**CS3352 FOUNDATIONS OF DATA SCIENCE L T P C**

**3 0 0 3**

**COURSE OBJECTIVES:**

 To understand the data science fundamentals and process.

 To learn to describe the data for the data science process.

 To learn to describe the relationship between data.

 To utilize the Python libraries for Data Wrangling.

 To present and interpret data using visualization libraries in Python

**UNIT I INTRODUCTION 9**

Data Science: Benefits and uses – facets of data - Data Science Process: Overview – Defining research goals – Retrieving data – Data preparation - Exploratory Data analysis – build the model– presenting findings and building applications - Data Mining - Data Warehousing – Basic Statistical descriptions of Data

**UNIT II DESCRIBING DATA 9**

Types of Data - Types of Variables -Describing Data with Tables and Graphs –Describing Data with Averages - Describing Variability - Normal Distributions and Standard (z) Scores

**UNIT III DESCRIBING RELATIONSHIPS 9**

Correlation –Scatter plots –correlation coefficient for quantitative data –computational formula for correlation coefficient – Regression –regression line –least squares regression line – Standard error of estimate – interpretation of r2 –multiple regression equations –regression towards the mean

**UNIT IV PYTHON LIBRARIES FOR DATA WRANGLING 9**

Basics of Numpy arrays –aggregations –computations on arrays –comparisons, masks, boolean logic – fancy indexing – structured arrays – Data manipulation with Pandas – data indexing and selection – operating on data – missing data – Hierarchical indexing – combining datasets – aggregation and grouping – pivot tables

**UNIT V DATA VISUALIZATION 9**

Importing Matplotlib – Line plots – Scatter plots – visualizing errors – density and contour plots – Histograms – legends – colors – subplots – text and annotation – customization – three dimensional plotting - Geographic Data with Basemap - Visualization with Seaborn.

**TOTAL:45 PERIODS**

**TEXT BOOKS**

**1.** David Cielen, Arno D. B. Meysman, and Mohamed Ali, “Introducing Data Science”, Manning Publications, 2016. (Unit I)

2. Robert S. Witte and John S. Witte, “Statistics”, Eleventh Edition, Wiley Publications, 2017. (Units II and III)

3. Jake VanderPlas, “Python Data Science Handbook”, O’Reilly, 2016. (Units IV and V)

**REFERENCES:**

Allen B. Downey, “Think Stats: Exploratory Data Analysis in Python”, Green Tea Press,2014

**CS3361 DATA SCIENCE LABORATORY L T P C**

**0 0 4 2**

**COURSE OBJECTIVES:**

 To understand the python libraries for data science

 To understand the basic Statistical and Probability measures for data science.

 To learn descriptive analytics on the benchmark data sets.

 To apply correlation and regression analytics on standard data sets.

 To present and interpret data using visualization packages in Python.

**LIST OF EXPERIMENTS:**

1. Download, install and explore the features of NumPy, SciPy, Jupyter, Statsmodels and Pandas packages.
2. Working with Numpy arrays
3. Working with Pandas data frames
4. Reading data from text files, Excel and the web and exploring various commands for doing descriptive analytics on the Iris data set.
5. Use the diabetes data set from UCI and Pima Indians Diabetes data set for performing the following:
   1. Univariate analysis: Frequency, Mean, Median, Mode, Variance, Standard Deviation, Skewness and Kurtosis.
   2. Bivariate analysis: Linear and logistic regression modeling
   3. Multiple Regression analysis
   4. Also compare the results of the above analysis for the two data sets.
6. Apply and explore various plotting functions on UCI data sets.
   1. Normal curves
   2. Density and contour plots
   3. Correlation and scatter plots
   4. Histograms
   5. Three dimensional plotting
7. Visualizing Geographic Data with Basemap

**TOTAL: 60 PERIODS**