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 : The Optimal Value for alpha in our case are as follows :

- Optimal Value for Ridge : 4.0
- Optima Value for Lasso : 0.0001

If we were to double the value of the alpha i.e 8.0 For Ridge and 0.0002 for Lasso the following changes will come up.

- **Ridge :** There is a slight drop in the R2 and the coefficients or beta values of each of the features. Importance of some of the features change as some betas fall slightly more than others . See the Changes mentioned below in the top 10 features and their Betas .

Alpha : 4.0		Alpha 8.0	
Column	**Beta**	**Column**	**Beta**
GrLivArea	0.234	GrLivArea	0.213
TotalBsmtSF	0.182	1stFlrSF	0.166
1stFlrSF	0.174	TotalBsmtSF	0.151
OverallCond_FA	-0.149	BsmtFinSF1	0.127
BsmtFinSF1	0.144	OverallCond_FA	-0.119
LogLotArea	0.134	OverallQual_EX	0.113
OverallQual_EX	0.124	LogLotArea	0.11
2ndFlrSF	0.118	Neighborhood_Crawfor	0.1
Age	-0.112	2ndFlrSF	0.1
Neighborhood_Crawfor	0.111	OverallQual_VG	0.096

- **Lasso:** There is a slight drop in the R2 value of the model and there are slight increases in the coefficients or Beta values of the features. Some of the features have a slightly higher increase than the others causing a change in their importance to change See the Changes mentioned below in the top 10 features and their Betas.

Alpha : 0.0001		Alpha : 0.0002	
Column	**Beta**	**Column**	**Beta**
GrLivArea	0.46	GrLivArea	0.482
TotalBsmtSF	0.42	TotalBsmtSF	0.406
MSZoning_FV	0.285	Age	-0.225
MSZoning_RH	0.281	LogLotArea	0.216
MSZoning_RL	0.232	OverallCond_FA	-0.205
Age	-0.232	OverallQual_EX	0.155
LogLotArea	0.214	MSZoning_FV	0.137
OverallCond_FA	-0.207	OverallQual_VE	0.134
MSZoning_RM	0.193	Neighborhood_Crawfor	0.133
OverallQual_EX	0.153	MSZoning_RH	0.132

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: After building the models we can determine the following things :
R2 of the lasso for the test is slightly better and the model seems more reliable .
Also the lasso algorithm is such that it has the ability to bring the Beta value all the way down to zero making creating a data selection of sorts that reduces a chance of high variance . Therefore I would prefer to use the lasso model as it makes a simpler more explainable 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?

The original 5 Most important predictors were :

1. GrLivArea
2. TotalBsmtSF
3. MSZoning_FV
4. MSZoning_RH
5. MSZoning_RL

If we do not have access to there in the incoming data we can drop these and rebuild the model and the new model will give us the 5 most important predictors as :

1. 1stFlrSF - 0.401
2. BsmtFinSF1 - 0.353
3. 2ndFlrSF - 0.349
4. BsmtUnfSF - 0.276
5. BsmtFinSF2 - 0.258

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?
Answers: While we build a model we make sure that the model is robust and more importantly generaliable as that is the only way to make the model immune to outliers and high variance. When it comes to accuracy a general model might not have as high accuracy (R2) on the train data as compared to a more high variance model but when the high

variance model runs on unseen data it becomes extremely unreliable and can’t predict well , we see a sharp drop in the R2 of the test .A generalized model can work well and predict well when shown data that is not from the data set that it was trained in and can mange unseen data well. To prevent the data from being high on variance we must perform outlier treatment .The outliers that do not fit into a general pattern or are exceptional cases should be dropped out of the training data so that the model can learn the regular patterns of the data and is not influenced by high variances.We can use confidence intervals that fall within 3-5 Standard Deviations of the median of the data to make sure that the model isn’t learning data that falls into exceptional cases .

Any model that is not robust or generalizable will not be able to make reliable predictions.