Subjective Questions

Question 1

A) What is the optimal value of alpha for ridge and lasso regression?
B) What will be the changes in the model if you choose double the value of alpha for both ridge and lasso?
C) What will be the most important predictor variables after the change is implemented?

Answer 1

A) Optimal Value :
- Ridge Regression : 5
- Lasso Regression : 0.0001

B) Changes on doubling the value of optimal alpha :

1. Ridge Regression

<table>
<tr><th></th><th>Alpha=10</th><th>Alpha=5</th><th>Comments</th></tr>
<tr><td>Model Metrices</td><td colspan="2">

	Metric	Doubled Alpha	Optimal Alpha
0	R2 Score (Train)	0.903336	0.914121
1	R2 Score (Test)	0.829566	0.831418
2	RSS (Train)	1.206097	1.071524
3	RSS (Test)	0.930204	0.920098
4	MSE (Train)	0.035556	0.033514
5	MSE (Test)	0.047632	0.047372

</td><td>For both train and test data the R2 score value has dropped and MSE has increased</td></tr>
<tr><td>Model Coefficients</td><td>

	Feaure	Coef
5	OverallCond	0.077117
24	BsmtFullBath	0.072381
21	2ndFlrSF	0.066169
14	BsmtFinType2	0.062598
18	HeatingQC	0.055564
22	LowQualFinSF	0.050051
32	Functional	0.040508
31	TotRmsAbvGrd	0.040274
91	Neighborhood_OldTown	0.039073
90	Neighborhood_NridgHt	0.038377

</td><td>

Feaure	Coef
OverallCond	0.092525
BsmtFullBath	0.089870
2ndFlrSF	0.078730
BsmtFinType2	0.074283
LowQualFinSF	0.066545
HeatingQC	0.066145
LotShape	0.048120
MasVnrArea	0.041537
Neighborhood_OldTown	0.040970
TotRmsAbvGrd	0.040548

</td><td>On doubling alpha the model coefficients have reduced.</td></tr>
</table>

2. Lasso Regression

<table>
<tr><th></th><th>Alpha=0.0002</th><th>Alpha=0.0001</th><th>Comments</th></tr>
<tr><td>Model Metrices</td><td colspan="2">

	Metric	Doubled Alpha	Optimal Alpha
0	R2 Score (Train)	0.910564	0.917985
1	R2 Score (Test)	0.812248	0.814004
2	RSS (Train)	1.115911	1.023321
3	RSS (Test)	1.024726	1.015142
4	MSE (Train)	0.034201	0.032752
5	MSE (Test)	0.049993	0.049759

</td><td>for both train and test data the R2 score has reduced slightly and the MSE has increased slightly</td></tr>
<tr><td>Model Coefficients</td><td>

	Feaure	Coef
24	BsmtFullBath	0.325532
5	OverallCond	0.139076
18	HeatingQC	0.099737
14	BsmtFinType2	0.075293
2	LotShape	0.065796
33	Fireplaces	0.044963
91	Neighborhood_OldTown	0.044452
6	MasVnrArea	0.043808
31	TotRmsAbvGrd	0.042036
90	Neighborhood_NridgHt	0.033688

</td><td>

	Feaure	Coef
24	BsmtFullBath	0.315240
5	OverallCond	0.131956
18	HeatingQC	0.115684
2	LotShape	0.085616
14	BsmtFinType2	0.076192
6	MasVnrArea	0.054726
33	Fireplaces	0.051725
91	Neighborhood_OldTown	0.044104
31	TotRmsAbvGrd	0.038309
7	ExterQual	0.036576

</td><td>model coefficients has increased</td></tr>
<tr><td>Number of features</td><td>77</td><td>100</td><td>reduced from 100 to 77</td></tr>
</table>

C) Most important predictor variables after change is implemented :
Top 10 predictors will be :

1. **Ridge Regression**

	Feaure	Coef
5	OverallCond	0.077117
24	BsmtFullBath	0.072381
21	2ndFlrSF	0.066169
14	BsmtFinType2	0.062598
18	HeatingQC	0.055564
22	LowQualFinSF	0.050051
32	Functional	0.040508
31	TotRmsAbvGrd	0.040274
91	Neighborhood_OldTown	0.039073
90	Neighborhood_NridgHt	0.038377

2. **Lasso Regression**

	Feaure	Coef
24	BsmtFullBath	0.325532
5	OverallCond	0.139076
18	HeatingQC	0.099737
14	BsmtFinType2	0.075293
2	LotShape	0.065796
33	Fireplaces	0.044963
91	Neighborhood_OldTown	0.044452
6	MasVnrArea	0.043808
31	TotRmsAbvGrd	0.042036
90	Neighborhood_NridgHt	0.033688

3.

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 2

I will choose **Lasso** as it gives the option of **feature selection** along with regularization. It removes unwanted features from the model without affecting the model accuracy. In Lasso, some of the coefficients become 0, thus resulting in feature selection and, hence, easier interpretation, particularly when the number of coefficients is very large.

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

The top 5 important variables in Lasso Regression are :

1. BsmtFullBath
2. OverallCond
3. HeatingQC
4. LotShape
5. BsmtFinType2

On excluding these five, the top 5 variables in the new Lasso model are :

Feaure	Coef
BsmtHalfBath	0.316481
MasVnrArea	0.136221
CentralAir	0.102046
LandSlope	0.082873
BsmtFinSF2	0.076203

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

A robust and generalisable model is one which has low training error and low testing error.
To make such model following things are essential:

1. **Removing outliers in the train data** → This was done as part of data cleaning in EDA process. Though outlier removal is not always significant, specially where the sample size is small. Doing a drastic outlier removal, may reduce the sample points.
2. **Avoiding overfitting by doing Regularization while model making**. In overfitting, a model fits the training data but fails to generalize and hence, cannot be used as the model to predict on new data or out-of-sample data. Regularization helps to avoid overfitting as well underfitting, keeping bias & variance trade off at its best. We use regularization because we want our models to work well with unseen data, without missing out on identifying underlying patterns in the data

Implications on model accuracy

By making robust and generalized model i.e. by introducing regularization we compromise accuracy to some extent as we allow a little bias for a significant reduction in variance.

Reason for this implication

This happens because **Regularization** introduces a penalty, which grows in relation to the size of the coefficients and reduces its impact, thus making the model less sensitive to small changes in the variables. More extreme model coefficients values gives better accuracy but lead to a large variance. Regularization prevents this by shrinking the coefficients towards 0.