Q.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?

Ans. For ridge: alpha = 4.64, For Lasso: alpha = 0.00046.

After doubling the lambda values for both the regressions (ridge and lasso). There is small deep in the r2 score of approximately 0.01 in both the train and test results of lasso regression. Whereas, in ridge regression, this change is seen only in train results.

The top 5 significant variables in both the regressions did not change. Just the 3rd and 4th most important variable interchanged in Ridge regression. The variables are listed below:

Ridge –

1. OverallQual
2. GrLivArea
3. Neighborhood
4. 1stFlrSf
5. 2ndFlrSf

Lasso –

1. GrLivArea
2. OverallQual
3. Neighborhood
4. KitchenQual
5. BsmtExposure

Q.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?

Ans. I will choose Lasso regression over Ridge regression. Because, the house price prediction data has a smaller number of records and a higher number of independent variables. As we know lasso eliminates some variables by assigning 0 to some insignificant variables (depending on the value of alpha), Although there is not much difference in r2 scores for both the regressions and the feature size was reduced from 67 to 29, A simpler model is always better for the sense of interpretability. That's why I will choose Lasso over Ridge.

Q.3 After building the model, you realized 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?

Ans. Now the five most important variables are:

1. 1stFlrSf
2. 2ndFlrSf
3. ExterQual
4. GarageCars
5. MasVnrArea

Q.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?

Ans. There are some important points that should be considered.

1. Missing Values Imputation: The columns which are having more than 80% null values only should be removed as a general rule of thumb, but it should be discussed with business. For the numerical columns data visuals could help in choosing the appropriate statistic to fill the values. For categorical columns mode imputation or custom imputation could work.
2. Feature scaling: Feature scaling brings the variables to similar scale; it controls the range of values for coefficients and hence improves the readability and stability of the model.
3. Lasso and Ridge Regression: These methods can be used to overcome overfitting of the model, Ridge can be used when there is lesser number of variables and all the variables are significant for the business, while lasso can be used if the dimensionality is very high and some features can be removed and not important.

Implication for accuracy:

1. Regularization can slightly decrease the training accuracy as it minimizes the coefficient as well with the error term. But it provided better generalizability and hence improve the model performance towards unseen data.
2. Properly executing the analysis for missing values and feature scaling will improve the model's stability.