Selecting the best neighborhood for life in New York

March 28, 2020

1. Introduction

1.1. Background

People desire to live in a convenient and safe place. However, it is not so easy to understand, which neighborhood has best match with a person's criteria: each person has a different requirements and there are a lot of data to analyze. Especially, choice is difficult, when people change city or even country of living.

1.2. Problem

The report target is to provide recommendation for best place for life in New York for a person, who doesn't live currently in New York. The following criteria are important for a customer:

- It should be a safe place
- There must be markets and parks
- Bakeries, pharmacies and gyms are preferable

1.3. Audience

There is a single person, who desires to change his place of living to New York and doesn't know which neighborhood will be most convenient for him.

2. Data

2.1. Data sources

New York crimes data will be obtained from official source:
https://data.cityofnewyork.us/api/views/qgea-i56i/rows.csv.

New York neighborhoods geo information will be downloaded from
https://cocl.us/new_york_dataset.

Forsquare will be used to obtain information about venues in each neighborhood.

2.2. Data cleaning

Crime data has the following look:

CMPLNT_TO_TM	ADDR_PCT_CD	RPT_DT	KY_CD	OFNS_DESC	PD_CD	PD_DESC	CRM_ATPT_CPTD_CD	LAW_CAT_CD	BORO_NM
NaN	73.0	04/10/2008	341	PETIT LARCENY	321.0	LARCENY,PETIT FROM AUTO	COMPLETED	MISDEMEANOR	BROOKLYN
NaN	28.0	06/03/2007	236	DANGEROUS WEAPONS	782.0	WEAPONS, POSSESSION, ETC	COMPLETED	MISDEMEANOR	MANHATTAN
20:50:00	102.0	02/16/2010	105	ROBBERY	375.0	ROBBERY,PHARMACY	COMPLETED	FELONY	QUEENS

The following data is available:

```
CMPLNT_NUM              int64
CMPLNT_FR_DT           object
CMPLNT_FR_TM           object
CMPLNT_TO_DT           object
CMPLNT_TO_TM           object
ADDR_PCT_CD           float64
RPT_DT                 object
KY_CD                   int64
OFNS_DESC              object
PD_CD                 float64
PD_DESC                object
CRM_ATPT_CPTD_CD       object
LAW_CAT_CD             object
BORO_NM                object
LOC_OF_OCCUR_DESC      object
PREM_TYP_DESC          object
JURIS_DESC             object
JURISDICTION_CODE     float64
PARKS_NM               object
HADEVELOPT             object
HOUSING_PSA            object
X_COORD_CD            float64
Y_COORD_CD            float64
SUSP_AGE_GROUP         object
SUSP_RACE              object
SUSP_SEX               object
TRANSIT_DISTRICT      float64
Latitude              float64
Longitude             float64
Lat_Lon                object
PATROL_BORO            object
STATION_NAME           object
VIC_AGE_GROUP          object
VIC_RACE               object
VIC_SEX                object
```

Customer didn't mention any special requirements for safety, so, the total number of crimes per borough (BORO_NM) will be as safety criteria. Other information will be not used.

New York geo data consists of following features:

```
{'type': 'Feature',
 'id': 'nyu_2451_34572.1',
 'geometry': {'type': 'Point',
  'coordinates': [-73.84720052054902, 40.89470517661]},
```

```
'geometry_name': 'geom',
'properties': {'name': 'Wakefield',
 'stacked': 1,
 'annoline1': 'Wakefield',
 'annoline2': None,
 'annoline3': None,
 'annoangle': 0.0,
 'borough': 'Bronx',
 'bbox': [-73.84720052054902,
  40.89470517661,
  -73.84720052054902,
  40.89470517661]}}
```


The following data will be used:

- Borough name
- Neighborhood name (properties.name)
- Coordinates

Forsquare data will be used to collect information of venues in each neighborhood. Following data will be used:

- Venue name
- Venue coordinates
- Venue category name

3. Methodology

3.1. Criminal situation analysis

To achieve crime rate information, let's count crimes accidents and group data per borough:

BORO_NM	crimes
BRONX	1484373
BROOKLYN	2035004
MANHATTAN	1645015
QUEENS	1351400
STATEN ISLAND	321394

Figure 1 Crimes count per borough

Let's visualize obtained data for better understanding:

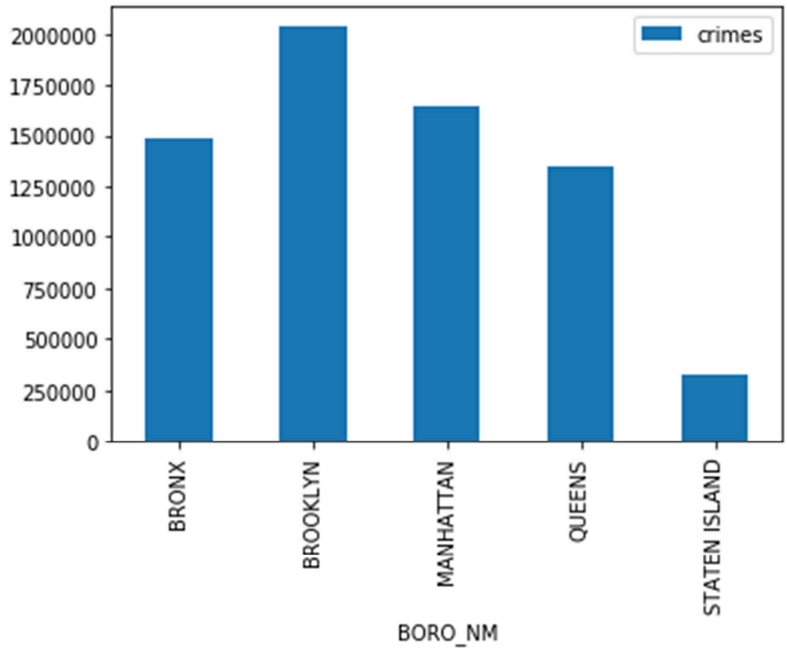


Figure 2 Crimes count per borough

As we can see, Staten Island has significantly smaller count of crimes. However, need to analyze existing venues for each neighborhood, so, let's assume that each neighborhood has the same crimes distribution. Let's calculate neighborhoods:

Borough	count
BRONX	52
BROOKLYN	70
MANHATTAN	40
QUEENS	81
STATEN ISLAND	63

Figure 3 Neighborhoods count per borough

Then we can divide each borough's total crimes count per neighborhoods count and obtain approximate crimes count per neighborhood.

3.2. Venues analysis

Firstly, it is good to understand our geo data. Let's visualize neighborhoods geo data:

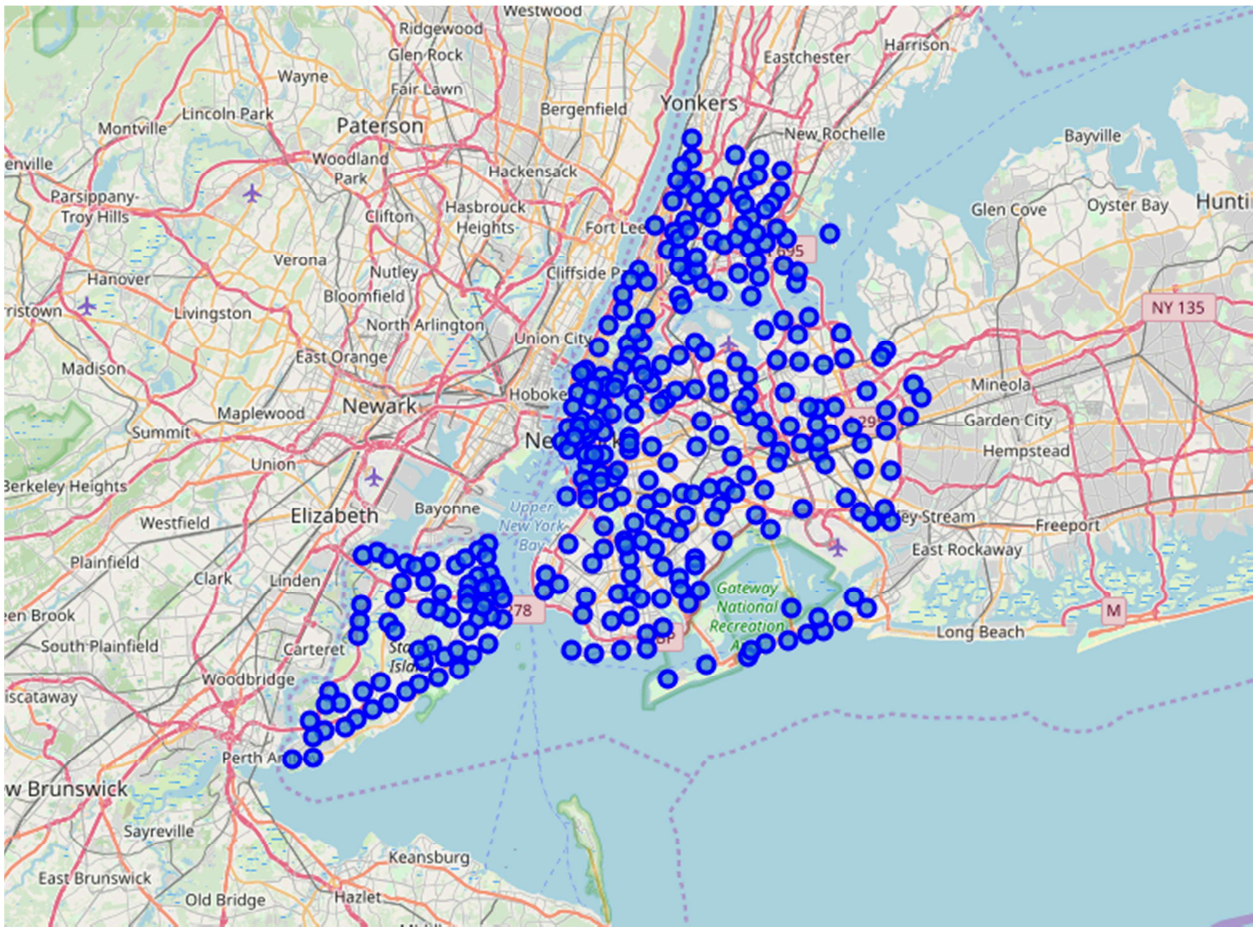


Figure 4 New York neighborhoods map

For each of neighborhoods we obtain venues list from Forsquare API with limit 100 per neighborhood. We create one-hot data frame for each venue category to have a possibility to analyze venues frequency by category in each neighborhood.

There are 434 unique venue categories. However, client is interested only in 5 of them. So let's rename some categories for more convenient further analysis:

- Athletics & Sports to Gym
- Gym / Fitness Center to Gym
- Gymnastics Gym to Gym
- Drugstore to Pharmacy
- Garden to Park
- Garden Center to Gym

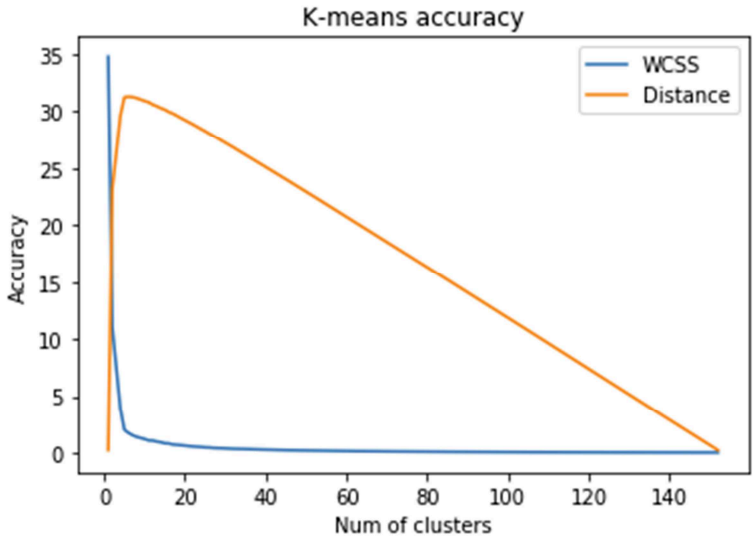
Then we filter Forsquare data leaving only following venue categories for one-hot data-frame:

- Bakery
- Pharmacy

- Park
- Gym
- Market

Next, let's group rows by neighborhood and by taking the mean of the frequency of occurrence of each category. Also, we add normalized crime count for further neighborhoods clustering. As a clustering algorithm K-means was chosen. The simplest algorithm was chosen. Then other algorithms may be tried if K-means will not do its job well.

We don't know optimal clusters amount so let's try "Elbow method" to find an optimum.



So, optimal cluster number is 6. Now we can label with cluster number each neighborhood and analyze each cluster. Firstly, let’s visualize clustered data:

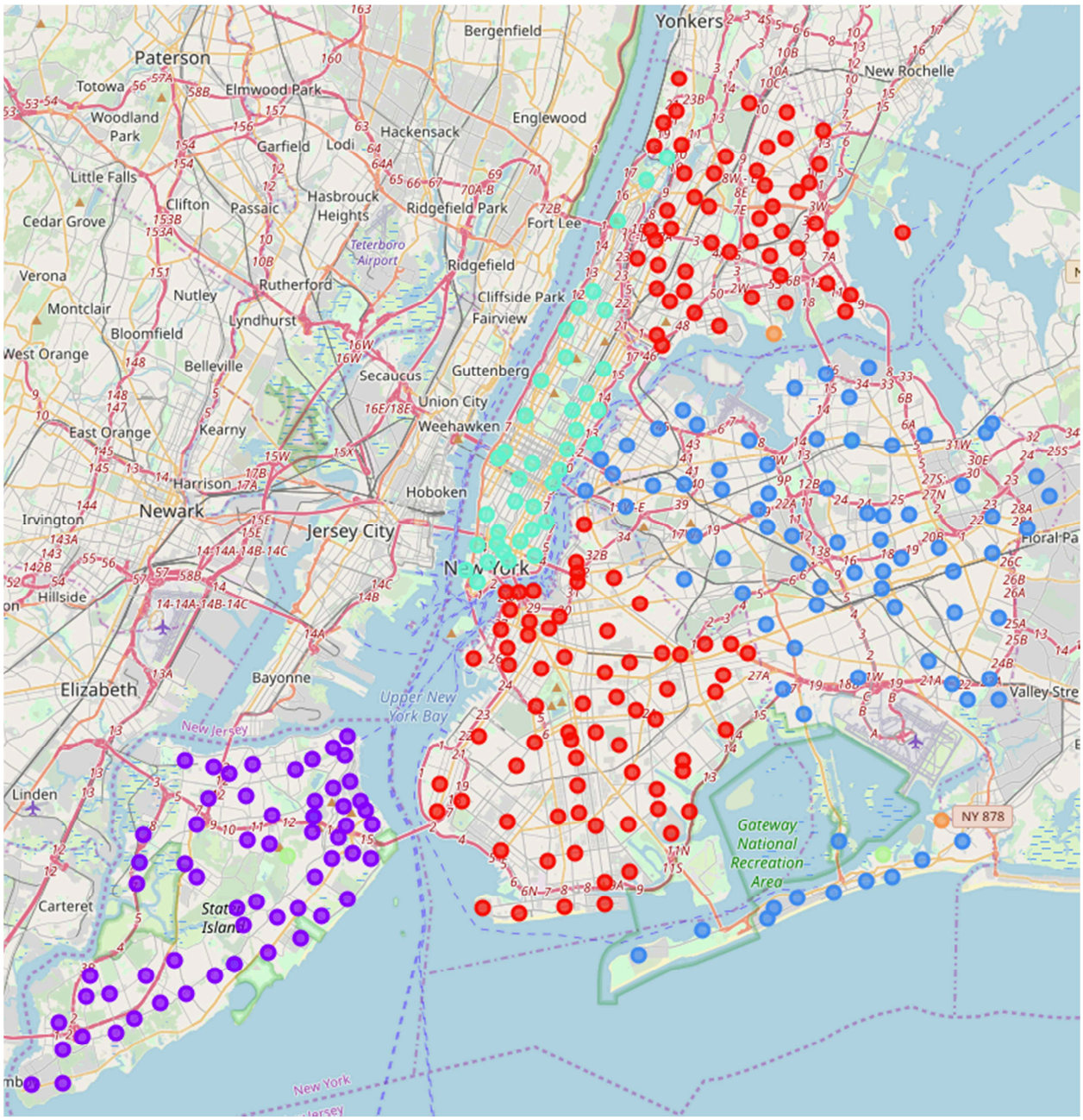


Figure 5 Clustered neighborhoods

It is obvious that data was clustered by boroughs because of crime rate. That means that each borough have more or less same situation with venues, required by customer and main borough choice criteria will be safety. So, recommended neighborhoods must be in Staten Island.

4. Results

Let's look on venues per each cluster:

```
Cluster No. 0
Bakery          0.016969
Pharmacy        0.032072
Park            0.018265
```

```
Gym           0.019960
Market        0.001696
crime_rate    0.659247

Cluster No. 1
Bakery        0.011877
Pharmacy      0.018296
Park          0.018673
Gym           0.020603
Market        0.002328
crime_rate    0.000000

Cluster No. 2
Bakery        0.024609
Pharmacy      0.021834
Park          0.015479
Gym           0.024158
Market        0.004273
crime_rate    0.321536

Cluster No. 3
Bakery        0.021606
Pharmacy      0.004812
Park          0.025618
Gym           0.037016
Market        0.001584
crime_rate    1.000000

Cluster No. 4
Bakery        0.000000
Pharmacy      0.000000
Park          1.000000
Gym           0.000000
Market        0.000000
crime_rate    0.160768

Cluster No. 5
Bakery        0.000000
Pharmacy      0.000000
Park          0.450000
Gym           0.000000
Market        0.000000
crime_rate    0.486176
```

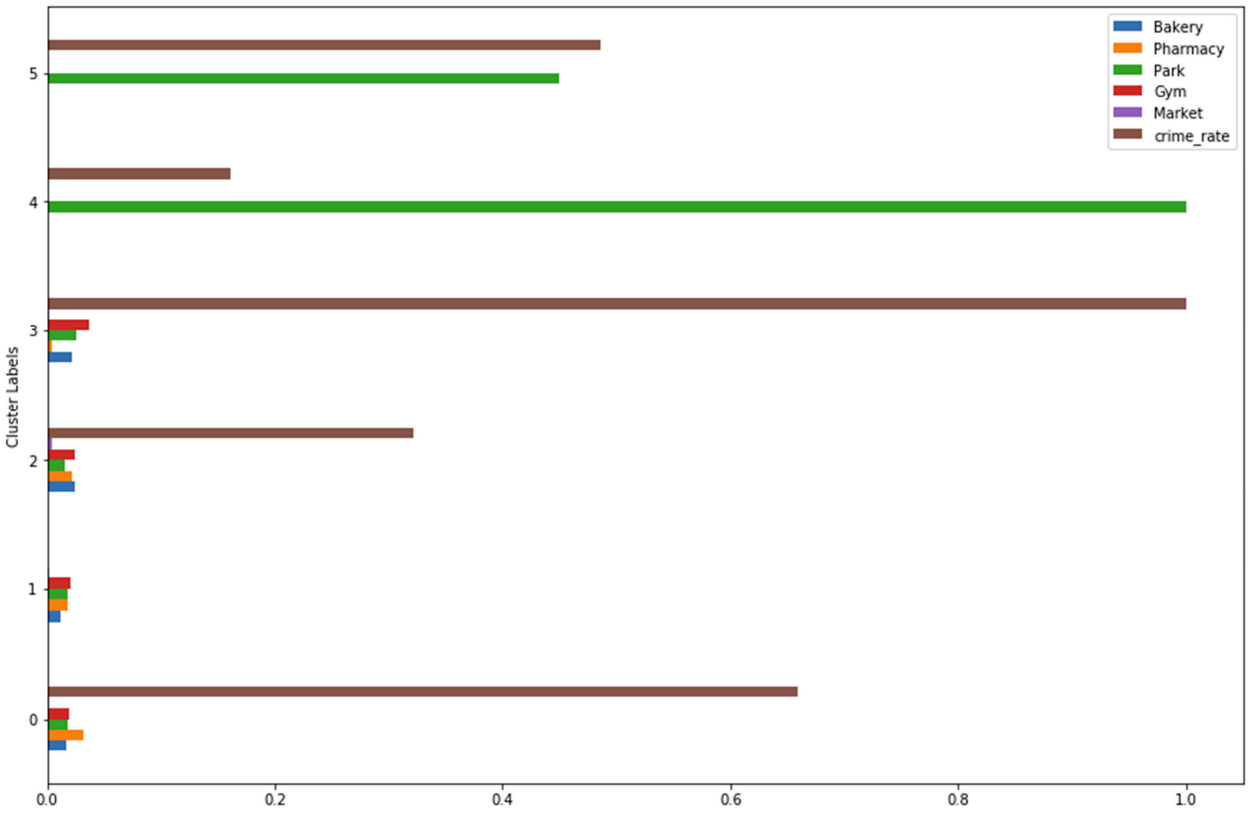


Figure 6 Mean venues and crime rate per cluster

Looks like the most attractive cluster in cluster No. 1. It has all required venues and low crime rate. Let's apply additional data filtering and visualize a result. We need only neighborhoods with parks and markets as they are mandatory for the customer. Other venues are preferable.

	Borough	Neighborhood	Latitude	Longitude	Bakery	Pharmacy	Park	Gym	Market	crime_rate
51	Staten Island	Chelsea	40.594726	-74.189560	0.056604	0.0	0.009434	0.009434	0.018868	0.0
266	Staten Island	Sunnyside	40.612760	-74.097126	0.041667	0.0	0.020833	0.041667	0.020833	0.0

Figure 7 Venues, which meets customer's requirements

So, the recommended neighborhoods are Chelsea and Sunnyside on Staten Island.

5. Discussion

So, we found two neighborhoods candidates that meet customer's requirement best and which we can recommend to the customer. Unfortunately, they don't have any pharmacy, but this venue was optional and other criteria are better than for other neighborhoods.

Finally, let's visualize results on a map:

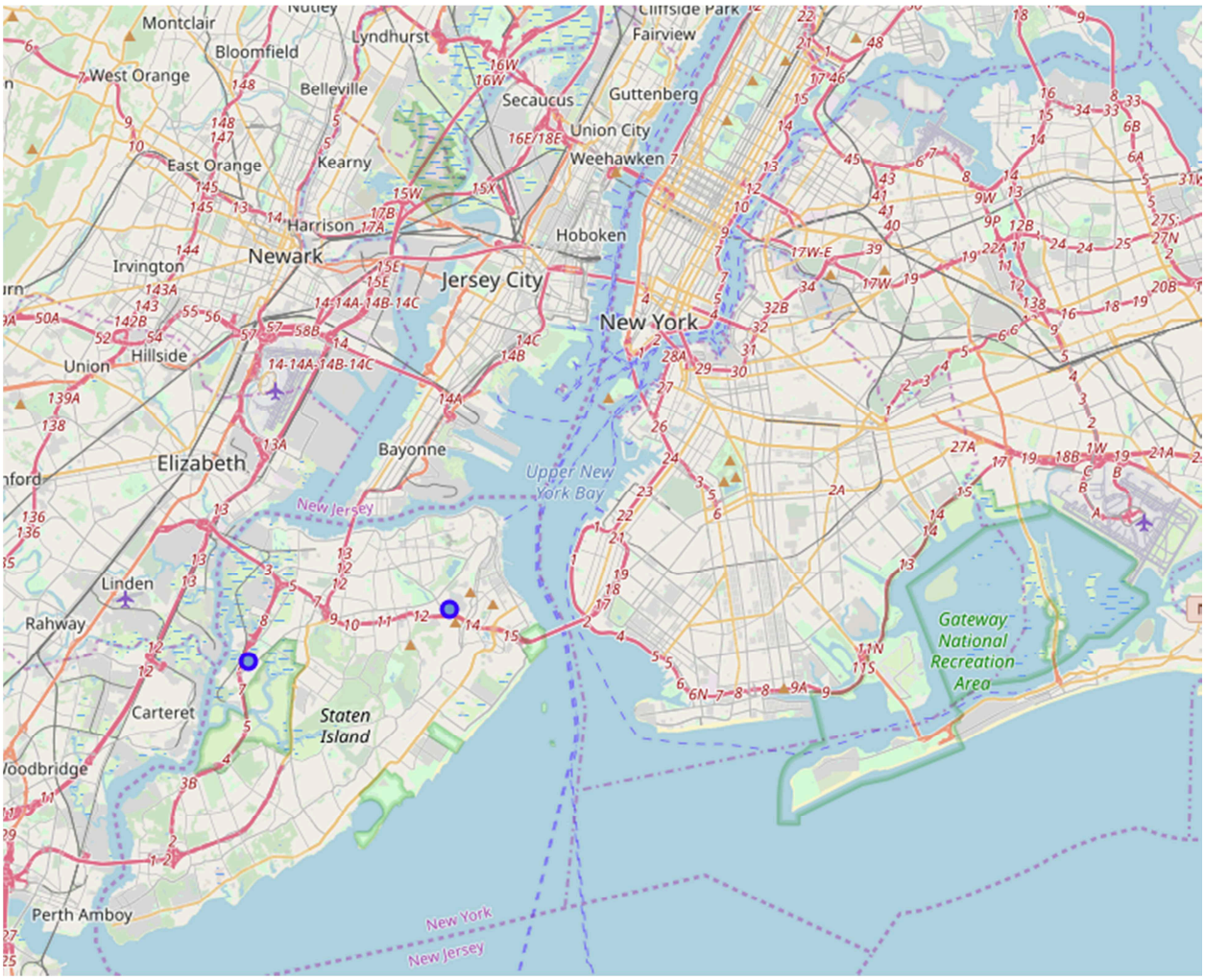


Figure 8 Recommended neighborhoods

6. Conclusion

In this study New York neighborhoods were analyzed to find best match for customer's requirement for a place to live. Official New York crime data and Forsquare API were used to combine data for an analysis. Then Neighborhoods clustering was performed using K-means algorithm. The results were analyzed and two neighborhoods candidates may be offered to the customer as the result of this study.