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:

- The optimal value of alpha for ridge and lasso regression are 7.0 and 0.001 respectively.
- After doubling the alpha for both regression the R2 value decreases. As shown below

```
- Old R2 value for Ridge:
    - R2 score (Train): 0.943686
    - R2 score (Test): 0.903403
- Old R2 value for Lasso:
    - R2 score (Train): 0.942149
    - R2 score (Test): 0.912092
- New R2 value for Ridge:
    - R2 score (Train): 0.9387009454584497
    - R2 score (Test): 0.9109831575244767
- New R2 value for Lasso:
    - R2 score (Train): 0.9353634972218796
    - R2 score (Test): 0.9123782463599925
```

- Most important predictor variable (Ridge):

```
'OverallQual_9', 'GrLivArea', 'Neighborhood_StoneBr',
'Neighborhood_Crawfor', 'BsmtExposure_Gd', 'Functional_Typ', '1stFlrSF',
'GarageCars_3', 'Condition1_Norm', 'OverallQual_10', 'OverallCond_8',
'OverallQual_8', 'BsmtFinSF1', 'TotalBsmtSF', 'OverallCond_7',
'Exterior2nd_CmentBd', 'FullBath_3', 'Neighborhood_Somerst',
'Neighborhood_NridgHt', '2ndFlrSF'
```

- Most important predictor variable (Lasso):

```
'OverallQual_9', 'OverallQual_10', 'GrLivArea', 'OverallQual_8',
'Neighborhood_Crawfor', 'Neighborhood_StoneBr', 'Functional_Typ',
'BsmtExposure_Gd', 'GarageCars_3', 'SaleCondition_Partial',
'OverallCond_8', 'Exterior2nd_CmentBd', 'OverallCond_7',
'Neighborhood_Somerst', 'BsmtFinSF1', 'Condition1_Norm',
'Exterior1st_BrkFace', 'Neighborhood_BrkSide', 'TotalBsmtSF',
'OverallCond_6'
```

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:

	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	0.951409	0.943686	0.942149
1	R2 Score (Test)	0.903403	0.912092	0.913241
2	RSS (Train)	38.095156	44.150091	45.354989
3	RSS (Test)	43.861859	39.916710	39.394735
4	RMSE (Train)	0.220433	0.237305	0.240522
5	RMSE (Test)	0.360768	0.344162	0.341904

- From the above two techniques of Lasso and Ridge Regression, we can say that both almost having the same r2 value.
- When comparing the complexity, it is better to use Lasso because as we have 238 variables, Lasso will make the feature selection among the present variables, but Ridge will not reduce columns, it will keep all 238 variables with the reducing the coefficient of variables.

So, will use the Lasso Regression model as the final model.

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:

Top 5 variables in the current model:

- OverallQual_9
- OverallQual_10
- GrLivArea
- OverallQual_8
- Neighborhood_StoneBr

The five most important predictor variables after removing the top 5 variable from final model are:

- 1stFlrSF
- 2ndFlrSF
- Neighborhood_Crawfor
- Exterior2nd_CmentBd
- SaleType_CWD

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?

Answer:

While creating the best model for any problem statement, we end up choosing from a set of models that would give us the least test error. Hence, the test error, and not only the training error, needs to be estimated in order to select the best model.

This can be done in the following two ways:

- Use metrics that take into account both the model fit and its simplicity. They penalise the model for being too complex (i.e., for overfitting) and, consequently, more representative of the unseen 'test error'. Some examples of such metrics are adjusted R2, AIC and BIC.
- Estimate the test error via a validation set or a cross-validation approach. In the validation set approach, we find the test error by training the model on a training set and fitting on an unseen validation set, while the in n-fold cross-validation approach, we take the mean of errors generated by training the model on all folds except the kth fold and testing the model on the kth fold, where k varies from 1 to n.

So far, you have understood that MSE of the training might not be a good estimate of the test error. The aforementioned parameters are a manipulation of the RSS (residual sum of squares), wherein a penalty term is introduced to compensate for the increase in complexity due to the increase in the number of predictors.