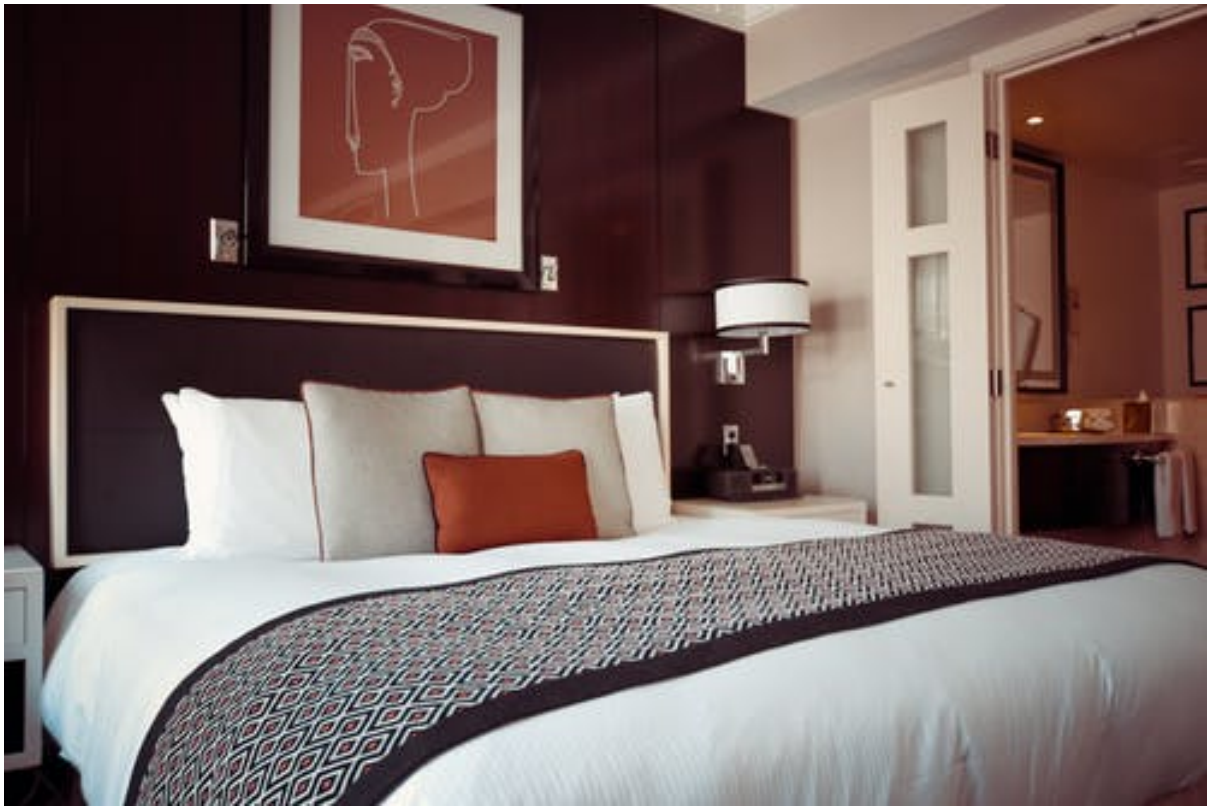
Sahil Sharma
Coursera Capstone
IBM Data Science Course
8 April 2020

Opening a New Hotel in New York



Opening a new business can be challenging. Choosing a location for the business can be even more challenging. Hotels are one of the most important thing in any neighborhood. When people visit any place, having a hotel at a convenient location is the best thing they can wish for. In this project, I will be creating a Machine Learning Algorithm which uses Foresquare's data and analyze where can a developer open a new hotel so that there is very little to no competition from existing hotels.

Business Problem

The objective of this project is to analyze and select the best locations in New York City where one can open a new Hotel. Using Data Science and Machine learning technology like Clustering, can we answer this simple question: Where to open a new Hotel in New York?

Data

To solve this challenge, I will be using the following data:

- List of neighborhoods in New York - This defines the scope of this project which is confined to New York City.
- Latitude and Longitude of the neighborhoods in New York City - This is required to visualize the places and the the venues.
- Venue data, Particularly related to the hotels in the neighborhoods.

Source of the data

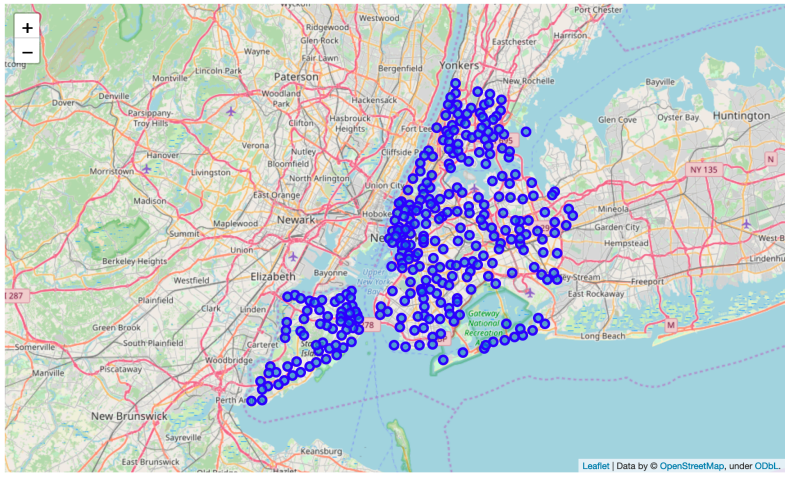
I will be using data from https://geo.nyu.edu/catalog/nyu_2451_34572 where we can find the List of neighborhoods in NY and other features as well. We will be only needing the neighborhoods and the boroughs data from it. The data is in the format of JSON so we will need to modify it ad well. Then we will get the Longitude and the Lattitude of the neighborhoods and use them.

From Foursquare, we can get the common areas nearby and check how many hotels are there in the neighborhood. Foursqaure has one of the largest database of places in the whole world and we can get a lot of information from it. Here we are only interested in the subcategory “Hotel”

Methodology

Firstly, we need the complete data of NY which includes the neighborhoods and the boroughs. After analyzing the dataset, few observations were made, such as, the data was in JSON format, it had other information also which was not required for the project. So we began the Data Preprocessing.

The very first step was to remove the unnecessary data. The only data which was required was in the *featured* key which was basically the list of the neighborhoods. So, we define a new variable that includes this data. After that, we had to transform the data into a *pandas* dataframe. So we create a new dataframe and define the columns. Then we iterate through our JSON data and put it in the pandas dataframe. We noticed that we had 5 Boroughs and 306 neighborhoods in the dataframe. Then we get the latitude and longitude values of New York City using geopy library.Then we create a map of New York with neighborhoods superimposed on top.



Let's simplify the above map and segment and cluster only the neighborhoods in Manhattan. So let's slice the original dataframe and create a new dataframe of the Manhattan data. Then we again create a map of Manhattan with neighborhoods superimposed on top.

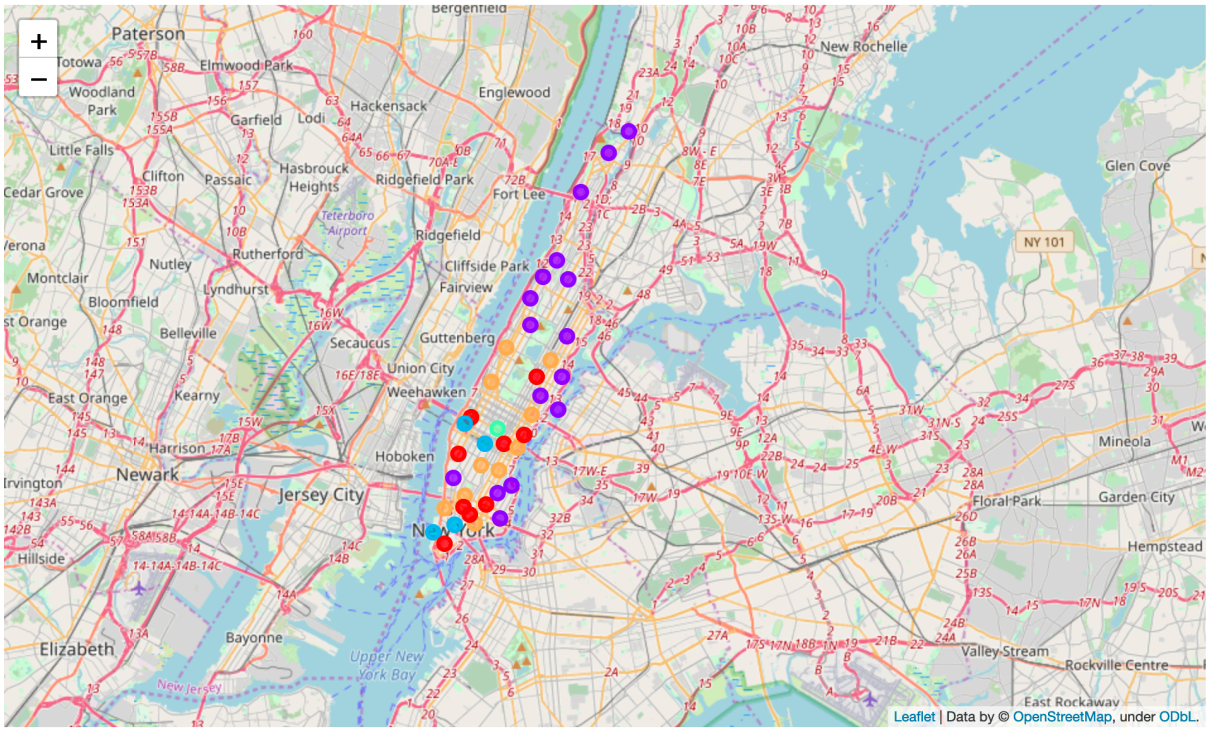
Then we define Foursquare Credentials and Version and use its api to get the top 100 venues that are in Marble Hill within a radius of 750 meters. Then we merge the data with the neighborhood data in a new dataframe.

Since we are concerned about the Hotel subcategory, we build a new dataframe from the above dataframe by grouping. Now we have a dataframe with the neighborhood and the hotels. Next we have to cluster the neighborhoods. We set the number of clusters to 5 and cluster all the neighborhoods and visualize it on a map.

Results

The results from the k-means clustering algorithm shows that we can categorize the neighborhoods:

- Cluster 1: Low number or no number of Hotels
- Cluster 2 and Cluster 3: Medium number of Hotels
- Cluster 0 and Cluster 4: High number of Hotels



From this we can figure out that Cluster 1 might be the best option to open a new Hotel in Manhattan, NY.

Discussion

From our observations, most of the hotels are in cluster 0 or cluster 4. On the other hand cluster 2 and cluster 3 have medium number of hotels and cluster 1 has the least or no number of hotels. It means that opening a new hotel in Cluster 1 would be a very good opportunity as there will be very less competition. Meanwhile Cluster 0 and Cluster 4 would be facing a lot of competition because of the oversupply and high concentration of the the hotels. There might be a possibility that there are more number of hotels in Cluster 0 and Cluster 4 because it is the center part of the borough.

Also, in this project we have only considered one factor which is occurance of Hotels. There can be so many other factors such as Airport distance or Famous places nearby and more. We can build a more better model in future considering all those things in mind.

In this model, the k in k-means clustering algorithm has been chosen randomly. We can use use the proper way to find the key which might result in better results.

Conclusion

After going through a business problem, specifying the data required and preparing it, and performing a clustering algorithm on the data, the best place to open a new hotel in Manhattan, NY is in Cluster 1 as there are less or no number of hotels in the area so there will be less competition.