# EXPERIMENT-1

# TOOLS USED IN DATA SCIENCE

While working with data Science projects, different tools are used in different Stages.

## 1. Python:

- Python is a simple & powerful language.
- It Enables you to provide solutions to the given problems.
- It is more Suitable for scripting and application development because of its features:
  - ✓ Unsophisticated
  - ✓ High-level
  - ✓ Portable
  - ✓ Interpreted
  - ✓ Object Oriented
  - ✓ Extensive Libraries

## 2. Jupytor Notebook

- It is a powerful Environment for data Science projects.
- It is an Interactive Environment for Writing Code, combining code & results, to write explanations in a single document.

## 3. R-programming

- It is a powerful tool in data science specially for Statistical Analysis & Graphics.
- R is Completely free and open Source.
- R is great tool for EDA analysis and its ggplotz library is used to create interactive visualizations from the data.

## 4. Automated Methods (Tools) for Data Collection

- It is the process of gathering data from various Sources without any manual Effort.
- It uses special tools & Software's to automate the data Capture.

### *a) Web Scraping*

- Web scraping is the process of automatically Extracting data from the websites using automated Scripts.
- HTML, CSS, JavaScript Knowledge is required.

- The python libraries (tools) Such as Beautiful Soup, Scrapy and browser Extension web scraper can be used to Extract data from the websites.

***b) API's***

- Application program Interfaces (API's) are the Communication gates to data from online Services & databases.
- The API's Such as Twitter API, Google maps API, allows to automated data access.
  *Ex: The Twitter API is useful in retrieving tweets for Sentiment Analysis*.

***c) Web Crawlers***:

- Also known as Web spyders or bots are the automated tools used to browse the websites and collect data
- Useful for Data Mining, Web Scraping, Search Engine indexing etc.
- The tools Nutch, Heritix, Selenium etc. are the examples for Web Crawlers.

## 5. Data Pre-processing & Cleaning Tools

There are various tools used in this step for Handling missing values, removing duplicates, Standardize data etc.

***a) Pandas***

- A powerful data manipulation library.
- It Can handle missing data, filtering, merging & grouping etc.

***b) Numpy***

- A python library for Numerical operations that is helpful for preparing data for ML Models.

***c) OpenRefine***

- An open Source tool for data cleaning, de-duplication & for data transformation.

***d) Scipy***

- An Advanced library for data pre-processing.

***e) Datawrangler***

- Developed by Stanford University.
- An interactive tool for data cleaning & Transformation.

## 6. Data Analysing & Visualization Tools

These tools are used to Analyse & Visualize data to see the insights, patterns & relationships.

***a) Pandas***

- Useful in EDA for quick Summaries of data.

***b) Matplotlib***

- The python library to Create Static & interactive Charts.

***c) Seaborn***

- Provides high level interface for drawing attractive and informative Statistical graphs.

## 7. Machine Learning Tools

These tools are used to train & deploy Machine Learning Models.

***a) Scikit Learn***

- A python library for Machine Learning. It has tools for data analysis & mining Such as regression, classification, clustering etc.

***b) Tensor Flow***

- An open Source library for deep learning applications developed by Google.
- They are used to build and train Neural Networks.

***c) Keras***

- A High-level Neural Networks API written in python.

*****