DAY 2

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

(i)Use cov() to calculate the sample covariance between B and C.
(ii)Use another call to cov() to calculate the sample covariance matrix for the preferences.
(iii)Use cor() to calculate the sample correlation between B and C.
(iv)Use another call to cor() to calculate the sample correlation matrix for the preferences.

CODE:

```
(i)b<-c(22, 28, 10)
c<-c(20, 40, 40)
cov(b,c)

(ii)a<-c(18, 2, 20)
b<-c(22, 28, 10)
c<-c(20, 40, 40)
pre<-cbind(a,b,c)
cov(pre)

(iii).b<-c(22, 28, 10)
c<-c(20, 40, 40)
cor(b,c)

(iv)a<-c(18, 2, 20)
b<-c(22, 28, 10)
c<-c(20, 40, 40)
pre<-cbind(a,b,c)
cor(pre)
```

OUTPUT:

```
> b<-c(22, 28, 10)
> c<-c(20, 40, 40)
> cov(b,c)
[1] -20
> a<-c(18, 2, 20)
> b<-c(22, 28, 10)
> c<-c(20, 40, 40)
> pre<-cbind(a,b,c)
> cov(pre)
          a   b         c
a  97.33333 -74 -46.66667
b -74.00000  84 -20.00000
c -46.66667 -20 133.33333
> b<-c(22, 28, 10)
> c<-c(20, 40, 40)
> cor(b,c)
[1] -0.1889822
> a<-c(18, 2, 20)
> b<-c(22, 28, 10)
> c<-c(20, 40, 40)
> pre<-cbind(a,b,c)
> cor(pre)
           a          b          c
a  1.0000000 -0.8183918 -0.4096440
b -0.8183918  1.0000000 -0.1889822
c -0.4096440 -0.1889822  1.0000000
```

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,

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

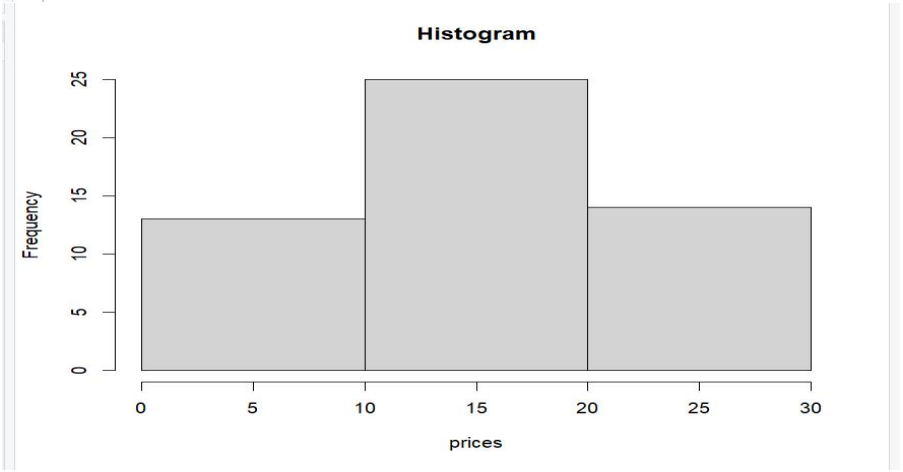
8, 8, 10, 10, 10, 10, 12, 14, 14, 14, 15, 15, 15, 15, 15, 15, 18, 18, 18, 18, 18,

CODE:

```
data<-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)
bin<-length(data)/3
bins<-cut(data, breaks = c(-Inf, quantile(data, probs = seq(0,1, 1/3)),Inf), include.lowest = TRUE)
tapply(data,bins,mean)
tapply(data, bins,function(x) c(min(x),max(x)))

hist(data,breaks = 3, main = "Histogram", xlab = "prices")
```

OUTPUT:

```
Console   Terminal   Background Jobs
R 4.2.2 · ~/
> data<-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, 1
8, 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)
> bin<-length(data)/3
> bins<-cut(data, breaks = c(-Inf, quantile(data, probs = seq(0,1, 1/3)),Inf), include.lowest = TRU
E)
> tapply(data,bins,mean)
 [-Inf,1]     (1,15]    (15,20]    (20,30] (30, Inf]
  1.00000  10.71429  18.93333  25.35714         NA
> tapply(data, bins,function(x) c(min(x),max(x)))
$`[-Inf,1]`
[1] 1 1

$`(1,15]`
[1]  5 15

$`(15,20]`
[1] 18 20

$`(20,30]`
[1] 21 30

$`(30, Inf]`
NULL

>
> hist(data,breaks = 3, main = "Histogram", xlab = "prices")
>
```



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

CODE:

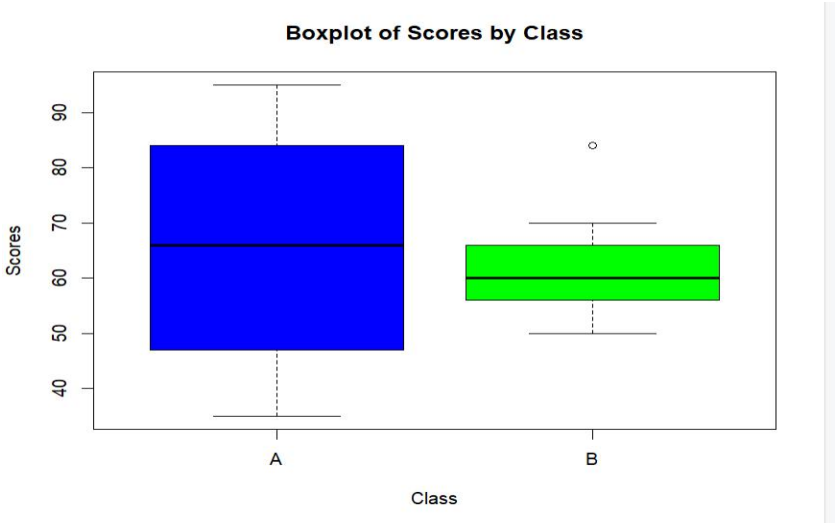
```
A <- c(76, 35, 47, 64, 95, 66, 89, 36, 84)
B <- c(51, 56, 84, 60, 59, 70, 63, 66, 50)

mean_A <- mean(A)
median_A <- median(A)
range_A <- max(A) - min(A)
mean_B <- mean(B)
```

```
median_B <- median(B)
range_B <- max(B) - min(B)

combined_data <- data.frame(Class = c(rep("A", length(A)), rep("B", length(B))), Score = c(A, B))
boxplot(Score ~ Class, data = combined_data, col = c("blue", "green"), xlab = "Class", ylab = "Scores", main = "Boxplot of Scores by Class")
```

(II)

```
combined_data <- data.frame(Class = c(rep("A", length(A)), rep("B", length(B))), Score = c(A, B))

boxplot(Score ~ Class, data = combined_data, col = c("blue", "green"), xlab = "Class", ylab = "Scores", main = "Boxplot of Scores by Class")
```

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

CODE:

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

min_value <- 50000
max_value <- 100000
v <- 80
min_max_normalized <- (v - min_value) / (max_value - min_value)
min_max_normalized
mean_value <- mean(data)
```

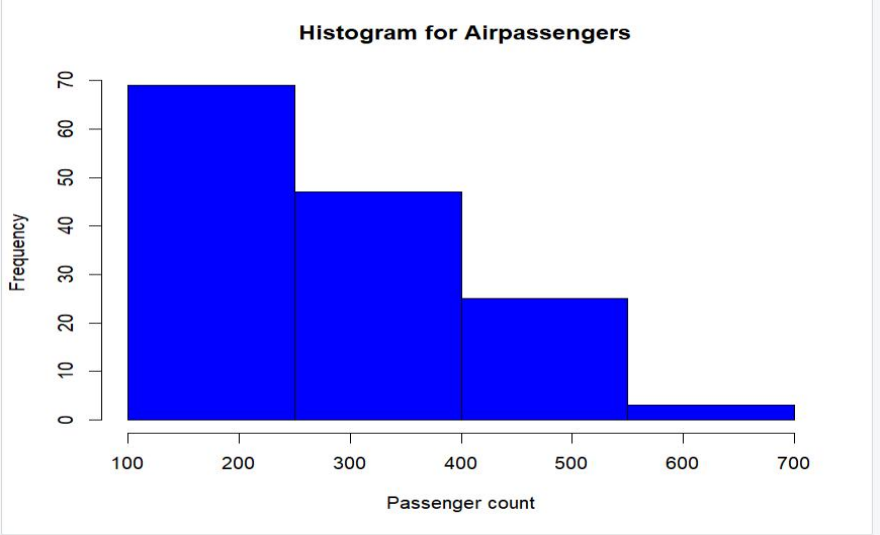
standard_deviation <- sd(data)
z_score_normalized <- (v - mean_value) / standard_deviation
z_score_normalized
OUTPUT:

```
> data <- c(200, 300, 400, 600, 1000)
>
> min_value <- 50000
> max_value <- 100000
> v <- 80
> min_max_normalized <- (v - min_value) / (max_value - min_value)
> min_max_normalized
[1] -0.9984
> mean_value <- mean(data)
> standard_deviation <- sd(data)
> z_score_normalized <- (v - mean_value) / standard_deviation
> z_score_normalized
[1] -1.328157
```

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

CODE:
data("AirPassengers")
hist(AirPassengers, breaks = seq(100, 700, by = 150), col = "blue", main=" Histogram for Airpassengers", xlab = "Passenger count", ylab = "Frequency")

OUTPUT:

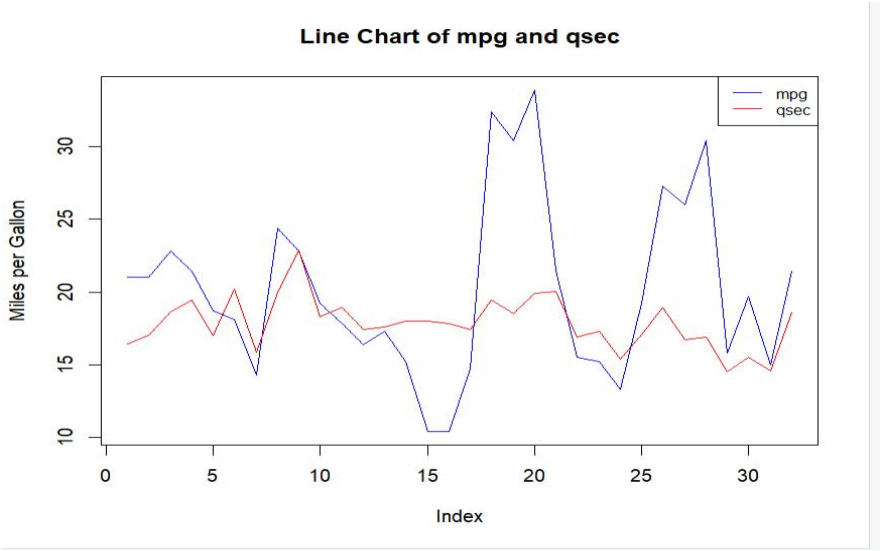


6.Obtain Multiple Lines in Line Chart using a single Plot Function in R.Use attributes“mpg”and“qsec”of the dataset “mtcars”

CODE:
data("mtcars")

plot(mtcars$mpg, type = "l", col = "blue", xlab = "Index", ylab = "Miles per Gallon", main = "Line Chart of mpg and qsec")
lines(mtcars$qsec, col = "red")
legend("topright", legend = c("mpg", "qsec"), col = c("blue", "red"), lty = 1, cex = 0.8)

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.

CODE:

```
data("iris")
str(iris)
plot(iris$Sepal.Length, iris$Petal.Length, main = "Scatter plot of Sepal.Length vs. Petal.Length",xlab = "Sepal.Length", ylab = "Petal.Length", col = "blue", pch = 16)
model <- lm(Petal.Length ~ Sepal.Length, data = iris)
abline(model, col = "red")
new_data <- data.frame(Sepal.Length = 5.5)
predicted_Petal_Length <- predict(model, newdata = new_data)
predicted_Petal_Length
```

OUTPUT:

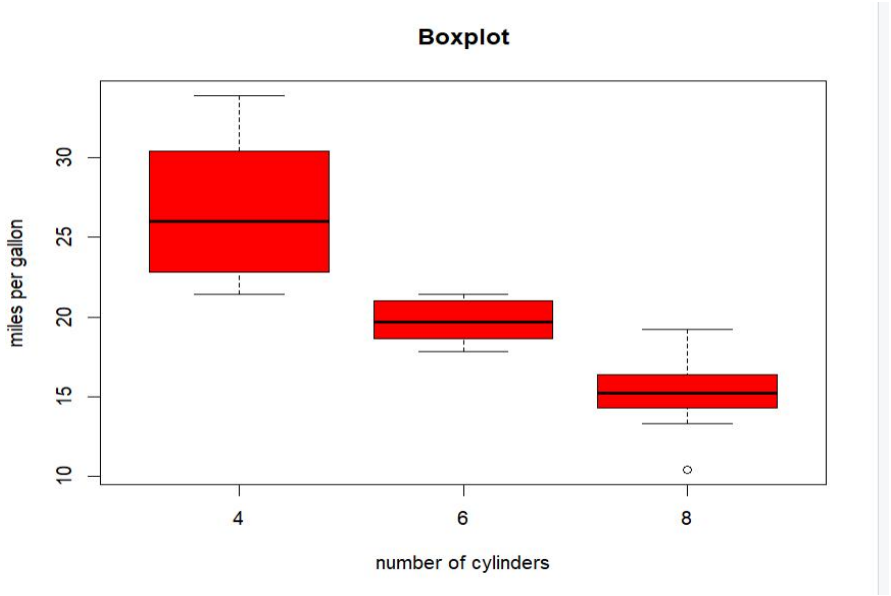
```
> str(iris)
'data.frame':   150 obs. of  5 variables:
 $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
 $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
 $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
 $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
 $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ...
> plot(iris$Sepal.Length, iris$Petal.Length, main = "Scatter plot of Sepal.Length vs. Petal.Lengt
h",xlab = "Sepal.Length", ylab = "Petal.Length", col = "blue", pch = 16)
> model <- lm(Petal.Length ~ Sepal.Length, data = iris)
> abline(model, col = "red")
> new_data <- data.frame(Sepal.Length = 5.5)
> predicted_Petal_Length <- predict(model, newdata = new_data)
> predicted_Petal_Length
       1
3.119938
>
```

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.

CODE:

```
data("mtcars")
boxplot(mpg ~ cyl, data = mtcars, main = "Boxplot", xlab = "number of cylinders", ylab = "miles per gallon", col= "red")
```

OUTPUT:

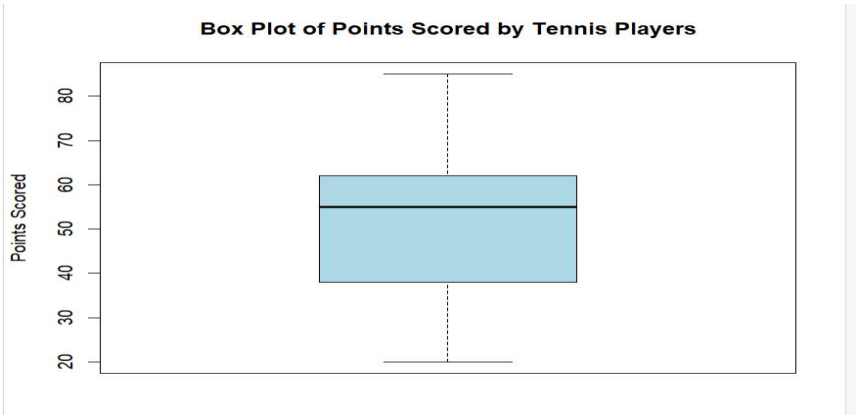


9. Assume the Tennis coach wants to determine if any of his team players are scoring outliers. To visualize the distribution of points scored by his players, then how can he decide to develop the box plot? Give suitable example using Boxplot visualization technique.

CODE:

```
score <- c(20, 25, 30, 32, 35, 38, 40, 45, 50, 52, 55, 56, 58, 59, 60, 62, 65, 70, 75, 80, 85)

boxplot(score, col = "lightblue", main = "Box Plot of Points Scored by Tennis Players", ylab = "Points Scored")
```

OUTPUT:

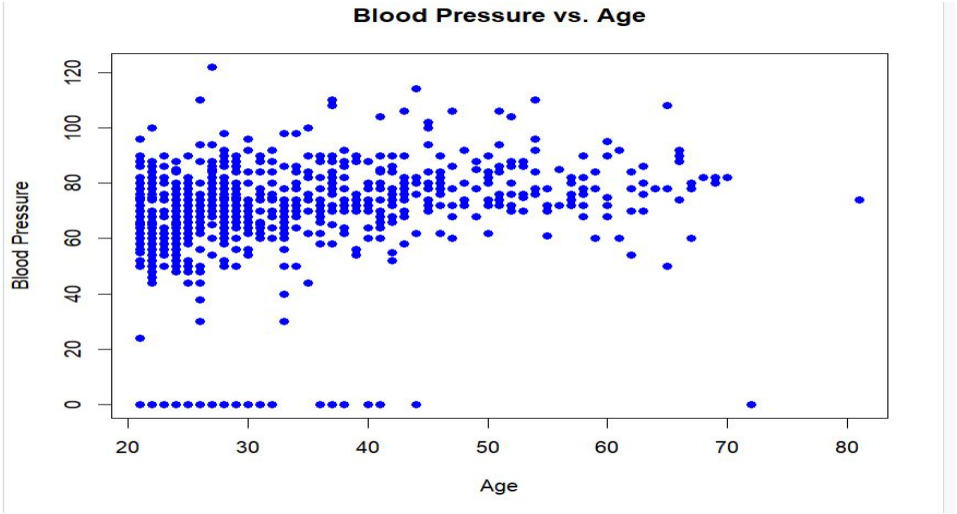


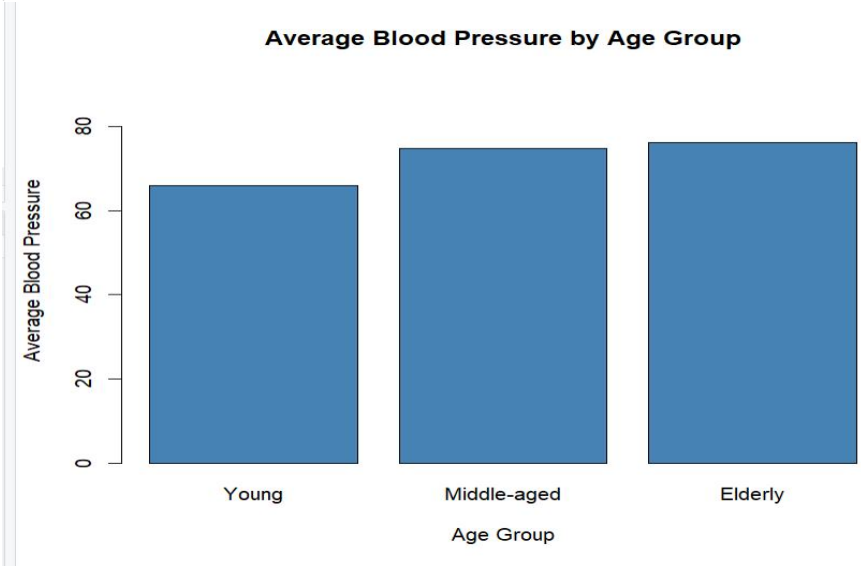
10. Implement using R language in which age group of people are affected by blood pressure based on the diabetes dataset show it using scatterplot and bar chart (that is Blood Pressure vs Age using dataset “diabetes.csv”)

CODE:

```
dia<-read.csv("C://Users//FLORENCIA ABEL//OneDrive//Documents//diabetes.csv")
View(dia)

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

age_group_labels <- cut(dia$Age, breaks = c(0, 35, 55, Inf), labels = c("Young", "Middle-aged", "Elderly"))

age_group_avg_bp <- tapply(dia$BloodPressure, age_group_labels, mean)

barplot(age_group_avg_bp, main = "Average Blood Pressure by Age Group",xlab = "Age Group",ylab = "Average Blood Pressure", col = "steelblue", ylim = c(0, max(age_group_avg_bp) * 1.2))
```

OUTPUT:


Blood Pressure vs. Age
Blood Pressure
0
20
40
60
80
100
120
20
30
40
50
60
70
80
Age


Average Blood Pressure by Age Group
Average Blood Pressure
0
20
40
60
80
Young
Middle-aged
Elderly
Age Group