Ravi Gupta

1. Executive Summary

This report summarizes the statistical modeling and analysis results of market value per sq. ft. modeling for Condominiums in New York across the 5 boroughs: Manhattan, Queens, Bronx, Brooklyn and Staten Island. The purpose of this report is to document the design and implementation of the data model with linear regression technique used for this statistical analysis to find out the factors that affects the market value per sq. ft. of condominiums.

The initial dataset for designing our model is taken from NYC open data. The dataset is prepared by merging 5 different datasets from each borough for the year 2011-2012. The dataset contained 2612 observations from 5 boroughs with 52 variables. After data cleaning and adding two extra calculated attributes, WalkScore and Zone, the final data set contains 2552 observations across 24 variables. 4 independent variables and one control variable (neighborhood) have been chosen from this final dataset to model the dependent variable, **Market value per sq. ft**., as per our research question.

We had similar rental pricing data with the dataset which we did not use in our model. The reason for not using them in our model is that they are the similar priced properties and would deliver a very high adjusted R-squared value overshadowing the effect of other independent variables. Further, comparing price of a property with a price of a similar housing does not bring out any rich analysis. That is why we mainly focused on other independent variables like building classification, neighborhood and gross area and discarded the pricing related variables for 'Rental1' and 'Rental2' data.

The density plot of the dependent variable is bi-modal. We tried to transform it into normally distributed variable by applying log, square and other transformations but we got the best distribution with its original state. The univariate and bivariate analysis has been carried out in details. The univariate analysis explains the distribution of each variable independently and the bivariate analysis explains the relationship between dependent and independent variables i.e. the distribution across the data and the effects of variables on the market value per sq. ft. of a condominium in New York has been presented.

After performing univariate analysis, few skewed variables like walk score and year built were categorized into factors. For the ease of interpretation, the variable Neighborhood was converted into Zone (North, South, East, West, and Central), by manually obtaining the relative position of each property within a Borough. This reduced the level of this variable to 5. These categorical variables and other dependent variables were used in multiple linear regression modeling, to get a final model with an adjusted R-square value of 67.77%.

The final model indicates that the dependent variables Borough, Building Classification, Year Built, Gross Sq. Ft, Zone, and walk score categories are the determining factors of the Condominium Market Value per Sq. Ft. The detailed findings are discussed further in this report.

2. Introduction

About our dataset

Condominiums and cooperatives are valued as if they were residential rental apartments by the Department of Finance in NYC. Income information from similar rental properties is applied to determine value with respect to their distance from the Condominiums, as well as in their respective neighborhood. Properties are selected based on a combination of factors such as land location (distance from the property), income levels, year of built (building age) and construction (area in sq. ft.) and exemptions and subsidies.

Data Provided by Department of Finance (DOF)

Dataset Owner: NYC OpenData

Characteristics

The initial dataset contained 2612 observations from 5 boroughs with 52 variables. For analysis, we used only variables that are related to the condominium properties. The reason for not using comparable property in our model is that they would deliver a very high adjusted R-squared value overshadowing the effect of other independent variables. Further, comparing price with the price of a similar housing does not bring out any rich analysis. That is why we mainly focused on other independent variables like building classification, neighborhood, and gross area and discarded the pricing related variables in Rental1 and Rental2 data.

The table below describes the final data used for analysis, after adding categorical variables (Explained in detail in the Univariate Analysis section). Also, outliers were removed from each dependent variable using the identify function. Final dataset contains 2552 Rows with 24 Columns.

Variable	**Data Type**	**Description**
BBL (Borough Block Lot)	Integer	the number system used to identify each unit of real estate in New York City
Condominiums Address	Character	Address of the given property
Postcode	Integer	Zip code of an area
Latitude	Numeric	Horizontal coordinates of a property
Longitude	Numeric	Vertical coordinates of a property
Council District	Integer	Area roughly equivalent to a neighborhood for analyzing populations.
Condominiums Boroughs	Factor	Building Identification Number
Condominiums Walkscore	Integer	Boro Block Lot

Zone	Factor	Manually transformed Neighborhoods into 5 categories (North, South, East, West and Center) w.r.t true north
Condominiums Neighborhood	Factor	Localities around the properties in different communities
Condominiums Building Classification	Factor	Building types- different types of residential units
Condominiums Total Units	Integer	Total number of residential units in the neighborhood
Condominiums Year Built	Integer	Year in which the building was constructed
Condominiums Gross Sq. Ft	Integer	Total square footage of a building.
Condominiums Estimated Gross Income	Integer	Estimated value of a piece of investment property
Condominiums Gross Income per Sq. Ft	Integer	Total value of a piece of investment property per square feet
Condominiums Estimated Expense	Integer	Estimated costs incurred to own the property.
Condominiums Expense per Sq. Ft	Integer	Expenses costs to own the property per square feet.
Condominiums Net Operating Income	Integer	Net income of property after deducting expenses from the Income.
Condominiums Full Market Value	Integer	Probable price that a property will bring in an open market transaction.
Condominiums Market Value per Sq. Ft	Integer	Price of a property per square feet in an open market transaction.
YearBuiltCategories	Factor	Year Built transformed into 2 categories
WalkscoreCategories	Factor	Walkscore transformed into 3 categories
Borough	Integer	Identification no. of borough

R2-CONDOMINIUM - RESIDENTIAL UNIT IN WALK-UP BUILDINGS

The R2 zone is a low density multi-dwelling zone. It allows approximately 21.8 dwelling units per acre. Density may be as high as 32 units per acre if amenity bonus provisions are used. Allowed housing is characterized by one to three story buildings, but at a slightly larger amount of building coverage than the R3 zone. The major types of new development will be duplexes, townhouses, rowhouses and garden apartments. These housing types are intended to be compatible with adjacent houses. Generally, R2 zoning will be applied near Major City Traffic Streets, Neighborhood Collector and District Collector streets, and local streets adjacent to commercial areas and transit streets

R4-CONDOMINIUM - RESIDENTIAL UNIT IN ELEVATOR BLDG

The R4 building class comprises residential units that are common in high buildings. That's why majority of these units can be found in Manhattan (61%). It is followed by Brooklyn (15%) and Queens (13%).

RR-CONDOMINIUM - CONDO RENTALS

The RR building class comprises rental condominiums. They can most frequently be found in Manhattan, in Midtown / Midtown South, Clinton, Battery Park City / Lower Manhattan and Lincoln Square, but also in the Bronx in Melrose South / Mott Haven North.

R9-CONDOMINIUM - CO-OP WITHIN A CONDOMINIUM
The R9 property class comprises of condominium co-operatives.

3. Steps taken to clean the data

Data Transformations

1. Expanded our datasets to generate more variables which will be required in statistical modelling. Initially we only had data for one borough, Manhattan. Then we gathered data for other 4 boroughs, Queens, Bronx, Brooklyn, Staten Island.
2. After merging datasets of other 4 boroughs with Manhattan's data, there were a lot of missing values for postal codes, which hindered our analysis. We searched for longitude and latitude values from previous datasets of 2010-2011 and filled in the missing postcodes.
3. Categorized the numeric data of an independent variable "Year Built", to predict the age of condominium. Hence, we categorized the data into factors of levels 'Before_2000' & 'After_2000'.
4. In a same fashion, we segmented the 5 boroughs into cardinal directions termed as 'Zone'. To analysis the market value of Condominium's properties zone wise (North, West, East, South, Central).
5. We identified the outliers laying in our dataset and removed them for different variables by plotting them on the graphs.
6. To attain normal distribution, we transformed highly skewed variables using log function.

7. We added ‘WalkScore’ for each postal code through WalkScore API packages installed in R. WalkScore helps you find the walkable place to live. Further, we divided the ratings into factors of 3 levels ‘Walker’s Paradise’, ’Car-dependent’ & ’Somewhat Walkable.

Data Cleaning

1) Dropped the column number 25 to 70 as it was not used in our Model
2) Removed outliers by plotting each numerical variable
3) Removed the rows containing entirely NAs

4. Research Question

What factors affect the market price per sq. ft. of condominiums in New York state across the 5 boroughs?

According to the “price-per-square-foot-comparison” ‘s analysis we got to know that this comparison is being used as PPSF comparison tool to set the norms of market value per square feet. Keeping in mind the Condominiums characteristics and Property’s value as in which year what type of buildings were made. Accordingly, the market value per square feet is set. New York city is divided into 5 main boroughs, so we have further divided them into zones. Through this we the market value of the properties in individual directions North, South, East & West.

5. Hypothesis

1. Condominium’s market value per sq. ft. increases with increase in Walk score
2. Condominium’s market value per sq. ft. is relatively higher at center of a borough area.
3. Condominium’s market value per sq. ft. is comparatively lower for low-rise condominiums.
4. Condominium’s market value per sq. ft. is higher for newer condominiums.

6. Limitations

1. We required all the values of Postal codes and their coordinates to incorporate the Walkscore API as R packages.

2. We did not have any variables like reviews of the property or crime heat of any borough.

3. Condominium’s market value was given as full market value as well as per sq. ft. too. So, we used market value per sq. ft. to get the normal distribution.

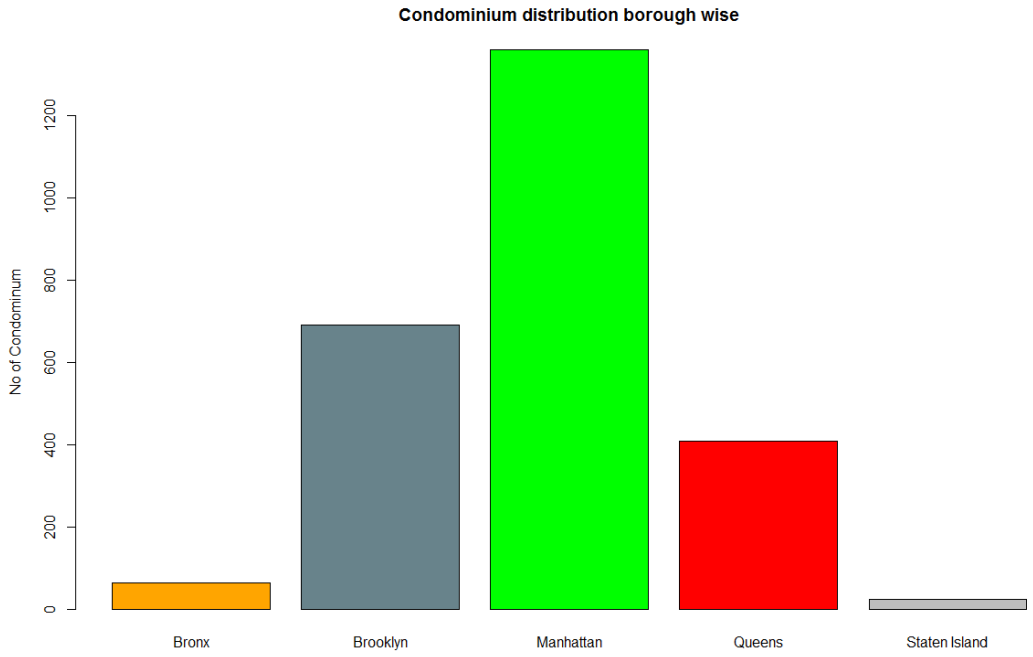
4. We divided the five main boroughs into five zones as it was limiting the analysis of market value per sq. ft. only to the boroughs. Now, we can get the market value of condominiums zones wise.

5. Similarly, time of year in which condos were built was categorized further as time after 2000 & before 2000. Also, because the mean values were approximately coming equal for before 2000 & after it.

6. Details about amenities like parking, swimming pool, gym etc. would have been helpful in alalysis.

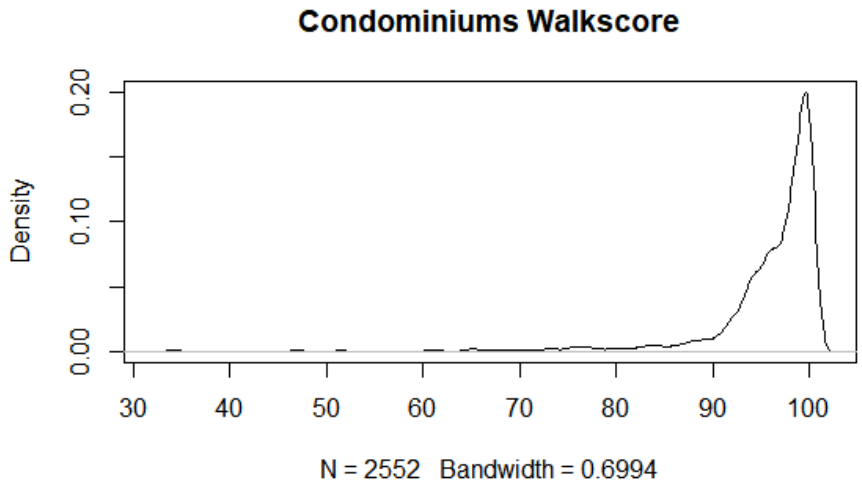
7. Univariate Analysis

Condominium Distribution Borough Wise:

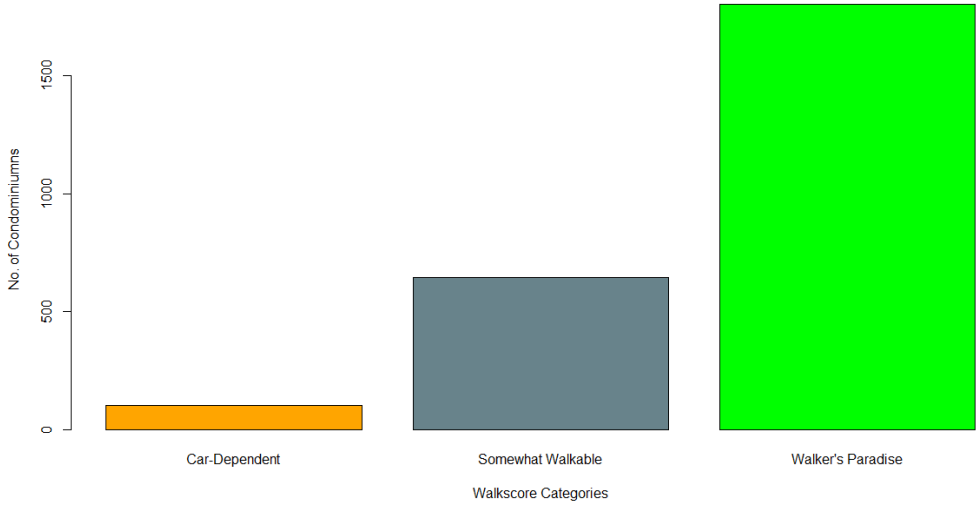


The dataset comprises of condominium property stats from all the 5 boroughs of New York. We got Bronx's condos as 66, Brooklyn's as 691, Manhattan with 1360 condominium units, Queens as 409 and Staten Island as 26. Manhattan is the center of New York.

Number of Condos with respect to WalkScore Categories

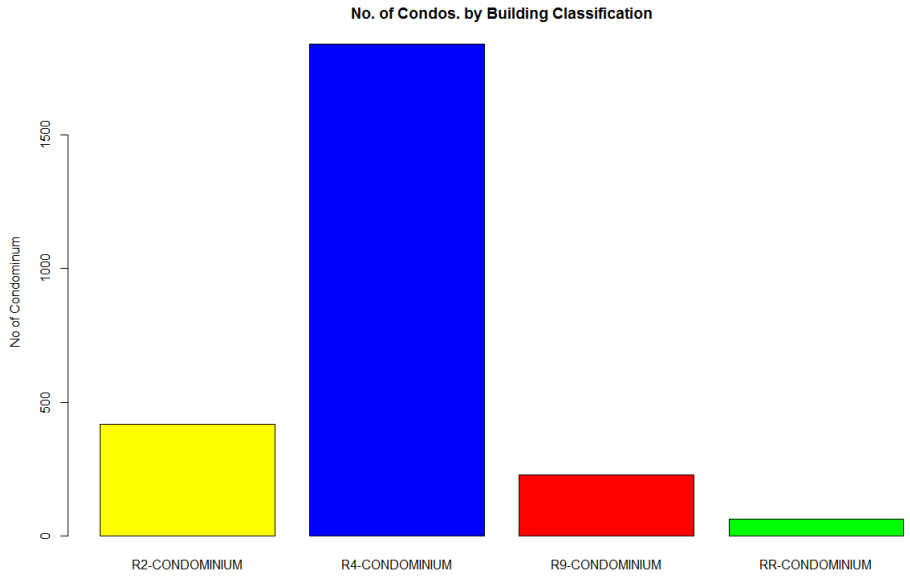


Even after performing log function on 'Condominiums WalkScore, the variable was highly skewed to the left. Hence, we converted the WalkScore to a categorical variable with three categories as 'Car-dependent', 'Somewhat walkable', 'Walker's Paradise' for our analysis. Below is the description of the new categorical variable 'WalkScore Categories'.



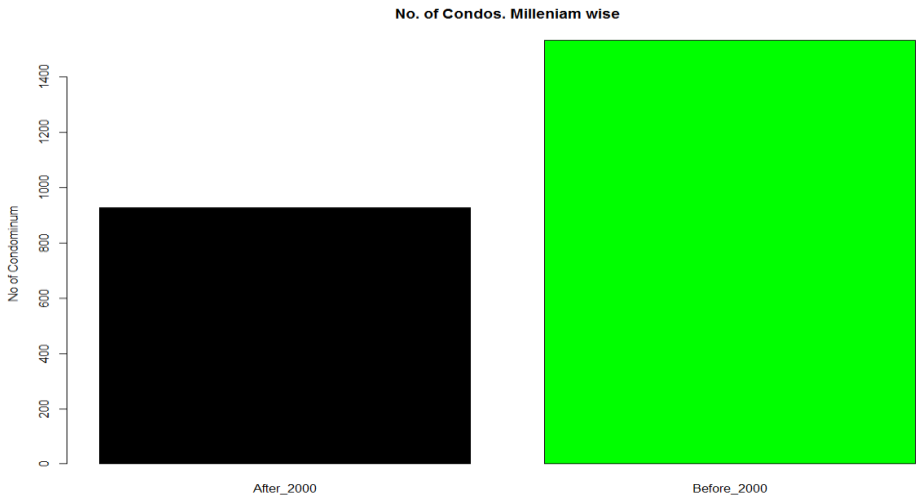
Number of Condominium units in each WalkScore categories like "Walker's Paradise" have maximum 1802 condominiums as WalkScore ratings of walker's paradise lies between the range of 96 to 100, "Somewhat Walkable" have 645 condo units & it lies in the range of 80 to 95. "Car dependent" have only 101 condominium units in the range of WalkScore ratings as less than 80.

Number of Condominiums by Building Classification



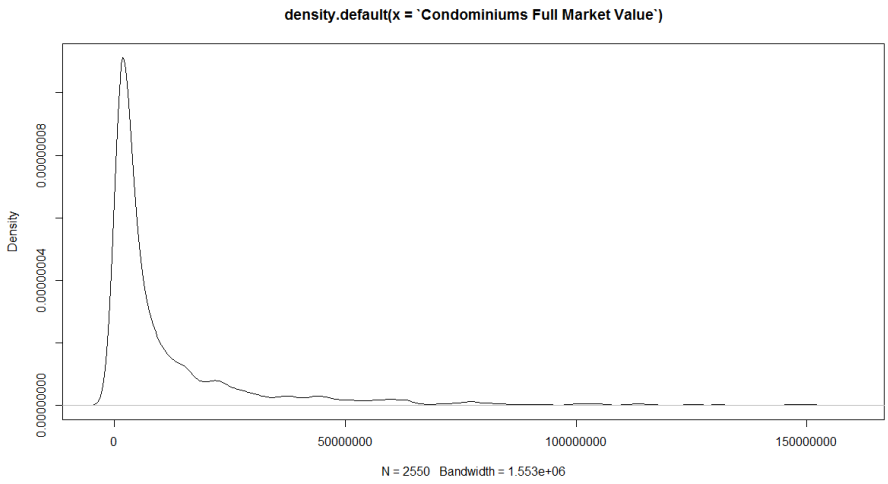
Building Classification according to the housing department's norms of the boroughs. We got R2-Condominium as 420 units, which is in a low density multi-dwelling zone. It allows approximately 21.8 dwelling units per acre. R4-Condominium's units are 1838 in number which comprises residential units that are common in tall buildings. That's why many of these units can be found in Manhattan (61%). R9-Condominium as 230 in number, they comprise of cond-ops. Cond-ops are condominium co-operatives. RR-Condominums are rental condominiums & they have around 64 units in total.

Number of Units according to the Millennium wise

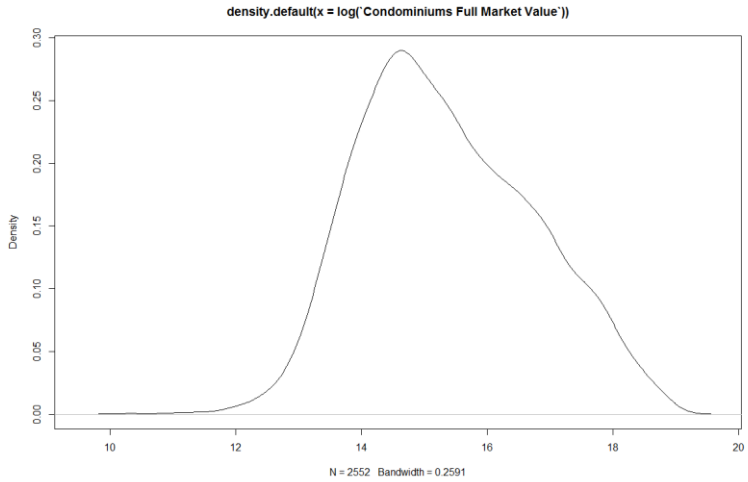


Data set have properties from year 1825 till 2011. Number of condominium units built before 2000 was 1588 while after 2000 their number decreased to 954. Mean of market value per sq. ft. for years before 2000 and after it was coming around same. Therefore, we segmented the time into two categories.

Condominiums Full Market Value

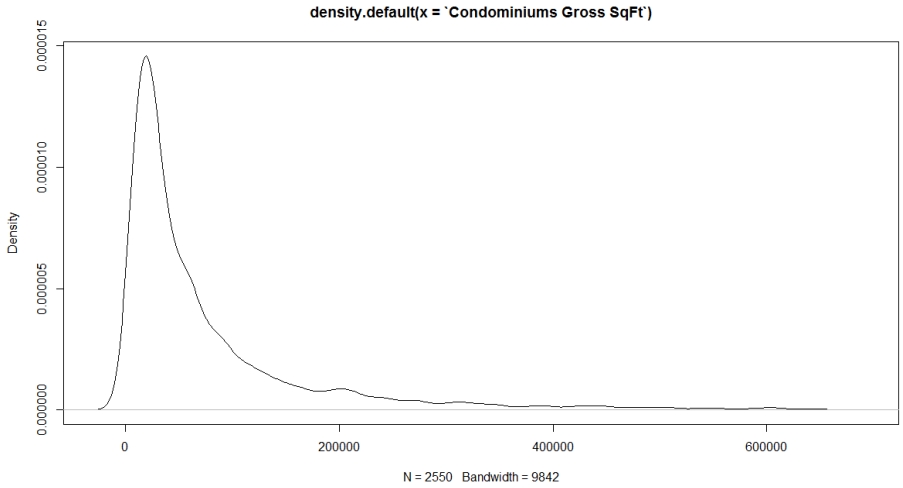



```
> describe(`Condominiums Full Market Value`)
   vars    n     mean       sd  median trimmed     mad   min       max     range skew kurtosis       se
X1    1 2550 12169875 20446212 4008499 7208356 4392208 30000 155560004 155530004 3.28    13.05 404895.3
> |
```



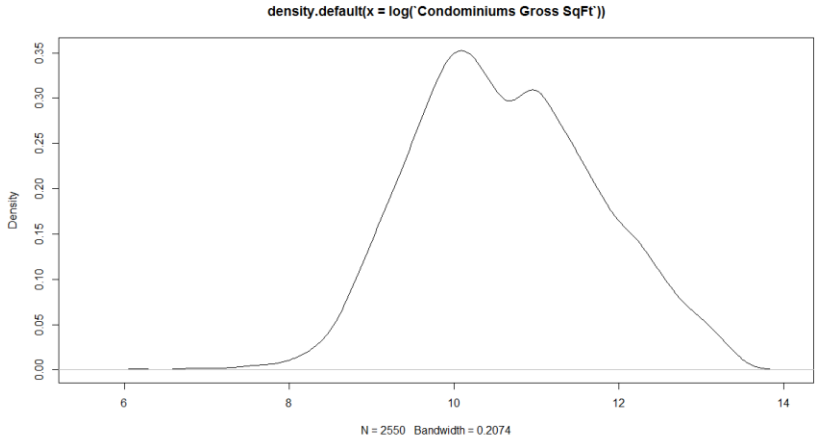
The data was highly right skewed. Skewness is 3.281 and kurtosis is 13.05. We transformed the variable by applying log and got normal distribution curve for the log (condominiums full market value).

Condominiums Gross per Sq. Ft.

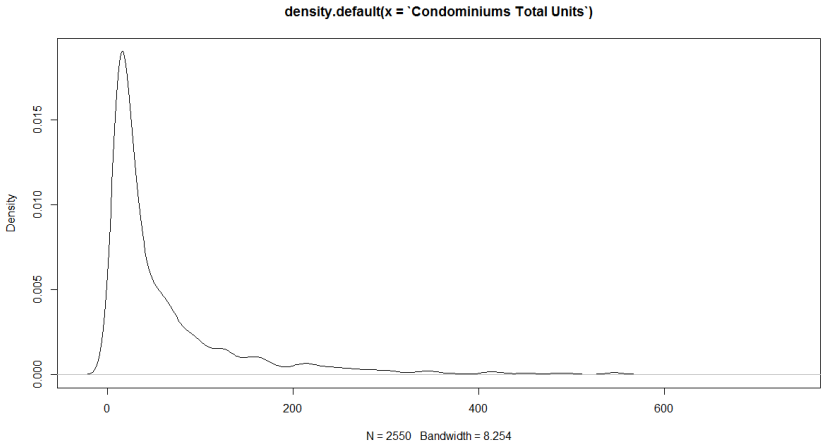



```
> describe(`Condominiums Gross SqFt`)
   vars    n     mean       sd  median  trimmed      mad min    max  range skew kurtosis      se
X1    1 2550 76241.73 98745.93 37682.5 54116.23 36488.27 478 665236 664758  2.7     8.47 1955.46
```

As we can see mean is larger than median, so the variable is right skewed. As this independent variable was highly skewed to the right, we performed log transformation to achieve normal distribution with a mean of **10.633** and a standard deviation of **1.11**.

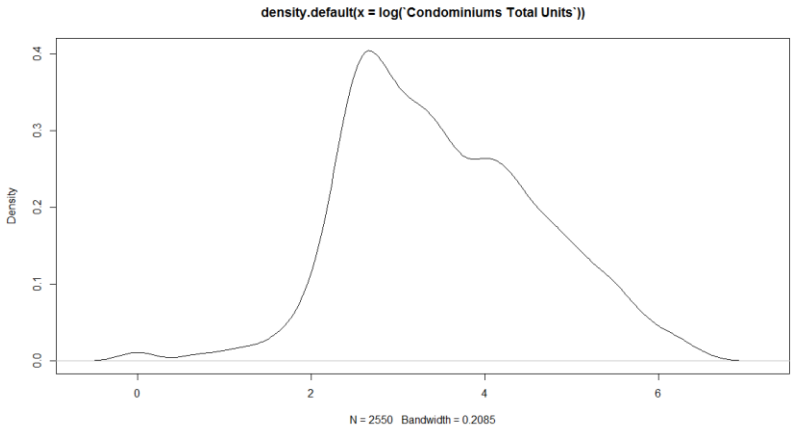


Condominiums Total Units

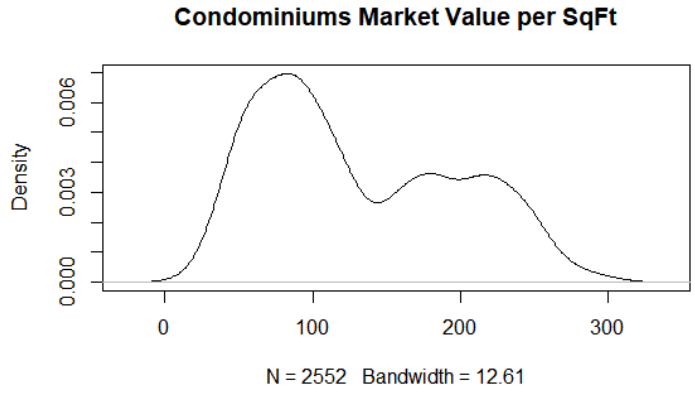



```
> describe(`Condominiums Total Units`)
   vars    n  mean    sd median trimmed   mad min max range skew kurtosis   se
X1    1 2550 64.62 89.02     30   44.35 26.69   1 713   712 3.04    11.54 1.76
```

As we can see mean is larger than median, so the variable is right skewed. As this independent variable was highly skewed to the right, we performed log transformation to achieve normal distribution.



Dependent Variable: Condominiums Market Value per Sq. Ft.

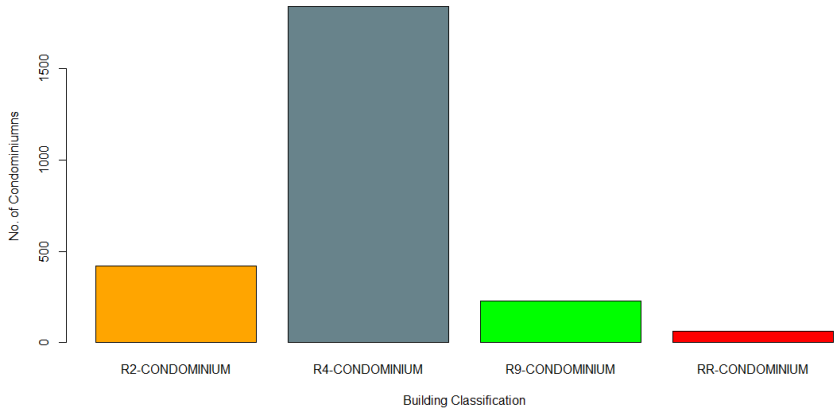


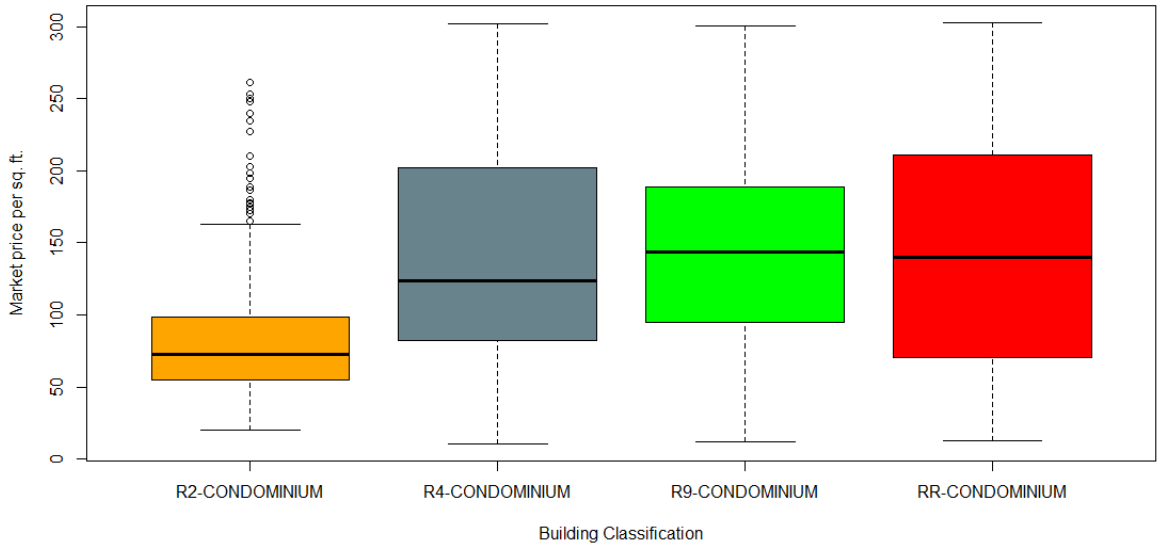
We tried to transform the variable by using log transformation, but we got the best distribution in its original state with a mean of **131.0108.**

8. Bivariate analysis

Market Price per sq. ft. and Condominium Building Classification

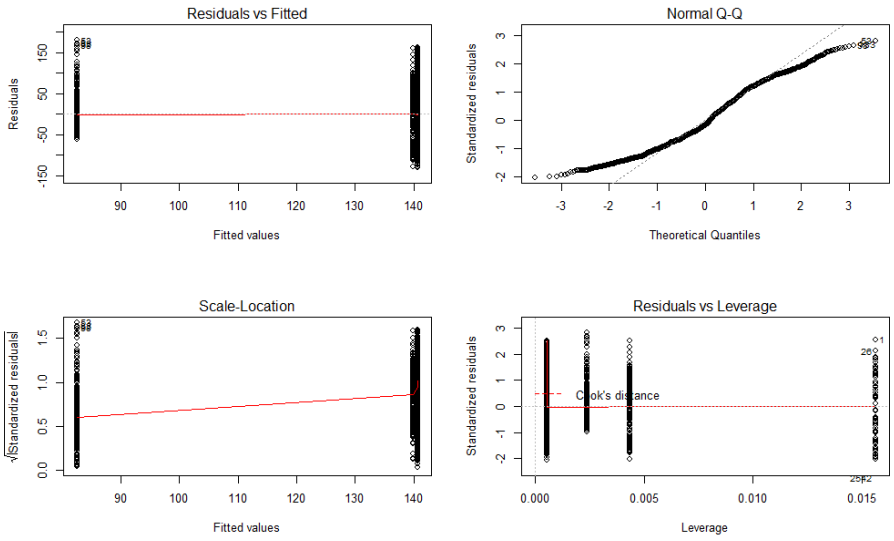
Building Classification	MEAN	MEDIAN	STD. DEVIATION	No. of Condominiums
R2-CONDOMINIUM	82.46714	72.665	40.25365	420
R4-CONDOMINIUM	140.65767	124.010	67.65189	1838
R9-CONDOMINIUM	139.86852	143.665	61.62104	230
RR-CONDOMINIUM	140.69984	140.245	79.38921	64



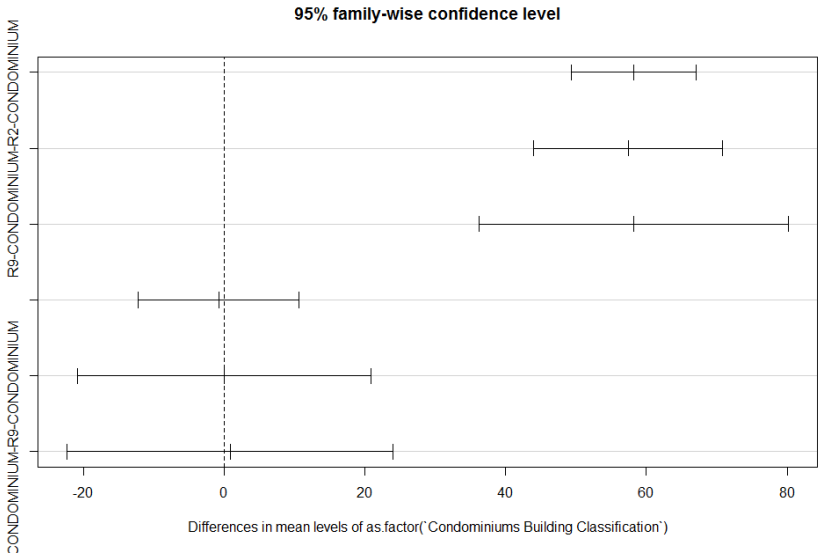


R4-Condominium are the residential units in elevator buildings(tall). Most of these units can be found in Manhattan (61%) whose Market Value per sq. ft. is higher than other boroughs. It is followed by Brooklyn (15%) and Queens (13%) in terms of number of R4- Condominiums.

Mean of market value per sq. ft. for RR- condominium is higher than the other condos, which implies that sale value of the properties having rental condominiums will be surely high but depends upon the number of units as the 'red' bar for RR-Condominium have quite few numbers of units.



```
> summary(price.aov)
                                                    Df   Sum Sq Mean Sq F value
as.factor(`Condominiums Building Classification`)    3  1184826  394942    97.2
Residuals                                         2548 10353086    4063
                                                             Pr(>F)
as.factor(`Condominiums Building Classification`) <0.0000000000000002 ***
Residuals
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

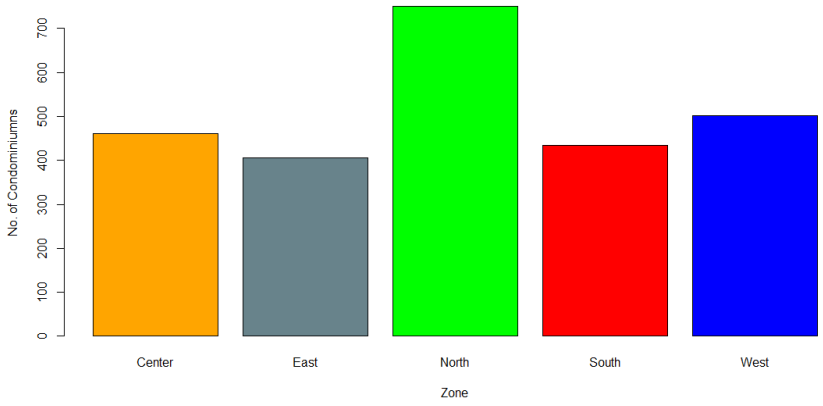


We performed Anova test and we can see from the P-value which is very less. This explains that there is a difference in market value across the Building classification. To determine the extent of the variation we further carried Tukey HSD test. Which showed that there is not much difference between the mean of R4, R9 and RR but there is difference between mean of R2 and other building types.

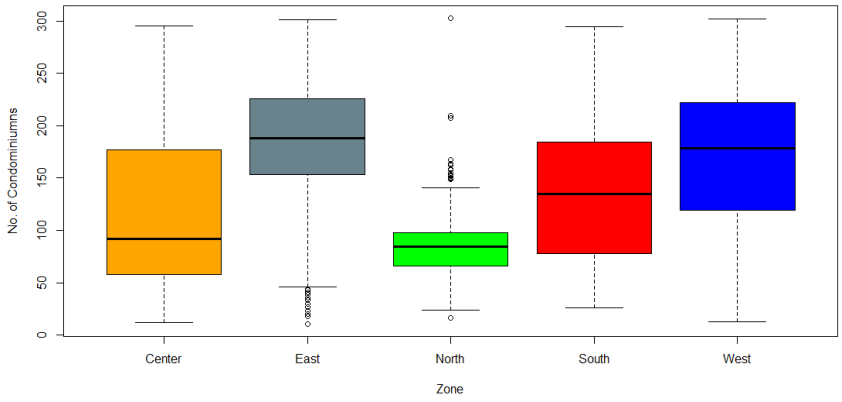
Taking variable Condominium building classification as factors with 4 levels of Condominiums such as R2, R4, R9 & RR, we get that the difference between the mean of each condo. Difference of means in R2 and R9 condominiums is huge that's why the graph for their difference of means is away from zero.

Market Price per sq. ft. and Zone

Building Zones	MEAN	MEDIAN	STD. DEVIATION	No. of Condominiums
CENTER	116.84178	91.965	69.43960	460
EAST	181.56022	187.870	58.25423	405
NORTH	84.09596	84.135	27.57437	750
SOUTH	136.73214	134.530	63.13812	435
WEST	168.34657	178.210	65.69291	502

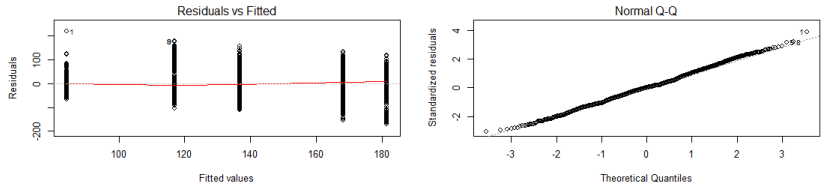


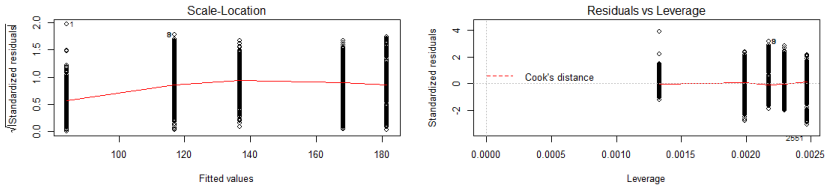
Market value per sq. ft. in each cardinal direction is evaluated where Mean & Median in East zone is seen higher than those in other zones. Though the number of condominium zones are higher in North. But the market value per sq. ft. for East zone is higher.

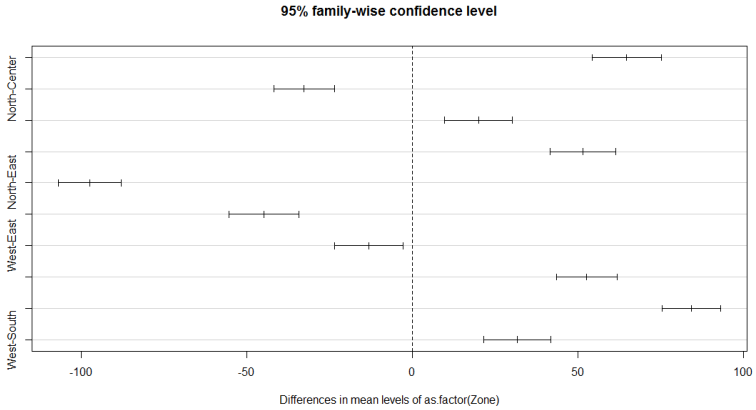



```
> summary(price.aov)
                Df  Sum Sq Mean Sq F value             Pr(>F)
as.factor(Zone)  4 3491983  872996   276.4 <0.0000000000000002 ***
Residuals     2547 8045930    3159
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

The Anova Diagonostic plot shows that the data is normally distribted and there are few outliars for the Northern zone. Also the low P-value suugests that there is differnce in market value w.r.t. zone. For further analysis we performed Tukey HSD test and found out that theere is differnec between the mean of each zone. Which makes is good fit for inculding in Model.





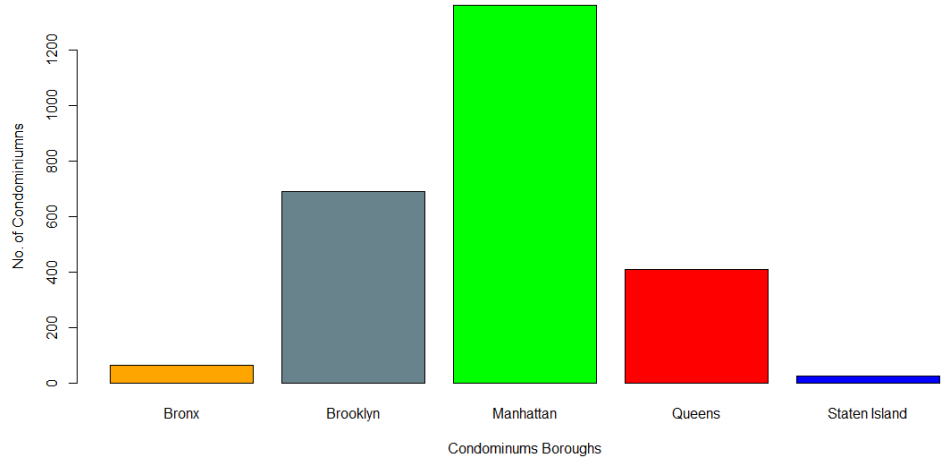


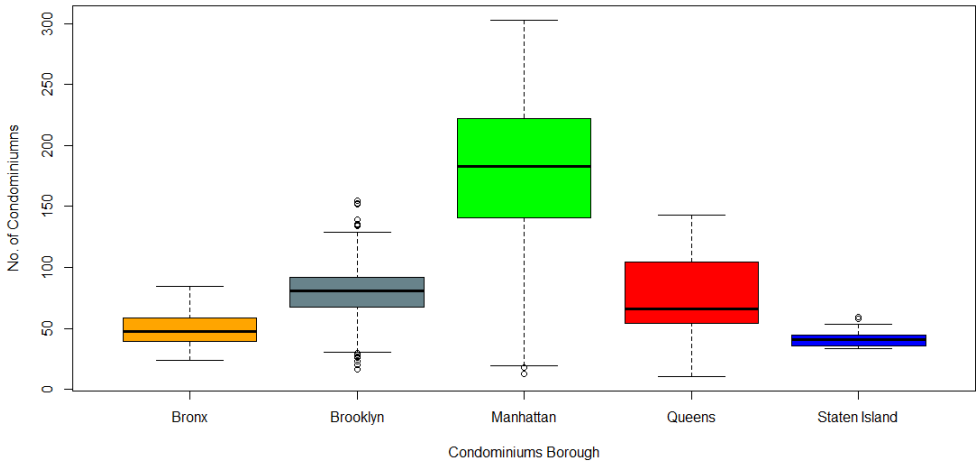
Market Price per sq. ft. and Condominiums Borough

Building Zones	MEAN	MEDIAN	STD. DEVIATION	No. of Condominiums
Bronx	48.68121	47.375	14.985328	66
Brooklyn	79.29486	80.820	21.376635	691
Manhattan	179.37640	182.955	54.227530	1360
Queens	76.51227	66.290	27.665125	409
Staten Island	41.86192	40.990	6.976115	26

After obtaining just the data of Manhattan for prediction of Market value per sq.ft. of the condominiums , we decided to merge data of other boroughs with in New York . We added data for Bronx, Brooklyn, Queens and Staten Island other than the data of Manhattan we we already had . As per our hypothesis, market value per sqft . with in the state of New York should be maximum for the borough that is in the center.

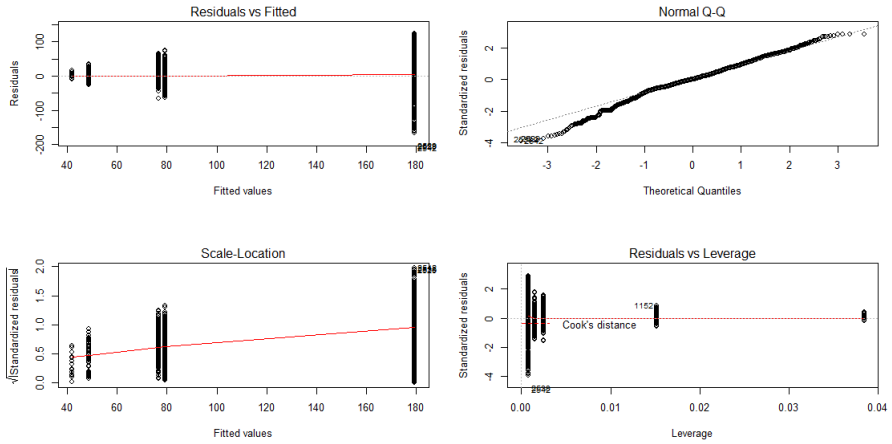
Here, we can see number of condominiums is maximum for Manhattan (1360) followed by Brooklyn that is close to the center Borough. Mean of market value per sq. ft. for condominiums in Manhattan is higher than the market value per sq ft. for condominiums in other Boroughs which implies market price per sq.ft. of condominiums is directly dependent on location of the boroughs.

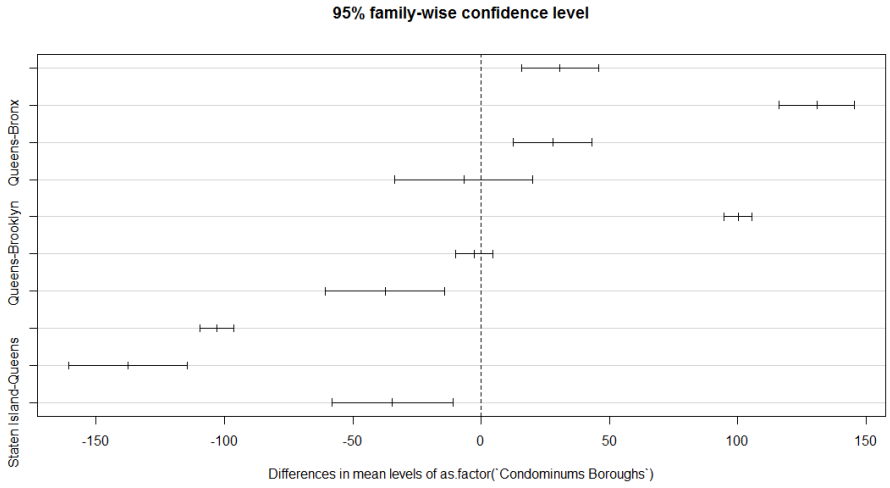





```
> summary(price.aov)
                                Df  Sum Sq Mean Sq F value             Pr(>F)
as.factor(`Condominums Boroughs`)    4 6898221 1724555   946.7 <0.0000000000000002
Residuals                         2547 4639692    1822

as.factor(`Condominums Boroughs`) ***
Residuals
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Box plotting gives out some outliers in Brooklyn and Staten Island and the median value of market value per sq. ft. for Manhattan is higher than other Boroughs.



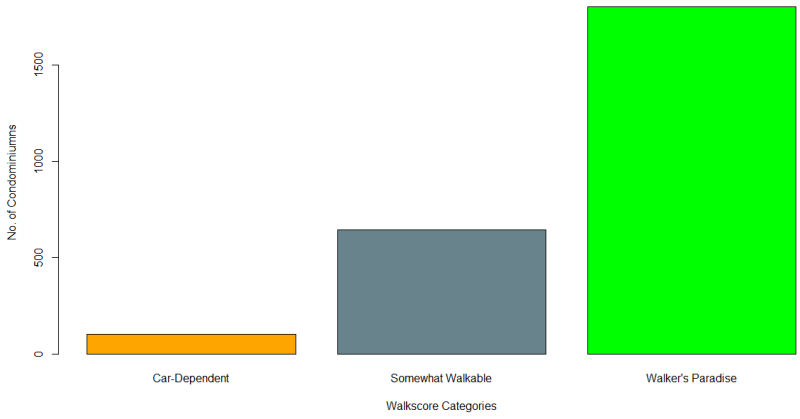


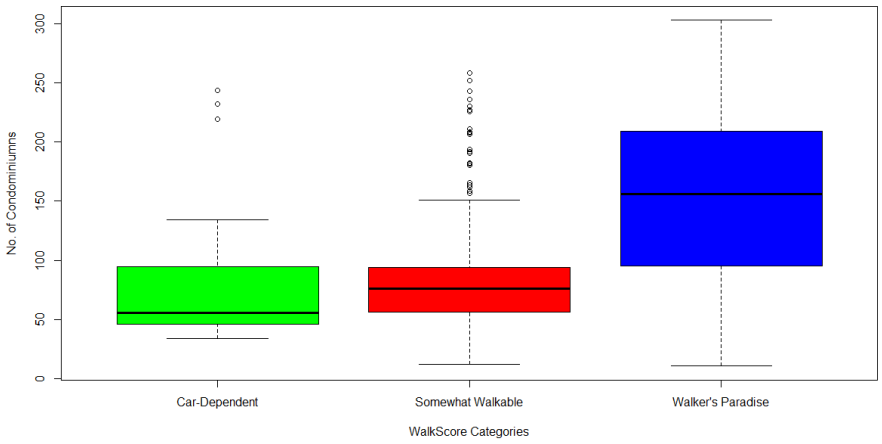
The Anova and Tukey HSD test together shows that there is significant difference in mean of market value across the five boroughs. Manhattan being the costliest and Staten Island being the cheapest.

Market Price per sq. ft. and WalkScore Categories

WalkScore Categories	MEAN	MEDIAN	STD. DEVIATION	No. of Condominiums
Car-Dependent	71.68050	55.700	42.50612	101
Somewhat Walkable	81.18622	76.040	37.00881	645
Walker's Paradise	152.28931	155.775	65.28442	1802

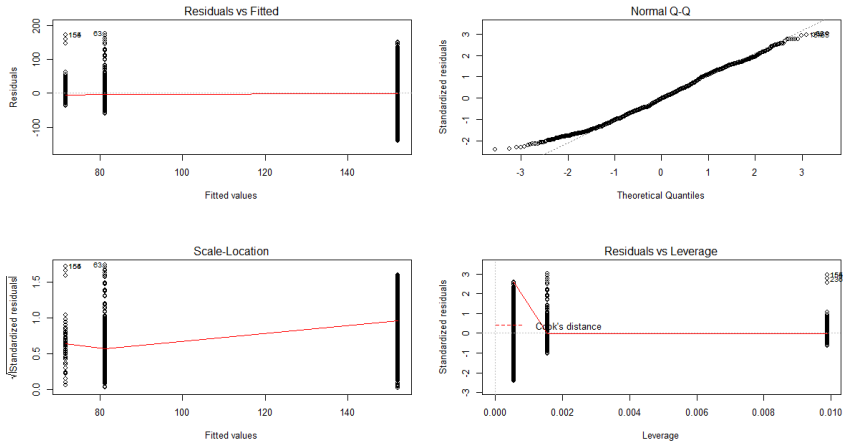
To improve our analysis and Model, we incorporated walk score through R API. Then we categorized Walk score into 3 categories viz-a-viz car-dependent, somewhat walkable and walker's paradise. Based on this factor, we can observe that number of condominiums in walker's paradise is the highest. Mean of market value per sq. ft. for condominiums is highest in Walker's paradise.

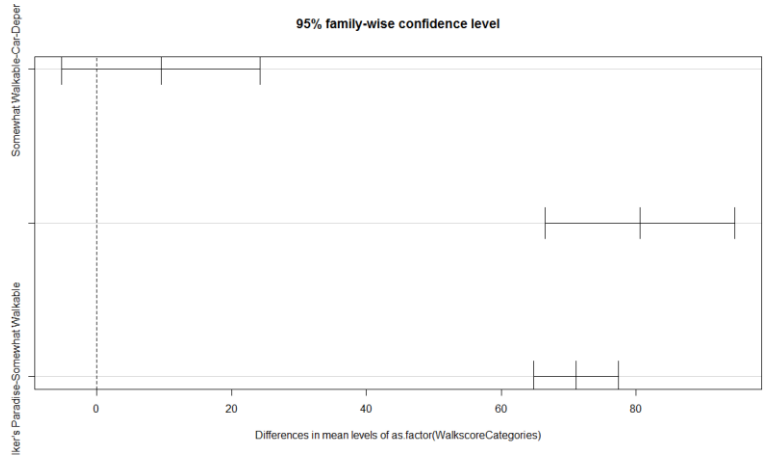





```
> summary(price.aov)
                              Df  Sum Sq Mean Sq F value             Pr(>F)
as.factor(WalkscoreCategories)    2 2772617 1386308   403.7 <0.0000000000000002 ***
Residuals                      2545 8738694    3434
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

From Anova test we can see that there is market value per sq. ft. is changing across the WalkScore categories. To substantiate this further we conducted TukeyHSD test and found out that there is no significant difference between the mean of Car-dependent and Somewhat Walkable but there is significant difference between the Walker’s paradise and other two categories.

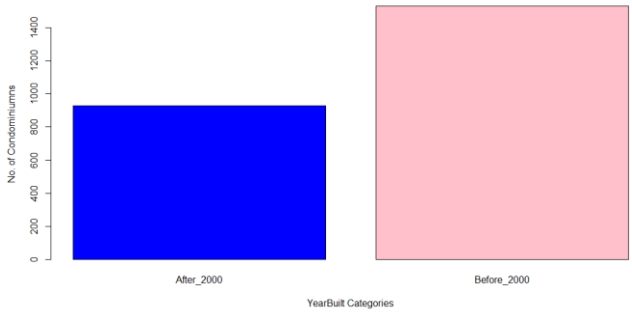


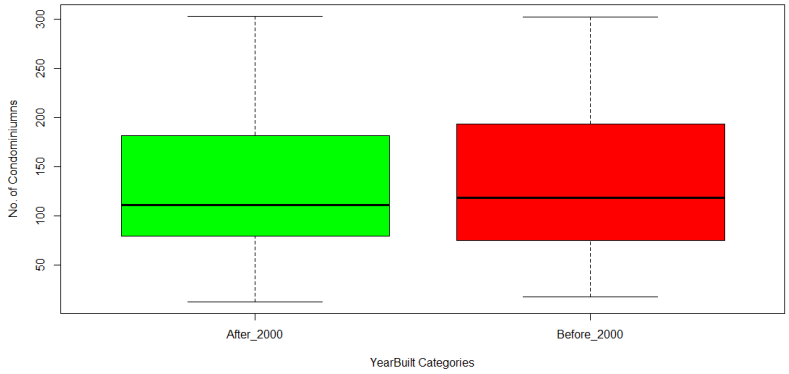


Market Price per sq. ft. and YearBuilt Categories

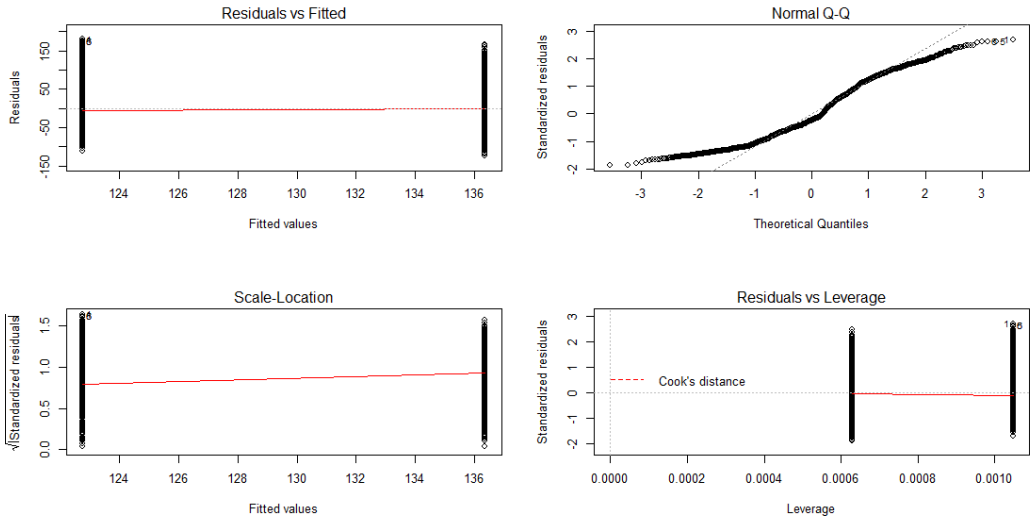
YearBuilt Categories	MEAN	MEDIAN	STD. DEVIATION	No. of Condominiums
After_2000	122.7403	96.905	63.13224	954
Before_2000	136.3647	134.340	69.05979	1588

Year Built category is a categorical variable created from the variable Condominiums Year Built. The aggregation shows that average market value per sq.ft of newer buildings is slightly lower than older building. Although, the difference is not very significant, there seems to be a difference in the market value with respect to the age of the building. To confirm this, we performed anova test on these variables.






```
> summary(price.aov)
                                Df   Sum Sq Mean Sq F value      Pr(>F)
as.factor(YearBuiltCategories)   1   110678  110678   24.73 0.000000704 ***
Residuals                     2542 11376698    4475
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
8 observations deleted due to missingness
```



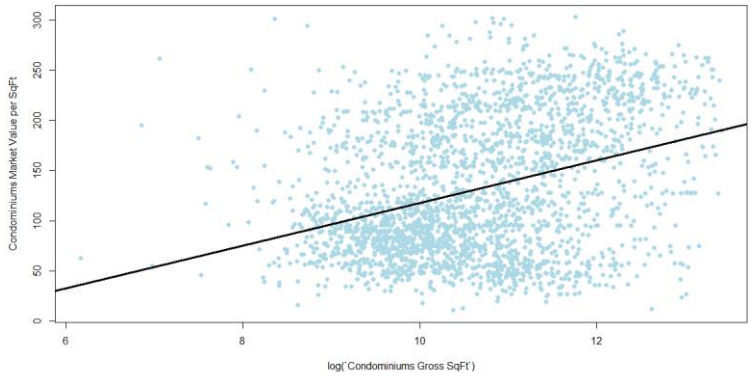
```
        Welch Two Sample t-test

data:  `Condominiums Market Value per SqFt`[YearBuiltCategories == "Before_2000"] and `Condominiums Market Val
ue per SqFt`[YearBuiltCategories == "After_2000"]
t = 5.1046, df = 2148.5, p-value = 0.0000001803
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 9.268553      Inf
sample estimates:
mean of x mean of y
 136.4182  122.7403
```

P-value of the Welch t.test is significant. Hence, we can't reject the null hypothesis that there is no difference in the dataset with respect to the age of the building. The observation is in line with our assumption that newer building will have higher market value per sq. ft. Hence, our hypothesis regarding the age and market value per sq. ft. of condominium properties holds true.

Market Price per sq. ft. and (Condominiums Gross Sq. Ft.)

```
> cor(`Condominiums Market Value per SqFt`,log(`Condominiums Gross SqFt`), use="complete.
obs", method="pearson")
[1] 0.3495494
```

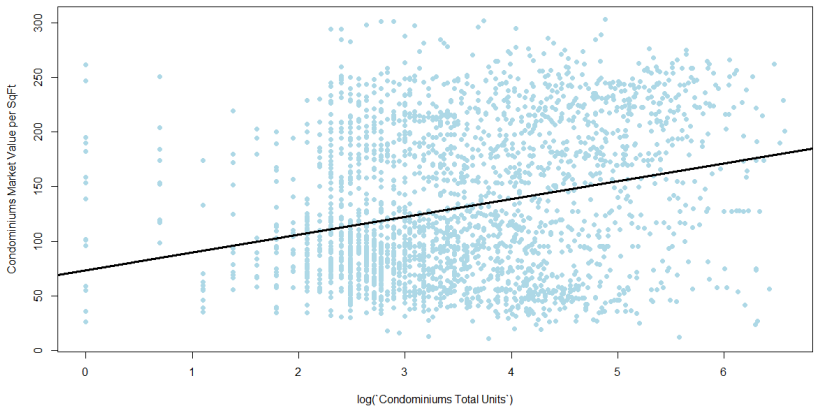
The correlation between the dependent variable, market value per sq. ft. and log (Condominium Gross SqFt) is 0.34.95. This is good correlation. Therefore, we can include this variable in our Model.



Market Price per sq. ft. and log(Condominiums Total Units)

```
> cor(`Condominiums Market Value per SqFt`,log(`Condominiums Total Units`), use="complete
.obs", method="pearson")
[1] 0.2693579
```

The correlation between the dependent variable, market value per sq. ft. and log (Condominium Total Units) is 0.26.93. This is good correlation. Therefore, we can include this variable in our Model.



9. Linear regression models

Model 1 – Predicting Market value per sq. ft. by selecting independent variable

First, we tried to model the dependent variable, market value per sq. ft. with the following independent variable: Condominiums Boroughs, Condominiums Building Classification, Condominiums Total Units and Condominiums Gross Sq. Ft.

```
Coefficients:
                                                          Estimate Std. Error
(Intercept)                                               -31.067     15.765
as.factor(`Condominiums Building Classification`)R4-CONDOMINIUM   14.729      2.459
as.factor(`Condominiums Building Classification`)R9-CONDOMINIUM   -7.204      3.690
as.factor(`Condominiums Building Classification`)RR-CONDOMINIUM  -22.261      5.784
as.factor(`Condominums Boroughs`)Brooklyn                  30.317      5.330
as.factor(`Condominums Boroughs`)Manhattan                123.987      5.241
as.factor(`Condominums Boroughs`)Queens                    24.972      5.497
as.factor(`Condominums Boroughs`)Staten Island            -19.303      9.582
log(`Condominiums Gross SqFt`)                             7.189      1.996
log(`Condominiums Total Units`)                           -0.508      1.940
                                                            
(Intercept)                                               *
as.factor(`Condominiums Building Classification`)R4-CONDOMINIUM ***
as.factor(`Condominiums Building Classification`)R9-CONDOMINIUM .
as.factor(`Condominiums Building Classification`)RR-CONDOMINIUM ***
as.factor(`Condominums Boroughs`)Brooklyn                  ***
as.factor(`Condominums Boroughs`)Manhattan                 ***
as.factor(`Condominums Boroughs`)Queens                    ***
as.factor(`Condominums Boroughs`)Staten Island             *
log(`Condominiums Gross SqFt`)                             ***
log(`Condominiums Total Units`)
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 41.08 on 2538 degrees of freedom
Multiple R-squared:  0.628, Adjusted R-squared:  0.6267
F-statistic: 476.1 on 9 and 2538 DF,  p-value: < 0.00000000000000022
```

We are considering it as our base model from which we will try to improve the adjusted R squared value further. Result of the model was that we got an adjusted r-squared value of 62.67%. As we can see from

the result borough Staten Island and Building type R9 condominium is displaying less significance. The reason is that Staten Island has very less observations (Total 26) compared to other boroughs. That is why it is not significant is predicting market value per sq. ft. For similar reason R9-CONDOMINUM is showing less significance.

We will try to improve our model further by categorizing the Condominium year built independent variable and adding it into the base model.

Model 2 - Predicting Market value per sq. ft. by adding YearBuiltCategories independent variable to base Model

For Model 2, We a new calculated variable, i.e., YearBuiltCategories and added it to the base Model (Model 1). We divided the variable into two parts i.e. buildings that are constructed before year 2000 and buildings that are constructed in and after year 2000.

Adding new categorical independent variable led to an improvement in the adjusted r-squared value from 63.33% to 64.65%.

```
Coefficients:
                                                    Estimate Std. Error
(Intercept)                                          -17.7773    15.4020
as.factor(`Condominiums Building Classification`)R4-CONDOMINIUM   8.0962     2.4636
as.factor(`Condominiums Building Classification`)R9-CONDOMINIUM -10.0190     3.6010
as.factor(`Condominiums Building Classification`)RR-CONDOMINIUM -39.0947     5.7947
as.factor(`Condominums Boroughs`)Brooklyn             30.1745     5.1857
as.factor(`Condominums Boroughs`)Manhattan           133.6023     5.1575
as.factor(`Condominums Boroughs`)Queens               29.2753     5.3577
as.factor(`Condominums Boroughs`)Staten Island       -6.1521     9.3812
log(`Condominiums Gross SqFt`)                        6.7938     1.9476
log(`Condominiums Total Units`)                       0.8029     1.8939
YearBuiltCategoriesBefore_2000                      -22.5914     1.8414
```

```
 (Intercept)
as.factor(`Condominiums Building Classification`)R4-CONDOMINIUM **
as.factor(`Condominiums Building Classification`)R9-CONDOMINIUM **
as.factor(`Condominiums Building Classification`)RR-CONDOMINIUM ***
as.factor(`Condominums Boroughs`)Brooklyn                      ***
as.factor(`Condominums Boroughs`)Manhattan                     ***
as.factor(`Condominums Boroughs`)Queens                        ***
as.factor(`Condominums Boroughs`)Staten Island
log(`Condominiums Gross SqFt`)                                 ***
log(`Condominiums Total Units`)
YearBuiltCategoriesBefore_2000                                 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 39.95 on 2529 degrees of freedom
  (8 observations deleted due to missingness)
Multiple R-squared:  0.6479,    Adjusted R-squared:  0.6465
F-statistic: 465.4 on 10 and 2529 DF,  p-value: < 0.00000000000000022
```

The adjusted r-squared value improved by nearly 2% but we further needed to fine improve our model. We further created one more categorical variable 'Zone' and added it into the Model 2. Condominiums market value per sq. ft. is $22 cheaper for before_2000(older) condominiums.

Model 3 - Predicting Market value per sq. ft. by adding Zone independent variable to Model 2

For model 3, We calculated a new independent variable Zone by dividing each borough w.r.t to cardinal directions east, west, north, south and center. We wanted to see if within each borough the direction at which condominiums are situated affects the market value.

```
Coefficients:
                                                                  Estimate Std. Error
(Intercept)                                                      -10.1753    15.0791
as.factor(`Condominiums Building Classification`)R4-CONDOMINIUM    7.8970     2.3744
as.factor(`Condominiums Building Classification`)R9-CONDOMINIUM  -11.3234     3.4732
as.factor(`Condominiums Building Classification`)RR-CONDOMINIUM  -37.8304     5.5836
as.factor(`Condominums Boroughs`)Brooklyn                         29.8825     5.0295
as.factor(`Condominums Boroughs`)Manhattan                       119.8299     5.0687
as.factor(`Condominums Boroughs`)Queens                           27.1324     5.2156
as.factor(`Condominums Boroughs`)Staten Island                   -10.8813     9.0855
log(`Condominiums Gross SqFt`)                                    7.4886     1.8878
log(`Condominiums Total Units`)                                  -0.7419     1.8478
as.factor(YearBuiltCategories)Before_2000                        -23.6924     1.7825
as.factor(Zone)East                                               17.1761     2.8001
as.factor(Zone)North                                             -18.0036     2.3538
as.factor(Zone)South                                              -3.0013     2.6601
as.factor(Zone)West                                               11.4885     2.6162
   (Intercept)
   as.factor(`Condominiums Building Classification`)R4-CONDOMINIUM ***
   as.factor(`Condominiums Building Classification`)R9-CONDOMINIUM **
   as.factor(`Condominiums Building Classification`)RR-CONDOMINIUM ***
   as.factor(`Condominums Boroughs`)Brooklyn                      ***
   as.factor(`Condominums Boroughs`)Manhattan                     ***
   as.factor(`Condominums Boroughs`)Queens                        ***
   as.factor(`Condominums Boroughs`)Staten Island
   log(`Condominiums Gross SqFt`)                                 ***
   log(`Condominiums Total Units`)
   as.factor(YearBuiltCategories)Before_2000                      ***
   as.factor(Zone)East                                            ***
   as.factor(Zone)North                                           ***
   as.factor(Zone)South
   as.factor(Zone)West                                            ***
   ---
   Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

   Residual standard error: 38.43 on 2525 degrees of freedom
     (8 observations deleted due to missingness)
   Multiple R-squared:  0.6746,    Adjusted R-squared:  0.6728
   F-statistic: 373.8 on 14 and 2525 DF,  p-value: < 0.00000000000000022
```

After considering this new independent variable in our model, the adjusted r-squared value has further increased to 67.28%. This approach made it clear that moving to east and west from center market value per sq. ft. is increasing $17 and $11 respectively. Similarly, it was decreasing if we move to north or south from center.

Model 4 - Predicting Market value per sq. ft. by adding WalkscoreCategories independent variable to Model 3

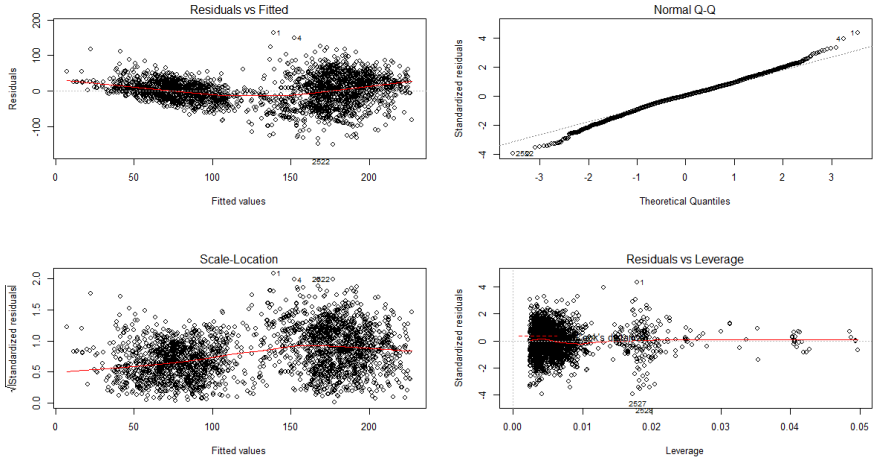
For Model 4, We further wanted to increase the adjusted R squared value further. We introduced a new variable called WalkscoreCategories. We installed the WalkScore packages into R and used longitude and latitude value of dataset to calculate the WalkScore value. We used the below code

```
Real_estate_NY_Project$Walkscore[i]<-
getWS(Longitude[i],Latitude[i],"24dde54ead291ea8d1c09476aa2ed2c2")$walkscore
```

Most of the values are above 90 and made the variable highly skewed. For better result we transformed the WalkScore variable to three categories: Walker's Paradise, Somewhat Walkable and Car needed.

```
Coefficients:
                                                          Estimate Std. Error
(Intercept)                                               -10.345     15.405
as.factor(`Condominiums Building Classification`)R4-CONDOMINIUM   7.018      2.373
as.factor(`Condominiums Building Classification`)R9-CONDOMINIUM  -12.279      3.463
as.factor(`Condominiums Building Classification`)RR-CONDOMINIUM  -38.685      5.543
as.factor(`Condominums Boroughs`)Brooklyn                  28.324      5.061
as.factor(`Condominums Boroughs`)Manhattan                112.106      5.251
as.factor(`Condominums Boroughs`)Queens                    26.699      5.192
as.factor(`Condominums Boroughs`)Staten Island             -9.900      9.378
log(`Condominiums Gross SqFt`)                             7.857      1.874
log(`Condominiums Total Units`)                           -1.120      1.835
YearBuiltCategoriesBefore_2000                            -24.729      1.776
as.factor(Zone)East                                        17.786      2.783
as.factor(Zone)North                                      -18.237      2.339
as.factor(Zone)South                                       -2.272      2.658
as.factor(Zone)West                                        12.554      2.607
as.factor(WalkscoreCategories)Somewhat Walkable            -6.692      4.486
as.factor(WalkscoreCategories)Walker's Paradise             6.952      4.557
(Intercept)
as.factor(`Condominiums Building Classification`)R4-CONDOMINIUM **
as.factor(`Condominiums Building Classification`)R9-CONDOMINIUM ***
as.factor(`Condominiums Building Classification`)RR-CONDOMINIUM ***
as.factor(`Condominums Boroughs`)Brooklyn                   ***
as.factor(`Condominums Boroughs`)Manhattan                  ***
as.factor(`Condominums Boroughs`)Queens                     ***
as.factor(`Condominums Boroughs`)Staten Island
log(`Condominiums Gross SqFt`)                              ***
log(`Condominiums Total Units`)
YearBuiltCategoriesBefore_2000                              ***
as.factor(Zone)East                                         ***
as.factor(Zone)North                                        ***
as.factor(Zone)South
as.factor(Zone)West                                         ***
as.factor(WalkscoreCategories)Somewhat Walkable
as.factor(WalkscoreCategories)Walker's Paradise
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 38.14 on 2523 degrees of freedom
  (8 observations deleted due to missingness)
Multiple R-squared:  0.6798,    Adjusted R-squared:  0.6777
F-statistic: 334.7 on 16 and 2523 DF,  p-value: < 0.00000000000000022
```

After transforming the variable, we got better adjusted R squared value of 67.77%. Since we could not increase the adjusted R squared value without adding any additional variable. We considered Model 4 as our final model.



10. Conclusion

After making four models we succeed to get adjusted R-square value as 67.77% by involving zone, WalkScore categories as well as the condominium total units. Final adjusted R-square value indicates the 67.77% of variance in Market Value per sq. ft. of condominiums over the 5 boroughs. Therefore, the factors constitute to our regression modelling to get the variance of Market Value per sq. ft. are Condominiums Boroughs, Condominiums Building Classification, Condominiums Total Units, Condominiums Gross Sq. Ft., WalkScore categories, Year built categories and Zones. Though through Anova and Tukey HSD test we got to know that one East and West zone contributes higher than the rest of zones. R2- Condominium contribute more to the model than rest of building. Similarly, rows with WalkScore category Walker's paradise contribute more to the model.

Hypothesis 1:

Condominium's market value per sq. ft. increases with increase in Walk score- Bivariate analysis explains that Market value per sq. ft. is highest in "Walker's paradise" category. Also, the number of condominiums in this category were found more in number with the mean value 147.21105 and variance of 67.49%. So, condominiums with high WalkScore ratings tend to have more market value per sq. ft. than others.

Hypothesis 2:

Condominium's market value per sq. ft. is relatively higher in East of borough area as from the bivariate analysis of price with respect to Zones we infer that of mean is 181.56022 and median is 187.870 which is higher than any other zones. Therefore, the market value per sq. ft. for eastern part of the boroughs was found to be much higher than it was predicted for center. Second most expensive was the West zone. So moving towards east or west from center will increase the market value of the condominiums.

Hypothesis 3:

Condominium's market value per sq. ft. is comparatively lower for low-rise condominiums as these buildings comprising of low-rise walk-up condos rather than having elevators. By bivariate analysis we get the mean of R2-Condominiums as 82.46714 which is much lower than the other building classifications. R2- condos are mostly 2 story or duplex so as compared to high rise buildings their market value is not that high. Therefore, hypothesis that Condominium's market value per sq. ft. is comparatively lower for low-rise condominiums holds good.

Hypothesis 4:

Condominium's market value per sq. ft. is higher for newer condominiums. Bivariate analysis of market value per sq. ft. for newly constructed condominiums i.e. with year built of after 2000 explains that mean price for newer buildings for years after 2000 was coming relatively lower than older built buildings but the difference is insignificant. Also, through t-test we get the relatively small p-value which indicates the truthfulness of our hypothesis.