**Analysis of Brooklyn Neighborhoods for Opening a Convenience Store**

Mikhail Gongadze

IBM Data Science Capstone

10/17/19

**Background and Introduction to the Business Problem**

Brooklyn has a highest population of all the New York boroughs, it was estimated to be 2,648,771 in 2017. It is also the second by land area, which was estimated as 70.82 sq. miles in 2017 and second in population density, which was estimated as 37,137 persons per sq. mi. in 2017 [Wikipedia]. Overall Brooklyn is mostly residential, which could be a great place to open up a business that would cater to local people, such as convenience store. Brooklyn has 80 neighborhoods, out of which I will look into 70. In this project, we will analyze all the neighborhoods in Brooklyn and try to find best locations where one can start a business by opening a convenience store that’ll be in demand. At the same time, we will try to analyze the prices of commercial real estate, to get the most for the buck and see if it’s going to be easier to lease or to purchase premises.

Having a high population density, Brooklyn can be a great place to open up a convenience store. Few things to note though is that because of location and extreme competition real estate prices, be it for rent or for purchase, can get very high. Brooklyn is a very diverse borough, many Brooklyn neighborhoods are ethnic enclaves, based on which, one can adjust to specific needs and interests of surrounding area as well as their own preferences. Therefore, a potential business owner should pick a right niche, either to cater to taste preferences of a specific nationality or to be more general and sell products that are more common to everyone.

To consider all of the aforementioned issues, we will generate a map of Brooklyn, which we will populate with different venues from Foursquare, then cluster it into different districts to look at the interests, prices, competition and other factors of different districts to determine best places to open a convenience store.

This project may be useful to potential entrepreneurs, who are thinking of taking a challenge of opening a business, such as convenience store in such a place as Brooklyn, New York. It also will benefit those, who are interested in exploring more about Brooklyn districts and type of venues it has to offer, including the potential to own or rent a property there.

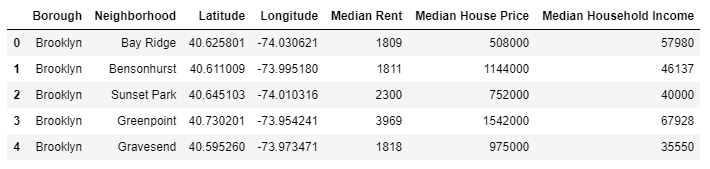
**Data Description**

* The data that was used came from the New York .json Dataset, which was downloaded from Coursera website.
* The \*.json file contains all of the coordinates for boroughs and districts of New York, from which we remove all the unnecessary coordinates except the ones for Brooklyn that further was used for choropleth map creation.
* The venues were taken from Foursquare API. The data represents coordinates and descriptions of most common venues that belong to all the neighborhoods in Brooklyn.
* We will look at residential and commercial property prices pulling data from Zillow and other retail web pages. Then that data will be clustered into each separate district, and the mean with the median prices for each district will be calculated.
* The features that are going to be extracted from the data consist of:
  + Districts in Brooklyn and real estate prices
  + Type and frequency of venues

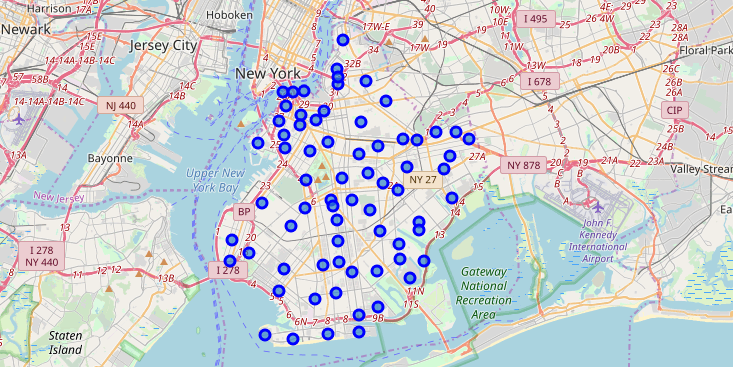
Further, the data will be used to find out which districts have the most potential for opening convenience store, or what kind of convenience stores can be opened there.

**Methodology**

The master dataset consisted of two joint datasets, the dataset of all the neighborhoods of Brooklyn [1] with their coordinates and the dataset of Median Household Income, House Price and House Rent [2-6]. GitHub repository was used as a database for this project and IBM Watson Studio as an environment for coding in Python.



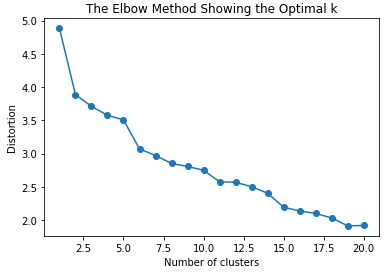
To visualize the map of Brooklyn borough in New York, I’ve used folium maps, Python’s library that provides geographic representation of the master dataset. Each circle on the map represent points of coordinates for each of the neighborhood in Brooklyn.



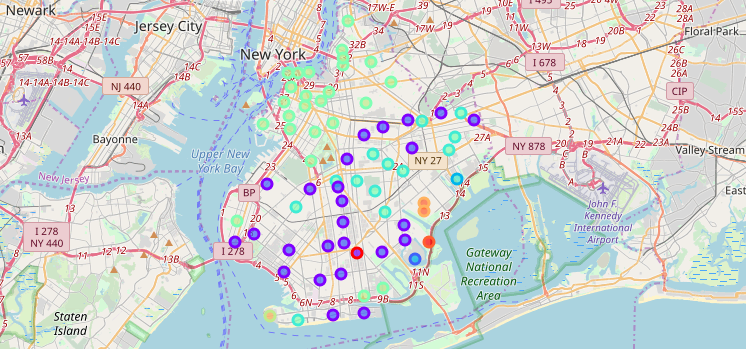
To look at the different venues in Brooklyn I used the Foursquare API. I set the limit of number of venues returned by Foursquare API to 100 and defined radius as 500 meters.



Since there are many similar venues in each of the neighborhoods, I used K-means clustering method to break up Brooklyn into several clusters. To find the best number of clusters I used the Elbow Method, which failed to provide very meaningful results.



Thus, I manually explored different number of clusters to find the most appropriate one. So far, the best number of clusters turned out to be nine. As there are 70 neighborhoods in total and 290 unique venue categories. The number of clusters didn’t seem extra high so I decided to use it in my further analysis. I plotted the clustered neighborhoods for Brooklyn Borough only, coloring each neighborhood according to its cluster.

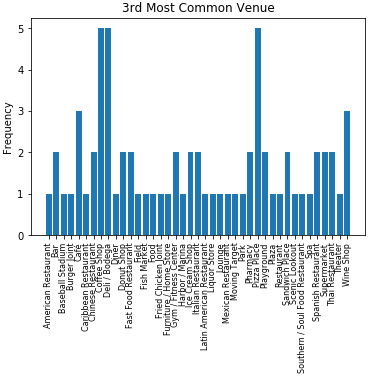
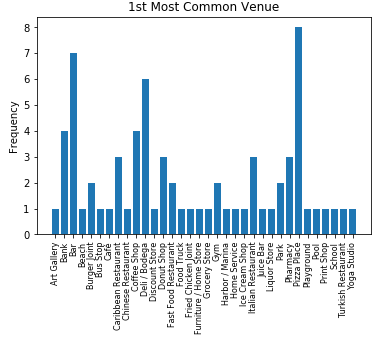
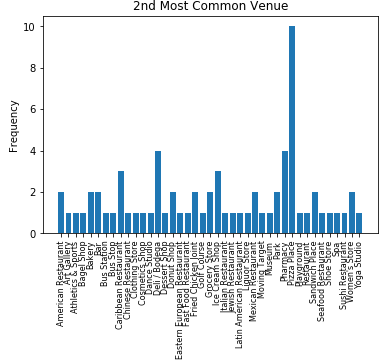


Then I joined the ten most common venues to the master dataset and broke it up into nine different sub datasets, corresponding to each specific cluster.



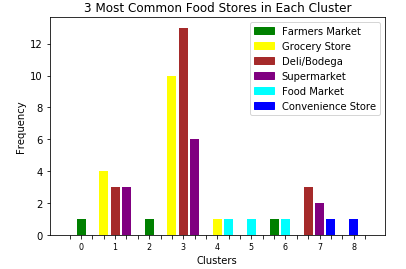
**Results**

I looked at the first, second and third most common venues for all of the neighborhoods in Brooklyn to see how many stores that sell food there are and how popular are they.



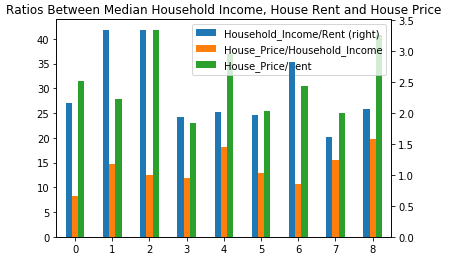
It looks like pizza places, banks, coffee shops and Deli/Bodegas are the most common venues throughout first, second and third most common places.

Furthermore, I looked at the three most common food stores in each of the cluster. According to the plot, cluster three has the most food stores among all clusters. Deli/Bodega is the most common in cluster three, followed by grocery store and supermarkets. Cluster three has similar situation, only smaller number of venues than cluster one.



As we can see from the graph, clusters zero, two, four, five, six and eight have the less number of the stores selling food, thus less competition. On the other hand it can represent smaller customer demand in those clusters, which in turn will yield clusters one and three as the most desirable for food stores, as they have proven to have higher demand, but higher supply at the same time.

My next step was to look and compare median household income, median house price and median rent for each of the clusters. For that I used the three different ratios of the variables. Household income to rent, which provided the descriptive information on how much money average household has after paying their monthly rent; the values were in the range of 1.62 and 3.35, where the higher value meant there were potentially more money left in the household for other expenses. House price to household income, which provided information on how affordable was buying a house in that cluster; values were in the range 8.34 to 19.8, where lower number represented higher affordability. House price to house rent was another ratio I looked at, which can be interpreted as another factor that can affect the willingness of residents to buy a house in a particular cluster; values were in the range of 22.92 to 41.86, where higher number represented bigger disparity between house price and house rent, which in turn can be translated to lesser desirability to buying a house.



Clusters one, three and five and seven had the lowest House price to rent ratio, which meant residents there are more willing to buy a house rather than rent. However, clusters one and seven had the lowest House price to household income ratio, which meant those clusters provided higher house purchase affordability. On the other hand, cluster one had the highest household income to rent ratio, which meant that households in that cluster had more money left after paying for their rent.

Based on the performed data analysis, the potential clusters where we can see the profit, are clusters one and five. Cluster one is because of the analysis of ratios, cluster five because of the lack of food stores in that cluster. Out of which cluster one would be my first suggestion for opening a convenience store.

**Discussion**

Brooklyn is considered a borough with a high population density. For entrepreneurs looking to get into business of grocery sales Brooklyn can be seen as a very desirable place. At the same time, competition is very high there, so it’s not as obvious where can one open up a convenience store and what kind of food to sell. For that matter I performed a simple data analysis to find out, which locations can be a good fit for starting the business of grocery sale.

I used Kmeans clustering with manually selecting the number of clusters, eventually selecting the optimal number as nine clusters. 70 neighborhoods were clustered and then utilizing Foursquare API I extracted the venues in all of the neighborhoods. Using external websites [2-6] I’ve calculated the median household income, house prices and house rents for each of the clusters.

I used several visualization methods to look at the data and results gained from the data. I used bar graphs and tables to look at data and the descriptive statistic, and I used folium maps to look at the neighborhoods from the geographical perspective.

Three types of ratios were used in this project to see how the median household income, house prices and house rents differ in Brooklyn and how they may affect the choice of the location to start a business.

Based on the descriptive statistic and clustering approach I analyzed what are the potential places to open up a convenience store. At the end I made suggestions to potential entrepreneurs where to open a store to sell groceries.

**Conclusion**

Making a business profitable can be a challenging task. According to the Small Business Association more than half of new businesses fail during the first year. Data/business analysis can be crucial for entrepreneur to become successful. Not only in the beginning but throughout the whole journey from start to end, regardless if the goal is to exit the business at some point, or to make it a source of permanent income.

Data can provide valuable insights, if analyzed properly, or it can potentially ruin the business if done incorrectly. I recommend every entrepreneur, new or existing, to learn how to analyze the data for current or future endeavors.

**References**

[1] https://cocl.us/new\_york\_dataset

[2] <https://www.rentcafe.com/average-rent-market-trends/us/ny/brooklyn/>

[3] <https://ny.curbed.com/2017/8/4/16099252/new-york-neighborhood-affordability>

[4] <https://www.census.gov/quickfacts/fact/table/kingscountybrooklynboroughnewyork/INC110217>

[5] <https://www.cnbc.com/2019/01/25/brooklyn-neighborhoods-stay-hot-as-new-yorks-real-estate-market-cools.html>

[6] <https://www.bloomberg.com/graphics/property-prices/nyc/>