Capstone Project – The Battle of Neighborhoods

**Introduction Section**

This final project explores the best locations for Chinese restaurants throughout the Queens of New York. New york is a major metropolitan area with more than 8.4 million (Quick Facts, 2018) people living within city limits. New York City is the largest city in the united states with a long history of international immigration. The New York metropolitan area is home to the largest and most prominent ethnic Chinese population outside of Asia, hosting Chinese populations representing all 34 provincial-level administrative units of China and constituting the largest metropolitan Asian American group in the United States as well as the largest Asian-national metropolitan diaspora in the Western Hemisphere. The Chinese American population of the New York City metropolitan area was an estimated 893,697 as of 2017. New York City itself contains by far the highest ethnic Chinese population of any individual city outside Asia, estimated at 628,763 as of 2017.

**Target Audience**

* Business personnel who wants to invest or open a restaurant in New York
* The freelancer who loves to have their own restaurant as a side business
* Finding the best location for opening a restaurant
* Budding data Scientists, who want to implement some of the most used

**Data Section**

For this project, we need the following data:

1. New York city data that contains Borough, Neighborhoods alongwith there latitude and longitude

* Data Souce: <https://cocl.us/new_york_dataset>
* Description: This data set contains the required information. We can use this data set to explore various neighborhoods of new York city.

1. Chinese restaurants in Queens neighborhood of New York City

* Data Source: Foursquare API
* Description: by using this API we can get all the venues in the Queens neighborhood. We can filter these venues to get only Chinese restaurants

**Approach**

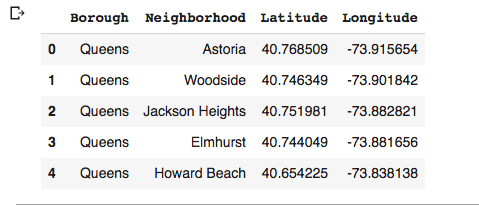
1. collect New York city data from <https://cocl.us/new_york_dataset>
2. Using Foursquare API, we will get all venues for each neighborhood
3. Filter out all venues which are Chinese restuarants
4. Data visualization and statistical analysis
5. Analyzing using KMeans clustering
6. compare the Neighborhoods to find the best place for starting up a restautant
7. Inference from these results and related conclusions

**Problem Statement**

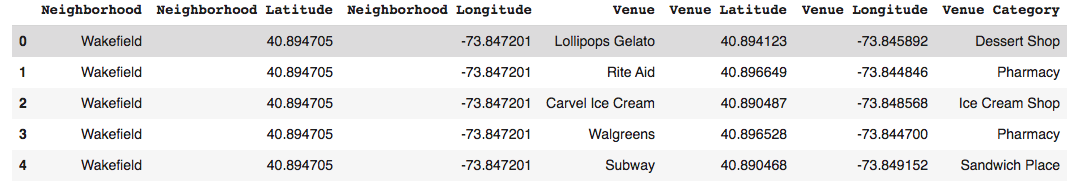
1. what is the best location for a Chinese restaurant in Queens, New York City?
2. In what Neighborhood should I open a Chinese restaurant to have the best chance of being successful?

**Analyzation**

1. Get Queens Borough geological data



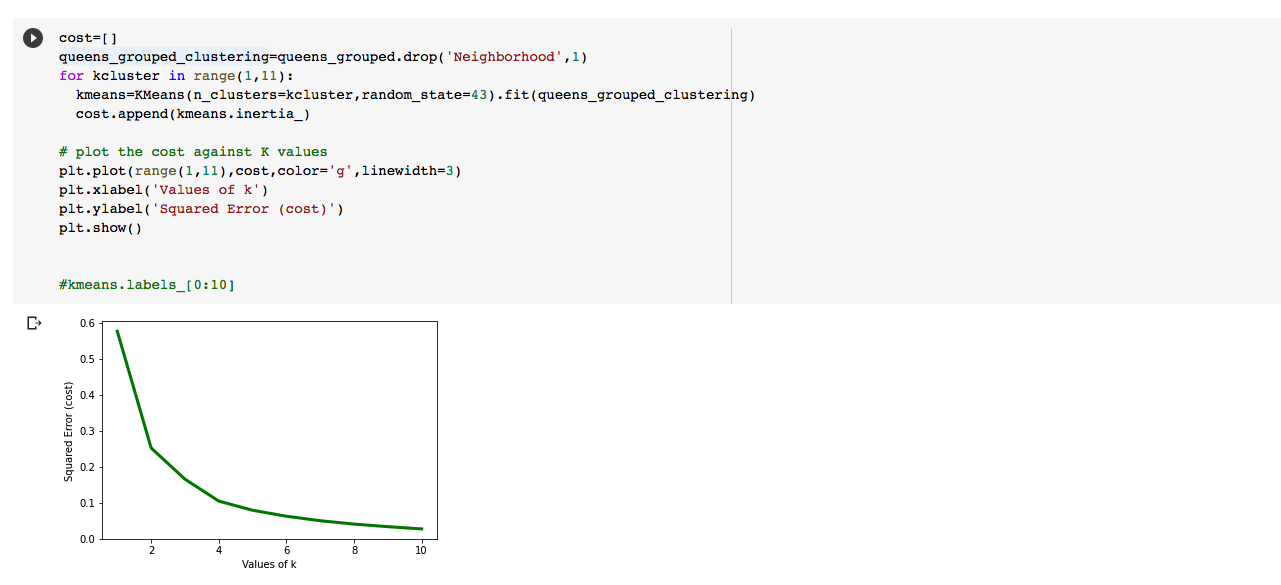
1. Using Foursquare to explore Neighborhood in Queens



1. I find the mainly Chinese restaurants belong to “Asian Restaurant” and “Chinese Restaurant” venues. I pick these two venues from the database and group them by “Neighborhood”. The below shows the mean value of each neighborhood for two restaurant venues.



1. I run KMean clustering to analyze the data. I find the data can be divided into 3 clusters

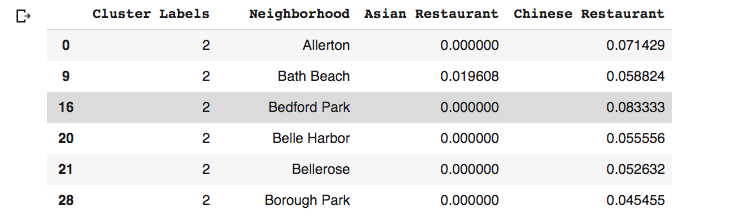


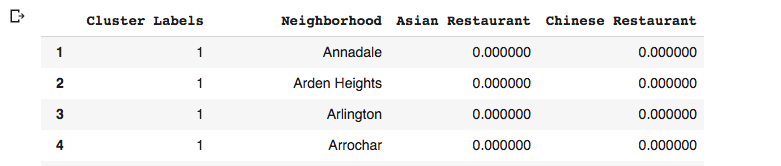
cluster 0 has the highest density of the Chinese restaurant

cluster 2 has the medium density of the Chinese restaurant

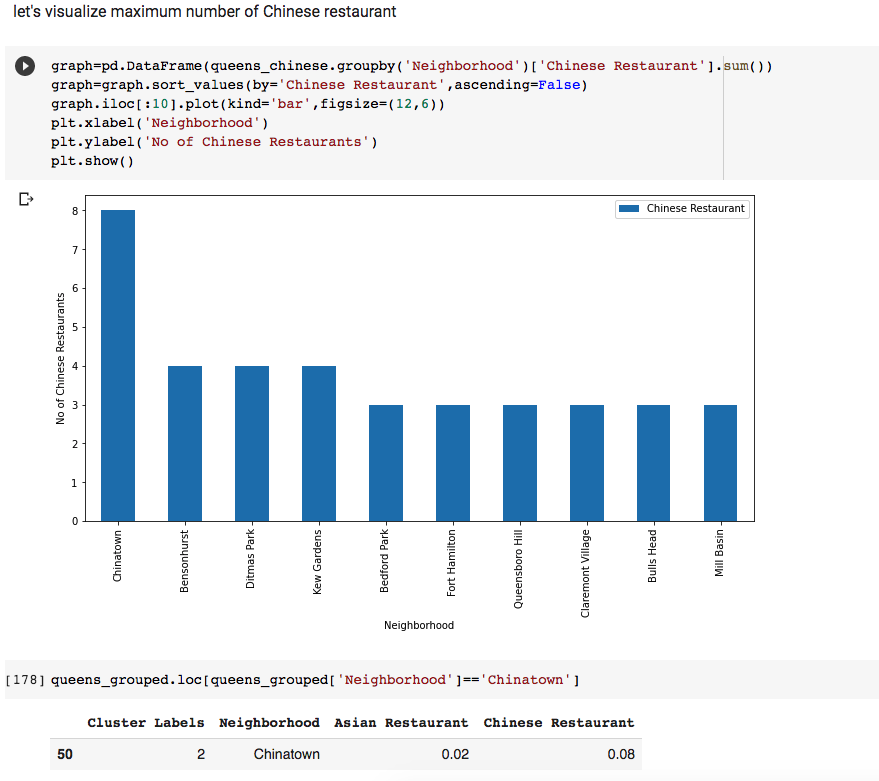
cluster 1 has the lowest density of the Chinese restaurant.



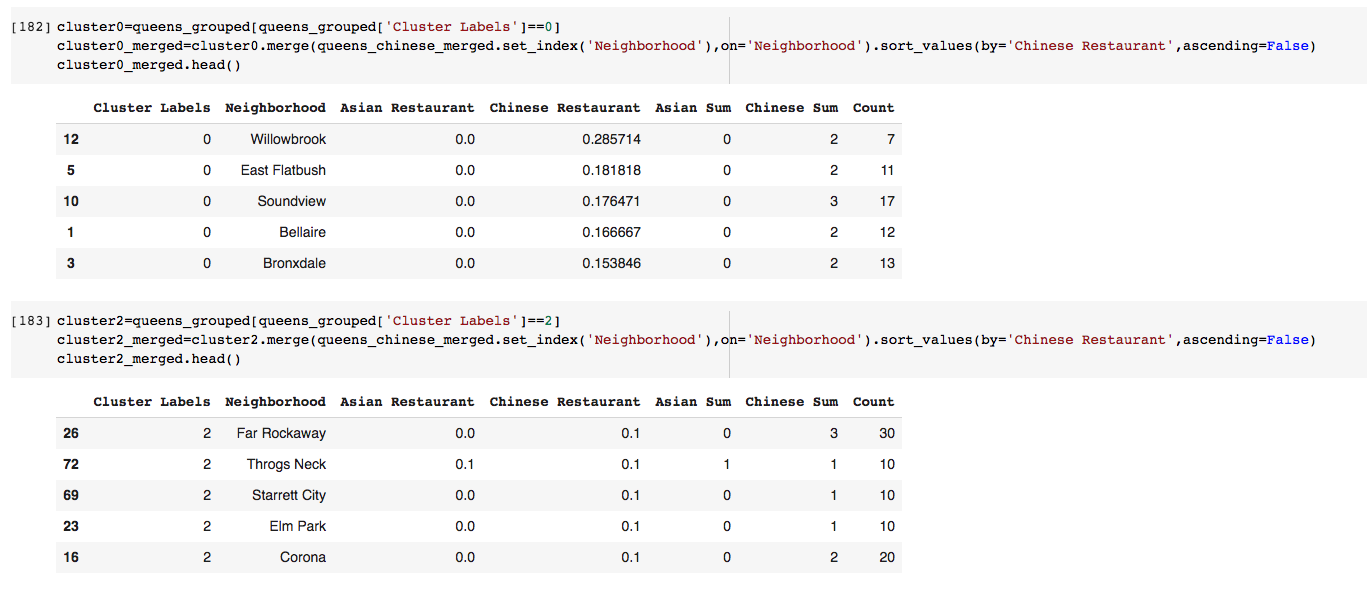


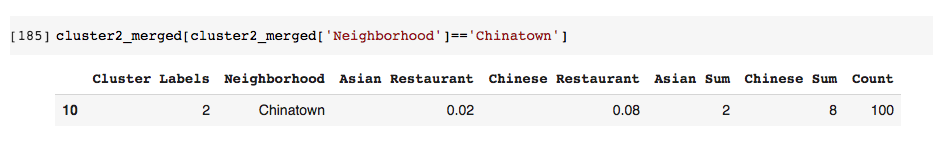


1. However, I find that Chinatown has the highest number of the Chinese restaurant, which belongs to the Cluster 2.



1. I suppose that the cluster 2 has a large number of shops, which decrease the density of the Chinese restaurants.





As we see, the number of the shop in cluster 2 is larger than that in cluster 0. Especially, for the Chinatown neighborhood, there are 100 shops, which is much larger than the number of shops in cluster 0.

1. Results

The results of the exploratory data analysis and clustering is summarized below:

* Willowbrook neighborhood has the highest density of Chinese restaurants
* Chinatown neighborhood has the highest number of Chinese restaurants
* Cluster 1 neighborhoods have the least number of Chinese restaurants.
* I will choose neighborhood in cluster 2 such as Far Rockaway and Throgs Neck to open a chinese restaurant. Because there are many shops and few Chinese restaurants. The large number of shops will attract many people come there. and the less restaurants mean a less competition.

**Discussion**

According to this analysis, Far Rockaway will be the best place to open a chinese restaurant. Because there are 30 shops in the Far Rockaway neighborhood, and there are only 3 restaurants. This small number of chinese restaurant could not satisfy so many people around these 30 shops.

Some drawbacks of analysis are: the clustering is completely based on the data provided by Foursquare API. Since land price, the distance of venues from the closest station, the number of potential customers, could all play a major role and thus, this analysis is definitely far from being conclusory. However, it definitely gives us some very important preliminary information on the possibilities of opening restaurants in the Queens borough of New York City. Also, another pitfall of this analysis could be the consideration of only one major borough of New York City, taking into account all the areas under the 5 major boroughs that would give us an even more realistic picture. Furthermore, these results also could potentially vary if we use some other clustering techniques like DBSCAN.

**Conclusion**

Finally, to conclude this project, we have got a small glimpse of how a real-life Data science project looks like. I have used some frequently used python libraries to handle JSON file, plotting graphs, and other exploratory data analysis. Use Foursquare API to major boroughs of New York City and their neighborhoods. The potential for this kind of analysis in a real-life business problem is discussed in great detail. Also, some of the drawbacks and chances for improvements to represent even more realistic pictures are mentioned. As a final note, all of the above analyses is depended on the adequacy and accuracy of Four Square data. A more comprehensive analysis and future work would need to incorporate data from other external databases.