Queen's\_Data\_Analysis\_Report

Bryan\_Kevin\_Ruhaab

June 18, 2017

# Cleaned Queens Data Exploratory data Analysis

This documents presents the analysis of queens dataset that has been cleaned

#Required packages  
packages <- c("ggplot2","plyr")  
  
#Check if the package is installed. if not, install it.  
  
is.installed <- function(mypkg){  
 is.element(mypkg, installed.packages()[,1])  
}  
  
for(package in packages){  
 # check if package is installed  
 if (!is.installed(package)){  
 install.packages(package, repos = "https://cran.r-project.org/")  
 }else {print(paste(package, " is library already installed"))}  
}

## [1] "ggplot2 is library already installed"  
## [1] "plyr is library already installed"

#load the required packages  
print("loading the required packages!")

## [1] "loading the required packages!"

suppressMessages(lapply(packages, require, character.only = TRUE))

## Warning: package 'ggplot2' was built under R version 3.3.3

## Warning: package 'plyr' was built under R version 3.3.3

## [[1]]  
## [1] TRUE  
##   
## [[2]]  
## [1] TRUE

## Loading clean Queens Dataset and exploratory

# read csv file  
queens<- data.frame(read.csv("C:/Users/kevolo/Desktop/smu/doing datascience/DoingDataScienceGroupProject/data/rollingsales\_queens\_clean.csv",skip=0,header=TRUE))  
  
# check the number of rows and columns that exist  
dim(queens)

## [1] 10299 27

# List the columns names  
names(queens)

## [1] "X" "borough"   
## [3] "neighborhood" "building.class.category"   
## [5] "tax.class.at.present" "block"   
## [7] "lot" "ease.ment"   
## [9] "building.class.at.present" "address"   
## [11] "apartment.number" "zip.code"   
## [13] "residential.units" "commercial.units"   
## [15] "total.units" "land.square.feet"   
## [17] "gross.square.feet" "year.built"   
## [19] "tax.class.at.time.of.sale" "building.class.at.time.of.sale"  
## [21] "sale.price" "sale.date"   
## [23] "sale.price.n" "gross.sqft"   
## [25] "land.sqft" "outliers"   
## [27] "missing\_Sqft"

#check what the data looks like  
head(queens, 3)

## X borough neighborhood  
## 1 1 4 AIRPORT LA GUARDIA   
## 2 2 4 AIRPORT LA GUARDIA   
## 3 3 4 AIRPORT LA GUARDIA   
## building.class.category tax.class.at.present block  
## 1 01 ONE FAMILY DWELLINGS 1 976  
## 2 01 ONE FAMILY DWELLINGS 1 976  
## 3 01 ONE FAMILY DWELLINGS 1 976  
## lot ease.ment building.class.at.present  
## 1 10 NA A5  
## 2 11 NA A5  
## 3 61 NA A5  
## address apartment.number zip.code  
## 1 21-10 81ST STREET 11370  
## 2 21-12 81ST STREET 11370  
## 3 21-21 80TH STREET 11370  
## residential.units commercial.units total.units land.square.feet  
## 1 1 - 1 1,800   
## 2 1 - 1 1,800   
## 3 1 - 1 1,800   
## gross.square.feet year.built tax.class.at.time.of.sale  
## 1 1,224 1950 1  
## 2 1,224 1950 1  
## 3 1,224 1950 1  
## building.class.at.time.of.sale sale.price sale.date sale.price.n  
## 1 A5 600,000 4/10/2017 600000  
## 2 A5 600,000 4/10/2017 600000  
## 3 A5 660,000 7/26/2016 660000  
## gross.sqft land.sqft outliers missing\_Sqft  
## 1 1224 1800 0 FALSE  
## 2 1224 1800 0 FALSE  
## 3 1224 1800 0 FALSE

tail(queens, 3)

## X borough neighborhood  
## 10297 25778 4 WOODSIDE   
## 10298 25782 4 WOODSIDE   
## 10299 25787 4 WOODSIDE   
## building.class.category tax.class.at.present  
## 10297 29 COMMERCIAL GARAGES 4  
## 10298 30 WAREHOUSES 4  
## 10299 30 WAREHOUSES 4  
## block lot ease.ment building.class.at.present  
## 10297 1321 43 NA G2  
## 10298 2287 1 NA E1  
## 10299 2287 10 NA E1  
## address apartment.number zip.code  
## 10297 52-19 QUEENS BOULEVARD 11377  
## 10298 47-59 49TH STREET 11377  
## 10299 47-39 49TH STREET 11377  
## residential.units commercial.units total.units land.square.feet  
## 10297 0 1 1 4,000   
## 10298 0 1 1 14,137   
## 10299 0 1 1 8,800   
## gross.square.feet year.built tax.class.at.time.of.sale  
## 10297 4,000 1939 4  
## 10298 14,137 1938 4  
## 10299 8,800 1932 4  
## building.class.at.time.of.sale sale.price sale.date sale.price.n  
## 10297 G2 2,925,000 11/15/2016 2925000  
## 10298 E1 1,700,000 12/16/2016 1700000  
## 10299 E1 1,700,000 12/16/2016 1700000  
## gross.sqft land.sqft outliers missing\_Sqft  
## 10297 4000 4000 0 FALSE  
## 10298 14137 14137 0 FALSE  
## 10299 8800 8800 0 FALSE

#check the structure of your dataset  
str(queens)

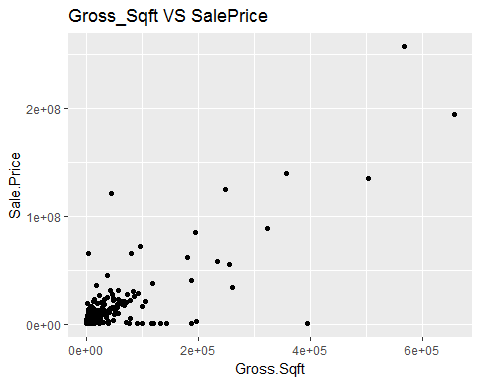
## 'data.frame': 10299 obs. of 27 variables:  
## $ X : int 1 2 3 4 9 10 11 12 14 15 ...  
## $ borough : int 4 4 4 4 4 4 4 4 4 4 ...  
## $ neighborhood : Factor w/ 58 levels "AIRPORT LA GUARDIA ",..: 1 1 1 1 2 2 2 2 2 2 ...  
## $ building.class.category : Factor w/ 27 levels "01 ONE FAMILY DWELLINGS ",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ tax.class.at.present : Factor w/ 5 levels "1","2","2A","2B",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ block : int 976 976 976 976 15829 15830 15831 15837 15849 15905 ...  
## $ lot : int 10 11 61 63 22 20 24 73 22 92 ...  
## $ ease.ment : logi NA NA NA NA NA NA ...  
## $ building.class.at.present : Factor w/ 87 levels "A0","A1","A2",..: 5 5 5 5 2 2 3 5 6 5 ...  
## $ address : Factor w/ 10087 levels "1-30 BEACH 134 STREET ",..: 4565 4568 4586 4576 6007 5917 6005 5725 6302 2804 ...  
## $ apartment.number : Factor w/ 7 levels " ",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ zip.code : int 11370 11370 11370 11370 11691 11691 11691 11691 11691 11692 ...  
## $ residential.units : Factor w/ 68 levels " - "," 1 ",..: 2 2 2 2 2 2 2 2 2 2 ...  
## $ commercial.units : Factor w/ 24 levels " - "," 1 ",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ total.units : num 1 1 1 1 1 1 1 1 1 1 ...  
## $ land.square.feet : Factor w/ 2288 levels " - "," 1,000 ",..: 279 279 279 279 914 802 1466 884 892 1409 ...  
## $ gross.square.feet : Factor w/ 2317 levels " 1,000 "," 1,004 ",..: 181 181 181 181 414 410 516 410 2084 1292 ...  
## $ year.built : int 1950 1950 1950 1950 2005 2005 1930 2006 1920 2014 ...  
## $ tax.class.at.time.of.sale : int 1 1 1 1 1 1 1 1 1 1 ...  
## $ building.class.at.time.of.sale: Factor w/ 88 levels "A0","A1","A2",..: 5 5 5 5 2 2 3 5 6 5 ...  
## $ sale.price : Factor w/ 2035 levels " 1,000,000 ",..: 1421 1421 1536 650 741 888 855 890 300 1663 ...  
## $ sale.date : Factor w/ 274 levels "1/10/2017","1/11/2017",..: 140 140 217 57 235 40 176 263 160 29 ...  
## $ sale.price.n : num 600000 600000 660000 275500 310000 ...  
## $ gross.sqft : int 1224 1224 1224 1224 1480 1476 1600 1476 618 2515 ...  
## $ land.sqft : int 1800 1800 1800 1800 2670 2469 3804 2612 2626 3667 ...  
## $ outliers : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ missing\_Sqft : logi FALSE FALSE FALSE FALSE FALSE FALSE ...

# get the summary of the data  
summary(queens)

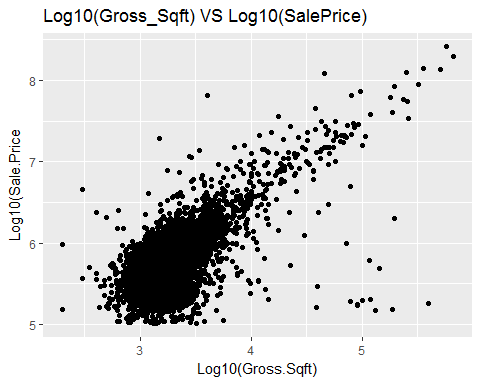
## X borough neighborhood   
## Min. : 1 Min. :4 FLUSHING-NORTH : 788   
## 1st Qu.: 6572 1st Qu.:4 BAYSIDE : 426   
## Median :14980 Median :4 RICHMOND HILL : 425   
## Mean :13917 Mean :4 SO. JAMAICA-BAISLEY PARK : 404   
## 3rd Qu.:21368 3rd Qu.:4 ST. ALBANS : 393   
## Max. :25787 Max. :4 SPRINGFIELD GARDENS : 379   
## (Other) :7484   
## building.class.category  
## 01 ONE FAMILY DWELLINGS :5514   
## 02 TWO FAMILY DWELLINGS :3329   
## 03 THREE FAMILY DWELLINGS : 631   
## 07 RENTALS - WALKUP APARTMENTS : 299   
## 22 STORE BUILDINGS : 151   
## 30 WAREHOUSES : 68   
## (Other) : 307   
## tax.class.at.present block lot ease.ment   
## 1 :9478 Min. : 13 Min. : 1.00 Mode:logical   
## 2 : 92 1st Qu.: 3939 1st Qu.: 18.00 NA's:10299   
## 2A: 271 Median : 8595 Median : 36.00   
## 2B: 42 Mean : 7894 Mean : 52.55   
## 4 : 416 3rd Qu.:11457 3rd Qu.: 59.00   
## Max. :16319 Max. :2687.00   
##   
## building.class.at.present  
## A1 :2626   
## A5 :1323   
## B3 :1241   
## B2 :1143   
## A2 : 984   
## B1 : 744   
## (Other):2238   
## address apartment.number  
## 184-03 GALWAY AVENUE : 3 :10293   
## 21-26 45TH AVENUE : 3 4 : 1   
## 25-34 JACKSON AVENUE : 3 5 : 1   
## 27-01 ASTORIA BOULEVARD : 3 54 : 1   
## 56-06 WALDRON STREET : 3 60 : 1   
## 87-23 160TH STREET : 3 OO4 : 1   
## (Other) :10281 RES : 1   
## zip.code residential.units commercial.units total.units   
## Min. : 0 1 :5465 0 :9362 Min. : 0.000   
## 1st Qu.:11365 2 :3194 1 : 507 1st Qu.: 1.000   
## Median :11411 3 : 642 - : 258 Median : 1.000   
## Mean :11390 0 : 353 2 : 71 Mean : 2.553   
## 3rd Qu.:11423 2 : 185 3 : 23 3rd Qu.: 2.000   
## Max. :11694 4 : 139 4 : 16 Max. :2261.000   
## (Other): 321 (Other): 62   
## land.square.feet gross.square.feet year.built   
## 4,000 :1189 1,224 : 117 Min. : 0   
## 2,500 : 778 1,280 : 103 1st Qu.:1925   
## 2,000 : 770 1,600 : 102 Median :1933   
## 3,000 : 446 1,440 : 92 Mean :1939   
## 1,800 : 271 1,152 : 89 3rd Qu.:1950   
## 5,000 : 239 1,200 : 88 Max. :2016   
## (Other):6606 (Other):9708   
## tax.class.at.time.of.sale building.class.at.time.of.sale sale.price   
## Min. :1.00 A1 :2629 650,000 : 139   
## 1st Qu.:1.00 A5 :1323 800,000 : 119   
## Median :1.00 B3 :1226 700,000 : 116   
## Mean :1.16 B2 :1136 450,000 : 113   
## 3rd Qu.:1.00 A2 : 993 750,000 : 110   
## Max. :4.00 B1 : 744 600,000 : 108   
## (Other):2248 (Other) :9594   
## sale.date sale.price.n gross.sqft land.sqft   
## 10/28/2016: 82 Min. : 101365 Min. : 200 Min. : 400   
## 12/22/2016: 77 1st Qu.: 440000 1st Qu.: 1280 1st Qu.: 2101   
## 6/30/2016 : 69 Median : 647900 Median : 1640 Median : 2800   
## 7/29/2016 : 69 Mean : 1029913 Mean : 2874 Mean : 3700   
## 9/30/2016 : 67 3rd Qu.: 875000 3rd Qu.: 2238 3rd Qu.: 4000   
## 11/22/2016: 66 Max. :257500000 Max. :656241 Max. :230400   
## (Other) :9869 NA's :1   
## outliers missing\_Sqft   
## Min. :0 Mode :logical   
## 1st Qu.:0 FALSE:10299   
## Median :0 NA's :0   
## Mean :0   
## 3rd Qu.:0   
## Max. :0   
##

## Queens data Visualization Gross\_Sqft VS SalePrice

#queens.sale$gross.sqft vs queens.sale$sale.price.n  
p <- ggplot(queens, aes(queens$gross.sqft, queens$sale.price.n))  
p + geom\_point()+ labs(title ="Gross\_Sqft VS SalePrice", x = "Gross.Sqft", y = "Sale.Price")

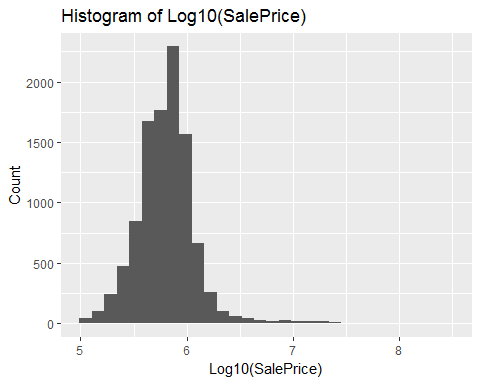


#log10(queens.sale$gross.sqft) vs log10(queens.sale$sale.price.n)  
y <- ggplot(queens, aes(log10(queens$gross.sqft), log10(queens$sale.price.n)))  
y + geom\_point()+labs(title ="Log10(Gross\_Sqft) VS Log10(SalePrice)", x = "Log10(Gross.Sqft)", y = "Log10(Sale.Price")

 ## Other Visualizations

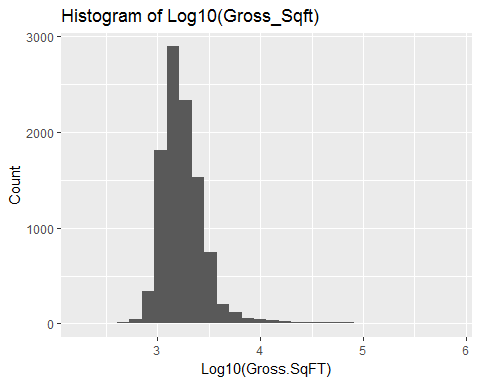
#Histogram of log saleprice  
ggplot(queens, aes(log10(queens$sale.price.n)))+ geom\_histogram()+  
 labs(title ="Histogram of Log10(SalePrice)", x = "Log10(SalePrice)", y = "Count")

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



#histogram of gross.sqft  
ggplot(queens, aes(log10(queens$gross.sqft)))+ geom\_histogram()+  
 labs(title ="Histogram of Log10(Gross\_Sqft)", x = "Log10(Gross.SqFT)", y = "Count")

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



sessionInfo()

## R version 3.3.2 (2016-10-31)  
## Platform: x86\_64-w64-mingw32/x64 (64-bit)  
## Running under: Windows 10 x64 (build 14393)  
##   
## locale:  
## [1] LC\_COLLATE=English\_United States.1252   
## [2] LC\_CTYPE=English\_United States.1252   
## [3] LC\_MONETARY=English\_United States.1252  
## [4] LC\_NUMERIC=C   
## [5] LC\_TIME=English\_United States.1252   
##   
## attached base packages:  
## [1] stats graphics grDevices utils datasets methods base   
##   
## other attached packages:  
## [1] plyr\_1.8.4 ggplot2\_2.2.1  
##   
## loaded via a namespace (and not attached):  
## [1] Rcpp\_0.12.8 assertthat\_0.1 digest\_0.6.10 rprojroot\_1.1   
## [5] grid\_3.3.2 gtable\_0.2.0 backports\_1.0.4 magrittr\_1.5   
## [9] evaluate\_0.10 scales\_0.4.1 stringi\_1.1.2 lazyeval\_0.2.0   
## [13] rmarkdown\_1.5 labeling\_0.3 tools\_3.3.2 stringr\_1.1.0   
## [17] munsell\_0.4.3 yaml\_2.1.14 colorspace\_1.3-1 htmltools\_0.3.5   
## [21] knitr\_1.16 tibble\_1.2