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?

During ridge and lasso usage, the optimal value of alpha depends on amount of regularization needed. As a part of regularization, the beta coefficients shrink. A trade-off is made between bias and variance to make this happen such that model overfitting is reduced a bit.

If value of alpha is doubled, there is no change in the predictor variables however their shrinkage increases a bit (in majority of cases) and the R2 score decreases with slight increase in RSS and MSE thereby decreasing effectiveness of model.

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?

For the given problem, I will choose ridge as most of the independent variables are retained even by lasso. Also, for ridge, the difference between train and test R2 is little compared to lasso. RSS and MSE are marginally better for ridge.

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?

The optimal model built initially had only 8 variables. Hence created new model for this question. The top 5 predictors for this were GrLivArea, OverallQual , TotalBsmtSF , MasVnrArea, GarageArea . After exclusion of these, the new 5 most important predictor variables are ExterQual , KitchenQual, BsmtQual , Fireplaces_2 , binned_LotArea_3 .

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 robust and generalizable works reasonably well on training and unseen data.

Such a model is evaluated based on performance criteria like R2 , RSS , MSE which give indication of discrepancy between training and testing results. In this process, things like proper variable selection are achieved through correlations, outliers and multicollinearity analysis.

To build such a model, bias-variance trade-off is made which decreases accuracy of the model on training data and variance increases a bit. This compromise is necessary as while training the model limited data is available however in production data might have different scale and variation.