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?

Ans: - The optimal value of alpha of lasso regression is 50 and r2 score for optimal value of alpha is given below:-

R2 score for train: 0.9372405328256925

R2 score for test: 0.9254664123086983

Optimal value of alpha of ridge regression is 4 and r2 score for optimal value of alpha is given below

R2 score for train: 0.9371096095852764

R2 score for test: 0.9253982765709686

I have got similar values for the first time for both Lasso and Ridge.

There is slight difference in the R2 score, RSS for Ridge regression when we double the value of alpha. A noticeable difference is there in the R2 score test and train, RSS in Lasso Regression when we double the value of alpha.

The important variables after this change is implemented are:MSZoning_FV

-> MSZoning_RL

-> GrLivArea

-> OverallQual

-> TotalBsmtSF

-> Neighborhood_Crawfor

-> Foundation_PConc

-> Neighborhood_NridgHt

-> SaleCondition_Normal

-> GarageCars

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?

Ans:- Optimal value of alpha/lambda is 1 for ridge regression on variables selected by lasso regession and r2 score for optimal value of alpha is given below:-

R2 score for train: 0.9392476999525955

R2 score for test: 0.9231948226312643

Lasso regression has successfully reduced variables by shrinking the variable coefficient to 0. There are 134 variables selected by lasso regression out of 255 variables.

Based on the alpha/lambda values I have got, Ridge regression does not zero any of the co efficient, Lasso zeroed one or two coefficients in the selected features, Lasso is better option and it also helps in some of the feature elimination. From the output of both models, it is evident that Lasso does feature selection and its output is much simpler than Ridge's, without any compromise in the accuracy over the test data. So, I will go with the 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?

Ans :- I have excluded the five most important variable I have got prior. Those are MSZoning_FV, GrLivArea, MSZoning_RL, OverallQual, Foundation_PConc. I have created a new model after removing these columns code is mentioned the Python notebook. After the Lasso Regression I have got the other important predictors are

Overall condition:-

- ➔ Lot area
- ➔ Lot shape
- ➔ Condition1
- ➔ IsRemodeled

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?

Ans :-The model should be as simple as possible. This might lead to a decrease in training accuracy but this will be making the model more robust and easier to generalisable. This can also be understood using the Bias variance trade-off.

The mode is robust and generalization when

- Test accuracy is not much lesser than the training score.
- The model should not be impacted by the outliers: Outlier treatment is most important to get the robust model. We can detect outliers in the dataset using box plots, Z score etc. Treating the outliers will not affect mean, median etc. so that we can impute correct values to missing values. The outlier analysis needs to be done and only those which are relevant. This would help standardize the predictions made by the model. If the model is not robust, it cannot be trusted for predictive analysis.
- **The predicted variables should be significant.**
  Model significance can be determined the P-values, R2 and adjusted R2.
  Always a simple model can be more robust.

Implications of Accuracy of a model:

- Gain the more data as much you can:
  Having more data allows the data to train itself, instead of depending on the weak correlations and assumption, it is good to have more data.
- Fix missing values and outliers:
  If the data has missing values and outliers can lead to inaccurate model. Outliers can affect the mean, median that we are imputing to continuous variables. You can get the outlier values using a boxplot, treating the outliers in the data will make our mode more accurate.
- Featuring Engineering or newly derived columns/Standardize the values:
  We can extract the new data from the existing data ex: from DOB we can get the Age of the person, after extracting the new data required, we can drop the existing features.
  Scaling the values: ex: one value is in meters, the other is Kilo meters, it is important to scale these feature into one standardized unit. If we did this, we can get accurate model.
- Feature Selection:
  It is purely based on the domain knowledge, so that we can select important features that have good impact on the target variable. Data visualization also helps the selecting the features. Statistical parameters like p-Values, VIF can give us significant variables.
- Applying the right algorithm
  Choosing the right machine learning algorithm is very important to get accurate model.