SUBJECTIVE QUESTIONS

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 0.2 and for Lasso it is 0.001. With these alphas the R2 of the model was approximately 0.81. After doubling the alpha values in the Ridge and Lasso, the prediction accuracy remains around 0.81 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.

Lasso Model with Alpha 0.001

	Featuere	Coef
5	MSZoning_RH	0.558070
7	MSZoning_RM	0.548168
10	Neighborhood_MeadowV	0.357946
3	GarageCars	0.337657
1	BsmtQual	0.311058
13	RoofStyle_Shed	0.309493
22	GarageType_Basment	0.299728
9	Neighborhood_Crawfor	0.278108
6	MSZoning_RL	0.249516
12	Neighborhood_Veenker	0.234882

Lasso Model with Alpha 0.002

	Featuere	Coef
7	MSZoning_RM	0.340005
3	GarageCars	0.337108
1	BsmtQual	0.326290
5	MSZoning_RH	0.317185
10	Neighborhood_MeadowV	0.310499
9	Neighborhood_Crawfor	0.229323
2	GrLivArea	0.217947
4	MSZoning_FV	0.199260
22	GarageType_Basment	0.181935
13	RoofStyle_Shed	0.180462

Overall since the alpha values are small, we do not see a huge change in the model after doubling the alpha. But the coefficient value is changed little bit.

Ridge Model with Alpha 0.2

Ridge Model with Alpha 0.4

	Feaure	Coef
5	MSZoning_RH	1.355513
7	MSZoning_RM	1.316307
6	MSZoning_RL	1.078677
8	Neighborhood_ClearCr	0.989553
22	GarageType_Basment	0.601444
24	GarageType_Detchd	0.536558
20	Foundation_Wood	0.429795
13	RoofStyle_Shed	0.423759
10	Neighborhood_MeadowV	0.406971
23	GarageType_BuiltIn	0.351718

	Feaure	Coef
5	MSZoning_RH	1.202830
7	MSZoning_RM	1.168964
6	MSZoning_RL	0.921518
8	Neighborhood_ClearCr	0.841547
22	GarageType_Basment	0.567883
24	GarageType_Detchd	0.501616
13	RoofStyle_Shed	0.415038
20	Foundation_Wood	0.414182
10	Neighborhood_MeadowV	0.403874
3	GarageCars	0.341317

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?

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

- Ridge – 0.2
- Lasso – 0.001

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

- Ridge - 0.273
- Lasso – 0.25

▪ 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), and lasso will penalize more on the dataset and can also help in feature elimination, I am going to consider that as my 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:-

- MSZoning_RH - Residential High Density
- MSZoning_RM - Residential Medium Density
- GarageCars - Size of garage in car capacity
- GarageType_Basment - Garage location Basement Garage
- Neighborhood_MeadowV - Physical locations within Ames city limits

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 .79 The Mean Squared Error increases to 0.2735 The new Top 5 predictos are:-

	Lasso Co-Efficient
MSZoning_FV	0.426716
MSZoning_RL	0.424939
OverallQual	0.380716
GrLivArea	0.375514
Neighborhood_Veenker	0.369899

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?

As Per, Occam's Razor— given two models that show similar 'performance' in the finite training or test data, we should pick the one that makes fewer on the test data due to following reasons:-

▪ Simpler models are usually more 'generic' and are more widely applicable. As simpler model is easy to interpret as compared to complex model.

▪ Simpler models require fewer training samples for effective training than the more complex ones and hence are easier to train and computationally fast as well.

▪ Simpler models are more robust.

- o Complex models tend to change wildly with changes in the training data set
- o Simple models have low variance, high bias and complex models have low bias, high variance
- o Simpler models make more errors in the training set.

Complex models lead to overfitting — they work very well for the training samples, fail miserably when applied to other test samples Therefore, to make the model more robust and generalizable, make the model simple but not simpler which will not be of any use. 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. Also, Making a model simple leads to Bias-Variance Trade-off:

• A complex model will need to change for every little change in the dataset and hence is very unstable and extremely sensitive to any changes in the training data.

• A simpler model that abstracts out some pattern followed by the data points given is unlikely to change wildly even if more points are added or removed.

Bias quantifies how accurate is the model likely to be on test data. A complex model can do an accurate job prediction provided there is enough training data. Models that are too naive, for e.g., one that gives same answer to all test inputs and makes no discrimination whatsoever has a very large bias as its expected error across all test inputs are very high. Variance refers to the degree of changes in the model itself with respect to changes in the training data.