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?

In our final model, the optimal value of alpha in Ridge regression is 1.0 and Lasso regression is 0.0001. When we double these values, the model performance is almost the same.
After the changes, the key predictor variables in Ridge regression are :
'LotFrontage, OverallCond, BsmtUnfSF, BsmtFinType2, OverallQual, BsmtFinSF1, BsmtExposure, GarageFinish, LotConfig_CulDSac, ExterCond'

In Lasso the key predictor variables are :
'LotFrontage, BsmtFullBath, OverallCond, CentralAir, OverallQual, Exterior1st_CBlock, MSZoning_RH, MSZoning_RM, Street_Pave, GarageQual'

The entire calculation for this question is shown in Subjective questions section of the python notebook at the end.

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?

Both Ridge regression and Lasso gave very similar results in terms of its performance. Ridge regression gave accuracies(r2 score in percentage) : 92.72 percent on the train set and 74. 5 percent on the test dataset. Lasso regression gave accuracies(r2 score in percentage) : 92.80 percent on the train data set and 73.73 percent on the test dataset. Despite the fact that ridge performs (1 percent)

better than lasso on the test dataset, We still choose the lasso model finally. The reason being, the Lasso model does feature elimination. In the current dataset, there are many(80+) columns. Hence feature elimination would be important to understand the key predictor variables. Therefore our final model is Lasso with an r2 score of 0.9280 on the train data set and 0.73 on the test dataset respectively. i.e 92.80 percent on the train dataset and 73.73 percent on the test data set.

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?

The five most important predictor variables in our Lasso model initially were - 'LotFrontage','BsmtFullBath','OverallCond','CentralAir','OverallQual'

After removing the top 5 predictor variables and rebuilding the model, the five most important predictor variables now are –
'LotArea, FullBath, 1stFlrSF, ExterCond, MSZoning_RH'

The entire calculation for this question is shown in Subjective questions section of the python notebook at the end.

Question 4. How can you make sure that a model is robust and generalizable? What are the implications of the same for the accuracy of the model and why?

We can have a model that is robust and generalizable by ensuring that the model is as simple as possible and it doesn't overfit. Regularization also helps in making the model simpler. The accuracy of the model increases on the train dataset if we are overfitting the model but it fails to generalize on the unseen data/test

data or the accuracy on the test data set is very low. When we say a model can generalize well and it is robust, then it means both the training and testing accuracies are pretty good.