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 optimal value of alpha for:
Ridge regression: Generally value of alpha is set closer to zero or very less value but it is essential to keep the balance between bias and variance. Cross validation can be use in choosing best alpha value, also value of alpha depends on data set.

Lasso regression: It is similar to ridge and smaller value of alpha is preferred and also by using lasso regression, it leads some of the coefficient to zero and it will make feature selection easier. Cross validation can be used to find the optimal value of alpha.

If we double the value of alpha for ridge regression then the predictors will be retained and in lasso regression if we double the value of alpha it will select only relevant features
Ridge retains all predictors, so the most important predictor variables will remain unchanged. However, their impact on the response variable diminishes due to regularization. Lasso selects relevant features, effectively performing feature selection. The remaining predictors having non-zero coefficients are the most influential.

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:
The choice between Ridge and Lasso should be based on the specific use case and the nature of the dataset. If we want to keep all the predictors, then Ridge is a better tool to use and if you are okay with excluding some variables, Lasso should be the choice. It's always a good idea to try both methods and see which one works best on the validation dataset.

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:
After building the model, the five most important predictor variables in the lasso model are:
1stFlrSF, 2ndFlrSF, OverallQual, OverallCond, SaleCondition_Partial.

After excluding the five most important predictor variables, the new five most important predictor variables are:
BsmtFinSF1, LotArea, BsmtUnfSF, GarageArea, KitchenQual

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:

- Diverse and Representative Dataset: Use a dataset that covers various scenarios and conditions to avoid overfitting.

- Data Augmentation: Enhance the model's ability to generalize by exposing it to different variations of the data.

- Cross-Validation: Employ methods like k-fold cross-validation to assess how well the model will generalize to an independent dataset.

- Regularization: Apply techniques such as L1 and L2 regularization to reduce overfitting by penalizing large coefficients.

- Hyperparameter Tuning: Use techniques like grid search or random search to find the optimal set of hyperparameters that yield a more generalized model.

- Evaluation Metrics: Don't solely rely on accuracy as it can be misleading, especially in imbalanced datasets. Use metrics like precision, recall, F1-score, and Area Under the ROC Curve (AUC-ROC) for a more holistic view of the model's performance.

- Ensemble Methods: Utilize methods that combine predictions from multiple models to enhance robustness and generalizability.