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:

Initial alpha for Ridge regression: 0.2
Initial alpha for Lasso regression: 0.0001

	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	9.456799e-01	0.936472	0.927912
1	R2 Score (Test)	-1.259406e+23	0.694098	0.640577
2	RSS (Train)	6.685318e-01	0.781861	0.887207
3	RSS (Test)	6.862939e+23	1.666965	1.958619
4	MSE (Train)	2.558870e-02	0.027673	0.029478
5	MSE (Test)	3.953874e+10	0.061621	0.066795

But R2 value shows huge difference b/w Ridge & Lasso Regression. This is an indication of overfitting & existence of multicollinearity. So, a lot of predictors were removed & another set of Ridge & Lasso Regression performed & below are the results.

Optimal alpha for Ridge Regression : 0.0001
Optimal alpha for Lasso Regression : 0.0001

R2 Square using Alpha 0.0001

	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	0.852754	0.852754	0.849573
1	R2 Score (Test)	0.807963	0.807969	0.816495
2	RSS (Train)	1.812197	1.812197	1.851346
3	RSS (Test)	1.046476	1.046446	0.999982
4	MSE (Train)	0.042130	0.042130	0.042582
5	MSE (Test)	0.048824	0.048823	0.047727

Doubling the alpha gives below results:
R2 Square using Alpha 0.0002

	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	0.852754	0.852754	0.842677
1	R2 Score (Test)	0.807963	0.807974	0.816719
2	RSS (Train)	1.812197	1.812197	1.936214
3	RSS (Test)	1.046476	1.046415	0.998764
4	MSE (Train)	0.042130	0.042130	0.043548
5	MSE (Test)	0.048824	0.048822	0.047698

The effect of alpha value on both ridge and lasso regression is the same in terms of value increase and decrease. In our case, changing alpha has very little effect on the R2 Score of Lasso Regression, whereas Ridge Regression value is nearly the same.
In case of Lasso and Ridge regression, as alpha value increases, the slope of the regression line reduces and becomes horizontal.

After the change:
In Ridge Regression : Column **Condition2_PosN** is most significant predictor variable.
In Lasso Regression : Column **OverallQual** is most significant predictor variable.

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?

Prefer to use Lasso regression. Because Lasso performs feature selection by setting coefficients of some of the predictors to zero. Hence it is easier to interpret models generated by Lasso compared to models generated by Ridge.
Also R2 square value difference b/w training set & test less compared to Ridge & LR models.

In our example : Condition2_PosA predictor's co-efficent is zero, where as Ridge regression has value for it.

Question 3

After building the model, you realized 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?

five most important predictor variables now are here :

LotArea,
MasVnrArea,
GarageArea ,
Neighborhood_NoRidge,
Condition2_PosA

Full list is here :

	LassoRemoved5Col
LotArea	0.173486
OverallCond	0.026395
YearBuilt	0.107466
YearRemodAdd	0.061665
MasVnrArea	0.112681
BsmtFullBath	0.024407
GarageArea	0.160469
WoodDeckSF	0.059784
Street_Pave	0.056428
LotConfig_CulDSac	0.017281
Neighborhood_Crawfor	0.087227
Neighborhood_Mitchel	-0.034656
Neighborhood_NoRidge	0.141954
Neighborhood_NridgHt	0.099220
Neighborhood_Somerst	0.037137
Neighborhood_StoneBr	0.096684
Condition1_Norm	0.010468
Condition2_PosA	0.116159
BldgType_Twnhs	-0.071344
BldgType_TwnhsE	-0.041208
Exterior1st_BrkFace	0.056329
Exterior2nd_CmentBd	0.028690
BsmtQual_Gd	-0.014420
BsmtExposure_Gd	0.048011
KitchenQual_Gd	-0.001526

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?

Making a predictive model more robust to outliers is important to ensure the model's stability and accuracy when dealing with extreme data points. Outliers can significantly impact the model's performance, leading to less reliable predictions. Here are few methods to enhance the robustness of a predictive model to outliers:

Data Preprocessing: Carefully preprocess the data by identifying and handling outliers. Some common techniques include:Trimming: Removing extreme values beyond a certain threshold.Capping the extreme values to a specified percentile.Imputation: Replacing outliers with more representative values, such as the mean or median.

Feature Scaling: Apply feature scaling techniques such as normalization or standardization to bring all features to a similar scale. Scaling can reduce the impact of extreme values on the model's performance.

Transformations: Apply data transformations to make the data less sensitive to outliers. For example, using logarithmic transformations can compress the range of extreme values.

Make sure following issues in the dataset are handled.

- Non-constant variance
- Autocorrelation and time series issue
- Multicollinearity
- Overfitting
- Extrapolation

Cross-Validation: Use robust cross-validation methods, such as k-fold cross-validation or stratified cross-validation, to evaluate the model's performance more accurately and minimize the effect of outliers on the validation process.

Remove Outliers: In some cases, it may be appropriate to remove extreme outliers if they are likely due to data entry errors or anomalies and not representative of the underlying pattern.

Data Augmentation: Consider data augmentation techniques that generate additional training samples based on existing data to make the model more robust.

Train on Robust Subsets: If possible, create subsets of the data with reduced outliers or remove outliers from the training set entirely when building the model.

It's essential to strike a balance between making the model robust to outliers while retaining the ability to capture valuable information from the data. Careful experimentation and evaluation of different techniques on a validation set will help identify the most effective approach for the specific predictive modeling task at hand.