June 2019

WHAT MAKES A HEALTHY CITY?

Machine learning clustering algorithms combined with Foursequare API massive databases and census data can help us better understand the factors that make cities healthy.

Dor Abelman
Coursera Capstone for the Professional Certificate in Data Science, authored by IBM

What makes a healthy city?

0. Abstract

What makes a healthy city? K-means clustering, a machine learning tool, was used to cluster neighbourhoods into different groups based on their venue properties hosted in the Foursquare API. These clusters were then compared to health indicators from the national census and the City of Toronto's health data portal to see if some clusters had better health indicators than others. Each cluster reported different health outcomes on things such as breast cancer screening to asthma rates. The study was not strong enough to yield recommendation significance due to limitations in the comparison methods that were outside of the scope of the course. Future studies with similar methods can yield promising results that can validate or challenge Public Health epidemiological observations and theory on the associations between neighbourhood properties and health indicators. These methods can also be used to better understand health disparities reported in the literature between reigions.

1. Introduction

What makes a healthy city? In this project we are going to use machine learning tools to cluster different neighbourhoods based on their characteristics in the Foursquare database. We will then compare it to regional health data from Canadian census data and see if there are trends between the health indicators of that neighbourhood and its characteristics from the Foursquare database. We will be able to report at the end of this investigation if certain characteristics in the Foursquare database are generally associated with better or worse health outcomes. Outcomes we can consider are average life expectancy, hospital admissions, rates of diabetes, asthma, high blood pressure, and premature mortality.

This kind of analysis is important because it may help to answer questions regarding health inequalities already readily reported in the literature. For example, see (Hou & Chen, 2003; Van Ingen et al., 2015). The city of Toronto is aware of health inequalities between neighbourhoods and adding light to what may be contributing to them is an important step forward to formulating policies that can aim to improve health outcomes in the communities that need them most by tackling the factors that lead to them at their source.

Shedding more light to the answers to these questions is important in aiding policy makers have better tools at their disposal to make effective decisions. Using machine learning tools and large newly available datasets holds a great opportunity to find answers from large amounts of data in ways that were previously impossible due to lack of data available, computing power, and system architectures that improve calculation efficiency.

2.0 Data

In this analysis we will be using two different kinds of data, and in that, analyses. The first type of data will be geolocation data from the Foursquare database that classifies items in a neighbourhood by their location, name and category. We will group neighborhoods in the same way as they were grouped in the most recent available census. We will then use clustering algorithms to differentiate these neighbourhoods based on their Foursquare attributes and generate figures that display the message in a clear and effective manner.

We will then use the census data to find neighbourhoods that have the worst health outcomes based on indicators identified in the previous section. Finally, we will compare the properties of the neighbourhoods or groups of neighbourhoods with the most severe health outcomes with the ones with the least severe health outcomes. We can use a regression model to show trends in the kinds of attributes presented in Foursquare that are generally associated with healthier or unhealthier neighbourhoods if there are indeed trends present. An example of the census data that can be used can be found here: http://www.torontohealthprofiles.ca/a_dataTables.php?varTab=HPDtbl. We also use data from here: https://www.toronto.ca/city-government/data-research-maps/open-data/open-data-catalogue/health/#ccc4fb43-9c06-8414-32fd-54ab7d5ae422

3.0 Methods

3.1 Background on Machine Learning

Machine learning in Python is powerful tool to understand what data means. With linear and logistic regression models, we can learn about the association between different data points to answer questions. Some examples of this are through using the sigmoid function to sort categorical data and the gradient descent approach to minimize the error. Another method is using a support vector machine to find a hyperplane that can separate data points in upper dimensions. These above are data classification protocols that are supervised, ie, compared to a known dataset to validate. The data can also be clustered in an unsupervised format in which the right answer is not unknown but found. Some examples are through partitioning the data using a k-means nearest neighbour method, sorting the data into hierarchies with a dendrogram, or sorting it based on its density with the DBSCAN approach. Any of these methods can be implemented relatively easily in Python using packages, or in R.

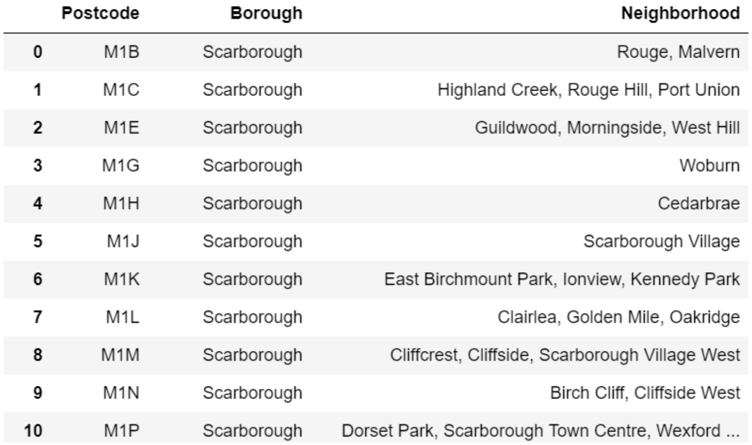
Machine learning is a powerful tool because it can automatically calculate the best parameters of a model to make sense of the data. We just need to pick the model we think would be the most representative of the work, split it into testing and training sets, and then leave it to the model to save us time in finding the most representative formulas that describe it. We can use these results to make predictions that help advance society.

3.2 Project Design

In this project, we first scrape neighbourhood data from the city of Toronto using a scraping code. Following this, we normalize the data, and split it into a training and testing set. We then train the set using a kmeans clustering algorithm to find different groups of neighbourhoods. Following this, we import and normalize census data on health from those neighbourhoods. We sort the data to rank the neighbourhoods from most to least healthy, and then apply our ranks to the clustered groups. We see which neighbourhoods fell in which clusters, and thus, if there were more beneficial health outcomes associated with specific attributes of clusters. This would help to provide a better understanding of the kinds of things that make a healthy neighbourhood.

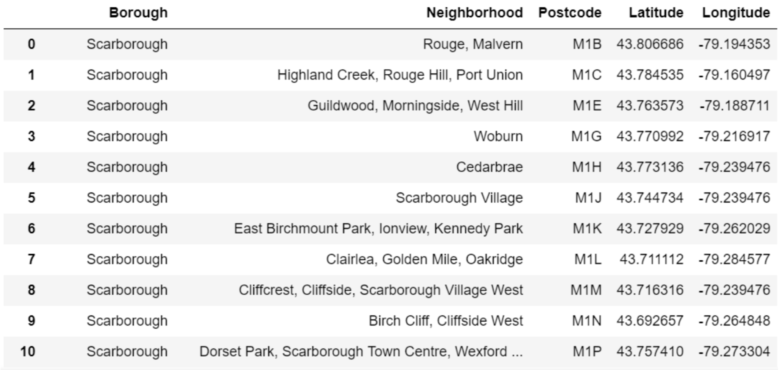
3.3 Steps

1. We start by importing data and sorting it into a Panda's dataframe to get something like this regarding our neighbourhoods:



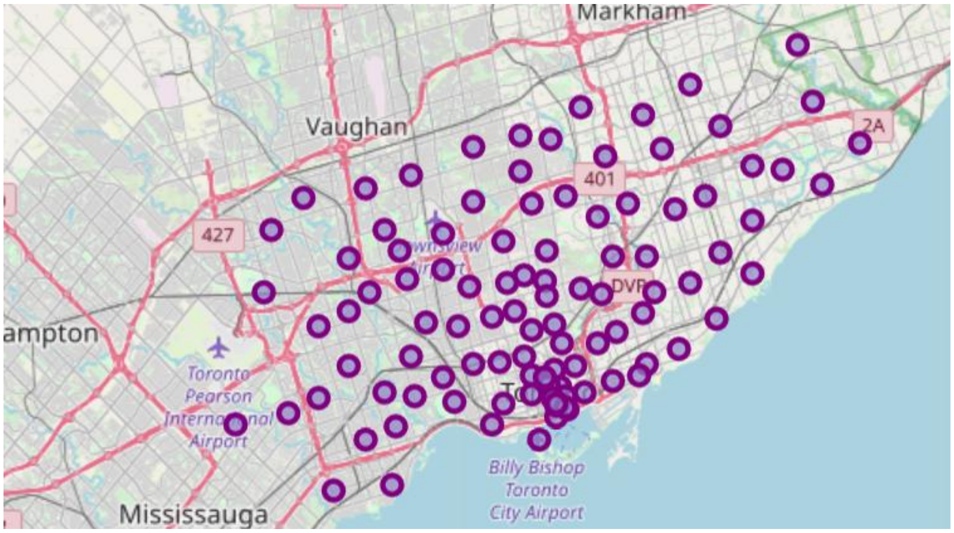
	Postcode	Borough	Neighborhood
0	M1B	Scarborough	Rouge, Malvern
1	M1C	Scarborough	Highland Creek, Rouge Hill, Port Union
2	M1E	Scarborough	Guildwood, Morningside, West Hill
3	M1G	Scarborough	Woburn
4	M1H	Scarborough	Cedarbrae
5	M1J	Scarborough	Scarborough Village
6	M1K	Scarborough	East Birchmount Park, Ionview, Kennedy Park
7	M1L	Scarborough	Clairlea, Golden Mile, Oakridge
8	M1M	Scarborough	Cliffcrest, Cliffside, Scarborough Village West
9	M1N	Scarborough	Birch Cliff, Cliffside West
10	M1P	Scarborough	Dorset Park, Scarborough Town Centre, Wexford ...

2. We then merge it with postal code data:

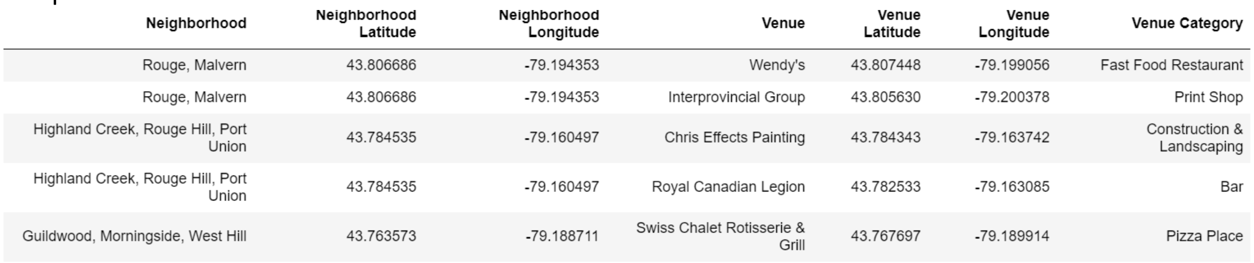


	Borough	Neighborhood	Postcode	Latitude	Longitude
0	Scarborough	Rouge, Malvern	M1B	43.806686	-79.194353
1	Scarborough	Highland Creek, Rouge Hill, Port Union	M1C	43.784535	-79.160497
2	Scarborough	Guildwood, Morningside, West Hill	M1E	43.763573	-79.188711
3	Scarborough	Woburn	M1G	43.770992	-79.216917
4	Scarborough	Cedarbrae	M1H	43.773136	-79.239476
5	Scarborough	Scarborough Village	M1J	43.744734	-79.239476
6	Scarborough	East Birchmount Park, Ionview, Kennedy Park	M1K	43.727929	-79.262029
7	Scarborough	Clairlea, Golden Mile, Oakridge	M1L	43.711112	-79.284577
8	Scarborough	Cliffcrest, Cliffside, Scarborough Village West	M1M	43.716316	-79.239476
9	Scarborough	Birch Cliff, Cliffside West	M1N	43.692657	-79.264848
10	Scarborough	Dorset Park, Scarborough Town Centre, Wexford ...	M1P	43.757410	-79.273304

3. And make a map:

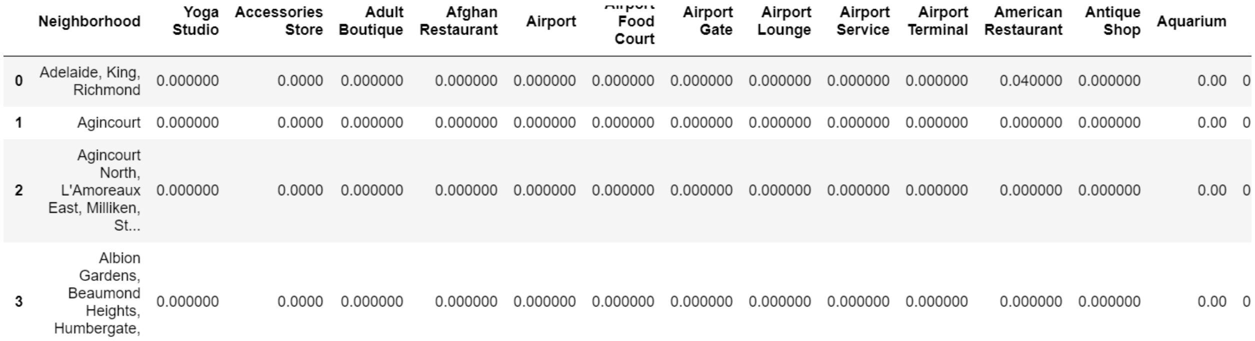

Markham
Vaughan
401
2A
427
DVP
ampton
Toronto Pearson International Airport
Billy Bishop Toronto City Airport
Mississauga

4. Next we import thousands of neighbourhood datapoints from the Foursquare API and merge it with our previous table:



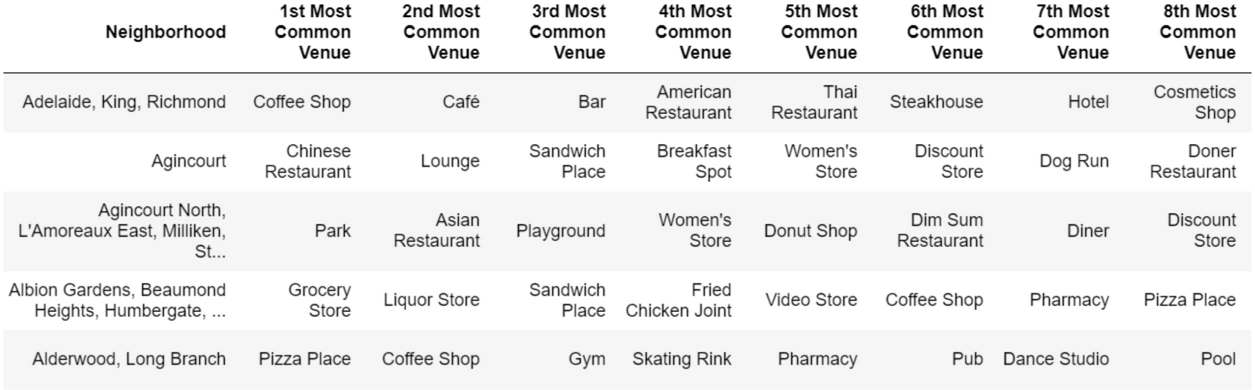
Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Rouge, Malvern	43.806686	-79.194353	Wendy's	43.807448	-79.199056	Fast Food Restaurant
Rouge, Malvern	43.806686	-79.194353	Interprovincial Group	43.805630	-79.200378	Print Shop
Highland Creek, Rouge Hill, Port Union	43.784535	-79.160497	Chris Effects Painting	43.784343	-79.163742	Construction & Landscaping
Highland Creek, Rouge Hill, Port Union	43.784535	-79.160497	Royal Canadian Legion	43.782533	-79.163085	Bar
Guildwood, Morningside, West Hill	43.763573	-79.188711	Swiss Chalet Rotisserie & Grill	43.767697	-79.189914	Pizza Place

5. Before dummy-coding it by frequency:



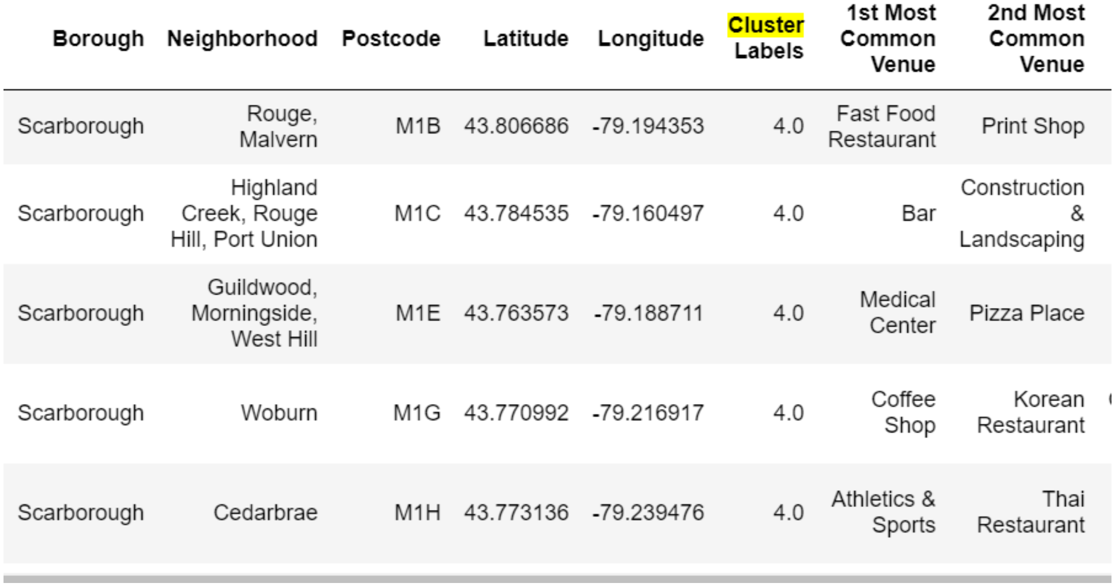
	Neighborhood	Yoga Studio	Accessories Store	Adult Boutique	Afghan Restaurant	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Antique Shop	Aquarium	
0	Adelaide, King, Richmond	0.000000	0.0000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.040000	0.000000	0.00	0
1	Agincourt	0.000000	0.0000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0
2	Agincourt North, L'Amoreaux East, Milliken, St...	0.000000	0.0000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0
3	Albion Gardens, Beaumond Heights, Humbergate,	0.000000	0.0000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0

6. We can find the top 10 most common venues in each neighbourhood:



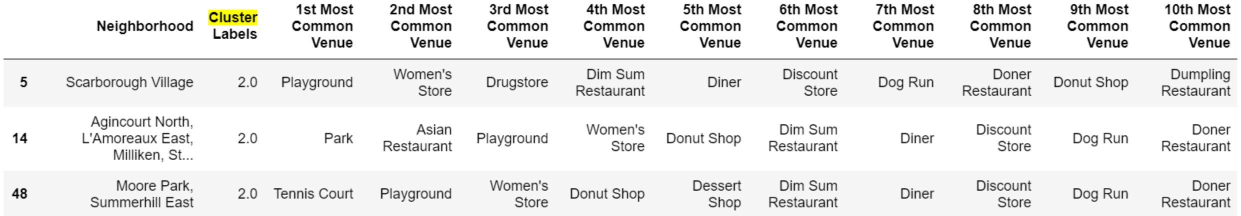
Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue
Adelaide, King, Richmond	Coffee Shop	Café	Bar	American Restaurant	Thai Restaurant	Steakhouse	Hotel	Cosmetics Shop
Agincourt	Chinese Restaurant	Lounge	Sandwich Place	Breakfast Spot	Women's Store	Discount Store	Dog Run	Doner Restaurant
Agincourt North, L'Amoreaux East, Milliken, St...	Park	Asian Restaurant	Playground	Women's Store	Donut Shop	Dim Sum Restaurant	Diner	Discount Store
Albion Gardens, Beaumond Heights, Humbergate, ...	Grocery Store	Liquor Store	Sandwich Place	Fried Chicken Joint	Video Store	Coffee Shop	Pharmacy	Pizza Place
Alderwood, Long Branch	Pizza Place	Coffee Shop	Gym	Skating Rink	Pharmacy	Pub	Dance Studio	Pool

7. And cluster neighbourhoods based on this information:



Borough	Neighborhood	Postcode	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue
Scarborough	Rouge, Malvern	M1B	43.806686	-79.194353	4.0	Fast Food Restaurant	Print Shop
Scarborough	Highland Creek, Rouge Hill, Port Union	M1C	43.784535	-79.160497	4.0	Bar	Construction & Landscaping
Scarborough	Guildwood, Morningside, West Hill	M1E	43.763573	-79.188711	4.0	Medical Center	Pizza Place
Scarborough	Woburn	M1G	43.770992	-79.216917	4.0	Coffee Shop	Korean Restaurant
Scarborough	Cedarbrae	M1H	43.773136	-79.239476	4.0	Athletics & Sports	Thai Restaurant

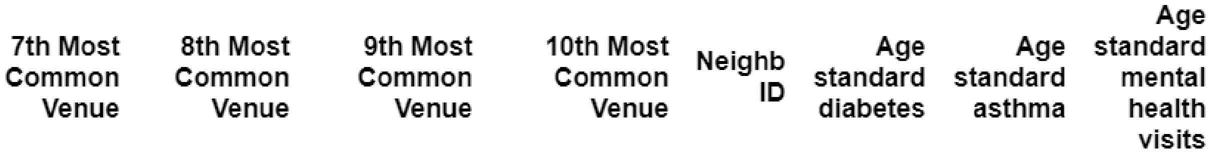
8. And finally, we can learn from each of the clusters:



	Neighborhood	Cluster Labels	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
5	Scarborough Village	2.0	Playground	Women's Store	Drugstore	Dim Sum Restaurant	Diner	Discount Store	Dog Run	Doner Restaurant	Donut Shop	Dumpling Restaurant
14	Agincourt North, L'Amoreaux East, Milliken, St...	2.0	Park	Asian Restaurant	Playground	Women's Store	Donut Shop	Dim Sum Restaurant	Diner	Discount Store	Dog Run	Doner Restaurant
48	Moore Park, Summerhill East	2.0	Tennis Court	Playground	Women's Store	Donut Shop	Dessert Shop	Dim Sum Restaurant	Diner	Discount Store	Dog Run	Doner Restaurant

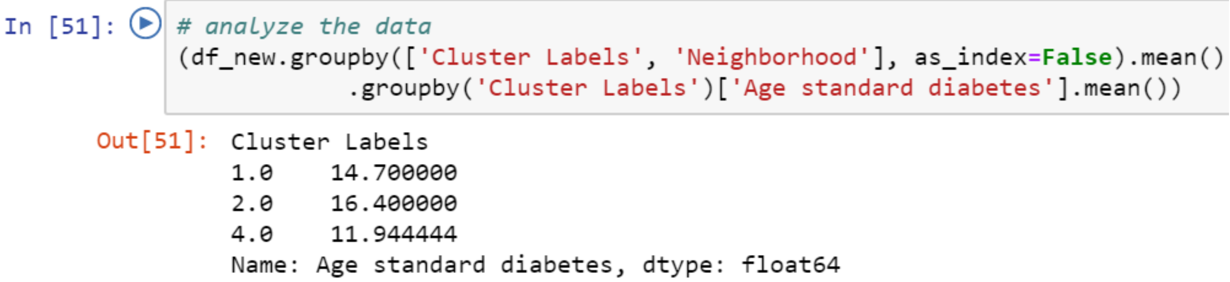
For example, it seems that in this neighbourhood, physical activity regions like playground, parks or tennis courts were the most common.

9. Next we imported information from the census and merged it with our table. We ended up with a new table that also had these columns at the end of the table:



7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue	Neighb ID	Age standard diabetes	Age standard asthma	Age standard mental health visits

10. Finally we were able to calculate the mean of our health indicators, such as age-standardized diabetes, between our clusters:

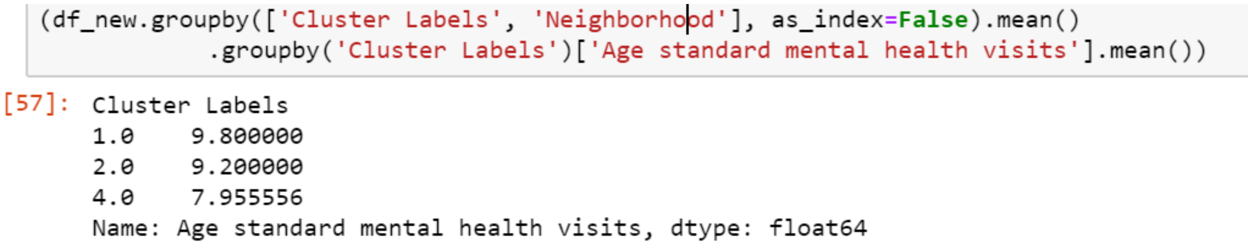


```
In [51]:  # analyze the data
          (df_new.groupby(['Cluster Labels', 'Neighborhood'], as_index=False).mean()
                  .groupby('Cluster Labels')['Age standard diabetes'].mean())

Out[51]: Cluster Labels
         1.0     14.700000
         2.0     16.400000
         4.0     11.944444
         Name: Age standard diabetes, dtype: float64
```

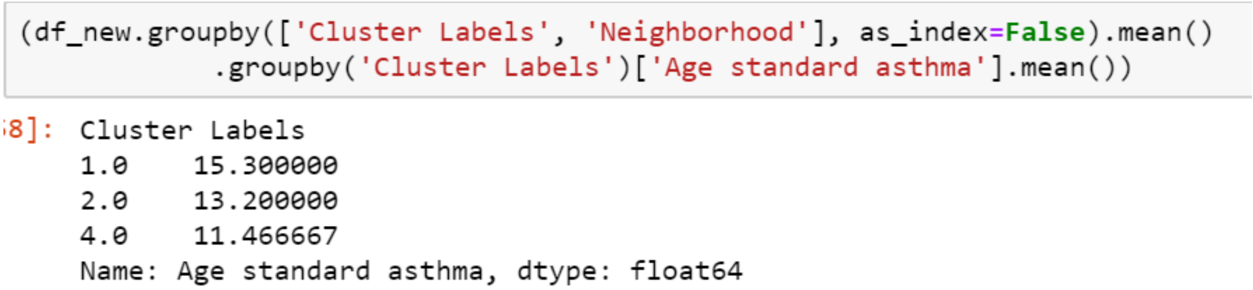
11. The problem that now arose was that not all the clusters were showing. This is because some of the neighbourhood names were different between the two files. To solve this problem, we looked for a

new census with neighbourhood classifiers that were the same as the ones we used for foursquare API. But before we did that, we calculated the cluster data for age-standardized mental health visit rates, and age-standardized asthma rates:



```
(df_new.groupby(['Cluster Labels', 'Neighborhood'], as_index=False).mean()
            .groupby('Cluster Labels')['Age standard mental health visits'].mean())
```

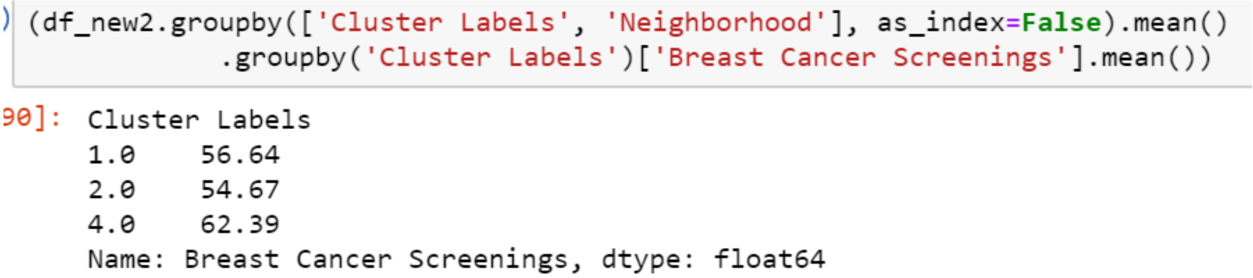
```
[57]: Cluster Labels
      1.0    9.800000
      2.0    9.200000
      4.0    7.955556
      Name: Age standard mental health visits, dtype: float64
```



```
(df_new.groupby(['Cluster Labels', 'Neighborhood'], as_index=False).mean()
            .groupby('Cluster Labels')['Age standard asthma'].mean())
```

```
8]: Cluster Labels
    1.0    15.300000
    2.0    13.200000
    4.0    11.466667
    Name: Age standard asthma, dtype: float64
```

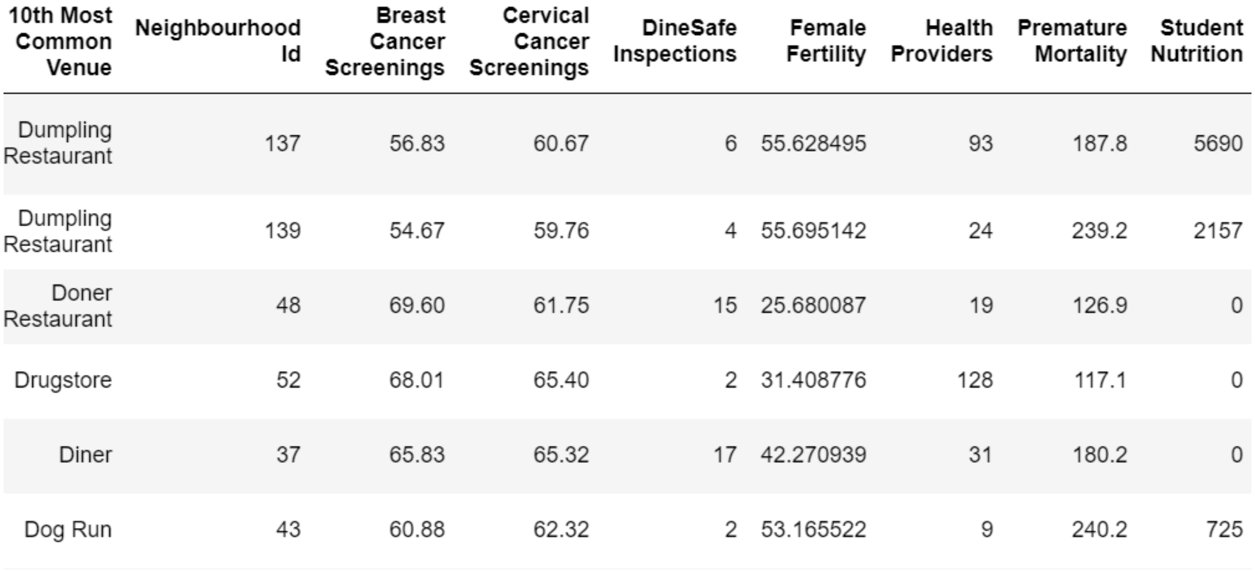
12. What's interesting to note at this point is that there are large differences between cluster groups, these conditions are not evenly distributed throughout the society. To learn about this further, we took data from a new source, the city of Toronto's health data catalogue and included it in our analysis: https://www.toronto.ca/city-government/data-research-maps/open-data/open-data-catalogue/health/#ccc4fb43-9c06-8414-32fd-54ab7d5ae422
13. This gave us new results such as breast cancer screenings:



```
(df_new2.groupby(['Cluster Labels', 'Neighborhood'], as_index=False).mean()
            .groupby('Cluster Labels')['Breast Cancer Screenings'].mean())
```

```
90]: Cluster Labels
     1.0    56.64
     2.0    54.67
     4.0    62.39
     Name: Breast Cancer Screenings, dtype: float64
```

14. Other data we retrieved includes:

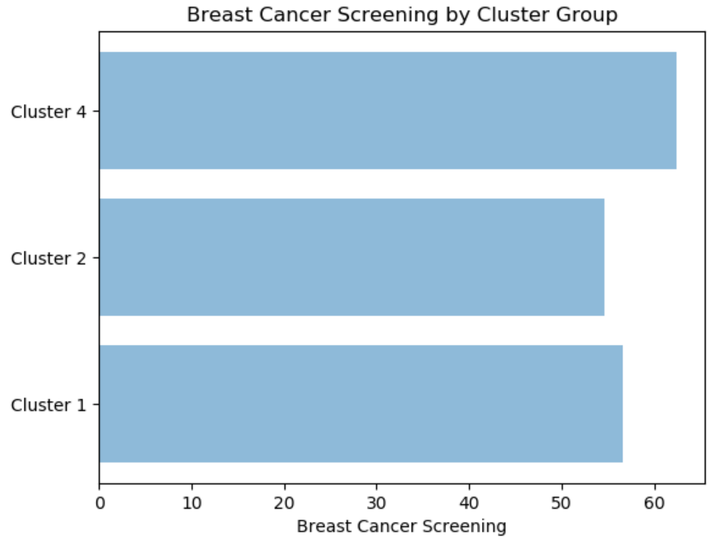


10th Most Common Venue	Neighbourhood Id	Breast Cancer Screenings	Cervical Cancer Screenings	DineSafe Inspections	Female Fertility	Health Providers	Premature Mortality	Student Nutrition
Dumpling Restaurant	137	56.83	60.67	6	55.628495	93	187.8	5690
Dumpling Restaurant	139	54.67	59.76	4	55.695142	24	239.2	2157
Doner Restaurant	48	69.60	61.75	15	25.680087	19	126.9	0
Drugstore	52	68.01	65.40	2	31.408776	128	117.1	0
Diner	37	65.83	65.32	17	42.270939	31	180.2	0
Dog Run	43	60.88	62.32	2	53.165522	9	240.2	725

15. We were still unable to have all the clusters labelled due to disagreement between neighbourhood names in the Toronto census and table that we used. We do have sufficient data, however, to start comparing the difference in health indicators between clusters. Now we use the matplotlib function to start learning about the clusters.

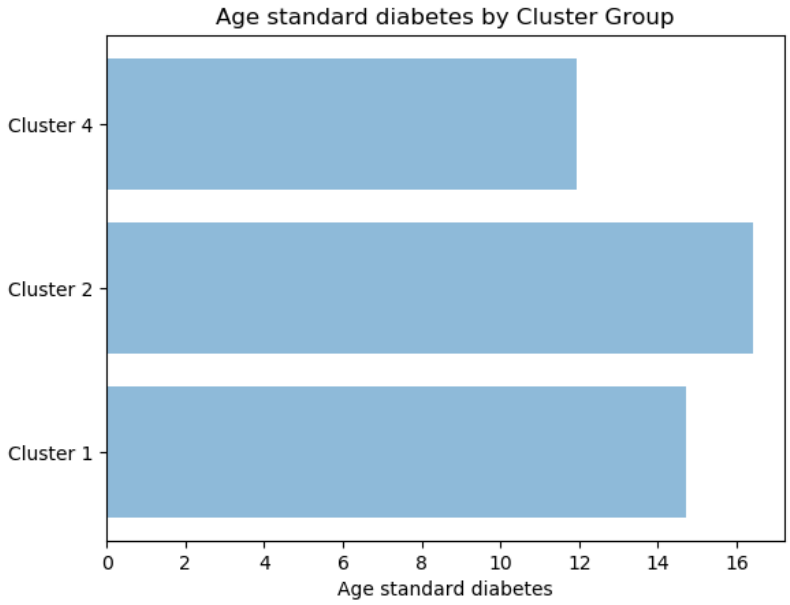
4.0 Results

4.1 Difference in breast cancer screening rates by cluster group



The results show that the inhabitants of cluster 4 had the greatest amount of breast cancer screening and that the inhabitants of cluster 2 had the least amount of this practice. Breast cancer screening is a positive health indicator, with more screening deemed in this report as better.

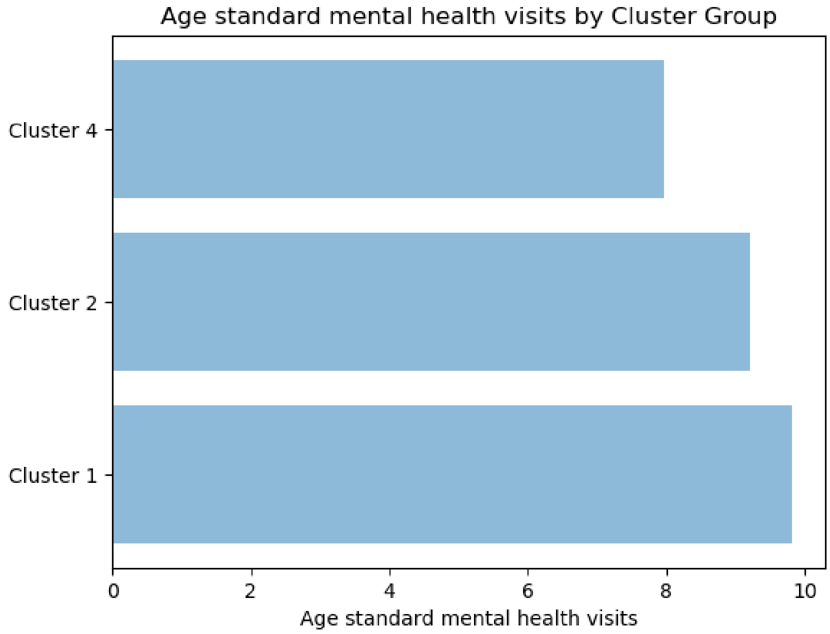
4.2 Difference in age standard diabetes by cluster group



The results show that the inhabitants of cluster 2 had the greatest amount diabetes, age and population size standardized and that the inhabitants of cluster 4 had the least amount. Diabetes is a negative health indicator, with more deemed worse in this report.

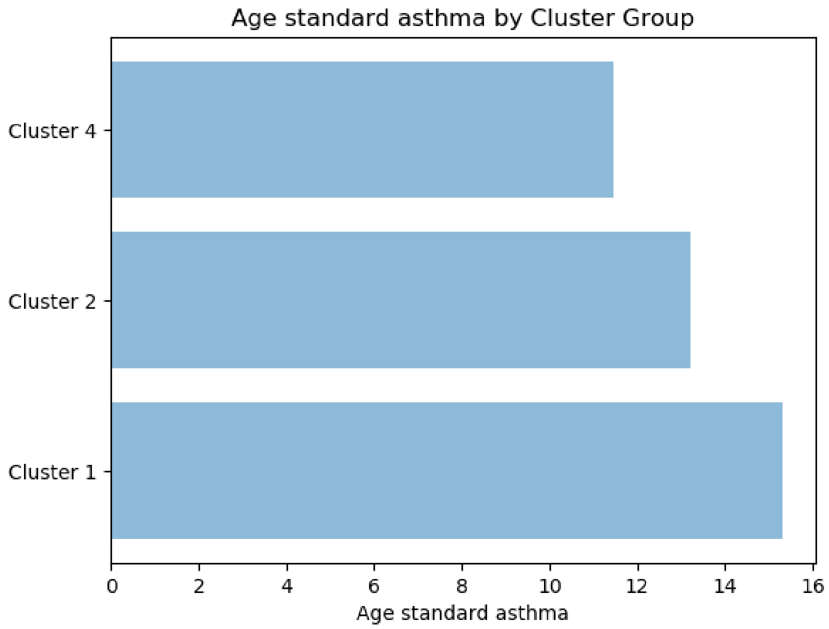
This is strange and interesting as this is the opposite of the previous figure. It seems that rates of diabetes are inversely proportional to rates of breast cancer screening. What in this cluster was different than the others based on the foursquare API data? The discussion will discuss this.

4.3 Difference in age standard mental health visits by cluster group



Cluster 1 had the greatest amount of mental health visits in healthcare centers, age and population-size standardized. Cluster 4 had the least. Mental health visits are considered a negative health factor in this report.

4.4 Difference in age standard asthma by cluster group



Cluster 1 had the greatest amount of cases of asthma reported, age and population-size standardized. Cluster 4 had the least. Mental health visits are considered a negative health factor in this report.

5.0 Discussion

So, what makes a healthy city? In this report we clustered neighbourhoods into groups, compared those groups with census health indicators, and then compared the health indicators to the groups that were separated based on the foursquare API data. Similar clusters should therefore have similar venue properties. Understanding the difference between these venue properties can help us understand why some clusters have different health outcomes than others.

The most common venue in cluster 1 was a park, for all but two cases. The most common venue in cluster 2 was an outdoor sporting area, such as a tennis court. In cluster 4, it was some sort of restaurant, including cafes. It is possible then that people in cluster 4 spend more time indoors or inactive than in the other clusters. There were no other common differences easily observable between cluster groups after the most common cluster. More clearly differentiated clusters in future studies could demonstrate clearer if there are trends between types of venues present and health indicators. There is not enough power in the venues returned from the foursquare API to make such conclusions in this study.

The largest limitations of the ability of this study to make conclusions is the discrepancy between neighbourhood names in the Foursquare API, national census data neighbourhood classifiers, and city of Toronto neighbourhood classifiers. They just don't all use the same names. Finding a way to integrate these names would yield more opportunities to compare between cluster values on health indicators.

Another limitation is the challenge in clearly separating different neighbourhood properties other than sorting their most and least popular venues. The type of regression model that would be required to do this effectively was outside the scope and teachings of this module. However, the concept of clustering neighbourhoods and comparing health indicators to them, holds much potential for future studies on the topic and can help urban planners and policy makers make decisions that improve health outcomes.

Health outcomes are becoming increasingly predictable with advents of data science, innovations in the personal and social determinants of health, and basic research on the cause of diseases and properties of the things that govern them. Public health has shed much light on the importance of outdoor play places, social cohesion, and physical activity in maintaining health and wellbeing (American Public Health Association, 2019; Knibbe, Biddiss, Gladstone, & McPherson, 2017; Mandigo, Francis, Lodewyk, & Lopez, 2009; Pedersen & Saltin, 2015; Public Health Agency of Canada, 2012). Studies such as this with more advanced tools can shine more light onto previous works and better validate them in the case of current health indicators and settings.

6.0 Conclusion

This study validates that there are much health indicator differences between neighbourhoods in Toronto. These differences have been reported in literature (City of Toronto, 2011; Hou & Chen, 2003; Van Ingen et al., 2015). Using Foursquare API can help better understand the differences between neighbourhoods based on their venue properties. This could help policy makers and researchers be better able to understand if there are significantly correlated associations between venue types, their densities, and health disparities. Future studies will shed more light onto this that either validates or challenges epidemiological evidence and public health theories on the matter.

7.0 Sources

American Public Health Association. (2019). What is Public Health? Retrieved June 3, 2019, from https://www.apha.org/what-is-public-health

City of Toronto. (2011). Health - Data Catalogue. *City of Toronto Data Catalogue*. Retrieved from https://www.toronto.ca/city-government/data-research-maps/open-data/open-data-catalogue/health/#ccc4fb43-9c06-8414-32fd-54ab7d5ae422

Hou, F., & Chen, J. (2003). Neighbourhood low income, income inequality and health in Toronto. *Health Reports*, *14*(2), 21–34. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12658862

Knibbe, T. J., Biddiss, E., Gladstone, B., & McPherson, A. C. (2017). Characterizing socially supportive environments relating to physical activity participation for young people with physical disabilities. *Developmental Neurorehabilitation*, *20*(5), 294–300. https://doi.org/10.1080/17518423.2016.1211190

Mandigo, J., Francis, N., Lodewyk, K., & Lopez, R. (2009). Physical Literacy for Educators. *Sport Research Intelligence Sportive*, *4*(2), 27–30. https://doi.org/http://books.scholarsportal.info/viewdoc.html?id=678422

Pedersen, B. K., & Saltin, B. (2015). Exercise as medicine - evidence for prescribing exercise as therapy in 26 different chronic diseases. *Scandinavian Journal of Medicine & Science in Sports*, *25*(S3), 1–72. https://doi.org/10.1111/sms.12581

Public Health Agency of Canada. (2012). Physical Activity Tips for Adults (18-64 years) - Tips to Get Active - Physical Activity - Public Health Agency of Canada. Retrieved March 23, 2017, from http://www.phac-aspc.gc.ca/hp-ps/hl-mvs/pa-ap/07paap-eng.php

Van Ingen, T., Khandor, E., Fleiszer, P., Mckeown, D., Houston, J., Fordham, J., … Public Health, T. (2015).

The Unequal City 2015: Income and Health Inequities in Toronto. Toronto Public Health. Retrieved from http://www.toronto.ca/health/