Venues Data Analysis of Los Angeles for setting up a new business

Amlan Abhisek Nayak

June 3, 2019

1. Introduction

1.1 Background

Los Angeles ,officially the City of Los Angeles and often known by its initials L.A., is the most populous city in California, the second most populous city in the United States, after New York City, and the third most populous city in North America. The area is a large market in its own right, and it is still growing. There are 17.9 million residents in the five-county area, and nearly 10 million in Los Angeles County. The population increases between 2000 and 2010 were 1.5 million and 283,087 respectively. Census Bureau projections to 2025 indicate significant additional growth. In addition, there is quick access to markets in San Diego and Northern California, as well as Arizona and Nevada.

1.2 Problem

In a city like Los Angeles, there is enormous opportunity to establish a business like restaurant, mall , office etc. One among many problems lies in identifying the place where to setup your business. There are various neighbourhoods in Los Angeles and they are already crowded with venues like restaurants, recreation places, mall etc . Apart from this, the rent price varies in each and every neighbourhood. The challenge here is in understanding the demand , existing competition and infrastructure cost like rent.

1.3 Interest

For a businessman who wants to setup a restaurant, his interest would lie in two factors. First, he would prefer neighbourhoods where intensity of restaurants or that particular category of restaurant, for example, Chinese or Italian , is less and hence the competition will be less. Second, he would also consider the average rent price in that neighbourhood. Whereas, for a company who wants to setup an office would opt for places with certain type of social venues and also consider the rent price of that area.

Keeping in mind these two important factors, we will create a map where we can visualize the neighbourhoods clustered according to venue category , and the average rent price of the neighbourhood is also listed.

2. Data Acquisition and Cleaning

2.1 Data Sources

To consider the problem we can list the datas as below:

A. The rent price for Los Angeles neighbourhoods can be found here . The dataset has the rent price for each and every neighbourhood of Los Angeles from year 2010 to 2016. I couldn't find any reliable source for more recent data than this. Therefore, I used the data from year 2016 for the work. The dataset also contains location coordinates of these neighbourhoods.

B. I used **FourSquare API** to get the most common venues for each neighbourhood of Los Angeles.

2.2 Data Cleaning and Feature Selection

There were several features or columns in the dataset mentioned above. First, I cleaned the dataset keeping only four features namely Neighborhood, Location, Rent($), Year. After that I extracted the data from year 2016 and used it for further processing. The location data was in a single column enclosed within parenthesis. I extracted the location coordinates from 'Location' feature and added two new columns namely Latitude and Longitude. I removed the 'Year' column since all the rows are from the year 2016. Finally, the dataset contributed to total four features i.e Neighborhood, Rent($), Latitude, Longitude. The rest of the features were related to venues which I got it by using FourSquare API for each neighbourhood. I got atmost 100 top venues for each neighbourhood in Los Angeles and their category. The list of venues were used as features for clustering of neighbourhoods according to venue density.