Choosing where to live when remote employment is an option

# Introduction

The COVID-19 pandemic normalized remote employment and made it possible for many workers to live almost anywhere without having to separate from their current employer. This project uses structured, agglomerative clustering to compare neighborhoods in two cities, Toronto and Philadelphia, using data from Foursquare. Neighborhoods from each city are associated with similar neighborhoods in the other city. Neighborhoods that are relatively dissimilar from neighborhoods in the other city are also identified. By characterizing the neighborhoods of each city relative to the other, someone familiar with only one city's neighborhoods (e.g., a current resident) could make a better decision about where in the other city to live. Though this project is limited to just two cities, the scope could be broadened to an arbitrary number of locales. Results are summarized geospatially and a brief qualitative review of Foursquare data by groups of neighborhoods is discussed.

# Data

Toronto and Philadelphia have 103 and 48 postal codes, respectively. These postal codes serve as proxies for each city's neighborhoods.

Github user AG2816 posted a geoJSON file containing Toronto's postal codes, available here: <https://github.com/ag2816/Visualizations/raw/master/data/Toronto2.geojson>

The organization OpenDataPhilly.org (https://opendataphilly.org) provides geoJSON files for Philadelphia postal codes, available here: <http://data.phl.opendata.arcgis.com/datasets/b54ec5210cee41c3a884c9086f7af1be_0.geojson>

I replaced the postal codes associated with neighborhood names are available in the Jupyter notebook associated with this report. For Toronto, I scraped neighborhood names from this Wikipedia page: https://en.wikipedia.org/wiki/List\_of\_postal\_codes\_of\_Canada:\_M) Wikipedia

For Philadelphia, I compiled neighborhood names from pages 18 and 19 of this Pew Charitable Trusts report: https://www.pewtrusts.org/en/research-and-analysis/reports/2014/07/09/homeownership-in-philadelphia-on-the-decline.

To compare neighborhoods in the two cities, I used data from Foursquare. I did not consider additional neighborhood data (e.g., housing prices, crime, access to public transportation, etc.), though it would be possible to expand the analysis to include an arbitrary number of features.

Copies of all data used in this analysis are available here: https://github.com/JaeAre/Coursera\_Capstone.git, including the results retrieved from Foursquare.

# Methodology

I chose an agglomerative, hierarchical clustering approach using a connectivity matrix. If the neighborhoods of Philadelphia all cluster together, and the neighborhoods of Toronto all cluster together, this analysis would not be useful for someone familiar with one city trying to learn about the other. The connectivity matrix defines which postal codes can merge to form a cluster. It has the following form:

1. It is \_n x n\_ with rows and columns representing each of the \_n\_ postal codes in Philadelphia and Toronto.

2. It contains only zeros and ones. The elements along the principal diagonal are always 1.

3. Off diagonal elements indexed by a row and column corresponding to two postal codes from the same city equal zero.

4. Off diagonal elements indexed by a row and column corresponding to two postal codes from different cities equal one.

Each postal code begins in its own cluster. At the first step, the connectivity matrix prevents two Toronto (Philadelphia) postal codes from merging. As long as a cluster contains at least one postal code from Toronto (Philadelpia) a postal code from Philadelphia (Toronto) may join that cluster. One way I thought to extend this analysis would be to choose a reference city, say Toronto, and then include more cities in the United States. I could construct a connectivity matrix that would only let postal codes in US cities join clusters that contained at least one postal code from Toronto. In this way, it would be a "Toronto-centric" analysis and would be most useful to a Toronto resident considering relocation to the USA.

When results are returned, all clusters will either contain a mixture of Toronto and Philadelphia neighborhoods or contain a single neighborhood. The former will be helpful for learning about collections of similar neighbohoods between the two cities. The latter will be helpful for identifying those neighborhoods that are unique relative to the neighborhoods in the other city.

The algorithm used is the \_AgglomerativeClustering\_ algorithm in \_scikit-learn\_. The metric used to compute the linkage is \_Euclidean\_ and a \_ward\_ criterion is used to minimize the variance of the clusters being merged. One challenge when working with agglomerative cluster models is to choose the number of clusters. Two common ways of choosing the number of clusters is the either visually inspect a dendrogram or compute a (see, for example, https://en.wikipedia.org/wiki/Silhouette\_(clustering)). I have found that a combined approach works well. First, determine the optimal number of clusters using the silhouette score. Set this as the upper limit on the number of clusters. Next, examine the dendrogram. If a relatively small reduction in the number of clusters leads to a relatively large reduction in complexity, reduce the number of clusters accordingly.

# Results

The results of the analysis are summarized in the following figure and table.

A picture containing icon

Description automatically generated

Table

Description automatically generated

The first cluster (cluster 0) contains 64 Toronto neighborhoods and 44 Philadelphia neighborhoods. This is by far the largest cluster in terms of number of neighborhoods it contains. Looking across its top ten venues, this cluster seems like an ideal residential cluster, with restaurants, bars and cafes, as well as a grocery store and pharmacy.

The second cluster (cluster 1) contains 23 Toronto neighborhoods and 2 Philadelphia neighborhoods. It seems like it might be a little less residential with fewer options neighborhood activities. However, parks and gyms figure more prominently in this cluster, suggesting it might be a better choice for someone who prefers more active recreational activities.

The remaining five clusters contain a single Toronto neighborhood each. These clusters all share a zoo exhibit and event space. From the Foursquare data, these Toronto neighborhoods do not look very residential.

# Discussion and Conclusion

The purpose of this project was to cluster the neighborhoods of two cities, Toronto and Philadelphia, so that someone familiar with one city could better understand the neighborhoods in the other city.

With more time, some additional avenues of exploration might be:

1. Expand the analysis to beyond two cities so that someone familiar with one city could choose one among many city's neighborhoods for relocation.
2. Incorporate data beyond those available from Foursquare (i.e., housing prices, crime, etc.)
3. Adjust the structure of the analysis (i.e., the connectivity matrix) so that results could be tailored to the reference city of the user.