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?

The optimal value of alpha for ridge and lasso regression is 500.
When we doubled the alpha value to 1000, the R2 scores decreased.
GrLivArea, OverallQual, GarageCars etc. the most important predictor variables after the change is implemented.

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?

We have determined that the optimum value of Alpha as 500 using GridSearchCV method. Using this value we can be sure that we are neither overfitting nor underfitting the model. Since Ridge regression shows better score in test set as well, we are choosing Ridge regression.

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?

As per our model building, we determined that we are going to use Ridge model for our predictions. Top five predictors from this model are:

GrLivArea
OverallQual
1stFlrSF
Neighborhood
RoofMatl

So if we can’t use data from these columns, we will use the next 5 most important predictors which are :

GarageCars
FullBath
2ndFlrSF
TotRmsAbvGrd
BsmtExposure

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?

The Ridge model which we chose has the R2 score of 90% on training data and 85% on test data. Since, these scores are very close, we can be sure that the model won’t overfit or underfit and will give us generalised results. Ridge also has helped us regularise the model by penalising the overfitting predictors. So we are not affected by outliers in the future test data. We have used different data for training and tested the model on a different set.
We can also split the given data in train, validate and test data sets, if we have enough data to test it.

If the model is not robust and generalisable, the accuracy of the model gets affected by any new outlier in the test data or unseen test scenarios.