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?

The optimal values of alpha for ridge and lasso regression were 10 and 0.0001 respectively. After doubling the optimal value of alpha for ridge regression, there was not a significant change in the train and test r-squared metrics. The train r-squared value dropped by merely 0.7 percentage points.

The variables contributing the most to the prediction of sale price are as follows:

- GrLivArea
- TotalBsmtSF
- OverallQual_9
- 1stFlrSF
- OverallQual_8

After doubling the optimal value of alpha for lasso regression, the test r-squared value increased from approximately 81% to approximately 85%. The variables contributing the most to the prediction of sale price are as follows:

- GrLivArea
- OverallQual_9
- TotalBsmtSF
- OverallQual_8
- YearBuilt

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?

In this case, the Ridge regression model seems to generalize much better on the test set than the Lasso regression model, shown by its higher r-squared value on the test set. Hence, the Ridge regression model will be chosen.

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?

After excluding the 5 most important predictor variables and creating another model, the 5 most important predictor variables are the following:

- 1stFlrSF
- 2ndFlrSF
- BsmtFinSF1
- GarageArea
- OverallCond_9

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?

To ensure that a model is robust and generalisable, there must be a balance between variance and bias. As the complexity of a model increases, the bias decreases and the variance increases. A low bias and high variance model will have overfit the training data, memorized the noise in the data and hence will not perform well on an unseen dataset. On the other hand, a high bias and low variance model would not have learnt the underlying patterns of the data distribution. Therefore, to ensure that the model is not too complex, regularization techniques such as ridge and lasso regression can be used. The implication of regularization is that the bias might increase as the variance decreases and hence the accuracy of the model could be lower.