Q.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 optimal value of alpha for,

1. Ridge Regression: *4.0*
2. Lasso Regression: *0.0001*

If we double the value of alpha, it would result in a stronger regularisation. In case of,

1. Ridge Regression: Higher alpha values are more likely to aggressively compress the coefficients, thus lowering the impact of less significant variables. It can also help to reduce overfitting by reducing model complexity, but if set too high, it can lead to underfitting.
2. Lasso Regression: Lasso tends to drive more coefficients to exactly zero as alpha increases, essentially accomplishing feature selection. As a result, the final model may contain fewer active features, resulting in a simpler and more interpretable model. Extremely high alpha values, on the other hand, may result in excessive feature deletion, perhaps resulting in underfitting.

When we double the value of alpha, the most important predictor variables would be,

1. Ridge Regression: *OverallQual* is the predictor variable with the highest absolute coefficient i.e. *0.087444*
2. Lasso Regression: *GrLivArea* is the predictor variable with the highest absolute coefficient i.e. *0.315266*

Q.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?

Consider `model_cv` being the `GridSearchCV` instance variable, then by using `model_cv.best_params_` we get the optimal value for the Ridge / Lasso Regression. This uses k-fold validation technique to identify the optimal value of the lambda. The value returned would be best suited for the regularisation as it would balance overfitting the model and underfitting the model.

Q.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?

The table below shows the five most important predictor variables in the lasso model before the change,

Feature	**Coefficient**
Condition2_PosN	-0.340414
GrLivArea	0.339689
OverallQual	0.123430
RoofMatl_WdShngl	0.074416
GarageCars	0.067505

The table below shows the five most important predictor variables in the lasso model after removing the above features and retraining the lasso model,

Feature	**Coefficient**
TotalBsmtSF	0.256882
2ndFlrSF	0.180041
Neighborhood_NoRidge	0.075534
Neighborhood_StoneBr	0.069575
Neighborhood_NridgHt	0.056958

Q.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?

To make sure that a model is robust and generalisable, we need to consider the following,

1. *Feature Engineering*: Dimensionality reduction, feature selection, and transformation strategies can assist the model focus on the most essential parts of the data, decreasing noise and boosting generalisation.
2. *Regularisation Techniques*: Regularisation approaches like Ridge/Lasso regression regularisation can help prevent overfitting, which occurs when a model becomes too particular to the training data and fails to generalise well to new cases. These strategies impose limits or penalties on the model, discouraging it from relying too heavily on any single feature or pattern in the training data.
3. *Cross-Validation*: Cross-validation involves splitting the data into many subsets and training the model on various combinations of these subsets. This assists in determining how effectively the model generalises to previously unseen data and provides insight into its performance on diverse data samples.
4. *Evaluation on Unseen Data*: It is critical to evaluate the model's generalisation capabilities by testing its performance on a distinct, unbiased test set that was not utilised during training. This ensures that the model's accuracy is evaluated on new and previously unseen cases, accurately representing its real-world performance.