VNR VIGNANA JYOTHI INSTITUTE OF ENGINEERING AND TECHNOLOGY HYDERABAD
B.TECH. MINOR IN DATA SCIENCE

COURSE STRUCTURE AND SYLLABUS
(Applicable from the academic year 2021-2022)

V SEMESTER (III YEAR I SEMESTER) R19

Course Code	Title of the Course	L	T	P/D	Contact Hours/ Week	Credits
19MC1DS01	Introduction to Data Science	3	0	0	3	3
19MC2DS01	R Programming Laboratory	0	0	3	3	1.5
Total		**3**	**0**	**3**	**6**	**4.5**

VI SEMESTER (III YEAR II SEMESTER) R19

Course Code	Title of the Course	L	T	P/D	Contact Hours/ Week	Credits
19MC1DS02	Data Science Applications	3	1	0	4	4
Total		**3**	**1**	**0**	**4**	**4**

VII SEMESTER (IV YEAR I SEMESTER) R19

Course Code	Title of the Course	L	T	P/D	Contact Hours/ Week	Credits
19MC1DS03	Data Wrangling and Visualization	3	0	0	3	3
19MC2DS02	Data Wrangling and Visualization Laboratory	0	0	3	3	1.5
Total		**3**	**0**	**3**	**6**	**4.5**

VIII SEMESTER (IV YEAR II SEMESTER) R19

Course Code	Title of the Course	L	T	P/D	Contact Hours/ Week	Credits
19MC1DS04	Time Series Analysis and Forecasting	3	0	0	3	3
19MP1DS01	Mini-Project	0	0	4	4	2
Total		**3**	**0**	**4**	**7**	**5**

VNR VIGNANA JYOTHI INSTITUTE OF ENGINEERING AND TECHNOLOGY

B.Tech. Minor (DS) V Semester	**L**	**T/P/D**	**C**
	3	**0**	**3**

(19MC1DS01) INTRODUCTION TO DATA SCIENCE

COURSE OBJECTIVES:

- To learn concepts, techniques, and tools they need to deal with various facets of data science practice, including data collection and integration
- To understand the basic types of data and basic statistics
- To exploring data analysis, predictive modeling, descriptive modeling, data product creation, evaluation, and effective communication
- To 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. Identify probability distributions commonly used as foundations for statistical modeling. Fit a model to data
CO-2: Describe the data using various statistical measures
CO-3: Utilize R elements for data handling
CO-4: Perform data reduction and apply 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 Package, 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.

UNIT – VI:

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

TEXT BOOKS:

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

REFERENCES:

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

VNR VIGNANA JYOTHI INSTITUTE OF ENGINEERING AND TECHNOLOGY

B.Tech. Minor (DS) V Semester	L	T/P/D	C
	0	3	1.5

(19MC2DS01) R PROGRAMMING LABORATORY

COURSE OBJECTIVES:

- To provide insights about the basic roles of various statistical methods in building computer applications
- To develop a greater understanding of the importance of Data Visualization techniques
- To make inferences about the box plots and histograms using sample data
- To provide an understanding on the importance and techniques of predicting a relationship between the two sets of data

COURSE OUTCOMES: After completion of the course, the student should be able to
CO-1: Install and use R for simple programming tasks
CO-2: Understanding the functionality of R by using add-on packages and extract data from files and other sources and perform various data manipulation tasks on them
CO-3: Applying R Graphics and Tables to visualize results of various statistical operations on data
CO-4: Apply the knowledge of R gained to data Analytics for real life applications

EXERCISES:

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 patient id, 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
    iii. 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 BOOKS:

1. Statistical programming in R, K. G. Srinivas, G. M. Siddesh, Oxford Publications
2. R for Beginners, Sandip Rakshit, 1st Edition, McGraw Hill Education, 2017

REFERENCES:

1. R-The Statistical Programming Language, Dr. Mark Gardner, Wiley India Pvt. Ltd., 2013
2. Introduction to the Theory of Statistics, A. M. Mood, F. A. Graybill and D. C. Boes, 3rd Edition, McGraw Hill Education, 2017

VNR VIGNANA JYOTHI INSTITUTE OF ENGINEERING AND TECHNOLOGY

B.Tech. Minor (DS) VI Semester	**L**	**T/P/D**	**C**
	3	**1**	**4**

(19MC1DS02) DATA SCIENCE APPLICATIONS

COURSE OBJECTIVE:

- To provide deep knowledge of Data Science and how it can be applied in various fields to make the life easy
- To obtain a Comprehensive knowledge of various applications of Data Science
- To understand the concept of PCA and its related applications
- To analyze the use of data in various Data Science applications i.e., in education, social media, health care, bioinformatics

COURSE OUTCOMES: After completion of the course, the student should be able to
CO-1: Correlate the data science and solutions to modern problem
CO-2: Decide when to use which type of technique in data science
CO-3: Applying various Data Science techniques to Education and social media
CO-4: Analyse various applications of Data Science in Health care and bio informatics

UNIT – I:
Data Science Applications in various domains, Challenges and opportunities, tools for data scientists. Recommender systems – Introduction, methods, application, challenges.
Principle Component Analysis: For dimensionality and its reduction, Data Visualization with MNIST Dataset, Geometric interpretation of Eigen values, column Standardization.

UNIT – II:
Time Series Data: Stock market index movement forecasting.
Supply Chain Management: Real world case study in logistics.
Support Vector Machines: Geometric interpretation, kernel trick, polynomial kernel, RBF kernel, Domain specific kernel. Decision tree, sample decision tree, building a decision tree with entropy, information gain, Gini-impurity.

UNIT – III:
Data Science in Education, social media.
Text Vectorization: Text2vector, Bag of Words (BOW), Stemming, Unigram, Bigram, TFIDF, word2vector, Avgw2vector, TFIDF Weight edWord2Vector, Count Vectorizing, Ngrams, Implementation of TFIDF, Word2Vector.

UNIT – IV:
Data Science in Healthcare, Bioinformatics,
Introduction to t-SNE, Geometrical interpretation of t-SNE, How to use t-SNE effectively t-SNE with MNIST dataset.

UNIT – V:
Logistic regression, K-nearest neighbour, Naive Bayes algorithm, imbalanced data vs balanced data set.

UNIT – VI:

Case studies in data optimization using Python

TEXT BOOKS:

1. Data Science and its applications, Aakanksha Sharaff, G. K. Sinha, CRC Press, 2021
2. Data Science: Theory, Analysis and Applications, Q. A. Menon, S. A. Khoja, CRC Press, 2020

REFERENCES:

1. Principal Component Analysis, I. T. Jolliffe, 2nd Edition, Springer
2. Machine Learning, Tom M. Mitchell, McGraw-Hill
3. The Elements of Statistical Learning, Trevor Hastie, Robert Tibshirani, Jerome Friedman, 2nd Edition, Springer

VNR VIGNANA JYOTHI INSTITUTE OF ENGINEERING AND TECHNOLOGY

B.Tech. Minor (DS) VII Semester	L	T/P/D	C
	3	0	3

(19MC1DS03) DATA WRANGLING AND DATA VISUALIZATION

COURSE OBJECTIVES:
- To learn data wrangling techniques
- To introduce and understand the concept of visual perception
- To acquire core skills for visual analysis
- To apply the visualization techniques in real-life applications

COURSE OUTCOMES: After completion of the course, the student should be able to
CO-1: Perform data wrangling
CO-2: Explain principles of visual perception
CO-3: Apply visualization techniques for various data analysis tasks
CO-4: Evaluate visualization techniques

UNIT - I:
Introduction: Introductory focus on Python Data structure for NumPy & pandas, Auditing data to improve quality.
Data Wrangling: Introduction, Need of data cleanup, data clean up basics.

UNIT - II
Tasks of Data Wrangling: Data wrangling tools with – formatting, outliers, duplicates, Normalizing and standardizing data. Importance of analytics and visualization in the era of data abundance.

UNIT - III:
Introduction of Visual Perception: visual representation of data, Gestalt principles, information overloads. Creating visual representations, visualization reference model, visual mapping, visual analytics, Design of visualization applications.
Basic Plotting: Line plot - Bar plot - Pie Chart - Scatter Plot - Histogram

UNIT - IV:
Classification of visualization systems, Interaction and visualization techniques misleading, Visualization of one, two and multi-dimensional data, text and text documents.

UNIT - V:
Visualization of groups, trees, graphs, clusters, networks, software, Metaphorical visualization

UNIT - VI:
Visualization of volumetric data, vector fields, processes and simulations, Visualization of maps, geographic information, GIS systems, collaborative visualizations, evaluating visualizations

TEXT BOOKS:

1. Data Wrangling with Python: Tips and Tools to Make Your Life Easier, Jacqueline Kazil and Katharine Jarmul, O'Reilly
2. Interactive Data Visualization: Foundations, Techniques, and Applications, Ward, Grinstein Keim, Natick A. K. Peters Ltd.

REFERENCES:

1. The Visual Display of Quantitative Information, E. Tufte, Graphics Press
2. Data Wrangling with Python: Creating Actionable Data from Raw Sources, Dr. Tirthajyoti Sarkar, Shubhadeep Roychowdhury, Packet Publishing Ltd., 2019

VNR VIGNANA JYOTHI INSTITUTE OF ENGINEERING AND TECHNOLOGY

B.Tech. Minor (DS) VII Semester	**L**	**T/P/D**	**C**
	0	**3**	**1.5**

(19MC2DS02) DATA WRANGLING AND DATA VISUALIZATION LABORATORY

COURSE OBJECTIVES:

- To learn data wrangling techniques
- To introduce and understand the concept of visual perception
- To acquire core skills for visual analysis
- To apply the visualization techniques in real-life applications

COURSE OUTCOMES: After completion of the course, the student should be able to
CO-1: Perform experimentation on data wrangling
CO-2: Explain principles of visual perception with datasets
CO-3: Apply visualization techniques for various data analysis tasks
CO-4: Evaluate visualization techniques

LIST OF EXPERIMENTS:

Implement the following experiments using Python

1. Find missing values and perform data imputation
2. Find outliers in a chosen dataset
3. Methods to handle duplicate data
4. Perform data normalization
5. Explore 2-D charts such as Clustered bar charts, connected dot plots, pictograms, bubble charts, radar charts, polar charts, Range chart, Box-and-whisker plots, univariate scatter plots, histograms word cloud, pie chart, waffle chart, stacked bar chart, tree map.
6. Multi-dimensional data visualization
7. Graph data visualization

TEXT BOOKS:

1. Data Wrangling with Python: Tips and Tools to Make Your Life Easier, Jacqueline Kazil and Katharine Jarmul, O'Reilly
2. Data Visualization A Handbook for Data Driven Design, Andy Kirk, Sage Publications, 2016
3. Understanding Data with Graphs, Philipp K. Janert, Gnuplot in Action, Manning Publications, 2010

VNR VIGNANA JYOTHI INSTITUTE OF ENGINEERING AND TECHNOLOGY

B.Tech. Minor (DS) VIII Semester	L	T/P/D	C
	3	0	3

(19MC1DS04) TIME SERIES ANALYSIS AND FORECASTING

COURSE OBJECTIVES:

- To learn basic analysis of time series data and time series regression
- To learn auto-regressive and model averaging models
- To learn basic concepts of forecasts for a time series data
- To learn how to perform the advanced computation like Multivariate Time Series Models and Forecasting

COURSE OUTCOMES:

CO-1: Apply ideas to real time series data and interpret outcomes of analyses.
CO-2: Learn how to construct a time-series plot &amp; identify the underlying patterns in the data
CO-3: Learn basic concepts featurization techniques in time series analysis
CO-4: Learn how to develop forecasts for a time series that has a seasonal pattern
CO-5: Demonstrate an advanced understanding the underlying concepts in the time series and frequency domains for computation, visualization, and analysis of time series data

UNIT – I:

Introduction of time series analysis, The Nature of Time Series Data, Internal structures of time series, Time Series Statistical Models, Moving Window for Time series data.

UNIT – II:

Fourier decomposition, deep learning features LSTM, Image Histogram, Deep learning feature CNN, Relational data, Graph data, Indicator Variable, Feature binning, interaction variable.

UNIT – III:

Mathematical transforms, Model specific featurizations, Feature orthogonality, Domain specific featurizations, Feature slicing.
Kaggle Winner Solutions: Students are suggested to carry out with at least 3 Kaggle winner competition case study solutions and present the report for that (https://www.kaggle.com/code/sudalairajkumar/winning-solutions-of-kaggle-competitions)

UNIT – IV:

Autocorrelation and Partial autocorrelation, General Approach to Time Series Modelling and Forecasting, Evaluating and Monitoring Forecasting Model Performance.

UNIT – V:

Introduction Least Squares Estimation in Linear Regression Models, Statistical Inference in Linear Regression, Model Adequacy Checking, Regression Models for General Time Series Data.

UNIT – VI:

Exponential Smoothing, First order and Second order, Multivariate Time Series Models and Forecasting, Multivariate Stationary Process, ARIMA model, Vector AR (VAR) Models.

TEXT BOOKS:

1. Introduction to Time Series Analysis and Forecasting, Douglas C. Montgomery, Cheryl L. Jen, 2nd Edition, Wiley Series in Probability and Statistics, 2015
2. Master Time Series Data Processing, Visualization, and Modeling Using Python, Dr. Avishek Pal, Dr. P. K. S. Prakash, 2017

REFERRENCES:

1. Introduction to Time Series Analysis, P. J. Brockwell and R. A. Davis
2. Time Series Analysis and Its Applications, Robert H. Shumway and David S. Stoffer
3. Introduction to Statistical Time Series, W. A. Fuller
4. Time Series Analysis, Wilfredo Palma