Data Science Capstone – IBM Data Science Specialization

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

Tatjana Kaiser

June 5th, 2020

Content

1 Business Problem

2 Data Selection

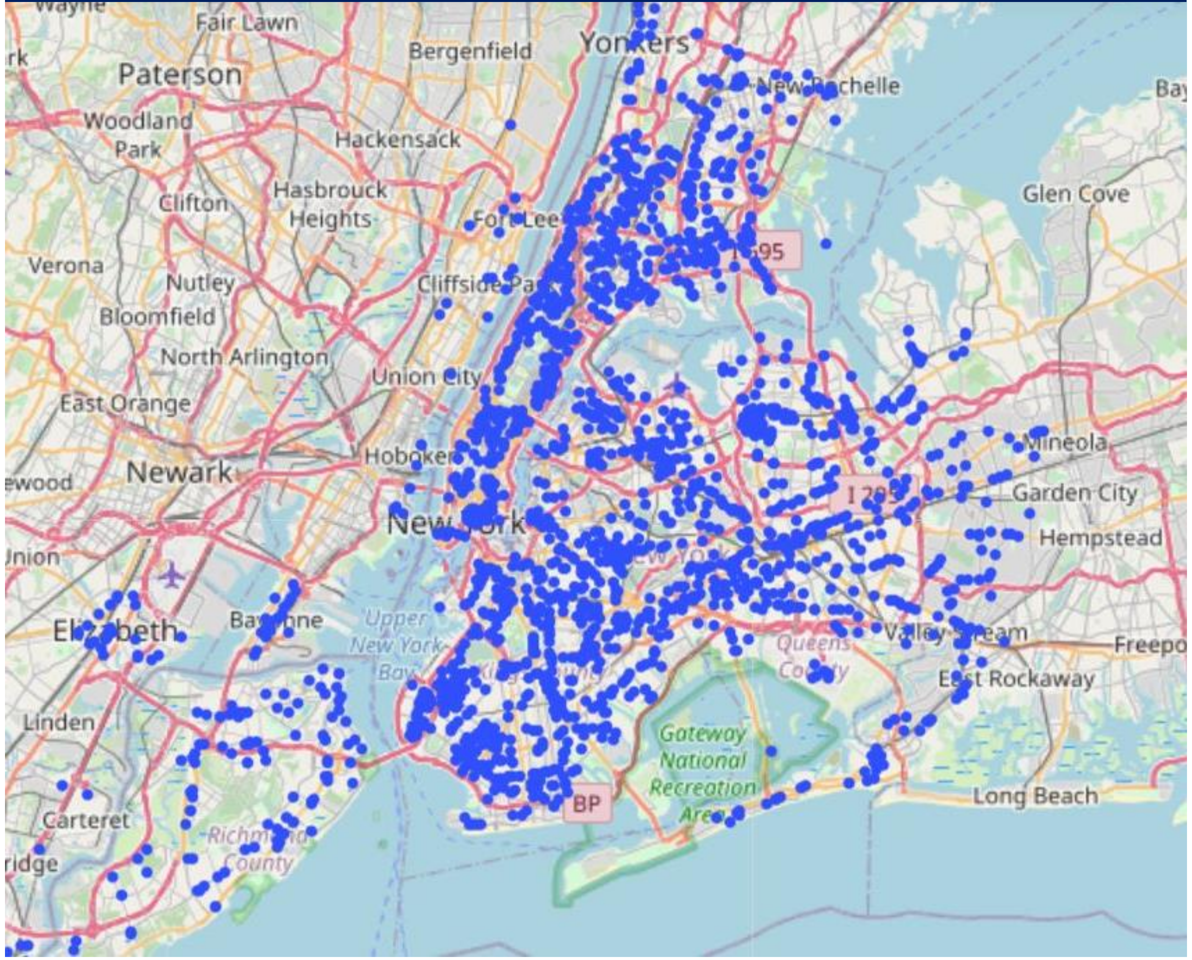
3 Exploratory Data Analysis

4 Results

5 Discussion

Business Problem

Chinese restaurants in New York



Description of business problem

A **Chinese restaurant chain** wants to gain a foothold in the New York restaurant scene opening **15 new restaurants** and **3 distribution centers** for these restaurants.

As there are already a multitude of restaurants in New York, fierce competition can be expected. Determining the right location is crucial for ensuring business success.

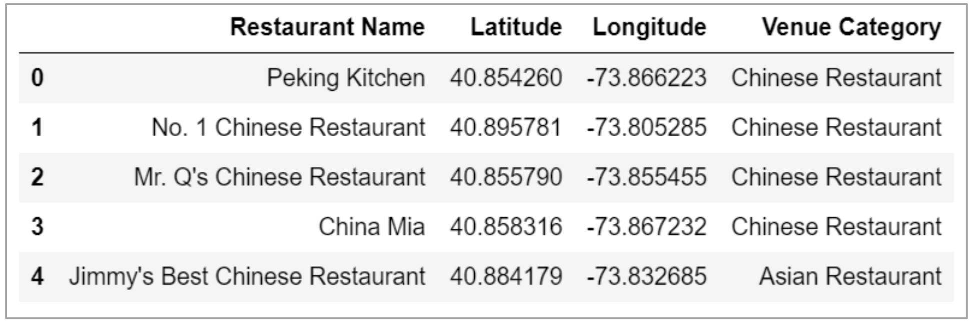
Main questions:

- **Which districts and neighborhoods are the most promising to open new Chinese restaurants?**
- **Which are the best places to set-up the distribution centers for supplying the restaurant chain?**

Data Selection

Foursquare Location Data

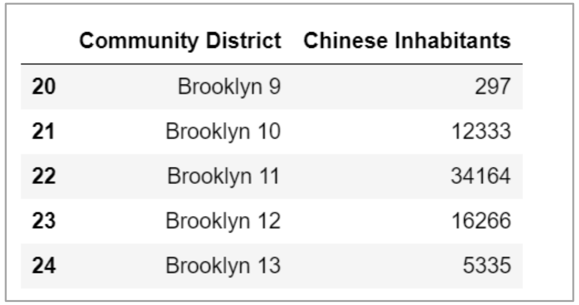
- Public Foursquare API is used
- Result table contains all Chinese restaurants in New York:
  - ❖ Restaurant name
  - ❖ Coordinates (longitude, latitude)
  - ❖ Restaurant category



	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

Census Data

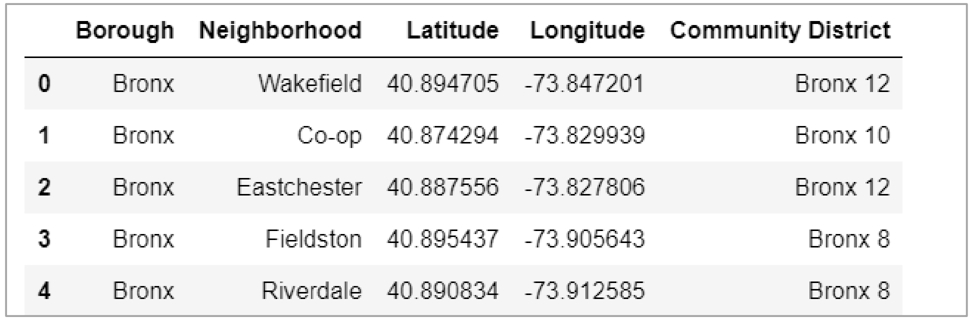
- Decennial census data (2000) from the official website of Ney York
- Data includes Asian population and selected subgroups per community district



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

Neighborhoods & Community Districts

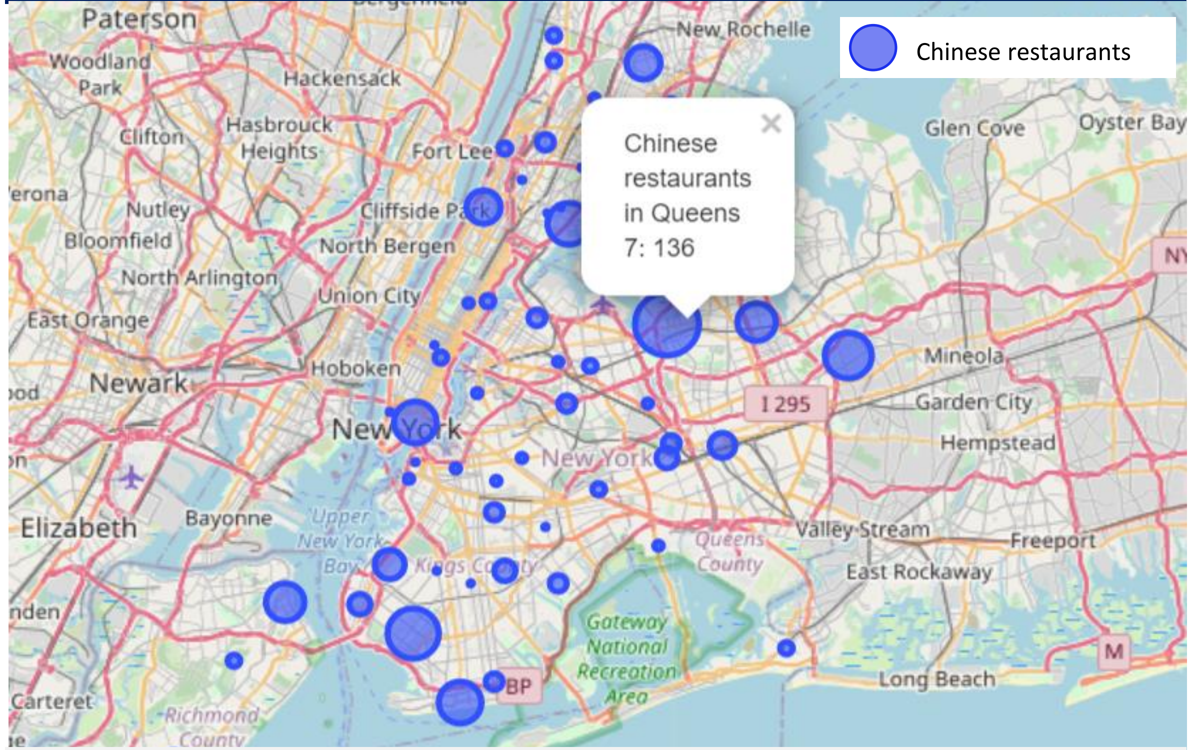
- Mapping table to map the Chinese restaurants with the census data, including:
  - ❖ Neighborhood incl. coordinates
  - ❖ Community district of the neighborhood



	Borough	Neighborhood	Latitude	Longitude	Community District
0	Bronx	Wakefield	40.894705	-73.847201	Bronx 12
1	Bronx	Co-op	40.874294	-73.829939	Bronx 10
2	Bronx	Eastchester	40.887556	-73.827806	Bronx 12
3	Bronx	Fieldston	40.895437	-73.905643	Bronx 8
4	Bronx	Riverdale	40.890834	-73.912585	Bronx 8

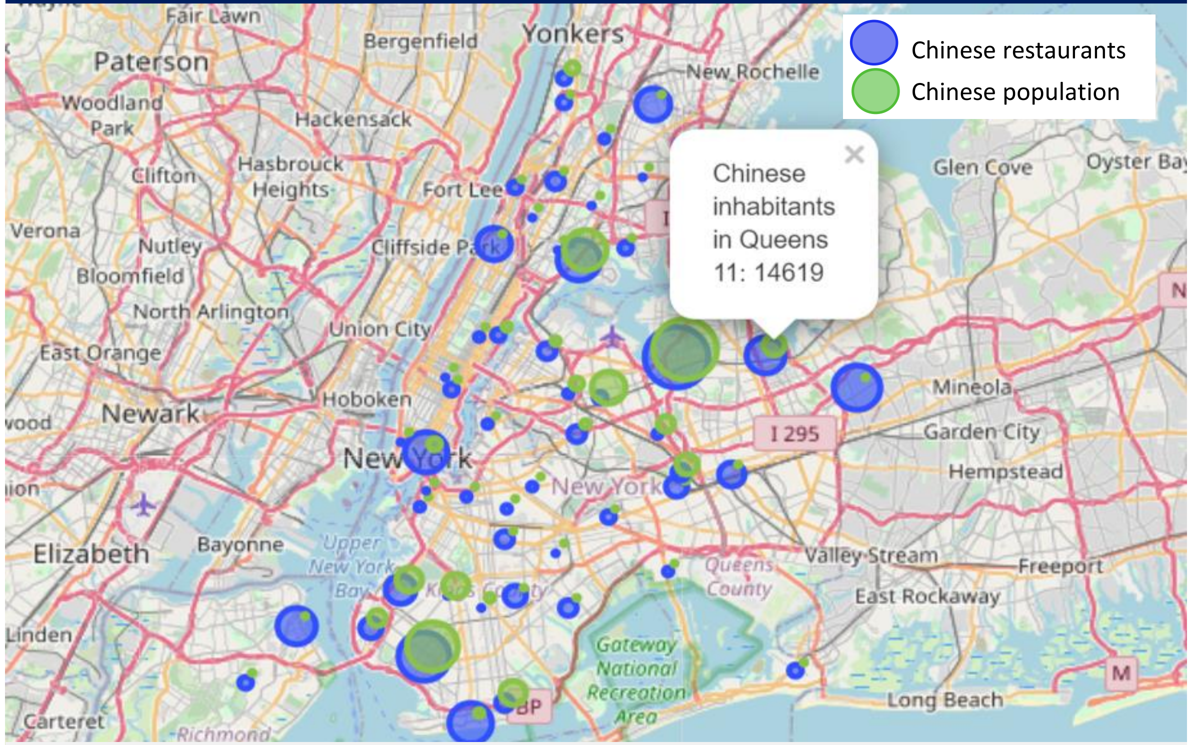
Exploratory Data Analysis

Chinese restaurants clustered by community district



- Using K-nearest neighbor the Chinese restaurants are assigned to the closest neighborhood
- Next, they are clustered by neighborhood and community district

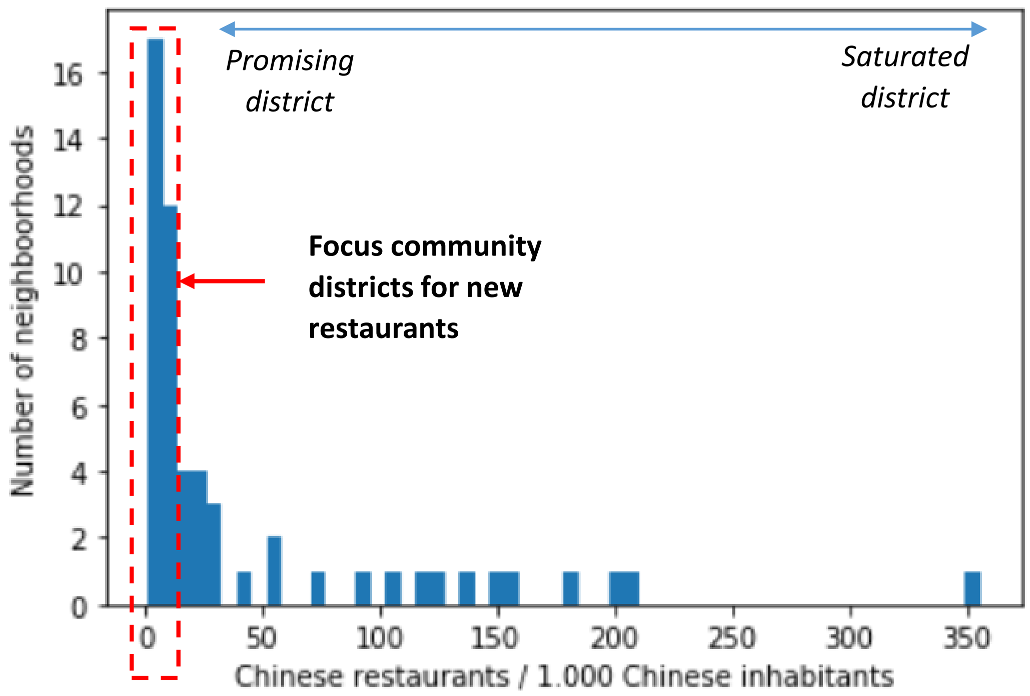
Chinese population per community district



- Chinese restaurants and Chinese population are positively correlated (Pearson Correlation Coefficient: 0.587)
- To determine promising locations, the following ratio is used:
  Chinese restaurants per 1,000 Chinese residents

Results

Distribution of Chinese restaurant - population ratio



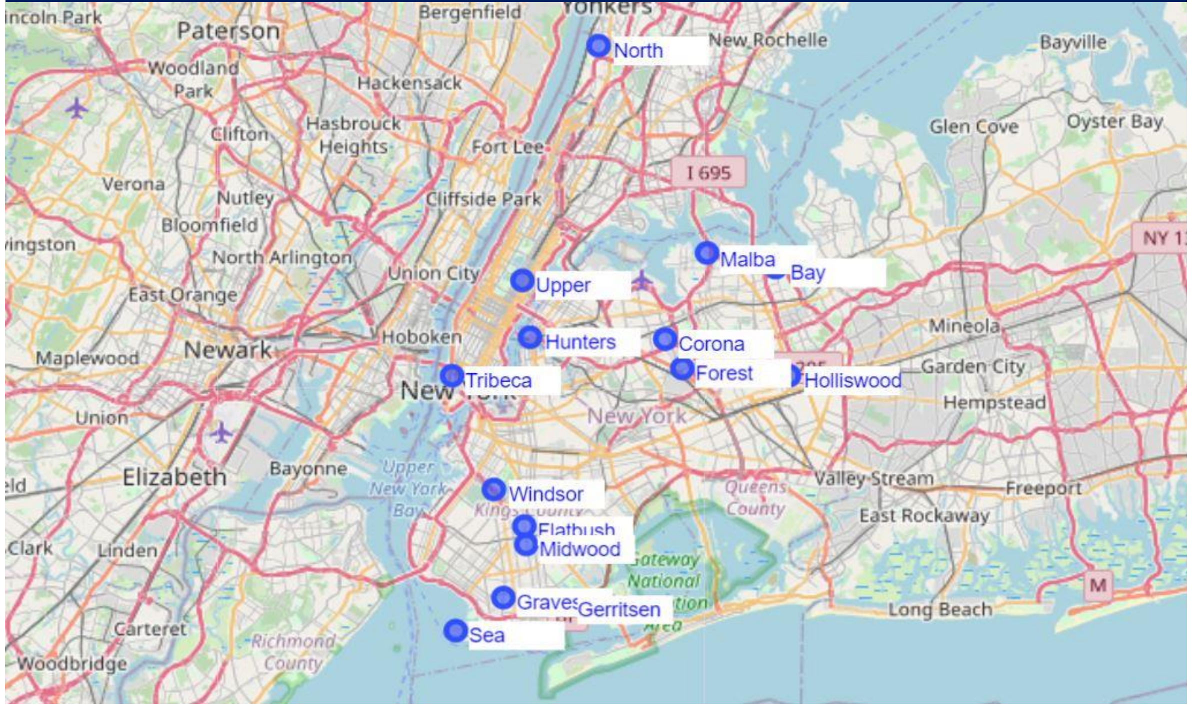
Promising district

Saturated district

Focus community districts for new restaurants

- Most districts have 1-30 restaurants per 1,000 Chinese residents
- A low ratio refers to a promising district where supply is rather low compared to the demand for Chinese restaurants

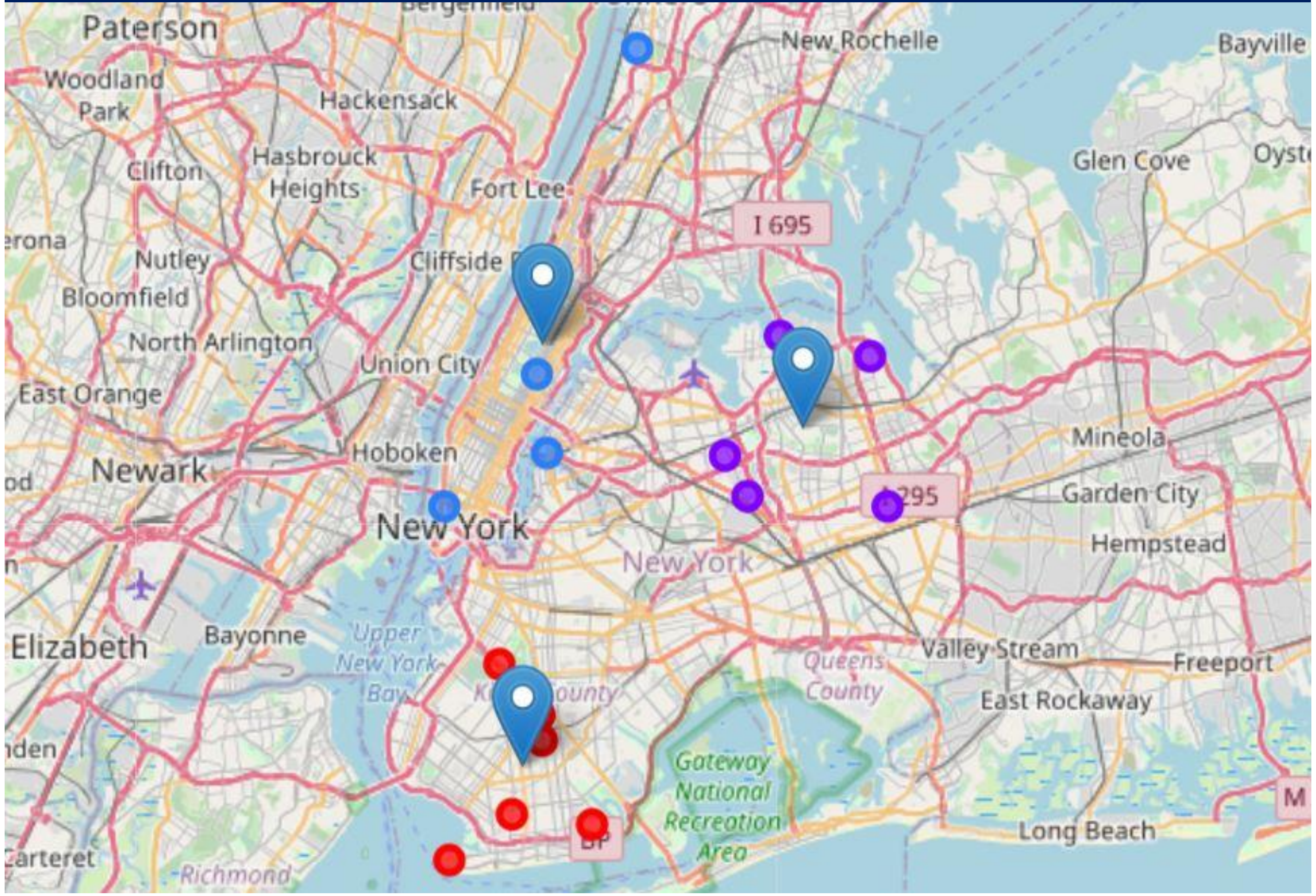
Locations for the new Chinese restaurants



- The 15 community districts with the lowest Chinese restaurant – population ratio are selected
- Within the districts, the neighborhood with the fewest Chinese restaurants is selected as new restaurant location

Results

Locations of the distribution centers



Using K-Means clustering

- Using **K-Means clustering** the best locations for the 3 new distribution centers are determined.
- The **15 new restaurant locations** are clustered in **3 groups**
- 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.

Discussion

Discussion & Recommendation for further investigation

- For reason of simplification, the analysis is based on the assumption that demand for Chinese restaurants only comes from the Chinese residents living in that area.
- In a more detailed analysis, the following factors should also be considered:
  - ❖ Number of tourists in that district
  - ❖ Demand of other population groups living in that district
  - ❖ Overall density of restaurants
  - ❖ Proximity to the adjacent Chinese restaurants
  - ❖ Proximity of public transport or parking options

- Using K-Means Clustering the selected restaurant locations where divided into 3 distinct clusters based on their location data. The centers of the clusters were selected as the distribution center locations.
- As a next step, the following factors should be also considered:
  - ❖ Proximity to high-ways and big distribution points
  - ❖ Average traffic volume in that area
  - ❖ Other factors as rental costs or availability of qualified workforce

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