Applied Data Science Capstone

The Battle of Neighborhoods

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.

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

DATA

DATA 1 :

Source: http://www1.nyc.gov/site/doh/services/restaurant-grades.page

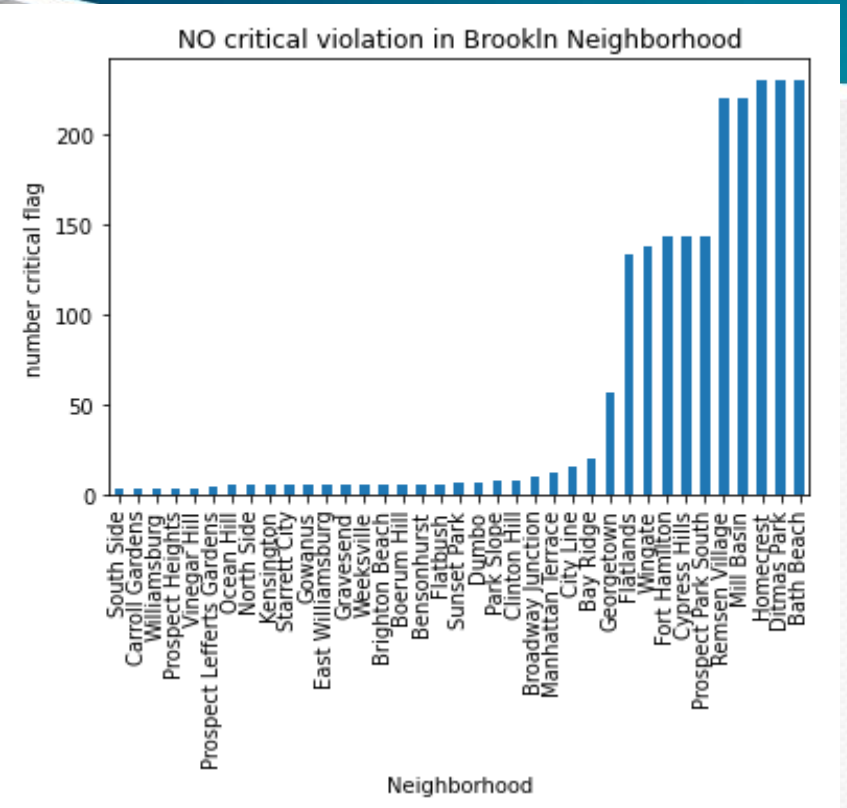
description : contains the findings of registrations by the Department of Health and Mental Hygiene of during inspections in restaurants

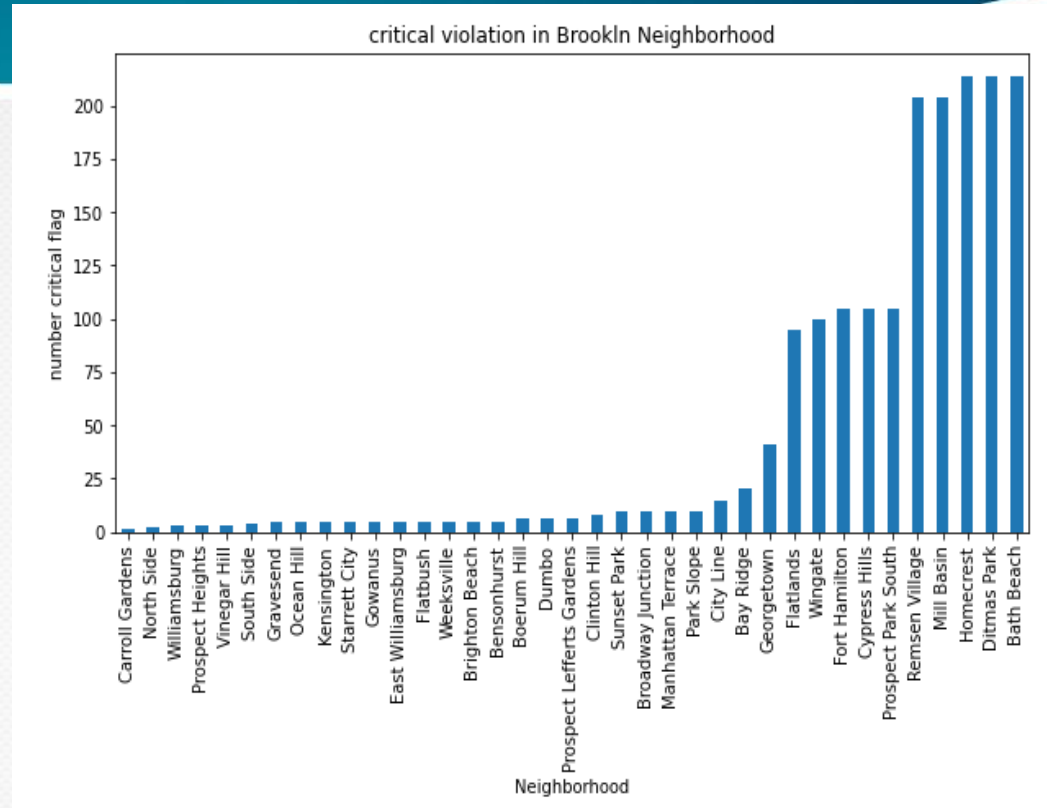
DATA 2 :

Source: https://geo.nyu.edu/catalog/nyu_2451_34572

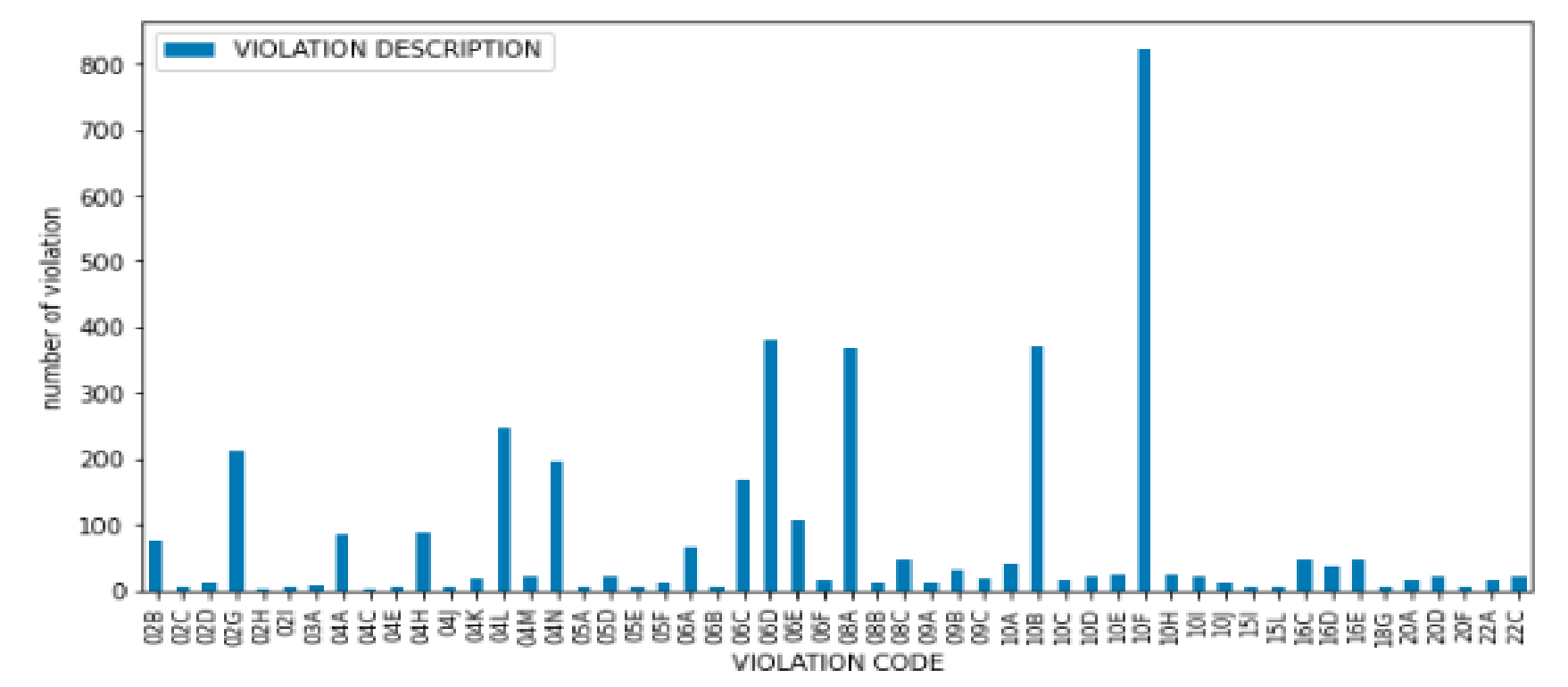
description : contains all the geographic coordinates of all New York neighborhoods

VIOLATION IN BROOKLN





Frequency of violations



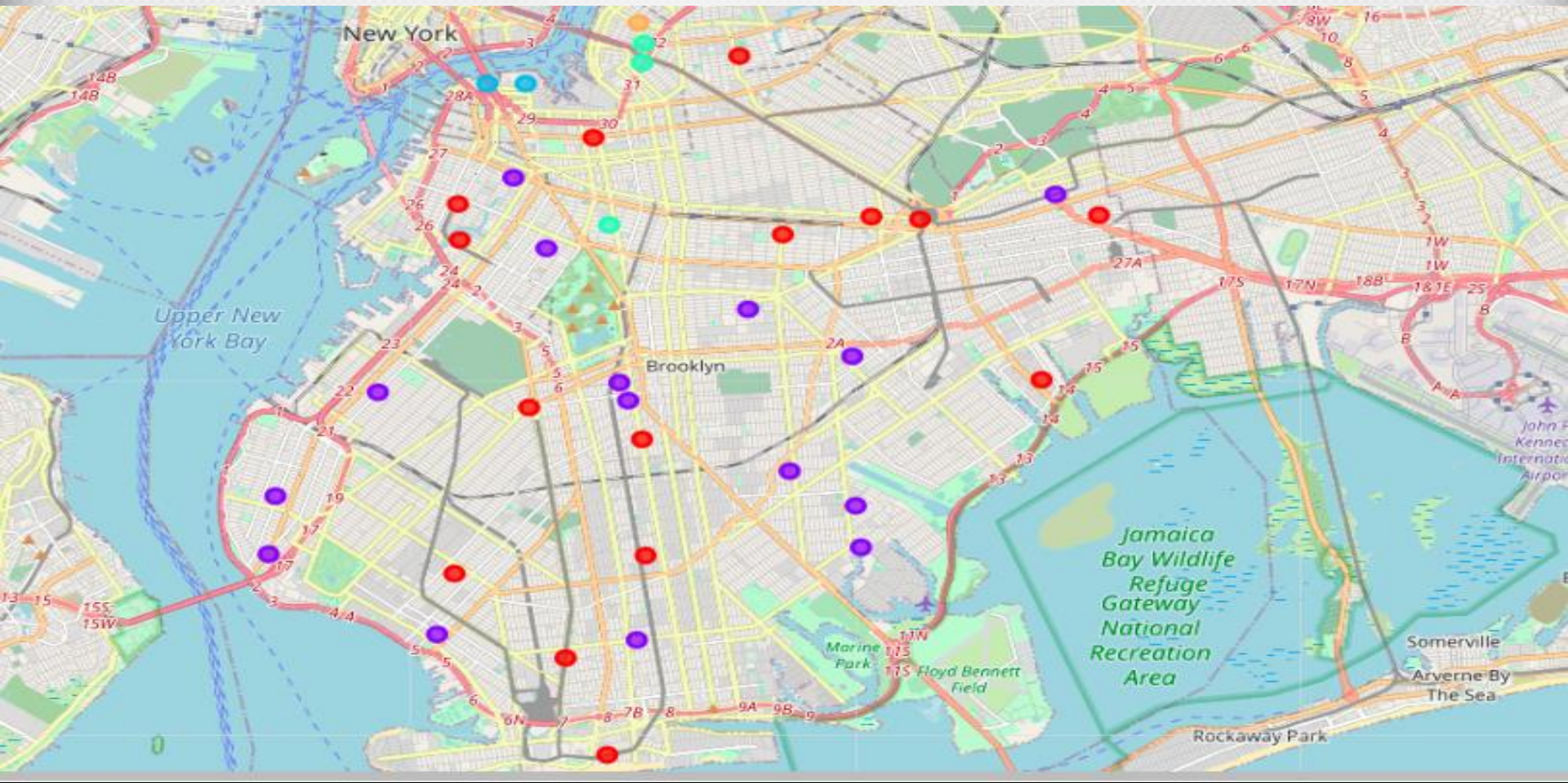
The 10 most common violations by neighborhood

	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	04N	08A	06F	02G	04H	06C	09A
2	Bensonhurst	10F	04H	06C	08A	10B	04N	06E	02G	03A	02C
3	Boerum Hill	08A	04K	10F	04L	04N	06D	02I	02D	02C	06C
4	Brighton Beach	10F	04H	06C	08A	10B	04N	06E	02G	03A	02C
5	Broadway Junction	10F	04H	06C	08A	10B	04N	06E	02G	03A	02C
6	Carroll Gardens	10F	06C	10B	22C	06B	06A	05F	05E	05D	05A
7	City Line	10F	04H	06C	08A	10B	04N	06E	02G	03A	02C
8	Clinton Hill	10F	06F	10H	02G	04H	04N	06C	06D	06E	08A
9	Cypress Hills	10F	08A	10B	06D	04N	04L	06C	02G	10A	04H

Cluster of Neighborhoods

- 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.

Map of Neighborhood Cluster



Presentation of Cluster

- 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

Presentation of Cluster

- 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

This cluster contains the neighborhoods with the least critical violation

Presentation of Cluster

- 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

Presentation of Cluster

- 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

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 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.

Library used

- **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.

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