IBM Data Science

Capstone Project: Demographic classification of Ottawa Neighbourhoods

Alexandre Poulin

Table of Contents

Introduction: Business Problem

So far in the capstone project class exercises, we have been clustering neighbourhoods based on similarities of various businesses in different neighbourhood. This can be useful if you live in a neighbourhood and want to move to a neighbourhood, or if you are a business owner and want to see what parts of a city has similar businesses to each other. The goal of this project will be to see if there are any connection to the demographics of the neighbourhood and the venues in the neighbourhood. If such a connection exists, then it can be useful to know for various reasons:

- Someone who is looking to move may be interested in finding an area near them which are popular among others who share a demographic. This can improve their overall happiness as opposed to finding similar neighbourhoods to where they already were.
- A business which may have a target demographic may want to know which neighbourhoods are already popular among those demographics.
- Business owners would like to know who their audience are so that they can focus more on a group
- Maybe a neighbourhood is popular to a group only because it has a high density of a type of business which is primarily enjoyed only by a single demographic.

The questions we will aim to answer are the following: Can we make a connection between the demographics of a neighbourhood and the venues in that neighbourhood? Because it would be interesting for me based on where I live, the city which will be analysed will be Ottawa, Canada.

Data Acquisition and initial manipulation.

To complete this project, we will use an open source GEOJSON file (https://gist.githubusercontent.com/mattleduc/10549018/raw/b76778e347fde6ae7cbec2066be1dcb8119aa93e/ons.geojson) which contains data on the outline of each neighbourhood, as well as demographic information on the population there. We can also use the location data to acquire data from Foursquare for the venues in the neighbourhoods. Using these data sets together, we can extract trends to better understand the state of each neighbourhood.

For each neighbourhood, we can calculate the centroid using the geometric data the response. To calculate the centroid, we use:

$$C_x = \frac{1}{6A}\sum_{i=0}^{n-1}(x_i + x_i + 1)(x_i y_{i+1} - x_{i+1} y_i)$$

$$C_y = \frac{1}{6A}\sum_{i=0}^{n-1}(y_i + y_i + 1)(x_i y_{i+1} - x_{i+1} y_i)$$

We could calculate the area of each neighbourhood using:

$$A = \frac{1}{2}\sum_{i=0}^{n-1}(x_i y_{i+1} - x_{i+1} y_i)$$

In the data, we have the AREASQKM column that we could use for A. We calculate this instead as a sanity check.

All above formulas from https://en.wikipedia.org/wiki/Centroid.

We will need to find the furthest point away from the centroid to make sure we select a radius at least that big when looking up Foursquare data. To do this, one could try to compute:

$$Rmax = \max_i \sqrt{(C_x - x_i)^2 + \left(C_y - y_i\right)^2}$$

However, since we have all the values in longitude and latitude, it doesn't give us an answer in km. Instead we need to use:

$$R = 2R_{Earth} \sin^{-1}\left[\sin^2\left(\frac{\phi_2 - \phi_1}{2}\right) + \cos(\phi_1)\cos(\phi 2)\sin^2\left(\frac{\lambda_2 - \phi\lambda_1}{2}\right)\right]$$

Where ϕ_1, ϕ_2, λ_1, and λ_2 are the two latitudes and two longitude, respectively; and R_{Earth}=6,371,000 meters is the radius of the Earth. This comes from the Haversine formula. More information can be found at https://en.wikipedia.org/wiki/Haversine_formula.

Because the radius we're using is going to be the biggest distance between the centroid and the border, the radius might include venues outside of the neighbourhood that we need to eliminate from the query results. We will thus need to write a function to test for this. The easiest way to check for this is to imagine sliding the coordinate of the venue along the latitude and see how many times it crosses the borders. If it crosses an odd number of times, it was in the neighbourhood, but if it crosses and odd number of times, then it wasn't. Say the center is at v_x and v_y. Then it will cross a section of the border if:

$$y_i < v_y < y_{i+1} \; and \; v_y - y_i > \frac{y_{i+1} - y_i}{x_{i+1} - x_i}(v_x - x_i)$$

or

$$y_i > v_y > y_{i+1} \; and \; v_y - y_i > \frac{y_{i+1} - y_i}{x_{i+1} - x_i}(v_x - x_i)$$

It should be noted that if we ever have a situation where $x_{i+1} = x_i$ then this is a horizontal line that would not be crossed, so we can ignore it.

We now have all the data frames that we need:

- Ottawa_Data: contains all the demographic data for the neighbourhoods
- geometry, shape: contains the boundary and type of boundary for each neighbourhood
- Ottawa_grouped: contains all the information on venues in each neighbourhood.

Methodology

In this project, we want to see if there are any connections between the demographics of a neighbourhood and the types of venues in neighbourhood. This analysis will follow 3 steps:

1. Cleaning: Both the Ottawa_Data and Ottawa_grouped have >100 columns. We will see if we can reduce the number of columns by seeing if any are similar, could be aggregated, or if they are not relevant.
2. Clustering: We will individually cluster the neighbourhoods based on demographics and on venues using k-means clustering.
3. Comparison: We will use a heat map to see the correlation between the clusters.

Analysis

Initial cleaning

Let's start by looking at columns in the Ottawa_Data data frame. They can be split up into a few groups. Let's go over the first group:

- OBJECTID
- names
- BLKPOP2011
- TOTDWL2011
- URBDWL2011
- AREASQKM
- NID
- ID
- Neighborhood,
- Population
- Total Area of Neighbourhood (km2)
- Total Population

From these, we will only use the Neighborhood, AREASQKM and Population column. The other columns are either duplicated, irrelevant, or of unknown meaning. Let's go and remove the other columns.
For the next categories, there are many subdivisions such as age for the population, income, etc. We use a heatmap of a correlation matrix to see if any variables behave the same and so we can use larger categories. Note that because of the small size of the heatmap, not all the column names appear.

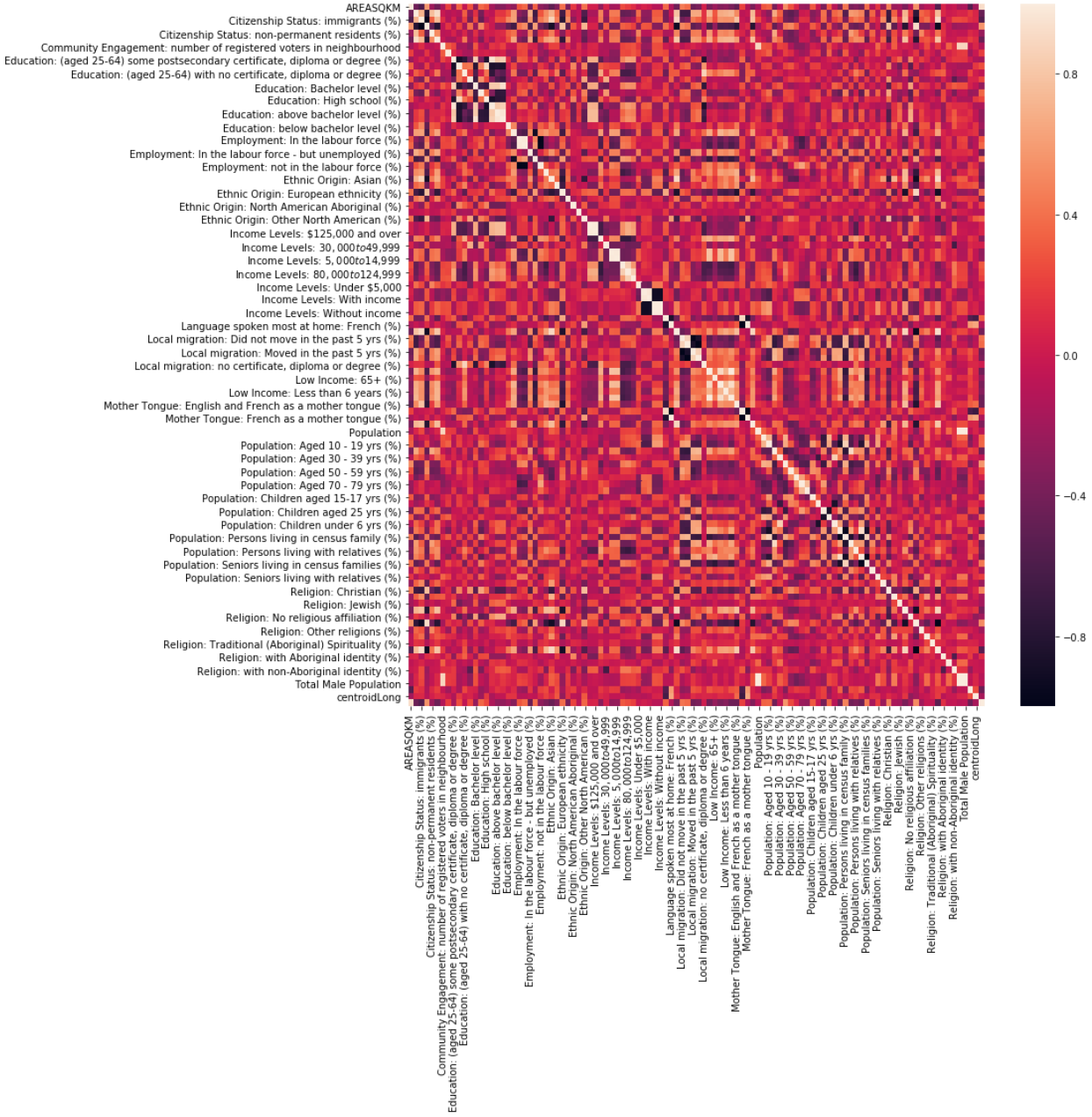

Figure 1: Heatmap of the correlation between the demographic data

From the heatmap, we will make a few changes.

With respect to Citizenship Status, we will keep:

- "Citizenship Status: immigrants (%)"
- "Citizenship Status: non-immigrants (%)"

and drop the rest. The dropped columns have similar correlations to the "Citizenship Status: immigrants (%)" column.

For the Community Engagement columns, these don't correlate much with any other variables so they will be dropped.

Next, we look at the education columns:

- New Column "Education: non-bachelor" defined as the sum of the "Education: Apprenticeship or trades (%)", "Education: College, CEGEP or other non-university (%)", and "Education: High school (%)" columns, which will be dropped
- "Education: bachelor level or above (%)"

We will drop all the other columns as they have the same or similar information as the two columns we will use.

For the Employment columns:

- "Employment: In the labour force (and employed) (%)": keep
- "Employment: In the labour force - but unemployed (%)": keep

and drop the rest. The dropped columns have similar correlations and information.

For the Income levels, we will make the following changes:

- New Column "Income Level: over $50,000" defined as the sum of the "Income Levels: $100,000 and over", "Income Levels: $80,000 to $99,999", and "Income Levels: $50,000 to $79,999" columns
- Keep "Income Levels: Under $5,000", "Income Levels: $15,000 to $29,999", "Income Levels: $30,000 to $49,999", "Income Levels: $5,000 to $14,999", "Income Levels: With income" and "Income Levels: Without income"

and drop the rest.

For the Local migration, we will keep:

- "Local migration: Did not move in the past 5 yrs (%)"
- "Local migration: Moved in the past 5 yrs (%)"
- "Local migration: no certificate, diploma or degree (%)"

And drop the rest.

For the Low Income, we will keep:

- "Low Income: 65+ (%)"

- "Low Income: Total population (LIM-AT) (%)"

And drop the rest.

We will drop the "Total Female Population" and "Total Male Population" as they have the same information as the total population. We will also combine "Population: Aged 70 - 79 yrs (%)" and "Population: Aged 80+ yrs (%)" into a single column "Population: Aged 70+ yrs (%)"

For the Ethnic Origin, Language and Mother Tongue, we will keep all the columns.

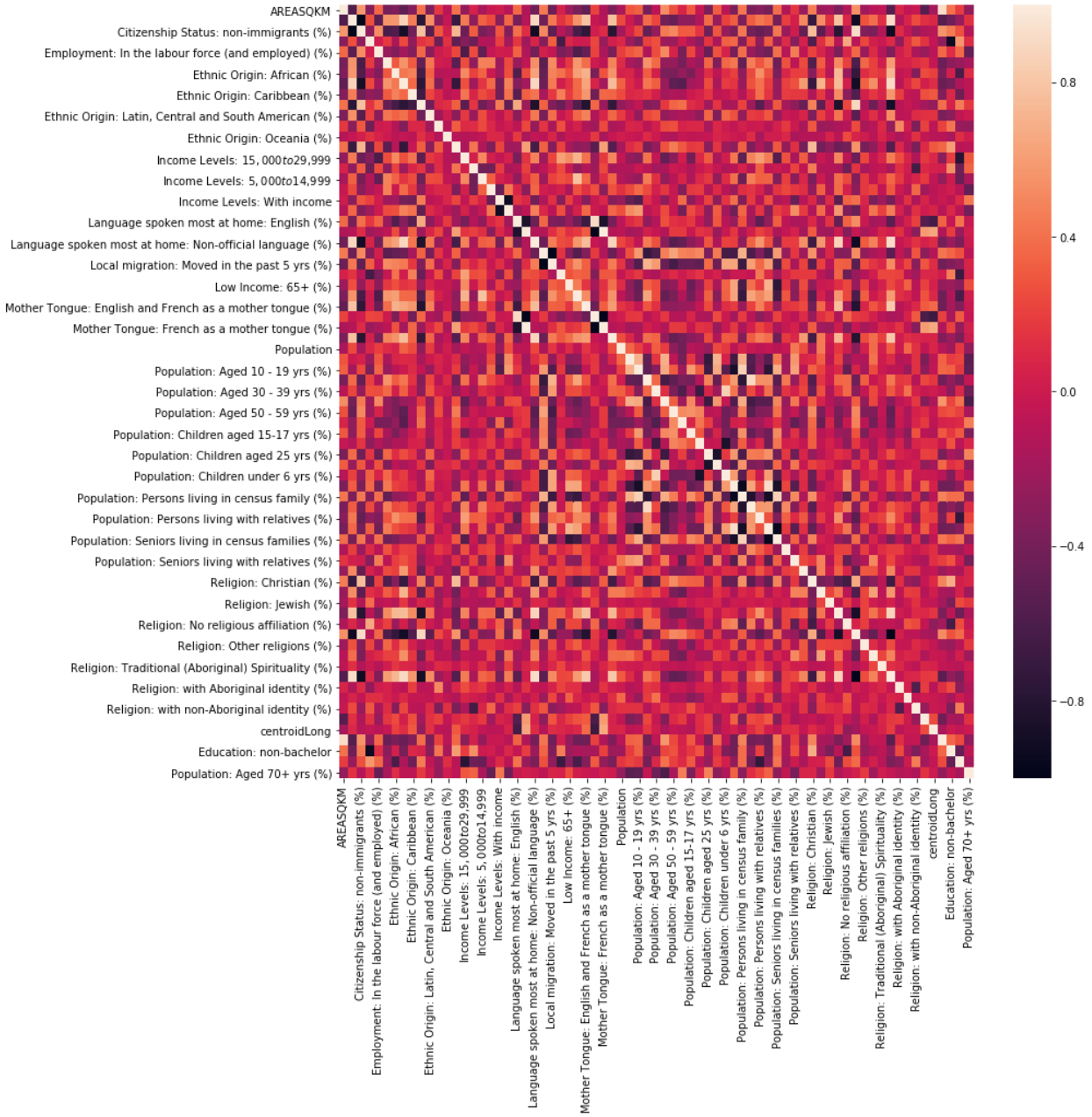

Figure 2: heatmap of the demographic information after initial cleaning.

Constructing the model

At this point we can apply the k-means algorithm to this data set. First, we will need to decide which number of clusters is correct using the elbow method.

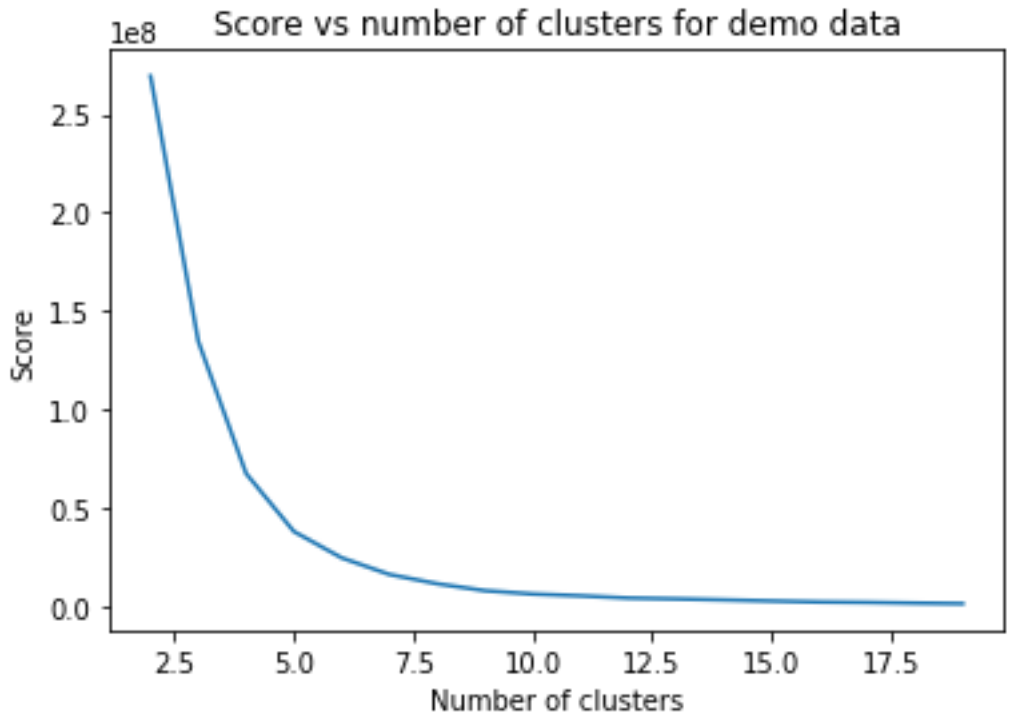

Figure 3:Finding the k parameter in k-means for the demographic data

We see that 5 seems to be the correct number of clusters. Let's also do the clustering for the venue data. First, we look at correct value of k.

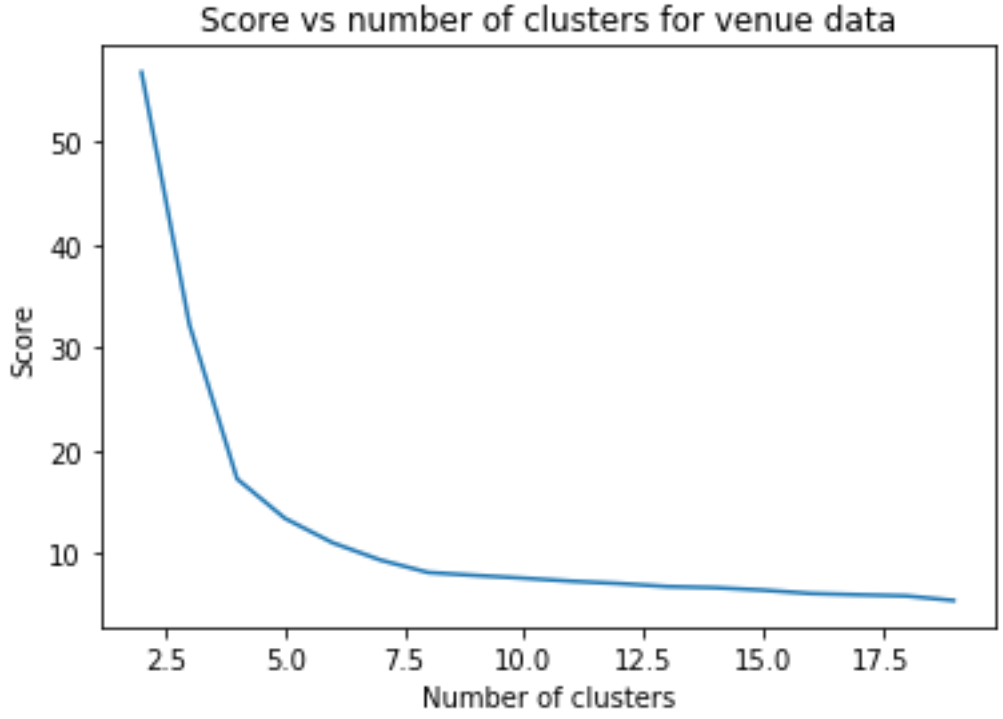

Figure 4: Finding the k parameter in k-means for the venue data

We see that there is an elbow at k=8 so that is what we will use for k.

Let's visualize how the distribution of clusters look on a map. Note that we split up the maps into two chunks each. This is because folium can't handle the whole set of neighbourhoods at once. First, cluster 1 and 3 from the demographic data.

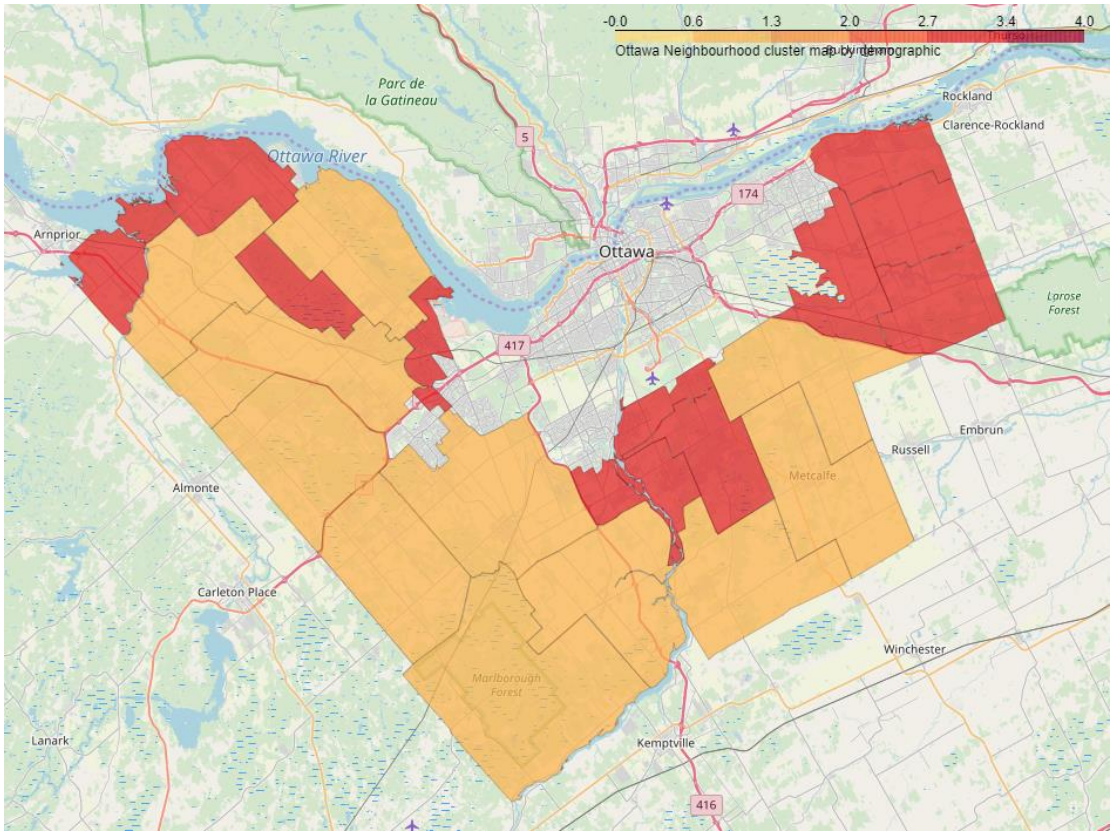

Figure 5: Cluster of neighbourhoods using the demographic data showing cluster 1 and 3

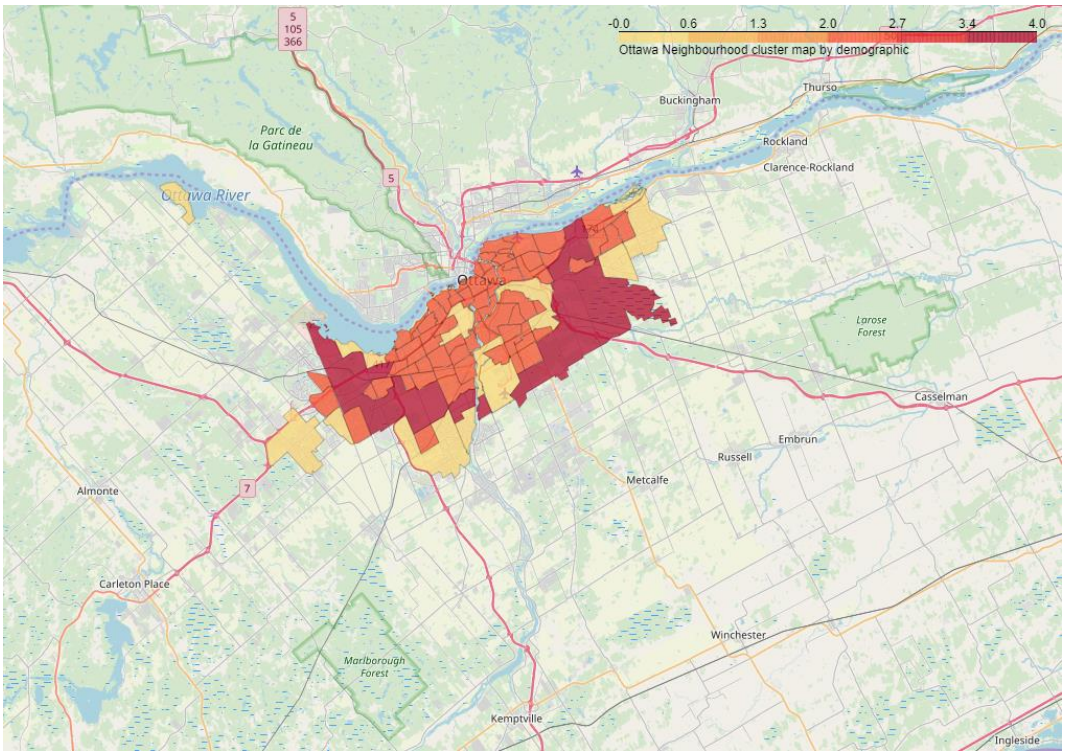

Figure 6: Cluster of neighbourhoods using the demographic data showing cluster 0, 2, and 4

Now for the venue clustering. The clusters are again slip up into two maps so that it could be handled by folium.

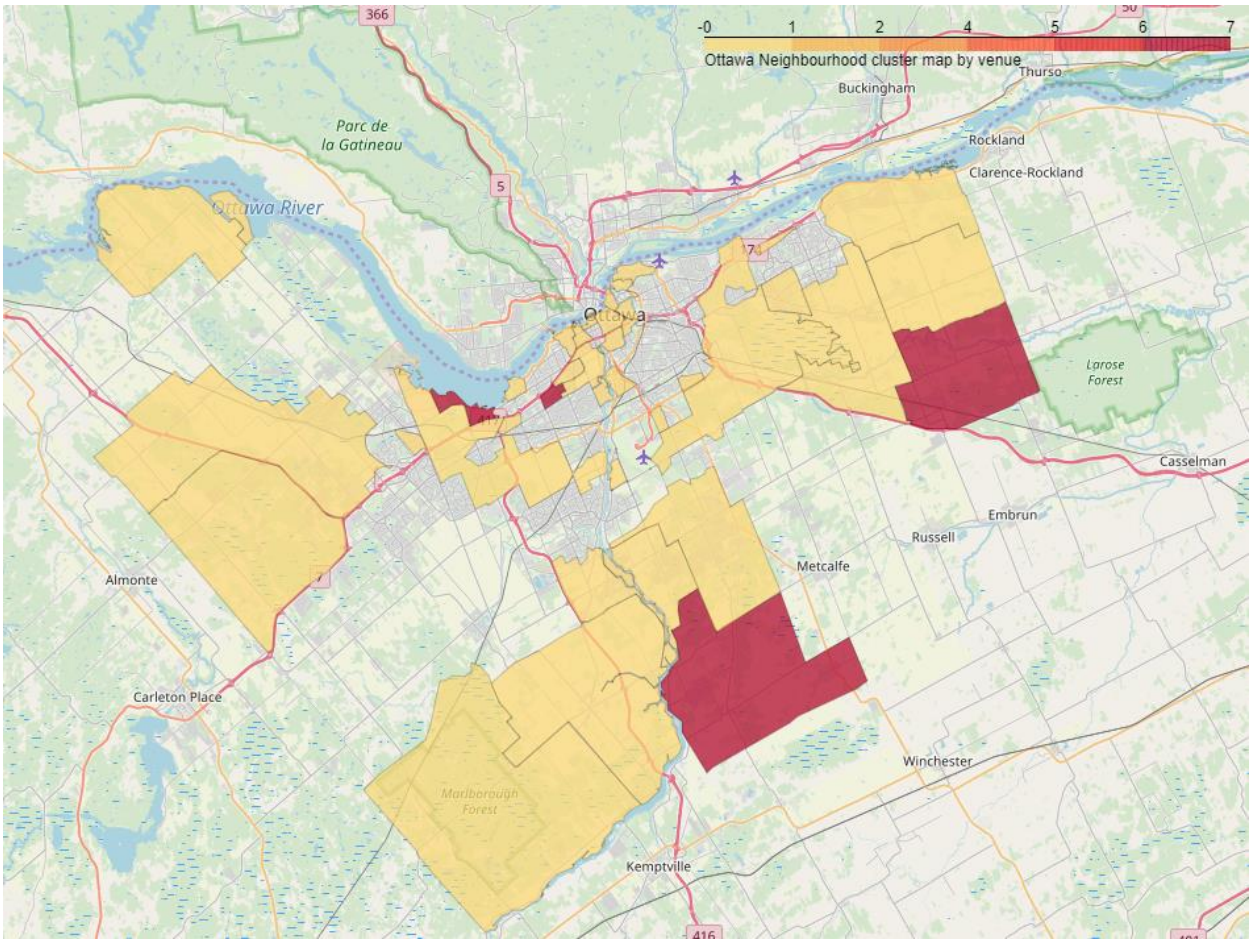

Figure 7: Cluster of neighbourhoods using the venue data showing cluster 0 and 6

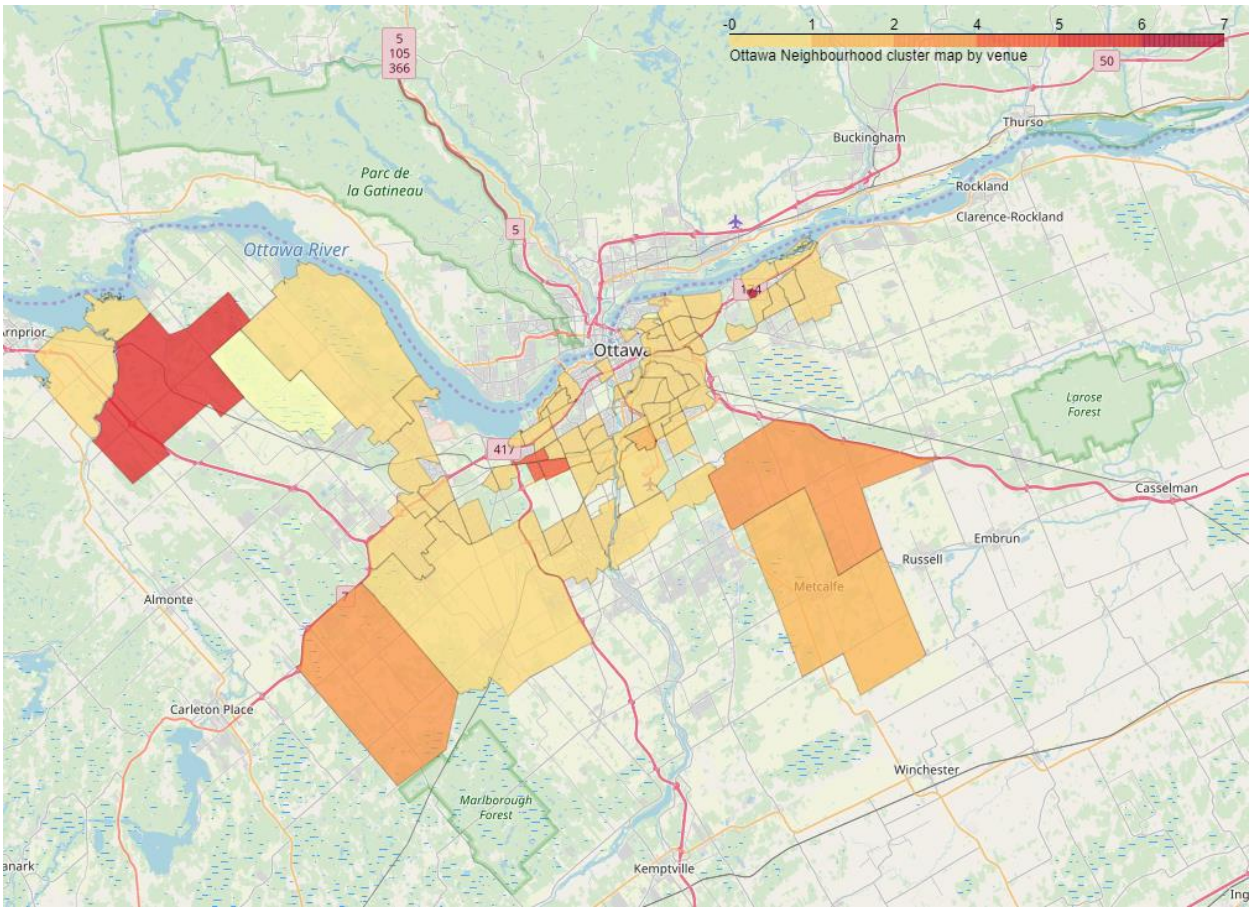

Figure 8: Cluster of neighbourhoods using the venue data showing cluster 1,2,3,4,5, and 7

Now let's combine the clustering information into a single data frame and compare them. We make a heat map using the pair of cluster values for each neighbourhood.

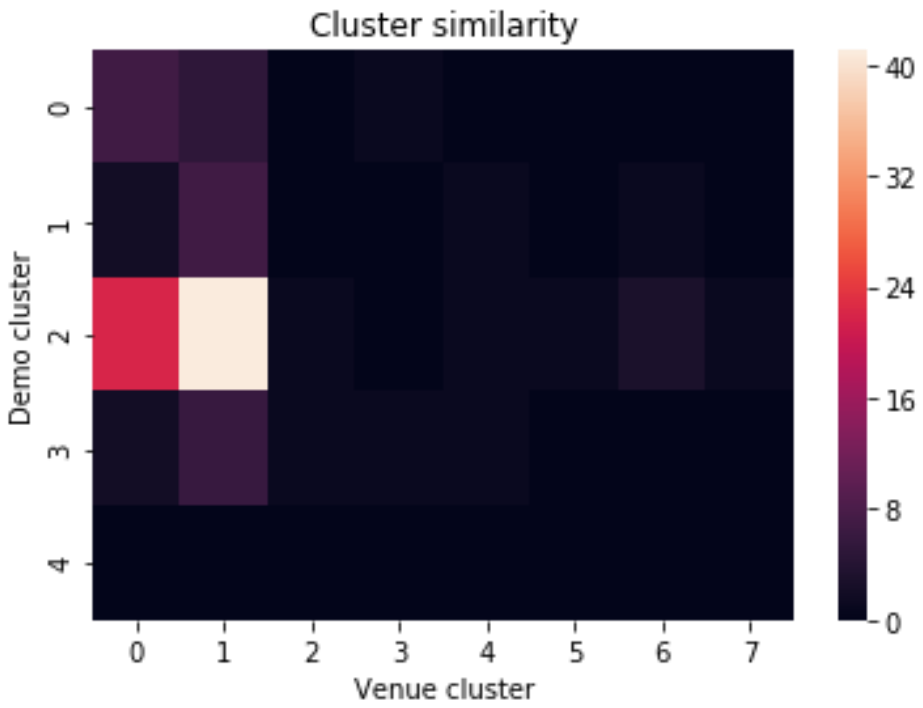

Figure 9: Heat map for the number of neighbourhoods which fall in each pair of cluster values

Results and Discussion

In this analysis, we clustered Ottawa neighbourhoods using two different methods: one using demographic information found in a geojson file and one using venue information from FourSquare. Using the demographic clustering, we see a clear division based on the geography with clusters forming sort of rings around the center of the city. This makes sense as it is typical for the population to spread outwards from a city. A fact about Ottawa is that until the 1960's, it was prohibited to build a building which was taller than the Peace Tower, the tower in Ottawa's parliament building. This caused the population to spread outwards as Ottawa grew as a city and there is a lingering effect. Another historical note is that the borders of Ottawa were changed in 2001 to include the more rural townships. We can see that the outermost neighbourhoods seem to belong to the same cluster.

For the venue clustering, there is much less geographical cohesion. Ottawa is a highly diverse city with a lil' italy, a Chinatown, lil' lebanon, french sections, and so on. These would be expected to be part of their own clusters. What we see is that although the business seems to be in those regions, this is not reflected in the demographic clustering. This suggest the represented demographic for a ghetto are not significantly higher in these areas compared to other places in the city, i.e. although there may be more Asian people in Chinatown, there are not statistically higher than other demographics.

There seems to be poor agreement between the different clustering methods with notable exceptions. Cluster 2 for the demographic cluster mainly include the downtown area and the area nearby. This area is also dense in similar bars, clubs, restaurants, museums, etc. Because of this, the area around downtown tends to be more homogeneous and thus part of the same cluster. This

means that downtown is not only clustered by its demographics, but also its venues. Overall, apart from the downtown area, there is little connection between the population and the businesses.

Conclusion

We set out to see if there was any link between the population and the venues in a neighbourhood. We found that only one part of the city, downtown, had a significant number of neighbourhoods which could be clustered together by venue and demographics, simultaneously. All other sections followed different patterns, the demographic clusters following rings around downtown, and venues following ethnic ghettos.