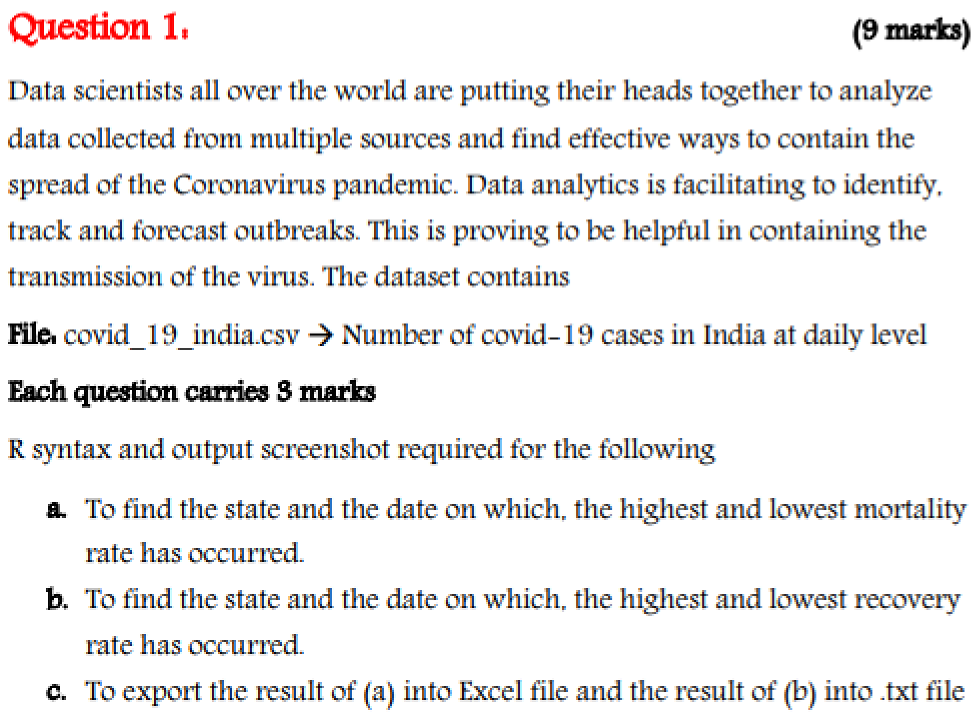
18BCE0272
Nitin Pramod Ranjan

CSE 3020 Data Visualization

Lab Assessment -4

18BCE0272



Question 1: **(9 marks)**

Data scientists all over the world are putting their heads together to analyze data collected from multiple sources and find effective ways to contain the spread of the Coronavirus pandemic. Data analytics is facilitating to identify, track and forecast outbreaks. This is proving to be helpful in containing the transmission of the virus. The dataset contains

File: covid_19_india.csv → Number of covid-19 cases in India at daily level

Each question carries 3 marks

R syntax and output screenshot required for the following

a. To find the state and the date on which, the highest and lowest mortality rate has occurred.

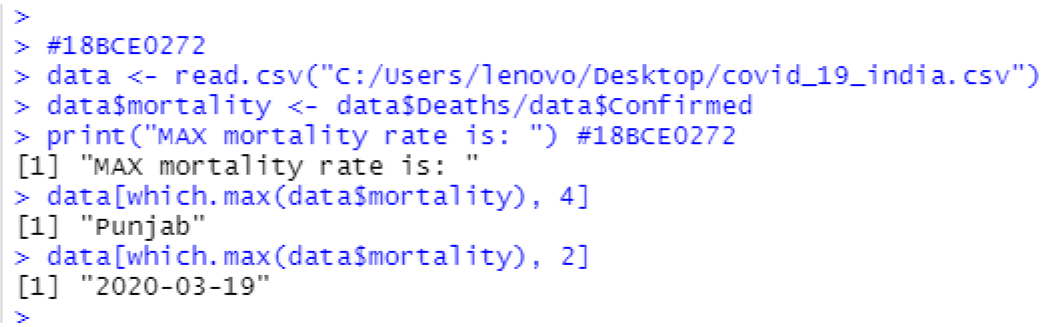
b. To find the state and the date on which, the highest and lowest recovery rate has occurred.

c. To export the result of (a) into Excel file and the result of (b) into .txt file

Question 1 (a):

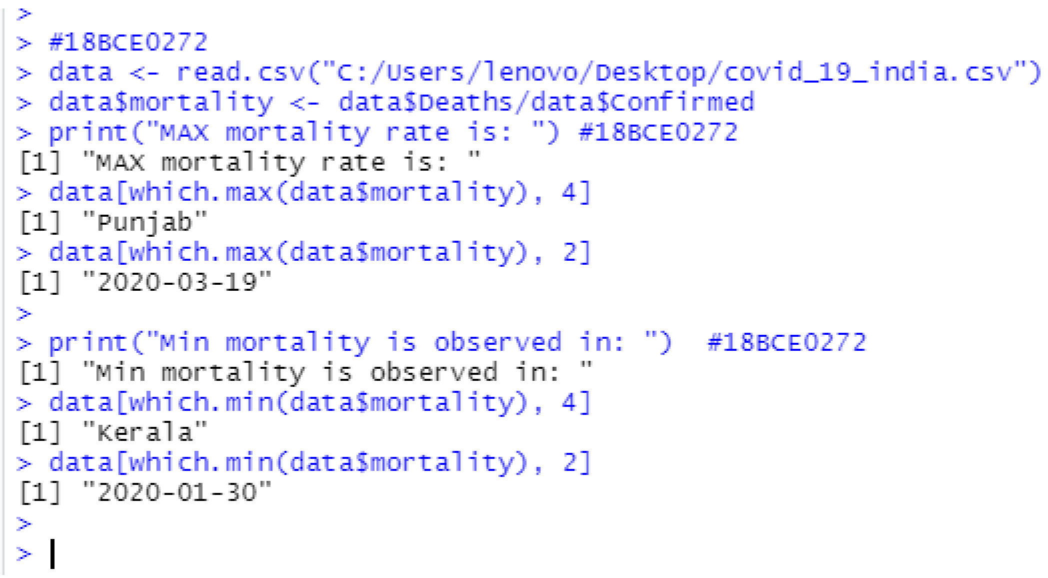
CODE AND OUTPUT:

Max and min mortality rate



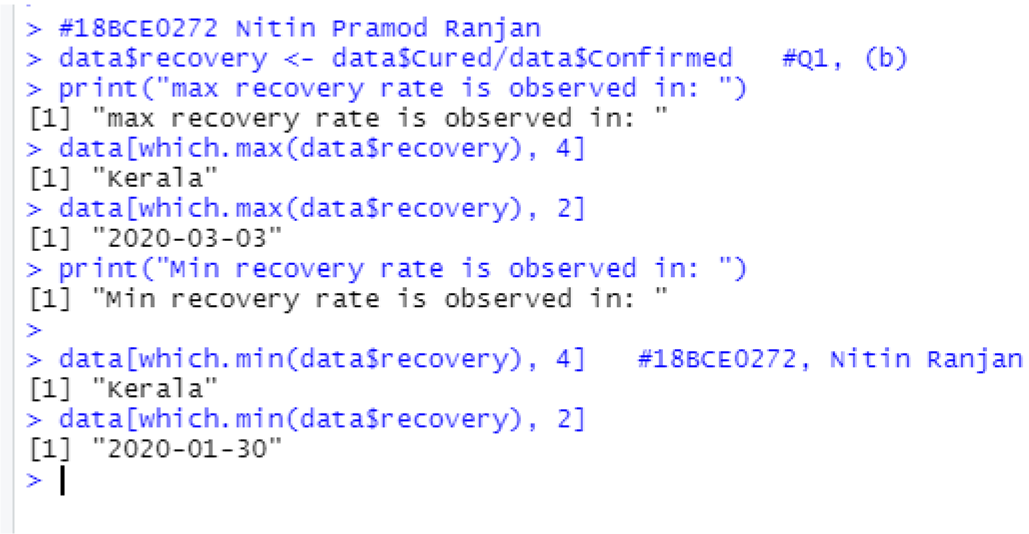
```
>
> #18BCE0272
> data <- read.csv("C:/Users/lenovo/Desktop/covid_19_india.csv")
> data$mortality <- data$Deaths/data$Confirmed
> print("MAX mortality rate is: ") #18BCE0272
[1] "MAX mortality rate is: "
> data[which.max(data$mortality), 4]
[1] "Punjab"
> data[which.max(data$mortality), 2]
[1] "2020-03-19"
>
```

18BCE0272
Nitin Pramod Ranjan



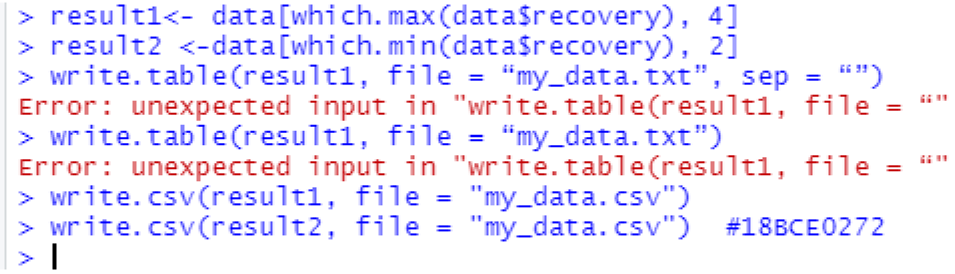
```
>
> #18BCE0272
> data <- read.csv("C:/Users/lenovo/Desktop/covid_19_india.csv")
> data$mortality <- data$Deaths/data$Confirmed
> print("MAX mortality rate is: ") #18BCE0272
[1] "MAX mortality rate is: "
> data[which.max(data$mortality), 4]
[1] "Punjab"
> data[which.max(data$mortality), 2]
[1] "2020-03-19"
>
> print("Min mortality is observed in: ")  #18BCE0272
[1] "Min mortality is observed in: "
> data[which.min(data$mortality), 4]
[1] "Kerala"
> data[which.min(data$mortality), 2]
[1] "2020-01-30"
>
> 
```

Question 1(b)



```
> #18BCE0272 Nitin Pramod Ranjan
> data$recovery <- data$Cured/data$Confirmed   #Q1, (b)
> print("max recovery rate is observed in: ")
[1] "max recovery rate is observed in: "
> data[which.max(data$recovery), 4]
[1] "Kerala"
> data[which.max(data$recovery), 2]
[1] "2020-03-03"
> print("Min recovery rate is observed in: ")
[1] "Min recovery rate is observed in: "
>
> data[which.min(data$recovery), 4]   #18BCE0272, Nitin Ranjan
[1] "Kerala"
> data[which.min(data$recovery), 2]
[1] "2020-01-30"
> 
```

Answer 1)**(c) EXPORTING TO CSV AND TXT**

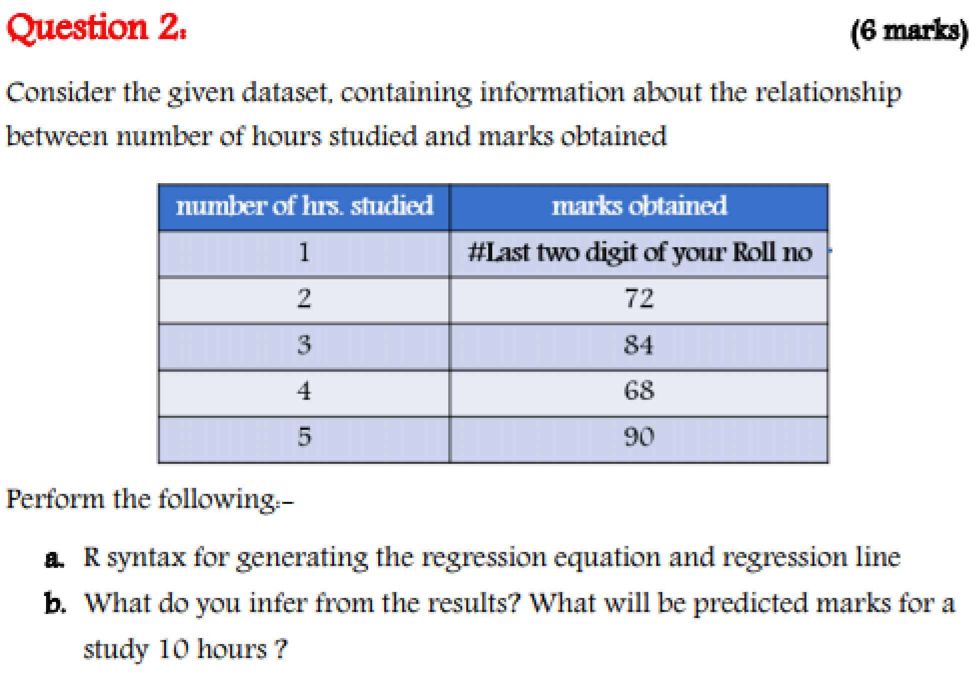


```
> result1<- data[which.max(data$recovery), 4]
> result2 <-data[which.min(data$recovery), 2]
> write.table(result1, file = “my_data.txt”, sep = “”)
Error: unexpected input in "write.table(result1, file = “"
> write.table(result1, file = “my_data.txt”)
Error: unexpected input in "write.table(result1, file = “"
> write.csv(result1, file = "my_data.csv")
> write.csv(result2, file = "my_data.csv")  #18BCE0272
> 
```

RESULT: Successfully exported to csv but export to txt failed.

18BCE0272
Nitin Pramod Ranjan

Question 2)



Question 2. (6 marks)

Consider the given dataset, containing information about the relationship between number of hours studied and marks obtained

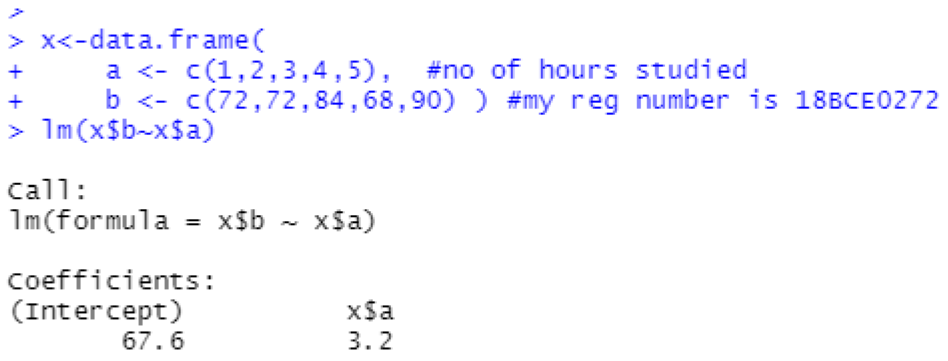
number of hrs. studied	marks obtained
1	#Last two digit of your Roll no
2	72
3	84
4	68
5	90

Perform the following:-

a. R syntax for generating the regression equation and regression line

b. What do you infer from the results? What will be predicted marks for a study 10 hours ?

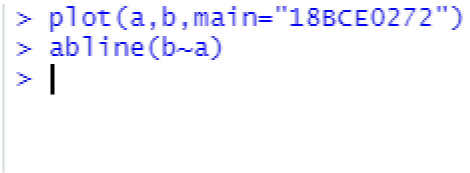
a) CODE AND OUTPUT:



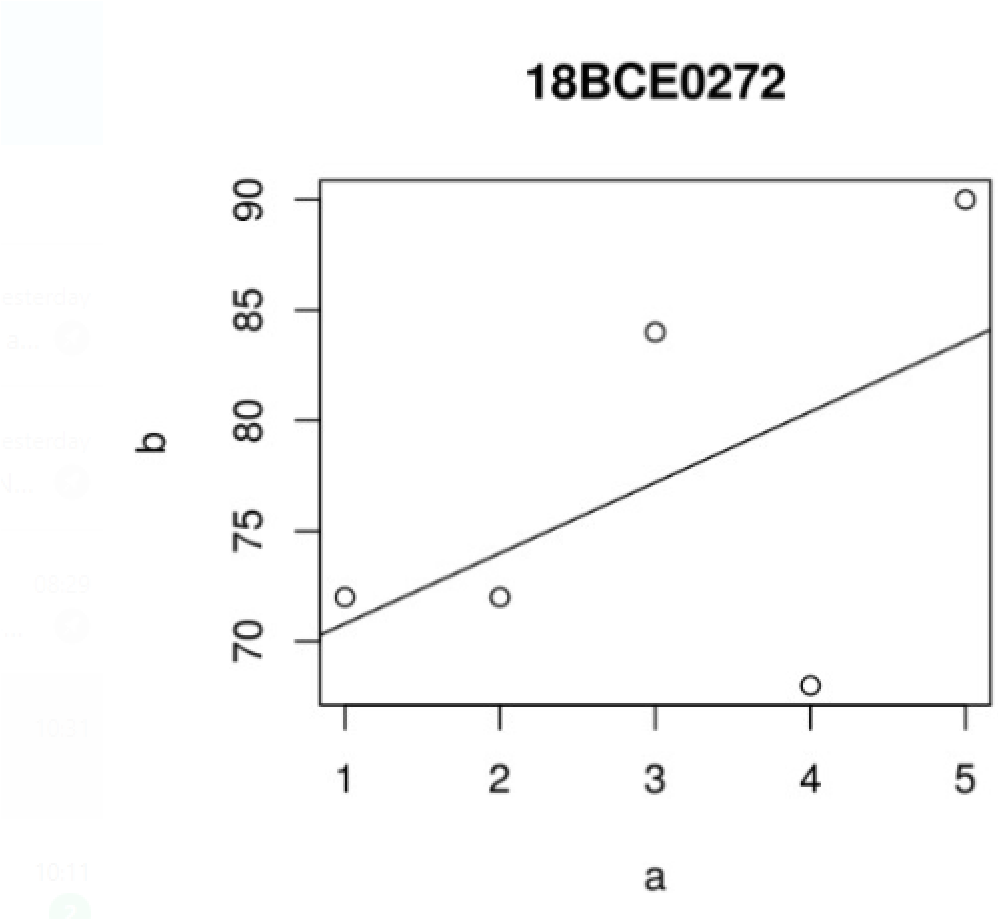
```
>
> x<-data.frame(
+     a <- c(1,2,3,4,5),  #no of hours studied
+     b <- c(72,72,84,68,90) ) #my reg number is 18BCE0272
> lm(x$b~x$a)

Call:
lm(formula = x$b ~ x$a)

Coefficients:
(Intercept)          x$a
       67.6          3.2
```



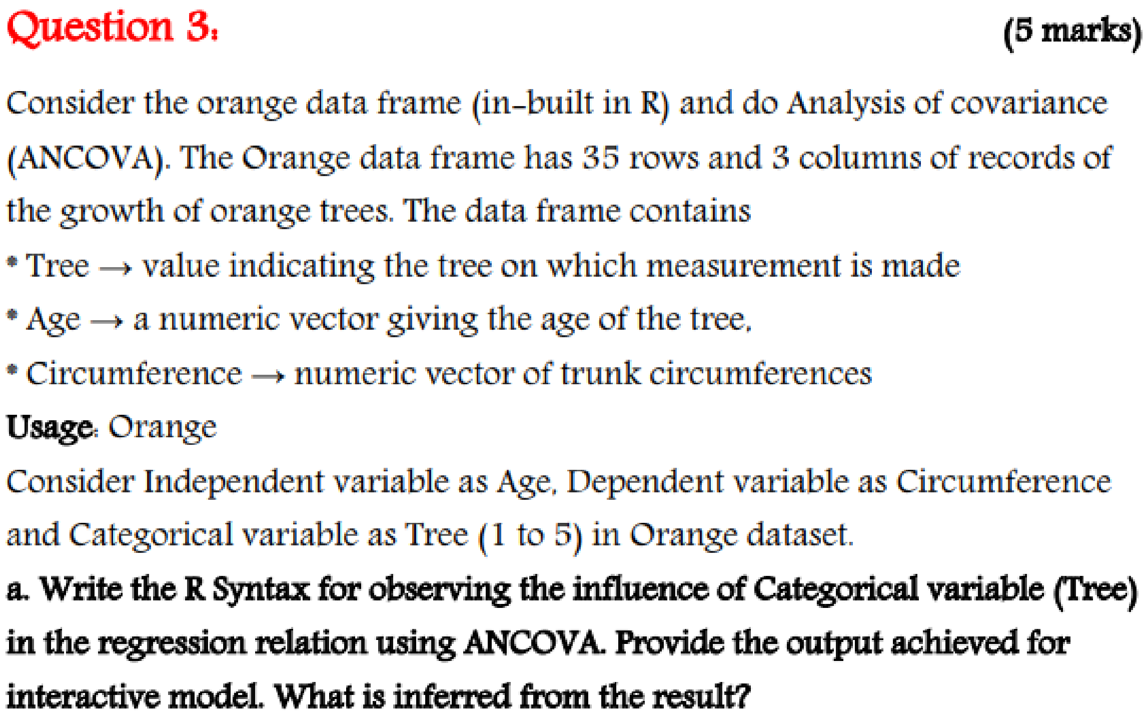
```
> plot(a,b,main="18BCE0272")
> abline(b~a)
> 
```



b) An observer can derive the inference that the two values are in a positive correlation with each other, i.e. it is highly likely that if a increases, or the number of hours increases, the marks scored will increase.
We derive the coefficient and intercept of regression line in (a) and the resultant equation is -
Marks = 3.2 * hours + 67.6
So, for 10 hours of study, marks obtained is: 99.6

18BCE0272
Nitin Pramod Ranjan

Q3)



Question 3: (5 marks)

Consider the orange data frame (in-built in R) and do Analysis of covariance (ANCOVA). The Orange data frame has 35 rows and 3 columns of records of the growth of orange trees. The data frame contains

* Tree → value indicating the tree on which measurement is made
* Age → a numeric vector giving the age of the tree,
* Circumference → numeric vector of trunk circumferences

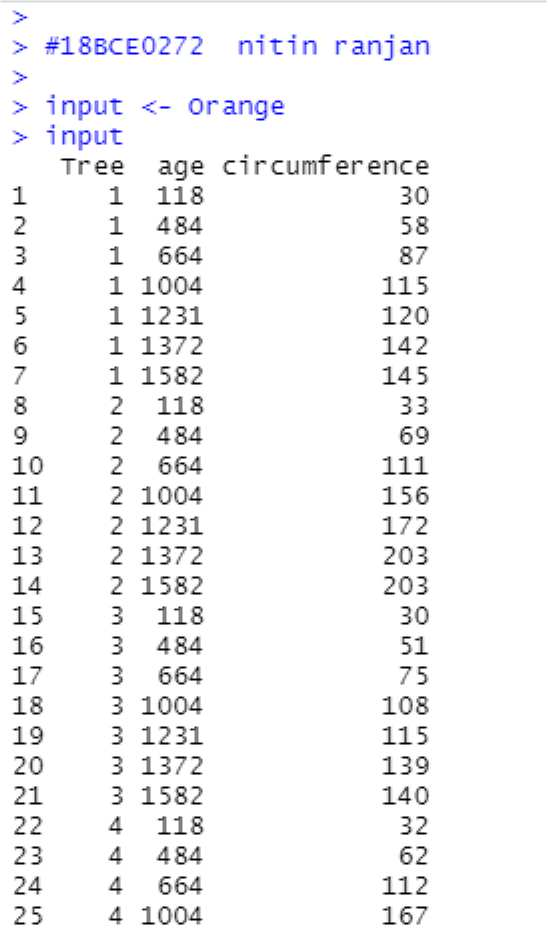
Usage: Orange

Consider Independent variable as Age, Dependent variable as Circumference and Categorical variable as Tree (1 to 5) in Orange dataset.

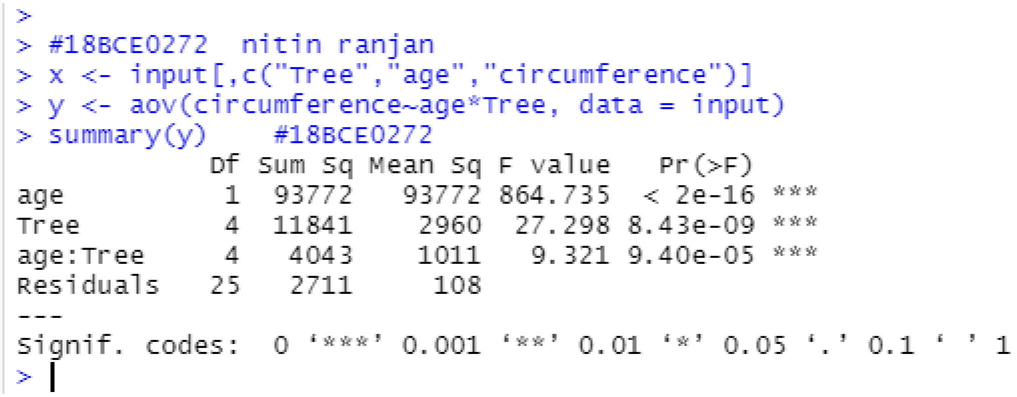
a. Write the R Syntax for observing the influence of Categorical variable (Tree) in the regression relation using ANCOVA. Provide the output achieved for interactive model. What is inferred from the result?

CODE AND OUTPUT:

18BCE0272
Nitin Pramod Ranjan



```
>
> #18BCE0272  nitin ranjan
>
> input <- Orange
> input
   Tree  age circumference
1     1  118            30
2     1  484            58
3     1  664            87
4     1 1004           115
5     1 1231           120
6     1 1372           142
7     1 1582           145
8     2  118            33
9     2  484            69
10    2  664           111
11    2 1004           156
12    2 1231           172
13    2 1372           203
14    2 1582           203
15    3  118            30
16    3  484            51
17    3  664            75
18    3 1004           108
19    3 1231           115
20    3 1372           139
21    3 1582           140
22    4  118            32
23    4  484            62
24    4  664           112
25    4 1004           167
```



```
>
> #18BCE0272  nitin ranjan
> x <- input[,c("Tree","age","circumference")]
> y <- aov(circumference~age*Tree, data = input)
> summary(y)    #18BCE0272
            Df Sum Sq Mean Sq F value   Pr(>F)    
age          1  93772   93772 864.735  < 2e-16 ***
Tree         4  11841    2960  27.298 8.43e-09 ***
age:Tree     4   4043    1011   9.321 9.40e-05 ***
Residuals   25   2711     108                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> 
```

Inference:

age(pvalue) is less than 0.05

Tree(pvalue) is less than 0.05

The summary shows that age and Tree have significant effect on the circumference of the tree.

And,

18BCE0272
Nitin Pramod Ranjan

Age:Tree(pvalue) is less than 0.05.

Ao, their interaction is also significant and hence the independnent variable and categorical variable obey some form of mathematical relationship.