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 = 1.0

Optimal Value of Alpha for Lasso = 0.001

After doubling the Alpha values for both Ridge and Lasso, the train and test accuracy becomes-

Train Accuracy (Ridge) – 82.69%

Test Accuracy (Ridge) – 83.14%

Train Accuracy (Lasso) – 82.47%

Test Accuracy (Lasso) – 83.34%

Top 5 Features after doubling the value:-

Lasso Features	Ridge Features
GrLivArea	GrLivArea
OverallQual	OverallQual
LotArea	LotArea
MasVnrArea	MasVnrArea
GarageArea	GarageArea

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-

Both the Ridge and Lasso have very similar accuracy on test and train data.

Train Accuracy (Ridge) – 82.91%

Test Accuracy (Ridge) – 83.38%

Train Accuracy (Lasso) – 82.85%

Test Accuracy (Lasso) – 83.70%

After determining the optimal values for the alpha, I would choose Lasso Regression model. The reason for it, Lasso is making many features Zero and hence the model will be less complex.

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-

According to current model 5 most important Predictor variables are-

Rank	Features
1	GrLivArea
2	OverallQual
3	LotArea
4	MasVnrArea
5	GarageArea

After dropping above 5 variables, 5 most important Predictor variables are-

Rank	Features
1	TotalBsmtSF
2	2ndFlrSF
3	Fireplaces
4	FullBath
5	OverallCond

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-

The model should be as simple as possible, though its accuracy will decrease but it will be more robust and generalisable. It can be also understood using the Bias-Variance trade-off. The simpler themodel the more the bias but less variance and more generalizable. Its implication in terms of accuracy is that a robust and generalisable model will perform equally well on both training and test data i.e. the accuracy does not change much for training and test data.

Bias: Bias is error in model, when the model is weak to learn from the data. High bias means model is unable to learn details in the data. Model performs poor on training and testing data.

Variance: Variance is error in model, when model tries to over learn from the data. High variance means model performs exceptionally well on training data as it has very well trained on this of data but performs very poor on testing data as it was unseen data for the model.

It is important to have balance in Bias and Variance to avoid overfitting and under-fitting of data.