The Battle of Neighborhoods -- Safe and Delicious Bronx

Introduction

Bronx, NY, is home to a population of 1.43M people (U.S. Census Bureau, 2018). Between 2010 and 2018, Bronx experienced the largest change in population (3.4%) among the five boroughs in New York City (New York City Planning Department analysis).

Meanwhile, Bronx remains to be one of the most diverse boroughs in New York City. In 2018, there were 1.12 times more Other (Hispanic) residents (464k people) in Bronx County, NY than any other race or ethnicity (Data USA). Hispanics (all races) accounts for 56.4% of the total population, while White non-Hispanic population were only 9% of the total. The most common foreign languages spoken in Bronx County, NY are Spanish, Yoruba, Twi, Igbo, or Other Languages of Western Africa, and French (Incl. Cajun).

Its diversity and fairly low rent (compared to other boroughs) made it a popular landing spot for immigrants. And imagine, for this study, a Hispanic immigrant family is look for settle in Bronx, and the criterion the family is considering, other than job location, transportation access, housing prices, are safety and places to eat. So which neighborhoods in Bronx stand out? — This is the problem this report is trying address.

Data

For this project we need the following data:

1. New York City Boroughs, Neighborhoods data
   - Data source: https://data.cityofnewyork.us/City-Government/Neighborhood-Tabulation-Areas/cpf4-rkhq
   - Description: I will use this dataset to get the neighborhood information of NYC, in particular Bronx.

2. New York City Crime data
   - Data source: https://data.cityofnewyork.us/Public-Safety/Crime-Map-/5jvd-shfj
   - Description: Crime data has location (lat and long) and categories

3. New York City Places
   - Data source: Foursquare API
   - Description: By using this API we will get all the venues in each neighborhood.

Methodology

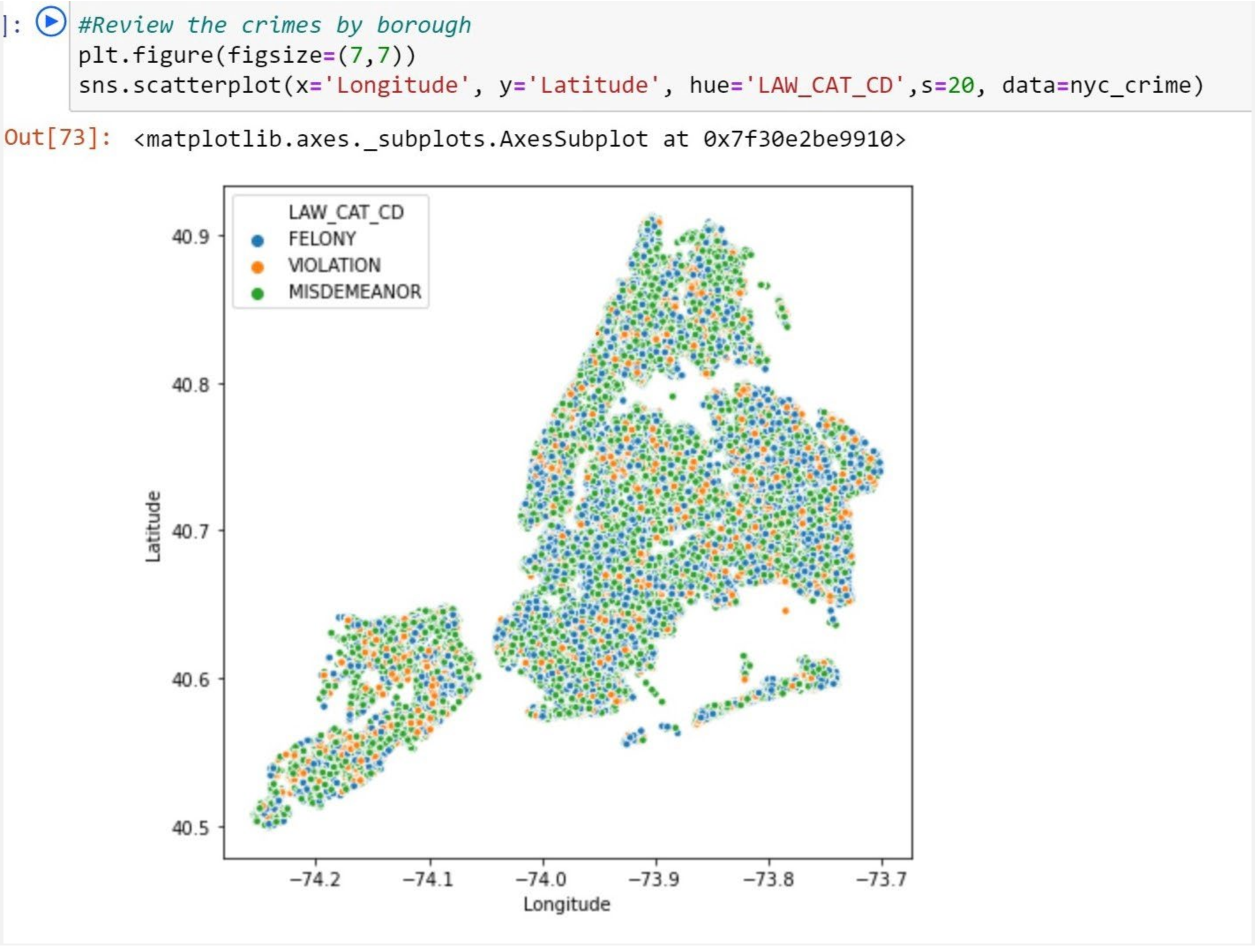
I will use different data visualization methods to explore the crime data and venue data. I will use GeoPandas package to perform various spatial analysis. I will also explore the neighborhoods clusters using K-means.

Exploration and Results

New York City Crime Data

The crime map data is updated by the New York Police Department quarterly. The dataset includes all valid felony, misdemeanor, and violation crimes reported to the New York City Police Department (NYPD) for all complete quarters, and was most recently updated with September 2020 incidents. It includes latitude and longitude of the incidents, the categories, and borough name, **but it doesn't not include neighborhood information**. I accessed the data on December 10th, 2020.

Again, we are interested the number of crimes since 2019 in Bronx Borough, New York City. But let's look at the citywide map first.

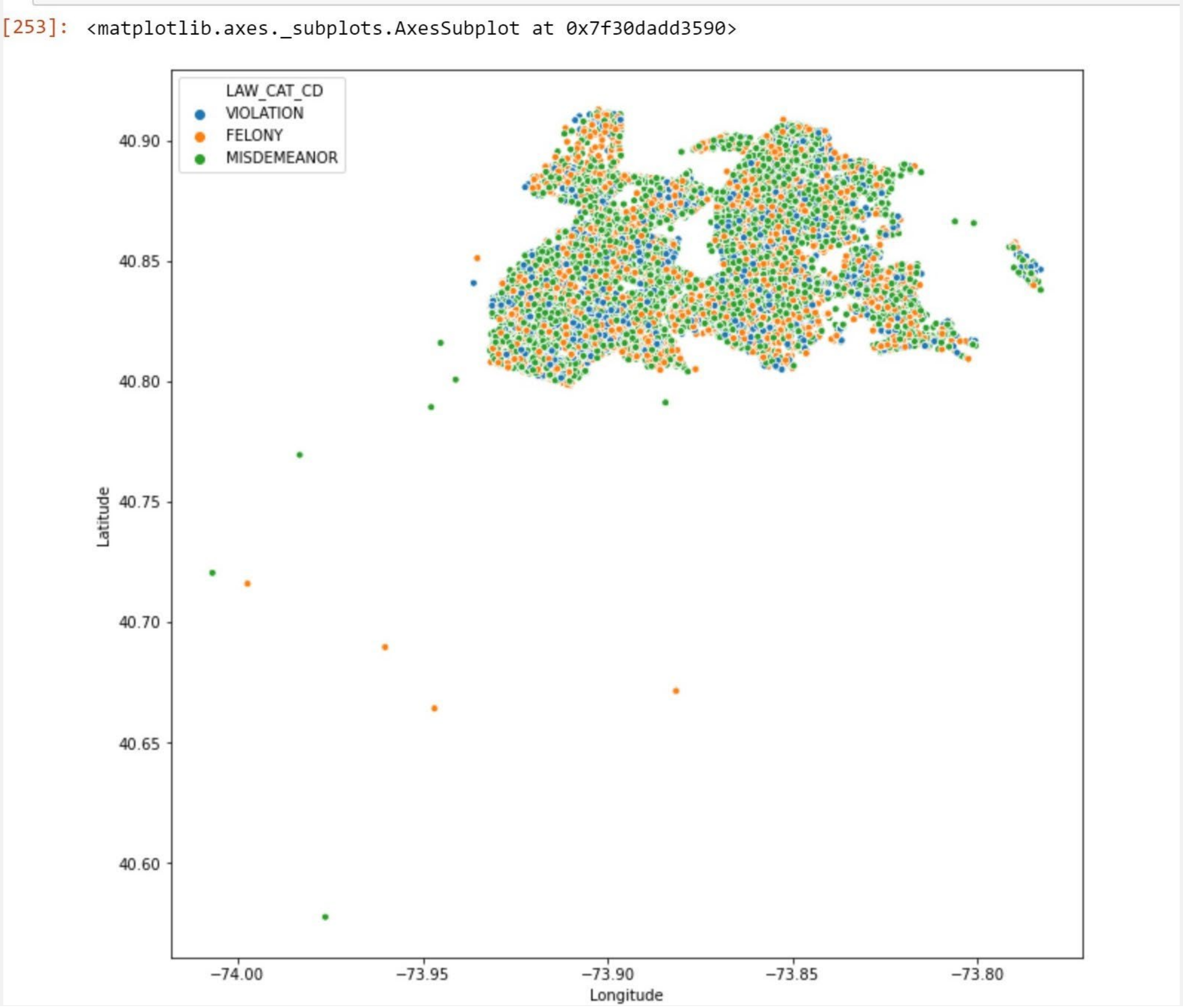


```
]: #Review the crimes by borough
   plt.figure(figsize=(7,7))
   sns.scatterplot(x='Longitude', y='Latitude', hue='LAW_CAT_CD',s=20, data=nyc_crime)

Out[73]:  <matplotlib.axes._subplots.AxesSubplot at 0x7f30e2be9910>
```

NYC crime data — Screenshot from Jupyter Notebook

Filter the data by borough names, we want to focus on Bronx.



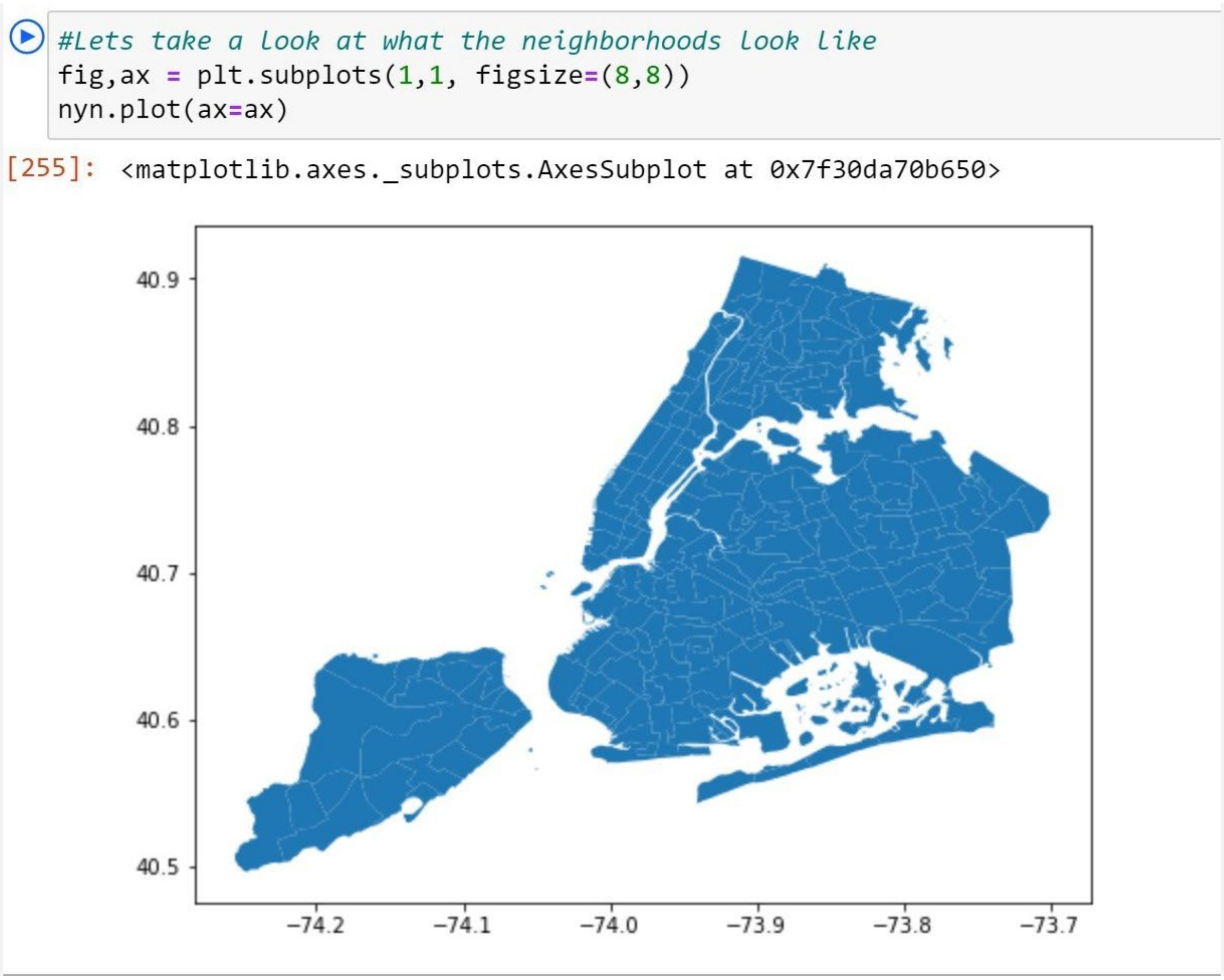
```
[253]: <matplotlib.axes._subplots.AxesSubplot at 0x7f30dadd3590>
```

Bronx crimes — Screenshot from Jupyter Notebook

We can see from the plot that there are some outliers that are outside the clusters — crime data coded wrong. But we don’t need to worry about it now, since we will be bring in neighborhood boundary map, and we can use the boundary to filter out those erroneous data.

Neighborhood Boundary Data

The boundary data is also provided by the city. The data includes borough name, neighborhood name, and most important, the geometry column that enables us to use GeoPandas package to perform various spatial analysis. I accessed the data on December 10th, 2020.
First, let’s take a look at the map.

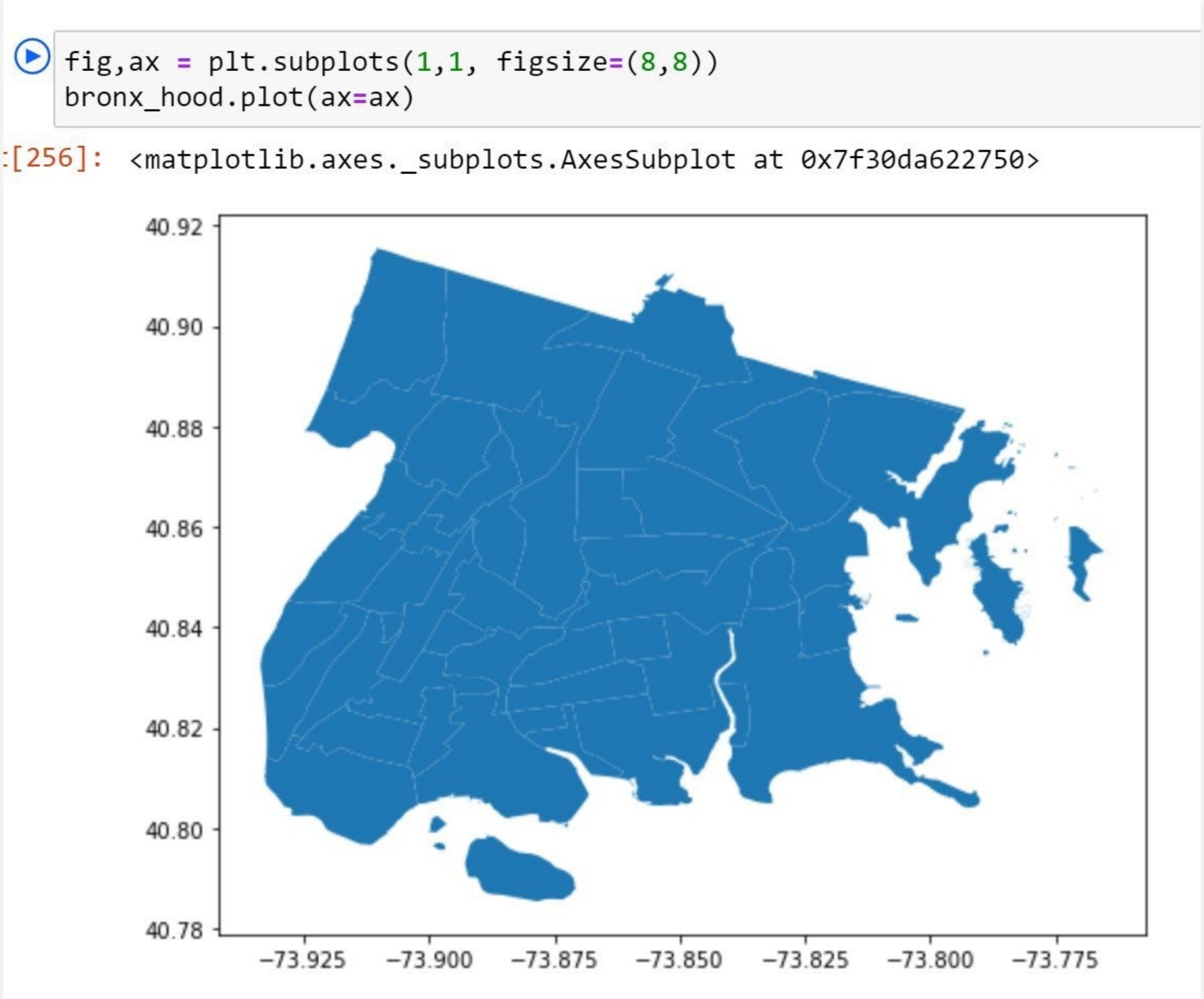


```
#Lets take a look at what the neighborhoods look like
fig,ax = plt.subplots(1,1, figsize=(8,8))
nyn.plot(ax=ax)
```

```
[255]: <matplotlib.axes._subplots.AxesSubplot at 0x7f30da70b650>
```

NYC neighborhoods — Screenshot from Jupyter Notebook

Then, filter and zoom into Bronx.



```
fig,ax = plt.subplots(1,1, figsize=(8,8))
bronx_hood.plot(ax=ax)
```

```
:[256]: <matplotlib.axes._subplots.AxesSubplot at 0x7f30da622750>
```

Bronx neighborhoods — Screenshot from Jupyter Notebook

Normally, using this following code would allow users to use geopandas.sjoin method to join two dataframes spatially.

```
geopandas.sjoin(df1, df2, how=””, op=’’)
```

But somehow my environment setup would not make this line of code work. As a workaround, I,

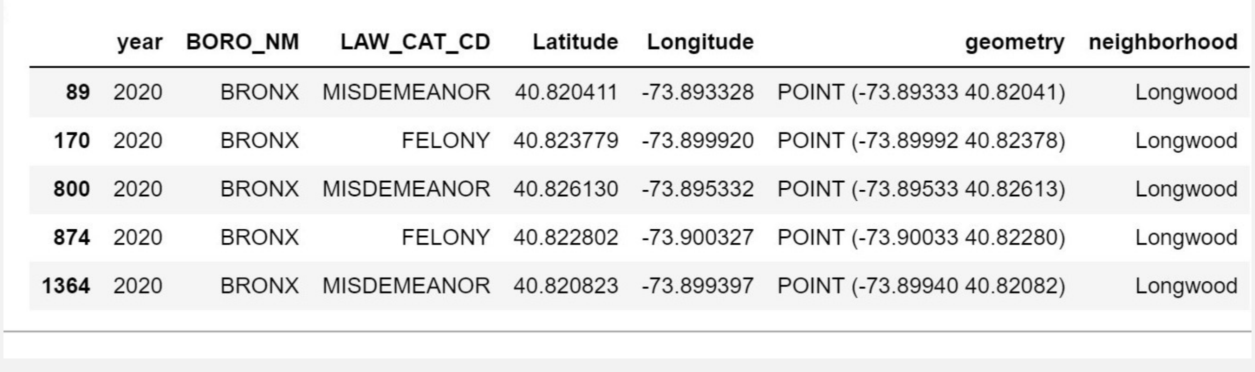
1. first turned the crime data file into GeoDataframe, using GeoPandas’s method:

```
bronx_crime_g = gpd.GeoDataFrame(bronx_crime,
geometry=gpd.points_from_xy(bronx_crime.Longitude, bronx_crime.Latitude))
```

2. I used geodataframe’s within function to test for each neighborhood, whether the crime location data falls inside its boundary or not. Here is the code:

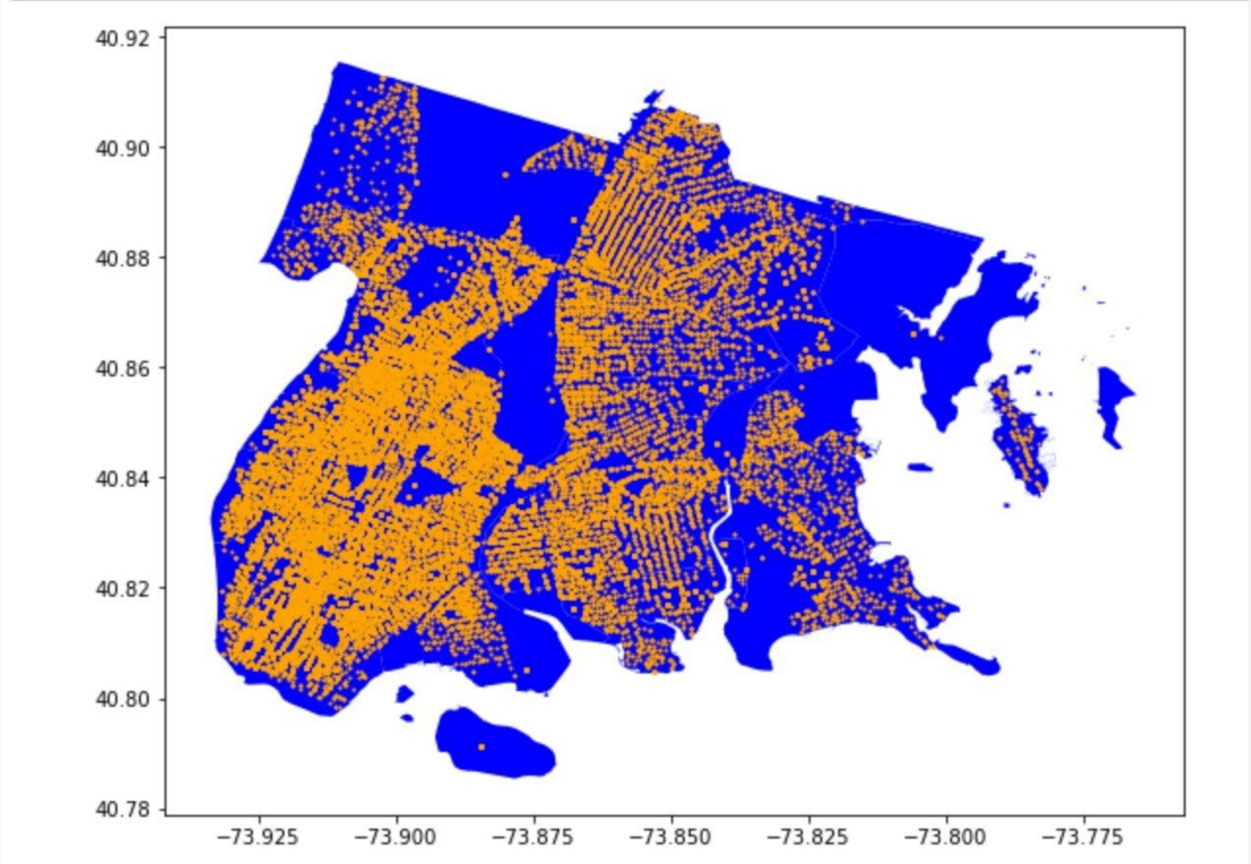
```
#geopandas spatial join is not successful, another way
crimedf_list = []
for i in range(len(bronx_hood)):
   mask = bronx_crime_g.within(bronx_hood.iloc[i, 9])
   crimedf = bronx_crime_g.loc[mask]
   print (bronx_hood.iloc[i, 6])
   crimedf['neighborhood'] = bronx_hood.iloc[i, 6]
   crimedf_list.append(crimedf)
bronx_data = pd.concat(crimedf_list)
```

And the result is — for each crime data, there is now a neighborhood value.



	year	BORO_NM	LAW_CAT_CD	Latitude	Longitude	geometry	neighborhood
89	2020	BRONX	MISDEMEANOR	40.820411	-73.893328	POINT (-73.89333 40.82041)	Longwood
170	2020	BRONX	FELONY	40.823779	-73.899920	POINT (-73.89992 40.82378)	Longwood
800	2020	BRONX	MISDEMEANOR	40.826130	-73.895332	POINT (-73.89533 40.82613)	Longwood
874	2020	BRONX	FELONY	40.822802	-73.900327	POINT (-73.90033 40.82280)	Longwood
1364	2020	BRONX	MISDEMEANOR	40.820823	-73.899397	POINT (-73.89940 40.82082)	Longwood

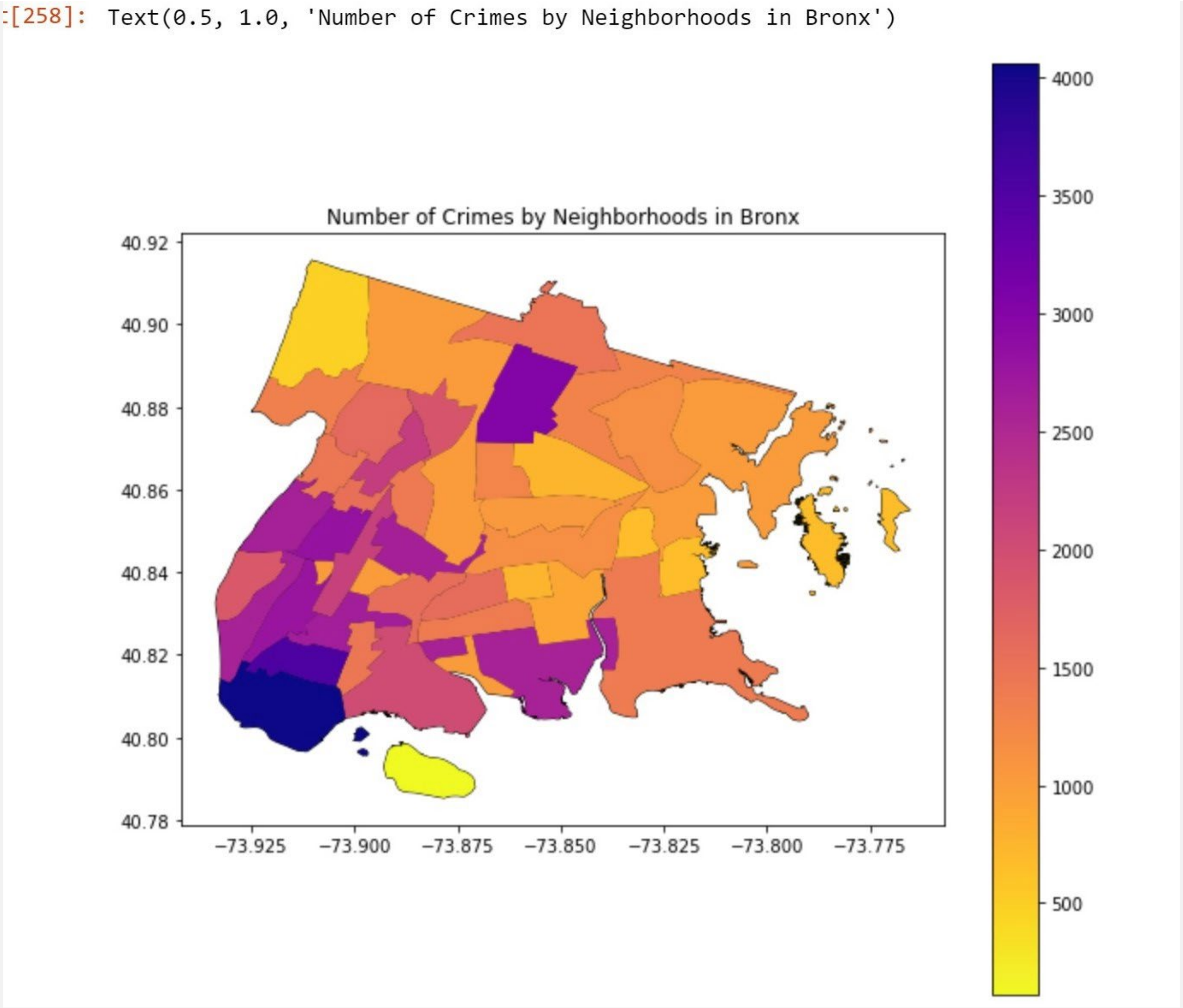
Now, let’s see the crime and neighborhood maps together.



Screenshot from Jupyter Notebook

As we can see, all the outliers are gone. The orange dots represent each crime.

And another map plot showing the intensity of crimes across al neighborhoods.



```
:[258]:  Text(0.5, 1.0, 'Number of Crimes by Neighborhoods in Bronx')
```

Number of Crimes by Neighborhoods in Bronx — Screenshot from Jupyter Notebook

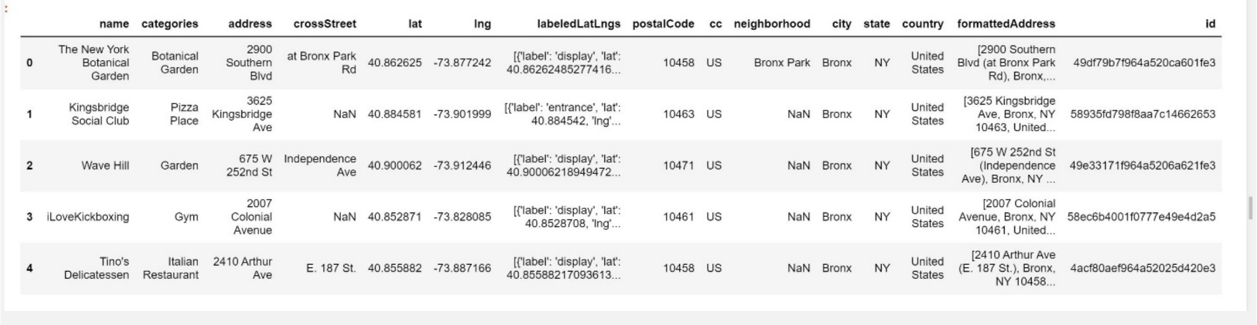
As we can see, southern and western part of Bronx experience more crimes than other parts the borough. One neighborhood in the north also has high crime rate.

Right now, let’s set the table and maps aside first, and turn our focus now to the amenities in Bronx, which we use Foursquare’s Place API.

Amenities in Bronx — using Foursquare data

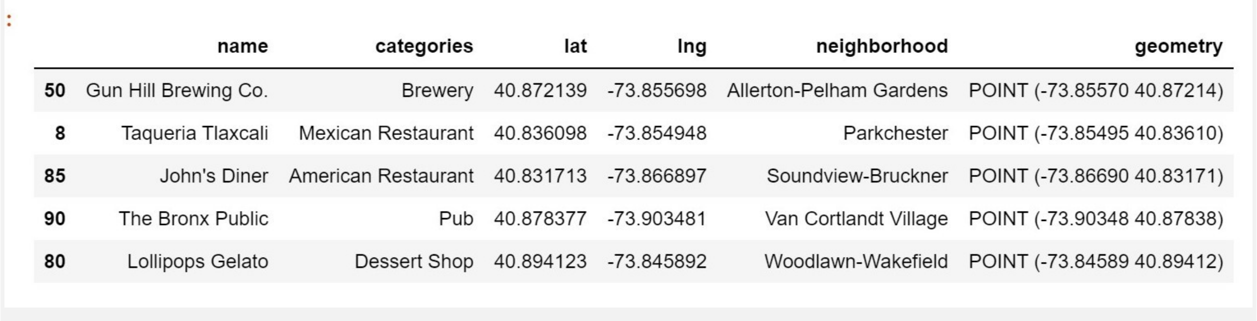
I used this the explore call in the Place API to get all the venues in Bronx, NY. I set “limit = 1000000” to get all the venues that are in Bronx. However, only 100 venues are returned from the request call.

How unfortunate! But for now, we will work with what we’ve got.



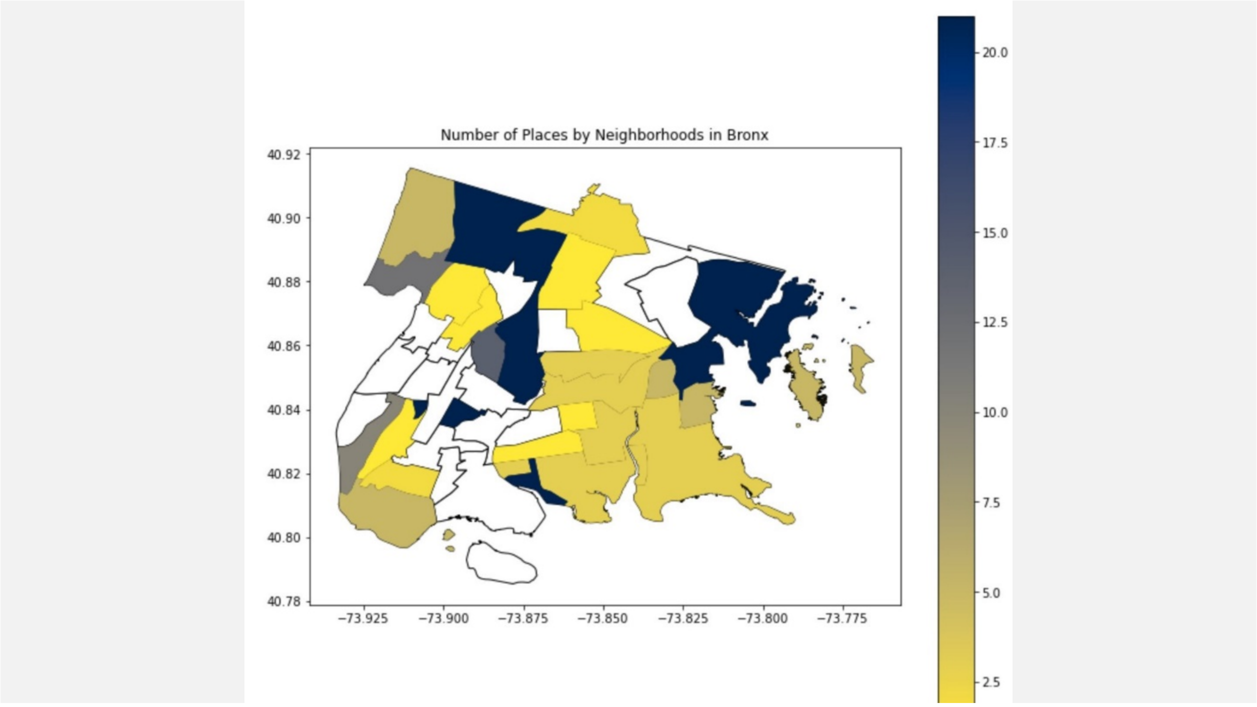
	name	categories	address	crossStreet	lat	lng	labeledLatLngs	postalCode	cc	neighborhood	city	state	country	formattedAddress	id
0	The New York Botanical Garden	Botanical Garden	2900 Southern Blvd	at Bronx Park Rd	40.862625	-73.877242	[{'label': 'display', 'lat': 40.86262485277416...	10458	US	Bronx Park	Bronx	NY	United States	[2900 Southern Blvd (at Bronx Park Rd), Bronx,...	49df79b7f964a520ca601fe3
1	Kingsbridge Social Club	Pizza Place	3625 Kingsbridge Ave	NaN	40.884581	-73.901999	[{'label': 'entrance', 'lat': 40.884542, 'lng'...	10463	US	NaN	Bronx	NY	United States	[3625 Kingsbridge Ave, Bronx, NY 10463, United...	58935fd798f8aa7c14662653
2	Wave Hill	Garden	675 W 252nd St	Independence Ave	40.900062	-73.912446	[{'label': 'display', 'lat': 40.90006218949472...	10471	US	NaN	Bronx	NY	United States	[675 W 252nd St (Independence Ave), Bronx, NY ...	49e33171f964a5206a621fe3
3	iLoveKickboxing	Gym	2007 Colonial Avenue	NaN	40.852871	-73.828085	[{'label': 'display', 'lat': 40.8528708, 'lng'...	10461	US	NaN	Bronx	NY	United States	[2007 Colonial Avenue, Bronx, NY 10461, United...	58ec6b4001f0777e49e4d2a5
4	Tino's Delicatessen	Italian Restaurant	2410 Arthur Ave	E. 187 St.	40.855882	-73.887166	[{'label': 'display', 'lat': 40.85588217093613...	10458	US	NaN	Bronx	NY	United States	[2410 Arthur Ave (E. 187 St.), Bronx, NY 10458...	4acf80aef964a52025d420e3

Looking closely, I found that many records missing neighborhood information in the “neighborhood” column. Let’s replicate what we did for the crime data, spatial join the data and add that info.



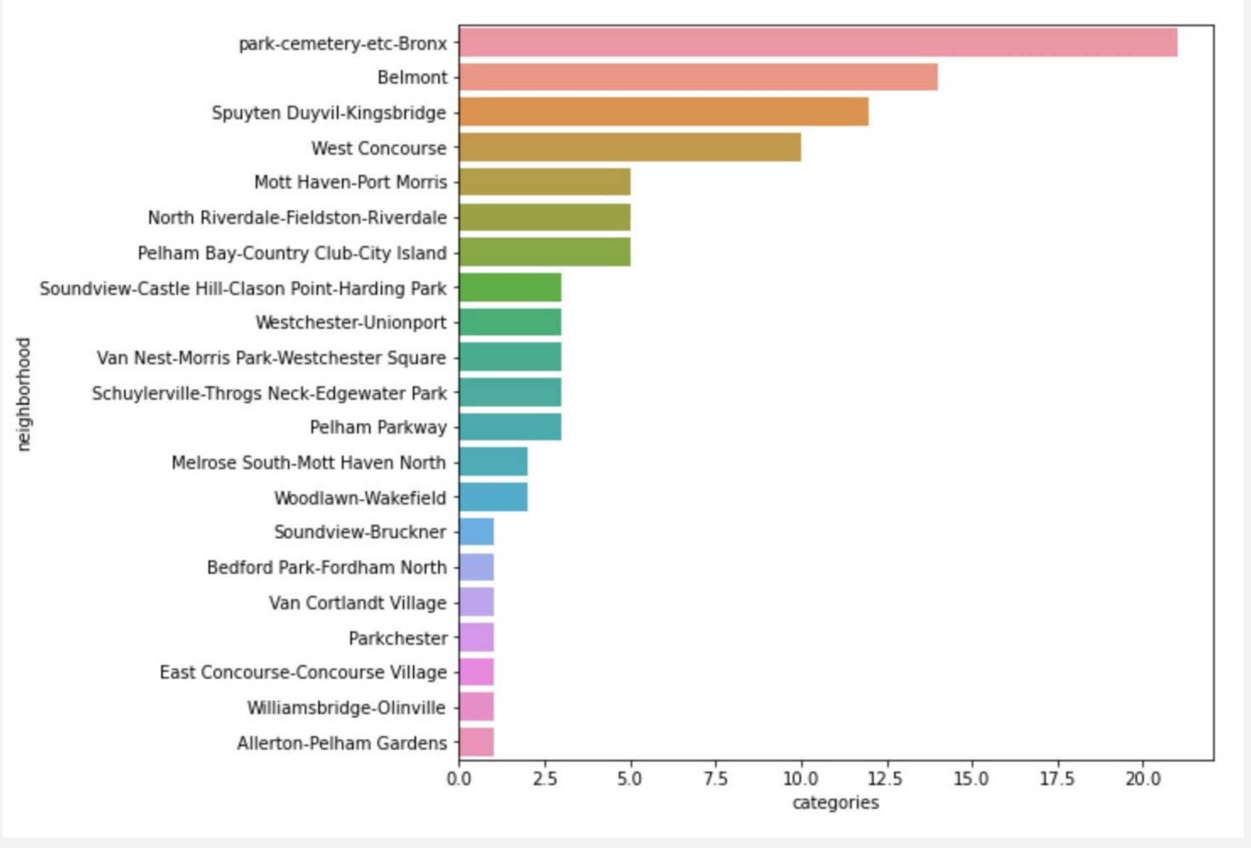
	name	categories	lat	lng	neighborhood	geometry
50	Gun Hill Brewing Co.	Brewery	40.872139	-73.855698	Allerton-Pelham Gardens	POINT (-73.85570 40.87214)
8	Taqueria Tlaxcali	Mexican Restaurant	40.836098	-73.854948	Parkchester	POINT (-73.85495 40.83610)
85	John's Diner	American Restaurant	40.831713	-73.866897	Soundview-Bruckner	POINT (-73.86690 40.83171)
90	The Bronx Public	Pub	40.878377	-73.903481	Van Cortlandt Village	POINT (-73.90348 40.87838)
80	Lollipops Gelato	Dessert Shop	40.894123	-73.845892	Woodlawn-Wakefield	POINT (-73.84589 40.89412)

Okay, now we have the neighborhood column filled. And it is consistent with the naming system in the crime data, which allows us to merge the table when needed.



Number of Places by Neighborhoods — Screenshot from Jupyter Notebook

Only 21 neighborhoods have foursquare venue information, and most venues are actually in the parks (neighborhood: park-cemetery-etc-Bronx). Again, we will use the data we have.



So this doesn’t tell us much about the neighborhood. Next, I did a simple cluster analysis using K-Means on this unlabeled venue data to see if there are any similarities among those neighborhoods.

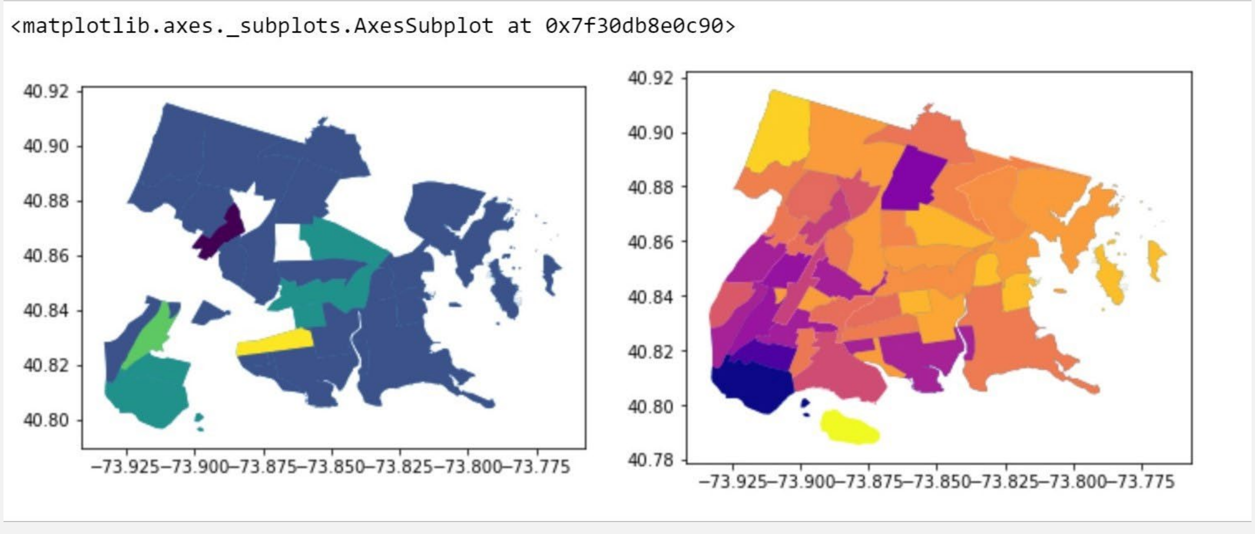
More specifically, I

1. used one hot code to preprocessing the data
2. used K-Means to cluster the neighborhoods
3. added the label, map it, and compare that with crime map

And here is my k-means code:

```
#clustering
from sklearn.cluster import KMeans
kclusters = 5bronx_clustering = bronx_place_grouped.drop('neighborhood', 1)# run k-means clustering
kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(bronx_clustering)# check cluster labels generated for each row in the dataframe
kmeans.labels_[0:10]
```

And let’s look at the clustered neighborhood map and crime map side by side.



```
<matplotlib.axes._subplots.AxesSubplot at 0x7f30db8e0c90>
```

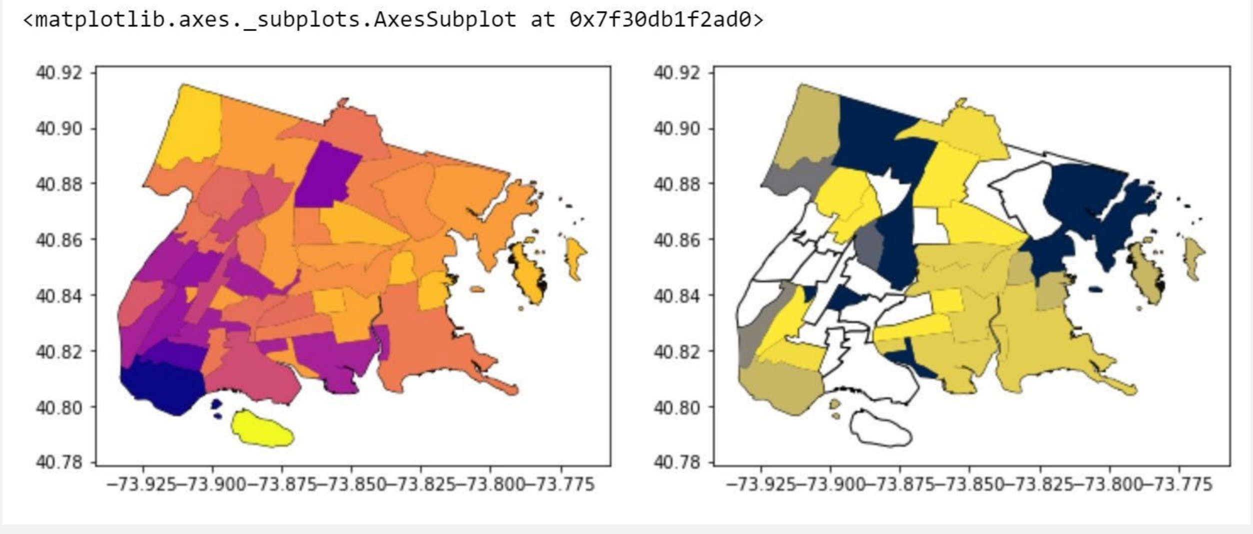
On the left side is the 5-cluster neighborhood map, again, only 21 neighborhoods have available venue data. And on the right side is the crime map data. We can see from the map plot on the left that most of the neighborhoods are similar to each other, and they are in dark blue. However, when looking at crime maps, those similar neighborhoods have very different crime patterns. And this does not help us narrow down to 1 or 2 neighborhoods that have relatively low crime rates, and relatively high concentration of amenities.

So, cluster analysis doesn't really helps. And with the crime data, venue data at hand, how do I choose the neighborhood for the immigrant family?

The answer: your second-grade algebra. — I calculated a weighted average score.

Let's go over this step by step.

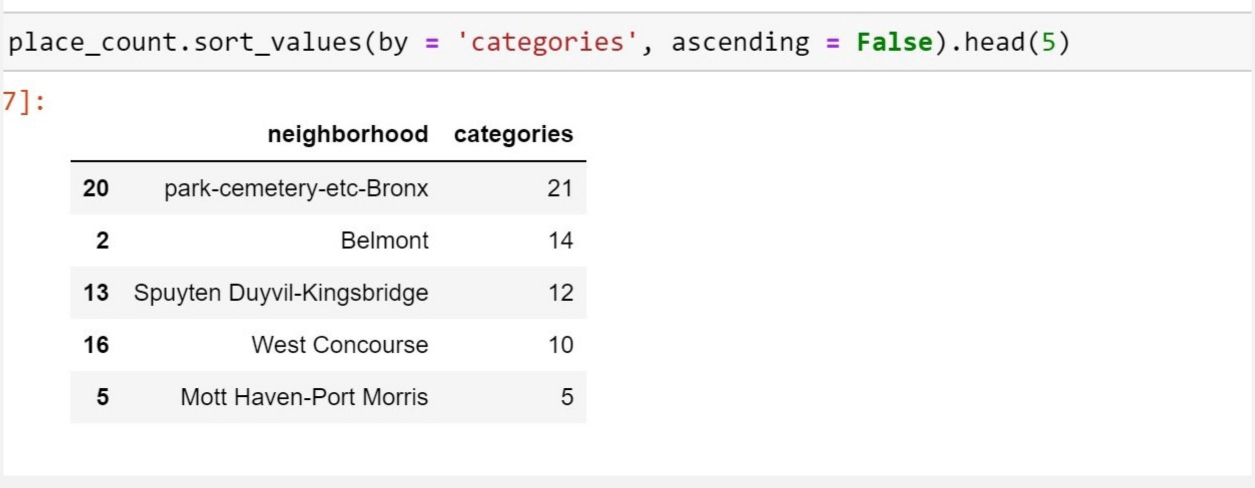
- Let's review two maps — crime map (right) and venue map (left):



```
<matplotlib.axes._subplots.AxesSubplot at 0x7f30db1f2ad0>
```

It seems that we might be able to choose the neighborhoods with highest number of venues and the lowest crime rates.

- Top 3 neighborhoods with the highest number of venues:



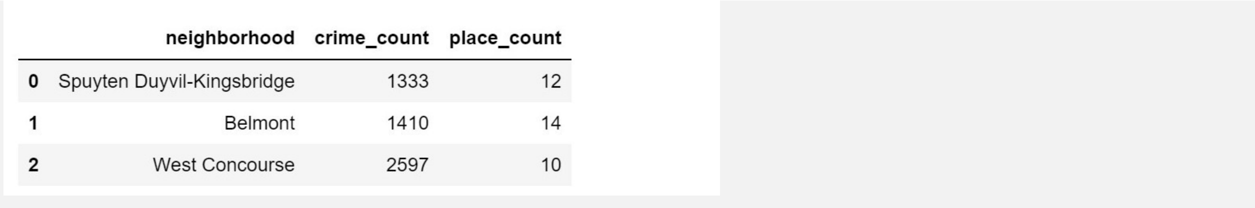
```
place_count.sort_values(by = 'categories', ascending = False).head(5)
```

7]:

	neighborhood	categories
20	park-cemetery-etc-Bronx	21
2	Belmont	14
13	Spuyten Duyvil-Kingsbridge	12
16	West Concourse	10
5	Mott Haven-Port Morris	5

As we see it again, the neighborhood that has the highest number is park-cemetery in Bronx. And of course we don't want to live in parks or cemeteries. And among the top5, between West Concourse and Mott Haven-Port Morris there is a significant gap. It seems naturally that we can choose Belmont, Spuyten Duyvil-Kingsbridge, and West Concourse, these three neighborhoods to further examine their crime data.

- Top3 neighborhood crime and place data:



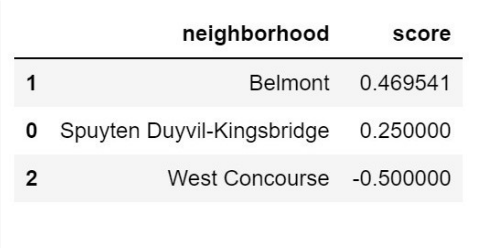
	neighborhood	crime_count	place_count
0	Spuyten Duyvil-Kingsbridge	1333	12
1	Belmont	1410	14
2	West Concourse	2597	10

Of course, we can't just choose Belmont as our winner, since it has more crimes than Spuyten Duyvil-Kingsbridge!

- Let's standardize and calculated a weighted score (assuming that safety and amenity is of equal importance for this family):

```
#let's do a simple weighted average calculation
bronx_m['crime_standarized'] = (bronx_m['crime_count'] - min(bronx_m['crime_count']))/
(max(bronx_m['crime_count']) - min(bronx_m['crime_count']))bronx_m['place_standarized'] =
(bronx_m['place_count'] - min(bronx_m['place_count']))/ (max(bronx_m['place_count']) -
min(bronx_m['place_count']))bronx_m['score'] = bronx_m['place_standarized'] *0.5 +
bronx_m['crime_standarized'] * -1 * 0.5 #crime is bad, so negative
bronx_m.sort_values(by = 'score', ascending = False, inplace = True)
bronx_m[['neighborhood', 'score']]
```

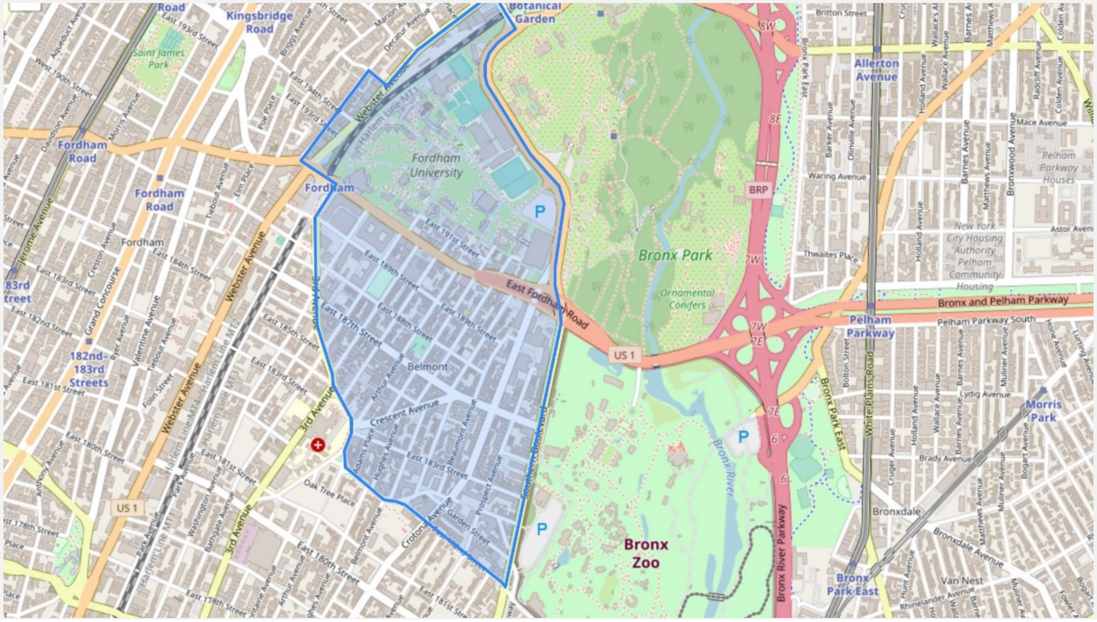
using the code above, we can produce the table below.



	neighborhood	score
1	Belmont	0.469541
0	Spuyten Duyvil-Kingsbridge	0.250000
2	West Concourse	-0.500000

Ah, we got a winner! It is Belmont indeed! Let’s use folium to map it!

```
latitude = 40.856673
longitude = -73.877499
Belmont= bronx_hood_3[bronx_hood_3.neighborhood == 'Belmont']
Belmont.crs = "EPSG:4326"Belmont_map = folium.Map(location=[latitude, longitude],
zoom_start=15)
folium.GeoJson(data=bronx_final["geom"]).add_to(Belmont_map)
# display map
Belmont_map
```



Belmont neighborhood, Bronx, NY

As we can see, Belmont is next to Bronx Park, Bronx Zoo, The New York Botanical Garden. It has the Fordham University. And there is hospital on 3rd Avenue. It seems to be the perfect neighborhood for our immigrant family!

Discussion

According to the analysis, Belmont, Bronx, NY will be a good choice for our new immigrant family who values safety and neighborhood amenities.

Belmont has high number of amenities and low number of crimes. And the weighted average indicates that looking at these two factors jointly, Belmont stands out.

One of the limitations that was mentioned throughout the post is that the analysis is based heavily on data provided by Foursquare API. It doesn't provide enough data points for our analysis. It only has its biases, venues reported by Foursquare users reflects the users' preferences (which we see that most venues are in parks and cemeteries).

In addition, although GeoPandas, and Scikit-Learn packages are very powerful, many of the tools that are available in QGIS or ArcGIS couldn't be easily replicated, which limits the number of tools we can use.

Conclusion

Not all safe neighborhoods have good amenities. Not all neighborhoods with good amenities are safe. I believe that a simple weighted average score method is a good method to choose the desirable neighborhood given the data availability and limitations, particularly data from Foursquare API.

As for our immigrant family, welcome to America! Welcome to Belmont, Bronx!