A Battle of Neighbors

Xiyue Wang

Introduction

New York City (NYC) and Tokyo - two major international megacities.
They share many common characteristics. First, of course they are all densely populated! NYC has more than 8 million residents with population density around 30000 per mile. Tokyo has 13 million residents, and the population density is about half of NYC. In terms of living quality, both cities have convenient public transportation systems, adequate schools and good health care system. Speaks of environment, they are all accessible to seaside beaches, with similar temperature and humidity rate all year around. What's the differences?

NYC is a multi-cultural city, known for its boroughs and the cultures of individual neighborhoods. English the spoken language for business. However, more than 40 percent of the NYC population was born outside of the United States who speaks more than 200 languages. NYC's boroughs include Manhattan, Queens, Brooklyn, Staten Island, and the nearby cities of Long Island. It is so spread-out that often when people refer to 'city' they just mean Manhattan. Within these boroughs, NYC can be divided into neighborhoods like Chinatown and Financial District, which are very distinct in nature. They have different lifestyles and cultures, populated by people with different ethnicities.

Tokyo, on the other hand, is divided up into various districts called wards. Even with growing number of people communicating in English, Japanese is still essential for living in Tokyo. It is more famous for its cultural landmarks than its diversity. Also, it is renowned for being an in-gathering of Japanese citizens from all over the country. This is because the city population is still around 90% Japanese, and many businesspeople and tourists who have lived there are only temporary residents. It is highly likely that people will be shocked by the stoic silence on the trains, the inevitable language barrier, and the differences in cultural etiquette when first come to Tokyo. However, with growing business and uniqueness of it, Tokyo now attract more and more global citizens.

Business Question

Since Tokyo and Manhattan have some similarities and some striking differences, one interesting thing to compare is that: how similar are they in terms of business owners? If we are profiting in Manhattan, which wards are we more likely to be successful in Tokyo

In this project we will next explore a sample question - finding an optimal location for a family-style Chinese restaurant we want to launch.

Data Acquisition

Manhattan's geography data was obtained from Coursera. The geojson data for Manhattan was forked from GitHub. The Tokyo wards information was obtained from Wikipedia. The coordinates of each city neighborhoods were requested using Google API. Additional environment, lifestyle, socio-economic data was collected from Wikipedi.
The venue information was requested from Foursquare.
Here is a sample API:
https://api.foursquare.com/v2/venues/explore?&client_id=###&client_secret=###&v=###&ll=#,#&radius=#

Methods

Clustering

First, the venue information for each neighborhood/ward in Manhattan and Tokyo was obtained from Foursquare, free open source API Foursquare. We set the radius for venue collection at 2000 meters. The venue name, category, coordinates and distance to neighborhood center were returned. We next applied onehot encoding to analyze venue categories for each neighborhood.
In this study, we performed KMeans clustering, we first identified the optimal cluster number for each city using elbow method. Next we performed clustering analysis and feed the results to original dataset to extract the cluster features. A folium map was generated for both cities.

Correlation

To calculate the correlations, we extracted the common categories for both cities. Since there are a lot of zero values, we will just use Spearman's correlation coefficient. Spearman's correlation is what is known as a non-parametric statistic, which is a statistic whose distribution doesn't depend parameters (statistics that follow normal distributions or binomial distributions are examples of parametric statistics). Frequently non-parametric statistics are based on the ranks of data rather than the original values collected. This happens to be the case with Spearman's correlation coefficient, which is calculated similarly to Pearson's correlation. However, instead of using the raw data, we use the rank of each value.

Chinese Restaurants Analysis

Shifting demographics and changing lifestyles are driving the surge in food-service businesses. Busy consumers don't have the time or inclination to cook. The reality is, many restaurants fail during their first year, frequently due to a lack of planning. Our target is to launch a family-style Chinese restaurant. Family style restaurants charge reasonable prices. They offer speedy service that falls somewhere between that of quick-service places and full-service restaurants. Based on the description we picked the best neighborhoods from our cluster's analysis, and we requested the food-related venues from Foursquare following two criteria:
1. Since there are lots of restaurants in both cities, we will try to detect locations that are not already crowded with restaurants (less than 5 restaurants per 3000 people).
2. We are also particularly interested in areas with no Chinese restaurants in vicinity (less than 3 Chinese restaurants per neighborhood and within 500 meters no Chinese restaurants).
The results were visualized using folium

Results

Tokyo can be separated by 3 clusters

From the folium map, it was clear that clusters in Tokyo are generally geographically separable. More specific, cluster 0 in Tokyo is located in the city center includes famous wards like Shinjuku and Shibuya. Then the 'middle circle', which is west to the city center, has wards like Sugumani and Setagaya. The northeastern wards form another cluster. We next examine the popular venues within each cluster.
It is interesting to see that the city center (cluster 0) has a lot japanese restaurants, hotels, sushi restaurants. The middle circle, however, contains more family-stlye restaurants like ramen joints, sake bar and cafe. The northestern cluster have much fewer food joints, with much more covenient stores/grocery stores.
see an interesting pattern that the clusters is formed almost surround the center of the city. The 'inner circle' includes famous wards like Shinjuku and Shibuya. The 'middle circle' are the west wards next to 'inner circle', such as Sugumani and Setagaya. Some northeastern wards form a cluster, including Kitasenju and Tateishi etc.

To further explore, we examined what's the most prominent features (venue categories) within each cluster. It was interesting to see that the cluster 0 (inner circle) has a lot of Japanese restaurants, hotels, sushi restaurants. The middle circle, however, contains a lot of ramen restaurants, sake bar, cafe and BBQ joints. The northeastern cluster has convenience stores and grocery shops.

cluster 0 (green):

Japanese Restaurant	0.052857
Sushi Restaurant	0.038571
Hotel	0.037143
Dessert Shop	0.035714
Sake Bar	0.030000

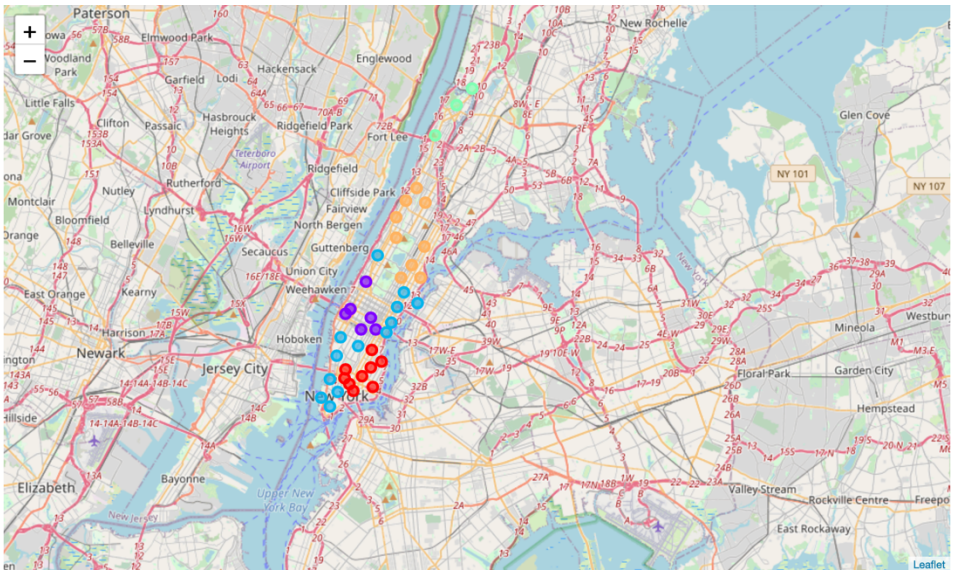
cluster 1 (red):

Ramen Restaurant	0.079
Sake Bar	0.077
Café	0.053
Japanese Restaurant	0.046
BBQ Joint	0.038

cluster 2 (blue):

Convenience Store	0.214
Ramen Restaurant	0.076
Discount Store	0.036
Grocery Store	0.034
Supermarket	0.032

Manhattan can be separated by 5 clusters



cluster 0 (red):

Pizza Place	0.035556
Café	0.028889
Hotel	0.026667
Coffee Shop	0.024444
Bakery	0.024444

cluster 1 (purple):

Theater	0.133333
Park	0.041667
Gym	0.031667
Hotel	0.031667
Gym / Fitness Center	0.028333

cluster 3 (green):

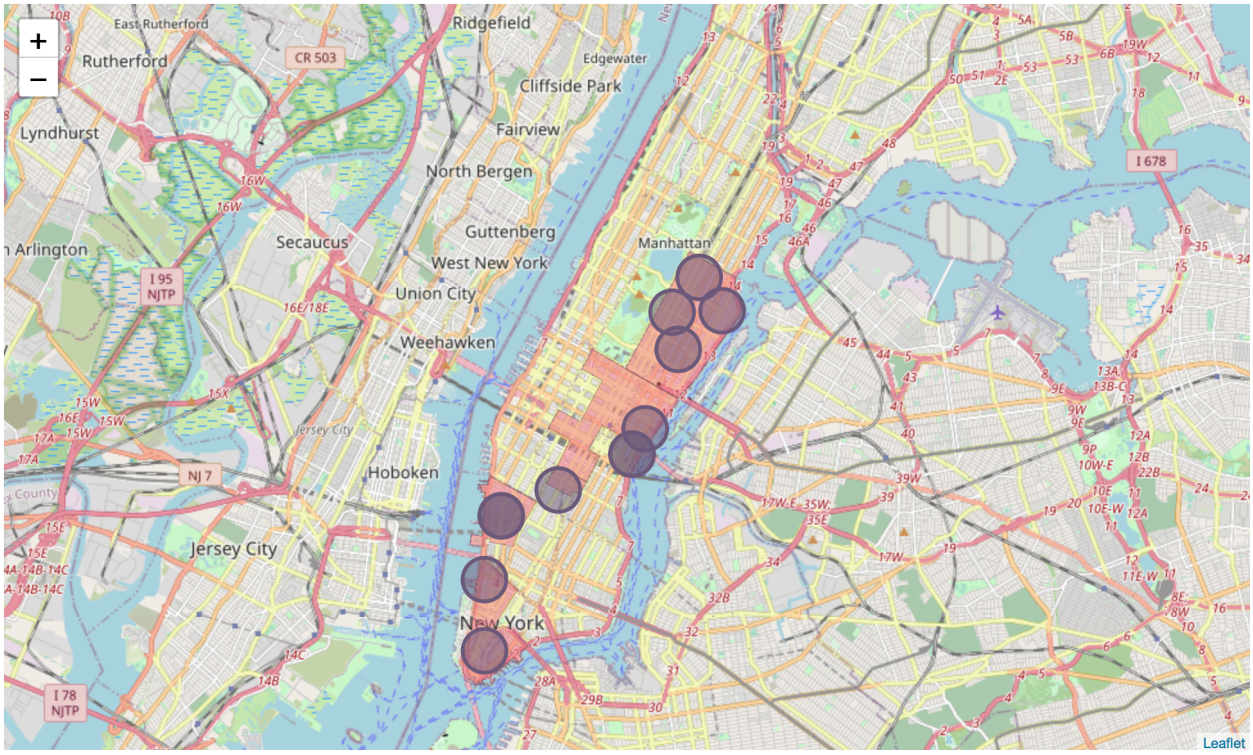
Pizza Place	0.063333
Mexican Restaurant	0.060000
Park	0.056667
Latin American Restaurant	0.043333
Café	0.040000

In Manhattan, clusters were formed follow geometrical pattern too, which was consistent with the known-city culture - neighborhoods are segregated by ethnicity and lifestyle. For example, cluster 0 (lower west, upper east) near financial district and residential area, has parks, coffee shops, gyms and bakeries, which offer convenient options for a quick meal. Cluster 3 (upper west), which was actually separated from the main island, has Mexican and Latin American restaurants, probably due to most of the residents are Hispanic. Cluster 3 (mid-town), is where most tourists visit and stay, contains gyms, parks, theaters and hotels!

Correlation: They are not very similar!

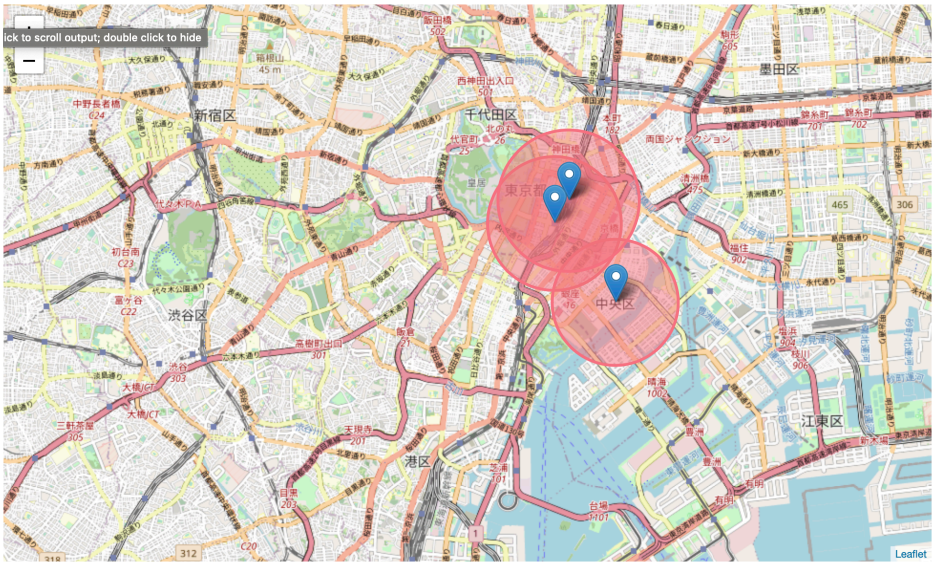
Most similar: Tokyo cluster 0 and Manhattan cluster 4, value is 0.3579922202913762. Most dissimilar: To kyo cluster 1 and Manhattan 1, value is 0.015357087258719685.

Manhattan's market is really limited for Chinese restaurants



The boundary for neighborhoods had less than 5 restaurants per 3000 people, as well as contains less th an 3 restaurants, were marked with red. They are Carnegie Hill, Financial District, Flatiron, Lenox Hill, Trib eca, Tudor City, Turtle Bay, Upper East Side, West Village and Yorkville. The Chinese restaurants and their 500 meters radius were marked with purple.

Tokyo is wide open for Chinese restaurants



Since we didn't obtain the GIS/geojson data for Tokyo, we used the area to calculate radius (sqrt(area)/pi). The wards had less than 5 restaurants per 3000 people, as well as contains less than 3 restaurants were marked with red. They are Chiyoda, Chuo and Toshima. The location for Chinese restaurants were marked. They are all close to the center of the wards.

Discussion

Our cluster analysis shows that in both Tokyo and Manhattan, clusters are generally geographically separable. For example, cluster 0 in Tokyo is located in the city center includes famous wards like Shinjuku and Shibuya. Then the ‘middle circle', which is west to the city center, has wards like Sugumani and Setagaya. It is interesting to see cluster 0 has a lot of Japanese restaurants, hotels, sushi restaurants. The middle circle, however, contains more family-style restaurants like ramen joints, sake bar and cafe. The northeastern cluster have much fewer food joints, with much more convenient stores/grocery stores.

In Manhattan, clusters are formed follow geometrical pattern too. For example, cluster 0 (lower west, upper east) near financial district and residential area, has parks, coffee shops, gyms and bakeries, which offer convenient options for a quick meal. Cluster 1 (upper west), which is actually separated from the main island, has Mexican and Latin American restaurants, corresponding to its ethnicity. Cluster 3 (mid-town), is where most tourists visit and stay, contains gyms, parks, theaters and hotels!

Our cluster results show some very distinct patterns. It is probably due to a combination of popularity among tourists, closeness to city center, socio-economic dynamics and many other facts. We should note that this analysis is based on the information provided by Foursquare, which is not include all possible venues. Also, the neighborhoods boundaries can be vague, and can change over time with the migration of residents/socio-economical change. We need to be careful when doing this type of cluster analysis.

We next ran a correlation analysis using spearman correlation. We can see that Tokyo cluster 1 are most similar to Manhattan cluster 4. They both have small restaurants like cafes, coffee shops, pizza place, which is not fancy but convenient food options. On the other hand, Tokyo cluster 2, which is the more 'suburban' region, is very different from Manhattan cluster 3, which is the center spot of Manhattan. In another sense, if you are a small restaurants owner in Manhattan and want to open a new restaurant in Tokyo, maybe neighborhoods such as Sugumani and Setagaya are the ones you could consider. However, in real world, open a new business can be complicated. Facts like residents age, gender, competitiveness are all important, but this result provide some interesting insights.

After directing our attention to this narrower area of interest (clusters favor family-style restaurants), we start to filter neighborhoods that are not densely populated with restaurants. We first obtained only food-related APIs from foursquare, and any neighborhoods with more than 5 restaurants per 3000 people were exclude. After initial screening, we have 5 wards in Tokyo and 21 neighborhoods in Manhattan. Those location candidates were then clustered to create zones of interest which contain less than 3 Chinese restaurants. After more filtering, only 10 neighborhoods of Manhattan (upper East, midtown East, SoHo-TriBeCa-Civic Center-Finantial) and 3 wards in Tokyo (Chuo, Chioda, Toshima) left.

Purpose of this analysis was to only provide info 1) business owner want to launch in the other city, 2) areas suitable for new restaurants but not crowded with existing restaurants - it is entirely possible that there is a very good reason for small number of restaurants in any of those areas, reasons which would make them unsuitable for a new restaurant regardless of lack of competition in the area. Notably, each existing Chinese restaurant has almost non-overlapping in Manhattan, but they are quite overlapped in Tokyo! There might be some interesting reasons we should dig deeper into. 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 no nearby competition, but also other factors taken into account and all other relevant conditions met.

Conclusion

Purpose of this project was to identify the similarity of Tokyo and Manhattan in terms of venues.
We will use our data science powers to generate a few most promising neighborhoods based on given criteria. By clustering neighborhood venues from Foursquare data, we have first identified neighborhood clusters that have different preferences to venues. We then investigated the similarity of Tokyo wards versus Manhattan neighborhoods. These results, with the final decision on optimal Chinese restaurant location will be made by stakeholders based on specific characteristics of neighborhoods and locations in every recommended zone, taking into consideration additional factors like attractiveness of each location (proximity to park or water), levels of noise / proximity to major roads, real estate availability, prices, social and economic dynamics of every neighborhood etc.