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?

Increasing the value of alpha in both Ridge and Lasso regression would lead to stronger regularization, which means:

The model coefficients will be more heavily penalized, leading to smaller coefficient values.

Some coefficients may be reduced to exactly zero, effectively performing feature selection.

The model will become more biased but fewer complexes, potentially reducing over fitting.

As for the most important predictor variables after doubling the alpha values, the previous attempt to identify them encountered an error due to missing columns in the Data Frame. To determine the most important predictors with the updated alpha values, we would need to:

Ensure the intended columns are present in the Data Frame.

Fit the Ridge and Lasso models with the doubled alpha values.

Examine the non-zero coefficients in the fitted models to identify the most important predictors.

Lasso MSE with alpha = 0.002: 0.020575272183313464 It has been observed that on increasing the alpha value, the Mean squared error is slightly increased for both Lasso and Ridge. Ridge R2 score with alpha =20: 0.8841699740253731 Ridge R2 score with alpha =40: 0.8817870719700953 Lasso R2 score with alpha = 0.001: 0.8771160127831338 Lasso R2 score with alpha = 0.002: 0.8749511815052119 With the increase in value of alpha, we can see slight decrease in the value of R2. After doubling the value of alpha, the 5 most important variables turned out to be as mentioned below: From Ridge Model:

1. MSSubClass

2. OverallCond

3. BsmtFullBath

4. Neighborhood_Crawfor

5. Neighborhood_NridgHt From Lasso Model:

    1. MSSubClass

2. MSZoning_RL

3. MSZoning_RH

4. MSZoning_FV

5. MSZoning_RM

Optimal Alpha Values:

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?

To determine which regression model (Ridge or Lasso) to choose based on the optimal lambda (alpha) values, we need to consider the following factors:

Model performance: Compare the Mean Squared Error (MSE) and R2 scores of both models at their optimal alpha values. The model with lower MSE and higher R2 score is generally preferred.

Feature selection: Lasso regression has the ability to perform feature selection by shrinking some coefficients to exactly zero. If feature selection is a priority, Lasso might be preferred.

Interpretability: Ridge regression keeps all variables in the model, which can be beneficial if interpretability is important. Lasso's feature selection may make the model more parsimonious but less interpretable.

Multicollinearity: Ridge regression is effective in handling multicollinearity among predictor variables, while Lasso may struggle in such cases.

Compare the performance of Ridge and Lasso regression at their optimal alpha values:

Ridge regression (alpha = 20):

MSE: 0.01905843126000864

R2 score: 0.8841699740253731

Lasso regression (alpha = 0.0001):

MSE: 0.020219075353064726

R2 score: 0.8771160127831338

The Ridge regression model has a slightly lower MSE and higher R2 score compared to the Lasso model.

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?

To address the issue of missing values and proceed with fitting a new Lasso model excluding the specified variables (LotFrontage, BsmtFullBath, Neighborhood_Crawfor, Neighborhood_Somerst, OverallCond), we need to take the following steps:

1. Handle missing values either by imputation or by removing rows/columns with missing values.
2. Encode categorical variables to numeric using one-hot encoding or another suitable method.
3. Exclude the specified variables from the dataset.
4. Fit the Lasso model with the updated dataset.

After excluding the specified variables and fitting a new Lasso model, the five most important predictor variables are related to the material of the roof, with RoofMatl_Membran having the highest coefficient, followed by RoofMatl_WdShngl, RoofMatl_Metal, RoofMatl_CompShg, and RoofMatl_Tar&Grv. These variables are now considered the most influential in predicting the house price in the updated model.

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?

The analysis of the updated Lasso model reveals that the five most important predictor variables are now related to the material of the roof:

	0
RoofMatl_Membran	739903.4349006016
RoofMatl_WdShngl	702317.2784714056
RoofMatl_Metal	700626.7897973518
RoofMatl_CompShg	648866.7320694894
RoofMatl_Tar&Grv	646671.2718439889

These variables, such as RoofMatl_Membran and RoofMatl_WdShngl, have the highest coefficients in the model, indicating their strong influence on predicting house prices. The exclusion of the previously identified variables (LotFrontage, BsmtFullBath, Neighborhood_Crawfor, Neighborhood_Somerst, OverallCond) has shifted the focus to the roof material as the key predictor.

To ensure the robustness and generalizability of a model, several techniques can be employed:

Cross-validation: By using techniques like k-fold cross-validation, the model's performance can be evaluated on different subsets of the data, providing a more reliable estimate of its generalization ability.

Regularization: Regularization techniques, such as Lasso (L1) or Ridge (L2) regularization, help prevent overfitting by adding a penalty term to the model's objective function. This encourages the model to learn more generalizable patterns rather than memorizing noise in the training data.

Feature selection: Identifying and selecting the most relevant features, as done in this analysis, can help improve the model's generalizability by focusing on the most informative variables and reducing the impact of irrelevant or noisy features.

Testing on unseen data: Evaluating the model's performance on a separate test set that was not used during training provides an unbiased assessment of its generalization ability.