Capstone Project

IBM Data Science Professional Certificate

Introduction

Background

Texas and California are considered to be significantly different states. Especially in terms of political approach, where California is a liberal-leaning state while Texas is a conservative-leaning state. The cost of living is also considerably higher in California than in Texas. Texas is known for its unpredictable weather. These, among other things, make moving between these states is a unique experience for movers. I attempted to find rudimentary comparability between cities in these two sates using K-mean clustering.

Business Problem

If a person wants to move between two polarizing states of California and Texas, they would like to see what are the cities that have similar characteristics in terms of amenities. It would give them peace of mind knowing that they could have similar access to the businesses and entertainment options once they moved to the new city. So, I tried to answer the problem: What are the similar cities between Texas and California?

Target Audience

The primary audience for this project would be the potential movers between states of California and Texas.

Data

I will use the list of cities of California and Texas from the following sources.

Texas: https://www.texas-demographics.com/cities_by_population

California: https://www.california-demographics.com/cities_by_population

Data will contain the list of cities and the respective population. I will remove big cities from the analysis as those would have an abundance of amenities and can easily be compared to big cities in the other state. Therefore, not suitable for this type of basic comparison.

I will follow the following steps for obtaining and preparing data.

1. Use web scraping techniques to obtain the list of cities from above publicly available pages. Python requests and beautifulsoup packages.

2. Obtain the latitude and longitude coordinates for the cities using Python Geocoder package or MapQuest Geocoding API.
3. Obtain the list of venues (amenities) for those cities using Foursquare API

Data Cleaning and Methodology

There are 1,354 total cities in California and 1,434 in Texas. However, this analysis and comparison are based on basic amenities found within the proximity to the cities. Very large cities like Los Angeles or Houston typically have all of these amenities and easily comparable due to the nature of the analysis. Therefore, I left out cities with a population of more than 250,000. On the other hand, both California and Texas have lots of smaller cities that are remote. Assuming that the target of the study is the potential movers who probably not looking to move into these remote smaller cities, I removed the cities with a population less than 10,000. After the adjustments, I have a total of 713 cities in California and Texas to compare.

Obtaining location data: I use https://www.mapquest.com/ to obtain the latitude and longitude data for each of the locations. After that, I used the Folium library to obtain the USA map and map all of the above cities.

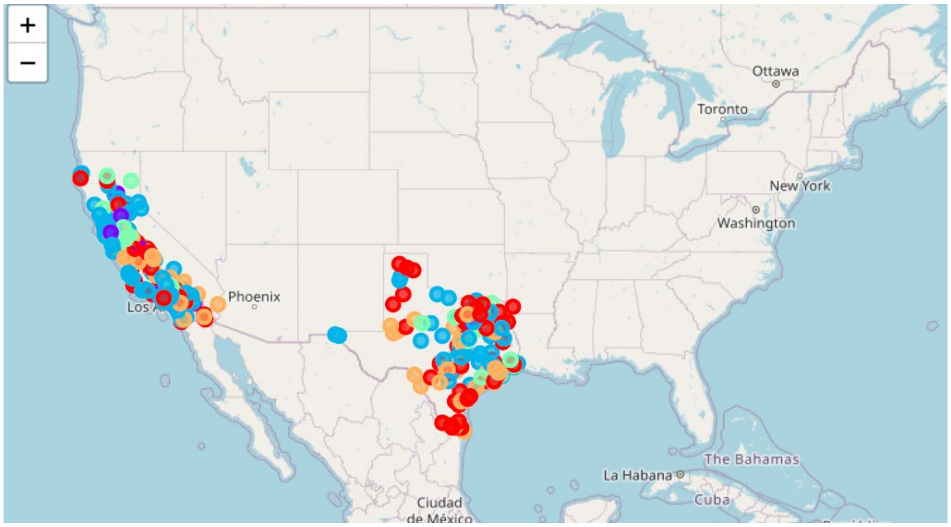
Information on amenities: I used Foursquare API to obtain the common amenities of each of the cities. For each city, I obtained 50 venues within a radius of 1 km. To understand the type of data I got, I looked at some aggregate data. First, I observed how many venues I got for each town. Due to the differences in sizes and availability of data in Foursquare API, towns might not return 50 venues all the time. Having a higher number of venues is better in order to make a better comparison. Therefore, I dropped all the cities which returned venues less than 5. After the adjustment, I have 674 cities from both states. I also have 498 unique venue categories in the data set. Moreover, I also looked at the most common venues in each city.

K-mean Clustering

I use a popular clustering method of K-mean clustering to put cities into similar groups. For the clustering exercise, I used five most common venues in each city. These five most common venues work as the features of the K-mean clustering. I also set the number of clusters as 5. Once I obtain the clusters, I used folium to map all the clusters in the US map for visualization.

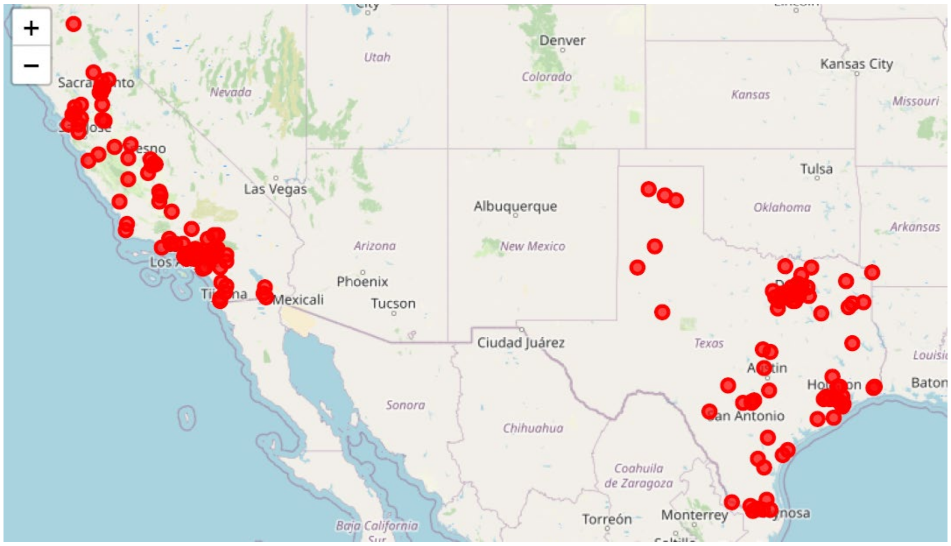
Results:

Results show five distinctive clusters of towns based on the top amenities they have. The figure below shows the resulting clusters. As it shows, clusters are scattered around two states.

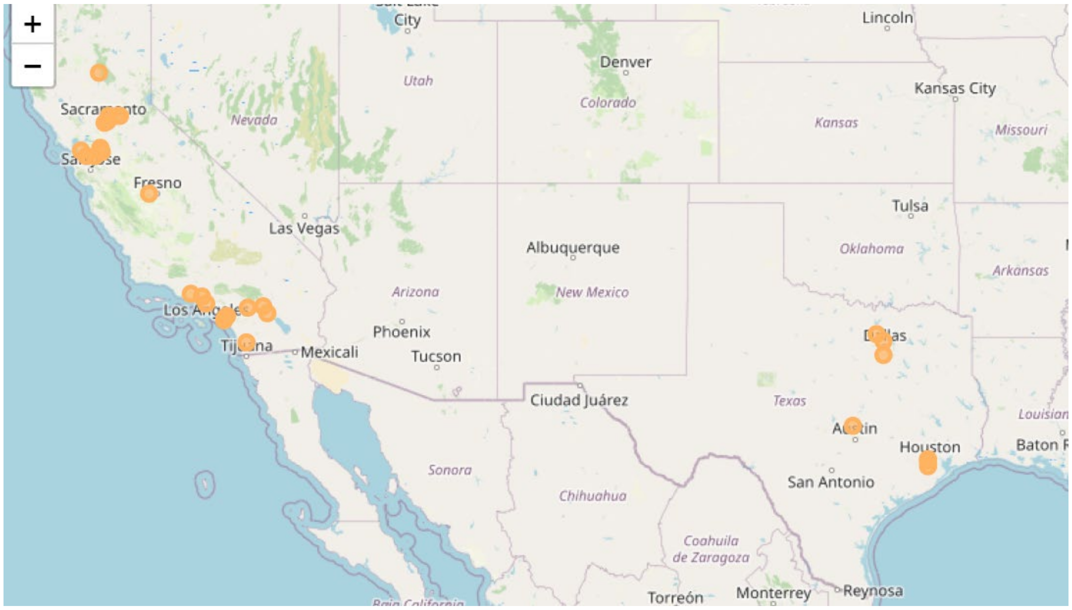


Discussion:

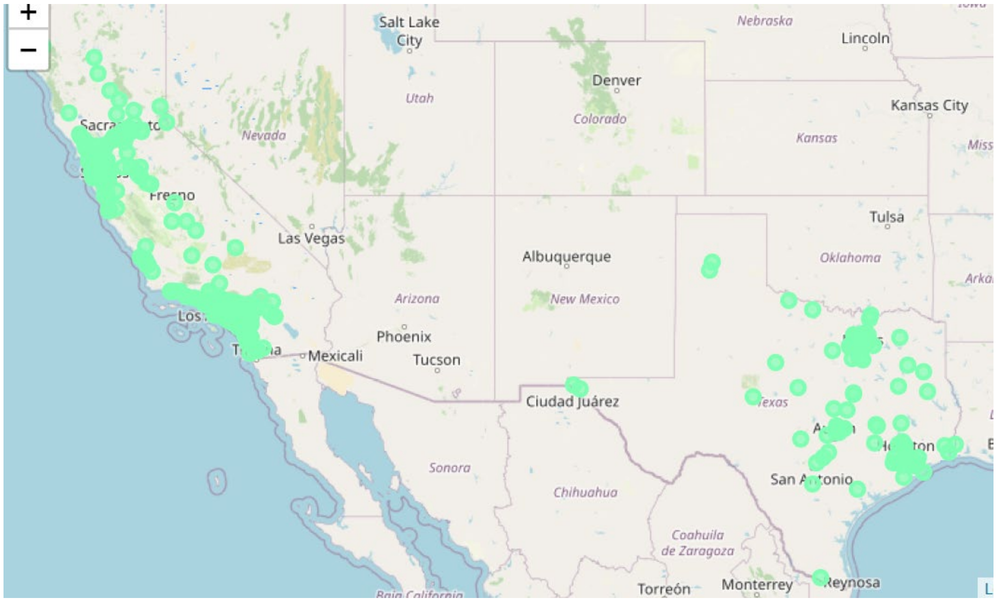
Cluster 1: This cluster could be named as “Fast Food Friendly.” The most common venues are fast food restaurants, pizza places, etc. These types of venues are looking prevalent across all top five venues in these cities.



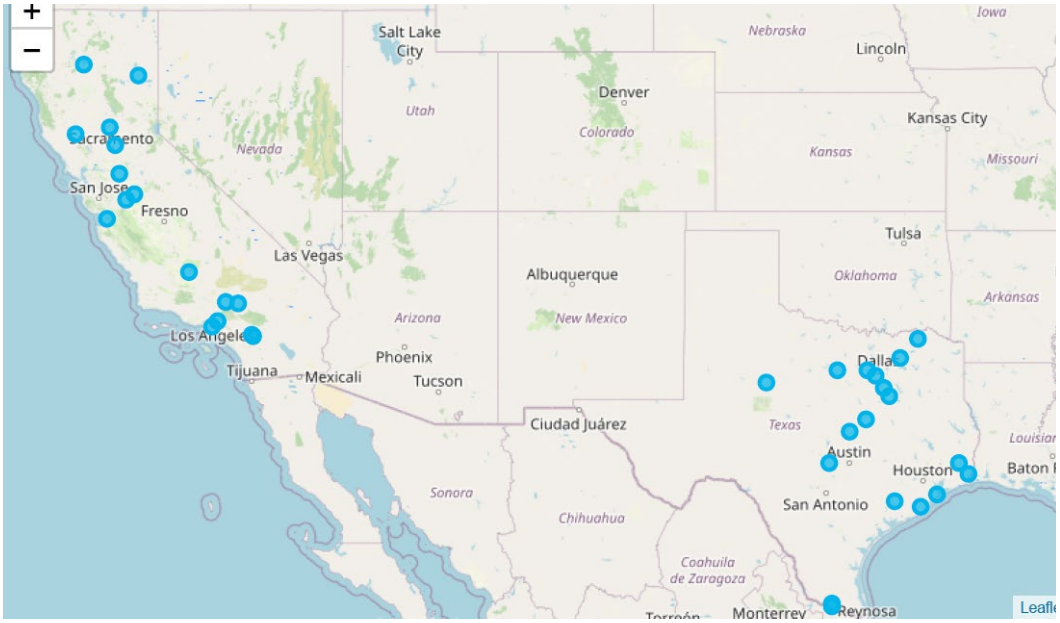
Cluster 2: This cluster could be named as "Urban Outdoor Friendly." The most common venue in this cluster is the parks. However, as the following figure shows, these cities are located near big cities like Dallas, Los Angeles, etc. This suggests that the parks are urban parks. So, people can be close to big cities at the same time have access to parks and recreation.



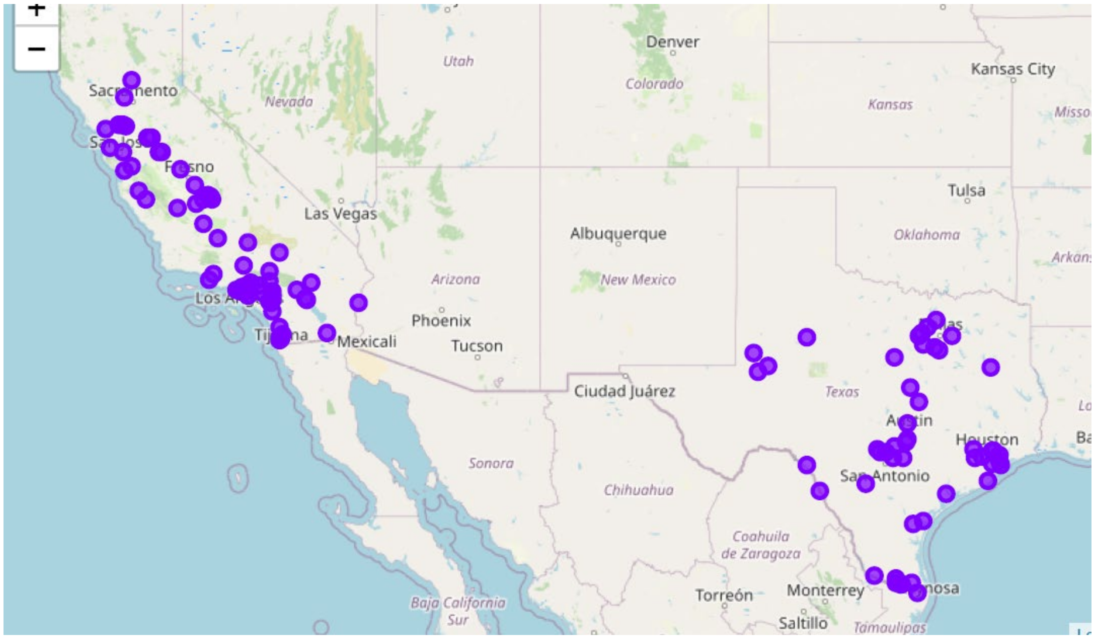
Cluster 3: This cluster could be named as "Night-Life Friendly." These places have a variety of places for people from bars to cafes to theaters.



Cluster 4: This cluster could be named as “Easy Access Friendly.” These cities have access to a lot of amenities at the same time, relatively close to a big city. However, not close enough to face traffic.



Cluster 5: This cluster could be named as “Latin American Food Friendly.” These cities show a clear identity common to both states. Both CA and TX have a large Hispanic population, thus representing the common availability of Mexican and Latin American restaurants.



Limitation and Recommendation

I used K-mean clustering to group similar cities across Texas and California. The main limitation of the project is the use of limited information. Clustering is entirely based on the most common venues in a particular city. However, a potential mover would require a lot more information than that. For example, types of schools, taxes, crime statistics, etc. needed to be added. Moreover, this analysis based on the availability of amenities that leaves out the quality of those places. Therefore, an improved analysis needs to include a wide variety of places and an assessment of the quality of those places.

Conclusion

In this project, I attempted to find similar cities across California and Texas, based on amenities they have. The main objective is to help people who are planning to move between two cities to find a comparable place to move. I limited my study to cities with at least 5,000 population and no more than 250, 000. I used several techniques to obtain the city data from two online sources, then to get the location data from MapQuest API. Then using venue information from Foursquare API and employing K-means clustering methodology, I divided cities into five similar clusters. Cluster were names based on their common characteristics, and they are “Fast Food Friendly,” “Urban Outdoor Friendly,” “Night-Life Friendly,” “Easy Access Friendly,” and “Latin American Food Friendly.”

A potential mover can use these clustering to find a similar city from another state, or even from your own state, based on the amenities they offer. However, the clustering is based on basic information of venue categories and can be significantly improved by incorporating much other information like schools, quality of schools, etc.