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 to 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, top ten predictor variables and the corresponding scores of the model are shown below.

Ridge: 10

Lasso: 0.001

:

	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	0.9237	0.9087	0.9067
1	R2 Score (Test)	0.9442	0.9137	0.9150
2	RSS (Train)	7.3919	8.8427	9.0370
3	RSS (Test)	3.0475	4.7093	4.6396
4	MSE (Train)	0.0929	0.1016	0.1027
5	MSE (Test)	0.0910	0.1131	0.1123

Top 10 features from Ridge	Top 10 features from Lasso
2ndFlrSF ---> 0.1090 OverallQual ---> 0.0792 1stFlrSF ---> 0.0720 MSZoning_RL ---> 0.0641 Age_of_house (yrs) ---> 0.0563 TotalBsmtSF ---> 0.0463 OverallCond ---> 0.0461 MSZoning_RM ---> 0.0395 MSSubClass ---> 0.0352 BsmtFinSF1 ---> 0.0341	2ndFlrSF ---> 0.1065 OverallQual ---> 0.0875 1stFlrSF ---> 0.0815 Age_of_house (yrs) ---> 0.0604 OverallCond ---> 0.0464 TotalBsmtSF ---> 0.0382 BsmtFinSF1 ---> 0.0314 GarageArea ---> 0.0305 SaleType_New ---> 0.0273 Neighborhood_Crawfor ---> 0.0238

If the values are doubled, i.e.

Ridge: 20

Lasso: 0.002

	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	0.9237	0.9070	0.9011
1	R2 Score (Test)	0.9442	0.9127	0.9126
2	RSS (Train)	7.3919	9.0087	9.5779
3	RSS (Test)	3.0475	4.7638	4.7690
4	MSE (Train)	0.0929	0.1025	0.1057
5	MSE (Test)	0.0910	0.1138	0.1138

Top 10 features from Ridge	Top 10 features from Lasso
2ndFlrSF -----> 0.1090 OverallQual -----> 0.0792 1stFlrSF -----> 0.0720 MSZoning_RL -----> 0.0641 Age_of_house (yrs) -----> 0.0563 TotalBsmtSF -----> 0.0463 OverallCond -----> 0.0461 MSZoning_RM -----> 0.0395 MSSubClass -----> 0.0352 BsmtFinSF1 -----> 0.0341	2ndFlrSF -----> 0.1065 OverallQual -----> 0.0875 1stFlrSF -----> 0.0815 Age_of_house (yrs) -----> 0.0604 OverallCond -----> 0.0464 TotalBsmtSF -----> 0.0382 BsmtFinSF1 -----> 0.0314 GarageArea -----> 0.0305 SaleType_New -----> 0.0273 Neighborhood_Crawfor -----> 0.0238

There is a slight variation in all scores compared to the scores obtained from the optimal value of alpha. The predictor variables and their coefficients are the same in both the values of alpha.

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?

The chosen model and lambda value will be Lasso model with value 0.001.

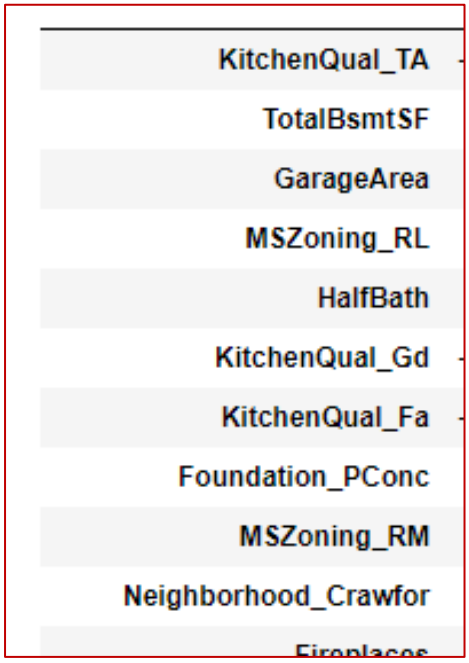
This is because the model has feature reduction technique, i.e. coefficients of predictors become 0. Also, the top ten features selected by lasso regression are more appropriate to the business problem.

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?

Once the initial top 5 variables are removed, and the model is built again on lasso, the next top ten variables are as follows:



However, R^2 value is lesser (82.1) compared to the initial (R^2 value 90.1) model and the mean square error is higher.

Question 4

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

A model is considered robust when the target variable is accurate consistently even though one or more predictor variables or even assumptions are changed due to unexpected circumstances. In clean words, model built should be simple. A simple model is comparatively more generic; and this type of model requires only few training samples. A simple model will have high bias, low variance. A weaker or complex model however will struggle to predict an accurate outcome for the target variable; hence will have a low bias and high variance.

Even though a simple model might make more errors, it does not lead to overfitting. A good model will have its R^2 value close to 1; and the mean square error of linear model should be much lesser than the mean square error of simple model. One of the main goals of a simple model is to achieve same performance every single time.

To make the model simple, outliers and missing values needs to be taken care of. Once that is done, regularization methods like ridge regression or lasso regression are used.