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?

The optimal alpha values for Ridge and Lasso are identified as 2 and 0.001, respectively, resulting in an R2 of approximately 0.83. Upon doubling these alpha values, the prediction accuracy holds steady at around 0.82, albeit with slight adjustments in the coefficient values. The updated model is implemented and showcased in the Jupyter notebook, illustrating the alterations in the coefficients.

Ridge Regression Model

	Ridge Co-Efficient
Total_sqr_footage	0.169122
GarageArea	0.101585
TotRmsAbvGrd	0.067348
OverallCond	0.047652
LotArea	0.043941
CentralAir_Y	0.032034
LotFrontage	0.031772
Total_porch_sf	0.031639
Neighborhood_StoneBr	0.029093
Alley_Pave	0.024270
OpenPorchSF	0.023148
MSSubClass_70	0.022995
RoofMatl_WdShngl	0.022586
Neighborhood_Veenker	0.022410
SaleType_Con	0.022293
HouseStyle_2.5Unf	0.021873
PavedDrive_P	0.020160
KitchenQual_Ex	0.019378
LandContour_HLS	0.018595
SaleType_Oth	0.018123

	Ridge Doubled Alpha Co-Efficient
Total_sqr_footage	0.149028
GarageArea	0.091803
TotRmsAbvGrd	0.068283
OverallCond	0.043303
LotArea	0.038824
Total_porch_sf	0.033870
CentralAir_Y	0.031832
LotFrontage	0.027526
Neighborhood_StoneBr	0.026581
OpenPorchSF	0.022713
MSSubClass_70	0.022189
Alley_Pave	0.021672
Neighborhood_Veenker	0.020098
BsmtQual_Ex	0.019949
KitchenQual_Ex	0.019787
HouseStyle_2.5Unf	0.018952
MasVnrType_Stone	0.018388
PavedDrive_P	0.017973
RoofMatl_WdShngl	0.017856
PavedDrive_Y	0.016840

Lasso Regression Model

	Lasso Co-Efficient
Total_sqr_footage	0.202244
GarageArea	0.110863
TotRmsAbvGrd	0.063161
OverallCond	0.046686
LotArea	0.044597
CentralAir_Y	0.033294
Total_porch_sf	0.028923
Neighborhood_StoneBr	0.023370
Alley_Pave	0.020848
OpenPorchSF	0.020776
MSSubClass_70	0.018898
LandContour_HLS	0.017279
KitchenQual_Ex	0.016795
BsmtQual_Ex	0.016710
Condition1_Norm	0.015551
Neighborhood_Veenker	0.014707
MasVnrType_Stone	0.014389
PavedDrive_P	0.013578
LotFrontage	0.013377
PavedDrive_Y	0.012363

	Lasso Doubled Alpha Co-Efficient
Total_sqr_footage	0.204642
GarageArea	0.103822
TotRmsAbvGrd	0.064902
OverallCond	0.042168
CentralAir_Y	0.033113
Total_porch_sf	0.030659
LotArea	0.025909
BsmtQual_Ex	0.018128
Neighborhood_StoneBr	0.017152
Alley_Pave	0.016628
OpenPorchSF	0.016490
KitchenQual_Ex	0.016359
LandContour_HLS	0.014793
MSSubClass_70	0.014495
MasVnrType_Stone	0.013292
Condition1_Norm	0.012674
BsmtCond_TA	0.011677
SaleCondition_Partial	0.011236
LotConfig_CulDSac	0.008776
PavedDrive_Y	0.008685

Overall since the alpha values are small, we do not see a huge change in the model after doubling the alpha.

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?

The optimal lambda values for Ridge and Lasso are determined to be 2 and 0.0001, respectively. The Mean Squared Error (MSE) for Ridge is found to be 0.0018396090787924262, and for Lasso, it is 0.0018634152629407766. Notably, the MSE values for both models are nearly identical. Given Lasso's ability to facilitate feature reduction by causing certain feature coefficients to become zero, it holds a distinct advantage over Ridge. Consequently, Lasso is recommended as the preferred final model.

The top five predictor variables identified in the current Lasso model are as follows:

Total_sqr_footage
GarageArea
TotRmsAbvGrd
OverallCond
LotArea

After removing these attributes and constructing a new Lasso model in the Jupyter notebook, the R2 of the updated model decreases to 0.73. Additionally, there is an increase in the Mean Squared Error, which rises to 0.0028575670906482538. This indicates that the excluded predictor variables were contributing significantly to the model's predictive performance, and their removal has resulted in a decrease in the model's accuracy.

The new Top 5 predictos are:

	Lasso Co-Efficient
LotFrontage	0.146535
Total_porch_sf	0.072445
HouseStyle_2.5Unf	0.062900
HouseStyle_2.5Fin	0.050487
Neighborhood_Veenker	0.042532

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?

In accordance with Occam's Razor, when comparing two models with similar performance on finite training or test data, the preference is for the one exhibiting fewer complexities. Several reasons support this inclination:

Generality and Applicability: Simpler models tend to be more generic and widely applicable.
Efficient Training: Simpler models require fewer training samples for effective training compared to more complex ones, making them easier to train.
Robustness: Simpler models demonstrate greater robustness. In contrast, complex models may undergo drastic changes with variations in the training dataset, leading to high variance and low bias.

To enhance a model's robustness and generalizability, it is advocated to keep the model simple but not excessively so. Regularization is a valuable technique for achieving this balance. In regression, regularization involves introducing a term to the cost function that penalizes the absolute values or squares of the model parameters.

Additionally, the pursuit of model simplicity contributes to the Bias-Variance Trade-off:

Model Stability: Complex models exhibit instability, changing significantly with even minor alterations in the dataset. Simpler models, abstracting patterns from given data points, are less likely to undergo wild changes, even with the addition or removal of data points.

Bias and Variance: A complex model, given sufficient training data, can accurately predict outcomes but might have a high bias. Conversely, overly simplistic models, behaving uniformly across all inputs, may have a large bias. Striking a balance between bias and variance is crucial for maintaining model accuracy.

By navigating the delicate equilibrium between bias and variance, model accuracy can be upheld, minimizing total error, as depicted in the accompanying graph.