# Comparing international cities for their similarities and differences

Patrick Bosch

April 2020

## 1 Introduction

### 1.1 Background

Since centuries urbanization and concentration of industry around centralized locations led to the evolution of ever-growing cities. In combination with globalization, this led to large and international cities that have millions of inhabitants. Usually, countries and even cities have their specific properties that are different from other cities and countries. International cities have an increase in the diversity of inhabitants, mostly due to immigration and ex-pats, either temporarily or permanently. This change in population might change the properties of the cities. Around the world, many such large cities formed, and we want to compare them according to their specific properties.

### 1.2 Problem

In this report, we will focus on three such cities, more concretely, New York, London, and Paris. We want to answer the question, what the similarities and differences of these cities are. We explore them in terms of social and shopping places, as well as how neighborhoods are structured.

### 1.3 Interest

Two main groups are interested in this report. First, the people who plan to move to such a large city and need to make a choice which one might be the best fit for them. Second, companies who want to move a branch to such a city and want to make sure their workers feel comfortable in a similar environment as their original location.

# 2 Data acquisition and cleaning

## 2.1 Data sources

We use three data sources for our analysis. First, we use Wikipedia to get data for neighborhoods and their boroughs for each city. Usually, a city is partitioned into boroughs or districts which are further partitioned into neighborhoods. We are interested in the neighborhoods and their properties so that we can cluster and compare them later on. Second, we use location data from OpenStreetMaps and Google to get coordinates for each neighborhood. The location data for each neighborhood is necessary so that we can get additional information (venue data) for each neighborhood. Third, we use Foursquare to get information about venues for each neighborhood. For each neighborhood, we get types of venues, as well as how many of each type exist. The clustering algorithm will use these venues to decide which neighborhoods are similar to each other.

## 2.2 Data cleaning and processing

### 2.2.1 Neighborhood and location data

While Wikipedia provides borough and neighborhood data for each city, we need to clean and process it differently for each of them.

A table provides the data for London readily[1]. Still, it does provide information for the surrounding areas as well. Therefore, we need to first filter out the data according to the *post town*. After that, we can drop additional information such as post town, postcode district, the dial code, or the OS grid ref. We do need to clean the borough information, though, as it contains references and is not fully comma separated. Similarly, the neighborhood section also contains alternate names for some of the neighborhoods, which we remove so that getting location coordinates later on works properly. Additionally, we correct one neighborhood that is misspelled: Somerstown instead of Somers Town. We also detected that several pairs of neighborhoods have the same name but a different borough. We need to take this into account later on when we aggregate data for each neighborhood. OpenStreetMap does not always provide location data for a neighborhood, so we additionally use the Google API. We also make sure with a sanity check that all of the neighborhoods are within 30km of the city center with one exception that is 30.3 km away and is still accepted.

The data for New York is also available in a singular table[2]. Similar to London, it also has additional not needed information that we remove. Specifically, the area, population, and population per area are not of interest to us. The data is structured according to the community board with aggregated neighborhoods. So we split it into one neighborhood per row while removing information about community boards and keeping only the borough information. Similar to London, we use OpenStreetMaps and the Google API to get location coordinates

[1] `https://en.wikipedia.org/wiki/List_of_areas_of_London`

[2] `https://en.wikipedia.org/wiki/Neighborhoods_in_New_York_City`

and use a sanity check to confirm all coordinates are within range of the city center. Again, there is one exception with 30.4 km distance.

The data for Seoul is also available in one table[3], although one district is missing information about its neighborhoods. We need to merge that information from the district website into our table[4]. We remove the data about general information of a district, or points of interests so that we end up with neighborhoods and boroughs only. The neighborhoods are aggregated by district, so we split them into single rows. Again, OpenStreetMaps and Google were used to acquire location information, and we perform a sanity check to make sure the coordinates are close to the city center.

### 2.2.2 Venue data

We acquire the venue data for each neighborhood in each city through Foursquare. For each neighborhood, we request 200 venues in a radius of 750 m around the center of it, but the API seems to limit it to 100 venues. The data does not require further cleaning but requires preprocessing. We use one-hot encoding to mark each venue and then group them by neighborhood and borough and take the mean value to get weights for each category. Grouping by neighborhood and borough prevents the merging of neighborhoods with the same name. With this, the data is ready for further processing.

[3] `https://en.wikipedia.org/wiki/List_of_districts_of_Seoul`
[4] `https://en.wikipedia.org/wiki/Jung_District,_Seoul`