Location of Manhattan's Next Neighborhood Breakfast Spot

IBM Applied Data Science Capstone
Capstone Project

Vineet Raichur
May, 2020

Background

Mr. John Snow is a 35 year old software engineer working in Manhattan, New York. John was born and brought up in the New York city. As a proud New Yorker, John's dream has been to own and operate a breakfast spot in Manhattan. Having saved some money, John has decided to finally retire from the software engineering job and pursue his dream of running a breakfast spot. John is motivated by the concept of a neighborhood business which will be a favorite among the locals. John has asked me to help him with finding the location for his breakfast spot.

Business Problem

Manhattan happens to be one of the most expensive real estate markets in the United States [1]. Therefore, choosing the right location will be critical to ensure profitability of the proposed breakfast spot. In choosing the right location, we will need to consider several factors including type of the neighborhood, size of the space, cost of renting/leasing and competition in the neighborhood. To help narrow down the search we decided that we will first identify a few neighborhoods suitable for opening the breakfast spot. Once suitable neighborhoods have been identified, further research could be conducted to identify suitable buildings/spaces that fit within the budget.

Audience

John Snow is a persona of an individual whose goal is to own and operate a neighborhood breakfast spot. This analysis can be useful to any individual thinking of owning and operating their small food and beverage business in a city where real estate is expensive, and competition can be intense. Individuals looking to start and establish a small business, compared to a chain or a franchise, will not have a brand image that they can benefit from. In this situation, small business owners will need to be mindful of the competition from the more established businesses in the vicinity and invest in creating a community around them that will support their business in the long run. Data-based insights can help small businesses in choosing the right location.

Method

In identifying suitable neighborhoods for establishing a new breakfast spot, we are looking for neighborhoods with no or a low number of existing breakfast spots and a high number of other types of venues that could indicate customer support for food and beverage businesses. We employed data science skills such as working with APIs, data wrangling, unsupervised machine learning (k-means clustering) and visualizing clusters (map, folium) for this analysis.

Data Sources and Acquisition

We used two data elements in this analysis.

1. List of neighborhoods in Manhattan
   We needed a dataset that contained a list of the neighborhoods in Manhattan along with their latitude and longitude coordinates. We used the dataset made available by NYU Spatial Data Repository [3]. NY city has a total of 5 boroughs and approximately 329 neighborhoods [2]. Because we were focusing on Manhattan specifically for this analysis, we filtered out data for the other four boroughs available in the dataset. We were finally left with the dataset with 40 neighborhoods as represented in Figure 1.

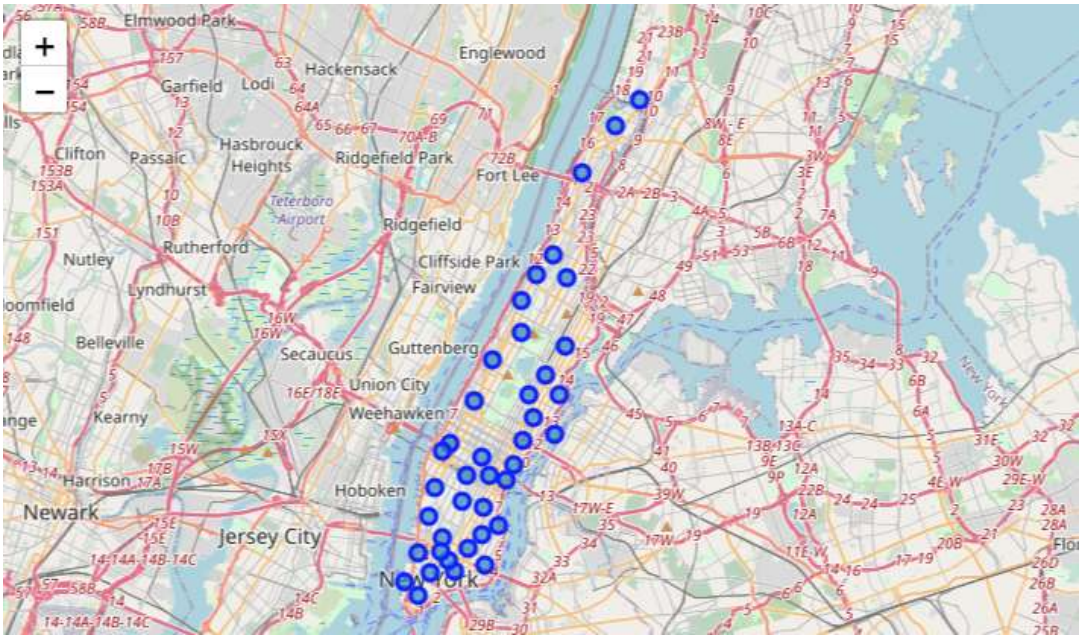

Figure 1: Map of Manhattan with the 40 neighborhoods available in the dataset

2. Data on the venues in each neighborhood of Manhattan
   We used the Foursquare data on the venues in each neighborhood. We gathered top 100 venues that are within a 500 meter radius of the neighborhood geographical coordinate. To gather venue data, API calls were made to Foursquare in a loop with geographical coordinates of each neighborhood. Foursquare data contained venue name, category, latitude and longitude for each venue.

Data Pre-processing

Data from Foursquare contained a list of top 100 venues within the specified range in each neighborhood. We converted the categorical data with venue name and category into quantitative format that can be used in machine learning algorithms during analysis. We pre-processed the venue dataset to calculate the number of venues present in each category in each neighborhood. We used the one hot encoding method to produce a dataframe in which each row represents a venue in a

specific neighborhood and each column represents the venue category. Each cell in this dataframe represents whether each venue belonged to a certain category. We then grouped the rows by neighborhood and averaged the numerical values. This resulted in a new dataframe that contained the average frequency of occurrence of venues in each category in each neighborhood.

For this analysis we are interested in identifying clusters of residential neighborhoods with low frequency of breakfast spot and high frequency of food and beverage businesses. We decided to consider the frequency of occurrence of the following types of venues in our analysis:

1. Breakfast spot
2. Restaurants
3. Park
4. Office
5. Bakery
6. Coffee shop
7. Residential building

Exploratory Data Analysis

As part of the exploratory data analysis (EDA) we summarized the frequency of occurrence of each venue category and checked the association between occurrence of breakfast spots and other venue types. Figure 2 shows the mean and standard deviation of each of the seven venue categories considered in the analysis. This figure shows how certain types of venues such as coffee shops, parks and bakeries occur a lot more often on average than other venues. Comparing against these average values in the overall dataset, we can tell if frequency of occurrence of certain types of venues in the clusters are actually high or low.

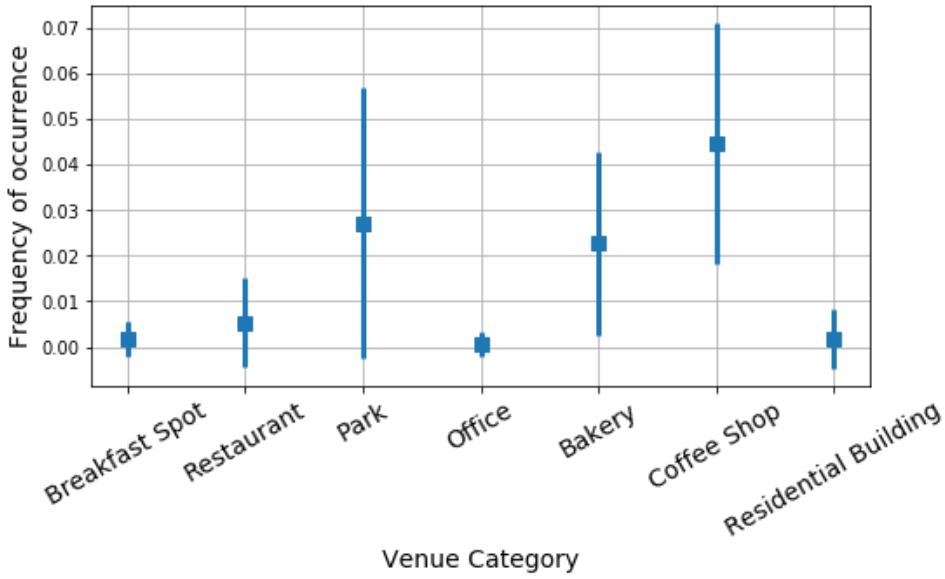

Figure 2: Mean and standard deviation of frequency of occurrence of each venue type

Figure 3 shows the scatter plots of frequency of occurrence of breakfast spots vs all other types of venues. We can immediately see two clusters in the data – one with and one without breakfast spots. Locations where no breakfast spots exist, there is a wide range in the frequency of occurrence of other venues.

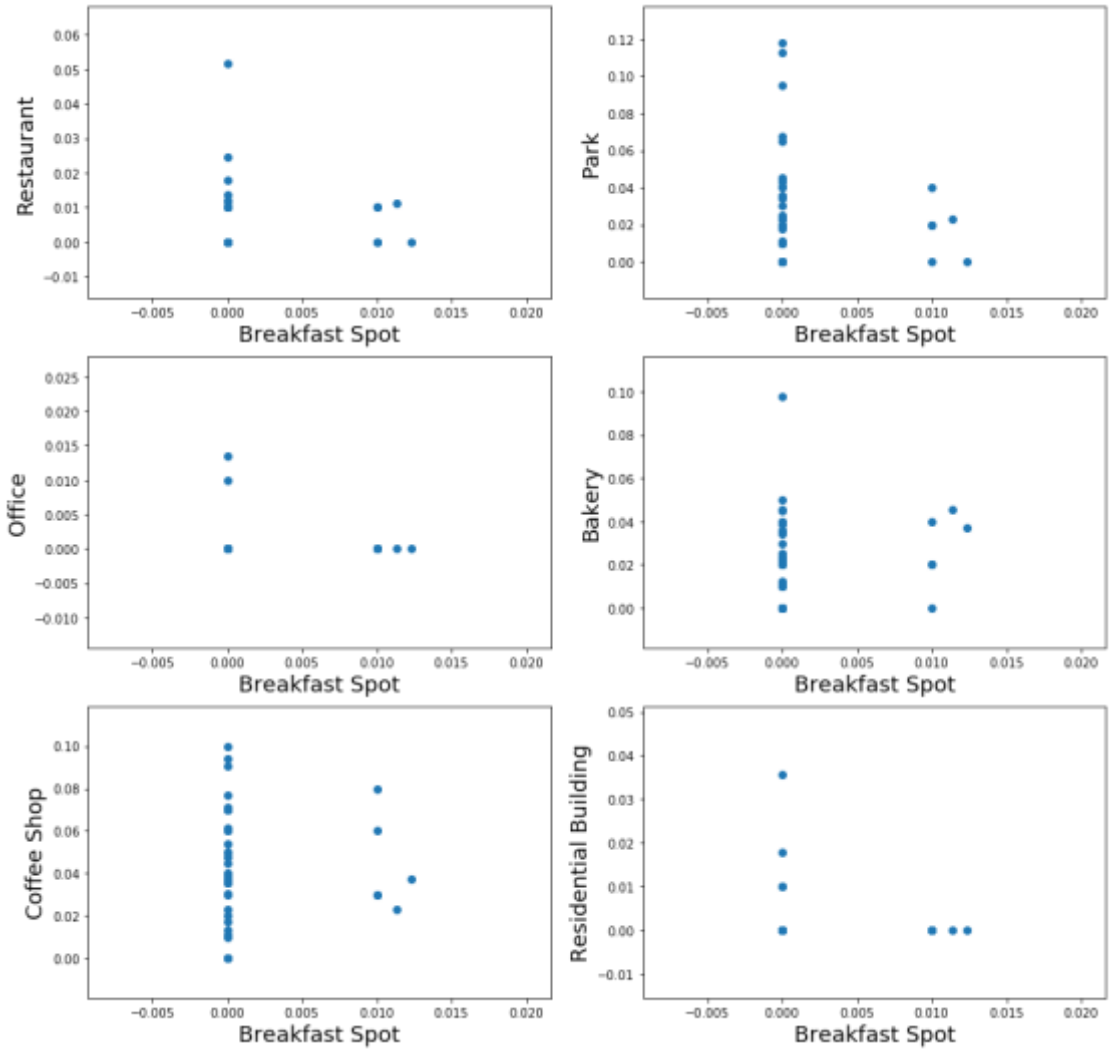

Figure 3: Scatter plots of frequency of occurrence of breakfast spots vs all other types of venues

Analysis

We performed clustering on the data using k-means clustering approach, which is one of the popular unsupervised machine learning algorithms. Unsupervised algorithms make inferences from datasets using only input data without using any labelled outcomes. This algorithm will group similar data points together into k clusters and discover underlying patterns. We first identified the ideal number of k clusters to split the data. We used Within Cluster Sum of Squares (WCSS), which measures the squared average distance of all the points within a cluster to the cluster centroid, to determine the optimal numbers clusters for the given data.

Results

Figure 4 shows the WCSS for each number of k clusters. Ideally, the "elbow" – k clusters after which the WCSS drops at a slower rate – is identified. K number of clusters at the elbow are deemed at the optimal number of clusters to split the data. In our case there is no clear elbow, but the slope beyond 6 clusters is the same. We therefore decided to run the k-means algorithm with 6 clusters.

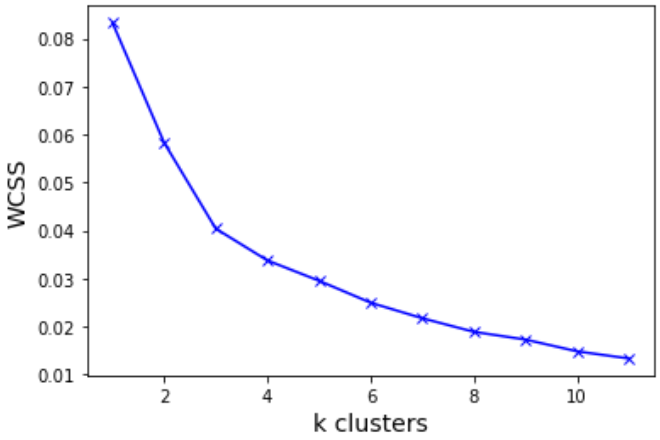

Figure 4: WCSS values at each clustering trial with k number of clusters

Table 1 shows the average frequency of occurrence of each venue type in each of the six clusters. Three of these clusters have no breakfast spots and could be ideal spots for the new breakfast spot. Because the goal is to open a neighborhood spot, we need a location with residential buildings and parks as well. Cluster 3 meets the requirements of no breakfast spots, no offices and high occurrence of residential buildings and parks.

Table 1: Average frequency of occurrence of each venue type in each cluster

Cluster Labels	Breakfast Spot	Restaurant	Park	Office	Bakery	Coffee Shop	Residential Building
0	0	0	0.1040	0	0	0.0709	0
1	0.0017	0.0035	0.0140	0.0008	0.0186	0.0735	0.0008
2	0	0.0062	0.0393	0	0.0141	0.0429	0.0079
3	0.0036	0.0146	0.0278	0	0.0476	0.0154	0
4	0	0.0068	0.0926	0.0068	0	0.0068	0
5	0.0022	0.0020	0.0041	0	0.0282	0.0306	0

Figure 5 shows the neighborhoods in Manhattan color coded on the basis of the clusters they were assigned to. We can see that the clusters are spread out geographically.

Figure 5: Map of Manhattan with the neighborhoods assigned to each of the six clusters

Discussion

Results of the k-means clustering showed that neighborhoods in the third cluster could be ideal for setting up the new breakfast spot. Table 2 shows the average frequency of occurrence of each of the 7 venue types in each neighborhood in cluster 3. All 8 neighborhoods in this cluster have no breakfast spots and no offices, but only 3 of them have residential buildings. We took a closer look at these 3 neighborhoods – Hudson Yards, Roosevelt Island and Turtle Bay. All 3 neighborhoods have a good occurrence of parks and coffee shops, but they vary in the occurrence of restaurants and bakeries. These three neighborhoods nevertheless, are good candidates for the breakfast spot.

Table 2: Average frequency of occurrence of each venue type in each neighborhood in Cluster 3

Cluster Labels	Neighborhood	Breakfast Spot	Restaurant	Park	Office	Bakery	Coffee Shop	Residential Building
2	Civic Center	0	0	0.0430	0	0.0215	0.0538	0
2	Flatiron	0	0.01	0.03	0	0.01	0.02	0
2	Gramercy	0	0.0119	0.0238	0	0.0119	0.0476	0
2	Hudson Yards	0	0.0179	0.0357	0	0	0.0357	0.0179
2	Roosevelt Island	0	0	0.0357	0	0	0.0357	0.0357
2	Sutton Place	0	0	0.0408	0	0.0204	0.0612	0
2	Tribeca	0	0	0.0649	0	0.0390	0.0390	0
2	Turtle Bay	0	0.01	0.04	0	0.01	0.05	0.01

Conclusion

Our aim in this project was to help narrow down the search for the location of a new neighborhood breakfast spot. We started by gathering necessary location data on the neighborhoods in Manhattan and venue data for these neighborhoods from Foursquare. We then used the k-means clustering approach to cluster the neighborhoods on the basis of the frequency of certain types of venues. We found one out of 6 clusters that met the requirements of no breakfast spots, no offices and high occurrence of residential buildings and parks. Within this cluster, we found that Hudson Yards, Roosevelt Island and Turtle Bay would be ideal candidates for the new breakfast spot.

Further research will need to look into the availability and cost to rent/lease spaces in these neighborhoods. We will also need to consider what type of businesses are located in the immediate vicinity of the candidate spaces. To further refine this analysis we could add population densities and average commercial space rents in each neighborhood to the data on frequency of occurrence of venue. Addition of these variables could help us cluster neighborhoods not only on the basis of venues in the vicinity but also on the basis of expected rent and the number of customers.

References

[1] CNBC, “Manhattan real estate is the most expensive in the US per square foot with some properties topping $10,000: Study,” 2018. [Online]. Available: https://www.cnbc.com/2018/08/11/manhattan-real-estate-is-the-most-expensive-in-the-us-per-square-foo.html. [Accessed: 23-May-2020].

[2] Baruch College, “New York City (NYC) Neighborhoods - By Borough,” 2020. [Online]. Available: https://www.baruch.cuny.edu/nycdata/population-geography/neighborhoods.htm. [Accessed: 24-May-2020].

[3] NYU Spatial Data Repository, “2014 New York City Neighborhood Names,” 2018. [Online]. Available: https://geo.nyu.edu/catalog/nyu_2451_34572. [Accessed: 23-May-2020].