Q1. What is the optimal value of alpha for ridge and lasso regression? What will be the changes in the model if you choose to double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

A1. The optimal value of alpha for ridge is 1.884 while that for lasso is 0.000195
Alpha is the regularization parameter. Higher the alpha, higher the penalty for adding extra variables to model. This will result in lower variance and higher bias. This also means that r2_score of the model will decrease slightly. This is evident in our model.

Regression	Alpha	Train Score	Test Score
Ridge	0.33	0.9233834143440316	0.8710509929597355
Ridge	0.66	0.9223764441317014	0.8731509099126763
Lasso	0.0001	0.9226107347077528	0.8742974432281361
lasso	0.0002	0.9199708623588315	0.8778816033450771

Top 5 features for ridge after doubling the value of alpha.

1. MSZoning_FV(Floating Village residential area)
2. MSZoning_RL (Residential Low-Density Zone)
3. MSZoning_RM
4. MSZoning_RH
5. OverallCond_Fair

Top 5 features for Lasso after doubling the value of alpha

1. MSZoning_FV
2. MSZoning_RL
3. OverallQual_Very_Poor
4. MSZoning_RM
5. MSZoning_RH

Q2. 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?

A2. Since both the models are performing similarly on given data, we will choose Lasso regression to apply on the data.
This is because, lasso helps us in reducing the number of variables and hence reduce the complexity of the model further. We can notice in our model that around 18 columns have been eliminated by lasso. Lasso also works well for high number of features in comparison to Ridge.

Q3. After building the model, you realized 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?

A3. The new top 5 variable are

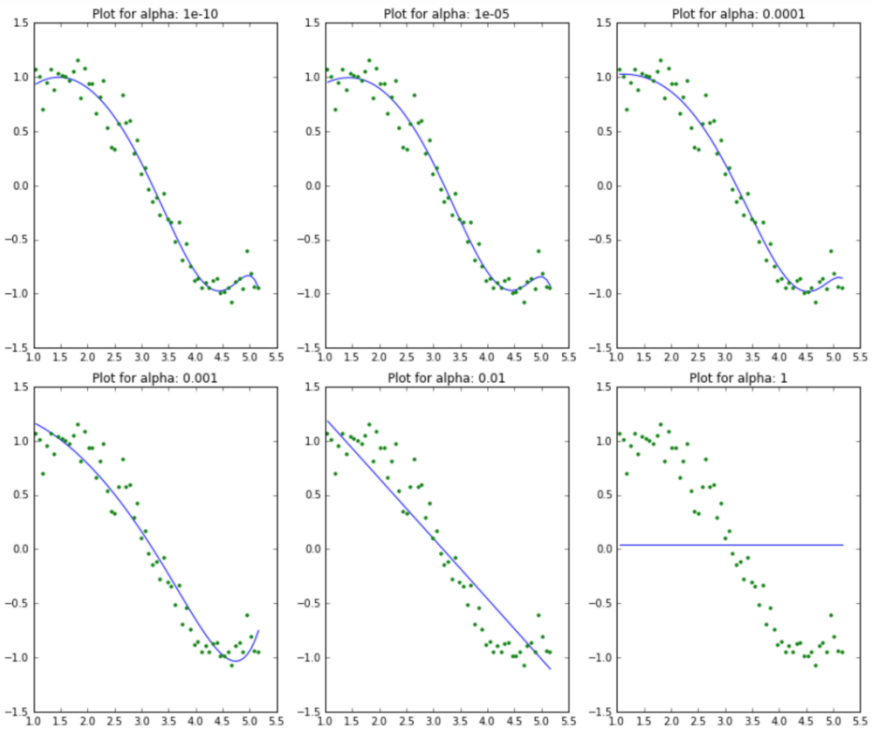
1. OverallQual_Very_Excellent
2. OverallQual_Excellent
3. OverallQual_Poor
4. OverallCond_Fair
5. Neighborhood_MeadowV

Q4. How can you make sure that a model is robust and generalizable? What are the implications of the same for t he accuracy of the model and why?

A4. A model is said to be robust and generalizable if the model can work equally well on unseen data. The test and tr aining error of the model should not be very different from each other.
A model fails to be robust and generalizable if it overfits the training data. This leads to high variance in the model. T herefore, to make sure that model is generalizable and robust, we should avoid model from overfitting the data.
This is done through regularization. It means we penalize the model for increasing complexity.
In regression problems, both ridge and lasso regularization can be used. It adds a regularization term to error term.

Objective Function = Error Term + (regularization parameter) * regularization term.

We try to minimize the objective function. Hi gher the regularization parameter, higher the penalty. This does not allow coefficients to b ecome very large.
In generalized regression, we try to not allow adding of higher order terms unnecessarily as the function may overfit the data.
As we can see in this plot, Image Source



For low value of alpha, the model is overfittin g. It tries to read and learn each point and th e curve passes through almost all points.
Due to this, if there are outliers in the data m odel may tend to deviate and be prone to ev en slight changes.

As we can clearly see model becomes straigh t line and underfits and will not be able to pr edict anything correctly if we make the value of alpha too high. **As an implication of this th e accuracy of the model will reduce.**

Therefore, we must choose alpha carefully. This is known as bias-variance trade-off.