Applied Data Science: Capstone

Shopping Mall and Department Store Location Selection in New York

Charles Xi Yan

27th March 2020

Contents

1 Introduction

As the biggest city in New York and one of the most famous tourism destination in the world, New York is always regarded as an ideal location for retail industry, especially for shopping malls and department stores. Shopping malls and department stores not only provide customers with an excellent platform for dining, shopping, film-seeing and other entertainment activities, but also offer retailers a good opportunity to expand their business and promote their products. As a fact, there are a lot of shopping malls and department stores in New York, say Macy's. However, as a complicated issue, openning a new shopping mall or department store needs quite a serious and careful consideration because of the business running cost, such as the rent and the management cost. To some extent, a relatively ideal location makes the business on the halfway to the success.

1.1 Problem Description

The object of this project is to analyze the neighbourhoods in New York so that finally the best location for a shopping mall or a department store is determined. With data science methods and machine learning techniques learnt from IBM Data Science course, we aim to find a solution to such a question: if a real estate developer is planning to open a new shopping mall or a department store in New York, where should be recommended?

1.2 Target of the Project

This project offers useful advice especially to property developers or real estate investors for investing shopping malls or department stores in New York. This project is opportune as the shopping mall or department store industry in New York, especially in Manhattan, has almost fully developed and some districts are suffering from oversupply of shopping malls and department stores. As Financial Times ever reported, New York has passed the peak era of shopping malls and department stores. However, the one with eye-catching features is still proved to have a great chance of success.

2 Data

We list the required dataset as follows:

1. List of neighbourhoods in New York.
2. Geographical coordinates of these neighbourhoods, i.e. the latitude and the longitude, in order to plot the map and collect the venue data.
3. Venue data, in order to perform clustering on the neighbourhoods.

2.1 Source of Data

The json data in the link (`https://cocl.us/new_york_dataset`) contains a list of neighbourhoods in New York with a total of 306 neighbourhoods in 5 boroughs. Wget command and relevant python packages are used to access the data. Geographical coordinates of neighbourhoods are obtained by using Geopy package.

Foursquare API is then used to collect the venue data of these neighbourhoods. As Foursquare API provides various categories of venue data, we directly focus on the shopping mall and department store category for the project object.

2.2 Skills and Techniques

The project involves a big range of data science skills and machine learning techniques, from loading data by wget command, using Foursquare API, data cleaning, data wrangling, to K-mean clustering technique and Folium map visualization. All detail of applying these skill and techniques will be shown in the next section.

3 Methodology

In order to find the solution to the problem, we need to obtain the neighbourhood data of New York city. Fortunately, we can find it directly from the link (`https://cocl.us/new_`

`york_dataset`) by using wget command in python. As this is a list of neighbourhood names, we add latitude and longitude of each neighbourhood to the dataframe through requesting Foursquare API. Geopy is a very useful python package for obtaining geographical coordiantes based on the address. Then, the required data above will be shown in python dataframe and thus each neighbourhood of the object is displayed on the map through Folium package. This gives us a double check on the location of each neighbourhood and a direct overview of the object.

Next, Foursquare API is used to get top 100 venues within a certain radius (500 metres) of each neighbourhood, which requires a Client ID and Client Secret in a Foursquare Developer App. After receiving the request, Foursquare returns the venue data in a JSON file. Thus, we extract the venue name, venue category, longitude and latitude. Based on the venue data, we check how many unique venue categories were returned. Then for each neighbourhood, we calculate the mean of the frequency of occurrence of each venue category. Thus we obtain the raw "raw" data before clustering neighbourhoods. Since the object is on shopping malls and department stores, we pick mean values of shopping malls and department stores up to make a new dataframe.

Finally, data analysis will be done by K-means clustering neighbourhoods. K-means clustering algorithm identifies k centroids, and then allocate every data point to nearest cluster, while keeping the centroids as small as possible. As one of the simplest and popular unsupervised machine learning algorithms, K-means clustering is very suitable for solving this problem. Here we choose k to be 4, since we aim to have clusters that are suitable for openning both shopping malls and department stores, more suitable for one of them compared with the other one, and not suitable for either of them. It should be noted that the criterion of openning a shopping mall or a department store only considers the occurrence of shopping malls or department stores in different neighbourhoods.

4 Results

The following bullet points are observations of k-means clustering for neighbourhoods in New York city, based on the frequency of occurrence for shopping malls and department stores:

1. Cluster 0 (Colour Red): Neighbourhoods with almost no existence of either shopping malls or department stores.

2. Cluster 1 (Colour Purple): Neighbourhoods with high concentration of shopping malls but almost no existence of department stores.

3. Cluster 2 (Colour Light Blue): Neighbourhoods with moderate number of both shopping malls and department stores.

4. Cluster 3 (Colour Gold): Neighbourhoods with high concentration of department stores but almost no existence of shopping malls.

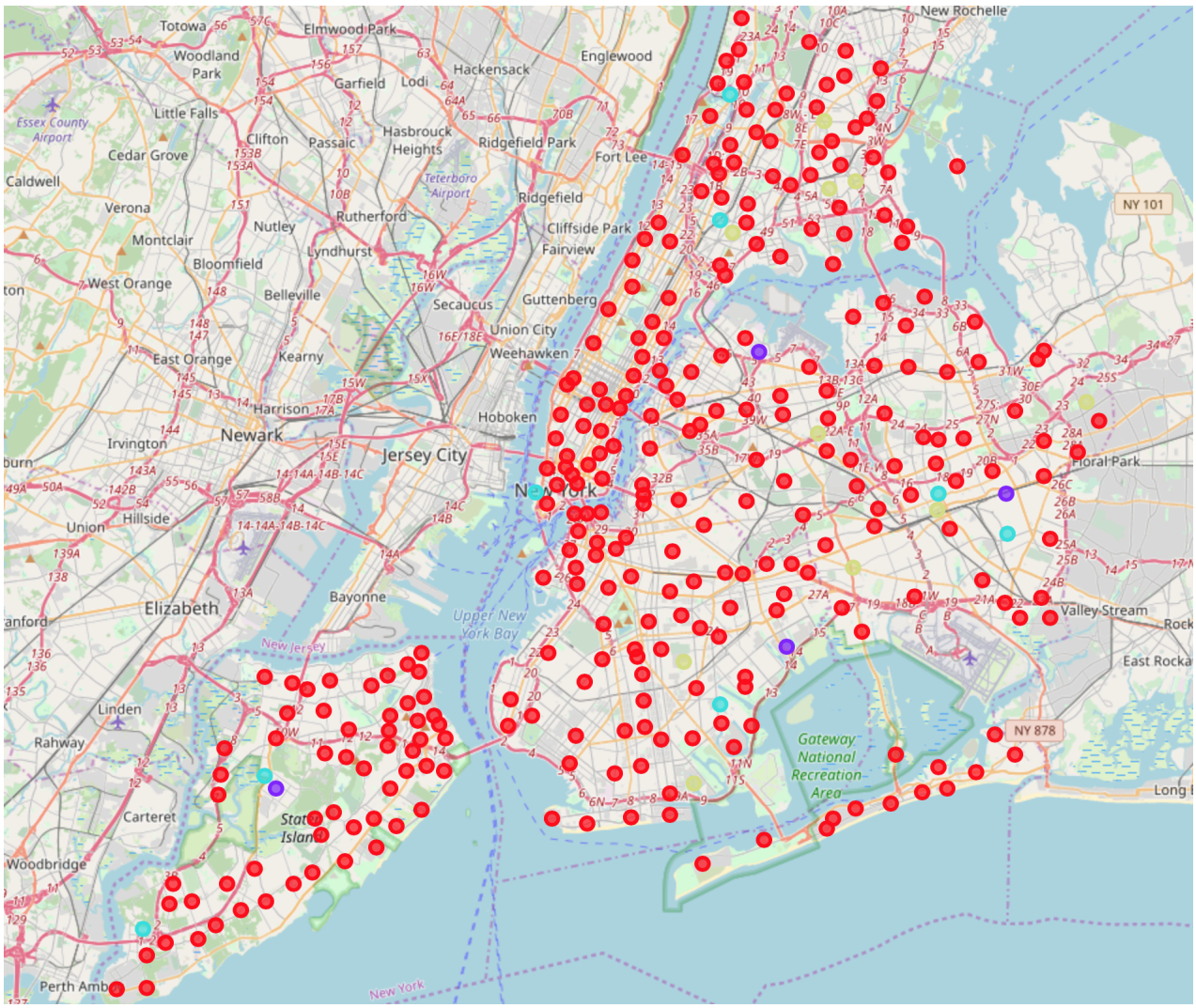

Figure 1: 4-mean Clustering on Neighbourhoods in New York City.

Based on observations above, we give the following results:

1. For the neighbourhoods belonging to Cluster 0, from the perspective of competition due to oversupply and high concentration, both shopping malls and department stores can be considered.

2. For the neighbourhoods belonging to Cluster 1, investing on a department store seems to have less pressure from the peer competition.

3. For the neighbourhoods belonging to Cluster 3, investing on a shopping mall seems to have less pressure from the peer competition.

4. For the neighbourhoods belonging to Cluster 2, the investment condition seems like a intermediate case of Case 2 and Case 3 above.

5 Discussion, Limitation and Improvement for Future Research

Based on the dataframe we used and the model we built, there are still several places that can be boosted. The corresponding improvements for future research are also attached to discussion and limitations as follows:

1. Considering the reality factors, the safety scaling data and the average income of neighbourhood residents need to be taken into consideration. Thus, from this perspective, whether cluster 0 is suitable for investment on shopping malls or department stores cannot be determined immediately.

2. For neighourhoods that already have shopping malls and department stores, the market has also come into a good form. That implies the safety and the consumption do not influence the decision very much. Thus, if investing shopping malls or department store there, we need more data about what features these shopping malls or department stores already have. Then after analysis, we can make the decision.

3. From the outcome of Cluster 3, there area still some neighbourhoods that do not have department stores. Hence, K=4 may not be the optimal value. If possible, we could try

to obtain a training data and testing data that already have cluster indices under our criteria. Then, we can run a few more cases of k-mean clustering analysis for different values of K and find the best K for analysis.

6 Conclusion

In this project, we have gone through the process of identifying the business problem, specifying the data required, extracting and preparing the data, performing machine learning by clusting the data into 4 clusters based on their similarities, and finally providing recommendations to the relevant stakeholders, i.e. property developers and real estate investors regarding an ideal location to open a new shopping mall or a new department store. To answer the business question that was raised in the introduction section, the answer proposed by this project is: neighoubhoods in cluster 2 are suitable for both shopping malls and department stores, considering the complete community facility and peer competition; neighbourhoods in cluster 1 are suitable for department stores considering the peer competition; neighbourhoods in cluster 3 are suitable for shopping malls considering the peer competition. The findings of this project will help the relevant stakeholders to capitalize on the opportunities on high potential locations while avoiding overcrowded areas in their decisions to open a new shopping mall.

Bibliography

[1] *2014 New York Neighbourhood Names*. Spatial Data Repository, New York University, 2014. Retrieved from `https://geo.nyu.edu/catalog/nyu_2451_34572`.

[2] *Foursquare Developers Documentation*. Foursquare. Retrieved from `https://developer.foursquare.com/docs`.