Assignment-based Subjective Questions

Q. 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 alpha for Ridge Regression will be 5.0
- The optimal value of alpha for Lasso Regression will be 0.0005
- On doubling the values of alpha :-
  - The R-Squared Values decrease.
  - The Coefficients of the variables change
- On implementing the changes for Lasso, the most important predictor variables will be
  - **GrLivArea**: Above grade (ground) living area square feet
  - **OverallQual**: Rates the overall material and finish of the house
  - **GarageCars**: Size of garage in car capacity
  - **TotRmsAbvGrd**: Total rooms above grade (does not include bathrooms)
  - **Fireplaces**: Number of fireplaces
- On implementing the changes for Ridge, the most important predictor variables will be
  - **OverallQual**: Rates the overall material and finish of the house
  - **GrLivArea**: Above grade (ground) living area square feet
  - **TotRmsAbvGrd**: Total rooms above grade (does not include bathrooms)
  - **GarageCars**: Size of garage in car capacity
  - **1stFlrSF**: First Floor square feet

Q. 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 to apply the Lasso Regression, as the difference between the R-Squared values is lesser compared to that of Ridge Regression.
- Furthermore, With Lasso Regression, we get to know what are the least important variables due to elimination and we will hence naturally get to know the most important variables, as they are the only ones that have coefficients less than or more than 0.

Q. 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?

- For Ridge Regression
  - **TotRmsAbvGrd** : Total rooms above grade (does not include bathrooms)
  - **2ndFlrSF** : Second floor square feet
  - **OverallCond** : Rates the overall condition of the house
  - **FullBath** : Basement full bathrooms
  - **TotalBsmtSF** : Total square feet of basement area
- For Lasso Regression
  - **TotRmsAbvGrd** : Total rooms above grade (does not include bathrooms)
  - **2ndFlrSF** : Second floor square feet

- ○ **YearBuilt** : Original construction date
- ○ **FullBath** : Basement full bathrooms
- ○ **HeatingQC** : Heating quality and condition

Q. 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?
- The first main step will involve cleaning the data. This will ensure that we remove outliers, handle missing data etc.
- From a Regression perspective, we can also execute pre-emptive feature selection to eliminate variables that are going to be irrelevant.
- After that, we go ahead and execute one-hot encoding of categorical variables and standardize the numerical variables.
- We then execute train test split and cross validation to test out the model on unseen performance.
- We can use regularization techniques to reduce overfitting and ensure that the model is neither too simple nor too complex and ensure that the bias and variance are as low as possible.