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?

Ans.: The value of alpha for Ridge regression is 7.924828983539186 and for Lasso regression is 0.0002848035868435802

Upon doubling the value of alpha, the following changes are noticed:

Ridge regression

- The R^2 value for the train set changes from 0.8835735526876723 to 0.8710463043588892
- The RMSE value for train set changes from 0.0375375904383299 to 0.039505491132225505
- The R^2 value for test set changes from 0.8826368614236625 to 0.8707388289952278
- The RMSE value for test set changes from 0.038145894661289245 to 0.04003280218667854

Lasso regression

- The R^2 value for the train set changes from 0.8724811593967157 to 0.8492077365163264
- The RMSE value for train set changes from 0.0392850894708999 to 0.04271988736655009
- The R^2 value for test set changes from 0.8814990058953969 to 0.855646175159684
- The RMSE value for test set changes from 0.03833036408253015 to 0.04230542822061536

After the change is implemented, the 5 most important predictor variables for Ridge regression changes to Functional_Min1, Functional_Maj2, Neighborhood_NoRidge, RoofMatl_WdShngl, Neighborhood_StoneBr. For Lasso, it changes to LandContour_Lvl, YearRemodAdd, RoofStyle_Gable, Neighborhood_StoneBr, MSSubClass

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?

Ans.: After building the Ridge and Lasso regression model, we find that the Lasso regression is performing better. The presence of a large number of features also helps the case for Lasso regression as it automatically does feature selection to remove the unnecessary or less significant features from the model, making the model relatively simpler.

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?

Ans.: Excluding the current five most important features, we are left with Neighborhood_Edwards, LotConfig_CulDSac, BsmtFullBath, Neighborhood_Somerst, MasVnrArea as the next five most important features for Lasso model

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?

Ans.: To make a model robust and generalisable, we have to balance the bias-vairance tradeoff. This can be in multiple ways

- Providing a clean and relevant data to the model
- Proper feature selection to remove redundant data and to avoid overfitting
- Choosing the appropriate model
- Cross-validation
- Regularization
- Hyperparameter tuning

Having a robust and generalisable model makes the accuracy of the model quite high. Even though a model which is highly complex may perform outstandingly on the training dataset, it may fail on the unseen data. However, a simpler, generalised model will have a lower performance compared to the complex model on the training dataset, but will provide a much better result on the unseen data.