Analyzing the best location for a new restaurant

Lakshmi Narasaiah

IBM Data Science Certificate Capstone

Introduction

The restaurant industry is difficult to succeed in. Aside from the operational challenges, there are two external factors that are required to survive:

1) Regular foot traffic
2) An underserved population that they can address

Foot traffic may come from locals and tourists visiting other nearby venues, from the people that live in that neighborhood, or in some cases: a busy work center/industrial area

A well organized restaurateur will know their particular cuisine style, price range, and desired target customers. A target customer in this case can be defined as underserved if there’s a lack of other options that serve this restaurant’s style, price, and theme (ie, an italian restaurant in little Italy will have a hard time differentiating itself)

Business Explanation

The restaurant in mind for this study is a new Hong Kong style cafe, modeled after one of my favorite restaurants in Boston : Double Chin Cafe

This restaurant caters to a younger crowd, mostly in their 20’s and 30’s. Their menu serves fast, cheap, and delicious eats with an emphasis on regularly creating fun dishes.

Their late-night menu begs for them to be the last stop after a long night on the town. The target customer is groups looking for something to eat in between visiting art shows, concerts, bars, and other big social events.

Business Problem

The issue here is, we don’t know where the restaurant should be located. The owners had success with their first restaurant in Chinatown, but the crowding of other places serving similar fare made it extremely difficult to stand out.

Aside from Chinatown, the owners don’t know where else to look. They need a neighborhood with a steady stream of customers visiting other nightlife options, with hopefully not too many other Chinese/Asian fusion restaurants in the area.

Can we use publicly available data to source potential new locations for their restaurant?

Data

To attempt a solution at this problem, we'll be using the Foursquare API to fetch venue information and analyze different neighborhoods.

To fulfill the requirements of the business problem stated, we'll specifically be looking for two types of information:

- Data on the neighborhoods in general, specifically:
    - The categories for each venue in each Brooklyn neighborhood
    - The popularity for each of the target venues, determined by:
        - The total number of tips
        - The number of likes
- Data on venues that might compete with ours, specifically:
    - The relative frequency of all restaurants
    - The relative frequency of specifically Chinese/Asian Restaurants

Data (fetching popular venues in a neighborhood)

In [15]:
```
uniques = venues['Category'].value_counts()
uniques
```

Out[15]:
```
Chinese               9
Bubble Tea            6
Cocktail              4
American              4
Dim Sum               4
Vietnamese            4
Bakery                3
Noodles               3
Hotpot                3
Salon / Barbershop    3
Optical               2
Bar                   2
Spa                   2
Coffee Shop           2
Malay                 2
Asian                 2
Dumplings             2
Ice Cream             2
Sandwiches            2
Boutique              1
Cosmetics             1
```

Data (fetching number of tips)

```
In [29]: VENUE_ID = '4db3374590a0843f295fb69b'
         url = 'https://api.foursquare.com/v2/venues/{}?&client_id={}&client_secret={}&v={}'.format(
             VENUE_ID,
             CLIENT_ID,
             CLIENT_SECRET,
             VERSION,
             )
         request2 = requests.get(url).json()
         attributes = request2['response']['venue']['tips']
         attributes
```


```
Out[29]: {'count': 172,
          'groups': [{'type': 'others',
            'name': 'All tips',
            'count': 172,
            'items': [{'id': '4df167a3b0fb807158b979f8',
              'createdAt': 1307666339,
              'text': 'Big tray chicken. Make sure you ask them to add hand-pulled noodles to it.',
              'type': 'user',
              'canonicalUrl': 'https://foursquare.com/item/4df167a3b0fb807158b979f8',
              'lang': 'en',
```

Methodology

After fetching data for all venues and neighborhoods in Brooklyn, I filtered out the specific venues I was interested in (Restaurants and Nightlife venues). I then picked venues from this list that had a high number of likes and tips (showing high popularity). This list was used to narrow down the neighborhood list to just 19 entries

I took the filtered list, and searched for venues again, this time with a little wider search radius and a higher limit on the # of venues

I took the information from this list of "hot" neighborhoods and pulled together a relative frequency of nightlife venues, restaurants, and then Chinese Restaurants

```
In [568]: popNeighborhoods.fillna(0.0)
Out[568]:
```

	Borough	Neighborhood	Latitude	Longitude	Night Counts	Restaurant Counts	Total Counts	Night Frequency	Eats Frequency	Chinese Restaurants	Chinese Frequency
0	Brooklyn	Boerum Hill	40.685683	-73.983748	7	13	84	0.083333	0.154762	2.0	0.023810
1	Brooklyn	Carroll Gardens	40.680540	-73.994654	5	12	100	0.050000	0.120000	0.0	0.000000
2	Brooklyn	Cobble Hill	40.687920	-73.998561	5	17	100	0.050000	0.170000	2.0	0.020000
3	Brooklyn	Crown Heights	40.670829	-73.943291	2	4	19	0.105263	0.210526	0.0	0.000000
4	Brooklyn	Ditmas Park	40.643675	-73.961013	2	14	50	0.040000	0.280000	3.0	0.060000
5	Brooklyn	Downtown	40.690844	-73.983463	7	15	100	0.070000	0.150000	5.0	0.050000

Methodology

After pulling all the data together in a single dataframe, I ran a clustering analysis to group similar neighborhoods together

The different clusters were analyzed to see which were most favorable for our new restaurant (a lot of nightlife, very few chinese restaurants)

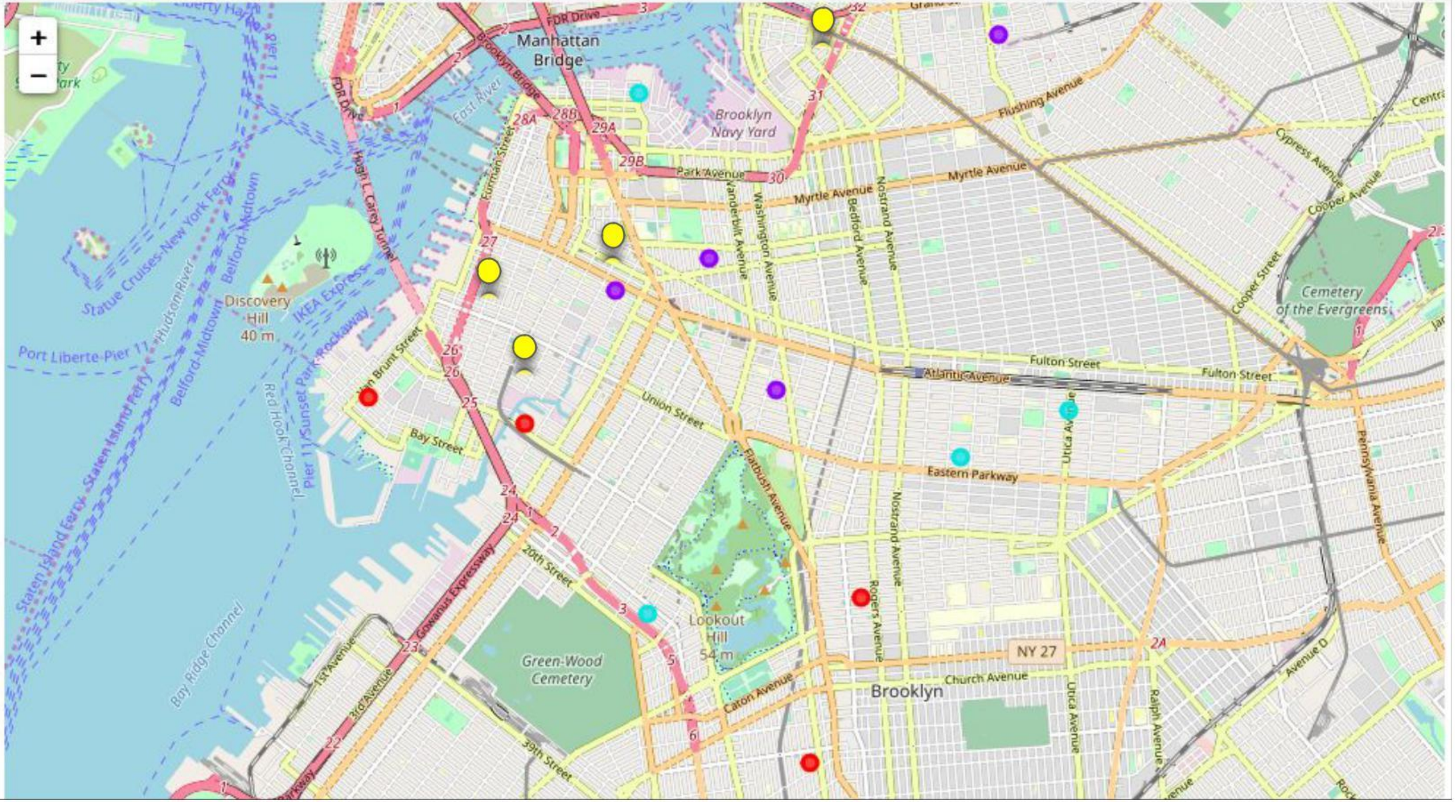
```
In [600]: popNeighborhoods
```

Out[600]:

	Borough	Neighborhood	Latitude	Longitude	Night Counts	Restaurant Counts	Total Counts	Night Frequency	Eats Frequency	Chinese Restaurants	Chinese Frequency	Cluster
0	Brooklyn	Boerum Hill	40.685683	-73.983748	7	13	84	0.083333	0.154762	2.0	0.023810	1
1	Brooklyn	Carroll Gardens	40.680540	-73.994654	5	12	100	0.050000	0.120000	NaN	NaN	3
2	Brooklyn	Cobble Hill	40.687920	-73.998561	5	17	100	0.050000	0.170000	2.0	0.020000	3
3	Brooklyn	Crown Heights	40.670829	-73.943291	2	4	19	0.105263	0.210526	NaN	NaN	2
4	Brooklyn	Ditmas Park	40.643675	-73.961013	2	14	50	0.040000	0.280000	3.0	0.060000	0
5	Brooklyn	Downtown	40.690844	-73.983463	7	15	100	0.070000	0.150000	5.0	0.050000	3
6	Brooklyn	East Williamsburg	40.708492	-73.938858	7	5	71	0.098592	0.070423	NaN	NaN	1
7	Brooklyn	Fort Greene	40.688527	-73.972906	6	20	79	0.075949	0.253165	1.0	0.012658	1

Results

Out[614]:



Results

The cluster shown in light blue are the most ideal spots for our new concept restaurant

This is based on the high number of nightlife venues and the almost complete absence of other Chinese Restaurants

The ideal neighborhoods for our new location are:

- Carroll Gardens
- Cobble Hill
- Downtown
- Greenpoint

Discussion

Basing our popularity decision based on number of likes and tips is biased, since it doesn’t take into account the length of time a place has been open

When you use the explore endpoint in the Foursquare API it returns search results based on your own user preferences. Obviously this doesn’t give perfectly fair data to inform our model

After picking the best neighborhood, there’s a lot more work that needs to be done to actually find the best location. However, this project gives a nice start to the process and narrowed down a very long list (from 70 to 4 choices)

References

For this project, I used some notes on clustering from the IBM Data Science Capstone labs

The Foursquare API documentation was also used to pull together this project

Acknowledgements

Thanks to the Coursera team for running the website

A special thanks as well to IBM for putting together the course materials