Using Location Data and K-Means Clustering to determine the locations for a new Restaurant Chain and its Distribution Centers in New York City

Introductory Section

There are already a multitude of restaurants, bars and lifestyle locations in New York City. Fierce competition can therefore be expected when opening new restaurants**.** To ensure business success it is crucial to determine the right location for the new venue taking into account the characteristics of the neighborhood, population and restaurant variety. The search for white spots where the expected demand is high, but the restaurant density that corresponds to the demand is still rather low increases the chances of a successful start.

This report examines how census and location data can be used to determine the locations of a new Chinese restaurant chain that wants to gain a foothold in the New York restaurant scene. These are the main questions that will be answered:

- **Which districts and neighborhoods are the most promising to open new Chinese restaurants?**
- Having determined the new restaurant locations, the next step is to deal with logistic questions. A well-organized supply chain will not only ensure smooth supply of the restaurants but also reduce costs. Therefore, the question will be answered: **Which are the best places to set-up the distribution centers for supplying the restaurant chain?**

Data

This section describes the data that is used for answering the business problem. To determine the best locations for opening new Chinese restaurant, the following data has been collected:

Census Data is used to examine the Chinese population density per New York District. The Official Website of the City of New York offers access to the decennial census data, which is collected once a decade. This data includes the Asian population and selected subgroups per community district. For the analysis the data from 2000 is used, as the decennial census data from 2010 has not been published yet. [1]

In total there are 55 New York community districts in the data set. The following table shows an example of the census data for selected Brooklyn districts:

	Community District	Chinese Inhabitants
20	Brooklyn 9	297
21	Brooklyn 10	12333
22	Brooklyn 11	34164
23	Brooklyn 12	16266
24	Brooklyn 13	5335

A community district consists of several neighborhoods. Determining a restaurant location only based on community districts would be rather vague. Therefore, also location data about the

neighborhoods in New York is collected [2]. This data contains the neighborhood name, the respective borough where the neighborhood is located and its coordinates:

	Borough	Neighborhood	Latitude	Longitude
0	Bronx	Wakefield	40.894705	-73.847201
1	Bronx	Co-op	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

To map these two tables, a mapping list of neighborhoods by community district can be retrieved from Wikipedia.[3]

	Community District	Neighborhoods
0	Bronx CB 1	Melrose, Mott Haven, Port Morris
1	Bronx CB 2	Hunts Point, Longwood
2	Bronx CB 3	Claremont, Concourse Village, Crotona Park, Mo...
3	Bronx CB 4	Concourse, Highbridge
4	Bronx CB 5	Fordham, Morris Heights, Mount Hope, Universit...

Foursquare location data is used to retrieve all existing Chinese restaurants in New York. This data can be accessed using the public Foursquare API. To limit the result set, a 'CategoryId' is passed with the API request to limit the results to only Chinese restaurants. As the result set of one foursquare request is limited to 50 results, the request is repeated for each neighborhood. This procedure leads to duplicates in the result set; therefore, the duplicates are removed during the data cleaning process.

In total, there are 2205 Chinese restaurants in New York. The data includes name of the restaurant, latitude, longitude and the category:

	Restaurant Name	Latitude	Longitude	Venue Category
0	Peking Kitchen	40.854260	-73.866223	Chinese Restaurant
1	No. 1 Chinese Restaurant	40.895781	-73.805285	Chinese Restaurant
2	Mr. Q's Chinese Restaurant	40.855790	-73.855455	Chinese Restaurant
3	China Mia	40.858316	-73.867232	Chinese Restaurant
4	Jimmy's Best Chinese Restaurant	40.884179	-73.832685	Asian Restaurant

Methodology

To get insights in the gathered data, the first step is using exploratory data analysis and data visualization. The locations of existing Chinese restaurants in New York are displayed on maps using the Folium Python library. K-nearest neighbor algorithm is used to determine the neighborhood of each Chinese restaurant. Then, the restaurants are clustered by neighborhood and community district to investigate the density of restaurants.

Subsequently, the restaurant density is compared to the Chinese residents living in the same community district. To estimate the likelihood of success for a new restaurant the following ratio is applied: **Number of Chinese restaurants per 1,000 Chinese inhabitants.** A low ratio refers to a promising district where supply is rather low compared to the demand, whereas a high ratio indicates a saturated restaurant – population relation.

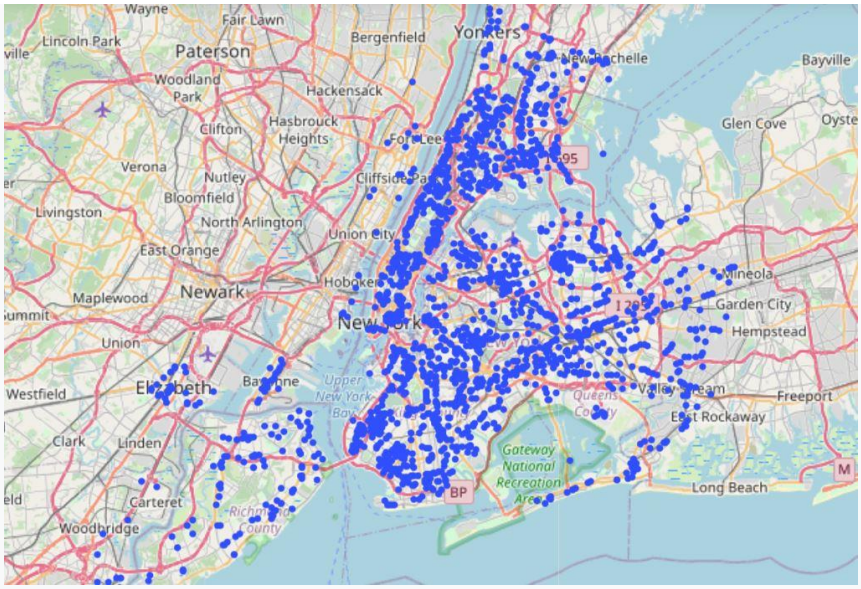
In total, the new restaurant chain wants to open 15 new Chinese restaurants in New York. Therefore, the 15 community districts with the lowest ratio of Chinese restaurants / 1,000 Chinese inhabitants are selected from the dataset. Within the selected community district, the neighborhood with the fewest Chinese restaurants is assigned as the new location for a Chinese restaurant.

Having determined the most promising locations to open the restaurants, the next step is to find the best locations to set-up the distribution centers for supplying the restaurant chain. This is done by applying K-means clustering. In total, 3 distribution centers shall be established. The coordinates of the selected restaurant locations are used to cluster the restaurants and define the best locations for the new distribution centers.

Exploratory Data Analysis

To get insights in the gathered data, the first step is using exploratory data analysis and data visualization. By using Folium Map we can see that the Chinese restaurants are almost equally spread all over New Yorks districts. As this does not allow drawing conclusions about promising restaurant locations, we will cluster the restaurants by neighborhoods and community district in the next step.

Figure 1: Chinese restaurants in New York



Using k-nearest neighbor algorithm each Chinese restaurant is assigned to the closest neighborhood. Then, the neighborhoods are grouped by community district and the results displayed in a map. Now we can clearly identify community districts, which have a higher density of Chinese restaurants. The community district with the highest density is Queens 7 where 136 Chinese restaurants are located.

Figure 2: Chinese restaurants in New York grouped by community district

To get insights about the Chinese population in these districts, the census data is also plotted on the map:

Figure 3: Chinese restaurants and Chinese population per community district

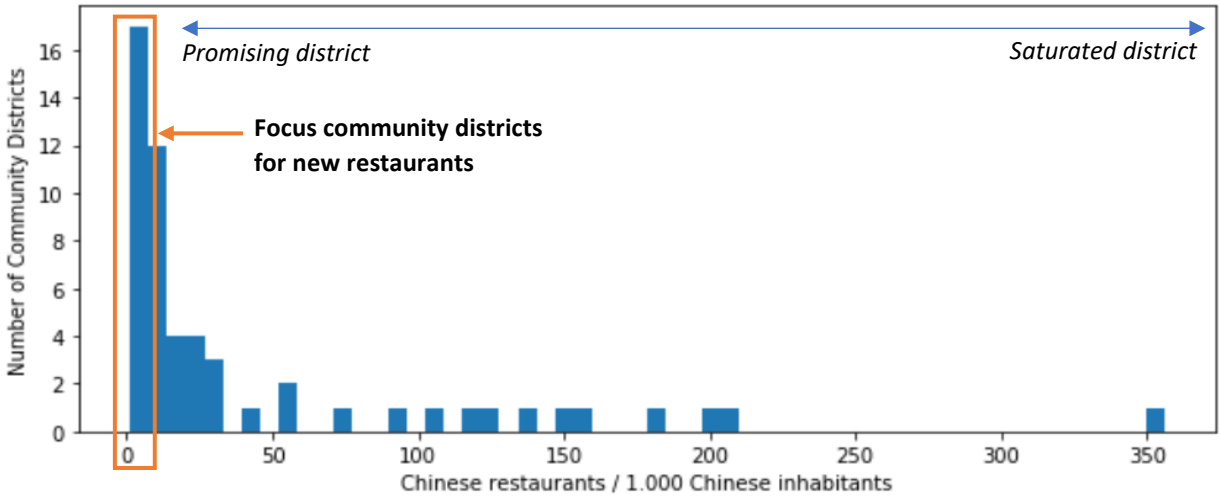
The following insights can be obtained by comparing the restaurant density and the population data:

1. There is a positive correlation between the Chinese population and restaurant density which is the natural result of demand and supply. To quantify the strength of this linear relationship, the Pearson Correlation Coefficient is used. The Pearson Correlation Coefficient between population and restaurants is 0.587 indicating a moderate positive correlation.
2. There are districts with a high density of Chinese restaurants, but comparably few Chinese residents. These districts can be considered as less promising for the establishment of new restaurants, as there is already a high supply for the existing demand in this area.

3. Thus, the focus of the new restaurant chain should be on districts with a small number of restaurants compared to the respective Chinese population. These districts may be seen as white spots where supply in the area has not yet fully met existing demand.

Therefore, for each community district the ratio between the Chinese restaurants per 1,000 Chinese inhabitants is calculated. The range of Chinese restaurants / 1,000 Chinese inhabitants ranges from 1 up to 350. However, the districts with ratios above 50 can be considered as outliers as most districts range between 1 and 30. The selection of locations for new Chinese restaurants should be therefore focused on districts with a ratio below 10.

Figure 4: Histogram of the distribution of restaurant - population ratio



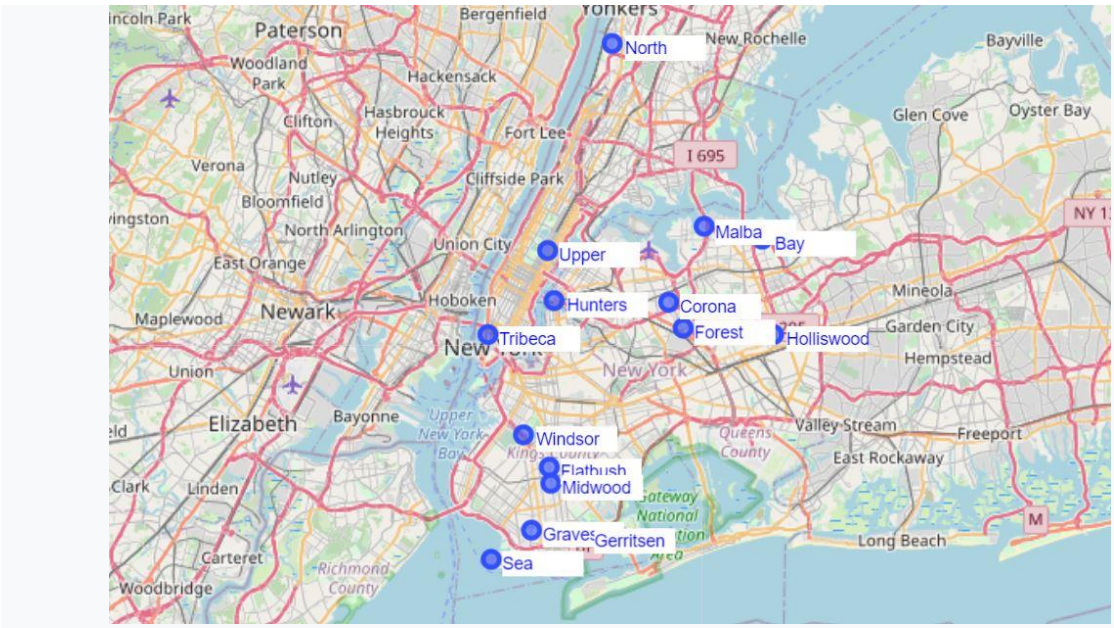
Results

The 15 community districts with the lowest ratio of Chinese restaurants / 1,000 Chinese inhabitants can be retrieved from the dataset:

	Community District	Number Chinese Restaurants	Chinese	Restaurants / 1.000 Chinese
0	Brooklyn 12	21	16266	1.291037
1	Queens 4	33	21714	1.519757
2	Brooklyn 13	9	5335	1.686973
3	Queens 6	25	12374	2.020365
4	Brooklyn 14	11	4235	2.597403
5	Manhattan 1	11	4040	2.722772
6	Brooklyn 15	44	15390	2.858999
7	Queens 2	22	7600	2.894737
8	Queens 7	136	41777	3.255380
9	Brooklyn 11	113	34164	3.307575
10	Queens 8	45	13192	3.411158
11	Brooklyn 7	65	17362	3.743808
12	Brooklyn 10	48	12333	3.891997
13	Queens 3	37	8260	4.479419
14	Manhattan 7	22	3846	5.720229

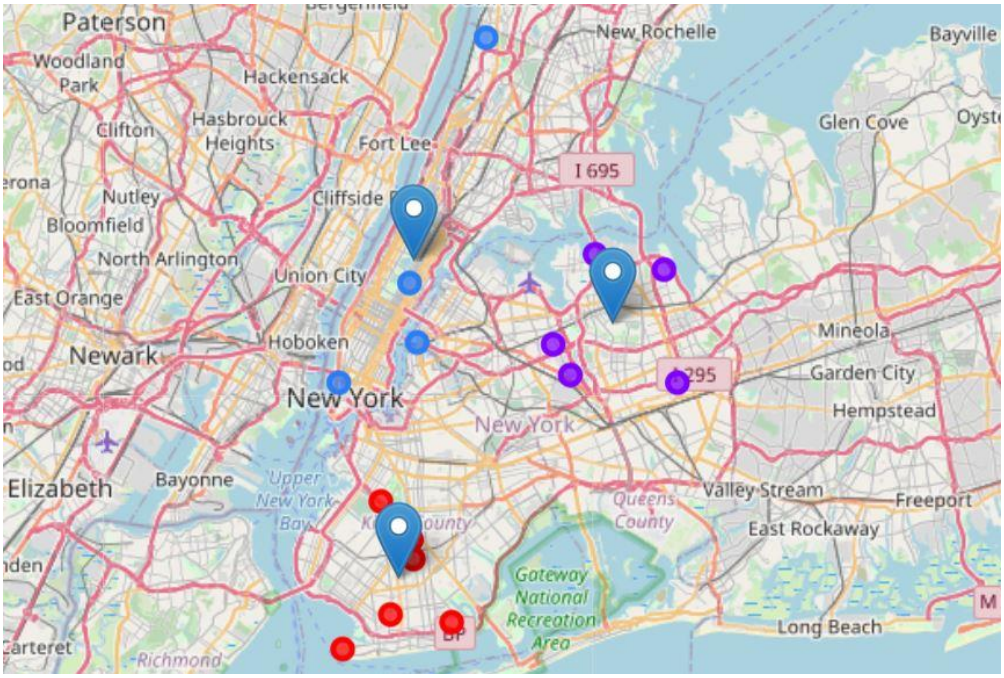
Within the community districts, the neighborhood with the fewest Chinese restaurants is selected as the new location. This is based on the assumption that competition in this neighborhood is relatively low which increases the chance of a successful launch. Finally, the selected neighborhoods for the new restaurant locations are plotted on the New York Map.

Figure 5: Selected neighborhoods for the new Chinese restaurants



To ensure supply of the selected restaurant locations, three distribution centers will be set-up. Using K-Means Clustering the best locations for these new centers are determined. The 15 restaurants are clustered in 3 groups and the center of each cluster represents the location of the distribution center. The distribution centers not only reduce delivery times and ensure reliable supply but also optimize costs for the restaurant chain.

Figure 6: Selected locations for the distribution centers of the Chinese restaurant chain



Discussion

In the methodology and result section it was shown how to use location data and enrich them by combining other datasets (e.g. census data). As indicator for selecting the new restaurant locations the ratio between Chinese restaurant density and the Chinese population was used. This ratio is a good starting point to select the most promising districts, but also has some limitations. For reason of simplification, it was assumed that demand for Chinese restaurants only comes from the Chinese residents living in that district. Tourists or the demand of other population groups living in that area were not considered. In a more detailed analysis, the factors that increase the demand for Chinese restaurants should therefore be diversified. This also includes the overall density of restaurants in the districts, the proximity to the adjacent Chinese restaurants as well the proximity of public transport or parking options.

Based on the selected restaurant locations, the distribution center locations were determined using K-Means Clustering, which is one the simplest and most popular unsupervised machine learning algorithms. The algorithm was used to divide the selected restaurant locations in 3 distinct non-overlapping subgroups which are called clusters. A cluster refers to a collection of data points aggregated together because of certain similarities. In this case, the location data of the selected restaurants was used as similarity to minimize the distance within the clusters. Using location data for determining the distribution centers helps determining the approximate area to set up the distribution center. As a next step, it is recommended to take further factors into consideration. These include the proximity to high-ways and big distribution points and the average traffic volume in that area. Other factors as rental costs or availability of qualified workforce might also be important.

Conclusion

This report showed how to leverage location data to determine new restaurant locations in a city. By enriching these data with additional data, the gain in knowledge can be significantly increased. In order to obtain first insights into the available data, an explorative data analysis was conducted. Machine learning algorithms such as K-nearest neighbor and K-means clustering were deployed to cluster the data and find the distribution center locations for the new restaurant chain. The limitations of the analysis were also outlined in the discussion section. It is recommended to overcome these limitations by including additional factors in the further analysis.

References

[1] https://www1.nyc.gov/site/planning/data-maps/nyc-population/demo-tables-2000.page.

[2] https://cocl.us/new_york_dataset

[3] https://en.wikipedia.org/wiki/Neighborhoods_in_New_York_City