#### 20BS4101 – STATISTICS WITH R

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Course Category:** | | Basic Science | | | | | | | | **Credits:** | | | | | | | 3 | |
| **Course Type:** | | Theory | | | | | | | | **Lecture-Tutorial-Practice:** | | | | | | | 2-0-2 | |
| **Prerequisites:** | | 20IT3302 Discrete Mathematical Structures | | | | | | | | **Continuous Evaluation:** | | | | | | | 30 | |
|  | | | | | | | | | | **Semester end Evaluation:** | | | | | | | 70 | |
| **Total Marks:** | | | | | | | 100 | |
| **Course Outcomes** | Upon successful completion of the course, the student will be able to: | | | | | | | | | | | | | | | | | |
| CO1 | | Understand the fundamental syntax of R through readings, practice exercises, demonstrations, writing R code and Visualize data attributes using ggplot2 and other R packages. | | | | | | | | | | | | | | | |
| CO2 | | Manipulate numeric and textual data types using the R programming language and RStudio. | | | | | | | | | | | | | | | |
| CO3 | | Apply the knowledge of Probability and conduct Tests of Hypothesis for Statistical Inference. | | | | | | | | | | | | | | | |
| CO4 | | Fit some basic types of Statistical Models. | | | | | | | | | | | | | | | |
| **Contribution of Course Outcomes towards achievement of Program Outcomes**  **(1-Low, 2- Medium, 3- High)** |  | | PO1 | PO2 | PO3 | PO 4 | PO 5 | PO 6 | PO 7 | | PO 8 | PO 9 | PO 10 | PO 11 | PO 12 | PSO1 | | PSO2 |
| CO1 | | 2 | 2 |  |  |  |  |  | |  |  |  |  |  | 2 | | 2 |
| CO2 | | 2 | 1 |  |  | 2 |  |  | |  |  |  |  |  | 2 | | 2 |
| CO3 | | 3 | 2 |  | 3 |  |  |  | |  |  |  | 2 |  | 3 | | 2 |
| CO4 | | 3 | 2 | 3 | 2 | 2 |  |  | |  |  |  | 3 |  | 3 | | 2 |
| **Course Content** | **UNIT I:**  **The R Environment** : Command Line interface, R Studio, Installing R Packages.  **Basics of R** : Basic math, variable, data types, vectors, calling function, missing data, data.frames, lists, matrices, arrays.  **Reading data into R :**Reading CSVs, Excel Data.  **Case Study:**Loading data from mysql into RStudio.  Writing R functions, control statements – if and else, switch, compound tests, for loops, while loops.  **Statistical Graphs :**Base Graphs, ggplot2. | | | | | | | | | | | | | | | | | |
| **UNIT II:**  **Group manipulation** : Apply Family, aggregate, plyr, data.table.  **Data Reshaping** : cbind, rbind, joins, reshape2.  **Strings :** paste, sprint, extracting text, regular expressions.  **Case Study: String Theory:** To focus on manipulating unstructured data, this in most cases means natural language texts. Tweets are again a useful source of data for this because tweets are mainly a short (140 characters or less) character strings.  **Math Functions :**Calculating a Probability, cumulative sums and products, minima and maxima, calculus, sorting, set operations. | | | | | | | | | | | | | | | | | |
| **UNIT III:**  **Probability Distributions** : Normal Distribution, Binomial Distribution, Poisson Distribution.  **Basics Statistics** : Summary statistics, correlation and covariance, t-tests, ANOVA.  **Case Study:Popularity Contest:** Develop a test to compare two different Twitter topics to see which one is most popular(or at least which one has a higher posting rate) | | | | | | | | | | | | | | | | | |
| **UNIT IV:**  **Linear Models** : Simple Linear Regression, Multiple Regression, Logistics Regression, Poisson Regression.  **Nonlinear Models** : Nonlinear least squares, splines, generalized additive models, decision trees, random forests.  **Time Series** : Autoregressive Moving Average, VAR, GARCH.  **Clustering** : K Means, PAM, Hierarchical Clustering  **Case Study:**   1. **Word Perfect:** Analyze the actual words that appear in text documents. 2. **Decision Tree:** Implement Decision Tree, Random Forest in R for party package. | | | | | | | | | | | | | | | | | |
| **Text books and Reference books** | **Text Book(s):**   1. Jared P. Lander, “R for Everyone, Addison Wesley Data &Analytics Series, Pearson”, 2014.**(UNIT-I,II(Except Math Functions), III &IV)** 2. Norman Matloff, “The Art of R Programming, No Strach Press”, San Francisco, 2011**.(UNIT-II Math Functions)**   **Reference Books:**   1. Jeffrey Stanton, “An Introduction To Data Science”, 2012 2. G. Jay Kerns, Introduction to Probability and Statistics using R, First Edition, 2010 | | | | | | | | | | | | | | | | | |
| **E-resources and other digital material** | 1. Rafael Irizarry, Michael Love, Statistics with R, Harvard University (18, May, 2021). Available: <https://www.edx.org/course/statistics-r-harvardx-ph525-1x-1> 2. Mine Çetinkaya-Rundel, David Banks, Colin Rundel, Merlise A Clyde, Duke University, (18, May, 2021). Statistics with R Specialization. Available: <https://www.coursera.org/specializations/statistics> | | | | | | | | | | | | | | | | | |