Coursera – IBM Data Science Professional Certification

Applied Data Science Capstone

Opening a Bakery/Patisserie in New York

By

Preetha Venkatasamy

January, 2021

Introduction & Background

New York is the most ethnically diverse, commercially driven, and the most attractive urban centre in the country. New York always meant possibility, for it was an urban centre on its way to something better, a metropolis too busy to be solicitous of those who stood in the way of progress. New York has remained prosperous even as it underwent change, its strength lying in its diversity. New York still continues to welcome many newcomers into the city's "golden door" to this day.

Every New York City neighborhood has one: A pastry shop where the air smells sweet and the coffee is hot. Open for breakfast, pastry shops serve warm cinnamon rolls or many-layered cherry danishes, tall biscuits, or soft muffins studded with berries. They're almost always open all day, too, so they also attract those looking for a sweet fix just before the sun starts to set.

The investor was fascinated with the city's café and pastry culture but noticed that it was all so focused on trends from people to the décor and even the food lacked a warm, welcoming environment - a homelike feeling. With the idea to offer a unique experience and a welcoming place to eat, drink and gather with friends, starting a patisserie in New York City seemed very challenging. As with any food and beverage industry, finding the best location is one of the key factors that will determine its success or failure.

Choosing a location is one of the more permanent choices a restaurant owner makes. You cannot move without significant expense and trouble. So, making a snap decision without doing any research may leave an owner with a location he or she may later regret.

Business Problem

This project aims to find the best locations in New York City to open a Patisserie that would attract the culturally diverse population. It should have great visibility, should be easy to find and should attract enough initial customer interest.

With the help of Data Science methodology and tools, this project aims at addressing the business problem by helping the investor find the best locations in New York City, to start the business keeping in mind all the essential factors.

Data Requirement

I will need the following data to address the business question:

1) The entire New York City has a total of 5 boroughs and 306 neighborhoods. In order to segment the neighborhoods and explore them, I will essentially need a dataset that contains the 5 boroughs and the neighborhoods that exist in each borough
2) The latitude and longitude coordinates for each neighborhood

3) Possibility to explore each neighborhood with venue categories relevant to bakery/pastry shops. This data will be used for further analysis and modelling to decide the best locations

Data Source and Extraction

New York city data containing the neighborhoods and boroughs was obtained from the open data source: https://cocl.us/new_york_dataset. Once this data is obtained in the required format, the corresponding latitude and longitude co-ordinates for each neighborhood can be obtained using the python Geocoder package.

I used the Foursquare API to explore neighborhoods in New York City. Here, we will construct a URL to send a request to the API to search for a specific type of venue that is relevant to bakery in each neighborhood, and then use this feature to group the neighborhoods into clusters. We will use the *k*-means clustering algorithm to complete this task.

Finally, we will use the Folium library to visualize the neighborhoods in New York City and their emerging clusters.

Methodology

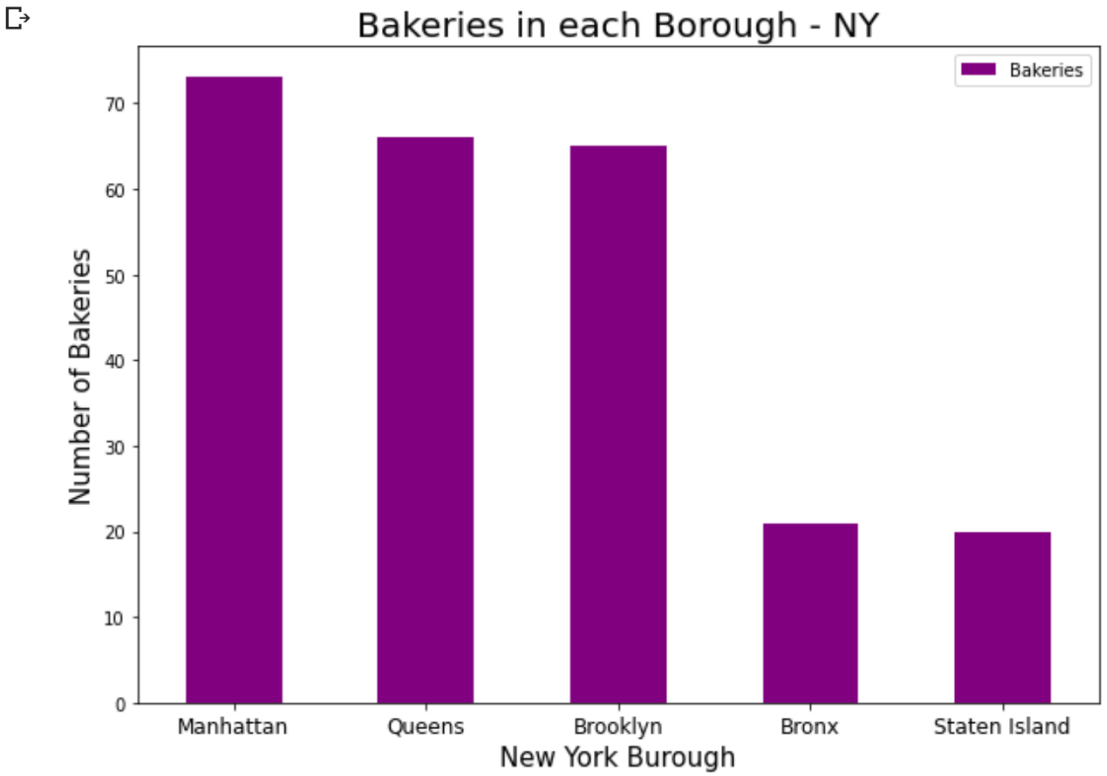
1) Collected New York data from https://cocl.us/new_york_dataset that has all the 5 boroughs and 306 neighborhoods
2) Accessed the required co-ordinates using python Geocoder package
3) Once all the boroughs and neighborhoods and corresponding latitude and longitude coordinates were obtained, I explored and cleaned the dataset.
4) Accessed Foursquare API to extract venues for all New York neighborhoods. Analyzed the venues specific to Bakery and located the most popular Boroughs and Neighborhoods with bakeries
5) Selected the most popular Borough for further analysis and clustering using k-Means algorithm

6) Prepared data for clustering by one-hot encoding, grouping and sorting the 10 most popular venues from each neighborhood in the chosen Borough
7) Used KMeans clustering algorithm to cluster the prepared data into 3 clusters and joined the resulting cluster labels with the neighborhoods data table showing their top 10 venues
8) Explored the resulting clusters with similar characteristics and chose the neighborhoods that best answers the business questions
9) Visualized data using maps and plot as needed throughout the project

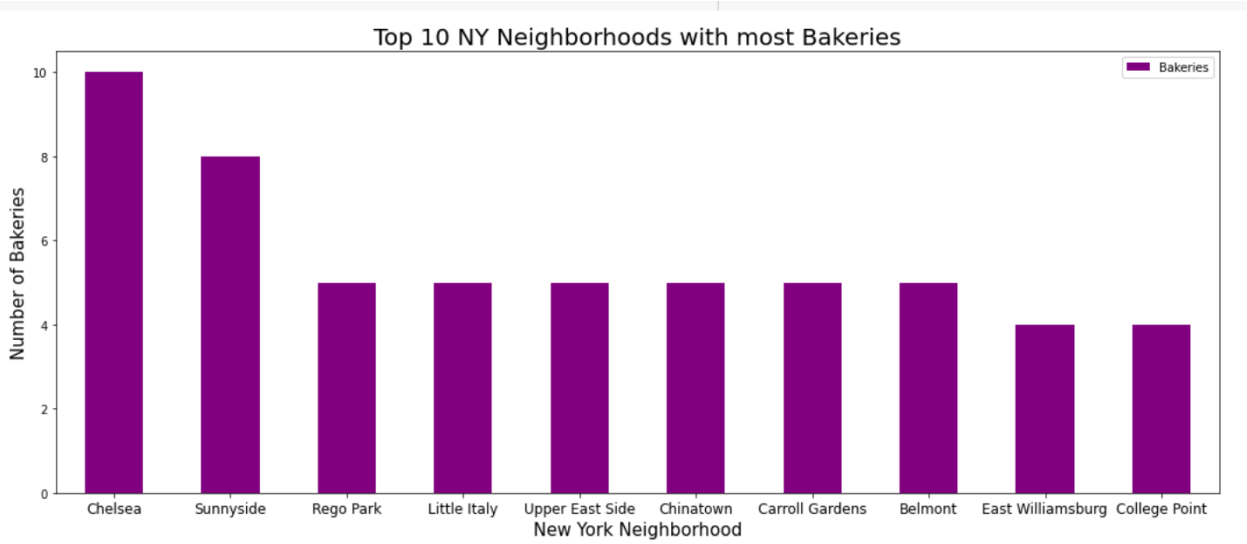
Results

I have shown all the results below based on which conclusions were drawn.

1) Based on the visualizations below, I chose Manhattan as the best Borough to open the new Bakery as it had most number of bakeries and was popular for such food.



2) Chelsea has the most number of bakery related venues and it is located in Manhattan, NY.

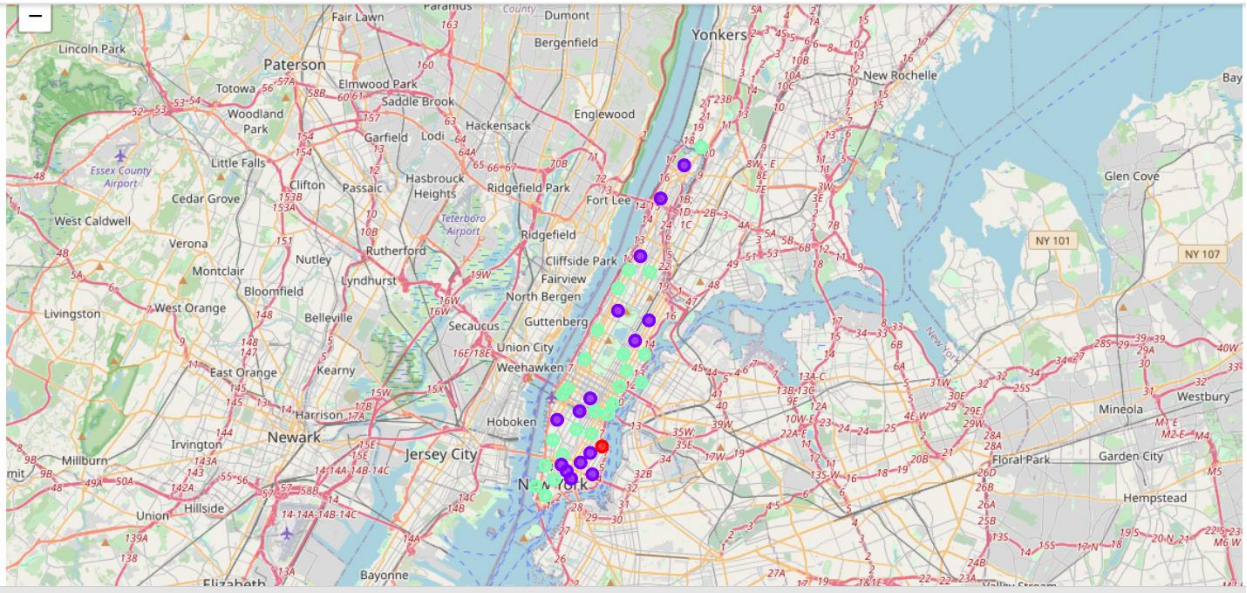


3) Based on above analysis, Manhattan is one of the best Boroughs to start a bakery. For further analysis, I am using KMeans Clustering algorithm to find similar neighborhoods in Manhattan to start a bakery. I am looking for a location that will answer the following business questions.

   - Must attract enough initial customer interest
   - Must have great visibility and should be easy to find

4) Visual representation of the resulting clusters. Further analyses of ach cluster results in the following:

   - Cluster 1 (red color) - **NOT a preferred cluster** to open a new Bakery/Patisserie. Does not answer the business questions as compared to Cluster 2 and 3.
   - Cluster 2 (green color) - Seems like the **MOST preferred cluster** to start a new Bakery/Patisserie. It has a lot of restaurants and bakeries that are among the top most venues in these Neighborhoods.
   - Cluster 3 (purple color) - Although this cluster has many restaurants among its top venues, it does not look as popular compared to cluster 2. As a result, this is **NOT a preferred cluster** to start a new Bakery/Patisserie.



Future enhancements

This data and model can be used for more detailed and comprehensive analysis in the future to find **the BEST neighborhood in Cluster 2** to open a new Patisserie/Bakery.

Conclusion

Cluster 2 in Manhattan has the **MOST preferred neighborhoods** with many restaurants and bakeries among the top 10 venues. Starting a Patisserie here will yield great success and consistent profit due to high visibility, easy to locate and ability to attract enough initial customer interest. As a final note, all of the above analysis is dependent on the accuracy of Foursquare data.