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 -

Optimal values of ridge & lasso –

1. Ridge : alpha value = 500
2. Lasso : alpha value = 400

When alpha values are increased for ridge & lasso, changes observed are –

1. Ridge : The R2 score decreased little bit when alpha value is doubled. Also the RSS & RSME has increased.
2. Lasso : The R2 score for test dataset increased when alpha value is doubled. Also the RSS & RMSE has decreased for test dataset.

Below are the predictor variables before & after changing value of alpha -

	Ridge Regression	Lasso Regression
Before	OverallQual_Qual_Excellant GrLivArea FullBath_3 OverallQual_Qual_VExcellant GarageCars_3	RoofMatl_CompShg GrLivArea RoofMatl_WdShngl RoofMatl_Tar&Grv OverallQual_Qual_Excellant
After	OverallQual_Qual_Excellant GrLivArea FullBath_3 GarageCars_3 1stFlrSF	GrLivArea OverallQual_Qual_Excellant OverallQual_Qual_VExcellant OverallQual_Qual_VGood GarageCars_3

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 -

In the dataset there are many independent variables that we need to deal with, so lasso regression will be used as it helps for automatic feature selection. Lasso regression not only reduces complexity of model, but it also helps to remove non relevant features by making coefficients 0. Ridge regression avoids the model from getting overfit. Both ridge & lasso have almost got same accuracy. So, taking into consideration this dataset lasso regression can be used. The R2 score is also good for training & test dataset for lasso regression. The value of alpha is also not too high which will lead to underfitting & not too low that will lead to underfitting.

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 -

If the five most important variables are not present in dataset, we need to drop those five important variables and retrain with remaining independent variables. After retraining below are the 5 predictor variables given by lasso regression model –

1stFlrSF
2ndFlrSF
PoolArea
GarageCars_3
Neighborhood_NridgHt

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 -

The model can be said generalizable when testing accuracy or the accuracy obtained on unseen data does not vary too much from training accuracy. So, the model should perform similar on unseen data as it had performed on training dataset. Outliers in the dataset should be removed in order build models which can be robust & generalizable. Simple model is considered to be more generalizable that is t doesn't vary more on unseen data. Complex model gives more training accuracy & less testing accuracy. So more generalizable model there is less difference in accuracy of training & test dataset.