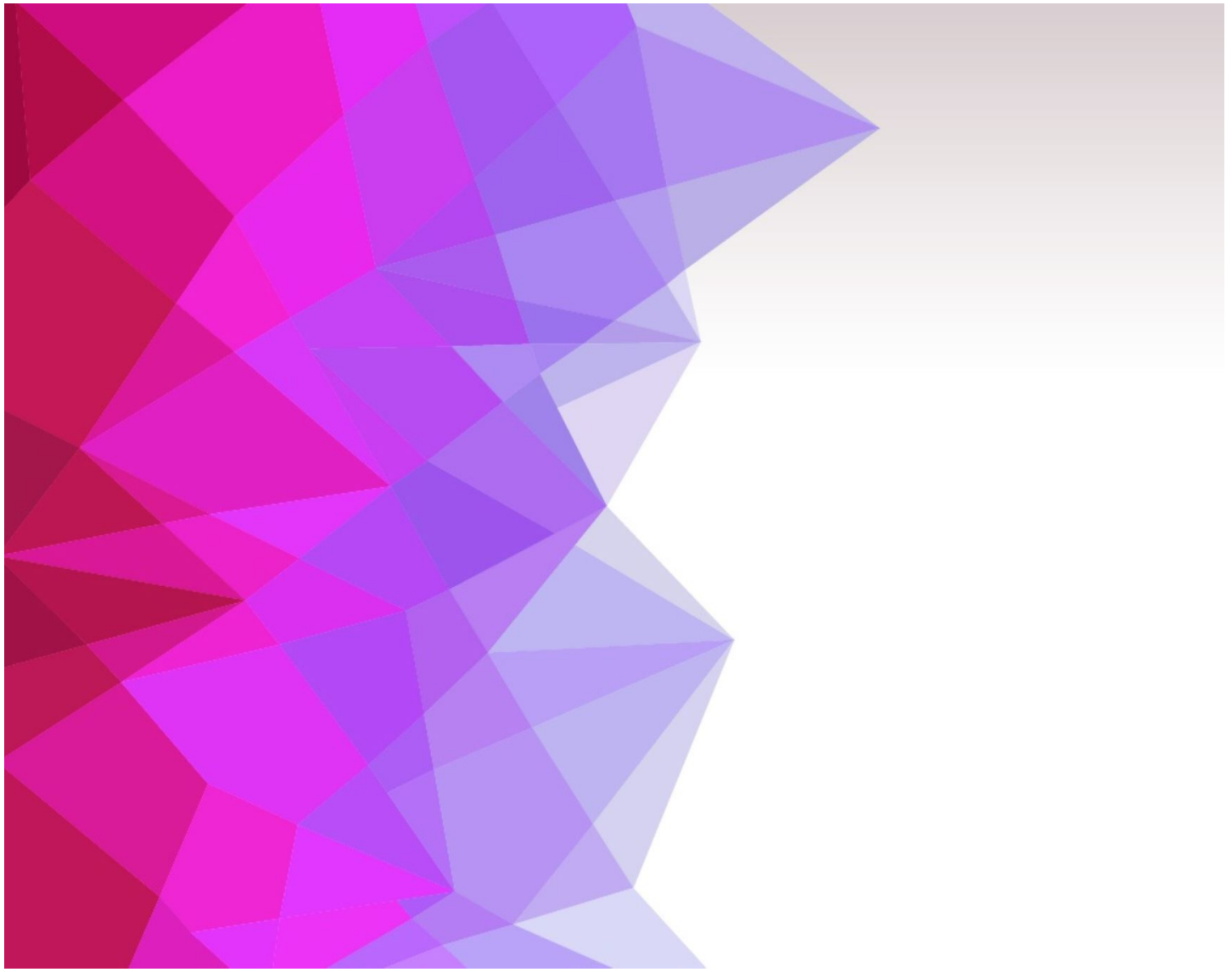
Data Science Capstone project



IBM Coursera

Firman insan M

Problem that can be solved with this dataset :

1. House-investment Analysis

Q : What is the best house investment and have good enough facility in the area ?

---

2. Favorite Food Venue Analysis.

Q : What is the most favorite food in Toronto based on its borough. And what is the variation.

Quick Background :

In a Big modern city, we can find a lot of open data that correspond our needs like location data of food and facilities venue.

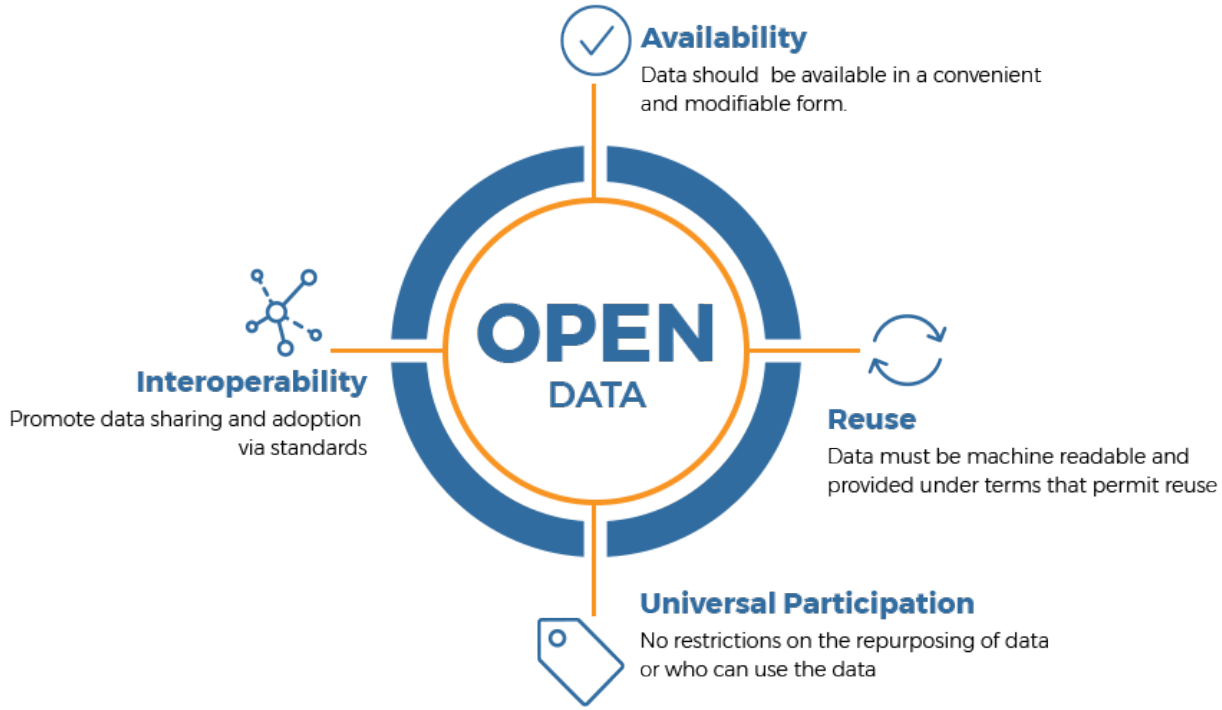
It may have varied in places from quiet to crowded one.

So, It is important find and segment the different food venues and facilities in a neighborhood according to venue category, and then we can group neighborhoods together that incorporate similar kind of neighborhoods to find pattern and similarities.

Pre-Process Open Data : Foursquare API + Wikipedia

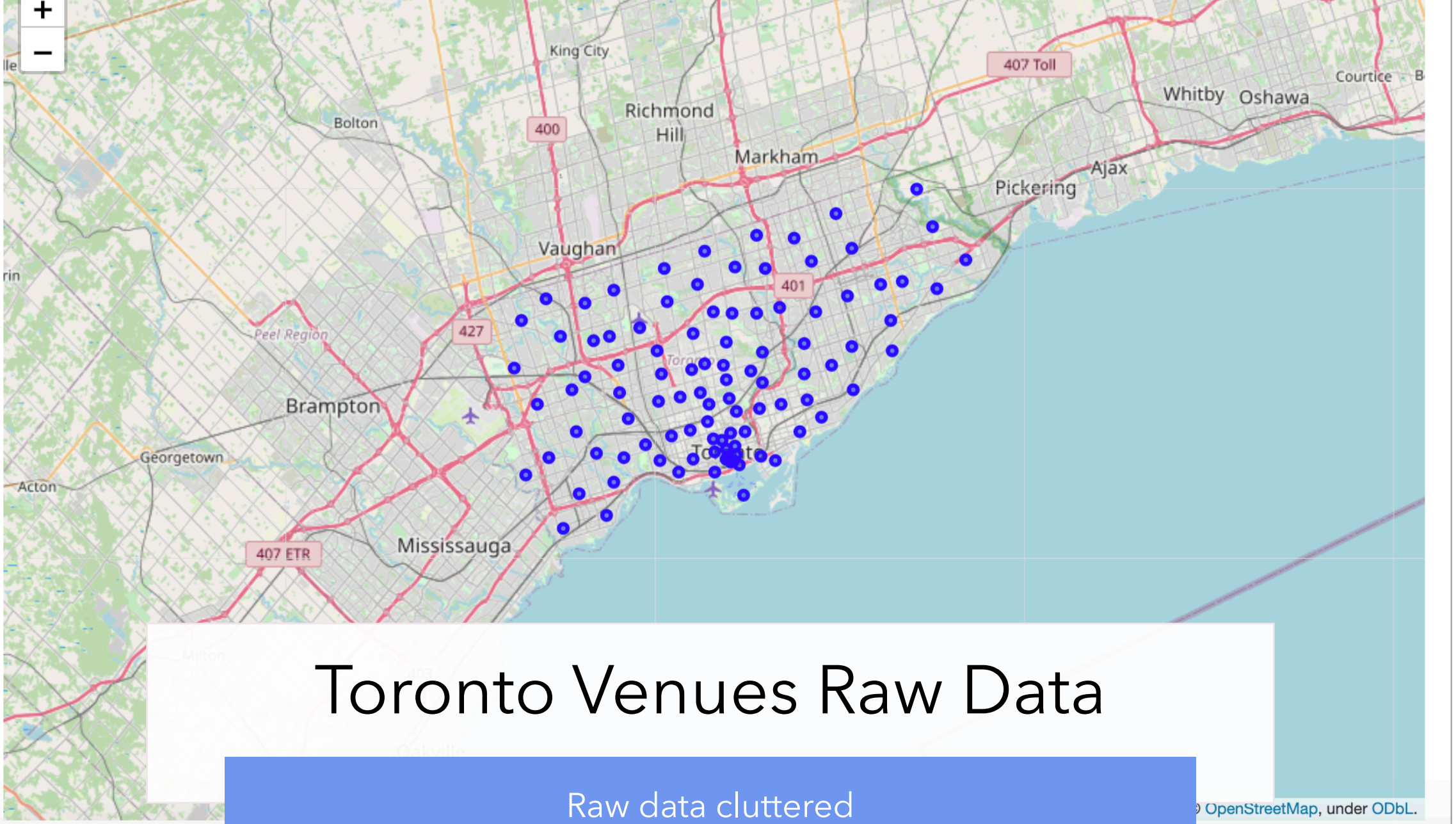
To get our first data sources which we can get data from Wikipedia using web scrapping, XPath Python Library, then wrangling it into pandas Data Frame.

And for coordinate data, which we get from foursquare Application Programming interface (API) in JSON format and the combine it with csv data from Toronto open data. Once the data are already tidied, we can see the pattern that we want to research on



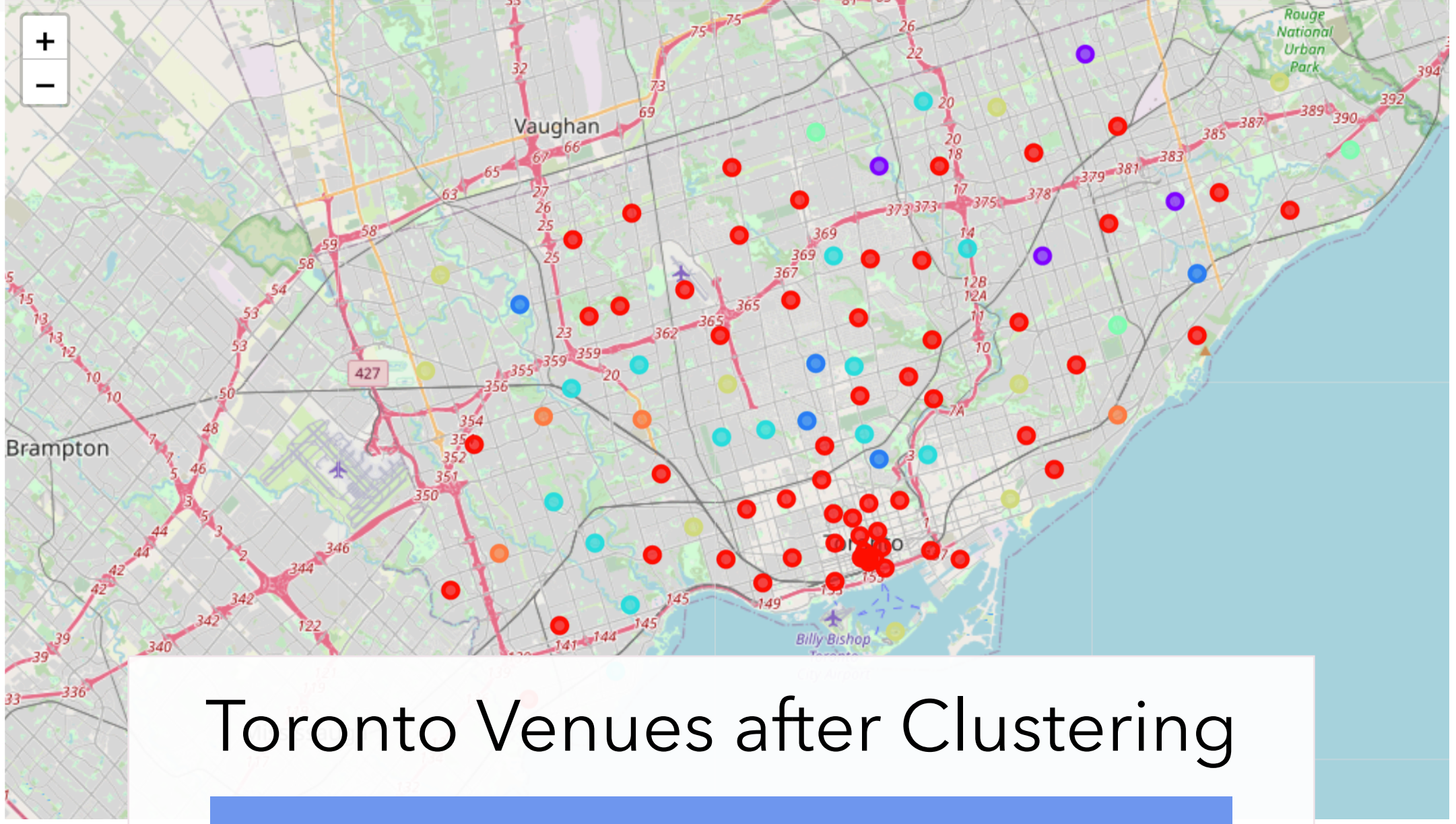
K- Means Clustering

Unsupervised learning



Toronto Venues Raw Data

Raw data cluttered



Toronto Venues after Clustering

Toronto with 7 Cluster

Current Result :

The different color that available on the Toronto map above is the cluster of venues from foursquare API based on the most common venue on its region / borough.

We can see there are 7 clusters that we can use to determine our decision whether to invest on property or to start a business.

For example, if we know food and beverages venue is too common in the area that marked red , we can try to start a business to complement the market there. And so on.

Or we can find the nearest facilities that related to Education venue that marked Green or Leisure venue that market Purple, It makes housing there can be good investment for the future.

- Further works

In an Unsupervised clustering, The room of improvement can be overlooked on venues there and enrich our data as much as possible so we can get the optimal results of our investigations.

Thank You

I welcome any feedback ☺