CSE 587: Project 1
Problem 4 -Data Product

Lalith Vikram Natarajan, UBIT ID: lalithvi, Person No: 50169243

March 5, 2016

1 Objective

Combine different data sets and perform statistical analysis for business recommendation

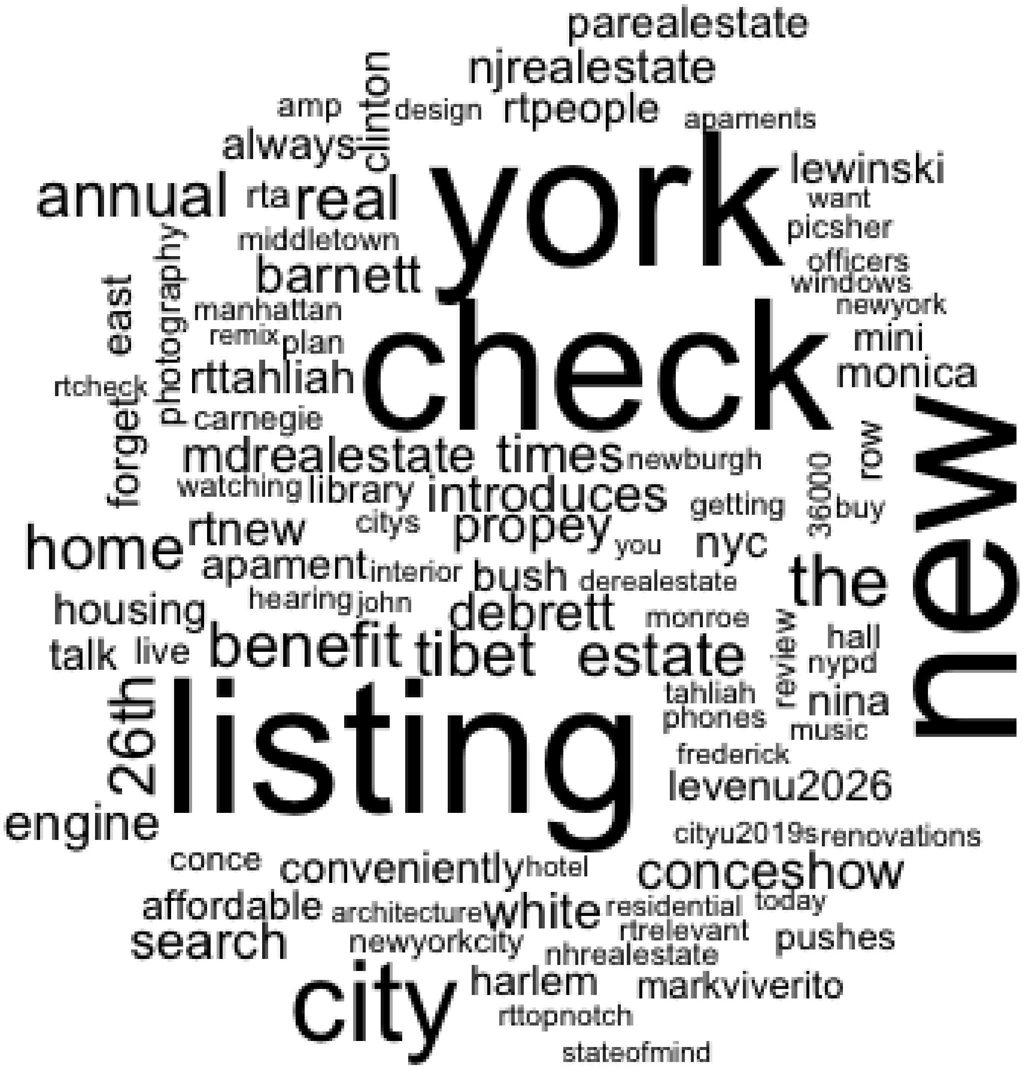
- Collect data of interest from Social Media (Twitter)
- Collect data from .csv file
- Combine them and make a recommendation

2 Data Source

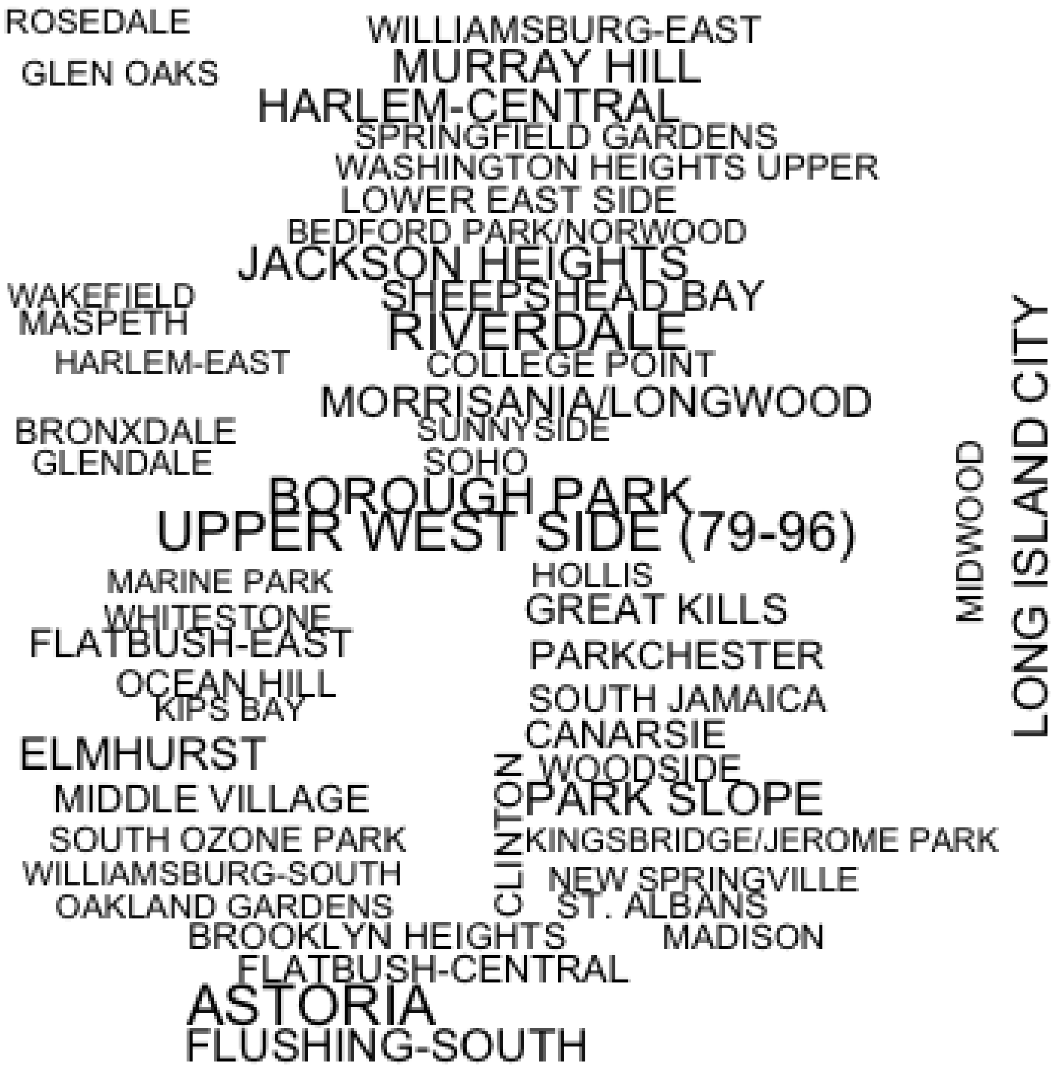
Twitter data was used as the source of data for this exercise. I came up with a list of keywords that would help narrow the search down. They are real estate, housing, residential, realtor, house, foreclosure, condos, apartments, fsbo, buy home., etc. New York, NYC, Manhattan were the other words added along with the above mentioned words. Also, a search was made with the location set to New York's co-ordinates.The data was steadily collected over 10 days. The other data set was the ones provided in the book "Doing Data Science" . It consisted of rolling sales details of five boroughs.

3 Analysis

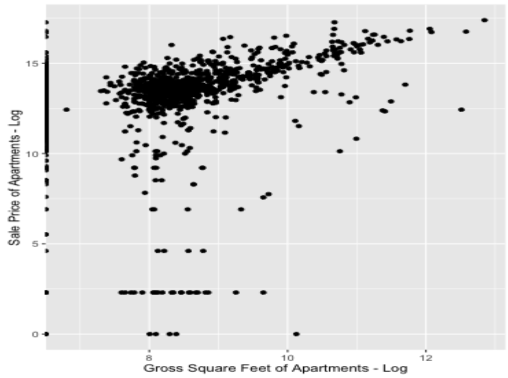
The data was from two different years and of varying fields, but the one field which was common to both was the query text and the neighborhood. Computing the most talked about words from the tweets and the .csv should help us arrive at a relationship. I decided to compute the word cloud of both data sets to establish a connection. This was accomplished using the tm package and the wordcloud package.
Following is the word cloud of the various tweets.



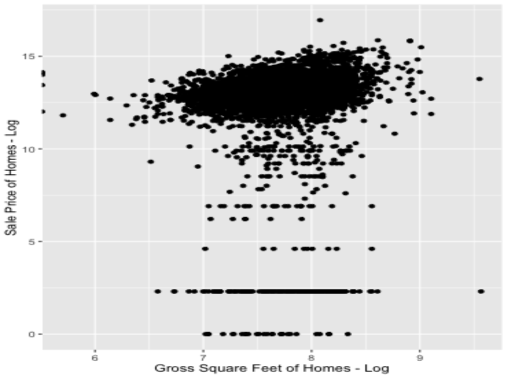
Following is the word cloud of the most popular neighborhoods.



From the above two graphs, one can figure that people do not necessarily talk about places, a better alternative would be the location data, but ideally people do not share location or tweet about it. A larger dataset with appropriate data would have been helpful to strengthen this relationship. However some tweets did contain location mentions such as york, new york city, harlem. etc.



(a) Apartments



(b) Homes

The above plots show map the sales price vs square foot area of an apartment and home. It is found that apartments give higher profit margins. Using this knowledge from the tweets and the .csv's got in the previous years, we understand that people's searching techniques have changed. People are more inclined in apartment renting or selling (shown by the presence of apatments and apatment in the wordcloud). From this we can give two recommendations to RealDirect, one would be to promote apartments than houses at this point. The second one would be to establish a social media presence, try to get people's attention using the top words from the tweet and try to push their best selling neighborhoods through the same medium.