April 16, 2021

The Battle of Neighborhoods

Comparison of neighbouring cities of US states.

Vishal Raj

IBM COURSERA CAPSTONE PROJECT

Introduction

As we all know shopping, visiting a particular cuisine restaurant/fast-food joint or movie theatres gives many people a way to relax and entertain themselves. But due to unavailability or services that are far away, they have to travel long distances which results in exhaustion before even they can enjoy properly. Also, many stakeholders, realtors & business peoples are searching for many better ways to find out the best possible solution in searching of a profitable investment or projects. They are also interested in getting an idea about what could be the demand of the general public by seeing the unavailability or low competition in a particular business. Opening the right kind of business at the right place allows them to earn consistent profit without the worry of going into loss and at the end closing the business due to too much competition or no demand. It also profits the common people as now they will have the luxury to visit those places that are now in their vicinity without giving it much thought. Any business decision requires serious consideration and is lot more complicated than it seems. Thus, here we come with our best solution possible within our means.

Business Problem

The objective of this capstone project is to analyse and divide a group of similar cities into clusters of a given US state and also select the best locations of the cities of that state to open a new particular business. Using data science methodology and machine learning techniques like clustering, this project aims to provide solutions to answer the business question: In any of the US states which city would be best to open which business?

Target audience

This project is particularly useful to stake holders, investors, realtors, property developers and other business people looking to open or invest in any particular business profile. According to Convergehub (2019) more than 50 percent of small enterprises fail in the very first year, and more than 95 percent of small startups fail within the first five years. And according to CB Insights (2019), the primary reason that new businesses fail is because of a lack of market demand. In fact, 42 percent of small businesses fail because of this reason.

Data

To solve the problem, we will need the following data:

- List of US States and its neighbouring cities. This defines the scope of this project which is confined to USA.
- Latitude and Longitude coordinates of those neighbouring cities. This is required to plot the map and get the venue data.
- Venue data, all the top venues in any particular city. We will use this data to perform clustering of the neighbouring cities according to their similarity.

Source of data and methods to extract them:

This Britannica Encyclopaedia page (https://www.britannica.com/topic/list-of-cities-and-towns-in-the-United-States-2023068) contains a list of all US states and their neighbouring cities. We will use web scraping techniques to extract the data from the webpage page, and transform that data into a CSV file for easy execution and uses. Then we will get the geographical coordinates of the neighbouring cities of any particular state using Python Geocoder package which will give us the latitude and longitude coordinates of the neighbouring cities.

After that, we will use Foursquare API to get the venue data for those cities. Foursquare has one of the largest databases of 105+ million places and is used by many developers. Foursquare API will provide many categories of the venue data, which we will later use to categorize the cities into similarities/dissimilarities in order to help us to solve the business problems put forward like what is the probability of making a profit of opening a business structure similar to another city when there is no availability of that business in the present city. This is a project that will make use of many data science skills, from web scraping (Britannica Encyclopaedia), data cleaning & wrangling, working with Foursquare API, to map visualization (Folium) and machine learning (K-means clustering). In the next section, we will present the Methodology section where we will discuss the steps taken in this project, the data analysis that we did and the machine learning technique that was used.