CALIFORNIA ⇄ TEXAS

Capstone Project

IBM Data Science Professional Certificate

WHAT I AM LOOKING AT ?

- Texas and California are considered to be significantly different states.
- 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:** potential movers between states of California and Texas.

DATA

- Source of Data:
  - Texas: https://www.texas-demographics.com/cities_by_population
  - California: https://www.california-demographics.com/cities_by_population

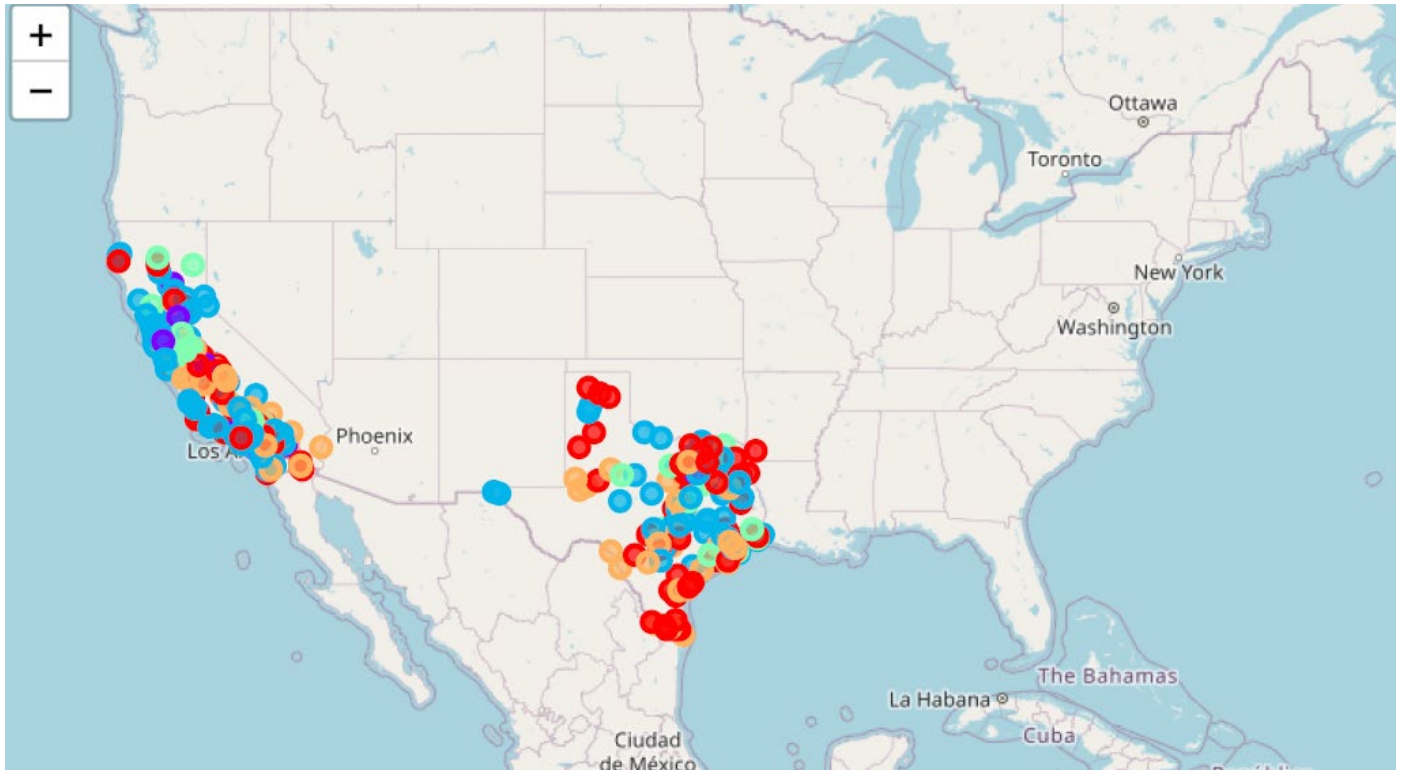
- Three steps to get the Data:
  - Use web scraping techniques to obtain the list of cities.
  - Obtain the latitude and longitude coordinates for the cities using the Mapquest Geocoding API.
  - Obtain the list of venues (amenities) for each city using the Foursquare API

METHODOLOGY

- 1,354 total cities in California and 1,434 in Texas.
- Only selected cities with 10,000 < population < 250,000
  - Very large cities like Los Angeles or Houston typically have all of these amenities and easily comparable due to the nature of the analysis. T
  - On the other hand, both California and Texas have lots of smaller cities that are remote. Assumes that the potential movers are probably not looking to move into these remote smaller cities
- MapQuest API to obtain the latitude and longitude data for each of the locations. Folium library to obtain the USA map.
- API to obtain the common amenities of each of the cities.
  - For each city, 100 venues within a radius of 1 km.
  - To understand the cities: First, I observed how many venues I got for each town. 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 10.
- After the adjustment: I have 706 cities from both states. I also have 498 unique venue categories.

METHODOLOGY

- **K-mean Clustering**
- I used five most common venues in each city. These five most common venues work as the features of the K-mean clustering. Set the number of clusters as 5.
- Folium to map all the clusters in the US map for visualization.



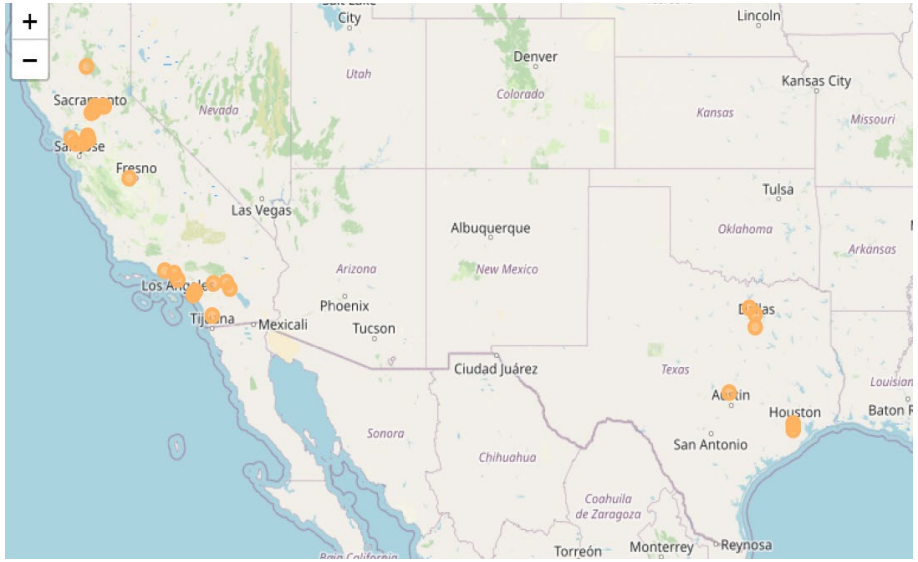
RESULTS

Results show five distinctive clusters of cities based on the top amenities they have.

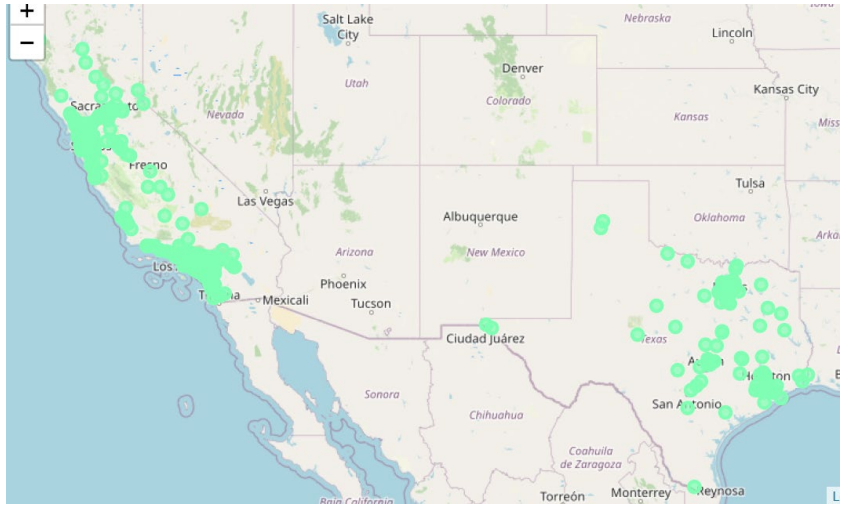
DISCUSSION

CLUSTER 01:

FAST FOOD FRIENDLY

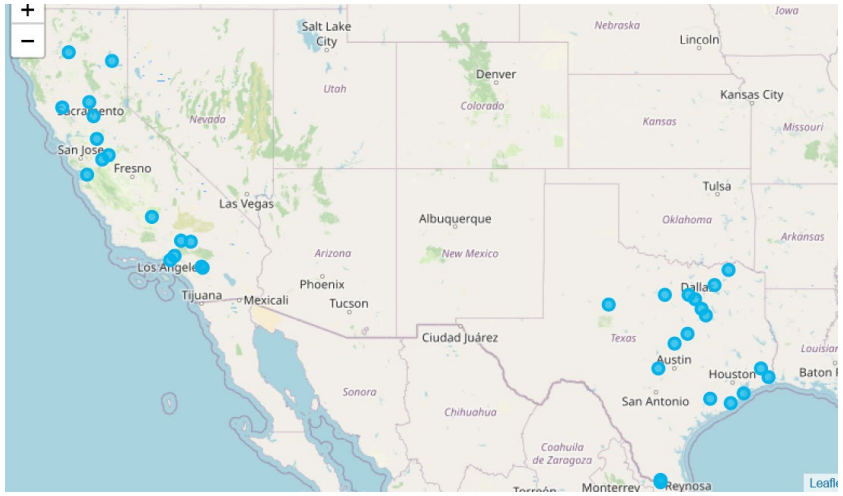


CLUSTER 2: URBAN OUTDOOR FRIENDLY



CLUSTER 2:

NIGHT-LIFE FRIENDLY



CLUSTER 2:

EASY ACCESS FRIENDLY

CLUSTER 2: LATIN AMERICAN FOOD FRIENDLY

LIMITATION AND RECOMMENDATION

- The main limitation of the project is the use of limited information.
- A potential mover would require a lot more information than that. For example, types of schools, taxes, crime statistics, etc. and quality of venues
- 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 10,000 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 gave 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.