Course Code : CST 451-3 **EZFY/RW – 22 / 1012**

Seventh Semester B. E. (Computer Science and Engineering) Examination

Elective – III

DATA VISUALIZATION AND ANALYTICS

Time : 3 Hours] [Max. Marks : 60

Instructions to Candidates :—

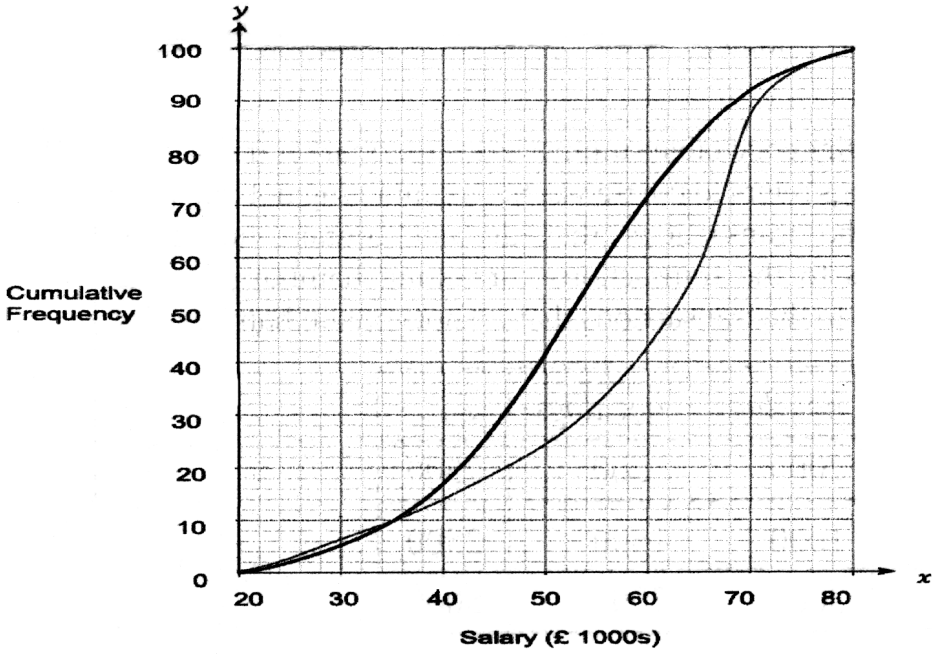
(1) Assume suitable data wherever necessary and clearly state your assumptions.
(2) Use graph paper and give examples wherever necessary.
(3) Write your paper in neat and clean handwriting.

1. (a) Describe the Data Analytics lifecycle. What kinds of tools would be used in the following phases, and for which kinds of use scenarios ?

(i) Phase 2 : Data preparation

(ii) Phase 4 : Model building 5(CO1)

(b) The cumulative frequency graph below shows the salary of 100 employees who work for Company A(black) and 100 employees who work for Company B(gray) :



(i) Draw two separate boxplots to represent the spread of salaries at each company.

(ii) From the box plots, make atleast two comparisons between the datasets. 5(CO1)

2. (a) A study conducted on pregnant women studied, the relationship between smoking and weight of the baby. The variable smoke is coded 1 if the mother is a smoker, and 0 if not. The summary table below shows the results of a linear regression model for predicting the average birth weight of babies, measured in ounces, based on the smoking status of the mother.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	123.05	0.65	189.60	0.0000
Smoke	-8.94	1.03	-8.65	0.0000

The variability within the smokers and non-smokers are about equal and the distributions are symmetric.

(a) Write the equation of the regression line.

(b) Interpret the slope in this context, and calculate the predicted birth weight of babies born to smoker and non-smoker mothers.

(c) Is there a statistically significant relationship between the average birth weight and smoking ? 6(CO2)

(b) The following contingency table summarizes supermarket transaction data, where hot dogs refers to the transactions containing hot dogs, hot dogs refers to the transactions that do not contain hot dogs, hamburgers refers to the transactions containing hamburgers, and hamburgers refers to the transactions that do not contain hamburgers.

	hot dogs	$\overline{\text{hotdogs}}$	Σ_{row}
hamburgers	2000	500	2500
$\overline{\text{hamburgers}}$	1000	1500	2500
Σ_{col}	3000	2000	5000

(i) Suppose that the association rule "hot dogs $\Rightarrow$ hamburgers" is mined. Given a minimum support threshold of 25% and a minimum confidence threshold of 50%, is this association rule strong ?

(ii) Based on the given data, is the purchase of hot dogs independent of the purchase of hamburgers ? If not, what kind of correlation relationship exists between the two ?

(iii) Compare the use of the all confidence, max confidence, Kulczynski, and cosine measures with lift and correlation on the given data. 4(CO2)

3. (a) Suppose that we want to select between two prediction models, M1 and M2. We have performed 10 rounds of 10-fold cross-validation on each model, where the same data partitioning in round i is used for both M1 and M2. The error rates obtained for M1 are 30.5, 32.2, 20.7, 20.6, 31.0, 41.0, 27.7, 26.0, 21.5, 26.0. The error rates for M2 are 22.4, 14.5, 22.4, 19.6, 20.7, 20.4, 22.1, 19.4, 16.2, 35.0. Comment on whether one model is significantly better than the other considering a significance level of 1%. 5(CO2)

(b) The data tuples of following table are sorted by decreasing probability value, as returned by a classifier. For each tuple, compute the value for the number of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). Compare the true positive rate (TPR) and false positive rate (FPR. Plot the ROC curve for the data.

Tuple #	Class	Probability
1	P	0.95
2	N	0.85
3	P	0.78
4	P	0.66
5	N	0.60
6	P	0.55
7	N	0.53
8	N	0.52
9	N	0.51
10	P	0.40

5(CO2)

4. (a) Consider an auto regressive AR (2) model

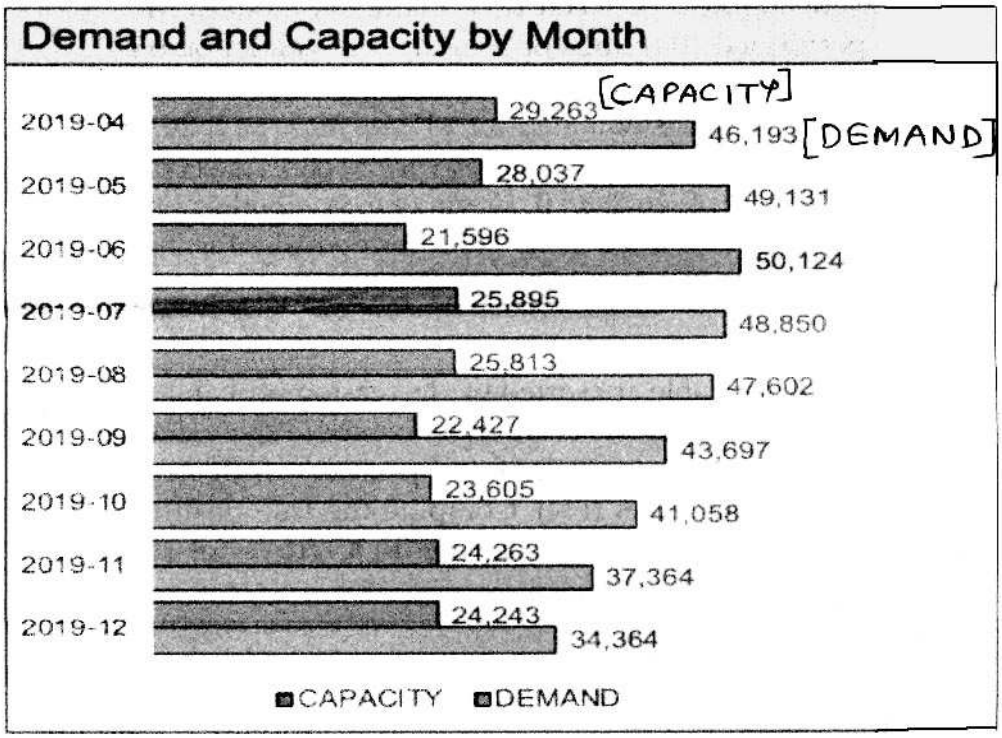
$X_t = 0.80X_{t-1} - 0.60X_{t-2} + a_t$

(i) Verify whether the series is stationary.

(ii) Obtain ρk for $k = 1, 2, ..., 5,$

(iii) Plot the correlogram. 10(CO3)

5. (a) The following graph shows capacity and demand measured in number of project hours over time. It is currently graphed as a horizontal bar chart.



Draw at least 3 different ways to potentially visualize this data. 5(CO4)

(b) Give an example from real life where each of the following Gestalt's principles are used :

(i) Connectedness

(ii) Common region

(iii) Figure and ground

(iv) Symmetry

(v) Focal point

(vi) Similarity

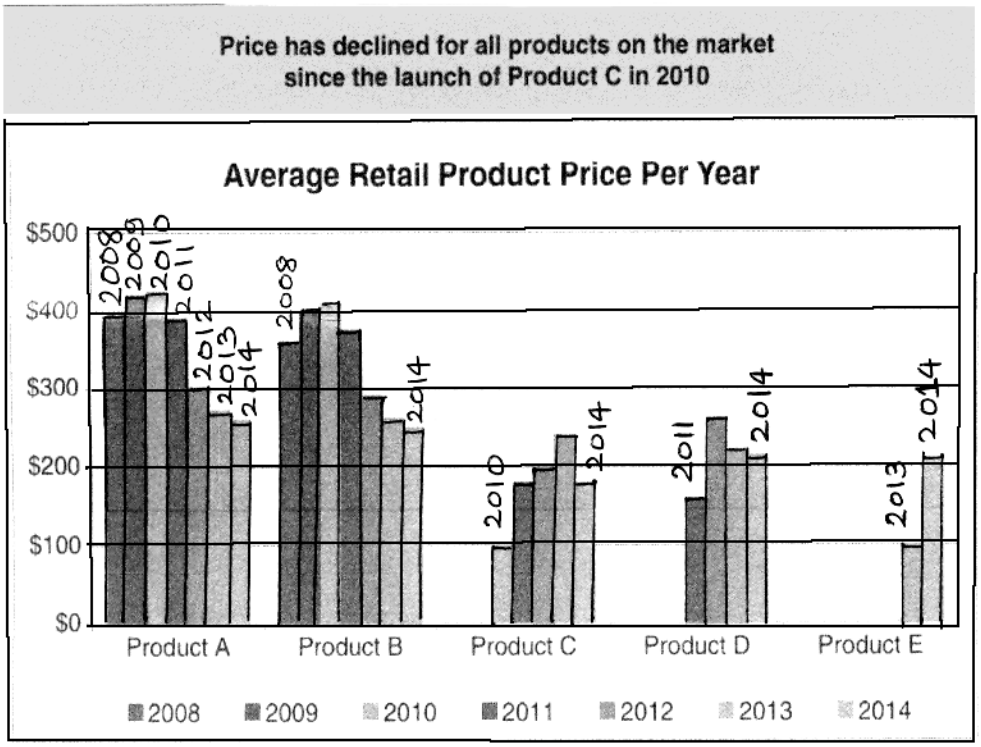
(vii) Closure

(viii) Continuity

(ix) Proximity

(x) Common fate. 5(CO4)

6. (a) Consider the following visual and craft your own data story by answering the following questions :



Target audience

Compelling narrative

Appropriate display/visualizations. 6(CO4)

(b) What is conflict in data story telling ? Expound with the help of examples. 4(CO4)