Question 1

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?

A: The most optimal value for the regularization parameter alpha (λ), in ridge and lasso regression are mostly determined by using techniques like cross-validation where the data is split into train and validation set to compare the model's performance under different regularization strength.

For Ridge Regression:

Objective function=$\sum_{i=1}^{n}(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij})^2+\alpha\sum_{j=1}^{p}\beta_j^2$

For Lasso Regression:

Objective function=$\sum_{i=1}^{n}(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij})^2+\alpha\sum_{j=1}^{p}|\beta_j|$

Alpha is the regularization parameters which controls the strength. Increase in alpha, increases the regularization strength.

For finding the optimal regularization, we can also use k-fold cross-validation.

If we double the value in both the Ridge and Lasso regression:

Ridge: More penalization, increased shrinkage, improved multicollinearity handling.

Lasso: Greater push towards zero, enhanced sparsity, potential feature selection.

The most important predictor variables after the changes:

Ridge : All predictors contribute with reduced impact.

Lasso : Some predictors become zero, selecting a subset.

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?

A. Ridge regression's optimal lambda is 2, reducing errors and minimizing test error. Doubling lambda to 10 increases penalty for a more generalized model. In Lasso regression, a small lambda of 0.01 is chosen, penalizing more and reducing more coefficients to zero, decreasing R2 square. Post-implementation, key variables are: Ridge Regression: MSZoning_FV, MSZoning_RL, Neighborhood_Crawfor, GrLivArea, etc. Lasso Regression: GrLivArea, OverallQual, TotalBsmtSF, LotArea, etc.

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?

A: After constructing the model, it was identified that the five most crucial predictor variables in the Lasso model are absent in the incoming data. Thus further excluded variables are: TotalBsmtSF,GarageArea, GrLivArea, OverallQual and OverallCond.

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?

A: To enhance model robustness, prioritize simplicity, strike a balance between bias and variance, employ cross-validation, and assess accuracy on a separate test set. Implications include a trade-off between training accuracy and generalization, with an emphasis on achieving a balance between bias and variance for overall improved accuracy