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.

2. 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

	Borough	Neighborhood	Latitude	Longitude
0	Queens	Astoria	40.768509	-73.915654
1	Queens	Woodside	40.746349	-73.901842
2	Queens	Jackson Heights	40.751981	-73.882821
3	Queens	Elmhurst	40.744049	-73.881656
4	Queens	Howard Beach	40.654225	-73.838138

---

2. Using Foursquare to explore Neighborhood in Queens

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Wakefield	40.894705	-73.847201	Lollipops Gelato	40.894123	-73.845892	Dessert Shop
1	Wakefield	40.894705	-73.847201	Rite Aid	40.896649	-73.844846	Pharmacy
2	Wakefield	40.894705	-73.847201	Carvel Ice Cream	40.890487	-73.848568	Ice Cream Shop
3	Wakefield	40.894705	-73.847201	Walgreens	40.896528	-73.844700	Pharmacy
4	Wakefield	40.894705	-73.847201	Subway	40.890468	-73.849152	Sandwich Place

3. 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.

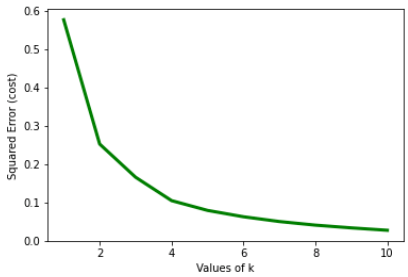
	Neighborhood	Asian Restaurant	Chinese Restaurant
0	Allerton	0.0	0.071429
1	Annadale	0.0	0.000000
2	Arden Heights	0.0	0.000000
3	Arlington	0.0	0.000000
4	Arrochar	0.0	0.000000

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

```
cost=[]
queens_grouped_clustering=queens_grouped.drop('Neighborhood',1)
for kcluster in range(1,11):
  kmeans=KMeans(n_clusters=kcluster,random_state=43).fit(queens_grouped_clustering)
  cost.append(kmeans.inertia_)

# plot the cost against K values
plt.plot(range(1,11),cost,color='g',linewidth=3)
plt.xlabel('Values of k')
plt.ylabel('Squared Error (cost)')
plt.show()


#kmeans.labels_[0:10]
```



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.

	Cluster Labels	Neighborhood	Asian Restaurant	Chinese Restaurant
18	0	Beechhurst	0.000000	0.117647
19	0	Bellaire	0.000000	0.166667
23	0	Bensonhurst	0.031250	0.125000
34	0	Bronxdale	0.000000	0.153846
54	0	Claremont Village	0.000000	0.142857

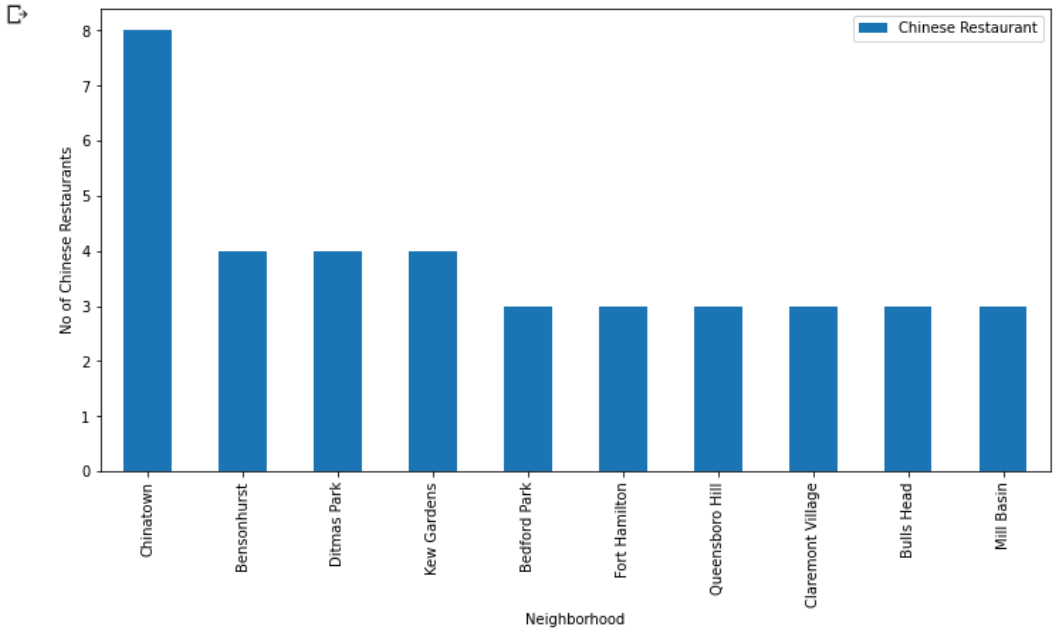
	Cluster Labels	Neighborhood	Asian Restaurant	Chinese Restaurant
0	2	Allerton	0.000000	0.071429
9	2	Bath Beach	0.019608	0.058824
16	2	Bedford Park	0.000000	0.083333
20	2	Belle Harbor	0.000000	0.055556
21	2	Bellerose	0.000000	0.052632
28	2	Borough Park	0.000000	0.045455

	Cluster Labels	Neighborhood	Asian Restaurant	Chinese Restaurant
1	1	Annadale	0.000000	0.000000
2	1	Arden Heights	0.000000	0.000000
3	1	Arlington	0.000000	0.000000
4	1	Arrochar	0.000000	0.000000

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

let's visualize maximum number of Chinese restaurant

```
graph=pd.DataFrame(queens_chinese.groupby('Neighborhood')['Chinese Restaurant'].sum())
graph=graph.sort_values(by='Chinese Restaurant',ascending=False)
graph.iloc[:10].plot(kind='bar',figsize=(12,6))
plt.xlabel('Neighborhood')
plt.ylabel('No of Chinese Restaurants')
plt.show()
```




```
[178] queens_grouped.loc[queens_grouped['Neighborhood']=='Chinatown']
```

	Cluster Labels	Neighborhood	Asian Restaurant	Chinese Restaurant
50	2	Chinatown	0.02	0.08

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

```
[182] cluster0=queens_grouped[queens_grouped['Cluster Labels']==0]
      cluster0_merged=cluster0.merge(queens_chinese_merged.set_index('Neighborhood'),on='Neighborhood').sort_values(by='Chinese Restaurant',ascending=False)
      cluster0_merged.head()
```

	Cluster Labels	Neighborhood	Asian Restaurant	Chinese Restaurant	Asian Sum	Chinese Sum	Count
12	0	Willowbrook	0.0	0.285714	0	2	7
5	0	East Flatbush	0.0	0.181818	0	2	11
10	0	Soundview	0.0	0.176471	0	3	17
1	0	Bellaire	0.0	0.166667	0	2	12
3	0	Bronxdale	0.0	0.153846	0	2	13

```
[183] cluster2=queens_grouped[queens_grouped['Cluster Labels']==2]
      cluster2_merged=cluster2.merge(queens_chinese_merged.set_index('Neighborhood'),on='Neighborhood').sort_values(by='Chinese Restaurant',ascending=False)
      cluster2_merged.head()
```

	Cluster Labels	Neighborhood	Asian Restaurant	Chinese Restaurant	Asian Sum	Chinese Sum	Count
26	2	Far Rockaway	0.0	0.1	0	3	30
72	2	Throgs Neck	0.1	0.1	1	1	10
69	2	Starrett City	0.0	0.1	0	1	10
23	2	Elm Park	0.0	0.1	0	1	10
16	2	Corona	0.0	0.1	0	2	20

```
[185] cluster2_merged[cluster2_merged['Neighborhood']=='Chinatown']
```

	Cluster Labels	Neighborhood	Asian Restaurant	Chinese Restaurant	Asian Sum	Chinese Sum	Count
10	2	Chinatown	0.02	0.08	2	8	100

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.

7. 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.