

IT462: Exploratory Data Analysis

Course placement: This course is a technical elective for the BTech (semester VI, VIII) and MSc-DS (semester II) students.

Credit Structure: 3-0-2-4

Course format: It is **3 hours lecture** and **2 hours Laboratory** every week.

Course content: This course provides students foundation for exploring data using the Python programming language. In particular, the numerical and graphical techniques for summarizing and displaying data. The course will cover the following contents: statistics and data, reading, cleaning, summarizing, and visualizing data using Python.

Textbook: Exploratory Data Analysis by J W Tukey

Reference books:

1. Visualizing Data by Cleveland
2. Think Stats: Exploratory Data Analysis by Allen B. Downey, 2014
3. Exploratory Data Analysis Using R by Ronald K. Pearson, CRC Press
4. Graphical Data Analysis with R by Unwin Antony
5. Graphical Exploratory Data Analysis by S. H. C. DuToit, A. G. W. Steyn, R. H. Stumpf

Assessment method: Homework/Assignments, Online quizzes/tests in Moodle, Online discussion in Moodle, term examination and a Final Project.

		Final letter grades will be assessed on the following scale:	
Grading Policy:	5% Attendance	A	[90% – 100%]
	25% Quizzes/Assignments	B	[80% – 90%)
	30% Project/Tests/Viva	C	[70% – 80%)
	40% Semester exam	D	[60% – 70%)
		F	[0% – 60%)

This grading policy is subject to change and the final grading policy will be updated once the class starts.

Course Outcomes: The course's primary aim is to introduce students to the graphical and modeling techniques for exploring data, emphasizing visualization, interpretation, and clear communication of findings. The course will use modern software tools for data manipulation and visualization. It will build a foundation for exploring data using Python. The intention is that the student develops an intuition for statistical data analysis.

After completing the course, the student will have the ability to

- Import, manipulate, clean, visualize, and export data in Python.
- Wrangle data from its original format into a fit-for-purpose format.
- Conduct a systematic exploratory data analysis of different types of data.
- Identify, formulate and analyze complex engineering problems reaching substantiated conclusions using the first principles of mathematics. **(P2)**
- Use research methods, including design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions. **(P4)**

- Create, select, and apply appropriate techniques, resources, and IT tools, including prediction and modeling, to complex engineering activities with an understanding of the limitations. (**P5**)
- Recognize the need for and have the preparation and ability to engage in independent and life-long learning in the broadest context of technological change. (**P12**)

P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12
	X		X	X							X

Lecture Schedule

Sl. No.	Description		No. of Lectures	Lab Work
1	Introduction to Data		05	Python Basics
	1.1	Statistics and Data		Python Installation
	1.2	Collection and Classification of Data		Jupyter Notebook/ Google Colab
	1.3	Variables and Measurement Scales		Packages, Libraries
	1.4	Resources for working with data		Data types
2	Reading Data		05	Import data from different sources
	2.1	Methods of Data Collection		Reading Data from different formats
	2.2	Import Data		Data Inspection
	2.3	Reading Data		Getting insights about the dataset
	2.4	Understanding data		Get values from specific locations
	2.5	Representing Data		
3	Cleaning Data		10	Operations on data
	3.1	Examining Data		Handling & removing outliers
	3.2	Finding Outliers		Checking missing values
	3.3	Finding Missing values		looking for incompatibility
	3.4	Correcting Data		Reorder, Reformat, Reshape
	3.5	Transform Data		
4	Summarizing and Visualizing Data		15	Identify variables, types and patterns
	4.1	Describe Data		Mean, median, variance, SD, percentile
	4.2	Summarize Data		Histogram, {Box, QQ, Scatter} plots
	4.3	Visualize Data		Measuring chance under distributions
	4.4	Distribution of Data		
5	Fitting Data		06	Inspect data, specify a fit
	5.1	Introduction		Linear & multiple regression
	5.2	Regression models		Curve fitting
	5.3	Goodness of fit		
	5.4	Limitations		
6	Final project			Collect your own data from data.gov.in
	6.1	Preparation		Perform EDA on the collected dataset
	6.2	Presentation		Discuss difficulties with peer/Instructor
	6.3	Discussion		Report submission, Presentation
	6.4	Communication, Submission		