SRN

PES University, Bangaiore
(Established under Kamataka Act No. 16 of 2013)

UE15CS322

ESA B.TECH. 5th SEMESTER - December 2017

UE15CS322- DATA ANALYTCIS (Dr. GS)

Time: 180 minutes | Answer All Questions | Max Marks: 100

There are three pages to this question paper with FIVE questions for a total of 100 marks. Read the questions carefully and answer all questions briefly and to the point.

Question 1 [20 marks]

(a) **[6 marks** (3+3)] Suppose S_1 and S_2 comprise a list of marks in an entrance test of two groups with n_1 and n_2 candidates respectively that are combined to have a single list S of marks for n students ($n = n_1 + n_2$), then are the following statements true or false? Substantiate with reasons. (Clearly state any assumption that is made about the data.)

(i) The mean of S is always equal to the average of S_1 and S_2

(ii) The median of S is between the median of S_1 and median of S_2 or equal to one of them, regardless of the range of S_1 and S_2

(b) **[8 marks** (4+4)] Given the marks secured (out of 10) by the top candidates in an entrance test, the time taken to complete that test and their percentage grade in undergrad, answer the following questions:

	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9
Marks	4	8	7	9	7	8	7	6	5
Time (min)	24	25	30	29	28	30	27	30	27
Avg. grades (%)	80	78	89	86	86	88	84	85	79

(i) The range of marks scored in the test and average grades in undergrad are different. Suggest a transformation that would bring data in both categories to a common range.

(ii) Suggest a method to visualize the data in a single plot or chart to be able to easily select the best performing students (i.e., those who have answered the maximum questions in the least time) by looking at the visualization.

(c) **[6 marks** (3+3 marks)] The testing agency wants to link the performance in the test to a candidate's mastery over a course they offer online. How should an experiment be set up to show the course indeed helps a candidate do better on the test? The question can be answered in two parts: (i) What data (any three attributes) could be collected? (ii) From whom should the data be collected and how (online survey, etc.) for the analysis to be meaningful?

Question 2 [20 points]

(a) **[6 marks]** The table below shows the number of samples for which data is available for five attributes:

Attribute	A	B	C	D	E
No. of samples	1,00,468	2500	44,000	1765	1,14,432

We intend to build a model to predict the value of attribute C in a test set based on one or more of the other attributes that are available for the same data. There are at most 2500 data points for

(ii) Is there any change that can be made to the model at the close of Tuesday to improve the accuracy of prediction for Wednesday?

(b) [**8 marks** (4+4)] (i) Suggest another component that can be included with AR(1) to improve the accuracy of prediction. (ii) The actual value of the stock at the close of Tuesday is found to be Rs. 4200 and at the close of Wednesday is found to be Rs. 2000 because the CEO resigned. How can a model be augmented to factor in such changes?

(c) [**6 marks**] Sketch a graph of a time series data that is multiplicative with a negative trend and suggest how each of the three components – trend, seasonal and residuals - can be extracted for this data.

Question 5 [20 marks]

(a) [**12 marks** (4+4+4)] We would like to analyze reviews for Airbnb hosts (vacation rentals/ home stay options) to provide automated ratings based on the reviews. Suggest a method to (i) extract the crux (or key words) of each review and (ii) cluster all reviews to five groups based on how similar they are. (iii) Suggest a method to assign each cluster with a numeric rating (between 1-5) based on the content of the reviews in that cluster.

(b) [**8 marks** (4+4)] (i) Describe the problem you sought to solve for the class project. (ii) List any one limitation you discovered when analyzing the data/ building the model and how you worked around it OR any one (interesting) inference you made through the analysis that was counter-intuitive to what you had expected at the start and how you checked your analysis to be sure there was no error.

*****