DISCIPLINE SPECIFIC ELECTIVE COURSE: Data Analysis and Visualization
Sem-III (Bsc(H) CS) UGCF

	TOPICS/UNITS	Chapter	Ref
Week 1 to 2	**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 6-29, pg 32-33** **Ch 1: 1.3 (pg 4-6)**	**[2]** **[1]**
Week 3 to 5	**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	**Ch4:4.1-4.2, Usage of rand(), randn() and randint() functions of NumPy**	**[1]**
Week 6 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 excluding JSON data and XML data,, 6.2 Reading Microsoft Excel files only** **Chapter 7 : 7.1, 7.2 till Detection and Filtering Outliers,7.3 till String object methods** **Chapter 8 : 8.1, 8.2 exclude combining data with overlap, 8.3 till Reshaping with Hierarchical Indexing**	**[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	**Chapter 9 : 9.1, 9.2 excluding Facet Grids and Categorical Data** **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 excluding example Group wise Linear Regression, 10.4**	**[1]**

References

1. McKinney W. *Python for Data Analysis: Data Wrangling with Pandas*, *NumPy and IPython*. 2nd edition. O'Reilly Media, 2018..
2. Molin S. *Hands-On Data Analysis with Pandas*, Packt Publishing, Second Edition, 2021.
3. Gupta S.C., Kapoor V.K., *Fundamentals of Mathematical Statistics*, Sultan Chand & Sons, 2020.

Suggested Practical List For Data Analysis and Visualization DSE Sem III

Note:

- Any platform for Python can be used for lab exercises
- Use a data set 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. Write programs in Python using NumPy library to do the following:
    a. Create a two dimensional array, ARR1 having random values from 0 to 1. Compute the mean, standard deviation, and variance of ARR1 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.
    f. Create an array with random values. Determine the size of the memory occupied by the array.
    g. Create a 2-dimensional array of size m x n having integer elements in the range (10,100). Write statements to swap any two rows, reverse a specified column and store updated array in another variable

2. Do the following using PANDAS Series:
    a. Create a series with 5 elements. Display the series sorted on index and also sorted on values seperately
    b. Create a series with N elements with some duplicate values. Find the minimum and maximum ranks assigned to the values using 'first' and 'max' methods
    c. Display the index value of the minimum and maximum element of a Series

3. Create a data frame having at least 3 columns and 50 rows to store numeric data generated using a random function. Replace 10% of the values by null values whose index positions are generated using random function. Do the following:
    a. Identify and count missing values in a data frame.
    b. Drop the column having more than 5 null values.
    c. Identify the row label having maximum of the sum of all values in a row and drop that row.
    d. Sort the data frame on the basis of the first column.
    e. Remove all duplicates from the first column.
    f. Find the correlation between first and second column and covariance between second and third column.
    g. Discretize the second column and create 5 bins.

4. Consider two excel files having attendance of two workshops, each of duration 5 days. Each file has three fields 'Name', 'Date, duration (in minutes) where names may be repetitve 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.

5. Using Iris data, plot the following with proper legend and axis labels: (Download IRIS data from: https://archive.ics.uci.edu/ml/datasets/iris or import it from sklearn datasets)
   a. Load data into pandas' data frame. Use pandas.info () method to look at the info on datatypes in the dataset.
   b. Find the number of missing values in each column (Check number of null values in a column using df.isnull().sum())
   c. Plot bar chart to show the frequency of each class label in the data.
   d. Draw a scatter plot for Petal Length vs Sepal Length and fit a regression line
   e. Plot density distribution for feature Petal width.
   f. Use a pair plot to show pairwise bivariate distribution in the Iris Dataset.
   g. Draw heatmap for any two numeric attributes
   h. Compute mean, mode, median, standard deviation, confidence interval and standard error for each numeric feature
   i. Compute correlation coefficients between each pair of features and plot heatmap

6. Using Titanic dataset, to do the following:
   a. Clean the data by dropping the column which has the largest number of missing values.
   b. Find total number of passengers with age more than 30
   c. Find total fare paid by passengers of second class
   d. Compare number of survivors of each passenger class
   e. Compute descriptive statistics for age attribute gender wise
   f. Draw a scatter plot for passenger fare paid by Female and Male passengers separately
   g. Compare density distribution for features age and passenger fare
   h. Draw the pie chart for three groups labelled as class 1, class 2, class 3 respectively displayed in different colours. The occurrence of each group converted into percentage should be displayed in the pie chart. Appropriately Label the chart.
   i. Find % of survived passengers for each class and answer the question "Did class play a role in survival?".

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

FamilyName	Gender	MonthlyIncome (Rs.)
Shah	Male	44000.00
Vats	Male	65000.00
Vats	Female	43150.00
Kumar	Female	66500.00
Vats	Female	255000.00
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

Project : Students are required to work on a good dataset in consultation with their faculty and apply the concepts learned in the course. Each project must include the following:

i. Download a dataset (Either web-scrap the data from various data sources like twitter, amazon, news sites or download a dataset from kaggle/UCI/data.gov.in etc.). Select a dataset which requires at least two steps of data cleaning and two steps of data pre-processing.
ii. Make an objective of the data analysis for that dataset. Depending on the dataset, perform the data cleaning, data pre-processing steps.
iii. The data cleaning steps may include handling of missing values, handling duplicate data, handling inconsistent data (e.g. height is given in feet for some objects and in inches in some other objects), removing redundant data (e.g. in some datasets, age and date of birth are given as two column while the analysis needs only the age), handling incomplete data (e.g. email address doesn't have @ symbol).
iv. The pre-processing steps may include transforming the data to some other format (e.g text comment converted to term vector, image files converted to various types of features), discretization and binning, standardization/normalization and outlier detection. Some string manipulations may also be required.
v. Prepare at least eight analysis questions for exploration of the data and at least two questions for the visualization of the data.

Prepared By:

1. Dr. Anamika Gupta (SSCBS) 2. Prof Arpita Sharma (DDUC) 3. Prof Hema Banati (DSC) 4. Prof. Sharanjit Kaur (ANDC)