# Neighborhood recommend for a new Chinese restaurant

XU Shuai

2020/7/29

## INTRODUCTION

## Background

New York City is one of the largest cities all around the world, there are millions of people lives here so the demand for the restaurant is always heavy. Among them, a Chinese restaurant is a popular branch and it will be a good chance to set up a new restaurant. To reach this goal, there is one important step to choose an address for the restaurant. The good place could have a great number of customers but a few competitors, the most important, it must be a safe place.

## Problem

To find the best address for a new restaurant, there must be a heavy demand of the restaurant which means the people lives there prefer to have a meal in the restaurant, then, there couldn't have a lot of similar restaurants to avoid the competition. At the same time, for the security of the owner and the customer, we want to find a place that has a low level of the crime.

To simplify the question, we search a neighborhood who has:

* A high level of **the restaurant;**
* A low level of **the Chinese restaurant;**
* A low level of the **crime.**

## Interest

This project is to find a neighborhood where is good for a new chines restaurant so it will be interesting for the person who wants to have a restaurant for himself in New York, with a few modifications, this project could meet the research for the other kind of restaurants.

## DATA

## data source

For our research, we need two kinds of data, the data of venues and of crimes, to get them, we could use the API of FourSquare with the coordinates of the neighborhoods from the <https://geo.nyu.edu/catalog/nyu_2451_34572> and NYPD Complaint Data Current who includes all valid felony, misdemeanor, and violation crimes reported to the New York City Police Department (NYPD) for all complete quarters so far this year (2020) <https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Current-Year-To-Date-/5uac-w243>

## data choosing

* + 1. restaurant data

with the site, we could get a JSON file and we could get the coordinates (latitude, longitude) for each neighborhood, then, with the coordinate, we can get a list of the venues with all kinds. Then, we could create the dataframe with the columns:

* **Neighborhood**

(as several neighborhoods have the same name in a different borough, I add the name of the borough in this column)

* **Latitude**
* **Longitude**
* **venues** 
  + 1. crime data

from the site, we could get la csv file with 196624 rows and 23 columns which is the report of all the complaints in New York in the year 2020 the columns are shown below:

|  |  |
| --- | --- |
| CMPLNT\_NUM | Randomly generated persistent ID for each complaint |
| CMPLNT\_FR\_DT | Exact date of occurrence for the reported event (or starting date of occurrence, if CMPLNT\_TO\_DT exists) |
| CMPLNT\_FR\_TM | Exact time of occurrence for the reported event (or starting time of occurrence, if CMPLNT\_TO\_TM exists) |
| CMPLNT\_TO\_DT | Ending date of occurrence for the reported event, if exact time of occurrence is unknown |
| CMPLNT\_TO\_TM | Ending time of occurrence for the reported event, if exact time of occurrence is unknown |
| RPT\_DT | Date event was reported to police |
| KY\_CD | Three digit offense classification code |
| OFNS\_DESC | Description of offense corresponding with key code |
| PD\_CD | Three digit internal classification code (more granular than Key Code) |
| PD\_DESC | Description of internal classification corresponding with PD code (more granular than Offense Description) |
| CRM\_ATPT\_CPTD\_CD | Indicator of whether crime was successfully completed or attempted, but failed or was interrupted prematurely |
| LAW\_CAT\_CD | Level of offense: felony, misdemeanor, violation |
| JURIS\_DESC | Jurisdiction responsible for incident. Either internal, like Police, Transit, and Housing; or external, like Correction, Port Authority, etc. |
| BORO\_NM | The name of the borough in which the incident occurred |
| ADDR\_PCT\_CD | The precinct in which the incident occurred |
| LOC\_OF\_OCCUR\_DESC | Specific location of occurrence in or around the premises; inside, opposite of, front of, rear of |
| PREM\_TYP\_DESC | Specific description of premises; grocery store, residence, street, etc. |
| PARKS\_NM | Name of NYC park, playground or greenspace of occurrence, if applicable (state parks are not included) |
| HADEVELOPT | Name of NYCHA housing development of occurrence, if applicable |
| X\_COORD\_CD | X-coordinate for New York State Plane Coordinate System, Long Island Zone, NAD 83, units feet (FIPS 3104) |
| Y\_COORD\_CD | Y-coordinate for New York State Plane Coordinate System, Long Island Zone, NAD 83, units feet (FIPS 3104) |
| Latitude | Latitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326) |
| Longitude | Longitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326) |

we don’t need all the information in this table such as the number of the complaint or the exact time of the call, we need only

* **LAW\_CAT\_CD:**
* **Latitude**
* **Longitude**

## Methodology

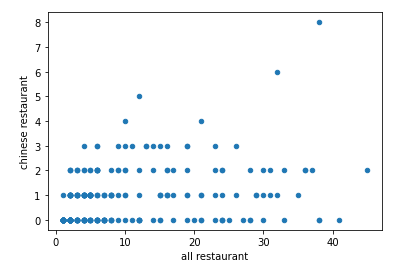
To get the information from the data, we need also do some operation and calculation

* 1. **Choose the neighborhood base on the number of the restaurant**

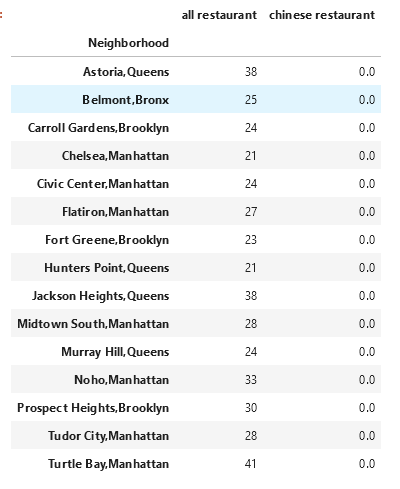
in the list venues, there are lots of the different kinds, among them, we choose the values who contain the word ‘restaurant’ as the data of ‘all restaurant’, then we count them to get a number of the restaurant around each neighborhood.

After that, we take only the venues with the catalog ‘Chinese restaurant’, and we can find if a neighborhood who has a Chinese restaurant nearby.

With these 2 columns, we can create a figure :



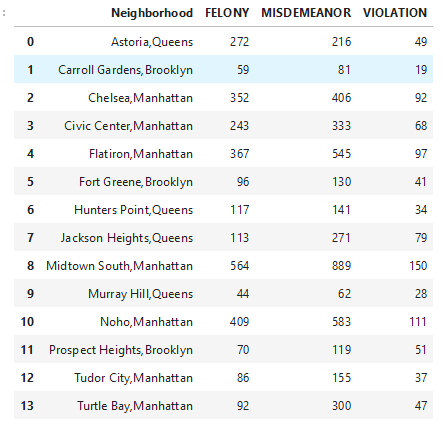
In this case, I'd like to choose the neighborhoods with the number of restaurants larger than 20 and there's no Chinese restaurant so we could have a list below:



* 1. **Count the number of crimes in a certain distance**

With the library, we can easily find the distance between two points with the coordinates, so, with the coordinates of the neighborhoods and the crimes, we could find the distances between them.

In this case, I choose a distance maximum as 500m between crimes and neighborhoods, with the level of offense in the dataset, we could count the number of the crimes near each neighborhood as the table::



* 1. **Merge the data and cluster**

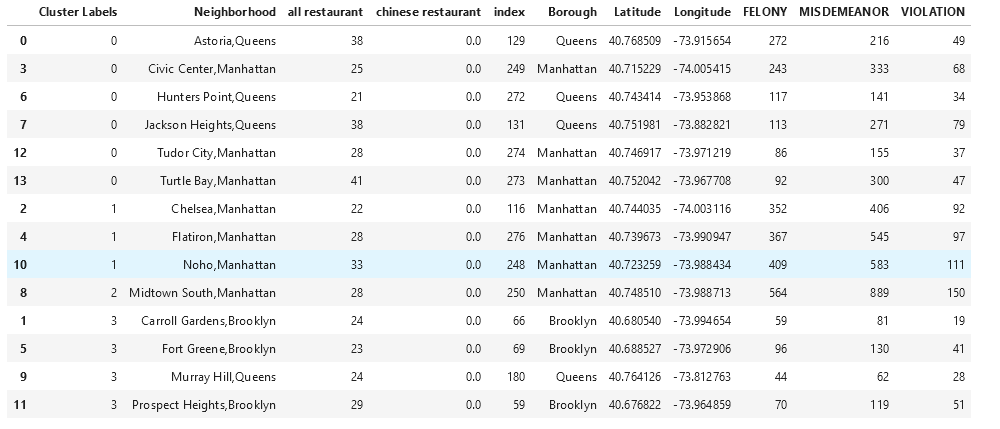
Then, we merge the number of restaurants and the crimes, and we could also drop the columns unnecessary to get a table with only the columns

* All restaurant
* Chinese restaurant
* Felony
* Misdemeanor
* Violation

After that, we could do the machine learning with the clustering algorithm KMeans to find a group of neighborhoods we want, in this case, I choose a k number as 4 so we will get 4 groups.

## Result

After the clustering, we could get the table:



To show the result and obvious their positions, we show these group with the different colors of the points in the map

Cluster 0: 6 red points

Cluster 1: 3 purple points

Cluster 2: 1 blue point

Cluster 3: 4 green points

|  |  |
| --- | --- |
|  |  |
|  |  |

## Discussion

For each neighborhood, there are 4 characters (we have chosen that the number of the Chinese restaurant is 0), we could have a rule as below

Cluster 0: high-level restaurant low crime

Cluster 1: low-level restaurant high serious crime

Cluster 2: low-level restaurant very high serious crime

Cluster 3: low-level restaurant low crime

Look at the locations of them, we could also find that the group 0 is a bit far from the center of the city, so there will be more residents and they have a heavy requirement of the restaurant, as there isn't any Chinese restaurant nearby, it will be a good choice to create one. Cluster 1 and 2, are located at the Manhattan island, there are a lot of visitors and it means that rent will be high, at the same time, there will be a lot of people every day so it will be a bit chaotic there. For the rest neighborhoods, they are further from the city center so they are more peaceful, but on the contrary, the people's lives there won't eat out a lot so it won't be a good place there.

## Conclusion

As the information we have got and analyzed before, we could have a conclusion that the cluster 0 will be good to create a new Chinese restaurant, there are heavy require of restaurants and low competition of the other Chinese restaurant, the rate of crime is also acceptable to compare with the other neighborhoods. They are:

* **Astoria, Queens**
* **Civic Center, Manhattan**
* **Hunters Point, Queens**
* **Jackson Heights, Queens**
* **Tudor City, Manhattan**
* **Turtle Bay, Manhattan**

Then, cluster 3 is also a choice, they are more peaceful but the requirement is low. We could think of them as an alternative, they are:

* **Chelsea, Manhattan**
* **Flatiron, Manhattan**
* **Noho, Manhattan**

The next step is to do the field visit for the further information, the objective of this project is to do the pre-search, we could find the first list for the choice in all the neighborhoods of New York City, we could save a lot of time by this research.

## To improve efficiency

In the future, we could approve the work in this project by the parts below:

1. **The list of the restaurant:** because there are some restaurant don't have this name in their catalogs such as bar or coffee shop, the list of the restaurant couldn't have all the places who sales the food, in the future, we could review the list by ourselves to classify;
2. **The research of the crimes:** in this case, for each coordinate of the neighborhood, we need to traverse all the list and to find the distance, it will take a long time of waste, we could find another way to solve this problem in the future;
3. **The algorithm of the clustering:** I have chosen the model KMeans in this project and the coefficient k as 4, we could improve the project by finding another model or coefficient, maybe there will be a result more clearly.