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:

Optimal value of alpha:

- For Ridge regression: 10

```
- Ridge
        - Ridge regression train r2: 0.9183
        - Ridge regression test r2:  0.9012
#top 10 variable coefficients are:
```

	Features	rfe_support	rfe_ranking	Coefficient
11	MSZoning_RL	True	1	0.0877
5	GrLivArea	True	1	0.0791
1	OverallQual	True	1	0.0684
9	MSZoning_FV	True	1	0.0594
12	MSZoning_RM	True	1	0.0580
4	TotalBsmtSF	True	1	0.0462
2	OverallCond	True	1	0.0450
14	Foundation_PConc	True	1	0.0434
7	GarageCars	True	1	0.0355
3	BsmtFinSF1	True	1	0.0328

- For Lasso Regression: 0.0004

```
Lasso Regression train r2 :  0.9187
Lasso Regression test r2 :  0.9026

#top 10 variable coefficients are:
```

	Features	rfe_support	rfe_ranking	Coefficient
11	MSZoning_RL	True	1	0.109840
5	GrLivArea	True	1	0.107541
12	MSZoning_RM	True	1	0.076511
9	MSZoning_FV	True	1	0.070579
1	OverallQual	True	1	0.069655
4	TotalBsmtSF	True	1	0.046112
2	OverallCond	True	1	0.044889
14	Foundation_PConc	True	1	0.043031
7	GarageCars	True	1	0.036338
3	BsmtFinSF1	True	1	0.033366

- For Ridge regression alpha is 10 and now doubling it and making it 20.

```
Ridge Regression train r2:  0.9173
Ridge Regression test r2:  0.9013
```

	Features	Coefficient	Mod
0	MSSubClass	11.995980	11.995980
28	d_HeatingQC	0.087714	0.087714
9	1stFlrSF	0.079058	0.079058
3	OverallQual	0.068363	0.068363
26	d_BsmtExposure	0.059410	0.059410
29	d_KitchenQual	0.058046	0.058046
6	BsmtFinSF1	0.046223	0.046223
4	OverallCond	0.045040	0.045040
47	Neighborhood_Crawfor	0.043378	0.043378
13	FullBath	0.035473	0.035473

- For Lasso regression alpha is 0.0004 and now doubling it and making it 0.0008.

```
Lasso Regression train r2:  0.9174
Lasso Regression test r2:  0.904
```

	Feature	Coef	mod
0	MSSubClass	11.995935	11.995935
28	MSZoning_RM	0.109840	0.109840
9	BsmtFullBath	0.107541	0.107541
29	Neighborhood_Crawfor	0.076511	0.076511
26	MSZoning_RH	0.070579	0.070579
3	OverallCond	0.069655	0.069655
6	1stFlrSF	0.046112	0.046112
4	BsmtFinSF1	0.044889	0.044889
47	Foundation_Slab	0.043031	0.043031
13	WoodDeckSF	0.036338	0.036338

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: As of my opinion I prefer LASSO Regression, because L1 tends to if there are small no.of.significant parameters and the others are close to 0's.

Also we can observe the Mean Squared Error (MSE) of Lasso is slightly lower than that of Ridge and we know that Lasso helps in feature reduction (as the coefficient value of one of the feature became 0), Lasso has a better edge over Ridge.

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: As I realised the 5 most imp predictors variables in the lasso model are not available in the data set, then i drop those first 5 predictors and apply the regression with rest of the data.

After dropping the 5 predictors my model has:

```
Lasso Regression train new r2:  0.9005
Lasso Regression test new r2:  0.875
```

Now the five most imp predictors:

	Features	rfe_support	rfe_ranking	Coefficient
2	2ndFlrSF	True	1	0.110301
1	1stFlrSF	True	1	0.091497
0	TotalBsmtSF	True	1	0.059860
4	Foundation_PConc	True	1	0.048156
3	GarageCars	True	1	0.038415

Question 4: How can you make sure that a model is robust and generalizable? What are the implications of the same for the accuracy of the model and why?

Ans: Simpler Models make more errors in the training data set. Complex models leads to Over fitting, they work very well for the training data, fail miserably when the applied to other test data.

For regression, regularization involves adding a regularization term to cost that adds up the absolute values or the squares of the parameters of the model.

To make the model more robust and generalizable,

- Make the model that the data are not impacted by outliers in the training data.

- Our model should be accurate for datasets other than the ones which were used during training the data.

Ex: Use the given data set as 70% as training the data and the use rest 30% of data as predicting the accurate estimates for the testing data.

- Our model should also be generalizable so that the test accuracy is not lesser than the training score.

In our problem: The Lasso Regression train r2: 0.9174, Lasso Regression test r2 : 0.904 by observing the results the above point has justified.

- To ensure that our Final Model is not the case, the outlier's analysis needs to be done and only those which are relevant to the dataset need to be retained.

- Make sure the Final Model leads to Bias-Variance Trade-off.

- A complex model leads to change for every little change in the dataset and hence it is unstable and extremely sensitive to have any changes in the training data.

- As in the case of simple model leads to abstracts out some pattern followed by the data points given is unlikely to change wildly even if more points are added or removed.