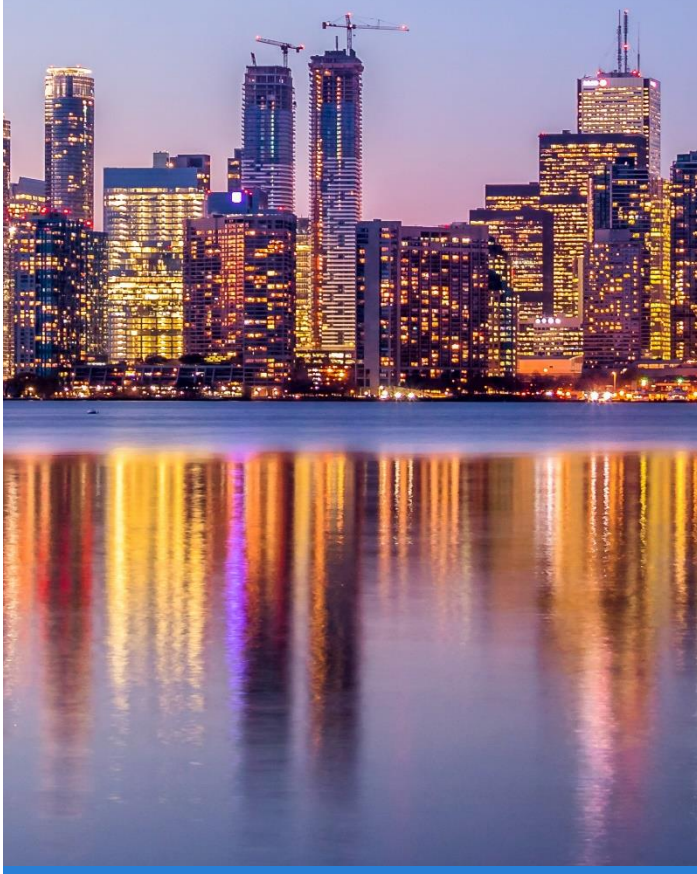
Examining Venues in Ontario Cities



JUNE 2

Authored by: Jacob Sharples

Clustering Ontario Cities by Venues

Introduction

The purpose of this research is to attempt to cluster cities in Ontario, Canada by categories of its most popular venue types. In doing so, we can distinguish cities by certain traits. The target audience for this report could be new businesses looking to enter cities or homeowners looking to settle somewhere. The problem entails the gathering of cities in Ontario with their venues and applying a machine learning algorithm to cluster similar cities. As a result, different types of cities can be classified (i.e. restaurant, bar cities) and generate insight into what cities in Ontario are based upon.

Data

The data is split into three parts. Firstly, the list of cities in Ontario is web scraped from the Wikipedia page: (https://en.wikipedia.org/wiki/List_of_cities_in_Ontario). This enables the creation of a DataFrame with all the cities. Secondly, using the previously created DataFrame, each city is located using the MapQuest API to add the longitude and latitude values to each city. Lastly, using the FourSquare API, each city has its venues added with its category noted. In doing so, the top five number of venue categories are then clustered in a K-means clustering algorithm. For example, the city of Toronto will have its closest 100 venues extracted using FourSquare then sorted by the five most popular venue categories.

Methodology

To analyze and cluster cities in Ontario based on venue category, I used the unsupervised machine learning method k-means clustering. The number of partitions

was chosen to be five. However, the data first had to be cleaned and organized in order to apply this clustering algorithm.

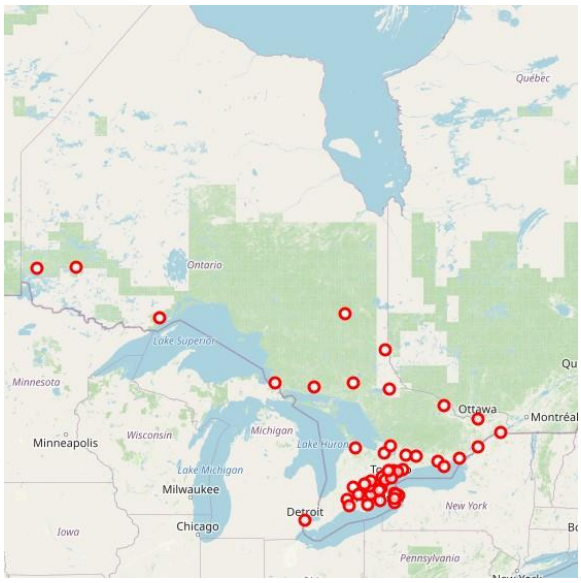
The list of cities is gathered through web scrapping the Wikipedia page containing all registered Ontario cities. Then, the longitude and latitude are obtained through each city with the MapQuest API. This allows retrieval of each city's venues given the location data of the city. From this, more data cleaning is applied as cities that have less than five total venues through FourSquare are removed from the DataFrame.

The venues are then categorized by one-hot encoding and tallied. As a result, each venue category for each city displays a proportion out of total venues on FourSquare. Only the top five in each city are kept.

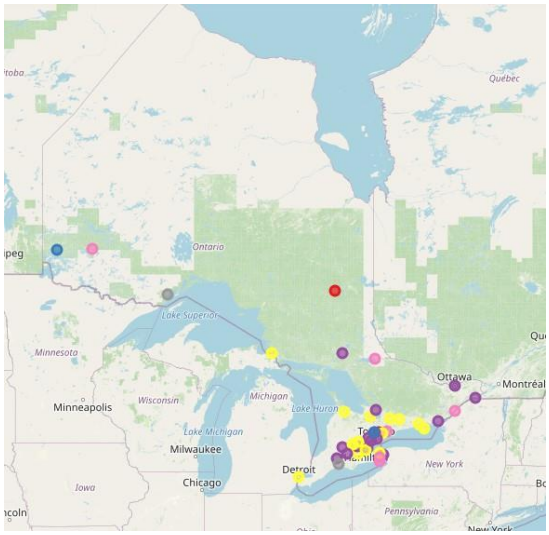
Finally, the k-means clustering algorithm is used where $k = 6$ and the results are displayed using a Folium map where the color of dot represents classification into a cluster.

Results/Discussion

The total number of cities used is 42 and is displayed below.

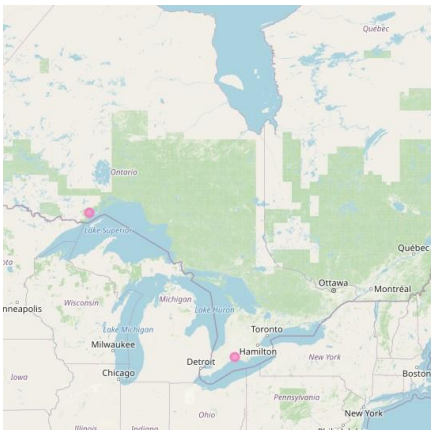


Here the cluster of cities are introduced and are categorized by their most common venues. Below is the map of the cities, each with their color representing the cluster that they are in.



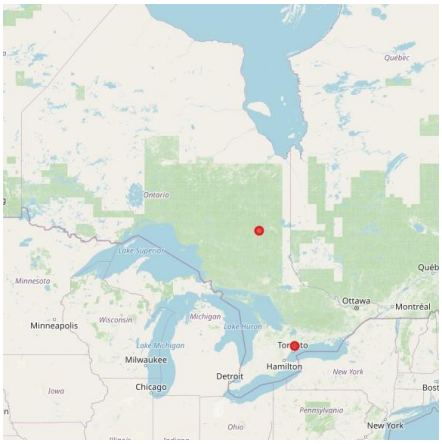
Cluster 0: Pizza Cities

These two cities both have Pizza Place as their most common venue type. This is interesting given that one is a northern Ontario city and the other is a Southern Ontario city, showing that closeness is not a factor.



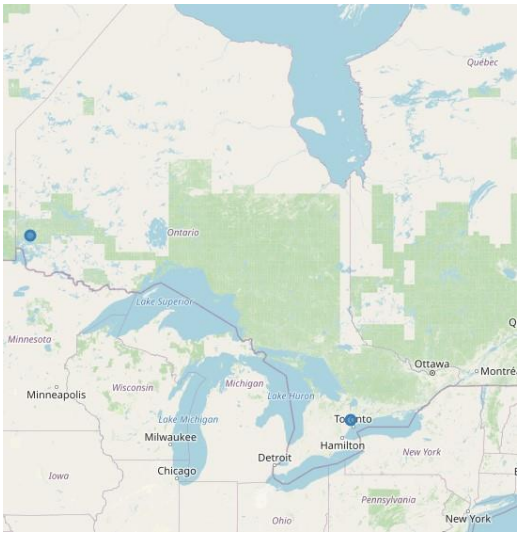
Cluster 1: Food Cities

This cluster also has two cities with Italian Restaurant and Ice Cream Shop as the common venues between them. Generalizing this cluster to a similar trait seems to run along as Food Cities.



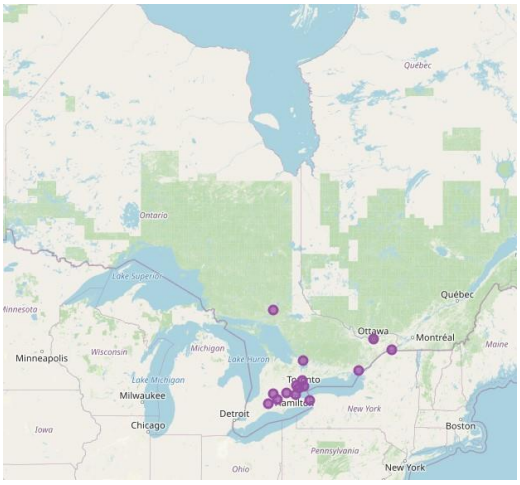
Cluster 2: Outlier Cities

These cities (Kenora and Vaughan) don't seem to have anything in common. As well, given their locations don't suggest any underlying relationship. Given this information, it is reasonable to classify these cities as outliers in this study.



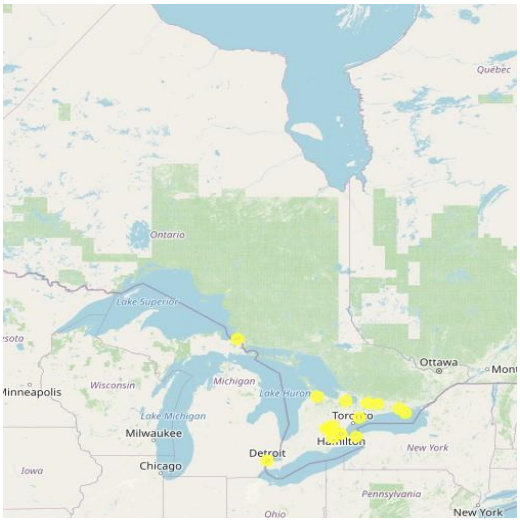
Cluster 3: Coffee Shop Cities

Coffee Shops is the most apt description to categorize these cities which include most cities in the Greater Toronto Area.



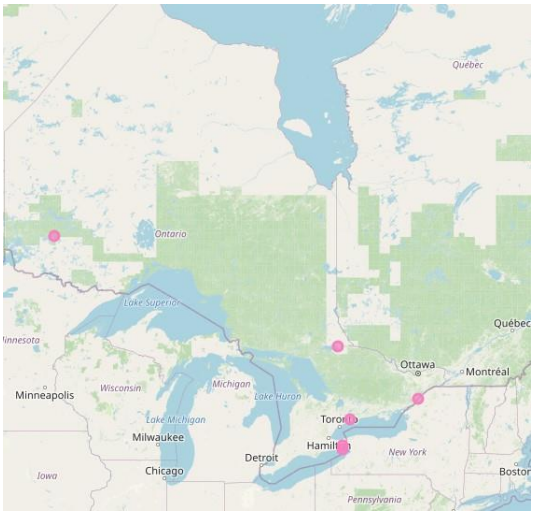
Cluster 4: Café and Pub/Bar Cities

These cities are odd to generalize as Café tends to be the popular venue category but it is also littered with Pub and Bars as well. A good description could be that they tend to be related to pleasure and leisure drinking.



Cluster 5: Restaurant Cities

Sandwich places, Indian restaurants, and Thai restaurants are the most common venues between each city.



Conclusion

Using 42 cities across Ontario and FourSquare data with it, the cities are able to be clustered into 6 clusters using the k-means clustering algorithm. We found that for the most part, the cities are able to clustered fairly well with an underlying common venue category. Insight can be generated from this as new businesses or people looking to move can identify a potential city with the trait that has been generated by this clustering. Although two cities were outliers, the analysis suffers from a lack of data from FourSquare here. Most cities here have limited venues supported by the FourSquare API so ranking them by the top five suffers from it. Coupled with the fact that there is a limited number of cities in Ontario, clustering such a little number of cities may harm this analysis as well. A solution for future analysis may be to analyze municipalities in Ontario instead of cities. However, this research still provides some answers to common venue traits found in cities across Ontario.