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?

In the case of ridge regression:- When we plot the curve between negative mean absolute error
and alpha we see that as the value of alpha increase from 0 the error term decrease and the train error
is showing increasing trend when value of alpha increases .when the value of alpha is 2 the test error is
minimum so we decided to go with the value of alpha equal to 2 for our ridge regression.

For lasso regression I have decided to keep very small value that is 0.01, when we increase the value of
Alpha the model tries to penalise more and try to make most of the coefficient value zero. Initially it came
as 0.4 in negative mean absolute error and alpha.
When we double the value of alpha for our ridge regression no we will take the value of alpha equal to
10 the model will apply more penalty on the curve and try to make the model more generalised that is
making model more simpler and no thinking to fit every data of the data set .from the graph we can see
that when alpha is 10 we get more error for both test and train.
Similarly when we increase the value of alpha for lasso we try to penalize more our model and more
coefficient of the variable will reduced to zero, when we increase the value of our r2 square also
decreases.
The most important variable after the changes has been implemented for ridge regression are as
follows:-
1. MSZoning_FV

2. MSZoning_RL
3. Neighborhood_Crawfor
4. MSZoning_RH
5. MSZoning_RM
6. SaleCondition_Partial
7. Neighborhood_StoneBr
8. GrLivArea
9. SaleCondition_Normal
10. Exterior1st_BrkFace
The most important variable after the changes has been implemented for lasso regression are as
follows:-
1. GrLivArea
2. OverallQual
3. OverallCond
4. TotalBsmtSF
5. BsmtFinSF1
6. GarageArea
7. Fireplaces
8. LotArea
9. LotArea
10. LotFrontage

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?

- Ensuring the regularisation of coefficients is paramount, not only for enhancing prediction accuracy but also for mitigating variance and fostering model interpretability.
- **Ridge regression** employs a tuning parameter known as lambda, applying a penalty proportional to the square of coefficient magnitudes, a value determined through cross-validation. This penalty, lambda multiplied by the sum of squared coefficients, penalises larger coefficients. Increasing lambda reduces model variance while maintaining bias.

- Unlike Lasso Regression, Ridge regression retains all variables in the final model. Conversely, Lasso regression utilises lambda as a penalty on the absolute value of coefficient magnitudes, determined via cross-validation.
- As lambda increases, Lasso progressively shrinks coefficients towards zero, effectively rendering some variables as precisely zero. Lasso regression also facilitates variable selection. With small lambda values, it resembles simple linear regression, but as lambda increases, shrinkage occurs, leading the model to disregard variables with zero coefficients.

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?

Those 5 most important predictor variables that will be excluded are :

- GrLivArea
- OverallQual
- OverallCond
- TotalBsmtSF
- GarageArea

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?

- The pursuit of simplicity in model construction is imperative, even if it means a decline in accuracy. A simpler model typically offers greater robustness and generalizability, a concept underscored by the Bias-Variance trade-off. In essence, simpler models exhibit higher bias but lower variance, resulting in enhanced generalizability. This principle implies that a robust and generalizable model should demonstrate consistent performance across

both training and test datasets, maintaining relatively stable accuracy levels across both domains.

- **Bias**: Bias is an error in a model, when the model is weak to learn from the data. High bias means the model is unable to learn details in the data. Model performs poorly on training and testing data.
- Variance: Variance is error in model, when model tries to over learn from the data. High variance means the model performs exceptionally well on training data as it has very well trained on this of data but performs very poor on testing data as it was unseen data for the model.It is important to have balance in Bias and Variance to avoid overfitting and under-fitting of data.