Finding the best place to open a Chinese Restaurant in Manhattan, NY (Report)

Raafat El Jamal

12-Mar-2021

1 Introduction

1.1 Background

Manhattan is a great place for starting a new business as many tourists visit it frequently as it has almost everything a tourist might need such as shops, restaurants, historical places, entertainment, and hotels. It has a strong economy and high population density. There is a high competition between business owners which lets anyone think about starting business there. In this project the investor is wants to start a Chinese Restaurant. It is not easy to find the best place in Manhattan and that is what we will work on.

1.2 Problem

I think the best way is to look at Neighborhoods that do not have Chinese restaurants so that this will be the first Chinese restaurant there which gives an advantage for a good start with less competition, then the highest rated venues there, then select the densest place which gives another advantage as this location is highly demanded. This is how to keep squeezing the search to find the golden location.

1.3 Interest

Any investor would be interested in the results as it saves a lot of research time.

2 Data

2.1 Data Sources

The main dataset is the New York data with all Boroughs and Neighborhoods which can be found here. I extracted Manhattan's data from it then from Foursquare API I explored all Neighborhoods in Manhattan to get Venues' names, categories, IDs, longitudes, and latitudes. After that I extracted Venues' ratings from Foursquare API to build the final table required for analysis.

2.2 Data Cleaning

New York data were extracted from the json file and converted to a dataframe, then I took specific columns (Borough, Neighborhood, Latitude, Longitude). Manhattan data were extracted from the dataframe, then connected to Foursquare API to explore Neighborhoods in Manhattan. Downloaded the json file from Foursquare API then extracted from it (Venue name, category, ID, longitude, latitude) and added them to the Manhattan dataframe as additional columns. I deleted all the Neighborhoods that have Chinese restaurants in it and kept the rest to explore, then based on the Venue ID I extracted Venue Ratings from Foursquare for each venue and added the ratings to the table. This operation was complicated as Foursquare limits API calls to ratings per day, so I divided a copy of the main table into three parts and ran the API call for each part in a separate day, then merged the three ratings results in one file and took the ID and rating columns only and removed the duplicates, then merged the ratings to the main table again and removed venues with empty ratings. Finally, the dataframe was ready for analysis.

2.3 Data Readiness

The final table has Manhattan venues with their coordinates, categories, neighborhoods and ratings, these data are required for my analysis to find the golden location for starting a new Chinese restaurant.

3 Methodology

3.1 Exploratory Data Analysis

The idea is to open the restaurant in a neighborhood that does not have any Chinese restaurant, so I searched for neighborhoods that have Chinese restaurants, collected the neighborhoods in a set then removed all rows related to these neighborhoods, the resulted dataframe was a list of venues in

neighborhoods without Chinese restaurants. After that I called venue ratings from Foursquare API as explained above. In the ratings dataframe there are duplicates because some venues have branches in different neighborhoods with the same venue IDs and ratings, so I removed the duplicates. I merged the ratings with the main dataframe. Some venues do not have ratings, so I removed them to refine the data for better evaluation. I calculated the mean value (Average) of ratings per neighborhood to see which neighborhoods have the highest ratings and found that Soho, has the highest rating. From the main dataframe I created another one with Soho data (Only Soho neighborhood), and then started analyzing each region to find the golden place for the restaurant.

3.2 Clustering

I used the DBSCAN clustering to see how venues are close to each other and the rating similarity to magnify the most competitive regions. DBSCAN clusters groups with similar characteristics and closeness and shows the outlier who do not belong to the group. I labeled the venues to indicate the outliers and check the cluster performance to see how data is centralized, then normalized data between 0 and 1, then calculated the average of each group to have a general idea. Finally, I visualized the clusters using folium to see the clusters in different colors, analyze them and find the best place based on colors and sizes, then plotted labels versus average rating to visualize the categories of the top cluster.

4 Results

In the analysis, I have clustered the venues based on their longitude, latitude and rating using DBSCAN. 5 clusters were created, and one is an outlier cluster. I found that cluster 4 (Red) located in the north has the highest average rating and is the densest. Cluster 3 (Orange) located in the east has a high average rating but less in density. The rest of clusters have low average rating. Cluster 2 (Light Green) located in the south has very low density. Cluster 0 (Blue) located in the north is the largest. The outcome is that cluster 4 is the golden place for the restaurant as it has the highest average rating and is the densest, it is composed of a Snack Place, Clothing Store, Cosmetics Shop and a Coffee Shop which makes the place suitable as these are common daily venues.

5 Discussion

During my analysis, I found that in Foursquare all the branches of any company or brand have the same rating which makes the analysis less accurate than desired, so I recommend rating every branch separately to increase the accuracy of calculations.

I also recommend enriching the API by rating more venues for more accuracy.

6 Conclusion

The purpose of this project was to analyze different parts of Manhattan, to understand the quality of each region and detect homogenous clusters and outlier based on the ratings. This aids me to make a final decision about opening a new Chinese restaurant. By clustering the venues using latitude, longitude, and rating from Foursquare data, I was able to detect the clusters within each part of the city and calculate an average rating for each cluster, as a result giving the ability to have a generic view of the city's quality of venues and an insight regarding the quality, luxury, and expectation of customers for each region.

Final decision on the optimal venue location was made based on specific characteristics which are rating and density in neighborhoods that are free of Chinese restaurants.