Targeting Prime Location for a Chinese Restaurant with K-Means Clustering

Shayne Williams

08/08/20

1.Introduction

1.1 Background

This project will use various sources of data to group specific Toronto neighbourhoods into categories identifying their potential for establishing a successful Chinese restaurant. The categories will be defined by aspects of the neighbourhood such as the number of Chinese restaurants, household income, and number of Chinese residents.

1.2 Problem

A client wishes to open an authentic Chinese restaurant in the city of Toronto, which will evidently target a customer base majoritively made up of Chinese residents. The choice of which neighbourhood to establish the restaurant should consider competition, income, and number of Chinese residents.

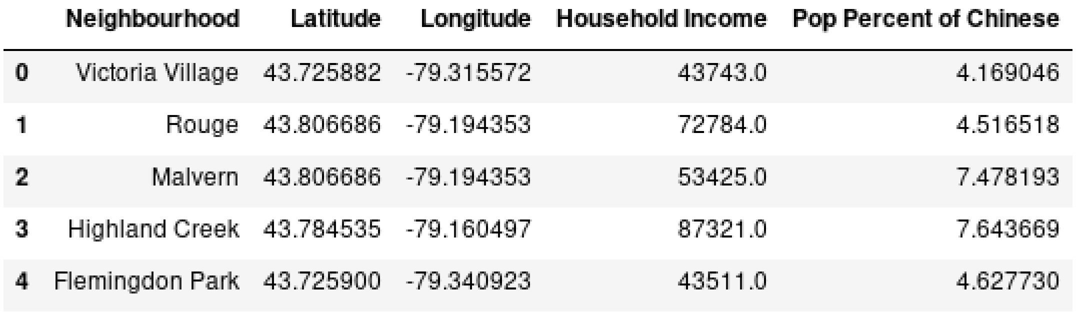
2. Data acquisition and cleaning

2.1 Data sources

The social demographic of Toronto neighbourhoods is obtained from City of Toronto open data source. The geographical coordinates of the neighbourhoods can be imported from here.The list of postal codes for each neighbourhood can be scraped from wikipedia. Foursquare API is used to obtain venue categories for each neighbourhood using its explore function. Registration for a foursquare account is required in order to access this data, which can be done here.

2.2 Data Processing

The data was scraped and imported from various sources and converted into data frames. The data scraped from Wikipedia contained cells with no assigned 'Boroughs' which had to be removed along with any irrelevant columns. The data frames produced from each source were merged with common 'neighbourhood'. The number of Chinese residents in each neighbourhood was converted to the percentage of total neighbourhood population consisting of Chinese residents. Figure 1 displays the data frame produced from the initial data cleaning.



	Neighbourhood	Latitude	Longitude	Household Income	Pop Percent of Chinese
0	Victoria Village	43.725882	-79.315572	43743.0	4.169046
1	Rouge	43.806686	-79.194353	72784.0	4.516518
2	Malvern	43.806686	-79.194353	53425.0	7.478193
3	Highland Creek	43.784535	-79.160497	87321.0	7.643669
4	Flemingdon Park	43.725900	-79.340923	43511.0	4.627730

Figure 1: Dataframe after initial data cleaning.

3. Exploratory Data Analysis

3.1 Folium Mapping

A folium map of the Toronto neighbourhoods was produced by firstly obtaining the geographical coordinates using the python client 'geopy', then importing the folium library to visualise the neighbourhoods.

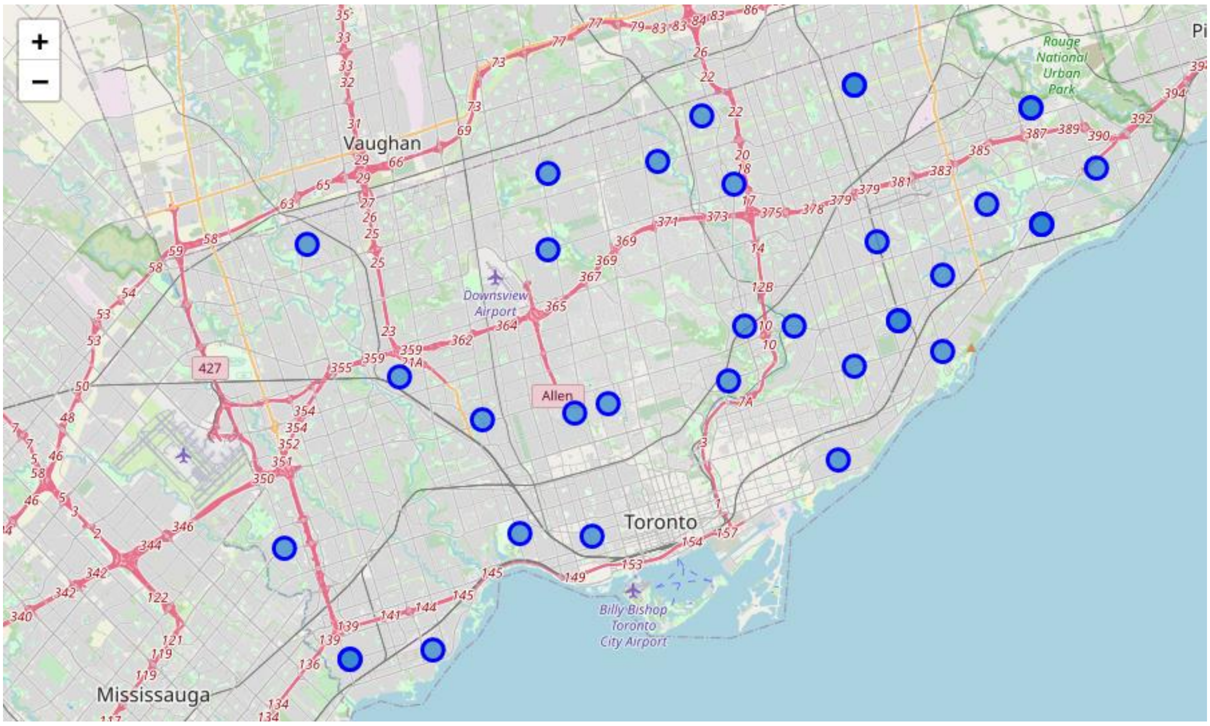


Figure 2: Folium map of Toronto neighbourhoods.

3.2 Analysis of Venue Categories for Each Neighbourhood

Applying the Foursquare API's explore function, the top venues within Toronto were obtained. One hot encoding was then applied to the venue categories to calculate the number of Chinese restaurants per total number of venues for that neighbourhood. The assumption here is that a high ratio of Chinese restaurants would indicate a high level of competition for this particular neighbourhood, however, this could also point to a high demand. The perspective of this ratio being a positive or negative would have to be decided by the investor and the confidence in their business plan to innovate and raise productivity against heavy competition, or alternatively establish and grow the business in a less competitive neighbourhood.

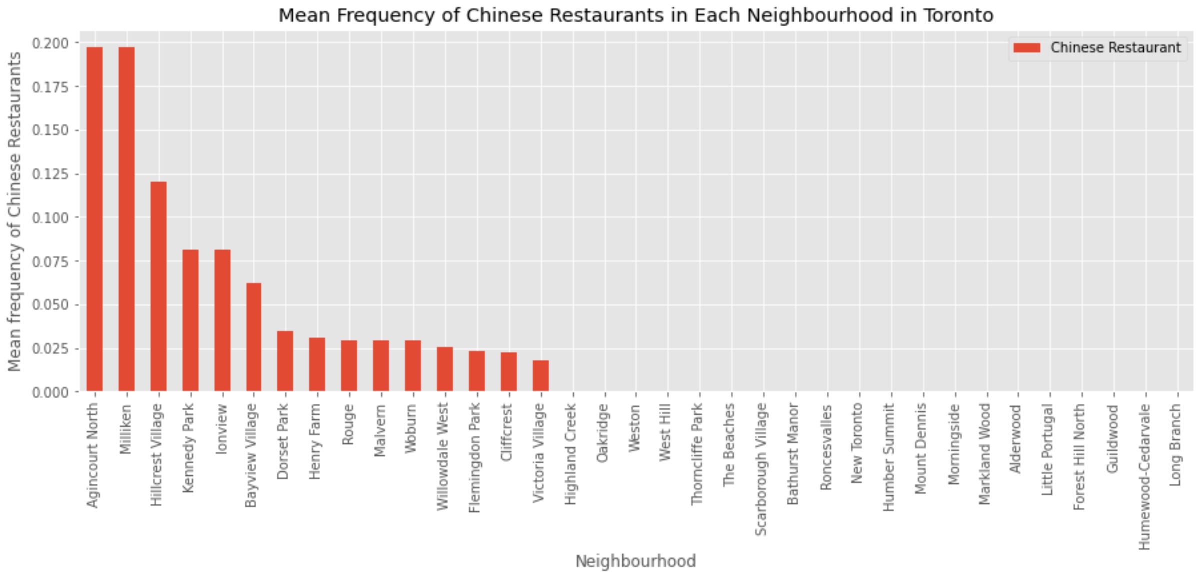
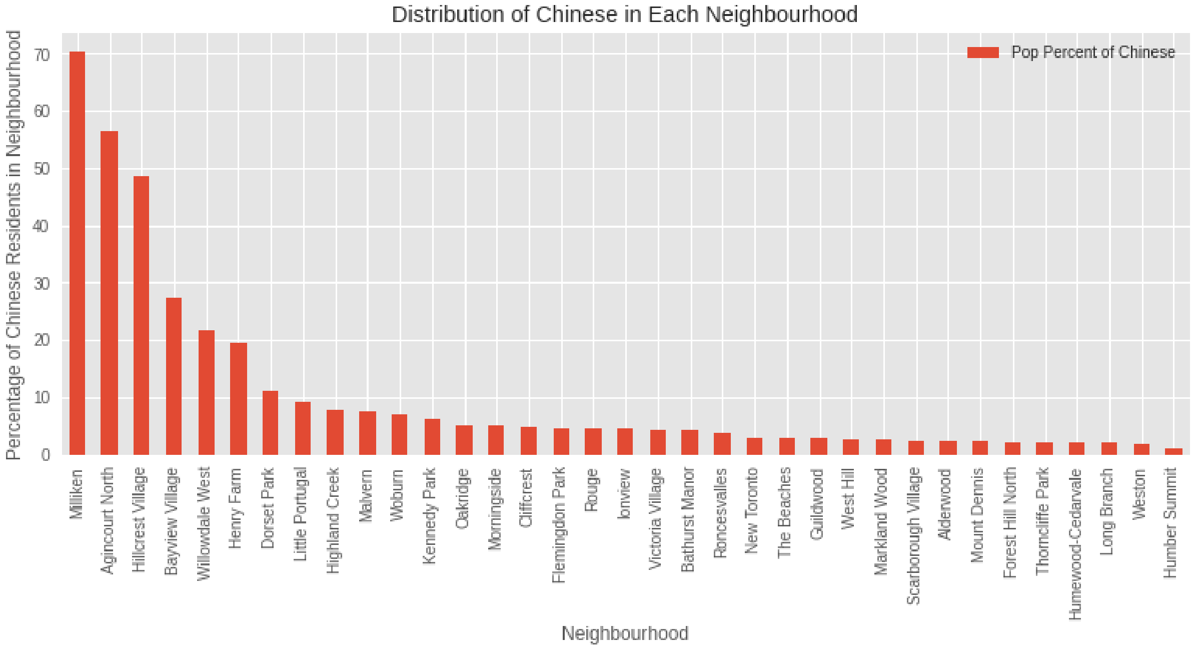


Figure 3: Normalised frequency of Chinese restaurants in each neighbourhood.

3.3 Distribution of Chinese and household income in each neighbourhood

The Number of Chinese Residents and median household income for each neighbourhood were also plotted in a bar chart. The assumption made for the number of Chinese residents is that this should exhibit a linear correlation with demand for authentic Chinese food. The income of the area would have to be considered by investors and whether their business model would permit low prices or would target a predominantly middle class customer base.



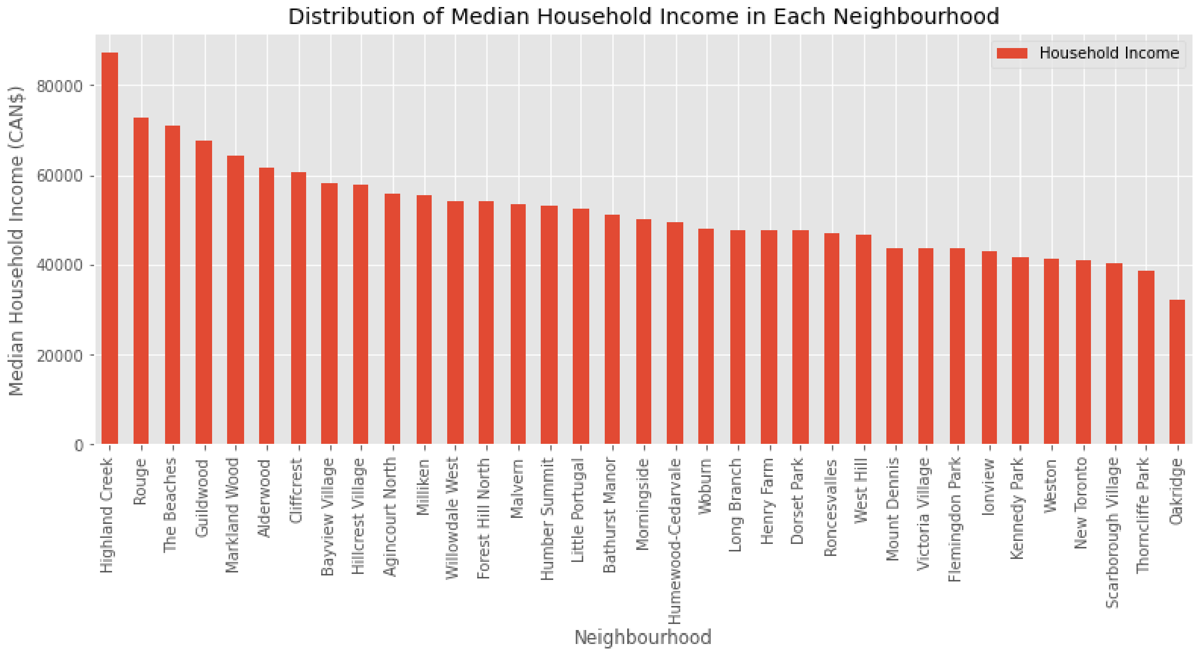


Figure 4 and 5: Bar charts of percentage of Chinese residents (4) and median after-tax household income (5) for each neighbourhood.

4. Predictive modelling

4.1 Elbow Method for predicting K number of clusters

To apply K-means clustering, a final database had to be produced with normalised values for each neighbourhoods income, number of Chinese restaurants, and number of Chinese residents. This was done with the data pre-processing model ‘StandardScalar’. The optimum number of assigned clusters then had to be estimated by means of the ‘elbow method’, performing an analysis with the ‘K-Elbow Visualiser’. This analysis iterates through the given range of clusters, calculating the average of the squared differences of the points to the centre of that respected cluster. The optimum value can be obtained by taking the point at which the plot begins to decrease linearly, given as 5 in this case.

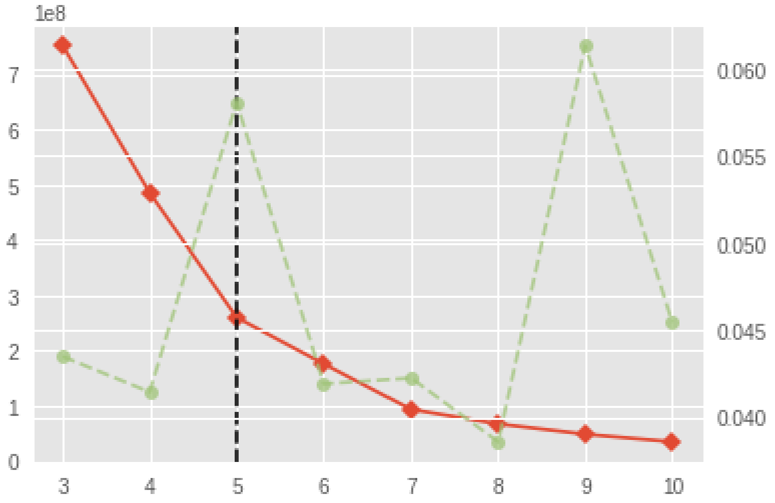


Figure 6 : The ‘elbow method’ which gives the average squared distances for the points from centre of the cluster for a range of clusters.

4.2 K-Means clustering

The K-means algorithm was then applied to the data points. This assigns each point to an initial cluster with a centroid, the average distance between the points within each cluster is calculated and given as the new centroid, with the points then being assigned to the cluster with the nearest centroid, this process is performed iteratively until the centroids become stabilised, assigning the points to the most relevant cluster.

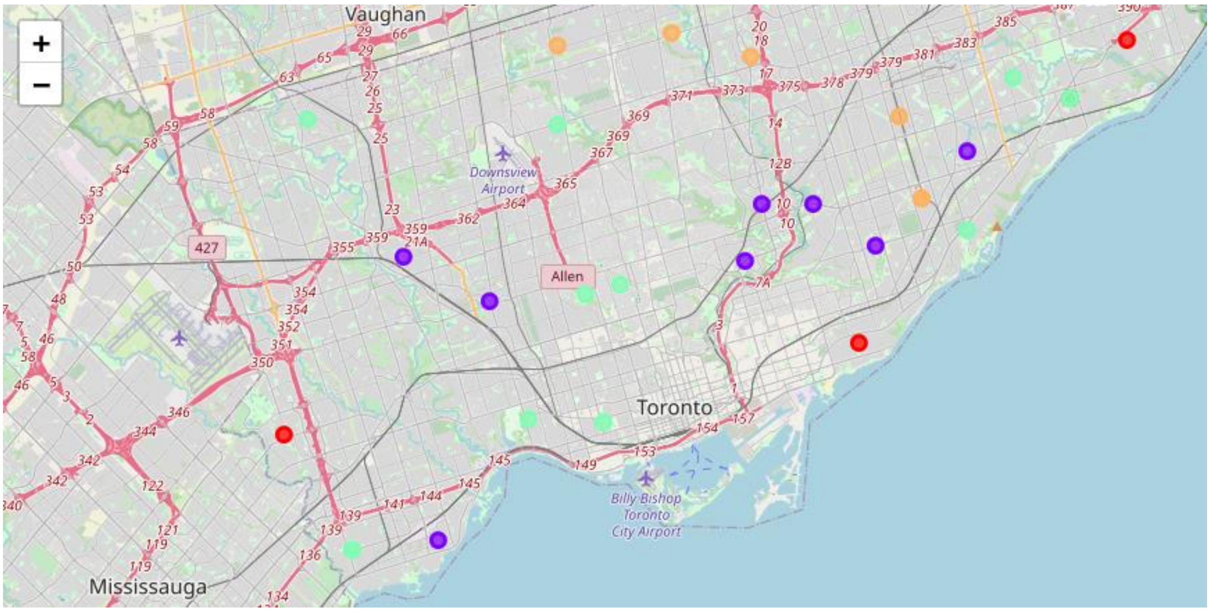


Figure 7: Map of clustered neighbourhoods around Toronto.

5. Results

5.1 Cluster Groups

Group 1: Cluster 0

- High income
- Low Chinese population
- Low number of Chinese restaurants

Group 2: Cluster 1, 3

- Low-mid income
- Low Chinese population
- Low number of Chinese restaurants

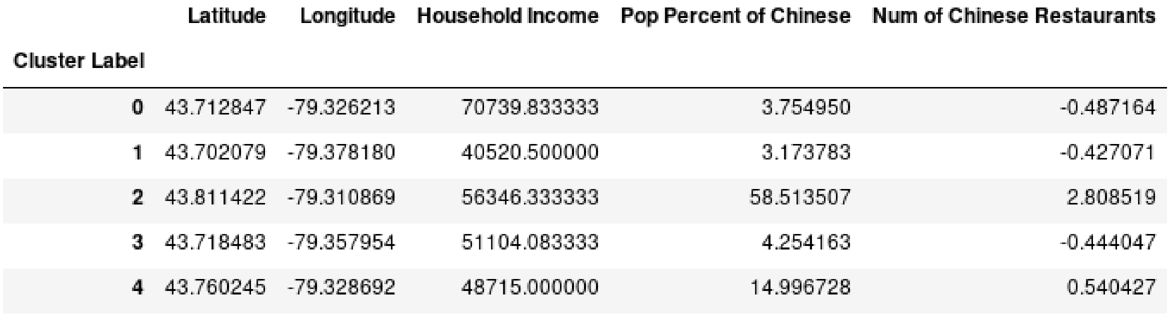
Group 3: Cluster 2

- Medium income
- High Chinese population
- High number of Chinese restaurants

Group 4: Cluster 4

- Medium income

- Medium Chinese population
- Medium number of Chinese restaurants



Cluster Label	Latitude	Longitude	Household Income	Pop Percent of Chinese	Num of Chinese Restaurants
0	43.712847	-79.326213	70739.833333	3.754950	-0.487164
1	43.702079	-79.378180	40520.500000	3.173783	-0.427071
2	43.811422	-79.310869	56346.333333	58.513507	2.808519
3	43.718483	-79.357954	51104.083333	4.254163	-0.444047
4	43.760245	-79.328692	48715.000000	14.996728	0.540427

Figure 6 1 Mean value for the features of each cluster

6. Discussion

Group 1: The high spending power of these neighbourhoods could be attractive for investment, however, the low population of Chinese residents and low number of Chinese restaurants indicate a low demand for authentic Chinese food.

Group 2: This group offers no justification for investment in an authentic Chinese restaurant.

Group 3: The high population of target customers and good spending power would make these neighbourhoods ideal for establishing the restaurant, however, the evident high demand has resulted in a high level of competitiveness which would need to considered in the decision of the investor.

Group 4: This is a promising group which has potential in its medium number of target customers, medium spending power, and medium level of competitiveness.

6.1 Conclusion

Investment should not be considered for groups 1 and 2, leaving the decision to be made on groups 3 or 4. This decision should be taken by the investor on the basis of their business plan, for example if the investor would like to open an already well-established franchise, then group 3 should be highly considered as the business could thrive in a neighbourhood with a high demand, taking customers from the many number of competitors. Group 4 should be chosen if the investor is starting a franchise and has some reservations about taking on serious competition.

6.3 Future Directions

To further this analysis, one could survey the Chinese restaurants within Toronto, to identify which restaurants offered authentic Chinese cuisine as opposed to localised Chinese cuisine. This information was not supplied from the API explore tool, but is required to give an accurate summary of the neighbourhoods.