Question1)

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 1)

Alpha is the regularisation hyper-parameter which decides how much of the additional error (regularisation error) will be added to error term of the equation and thus prevents us from overfitting. The value of Alpha needs to be decided by the user such that we are able to overcome the issue of overfitting but at the same time doesn't make our model to underfit which could happen in the case of higher alpha values. The value of alpha therefore should such that we compromise with low bias a little and the same time reduce the variance in large.

Both Lasso and Ridge tries to make the value the coefficients tend to zero, with lasso even helping to remove various features. As per the formula, if alpha is doubled the regularization error will be doubled.

The most important predictor variables would be the values with their **absolute coefficients** higher than others.

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 2)

As both Ridge and lasso models have similar R2 score on both train and test set, and also both have their error terms around the 0 and follow normal graph. So, both the models are valid and results in similar learning.

	Ridge	**Lasso**
R2 (train)	0.897964	0.898037
R2 (test)	0.852446	0.857896
No of Features (non-zero coef)	137	111

But as Lasso has helped us to remove some of the features, thus making our model less complex with no significant changes in the learning outcome, I will go with Lasso Regression to save some of the computational power.

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 3)

Currently, the five most important features in Lasso are: TotalBsmtSF, YearBuilt, GarageArea, TotRmsAbvGrd, KitchenQual_TA.

Now, after removing these features, the five important features are:

BsmtFinSF1, BsmtUnfSF, FullBath, Neighborhood_StoneBr, LotArea

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 4)

While training a model, the goal is always to decrease the error, but this error reduction on a train data can lead to over fitting of the model over the training data, that is, the model tends to work really well on the training data but fails to perform when fitted on unknown data. This leads to low bias but high variance model, which is not the desired model. A model should be such that it is robust and perform similar on unknown data and for this, the model should be regularised. Also, if we have a large data set, we can keep some data, validation data, outside train and test range, which is to be applied to the model after finalisation.

When we regularize the model, we tend to gain some bias but at the same time loose very much of variance. Thus, it might affect little accuracy, but the same model tends to outperform in the real world, that is, it is much more accurate in real world problem.