IBM – Coursera

Data Science Specialization

Capstone project - Final report

Analysing and comparing student friendly neighbourhoods in USA

Kripa Shankar Muthu Kumar– 2020

I. Introduction:

The students (all graduate levels) who decides to go to a university strive to know the various characteristics and benefits of the city/location to determine the ease of living.

Both international and citizens search for common parameters in the city/state:

1. Housing

2. Transport facilities (public transport), traffic incidents (crash, weather)

3. Corporate companies and tech startups (internship/full time)

4. Sports and recreation

5. Restaurants, malls, music concerts and game night pubs/bars

6. Libraries, medical centres and more

I have added traffic incidents as the students/parents might like to know the safety of driving through the neighbourhoods.

Problem

We must enable the user to determine the best neighbourhood cluster that satisfies most of their parameters

In this project, I try to solve this problem using foursquare location data, New York crash data and machine learning algorithms to compare and find the best places for students to find universities and settle in.

Stakeholders

This project will be useful to students, parents and graduate students (job seekers). This will be useful to other third parties who want to see the behaviour and psychology behind students’ requirements that must be around the universities after learning what clusters usually those people fall into.

II. Data description:

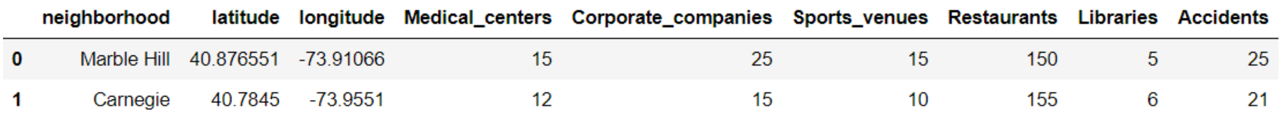
We pull location/venue data that belong to above categories from foursquare and traffic incidents data from the selected state website. In this project, we consider only one state, New York city and its neighborhoods.

I obtain Incident/crash data from New York Motor Vehicle Collisions – Crashes dataset. The dataset contains a vast number of features like latitude, longitude, number of persons injured/killed, number of pedestrians/cyclists injured, contributing factor vehicle 1 and more. But we just count the accidents per location or borough and append it to the dataset.

Regarding the venue data, it is same as what we saw in our previous exercise (Week 3). I will extract locations like universities, tech startups/corporate companies, restaurants/pubs, medical centers, events, etc. And, combine with the crash data.

For example, we extract venues using foursquare API: **https://api.foursquare.com/v2/venues/explore.** We get the data, normalize, find the frequency and append to the dataset.

Below is the example of the resultant data.



	neighborhood	latitude	longitude	Medical_centers	Corporate_companies	Sports_venues	Restaurants	Libraries	Accidents
0	Marble Hill	40.876551	-73.91066	15	25	15	150	5	25
1	Carnegie	40.7845	-73.9551	12	15	10	155	6	21

Also, this is a sample dataset for now. Dataset features can get modified as the project gets progressed.