1) What is the optimal value of alpha for ridge and lasso regression? What will be the changes in the model if you choose to double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

- Ridge Regression is a technique for analyzing multiple regression data that suffer from multicollinearity. When multicollinearity occurs, least squares estimates are unbiased, but their variances are large so they may be far from the true value. For ridge regression, the optimal value of alpha is 20.
- Lasso regression is a type of linear regression that uses shrinkage. Shrinkage is where data values are shrunk towards a central point, like the mean. The lasso procedure encourages simple, sparse models. In the case of Lasso regression, the optimal value for alpha is 1.
- If we choose to double the value of alpha for both ridge and lasso regression, model complexity will have a greater contribution to the cost. Because the minimum cost hypothesis is selected, this means that higher λ will bias the selection toward models with lower complexity.
- The model which is having high r square of test and train dataset, we will select the features/variables from that model. The variable is selected based on the high coefficient value.

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?

- Lasso regression would be a better option it would help in feature elimination and the model will be more robust.

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?

The five Most important predictor variables are as follows:-

1) GrLivArea
2) OverallQual
3) OverallCond
4) TotalBsmtSF
5) GarageArea

4) How can you make sure that a model is robust and generalizable? What are the implications of the same for the accuracy of the model and why?

- A model needs to be made robust and generalizable so that they are not impacted by outliers in the training data. The model should also be generalizable so that the test accuracy is not lesser than the training score. The model should be accurate for datasets other than the ones that were used during training. Too much weightage should not given to the outliers so that the accuracy predicted by the model is high. To ensure that this is not

the case, the outlier analysis needs to be done and only those which are relevant to the dataset need to be retained. Those outliers which it does not make sense to keep must be removed from the dataset. This would help increase the accuracy of the predictions made bythe model.