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)

After performing the all the steps required to build the model, below are the optimal values that we got from the model.

Ridge – Alpha Value : 7.0

Lasso – Alpha Value : 100

When we increase the Alpha value (especially double it), we see a great impact on the model as the coefficients gets penalized

Ridge : RSS + α * (sum of square of coefficients)

Lasso : RSS + α * (sum of absolute value of coefficients)

So, as the alpha increases, the coefficients would move towards zero for Ridge and coefficients become zero for Lasso.

Below are the final list of coefficients from the models built as part of the assignment.

Ridge Coefficients:

	Variable	Coeff
5	OverallQual	68857.54
20	2ndFlrSF	57643.34
63	Neighborhood_NoRidge	44952.46
32	GarageCars	42637.59

Variable		Coeff
28	TotRmsAbvGrd	42160.80

Lasso Coefficients:

Variable		Coeff
5	OverallQual	116999.91
20	2ndFlrSF	96016.85
18	TotalBsmtSF	69588.86
63	Neighborhood_NoRidge	52787.27
32	GarageCars	46467.31

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)

Below are the alpha values from the assignment.

Ridge – Alpha Value : 7.0

Lasso – Alpha Value : 100

As a principle of rule, the higher the alpha value the coefficients tends to move towards zero in case of Ridge and in case of Lasso, the coefficients would become zero.

So, Alpha value should be such that it strikes a balance the bias to variance in the model.

Having said, in this problem I have not been able to conclusively determine the appropriate model in comparison to the resultant values due the variance outcome and might need further analysis.

But in the current problem, there are many features and we need to select the once, which helps us build a better model and lasso is better than Ridge in such cases as they would also do feature elimination

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) In the existing model, below are the top 5 co-efficient.

Lasso Coefficients:

	Variable	Coeff
5	OverallQual	116999.91
20	2ndFlrSF	96016.85
18	TotalBsmtSF	69588.86
63	Neighborhood_NoRidge	52787.27
32	GarageCars	46467.31

In case these features are not available, we would go ahead with next 5 co-efficient.

64	Neighborhood_NridgHt	42789.93
28	TotRmsAbvGrd	42608.92
4	LotArea	38190.64
9	MasVnrArea	35410.81
13	BsmtExposure	34739.04

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)

To understand better about how to arrive at a robust model, its important for us to understand below images.

Both the images speak about the same concept – Bias and Variance. An optimum model balances the bias and variance in such a way that the model is not very complex and yet gives a decent level of accuracy.

Fig 1

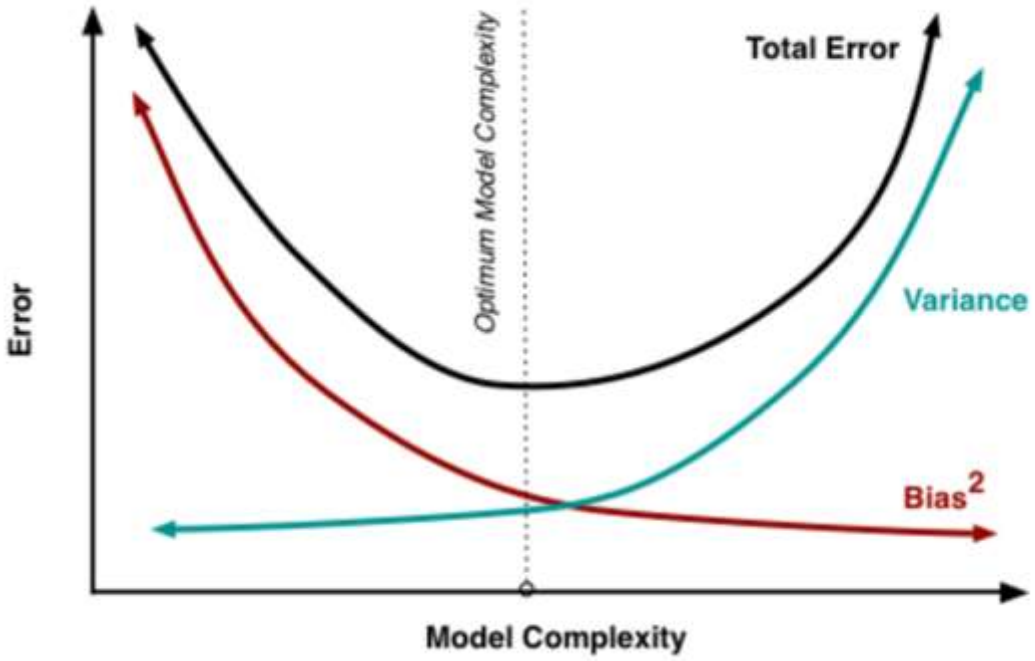
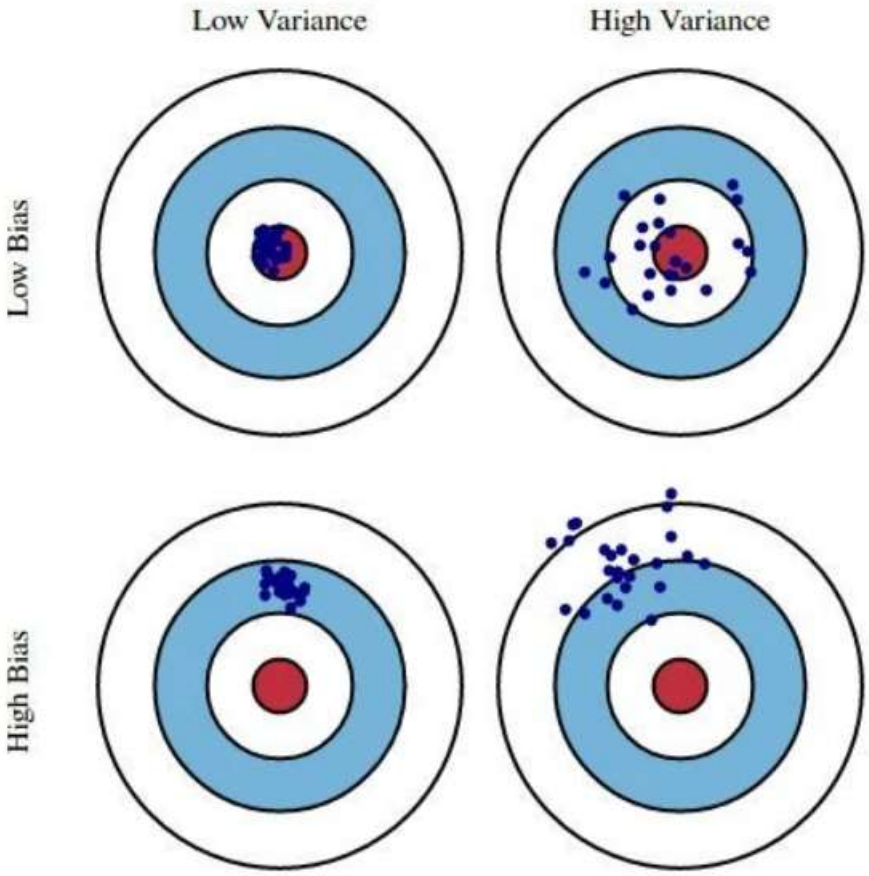

Error
Optimum Model Complexity
Total Error
Variance
Bias^2
Model Complexity

Fig 2


Low Variance
High Variance
Low Bias
High Bias

We would desire to have low bias and low variance in our model, but that is practically not possible.

When the data has been learnt by the model, we can expect the model to remember each and every data point and in test environment, we can expect high degree of variance in the data and low bias. When the model is too simple, we see the predictions are always closer and hence high bias and low variance.

In both the cases the predicted values have significant error terms which reduces the accuracy of the model. So, regularization is very important to maintain a balance between bias and variance in such a way that, we have a model which can learn efficiently in training set and predict well in the test environment too.

Regularization approach would primarily keep a check on the complexity of the model and drive towards it being more robust and efficient.