Lab Assignment 2 – Statistical Analysis and Pandas (Part 1)

Objectives:

1. Explore Pandas
    a. Importing data
    b. Dataframes
    c. Indexing, data selection and subsetting
    d. Sorting and ranking
    e. Series
2. Explore statistical methods

STEPS:

1. **Working with sheets:** Download retail data from https://archive.ics.uci.edu/ml/datasets/online+retail and perform:
    a. **Import** complete data to pandas dataframe
        i. **Indexing**: Create subset of data frame containing rows where value is "536544" in voice field and containing first 4 columns only.
    b. **Import**: Read from sheet, only stock code and price
        i. **Statistical analysis**: Apply simple statistics on the "price" column. Like, find min, max, median, mean, count, sum, standard deviation, and variance.
    c. **Import** sheet, by skipping first 100 rows.
        i. **Sort** the records by the "InvoiceDate" column
        ii. Find details where " Description " starts with letter "D", You can use Map Function
2. **Creating a dataframe (pandas):**
    a. Create a student database with 5 entries, created randomly. List of features would be "student name","date of birth", "marks in maths", store the information in pandas dataframe. Create this dataframe from
        i. lists of lists.
        ii. dict of narray/lists
        iii. list of dicts
        iv. dict of series
    b. Create an empty DataFrame with columns name only then appending rows one by one to it using append() method. Empty data frame should be having 3 columns "student name","date of birth", "marks in english". Append one by one 5 random rows to this dataframe.
3. **Working with series:**
    a. Create two panda series of 100 random numbers
    b. Write a Pandas program to add, subtract, multiply and divide two Pandas Series
    c. Store all the 6 series to a dataframe with column names as "number 1" "number 2", "add", "sub", "mult" and "div"

Reference: https://pandas.pydata.org/docs/user_guide/index.html

Suggested Platform: Python: Azure Notebook/Google Colab Notebook, packages such as pandas >= 0.11.1 and its dependencies

NumPy >= 1.6.1
stats/statistics
Other packages that you can explore: pytz, IPython >= 0.12, pyzmq, tornado, statsmodels, xlrd and openpyxl, matplotlib >= 1.0.0

Submission: Submit your files (Both Part1 and Part 2) in Single ipython Notebook in LMS before Sunday 23 Aug, 11.59 pm.

Marking: Marking is based on both performance during the lab hours as well as complete submission in LMS.

Important Note: Please feel free to think out of the box by exploring all the possibilities in the web. Objective of any assignment is only to improve your learning experience, not just about getting output!