IBM Applied Data Science Capstone

Latino Restaurant opportunity in Manhattan, New York, NY

By: Rodrigo Martins

May 2020

Introduction

In this Capstone project, I am creating a hypothetical scenario for an entrepreneur who wants to explore opening a Latino Restaurant in Manhattan, New York, NY, such as Brazilian, Peruvian, etc. The idea behind this project is that there may not be enough Latino Restaurants in Manhattan, and it might be a great opportunity for this entrepreneur since there are a lot of Hispanic / Latinos living in this city. It would be better if the entrepreneur opens this kind of restaurant in locations where Latino food is popular. Finding the best location to open a restaurant is one of the most important decisions for an entrepreneur, and because of that this project has an objective of help him to find the most suitable location.

- **Business Problem**

The objective of this capstone project is to analyze and select the best location in Manhattan, New York, NY to open a Latino Restaurant. This project will use data science methodology and machine learning algorithms like clustering to help an entrepreneur solve the following question: Where would be the best locations to open a Latino Restaurant in Manhattan, New York?

- **Target Audience**

This project is particularly useful to entrepreneurs looking to open or invest in a new Latino Restaurant in Manhattan, New York since there is a big community of Latinos living there. According to the data released on Wikipedia, Manhattan, New York is populated by 27.5% of Hispanic or Latino, or approximately 2,287,905 people from this ethnicity, which give a good opportunity to entrepreneurs who want to open a Latino Restaurant and explore the lack of Latino cuisine in the city.

Data

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

I. List of neighborhoods in Manhattan, New York. This defines the scope of this project.

II. Latitude and longitude coordinates of those neighborhoods. This is required to plot the map and to get the venue data.

III. Venue data, particularly data related to restaurants. We will use this data to perform clustering on the neighborhoods.

- **Extracting the data**

I. This webpage (https://geo.nyu.edu/catalog/nyu_2451_34572) contains a list of neighborhoods in Manhattan, New York. We will download the data in a json file format, and use Python packages such as Pandas, to clean the data.

II. The geographical coordinates of the neighborhoods will be obtained by cleaning the json file, which already contains the latitude and longitude coordinates of the neighborhoods, using geocoder to obtain more specific locations.

III. Moreover, we will use the Foursquare API to get the venue data for all neighborhoods.

Methodology

The first thing to do is getting the list of boroughs and neighborhoods in New York, NY. This is possible by downloading the JSON file that contains the list of all neighborhoods with many information regarding each neighborhood in New York, NY, and then clean the file to obtain a data frame with all neighborhoods organized in a table. Link of data: https://geo.nyu.edu/catalog/nyu_2451_34572. After that, as I wanted only data from Manhattan, I had to clean the data to a data frame which contains only neighborhoods of Manhattan.

Since the file is formatted in a JSON file, I used Pandas library to clean the data and return a data frame containing only the boroughs, neighborhoods, latitude and longitude of each neighborhood in Manhattan. The result was obtained the following data frame:

```
[14]
manhattan_data = neighborhoods[neighborhoods['Borough'] == 'Manhattan'].reset_index(drop=True)
manhattan_data.head()
```

	Borough	Neighborhood	Latitude	Longitude
0	Manhattan	Marble Hill	40.876551	-73.910660
1	Manhattan	Chinatown	40.715618	-73.994279
2	Manhattan	Washington Heights	40.851903	-73.936900
3	Manhattan	Inwood	40.867684	-73.921210
4	Manhattan	Hamilton Heights	40.823604	-73.949688

After gathering all that information, I visualized the map of Manhattan using Folium package to verify whether these are correct coordinates.

Next, I used Foursquare API to pull the list of top 100 venues within 500 meters radius. This could be done by creating a Foursquare developer account to obtain client ID and client secret which will be used to pull the data. The result data frame is the following:

```
[19]
manhattan_venues_df = pd.DataFrame(venues)

manhattan_venues_df.columns = ['Neighborhood', 'Latitude', 'Longitude', 'Venue', 'Venue Latitude', 'Venue Longitude', 'Venue Category']

print(manhattan_venues_df.shape)
manhattan_venues_df.head()
```

(3093, 7)

	Neighborhood	Latitude	Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Marble Hill	40.876551	-73.91066	Arturo's	40.874412	-73.910271	Pizza Place
1	Marble Hill	40.876551	-73.91066	Bikram Yoga	40.876844	-73.906204	Yoga Studio
2	Marble Hill	40.876551	-73.91066	Tibbett Diner	40.880404	-73.908937	Diner
3	Marble Hill	40.876551	-73.91066	Starbucks	40.877531	-73.905582	Coffee Shop
4	Marble Hill	40.876551	-73.91066	Dunkin'	40.877136	-73.906666	Donut Shop

Moreover, I had to analyze each neighborhood in Manhattan by grouping the rows by neighborhood and take the mean on the frequency of occurrence of each venue category. The data was being prepared to be clustered in the sequence.

Then I used One-Hot Encoding to transform the categorical values into a numerical value that could be calculated by the algorithm when finding the best result of the clusters.

As I was looking data for Latino Restaurants, I created another data frame with the data containing the mean of Latino Restaurants by each neighborhood in Manhattan:

```
[36]
manhattan_latin_rest = manhattan_grouped[["Neighborhood","Latin American Restaurant"]]
manhattan_latin_rest
```

	Neighborhood	Latin American Restaurant
0	Battery Park City	0.000000
1	Carnegie Hill	0.000000
2	Central Harlem	0.000000
3	Chelsea	0.000000
4	Chinatown	0.000000

Finally, I used the k-means clustering algorithm to replicate the answer for the business problem. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the

simplest and popular unsupervised machine learning algorithms and it is highly suited for this project as well. I used the cluster to visualize the neighborhoods in Manhattan into 5 clusters based on their frequency of occurrence for “Latin American Restaurant”. According to the concentration of cluster in the following figure, I can recommend the ideal location to open the Latino Restaurant.

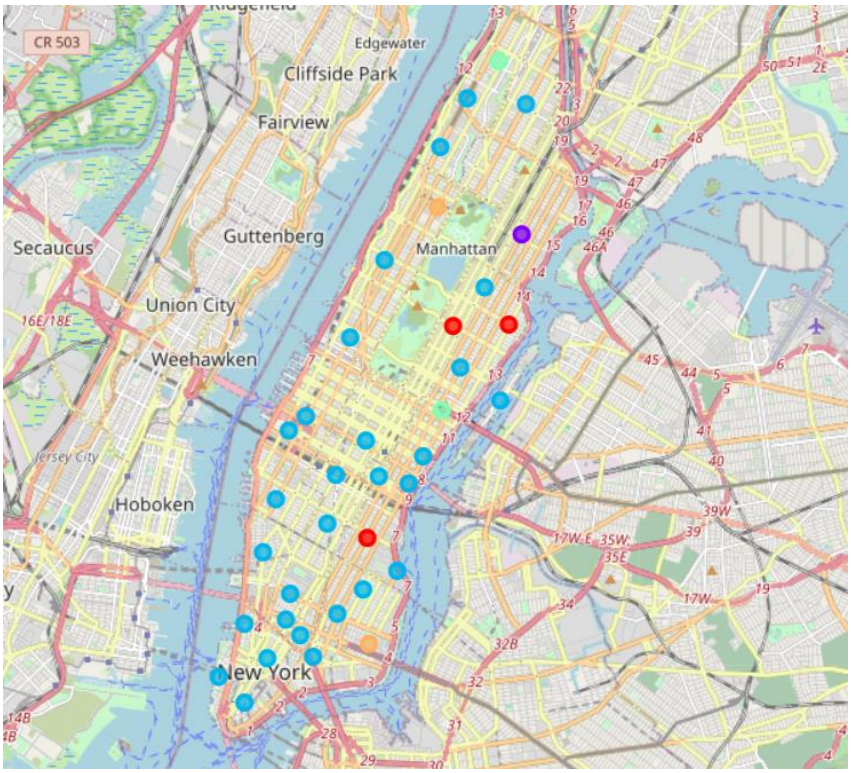
```
[44]  M↓
print(manhattan_merged.shape)
manhattan_merged.sort_values(["Cluster Labels"], inplace=True)
manhattan_merged
```

```
(40, 6)
```

	Neighborhood	Latin American Restaurant	Cluster Labels	Borough	Latitude	Longitude
39	Yorkville	0.010000	0	Manhattan	40.775930	-73.947118
35	Upper East Side	0.011236	0	Manhattan	40.775639	-73.960508
11	Gramercy	0.011905	0	Manhattan	40.737210	-73.981376
7	East Harlem	0.073171	1	Manhattan	40.792249	-73.944182
22	Marble Hill	0.000000	2	Manhattan	40.876551	-73.910660
23	Midtown	0.000000	2	Manhattan	40.754691	-73.981669
24	Midtown South	0.000000	2	Manhattan	40.748510	-73.988713
25	Morningside Heights	0.000000	2	Manhattan	40.808000	-73.963896

Results

With the image of the cluster, I had analyzed it:



The results obtained with k-means show Manhattan neighborhoods divided into 5 clusters based on how many Latino Restaurants are in each neighborhood. As we can see in the figure, where Cluster 0 appears in red color, Cluster 1 in purple color, Cluster 2 in blue color, Cluster 3 in green and Cluster 4 in orange:

a) Cluster 0: Neighborhoods with a few Latino Restaurants.

b) Cluster 1: Only one neighborhood with high concentration of Latino Restaurants.

c) Cluster 2: Neighborhoods with zero Latino Restaurants.

d) Cluster 3: Neighborhoods with many Latino Restaurants.

e) Cluster 4: Neighborhoods with high concentration of Latino Restaurants.

Recommendations

According to the result, I may say that most of the Latino Restaurants located in Manhattan is around the Grand Central Park, in Washington Heights or Lower East Side, and Hamilton Heights or Sutton Place, for example, as we could see in clusters 3 and 4.

It is not a good idea opening a Latino Restaurant in the Downtown or near the Brooklyn bridge. As we could see in cluster 2, there is almost zero concentration of Latino Restaurant in those area, which could show low interest by people that living or working in that area. The same with cluster 1 where appears only one neighborhood with a high concentration of Latino Restaurants, which is not a good idea because of concurrency.

It seems Cluster 0 might be a good location as well, since there are few Latino Restaurants in these areas.

Conclusion

To conclude, I have worked in identifying the business problem, specifying the data required to solve the business problem, extracting, preparing and cleaning the data, utilizing one of the machine learning algorithms to find the best solution for the problem, and in this case I have used k-means clustering, and providing recommendation to the entrepreneurs who might want to open a Latino Restaurant in Manhattan, NY.