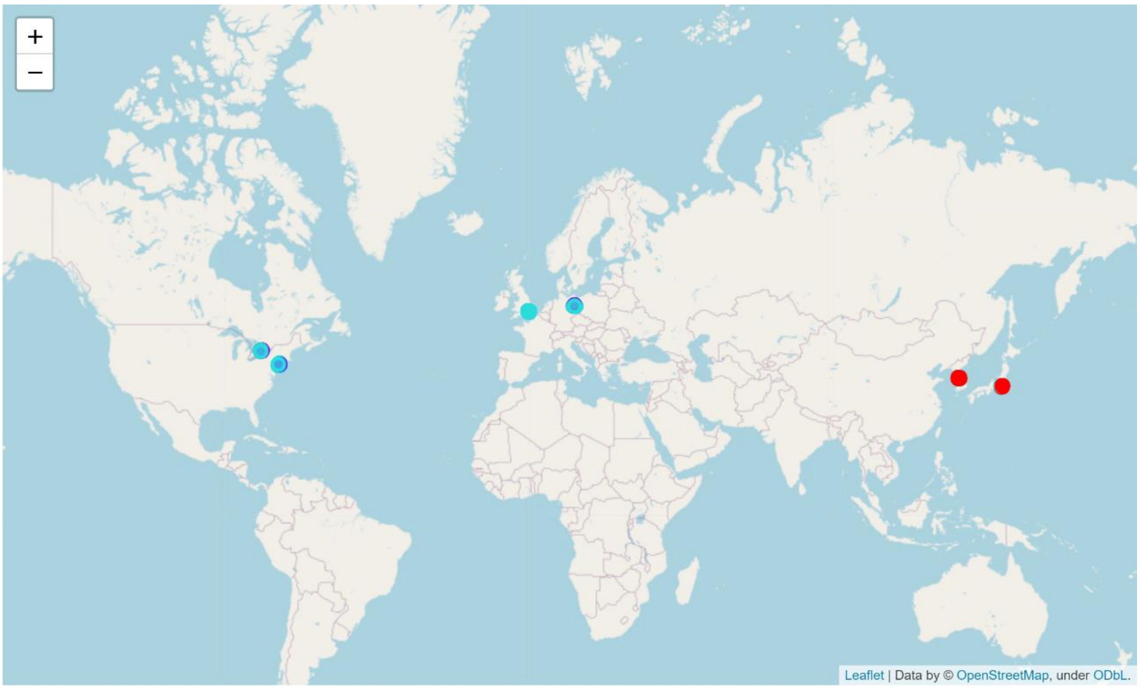
The Battle of the Neighborhoods

Clustering of Cities

Maxime Menu

April 18, 2020



1 Introduction

1.1 Background

Cheap air travel and globalization has made overseas traveling easier. Another consequence of globalization has been the cross-influencing of different cultures. This means that a lot of countries and their capital-level cities share many features in the kind of venues they manifest. This information can be used by travel agencies and alike to easily recommend and advert new travel destinations to their new customers.

1.2 Problem

Given the preference a customer might have for New York City, which city should a travel agency recommend to their customer with the assumption that travel distance is irrelevant. This project aims to answer this question.

1.3 Interest

Travel agencies, airline companies and other parties involved in the travel industry would be interested to know how to recommend next travel destinations to (potential) customers.

2 Materials & Methods

2.1 Data Gathering

Three types of data were gathered to carry out this analysis.

- Boroughs and neighborhoods for 6 selected capitals (New York, Toronto, Tokyo, Seoul, Berlin and London)
- Coordinates for neighborhoods
- Venue categories of FourSquare
- Venues in a set radius for each of the coordinates

Boroughs and neighborhoods

The purpose of boroughs and neighborhoods was to gather subdivisions of the cities which would later be clustered based on venues. In order to bring in a diversity of neighborhoods 6 capital-like cities were selected from 3 parts of the world (North America, Europe, Far East Asia).

This data was gathered from different sources, depending on the city.

New York: The data was downloaded from a public dataset provided by the Applied Data Science Capstone course as a JSON file. (https://cocl.us/new_york_dataset)

Toronto: The data was scraped from the Wikipedia page for Canadian postal codes. (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M)

Tokyo: The data was scraped from the Wikipedia page for Tokyo special wards. (https://en.wikipedia.org/wiki/Special_wards_of_Tokyo)

Seoul: The data was scraped from the Wikipedia page for the districts of Seoul. (https://en.wikipedia.org/wiki/List_of_districts_of_Seoul)

Berlin: The data was scraped from the Wikipedia page for boroughs and neighborhoods of Berlin. (https://en.wikipedia.org/wiki/Boroughs_and_neighborhoods_of_Berlin)

London: The data was scraped from the Wikipedia page for London postal districts. (https://en.wikipedia.org/wiki/London_postal_district)

Coordinates

The purpose of coordinates data was to be able to gather venue data with the Foursquare API in the next step, as the API requires coordinates.

This data was gathered from different sources, depending on the city.

New York: Coordinates were contained in the same file gathered for the boroughs and neighborhoods data or New York.

Toronto: The data was downloaded from a public dataset provided by the Applied Data Science Capstone course as a csv file. (http://cocl.us/Geospatial_data) This dataset was merged with the neighborhoods data for Toronto based on postal codes.

Tokyo: The coordinates were gathered by using the geopy package for python with the Nominatim geocoder for OpenStreetMap data.

Seoul: The coordinates were gathered by using the geopy package for python with the Nominatim geocoder for OpenStreetMap data. In case no result could be returned for the neighborhood names, coordinates were gathered by manually inputting the neighborhood names in google maps. (https://www.google.com/maps)

Berlin: The coordinates were gathered by using the geopy package for python with the Nominatim geocoder for OpenStreetMap data. In case no result could be returned for the neighborhood names, coordinates were gathered by manually inputting the neighborhood names in google maps. (https://www.google.com/maps)

London: The coordinates were downloaded from a public dataset containing coordinates for every postal code in the United Kingdom. (url: https://www.doogal.co.uk/files/postcodes.zip) This dataset was merged with the neighborhoods data for London based on postal codes.

Coordinates were visually confirmed by mapping them with the Folium package for python.

Venue Categories

The foursquare API returns the most specific category for a given venue. Spatial bias can however be included in the categories if they are too specific (e.g. Japanese restaurant, Korean restaurants etc.) which lead to the need to gather all Foursquare categories with their hierarchies as to be able to infer super categories of venues.

This data was gathered by using the categories Foursquare API.

Venues

The purpose of venues data was to assign features to neighborhoods based on the categories of venues found in these neighborhoods. These features were then used in the clustering steps in the analysis part.

This data was gathered by using the explore API of Foursquare. A limit of 100 venues in a radius of 500 meters for every set of coordinates was gathered.

2.2 Exploratory Analysis

In order to perform clustering of the neighborhoods, new features were created. These features are the most common venue categories per neighborhood.

2.3 Neighborhood Clustering Analysis

The neighborhoods were clustered based on the most common venues features created in the previous step.

The clustering algorithm used in this analysis is the k-means algorithm. In order to find the optimal number of clusters k, silhouette scores and inertia were calculated for clusters in a range of values 2 to 25 for the number of clusters.

The data was clustered with the optimal k value. The cluster labels resulting from the k-means clustering were assigned to the most common venues per neighborhood data. Clustered neighborhoods were visualized with Folium maps with color encoding for the clusters

The top 10 most present category values per cluster label were then calculated in order to label these categories.

2.4 City Clustering Analysis

The proportions of cluster labels were calculated for each city. These proportions were visualized with a stacked bar chart.

In the final step of the analysis this proportions data was hierarchically clustered and visualized with a dendrogram in order to which cities were similar to each other, especially to New York.

3 Results

3.1 City coordinates

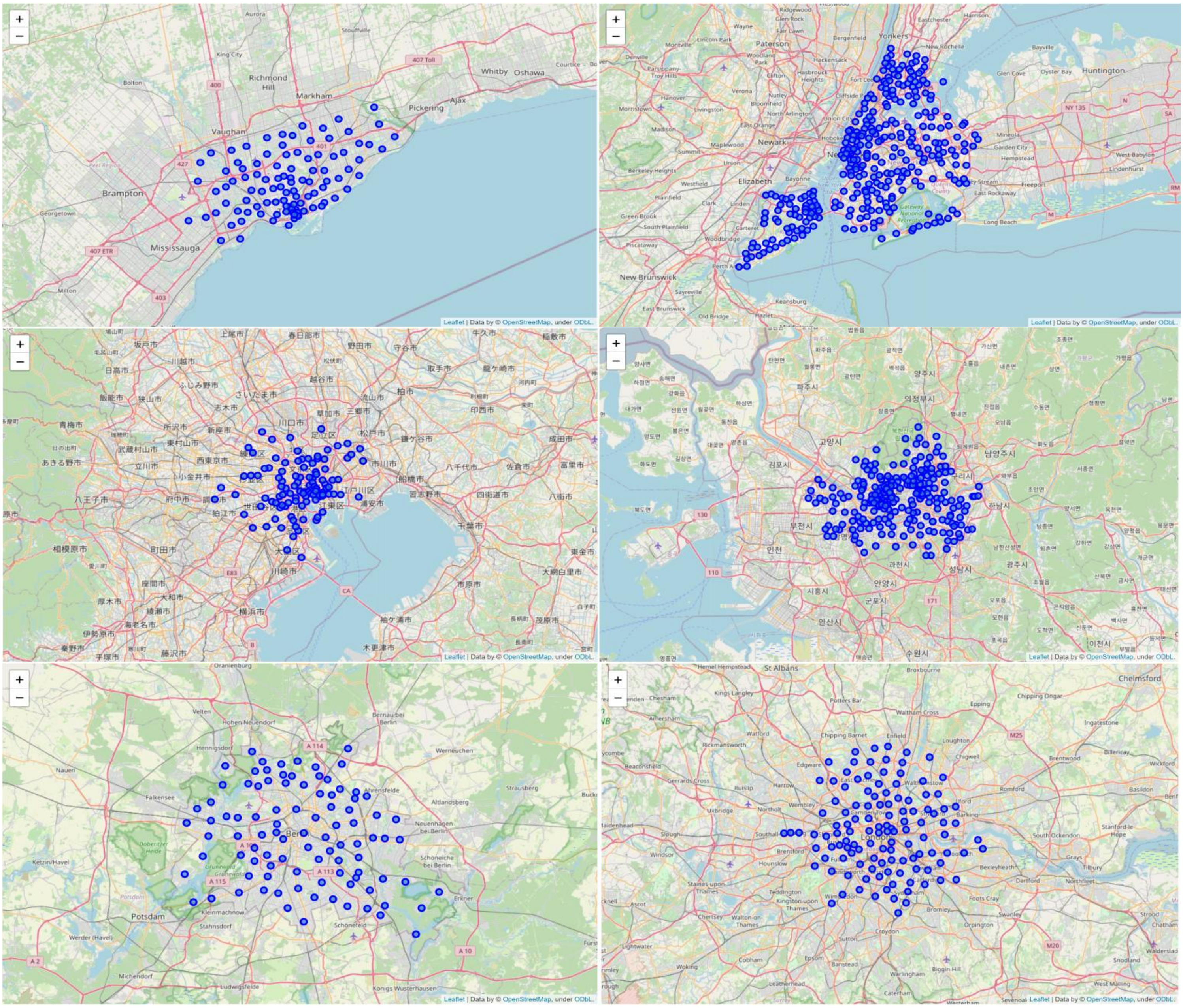


Figure 1: Coordinates of neighborhoods of 6 capital cities plotted on a Folium map.

(Upper left: Toronto, upper right: New York City, middle left: Tokyo, middle right: Seoul, lower left: Berlin, lower right: London)

3.2 Neighborhood Clusters

The silhouette scores returned an optimal value of k = 4. The inertia values didn’t show a clear/strong elbow point, therefore only the silhouette score was taken into account when selecting the optimal k number of clusters.

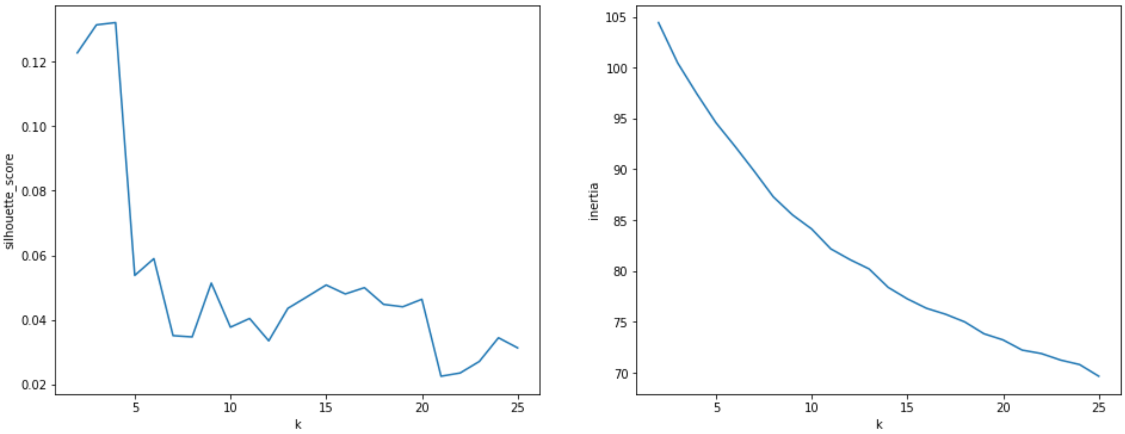


Figure 2: Silhoutte scores (left) and inertia values (right) for k clusters in range 2 to 25. The silhouette score shows an optimum at k=4. The inertia doesn't show a clear elbow point.

The cluster colored map shows that clusters 0 and 2 contain the most points, making those the most prevalent kind of neighborhoods in the analyzed capital cities.

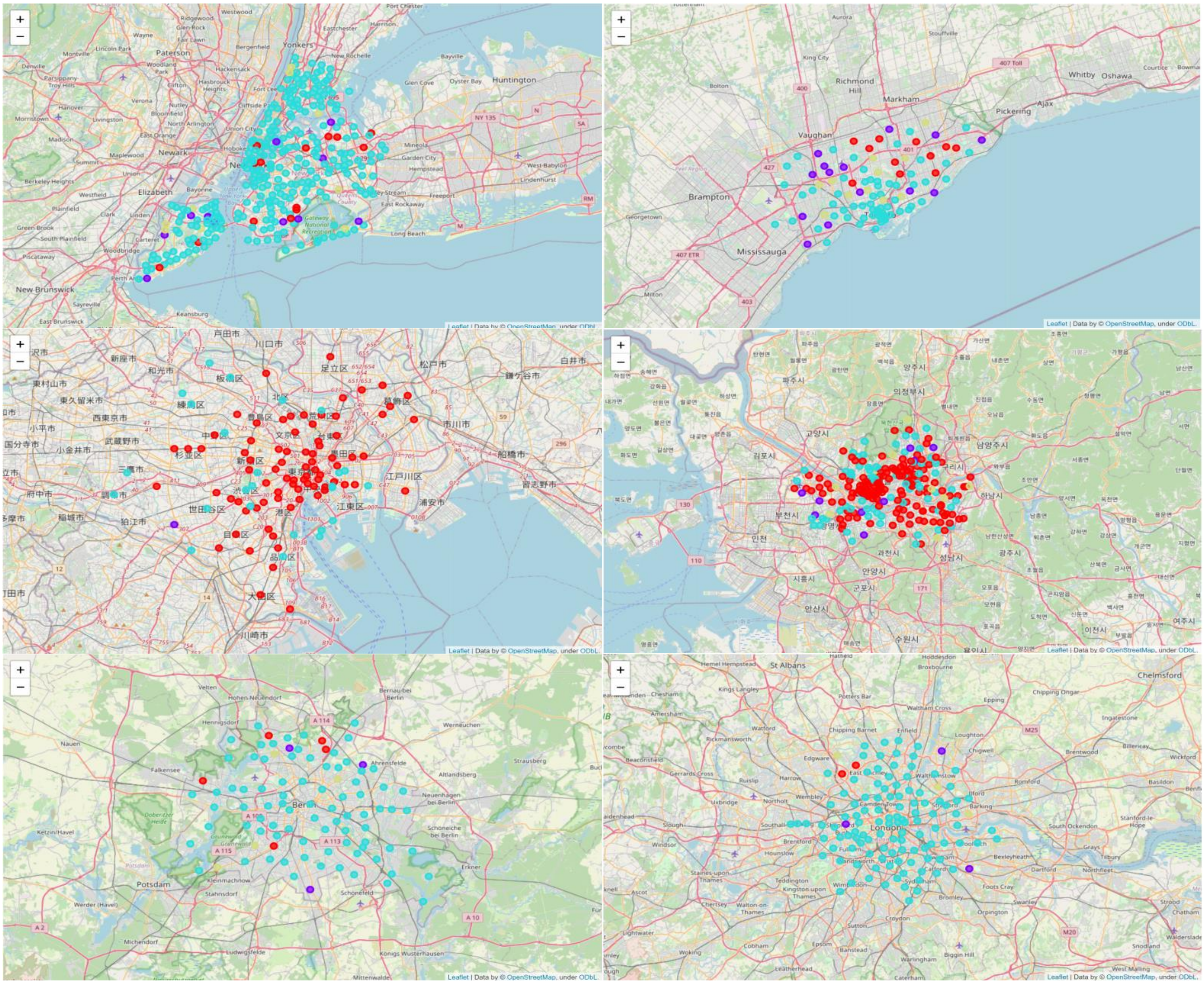


Figure 3: Neighborhoods after clustering colored by cluster label.

(Cluster 0: red, cluster 1: purple, cluster 2: blue, cluster 3: yellow)

(Upper left: Toronto, upper right: New York City, middle left: Tokyo, middle right: Seoul, lower left: Berlin, lower right: London)

Based on the most common venue categories in the clusters, the 4 types of neighborhoods were identified being: Asian restaurants Neighborhoods, Athletics & Sports Neighborhoods, Bars Neighborhoods and Parks Neighborhoods.

Table 1: Top 2 Most common venue counts per cluster labels.

Cluster Label	Most Common Venues	Count
3 **Parks and trails**	Park	22
	Trail	4
2 **Bars and Food & Drink shops**	Bar	77
	Food & Drink Shop	77
1 **Athletics & Sports and Food & drink shops**	Athletics & Sports	14
	Food & Drink Shop	4
0 **Asian Restaurants**	Asian Restaurant	223
	Coffee Shop	10

3.3 Neighborhood Cluster Proportions

With high rates of Asian Restaurants Neighborhoods, Seoul and Tokyo are similar to each other, while the western capitals with higher varieties in neighborhood types show some similarity.

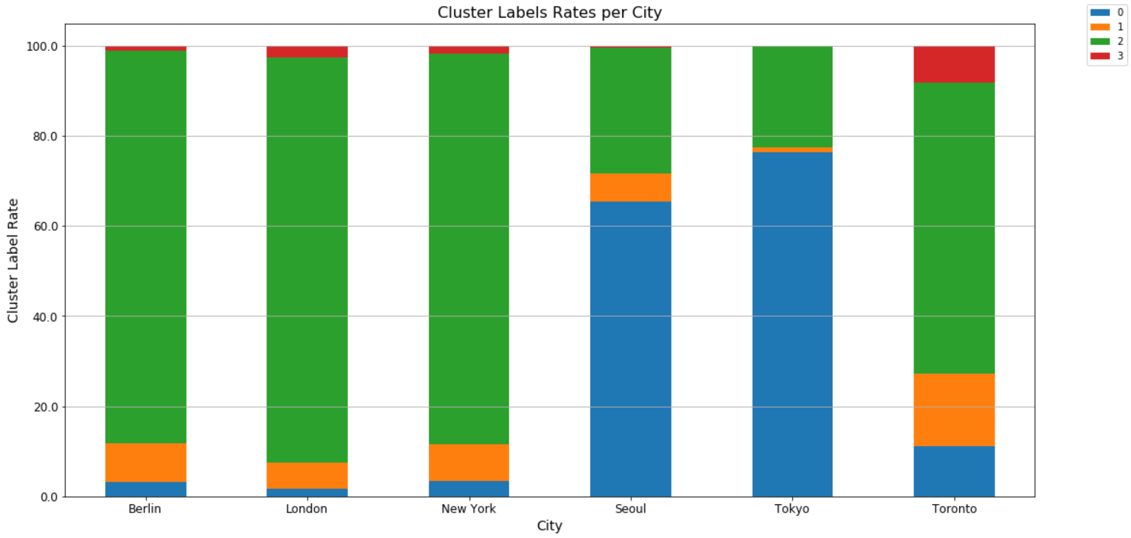


Figure 4: Proportions of cluster labels per city. Seoul and Tokyo have a majority of neighborhoods with Asian Restaurants while the 4 western cities majorly contain neighborhoods with bars and food & drink shops.

3.4 City Clusters

The hierarchical clustering of the cities reveals that Seoul and Tokyo form one cluster while the remaining 4 cities form a second cluster.

Out of the 4 western cities, Berlin and New York are the most similar followed by London and Toronto. Toronto is however more distant from the other 3 western cities than Tokyo and Seoul are from each other, suggesting Toronto might be its own singleton cluster or part of a cluster with cities other than the ones used in this analysis.

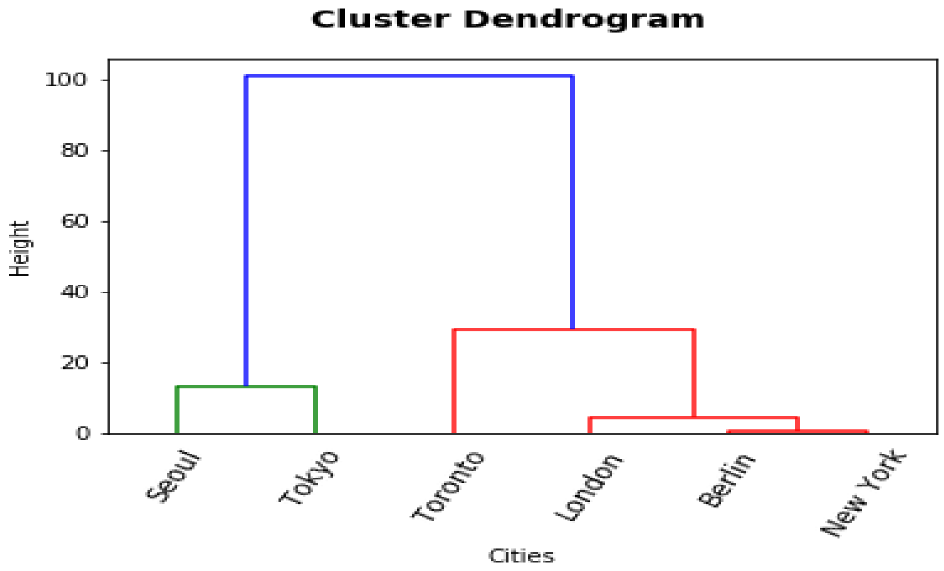


Figure 5: Dendrogram of the 6 cities. Seoul and Tokyo form one distinct cluster, while the western cities form another cluster. Toronto stands out as the most different from the western cities, New York is most similar to Berlin

4 Discussion

Based on the rates of the neighborhoods in the capital cities, capital cities were clustered to reveal the similarities between the capitals. With high rates of Asian Restaurants Neighborhoods, Seoul and Tokyo are similar to each other, while the western capital with higher varieties in neighborhood types cluster together. Out of the western capitals, surprisingly Berlin and New York were the most similar, while Toronto stands out in what it could be a possibly different cluster.

To the question of which capital city travel agencies could recommend to their clients given they liked New York and given our current data, the answer would be Berlin. This question could be expanded into a city recommendation engine for travel agencies, if we could gather data for not only other capital cities, but any city with the methods used in this study.

However great challenges lie in the data gathering steps given language barriers, non-systemized datasets and different definitions for neighborhoods depending on countries.

5 Conclusion

The purpose of this analysis was to identify the most similar out of 5 capital cities to New York in order to aid travel agencies who wish to recommend travel places more efficiently to their customers who liked New York.

Neighborhood data and geospatial data were gathered from different sources such as, Wikipedia, large csv files and JSON files.

Venues data were gathered using the foursquare API. These datasets were merged and the most common venues were identified for each neighborhood. These neighborhoods were then clustered based on their most common venues.

Given these new cluster labels, Capitals were further clustered to reveal similarities between capital cities.

Expansion of this analysis into a full recommendation engine could be built by gathering data from multiple cities and carrying out a similar analysis.