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?

Ans: The optimal value of alpha for Ridge is 2 and for Lasso it is 0.0005. With these alphas the R2 of the model was approximately 0.85. After doubling the alpha values in the Ridge and Lasso, the prediction accuracy remains around 0.8 but there is a small change in the co-efficient values. The new model is created and demonstrated in the Jupiter notebook.

Below are the changes in the co-efficients.

	Ridge Co-Efficient
LotArea	0.305458
OverallQual_10	0.287260
OverallQual_9	0.271877
GarageArea	0.242461
FullBath	0.233263
TotRmsAbvGrd	0.233186
Neighborhood_Crawfor	0.200674
LotFrontage	0.189872
OverallQual_8	0.172335
Fireplaces	0.164290
WoodDeckSF	0.135467
Neighborhood_StoneBr	0.132718
GarageCars	0.123897
OpenPorchSF	0.122857
Condition1_PosN	0.114455
YearRemodAdd	0.109267
ScreenPorch	0.109102
BedroomAbvGr	0.107873
PoolArea	0.106615
HalfBath	0.105333

	Ridge Doubled Alpha Co-Efficient
LotArea	0.248086
OverallQual_9	0.247373
TotRmsAbvGrd	0.240529
OverallQual_10	0.239825
FullBath	0.221067
GarageArea	0.219435
Neighborhood_Crawfor	0.188898
Fireplaces	0.169610
LotFrontage	0.165822
OverallQual_8	0.164717
GarageCars	0.144385
WoodDeckSF	0.132021
Neighborhood_StoneBr	0.121409
OpenPorchSF	0.117769
BedroomAbvGr	0.113068
YearRemodAdd	0.111127
HalfBath	0.104218
ScreenPorch	0.098520
Condition1_PosN	0.095393
Foundation_PConc	0.092875

	Lasso Co-Efficient
LotArea	0.398696
OverallQual_10	0.323633
OverallQual_9	0.300729
GarageArea	0.282321
TotRmsAbvGrd	0.281028
FullBath	0.230077
Neighborhood_Crawfor	0.188842
OverallQual_8	0.187073
Fireplaces	0.172857
LotFrontage	0.168334
WoodDeckSF	0.123574
YearRemodAdd	0.108176
Neighborhood_StoneBr	0.106268
HalfBath	0.103487
OpenPorchSF	0.098663
ScreenPorch	0.093594
Condition1_PosN	0.089163
Foundation_PConc	0.086146
OverallQual_7	0.084210
GarageCars	0.079243

	Lasso Doubled Alpha Co-Efficient
LotArea	0.422067
OverallQual_10	0.346174
OverallQual_9	0.304160
GarageArea	0.287176
TotRmsAbvGrd	0.247950
FullBath	0.239864
Neighborhood_Crawfor	0.202760
LotFrontage	0.197958
OverallQual_8	0.185755
Fireplaces	0.161886
Neighborhood_StoneBr	0.129217
WoodDeckSF	0.125754
Condition1_PosN	0.123709
PoolArea	0.122403
OpenPorchSF	0.111789
ScreenPorch	0.108992
YearRemodAdd	0.107236
HalfBath	0.104285
Foundation_PConc	0.092558
3SsnPorch	0.088179

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?

Ans:
The optimum lambda value in case of Ridge and Lasso is as follows:-

- Ridge – 2
- Lasso – 0.0005

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

- Ridge - 0.025454596256614572
- Lasso - 0.026171161948074582

The Mean Squared Error of both the models are almost same.
Since Lasso helps in feature reduction (as the coefficient value of some of the features become zero), Lasso has a better edge over Ridge and should be used as the final model.

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?

Ans:

The five most important predictor variables in the current lasso model is:-

1. Overall Qual
2. Total Garage Area
3. Total rooms above grade
4. Full bathrooms above grade
5. Neighbourhod Crawford

We build a Lasso model in the Jupiter notebook after removing these attributes from the dataset. The R2 of the new model without the top 5 predictors drops to 0.81 The Mean Squared Error increases to 0.03186

The new Top 5 predictos are:-

	Lasso Co-Efficient
LotFrontage	0.462279
FullBath	0.364248
GarageCars	0.361037
Exterior1st_Stone	0.236592
BedroomAbvGr	0.236231

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?

Ans:

Model should be as simple as possible, though its accuracy will decrease but it will be more robust and generalizable. It can be understood by bias-variance tradeoff. The simpler the model, more will bias but less variance and more generalizable will be the model. It will perform equally on both training and test data set It is important to have balance between bias and variance to avoid overfitting and underfitting.

Regularization can be used to make the model simpler. Regularization helps to strike the delicate balance between keeping the model simple and not making it too naive to be of any use. For regression, regularization involves adding a regularization term to the cost that adds up the absolute values or the squares of the parameters of the model.