# Problem Statement - Part 2

**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:**

For the Lasso Regression, I used a very low value of alpha, 0.001. The model will try to penalise more and attempt to make the majority of the coefficients zero as we increase the value of alpha.

In the case of the Ridge Regression, I used an alpha value of 5.0. When we analyse the relationship between negative mean absolute error and alpha, we can observe that the error term decreases as alpha climbs from 0 while the train error shows an increasing trend. We chose to use a value of alpha equal to 5 for our ridge regression since the test error is lowest when alpha is equal to 5.

The most important predictor variables after Lasso Regression are:

- GarageCars
- GrLivArea
- 1stFlrSF
- OverallQual
- OverallCond
- LotArea
- LotFrontage

The most important predictor variable after applying Ridge Regression are:

- SaleCondition_Normal
- SaleCondition_Partial
- GrLivArea
- MSZoning_FV
- MSZoning_RL
- MSZoning_RH
- MSZoning_RM

**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:**

Regularizing coefficients is crucial for increasing prediction accuracy, reducing variation, and making the model understandable.

Ridge regression, which employs cross validation to identify the penalty is square of magnitude of coefficients, requires a tuning parameter called lambda. By applying the penalty, the residual sum or squares should be minimal. The coefficients with higher values are penalised because the penalty is equal to lambda times the sum of the squares of the coefficients. The variance in the model is lost when we raise the value of lambda, while bias stays constant. In contrast to Lasso Regression, Ridge Regression incorporates all variables in the final model.

When performing a lasso regression, the lambda tuning parameter is used as the penalty, which is the absolute magnitude of the coefficients as determined by cross validation. As the lambda value rises, Lasso reduces the coefficient in the direction of zero, bringing the variables exactly to zero. Lasso performs variable selection as well.

When lambda is small, straightforward linear regression is performed; however, as lambda rises, shrinkage occurs and variables with a value of 0 are ignored by the 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:**

Those 5 most important predictor variables that will be excluded are :-

1. GrLivArea
2. OverallQual
3. OverallCond
4. TotalBsmtSF
5. GarageCars

**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:**

The model should be as straightforward as feasible because this will increase its robustness and generalizability while reducing accuracy. The Bias-Variance trade-off can also be used to understand it. The bias increases with model complexity despite decreasing variance and increasing generalizability. Its accuracy implication is that a robust and generalizable model will perform similarly on both training and test data, i.e., the accuracy does not change significantly for training and test data.

Bias: When a model is unable to learn from the data, it makes a mistake. High bias prevents the model from learning specifics from the data. Model's performance on training and test data is subpar.

Variance: Variance is a model error that results from the model trying to learn too much from the data. High variance indicates that the model performs remarkably well on training data since it was well trained on those data, but it performs dreadfully on testing data because that data was unknown to the model.

To prevent the overfitting and underfitting of data, bias and variance must be in balance.