# House Price Prediction – Subjective Questions

**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 optimum value of Alpha for Ridge is 5 and Lasso is 0.0001.
- Upon doubling the alpha value for both ridge and lasso, the R2 Score of model is decreased and Residual Sum of Square(RSS) and Mean Root Square Error is increased.
- The top 10 important predictor variables after the change

    Ridge:
    1. GrLivArea
    2. 1stFlrSF
    3. OverallQual_Excellent
    4. TotalBsmtSF
    5. BsmtFinSF1
    6. OverallQual_Very Excellent
    7. 2ndFlrSF
    8. FullBath
    9. Neighborhood_StoneBr
    10. Neighborhood_NoRidge

    Lasso:
    1. GrLivArea
    2. OverallQual_Very Excellent
    3. OverallQual_Excellent
    4. TotalBsmtSF
    5. BsmtFinSF1
    6. Neighborhood_StoneBr
    7. Neighborhood_NoRidge
    8. LotArea
    9. Neighborhood_Crawfor
    10. SaleCondition_Partial

**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:**

After determining the optimal value of lambda for ridge and lasso I will choose Ridge to apply. Ridge gave minimum difference in R2 score for Train and Test data. Also Ridge regression helped to remove

156 features and only selected 116 features for the model by giving same R2 score as of Ridge. It helped to make Lasso model less complex than Ridge.
Below is the results obtained after building all 3 Models.

| | Model1_LinearRegression | Model2_Ridge_alpha_5.0 | Model3_Lasso_alpha_0.0001 |
|---|---|---|---|
| r2_score_train | 0.951958 | 0.940701 | 0.942063 |
| r2_score_test | 0.900242 | 0.908014 | 0.914318 |
| rss_train | 0.502649 | 0.620434 | 0.606181 |
| rss_test | 0.488821 | 0.450739 | 0.419846 |
| mse_train | 0.022823 | 0.025356 | 0.025063 |
| mse_test | 0.034362 | 0.032996 | 0.031845 |

## 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:**

Top 5 features selected by Lasso Model with alpha 0.0001 are :
1. GrLivArea
2. OverallQual_Very Excellent
3. TotalBsmtSF
4. OverallQual_Excellent
5. PoolArea

After dropping these features from training set and retraining the model, we got following 5 top features:

1. 1stFlrSF
2. 2ndFlrSF
3. BsmtFinSF1
4. BsmtUnfSF
5. Neighborhood_StoneBr

## 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:**

Below is how we can make sure model is robust and generalisable
- The Model should be simple enough to explain the variation in target variable with predictive variables.

- The Model should perform well on training and test set both. The difference in accuracy of model on train and test set should not have much of difference.
- The model should produce similar accuracy with new data which it have not seen earlier.
- The model should not be overfitted or underfitted.

Implications of the same on accuracy
- More robust and generalisable model tend to produce more errors.
- More robust and generalisable model can have less accuracy than a complex model on training set. But it works well with test data.