**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 values of lambda for Ridge and Lasso regression are as follows:

- Ridge Regression, it is **0.001**
- Lasso Regression, it is **20**

After doubling of values, the r2 score when regularized using **Ridge** regression **decreased** from

*TRAINING DATA: 92.59 to 92.15 TEST DATA:  88.98 to 88.73*

After doubling of values, the r2 score when regularized using **Lasso** regression **decreased** from to

*TRAINING DATA: 91.91 to 90.96 TEST DATA:  88.67 to 87.79*

Therefore, the important predictor variables after doubling the values of lambda using Ridge are,

```
OverallQual              1.050918
CentralAir_Y             1.046316
OverallCond              1.046073
Neighborhood_Crawfor     1.045878
Exterior1st_BrkFace      1.041742
Condition1_Norm          1.041628
MSZoning_FV              1.030898
SaleCondition_Normal     1.027450
Fireplaces               1.027038
Neighborhood_BrkSide     1.026143
```

The important predictor variables after doubling the values of lambda using Lambda are,

```
OverallQual              1.059503
CentralAir_Y             1.050898
Neighborhood_Crawfor     1.049571
OverallCond              1.046932
Exterior1st_BrkFace      1.041103
Condition1_Norm          1.040686
Fireplaces               1.027885
MSZoning_RL              1.026485
MSZoning_FV              1.023348
Foundation_PConc         1.020174
```

(Please refer coding document for visualizing the interpretation of important predictor variables after the lambda is doubled)

**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 r2 score value obtained for Ridge regression seems to be slightly high compared to the Lasso regression. As per our business use case, the goal is to find the important predictor variables used in predicting the Sales Price of house. So I will go with Lasso.

**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:**

```
Neighborhood_Crawfor     0.088122
CentralAir_Y             0.070992
Exterior1st_BrkFace      0.069739
MSZoning_FV              0.055165
OverallQual              0.052970
```

**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:**

Robustness refers to the model's capability to perform well across different datasets or in the presence of noise and perturbations. While Generalisability refers to the model's capability to make accurate predictions on unseen data.

If a model's complexity increases, the bias in the model keeps going down. Initially it identifies all the pattern in the data, but as the bias reduces, it starts identifying the noise present in the data as well.

On the other side, when we keep the model very simple, the variance is low. Initially it identifies the basic pattern present in the data, but the bias of the model goes very high.

So, High Bias -> model fails on training data itself

High Variance -> model fails to fit on testing data

Therefore, Bias and Variance are in a trade-off relationship over model complexity. We need a balance between bias and variance such that both are as low as possible.

One way, we can manage Overfitting is through Regularization

Regularization helps to lower model complexity by shrinking the model coefficients. This discourages the model from becoming very complex and avoids overfitting.

For Linear Regression models, model complexity mainly depends on magnitude and number of coefficients.

While building an OLS model, we will want to estimate the coefficients for which cost function is minimum. Optimizing this cost function results in model coefficients with the least possible bias, although model may have overfitted and have high variance.

In case of Overfitting, we know that we need to manage the model's complexity by primarily taking care of the magnitude of coefficients. The more extreme values of coefficients are (high positive/negative values) the more complex the model is and so higher chance of overfitting.

So, Ridge and Lasso techniques helps to add a penalty term to the model's cost function. Adding penalty in cost function helps shrink magnitude of model coefficients towards 0 and discourages creation of more complex model preventing overfitting.