Answer ALL questions.

Before you begin answering the questions, you are required to install 1 package by using the following:

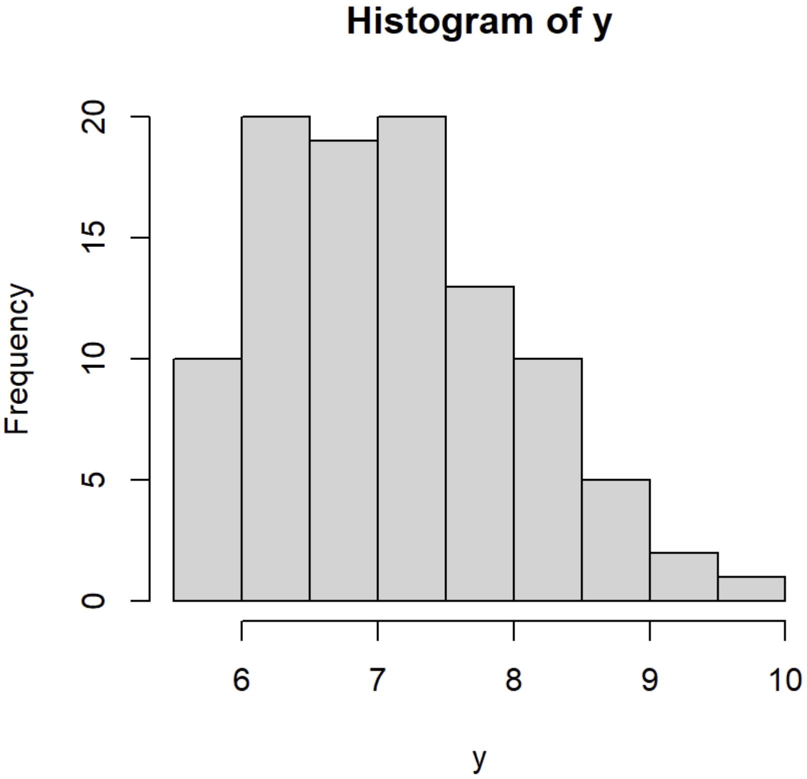
```
install.packages(“moments”)
library(“moments”)
```

The questions below are based on the given dataset in CSV format, **LabTestData.csv**. Answer all the questions by using **R programming language** whenever necessary. Show the main **R code** used and display the R results, in each relevant part of the questions.

QUESTION 1 (25 marks)

(a) (i) R code: hist(y)

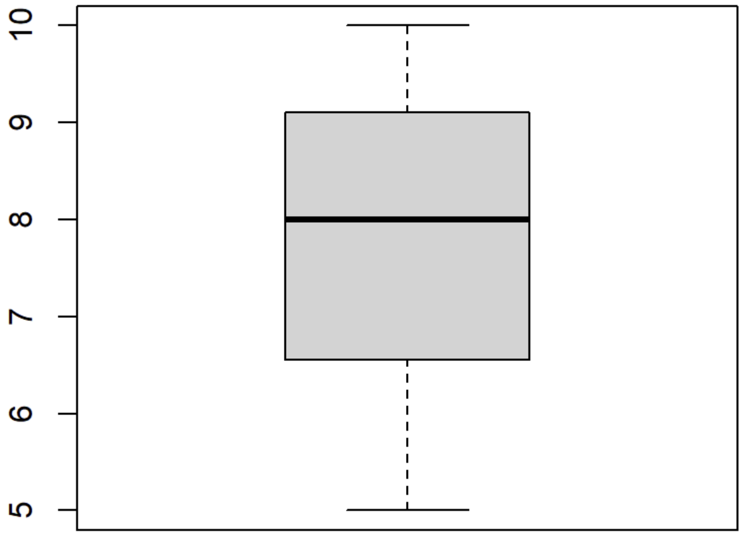
<<Insert R output here>>



(ii) Majority of values lie on the left side of the histogram while only a small number of extreme high values lie on the right side. The histogram of y has a positively skewed distribution.

(b) (i) R code: boxplot(x2)

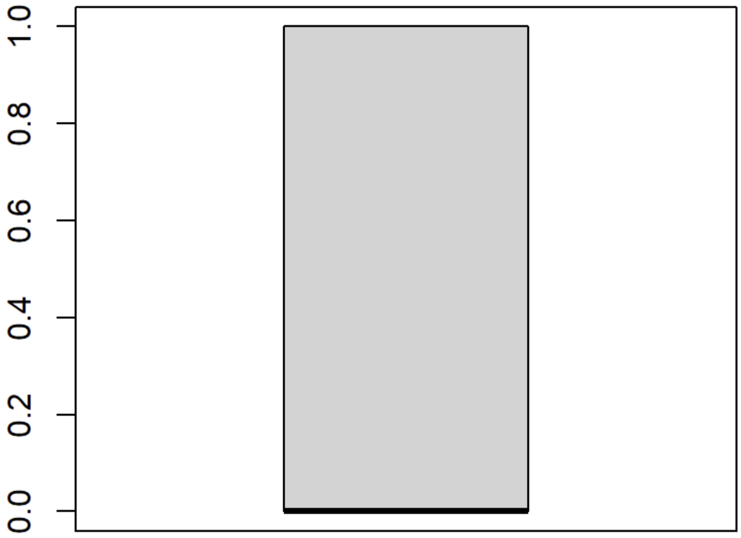
<<Insert R output here>>



(ii) The upper portion of the box is slightly smaller than the lower portion. x2 has a negatively skewed distribution.

(iii) R code: boxplot(x1)

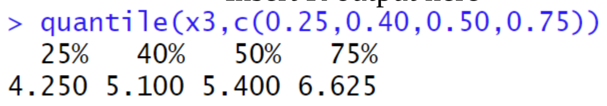
<<Insert R output here>>



The boxplot cannot be used to describe the shape of distribution of x1 because it has an entirely positive distribution.

(c) (i) R code: quantile(x3,c(0.25,0.40,0.50,0.75))

<<Insert R output here>>

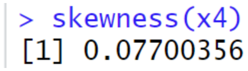


```
> quantile(x3,c(0.25,0.40,0.50,0.75))
  25%   40%   50%   75% 
4.250 5.100 5.400 6.625 
```

(ii) It means that 40% of the values are below 5.100.

(iii) R code: skewness(x4)

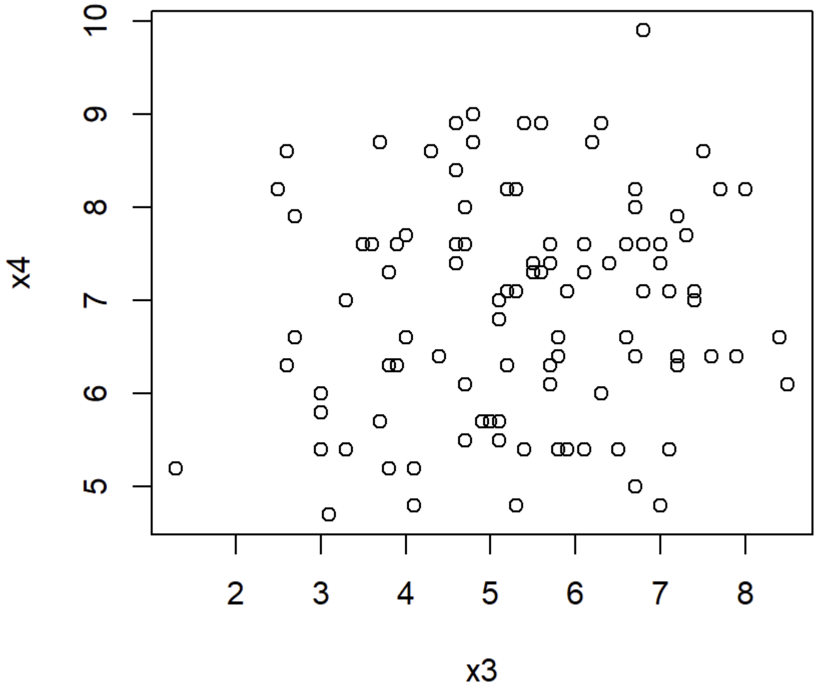
<<Insert R output here>>



```
> skewness(x4)
[1] 0.07700356
```

(d) (i) R code: plot(x3,x4)

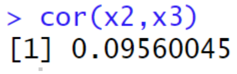
<<Insert R output here>>



(ii) The correlation between x3 and x4 is weak as the arrangement of the points in the scatter plot does not show a clear straight line.

(iii) R code: cor(x2,x3)

<<Insert R output here>>

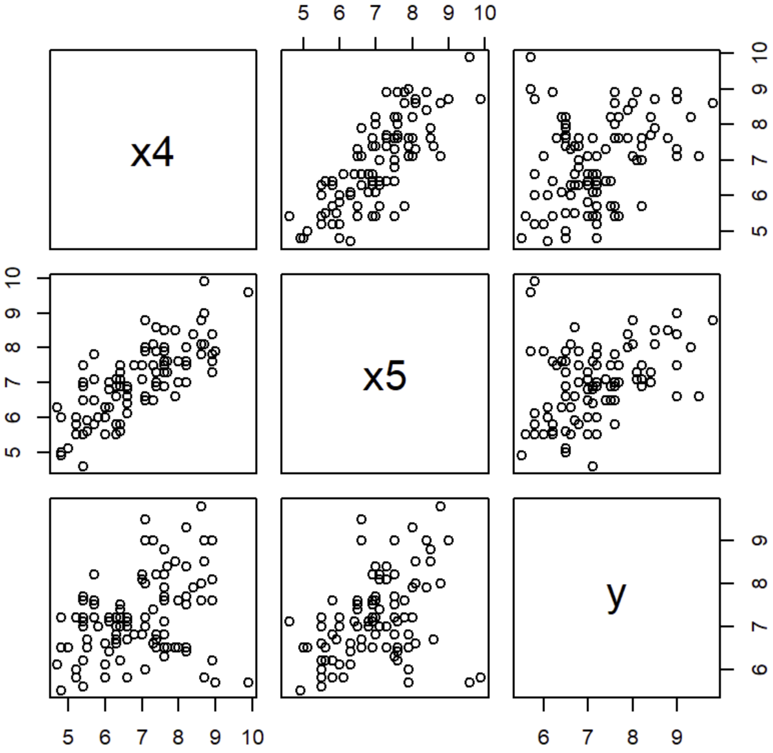


```
> cor(x2,x3)
[1] 0.09560045
```

QUESTION 2 (25 marks)

(a) (i) R code: plot(LabTestData[4:6])

<<Insert R output here>>

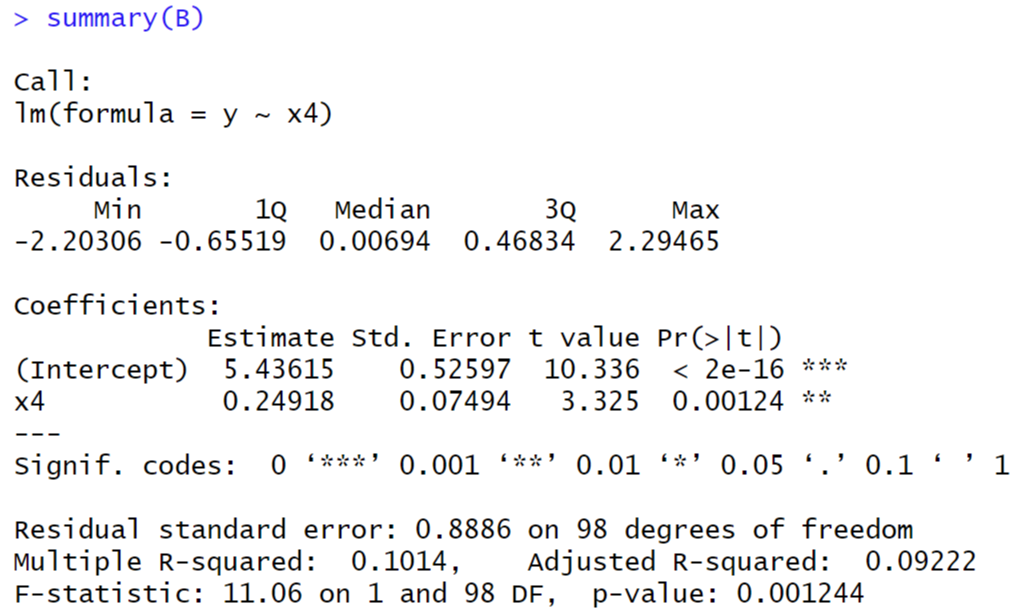


(ii) Variables x4 and x5 have the highest correlation because the points in the scatterplot for these two variables show a clear straight line pattern.

(b) (i) R code:

```
B=lm(y~x4)
B
summary(B)
```

<<Insert R output here>>



```
> summary(B)

Call:
lm(formula = y ~ x4)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.20306 -0.65519  0.00694  0.46834  2.29465 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  5.43615    0.52597  10.336  < 2e-16 ***
x4           0.24918    0.07494   3.325  0.00124 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8886 on 98 degrees of freedom
Multiple R-squared:  0.1014,	Adjusted R-squared:  0.09222 
F-statistic: 11.06 on 1 and 98 DF,  p-value: 0.001244
```

(ii) y=5.43615+0.24918x4

(iii) a=5.43615 means that the predicted value of y will be 5.43615 if x4=0.
b=0.24918 means that the value of y increases by 0.24918 if x4 increases by 1.

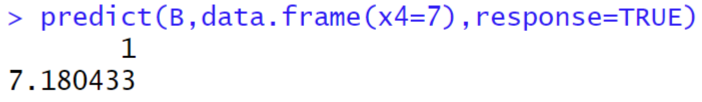
(iv) R^2=0.1014 means that the variance is 0.1014.

(v) H_0: x4 does not significantly affect y.
H_1: x4 does significantly affect y.

(vi) Since the p-value < 0.05, H_0 is rejected. x4 is a significant factor of y.

(vii) R code: predict(B,data.frame(x4=7),response=TRUE)

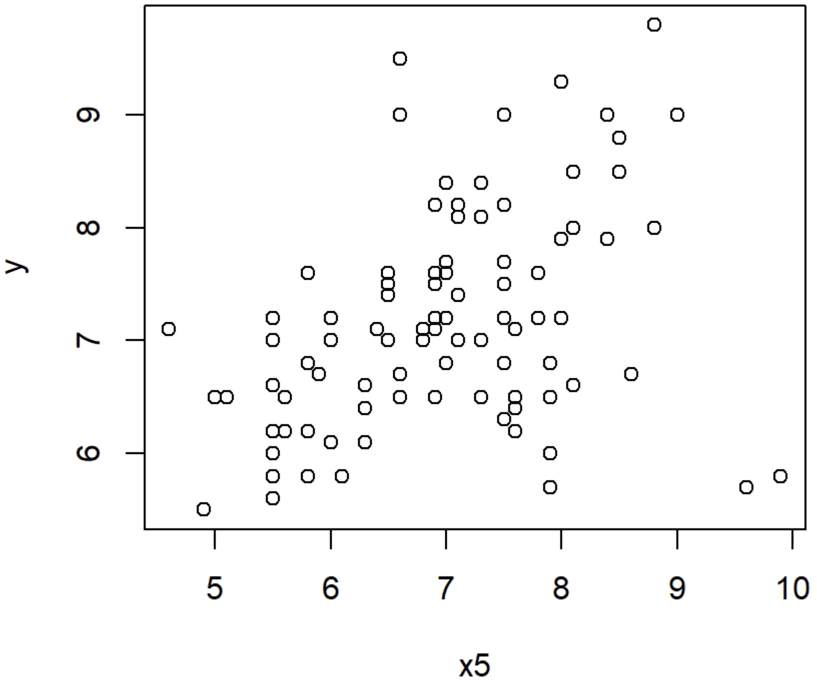
<<Insert R output here>>



```
> predict(B,data.frame(x4=7),response=TRUE)
       1 
7.180433 
```

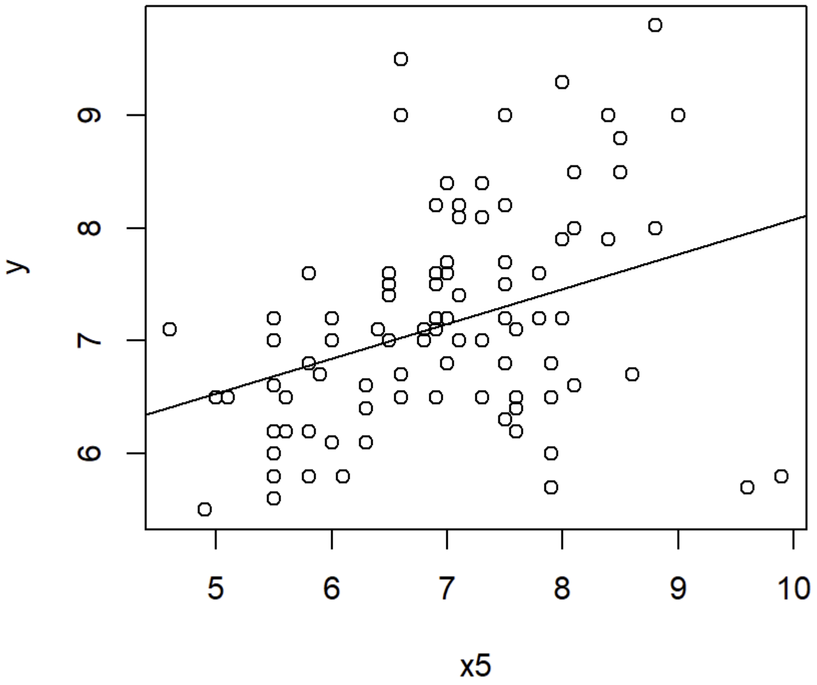
(c) (i) R code: plot(x5,y)

<<Insert R output here>>



(ii) R code: C=lm(y~x5)
abline(C)

<<Insert R output here>>


y
6
7
8
9
5
6
7
8
9
10
x5