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: Optimal Values of Alpha

Ridge - 0.3
LAsso - 0.0001

What will be changes when we choose double value of alpha for both ridge and lasso?

When double the value of alpha

Ridge - 0.6
LAsso - 0.0002

The important variables after the change is implemented are:

MSZoning_FV
MSZoning_RL
GrLivArea
OverallQual
TotalBsmtSF
Neighborhood_Crawfor
Foundation_PConc
Neighborhood_NridgHt
SaleCondition_Normal
GarageCars

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. I got similar values for both 'Ridge' and 'Lasso'
b. Alpha values are '0.3' and '0.0001'
c. Lasso Zerod One or two coefficients so Lasso is better.

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: I excluded 5 important variables :

a) MSZoning_FV
b) GrLivArea
c) MsZoning_RL
d) OverallQual
e) Foundation_Pnoc

After the Lasso regression, I got another Prediators :

a) OverAll condition
b) Lot Area
c) Lot Shape

d) Condition 1
e) Is Remodeled

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?

Test accuracy close to the training score. Ensuring the model's resilience against outliers is paramount for its robustness. Detecting outliers through methods like box plots and Z scores is essential. Addressing outliers maintains the integrity of statistical measures such as mean and median, aiding in accurate imputation of missing values. Outlier analysis should be selective, focusing only on pertinent outliers to standardize model predictions. A robust model is imperative for reliable predictive analysis.
Significance of predicted variables is crucial. Model significance can be assessed through metrics such as P-values, R2, and adjusted R2. Simplicity in model design often enhances robustness.
Implications of Model Accuracy:

Maximize data acquisition:
Leveraging larger datasets enables the model to train more effectively, reducing reliance on weak correlations and assumptions.
Address missing values and outliers:
Incomplete or erroneous data can compromise model accuracy. Outliers, in particular, can skew statistical measures used for imputation. Identifying outliers, perhaps through box plots, and subsequently treating them enhances model accuracy.

Feature engineering or standardization:

Feature Selection:
Feature selection relies heavily on domain knowledge to identify impactful variables affecting the target variable. Data visualization aids in this process by providing insights into feature relevance. Statistical parameters like p-values and Variance Inflation Factor (VIF) help determine the significance of variables.

Choosing the Right Algorithm:
Selecting an appropriate machine learning algorithm is paramount for achieving accurate models. This decision often requires experience and experimentation to determine the algorithm best suited to the data and problem at hand.

Cross Validation:
Sometimes, striving for higher accuracy can lead to overfitting. Cross-validation techniques help mitigate this risk by leaving out a portion of the data from model training and testing it separately before finalizing the model. This approach ensures that the model's performance is robust and generalizable.