# Problem Statement 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 1:**
The optimal value of alpha is:
Ridge Regression- 2
lasso Regression- 50
Ridge  regression:  Train r2: 0.89 Test r2: 0.868
Lasso  regression:  Train r2: 0.891 Test r2: 0.87

After doubling the alpha for Ridge regression:
Ridge train r2: 0.8863020351786519
Ridge test r2: 0.8658961181909242

After doubling the alpha for Lasso regression:
Lasso train r2: 0.8876692764313839
Lasso test r2: 0.8683327746966043

r2score of training data has decreased slightly for both ridge and lasso regression

Important variables:

1. LotArea (Lot size in square feet)
2. OverallQual (Rates the overall material and finish of the house)
3. OverallCond (Rates the overall condition of the house)
4. YearBuilt (Original construction date)
5. Basement floor-related variables
6. GrLivArea (Above grade (ground) living area square feet)
7. TotRmsAbvGrd (Total rooms above grade (does not include bathrooms))
8. GarageCars (Size of garage in car capacity)

It gives the same variables as important features but with different coefficients

**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 2:** As we can see from the model results, the r2 score of lasso regression is slightly better than the ridge regression for the test dataset so we will choose lasso regression to solve the prediction problem statement.

Ridge regression, which employs cross-validation to identify the penalty as a square of the magnitude of coefficients, requires a tuning parameter called lambda. By applying the penalty, the residual sum of squares should be minimal. Since the penalty is equal to lambda times the sum of the squares of the coefficients, the coefficients with higher values suffer a penalty. The variance in the model is lost when we raise the value of lambda, while the 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 3**: The top five variables now are:

1. GrLivArea (Above grade (ground) living area square feet)
2. Street_Pave (Pave road access to property)
3. 11stFlrSF (First Floor square feet)
4. BsmtFinSF1 (Type 1 finished square feet)
5. RoofMatl_Metal (Roof material_Metal)

**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 4:** The model needs to be generalised to ensure that test accuracy does not fall short of training results. For datasets other than the ones that were used during training, the model should be accurate. The outliers shouldn't be given an excessive amount of weight in order to maintain a high level of model accuracy. Only those outliers that are pertinent to the dataset should be preserved after conducting the outliers analysis to verify that this is not the case. The dataset must be cleaned up of any outliers that don't make sense to preserve. Predictive analysis cannot be believed if the model is not robust.

To prevent overfitting and underfitting of data, bias and variance must be balanced.

Bias is 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. The model's performance on training and test data is subpar. 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 that data, but it performs dreadfully on testing data because it was uncharted territory for the model.