DATA INTENSIVE COMPUTING

Statistical Analysis to support new data product

Ramanpreet Singh Khinda | DIC 587 | March 5, 2016
rkhinda@buffalo.edu
50169622

Problem 4: Collecting and Analyzing data on NYC Apartments and recommend a plan to Real Direct to offer Apartment Rentals as a product

Step 1: Parsing Tweets

In this step we will parse the tweets on NYC Apartments (tweets on apartments and home sales in Manhattan, Brooklyn, Bronx and Staten Island) which we collected over 1 week during problem 1

First we parsed the tweets and tweets and extracted meaningful information and stored it as a Data Frame

```
tweets_manhattan <- fromJSON(file = "Tweets_Manhattan.json")

df_manhattan <- data.frame(tweets_manhattan[[1]]$id_str, tweets_manhattan[[1]]$created_at,
                           tweets_manhattan[[1]]$text, tweets_manhattan[[1]]$favorite_count,
                           tweets_manhattan[[1]]$retweet_count, tweets_manhattan[[1]]$user$id_str,
                           tweets_manhattan[[1]]$user$verified, tweets_manhattan[[1]]$user$followers_count,
                           tweets_manhattan[[1]]$user$friends_count, tweets_manhattan[[1]]$user$listed_count,
                           tweets_manhattan[[1]]$user$statuses_count, tweets_manhattan[[1]]$user$location)
df_manhattan$data_location <- "Manhattan"
df_manhattan$sentiment_buy_home <- "Neutral"
df_manhattan$sentiment_rent_apt <- "Neutral"
```

Step 2: Performing Sentiment Analysis on the tweets

In this step we will perform sentiment analysis on each tweet and analyze whether the user is talking about renting an apartment or buying an apartment.

This analyses will be our backing information for suggesting a new product to Real Direct.

Below is the script to perform sentiment analysis.

```
repeat {
  print(paste("Creating Data Frame for Manhattan Tweets Data. Collected ", x," Tweets" ,sep=""))
  temp_df_manhattan <- data.frame(tweets_manhattan[[x]]$id_str, tweets_manhattan[[x]]$created_at, tweets_manhattan[[x]]$text
  temp_df_manhattan$data_location <- "Manhattan"
  temp_df_manhattan$sentiment_buy_home <- "Neutral"
  temp_df_manhattan$sentiment_rent_apt <- "Neutral"

  colnames(temp_df_manhattan) <- c("tweet_id","tweet_created_at","tweet_text","tweet_favorite_count","tweet_retweet_count","

  if((grepl("buy", temp_df_manhattan$tweet_text, ignore.case = TRUE) | grepl("purchase", temp_df_manhattan$tweet_text, ignor
     & (grepl("house", temp_df_manhattan$tweet_text, ignore.case = TRUE)   | grepl("housing", temp_df_manhattan$tweet_text,
     & (grepl("expensive", temp_df_manhattan$tweet_text, ignore.case = TRUE) | grepl("high price", temp_df_manhattan$tweet_t
    temp_df_manhattan$sentiment_buy_home <- "Expensive"
  } else if((grepl("buy", temp_df_manhattan$tweet_text, ignore.case = TRUE) | grepl("purchase", temp_df_manhattan$tweet_text
            & (grepl("house", temp_df_manhattan$tweet_text, ignore.case = TRUE)  | grepl("housing", temp_df_manhattan$tweet_
    temp_df_manhattan$sentiment_buy_home <- "Affordable"
  }

  # sentiment for renting home
  if((grepl("rent", temp_df_manhattan$tweet_text, ignore.case = TRUE) | grepl("rental", temp_df_manhattan$tweet_text, ignore
     | (grepl("house", temp_df_manhattan$tweet_text, ignore.case = TRUE) | grepl("housing", temp_df_manhattan$tweet_text, ig
     & (grepl("expensive", temp_df_manhattan$tweet_text, ignore.case = TRUE) | grepl("high price", temp_df_manhattan$tweet_t
    temp_df_manhattan$sentiment_rent_apt <- "Expensive"
  } else if((grepl("rent", temp_df_manhattan$tweet_text, ignore.case = TRUE) | grepl("rental", temp_df_manhattan$tweet_text,
            | (grepl("house", temp_df_manhattan$tweet_text, ignore.case = TRUE) | grepl("housing", temp_df_manhattan$tweet_t
    temp_df_manhattan$sentiment_rent_apt <- "Affordable"
  }

  df_manhattan <- rbind(df_manhattan, temp_df_manhattan)
  x = x+1
  if (x > count_manhattan_tweets){
    break
  }
}
df_manhattan = unique(df_manhattan)
```

After performing the analysis, we get to know how many users are talking about renting vs buying an apartment and also their response on cost of living at a particular place

This is the result of Sentiment analysis on individual locations

```
> head(df_manhattan)
            tweet_id               tweet_created_at
1 705536615440056321 Thu Mar 03 23:33:32 +0000 2016
2 704297853708079104 Mon Feb 29 13:31:08 +0000 2016
3 703942120583860224 Sun Feb 28 13:57:34 +0000 2016
4 705529062328897536 Thu Mar 03 23:03:31 +0000 2016
5 705503297952735232 Thu Mar 03 21:21:08 +0000 2016
6 703941373196529664 Sun Feb 28 13:54:36 +0000 2016

tweet_text
1                                  how does anyone move in manhattan like ever? this is so expensive and hard to find a
not-shithole :O
2 The rent that this Manhattan office asked in January hit a record high. #nyrealestate https://t.co/TKuLG2BKth https:/
/t.co/t7Cx61rGrV
3                        RT @archangelcrew: Fiddler On The High Rent District Roof (Manhattan Beach Opening only) #Rejec
tedBroadwayPlays
4                                Most Expensive Residence in #Manhattan - Penthouse at the #RitzCarlton #Serrini https:/
/t.co/ggfeIS5Xki
5                     Smaller Manhattan #condos Keep Getting More Expensive https://t.co/kqCZTQFKgi via @CurbedNY https:/
/t.co/kEwuszPwqQ
6                                          Fiddler On The High Rent District Roof (Manhattan Beach Opening only) #Rejec
tedBroadwayPlays
  tweet_favorite_count tweet_retweet_count    user_id user_verified user_followers_count user_friends_count
1                    0                   0  213180708         FALSE                   62                 88
2                    0                   0 3418262054         FALSE                  106                533
3                    0                   2   31589284         FALSE                 1763               2019
4                    1                   0 2784747340         FALSE                   24                 15
5                    0                   0  126354240         FALSE                  665                744
6                    1                   2 2728424702         FALSE                 9322               5836
  user_listed_count user_statuses_count    user_location data_location sentiment_buy_home sentiment_rent_apt
1                 2                 162              nyc     Manhattan            Neutral          Expensive
2                17                 436                      Manhattan            Neutral          Expensive
3               254               19051         New York     Manhattan            Neutral          Expensive
4                16                1107   the interwebs      Manhattan            Neutral          Expensive
5                60                5048 White Plains, NY     Manhattan            Neutral            Neutral
6               158               36283  Bakersfield, CA     Manhattan            Neutral          Expensive
```

```
  tweet_favorite_count tweet_retweet_count    user_id user_verified user_followers_count user_friends_count
2                    0                   1 2786669724             0                1366                343
3                    0                   0 4784683639             0                 183                  0
4                    0                   0 2868205074             0                 141                  0
5                    0                   1  480002469             0                 536               2269
6                    3                   1    1797991             1               53080               5038
7                    6                   1    9917512             1               59095                185
  user_listed_count user_statuses_count           user_location data_location sentiment_buy_home sentiment_rent_apt
2                52                8453            New York, NY      brooklyn         Affordable         Affordable
3               142               42126                              brooklyn         Affordable         Affordable
4               135               51553                              brooklyn            Neutral            Neutral
5                26                8188 New York Public Library      brooklyn            Neutral            Neutral
6              1938                 143            Brooklyn, NY      brooklyn            Neutral            Neutral
7              1857               15681            New York, NY      brooklyn         Affordable         Affordable
```

Step 3: Collective Analysis

In this step we will do collective analysis on tweets of all locations. For this we done simple aggregation based on these 4 criteria: -

1. How many users talk about buying home at a particular location in NYC?
2. How many users talk about renting an apartment?
3. How many of users think its expensive/affordable to rent an apartment?
4. How many of users think its expensive/affordable to buy a home?

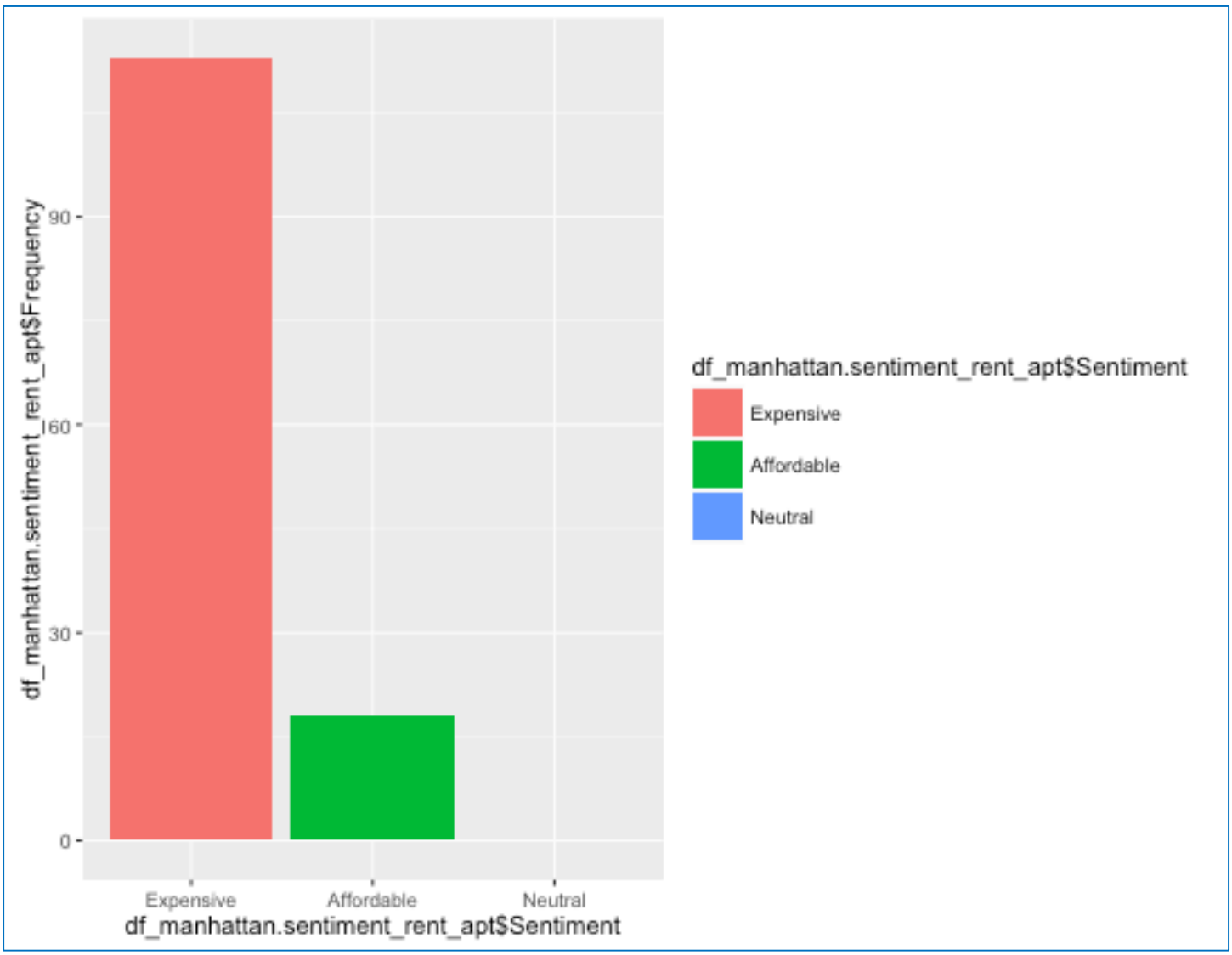
Cumulative Sentiment Analysis

```
> df_sentiment
                  location freq_sentiment_buy_home freq_sentiment_rent_apt
1      Manhattan_Expensive                       0                      113
2              M_Affordable                      3                       18
3                 M_Neutral                    180                       52
4       Brooklyn_Expensive                       0                       49
5            Br_Affordable                       4                      116
6               Br_Neutral                     238                       77
7          Bronx_Expensive                       0                        9
8              B_Affordable                      0                       39
9                 B_Neutral                     94                       46
10 Staten_Island_Expensive                       0                        1
11             S_Affordable                      0                        0
12                S_Neutral                     17                       16
```

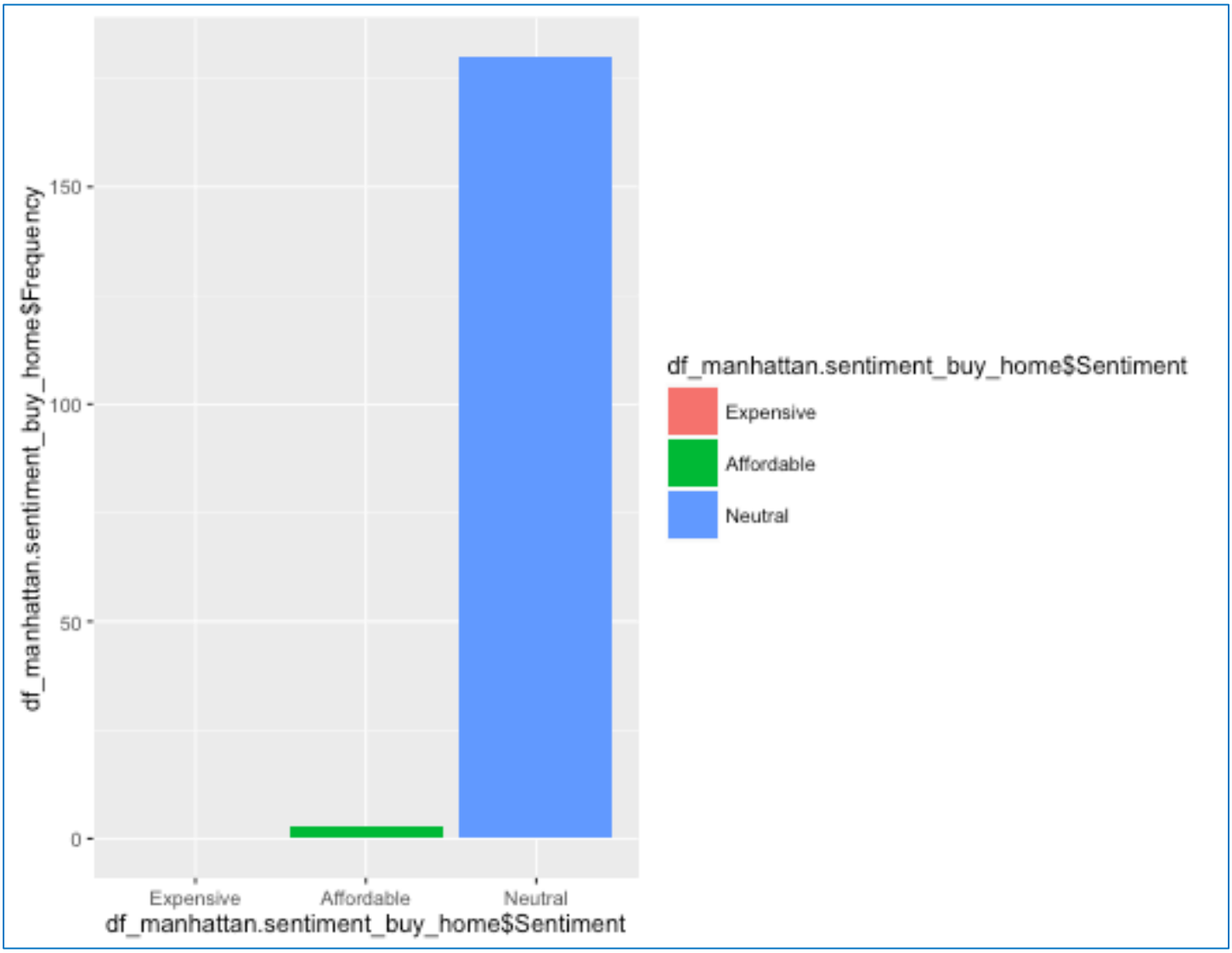
Step 4: Plotting Graphs

Now will backup our data and analysis by graphically

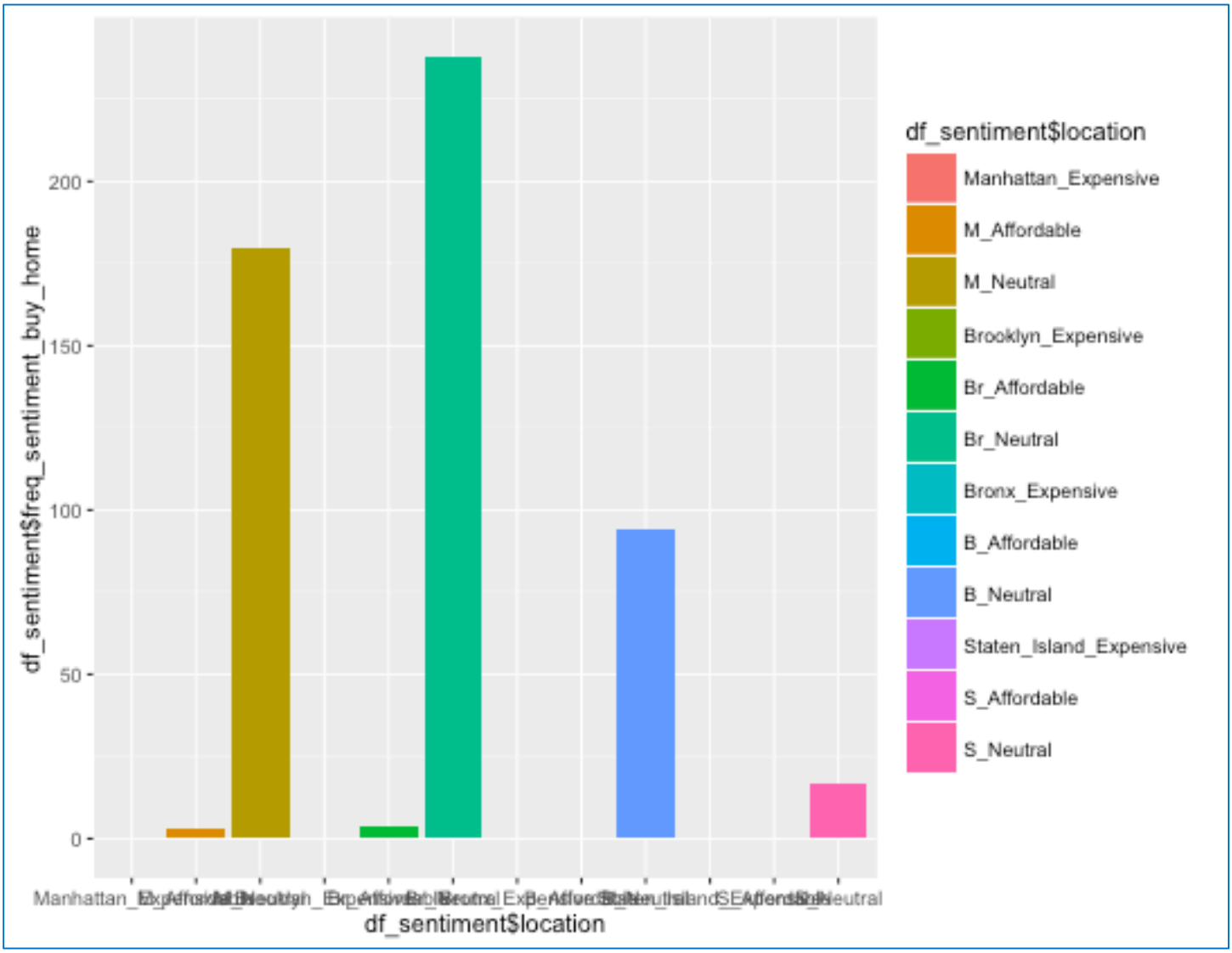
Graph on Users Sentiment on Renting an Apartment in Manhattan vs No. of users



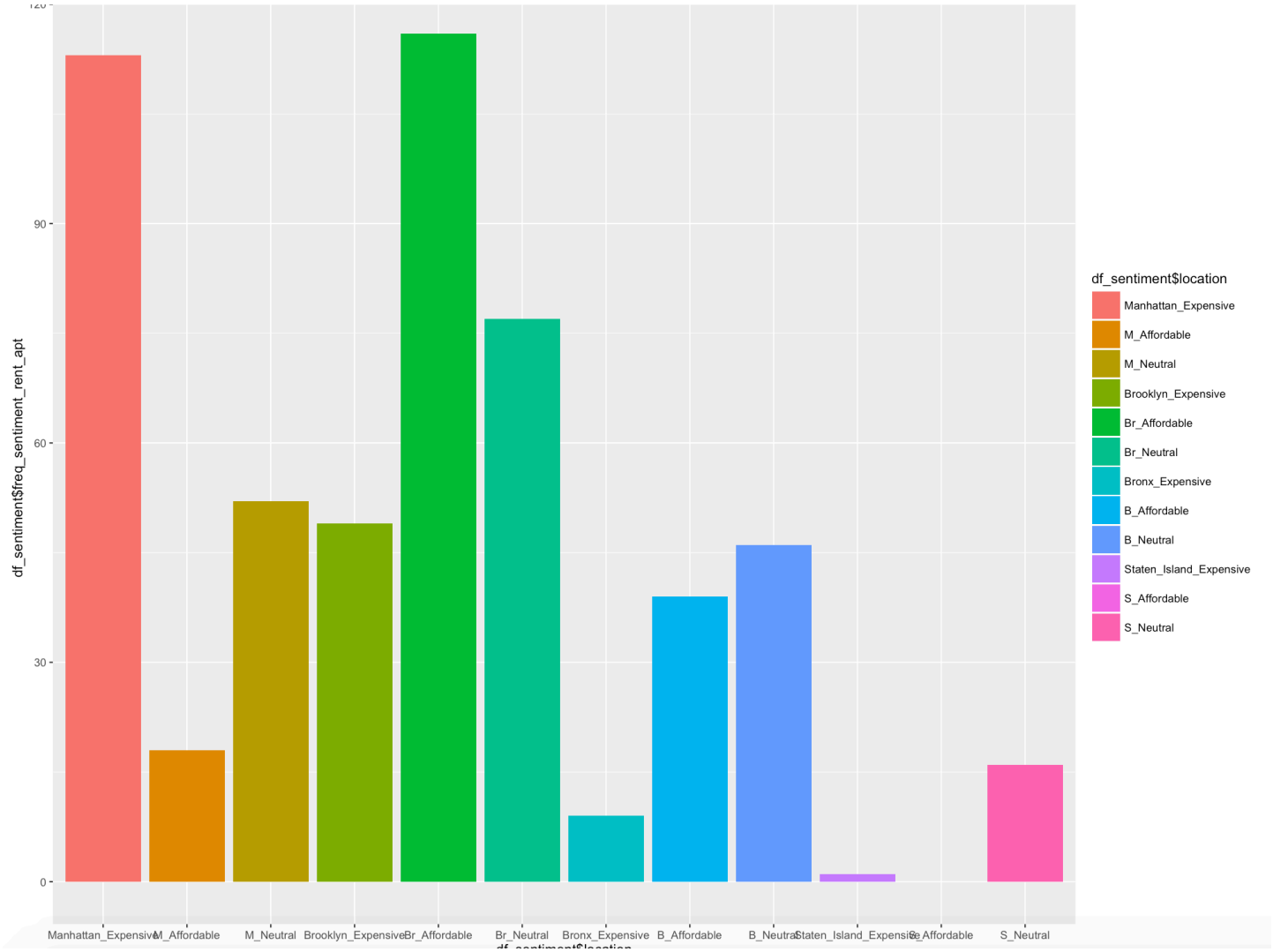
Graph on Users Sentiment on Buying home in Manhattan vs No. of users



Graph on Users Sentiment on buying home at different locations in NYC



Graph on Users Sentiment on Renting an Apartment at different locations in NYC



Results

After analyzing these graphs and watching users sentiments on Renting an Apartment in NYC vs. Buying a home, we came to know that more no. of users are talking about Renting an apartment.

This data was collected for 1 week and we can clearly see the results. Even if we have collected data over past 1 month or year we can easily figure out that more no. of people are towards renting instead of buying.

These graphs also talk about Cost of Living at particular location in NYC and the data support this claim. Based on our analysis we recommend CEO at Real Direct to consider offering Apartment Rentals as a product.