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:

- The best alpha value for Ridge regression is 1.49, indicating moderate regularization, while for Lasso regression, the best alpha is 0.0006, suggesting very weak regularization.
- After changing the value of Ridge regression's alpha, the coefficient values are increasing, potentially leading to overfitting. Similarly, increasing the alpha value of Lasso regression results in more features having a coefficient of zero, which means they are not given weight in the calculation.
- The most important predictor variables in Lasso regression's top five are OverallCond, BsmtFullBath, YearBuilt, and LotShape.

Ridge (alpha=1.48):

- 2ndFlrSF (Coefficient: 0.480840)
- OverallCond (Coefficient: 0.451306)
- BsmtFullBath (Coefficient: 0.347978)
- YearBuilt (Coefficient: 0.182533)
- BsmtFinType2 (Coefficient: 0.179324)

Lasso (alpha=0.0006):

- 2ndFlrSF (Coefficient: 0.591979)
- OverallCond (Coefficient: 0.501329)
- BsmtFullBath (Coefficient: 0.407713)
- YearBuilt (Coefficient: 0.175503)
- LotShape (Coefficient: 0.128970)

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:

Based on these results, I would choose Lasso Regression for the following reasons:

- Higher Test R^2 Score: Lasso Regression has a slightly higher Test R^2 Score (0.8967) compared to Ridge Regression (0.8950), indicating better predictive performance on unseen data.
- Lower RMSE: Lasso Regression also has a slightly lower Root Mean Squared Error (RMSE) of 0.1281 compared to Ridge Regression's RMSE of 0.1295, indicating better accuracy in predicting the target variable.
- Better Feature Selection: Lasso Regression tends to perform feature selection by shrinking some coefficients to exactly zero. This means it may provide a simpler and more interpretable model by automatically selecting the most important features while still maintaining good predictive performance.

Model Evaluation: Ridge Regression, alpha=1.48

- R2 score (train) : 0.9142
- R2 score (test) : 0.8948
- RMSE (train) : 0.117
- RMSE (test) : 0.1292

Model Evaluation: Lasso Regression, alpha=0.006

- R2 score (train) : 0.9106
- R2 score (test) : 0.8967
- RMSE (train) : 0.1195
- RMSE (test) : 0.1281

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?

Ans:

Ridge (alpha=1.48):

- 2ndFlrSF (Coefficient: 0.480840)
- OverallCond (Coefficient: 0.451306)
- BsmtFullBath (Coefficient: 0.347978)
- YearBuilt (Coefficient: 0.182533)
- BsmtFinType2 (Coefficient: 0.179324)

Lasso (alpha=0.0006):

- 2ndFlrSF (Coefficient: 0.591979)
- OverallCond (Coefficient: 0.501329)
- BsmtFullBath (Coefficient: 0.407713)
- YearBuilt (Coefficient: 0.175503)
- LotShape (Coefficient: 0.128970)

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:

- **Cross-Validation**: Use techniques like k-fold cross-validation to assess the model's performance across multiple subsets of the data. This helps in evaluating how well the model generalizes to unseen data.
- **Train-Test Split**: Split the dataset into training and testing sets. Train the model on the training set and evaluate its performance on the testing set. This helps to simulate how well the model performs on new, unseen data.
- **Feature Selection and Regularization:** Employ techniques like L1 (Lasso) and L2 (Ridge) regularization to prevent overfitting and reduce the impact of irrelevant features. Feature selection methods can also be used to choose the most relevant features for the model.
- **Hyperparameter Tuning:** Fine-tune the model's hyperparameters using techniques like grid search or random search. This ensures that the model is optimized for performance while avoiding overfitting.
- **Ensemble Methods:** Utilize ensemble methods like bagging, boosting, or stacking to combine the predictions of multiple models. Ensemble methods often lead to more robust and accurate models by leveraging the strengths of individual models.
- **Out-of-Sample Testing**: Validate the model's performance on completely unseen data, separate from both the training and testing sets. This provides a more stringent test of the model's generalization capability.