Questions

Q1

1. The optimal value of alpha in Ridge and Lasso regression is not a fixed value and depends on the specific dataset and problem. Alpha is a hyperparameter that controls the strength of regularisation in these regression techniques. It is usually determined that different values of alpha are tested, and the one that gives the best performance on a validation set is selected.

   In Ridge regression, increasing the value of alpha penalises the sum of squared coefficients, leading to a more pronounced shrinkage of the coefficients. In Lasso regression, increasing alpha not only penalises the sum of squared coefficients but also encourages sparsity by pushing some coefficients to exactly zero.

   If double the value of alpha for both Ridge and Lasso regression, the regularisation effect would be stronger. This would result in more aggressive shrinking of the coefficients and potentially lead to a simpler model with fewer predictors. In the case of Lasso, some coefficients might even be forced to exactly zero, effectively removing certain predictors from the model.

   The most important predictor variables after implementing such a change would be the ones whose coefficients survive the increased regularisation. These would be the variables that still contribute significantly to the model despite the stronger penalty on the coefficients.
   The 5 important predictors are: Overall Quality of house : Very Excellent, LotArea - Lot size in square feet, Type of foundation - Wood, Neighborhood NoSeWa, Masonry veneer area in square feet

Q2

2. In this specific dataset I use Lasso over Ridge because this dataset contains too many features after dummy variable creation, and also in my case both Lasso and ridge have almost similar r2 score. With the use of Lasso we can reduce the irrelevant features significantly without compromising the accuracy much, so I would prefer Lasso.

Q3

3. After excluding the 5 most important predictors, I will choose the next 5 important variables present based on their coefficient value. Those are Neighborhood NoRidge, Neighborhood StoneBr, Exterior covering on house: Imitation Stucco, Number of fireplaces, Overall housing quality - good.

Q4

4. A robust and generalizable model is likely to perform well on both the training and testing/validation datasets. a balance needs to be struck between model complexity and the ability to generalise to new, unseen data. Techniques such as cross-validation, proper feature engineering, regularisation, and hyperparameter tuning play essential roles in achieving a robust and generalizable model.