COURSERA CAPSTONE

IBM Applied Data Science Capstone

Opening a New Hotel in New York City

2020

JULY 19

The University of Texas at Austin
Authored by: Lu Wang

Introduction

For many business or vocation visitors, Hotels play an important role in their trip. Customers of the hotel can get easy access to the meeting center or sightseeing spots, have authentic food, enjoy pleasing view in front of their windows directly. Hotels are like a second for various visitors. For hotel owners, the central location and convenient transportation provides a great distribution channel to market their products and services. Property developers are also taking advantage of this trend to build more hotels to cater to the demand. As a result, there are many hotels in the New York city and many more are being built. Opening hotels allows property developers to earn consistent rental income. Of course, as with any business decision, opening a new hotel requires serious consideration and is a lot more complicated than it seems. Particularly, the location of the hotel is one of the most important decisions that will determine whether the hotel will be a success or a failure. It's no secret, location is everything in the hospitality business. The advantageous location for a new restaurant or hospitality business guarantees its long-term success. Advantageous location usually means it's easy to get found, followed and engaged.

Business Problem

The objective of this capstone project is to analyze and select the best locations in the New York City, USA to open a new Hotel. Using data science methodology and machine learning techniques like clustering, this project aims to provide solutions to answer the business question: In the New York City, USA, if a property developer is looking to open a new hotel, where would you recommend that they open it?

Target Audience of this project

This project is particularly useful to property developers and investors looking to open or invest in new hotels in the New York City. This project is timely as the city is currently suffering from oversupply of hotels.

Data

To solve the problem, we will need the following data:

- List of neighborhoods in New York City. This defines the scope of this project which is confined to the New York City, the largest metropolitan area in the world by urban landmass.
- Latitude and longitude coordinates of those neighborhoods. This is required in order to plot the map and also to get the venue data.
- Venue data, particularly data related to Hotels. We will use this data to perform clustering on the neighborhoods.

Sources of data and methods to extract them

This Wikipedia page (https://en.wikipedia.org/wiki/Category:New_York_City) contains a list of neighborhoods in New York City, with a total of 31 neighborhoods. We will use web scraping techniques to extract the data from the Wikipedia page, with the help of Python requests and beautifulsoup packages. Then we will get the geographical coordinates of the neighborhoods using Python Geocoder package which will give us the latitude and longitude coordinates of the neighborhoods.

After that, we will use Foursquare API to get the venue data for those neighborhoods. Foursquare has one of the largest databases of 105+ million places and is used by over 125,000 developers.

Foursquare API will provide many categories of the venue data, we are particularly interested in the hotel category in order to help us to solve the business problem put forward. This is a project that will make use of many data science skills, from web scraping (Wikipedia), working with API (Foursquare), data cleaning, data wrangling, to machine learning (K-means clustering) and map visualization (Folium). In the next section, we will present the Methodology section where we will discuss the steps taken in this project, the data analysis that we did and the machine learning technique that was used.

Methodology

Firstly, we need to get the list of neighborhoods in the New York City. Fortunately, the list is available in the Wikipedia page (https://en.wikipedia.org/wiki/Category:New_York_City). We will do web scraping using Python requests and beautifulsoup packages to extract the list of neighborhoods data. However, this is just a list of names. We need to get the geographical coordinates in the form of latitude and longitude in order to be able to use Foursquare API. To do so, we will use the wonderful Geocoder package that will allow us to convert address into geographical coordinates in the form of latitude and longitude. After gathering the data, we will populate the data into a pandas DataFrame and then visualize the neighborhoods in a map using Folium package. This allows us to perform a sanity check to make sure that the geographical coordinates data returned by Geocoder are correctly plotted in the New York City.

Next, we will use Foursquare API to get the top 100 venues that are within a radius of 2000 meters. We need to register a Foursquare Developer Account in order to obtain the Foursquare ID and Foursquare secret key. We then make API calls to Foursquare passing in the geographical coordinates of the neighborhoods in a Python loop. Foursquare will return the venue data in JSON format and we will extract the venue name, venue category, venue latitude and longitude. With the data, we can check how many venues were returned for each neighborhood and examine how many unique categories can be curated from all the returned venues. Then, we will analyze each neighborhood by grouping the rows by neighborhood and taking the mean of the frequency of occurrence of each venue category. By doing so, we are also preparing the data for use in clustering. Since we are analyzing the “Hotel” data, we will filter the “Hotel” as venue category for the neighborhoods.

Lastly, we will perform clustering on the data by using k-means clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as Hotel as possible. It is one of the simplest and popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project. We will cluster the

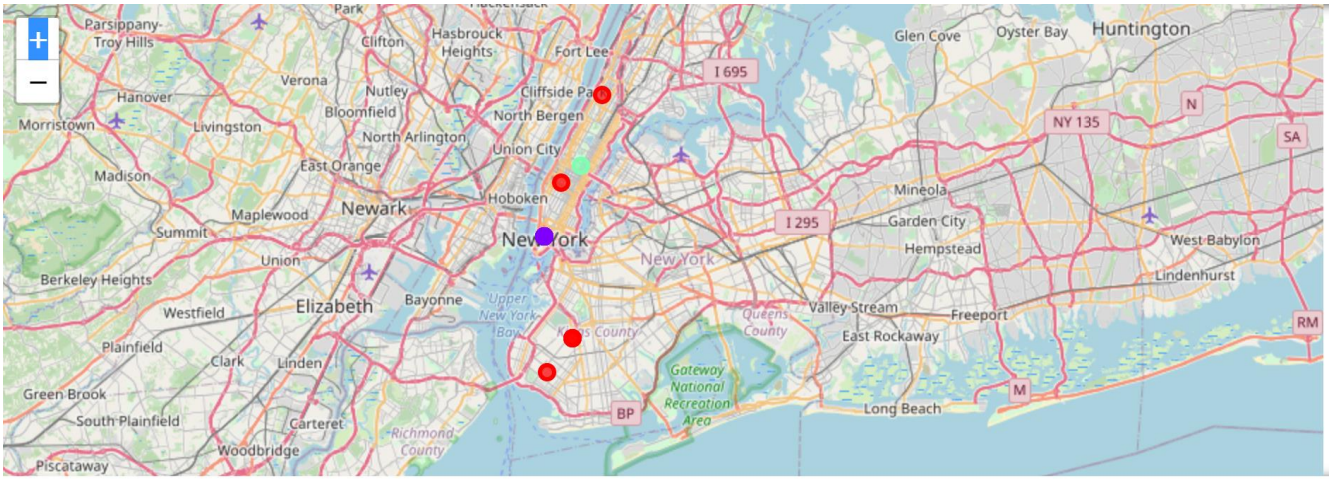
neighborhoods into 3 clusters based on their frequency of occurrence for “Hotel”. The results will allow us to identify which neighborhoods have higher concentration of Hotels while which neighborhoods have fewer number of Hotels Based on the occurrence of Hotels in different neighborhoods, it will help us to answer the question as to which neighborhoods are most suitable to open new Hotels.

Results

The results from the k-means clustering show that we can categorize the neighborhoods into 3 clusters based on the frequency of occurrence for “Hotel”:

- Cluster 0: Neighborhoods with low number to no existence of Hotels
- Cluster 1: Neighborhoods with moderate number of Hotels
- Cluster 2: Neighborhoods with high concentration of Hotels

The results of the clustering are visualized in the map below with cluster 0 in green color, cluster 1 in red color, and cluster 2 in purple color



Discussion

Most of the Hotels are concentrated in the central area of New York city, with the highest number in cluster 2 and moderate number in cluster 1. On the other hand, cluster 0 has very low number to totally no hotel in the neighborhoods. This represents a great opportunity and high potential areas to open new Hotels as there is very little to no competition from existing hotels. Meanwhile, hotels in cluster 2 are likely suffering from intense competition due to oversupply and high concentration of Hotels. From another perspective, this also shows that the oversupply mostly happened in the central area of the city, with the suburb area still have very few. Therefore, this project recommends property developers to capitalize on these findings to open new hotels in neighborhoods in cluster 0 with little to no competition. Property developers with unique selling propositions to stand out from the competition can also open hotels in neighborhoods in cluster 1 with moderate competition. Lastly, property developers are advised to avoid neighborhoods in cluster 2 which already have high concentration of hotels and suffering from intense competition.

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 clustering the data into 3 clusters based on their similarities, and lastly providing recommendations to the relevant stakeholders i.e. property developers and investors regarding the best locations to open a new Hotel. To answer the business question that was raised in the introduction section, the answer proposed by this project is: The neighborhoods in cluster 1 are the most preferred locations to open a new Hotel. 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 Hotel.

References

Category: New York City. Wikipedia. Retrieved from https://en.wikipedia.org/wiki/Category:New_York_City

Foursquare Developers Documentation. Foursquare. Retrieved from https://developer.foursquare.com/docs

Appendix

Cluster 0			Communications in New York City (1 C, 9 P)
Transportation in New York City (24 C, 22 P)			Former municipalities in New York (city) (...
New York City society (3 C)			Images of New York City (6 C, 2 F)
Buildings and structures in New York City ...			Education in New York City (18 C, 45 P)
Cluster 1			Wikipedia books on New York City (24 P)
Architecture in New York City (33 C, 2 P)			Healthcare in New York City (7 C, 34 P)
New York City stubs (4 C, 356 P)			Government of New York City (24 C, 81 P)
New York City templates (4 C, 106 P)			Geography of New York City (14 C, 8 P)
Organizations based in New York City			Environment of New York City (6 C, 10 P)
People from New York City (8 C, 645 P)			Economy of New York City (16 C, 35 P)
Politics of New York City (3 C, 8 P)	0.04	1	Demographics of New York City (1 C, 10 P)
Professional wrestling in New York City			Death in New York City (8 C, 4 P)
Sports in New York City (28 C, 29 P)			Culture of New York City (38 C, 245 P)
Tourist attractions in New York City			Boroughs of New York City (5 C, 12 P)
New York City-related lists (13 C, 66 P)			Books about New York City (3 C, 48 P)
New York City infrastructure (2 C, 13 P)			Mass media in New York City (7 C, 24 P)
Cluster 2			History of New York City (39 C, 244 P)