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 alph a for both ridge and lasso? What will be the most important predictor v ariables after the change is implemented?

Ans) Optimal Alpha Values:
For Ridge Regression: The optimal alpha is 100.
For Lasso Regression: The optimal alpha is 0.1.

Impact of Doubling Alpha:
If we double alpha to 200 for Ridge regression, it makes the model more simple and less likely to overfit, but it might not predict as well.
If we double the alpha to 0.2 for Losso regression it makes the model e ven simpler, focusing only on the most important factors and ignoring l ess important ones.

Important Predictor Variables:

In Ridge regression, doubling alpha leads to smaller coefficients for a ll predictors compared to their values with the original alpha. However , the importance of predictors remains similar, with GrLivArea and Over allQual still being the top predictors.

In Lasso regression, doubling alpha resulting in coefficient shrinkage, leading to even smaller coefficients. The predictor OverallQual becomes the most important predictor, followed by GrLivAreav and TotalBsmtSF. Other predictors coefficients are reduced significantly indicating stro nger feature selection.

Top predictors after doubling alpha in Ridge regression:
GrLivArea: 0.11853310912926272
OverallQual: 0.09513465220914079
TotalBsmtSF: 0.08013810343001075
1stFlrSF: 0.07832319378584823
BsmtFinSF1: 0.06921665983390492
GarageArea: 0.06623276119594229
BsmtQual_Ex: 0.0656964299734938
KitchenQual_Ex: 0.06414401427069984
2ndFlrSF: 0.06373885213891518
Neighborhood_StoneBr: 0.05952595391211318

Top predictors after doubling alpha in Lasso regression:
OverallQual: 0.2974164832498771
GrLivArea: 0.27425976277623787
TotalBsmtSF: 0.10295165888204998
GarageArea: 0.08047811991948153
BsmtQual_Ex: 0.04730988020720439
YearBuilt: 0.03495810167444036
GarageCars: 0.026666451027454907
Id: -0.0
MSSubClass: -0.0
LotFrontage: 0.0

Question 2
You have determined the optimal value of lambda for ridge and lasso reg
ression during the assignment. Now, which one will you choose to apply
and why?

Ans) Ridge Regression:
Optimal alpha: 100
Mean Squared Error: 0.098
R-squared: 0.887
Lasso Regression:
Optimal alpha: 0.1
Mean Squared Error: 0.145
R-squared: 0.833
Based on these metrics, Ridge Regression appears to perform better than
Lasso Regression for this dataset.
Ridge Regression has a lower mean squared error (0.098) compared to Las
so Regression (0.145), indicating that Ridge Regression's predictions a
re closer to the actual values on average. Ridge Regression achieves a
higher R-squared value (0.887) compared to Lasso Regression (0.833), in
dicating that Ridge Regression explains more of the variance in the tar
get variable and provides a better fit to the data.

Therefore, I would choose Ridge regression for the model for this data
set.

Question 3
After building the model, you realised that the five most important pre
dictor variables in the lasso model are not available in the incoming d
ata. You will now have to create another model excluding the five most
important predictor variables. Which are the five most important predic
tor variables now?

Ans) The five most important predictor variables in the new Lasso model
, after excluding the top contributors from the previous model, are:
1stFlrSF
GarageCars
YearBuilt
TotRmsAbvGrd
2ndFlrSF
This means that these variables are now considered the most influential
in predicting the target variable based on the updated Lasso regression
model.

Question 4
How can you make sure that a model is robust and generalisable? What ar
e the implications of the same for the accuracy of the model and why?

Ans) To ensure a model is robust and generalizable, use techniques like
cross-validation, train-test splitting, feature selection, regularizati
on, hyperparameter tuning, and evaluating on different datasets. This h
elps improve accuracy by ensuring the model performs well on new, unsee
n data, avoids overfitting, and maintains consistent performance across
various data distributions.