SUBJECTIVE QUESTIONS

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: - With an optimal value of alpha equal to 2 for Ridge and 0.0001 for Lasso, the model achieved an R2 score of approximately 0.82.

   After doubling the alpha values in the Ridge and Lasso models, the prediction accuracy remains around 0.82, but there is a small change in the coefficient values. The new model has been created and demonstrated in the Jupyter notebook.

   The changes in the coefficients are as follows:

	Ridge Co-Efficient
Total_sqr_footage	0.176041
GarageArea	0.105032
TotRmsAbvGrd	0.068524
LotArea	0.051735
OverallCond	0.048864
CentralAir_Y	0.032595
LotFrontage	0.031254
Neighborhood_StoneBr	0.029776
HouseStyle_2.5Unf	0.029395
Alley_Pave	0.026168
RoofMatl_WdShngl	0.024940
Neighborhood_Veenker	0.024193
MSSubClass_70	0.022818
Condition1_PosN	0.021967
Condition2_PosA	0.021145
PavedDrive_P	0.020700
SaleType_Con	0.020446
ExterCond_Ex	0.019837
BsmtQual_Ex	0.019191
KitchenQual_Ex	0.018763

	Ridge Doubled Alpha Co-Efficient
Total_sqr_footage	0.155621
GarageArea	0.095279
TotRmsAbvGrd	0.069923
LotArea	0.044952
OverallCond	0.044792
CentralAir_Y	0.032255
Neighborhood_StoneBr	0.027367
LotFrontage	0.027234
HouseStyle_2.5Unf	0.025527
Alley_Pave	0.023545
MSSubClass_70	0.021947
Neighborhood_Veenker	0.021862
BsmtQual_Ex	0.021273
RoofMatl_WdShngl	0.020224
Condition1_PosN	0.019441
KitchenQual_Ex	0.019310
MasVnrType_Stone	0.018827
PavedDrive_P	0.018461
PavedDrive_Y	0.015740
Condition1_Norm	0.015182

	Lasso Co-Efficient		Lasso Doubled Alpha Co-Efficient
Total_sqr_footage	0.209067	Total_sqr_footage	0.211765
GarageArea	0.114571	GarageArea	0.106625
TotRmsAbvGrd	0.064219	TotRmsAbvGrd	0.066123
LotArea	0.056531	OverallCond	0.044459
OverallCond	0.048484	LotArea	0.038191
CentralAir_Y	0.033568	CentralAir_Y	0.033482
Neighborhood_StoneBr	0.024157	BsmtQual_Ex	0.019367
Alley_Pave	0.022826	Alley_Pave	0.018900
HouseStyle_2.5Unf	0.020672	Neighborhood_StoneBr	0.018490
MSSubClass_70	0.018613	KitchenQual_Ex	0.015456
BsmtQual_Ex	0.018150	MSSubClass_70	0.014084
KitchenQual_Ex	0.015777	MasVnrType_Stone	0.013712
Neighborhood_Veenker	0.015695	Condition1_Norm	0.012867
LandContour_HLS	0.015332	LandContour_HLS	0.012794
Condition1_PosN	0.015328	BsmtCond_TA	0.012041
Condition1_Norm	0.014940	SaleCondition_Partial	0.010618
MasVnrType_Stone	0.014891	LotConfig_CulDSac	0.009422
PavedDrive_P	0.013511	PavedDrive_Y	0.007679
BsmtCond_TA	0.011741	ExterQual_Ex	0.007250
PavedDrive_Y	0.011409	MasVnrType_BrkFace	0.007204

Overall, since the alpha values are small, we do not observe significant changes in the model after doubling the alpha value.

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: -
   - The optimal lambda values for Ridge and Lasso are as follows:
     - ★ Ridge – 2
     - ★ Lasso – 0.0001
   - The Mean Squared Error in case of Ridge and Lasso are as follows:
     - ★ Ridge – 0.0018536041068455062
     - ★ Lasso – 0.0018777753340302228
   - The Mean Squared Error of both the models are almost same.
   - Since Lasso helps in feature reduction, Lasso has a better edge over Ridge and should be used as the final model.

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: - The five most important predictor variables in the current Lasso model are listed below: -
   a. Total_sqr_footage
   b. GarageArea
   c. TotRmsAbvGrd

d. LotArea
e. OverallCond

The New Top 5 predictors are: -

	Lasso Co-Efficient
LotFrontage	0.151172
HouseStyle_2.5Unf	0.084901
HouseStyle_2.5Fin	0.065238
Neighborhood_Veenker	0.050660
Neighborhood_StoneBr	0.046351

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: -
As per, Occam's Razor – given two models that show similar 'performance' in the finite training or test dataset, we should pick the one that makes fewer on the test dataset due to the following reasons: -
   - Simpler models are typically more "generic" and have broader applicability.
   - Simpler models require fewer training samples for effective training compared to more complex models, making them easier to train.
   - Simpler models tend to be more robust.
     - Complex models may exhibit significant changes with variations in the training dataset.
     - Simple models have lower variance and higher bias, whereas complex models have lower bias and higher variance.
     - Simpler models may make more errors on the training set, but complex models are prone to overfitting – they work very well for the training samples, fail miserably when applied to other test samples.

Therefore, to ensure that the model is both robust and generalizable, it's important to strike a balance and make the model simple enough to avoid overfitting but not so simple that it becomes ineffective.

Regularization can indeed be used to simplify the model. It helps in striking a delicate balance between keeping the model simple and avoiding making it too naive to be useful. In regression, regularization involves adding a regularization term to the cost function that penalizes the absolute values or squares of the parameters of the model.

Also, Making a model simple lead to Bias-Variance Trade-off:

- A complex model is highly sensitive to changes in the dataset and tends to be unstable, requiring adjustments for even minor variations in the training data.
- A simpler model, which abstracts some patterns followed by the data points, is less likely to change drastically even when more points are added or removed.

Bias quantifies the accuracy of the model on test data. A complex model can accurately predict outcomes given sufficient training data. Models that are too naïve, for e.g., one that fives same answer to all test inputs and makes no discrimination whatever has a very large bias as its expected error across all test inputs are very high.

Variance refers to the extent of changes in the model itself in response to changes in the training data.

Thus, maintaining a balance between bias and variance is crucial for ensuring the accuracy of the model. This balance minimizes the total error, as illustrated in the graph below: