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 values of alpha for Ridge and Lasso are respectively: **1.0 and 0.0001.**

When you double the values, we get: **2.0 and 0.0002.**

In Ridge Regression:

	Feaure	Coef
7	GarageCars	0.224389
3	OverallCond	0.196580
5	1stFlrSF	0.130284
6	GrLivArea	0.098723
39	BrDale	0.080271
2	OverallQual	0.075492
40	Crawfor	0.064669
42	NoRidge	0.055220
32	ImStucc	0.051080
4	BsmtFinSF1	0.048343

The only change is the position of features changing and the addition of **BsmtFinSF1**.

Now in Lasso Regression

	Features	rfe_support	rfe_ranking	Coefficient
4	GrLivArea	TRUE	1	0.367935
1	OverallQual	TRUE	1	0.224231
3	BsmtFinSF1	TRUE	1	0.165624
7	NoRidge	TRUE	1	0.075015
8	NridgHt	TRUE	1	0.056637
2	OverallCond	TRUE	1	0.048697
9	New	TRUE	1	0.037538
6	CmentBd	TRUE	1	0.030385
5	GarageCars	TRUE	1	0.029250
0	MSSubClass	TRUE	1	-0.029771

There is nothing much change in R2 score, RSS, MSE values in Ridge and Lasso after doubling the alpha value. However, the top features are listed above for the Lasso Regression.

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:

Well, I prefer the Lasso Regression. I have two reasons for it.

- The first one is feature elimination in the Regression
- The second one is the R2 score for the train and test sets are pretty decent with the optimal alpha value

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:

We dropped the top five important variables. And built the Lasso Regression again.

	Featuere	Coef
7	BsmtUnfSF	0.491422
3	ExterCond	0.200746
5	BsmtFinSF1	0.180619
2	OverallCond	0.136244
49	Twnhs	0.084249
48	Duplex	0.078103
30	RH	0.077594
40	Norm	0.072597
39	Feedr	0.070625
42	PosN	0.068737

The top five important features are: BsmtUnfSF, ExterCond, BsmtFinSF1, OverallCond, Twnhs.

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?

Ans:

A machine learning model is robust when its dependent variable doesn't affect much if there is a change in one or two of its independent variables. And this robustness of a model is generalisable in such scenarios.

So one good way to make sure your model is robust and generalisable is by properly treating the missing values and the outliers in the data. The outliers make the predictions swing widely. With the decent R2 score and Adjusted R2 scores, you will make sure whether a model is generalisable. Or it is learning patterns and thus showing us the overfitting model.

Hence a good test results, or at least not much lesser than the train results, will make sure the model we built is robust and generalisable.

The implications (or conclusions) for the model accuracy is simple as mentioned below.

1. Train the models on enough data. So having enough data is crucial here.
2. Missing values and the outliers in the data cause inaccurate predictions when you employ the model on the test data. Also we need to standardise the data, derive the new columns, scale the variables, etc.
3. Based on the domain knowledge, selecting the features and the right Algorithm help us.
4. And testing the model on some untrained data for cross-validation of the model will help us know how accurate the machine learning model is.