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?

Answer1:

Optimal value of alpha for ridge: 20 **(if we use selected (after VIF and RFE) variables)**

Optimal value of alpha for ridge: 500 (if we use all variables)

Optimal value of alpha for ridge: 100 **(if we use selected (after VIF and RFE) variables)**

Optimal value of alpha for ridge: 1000 (if we use all variables)

After doubling the alpha values for Ridge and Lasso regression to 40 and 200 (1000 and 2000):

For Ridge regression, increasing the alpha parameter leads to more significant shrinkage of the coefficients. This helps in reducing the impact of multicollinearity by effectively penalizing large coefficients. As alpha increases, the model becomes more robust to multicollinearity, but there's a trade-off because too much regularization can lead to underfitting.

For Lasso regression, increasing the alpha value results in the removal of more features from the model. This effect is particularly pronounced when working with a larger number of variables, as higher alpha values lead to more aggressive feature selection.

Top features using selected variables after RFE and VIF:

Top Features: '2ndFlrSF', '1stFlrSF', 'OverallQual', 'Neighborhood_NridgHt', 'BsmtFinSF1'

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?

Answer 2:

The choice between Ridge and Lasso regression depends on the specific requirements and objectives of your analysis. Both techniques have their strengths and weaknesses, so it's essential to carefully consider your dataset and objectives before making a decision. If interpretability and retaining all features are crucial, Ridge regression might be the better choice. If feature selection and identifying a sparse set of predictors are more important, Lasso regression might be preferable.

In this case, we will choose **Lasso** as its giving **feature selection** option also. It has removed unwanted features from model without affecting the model accuracy. Which makes are model generalized and simple and accurate.

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?

The top 5 features are **'2ndFlrSF', '1stFlrSF', 'OverallQual', 'Neighborhood_NridgHt', 'BsmtFinSF1'**. After dropping them model accuracy has reduced significantly. Now topmost features are: Next top 5 features after droping 5 main predictors **'TotalBsmtSF', 'KitchenQual', 'GarageCars', 'BsmtQual', 'Neighborhood_NoRidge'**

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?

Model robustness and generalizability involves a combination of careful data preprocessing, model selection, hyperparameter tuning, and evaluation on unseen data. While this may result in a slight trade-off in accuracy on the training set, it ultimately leads to more reliable and trustworthy predictions in real-world applications.

To make the model robust and generalize below features are required:

1. **Model accuracy** should be in cross validated and consistent across test and train data.
2. **P-value** of all the features is < 0.05. (We have checked that)
3. **VIF** of all the features are < 5. (checked that)
4. **Residuals analysis**

Thus we are sure that model is robust and generalizable.