Data acquisition and cleaning

1 Data sources

Neighborhood has a total of 5 boroughs and 306 neighborhoods. In order to segment the neighborhoods and explore them, we will essentially need a dataset that contains the 5 boroughs and the neighborhoods that exist in each borough as well as the latitude and longitude coordinates of each neighborhood.

Luckily, this dataset exists for free on the web; here are the links of the dataset https://geo.nyu.edu/catalog/nyu_2451_34572, https://cocl.us/new_york_dataset . Further we will use Foursquare API to fetch the most common places of every neighborhood to predict the best place to open restaurant.

2 Data cleaning and Analyzing

Data downloaded or scraped from multiple sources were combined into one table. There were a lot of missing values from earlier seasons, because of lack of record keeping.

Since it is a json data we need all relevant data is in features key, which is basically a list of neighborhoods. So we defined a new variable that includes this data. After that we transform the data into Pandas dataframe as it is initially in Python Dictionaries. After cleaning dataframe will look like:

	Borough	Neighborhood	Latitude	Longitude
0	Bronx	Wakefield	40.894705	-73.847201
1	Bronx	Co-op City	40.874294	-73.829939
2	Bronx	Eastchester	40.887556	-73.827806
3	Bronx	Fieldston	40.895437	-73.905643
4	Bronx	Riverdale	40.890834	-73.912585

Neighborhood has a total of 5 boroughs and 306 neighborhoods.