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 Regression is: 10
The optimal value of alpha for Lasso Regression is: 0.001

After retraining the model with double alpha value for Ridge:
R2 for train decreased from .93 to .92
The RMSE on train data remains 0.1
R2 for test dataset remains unchanged i.e .87
The RMSE on test data remains 0.1

Top 10 significant features for Ridge Regression after changes implemented:

	features	Beta_coef_ridge
0	Neighborhood_Crawfor	0.073291
1	OverallQual_8	0.072464
2	OverallQual_9	0.072414
3	CentralAir_Y	0.067419
4	totSF	0.067031
5	Functional_Typ	0.062139
6	TotalBsmtSF	0.061554
7	GrLivArea	0.056841
8	Neighborhood_Somerst	0.055061
9	OverallCond_7	0.053128

Top 10 significant features for Lasso Regression after changes implemented:

	features	Beta_coef_lasso
0	totSF	0.106695
1	OverallQual_9	0.092264
2	CentralAir_Y	0.086028
3	Neighborhood_Crawfor	0.083716
4	OverallQual_8	0.083284
5	TotalBsmtSF	0.061882
6	Functional_Typ	0.058206
7	Neighborhood_Somerst	0.057244
8	Condition1_Norm	0.045813
9	OverallQual_7	0.036968

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
The concept of regularization used to deal with overfitting issues due to complex model. When the dataset have high number of predictors variables then the model learn too much on training dataset resulting overfitting. The same can be controlled by penalizing the less significant features.
Cost = RSS + Penalty
Ridge Regression: Ridge regression uses sum of square of beta coefficients to compute the penalty terms. Hence, it is called as L2 regularization.
Penalty = lambda * (sum of square of coefficients)
Ridge regression minimises the coefficients value to keep overall cost function low, but it never makes any coefficients 0 to discard the non-significant variables completely. Therefore, Ridge is good when we have no too many features there is multicollinearity in dataset.
Lasso Regression: Lasso regression uses the sum of absolute value of beta coefficients to compute the penalty terms. Hence, it is called L1 regularization.
Penalty = lambda * (sum of abs of coefficients)
It also makes the few coefficients to 0 for non-significant features, so in this way it also works as a feature selection model.
In our case the performance of Ridge(R2_train = .93, R2_test = .87, RMSE_train = .1, RMSE_test = .13) is slightly better than Lasso (R2_train = .91, R2_test = .86, RMSE_train = .11, RMSE_test = .13), Also it is obvious there are few cases of multicollinearity exist in predictors, so I will prefer Ridge over Lasso for our case.

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-predictor variables after removing first top 5-predictor from original model variable is below.

	features	Beta_coef_lasso1
0	MSZoning_FV	0.360750
1	Condition2_PosA	0.355388
2	MSZoning_RL	0.323914
3	MSZoning_RH	0.322419
4	MSZoning_RM	0.276796

Question 4

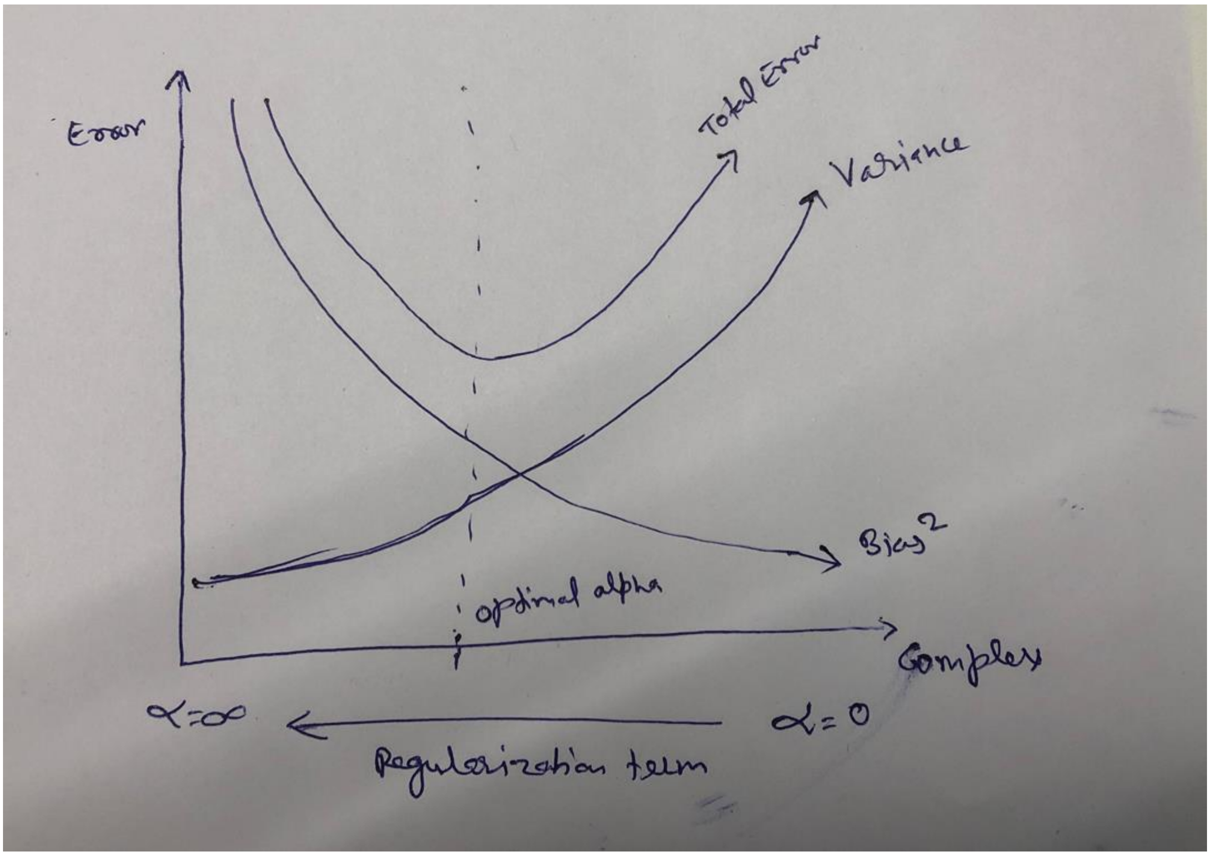
How can you make sure that a model is robust and generalizable? What are the implications of the same for the accuracy of the model and why?

Answer 4:

The model should be as simple as possible to generalize the prediction that means model should perform equally well on unseen data as compare to training data, at the same it should be as complex enough to predict close to actual value, that means the error (Bias) should be minimized.

The most general model is it ignores all coefficients and predict everything as the mean of the response variable (y-intercept) but in this case the Model error (RSS) will be maximum (Variance is low but Bias is high). Moreover, most complex model will try to remember every training data points and failed drastically on unseen dataset (Bias is low but variance is too high)

We can explain the above by plotting a trade-off between bias and variance between complex model and generalized model.



Here we can see when the alpha is 0, i.e. the model ignores all regularization and trained as complex model (Bias is high, and variance is low), when we keep increasing the alpha (regularization term) value the model the variance decrease and bias increases, but the overall error decreases. After a certain points model started generalizing itself and overall error again started increasing, this point we called as optimal value of alpha.