Comparing the similarities of different cities across the globe

Priya Dwivedi

May 2020

1 Introduction

An ill planned city can be a pain for people living in that city. For example, it is important to have enough public space, parks so that families can spend their quality time or to do any exercise. Similarly, for young people, it becomes important to have enough bar, pub, cafe where they can spend their quality night life. Until now, people just move to the city where they find good jobs. However, living in a city without enough amenities can be mentally challenging.

At the same time, municipalities are trying to attract young and skilled labours to their city in order to improve the city's economic activity. Therefore, the department of marketing works together with the urban planner to improve the city and utilise the space efficiently.

To help the urban planner or the person who is looking for a new city to move, we will examine the city's neighborhood and assess its similarities with other cities across the globe. In this project, we will answer the following question:

- Which of the cities falls in the same cluster?
- Which of the venue categories are dominated in a particular city or cluster?
- What actions need to be taken in order to improve the resident lifestyle of that particular city?

We will be investigating multiple cities from Europe, Asia and Australia. This analysis will be performed using Foursquare API.

The targeted audience for this project will the urban planner as well as the people looking for a new city to move.

2 Data

After defining the problem, the next step for data science project is to find the relevant data in order to solve the problem. This step is detailed in the sub-section Data acquisition. Furthermore, this will be followed by data pre-processing.

2.1 Data acquisition

First, we selected a number of cities from Europe, Asia, and Australia. In order to find their latitudes and longitudes, we used geopy library in python. These latitudes and longitudes were used as an input to Foursquare API in order to explore the near by venues. We chose to explore the area around the city center by defining the radius of 8 km and limited the number of venues to 100.

For example, if x y and z are the cities to be investigated, the following steps were performed to extract the relevant data.

1. Found the latitude and longitude of cities using geopy python library.
2. These latitudes and longitudes would be the input for foursquare API. Using foursquare API, explored the nearby venues of the city.
3. Clean and pre-process the data.

In the following sub-section, we will detail about the pre-processing step.

2.2 Data pre-processing

The collected data via Foursquare API had venue name, latitude, longitude and venue category. After dropping the non-significant columns e.g., venue name, latitude and longitude and adding a significant column, e.g., city name, we created a dataframe containing the cities and venue categories. These venue categories were converted to dummy variables corresponding to each

city. We grouped the data by cities and took average of dummy variables corresponding to the same cities. In the next step, we selected the top 15th most occurred variables for each cities. This clean data was used for k-mean clustering algorithm to find the similarities between the cities.