**Capstone Project Report – Opening a Restaurant in Toronto**

1. **Introduction**

Location, location, location. Similar to looking for a place to live, the location of a new business is the most important decision an entrepreneur can make, especially for a service industry such as a restaurant or hairstylist where people will generally not travel too far from their homes to go. Where would be a good place to start a new business? Where would be a good place to start a new franchise of an existing successful business? Specifically, this assessment is targeted to help entrepreneurs who want to open a new restaurant in the City of Toronto and are looking to answer the question of where best to open the restaurant and what type of restaurant to open.

Census data provides a wealth of information including family sizes, household income, ages, etc. This information can be used to group neighbourhoods and people with similar characteristics. Neighbourhoods with similar characteristics can then be used to compare what types of restaurants are prevalent in each area and what type of restaurants may be lacking representing a good opportunity for a new business. For example, if five out of six neighbourhoods that have a high percentage of young singles with high incomes shows a prevalence of Japanese restaurants, then a Japanese restaurant might be a good business to open in the sixth neighbourhood.

1. **Data**
   1. **Census Data**

The Government of Canada's last census data was undertaken in 2016 and has been provided on the Statistics Canada website. The data can be obtained broken down by forward sortation areas (FSA), which is the first 3 characters of a Canada postal code, and thus allows for the grouping of information by neighbourhoods. However, all the pieces of information for each FSA is stored as name/value pairs and need to be transformed into tables that can be used for neighbourhood analysis. Also, certain FSA are not residential areas so they will need to be removed.

From the Canada-wide census data, a subset of this data from Toronto FSA's can be obtained. There are over 2000 features per FSA, and a subset, specifically age groups, family sizes, and household income will be used to group similar neighbourhoods. Because each FSA in Toronto will have different population sizes, a percentage distribution will give us a better measure of similarity amongst neighbourhoods.

**2.2 Restaurant (Venue) Data**

Foursquare data will be used to obtain common venues for grouped Toronto neighbourhoods. This can then be used to analyze whether similar neighbourhoods have a lack of venue (restaurant) and hence a potential business opportunity.

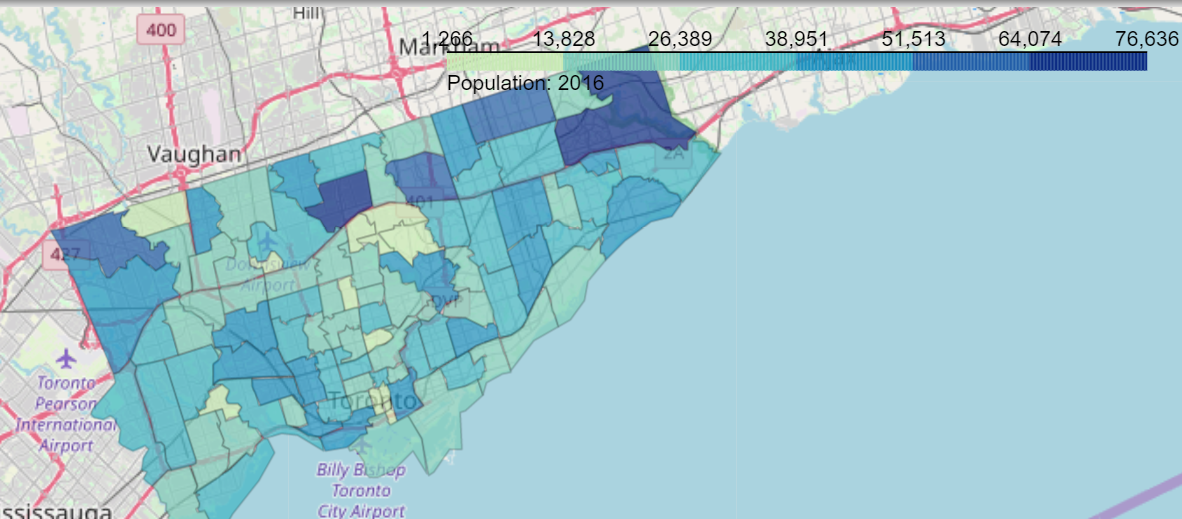
**2.3 Supporting Data**

Geo coordinate information for the forward sortation areas of Toronto is also used to be able to visualize/map the results of the analysis.

A geojson file was also obtained to also help visualize the FSA areas of Toronto.

1. **Methodology**
   1. **Population Distribution**

We can see from the data, Toronto consists of 96 FSAs (postal codes) not including the FSA’s that are not residential areas. Looking at the population, the average size of each postal area is around 28,500 people, but range quite considerable from 2,005 to 75,897. This is why getting percentage distributions would be a better comparison of neighbourhoods rather than absolute numbers.



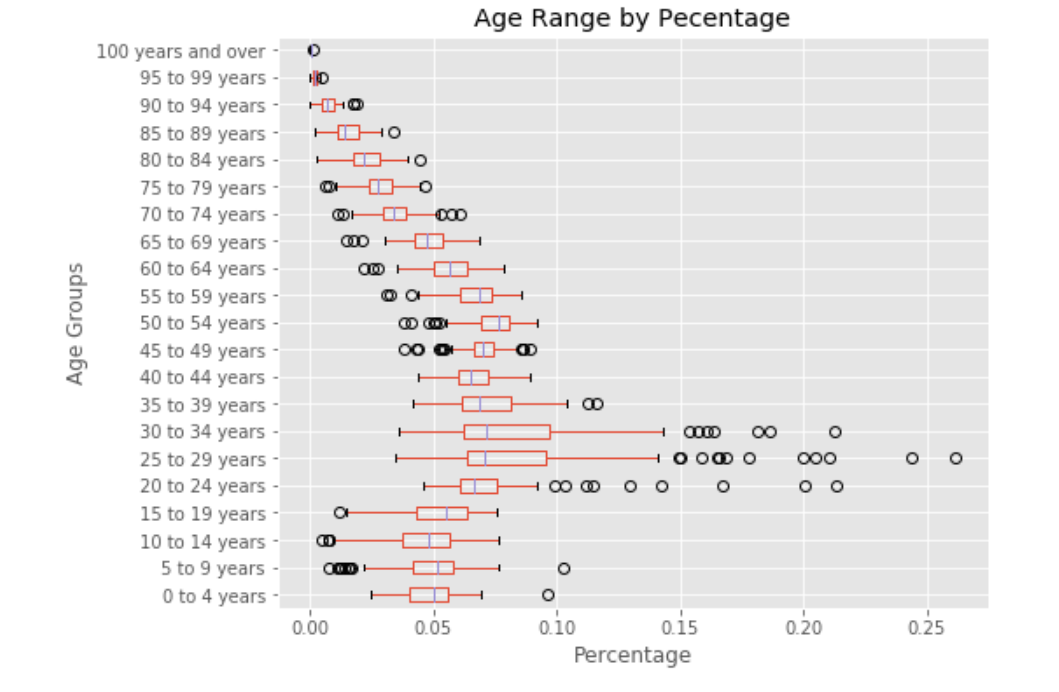
* 1. **Selecting Census Data to Cluster Neighbourhoods**

As described above, the features of age distribution, household income (after-tax) distribution, and family types will be used to cluster neighbourhoods. The census data contains several income categories such as income before tax, income of individuals; however, after tax income of a household would best represent the disposable income available for discretionary spending such as for meals.

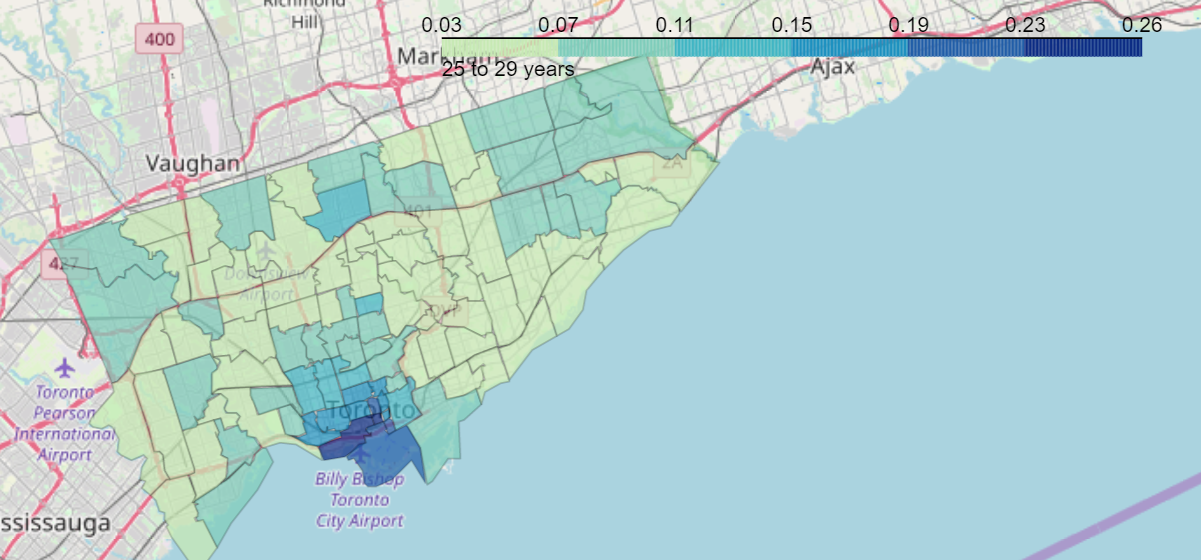
Looking at each census data individually:

* 1. **Age Distribution**

We can see that the age range of each FSA sits within a tight range except for those between 20 to 39 years old where the range is wider and many outliers.

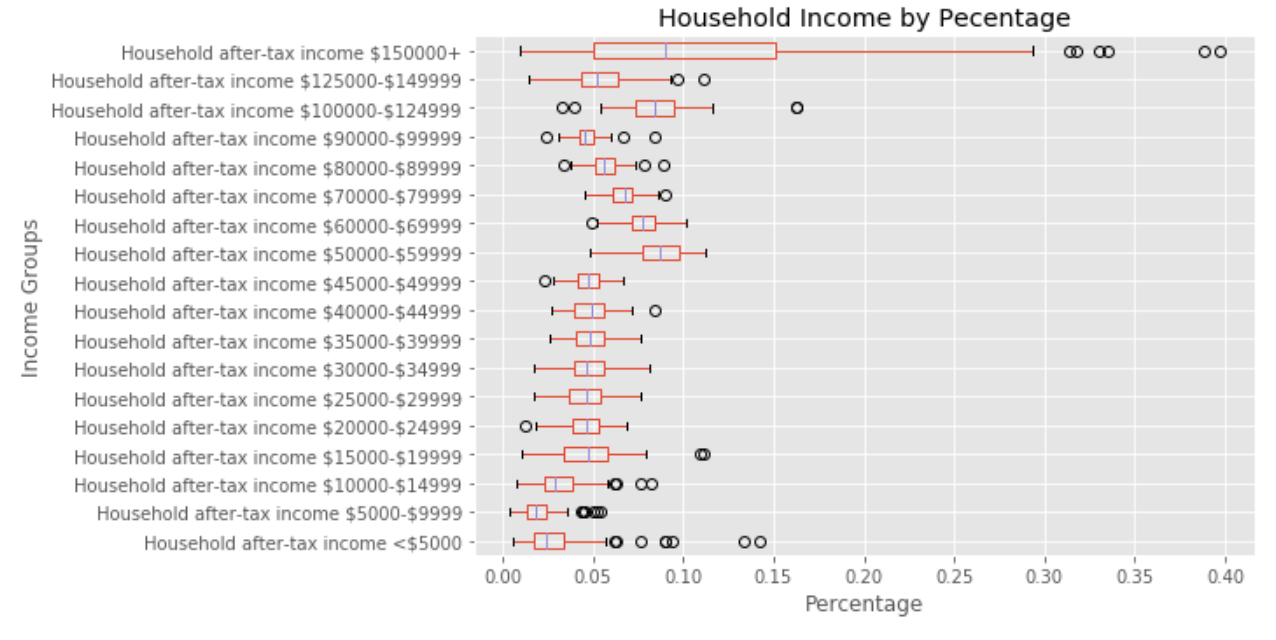


Looking specifically at the percentage that are 25 to 29 years old, there is a higher concentration in FSA/postal codes in the downtown area close to the waterfront.

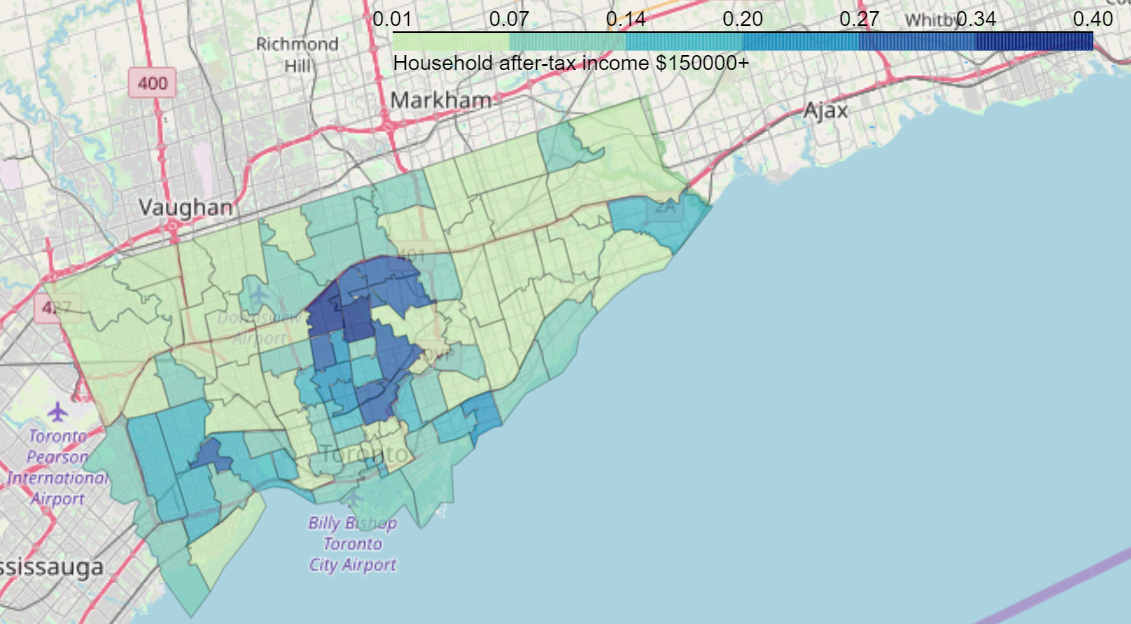


* 1. **Household (After-Tax) Income**

The household income distribution varies quite considerable above $150,000.

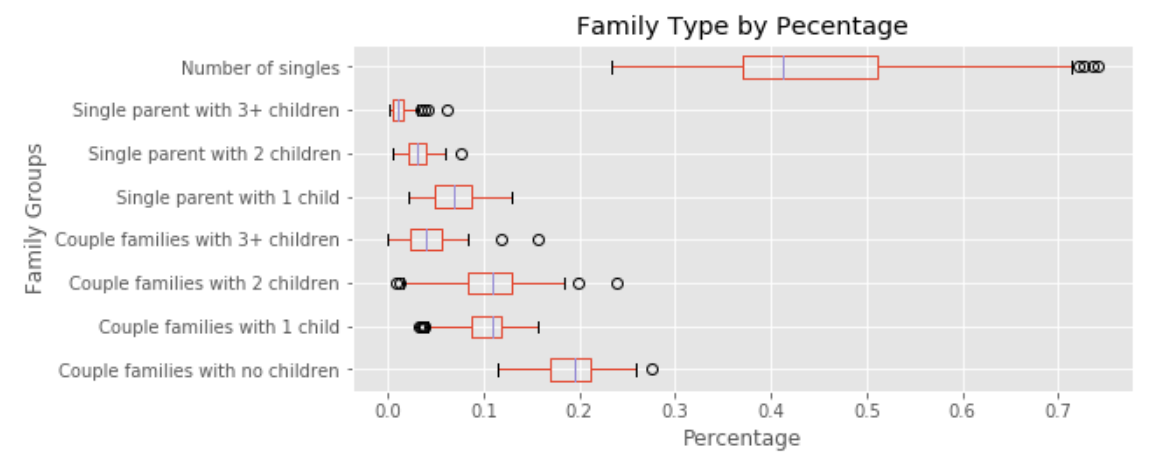
****

From the choropleth map, we can see that there is a high concentration of FSA’s with a high percentage whose income is above $150,000 in the centre of Toronto.

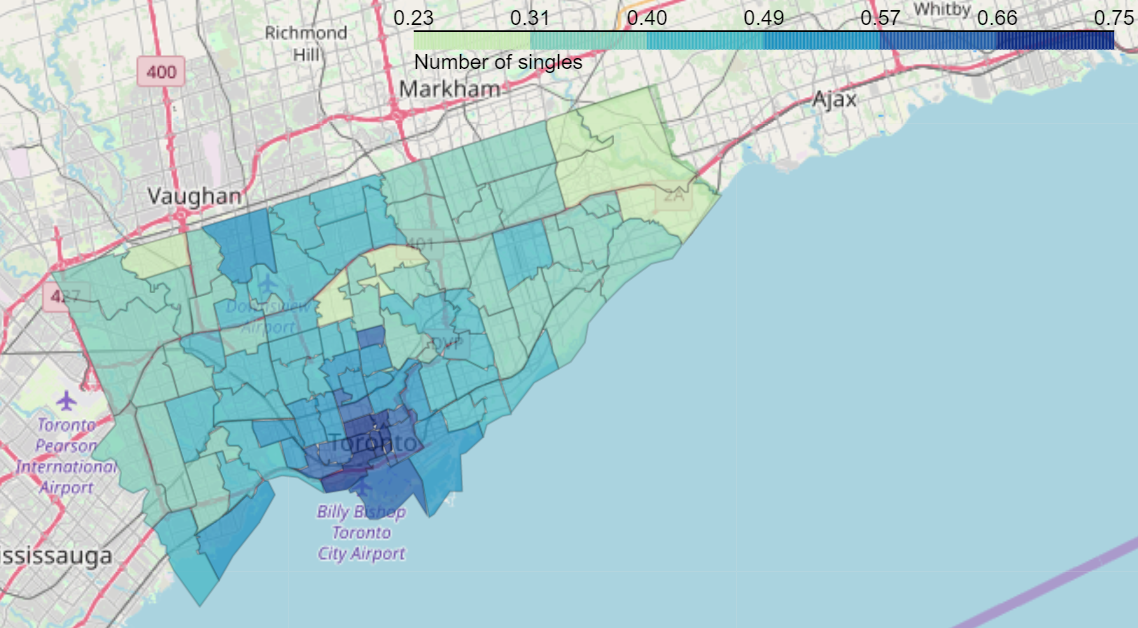
****

* 1. **Family Type Distribution**

Singles make up the most common family type.



The concentration of FSA’s with a high percentage of single “families” are located around the downtown core near the lake.

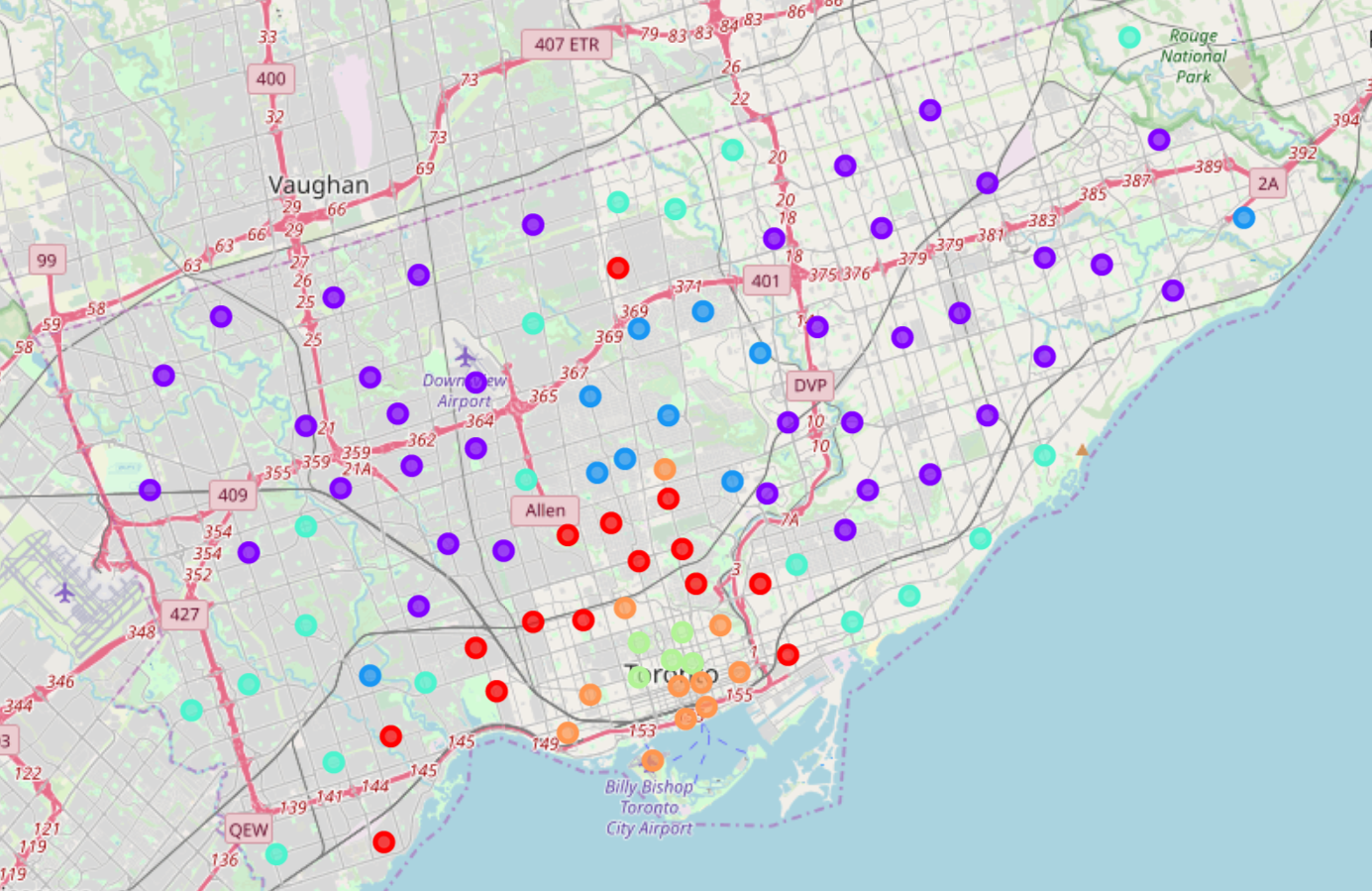


* 1. **Neighbourhood Clustering Using kmeans**

The three datasets of the age distribution, household income distribution, and family type distribution (all in percentages) were combined into one dataset and kmeans was used to cluster the FSA’s into six clusters.

1. **Results & Discussion**

The following map shows the six clusters.



Cluster 0 (red): high income and high ratio of singles and families without children

Cluster 1 (purple): balanced distribution across income, family type and age groups

Cluster 2 (blue): very high income and high ratio of families with children

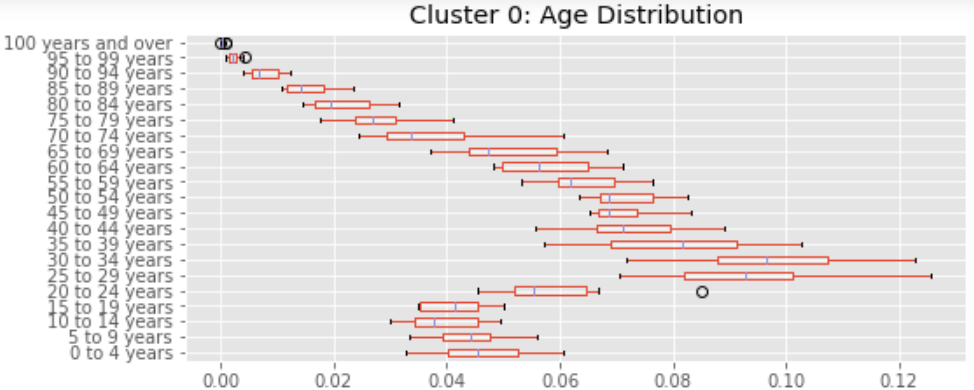
Cluster 3 (cyan): very similar to Cluster 1 (balanced distribution)

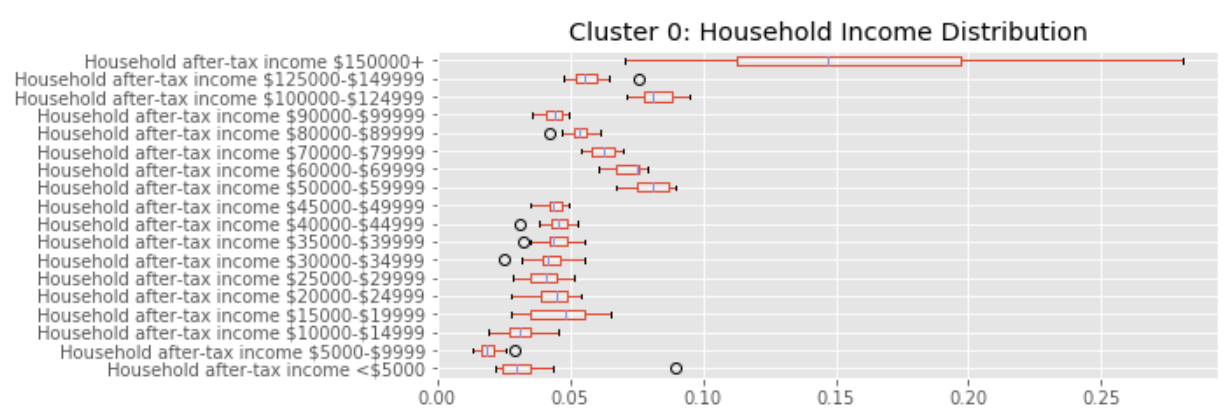
Cluster 4 (green): high ratio of young, single, adults with low income (predominantly a university area)

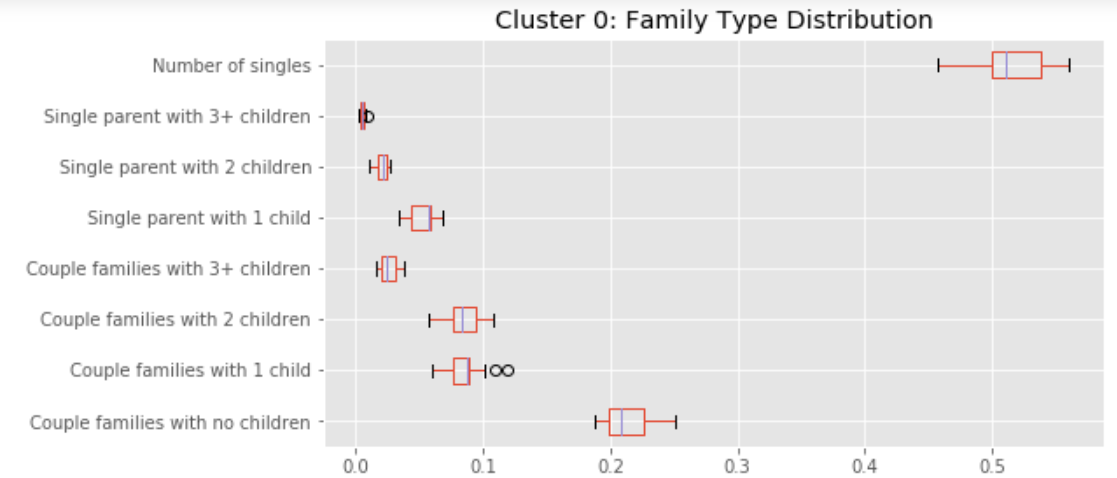
Cluster 5 (orange): high ratio of young adults and families without children but with a regular income distribution

* 1. **Cluster 0 (red)**

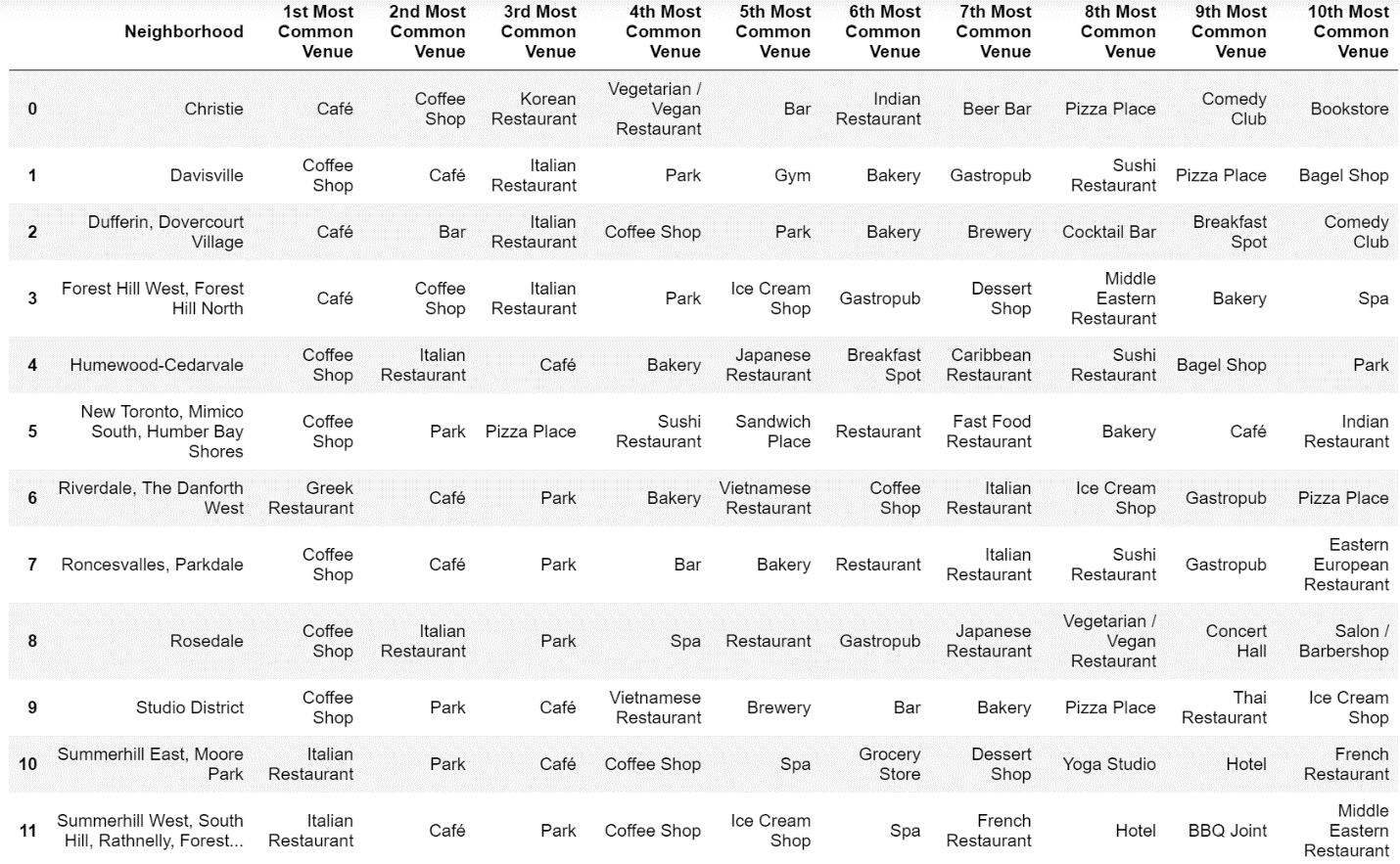
Cluster 0 is geographically defined immediately surrounding the downtown core of Toronto and consists of 15 neighbourhoods. From the bar charts below, this cluster’s age distribution is fairly typical and has a high percentage of households whose income is greater than $150,000. Singles and couples without children make up the family types of this area.







From the Four Square data, Italian restaurants appear in the top 10 venues in 11 out of the 15 neighbourhoods, and even in the top 3 in 8 neighbourhoods. However, Mimico South, Christie, Willowdale South, and Studio District do not have any in top 10. It is an opportunity to open an Italian restaurant in these neighbourhoods.

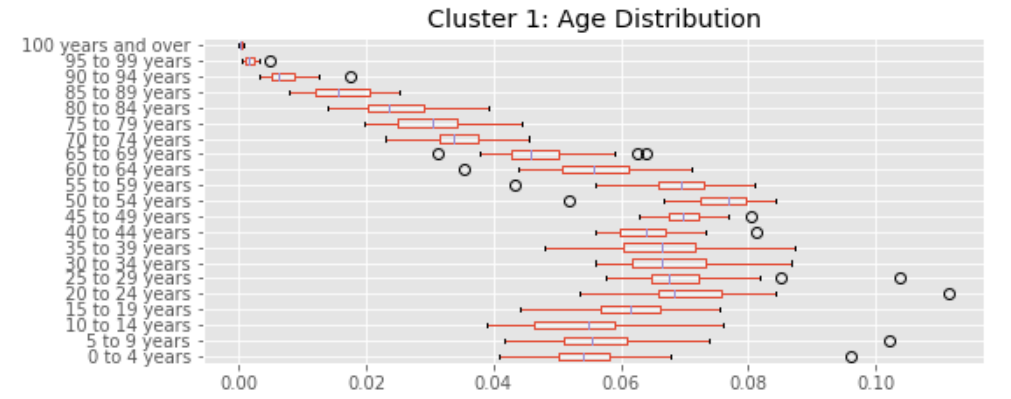


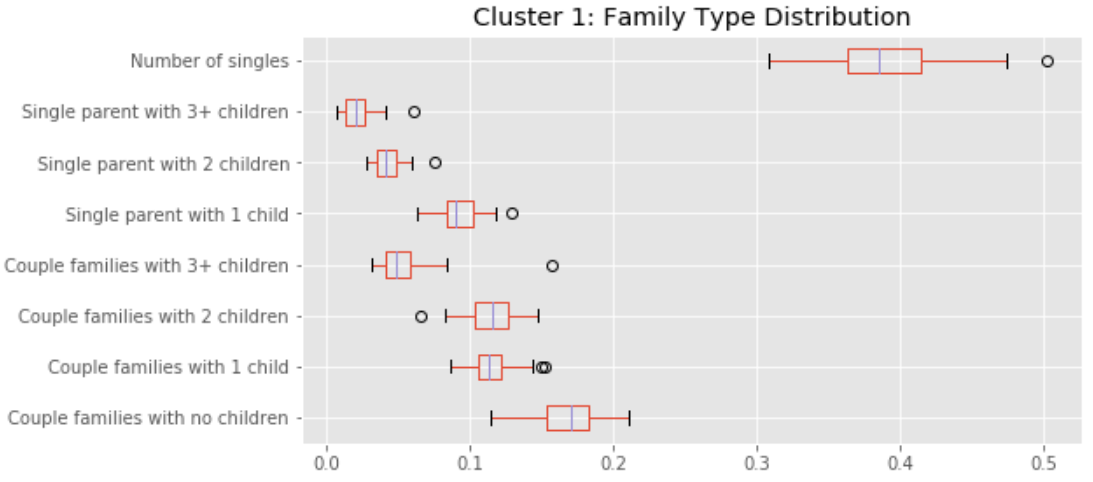


* 1. **Cluster 1 (purple)**

Cluster 1 encompasses the entire area of Toronto outside of the core and consists of 37 neighbourhoods which is the biggest cluster by far. The age, household income, and family type distributions all have typical patterns as compared to the other clusters.

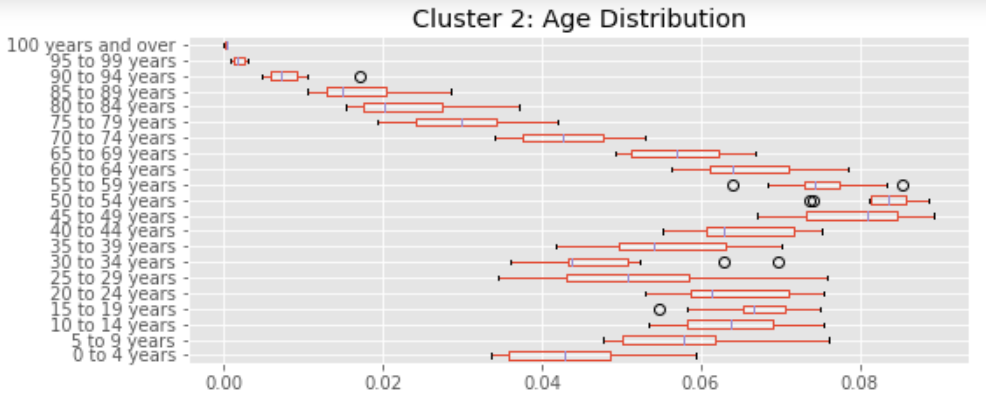
With such a large cluster size, the top 10 venues vary quite widely and there is no obvious pattern of restaurants to make a recommendation. To better recommend restaurants, more census data, such as the ethnic backgrounds, could be incorporated to further segment these areas.

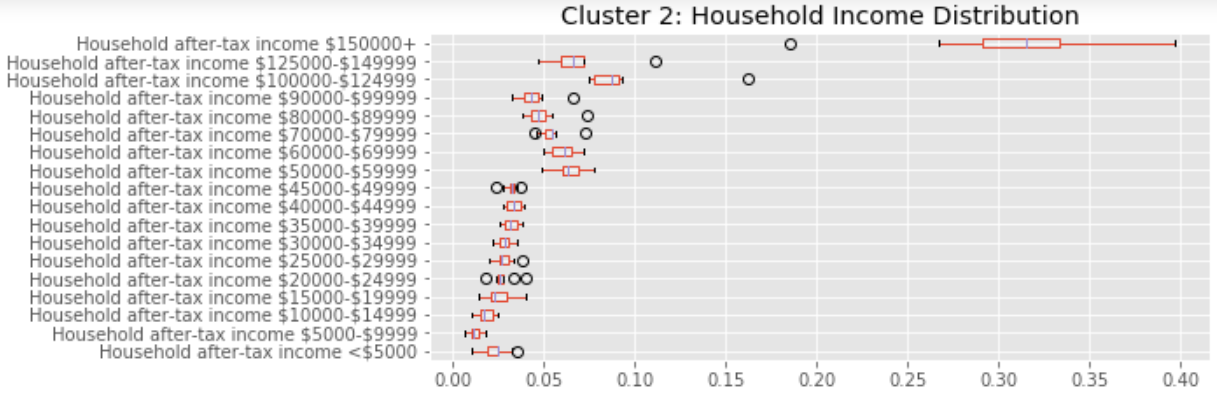


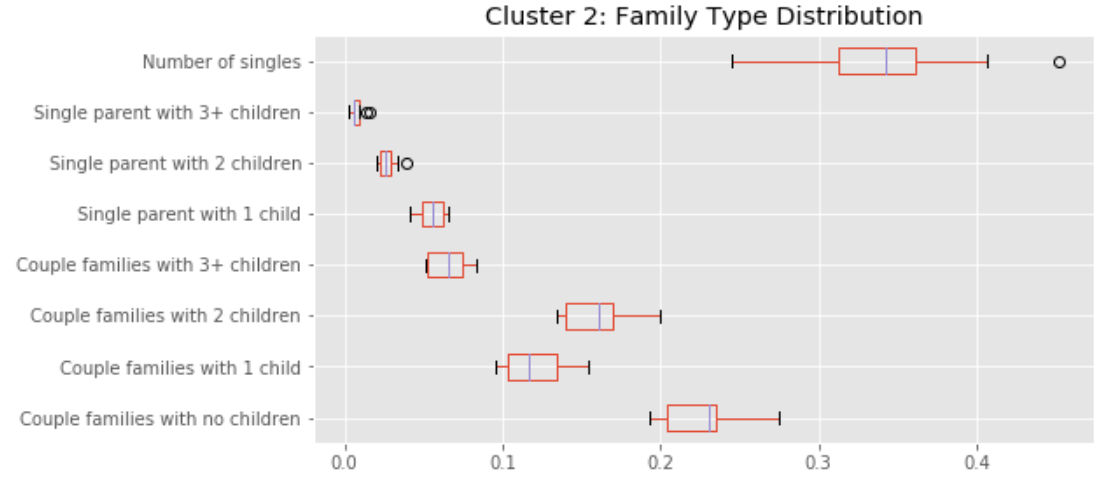


* 1. **Cluster 2 (blue)**

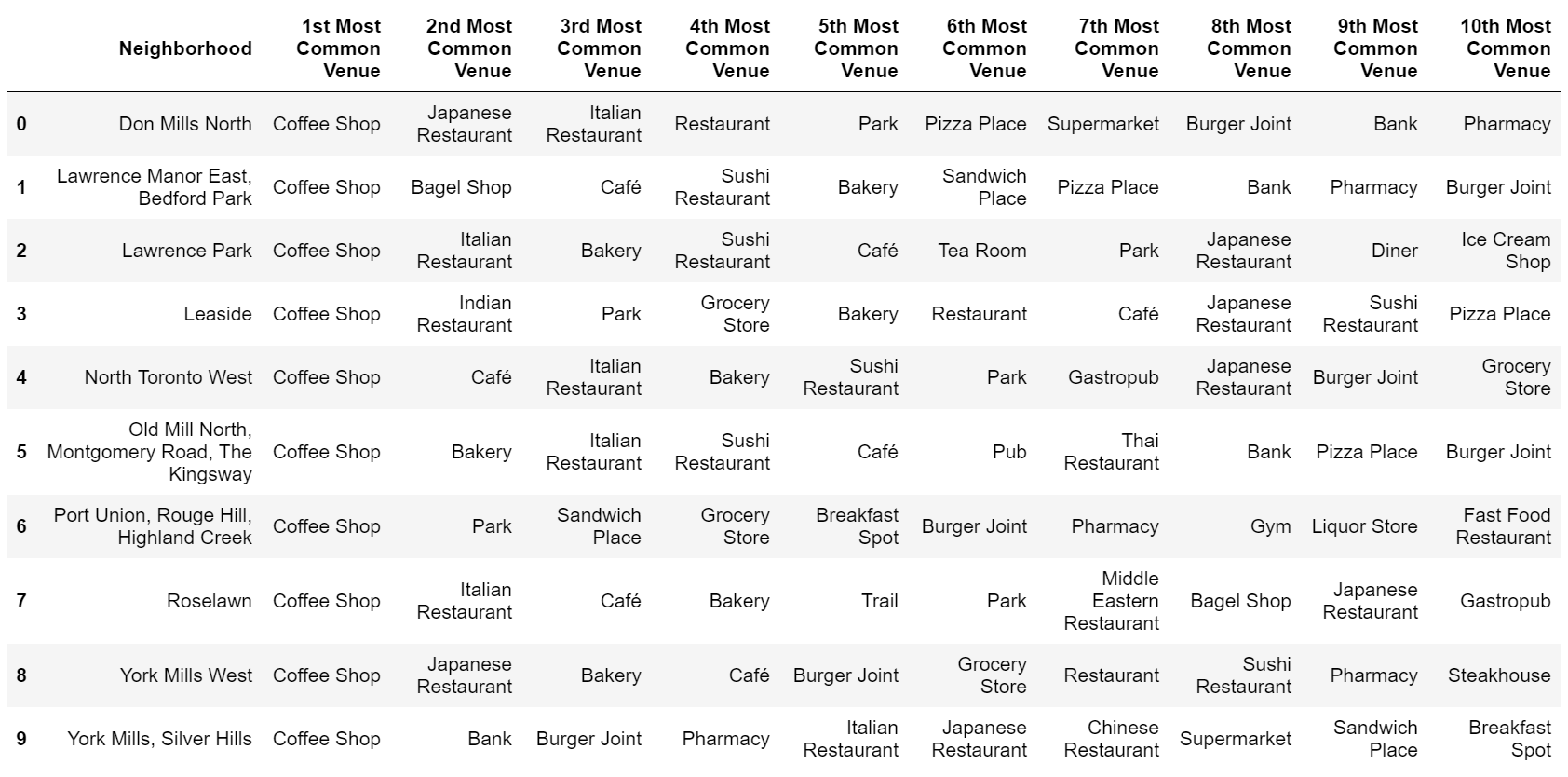
Cluster 2 is geographically in the centre of Toronto and consists of 10 neighbourhoods. From the bar charts below, the number of households with income over $150,000 is very high compared to the rest of the clusters. There is also a relatively higher percentage of families with children.







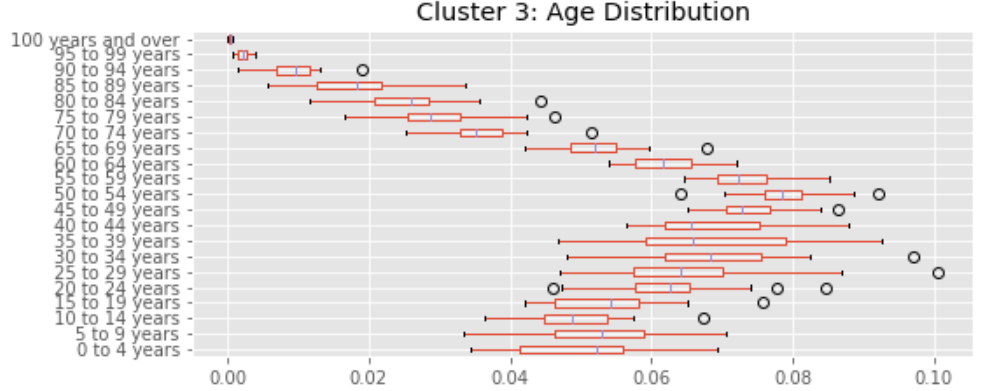
From the Four Square data around Cluster 1, Japanese or Sushi restaurants appear to be a common venue for these neighbourhoods. Nine out of 10 neighbourhoods have a Japanese/Sushi restaurant. Rouge Hill would be a recommended neighbourhood to open a Japanese/Sushi restaurant.

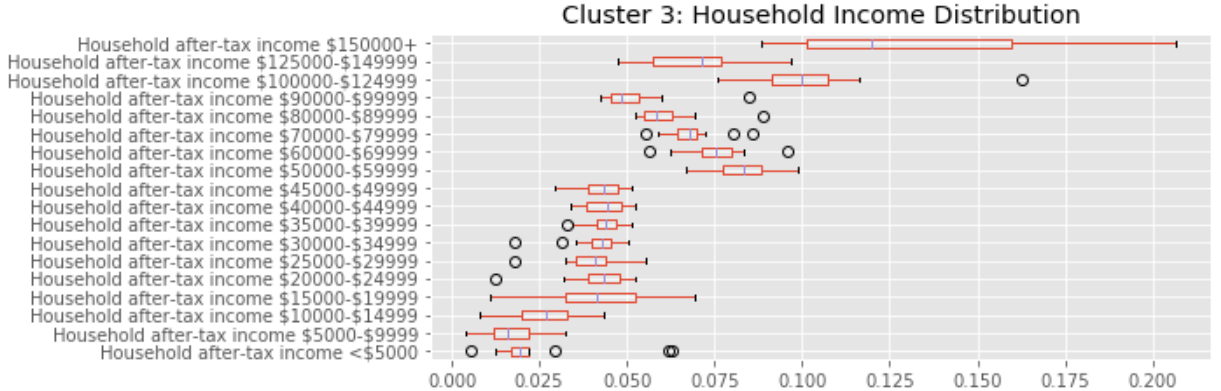


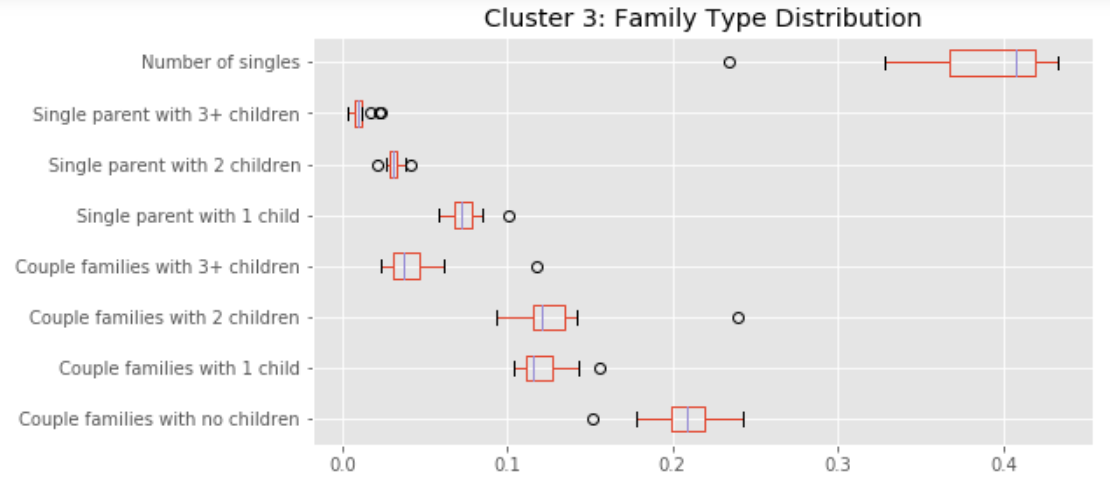
* 1. **Cluster 3 (cyan)**

Cluster 3 and Cluster 1 share similar characteristics. Cluster 3 is geographically surrounding Toronto outside the central core consisting of 18 neighbourhoods. Nothing really stands out when comparing the distributions of age, household income and family types with the other clusters.

Similar to Cluster 1, the top venues for these neighbourhoods vary widely. Again, To better recommend restaurants in this cluster, more census data, such as the ethnic backgrounds, could be incorporated to further segment.

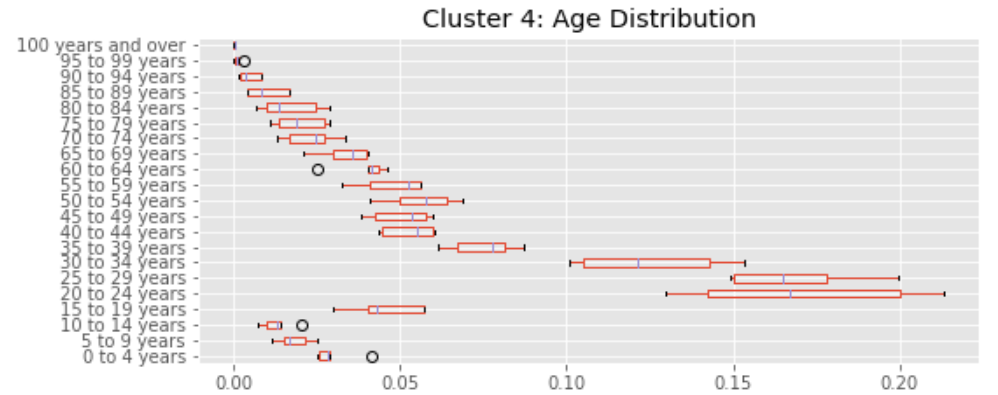


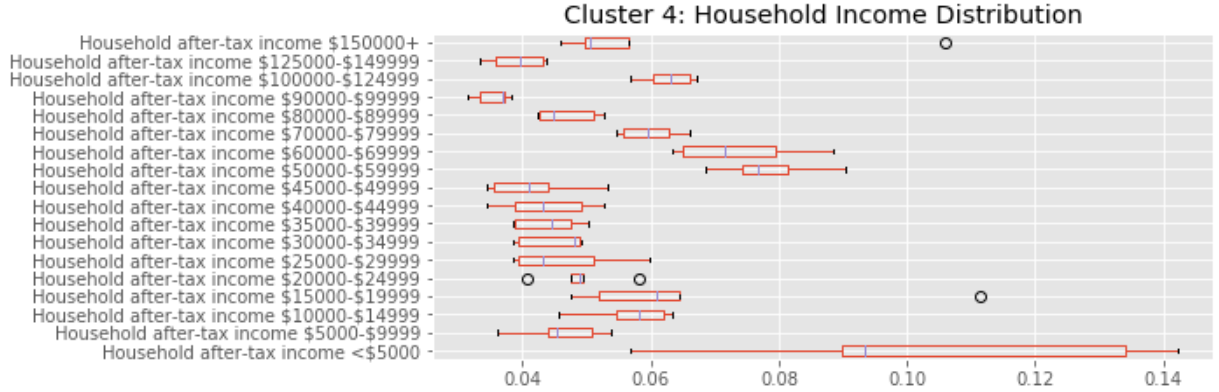


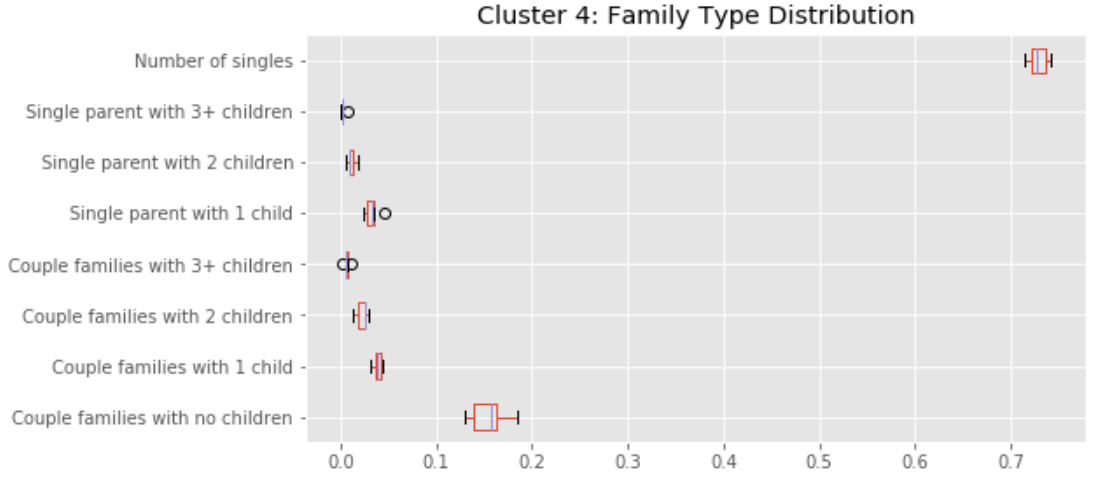


* 1. **Cluster 4 (green)**

Cluster 4 only consists of 5 neighbourhoods that are located in the downtown core (mostly surrounding the University of Toronto campus). The age distribution is heavily skewed to 20 – 34 years old and there is a relatively high percentage of households whose income is less than $5000. There are also a small percentage of families with children.





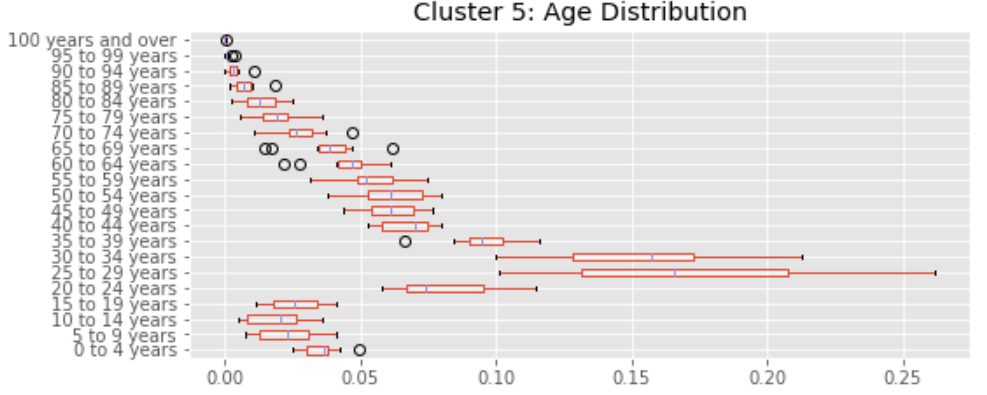


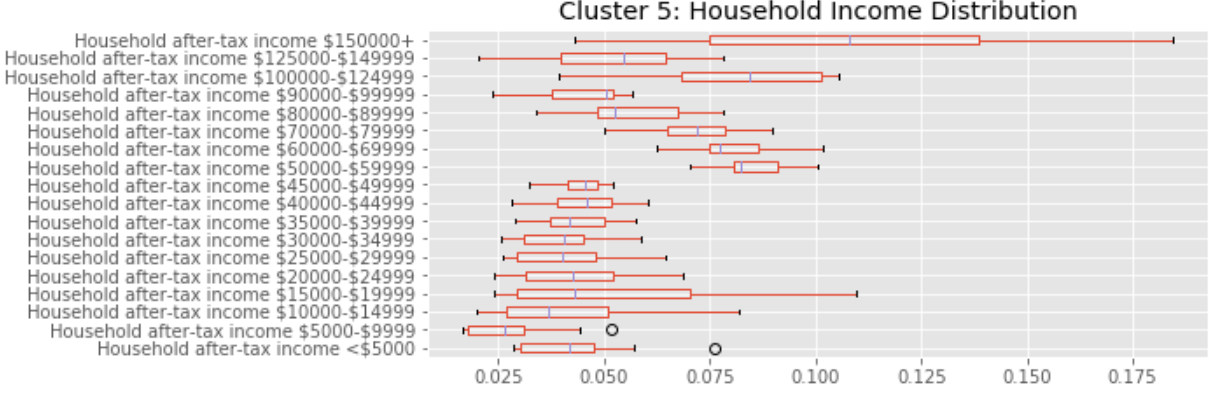
The Four Square data for Cluster 4 shows that 4 out of the 5 neighbourhoods have a Vegan/Vegetarian restaurant although not high on the list of top 10 venues. However, a Vegan/Vegetarian restaurant in the Garden District/Ryerson would be a recommendation.

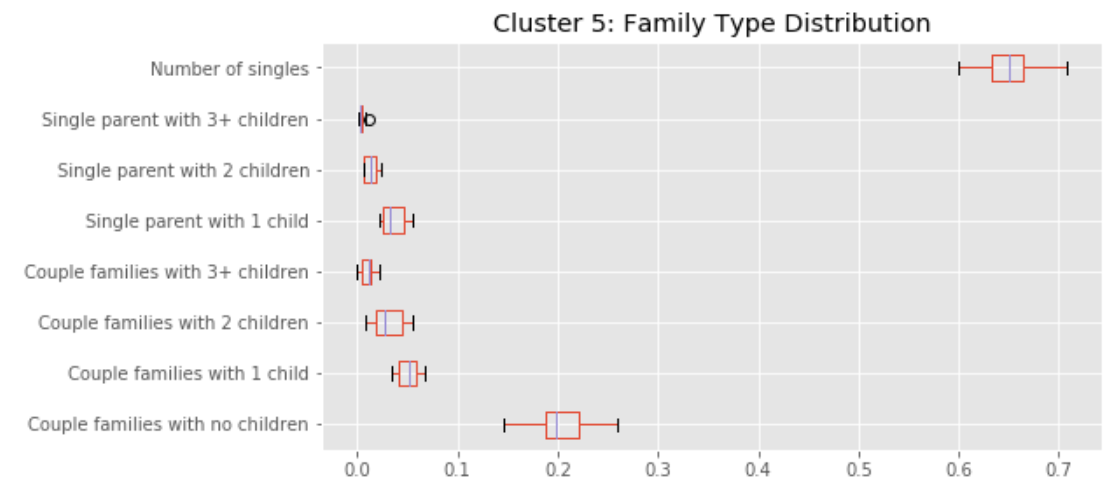


* 1. **Cluster 5 (orange)**

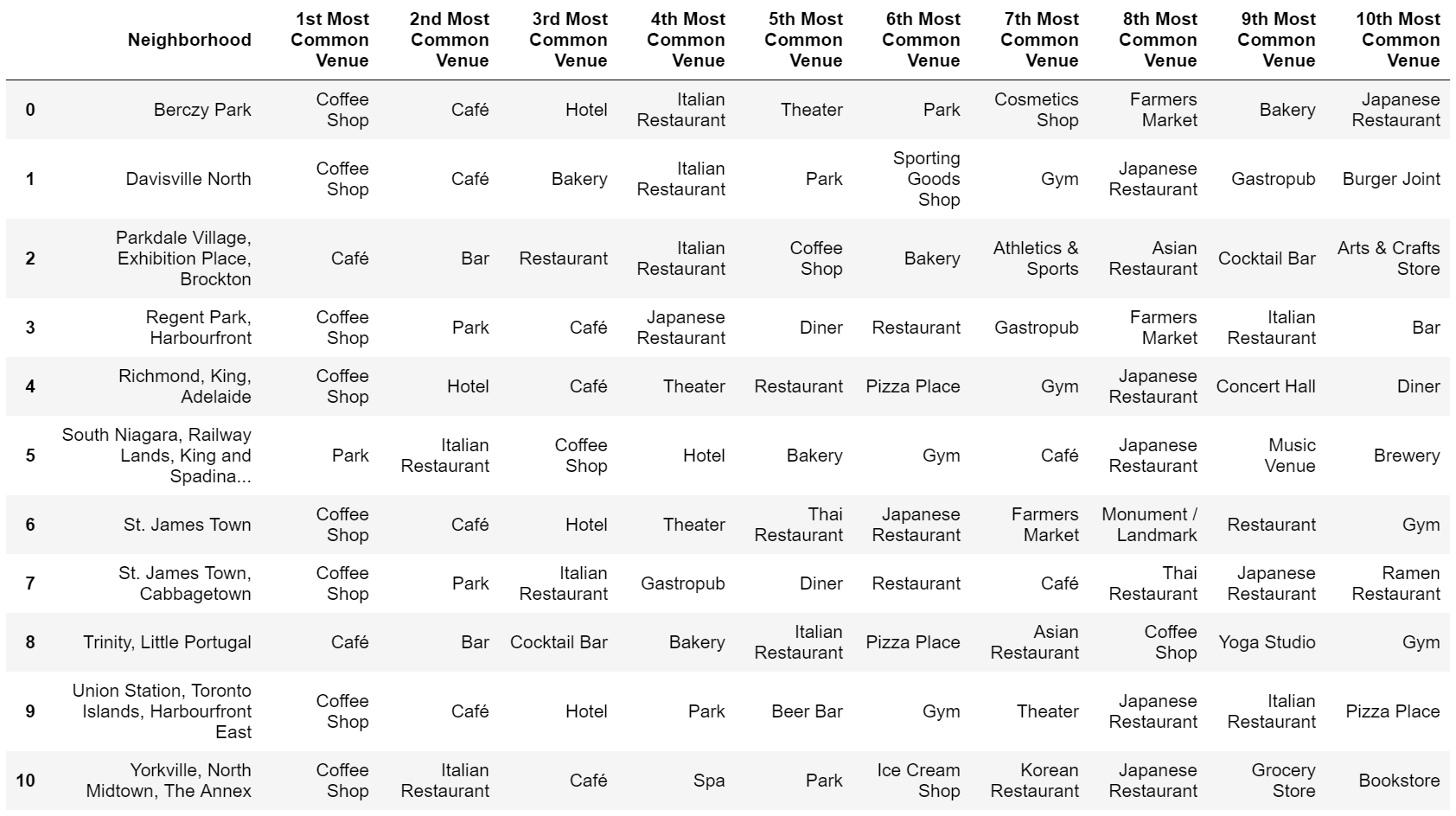
Cluster 5 consists of 11 neighbourhoods that are located mostly in the downtown core. Cluster 5 is similar to Cluster 4 as the age distribution is skewed to 20 – 34 year olds and there is a small percentage of families with children. However, the household income is fairly typical.







The Four Square data from the neighbourhoods of Cluster 4 show Italian restaurants as a common venue in 9 out of the 11 neighbourhoods. Italian restaurants would be recommended to be opened in the Richmond/King/Adelaide and St. James Town neighbourhoods.



1. **Conclusions**

Leveraging Canada census data, Toronto neighbourhoods were clustered by similarity using age, income, and family type distribution. Four Square data was obtained for each of the clusters to compare which restaurants were common amongst most of neighbours and were also not common in the other neighbourhoods of the cluster. Recommendations were provided for:

* Italian restaurants in 6 neighbourhoods
* Vegan/Vegetarian restaurant in 1 neighbourhood
* Japanese/Sushi restaurant in 2 neighbourhoods

The Canada census information also has additional data that could be used to provide additional information for clustering, such as language spoken, immigration, and ethnic background. This information could also be used to further segment Clusters 1 and 3 which had a high number of neighbourhoods.