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?

Based on the investigation, we came up with the following LAMBDA values for Ridge and Lasso regression:

The ideal LAMBDA value for a ridge regression is 2.0.

Lasso Regression: 0.0001 is the ideal LAMBDA value.

The following alterations may be seen if we opt to double the value of alpha for both Ridge and Lasso regression:

Ridge Regression: In Ridge regression, doubling the alpha value causes the mean squared error to slightly rise. The R-squared values for the train and test sets, however, are mostly unaltered. This shows that increasing the alpha value has no appreciable impact on the model's performance.

Lasso Regression: In Lasso regression, doubling the alpha value causes the mean squared error to slightly rise. Additionally, the train set's R-squared value marginally declines, indicating a worse fit. The test set's R-squared value, which significantly decreases, experiences the most shift, though. This shows that increasing the alpha value by two degrades the model's capacity for prediction.

In addition, increasing the alpha value by two in both Ridge and Lasso regression penalises the models even more, pushing more coefficients in the direction of zero. This may lead to more characteristics being ignored or having less of an influence on the predictions made by the model.

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?

For Ridge and Lasso, the best values of lambda were as follows:

Lasso: 0.0001; Ridge: 2.0

The R-squared values for Ridge and Lasso are as follows:

Ridge: Test = 0.896, Ridge: 0.930, difference: 0.046

Lasso: Train=0.927, Test=0.902, difference=0.025

Ridge and Lasso's Mean Squared Errors are: - Ridge: 0.00297

- Lasso: 0.00280

We can see from the findings that Lasso has a little smaller Mean Squared Error than Ridge. Additionally, Lasso has a smaller R-squared difference between the train and test sets than Ridge does. In addition, Lasso has an advantage over Ridge due to its capacity to condense information and interpret the model by reducing coefficients in the direction of zero.

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?

The top five predictors were discovered to be TotalBsmtSF, TotRmsAbvGrd, OverallCond, Total_Bathrooms, and LotArea by eliminating the top five predictors from the Lasso 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?

There are various crucial procedures to think about in order to guarantee that a model is reliable and generalizable. In the beginning, applying cross-validation techniques can assist in evaluating the model's performance on various subsets of the data, lowering the danger of the model becoming overfit to particular patterns in the training data. This encourages the model's successful generalisation to new data.

Second, it's critical to employ a broad and representative dataset. The model can discover patterns that are more relevant and generalizable when the dataset has a variety of data points and accurately depicts the real-world event.

The robustness of the model can then be increased by using regularisation techniques and careful feature selection. The model may concentrate on the most useful characteristics, decrease the influence of noise, and become more generalizable by choosing pertinent features and implementing regularisation techniques like L1 or L2 regularisation.

Another crucial element is preventing data leaks. A meaningful assessment of the model's performance on unobserved data may be achieved by ensuring a clear division between training and testing data and preventing any unintentional usage of test data during model training.

The model's generalizability may also be estimated by testing it on unobserved data using a hold-out validation set or cross-validation. We learn more about the model's potential for success in real-world circumstances by evaluating how well it performs on data that it has never seen before.

The model's accuracy is impacted by these procedures. The model becomes more dependable and accurate while handling fresh, untested data by encouraging generalizability and minimising overfitting. Instead of memorising individual instances, it learns fundamental patterns, which enhances performance on real-world tasks. Through assessment on unseen data, reliable performance estimates are established, guaranteeing that stated metrics accurately represent the model's capabilities.

In conclusion, cross-validation, a variety of datasets, feature selection, regularisation, preventing data leaking, and assessing on unseen data all contribute to model robustness and generalizability. By allowing the model to successfully manage fresh data and offer accurate predictions in real-world applications, these procedures increase the model's accuracy.