Subjective Questions

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. According to our modeling the optimum alpha value for ridge and lasso regression, obtained by Grid Search is as follows:

- Ridge - 15
- Lasso - 120

As observed in the notebook, when we double the optimum alpha value for ridge regression from 15 to 30, the following happens:
Our evaluation metrics, r-squared and mse take a drop. R-squared is reduced by roughly 3 points on train and test sets. We also observe that the value of the most impacting negative and positive coefficients has also changed. Although the change is not very significant, there is change nonetheless.
As observed in the notebook, when we double the optimum alpha value for lasso regression from 120 to 240, the following happens:
Our evaluation metrics, r-squared and mse take a drop. R-squared is reduced by less than a point on train and test sets. We also observe that the value of the most impacting negative and positive coefficients has also changed. Although the change is not very significant, there is change nonetheless. We also observe after.
After running the tests on both the models, we observe that although our accuracy did take a hit after we doubled the alpha, the approximate top features remained the same, although some change in their coefficients were observed.

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. According to our modeling the optimum alpha value for ridge and lasso regression, obtained by Grid Search is as follows:

- Ridge - 15
- Lasso - 120

While modeling the sales price using ridge regression, we used RFE and VIF for feature elimination and multicollinearity reduction. After building the model we get an r-squared value of ~0.60 on the test set.
While modeling the sales price using ridge regression, we did not perform any manual or automatic feature selection process as lasso regression pushes the coefficient values of insignificant features to 0. After building the model we get an r-squared value of ~0.89 on the test set.
With the evidence of a better r-squared and no need of special feature elimination, we'll be choosing the model built using lasso regression.

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?

For the model built using lasso regression, the 5 most important predictor features are:

OverallQual	12282.811
BsmtExposure_Gd	12790.763
SaleCondition_Partial	16568.669
GrLivArea	18315.842
KitchenAbvGr	-18076.763

After removing these 5 predictors from the model, the most important predictor features are as follows:

SaleType_New	18341.458
2ndFlrSF	22597.902
Neighborhood_NoRidge	28557.180
Neighborhood_NridgHt	34991.987
Neighborhood_StoneBr	41485.555

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 - A robust model implies that it makes few or acceptable errors/biases on the test and other unseen data. Whereas a generalisable model implies that a model shouldn't suffer from high variance. It should not change a lot when the input data changes.

Occam's Razor suggests that machine learning should prefer simple models with few coefficients over complex models like ensembles. At face value, Occam's Razor is a heuristic, suggesting that more complex hypotheses create more assumptions.

We can achieve this by using techniques like regularization to keep the models complexity in check and prevent it from growing very complex.