==========

Problem Statement - Part II of the Assignment

==========

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 1:

- Optimal value of alpha for Ridge and Lasso Regression is below:
  - **Alpha value for Ridge Regression:** 2
  - **Alpha value for Lasso Regression:** 0.0001

- Doubling the Alpha value of **Ridge Regression** from 2 to 4
  - Train R2 score reduces little bit from 89.74% to 89.18% and Test R2 score also reduce from 87.67% to 87.11% as shown below.

```
**********Data after Ridge Regression Alpha Value = 4**********
Train R2 score:  0.8918465964512801
Test R2 score:  0.8711235502766299
Train RSS score:  1.8360807363940839
Test RSS score:  0.9866457629426572
Train MSE score:  0.0017983160983291713
Test MSE score:  0.0022474846536279206
Train RMSE score:  0.042406557256268414
Test RMSE score:  0.04740764340934825
**********
```

- Doubling Alpha value of **Lasso Regression** from 0.0001 to 0.0002
  - Train R2 score reduces little bit from 89.87% to 89.04% and Test R2 score also reduce from 88.23% to 87.64% as shown below.

```
**********Data after Lasso Regression with Value = 0.0002**********
Train R2 score:  0.8904249243822818
Test R2 score:  0.8764481675946416
Train RSS score:  1.860215942626232
Test RSS score:  0.9458818287453397
Train MSE score:  0.0018219548899375436
Test MSE score:  0.002154628311492801
Train RMSE score:  0.04268436352972296
Test RMSE score:  0.04641797401322898
**********
```

Combined comparison of original and new Alpha values for both Ridge and Lasso Regression

Metrics	Ridge Regression_2	Lasso Regression_0.0001	Ridge Regression_4	Lasso Regression_0.0002
R2 Score (Train)	8.97e-01	8.99e-01	8.92e-01	8.90e-01
R2 Score (Test)	8.77e-01	8.82e-01	8.71e-01	8.76e-01
RSS (Train)	1.74e+00	1.72e+00	1.84e+00	1.86e+00
RSS (Test)	9.44e-01	9.01e-01	9.87e-01	9.46e-01
MSE (Train)	1.71e-03	1.68e-03	1.80e-03	1.82e-03
MSE (Test)	2.15e-03	2.05e-03	2.25e-03	2.15e-03
RMSE (Train)	4.13e-02	4.10e-02	4.24e-02	4.27e-02
RMSE (Test)	4.64e-02	4.53e-02	4.74e-02	4.64e-02

Here,

- "Ridge Regression**_2**" and "Lasso Regression**_0.0001**" column contains metric for original value of Alpha.
- "Ridge Regression**_4**" and "Lasso Regression**_0.0002**" column contains double the original value of Alpha.

Overall outcome:
After doubling the Alpha Values for Ridge and Lasso, both Train and Test R2 score reduces slightly.

❖ Top 10 Most important predictor in predicting Sale Price after doubling the Alpha values are below:

- **GrLivArea:** Above grade (ground) living area square feet.
- **OverallQual:** Rates the overall material and finish of the house.
- **GarageCars:** Size of garage in car capacity.
- **OverallCond:** Rates the overall condition of the house.
- **FullBath:** Full bathrooms above grade
- **BedroomAbvGr:** Bedrooms above grade (does NOT include basement bedrooms)
- **MSZoning_RL:** Identifies residential with Low Density zone.
- **TotalBsmtSF:** Total square feet of basement area
- **BsmtFullBath:** Basement full bathrooms
- **Neighborhood_Crawfor:** Physical locations within Ames city limits is Crawford.

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?**

Answer 2:

R2 Score of Lasso is better than Ridge for Test Data, so we will prefer to go for Lasso regression.

Metrics	Ridge Regression_2	Lasso Regression_0.0001	Ridge Regression_4	Lasso Regression_0.0002
R2 Score (Train)	8.97e-01	8.99e-01	8.92e-01	8.90e-01
R2 Score (Test)	8.77e-01	8.82e-01	8.71e-01	8.76e-01
RSS (Train)	1.74e+00	1.72e+00	1.84e+00	1.86e+00
RSS (Test)	9.44e-01	9.01e-01	9.87e-01	9.46e-01
MSE (Train)	1.71e-03	1.68e-03	1.80e-03	1.82e-03
MSE (Test)	2.15e-03	2.05e-03	2.25e-03	2.15e-03
RMSE (Train)	4.13e-02	4.10e-02	4.24e-02	4.27e-02
RMSE (Test)	4.64e-02	4.53e-02	4.74e-02	4.64e-02

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?**

Answer 3:

After removing Top 5 most predictor from current model and then doing re-modelling we will get new R2 score and predictors. Below is the outcome:

- Train R2 score reduces drastically bit from 89.87% to 84.09% and Test R2 score also reduces from 88.23% to 80.24% after removal of Top 5 predictive variables.

  **********Data after Lasso Regression with Value = 0.0001 after removal top 5 variables**********
  Train R2 score: 0.8409328271411474
  Test R2 score: 0.8024966594896752
  Train RSS score: 2.700425158115736
  Test RSS score: 1.5120360197678266
  Train MSE score: 0.002644882623032063
  Test MSE score: 0.00344427339354858
  Train RMSE score: 0.0514284223268813
  Test RMSE score: 0.05868793226506264
  ****

- After removing Top 5 predictive variables values, both Train and Test R2 score drastically decreases in Lasso Regression.

Metrics	Lasso Regression_0.0001	Lasso Regression_subj_0.0001
R2 Score (Train)	8.99e-01	8.41e-01
R2 Score (Test)	8.82e-01	8.02e-01
RSS (Train)	1.72e+00	2.70e+00
RSS (Test)	9.01e-01	1.51e+00
MSE (Train)	1.68e-03	2.64e-03
MSE (Test)	2.05e-03	3.44e-03
RMSE (Train)	4.10e-02	5.14e-02
RMSE (Test)	4.53e-02	5.87e-02

- Below are the New Top 5 Most important predictor in predicting Sale Price after removing Top 5 predictive variables values from ORIGINAL model

    - **TotalBsmtSF:** Total square feet of basement area
    - **BedroomAbvGr:** Bedrooms above grade (does NOT include basement bedrooms)
    - **LotArea:** Lot size in square feet
    - **MSZoning_RL:** Identifies residential with Low Density zone.
    - **MSZoning_RH:** Identifies residential with High Density zone.

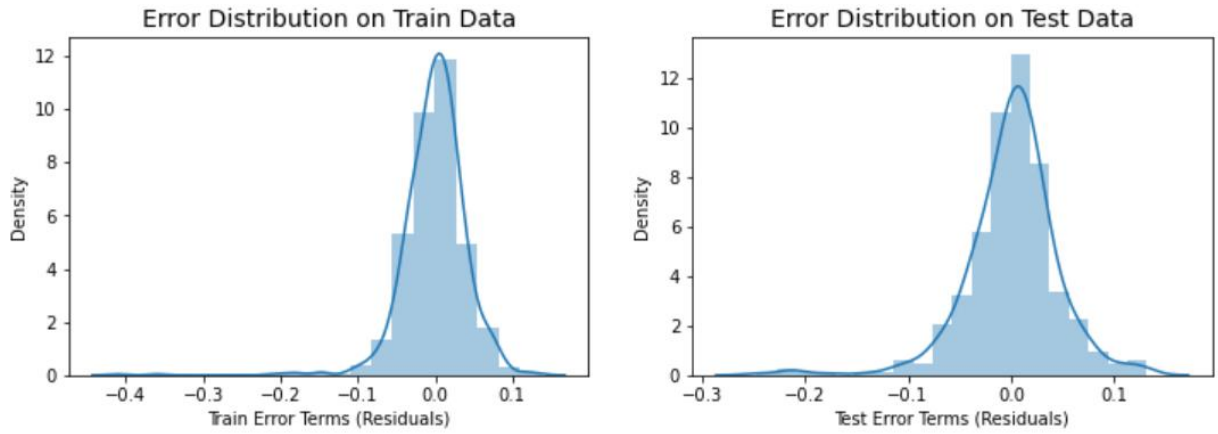
Question 4. **How can you make sure that a model is robust and generalizable? What are the implications of the same for the accuracy of the model and why?**

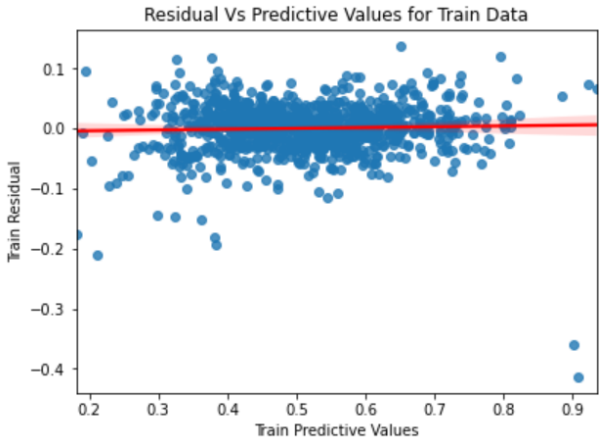
Answer 4:

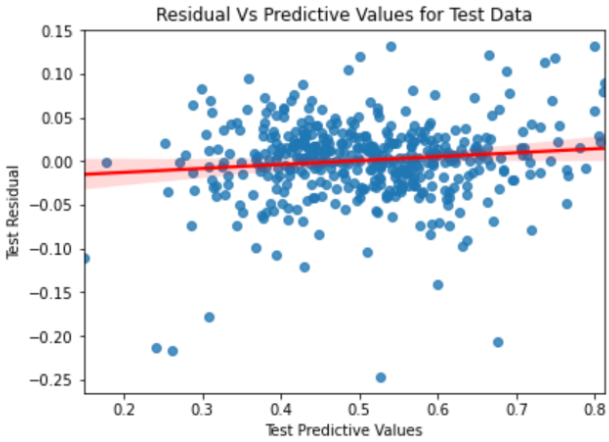
- Robust refers the model works for a broad range of inputs. If the model gets really good results at training time (it seems “more accurate”) but won’t generalize to out-of-sample data (i.e., it isn’t robust) then we call it overfitting.
- The model should be generalized so that the test accuracy is not lesser than the training score.
- Here in our case, based on all data and modelling both Ridge and Lasso performed good on Train and Test Data which shows our model with Alpha value "2" for Ridge and "0.0001" for Lasso is Robust and more Generalized model.
    - Simpler models are more generic.
    - Simpler model is more robust.
- Too much importance should not be given to the outliers so that the accuracy predicted by the model is high. But outlier’s analysis needs to be done and only those which are relevant to the dataset need to be retained and rest should be dropped.

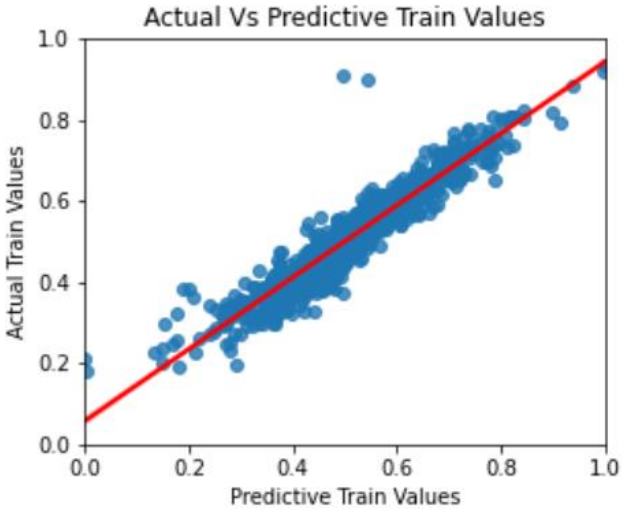
- If the accuracy of the Train and Test are same then that means model is overfitted and it learnt all the Train and Test data and model is not robust and generalized. So, it will drastically be failed and will not work on broad range of unseen data.

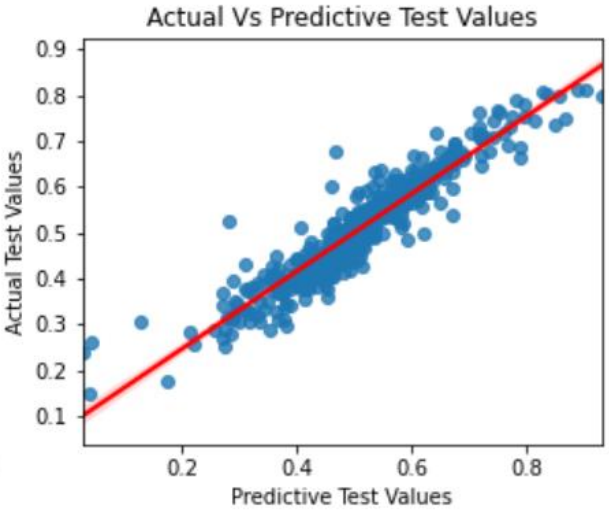
Graphical representation of residual analysis on Train and Test Data.











Conclusion:

- The residual analysis for both test and train data seem to fit the assumptions of the Linear Regression.
- Residuals have mean of zero and closely normally distributed.
- Residuals do not have any pattern hence it has homoscedasticity.