Guidelines of DSE Semester III / B.A. Programme II Semester / GE Semester II (NEP-UGCF 2022)

Data Analysis and Visualization using Python

DSE/A2/GE2a

(Effective from Academic Year 2024-25)

	TOPICS/UNITS	Chapter	Ref
Week 1 to 3	**Unit 1 Introduction to basic statistics and analysis:** Fundamentals of Data Analysis, Statistical foundations for Data Analysis, Types of data, Descriptive Statistics, Correlation and covariance, Linear Regression, Statistical Hypothesis Generation and Testing Python Libraries: NumPy, Pandas, Matplotlib	**Ch1: pg 11-24, pg 29-35, pg 37-p38** **Ch 1: 1.3 (pg 4-6)**	**[2]** **[1]**
Week 4 to 6	**Unit 2 Array manipulation using Numpy:** NumPy array: Creating NumPy arrays, various data types of NumPy arrays Indexing and slicing, swapping axes, transposing arrays, data processing using Numpy arrays	**Ch4:4.1. Usage of rand(), randn() and randint() functions of NumPy**	**[1]**
Week 7 to 10	**Unit 3 Data Manipulation using Pandas:** Data Structures in Pandas: Series, Data Frame, Index objects, loading data into Panda's data frame, Working with Data Frames: Arithmetics, Statistics, Binning, Indexing, Reindexing, Filtering, Handling missing data, Hierarchical indexing, Data wrangling: Data cleaning, transforming, merging and reshaping	**Ch 5: 5.1, 5.2 excluding Arithmetic and data alignment, axis indexes with duplicate labels, 5.3** **Ch 6: 6.1 (pg 177-181,184)** **Ch 7: 7.1, 7.2 till binning (pg 203-217)** **Ch 8: 8.1 (pg 247-253), 8.2 (pg 253-258) 8.3 (pg 270-273)**	**[1]**
Week 11 to 13	**Unit 4 Plotting and Visualization:** Using Matplotlib to plot data: figures, subplots, markings, color and line styles, labels and legends, Plotting functions in Pandas: Lines, bar, Scatter plots, histograms, stacked bars, Heatmap, 3D Plotting, interactive plotting using Bokeh and Plotly	**Ch 9: 9.1 (pg 281-296), 9.2 (pg 298-313), 9.3** **Ch 5 : pg 281-282**	**[1]** **[2]**
Week 14 to 15	**Data Aggregation and Group operations:** Group by mechanics, Data aggregation, General split-apply-combine, Pivot tables and cross tabulation	**Chapter 10: 10.1, 10.2, 10.3 (till pg 337), 10.5**	**[1]**

Essential/recommended readings

1. McKinney W. *Python for Data Analysis: Data Wrangling with Pandas*, *NumPy and IPython*. 3rd edition. O'Reilly Media, 2022

2. Molin S. *Hands-On Data Analysis with Pandas*, Packt Publishing, 2019.
3. Gupta S.C., Kapoor V.K., *Fundamentals of Mathematical Statistics*, Sultan Chand & Sons, 2020.

Suggested Practical List (If any): (30 Hours)

Practical exercises such as
Use a dataset of your choice from Open Data Portal (https:// data.gov.in/, UCI repository) or load from scikit, seaborn library for the following exercises to practice the concepts learnt.

1. Load a Pandas dataframe with a selected dataset. Identify and count the missing values in a dataframe. Clean the data after removing noise as follows
   a) Drop duplicate rows.
   b) Detect the outliers and remove the rows having more than two outliers identified using boxplot.
   c) Identify the most correlated positively correlated attributes and negatively correlated attributes

2. Import iris data using sklearn library or (Download IRIS data from: https://archive.ics.uci.edu/ml/datasets/iris or import it from sklearn.datasets)
a. Compute mean, mode, median, standard deviation, confidence interval and standard error for each feature
b. Compute correlation coefficients between each pair of features and plot heatmap
c. Find covariance between length of sepal and petal iv. Build contingency table for class feature

3. Load Titanic data from sklearn library , plot the following with proper legend and axis labels:
a. Plot bar chart to show the frequency of survivors and non-survivors for male and female passengers separately
b. Draw a scatter plot for any two selected features
c. Compare density distribution for features age and passenger fare
d. Use a pair plot to show pairwise bivariate distribution

4. Using Titanic dataset, do the following
a. Find total number of passengers with age less than 30
b. Find total fare paid by passengers of first class
c. Compare number of survivors of each passenger class

5. Download any dataset and do the following
a. Count number of categorical and numeric features
b. Remove one correlated attribute (if any)
c. Display five-number summary of each attribute and show it visually

Project: Students are encouraged to work on a good dataset in consultation with their faculty and apply the concepts learned in the course.

Additional Practice Exercises:

1. Write programs in Python using NumPy library to do the following:
    a. Compute the mean, standard deviation, and variance of a two dimensional random integer array along the second axis.
    b. Create a 2-dimensional array of size m x n integer elements, also print the shape, type and data type of the array and then reshape it into an n x m array, where n and m are user inputs given at the run time.
    c. Test whether the elements of a given 1D array are zero, non-zero and NaN. Record the indices of these elements in three separate arrays.
    d. Create three random arrays of the same size: Array1, Array2 and Array3. Subtract Array 2 from Array3 and store in Array4. Create another array Array5 having two times the values in Array1. Find Co-variance and Correlation of Array1 with Array4 and Array5 respectively.
    e. Create two random arrays of the same size 10: Array1, and Array2. Find the sum of the first half of both the arrays and product of the second half of both the arrays.

2. Consider two data files (in CSV format) having attendance of two workshops. Each file has three fields 'Name', 'Date, duration (in minutes) where names are unique within a file. Note that duration may take one of three values (30, 40, 50) only. Import the data into two data frames and do the following:
    a. Perform merging of the two data frames to find the names of students who had attended both workshops.
    b. Find names of all students who have attended a single workshop only.
    c. Merge two data frames row-wise and find the total number of records in the data frame.
    d. Merge two data frames row-wise and use two columns viz. names and dates as multi-row indexes. Generate descriptive statistics for this hierarchical data frame.

3. Consider the following data frame containing a family name, gender of the family member and her/his monthly income in each record.

Name	Gender	MonthlyIncome (Rs.)
Shah	Male	114000.00
Vats	Male	65000.00
Vats	Female	43150.00
Kumar	Female	69500.00
Vats	Female	155000.00
Kumar	Male	103000.00
Shah	Male	55000.00
Shah	Female	112400.00
Kumar	Female	81030.00
Vats	Male	71900.00

Write a program in Python using Pandas to perform the following:
    a. Calculate and display familywise gross monthly income.
    b. Display the highest and lowest monthly income for each family name
    c. Calculate and display monthly income of all members earning income less than Rs. 80000.00.
    d. Display total number of females along with their average monthly income
    e. Delete rows with Monthly income less than the average income of all members