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?

Optimal Value of α for Ridge: 3.0
Optimal Value of α for Lasso: 0.001

When the alpha is doubled for Ridge regression, the features remain the same but the coefficient values reduces. While in Lasso regression the coefficient values are getting reduced but many other variables are reduced to 0 (or eliminated). The important predictor remains the same.

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?

Out of the two, it would be the Lasso regression as it eliminates non related variables. Hence a model with limited variables is easy for the business to interpret and make significant decision based on those variables.

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 new significant features are

MSSubClass_MSC60 : Type of dwelling involved in the sale - 2-STORY 1946 & NEWER
2ndFlrSF : Second floor square feet
Exterior1st_BrkComm : Exterior covering on house - Brick Common
1stFlrSF : First Floor square feet
Neighborhood_Crawfor : Crawford

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?

Robustness:

The model is built using regularization in order to be robust. In regression, robustness is usually affected by data that does not satisfy some of the assumptions/expectations like

- Linear Independence
- Outliers
- Data that is not normally distributed (Skew and Kurtosis)
- Homoscedasticity (expected) - Heteroscedastic Errors (High variance) – Usually caused by outliers

These are corrected using regularization.

Generalisable:
The property of the model not to overfit and have similar errors in both train and test datasets. Model is built on these principles and will not overfit.

The accuracy of the model may be affected if the model becomes too generic. Hence a proper bias-variance threshold is identified to balance between accuracy and robustness.