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 shopping mall. 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 shopping malls. 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.