Estimating the Best New Venue

Yegang Wu

July 8, 2020

1. Introduction

1.1 Background

It is never easy to start a new business, especially in New York – a highly competitive area. In order to start a new business, there are two things we need to determine beforehand: what business should I do and where should the business be located at. In this paper, we will focus on the first question: assume I would want to start a new business in a given neighborhood, say, Newport (or any other neighborhood), but I am not sure about what is the best business in this neighborhood that is most needed. For example, should I open a restaurant, or should I open a pharmacy, or maybe I should open a coffee bar? Therefore, the goal of this study is to identify the best new businesses, which are measured by Foursquare venue categories, to open in this neighborhood (or any other neighborhood).

1.2 Problem

As is described above, the problem is to find the best new businesses (measured by Foursquare venue categories) to open in a given neighborhood. It could be a restaurant, a gym, a bar, or anything. The goal is to find best venue category using a machine learning algorithm. In this study, this problem is solved by first running an unsupervised clustering analysis, then perform a sensitivity test on the intra-cluster similarity.

1.3 Interest

Clearly this would be an interesting topic to many investors and entrepreneurs. It would also be a useful information to the local government, as it can help them with better future city design.

2. Data acquisition and cleaning

2.1 Data sources

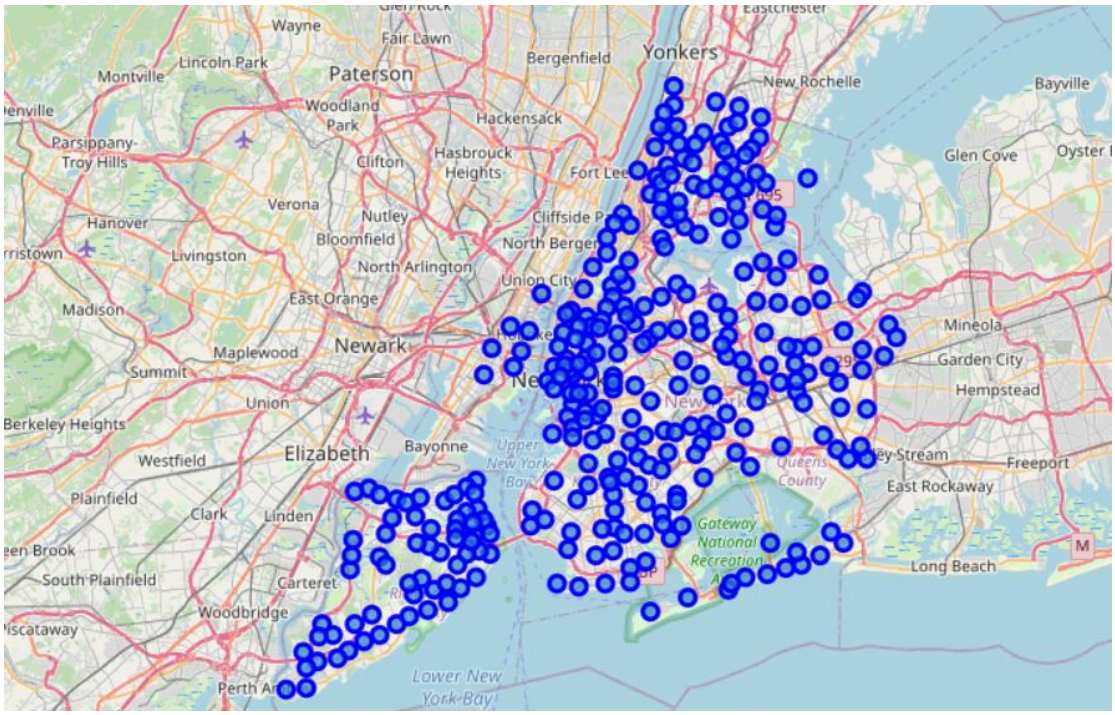
There are two data sources for this project:

1. Venue data for New York City from Foursquare. In addition to the standard venue data that can be derived by calling “explore” API request, we also fetch the venue categories definition data from Foursquare. The venue categories definition table allows us to map the original venue data returned from the “explore” API request to higher group levels.

   For example, the original venue could be “Hunan restaurant”. This is actually a level IV category. The corresponding Level III category is “Chinese restaurant”, and similarly, the level II category is “Asian restaurant”, and Level I category is “Food”.
2. New York Neighborhood Latitude and Longitude data from Class Lab - Segmenting and Clustering Neighborhoods in New York City.

We derived the venue data of neighborhoods in New York City from Foursquare using “explore” API requests with latitude and longitude data from the class lab listed above.

Neighborhoods included in our analysis:



2.2 Data cleaning and feature selection

First, we will merge the venue category table with our venue data table so that we can retrieve the higher-level venue information for each row of the original venue data. This is particularly useful as the original venue data returned by "explore" API calls are too granular. Mapping to higher level would help us to reduce the dimensionality of our data set and therefore improve the robustness of our study.

Example of original venue data:

```
nearby_venues.head()
```

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue Name	Venue Latitude	Venue Longitude	Venue Category
0	Manhattan_Marble Hill	40.876551	-73.91066	Arturo's	40.874412	-73.910271	Pizza Place
1	Manhattan_Marble Hill	40.876551	-73.91066	Bikram Yoga	40.876844	-73.906204	Yoga Studio
2	Manhattan_Marble Hill	40.876551	-73.91066	Tibbett Diner	40.880404	-73.908937	Diner
3	Manhattan_Marble Hill	40.876551	-73.91066	Starbucks	40.877531	-73.905582	Coffee Shop
4	Manhattan_Marble Hill	40.876551	-73.91066	Dunkin'	40.877136	-73.906666	Donut Shop

Example of the new venue data with higher level information:

```
nearby_venues_full.head()
```

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue Name	Venue Latitude	Venue Longitude	Venue Category	Level I	Level II	Level III	Level IV
0	Manhattan_Marble Hill	40.876551	-73.91066	Arturo's	40.874412	-73.910271	Pizza Place	Food	Pizza Place	Pizza Place	Pizza Place
1	Manhattan_Marble Hill	40.876551	-73.91066	Bikram Yoga	40.876844	-73.906204	Yoga Studio	Outdoors & Recreation	Athletics & Sports	Gym / Fitness Center	Yoga Studio
2	Manhattan_Marble Hill	40.876551	-73.91066	Tibbett Diner	40.880404	-73.908937	Diner	Food	Diner	Diner	Diner
3	Manhattan_Marble Hill	40.876551	-73.91066	Starbucks	40.877531	-73.905582	Coffee Shop	Food	Coffee Shop	Coffee Shop	Coffee Shop
4	Manhattan_Marble Hill	40.876551	-73.91066	Dunkin'	40.877136	-73.906666	Donut Shop	Food	Donut Shop	Donut Shop	Donut Shop

Next, consistent with the data reformatting method in the class lab -- Segmenting and Clustering Neighborhoods in New York City, we will reformat the venue data for each neighborhood using one hot encoding, then aggregate the venue data within each of the neighborhood by taking average. By doing that, we can achieve a table with each row representing the distribution of different venues in a neighborhood.

Regarding on feature selection, we are focusing on the above distribution numbers. Note that in this step, we are working on Level II venue category. Level III and Level IV venue categories are too granular. Moving to Level II can help us to significantly reduce the number of features in the data, and therefore improve the robustness of our model.

Example of the data post-reformatting:

	Neighborhood	Adult Boutique	Afghan Restaurant	African Restaurant	American Restaurant	Antique Shop	Arcade	Art Gallery	Arts & Crafts Store	Arts & Entertainment	Asian Restaurant	Athletics & Sports	A
0	Bronx_Allerton	0.0	0.0	0.0	0.034483	0.0	0.00	0.0	0.0	0.0	0.068966	0.034483	
1	Bronx_Baychester	0.0	0.0	0.0	0.000000	0.0	0.05	0.0	0.0	0.0	0.000000	0.000000	
2	Bronx_Bedford Park	0.0	0.0	0.0	0.000000	0.0	0.00	0.0	0.0	0.0	0.062500	0.062500	
3	Bronx_Belmont	0.0	0.0	0.0	0.010526	0.0	0.00	0.0	0.0	0.0	0.010526	0.010526	
4	Bronx_Bronxdale	0.0	0.0	0.0	0.000000	0.0	0.00	0.0	0.0	0.0	0.071429	0.071429	

The above matrix (after dropping the "Neighborhood" column) will be our input data for our analysis today. The matrix is 311 x 286, which means there are 311 neighborhoods and 286 different level II venue categories. Again, each row in the matrix corresponds to the distribution of level II venue categories within a neighborhood

3. Methodology

3.1 Exploratory data analysis – Neighborhood Level

If we would want to start a new business in any given neighborhood, then the first thing we should take a look is the current most popular venues in that neighborhood.

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Bronx_Allerton	Pizza Place	Deli / Bodega	Food & Drink Shop	Department Store	Asian Restaurant
1	Bronx_Baychester	Donut Shop	Bank	Convenience Store	Sandwich Place	Electronics Store
2	Bronx_Bedford Park	Diner	Mexican Restaurant	Pizza Place	Deli / Bodega	Sandwich Place
3	Bronx_Belmont	Italian Restaurant	Food & Drink Shop	Pizza Place	Deli / Bodega	Bakery
4	Bronx_Bronxdale	Food & Drink Shop	Italian Restaurant	Paper / Office Supplies Store	Bank	Mexican Restaurant
5	Bronx_Castle Hill	Southern / Soul Food Restaurant	Bus Station	Market	Pharmacy	Bank
6	Bronx_City Island	Spanish Restaurant	Seafood Restaurant	Thrift / Vintage Store	Jewelry Store	Food & Drink Shop
7	Bronx_Claremont Village	Food & Drink Shop	Asian Restaurant	Bus Station	Bakery	Pizza Place
8	Bronx_Clason Point	Park	Pool	Food & Drink Shop	Boat or Ferry	Latin American Restaurant
9	Bronx_Co-op City	Bus Station	Food & Drink Shop	Discount Store	Salon / Barbershop	Bagel Shop

Let's take Allerton as an example, you can verify from the table above, the most popular venue is pizza place. So, regarding on our question of which business/venue should we open, a quick recommendation could be: Pizza it is! Let's just open a new pizza place in Allerton.

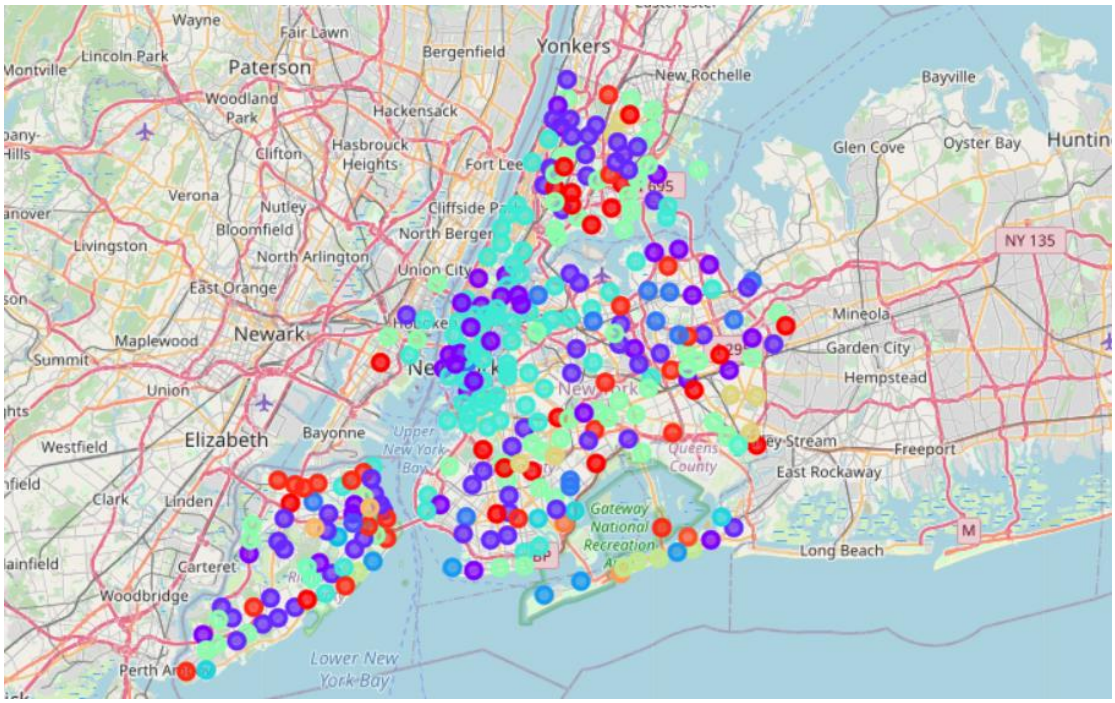
Of course, this is just a naïve answer, as it is quite possible that the pizza places in

Allerton are already under heavy competition. In such cases, it would be hard for a new pizza place to enter into the business.

3.2 Exploratory data analysis – Cluster Level

To perform a more comprehensive analysis, let's take a look at other neighborhoods that are similar to Allerton so we would have more data to analyze. In order to find the other neighborhoods that are similar to Allerton, we will perform an unsupervised clustering analysis using the K-Means algorithm. The input data are constructed in the data section above, and as a starting point, given we have 300+ neighborhoods, let's first randomly set K = 20.

K Means Outcome with K = 20



In particular, there are 60 neighborhoods that are grouped into the same cluster as Allerton.

Examples of some of the neighborhoods in the same cluster:

Borough	Neighborhood	Latitude	Longitude	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	Cl L
Jersey City	Jersey City_JC07307	40.748001	-74.049430	Pizza Place	Spanish Restaurant	Sandwich Place	Asian Restaurant	Donut Shop	
Bronx	Bronx_Kingsbridge	40.881687	-73.902818	Pizza Place	Bar	Food & Drink Shop	Latin American Restaurant	Mexican Restaurant	
Bronx	Bronx_Norwood	40.877224	-73.879391	Pizza Place	Bank	Park	Food & Drink Shop	American Restaurant	
Bronx	Bronx_Pelham Parkway	40.857413	-73.854756	Dessert Shop	Italian Restaurant	Pizza Place	Deli / Bodega	Asian Restaurant	
Bronx	Bronx_Bedford Park	40.870185	-73.885512	Diner	Mexican Restaurant	Pizza Place	Deli / Bodega	Sandwich Place	
Bronx	Bronx_Morris Heights	40.847898	-73.919672	Spanish Restaurant	Pizza Place	Bus Station	Bank	Food & Drink Shop	

Now if we aggregate the data by running value_count() in each column, we can achieve the following table:

The most popular venues in the cluster:

	Venue	Most Popular	2nd Popular	3rd Popular
0	Pizza Place	15.0	13.0	7.0
1	Asian Restaurant	11.0	8.0	7.0
2	Food & Drink Shop	11.0	7.0	8.0
3	Italian Restaurant	5.0	4.0	2.0
4	Deli / Bodega	5.0	3.0	1.0
5	Bus Station	2.0	2.0	1.0
6	Bank	2.0	2.0	2.0
7	Spanish Restaurant	1.0	1.0	0.0
8	Coffee Shop	1.0	0.0	0.0
9	Professional & Other Places	1.0	0.0	0.0
10	Pharmacy	1.0	6.0	2.0
11	Diner	1.0	1.0	1.0
12	Bagel Shop	1.0	0.0	0.0
13	Dessert Shop	1.0	1.0	2.0

From this table, you can see that pizza place is indeed the most popular venue category with in the neighborhoods that are similar to Allerton. Therefore, it could indeed be a good choice for us if we want to start a business in Allerton. In addition, Asian Restaurant and Food & Drink Shop are also good choices, as their ranks are actually quite close the pizza places.

3.3 Exploratory data analysis – Intra-cluster Similarity

Now if we want to go one step deeper, we can view our problem in a different way: what new business/venues does Allerton need? In other words, if we compare Allerton to other neighborhoods in the same clusters, can we tell what is missing in Allerton? Clearly, starting a business in which Allerton is missing currently would be a great choice.

To solve this problem, we can actually take a look at intra-cluster similarity of the cluster and Allerton belongs to. The intra-cluster similarity tells us how similar the neighborhoods are within the same cluster.

Now we will perform a sensitivity analysis on each of the venue category in Allerton by a positive shock of +1 (assuming we will start a venue).

For example, for the sensitivity on American restaurants, we will just add one more American restaurant to the count of American restaurants. That will slightly change the distribution of the venue categories in Allerton.

Example of original distribution of venue categories:

```
: data_with_in_group.iloc[ind_Neighborhood,3:10]
```

```
: American Restaurant      0.026316
  Antique Shop             0.000000
  Arcade                   0.000000
  Art Gallery              0.000000
  Arts & Crafts Store      0.000000
  Arts & Entertainment     0.000000
  Asian Restaurant         0.052632
  Name: 127, dtype: float64
```

Example of post-shock distribution of venue categories:

```
American Restaurant      0.051939
Antique Shop             0.000000
Arcade                   0.000000
Art Gallery              0.000000
Arts & Crafts Store      0.000000
Arts & Entertainment     0.000000
Asian Restaurant         0.051247
Name: 127, dtype: float64
```

Then we will analyze the impact on the intra-cluster similarity resulted by this shock. An increment in the intra-cluster similarity means that by starting a new business in American restaurants, Allerton is getting more similar to other neighborhoods in the same cluster, which means American restaurants is what Allerton needs.

We will repeat the above shocking-testing process on all of the venue categories, and the one result in the largest increment of intra-cluster similarity is the venue category that Allerton needs the most, which is also the one that our study would propose.

We will calculate the intra-cluster similarity using the distance matrix. Note that they are inversely correlated: higher distance means lower similarity.

Output:

	index	Venue	Distance	Delta in Distance
0	21	Bank	15.549499	-0.372283
1	141	Italian Restaurant	15.576119	-0.345663
2	20	Bakery	15.632218	-0.289564
3	230	Sandwich Place	15.635771	-0.286011
4	163	Mexican Restaurant	15.666826	-0.254956
5	19	Bagel Shop	15.676339	-0.245443
6	22	Bar	15.681945	-0.239837
7	78	Diner	15.712208	-0.209574
8	59	Coffee Shop	15.719045	-0.202737
9	46	Café	15.727091	-0.194691
10	191	Park	15.733023	-0.188759

According to the output table above, the top three best venue categories are: Bank, Italian Restaurant, and Bakery, as they will results in the largest reduction in the intra-

cluster distance, which corresponds to the largest increment in the intra-cluster similarity. Intuitively, if you compare Allerton to its similar neighborhoods, which corresponds to the other neighborhoods in the same cluster, those are the venues that are most needed in Allerton.

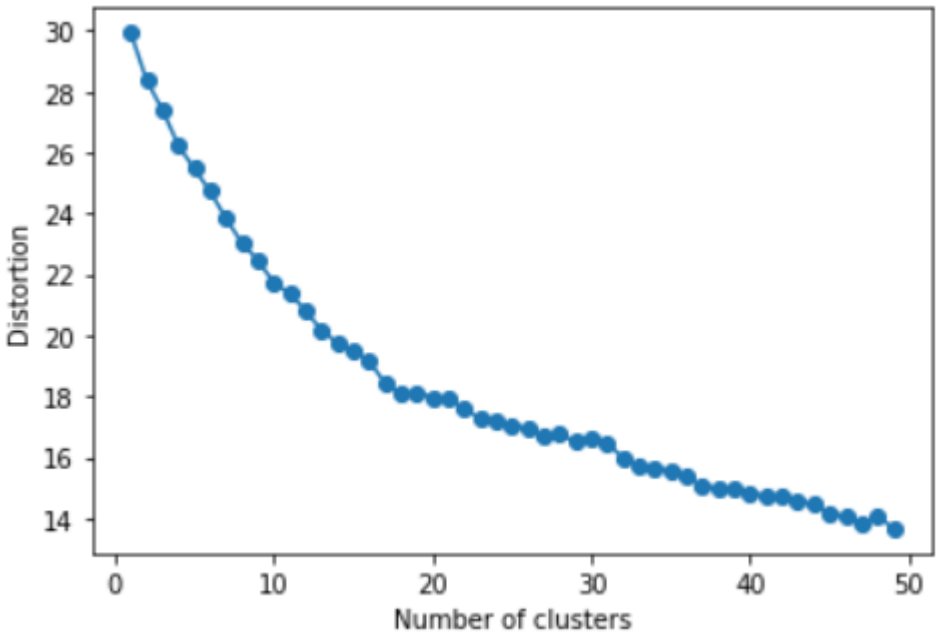
Now let's take a look at our previous recommendation, pizza places:

281	202	Pizza Place	16.062165	0.140383

We got very different answer. The rank of pizza place is 281, and it will increase the intra-cluster distance. That implies there are already too many pizza places in Allerton, which means it would not be a good choice to start a new pizza place in Allerton. Note that this completely denied our previous naïve proposal.

3.4 Exploratory data analysis – Different choices of K

The above result tables were generated using K means with K = 20. 20 is a random choice, so let's take a look at the algorithm performance with different K.



Unfortunately, we were unable to find any clear elbow point. In such scenario, in order to improve the robustness of our model, we will repeat the analysis in 3.3 for all choices of K, then aggregate the results.

4. Results

We repeat the process of 3.3 with different choices of K in K Means algorithm.

Outputs:

	K	Neiborhoods in the Group	Best Business	2nd Best Business	3rd Best Business
0	5	135	Bank	Sandwich Place	Bar
1	6	69	Food & Drink Shop	Caribbean Restaurant	Bakery
2	7	157	Bank	Sandwich Place	Bakery
3	8	101	Bank	Sandwich Place	Beach
4	9	72	Food & Drink Shop	Caribbean Restaurant	Bakery
5	10	72	Bank	Bus Stop	Sandwich Place
6	11	116	Bank	Sandwich Place	Bakery
7	12	45	Bank	Bar	Sandwich Place
8	13	67	Bank	Sandwich Place	Italian Restaurant
9	14	46	Bank	Bar	Sandwich Place
10	15	97	Bank	Sandwich Place	Park
11	16	46	Bank	Italian Restaurant	Bar
12	17	61	Bank	Caribbean Restaurant	Clothing Store
13	18	36	Bank	Bar	Sandwich Place
14	19	79	Bank	Sandwich Place	Mexican Restaurant
15	20	61	Bank	Italian Restaurant	Bakery
16	21	38	Bank	Sandwich Place	Bar
17	22	41	Food & Drink Shop	Bakery	Sandwich Place
18	23	30	Bank	Bar	Pizza Place
19	24	23	Pizza Place	Bar	Coffee Shop
20	25	43	Bank	Latin American Restaurant	Sandwich Place
21	26	25	Bar	Coffee Shop	Bakery
22	27	48	Bank	Sandwich Place	Bar
23	28	43	Bank	Bakery	Bar
24	29	21	Deli / Bodega	Park	Bank
25	30	22	Bank	Pizza Place	Bagel Shop

You can see the outcomes (most recommended venue categories) are overall quite stable with different Ks.

Now we aggregate the results above by summing over different Ks. Let's also define a score for each venue category:

$$Score_{venue(i)} =$$

$$\sum_{k} \{[3 * I(if\ venue(i)\ is\ the\ most\ recommended\ venue)]$$
$$+ [2 * I(if\ venue(i)\ is\ the\ 2nd\ most\ recommended\ venue)]$$
$$+ [1 * I(if\ venue(i)\ is\ the\ 3rd\ most\ recommended\ venue)]\}$$

Final output:

	Venue	Best	2nd	3rd	Final Score
0	Bank	25.0	1.0	2.0	79.0
1	Sandwich Place	0.0	10.0	8.0	28.0
2	Bar	3.0	5.0	6.0	25.0
3	Bakery	0.0	5.0	6.0	16.0

The top 4 best choices are Bank, Sandwich Place, Bar and Bakery.

5. Discussion

According to the summary table above, we found the top 4 best choices for new business in Allerton are Bank, Sandwich Place, Bar and Bakery. In particular, Bank is the top one choice, with its score much higher than all other venue categories. There is currently no bank in Allerton, while there are banks in most of its similar neighborhoods. Therefore, bank is indeed one of the most needed business in Allerton – the output of our model is consistent with intuition.

Of course, opening a bank is not a practical choice to most people. Therefore, the top three best choices of new businesses in Allerton would be Sandwich Place, Bar and Bakery.

However, the fact that bank is the most needed business in Allerton is quite useful information -- it might be a good choice for banks to open a branch in Allerton

6. Conclusions

Based on our study above, we found the top 4 best choices for new business in Allerton are Bank, Sandwich Place, Bar and Bakery. In our study, we used Allerton as an example, but the same algorithm can be applied to any other neighborhood.