**Predicting optimal location for establishing a Chinese restaurant**

1. **Introduction**
   1. **Background**

The one who considers opening a new restaurant should keep in mind several things such as menu, price, and type of cuisine and so on. However, one of the vitally important factors for the success of any restaurant is the location. For example, if you are considering establishing restaurant with specific national cuisine, you have to open it near people with the same nationality, so that demand for your meal will be high in order to maximize the profits.

* 1. **Business Problem**

Throughout this project, optimal places for establishing Chinese restaurant will be found out specifically in Toronto, Canada using different data and data science technique and models.

Independent variables used for finding the best locations for the restaurant are spending power of households, distribution of nationality across neighborhoods of Toronto and all the Asian restaurants located in each neighborhood so that tough competition will be avoided.

* 1. **Interests**

The results of this project will help any business person who has interests in opening Chinese restaurant will be helpful in identifying potential places in Toronto.

1. **Data Acquisition and Cleaning**

**2.1 Data Sources**

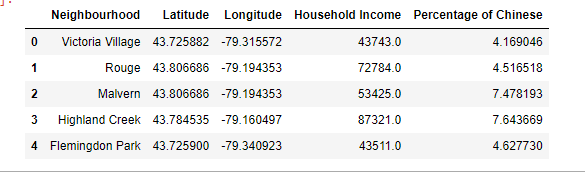
First of all, census data containing information about household income of Toronto’s population, Chinese population living in each neighborhood of Toronto was obtained from Toronto’s Open Data Portal (link: <http://map.toronto.ca/wellbeing>). Secondly, neighborhoods with postal codes and boroughs were scraped from Wikipedia page (link: <https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M>). Thirdly, geographical coordinates of each neighborhood were acquired using the following link: (<https://cocl.us/Geospatial_data>).

Furthermore, all Asian restaurants located in each neighborhood were identified using Foursquare API.

**2.2 Data Cleaning**

All obtained data were merged together into one table. There were a lot of missing values for certain neighborhoods, due to lack of record keeping. Few assumptions were made to achieve the data frame shown below:

* Only the cells that have an assigned borough will be processed; boroughs that were not assigned were ignored.
* Neighborhoods missing more than two census data value were dropped.
* A column that features the percentage of distribution of Chinese population across each neighborhood was calculated by dividing the population of the Chinese demographic by the total population of each neighborhood; the two latter columns were then made redundant and dropped.



1. **Exploratory Data analysis**

**3.1 Folium mapping**

For visualizing Toronto’s map by specifically showing its neighborhoods, Folium library was used.

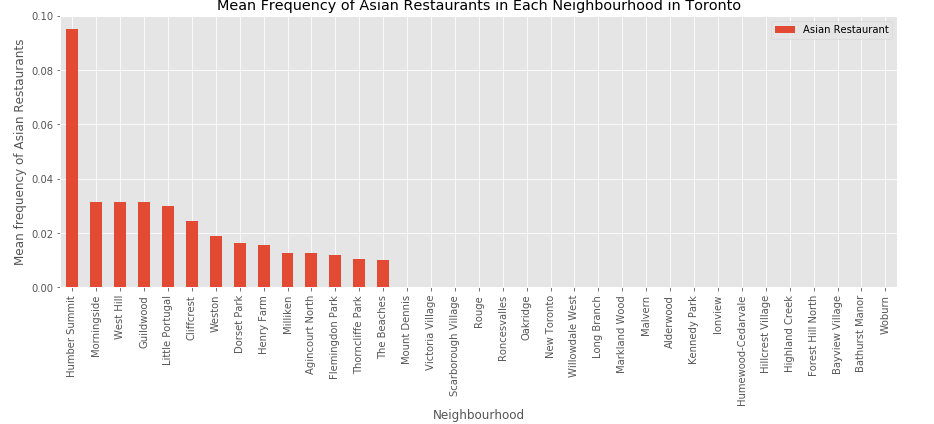


**3.2 Frequency Distribution of Asian Restaurants**

Using the Foursquare API’s explore function, we could return the number of Asian restaurants located in each neighborhood. By calculating the mean respectively, it can give us a better understanding of the frequency of occurrence in each neighborhood. The argument for the use of frequency of Asian restaurants is that I hypothesize that there would be a correlation between the number of Asian restaurants and competition. The higher the number of Asian restaurants in a neighborhood, the stronger the competition. The assumption of our analysis is that the barrier of entry to establish a new restaurant in a competitive market is high as existing Asian restaurants may have the competitive advantage of brand loyalty.

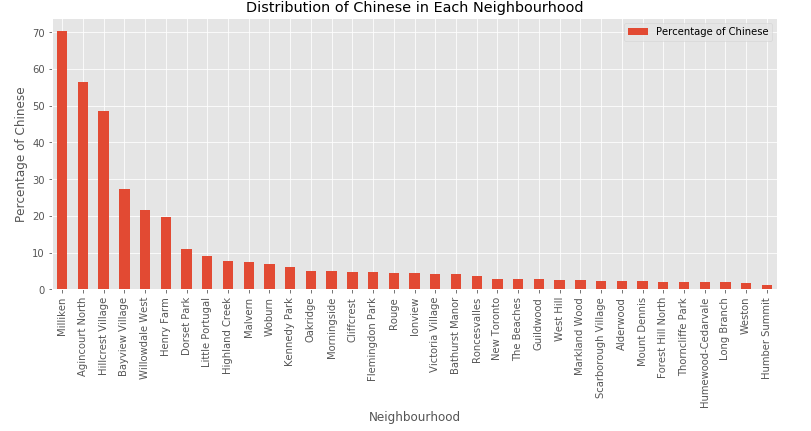
Though, counter-intuitively, the presence of these restaurants may be an indicator of demand for Asian cuisine; the presence of competition may even incentivize innovation to reduce cost and increase productivity.

Hence, it would be sound to establish business operations in a neighborhood that consists of a number of restaurants around the median value.



**3.3 Distribution of the Chinese Demographic**

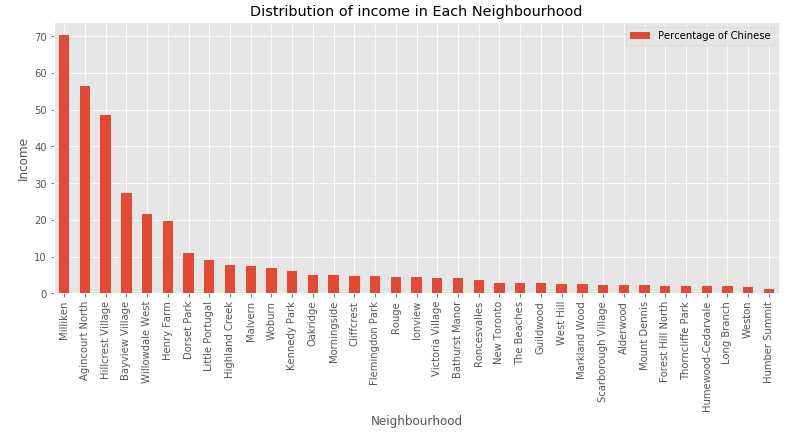
I hypothesize that there would too exist a linear relationship between the population of a specific ethnic group and the demand for its respective cultural cuisine. Hence, it would only be sound for our clients to carry out business operations in neighborhoods that are relatively more densely populated with Chinese.



**3.4 Distribution of Household Income**

As the franchised Asian restaurant could be categorized as casual dining, the target audience is more geared towards the middle class income earners. As can be inferred from the bar chart below, neighborhoods distributed around the mean of household income can readily afford and indulge themselves in the aforementioned Asian cuisine.

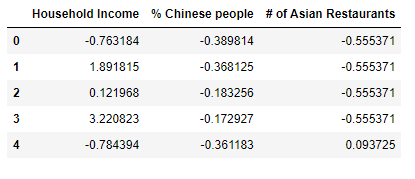
This study also assumes that the cost of living in all neighborhoods in this study is relatively equal in determining their respective spending power.



1. **Predictive Modeling**

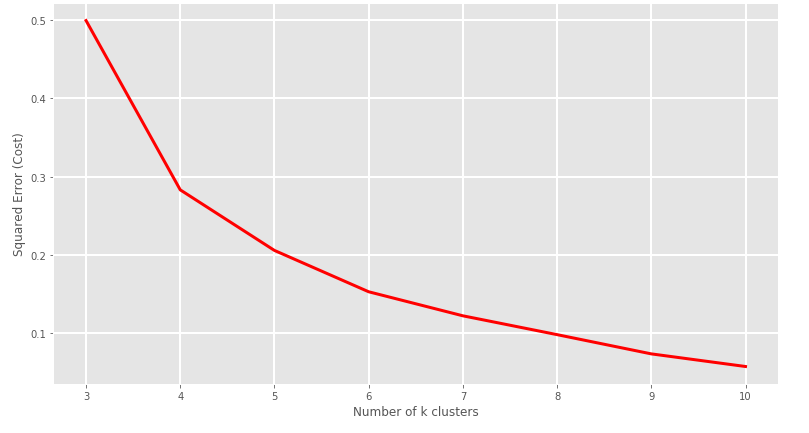
**4.1 Data Pre-processing**

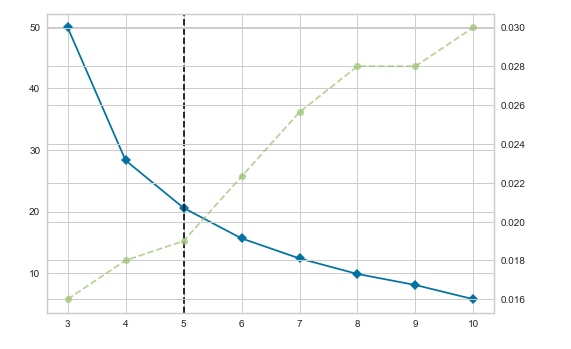
To help mathematical-based algorithms — like our k-Means algorithm in this case — to interpret features with different magnitudes and distributions equally, we will have to normalize our data; as these feature columns are different in scale, we will standardize the values to a common scale. One approach of data normalization is StandardScaler.



**4.2 k-Means Clustering**

Before we fit the feature values into our model, we have to pre-assign the number of clusters the algorithm should label. In order to identify the optimal number clusters to use, a range of 3 to 10 clusters were used, then the squared error calculated respectively were used as metrics of their performances.

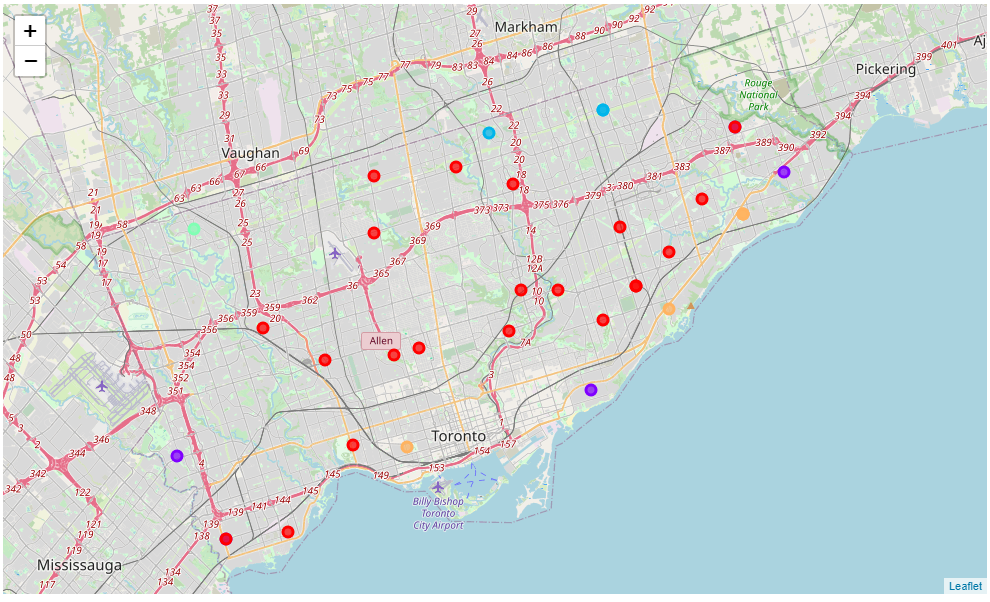




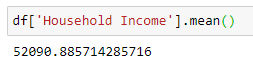
An analysis using K Elbow Visualizer and Squared error for each k value evident shows that k = 5 would the best value.

After identifying the number of clusters, we will fit the standardized feature values into our k-Means algorithm; resulting in 5 clustered neighborhoods of similar characteristics.

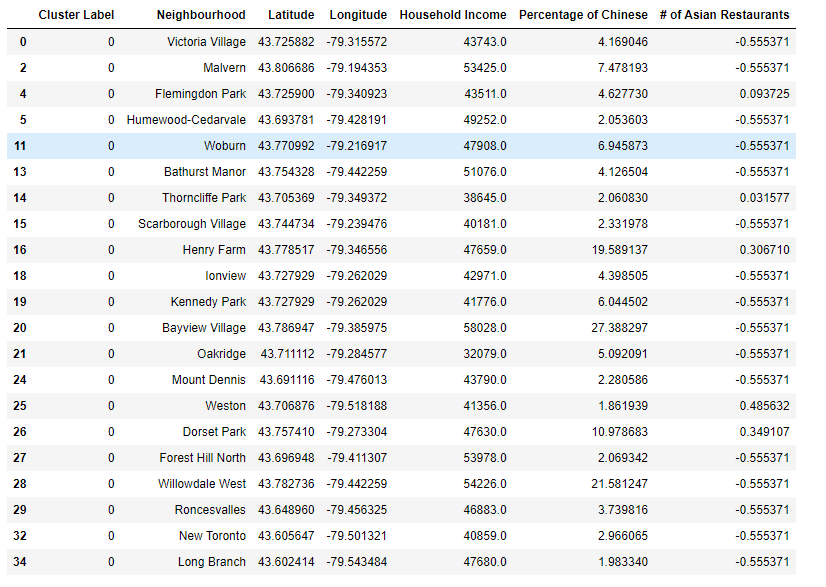
**4.3 Cluster Labeling**



1. **Examining Results**

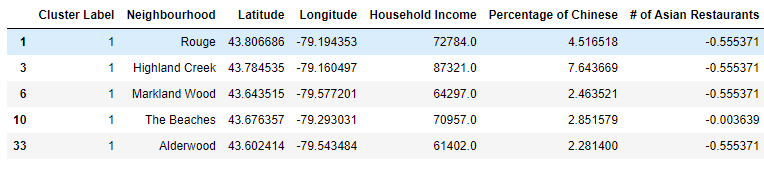


**Cluster 0**



Neighborhoods belonging to cluster 0 shows that distribution of income is mostly below average. However, percentage of Chinese people and density of Asian restaurants are appropriate for establishing a restaurant.

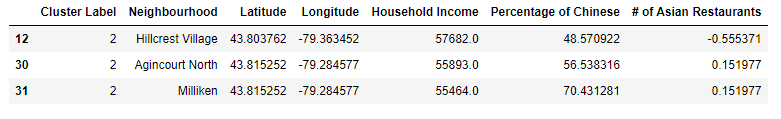
**Cluster 1**



Cluster 1:

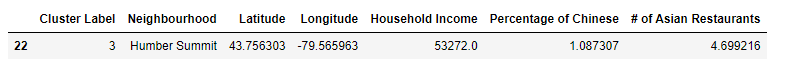
* Household income: above average
* Percentage of Chinese: low
* # of Asian restaurants: low

**Cluster 2**



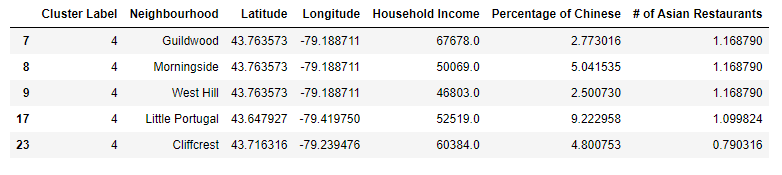
* Household income: average
* Percentage of Chinese: high
* # of Asian restaurants: average

**Cluster 3**



* Household income: average
* Percentage of Chinese: low
* # of Asian restaurants: high

**Cluster 4**



* Household income: above average
* Percentage of Chinese: low
* # of Asian restaurants: high

1. **Conclusion**

In conclusion, after analyzing all the clusters, cluster 2 is the best area to build Chinese restaurant. All the areas in the cluster 0, potentially good places for building Chinese restaurant, because density of Chinese people living in the areas belonging to cluster 2 is high and density of Asian restaurants is low. Moreover, specifically, Hillcrest Village is the best one as there are a lot of Chinese people but no Chinese restaurants.