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?

For Ridge the optimal value of alpha is 2 and For Lasso the optimal value of alpha is 0.01

When the value of alpha is doubled,

In case of ridge the model gets more generalized as we can see the r2 score for train and test are almost similar

In case of lasso more variables are penalized and In my case the model is generalized more compared to the previous value, If it is further increased more variable will be penalized resulting in underfitting

For Ridge, **OverallQual, RoofMatl, Neighborhood, FullBath, RoofStyle**

For Lasso **OverallQual, RoofMatl, Neighborhood, GrLivArea, SaleType**

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?

When we compare the r2 score for both Lasso and Ridge, Ridge is more generalized compared to Lasso and its doing well for both train and test

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?

OverallQual

RoofMatl

Neighborhood

GrLivArea

SaleType

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?

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 generalisable so that the test accuracy is not lesser than the training score. The model should be accurate for datasets other than the ones which 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 by the model. Confidence intervals can be used (typically 3-5 standard deviations). This would help standardize the predictions made by the model. If the model is not robust , it cannot be trusted for predictive analysis