Subjective Questions

Q1 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?

Ans 1 The optimal value of alpha for ridge and lasso regression is 500 for both.

The changes in the model if we choose to double the values of alpha to 1000 for both ridge and lasso are as follows:

	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	9.405150e-01	8.604067e-01	8.857179e-01
1	R2 Score (Test)	-5.890046e+20	8.598371e-01	8.566159e-01
2	RSS (Train)	4.360306e+11	1.023232e+12	8.376986e+11
3	RSS (Test)	1.105735e+33	2.631270e+11	2.691742e+11
4	MSE (Train)	1.932133e+04	2.959822e+04	2.678073e+04
5	MSE (Test)	1.945961e+15	3.001866e+04	3.036165e+04

We observe that r2scores remains similar to what it was for alpha = 500 for alpha = 1000 also.

The most important predictor variables after the change is implemented are:

For ridge regression (Top 5 variable with highest absolute coefficients):
OverallQual
GrLivArea
Neighborhood_NridgHt
1stFlrSF
Neighborhood_NoRidge

For lasso regression (Top 5 variables with highest absolute coefficients):
GrLivArea
OverallQual
Condition2_PosN
Neighborhood_NridgHt

Neighborhood_NoRidge

Q2 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?

We have determined the optimal value of lambda for ridge and lasso regression during the assignment as 500 for both.

The following are the metrics obtained for ridge and lasso regression:

	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	9.405150e-01	8.817133e-01	9.140455e-01
1	R2 Score (Test)	-5.890046e+20	8.707261e-01	8.482173e-01
2	RSS (Train)	4.360306e+11	8.670527e+11	6.300546e+11
3	RSS (Test)	1.105735e+33	2.426851e+11	2.849408e+11
4	MSE (Train)	1.932133e+04	2.724591e+04	2.322564e+04
5	MSE (Test)	1.945961e+15	2.882904e+04	3.123819e+04

The optimal value of lambda for ridge and lasso regression is 500 for both.
Ridge and Lasso regression perform well both on training and test set.
Total error i.e. sum of training and test error is approximately equal for both ridge and lasso regression.
However, difference between train and test errors is lesser for ridge regression than lasso regression, and also test error is lesser for ridge regression.
Thus, ridge regression seems to perform the best.

Therefore, I will choose to apply ridge reasons because of the above reasons.

Q3 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?

After creating another model excluding the five most important predictor variables which were GrLivArea, OverallQual, Condition2_PosN, Neighborhood_NridgHt, Neighborhood_NoRidge we obtained the following results.
Train r2_score of new lasso model = 0.896
Test r2_score of new lasso model = 0.885

The five most important predictor variables now are (Top 5 variable with highest absolute coefficients):
RoofMatl_CompShg
2ndFlrSF
1stFlrSF
RoofMatl_WdShngl
RoofMatl_Tar&Grv

Q4 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?

We can make sure that the model is robust and generalisable by using the bias variance tradeoff. We will select the model which has both low variance and low bias so that the total error for the selected model is minimum.

The results for the different models are as follows:

	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	9.405150e-01	8.817133e-01	9.140455e-01
1	R2 Score (Test)	-5.890046e+20	8.707261e-01	8.482173e-01
2	RSS (Train)	4.360306e+11	8.670527e+11	6.300546e+11
3	RSS (Test)	1.105735e+33	2.426851e+11	2.849408e+11
4	MSE (Train)	1.932133e+04	2.724591e+04	2.322564e+04
5	MSE (Test)	1.945961e+15	2.882904e+04	3.123819e+04

Total error i.e. sum of training and test error is approximately equal for both ridge and lasso regression.
However, difference between train and test errors is lesser for ridge regression than lasso regression, and also test error is lesser for ridge regression.

Thus, ridge regression seems to be more robust and generalisable.
This is because total error which is sum of bias error and variance error in bias variance tradeoff seems to be minimum for ridge regression.

Implications for accuracy of the model are that ridge regression has lower test error and difference in train and test error is also lesser for ridge regression.