LOYOLA COLLEGE (AUTONOMOUS), CHENNAI - 600 034

M.A. DEGREE EXAMINATION – ECONOMICS

THIRD SEMESTER – NOVEMBER 2019

18PEC3ID01 – DATA ANALYTICS FOR ECONOMISTS

Date: 08-11-2019 Time: 09:00-12:00

PART- A

Answer any **FIVE** questions in 75 words each.

1. Differentiate KDD process and Data mining.

- 2. List two criticisms on data mining.
- 3. State two applications where Big Data is widely used.
- 4. Write short note on Noisy data.
- 5. What are variables in R language?
- 6. Name some Arithmetic operators used in R.
- 7. Give the syntax for creating an R-Dataframe.

PART- B

Answer any FOUR questions in 300 words each.

- 8. Examine the key steps involved in the Knowledge Discovery in Databases (KDD) process.
- 9. Illustrate the criteria a data analyst has to consider for building an analytical model.

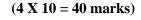
Dept. No.

- 10. Differentiate between Logistic Regression and Simple Linear Regression.
- 11. Explain the concept of Sequence Rules with an example
- 12. Explain the concept of One way and Two way ANOVA.
- 13. What is statistical Inference? Explain the concept of hypothesis testing and the steps involved in testing a statistical hypothesis.
- 14. Two parcel-packaging operators "Madras Parcel Services" and "Dunzo Packers" transport goods between two cities on a regular basis. The amounts charged by these operators vary according to the timing of services. Customers who have used both services claim that, on an average, MPS charges are significantly higher than DP charges. The charges (in '000 of Rupees) for 10 randomly chosen customers of SPS and 12 randomly chosen customers of NP are reported below:

MPS: 4.47, 3.87, 4.25, 4.92, 5.11, 5.24, 5.62, 4.80, 3.94, 5.21 **DP**: 4.13, 4.96, 3.73, 4.61, 5.28, 4.75, 5.02, 4.84, 4.10, 3.90, 3.64, 3.94

Output: Welch Two Sample t-test

data: MPS and DP
t = 1.3535, df = 18.927, p-value = 0.09592
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
-0.09297287 Inf
sample estimates:
mean of x mean of y
4.743000 4.408333



(5 X 4 = 20 marks)

Max.: 100 Marks



(i) Write the relevant null hypothesis and alternative hypothesis.

(ii) Write the R code to carry out the test

(iii) Draw your conclusion on the claim of customers and write interpretation.

PART- C

Answer any **TWO** questions in 1200 words each.

(2 X 20 = 40 marks)

- 15. Explain the concept of Simple linear regression and multiple linear regression analysis in statistical modeling.
- 16. Describe the basic data mining tasks and functions. Illustrate your answer with suitable example.
- 17. Explain the types of survival analysis in detail.
- 18. The finance team of an insurance company would like to see how the weekly premium collections from new customers are related to the number of customers contacted by the marketing team. The following data over a ten weeks period are available:

Customers contacted	120	133	127	102	99	127	112	86	110	77
Premium collected (in	6.54	8.58	6.48	5.43	5.19	6.85	5.76	4.96	5.47	5.06
lakhs)										

Build the appropriate linear regression model and interpret the results.

Output

Call: lm(formula = cust ~ premium)

Residuals:

Min 1Q Median 3Q Max -18.684 -5.929 2.539 6.443 11.424

Coefficients:

Estimate Std. Error t value Pr(>|t|) (Intercept) 24.804 19.029 1.303 0.22867 premium 14.008 3.107 4.508 0.00198 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.39 on 8 degrees of freedom Multiple R-squared: 0.7175, Adjusted R-squared: 0.6822 F-statistic: 20.32 on 1 and 8 DF, p-value: 0.001981

(a) Formulate the null and alternate hypotheses

(b) Write the code to get the simple linear regression model equation.

(c) Write the simple linear regression model equation.

(d) Give the interpretation of the coefficients in the R-Output.

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

(f) Give your final interpretations.

@@@@@@@