

Teaching Guidelines for
Statistical analysis and data handling using Python
ACC-HPC June 2025

Duration: 48 hours (24 theory hours + 24 lab hours)

Prerequisites: Familiarity with Python is essential (especially for NumPy, Pandas, and web scraping).

Basic Understanding of Mathematics

Evaluation: 100 marks

Weightage: CCEE – 40%, Lab exam – 40%, Internals – 20%

Textbook:

- Learn Python the Hard Way, Zed A. Shaw, Pearson
 - Statistics for Engineers and Scientists
 - All of Statistics: A Concise Course in Statistical Inference
 - Web Scraping with Python
 - Storytelling with Data: A Data Visualization Guide for Business Professionals
-

Session 1:

Lecture:

- Basics of Statistics
- Statistical Analytics
- Descriptive Statistical Measures
- Statistics - Central Tendency & Dispersion (Mean, Median, Mode, Quartiles, Percentiles, Range, Interquartile Range, Standard Deviation, Variance, and Coefficient of Variation)

Lab:

- Load any dataset and find out the mean, median, mode, and other central tendencies of the dataset.

Session 2:

Lecture:

- Probability and Distribution functions.
- Random Variables
- Central Limit Theorem.
- Cumulative Distribution Function (CDF) & Probability density functions (PDF).

Lab:

- Generate random numbers and check if they are in normal distribution using the scipy libraries.

Session 3:

Lecture:

- Inferential Statistics
- Sampling and sampling distributions
- Estimation: Point and interval estimates

- Hypothesis testing
- Types of Hypothesis Tests.
- Confidence Interval

Lab:

- Two different batches of students were taught using different methods. Their final exam scores are recorded.
- Tasks:
 - Conduct a two-sample t-test to determine if there's a significant difference in the mean scores.
 - Check if the variances are equal or unequal and choose the appropriate t-test accordingly.

Session 4:**Lecture:****Exploratory Data Analysis (EDA) using NumPy and Pandas****EDA (Exploratory Data Analysis)**

- Understanding Data Distribution
- Identify Outliers
- Relationships between Variables
- Data Cleaning

NumPy

- Array Manipulation
- Statistical Calculation
- Mathematical Operation

Pandas

- Data Wrangling
- Data Transformation
- Inspection

Lab:

- Using NumPy, perform the following tasks:
- Create a 2D array of size (5, 4) with random integers between 10 and 100.
- Calculate the mean and standard deviation for each row.
- Replace all elements greater than 50 with -1.
- Normalize the array (zero mean, unit variance).

Session 5:**Lecture:****Data Manipulation (Text, Image, Audio, Video, Webscraping)**

- Reading and manipulating text/tabular data (CSV, Excel, JSON)
- Basics of web scraping using requests and BeautifulSoup
- Introduction to handling (Images with PIL/OpenCV, Audio with Librosa or PyDub, Video with OpenCV)
- Converting and cleaning formats

Lab:

- A zip folder named media_lab_data.zip containing:
- students.json – Student data with names and grades.
- attendance.csv – Attendance records.
- college_info.xlsx – General college info.
- sample.html – A saved HTML page with event announcements.
- student_photo.jpg – A scanned student photo.
- greeting_audio.wav – A greeting audio file.
- event_clip.mp4 – A short event recording.

Tasks:**1. Text/Tabular Data**

- Load all three files (students.json, attendance.csv, and college_info.xlsx) using **Pandas**.
- Merge student grades and attendance based on a common key (like Student ID).
- Clean column names (lowercase, remove spaces).
- Export the merged file as a new CSV.

2. Web Scraping

- Using BeautifulSoup, extract all <h2> and <p> tags from sample.html.
- Display them in a readable format.

3. Image

- Load the student_photo.jpg using **PIL** or **OpenCV**.
- Convert it to grayscale and resize it to 100x100 pixels.

4. Video

- Load event_clip.mp4 using OpenCV.
- Extract the **first frame** and save it as first_frame.jpg.

Session 6:**Lecture:****Data Visualization using Power BI**

- Introduction to the Power BI interface
- Importing datasets (CSV, Excel, SQL, web)
- Creating visuals: Bar, Line, Pie, Map, etc.
- Customizing visuals and dashboards
- Filters, slicers, drill-throughs
- Publishing and sharing reports

Lab:

- You are working as a data analyst at a retail company. The company has provided you with a dataset, retail_sales.xlsx, containing the following sheets:
- Sales (Date, Region, Product Category, Sales Amount, Units Sold)

- Targets (Region, Monthly Target)
- Products (Product ID, Product Name, Category, Price)

Tasks:

1. Import Data

- Load all three sheets into Power BI.
- Establish relationships between tables where necessary (e.g., Products to Sales via Product Category).

2. Visual Creation

- Create the following visuals:
 - A bar chart showing total sales by product category.
 - A line chart of monthly sales trends.
 - A pie chart of sales distribution across regions.
 - A map visualizing total sales by region.

3. Interactivity

- Add **slicers** for:
 - Region
 - Product Category
- Enable drill-through to show detailed sales per product when clicking on a category.

4. Customization

- Format your visuals with appropriate labels, colors, and titles.
- Add KPIs for:
 - Total Sales
 - Units Sold
 - Highest-selling Product

5. Publish

- Save the Power BI file as Retail_Sales_Report.pbix.
- (Optional if Power BI Service is available) Publish it to Power BI Service and share the link.