Project Report Part3

Presented by:

Anuj Rastogi

anujrast@buffalo.edu

Person# 50134324

Let's begin the report by answering the questions given on page 48 and 49 of the book. The book in these pages talks about a website for buying and selling properties, namely http://www.realdirect.com/.

The success story of http://www.realdirect.com/ is based totally on finding patterns in data regarding how people approach property and how the demand for the same is created. Hence, data mining is the core phenomenon or the platform that has launched http://www.realdirect.com/. It till now forms an integral part of the strategy they deploy in marketing products in their website. Also, it is an integral part of the way they maintain their own service for people.

QUESTION 1:

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:

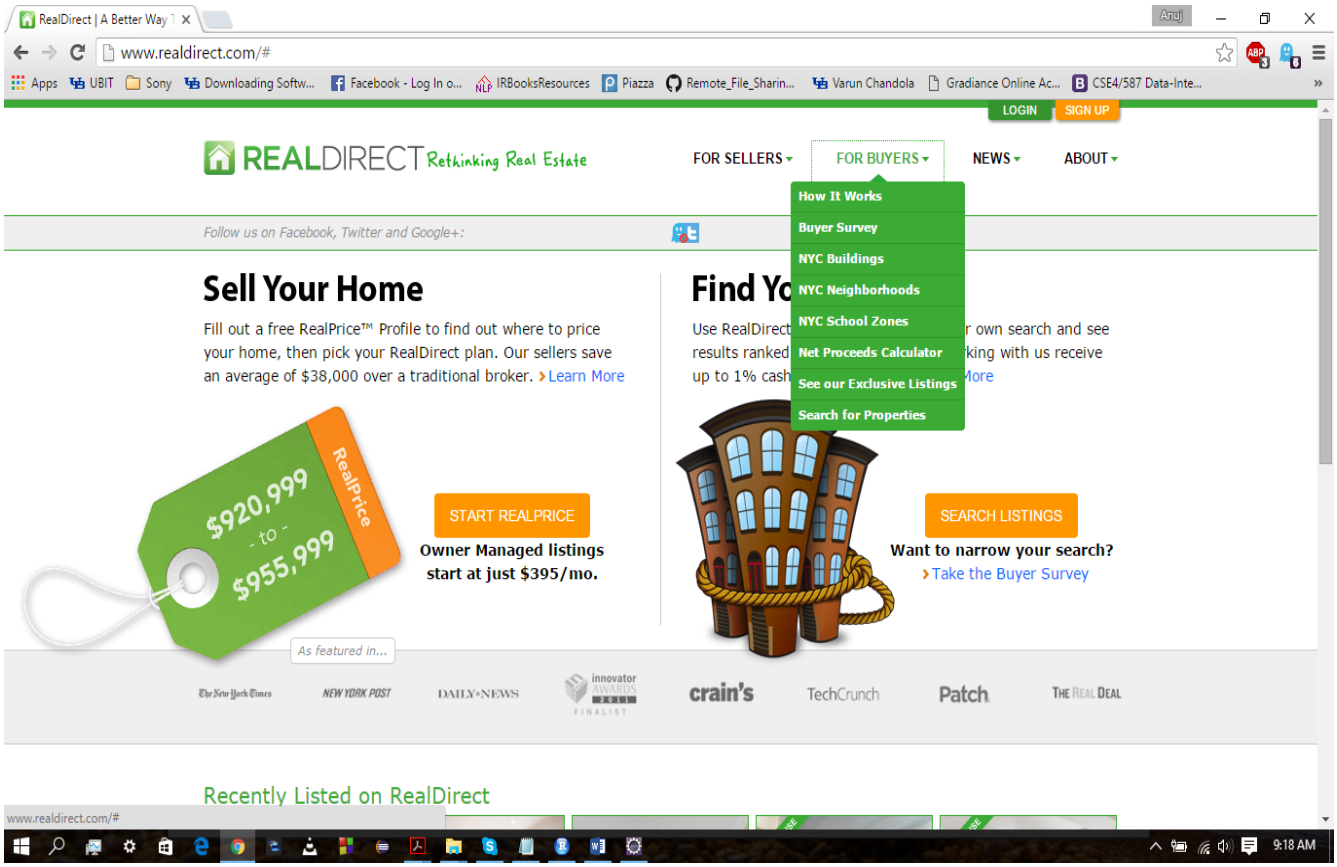
- What data would you advise the engineers log and what would your ideal datasets look like?
- How would data be used for reporting and monitoring product usage?
- How would data be built back into the product/website?

ANSWER 1:

Real Direct has a number of options in their website for both buyers and sellers. The idea of the website itself seemed to have emerged from the people's demand of a forum where in they can trade their property or possession in an effective and friendly way.

It has several options of navigation for both buyers and sellers. They are listed below:

- **Buyers**: The buyers can go to a separate link for buyers having ample number of options for buying property based on user demands. Below is the sample page of the website. The buyers can even buy property based on which areas have good school for their children. The UI is interactive where the buyers can have a first look at the property which may seem impressive to them.



- **Sellers:** Like buyers, sellers also do have multiple options to explore the website. The website provides sellers with the option of estimating the prices for their properties. Also, it allows sellers to display their property photos to effectively display their product.

THE EXISTING BUSINESS MODEL: The existing business model of RealDirect is based on coming up with a website that is adaptable to the changing user demands. To study this demand they keep in touch with the real world data, related to property demand of people. With the help of this data analysis they come to know what more the people demand of or what is the new trend in property and henceforth, they try to align with that changing trend. The demand or the business model is such that it is a service oriented website as they provide a platform where buyers meet sellers.

Hence, user behavior and especially the behavior on how buyers or sellers perceive property is of supreme importance to RealDirect. As it is because of them the website is running, hence, user behavior and changing taste and need towards property is of extreme importance for this company. What people want or what people think about investment in property begets the idea behind the work order of the website.

A Few Research questions that can be answered by data are:

- What is the trend in property demand?
- Do people want property or they want to rent it?
- For whom or which class of people the advertisements should be most intense?
- What should be the price of the product based on user perception?

What data would you advise the engineers log and what would your ideal datasets look like?

The ideal dataset should be such that expresses the sentiment of people towards a particular product. A dataset that should capture the people's opinion and subtle demands. Such a dataset is helpful for the engineers ad can help them come up with something trending and of requirement to people as a whole.

How would data be used for reporting and monitoring product usage?

With the help of data we can capture people's thinking about a product and also what is the trend of usage of such a product. We can do a sentiment analysis describing as to how people feel about a particular product. If they feel good what purpose this data is good for and similarly if bad what was wrong with the product. With all these questions being answered from data, one can report and monitor product usage.

How would data be built back into the product/website?

With the help of data we can understand the current trends of demand and fashion. With such an understanding we can enhance our product.

QUESTION 2:

Because there is no data yet for you to analyze (typical in a startup when it's still building its product), you should get some auxiliary data to help gain intuition about this market. For example, go to *https://github.com/oreillymedia/doing_data_science*. Click on Rolling Sales Update (after the fifth paragraph). You can use any or all of the datasets here—start with Manhattan August, 2012–August 2013.

- First challenge: load in and clean up the data. Next, conduct exploratory data analysis in order to find out where there are outliers or missing values, decide how you will treat them, make sure the dates are formatted correctly, make sure values you think are numerical are being treated as such, etc.
- Once the data is in good shape, conduct exploratory data analysis to visualize and make comparisons (i) across neighborhoods, and (ii) across time. If you have time, start looking for meaningful patterns in this dataset.

ANSWER 2:

BROOKLYN DATA

Let us first read and clean the data and convert factor columns into normal numeric columns.

```
bk <- read.xls(perl = 'C:/Perl64/bin/perl.exe',xls = "rollingsales_brooklyn.xls",pattern="BOROUGH")
head(bk)
summary(bk)

#Converting sales price from factor to numeric
bk$SALE.PRICE.N <- as.numeric(gsub("[^[:digit:]]","",bk$SALE.PRICE))

#Converting the names to lower order
names(bk) <- tolower(names(bk))

bk$gross.sqft <- as.numeric(gsub("[^[:digit:]]","",bk$gross.square.feet))
bk$land.sqft <- as.numeric(gsub("[^[:digit:]]","",bk$land.square.feet))

bk$sale.date <- as.Date(bk$sale.date)
bk$year.built <- as.numeric(as.character(bk$year.built))

summary(bk$sale.price)
## do a bit of exploration to make sure there's not anything
```

The above script cleans the data and produces a summary of the data. The **as.numeric()** function converts the data into numeric form from original factor form. Also, it specifies a pattern inside **gsub() ,** which is used to withdraw a subset from that data based on the pattern provided in **gsub().**

The below command plots three histograms to analyze the sales price and the square feet area of the property which has 0 sales price.

```
hist(bk$sale.price.n, main = 'hist(bk$sale.price.n)')
hist(bk$sale.price.n[bk$sale.price.n>0], main = 'hist(bk$sale.price.n[bk$sale.price.n>0])')
hist(bk$gross.sqft[bk$sale.price.n==0], main = 'hist(bk$gross.sqft[bk$sale.price.n==0])')
```

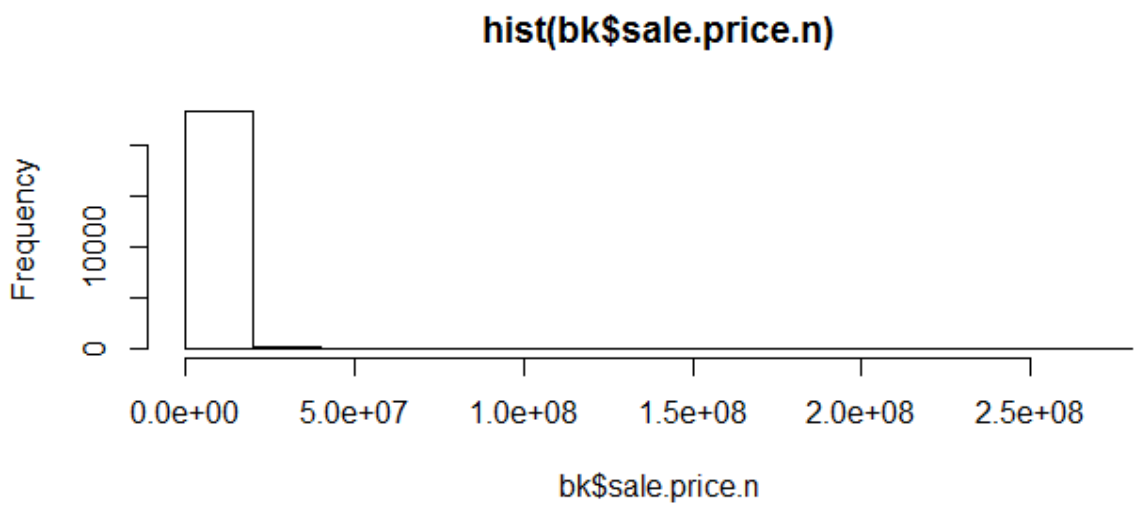


Figure 1

The above histogram shows the total number of apartments with all the price included in x-axis.

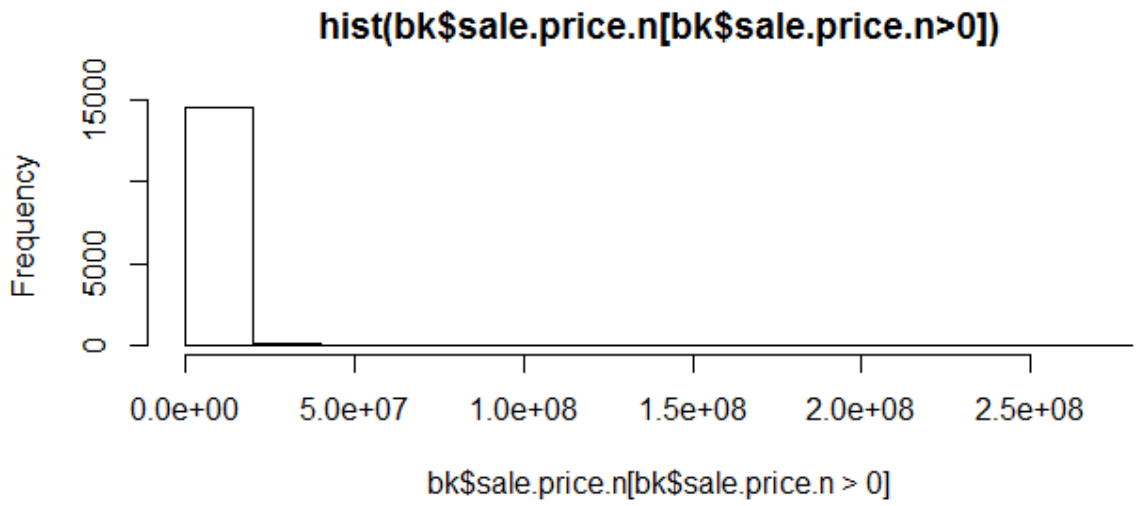


Figure 2

The above histogram represents the number of houses or properties for which the selling price is not 0. Hence, there are around 15000 of such properties in Brooklyn.

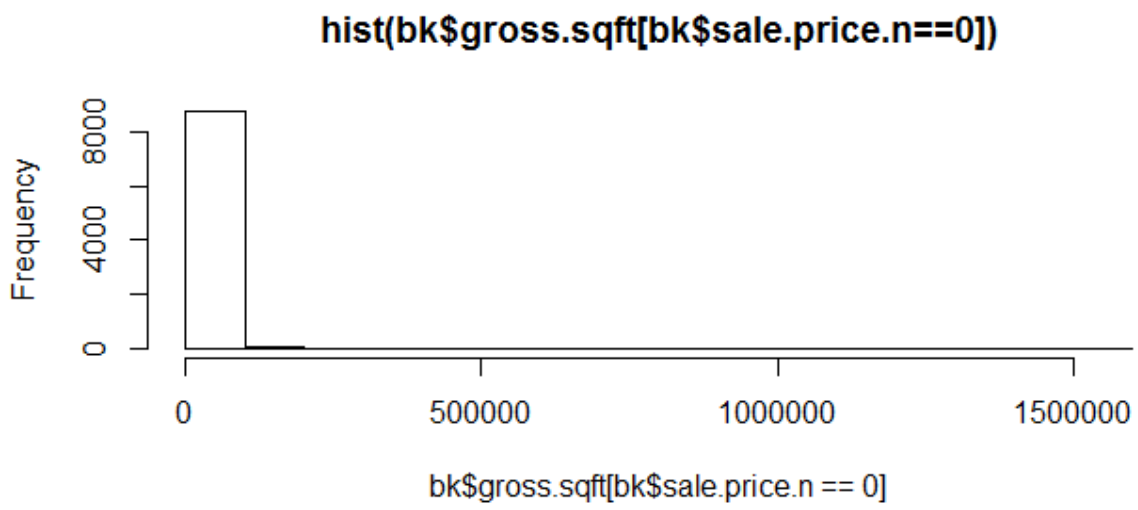


Figure 3

The above histogram represents the total number of properties in Brooklyn for which the total selling price is 0.

Moving further, let's plot some more plots for analysis. The below script cleans the data and stores all those properties for which the sale prices are not 0 into one table named **bk.sale**.

```
## keep only the actual sales
bk.sale <- bk[bk$sale.price.n!=0,]

plot(bk.sale$gross.sqft,bk.sale$sale.price.n, main = 'plot(bk.sale$gross.sqft,bk.sale$sale.price.n)')
plot(log(bk.sale$gross.sqft),log(bk.sale$sale.price.n), main = 'plot(log(bk.sale$gross.sqft),log(bk.sale$sale.price.n))')
```

It also plots two graphs. One between property prices and square feet and the other between gross area and square feet.

The below graph represents gross square feet area and sale prices.

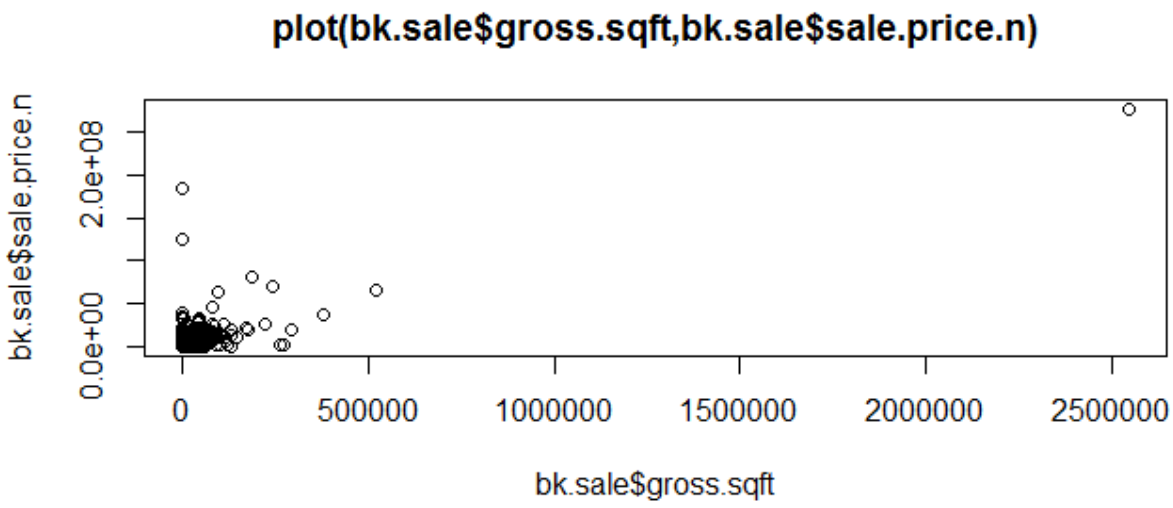


Figure 4

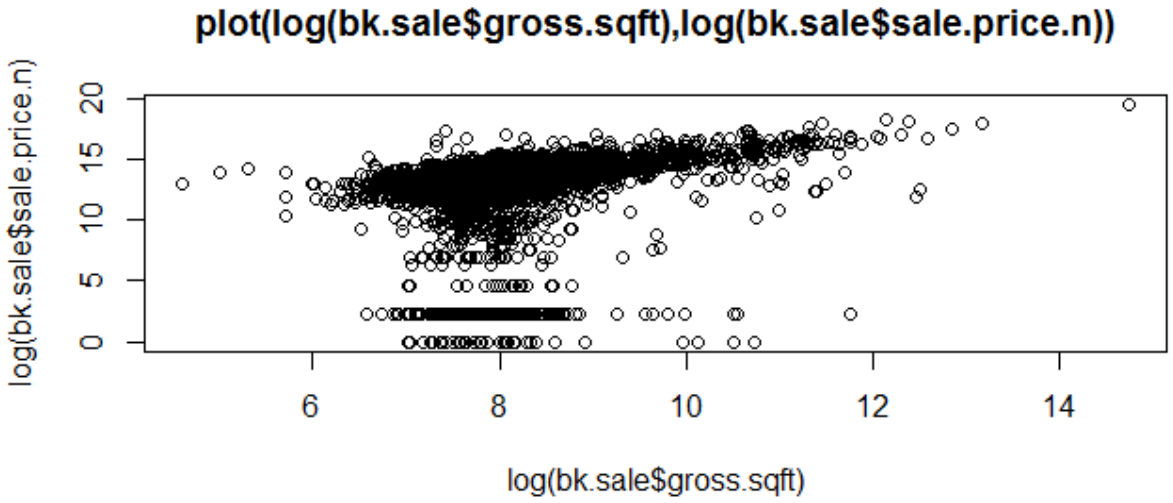


Figure 5

The above figure represents the log of the gross square feet area and the property sale prices. This graph can help a company decide on its prices in Brooklyn, by analyzing and fixing a median price for its property, based on square feet.

Since, the above graph is for all the properties and the variation in their rates based on the area, let's now consider only the family homes and draw a plot that represents the change in the prices of the family house based on area. The below script does this.

```
## for now, let's look at 1-, 2-, and 3-family homes
bk.homes <- bk.sale[which(grepl("FAMILY", bk.sale$building.class.category)),]

plot(log(bk.homes$gross.sqft),log(bk.homes$sale.price.n) ,main = 'plot(log(bk.homes$gross.sqft),log(bk.homes$sale.price.n))')
```

The below plot reflects the change in the prices of the family houses based upon their area.

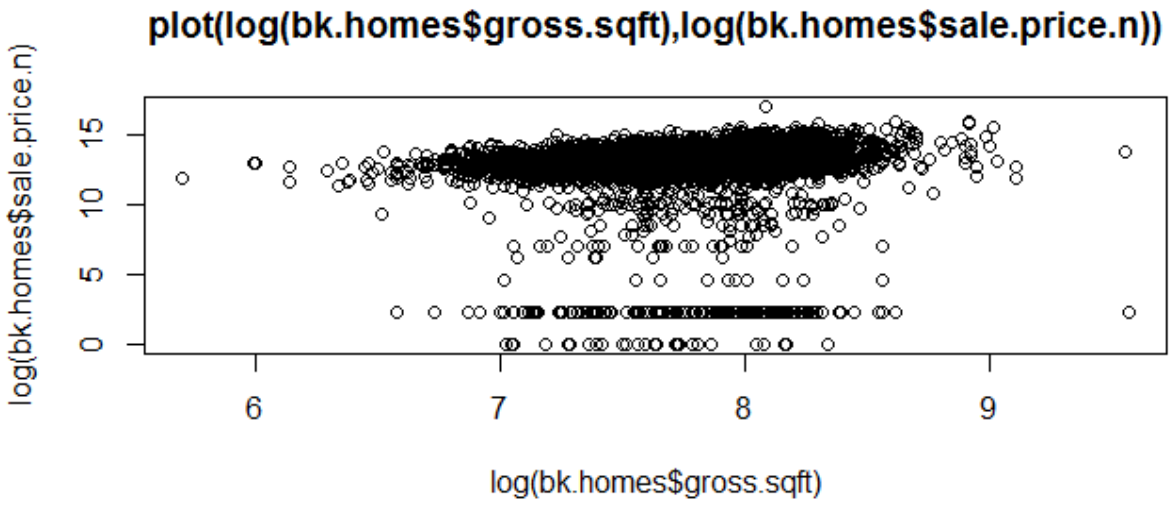


Figure 6

Going ahead we can now start understanding the trends of population distribution and the fluctuation in prices, if any, over a period of time, in Brooklyn. The below script identifies the change in prices in Brooklyn as comprehended by the given data.

```
#Plotting the graph to view the change in property sale prices
plot(bk.sale$sale.date, bk.sale$sale.price.n ,  main = 'Variation in Property Prices', xlab = 'TimeSpan', ylab = 'sale prices', col = 'blue', pch = 15)
abline(lm(bk.sale$sale.date~bk.sale$sale.price.n), col = 'red')
```

The below plot shows the variation in property prices over a timespan for which the data was collected. Also, it draws a correlation Line with red color.

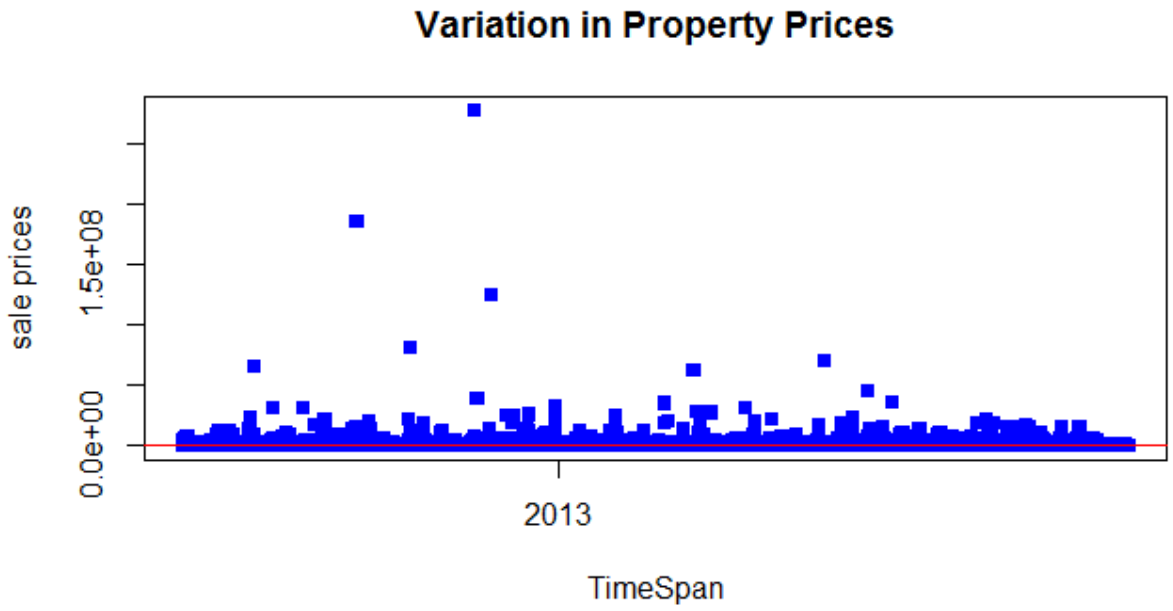


Figure 7

Let us now see the distribution of people based on their Tax class category and see which Tax category has the maximum population in Brooklyn. The below script and plot helps us understand that.

```
#Determining the distribution of tax class category in Buildings
TaxClassvsBuildings <- table(bk.sale$building.class.category, bk.sale$tax.class.at.present)
TaxPayerDistribution <- table(bk.sale$tax.class.at.present)
barplot(TaxPayerDistribution, las = 1, main = 'Tax Payer Category Distribution')
box()
```

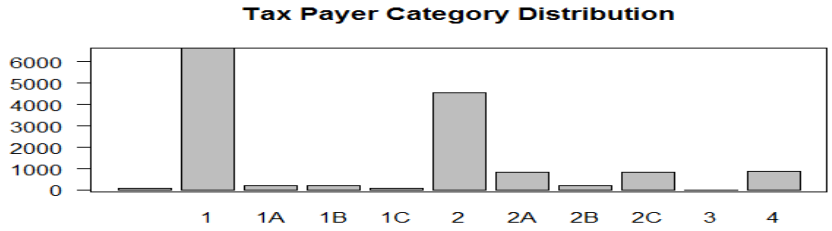


Figure 8

Since now we know that the maximum people who reside in Brooklyn belong to Tax Class Category 1 therefore lets now try to find some facts about these people. The below script tells us the neighborhoods where in the Tax payer category 1 population resides and their residential population is more than 150.

```
#Tax Class Category 1 Distribution in NeighbourHoods
TaxClassCategory1 <- table(bk.sale$neighborhood[bk.sale$tax.class.at.present == 1])
TaxCategoryDistribution <- subset(TaxClassCategory1, TaxClassCategory1>150)
barplot(TaxCategoryDistribution, las = 2, cex.names = 0.5, main = 'Areas With maximum Population of Tax Class Category1')
box()
```

Also, the plot below shows the residential areas or the properties where in the population of tax class category 1 is maximum, within Brooklyn.

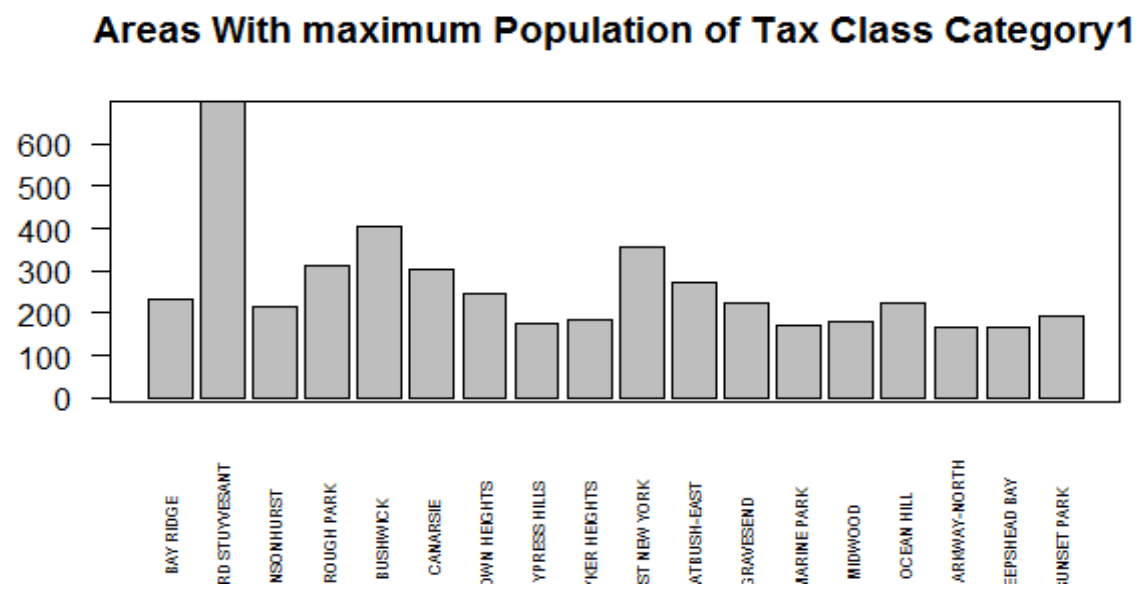


Figure 9

With this let us now find the areas where the population of Tax Class category 1 stays and pays more than their average selling price. Also the script is attached to find the mean selling price for Tax Class Category 1.

```
#Mean Salary For Tax Class Category 1
taxCat1 = subset(bk.sale, bk.sale$tax.class.at.present == 1)
mean(taxCat1$sale.price.n)
```

```
> mean(taxCat1$sale.price.n)
[1] 577176
>
```

The above script is for finding the mean selling price for tax Class category 1. The below script determines the areas where the population of Tax Class category 1 stays and pays more than their average selling price.

```
#Analysing the sales prices for Tax Class Category1 where sales price are more than average
AreasWithHighPrice <- subset(bk.sale, bk.sale$sale.price.n> mean(bk.sale$sale.price.n) & bk.sale$tax.class.at.present == 1)
MoreThanAvgNeighbour <- table(AreasWithHighPrice$neighborhood)
Result1 <- subset(MoreThanAvgNeighbour, MoreThanAvgNeighbour > 50)
barplot(Result1, las = 2, cex.names = 0.5, main = 'Areas With above average prices occupied by Tax Category 1')
```

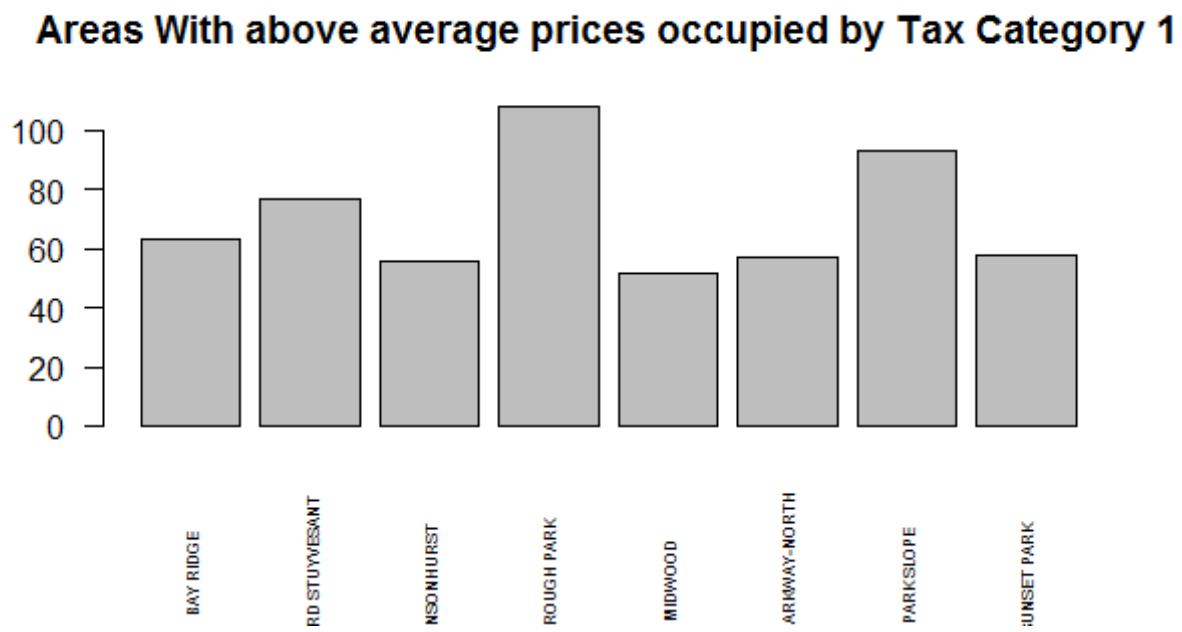


Figure 10

Also, let us now find the areas where the Tax Class category 1 lives and have selling price lesser than average price for Tax Class Category1. The below script helps plot a graph and from there we can see this conclusion.

```
#Analysing the sales prices for Tax Class Category1 where sales prices are least

AreasWithLeastPrice <- subset(bk.sale, bk.sale$sale.price.n < mean(bk.sale$sale.price.n) & bk.sale$tax.class.at.present == 1)
LessThanAvgNeighbour <- table(AreasWithLeastPrice$neighborhood)
Result2 <- subset(LessThanAvgNeighbour, LessThanAvgNeighbour > 50)
barplot(Result1, las = 2, cex.names = 0.5, main = 'Areas With less prices occupied by Tax Category 1')
```

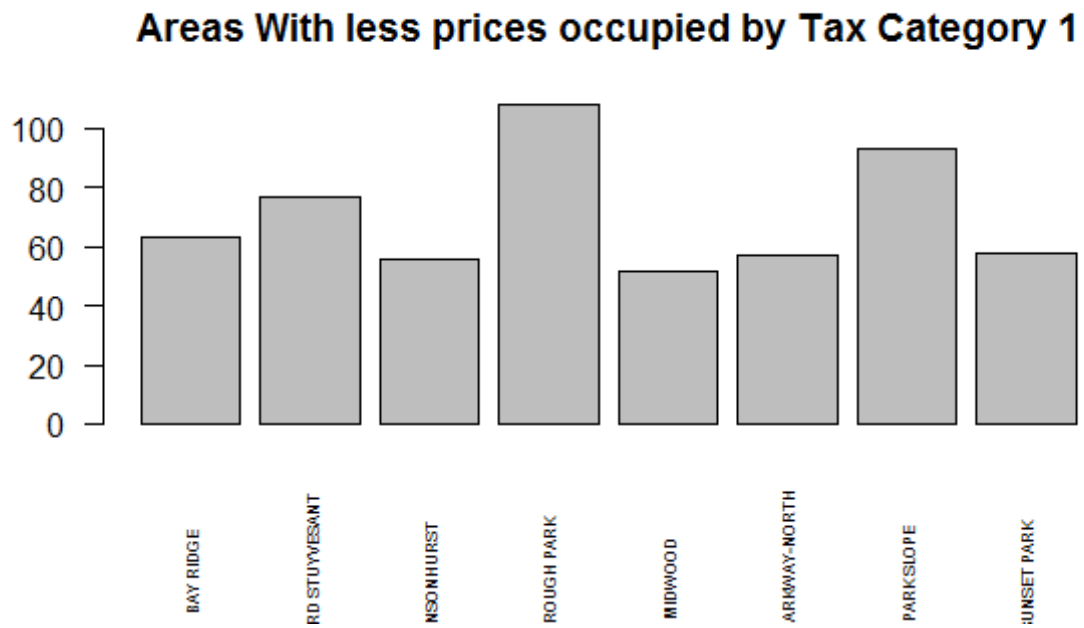
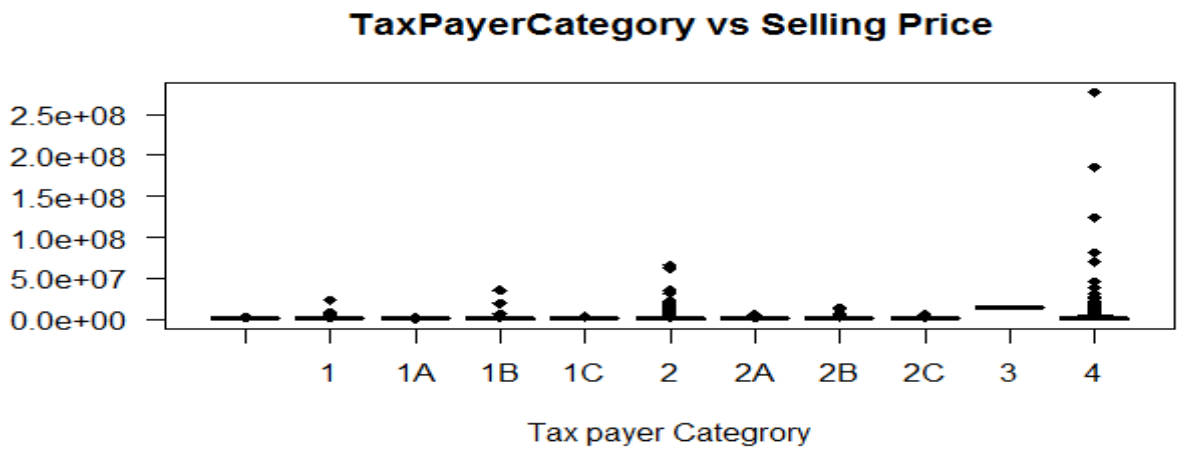


Figure 11

Finally let us find out which Tax category pays the maximum selling price for property in Brooklyn. The below script and graph help us to depict this.

```
#Maximum Selling price Correspondng to TaxPayer
plot(bk.sale$tax.class.at.present, bk.sale$sale.price.n, las = 1, pch = 18, xlab = 'Tax payer Categrory', main = 'TaxPayerCategory vs Selling Price' )
```



ANALYSIS OF THE BROOKLYN DATA:

From the Brooklyn data Figure 8 it is evident that the maximum number of people residing in Brooklyn belong to Tax Class Category 1. Therefore, if we do economics of scale and focus on these payers there is a potential of finding more customers compared to all the others. Also, from the analysis of figure 9 we can deduce areas where in these people are concentrated. Therefore, from Brooklyn if we focus on these people and the areas showcased in Figure 9, we can get a good number of customers.

To decide on the prices we first find the mean selling price of Tax Class category 1. That is something below.

```
> mean(taxCat1$sale.price.n)
[1] 577176
>
```

Therefore we can now decide the pricing in different areas based on this mean price. In the areas where people are living and are paying selling price more than average means those people are financially pretty well off. Therefore property prices in those areas can be fixed above the mean price. This is shown in Figure 10. Also, for people living in areas where they are paying lesser than the mean price we can fix the price in those areas to be lower than average as in Figure 11. If there areas where there is a mixed population, there in keeping the price to mean will suffice for at least 50% of population.

EXTENDING THE ANALYSIS TO ALL THE BOROUHGS

Above we have done the analysis just for one Brooklyn Borough. Now let us combine all the data for all the boroughs and then come to some conclusion. Let us first see which Tax Class category is in majority in the city. This can be found out by the script below and can be deciphered from the graph below.

```
#Determining the distribution of tax class category in Buildings
TaxClassvsBuildings <- table(bk.sale$building.class.category, bk.sale$tax.class.at.present)
TaxPayerDistribution <- table(bk.sale$tax.class.at.present)
barplot(TaxPayerDistribution, las = 1, main = 'Tax Payer Category Distribution')
box()
```

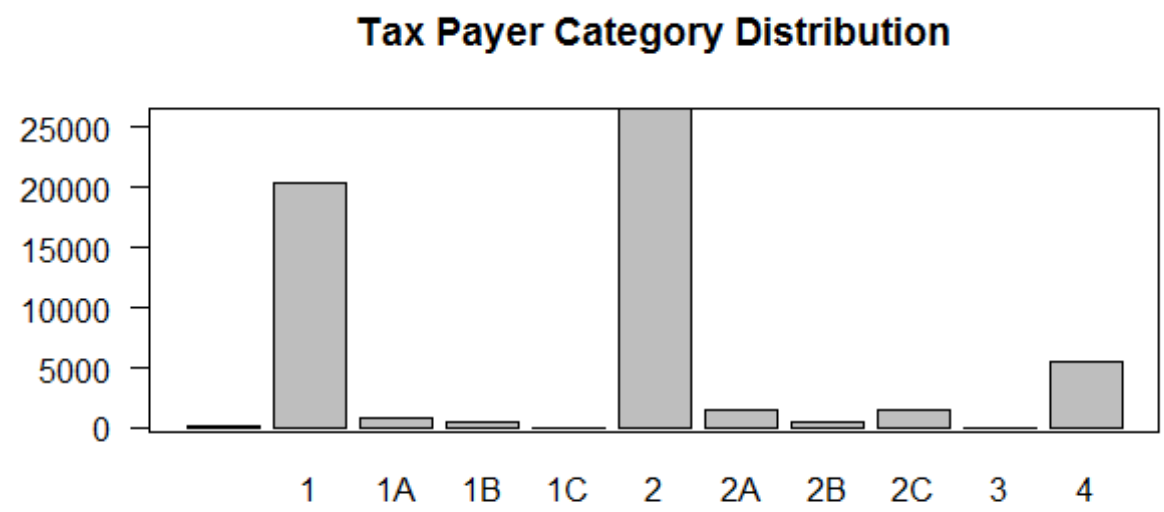


Figure 12

Also, let us see the variation in the property prices for the period in which data was collected. The below script suffices that and then we can see this in the plot below.

```
#Plotting the graph to view the change in property sale prices
plot(bk.sale$sale.date, bk.sale$sale.price.n ,  main = 'Variation in Property Prices', xlab = 'TimeSpan', ylab = 'sale prices', col = 'blue', pch = 15)
abline(lm(bk.sale$sale.date~bk.sale$sale.price.n), col = 'red')
```

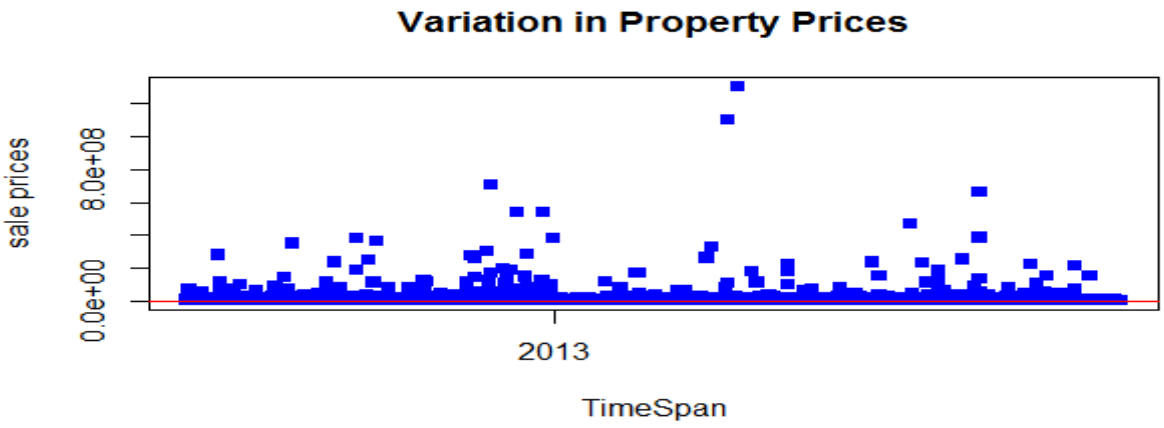


Figure 13

Since the maximum people belong to the Tax Class category 2. So let's now focus on their concentration areas and the process in different areas in which they reside. The below plot and script helps us understand the Tax Class 2 category people in neighborhoods.

```
#Tax Class Category 2 Distribution in NeighbourHoods
TaxClassCategory1 <- table(bk.sale$neighborhood[bk.sale$tax.class.at.present == 2])
TaxCategoryDistribution <- subset(TaxClassCategory1, TaxClassCategory1>150)
barplot(TaxCategoryDistribution, las = 2, cex.names = 0.5, main = 'Areas With maximum Population of Tax Class Category2')
box()
```

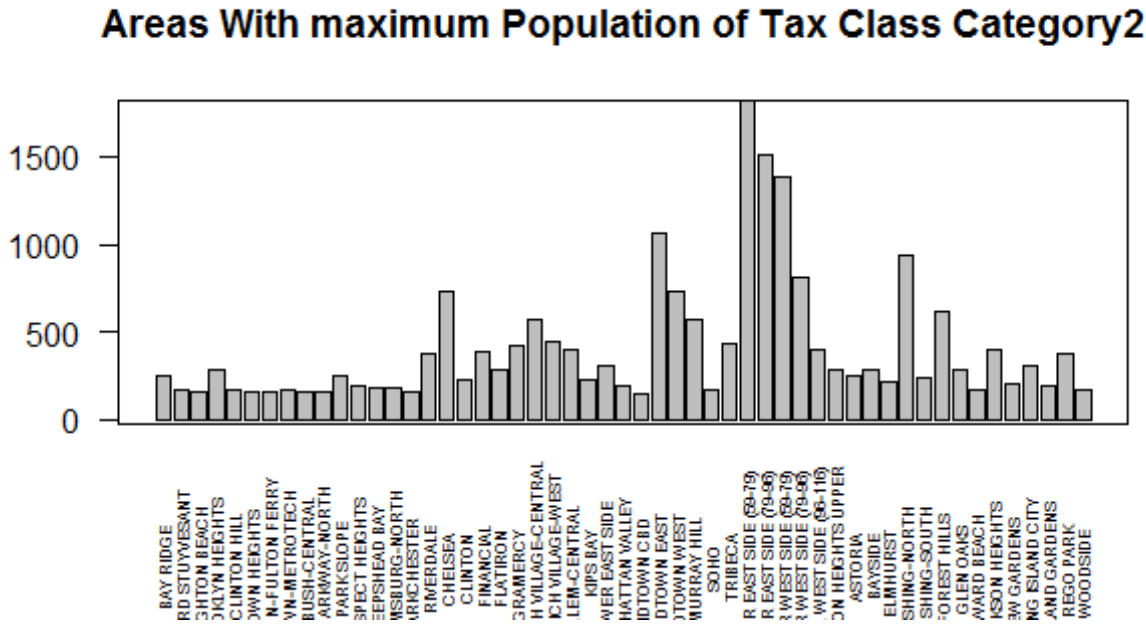


Figure 14

Let's us now focus on where the price of property for Tax Class Category 2 is more than average for Tax Class Category 2. The below script helps to find out those areas where Tax Class category 2 people live and have selling price more than average. Also the plot help to identify those areas.

```
#Analysing the sales prices for Tax Class Category2 where sales price are more than average
AreasWithHighPrice <- subset(bk.sale, bk.sale$sale.price.n> mean(bk.sale$sale.price.n) & bk.sale$tax.class.at.present == 2)
MoreThanAvgNeighbour <- table(AreasWithHighPrice$neighborhood)
Result1 <- subset(MoreThanAvgNeighbour, MoreThanAvgNeighbour > 50)
barplot(Result1, las = 2, cex.names = 0.5, main = 'Areas above average prices occupied by Tax Category 2>50')
```

The below is the plot from the above script.

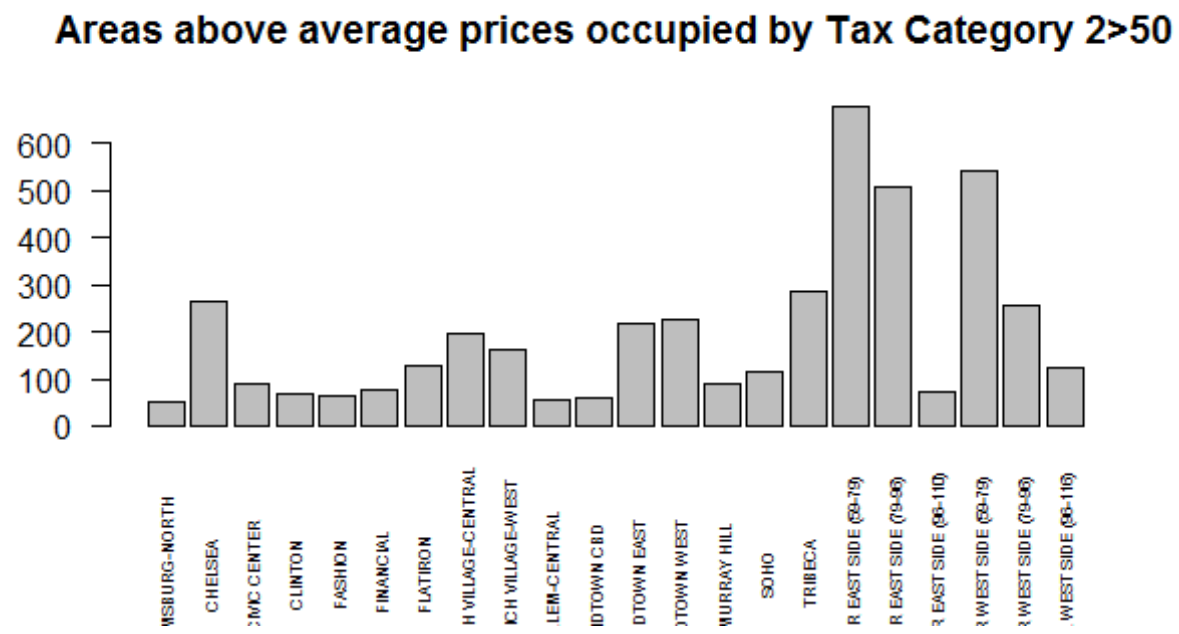


Figure 15

Let's us now focus on where the price of property for Tax Class Category 2 is less than average for Tax Class Category 2. The below script helps to find out those areas where Tax Class category 2 people live and have selling price less than average. Also the plot help to identify those areas.

```
#Analysing the sales prices for Tax Class Category2 where sales prices are least

AreasWithLeastPrice <- subset(bk.sale, bk.sale$sale.price.n < mean(bk.sale$sale.price.n) & bk.sale$tax.class.at.present == 2)
LessThanAvgNeighbour <- table(AreasWithLeastPrice$neighborhood)
Result2 <- subset(LessThanAvgNeighbour, LessThanAvgNeighbour > 50)
barplot(Result1, las = 2, cex.names = 0.5, main = '>50 population less prices Areas by Tax Category 2')
```

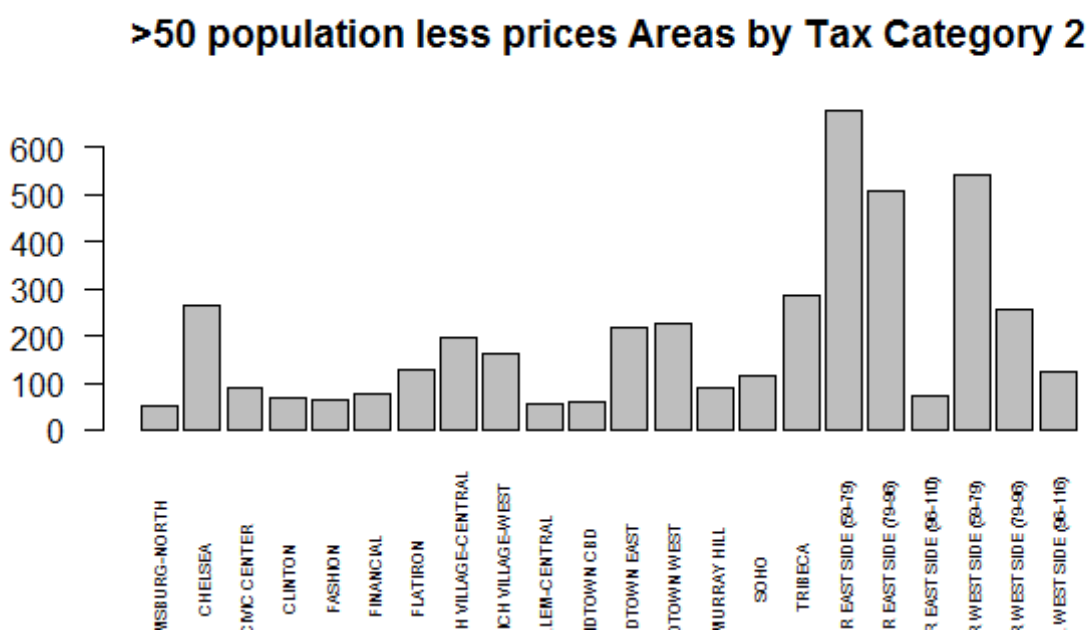


Figure 16

The mean prices for Tax Class Category 2 can be found out using the below script.

```
#Determining The Mean
TaxCat2 = subset(bk.sale, bk.sale$tax.class.at.present == 2)
mean(TaxCat2$sale.price.n)
```

```
> mean(TaxCat2$sale.price.n)
[1] 1240691
```

Let's also analyze the maximum selling price that has been recorded. Also lets try to visualize that this has been recorded corresponding to which category.

```
#Maximum Selling price Corresponding to TaxPayer
plot(bk.sale$tax.class.at.present, bk.sale$sale.price.n, las = 1, pch = 18, xlab = 'Tax payer Categrory', main = 'TaxPayerCategory vs Selling Price' )
```

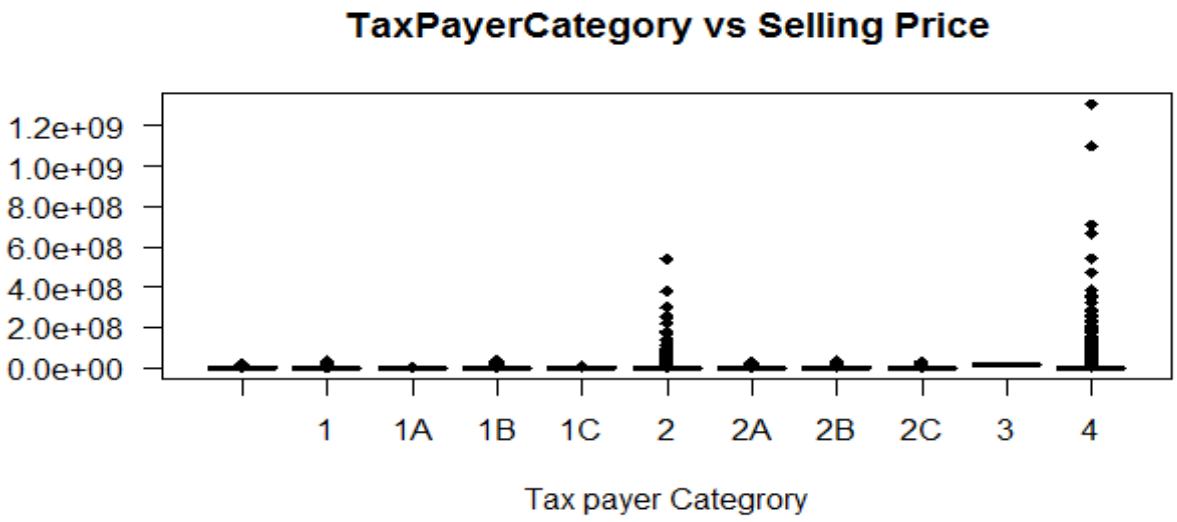


Figure 17

ANALYSIS OF THE MONTHLY DATA:

From the analysis of the monthly data we see that the maximum number of people correspond to the Tax Class category 2. Thus, focusing on these people will give more probability of success. Also, in Figure 14 we can see the concentration of these people and can focus on these areas for advertisements. Also, from Figure 15 and Figure 16 we can have an idea about pricing for Tax class category 2. From figure 15 we can identify areas where to keep the prices of property above average for Tax Class Category 2. The reverse can be identified from Figure 16.

From Figure 17 we, get an idea about which Tax Category to focus for luxury apartments because this Figure tells us about maximum selling price corresponding to the Tax Class Category.

COMPARISION OF BROOKLYN AND MONTHLY DATA

ACROSS NEIGHBORHOODS

From the Brooklyn data we deciphered that the maximum people residing in Brooklyn belong to Tax Class Category 1. Hence, in this area we can focus on Tax Class Category 1 but from the monthly data we have figured out that, in the city the maximum population corresponds to the Tax Class Category 2 and not 1. So for the city we will have to focus on Tax Class Category 2 if we want to do the economics of scale. But from the above 1 more thing is clear that the data in individual areas can fluctuate significantly when compared with the whole city. Figure 8 and 12 give us an idea on the above argument.

Also, from the city analysis we get to know a lot of different areas where people from Tax Class category 2 reside and hence, focusing on these areas is going to bring more customers. This is shown in Figure 14. Also, the monthly analysis done in Figure 15 and Figure 16 will help us to identify prices in those areas where Tax Class Category 2 reside and have prices more than average or lesser than average.

Also, the Brooklyn data produces a graph of maximum selling price against the tax category 4, which is the same for the monthly analysis, as well, from Figure 17. Therefore, this statistics is concurrent. Hence, for luxury housing in all areas we can focus on category 4.

ACROSS TIME:

Both the data from Brooklyn and monthly analysis project less change in prices of property. Although the graph obtained from monthly analysis is much more scattered showing that a lot of properties throughout the city had an increase in selling price. Figure 7 and 13 helps us visualize this.

QUESTION 3:

Summarize your findings in a brief report aimed at the CEO.

ANSWER:

The above analysis can be submitted to the CEO for the purpose of review and business profit.

QUESTION 4:

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?

ANSWER:

Along with talking to people involved in buying and purchasing properties, we can also talk to potential customers and common people. This is to get an idea on how their virtual demand will be once they think about buying and selling properties.
But one thing is certain that, we should be absolutely sure about the information we are seeking and this will help up frame questions for people in such a way that will extract some meaningful information from them.

QUESTION 5:

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.

ANSWER:

- Once stepping out of your domain you are not much familiar with the parlance of that field. So searching about data with those exact words become a bit difficult. But having said that both these situations still require clear understanding of the demand you want to collect the data for.
- It was initially bit difficult but later on with the passage of time and with a motive of understanding at least the technical terms, and reading about them on google, provided an insight about those jargons. Of course understanding plays a significant role in comprehending the correlation between data attributes.

QUESTION 6:

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.

ANSWER:

As already talked about, the best practices in data strategy include forming a consolidated plan as to which field the data should be related to. If there is confusion regarding field of data retrieval, then it becomes hard to focus on data acquisition resources. Also, before data collection the potential customer for the product should be known and maximum effort should be put to acquire information from those potential customers since they are the most promising source of business return.