Statistical Analysis of the House Price in New York City

Deyoz Rayamajhi, Yaw Mawuli, Rahat Mahmud

2022-05-20

# Introduction

New York City (NYC), which has five boroughs - New York, Kings, Queens, Bronx, and Staten Island - is considered one of the most expensive cities in the world, particularly for housing. However, the city is so diverse that the house prices vary significantly across its neighborhoods. It would be, therefore, very interesting to know the factors that influence the house prices, and also to determine the degree of variations in the house price across the regions within the city. For this purpose, we studied and performed the statistical analysis and developed the multivariate regression model to understand multiple explanatory variables that can predict the house price in NYC reasonably. Moreover, we also performed regression analysis to examine the relationship between average house price per zip code and total population per zip code in NYC. Furthermore, we performed the exploratory and descriptive analysis to determine the average house price by county, zip code, and also house type. We have also developed machine learning algorithms to predict the crime level.

From the multivariate regression, we found that the number of bathrooms, bedrooms, and level of the nightlife activities are the most important determinants of the house price. The number of schools in the area and square feet is other explanatory variables. Surprisingly, we found that there is no correlation between average house prices and population per zip code. As expected the most expensive zip codes for the house price are in New York County, but we have found that house price in kings is also high, particularly in Western and Northern Brooklyn. Interestingly, we have found that the house price in Staten Island is least among all the boroughs, followed by Bronx and Queens.

The first section of this document discusses the data set we used to perform the analysis. Second, the cleaning and preparation. Third, some exploratory and descriptive analysis to examine the distribution, outliers, and relationship of the variables. The fourth section is the multivariate Regression Model. In the fifth section, we have developed the machine learning model to predict the crime level. The sixth section is uni-variate regression analysis. Seventh is descriptive statistics which presents the aggregated data by the different categorical variables. The last section presents the average price per zipcode visualized in the NYC map. In the last section, we conclude.

This document is prepared as a final project for the partial completion of the graduate course STA/OPR 9750 at Baruch College, New York. This document is based on the general instructions set in the project guidelines posted on the blackboard by Prof. Jean Francois Collard (Jeff). We also declare that all the work done for the project is original work.

# Section 1: The Datasets

For the analysis, we used the housing data that was available on Kaggle, an open-source platform for data analysis. The link to the data set is embedded [here](https://www.kaggle.com/datasets/parv619/newyork-house-sale-data). The second data that we have used is the population of NYC by zipcode which we have extracted from [this](https://data.cityofnewyork.us/Business/Zip-Code-Boundaries/i8iw-xf4u) link. We have used this dataset also for our map visualization.

housing\_data = read.csv('D://MSBA//Second Sem//OPR 9750//Project//Housesell//data\_lat\_long2.csv//project\_opr.csv')

This uploaded dataset contains information on New York City neighborhood housing. The raw data contains some variables we do not need for our analysis, so we will begin by viewing and reading the structure of the data, then we clean up the dating after reading it to keep only variables that are needed for our work. Also, variables that are characters may be converted into numeric and vice versa when necessary for our analysis.

# Section 2: Data Cleaning and Preparation

In this section, we read, examine and clean to make the data set useful for our analysis.

***Import required basic libraries***

library(tidyverse)  
library(dplyr)  
library(stringi)  
library(miscset)

***Examining the structure of the dataset***

It is observed that the data is very raw and needs massive cleaning and in some instances needs extractions, We, therefore, begin with the cleaning for our data preparation.

***Removing Unwanted characters from the bath, bed, and sqft variables and changing their type to numeric***

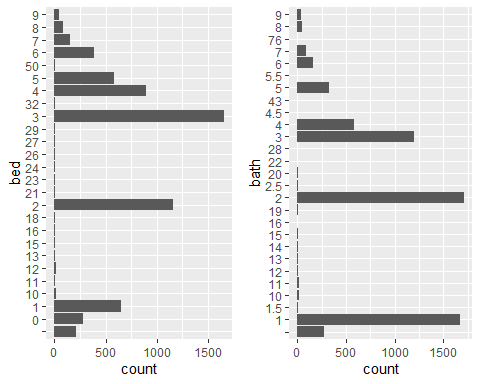
As it is observed from the structure of the data frame that bath, bed, and sqft variables have type characters instead of numeric, and is because of the additional characters in the values. So, we first remove those unwanted characters.

housing\_data$bath = gsub(" .\*",'',housing\_data$bath)  
  
housing\_data$bed = gsub(" .\*",'',  
 gsub("Studio", "0",housing\_data$bed))  
  
housing\_data$sqft = gsub(" .\*",'',housing\_data$sqft)

Remove the $ sign and comma separator in the values of the variables tax, land.assessment.cost, improvement.cost, total.cost and sqft.

housing\_data$tax = gsub("\\$",'',  
 gsub(',','',housing\_data$tax))  
  
housing\_data$land.assessment.cost = gsub("\\$",'',  
 gsub(',','',housing\_data$land.assessment.cost))  
  
housing\_data$improvement.cost = gsub("\\$",'',  
 gsub(',','',housing\_data$improvement.cost))  
  
housing\_data$total.cost = gsub("\\$",'',  
 gsub(',','',housing\_data$total.cost))  
  
housing\_data$sqft = gsub("\\$",'',  
 gsub(',','',housing\_data$sqft))

Counting Unique values of bath and bed variables



As it can be seen from the above bar diagrams that there are some missing values, so we assign NA to those missing values by converting those to numeric.

***Extracting ZIP Codes from address variable***

housing\_data['zipcode'] = gsub(".\*NY ","", housing\_data$address)

***Extracting Number of schools from school.information***  
  
housing\_data['schools'] = gsub("[^0-9]","",housing\_data$school.information)   
housing\_data$schools = as.numeric(housing\_data$schools)

***Removing unwanted characters from estimated.mortgage***

housing\_data$estimated.mortage = gsub(".\*\\$", '',   
 gsub("/.\*",'',   
 gsub(",", "", housing\_data$estimated.mortage)))  
housing\_data$estimated.mortage = as.numeric(housing\_data$estimated.mortage)

***Extracting Number of Restaurant, Groceries and Nightlife***

housing\_data['Restaurant'] = gsub(".\*Eat", '',   
 gsub('Restaurants.\*','', housing\_data$shop.and.eat.information))  
   
  
housing\_data['Groceries'] = gsub(".\*Restaurants", '',   
 gsub('Groceries.\*','', housing\_data$shop.and.eat.information))  
  
housing\_data['Nightlife'] = gsub(".\*Groceries", '',   
 gsub('Nightlife.\*','', housing\_data$shop.and.eat.information))  
  
  
housing\_data[, c('Restaurant', 'Groceries', 'Nightlife')]= sapply(housing\_data[, c('Restaurant', 'Groceries', 'Nightlife')],as.numeric)

***Extracting Crime information***

housing\_data['crime.level'] = gsub(".\*Crime", '',  
 gsub('crime.\*', '',   
 gsub('Learn about.\*', 'Unknown', housing\_data$crime.information)))

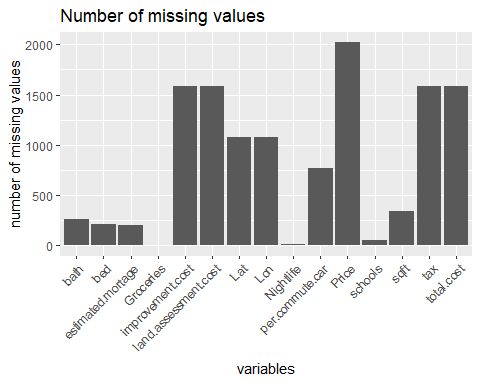
***Extracting Commute information***

housing\_data['per.commute.car'] = gsub(".\*Commute", '',  
 gsub('%.\*', '', housing\_data$commute.information))  
housing\_data$per.commute.car = as.numeric(housing\_data$per.commute.car)

***Extracting Housing Type***

housing\_data['housing.type'] = gsub('/sqft.\*', '',  
 gsub(',.\*', '',   
 gsub('Home.\*', '', housing\_data$home.details)))

## Bar plot of missing values



***Omitting records of missing values in price***

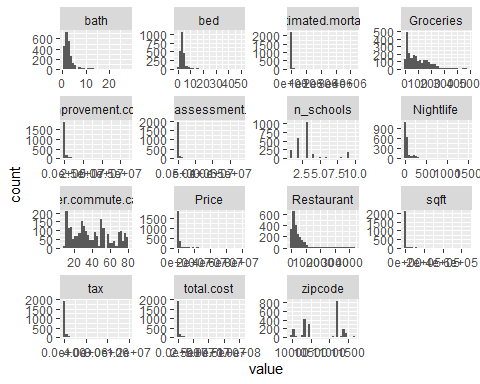
Because the price has the maximum number of missing values, and it is both difficult and unwarranted to estimate the price of the missing values, so we will omit all the records, which has missing values in Price.

Now, we do not have any missing values. So, let us explore the data using descriptive statistics.

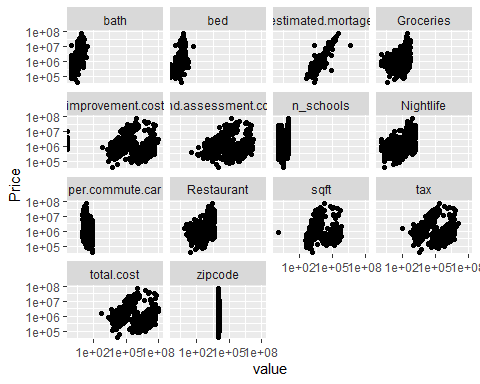
# Section 3: Data Explorations and Descriptive Statistics

In this section, we explore the data set to check some descriptive statistics. We examine the correlation between the variables using a scatter plot, and check outliers and summaries of data using a box plot . We further examine the scatter plots also using categorical variables.

It can be observed from the summary statistics of the data set that there are outliers. for example square feet, the minimum is 1 and is outlier. We will examine the outliers in more detail in subsequent subsections. Let’s examine the distribution of all numeric variables.



It is observed that the distributions of most of the numeric variables are highly skewed to the right and therefore required transformations. let’s examine the scatter plot of all the numeric variables against the price to see the correlations.



From the scatter plot, it can be observed that there is a positive correlation between house price and these numeric variables.

***Checking the Correlation between numeric variables and price***

| name | corr |
| --- | --- |
| Restaurant | 0.4541575 |
| Nightlife | 0.4408859 |
| bath | 0.4267392 |
| Groceries | 0.3709846 |
| n\_schools | 0.3485888 |
| zipcode | -0.2400169 |
| per.commute.car | -0.2373345 |
| bed | 0.2169726 |
| land.assessment.cost | 0.1155505 |
| tax | 0.1016339 |
| total.cost | 0.0985291 |
| improvement.cost | 0.0928168 |
| sqft | 0.0894373 |

We ignore the zip code in the correlation above as it is the categorical variable. Apart from the per.commute.car, all other variables are positively correlated with price.

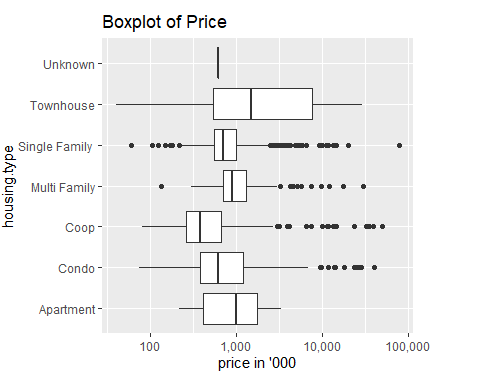
## Examining outliers and the distribution

***Box Plot of house price by several categorical variables***

boxplot\_price = function(x){  
 (df\_housing.nona %>%  
 mutate(Price = Price/1e3) %>%  
 ggplot(aes(Price, {{x}}))+ geom\_boxplot()+  
 scale\_x\_log10(labels = scales ::comma)+  
 labs(x = "price in '000", title = "Boxplot of Price ")+  
 theme(plot.margin = margin(10,50,10,0)))  
  
}

### Boxplot of Price by housing Type

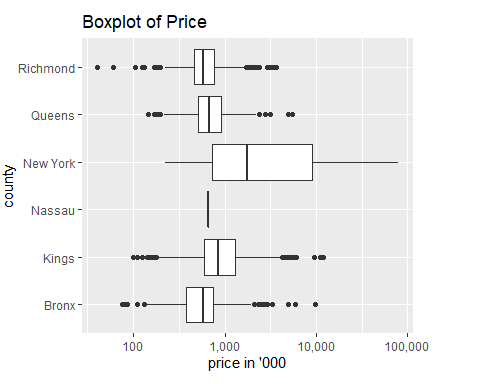
boxplot\_price(housing.type)



As it can be seen that there are points that lie below and above of whisker of each boxplot. However, these points may not be the outliers, because it is possible that single-family house prices can range from $40,000 to $79,000,000, given the diverse location within New York City and other factors such as size, type, etc. Therefore, we do not treat these points as outliers. Among the median price of all the house types, the median price of the townhouse is the highest. This is consistent with the fact that townhouses are the most expensive on average. There is another house type, which price is much more than the townhouse, this is because of the prime location in Manhattan and Southern Brooklyn.

### Boxplot of Price by county

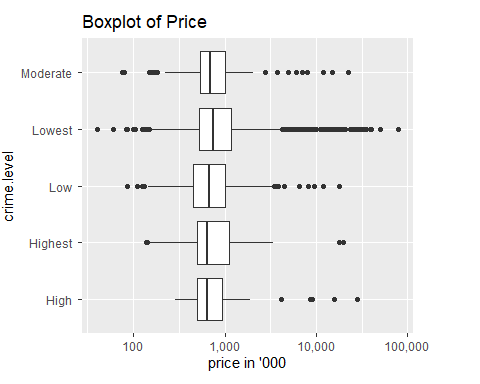
boxplot\_price(county)



From the above boxplot of price by county in NYC, the median price of the houses in New York county is way more than the median prices of the houses in other counties. However, there are houses in counties that are also very expensive, as indicated by the fourth quantile and points above the whiskers of these box plots. This is reasonable because of some prime locations within these counties.

### Box Plot of Price by cri.level

boxplot\_price(crime.level)



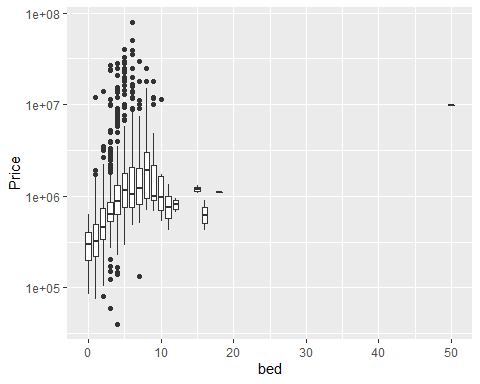
It can be observed that the median price of the house in the area where crime is lowest is higher than that in the area where crime is high.

***Box Plot of house price by numerical variables***

boxplot\_numerics = function (x){  
 (df\_housing.nona %>%  
 ggplot(aes({{x}}, Price, group = {{x}}))+  
 geom\_boxplot()+scale\_y\_log10())  
   
}

### Box plot of Price by number of bed

boxplot\_numerics (bed)



It is observed that the high median price of the house, which has more number of bedrooms. However, median prices are relatively low for the houses, which have more than 10 bedrooms, as compared to the houses, which have 8 or 9 bedrooms. This is likely due to the drop in house price, which has an increasing number of sets of bedrooms than single bedrooms.

### Boxplot of Price by the number of baths

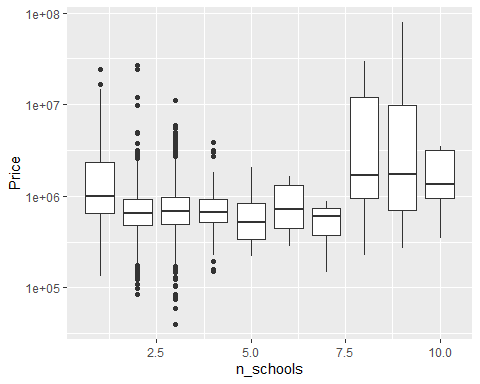
boxplot\_numerics(bath)



It is observed that the high median price of the house, which has more number of bathrooms. However, median prices are relatively low for the houses, which have more than 10 bathrooms, as compared to the houses, which have 8 or 9 bathrooms. This is likely due to the drop in house price, which has an increasing number of sets of bathrooms than the single bathrooms.

### Boxplot of Price by the number of schools

boxplot\_numerics (n\_schools)

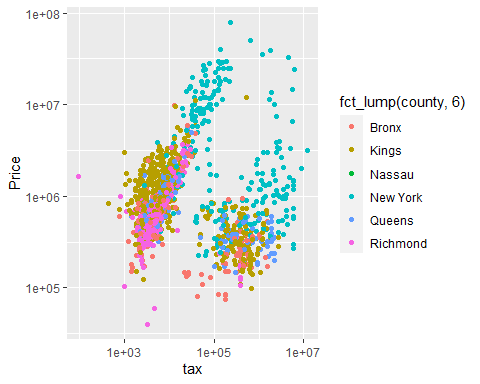


It is observed that the median prices of houses are high, in the areas, where there are more than 7 schools. More often than not, the higher the number of schools in the neighborhood, higher the house prices.

## Scatterplot of Price against numerical and categorical variables

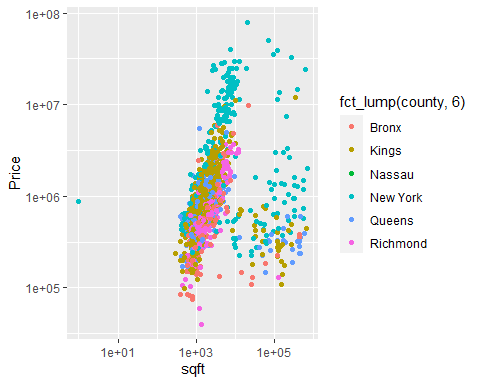
lmplot\_numerics = function (x){  
 df\_housing.nona %>%  
 ggplot(aes({{x}}, Price, color = fct\_lump(county, 6))) +  
 geom\_point() +  
 scale\_x\_log10() + scale\_y\_log10()  
}

lmplot\_numerics(tax)



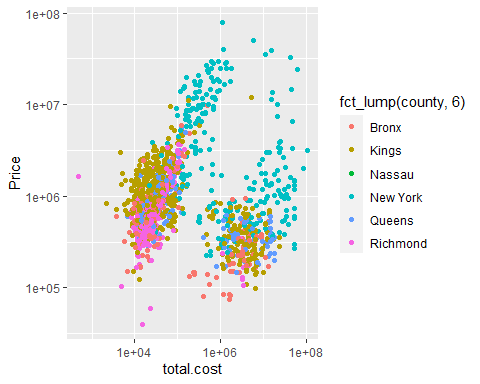
It is observed that there is a positive correlation between tax and house price. However, since tax depend on house prices, we will not use tax as predictor for house price.

lmplot\_numerics(sqft)



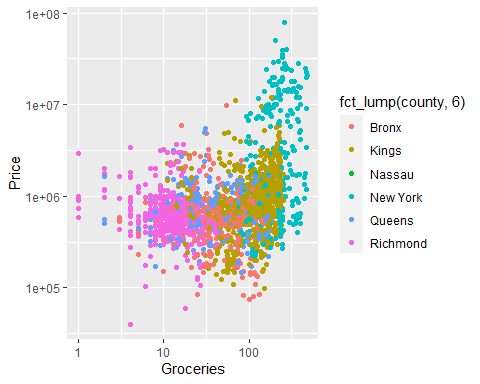
We observed that there is strong positive correlation between house price and square feet. We also see that the house prices based in square feet is higher in New York County, followed by kings, Richmond, the Bronx and then Queens.

lmplot\_numerics(total.cost)



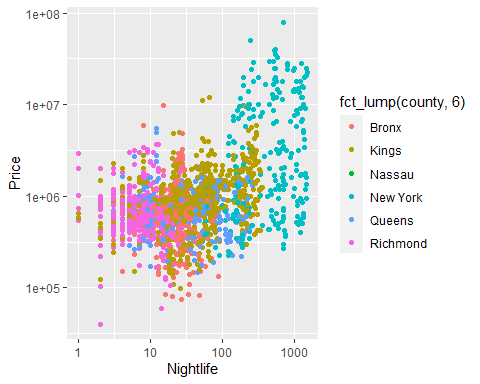
We observed that there is positive correlation between total cost, which is the sum of land assessment cost and house improvement cost, and house price.

lmplot\_numerics(Groceries)



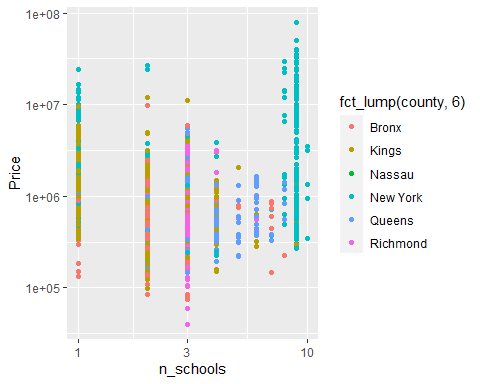
We observed that there is positive relationship between house price and number of groceries in a neighborhood.

lmplot\_numerics(Nightlife)



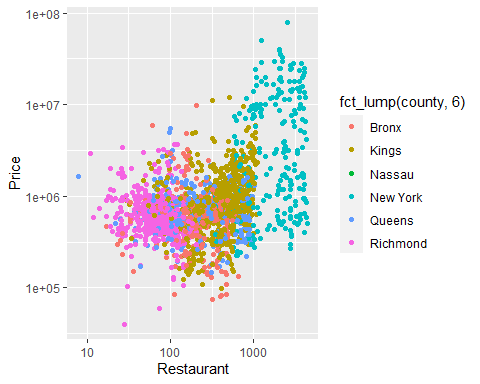
We observed that there is positive relationship between house price and number of nightlife in a neighborhood.

lmplot\_numerics(n\_schools)



We observed that there is no clear pattern of correlation between house prices in a neighborhood and the number of schools in that neighborhood from our multiple regression analysis.

lmplot\_numerics(Restaurant)

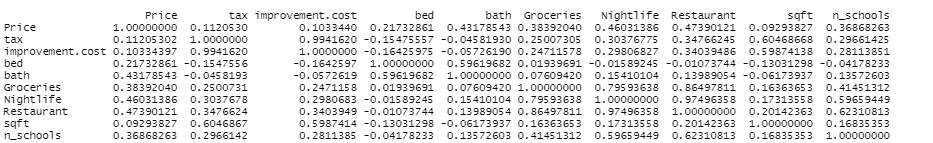


We observed that there is a positive relationship between house price and the number of restaurants in a neighborhood

# Section 4: Development of a model to predict the house price in NYC using Multi-Linear Regression

Since, n\_schools, Nightlife, Restaurants, Groceries, sqft, tax, bed, bath and total cost all affect house Price. We will develop a model using a multivariate regression model that best predicts the house price.

### Examining the correlation between the independent variables



As it can be seen that tax is highly correlated with improvement.cost, so we exclude tax. Likewise, bed and bath have a positive correlation but we include both of them as they are different variables. Similarly, Nightlife, Restaurant, and Groceries, we include only Nightlife.

### Updated model

model.updated = lm(log(model$Price) ~   
 log(model$bath) +  
 log(model$improvement.cost) +   
 log(model$n\_schools) +  
 log(model$Nightlife) +  
 log(model$bed)+   
 log(model$sqft))  
   
   
summary(model.updated)

##   
## Call:  
## lm(formula = log(model$Price) ~ log(model$bath) + log(model$improvement.cost) +   
## log(model$n\_schools) + log(model$Nightlife) + log(model$bed) +   
## log(model$sqft))  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.1535 -0.3030 -0.0199 0.2856 3.2599   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 11.065162 0.091443 121.006 < 2e-16 \*\*\*  
## log(model$bath) 0.739455 0.029307 25.231 < 2e-16 \*\*\*  
## log(model$improvement.cost) -0.001733 0.007225 -0.240 0.810434   
## log(model$n\_schools) 0.076092 0.022023 3.455 0.000561 \*\*\*  
## log(model$Nightlife) 0.291566 0.008852 32.937 < 2e-16 \*\*\*  
## log(model$bed) 0.344929 0.033727 10.227 < 2e-16 \*\*\*  
## log(model$sqft) 0.056542 0.011240 5.030 5.3e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.5165 on 2169 degrees of freedom  
## Multiple R-squared: 0.6587, Adjusted R-squared: 0.6577   
## F-statistic: 697.5 on 6 and 2169 DF, p-value: < 2.2e-16

The p - values for total cost is very high, so we do not reject the hypothesis that their betas’s values are 0. But, those for the other predictors (n\_schools, Nightlife, bath and sqft, bed) are very small, hence we can conclude that their slopes are not 0.

Now, lets run the multiple regression with remaining five variables.

model.updated = lm(log(model$Price) ~   
 log(model$bath) +  
 log(model$n\_schools) +  
 log(model$Nightlife) +  
 log(model$bed)+   
 log(model$sqft))  
   
   
summary(model.updated)

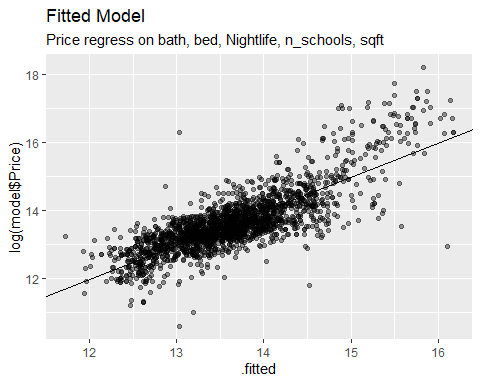
##   
## Call:  
## lm(formula = log(model$Price) ~ log(model$bath) + log(model$n\_schools) +   
## log(model$Nightlife) + log(model$bed) + log(model$sqft))  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.1568 -0.3028 -0.0192 0.2860 3.2620   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 11.054837 0.080663 137.050 < 2e-16 \*\*\*  
## log(model$bath) 0.739799 0.029265 25.279 < 2e-16 \*\*\*  
## log(model$n\_schools) 0.075407 0.021832 3.454 0.000563 \*\*\*  
## log(model$Nightlife) 0.290772 0.008209 35.423 < 2e-16 \*\*\*  
## log(model$bed) 0.347774 0.031566 11.017 < 2e-16 \*\*\*  
## log(model$sqft) 0.055398 0.010176 5.444 5.8e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.5163 on 2170 degrees of freedom  
## Multiple R-squared: 0.6586, Adjusted R-squared: 0.6579   
## F-statistic: 837.4 on 5 and 2170 DF, p-value: < 2.2e-16

from the revised summary of multiple regression analysis, we observe that p values of all the variables are less than 0.05, indicating that they are highly significant. This implies that we do not accept the null hypothesis that the house price is not dependent on these variables. Hence, we conclude that these variables are very strong predictors of house prices.

library(broom)

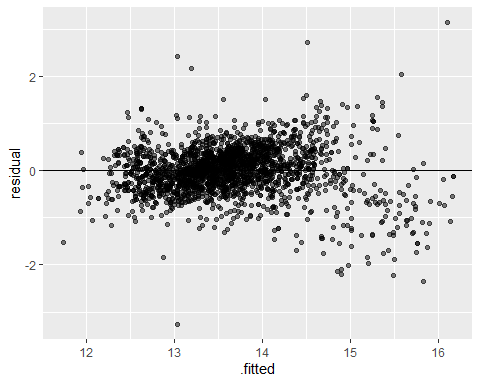
# Regression Plot of the fitted model

reg.model = augment(model.updated) %>%   
 ggplot(aes(x = .fitted,   
 y =`log(model$Price)`)) +   
 geom\_point(alpha = 0.4) +   
 geom\_abline() +  
 labs (title ="Fitted Model",  
 subtitle = "Price regress on bath, bed, Nightlife, n\_schools, sqft")  
reg.model



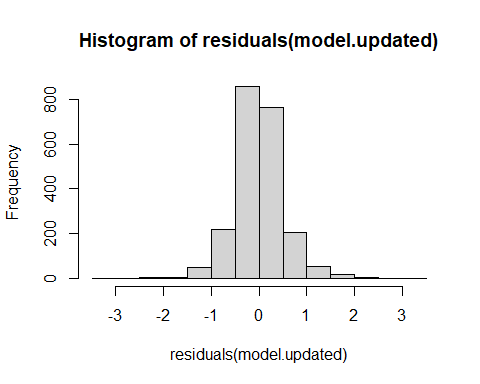
Obviously, there is a high positive correlation between house prices and the dependent variables which are bath, bed, square feet, number of schools, and nightlife. These predictors are the necessary variables from the data in estimating house prices in neighborhoods of NYC. The implication here is, that knowing this information which are easy to come by without having access to the data will educate the general public on a better estimation of house price with little to no knowledge of data analysis.

### Residual plot of the fitted model



From the diagnostic test of the residuals, it is observed that the variance of the residuals is nicely stabilized, even though there are a few points that are scattered to the right. Therefore, the model is valid and is a good estimator of house prices.

### Histogram of residuals of fitted model



The nicely normally distributed histogram of residuals is evidence of a valid model.

# Section 5: Machine learning Model

In this section we build a machine learning model to predict the crime level of the location in NYC based on nightlife, n\_schools, housing.type, and per.commute.car.

### Prediction of Crime level (crime.level) based in the other categorical variable

***Data Preparation for KNN Model*** Since KNN in the distance-based algorithm, it requires all the predictive variables in numeric type. so we convert all the variables to numeric. Also, we create a dummy variable of all the nominal type categorical variables and scale the variables to standardize using a standard scaler.

***Split the data into the training and test sets***

We partition 75% of the data into the training set and the remaining 25% into the test set

set.seed (3000)  
smp\_size = floor (0.75 \* nrow(knn\_dummy))  
  
train\_ind = sample(seq\_len(nrow(knn\_dummy)), size = smp\_size)  
  
class\_pred\_train = knn\_dummy[train\_ind, ]  
class\_pred\_test = knn\_dummy[-train\_ind, ]

Split outcome variable into training and test sets using the same partition as above

cl\_outcome\_train = cl\_outcome[train\_ind, ]  
cl\_outcome\_test = cl\_outcome[-train\_ind, ]

***Run K-NN Classification***

cl\_pred = knn(train = class\_pred\_train, test = class\_pred\_test, cl = cl\_outcome\_train, k = 5)

**Model evaluation**

The table below is the head of the predicted vs actual class.

## Predicted\_Cl Observed\_Cl  
## 1 Low Moderate   
## 2 Low Low   
## 3 Low Lowest   
## 4 Low Lowest   
## 5 Lowest Lowest   
## 6 Lowest Lowest

**Confusion table**

## Predicted\_Value  
## Actual\_Value High Highest Low Lowest Moderate   
## High 5 1 5 10 3  
## Highest 2 0 3 4 2  
## Low 5 0 18 55 7  
## Lowest 5 1 24 346 12  
## Moderate

Our model did not perform well, because it classified most of the data points as the lowest. One reason for this poor performance of the model is class imbalance. Since most of the data points are of class lowest (about 70 percent of data points), the model learned to classify the lowest class but not other classes. The solution to improve the model could be to take the balance sample of each class, say 20 percent from each class and train the model, using a new balanced sample.

# Section 6: Average house price by zip code regress on total population by zip code

Now in this section, we will examine the relationship between average house price by zip code and total population by zip code in New York City. further we ,will develop the model, using bivariate regression, to predict the house price based on the population. For this analysis, we will calculate the average house price by zip code, using our main cleaned data set (df\_housing.nona), and merge this data set with the another data set called population1 to get the variable called population by zipcode.

***Loading datasets of Population by zipcode***

pop\_zip = read.csv('D://MSBA//Second Sem//OPR 9750//Project//Housesell//data\_lat\_long2.csv//population1.csv')  
pop\_zip['zipcode'] = pop\_zip$ï..zip

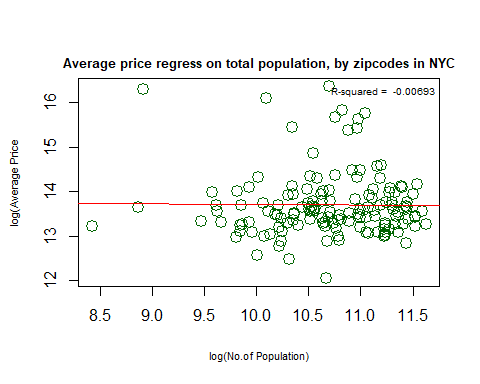
***Merging two datasets (avg.price\_zip and pop\_zip)***

## Joining, by = "zipcode"

### Regression Analysis

### Develop the regression model

plot(log (merge.df$pop) , log(merge.df$avg\_price),  
 xlab = "log(No.of Population)",   
 ylab = "log(Average Price",   
 col = "darkgreen", cex = 1.7, lwd = 0.6, cex.lab = 0.7)  
linear\_regr = lm( log(merge.df$avg\_price)~log(merge.df$pop))  
legend("topright", legend = paste('R-squared = ',   
 format(summary(linear\_regr)  
 $adj.r.squared,   
 digit = 3)), bty = 'n', cex = 0.6)  
abline(linear\_regr, col = 'red')  
title (main = 'Average price regress on total population, by zipcodes in NYC',  
 line = 0.6, cex.main = 0.8)



**Summary of Regression**

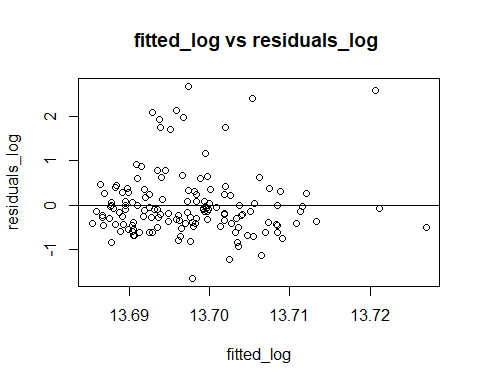
##   
## Call:  
## lm(formula = log(merge.df$avg\_price) ~ log(merge.df$pop))  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.6528 -0.4205 -0.1433 0.2386 2.6732   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 13.83632 1.10126 12.564 <2e-16 \*\*\*  
## log(merge.df$pop) -0.01299 0.10278 -0.126 0.9   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.7289 on 142 degrees of freedom  
## Multiple R-squared: 0.0001124, Adjusted R-squared: -0.006929   
## F-statistic: 0.01596 on 1 and 142 DF, p-value: 0.8996

The intercept has a significantly low p-value, however, the slope has high p-value of 0.8996. Hence we fail to reject the null hypothesis that the beta (coefficient) of the population is equal to zero, even though we observe a slightly negative correlation between population and house price. Normally, we understand that the house price tends to be positively correlated with the population because an increased population will increase economic activity, which pushes the price of houses up. However, in the case of New York City, particularly, New York County, the true number of its residents is less as compared to that of other counties of NYC. People commute from other boroughs or even from different states to work and do other economic activities in Manhattan, but they do not live in Manhattan. As an effect, Manhattan has high economic activities which continuously increase house prices, despite a low number of actual residents. According to the census of 2020, Kings County has a population of 2,736,074. Likewise, Queens has 2,405,464. The Bronx has 1,472,654 and Richmond has 495,747. It could be true that there could be a positive correlation between house prices and population if we do not include New York. But as said, New York City is such a diverse city, even the less densely populated regions are very expensive. For example, Zip Code 11213 is Crown Heights and is one of the most densely populated regions in Kings and overall NYC, however, the average house price is much less than other zip codes in Kings that have low population.

It is, therefore, the population is not a good predictor of house prices in NYC.

We can further examine by plotting the residuals of fitted line.

**Residuals vs fitted line**

As it can be observed that despite the stabilized variances, they are very scattered, indicating the high degree of uncertainty. Therefore, the population is not very good predictor of house price.

# Section 7: Summarization of House Price by various categorical variables

In this section, we calculate the average house price based on various categorical variables such as housing type, county, zip code, and crime level. We also determine the top 10 expensive zip codes and 10 least expensive zip codes for house prices in New York City.

### Summarization of house price by County

The table below shows the average house price by County

| County | Average Price |
| --- | --- |
| Bronx | 713026.8 |
| Kings | 1071977.3 |
| Nassau | 662498.0 |
| New York | 6533492.9 |
| Queens | 795756.3 |
| Richmond | 714623.8 |

New York County is the most expensive in NYC for housing, followed by Kings, Queens, Richmond, and Bronx. Bronx is the cheapest in terms of housing in NYC. it can also be observed that the difference in average house prices in New York County and other counties is huge.

### Summarization of House Price by House Type

The table below shows the average house price by housing type

| Housing Type | Average Price |
| --- | --- |
| Apartment | 1343833 |
| Condo | 2399665 |
| Coop | 1328201 |
| Multi Family | 1262291 |
| Single Family | 1083426 |
| Townhouse | 4823210 |
| Unknown | 610000 |

The most expensive type of housing in NYC is townhouse, followed by condos and apartments.

### Summarization of House Price by crime level

The table below shows the average house price by crime level

| Crime Level | Average Price |
| --- | --- |
| High | 1326483 |
| Highest | 1524702 |
| Low | 1055228 |
| Lowest | 1734461 |
| Moderate | 1194541 |

Average price of houses is high, in the places, where the crime level is lowest. However, there are no significant variations in the house prices bases on crime level.

### 10 most expensive average house prices by County, zipcode

The table below shows the 10 most expensive zip codes in NYC for housing

|  | County | Zipcode | Average Price |
| --- | --- | --- | --- |
| 5 | New York | 10021 | 12872927 |
| 2 | New York | 10007 | 12110000 |
| 1 | New York | 10001 | 9807143 |
| 3 | New York | 10011 | 7399150 |
| 7 | New York | 10023 | 7037949 |
| 9 | New York | 10028 | 6362118 |
| 10 | New York | 10128 | 6072509 |
| 6 | New York | 10022 | 5146367 |
| 8 | New York | 10024 | 5028224 |
| 4 | New York | 10016 | 4818998 |

the 10 most expensive zip codes for the house price are all in New York County. 10021, which is on the eastern side of the Central Park, and 10007, which is in lower Manhattan are the most expensive zip codes for housing in NYC.

## 10 least expensive average house prices by zip code

The table below shows the 10 most expensive zip codes in NYC for housing

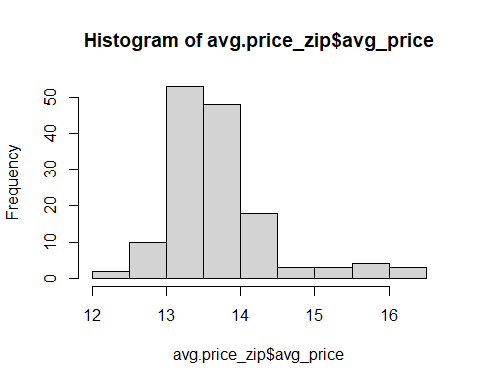
|  | County | Zipcode | Average Price |
| --- | --- | --- | --- |
| 7 | Bronx | 10475 | 170245.0 |
| 9 | Queens | 11423 | 260000.0 |
| 10 | Queens | 11692 | 287422.0 |
| 3 | Richmond | 10303 | 347565.1 |
| 5 | Bronx | 10456 | 377431.0 |
| 1 | New York | 10039 | 387000.0 |
| 8 | Queens | 11374 | 393380.5 |
| 6 | Bronx | 10459 | 397760.4 |
| 2 | Richmond | 10302 | 427515.7 |
| 4 | Bronx | 10451 | 433032.2 |

Zip code 10475 in the Bronx, which is the western side of Pelham Bay Park is the cheapest zip code for housing in NYC. The Zip code 11432 in Queens, which is situated in Hollis in the Jamaica region is also cheap for housing. The zip code 10039 is in Upper Manhattan is also cheap, which came out as a surprise.

# Section 8: Visualization of Average House Price by Zip code in NYC Map

In this section, we plot the average house price in NYC by zip code in the NYC map to visualize the area with varying average house prices.

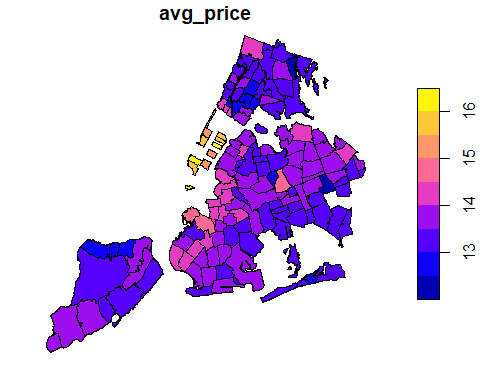
# Examining the distribution of average price of houses by zip code



### After taking the log, the distribution of the average price by zip code looks like normal.

### Plotting average house price by zipcode in NYC map

plot(map.df['avg\_price'])



The map of New York city depicts the average house price by zip codes in NYC. As expected, the most expensive zip codes for houses are in Manhattan. The area is just on the west side of the central park as indicated by the brighter yellow color on the map, which is a zip code of 10021 and is the most expensive zip code in terms of housing. Another Zip code 1000, the second most expensive in NYC, is situated in lower Manhattan and is also indicated in brighter yellow on the map. The third most expensive zip code for house price is 10001, which is across the 11th to 5th Avenue from 25th to 33rd street. This is also highlighted in brighter yellow on the map.

Interestingly, in Brooklyn, 11231 and 11215 are the most expensive zip codes for housing and they both are situated in western Brooklyn, just below the lower Manhattan, and are highlighted in brighter red on the map. Then after the regions in Northern Brooklyn, in the downtown and Williamsburg area. Eastern Brooklyn is the cheapest region. Likewise, on the map, there are regions in upper queens, the house price is expensive, but the most expensive zip code in queens is 11375, which is Forest Hills. The Staten Island and the Bronx, in particular, mid and the east Bronx have many zip codes that offer cheap housing as compared to the other regions of NYC.

### House prices by zip code in Richmond County

The table below shows the average house prices by zip codes of Richmond County.

| County | Zipcode | Average Price |
| --- | --- | --- |
| Richmond | 10301 | 770206.2 |
| Richmond | 10302 | 427515.7 |
| Richmond | 10303 | 347565.1 |
| Richmond | 10304 | 1122182.3 |
| Richmond | 10305 | 593549.9 |
| Richmond | 10306 | 614181.1 |
| Richmond | 10307 | 876955.5 |
| Richmond | 10308 | 592588.2 |
| Richmond | 10309 | 841996.8 |
| Richmond | 10310 | 441023.2 |
| Richmond | 10312 | 794494.3 |
| Richmond | 10314 | 569210.3 |

### House prices by ten most expensive zip codes in Kings County

The table below shows the top ten zip codes in Kings County for average house price

|  | County | Zipcode | Average Price |
| --- | --- | --- | --- |
| 9 | Kings | 11231 | 2857591 |
| 4 | Kings | 11215 | 2125536 |
| 10 | Kings | 11238 | 1930735 |
| 1 | Kings | 11205 | 1743005 |
| 8 | Kings | 11222 | 1684028 |
| 5 | Kings | 11216 | 1638339 |
| 6 | Kings | 11218 | 1612498 |
| 3 | Kings | 11211 | 1399525 |
| 2 | Kings | 11206 | 1356776 |
| 7 | Kings | 11219 | 1315510 |

**House prices by least 10 expensive zip codes in Kings County**

The table below shows the least ten zip codes in Kings County for average house price

|  | County | Zipcode | Average Price |
| --- | --- | --- | --- |
| 9 | Kings | 11236 | 553631.6 |
| 1 | Kings | 11203 | 558504.6 |
| 6 | Kings | 11229 | 574791.0 |
| 5 | Kings | 11224 | 591083.6 |
| 4 | Kings | 11212 | 591890.4 |
| 3 | Kings | 11208 | 689074.0 |
| 10 | Kings | 11237 | 708200.0 |
| 8 | Kings | 11232 | 723000.0 |
| 7 | Kings | 11230 | 749299.7 |
| 2 | Kings | 11207 | 814016.4 |
|  |  |  |  |

**House prices by least 10 expensive zip codes in Queens County**

The table below shows the top ten zip codes in Queens County for average house price

|  | County | Zipcode | Average Price |
| --- | --- | --- | --- |
| 9 | Queens | 11375 | 2180000.0 |
| 4 | Queens | 11357 | 1628003.3 |
| 3 | Queens | 11103 | 1268250.0 |
| 5 | Queens | 11362 | 1207022.5 |
| 7 | Queens | 11366 | 1178976.0 |
| 1 | Queens | 11101 | 1120000.0 |
| 2 | Queens | 11102 | 1098264.5 |
| 6 | Queens | 11365 | 966069.5 |
| 8 | Queens | 11367 | 960358.5 |
| 10 | Queens | 11379 | 956142.0 |

# Conclusion

New York City is truly very diverse in terms of the house price, which ranges from the US $ 40,000 to the US $ 79,000,000. As expected, the most expensive county is New York, where the average house price is about above $ 6.5 million, followed by Brooklyn, where the average house price is above 1 million. The average price of the townhouse, which is approximately $ 4.8 million is the highest among all other types of houses, followed by condos, which cost about $ 2.4 million on an average. However, the price also depends on many other variables such as location, crime levels, number of schools, number of restaurants, groceries, etc. Based on our multivariate regression model, the number of bedrooms, and bathrooms, square feet of the house, number of schools, and level of nightlife in the neighborhood are the most important explanatory variables that can estimate the house price reasonably. The implication here is, that knowing this information which is easy to come by without having access to the data will educate the general public on a better estimation of house prices with little to no knowledge of data analysis. The level of nightlife came out to be an important variable not only because it is highly correlated with the number of restaurants and groceries in the neighborhood but also because it is the proxy of locations. Houses in the areas in such as Manhattan, and western Brooklyn, where the nightlife is very high, the price is very high in general.

Interestingly, our study, based on regression analysis, showed that there is no relationship between population and house prices in New York City. Generally, it is expected that the house prices also depend on the population because the population is the driving variable of economic activity that in turn affects house prices. However, this came out to be an interesting fact that contradicts our beliefs that the prices of houses are high where the population is dense. The one reason could be in New York County, where the house price is highest, the true number of its residents is less compared to that of other counties of NYC. People commute from other boroughs or even from different states to work and do other economic activities in Manhattan, but they do not live in Manhattan. Consequently, Manhattan has high economic activities which continuously increase house prices, despite a low number of actual residents. According to the census of 2020, Kings County has a population of 2,736,074. Likewise, Queens has 2,405,464. The Bronx has 1,472,654 and Richmond has 495,747. It could be true that there could be a positive correlation between house prices and population if we do not include New York County in the study. But as said, New York City is such a diverse city, even the less densely populated regions are very expensive. For example, Zip Code 11213 is Crown Heights and is one of the most densely populated regions in Kings and overall NYC, however, the average house price is much less than other zip codes in Kings that have a low population.

From our map analysis of average house prices by zip codes in New York City, Interestingly, there are three areas in Manhattan, which are the west side of Central Park, Lower Manhattan having the World Trade Center and the area where the Empire State Building is located are the most expensive areas for housing. After Manhattan, where the average house price is highest, Brooklyn is the second most expensive in terms of housing, particularly, in western and northern Brooklyn, the price of houses is expensive. In upper Queens and some areas in lower and upper-west Bronx, houses are expensive. Staten Island is cheaper, followed by Bronx and Queens. The cheapest area is on the west side of Pelham Bay Park in the Bronx. We have also found that even in New York County, there is an area in upper Manhattan, where house price is below half a million.

Our machine learning model based on the K Nearest Neighbor algorithm to predict the crime level in New York City by using variables nightlife, the number of schools, percentage of commute by car, and housing type did not perform well to predict the class of crime level. One reason for the poor performance is class imbalance.

Housing data analysis is applicable in the real world as it gives an insight to potential homeowners who are looking to buy a home with some specifications like house type, location, and how to match various prices with their budget. Also, knowledge about key variables that impact house prices helps regular buyers become better at evaluating house prices without resulting to real estate agencies and other brokers. Not only do buyers save some change on the capital, but they also save time as the information from this analysis is a road map that leads them to the destination of their housing type with specifics they are looking to own. Furthermore, real estate developers can take a cue from the housing demographics of NYC, how house prices are changing with time, and where and how to invest their capital. The City and Federal governments will need the results or information from the analysis of such data to decide on myriad housing development projects in various neighborhoods, to best tackle the NYC housing demand crisis. Finally, potential tenants will have the privilege of knowing which neighborhood matches their budget and where to look for what is affordable. There is a high correlation between house prices and monthly rent or mortgage. Housing in NYC, just like most developed cities is expensive and hard to come by, so vital information about house prices is priceless for lessees. To summarize, information from the analysis of housing data comes in handy ranging from government developmental projects to the lessee, to the real estate developer, and also to the potential homeowner -new to the housing market who wants to fulfill the objective of buying assets at the right price with less to no headaches.