Course	DS 203: Programming for Data Science
Activity	Type: Exercise Title: E05 – MLR and Backward Elimination using Python
Background	• Concepts of MLR and Feature Engineering / Feature Selection have already been covered in the class (see : **DS203-2023-08-18-MLR-of-Non-Linear-Y.pdf**) • Regression problems often involve tens or hundreds of variables. Not all of these independent variables are significant. The Feature Eliminatio process ensures that only the most significant independent variables are included in the final model. Backward Elimination involves starting with all the independent variables and then sequentially eliminating some of them to progressively improve the model. • A Python Notebook outlining the use of OLS for MLR has been uploaded. See: **MLR-using-OLS.ipynb**
Expected outcomes of this exercise	• Understand the Backward Elimination process. • **Get used to the iterative nature of solving Data Science problems!**
Tools	• Python Notebook using either Jupyter or VSC
Effort estimate	• **3 hours**
Submission type	• **Mandatory submission**
Due date and time	• **September 2, 2023, 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. • Lecture notes • Excel spread-sheets uploaded to Moodle • Help documentation for Python, Numpy and Matplotlib. • Articles, blogs and other sources
NOTE	• **The PDF should have your roll number as the filename. You may additionally include your name. No credit will be given if this is not followed!**

Prerequisite:

- Review the uploaded Python Notebook **MLR-using-OLS.ipynb** and understand the steps
- Review the documentation for the **OLS** function defined in the **statsmodels** package

The data set and the problem:

- This exercise refers to the data set in **MLR-Feature-Elimination.csv**
- This data set is a small part of the data acquired by a data acquisition system of a factory involved in the manufacture of chemicals. The data has already gone through the visualization and cleaning steps – so it can be considered 'good data'
- The data set contains 41 columns and 1025 rows (observations)
- The column c52 is the output parameter that needs to be predicted
- All other columns, **except** c1, c2 and c241 are deemed to be relevant process parameters (such as flow, temperature, pressure, etc.). Many of them can be controlled to impact the output.
- It is required to create a regression model to predict c52, **and to understand the relative importance of the variables on the output**.

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

1. Using the above data, and the technique outlined in **MLR-using-OLS.ipynb** set up the code for MLR.
2. Select y and X based on the information given above.
3. Carry out MLR and review the regression parameters. Progressively use the backward elimination process outlined in **DS203-2023-08-18-MLR-of-Non-Linear-Y.pdf** to drop inappropriate variables from the model. At every step, in a Table, keep track of the variable dropped and the resulting R2 and MSE values
4. Once you have the final model, study the coefficients. Based on their values can you identify the independent variables that potentially have a large impact on the output – in terms of either an increase or decrease of the output? Identify the most important variables and document them, stating your reasons for selecting them.
5. Submit your final MLR model, the Table, and your analysis of the coefficients.

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

*****