Applied Data Science Capstone: Battle of Neighborhoods

Selecting Neighborhoods in New York City

Arnab Basu
March 2019

1. Purpose:

   This document is a report of the capstone project done as part of the Applied Data Science course and is a mandatory requirement for the IBM Data Science Professional Certificate on Coursera.

2. Introduction:

   New York City is one of the most populous cities in the United States population of almost 9 million people distributed over a land area of about 302.6 square miles. New York is also the most densely populated major city in the United States and also the financial capital of the United States.

   Now, imagine a scenario where you have received a new job opportunity which requires you to relocate to New York City. You have never been to NYC before but need to find a neighborhood to find a rental apartment that meets your needs. As you can imagine, for someone new to any big city this can be daunting task, let along New York.

   Since I have personally faced a similar situation in the past, I am trying to come up with a model that helps others like me make an informed decision.

   In this exercise I am attempting to build a model that helps the user to:

   - Short-list a few neighborhoods in NYC given the following conditions:
     - Average apartment rents are within the budgeted range
     - Preferred minimum school rating

- Neighborhood with more Indian population since this will mean proximity to Indian grocery stores and restaurants
- The neighborhood has a good mix of venues as per the individual's preference

Note: For this exercise I am assuming that user prefers a neighborhood that has more Indian population, grocery stores and restaurants.

3. Data Acquisition and Cleaning:

A. NYC Boroughs and Neighborhoods:

Data on NYC boroughs and neighborhoods were downloaded as JSON from https://geo.nyu.edu

B. NYC Schools:

Data on NYC schools were downloaded from https://www.greatschools.org

C. NYC Demographics and Ethnicity:

Data on NYC neighborhood demographics and ethnicity were obtained from https://data.cityofnewyork.us

D. Geographical Coordinates:

Geographical coordinates for NYC as well as its neighborhoods and venues were retrieved using Geopy.

E. List of popular venues:

Foursquare API was used to explore the neighborhoods in the city and retrieve the data pertaining to venues in each neighborhood. The data obtained from Foursquare was used to get the most common venues in each neighborhood and in turn group the neighborhoods into clusters.

For the sake of simplicity some parts of the data wrangling were done offline.
The original data-frame has 5 boroughs and 306 neighborhoods

I observed that some neighborhood names (for e.g. both Manhattan and Staten Island has neighborhoods called Chelsea) are common between more than one borough. In order to overcome this, I introduced another column that is a combination of neighborhoods and boroughs. Thus, giving us 306 unique combinations of neighborhoods and boroughs.

The school rating, average rent and the demographics/ethnicity data was then merged to form another data set.

4. Methodology:

A. Data Grouping:

The NYC Boroughs and Neighborhoods data was downloaded as JSON and the same was then converted into a data-frame with the following columns:

1. Neighborhood
2. Borough
3. Latitude
4. Longitude

As mentioned above, since there are instances of neighborhoods that have the same name but are present in more than one borough, another column ("Neighborhood with Borough") was added to the data-frame where the borough was appended to the neighborhood name in order to arrive at the 306 unique combinations of neighborhoods and boroughs. Next, the 'Neighborhood' and 'Borough' columns were dropped since these were not be needed individually.

For the data on average rent, school rating and demographics the data wrangling was done off-line and merged into a CSV file for easier processing.

B. *Fetch geographical coordinates:*

Geopy was used to get the latitude and longitude values for New York City.

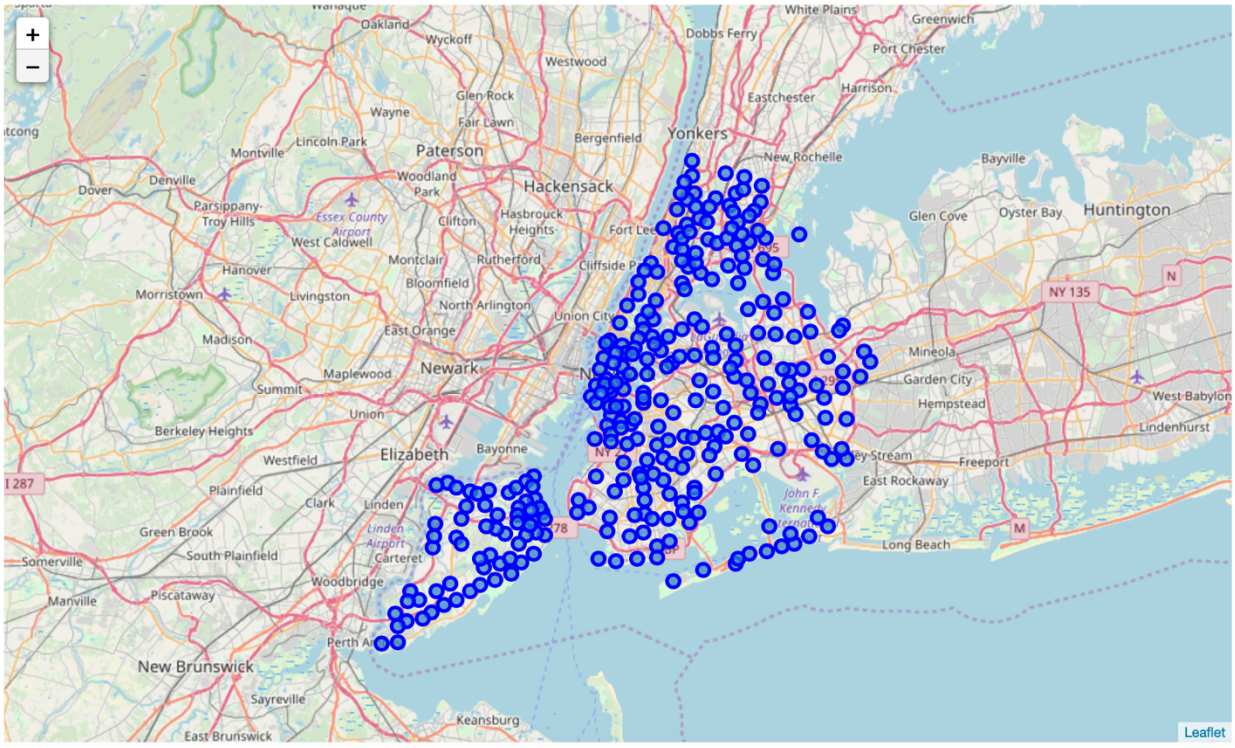
```
# Get the latitude and longitude values of New York City
address = 'New York City, NY'

geolocator = Nominatim(user_agent="BattleOfNeighborhoods_NYC")
location = geolocator.geocode(address)
latitudeNYC = location.latitude
longitudeNYC = location.longitude
print('The geograpical coordinate of New York City are {}, {}.'.format(latitudeNYC, longitudeNYC))
```

```
The geograpical coordinate of New York City are 40.7308619, -73.9871558.
```

C. *Processing the neighborhood data:*

We created a map of New York to show the neighborhoods using latitude and longitude values.



Then, the Foursquare API was invoked to retrieve the details (name, category, latitude and longitude) of nearby venues pertaining to each neighborhood.

```
NYC_venues.head()
```

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Wakefield * Bronx	40.894705	-73.847201	Lollipops Gelato	40.894123	-73.845892	Dessert Shop
1	Wakefield * Bronx	40.894705	-73.847201	Rite Aid	40.896649	-73.844846	Pharmacy
2	Wakefield * Bronx	40.894705	-73.847201	Jackie's West Indian Bakery	40.889283	-73.843310	Caribbean Restaurant
3	Wakefield * Bronx	40.894705	-73.847201	Carvel Ice Cream	40.890487	-73.848568	Ice Cream Shop
4	Wakefield * Bronx	40.894705	-73.847201	Cooler Runnings Jamaican Restaurant Inc	40.898276	-73.850381	Caribbean Restaurant

There are 449 unique categories.

The venues were then encoded followed by finding out the list of top 10 venues for each neighborhood.

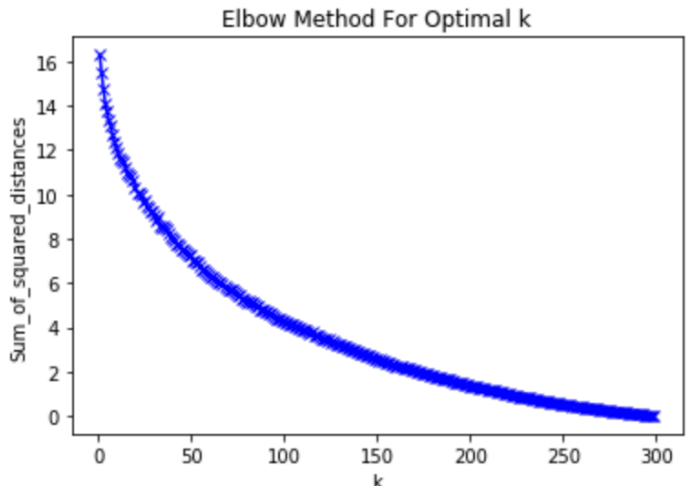
```
----1----Allerton * Bronx----
                  venue  freq
0            Donut Shop  0.12
1        Sandwich Place  0.09
2           Supermarket  0.06
3           Bus Station  0.06
4                  Bank  0.06
5           Pizza Place  0.06
6  Fast Food Restaurant  0.06
7              Pharmacy  0.06
8       Clothing Store  0.03
9     Mobile Phone Shop  0.03


----2----Annadale * Staten Island----
                         venue  freq
0                  Pizza Place   0.3
1                   Restaurant   0.2
2                Cosmetics Shop   0.1
3                    Sports Bar   0.1
4                 Train Station   0.1
5                        Diner   0.1
6                  Dance Studio   0.1
7           Rental Car Location   0.0
8  Paper / Office Supplies Store   0.0
9                     Pet Store   0.0
```

This was then transformed into a data-frame:

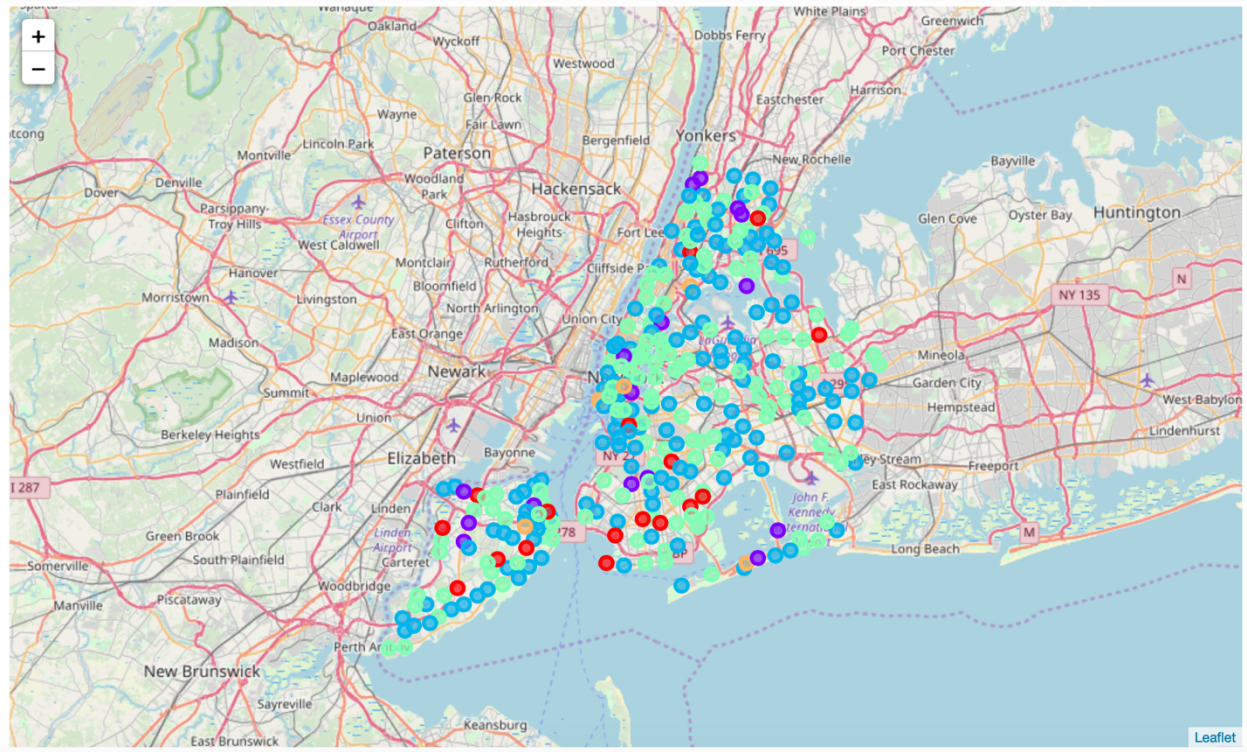
	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Allerton * Bronx	Donut Shop	Sandwich Place	Supermarket	Bus Station	Fast Food Restaurant	Pharmacy	Pizza Place	Bank	Fried Chicken Joint	Dessert Shop
1	Annadale * Staten Island	Pizza Place	Restaurant	Cosmetics Shop	Dance Studio	Diner	Train Station	Sports Bar	Zoo	Farm	Fast Food Restaurant
2	Arden Heights * Staten Island	Pharmacy	Garden Center	Home Service	Bus Stop	Pizza Place	Food	Coffee Shop	Elementary School	Deli / Bodega	Factory
3	Arlington * Staten Island	Bus Stop	Polish Restaurant	Grocery Store	Home Service	Coffee Shop	Business Service	Boat or Ferry	American Restaurant	Intersection	Snack Place
4	Arrochar * Staten Island	Bus Stop	Pizza Place	Deli / Bodega	Italian Restaurant	Beach	Cosmetics Shop	Playground	Sandwich Place	Sculpture Garden	Outdoors & Recreation

Next, I ran the unsupervised machine learning algorithm k-means to cluster the neighborhood based on different categories of places in each neighborhood. But before doing that I needed to determine the determine the optimal number of clusters for k-means clustering using elbow methods.



Since there are 306 neighborhoods and even k = 300 does not yield a well-defined elbow, for the sake of simplicity we will assume that for this exercise k = 5.

The neighborhoods were then clustered and the corresponding map generated.



After this, the neighborhood venues data was analyzed along with the data on average rent, school rating and ethnic diversity.

Here is a snap-shot of the data set:

	Neighborhood with Borough	School Rating	Population - White	Population - Black	Population - Asian	Population - Hispanic	Population - Hawaiian	Population - Indian	Average Rent
0	Wakefield * Bronx	5	50481.0	4369.0	4995.0	3278.0	924.0	991.0	2664.0
1	Co-op City * Bronx	10	19200.0	4479.0	4376.0	5702.0	996.0	815.0	2564.0
2	Eastchester * Bronx	7	14354.0	3788.0	5554.0	8710.0	682.0	617.0	5158.0
3	Fieldston * Bronx	5	24809.0	1257.0	4052.0	9772.0	171.0	997.0	3868.0
4	Riverdale * Bronx	10	43101.0	5624.0	3387.0	13538.0	136.0	409.0	4557.0

D. *Exploratory Analysis of the rent data:*

	Average Rent
count	307.000000
mean	3971.322476
std	1184.971852
min	1805.000000
25%	2974.500000
50%	4015.000000
75%	5044.000000
max	5977.000000

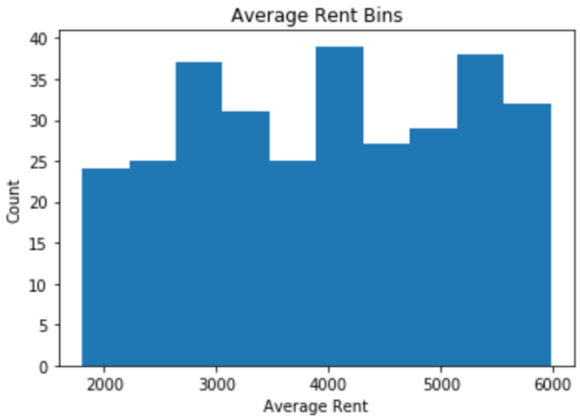
```
#Median
NYC_avg_rent['Average Rent'].median()
```

```
4015.0
```

```
#Mode
NYC_avg_rent['Average Rent'].mode()
```

```
0    1833.0
1    3281.0
2    3867.0
3    3985.0
4    3986.0
5    4441.0
6    4602.0
7    4833.0
8    5289.0
9    5420.0
dtype: float64
```

The histogram of average rent, to see what the distribution looks like:

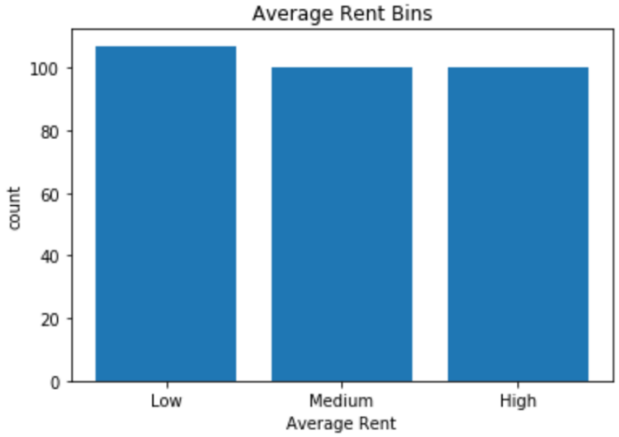


Binning was done on the average monthly rent data as follows:

- Bins = array([1805. , 3195.66666667, 4586.33333333, 5977.])
- Group Names = [Low, Medium, High]
- "Binned Average Rent" Counts per Group:

Group Name	Count
High	107
Medium	100

Low	100

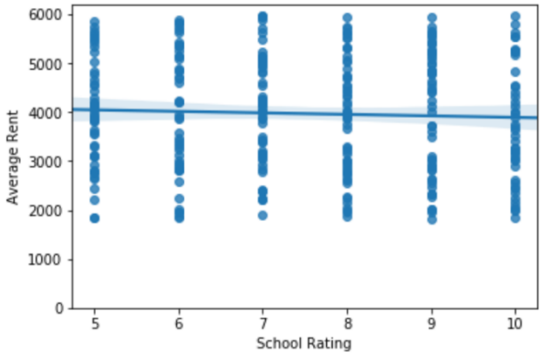


E. Relationship between School Rating and Rent:

This was done to see if School Rating can be potential predictor variable of rent.

```
NYC_schoolrating_rent[["School Rating", "Average Rent"]].corr()
```

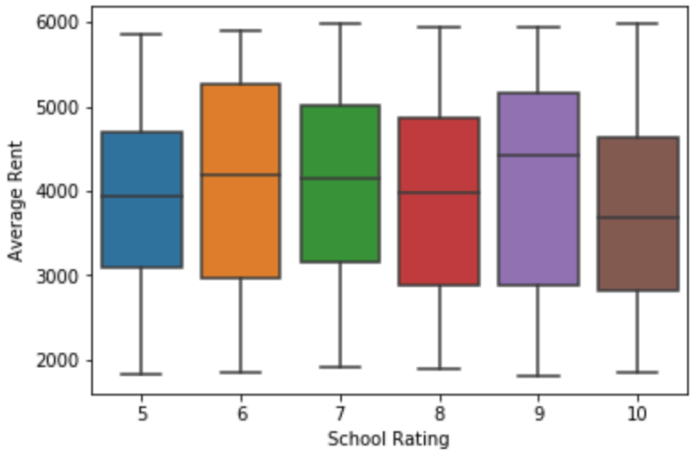
	School Rating	Average Rent
School Rating	1.000000	-0.042825
Average Rent	-0.042825	1.000000



For the given data, school rating does not seem like a good predictor of the rent at all since the regression line is close to horizontal. Also, the data points are very scattered and far from the fitted line, showing lots of variability. Therefore, it is not a reliable variable.

F. Box-plot Analysis:

Distribution of average rent significantly differs with school rating.



5. Results:

For this exercise, I assumed the following user preferences (in the given order) which were taken as inputs while running the notebook:

- Monthly Rent Budget:
  - Min = US$ 3500
  - Max = US$ 3750
- Minimum preferred school rating: 9
- Neighborhood with a sizable Indian population for reasons explained above

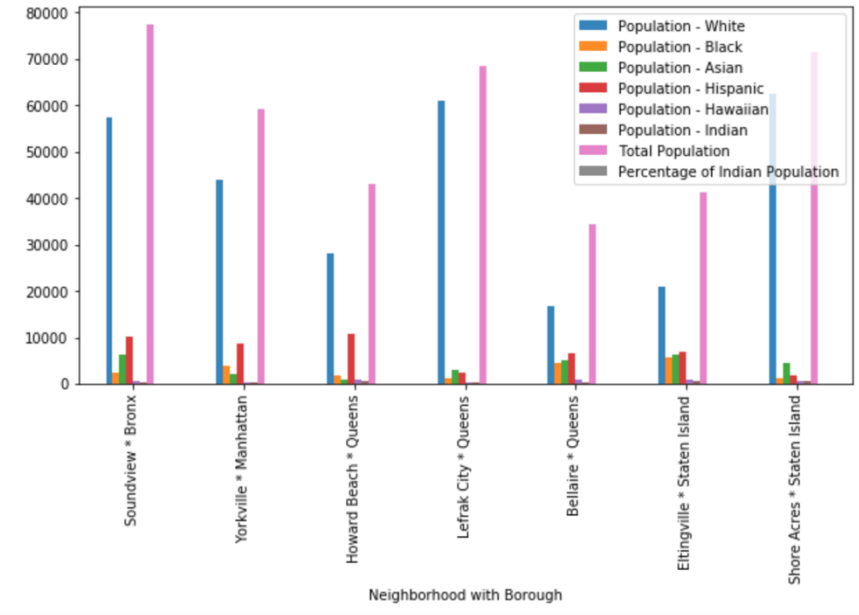
Based on the above inputs, the following neighborhoods were shortlisted:

	Neighborhood with Borough	School Rating	Average Rent	Binned Average Rent	Percentage of Indian Population
5	Howard Beach * Queens	10	3653	Medium	1.447054
10	Eltingville * Staten Island	9	3729	Medium	1.413067
8	Bellaire * Queens	9	3622	Medium	0.954466
12	Shore Acres * Staten Island	10	3694	Medium	0.827706
7	Lefrak City * Queens	10	3524	Medium	0.655047
0	Soundview * Bronx	10	3750	Medium	0.495543
3	Yorkville * Manhattan	9	3714	Medium	0.334031

Ten most common venues in the short-listed neighborhoods along with cluster labels:

Neighborhood with Borough	Soundview * Bronx	Yorkville * Manhattan	Howard Beach * Queens	Lefrak City * Queens	Bellaire * Queens	Eltingville * Staten Island	Shore Acres * Staten Island
Latitude	40.821	40.7759	40.6542	40.7361	40.733	40.5422	40.6097
Longitude	-73.8657	-73.9471	-73.8381	-73.8625	-73.7389	-74.1643	-74.0667
Cluster Labels	0	3	1	0	3	2	0
1st Most Common Venue	Fried Chicken Joint	Italian Restaurant	Italian Restaurant	Clothing Store	Pizza Place	Sushi Restaurant	Italian Restaurant
2nd Most Common Venue	Fast Food Restaurant	Gym	Park	Bakery	Chinese Restaurant	Pizza Place	Deli / Bodega
3rd Most Common Venue	Deli / Bodega	Coffee Shop	Sandwich Place	Cosmetics Shop	Deli / Bodega	Bank	Bus Stop
4th Most Common Venue	Chinese Restaurant	Bar	Chinese Restaurant	Shoe Store	Convenience Store	Sandwich Place	Bar
5th Most Common Venue	Clothing Store	Mexican Restaurant	Bagel Shop	Fast Food Restaurant	Fast Food Restaurant	Pharmacy	Intersection
6th Most Common Venue	Breakfast Spot	Pizza Place	Sushi Restaurant	Coffee Shop	Italian Restaurant	Italian Restaurant	Gastropub
7th Most Common Venue	Pizza Place	Japanese Restaurant	Ice Cream Shop	Pizza Place	Moving Target	Diner	Furniture / Home Store
8th Most Common Venue	Food	Bagel Shop	Fast Food Restaurant	Department Store	Sushi Restaurant	Fast Food Restaurant	Filipino Restaurant
9th Most Common Venue	Southern / Soul Food Restaurant	Ice Cream Shop	Fried Chicken Joint	Ice Cream Shop	Bus Station	Bagel Shop	Pharmacy
10th Most Common Venue	Grocery Store	Deli / Bodega	Pharmacy	Sandwich Place	Bus Stop	Chinese Restaurant	Donut Shop

Ethnic mix amongst the population in the short-listed neighborhoods:



6. Discussion:

   Based on the input data and the user preference, the we arrived at the following neighborhoods:

   1. Soundview * Bronx
   2. Yorkville * Manhattan
   3. Howard Beach * Queens
   4. Lefrak City * Queens
   5. Bellaire * Queens
   6. Eltingville * Staten Island
   7. Shore Acres * Staten Island

   From 306 neighborhoods we have been able to short-list 7 neighborhoods.
   Now it is up to the user to finalize on a neighborhood based on individual preference, venues, apartment availability etc.

7. Conclusion:

   Here, I want to highlight the possible ways that this model can be enriched and made more generic:

   1. Accuracy of the model can be improved
   2. Other algorithms can be explored to see if they yield more accurate results
   3. The model can be made generic in terms neighborhood ethnicity selection
   4. The model can be extended to take user's venue preference as inputs and use that to come up with the short-listed neighborhoods