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?

Answer:

Ridge regression: Optimal value of alpha obtained for ridge regression is 10.0.
Lasso regression: Optimal value of alpha obtained for lasso regression is .001

As the value of alpha increases, there is a significant bias and variance trade-off is observed. As the value of alpha (lambda) increases it will try to shrink the co-efficient towards zero which will make y predicted value less i.e., "Y = Bo + B1*X" and hence RSS will increase.

As the change is not significant though r2_score of the model is dropped due to an increase in RSS.

The most significant parameter in lasso regression is "GrLivArea", where as

The most significant parameter in ridge regression is "OverallQual".

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:

Lasso and ridge regression both are used for regularization. Lasso regression helps to perform feature selection in model by shrinking co-efficient towards zero as well as makes some of the coefficient zero. Ridge regression is preferred when all variables have same level of significance, and it shrinks the coefficients towards zero not zero.

In our analysis lasso and ridge both have similar results, but lasso results tend to seem better as r2_score and RSS, RMSE results both are same for training and test data whereas for ridge regression we can notice slight variation in results.

In addition, from EDA we found that not all variables have same level of impact on response variable and only few has higher co-efficient, we choose lasso over ridge as it will help the model in feature selection.

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:

Lasso model tends to perform better as r2_score is very close to each other, whereas and RSS & RMSE are very close to each other in both the models.

Answers below are from lasso regression.

Most important features obtained for the lasso regression before were:

Five most important features are highlighted in red (in terms of absolute coeff)

Features with (+) ve coefficient		Features with (-) ve coefficient	
GrLivArea	0.121	Property Age	-0.066
Sale Type (type: New)	0.094	BldgType_Twnhs	-0.031
OverallQual	0.090	HeatingQC_FA	-0.028
Neighborhood_Crawfor	0.071	HeatinQC_TA	-0.028
Overall Cond	0.059	GarageType_none	-0.028

Where Property Age (derived column) YrSold - YearBuilt

After dropping the five most important features:

Assumption: As five most important feature contains some of subtype of feature (for ex: saleType = new and Neighborhood = Crawfor). It has been assumed that feature column itself is not available in solution.

Five most important features are highlighted in red (in terms of absolute coeff).

Features with (+) ve coefficient		Features with (-) ve coefficient	
Exterior1st_BrkFace	0.104	BsmtQual_TA	-0.092
2ndFlrSF	0.092	KitchenQual_TA	-0.087
Central Air (Type: Y)	0.077	MSZoning_RM	-0.070
1stFlrSF	0.073	KitchenQual_Gd	-0.069
LandContour_HLS	0.069	GarageType_none	-0.069

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:

Model tends to perform better if it has low variance and low bias, and it does not overfit/underfit the data. Model over-fitting can be observed by analysing the r2_socre, comparing RSS & RMSE in training and test data. For ex: In housing price prediction we clearly saw that r2_score for training data was approx. 90 and for test data it was < 0. It was a sign of over fitting of the model.

In addition, model residuals should be plotted against Y and there should not be any variance observed in the plot. If a variance is observed in histogram of error
or residual plots against Y, it may indicate non-linearity of the data and a model may be overfitted. To solve the case polynomial regression, data-transformation and non-linear regression can be used.

To make a sure a model is robust and generalisable, model should have less features in final model. In addition, r2_score for training and test data should be same. **K fold cross validation should be used to find out** the hidden patter in underlying data and a best model should be chosen.

Model is considered robust and generalisable if it has less variance and tends to give similar result on test data. In addition, it should be simple enough for business to understand.