Capstone Project - Queens

Table of Contents

1. Introduction: Business Problem
2. Data
3. Methodology
4. Analysis
5. Results and Discussion
6. Conclusion

Introduction: Business Problem

A fastfood chain is interested in expanding to Queens. They want to open a few new restaurants and would like to find out the best neighborhoods to do so.

In order to answer that question, we first need to analyse the different neighborhoods in Queens, New York. We would like to understand the most common venues in each neighborhood. Knowing this information, we can group the neighborhoods which have similar characteristics. This helps to analyse each group to try to get insights or see patterns to help us to draw our conclusions and make good suggestions to the stakeholders.

Data

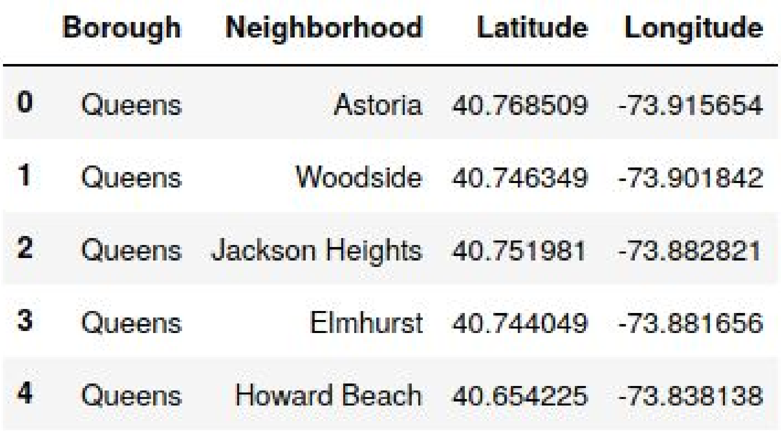
The data we need must contain the following information:

- Names of the neighborhoods in Queens, New York
- Their latitude and longitude
- Names of the 10 most common venues for each neighborhood

This dataset is accessible for free on the internet. It is acquired from https://geo.nyu.edu/catalog/nyu_2451_34572

The dataset contains data for all of New York, so we download the full set and the extract the data for Queens.

The latitude and longitude values of the different neighborhoods in Queens are determined using the **Geopy Library**. In this way, we find the data frame we will continue to work with. Here only the first 5 neighborhoods of Queens are shown as an example.

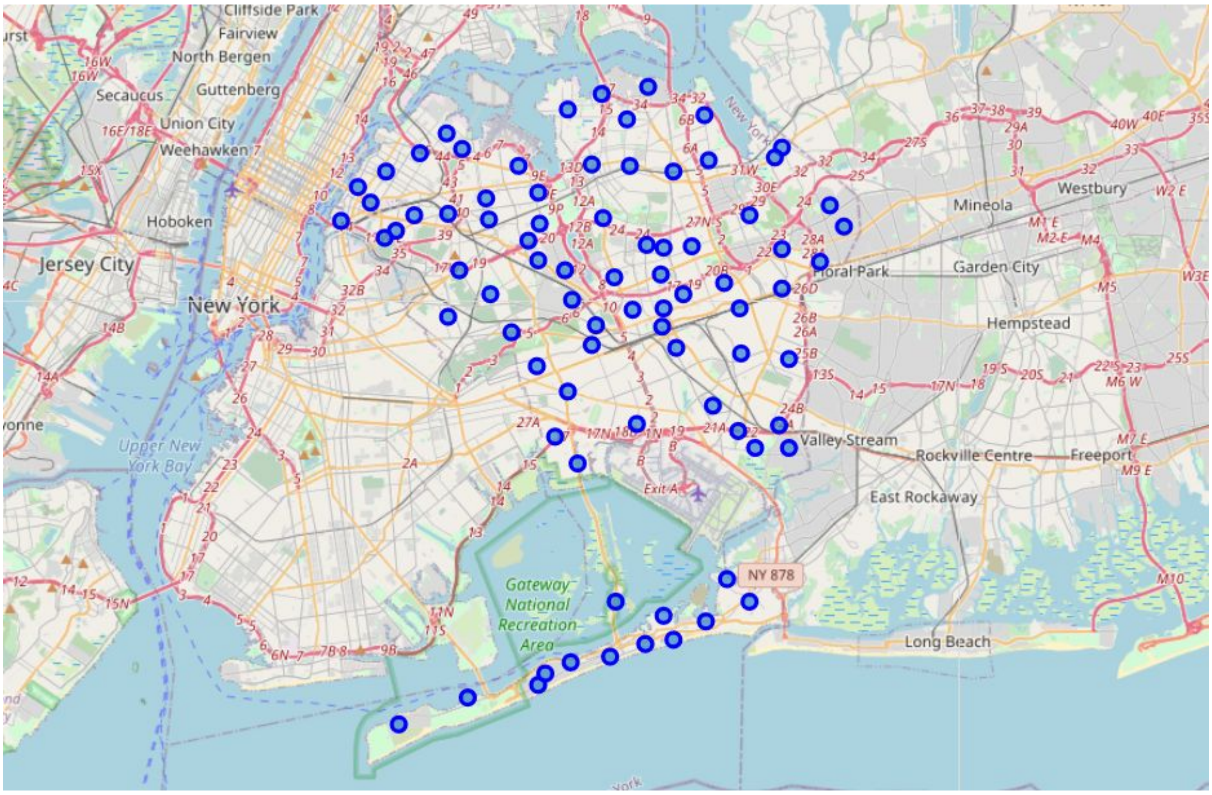


	Borough	Neighborhood	Latitude	Longitude
0	Queens	Astoria	40.768509	-73.915654
1	Queens	Woodside	40.746349	-73.901842
2	Queens	Jackson Heights	40.751981	-73.882821
3	Queens	Elmhurst	40.744049	-73.881656
4	Queens	Howard Beach	40.654225	-73.838138

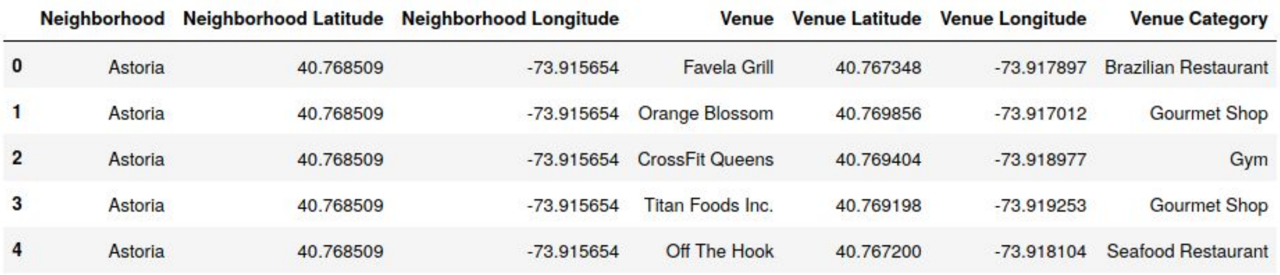
Methodology

We use the folium library to create the geographical map of Queens, in order to select Queens, York City as the central location in our map we need the coordinates. The geographical coordinates of Queens are 40.7498243, -73.7976337.

On the map we show all neighborhoods of Queens as blue dots.



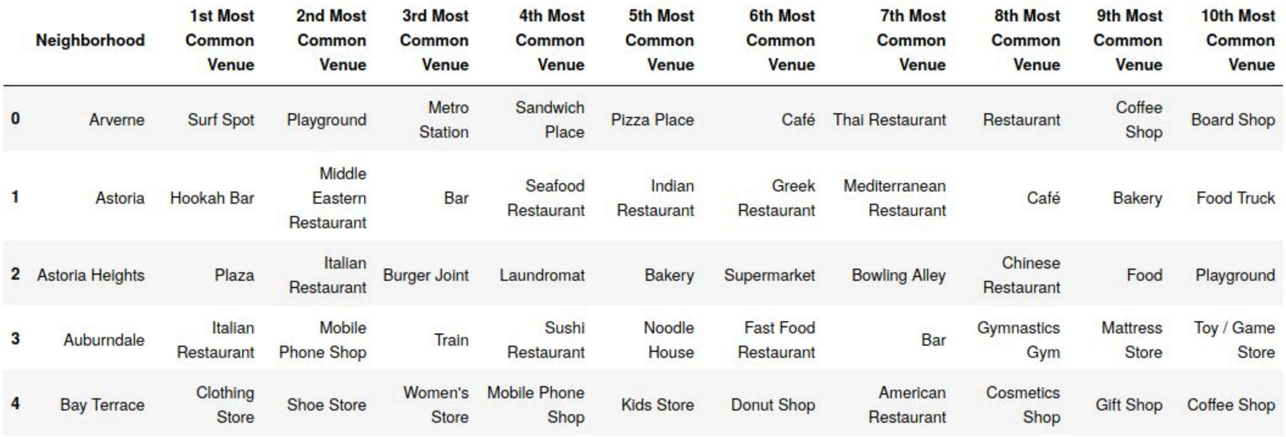
Now, the most common venues of each neighborhood are obtained using the **Foursquare API**. As an example we look at the first 5 entries in our new dataframe.



	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Astoria	40.768509	-73.915654	Favela Grill	40.767348	-73.917897	Brazilian Restaurant
1	Astoria	40.768509	-73.915654	Orange Blossom	40.769856	-73.917012	Gourmet Shop
2	Astoria	40.768509	-73.915654	CrossFit Queens	40.769404	-73.918977	Gym
3	Astoria	40.768509	-73.915654	Titan Foods Inc.	40.769198	-73.919253	Gourmet Shop
4	Astoria	40.768509	-73.915654	Off The Hook	40.767200	-73.918104	Seafood Restaurant

We find that there are 290 unique categories of venues in the data fame.

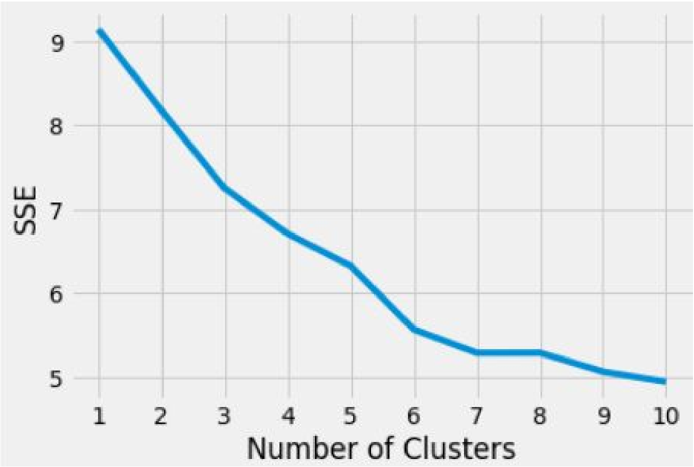
Next, utilizing the one-hot encoding technique, we create a new dataframe displaying the top 10 venues for each neighborhood.



	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Arverne	Surf Spot	Playground	Metro Station	Sandwich Place	Pizza Place	Café	Thai Restaurant	Restaurant	Coffee Shop	Board Shop
1	Astoria	Hookah Bar	Middle Eastern Restaurant	Bar	Seafood Restaurant	Indian Restaurant	Greek Restaurant	Mediterranean Restaurant	Café	Bakery	Food Truck
2	Astoria Heights	Plaza	Italian Restaurant	Burger Joint	Laundromat	Bakery	Supermarket	Bowling Alley	Chinese Restaurant	Food	Playground
3	Auburndale	Italian Restaurant	Mobile Phone Shop	Train	Sushi Restaurant	Noodle House	Fast Food Restaurant	Bar	Gymnastics Gym	Mattress Store	Toy / Game Store
4	Bay Terrace	Clothing Store	Shoe Store	Women's Store	Mobile Phone Shop	Kids Store	Donut Shop	American Restaurant	Cosmetics Shop	Gift Shop	Coffee Shop

This enables us to cluster the neighborhoods according to their similarities in venues using the *k*-means method.

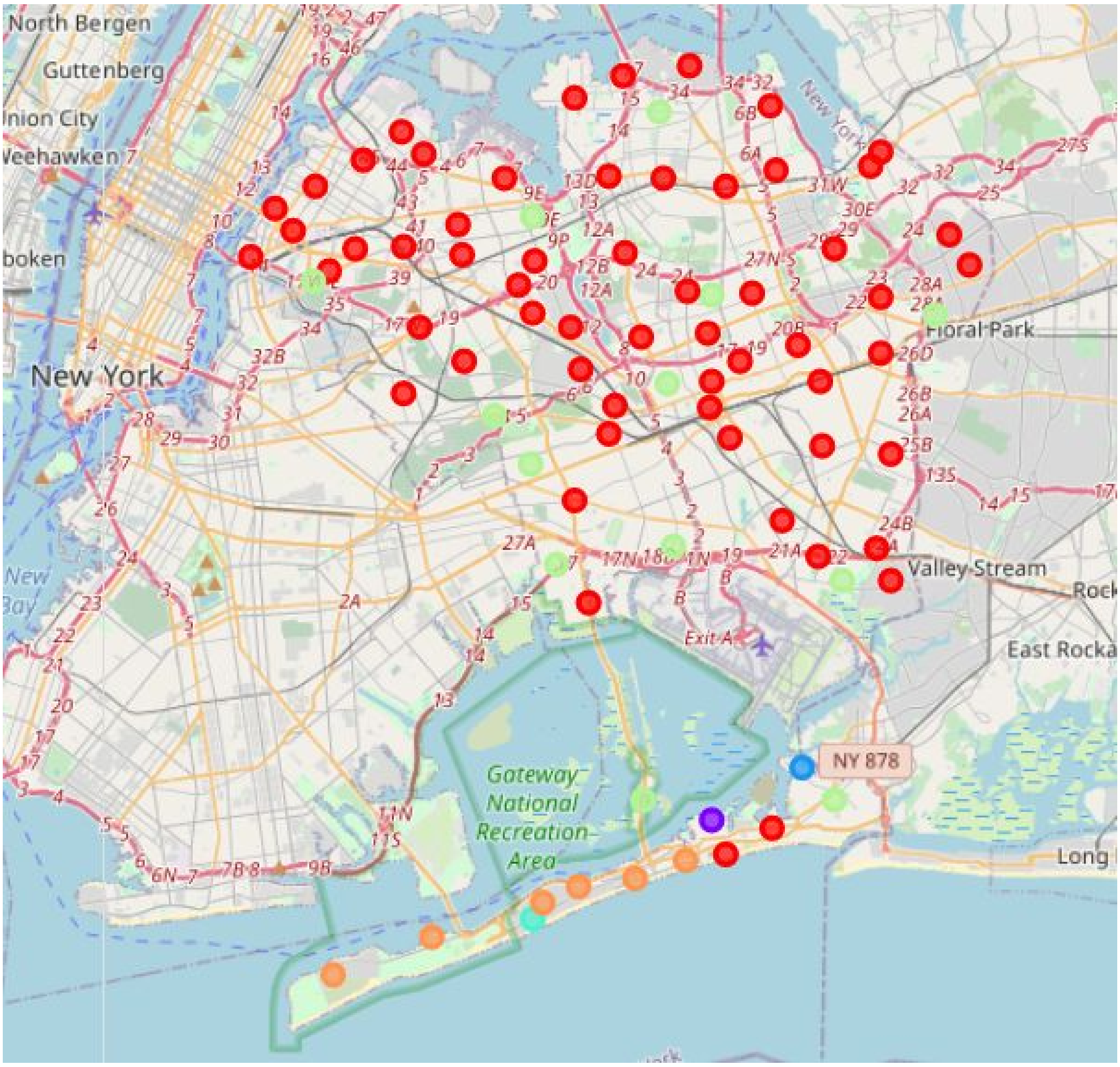
Let's find first the best number of clusters based on the elbow method.



The optimal group number is determined by the convexity of the curve, here we find k=6.

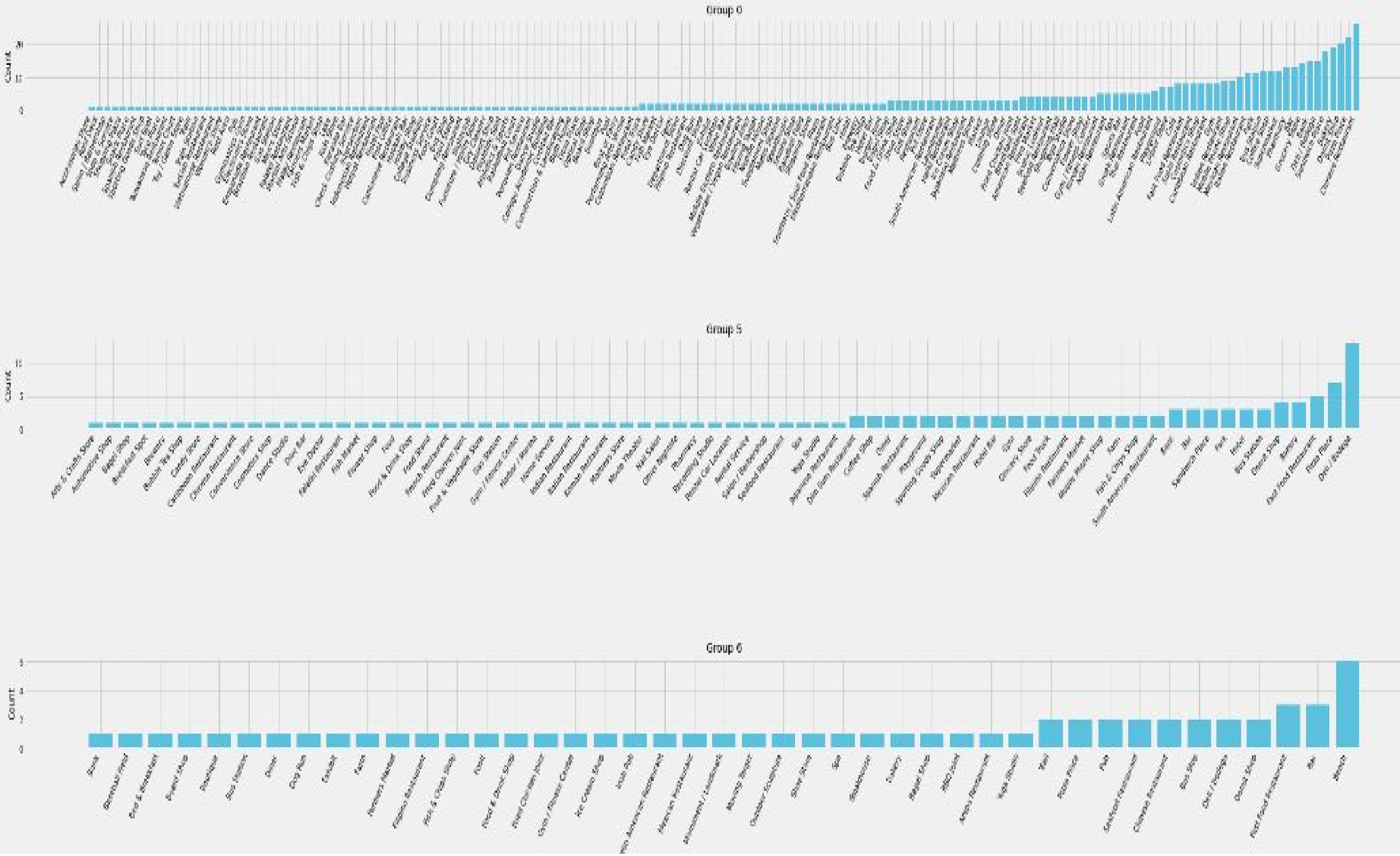
Results and Discussion

Finally, we can visualize the resulting clusters, where the neighborhoods with the same color belong to the same group.



Now, we examine each cluster and determine the discriminating venue categories that distinguish each cluster.

We have used the k-means method to group the cities based on the common venues. Now we want to count the number of unique venues of each cluster among these most common venues. Notice that now we do not care about the order of the most common venues anymore. To get an overview we will make bar charts for the clusters 1,5,6. We exclude the other clusters from the further analysis, since they only contain one neighborhood.


Group 0
Count
0
10
20
Group 5
Count
0
5
10
Arts & Crafts Store
Automotive Shop
Bagel Shop
Breakfast Spot
Brewery
Bubble Tea Shop
Candy Store
Caribbean Restaurant
Chinese Restaurant
Convenience Store
Cosmetics Shop
Dance Studio
Dive Bar
Eye Doctor
Falafel Restaurant
Fish Market
Flower Shop
Food
Food & Drink Shop
Food Stand
French Restaurant
Fried Chicken Joint
Fruit & Vegetable Store
Gas Station
Gym / Fitness Center
Home Service
Indian Restaurant
Italian Restaurant
Korean Restaurant
Mattress Store
Movie Theater
Nail Salon
Other Repair Shop
Pharmacy
Recording Studio
Rental Car Location
Rental Service
Salon / Barbershop
Seafood Restaurant
Japanese Restaurant
Dim Sum Restaurant
Coffee Shop
Diner
Spanish Restaurant
Playground
Sporting Goods Shop
Supermarket
Mexican Restaurant
Hotel Bar
Gym
Grocery Store
Food Truck
Farmers Market
Mobile Phone Shop
Café
Fish & Chips Shop
South American Restaurant
Bank
Bar
Sandwich Place
Park
Hotel
Bus Station
Donut Shop
Bakery
Fast Food Restaurant
Pizza Place
Deli / Bodega
Group 6
Count
0
2
4
6
Bar
Pizza Place
Park
Seafood Restaurant
Chinese Restaurant
Bus Stop
Deli / Bodega
Donut Shop
Fast Food Restaurant

We can observe that in cluster 5 and 6 fast food restaurants seem to be very popular, they appear with a high frequency among the most 10 common venues. Since the cities in the same group are similar, this leads to the conclusion that it could be a good choice to open a fast food restaurant in the neighborhoods belonging to these two clusters, which do not already show fast food restaurants among the 10 most common venues.

Conclusion

In this project we have analysed the neighborhoods in Queens and their most common venues. We have clustered the neighborhoods and found out which ones are similar. We found that in two clusters, fast food restaurants are especially popular.

After analysing the data we have the following recommendation: It would be a good choice to open new fast food restaurants in those neighborhoods that belong to cluster 5 and 6, but don't have a fast food restaurant among their 10 most common venues yet.