Locating an optimal place for business ventures in Manhattan, NY

Shrey Sharma

July 17, 2020

1. Introduction

1.1 Background

From majestic mountains to vineyards to major cities the state of New York has just about everything. Hundreds of corporations make New York City their home from advertising to finance to insurance. The city is a melting pot of both industries and businesses. New York City is a major focal point for culture and arts. The city is also a major financial hub and continues to be the global headquarters for finance, entertainment, media, telecommunications and manufacturing. Manhattan has one of the largest business districts in the world.

1.2 Problem

In this project, we will try to look for an optimal location for a legitimate business venture. This report will be targeted to stakeholders planning to start their business or interested in setting up a corporate office in Manhattan, New York**.**

Since there are a lot of places in Manhattan to consider for, while planning to set up a business, we will try to find out areas where businesses are operating legally as authorized by the Department of Consumer Affairs (DCA)**.** We will also take into consideration the count of venue likes reviewed by Foursquare users. This will provide an insight into the attractiveness of areas in the borough and so will aid the search.

2. Data acquisition and cleaning

2.1 Data sources

- Businesses/individuals holding a DCA license so that they may legally operate in Manhattan will be obtained using data provided by the Department of Consumer Affairs (DCA) on https://data.cityofnewyork.us/Business/Legally-Operating-Businesses/w7w3-xahh
- Data involving venues in and around Manhattan borough will be obtained using Foursquare API.
- Boundaries for NYC zip codes will be extracted using https://github.com/fedhere/PUI2015_EC/blob/master/mam1612_EC/nyc-zip-code-tabulation-areas-polygons.geojson data.

2.2 Data cleaning

Data downloaded or scraped from multiple sources were combined into adjacent tables. There were a lot of missing values in the legally operating businesses in New York City data, which were removed later as we are mostly interested to group businesses as per the industry category.

Firstly, venues data were derived from foursquare API using explore and venue/likes endpoints. This data was in a json file format, which I converted into a pandas data frame for convenience. For which I used reverse geocoding.

Further, to get the popularity of venues based on user likes count, other set of data was also imported using foursquare API that basically gave user likes for each venue id derived earlier. These data frames were merged into one.

Later on, information about legally operating businesses in New York City was obtained from open data source provided by the DCA. I narrowed down the search for DCA-issued license holders to only Manhattan borough among all other boroughs included in the initial data. To check for areas per zip code where businesses/individuals operate in a legit way, I accessed the boundaries data for NYC from GitHub for visualization using the choropleth map. The active/inactive license status was transformed- using one hot encoding, into 1/0 labels respectively.

Finally, I eliminated postal areas outside Manhattan and concatenated license status of each given business with the zip code data derived from geojson file.

2.3 Feature selection

After cleaning the venues data, there were about 30 derived venues returned by explore foursquare endpoint with 10 features. Among these features, I have selected 5 relevant features which includes venue id, name, category, latitude & longitude. Then, using venue likes data per venue id as per foursquare users, I merged both of these data sets into one with total 6 features and 30 samples.

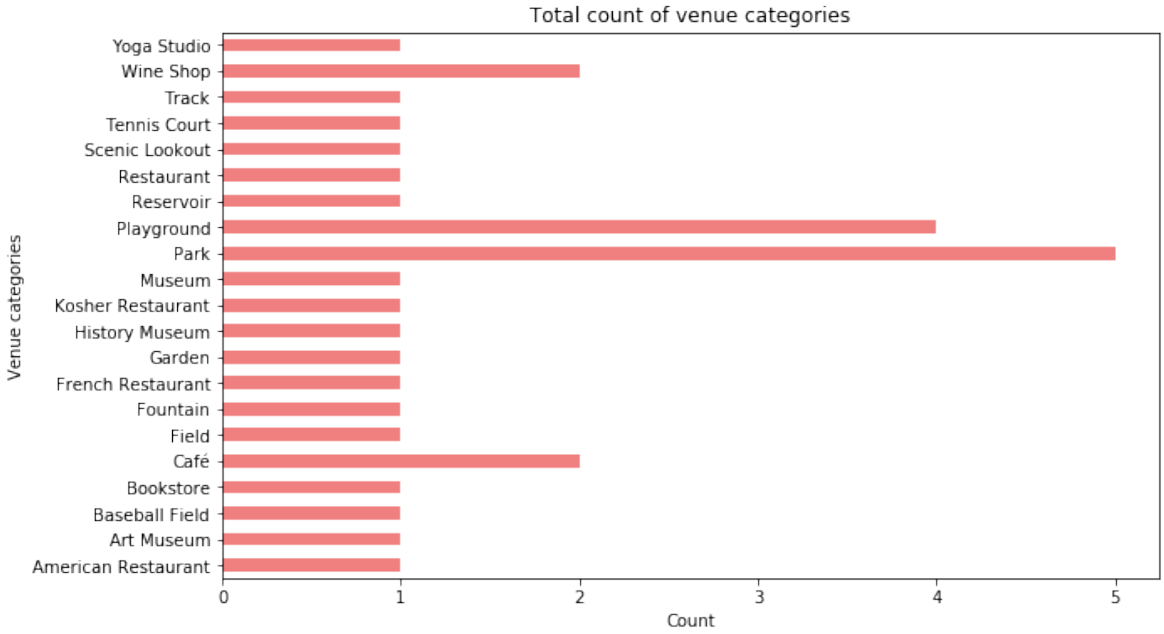
After cleaning the legally operating businesses data, which earlier had 204K samples and 27 features of whole NYC, I filtered the samples and sliced only those included in Manhattan borough which reduced the data instances to 28K. Among various given features, I have chosen 6 relevant features which includes business name, address, zip code, latitude, longitude and license status. Moreover, data from geojson file about NYC boundaries distinguished as per zip codes was merged with businesses data which consist of zip codes and the corresponding license status.

There are total 3 data frames that I have used in this project for proper analysis.

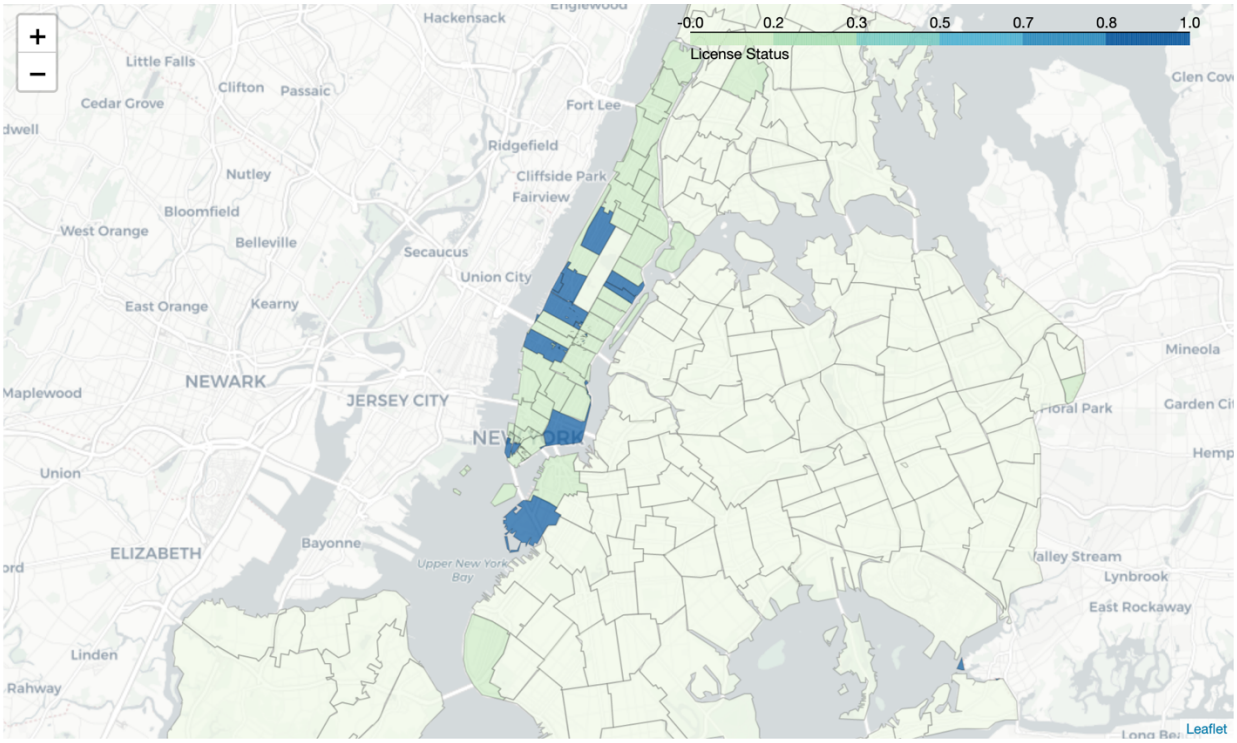
3. Exploratory data analysis

3.1 Venues category

From the data obtained about venues, we can see that there are different categories of venues like parks, playgrounds, wine shops & cafes, and so on. Categorization of venues popular among visitors might count as one important factor for stakeholders interested to know more about places that attracts visitors and its completely on the basis of user reviews which makes it pretty obvious that the surrounding have visitors quite often. It basically makes one familiar with the surroundings.



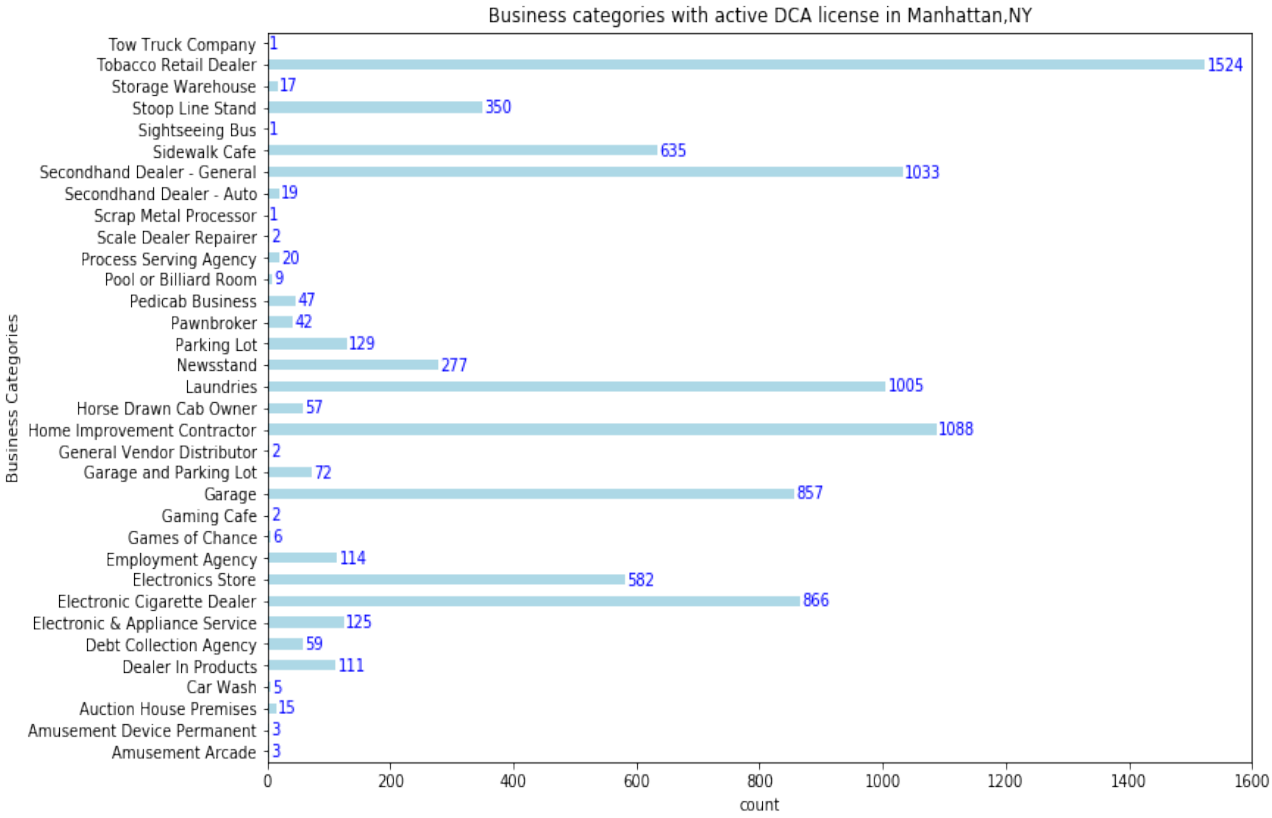
3.2 Businesses license status in Manhattan per postal code



The license status was converted to one hot labels where 0: inactive status and 1: active status. From the above map certain areas can be observed that hold active license status in Uptown, Midtown and Downtown region. Since most of the DCA-approved businesses are located in Uptown Manhattan so I have kept my focus mostly on these areas.

3.3 Business categories with active DCA-issued license

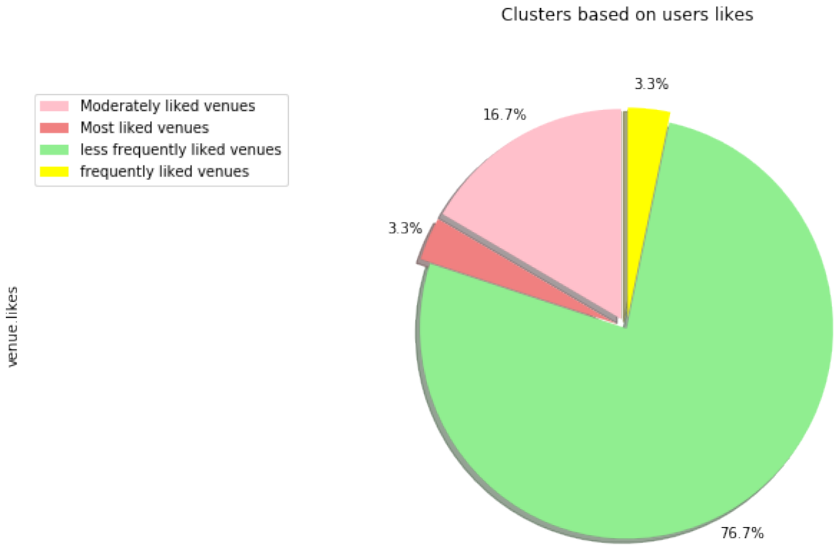
There were about 28 K ongoing businesses in Manhattan as per the data and to plot each business location in maps makes it a bit complicated. Also, we don't really need information regarding each one of them rather we try to generalize our results as per industry category which will give a clearer idea to the stakeholders about the preferred locality.
The plot below shows the total count of business categories in Manhattan.



4. Clustering

4.1 K-Means clustering on venues data based on user likes

Dividing the whole dataset into clusters increases interpretability and will help us to distinguish among venues' attractiveness for this dataset. We will randomly select the value of K as 4. That is, for further analysis we have 4 categories of venues based on user reviews namely 'Most liked venues', 'Moderately liked venues', 'Frequently liked venues' and 'Less frequently liked venues'. Proportion of each cluster can be observed as follows-



4.2 Marker clustering for businesses data

We will cluster businesses according to industry type using marker clustering which combine markers of close proximity into clusters and simplify the display of markers on the map. This allows us to get locations of businesses in the surrounding and basically displays the count of legal operating businesses on the map, which again aid the search for an optimal location. When we zoom out all the way, all markers are grouped into one cluster, the global cluster, of 9057 markers or businesses, which is the total number of businesses with active DCA approved license in our data frame. Once we start zooming in, the global cluster will start breaking up into smaller clusters. Zooming in all the way will result in individual markers.

Note: purple marker -> **"Moderately liked venues"**, pink marker -> **"Frequently liked venues"**, black marker -> **"Most liked venues"** and blue marker -> **"Less frequently liked venues"**
Also, the number on a cluster of businesses indicates how many markers it contains and the red markers with information icon (**i**) displays the label for industry type.

5. Results

The analysis shows that there is a great number of places in Manhattan to choose for business ventures. Areas in **Uptown, Midtown and Downtown** are shown, taking into consideration their license status as issued by DCA. The choropleth map takes into account all the areas (as per the zip codes) though, we will keep our focus on areas in Uptown Manhattan since the API derived venues are highly concentrated over there. As we can observe there are comparatively more pockets of businesses/individuals, with legal authorization, operating fairly close to **Upper east side**. Dense gathering was detected in between **east 89th Street and east 98th Street**, so we focused our attention to areas around these streets along **Madison Avenue, 5th Avenue and Park Avenue** with majority of **Moderately liked venues** and **Less Frequently liked venues** as per Foursquare user's review. Another locality was identified as potentially interesting (on **Columbus Avenue between West 97th Street and West 100th Street**). Anyways, we narrowed down the search to **Upper east side** only, which offer a combination of popularity among visitors, strong socio-economic dynamics *and* a number of pockets of DCA-issued active license holding businesses.

5. Conclusion

Purpose of this project was to identify areas in Manhattan with more popularity among visitors and legitimacy of businesses around different postal areas in Manhattan borough in order to facilitate stakeholders in narrowing down the search for optimal location for a business setup. By gathering information about venues and merging it with the user likes distribution from Foursquare data we have first identified certain locations that are distinguished as per user reviews and venue categories which justifies further analysis, and then generated extensive collection of locations- distributed on the basis of zip code, which satisfy some basic requirements to run a business. Clustering of those locations was then performed in order to create major marks of interest (containing total number of businesses that satisfy legal conditions and categorized by the industry type) , addresses of API derived venues were pinpointed which is to be used as starting points for final exploration by stakeholders.
Final decision on optimal location will be made by stakeholders based on specific characteristics of neighborhoods and locations in every recommended place, taking into consideration additional factors like pricing, commutation, real estate availability, social and economic dynamics of every neighborhood etc.

6. Further directions

This, of course, does not imply that these are the only factors to be considered while searching for an optimal location for a business set up or starting a new branch office. Purpose of this analysis was to only provide information about areas in Manhattan (located in Upper east side) that might fit stakeholders' search criteria- it is entirely possible that there might be a very good reason to not choose these areas based on the high land/buildings prices, which would make them unsuitable for a startup regardless of popularity among visitors in the area. Other than that, I think more data related to venues can be retrieved from other APIs (in this case, it was limited to 30 samples). Recommended areas should therefore be considered only as a starting point for more detailed analysis which could eventually result in location which has not only attractiveness of venues and legal operating businesses around, but also other factors taken into account and all other relevant conditions met.