SUBJECTIVE QUESTIONS

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 1:
The optimal value of alpha for ridge regression is 2 and the optimal value of alpha for lasso regression is 0.0005.

Increasing the value of alpha for both ridge and lasso regression will increase the bias of the model, which means that model will not be able to learn the patterns in the training data well(underfitting).

For alpha = 4 in ridge regression, the most important predictor variables are Living Area, Overall Quality, Basement Area, Overall Condition, and Lot Area.

For alpha = 0.001 in lasso regression, the most important predictor variables are Living Area, Overall Quality, Overall Condition, Lot Area, and Basement Area.

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?

I will be using alpha = 0.0005 for lasso regression since the model built using that had output the predictor variables which made more business sense. The most important predictor variables are Living Area, Overall Quality, Overall Condition, Lot Area, and Basement Area.

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 3:
The next set of important predictor variables returned by the lasso regression model are Finished Basement Area, Neighbourhood(Crawford), Neighbourhood(Stone Brook), Neighbourhood(Northridge Heights), and Sale Type(new).

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?

A robust and generalizable model requires, in technical terms, a bias-variance tradeoff. The model should not be too simple(high bias) such that it performs badly on the training data(underfitting) and the model should not be too complex(high variance) such that it performs badly on the test/unseen data(overfitting).

To achieve this, for regression models, we use regularisation which simply adds a penalty to the model for using too many features. The implications of this on the training data is that we compromise a bit on the training accuracy to achieve reasonable test accuracy, as in the end, we want our model to make reliable predictions on the new unseen data that is fed to it.

However, in regularization, choosing the appropriate value of alpha(penalty term weight) is important. Choosing too less a value will not benefit from the power of regularization and hence lead to overfitting and choosing too large a value will overpower the penalty term on the error term which leads to underfitting.