Battle of the Neighborhoods
New York City vs. Toronto

Lian Wang

August, 2020

Relocating: New York City ⇔ Toronto

- The old way
  - Facing unknown, anxiety
  - Limited information and resource, trial and error
  - Hassle of multiple moving
- What could be better?
  - “Learn” about the neighborhoods in the new city in advance
  - Compare the neighborhoods in the two cities
  - Recommend/pick target neighborhoods based on knowledge of the old city
- Location data and machine learning could help

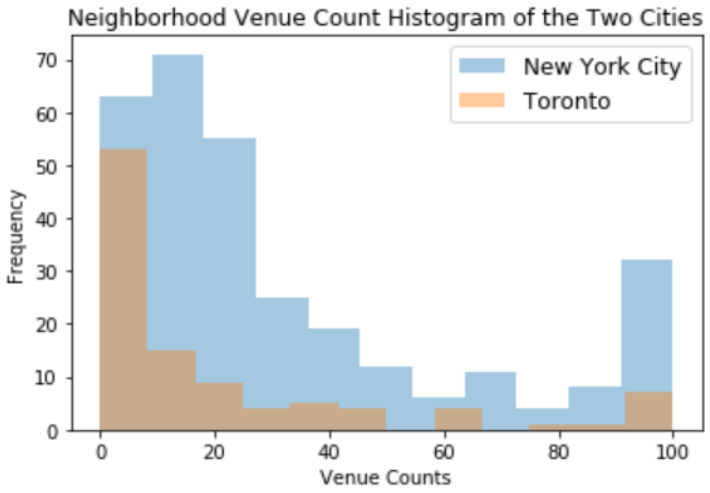
Data Sources

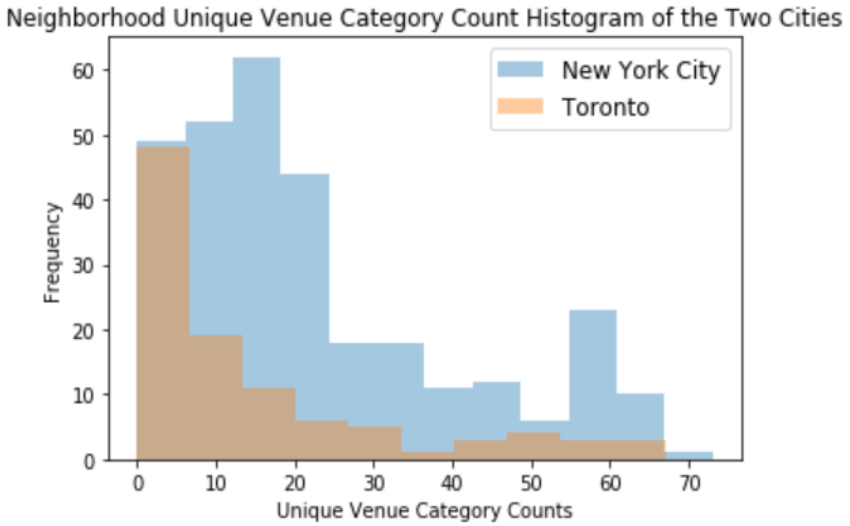
- New York City neighborhood data with geographical details
  - At https://cocl.us/new_york_dataset from this course
- Toronto neighborhood data with geographical details
  - From Wikipedia page at https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
  - At http://cocl.us/Geospatial_data from this course
- Foursquare API to extract neighborhood venues details within 500 meters of radius, limit up to 100 retrieves.

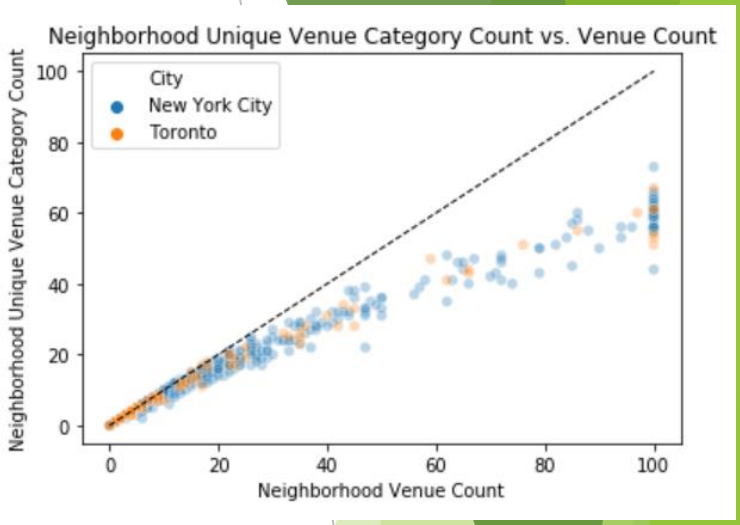
Data Summary

- 409 neighborhoods
  - 306 in New York City
  - 103 in Toronto
- 458 unique venue categories from **Foursquare**
- Neighborhood summary measures
  - Number of venue counts – how busy
  - Number of unique venue category counts – how diverse/convenient
- Clustering analysis features
  - Number of venues in each unique venue category in each neighborhood

Overall, different yet similar…

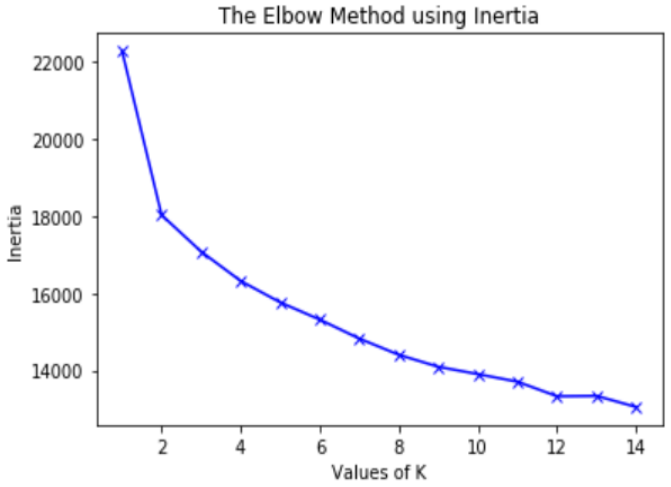


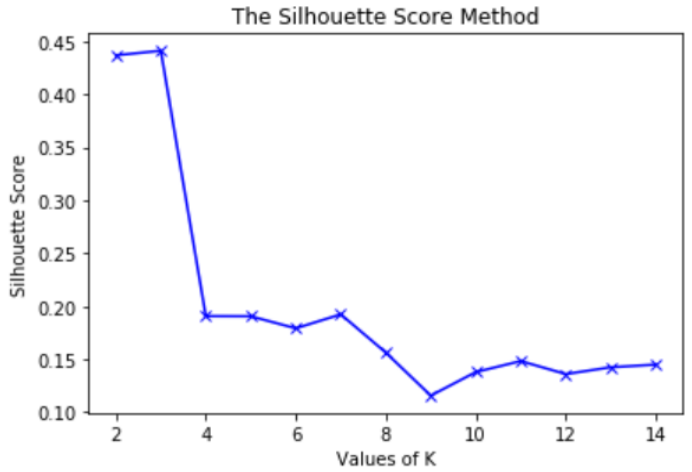


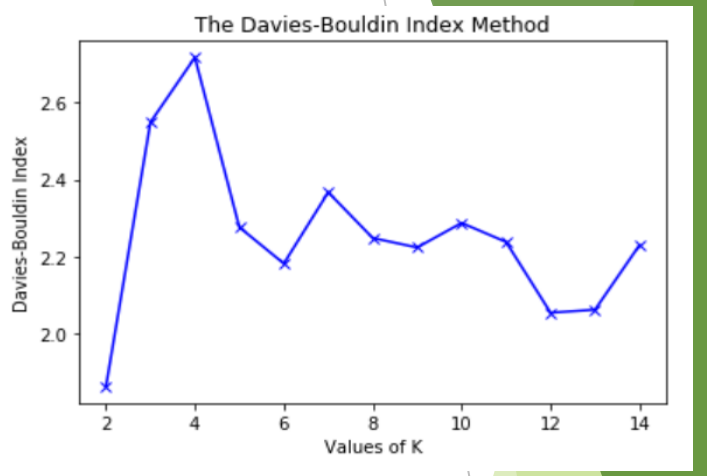


- New York City has more neighborhoods than Toronto across the whole spectrum of venue counts or unique venue category counts, no surprise since it has about 3 times more neighborhoods
- Similar histogram patterns follow a somewhat lopsided U-shape for both cities
- The neighborhoods from the two cities are relatively well mixed across the whole spectrum of these two measures in the scatter plot

How many clusters to use for K-means?

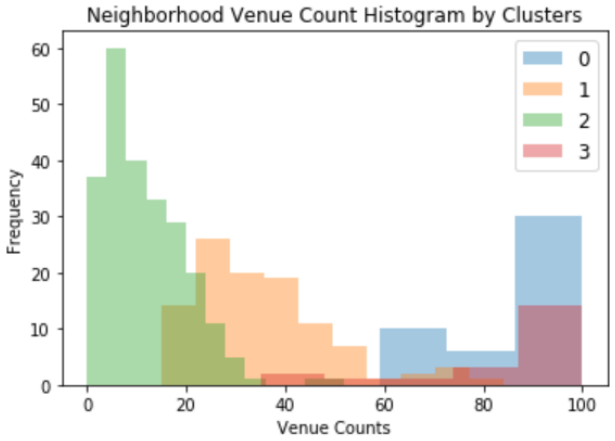


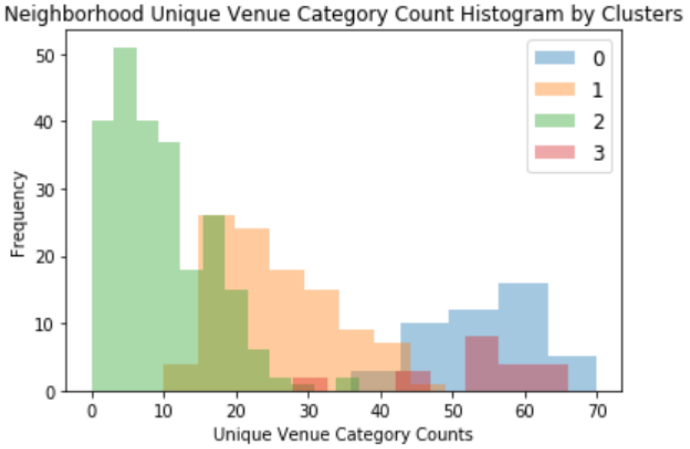




- Elbow-method not clear, could do 3-6
- Silhouette-method suggests 3
- Davies-Bouldin Index suggests 4
- Decide on 4 (3 resulted in very uneven clusters, 5-6 resulted in one very small cluster of 2 neighborhoods)

Meet the Four Clusters

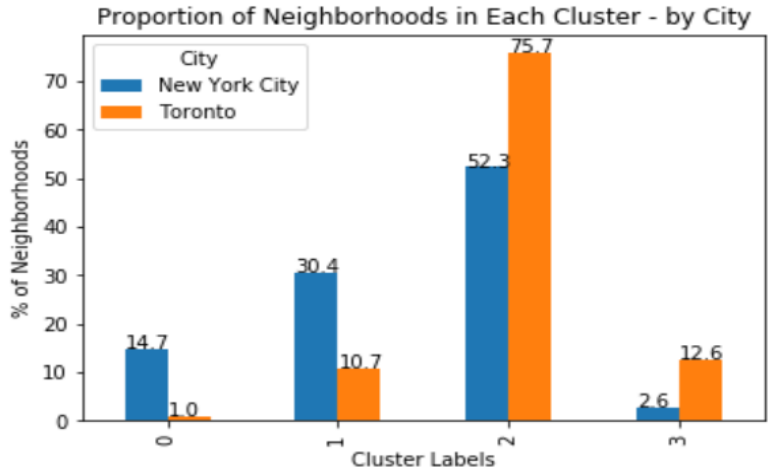




- Cluster 0: Most with high venue counts (>60) and high unique venue category counts (>40)
- Cluster 1: Most with medium venue counts (20-60) and medium unique venue category counts (20-40)
- Cluster 2: Most with low venue counts (<20) and low unique venue category counts (<20)
- Cluster 3: Most with high venue counts (>60) and high unique venue category counts (>40)

* Can't tell what is different between cluster 0 and 3, both being the high venue count/high unique venue category count type using these two measures only.

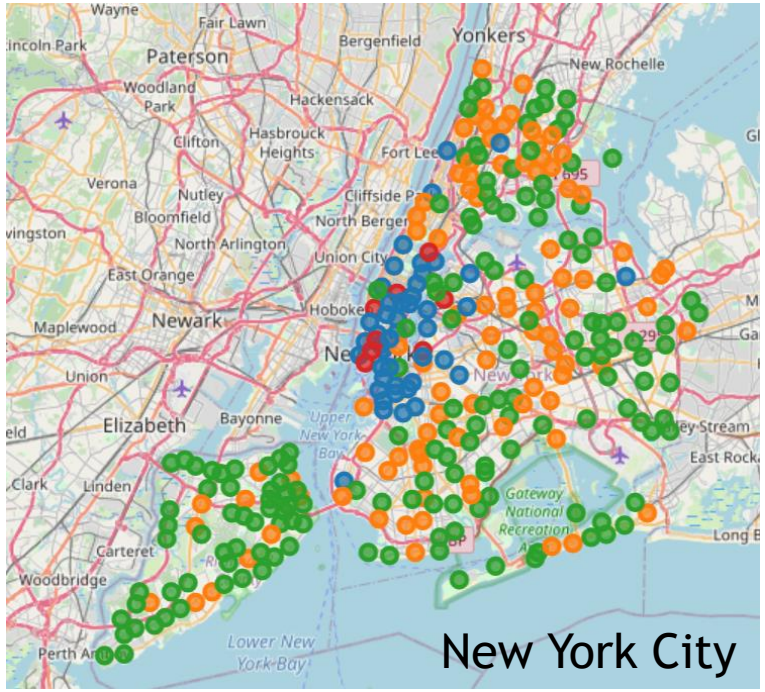
Compare Clusters by City

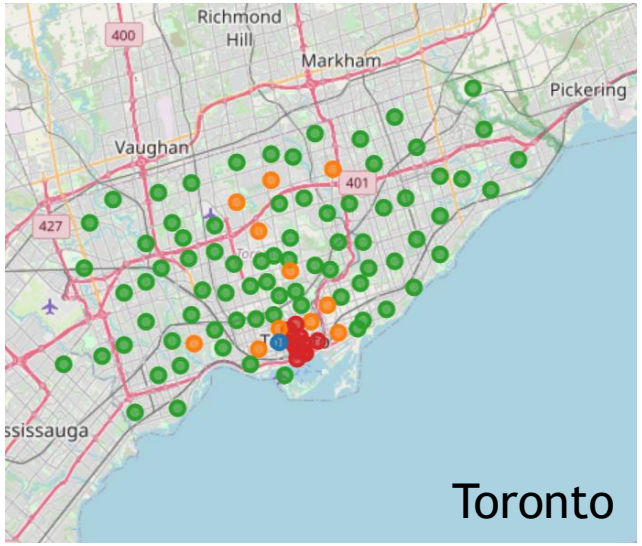


City	New York City	Toronto
Cluster Labels		
0	45	1
1	93	11
2	160	78
3	8	13

- Chi-squared test based on the table indicates shows difference in the neighborhood cluster distribution between the two cities ($p<0.001$)
- Toronto
  - Higher proportion of neighborhoods in cluster 2 (the low venue counts/low unique venue category count cluster)
  - Lower proportion of neighborhoods in cluster 1 (the medium venue counts/medium unique venue category counts cluster)
- Cluster 0, almost exclusive to New York City neighborhoods
- Both cities have at least one neighborhood in each cluster

Map view





- New York City, more neighborhoods and more densely packed
- Similar cluster distribution patterns
  - High venue count/unique venue category count (Clusters 0 and 3, blue and red) are aggregated around city center (Manhattan for NYC, Downtown Toronto for Toronto)
  - Others spread out and farther from the center

Recommending target neighborhoods

- When target cluster in a city has 5-10 neighborhoods, recommend those
- When target cluster in a city has <5 neighborhoods, expand the target to clusters that are most close to the initial target cluster
- When target cluster in a city has >10 neighborhoods
  - Slightly >10, use further restrictions to narrow down the target, say, must contain work-out venues or other criteria
  - >>10, repeat the clustering analysis using the target cluster neighborhoods in both cities
    - Could use different features (say, top 100 most retrieved venue categories only, or a prespecified focus category set).
    - Iterate if needed until the final target cluster contains 5-10 neighborhoods

Summary

- New York City and Toronto are different
  - Number of neighborhoods (NYC has more)
  - Density of neighborhoods (NYC neighborhoods are more packed)
  - Cluster distribution (Toronto has higher proportion of low venue count/unique venue category count)
  - One high venue count/unique venue category count cluster contains almost exclusively NYC neighborhoods
- Yet, their neighborhoods are similar
  - Venue count and unique venue category count distribution patterns somewhat lopsided U-shape for both cities
  - Cluster distributions on the map show similar pattern as well
  - Most importantly, both cities have at least one neighborhood in every cluster, meaning one could always find similar neighborhoods base on their ideal neighborhood in the old city and feel welcome and comfortable relocating!

Future Work

- Other unsupervised machine learning techniques could be used other than k-means
- Features could be retrieved and extracted differently from locational data
- Other data sources could be incorporated
- More user input could be added as restrictions or guide the feature formulation
- With the larger amount of diverse data and the power of machine learning, relocating across the global could be a better and easier experience…