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?

Alpha for Ridge: 100
Alpha for Lasso: 1000

R2 is decreased when we double the value of Alpha for both training and test dataset but not much significant changes. RSS is increased for both Ridge and Lasso.

After the change most important predictor variables are for lasso:

GrLivArea
OverallQual
Neighborhood_NridgHt
ExterQual
Neighborhood_NoRidge

After the change most important predictor variables are for Ridge:

GrLivArea
OverallQual
Neighborhood_NridgHt
Neighborhood_NoRidge
ExterQual

Neighborhood_NoRidge	8308.029	8015.157
ExterQual	7864.222	8676.448
Neighborhood_NridgHt	9589.815	9692.028
OverallQual	16442.25	22983.29
GrLivArea	20455.27	25045.54

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?

I will use lasso regression as both Train and Test R2 score is same. So it is more generalised on unseen data compare to ridge and also it has feature selection when more features involved in the modelling. RMSE and RSS on test data is less compared to Ridge regression. That means it produces less errors on unseen data.

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?

1. ExterQual
2. TotalBsmtSF
3. GarageArea
4. Fireplaces
5. MasVnrArea

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?

Checking the difference between test and train r2 score. The difference between both should be less. The difference is less means the model performs well on the unseen data. The model should be robust and accurate for datasets other than the ones which were used during training. The outliers in the dataset should be handled properly to get the more accuracy of the model.