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?

Ans: The optimal values of alpha for Ridge is 0.3 and for Lasso alpha is 0.0001

Before Changes - Alpha for Ridge is 0.3 and for Lasso alpha is 0.0001

	Metric	Linear Regression	Ridge	Lasso
0	R2 Score(Train)	9.572093e-01	0.953939	0.952196
1	R2 Score(Test)	-3.460182e+21	0.773461	0.767728
2	RSS(Train)	7.061853e+00	7.601641	7.889285
3	RSS(Test)	2.341681e+23	15.331041	15.719027
4	MSE(Train)	8.316612e-02	0.086286	0.087903
5	MSE(Test)	2.312207e+10	0.187089	0.189442

After Changes - Alpha for Ridge is 0.6 and for Lasso alpha is 0.0002

	Metric	Ridge_New	Lasso_New
0	R2 Score(Train)	0.950662	0.947273
1	R2 Score(Test)	0.794949	0.790065
2	RSS(Train)	7.601641	7.601641
3	RSS(Test)	15.331041	15.331041
4	MSE(Train)	0.089302	0.092319
5	MSE(Test)	0.177995	0.180102

When alpha value is doubled, R2score on Training set there is minor changes. However on test set, R2 score is improved. RSS on training/Test set didn't change for Ridge, but for Lasso there is minor decrease in the RSS value. The mean squared error is reduced by a very small values

Top Predictors after doubling alpha Values

Features	Lasso_Coefficients
GrLivArea	0.985721
OverallQual	0.453984
TotalBsmtSF	0.333838
LotArea	0.325797
MSZoning_FV	0.314227
OverallCond	0.308098
MSZoning_RL	0.265362
MSZoning_RM	0.240925
MSZoning_RH	0.232551
SaleType_ConLD	0.179756
Neighborhood_Crawfor	0.117666
SaleType_New	0.111212

Features	Ridge_Coefficients
GrLivArea	0.657613
OverallQual	0.369039
MSZoning_FV	0.328552
LotArea	0.321569
TotalBsmtSF	0.299879
MSZoning_RM	0.292483
MSZoning_RL	0.291938
OverallCond	0.285349
Condition2_Feedr	0.269962
MSZoning_RH	0.266168
Condition2_Norm	0.233595
SaleType_ConLD	0.231399

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: I would prefer to use Lasso as it will have inbuilt feature to eliminate the least relevant column by making the coefficient 0 or negative values.

	Metric	Linear Regression	Ridge	Lasso
0	R2 Score(Train)	9.572093e-01	0.953939	0.952196
1	R2 Score(Test)	-3.460182e+21	0.773461	0.767728
2	RSS(Train)	7.061853e+00	7.601641	7.889285
3	RSS(Test)	2.341681e+23	15.331041	15.719027
4	MSE(Train)	8.316612e-02	0.086286	0.087903
5	MSE(Test)	2.312207e+10	0.187089	0.189442

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?

Top 5 most important predictor variables in Lasso model was GrLivArea, TotalBsmtSF, MSZoning_FV, OverallQual, LotArea

After removing these top predictor variables we got below TotRmsAbvGrd, BsmtFinSF1, OverallCond GarageArea and FullBath are the top predictors

Features	Lasso
TotRmsAbvGrd	0.481110
BsmtFinSF1	0.340347
OverallCond	0.307525
GarageArea	0.262322
FullBath	0.243634
BsmtUnfSF	0.243135
HalfBath	0.135809
Fireplaces	0.129160
Neighborhood_StoneBr	0.125861
Neighborhood_Crawfor	0.116457
CentralAir_Y	0.113641

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?

Model robustness means how well a model can handle variations in the data, such as noise, outliers, or changes in distribution. A robust model should be able to generalize well to new or unseen data and avoid overfitting or underfitting. Model accuracy measures how well a model predicts the correct outcome for a given input data. A high accuracy model should be able to capture the true relationship between the input and the output variables and minimize the error rate.

1. Identification of Outliers and removing it
2. Imputing the missing values during data Pre-processing
3. If the distribution curve is not a normal distribution applying log scale will help
4. Feature Engineering helps to derive the values or add more relevant features
5. Feature Selection helps to select the relevant features for model prediction
6. Cross Validation strategies – Apply k-fold or leave-one-out cross-validation to ensure that the model generalizes well to different data subsets, including those containing outliers.
7. Regularisation - Optimise or try different values of alphas while using Lasso/Ridge and remove Overfitting.
8. R2 Scoring
9. Mean Square Error

To make sure that the model is robust and generalisable the model needs to be simple. We need to adjust the hyperparameters of the models to find the optimal values that minimize the error or maximize the R2Score of the Training and Test data. Accuracy can be viewed as the complement of the error rate. In other words, accuracy is equal to 1 minus the error rate. We need to use various methods to optimize and tune your models, such as grid search, regularization techniques, such as lasso, ridge and prevent overfitting.