IBM Applied Data Science Capstone

Capstone Project - The Battle of Neighborhoods

Opening a bakery in Bronx, New York

Author: Alexandros Vellios

Introduction

With an estimated population (2018) of 8,398,748 distributed over about 302.6 square miles New York is also the most densely populated major city in the United States. It is also the most populous city in the United States.

The city has been described by many as the cultural, financial, and media capital of the world, significantly influencing commerce, entertainment, research, technology, education, politics, tourism, art, fashion, and sports.

Given how densely populated and economically developed the city is, space is limited and the cost of doing business high. However the big population makes the area very appealing for new businesses, and the rewards high if one strategically plan the business venture.

In this project we will work for a client who is looking to open a bakery in Bronx, NYC after he identified a gap in the market around a neighborhood of New York.

Business Problem

Our client, an independent business owner, has come to us to advise them regarding which neighborhood within Bronx in NYC is the best location to open a bakery.

According to our client we should focus on two features to find the best location:

1. Lack of competition (bakeries)
2. Presence of entertainment businesses near the location

In order to find the best location we will segment and cluster the different neighborhoods of Bronx and come up with the best possible location for our client to open his business.

Target Audience

The target audience of this report will be our client, who wants to open a bakery in NYC

Data Selection

For this project we will utilise two data sources. The first data source (https://geo.nyu.edu/catalog/nyu_2451_34572) has a total of 5 boroughs and 306 neighborhoods. It also contains the latitude and longitude coordinates of each neighborhood. Example of the data:

	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

Furthermore, using the geographical coordinates data as input for the Foursquare API, we will extract venue information for each neighborhood.Therefore, we will use the Foursquare API to explore neighborhoods in New York City. The below is an example of the Foursquare API data:

	Neighborhood	Neighborhood Latitude	Neighborhood 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

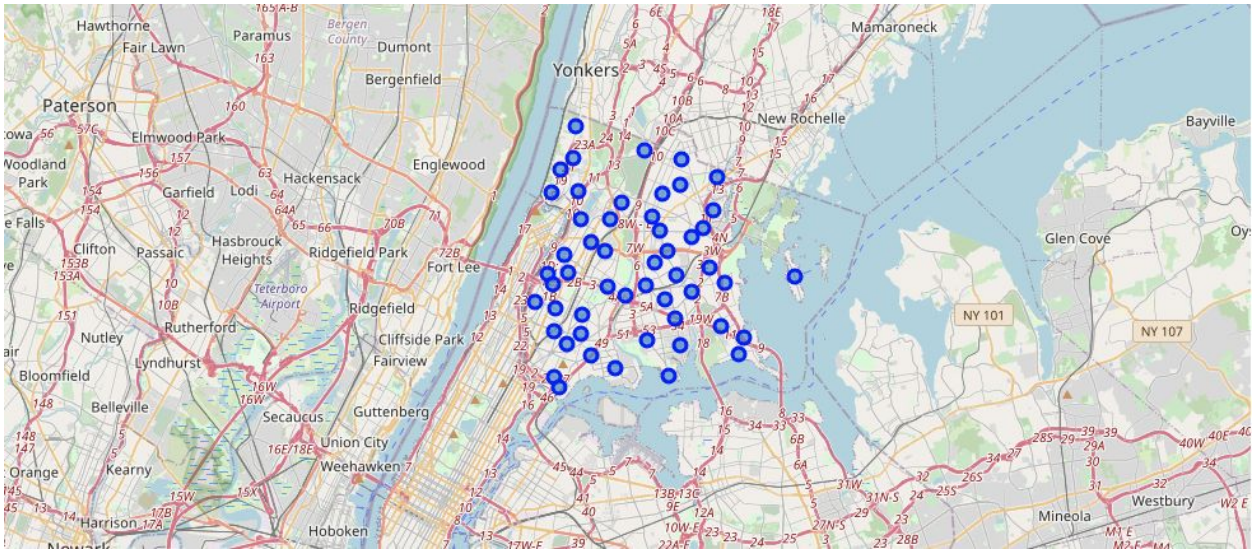
With data collected from the two sources above we will be able to tackle our client's issue.

Methodology

Our goal is to find the best location to open a bakery in Bronx for our client.

Exploratory Data Analysis :

Our first task was to get the New york city Geographical Coordinates Data. To do so, first we load and explore data from newyork_data.json file. We then transform the data into a dataframe, which contains the geographical coordinates of New York city neighborhoods. Then we created a new dataframe from the existing one, which contained data only for the borough of Bronx (this data will be used at the next stage to get Venues data from Foursquare). Finally, we used geopy and folium libraries to create a map of New York city with neighborhoods superimposed on top.



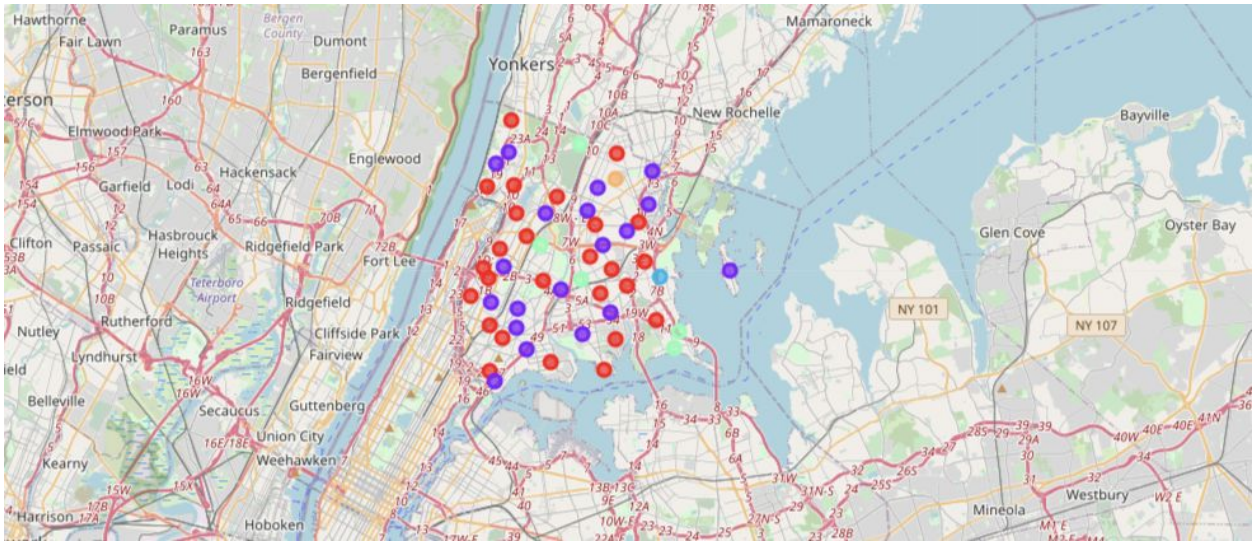
After the above data was collected, we used the geographical coordinates as input for the Foursquare API, in order to obtain venue information for each neighborhood. Then we used this data to cluster the neighborhoods and find the best location for our client's business venture.

From the Foursquare API, we got the top 100 venues in a radius of 500 meters for each neighborhood. Example of the data below.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Wakefield	40.894705	-73.847201	Lollipops Gelato	40.894123	-73.845892	Dessert Shop
1	Wakefield	40.894705	-73.847201	Carvel Ice Cream	40.890487	-73.848568	Ice Cream Shop
2	Wakefield	40.894705	-73.847201	Walgreens	40.896528	-73.844700	Pharmacy
3	Wakefield	40.894705	-73.847201	Rite Aid	40.896649	-73.844846	Pharmacy
4	Wakefield	40.894705	-73.847201	Dunkin'	40.890459	-73.849089	Donut Shop

Results

Using the venue data, we decided to cluster Bronx neighborhoods into 5 clusters using K-means clustering. Below, a visualisation of the clusters that were created.



Examining the clusters:

Cluster 1

This cluster contains a lot of competition (donut shops, cafes etc.) therefore it is excluded from further analysis

Cluster 2

Again, like the previous one, this cluster contains a lot of competition (donut shops, cafes etc.) therefore it is excluded from further analysis

Cluster 3

For this cluster out of the top 10 venues, 2 are competitors (sandwich and donut shop), while 3 out 10 venues are entertainment businesses

Cluster 4

For this cluster out of the top 10 venues, on average 2.3 are competitors (sandwich shop, cafe, donut shop etc), while 2.5 out 10 venues are entertainment businesses

Cluster 5

The final cluster contains no competitors on the top 10 venues, while 1 out of 10 venues are entertainment businesses

Therefore given the results of the above analysis, we will propose to our client that the best neighborhood in bronx to open his bakery is Edenwald (the only neighborhood included in cluster 5)

Discussion

1. We observed that in most neighborhoods in Bronx, there were a lot of competition present, making a not so compiling case for our client to set shop in any of those locations.
2. However, due to our competitors analysis (using FourSquare data) we were able to find the best neighborhood (according to the clients wishes) to open their business in Edenwald, Bronx, NY

Conclusions

This analysis was performed on limited data and a narrow focus (in terms of location).
Future research should focus on a bigger location, and also utilise more data, such as number of residents in the neighborhood, number of tourist, data on foot data, resident demographics (including age, income etc.), in order to take into account more issues that might influence the choice of the best location for a new business venture (regardless of type of business).