# BSc(P) CSc-Data Analysis using Python Programming (SEC-1A)

(Guidelines, August 2020)

| UNIT | Chapter | References | No. of Lectures |
|---|---|---|---|
| **UNIT I - Introduction to Pandas, NumPy, SciPy:**<br>Introduction to Pandas DataFrames, Numpy multi-dimentional arrays, and SciPy libraries to work with different datasets | **1.3**<br>**5.1-5.2**<br>4.1 to 4.4 | **1** | **7** |
| **UNIT II - Import and Export of Data:**Installing, loading and using packages for importing and exporting data in Python | 6.1 upto page no. 176 (excluding Tables 6.1 and 6.2) | 1 | **6** |
| **UNIT III - Data Preprocessing and Transformation:**<br>Handling of missing data, Data cleaning and transformation | 7 (upto page No. 213) | **1** | **5** |
| **UNIT IV - Data Exploration**<br>Exploring data using statistical methods: mean, median, mode[1], quantiles. Building contingency table 2. Basics of grouping data and Correlation. | **5.3**<br>10.1 (upto page 293)<br>[1]use mode() | **1** | **6** |
| **Unit 5 - Data Visualization:**<br>Scatter Plot, line graph, histogram, boxplot, line plots regression, word clouds[2], exporting plots as images. | 9.1-9.2<br>[2] use wordcloud package | **1** | **4** |

**Text book:**
1. Mckinney, W. (2017). Python for Data Analysis. Second edition, O’reilly (SPD).

Additional Resources
2. Grus, J. (2016). Data Science from scratch. First edition, O’reilly (SPD).
3. VanderPlas, J. (2016). Python Data Science Handbook: Essential Tools for Working with Data. Second edition, O’reilly (SPD).

1 Mode: use mode function of pandas (https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.mode.html)
2 Contigency table using crosstab function : use crosstab function https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.crosstab.html

Links for Examples on Word clouds:
https://www.datacamp.com/community/tutorials/wordcloud-python
https://www.tutorialspoint.com/create-word-cloud-using-python
https://www.geeksforgeeks.org/generating-word-cloud-python/

Links for Examples on Contigency table:
https://www.geeksforgeeks.org/contingency-table-in-python/
https://www.tutorialspoint.com/contingency-table-in-python

Additional daratsets for practice: Chapter 14: Data analysis on datasets [1]

**Specimen list for practicals:**
Use data set of your choice from Open Data Portal (https://data.gov.in/) for the following exercises, wherever datasets are not mentioned explicitly.

1. Make visual representations of data using libraray Matplotlib and apply basic principles of data graphics to create rich analytic graphs for available datasets.

2. Use boston house-prices dataset avaiable with sklearn library to do the following for:
   i. Generate box whisker plots for price and age of the owner
   ii. Identify outliers, if any
   iii. Display 5 point summary of data distribution for all attributes
   iv. Find if there is any missing value in data or not
   v. Find pairwise correlation between attributes
   vi. Use scatterplot to show relationship between each feature w.r.t target class in a single panel for comparison

3. Create a CSV file having employee data records. Each employee record has three features viz. age, home city and salary. Import employee file and :
   i. Draw scatter plot for age vs salary
   ii. Plot histogram for features age and salary
   iii. Plot Pie chart for the qualitative attribute city
   iv. Generate box plots for salary and age

4. Import iris data using sklearn library to:
   i. Compute mean, mode, median, standard deviation, confidence interval and standard error for each feature
   ii. Compute correlation between length and width of sepal feature
   iii. Find covariance between length of sepal and petal
   iv. Build contingency table for class feature

5. Download datasets Hepatitis and automobile from UCI repository
   i. Find the number of records which are noise free
   ii. Clean data after removing noise
   iii. Normalize quantitative features in range of [0,1]
   iv. Compare frequency distribution for any two columns by plotting histograms for any two columns in the same plot
6. Do the following using iris CSV file (use of Pandas/NumPy/SciPy)
   i. Find total number of records and columns in a csv file
   ii. Find correlation and contingency table for any two variables
   iii. Find the coulmn with maximum variance
   iv. Draw scatter plot for any two columns and also write their correlation in the caption of scatter plot

7. Use car dataset from UCI repository (https://archive.ics.uci.edu/ml/machine-learning-databases/car/)
   i. Find the most popular car and draw appropriate plot to justify your answer
   ii. Plot barchart to compare capacity of any two cars alongwith their cost
   iii. Draw word cloud for car names and export to a file