Lab Assignment 2 – Part 2 Statistical Analysis and Pandas in Python

1. Explore Pandas
    a. Missing data
    b. Data summarization
2. Data Wrangling with Pandas
    a. Types
    b. Merging and joining DataFrame objects
    c. Concatenation
    d. Reshaping and Pivoting
    e. Data transformation
    f. Permutation and sampling
    g. Data aggregation and GroupBy operations
3. Statistical Data Modeling
    a. Statistical modeling
    b. Fitting regression models
    c. Model selection

Steps:

Part a) Data Wrangling:

1. **Dataset**: Download any the dataset of your choice from online and select appropiate x and y label values. Dataset should contain more than 100 samples. Both x and y values should be numerical. You can also use **https://tinyurl.com/y5ux9ukz** data
2. **Missing Data:** Check for missing data in all the cells, print True where you find the missing data information otherwise print False.
3. **Categorize:** Create a new list/array/dataframe containing type corresponding to sales, name this as "sales type". If sales is >=15, set value of sales type to "high", if sales is <15 and >=5, set value of sales type to "medium", else set sales to "low". (If you chose data of your choice, create a new category with respect to your data and follow this in further steps as well.)
4. **Concatenation:** Concatenate the "sales type" column to the original downloaded dataset.
5. **Groupby and Aggregation:** Aggregate for each column, grouped by "sales type column", perform sum and average on "TV", "newspaper" and "radio".
6. **Reshaping and Pivoting:** Explore these two on pandas and do 1-2 sample examples.

Part b) Data Modeling:

7. **Dataset**: Use the original downloaded dataset and perform the following: Select every 6^{th} row of column "TV" and set it to NaN/NULL and select every 10^{th} row of column "Newspaper" and set it to NaN/NULL. Print only rows containing no Nan/NULL value.
8. **Fill in missing values:** There are a variety of methods which can be used to fill in the missing values (Recently Created). Explore them
    a. Fill missing(NaN/NULL) values with 0
    b. Fill missing(NaN/NULL) values with average of column

9. **Normalization**: Each column can be normalized into 0 to 1 using different methods such as scaling, standardizing etc. This part you have to explore. Note: Normalization should not be done for the target feature(which is sales here).
10. **Data Splitting**: After the range normalization, its time to split the data into training and testing. Split the data and keep any randomly selected rows 40% samples for testing and rest of them for training.
11. **Linear Regression Training**: Now we have to train this model based on gradient descent algorithm. Train model over linear regression. Read the training samples one by one. Compute the error for each training sample, update the the parameters. Train the model for many iteration in the training dataset. Plot the error of the model as a line graph for training iterations. (You can use sklearn for regression)
12. **Testing error**: Apply the model with the test data and compute the mean squared error (MSE) for test data. Also use other metrics like R2 score to calculate the error on the testing data.
13. **Playing with the Model**: You can try different strategies to see whether testing error comes down or not. Strategies can be different 1. Initialization of parameters like learning rate etc, 2. Different normalization methods, 3. Shuffling of training samples, 4, handling of missing values. Check the model error for the testing data for each setup.

Suggested Platform: Python: Azure Notebook/Google Colab Notebook, packages such as
pandas >= 0.11.1 and its dependencies
NumPy >= 1.6.1
matplotlib >= 1.0.0
statistics
sklearn
Other packages that you can explore: pytz, IPython >= 0.12, pyzmq, tornado, statsmodels, xlrd and openpyxl

Submission: Submit your files in Single ipython Notebook in LMS before Sunday 30th Aug, 11.59 pm.

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!