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 city in the world, particularly for the housing. However, the city is so diverse that the house prices vary significantly across the 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 codes and also house type. We have also developed the 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 are other explanatory variables. Surprisingly, we found that there is no correlation between average house price 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.

First section of this document discusses about 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 of the variables. Fourth section is the multivariate Regression Model. In fifth section, we have developed the machine learning model to predict the crime level. Sixth section is uni-variate regression analysis. Seventh is descriptive statistics that present the aggregated data by by the different categorical variables. The last section present 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 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 in 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 purpose of the analysis, we used the housing data that was available in Kaggle, an open source platform for data analysis. The link of the data set is embedded [here](https://www.kaggle.com/datasets/parv619/newyork-house-sale-data). Second data that we have used is population of NYC by zipcode which we have extracted from [this](https://data.cityofnewyork.us/Business/Zip-Code-Boundaries/i8iw-xf4u). We have use this dataset also for our map visualization

This uploaded dataset contains information of New York City neighborhood housings. 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 the purpose of our analysis.

# Section 2: Data Cleaning and Preparation

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

***import required basic libraries***

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

\***loading dataset**

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

#head(housing\_data,30)

***Examining the struture of the dataset***

#str(housing\_data)

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

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

As it is observed from the structure of dataframe that bath, bed and sqft variables have type character 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))

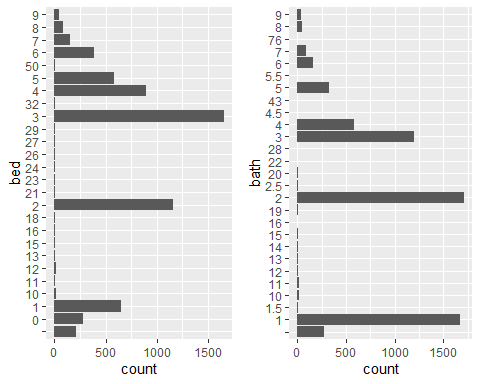
#Counting NA before conversion  
#colSums(is.na(housing\_data))

***Convert variables in to numeric***

housing\_data[, c('sqft', 'tax', 'land.assessment.cost', 'improvement.cost', 'total.cost')]= sapply(housing\_data[, c('sqft', 'tax', 'land.assessment.cost', 'improvement.cost', 'total.cost')],as.numeric)

Counting Unique values of bath and bed variables

ggplotGrid(ncol = 2,  
 lapply(c("bed", "bath"),  
 function(col) {  
 ggplot(housing\_data,aes\_string(col)) + geom\_bar() + coord\_flip()   
 }))

 As it can bee 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.

housing\_data[, c('bed', 'bath')]= sapply(housing\_data[, c('bed', 'bath')],as.numeric)

#counting NAs after the conversion  
#colSums(is.na(housing\_data))

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

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

***Extracting Number of schools from school.information***

#housing\_data$school.information = str\_extract\_all(housing\_data$school.information,"[1-9]{1}")  
  
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))  
  
  
# converting to Numeric data type  
  
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)))  
  
  
#table(housing\_data$crime.level)

***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$home.details)))  
#table(housing\_data$housing.type)

***Merging Datasets to get the county variables***

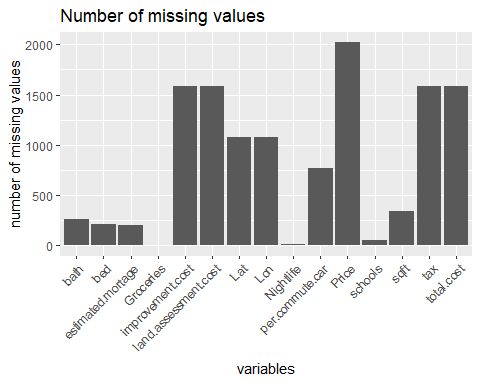
new\_data = read.csv("D://MSBA//Second Sem//OPR 9750//Project//Housesell//data\_lat\_long2.csv//population1.csv")  
new\_data['zipcode'] = new\_data$ï..zip  
housing\_data$zipcode = as.integer(housing\_data$zipcode)  
housing\_data = inner\_join(housing\_data, new\_data, by = 'zipcode')  
housing\_data = housing\_data %>% select(c(-ï..zip, -pop))

***Dropping unnecessary columns from the dataSet***

drop = c('X', "linktoproperty", 'provider.info', 'what.local.say.about.the.neighborhood','neighborhood.name', "shop.and.eat.information", "school.information", "commute.information","seo.description", "crime.information", "description","comparable.properties", "address", 'comments.of.residents.and.previous.residents', 'home.details', 'price.details')  
  
df\_housing = housing\_data[,!(names(housing\_data) %in% drop)]

***Checking Missing Values*** #### barplot of missing values

missing.values <- df\_housing %>%  
 gather(key = "key", value = "val") %>%  
 mutate(is.missing = is.na(val)) %>%  
 group\_by(key, is.missing) %>%  
 summarise(num.missing = n()) %>%  
 filter(is.missing==T) %>%  
 select(-is.missing) %>%  
 arrange(desc(num.missing))   
  
missing.values %>%  
 ggplot() +  
 geom\_bar(aes(x=key, y=num.missing), stat = 'identity') +  
 labs(x='variables', y="number of missing values", title='Number of missing values') +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))

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

df\_housing.nona = na.omit(df\_housing)  
#colSums(is.na(df\_housing.nona))

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

***Converting the schools variable to n\_schools by taking the sum of digits***

total\_schools = c()  
digitsum2 <- function(x) sum(floor(x / 10^(0:(nchar(x) - 1))) %% 10)  
for (i in df\_housing.nona$schools){  
 total\_schools = append(total\_schools, digitsum2(i))  
}  
  
df\_housing.nona$n\_schools = total\_schools

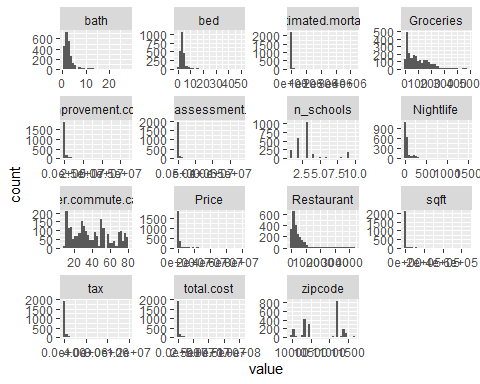
# Section 3: Data Explorations and Descriptive Statistics

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

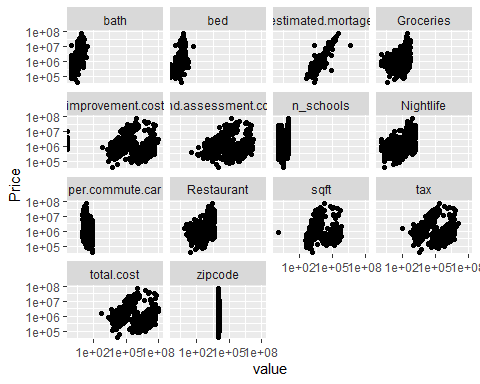
summary(df\_housing.nona)

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

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

 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.

## Warning: Transformation introduced infinite values in continuous x-axis

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

***Checking 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.

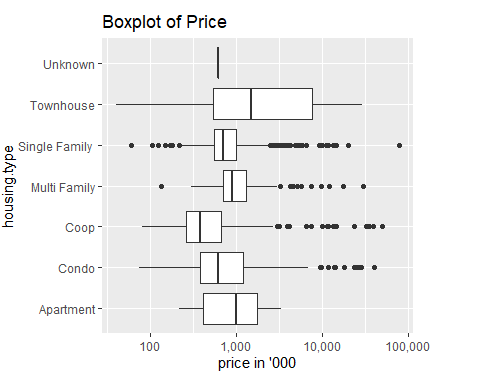
## Examining outliers and the distribution

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

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

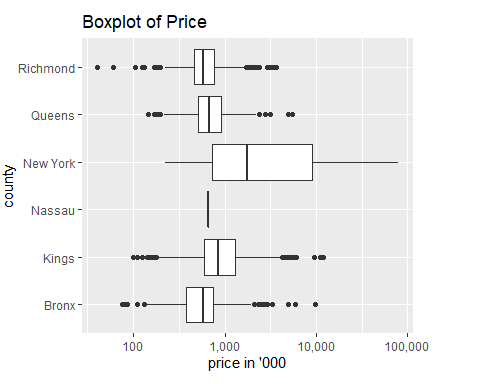
boxplot\_price(housing.type)



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

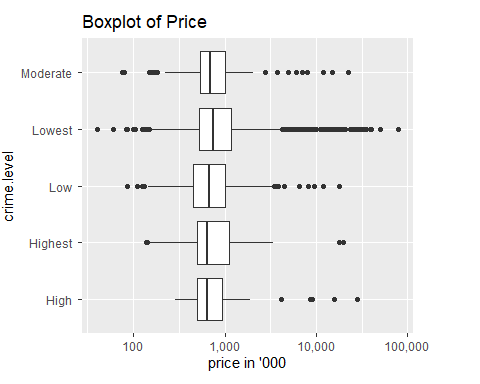
### Boxplot of Price by county

boxplot\_price(county)

 From the above the 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 with in this counties.

### Box Plot of Price by cri.level

boxplot\_price(crime.level)

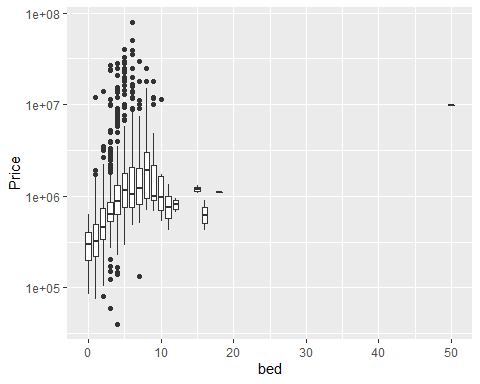
 It can be observed that the median price of the house in the area where crime is lowest is higher than that in the area where crime is high.

\*\*\* Box Plot of house price by numeraical 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 house, which has more number of bedrooms. However, median prices are relatively low for the houses, which has more than 10 bedrooms, as compare to the houses, which has 8 or 9 bedrooms. This is likely due to the drop in house price, which has increasingly number of set of bedrooms than single bedrooms.

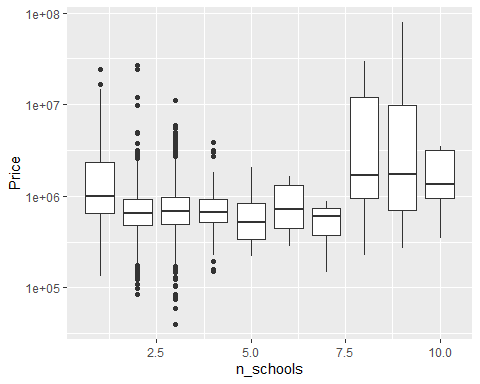
### Boxplot of Price by number of bath

boxplot\_numerics(bath)

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

### Boxplot of Price by number of schools

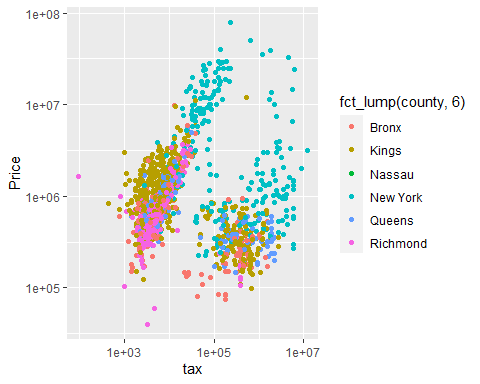
boxplot\_numerics (n\_schools)

 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.

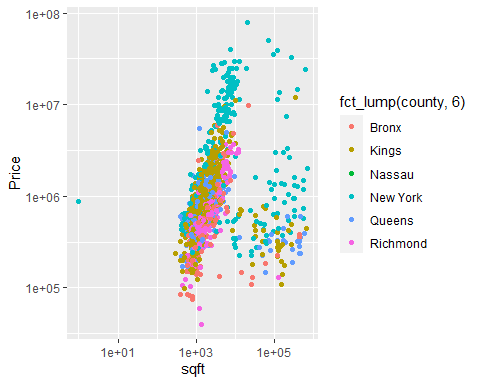
## Scatterplot of Price against numerical and categorical variables

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

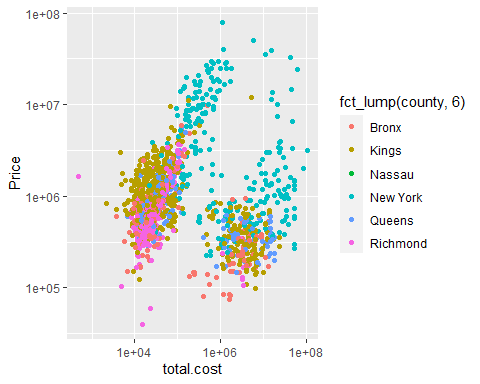
lmplot\_numerics(tax)

 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)

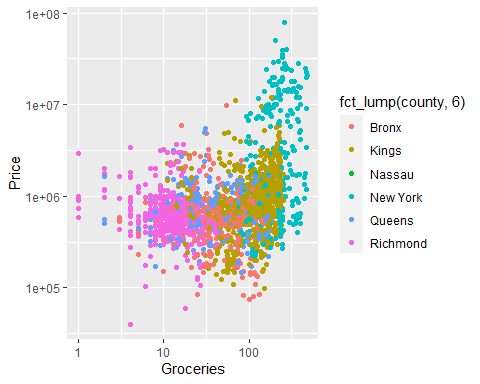
 We observed that there is strong positive correlation between house price and square feet. We also see that the house prices based in square feet is higher in New York county, followed by kings, Richmond, the Bronx and then Queens.

lmplot\_numerics(total.cost)

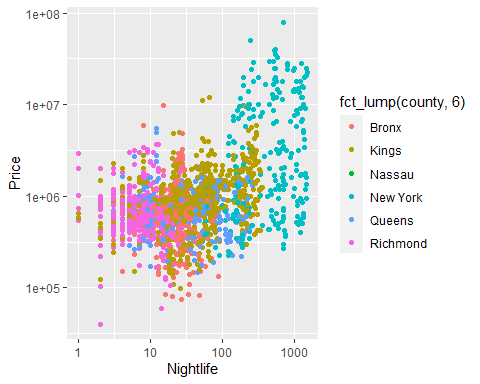


We observed that there is positive correlation between total cost, which is the sum of land assessment cost and house improvement cost, and house price.

lmplot\_numerics(Groceries)

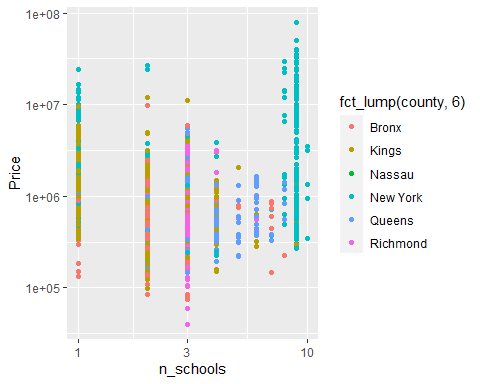
 We observed that there is positive relationship between house price and number of groceries in a neighborhood.

lmplot\_numerics(Nightlife)

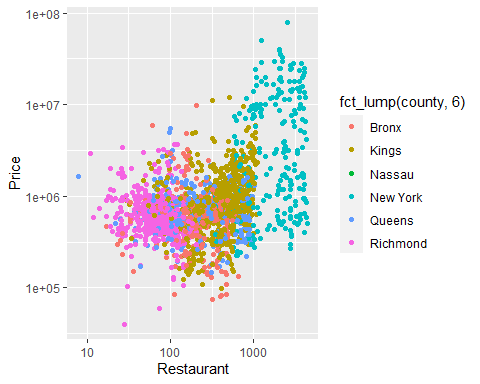


We observed that there is positive relationship between house price and number of nightlife in a neighborhood.

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 positive relationship between house price and number of restaurant in a neighborhood

# Section 4: Development of 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 multi-variate 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 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)

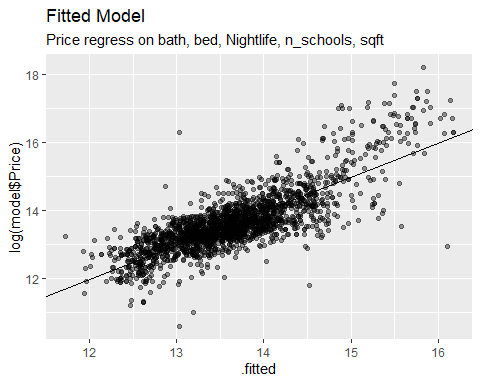
##   
## Call:  
## lm(formula = log(model$Price) ~ log(model$bath) + log(model$n\_schools) +   
## log(model$Nightlife) + log(model$bed) + log(model$sqft))  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.1568 -0.3028 -0.0192 0.2860 3.2620   
##   
## Coefficients:  
## Estimate Std. Error t 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 observer 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 for house prices.

library(broom)

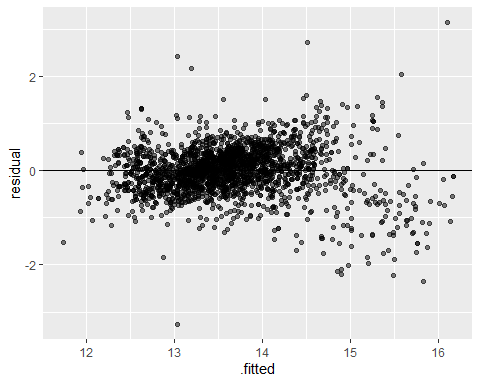
# Regression Plot of the fitted model

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

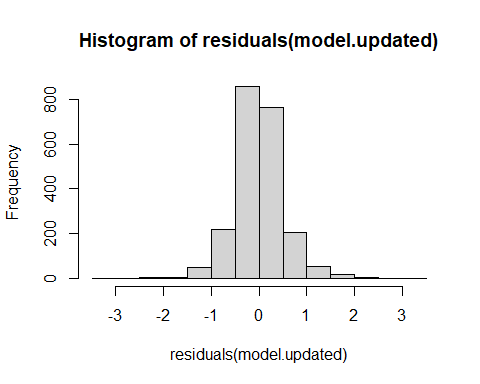


Obviously, there is a high positive correlation between house prices and the dependent variables which are bath, bed, square feet, number of schools, nightlife.These predictors are the necessary variables from the data in estimating house prices in neighborhoods of NYC. The implication here is, knowing these 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.

### Ressidual plot of the fitted model

 From the diagnostic test of the residuals, it is observed that the variance of the residuals are nicely stabilized, even though there are few points that are scattered to the right. Therefore, 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 an evidence that the 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

***copy the cleaned dataset for KNN***

knn = df\_housing.nona

***change all the variables names to lower case.***

***Import required libraries***

##   
## Attaching package: 'ggvis'

## The following object is masked from 'package:ggplot2':  
##   
## resolution

##   
## Attaching package: 'recipes'

## The following object is masked from 'package:stringr':  
##   
## fixed

## The following object is masked from 'package:stats':  
##   
## step

##   
## Attaching package: 'crosstable'

## The following object is masked from 'package:purrr':  
##   
## compact

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

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

We partition 70% 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*** we take k = 41 (approx value of the square root of the 1171, which is the total number of training sets)

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

***Model evaluation***

***Create confusion table***

Our model did not performed well, because it classified most of the data points as a 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), model learned to classify lowest class but not other class. The solution to improve the model could be to take the balance sample of each class, say 20 percent from each class and tarin the model, using new balanced sample.

# Section 6: Average house price by zipccode regress on total population by zipcode

Now in this section, we will examine the the relationship between average house price by zipcode and total population by zipcode 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.

\*\*\* calculation of average house price by zip code.\*\*\*

***Convert avg.price\_zip to a dataframe***

***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)***

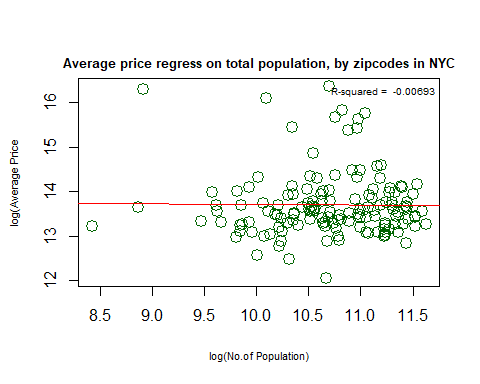
## Joining, by = "zipcode"

### Regression Analysis

***Convert all the integer type variable to the numeric***

### Develop the regression model

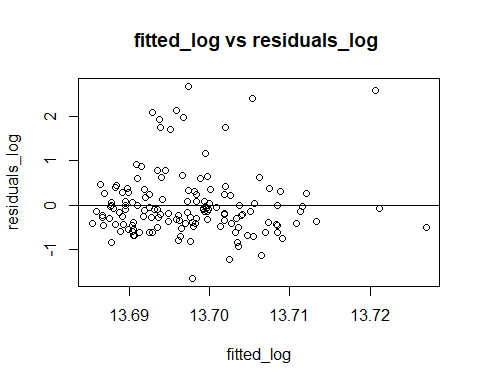
plot(log (merge.df$pop) , log(merge.df$avg\_price),  
 xlab = "log(No.of Population)",   
 ylab = "log(Average Price",   
 col = "darkgreen", cex = 1.7, lwd = 0.6, cex.lab = 0.7)  
linear\_regr = lm( log(merge.df$avg\_price)~log(merge.df$pop))  
legend("topright", legend = paste('R-squared = ',   
 format(summary(linear\_regr)  
 $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 significant low p-value, however slope has high p-value of 0.8996. Hence we fail to reject null hypothesis that the beta (coefficient) of population is equal to zero, even though we observe slightly negative correlation between population and house price. Normally, we understand that the house price tends to positively correlated with the population, because increase population will increase economic activity, which push price of houses up. However, in the case of New York City, particularly, the New York County, true number of its residents is less as compare to that of others 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 increases house prices, despite low number of actual residents. According to the census of 2020, Kings County has population of 2,736,074. Likewise, Queens has 2,405,464. Bronx has 1,472,654 and Richmond has 495,747. It could be true that there could be positive correlation between house price 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 an example,Zip Code 11213 is Crown Heights and is one of the most densely populated regions in Kings and overall NYC, however, average house price is much less than other zipcodes in Kings that has low population.

it is therefore, population is not a good predictor of house price in NYC.

We can futher 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 variable

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

### Summarization of house price by County

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

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

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

## Selecting by Average Price

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

## 10 least expensive average house price by zipcode

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

## Selecting by Average Price

|  | 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 |

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

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

In this section we plot the average house price in NYC by zipcode in the NYC map to visualize the area with varying average house prices. # load NYC map data

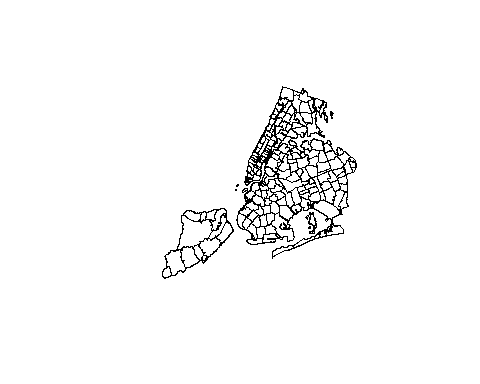
library(sf)

## Linking to GEOS 3.9.1, GDAL 3.2.1, PROJ 7.2.1; sf\_use\_s2() is TRUE

loc = "D://MSBA//second Sem//OPR 9750//DataSets//ZIP\_CODE\_040114"  
NYCmap = st\_read(dsn = loc, layer = "ZIP\_CODE\_040114")

## Reading layer `ZIP\_CODE\_040114' from data source   
## `D:\MSBA\Second Sem\OPR 9750\DataSets\ZIP\_CODE\_040114' using driver `ESRI Shapefile'  
## Simple feature collection with 263 features and 12 fields  
## Geometry type: POLYGON  
## Dimension: XY  
## Bounding box: xmin: 913129 ymin: 120020.9 xmax: 1067494 ymax: 272710.9  
## Projected CRS: NAD83 / New York Long Island (ftUS)

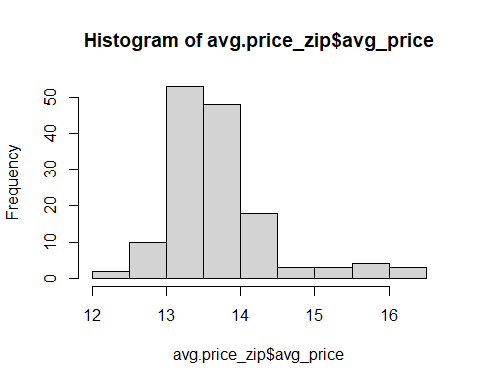
#str(NYCmap)  
plot(st\_geometry(NYCmap))



## Merging avg.price\_zip and NYCmap

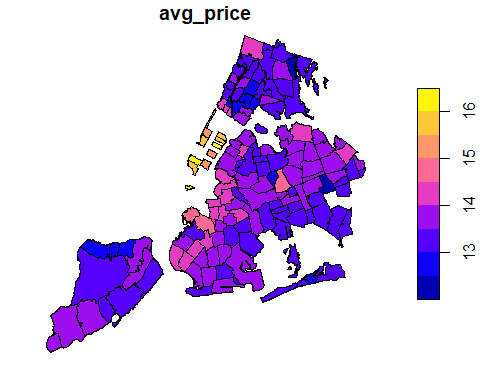
# creating ZIPCODE variable and converting it to numeric

hist(avg.price\_zip$avg\_price)

 # merging avg.price\_zip and NYCmap

### Ploting 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 zipcodes in the NYC. As expected the most expensive zip codes for houses are in the Manhattan. The area is just in west side of the central park as indicated by the brighter yellow color in the map, which is a zipcode of 10021, and is the most expensive zipcode in terms of housing. Another Zipcode 1000, second most expensive in NYC, which is situated in lower Manhattan, and is also indicated in brighter yellow in the map. The third most expensive zipcode 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 in the map.

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

### House price by zipcode in Richmond County

the table below shows the average house prices by zipcodes 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 price by ten most expensive zipcodesin Kings county

the table below shows the top ten zipcodes in Kings County for average house price

## Selecting by Average 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 price by least 10 expensive zipcodes in Kings county the table below shows the least ten zipcodes in Kings County for average house price

## Selecting by Average 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 |

the table below shows the top ten zipcodes in Queens Coonty for average house price

## Selecting by Average 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 US $ 40,000 to the US $ 80,000,000. As expected, the most expensive county is New York, where the average house price is about above $ 6.5 million, followed by the 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. 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, our study, based on regression analysis, showed that there is no relationship between population and house prices in New York City. 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 Manhatten, which are west side of Central Park, Lower Manhatten having World Trade Center and the area where 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. Our machine learning model based on the K Nearest Neighbor algorithm to predict the crime level in New York City by using variables night life, 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. It is obvious that 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.