# Terna Engineering College

## Department of Artificial Intelligence and Data Science

## Program: Sem VI

## Course: Data Analytics and Visualization Lab

# Experiment No.02

## PART A

(PART A: TO BE REFERRED BY STUDENTS)

**A.1 Aim:** To study data analytics libraries in Python and R.

## A.2 Theory:

Data analytics is the process of analyzing raw data to find trends and answer questions. It has a broad scope across the field. This process includes many different techniques and goals that can shift from industry to industry.

The data analytics process has components that can help a variety of initiatives. By combining these components, a successful data analytics initiative can help answer business questions related to historical trends, future predictions and decision-making.

Python and R are preferred languages for data analytics due to their rich ecosystem of libraries, extensive community support, and powerful tools for statistical analysis, data manipulation, machine learning, and visualization. Python’s syntax and readability make it beginner-friendly and conducive to collaborative coding. R’s ability to produce publication-quality graphs is advantageous for data exploration and presentation.

## Python Libraries:

### Numpy and Scipy – Fundamental Scientific Computing

NumPy stands for Numerical Python. The most powerful feature of NumPy is n-dimensional array. This library also contains basic linear algebra functions, Fourier transforms, advanced random number capabilities and tools for integration with other low level languages like Fortran, C and C++.

SciPy stands for Scientific Python. It is built on NumPy. Scipy is one of the most useful library for variety of high level science and engineering modules like discrete Fourier transform, Linear Algebra, Optimization and Sparse matrices.

### Pandas – Data Manipulation and Analysis

Pandas for structured data operations and manipulations. It is extensively used for data munging and preparation. Pandas were added relatively recently to Python and have been instrumental in boosting Python's usage in data scientist community.

### Matplotlib – Plotting and Visualization

Matplotlib for plotting vast variety of graphs, starting from histograms to line plots to heat plots.. You can use Pylab feature in ipython notebook (ipython notebook –pylab = inline) to use these plotting features inline. If you ignore the inline option, then pylab converts ipython environment to an environment, very similar to Matlab.

### Scikit-learn – Machine Learning and Data Mining

Scikit Learn for machine learning. Built on NumPy, SciPy and matplotlib, this library contains a lot of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensional reduction.

### Seaborn – For Statistical Data Visualization

Seaborn for statistical data visualization. It is a library for making attractive and informative statistical graphics in Python. It is based on matplotlib. Seaborn aims to make visualization a central part of exploring and understanding data.

### StatsModels – Statistical Modeling, Testing, and Analysis

Statsmodels for statistical modeling. It is a Python module that allows users to explore data, estimate statistical models, and perform statistical tests. An extensive list of descriptive statistics, statistical tests, plotting functions, and result statistics are available for different types of data and each estimator.

## R Libraries:

### ggplot2

It is one of the most popular and widely used R package for data visualization and exploratory data analysis. we can create interactive data visualizations using this package. It provides a wide range of pretty plots that take care of minute details as well as drawing legends. This package works under deep grammar called “Grammar of graphics”. It provides a wide range of plots like scatterplot, bubble plots. Jitter plots are charts, histograms, density plots, box plots, violin plots, dendrograms, and many more.

## dplyr

It is one of the most used R packages for data science and machine learning tasks. This package is written by Hadley Wickham. It is used to solve data manipulation tasks. It has a set of functions for data manipulation. It is also called a grammar of data manipulation. It has s set of verbs that help us to solve the most challenging data manipulation tasks such as mutate(), select(), filter(), summarise(), arrange().

## Caret

Caret stands for classification and regression training. One of the primary tools in the package is the train function which can be used to. evaluate, using re-sampling, the effect of model tuning parameters on performance. Caret has several functions that attempt to streamline the model building and evaluation process, as well as feature selection and other techniques. This package alone is all you need to know for solve almost any supervised machine learning problem. It provides a uniform interface to several machine learning algorithms and standardizes various other tasks such as Data splitting, pre-processing, feature selection, variable importance estimation etc.

## Lubridate

Another fantastic R library which gets a lot of use, especially in real life applied situations is Lubridate. Lubridate is a great library for wrangling and cleaning time series data and managing any time related variables which you are working with. we can do everything with date arithmetic using this library, although understanding & using available functionality can be somewhat complex here.

## Shiny

Shiny brings together the computational power of R and the interactivity of the modern web. Shiny lets you interact and communicate with your team on the same platform for greater transparency and collaboration. It is the perfect tool for building interactive web apps straight from R. You can either host standalone apps on a webpage, or you can embed them in R Markdown documents. Not just that, Shiny also lets you build interactive dashboards. It is packed with a wide range of built-in input widgets. Once your Shiny apps are created, you can extend them using HTML widgets, CSS themes, and JavaScript actions.

## Knitr

Knitr is essential publishing software for R. It’s purpose is to produce reproducible reports in a variety of formats. This package also enables integration of R code into LaTeX, Markdown, LyX, HTML, AsciiDoc, and reStructuredText documents. You can add R to a markdown document and easily generate reports in HTML, Word and other formats.

**PART B**

(PART B: TO BE COMPLETED BY STUDENTS)

***(Students must submit the soft copy as per following segments within two hours of the practical. The soft copy must be uploaded on the Blackboard or emailed to the concerned lab in charge faculties at the end of the practical in case the there is no Black board access available)***

| | |
|---|---|
| Roll. No. A12 | Name: Sufiyan Khan |
| Class: TE – AI & DS | Batch: A1 |
| Date of Experiment: | Date of Submission: 18/02/24 |
| Grade: | |

## B.1 Conclusion:

Thus we have successfully explored and studied data analytics libraries in Python and R languages and understood their purpose and use cases in data analytics.