Finding a neighborhood of interest to move in/visit in Newyork city.

Abstract
Machine learning allows for the creation of computational models capable of identifying patterns in multi-dimensional datasets. This project aims to leverage venue data from Foursquare's 'Places API' and a machine learning algorithm called 'k-means clustering' to identify 'New York City' neighborhoods of similar tastes.

Introduction

Background

Every single person is unique in their own way. Different people choose different parameters when deciding to move to a new place, be it a local deciding to relocate or a tourist just looking for a place to visit. Classifying like-minded neighborhoods would help a new person to decide where to move or to just have some fun. Several entities like coffee shops, movie theatres bring together large amounts of like-minded people, influence popular culture, and contribute to the growth of the neighborhood in general.

Problem

"Suggest whether or not to move to a neighborhood based on its neighborhood profile. Suggest a tourist which neighborhood to visit based on their interest. "
Cities are, in part, entities varying from coffee shops to operas that not only provide to the needs of local citizens but also to tourists from around the world. For bigger cities, the entities can be spread apart or concentrated based on its geographical proximity to a place of significance like a monument or a memorial, resulting in an ecosystem of neighborhoods that evolve and change over time. This ecosystem is often learned by people through either natural life experience (wandering) or recommendations in the form of internet reviews, comments, and conversations with people in-real-life.

Stakeholders

Different parties may be interested in a model that is able to quantify neighborhood similarity based on the types of outlets available. Such a model would be able to inform renters and home buyers who prefer to live according to their taste. Future venue start-ups can utilize the model to identify neighborhoods lacking venues and ensure they are investing in an area that is not saturated. Future retail vendors, sellers can similarly utilize the model to ensure they are launching a business where competition is in their favor.

Methodology

Data Sources

NYU Spatial Data Repository: I will be using the ‘2014 New York City Neighborhood Names’ dataset hosted by NYU’s Spatial Data Repository as the basis for the neighborhood information. https://geo.nyu.edu/catalog/nyu_2451_34572 I will be using Foursquare’s ‘Places API’ to acquire data related to ‘venues’.It is important to note that Foursquare defines a ‘venue’ as a place that one can go to, or check-in to, can be any establishment such as a restaurant or type of retail shop. Each Foursquare ‘venue’ is assigned a ‘category’ and each ‘category’ is associated with a particular ‘categoryID’. We will be grouping the neighborhood venues based on its categories and try to cluster them .

[4]:

	Borough	Neighborhood	Latitude	Longitude
0	Bronx	Wakefield	40.894705	-73.847201
1	Bronx	Co-op City	40.874294	-73.829939
2	Bronx	Eastchester	40.887556	-73.827806
3	Bronx	Fieldston	40.895437	-73.905643
4	Bronx	Riverdale	40.890834	-73.912585

[32]:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue Name	Venue Category	Venue Latitude	Venue Longitude	Venue City	Venue State
0	Wakefield	40.894705	-73.847201	The Upper Room	Music Venue	40.892567	-73.846406	NaN	New York
1	Wakefield	40.894705	-73.847201	241st Liquor Store	Other Nightlife	40.902771	-73.849898	Bronx	NY
2	Wakefield	40.894705	-73.847201	Dyme Life Radio	Music Venue	40.894541	-73.843266	Bronx	NY
3	Wakefield	40.894705	-73.847201	Par-City	Music Venue	40.890211	-73.847002	Bronx	NY
4	Wakefield	40.894705	-73.847201	Tavern	Bar	40.895898	-73.855731	Bronx	NY

Data Analysis

The series of images below are meant to capture my process for exploring the data retrieved from Foursquare in an effort to better understand what kind of venues were actually pulled during my requests. In a perfect world, each entry would be -located in New York City, but that needed to be verified.

Most entries pulled from the API request included a ‘state’ parameter equal to either ‘New York’ or ‘NY.’ Some entries included a ‘state’ parameter equal to ‘CA’, ‘MA’, and ‘NJ’ and will need to be removed

What states are the venues in?

```
[33]: prelim_venue_data.groupby('Venue State')['Venue State'].count()
```

```
[33]: Venue State
      MA             2
      NJ            17
      NY          8817
      New York     617
      Name: Venue State, dtype: int64
```

What venue categories are the entries in?

```
[34]: n_unique = len(prelim_venue_data['Venue Category'].unique())
      print(f'There are {n_unique} unique venue categories in this dataframe')
      prelim_venue_data.groupby('Venue Category')['Venue Category'].count().sort_values(ascending=False)
```

```
      There are 149 unique venue categories in this dataframe
[34]: Venue Category
      Bar                                    2094
      Lounge                                  929
      Cocktail Bar                            500
      Nightclub                               481
      Other Nightlife                         476
      Music Venue                             470
      Pub                                     341
      Wine Bar                                278
      Sports Bar                              264
```

In the preliminary dataset, there are less unique venue names than there are entries in total. This means that there are venues associated with more than one neighborhood, which is the result of queries that overlapped because of radius being set to 1000m in the API request. This will be accepted because the venue is within walking distance of the neighborhood and can influence that neighborhood’s scene.

Data Pre--Processing

The preliminary dataset was cleaned according to the Exploratory Data Analysis section above. First, venues located in states other than “New York” or “NY” were removed. Entries with “Venue State” equal to “New York” were changed to “NY.

[39]:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue Name	Venue Category	Venue Latitude	Venue Longitude	Venue City	Venue State
0	Wakefield	40.894705	-73.847201	The Upper Room	Music Venue	40.892567	-73.846406	NaN	NY
1	Wakefield	40.894705	-73.847201	241st Liquor Store	Other Nightlife	40.902771	-73.849898	Bronx	NY
2	Wakefield	40.894705	-73.847201	Dyme Life Radio	Music Venue	40.894541	-73.843266	Bronx	NY
3	Wakefield	40.894705	-73.847201	Par-City	Music Venue	40.890211	-73.847002	Bronx	NY
4	Wakefield	40.894705	-73.847201	Tavern	Bar	40.895898	-73.855731	Bronx	NY

Entries returned by Foursquare with no ‘Venue City’ and given the ‘N/A’ treatment were also removed:

Remove entries with "Venue City" equal to "N/A"

```
[40]: ny_venue_data_with_city = ny_venue_data[(ny_venue_data['Venue City'] != "N/A")]
delta = ny_venue_data.shape[0] - ny_venue_data_with_city.shape[0]
print(f'{delta} entries were removed based on "Venue City"')
ny_venue_data_with_city.head(5)
```

```
0 entries were removed based on "Venue City"
```

[40]:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue Name	Venue Category	Venue Latitude	Venue Longitude	Venue City	Venue State
0	Wakefield	40.894705	-73.847201	The Upper Room	Music Venue	40.892567	-73.846406	NaN	NY
1	Wakefield	40.894705	-73.847201	241st Liquor Store	Other Nightlife	40.902771	-73.849898	Bronx	NY
2	Wakefield	40.894705	-73.847201	Dyme Life Radio	Music Venue	40.894541	-73.843266	Bronx	NY
3	Wakefield	40.894705	-73.847201	Par-City	Music Venue	40.890211	-73.847002	Bronx	NY
4	Wakefield	40.894705	-73.847201	Tavern	Bar	40.895898	-73.855731	Bronx	NY

One-Hot-Encoding Venue Categories

In order to use Foursquare’s category values to find similar neighborhoods based on venues, a one-hot-encoding representation of each entry was created using Pandas’ ‘get_dummies’ function. The result was a dataframe of New York City -venues
where entry venue category is represented by a value of 1 in the column of matching venue category, as shown below:

```
(9434, 150)
```

[42]:

	Neighborhood	African Restaurant	American Restaurant	Amphitheater	Arcade	Arepa Restaurant	Art Gallery	Asian Restaurant	Australian Restaurant	BBQ Joint	Bar	Basketball Stadium	Beach	Beach Bar	Beer Bar	Beer Garden
0	Wakefield	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	Wakefield	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	Wakefield	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	Wakefield	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	Wakefield	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
5	Wakefield	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	Wakefield	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	Wakefield	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	Wakefield	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
9	Wakefield	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Data Visualization

Venue counts were determined for each venue category using the one hot encoded dataframe:

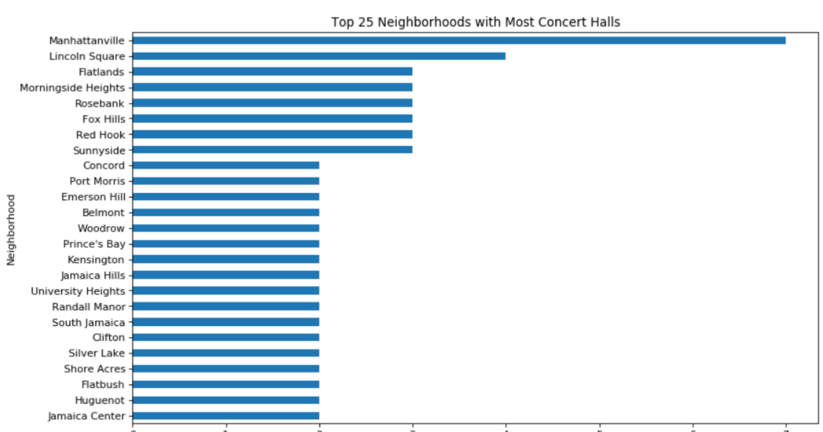
Determine the total amount of venues of each category in each neighborhood

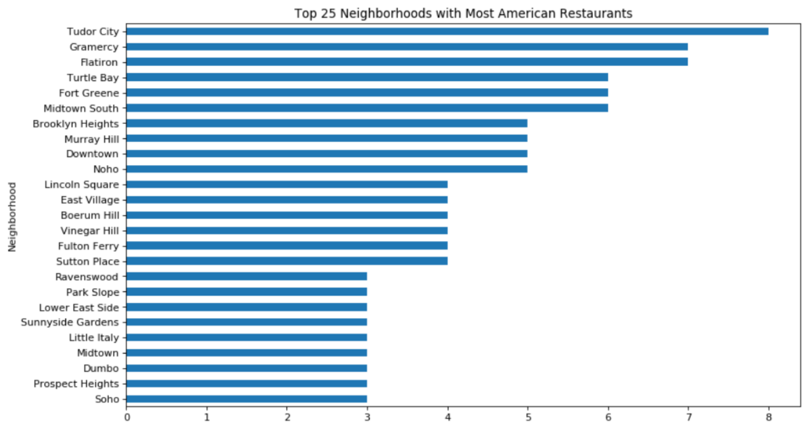
```
venue_counts = ny_venue_category_onehot.groupby('Neighborhood').sum()
venue_counts.head(10)
```

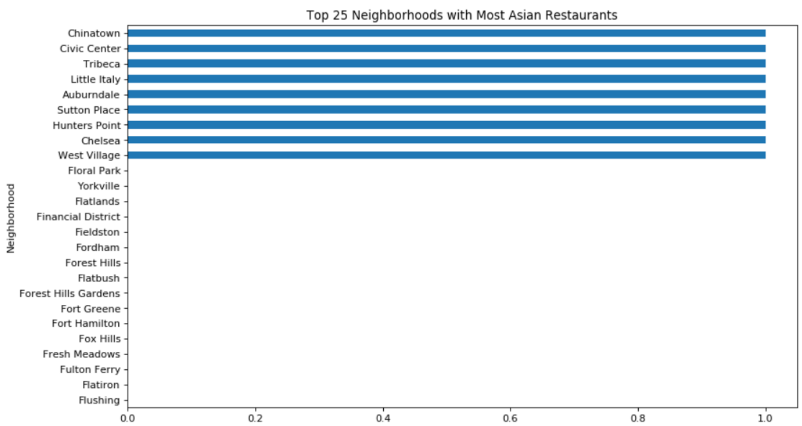
Neighborhood	African Restaurant	American Restaurant	Amphitheater	Arcade	Arepa Restaurant	Art Gallery	Asian Restaurant	Australian Restaurant	BBQ Joint	Bar	Basketball Stadium	Beach	Beach Bar	Beer Bar	Bee Garde
Allerton	0	0	0	0	0	0	0	0	0	7	0	0	0	0	
Annadale	0	0	0	0	0	0	0	0	0	2	0	0	0	0	
Arden Heights	0	0	0	0	0	0	0	0	0	3	0	0	0	0	
Arlington	0	0	0	0	0	0	0	0	0	3	0	0	0	0	
Arrochar	0	0	0	0	0	0	0	0	0	3	0	0	0	0	
Arverne	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Astoria	0	0	0	0	0	0	0	0	0	19	0	0	0	0	
Astoria Heights	0	0	0	0	0	0	0	0	0	5	0	0	1	0	
Auburndale	0	0	0	0	0	0	1	0	0	14	0	0	0	0	
Bath Beach	0	0	0	0	0	0	0	0	1	4	0	0	0	0	

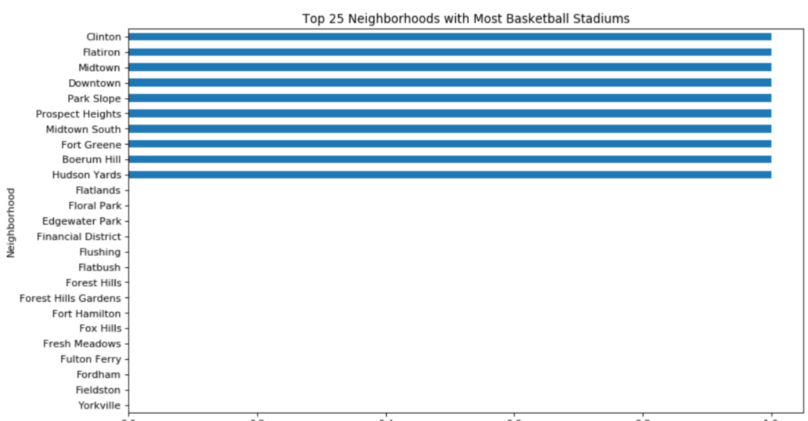
Using the dataframe of venue counts shown above, horizontal bar plots were created for select venue categories to help visualize the top 25 neighborhoods with the most of each particular venue.

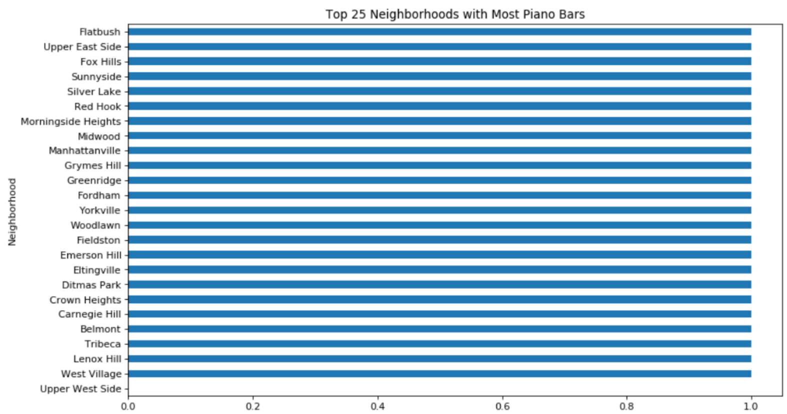
```
plot_categories = ['Concert Hall', 'American Restaurant', 'Asian Restaurant', 'Basketball Stadium', 'Piano Bar']
n = 25
for category in plot_categories:
    plt.figure(num=None, figsize=(12, 7), dpi=80, facecolor='w', edgecolor='k')
    plt.title(f'Top {n} Neighborhoods with Most {category}s')
    top_category_neighborhoods = venue_counts[category].sort_values(ascending=False)[0:n]
    top_category_neighborhoods = top_category_neighborhoods.sort_values(ascending=True)
    top_category_neighborhoods.plot.barh(y=category, rot=0)
```




Top 25 Neighborhoods with Most American Restaurants
Neighborhood
Tudor City
Gramercy
Flatiron
Turtle Bay
Fort Greene
Midtown South
Brooklyn Heights
Murray Hill
Downtown
Noho
Lincoln Square
East Village
Boerum Hill
Vinegar Hill
Fulton Ferry
Sutton Place
Ravenswood
Park Slope
Lower East Side
Sunnyside Gardens
Little Italy
Midtown
Dumbo
Prospect Heights
Soho
0
1
2
3
4
5
6
7
8


Top 25 Neighborhoods with Most Asian Restaurants
Neighborhood
Chinatown
Civic Center
Tribeca
Little Italy
Auburndale
Sutton Place
Hunters Point
Chelsea
West Village
Floral Park
Yorkville
Flatlands
Financial District
Fieldston
Fordham
Forest Hills
Flatbush
Forest Hills Gardens
Fort Greene
Fort Hamilton
Fox Hills
Fresh Meadows
Fulton Ferry
Flatiron
Flushing
0.0
0.2
0.4
0.6
0.8
1.0


Top 25 Neighborhoods with Most Basketball Stadiums
Neighborhood
Clinton
Flatiron
Midtown
Downtown
Park Slope
Prospect Heights
Midtown South
Fort Greene
Boerum Hill
Hudson Yards
Flatlands
Floral Park
Edgewater Park
Financial District
Flushing
Flatbush
Forest Hills
Forest Hills Gardens
Fort Hamilton
Fox Hills
Fresh Meadows
Fulton Ferry
Fordham
Fieldston
Yorkville
0.0
0.2
0.4
0.6
0.8
1.0



Feature Selection:

The encoded dataset of -related venues in New York City was then used to quantify a profile for each neighborhood. For each venue category, the percent distribution of venues across each neighborhood was calculated. This information would then be used to fit a K-Means clustering algorithm to the data in an effort to determine neighborhoods of similar venue profile .

Finally, the percentage of venues in each neighborhood was calculated with respect to the total amount of venues in the dataset, by venue category. So it's clear, the value shown in the "Bar" column for Astoria represents the percentage of Bars in the dataset that are located in Astoria.

	Neighborhood	African Restaurant	American Restaurant	Amphitheater	Arcade	Arepa Restaurant	Art Gallery	Asian Restaurant	Australian Restaurant	BBQ Joint	Bar	Basketball Stadium	Beach	Beach Bar	Beer Bar	Beer Garden
0	Allerton	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.003348	0.0	0.0	0.0	0.0	0.004673
1	Annadale	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000956	0.0	0.0	0.0	0.0	0.004673
2	Arden Heights	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.001435	0.0	0.0	0.0	0.0	0.000000
3	Arlington	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.001435	0.0	0.0	0.0	0.0	0.000000
4	Arrochar	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.001435	0.0	0.0	0.0	0.0	0.000000

With the above, a dataframe showing the top five music venue categories for each neighborhood was created:

	Neighborhood	1st Top Venue Category	2nd Top Venue Category	3rd Top Venue Category	4th Top Venue Category	5th Top Venue Category
0	Allerton	Liquor Store	Brewery	Music Venue	Nightclub	Beer Garden
1	Annadale	Beer Garden	Sports Bar	Wine Bar	Pub	Other Nightlife
2	Arden Heights	Concert Hall	Pub	Hookah Bar	Cocktail Bar	Sports Bar
3	Arlington	Nightlife Spot	Gay Bar	Brewery	Sports Bar	Bar
4	Arrochar	Steakhouse	Record Shop	Nightlife Spot	Other Nightlife	Nightclub

Results

Cluster Modeling

Scikit-learn's K-Means clustering was used to determine similar neighborhoods based on venue percentage. The image below shows the data being scaled and the K-Means model being created:

```
[51]: array([ 0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,
            3,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,
            0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,
            0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  1,  0,  0,  0,
            0,  0,  0,  0,  0,  5,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,
            0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,
            0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,
            2,  0,  0,  0, 14,  0, 13,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,
            0,  0,  0, 11,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  7,  0,
            0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,
            0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  4,  0,  0, 10,  0,  0,
            0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,
            0,  9,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,
            0,  0,  0,  8,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,
            0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,
            0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,
            0,  0,  0,  0,  0,  6,  0,  0,  0,  0, 12,  0,  0,  0,  0,  0,  0,
            0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0], dtype=int32)
```

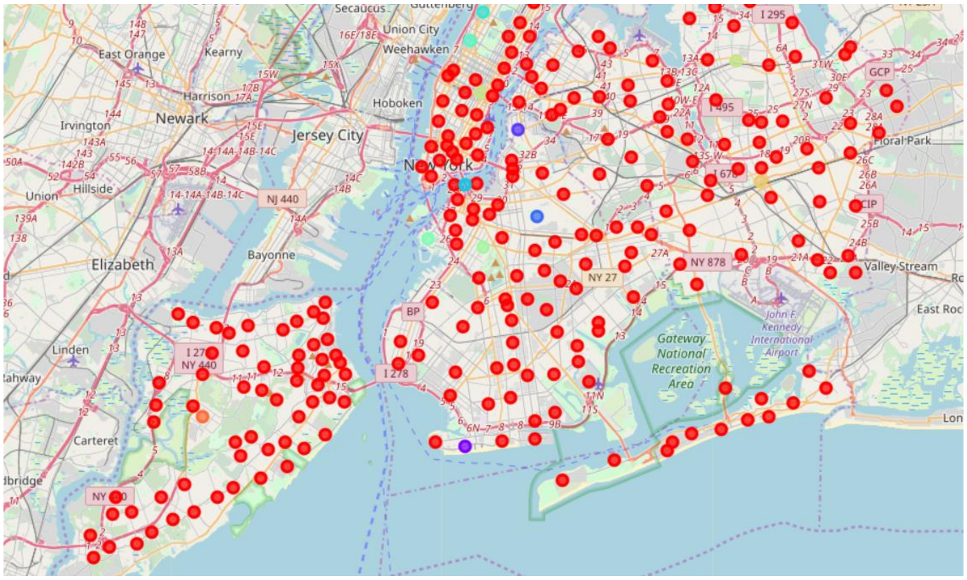
A new dataframe was created by merging neighborhood location data with cluster labels and top venue categories

[52]:

	Neighborhood	Latitude	Longitude	Cluster Labels	1st Top Venue Category	2nd Top Venue Category	3rd Top Venue Category	4th Top Venue Category	5th Top Venue Category
0	Wakefield	40.894705	-73.847201	0.0	Smoke Shop	Music Venue	Nightlife Spot	Speakeasy	Other Nightlife
1	Co-op City	40.874294	-73.829939	0.0	Jazz Club	Recording Studio	Sports Bar	Nightclub	Pub
2	Eastchester	40.887556	-73.827806	0.0	Recording Studio	Jazz Club	Sports Bar	Nightclub	Cocktail Bar
3	Fieldston	40.895437	-73.905643	0.0	Piano Bar	Nightlife Spot	Sports Bar	Record Shop	Pub
4	Riverdale	40.890834	-73.912585	0.0	Nightlife Spot	Speakeasy	Music Venue	Hotel Bar	Sports Bar

Cluster Visualization

The following code uses folium to visualize neighborhoods of similar profile by coloring each neighborhood point based on cluster label:



Conclusion

Machine learning and clustering algorithms can be applied to multi-dimensional datasets to find similarities and patterns in the data. Clusters of neighborhoods of similar profile can be generated using high- quality venue location data. There is a preface on high- quality because analysis models are only as good as the input into them . Foursquare offers a robust 'Places API' service that can be leveragd s in similar studies and model-making. This project could be expanded on in a number of different ways. Foursquare's API could be further interrogated to retrieve and consider more related venues in New York City. New datasets of -venues can be acquired and potentially merged with what was retrieved from Foursquare. The DBSCAN clustering algorithm, better at maintaining dense clusters and ignoring outliers, could be implemented and compared to KMeans. The clustering model could become the basis for a recommendation system aimed to provide neighborhoods of similar profile to users.