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:

Model Evaluation metrics with best alpha values:

- Best Alfa value using Ridge regression is: 0.9
- Best Alfa value using Lasso regression is: 0.00001

```
Score Type     | Lasso   | Ridge
--------------------------------
R2_Score_Train | 0.9183  | 0.9178
R2_Score_Test  | 0.8714  | 0.8727
RSS_Train      | 5.7673  | 5.8025
RSS_Test       | 4.2149  | 4.2149
MSE_Train      | 0.0058  | 0.0059
MSE_Test       | 0.0099  | 0.0098
RMSE_Train     | 0.0764  | 0.0766
RMSE_Test      | 0.0997  | 0.0992
```

Model Evaluation Metrics with double the best alpha values:

```
Score Type     | Lasso   | Ridge
--------------------------------
R2_Score_Train | 0.9182  | 0.9182
R2_Score_Test  | 0.8715  | 0.8716
RSS_Train      | 5.7694  | 5.7693
RSS_Test       | 4.211   | 4.211
MSE_Train      | 0.0058  | 0.0058
MSE_Test       | 0.0099  | 0.0099
RMSE_Train     | 0.0764  | 0.0764
RMSE_Test      | 0.0997  | 0.0996
```

When the alpha value is doubled, the R2 values reduced, but in insignificant amount. And below are the top 10 most important predictors based on beta values. Even the alpha values are doubled, still they are the top 10 variables with different order.

- GrLivArea,
- TotalBsmtSF,
- GarageQual,
- OverallQual,
- Exterior2nd_CmentBd,
- Neighborhood_NoRidge,
- OverallCond,
- GarageYrBlt_2008.0,
- Neighborhood_Somerst,
- Fireplaces

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:
As the R2 values and other metrics are almost same, its tough decision to take. However I will go with Lasso regression as it gives option for feature selection. It will give similar accuracy with reduced number of features.

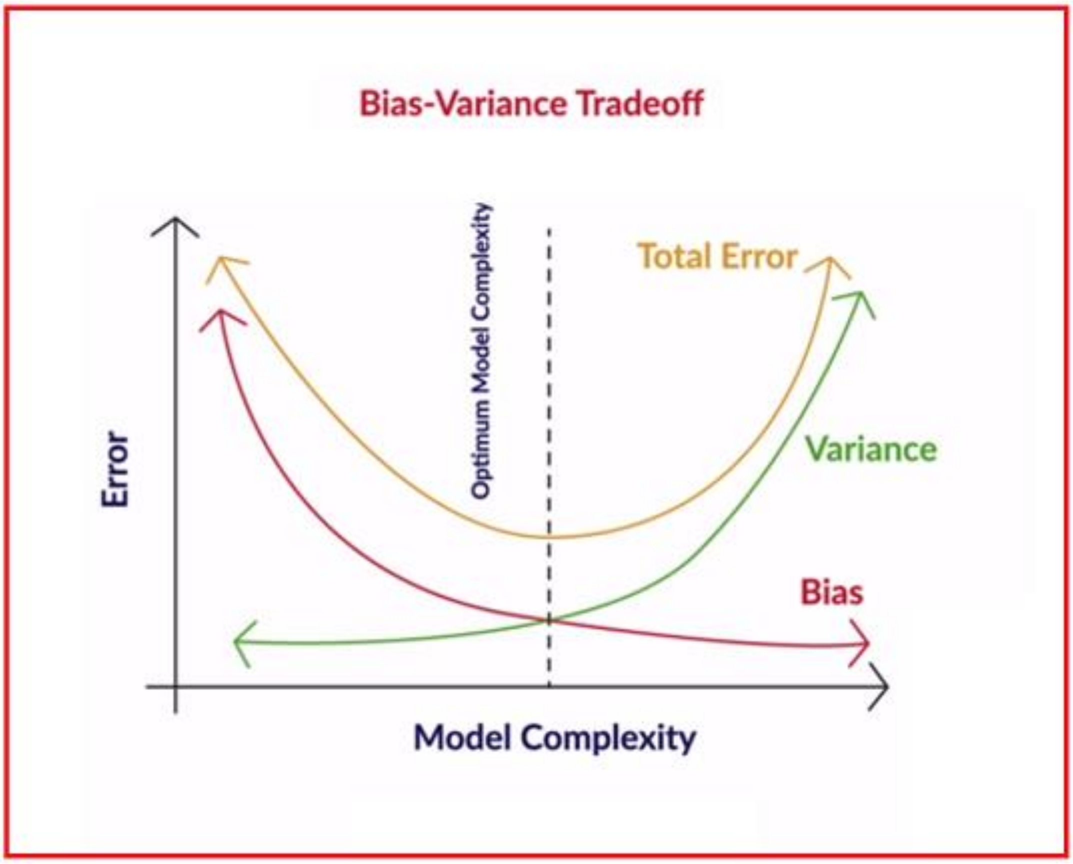
Question 3:
After building the model, you realised that the five most important predictor variables in the lasso model are not available in the incoming data. You will now have to create another model excluding the five most important predictor variables. Which are the five most important predictor variables now?
Answer:
Below are the 5 most important features that we can exclude

- Neighborhood_NoRidge,
- OverallCond,
- GarageYrBlt_2008.0,
- Neighborhood_Somerst,
- Fireplaces

After removing these variables, R2 score slightly reduced. We get the R2 value as 0.9031 for training & 0.8642 for test

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:
A simpler model is usually more generic than a complex model. Generic models are bound to perform better on unseen data sets. A simple model is more robust and does not change significantly if the training data points undergo small changes. A simple model may make more errors in the training phase but is bound to outperform complex models when it views new data.

Here comes the Bias-Variance trade-off



Variance: How sensitive is the model to the input data.
Bias: How much error the model is likely to make in test data

When the model is very simple, it has high bias and low variance. It will give low r2 score on train and test.
But when the model becomes very complex, it will be overfit and has high training accuracy, but lower test accuracy.
So, by using regularization, we deliberately simplify the model to achieve correct balance between keeping the model simple and too naïve.

This can be achieved by Ridge and Lasso Regression by adding regularization to error term.