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?

The optimal value of alpha for ridge and lasso regression from the model creation turned out to be 500. If the value of alpha is doubled, this would lead to underfitting of the model. The variance of the model would decrease and the bias would increase. The slope of the best fit line would slowly reduce and become horizontal or parallel to the x-axis as the alpha increases.

The coefficients would become smaller and smaller

If the value of alpha is doubled, which will be 1000, then

In ridge regression, the important predictors would be

```
Top 10 predictors:
GrLivArea: Coefficient = 5838.945881500783
OverallQual_9: Coefficient = 5034.59857753953
1stFlrSF: Coefficient = 4796.238032305292
GarageCars_3: Coefficient = 4794.0627082455685
OverallQual_10: Coefficient = 4757.49115774285
FullBath_3: Coefficient = 4504.581307199981
TotRmsAbvGrd_10: Coefficient = 4044.539485868117
Condition2_PosN: Coefficient = 4030.9905365543236
TotalBsmtSF: Coefficient = 3908.6134863986986
Neighborhood_NridgHt: Coefficient = 3867.3336856571304
```

In lasso regression, the top 10 predictors would be

```
Top 10 predictors:
GrLivArea: Coefficient = 25162.519486418623
Condition2_PosN: Coefficient = 12318.961836834551
OverallQual_9: Coefficient = 12281.445432322613
OverallQual_10: Coefficient = 9711.991737949855
OverallQual_8: Coefficient = 8630.25864475213
GarageCars_3: Coefficient = 7867.261931859171
FullBath_3: Coefficient = 6009.324190185531
BsmtExposure_Gd: Coefficient = 5007.68396446389
ExterQual_TA: Coefficient = 4613.075094704343
Neighborhood_NridgHt: Coefficient = 4351.711496577023
```

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?

There are lot of predictors in the equation and ridge regression would bring the coefficients would bring their values close to zero while lasso regression helps with feature selection.

Lets look at the metrics

	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	9.69E-01	9.15E-01	9.36E-01
1	R2 Score (Test)	-2.18E+21	8.57E-01	8.34E-01
2	RSS (Train)	1.98E+11	5.41E+11	4.06E+11
3	RSS (Test)	6.15E+33	4.04E+11	4.68E+11
4	MSE (Train)	1.39E+04	2.30E+04	1.99E+04
5	MSE (Test)	3.75E+15	3.04E+04	3.27E+04

Ridge and Lasso have very close R2 train and test values. While Ridge does slightly better than Lasso. The Residual Sum of Squares is also quite comparable. Lasso seems to have done better on the Train Mean Squared Error and Ridge to have done better on the Test data.

Looking at the data, I would like to go with ridge regression as it seems to perform consistently with both Train and Test data.

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?

I have removed the first five predictors from the initial lasso regression model equation. The alpha came out to be 1000. The top five coefficients that came out of the model are

```
1stFlrSF: Coefficient = 17061.95181063459
2ndFlrSF: Coefficient = 14878.32902308476
OverallQual_7: Coefficient = 13169.029171517459
OverallQual_6: Coefficient = 9245.380359974513
OverallQual_8: Coefficient = 8806.516872639098
```

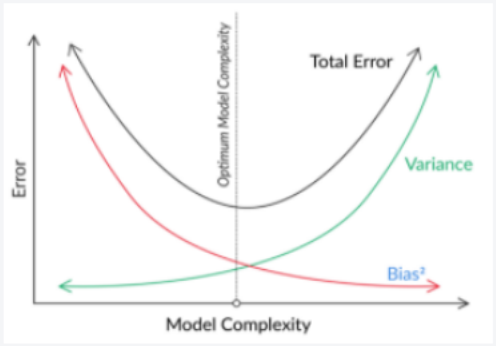
Question 4

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?

When a model is too simple, it shows that the model can perform very bad on the training set but will be able to perform well on unseen test data. When the complexity of the model is high, then the model performs well on the training data but may not perform well on unseen test data.

So, the variance is low whereas the bias is higher for simple model and the case of a complex model, the variance will be higher and bias will be lower. We have to find an optimal balance between bias, variance and model complexity.

So, regularization helps with managing the model complexity by imposing a penalty on the coefficients by shrinking the model coefficient estimates towards 0.



Image

This discourages the model from becoming too complex avoiding the risk of overfitting and too simple to avoid the risk of underfitting.

Here the cost function would be Cost = RSS + Penalty. Adding this penalty in the cost function helps suppress or shrink the magnitude of the model coefficients to 0.

When building a Ordinary Least Squares (OLS) model, the coefficients for whatever cost/loss , RSS is minimum. Optimising the cost fuction results in model coefficients with the least possible bias, although the model may have overfitted and hence have high variance.

2. Using Cross validation like k-fold cross validation to assess the model's performance across different subsets of the same data. This ensures that the model's performance is consistent and is not dependent on the split of the data.

Effect on Accuracy: Using Regularization and optimized hyperparameters lead to higher accuracy on unseen data by finding the right balance between complexity and overfitting.

Using cross validation, we can get more accuracy of its generalization performance by ensuring that the model estimates well on train data.