

DATA SCIENCE LABORATORY

II Semester: CSE (Data Science)

Course Code	Category	Hours / Week			Credits	Maximum Marks		
ACDC02	Foundation	L	T	P	C	CIA	SEE	Total
		1	0	2	2	30	70	100
Contact Classes: 12	Tutorial Classes: Nil	Practical Classes: 24			Total Classes:36			

Prerequisite: There is no prerequisites required

I. COURSE OVERVIEW:

This course is an inter-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from many structural and unstructured data. Data science is related to data mining, machine learning and big data.

II. COURSE OBJECTIVES:

The students will try to learn:

- I. The graphs and charts with the help of statistics in R Programming.
- II. The given data with different distribution functions.
- III. The relevance and importance of the theory in solving practical problems in the real world.

III. COURSE SYLLABUS:

WEEK – 1 : INTRODUCTION TO COMPUTING

- a) Installation of R
- b) The basics of R syntax, workspace
- c) Matrices and lists
- d) Subsetting
- e) System-defined functions; the help system
- f) Errors and warnings; coherence of the workspace

WEEK – 2 GETTING USED TO R: DESCRIBING DATA

- a) Viewing and Manipulating Data
- b) Plotting Data
- c) Reading the Data from console, file (.csv) local disk and Web
- d) Working with larger datasets

WEEK – 3 VISUALIZING DATA

- a) Tables, charts and plots.
- b) Visualizing Measures of Central Tendency, Variation, and Shape.
- c) Box plots, Pareto diagrams.
- d) Find the mean, media, standard deviation and quantiles of a set of observations.

Note: Experiment with real as well as artificial data sets.

WEEK – 4 BINOMIAL DISTRIBUTION

- a) Study of binomial distribution.
- b) Plots of density and distribution functions.
- c) Normal approximation to the Binomial distribution.

WEEK – 5 PROBABILITY DISTRIBUTIONS

- a) Random number generation Distributions, the practice of simulation
- b) Generate and Visualize Discrete and continuous distributions using the statistical environment.
- c) Demonstration of CDF and PDF uniform and normal, binomial Poisson distributions.
- d) Generate artificial data using and explore various distribution and its properties. Various parameter changes may be studied.

WEEK - 6 EXPLORATORY DATA ANALYSIS

Demonstrate Range, summary, mean, variance, median, standard deviation, histogram, box plot, scatterplot

WEEK – 7 DENSITIES OF RANDOM VARIABLES

- a) Distributions in R
- b) Matching a Density to Data
- c) Making Histograms

WEEK - 8 CORRELATION

- a) How to calculate the correlation between two variables.
- b) How to make scatter plots.
- c) Use the scatter plot to investigate the relationship between two variables

WEEK - 9 TESTS OF HYPOTHESES

- a) Perform tests of hypotheses about the mean when the variance is known.
- b) Compute the p-value.
- c) Explore the connection between the critical region, the test statistic, and the p-value

WEEK – 10 ESTIMATING A LINEAR RELATIONSHIP

Demonstration on a Statistical Model for a Linear Relationship

- a) Least Squares Estimates
- b) The R Function lm
- c) Scrutinizing the Residuals

WEEK – 11 APPLY-TYPE FUNCTIONS

- a) Defining user defined classes and operations, Models and methods in R
- b) Customizing the user's environment
- c) Conditional statements
- d) Loops and iterations

WEEK – 12 STATISTICAL FUNCTIONS IN R

- a) Demonstrate Statistical functions in R
- b) Statistical inference, contingency tables, chi-square goodness of fit, regression, generalized linear models, advanced modeling methods

IV. REFERENCE BOOKS:

1. Maria Dolores Ugarte , Ana F. Militino , Alan T. Arnholt “Probability and Statistics with R” 2nd Edition on, CRC Press, 2016.
2. P. Dalgaard. “Introductory Statistics with R” Springer, 2nd Edition, 2008.

V. WEB REFERENCES:

1. <http://nptel.ac.in/courses/106104135/48>
2. <http://nptel.ac.in/courses/110106064/>