**Estimating the Best New Venue**

Yegang Wu

July 8, 2020

1. **Introduction**
   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. 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. 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.

1. **Data acquisition and cleaning**
   1. Data sources

There are two data sources for this project:

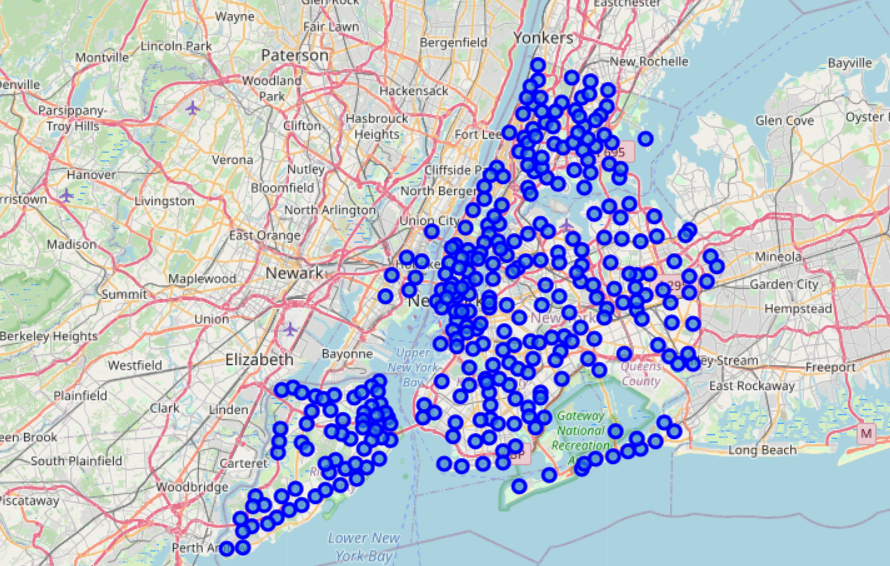
1. Venue data for New York City from Foursquare. In addition to the standard venue data that can be derived by calling “explore” API request, we also fetch the venue categories definition data from Foursquare. The venue categories definition table allows us to map the original venue data returned from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

1. 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.

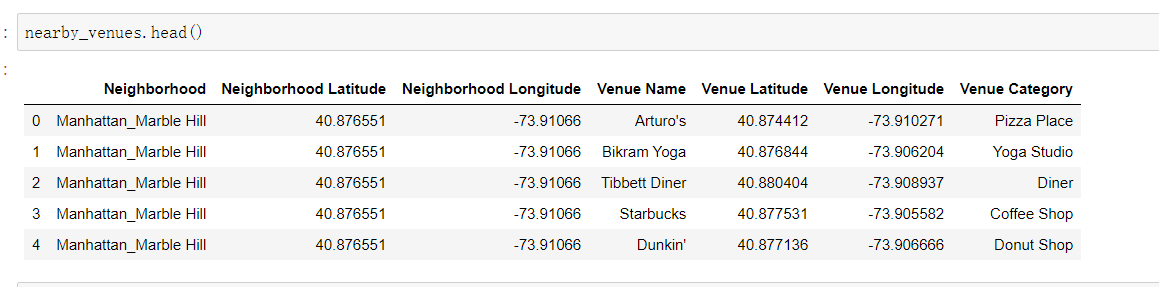
Neighborhoods included in our analysis:



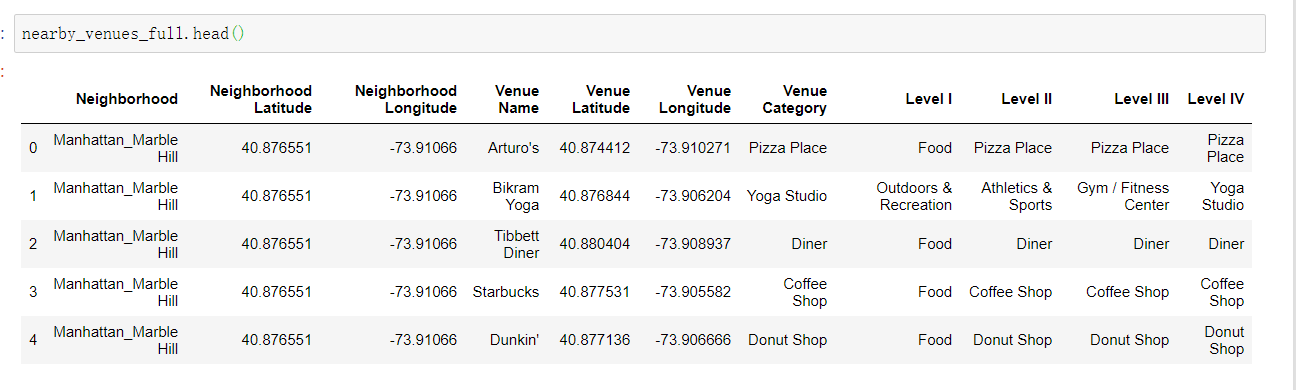
* 1. Data cleaning and feature selection

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:



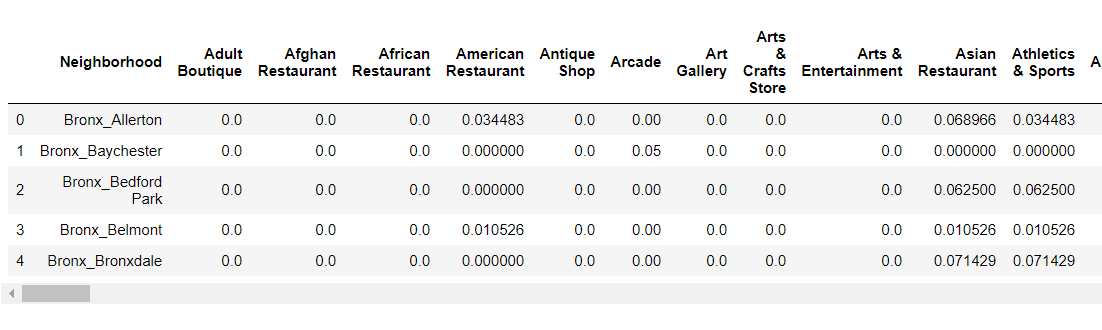
Example of the new venue data with higher level information:



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:



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

1. **Methodology**
   1. Exploratory data analysis

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.

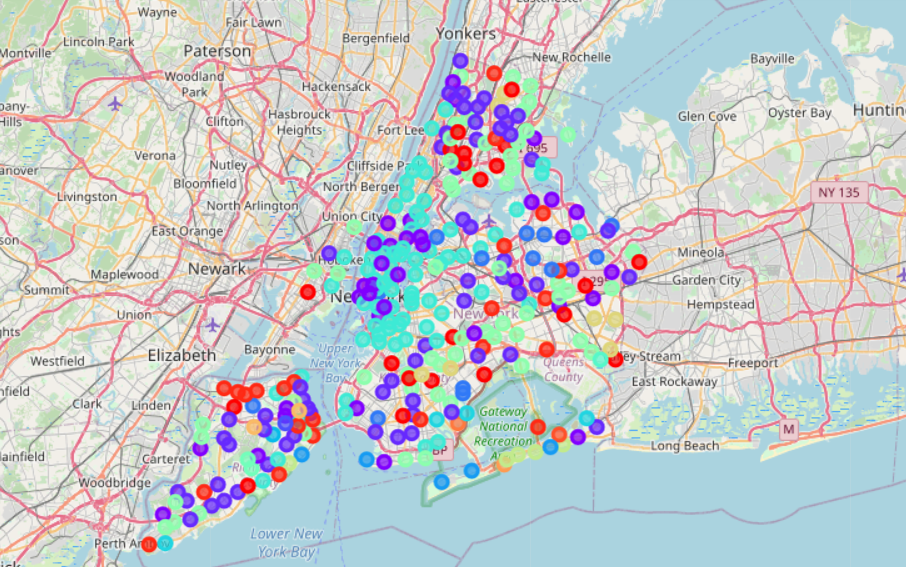


Let’s take Allerton as an example, you can verify from the table above, the most popular venue is pizza place. So, regarding on our question of which business/venue should we open, a quick recommendation could be: Pizza it is! Let’s just open a new pizza place in Allerton.

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.

Therefore, 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 sector 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]:



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:



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. So 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.

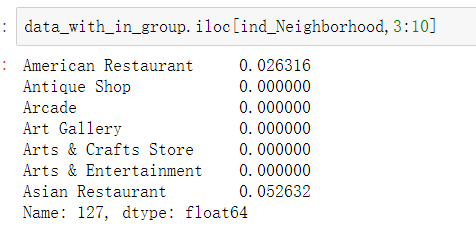
Now if we want to go one step deeper, we can view our problem in a different way: what new business/venues does Allerton need? In other words, if we compare Allerton to other neighborhoods in the same clusters, can we tell what is missing in Allerton? Clearly, starting a business in which Allerton is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

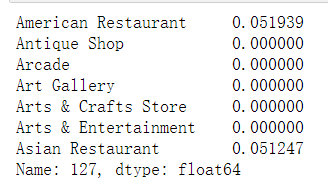
Now we will perform a sensitivity analysis on each of the venue category in Allerton by a positive shock of +1 (assuming we will start a venue).

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:



Example of post-shock distribution of venue categories:



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.

1. **Results**
2. **Discussion**
3. **Conclusions**