Assignment 1:

1. **Which variables are significant in predicting the price of a house?**

   The Mean Squared Error of Lasso is slightly lower than that of Ridge and also the R-squared score in Lasso is slightly better. Also, since Lasso helps in feature reduction (as the coefficient value of one of the feature became 0), Lasso has a better edge over Ridge. So, I will go with Lasso model. So, the top 10 most significant variables for my model are:
   a. MSZoning_FV
   b. MSZoning_RL
   c. MSZoning_RH
   d. MSZoning_RM
   e. OverallQual_7
   f. OverallQual_8
   g. BsmtCond_Gd
   h. Neighborhood_Crawfor
   i. Neighborhood_ClearCr
   j. FullBath_2
2. **How well those variables describe the price of a house?**
   For my model – Lasso, the top 10 most significant variables and their coefficients are given below:

Features	Coefficients
MSZoning_FV	0.0890
MSZoning_RL	0.0761
MSZoning_RH	0.0618
MSZoning_RM	0.0517
OverallQual_7	0.0438
OverallQual_8	0.0423
BsmtCond_Gd	0.0378
Neighborhood_Crawfor	0.0373
Neighborhood_ClearCr	0.0368
FullBath_2	0.0353

   The value of the target variable Sale Price increases or decreases depending on the value of the coefficients of these above specified features. For every change in the value of any of the coefficients specified above, the Sale Price increases or decrease by the value of the coefficient depending on the sign and also provided the other coefficients are constant then. Here, the coeeficients of all the variables are positive, so change in these features will increase the value of Sale Price.

3. **Determine the optimal value of lambda for ridge and lasso regression.**
   The Optimal value of lambda for Ridge is 20 and for Lasso it is 0.0003.

Assignment 2:

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?**

The Optimal value of lambda for Ridge is 20 and for Lasso it is 0.0003.

When the value of the alpha is doubled for both Ridge and Lasso, the value of the coefficients of some of features starts to decrease, while it increases for some of the features. Also, the top variables change and for Lasso the best variable changes. Also, the R2 score starts to decrease and the Mean Squared Error increases. It happens because the more we increase the value of alpha the more the model will be penalized for each term in the model.

The most important predictor after the changes are implemented is given below:

Lasso - OverallQual_7
Ridge - OverallQual_7

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?**

I will choose Lasso over Ridge and the reasons are given below:

a. The Mean Squared Error of Lasso is slightly lower than that of Ridge and also the R-squared score in Lasso is slightly better
b. Also, Lasso helps in feature reduction i.e. the coefficient value of the features, which are redundant becomes 0. This feature is not available in Ridge. So, Lasso has a better edge over Ridge.

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?**

After the changes, the five most important predictors along with the coefficients are given below:

Features	Coefficients
BsmtCond_Gd	0.0377
Neighborhood_Crawfor	0.0351
BsmtQual_Gd	0.0281
BsmtCond_TA	0.0274
Neighborhood_ClearCr	0.0258

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?**

A model is considered to be robust and generalized if its output dependent variable (label) is consistently accurate even if one or more of the input independent variables (features) or assumptions are drastically changed due to unforeseen circumstances. In order to do that, we use regularization, the process of deliberately simplifying the model to achieve the correct balance between keeping the model simple and yet not too naive.

To generalize a model, we have to make sure that we test our model for different sets of values of the independent variables and still our model predicts with almost same accuracy. It can be achieved through Cross-validation. Here, we can test our model on different sets of values of the independent variables before actually testing it on Test data. This way, we can make sure that our model is generalized, and it will consistently give accurate results for different values of independent variables.

There will implication of the accuracy to achieve this, because we have to make balance between the bias and variance. If we try to achieve too much accuracy, we may end up making the model too complex and mugging the input. It will have very high accuracy. But, when there is change in the input, the model will fail to predict the target variable accurately. This is called overfitting. At the same time we also do not want our model too simple so that it predicts the target variable too inaccurately on train data. It is called underfitting. We have to make sure that we keep a balance between the accuracy of the model and also the variance. This is achieved through Regularization. So, we may have to give up some accuracy to make the model robust and generalized.