Open an authentic Chinese restaurant in Staten Island, NYC

Introduction & Business Problem

Problem Description

The City of New York is the financial capital and the largest city of the United States. The city is well known as the "melting pot" of diverse cultures. It provides lot of business opportunities and business friendly environment.

New York City has 5 boroughs Bronx, Brooklyn, Manhattan, Queens and Staten Island-Each with dozens of neighborhoods lending their own local flavor. The present project is going to focus on studying the possibility to open a new Chinese restaurant in the relatively underexplored and underserved New York City Staten Island borough.

Problem Statement

1. What is the best Neighborhood/Location for a new Chinese restaurant in Staten Island, New York City?
2. Which Neighborhood/Location offers the best chance of success to a new Chinese restaurant?

Target Audience

- Business starters who plan to open their first restaurant.
- Restaurant owners who want to invest in a chain store.
- Data Scientists/Analysts who want to design a ML flow for finding the best location to open a business at a given location.
- Students or professionals interested in the Data Science field who want to learn and implement some of the most used Data Analysis techniques.

Success Criteria

The success criteria of the project will be a good recommendation of Neighborhood choice to interested party based on scarcity of any or Chinese restaurant and a sizeable potential clients pool surrounding the chosen location.

Data Section

For the purpose of this project, the following data are going to be used

1. New York City dataset

- The NYC data contains all 5 Boroughs and their Neighborhoods, as well as the coordinates that define their location. The dataset provide the ability to explorer all the Neighborhoods in Staten Island.
- Data Source https//cocl.us/new_york_dataset
- Data Example

]:

	Borough	Neighborhood	Latitude	Longitude
0	Staten Island	St. George	40.644982	-74.079353
1	Staten Island	New Brighton	40.640615	-74.087017
2	Staten Island	Stapleton	40.626928	-74.077902
3	Staten Island	Rosebank	40.615305	-74.069805
4	Staten Island	West Brighton	40.631879	-74.107182

2. Foursquare Location data

- The Foursquare dataset has the Venues location at each Neighborhood in Staten Island. The venues data are further drill down to Restaurant and then Chinese Restaurant as sub-category.
- Data Source https//api.foursquare.com/
- Data Example

]:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	St. George	40.644982	-74.079353	Beso	40.643306	-74.076508	Tapas Restaurant
1	St. George	40.644982	-74.079353	Staten Island September 11 Memorial	40.646767	-74.076510	Monument / Landmark
2	St. George	40.644982	-74.079353	A&S Pizzeria	40.643940	-74.077626	Pizza Place
3	St. George	40.644982	-74.079353	Richmond County Bank Ballpark	40.645056	-74.076864	Baseball Stadium
4	St. George	40.644982	-74.079353	Shake Shack	40.643660	-74.075891	Burger Joint

3. Asian Population dataset

- The Asian Population data provide the Asian population in each Staten Island Neighborhood and the percentage it represent in respect to the total population. These population data serve as background information to access whether a neighborhood has the potential customer base to support a new Chinese Restaurant.
- Data Source https//statisticalatlas.com/
- Data Example

	Neighborhood	% of Asian Population	Count	Rank
0	Emerson Hill	23.40%	3,092	1.0
1	Clifton	20.70%	584	2.0
2	New Springville	19.30%	1,096	3.0
3	Graniteville	18.00%	2,035	4.0
4	Heartland Vlg	18.00%	3,728	5.0

Methodology

From the three datasets described in the Data Section, the idea is to obtain the neighborhoods information, venues (Chinese restaurants), and populations and run a Classification analysis with K-mean Clustering.

The "Elbow" method is used to determine the best number of k for the K-mean engine. Once k is determined, we will proceed to classify the neighborhoods and label them by cluster. The classified neighborhoods will be plotted on a Folium map for Analysis.

In addition to the Folium map, an Asian population per neighborhood with its number of Chinese restaurants will be plotted in a bar plot to visualize the top neighborhood with high Asian population and low numbers of Chinese restaurant.

The use of these two types of visualization aid will help to find the neighborhood that lack of Chinese restaurant and has a sizeable potential client's pool. Such location has a greater success rate to start a new business.

Exploratory Data Analysis

Foursquare data is very comprehensive tool and it powers location data for developers. The Foursquare API was used to retrieve information about the Venue, Venue category with their longitudes and latitudes. The call returns a JSON file and it was put into a data-frame. Here 100 popular spots for each neighborhood were chosen with a radius of 750 meters. A data-frame named "staten_island_venues" was obtained from the JSON file and showed in the Data Section.

The "staten_island_venues" data-frame returned 1515 total and 212 unique venue categories.

```
: print(staten_island_venues.shape)
  staten_island_venues.head()
```

```
(1515, 7)
```

```
: print('There are {} uniques categories.'.format(len(staten_island_venues['Venue Category'].unique())))
```

```
There are 212 uniques categories.
```

The one hot encoding for getting dummies of the venue category was ran, and rows were grouped by neighborhood to calculate the mean of the frequency of occurrence of each category.

	Neighborhood	ATM	American Restaurant	Arcade	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	Automotive Shop	BBQ Joint	Bagel Shop	Bakery	Bank
0	Annadale	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.076923	0.000000
1	Arden Heights	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.076923	0.000000	0.000000
2	Arlington	0.000000	0.090909	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
3	Arrochar	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.030303	0.000000	0.000000	0.030303	0.000000	0.000000
4	Bay Terrace	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
5	Bloomfield	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.076923	0.000000	0.000000
6	Bulls Head	0.000000	0.020000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.040000	0.000000	0.020000
7	Butler Manor	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
8	Castleton	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.027027	0.000000	0.054054

Next, the venues dataset was further drilled down to Restaurant and Chinese Restaurant category for further analysis, and a new data-frame "SI_restaurants" was created as below:

	Neighborhood	Total Restaurants	Total Chinese Restaurants
0	Annadale	2	0
1	Arden Heights	2	1
2	Arlington	2	0
3	Arrochar	5	0
4	Bay Terrace	3	0

A Staten Island Asian Population dataset was cleaned and prepared. The data-frame is called "asian_pop_data" and it has the total and Asian population's information for each neighborhood.

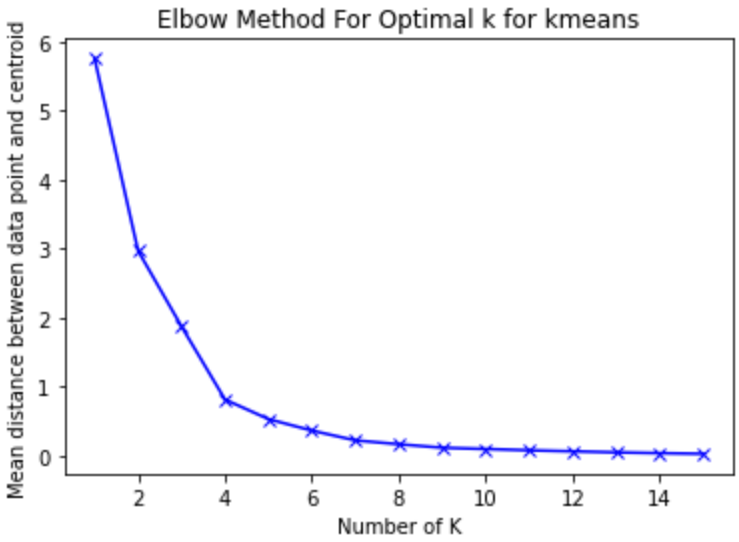
	Neighborhood	Asian_Population	Total_Population
0	Annadale	270.0	7297
1	Arden Heights	741.0	13722
2	Arrochar	1025.0	8471
3	Bay Terrace	321.0	7295
4	Bulls Head	821.0	13241

A final data-frame was obtained by merging "SI_restaurant" with "asian_pop_data". The resulting data-frame is called "staten_island_rest_pop_data" as shown below.

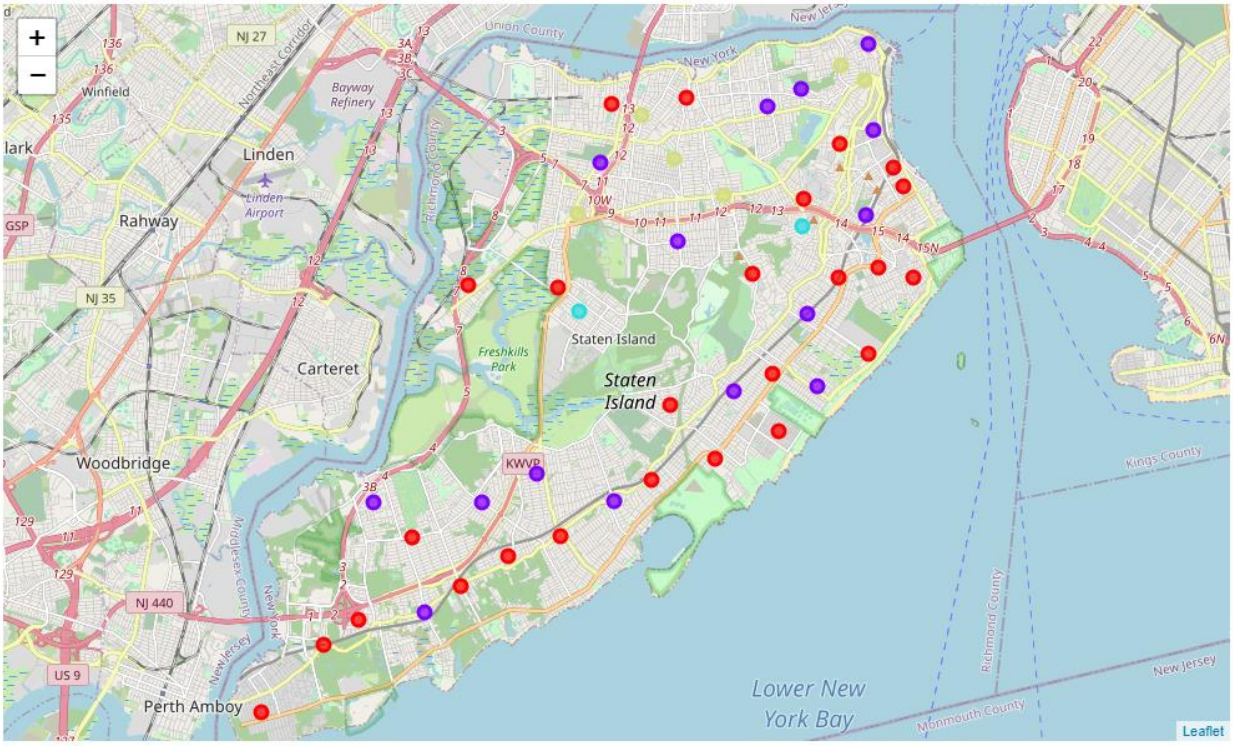
	Neighborhood	Total Restaurants	Total Chinese Restaurants	Asian_Population	Total_Population
0	Annadale	2.0	0.0	270.0	7297.0
1	Arden Heights	2.0	1.0	741.0	13722.0
2	Arrochar	5.0	0.0	1025.0	8471.0
3	Bay Terrace	3.0	0.0	321.0	7295.0
4	Bulls Head	13.0	3.0	821.0	13241.0

From the obtained neighborhoods, venues (Chinese restaurants), and Asian populations information, a classification analysis with K-mean Clustering was carried out by using the available data.

The first part of K-mean clustering was to perform the "Elbow" method in order to determine the best number of k, the numbers of centroid. To do this an iterative calculation of mean distance between data point (SSE) and centroid for k values from 1 to 15. In the SSE vs. K scatter point plot, the best k value is defined by the inflection point in the curve where the mean square error starts to approach to zero. From the "Elbow Method For Optimal k for kmeans" plot it was determined that the best k had to be 4.

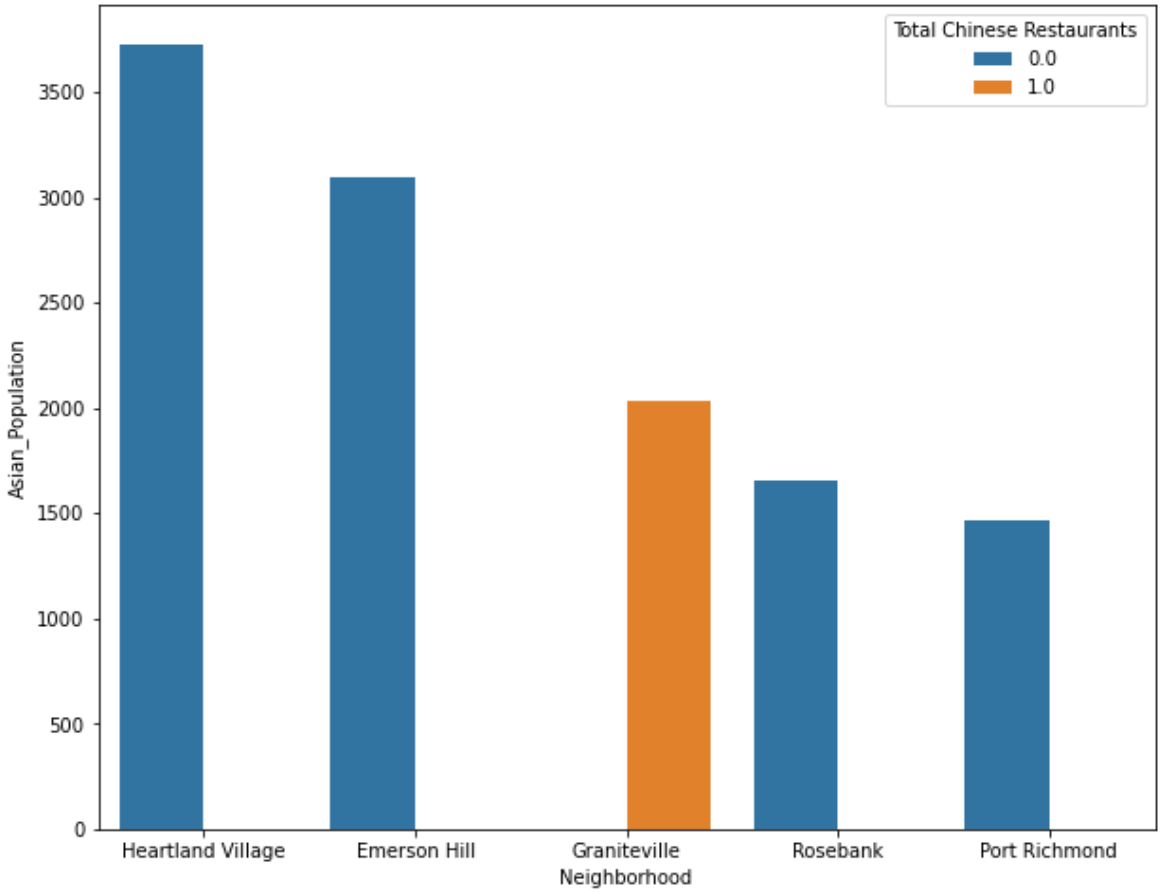


We classified the neighborhoods and label them by clusters 0 to 3. The classified neighborhoods were plotted using Folium for Analysis.



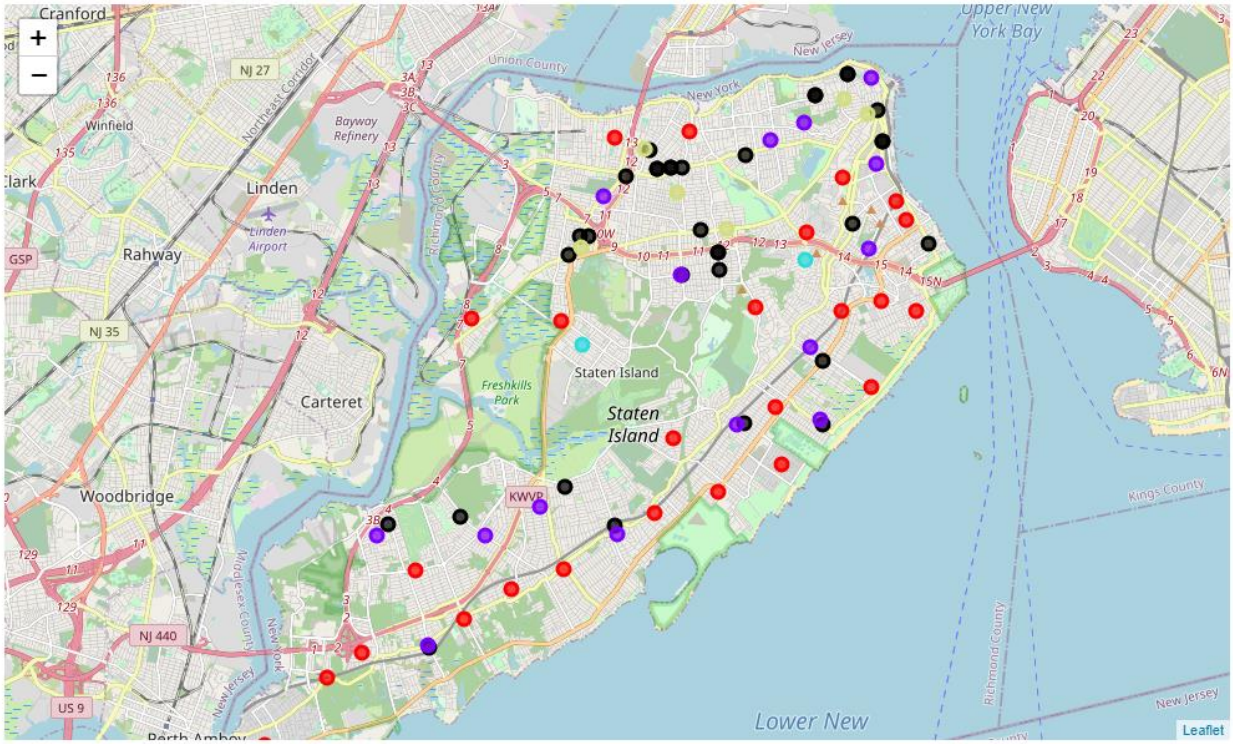
The above neighborhood classification map showed that Cluster 1 (C0 represented by red dot) have 0 restaurants and the low to medium low Asian population. Cluster 2 (C1 represented by purple dot) have 1 restaurants and the population numbers are similar to C1. Cluster 3 (C2 represented by light blue dot) have 0 restaurants and the highest Asian population, they indicate an underserved community and thus they have the most potential to host new Chinese restaurants. Cluster 4 (C3 represented by light yellow dot) have 2 or 3 restaurants and low to medium low Asian population which suggest that these neighborhoods are well served and probably offer low potential for new restaurant.

Next step was to focus more on Cluster 1 (C0), Cluster 2 (C1), and Cluster 3 (C2), doing more analysis and pick our top recommendations. Since the success criteria of the project is to recommend good Neighborhood choice to interested party based on scarcity of any or Chinese restaurant and a sizeable potential clients pool. The first analysis was to sort the focused data by Asian population from high to low, and plot the top 5 highest Asian population neighborhood with its number of Chinese restaurant in the third dimension (color coded bar).



The above graph (Asian population bar plot) showed that Heartland Village and Emerson Hill have 0 Chinese restaurants and the first and second highest Asian population respectively. Graniteville has 1 restaurant and the third highest population numbers.

Based on this finding, it was decided to further focus on these three neighborhoods. In the following map (Map Restaurant), a second map of the classified neighborhoods with all the Chinese restaurants was plotted using Folium to visualize the locations of interest.



On the above Folium map (Map Restaurant): black dot represent Chinese restaurants. Red dot is Cluster 1 (C0), purple dot is Cluster 2 (C1), light blue dot is Cluster 3 (C2), and light yellow dot is Cluster 4 (C3).

Results and Discussion

The results show that there are a great number of Chinese restaurants in the northeastern part of the Staten Island (C4 in the clustering analysis). These communities are relative well served and are excluded from the recommended locations. Surrounding the island there are the neighborhoods with 0 to 1 Chinese restaurant and low to moderate Asian population (C1, C2 in the clustering analysis). Their potential is interpreted as average; on the center of the island (C3 in the clustering analysis) the two neighborhoods with the most potential was identified as Heartland Village and Emerson Hill. Both neighborhoods have two of the highest Asian population, and none Chinese restaurant to serve the community.

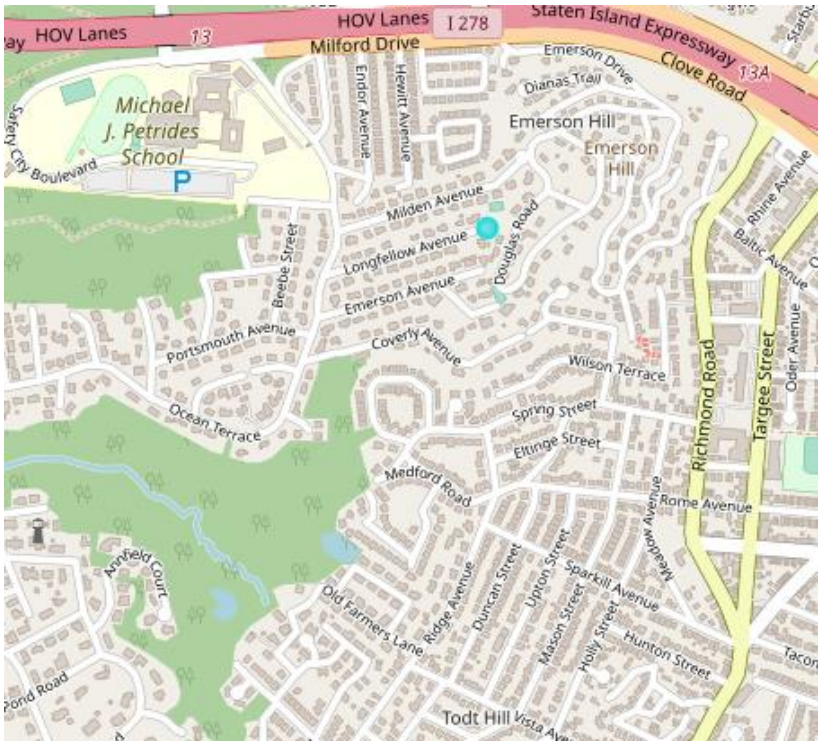
When evaluating the Asian population bar plot and the Map Restaurant Folium map, a third candidate has emerged. Graniteville located in the northwestern part of the island has the third highest Asian population and only one Chinese restaurant serving the area.

The top recommendation for the most successful place to open a new Chinese restaurant is at Heartland Village. As for the second place, both Emerson Hill and Graniteville have pros and cons; they are in a close call or even could tie for this tier.

Conclusion

The top recommendation is Heartland Village. It has the highest Asian population and no Chinese restaurant serving the area. It is also highly suggested to choose a location close to Staten Island Mall for the best place to set up a new business. Another factor to consider is the fact there are three other kinds of restaurants near the Mall. If a new Chinese restaurant is opened nearby there could be some competition between the four. But since they are not of the same kind of cuisine what they will offer is a more diverse choice of meal to customers. A place with four restaurants near the Mall will be remembered by people as the "Restaurant Corner" and it will attract more and more food passioners to come out and eat here. In short, a new Chinese restaurant in Heartland Village near the Staten Island Mall offers little or no competition from other restaurants, and it will also enjoy a broad customer base from both the Asian folks living in the area and people who visit the Mall.

For the second place of best success chance, the choice is between Emerson Hill and Graniteville. Emerson Hill has the second most Asian population and none of the Chinese restaurants. However, is a residential neighborhood with few other business and point of attractions, thus the potential customer base narrowly relies on the residents in the area; Graniteville on the other hand has the third most Asian population, and only one Chinese restaurant. The local shopping plaza being a good place to set up a new Chinese restaurant, this location could provide a similar broad potential customer base just like the Heartland Village case.



Considering all the pros and cons, overall the second place for the best success chance is given to Graniteville. As for the competition posted by the other restaurant of the same kind, they are the two sides of one coin. Healthy competition can drive business to thrilled, and besides the fact that there is a Chinese restaurant in the area means a pro-Chinese cuisine customer base is already built. A new restaurant of the same kind can start by giving a more diverse choice to Chinese food lovers rather than building customer preference from scratch.