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?

Optimal Value of Alpha:

For Ridge = 2

For Lasso = 0.0004

Below is the table that shows change in value if we double the value of alpha:

	Alpha	R Square	Error
Ridge	2	0.8625	0.016013
Lasso	0.004	0.8679	0.0153829
Ridge	4	0.865866536	0.0156279
Lasso	0.008	0.87320822	0.0147726

Also along with that, we have got the change in value for coefficients too.

So, when we double the alpha for both ridge and lasso, then model will start pushing coefficient towards 0. In case of ridge it may shrink close to zero while in case of lasso, few variables may become exactly 0. It may result into underfitting too. In this case, we have found better results without compromising any metrics since the value of alpha we given for parameter tuning was not there and while doubling the alpha value, we got more suitable alpha value. But if we keep doing this , then point will come where the R square value will start dropping towards 0 too.

Most Important variable after the change is implement are still same that are –

- MSZoning
- Overallqual
- OverallCond
- GrLivArea
- Garage Type

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?

	Alpha	R Square	Error
Ridge	2	0.86	0.016013
Lasso	0.004	0.86	0.0153829
Ridge	4	0.86	0.0156279
Lasso	0.008	0.87	0.0147726

I will choose to apply lasso here since by selecting less variables or making model less complex we are able to get better RMSE value along with sample variance explained

Question 3

After building the model, you realized 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?

After removing the following variables (one more python file_ques_3.):

- MSZoning
- Overallqual
- OverallCond
- GrLivArea
- Garage Type

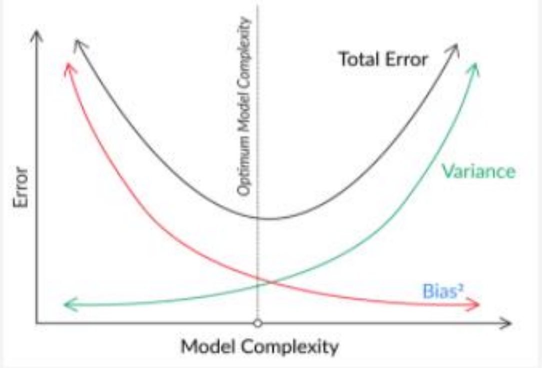
Five Most important Predictor Variable now are :

- 2ndFlrSF
- 1stFlrSF
- BsmtFinSF1
- GarageArea
- Neighborhood

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?

We know that whenever our train accuracy is high and test accuracy is low(overfitting) , out model is not robust and generalizable. We need tradeoff between bias and variance to have a generalize model which is going to perform better for unseen datasets with less variance. Below is the diagram that show trade off between bias and variance:



Image

Here we can see as per diagram that if we need optimum model complexity or robust model, we need lowest total error, i.e., low bias and low variance, such that the model identifies all the patterns that it should and is also able to perform well with unseen data. For this, we need to manage model complexity: It should neither be too high, which would lead to overfitting, nor too low, which would lead to a model with high bias (a biased model) that does not even identify necessary patterns in the data. There comes the usage of regularization which helps in managing model complexity by essentially shrinking the model coefficient estimates towards 0. This prevents the model from becoming too complex, thus avoiding the risk of overfitting. Implication of using this technique will result into compromising train accuracy because we are going to make our model less complex by losing on explain ability part but at the same time, we will prevent our model to overfit. Here We are going to shrink the variables coefficient to make our model less complex which will compensate with the accuracy of the model.