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

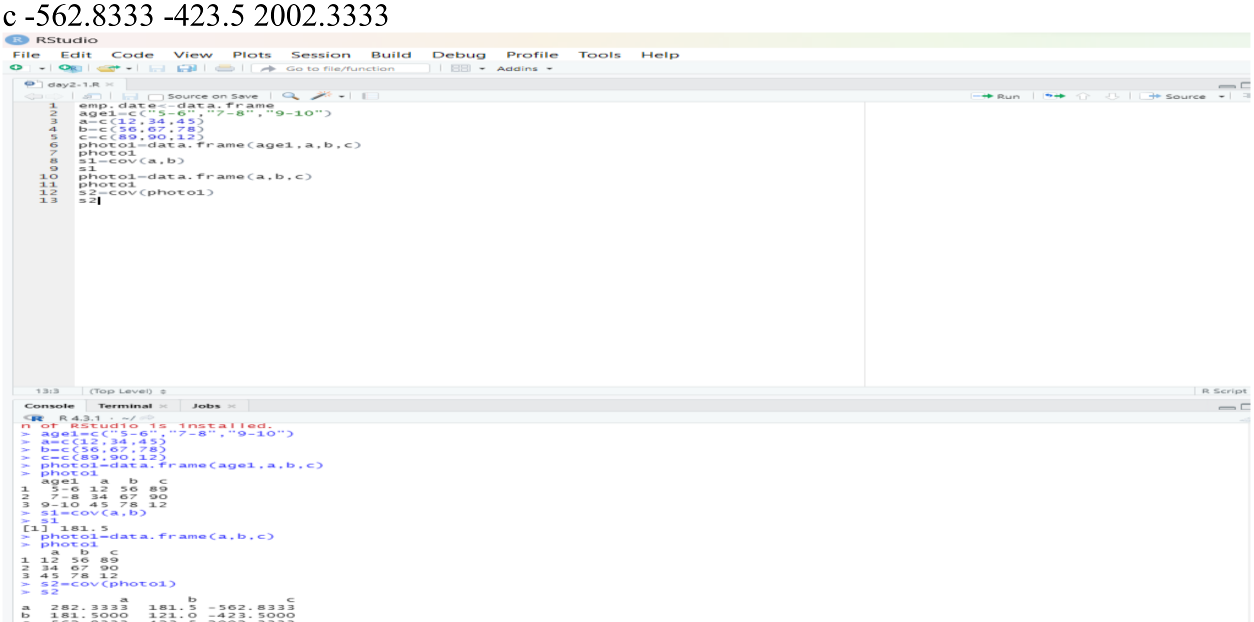
1. Use cov() to calculate the sample covariance between B and C.
2. Use another call to cov() to calculate the sample covariance matrix for the preferences.
3. Use cor() to calculate the sample correlation between B and C.
4. Use another call to cor() to calculate the sample correlation matrix for the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

INPUT

```
emp.date<-data.frame
age1=c("5-6","7-8","9-10")
a=c(12,34,45)
b=c(56,67,78)
c=c(89,90,12)
photo1=data.frame(age1,a,b,c)
photo1
s1=cov(a,b)
s1
photo1=data.frame(a,b,c)
photo1
s2=cov(photo1)
s2
```

OUTPUT

```
> emp.date<-data.frame
Warning message:
R graphics engine version 16 is not supported by this version of RStudio. The Plots tab will be disabled until a newer version of RStudio is installed.
> age1=c("5-6","7-8","9-10")
> a=c(12,34,45)
> b=c(56,67,78)
> c=c(89,90,12)
> photo1=data.frame(age1,a,b,c)
> photo1
  age1  a  b  c
1  5-6 12 56 89
2  7-8 34 67 90
3 9-10 45 78 12
> s1=cov(a,b)
> s1
[1] 181.5
> photo1=data.frame(a,b,c)
> photo1
   a  b  c
1 12 56 89
2 34 67 90
3 45 78 12
> s2=cov(photo1)
> s2
          a      b         c
a  282.3333  181.5 -562.8333
b  181.5000  121.0 -423.5000
c -562.8333 -423.5 2002.3333
```



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,

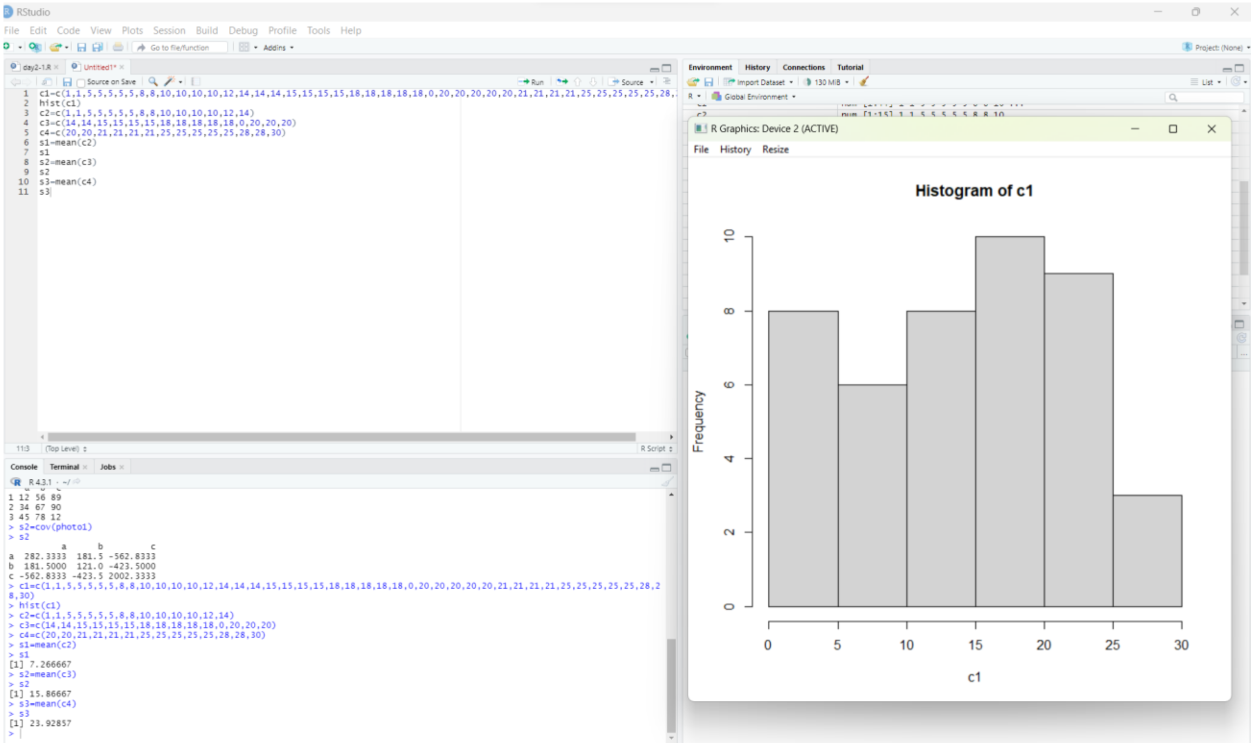
dataset using an equal-frequency partitioning method with bin equal to 3 (ii) apply data smoothing using bin means and bin boundary.
(iii) Plot Histogram for the above frequency division

INPUT

```
c1=c(1,1,5,5,5,5,5,8,8,10,10,10,10,12,14,14,14,15,15,15,15,18,18,18,18,18,0,20,20,20,20,20,21,
21,21,21,25,25,25,25,25,28,28,30)
hist(c1)
c2=c(1,1,5,5,5,5,5,8,8,10,10,10,10,12,14)
c3=c(14,14,15,15,15,15,18,18,18,18,18,0,20,20,20)
c4=c(20,20,21,21,21,21,25,25,25,25,25,28,28,30)
s1=mean(c2)
s1
s2=mean(c3)
s2
s3=mean(c4)
s3
```

OUTPUT

```
>c1=c(1,1,5,5,5,5,5,8,8,10,10,10,10,12,14,14,14,15,15,15,15,18,18,18,18,18,0,20,20,20,20,20,21
,21,21,21,25,25,25,25,25,28,28,30)
> hist(c1)
> c2=c(1,1,5,5,5,5,5,8,8,10,10,10,10,12,14)
> c3=c(14,14,15,15,15,15,18,18,18,18,18,0,20,20,20)
> c4=c(20,20,21,21,21,21,25,25,25,25,25,28,28,30)
> s1=mean(c2)
> s1
[1] 7.266667
> s2=mean(c3)
> s2
[1] 15.86667
> s3=mean(c4)
> s3
[1] 23.92857
```



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

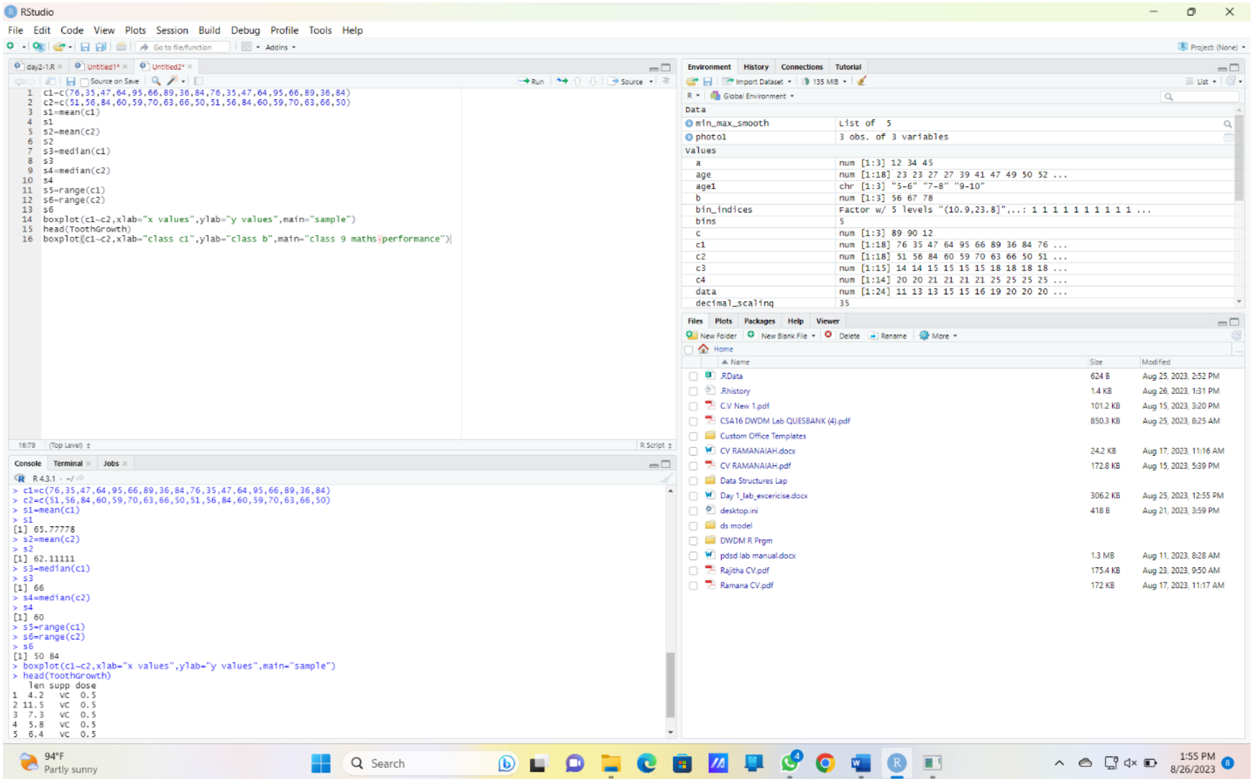
INPUT

```
c1=c(76,35,47,64,95,66,89,36,84,76,35,47,64,95,66,89,36,84)
c2=c(51,56,84,60,59,70,63,66,50,51,56,84,60,59,70,63,66,50)
s1=mean(c1)
s1
s2=mean(c2)
s2
```

```
s3=median(c1)
s3
s4=median(c2)
s4
s5=range(c1)
s6=range(c2)
s6
boxplot(c1~c2,xlab="x values",ylab="y values",main="sample")
head(ToothGrowth)
boxplot(c1~c2,xlab="class c1",ylab="class b",main="class 9 maths performance")
```

OUTPUT

```
> c1=c(76,35,47,64,95,66,89,36,84,76,35,47,64,95,66,89,36,84)
> c2=c(51,56,84,60,59,70,63,66,50,51,56,84,60,59,70,63,66,50)
> s1=mean(c1)
> s1
[1] 65.77778
> s2=mean(c2)
> s2
[1] 62.11111
> s3=median(c1)
> s3
[1] 66
> s4=median(c2)
> s4
[1] 60
> s5=range(c1)
> s6=range(c2)
> s6
[1] 50 84
> boxplot(c1~c2,xlab="x values",ylab="y values",main="sample")
> head(ToothGrowth)
   len supp dose
1  4.2   VC  0.5
2 11.5   VC  0.5
3  7.3   VC  0.5
4  5.8   VC  0.5
5  6.4   VC  0.5
6 10.0   VC  0.5
> boxplot(c1~c2,xlab="class c1",ylab="class b",main="class 9 maths performance")
```



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

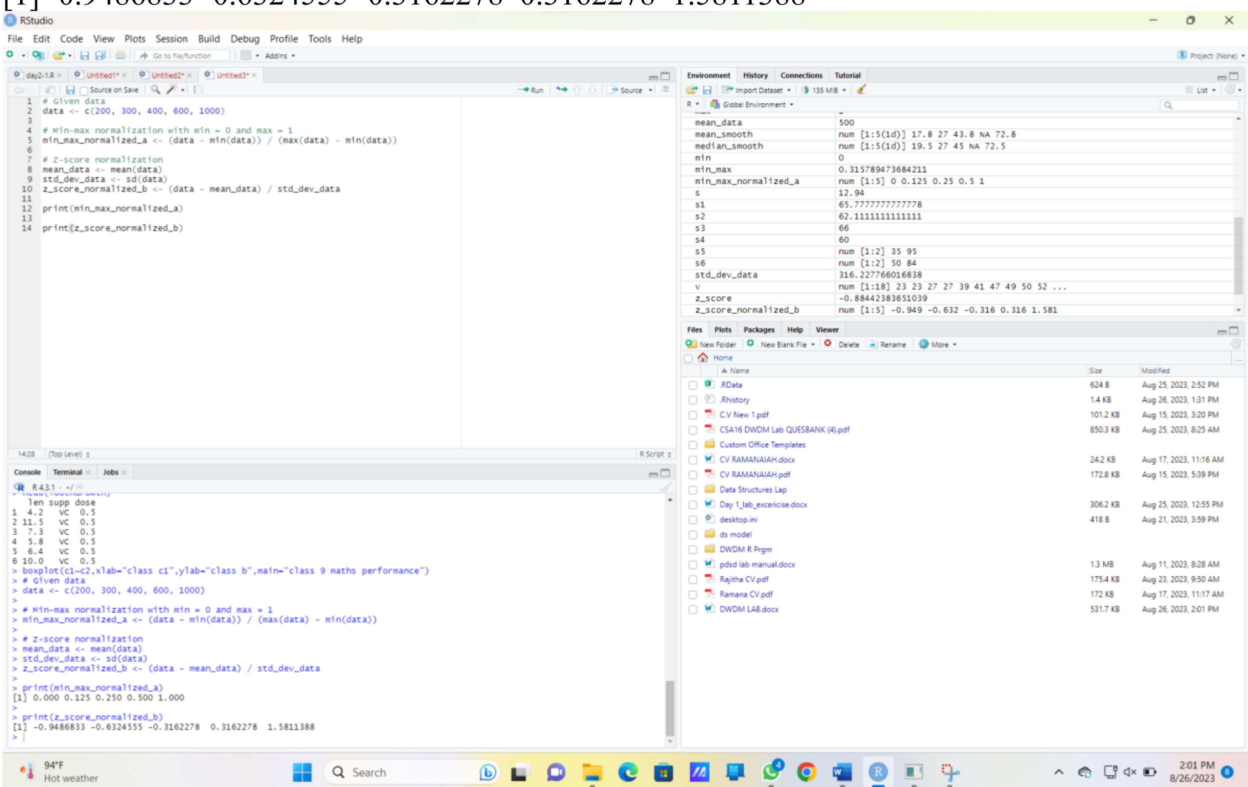
(b) z-score normalization

INPUT

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

# Min-max normalization with min = 0 and max = 1
min_max_normalized_a <- (data - min(data)) / (max(data) - min(data))

# Z-score normalization
mean_data <- mean(data)
std_dev_data <- sd(data)
z_score_normalized_b <- (data - mean_data) / std_dev_data

print(min_max_normalized_a)

print(z_score_normalized_b)
```

OUTPUT

```
> boxplot(c1~c2,xlab="class c1",ylab="class b",main="class 9 maths performance")
> # Given data
> data <- c(200, 300, 400, 600, 1000)
>
> # Min-max normalization with min = 0 and max = 1
> min_max_normalized_a <- (data - min(data)) / (max(data) - min(data))
>
> # Z-score normalization
> mean_data <- mean(data)
> std_dev_data <- sd(data)
> z_score_normalized_b <- (data - mean_data) / std_dev_data
>
> print(min_max_normalized_a)
[1] 0.000 0.125 0.250 0.500 1.000
>
> print(z_score_normalized_b)
[1] -0.9486833 -0.6324555 -0.3162278  0.3162278  1.5811388
```

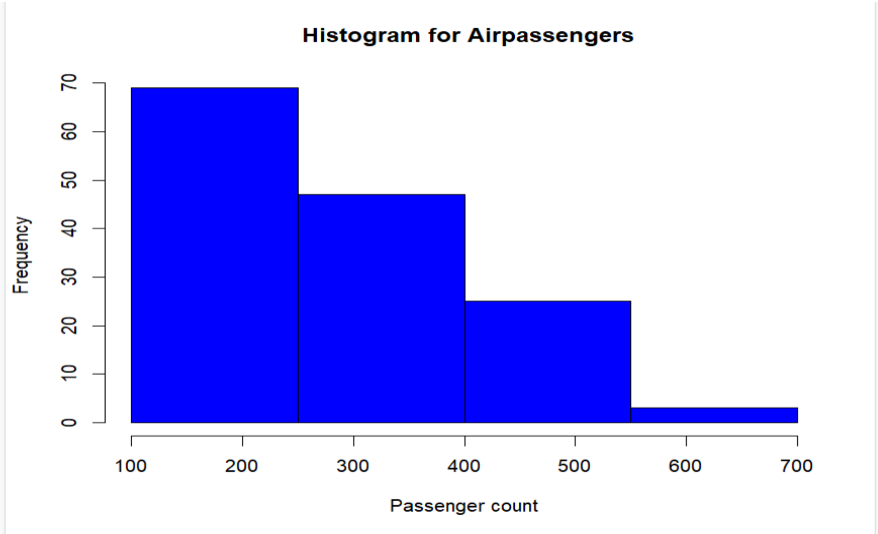


5.Make a histogram for the “AirPassengers “dataset, start at 100 on the x-axis, and from values 200 to 700, make the bins 150 wide

INPUT

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

OUTPUT

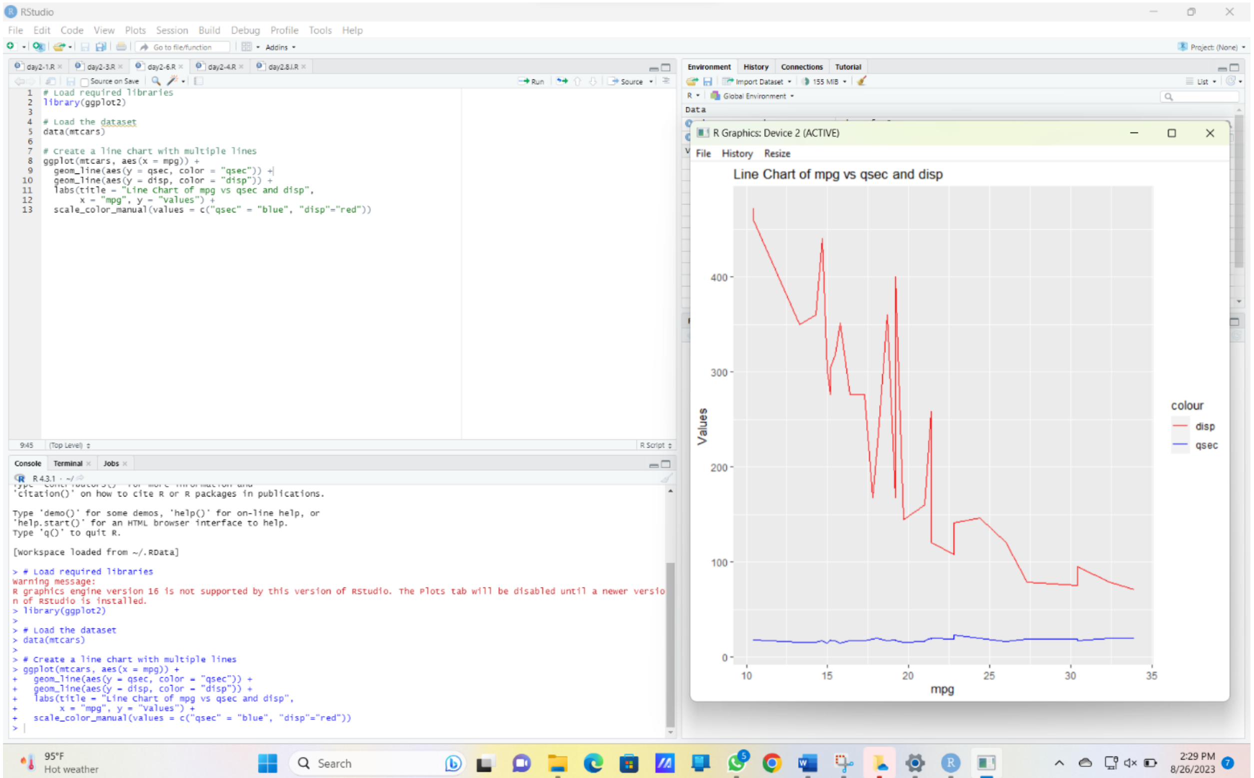


6.Obtain Multiple Lines in Line Chart using a single Plot Function in R.Use attributes"mpg"and"qsec"of the dataset "mtcars"

INPUT

```
# Load required libraries
library(ggplot2)

# Load the dataset
data(mtcars)

# Create a line chart with multiple lines
ggplot(mtcars, aes(x = mpg)) +
  geom_line(aes(y = qsec, color = "qsec")) +
  geom_line(aes(y = disp, color = "disp")) +
  labs(title = "Line Chart of mpg vs qsec and disp",
       x = "mpg", y = "Values") +
  scale_color_manual(values = c("qsec" = "blue", "disp"="red"))
```

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("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

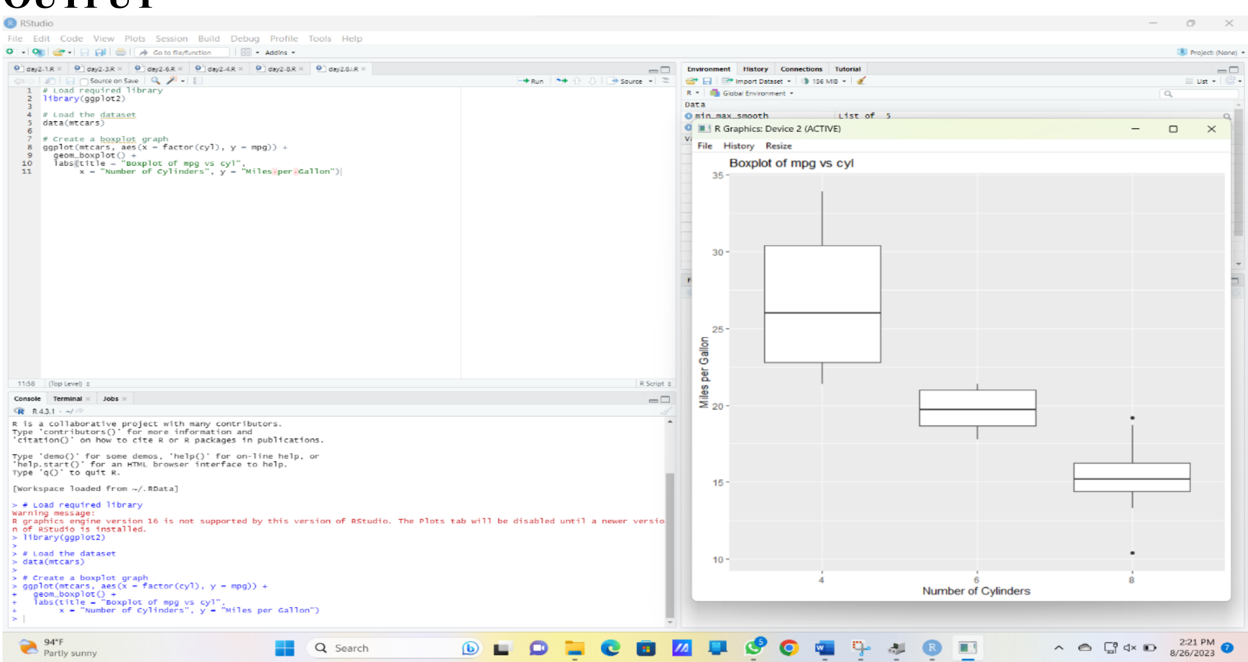
```
> str(iris)
'data.frame':   150 obs. of  5 variables:
 $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
 $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
 $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
 $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 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.

INPUT

```
# Load required library
library(ggplot2)

# Load the dataset
data(mtcars)

# Create a boxplot graph
ggplot(mtcars, aes(x = factor(cyl), y = mpg)) +
  geom_boxplot() +
  labs(title = "Boxplot of mpg vs cyl",
       x = "Number of Cylinders", y = "Miles per Gallon")
```

OUTPUT



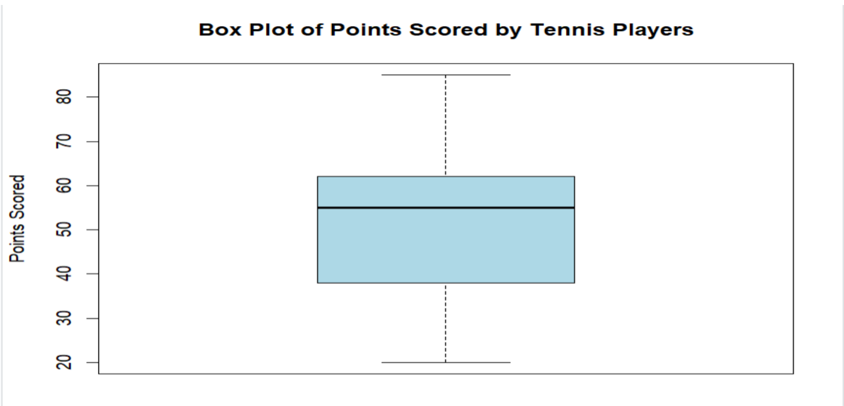
9. Assume the Tennis coach wants to determine if any of his team players are scoring outliers. To visualize the distribution of points scored by his players, then how can he decide to develop the box plot? Give suitable example using Boxplot visualization technique.

INPUT

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”)

INPUT

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

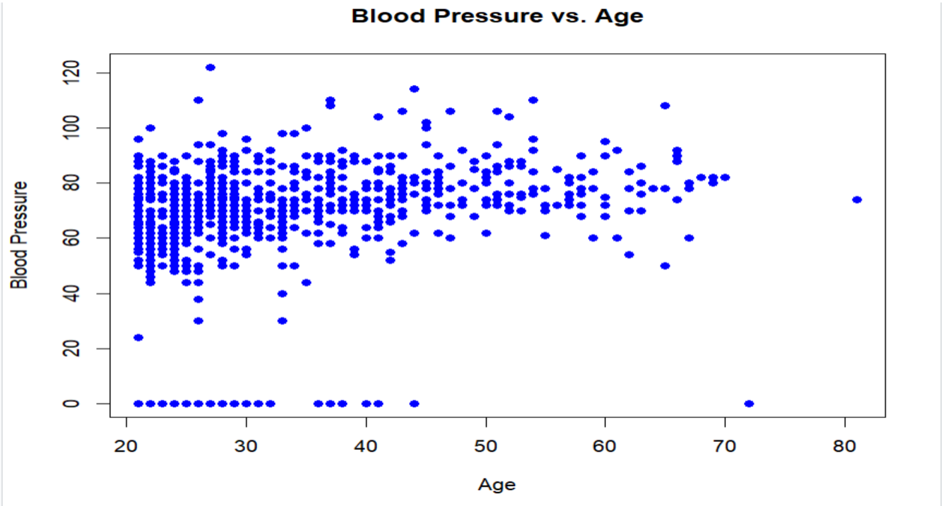
plot(diaAge, diaBloodPressure, xlab = "Age", ylab = "Blood Pressure", main = "Blood Pressure vs. Age", col = "blue",pch = 16)

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

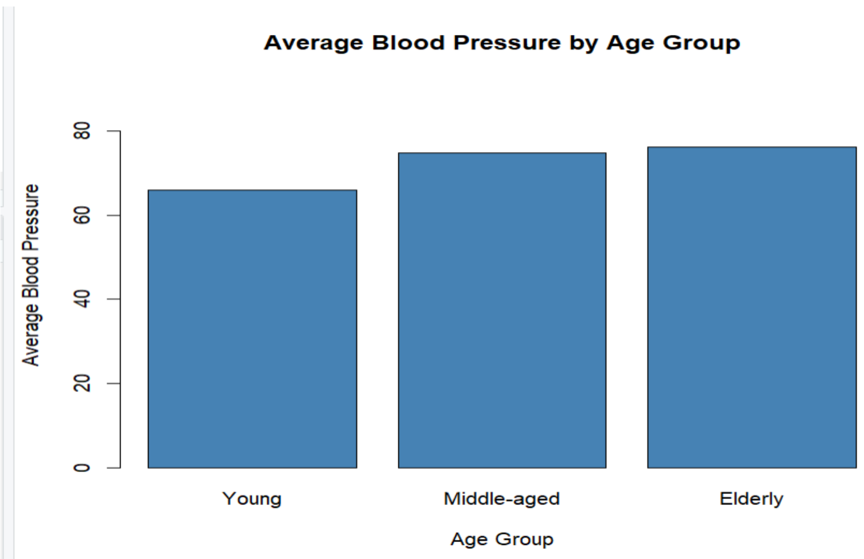
age_group_avg_bp <- tapply(dia$BloodPressure, age_group_labels, mean)

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