INSTITUTE OF AERONAUTICAL ENGINEERING

(Autonomous)

Dundigal, Hyderabad - 500 043

COMPUTER SCIENCE AND ENGINEERING

TUTORIAL QUESTION BANK

Course Title	DATA SCIENCE	
Course Code	BCSB06	
Programme	M.Tech	
Semester	I	CSE
Course Type	Elective	
Regulation	IARE-R18	
Course Structure	**Theory**			**Practical**	
	Lectures	Tutorials	Credits	Laboratory	Credits
	3	-	3	-	-
Course Faculty	Dr. Chukka Santhaiah, Associate Professor				

COURSE OBJECTIVES:

	The course should enable the students to:
I	Summarize the fundamental knowledge on basics of data science and R programming.
II	Develop programs in R language for understanding and visualization of data using statistical functions and plots.
III	Learn to apply hypotheses and data into actionable Predictions.
IV	Understand a range of machine learning algorithms along with their strengths and weaknesses.
V	Able to document and transfer the results and effectively communicate the findings using visualization techniques.

COURSE OUTCOMES (COs):

CO 1	Understand the process and different stages of data science and relevant data descriptions in R language.
CO 2	Illustrate various SQL, NOSQL databases connecting with R and perform correlation and regression analysis.
CO 3	Evaluate different data models and perform clustering analysis.
CO 4	Solve various real time problems using artificial neural networks techniques and comparing different learning algorithms.
CO 5	Explore on various ways to deliver results through documentation and plots of multivariate data and matrix data.

COURSE LEARNING OUTCOMES (CLOs):

BCSB06.01	Understand and develop relevant programming abilities.
BCSB06.02	Understand and intuition of the whole process line of extracting knowledge from data.
BCSB06.03	Equip with the fundamental knowledge on basics of data science and R programming
BCSB06.04	Critically analyze and evaluate variety of NoSQL databases.
BCSB06.05	Develop the ability to build and assess Data-based models.
BCSB06.06	Analyze data analysis and make models using regression analysis.
BCSB06.07	Familiarize with variety of machine learning tasks: clustering, dimensionality reduction, regression and classification.
BCSB06.08	Understand how to formalize practical problems using methods of machine learning.
BCSB06.09	Understand neural networks techniques solve real time problems.
BCSB06.10	Understand the different learning algorithms.
BCSB06.11	Chose an appropriate learning Algorisms to solve particular problems.
BCSB06.12	Based on delivering results make documentation for various results sets.
BCSB06.13	Understand how to plot graphs for multivariate and matrix data.

TUTORIAL QUESTION BANK

UNIT – I INTRODUCTION				
PART – A (SHORT ANSWER QUESTIONS)				
S. No	**QUESTIONS**	**Blooms Taxonomy Level**	**Course Outcomes**	**Course Learning Outcomes (CLOs)**
1	List the major steps involved in data science project.	Remember	CO 1	BCSB06.03
2	Identify the steps in creating and running a R script?	Remember	CO 1	BCSB06.02
3	Write the steps to create a calculator application with R objects?	Understand	CO 1	BCSB06.02
4	Write the basic set of steps followed in data management?	Understand	CO 1	BCSB06.02
5	State the any few data sources in data collection step.	Remember	CO 1	BCSB06.02
6	Draw the block diagram of data science project process.	Understand	CO 1	BCSB06.02
7	State the features of R Language.		CO 1	BCSB06.02
8	Write the various classes of data types ANSI C supports?	Remember	CO 1	BCSB06.04
9	State which of the following are valid identifiers. If not invalid, state the reason. a. Q b. C c. c d. t e. pi	Understand	CO 1	BCSB06.02
10	Write expression in R a)e4+log2 b)24X53 c)log2 10 c)log10 2	Remember	CO 1	BCSB06.01
11	What are the different forms of data types and how to test the data type in 'R'? Give one example for each	Understand	CO 1	BCSB06.03
12	Differentiate 'R' while comparing with other programming languages?	Remember	CO 1	BCSB06.02
13	How to set up the 'R' environment?	Understand	CO 1	BCSB06.02
14	Explain R as calculator using basic operations and inbuilt functions with suitable example?	Remember	CO 1	BCSB06.02
15	What are different basic components in 'R'?	Remember	CO 1	BCSB06.02

	PART – B (LONG ANSWER QUESTIONS)			
1	Elaborate the following R objects. a)vector b)data frame c)matrix d)list	Understand	CO 1	BCSB06.03
2	Compute the given mathematical formula and display on console in R. X=	Remember	CO 1	BCSB06.02
3	Describe the multiple ways to read and write data from disc and web. Elaborate R functions to choose character, numerical input dynamically from user.	Understand	CO 1	BCSB06.02
4	List the inbuilt summary functions to apply on vectors. Create a vector and apply all functions on it.	Remember	CO 1	BCSB06.01
5	Identify the different ways to access the R objects. List the different data types in R with suitable example.	Remember	CO 1	BCSB06.03
6	The price of one kg of rice is Rs. 40.75 and one kg of sugar is Rs. 30. **Write** R program to get the total amount of 2kg rice and 5kg sugar purchase.	Understand	CO 1	BCSB06.02
7	Elaborate the process of working with different data files. Write about filechoose() function.	Remember	CO 1	BCSB06.04
8	List the inbuilt summary functions to apply on vectors. Create a vector and apply all functions on it.	Understand	CO 1	BCSB06.02
9	How to get system date in R? Generate sequence of previous and coming 10 dates from today in R.	Understand	CO 1	BCSB06.01
10	What are the different ways to read the dataset? How to create and rename a variable in R? What are the read write methods available in R and explain?	Understand	CO 1	BCSB06.03
	PART – C (PROBLEM SOLVING AND CRITICAL THINKING QUESTIONS)			
	UNIT – I INTRODUCTION			
1.	**Create** an Array with name "MySales" with 30 observations using following methods: a) By using the array with dimensions 3, 5 and 2. b) By using Vector method.	Remember	CO 1	BCSB06.01

2	**That** is data frame and how to create a data frame using the following data: Height GPA
	    66 3.80
	    62 3.78
	    63 3.88
	    70 3.72
	    74 3.69
		Analyze	CO 1	BCSB06.03

3	**How** will you identify and treat the missing value and outlier data in R?	Remember	CO 1	BCSB06.03
4	**Compare** the different forms of data types and create a list object with restaurant menu items. Give inbuilt functions on list object with example Write R script to create and display list object of stores items having: {Fruits:{orange,mango,apple,watermelon,banana}Juices:{appy,fruty, slice} Milkshakes:{Mango, papaya, sapota, pineapple}}.	Remember	CO 1	BCSB06.03
5	Let two vectors x <- c(1,3, 5) and y <- c(3, 2, 10), **what** is size of the expression rbind(x, y) output? Differences between rbind and cbind. Create a matrix using rbind and cbind.	Remember	CO 1	BCSB06.01

6	**Create** a data frame with a = c(1, 2, 3), b = c(4, 5, 6), c(7, 8, 9) and find the value of the following 1a.) How do I select the c(4, 5, 6)? 1b.) How do iselect the 1? 1c.) How do I select the 5? 1d.) What is df[, 3]? 1e.) What is df[1,]? 1f.) What is df[2, 2]?	Understand	CO 1	BCSB06.03
7	**Find** the output in R: i) If x is defined as x <- list(2, "a", "b", TRUE). What does x[[1]] ? ii) If x is defined as x <- list(2, "a", "b", TRUE). What does x[[2]] give me? iii) if x ,y are two vectors x <- 1:4 and y <- 2:3. What is produced by the expression x + y? iv) if x ,y are two vectors, x <- 1:4 and a vector y <- 2. What is produced by the expression x + y? v) Find x^3	Understand	CO 1	BCSB06.02
8	**Analyze** the control structures with conditional statements in R with suitable example. Create a user defined function fact(j) to return the the factorial of j using functions in R.	Understand	CO 1	BCSB06.01

UNIT – II

SQL, NOSQL AND DATA ANALYSIS

PART – A (SHORT ANSWER QUESTIONS)

1	State the packages and function required to run SQL queries in R?	Understand	CO 2	BCSB06.04
2	List the packages and function required to read Excel sheet data to R object?	Remember	CO 2	BCSB06.05
3	Explain the drawbacks in different ways to access Excel sheet in R.	Understand	CO 2	BCSB06.06
4	What are other names of NoSQL?	Understand	CO 2	BCSB06.07
5	List out the features of NoSQL.	Remember	CO 2	BCSB06.04
6	Give any five examples of NoSQL databases.	Understand	CO 2	BCSB06.05
7	Differentiate between SQL and No SQL databases.	Understand	CO 2	BCSB06.06
8	Compare the SQL and NoSQL in terms of Data storage model and Schema.	Remember	CO 2	BCSB06.07
9	What are the assumptions of regression Modeling?	Remember	CO 2	BCSB06.04
10	Write short note on ANOVA?	Understand	CO 2	BCSB06.05

PART – B (LONG ANSWER QUESTIONS)

1	Write the R script which include relevant packages and procedure to access .csv and .exl files. Elaborate with suitable example.	Understand	CO 2	BCSB06.04
2	What is the difference between SQL database and NoSQL Database?	Remember	CO 2	BCSB06.05
3	Explain different approaches in R to connect with Excel	Understand	CO 2	BCSB06.06
4	Define basic classification based on data model, with examples.	Remember	CO 2	BCSB06.07
5	Explain Covariance and Correlation with example. Infer the relation among attributes with respect to correlation coefficient.	Remember	CO 2	BCSB06.04
6	Calculate the correlation coefficient of iris dataset. What preprocess steps is needed for iris data set.	Analyze	CO 2	BCSB06.05
7	Write a R program to find the correlation coefficient of iris data set and find the exact relation by using linear regression model.	Analyze	CO 2	BCSB06.06
8	Distinguish simple and multiple regression analysis and its applications working with numerical and categorical data?	Remember	CO 2	BCSB06.07
9	Differentiate SQL and NoSQL databases in detail?	Understand	CO 2	BCSB06.04
10	What is forecasting give examples?	Remember	CO 2	BCSB06.05

		PART – C (PROBLEM SOLVING AND CRITICAL THINKING QUESTIONS)			
1	Write R program to extract sample XML data from web and steps to convert as dataframe. Specify the needed packages and functions.	Apply	CO 2	BCSB06.04
2	Distinguish simple and multiple regression analysis and its applications working with numerical and categorical data?	Remember	CO 2	BCSB06.05
3	What are residuals? Define briefly about Regression analysis.	Understand	CO 2	BCSB06.06
4	Generate prediction model using linear regression for finding relative relation among variables. Write a R script to get a linear equation y=mx+c form for the heart weight and body weight in cats dataset.	Analyze	CO 2	BCSB06.04
5	Compute the covariance matrix and correlation matrix for the four numerical attributes. Interpret the statistical findings to know more about hidden nature in data.	Remember	CO 2	BCSB06.05
6	Write an R script to connect with Excel, read the contents of sheet and load into R object.	Remember	CO 2	BCSB06.06
7	Write a R program to perform the following: i. Import a data from web storage. ii. Name the dataset with suitable identifier iii. Perform Logistic Regression to find out relation between variables that are affecting the admission of a student in a institute based on his or her GRE score, GPA obtained and rank of the student. iv. Check the model is fit or not.	Analyze	CO 2	BCSB06.07
8	Write R program to perform the following: i. Find the correlation matrix of iris data set ii. Plot the correlation plot on dataset and visualize giving an overview of relationships among data on iris data. iii. Perform analysis of covariance	Remember	CO 2	BCSB06.04

UNIT – III
DATA MODELS

PART – A (SHORT ANSWER QUESTIONS)

1	State the applications of regression models?	Remember	CO 3	BCSB06.09
2	List out various types of regression models?	Understand	CO 3	BCSB06.10
3	State the difference between numerical and categorical parameters?	Understand	CO 3	BCSB06.08
4	Justify a single model on data is suggestible.	Remember	CO 3	BCSB06.10
5	State the various types of functions in R to support linear regression?	Remember	CO 3	BCSB06.09
6	State the various attributes to evaluate the multiple regression?	Understand	CO 3	BCSB06.07
7	State the residuals impact in linear model.	Remember	CO 3	BCSB06.08
8	What are the evaluating measures in regression models.	Remember	CO 3	BCSB06.10
9	What is the role of machine learning algorithms in data model.	Remember	CO 3	BCSB06.09
10	State the steps to evaluate the data model.	Understand	CO 3	BCSB06.08

PART – B (LONG ANSWER QUESTIONS)

1	**Outline** about the learning of a model? Write any four learning techniques and in each case give the expression for weight-updating.	Remember	CO 3	BCSB06.07
2	**Describe** the limitations of the perception model. How to create and evaluate a data model. Describe with one case study.	Understand	CO 3	BCSB06.10

3	**List** out the applications of Machine learning with example. Discuss about a consistent learner and what it means for a set of training examples to be linearly separable.	Understand	CO 3	BCSB06.09
4	**Write** the steps to describe the process to create and evaluate the data model for the given data.	Remember	CO 3	BCSB06.07
5	**Predict** whether an email is a spam and should be delivered to the Junk folder. Suggest suitable data model.	Remember	CO 3	BCSB06.10
6	**Discuss** about clustering in detail? Elaborate k means algorithm.	Understand	CO 3	BCSB06.09
7	Write a R script to implement	Remember	CO 3	BCSB06.11
8	**State** about conditional probability in the bayes theorem for different type of data classification problem with suitable example.	Remember	CO 3	BCSB06.10
9	**Calculate** the Jaccard coefficient for the given data p = 1 0 0 0 0 0 0 0 0 0, q = 0 0 0 0 0 0 1 0 0 1;	Remember	CO 3	BCSB06.07
10	**List** the different types clustering. Write about k-nn algorithm. Write a R script to cluster the mtcars dataset using k-nn algorithm.	Understand	CO 3	BCSB06.09
PART – C (PROBLEM SOLVING AND CRITICAL THINKING QUESTIONS)			

1	Find the suitable Machine learning algorithm for the following real time problems:
	   i. Predicting Housing Prices
	   ii. Analyzing Sentiment
	   iii. Finding Similar Documents
	   iv. Recommending Products
		Understand	CO 3	BCSB06.07

2	How can you identify, the best fit data model from the given dataset.	Remember	CO 3	BCSB06.08
3	Describe the prediction model in terms of the following measures for best fit: Residual standard error, Multiple R-squared, F-statistic, p-value	Remember	CO 3	BCSB06.08
4	State the suitable classification algorithm for pima Indian diabetes dataset classification. Improve the accuracy by performing multiple preprocess steps.	Understand	CO 3	BCSB06.10

UNIT – IV
ARTIFICIAL NEURAL NETWORKS

PART – A (SHORT ANSWER QUESTIONS)

1	Draw the basic structure of neural network.	Remember	CO 4	BCSB06.09
2	List various types of problems for neural network learning?	Understand	CO 4	BCSB06.12
3	State the difference between neural network and multi-layer neural network?	Understand	CO 4	BCSB06.11
4	What is meant by perceptions in neural network?	Remember	CO 4	BCSB06.11
5	State the various types of layers in neural network design?	Remember	CO 4	BCSB06.09
6	State the remarks in multi-layer perceptions?	Understand	CO 4	BCSB06.09
7	State the role of error function in neural network.	Remember	CO 4	BCSB06.11
8	What are the evaluating measures of neural network model?	Understand	CO 4	BCSB06.10
9	Define Multi-layer networks	Remember	CO 4	BCSB06.09
10	Give a short note on feed forward networks	Understand	CO 4	BCSB06.11

PART – B (LONG ANSWER QUESTIONS)

1	**Outline** about the learning of a model? Write any four learning techniques and in each case give the expression for weight-updating.	Remember	CO 4	BCSB06.09
2	**Describe** the limitations of the perception model. How to create and evaluate a data model. Describe with one case study.	Understand	CO 4	BCSB06.11

3	**List** out the applications of Machine learning with example. Discuss about a consistent learner and what it means for a set of training examples to be linearly separable.	Understand	CO 4	BCSB06.10
4	**Write** the steps to describe the process to create and evaluate the data model for the given data.	Remember	CO 4	BCSB06.12
5	**Predict** whether an email is a spam and should be delivered to the Junk folder. Suggest suitable data model.	Remember	CO 4	BCSB06.08
6	**Give** the basic structure of neural network and different types of ANN with real time examples.	Remember	CO 4	BCSB06.10
7	**State** the suitable classification algorithm for pima Indian diabetes dataset classification. Improve the accuracy by performing multiple preprocess steps.	Understand	CO 4	BCSB06.08
8	**How** to evaluate hypothesis of different types with examples? Explain the basics of sampling theory.	Understand	CO 4	BCSB06.10
9	**Discuss** the difference of error in two hypotheses. Differentiate the MAP (maximum a posteriori) and ML (maximum likelihood) hypothesis. Give an example of a scenario in which a MAP hypothesis is preferable to an ML hypothesis.	Remember	CO 4	BCSB06.09
10	**Compare** the learning algorithms with example in terms of problem nature, accuracy and error rate.	Remember	CO 4	BCSB06.11
	PART – C (PROBLEM SOLVING AND CRITICAL THINKING QUESTIONS			
1	Find the suitable Machine learning algorithm for the following real time problems: i. Predicting Housing Prices ii. Analyzing Sentiment iii. Finding Similar Documents iv. Recommending Products	Understand	CO 4	BCSB06.11
2	How can you identify, the best fit data model from the given dataset.	Understand	CO 4	BCSB06.10
3	Describe the prediction model in terms of the following measures for best fit:Residual standard error, Multiple R-squared, F-statistic, p-value	Understand	CO 4	BCSB06.12
4	How to evaluate hypothesis of the given problem. Describe the basic principle of sampling theory.	Understand	CO 4	BCSB06.11
5	List the steps to improve the accuracy of neural network data model.	Remember	CO 4	BCSB06.09
	UNIT – V **DELIVERING RESULTS**			
	PART – A (SHORT ANSWER QUESTIONS)			
1	State the applications of regression models?	Remember	CO 5	BCSB06.11
2	List various types of regression models?	Understand	CO 5	BCSB06.10
3	State the difference between numerical and categorical parameters?	Understand	CO 5	BCSB06.13
4	Justify a single model on data is suggestible.	Remember	CO 5	BCSB06.11
5	State the various types of functions in R to support linear regression?	Remember	CO 5	BCSB06.10
6	State the various attributes to evaluate the multiple regression?	Remember	CO 5	BCSB06.13
7	State the residuals impact in linear model.	Remember	CO 5	BCSB06.12
8	What are the evaluating measures in regression models.	Remember	CO 5	BCSB06.11
9	What is the role of machine learning algorithms in data model.	Remember	CO 5	BCSB06.10
10.	State the steps to evaluate the data model.	Understand	CO 5	BCSB06.12
	PART – B (LONG ANSWER QUESTIONS)			
1	Generalize the graphical analysis in data analysis? List the various plots in R and explain in detail.	Remember	CO 5	BCSB06.10

2	How to plot the word (text) data based on frequency of words. Write R script to plot a data frame having: {df1: {sea,river,pond,lake,pool} df2: {3,5,8,10,34}} using relevant plot.	Understand	CO 5	BCSB06.12
3	List out the applications of Machine learning with example. Discuss about a consistent learner and what it means for a set of training examples to be linearly separable.	Understand	CO 5	BCSB06.10
4	Write the steps to plot the numerical data at different granularity levels.	Remember	CO 5	BCSB06.13
5	Plot the k-means clustering algorithm outcome. State the relevant packages and functions in R	Remember	CO 5	BCSB06.10
6	How would you get the multiple plots in single window?	Remember	CO 5	BCSB06.13
7	Elaborate how to export a graph using graphics parameters. How to export the text data to plot with example.	Understand	CO 5	BCSB06.12
8	Describe boxplot() of iris$petal length attribute. Specify the observations of plot.	Understand	CO 5	BCSB06.11
9	Plot the regression model along with residuals.	Remember	CO 5	BCSB06.12
10	Write a R script for creating a boxplot of iris sepal length attribute.	Remember	CO 5	BCSB06.11
	PART – C (PROBLEM SOLVING AND CRITICAL THINKING QUESTIONS)			
1	Describe the various plots in R to visualize the data and explain the purpose of each plot in detail.	Understand	CO 5	BCSB06.10
2	Write R script to plot a data frame having: {df1: {red,green,blue,pink,black} df2: {3,5,8,10,34}} using relevant plot.	Understand	CO 5	BCSB06.12
3	List out the steps to plot the data models with relevant packages.	Remember	CO 5	BCSB06.11
4	How to partition the window to get more number of plots. Discuss on single and multi object plots in R.	Understand	CO 5	BCSB06.13
5	Discuss about the residuals with respect to observed values? State a case study to show the fitted line and residuals in logistic regression.	Understand	CO 5	BCSB06.10

Prepared by: **Dr. Chukka Santhaiah, Associate Professor**   **HOD, CSE**