Dheeraj Goud Borlla (dxb160130)

MIDTERM PROJECT: (BUILDING A PREDICTIVE MODEL)

***PURPOSE*:** In this assignment, we analyse the Ames Housing dataset and try to build a Predictive model. First, we clean up the dataset and apply various techniques to achieve an accurate model that predicts the Sales Price of the houses.

DATASETS: Ames Housing Dataset.

APPROACH:

Data Cleaning / Data Scrubbing -

- The dataset has of total of 82 variables. Of these variables, I tried to build a model to predict 'SalePrice' variable.
- In the dataset, there are few variables those are factors. So, we need to apply the unclass function to change to factors.
- Now, we look through the Summary of the dataset. We see that there are lot of 'NA' in the Data. We can't just omit the NA's from the dataset, this would leave 0 rows in the dataset. So, we need to remove the variables that have lot of NA's.
- After removing the variables with the NA's, we checked again the summary of the updated dataset. There are lot of variables which don't have lot of variation in them. So, they don't add much to build a predictive model. So they are removed.
- The updated dataset after the removal of the unwanted variables, we see that there are still few NA's that are to be eliminated. We can use the function to eliminate those.
- After the process, the dataset is reduced to 2219 row from 2930 and to 56 variables from 82. So, there was lot of unwanted data that we removed.
- There may be still some of the variables that may be linearly dependent with other variables. So, I applied the collinearity matrix, from this matrix we can eliminate variables that are almost collinear to other means they are linearly dependent.
- I have gone through the summary of the dataset. And removed the variables that are collinear to other variables (in variation and in values) in summary. And the dataset is reduced to 49 variables.
- Finally, the plot between 'SalePrice' and 'Gr.Liv.Area' show that there are few outliers. The houses with 4000 sq ft in size are to be removed. There are hardly five of them and to be removed.
- So, Now the Dataset has been cleaned with the unwanted variables that has more NA's or has no variation in it that could help for a model or other variables that are collinear to other variables. The final updated dataset has 2214 rows and 49 variables.
- Now we go on with building the model for predicting 'SalePrice'.

***B*uilding the Model –**

- **### Basic Linear Regression Model ###**
- First, I have applied the linear regression model which gave a pretty good value of adjected R sq with value of 0.9126. But, we should not forget that adding more variables to the model can also increase the value of adjusted R sq. The MSE of the linear model is 548622452.
- We look at the residual plot of the linear model in graph 1. From the Normal Q-Q plot, we see non-linearity and the residuals and fitted values plot also shows non-linearity.
- So, we should try other techniques to build the more accurate model.
- **### Forward and Backward Subset Selection ###**
- In this process, we are using 'leaps' package. From this package, we will make use of 'regsubsets' function for model selection.
- First, I have applied the *forward* method. The plot of values of adj R sq and 'cp' are shown in the graph 2.
- From the graph, we see that the 46 variables count got the least 'cp' value.
- Next, we look at the *backward* selection method. The plot of the adj R Sq and cp values are shown in the graph 3.
- In this method, the variable count that got least 'cp' value is 47 variables.
- From the plot in graph 4, we see that removing few more variables will give us accurate model. That is even removal of few variables will not affect the variation in the 'SalePrice' of the houses. Because adding more variables can cause overfitting.
- We will go with some other techniques like 'Lasso' and 'Ridge'.
- **### RIDGE ###**
- This technique help us with the problem of overfitting by introducing additional information. In this method, we use lambda which controls the importance of regularization term.
- In this method, we use '*glmnet*' package. The input parameters to this function is a model matrix, response variable and a user specified 'lambda' sequence.
- When the *glmnet function* is applied with x and y and sequence of lambda alpha is 0 for ridge y default, we get many values for the lambda.
- To get the optimal lambda value, we used the Cross-Validation technique.
- We divide the data into training and testing sample. Then we apply the function '*cv.glmnet*' which does a k-fold cross validation, where k is 10 by default.
- The value for the optimal lambda is 7238.43. We find the MSE value for best lambda and the value is '487032310'.
- With the best lambda, we tried to build the accurate model. The coefficients that are included are as follows:
  Lot.Frontage, Lot.Area, Lot.Shape, Land.Contour, Lot.Config, Land.Slope, Bldg.Type, House.Style, Overall.Qual, Overall.Cond, Year.Built, Year.Remod.Add, Mas.Vnr.Area, Exter.Qual, Bsmt.Exposure, BsmtFin.Type.1, BsmtFin.SF.1, BsmtFin.Type.2, Bsmt.Unf.SF,

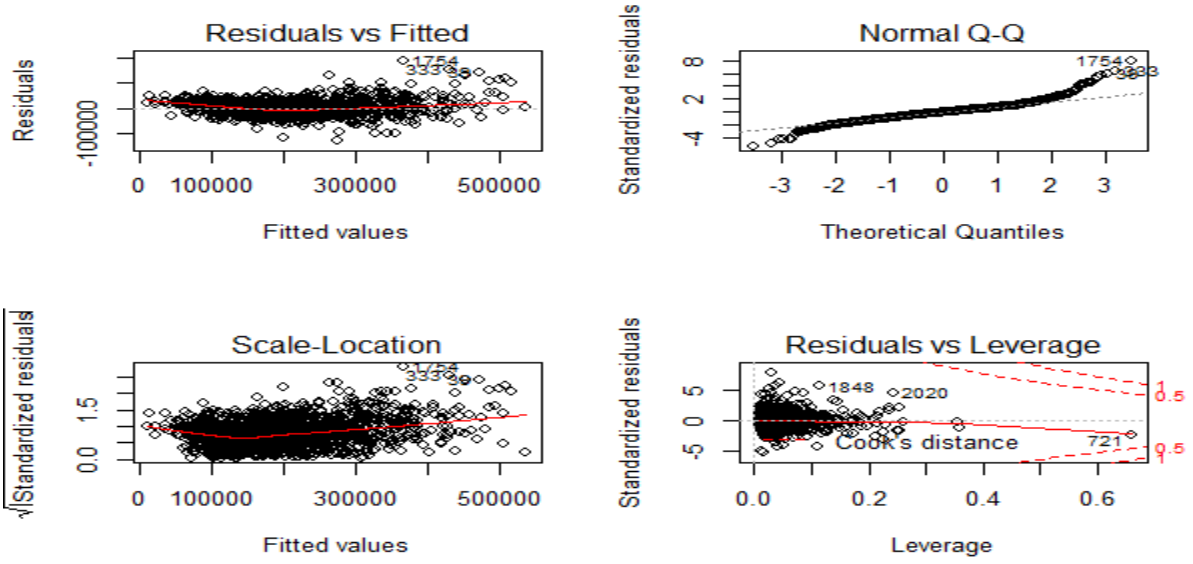
Total.Bsmt.SF, X2nd.Flr.SF, Gr.Liv.Area, Bsmt.Full.Bath, Bsmt.Half.Bath, Full.Bath, Half.Bath, Bedroom.AbvGr, Kitchen.AbvGr, TotRms.AbvGrd, Fireplaces, Garage.Type, Garage.Yr.Blt, Garage.Finish, Garage.Cars, Garage.Area, Wood.Deck.SF, Open.Porch.SF, Screen.Porch, Pool.Area, Yr.Sold.

- The Ridge technique gave model with 40 variables.

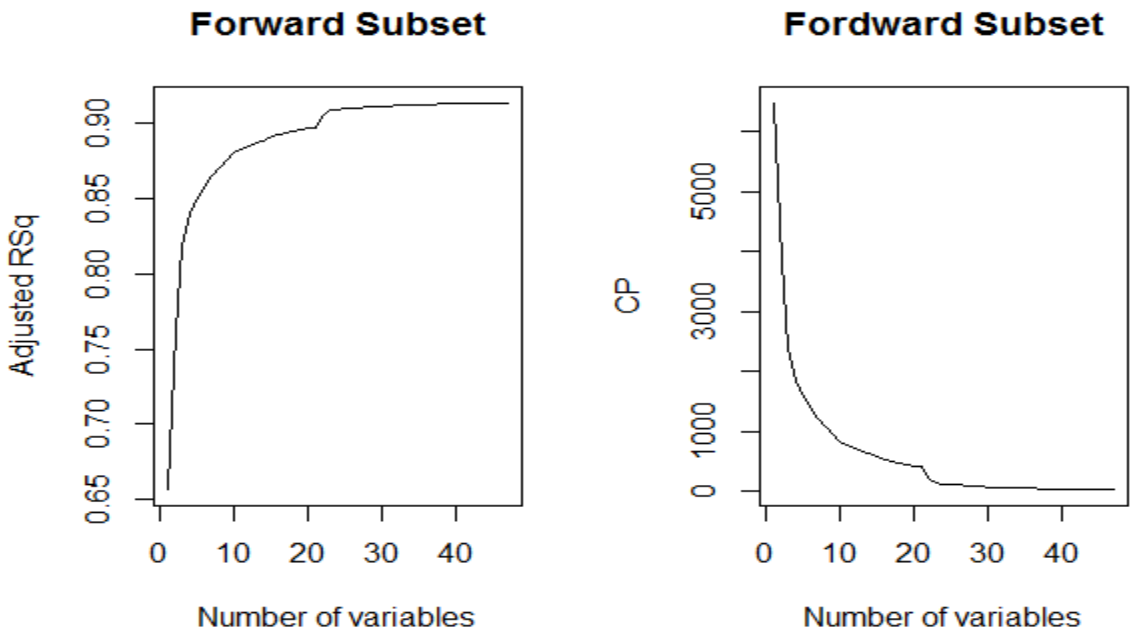
➢ **### LASSO ###**

- For this we use 'glmnet' function but the difference is with the value of alpha that is alpha value is 1 for lasso by default.
- The plot function when applied on the lasso mode gave the plot shown in graph 5.
- We should not forget that Lasso takes certain coefficient to zero for large enough lambda. So, we used Cross Validation technique to get best lambda. For this, we used 'cv.glmnet' function.
- The value of the optimal lambda is 272.52 and the MSE value for the obtained lambda is 462918567.
- With the best lambda value, we build a model. The coefficients that are not zero are: Lot.Frontage, Lot.Area, Lot.Shape, Land.Contour, Lot.Config, Land.Slope, Bldg.Type, House.Style, Overall.Qual, Overall.Cond, Year.Built, Year.Remod.Add, Mas.Vnr.Area, Exter.Qual, Bsmt.Exposure, BsmtFin.Type.1, BsmtFin.SF.1, BsmtFin.Type.2, Bsmt.Unf.SF, Total.Bsmt.SF, X2nd.Flr.SF, Gr.Liv.Area, Bsmt.Full.Bath, Bsmt.half.Bath, Half.Bath, Bedroom.Abvgr, Kitchen.Abvgr, TotRms.AbvGrd, FirePlaces, Garage.Type, Garage.Finish, Garage.Cars, Garage.Area, Wood.Deck.SF, Open.Porch.SF, Screen.Porch, Pool.Area, Yr.Sold.
- The final model obtained using the 'LASSO' technique has 38 Variables and explaining about 90% of the variation in SalePrice.

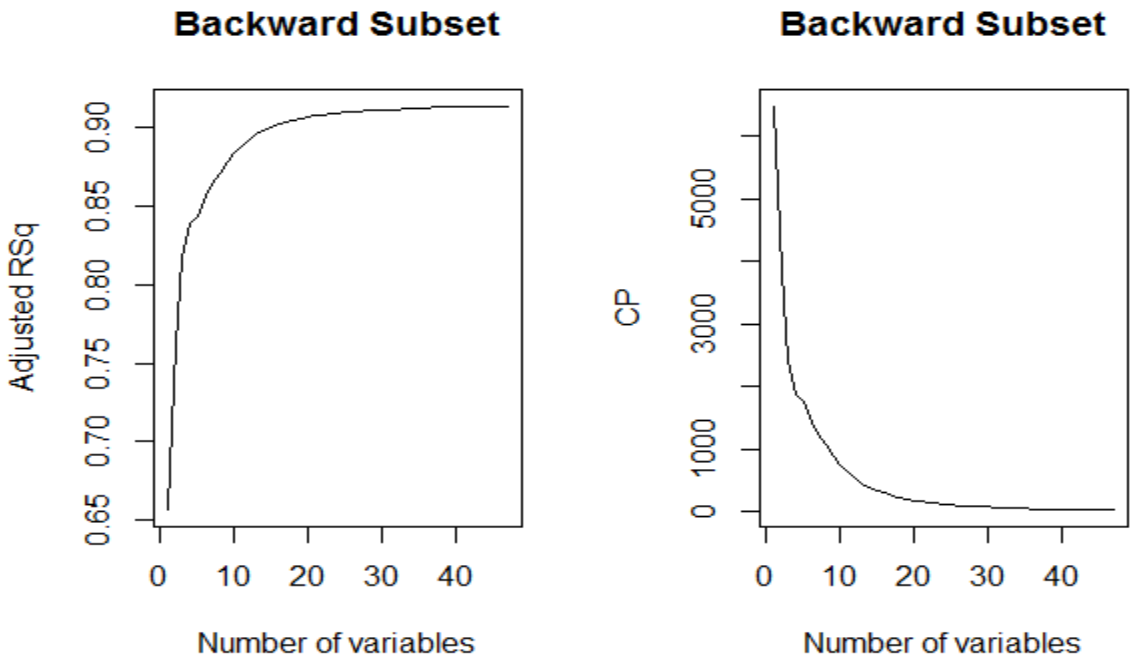
GRAPHS:



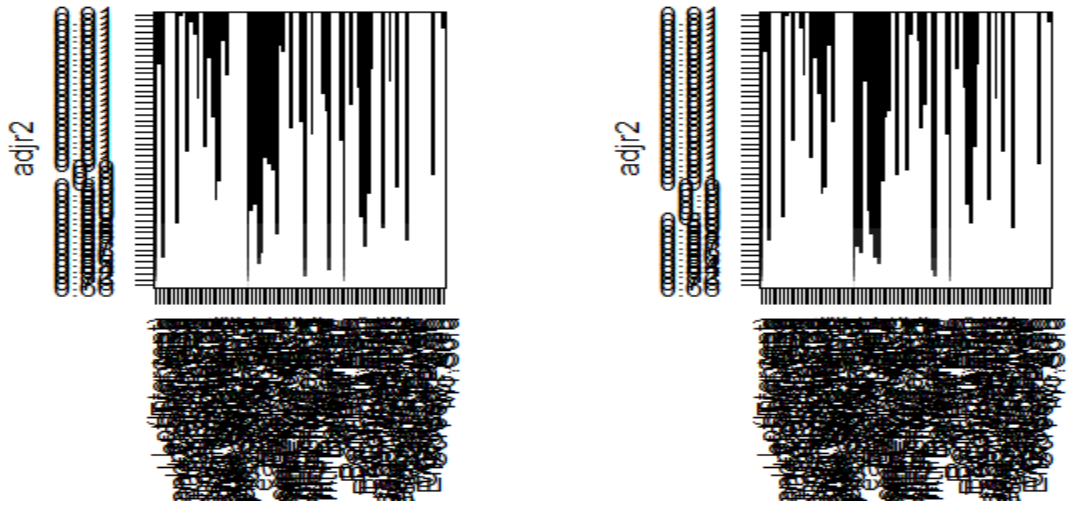
(Graph 1)



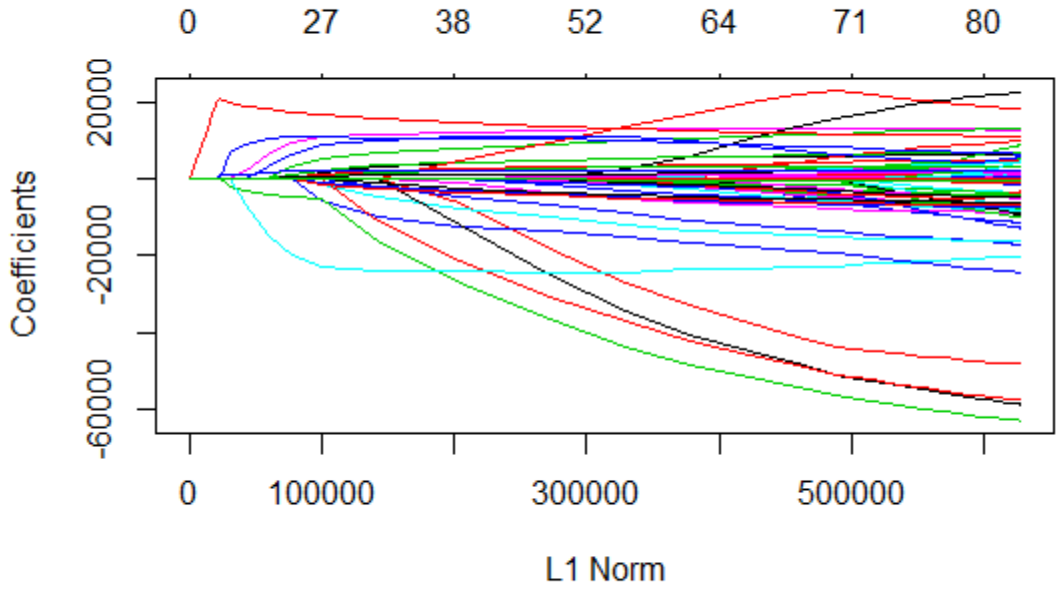
(Graph 2)



(Graph 3)



(Graph 4)



(Graph 5)

SUMMARY:

- In this Assignment, we tried to build a predictive model for the Ames Housing Dataset. The target variable is 'SalePrice'.
- First, we did data cleansing. The process of cleaning the data from unwanted variables. The dataset has lot of NA's in it. We removed the variables with many NA's, or the variables that does not have much variation in them and the variables that are collinear with other variables.
- Next, we applied linear regression model to the updated dataset. The adj R Sq of the model is 0.912 and MSE of the model is 548622452. Still we can get more accurate model using other techniques.
- Next, the forward subset and backward selection methods are used. The results that are obtained from these models doesn't seem accurate. So, I tried using other techniques.
- Finally, I applied Lasso and Ridge techniques to the dataset. From the Ridge model and Lasso Model, we see that the model obtained from Lasso is more accurate. The final model's MSE is less than the previous model and have 38 variables.