An insight on two neighborhood types in New York City based on their most prevalent venue categories

Introduction and business problem

The 2020 Global Cities index firmly positions New York City (NYC) as the top global metropolis, remaining unchanged in this position from the previous year ranking (source: https://www.kearney.com/global-cities/2020). Therefore, as a global city, NYC is a prime location for new startup businesses. However, lease space in global cities is limited and expensive, so prospective business owners should use quality data to inform their venue-opening decisions according to their needs. This report attempts to provide guidance by offering a segmentation of the neighborhoods in NYC based on their venue composition. A prospective entrepreneur should leverage this information to make a business decision that optimizes resources.

The present document is thus specifically meant for entrepreneurs looking to open a business in the city of New York. The data analyses done here address the question: Which type of neighborhood should I choose to open my business based on the neighborhood features?

In order to provide a recommendation, the data analyses offer a machine learning-based segmentation of neighborhoods into two main categories based on their most prevalent types of venues and services. Information on said venues and services was collected for every neighborhood of NYC using the Foursquare service. Readers are thus encouraged to use this information to inform their buying or leasing decisions.

Please note that an analysis of the property prices per neighborhood is not in the scope of this document. It is expected that, after identifying their preferred type of neighborhood based on venue composition, prospective business owners should seek additional data sources to identify their most cost-effective solution.

Data

The data used on this project comes from two main sources:

1. A dataset that offers a breakdown of boroughs and neighborhoods of NYC, with coordinates information (see attached Jupyter notebook for the corresponding URL). This dataset is offered by IBM and was used previously on the capstone course. The main purpose of this dataset is to categorize NYC into its component neighborhoods in order to access venue data for each of these locations.

2. The second source of data consists of information on venues obtained through Foursquare (https://foursquare.com) using the site's API available through Python's requests library. All neighborhoods in NYC were queried by providing latitude and longitude information for each, and 100 venues on a 500 m radius per neighborhood were fetched.

Methodology

Neighborhood-level data was first obtained through a dataset made available over the internet by IBM in json format. A pandas dataframe was constructed from this json file, containing the following variables: borough, neighborhood, latitude, and longitude. Then the Nominatim Python package was used to obtain the coordinates of NYC, and folium was employed to generate a map of the city and to assign the location of each neighborhood into this map (Figure 1).

After generating the neighborhood map, venues for each of those neighborhoods were fetched from Foursquare using the /venues/explore parameter . For each neighborhood, 100 venues in a 500 m radius were acquired. A dataframe was constructed that correlates each neighborhood with the average number of venues on every category for each of said neighborhoods. Data were sorted so that for every neighborhood the top 10 most common venue categories were displayed.

The aforementioned venue categories were used to train a K-means clustering algorithm, setting the number of clusters to 5. Fitting was successful, but upon analysis, it was revealed that only two of the clusters consisted of more than two neighborhoods. Upon visualization of the clusters on a folium map, it became apparent that, according to the selected K-means clustering algorithm, the city of New York can be broken down into two types of neighborhoods (Figure 2). So the next task was to investigate said types further.

Therefore, these two main clusters, named **A** (labeled as cluster 0 by the algorithm; 16 neighborhoods) and **B** (labeled as cluster 2 by the algorithm; 284 neighborhoods) were used for further analyses.

For clusters A and B, the mode for each of the top 10 most common venue categories per neighborhood was obtained, and the resulting 10 categories were used to describe each cluster in terms of three descriptors: **food venues, stores, and transportation services**. These descriptors were manually assigned by looking at the modes and allocating each one to one of the three categories (Table 1).

The descriptors were then used to generate the main visualization of this report: a bar chart reporting the share of the 10 category modes allocated to each descriptor (Figure 3). This chart reveals a very distinct profile for each of these two clusters in terms of their most common services.

Results

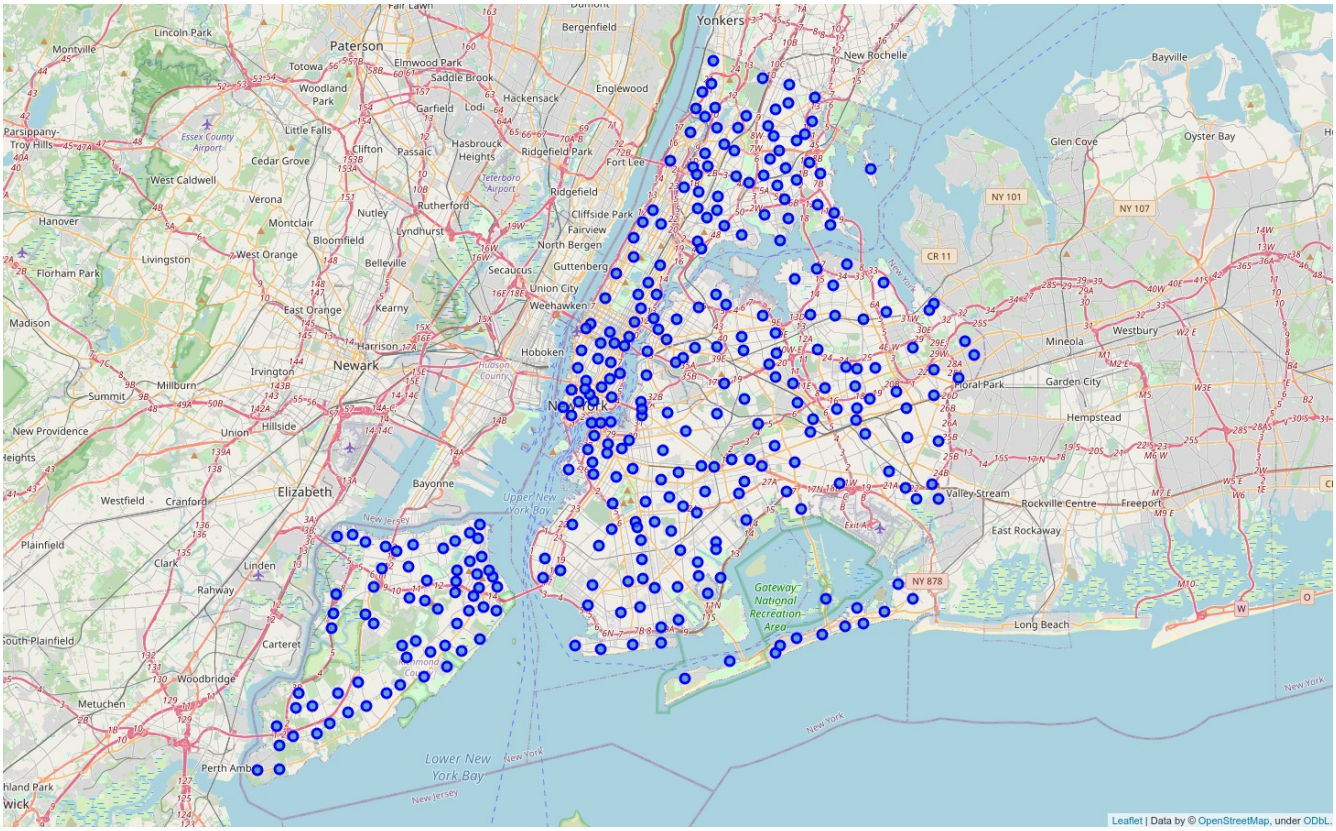

Figure 1: Map of New York City neighborhoods. Each blue dot represents a neighborhood

Through the neighborhood data and Folium's visualization capabilities, a quick overview of NYC's neighborhoods can be seen in a very visually digestible format (Figure 1). However, this visualization doesn't provide any further insights on venue features that might be attractive to a prospective entrepreneur.

Through its "explore" mode, Foursquare allows fetching data in json format, which details of nearby venues around a queried location. In the present project, up to 100 venues per neighborhood, in a 500 m radius limit, were incorporated and appended to the neighborhoods dataset. For the purposes of this study, the categories variable is of crucial interest. In order to train a clustering algorithm, the mean quantity of each venue category was reported for every neighborhood. This aggregate data was employed as the fitting dataset of a K-means clustering algorithm. The K number for this process was set to 5 in order to provide a compromise between various interesting clusters, if available, yet not too much information which might end up being confusing for readers.

However, after clustering, it became apparent that the bulk of NYC can be broken down into two main clusters, with three "outlier" clusters being present but comprising no more than two neighborhoods each (Figure 2). Notice that, while most of NYC is mostly homogeneous towards cluster B, Staten Island in particular (Figure 2, lower left) features a mixture of cluster A and cluster B neighborhoods,

suggesting that this area of NYC is more diverse in terms of its neighborhood-level venue composition.

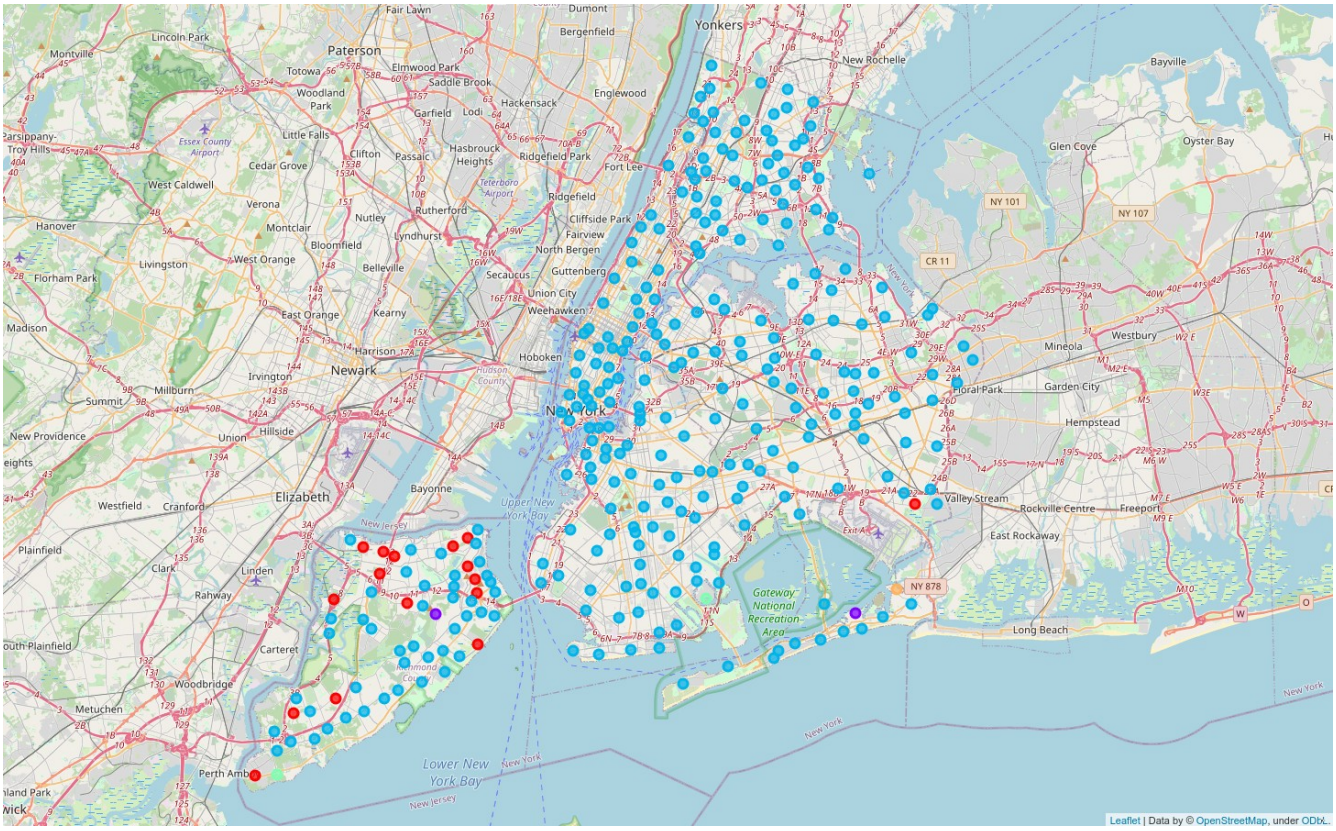

Figure 2: Map of New York City neighborhoods segmented into clusters based on most common venue types. Red: cluster 0 (renamed as cluster A). Purple: cluster 1. Cyan: cluster 2 (renamed as cluster B). Green: cluster 3. Orange: cluster 4

Simply visualizing that there are two main clusters is not very informative, so an exploration of what makes up each category was desirable. For this, it was decided to obtain the top 10 most common categories of venue for each neighborhood. Then the neighborhood mode for each of the 10 spots was computed. It was expected that these 10 mode values would confer a gauge of what categories are the most prevalent across an entire cluster. In order to provide a more actionable insight, the 10 modes for each cluster were grouped into three more general venue descriptors: **food**, **shops**, and **transportation.** These descriptors, from now on referred to as **types of venue,** were assigned to each cluster manually, simply by looking at each of the ten modes and deciding which category they fit best. The results are summarized on Table 1.

A bar chart provides a more visually informative way of understanding these values. Figure 3 summarizes the venue type composition of each cluster. An exploration of this graph reveals that these two clusters are quite different in their offerings: cluster A consists of areas dominated by shops and abundant transportation services, with comparatively few eating venues. Cluster B, in contrast, is heavily dominated by eating places. Notice that transportation venues feature in none of the top 10

spots for the neighborhood modes for this cluster, suggesting that these areas are much more defined by abundant access to restaurants and cafeterias than by great connectivity.

Table 1: Cluster share of type of venue for top 10 neighborhood venue category modes

Types of venue	Cluster A share of top 10 modes	Cluster B share of top 10 modes
Food	2	8
Shops	5	2
Transportation	3	0

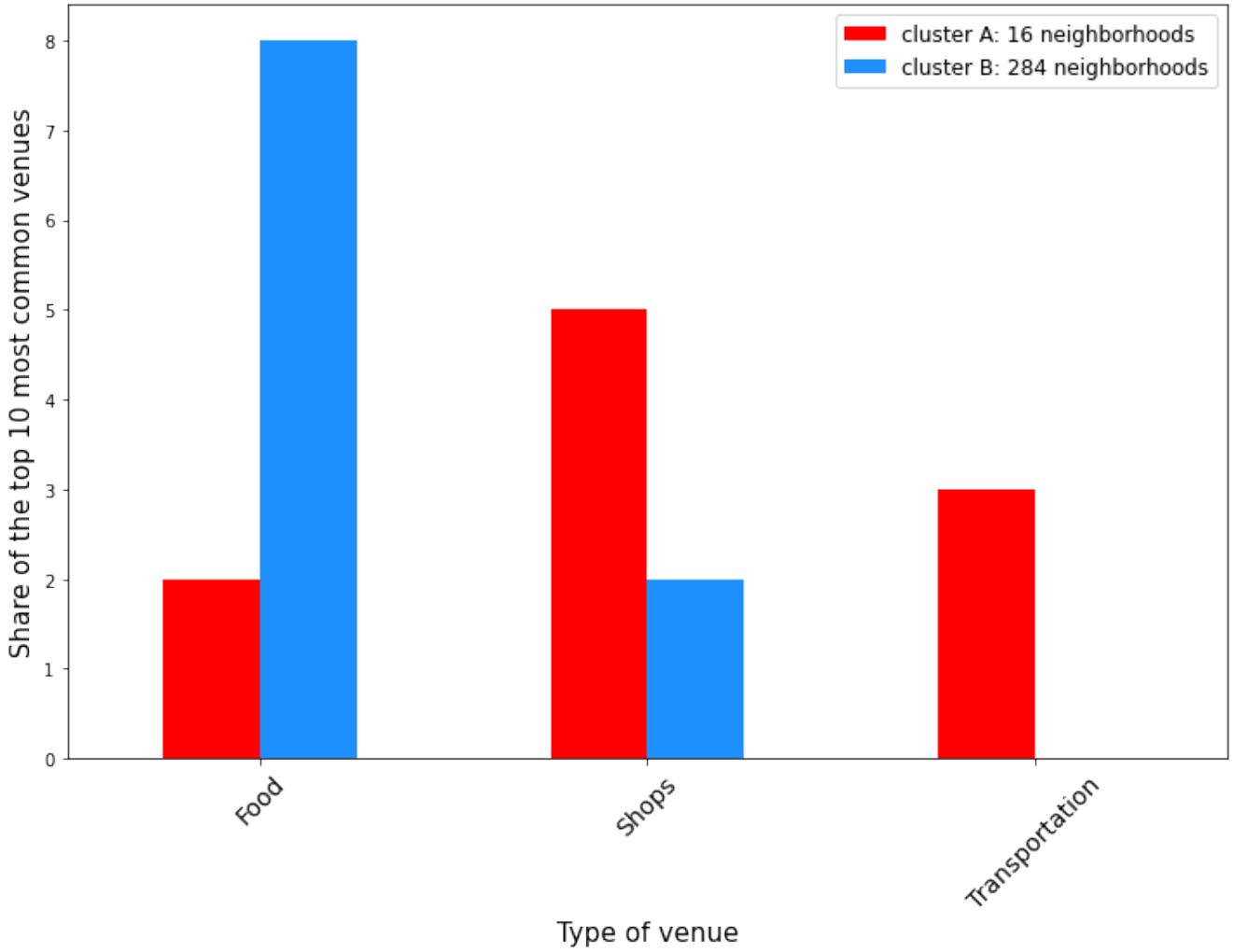

Figure 3: Venue type composition of the top 10 neighborhood-level category modes

Discussion

Opening a business in the world's top global city is no trivial matter, and prospective owners need to juggle several factors. This document is meant to provide a particular insight in the form of dividing the neighborhoods of NYC into two broad types based on venue composition. For this clustering, the K-means algorithm was chosen, as it is a simple, easy to deploy, memory efficient method that can potentially yield insights and patterns that might otherwise escape human observation. The deployed algorithm in this work revealed that, after removing "outlier" clusters that provide little insights, two main types of neighborhoods can be defined: first, neighborhoods that are heavily dominated by shops and frequent transportation, which were grouped into cluster A, and second, neighborhoods that chiefly feature eating places, such as cafeterias, restaurants, and bars, which were grouped into cluster B. It is important to note that the results shown in Figure 3 and Table 1 do not necessarily imply that neighborhoods in cluster B have poor access to transportation, but rather that eating places are far more ubiquitous.

From a geographic perspective, Figure 2 reveals an interesting pattern: the majority of neighborhoods in cluster A (i.e. those with an important presence of shops and transportation) are located in Staten Island, while the rest of NYC consists chiefly of cluster B neighborhoods (that is, dominated by eating places). Indeed, Staten Island seems to be a standout within NYC as far as its neighborhood category composition is concerned. A possible explanation for the way cluster A emerged through the algorithm could be that, given that Staten Island is the only borough not connected to the NYC subway system, residents there need to rely more heavily on bus stops, thus possibly explaining the abundance of transportation venues.

When it comes to providing recommendations, owners should be aware of what types of customers might visit their chosen neighborhood. When using online recommendation applications such as Foursquare, customers looking for restaurants and cafeterias might be more likely to be suggested locations within cluster B. In contrast, customers looking to shop might find a broader offering among neighborhoods within cluster A. Therefore, business owners can take advantage of this in order to place their business in an area that maximizes their likelihood of attracting their target customers, and thus increasing their chances of entrepreneurial success.

Conclusion

The present project provided a brief exploration of a possible categorization system of the neighborhoods of NYC based on their most common venues. Overall, two distinct clusters were identified. The far more numerous cluster B is dominant through the city, and features a majority of eating venues. Cluster A, in contrast, is particularly associated with Staten Island, and comprises a mixture of shops and intensive transportation services. The information presented herein could be used as part of a broader research effort by businesses looking to set up shop on the city. It is important to emphasize that the present insights are but a small part of the complex collection of discerning factors needed to make such a consequential decision as choosing a business location within a global city, and so it is not meant to serve as the only piece of information for prospective entrepreneurs.