Capstone project report week 4

Priya Dwivedi

May 2020

1 Introduction and background

To keep a happy and healthy life style, urban planning has become significant. It creates a balance between economic activities and sufficient space to nurture a healthy and happy lifestyle. While planning a new city or improving the infrastructure of an existing city, it becomes necessary to assess other cities across the globe. Moreover, if a person is planning to move to a new city, examining the neighborhood and comparison with the current city becomes necessary to take an informed judgement. Therefore, this project is targeting the urban planners and people who are planning to move or possibly to open a business in a new city.

In this work, we will examine the city's neighborhood and assess its similarities with other cities across the globe. We will be investigating multiple cities from Europe, Asia and Australia. This analysis will be performed using Foursquare API.

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.

2.2 Data pre-processing

The collected data via Foursquare API had venue name, latitude, longitude and venue category. After cleaning and pre-processing the data, 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.