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?

Optimal Values of Ridge and Lasso Given by the model is 3.0 and 1.0 respectively.
f we increase the alpha(hyper parameter value) the accuracy of the model starts dropping gradually. It might increase a bit till the optimal hyper parameter value but the accuracy will decrease with the increase in alpha and model will become more parse

```
Top predictor features for ridge when alpha is 6 are :

                      Coefficient
Total_sqr_footage        0.134351
OverallQual              0.088469
TotalBsmtSF              0.079235
Neighborhood_StoneBr     0.058877
TotRmsAbvGrd             0.046711
Total_Bathrooms          0.045971
OverallCond              0.044430
GarageArea               0.044175
LotArea                  0.038149
Neighborhood_NoRidge     0.035671
```

```
Top correlated features of Lasso when alpha is 0.0002 are:

                        Coefficient
Total_sqr_footage          0.225682
OverallQual                0.133786
Neighborhood_StoneBr       0.065950
TotalBsmtSF                0.060540
OverallCond                0.058713
Neighborhood_NridgHt       0.044229
GarageArea                 0.043697
Neighborhood_NoRidge       0.035138
SaleCondition_Partial      0.032983
BsmtExposure_Gd            0.030938
```

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?

Although the R-squared scores for Ridge and Lasso models are nearly identical, we favor Lasso as our final model. Lasso imposes a stronger penalty on the dataset, potentially aiding in feature elimination. Given its ability to perform feature selection, we find Lasso more appealing for our specific modeling objectives.

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?

Before removing the top5 predictors, model was performing well than after dropping the top5 predictors.

```
                        Coefficient
Total_sqr_footage          0.225682
OverallQual                0.133786
Neighborhood_StoneBr       0.065950
TotalBsmtSF                0.060540
OverallCond                0.058713
Neighborhood_NridgHt       0.044229
GarageArea                 0.043697
Neighborhood_NoRidge       0.035138
SaleCondition_Partial      0.032983
BsmtExposure_Gd            0.030938
r2 score 0.9369076583225618
```

```
Applying Lasso after dropping 5 predictor variables
                        Coefficient
TotRmsAbvGrd               0.129455
Total_Bathrooms            0.112029
GarageArea                 0.105183
Fireplaces                 0.051469
LotArea                    0.048502
Street_Pave                0.047245
Neighborhood_NoRidge       0.044297
BsmtExposure_Gd            0.042970
BsmtUnfSF                  0.040704
SaleCondition_Partial      0.033489
r2 score 0.9050034157050509
```

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?

Ensuring that a model is robust and generalizable is crucial for its performance on new, unseen data.

1.Use techniques like k-fold cross-validation to assess how well the model performs on different subsets of the training data. This helps in detecting overfitting and provides a more reliable estimate of the model's performance.

2.Use regularization techniques (e.g., Ridge or Lasso regularization) to prevent the model from becoming too complex and overfitting the training data.

3. Tune the hyperparameters of the model using techniques like grid search or random search.

4. Besides accuracy, evaluate the model on various metrics such as precision, recall, F1 score, or area under the ROC curve. This provides a more comprehensive understanding of the model's performance.

5.Striking the right balance between bias and variance is essential. High bias (underfitting) and high variance (overfitting) both lead to poor generalization. Techniques like regularization help control this trade-off.