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 1

Optimal values of alpha are as follows:

For Ridge:100

For Lasso:0.001

Below is the comparison of Ridge and lasso before and after doubling alpha value:

	Metric	Ridge Regression	Lasso Regression	Ridge Regression 2*Alpha	Lasso Regression 2*Alpha
0	R2 Score (Train)	0.890675	0.892416	0.886629	0.891396
1	R2 Score (Test)	0.820788	0.826514	0.811651	0.828895
2	RSS (Train)	17.540602	17.261300	18.189682	17.424935
3	RSS (Test)	12.966679	12.552343	13.627745	12.380057
4	MSE (Train)	0.131072	0.130024	0.133475	0.130639
5	MSE (Test)	0.171863	0.169095	0.176189	0.167930

From above table it was clear that after doubling value of alpha R2 score (Test as well as train) of both models (Ridge and lasso) goes down. Which tells us that performance of model goes down which was also implied by RSS and MSE values as both these values goes up which was not good indication for a model.

Below are the top 5 features after doubling the alpha value:

Ridge:

	Ridge	Lasso	Ridge1	Lasso1
OverallQual	0.086788	0.095713	0.081362	0.097779
GrLivArea	0.071769	0.132441	0.066666	0.124936
1stFlrSF	0.051947	0.014167	0.050636	0.019228
OverallCond	0.052698	0.059190	0.046870	0.057599
GarageCars	0.047496	0.051180	0.045550	0.051775

1. OverallQual: Rates the overall material and finish of the house
2. GrLivArea: Above grade (ground) living area square feet
3. 1stFlrSF: First Floor square feet

4. OverallCond: Rates the overall condition of the house
5. GarageCars: Size of garage in car capacity.

Lasso:

	Ridge	Lasso	Ridge1	Lasso1
GrLivArea	0.071769	0.132441	0.066666	0.124936
OverallQual	0.086788	0.095713	0.081362	0.097779
OverallCond	0.052698	0.059190	0.046870	0.057599
GarageCars	0.047496	0.051180	0.045550	0.051775
FireplaceQu	0.035821	0.033263	0.037292	0.033392

1. GrLivArea: Above grade (ground) living area square feet
2. OverallQual: Rates the overall material and finish of the house
3. OverallCond: Rates the overall condition of the house
4. GarageCars: Size of garage in car capacity.
5. FireplaceQu: Fireplace quality

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?

	Metric	Ridge Regression	Lasso Regression
0	R2 Score (Train)	0.890675	0.892416
1	R2 Score (Test)	0.820788	0.826514
2	RSS (Train)	17.540602	17.261300
3	RSS (Test)	12.966679	12.552343
4	MSE (Train)	0.131072	0.130024
5	MSE (Test)	0.171863	0.169095

Based on above metrics we will choose Lasso regression as it will be giving better results on test data as compared to ridge regression model.

R2 score of Lasso regression is better than ridge regression and in case of lasso we have less number of dependent variables.

RSS and MSE values of Lasso regression model is lower than ridge one which is better in terms of performance.

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 3

After removing top 5 features from our model we get below R2 score and RSS, MSE values, which are less from actual model.

```
R2 Score Train :  0.84
R2 Score Test :  0.72
RSS of Train :  25.22
RSS of Test :  20.37
MSE of Train :  0.02
MSE of Test :  0.05
```

After building model we get below 5 features as most important features:

	Lasso
1stFlrSF	0.153015
2ndFlrSF	0.141959
FireplaceQu	0.050811
Neighborhood_NridgHt	0.046382
BsmtFullBath	0.032180

1stFlrSF (First Floor square feet) is the most important feature as it has biggest coefficient value. After 1stFlrSF, 2ndFlrSF (Second floor square feet) is the most

important one. Followed by FireplaceQu (Fireplace quality), Neighborhood_NridgHt (Physical locations within Ames city limits- Northridge Heights) and BsmtFullBath (Basement full bathrooms).

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 4

For making a robust model we should take care below mentioned points:

1. Our model does not contain Outliers.
2. Using robust evaluation metric
3. Transform data to reduce impact of outliers
4. Over fitting on training data must be avoided

All of these will have very much impact on accuracy of model:

1. If our data have outliers then our model tries to fit them, in fitting them it will not bale to predict normal data hence accuracy will be low.
2. Robust evaluation metrics will help in building model more accurate.
3. Transforming data will help in reducing impact of outliers, which will help in building better model.
4. As our model tries to overfit train data or memorising training data will increase chance of model failure on unseen data very much hence accuracy will be impacted very much.