IBM Data Science Capstone Project

The Battle of Neighborhoods

Potential locations for the sharing economy of power banks in New York

I. Introduction

It's a Friday night, after a long and tedious week, you decide to have some fun with friends. After flicking through a few Facebook pages, you and your friends decide to go to a newly opened restaurant-bar, that seem to be quite popular among your Facebook acquaintances. You guys were having a great time taking photos of the food, of each other and instagraming these all night long. Suddenly a friend says, do you guys have chargers, I didn't bring mine because they are so heavy and inconvenient. Me neither, you replied, and looking at your phone's battery going red, you start worrying about how to get an Uber later when your phone goes out!

Although seeming a little dramatic, but the above scenario is quite common nowadays as mobile devices become almost necessities for people (especially the younger generation). People rarely carry mobile chargers or power banks, not because they are pricey but because they are inconvenient and heavy. Wouldn't it be nice to have easily accessible power banks when you need them the most?

Indeed, a number of companies have started market testing power banks for rent. They typically cost the user less than 50 cents per hour of charging.

Figure 1.1 A user returning the power bank to the rental station after usage. Fees can be automatically charged through Paypal or Alipay.

II. The Business Problem

Observing some business success of the power bank rental business in Asia, our company (hypothetically) decided that it could be a business opportunity in the US. However, as a start-up company, we are tight on marketing budget and can only test market in a handful of locations in New York.

After considering our constraints and potential users, we have summarized the following information, and we will use these as criteria for selecting the neighborhoods and locations for launching our pilot testing:

1. Our potential customers are 20~35 year olds.
2. Potential customers are more likely to use our service, the more time they spend at a venue (thus the more likely that their phone's battery running out).
3. The power bank rental stations need some but not extensive maintenance (assuming a monthly checkup). So they shouldn't be too spread out.
4. The power bank rental stations cost $3,000 each to produce, and we only have a budget to produce a limited amount for pilot testing.
5. Our marketing team would like to be able to go frequently around the rental stations to do promotions and observe how people actually adopt or interact with our service.

Therefore, our business problem can be described as a two-stage process:

1. Maximizing potential customer exposure:
    a) Identifying which neighborhoods and venues are the most popular in New York.
    b) Segmenting these venues as clusters, based on the similar types of venues.
    c) Select the types that best suit our potential customer criteria. For example, it is unlikely for a pharmacy to be a good venue for our power bank rental business, even if it is very popular, because people may visit pharmacy regularly, but only spend very little time on each visit. On the other hand, a popular bar may be the perfect spot for our business.
2. Minimizing marketing budget:
    a) From the venues that meet our criteria, select those in neighborhoods that are most densely concentrated, and geographically close-by. So to minimize the time, effort and cost of the marketing team going around for promotion and maintenance.

II. Data Description

In order to address our business problem, we will need data on popular places among our target customers (20~35 year olds). A good source of this kind of data is the Foursquare location data.

I believe the Foursquare data is appropriate for a number of reasons:

1. Foursquare users overlap with our target customers in terms of age-group.[1]
2. Foursquare data are user generated and updated, which is more relevant to our business, because we don't want outdated places that are recommended by critics but irrelevant for our actual target customers.

Specifically, we will need to explore the following aspects of the Foursquare data:

1. The most popular venues in all neighborhoods in New York;
2. These venues will contain information about their types (for example, restaurants, shops, bars and pharmacies);
3. We will also have to longitude and latitude of these venues.
4. After identifying the venues in particular neighborhoods, we will group them and explore neighborhoods that have similar popular venues.

III. Methodology

A. Libraries

We will be using the following libraries throughout our analysis:

```
import numpy as np # library to handle data in a vectorized manner
import pandas as pd # library for data analsysis
import json # library to handle JSON files
from geopy.geocoders import Nominatim # convert an address into lat
itude and longitude values
import requests # library to handle requests
from pandas.io.json import json_normalize # tranform JSON file into
 a pandas dataframe
# Matplotlib and associated plotting modules
import matplotlib.cm as cm
import matplotlib.colors as colors
# import k-means from clustering stage
from sklearn.cluster import KMeans  #for clustering
import folium # map rendering library
```

B. Data Preprocessing

We first get the necessary data from https://cocl.us/new_york_dataset.

From this, we have extracted the 'features' and assigned it to nb_data, which is a nested list.

---

[1] The average age of users on Foursquare is 32. Of the total users, 44% are less than 30, 42% are 30~43. http://www.ignitesocialmedia.com/2010-social-network-analysis-report/#Foursquare

In order to work with this data, we first transform it into a DataFrame using pandas.

Taking a look at the top 10 entries.

```
nb.head(10)
```

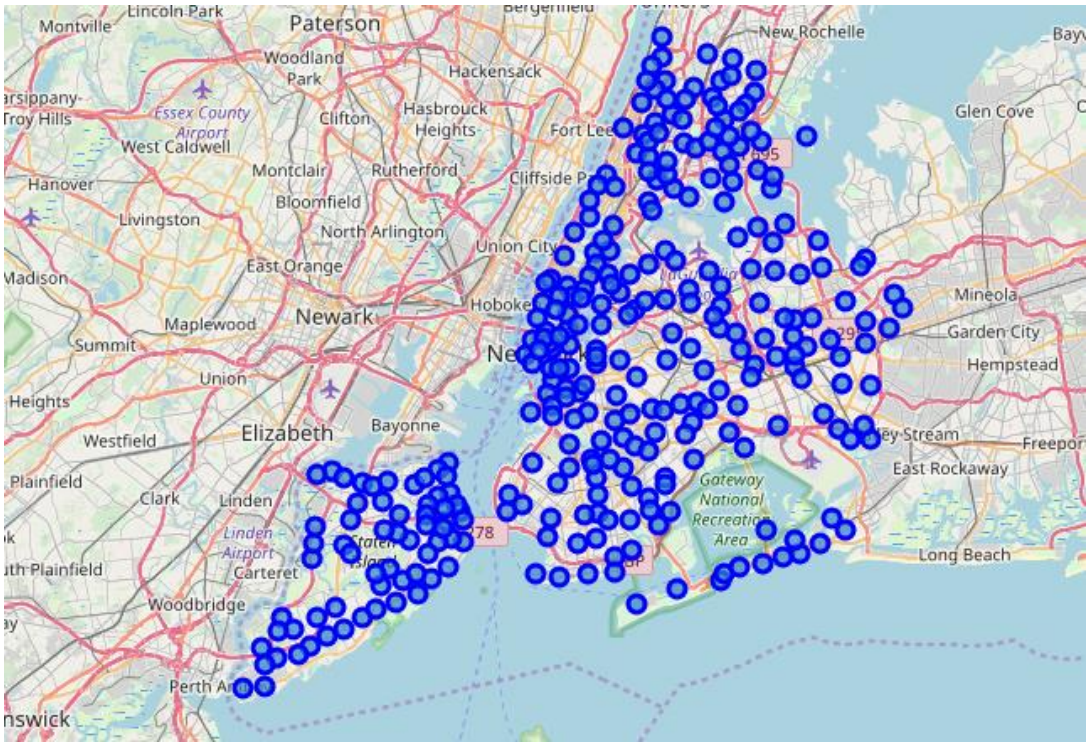
	Borough	Neighborhood	Latitude	Longitude
0	Bronx	Wakefield	40.894705	-73.847201
1	Bronx	Co-op City	40.874294	-73.829939
2	Bronx	Eastchester	40.887556	-73.827806
3	Bronx	Fieldston	40.895437	-73.905643
4	Bronx	Riverdale	40.890834	-73.912585
5	Bronx	Kingsbridge	40.881687	-73.902818
6	Manhattan	Marble Hill	40.876551	-73.910660
7	Bronx	Woodlawn	40.898273	-73.867315
8	Bronx	Norwood	40.877224	-73.879391
9	Bronx	Williamsbridge	40.881039	-73.857446

Using the .shape attribute, we see that there are 306 neighborhoods in total.

C. Data Visualization

It is always useful to be able to visualize the data and get a sense of what we will be working with.

We will use the geopy library to obtain the latitude and longitude information of New York and use folium to create a map of New York with neighborhoods flagged on top, the result is shown below.



D. Foursquare API

After defining Foursquare credentials and version, we can explore the venue information through the Foursquare API.

Although we can jump straight into exploring all neighborhoods, I started by just looking at the first one, the same way as was done in the lab. I find it useful to look at one example, and get a feel for the results we are getting, before moving onto the big dataset.

After the standard json and pandas transformations, here is the top 5 most popular venues for Wakefield neighborhood within a 500 meter radius:

	name	categories	lat	lng
0	Lollipops Gelato	Dessert Shop	40.894123	-73.845892
1	Rite Aid	Pharmacy	40.896649	-73.844846
2	Carvel Ice Cream	Ice Cream Shop	40.890487	-73.848568
3	Cooler Runnings Jamaican Restaurant Inc	Caribbean Restaurant	40.898276	-73.850381
4	Dunkin'	Donut Shop	40.890459	-73.849089

Of the top 5 places, three are sweets/snack shops, one Caribbean restaurant, and a pharmacy! Although premature, but I feel that this already is telling us a lot about this neighborhood. Its occupants are likely to be a lot of younger people, and maybe income levels here are relatively low.

E. Exploring all of the neighborhoods

Using the getNearbyVenues function defined in the labs, and using the location information from earlier we have created a ny_venues DataFrame.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Wakefield	40.894705	-73.847201	Lollipops Gelato	40.894123	-73.845892	Dessert Shop
1	Wakefield	40.894705	-73.847201	Rite Aid	40.896649	-73.844846	Pharmacy
2	Wakefield	40.894705	-73.847201	Carvel Ice Cream	40.890487	-73.848568	Ice Cream Shop
3	Wakefield	40.894705	-73.847201	Cooler Runnings Jamaican Restaurant Inc	40.898276	-73.850381	Caribbean Restaurant
4	Wakefield	40.894705	-73.847201	Dunkin'	40.890459	-73.849089	Donut Shop

We then use the one-hot encoding to create dummies for the categories of venues.

	Yoga Studio	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	American Restaurant	Antique Shop	Arcade	Arepa Restaurant	Argentinian Restaurant	Art Gallery	Art Museum
0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0	0
2	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
4	0	0	0	0	0	0	0	0	0	0	0	0

F. Neighborhood examples and discussion

After grouping all the rows by neighborhood, and taking the means of the frequency of occurrence of each category:

	Neighborhood	Yoga Studio	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	American Restaurant	Antique Shop	Arcade	Arepa Restaurant	Argentinian Restaurant	Art Gallery
0	Allerton	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0
1	Annadale	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0
2	Arden Heights	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0
3	Arlington	0.0	0.0	0.0	0.0	0.0	0.166667	0.0	0.0	0.0	0.0	0.0
4	Arrochar	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0

We can now print out the top 5 most common venues in each neighborhood. I find that this information can already tell us a lot about the nature of neighborhoods. There are more than 300 results, so I am not going to include all of them here, but I am going to present some representative and interesting cases here for discussion.

```
----Emerson Hill----
                 venue  freq
0               Casino   0.5
1       Sculpture Garden   0.5
2      Outdoor Sculpture   0.0
3    Peruvian Restaurant   0.0
4  Performing Arts Venue   0.0
```

Emerson Hill: my first impression just looking at these, is that I am literally walking into a classic movie scene. This seems like a more affluent and a more culturally strong (in terms of the classical arts) neighborhood. I can see some opportunities with power banks rental business in the casinos, but that it'd be useful for sculpture and performing arts venues.

```
----Hammels----
                 venue  freq
0                Beach   0.4
1                 Café   0.1
2           Food Truck   0.1
3                Diner   0.1
4  Fast Food Restaurant   0.1
```

Hammels: the Café, food truck, and fast food restaurants all suggest that is more of a grab-a-bite and go type of place. It makes sense as the top venue here is the beach, where people is just here to relax and play around outdoors rather than constantly be on their phones. Furthermore, a beach would seem too distant for our marketing and promotion teams.

```
----Greenridge----
           venue  freq
0     Playground  0.17
1  Bowling Alley  0.17
2          Diner  0.17
3            Pub  0.17
4    Pizza Place  0.17
```

Greenridge: the pub, the pizza place, the bowling alley, the playground, the diner! They all seem perfectly reasonable for people to stay for a while, thus increasing the possibility of needing to recharge their phones. Also these places seem like where younger people would like to go.

G. Clustering the neighborhoods

We will use the k-means method for clustering. It is a simple but useful unsupervised machine learning algorithm. One of the most important hyper-parameter is K, which represents the number of clusters we want to set.

Based on previous discussion in part F, we can see different neighborhoods indeed have different preferences in terms of the venues that they host. I propose to use the K of 5, which is the same in the lab, but I do believe that it is a reasonable choice.

I think generally, there are 5 types of neighborhoods. The really affluent living neighborhood, the highly commercial neighborhood with lots of workspace and entertainments, the leisure type neighborhoods such beaches and sports grounds, the working class/factory neighborhoods, and a good number of 'average' neighborhoods.

So let's use k-means and merge the venue table with our previous data. And visualize these in maps to see if the result meet with our expectations. Let's explore these in the next section.

IV Results

A. Cluster 1:

Cluster 1 have many outdoor parks and also supermarkets, and also commonly have fast food restaurants. Combing these information with the graphical representation, we can see that these neighborhoods are spread-out and are on the outskirt of New York (far from the city's center). It could be some of the lower income neighborhoods.

B. Cluster 2：

Cluster 2 are characterized by lots of pizza places and fast food restaurants. The difference from cluster 1 is that they are very concentrated and also nearer the centre of the city, in fact one concentration on the top, and the other on the bottom, with Manhattan just sitting in between. Reading from the map, the spots are close to Bronx and Brooklyn.

C. Cluster 3:

Cluster 3 is the most spread-out so far, and has so many more neighbors contained. Examining the result, we see a lot of restaurants, banks, and shops. They all suggest that these could be 'average' neighborhoods where most people live.

D. Cluster 4:

Ignoring a few very isolated spots, cluster 4 is by far the most densely concentrated areas. It is characterized by lots of theatres, restaurants and bars. In fact, looking at the map, this is Manhattan and its surrounding neighborhoods.

E. Cluster 5:

Cluster 5 seem to be all quite close to the beach, like Brighton beach, these locations have restaurants, bed & breakfast. So they appear to be more tourist and resort like neighborhoods.

V Discussions

To remind ourselves of our selection criteria, I will paste them again here:

1. Our potential customers are 20~35 year olds.
2. Potential customers are more likely to use our service, the more time they spend at a venue (thus the more likely that their phone's battery running out).
3. The power bank rental stations need some but not extensive maintenance (assuming a monthly checkup). So they shouldn't be too spread out.
4. The power bank rental stations cost $3,000 each to produce, and we only have a budget to produce a limited amount for pilot testing.
5. Our marketing team would like to be able to go frequently around the rental stations to do promotions and observe how people actually adopt or interact with our service.

Based on previous findings and our criteria, I will first rule out cluster 1 and cluster 3, the reason is they are all too spread-out and not characteristic of younger demographics, who are more likely to adopt new innovations such as our service.

Cluster 5 is also quite spread-out. However, based on its characteristic and tourist-related functions, I think this is has potentials and is worth saving out for later. If our power bank rental business comes out to be successful, then in the second round, we can launch the service here, working with local restaurants and resorts to minimize our own efforts. But at this stage, what we really want is quick feedback, cluster 5 is too difficult for our marketing team to promote and maintain.

Cluster 2 and 4 both seem like promising places. They are full of restaurants and bars, which are places people really would like to spend a whole night in, so there's certainly a lot of potential for needing to charge their phones. These places are also very densely concentrated, so it is easy for our marketing team to go from one place to the next.

VI Recommendation and conclusion

I would recommend our company to start piloting in cluster 4 neighborhoods. And if successful, in the next testing stage, can move onto either one section of cluster 2 first, and then later use the other for cross-checking to see the marketing results are as expected.