Cleaning

Kaushik Mohan

11/14/2018

## Datasets

1. NYC Annualized Property Sales Data (2012-2017)
2. MapPLUTO (18v1)
3. Geoclient API v1.1

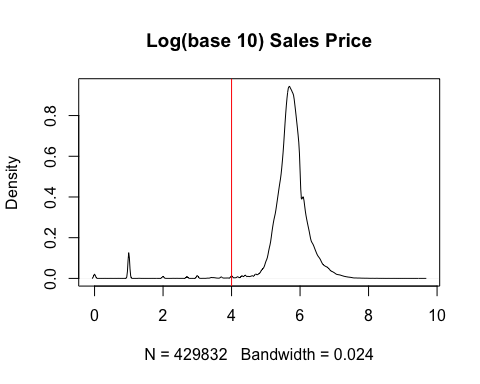
## Merging Process

1. Subset building classes A,B,C,D,R,S as these are the ones coming under the residential tax class
2. Merge with MapPLUTO data on Borough,Block and Lot
3. Condo listings don’t merge as there is a mismatch in BBL with MapPLUTO data. Therefore, we use Geoclient API to get the condominiumBillingBbl for the condos and then merge with MapPLUTO data on the BBL.

## Cleaning Data

### 1. Sale Price

First, we analyse the distribution of the sale prices. We note a sharp peak at 1, which is a price of . Also, from the overall distribution, we observe a lot of spikes below (red line). Hence, we choose to remove all the cases where the Price is less than .



##   
## A B C D R S   
## 100271 87121 66715 84443 79037 12245

##   
## A B C D R S   
## 31190 32459 23519 8777 4 5520

##   
## A B C D R S   
## 33238 34543 24777 9177 4 5783

## [1] 0.2360666

## [1] 0.2501489

## [1] "25.01% rows removed"

### 2. Area

We need the area of the building because we would calculate the price per sq. ft which is a better measure than the sale price itself.

We use the GR.SQFT from the Property Assessment Roll.

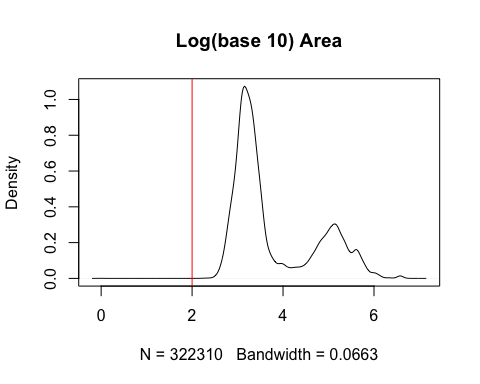
We have Gross Sq. Ft in the Property Sales Data and Building Floor Area (BldgArea) Residential Floor Area (ResArea) in the MapPLUTO Data. We note that Gross Sq. Ft strongly associated and predicted by with both these with a slope of nearly 1, except for the cases when the Gross Sq. Ft is 0, of course. Given Bldg Area has way fewer missing values, we choose to predict the missing values in Gross Sq. Ft with the using the Bldg Area from MapPLUTO data based on the simple linear model we fit on the non-missing ones.

##   
## A B C D R S   
## 1.16 0.75 0.33 0.01 6.63 1.15

## [1] 0.02056405

### 3. Filtering by area

We see a few cases with area less than 100 sq.ft (red line). We remove these.

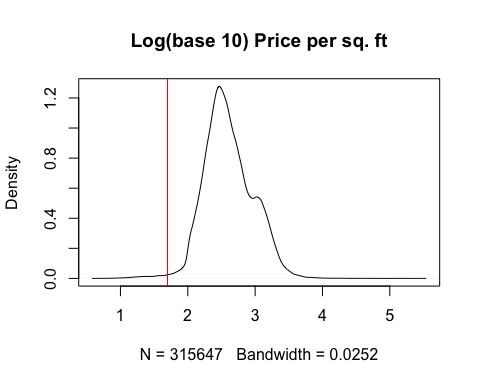


## [1] "1.55% rows removed"

### 4. Price per sq. ft

We compute Price per square feet from the sale price and above computed area. Apartments and Coops don’t have area for each unit and therefore we approximate those to be total area of the building divided by the total units in the building. Any recod with price per sq. ft < 100 is multiplied by total number of units.

Looking at the distribution, we see a two modes around per sq. feet and per sq.ft with a dip around (red line).



##   
## A B C D R S   
## 66243 52186 41793 75261 73776 6388

## s2  
## A B C D R S   
## 642 559 531 736 233 82

## [1] "0.65% rows removed"

#### 4.1 Checking Price Distributions

Below are the marginals by Borough and Building class and distributions of price per square feet by building class.

## # A tibble: 5 x 2  
## borough median\_price\_per\_sqft  
## <dbl> <dbl>  
## 1 1 973.  
## 2 2 194.  
## 3 3 366.  
## 4 4 303.  
## 5 5 274.

## # A tibble: 6 x 2  
## building\_class median\_price\_per\_sqft  
## <chr> <dbl+lbl>   
## 1 A 327.5530   
## 2 B 281.5657   
## 3 C 267.3636   
## 4 D 394.8000   
## 5 R 955.1777   
## 6 S 273.0088

## # A tibble: 30 x 3  
## # Groups: borough [?]  
## borough building\_class median\_price\_per\_sqft  
## <dbl> <chr> <dbl>  
## 1 1 A 1737.  
## 2 1 B 830.  
## 3 1 C 602.  
## 4 1 D 691.  
## 5 1 R 1328.  
## 6 1 S 865.  
## 7 2 A 246.  
## 8 2 B 197.  
## 9 2 C 175.  
## 10 2 D 163.  
## # ... with 20 more rows