Course	DS 203: Programming for Data Science
Activity	Type: Exercise Title: E03 – Simple Linear Regression using the Python Programming Language
Background	• The previous exercises have introduced various SLR concepts using Excel as a tool • The Python programming environment provides rich features to do all that spread-sheets enable – and much more. • This exercise is part of acquiring Python programming skills in the context of solving Data Science problems
Expected outcomes of this exercise	• Get familiar with the Python packages Numpy and Matplotlib • Start solving data problems using arrays and their compact representations, backed by powerful and efficient operations. • Start using Matplotlib functions for data visualization
Tools	• Python Notebook using either Jupyter or VSC
Effort estimate	• **3 hours**
Submission type	• **Mandatory submission**
Due date and time	• **24th August, 23:55 Hrs**
Submission instruction	• Submit to the appropriate Moodle submission point • Your final submission should be in the form of a single PDF file. • Your solutions and answers should include explanations / graphs / charts / Tables that are necessary to fully explain the solution(s). • **Complete this exercise and submit well within time. No extensions will be granted, no email submissions will be accepted.**
Marks for the exercise	• Credit will be given for **complete and timely** submission to Moodle • The exercise itself will not carry marks. • **Your understanding and skill – expected to be gained by completing this exercise – will be gauged in a quiz, test, or viva that will be conducted subsequently.**
References	• Python Notebooks uploaded to course page on Moodle. Especially units 5 and 6. • Lecture notes • Excel spread-sheets uploaded to Moodle • Help documentation for Python, Numpy and Matplotlib. • Articles, blogs and other sources

Note:

- This exercise requires skills related to Numpy and Matplotlib libraries. Peruse the Python Notebooks 5 (Numpy) and 6 (Matplotlib) uploaded to the course page on Moodle for an introduction.

Create a Python Notebook to program and complete the following tasks (estimated effort: 3 hours):

1. Load the data file **linear-data-set-for-regression.csv** into a numpy array
2. Create a scatter plot of the data using Python matplotlib functions
3. Using numpy functions calculate the regression coefficients a, b for this data set
4. Create an array of predicted values based on the regression model
5. Create an array of prediction errors
    a) There exists a function '**scipy.stats.normaltest'.** Understand this function and use the p-value returned by it to check if the prediction errors follow normal distribution. Print the results.
6. Calculate SST, SSR, SSE and R2 values
7. Validate the expression: SST = SSR + SSE
8. Create a plot with two sub-plots that have aligned horizontal scales:
    a) Scatter plot of original data + predicted points
    b) Error plot
9. Create a nicely formatted output text file with all the metrics you have created. Open this file in a text editor, capture a screen-shot and add it to the Notebook.
10. Compare all the numbers with those in the file **linear-data-set-for-regression-soln-aug-11.xlsx** and ensure that the outputs of this exercise match with them! Mention the result of this check in the Notebook.

Convert the Notebook into PDF (your program segments and all the generated / added outputs) and upload it as your submission for this assignment.

*****