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 – Optimal value of alpha for ridge and lasso regression

- Ridge Alpha 1
- lasso Alpha 10

For alpha =1 (Optimal values) below are the evaluation metrics

- Adj R2 Score for train data is using Ridge Reg: 0.8877179065591213
- Adj R2 Score for test data is using Ridge Reg: 0.8666735446964924
- Sum of squared error train: 599030263305.9067
- Sum of Mean squared error train: 676870354.0179737
- Sum of Root Mean squared error train: 26016.732193301556
- Sum of squared error test: 227787554132.0808
- Sum of Mean squared error test: 599440931.9265285
- Sum of Root Mean squared error test: 24483.482838977965

For alpha =2 (Doubling) below are the evaluation metrics

- Adj R2 Score for train data is using Ridge Reg: 0.8857904700835554
- Adj R2 Score for test data is using Ridge Reg: 0.8681741109142654
- Sum of squared error train: 609313227793.6643
- Sum of Mean squared error train: 688489522.9306941
- Sum of Root Mean squared error train: 25850.742981398016
- Sum of squared error test: 225223844568.3513
- Sum of Mean squared error test: 592694327.8114507
- Sum of Root Mean squared error test: 25074.768421149278

Not much difference observed in doubling the alpha value for Ridge regression, only slight variation in RMSE. RMSE is increasing for test dataset and decreasing in train set when alpha is doubled, not enough difference in r2_score for test and train dataset

Lot Area most important predictor value

For alpha =10 (Optimal values) below are the evaluation metrics

- Adj R2 Score for train data is using Lasso Reg: 0.8890777114075131
- Adj R2 Score for test data is using Lasso Reg: 0.8630238156295627

For alpha =20 (Doubling) below are the evaluation metrics

- Adj R2 Score for train data is using Lasso Reg: 0.8888717786145917
- Adj R2 Score for test data is using Lasso Reg: 0.8655076046027884

Not Enough differences observed in the R2_score, Lot Area is again most important predictor

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 – R2_Score for test dataset is slightly better for Ridge Regression so I will be using Ridge Regression as the final model to find the features

Question 3

After building the model, you realized 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 – Top 5 predictors after removing the earlier top 5 are

- TotalBsmtSF
- GrLivArea
- GarageArea
- TotRmsAbvGrd
- Foundation_Slab

Question 4

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

Answer – The model should be generic, but it shouldn't underfit as well, a trade-off needs to be done in bias and variance and to get an optimal solution Regularization can be done. There are 2 regularization methods Ridge and Lasso both helps in getting the optimal features.

If the model is overfitting the train dataset then it won't behave well with the test dataset , similar if the model in underfitting than the train data itself doesn't give a very good r2 score which means it not the best model as it doesn't explain the variation in dependent variable done by the independent variable