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 optimal values of Ridge and Lasso Regression are :-

Ridge : 20

Lasso : 0.001

If the Alpha values for both the regression models are doubled then there seems to be changes in the R^2 values.

Regression	Alpha	R^2
Ridge	20 40	89.4 89.39 ~ 89.4
Lasso	0.001 0.002	89.3 88.9

As we can see, when the values of alphas were doubled there seems to be no change for the Ridge regression R^2 value where as a very little change for Lasso regression.
The most important predictor variables after the change is implemented are :

1. OverallQual
2. GrLivArea
3. Neighborhood_Crawfor
4. Neighborhood_NridgHt
5. Condition1_Norm
6. OverallCond
7. Foundation_PConc
8. GarageArea
9. CentralAir_Y
10. Exterior1st_BrkFace

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 :- We all can see from the model that Ridge model has given the better result that Lasso model so one can go with the Ridge model. But, here I have chosen the **Lasso model** as it is very helpful when it comes to feature selection.

In our current situation, feature selections are more important when it comes to predict prices of the said houses and they play an important role in driving the prices up or down.

If we don't want to get too large coefficients and reduction of coefficient magnitude is one of our prime goals, then we can use **Ridge Regression.**

The model we will choose to apply will depend on the use case. However in our case feature selection plays important role thus we will choose **Lasso regression 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?

Ans :- To solve such discrepancy we can drop the top 5 features in Lasso model and build the model again. We know that our top 5 Lasso predictors were :-

1. OverallQual
2. GrLivArea
3. Neighborhood_Crawfor
4. Neighborhood_NridgHt
5. Condition1_Norm

The new model after dropping these columns gave an alpha of value 0.001 whereas the R^2 score came out to be 87.4%.

Considering this newfound values the next top 5 predictor variables came out to be :-

1. Neighborhood_StoneBr
2. 2ndFlrSF
3. 1stFlrSF
4. Exterior1st_BrkFace
5. LandContour_HLS

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 :- **Accuracy**, a too complex model will have a very high accuracy, overfitting so to speak. To make our model more robust and generalizable, we will have to decrease variance which will lead to some bias. Addition of bias means that accuracy will decrease and we can try to find some bitter sweet balance between both characteristics. The best of both worlds.

We have to make sure a model is robust and generalizable & we have to take care it doesn't **overfit**. Very high variance and a smallest change in data affects the model prediction heavily. Such a model learns the patterns of a training data, but fails to pick up the patterns in unseen test data.

A model is **robust** when any variation in the data does not affect its performance much. In other words, the model should not be too complex in order to be robust and generalizable.

A **generalizable** model is able to adapt properly to new, unseen data, drawn from the same distribution as the one used to create the model.

In general, we have to find strike some balance between model accuracy and complexity. This can be achieved by Regularization techniques like Ridge and Lasso.