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 optimal value of the regularisation parameter (alpha) was determined using cross-validation. For Ridge regression, the optimal alpha value is 10, while for Lasso regression, it is 100. These values provide the best balance between model complexity and prediction error on unseen data.

If the value of alpha is doubled (Ridge = 20, Lasso = 200), the strength of regularisation increases. In Ridge regression, this results in further shrinking of coefficient magnitudes, but all predictors continue to remain in the model. The model becomes slightly more biased but more stable, with reduced sensitivity to multicollinearity.

In contrast, doubling alpha in Lasso regression has a stronger effect. More coefficients are driven exactly to zero, leading to a simpler model with fewer active predictors. While this improves interpretability and reduces variance, it may also remove some moderately important variables, potentially increasing bias.

After doubling alpha, the most important predictor variables remain largely consistent, though fewer in number for Lasso. Variables such as OverallQual, GrLivArea, TotRmsAbvGrd, GarageCars, and premium Neighborhood indicators continue to have the strongest influence on house prices. This indicates that quality, size, and location remain dominant drivers of pricing even under stronger regularisation.

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?

Between Ridge and Lasso regression, Lasso regression is the preferred choice for this problem.

Lasso not only performs regularisation but also automatically performs feature selection by shrinking some coefficients exactly to zero. This results in a simpler and more interpretable model, which is highly valuable for business stakeholders who want to understand which factors truly drive house prices.

Although Ridge regression performs well in handling multicollinearity and often produces slightly better predictive stability, it retains all features in the model. This makes interpretation more difficult, especially when dealing with a large number of correlated predictors created through one-hot encoding.

Given the business objective of identifying key price-driving variables and understanding pricing dynamics in a new market, Lasso regression offers a better trade-off between accuracy and interpretability.

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?

After removing the original five most important predictor variables from the Lasso model and rebuilding the model using the remaining features, the importance shifts to other variables that still capture strong price-related signals.

Based on the coefficients of the refitted Lasso model, the five most important predictor variables are:

1. 1stFlrSF – First floor living area
2. 2ndFlrSF – Second floor living area
3. MasVnrArea – Masonry veneer area
4. GarageCars – Garage capacity in number of cars
5. Neighborhood_StoneBr – Indicator for the Stone Brook neighborhood

Question 4

How can you make sure that a model is robust and generalizable? What are the implications of the same for the accuracy of the model and why?

A model can be made robust and generalizable by ensuring that it performs well not only on training data but also on unseen data. This can be achieved through techniques such as train-test splitting, cross-validation, and regularisation.

Cross-validation helps assess model performance across multiple data subsets, reducing the risk of overfitting to a specific sample. Regularisation techniques like Ridge and Lasso penalise overly complex models, preventing large coefficient values that may capture noise instead of true patterns. Proper preprocessing, such as scaling numerical variables and avoiding data leakage, further improves generalizability.

The implication of building a robust model is that training accuracy may be slightly lower than that of an overfitted model. However, this trade-off is desirable because the model performs more consistently on new data. A slightly less accurate but stable model is far more valuable in real-world decision-making