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?

The optimum values are
Ridge: 9
Lasso: 0.0001

If the values are doubled we see a decrease in both R^2 values of train & test datasets

Important features after change

Ridge

	coeff
OverallQual	0.069220
GrLivArea	0.056240
GarageCars	0.052744
Neighborhood_NoRidge	0.051368
Neighborhood_NridgHt	0.049653

Lasso

	coeff
GrLivArea	0.304106
OverallQual	0.151445
Neighborhood_NoRidge	0.076612
Neighborhood_NridgHt	0.072056
RoofMatl_WdShngl	0.069637

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?

We will choose model created using Lasso regression. This is because

- R^2 for training and test sets is higher than other models
- Lasso eliminates less important features from the model by making their coefficients 0. This makes the model simple by reducing variance.
- The model is more resilient in nature compared to ridge

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?

Assuming we have chosen Lasso model, the most important features are

	coeff
GrLivArea	0.335073
OverallQual	0.136617
RoofMatl_WdShngl	0.111923
Neighborhood_NoRidge	0.072496
Neighborhood_NridgHt	0.072127

After removing the features and retraining the model (including finding new lambda for lasso), new Important features are

	coeff
TotalBsmtSF	0.319153
2ndFlrSF	0.166349
GarageCars	0.100258
LotArea	0.088457
Exterior2nd_ImStucc	0.063517

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 shouldn't be a substantial disparity between the training R^2 and test R^2. If the training R^2 is considerably higher than the test R^2, it indicates that the model has overfitted and is incapable of generalizing. Such models tend to be complex with very high variance, meaning that even minor changes to the inputs can lead to significant deviations in the output.

Regularization techniques, such as Lasso, help eliminate non-significant features from the dataset by setting their coefficients to zero. This simplifies the model, reducing variance. It's important to note, however, that an excessively simple model may lack the robustness to capture the underlying patterns in the training data, leading to what is known as a high bias model.

Therefore, striking a balance between variance and bias is essential to achieving a generalized and robust model. An optimal model complexity, achieved through techniques like regularization, ensures that the model neither overfits nor oversimplifies, allowing it to generalize well to new data while still capturing meaningful patterns.