B. Sc. (H) Computer Science Semester V

BHCS15A: Data Analysis and Visualization

Unit No.	Chapters	Reference	No. of Lectures
1	Chapter 1 (1.1, 1.2, 1.3) Chapter 2 (2.1, 2.3) Chapter 3 (3.1, 3.2)	1	8
	Chapter 2 (Datasets mentioned in Exercise: EDA pg. 37, 38 can be considered for possible project work, *pg. 41, 42*)	2	
2	Chapter 4 (4.1) Chapter 5 (5.1, 5.2 excluding Arithmetic and data alignment, 5.3) Chapter 6 (6.1 excluding JSON data and XML data, 6.2 *Reading Microsoft Excel files only*, 6.3,6.4) Chapter 7 (7.1, 7.2 till *Detection and Filtering Outliers*,7.3 till *String object methods*)	1	20
3	Chapter 8 (8.1, 8.2 Exclude combining data with overlap, 8.3 till Reshaping with Hierarchical Indexing) Chapter 9 (9.1, 9.2 Excluding Facet Grids and Categorical Data)	1	14
4	Chapter 10 (10.1, 10.2, 10.3 excluding example Group wise Linear Regression,10.4), Chapter 11 (11.1, 11.2, 11.3, 11.4 only Introduction to Timezone handling page 335, 11.5 till Period Frequency conversion page 342, 11.6, 11.7 before subsection Exponentially Weighted Functions page 358)	1	14
5	Chapter 12 (12.1 excluding Computations with categorical, 12.2 excluding Grouped Time resampling, 12.3 Excluding Pipe Method)	1	4

References

1. McKinney, W. (2017). Python for Data Analysis: Data Wrangling with Pandas, NumPy and IPython. 2nd edition. O'Reilly Media.
2. O'Neil, C., & Schutt, R. (2013). Doing Data Science: Straight Talk from the Frontline O'Reilly Media

Practical List

(Note: Any platform for Python can be used for lab exercises)

1. Given below is a dictionary having two keys ‘Boys’ and ‘Girls’ and having two lists of heights of five Boys and Five Girls respectively as values associated with these keys

   Original dictionary of lists:

   {'Boys': [72, 68, 70, 69, 74], 'Girls': [63, 65, 69, 62, 61]}

   From the given dictionary of lists create the following list of dictionaries:

   [{'Boys': 72, 'Girls': 63}, {'Boys': 68, 'Girls': 65}, {'Boys': 70, 'Girls': 69}, {'Boys': 69, 'Girls': 62}, {‘Boys’:74, ‘Girls’:61]

2. 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. Get the indices of the sorted elements of a given array.
      a. B = [56, 48, 22, 41, 78, 91, 24, 46, 8, 33]
   c. 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 nx m array, n and m are user inputs given at the run time.
   d. Test whether the elements of a given array are zero, non-zero and NaN. Record the indices of these elements in three separate arrays.

3. Create a dataframe 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 dataframe.
   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 dataframe 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. Detect the outliers and remove the rows having outliers.
   h. Discretize second column and create 5 bins

4. Consider two excel files having attendance of a workshop’s participants for two days. Each file has three fields ‘Name’, ‘Time of joining’, 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 dataframes and do the following:

   a. Perform merging of the two dataframes to find the names of students who had attended the workshop on both days.
   b. Find names of all students who have attended workshop on either of the days.
   c. Merge two data frames row-wise and find the total number of records in the data frame.
   d. Merge two data frames and use two columns names and duration as multi-row indexes. Generate descriptive statistics for this multi-index.

5. Taking 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. Plot bar chart to show the frequency of each class label in the data.
b. Draw a scatter plot for Petal width vs sepal width.
c. Plot density distribution for feature petal length.
d. Use a pair plot to show pairwise bivariate distribution in the Iris Dataset.

6. Consider any sales training/ weather forecasting dataset

a. Compute mean of a series grouped by another series
b. Fill an intermittent time series to replace all missing dates with values of previous non-missing date.
c. Perform appropriate year-month string to dates conversion.
d. Split a dataset to group by two columns and then sort the aggregated results within the groups.
e. Split a given dataframe into groups with bin counts.

7. Consider a data frame containing data about students i.e. name, gender and passing division:

	Name	Birth_Month	Gender	Pass_Division
0	Mudit Chauhan	December	M	III
1	Seema Chopra	January	F	II
2	Rani Gupta	March	F	I
3	Aditya Narayan	October	M	I
4	Sanjeev Sahni	February	M	II
5	Prakash Kumar	December	M	III
6	Ritu Agarwal	September	F	I
7	Akshay Goel	August	M	I
8	Meeta Kulkarni	July	F	II
9	Preeti Ahuja	November	F	II
10	Sunil Das Gupta	April	M	III
11	Sonali Sapre	January	F	I
12	Rashmi Talwar	June	F	III
13	Ashish Dubey	May	M	II
14	Kiran Sharma	February	F	II
15	Sameer Bansal	October	M	I

a. Perform one hot encoding of the last two columns of categorical data using the get_dummies() function.
b. Sort this data frame on the “Birth Month” column (i.e. January to December). Hint: Convert Month to Categorical.

8. 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. Calculate and display the member with the highest monthly income in a family.
c. Calculate and display monthly income of all members with income greater than Rs. 60000.00.
d. Calculate and display the average monthly income of the female members in the Shah family.

The students are encouraged to work on a good dataset in consultation with their faculty and apply the concepts learned in the course. Datasets mentioned in Ref 2, chapter 2 pg 37,38 may be consulted.The following is a sample of the kind of work expected in the project.

Sample Project

Download Covid_19_India Dataset named **"covid_19_india.csv"** from https://www.kaggle.com/sudalairajkumar/covid19-in-india?select=covid_19_india.csv and perform the following, with proper annotations of the legend and axes labels:

Use **covid_19_india.csv** to do the following:

- For each Indian state, find maximum cases reported for confirmed, deaths and recovered individually along with date on which these cases were reported for any three months of year 2020. Display the result in the self-explanatory format.
- Use appropriate year-month string date conversions for example: Identify the no. of cases on the 6th day of the month by converting year-month string to dates.
- Create subplots (line graph) for showing total number of cured cases month-wise from April 2020 to March 2021 in four states namely Karnataka, Gujarat, Haryana, and Uttar Pradesh.
- Compare the deaths due to Covid-19 in the months of May 2020 and May 2021 for the states namely Karnataka, Delhi, and Madhya Pradesh using stacked bars.
- Make a graph to show the month wise relation (Positive/Negative/Neutral) between number of confirmed Covid-19 cases and Deaths in Uttar Pradesh. Display correlation value too in the graph.