Coursera Data Science Capstone Project

Analyzing restaurant and demographic distribution in New York City

May 01, 2020

1. Introduction

1.1 Description of the Background

New York City is the top international and densest city in the US. The 5 boroughs - Manhattan, Bronx, Queens, Brooklyn, Staten Island - differ a lot in their variety. The city is ethnically diverse and famous for its commercial city center, Times Square. It is one of the largest megacities in the world, considering the landmass New York City is the largest one. [1] [2]
As New York City is such a large international city, there are also a lot of different restaurants around the city. Just in the year 2018, the number of restaurants in New York City grew to a number of **27.043** and the number is still growing. [3]
With the growing number of restaurants, also the competition grows. For a client, opening up a restaurant in New York City, it is important to look at the existing data to help make a better decision about which kind of restaurant and where to open it.
For that case the favorite cuisine style of all 5 boroughs of New York City will be examined, to get a detailed look at the favorite tastes. Also, the demographic distribution will be considered to get a look at cultural diversity.
The study will focus on Asian and Hispanic demographic analysis.
Asian and Hispanic restaurants have a high potential for a new restaurant, as the population is fast growing. Asians are in fact the fastest-growing ethnic group and New York City owns the biggest population of Asians in the US[4]. Hispanics are also a fast-growing population in New York City and makeup 29% of the total New York City population.

1.2 Data

For analyzing potential Hispanic and Asian restaurants and the place to open one, different data sources will be used. For a profound decision multiple factors influence the decision:

- Distribution of different restaurants in all boroughs of NYC
- Demographic distribution of Hispanics and Asian in the boroughs of NYC

The following data sources will be used:

- A Geojson file to get the location data of NYC. This file is important for visualization and the coordinates of the boroughs in NYC.
- The number, type, and location of all restaurants will be obtained using the **Foursquare API.** The Foursquare API makes it possible to receive information and coordinates of a venue. [7] Only the latitude and longitude are needed to get the data from Foursquare. The complete list is after that filtered for restaurants, to get only relevant venues for this problem.
- Census data of NYC to get the demographic distribution of Hispanics and Asians. The data will be obtained from **Kaggle.** The Census data is built up of two CSV files. One contains the percentage of ethnic demographic information (for example Hispanics) for a specific latitude/longitude. The other CSV file contains the coordinates with Census numbers. These two files will be cleaned and merged, to get one dataframe. [6]

2. Methodology

2.1 Get the coordinates of NYC

First, a Geojson file with coordinates of New York City is opened and written into a dataframe. The resulting dataframe in Table 1 consists of the **Borough, Neighborhood, Latitude, Longitude.**
In total there are **5 boroughs** and **306 neighborhoods**.

	Borough	Neighborhood	Latitude	Longitude
0	Bronx	Wakefield	40.894705	-73.847201
1	Bronx	Co-op City	40.874294	-73.829939
2	Bronx	Eastchester	40.887556	-73.827806
3	Bronx	Fieldston	40.895437	-73.905643
4	Bronx	Riverdale	40.890834	-73.912585

Table 1: Dataframe of the New York City boroughs with latitude, longitude

The pandas dataframe will be filtered for every single borough so that in the end every borough got his own dataframe. For the first visualization, **Folium** is used. This is a Python package, which makes it possible to visualize data on an interactive map [5]. Every borough is visualized on a map in Image 1 with a different color so that the different boroughs can be easily seen. The single points on the map are the neighborhoods in a borough with a specific latitude, longitude.

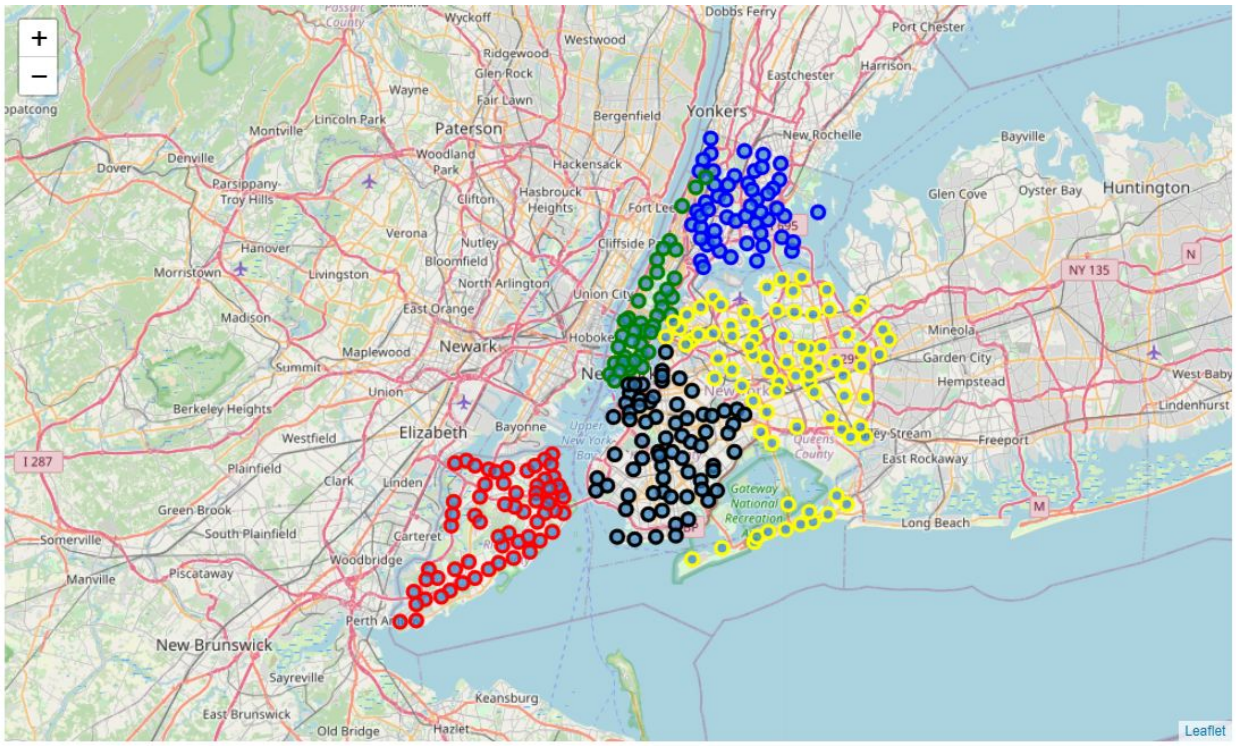

Image 1: Visualization of all boroughs in New York City. Green is Manhattan, blue is the Bronx, red is Staten Island, black is Brooklyn, and yellow is Queens.

2.2 Handle venue data from Foursquare and explore restaurant distributions in boroughs

In the next step, the venues from Foursquare are retrieved. For that case the API endpoint **explore** is addressed with the latitude and longitude of all neighborhoods and responses with the venue name, venue location, and venue category. This has to be done with all 5 boroughs. All the data from Foursquare is saved in 5 different dataframes for every borough. In Table 2 the head data for Manhattan is visible and only for Manhattan **2965** venues were retrieved.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Marble Hill	40.876551	-73.91066	Golden City Chinese Restaurant	40.879319	-73.906187	Chinese Restaurant
1	Marble Hill	40.876551	-73.91066	blue bay restaurant	40.879181	-73.910093	Diner
2	Marble Hill	40.876551	-73.91066	China Wang Restaurant	40.877432	-73.906981	Food
3	Marble Hill	40.876551	-73.91066	El Economico Restaurant	40.879330	-73.904597	Spanish Restaurant
4	Marble Hill	40.876551	-73.91066	Medio Restaurant Coffee Shop	40.880832	-73.908419	Food

Table 2: The dataframe output from manhattan_venues.head() with the venue data from Foursquare

Every dataframe with venue data from Foursquare is subsequently processed. The dataframes are filtered after the **Venue Category** to receive only categories with the name *restaurant* in it. After we get the restaurant filtered dataframes, the different restaurant types will be counted. This has to be done to get an impression of which restaurant type is more common and recommended. In the end a bar plot will show the top 5 restaurant types for every borough. This will lead to the first impression of restaurant differences in the boroughs.

Manhattan (Image 2) has got a high number of Italian restaurants with **117**. Also, Asian style restaurants (Chinese and Japanese) are often recommended, in a total of **88** restaurants.
In the Bronx (Image 3) there are a lot of Hispanic style restaurants. In total there are **57 Hispanic restaurants**.

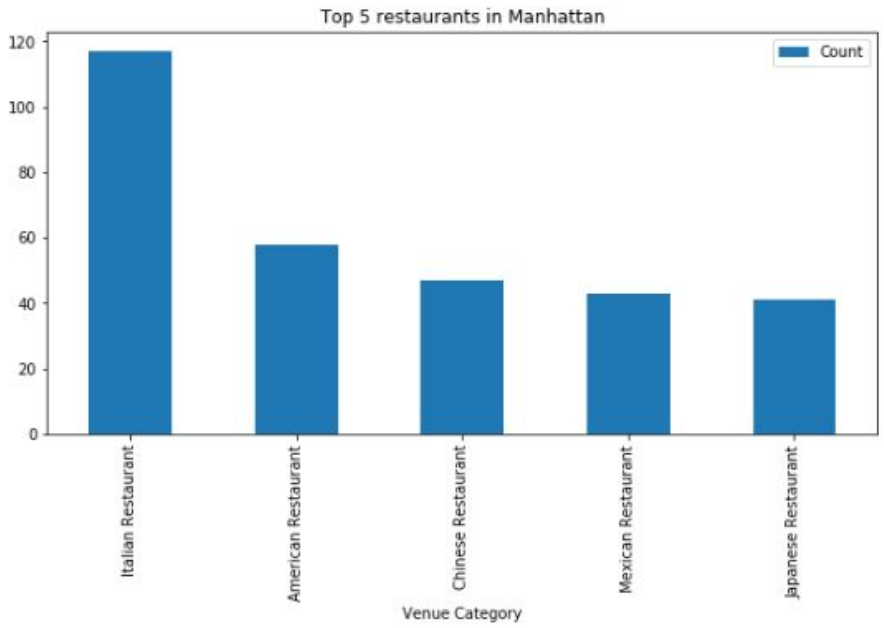

Image 2: The top 5 restaurant types in Manhattan

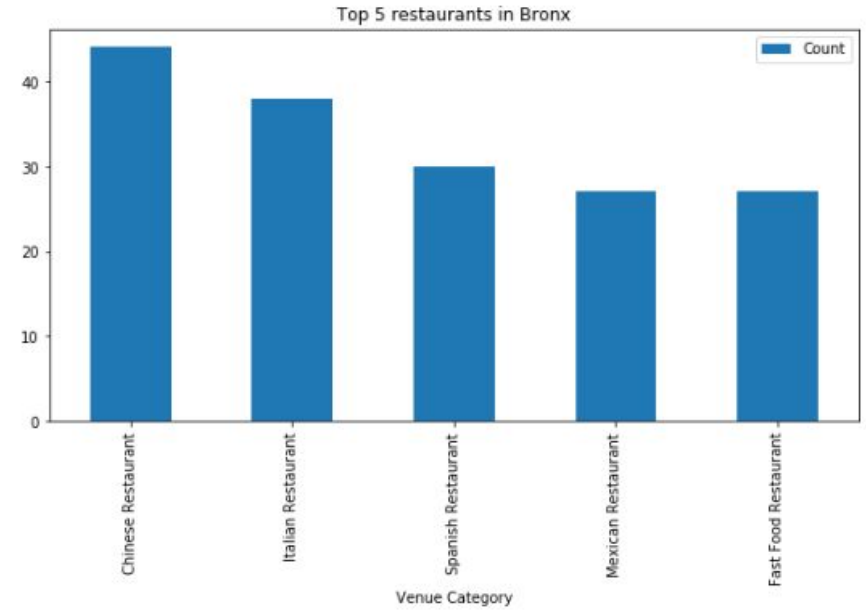

Image 3: The top 5 restaurant types in the Bronx

In Queens (Image 4) Asian restaurants are more often recommended. In total, Chinese and Korean restaurants, there are **96 Asian restaurants**.

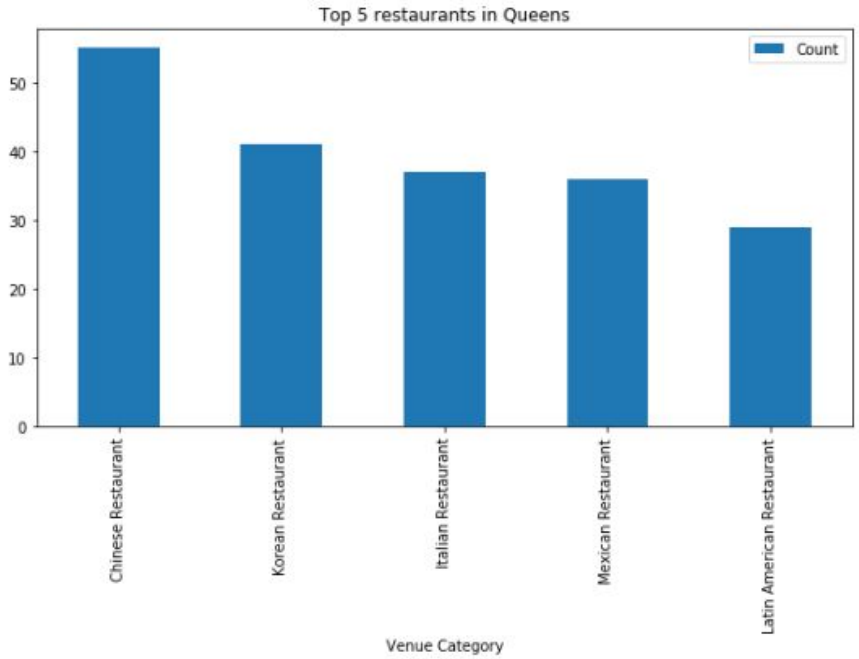

Image 4: The top 5 restaurant types in Queens

Brooklyn (Image 5) has got the highest number of Italian restaurants after Manhattan. Also, there are **46 Hispanic style restaurant**s and **50 Asian style restaurants**.

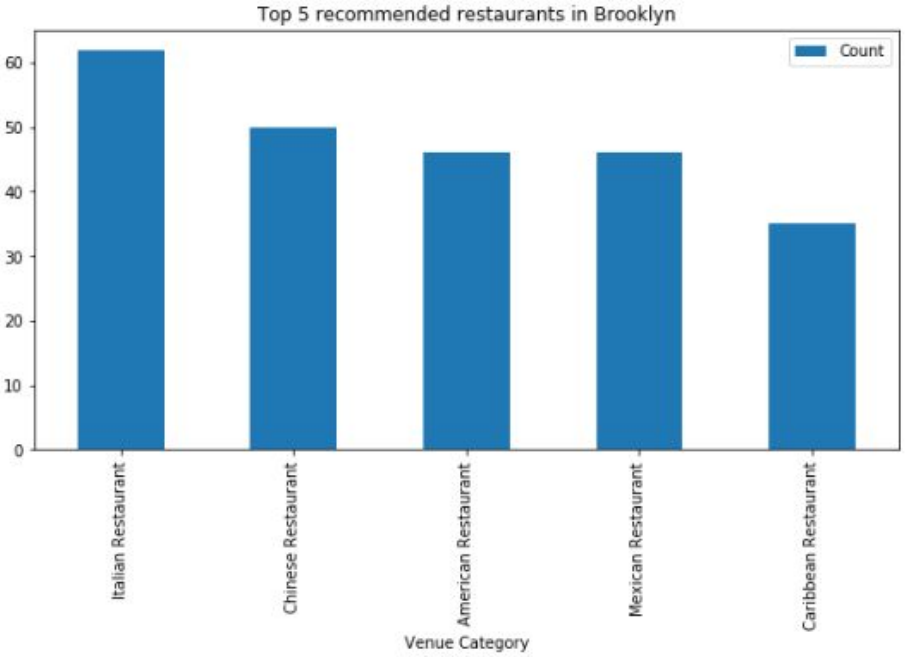

Image 5: The top 5 restaurant types in Brooklyn

Staten Island (Image 6) is the borough with fewer Chinese restaurants and less Spanish/Mexican restaurants.

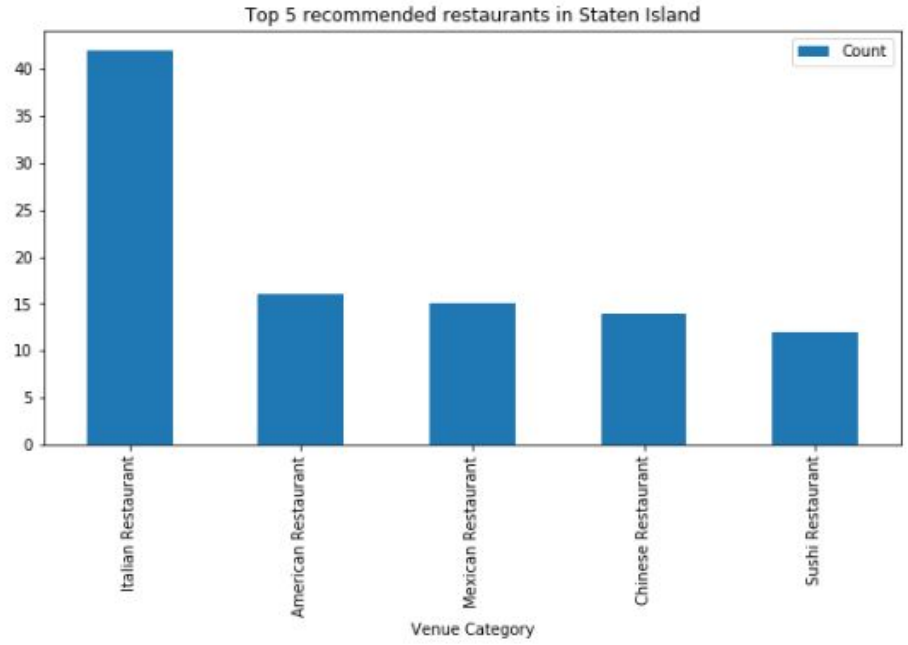

Image 6: The top 5 restaurants in Staten Island

2.3 Use Kaggle Census Data to analyze the demographic distribution in the Boroughs

The Census data from Kaggle is used to make a statistical analysis of the boroughs in NYC. The **Kaggle Dataset** offers two CSV files, which are read into a dataframe. The first one contains the demographic data for every latitude and longitude in NYC. It shows the percentage of for example Hispanics in a specific latitude and longitude of the boroughs, please refer to Table 2. The second CSV file contains the coordinates and Blockcode. Both dataframes will be merged in the next step with the column CensusTract and BlockCode to get only one dataframe.
The resulting dataframe is cleaned so that every **NAN** in the rows is removed.

	CensusTract	County	Borough	TotalPop	Men	Women	Hispanic	White	Black	Native	Asian	Citizen	Income	IncomeErr	IncomePerCap	IncomePerCap
0	36005000100	Bronx	Bronx	7703	7133	570	29.9	6.1	60.9	0.2	1.6	6476	NaN	NaN	2440.0	37
1	36005000200	Bronx	Bronx	5403	2659	2744	75.8	2.3	16.0	0.0	4.2	3639	72034.0	13991.0	22180.0	220
2	36005000400	Bronx	Bronx	5915	2896	3019	62.7	3.6	30.7	0.0	0.3	4100	74836.0	8407.0	27700.0	244
3	36005001600	Bronx	Bronx	5879	2558	3321	65.1	1.6	32.4	0.0	0.0	3536	32312.0	6859.0	17526.0	294
4	36005001900	Bronx	Bronx	2591	1206	1385	55.4	9.0	29.0	0.0	2.1	1557	37936.0	3771.0	17986.0	269

Table 2: The dataframe with demographic data from the Census data

	Latitude	Longitude	BlockCode	County	State
0	40.48	-74.280000	340230076002012	Middlesex	NJ
1	40.48	-74.276834	340230076005000	Middlesex	NJ
2	40.48	-74.273668	340230076003018	Middlesex	NJ
3	40.48	-74.270503	340230076003004	Middlesex	NJ
4	40.48	-74.267337	340230074021000	Middlesex	NJ

Table 3: The dataframe with coordinates of NYC

	CensusTract	County_x	Borough	TotalPop	Men	Women	Hispanic	White	Black	Native	Asian	Citizen	Income	IncomeErr	IncomePerCap	IncomePerC
0	36005000200	Bronx	Bronx	5403	2659	2744	75.8	2.3	16.0	0.0	4.2	3639	72034.0	13991.0	22180.0	2
1	36005000200	Bronx	Bronx	5403	2659	2744	75.8	2.3	16.0	0.0	4.2	3639	72034.0	13991.0	22180.0	2
2	36005000200	Bronx	Bronx	5403	2659	2744	75.8	2.3	16.0	0.0	4.2	3639	72034.0	13991.0	22180.0	2
3	36005000200	Bronx	Bronx	5403	2659	2744	75.8	2.3	16.0	0.0	4.2	3639	72034.0	13991.0	22180.0	2
4	36005000200	Bronx	Bronx	5403	2659	2744	75.8	2.3	16.0	0.0	4.2	3639	72034.0	13991.0	22180.0	2

Table 4: The merged and cleaned dataframe for the Census Data

The data for Hispanics and Asians are analyzed with the merged dataframe. The data for every borough is aggregated for Hispanics and Asians to receive the mean value. The aggregated value for every borough is then used for visualization of a **choropleth** map with Folium. This is done for the Hispanic and Asian demographic distribution.

Results of demographic analysis

Hispanics are on average mostly in the Bronx, as the dark red color of the map in Image 7 indicates. Also, Manhattan and Queens are showing more Hispanic population on average, than the boroughs Brooklyn and Staten Island.

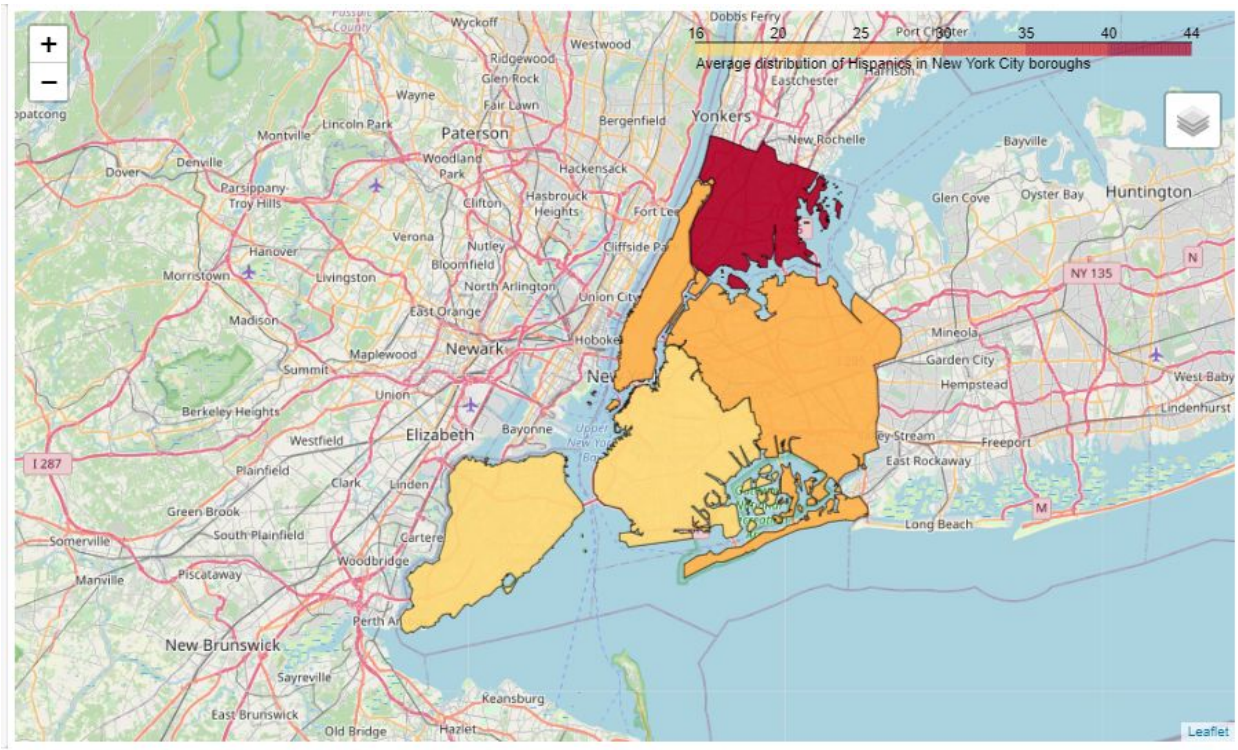

Image 7: Choropleth map of the Hispanic distribution in the boroughs

The Asian population is on average especially in Queens represented. After that Manhattan has got a high volume of Asians. This can be easily seen in Image 8.

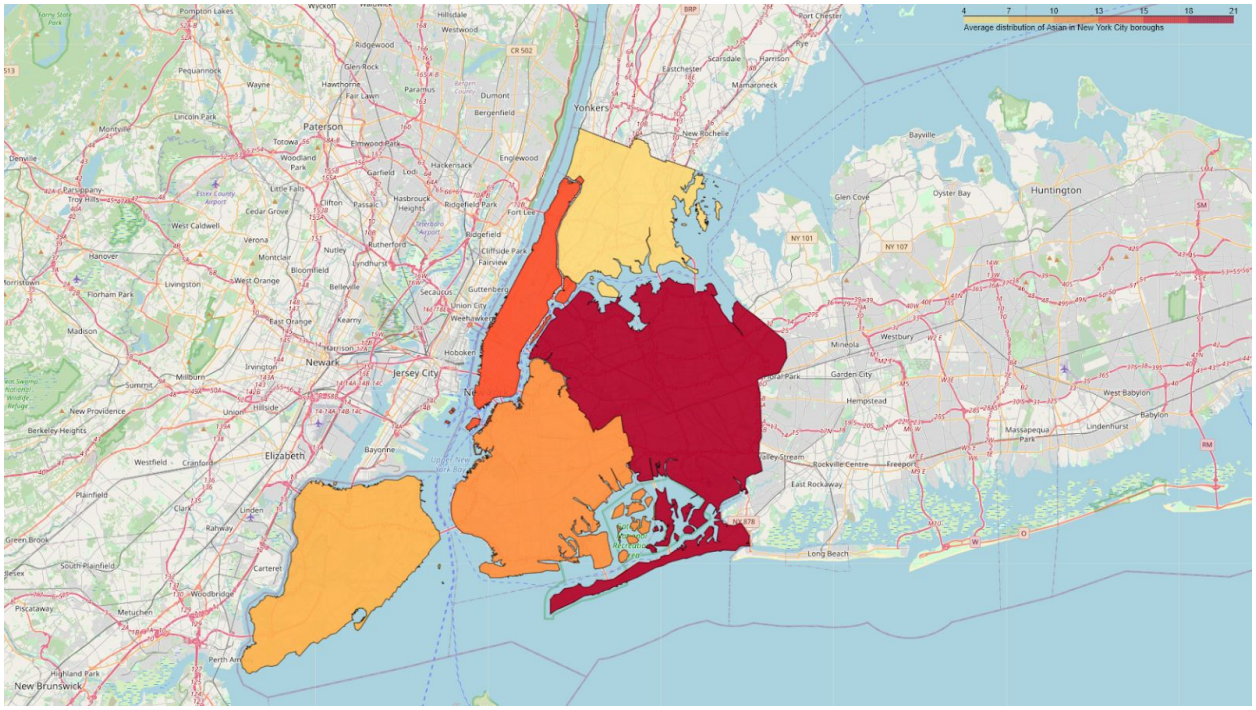

Image 8: Choropleth map of the Asian distribution in the boroughs

3.Cluster the Neighborhoods in the Boroughs with k-means

For Clustering, the neighborhoods in all 5 boroughs some further preprocessing steps have to be done. First, all Venue Categories from the venue data from Foursquare needs to be one-hot encoded, to get only numerical data for using k-means.

3.1 One hot encoding and sorting the most common

All the restaurant-style venues are one hot encoded with the pandas function pd.dummies(). The resulting dataframe in Table 5 shows for every neighborhood a 1 if the specific restaurant is available and a 0 if not. This dataframe is then grouped for all Neighborhoods and taking the mean of the frequency of occurrence of each category. This results in a frequency table (Table 6) for the occurrence of restaurant types.

	Neighborhood	Afghan Restaurant	African Restaurant	American Restaurant	Arepa Restaurant	Argentinian Restaurant	Asian Restaurant	Australian Restaurant	Austrian Restaurant	Belgian Restaurant	Cambodian Restaurant	Cantonese Restaurant
0	Marble Hill	0	0	0	0	0	0	0	0	0	0	0
1	Marble Hill	0	0	1	0	0	0	0	0	0	0	0
2	Marble Hill	0	0	1	0	0	0	0	0	0	0	0
3	Chinatown	0	0	0	0	0	0	0	0	0	0	0
4	Chinatown	0	0	0	0	0	0	0	0	0	0	0

Table 5: One-hot encoded venue dataframe

	Neighborhood	Afghan Restaurant	African Restaurant	American Restaurant	Arepa Restaurant	Argentinian Restaurant	Asian Restaurant	Australian Restaurant	Austrian Restaurant
0	Battery Park City	0.0	0.000000	0.000000	0.0	0.00	0.0	0.000000	0.000000
1	Carnegie Hill	0.0	0.000000	0.050000	0.0	0.05	0.0	0.000000	0.000000
2	Central Harlem	0.0	0.142857	0.142857	0.0	0.00	0.0	0.000000	0.000000
3	Chelsea	0.0	0.000000	0.111111	0.0	0.00	0.0	0.000000	0.000000
4	Chinatown	0.0	0.000000	0.023256	0.0	0.00	0.0	0.023256	0.023256

Table 6: Frequency table of restaurants in the neighborhoods

The knowledge of the frequency of restaurant venues can be used to sort the 10 most common venues for all the neighborhoods. This can be seen in Table 7 for Manhattan.

	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	Battery Park City	Chinese Restaurant	Mexican Restaurant	Vietnamese Restaurant	Greek Restaurant	Ethiopian Restaurant	Falafel Restaurant	Fast Food Restaurant	Filipino Restaurant	French Restaurant	German Restaurant
1	Carnegie Hill	Italian Restaurant	Japanese Restaurant	Vietnamese Restaurant	French Restaurant	Restaurant	American Restaurant	Argentinian Restaurant	Fast Food Restaurant	Indian Restaurant	Mexican Restaurant
2	Central Harlem	African Restaurant	American Restaurant	Seafood Restaurant	Chinese Restaurant	French Restaurant	Ethiopian Restaurant	Caribbean Restaurant	Tapas Restaurant	Southern / Soul Food Restaurant	Falafel Restaurant
3	Chelsea	Italian Restaurant	Seafood Restaurant	American Restaurant	Mediterranean Restaurant	Japanese Restaurant	Indian Restaurant	Thai Restaurant	Tapas Restaurant	Ramen Restaurant	Restaurant
4	Chinatown	Chinese Restaurant	Malay Restaurant	Vietnamese Restaurant	Vegetarian / Vegan Restaurant	Dumpling Restaurant	Dim Sum Restaurant	Shanghai Restaurant	Taiwanese Restaurant	Cantonese Restaurant	Japanese Restaurant

Table 7: The most common restaurants in the neighborhoods of Manhattan

3.2 K-means clustering the neighborhoods

For clustering, the neighborhoods in New York City the *K-means cluster algorithm* is used. This machine-learning algorithm is used to cluster similar neighborhoods into one. This makes it possible to see the distribution of restaurants in the boroughs and makes it possible to compare the boroughs. In this case, K-means is executed with 5 clusters, so that the algorithm clusters the restaurants into 5 different clusters. Also as input for the algorithm is the frequency dataframe (Table 6) used, but the **Neighborhood** column is removed beforehand so that the dataframe only contains numeric data. The K-means algorithm is executed on all 5 boroughs and outputs the *Cluster Labels*.

3.3 Visualize the Cluster Labels with Folium

The Cluster Labels can now be visualized on a Folium map for every borough. In the following images 9 -13, the different boroughs with the 5 clusters from the k-means cluster algorithm are visualized.

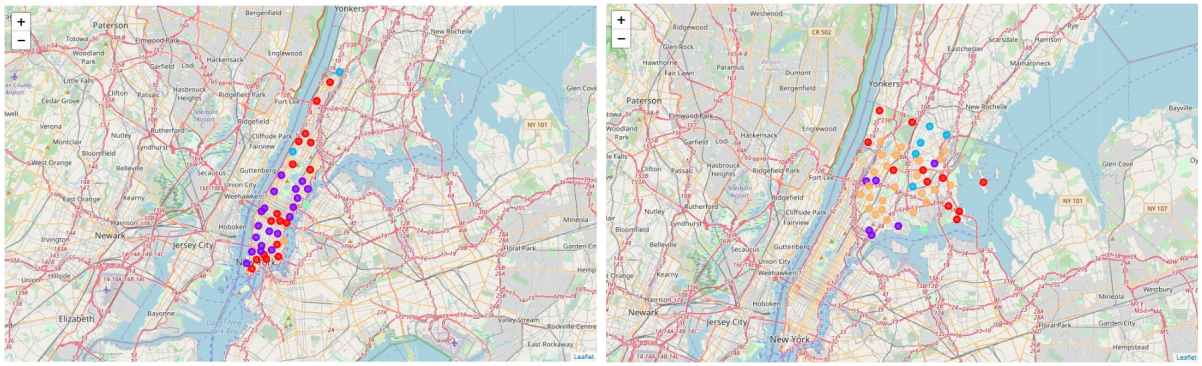

Image 9 and Image 10: The cluster distribution in Manhattan (left) and the cluster in the Bronx (right)

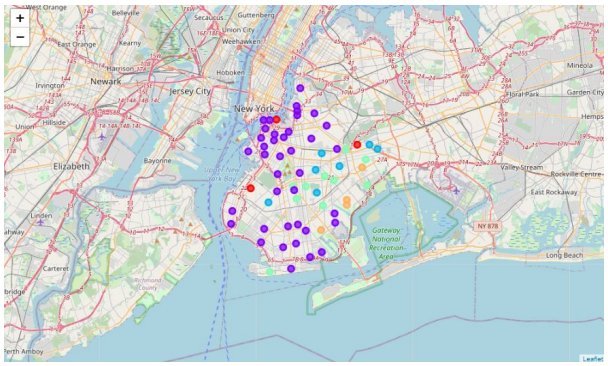
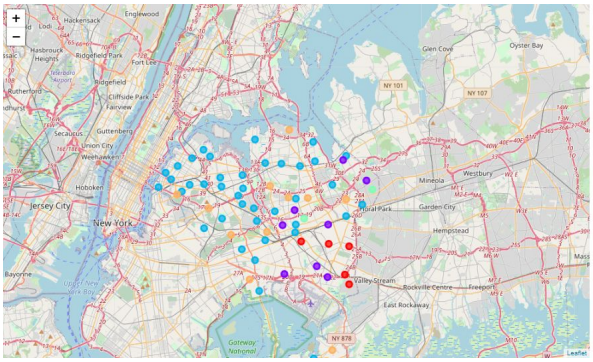

Image 11 and Image 12: Cluster Label in Brooklyn (left) and in Queens (right)

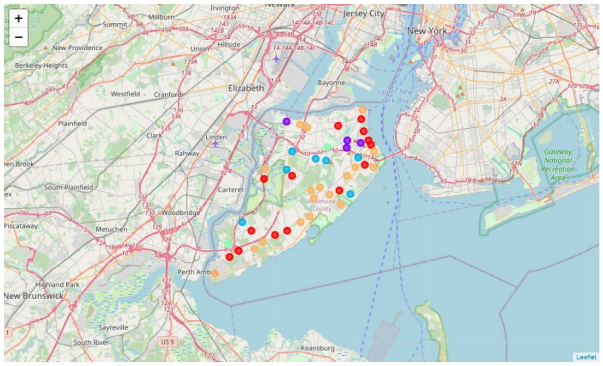

Image 13: Cluster Labels in Staten Island

3. Results

Multiple analysis has been made to help clients open up a restaurant in New York City. This analysis is relevant, as it gives more insights into the actual situation, which makes decisions easier. The first analysis in Section 2.2 was made about the most common restaurants in the boroughs. In section 2.3 the demographic distribution of Hispanics and Asians in New York City was visualized with a choropleth map.

The demographic and common restaurant information can be combined. The following list shows the boroughs with most Hispanics in descending order and the number of **Hispanic** restaurants that are actually there.

1. Bronx: **57** Hispanic restaurants
2. Manhattan: **43** Hispanic restaurants
3. Queens: **65** Hispanic restaurants
4. Brooklyn: **46** Hispanic restaurants
5. Staten Island: **15** Hispanic restaurants

The list makes it more clear that Manhattan has got more Hispanics on average than all the other boroughs except the Bronx, but not that many Hispanic restaurants than other boroughs. In Queens, there are the most Hispanic restaurants with **65** in total, but the Bronx and Manhattan with more Hispanics on average got less with **57 and 43** restaurants.

The following list shows the boroughs with most Asians in descending order and the number of **Asian** restaurants that are actually there.

1. Queens: **96** Asian restaurants
2. Manhattan: **88** Asian restaurants
3. Brooklyn: **50** Asian restaurants
4. Bronx: **44** Asian restaurants
5. Staten Island: **26** Asian restaurants

The amount of Asian restaurants correlates with the demographic distribution in the boroughs. The borough with the most Asian, Queens, also got the most Asian restaurants.
Most common restaurants in all boroughs, independent of demographic distribution, are Italian restaurants.

In section 3 a k-means cluster algorithm was executed to cluster the neighborhoods in the boroughs in 5 different clusters. The Bronx is not uniform and shows a lot of different clusters in one area. Just like the Bronx, Queens is not uniform. Manhattan and Brooklyn have 2 bigger clusters, especially in Manhattan it is visible. Staten Island has a less high density and no bigger cluster.

4. Discussion

The Hispanic restaurant distribution is very interesting. For a client, this information seems to be a good starting point for opening up a restaurant. There are fewer Hispanic restaurants in the Bronx or in Manhattan, despite the fact that there are more Hispanics than in other boroughs. So it could be a good idea to open up a Spanish/Mexican restaurant in Manhattan, as Manhattan only got 43 Hispanic restaurants. Also in the Bronx with the most Hispanics opening up a Hispanic restaurant could be profitable.

5. Conclusion

In this report the 5 boroughs of New York City have been analyzed in terms of restaurant and demographic distribution. This analysis will help clients, who want to open up an Asian or Hispanic restaurant, to easily decide which place is the best fit. An interesting finding was found when analyzing the demographic distribution of Hispanics and the restaurant distribution. Further analysis can be drawn to a finer analysis of the neighborhoods of the boroughs to determine the exact neighborhoods to open up a profitable restaurant.

Bibliography

[1]
http://demographia.com/db-worldua.pdf
[2]
http://www.ameredia.com/resources/demographics/asian_american.html
[3]
https://www.statista.com/statistics/259776/number-of-people-who-went-to-restaurants-in-new-york-by-type/
[4]
https://www.asanet.org/news-events/footnotes/jun-jul-aug-2019/features/latino-population-new-york-city
[5]
https://python-visualization.github.io/folium/
[6]
https://www.kaggle.com/muonneutrino/new-york-city-census-data
[7]
https://developer.foursquare.com/