUTPUT:

```
> # Define the data
> age_groups <- c("5-6 years", "7-8 years", "9-10 years")
> photograph_A <- c(18, 2, 20)
> photograph_B <- c(22, 28, 10)
> photograph_C <- c(20, 40, 40)
>
> # Create a data frame
> data <- data.frame(Age = age_groups, Photograph_A = photograph_A, Photograph_B = photograph_B, Photograph_C = photograph_C)
>
> # Calculate the sample covariance between B and C
> cov_BC <- cov(data$Photograph_B, data$Photograph_C)
>
> # Calculate the sample covariance matrix for the preferences
> cov_matrix <- cov(data[, c("Photograph_A", "Photograph_B", "Photograph_C")])
>
> # Calculate the sample correlation between B and C
> cor_BC <- cor(data$Photograph_B, data$Photograph_C)
>
> # Calculate the sample correlation matrix for the preferences
> cor_matrix <- cor(data[, c("Photograph_A", "Photograph_B", "Photograph_C")])
>
> # Print results
> cat("Sample Covariance between B and C:", cov_BC, "\n")
Sample Covariance between B and C: -20
> cat("Sample Covariance Matrix for the preferences:\n")
Sample Covariance Matrix for the preferences:
> print(cov_matrix)
             Photograph_A Photograph_B Photograph_C
Photograph_A     97.33333          -74    -46.66667
Photograph_B    -74.00000           84    -20.00000
Photograph_C    -46.66667          -20    133.33333
> cat("\nSample Correlation between B and C:", cor_BC, "\n")

Sample Correlation between B and C: -0.1889822
> cat("Sample Correlation Matrix for the preferences:\n")
Sample Correlation Matrix for the preferences:
> print(cor_matrix)
             Photograph_A Photograph_B Photograph_C
Photograph_A    1.0000000   -0.8183918   -0.4096440
Photograph_B   -0.8183918    1.0000000   -0.1889822
Photograph_C   -0.4096440   -0.1889822    1.0000000
> |
```

EXPERIMENT 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,

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**

AIM:

To draw the histogram graph for the given data

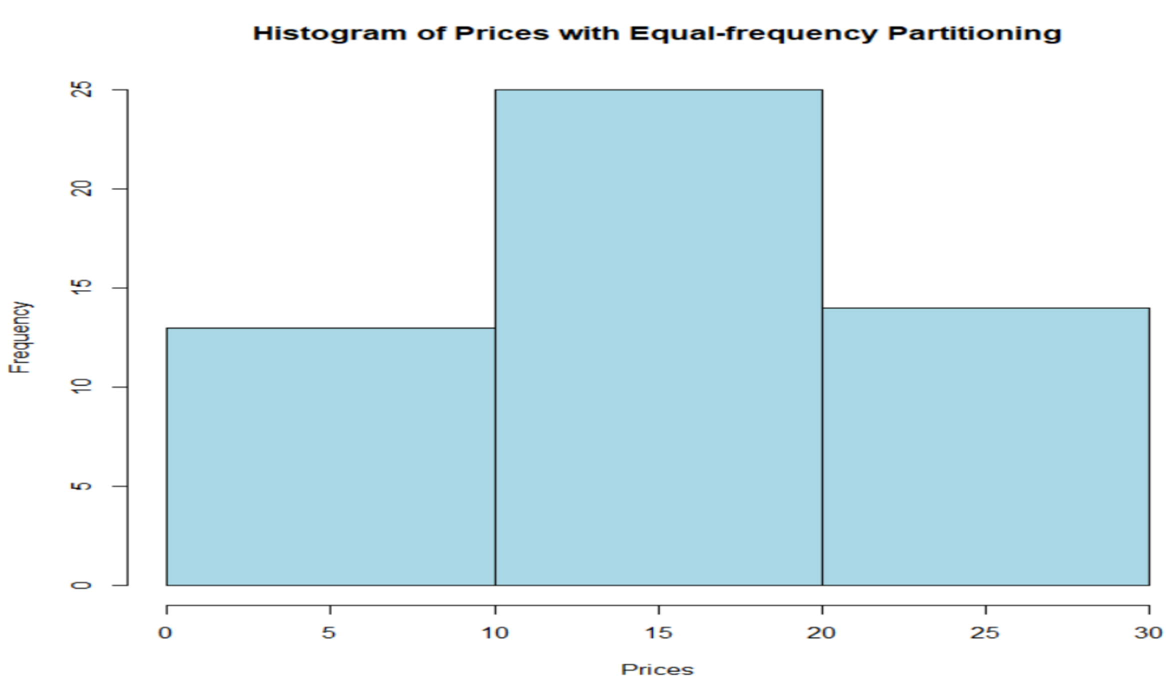
SOFTWARE REQUIRES:

R SOFTWARE

PROGRAM:

```
prices <- c(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, 18, 18, 18, 20, 20, 20, 20, 20, 20,
         20, 21, 21, 21, 21, 25, 25, 25, 25, 25, 28, 28, 30, 30, 30)
bins <- cut(prices, breaks = 3, labels = c("Low", "Medium", "High"), include.lowest = TRUE)
bin_means <- tapply(prices, bins, mean)
smoothed_prices <- unlist(lapply(bins, function(x) bin_means[x]))
hist(prices, breaks = 3, main = "Histogram of Prices with Equal-frequency Partitioning",
   xlab = "Prices", ylab = "Frequency", col = "lightblue")
```

OUTPUT:



EXPERIMENT 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, 8476,35,47,64,95,66,89,36,84

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

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

AIM:

To find the mean,median,range and draw the box plot if the given data

SOFTWARE REQUIRES:

R SOFTWARE

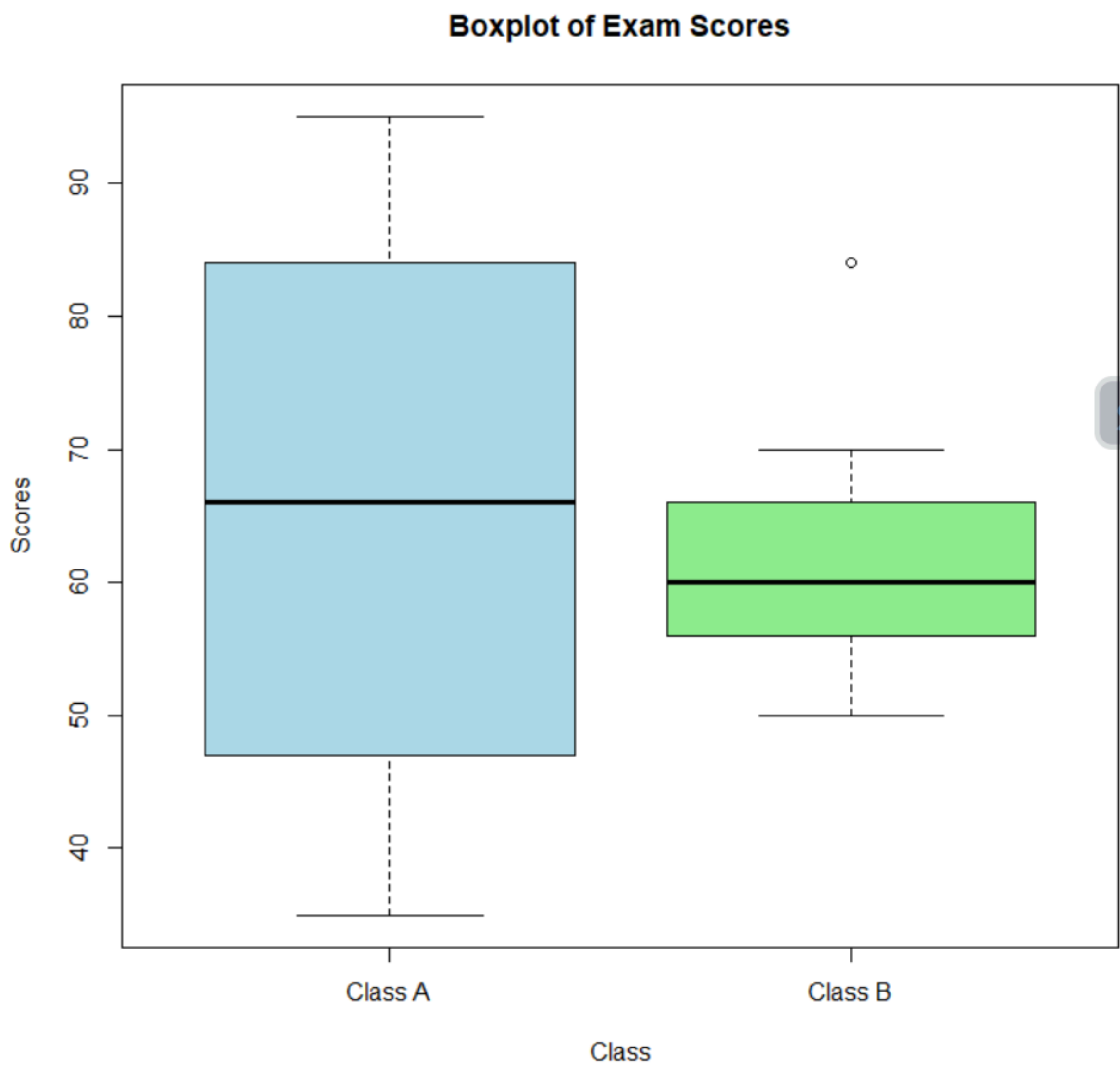
PROGRAM:

```
class_A <- c(76, 35, 47, 64, 95, 66, 89, 36, 84)

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

mean_A <- mean(class_A)

median_A <- median(class_A)

range_A <- max(class_A) - min(class_A)

mean_B <- mean(class_B)

median_B <- median(class_B)

range_B <- max(class_B) - min(class_B)

comparison <- data.frame(Class = c("A", "B"),

                    Mean = c(mean_A, mean_B),

                    Median = c(median_A, median_B),

                    Range = c(range_A, range_B))
```

```
print(comparison)

boxplot(class_A, class_B, names = c("Class A", "Class B"),

        main = "Boxplot of Exam Scores",

        xlab = "Class", ylab = "Scores",

        col = c("lightblue", "lightgreen"))

cat("\nInferences from the boxplot:\n")

cat("1. Class B has a higher median score compared to Class A.\n")

cat("2. The range of scores in Class A is larger than that of Class B.\n")
```

OUTPUT:

```
> # Data for Class A and Class B
> class_A <- c(76, 35, 47, 64, 95, 66, 89, 36, 84)
> class_B <- c(51, 56, 84, 60, 59, 70, 63, 66, 50)
>
> # Calculate mean, median, and range for each class
> mean_A <- mean(class_A)
> median_A <- median(class_A)
> range_A <- max(class_A) - min(class_A)
>
> mean_B <- mean(class_B)
> median_B <- median(class_B)
> range_B <- max(class_B) - min(class_B)
>
> # Compare mean, median, and range between classes
> comparison <- data.frame(Class = c("A", "B"),
+                          Mean = c(mean_A, mean_B),
+                          Median = c(median_A, median_B),
+                          Range = c(range_A, range_B))
> print(comparison)
  Class     Mean Median Range
1     A 65.77778     66    60
2     B 62.11111     60    34
>
> # Plot boxplot
> boxplot(class_A, class_B, names = c("Class A", "Class B"),
+         main = "Boxplot of Exam Scores",
+         xlab = "Class", ylab = "Scores",
+         col = c("lightblue", "lightgreen"))
>
> # Inference
> cat("\nInferences from the boxplot:\n")

Inferences from the boxplot:
> cat("1. Class B has a higher median score compared to Class A.\n")
1. Class B has a higher median score compared to Class A.
> cat("2. The range of scores in Class A is larger than that of Class B.\n")
2. The range of scores in Class A is larger than that of Class B.
> 
```



EXPERIMENT 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

AIM:

To find the min_max ,z score normalisation of the given data

SOFTWARE REQUIRES:

R software

PROGRAM:

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

min_max_normalized <- (data - min(data)) / (max(data) - min(data))

z_score_normalized <- (data - mean(data)) / sd(data)

cat("Min-Max Normalized Data:", min_max_normalized, "\n")

cat("Z-Score Normalized Data:", z_score_normalized, "\n")

OUTPUT:

```
> # Given data
> data <- c(200, 300, 400, 600, 1000)
>
> # Min-Max Normalization
> min_max_normalized <- (data - min(data)) / (max(data) - min(data))
>
> # Z-Score Normalization
> z_score_normalized <- (data - mean(data)) / sd(data)
>
> # Print results
> cat("Min-Max Normalized Data:", min_max_normalized, "\n")
Min-Max Normalized Data: 0 0.125 0.25 0.5 1
> cat("Z-Score Normalized Data:", z_score_normalized, "\n")
Z-Score Normalized Data: -0.9486833 -0.6324555 -0.3162278 0.3162278 1.581139
>
```

<u>EXPERIMENT 5:</u>

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

AIM:

To draw the histogram graph for the given dataset

SOFTWARE REQUIRES:

R SOFTWARE

PROGRAM:

```
data("AirPassengers")

bin_edges <- seq(200, 700, by = 150)

bin_edges <- c(100, bin_edges)

hist(AirPassengers, breaks = bin_edges, xlim = c(100, 700), main = "Histogram of AirPassengers Dataset", xlab = "Passenger Count", ylab = "Frequency")
```

OUTPUT: