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 optimum value of alpha for Ridge Regression is 7 and for Lasso, it is 0.0001.

If we double the value of alpha in Ridge, the model will apply more penalty, thereby making the model more generalized and simpler, thereby increasing the error in the test dataset. The variable coefficient decreases too. The number of significant predictors (with non-zero coefficient) will also reduce from 207 to 206

In Lasso too, when we double the alpha value, the model will penalize more and reduce the variable coefficient to zero. R-square value decreases too. The number of significant predictors (with non zero coefficient) will also reduce from 151 to 126

For Ridge, top 20 most important predictors will be

Variable	**Coeff**
Neighborhood_Crawfor	0.092
OverallQual	0.072
GrLivArea	0.066
SaleCondition_Normal	0.065
Neighborhood_MeadowV	-0.06
CentralAir_Y	0.057
Neighborhood_StoneBr	0.054
PropAge	-0.054
OverallCond	0.053
MSZoning_RL	0.051
SaleType_New	0.049
Foundation_PConc	0.047
Exterior1st_BrkFace	0.046
Condition1_Norm	0.044
TotalBsmtSF	0.043
MSZoning_FV	0.043
LandSlope_Sev	-0.043
Neighborhood_IDOTRR	-0.043
KitchenQual_Fa	-0.041
2ndFlrSF	0.04

For Lasso, the top variables will be

Variable	Coeff
MSZoning_FV	0.162
MSZoning_RL	0.145
Neighborhood_Crawfor	0.142
MSZoning_RH	0.134
Neighborhood_MeadowV	-0.111
GrLivArea	0.11
SaleType_New	0.103
LandSlope_Sev	-0.1
Neighborhood_StoneBr	0.096
HouseStyle_2.5Fin	-0.092
Exterior1st_BrkComm	-0.086
SaleCondition_Normal	0.081
MSZoning_RM	0.08
KitchenQual_Fa	-0.079
SaleCondition_Alloca	0.077
Exterior1st_BrkFace	0.068
OverallQual	0.066
CentralAir_Y	0.064
Neighborhood_BrkSide	0.063
PropAge	-0.059

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 lambda or ridge and lasso regression has been found using list of alpha values to tune the Lasso and Ridge model respectively and then cross validating the same using Grid Search. We then plotted mean test score and mean train score for different values of alpha(x axis) against negative mean absolute error(y axis). We also used the command **print(model_cv.best_params_) and print(model_cv.best_score_)** to print the optimal alpha value having best score.

As we can see, we have too many variables in the data and therefore, feature selection should be preferred here. Therefore, Lasso method should be preferred. Lasso method tries to tune the lambda value with the penalty of reducing the absolute value of the predicting variable coefficient. As Lamda increase while tuning, it tends to move the coefficient towards 0 and make it 0. This way, Lasso does variable selection. If Lambda is increased, it neglects variable with 0 coefficient in the model.

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:

Variables	**Lasso**
GrLivArea	0.121963
SaleType_New	0.101769
OverallQual	0.089083
MSZoning_RM	-0.06898
PropAge	-0.05784

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 the model robust and generalisable, we need to make sure that any variation in data should not affect the model performance much. Also, the model should be able to perform well on the unseen data too. To ensure that it is robust and generalisable, we have to ensure that the model is not complex or else it will overfit the data. This might decrease the accuracy of the model but we need to tune the model to a point that there is a good balance between accuracy and complexity. For this, we can use Ridge and Lasso Regression methods.