**Question 1**

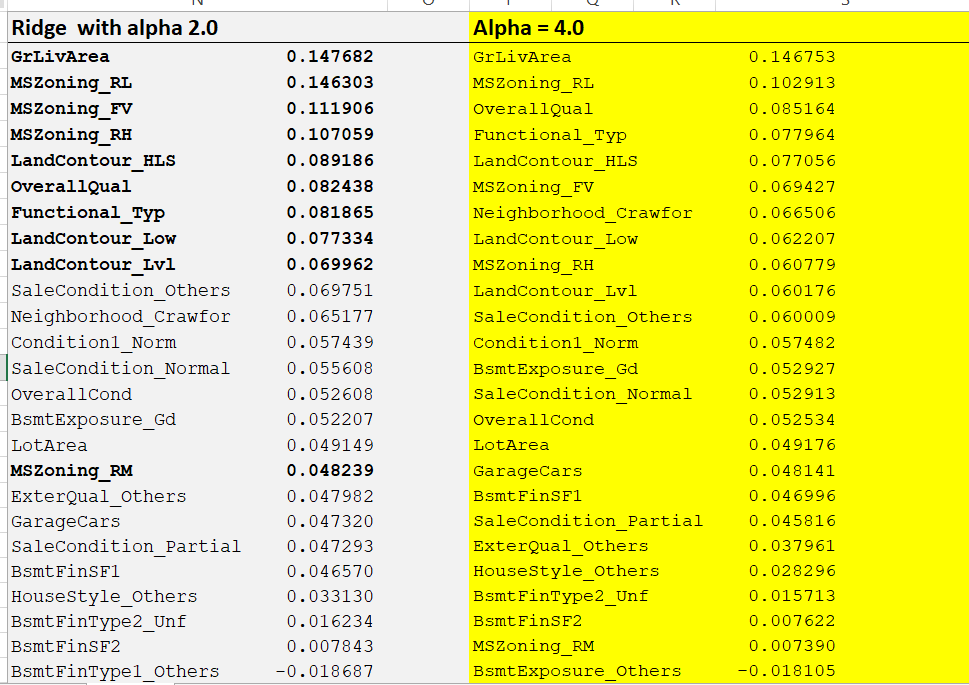
What is the optimal value of alpha for ridge and lasso regression? What will be the changes in the model if you choose double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented.

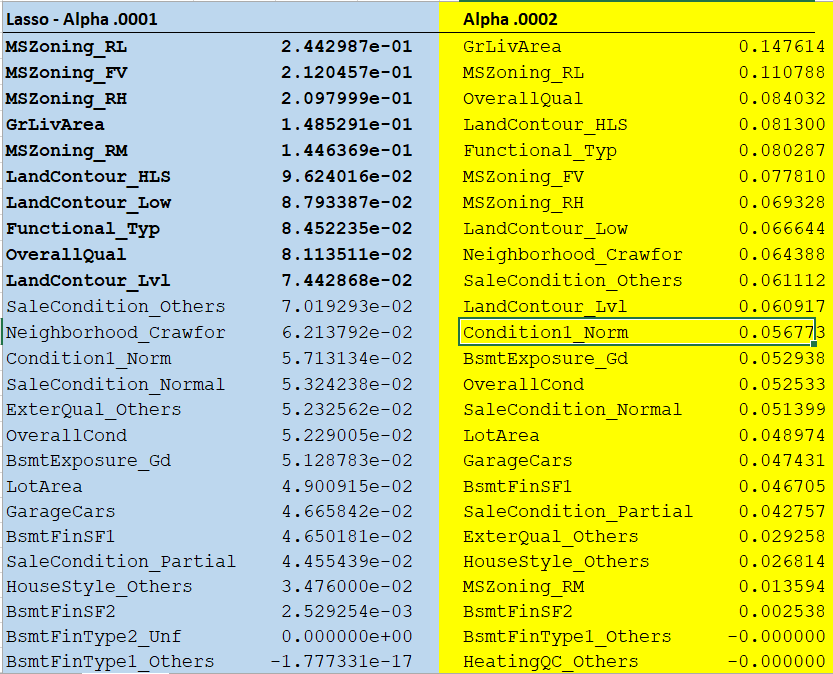
**Answer:**

* Optimal value of alpha for lasso regression = 0.0001
* Optimal value of alpha for ridge regression = 2.0

Post the change the optimal alpha value we see that the minor change in R2 values across train and test sets. **Bias** seems to have been introduced in the model.

Also with the increase in the alpha values the coefficients values have gone down (minimise the penalty term). Coefficients (not all) Pre and Post change in alpha has been indicated in the snapshot below.



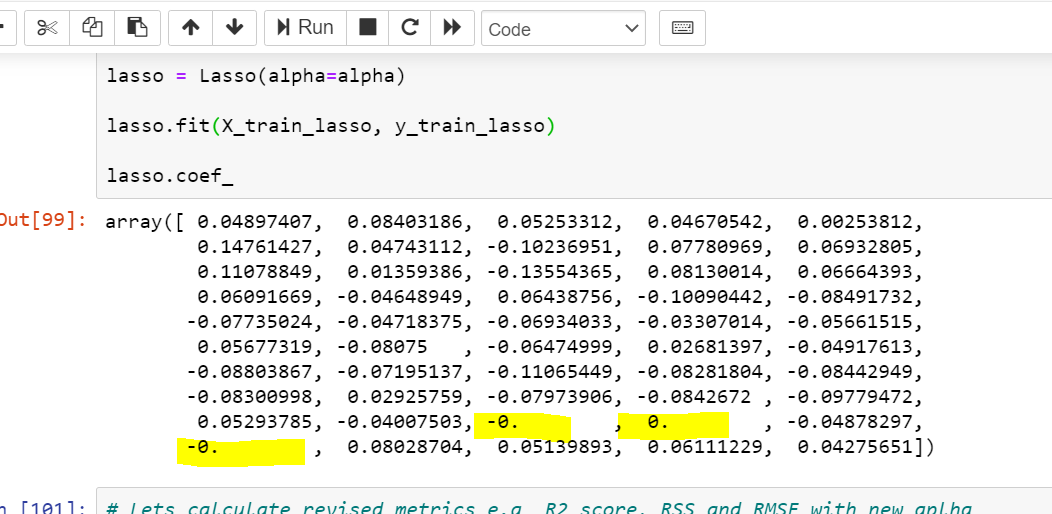


* **Post change to alpha it seems GrLivArea to be the top predictor emerging from both Lasso and Ridge regression models followed by**
* **MSZoning\_\***
* **LandContours\_\***

**Question 2**

You have determined the optimal value of lambda for ridge and lasso regression during the assignment. Now, which one will you choose to apply and why?

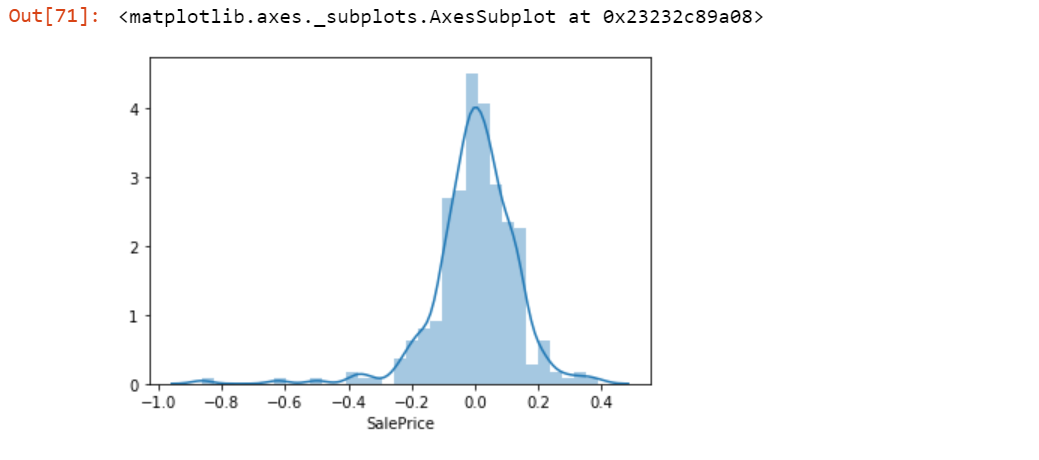
**Ans** : I will go with Lasso regression, as with lasso the penalty has pushed many of the coefficients towards 0. The model explanation with lesser variables is relatively easy, also the model complexity is reduced with lesser number of predictors.



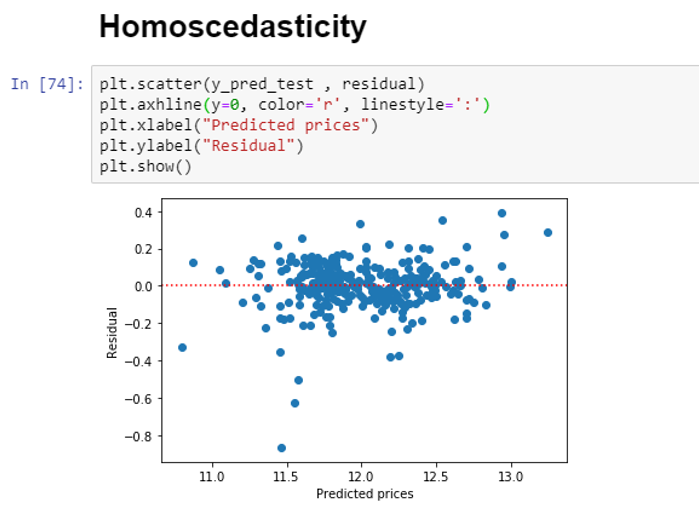
Identifying and Applying the Lambda (alpha) value is with trial and error which results in optimal regression matrices specially RMSE and **which helps ensure all the regression assumptions are held intact or met i**.e. Normality of residual teams, Homoscedasticity etc..

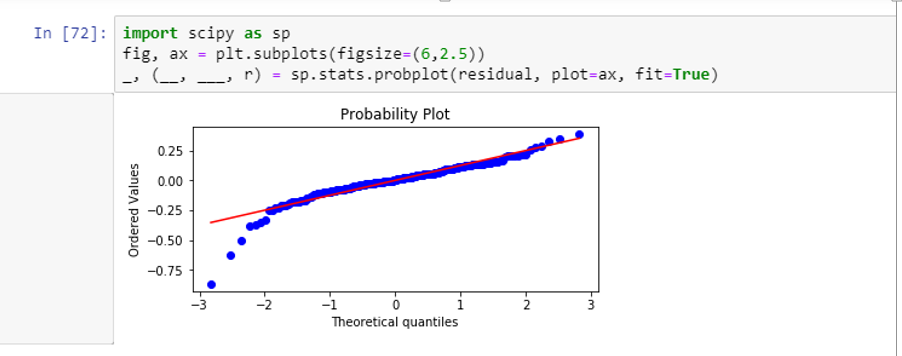
In this case we have chosen the value of alpha = 0.0001. Although the value is low however we could see the model performance is at 90%+ across both train and test sets.

The residual are normally distributed and they are randomly distributed. Variance in the error terms / residuals do not exhibit any patterns. The variance between the error terms is Homoscedastic.



*Error terms are normally distributed*





**Question 3**

After building the model, you realised that the five most important predictor variables in the lasso model are not available in the incoming data. You will now have to create another model excluding the five most important predictor variables. Which are the five most important predictor variables now?

**Ans:**

**Top Predictors with original dataset include following predictors:**

**MSZoning**: Identifies the general zoning classification of the sale.

A Agriculture

C Commercial

FV Floating Village Residential

I Industrial

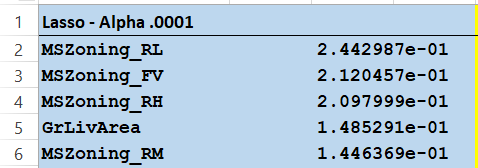
RH Residential High Density

RL Residential Low Density

RP Residential Low Density Park

RM Residential Medium Density

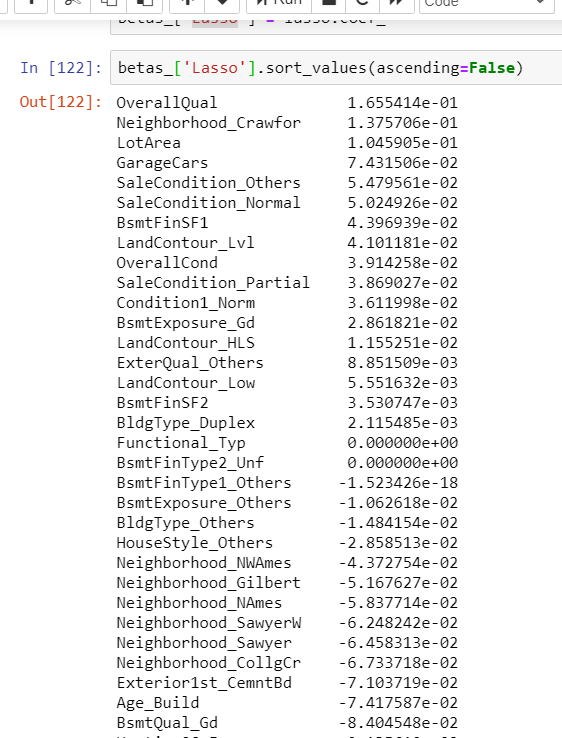
**GrLivArea**: Above grade (ground) living area square feet



We will drop these two fields from the predictor variable list and then rebuild the model. With the new model the top 4-5 predictors are

* **OverallQual**
* **Neighborhood\_Crawfor**
* **LotArea**
* **GarageCars**

**Snapshot of 20+ top predictors in order is given below.**



**Question 4**

How can you make sure that a model is robust and generalizable? What are the implications of the same for the accuracy of the model and why?

Model should be is robust and generalizable so that it is not impacted by the variance in the data wrt outliers and ranges in the training data. It’s important because the test accuracy can’t be drastically compromised as against the train accuracy. This leads to models being overfitting on Train set but under fitting on the test set. Overall model is over fitted and can’t be reliably used for unseen data. We adopt multiple strategies to make the model robust and generalizable:

* Outlier treatment :
  + The outlier analysis needs to be done and only those which are relevant to the dataset need to be retained i.e which make business sense. Those outliers which it does not make sense to keep must be removed from the dataset to ensure data stability.
* Data Transformation :
  + Transform the data (e.g. Log, inverse, sq , sqrt etc to ensure that the data can be transformed to follow a pattern.
* Scaling and standardisation: Same unit across all numeric variables helps model perform better and also standardisation helps make the variable more Gaussian which is what the most of the model expect and perform better on.
* Data imbalance – this is one issues which we cannot remediate easily (unless we manufacture data!) hence splitting into test and train can be avoided and we can use GRIDCV approach to use the dataset randomly multi-fold to as training data.
* Predictor Variability: Remove those variables which have near zero variability and doesn’t have much implications on decision making.

Too much weightage should not give to the predictors with outliers so that the accuracy predicted by the model is high

This has to be noted that we use to train data to fit once and only do transform on the train data (we don’t fit again)