Course Code : CST 451-3 **GVHW/RW – 23 / 1012**

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

DATA VISUALIZATION AND ANALYTICS

Time : 3 Hours] [Max. Marks : 60

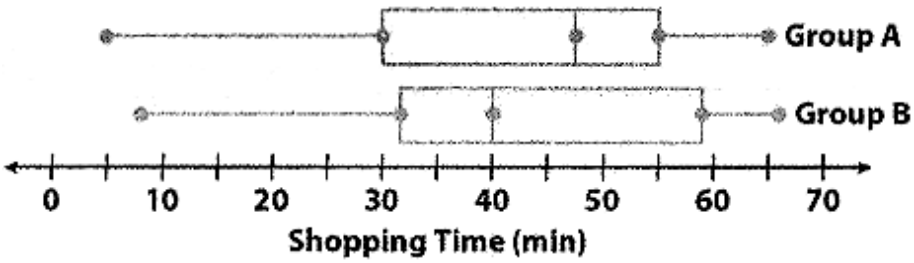
Instructions to Candidates :—

(1) All questions carry marks as indicated against them.

(2) Assume suitable data wherever necessary and clearly state your assumptions.

1. (a) Briefly describe the different phases of Data Analytics life cycle. 3(CO1)

(b) The box plots show the distribution of times spent shopping by two different groups.



(i) Compare the shapes of the box plots. Write your observations.

(ii) Compare the centers of the box plots. Write your observations.

(iii) Compare the spreads of the box plots. Write your observations.

(iv) Which group has the greater variability in the bottom 50% of shopping times, The top 50% of shopping times ? Explain your answer. 4(CO1)

(c) Construct MaxDifference histogram for the given data points using $\beta = 3$. Data points : 6, 6, 6, 6, 6, 9, 9, 9, 14, 14, 17, 17, 17, 17, 17, 17, 19, 19, 19, 19, 31, 31, 31, 31, 31, 31, 31, 54, 54, 66, 66, 81, 97, 97, 97, 164, 164, 164, 164, 164, 189, 189. 3(CO1)

2. (a) Data on Scholastic Aptitude Test (SAT) scores are published by the College Entrance Examination. SAT scores for randomly selected students from each of four high-school rank categories are displayed in the following table :

Top Tenth	Second Tenth	Second Fifth	Third Fifth
528	514	649	372
586	457	506	440
680	521	556	495
718	370	413	321
	532	470	424
			330

(Note : Mean for Top Tenth = 628·0, Mean for Second Tenth = 478·8, Mean for Second Fifth = 518·8, Mean for Third Fifth = 397·0 and mean for the entire data is 494·1)

Construct the one-way ANOVA table for the data. Conclude if the means of all the categories are similar or not. Given the critical value is 3·41. 7(CO2)

(b) Consider the given data of employee working hours and if they get increment or not.

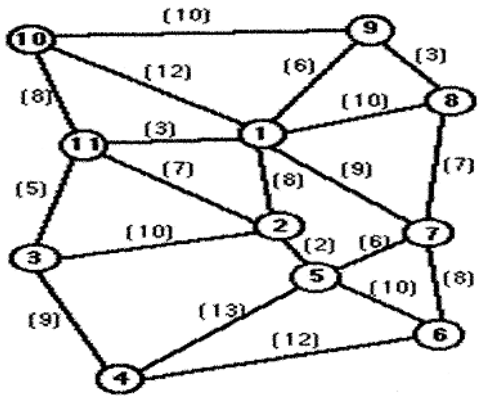
Hours	**Increment**
29	0
15	0
33	1
28	1
39	1

Apply logistic regression to the given data using the model suggested by optimizer for odds of increment as
log(odds) = −64 + 2*Hours.
Compute :

(i) The probability of getting increment if working hours is 33 hours.

(ii) At least how much time should the employee work to get an increment with probability of 90% ? 3(CO2)

3. (a) Consider the given webgraph and show which webpages are similar using Graph based clustering. Assume "1" as the starting node.



5(CO2)

(b) Consider the training dataset shown below :

A	B	Class Label
0	1	c1
0	0	c2
1	1	c1
0	1	c1
1	0	c1
0	0	c2
1	1	c1
0	0	c2
1	0	c1
1	0	c2

(i) Compute the conditional probabilities :

P(A = 1 | C = c1),

P(A = 0 | C = c1),

P(B = 1 | C = c1),

P(B = 0 | C = c1),

P(A = 1 | C = c2),

P(A = 0 | C = c2),

P(B = 1 | C = c2), and

P(B = 0 | C = c2).

(ii) Use the computed conditional probabilities to predict the class label for a test sample (A = 1, B = 0) using the Naive Bayes approach. 5(CO2)

4. (a) Find the pacf of the AR(2) process :

$X_t = 0{\cdot}333X_{t-1} + 0{\cdot}222X_{t-2} + a_t$ 5(CO3)

OR

(b) Check whether the given time series is stationary or not :

$X_t = 0{\cdot}80X_{t-1} - 0{\cdot}60X_{t-2} + \varepsilon_t$ 5(CO3)

(c) What is trend ? Name the various methods of finding trend in time series. Construct 3 yearly moving average from the following data and show on a graph against the original data :

Year	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Sales (in lakhs)	18	22	20	26	30	22	24	28	32	35

Also, check if there is seasonal variation in the sales. 5(CO3)

5. (a) A work schedule for a development project received by an IT company is as given. Elaborate on the use of Monte Carlo Simulation in this situation. Also, compute the duration of each activity using the PERT formula.

Activity	**Optimistic**	**Pessimistic**	**Most Likely**
Requirement analysis	1	3	4
Design	2	4	6
Web site development	7	11	9
Testing	3	6	6

What is the estimated total time for completion of the project ? Also, find the best case and worst case completion time. 4(CO2)

(b) List the pros and cons of common approaches to visualizing proportions : pie charts, stacked bars and side-by-side bars.
Consider the given data and specify which of the above approach should be used for visualization. Show a rough visualization as well.

Company	Percentage of People
Super hero	50%
Animated	25%
Comedy	20%
Romance	5%

3(CO4)

(c) Time series data of Monthly submissions to three preprint servers covering biomedical research : bioRxiv, the q-bio section of arXiv, and PeerJ Preprints is presented as a visualization in two Figures below :

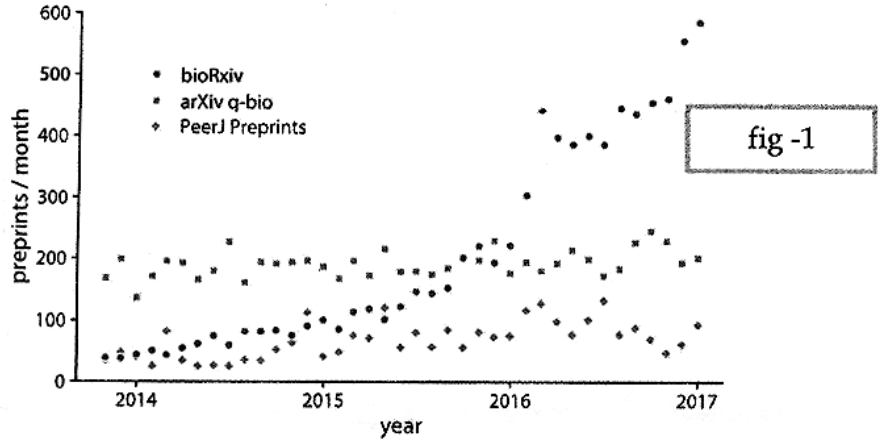


fig -1

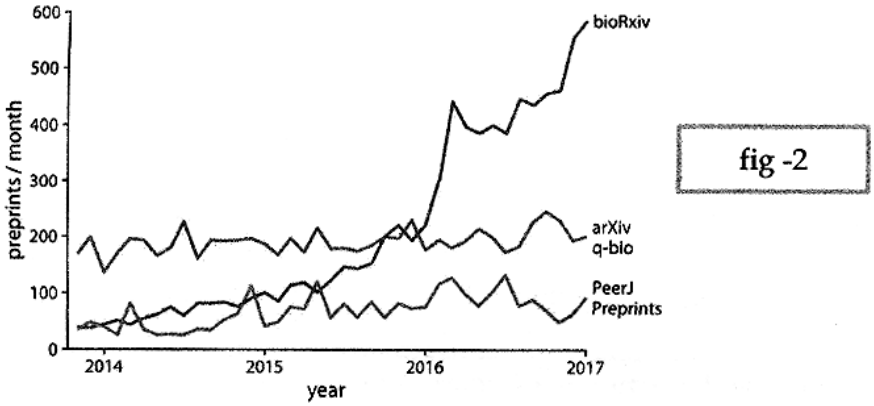


fig -2

Which Figure better represents the data ? Also, present your analysis about the data based on the Figure selected. 3(CO4)

6. (a) Table below shows a confusion matrix and a cost matrix for a two-class problem. Calculate the following measures :

	Predicted +	Predicted –
True +	100	40
True –	60	300

(a) Confusion Matrix

	Predicted +	Predicted –
True +	–1	100
True –	20	0

(a) Cost Matrix

(i) Accuracy,

(ii) Misclassification cost,

(iii) Precision,

(iv) Recall,

(v) F-measure. 5(CO1,2)

(b) Which are the three main components in data storytelling ? Explain Reader-driven and Author-driven Narratives. Which according to you are better for data story telling ? 5(CO4)