New York City - Segmentation On Basis Of NightLife

Sahil Sanwal

Introduction

Background

Nightlife of a city influences various businesses in a city and has a target audience that is growing very fast . Activities that a city has to offer can vary from lounges to night shows to beach bars to cafes all these sectors contribute to the economy and have various interconnected ecosystems within them . Here we study and organise clusters of neighbourhoods based on these nightlife activities to study the patterns in data so we can suggest a hotel chain on where they can open their next hotel

Problem

Cities are composed of various nightlife activities for local residents and tourists alike. A hotel chain firm is planning to expand their business and open new hotel in New York.

Chain is not bothered about the pricing and is focused mostly on nightlife activities in the city since most of their clients are tourists looking for nightlife activities .

They approach us to use machine learning to segment various neighbourhoods based on two basic parameters-

a) how many nightlife activities are present in 1km radius of the central location of the neighbourhood

b) what are the primary nightlife activities in these neighbourhoods

This project aims to quantify and monitor the state of neighbourhoods in a major metropolitan city, New York City, and identify clusters of similar Nightlife activity scenes.

Stakeholders

Different parties may be interested in a model that is able to quantify neighborhood similarity based on the types of nightlife available. Such a model would be able to inform renters and home buyers who prefer to live where the nightlife is happening that their next home is properly located. Future nighlife activity start-ups can utilize the model to identify neighborhoods lacking nightlife and ensure they are investing in an area that is not saturated.

Methodology

Data Sources

NYU Spatial Data Repository I am using the '2014 New York City Neighbourhood Names' dataset hosted by NYU's Spatial Data Repository as the basis for the neighbourhood names and associated location centroids [0]. The image to the right shows a sample of this information:

	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

Foursquare 'Places API'

I will be using Foursquare's 'Places API' to acquire data related to 'venues' (as defined by Foursquare) categorized to be somehow associated with music [1]. It is important to note that Foursquare defines a 'venue' as a place that one can go to, or checkin to, and that a 'venue' is not necessarily a music venue but 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'. The image to the right shows the 'categoryID' values provided by Foursquare that will be used to acquire music related venues within New York City:

Beach Bar : '4bf58dd8d48988d116941735'
Beer Bar : 56aa371ce4b08b9a8d57356c
Beer Garden : 4bf58dd8d48988d117941735
Champagne Bar : 52e81612bcbc57f1066b7a0e
Cocktail Bar : 4bf58dd8d48988d11e941735
Dive Bar : 4bf58dd8d48988d118941735
Hookah Bar : 4bf58dd8d48988d119941735
Karaoke Bar : 4bf58dd8d48988d120941735
Sports Bar : 4bf58dd8d48988d11d941735
Whisky Bar : 4bf58dd8d48988d122941735
Wine Bar : 4bf58dd8d48988d123941735
Brewery : 50327c8591d4c4b30a586d5d
Lounge : 4bf58dd8d48988d121941735
Night Market : 53e510b7498ebcb1801b55d4
Nightclub : 4bf58dd8d48988d11f941735
Other Nightlife : 4bf58dd8d48988d11a941735
Strip Club : 4bf58dd8d48988d1d6941735

Data Retrieval

Neighborhood Name & Location Centroid Data

The '2014 New York City Neighborhood Names' dataset hosted by NYU's Spatial Data Repository was easy to download as a JSON file and import into a Jupyter Notebook

The 'Borough', 'Neighborhood', 'Latitude', and 'Longitude' values associated with each neighborhood were then converted from JSON to a Pandas DataFrame that serves as the foundation of the analysis.

Foursquare Nightlife Related Venue Data

As mentioned in the Data Sources section of this report, Foursquare has numerous 'Venue Categories' that are used to identify each type of venue. A 'get' request to the 'api.foursquare.com/v2/venues/search?' endpoint that provides a category ID will return venues of that category

Exploratory Data Analysis

Exploratory data analysis was done to answer multiple questions regarding nighlife, the questions and findings are mentioned below-

Q-What states are venues located in ?

A-Refer to image below, following are states and number of venues

```
Out[110]: Venue State
          NJ              648
          NY            14328
          New Jersey       26
          New York        236
          Name: Venue State, dtype: int64
```

Q-How many categories are there that are unique for nightlife

A-There are 104 unique categories

Data Preprocessing

Data cleaning

The preliminary dataset was cleaned according to the answers listed in 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."

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

A list of nightlife related venue categories was created based on the unique venue categories included in the preliminary dataset. This list was used to filter out the non nighlife related entries that snuck into our request

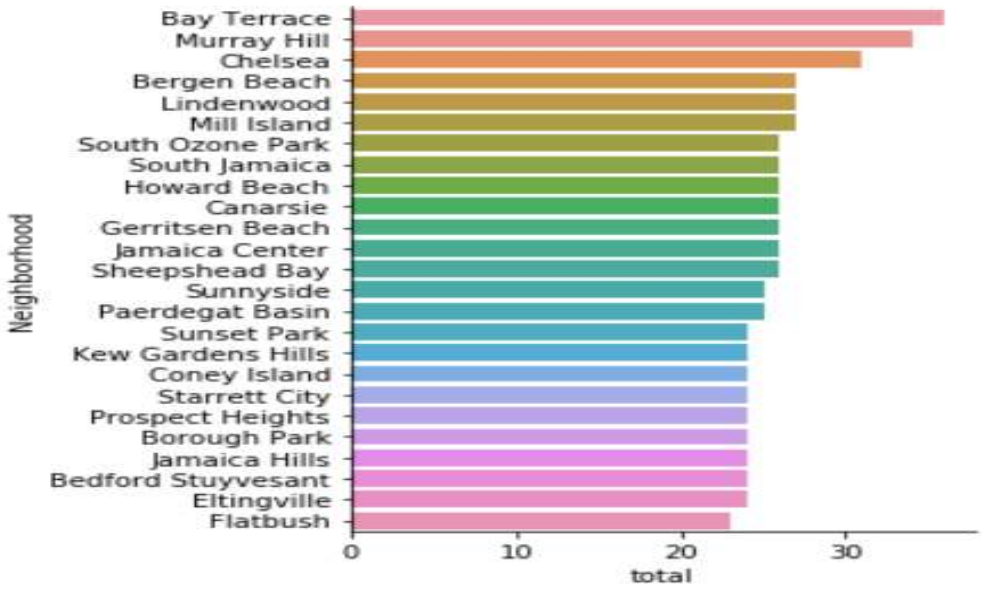
One Hot Encoding Venue Categories

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

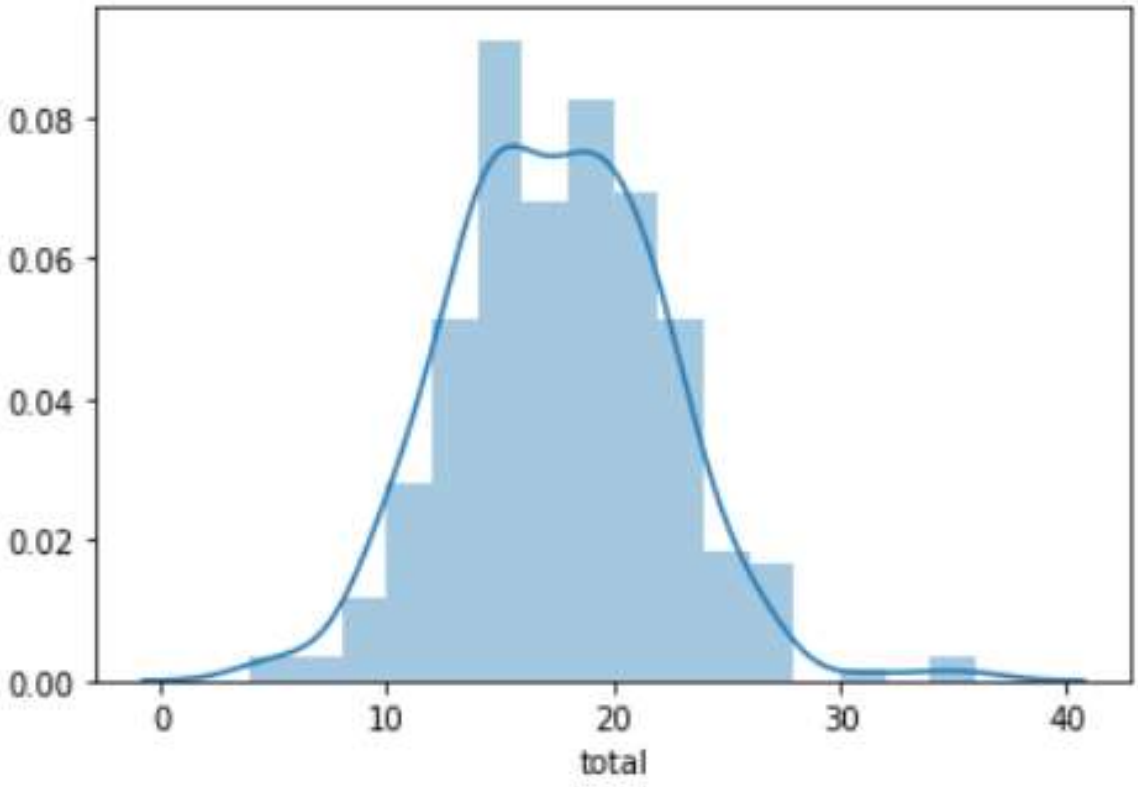
	Neighborhood	Beach Bar	Beer Bar	Beer Garden	Brewery	Cocktail Bar	Dive Bar	Hookah Bar	Karaoke Bar	Lounge
1	Wakefield	0	0	0	0	0	0	0	0	1
2	Wakefield	0	0	0	0	0	0	0	0	1
5	Wakefield	0	0	0	0	0	0	0	0	0
8	Wakefield	0	1	0	0	0	0	0	0	0
9	Wakefield	0	1	0	0	0	0	0	0	0
17	Wakefield	0	0	0	0	0	0	0	0	0
19	Wakefield	0	0	1	0	0	0	0	0	0

Data Visualization

Description of top 25 neighbourhoods nightlife



Description of density variations of nightlife in New York city



Feature Generation

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

determine amount of venues of each category

In [178]:
```
venue_totals = {}
for category in nightlife_related_categories:
    if category in venue_counts.columns:
        venue_totals[category] = venue_counts[category].sum()


venue_totals_sorted = sorted(venue_totals.items(), key=lambda x: x[1],reverse=
venue_totals_sorted
```

Out[178]:
```
[('Cocktail Bar', 907),
 ('Lounge', 692),
 ('Brewery', 563),
 ('Sports Bar', 528),
 ('Beer Garden', 455),
 ('Beer Bar', 420),
 ('Nightclub', 367),
 ('Wine Bar', 330),
 ('Other Nightlife', 293),
 ('Dive Bar', 235),
 ('Karaoke Bar', 214),
 ('Hookah Bar', 156),
 ('Beach Bar', 35),
 ('Whisky Bar', 34),
 ('Strip Club', 18),
 ('Night Market', 8)]
```

Finally, the percentage of venues in each neighbourhood was calculated with respect to the total amount of venues in the dataset, by venue category

Results

Cluster Modelling

Scikitlearn's KMeans clustering was used to determine similar neighborhoods based on music venue percentage. The image below shows the data being scaled and the KMeans model being created:

```
In [200]: kclusters = 8
          kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(venue_grouped_cluste

          kmeans.labels_

Out[200]: array([5, 4, 5, 0, 5, 2, 0, 0, 3, 0, 0, 0, 3, 5, 3, 7, 5, 0, 3, 5, 2, 5,
                 6, 0, 5, 0, 5, 0, 0, 7, 3, 5, 7, 0, 5, 0, 5, 0, 5, 0, 4, 5, 5, 0,
                 0, 4, 0, 0, 5, 5, 0, 6, 0, 0, 0, 5, 5, 0, 0, 5, 0, 3, 5, 0, 0, 5,
                 0, 5, 1, 0, 0, 5, 5, 0, 0, 0, 5, 0, 0, 0, 0, 0, 0, 5, 5, 7, 4, 5,
                 5, 6, 4, 5, 0, 7, 0, 0, 0, 0, 0, 5, 3, 6, 3, 0, 0, 5, 5, 3, 0, 0,
                 5, 5, 0, 0, 0, 5, 5, 5, 5, 5, 0, 5, 0, 5, 0, 7, 5, 0, 0, 3, 6, 6,
                 5, 5, 0, 0, 4, 0, 5, 0, 6, 6, 6, 1, 0, 6, 3, 0, 6, 5, 0, 0, 5, 0,
                 0, 0, 5, 0, 5, 0, 0, 3, 5, 0, 0, 0, 5, 6, 5, 0, 1, 0, 0, 5, 0, 0,
                 0, 0, 5, 0, 0, 4, 5, 0, 0, 0, 3, 2, 5, 5, 5, 0, 5, 0, 5, 0, 0, 5,
                 6, 5, 1, 0, 5, 5, 0, 0, 5, 0, 4, 5, 5, 4, 5, 3, 0, 0, 0, 4, 0, 0,
                 0, 5, 3, 0, 5, 0, 0, 4, 0, 6, 5, 5, 1, 0, 5, 2, 2, 0, 5, 5, 5, 7,
                 0, 5, 4, 5, 5, 5, 5, 0, 7, 5, 5, 1, 1, 0, 5, 0, 5, 5, 0, 0, 0, 0,
                 0, 0, 0, 0, 4, 5, 0, 5, 5, 0, 0, 0, 4, 0, 0, 0, 3, 5, 0, 5, 0, 1,
                 5, 6, 0, 4, 0, 3, 5, 0, 5, 0, 0, 0, 5, 4, 0, 0])
```

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

1]:

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
Wakefield	40.894705	-73.847201	5	Other Nightlife	Lounge	Beer Bar	Sports Bar	Beer Garden
Co-op City	40.874294	-73.829939	5	Other Nightlife	Lounge	Brewery	Nightclub	Beer Bar
Eastchester	40.887556	-73.827806	5	Other Nightlife	Lounge	Brewery	Beer Bar	Beer Garden
Fieldston	40.895437	-73.905643	0	Wine Bar	Beer Bar	Lounge	Sports Bar	Brewery
Riverdale	40.890834	-73.912585	0	Beer Bar	Wine Bar	Sports Bar	Brewery	Other Nightlife

Cluster Visualization:

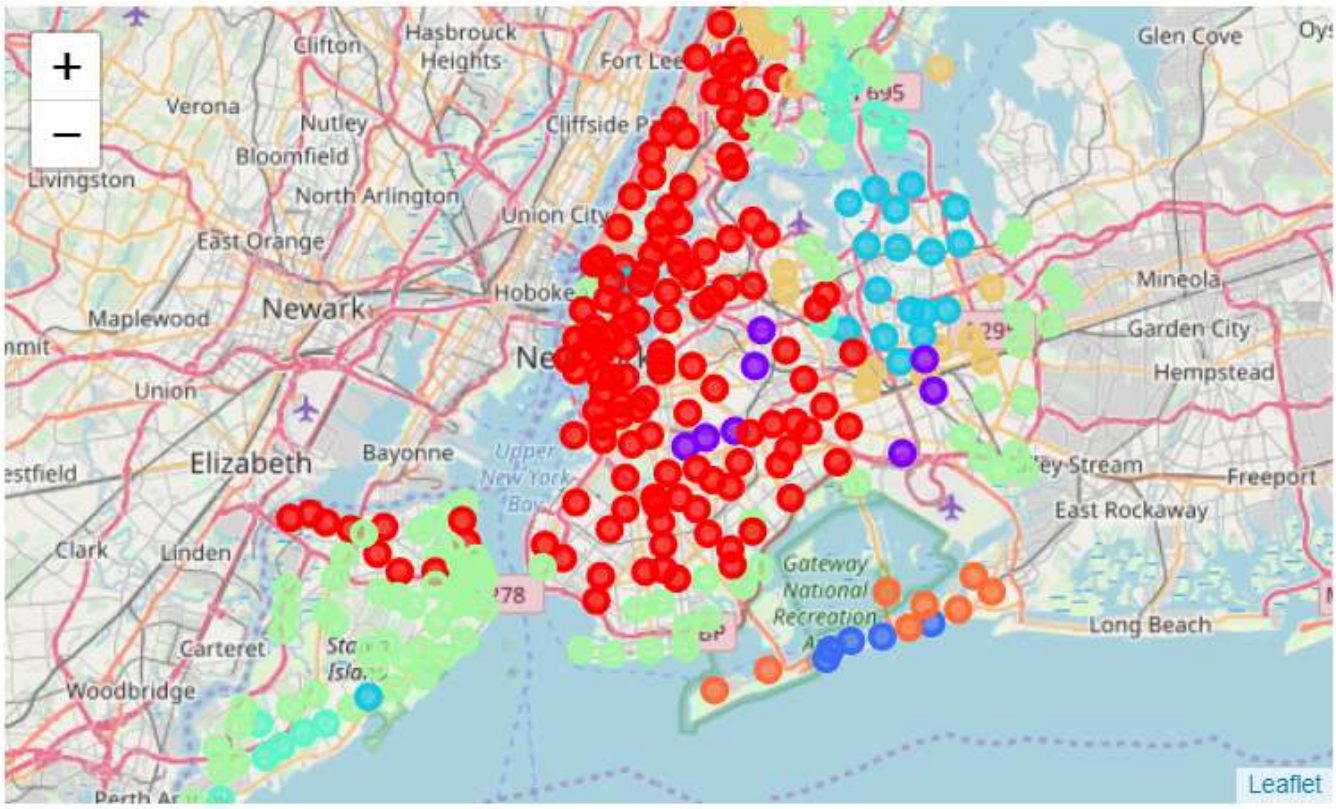
The following code uses folium to visualize neighborhoods of similar nightlife profile by coloring each neighborhood point based on cluster label:

```
# create map
import matplotlib.cm as cm
import matplotlib.colors as colors
latitude = 40.730610
longitude = -73.935242
map_clusters = folium.Map(location=[latitude, longitude], zoom_start=10)

# set color scheme for the clusters
x = np.arange(kclusters)
ys = [i + x + (i*x)**2 for i in range(kclusters)]
colors_array = cm.rainbow(np.linspace(0, 1, len(ys)))
rainbow = [colors.rgb2hex(i) for i in colors_array]

# add markers to the map
markers_colors = []
for lat, lon, poi, cluster in zip(ny_neighborhood_music_profile['Latitude'], r
    label = folium.Popup(str(poi) + ' Cluster ' + str(cluster), parse_html=Tru
    folium.CircleMarker(
        [lat, lon],
        radius=5,
        popup=label,
        color=rainbow[int(cluster)-1],
        fill=True,
        fill_color=rainbow[int(cluster)-1],
        fill_opacity=0.7).add_to(map_clusters)

map_clusters
```



Conclusion

Machine learning and clustering algorithms can be applied to multidimensional datasets to find similarities and patterns in the data. Clusters of neighborhoods of similar nightlife profile, or any profile, can be generated using highquality venue location data. There is a preface on highquality because analysis models are only as good as the input into them (garbage in, garbage out). Luckily, Foursquare offers a robust ‘Places API’ service that, although (as we have seen) not perfect (nothing is), can be leverages in similar studies and modelmaking. This project could be expanded on in a number of different ways. Foursquare’s API could be further interrogated to retrieve and consider more musicrelated venues in New York City. New datasets of nightlife related 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 music profile to users.