**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?

**Answer 1:**

Optimal Value of alpha for ridge and lasso regression are:

• Optimal Value of lambda for Ridge: 5

• Optimal Value of lambda for Lasso: 0.0004

If we choose to double the value of alpha for both ridge and lasso:

In case of ridge that will lower the coefficients and in case of Lasso there would be more less important features coefficients turning 0.

The most important predictor variable after the change is implemented are those which are significant.

**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:**

The optimal lambda value in case of Ridge and Lasso is as below:

- Ridge - 5
- Lasso - 0.0004

The Mean Squared error in case of Ridge and Lasso are:

- Ridge - 0.013743
- Lasso - 0.013556

The Mean Squared Error of Lasso is slightly lower than that of Ridge.

Also, since Lasso helps in feature reduction (as the coefficient value of one of the features became 0), Lasso has a better edge over Ridge.

Therefore, the variables predicted by Lasso can be applied to choose significant variables for

predicting the price of a house.

**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:**

Drooped most 5 important predictor variables 'MSZoning_RL', 'MSZoning_RM', 'GrLivArea', 'OverallQual', 'MSZoning_FV'

Post dropping most five important now we found below five most important predictor.

1. 2ndFlrSF
2. TotalBsmtSF
3. 1stFlrSF
4. OverallCond
5. Foundation_PConc

**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 creating the best model for any problem statement, we end up choosing from a set of models that would give us the least test error. Hence, the test error, and not only the training error, needs to be estimated in order to select the best model. This can be done in the following two ways:

1) Use metrics that take into account both the model fit and its simplicity. They penalise the model for being too complex (i.e., for overfitting) and, consequently, more representative of the unseen 'test error'. Some examples of such metrics are adjusted R2, AIC and BIC.
2) Estimate the test error via a validation set or a cross-validation approach. In the validation set approach, we find the test error by training the model on a training set and fitting on an unseen validation set, while the in n-fold cross-validation approach, we take the mean of errors generated by training the model on all folds except the kth fold and testing the model on the kth fold, where k varies from 1 to n.

Here are some changes you can make to your model:

1) Use a model that's resistant to outliers. Tree-based models are generally not as affected by outliers, while regression-based models are. If you're performing a statistical test, try a non-parametric test instead of a parametric one.
2) Use a more robust error metric. switching from mean squared error to mean absolute difference (or something like Huber Loss) reduces the influence of outliers.