Locating Potential Venues for Homeless Shelters in the City of Toronto

by

Yael Valdez Navarro, MASc.

A capstone project
presented to Coursera
in the completion of
the
IBM Data Science Certificate

Contents

1. Introduction

1.1 Background

There has been a rising trend in Canada, a vulnerable group of people with diverse issues such as drug abuse, mental health and fleeing abusive situations face the problem of homelessness.

Some statistics reveal the severity of the problem for homeless people:

- 26% had been hospitalized with an average of 5 times for an emotional or mental health problem
- 21% indicated that they were receiving help on drug treatment
- 43% of respondents indicated that addressing their health needs was important
- 35% of respondents had been diagnosed with at least one mental health condition
- Toronto's shelters operate at **98% capacity** every day

Multiple charity groups organize soup kitchens and shelters to alleviate the daily hazards faced by this community. The fact is most of these locations operate at almost full occupancy rate every day. To tackle this, a business problem is proposed.

1.2 Business Problem

By considering the challenges presented in the previous section, the problem statement is then, by analyzing multiple sources of data (i.e. socio-economic and geographic), find the best possible locations for establishing new homeless shelters on the city of Toronto.

There are multiple criteria that are given more importance to the establishment of the best location. These following factors can also be considered our assumptions for the problem definition:

- **Accessibility:** The ideal venue should be close to public transport, train stations or bus stops.

- **Socio-economic:** It has been shown that low income areas are hot-spots for homelessness, specially with increasingly prohibitive rent costs.

- **Geographic:** Identify "trends" in neighborhoods which have high prevalence of homelessness. To maximize area covered, the ideal venue should not be too close to other venues.

1.3 Target audience

The information generated by the completion of this project will provide with relevant knowledge to charities, especially those considering an expansion to their current operating shelters. Non-profit organizations looking to establish their initial venue will also benefit from this project.

The goal will be to provide a venue location, pinpointing general latitude, longitude and neighbourhood.

2. Data

2.1 Datasets:

The data compiled to solve the proposed problem is gathered from various sources. An outline is displayed according to the criteria chosen.

Socio-Economic:

- Neighbourhood profiles from Toronto Open Data initiative (TOD):
  - *Original size (*146 columns x 946 rows): The original file contained an exhaustive number of features pertaining to socio-demographic indicators, we selected those relevant only to our project, i.e. income and labour features.
  - *Reduced size (*25 columns x 140 rows): The wrangling procedure consisted in eliminating categories which did not pertain to this project, keeping only income and labour indicators. The dataset originally had neighbourhoods as columns, we decided to transpose the matrix to keep consistency with the rest of our data. Moreover, multiple descriptors were eliminated, due to their high correlation to one another and irrelevant for our analysis.

Geographic:

- Toronto geojson - This file was originally forked from a project by Github user adamw523. A geojson file is contains a list of features, in this case, neighbourhoods, along with coordinates that once mapped will display a polygon pertaining to the shape of the neighbourhood. This file was modified slightly, as the neighbourhood names did not fully comply with those in the socio-economic file.

- Shelter locations obtained from TOD, originally of file type shp, which are typically used with GIS type software, however, for our application it was converted to a comma separated values file (csv) with online tool. This dataset provides the longitude and latitude of most homeless shelters in the city of Toronto.

Accessibility:

- Foursquare – An API that allows queries to their platform to obtain venue information, be it restaurants, bars, museums, etc. For our application, it will be applied to retrieve public transportation stations. Mainly, to query the latitude and longitude of train, bus and streetcar stations

2.2 Libraries

Alongside the datasets, several Python libraries will be used to exploit their functionality and their integration with the project.

- Pandas: Used to store large table-like arrays of data
- Requests: Used to handle external queries and requests.
- Numpy: Used to handle vectors and matrices in a more intuitive manner.
- Matplotlib: Used to graphically display different data.
- Folium: Used to display maps and visualize location-based data.
- Scikit-learn: Used to perform clustering analysis, such as, k-means.

3. Methodology

This section describes the procedure followed for the exploratory analysis of the data, the representation of the socio-economic data and the clustering of the neighbourhoods based on their features.

3.1 Preliminary exploration of the data

To get acquainted with the agglomeration of the neighbourhoods and shelters we display their location information on a map. Image 1 helps to visualize the scale and limits of the area we are working with. Even if we are unfamiliar with the conditions of homelessness in Toronto, we would quickly notice with this display of the data that the shelters tend to agglomerate on a certain area of the city, as seen in Image 2. The reason for this at this point of the analysis, remains unknown. Additional analysis can be found in the Appendix.

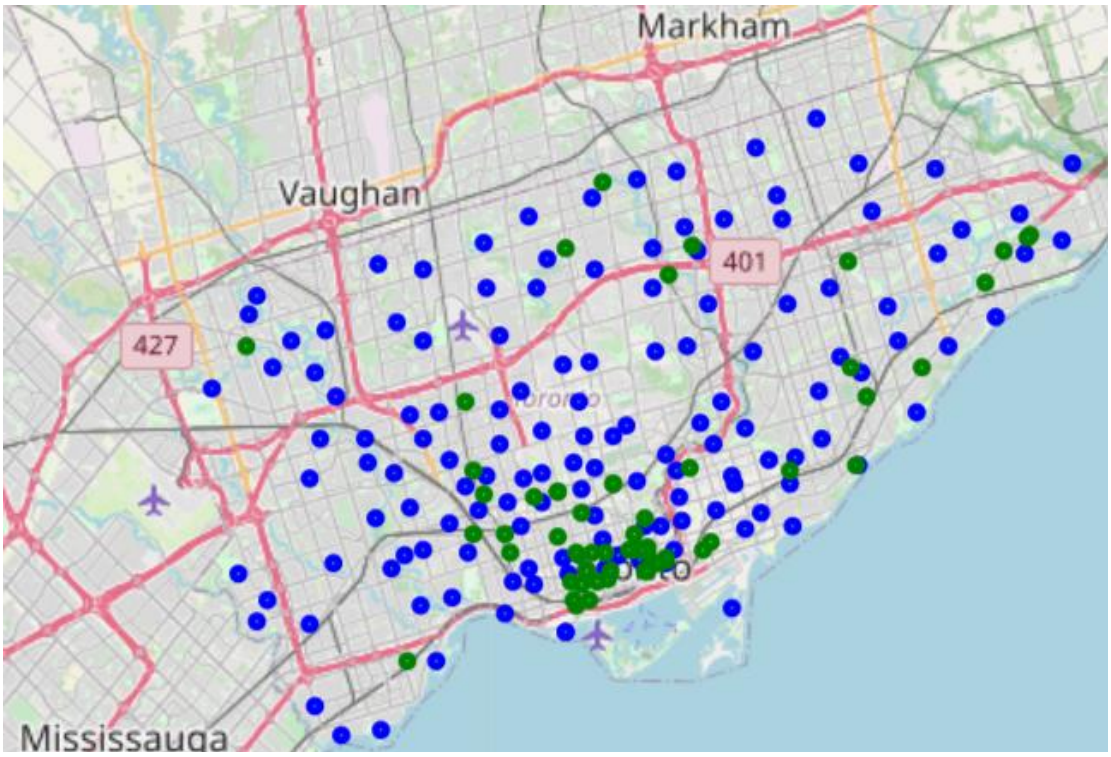

Image 1: A mapping of Toronto's neighbourhoods (blue) and homeless shelters (green)

Image 2: A closer look at the agglomeration in downtown Toronto's area.

A hypothesis can be established, mainly that this type of agglomeration of homeless shelters can be found in areas of low-income and high unemployment. We will determine if this is the case in the following section.

3.2 Socio-economic data

To evaluate the hypothesis stated in the previous section, we will evaluate 3 features and display them into our map.

- Fraction of people living under 10,000 CAD a year
- Fraction of people living over 80,000 CAD a year
- Fraction of people with low-income after taxes (LICO-AT indicator)

As seen in Image 3, we can see that the Baystreet corridor is one of the neighbourhoods with the highest fraction of people living with less than 10,000 CAD a year, somewhat paradoxically, the neighbourhood with the lowest fraction is the waterfront communities, indicating a presence of income inequality.

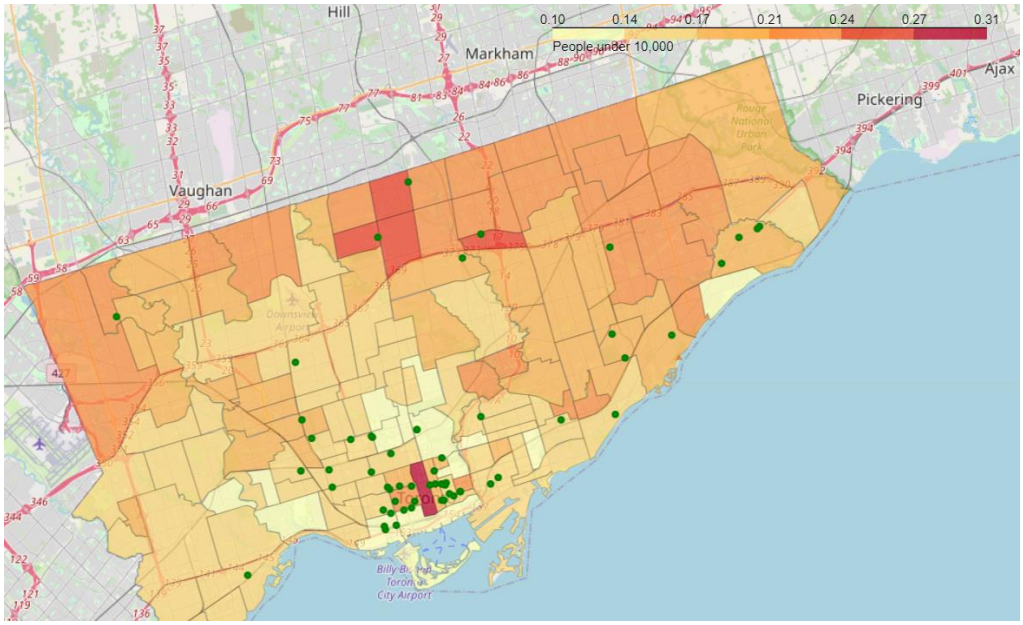

Image 3: Map displaying the fraction of people living under 10,000 CAD a year.

We can see on Image 4 that neighbourhoods north, but still close to downtown Toronto correspond to high-income areas. Our hypothesis seems to be confirmed with Image 5, as we can see that the location where homeless shelters concentrate more strongly, correspond to a group of neighbourhoods with a high indicator of low income (LICO-AT). Meaning there is a high prevalence of families or persons spending 20 percentage points more than average of their after-tax income on essentials (food, shelter and clothing).

In the next section we cluster neighbourhoods according to a k-means method.

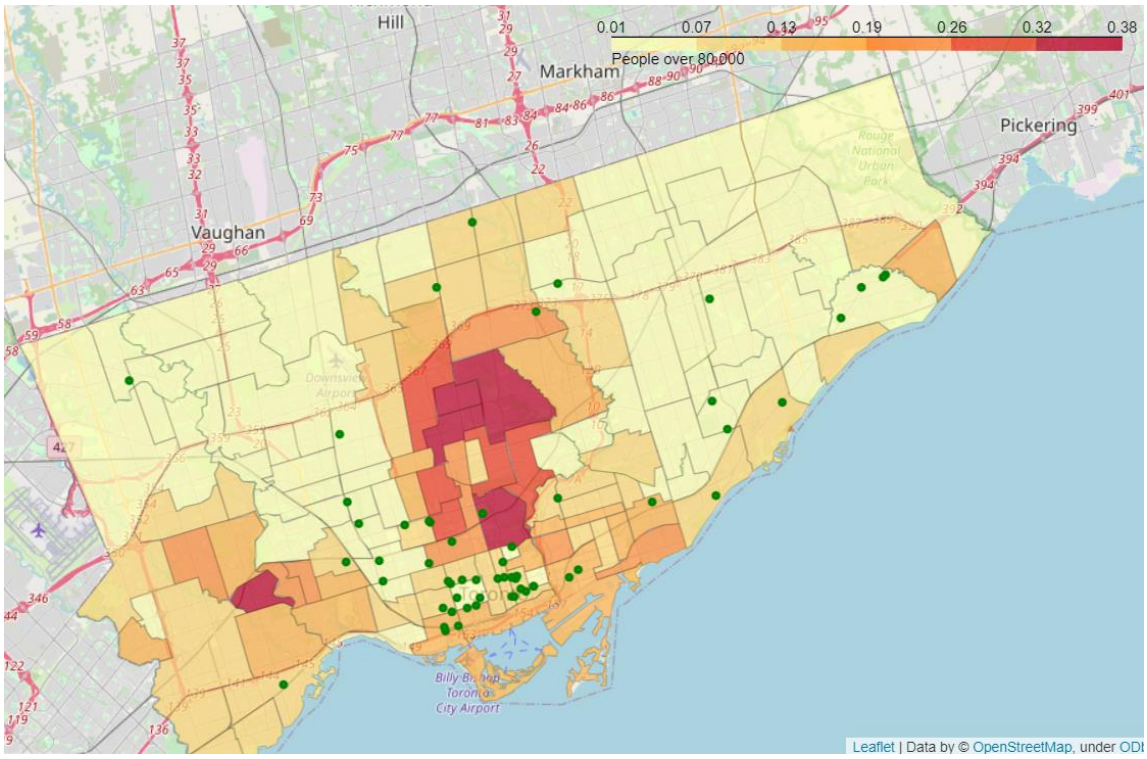

Image 4: Mapping of fraction of people living over 80,000 CAD a year.

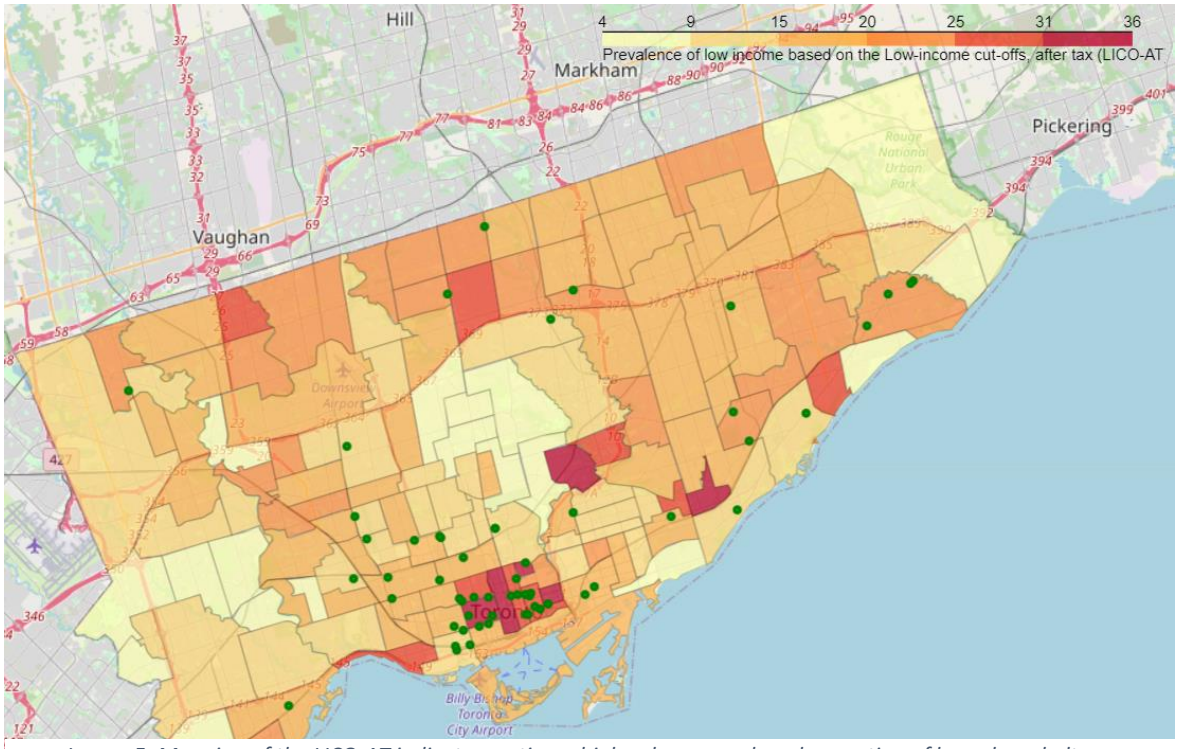

Image 5: Mapping of the LICO-AT indicator, noting a high value around agglomeration of homeless shelters.

3.3 K-means clustering

We apply a k-means clustering method to the neighbourhood's socio-economic dataset. We establish a priori the number of desired clusters, three was selected.

Resulting clusters are shown in Image 6, we find a peculiarity, mainly that the socio-economic dataset has no geographic information (latitude and latitude values), as such, the clustering is not dependent on neighbourhood's positions. Nevertheless, we find that neighbourhood clusters more or less tend to divide the map into "zones", which are entirely dependent on socio-economic information.

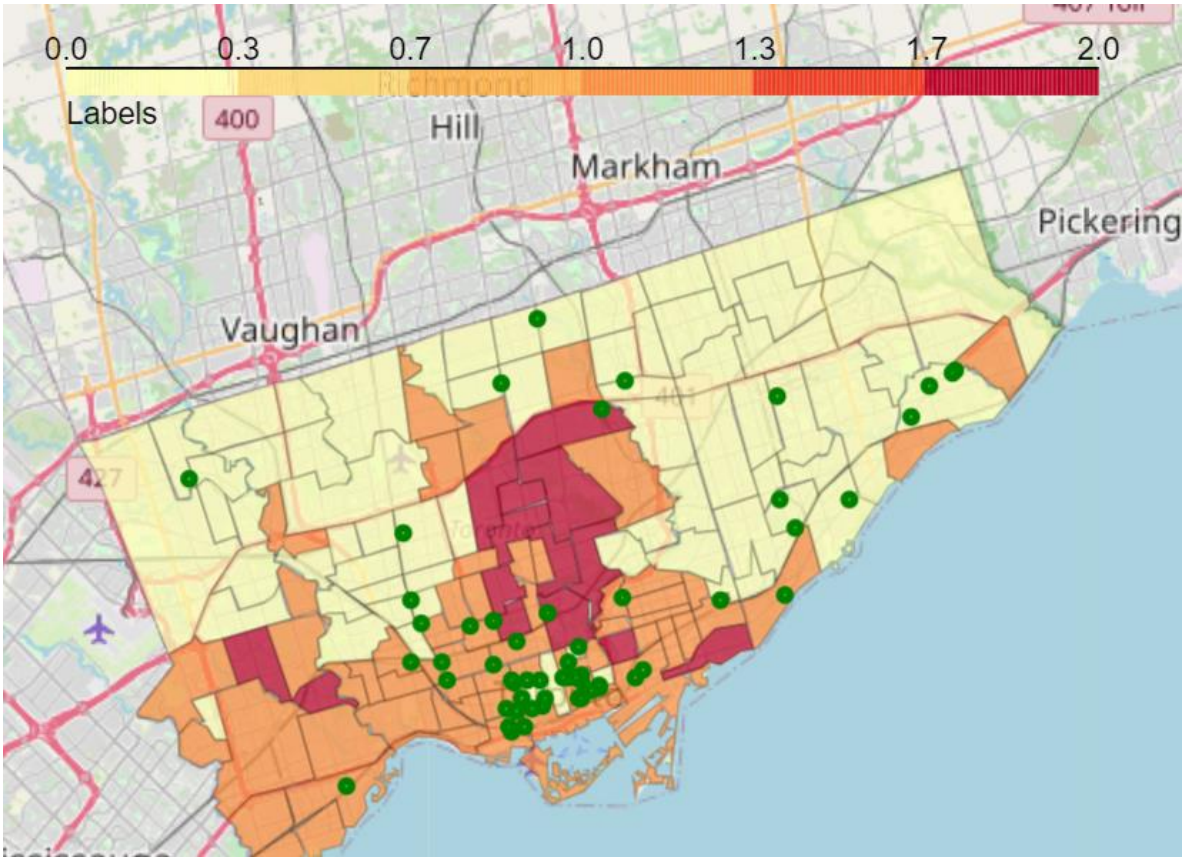

Image 6: K-means clusters (0,1 and 2) and homeless shelters (green dots) displayed in a map.

To differentiate between clusters, we do some exploratory data analysis with box plots. The images presented are not exhaustive, but will give us a good idea as to how the different clusters are classified. Image 7 helps us visualize this:

- Cluster 0: Low income, high unemployment, high prevalence of young people with low income
- Cluster 1: medium income, unemployment similar to cluster 2, medium prevalence of older people with low income
- Cluster 2: High income, unemployment similar to cluster 1, very low prevalence of low income

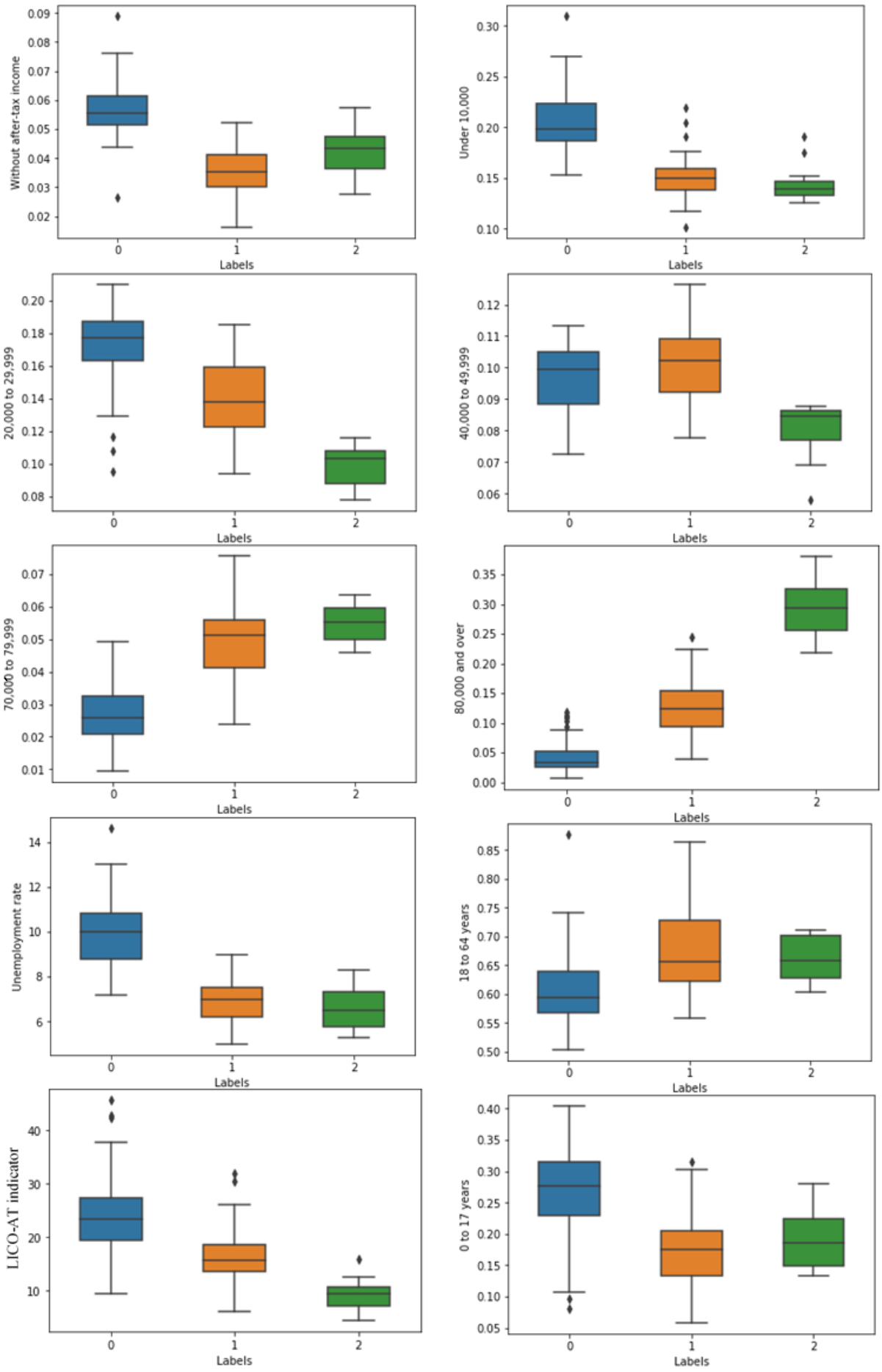

Image 7: K-means clusters box-plots to help visualize differentiation between them.

3.4 Public transit stations

In this section we query Foursquare to obtain a list of public transit stations focused around the different neighbourhoods.

A requirement set by Coursera was to use this API to obtain a dataset, although a better alternative would have been to obtain a more exhaustive dataset, which includes all stations in Toronto. Foursquare has the limitation of returning venues which correspond to a search string, we cannot be sure the string used to obtain the list of venues actually returns all public transit venues stored in Foursquare's data servers.

With this considered, we display the obtained stations in a map, alongside shelter locations and k-means clusters, as seen in Image 8. This data will help us in the selection of the new potential venues for the final section.

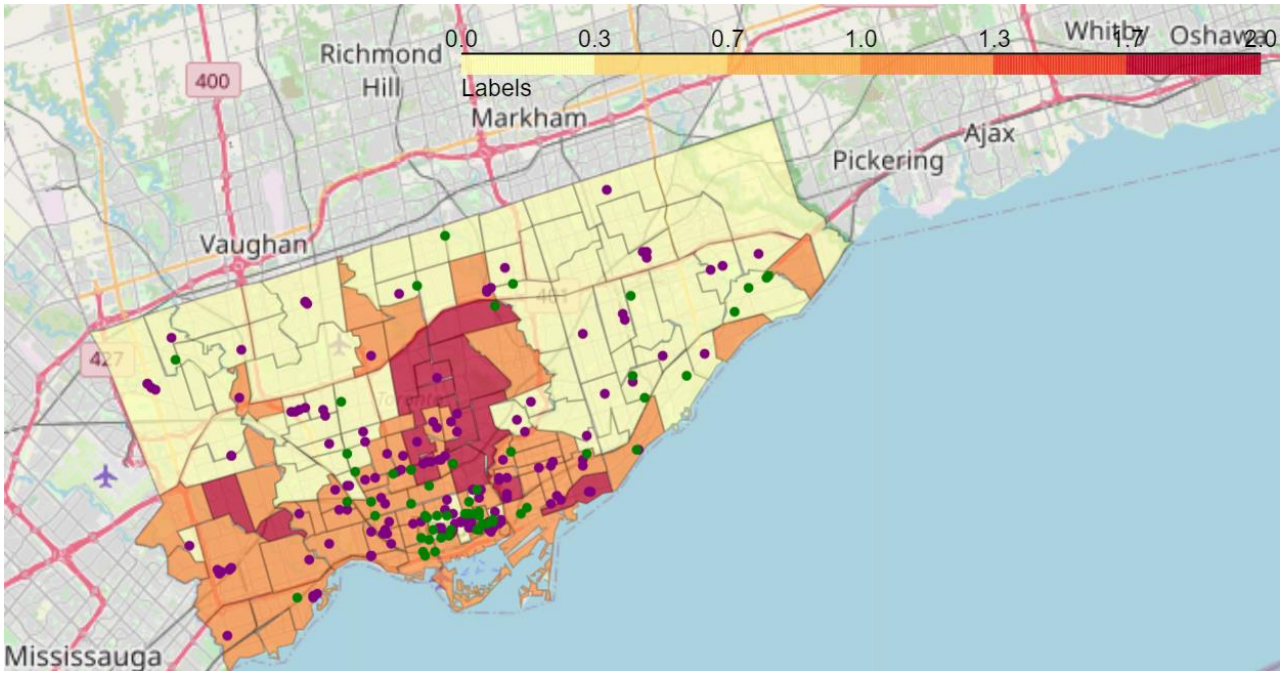

Image 8: Public transit stations (purple dots) mapped along with shelters (green dots) and k-means clusters.

3.5 Selection of new venues

As seen in Image 9, selection of the new venues was based around several key points, mainly the clusters where they are to be located have to be primarily from cluster 0 & 1 (low income, high LICO-AT). The second criteria considered was its closeness to public transit stations, while also considering their distance between shelters already established.

The five new venues selected are located in the following coordinates:

Venue	Latitude	Longitude
1	43.631993	-79.547807
2	43.710946	-79.49523
3	43.653341	-79.434522
4	43.787051	-79.25565
5	43.65275	-79.3980

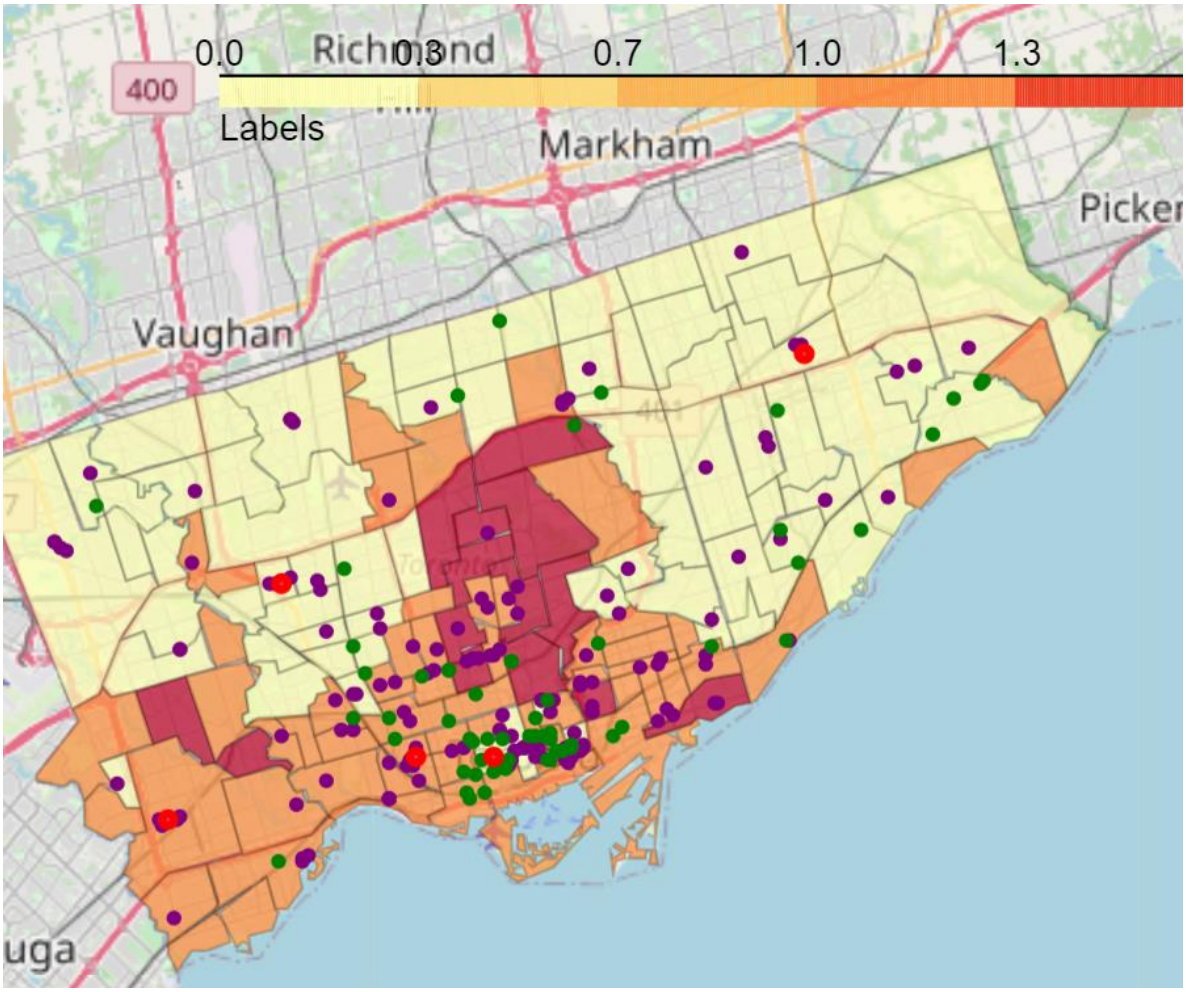

Image 9: Map of new venues (red dots) displayed with public transit (purple dots), shelters (green dots) and clusters.

4. Conclusions

We presented five new venues where homeless shelters could be established, based on closeness to public transit stations, socio-economic data and geographic information. To improve the clustering, features which were not used could be dropped, as to retain only the critical information considered for the analysis. A potential improvement would be to include an optimization method to analytically determine best venue location based on an objective function, this modification would require a more reliable method of obtaining public transit stations, such as a large government provided dataset.

Appendix

Socio-economic features