



**SIDDHARTHA INSTITUTE OF SCIENCE AND TECHNOLOGY:: PUTTUR
(AUTONOMOUS)**

Siddharth Nagar, Narayanavanam Road – 517583

QUESTION BANK (DESCRIPTIVE)

Subject with Code: DATA SCIENCE (19CS0542)

Course & Branch: B.Tech - CSE

Year & Sem: IV-B.Tech & I-Sem

Regulation: R19

**UNIT –I
INTRODUCTION TO R & DATA ANALYSIS**

1	a	What is exploratory data analysis ?List the five main skills and behavioral characteristics of Data scientists.	[L1][CO1]	[6M]
	b	Define the window panes of RStudio GUI .Give the difference between BI and Data science	[L4][CO1]	[6M]
2	a	Sketch Date Analytics Lifecycle diagram.	[L3][CO1]	[4M]
	b	Explain in detail each stages of Data Analytics Lifecycle	[L2][CO1]	[8M]
3		Illustrate in detail the data types of R.	[L3][CO1]	[12M]
4	a	Expand and Distinguish between NOIR attributes	[L4][CO1]	[6M]
	b	Describe the Array function in R	[L2][CO1]	[6M]
5		Examine the functions used for Visualizing a Single Variable	[L3][CO5]	[12M]
6		Discriminate about the functions used for examining Multiple Variables	[L5][CO1]	[12M]
7		Give example and define the following function of R a) read() b) head() c) summary() d) plot() e) summary()	[L2][CO1]	[12M]
8	a	Discuss the Data frames utilization in R	[L2][CO1]	[6M]
	b	Describe the List function in R	[L2][CO1]	[6M]
9	a	Define dirty data	[L1][CO5]	[4M]
	b	How dirty data can be detected in the data exploration phase with Visualizations	[L2][CO5]	[8M]
10	a	Illustrate the importance of visualizing data before analysis	[L2][CO5]	[6M]
	b	Justify “Using visualization for data exploration is different from presenting results to stakeholders”	[L6][CO5]	[6M]

UNIT –II
STATISTICAL METHODS FOR EVALUATION & ASSOCIATION RULES

1	a	Categorize the Hypothesis Testing.	[L1][CO3]	[6M]
	b	What is Power of test? Explain the Sample Size and effect size are related.	[L2][CO3]	[6M]
2	a	Differentiate Null Hypotheses and Alternative Hypotheses	[L4][CO3]	[6M]
	b	Examine the application property of Wilcoxon rank-sum test	[L3][CO3]	[6M]
3		Discriminate about Difference of Means	[L5][CO2]	[12M]
4		Suppose everyone who visits a retail website gets one promotional offer or no promotion at all. We want to see if making a promotional offer makes a difference. What statistical method would you recommend for this analysis?	[L5][CO2]	[12M]
5		Explain the following a) Student's t-test b) Welch's t-test	[L2][CO2]	[12M]
6	a	Define and Detail ANOVA.	[L1][CO3]	[6M]
	b	How evaluation of Candidate Rules are done?	[L2][CO3]	[6M]
7	a	What is a type I error? What is a type II error? Is one always more serious than the other? Why?	[L1][CO3]	[6M]
	b	Give the difference between Validation and Testing	[L4][CO5]	[6M]
8	a	State Apriori Algorithm	[L1][CO3]	[4M]
	b	Explain Apriori Algorithm with example	[L2][CO3]	[8M]
9	a	List and Discuss the four measures of significance of Association rules	[L1][CO3]	[6M]
	b	Give the Applications of Association Rules	[L1][CO3]	[6M]
10		Illustrate any five approaches to improve Apriori's efficiency when the dataset is large.	[L3][CO3]	[12M]

UNIT –III
REGRESSION & CLASSIFICATION

1	a	Discuss confusion matrix with suitable example.	[L1][CO4]	[6M]																							
	b	What is accuracy? Which two basic measures does the entropy methods select the most informative attribute?	[L1][CO4]	[6M]																							
2		Explain the analytical technique Linear Regression with its model description.	[L2][CO4]	[12M]																							
3		Discuss the following with respect to linear regression a) Categorical Variables b) Confidence Intervals on the Parameters c) Confidence Interval on the Expected Outcome d) Prediction Interval on a Particular Outcome	[L2][CO4]	[12M]																							
4	a	Justify the usage of linear regression and logistic regression.	[L6][CO4]	[4M]																							
	b	Illustrate Logistic Regression Model.	[L3][CO4]	[8M]																							
5	a	Describe Decision Trees in detail with example.	[L2][CO4]	[6M]																							
	b	Explain the algorithm of decision tree	[L3][CO4]	[6M]																							
6		Intercept the decision trees algorithms	[L4][CO4]	[12M]																							
7	a	State Bayes' Theorem	[L1][CO4]	[4M]																							
	b	Discuss Naïve Bayes classification method considering an example	[L2][CO4]	[8M]																							
8		How does one pick the most suitable method for a given classification problem?	[L2][CO4]	[12M]																							
9	a	Compare the C4.5 and CART algorithm of decision tree.	[L4][CO4]	[4M]																							
	b	Discriminate the way how the evaluation of decision tree is done	[L5][CO4]	[4M]																							
	c	Give the two approaches that help avoid over fitting in decision tree learning.	[L2][CO4]	[4M]																							
10		Consider the given confusion Matrix of Naïve Bayes from the Bank Marketing Dataset	[L4][CO4]	[12M]																							
		<table border="1" style="margin-left: 20px;"> <thead> <tr> <th colspan="2"></th> <th colspan="3">Predicted Class</th> </tr> <tr> <th colspan="2"></th> <th>Subscribed</th> <th>Not Subscribed</th> <th>Total</th> </tr> </thead> <tbody> <tr> <th rowspan="2">Actual Class</th> <th>Subscribed</th> <td>3</td> <td>8</td> <td>11</td> </tr> <tr> <th>Not Subscribed</th> <td>2</td> <td>87</td> <td>89</td> </tr> <tr> <th colspan="2">Total</th> <td>5</td> <td>95</td> <td>100</td> </tr> </tbody> </table> <p>Calculate the following defining them</p> <p>a) Accuracy b) TPR c) FPR d) FNR e) Precision</p>			Predicted Class					Subscribed	Not Subscribed	Total	Actual Class	Subscribed	3	8	11	Not Subscribed	2	87	89	Total		5	95	100	
		Predicted Class																									
		Subscribed	Not Subscribed	Total																							
Actual Class	Subscribed	3	8	11																							
	Not Subscribed	2	87	89																							
Total		5	95	100																							

UNIT –IV
CLUSTERING & TIME SERIES ANALYSIS

1	a	What is clustering? State the advantage of using PAM.	[L1][CO4]	[8M]
	b	List the various parts of ARIMA model	[L1][CO4]	[4M]
2		Illustrate the method to find k clusters from a collection of M objects with n attributes.	[L3][CO6]	[12M]
3	a	Explain any one case study for time series analysis	[L2][CO5]	[6M]
	b	What is forecasting in association with time series. Explain	[L1][CO6]	[6M]
4	a	Indicate when the time series y_t for $t=1,2,3,\dots$ is said to be stationary time series.	[L2][CO6]	[4M]
	b	Express the stationary time series conditions in detail.	[L6][CO6]	[8M]
5		Discuss in detail each part of the ARIMA model	[L2][CO4]	[12M]
6	a	List and explain time series components	[L1][CO6]	[6M]
	b	Discriminate the steps involved in Box-Jenkins Methodology	[L5][CO6]	[6M]
7	a	What is meant by k-means	[L1][CO4]	[4M]
	b	Describe k-means algorithm to find k clusters	[L2][CO4]	[8M]
8		Correlate ARMA and ARIMA Models	[L4][CO6]	[12M]
9		Express the following	[L2][CO6]	[12M]
		a) Autocorrelation Function b) Autoregressive Models		
10		List and describe Additional time series methods	[L2][CO6]	[12M]

UNIT –V
TEXT ANALYSIS

1	a	What is text analytics? Explain meant by word cloud?	[L1][CO6]	[6M]
	b	Define Confusion matrix.And Illustrate the Porter's stemming algorithm.	[L3][CO6]	[6M]
2		Explain the three important steps of the text analysis	[L2][CO6]	[12M]
3	a	Sketch the flow diagram of Text analysis process	[L5][CO6]	[4M]
	b	Illustrate in detail the steps involved in the process of Text Analysis done by organizations	[L3][CO6]	[8M]
4	a	Define TFIDF.	[L1][CO6]	[4M]
	b	Describe the usage of TFIDF to compute the usefulness of each word in the texts.	[L2][CO6]	[8M]
5		Explain how the data science team will categorize the reviews by topics	[L2][CO6]	[12M]
6		Illustrate the main challenges of text analysis	[L3][CO6]	[12M]
7	a	Define Topic model. Describe LDA.	[L2][CO6]	[6M]
	b	Justify the process of topic modeling simplification.	[L6][CO6]	[6M]
8		Explain the following a) Tokenization b) Case folding	[L3][CO6]	[12M]
	a	Explain how categorizing documents by topics is done.	[L2][CO6]	[6M]
	b	Interpret the procedure used in data science to gain insights into customer opinions	[L3][CO6]	[6M]
10	a	What is meant by sentiment analysis	[L1][CO6]	[4M]
	b	Discriminate the methods used for sentiment analysis	[L5][CO6]	[8M]

Prepared by:
Mr. G Ravi Kumar
Assistant Professor