

Coursera Capstone

IBM Applied Data Science

***Setting up a new Bar in the city of Los Angeles,
United States of America***

By: Aniruddh Mallya
8 August 2020



Introduction:

For many drinkers as well as nondrinkers, visiting bars is a great way to relax and enjoy themselves after a long day of work.

They can spend quality time with the folks and also get to dine at a few places. It's like a one-stop destination for all get togethers.

For people who own these bars, a larger crowd provides a good source of income, in order to run their businesses successfully.

Owing to these reasons, we observe there are many bars in locations of a city with a larger crowd, especially the central locations. Many more are being built.

The location plays a major role in determining success in most of the businesses, same is true with bars as well.

Business Problem

The location of setting up public infrastructure/venues is a major factor that decides as to whether a business would be successful or not.

Problem Statement:

The main objective of this Capstone is to answer the question as to **which locations in Los Angeles (LA) would be most preferable to set up a new bar business.**

- On visualizing the spread of bars across various locations in LA, we observe patterns that make this project relevant and timely.

- By making use of various data science methodologies and Machine Learning tools, this paper aims to provide a conclusive proof to the underlying problem statement.
-

Data Required

- List of neighborhoods/districts in Los Angeles. This confines the locations to a specific boundary.
- Latitude and Longitudes coordinates, to plot maps and to get information about the venues.
- Venue data corresponding to bars will be helpful to perform clustering based on various categories.

Sources of Data

- Wikipedia page of neighborhoods

https://en.wikipedia.org/wiki/List_of_districts_and_neighborhoods_of_Los_Angeles

The above site contains neighbourhood names of Los Angeles from A-L and M-Z.

We make use of the flexibility of Python programming language and make use of the BeautifulSoup library to parse the html document and extract what is required.

➤ Foursquare API:

This API consists of a huge database consisting of venues, tips corresponding to various locations. We use this to obtain the most popular venues at various neighborhoods in Los Angeles, USA.

Methodology

Web Scraping is done from Wikipedia to get a list of neighborhoods in Los Angeles. But before that we need to import various Python Libraries which makes programming a lot simpler.

We make use of the Geopy and Geocoder packages which are Python libraries that we can use to obtain latitude and longitude coordinates from the corresponding addresses.

Later we perform data extraction by passing in the latitude and longitude coordinates to the Foursquare API, that returns a list consisting of top 100 venues, lying in the vicinity of each of the specific coordinates.

For this purpose, we take a radius of 5000 m at a particular location, that covers a significant area and gives better results.

Foursquare will return the data in JSON format and then we extract the names, latitude, longitude and category of the venue. We check the number of unique venues, the most popular ones

and how frequently they occur at various places. We then filter out locations based on the relative occurrence of one of the top venues, which in our case is “Bars”.

We make use of Data Science techniques such as Data cleaning and Data wrangling to reduce the size and improve readability of the data.

Lastly, we perform KMeans clustering in order to segment the city into a finite number of clusters, which is determined by making use of the “Elbow Curve” method.

KMeans clustering algorithm identifies k number of centroids, then allocates every data point to the nearest cluster, while keeping the centroids as small as possible.

We will cluster the neighbourhoods into 3 clusters based on the frequency of occurrence of a “Bar”.

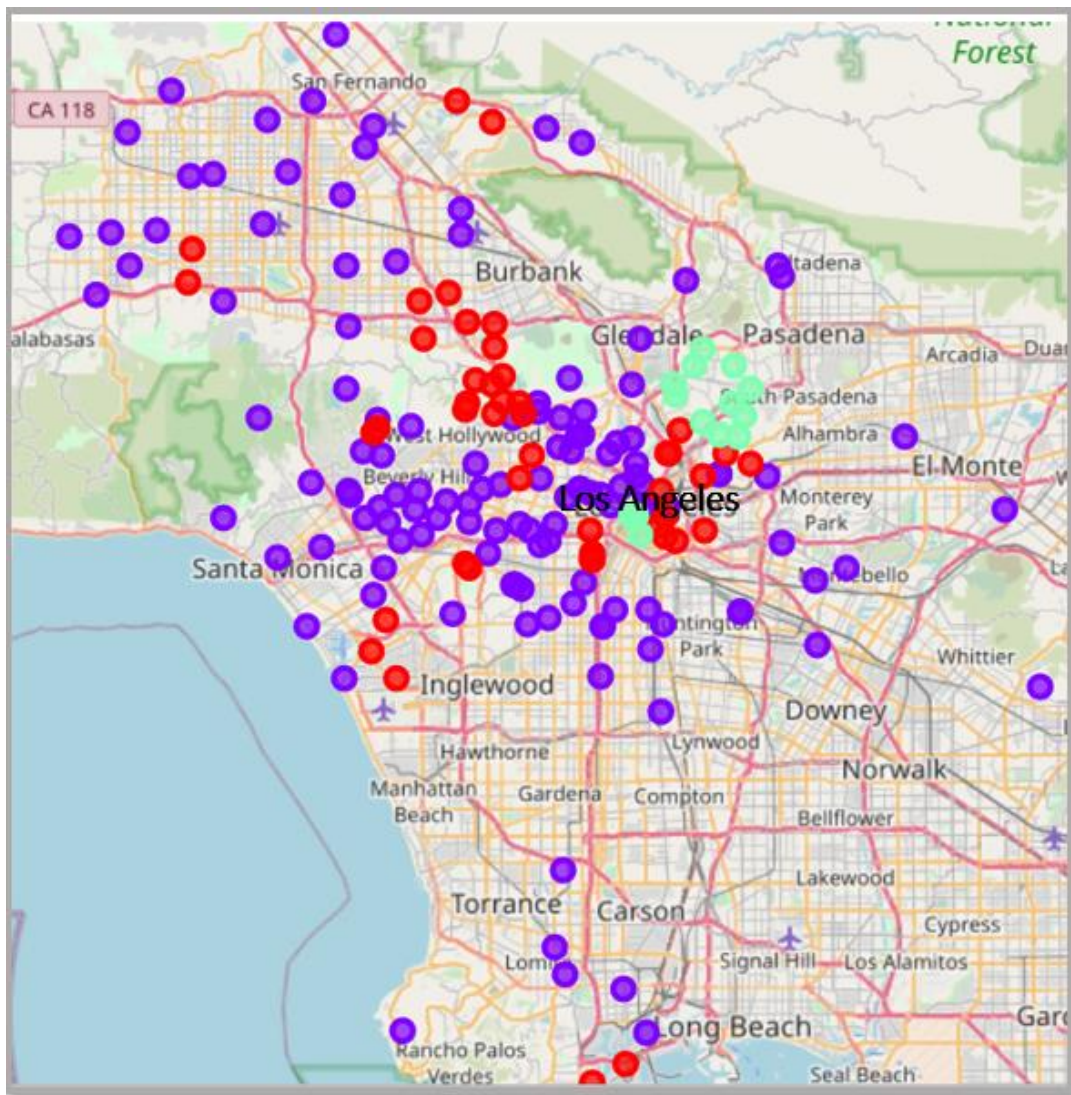
The results will help us identify which neighbourhoods have higher concentration of shopping malls and the ones that don't.

This will ultimately help us decide as to which neighbourhoods is most suitable to set up a bar.

Results

The neighborhoods were segmented to 3 clusters:

- Cluster 0: Moderate density – Red color
- Cluster 1: Low density – Purple color
- Cluster 2: High density – Pale Green color



Discussions

Most of the bars are concentrated in the central area of the city. Suburbs have fewer number of bars, so that people residing here have to travel farther for their drinks. Sometimes this can lead to the situation of unnecessary crowding at specific locations, oversupply of items such as drinks and lead to an imbalance in the supply chain.

- As far as possible, try and avoid neighborhoods in ‘Cluster 2’, since there’s already a high concentration of the number of bars implying very high competition.
- Since there is far less competition in ‘Cluster 1’ neighborhoods, new businesses have a great opportunity to quickly establish themselves and capitalize.
- For the existing businesses that are well established, diversifying can contribute to improved productivity. However, in order to avoid taking too much of a risk, ‘Cluster 0’ with low-moderate density would be seemingly a good option.

Conclusion

- In my opinion, cluster 1 would be best preferred location to set up a ‘bar’ business, which answers the main business question.
- Thus, the findings of this paper will help the relevant stakeholders to capitalize on their opportunities on high potential locations while avoiding overcrowded areas with high competition to open a new bar.

References:

- Wikipedia page of neighborhoods
https://en.wikipedia.org/wiki/List_of_districts_and_neighborhoods_of_Los_Angeles
- Foursquare Documentation
<https://developer.foursquare.com/docs>

