Battle of the Neighborhoods
Analysis of Neighborhoods in Toronto

By Zeyar Oo
July, 2020

Table of Contents

Table of Figures

1. Introduction

1.1 Business Problem

For investors, it is useful to know what is likely to be popular in which neighborhood based on demographics such as age, education levels, marital status, income, ethnic background, and other factors that can explain and influence consumer behavior. There are some restaurants, for example, that open in a neighborhood with considerable sunk costs and only after a few months end up closing down because they fail to attract enough customers. If they can understand the market better in the first place, such failures could possibly be avoided.

1.2 Background

With a population of close to 3 million, Toronto is an important financial center for Canada. Being a cosmopolitan city with diverse demographics, it is only natural that it offers everything you need in terms of restaurants, parks, spas, pubs, gyms, business services and so on. Like in every major city, certain venues such as corporate offices can be concentrated in one area such as in a financial district. Households with higher income and education levels can also be concentrated in certain neighborhoods. Using two datasets on Toronto neighborhoods, I will test my hypothesis. If different patterns do exist, I will attempt to explain why certain venues are more prevalent in particular neighborhoods. I will also make recommendations for investment opportunities in case there are places with little competition but potential for attracting significant number of a particular group of customers.

2. Data Sources

The data used in this research comes from Foursquare and Toronto open data portal (https://open.toronto.ca/catalogue/). Foursquare API gives information on each neighborhood with venues and their location coordinates and category. The 2016 census for Toronto (https://open.toronto.ca/dataset/neighbourhood-profiles/) contains aforementioned important demographics for each neighborhood. By combining the two datasets, I plan to come up with recommendations for type of venues that are not yet fully saturated in selected neighborhoods.

With the use of clustering algorithms, it is possible to get an overview of the similarities or differences between the neighborhoods. Similar neighborhoods can be further analyzed to find out the concentration of certain venues, income levels, education, age, marital status, ethno-linguistic backgrounds and other demographic factors.

3. Methodology

I used venues data from Foursquare to create a K-means clustering model and compare it with the clusters obtained from the census dataset. This approach helps me figure out if the differences among neighborhoods can be explained by differences in demographics. The clustering labels from two different datasets can be joined using the “Neighborhood” index and compared using rank correlation methods such as Kendall’s Tau. If the differences between two clustering groups are not statistically significant, we can infer that the demographics from the census data indeed explain the variations among the neighborhoods.

3.1 Feature selection and data transformation

The census data for Toronto contains over 2300 columns with information such as income taxes, languages spoken at home, and mobility among others which are not very relevant in this study.

	_id	Category	Topic	Data Source	Characteristic	City of Toronto	Agincourt North	Agincourt South-Malvern West	Alderwood	Annex	Banbury-Don Mills	Bathurst Manor	Bay Street Corridor
0	1	Neighbourhood Information	Neighbourhood Information	City of Toronto	Neighbourhood Number	NaN	129	128	20	95	42	34	76
1	2	Neighbourhood Information	Neighbourhood Information	City of Toronto	TSNS2020 Designation	NaN	No Designation	No Designation	No Designation	No Designation	No Designation	No Designation	No Designation
2	3	Population	Population and dwellings	Census Profile 98-316-X2016001	Population, 2016	2,731,571	29,113	23,757	12,054	30,526	27,695	15,873	25,797
3	4	Population	Population and dwellings	Census Profile 98-316-X2016001	Population, 2011	2,615,060	30,279	21,988	11,904	29,177	26,918	15,434	19,348
4	5	Population	Population and dwellings	Census Profile 98-316-X2016001	Population Change 2011-2016	4.50%	-3.90%	8.00%	1.30%	4.60%	2.90%	2.80%	33.30%

Figure 1 (Original Data Frame)

Feature selection was challenging due to high dimensionality in this dataset. Fortunately, some of the features could be manually selected and this resulted in a data frame with 64 columns or variables. Afterwards, I transformed the numerical values by removing the white spaces, currency symbols, percentage signs and commas.

	Neighborhood	Population, 2016	Population Change 2011-2016	Total private dwellings	Population density per square kilometre	Children (0-14 years)	Youth (15-24 years)	Working Age (25-54 years)	Pre-retirement (55-64 years)	Seniors (65+ years)	Average household size	Persons living alone (total)	5,000 to 9,999	20,000 to 24,999	25,000 to 29,999	30,(34,(
0	Agincourt North	29113.0	-3.9	9371.0	3929.0	3840.0	3705.0	11305.0	4230.0	6045.0	3.16	1355.0	105.0	340.0	565.0	45
1	Agincourt South-Malvern West	23757.0	8.0	8535.0	3034.0	3075.0	3360.0	9965.0	3265.0	4105.0	2.88	1625.0	130.0	350.0	415.0	39
2	Alderwood	12054.0	1.3	4732.0	2435.0	1760.0	1235.0	5220.0	1825.0	2015.0	2.60	1105.0	35.0	160.0	175.0	17
3	Annex	30526.0	4.6	18109.0	10863.0	2360.0	3750.0	15040.0	3480.0	5910.0	1.80	7880.0	485.0	680.0	625.0	63
4	Banbury-Don Mills	27695.0	2.9	12473.0	2775.0	3605.0	2730.0	10810.0	3555.0	6975.0	2.23	4360.0	150.0	430.0	495.0	50

Figure 2 (Transformed Data)

Although the following correlation matrix with selected columns shows collinearity between some variables, it is not very strong. High collinearity does not significantly affect K-means because the algorithm computes distances between the samples to create clusters. However, high dimensionality does not necessarily add more information and can be noisy. To avoid this, I decided to use principal component analysis to reduce dimensions before K-means. Therefore, the previous data frame manually selected columns (variables) was discarded.

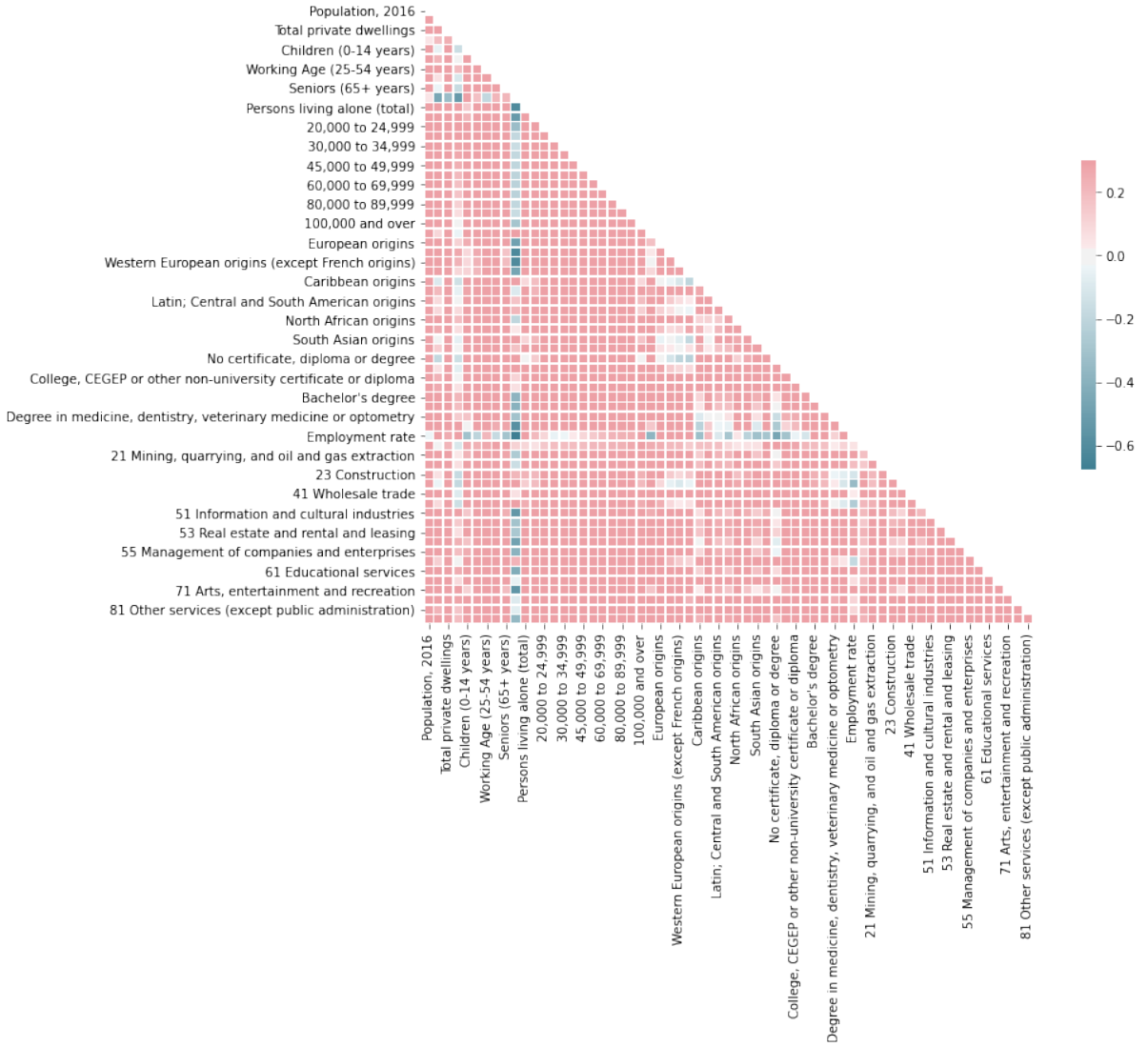


Figure 3 (Correlation Matrix)

3.2 Feature Extraction

As stated earlier, the original census data includes over 2300 variables and 140 neighborhoods.

```
df=pd.read_csv("../Downloads/toronto.csv",thousands=',',encoding='utf-8')
df.shape
```

```
(2314, 146)
```

Figure 4 (Original Data Shape)

To perform Principal Component Analysis, I started again with the original data set and performed data transformation steps mentioned in section 3.2. Only the columns with numerical values were kept to standardize data with StandardScaler() in Sci-kit Learn which uses mean and standard deviation to produce a set with mean of zero and variance of 1.

Afterwards, the scaled data was fed into the PCA function. The following graph shows that the first 30 components explain about 90% of the variance in the data.

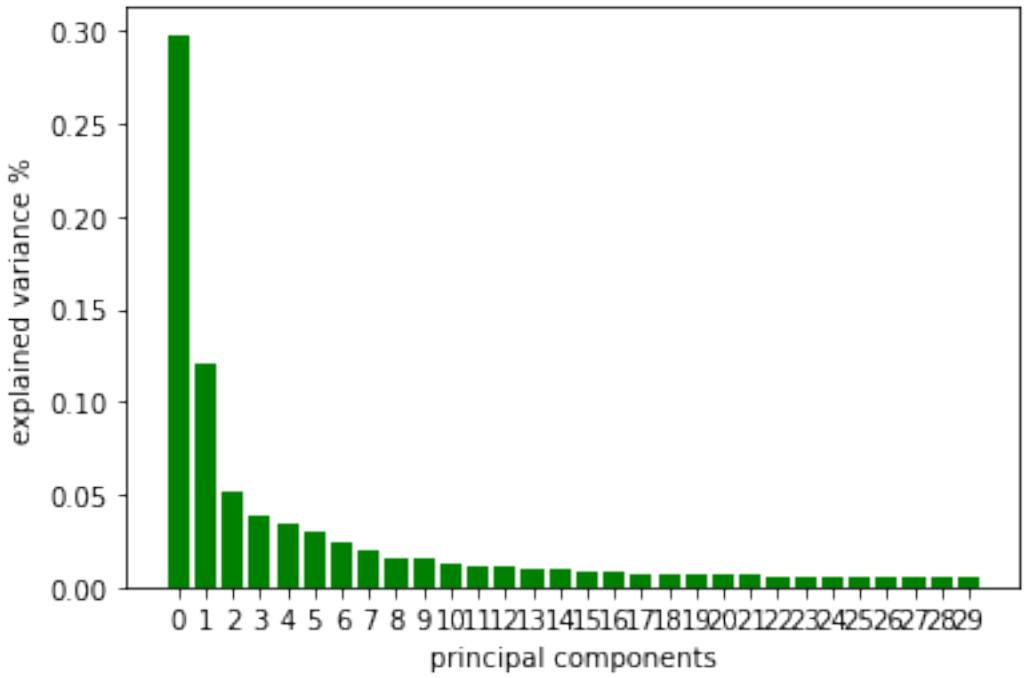


Figure 5 (Principal Components)

Figure 6 shows that it is not obvious if PCA 1 and PCA 2 are good for clustering. However, because there are 20 components, we need to run K-means to figure this out.

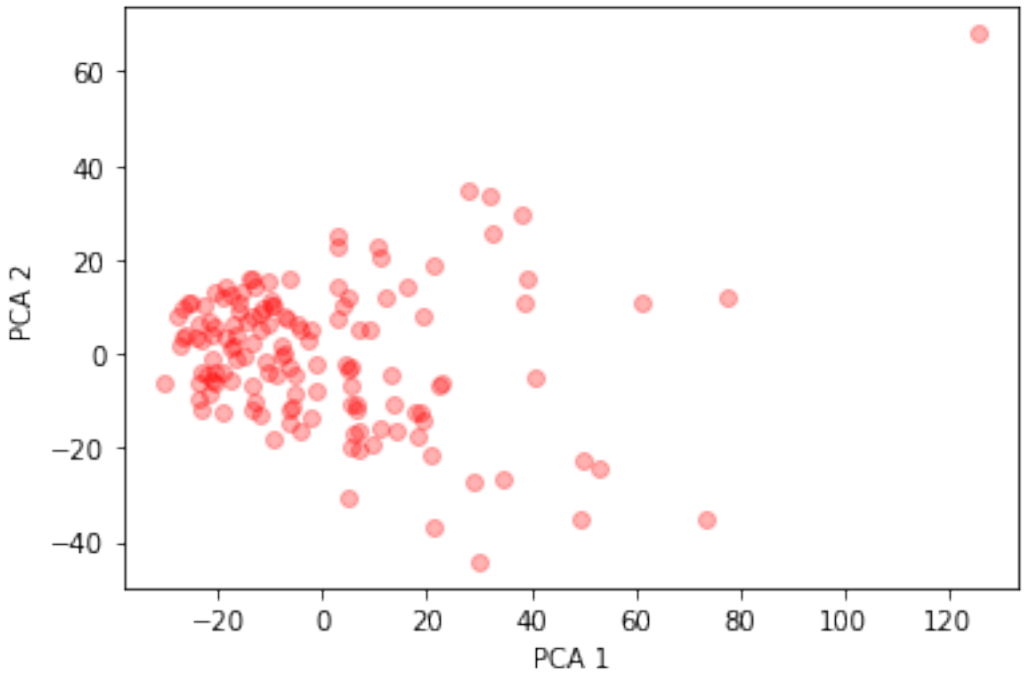


Figure 6 (PCA 1 and PCA 2)

After feeding the principal components through K-means algorithm, I observed that there is no clear elbow point for picking n clusters. Looking closely at the graph, 5 could be a good breaking point with a 'little' elbow.

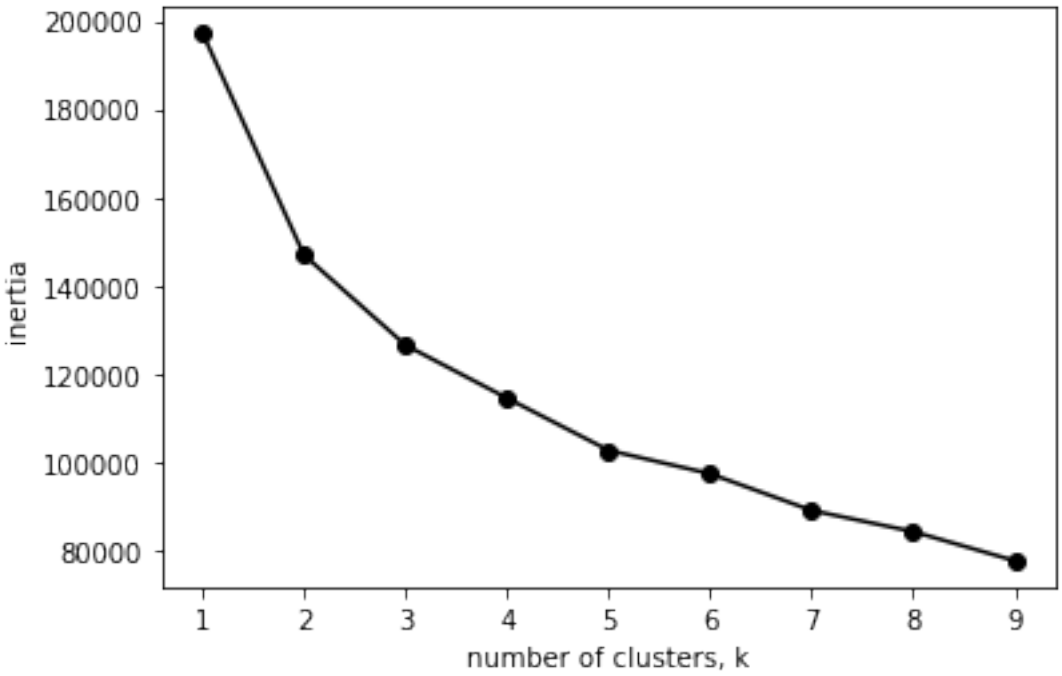


Figure 7 (K-means elbow)

Therefore, I run again with 5 clusters for K-means. It seems to be a good choice so far as we can clearly see five clusters in the following graph.

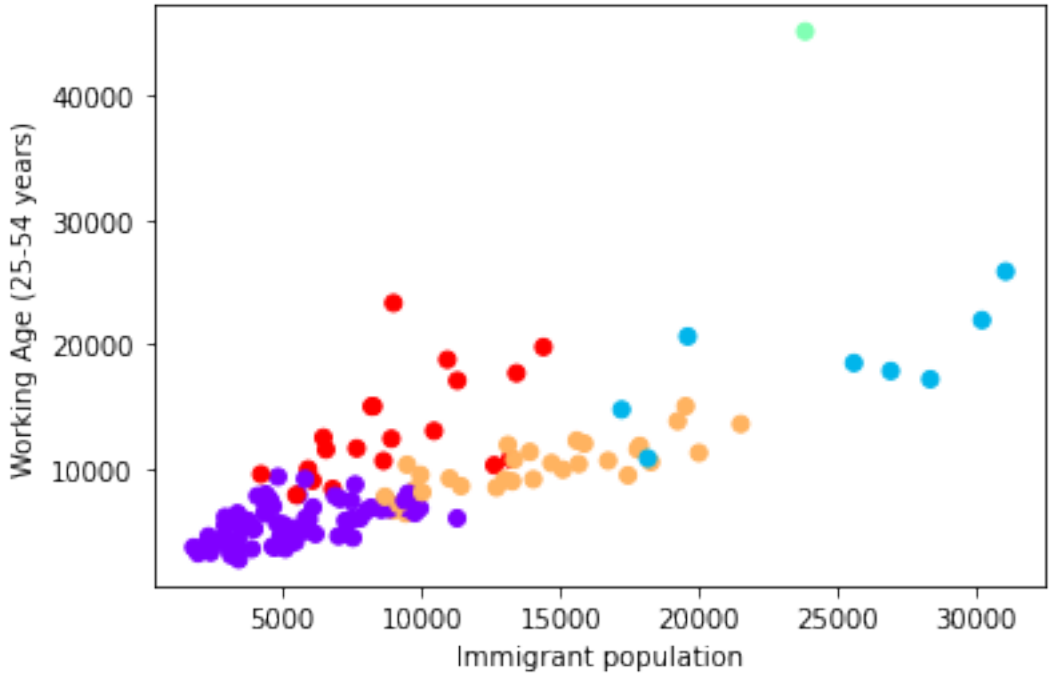


Figure 8 (Immigrant population vs Count of Working Age (25-54)

3.3 Separation of Outliers

Unlike DBSCAN algorithm, K-means is sensitive to outliers in the sample. Therefore, it is a good idea to separate outliers and treat them as a topic on their own. In this study, I am

interested in outliers especially because they are different from other neighborhoods in Toronto and, in theory, can present unique challenges or opportunities for a business.

Demographic outliers such as a high number of immigrants from South Asia or people of age 18-35, for example, could indicate an opportunity for offering services to that particular group. To determine if such opportunities exist, I can use Foursquare data and find out the density of venues by category.

I put all observations that deviate three standard deviations from the mean into a separate data frame. These values were obtained using z-score.

In the following two graphs, we can see that the outliers on average have higher number of immigrant population. In figure 3 for the normal group, the average number of immigrants in each neighborhood is around 7000.

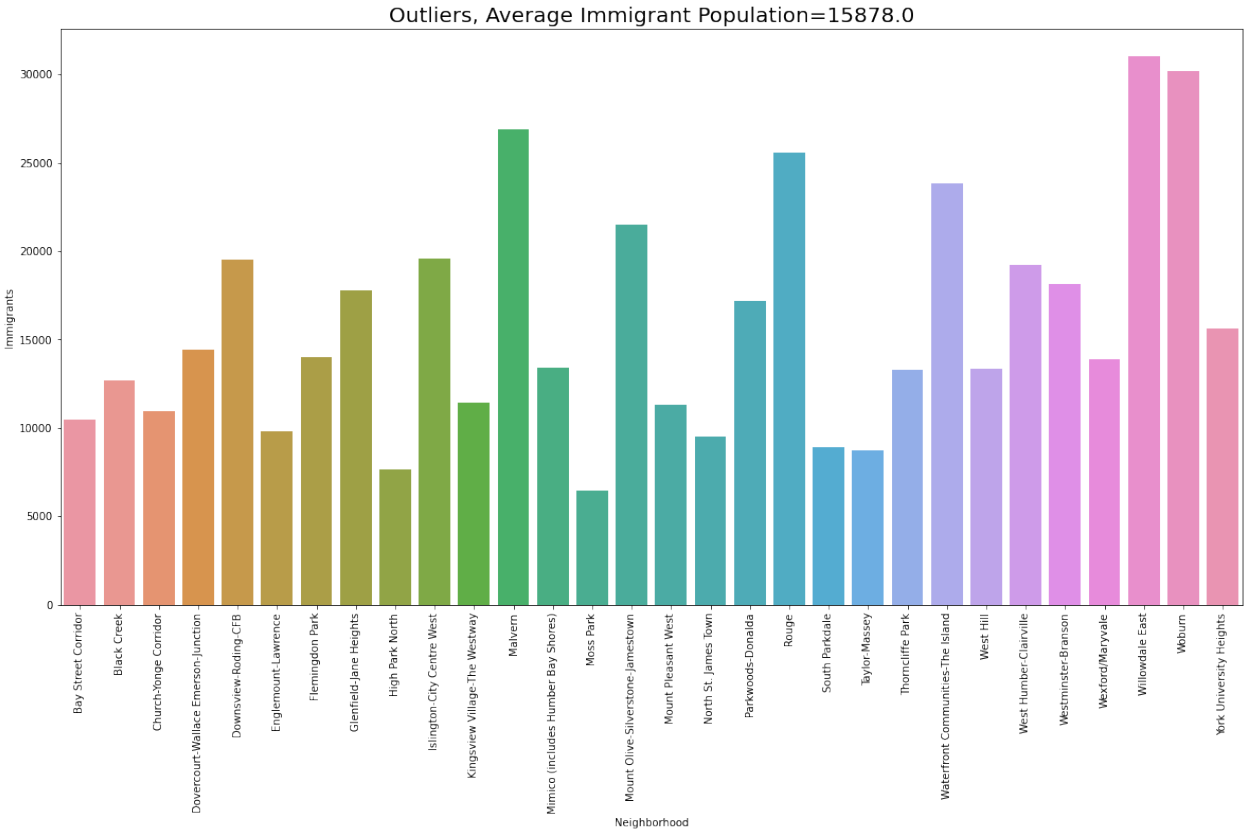


Figure 9 (Immigrant Population – Outliers)

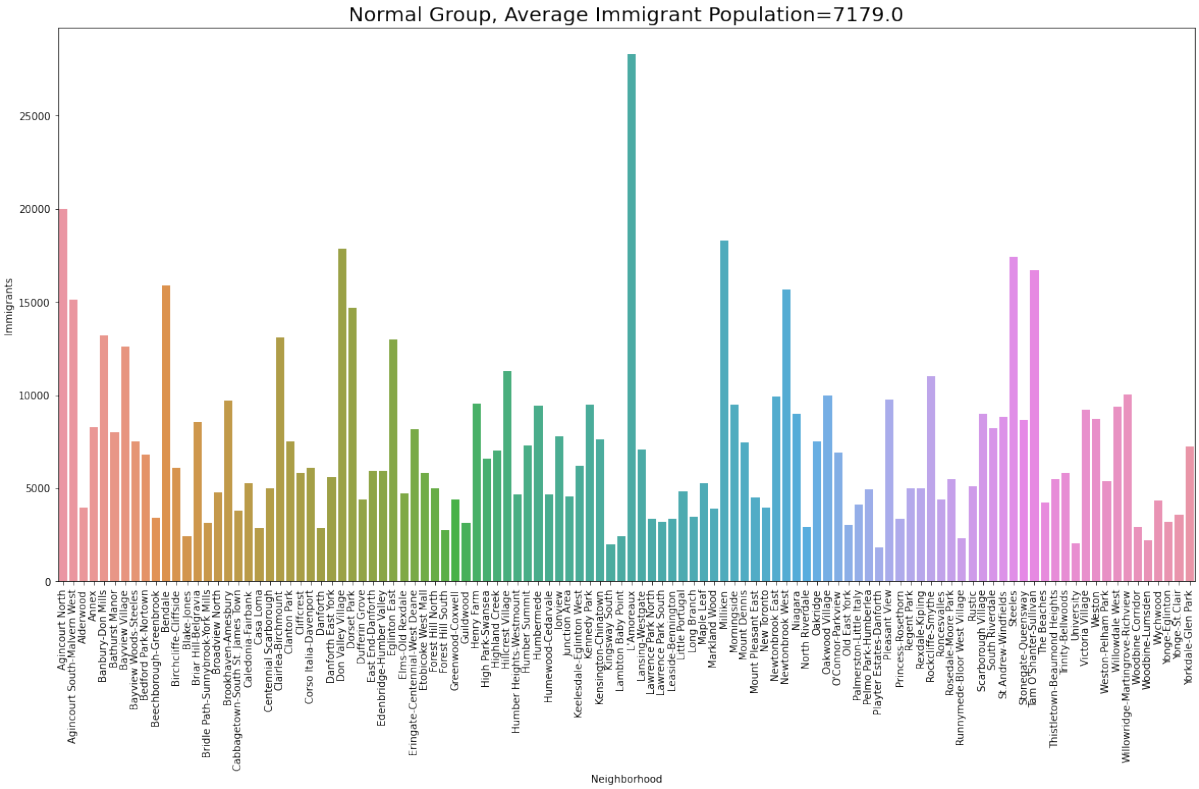


Figure 10 (Immigrant Population – Normal Group)

The differences disappear when we look at the proportion and not the absolute numbers like the previous graphs.

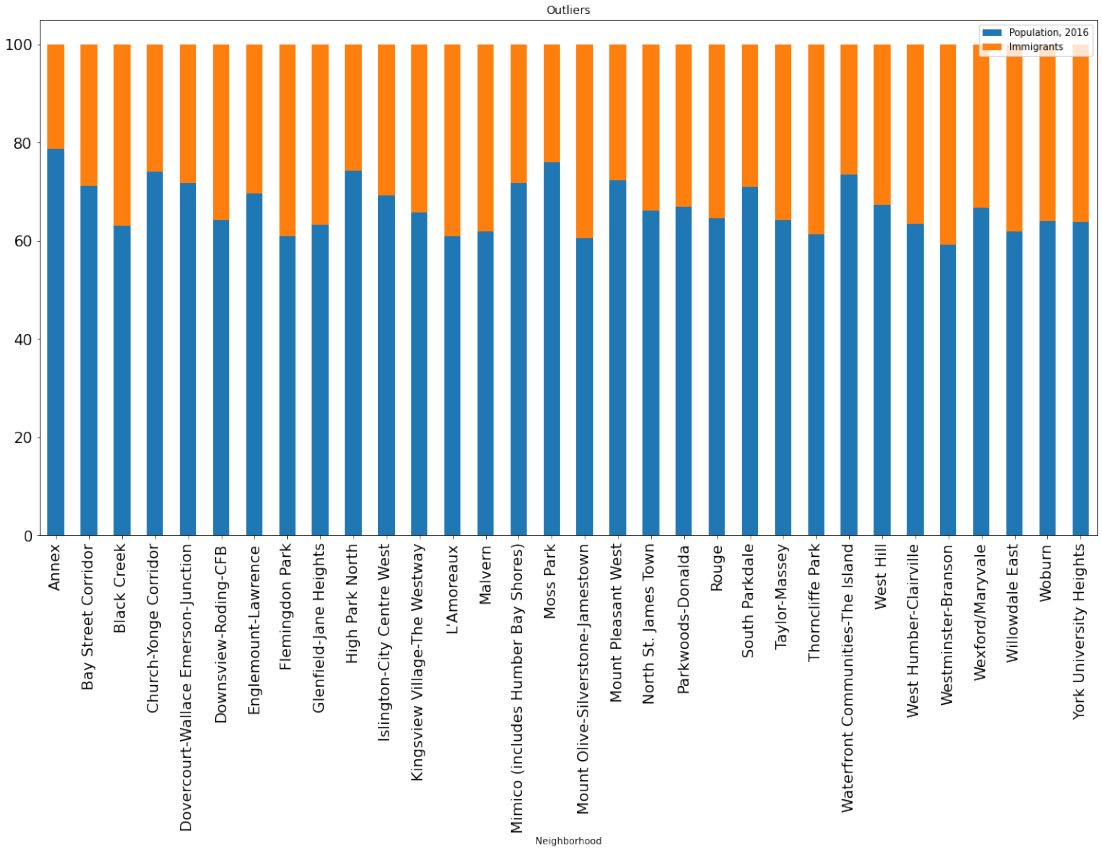


Figure 11 (Population Stack - Outliers)

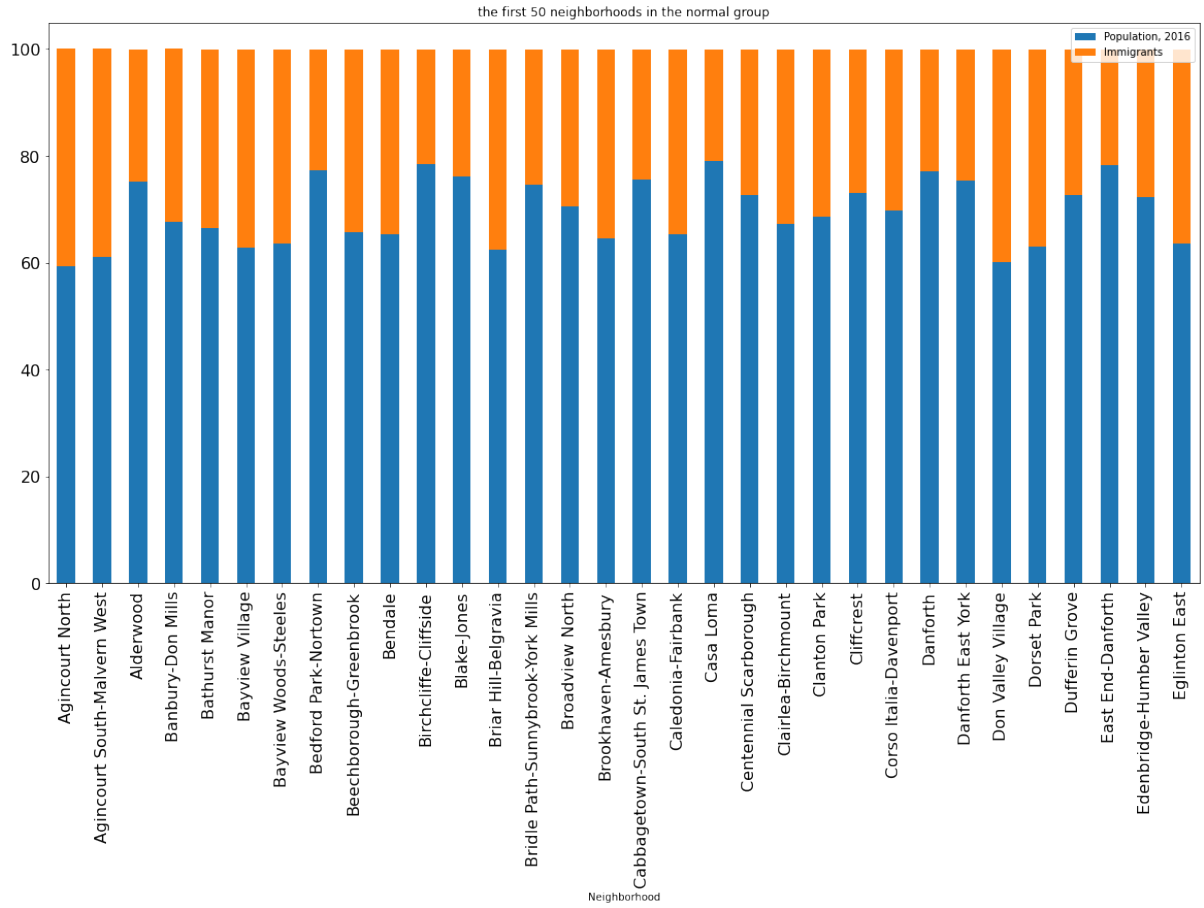


Figure 12 (Population Stack - normal group)

This is because outlier group contains neighborhoods with larger population. However, absolute numbers can also make a difference when they pass certain demand thresholds where it allows a business operation surpasses the breakeven point and be profitable. For example, a neighborhood with 40% millennials and Gen Z may not still be very attractive for a youth fashion brand if the whole population is only 500.

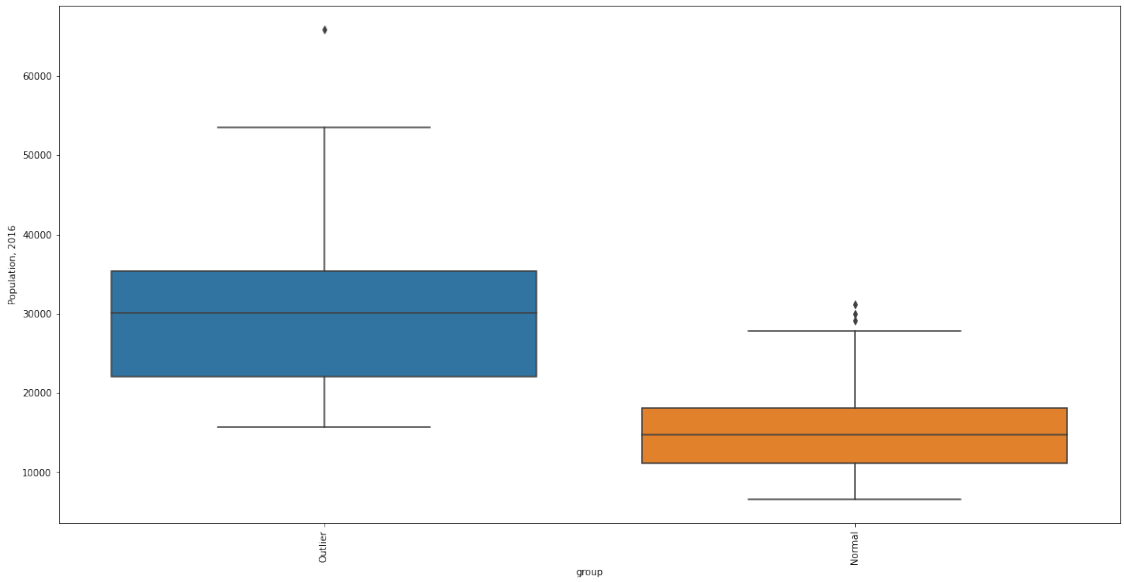


Figure 13 (Total Population Boxplot - outliers vs normal)

In addition, there does not seem to be much difference when it comes to education levels. The pattern is quite similar between the outliers and normal group (see Figure 5 and Figure 6).

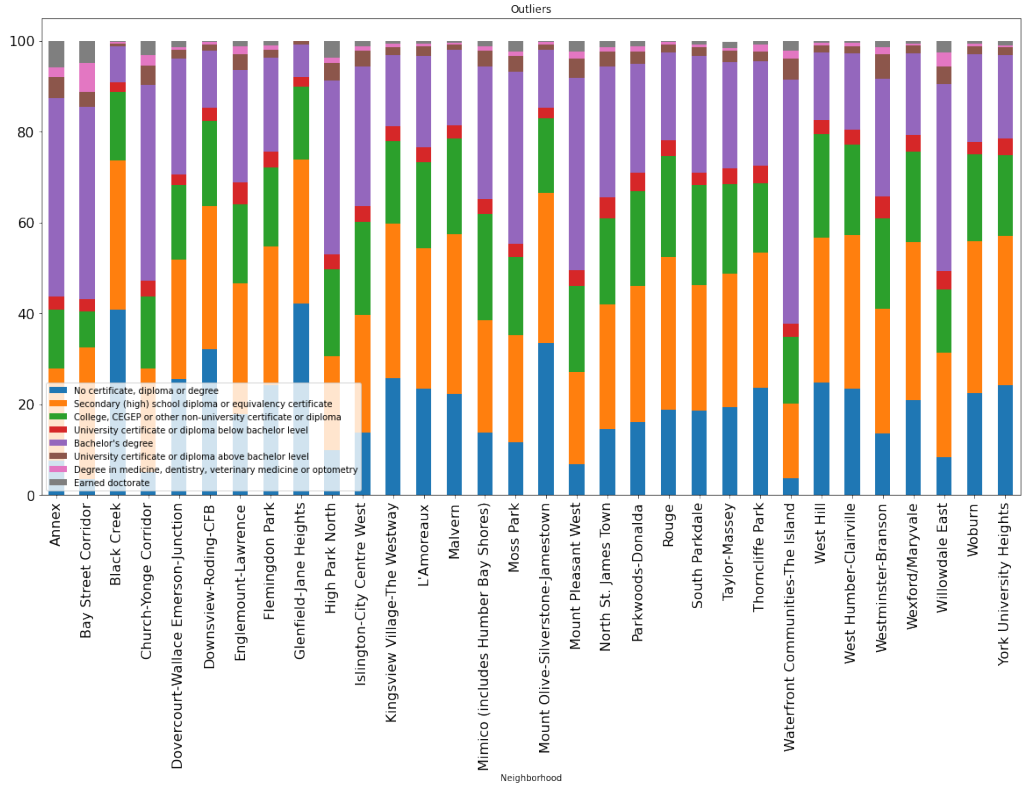


Figure 14 (Education Levels – Outliers)

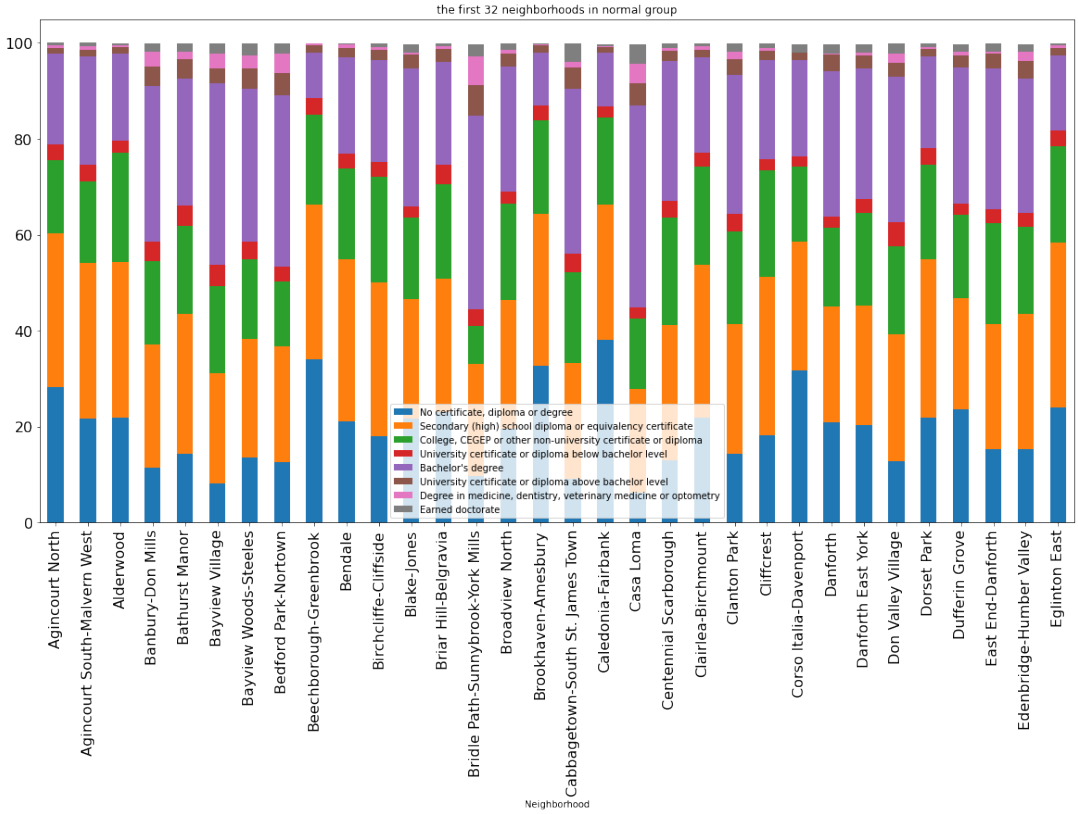


Figure 15 (Education Levels -Normal Group)

However, I found an interesting pattern when it comes to income groups. In the following two figures (13 and 14), you can observe that high income group (100,000 and over) is more prevalent in the normal group. The outlier group consist of less households with income 100,000 and over.

In Bridle Path-Sunny Brook-York Mills, it is especially higher than other neighborhoods. It is not clear why three neighborhoods are combined into one in the census data. It is probably due to the fact that all three combined only represent 9266 of the population. This wealthy neighborhood seems to be a suburban style area with low population density of 1040/sqkm while an average neighborhood in Toronto has 6261/sqkm.

Comparing the outliers and normal group based on other features would likely reveal more differences between the two. This is because differences in income—even when social security systems maybe exceptionally good—result in disparities in other aspects of life, and thus, divergent consumer behaviors.

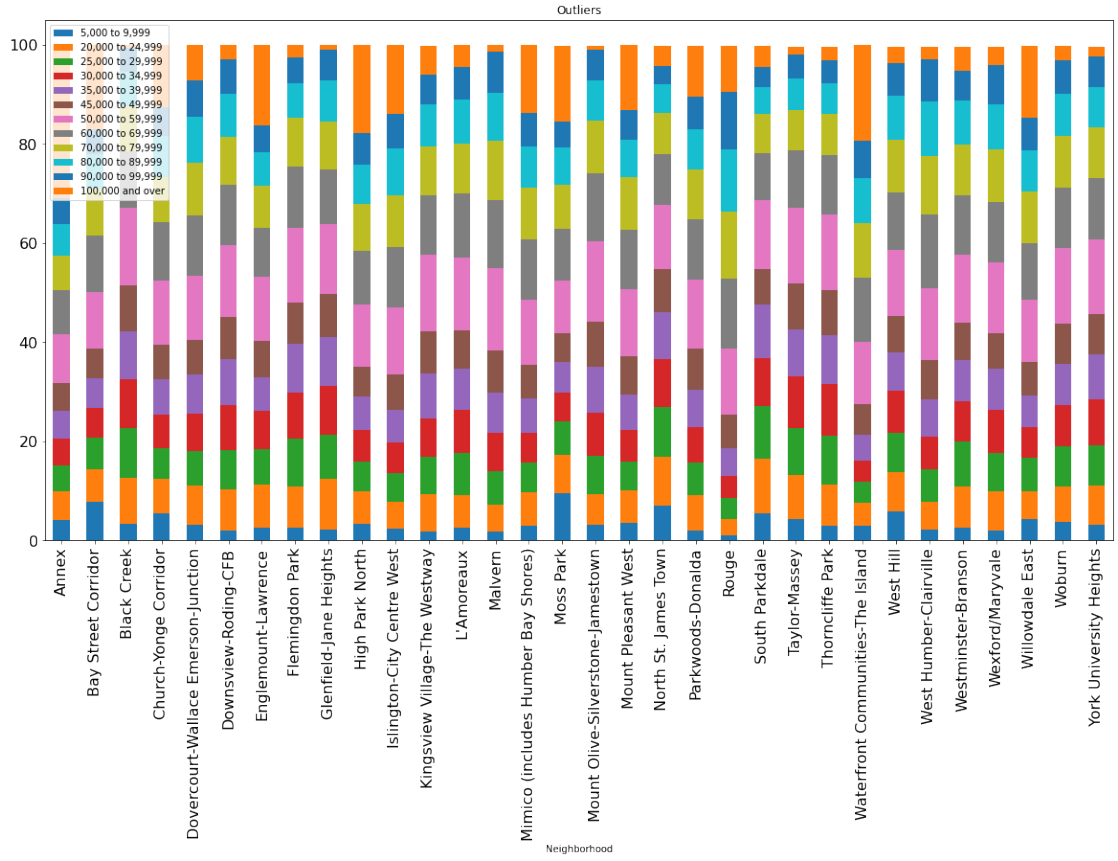


Figure 16 (Income groups in outlier neighborhoods)

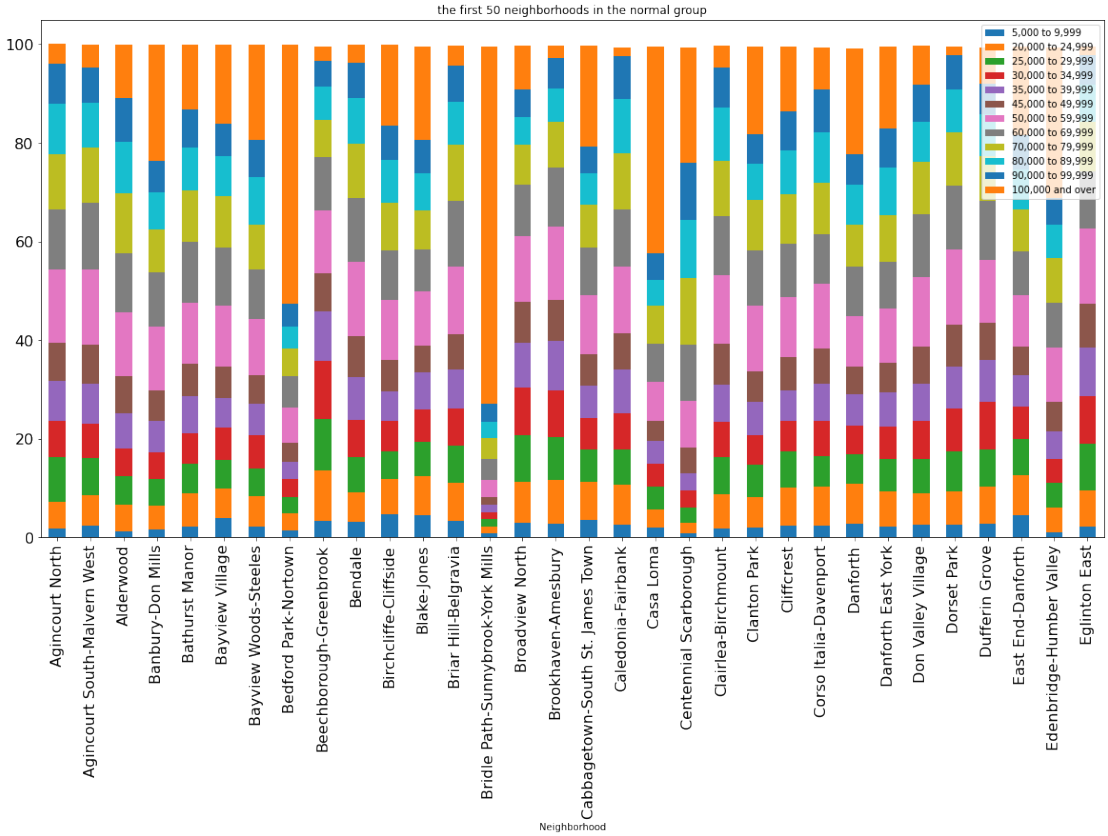


Figure 17 (Income groups in the normal group)

Another important assumption: due to large proportion of migrant in outlier neighborhoods, they would be qualitatively different from the 'normal' neighborhoods. An obvious example would be the case of Chinatowns across the world in which the majority ethnic group is Chinese and consequently we see more Chinese shops and restaurants in those neighborhoods.

In the following boxplots, we can see that South Asian origin is far higher in outlier group than normal group.

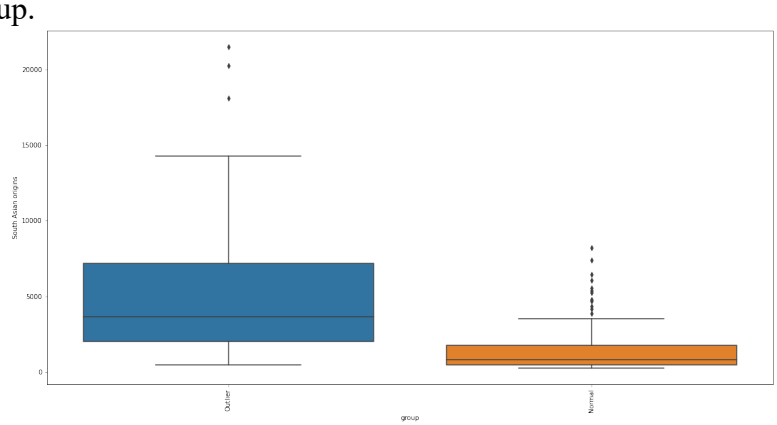


Figure 18 (South Asian origins)

It is not only in absolute numbers but also in proportion as seen from the calculations below:

```
People of South Asian origins as percentage of total population:
outliers : 18.14%
normal group : 9.99%
```

It is to be noted that outlier groups not only differ in population of migrants but also in other factors. I am using immigrant population variable as an example because it is one of the salient features that set the two groups apart. I will be examining other dimensions later. Due to differences outlined above, it is worth exploring normal and outlier groups separately.

4. Limitations

In this report, I will be only focusing on outlier group and it is one of the limitations of this exercise. Comparisons between outlier and normal group in more detail and that of between neighborhoods in each respective group should be done for a more thorough analysis.

5. Results

Finally, outlier group was selected for further analysis. Using Nominatim, I retrieved GPS coordinates for each neighborhood. However, some values were missing and I compensated for that by typing in manually. There are 30 outliers.
Correlation matrix on outliers return an interesting fact that French origins is highly correlated to jobs in arts and entertainment as seen in figure 16. It is worth looking more closely at this finding and see if it was mere a chance or not. This can be done by getting p-value for statistical significance. For the moment, real world information supports this finding because many French-Canadians are employed in Quebec's film industry.

```
# high correlation between French Origins and Arts/Entertainment is an interesting bit
t.loc['French origins'].loc['71 Arts, entertainment and recreation']
```

```
0.937286748043049
```

Figure 19 (French Origin correlation)

5.1 Clustering with census data

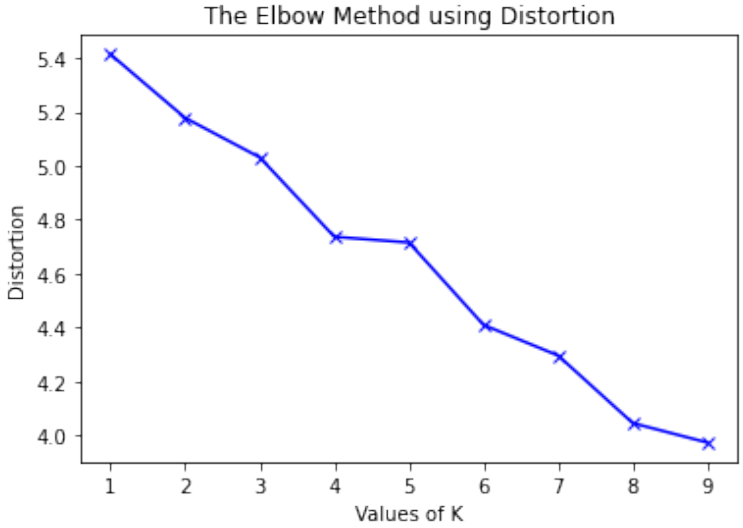
To perform K-means clustering, the values in outliers group obtained Principal Component Analysis were standardized using StandardScaler().

Iterating with "n" number of clusters again resulted in following elbow pattern. In this case, 4 seems to be the optimal number of clusters.

```
array([[-3.13239642e-01,  1.00468643e+00,  1.00638992e+00,
        -1.45700710e+00,  7.76324378e-01,  8.28348183e-01,
         4.63713599e-01,  1.73556171e-01,  1.69063978e-01,
        -2.74749606e-01, -7.32388586e-01, -2.35677992e-01,
        -3.63122097e-01, -1.72692129e-02,  1.29593437e+00,
         1.52121642e+00, -8.38212803e-02, -3.55818468e-01,
        -6.95088945e-02,  3.10260837e-01, -5.46192275e-02,
        -1.57217599e+00, -1.56488084e+00,  1.61267033e+00,
        -1.21788449e+00, -1.31609131e+00,  6.48128621e-01,
         2.14229097e+00, -1.40776975e+00, -6.39567503e-02],
       [-9.28071507e-01, -1.02988621e+00, -1.26405114e+00,
        -2.94850238e-01,  3.87186010e-01,  1.69786772e+00,
         9.84482215e-01, -2.99948369e-01,  2.51177682e-01,
         1.79915112e+00, -8.34963305e-02, -3.01210754e-01,
        -1.31548305e-01, -1.51517675e+00, -1.86659314e+00,
         2.33867800e-01, -7.35473461e-01, -1.41283797e+00,
         1.45907261e+00,  9.36516050e-02,  2.09885888e-02,
         4.33039041e-01,  4.96025097e-01,  4.31010359e-02,
        -1.25867435e+00, -7.96047399e-01, -1.29298229e+00,
```

Figure 20 (standardized scales)

Figure 21 (Elbow graph & Clusters - outliers)



	Neighborhood	Cluster Labels
0	Bay Street Corridor	1
1	Black Creek	3
2	Church-Yonge Corridor	0
3	Dovercourt-Wallace Emerson-Junction	3
4	Downsview-Roding-CFB	1
5	Englemount-Lawrence	1
6	Flemingdon Park	1
7	Glenfield-Jane Heights	3
8	High Park North	3
9	Islington-City Centre West	1
10	Kingsview Village-The Westway	0
11	Malvern	0
12	Mimico (includes Humber Bay Shores)	0
13	Moss Park	0
14	Mount Olive-Silverstone-Jamestown	3
15	Mount Pleasant West	0
16	North St. James Town	0
17	Parkwoods-Donalda	0
18	Rouge	0

Plotting the clusters on the map didn't result in clear geographical clusters. This could indicate a problem with the data or it could be that number of clusters is not optimal.

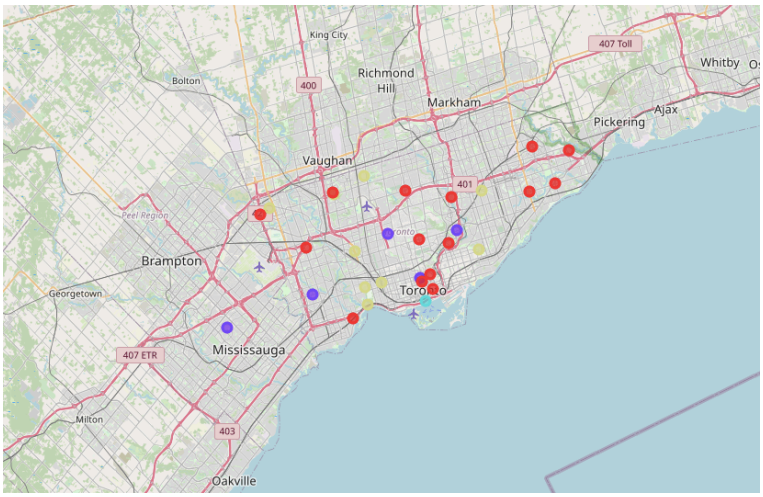


Figure 22 (Cluster map- Census Data)

5.2 Clustering with Foursquare Data

The outliers were further analyzed with venues data obtained from Foursquare API. In total, data for 661 venues were obtained for 30 neighborhoods with 172 unique categories.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Bay Street Corridor	43.667342	-79.388457	Indigo	43.669065	-79.389057	Bookstore
1	Bay Street Corridor	43.667342	-79.388457	DanceLifeX Centre	43.666956	-79.385297	Dance Studio
2	Bay Street Corridor	43.667342	-79.388457	Toronto Hemp Company	43.668419	-79.385848	Smoke Shop
3	Bay Street Corridor	43.667342	-79.388457	Tokyo Sushi	43.665885	-79.386977	Sushi Restaurant
4	Bay Street Corridor	43.667342	-79.388457	Bay Street Video	43.668890	-79.389247	Video Store

Figure 23 (Venues Foursquare Data)

In the next step, all categories in the data frame were one-hot coded and then mean value for each neighborhood was computed. The mean values were then used to rank most popular categories in each neighborhood.

```
toronto_grouped = toronto_onehot.groupby('Neighborhood').mean().reset_index()
toronto_grouped
```

	Neighborhood	Yoga Studio	Afghan Restaurant	American Restaurant	Aquarium	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	Automotive Shop	BBQ Joint	Bakery	Bank	Bar	Ba
0	Bay Street Corridor	0.010000	0.000000	0.000000	0.00	0.01	0.01	0.000000	0.01	0.000000	0.00	0.000000	0.000000	0.000000	0.0
1	Black Creek	0.000000	0.000000	0.000000	0.00	0.00	0.00	0.000000	0.00	0.000000	0.00	0.000000	0.000000	0.000000	0.0
2	Church-Yonge Corridor	0.032258	0.000000	0.000000	0.00	0.00	0.00	0.032258	0.00	0.000000	0.00	0.000000	0.032258	0.000000	0.0
3	Dovercourt-Wallace Emerson-Junction	0.000000	0.000000	0.000000	0.00	0.00	0.00	0.000000	0.00	0.000000	0.00	0.000000	0.000000	0.000000	0.0

Figure 24 (Venues One-Hot Coding and Mean)

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Bay Street Corridor	Coffee Shop	Clothing Store	Café	French Restaurant	Japanese Restaurant	Italian Restaurant	Boutique	Sushi Restaurant	Mediterranean Restaurant	Park
1	Black Creek	Construction & Landscaping	Playground	Coffee Shop	Diner	Fast Food Restaurant	Falafel Restaurant	Event Space	Ethiopian Restaurant	Electronics Store	Distribution Center
2	Church-Yonge Corridor	Coffee Shop	Diner	Park	Dance Studio	Burger Joint	Sculpture Garden	Burrito Place	Ramen Restaurant	Creperie	Yoga Studio
3	Dovercourt-Wallace Emerson-Junction	Coffee Shop	Café	Brewery	Pharmacy	Skating Rink	Portuguese Restaurant	Italian Restaurant	Grocery Store	Park	Discount Store
4	Downsview-Roding-CFB	Electronics Store	Park	Women's Store	Dim Sum Restaurant	Fast Food Restaurant	Falafel Restaurant	Event Space	Ethiopian Restaurant	Distribution Center	Discount Store

Figure 25 (Most popular venue categories by neighborhood)

In Figure 21 above, we can see that coffee shops, Cafes, and restaurants are most popular categories in Bay Street area. It is not surprising since Bay Street is a financial district with corporate offices.

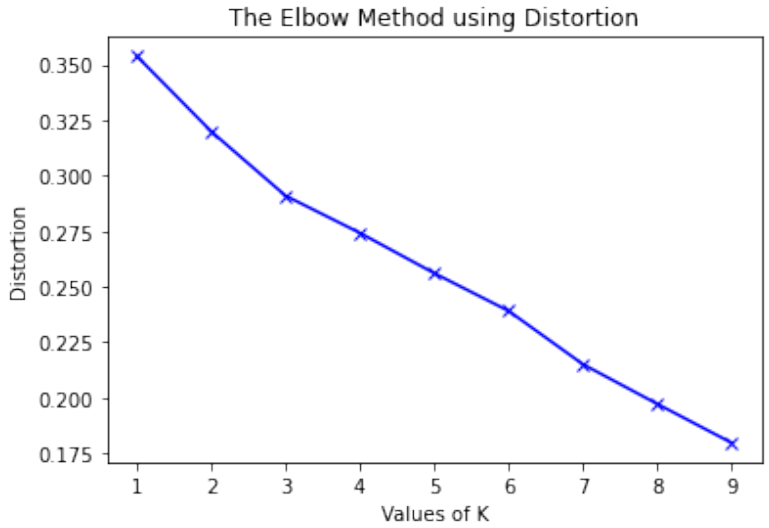


Figure 26 (Elbow graph- venues clusters)

With venues data, the optimal number of clusters was 3. However, 4 clusters were needed to compare with the previous K-means result for census data. Performing K-means with 4 clusters, the following clusters were obtained as shown in the map.

Comparing the two maps below show that the census data does not completely explain the differences between the clusters. However, there are still a few similarities indicating that

some of the data in the census such as income categories, age and other demographic factors do influence clustering of venues.

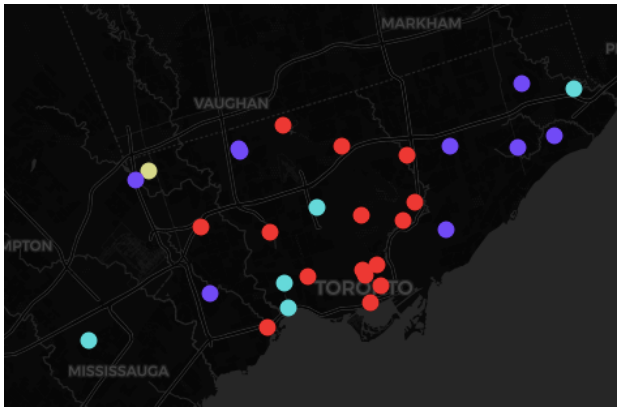


Figure 27 (Cluster Map - Venues)

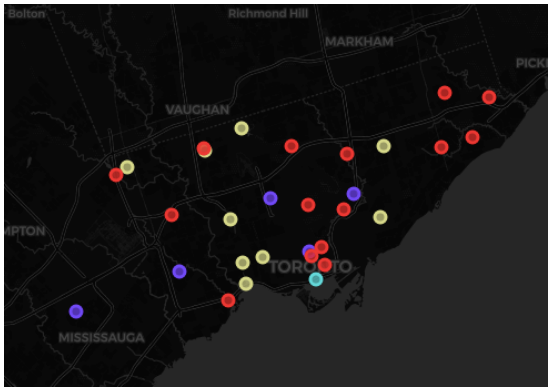


Figure 28 (cluster map - census data 2)

	Neighborhood	Cluster_census	Cluster_venues
0	Bay Street Corridor	1	0
1	Black Creek	3	0
2	Church-Yonge Corridor	0	0
3	Dovercourt-Wallace Emerson-Junction	3	0
4	Downsview-Roding-CFB	1	2
5	Englemount-Lawrence	1	2
6	Flemingdon Park	1	0
7	Glenfield-Jane Heights	3	1
8	High Park North	3	2

Figure 29 (comparing cluster labels)

Because the two clusters represent the same entities, i.e. neighborhoods, if the two datasets used for clustering are highly correlated to each other, the clusters should be roughly the same. At a first glance, we can see on the map that it is not the case.

To see if the different is statistically significant, I computed Kendall's tau and results were as follows:

```
tau:0.2535211267605633, p-value:0.12559254742606987
```

Tau value indicates they are only 25% similar. The difference is not statistically because p-value is above the usual 5% threshold at 12.6%.

6. Discussion

What can be inferred from this result is that the principal components from extracted from census data can still explain, albeit to a small degree, the differences between venue clusters. However, it is still arguable whether it is even meaningful to compare the two clusters that come from two entirely different datasets. For that, I will not be exploring further in this report.

As seen in Figure 18 and following paragraph, outlier group deserves further analysis because of its unique characteristics. Figure 18 shows that South Asian population is higher in both and absolute numbers. However, from Figure 25, we can see that South Asian restaurants are not very common in those neighborhoods and may present business opportunities.

Just like the South Asian restaurant example, we can use exploratory data analysis on other categories to uncover various opportunities where the market is not yet saturated for certain business types.

7. Conclusion

In this report, I attempted a solution for identifying business opportunities in a geographical area using census data and Foursquare API. I first explored the Toronto census data to uncover the demographic characteristics. Exploratory data analysis and separating outliers revealed interesting patterns that can be used to draw valuable insights. In order to see how much census data can explain density of venues in neighborhoods, I used K-means clustering on both datasets and compared them visually and used Kendall's tau to see if the difference is statistically significant. To conclude, further analysis is necessary to ensure that the statistically and machine learning methods are sound and appropriate. This can be done through the use of other existing models such as Kolmogorov-Smirnov test, Tukey's HSD and DBSCAN, K-medoids or other clustering algorithms.