Project -1

Problem 3 – EDA on Rolling Sales Dataset

CSE – 587

Instructor – Bina Ramammurthy

Submitted By:-

NALIN KUMAR

Person - 50170479

Problem1

Explore its existing website, thinking about how buyers and sellers would navigate through it, and how the website is structured/organized. Try to understand the existing business model, and think about how analysis of RealDirect user-behavior data could be used to inform decision-making and product development. Come up with a list of research questions you think could be answered by data:

- **What data would you advise the engineers log and what would your ideal datasets look like?**
- **How would data be used for reporting and monitoring product usage?**
- **How would data be built back into the product/website?**

Solution

1) Data is the most important aspect in analysing the user behaviour and trends. Firstly, we can track the prices quoted by buyers and sellers in a particular area and average those prices for a particular geographical area. This would help us in determining the correct price for a property located in that area. Based on this information, we can help in negotiations between potential buyers and suppliers and quote both of them with correct well deserved price for that property. Also, based on this price information, we can send regular emails to our customers recommending them some other properties whose total worth lies near the correct price we have calculated in the previous step. In addition to this, we can grab some more personal information about the user such as his age, income, user budget etc. and in the future provide him with regular recommendations regarding properties aligning with his personal interests as well as budget. This will provide the user with a personalized experience. Secondly, we can track in which particular area most of the buying and selling activities are happening. We can get this information based on the data related to total cumulative sales happening in different areas and then shortlisting the best areas. Once we get this information, we can get an estimate on the total number of agents required to service Real Direct in a specific area based on this key information. Along with this, we can also appoint agents based on the types of customers. For example, properties in expensive parts of New York need extremely trained agents since these customers are estimated to perform the most expensive transactions. Appointing such agents will help us in maintain good customer trust which is very essential for the profitability of Real Direct. We need to collect these data in a structured manner so that we can easily perform analysis on such data. Our ideal dataset should be data collected in a JSON format which can easily be rendered on a plot to give us quick estimation of customer's trend and behaviour.

2) As stated in the previous solution, we can fetch data in JSON format which can easily be used to plot charts based on various libraries like Fusion Charts or graphs plotted using RShiny. These real time graphs can help us capture real time customer behaviour and trends by performing EDA and sentiment analysis on the collected data. Also, the real time data can be saved in some kind of Reports such as JRXML which organizes data into meaningful columns and analysis can easily be performed by any profession from the top management. These reports can depict what volume of sales were conducted in a particular area, who are the potential buyers and sellers for that week or month etc. We can generate such daily, biweekly, weekly reports containing user trends and implement companies decisions in

accordance with the user trends. This will also help us in automating various processes related to capturing customer trends

3) As stated in the previous solution, the data captured in some structured format can be built back into product website and can provide potential customers with some important information like the correct price for properties in specific areas, some analysis on how many people bought or sold properties in an area in the past one month and were they satisfied or not etc. These kinds of simple analysis will prompt the user to buy or sell properties in his/her desired area. And this analysis can easily be built back into the website if we automate the whole process of data capturing in a structured manner such as JSON format or JRXML format. Whatever it might be, the only focus should be to easily analyse user trends and store this data in such a manner that it can be meaningful to be built back into our product/website.

Problem2 and 3

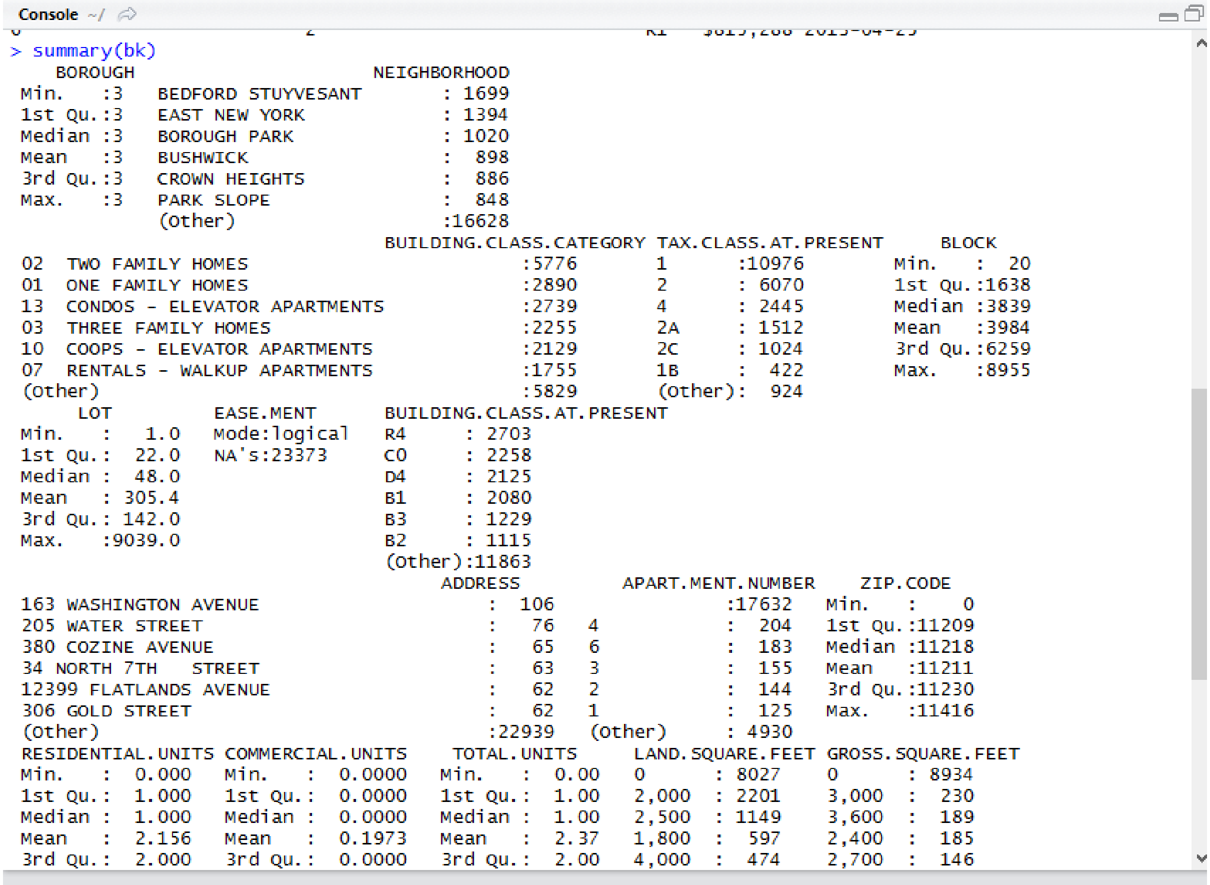
Load and clean up data from a single file as well as perform EDA after combining data from all the files(extended dataset). Compare the two analysis and based on the data and plots, recommend a report to the CEO of Real Direct which should contain solutions and suggestions based on your analysis

Solution

The data is manipulated in two parts. In the first part, a single file is read and EDA is performed on that single file data and saved into RD3nalinkum.R. In the next part, multiple files are read at once, EDA is performed on the combined data in all the files and the results are plotted in RDP3Exnalinkum.R. These two analysis are described as follows:-

RD3nalinkum.R

- Intially, I loaded the single file and read all of its data in a single variable named as bk
bk <- read.xls("D:/DIC/dds_datasets/rollingsales_brooklyn.xls",pattern="BOROUGH", perl = "C:\\Perl64\\bin\\perl.exe")
- Then, I analysed the summary of the loaded dataset from a single file named as rollingsales_brooklyn.xls which is shown as follows
head(bk)
summary(bk)



```
> summary(bk)
    BOROUGH                     NEIGHBORHOOD
 Min.   :3   BEDFORD STUYVESANT       : 1699
 1st Qu.:3   EAST NEW YORK            : 1394
 Median :3   BOROUGH PARK             : 1020
 Mean   :3   BUSHWICK                 :  898
 3rd Qu.:3   CROWN HEIGHTS            :  886
 Max.   :3   PARK SLOPE               :  848
             (Other)                  :16628
                             BUILDING.CLASS.CATEGORY TAX.CLASS.AT.PRESENT     BLOCK
 02   TWO FAMILY HOMES                   :5776        1      :10976        Min.   :  20
 01   ONE FAMILY HOMES                   :2890        2      : 6070        1st Qu.:1638
 13   CONDOS - ELEVATOR APARTMENTS       :2739        4      : 2445        Median :3839
 03   THREE FAMILY HOMES                 :2255        2A     : 1512        Mean   :3984
 10   COOPS - ELEVATOR APARTMENTS        :2129        2C     : 1024        3rd Qu.:6259
 07   RENTALS - WALKUP APARTMENTS        :1755        1B     :  422        Max.   :8955
 (Other)                                 :5829        (Other):  924
      LOT          EASE.MENT      BUILDING.CLASS.AT.PRESENT
 Min.   :   1.0   Mode:logical   R4     : 2703
 1st Qu.:  22.0   NA's:23373     C0     : 2258
 Median :  48.0                  D4     : 2125
 Mean   : 305.4                  B1     : 2080
 3rd Qu.: 142.0                  B3     : 1229
 Max.   :9039.0                  B2     : 1115
                                 (Other):11863
                                      ADDRESS           APART.MENT.NUMBER    ZIP.CODE
 163 WASHINGTON AVENUE                    :  106                 :17632   Min.   :    0
 205 WATER STREET                         :   76    4            :  204   1st Qu.:11209
 380 COZINE AVENUE                        :   65    6            :  183   Median :11218
 34 NORTH 7TH    STREET                   :   63    3            :  155   Mean   :11211
 12399 FLATLANDS AVENUE                   :   62    2            :  144   3rd Qu.:11230
 306 GOLD STREET                          :   62    1            :  125   Max.   :11416
 (Other)                                  :22939    (Other)      : 4930
 RESIDENTIAL.UNITS COMMERCIAL.UNITS     TOTAL.UNITS       LAND.SQUARE.FEET GROSS.SQUARE.FEET
 Min.   :  0.000   Min.   :  0.0000   Min.   :  0.00   0        : 8027   0        : 8934
 1st Qu.:  1.000   1st Qu.:  0.0000   1st Qu.:  1.00   2,000    : 2201   3,000    :  230
 Median :  1.000   Median :  0.0000   Median :  1.00   2,500    : 1149   3,600    :  189
 Mean   :  2.156   Mean   :  0.1973   Mean   :  2.37   1,800    :  597   2,400    :  185
 3rd Qu.:  2.000   3rd Qu.:  0.0000   3rd Qu.:  2.00   4,000    :  474   2,700    :  146
```

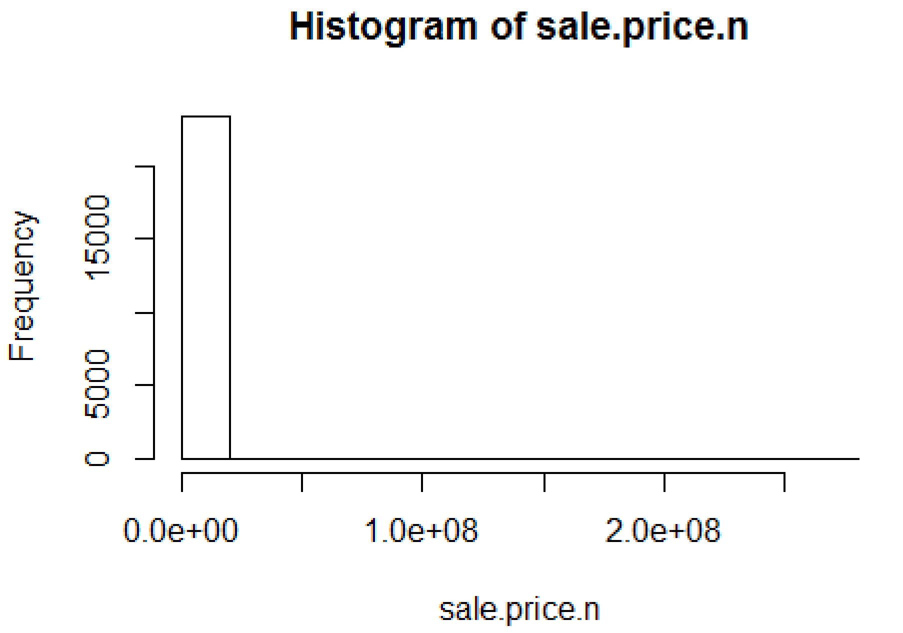
- In the next part, as mentioned in the book, I cleaned up the data using some regular expressions and converting some of the column values into numeric values as follows

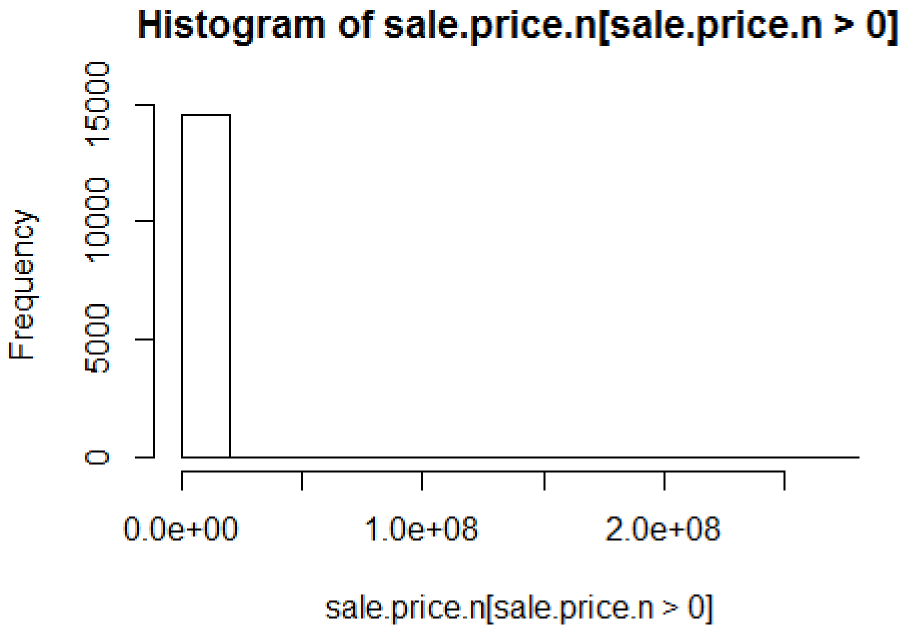
```
names(bk) <- tolower(names(bk))
names(bk)

bk$sale.price.n <- as.numeric(gsub("[^[:digit:]]","",
                          bk$sale.price))
bk$gross.sqft <- as.numeric(gsub("[^[:digit:]]","",
                         bk$gross.square.feet))
bk$land.sqft <- as.numeric(gsub("[^[:digit:]]","",
                        bk$land.square.feet))
bk$tax.class.at.time.of.sale <- as.numeric(gsub("[^[:digit:]]","",
                                   bk$tax.class.at.time.of.sale))
bk$sale.date <- as.Date(bk$sale.date)
bk$year.built <- as.numeric(as.character(bk$year.built))
bk$address <- as.character(bk$address)
bk$block <- as.numeric(as.character(bk$block))
bk$lot <- as.numeric(as.character(bk$lot))
bk$zip.code <- as.numeric(as.character(bk$zip.code))
```

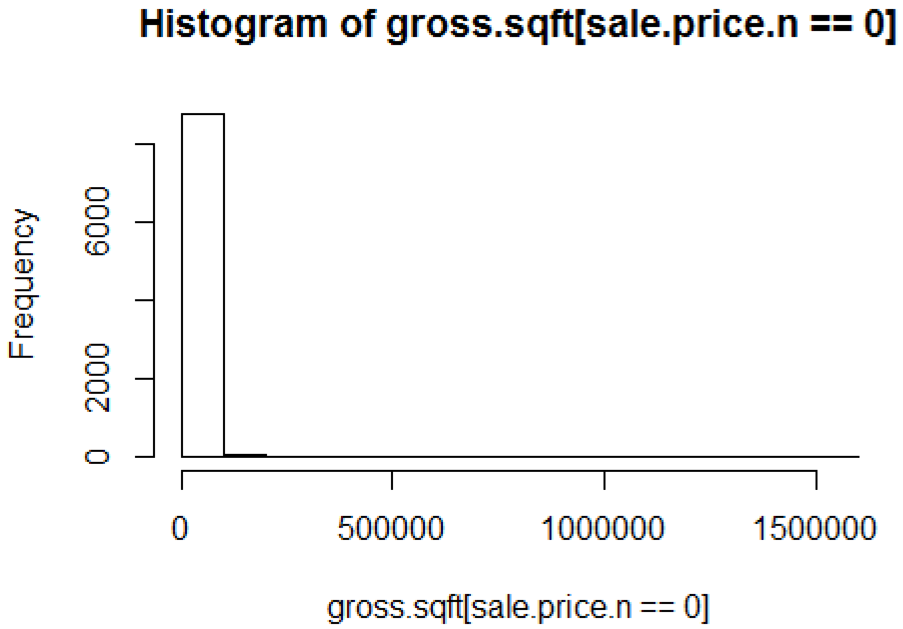
- Next, I verified whether there are no irregular variations in the new variable sale.price.n. I plotted different histograms for different values of sale.price.n and observed the variations

```
attach(bk)
```

```
hist(sale.price.n)
hist(sale.price.n[sale.price.n>0])
hist(gross.sqft[sale.price.n==0])
detach(bk)
```

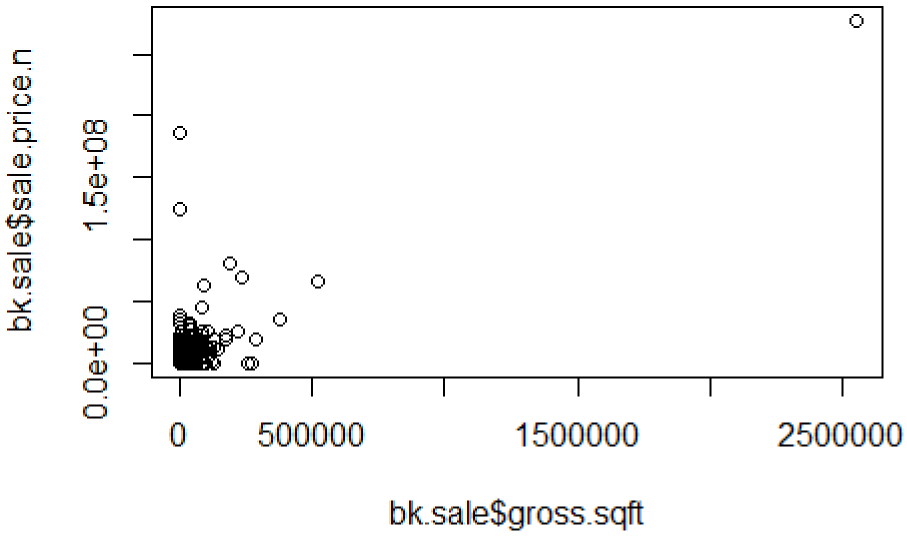


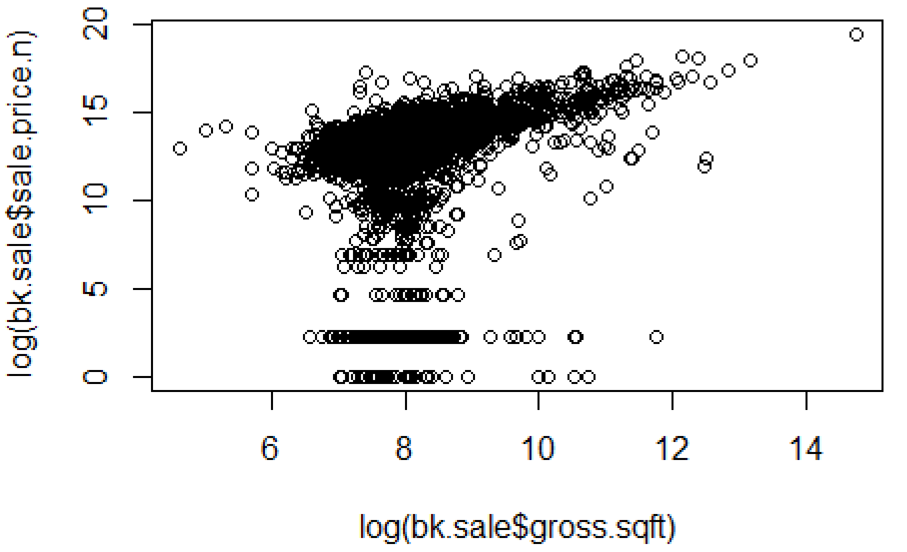




- Subsequently, I constructed 2 plots of sale price vs gross square feet and log(sale price) vs log(gross square feet) for only those values of sale price for which sale price >0. As can be seen from the plots below, the logarithmic plot is more clearer in the sense that the values are evenly scattered whereas in the non-logarithmic plot, the values are scattered only around a single point which does not give a true sense of the data as much as logarithmic plot gives

```
bk.sale <- bk[bk$sale.price.n!=0,]
plot(bk.sale$gross.sqft,bk.sale$sale.price.n)
plot(log(bk.sale$gross.sqft),log(bk.sale$sale.price.n))
```

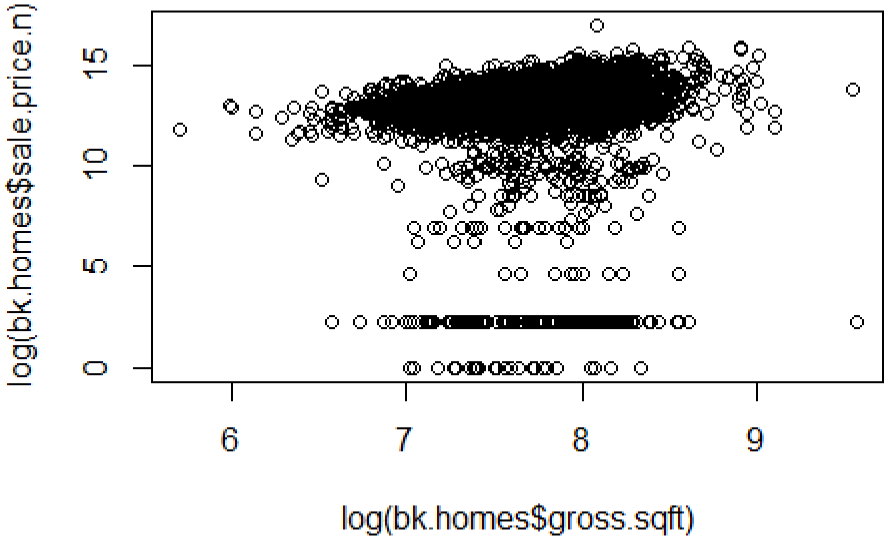




- Next, I only analysed dataset for which building class category is ‘FAMILY’ and sale price >0 (from previous part). Then, I constructed a plot of log(sale price) vs log(gross square feet). Finally, as mentioned in the textbook, I sorted the dataset for building class as ‘FAMILY’ based on the value of sale price where sale price lies in the range (0, 100000)

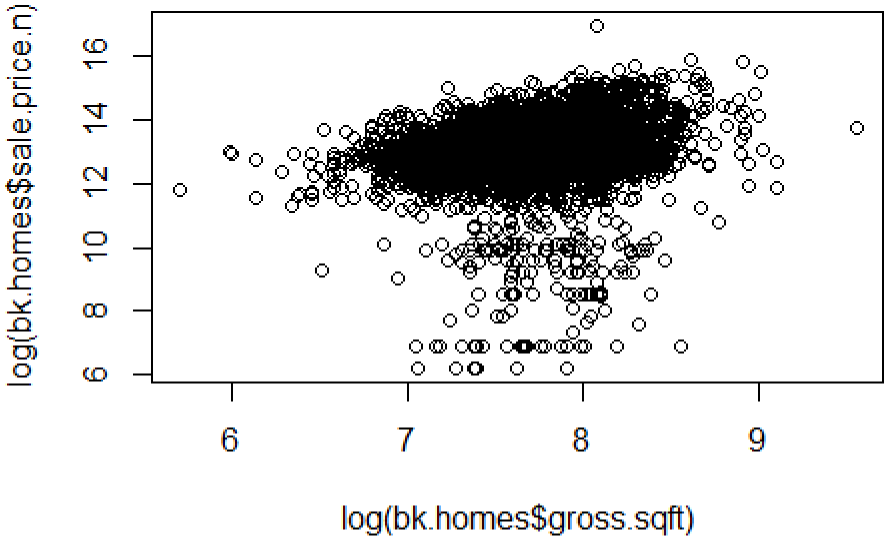
```
bk.homes <- bk.sale[which(grepl("FAMILY",
                       bk.sale$building.class.category)),]
plot(log(bk.homes$gross.sqft),log(bk.homes$sale.price.n))
bk.homes[which(bk.homes$sale.price.n<100000),]
```

```
order(bk.homes[which(bk.homes$sale.price.n<100000),]$sale.price.n)
```



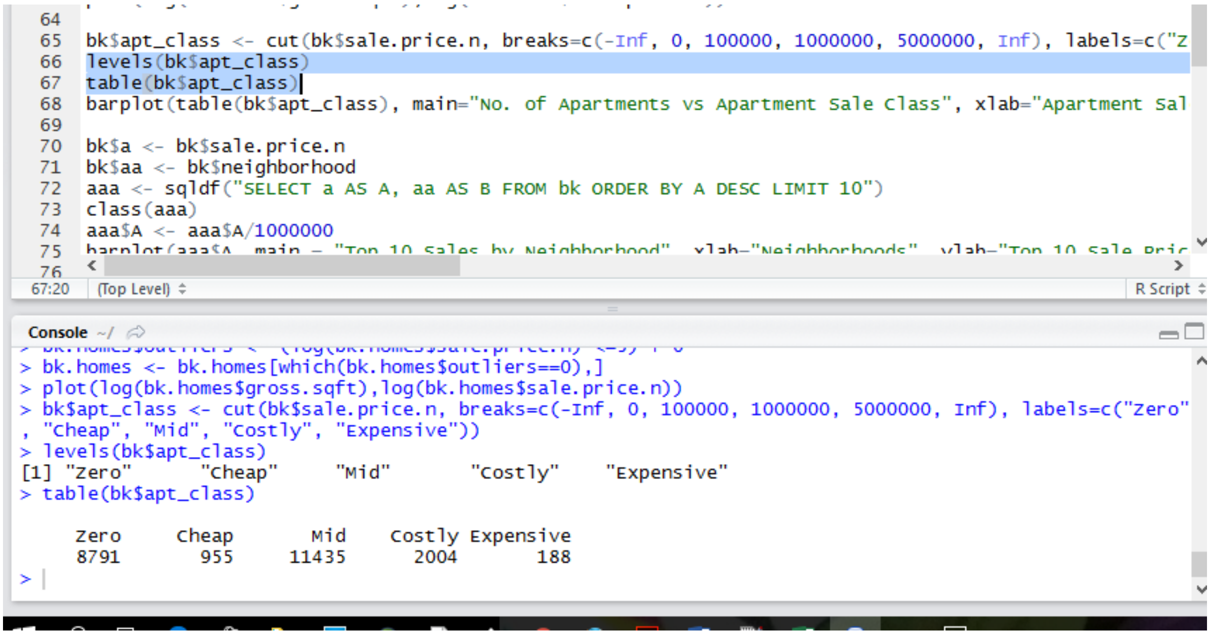
- In the next part, I trimmed certain portions of the dataset based on the value of newly declared variable named as outliers and finally plotted log(sale price) vs log(gross square feet) as follows

```
bk.homes$outliers <- (log(bk.homes$sale.price.n) <=5) + 0
bk.homes <- bk.homes[which(bk.homes$outliers==0),]
plot(log(bk.homes$gross.sqft),log(bk.homes$sale.price.n))
```



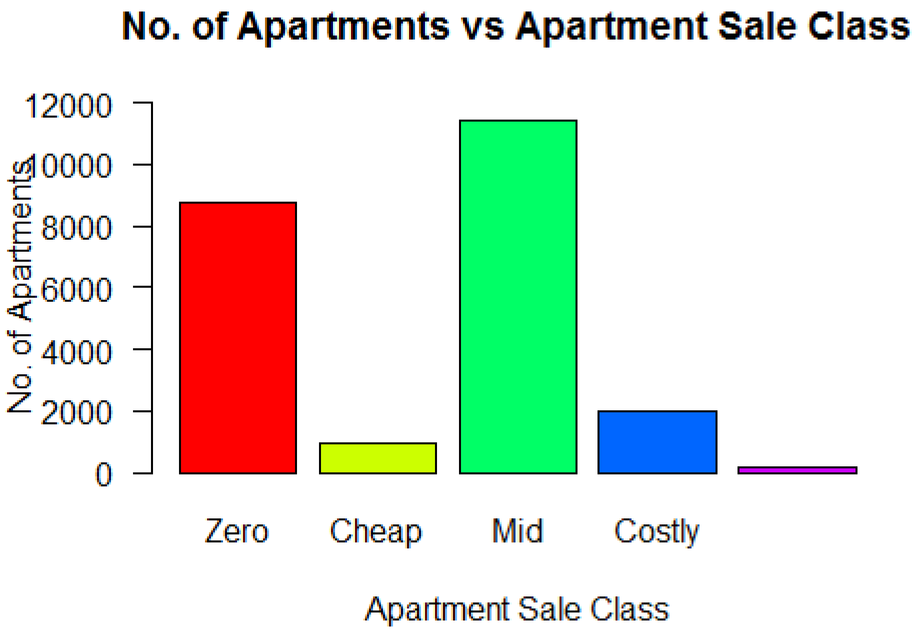
- The problem statement in the book suggests to plot and analyse based on neighbourhoods and time. Finally, we need to submit a report to the CEO of RealDirect suggesting our recommendations based on the relevant data and plot. For this, I constructed several plots which I would like to highlight one by one. In the first plot, I declared a categorical variable named as apt_class which divides all the sales into different categories like Zero, Cheap, Mid, Costly, Expensive etc. based on the value of sale price in the dataset. Subsequent to this, I made a bar plot in order to analyse of all the total apartments sold, how many of these sales belonged to which category. This report will help the CEO in understanding how many sales belonged to which of the sale classes outlined below and then he can focus on any particular sale class in the future. As can be seen, most of the apartments were sold in the Mid class category

```
bk$apt_class <- cut(bk$sale.price.n, breaks=c(-Inf, 0, 100000, 1000000, 5000000, Inf), labels=c("Zero", "Cheap", "Mid", "Costly", "Expensive"))
levels(bk$apt_class)
table(bk$apt_class)
barplot(table(bk$apt_class), main="No. of Apartments vs Apartment Sale Class", xlab="Apartment Sale Class", ylab="No. of Apartments", col=rainbow(5), ylim=c(0,12000), las=1)
```



```
bk$apt_class <- cut(bk$sale.price.n, breaks=c(-Inf, 0, 100000, 1000000, 5000000, Inf), labels=c("Z
levels(bk$apt_class)
table(bk$apt_class)
barplot(table(bk$apt_class), main="No. of Apartments vs Apartment Sale Class", xlab="Apartment Sal

bk$a <- bk$sale.price.n
bk$aa <- bk$neighborhood
aaa <- sqldf("SELECT a AS A, aa AS B FROM bk ORDER BY A DESC LIMIT 10")
class(aaa)
aaa$A <- aaa$A/1000000
barplot(aaa$A, main = "Top 10 Sales by Neighborhood", xlab="Neighborhoods", ylab="Top 10 Sale Pric
```

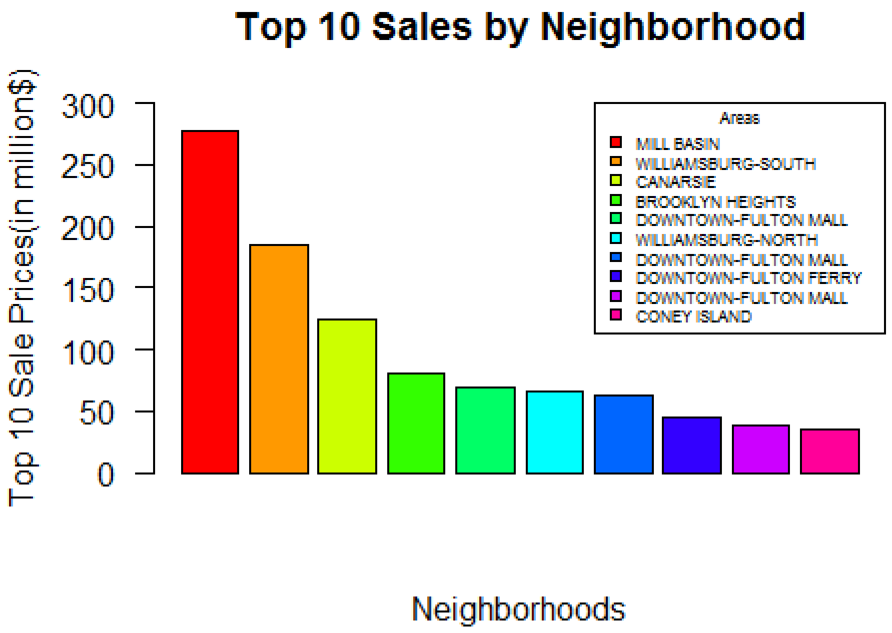
```
> bk.homes <- bk.homes[which(bk.homes$outliers==0),]
> plot(log(bk.homes$gross.sqft),log(bk.homes$sale.price.n))
> bk$apt_class <- cut(bk$sale.price.n, breaks=c(-Inf, 0, 100000, 1000000, 5000000, Inf), labels=c("Zero"
, "Cheap", "Mid", "Costly", "Expensive"))
> levels(bk$apt_class)
[1] "Zero"      "Cheap"     "Mid"       "Costly"    "Expensive"
> table(bk$apt_class)

     Zero     Cheap       Mid    Costly Expensive 
     8791       955     11435      2004       188 
> 
```



- In the next analysis, I grabbed the top 10 sales by neighbourhoods and made a bar plot between 'Top 10 Sale Prices' vs 'Neighborhoods'. This analysis will help the CEO in visualizing the fact that which are most prominent neighborhoods where the most expensive sales took place. As can be seen, neighborhoods where most prominent sales took place were 'Mill Basin', 'Williamsburg', 'Canarsie' etc.

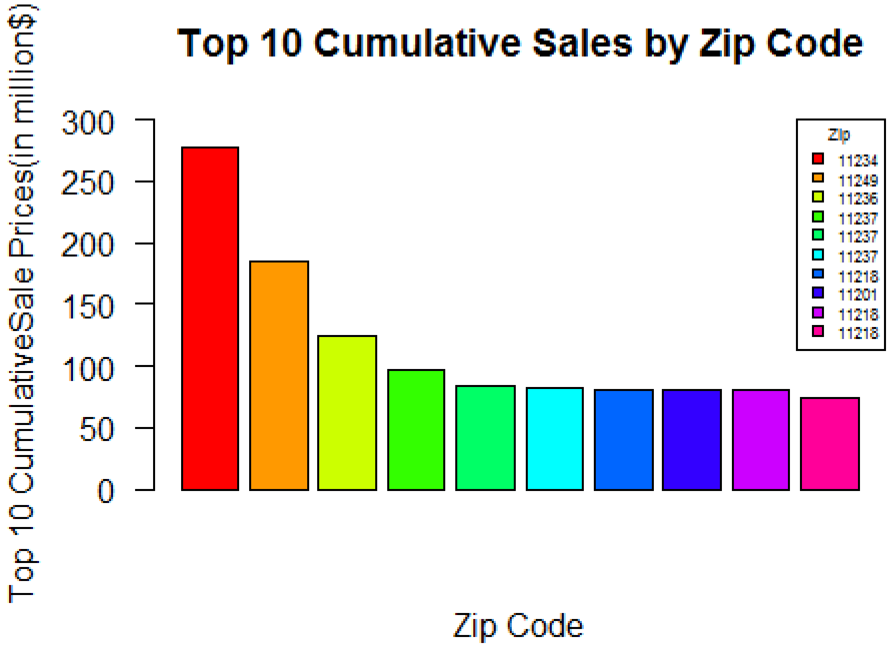
bk$a <- bk$sale.price.n
bk$aa <- bk$neighborhood
aaa <- sqldf("SELECT a AS A, aa AS B FROM bk ORDER BY A DESC LIMIT 10")
class(aaa)
aaa$A <- aaa$A/1000000

```
barplot(aaa$A, main = "Top 10 Sales by Neighborhood", xlab="Neighborhoods", ylab="Top 10 Sale Prices(in million$)",ylim=c(0,300), las=2, col=rainbow(10), beside=TRUE, legend = c(aaa$B), args.legend = list(title = "Areas", x = "topright", cex = .5))
```



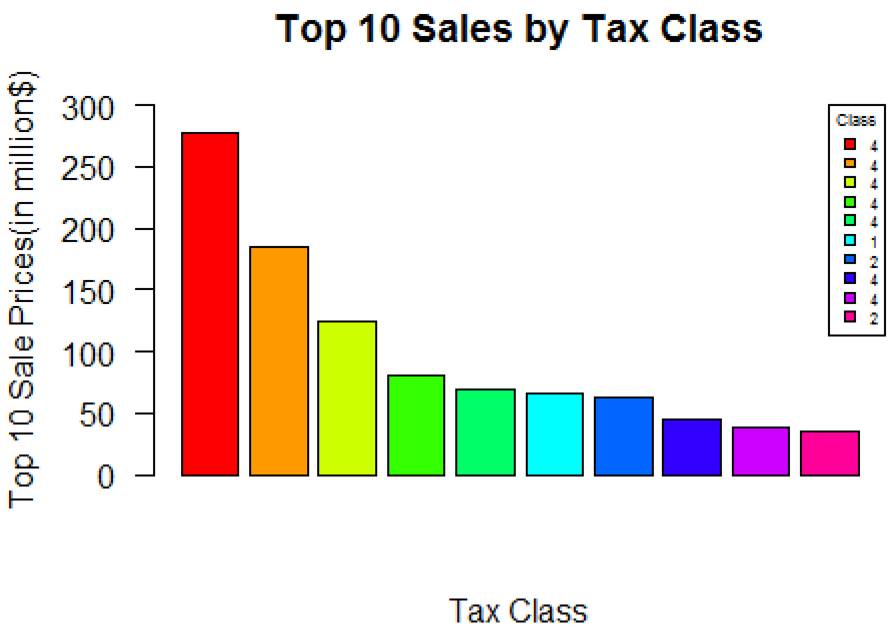
- In the next analysis, I grabbed the top 10 sales by zip-codes and made a bar plot between ‘Top 10 Sale Prices’ vs ‘Zip Codes’. This analysis will help the CEO in visualizing the fact that in which zip-codes the most expensive sales took place. In this manner, he can shortlist certain zip-codes and focus only on these. As can be seen, zip-codes where most prominent sales took place were 11234, 11249 etc.

```
bk$b <- bk$sale.price.n
bk$bb <- bk$zip.code
bbb <- sqldf("SELECT b AS A, bb AS B FROM bk GROUP BY B ORDER BY A DESC LIMIT 10")
class(bbb)
bbb$A <- bbb$A/1000000
barplot(bbb$A, main = "Top 10 Cumulative Sales by Zip Code", xlab="Zip Code", ylab="Top 10 CumulativeSale Prices(in million$)",ylim=c(0,300), las=2, col=rainbow(10), beside=TRUE, legend = c(bbb$B), args.legend = list(title = "Zip", x = "topright", cex = .5))
```



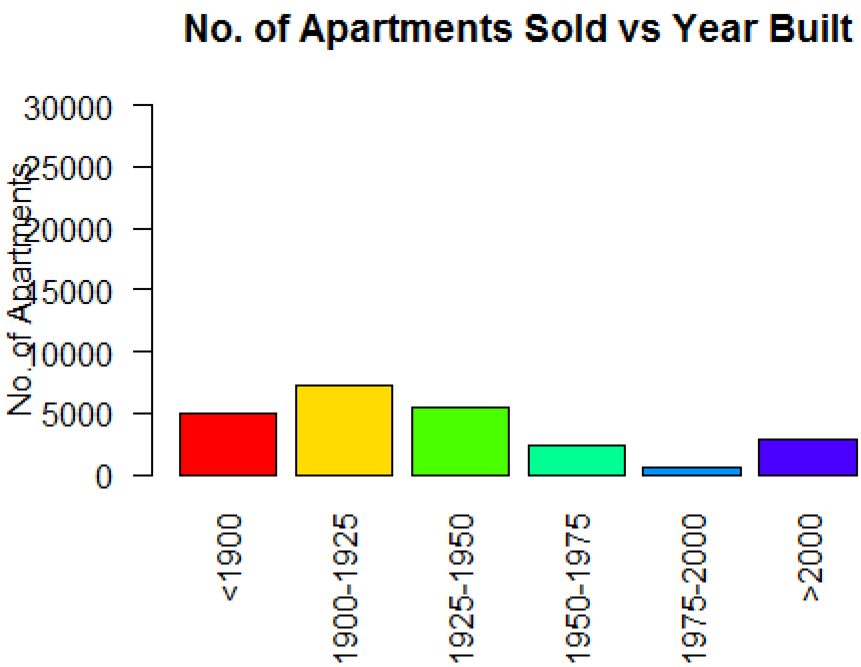
- In the next analysis, I grabbed the top 10 sales by tax classes at time of sale and made a bar plot between 'Top 10 Sale Prices' vs 'Tax Class'. This analysis will help the CEO in visualizing the fact that in which tax class the most expensive sales took place. In this manner, he can shortlist certain tax classes and focus only on these. As can be seen, tax classes where most prominent sales took place was '4'.

```
bk$c <- bk$sale.price.n
bk$cc <- bk$tax.class.at.time.of.sale
ccc <- sqldf("SELECT c AS A, cc AS B FROM bk ORDER BY A DESC LIMIT 10")
class(ccc)
ccc$A <- ccc$A/1000000
barplot(ccc$A, main = "Top 10 Sales by Tax Class", xlab="Tax Class", ylab="Top 10 Sale Prices(in million$)",ylim=c(0,300), las=2, col=rainbow(10), beside=TRUE, legend = c(ccc$B), args.legend = list(title = "Class", x = "topright", cex = .5))
```



- Subsequently, I defined a new categorical variable named as apt_year_built and tried to visualize out of all the apartments sold, how many of these were built in which year groups. As per apt_year_built, whole data was split into different categories based on breaks=c(-Inf,1900,1925,1950,1975,2000,Inf). Finally, I constructed a plot of Number of Apartments Sold vs Year Group to which it belonged i.e. it was built in which year group. This analysis will help the CEO in visualizing what number of apartments were sold in each year group. As can be seen, most of the apartments were sold out in the year group 1900-1925

```
bk$apt_year_built <- cut(bk$year.built, breaks=c(-Inf,1900,1925,1950,1975,2000,Inf), labels=c("<1900", "1900-1925", "1925-1950", "1950-1975", "1975-2000", ">2000"))
levels(bk$apt_year_built)
table(bk$apt_year_built)
barplot(table(bk$apt_year_built), main="No. of Apartments Sold vs Year Built", ylab="No. of Apartments", col=rainbow(7), ylim=c(0,30000), las=2)
```



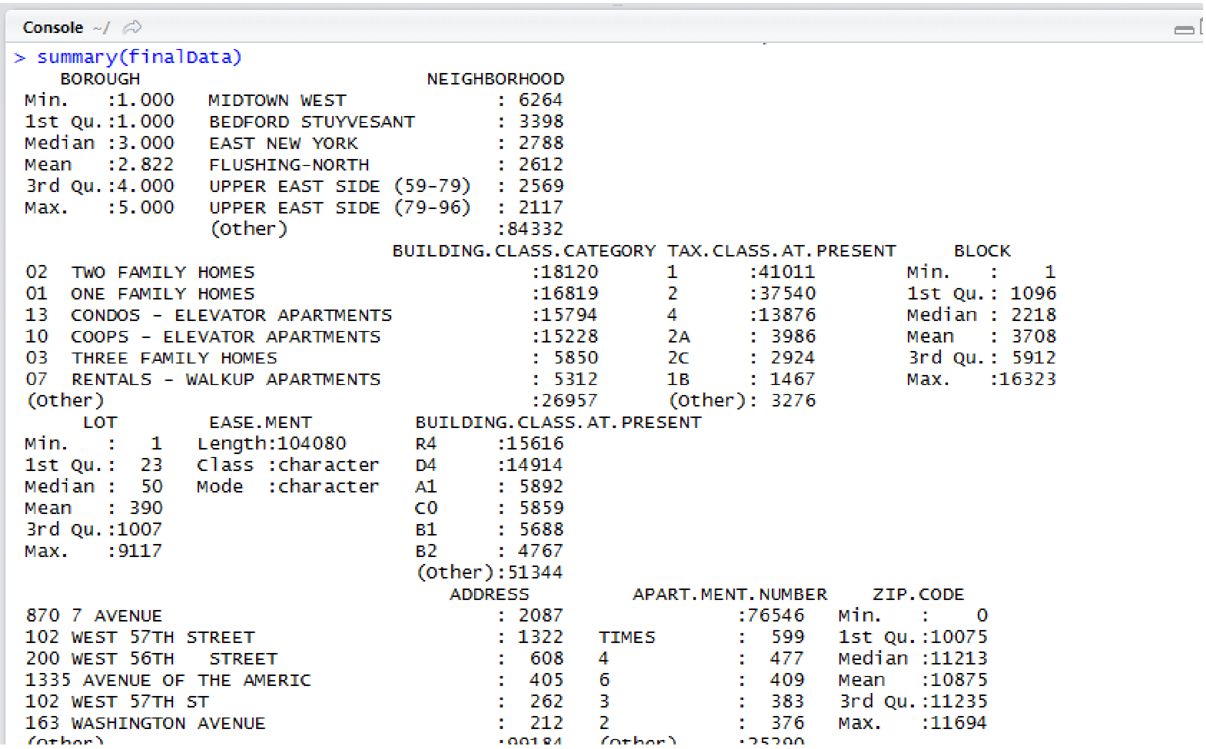
RDP3Exnalinkum.R

- Intially, I loaded and stored the dataset inside all the files in a single variable finalData using a for loop as done in Problem2 (Code snippet below)

```
fileDir <- 'D:/dic_data/problem3/'
fileDir1 <- 'D:/dic_data/problem3/rollingsales_brooklyn.xls'
files <- list.files(path = fileDir, pattern = "\\.xls$")
finalData <- read.xls(fileDir1,pattern="BOROUGH", perl = "C:\\Perl64\\bin\\perl.exe")


for(i in 2:5){
  tempDir = paste(fileDir,files[i],sep = "")
          loopFileData <- read.xls(tempDir,pattern="BOROUGH", perl =
"C:\\Perl64\\bin\\perl.exe")
  finalData = rbind(finalData, loopFileData)
}
```

- Then, I analysed the summary of the loaded dataset from all the files

```
head(finalData)
summary(finalData)
```



```
> summary(finalData)
    BOROUGH                         NEIGHBORHOOD
 Min.   :1.000   MIDTOWN WEST             : 6264
 1st Qu.:1.000   BEDFORD STUYVESANT       : 3398
 Median :3.000   EAST NEW YORK            : 2788
 Mean   :2.822   FLUSHING-NORTH           : 2612
 3rd Qu.:4.000   UPPER EAST SIDE (59-79)  : 2569
 Max.   :5.000   UPPER EAST SIDE (79-96)  : 2117
                 (Other)                  :84332
                                      BUILDING.CLASS.CATEGORY TAX.CLASS.AT.PRESENT     BLOCK
 02  TWO FAMILY HOMES                             :18120      1      :41011        Min.   :    1
 01  ONE FAMILY HOMES                             :16819      2      :37540        1st Qu.: 1096
 13  CONDOS - ELEVATOR APARTMENTS                 :15794      4      :13876        Median : 2218
 10  COOPS - ELEVATOR APARTMENTS                  :15228      2A     : 3986        Mean   : 3708
 03  THREE FAMILY HOMES                           : 5850      2C     : 2924        3rd Qu.: 5912
 07  RENTALS - WALKUP APARTMENTS                  : 5312      1B     : 1467        Max.   :16323
 (Other)                                          :26957      (Other): 3276
      LOT          EASE.MENT          BUILDING.CLASS.AT.PRESENT
 Min.   :   1   Length:104080       R4     :15616
 1st Qu.:  23   Class :character    D4     :14914
 Median :  50   Mode  :character    A1     : 5892
 Mean   : 390                       C0     : 5859
 3rd Qu.:1007                       B1     : 5688
 Max.   :9117                       B2     : 4767
                                    (Other):51344
                                         ADDRESS          APART.MENT.NUMBER     ZIP.CODE
 870 7 AVENUE                                : 2087                  :76546   Min.   :    0
 102 WEST 57TH STREET                        : 1322   TIMES          :   599   1st Qu.:10075
 200 WEST 56TH    STREET                     :  608   4              :   477   Median :11213
 1335 AVENUE OF THE AMERIC                   :  405   6              :   409   Mean   :10875
 102 WEST 57TH ST                            :  262   3              :   383   3rd Qu.:11235
 163 WASHINGTON AVENUE                       :  212   2              :   376   Max.   :11694
 (Other)                                     :99184   (Other)        :25290
```

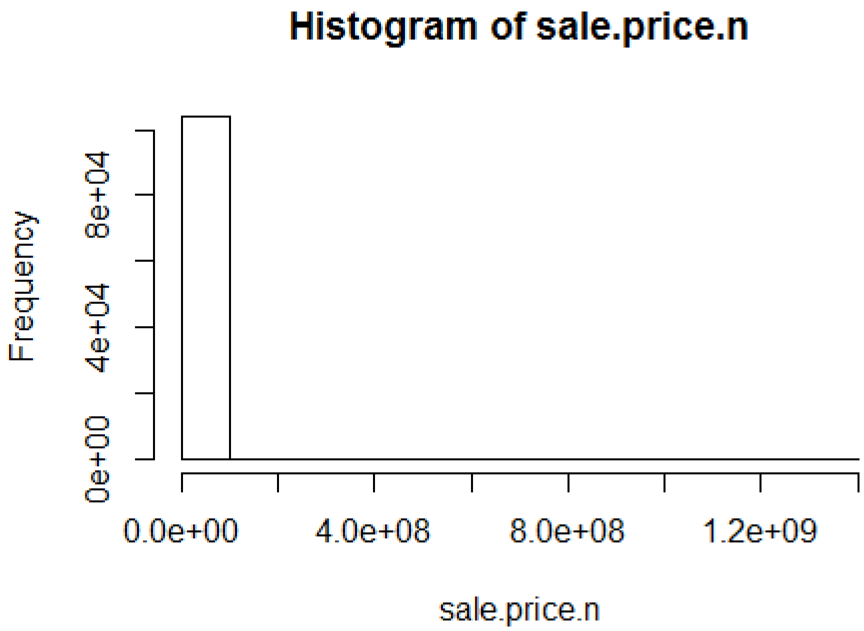
- In the next part, as mentioned in the book, I cleaned up the data using some regular expressions and converting some of the column values into numeric values as follows

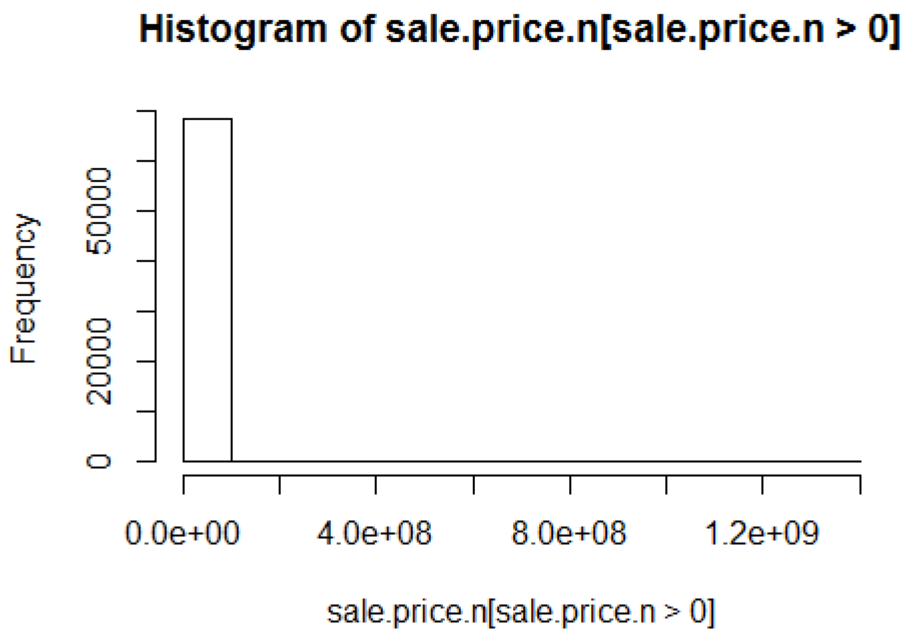
```
names(finalData) <- tolower(names(finalData))
names(finalData)
class(finalData$neighborhood)
class(finalData$block)
class(finalData$land.square.feet)
class(finalData$gross.square.feet)
class(finalData$sale.price)
class(finalData$sale.date)
class(finalData$zip.code)
class(finalData$address)
class(finalData$address)
class(finalData$tax.class.at.time.of.sale)
finalData$sale.price.n <- as.numeric(gsub("[^[:digit:]]","",
                          finalData$sale.price))
finalData$gross.sqft <- as.numeric(gsub("[^[:digit:]]","",
                         finalData$gross.square.feet))
finalData$land.sqft <- as.numeric(gsub("[^[:digit:]]","",
                        finalData$land.square.feet))
finalData$tax.class.at.time.of.sale <- as.numeric(gsub("[^[:digit:]]","",
                          finalData$tax.class.at.time.of.sale))
finalData$sale.date <- as.Date(finalData$sale.date)
finalData$year.built <- as.numeric(as.character(finalData$year.built))
finalData$address <- as.character(finalData$address)
finalData$block <- as.numeric(as.character(finalData$block))
```

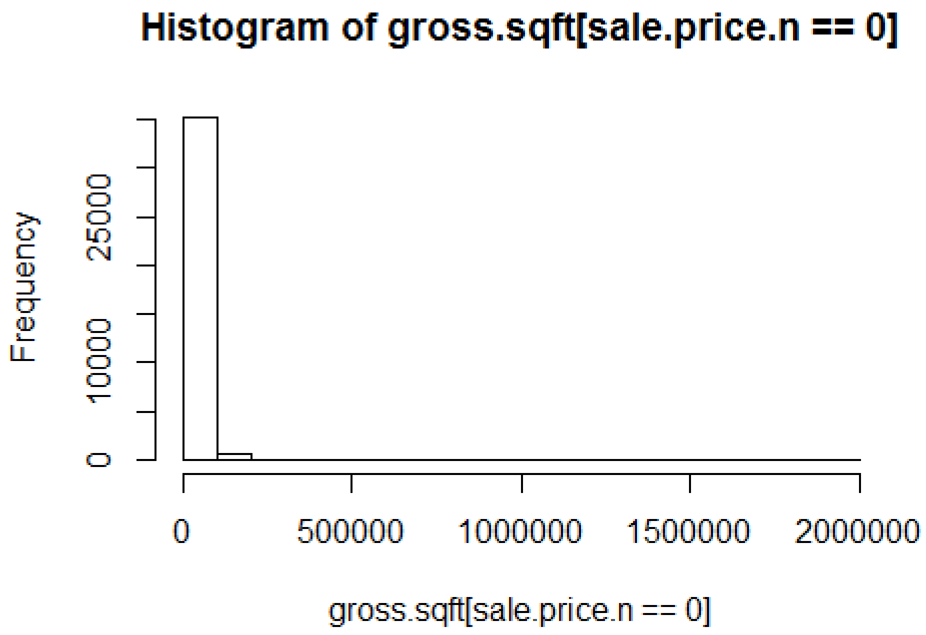
```
finalData$lot <- as.numeric(as.character(finalData$lot))
finalData$zip.code <- as.numeric(as.character(finalData$zip.code))

class(finalData$sale.price.n)
class(finalData$gross.sqft)
class(finalData$land.sqft)
```

- Next, I verified whether there are no irregular variations in the new variable sale.price.n. I plotted different histograms for different values of sale.price.n and observed the variations. The variations are almost the same as in the first part for a single file dataset

```
attach(finalData)
hist(sale.price.n)
hist(sale.price.n[sale.price.n>0])
hist(gross.sqft[sale.price.n==0])
detach(finalData)
```



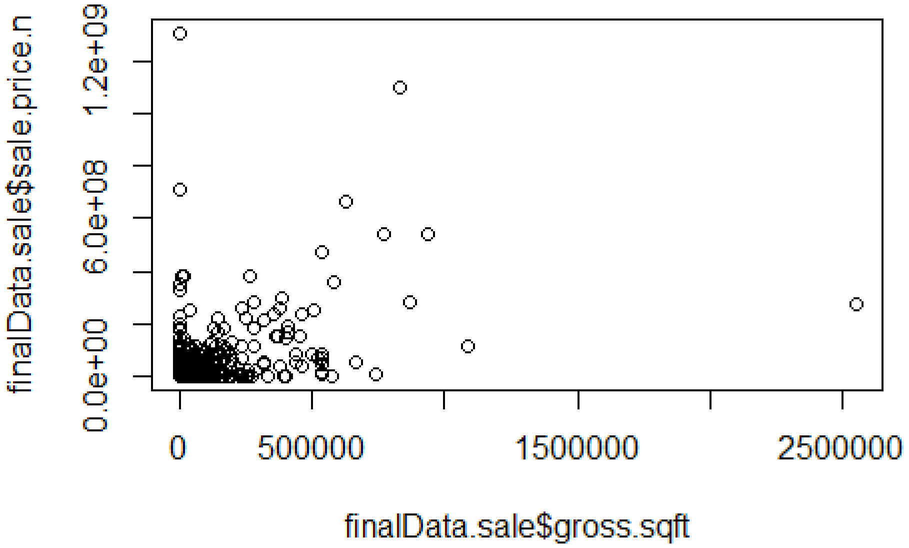


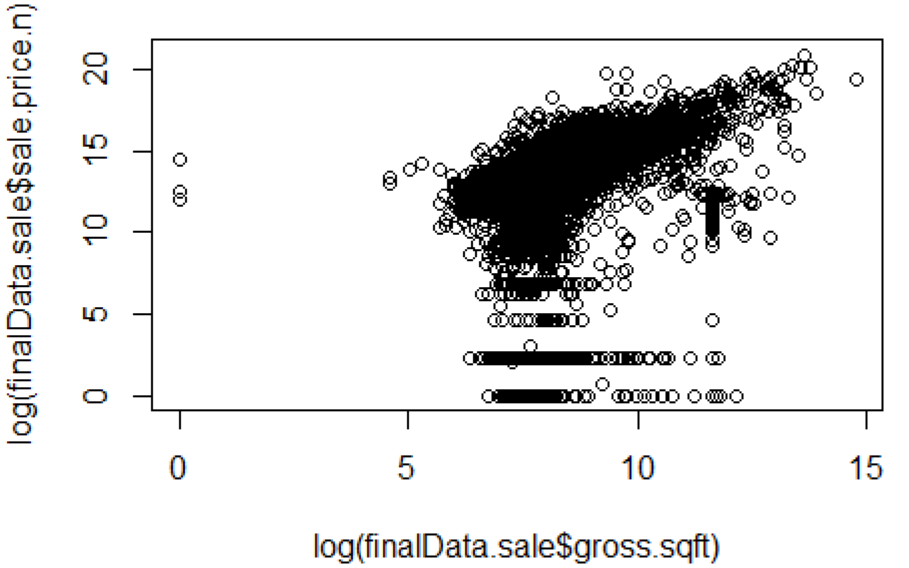


- Subsequently, I constructed 2 plots of sale price vs gross square feet and log(sale price) vs log(gross square feet) for only those values of sale price for which sale price >0. As can be seen from the plots below, the logarithmic plot is more clearer in the sense that the values are evenly scattered whereas in the non-logarithmic plot, the values are scattered only around a single point which does not give a true sense of the data as much as logarithmic plot gives. The plot we obtained is almost the same as obtained for a single file dataset

```
finalData.sale <- finalData[finalData$sale.price.n!=0,]
plot(finalData.sale$gross.sqft,finalData.sale$sale.price.n)
```

plot(log(finalData.sale$gross.sqft),log(finalData.sale$sale.price.n))

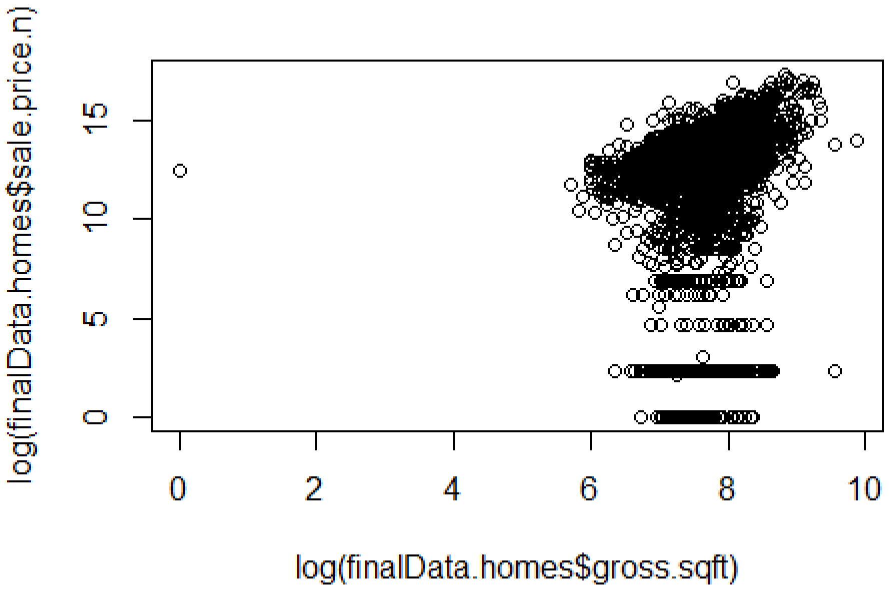




- Next, I only analysed dataset for which building class category is 'FAMILY' and sale price >0 (from previous part). Then, I constructed a plot of log(sale price) vs log(gross square feet). Finally, as mentioned in the textbook, I sorted the dataset for building class as 'FAMILY' based on the value of sale price where sale price lies in the range (0, 100000). The plot we obtained is almost the same as obtained for a single file dataset except that it is a bit more scattered towards right

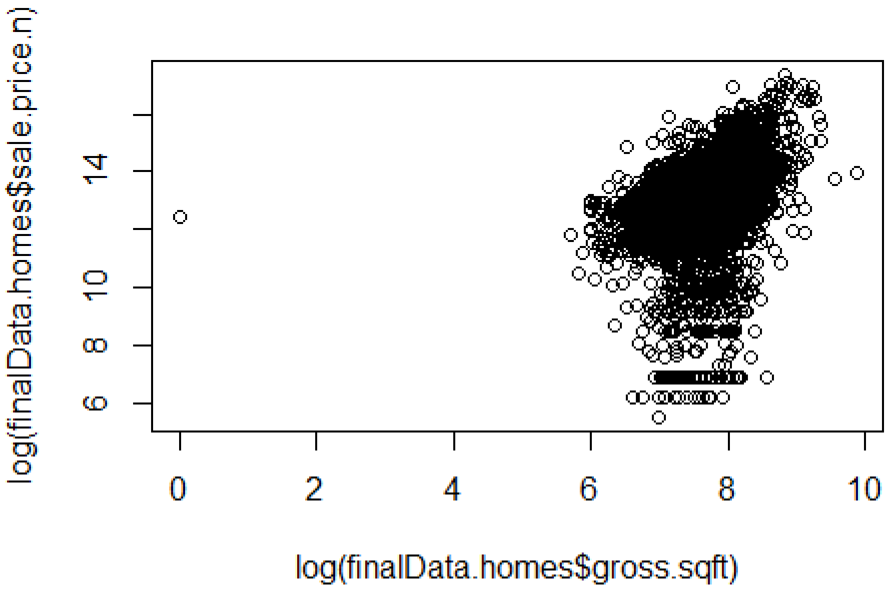
finalData.homes <- finalData.sale[which(grepl("FAMILY",

```
                      finalData.sale$building.class.category)),]
plot(log(finalData.homes$gross.sqft),log(finalData.homes$sale.price.n))
finalData.homes[which(finalData.homes$sale.price.n<100000),]
order(finalData.homes[which(finalData.homes$sale.price.n<100000),]$sale.price.n)
```



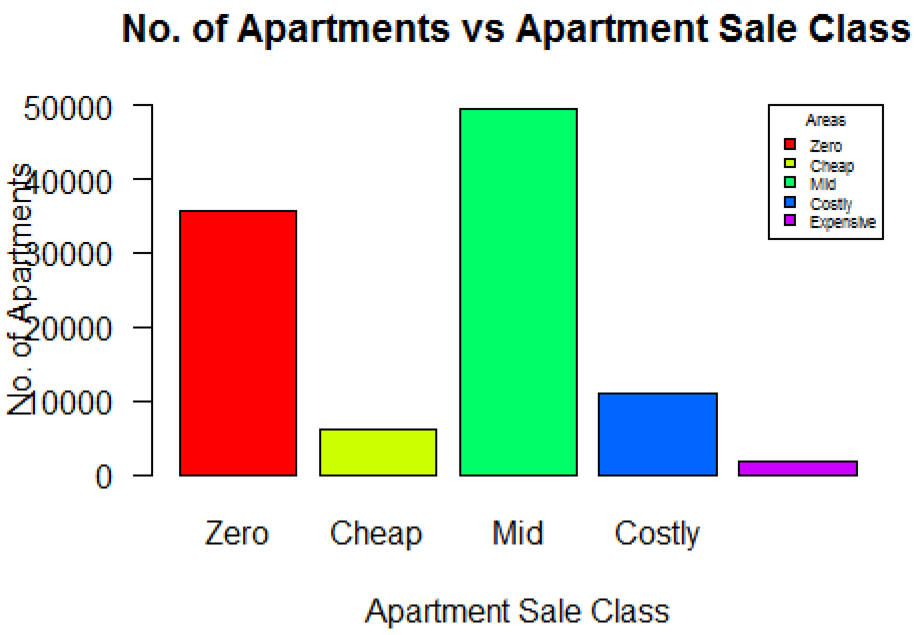
- In the next part, I trimmed certain portions of the dataset based on the value of newly declared variable named as outliers and finally plotted log(sale price) vs log(gross square feet) as shown below. The plot we obtained is almost the same as obtained for a single file dataset except that it is a bit more scattered towards right

```
finalData.homes$outliers <- (log(finalData.homes$sale.price.n) <=5) + 0
finalData.homes <- finalData.homes[which(finalData.homes$outliers==0),]
plot(log(finalData.homes$gross.sqft),log(finalData.homes$sale.price.n))
```



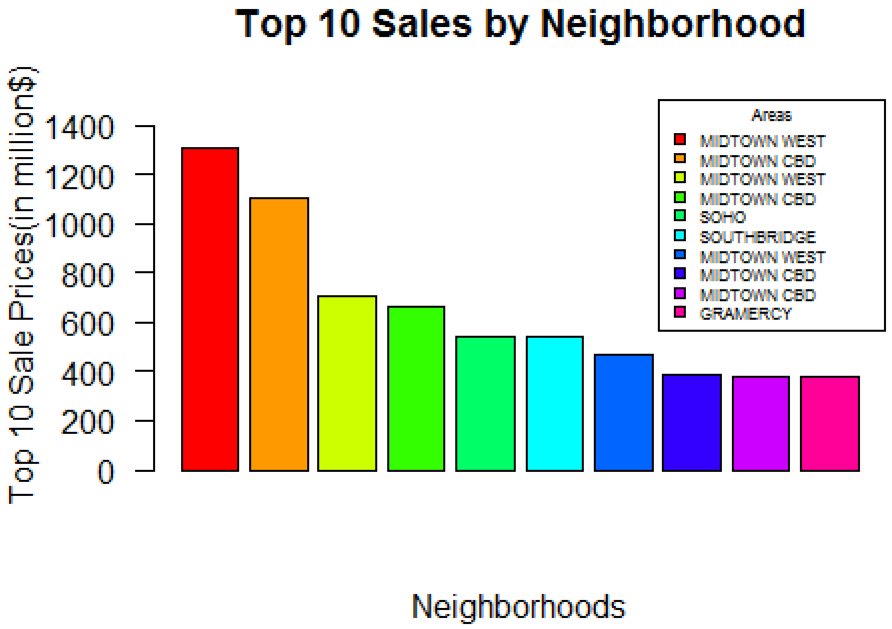
- The problem statement in the book suggests to plot and analyse based on neighbourhoods and time. Finally, we need to submit a report to the CEO of RealDirect suggesting our recommendations based on the relevant data and plot. For this, I constructed several plots for the extended dataset which I would like to highlight one by one. In the first plot, I declared a categorical variable named as apt_class which divides all the sales into different categories like Zero, Cheap, Mid, Costly, Expensive etc. based on the value of sale price in the dataset. Subsequent to this, I made a bar plot in order to analyse of all the total apartments sold, how many of these sales belonged to which category. This report will help the CEO in understanding how many sales belonged to which of the sale classes outlined below and then he can focus on any particular sale class in the future. As can be seen, most of the apartments were sold in the Mid class category similar to the previous part

```
finalData$apt_class <- cut(finalData$sale.price.n, breaks=c(-Inf, 0, 100000, 1000000, 5000000, Inf), labels=c("Zero", "Cheap", "Mid", "Costly", "Expensive"))
levels(finalData$apt_class)
table(finalData$apt_class)
barplot(table(finalData$apt_class), main="No. of Apartments vs Apartment Sale Class", xlab="Apartment Sale Class", ylab="No. of Apartments", col=rainbow(5), ylim=c(0,50000), las=1, beside=TRUE, legend = c(levels(finalData$apt_class)), args.legend = list(title = "Areas", x = "topright", cex = .5)
)
```



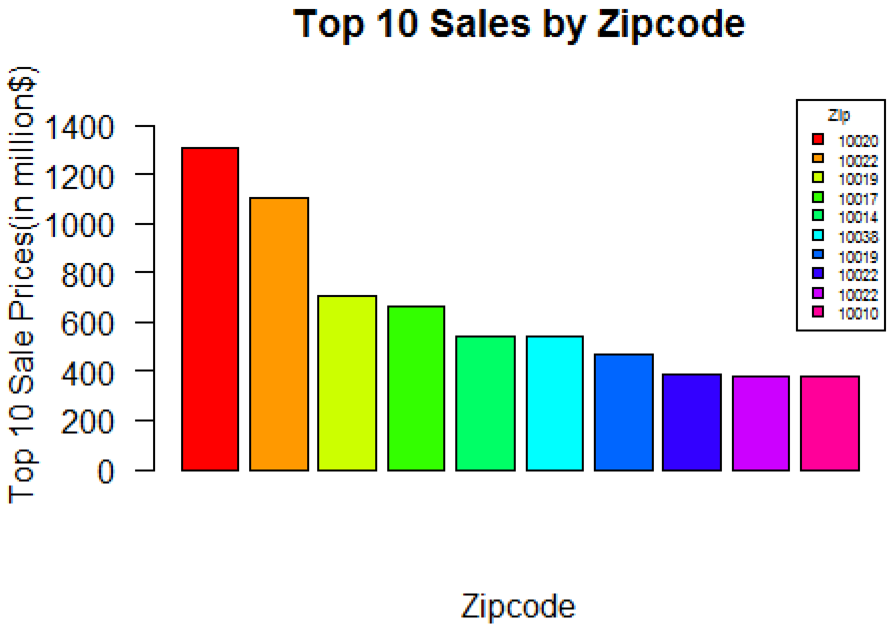
- In the next analysis, I grabbed the top 10 sales by neighbourhoods and made a bar plot between 'Top 10 Sale Prices' vs 'Neighborhoods'. This analysis will help the CEO in visualizing the fact that which are most prominent neighborhoods where the most expensive sales took place. As can be seen, neighborhoods where most prominent sales took place were 'MidTown West', 'MidTown CBD', 'SOHO' in the extended dataset unlike the previous part where we did an analysis on a single file

```
finalData$a <- finalData$sale.price.n
finalData$aa <- finalData$neighborhood
aaa <- sqldf("SELECT a AS A, aa AS B FROM finalData ORDER BY A DESC LIMIT 10")
class(aaa)
aaa$A <- aaa$A/1000000
barplot(aaa$A, main = "Top 10 Sales by Neighborhood", xlab="Neighborhoods", ylab="Top 10 Sale Prices(in million$)",ylim=c(0,1500), las=2, col=rainbow(10), beside=TRUE, legend = c(aaa$B), args.legend = list(title = "Areas", x = "topright", cex = .5))
```



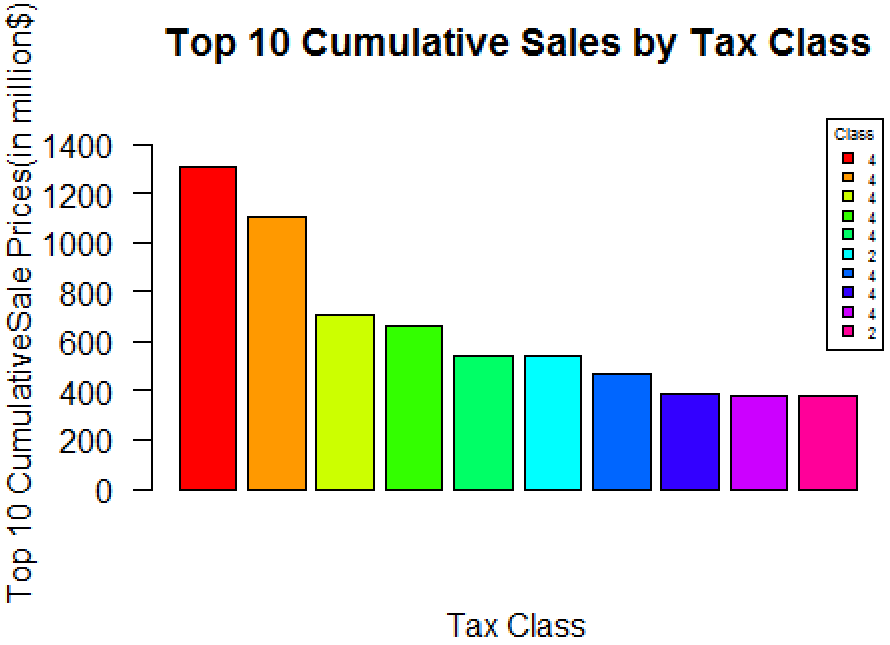
- In the next analysis, I grabbed the top 10 sales by zip-codes and made a bar plot between ‘Top 10 Sale Prices’ vs ‘Zip Codes’. This analysis will help the CEO in visualizing the fact that in which zip-codes the most expensive sales took place. In this manner, he can shortlist certain zip-codes and focus only on these. As can be seen, zip-codes where most prominent sales took place were 10020, 10022 etc. in the extended dataset unlike the previous part where we did an analysis on a single file

```
finalData$c <- finalData$sale.price.n
finalData$cc <- finalData$zip.code
ccc <- sqldf("SELECT c AS A, cc AS B FROM finalData ORDER BY A DESC LIMIT 10")
class(ccc)
ccc$A <- ccc$A/1000000
barplot(ccc$A, main = "Top 10 Sales by Zipcode", xlab="Zipcode", ylab="Top 10 Sale Prices(in million$)",ylim=c(0,1500), las=2, col=rainbow(10), beside=TRUE, legend = c(ccc$B), args.legend = list(title = "Zip", x = "topright", cex = .5))
```



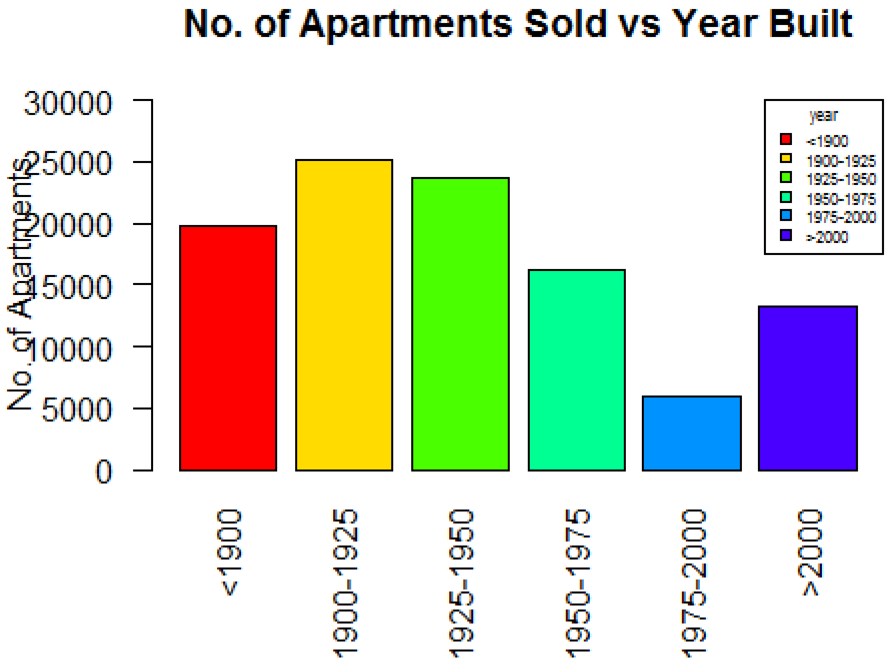
- In the next analysis, I grabbed the top 10 sales by tax classes at time of sale and made a bar plot between ‘Top 10 Sale Prices’ vs ‘Tax Class’. This analysis will help the CEO in visualizing the fact that in which tax class the most expensive sales took place. In this manner, he can shortlist certain tax classes and focus only on these. As can be seen, tax classes where most prominent sales took place was ‘4’ SOHO’ in the extended dataset similar to the previous part where we did an analysis on a single file

```
finalData$e <- finalData$sale.price.n
finalData$ee <- finalData$tax.class.at.time.of.sale
eee <- sqldf("SELECT e AS A, ee AS B FROM finalData ORDER BY A DESC LIMIT 10")
class(eee)
eee$A <- eee$A/1000000
barplot(eee$A, main = "Top 10 Cumulative Sales by Tax Class", xlab="Tax Class",
ylab="Top 10 CumulativeSale Prices(in million$)",ylim=c(0,1500), las=2, col=rainbow(10),
beside=TRUE, legend = c(eee$B), args.legend = list(title = "Class", x = "topright", cex = .5))
```



- Subsequently, I defined a new categorical variable named as apt_year_built and tried to visualize out of all the apartments sold, how many of these were built in which year groups. As per apt_year_built, whole data was split into different categories based on breaks=c(-Inf,1900,1925,1950,1975,2000,Inf). Finally, I constructed a plot of Number of Apartments Sold vs Year Group to which it belonged i.e. it was built in which year group. This analysis will help the CEO in visualizing what number of apartments were sold in each year group. As can be seen, most of the apartments were sold out in the year group 1900-1925 and 1925-1950 analogous to the previous part on a single file

```
finalData$apt_year_built <- cut(finalData$year.built, breaks=c(-Inf,1900,1925,1950,1975,2000,Inf), labels=c("<1900", "1900-1925", "1925-1950", "1950-1975", "1975-2000", ">2000"))
levels(finalData$apt_year_built)
table(finalData$apt_year_built)
barplot(table(finalData$apt_year_built), main="No. of Apartments Sold vs Year Built", ylab="No. of Apartments", col=rainbow(7), ylim=c(0,30000), las=2, beside=TRUE, legend = c(levels(finalData$apt_year_built)), args.legend = list(title = "year", x = "topright", cex = .5))
```



Conclusions to be provided to the CEO based on Report Analysis in Problem2

In order to display the potential of our recommendations to the CEO, we need to provide him with fruitful results based on the available dataset. As pointed out earlier, I plotted a number of graphs to draw some interesting observations from the data which are as follows:-

1) In the first graph, I initially categorized the whole dataset into different classes of sale based on the value of sale price. The results are then plotted with No. of apartments sold on the y-axis and their respective classes of sale on x-axis. When I performed EDA on a single file and extended dataset, I came to the conclusion that maximum sale occurred in the Mid Range of sale class which corresponds to the range (100000$, 1000000$).
2) In the next graph, I plotted Top 10 sales (in terms of individual sale value) on y-axis vs respective Neighborhoods on x-axis on a bar plot. This drew an interesting observation that neighborhoods where most prominent sales took place were 'MidTown West', 'MidTown CBD', 'SOHO' in the extended dataset and neighborhoods where most prominent sales took place were 'Mill Basin', 'Williamsburg', 'Canarsie' in the single Brooklyn file. This observation can allow the top management to focus more on these specific neighborhoods in the future
3) In the next graph, I plotted Top 10 sales (in terms of individual sale value) on y-axis vs respective zip-codes on x-axis on a bar plot. This drew an interesting observation that zip-codes where most prominent sales took place were '10020', '10022' in the extended dataset and zip-codes where most prominent sales took place were '11234', '11249' in the single Brooklyn file. This observation can allow the top management to focus more on these specific zip-codes in the future
4) In the next graph, I plotted Top 10 sales (in terms of individual sale value) on y-axis vs respective tax classes at time of sale on x-axis on a bar plot. This drew an interesting observation that tax class where most prominent sales took place was '4' in both the

extended dataset and in the single Brooklyn file. This observation can allow the top management to focus more on these specific tax classes in the future

5) In the final graph, I categorized the whole dataset into different year periods based on in which year the house was built. The results are then plotted with No. of apartments sold on the y-axis and their respective year built period on x-axis. When I performed EDA on a single file and extended dataset, I came to the conclusion that maximum sale occurred in the year built period '1900-1925'. This observation can allow the top management to focus more on these specific year built periods in the future

Problem4

Being the "data scientist" often involves speaking to people who aren't also data scientists, so it would be ideal to have a set of communication strategies for getting to the information you need about the data. Can you think of any other people you should talk to?
Solution

Obviously, the only communication strategy would be to prepare a detailed report after performing exploratory data analysis and displaying crucial observations to the concerned persons through a presentation. Besides data scientists, we need to effectively convey our ideas to the other members in the team like Product management, development, QA, operations. All these teams have a role to play in efficient implementation of data analysis. Product management can decide how to use the data to influence his/her decisions. Developers and QA need to understand how to implement data analysis in their current code and what user interface should be provided to the user after consuming the user behaviour data. Operations team can interact with the users based on their data and can provide a personalized experience. All in all, all of these team members are the ones with whom I need to interact, discuss and efficiently implement ideas based on user data

Problem5

Most of you are not "domain experts" in real estate or online businesses.

- **Does stepping out of your comfort zone and figuring out how you would go about "collecting data" in a different setting give you insight into how you do it in your own field?**
- **Sometimes "domain experts" have their own set of vocabulary. Did Doug use vocabulary specific to his domain that you didn't understand ("comps," "open houses," "CPC")? Sometimes if you don't understand vocabulary that an expert is using, it can prevent you from understanding the problem. It's good to get in the habit of asking questions because eventually you will get to something you do understand. This involves persistence and is a habit to cultivate.**

Solution

1) Although it's difficult to collect data in an altogether different setting. But the most important concern while collecting data is common to all fields. It is the tracking of customer usage trends and analysis and how to leverage the experience of customers using our product/website. The data which we collect might lie in a different domain altogether. But once we grasp the problem statement, we will be in a position to efficiently demarcate between what's relevant and irrelevant since the customer experience and satisfaction will be same in every domain. But this experience of data collection and analysis in a different domain which seems to be a bit challenging, will eventually help us in collecting data in our own field since this experience will teach us to perform market analysis, customer trend analysis and will make us proficient in identifying the most critical dataset in that domain based on these analyses.

2) As pointed out earlier, this domain specific vocabulary is a bit difficult to grasp initially. For getting a strong hold on such domain specific vocabulary, we can arrange regular meetings with domain experts in addition to studying about how the overall market works, what kind of users it has and what are all the possible terminologies used in this particular type of market. Once we get accustomed to basic vocabulary after trying to converse with domain experts for a few days, we can then put up some basic questions and understand the solutions given by domain experts. But this requires patience and regular interaction with domain experts along with study of the overall market on an individual basis

Problem6

Doug mentioned the company didn't necessarily have a data strategy. There is no industry standard for creating one. As you work through this assignment, think about whether there is a set of best practices you would recommend with respect to developing a data strategy for an online business, or in your own domain.

Solution

There are a number of best practices we can implement in data collection:-

1) Initially we can have a discussion with the domain experts which can help us understand the market. Such regular meetings can help us understand the domain and thereby perform market analysis
2) Also we can collect different possible datasets initially and choose only those which provide the most meaningful results based on data analysis. Again, this can be achieved by regular meetings with Product management and developers and user feedback.
3) With time, we can implement the best practices such as automating the whole workflow of collecting data automatically based on user input in a database or a reporting tool subsequent to which an automated analysis can be performed. This analysis can then be sent to all the concerned members in the team on a daily or biweekly basis.
4) Once we start capturing the best possible datasets, we can feed this data into our product/website which can provide users recommendations or some kind of personalized experience based on their product/website usage trends. This will increase customer satisfaction and trust