Comparing Neighborhoods of New York City and Toronto

Introduction

Given the current rate of greenhouse gas emissions (GGEs), the global average temperature might hit 2 degrees Celsius above pre-industrial levels within 15 years, a threshold that will likely cause serious harm (Mann, 2014). The transportation sector is the largest culprit of GGEs and stands for over a fourth of emissions (EPA, n.d.).

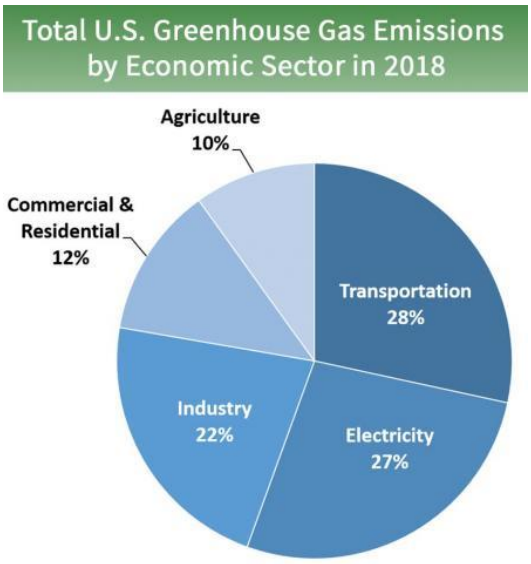

Figure 1: Pie chart of US GGEs by sector (EPA, n.d.).

A place for improvement could be business-related travels, for instance real estate scouting. Is there a way to assess a neighborhood in one city that is analogous to a neighborhood in another city? And if so, to which accuracy? Building such a model could be useful to real estate investors who cannot or does not want to travel but would still like to assess the neighborhood and its characteristics.

Data

I will use the New York data we used in week 3 and the Toronto dataset in week 4. Then I will use Foursquare to get data on nearby venues. Features will include different venues in the close distance to each neighborhood, e.g. Yoga Studio, ATM, Lebanese Restaurant.

Methodology

I will use K-means classifier to group the neighborhoods in both cities into clusters. 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. Objects within a cluster are very similar, and objects across different clusters are very different or dissimilar. The distance of samples from each other is used to shape the clusters. So we can say K-Means tries to minimize the intra-cluster distances and maximize the inter-cluster distances. You may use Euclidean distance, Cosine similarity, Average distance, and so on. Thus, I can use these clusters to find similar neighborhoods. In addition, I can use the data on number of venue types as a sorting key. I.e. I can find a similar neighborhood in the other city with a specific requirement to nearby venues.

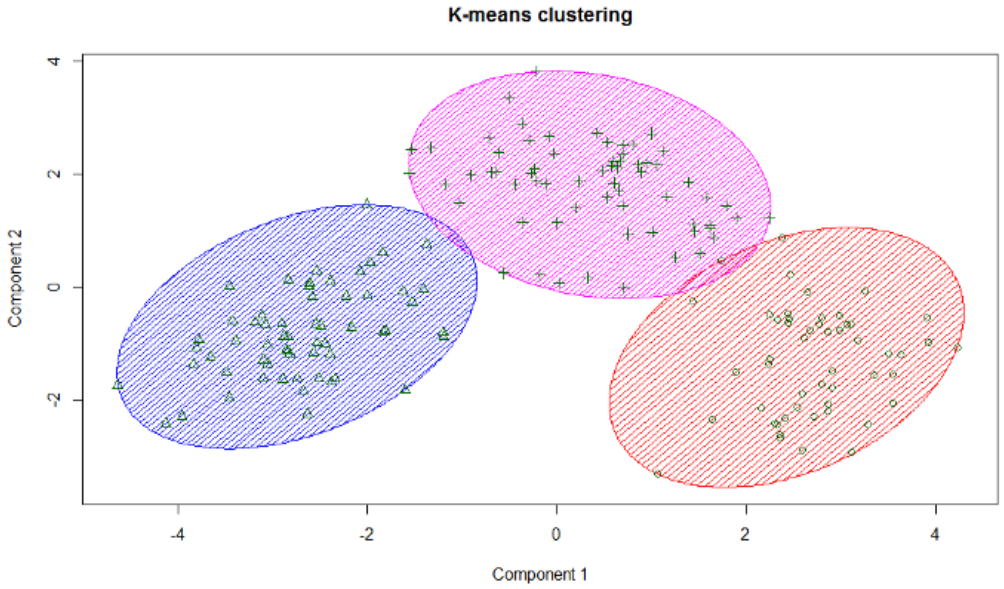

Figure 2: Visual example of K-mean clustering (Khan, 2017).

Results

After running the model, we get the following data frame:

	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	American Restaurant	...
6	Manhattan	Marble Hill	40.876551	-73.910660	1	0.000000	0.000000	0.0	0.0	0.041667	...
100	Manhattan	Chinatown	40.715618	-73.994279	0	0.000000	0.000000	0.0	0.0	0.040000	...
101	Manhattan	Washington Heights	40.851903	-73.936900	0	0.012346	0.000000	0.0	0.0	0.012346	...
102	Manhattan	Inwood	40.867684	-73.921210	3	0.000000	0.000000	0.0	0.0	0.018868	...
103	Manhattan	Hamilton Heights	40.823604	-73.949688	3	0.000000	0.015873	0.0	0.0	0.000000	...

Figure 3: First five rows of merged dataset with clusters.

We can see that we get clusters related to each neighborhood. The data frame consists of both New York City neighborhoods as well as Toronto neighborhoods. Parsing this data frame through the “find_similar_nbs()” function, we get similar neighborhoods (meaning in the same cluster) in the other city from which we are in. In addition, these entries are sorted by a key parameter linked to the venue type. In the following example, we want to find similar

neighborhoods to Somerville, Manhattan in Toronto where the number of stores for women is a key indicator:

```
0                                    Caledonia-Fairbanks
1          East Toronto, Broadview North (Old East York)
2                                 Willowdale, Newtonbrook
3                                          Lawrence Park
4                                                 Weston
5                                          York Mills West
6    Milliken, Agincourt North, Steeles East, L'Amo...
Name: Neighborhood, dtype: object
```

Figure 4: Results from running "find_similar_nbs" where neighborhood = "somerville", city = "Manhattan", and sort_values_by = "Women's Store".

If we map these locations with our geospatial data, we get the following map:

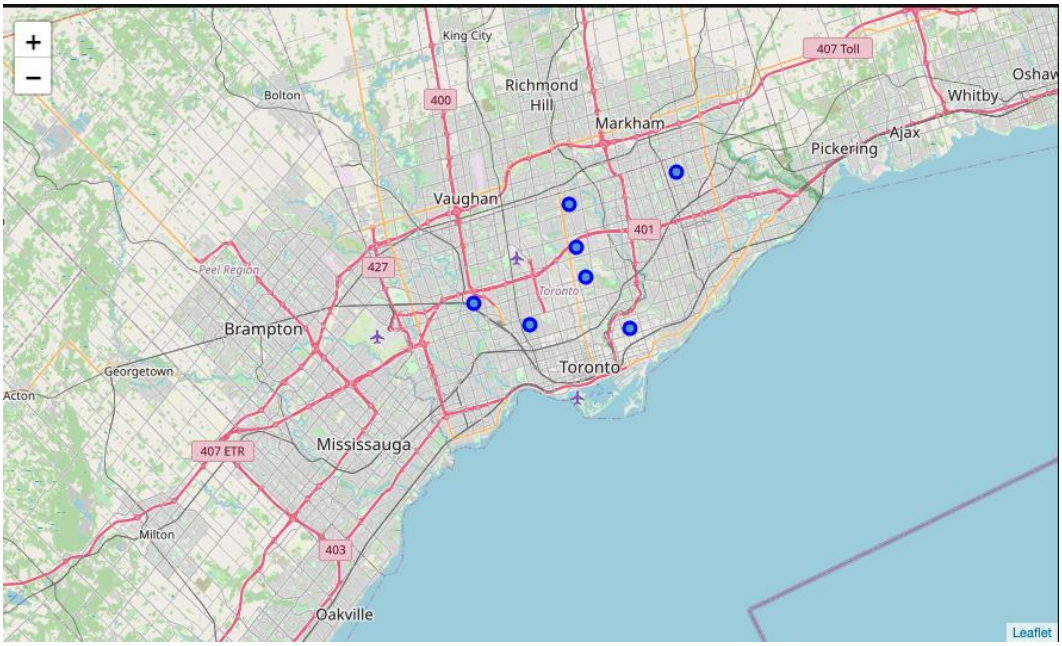

Figure 5: Map from running "find_similar_nbs" where neighborhood = "somerville", city = "Manhattan", and sort_values_by = "Women's Store".

Discussion

This is only a sparse model that shows the potential of comparisons models between cities. To further improve the model, there is a need for a magnitude of more data such as population, population density, median income, district type, and so forth. The climate benefit will be a function of the distance between the cities.

Conclusion

Solving the climate crisis is a difficult problem that necessitates out-of-the-box-thinking. City comparison models using K-means classifiers could be a part of the solutions by allowing for neighborhood comparisons without having to travel.

References

EPA. (n.d.). Sources of Greenhouse Gas Emissions. *United States Environmental Protection Agency*. https://www.epa.gov/ghgemissions/sources-greenhouse-gas-emissions

Khan, M. (2017). Kmeans clustering for classification. *Towardsdatascience*. https://towardsdatascience.com/kmeans-clustering-for-classification-74b992405d0a

Mann, M. E. (2014). Earth Will Cross the Climate Danger Threshold by 2036. *Scientific American*. https://www.scientificamerican.com/article/earth-will-cross-the-climate-danger-threshold-by-2036/