

CMR COLLEGE OF ENGINEERING & TECHNOLOGY
(UGC AUTONOMOUS)
B.Tech. Data Science (Minor)

CBCS & OUTCOME BASED COURSE STRUCTURE

III-I SEMESTER								
Course Code	Course Title	Course Delivery	Hours per Week			Credits	Maximum Marks	
			L	T	P		CIE	SEE
	Data Science using R	Offline	3	0	0	3	30	70
	R-Programming Laboratory		0	0	3	1.5	30	70
Total:			3	0	3	4.5		

	III-II SEMESTER							
Course Code	Course Title	Course Delivery	Hours per Week			Credits	Maximum Marks	
			L	T	P		CIE	SEE
	Data Wrangling and Visualization	Offline	3	0	0	3	30	70
	Data Wrangling and Visualization- Lab		0	0	3	1.5	30	70
	OR		OR					
	Big Data Analytics		3	0	0	3	30	70
	Big Data Analytics Lab		0	0	3	1.5	30	70
Total:			3	0	3	4.5		

IV-I SEMESTER								
Course Code	Course Title	Course Delivery	Hours per Week			Credits	Maximum Marks	
			L	T	P		CIE	SEE
	Electives <ol style="list-style-type: none"> Exploratory Data Analysis Mining Massive Databases Social Network Analysis Web & Social Media Analytics Video Analytics Predictive Analytics 	Either equivalent online Course through MOOCS or off-line Class	4	0	0	4	30	70
Total:			4	0	0	4		

IV-II SEMESTER								
Course Code	Course Title	Course Delivery	Hours per Week			Credits	Maximum Marks	
			L	T	P		CIE	SEE
	Data Science Applications	offline	3	0	0	3	30	70
	Mini-Project related to Data science	Project	0	0	4	2	30	70
Total:			3	0	4	5		

DATA SCIENCE USING R

B.Tech. Data Science (Minor) III Year I Sem.

L	T	P	C
3	0	0	3

Course Objectives:

- Learn concepts, techniques and tools they need to deal with various facets of data science practice, including data collection and integration
- Understand the basic types of data and basic statistics
- Identify the importance of data reduction and data visualization techniques

Course Outcomes: After completion of the course, the student should be able to

CO-1: Understand basic terms what Statistical Inference means.

CO-2: Identify probability distributions commonly used as foundations for statistical modeling & Fit a model to data

CO-3: Describe the data using various statistical measures

CO-4: Utilize R elements for data handling

CO-5: Describe Data reduction & visualization techniques.

UNIT-I: Introduction

What is Data Science? - Big Data and Data Science hype – and getting past the hype - Datafication - Current landscape of perspectives - Statistical Inference - Populations and samples - Statistical modeling, probability distributions, fitting a model –Over fitting. Basics of R: Introduction, R-Environment Setup, Programming with R, Basic Data Types.

UNIT-II: Data Types & Statistical Description

Types of Data: Attributes and Measurement, What is an Attribute? The Type of an Attribute, The Different Types of Attributes, Describing Attributes by the Number of Values, Asymmetric Attributes, Binary Attribute, Nominal Attributes, Ordinal Attributes, Numeric Attributes, Discrete versus Continuous Attributes. Basic Statistical Descriptions of Data: Measuring the Central Tendency: Mean, Median, and Mode, Measuring the Dispersion of Data: Range, Quartiles, Variance, Standard Deviation, and Inter-quartile Range, Graphic Displays of Basic Statistical Descriptions of Data.

UNIT-III

Vectors: Creating and Naming Vectors, Vector Arithmetic, Vector sub setting, Matrices: Creating and Naming Matrices, Matrix Sub setting, Arrays, Class. Factors and Data Frames: Introduction to Factors: Factor Levels, summarizing a Factor, Ordered Factors, Comparing Ordered Factors, Introduction to Data Frame, sub setting of Data Frames, Extending Data Frames, Sorting Data Frames. Lists: Introduction, creating a List: Creating a Named List, Accessing List Elements, Manipulating List Elements, Merging Lists, Converting Lists to Vectors

UNIT-IV

Conditionals and Control Flow: Relational Operators, Relational Operators and Vectors, Logical Operators, Logical Operators and Vectors, Conditional Statements. Iterative Programming in R: Introduction, While Loop, For Loop, Looping Over List. Functions in R: Introduction, writing a Function in R, Nested Functions, Function Scoping, Recursion, Loading an R Packag, Mathematical Functions in R.

UNIT-V

Data Reduction: Overview of Data Reduction Strategies, Wavelet Transforms, Principal Components Analysis, Attribute Subset Selection, Regression and Log-Linear Models: Parametric Data Reduction, Histograms, Clustering, Sampling, Data Cube Aggregation.

Data Visualization: Pixel-Oriented Visualization Techniques, Geometric Projection Visualization Techniques, Icon-Based Visualization Techniques, Hierarchical Visualization Techniques, Visualizing Complex Data and Relations.

TEXTBOOKS:

1. Doing Data Science, Straight Talk from The Frontline. Cathy O’Neil and Rachel Schutt, O’Reilly, 2014.
2. Jiawei Han, Micheline Kamber and Jian Pei. Data Mining: Concepts and Techniques, 3rd ed. The Morgan Kaufmann Series in Data Management Systems.
3. K G Srinivas, G M Siddesh, “Statistical programming in R”, Oxford Publications

REFERENCE BOOKS:

1. Introduction to Data Mining, Pang-Ning Tan, Vipin Kumar, Michael Steinbanch, Pearson Education.
2. Brain S. Everitt, “A Handbook of Statistical Analysis Using R”, Second Edition, 4 LLC, 2014.
3. Dalgaard, Peter, “Introductory statistics with R”, Springer Science & Business Media, 2008.
4. Paul Teetor, “R Cookbook”, O’Reilly, 2011.

****END****

R PROGRAMMING LABORATORY

B.Tech. Data Science (Minor) III Year I Sem.

L	T	P	C
0	0	3	1.5

Lab requirements: R-Studio with R compiler

Course Objectives:

- Learn to develop programs in R-environment
- Learn to install and use R-built in packages

Course Outcomes: After completion of the course, the student should be able to

CO-1: Install & Use R and R-Studio on various environments.

CO-2: Write simple R-scripts to manipulate data using R-programming language constructs.

CO-3: Utilize R elements for data handling

CO-4: Develop programs using R-operators and control structures.

CO-5: Develop R-Scripts for data visualization

1. R Environment setup: Installation of R and RStudio in Windows
2. Write R commands for
 - i. Variable declaration and retrieving the value of the stored variables,
 - ii. Write an R script with comments,
 - iii. Type of a variable using class () Function.
3. Write R command to
 - i. illustrate summation, subtraction, multiplication, and division operations on vectors using vectors.
 - ii. Enumerate multiplication and division operations between matrices and vectors in R console
4. Write R command to
 - i. Illustrate the usage of Vector sub setting& Matrix sub setting
 - ii. Write a program to create an array of 3×3 matrixes with 3 rows and 3 columns.
 - iii. Write a program to create a class, object, and function
5. Write a command in R console
 - i. to create a tshirt_factor, which is ordered with levels 'S', 'M', and 'L'. Is it possible to identify from the examples discussed earlier, if blood type 'O' is greater or less than blood type 'A'?
 - ii. Write the command in R console to create a new data frame containing the 'age' parameter from the existing data frame. Check if the result is a data frame or not. Also R commands for data frame functions cbind(), rbind(), sort()
6. Write R command for
 - i. Create a list containing strings, numbers, vectors and logical values
 - ii. To create a list containing a vector, a matrix, and a list. Also give names to the elements in the list and display the list also access the list elements
 - iii. To add a new element at the end of the list and delete the element from the middle display the same
 - iv. To create two lists, merge two lists. Convert the lists into vectors and perform addition on the two vectors. Display the resultant vector.

7. Write R command for
 - i. logical operators—AND (&), OR (|) and NOT (!).
 - ii. Conditional Statements
 - iii. Create four vectors namely patientid, age, diabetes, and status. Put these four vectors into a data frame patient data and print the values using a for loop & While loop
 - iv. Create a user-defined function to compute the square of an integer in R
 - v. Create a user-defined function to compute the square of an integer in R
 - vi. Recursion function for a) factorial of a number b) find nth Fibonacci number
8. Write R code for i) Illustrate Quick Sort ii) Illustrate Binary Search Tree
9. Write R command to
 - i. illustrate Mathematical functions & I/O functions
 - ii. Illustrate Naming of functions and sapply(), lapply(), tapply() & mapply()
10. Write R command for
 - i. Pie chart & 3D Pie Chart, Bar Chart to demonstrate the percentage conveyance of various ways for traveling to office such as walking, car, bus, cycle, and train
 - ii. Using a chart legend, show the percentage conveyance of various ways for traveling to office such as walking, car, bus, cycle, and train.
 - a. Walking is assigned red color, car – blue color, bus – yellow color, cycle – green color, and train – white color; all these values are assigned through cols and lbls variables and the legend function.
 - b. The fill parameter is used to assign colors to the legend.
 - c. Legend is added to the top-right side of the chart, by assigning
11. Using box plots, Histogram, Line Graph, Multiple line graphs and scatter plot to demonstrate the relation between the cars speed and the distance taken to stop, Consider the parameters data and x Display the speed and dist parameter of Cars data set using x and data parameters

TEXT BOOK:

1. K G Srinivas, G M Siddesh, “Statistical programming in R”, Oxford Publications

****END****

Data Wrangling and Visualization

B.Tech. Data Science (Minor) III Year II Sem

L	T	P	C
3	0	0	3

Course Objectives:

- Learn concepts, techniques and tools they need to deal with various facets of data science practice, including data collection and integration
- Identify the importance of data reduction and data visualization techniques

Course Outcomes: After completion of the course, the student should be able to

CO-1: Perform scraping of data from multiple data resources

CO-2: Apply data transformation techniques and also handle missing values in the data.

CO-3: Design interactive plots to describe the data

CO-4: Create Heatmap for finding the correlations among various feature vectors.

CO-5: Generate visualizations for continuous variables

UNIT-1

Importing Data- Reading Data from Text Files, Reading Data from Excel Files. Scraping Data- Importing Tabular and Excel Files Stored Online, Scraping HTML Text, Scraping HTML Table Data. Exporting Data- Writing Data to Text Files, Writing Data to Excel Files, excel Package, Saving Data as an R Object File.

UNIT-2

Managing Data Structures in R using packages: Data Structure Basics, Managing Vectors, Managing Lists, Managing Matrices, Managing Data Frames, Dealing with Missing Values, Reshaping Data with *tidyr* package, Transforming data with *dplyr* package.

UNIT-3

Basic and Interactive Plots-scatter plot-Scatter plots with texts, labels, and lines , Connecting points in a scatter plot
Generating an interactive scatter plot Bar plot- A simple bar plot , An interactive bar plot Line plot-A simple line plot
Line plot to tell an effective story. Generating an interactive Gantt/timeline chart in R , Merging histograms , Making an interactive bubble plot .

UNIT-4

Heat Maps and Dendrograms: Introduction, Constructing a simple dendrogram, Creating dendrograms with colors and labels, Creating a heat map, Generating a heat map with customized colors , Generating an integrated dendrogram and a heat map, Creating a three-dimensional heat map and a stereo map , Constructing a tree map in R

UNIT-5

Visualizing Continuous Data: Introduction- Generating a candlestick plot, Generating interactive candlestick plots, Generating a decomposed time series , Plotting a regression line , Constructing a box and whiskers plot , Generating a violin plot , Generating a quantile-quantile plot (QQ plot), Generating a density plot ,Generating a simple correlation plot.

TEXTBOOKS:

1. Data Wrangling with R- Bradley C. Boehmke, Springer publisher
2. R Data Visualization Cookbook- Atmajitsinh Gohi, Packt Publishing

REFERENCE BOOKS:

1. Brain S. Everitt, "A Handbook of Statistical Analysis Using R", Second Edition, 4 LLC, 2014.
- 2.R for Datascience, Hadley Wickham, Garrett Golemund , O'Reilly Media
3. Paul Teetor, "R Cookbook", O'Reilly, 2011.

****END****

Data Wrangling and Visualization Lab

B.Tech. Data Science (Minor) III Year II Sem

L	T	P	C
0	0	3	1.5

Lab Requirements: R-Studio with R-Compiler

Course Objectives:

- Learn to retrieve useful insights from datasets.
- Identify the relations among various attributes using visualization charts.

Course Outcomes: After completion of the course, the student should be able to

CO-1: Perform exploratory data analysis
CO-2: Infer the hidden relations in the data.
CO-3: Create Dendrograms for clustering the data.
CO-4: Create heat map to know the correlation among the attributes.
CO5: Use various data visualization tools

Experiment-1

Download World Happiness dataset from the following link:

https://raw.githubusercontent.com/uopsych/psy611/master/labs/resources/lab5/data/world_happiness.csv

Perform the following tasks

- 1.Convert names of all the variables to lowercase
2. Filter rows to retain data only for the India
3. select observations for the United States, Mexico and India
- 4 Filter observations that are greater than the mean of happiness
5. Filter for observations that are greater than the mean of happiness but less than the mean of gdp.
- 6.Order observations by happiness attribute
- 7.use select() to select one or more variables
8. Produce a data frame of the variables country, gdp, and happiness for countries whose gdp is greater than average.

Experiment-2

Download Airquality data set from the following link

<https://www.kaggle.com/datasets/rohanrao/air-quality-data-in-india>

Perform the following tasks

- 1.Plot Horizontal Bar Plot for Ozone concentration in air
2. Vertical Bar Plot for Ozone concentration in air
3. Histogram for Maximum Daily Temperature
4. Box plot for average wind speed
- 5.Multiple Box plots, each representing an Air Quality Parameter
6. Scatter plot for Ozone Concentration per month

Experiment-3

Current Population Survey (CPS) data. These particular data consist of a random sample of 534 people from the CPS in 1985, with information on wages and other characteristics of the workers, including sex, number of years of education, years of work experience, occupational status, region of residence and union membership.

Download the dataset from the following link

<http://math.montana.edu/shancock/data/cps.csv>

Perform the following tasks

1. Is there an association between number of years of education and wage?
2. Is there an association between age and union membership?
3. Do men make more than women?

Use both plots and summary statistics to investigate these questions.

Experiment -4

Download the USAarrests dataset from the following link

<https://www.kaggle.com/code/kernelel/starter-usarrests-af149109-c/data>

Perform the following link

1. Create hierarchical clustering dendrogram
2. Create a Triangle plot
3. Creating dendrograms with colors and labels

Experiment-5

Download the Iraq body count dataset from the following link

www.iraqbodycount.org

Perform the following tasks

1. Creating a heat map
2. Generate a heat map with customized colors
3. Generate an integrated dendrogram and a heat map

Experiment-6

Download The United Nations voting dataset from the following link

<https://www.kaggle.com/datasets/unitednations/general-assembly>

Perform the following tasks

1. Filtering rows
2. Adding a year column
3. Adding a country column
4. Grouping and summarizing
5. Summarizing the full dataset
6. Summarizing by year
7. Summarizing by country
8. Sorting and filtering summarized data
9. Sorting by percentage of “yes” votes
10. Filtering summarized output

Experiment-7.

Explore the following Data visualization tools to draw basic charts

1. Tableau
2. Microsoft Power BI
3. Plotly
4. Candela

****END****

BIG DATA ANALYTICS

B.Tech. Data Science (Minor) III Year II Sem

L	T	P	C
3	0	0	3

Course Objectives

- Learn the concepts of Bigdata & Hadoop
- Learn to learn process Bigdata application using Bigdata technologies

Course Outcomes:

On completion of this course, the students are able to:

1. Describe the Big-Data and Big Data Analytics.
2. Illustrate the Hadoop Software Framework and Its Core components (HDFS and Map-Reduce).
3. Demonstrate loading and reading Data from HDFS and Processing using Map-Reduce.
4. Implement Pig Latin Scripts for processing Data.
5. Use Hive Query language for creating and querying tables.

Unit 1

Introduction to Big Data: Introduction- Big Data, Characteristics & Importance of Big Data – Four V's, Relational Database Vs Big Data, Big Data Analytics, Big Data Applications, Introduction to NoSQL Database Systems

Unit-2

Hadoop: Introduction to Hadoop, History and future of Hadoop HDFS- HDFS Architecture and How to load data into HDFS, Rack Awareness, Data node to name node communication, fault- tolerance feature of HDFS, Read data from HDFS, Block Size concept of HDFS,

Unit-3

Map Reduce: Introduction to Map Reduce and its Architecture, Hadoop Eco System, Setup Hadoop on a Single node, Simple Map Reduce Program, Executing Map Phase – Shuffling and Sorting, Reducing Phase Execution

Unit-4

HIVE: Introduction to HIVE & its Architecture, HIVE Data Types and Table Creation, loading data in HIVE Tables, Managed Tables and External Tables, Querying HIVE Tables.

Unit-5

Hbase and Cassandra: Introduction to HBase, Row-Oriented vs Column-Oriented data stores, HBase Architecture, Understanding HBase Data Model, Casandra: Introduction, Features of Cassandra, Data Replication in Cassandra, Cassandra Query language (CQL), Cassandra Data Model.

Text Books:

1. Big Data, Black Book: Covers Hadoop 2, MapReduce, Hive, YARN, Pig, R and Data Visualization, DT Editorial Services, DreamTech
2. Programming Pig by Alan Gates, O'Reilly; 2nd Revised edition
3. Programming Hive by Edward Capriolo, Dean Wampler, Jason Rutherglen, O'Reilly; First edition

References

1. V K Jain, "BIG DATA and HADOOP", 2017 edition, Khanna Book Publishing. ISBN:978-93-82609-13-1
2. Pramod J. Sadalage, Martin Fowler, "NoSQL Distilled: A Brief Guide to the Emerging World of Polyglot Persistence", Addison-Wesley. ISBN: 9780133036121
3. Vignesh Prajapati, "Big data analytics with R and Hadoop", 2013, SPD.
4. Tom White, "Hadoop: The Definitive Guide", Third Edition, O'Reilley, 2012.
5. Lars George, "HBase: The Definitive Guide", O'Reilley, 2011.

****END****

BIG DATA ANALYTICS LAB

B.Tech. Data Science (Minor) III Year II Sem

L	T	P	C
0	0	3	1.5

Note: Linux or windows operating system, Java 1.6 or higher and Hadoop-1.2.1 or higher version

Course Objectives

- Learn to create Bigdata & Hadoop environment
- Learn to learn process Bigdata application using Bigdata technologies

Course Outcomes:

On completion of this course, the students shall be able to:

- CO-1. Install and configure Hadoop software framework for Bigdata applications.
- CO-2. Operate Hadoop system and manage files & resources on Hadoop machine.
- CO-3. Develop Map-Reduce applications and execute them on Hadoop Machine.
- CO-4. Explore bigdata to retrieve useful insights.
- CO-5. Create Map -Reduce application for various distributed applications.

Week1&2: Hadoop Installation

Install and configure Hadoop software framework in any one of the modes (standalone, sudo or fully).

Week3: starting Hadoop server

- a) Check which processes are running using jps.
- b) Format NameNode
- c) Start Hadoop processes
- d) Use HDFS Web interface to monitor Hadoop cluster.

Week4: Hadoop commands

- a) Introducing Hadoop command
- b) Navigating the location where the DataNodes store data
- c) Checking the current status of HDFS by using the “fsck” command
- d) Loading the small size data by using the “copyFromLocal”
- e) Show the output of the copyFromLocal command
- f) Reading the data from HDFS.
- g) Creating a Directory in HDFS.
- h) Removing files from HDFS.

Week5&6: MapReduce Program

- a) Write a MapReduce program for counting number of words in a given file or document.
- b) Create a jar file
- c) Run the jar file and observe mapper process and reducer process.
- d) Read the output file and display the results.

Week7&8: MapReduce Application

Using movie lens data (<https://www.kaggle.com/grouplens/movielens-20m-dataset>)

1. List all the movies and the number of ratings
2. List all the users and the number of ratings they have done for a movie

Week9&10:

1. List all the Movie IDs which have been rated (Movie Id with at least one user rating it)
2. List all the Users who have rated the movies (Users who have rated at least one movie)

Week11&12:

1. List of all the User with the max, min, average ratings they have given against any movie
2. List all the Movies with the max, min, average ratings given by any user.

Additional Experiments:

Create a MapReduce Application for the following:

1. Social Networks
2. Entertainment
3. Electronic Commerce
4. Fraud Detection
5. Search and Advertisement Mechanisms

****END****