IBM Applied DATA SCIENCE CAPSTONE

The Mexican Boys Are Back In Town!!!!!!

Sylvain Carlevato July 2021-07-23

I Introduction:

New York City's demographics show that it is a large and ethnically diverse city.

It is the largest city in the USA with a long history of International Immigration.

New York City was home to nearly 8 million people in 2021, accounting for over 40% of the population of New York State and a slightly lower percentage of the New York metropolitan area, home to approximately 23.6 million.

Over the last decade, the city has been growing faster than the region. The New York region continues to be by far the leading metropolitan gateway for legal immigrants admitted into the United States.

My final project explores the best locations for Mexican restaurants throughout the city of New York. Potentially the owner of the New Mexican restaurant can have great success and consistent profit.

However, as with any business, opening a new restaurant requires serious considerations and is more complicated than it seems from first glance. In particular, the location of the restaurant is one of the most important factors that will affect whether it will have success or a failure.

So our project will attempt to answer the questions:

"Where should the investor open a Mexican Restaurant?"

"Where should I go if I want great Mexican food?"

II Business Problem:

The objective of this Capstone project is to analyze and select the best locations in the city of New York to open a New Mexican restaurant.

Using Data Science Methodology and Instruments such as Data Analysis and Visualization, this project aims to provide solutions to answer the business question:

Where in the city of New York, should the investor open a Mexican Restaurant?

III Target Audience Of This Project:

This project is particularly useful to developers and investors looking to open or invest in a Mexican restaurant in the city of New York.

Overall, New York is a great place to open a restaurant with ethic cuisine.

New York is the most diverse city in the world (800 languages are spoken in New York).

With its diverse culture, comes diversity in the food items. There are many restaurants in New York City, each belonging to different categories like Chinese, Indian, Japanese, etc.

Why did we decide to focus on Mexican cuisine in our project? Now when the idea of a healthy lifestyle conquered the minds of people all over the country, Mexican restaurants became extremely popular, as they offer a healthy alternative to regular American eating habits.

IV Data:

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

· New York City data containing the neighborhoods and boroughs.

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

· Venue data, particularly data related to restaurants. We are going to use this data to perform further analysis of the neighborhoods.

New York City data containing the neighborhoods and boroughs will be obtained form the open data source: https://cocl.us/new_york_dataset. After it, we will get the geographical coordinates of the neighborhoods (latitude and longitude) using Python Geocoder package.

Finally, we will use Foursquare API to get the venue data for the neighborhoods defined at the previous step. Foursquare has one of the largest databases of 105+ million places and over 125,000 developers use this application. Foursquare API provides many categories of the venue data; we are particularly interested in the restaurant data to solve the business problem defined above.

This project will require using of many data science skills, from web scrapping (open-source dataset), working with API (Foursquare), data cleaning, data wrangling, to map visualization (Folium). In the next Methodology section, we will discuss and describe any exploratory data analysis that we did, any inferential statistical testing that we performed, and what machine learning techniques were used.

V Methodology:

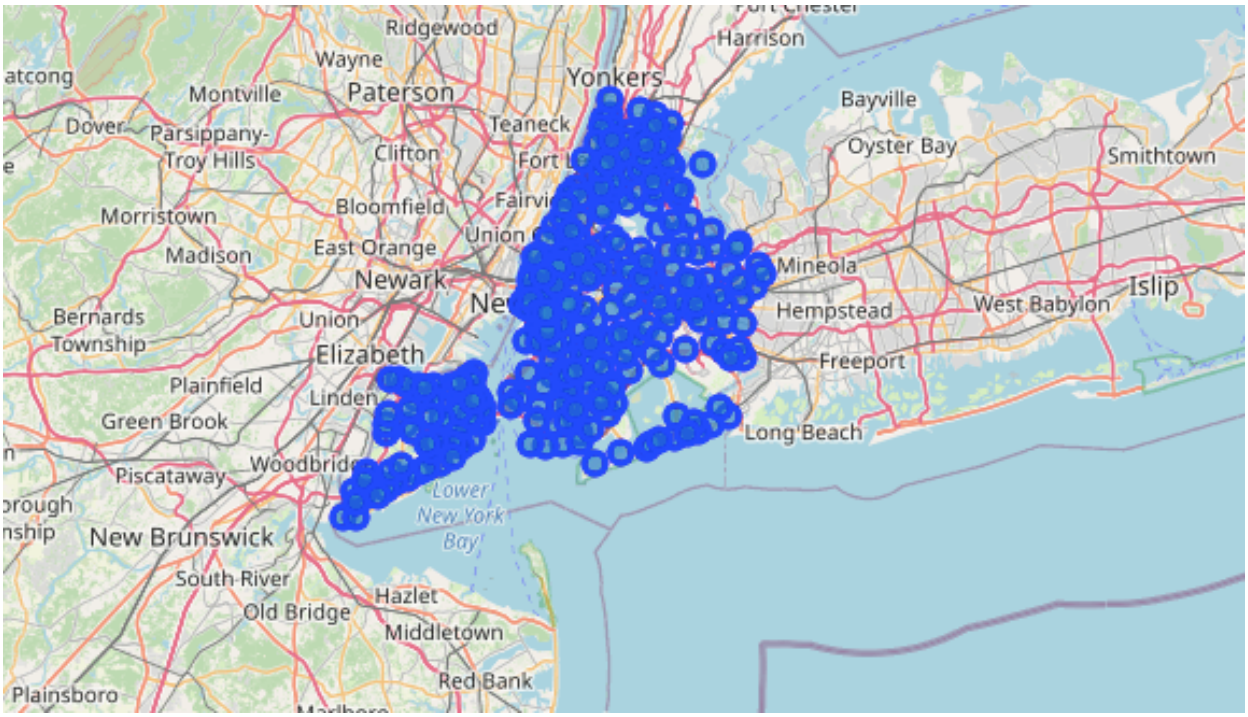
The data that was obtained from the "new_york.json" data is transformed into pandas dataframe that contains the borough and neighborhoods of New York City along with it coordinates.

These coordinates are plotted using Folium Map to see how the neighborhoods are distributed over the city.

Folium is a powerful Python library that helps create several types of Leaflet maps. The fact that the Folium results are interactive makes this library very useful for dashboard building.

By default, Folium creates a map in a separate HTML file but in case of using Jupyter Notebook inline maps can be created.

	Borough	Neighbourhood	Latitude	Longitude
301	Manhattan	Hudson Yards	40.756658	-74.000111
302	Queens	Hammels	40.587338	-73.805530
303	Queens	Bayswater	40.611322	-73.765968
304	Queens	Queensbridge	40.756091	-73.945631
305	Staten Island	Fox Hills	40.617311	-74.081740



VI Foursquare API:

As mentioned earlier since Foursquare API already has a well-maintained collection of data once the data is imported all that had to be done is to extract the necessary features into a pandas dataframe.

The data is imported using the Foursquare API request URL, a URL that is created by us in a specific format to import the required data. The Foursquare API has lots of data on a location but for this analysis only the venue data (shops) is imported. A python function is defined to import all the data using the Foursquare API and append it to a dataframe.

	Neighbourhood	Neighbourhood Latitude	Neighbourhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	Venue Category Id	Venue Id
0	Astoria	40.768509	-73.915654	Favela Grill	40.767348	-73.917897	Brazilian Restaurant	4bf58dd8d48988d16b941735	4bdf502a89ca76b062b75d5e
1	Astoria	40.768509	-73.915654	Orange Blossom	40.769856	-73.917012	Gourmet Shop	4bf58dd8d48988d1f5941735	52c580e8498eddd52d925dd9
2	Astoria	40.768509	-73.915654	Titan Foods Inc.	40.769198	-73.919253	Gourmet Shop	4bf58dd8d48988d1f5941735	4a9c0105f964a520b03520e3
3	Astoria	40.768509	-73.915654	CrossFit Queens	40.769404	-73.918977	Gym	4bf58dd8d48988d176941735	4c94d26d58d4b60c40fc2b29
4	Astoria	40.768509	-73.915654	Off The Hook	40.767200	-73.918104	Seafood Restaurant	4bf58dd8d48988d1ce941735	514f9fd5e4b023ae1edd4a68

Since the analysis is to determine the location for a Mexican restaurant only those venues are filtered using the category id for Mexican Restaurants that was assigned by Foursquare API.

Mexican Restaurants based on New York City

VII Machine Learning:

One hot encoding:

It is a process by which categorical variables are converted into a form that could be provided to ML algorithms to do a better job in prediction. Hence dataframe that is shown above is one hot encoded to convert all the categorical values to numerical values (0 & 1) for further analysis using clustering.

K mean clustering:

K-means clustering is one of the simplest and popular unsupervised machine learning algorithms. A cluster refers to a collection of data points aggregated together because of certain similarities. The value of *k* is initially defined, which refers to the number of centroids needed in the dataset.

A centroid is the imaginary or real location representing the center of the cluster.

Every data point is allocated to each of the clusters through reducing the in-cluster sum of squares.

In other words, the K-means algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible.

The 'means' in the K-means refers to averaging of the data; that is, finding the centroid.

How K-means forms cluster:

1. K-means picks k number of points for each cluster known as centroids.
2. Each data point forms a cluster with the closest centroids i.e. k clusters.
3. Finds the centroid of each cluster based on existing cluster members. Here we have new centroids.
4. As we have new centroids, repeat step 2 and 3. Find the closest distance for each data point from new centroids and get associated with new k-clusters. Repeat this process until convergence occurs i.e. centroids does not change.

For this scenario, we fit the model using the value of *k=5* and determine how all the Mexican restaurants are formed into clusters. The clusters are then plotted using a folium map to visualize how they are spread out across the city.

VIII RESULT:

From the above map we can decide that the new location to open a Mexican restaurant would be in one of the three sparsely populated clusters as each of them have only one Mexican restaurant in the neighborhoods. These neighborhoods are Hunters Point, Cambria Heights and Belle Harbor.

IX DISCUSSION:

It is important to mention that these results in real life might not be ideal as we did not include so many other factors as the article was focused only on learning and not the accurate results. The scope of this analysis can be further extended by including those various other factors that was omitted such as how easy is to access the location from other parts of the city especially from other popular spots, are there any particular culture thriving in that location, is it easy to procure fresh supplies from that location and so on. Also Elbow method can be used to determine the optimal value of *k* for clustering.

X CONCLUSION:

The article described the process involved in segmentation & clustering and introduced how to use Foursquare API to obtain location data. It showed how the data is imported and transformed to be used for clustering. The final results are the clusters from which we can interpret the necessary information and the scope of this entire article can be increased by including more data and performing an in-depth analysis.