Assignment on regression (If you did not do the previous assignment, please do it first, then complete this assignment):

1- One change in the previous assignment: instead of label encoding of the ordinal features, use ordinal encoding in Sklearn; the main difference is that it takes a list of categories with their logical order.

For example: if your sub-categories are:

a. Bad
b. Good
c. Excellent

You need to put these categories in a list with their logical order from the least sub-category to the highest sub-category as follows:

['Bad', 'Good', 'Excellent']

Then put this list as an input to the ordinal encoder in Sklearn

2- Instead of using the training data only, please use the training and test data in all the pre-processing steps as follows:

a. In all the pre-processing steps, transform the test-data using the transformation fitted on the training data.
But please note that: we never fit the transformation on the test data; we transform the test data using the fitted values of the training data.
For example: you standardized your training data using the mean and std of the training data. When you standardize the test data, you need to use the mean and std of the training data as well.
[so please modify your automated version of the pre-processing step to handle the training and test data]

3- In the modeling part, use the train and test data in evaluating the performance of these 4 models:

a. Linear regression (that used all the features coming from the pre-processing step, without any subset selection)
b. Linear regression (that used the features with no multi-collinearity, and filtered from the insignificant features using the backward stepwise subset selection; and remember to remove these features from the test data as well before evaluating this model)
c. Ridge and Lasso regression (that used all the features coming from the pre-processing step, without any subset selection), but we need to select the best lambda parameter (regularization parameter) for each model, by making a for loop that tries a range of lambda [0, 1, 0.1] between 0 and 1, with step 0.1.

Note that:

i. The best lambda parameter for lasso and ridge is the one that resulted in the least test MSE
ii. The evaluation metrics that need to be used in every step (a, b, c) are those which were taken in the lecture:
MSE, RMSE, MAE, MAPE, RAE, R-2, R-2 adjusted

4- Check the assumptions of linear regression as we done in the lab, but for the normality test, (beside the QQ-plot and the histogram), please use the Shapiro-wilk test with significance (alpha = 0.01) to decide whether the normality of residuals is fulfilled or not

5- Try to perform basis expansion of the input features with polynomial degree = 2, and repeat all the previous steps (step 3, and step 4) again.

6- Make a summary that tells us what the best model among all the models is.
(The best model is the one that resulted in the best R2 adjusted)