Introduction

The objective of this project is to help Department of Health and Mental Hygiene explore neighborhoods with the worst restaurants. It will help make a smart and efficient decision to choose to inspect restaurants in Brooklyn neighborhoods.

This project aims to create an analysis of the characteristics of restaurants in Brooklyn neighborhoods according to the violations frequently recorded in each establishment by Department of Health and Mental Hygiene.

Problem Which Tried to Solve

The problem tried to solve is that of facilitating the targeted planning of inspections by the Department of Health and Mental Hygiene

The main objective of this project is to recommend the areas of Brooklyn where the violations by restaurants are the most frequent

Description of data set

The dataset contains every sustained or not yet adjudicated violation citation from every full or special program inspection conducted up to three years prior to the most recent inspection for restaurants and college cafeterias in an active status on the RECORD DATE (date of the data pull). When an inspection results in more than one violation, values for associated fields are repeated for each additional violation record. Establishments are uniquely identified by their CAMIS (record ID) number. Keep in mind that thousands of restaurants start business and go out of business every year; only restaurants in an active status are included

In the dataset.

Records are also included for each restaurant that has applied for a permit but has not yet been inspected and for inspections resulting in no violations. Establishments with inspection date of 1/1/1900 are new establishments that have not yet received an inspection. Restaurants that received no violations are represented by a single row and coded as having no violations using the ACTION field. Because this dataset is compiled from several large administrative data systems, it contains some illogical values that could be a result of data entry or transfer errors. Data may also be missing. This dataset and the information on the Health Department's Restaurant Grading website come from the same data source. The Health Department's Restaurant Grading website is here: http://www1.nyc.gov/site/doh/services/restaurant-grades.page

Column Name	Description	Type
CAMIS	This is an unique identifier for the entity (restaurant); 10-digit integer, static per restaurant permit	Plain Text
DBA	This field represents the name (doing business as) of the entity (restaurant); Public business name, may change at discretion of restaurant owner	Plain Text
BORO	Borough in which the entity (restaurant) is located.;• 1 = MANHATTAN • 2 = BRONX • 3 = BROOKLYN • 4 = QUEENS • 5 = STATEN ISLAND • Missing; NOTE: There may be discrepancies between zip code and listed boro due to differences in an establishment's mailing address and physical location	Plain Text
BUILDING	Building number for establishment (restaurant) location	Plain Text
STREET	Street name for establishment (restaurant) location	Plain Text
ZIPCODE	Zip code of establishment (restaurant) location	Plain Text
PHONE	Phone Number; Phone number provided by restaurant owner/manager	Plain Text
CUISINE DESCRIPTION	This field describes the entity (restaurant) cuisine. ; Optional field provided by provided by restaurant owner/manager	Plain Text
INSPECTION DATE	This field represents the date of inspection; NOTE: Inspection dates of 1/1/1900 mean an establishment has not yet had an inspection	Date & Time
ACTION	This field represents the actions that is associated with each restaurant inspection. ; • Violations were cited in the following area(s). • No violations were recorded at the time of this inspection. • Establishment re-opened by DOHMH • Establishment re-closed by DOHMH • Establishment Closed by DOHMH. Violations were cited in the following area(s) and those requiring immediate action were addressed. • "Missing" = not yet inspected;	Plain Text
VIOLATION CODE	Violation code associated with an establishment (restaurant) inspection	Plain Text
VIOLATION DESCRIPTION	Violation description associated with an establishment (restaurant) inspection	Plain Text
CRITICAL FLAG	Indicator of critical violation; "• Critical • Not Critical • Not Applicable"; Critical violations are those most likely to contribute to food-borne illness	Plain Text

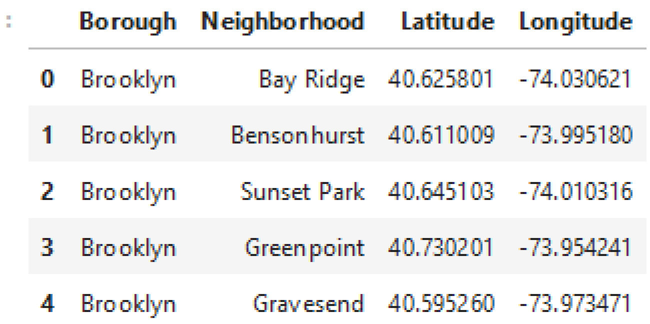
Column Name	Description	Type
SCORE	Total score for a particular inspection; Scores are updated based on adjudication results	Number
GRADE	Grade associated with the inspection; • N = Not Yet Graded• A = Grade A• B = Grade B• C = Grade C• Z = Grade Pending• P= Grade Pending issued on re-opening following an initial inspection that resulted in a closure	Plain Text
GRADE DATE	The date when the current grade was issued to the entity (restaurant)	Date & Time
RECORD DATE	The date when the extract was run to produce this data set	Date & Time
INSPECTION TYPE	A combination of the inspection program and the type of inspection performed; See Data Dictionary for full list of expected values	Plain Text
Latitude		Number
Longitude		Number
Community Board		Plain Text
Council District		Plain Text
Census Tract		Plain Text
BIN		Plain Text
BBL		Plain Text
NTA		Plain Text

We will need reliable location data from locations in different neighborhoods in the Brooklyn borough. In order to obtain this information, we will use the "Foursquare" location information. Foursquare is a location data provider with information on all kinds of locations and events in an area of interest. This information includes place names, locations, menus and even photos. After finding the list of neighborhoods, we then log on to the Foursquare API to collect information about the locations in each neighborhood. For each neighbourhood, we chose the 100-metre radius The data extracted from Foursquare contained information about the sites at a specified distance from the longitude and latitude of the postcodes. The information obtained by site as follows:

- Neighborhood
- Neighborhood Latitude
- Neighborhood Longitude
- Venue
- Name of the venue e.g. the name of a store or restaurant
- Venue Latitude
- Venue Longitude
- Venue Category

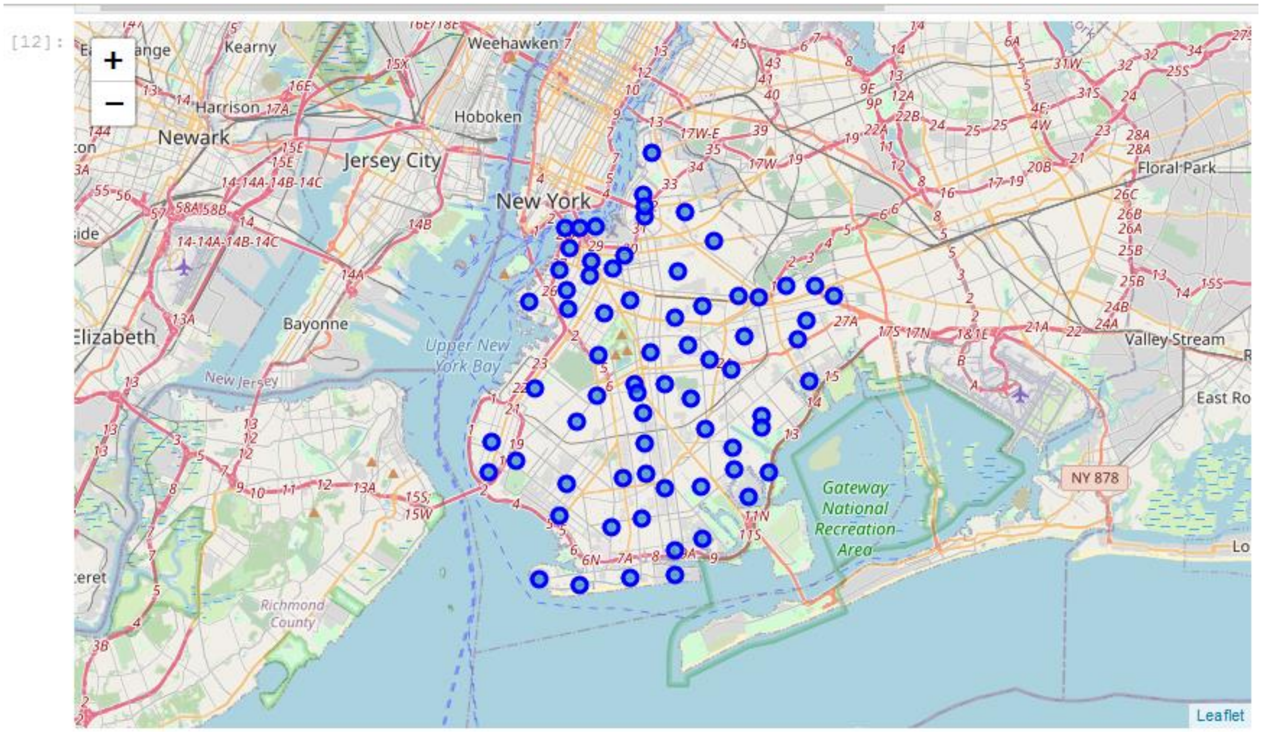
Methodology

From the data available on the sitehttps://geo.nyu.edu/catalog/nyu_2451_34572, I downloaded the latitude and longitude coordinates of all neighborhoods in the borough of Brooklyn.



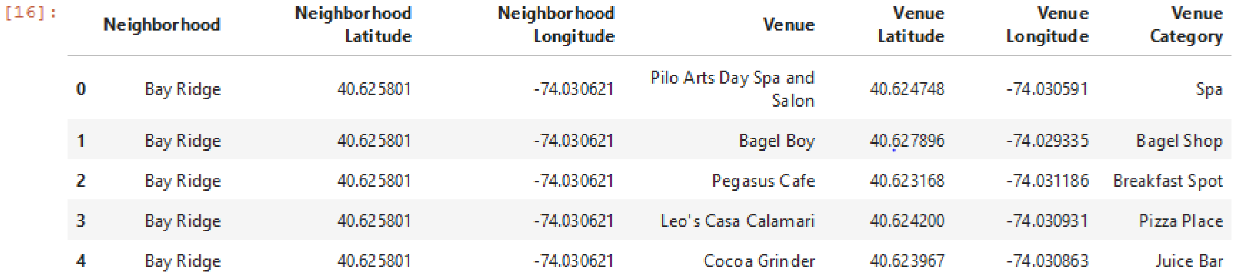
	Borough	Neighborhood	Latitude	Longitude
0	Brooklyn	Bay Ridge	40.625801	-74.030621
1	Brooklyn	Bensonhurst	40.611009	-73.995180
2	Brooklyn	Sunset Park	40.645103	-74.010316
3	Brooklyn	Greenpoint	40.730201	-73.954241
4	Brooklyn	Gravesend	40.595260	-73.973471

I used the folium python library to visualize the geographic details of Brooklyn and its neighborhoods and created a map of Brooklyn with neighborhoods overlaid. I used the latitude and longitude values to get the visual as below



[12]:

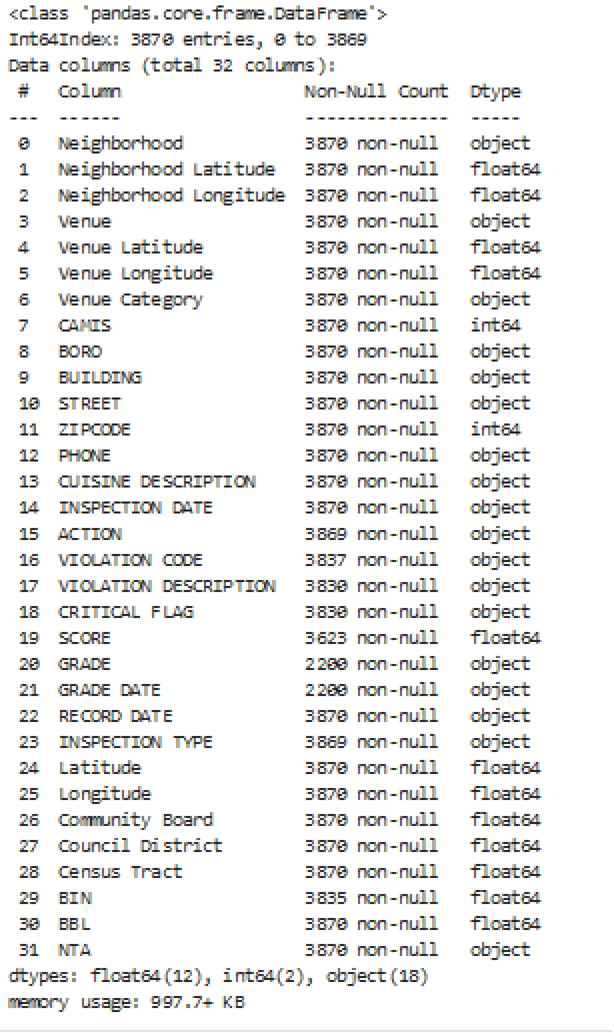
I used the Foursquare API to explore neighborhoods and segment them. I designed the limit as 100 sites and the radius as 500 meters for each district from their given latitude and longitude information. Here is a header of the Forsquare API location name, category, latitude and longitude information.



[16]:

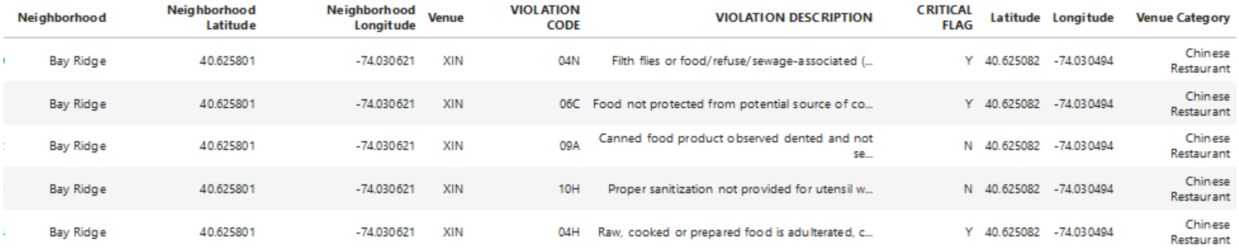
	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Bay Ridge	40.625801	-74.030621	Pilo Arts Day Spa and Salon	40.624748	-74.030591	Spa
1	Bay Ridge	40.625801	-74.030621	Bagel Boy	40.627896	-74.029335	Bagel Shop
2	Bay Ridge	40.625801	-74.030621	Pegasus Cafe	40.623168	-74.031186	Breakfast Spot
3	Bay Ridge	40.625801	-74.030621	Leo's Casa Calamari	40.624200	-74.030931	Pizza Place
4	Bay Ridge	40.625801	-74.030621	Cocoa Grinder	40.623967	-74.030863	Juice Bar

from the places I made the join with the dataset of the violations committed by the restaurants provided by the Department of Health and Mental Hygiene. I get



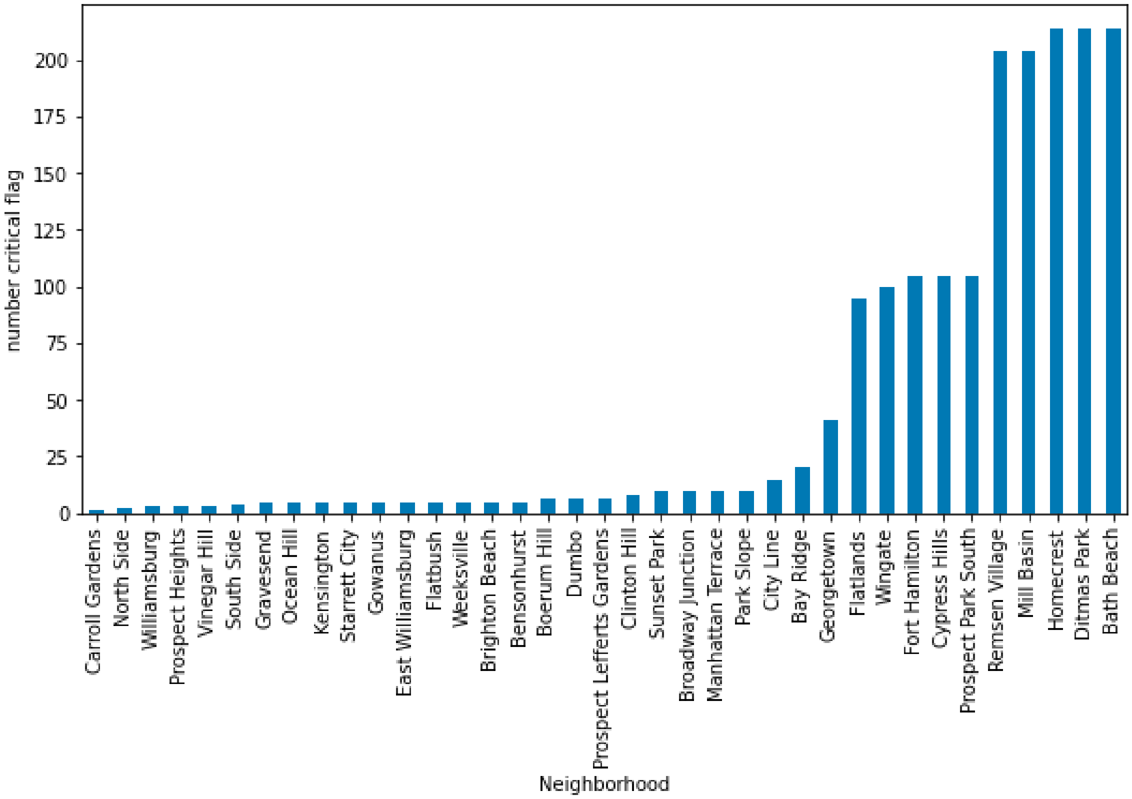
```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 3870 entries, 0 to 3869
Data columns (total 32 columns):
 #   Column                  Non-Null Count  Dtype  
---  ------                  --------------  -----  
 0   Neighborhood            3870 non-null   object 
 1   Neighborhood Latitude   3870 non-null   float64
 2   Neighborhood Longitude  3870 non-null   float64
 3   Venue                   3870 non-null   object 
 4   Venue Latitude          3870 non-null   float64
 5   Venue Longitude         3870 non-null   float64
 6   Venue Category          3870 non-null   object 
 7   CAMIS                   3870 non-null   int64  
 8   BORO                    3870 non-null   object 
 9   BUILDING                3870 non-null   object 
 10  STREET                  3870 non-null   object 
 11  ZIPCODE                 3870 non-null   int64  
 12  PHONE                   3870 non-null   object 
 13  CUISINE DESCRIPTION     3870 non-null   object 
 14  INSPECTION DATE         3870 non-null   object 
 15  ACTION                  3869 non-null   object 
 16  VIOLATION CODE          3837 non-null   object 
 17  VIOLATION DESCRIPTION   3830 non-null   object 
 18  CRITICAL FLAG           3830 non-null   object 
 19  SCORE                   3623 non-null   float64
 20  GRADE                   2200 non-null   object 
 21  GRADE DATE              2200 non-null   object 
 22  RECORD DATE             3870 non-null   object 
 23  INSPECTION TYPE         3869 non-null   object 
 24  Latitude                3870 non-null   float64
 25  Longitude               3870 non-null   float64
 26  Community Board         3870 non-null   float64
 27  Council District        3870 non-null   float64
 28  Census Tract            3870 non-null   float64
 29  BIN                     3835 non-null   float64
 30  BBL                     3870 non-null   float64
 31  NTA                     3870 non-null   object 
dtypes: float64(12), int64(2), object(18)
memory usage: 997.7+ KB
```

I filter this dataset to have my final dataset

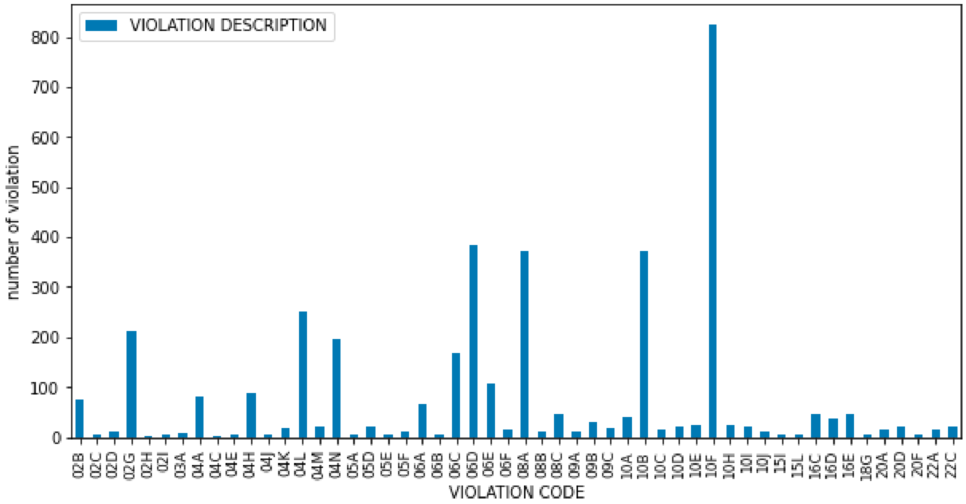


Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	VIOLATION CODE	VIOLATION DESCRIPTION	CRITICAL FLAG	Latitude	Longitude	Venue Category
Bay Ridge	40.625801	-74.030621	XIN	04N	Filth flies or food/refuse/sewage-associated (...	Y	40.625082	-74.030494	Chinese Restaurant
Bay Ridge	40.625801	-74.030621	XIN	06C	Food not protected from potential source of co...	Y	40.625082	-74.030494	Chinese Restaurant
Bay Ridge	40.625801	-74.030621	XIN	09A	Canned food product observed dented and not se...	N	40.625082	-74.030494	Chinese Restaurant
Bay Ridge	40.625801	-74.030621	XIN	10H	Proper sanitization not provided for utensil w...	N	40.625082	-74.030494	Chinese Restaurant
Bay Ridge	40.625801	-74.030621	XIN	04H	Raw, cooked or prepared food is adulterated, c...	Y	40.625082	-74.030494	Chinese Restaurant

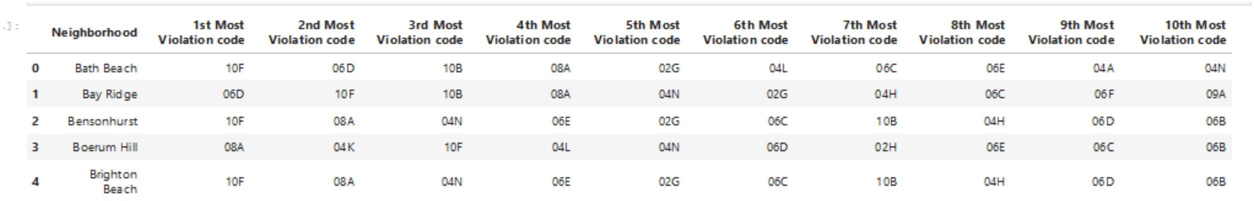
I visualized Brooklyn neighborhoods based on the number of code violations recorded



Let us represent the frequency of the types of violations in the city of Brooklyn



I have listed the 10 most common violations by neighborhood



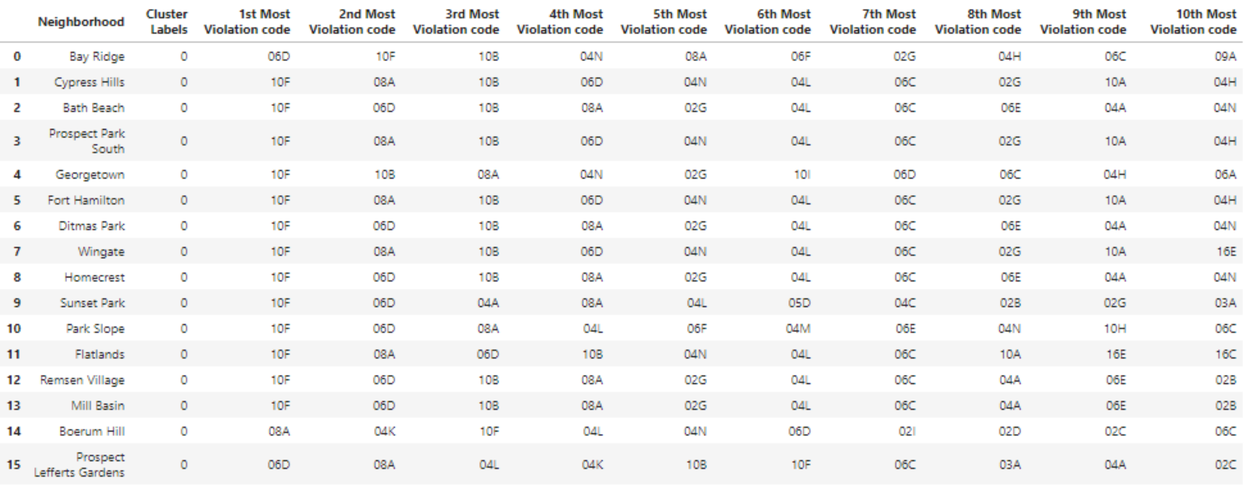
	Neighborhood	1st Most Violation code	2nd Most Violation code	3rd Most Violation code	4th Most Violation code	5th Most Violation code	6th Most Violation code	7th Most Violation code	8th Most Violation code	9th Most Violation code	10th Most Violation code
0	Bath Beach	10F	06D	10B	08A	02G	04L	06C	06E	04A	04N
1	Bay Ridge	06D	10F	10B	08A	04N	02G	04H	06C	06F	09A
2	Bensonhurst	10F	08A	04N	06E	02G	06C	10B	04H	06D	06B
3	Boerum Hill	08A	04K	10F	04L	04N	06D	02H	06E	06C	06B
4	Brighton Beach	10F	08A	04N	06E	02G	06C	10B	04H	06D	06B

We have common violation code categories in neighborhoods. For this reason, I used an unsupervised K-means learning algorithm to group neighborhoods. The K-Means algorithm is one of the most common cluster methods of unsupervised learning.

Results

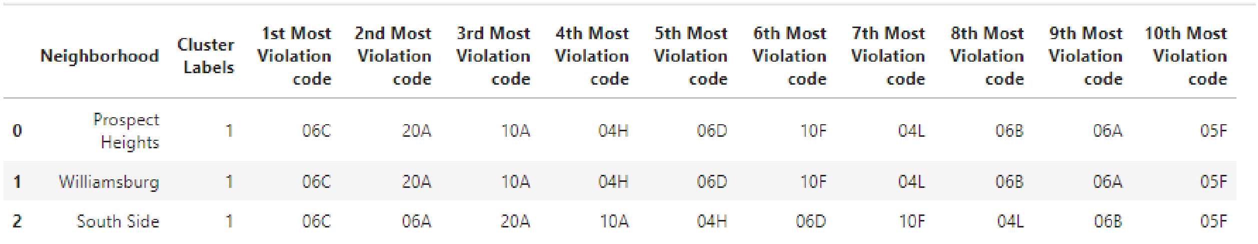
The clusters formed are as follows:

Cluster 1



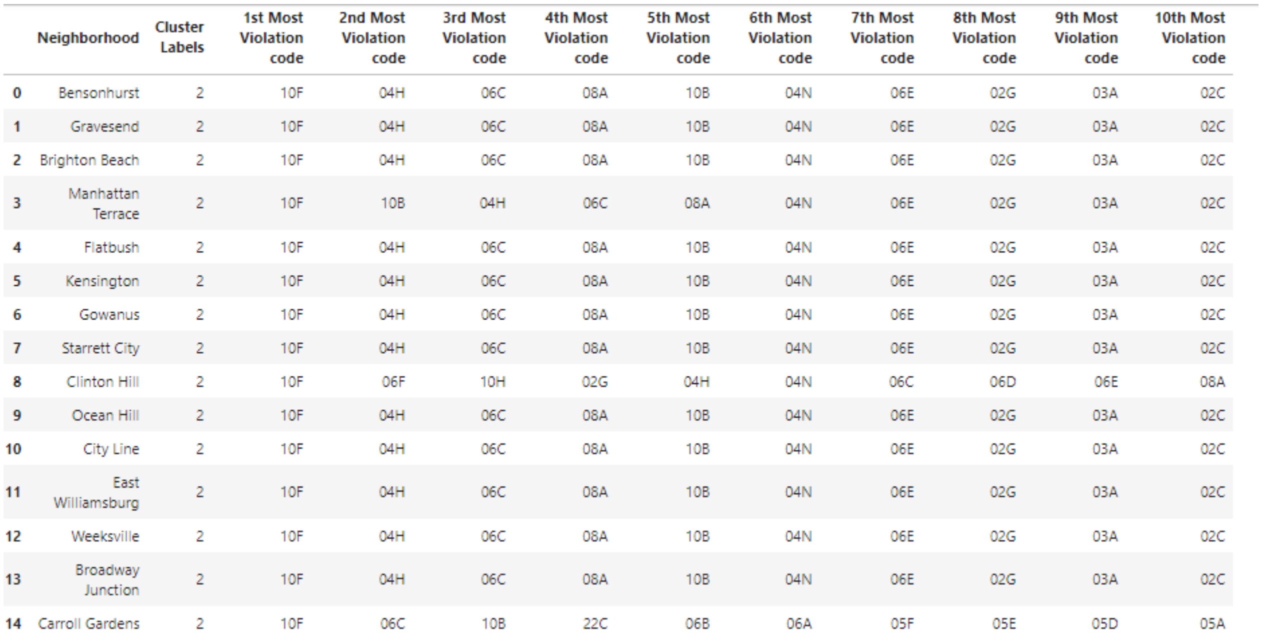
	Neighborhood	Cluster Labels	1st Most Violation code	2nd Most Violation code	3rd Most Violation code	4th Most Violation code	5th Most Violation code	6th Most Violation code	7th Most Violation code	8th Most Violation code	9th Most Violation code	10th Most Violation code
0	Bay Ridge	0	06D	10F	10B	04N	08A	06F	02G	04H	06C	09A
1	Cypress Hills	0	10F	08A	10B	06D	04N	04L	06C	02G	10A	04H
2	Bath Beach	0	10F	06D	10B	08A	02G	04L	06C	06E	04A	04N
3	Prospect Park South	0	10F	08A	10B	06D	04N	04L	06C	02G	10A	04H
4	Georgetown	0	10F	10B	08A	04N	02G	10I	06D	06C	04H	06A
5	Fort Hamilton	0	10F	08A	10B	06D	04N	04L	06C	02G	10A	04H
6	Ditmas Park	0	10F	06D	10B	08A	02G	04L	06C	06E	04A	04N
7	Wingate	0	10F	08A	10B	06D	04N	04L	06C	02G	10A	16E
8	Homecrest	0	10F	06D	10B	08A	02G	04L	06C	06E	04A	04N
9	Sunset Park	0	10F	06D	04A	08A	04L	05D	04C	02B	02G	03A
10	Park Slope	0	10F	06D	08A	04L	06F	04M	06E	04N	10H	06C
11	Flatlands	0	10F	08A	06D	10B	04N	04L	06C	10A	16E	16C
12	Remsen Village	0	10F	06D	10B	08A	02G	04L	06C	04A	06E	02B
13	Mill Basin	0	10F	06D	10B	08A	02G	04L	06C	04A	06E	02B
14	Boerum Hill	0	08A	04K	10F	04L	04N	06D	02I	02D	02C	06C
15	Prospect Lefferts Gardens	0	06D	08A	04L	04K	10B	10F	06C	03A	04A	02C

Cluster 2



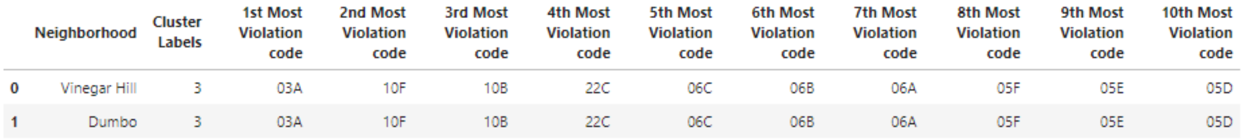
	Neighborhood	Cluster Labels	1st Most Violation code	2nd Most Violation code	3rd Most Violation code	4th Most Violation code	5th Most Violation code	6th Most Violation code	7th Most Violation code	8th Most Violation code	9th Most Violation code	10th Most Violation code
0	Prospect Heights	1	06C	20A	10A	04H	06D	10F	04L	06B	06A	05F
1	Williamsburg	1	06C	20A	10A	04H	06D	10F	04L	06B	06A	05F
2	South Side	1	06C	06A	20A	10A	04H	06D	10F	04L	06B	05F

Cluster 3



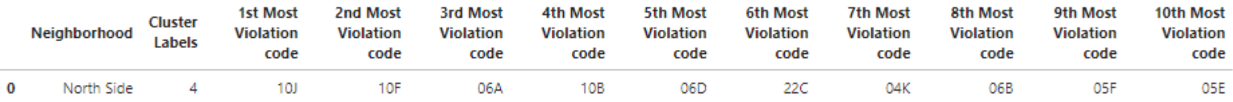
	Neighborhood	Cluster Labels	1st Most Violation code	2nd Most Violation code	3rd Most Violation code	4th Most Violation code	5th Most Violation code	6th Most Violation code	7th Most Violation code	8th Most Violation code	9th Most Violation code	10th Most Violation code
0	Bensonhurst	2	10F	04H	06C	08A	10B	04N	06E	02G	03A	02C
1	Gravesend	2	10F	04H	06C	08A	10B	04N	06E	02G	03A	02C
2	Brighton Beach	2	10F	04H	06C	08A	10B	04N	06E	02G	03A	02C
3	Manhattan Terrace	2	10F	10B	04H	06C	08A	04N	06E	02G	03A	02C
4	Flatbush	2	10F	04H	06C	08A	10B	04N	06E	02G	03A	02C
5	Kensington	2	10F	04H	06C	08A	10B	04N	06E	02G	03A	02C
6	Gowanus	2	10F	04H	06C	08A	10B	04N	06E	02G	03A	02C
7	Starrett City	2	10F	04H	06C	08A	10B	04N	06E	02G	03A	02C
8	Clinton Hill	2	10F	06F	10H	02G	04H	04N	06C	06D	06E	08A
9	Ocean Hill	2	10F	04H	06C	08A	10B	04N	06E	02G	03A	02C
10	City Line	2	10F	04H	06C	08A	10B	04N	06E	02G	03A	02C
11	East Williamsburg	2	10F	04H	06C	08A	10B	04N	06E	02G	03A	02C
12	Weeksville	2	10F	04H	06C	08A	10B	04N	06E	02G	03A	02C
13	Broadway Junction	2	10F	04H	06C	08A	10B	04N	06E	02G	03A	02C
14	Carroll Gardens	2	10F	06C	10B	22C	06B	06A	05F	05E	05D	05A

Cluster 4



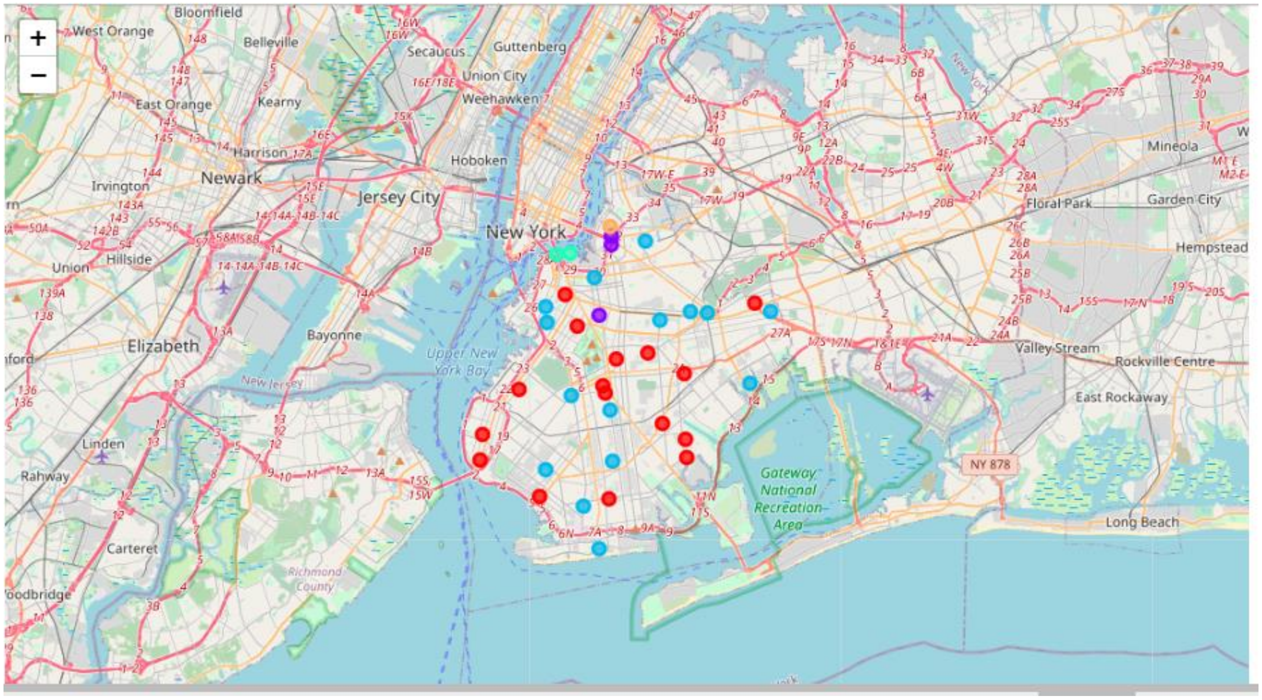
	Neighborhood	Cluster Labels	1st Most Violation code	2nd Most Violation code	3rd Most Violation code	4th Most Violation code	5th Most Violation code	6th Most Violation code	7th Most Violation code	8th Most Violation code	9th Most Violation code	10th Most Violation code
0	Vinegar Hill	3	03A	10F	10B	22C	06C	06B	06A	05F	05E	05D
1	Dumbo	3	03A	10F	10B	22C	06C	06B	06A	05F	05E	05D

Cluster 5



	Neighborhood	Cluster Labels	1st Most Violation code	2nd Most Violation code	3rd Most Violation code	4th Most Violation code	5th Most Violation code	6th Most Violation code	7th Most Violation code	8th Most Violation code	9th Most Violation code	10th Most Violation code
0	North Side	4	10J	10F	06A	10B	06D	22C	04K	06B	05F	05E

You can also see a grouped map of Brooklyn neighborhoods below.



Discussion

Bath Beach is where we find the most critical violation in restaurants. Cluster 1 cuts across neighborhoods similar to Bath Beach. It is therefore recommended that the Department of Health and Mental Hygiene increase inspections in these areas.

We can see Carroll Gardens and North Side are the neighborhoods with the least critical violation.

we also note that the most common violation are as follows: Non-food contact surface improperly constructed. Unacceptable material used. Non-food contact surface or equipment improperly maintained and/or not properly sealed, raised, spaced or movable to allow accessibility for cleaning on all sides, above and underneath the unit.

Conclusion

The department carries out regular inspections. To optimize and specialized it is inspection, it is therefore important to resort to analysis techniques.

Our analysis therefore proposed the possibility of inspection optimization.

In the future we will improve this analysis so that it covers the whole of New York city

Principal Libraries Which are Used to Develop the Project

Pandas: For creating and manipulating dataframes.

Folium: Python visualization library would be used to visualize the neighborhoods cluster distribution of using interactive leaflet map.

Scikit Learn: For importing k-means clustering.

JSON: Library to handle JSON files.

XML: To separate data from presentation and XML stores data in plain text format.

Geocoder: To retrieve Location Data.

Beautiful Soup and Requests: To scrap and library to handle http requests.

Matplotlib: Python Plotting Module.