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: The optimal value of alpha for ridge is **477.85** and the optimal value of alpha for lasso is **0.0019**.

This is how the importance of variables change if we double the value of alpha in case of ridge. Note that the following ranking of feature is obtained using RFE.

	features_alpha	ranking_alpha	features_2alpha	ranking_2alpha
0	OverallQual	1	OverallQual	1
1	GrLivArea	2	GrLivArea	2
2	GarageCars	3	GarageCars	3
3	KitchenQual	4	KitchenQual	4
4	BsmtQual	5	BsmtQual	5
5	1stFlrSF	6	1stFlrSF	6
6	YearBuilt	7	YearBuilt	7
7	FireplaceQu	8	Fireplaces	8
8	OverallCond	9	ExterQual	9
9	MSZoning_RL	10	TotRmsAbvGrd	10

Here the top 7 important features are exactly same for both the versions of Ridge, though the ranking of other features have changed slightly.

This is how the importance of variables change if we double the value of alpha in case of lasso. Note that the following ranking of feature is obtained using RFE.

	features_alpha	ranking_alpha	features_2alpha	ranking_2alpha
0	OverallQual	1	OverallQual	1
1	GrLivArea	2	GrLivArea	2
2	YearBuilt	3	YearBuilt	3
3	OverallCond	4	GarageCars	4
4	TotalBsmtSF	5	OverallCond	5
5	RoofMatl_CompShg	6	TotalBsmtSF	6
6	RoofMatl_Tar&Grv	7	MSZoning_RL	7
7	RoofMatl_WdShngl	8	RoofMatl_CompShg	8
8	RoofMatl_Metal	9	RoofMatl_Tar&Grv	9
9	RoofMatl_Membran	10	RoofMatl_WdShngl	10

The top 3 most important features for optimal λ version of Lasso and 2λ version of Lasso are exactly same. They are OverallQual, GrLivArea & YearBuilt. The ranking of all other features are quite different between both the versions of Lasso.

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: I chose to apply lasso model, because its predictive r2_score is 90% (which is slightly better than that of ridge with 88%) and it uses only 161 features thus it is significantly lighter than the ridge 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?

Answer:

	features_alpha	ranking_alpha
0	OverallQual	1
1	GrLivArea	2
2	YearBuilt	3
3	OverallCond	4
4	TotalBsmtSF	5
5	RoofMatl_CompShg	6
6	RoofMatl_Tar&Grv	7
7	RoofMatl_WdShngl	8
8	RoofMatl_Metal	9
9	RoofMatl_Membran	10

There are the top 10 most important variables of lasso model as per RFE. If the first 5 variables are not available in the new incoming data. We'll use the next 5 variables in the above list starting from RoofMatl_CompShg to RoofMatl_Membran.

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:

We can make our regression model more robust and generalizable by reducing its variance by artificially injecting some bias into the model by means of shrinking the length of parameter vector towards 0. This is treat the problem of overfitting and the model predictive power on the unseen data will improve significantly. In case of OLS, we achieve the above by using l1_norm penalty term or the square of l2_norm penalty term. It will give us a suboptimal fit on the training set, but a better fit on the test set.