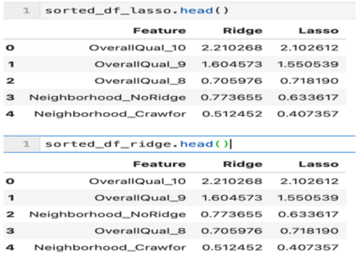
Question 1

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?

A) the optimal value for alpha is 0.1 and 0.001 for ridge and lasso respectively. For Lasso as we double the alpha the R2 score decreases for test and train data. While for the ridge regression the Test and Train R2 score remain the same.
For both ridge and lasso regression the most important predictor remain the same which is OverallQual_10



```
sorted_df_lasso.head()
```

	Feature	Ridge	Lasso
0	OverallQual_10	2.210268	2.102612
1	OverallQual_9	1.604573	1.550539
2	OverallQual_8	0.705976	0.718190
3	Neighborhood_NoRidge	0.773655	0.633617
4	Neighborhood_Crawfor	0.512452	0.407357

```
sorted_df_ridge.head()
```

	Feature	Ridge	Lasso
0	OverallQual_10	2.210268	2.102612
1	OverallQual_9	1.604573	1.550539
2	Neighborhood_NoRidge	0.773655	0.633617
3	OverallQual_8	0.705976	0.718190
4	Neighborhood_Crawfor	0.512452	0.407357

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?

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

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?

A) 5 most important predictors are:

1. OverallQual_10,
2. OverallQual_9,
3. OverallQual_8,
4. Neighborhood_NoRidge,
5. Neighborhood_Crawfor

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) 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.