LOYOLA COLLEGE (AUTONOMOUS), CHENNAI – 600 034

M.A. DEGREE EXAMINATION – **ECONOMICS**

THIRD SEMESTER – **NOVEMBER 2018**

16/17PEC3ID01 – DATA ANALYTICS FOR ECONOMISTS

Date: 03-11-2018 Dept. No. □ Max. : 100 Marks
Time: 09:00-12:00

PART- A (5 X 4 = 20 marks)

Answer any FIVE questions in 75 words each. Each question carries FOUR marks.

1. Identify the social implications of data mining.
2. What is a scatter diagram?
3. Mention the characteristic features of a predictive model.
4. Mention the purposes for which decision trees can be used in analytics.
5. State the various data types in R.
6. Write the codes to import and export a dataset from and to R.
7. Give the syntax for creating an R-Dataframe.

PART- B (4 X 10 = 40 marks)

Answer any FOUR questions in 300 words each. Each question carries TEN marks.

8. Explain the requirements of a good analytical model.
9. Examine the key steps involved in the Knowledge Discovery in Databases (KDD) process.
10. How does a logistic regression model differ from a linear regression model?
11. Illustrate the use of decision trees with the help of a suitable example.
12. Consider a text file with the name ‘pay commission’ listing out the years of the Central Pay Commission implementation from 1956 to 2016 with uniform 10-year interregnums, separated by commas, and suppose these years are to be treated as ‘string characters’.
    Write the **scan ( )** command to read the file into a data object with the name ‘cpcyears’.
    Give the command to R to display the contents of the object and write down what is displayed.
13. Two bus-operators **Delite Travels** and **Royal Roadking** operate regular bus services between two cities and the ticket rates offered by the operators vary according to the timing of services. Customers who have used both services in the past claim that, on an average, **DT** prices are significantly higher than **RR** prices. The ticket rates for 12 randomly chosen (recent) services of **DT** and 16 randomly chosen (recent) services of **RR** are available in two data objects with the names **dtprices** and **rrprices**. An investigator wishes to test the above claim of customers, assuming that

the variances in the two sets of prices are equal. It is found that the mean price for the 12 services of DT was Rs. 716.67 and the mean price for the 16 services of RR was Rs. 512.50.

Write the R-code for carrying out the relevant t-test.

Fill up the missing entries in the following output from the above code:

________ sample t-test

data: dtprices and ________

t = 2.7908, df = _____ , p-value = 0.0049

_________ hypothesis: true difference in mean is ___________ than 0

95 percent confidence interval

0.5379101 3.5454233

sample estimates:

mean of x ________

________ 512.5

What is your conclusion on the hypotheses being tested?

14. The stress-strength of three brands of electronic products (a , b, c) were measured on six items each and the following observations were made:

Brand a	**Brand b**	**Brand c**
24	14	6
10	16	16
8	22	22
6	12	38
12	18	16
6	16	18

The investigators wish to test whether the average stress-strength of the three brands are equal. Suppose that the data are stored in the data frame named **ss**.

If one wants to stack the data of all the three brands, write down the **stack ( )** command for the purpose. Name the stacked data frame as **sss** and apply the **name ( )** command to give the column names ‘strength’ and ‘brand’ to the stacked data. Write down the R-code for displaying the stacked dataframe with the above column names and write down the actual display. Finally, write down the command to carry out the ANOVA using the stacked dataframe.

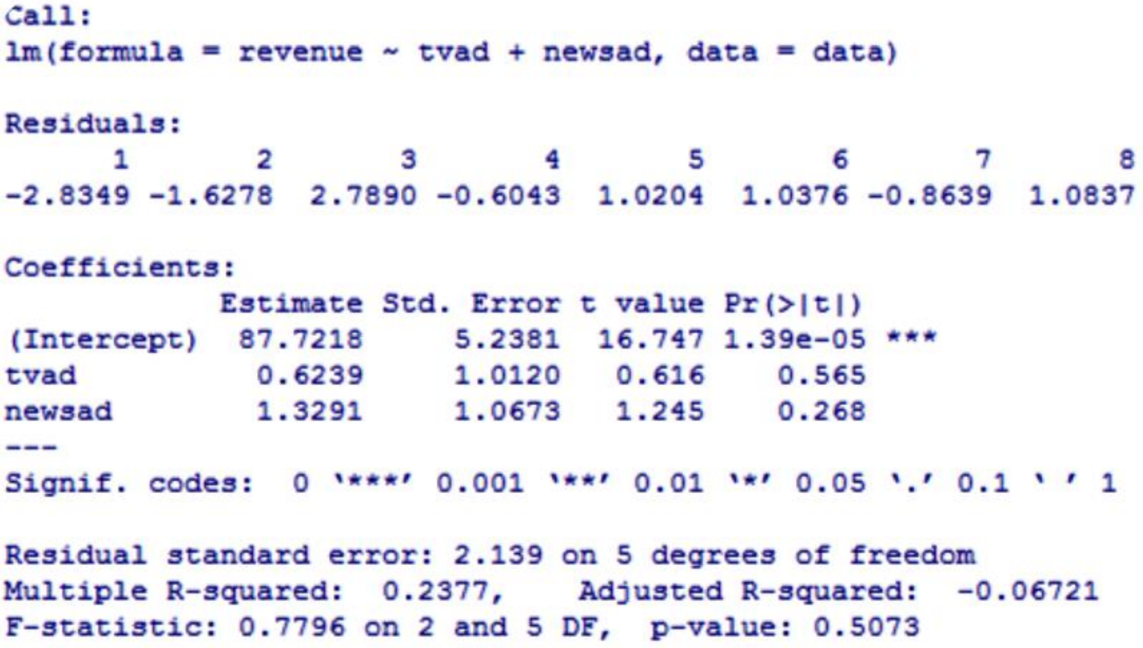
PART- C (2 X 20 = 40 marks)

Answer any TWO questions in 1200 words each. Each question carries TWENTY marks.

15. The owner of Show Time Movie Theatres Inc., would like to estimate weekly gross revenue as a function of advertising expenditures on TV and newspapers. Data for a sample of eight weeks are given below (Amount given in INR Crores):

Revenue	90	90	95	92	95	94	94	94
TV ad	5.0	2.0	4.0	2.5	3.0	3.5	2.5	3.0
News ad	1.5	2.0	1.5	2.5	3.3	2.3	4.2	2.5

#R Output



```
Call:
lm(formula = revenue ~ tvad + newsad, data = data)

Residuals:
      1       2       3       4       5       6       7       8 
-2.8349 -1.6278  2.7890 -0.6043  1.0204  1.0376 -0.8639  1.0837 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  87.7218     5.2381  16.747 1.39e-05 ***
tvad          0.6239     1.0120   0.616    0.565    
newsad        1.3291     1.0673   1.245    0.268    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.139 on 5 degrees of freedom
Multiple R-squared:  0.2377,    Adjusted R-squared:  -0.06721 
F-statistic: 0.7796 on 2 and 5 DF,  p-value: 0.5073
```

(a) Formulate the two sets of null and alternate hypotheses

(b) Write the multiple linear regression model equation.

(c) Give the interpretation of the coefficients in the R-Output and bring out the significance of the two expenditures on revenue.

(d) Identify outliers, if any, in the data.

(e) Give the clear interpretations of the different p-values in the output, the R^2 and Adj-R^2 .

(f) Give your final interpretations and recommendations to the company on their study objective.

16. Provide a diagrammatic overview of the analytics process model.

17. With the help of a suitable example illustrate the use of the Bayes Theorem.

18. Describe the basic data mining tasks and functions. Illustrate your answer with suitable examples.

*****