# REPORT PROBLEM 4

Prepared By:

Anuj Rastogi

anujrast@buffalo.edu

PersonNumber# 50134324

*INTRODUCTION*

New York has a huge population and therefore there is a good chance of a real estate business in the city. But whether to open a business that helps people to find apartment rentals or help people to buy apartments is a question. This question needs understanding of user demands and priorities. Although, one can open a real state agency to do both but which will yield more profit is the question over here.

Facts have reviled over time that apartment rentals are in more demand in New York because of the high buying rates. But this needs to be proved through valid datasets representing people's opinion. Hence, this part is an attempt to bolster the fact that apartment rentals are in more demand in New York compared with buying a house.

*UNDERSTANDING THE WORK DONE IN SCRIPT*

Let's first start by understanding what work has been done and then we will make an attempt to understand the possible outcomes and recommendations. Initially data from twitter in JSON format was collected using twitter REST API. Post this, the data was subjected to R environment for analysis.

First step is to import the data into R. This can be done by the below code.

```
#Reading the
fileRead = readLines('Total.json', warn = FALSE)
fileRead_df = jsonlite::fromJSON(fileRead)
```

```
> sum(unlist(rentCases))
[1] 599
```

This JSON file contains weekly data for property analysis done for a week related to New York City.

Post this let us extract only the text field of this data and see some statistics regarding buying or renting of apartments. The below code takes only the text field of these tweets and then performs some analysis.

```
#Extracting only text
onlyText <- fileRead_df$text


#Counting the number of tweets with keyWord Rent or Buy
rentCases <- lapply(onlyText,function(x) { grep(as.character(x), pattern = "\\<rent")} )
sum(unlist(rentCases))

buyCases <- lapply(onlyText,function(x) { grep(as.character(x), pattern = "\\<buy")} )
sum(unlist(buyCases))
```

From the above code we try to find out the number of tweets which have either rent or buy key word. This will give an idea on which is a more popular topic on Twitter. The below are the results of the above query.

```
> sum(unlist(rentCases))
[1] 599
```

Figure 1

The above figure shows that in 599 tweets in a week rent Keyword occurred. Also, the below figure shows the number of tweets in which the buy keyword occurred.

```
> sum(unlist(buyCases))
[1] 163
```

Figure 2

The above figure shows that in 163 tweets, out of the total number of tweets, buy keyword occurs. Let us now do some more interesting analysis to corroborate the claim that renting in New York is more common. For proceeding further, let us first convert our text vector into a corpus of words and then clean it. The below code is able to do these tasks.

```
#Installing NLP Library Function
install.packages('tm')
library(tm)

#CLeaning the data
myCorpus = Corpus(VectorSource(onlyText))
myCorpus2 = tm_map(myCorpus, content_transformer(tolower))
myCorpus3 = tm_map(myCorpus2, removeWords, stopwords('English'))
myCorpus4 = tm_map(myCorpus3, removePunctuation)
```

Here we have converted all the text to lower case and then we have removed the Stop words and Punctuations in the text. The '**tm_map'** is the function of NLP library which comes with the name **'tm'.** Post this let us now convert our data into the form of term document matrix and start doing some detailed and conclusive analysis. The below code converts the data into the Term Document Matrix.

```
#Converting The corpus in the form of TermDocumentMatrix
tdm = TermDocumentMatrix(myCorpus4)
```

Post this lets analyze the maximum frequency terms in our corpus. This will give us a feel of what is the common topic when it comes to real estate. Also, this will develop our understanding of comprehending user demands. The following Script helps us to find the most frequent words in the tweets.

```
#finding the frequent Terms
maxFreqTerms = findFreqTerms(tdm, lowfreq = 150, highfreq = Inf)
maxFreqTerms
```

The below figure is the result obtained when the value of low frequency is more than 200

```
> maxFreqTerms = findFreqTerms(tdm, lowfreq = 200, highfreq = Inf)
> maxFreqTerms
[1] "apartment" "buy"       "city"      "house"     "manhattan" "new"
[7] "nyc"       "rent"      "york"
```

Figure 3

The above analysis show that out of all the tweets a lot of tweets have both buy and rent, pertaining to New York city. Hence, people are interested in real estate and they are willing to buy properties like apartments.

Let us now find the closely related words to both buy and rent to know what is that people want to buy and rent. This will give an idea and differentiation of demands emerging from both buy and rent. The below script used finds the correlated terms to both buy and rent.

```
#Finding the closely related words to rent and buy
findAssocs(tdm, 'rent' , 0.2)
findAssocs(tdm, 'buy', 0.2)
```

The below is the derived statistics from the above scripts.

```
> findAssocs(tdm, 'rent' , 0.2)
$rent
          3400        average     onebedroom neverknownfacts      apartment
          0.33           0.33           0.33           0.31           0.23
           nyc      searching        bedroom
          0.21           0.21           0.20
```

Figure 4

The above analysis show that people have been searching for rented apartments which are one bed room. Also it looks like a lot of people have found and quoted a figure 3400 which seem to be the average rent in New York.

```
> findAssocs(tdm, 'buy', 0.2)
$buy
                  500k edu00a0u00bdedu00b2u0080                  even
                  0.44                  0.44                  0.43
                  half                   lmao           mikaarianna
                  0.43                  0.43                  0.43
                   lot                 think                 house
                  0.42                  0.42                  0.40
                   'll                739610             cocktail
                  0.35                  0.35                  0.35
                craft                scrape               dollars
                  0.35                  0.35                  0.32
              thirteen              children                home…"
                  0.30                  0.29                  0.29
            incredibly               married                second
                  0.29                  0.29                  0.29
                   dog       advicetowriters                  htt…
                  0.28                  0.27                  0.27
                  will              shedding              together
                  0.26                  0.25                  0.24
                  move                 upper
                  0.23                  0.20
```

Figure 5

The above analysis shows the closely related words to buy. Some interesting facts emerge showing Children, married, 500k, house as related terms apart from the others. Hence, these suggests the nature and the type of people who are in general looking for apartments to buy or houses to buy.

Let us now plot the graph for the most frequent terms and we will then see how the word count is present in our corpus of words. The below script does this for us.

```
#plotting the ggPlot for the most frequent words
library(ggplot2)
termFrequency <- rowSums(as.matrix(tdm))
termFrequency <- subset(termFrequency, termFrequency >= 100)
df <- data.frame(term = names(termFrequency), freq = termFrequency)
ggplot(df, aes(x = term, y = freq)) + geom_bar(stat = "identity")+xlab("Terms") + ylab("Frequency") + coord_flip()
```

The graph from the above script is shown below. It represents the number of words appearing and also, the number of documents they are contained in. We find the graph for words having term frequency>100.

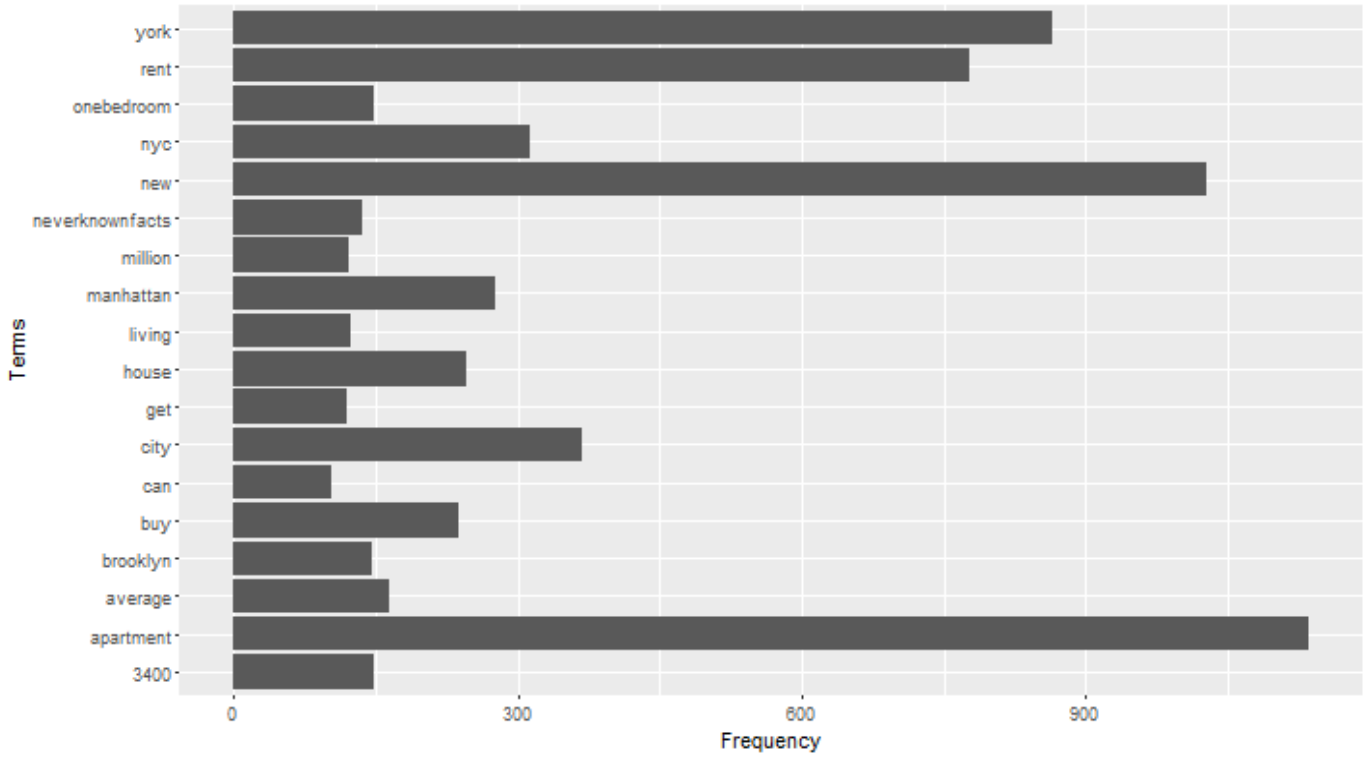


Figure 6

The above graph shows that how the word frequency is present in different documents. It shows that a lot of people have tweeted for apartments rent in New York. Lesser number of tweets been found for buy. This is ow becoming clearer that the inclination of people is more towards rent and not buy.

Let us try to do the clustering of words. This clustering will help us understand the co-relation between the words in the tweets in a better way and will thus, help us reach to some conclusion about the search behavior of people.

The below script helps us to plot a dendro-gram which is useful in situations where relational analysis between the words is expected.

```
#Clustering the terms and drawing the co-relation
tdm2 <- removeSparseTerms(tdm, sparse = 0.95)
m2 <- as.matrix(tdm2)
distMatrix <- dist(scale(m2))
fit <- hclust(distMatrix, method = "ward.D2")
plot(fit)
```

The below is the figure for the dendro-gram. This shows clearly that what are the closely related words to rent and gives an idea about people's demand towards rent and also, buy.

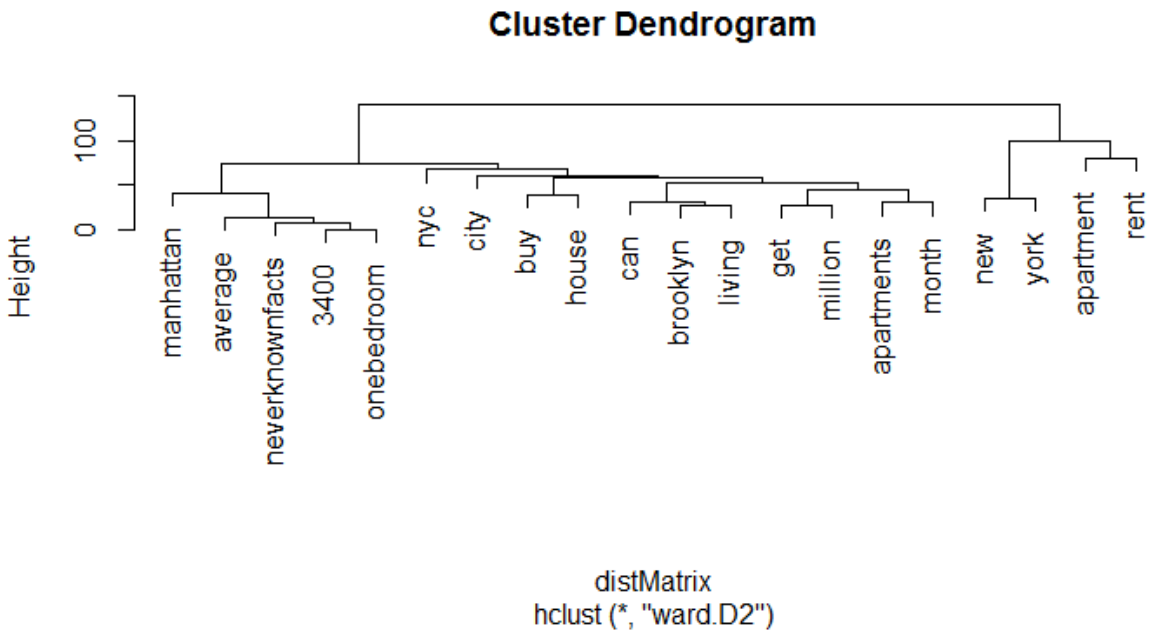


Figure 7

## *ANALYSIS OF THE ABOVE GRAPHS:*

Let us delve into the analysis of the above graphs. From the above data plots it is clear that majority of the people through the week have focused more on rents than buying. Hence, a business model aimed at rent is likely to produce more customers when compared to buy.

Even in producing a website that focuses or advertises on rented apartment we will have to focus on the following things. And similarly for buying as well we will have to focus on the following parameters:

**Rent:**

From our tweet analysis through the week it is pretty much clear that a lot of people in New York look for apartments. This is evident from Figure 1, 3 and 6. But which areas to look for and which parameters to focus on while building a rent website is a question. From the analysis of Figure 3 and Figure 4, we see that along with rent, Manhattan, one bedroom and apartment also come with high frequency. Figure 4 shows their correlation with rent. Thus, we can say with some verity that people who are looking for rented apartments look for in in mainly Manhattan. Also, for rent people prefer apartment over houses. Thus, if we focus on advertising rented apartment in New York, that is going to bring more customers. Along with this in Figure 7 we see that rent is also related to words like average and Brooklyn. This shows us that people are willing to know the average prices for rent in New York in locations like Manhattan and Brooklyn. This analysis will help us later in the report to predict the pricing model for our website.

**Buy:**

From the analysis of Buy keyword we see in Figure 2, 3 and 6 that people are also interested in buying the property although not so much as they are interested in renting. When buying a house people see various parameters and this is conveyed in Tweets. We see in Figure 5 that buy has a correlation with the words such as marriage, children and house. Also, it is related to words such as 500k. Which may be the average value of a house in New York. Also, we see from Figure 7 that house falls closely with Buying. This projects that people interested in buying mainly focus on buying a house than apartment. Hence, for buying segment it would be better to display more advertisements related to buying a house. Along with this a figure occurrence of the sort 500K will help us later in the analysis when we decide on pricing model.

From the above analysis it can be said that, mainly people who have a family prefer to buy houses. Along with this, people look for facilities related to their children. Hence, they actually indicate that they want to buy a house in an area with amenities related to their family, like schooling shops etc. Therefore, business model on buying should focus on areas with good amenities. Advertisements for houses in such areas will have a higher probability of getting a customer and hence, will increase the quality of our website or property search engine.

## *RECOMMENDATIONS TO THE CEO OF THE COMPANY*

As we know that, property business like any other business is the business of demand and supply. Hence, continuous customer engagement is necessary. Looking and analyzing the above figures tells us that majority of the interest of the people is towards rent. This is evident from Figure 1, 3 and 6. Hence, this source will provide a better customer bank. **Therefore, the company should focus more on rent advertisements than buying.**

While focusing on renting company should basically concentrate its research and advertisements on Brooklyn and Manhattan going with the word correlation obtained from Tweets. This is evident from figure 7. Also, advertisements for rent should be focused on apartments rather than houses. This will provide a better search result as the people prefer rented apartments as clear from Figure 6 and 4 which indicate the correlation of rent with apartments.

For the average prices of rented apartments in New York Figure 7 shows a figure of 3400 and average indicating the rent is related to average and rent 3400. This is an indicative of the fact that the average prices of apartments in locations like Manhattan is around $3400. Hence, we can keep the advertisements for Manhattan with the average price of $3400.

## *PRICING MODEL*

599 occurrences of rent and similarly 170 odd occurrences of buy, in a week's tweet suggest that a lot of people weekly look for renting an apartment. Hence, providing a separate website where in they can easily look for apartment will be a great help for them. Going with the quantity of words in a week's tweet suggest that they will not only cherish this idea behind website but are also, in dire need of a platform to look for rented apartments.

Also, people who stay on rent are vivacious and may change their apartments quite often. Hence, they can be a permanent customers to our website.

Therefore, we can have a one-time subscription fee for such people and then they can keep on searching for rented apartments through a passage of time after which they will get their subscription renewed in less prices.

For people who want to buy a house, we can have a onetime subscription and breakage fee, since they may not be permanent customers. Because people do not generally look for buying a house often, hence, one time breakage will earn good profit for the company.