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?

Answer 1 - The optimal alpha value for Ridge regression was initially determined to be 10, and for Lasso regression, it was 100.
Subsequently, we doubled these alpha values to 20 for Ridge and 200 for Lasso.
In the case of Ridge regression, as the alpha value increased to 20, the coefficients of the features also increased. However, this change led to a noticeable decrease in the R2 score for the training data, dropping from 0.807 to 0.45.
On the other hand, for Lasso regression, the higher alpha value of 200 resulted in the removal of more features from the model. Although this feature selection process led to a slight decrease in the R2 score by about 1% for both the test and training data, the model still maintains its predictive capabilities.
The top influential features in the model, despite the alpha adjustment, include Neighborhood_NoRidge, Neighborhood_NridgHt, OverallQual, and Neighborhood_Veenker, indicating their continued significance in explaining the target variable.

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 - We have decided to opt for Lasso regression due to its unique feature selection capability. Lasso effectively eliminates unnecessary features from the model while maintaining its accuracy. This results in a more generalized, simplified, and accurate model. The removal of redundant or less influential features enhances model interpretability and efficiency, making it a preferred choice for our predictive analysis.

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?

Answer 3 - After identifying and removing the top 5 features, namely Neighborhood_NoRidge, Neighborhood_NridgHt, 2ndFlrSF, OverallQual, and Neighborhood_Veenker, from the model, there was a notable reduction in model accuracy. The accuracy decreased from approximately 80% and 81% to 55% and 58% for the training and test data, respectively.
Subsequently, the top 5 features that emerged as the most influential after dropping the initial 5 main predictors are 1stFlrSF, MSSubClass_90, MSSubClass_120, TotalBsmtSF, and HouseStyle_1Story. These features have taken precedence in explaining the target variable and have become the primary drivers of the model's predictive power.

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?

Answer 4 - To ensure that our model is robust and capable of generalization, we've established three key criteria:

1. **Model Accuracy**: The model's accuracy should exceed 70-75%, which is indeed the case in our scenario, with an accuracy of 80% for the training dataset and 81% for the test dataset. This high accuracy indicates that the model is performing well on both the data it was trained on and unseen data.

2. **P-Values**: The p-values of all the features should be less than 0.05. This condition ensures that the features included in the model are statistically significant and have a meaningful impact on the target variable. By having low p-values for all features, we can be confident in the relevance of the chosen predictors.
3. **VIF (Variance Inflation Factor)**: The VIF for all features should be less than 5. VIF measures multicollinearity, which can lead to unstable coefficient estimates. A VIF value below 5 indicates that the features are relatively independent of each other, contributing to a more stable and reliable model.

Meeting these three criteria collectively assures us that our model is both robust and capable of generalization, as it exhibits strong predictive performance, statistical significance of features, and low multicollinearity among predictors.