Capstone Project – Applied Data Science

Battle of the Neighborhood

Introduction

The CEO of a Fortune 500 Company wants to open a new branch in New York. The CEO and most of the company's employees are food enthusiasts with a strong preference for Chinese food. The CEO would like to know where in New York to open the new branch in order to have the best variety of Chinese restaurants in proximity.

Therefore, we are going to determine which boroughs in New York are potentially interesting and which neighborhoods exactly could be considered.

Data

We will use the Foursquare location data accessed by the Foursquare API using a free account. Therefore, we will limit the max amount of searched venues per neighborhood to 50 in order to not exceed the limits of the 'freemium'-service.

Since we are specifically interested in food-venues, we will further add the parameter 'categoryId' to the request URL:

https://api.foursquare.com/v2/venues/explore?categoryId={}&client_id={}&client_secret={}&v={}&ll={},{}&radius={}&limit={}

By using the value '4d4b7105d754a06374d81259' as 'categoryId', we can filter directly for food venues.

To get the spatial information about the boroughs and neighborhoods in New York, we will use the '2014 New York City Neighborhood Names' (LINK) dataset from the NYU Spatial Repository (LINK).

Using the dataset and the Foursquare API we can use the k-means algorithm to cluster the Neighborhoods in an example borough to identify potentially interesting locations for our food-enthusiastic CEO.

Results

Using the NYU Spatial Repository dataset on New York we can get information on all the boroughs and their respective neighborhoods (Fig1.). We can see that the number of neighborhoods in the boroughs ranges from 40 to 80. With Manhattan having the fewest neighborhoods and Queens having the most.

We can also use Folium to display the individual neighborhoods on the map (Fig1., color-coded for boroughs).

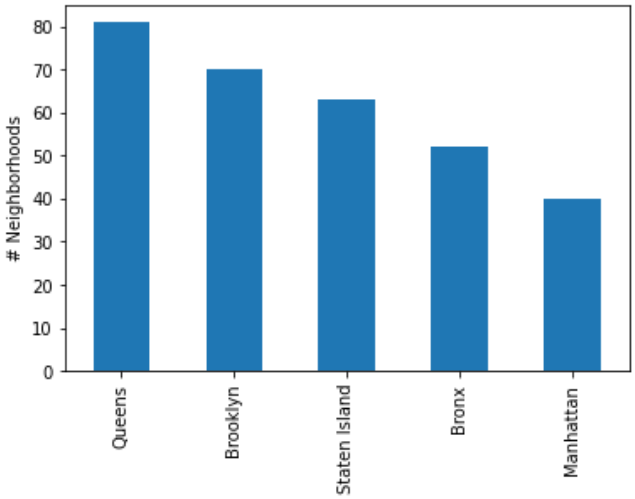

Figure 1 - Number of neighborhoods in each borough.

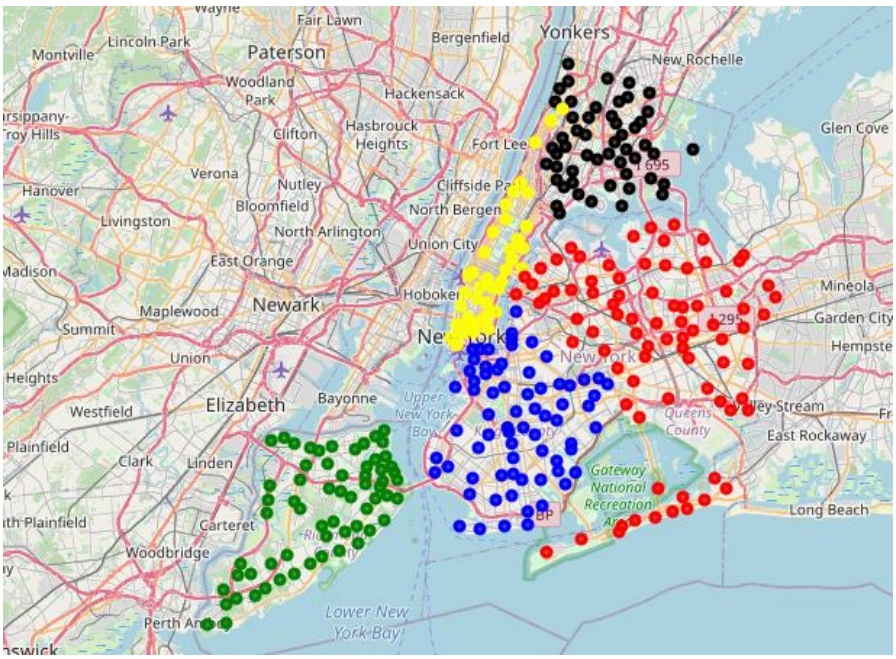

Figure 2 - Neighborhoods color-coded by their respective boroughs. Manhattan=Yellow, Bronx=Black, Queens=Red, Brooklyn=Blue, Staten Island=Green.

For the following analysis we are going to take a closer look at Queens (Fig 3.), since it has the most neighborhoods and as such we might get a bigger variety between the individual neighborhoods.

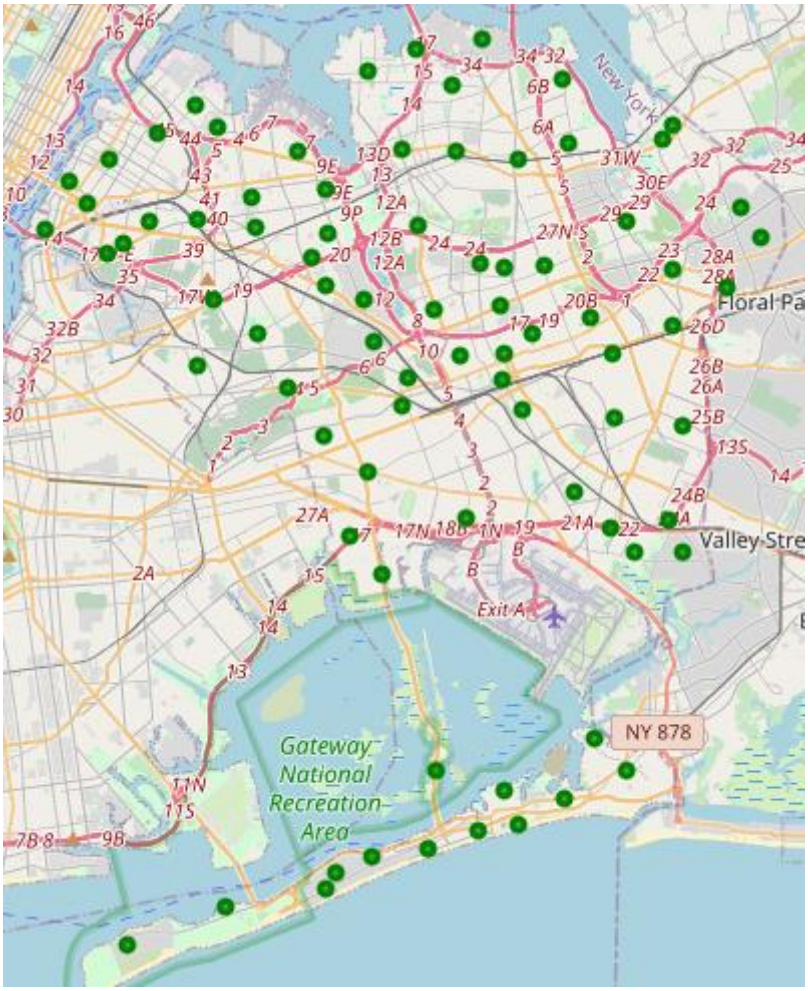

Figure 3 - Detailed overview over the neighborhoods in Queens.

Using the Foursquare API with the ‘categoryId’ parameter '4d4b7105d754a06374d81259' to filter for food venues we can get insights on the amount of food places in each neighborhood (Fig4.). We can see that there is an enormous range between the top six neighborhoods (Woodside, Bayside, Jackson Heights, Astoria, Flushing, Sunnyside Gardens) that cap out at the set limit of 50 venues/neighborhood and Malba, Whitestone and Jamaica Estades that only have a single food venue listed. This can of course be due to the size of the neighborhood. For example, Malba is only 1/3 of the size of Woodsize, but has only 1/50 the number of food places. So here it is likely that but additional factor like location, demographics, population etc. play an important role. Most neighborhoods have about 10-30 places to choose from.

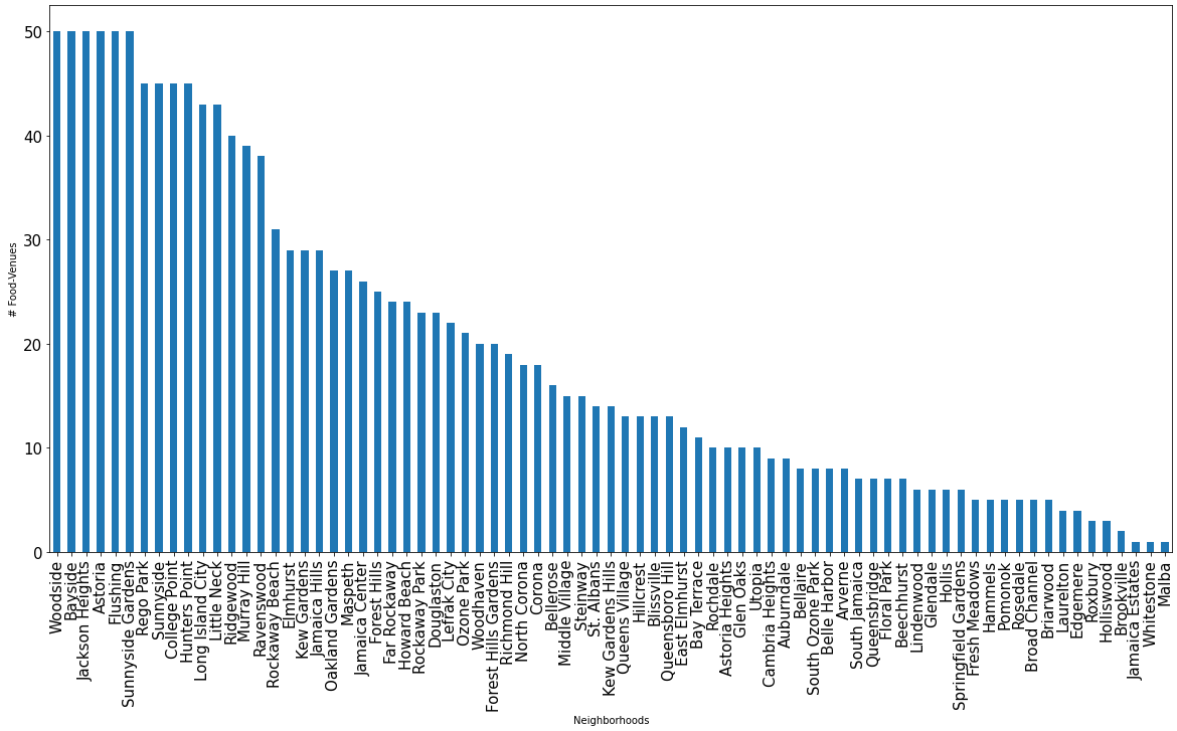

Figure 4 - Number of food venues per neighborhood (ranked).

Instead of looking at the total number of venues, we can also take a look at the frequency of the different food venue types. In total 96 different food categories were identified across all neighborhoods in Queens. The top 10 are listed below (Fig5.). The most frequent food location in Queens are Deli/Bodega. These are basically grocery stores that also offer prepared food. They originated in New York and the differ in their focus. Bodega is Spanish for 'grocery store' and this is its main focus, but also offers food (e.g. packaged or prepared). Deli comes from the German word 'Delikatessen' and has its focus on prepared food like Sandwiches. The next most frequent type is 'Chinese restaurant', so Queens seems to be already a good choice for our CEO.

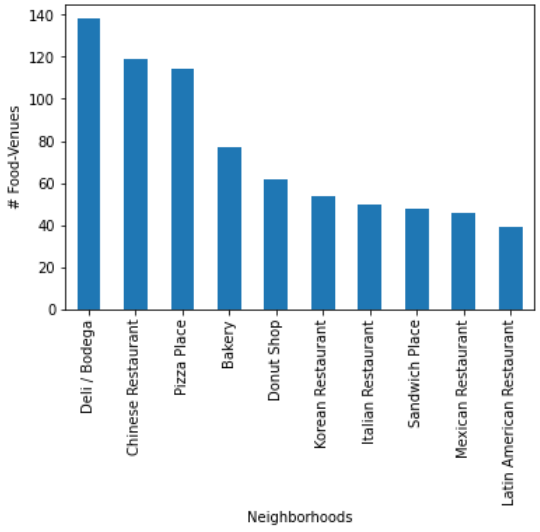

Figure 5 - The most frequent types of food locations (ranked).

Now its time to use the k-means algorithm to cluster the neighborhoods in Queens based on their most frequent food types. In order to use suitable amounts of cluster the elbow method (LINK) was used to determine the optimal number. In our case this is 5 (Fig6.).

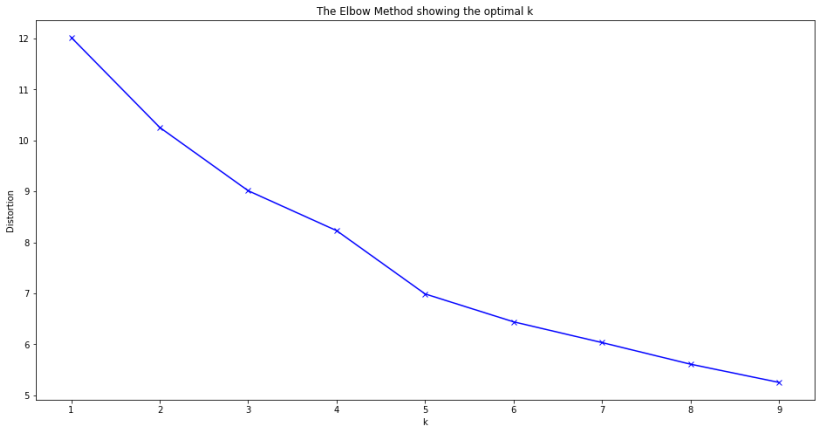

Figure 6 - Elbow Method to determine the best value for 'k'.

Using the k-means algorithm with 5 clusters and plotting the results via folium we can see the distribution of neighborhoods below (Fig7.).

Cluster 1 comprises 19 neighborhoods and mainly features deli/bodega, bakeries and pizza places. Cluster 2 is the largest cluster with 52 neighborhoods and focuses on Chinese kitchen, but also has pizza. Cluster 3 has Vietnamese food places and dumpling restaurants, but only consists of 3 neighborhoods. Cluster 4 has two entries and a focus on Indian food, whereas the last cluster only has one entry and specializes in Caribbean restaurants.

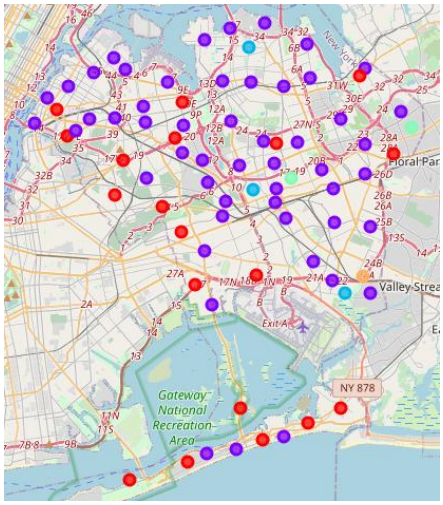

Figure 7 - Neighborhoods clustered by food-types. 5 clusters.

Discussion

In this small case study we have taken a closer look at the borough Queens of New York and used spatial location data, the Foursquare API and k-means clustering to determine potentially interesting neighborhoods for our client who seeks to find a new location to open a company branch. The CEO seeks to find a neighborhood that has many Chinese restaurants nearby, since she/he and the majority of the company's employees prefer Chinese food.

With this in mind, we can recommend neighborhoods from cluster 2. It contains 52 neighborhoods and has a focus on Chinese food places. The cluster also offers other food types like pizza or the New York typical deli/bodega places. This might be interesting for the minority of employees that prefer other types of food.

To determine which of these 52 neighborhoods should be chosen for the new branch, one could conduct an additional clustering. This clustering could, for example, focus on property costs (company building), affordable rents (for the employees) or other venues. The CEO might also like to look at the other boroughs in New York.