IBM DATA SCIENCE CAPSTONE
THE BATTLE OF THE NEIGHBORHOODS

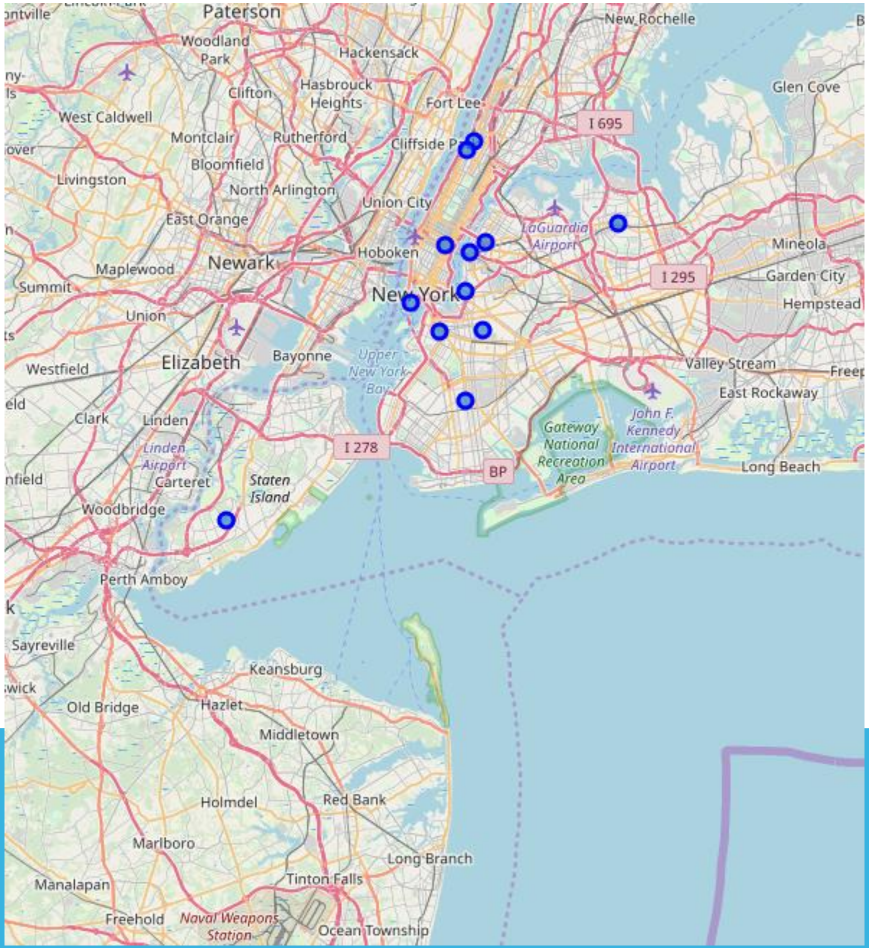
PROJECT: NEIGHBORHOOD SIMILARITIES BETWEEN CITIES

Paula Orellana
September 2019

BUSINESS PROBLEM

Our client is a successful owner of several boutique French Cafes in New York City, Currently evaluating the possibility to start a new venture in Toronto, but is not sure in what neighborhood the new venue should be located and be successful.

Customer Locations in NY



DATA

Requirement

- Data need it to answer the business question:
  - ✓ Neighborhoods in New York City
  - ✓ Neighborhoods in Toronto
  - ✓ Venues in each city per neighborhood
  - ✓ Successful locations already in place in New York City

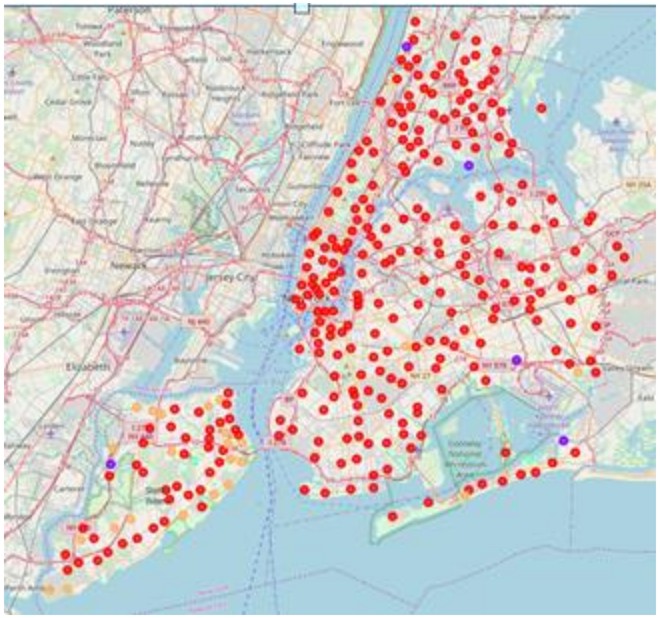
Source

- Wikipedia
  - ✓ https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M,
- NYU website
  - ✓ https://geo.nyu.edu/catalog/nyu_2451_34572
- GeoPy package, for Longitude and Latitudes
- FourSquare API, for venue information in each city

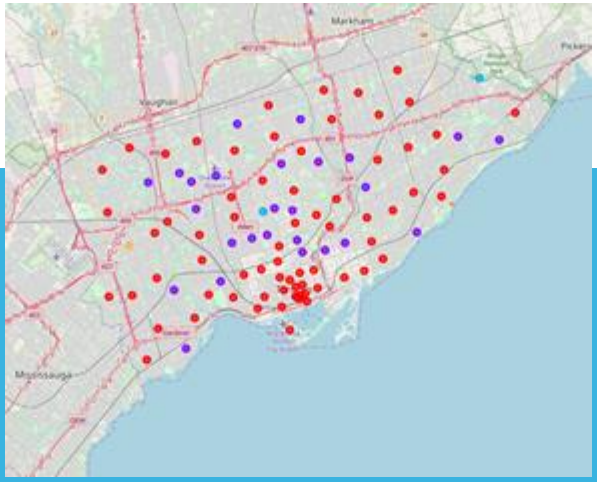
METHODOLOGY

1. Scrape Wikipedia (and other sources) for list of Neighborhoods in each City
2. Dataframe for each city with coordinates and venue information for these Neighborhoods.
3. Combine them to form a single set of Categories
4. Perform Clustering on the data with the help of Unsupervised Machine Learning Algorithm (K Means Algorithm)
5. Bring known successful venues data into play and select Toronto neighborhoods that match cluster and venue characteristics from known successful locations

New York Common Clusters



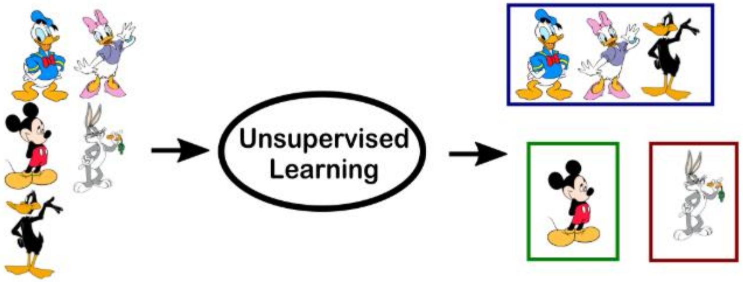
Toronto Common Clusters



CLUSTERING

K-Means Clustering (Unsupervised Machine Learning)

K-means can group data only unsupervised based on the similarity of elements (Neighborhoods in this case) to each other. We wish to learn the inherent structure of our data without using explicitly-provided labels



What is a Cluster?

A cluster is a group of data points or objects in a dataset that are similar to other objects in the group, and dissimilar to datapoints in other clusters.

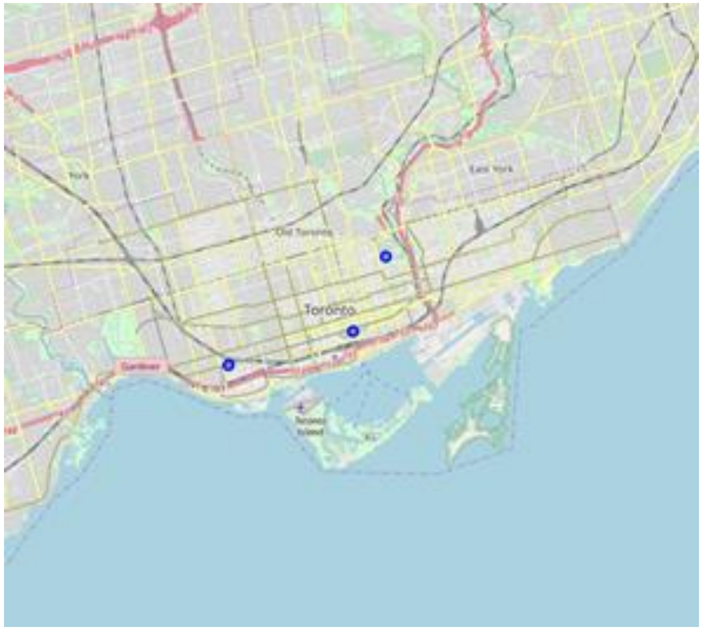
K-Means Clustering

K-Means is a type of partitioning clustering, that is, it divides the data into K non-overlapping subsets or clusters without any cluster internal structure or labels. This means, it's an unsupervised algorithm.

RESULTS

Our analysis come up with 3 potential locations for a new French Café in Toronto

- Cabbagetown,St. James Town in Downtown Toronto
- Commerce Court,Victoria Hotel in Downtown Toronto
- Brockton,Exhibition Place,Parkdale Village in West Toronto



DISCUSSION

1. Since two of the neighborhoods are in Downtown Toronto, we can point that borough as the one most promising to our client.
2. Would be interesting to explore other Clusters that also seem to have a high presence of food services business, where the market can be less saturated than our current selection.
3. A more rich analysis can be done if other sources of data were to be included such as: population density, demographics, and cuisine type distribution.

Je voudrais un café,
s'il vous plaît!
Merci