STAT 425- Final Project

Housing Price Prediction

Umesh Karamchandani (umeshk2)

Table of Contents

INTRODUCTION **2**

EXPLORATORY DATA ANALYSIS **2**

Univariate Analysis 3

Bivariate Analysis 5

MODELLING: METHODS **8**

Multiple Linear Regression 8

Multiple Linear Regression (Part 2) 10

3.3 Ridge Regression 11

3.4 Random Forest 12

CONCLUSION **12**

1. INTRODUCTION

Data Description: The dataset is extracted from UCI Machine Learning Repository and it consists of a market historical dataset of real estate valuation collected from Sindian District, New Taipei City, Taiwan. The variables in the original dataset are :

X1 - Transaction Date
X2 - The House Age
X3 - The Distance to the nearest MRT station
X4 - The number of convenience stores in the living circle
X5 - The latitude
X5 - The longitude
Y - House Price per unit area

The '*Transaction Date'* for example, is the date on which the house was sold. The *house age* is also one of the predictors that is used. This assumption is confirmed in our analysis. Variables Longitude and Latitude are the coordinates of the house in the locality. For the analysis, we divided the dataset into train and test with a ratio of 70 is to 30. The seed is set at 7 to maintain reproducibility.

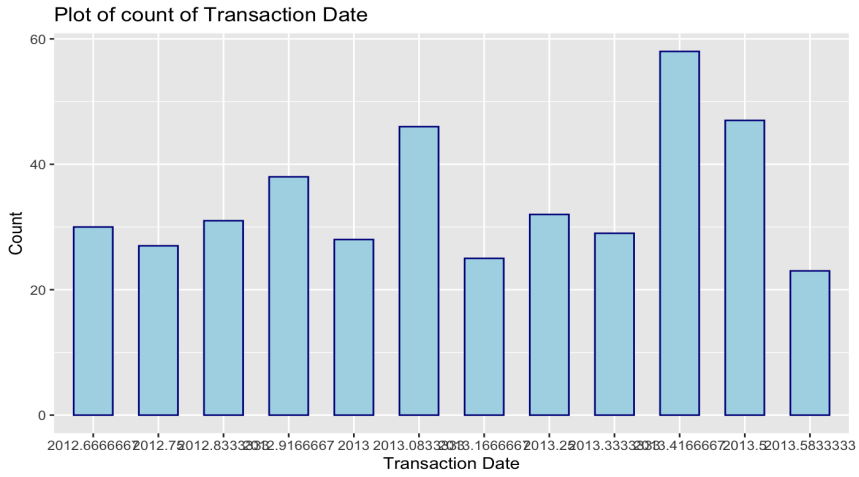
Project Goal: The aim of our project is to predict the price of a house in a locality. We are given certain features for each house and our aim is to build a machine learning model that is able to predict house price. Goal of our project is also to ensure that our model is interpretable.

2. EXPLORATORY DATA ANALYSIS

The total number of records in the dataset are 414 and the total number of features in the dataset are 7. There are 6 predictor variables and 1 response variable. One of the predictor variables is Transaction Date which has data values of float type. It contains 12 unique values starting from 2012.667 to 2013.583 which denote month and year ranging from August, 2012 to July, 2013.

Univariate Analysis

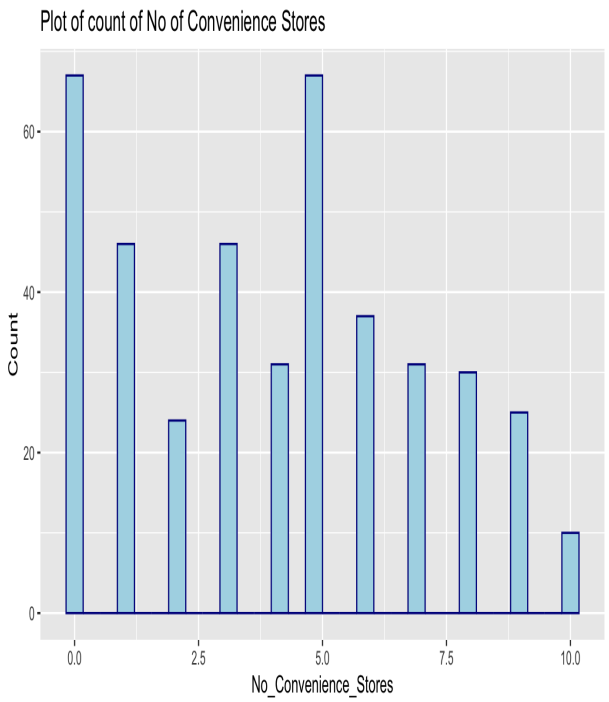
1) Transaction Date

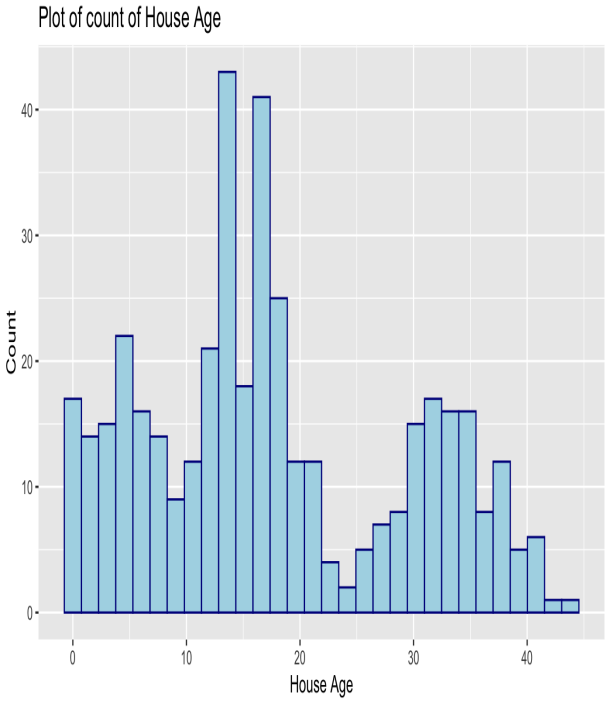


The Transaction Date has a heavy tailed distribution with the highest being count being that of 2013.417 and the lowest count being that of 2013.167. Hence we can apply log transformation on it and check the results.

2) House Age

The House Age refers to the age of the house in a year. The distribution is a bimodal distribution where there are 2 peaks. This can be normalised using the absolute value function. The largest number of the houses are aged at 14 years and the lowest number of the houses are aged at 24 years.

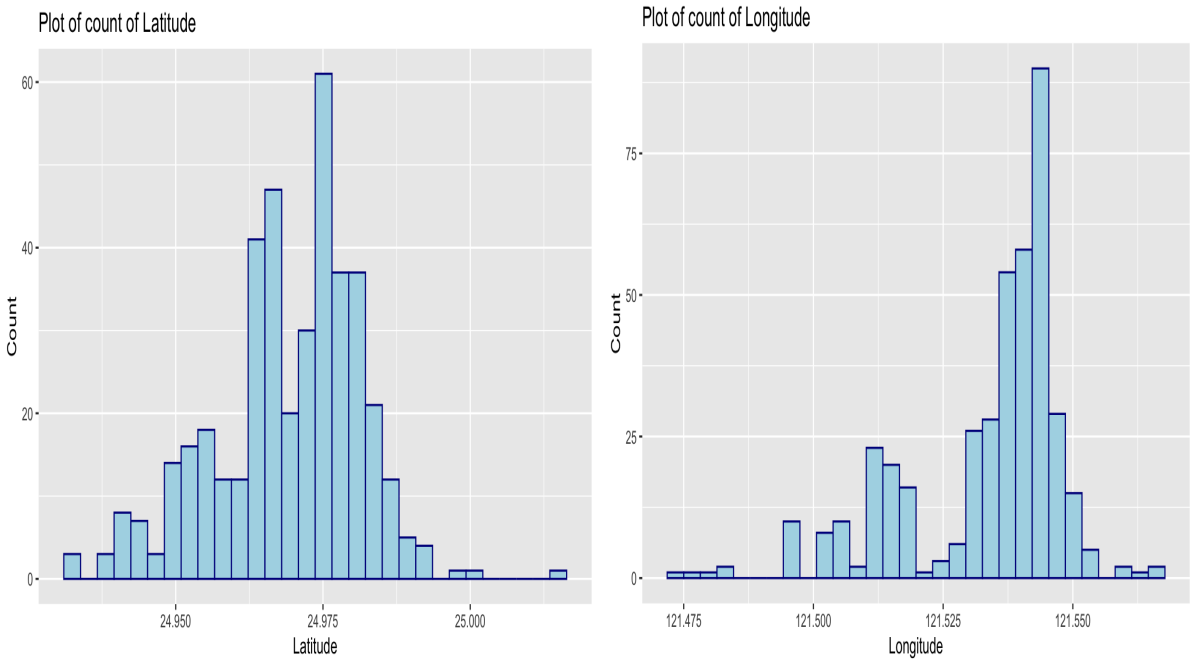




3) No. of Convenience Stores

The No_Convenience_Stores refers to the number of convenience stores in the living circle on foot. It is of the integer data type. There are 11 unique values to it in the dataset ranging from 0 to 10. The maximum number of houses have either 0 or 5 convenience stores while the lowest number of houses have 10 convenience stores.

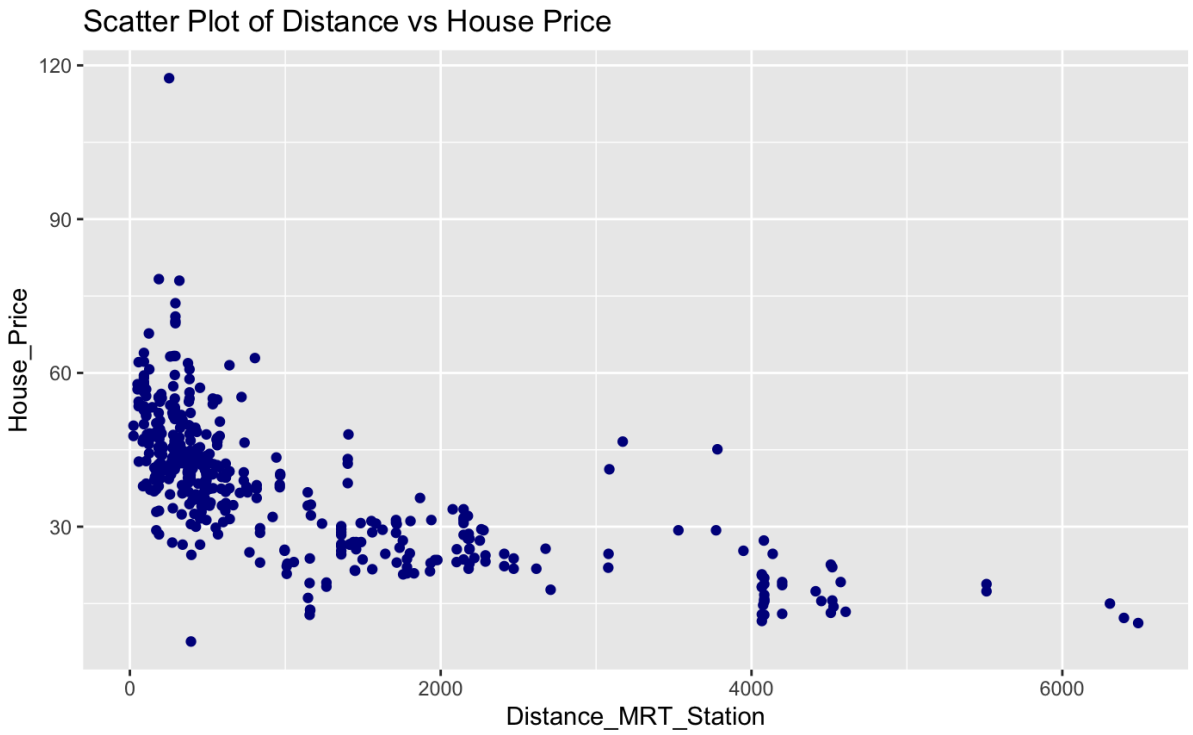
4) Longitude & Latitude



The latitude is one of the 2 geographical coordinates. It is a continuous variable with the lowest value being 24.93207 and the largest value being 25.01459. We do not need to transform this as it has an almost normal distribution. The longitude is the other geographical coordinate which is also a continuous variable. Its value ranges from 121.4735 to 121.5663. It has a slightly right skewed distribution but since this is a cyclic variable we do not need to transform it.

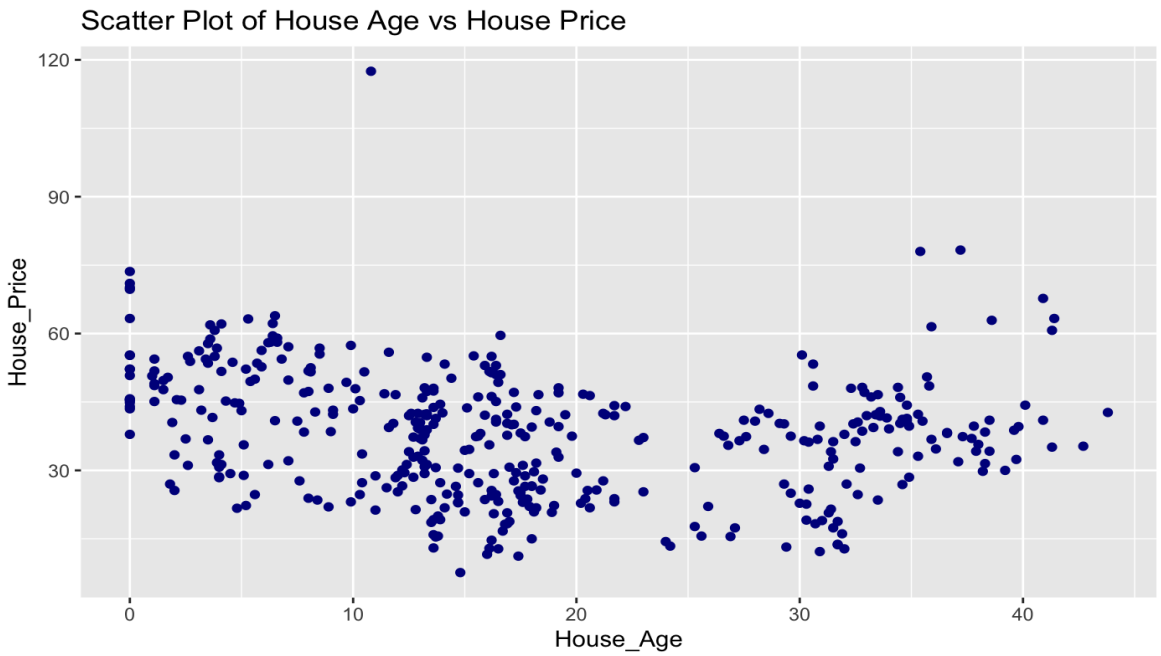
Bivariate Analysis

1) House Price Vs Distance_MRT_Station



In the above scatter plot Distance_MRT_Station and House Price have been plotted. It can be interpreted that this distribution is similar to a power-law distribution. On applying a 'log' transformation on the 'House price', which observed that the relationship between $log(House\ Price)$ and Distance to the MRT Station becomes linear.

2) House Price Vs House Age

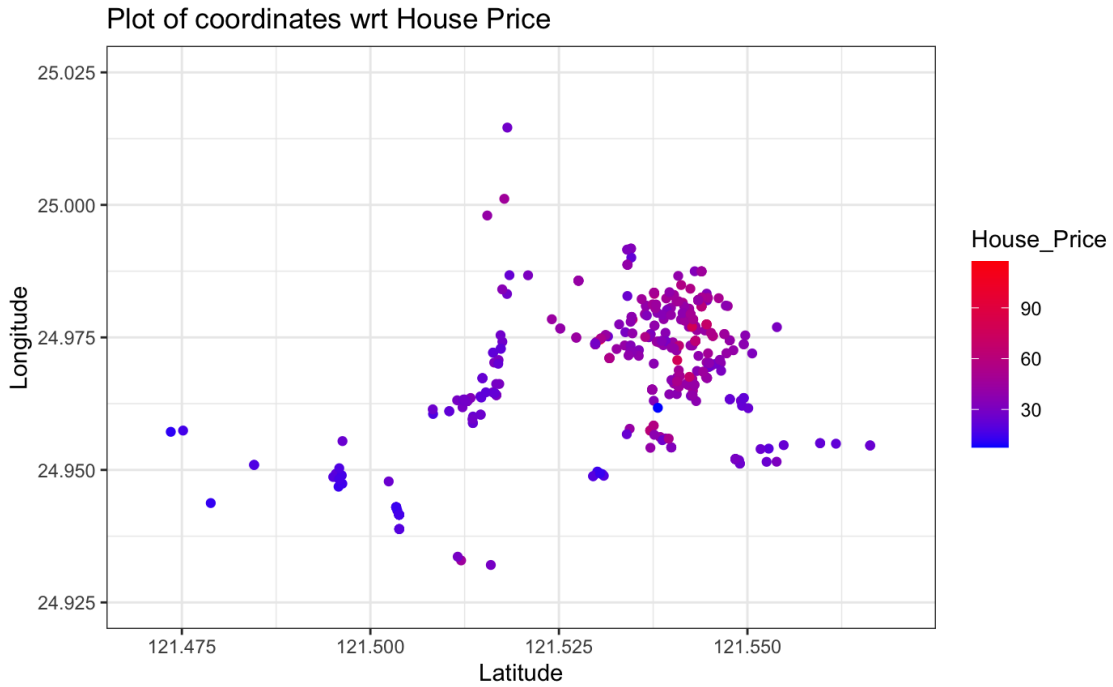


Through the above plot, of House Age versus House Price we can observe as the age increases the house price reduces but then increases after a certain point. This may be due to the fact that there are certain vintage houses that have a high price even when their house age is above average.

3) House Price Vs. No. of Convenience Stores

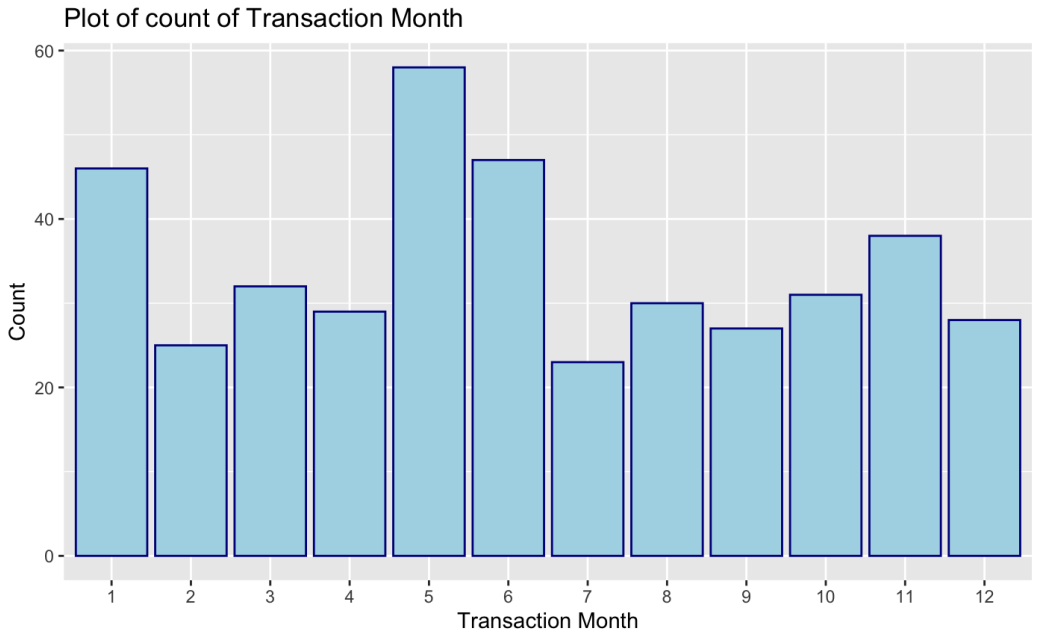
The houses with higher number of convenience stores have less variation in terms of house price as compared to the houses with low number of convenience stores. The highest variation in house price is shown in houses with 0 convenience stores.

4) House Price Vs Longitudes & Latitudes



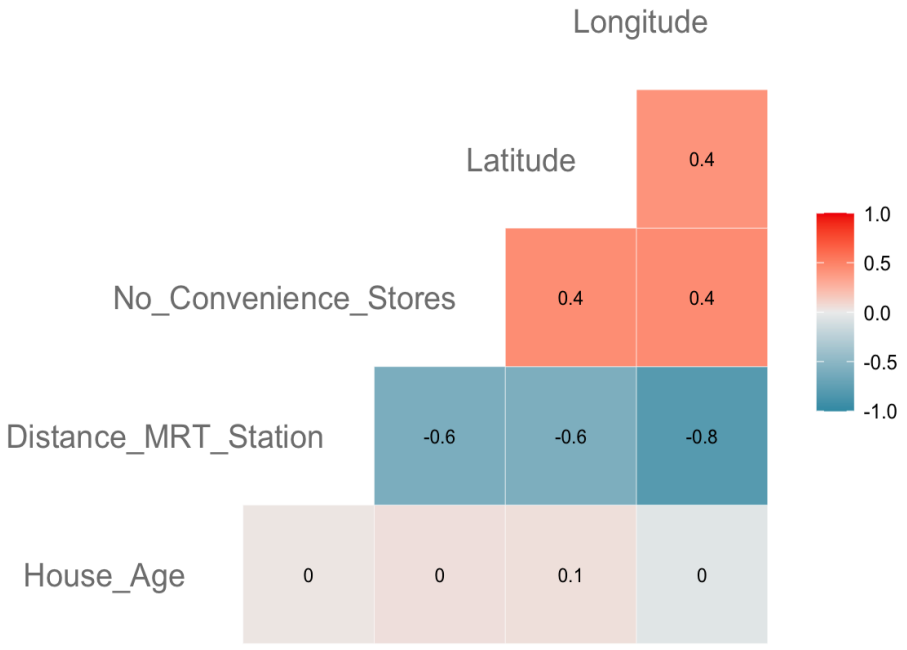
The plot of coordinates with respect to House price shows that Houses are purchased in a cluster and they are way more expensive than the houses in the outskirts of the city.

5) Derived Feature (Transaction Month)



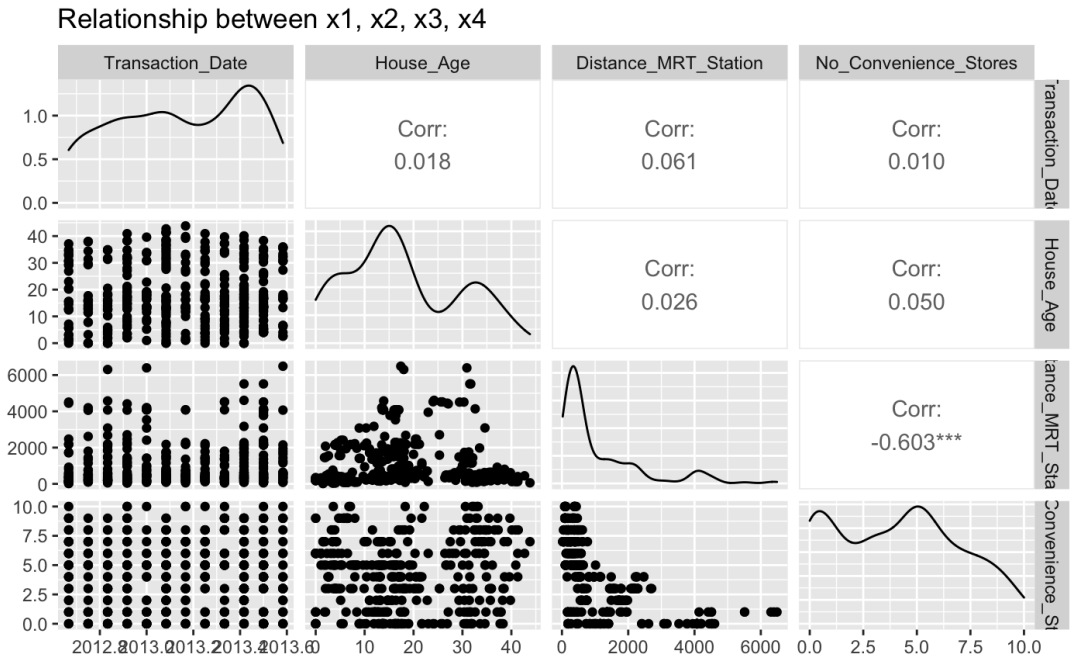
The distribution of Transaction Month is similar to that of Transaction Date wherein the 5th month of the year has the most transactions and the 7th month of the year has the lowest number of transactions.

Correlation Matrix



According to the correlation plot, Distance_MRT_Station and Longitude are highly correlated which we later tested using the Durbin-Watson test. Distance_MRT_Station is also moderately correlated with No of Convenience Stores and Latitude. Also No of Convenience Store is moderately correlated to Latitude and Longitude. Latitude and Longitude are moderately correlated to each other.

Pair Plot



The pair plot shows relationships between Transaction_Date, House_Age, Distance_MRT_Station and No_Convenience_Store. Transaction_Date and Convenience_Store are heavy tailed whereas House_Age is bimodal and Distance_MRT_Station is right skewed. Distance_MRT_Station is moderately correlated with No_Convenience_Store.
For the modeling We have taken Transaction month and Transaction

3. MODELLING: METHODS

1) Multiple Linear Regression

We began our modeling with a Multiple Regression model. Through the model, We were able to get an R-Square of 0.5611 which was moderate**.** Through the model's summary we were able to interpret the coefficients of different features. These Coefficients were in line with our initial hypothesis and the observations from the Exploratory Data Analysis. We then began analysing the model assumptions. We also observed that for the transaction month feature "NA" was being displayed for every field in the summary. This meant that we were having rank deficiency and this could be due to the fact that Transaction Date and Transaction month were simultaneously present in the model matrix.

Following Tests were done to check for assumptions:

Test for Constant Variance - Breusch-Pagan Test

The formal test to check if the errors have constant variance i.e. Homoscedasticity is Breush-Pagan Test. The null and the alternate hypothesis for this test are mentioned below:

H_0: *The residuals have a constant variance(Homoscedasticity)*

H_a: *The residuals do not have a constant variance(Heteroscedasticity)*

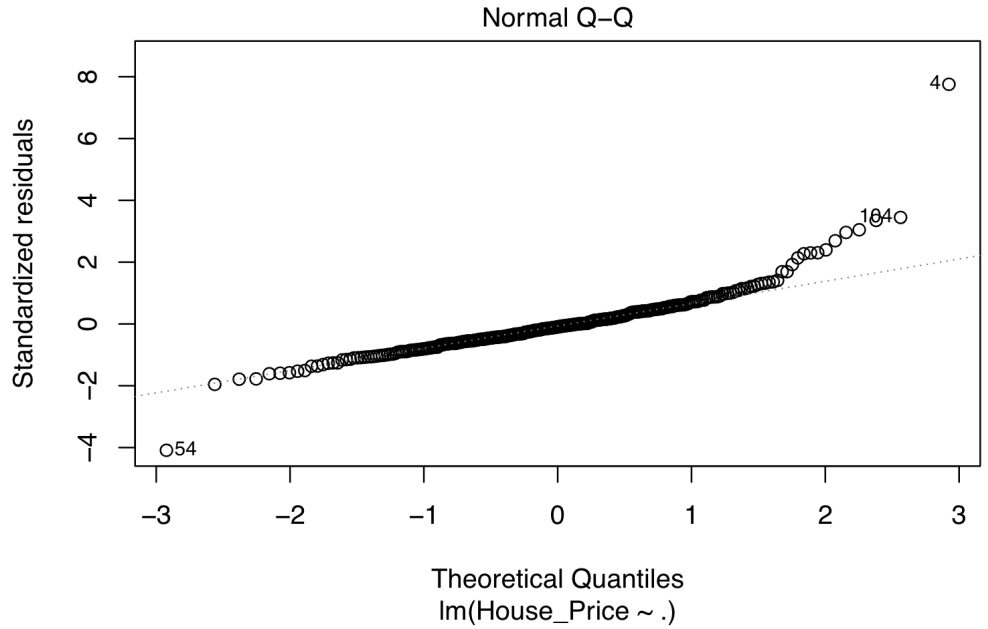
We carried out Breush-Pagan test and We observed Heteroscedasticity since the p-value was much lower than 0.05. Assumption Failed !

Test for Normality of Errors - Shapiro-Wilks Test

To check if the residuals have a Normal Distribution, we carried out the Shapiro-Wilks Test. The null and the alternate Hypothesis for this test are:

H_0: *The residuals have a normal distribution.*

H_a: *The residuals do not have a normal distribution.*



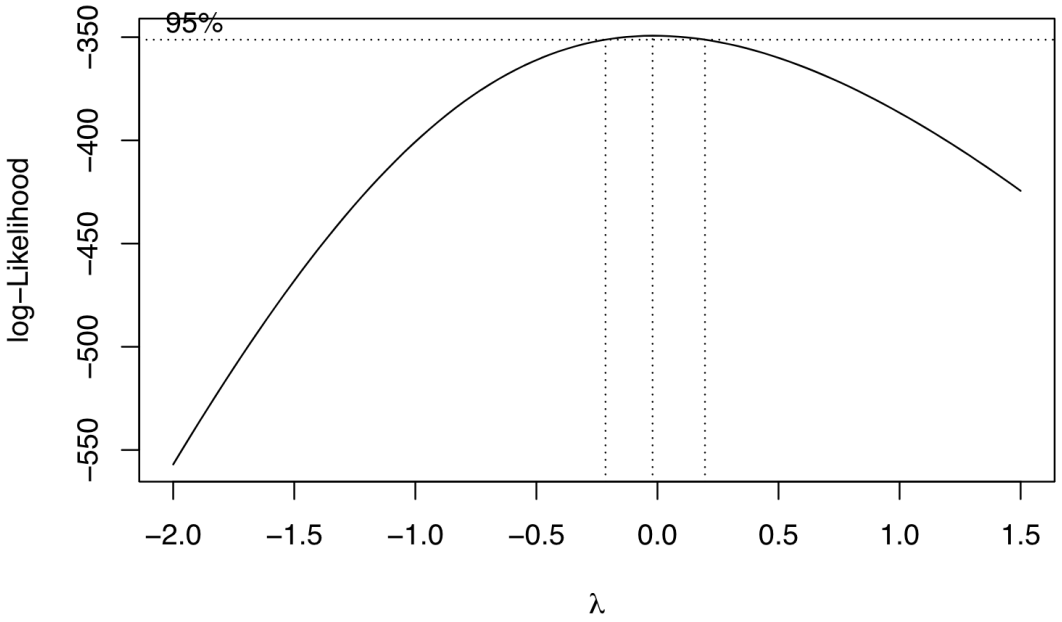
The p-value which we got after performing this test was lower than 0.05. Hence, we concluded that the residuals are not normally distributed. Assumption Failed !

Test for Autocorrelation of Errors

We used Durbin-Watson to test if the errors were autocorrelated. On doing Durbin - Watson test on the multiple linear regression model, the p-value was equal to 0.84. Thus confirming that there was no correlation between the errors. Assumption Passed !

After Performing the model diagnostics, We realised that We needed to do certain transformations on our response variables. We also needed to verify if there were any outliers and observations which were influencing our fitted model. For the next phase of the modelling, We Transformed Distance_MRT_station to log scale. This was done on the basis of observation from EDA.

To address the issue of Non-Constant Variance and Non-normality, We performed Box-Cox transformation on the response variable. The intent of these transformations is to maximize the likelihood of the data under normal assumption. As per the results from the box-cox plot, We decided to log transform our response variable.



Test for Influential Points and Outliers

We carried out standard tests to detect if there were outliers in our dataset. We were able to observe two outliers. However, To confirm if these were influential points as well We checked Cook's distance. The criteria we used to identify the influential points is:

$$Cook's\ Distance(CD) \geq 1$$

None of the observations in our data were influential points. However, since our model was failing assumptions we removed Observation 4 and Observation 54 from the training dataset

1.1) Multiple Linear Regression (Part 2)

With the inputs from Model diagnostics, model summary and observations from the Exploratory Data Analysis, We decided to fit a log-linear multiple regression model. The obtained Model was able to explain 77 percent of the variance in House Price (R-Squared- 0.77). We obtained a testing error of 6.75 as compared to 6.99 from the previous model. We again analysed model diagnostics to ensure whether regression assumptions were being passed.

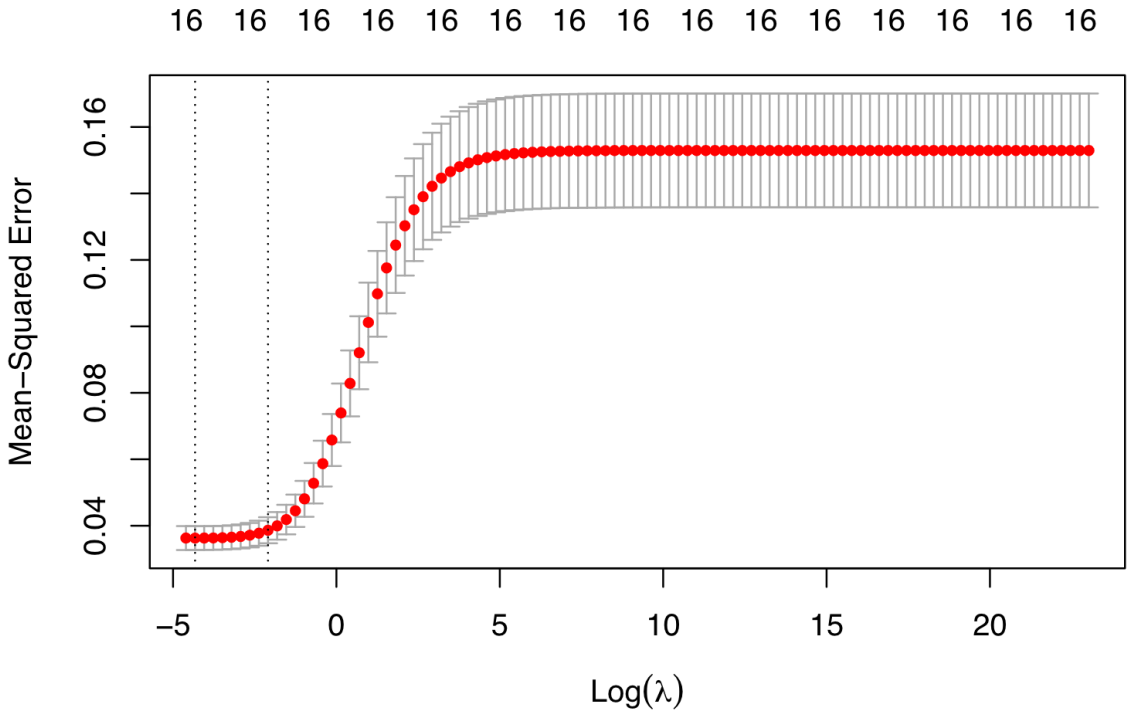
Test for Constant Variance - Breusch-Pagan Test Assumption Passed !
Test for Normality of Errors - Shapiro-Wilks Test Assumption Failed !
Test for Autocorrelation of Errors - Durbin -Watson Test Assumption Passed !

We again checked if there were any outliers in our model. Based on it we removed Observation No. 108 from the training Data
Even after these transformations, We observed that our model was not able to fulfil the assumption of normality since the p-value for the Shapiro-Wilks test was less than the significance level (0.05).

Model Type	R-Squared	Testing Error	Normality Assumption	Autocorrelation	Constant Variance
Multiple Linear Regression 1	0.5353	6.999604	Failed	No Autocorrelation	Failed
Multiple Linear Regression 2	0.7572	6.752118	Failed	No Autocorrelation	Passed

2) Ridge Regression

After applying the Multiple Regression Model, We turned towards Ridge regression which is particularly very useful when a unique solution for the data does not exist or when there is correlation among the variables. Ridge regression applies a penalty to the cost functions and limits their magnitude. We tuned our model by running our model on a grid of ƛ values. We chose the lambda value at which the Root Mean Square Error for the testing data was the least. We used the 'glmnet' package from R to train our dataset.

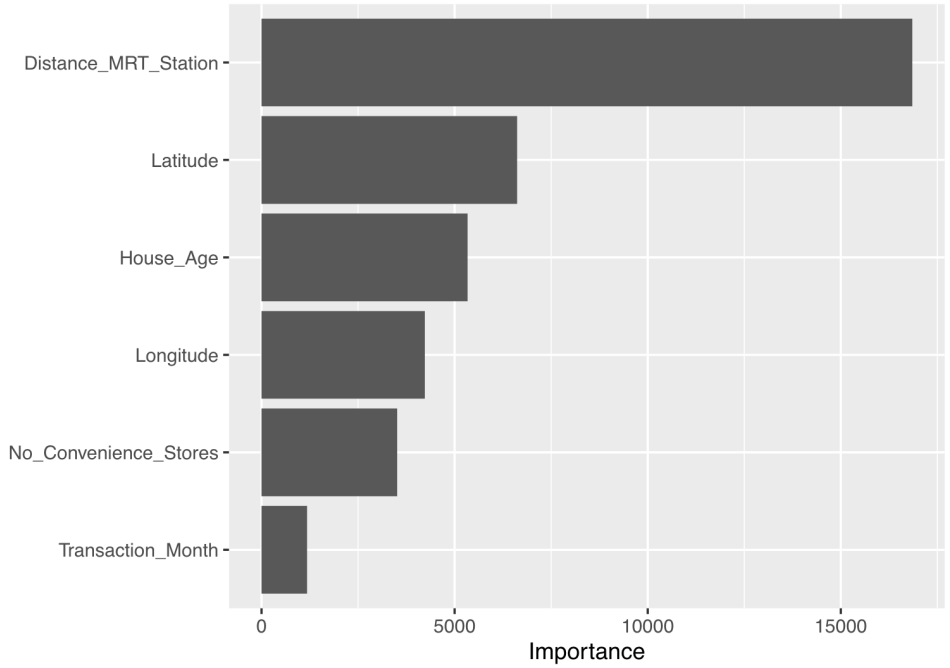


The **testing error that we achieved is 6.58.** The value at lambda equal to 0.017 minimum RMSE was achieved.

3.4 Random Forest

After applying Ridge regression on our data, we used the 'Random Forest Regressor' algorithm to predict the housing prices. Since there are a lot of parameters to tune in the random forest, We used cross-validation to get the best parameters. We obtained a **testing error of 5.55** to our using the most optimised model.

We obtained the feature importance from the model. As per the variable importance, Distance_MRT_stations was the variable which resulted in maximum reduction in Impurity, Followed by Latitude and House_Age, Longitude , No. of Convenience Stores and Transaction Month.



From the results which we got after applying the random forest algorithm, 'Distance to the MRT station' was the most important variable. Among others, 'Latitude' and 'House Age' are also very important.

4. CONCLUSION

Model Name	Testing Error
MLR 1	6.99
MLR 2	6.75
Ridge Regression	6.58
Random Forest Regressor	5.55

Final Table for Model Comparison

Following were the observations and conclusions from the dataset:

- Through Extensive modelling on the dataset, We observed that from all the algorithms, Distance_MRT_Station was the most significant feature of the dataset.
- We also observed a positive association between No. of Convenience Stores and house price.
- For Modelling we removed Transaction Date from the analysis and replaced that with Transaction month.
- We also observed that there was significant increase in R-Square When we used a log-linear model in regression than a linear - linear model in regression
- Though Random Forest gave the lowest error, It was the most difficult to tune due to the large number of parameters. Also it presented a more interpretable model.