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:

The optimal value of alpha for ridge regression is 10.0 and for Lasso regression is 0.001.

If we choose to double the value of alpha, it will increase the regularization strength of the model.

The R2 score decreases a little for both the models in this case.

The most important predictor variable after the change is implemented will be GrLivArea which is same as before.

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:

The R2-score for both the Ridge and Lasso model is almost same for train and test data. But as Lasso regression model performs feature selection and reduces the number of independent variables, lasso regression model is preferable as it will be simple and robust.

Question 3:

After building the model, you realised that the five most important predictor variables in the lasso model is 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:

The five most important predictor variables after dropping the 5 most important predictor variables in the lasso model are:

- Condition2_PosA
- RoofMatl_Membran
- SaleType_Oth
- SaleType_New
- RoofMatl_WdShngl

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:

To make sure that the model is robust and generalisable we perform the below steps:

- Handling Missing data and outliers to mitigate the impact of the same on the model.
- Cross-validation helps in evaluating the model's generalization ability and reduces the risk of overfitting to the training data.
- Train-Test Split to evaluate the data on unseen test data and train it on the train data set.
- Tuning of hyperparameters to get the optimal value for the model building.
- Feature selection to sleect features which are strongly correlated with the target variable.
- Regularization to prevent overfitting and to create robust model.
- The implication of the same is that we have more accurate model which can perform better on unseen data as the model has been trained on test data but has not learnt all the noises in the data
  as well. Avoiding overfitting of the model using cross-validation and evaluation on test data helps in the accuracy of model.