Question Q1: 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:

1. Linear model was given parameters ranging from lambda/alpha = 0.0001 to 2000, Optimal value of ridge coefficient is lambda = 300, and Optimal value of Lasso coefficient is = 400.
2. If we double value of ridge coefficient from 300 to 600 we see coefficients tend to decrease and order of most important predictor change, this is because lambda/alpha value if large, below is table of important predictor variables after lambda is doubled for ridge:

	Variable name	Ridge	Ridge(Aplha_Double)
2	OverallQual	9821.296938	8375.181620
14	GrLivArea	9484.586120	7816.097587
78	Neighborhood_NoRidge	6771.748191	5892.749054
11	1stFlrSF	6142.900843	5622.106778
99	Condition2_PosN	-7018.787295	-5359.869364
79	Neighborhood_NridgHt	6033.189332	5322.537831
125	RoofMatl_WdShngl	6243.514850	4844.810090
24	GarageCars	5210.520651	4516.514287
12	2ndFlrSF	6033.225258	4452.095187
178	BsmtExposure_Gd	4769.992434	4321.676873

3. Similarly by changing the value of Lass coefficient from 400 to 800 we see coefficients change and order of important predictors change, below is table of important predictor variables after lambda is doubled for Lasso :

	Variable name	Lasso	Lasso(Aplha_Double)
14	GrLivArea	30212.011346	28575.481489
2	OverallQual	12800.431978	16061.472426
99	Condition2_PosN	-11449.009594	-9504.578041
171	BsmtQual_Gd	-11013.008219	-8179.875680
78	Neighborhood_NoRidge	6694.311184	7708.363352
209	KitchenQual_Gd	-11283.968983	-7696.944092
79	Neighborhood_NridgHt	5684.124158	7297.881107
24	GarageCars	5990.074940	7027.251105
210	KitchenQual_TA	-10109.632749	-6722.361512
178	BsmtExposure_Gd	5699.670998	6195.737290

Question Q2: 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:

Model has too many variables and while doing EDA we see that some variables are having high correlation coefficient with other predictor variables, also few variables how very poor or no correlation with target variables.

While doing ridge regression all predictor variables are having coefficients but in case of lasso we see some predictor variables are having zero as its coefficients suggesting they do not exhibit any linear relationship with target variables (SalePrice)

Also we see that test score of both ridge and lasso is in similar range I will prefer using lasso since no of predictor variables are lower and model is more generalizable and also robust compared to ridge regression model

Question Q3: After building the model, you realized 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:
Actual top 5 variables for Lasso regression model are:

	Variable name
119	RoofMatl_CompShg
14	GrLivArea
125	RoofMatl_WdShngl
123	RoofMatl_Tar&Grv
124	RoofMatl_WdShake

Post removal of above top 5 variables from model we see next top 5 variables are:

	Variable name
12	2ndFlrSF
11	1stFlrSF
2	OverallQual
77	Neighborhood_NoRidge
23	GarageCars

Question Q4: 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:
Model is considered robust if small changes in training data doesn't change model coefficients drastically, Model is considered generalizable if it is simple and works well with new data.

If a model is robust and generalizable accuracy score of test and train data should be similar i.e. model should not overfit on train data it should generalize train data so that it is able to do good even on unseen test data.