                                                  DAY 1

1.The intervals and corresponding frequencies are as follows. age frequency

1-5. 200

5-15 450

15-20 300

20-50 1500

50-80 700

80-110 44

Compute an approximate median value for the data

Input:

#age, frequency

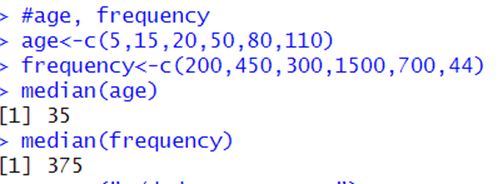
age<-c(5,15,20,50,80,110)

frequency<-c(200,450,300,1500,700,44)

median(age)

median(frequency)

Output:



2.Suppose that the data for analysis includes the attribute age. The age values for the data tuples are (in increasing order) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.

(a) What is the mean of the data? What is the median?

(b) What is the mode of the data? Comment on the data’s modality (i.e., bimodal, trimodal, etc.).

(c) What is the midrange of the data?

(d) Can you find (roughly) the first quartile (Q1) and the third quartile (Q3) of the data?

Input:

#mean,median,mode,quatile

age<-c(13,15,16,16,19,20,20,21,22,22,25,25,25,25,30,33,33,35,35,35,35,36,40,45,46,52,70)

mean(age)

median(age)

mode\_age<-names(table(age))[table(age)==max(table(age))]

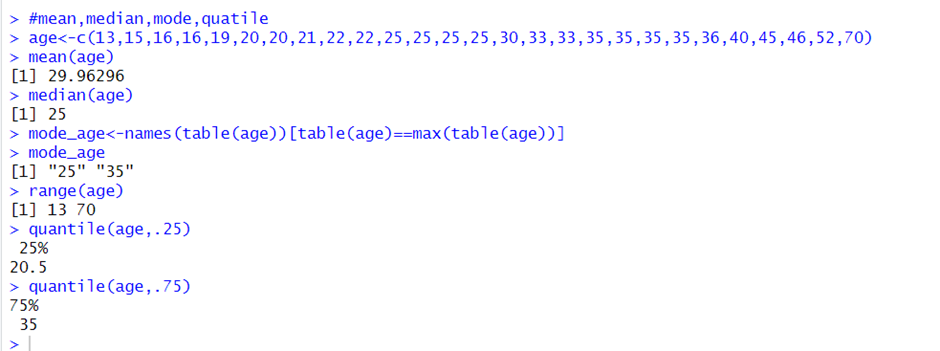
mode\_age

range(age)

quantile(age,.25)

quantile(age,.75)

Output:



  4.Data:11,13,13,15,15,16,19,20,20,20,21,21,22,23,24,30,40,45,45,45,71,

72,73,75

a) Smoothing by bin mean

b) Smoothing by bin median

c) Smoothing by bin boundaries

input:

data <- c(11,13,13,15,15,16,19,20,20,20,21,21,22,23,24,30,40,45,45,45,71,72,73,75)

bins <- 5

bin\_indices <- cut(data, bins)

mean\_smooth <- tapply(data, bin\_indices, mean)

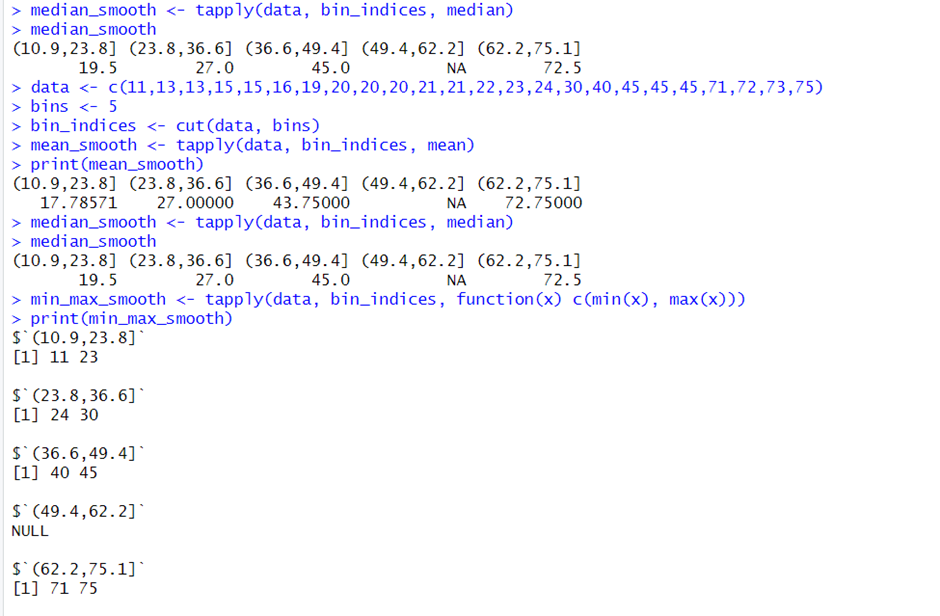
print(mean\_smooth)

median\_smooth <- tapply(data, bin\_indices, median)

median\_smooth

min\_max\_smooth <- tapply(data, bin\_indices, function(x) c(min(x), max(x)))

print(min\_max\_smooth)

Output:

5. Suppose that a hospital tested the age and body fat data for 18 randomly selected adults with the following results:

(a)   Calculate the mean, median, and standard deviation of age and %fat.

(b) Draw the boxplots for age and %fat.

(c) Draw a scatter plot and a q-q plot based on these two variables.

Input:

age<-c(23,23,27,27,39,41,47,49,50,52,54,54,56,57,58,58,60,61)

fat<-c(9.5,26.5,7.8,17.8,31.4,25.9,27.4,27.2,31.2,34.6,42.5,28.8,33.4,30.2,34.1,32.9,41.2,35.7)

mean(age)

median(age)

sd(age)

mean(fat)

median(fat)

sd(fat)

#boxplot

boxplot(age,fat)

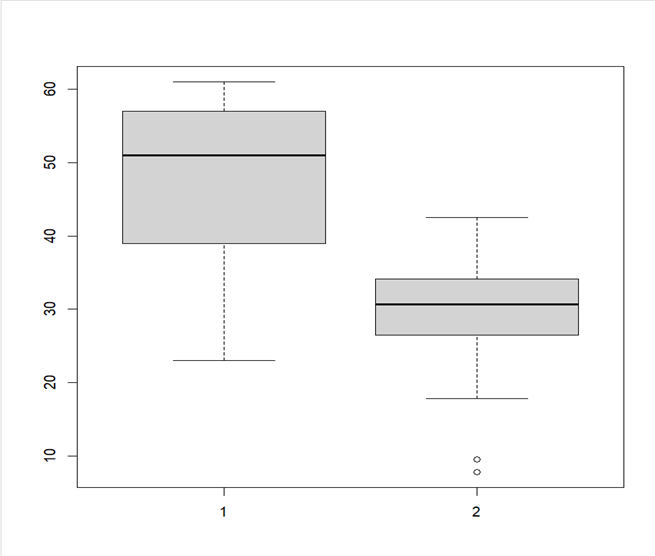
#scatter plot

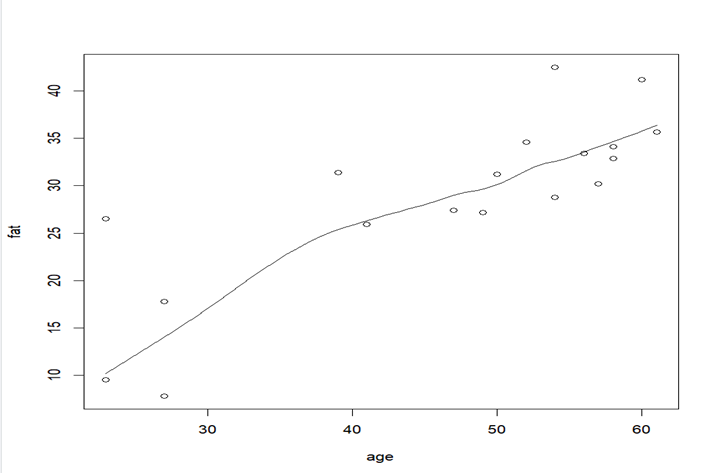
scatter.smooth(age,fat)

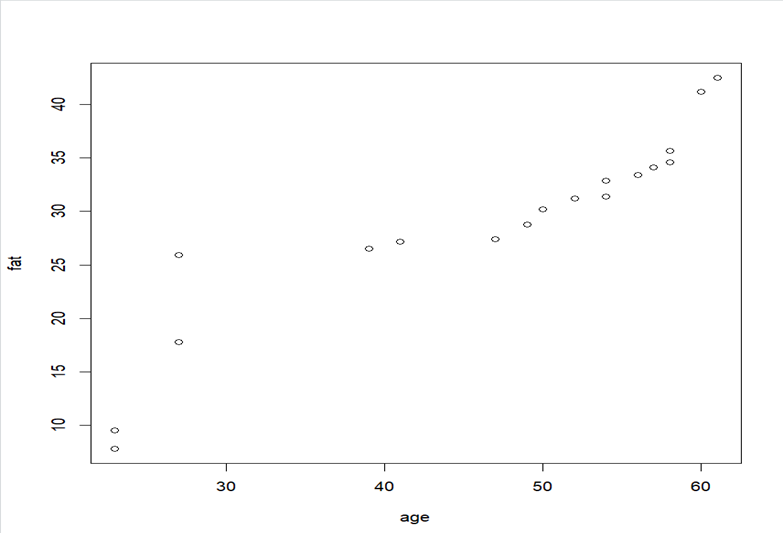
#qplot

qqplot(age,fat)

Output:







6.Suppose that a hospital tested the age and body fat data for 18 randomly selected adults with the following results:

(i) Use min-max normalization to transform the value 35 for age onto the range [0.0, 1.0].  
 (ii) Use z-score normalization to transform the value 35 for age, where the standard deviation of age is 12.94 years.  
 (iii) Use normalization by decimal scaling to transform the value 35 for age. Perform the above functions using R – tool

Input:

v<-c(23,23,27,27,39,41,47,49,50,52,54,54,56,57,58,58,60,61)

min<-0

max<-1

#min\_max

min\_max=((35-min(v))/(max(v)-min(v)))

print(min\_max)

#z-score

m=mean(v)

s<-12.94

z\_score=(35-m)/s

print(z\_score)

#decimal scaling

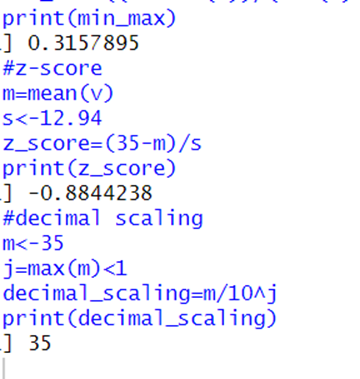
m<-35

j=max(m)<1

decimal\_scaling=m/10^j

print(decimal\_scaling)

output:



7.The following values are the number of pencils available in the different boxes. Create a vector and find out the mean, median and mode values of set of pencils in the given data.

Box1 Box2 Box3 Box4 Box5 Box6 Box7 Box8 Box9 Box 10

9          25  23 12      11  6  7        8    9         10

Input:

pencils<-c(9,25,23,12,11,6,7,8,9,10)

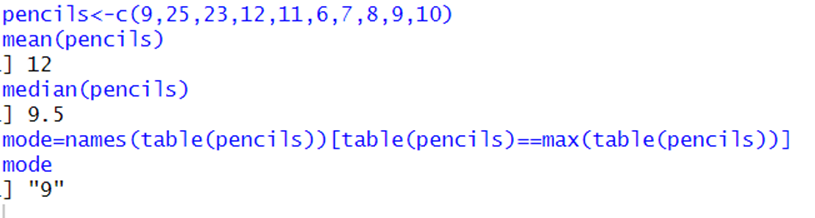
mean(pencils)

median(pencils)

mode=names(table(pencils))[table(pencils)==max(table(pencils))]

mode

Output:



 8. the following table would be plotted as (x,y) points, with the first column being the x values as number of mobile phones sold and the second column being the y values as money. To use the scatter plot for how many mobile phones sold.

x :4 1 5 7 10 2 50 25 90 36

y :12 5 13 19 31 7 153 72 275 110

input:

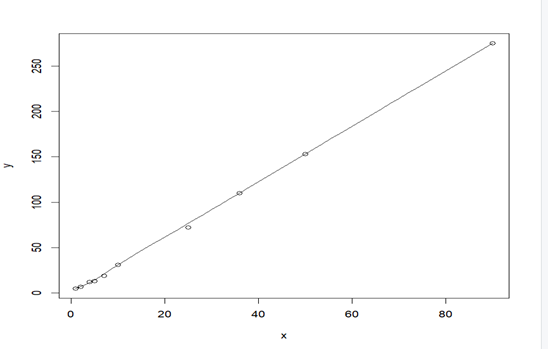
#scatterplot

x<-c(4,1,5,7,10,2,50,25,90,36)

y<-c(12,5,13,19,31,7,153,72,275,110)

scatter.smooth(x,y)

Output:



9. Implement of the R script using marks scored by a student in his model exam has been sorted as follows: 55, 60, 71, 63, 55, 65, 50, 55,58,59,61,63,65,67,71,72,75. Partition them into three bins by each of the following methods. Plot the data points using histogram.

(a) equal-frequency (equi-depth) partitioning (b) equal-width partitioning

Input:

marks <- c(55, 60, 71, 63, 55, 65, 50, 55, 58, 59, 61, 63, 65, 67, 71, 72, 75)

num\_bins <- 3

bins\_eq\_frequency <- cut(marks, breaks = num\_bins, labels = FALSE)

hist(marks, breaks = num\_bins, col = "lightblue", xlab = "Marks", main = "Equal-Frequency (Equi-Depth) Partitioning")

marks <- c(55, 60, 71, 63, 55, 65, 50, 55, 58, 59, 61, 63, 65, 67, 71, 72, 75)

bin\_mean <- tapply(data, cut(data, num\_bins), mean)

smoothed\_data\_by\_mean <- unname(bin\_mean[as.character(cut(data, num\_bins))])

bin\_median <- tapply(data, cut(data, num\_bins), median)

smoothed\_data\_by\_median <- unname(bin\_median[as.character(cut(data, num\_bins))])

bin\_boundaries <- tapply(data, cut(data, num\_bins), function(x) c(min(x), max(x)))

smoothed\_data\_by\_boundaries <- unlist(bin\_boundaries[as.character(cut(data, num\_bins))])

print("Original data:")

print(data)

print("Smoothed data by bin mean:")

print(smoothed\_data\_by\_mean)

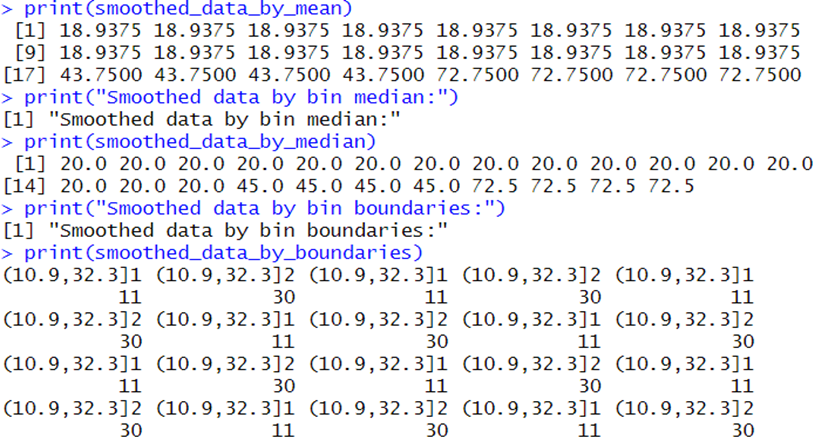
print("Smoothed data by bin median:")

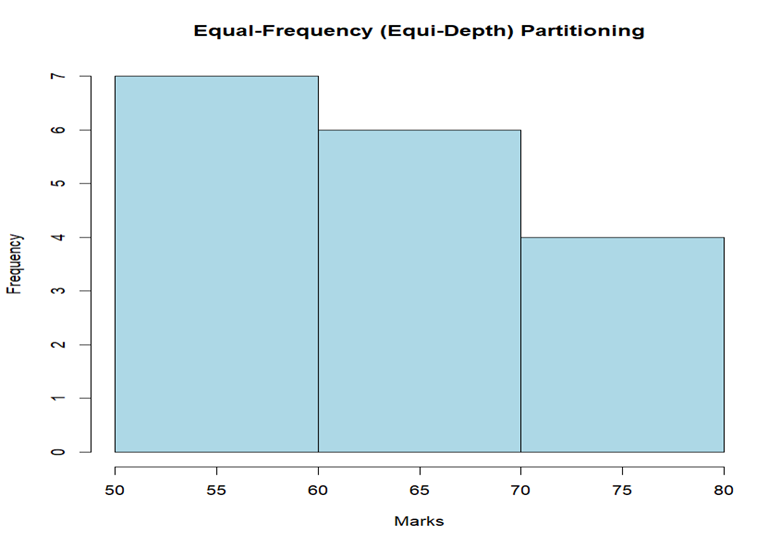
print(smoothed\_data\_by\_median)

print("Smoothed data by bin boundaries:")

print(smoothed\_data\_by\_boundaries)

Output:





10. Suppose that the speed car is mentioned in different driving style.

Regular 78.3 81.8 82 74.2 83.4 84.5 82.9 77.5 80.9 70.6 Speed

Calculate the Inter quantile and standard deviation of the given data.

Input:

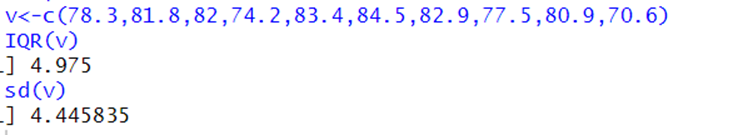
#IQR, SD

v<-c(78.3,81.8,82,74.2,83.4,84.5,82.9,77.5,80.9,70.6)

IQR(v)

sd(v)

Output:



11.Suppose that the data for analysis includes the attribute age. The age values for the data tuples are (in increasing order) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.

Can you find (roughly) the first quartile (Q1) and the third quartile (Q3) of the data?

Input:

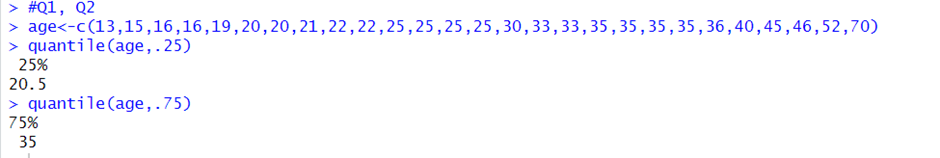
#Q1, Q2

age<-c(13,15,16,16,19,20,20,21,22,22,25,25,25,25,30,33,33,35,35,35,35,36,40,45,46,52,70)

quantile(age,.25)

quantile(age,.75)

output:



Day 2

12.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.

Input:

data <- data.frame(

Age = rep(c("5-6 years", "7-8 years", "9-10 years"), each = 3),

A = c(18, 2, 20, 22, 28, 10, 20, 40, 40),

B = c(22, 28, 10, 20, 40, 40, 30, 45, 50),

C = c(20, 40, 40, 30, 45, 50, 15, 35, 25)

)

covariance\_BC <- cov(data$B, data$C)

cat("Covariance between B and C:", covariance\_BC, "\n")

covariance\_matrix <- cov(data[, c("A", "B", "C")])

cat("Covariance matrix:\n", covariance\_matrix, "\n")

correlation\_BC <- cor(data$B, data$C)

cat("Correlation between B and C:", correlation\_BC, "\n")

correlation\_matrix <- cor(data[, c("A", "B", "C")])

cat("Correlation matrix:\n", correlation\_matrix, "\n")

output:

> data <- data.frame(

+ Age = rep(c("5-6 years", "7-8 years", "9-10 years"), each = 3),

+ A = c(18, 2, 20, 22, 28, 10, 20, 40, 40),

+ B = c(22, 28, 10, 20, 40, 40, 30, 45, 50),

+ C = c(20, 40, 40, 30, 45, 50, 15, 35, 25)

+ )

> covariance\_BC <- cov(data$B, data$C)

> cat("Covariance between B and C:", covariance\_BC, "\n")

Covariance between B and C: 16.875

> covariance\_matrix <- cov(data[, c("A", "B", "C")])

> cat("Covariance matrix:\n", covariance\_matrix, "\n")

Covariance matrix:

156.4444 84.83333 -38.33333 84.83333 171 16.875 -38.33333 16.875 137.5

> correlation\_BC <- cor(data$B, data$C)

> cat("Correlation between B and C:", correlation\_BC, "\n")

Correlation between B and C: 0.1100511

> correlation\_matrix <- cor(data[, c("A", "B", "C")])

> cat("Correlation matrix:\n", correlation\_matrix, "\n")

Correlation matrix:

1 0.5186667 -0.2613636 0.5186667 1 0.1100511 -0.2613636 0.1100511 1

13.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,

Input:

# Given data

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)

# (i) Equal-frequency partitioning with bin equal to 3

equal\_freq\_bins <- cut(prices, breaks = 3, labels = FALSE)

cat("(i) Equal-frequency partitioning bins:\n", equal\_freq\_bins, "\n")

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

# (ii) Data smoothing using bin means and bin boundary

bin\_means <- tapply(prices, equal\_freq\_bins, mean)

bin\_boundaries <- unique(cut(prices, breaks = 3, labels = FALSE, include.lowest = TRUE))

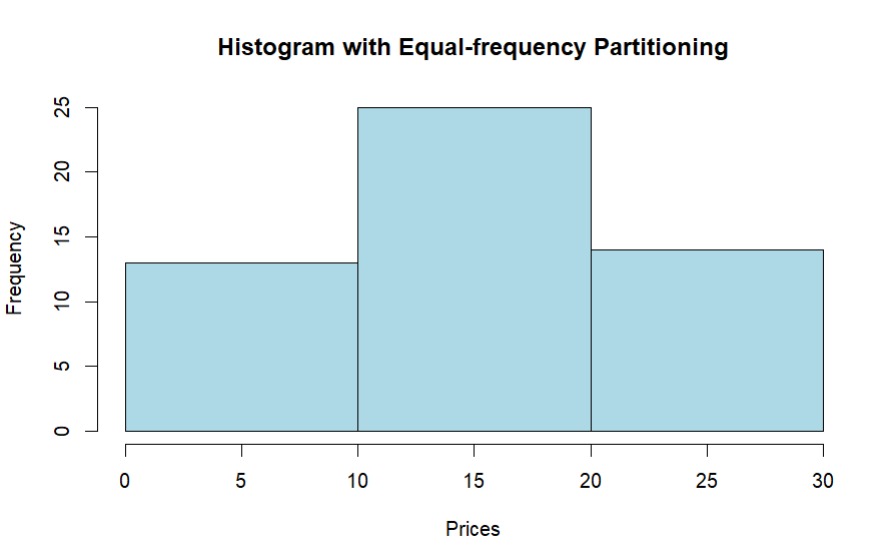
cat("(ii) Bin Means:\n", bin\_means, "\n")

cat("Bin Boundaries:\n", bin\_boundaries, "\n")

# (iii) Plot Histogram

hist(prices, breaks = 3, main = "Histogram with Equal-frequency Partitioning", xlab = "Prices", col = "lightblue", border = "black")

Output:



14.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.

INPUT:

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

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

meanA <- mean(classA)

meanB <- mean(classB)

medianA <- median(classA)

medianB <- median(classB)

rangeA <- range(classA)

rangeB <- range(classB)

cat("(i) Class A vs Class B:\n")

cat("Mean: Class A -", meanA, " Class B -", meanB, "\n")

cat("Median: Class A -", medianA, " Class B -", medianB, "\n")

cat("Range: Class A -", diff(rangeA), " Class B -", diff(rangeB), "\n")

boxplot(classA, classB, names = c("Class A", "Class B"), col = c("lightblue", "lightgreen"), main = "Boxplot - Class A vs Class B", ylab = "Scores")

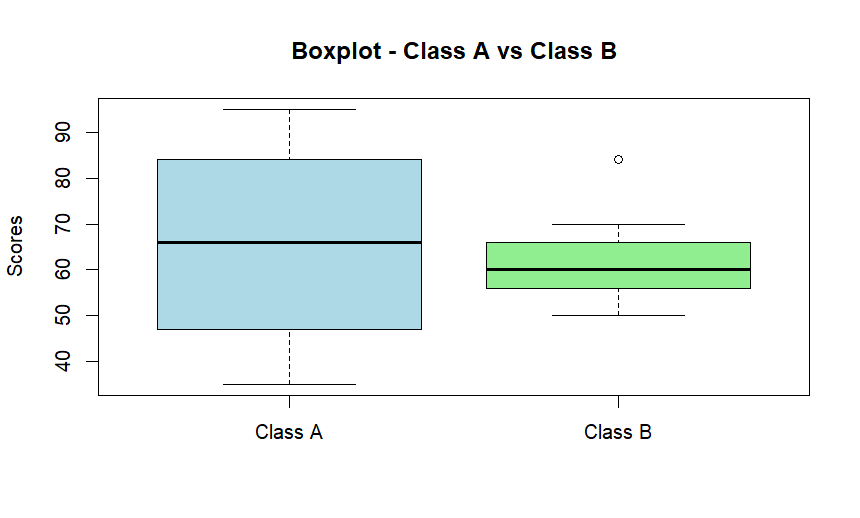
cat("(ii) Inferences from Boxplot:\n")

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

cat(" - The median score for Class A is higher than that for Class B.\n")

cat(" - Class A has an outlier which affects the mean.\n")

output:



15.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

Input:

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

min\_max\_custom <- function(x, min\_val, max\_val) {

return (x - min\_val) / (max\_val - min\_val)

}

min\_max\_normalized\_custom <- min\_max\_custom(data, 200, 1000)

min\_max\_normalized\_default <- scale(data, center = min(data), scale = diff(range(data)))

z\_score\_normalized <- scale(data)

cat("Original Data: ", data, "\n\n")

cat("(a) Min-Max normalization with custom min and max values:\n")

cat("Normalized Data: ", min\_max\_normalized\_custom, "\n\n")

cat("(b) Min-Max normalization with min = 0 and max = 1:\n")

cat("Normalized Data: ", min\_max\_normalized\_default, "\n\n")

cat("(c) Z-score normalization:\n")

cat("Normalized Data: ", z\_score\_normalized, "\n")

output:

Original Data: 200 300 400 600 1000

> cat("(a) Min-Max normalization with custom min and max values:\n")

(a) Min-Max normalization with custom min and max values:

> cat("Normalized Data: ", min\_max\_normalized\_custom, "\n\n")

Normalized Data: 0 100 200 400 800

> cat("(b) Min-Max normalization with min = 0 and max = 1:\n")

(b) Min-Max normalization with min = 0 and max = 1:

> cat("Normalized Data: ", min\_max\_normalized\_default, "\n\n")

Normalized Data: 0 0.125 0.25 0.5 1

> cat("(c) Z-score normalization:\n")

(c) Z-score normalization:

> cat("Normalized Data: ", z\_score\_normalized, "\n")

Normalized Data: -0.9486833 -0.6324555 -0.3162278 0.3162278 1.581139

16.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

Input:

data("AirPassengers")

start\_value <- 100

bin\_width <- 150

bin\_breaks <- seq(start\_value, 700, by = bin\_width)

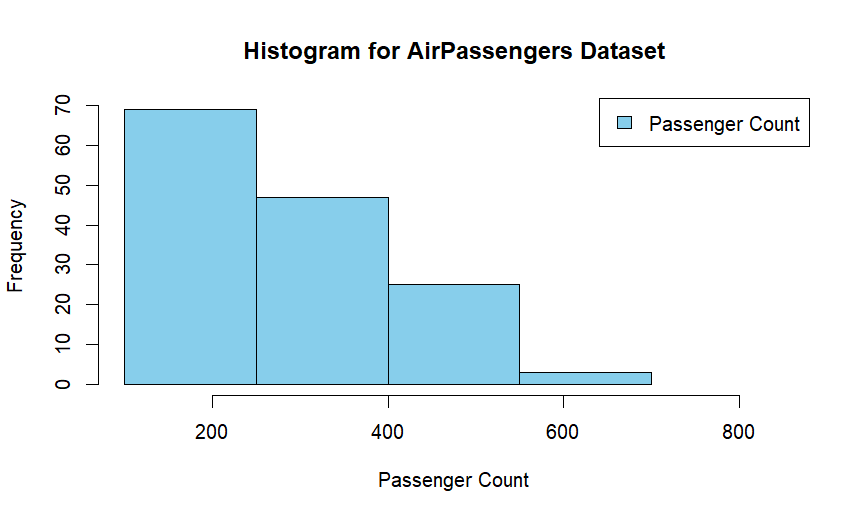
hist(AirPassengers, breaks = bin\_breaks, xlim = c(start\_value, max(bin\_breaks) + bin\_width),

main = "Histogram for AirPassengers Dataset",

xlab = "Passenger Count", ylab = "Frequency", col = "skyblue", border = "black")

legend("topright", legend = c("Passenger Count"), fill = c("skyblue"))

OUTPUT:



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

INPUT

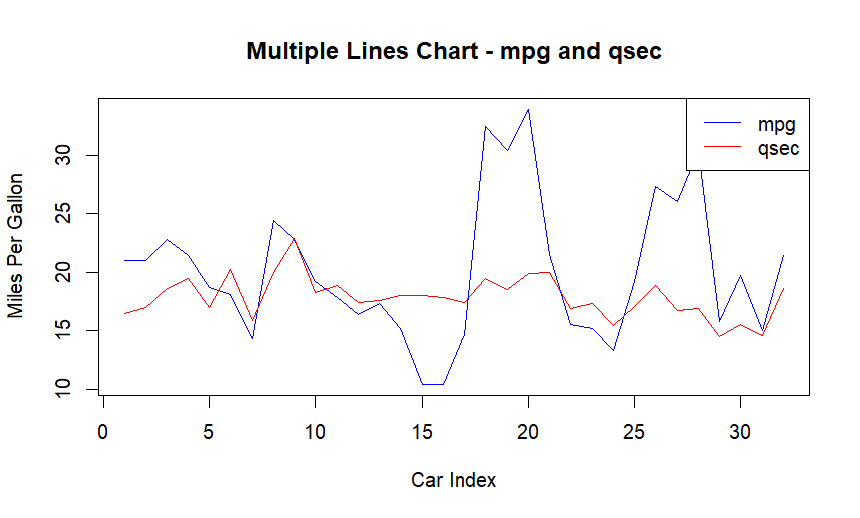
data(mtcars)

plot(mtcars$mpg, type = "l", col = "blue", xlab = "Car Index", ylab = "Miles Per Gallon", main = "Multiple Lines Chart - mpg and qsec")

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

legend("topright", legend = c("mpg", "qsec"), col = c("blue", "red"), lty = 1)

OUTPUT



18.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.

INPUT

data(mtcars)

mortality <- mtcars$mpg

hardness <- mtcars$hp

plot(hardness, mortality, main = "Linear Regression: Mortality vs. Hardness",

xlab = "Hardness", ylab = "Mortality", pch = 16, col = "blue")

linear\_model <- lm(mortality ~ hardness)

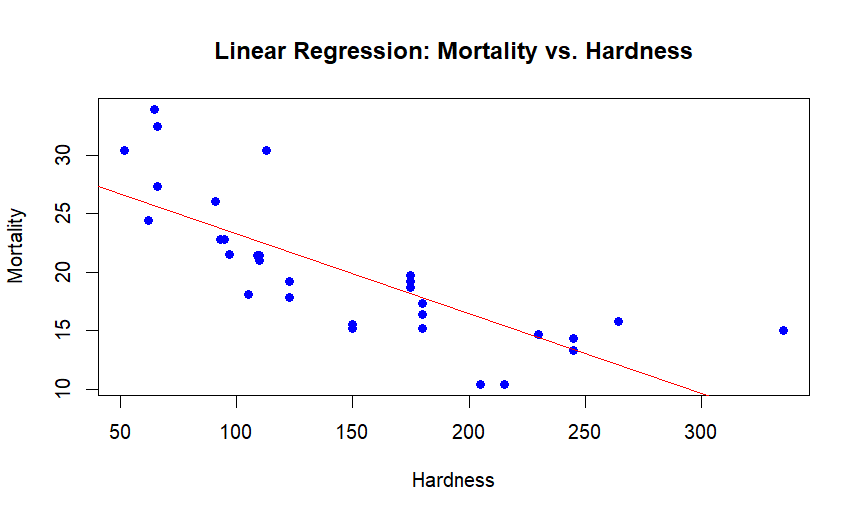
abline(linear\_model, col = "red")

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

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

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

Output:



19.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.

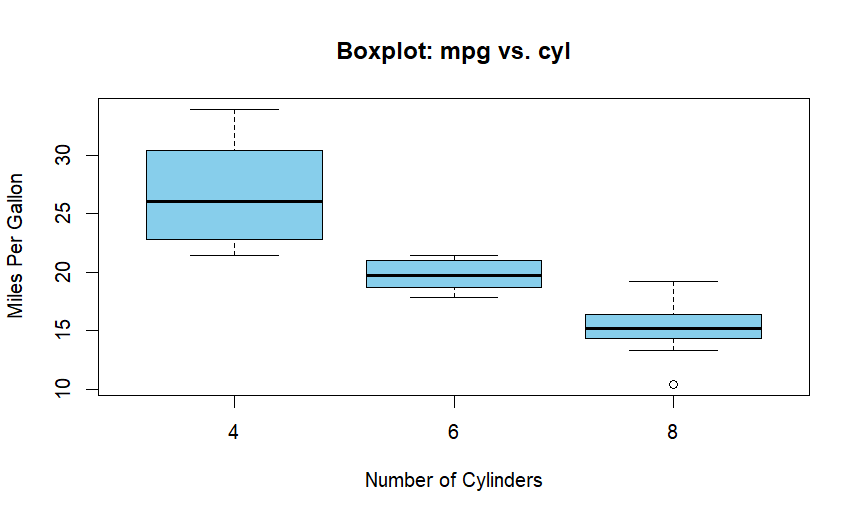
Input:

data(mtcars)

boxplot(mpg ~ cyl, data = mtcars, main = "Boxplot: mpg vs. cyl",

xlab = "Number of Cylinders", ylab = "Miles Per Gallon", col = "skyblue")

OUTPUT



20. 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.

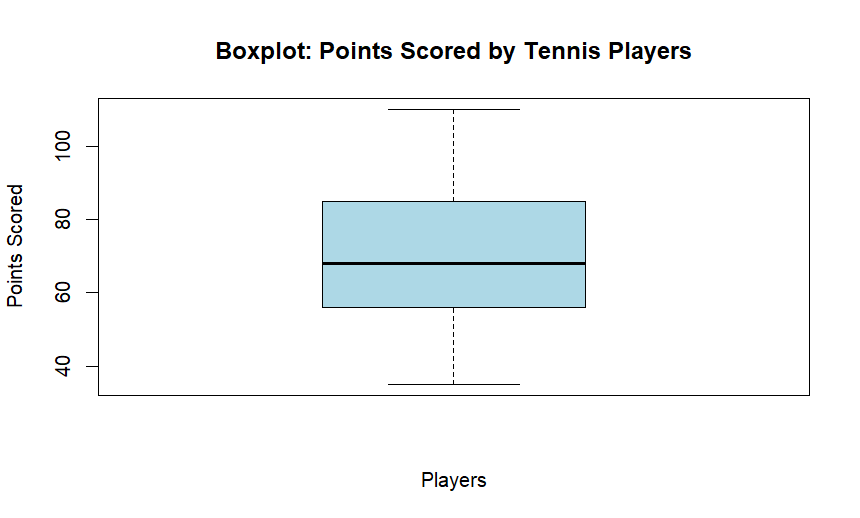
Input:

points\_scored <- c(35, 42, 48, 52, 56, 60, 62, 65, 68, 72, 76, 80, 85, 88, 92, 100, 110)

boxplot(points\_scored, main = "Boxplot: Points Scored by Tennis Players",

xlab = "Players", ylab = "Points Scored", col = "lightblue", border = "black")

OUTPUT



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

**Input:**

# Sample data (assuming the structure of your dataset)

data <- data.frame(

  Age = c(25, 30, 35, 40, 45, 50, 55, 60, 65, 70),

  BloodPressure = c(120, 130, 140, 150, 135, 145, 155, 160, 150, 140)

)

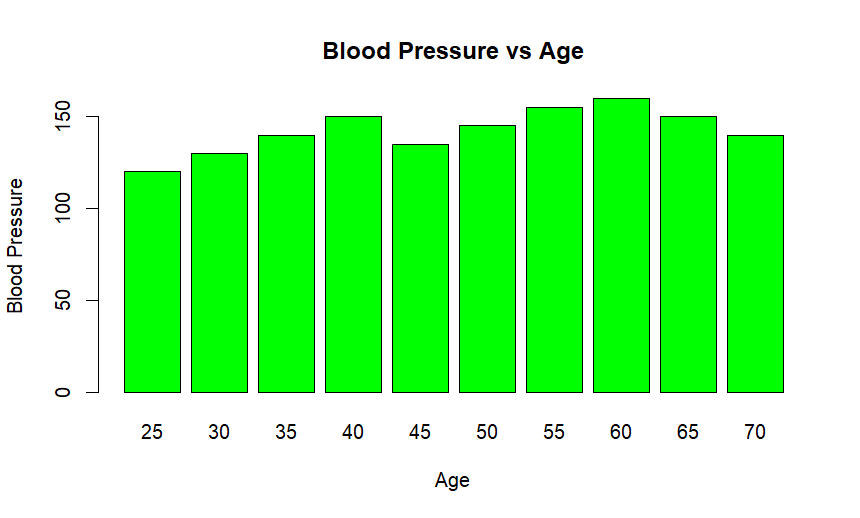
# Scatterplot

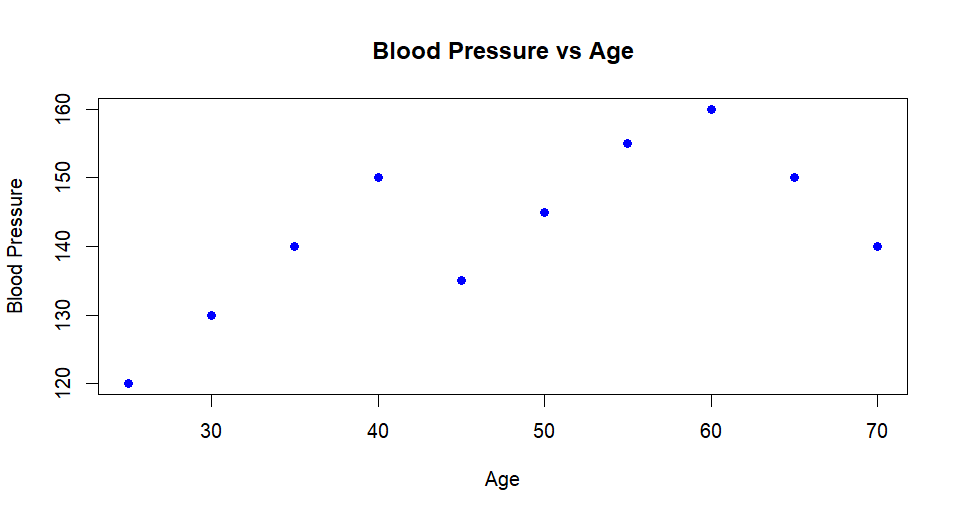
plot(data$Age, data$BloodPressure, main = "Blood Pressure vs Age", xlab = "Age", ylab = "Blood Pressure", pch = 16, col = "blue")

# Bar chart

barplot(data$BloodPressure, names.arg = data$Age, main = "Blood Pressure vs Age", xlab = "Age", ylab = "Blood Pressure", col = "green", border = "black")

**Output:**

****

****

**DAY 3:**

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.

**Input:**

data <- data.frame(

Age = rep(c("5-6 years", "7-8 years", "9-10 years"), each = 3),

A = c(18, 2, 20, 22, 28, 10, 20, 40, 40),

B = c(22, 28, 10, 20, 40, 40, 30, 45, 50),

C = c(20, 40, 40, 30, 45, 50, 15, 35, 25)

)

covariance\_BC <- cov(data$B, data$C)

cat("Covariance between B and C:", covariance\_BC, "\n")

covariance\_matrix <- cov(data[, c("A", "B", "C")])

cat("Covariance matrix:\n", covariance\_matrix, "\n")

correlation\_BC <- cor(data$B, data$C)

cat("Correlation between B and C:", correlation\_BC, "\n")

correlation\_matrix <- cor(data[, c("A", "B", "C")])

cat("Correlation matrix:\n", correlation\_matrix, "\n")

**output:**

> data <- data.frame(

+ Age = rep(c("5-6 years", "7-8 years", "9-10 years"), each = 3),

+ A = c(18, 2, 20, 22, 28, 10, 20, 40, 40),

+ B = c(22, 28, 10, 20, 40, 40, 30, 45, 50),

+ C = c(20, 40, 40, 30, 45, 50, 15, 35, 25)

+ )

> covariance\_BC <- cov(data$B, data$C)

> cat("Covariance between B and C:", covariance\_BC, "\n")

Covariance between B and C: 16.875

> covariance\_matrix <- cov(data[, c("A", "B", "C")])

> cat("Covariance matrix:\n", covariance\_matrix, "\n")

Covariance matrix:

156.4444 84.83333 -38.33333 84.83333 171 16.875 -38.33333 16.875 137.5

> correlation\_BC <- cor(data$B, data$C)

> cat("Correlation between B and C:", correlation\_BC, "\n")

Correlation between B and C: 0.1100511

> correlation\_matrix <- cor(data[, c("A", "B", "C")])

> cat("Correlation matrix:\n", correlation\_matrix, "\n")

Correlation matrix:

1 0.5186667 -0.2613636 0.5186667 1 0.1100511 -0.2613636 0.1100511 1

|  |
| --- |
| 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,

**Input:**

# Given data

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)

# (i) Equal-frequency partitioning with bin equal to 3

equal\_freq\_bins <- cut(prices, breaks = 3, labels = FALSE)

cat("(i) Equal-frequency partitioning bins:\n", equal\_freq\_bins, "\n")

# (ii) Data smoothing using bin means and bin boundary

bin\_means <- tapply(prices, equal\_freq\_bins, mean)

bin\_boundaries <- unique(cut(prices, breaks = 3, labels = FALSE, include.lowest = TRUE))

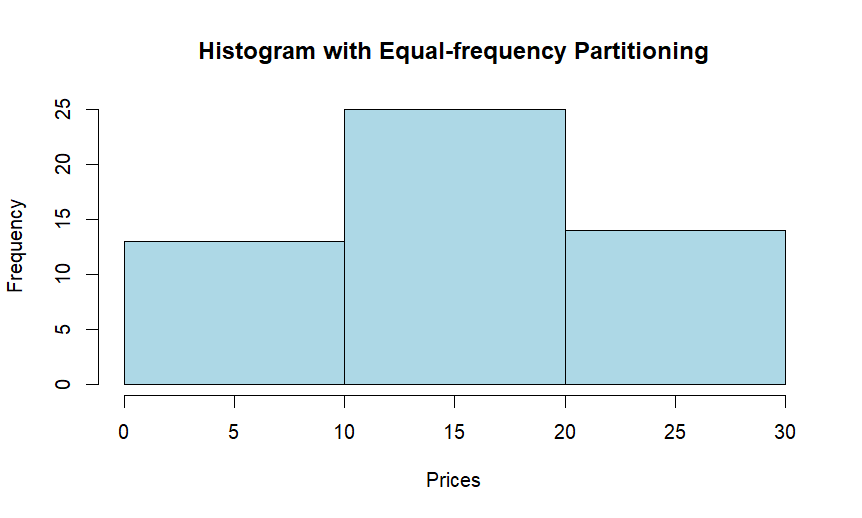
cat("(ii) Bin Means:\n", bin\_means, "\n")

cat("Bin Boundaries:\n", bin\_boundaries, "\n")

# (iii) Plot Histogram

hist(prices, breaks = 3, main = "Histogram with Equal-frequency Partitioning", xlab = "Prices", col = "lightblue", border = "black")

**output:**



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

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

**Input:**

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

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

meanA <- mean(classA)

meanB <- mean(classB)

medianA <- median(classA)

medianB <- median(classB)

rangeA <- range(classA)

rangeB <- range(classB)

cat("(i) Class A vs Class B:\n")

cat("Mean: Class A -", meanA, " Class B -", meanB, "\n")

cat("Median: Class A -", medianA, " Class B -", medianB, "\n")

cat("Range: Class A -", diff(rangeA), " Class B -", diff(rangeB), "\n")

boxplot(classA, classB, names = c("Class A", "Class B"), col = c("lightblue", "lightgreen"), main = "Boxplot - Class A vs Class B", ylab = "Scores")

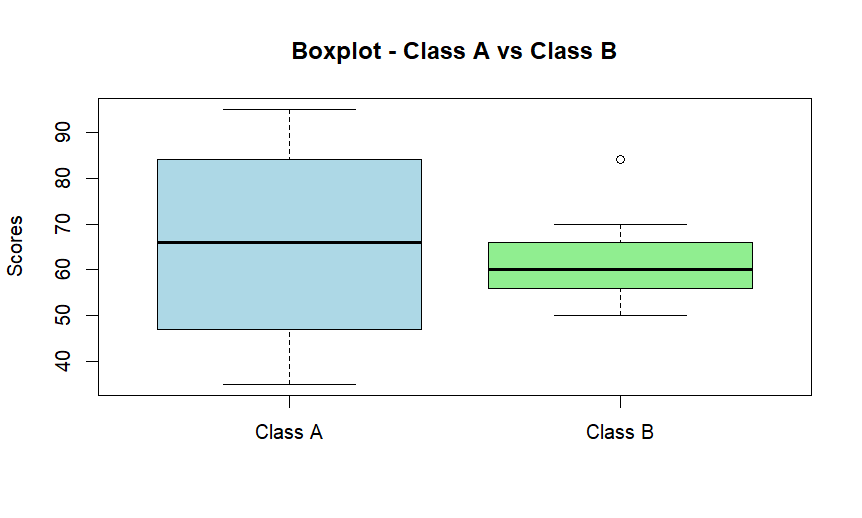
cat("(ii) Inferences from Boxplot:\n")

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

cat(" - The median score for Class A is higher than that for Class B.\n")

cat(" - Class A has an outlier which affects the mean.\n")

**output:**

****

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

**Input:**

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

min\_max\_custom <- function(x, min\_val, max\_val) {

return (x - min\_val) / (max\_val - min\_val)

}

min\_max\_normalized\_custom <- min\_max\_custom(data, 200, 1000)

min\_max\_normalized\_default <- scale(data, center = min(data), scale = diff(range(data)))

z\_score\_normalized <- scale(data)

cat("Original Data: ", data, "\n\n")

cat("(a) Min-Max normalization with custom min and max values:\n")

cat("Normalized Data: ", min\_max\_normalized\_custom, "\n\n")

cat("(b) Min-Max normalization with min = 0 and max = 1:\n")

cat("Normalized Data: ", min\_max\_normalized\_default, "\n\n")

cat("(c) Z-score normalization:\n")

cat("Normalized Data: ", z\_score\_normalized, "\n")

**output:**

Original Data: 200 300 400 600 1000

> cat("(a) Min-Max normalization with custom min and max values:\n")

(a) Min-Max normalization with custom min and max values:

> cat("Normalized Data: ", min\_max\_normalized\_custom, "\n\n")

Normalized Data: 0 100 200 400 800

> cat("(b) Min-Max normalization with min = 0 and max = 1:\n")

(b) Min-Max normalization with min = 0 and max = 1:

> cat("Normalized Data: ", min\_max\_normalized\_default, "\n\n")

Normalized Data: 0 0.125 0.25 0.5 1

> cat("(c) Z-score normalization:\n")

(c) Z-score normalization:

> cat("Normalized Data: ", z\_score\_normalized, "\n")

Normalized Data: -0.9486833 -0.6324555 -0.3162278 0.3162278 1.581139

5.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

**Input:**

data("AirPassengers")

start\_value <- 100

bin\_width <- 150

bin\_breaks <- seq(start\_value, 700, by = bin\_width)

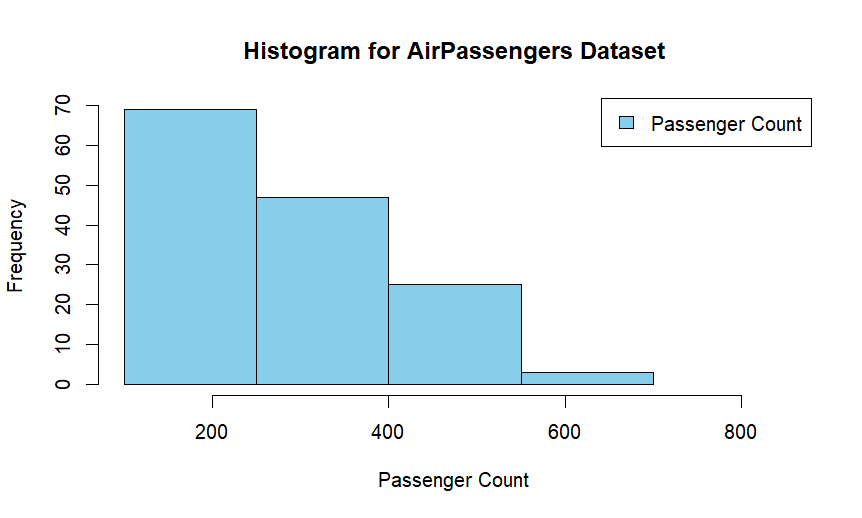
hist(AirPassengers, breaks = bin\_breaks, xlim = c(start\_value, max(bin\_breaks) + bin\_width),

main = "Histogram for AirPassengers Dataset",

xlab = "Passenger Count", ylab = "Frequency", col = "skyblue", border = "black")

legend("topright", legend = c("Passenger Count"), fill = c("skyblue"))

**output:**

****

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

**Input:**

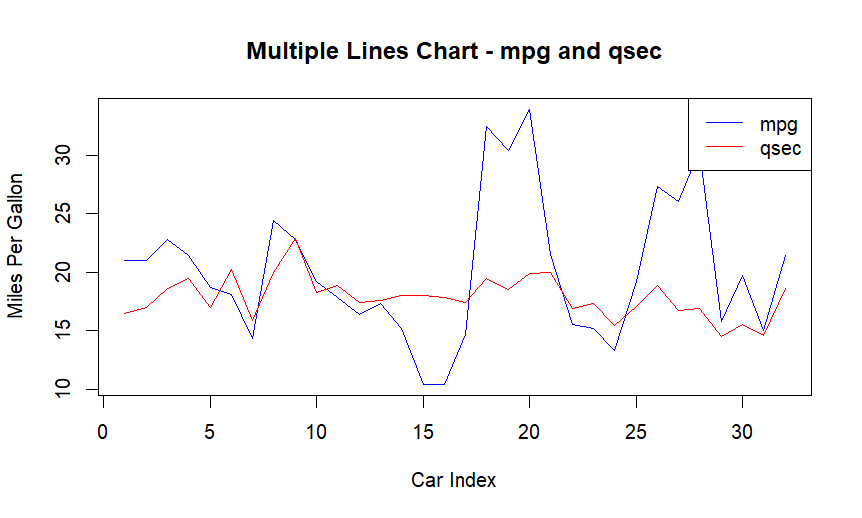
data(mtcars)

plot(mtcars$mpg, type = "l", col = "blue", xlab = "Car Index", ylab = "Miles Per Gallon", main = "Multiple Lines Chart - mpg and qsec")

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

legend("topright", legend = c("mpg", "qsec"), col = c("blue", "red"), lty = 1)

**output:**

****

7.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.

**Input:**

data(mtcars)

mortality <- mtcars$mpg

hardness <- mtcars$hp

plot(hardness, mortality, main = "Linear Regression: Mortality vs. Hardness",

xlab = "Hardness", ylab = "Mortality", pch = 16, col = "blue")

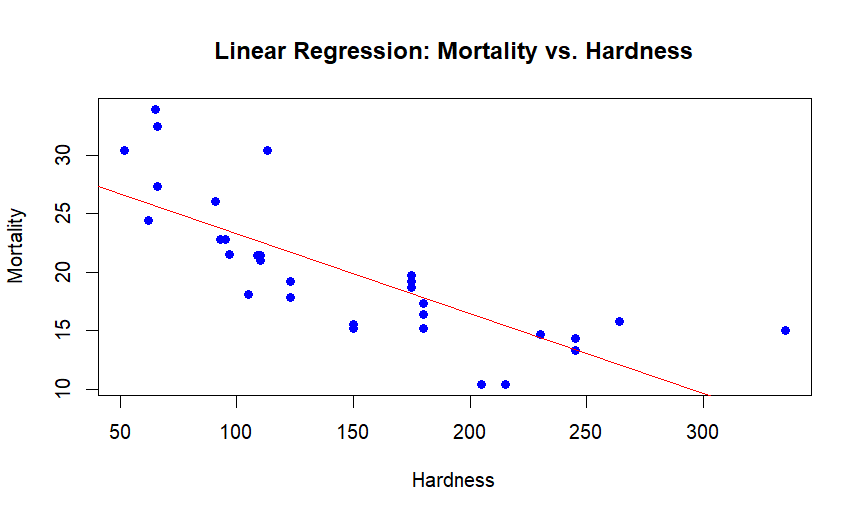
linear\_model <- lm(mortality ~ hardness)

abline(linear\_model, col = "red")

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

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

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

**output:  
**

**8.**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.

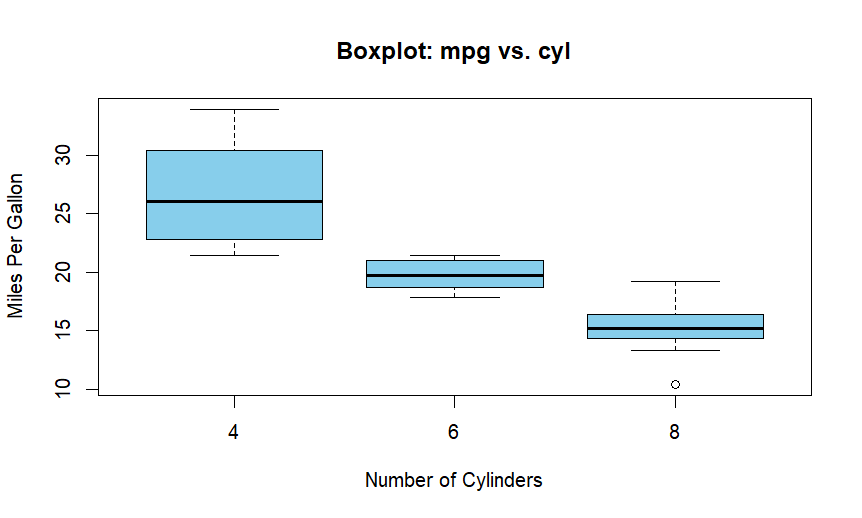
**Input:**

data(mtcars)

boxplot(mpg ~ cyl, data = mtcars, main = "Boxplot: mpg vs. cyl",

xlab = "Number of Cylinders", ylab = "Miles Per Gallon", col = "skyblue")

**output:**

****

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.

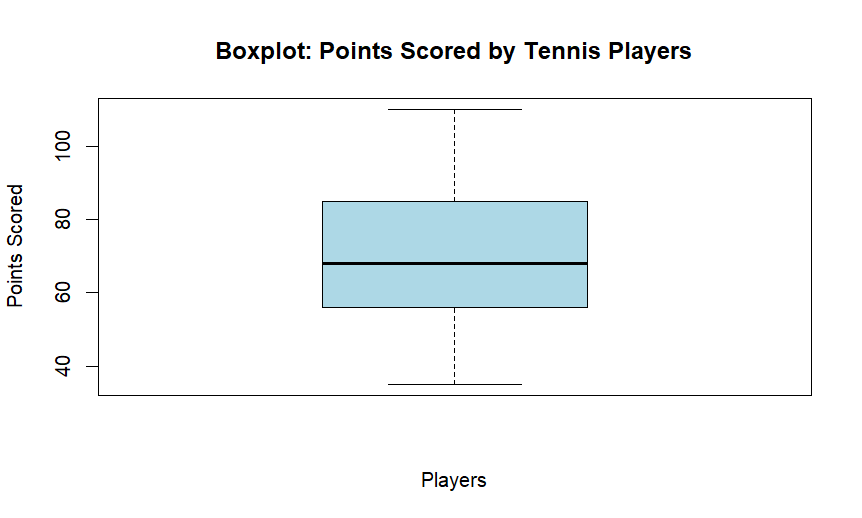
**Input:**

points\_scored <- c(35, 42, 48, 52, 56, 60, 62, 65, 68, 72, 76, 80, 85, 88, 92, 100, 110)

boxplot(points\_scored, main = "Boxplot: Points Scored by Tennis Players",

xlab = "Players", ylab = "Points Scored", col = "lightblue", border = "black")

**Output:**

****

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”)