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:
After considering Ridge and Lasso regularization, we obtained the following optimal values for LAMBDA:
1. Ridge: 2.0
2. Lasso: 0.0001

If we decide to double the value of alpha for both Ridge and Lasso, the model will undergo the following changes:
1. Ridge:
    a. There would be a slight increase in the mean squared error.
    b. The R-squared values for both the training and test datasets would remain nearly the same.
2. Lasso:
    a. There would be a slight increase in the mean squared error.
    b. The R-squared value for the training dataset would slightly decrease.
    c. However, there would be a significant decline in the R-squared value for the test dataset, indicating a deterioration in model performance and predictions.
    d. The increased alpha further penalizes the model, causing more coefficients of variables to shrink closer to zero.

Considering these changes, the most important predictor variables are:
Ridge:
1. Total_sqr_footage
2. OverallQual
3. GrLivArea
4. Neighborhood_StoneBr
5. OverallCond
6. TotalBsmtSF
7. LotArea
8. YearBuilt
9. Neighborhood_Crawfor
10. Fireplaces

Lasso:
1. Total_sqr_footage
2. OverallQual
3. YearBuilt
4. GrLivArea
5. Neighborhood_StoneBr
6. OverallCond
7. LotArea
8. Neighborhood_Crawfor
9. Neighborhood_NridgHt
10. GarageCars

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?
Ans:
In the case of Ridge and Lasso regularization, we obtained the following optimal values for LAMBDA:
1. Ridge: 2.0
2. Lasso: 0.0001

The R-squared (r2) values we obtained for Ridge and Lasso are as follows:
1. Ridge: Train = 0.930, Test = 0.896, difference = 0.046
2. Lasso: Train = 0.927, Test = 0.902, difference = 0.025

The Mean Squared Error (MSE) for Ridge and Lasso is:
- Ridge: 0.00297
- Lasso: 0.00280

From these results, several observations can be made:
1. The Mean Squared Error (MSE) of Lasso is slightly lower than that of Ridge, indicating better performance in terms of prediction accuracy.
2. The difference in R-squared (r2) values between the training and test datasets is smaller for Lasso compared to Ridge, suggesting that Lasso generalizes better to unseen data.
3. Lasso's ability to perform feature reduction by shrinking the coefficient values towards zero makes it more interpretable and advantageous over Ridge.

Overall, based on these findings, Lasso regularization demonstrates a competitive edge over Ridge regularization due to its lower MSE, better generalization, and enhanced interpretability.

Question 3
Ans: 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?
Ans:
We dropped the top 5 most important predictor variables in the lasso model and again created again model and got the below five most important predictor variables:
1. TotalBsmtSF
2. TotRmsAbvGrd
3. OverallCond
4. Total_Bathrooms
5. LotArea

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?

Ans:

Simplicity is a desirable characteristic for a model, as it enhances its generalizability and robustness, even at the cost of decreased accuracy. This concept is rooted in the Bias-Variance trade-off, where simpler models exhibit higher bias but lower variance, resulting in greater generalizability. Conversely, complex models tend to have high variance and low bias. Occasionally, models encounter problems such as underfitting or overfitting. Therefore, striking a balance between bias and variance becomes crucial, and this is where "Regularization" comes into play.

Regularization effectively manages model complexity by shrinking coefficients towards zero. By doing so, it prevents the model from becoming excessively complex, mitigating the risk of overfitting. To maintain an optimal level of simplicity, it is advisable to employ regularization methods, which penalize the model if it becomes too complex.

Regularization enables the achievement of the Bias-Variance trade-off. It compromises by increasing bias to an optimal level where the total error is minimized. This optimal model complexity, also known as the Optimum Model Complexity, strikes a balance between simplicity for generalizability and complexity for robustness.

In summary, pursuing model simplicity entails navigating the Bias-Variance trade-off.