Question No. 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?

Answer: The optimal value for ridge regression is 10 and the optimal value for lasso regression is 100. As we know the lambda parameter used to control how much the regularization needs to be applied on the model. Higher the value of lambda, a greater number of coefficients will get pushes towards the zero as we know, regularization parameter increases, it reduces the value of coefficients. Along with we need to also keep in mind that if your lambda values are too high, your model could become simpler, it won't explore or learn enough on the training data for making the prediction, which leads to underfitting. And if we keep the value of lambda too low, the model becomes complex and it will lead to overfitting.

The predictor are as follows:

1. LotArea
2. GrLiveArea
3. 1stFlrSF
4. OverallQual
5. OverallCond
6. MSZoning_RH
7. MSZoning_RL

Question No. 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: I would go with Lasso regression as the value of the lambda for the same regression looks decent enough which is 0.0001 and it has advantage of doing the important feature selection too.

Having said that, considering the assignment part, it has a greater number of features almost 250, looking at such huge number of feature Lasso performs better.

Question No. 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: The five most important predictor variables for excluding are listed below:

1. GrLivArea
2. LotArea
3. OverallQual
4. 1stFlrSF
5. TotalBsmtSF

Question No. 4

How can you make sure that a model is robust and generalized? What are the implications of the same for the accuracy of the model and why?

Answer: The robustness of the model is when the accuracy rate of the model does not change drastically when there is a change in the input variables. The model should be as simple as possible for making it more generalisable, post building the model as simple as possible which will have a high bias and low variance.

Having said that, high bias means the model will likely to make more error as the bias is nothing but the error in the model, which will lead to perform the model on training data and test data as well. Having high bias mean the model will struggle or will not be able to explore the data or learn the details of data very well which leads to low performance of the model.

However, low variance means that the model will perform pretty good on test and train data as the Variance in the model is nothing but the fact of how well the data perform on test and train. Higher variance means the model will perform best on the training data set but when it comes to run on the test data set it will perform poor. Which means the variance of the model is where the data performs well in the seen data (training data) and gives huge difference in the unseen or new data (test data).

The robustness of the model plays good role when it comes to deploy the model in production.