COURSERA CAPSTONE FINAL ASSIGNMENT: CLUSTERING NEW YORK FOR OPENING ORGANIC GROCERY STORES

INTRODUCTION

In the previous labs we examined how the neighborhoods of cities could be segmented by the types of business locations in each neighborhood. In this project, I am interested to explore New York again. The aim of my project is to examine NY Boroughs for opening organic grocery stores.

BUSINESS PROBLEM / BACKGROUND

In the last years, the consumers' interest and market for organic products, stores and restaurants is continuously growing. For this Project I work for a organic grocery franchise. The franchises aim is to open new stores in metropole regions like New York. But of course, other retailers realized the potential of this market as well. So in New York, like in almost every city, the density of organic grocery stores is rising and the earnings are increasingly shared by many providers. The Business Problem is to find out which Borough of New York is best fitting for opening new organic based grocery stores.

To answer the question which areas show the best market potentials for opening these stores I consider the following variables:

- Population Data of the five Boroughs of New York
- Land area sizes of the five Boroughs of New York
- Average per Capita income of the five Boroughs of New York
- Number of existing organic based grocery stores in the five Boroughs of New York

Other variables, like rentals or leases were not taken into account to answer the question but will be discussed in the Discussion/Conclusion section.

DATA

In the following lines I will provide information about the Data I use for my analyzes and from where I got these information. Also I will give an overview how the Data will be cleaned up and prepared to drive these analyzes.

DATA 1: NY POPULATION DATA

In this first step I gather population data of New York by the five Boroughs (Bronx, Brooklyn, Manhattan, Queens, Staten Island).

The Dataset is downloaded from http://app.coredata.nyc as .csv and uploaded to github repository.

This Dataframe includes Population Data from 2000 to 2018 for the five Boroughs of New York. Below you see the Dataframe (only including Data from 2018):

Borough	2018
Staten Island	476179
Manhattan	1628700
Bronx	1432130
Brooklyn	2582830
Queens	2278910

DATA 2: BOROUGH AREA SIZE

In this step, the NY area per Borough is being scraped from Wikipedia: https://en.wikipedia.org/wiki/Boroughs_of_New_York_City

For my analyzes only the column of sqaurekm is relevant. Below you see these Data:

Borough	squarekm
The Bronx	109.04
Brooklyn	183.42
Manhattan	59.13
Queens	281.09
Staten Island	151.18

DATA 3: NY MEAN HOUSEHOLD INCOME DATA

In this section data will be downloaded from https://www.census.gov and uploaded to github repository.

The internet source include many information about New York population, income, education and so on. For my purpose I will just read the rows with the average per capita income for the five Boroughs:

Borough	**Manhattan**	**Bronx**	**Queens**	**Brooklyn**	**Staten Island**
Per capita income in past 12 months (in 2018 d...	72832	20850	30289	31984	34987

Data 4: Get venues per Borough from Foursquare

Here I will use Foursquare to get information about existing organic grocery stores in the Boroughs of New York. In the first step I gather information where the current stores are located. In a second step I will use the API the check, how many other organic grocery stores are located within 500 meters to each of the stores:

Borough	Venue	Venue_Lat	Venue_Long	no_venues_500
Manhattan	Food Story Natural Market	40.773664	-73.913919	15
Staten Island	Food Story Natural Market	40.773664	-73.913919	15
Queens	Food Story Natural Market	40.773664	-73.913919	15
Bronx	Food Story Natural Market	40.773664	-73.913919	15
Staten Island	MidCity Farms	40.773799	-73.914041	11

Methodology

In this section I will perform the following tasks:

- Organize and prepare gathered Dataframes to merge them into one Dataframe
- Merging
- Visualizing the Data
- K-Means Clustering

Step 1: Organizing DataFrames

In all of these steps, the Dataframes are cleaned up (removing $ from values, organizing column names and so on) in order to merge the Data into one Dataframe which includes the relevant Data for the upcoming analyzes.

STEP 2: MERGING DATA

In the second step, the data from the different sources are being merged to have one DataFrame to work with. Furthermore some calculations will be made on the Data:

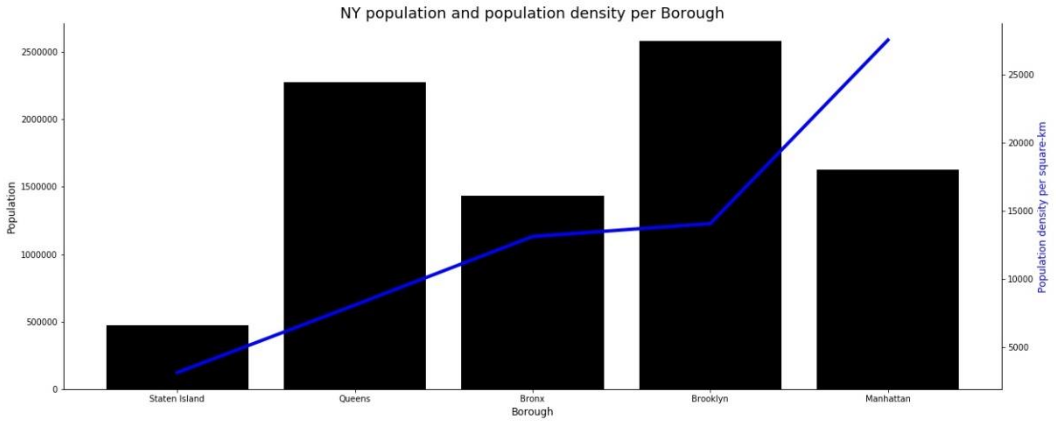
- Calculation of the population density per km^2
- Calculation of the estimated total income per Borough
- Calculation of the density of organic stores per km^2
- Calculation of the Population-Store-ratio for each Borough

Borough	Population_density	Total estimated income per Borough	No_organic	organic_density_per_km^2	Population-Store-ratio
Staten Island	3149.75	16660074673	96	0.64	4960
Manhattan	27544.39	118621478400	64	1.08	25448
Bronx	13133.99	29859910500	94	0.86	15235
Brooklyn	14081.51	82609234720	52	0.28	49669
Queens	8107.40	69025904990	13	0.05	175300

STEP 3: EXPLORATORY DATA ANALYSIS

In this section Visualization will be used to explore the Data and understand the distribution of population, income and organic grocery stores in the Boroughs of New York.

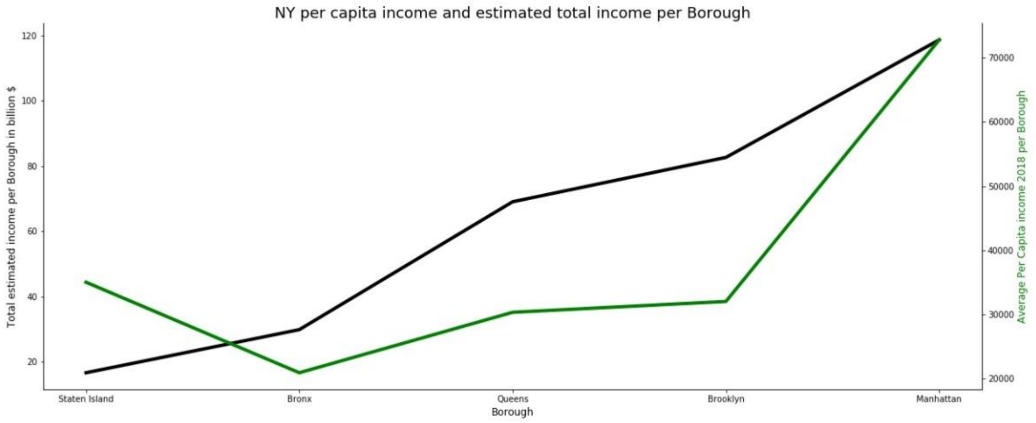
VISUALIZING DATA: POPULATION



As seen in the above population chart, the five Boroughs of New York differ significantly in their total population as well as in population density per km²:

- Staten Island has the lowest total population and the lowest population density.
- Manhattan has the highest population density with an average total population.
- Bronx, Brooklyn and Queens have average population densities but high total population, especially Brooklyn and Queens.

Visualizing Data: income

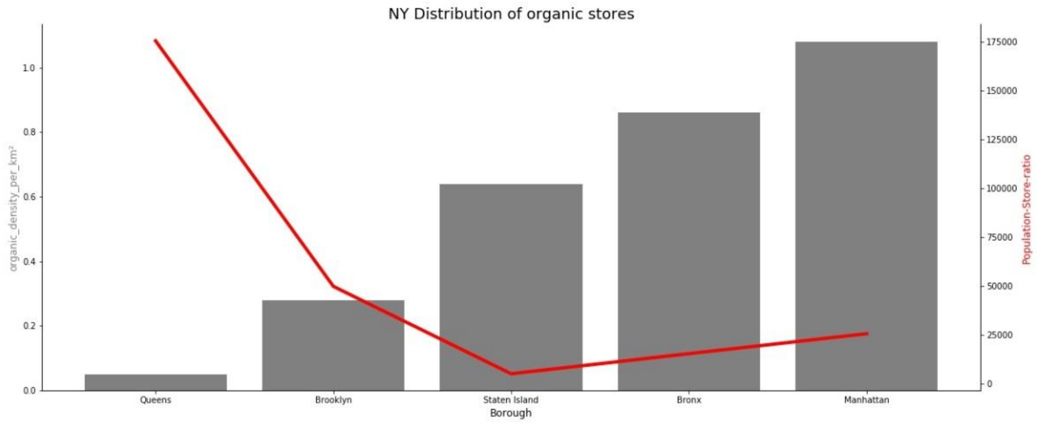


The above Chart shows that in most Boroughs in New York, the per capita income and the estimated total income in each Borough are correlated, except Staten Island. This means that in Staten Island the total income is relatively low and is shared by only a relatively small number or people. Staten Island has also the second highest per capita income in New York. In Manhattan people have the highest per capita incomes and its the Borough with the highest total purchase power in New York.

VISUALIZING DATA: VENUES

In this section I will show the venues per capita and venues per km^2 in a bar chart as well as the distribution of the organic stores in maps.

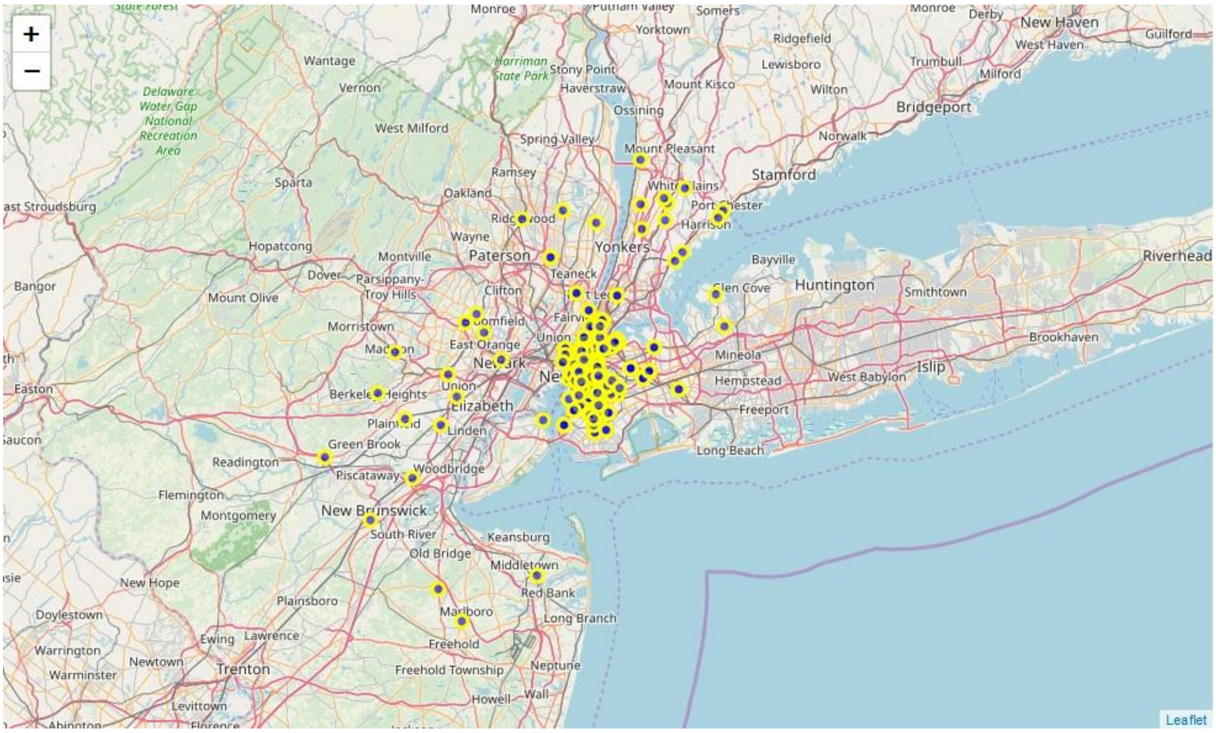
Venues Chart

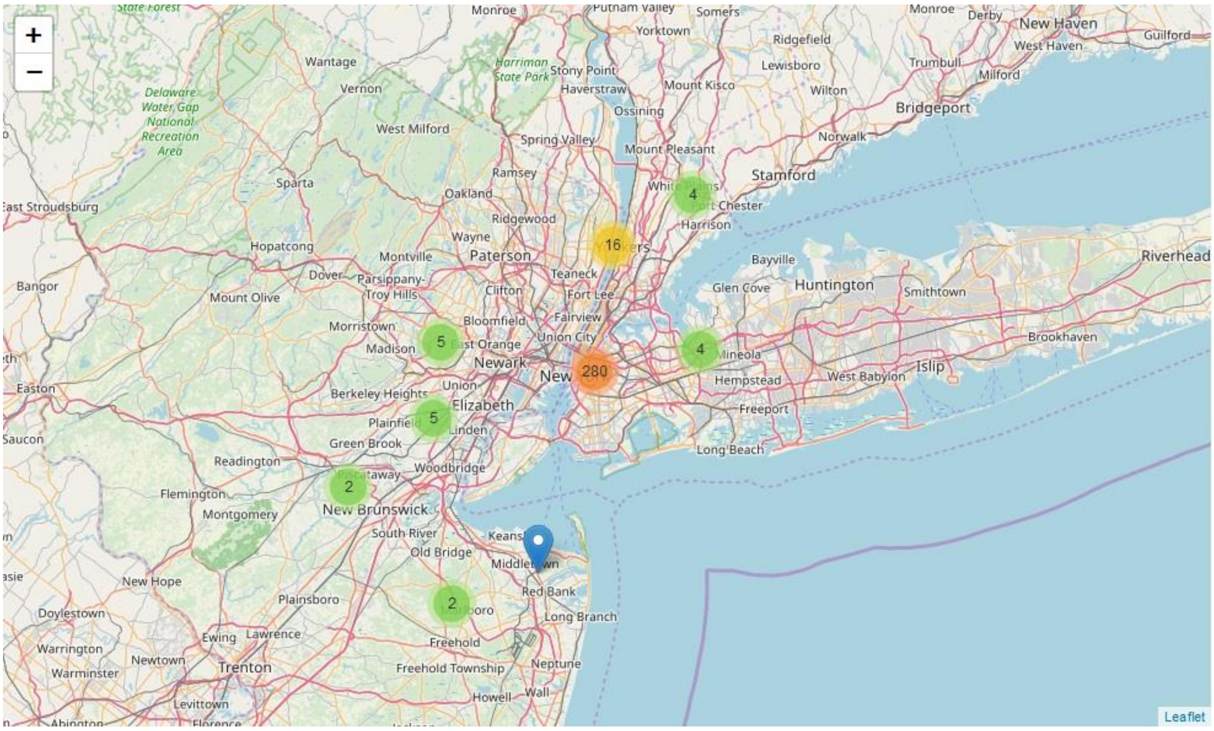


In this Chart we see that the Boroughs differ in their distribution of organic stores. While in Queens there are just a few stores, in Manhattan there is a plethora of organic stores. On the other hand, in Queens there comes one organic store on about 175,000 people while in Manhattan one store comes for about 25,000 people.

Venues maps

In the following maps we see the distribution of existing organic grocery stores in the Boroughs of New York. The second map shows merged Dataspots the visualize areas with high density of stores.





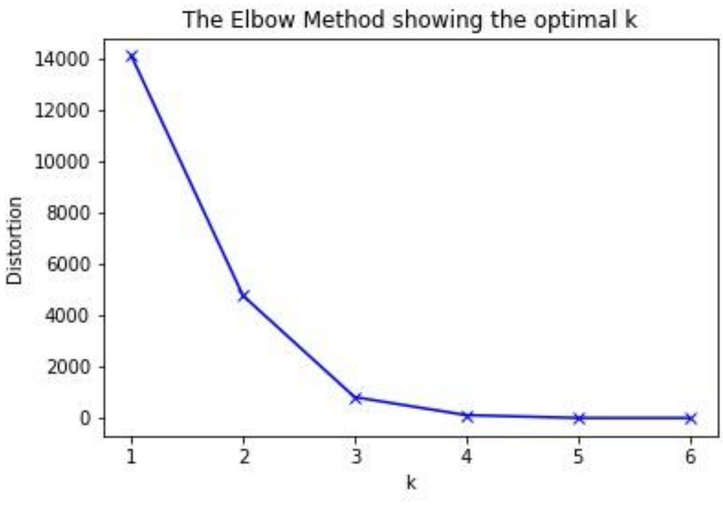
STEP 4: K-MEANS CLUSTERING

As these analyzes show, per capita income as well as the distribution of organic stores do differ in the Boroughs of New York. Therefore, for the next step of clustering the Data it is helpful to create a new DataFrame with only the relevant Data. Information for Clustering are

- Average per Capita income
- density of stores per km^2
- No of stores around 500 meters

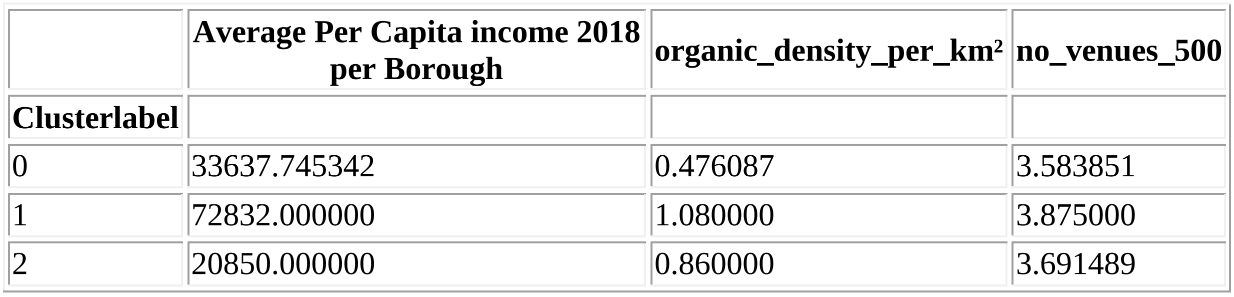
To check how many Clusters do exist based on these information the elbow method is performed as step one. In step 2 a map is being generated to visualize the different Clusters.

Step 1: Determine Elbow-Point



As this Chart shows, k = 3 fits best for the Data.

Step 2: Clustering



	Average Per Capita income 2018 per Borough	**organic_density_per_km^2**	**no_venues_500**
Clusterlabel			
0	33637.745342	0.476087	3.583851
1	72832.000000	1.080000	3.875000
2	20850.000000	0.860000	3.691489

Based on this information we see 3 Clusters which can be described as follows:

- Cluster 0: middle income; low density of existing stores; low number of stores around 500 meters

- Cluster 1: high income; high density of existing stores; high number of stores around 500 meters

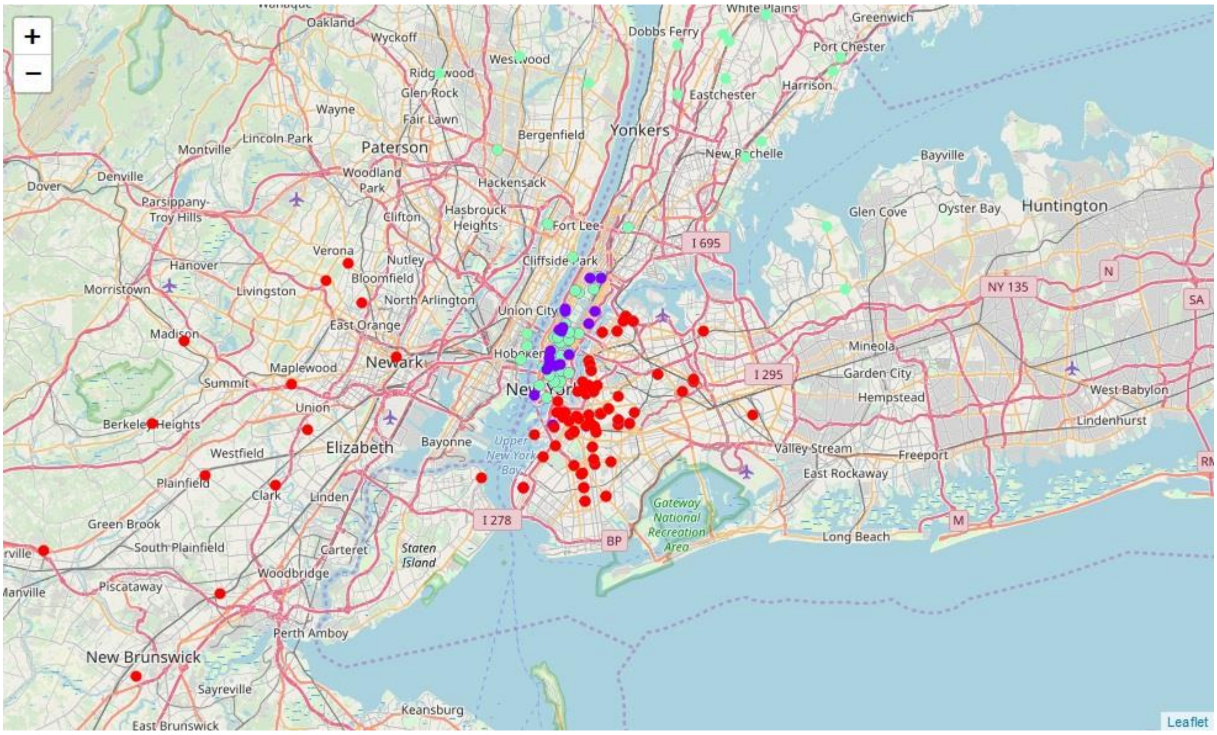
- Cluster 2: low income; relatively high density of existing stores; relatively high number of stores around 500 meters

As we seek for locations with high per capita income, low density of existing stores and low number of stores around 500 meters, two Clusters show good potential for opening new stores:

- Cluster 0 has the lowest density and number of stores around as well as a solid income-structure

- Cluster 1 has already an existing market for organic stores but shows a very high-income structure

To sum up, locations for new stores should be searched nearby stores of Clusters 0 and 1.



RESULTS

As shown in the previous analysis, New York organic stores can be clustered into 3 Cluster considering income structure, density of existing organic stores and number of organic stores around 500 meters of existing stores.

For opening new organic stores, I would suggest to search for locations with high income structures and low density of organic stores. The analysis shows that this situation is best met in Cluster 0 and 1.

These two Clusters extend mainly over Staten Island, Brooklyn (red dots in map) and parts of Manhattan (purple dots in map). Considering the population in the Boroughs (see Chart: NY population and population density per Borough), Brooklyn with its high population seems to be the Borough with highest potentials for opening new organic based grocery stores.

DISCUSSION

To sum up, organic products are especially bought by people with relatively high incomes. Another factor is clearly the already existing market (saturated markets vs. potential markets). These factors were included in the above analysis and showed that 3 clusters show up in New York. All of those Clusters do differ but two of them can be seen as potential Clusters. Cluster 0 has a low density of existing stores and a solid income structure while Cluster 1 has very high-income structure. Contrary, Cluster 2 has low income structures as well as relatively saturated markets and should not be preferred to open new stores.

CONCLUSION

To conclude this report it is necessary to say that the above analysis is based on just a simple set of factors being considered. Factors like rentals, leases or even purchase prices for new stores are not considered. Also one could even go deeper into each Neighborhood or even streets. This analysis was meant to search for top-level suggestions and future analyzes should consider these recommendations. Nevertheless, this report gives a first impression of the structures of organic stores in the NY Boroughs in order to recommend deeper studies within this Borough or seek for possible Retail outlets.