The Battle of Neighborhoods Finding a good place in Brooklyn, New York City

1. Introduction

- This project explores the city of New York, specifically one of its boroughs: Brooklyn. The neighborhoods in this district are segmented and grouped to see the most common venues.

- Brooklyn is very diverse and very interrelated with Manhattan, perhaps the best known. We will look at the most common venues for neighborhoods in this borough and visualize some indicators in this study.

- This Project aim to create an analysis of features for a people traveling to Brooklyn to search a best neighborhood as a comparative analysis between neighborhoods. The features include average sale price and population.

2. Data Section - part 1/3

2.1. List of neighborhood data of Brooklyn in New York

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

It contains the following data:

- Borough
- Neighborhood
- Latitude
- Longitude

Borough	Neighborhood	Latitude	Longitude
Brooklyn	Bay Ridge	40.625801	-74.030621
Brooklyn	Bensonhurst	40.611009	-73.995180
Brooklyn	Sunset Park	40.645103	-74.010316
Brooklyn	Greenpoint	40.730201	-73.954241
Brooklyn	Gravesend	40.595260	-73.973471
Brooklyn	Brighton Beach	40.576825	-73.965094
Brooklyn	Sheepshead Bay	40.586890	-73.943186
Brooklyn	Manhattan Terrace	40.614433	-73.957438
Brooklyn	Flatbush	40.636326	-73.958401
Brooklyn	Crown Heights	40.670829	-73.943291
Brooklyn	East Flatbush	40.641718	-73.936103
Brooklyn	Kensington	40.642382	-73.980421
Brooklyn	Windsor Terrace	40.656946	-73.980073
Brooklyn	Prospect Heights	40.676822	-73.964859
Brooklyn	Brownsville	40.663950	-73.910235

2. Data Section - part 2/3

2.2 Foursquare database, to be used in order to explore the desired neighborhood data for various venues details and access the JSON files.

https://Foursquare.com

The data retrieved from Foursquare contained information of venues within a specified distance of the longitude and latitude of the neighborhoods. The information obtained per venue as follows:

1. Neighborhood
2. Neighborhood Latitude
3. Neighborhood Longitude
4. Name of the venue e.g. the name of a store or restaurant
5. Venue Latitude
6. Venue Longitude
7. Venue Category

2. Data Section - part 3/3

2.3. CSV file with Average Sales Price and Population indicators for New York City.

This file has been generated with the following data sets downloaded from the NYC Open Data website (opendata.cityofnewyork.us):

DOF: Summary of Neighborhood Sales by Neighborhood Citywide by Borough

https://data.cityofnewyork.us/City-Government/DOF-Summary-of-Neighborhood-Sales-by-Neighborhood-/5ebm-myj7

New York City Population By Neighborhood Tabulation Areas

https://data.cityofnewyork.us/City-Government/New-York-City-Population-By-Neighborhood-Tabulatio/swpk-hqdp

It contains the following data:

- Neighborhood
- Average Sales Price indicator
- Population indicator

3. Methodology Section

3.1 Clustering Approach:

I decided to explore neighborhoods, segment them, and group them into clusters to find similar neighborhoods in a big city like New York (Brooklyn). To be able to do that, we need to cluster data which is a form of unsupervised machine learning: k-means clustering algorithm.

3.2 Using K-Means Clustering Approach

	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	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	Brooklyn	Bay Ridge	40.625801	-74.030621	0	Italian Restaurant	Spa	Pizza Place	American Restaurant	Bar	Greek Restaurant	Bagel Shop	Hookah Bar	Thai Restaurant	Ice Cream Shop
1	Brooklyn	Bensonhurst	40.611009	-73.995180	4	Chinese Restaurant	Italian Restaurant	Donut Shop	Ice Cream Shop	Sushi Restaurant	Park	Pizza Place	Shabu-Shabu Restaurant	Sporting Goods Shop	Smoke Shop
2	Brooklyn	Sunset Park	40.645103	-74.010316	4	Mobile Phone Shop	Latin American Restaurant	Bank	Pizza Place	Bakery	Mexican Restaurant	Fried Chicken Joint	Gym	Deli / Bodega	Ice Cream Shop
3	Brooklyn	Greenpoint	40.730201	-73.954241	0	Bar	Pizza Place	Coffee Shop	Cocktail Bar	Yoga Studio	French Restaurant	Record Shop	Mexican Restaurant	Deli / Bodega	Bakery
4	Brooklyn	Gravesend	40.595260	-73.973471	4	Lounge	Pizza Place	Bakery	Chinese Restaurant	Gym	Italian Restaurant	Cosmetics Shop	Farmers Market	Men's Store	Baseball Field

3. Methodology Section

3.3 Most Common venues near Neighborhood

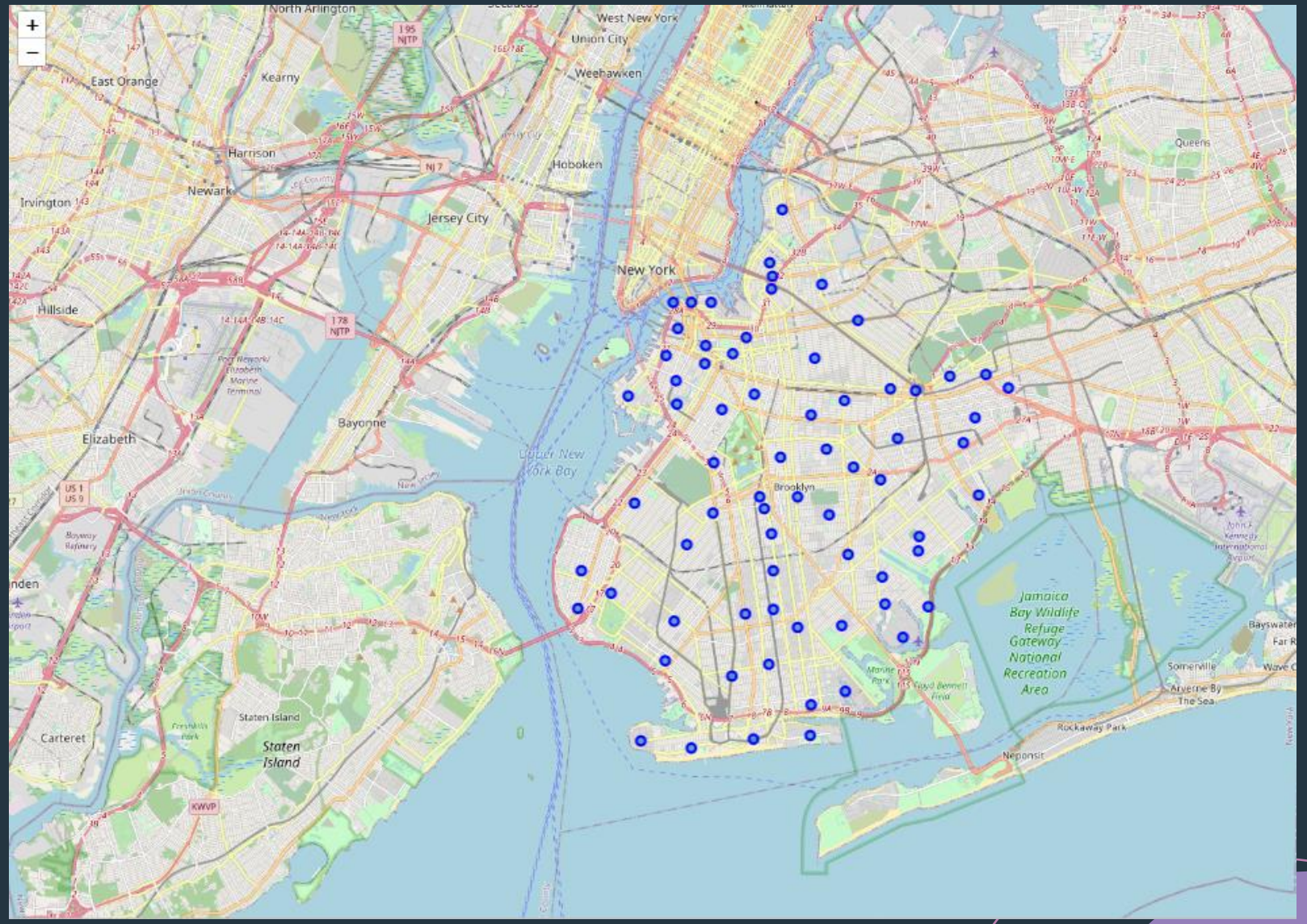
```
num_top_venues = 10

indicators = ['st', 'nd', 'rd']

# create columns according to number of top venues
columns = ['Neighborhood']
for ind in np.arange(num_top_venues):
    try:
        columns.append('{}{} Most Common Venue'.format(ind+1, indicators[ind]))
    except:
        columns.append('{}th Most Common Venue'.format(ind+1))

# create a new dataframe
neighborhoods_venues_sorted = pd.DataFrame(columns=columns)
neighborhoods_venues_sorted['Neighborhood'] = brooklyn_grouped['Neighborhood']

for ind in np.arange(brooklyn_grouped.shape[0]):
    neighborhoods_venues_sorted.iloc[ind, 1:] = return_most_common_venues(brooklyn_grouped.iloc[ind, :], num_top_venues)

neighborhoods_venues_sorted.head()
```

	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	Bath Beach	Bubble Tea Shop	Italian Restaurant	Cantonese Restaurant	Pizza Place	Gas Station	Pharmacy	Dessert Shop	Fast Food Restaurant	Chinese Restaurant	Donut Shop
1	Bay Ridge	Italian Restaurant	Spa	Pizza Place	American Restaurant	Bar	Greek Restaurant	Bagel Shop	Hookah Bar	Thai Restaurant	Ice Cream Shop
2	Bedford Stuyvesant	Coffee Shop	Pizza Place	Café	Bar	Deli / Bodega	Tiki Bar	Park	Cocktail Bar	New American Restaurant	Bus Stop
3	Bensonhurst	Chinese Restaurant	Italian Restaurant	Donut Shop	Ice Cream Shop	Sushi Restaurant	Park	Pizza Place	Shabu-Shabu Restaurant	Sporting Goods Shop	Smoke Shop
4	Bergen Beach	Harbor / Marina	Hockey Field	Baseball Field	Athletics & Sports	Playground	Food Court	Food & Drink Shop	Food	Flower Shop	Food Stand

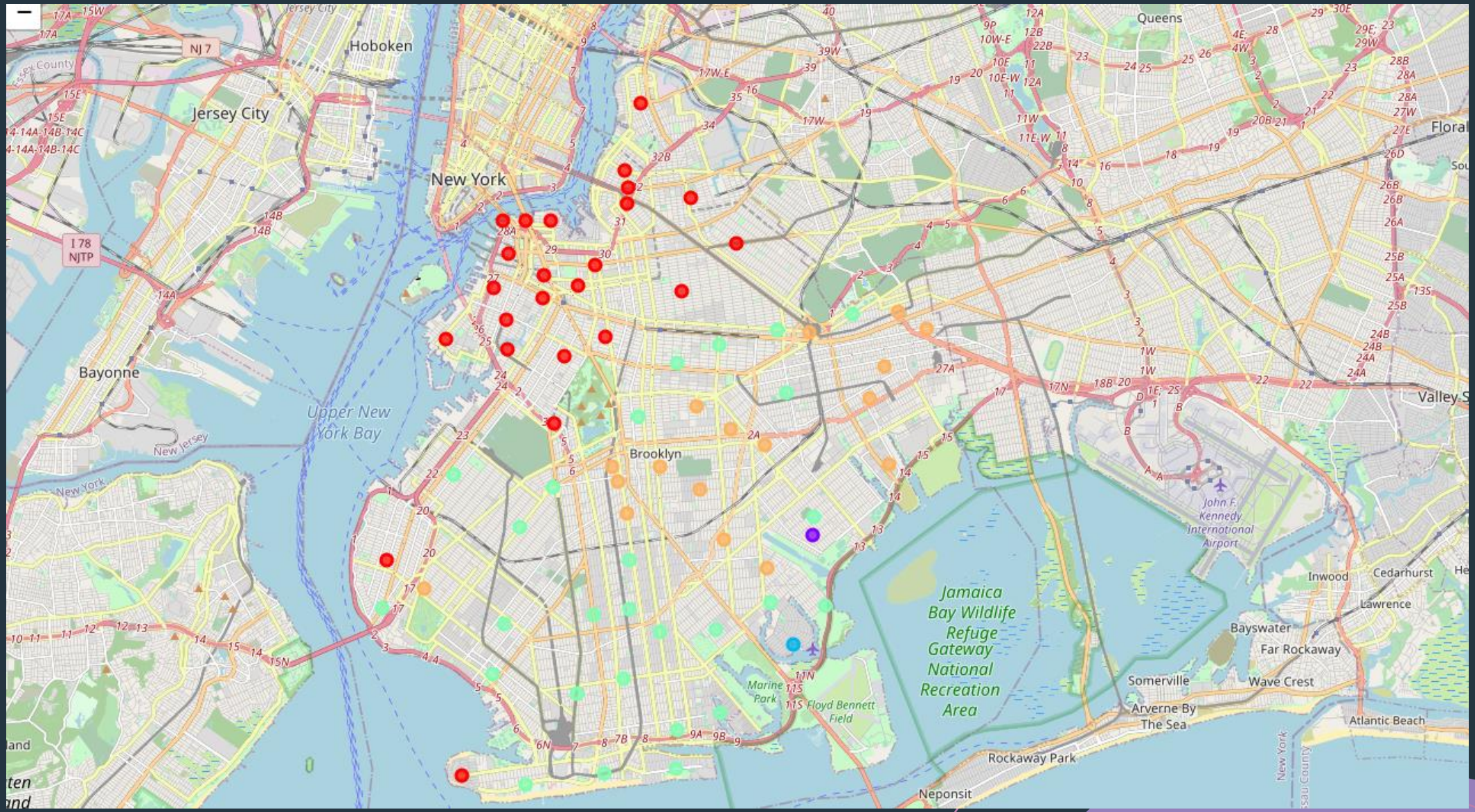
3.4 Workflow

Using credentials of Foursquare API features of near-by places of the neighborhoods would be mined. Due to http request limitations the number of places per neighborhood parameter would reasonably be set to 100 and the radius parameter would be set to 500.

4. Results Section – Map of Brooklyn



4. Results Section – Map of Clusters

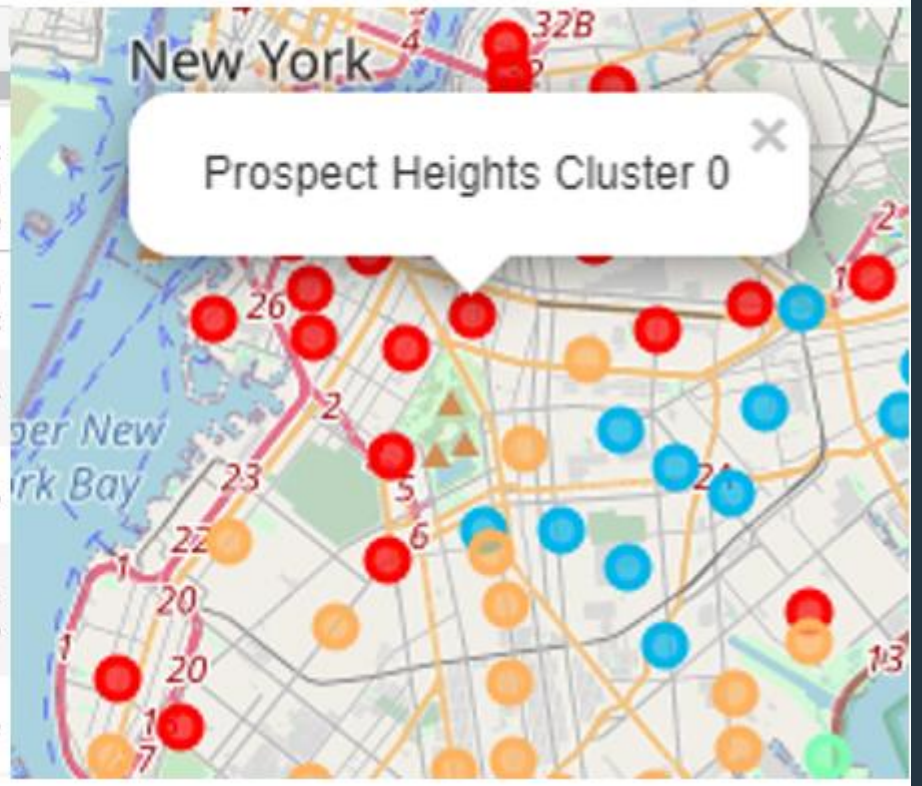


4. Results Section – Cluster 0

Cluster 0

```
brooklyn_merged.loc[brooklyn_merged['Cluster Labels'] == 0,
```

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue
0	Bay Ridge	Italian Restaurant	Spa	Pizza Place	American Restaurant
3	Greenpoint	Bar	Pizza Place	Coffee Shop	Cocktail Bar
6	Sheepshead Bay	Turkish Restaurant	Dessert Shop	Sandwich Place	Yoga Studio
11	Kensington	Thai Restaurant	Grocery Store	Pizza Place	Ice Cream Shop
12	Windsor Terrace	Diner	Park	Plaza	Grocery Store



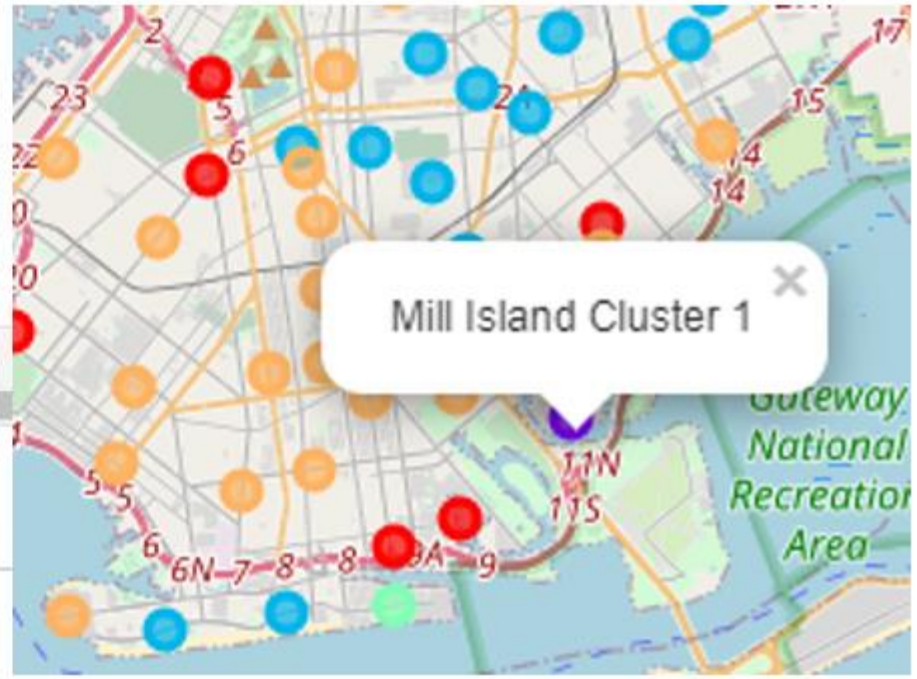
In cluster 0 in red, the most common venues are restaurants and shops. Perhaps because of its proximity to Manhattan. It is the one with the most neighborhoods, 31 in total.

4. Results Section – Cluster 1

Cluster 1

```
brooklyn_merged.loc[brooklyn_merged['Cluster Labels'] == 1,
```

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue
30	Mill Island	Pool	Women's Store	Flower Shop	Fast Food Restaurant



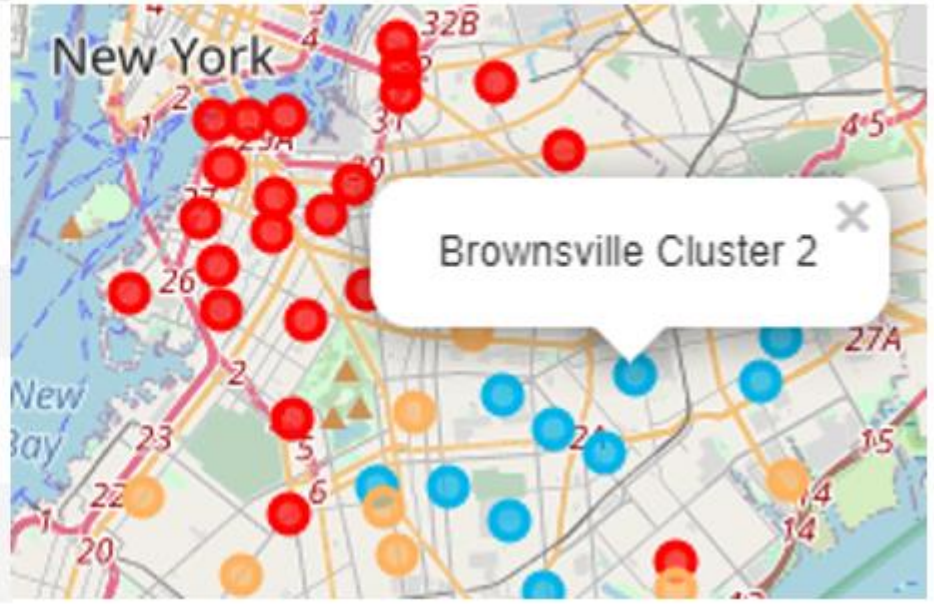
Cluster 1 in purple is the one with the fewest neighborhoods. The most common venues are stores.

4. Results Section – Cluster 2

Cluster 2

```
brooklyn_merged.loc[brooklyn_merged['Cluster Labels'] == 2,
```

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue
5	Brighton Beach	Restaurant	Russian Restaurant	Eastern European Restaurant	Gourmet Shop
10	East Flatbush	Department Store	Pharmacy	Moving Target	Print Shop
14	Brownsville	Moving Target	Fried Chicken Joint	Restaurant	Farmers Market
25	Cypress Hills	Latin American Restaurant	Fried Chicken Joint	Ice Cream Shop	Donut Shop



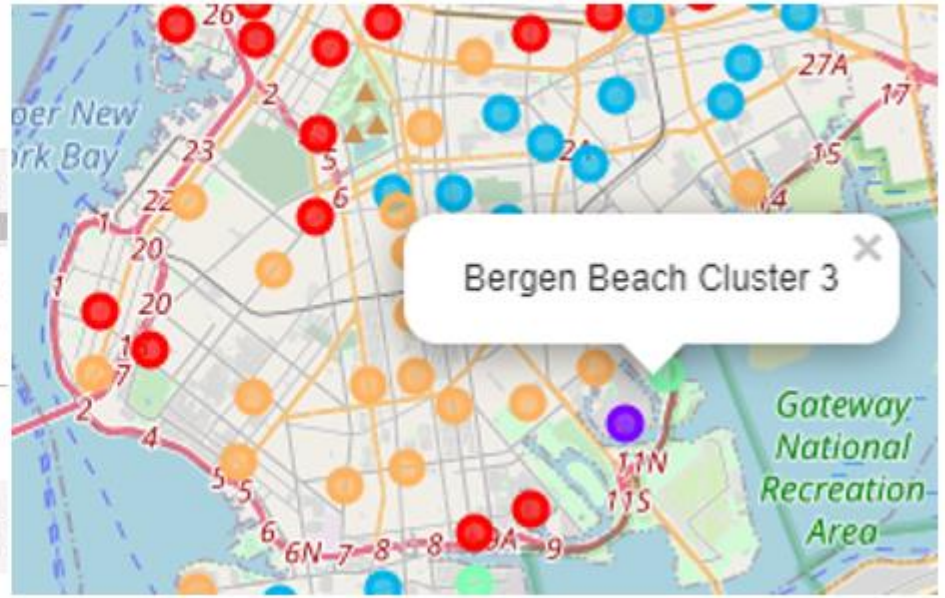
Cluster 2 in blue with 15 neighborhoods, the third with the most neighborhoods. It has more varied venues with shops, restaurants, supermarkets, gyms.

4. Results Section – Cluster 3

Cluster 3

```
brooklyn_merged.loc[brooklyn_merged['Cluster Labels'] == 3,
```

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue
31	Manhattan Beach	Café	Ice Cream Shop	Sandwich Place	Bus Stop
45	Bergen Beach	Harbor / Marina	Hockey Field	Baseball Field	Athletics & Sports



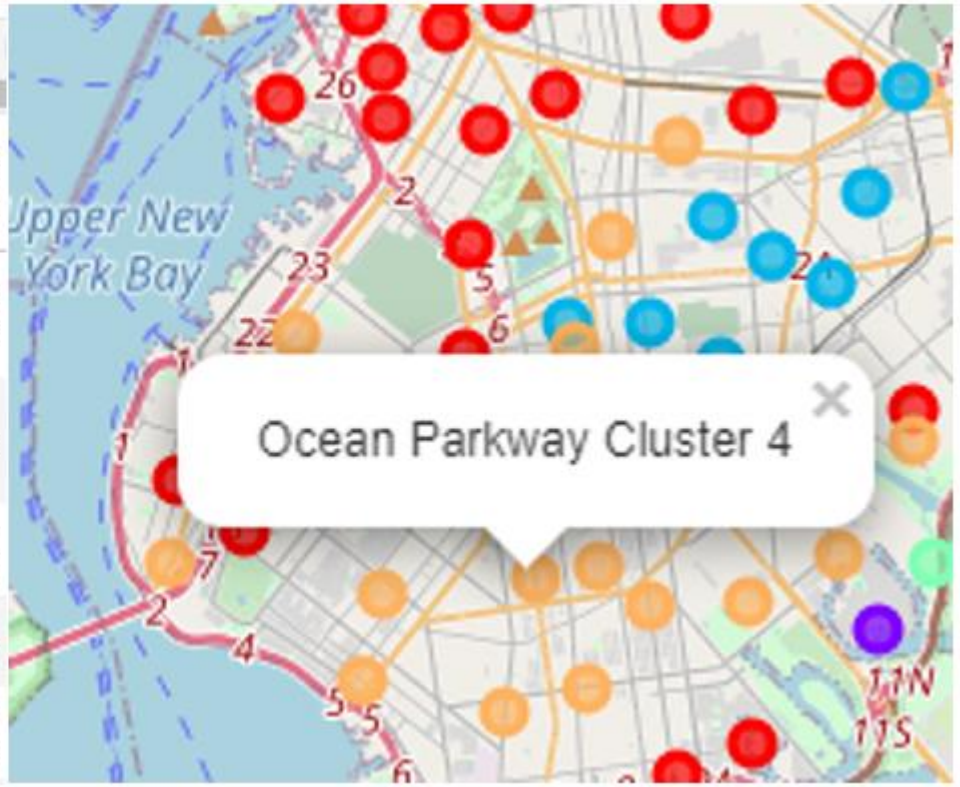
Cluster 3 in green with 2 neighborhoods. Maybe with places more focused on sports activities.

4. Results Section – Cluster 4

Cluster 4

```
brooklyn_merged.loc[brooklyn_merged['Cluster Labels'] == 4,
```

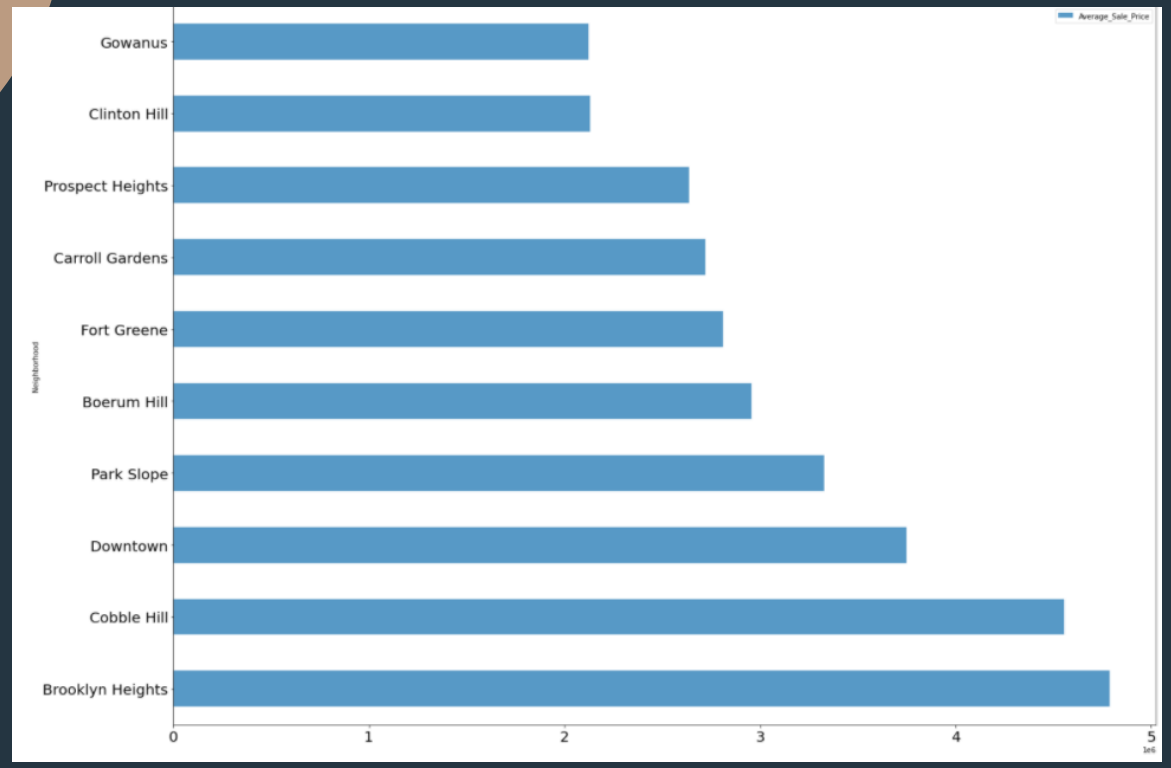
	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue
1	Bensonhurst	Chinese Restaurant	Italian Restaurant	Donut Shop	Ice Cream Shop
2	Sunset Park	Mobile Phone Shop	Latin American Restaurant	Bank	Pizza Place
4	Gravesend	Lounge	Pizza Place	Bakery	Chinese Restaurant
7	Manhattan Terrace	Ice Cream Shop	Pizza Place	Chinese Restaurant	Donut Shop
8	Flatbush	Mexican Restaurant	Deli / Bodega	Coffee Shop	Caribbean Restaurant



Cluster 4 in beige with 21 neighborhoods, the second with the most neighborhoods. Perhaps with more varied places, restaurants, shops, banks, etc.

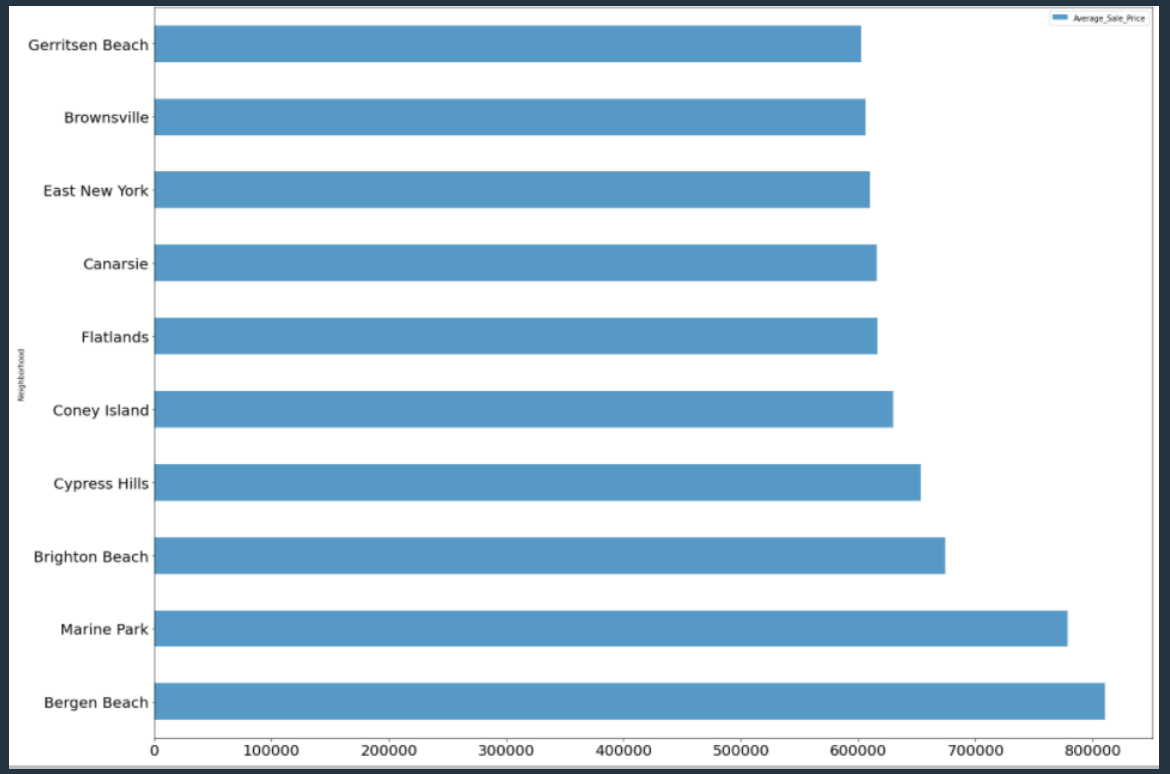
4. Results Section – Average Sales price indicator

The 10 neighborhoods with the highest average sales price



The top 10 neighborhoods are in cluster 0. We can assume that cluster 0 has the neighborhoods with the highest average sales price.

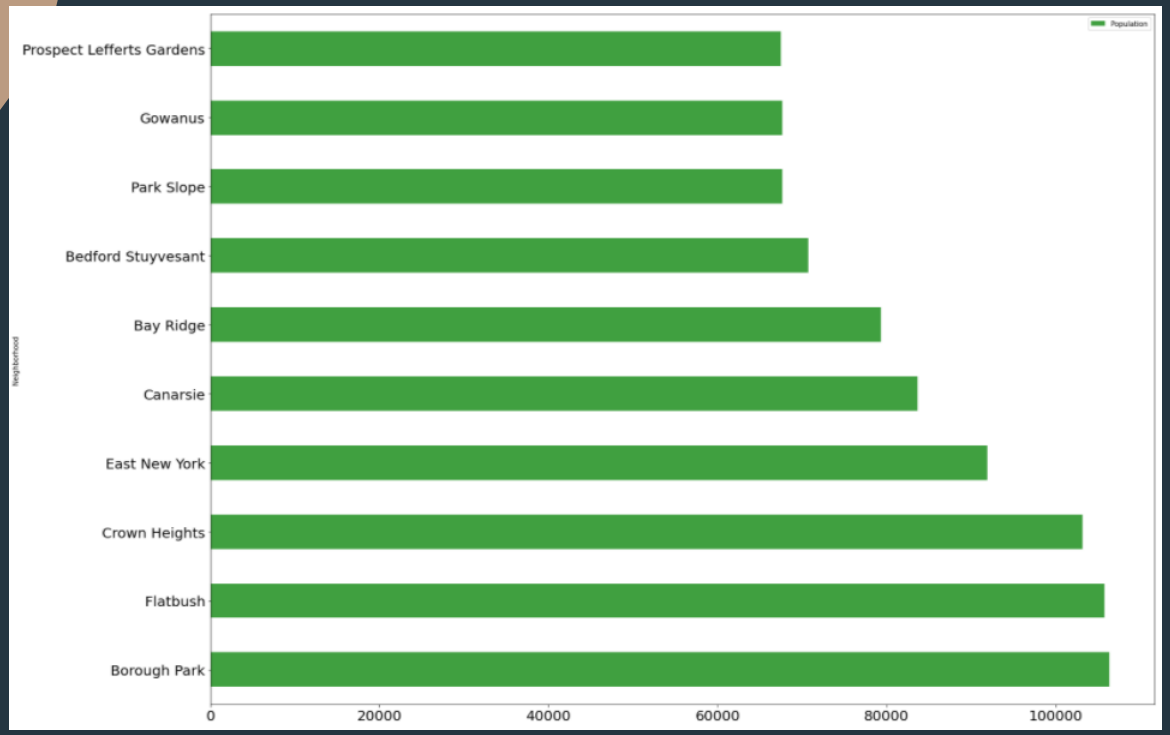
The 10 neighborhoods with the lowest average sales price



The bottom 10 neighborhoods, 60% are in cluster 2. We can assume that cluster 2 has the most neighborhoods with the lowest average sales price.

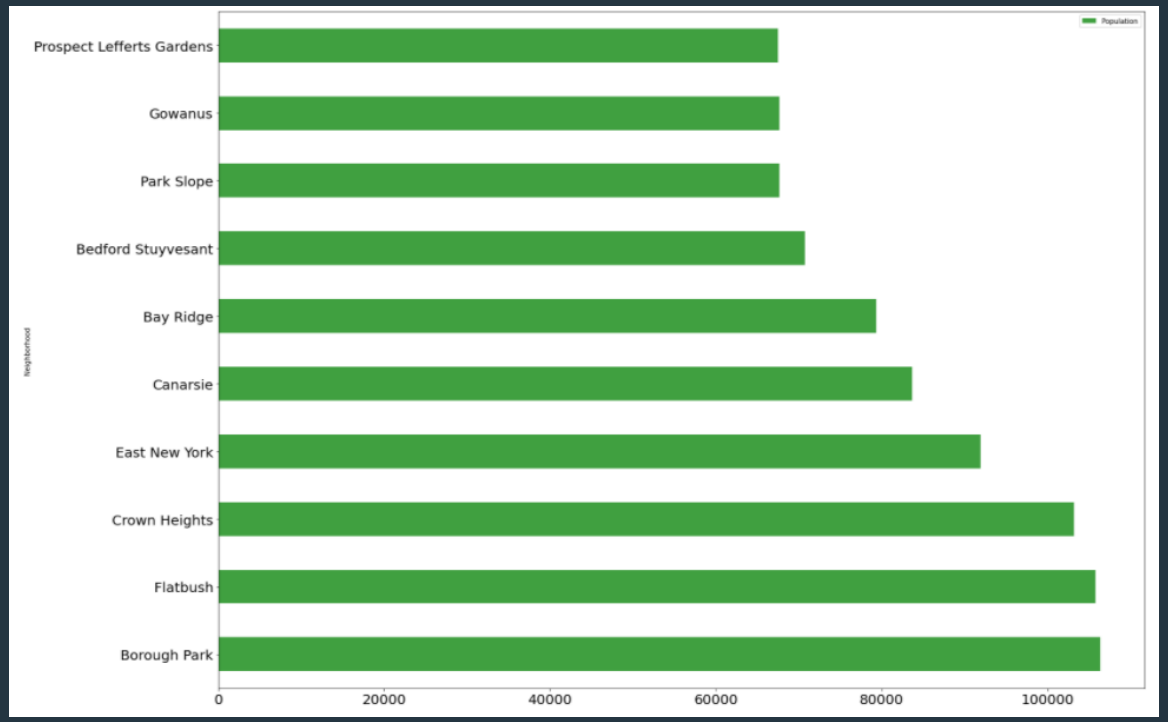
4. Results Section – Population indicator

The 10 neighborhoods with the largest population



The top 10 most populated neighborhoods are in clusters 0 and 4.

The 10 neighborhoods with the smallest population



Similarly, the last 10 neighborhoods with the lowest population are in clusters 0 and 4

5. Discussion Section

5.1 Problem to be solved

The main purpose of this project is to suggest a better neighborhood in Brooklyn (one of the boroughs of New York City) for the person who moves or wants to reside there. Issues such as social presence in society in terms of like-minded people, communications, house prices, shops, schools, sports centers and other daily needs.

5.2 Using the average sales price and population indicators. We have the views of:

- Top 10 Neighborhoods with Highest Average Sales Price
- Top 10 Neighborhoods with Lowest Average Sales Price
- Top 10 neighborhoods with the largest population
- Top 10 neighborhoods with the smallest population

5. Discussion Section

5.3 Conclusions:

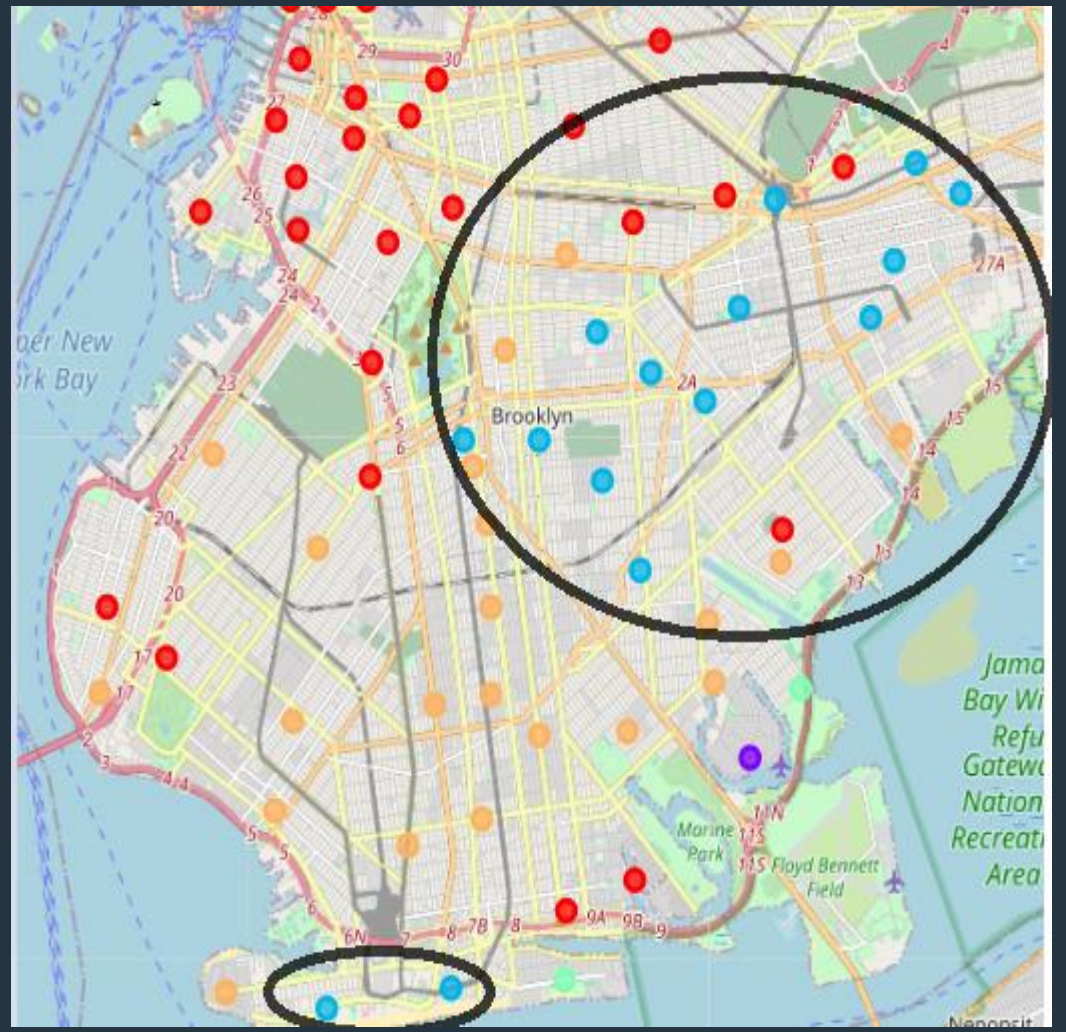
Clusters 2 and 4 contain the most varied neighborhoods with shops, restaurants, schools, gyms, etc.

The average of the highest sale price is in cluster 0. While the cheapest are in cluster 2 and also in 0.

Regarding Population, clusters 0 and 4 have quite a few neighborhoods with the largest and smallest population.

5.4 Recommendation

If you are looking for a neighborhood with the most varied sites, that the sale price averages are inexpensive, and that have a population that is neither too high nor too low, the recommendation would be neighborhoods in cluster 2 (neighborhoods that are in the black circles).



6. Conclusion Section

In this project I have used the k-means cluster algorithm, I have separated the neighborhood into 5 different clusters and for 70 different latitude and longitude of the dataset, which have very similar neighborhoods around them. Using the graphs above, the results presented to a particular neighborhood are based on average selling prices and population indicators.

I think this course with all the topics covered is worthy of appreciation and I feel rewarded for the effort made. This project presents a practical case to solve a real situation that has personal and financial impact using Data Science tools. Folium mapping is a very powerful technique to consolidate information and improve analysis and decision with confidence.

Future work
his project can continue by adding new indicators for the different neighborhoods in Brooklyn that allow finding the best house to live in, taking into consideration other criteria for current needs.

Appendix

Libraries Which are Used to Develope 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.
- Geocoder: To retrieve Location Data.
- Matplotlib: Python Plotting Module.