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?

The optimal values of alpha for Ridge is 6.0, and for Lasso is 20.

If we double the Ridge to 12 and Lasso to 20 the change is as follows, the model performance for both Ridge and Lasso reduced a bit.

	Ridge (actual)	Ridge (double)	Lasso (actual)	Lasso (double)
r2_train	0.8900312618894491	0.8767871365387543	0.9361578421952437	0.912246658182735
r2_test	0.8662370118938432	0.8588581382151828	0.8487416428029395	0.8512153096257837
rss_train	701677395242.0	786184169967.3022	407357579660.83417	559927642783.5236
rss_test	377039410671.92505	397838334369.83716	426353826751.55896	419381271082.0961
mse_train	687245245.0950049	770013878.5184156	398979020.2358807	548411011.5411593
mse_test	860820572.3103311	908306699.474514	973410563.3597237	957491486.488804

The most important predictor variable are:

- For Ridge: **OverallQual**
- For Lasso: **GrLivArea**

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?

Ridge works better if there are too many parameters. Lasso performs better if there a fewer number of parameters.

To choose between Ridge and Lasso perform cross-validation to select the performing model.

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?

If we just consider the already built model then it will be as shown below, taking the rest of the fields.

Sno	Predictors	Lasso
1	GrLivArea	246784.5743
2	RoofMatl_WdShngl	136009.1364
3	OverallQual	84325.57165
4	RoofMatl_CompShg	64044.4613
5	RoofMatl_WdShake	48766.63508
6	Neighborhood_NoRidge	**48155.48812**
7	LotArea	**45663.86133**
8	GarageCars	**41930.51733**
9	2ndFlrSF	**40892.28896**
10	OverallCond	**31399.30808**

But in absence of the top five fields ('GrLivArea', 'RoofMatl_WdShngl', 'OverallQual', 'RoofMatl_CompShg', 'RoofMatl_WdShake') and rebuilding without these top five fields we will get the following fields in top place:

Sno	Predictor	Lasso
1	1stFlrSF	213617.6958
2	2ndFlrSF	133944.8264
3	Neighborhood_NoRidge	52794.45352
4	GarageCars	46714.09337
5	Neighborhood_NridgHt	32862.58265

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?

1. Model is able to work with unseen data and able to adapt with new data.
2. Try to keep the model simple and make sure that the model is not overfit and just remember the data.
3. Model should generalize the try to find patterns in data and train and test accuracy are close.