GR5291 Project Report

Analysis and Prediction of New York City Housing Price

Group Members: Hongwei Chen(hc3272)
Huiying Wang(hw2816)
Jiyao Liu(jl5985)
Kexin Shi(hs3296)
Lili Tan(lt2846)

I. Introduction

New York City real estate is booming as an increasing number of people are moving to one of the largest, most modern and influential cities in the world. Since more people here want to find a house to dwell in, we intend to have a better understanding of the factors affecting the house prices as well as provide people with information about predictions of future pricing.

The model we are using takes into consideration the effects boroughs in New York, house categories, build year of the property, number of crimes, mortgage rate and median household income in New York have on the housing sale prices in New York. We expect Manhattan of all boroughs, three-family homes or rentals and the most recently built houses to have the largest positive impacts on house prices in New York. Apart from that, we believe an increase in the number of crimes and decreases in mortgage rates and median household income are likely to negatively affect the house prices. We hope to use this model to predict housing prices in New York for future years.

II. Data Collection and Data Description

Our main dataset includes information of house pricing in New York City in 2006-2019. The dataset primarily contains some categories about houses in New York City: Borough, neighborhood, building class, address, zip code, address, gross square

feet, land square feet, year built, sale price and sale date. Originally, we planned to utilize variables of Borough, gross square feet, year built, and sale date in this dataset to predict the house's sale price for the next year. However, we realized that the house prices will always be increasing if we simply use these variables and will not reflect the real situations behind the house pricing trend. Being mindful of the concerns aforementioned, we decided to combine other factors that are relevant to house prices in our model to improve our prediction: crime counts, mortgage rate, and median household income in New York.

We believe the crime rate is a significant factor that influences the house price: if the crime rate in a region is relatively high, fewer people would be willing to live in that area, leading to a drop in the house price. Therefore, the second dataset we used is New York City Crime. The dataset includes all valid felony, misdemeanor, and violation crimes reported to the New York City Police Department (NYPD) from 2006 to the end of 2019. The dataset has over 6 million crime reports from 2006 to 2019, with a total of 35 variables each. After going through the dataset, we decided to pick two variables: **BORO_NM** - The name of the borough in which the incident occurred, **CMPLNT_FR_DT** - Exact date of occurrence for the reported event. With information of borough and date of occurrence, we can merge crime information with our main dataset.

Mortgage rate is also an important factor of changes in house price since it's a rate of interest a homebuyer pays to finance a mortgage purchase and it takes into account larger economic factors such as inflation rate, employment rate, stock and bond market, 10-year Treasury yields, as well as the borrowers' financial and credit situations. Mortgage rates in a year reflect how likely people will buy houses; normally, high mortgage rates will make houses less affordable and homebuyers less inclined to buy houses. We obtained weekly data of the Freddie Mac 30-year fixed mortgage average rate from 2003 to 2019 in the United States from Federal Reserve Economic Data.

We also take real median household income into account since the purchase of a house is also related to the purchasing power of the household and the household income can explain that. A high median household income can show that homebuyers are more affordable to buy a house compared with a low median household income. We use the data of real median household income in New York from 1984-2020 and we mainly focus on the year we want to investigate. The source of this data comes from the U.S Census Bureau.

III. Data Cleaning

Reporting the data-cleaning efforts is essential for tracking alterations to the data given that there will be some level of error no matter how the data is collected. While some of the differences are justified because they reflect differences in the environment, others may reflect measurement or entry errors. These errors may be due to human error, poorly designed recording systems, or simply not having complete control over the format and type of data imported from external data sources.

In our dataset, we chose the data from 2006 to 2018 as our train data and the data of 2019 as our test data after prediction. Then, we plotted the boxplot of sale price during 2006 and 2018 and removed some outliers (i.e. extremely small data or 0) in the dataset, which consist of approximately 3 percent of our total data. In addition, since we needed to combine several more datasets, we merged our original housing data with crime, mortgage rate, and median household income in the New York City dataset based on date and borough. Finally, we got a complete dataset with date, sale price, crime, mortgage rate, and median household income in New York City.

To be more specific, we first processed the original housing price dataset. One explanatory variable **BUILDING.CLASS.CATEGORY** contains around 50 different building types such as “01 ONE FAMILY HOMES”, "07 RENTALS - WALKUP APARTMENTS", "27 FACTORIES". In R, categorical variables will be recoded into a series of dummy variables. To avoid the model being too complex, we selected the

categories that had predictive value and appeared most frequently, and further recoded classes where similarities exist into new ones. The details are shown below:

- CONDOS: "12 CONDOS - WALKUP APARTMENTS", "13 CONDOS - ELEVATOR APARTMENTS", "15 CONDOS - 2-10 UNIT RESIDENTIAL";
- COOPS: "09 COOPS - WALKUP APARTMENTS", "10 COOPS - ELEVATOR APARTMENTS";
- ONE FAMILY: "01 ONE FAMILY HOMES", "01 ONE FAMILY DWELLINGS";
- TWO FAMILY: "02 TWO FAMILY HOMES", "02 TWO FAMILY DWELLINGS";
- THREE FAMILY: "03 THREE FAMILY HOMES", "03 THREE FAMILY DWELLINGS";
- RENTAL: "07 RENTALS - WALKUP APARTMENTS", "08 RENTALS - ELEVATOR APARTMENTS", "14 RENTALS - 4-10 UNIT"

Next, to get the crime data frame containing the number of reports per day for each borough from 2006 to 2019, we grouped and summarized the NYPD_Complaint_Data_Historic dataset by two variables -- **BORO_NM** and **CMPLNT_FR_DT**, which indicated the name of the borough in which the incident occurred, and exact date of occurrence for the reported event separately.

Then, merge the processed housing price dataset and the crime dataset by c("**borough_name**"="**BORO_NM**","**SALE.DATE**"="**CMPLNT_FR_DT**"), i.e. by the borough name and the date (for the former, the date of sale, for the latter, the date when events occurred). Similarly, we merged this new data frame with the mortgage rate and median household income in New York by date.

Compared to the original dataset, we have simplified categories of the housing buildings and added three more explanatory variables – **crime_counts**, **mortgage**, and **income**.

The final dataset is shown below:

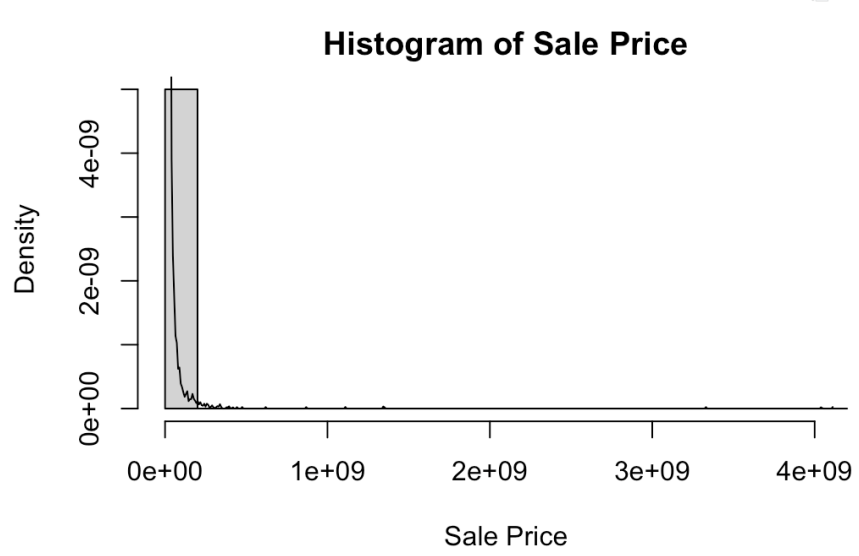
A tibble: 662,162 × 8

borough_name <chr>	category <chr>	YEAR.BUILT <dbl>	SALE.DATE <date>	crime_counts <dbl>	mortgage <dbl>	income <dbl>	SALE.PRICE <dbl>
MANHATTAN	CONDOS	1928	2006-07-20	424	6.80	62060	1200000
MANHATTAN	CONDOS	1928	2006-02-10	409	6.24	62060	695000
MANHATTAN	CONDOS	1928	2006-10-25	376	6.36	62060	585000
MANHATTAN	CONDOS	1928	2006-01-27	394	6.12	62060	460000
MANHATTAN	CONDOS	1928	2006-06-23	409	6.71	62060	725000
MANHATTAN	CONDOS	1920	2006-05-25	403	6.62	62060	715000
MANHATTAN	CONDOS	1905	2006-11-15	391	6.33	62060	763687
MANHATTAN	CONDOS	1905	2006-09-25	375	6.40	62060	1069162
MANHATTAN	CONDOS	2005	2006-04-25	321	6.53	62060	1420458
MANHATTAN	CONDOS	2005	2006-05-10	353	6.59	62060	1298268

1–10 of 662,162 rows Previous 1 2 3 4 5 6 ... 100 Next

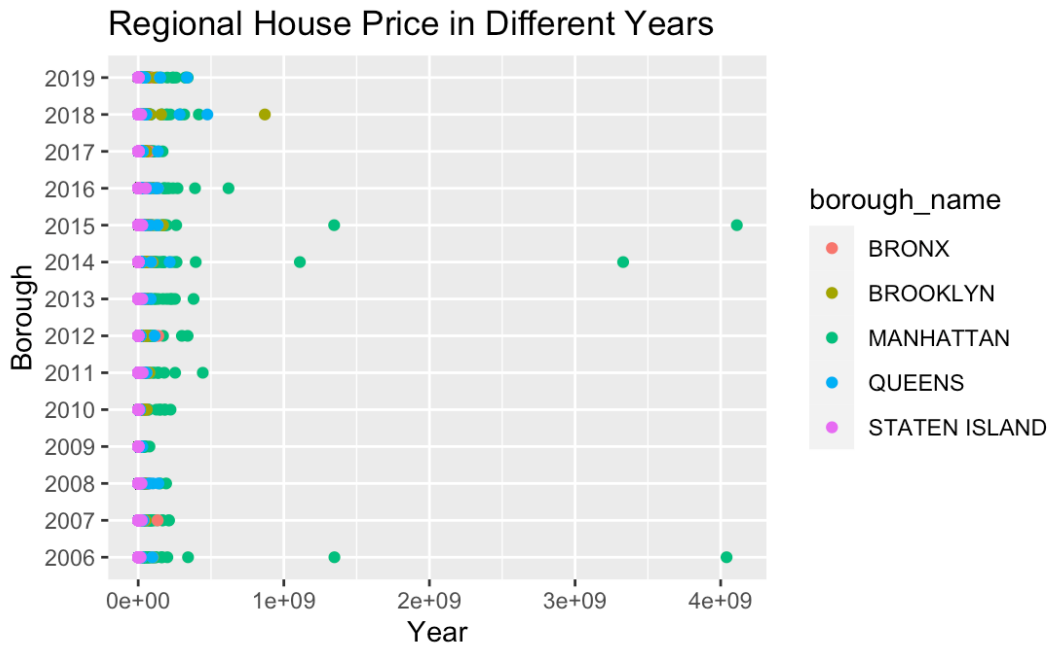
IV. Data Visualization

In this section, we will use plots to show whether there is a relationship between sale price and different variables. Firstly, in our data, the sales date is presented in the form of year-month-day. However, due to the large amount of data, it is impossible to present the housing sales price of each day in the plot, so we processed the data first. We extracted the year corresponding to the sales date as the new column, and then did the data visualization.



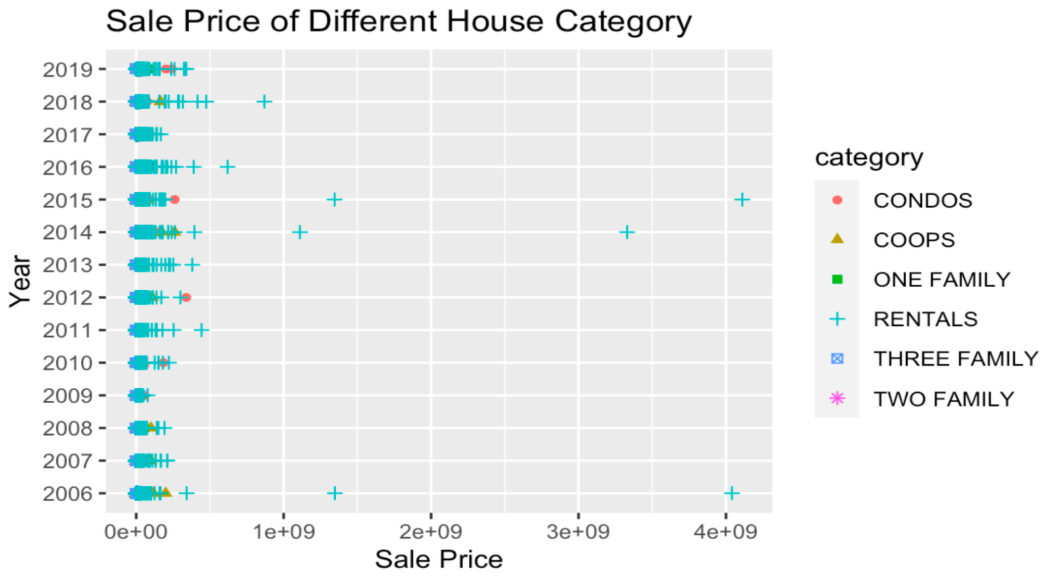
This figure shows that most of the sales price is under 1,000,000,000.

- Borough Name



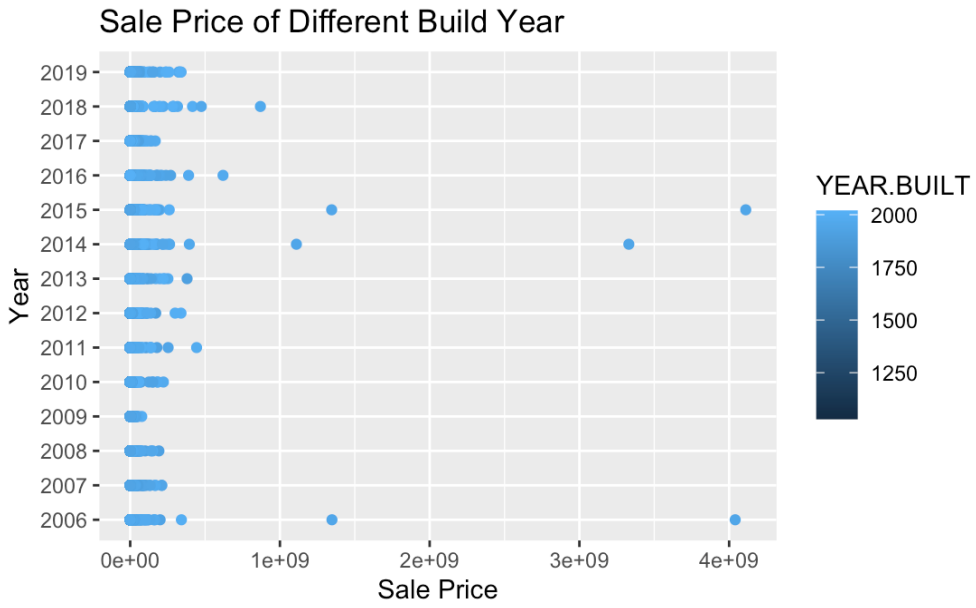
It is obvious to show that the housing price in Manhattan is the highest in most of the year, and the housing price in Staten Island is the lowest in all of the year in this graph. And Brooklyn holds the highest sale price in 2018.

- Category



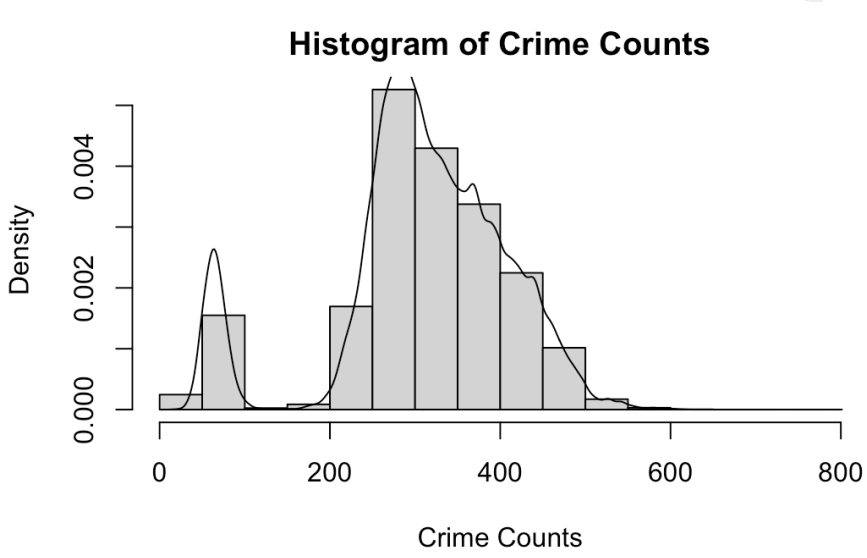
In this plot, the category of three families holds a relatively low price every year, and the most popular type of housing in NYC is rental, but it is difficult to find the relationship between the housing category and sale price because the data is scattered.

- Build Year

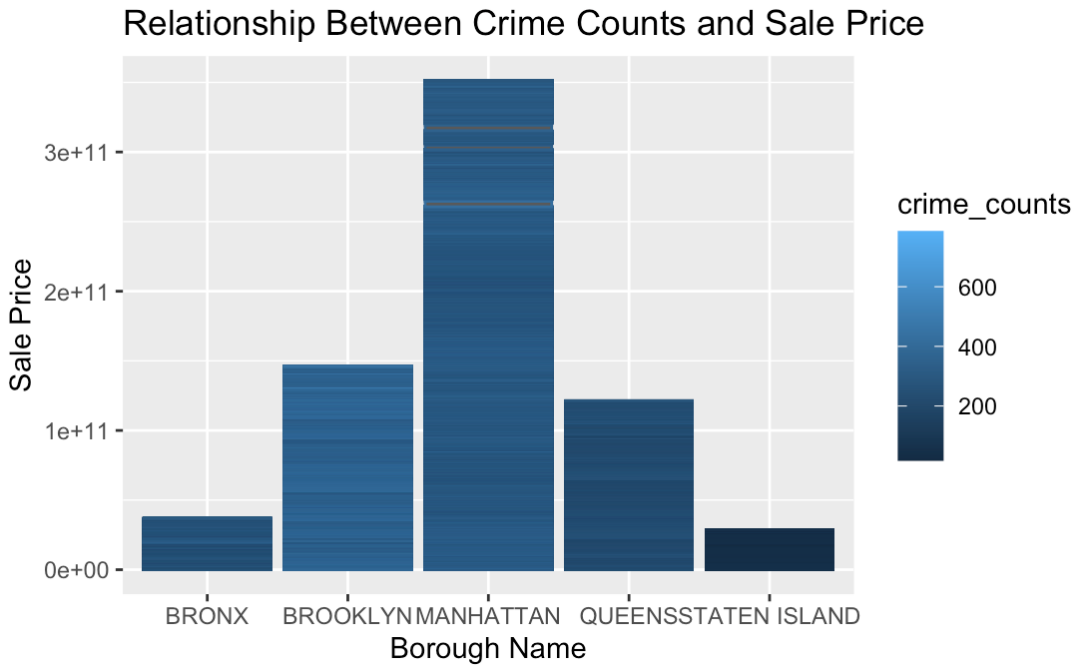


Most of the houses were built between 1750 and 2000, but we cannot see the relationship between the year the houses were built and the price of houses.

- Crime

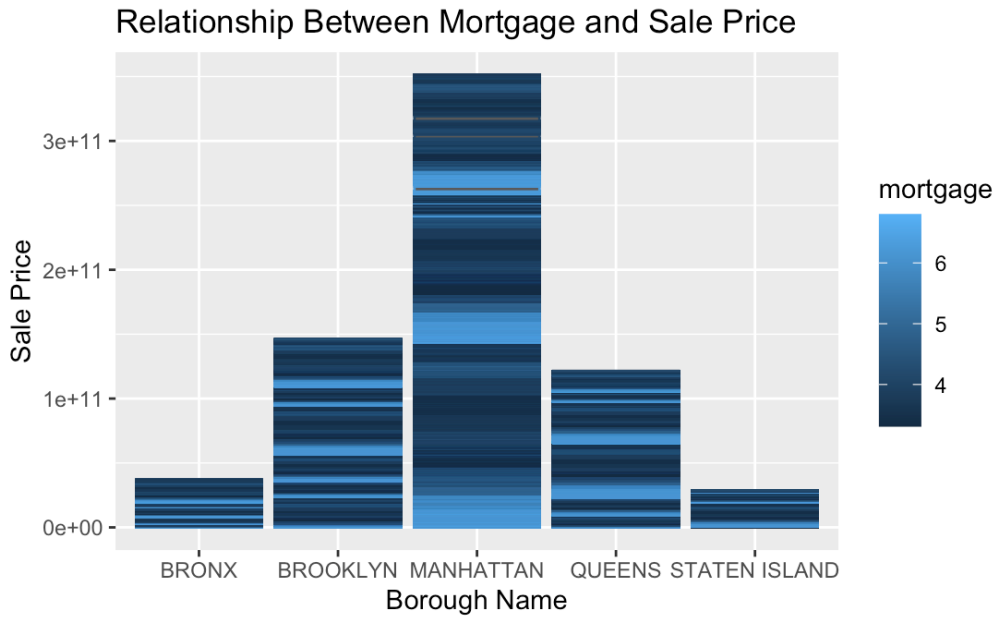


Most of the crime counts are between 200 and 400.



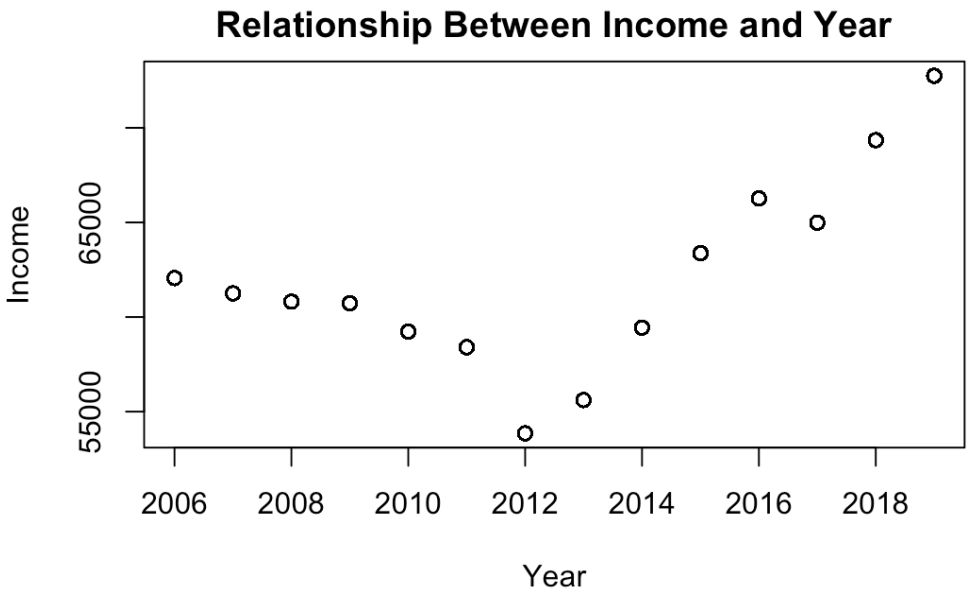
There is no evidence showing that areas with more crime counts have lower sale prices at this stage.

- Mortgage rate



By the plot of the relationship between mortgage and sale price, there is no positive or negative correlation between mortgage and housing price.

- Income



It seems like from 2006 to 2012, income in NYC showed a downward trend; And after 2012, income showed an upward trend.

It's not obvious that we can extrapolate the relationship between income and sale price by this plot.

V. Statistical Model

For the modeling part, we first split our cleaned dataset into two parts. One is the training set which contains data from 2006 to 2018, and the other part is the test set which is data from 2019.

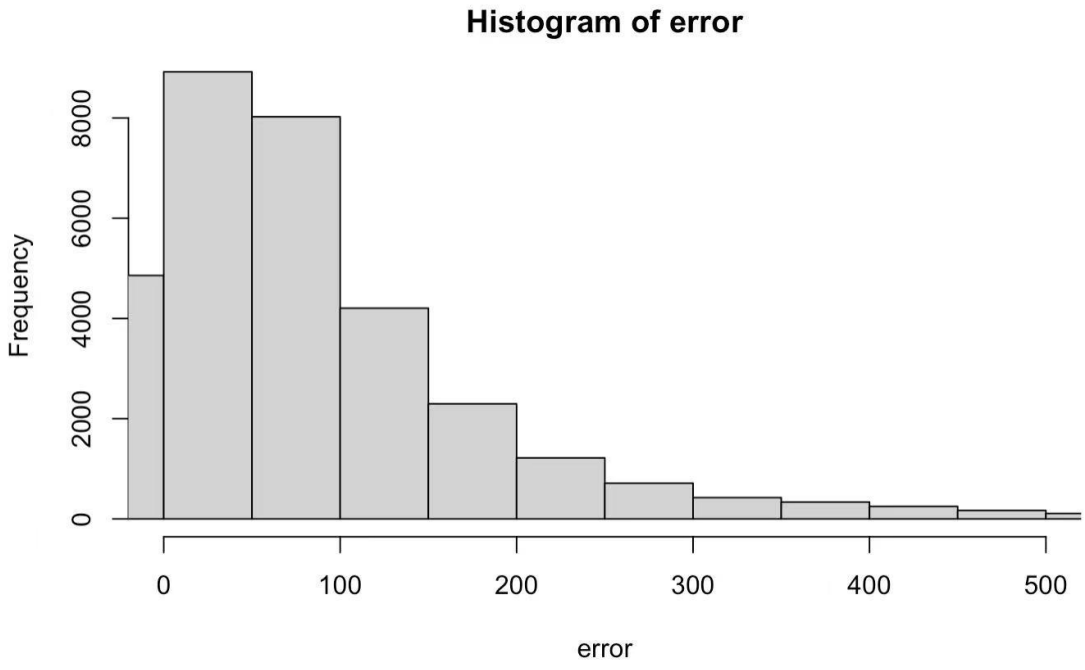
To better understand the model, we fit a linear regression model based on our train set. To sum up, we get a table of summary of the lm model.

```
Call:
lm(formula = SALE.PRICE ~ ., data = train)

Residuals:
       Min         1Q     Median         3Q        Max
  -6459628    -485178     -47973     258500 4104963035

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)               -1.030e+07  9.243e+05 -11.147  < 2e-16 ***
borough_nameBROOKLYN       1.846e+05  6.146e+04   3.003  0.00267 **
borough_nameMANHATTAN      2.010e+06  5.522e+04  36.398  < 2e-16 ***
borough_nameQUEENS         2.106e+05  4.944e+04   4.259 2.06e-05 ***
borough_nameSTATEN ISLAND  2.647e+04  9.917e+04   0.267  0.78950
categoryCOOPS             -3.710e+05  4.023e+04  -9.222  < 2e-16 ***
categoryONE FAMILY         4.094e+05  4.955e+04   8.262  < 2e-16 ***
categoryRENTALS            3.878e+06  6.534e+04  59.351  < 2e-16 ***
categoryTHREE FAMILY       5.500e+05  6.636e+04   8.288  < 2e-16 ***
categoryTWO FAMILY         4.600e+05  4.929e+04   9.333  < 2e-16 ***
YEAR.BUILT                 3.582e+03  4.300e+02   8.330  < 2e-16 ***
SALE.DATE                  1.555e-03  2.800e-04   5.554 2.80e-08 ***
crime_counts               5.805e+02  3.233e+02   1.795  0.07260 .
mortgage                   2.751e+04  3.011e+04   0.914  0.36085
income                     1.620e+01  5.065e+00   3.198  0.00138 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9540000 on 627995 degrees of freedom
Multiple R-squared:  0.01329,   Adjusted R-squared:  0.01327
F-statistic: 604.3 on 14 and 627995 DF,  p-value: < 2.2e-16
```

As we can see, most of the variables have a significant impact in the model except **borough “Staten Island”** and **mortgage**. Most of the variables have 95% significance level but **crime_counts** do not. If alpha is 0.1, it becomes significant. In the t-test hypothesis test, our null hypothesis is that the coefficient of borough “Staten

Island" and mortgage are zero and the alternative hypothesis is otherwise. With 0.05 alpha value, we fail to reject the null hypothesis which means we conclude that borough "Staten Island" and mortgage are insignificant variables. Also, except intercept and Coops house, all other variables have a positive relationship with the sale price.

Meanwhile, to evaluate the model we fitted, we use *predict()* function to check the accuracy of the model using the test data set. We use $\frac{(prediction - true\ value)}{true\ value} * 100$ to get our error rate. We set the threshold as 50 and calculate the accuracy rate as the proportion of error < =50 in our test set which is around 45.74%. Although it is lower than we expect, it can still evaluate the model as a relatively good fit since the dataset is quite big and the house pricing fluctuation is hard to predict.




```r
sum(error<=50)/length(error)
```

```
[1] 0.4574256
```

VI. Diagnostics and Remedial Measures

- Diagnostics Tests

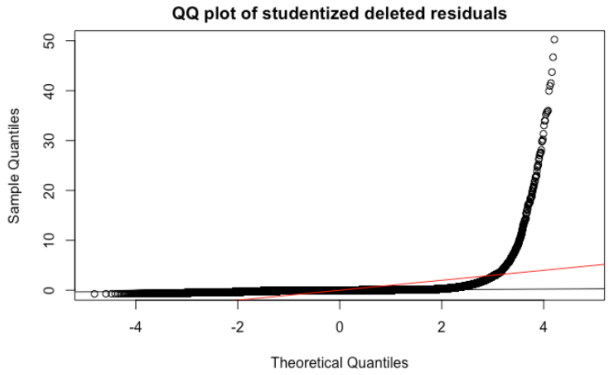
1. Testing for multicollinearity (VIF test)

```
> vif(model)
                  GVIF Df GVIF^(1/(2*Df))
borough_name 11.706181  4        1.360041
category      2.362301  5        1.089767
YEAR.BUILT    1.405123  1        1.185379
SALE.DATE     9.606195  1        3.099386
crime_counts  7.596942  1        2.756255
mortgage      6.868426  1        2.620768
income        3.339755  1        1.827500
```

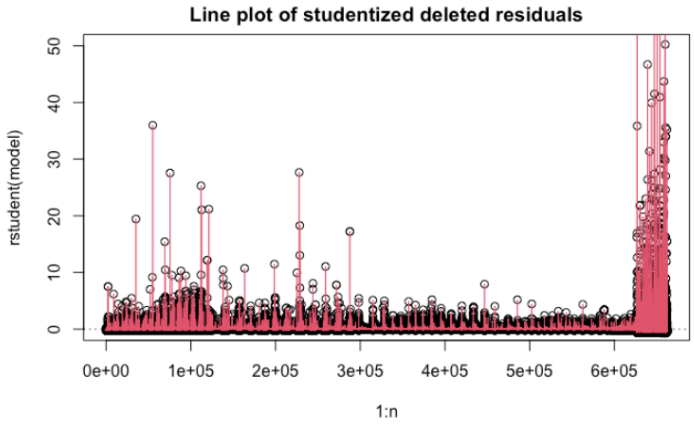
Normally, a GVIF value that exceeds 10 indicates a problematic amount of collinearity and the variable “borough_name” shows multicollinearity. However, we should check the column of GVIF^(1/(2*Df)) for boroughs, which takes into account the number of coefficients in “borough_name”, to make GVIFs comparable across dimensions. This column shows that the value for borough_name is less than 2. Hence, multicollinearity doesn't occur in this regression.

2. Testing for Normality and Equal Variance:

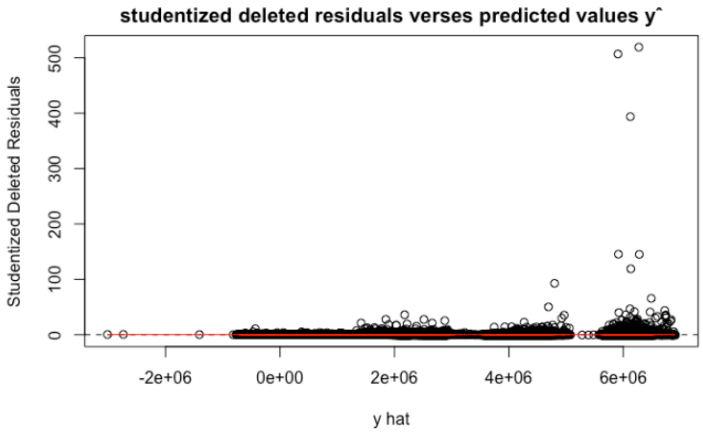
(a) QQ plot of the studentized deleted residuals



(b) line plot of the studentized deleted residuals

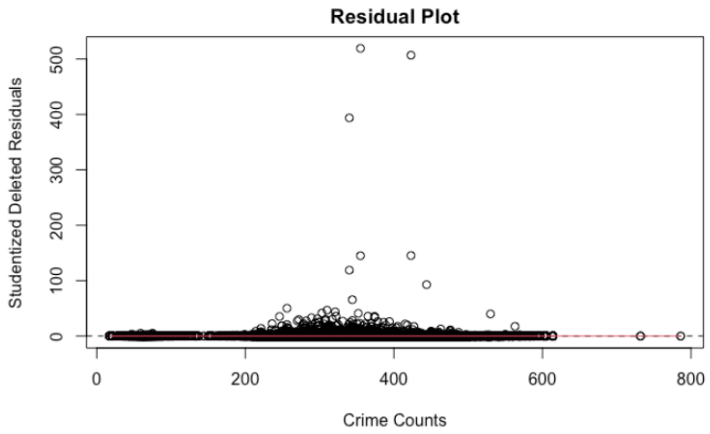


(c) studentized deleted residuals versus predicted values yˆ

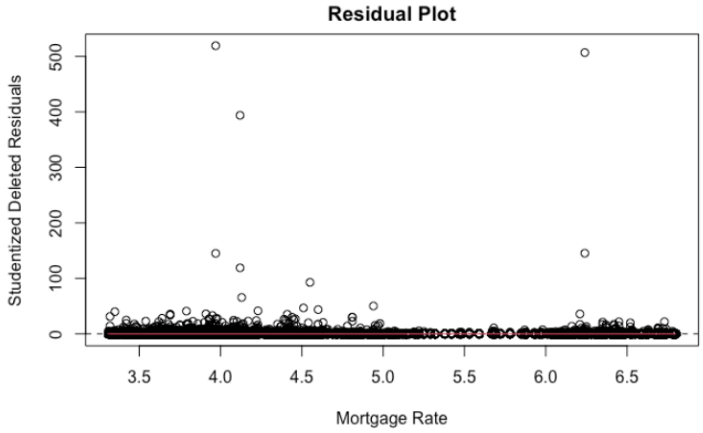


(d) Studentized deleted residuals versus predictor variable x

(i) Studentized deleted residuals versus crime_counts



(ii) Studentized deleted residuals versus mortgage rate



Interpretation from plot (a) - (d):

QQ-plot is highly right-skewed. Line plot, residual plots versus predicted fitted value of y and residual plots versus crime counts and mortgage rates are heavy-tailed. The plots show non-normality of the model due to heavy-tailed distribution, which can produce outlying and influential observations. Although there are not “megaphone”-shaped residual plots, heteroskedasticity often comes with non-normality.

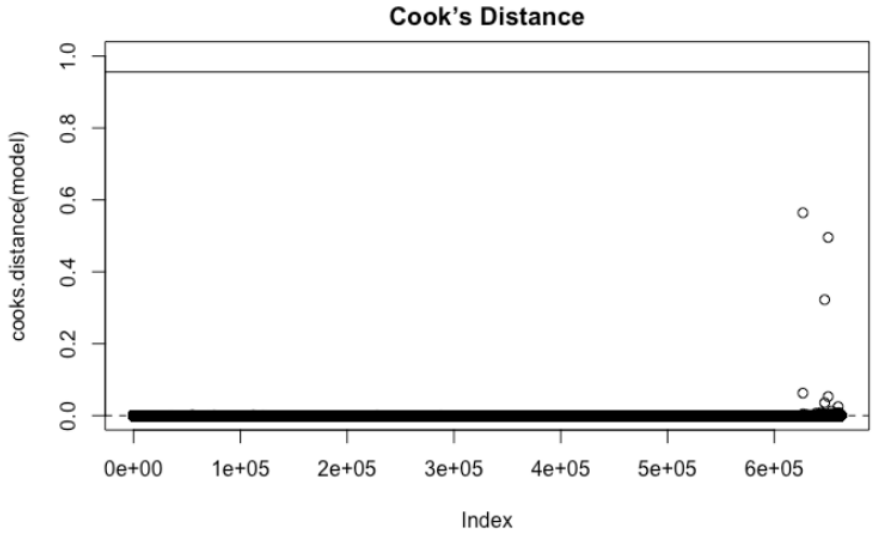
We can also apply Kolmogorov-Smirnov test to see if the data come from a normally distributed population.

```
        Lilliefors (Kolmogorov-Smirnov) normality test

data:  residuals(model)
D = 0.40413, p-value < 2.2e-16
```

Since p-value is smaller than 0.05, we reject the null hypothesis that data come from a normally distributed population.

- Identifying Influential Observations
  1. Cook's Distance

     Cook's distance can be used to assess the influence of data points on all fitted values. Given that the data is heavy-tailed, we can tell from the plot there are several outliers even though they have not passed the 50th percentile of the F-distribution, and we can quantify them by filtering out values greater than 3 times the mean, which are 1430 of data points out of the total data points in our dataset.

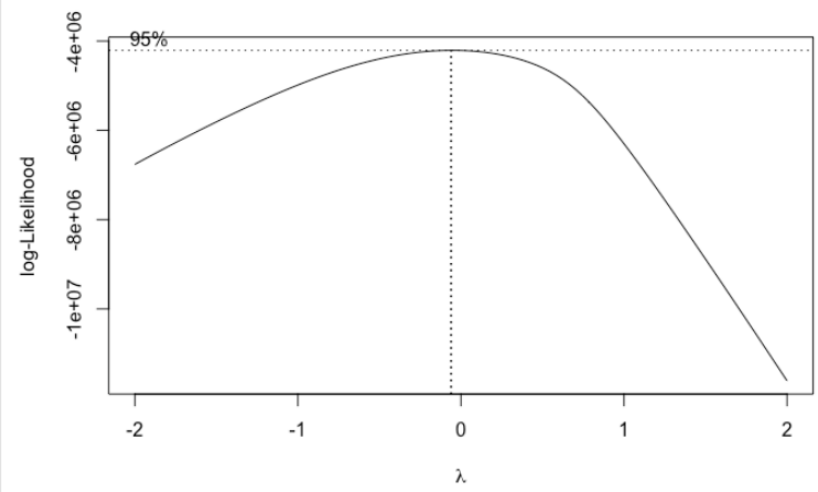


- Remedial Measures

  1. Remedy for non-normality and non-constant variance

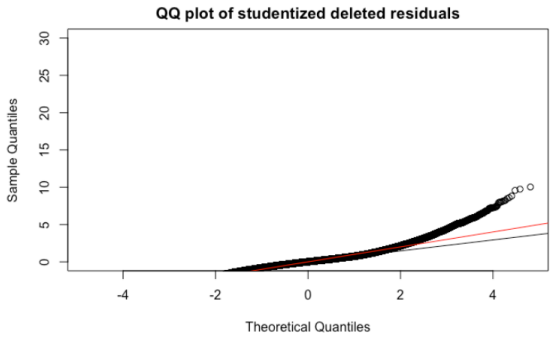
     Box-Cox transformation can be used to transform the original dataset into a normally distributed dataset with equal variance. The lambda calculated is

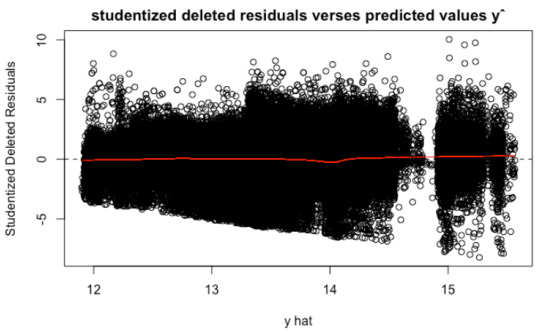
around -0.06, much closer to the value of 0 as shown below. Hence, we should instead use the logarithm of Y in our model.

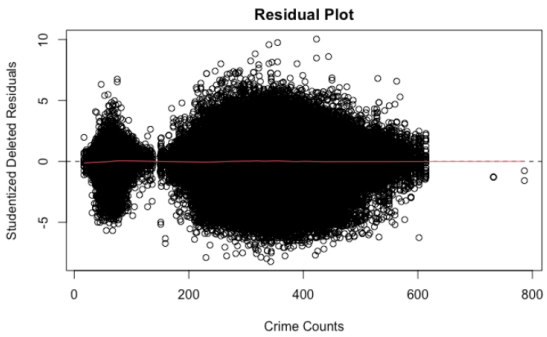


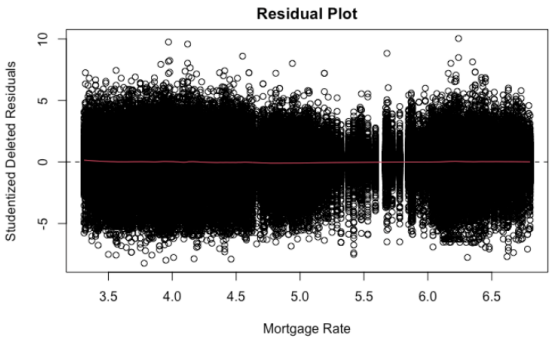
The regression output of the new model shown below has significantly larger R-squared value (0.3975) than the original model, indicating a larger proportion of variance of Y is explained by the independent variables. The QQ-plot and the residual plots versus predicted values and the covariates of the new model are also less skewed compared to the plots from the original model.

```
lm(formula = log(SALE.PRICE) ~ ., data = df3)

Residuals:
    Min      1Q  Median      3Q     Max
-5.8319 -0.3541  0.0120  0.3477  7.1104

Coefficients:
                           Estimate Std. Error  t value Pr(>|t|)
(Intercept)               8.014e+00  6.651e-02  120.492  < 2e-16 ***
borough_nameBROOKLYN      4.445e-01  4.411e-03  100.786  < 2e-16 ***
borough_nameMANHATTAN     1.402e+00  4.002e-03  350.344  < 2e-16 ***
borough_nameQUEENS        2.093e-01  3.561e-03   58.778  < 2e-16 ***
borough_nameSTATEN ISLAND -1.342e-02  7.126e-03   -1.883 0.059717 .
categoryCOOPS             -6.117e-01  2.924e-03 -209.181  < 2e-16 ***
categoryONE FAMILY         1.277e-01  3.563e-03   35.843  < 2e-16 ***
categoryRENTALS            1.017e+00  4.712e-03  215.775  < 2e-16 ***
categoryTHREE FAMILY       3.515e-01  4.775e-03   73.614  < 2e-16 ***
categoryTWO FAMILY         2.335e-01  3.546e-03   65.846  < 2e-16 ***
YEAR.BUILT                 1.062e-03  3.090e-05   34.382  < 2e-16 ***
SALE.DATE                  1.252e-09  2.057e-11   60.875  < 2e-16 ***
crime_counts              -8.052e-05  2.339e-05   -3.442 0.000577 ***
mortgage                   6.934e-02  2.189e-03   31.671  < 2e-16 ***
income                     1.054e-05  3.444e-07   30.621  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7087 on 662147 degrees of freedom
Multiple R-squared:  0.3975,    Adjusted R-squared:  0.3975
F-statistic: 3.12e+04 on 14 and 662147 DF,  p-value: < 2.2e-16
```









2. Remedy for influential observations

After transforming the data, robust regression by minimizing the objective function with respect to beta can be used for fixing our model with influential observations. We use the rlm() to perform robust regression.

```
Call: rlm(formula = log(SALE.PRICE) ~ ., data = df3)
Residuals:
      Min        1Q    Median        3Q       Max
-5.704040 -0.345742  0.006213  0.337935  7.212241

Coefficients:
                          Value     Std. Error t value
(Intercept)                 7.9748    0.0534   149.4378
borough_nameBROOKLYN        0.4649    0.0035   131.3717
borough_nameMANHATTAN       1.3740    0.0032   427.8729
borough_nameQUEENS          0.2425    0.0029    84.8684
borough_nameSTATEN ISLAND   0.0138    0.0057     2.4153
categoryCOOPS              -0.6191    0.0023  -263.8670
categoryONE FAMILY          0.1468    0.0029    51.3331
categoryRENTALS             0.9583    0.0038   253.4699
categoryTHREE FAMILY        0.4165    0.0038   108.7002
categoryTWO FAMILY          0.2794    0.0028    98.1924
YEAR.BUILT                  0.0011    0.0000    44.4687
SALE.DATE                   0.0000    0.0000    72.0339
crime_counts               -0.0001    0.0000    -4.6764
mortgage                    0.0710    0.0018    40.4196
income                      0.0000    0.0000    38.8641

Residual standard error: 0.5064 on 662147 degrees of freedom
```

The Residual standard error of the updated model is 0.5064, much smaller than the original model with RSE of 9378384 and the log-transformed model with RSE of 0.7087, suggesting the robust regression has a better fit to the data.

VII. Updated Model

Based on the previous part, we choose to transform the *y* into *log(y)* in our model to make our data normally distributed.

```
fit <- lm(log(SALE.PRICE)~., data=train)
summary(fit)
```

```
##
## Call:
## lm(formula = log(SALE.PRICE) ~ ., data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.8345 -0.3559  0.0124  0.3494  7.1110
##
## Coefficients:
##                              Estimate Std. Error  t value Pr(>|t|)
## (Intercept)                 8.101e+00  6.868e-02  117.950  < 2e-16 ***
## borough_nameBROOKLYN        4.373e-01  4.567e-03   95.754  < 2e-16 ***
## borough_nameMANHATTAN       1.401e+00  4.103e-03  341.486  < 2e-16 ***
## borough_nameQUEENS          2.065e-01  3.674e-03   56.218  < 2e-16 ***
## borough_nameSTATEN ISLAND -1.077e-02  7.369e-03   -1.462  0.14376
## categoryCOOPS              -6.112e-01  2.989e-03 -204.466  < 2e-16 ***
## categoryONE FAMILY          1.339e-01  3.682e-03   36.369  < 2e-16 ***
## categoryRENTALS             1.015e+00  4.855e-03  209.034  < 2e-16 ***
## categoryTHREE FAMILY        3.493e-01  4.931e-03   70.851  < 2e-16 ***
## categoryTWO FAMILY          2.340e-01  3.663e-03   63.886  < 2e-16 ***
## YEAR.BUILT                  1.003e-03  3.195e-05   31.377  < 2e-16 ***
## SALE.DATE                   1.265e-09  2.081e-11   60.808  < 2e-16 ***
## crime_counts               -7.171e-05  2.402e-05   -2.985  0.00284 **
## mortgage                    7.052e-02  2.237e-03   31.525  < 2e-16 ***
## income                      1.063e-05  3.763e-07   28.260  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7089 on 627995 degrees of freedom
## Multiple R-squared:  0.394,  Adjusted R-squared:  0.394
## F-statistic: 2.916e+04 on 14 and 627995 DF,  p-value: < 2.2e-16
```

When we observe the p-value column of all variables, we find out that **mortgage** becomes statistically significant. Besides, almost all variables are significant except for the STATEN_ISLAND category in borough with an alpha of 0.05. It might be due to the smaller proportion of data belonging to **borough_nameSTATEN ISLAND** compared to other boroughs. Our verification of this speculation is shown below:

```
sum(df3$borough_name=="STATEN ISLAND")
```

```
## [1] 60131
```

```
sum(df3$borough_name=="STATEN ISLAND")/nrow(df3)
```

```
## [1] 0.0908101
```

Only 9.08% of the data is related to **borough_nameSTATEN ISLAND**, which confirms our speculation. However, since there might be a proportion of people in New York considering living in Staten Island, we kept this factor in our model.

Besides, if we take alpha equals to 0.001, **crime_counts** also becomes insignificant. In the t-test hypothesis test, our null hypothesis is that the coefficient of **crime_counts** is zero and the alternative hypothesis is otherwise. With an alpha value of 0.001, p-value of crime_counts is 0.00284, which is smaller than 0.001. Hence, we fail to reject the null hypothesis, meaning that **crime_counts** is an insignificant variable.

In the next section, we did the variable selection to help us figure out which variables in our model are necessary and which are not.

VIII. Variable Selection

Until now, the linear regression model we have is:

$$log(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \ldots\ldots + \beta_{14} x_{14}$$

There are a total of seven variables, but two of them are categorical variables and others are continuous variables. Hence, taking into consideration the categorical variable, there are 14 variables in our model. We intend to apply variable selection methods in order to help us construct a more efficient and simplified model.
We chose to use forward and backward stepwise method with AIC criteria.

We started with backward elimination:

```
ols_step_backward_aic(fit)
```

```
## [1] "No variables have been removed from the model."
```

```
stepAIC(fit,direction = "backward")
```

```
## Start:  AIC=-432163.4
## log(SALE.PRICE) ~ borough_name + category + YEAR.BUILT + SALE.DATE +
##     crime_counts + mortgage + income
##
##                Df Sum of Sq    RSS     AIC
## <none>                      315564 -432163
## - crime_counts  1         4 315568 -432156
## - income        1       401 315965 -431367
## - YEAR.BUILT    1       495 316059 -431182
## - mortgage      1       499 316063 -431172
## - SALE.DATE     1      1858 317422 -428478
## - category      5    100162 415726 -259053
## - borough_name  4    101205 416769 -257477
```

```
##
## Call:
## lm(formula = log(SALE.PRICE) ~ borough_name + category + YEAR.BUILT +
##     SALE.DATE + crime_counts + mortgage + income, data = train)
##
## Coefficients:
##               (Intercept)       borough_nameBROOKLYN
##                 8.101e+00                  4.373e-01
##     borough_nameMANHATTAN         borough_nameQUEENS
##                 1.401e+00                  2.065e-01
## borough_nameSTATEN ISLAND              categoryCOOPS
##                -1.077e-02                 -6.112e-01
##        categoryONE FAMILY            categoryRENTALS
##                 1.339e-01                  1.015e+00
##      categoryTHREE FAMILY         categoryTWO FAMILY
##                 3.493e-01                  2.340e-01
##                YEAR.BUILT                  SALE.DATE
##                 1.003e-03                  1.265e-09
##              crime_counts                   mortgage
##                -7.171e-05                  7.052e-02
##                    income
##                 1.063e-05
```

Then, we also applied forward selection:

```
fit.none <- lm(log(SALE.PRICE)~1, data=train)
stepAIC(fit.none,scope = list(lower = fit.none, upper = fit),direction = "forward")
```

```
## Start:  AIC=-117667.3
## log(SALE.PRICE) ~ 1
##
##                Df Sum of Sq    RSS     AIC
## + borough_name  4     94323 426384 -243165
## + category      5     86705 434002 -232042
## + crime_counts  1     12365 508342 -132759
## + SALE.DATE     1      7001 513706 -126167
## + income        1      4982 515725 -123703
## + mortgage      1      2036 518671 -120126
## + YEAR.BUILT    1         6 520701 -117673
## <none>                      520707 -117667
##
## Step:  AIC=-243164.7
## log(SALE.PRICE) ~ borough_name
##
##                Df Sum of Sq    RSS     AIC
## + category      5     98224 328161 -407592
## + income        1      7493 418892 -254297
## + SALE.DATE     1      7466 418918 -254257
## + mortgage      1      1721 424664 -245702
## + crime_counts  1       798 425587 -244339
## + YEAR.BUILT    1         3 426382 -243167
## <none>                      426384 -243165
##
```

```
## Step:  AIC=-407591.7
## log(SALE.PRICE) ~ borough_name + category
##
##                Df Sum of Sq    RSS     AIC
## + SALE.DATE     1    8500.7 319660 -424072
## + income        1    8158.2 320002 -423400
## + mortgage      1    1848.3 326312 -411137
## + crime_counts  1    1077.0 327084 -409654
## + YEAR.BUILT    1     494.0 327667 -408536
## <none>                     328161 -407592
##
## Step:  AIC=-424072.2
## log(SALE.PRICE) ~ borough_name + category + SALE.DATE
##
##                Df Sum of Sq    RSS     AIC
## + mortgage      1    3172.9 316487 -430335
## + income        1    3105.7 316554 -430202
## + YEAR.BUILT    1     530.0 319130 -425112
## <none>                     319660 -424072
## + crime_counts  1       0.3 319660 -424071
##
## Step:  AIC=-430334.8
## log(SALE.PRICE) ~ borough_name + category + SALE.DATE + mortgage
##
##                Df Sum of Sq    RSS     AIC
## + YEAR.BUILT    1    507.48 315980 -431341
## + income        1    423.68 316063 -431174
## + crime_counts  1     15.13 316472 -430363
## <none>                     316487 -430335
##
## Step:  AIC=-431340.6
## log(SALE.PRICE) ~ borough_name + category + SALE.DATE + mortgage +
##     YEAR.BUILT
##
```

```
##                Df Sum of Sq    RSS     AIC
## + income        1    411.21 315568 -432156
## + crime_counts  1     14.39 315965 -431367
## <none>                      315980 -431341
##
## Step:  AIC=-432156.5
## log(SALE.PRICE) ~ borough_name + category + SALE.DATE + mortgage +
##     YEAR.BUILT + income
##
##                Df Sum of Sq    RSS     AIC
## + crime_counts  1    4.4773 315564 -432163
## <none>                      315568 -432156
##
## Step:  AIC=-432163.4
## log(SALE.PRICE) ~ borough_name + category + SALE.DATE + mortgage +
##     YEAR.BUILT + income + crime_counts
```

```
##
## Call:
## lm(formula = log(SALE.PRICE) ~ borough_name + category + SALE.DATE +
##     mortgage + YEAR.BUILT + income + crime_counts, data = train)
##
## Coefficients:
##               (Intercept)       borough_nameBROOKLYN
##                 8.101e+00                  4.373e-01
##     borough_nameMANHATTAN         borough_nameQUEENS
##                 1.401e+00                  2.065e-01
## borough_nameSTATEN ISLAND              categoryCOOPS
##                -1.077e-02                 -6.112e-01
##        categoryONE FAMILY            categoryRENTALS
##                 1.339e-01                  1.015e+00
##      categoryTHREE FAMILY         categoryTWO FAMILY
##                 3.493e-01                  2.340e-01
##                 SALE.DATE                   mortgage
##                 1.265e-09                  7.052e-02
##                YEAR.BUILT                     income
##                 1.003e-03                  1.063e-05
##              crime_counts
##                -7.171e-05
```

The results from both forward and backward stepwise methods with AIC criteria indicate that we should keep all the variables.

IX. Conclusion

The final model:

```
fit <- lm(log(SALE.PRICE)~., data=train)
summary(fit)
```

```
##
## Call:
## lm(formula = log(SALE.PRICE) ~ ., data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.8345 -0.3559  0.0124  0.3494  7.1110
##
## Coefficients:
##                               Estimate Std. Error  t value Pr(>|t|)
## (Intercept)                  8.101e+00  6.868e-02  117.950  < 2e-16 ***
## borough_nameBROOKLYN         4.373e-01  4.567e-03   95.754  < 2e-16 ***
## borough_nameMANHATTAN        1.401e+00  4.103e-03  341.486  < 2e-16 ***
## borough_nameQUEENS           2.065e-01  3.674e-03   56.218  < 2e-16 ***
## borough_nameSTATEN ISLAND -1.077e-02  7.369e-03   -1.462  0.14376
## categoryCOOPS             -6.112e-01  2.989e-03 -204.466  < 2e-16 ***
## categoryONE FAMILY           1.339e-01  3.682e-03   36.369  < 2e-16 ***
## categoryRENTALS              1.015e+00  4.855e-03  209.034  < 2e-16 ***
## categoryTHREE FAMILY         3.493e-01  4.931e-03   70.851  < 2e-16 ***
## categoryTWO FAMILY           2.340e-01  3.663e-03   63.886  < 2e-16 ***
## YEAR.BUILT                   1.003e-03  3.195e-05   31.377  < 2e-16 ***
## SALE.DATE                    1.265e-09  2.081e-11   60.808  < 2e-16 ***
## crime_counts                -7.171e-05  2.402e-05   -2.985  0.00284 **
## mortgage                     7.052e-02  2.237e-03   31.525  < 2e-16 ***
## income                       1.063e-05  3.763e-07   28.260  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7089 on 627995 degrees of freedom
## Multiple R-squared:  0.394,  Adjusted R-squared:  0.394
## F-statistic: 2.916e+04 on 14 and 627995 DF,  p-value: < 2.2e-16
```

The regression output we computed from our final model matches our expectations of the effect the variables have on the logarithm of **SALE.PRICE**.

Borough_name:

For all the boroughs we are considering, living in Manhattan has the highest positive impact on the house price: there will be a 140.1% increase in **SALE.PRICE** if we increase **borough_nameMANHATTAN** by one unit. Living in Brooklyn increases **SALE.PRICE** by 43.71% and living in Queens increases **SALE.PRICE** by 20.65%, the lowest of the three statistically significant borough factors. (**borough_nameSTATENISLAND** is not statistically significant at 95% confidence level as indicated by the p-value).

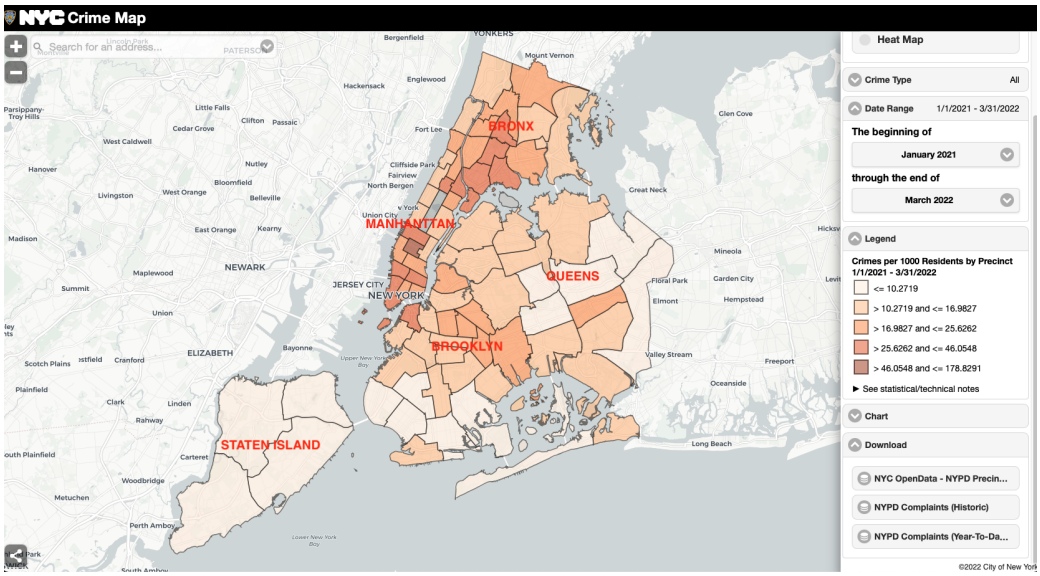
Category:

The categories of houses we include in our analysis all have positive effect on the house price with a descending order of rentals (101.5%), three-family (34.93%), two-family (23.4%) and one-family (13.39%), except for Co-op apartments (-61.12%).

Continuous variables:

YEAR.BUILT and **SALE.DATE** from our housing dataset don't have large impacts on the logarithm of **SALE.PRICE**. There is only 0.1% increase in the house price if we increase the build year by one unit and almost negligible increase in the house price with a one unit increase in the sale date of the property.

The variable **crime_counts** has a negative effect on the houses' sale prices. Boroughs with a higher number of crimes will likely have a decrease in the regional house prices. However, the coefficient of **crime-counts** is relatively smaller compared to those of other variables, which can be explained by New York's Crime Map as shown below:

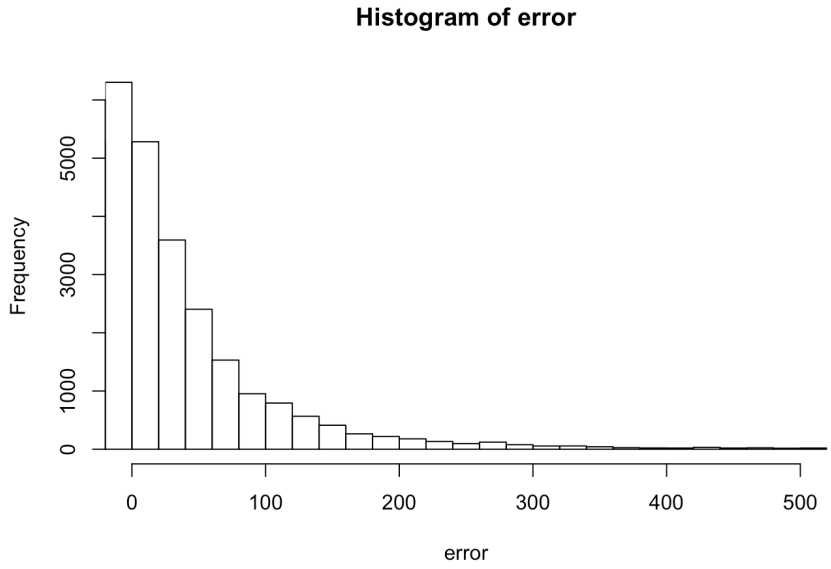


From the map, we can tell that Manhattan has the highest crime rates, Bronx is the second highest, while Brooklyn and Queens have similar crime rates, and Staten Island has the lowest crime rate. Using our previous result, Manhattan has the highest house price and highest crime rate at the same time that might lead to conflicting results.

Mortgage rate and median income of households in New York positively affects the logarithm of the house prices. A one unit increase in **mortgage** will increase **SALE.PRICE** by 0.7%. This is reasonable because the higher rates add hundreds of dollars or even more to the mortgage payments on houses. The effect **income**has on the logarithm of **SALE.PRICE** is trivial but positive. As median income rises, more people are considering buying houses, thus driving up the demand and leading to increase in houses' sale prices.

We then used the training set (data from 2003 to 2018) and test set (data in 2019) to validate our model. We utilized the formula: $\frac{(prediction - true\ value)}{true\ value} * 100$ to calculate our errors. We want to compare the size of bias between our prediction and the true value of house sale prices. From the below histogram, we can tell that the majority of the errors are between 0% to 100%. The proportion of error below 50% is 78.8%. This number has greatly improved compared to our first model with a 45.74% of error below 50%. Therefore, transforming the *sale price* into logarithm can significantly improve our model.

```
target <- df3[df3$time==2019,8]
error <- ((exp(predict(fit,test))-target)/target)*100
hist(error,freq=T,breaks=1000,xlim = c(0,500))
```




```
sum(error<=50)/length(error)
```

```
## [1] 0.7880066
```

Admittedly, there still exist errors in our prediction. The fact that we have many dummy variables in our model would reduce the accuracy in measuring the effects the variables have on house sale prices in New York. There might be other factors affecting house prices that are not included in this dataset; for example, education resources near the houses, convenience level, and view of the houses, etc. Apart from that, we cannot anticipate any force majeure such as the COVID situation, which greatly impacts the entire housing industry.

X. References

- House price dataset: https://www.kaggle.com/datasets/kaijiechen/nyc-housing-data-20032019?resource=download
- Real median household Income: https://fred.stlouisfed.org/series/MEHOINUSNYA672N
- Crime: https://www.kaggle.com/code/brunacmendes/new-york-crime-analysis/data
- Mortgage: https://fred.stlouisfed.org/series/MORTGAGE30US