Locating A Chinese Restaurant in New York City or Toronto

July 29, 2020

Table of Contents

1. Introduction

Mr. Chen would like to open a new Chinese restaurant in either of Toronto or New York City and he would like to know which of Toronto or New York is a better location for a Chinese restaurant. To ensure the success of this new Chinese restaurant business, Mr. Chen wants to know which neighborhood in either of Toronto or New York City would be an ideal neighborhood for his new Chinese restaurant.

Mr. Chen has recently been granted the Canadian immigration visa, and with a stroke of luck, also won the U.S Green Card lottery. Mr. Chen has an extensive experience running Chinese restaurants in his native country of China and he would like to continue running such restaurants in Canada or the U.S. Before deciding on whether to move to Canada or the U.S, Mr. Chen did an initial research that concluded that either New York or Toronto would be a good place to start a Chinese restaurant, allowing him the opportunity to put his vast experience running such a restaurant in China to good use in his newly adopted country of Canada or the U.S.

The goal would be to develop a model that can help Mr. Chen or any investor determine the ideal neighborhood to locate a restaurant in Toronto or New York City to ensure success. An ideal neighborhood is one that meets the following three criteria:

- Residents with above average income.
- Currently under-saturated by Chinese restaurants.
- Have a sizeable population that could patronize the new restaurant.

The model that would be built would be useful for any investor setting up any type of restaurant in New York City or Toronto. Furthermore, anyone with a basic understanding of Data Science would be able update the model to find an ideal location for restaurants in other cities.

2. Methodology

To complete the exercise requested, a search was conducted to collect the required data for the exercise. Some location data would be needed for both Toronto and New York City to use in Foursquare to investigate other restaurants in the cities to investigate potential for market saturation with the same type of restaurants. Furthermore, some demographics information, particularly population and income per neighborhood, would also be needed to get a measure of the market size and potential clientele of the new restaurant. The data collected is described in the following section.

2.1 Data Collection

The following key data were collected to facilitate the analysis required for the building of the model.

1. **New York Neighborhood Location Data:** The neighborhood location data for New York City came from the site https://cocl.us/new_york_dataset. The link to the site was provided as part of the Capstone Project class. The data from the site is a json format that had to be scraped with the JSON library in Python. After scraping with the library, a data frame of New York neighborhood location data was created from the data.
2. **Toronto Neighborhood Location Data:** The neighborhood location data was found on the Wikipedia site https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M. The site was provided as part of the Capstone Project course. The data table from the site was scraped with BeautifulSoup library in Python and loaded into a data frame. The data from this site only had postal code information for each Toronto neighborhood and the latitude and longitude data had to sourced differently. The latitude and longitude information was sourced from the site https://cocl.us/Geospatial_data, and read directly as csv format. The two files were combined to create a data frame of neighborhoods location data.
3. **New York Neighborhood Population Data:** Population data for New York neighborhoods was sourced from the site https://data.cityofnewyork.us/api/views/swpk-hqdp/rows.csv?accessType=DOWNLOAD, owned by the City of New York. The link provided neighborhood population data as downloadable csv format. The data was directly downloaded into a data frame.
4. **New York Neighborhood Income Data:** Income data for New York neighborhoods was sourced from the site https://ny.curbed.com/2017/8/4/16099252/new-york-neighborhood-affordability, a NYC real estate report provided by Curbed New York. Scraping the site with BeautifulSoup was impossible. Instead, the table was downloaded into a csv and loaded into a data frame. The income and population data were combined into a single demographic data frame.
5. **Toronto Neighborhood Population and Income Data:** Neighborhood income and population data for Toronto were sourced from the Wikipedia site https://en.wikipedia.org/wiki/Demographics_of_Toronto_neighbourhoods. Information on the site included multiple tables, which were scraped with BeautifulSoup in Python and put in a data frame.

2.2 Exploratory Data Analysis Methodology

After the location and demographics data frames for both cities were formatted, the following exploratory analyses were completed to get an understanding of the data, as well as to prepare the data for analysis and modeling.

2.2.1 Mapping Analysis

To show the relative distribution of neighborhoods in terms of population and income, bubble maps were created for both cities as shown below with bubbles sized based on population and income.

Figure 1A and Figure 1B for Toronto show that neighborhoods with higher average income generally have lower population and the converse is also true. These plots also show a wide distribution in population and income. Average incomes range from $22,000/year to $215,000/year in Toronto neighborhoods while neighborhood populations range from ~6,000 to ~49,000 per neighborhood. A descriptive statistics on the data frame showed significant population and income values for P25 and P27. Consequently, Boolean masking filtering was applied to the data frame based on average income inter-quartile range to remove extreme values of population and income. The data frame after filtering resulted in neighborhoods with sizeable population and average income, as shown in Figure 2A and Figure 2B. Another outcome of the filtering is that there is now a sizeable population for every neighborhood with significant income in the resulting data frame.

This process was repeated for the data from for New York City. Figure 3A and Figure 3B for New York City show that neighborhoods with higher average income also have lower population and the converse is true as found for Toronto. These plots also show a wide distribution in population and income. Average incomes range from $20,000/year to $130,000/year in Toronto neighborhoods while neighborhood populations range from ~16,000 to ~140,000 per neighborhood. A descriptive statistics on the data frame showed significant population and income values for P25 and P27. Consequently, Boolean masking filtering was applied to the data frame based on average income inter-quartile range to remove extreme values of population and income. The data frame after filtering resulted in neighborhoods with sizeable population and average income, as shown in Figure 2A and Figure 2B. Another outcome of the filtering is that there is now a sizeable population for every New York City neighborhood with significant income in the resulting data frame.

After all the exploratory data analysis and filtering, 19 neighborhoods in Toronto and 66 neighborhoods in Toronto have complete dataset for location, population and income.

Toronto Neighborhood Maps Before Inter-Quartile Income Filtering

Figure 1A: Bubble Map of TO Neighborhoods and Population (Before)

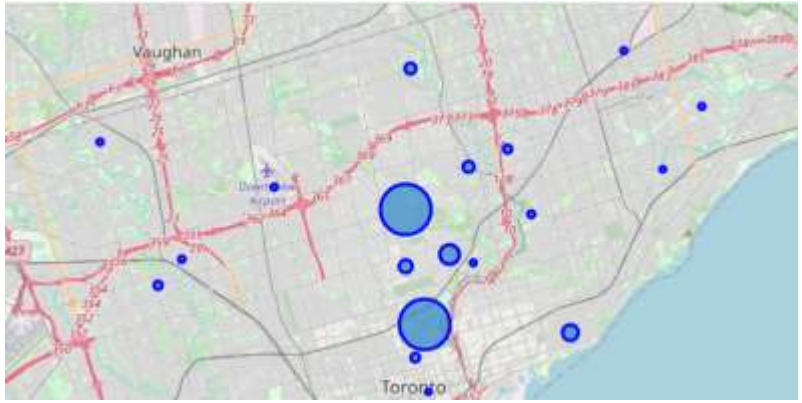

Figure 1B: Bubble Map of TO Neighborhoods and Incomes (Before)

Toronto Neighborhood Maps After Inter-Quartile Income Filtering

Figure 2A: Bubble Map of TO Neighborhoods and Populations (After)

Figure 2B: Bubble Map of TO Neighborhoods and Incomes (After)

New York City Neighborhood Maps Before Inter-Quartile Income Filtering

Figure 3A: Bubble Map of NYC Neighborhoods and Populations (Before)

Figure 3B: Bubble Map of NYC Neighborhoods and Incomes (Before)

New York City Neighborhood Maps After Inter-Quartile Income Filtering

Figure 4A: Bubble Map of NYC Neighborhoods and Populations (After)

Figure 4B: Bubble Map of NYC Neighborhoods and Incomes (After)

2.2.2 Population Normalization and Outlier Analysis

The post-filtering data frames for both cities were combined with columns added for cities to facilitate outlier analysis. Given the significant difference in population in Toronto and New York City, the population for both cities was normalized in a new column named 'restaurants per capita', which represents the number of restaurants in each neighborhood per 100,000 residents of each neighborhood. The restaurants per capita and income for both tables were checked for outliers using box plots. As shown in Figure 5, a Toronto neighborhood showing a restaurant per capital of ~500 was found to be an outlier and was removed from the data frame.

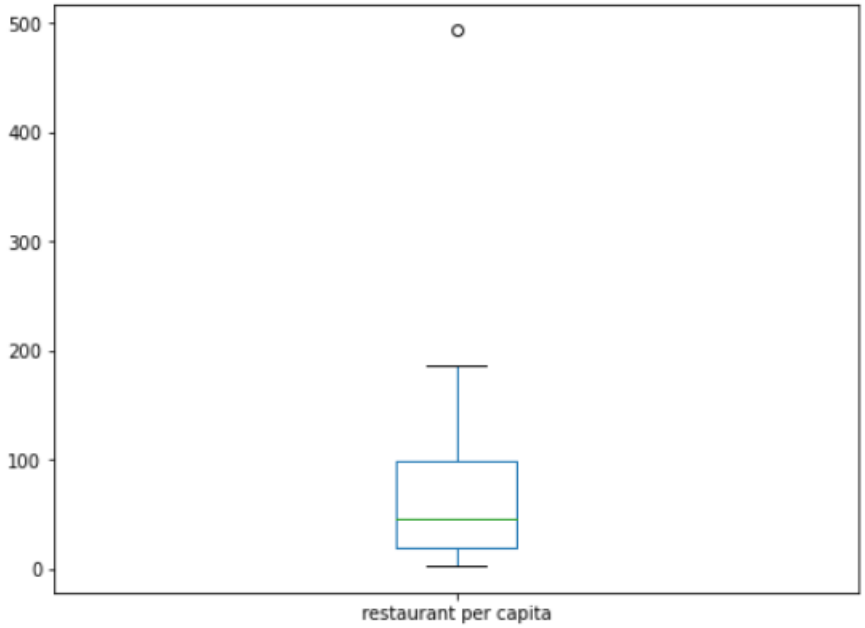

Figure 5: Outlier Analysis of Restaurant per Capita for Both Cities

2.2.3 Comparative Neighborhood Analysis

The neighborhoods in both cities were compared in a scatter plot with bubble sizing to narrow down neighborhoods with lowest restaurant per capital and high income. The scatter plot is colored by neighborhoods and the bubbles are sized to the population of the neighborhoods.

Figure 6 shows the comparative neighborhood analysis for Toronto, where the neighborhoods of Don Mills and Bayview Village show low restaurants per capita, high average neighborhood income and sizeable population. Figure 7 shows the comparative neighborhood analysis for New York City, where the neighborhoods of Flatlands, Glendale, Canarsie and Steinway show low restaurants per capita, high average neighborhood income and sizeable population.

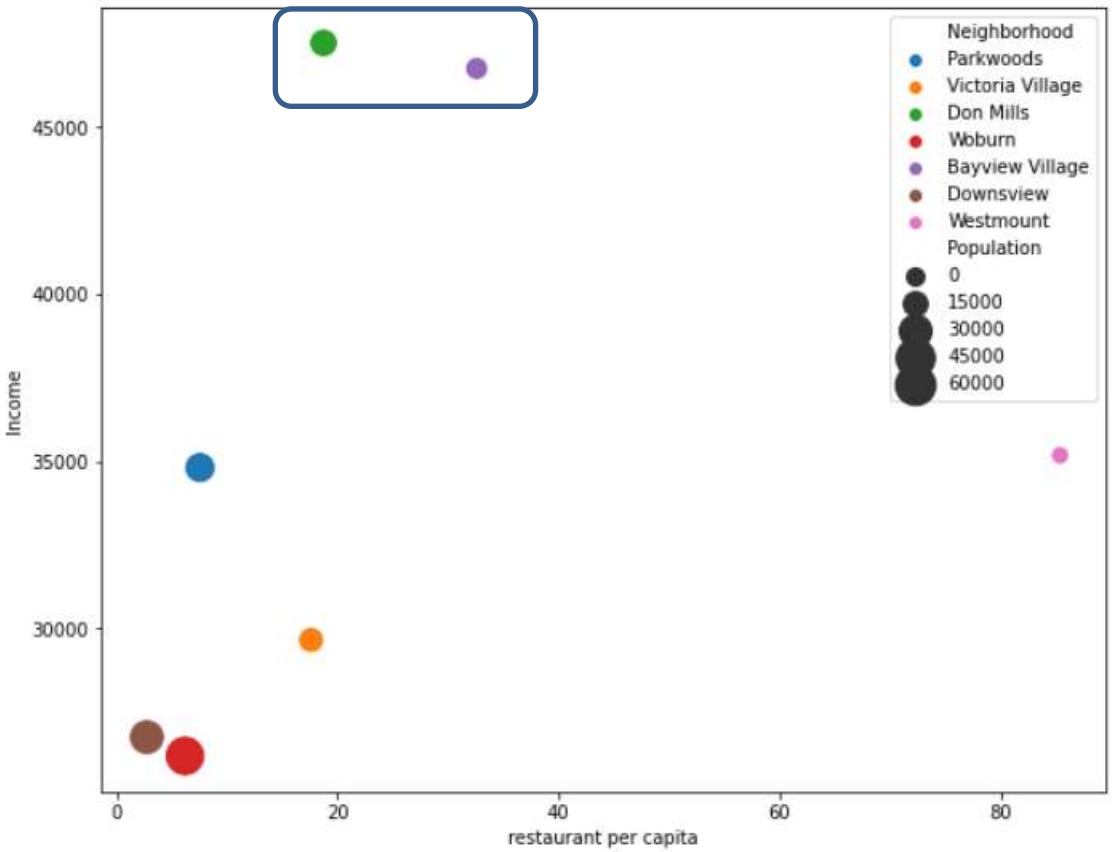

Figure 6: Comparative Neighborhood Analysis for Toronto

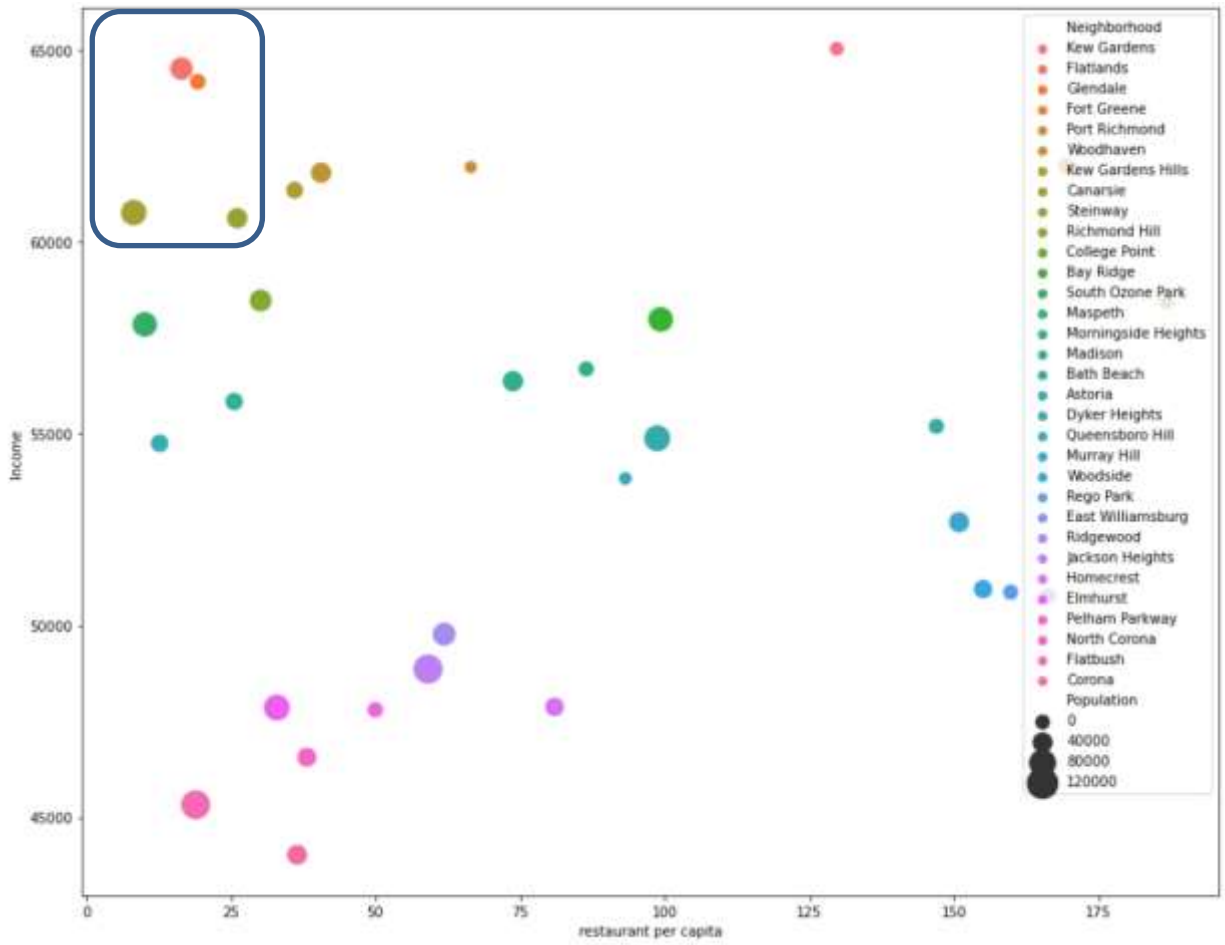

Figure 7: Comparative Neighborhood Analysis for New York City

2.2.4 Neighborhood Restaurants Analysis with Foursquare

The next step of the data analysis is to find potential competing restaurants in each of the neighborhoods, to ensure that the neighborhoods do not have too many Chinese restaurants. Post-filtering data for Toronto and New York City were analyzed with Foursquare to find top 100 nearby restaurants that are within 500 meters radius of each neighborhood.

2.2.5 Machine Learning and Cluster Analysis

The data from the neighborhood restaurants analysis was clustered into four using the K-means method. The top ten restaurant category in each neighborhood is ranked based on frequency of occurrence. The clusters allowed for the final analysis of the top ranked neighborhoods in Section 2.2.3. This is to ensure that those top ranked neighborhoods are not oversaturation by Chinese restaurants.

3. Results

Based on the analysis in the methodology section showing the top restaurants with residents earning a high income, low restaurant per capita and significant population were preliminarily selected for Toronto and New York City. Those neighborhoods were closely evaluated below in Figure 8. As shown the figure below, the top neighborhoods for a new Chinese restaurant would be Canarsie and Flatlands in New York City. Canarsie appears to be the best location given the comparatively lower restaurant per capita.

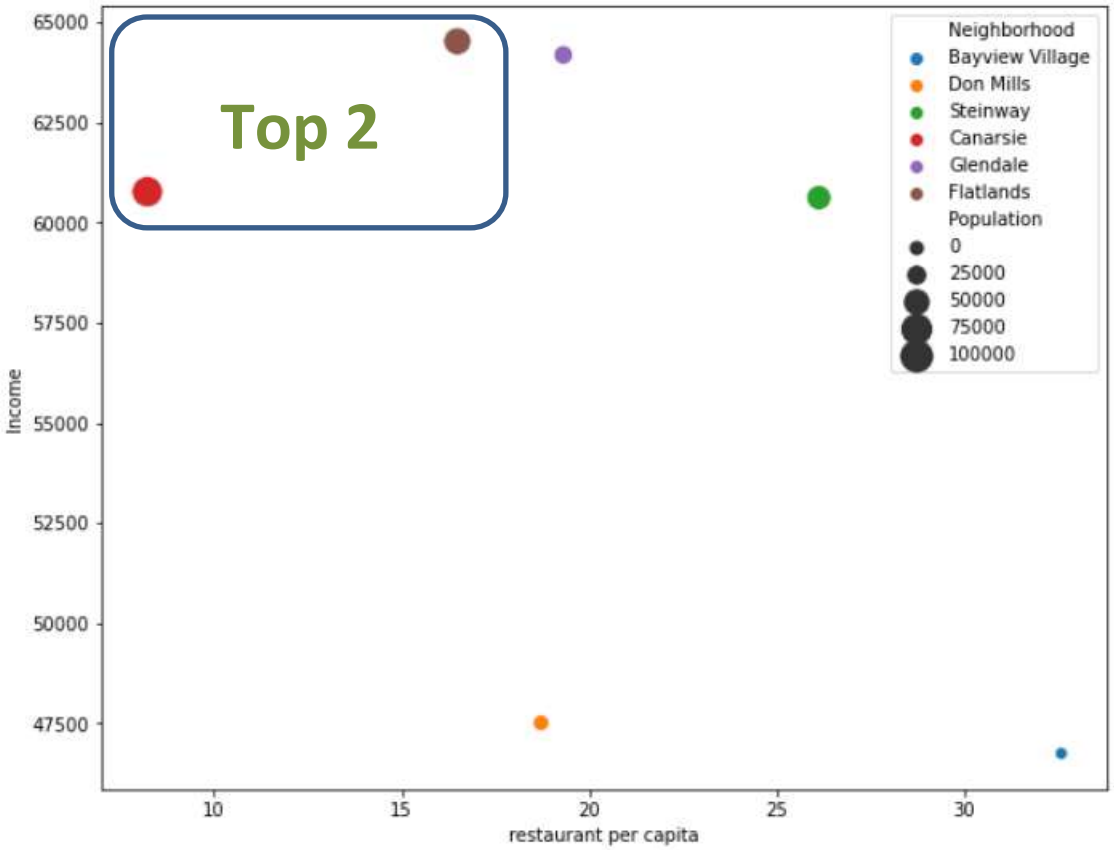

Figure 8: Comparative Analysis of Top Neighborhoods from Toronto and New York City

3.1 Confirmation with Cluster Analysis

Cluster analysis was completed for both Toronto and New York City and result mapped and tabulated.

3.1.1 Mapping of Cluster Analysis Results

Figure 9 shows the clusters for Toronto and Figure 10 shows the clusters for New York City. The clusters show the leading neighborhoods in Toronto in the same cluster, which makes sense given that both neighborhoods met the criteria for setting up a new Chinese restaurant. For New York City, the two leading neighborhoods fall into two different clusters. However the two leading neighborhoods are geographically close to each other. This again supports the reason both meet our criteria for setting up a new Chinese restaurant.

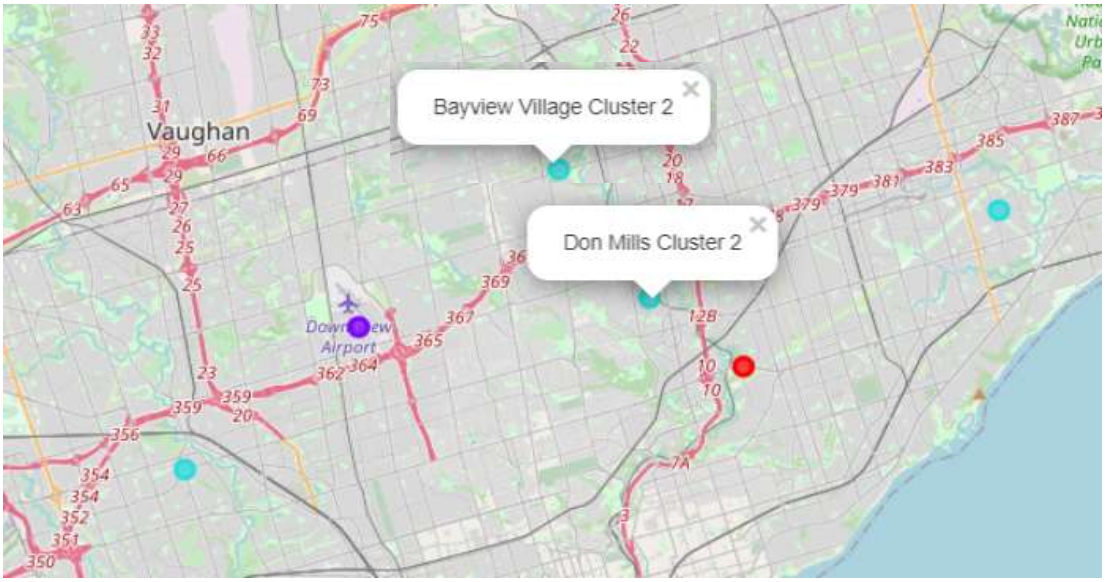

Figure 9: Mapping of Clusters for Toronto Neighborhoods

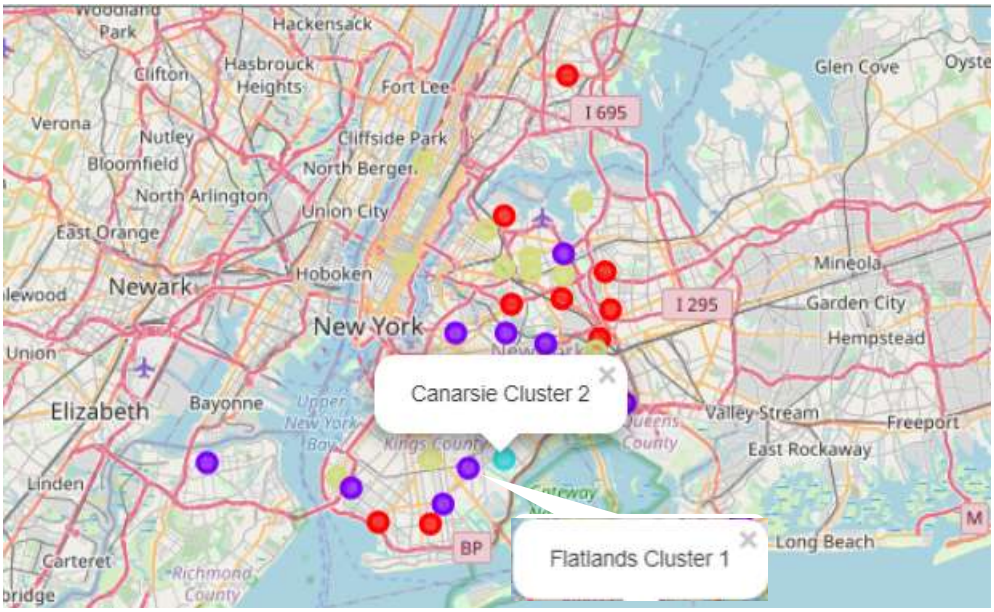

Figure 10: Mapping of Clusters for NYC Neighborhoods

3.1.2 Restaurant Saturation Analysis

Table 9 shows the cluster details of the top two neighborhoods for setting up a Chinese restaurant in Toronto. While they are not the final winning neighborhoods, the table shows how the two neighborhoods stack up against each other. Don Mill is the clear winner with Chinese restaurants being 8th most common restaurant in the neighborhood, which Chinese restaurants are second most common restaurants in Bayview Village. While residents of both Don Mill and Bayview have similar income, Don Mill also has almost half as many restaurants per capita as Bayview village with twice the population of Bayview Village. Don Mills has an income that is slight below Toronto average per neighborhood but a population that is above Toronto average per neighborhood, Bayview Village is below average on both front. Therefore, the model shows that Don Mill is the location of choice in Toronto for setting up a new Chinese restaurant.

Table 10 shows the cluster details of the top two neighborhoods for setting up a Chinese restaurant in New York. **As shown in Figure 8, Canarsie and Flatlands in New York City are also the top two preferred locations to set up a Chinese restaurant overall.** These two neighborhoods have income and population that are above neighborhood average for New York City. The table shows how the two neighborhoods stack up against each other. Chinese restaurants are approximately equally common in both Canarsie and Flatlands neighborhoods. While residents of both Canarsie and Flatlands have similar income, Canarsie also has almost half as many restaurants per capita as Flatlands with 30% higher population than Flatlands. Therefore, the model shows that Canarsie is the location of choice in New York City for setting up a new Chinese restaurant.

Table 1: Top Neighborhoods in Toronto

	1st	2nd
Neighborhood	Don Mills	Bayview Village
Population	21372	12280
Average Income	47515	46752
Cluster Labels	2	2
1st Most Common Restaurants	Japanese Restaurant	Japanese Restaurant
2nd Most Common Restaurants	Café	Chinese Restaurant
3rd Most Common Restaurants	Caribbean Restaurant	Café
4th Most Common Restaurants	Korean Restaurant	Asian Restaurant
5th Most Common Restaurants	Sandwich Place	Pizza Place
6th Most Common Restaurants	French Restaurant	Middle Eastern Restaurant
7th Most Common Restaurants	BBQ Joint	Mexican Restaurant
8th Most Common Restaurants	Chinese Restaurant	Korean Restaurant
9th Most Common Restaurants	Fast Food Restaurant	Sandwich Place
10th Most Common Restaurants	Snack Place	Indian Restaurant

Table 2: Best Overall Neighborhoods and Top Neighborhoods in New York City

	1st	2nd
Neighborhood	Canarsie	Flatlands
Population	85058	66726
Income	60766	64519
Cluster Labels	2	1
1st Most Common Restaurants	Food	Deli / Bodega
2nd Most Common Restaurants	Deli / Bodega	Caribbean Restaurant
3rd Most Common Restaurants	Thai Restaurant	Fast Food Restaurant
4th Most Common Restaurants	Asian Restaurant	Asian Restaurant
5th Most Common Restaurants	Chinese Restaurant	Restaurant
6th Most Common Restaurants	Caribbean Restaurant	Chinese Restaurant
7th Most Common Restaurants	Eastern European Restaurant	Fried Chicken Joint
8th Most Common Restaurants	Egyptian Restaurant	Seafood Restaurant
9th Most Common Restaurants	Empanada Restaurant	Filipino Restaurant
10th Most Common Restaurants	Ethiopian Restaurant	Eastern European Restaurant

4. Discussion

Mr. Chen plans to set up a new Chinese restaurant in either of Toronto or New York City, after initial analysis he had completed concluded that either of these cities would be ideal location for his restaurant. He wanted an analytic data science approach to determine which neighborhood(s) in either of Toronto or New York City would be an ideal neighborhood for his new Chinese restaurant to ensure success. Mr. Chen set the goal of developing a model that could help him or any investor to determine the ideal neighborhood to locate a restaurant in Toronto or New York City. It was agreed with Mr. Chen that an ideal neighborhood is one that meets the following three criteria:

- Residents with above average income.
- Currently under-saturated by Chinese restaurants.
- Have a sizeable population that could patronize the new restaurant.

The process of setting up the model got under with gathering the required data for the exercise. Neighborhood location data, neighborhood population data and neighborhood income data were collected for both cities to enable initial analysis. The data were cleaned and analyzed to narrow down the list of neighborhood to the most relevant in both cities. Population normalization was done for restaurant count in each neighborhood, by creating a restaurant per capita metric that helped to narrow down the best neighborhood identification process. Foursquare was used to explore the restaurants in neighborhoods in both cities, to check out competition and to evaluate oversaturation of neighborhoods by particular restaurants. K-means was finally used to complete clustering of neighborhoods and finally confirm restaurant saturation in each neighborhood.

4.1 Model Outcome: Best Neighborhood

As shown in Figure 8 and Table 2, Canarsie is the best neighborhood for Mr. Chen to set up a new Chinese restaurant, of all the neighborhoods in Toronto and New York City. Canarsie came out on top because of the high neighborhood income of ~$61,000, a restaurant per capita of 8 restaurants per 100,000 residents and a respectably high neighborhood population of 85,000. Canarsie's restaurant per capita is only half of that of the nearest top city in Toronto or New York City, which means that Mr. Chen's new restaurant would compete with fewer restaurants in the neighborhood than in any other neighborhood. Furthermore, the cluster analysis results showed that Chinese restaurants are the fifth most common restaurants in the neighborhood, guaranteeing less competition from other Chinese restaurants. Locating in Canarsie will give Mr. Chen's restaurant the best chance of sucess

4.2 Other Top Neighborhoods

As shown in Figure 8 and Table 2, the other top New York City neighborhood second only to Canarsie is Flatlands. It has a similar neighborhood income to Canarsie and 25% less population than Canarsie but with significantly higher restaurant per capita, about double that of Canarsie. For this reason, it only came second after Canarsie.

As shown in Figure 8 and Table 1, the top Toronto neighborhoods to locate a new Chinese restaurant are Don Mills and Bayview Village. Don Mills is the better of the two Toronto neighborhoods because of its population of 47,000, income of over $22,000 and Chinese restaurants being only 8th most popular in the neighborhood. However, its restaurant per capita is almost thrice that of Canarsie at half the population of Canarsie and a third of the neighborhood income of Canarsie. If Mr. Chen, for reasons other than economic, decides to locate his restaurants in Toronto, Don Mills neighborhood would be an ideal location.

4.3 Results Caveats

The model has done in this, and would do, an excellent job of locating an ideal location of locating a restaurant in either of Toronto or New York City given the defined parameters. However, other considerations might play into deciding to locate a restaurant in a particular location. A good example is the price of real estate property which may be more expensive in New York City than in Toronto. If the is an order of magnitude difference in the property price between New York City and Toronto, it might still make more sense to locate the new Chinese restaurant in Toronto depending on what Mr. Chen is willing to invest in his new restaurant. There was no indication from Mr. Chen the amount of money he has to invest in his restaurant venture and there is not easy way to find restaurant real estate prices in New York City and Toronto to incorporate into the model. The model has identified the top neighborhoods in both Toronto and New York City. By doing so, the model has made it easier for Mr. Chen to do further analysis on a narrow set of locations before making the final decision on exactly where to locate his new Chinese restaurant.

Another consideration is the data quality. Best efforts were exerted to obtain the best data possible for the analysis. However, some of the data were from 2007, an update of those data might change the outcome of the analysis. While unlikely, It is quite conceivable that some neighborhoods might have seen an increase in neighborhood income or population since 2007 which might affect the outcome of this analysis.

5. Conclusion

The analysis completed here has done a good job of locating ideal neighborhoods for setting up a Chinese restaurant in Toronto or New York City. If also helped to narrow down the list to a single neighborhood that provides Mr. Chen the best chance of success in his new venture.

Mr. Chen's goal of developing a model that could help him or any investor to determine the ideal neighborhood to locate a restaurant in Toronto or New York City was met. The analysis completed identified the neighborhoods that met Mr. Chen criteria for an ideal neighborhood below:

- Residents with above average income.
- Currently under-saturated by Chinese restaurants.
- Have a sizeable population that could patronize the new restaurant.

Many of the concepts learnt from the program were applied to solve the problem posed by Mr. Chen. Data extraction and scraping were completed, followed by exploratory data analysis. Statistical analysis and machine learning cluster analyses were also deployed to complete the task.

The model developed is not only useful for Mr. Chen, but could also be used by any other entrepreneur trying to locate new restaurant in cities given location and demographic data for those cities.

References

1. New York City neighborhood location data, https://cocl.us/new_york_dataset
2. Toronto neighborhood postal code data, https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
3. Toronto neighborhood geospatial data, https://cocl.us/Geospatial_data
4. New York City neighborhood population data, https://data.cityofnewyork.us/api/views/swpk-hqdp/rows.csv?accessType=DOWNLOAD
5. New York City neighborhood income data, https://ny.curbed.com/2017/8/4/16099252/new-york-neighborhood-affordability
6. Toronto neighborhood population and income data, https://en.wikipedia.org/wiki/Demographics_of_Toronto_neighbourhoods
7. Getting Started with Data Science, Murtaza Haider, IBM Press, 1 edition, Dec 13 2015