Segmenting and Clustering neighborhoods in New York City and Toronto

Collin Slack

1. **Introduction and Business Problem**

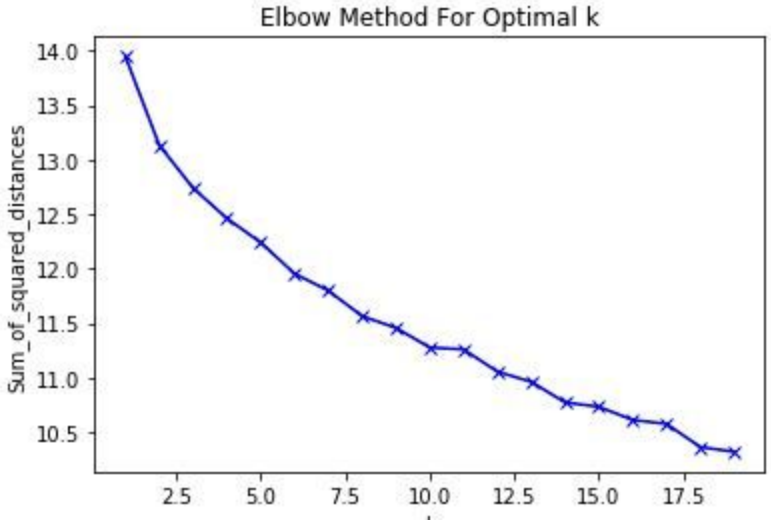
A business owner of several stores in New York City is looking to expand into other cities.
After some consideration he has chosen Toronto as one of the cities to expand into, however he
does not know much about different areas of Toronto. We must cluster neighborhoods of both New York and Toronto to determine what neighborhoods in Toronto are similar to neighborhoods that a store in currently located in in New York, then Decide the best candidates for the new
store.

2. **Data**

For this project we will use the Foursquare API location data on New York and Toronto. Foursquare's API allows us to find venues and their attributes such as rating, type of venue, and address from location Data. We will use the API to determine what types of venues are in each neighborhood, then use that data to perform k-means clustering. k-means clustering will cluster unlabeled data into n clusters where each observation belongs to the cluster with the closest mean. Any neighborhoods that returned less than ten venues from the FourSquare API were removed from the data set to prevent displacement due to small sample size. Data was gathered into a table containing Neighborhood names, what city they are in, location by latitude and longitude, and what venues were nearby.

3. **Methodology**

K-means clustering was used to group neighborhoods. To find the optimal K, the sum of squared distances was found at various k values as shown in the graph below.

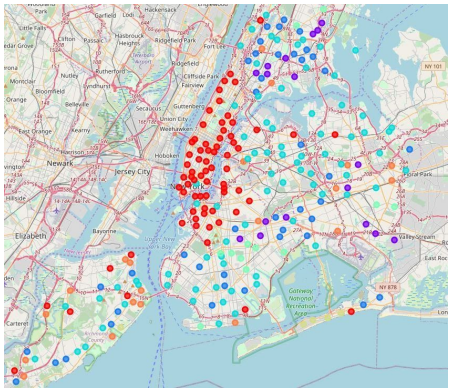


Unfortunately there is now clear elbow in the graph, making the elbow method a poor way to determine the optimal k for this data set. A k of 6 was chosen for this data.

4. **Results**

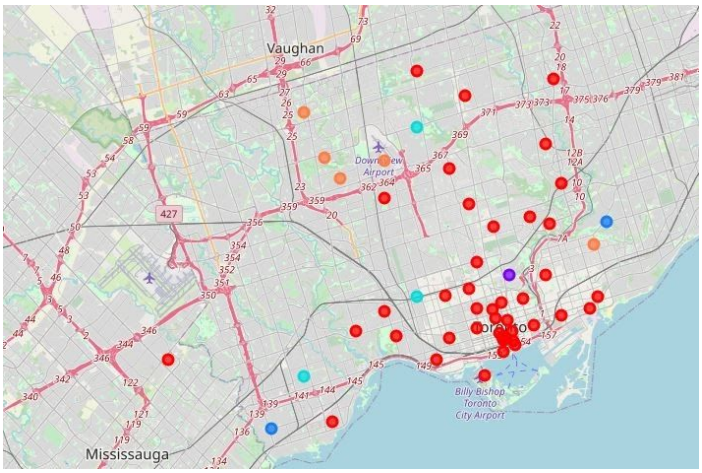
Figure 1 contains the map of New York City with each neighborhood colored according to its cluster.

Figure 1



Most neighborhoods fell into cluster 0 in cyan and cluster 3 in red, with most other clusters having more limited representation. The map of Toronto shows the same thing to an even bigger extreme, with almost every neighborhood being in 3, with a few in 0, 1 and 4 as shown in Figure 2 below.

Figure 2.



5. **Discussion**

Because the majority of neighborhoods in the target city, Toronto were in cluster 3 it may be difficult to decide the location of the new store if the best cluster based on new york locations is also cluster 3. In that case, a more in depth look into the neighborhoods may be

necessary. On the other hand, if the best cluster is one with a more limited amount in Toronto the potential locations will have been greatly narrowed down. In general across both cities a general pattern can be seen where most downtown neighborhoods belong to the same cluster. In future studies, it would be useful to find other data involving these neighborhoods such as their residents demographics, population, and growth.

6. **Conclusion**

In this report we looked at a clustering of neighborhoods in Toronto and New York. K-means clustering was used on data from the FourSquare API to cluster said neighborhoods. Neighborhoods in both cities downtown areas were clustered together, while clusters outside of downtown were much more spread out, with neighborhoods of different clusters close together and neighborhoods of the same cluster spread further apart. Stores that find the most success with locations downtown in one city will likely find success in locations that are downtown in the other; while stores that find success outside of downtown might have to be much more picky with their location selections due to the more spread out nature of all of the other clusters.