Starting a Coffeeshop in New York City

Predicting the right borough and neighbourhood to open a shop

KP Smit – 1st July 2020

1. Introduction

1.1 Background

New York City, NY, United States of America, is undoubtedly one of the (if not the) most iconic cities in the world. With a population of over 8 million people divided over five boroughs there are plenty of opportunities to start a business that caters to the needs of New Yorkers. Many of the citizens work in Manhattan, but live elsewhere and therefore have to commute to their work. Even the people that live in Manhattan are still likely to catch a form of public transport to get to their work. Because people are always on the go in the city that never sleeps, there is a market to serve people food and drinks on the go. In our case, we are trying to figure out what the best area is to start a coffeeshop to serve commuters their daily coffee and breakfast while getting to work.

1.2 Problem

We want to be able to predict the best location for a new coffeeshop, data that might help us to predict the best location is the distribution of the population in New York City, data about the different neighbourhoods, information about the New York Subway and the availability of similar venues in each neighbourhood.

1.3 Interest

This study may be of interest for future entrepreneurs that want to start their own business (in this case particular a business that serves food and drinks). This study could be altered to explore data for different kind of restaurants, or even different cities and neighbourhoods.

2. Data Acquisition and Cleaning

2.1 Data Sources

There is a rich variety of data sources available that we are able to use in answering our research question, there are four main data sources that are selected that prove to be helpful.

1) A dataset of all the different neighbourhoods and boroughs in New York City is used. This data is freely available via the following link (https://cocl.us/new_york_dataset).
2) A dataset with the population distribution over New York city is available via this link: https://en.wikipedia.org/wiki/Template:NYC_boroughs
3) A dataset with all the subway stations in New York city including coordinates and boroughs is available here: http://web.mta.info/developers/data/nyct/subway/Stations.csv

4) The last data set that is used is Foursquare, their API will allow us to see various venues in the different neighbourhoods in New York. We will retrieve the following information about venues:
   - Name: The name of the venue.
   - Category: The category type as defined by the API.
   - Latitude: The latitude value of the venue.
   - Longitude: The longitude value of the venue.

2.2 Data Cleaning

The dataset with the different boroughs and neighbourhoods of New York City, was ready to be used and didn’t require and data cleaning. The dataset used gives a clear list of different boroughs and neighbourhoods within those boroughs, as well as coordinates that show the location of these neighbourhoods.

The dataset analysing the population of New York required some data cleaning, the first problem that needed to be corrected was that there were multiple “header rows”, these needed to be overwritten and give a more descriptive name. In addition to this the last three rows needed to be dropped as these gave information about the entire city of New York, the state of New York and mentioned the sources.

The dataset with the different subway stations in New York City was ready to be used, and did not require any additional cleaning.

Foursquare API is an online application used my many developers & other applications. It can be used to retrieve information (data) about the places present in the neighbourhoods of New York City. The API returns a JSON file which can be turned into a data-frame.

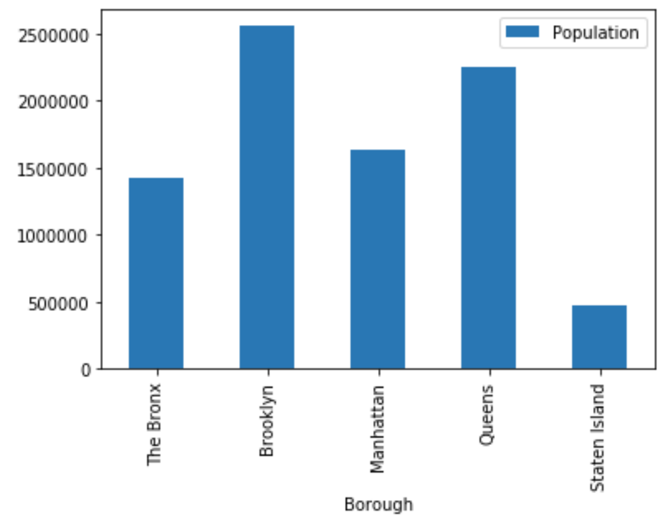
3. Exploratory Data Analysis

3.1 Population Distribution

Our hypothesis is that people who commute into work are interested in our coffee, to capture the biggest market of potential customers the focus is on the boroughs with the highest population and the highest number of subway stations.

Our data analysis shows (Figure 1) that Brooklyn has the largest population of all the New York boroughs. This indicates that this may be a good borough to start our coffeeshop, after all we want to capture as much customers as possible.

Figure 1: Population Distribution of New York Boroughs



3.2 Subway Stations in different Boroughs

Another indication of Brooklyn being a great location to start our coffeeshop is that Brooklyn has the most subway stations of any borough in New York City (as indicated in Figure 2). However, not all neighbourhoods in Brooklyn have access to a subway station, therefore we want to check the location of each subway station and determine in which neighbourhood it is located. Figure 3 shows a map of Brooklyn, where the blue dots indicate the neighbourhoods which have access to subway lines, the white dots are neighbourhoods which don't have subway stations, these neighbourhoods are therefore excluded from further analysis. The small black dots are all the locations of subway stations in Brooklyn.

Figure 2: Subway Stations per New York Borough

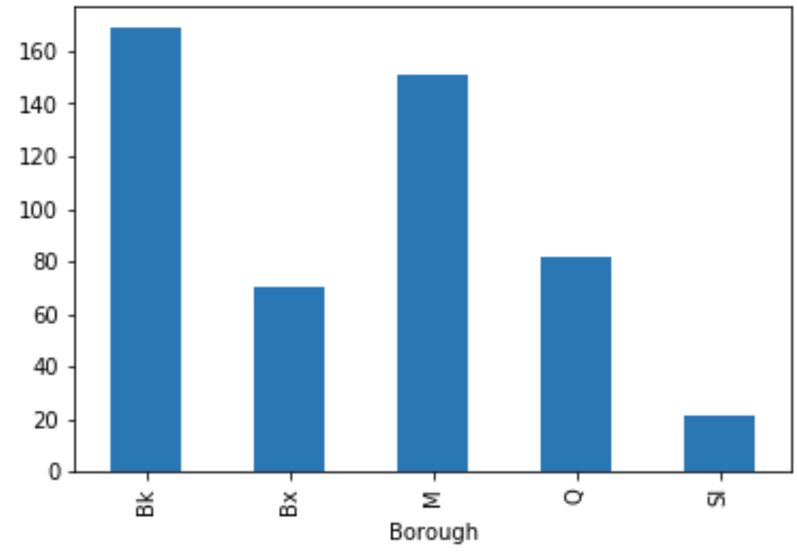
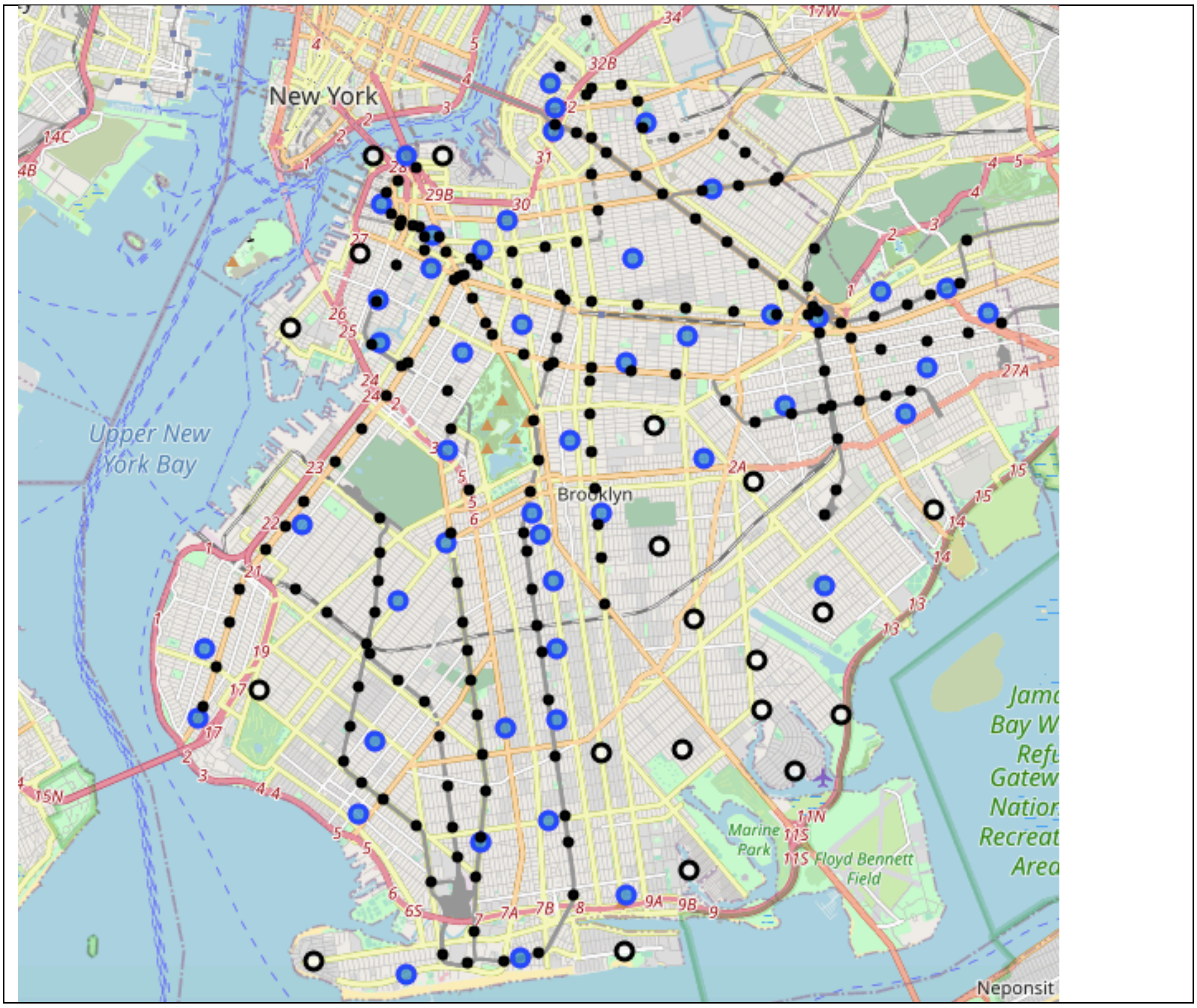


Figure 3: Neighbourhoods and Subway Stations in Brooklyn



3.3 Foursquare

Now that we have selected a list of suitable neighbourhoods within Brooklyn we can explore the most suitable neighbourhood to start a coffee shop in. To do this, we will explore the different neighbourhoods via Foursquare to see if there are competing businesses in the neighbourhoods. We will take a broader look at different venue categories which may also serve food and drinks for on the go. These categories are listed below:

- Bagel Shops
- Bakeries
- Breakfast spots
- Cafes
- Coffee Shops
- Donut Shops
- Fast Food Restaurants
- Food Courts
- Food Stands
- Food Trucks
- Juice Bars
- Markets
- Salad Bars
- Sandwich Bars
- Snack Stops
- Tea Rooms

These categories all serve food for on the go, and are therefore likely competition for our new coffeeshop. We want to explore which neighbourhood in Brooklyn has the least competition in this market, as that may be a good indication of a suitable neighbourhood to open our venue. We will try an achieve this by K-means clustering.

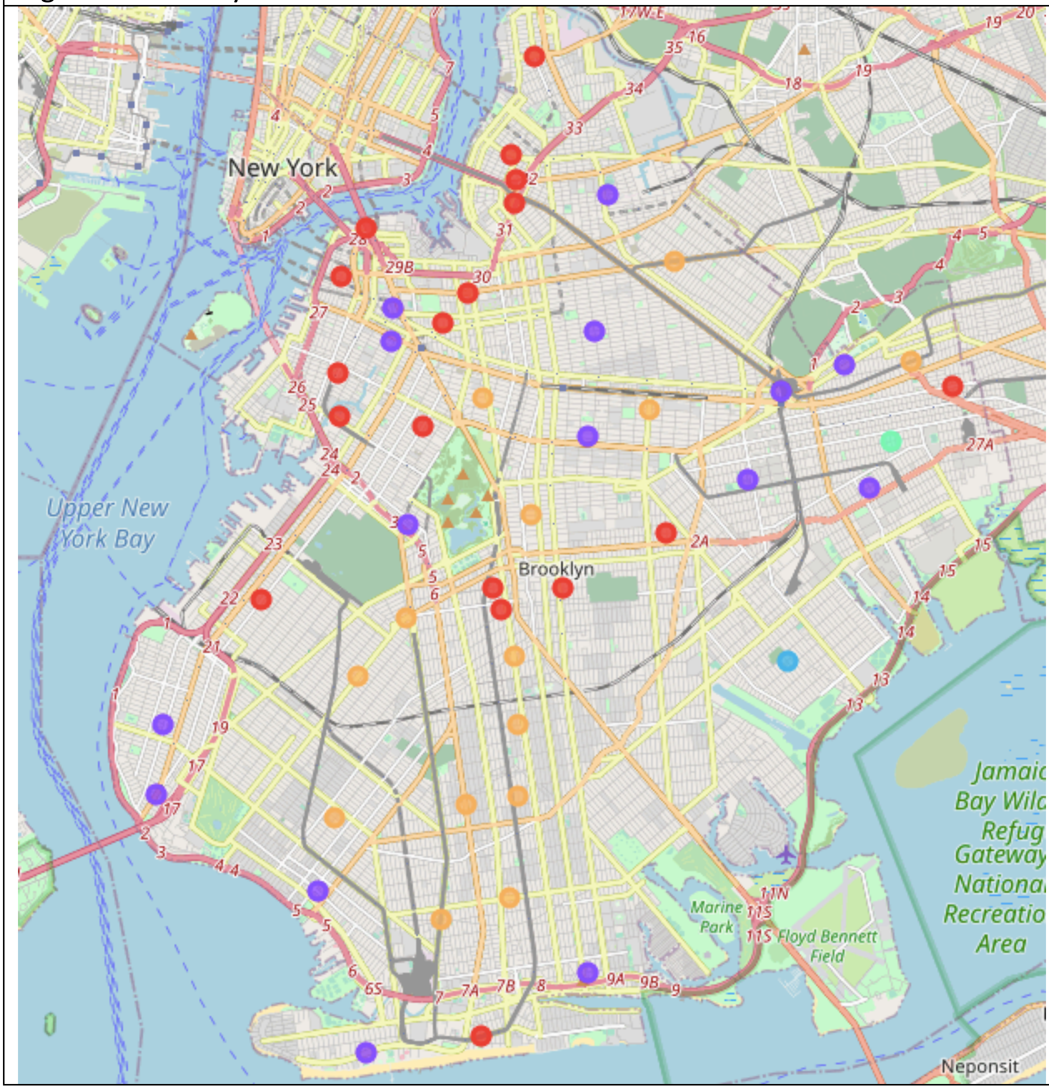
We will filter out the venues that fit into the categories in the list above, and disregard the other venues. We will group the venues by the different neighbourhoods and set a distribution of the venues in each neighbourhood.

4. Predictive Modelling

4.1 K-means clustering

We are dividing the different neighbourhoods into 5 different clusters based on the provided information we got from Foursquare. We project this information on the map of Brooklyn we created earlier.

Figure 4: Brooklyn in Clusters



4.1.1 Cluster 1

Cluster 1, represented by the red dots in the figure above has loads of Food Trucks and Breakfast Spots. The majority of these neighbourhoods are close to Northern and Central Brooklyn, which are the business and downtown areas of Brooklyn, which will see lots of people looking for a quick bite and drink.

4.1.2 Cluster 2

Cluster 2, has mostly Sandwich and Fast Food Restaurants. This cluster is represented by the purple dots in the figure above. They are scattered around the borough, and are mostly residential areas.

4.1.3 Cluster 3 & 4

Cluster 3, has Coffee Shops and Tea Rooms. It is depicted by the blue dot on the map, and comprises a single neighbourhood: 'Canarsie'. Cluster 4: comprised of the neighbourhood of East New York is the cluster which is predominantly Fast Food Restaurants and Tea Rooms.

4.1.4 Cluster 5

The last cluster is the cluster that is represented by the orange dots on the map in Figure 4. This cluster has the following venues as being the most common type of venues: 'Bakery, Café, Bagel Shops and Sandwich venues.

5. Results and Discussion:

5.1 Results

We have reached the end of the cluster analysis, in this section we will document all the findings from above clustering & visualization of the dataset. We started off with the business problem of identifying a good neighbourhood to open a Coffee Shop in New York City. To achieve that we looked into the population distribution and number of subway stations in each of the boroughs in New York. We identified that Brooklyn has the highest population and the highest number of subway stations.
We excluded the neighbourhoods that did not have access to a subway station as these are not of interest to our target market: 'commuters that want a coffee on their way to work'. After identifying the different suitable neighbourhoods we used foursquare data to analyse competing businesses. To get a realistic view of the market we included competing breakfast and fast food restaurants that may cater to the similar consumer market.

Via K-means clustering we identified that the areas with the most likely competition are the neighbourhoods that are part of cluster 5. In the 14 neighbourhoods in this cluster, the most common venues that may target similar consumers are: 'Bakeries, Bagel Shops, Cafes and venues that sell Sandwiches. We believe that these 14 neighbourhoods are unsuitable to open a Coffee Shop, due to the high volume of similar type venues.

Cluster 1, which comprises 18 different neighbourhoods shows a high volume of Food Trucks and Breakfast Spots venues. This is a mix of venues which could cater to customers that are looking for either breakfast or lunch and might prove strong competition. These neighbourhoods are therefore not the ideal location to open a new Coffee Shop.

Clusters 3 & 4 are not of interest due to their high focus on Coffee and Tea and hence are likely to have a strong competitor in place.

Cluster 2, which has 15 different neighbourhoods shows a lot of Sandwich focussed venues. Starting a Coffee Shop in one of these neighbourhoods, could offer something very different to the market and may therefore be an interesting opportunity. There are a number of neighbourhoods with a single subway station, these include Coney Island, Sheepshead Bay and Fort Hamilton. Of these three Sheepshead Bay appears the most suitable to start a Coffee Shop, when looking at the other available venues.

5.2 Discussion

According to this analysis, the neighbourhood of Sheepshead Bay provides the least competition for the new upcoming Coffee Shop. Looking at the population distribution, and the number of subways stations it looks like all people that use the subway from this neighbourhood will have to use a single station. This means the entire customer base needs to come towards this station which is a good opportunity to tap into the demand this generates. Some of the drawbacks of this analysis are — the clustering is completely based on data obtained from Foursquare API and the data about the population distribution between the different boroughs of New York, not the different neighbourhoods. Thus, there is a gap in the population distribution. Although there are areas where it can be improved, this analysis has provided us with some good insights, preliminary information on possibilities & a head start into this business problem by setting the step stones properly.

6. Conclusion

During this project, we have looked into a business problem for an entrepreneur who wants to start a Coffee Shop in New York City. To answer the question in which neighbourhood he should start he should start the business we have used python libraries to fetch and analysis data. We have used the Foursquare API to explore the competing venues in each neighbourhood.

Some of the drawbacks or areas of improvements shows us that this analysis can be further improved with the help of more data and different machine learning technique. Similarly we can use this project to analysis any scenario such as opening a different type of restaurants or opening of a new gym and etc.