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 and for Lasso regression is 0.00006.

If we double the of alpha, the value of coefficients would decrease.

After the value of alpha is double, below are the top 5 predictor variables-

1. Ridge Regression (alpha: 20)
    a. GarageFinish_No Garage
    b. GarageFinish_RFn
    c. GarageFinish_Unf
    d. SaleCondition_Normal
    e. SaleCondition_Others
2. Lasso Regression (alpha: 0.00012)
    a. GarageFinish_No Garage
    b. GarageFinish_RFn
    c. GarageFinish_Unf
    d. SaleCondition_Normal
    e. SaleCondition_Others

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 optimal value of alpha for ridge regression is 10 and for Lasso regression is 0.00006. With this value of alpha, the test score for ridge regression is found to be 0.8762875068524612 and test score for Lasso regression is found to be 0.883834546544181. Here, the score for Lasso regression is more than Lasso regression.

Also, some of the coefficients in Lasso regression is absolutely 0 which reduces computation by reducing number of features.

Considering the above points, I will chose to apply Lasso regression in my case.

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: The important predictor variables after removing the earlier important variables GarageFinish_No Garage,GarageFinish_RFn,GarageFinish_Unf,SaleCondition_Normal and SaleCondition_Others are as follows.

1. Ridge Regression (alpha: 10)
    a. ('KitchenQual_TA', 0.048),
    b. ('GarageType_BuiltIn', 0.051),
    c. ('GarageType_Detchd', 0.052),
    d. ('GarageType_No Garage', 0.054),
    e. ('GarageType_Others', 0.057)
2. Lasso Regression (alpha: 0.00006)
    a. ('KitchenQual_TA', 0.055),
    b. ('GarageType_BuiltIn', 0.065),
    c. ('GarageType_Detchd', 0.071),
    d. ('GarageType_No Garage', 0.076),
    e. ('GarageType_Others', 0.299)

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: The following operations are done on the model that makes it more robust and generalisable-

1. The skewed columns are removed and processed to make normal
2. Null or empty values are replaces with median for columns having outliers and mean for other continuous variables.
3. The columns having less correlation to target variable have been removed
4. The best model out of 5 fold cross validated model using Ridge and Lasso regularization has been chosen as final model. This shows the best model out of the available training set.
5. On plotting mean test and train scores with alpha we could observe similar curve for both train and test data.

As a result of the above points, I could observe a decent score for Ridge and Lasso Regression which are as follows-

1. Ridge Regression Score:
    a. Train Score : 0.9048825126844456
    b. Test Score : 0.8762875068524612
2. Lasso Regression Score:
    a. Train Score : 0.9178195810624109
    b. Test Score : 0.883834546544181

The values of train and test score are very near which shows the model is properly fit on the data and is supposed to predict target values with higher confidence.