RENT DATA ANALYSIS FOR FUTURE PRICE PREDICTIONS

Singapore Real-estate House Prices

The **intended audience** for this analytical research would be the decent apartment owners, Singapore citizens who are interested in buying a house or an apartment, realtors, and other related parties who would be looking to know the quoting on a place.

The analysis is to predict the future cost of the place in Singapore based on the present prices and various other factors.

Data and Background:

The data was gathered and collected by keeping a track of transactions done by the private residential properties in between the years 1995 and 2015 from the Urban Redevelopment Authority Real Estate Information System (REALIS). It was mainly concentrated on the Singapore's residential housing market. Some of the features that were recorded in the data collection are the type of the project (whether it is a loft, suite, finished, unfinished, and many others), longitudinal and latitudinal location of the project, completion year of the project, contract year, size, number of floors, area in sqm, building height, how far it is to a mall, if it has facilities like gym, pool, tennis court and others.

It is said that the Singapore's residential market is classified into three types: Apartments (high buildings) by Housing and Development Board (government), Apartments (high-rise buildings) by private developers, and semi implemented buildings by private developers.

Target Variable:

Since, we are trying to predict the future price of the plot we take the price per transaction as the target variable (pricetransaction_nominal). It is continuous variable. It is the cost of the residential plot in US Dollars ($), nominal.

Predictor Variable:

Using all the predictors in training a model can cost it to be biased and it would be a result of underspecified regression model. So, we can get the most effective predictors in a few ways like by plotting one predictor against all others and comparing, by plotting a correlation plot and consider the highly corelated ones. We concentrate on using only the size per square meter to see how it affects the price of the house.

Lead Statement:

Most of the landlords/vendors when entered the real estate business are finding it difficult to plan on purchasing or leasing a flat because of the insufficient, non-uniform data that is available which leads to the audience ending up making bad decisions.

Methodology:

Since we are trying to predict a continuous variable so, regression method should be used. Regression is a type of Supervised Machine Learning techniques, and it can be used to predict a numeric value. There are various types of Regression like one which predicts the Binary class, one which predicts the numerical value and others.

In this case, we use the linear regression, it gives a linear relation between the response variable and one or more predictor variables. LR is one of most common forms of the regression.

In R it can be achieved using a command 'lm'. Using the syntax:

lm([response] ~ [predictors], data = [dataframe])

with the 'summary ()' we can get the detailed report on the performance of the regression model.

First, we have imported all the necessary libraries

```
{r}
# import libraries
library(corrplot)
library(ggplot2)
library(dplyr)
library(broom)
library(ggpubr)
```

```
corrplot 0.92 loaded

Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

Then, check the working directory and loaded the dataset into a variable 'RENTDATA'. And then considered few important features ignoring the unnecessary ones to create a data frame named 'df'.

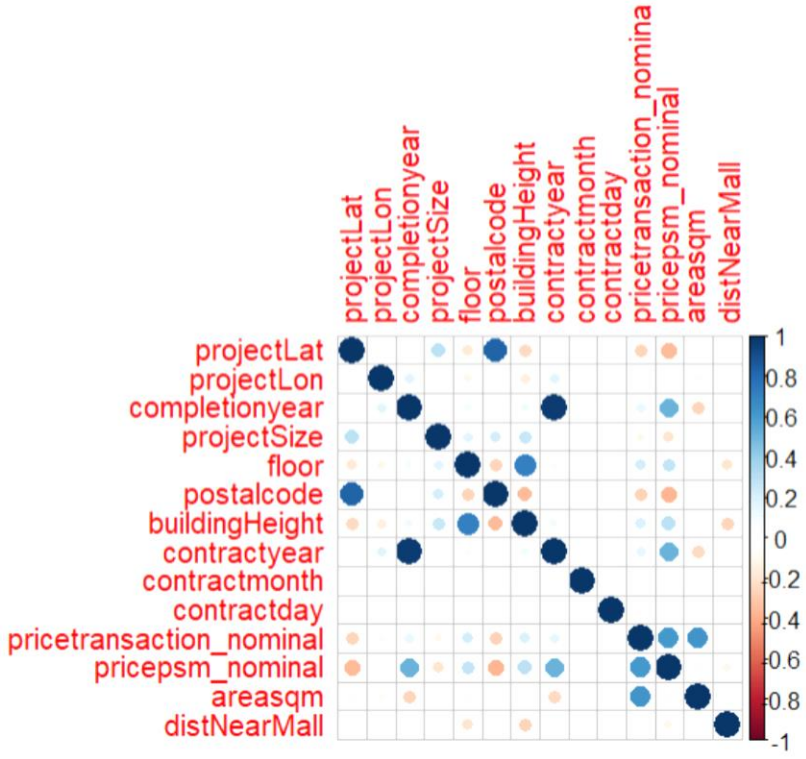
```r
{r}
# see the working directory
getwd()
# extract appropriate columns from the dataset and map into dataframe
RENTDATA <- read.csv("data_transactions.csv",strings = T)
df <- subset(RENTDATA, select=c("projectLat", "projectLon", "completionyear",
"projectSize", "floor", "postalcode", "buildingHeight", "contractyear",
"contractmonth", "contractday", "pricetransaction_nominal", "pricepsm_nominal",
"areasqm", "distNearMall"))
```

```
[1] "C:/Users/nihar/OneDrive/Desktop/unt/ADTA 5230/ProjectShowcase"
```

To check the structure of the object in detail we use the 'str()'.

To see which features are highly correlated we use the correlation matrix and plot it.

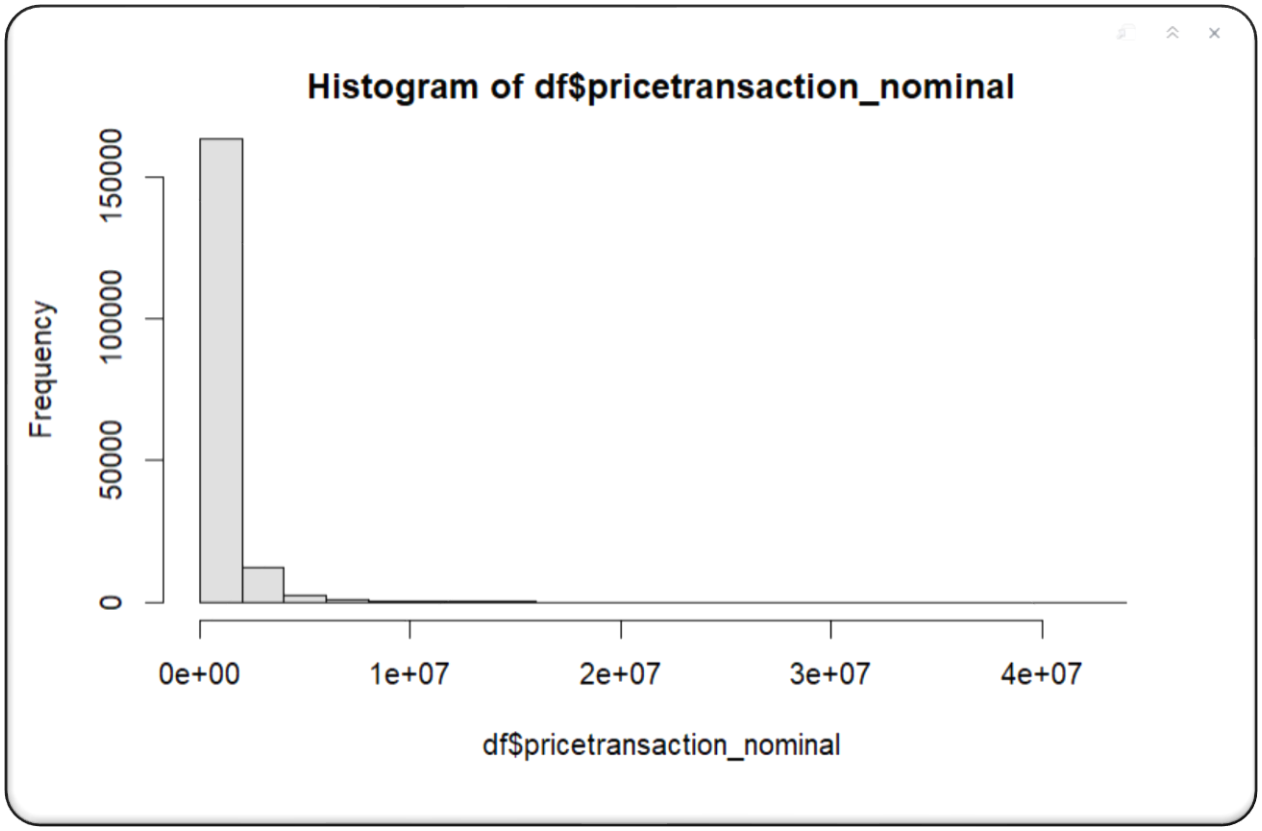
```r
{r}
corrplot(df.cor)
```



From the correlation matrix it is evident that the price nominal, project Latitude and price nominal, postal code are highly positively correlated to each other. And the postal code, project Latitude and contract year, completion year are highly negatively correlated. In all the pairs, either one variable can be considered for the model instead of both.

For, our regression model we have first considered only the area per square meter and the price of transaction, nominal. And it turned out to be approximately 67%.

Since, the price per transaction nominal is the dependent variable we also checked if it follows the normal distribution or not.

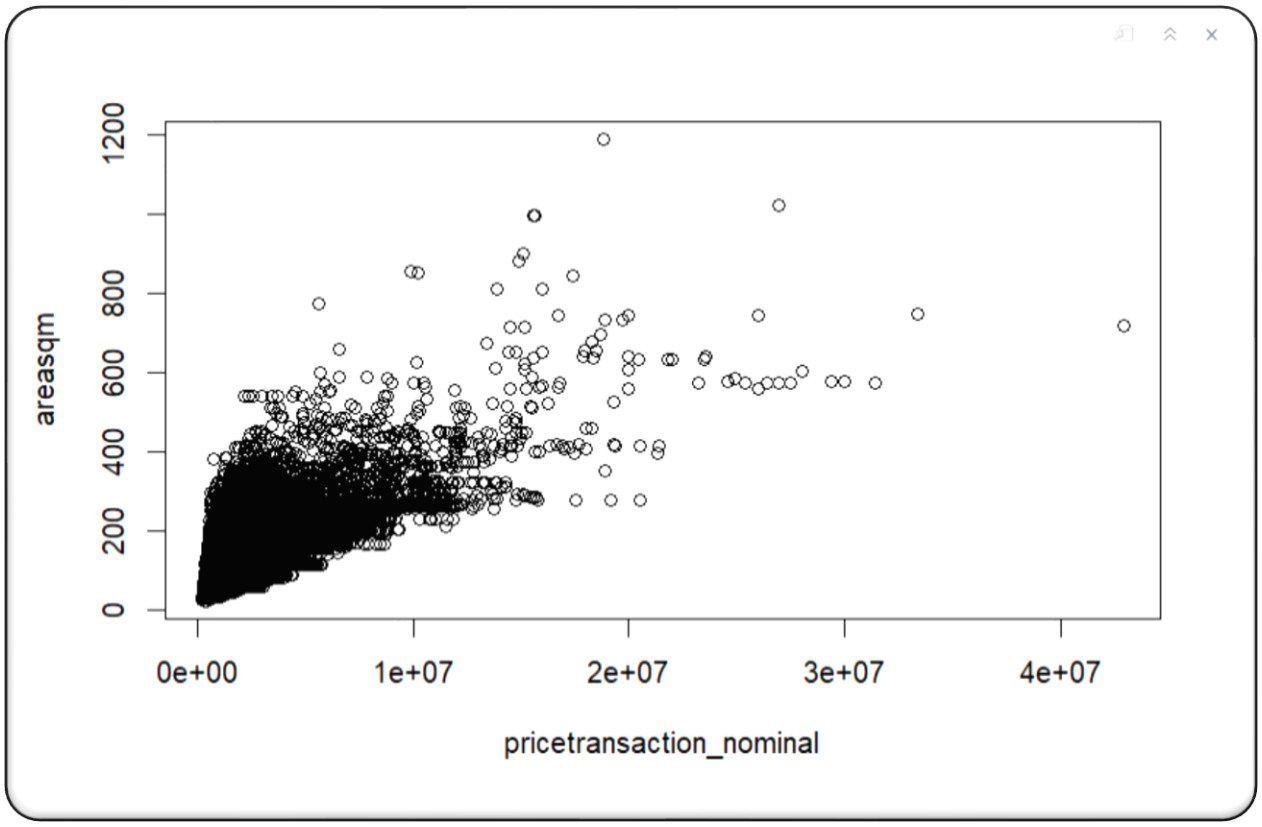


To the check the relation between the response and the target variable we use the linearity plot. If the relation is linear or not? Usually there are few assumptions about the data they are:

- Linearity of the data.
- Normality of residuals.
- Homogeneity of residuals variance. (homoscedasticity)
- Independence of residuals error terms.

When we check if these assumptions are true or not some of the common problems occurred are:

- Non-linearity
- Heteroscedasticity
- Presence of influential values

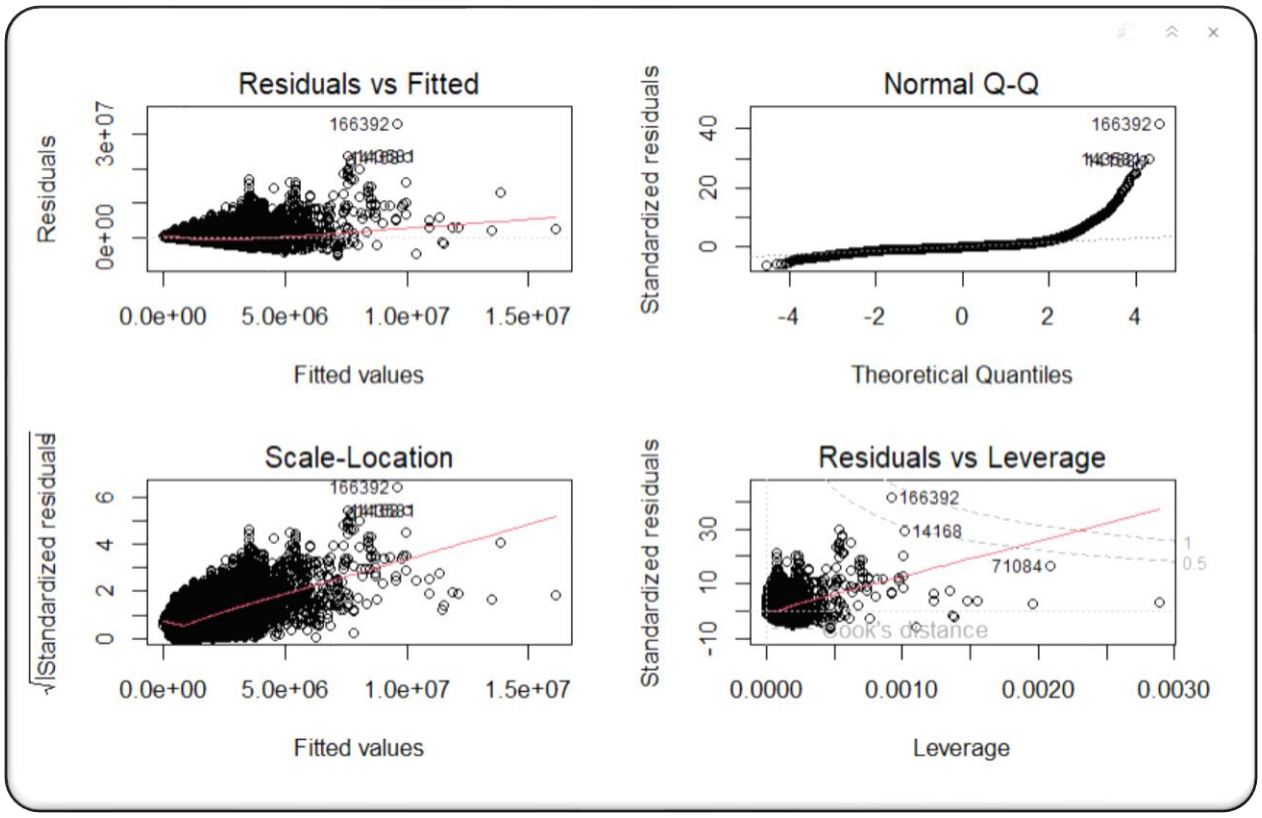


It looks like the area per square meter and the price are highly correlated to each other.

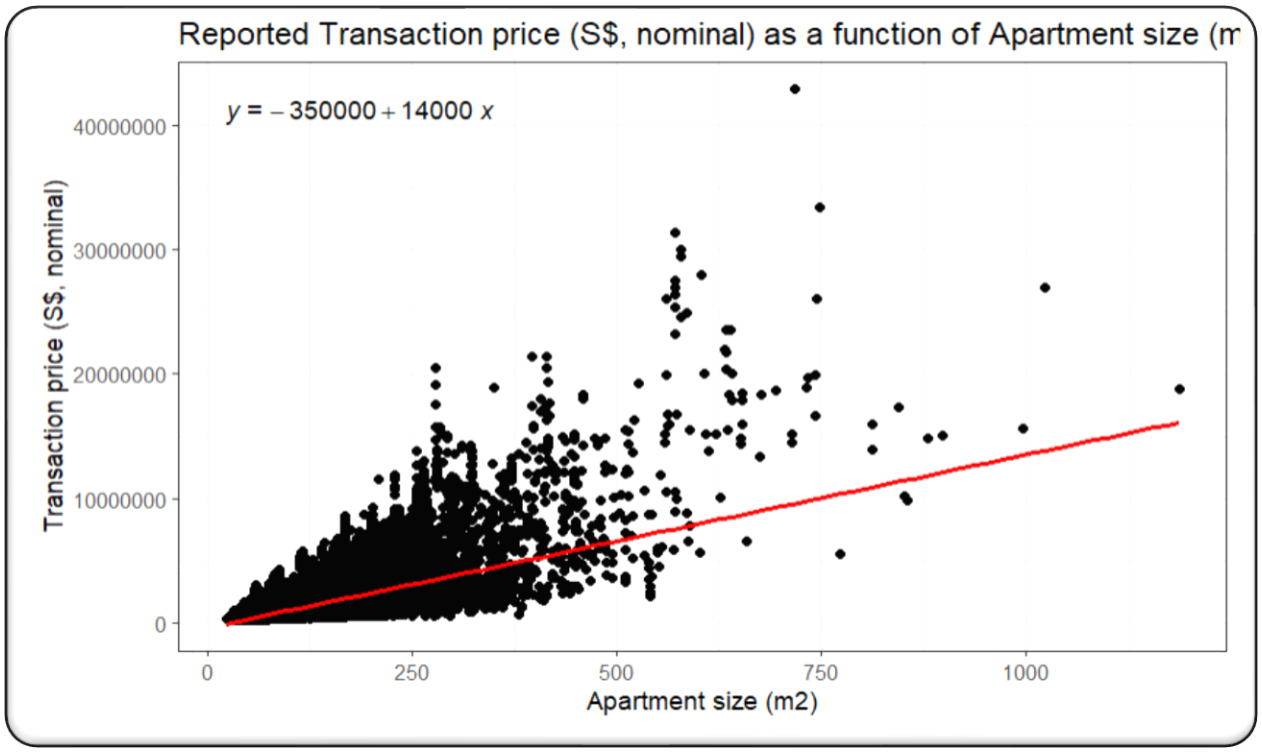
From the regression model with only the independent variable as 'areasqm' the model resulted with the residuals of minimum of a negative 4967036, the 1st Quartile with a negative of 450813 and a third quartile of negative 94711. The most important numbers are the coefficients, and they are for the intercept it is -352735.6 and the areasqm it is 13890.5. Making the regression equation as:

Price = -352735.6 + 13890.5 (areasqm)

The Adjusted R-Square value is 0.4058 and the F1-statistic is 1.224e+05.
To check for the Homogeneity of residuals variance (homoscedasticity)

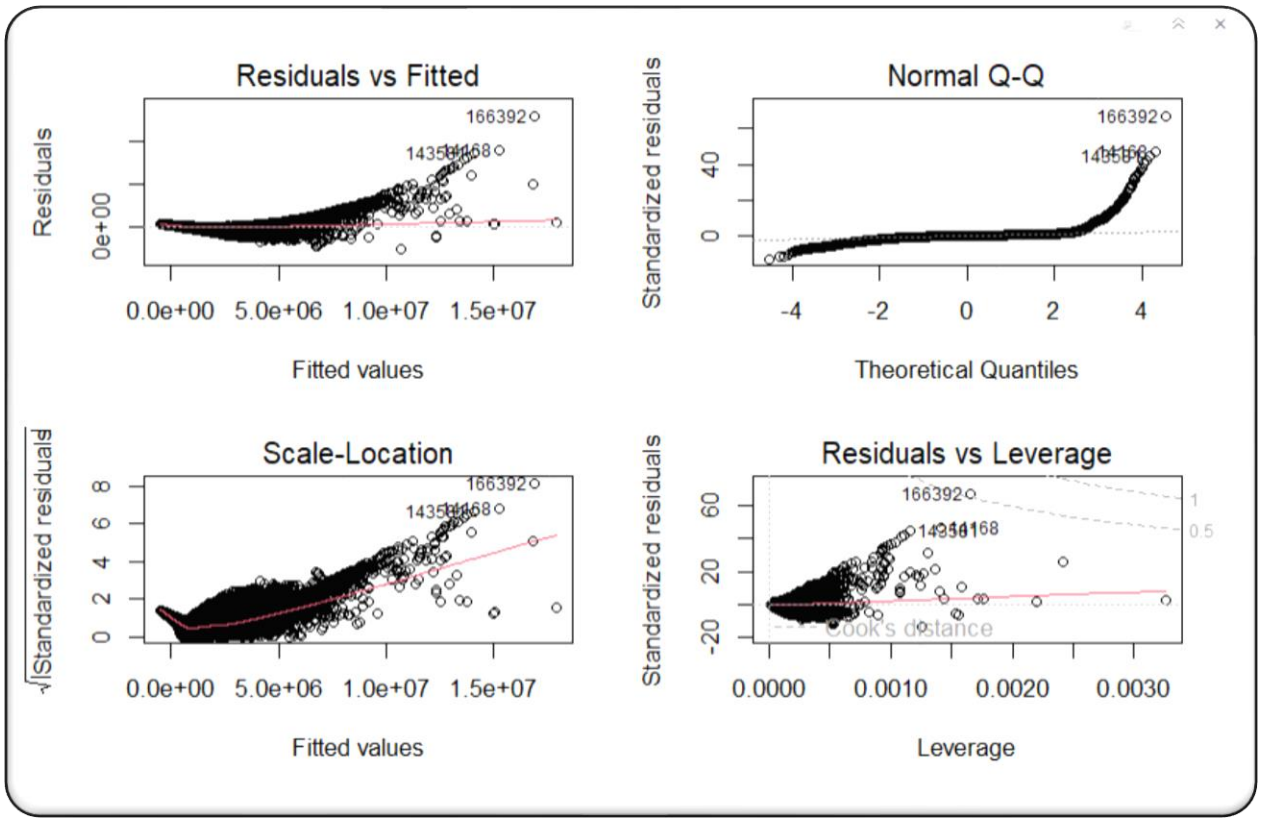


To visualize the model by a linear equation along with the points we need to use the ggplot and add the regression line to the plot and the result looks like:



We have used another model for linear regression but this time with all the independent variables

The intercept looks like -1.228e+0 and the residuals minimum is -5036641 and maximum is 26041339. The adjusted R-square value is 0.8593 and the F1-statistic is 8.422e+04.



Out of the two models the one model with all the independent variable has more R-squared value compared to the other model. The second model's regression equation would look like:

Price = -1.228e+07 + 1.873e+06(projectLat) + projectLon(2.554e+05) + completionyear(-3.007e+03) + projectSize(5.089e+01) + floor(-3.037e+03) + postalcode(-1.119e-01) + buildingHeight(-1.355e+02) + contractyear(-6.280e+03) + contractmonth(-4.658e+02) + contractday(-3.664e+02) + pricepsm_nominal(1.386e+02) + areasqm(1.498e+04) + distNearMall(-7.483e+03)

Results:

The prices of residential lands/houses in the Singapore region were predicted based on the features in this analysis. As we have mentioned in the methodology that we are using the two models as comparison and the numbers say that the model with only one predictor that is the area in square meters is not recommended over the other model. The model with only one predictor has a r squared value of 40 percent and the model with more predictors have an r squared value of 85 percent.

Conclusion:

We can say that the best model out of the two is the second one and in future we can also try modelling with changing and playing with the list of predictors so that it can improve the model even better with good performance.

So, by using this analysis we can predict the house rent in coming days. This would benefit the people who are looking to buy or rent a place without misplacing their money in a wrong place or by ending up paying high rents they can ease out this problem.

Some of the recommendations would be:

It would be better if the information is updated every now and then because it would provide accurate results. The model can be improved by training and testing including cross-validation layer while processing.