April 19, 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.

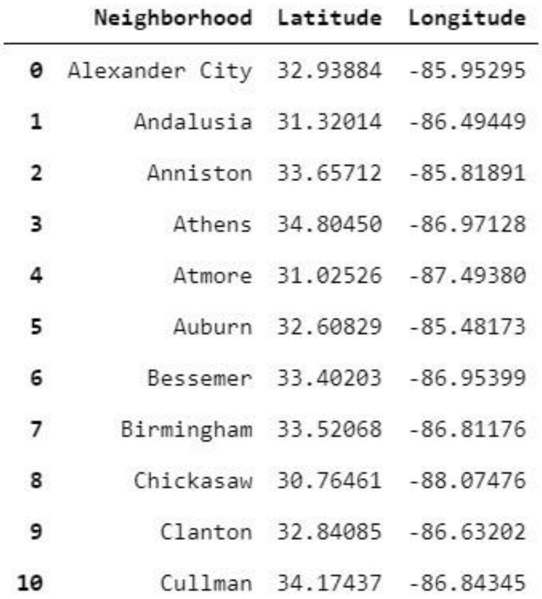
Source of data and methods to extract them:

This Britannica Encyclopedia 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 Encyclopedia), 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.

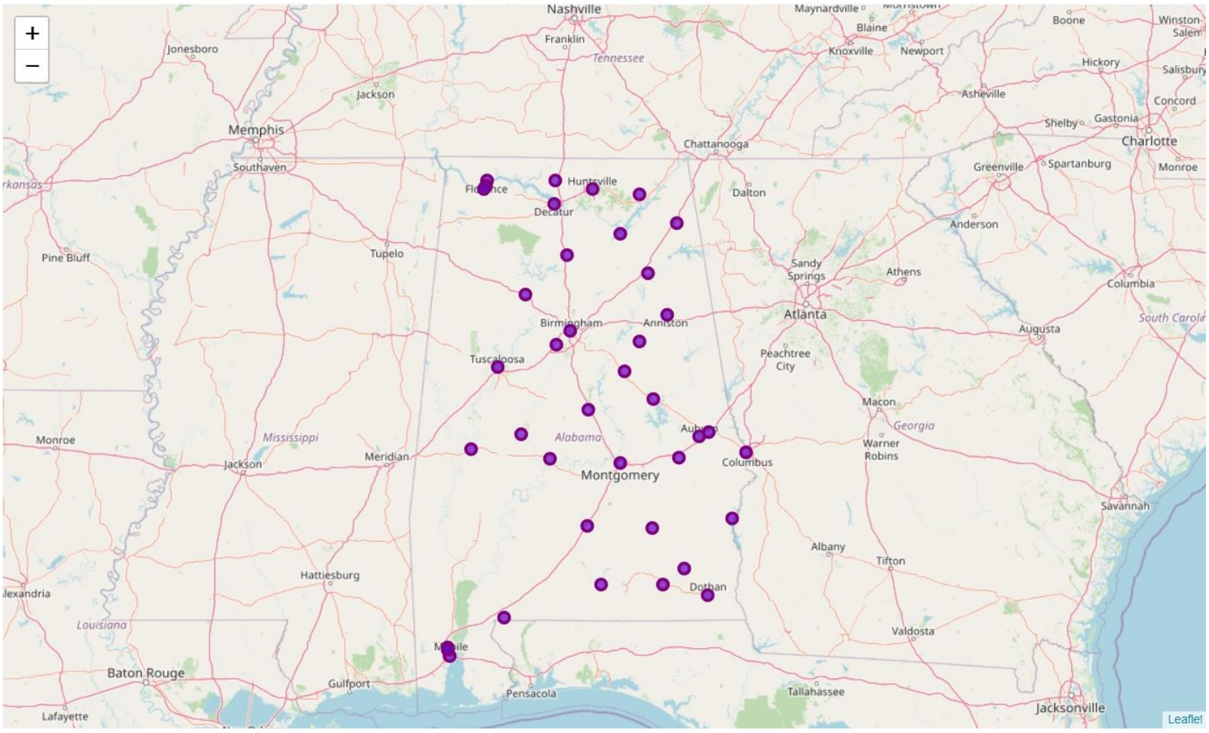
Methodology

At first, we need to get the list of all US states and their neighbouring cities. Fortunately, this list is available at the Britannica Encyclopedia page (https://www.britannica.com/topic/list-of-cities-and-towns-in-the-United-States-2023068). After that, we will do web scraping using Python requests & beautifulsoup packages to extract the list of neighbouring cities data. Then we will convert that into a DataFrame and CSV file for easy execution and use. We also need to get the geographical coordinates in the form of latitude and longitude of all the cities of the selected state (Here I have taken Alabama) in order to be able to use Foursquare API. To do so, we will use the Geocoder package that will allow us to convert address into geographical coordinates in the form of latitude and longitude.



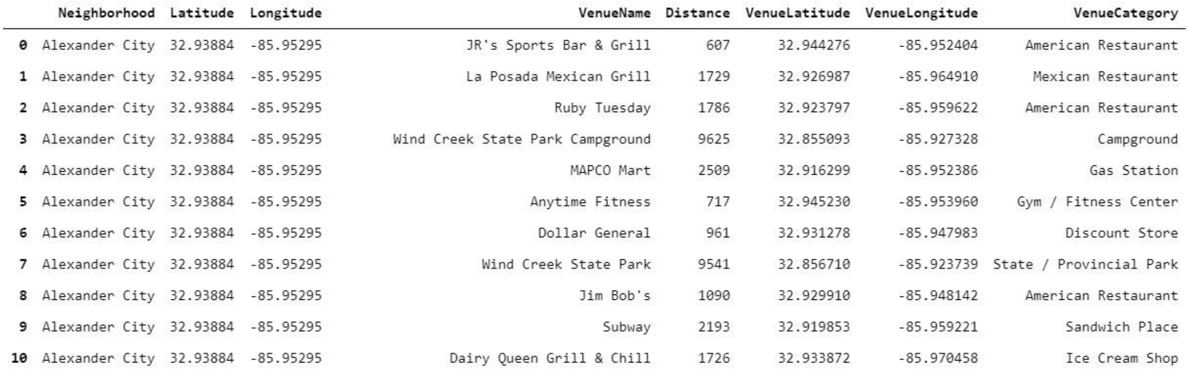
	Neighborhood	Latitude	Longitude
0	Alexander City	32.93884	-85.95295
1	Andalusia	31.32014	-86.49449
2	Anniston	33.65712	-85.81891
3	Athens	34.80450	-86.97128
4	Atmore	31.02526	-87.49380
5	Auburn	32.60829	-85.48173
6	Bessemer	33.40203	-86.95399
7	Birmingham	33.52068	-86.81176
8	Chickasaw	30.76461	-88.07476
9	Clanton	32.84085	-86.63202
10	Cullman	34.17437	-86.84345

After gathering the data, we will put the data into a pandas DataFrame and then visualize the neighbouring cities in a map using Folium package with the coordinates superimposed on top. This allows us to perform a sanity check to make sure that the geographical coordinates data returned by Geocoder are correctly plotted in the state of Alabama. Following visual was obtained from the DataFrame:



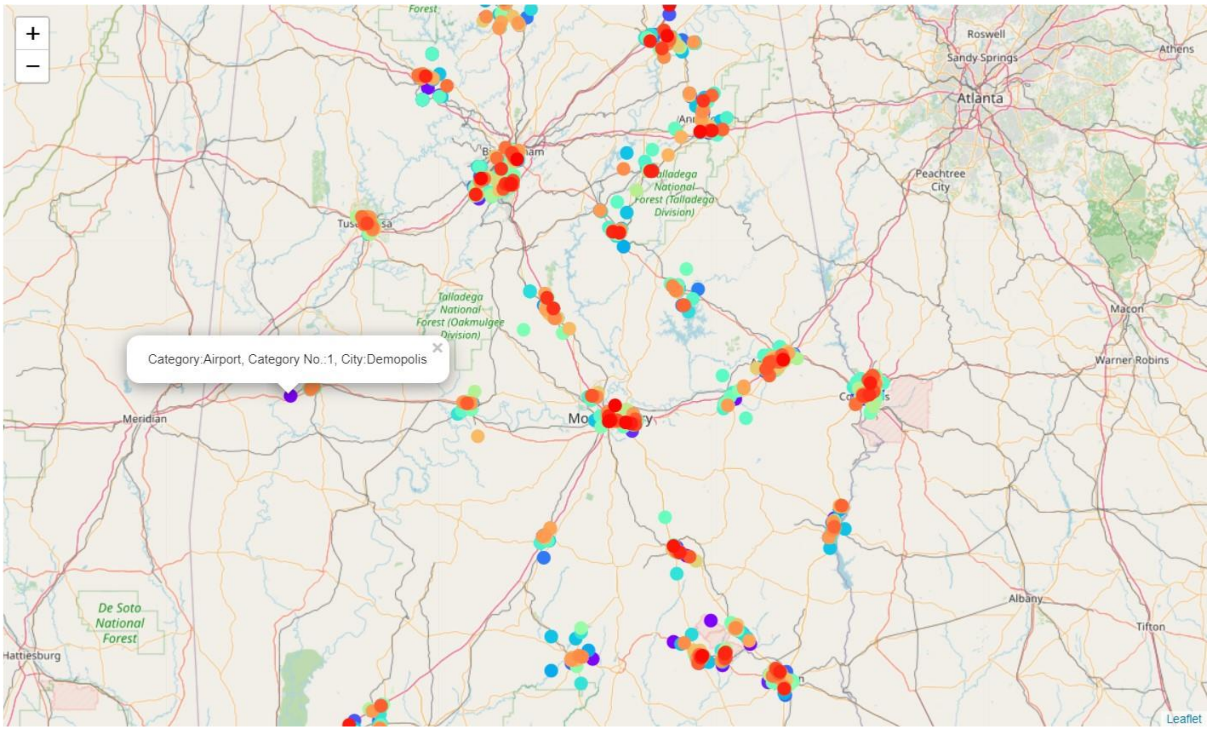
State of Alabama

Next, we will use Foursquare API to get the top 1400 venues that are within a radius of 15 KMs. (You can change the limit of venues & radius according to your preferences.) We need to register a Foursquare Developer Account in order to obtain the Foursquare Credentials. We then make API calls to Foursquare passing in the geographical coordinates of the neighbourhoods in a Python loop. Foursquare will return the venue data in JSON format and we will extract the venue name, venue category, venue latitude & longitude and distance. Here is a head of the list Venue names, categories, latitude & longitude information from Foursquare API.



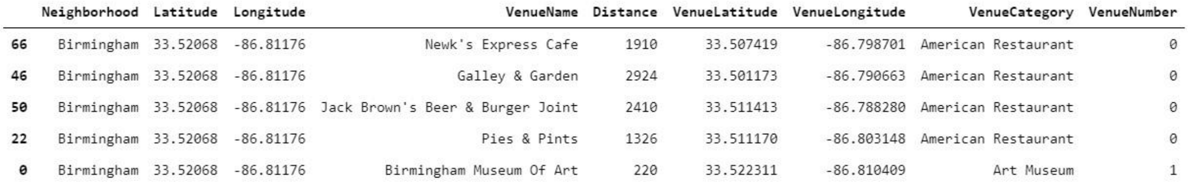
	Neighborhood	Latitude	Longitude	VenueName	Distance	VenueLatitude	VenueLongitude	VenueCategory
0	Alexander City	32.93884	-85.95295	JR's Sports Bar & Grill	607	32.944276	-85.952404	American Restaurant
1	Alexander City	32.93884	-85.95295	La Posada Mexican Grill	1729	32.926987	-85.964910	Mexican Restaurant
2	Alexander City	32.93884	-85.95295	Ruby Tuesday	1786	32.923797	-85.959622	American Restaurant
3	Alexander City	32.93884	-85.95295	Wind Creek State Park Campground	9625	32.855093	-85.927328	Campground
4	Alexander City	32.93884	-85.95295	MAPCO Mart	2509	32.916299	-85.952386	Gas Station
5	Alexander City	32.93884	-85.95295	Anytime Fitness	717	32.945230	-85.953960	Gym / Fitness Center
6	Alexander City	32.93884	-85.95295	Dollar General	961	32.931278	-85.947983	Discount Store
7	Alexander City	32.93884	-85.95295	Wind Creek State Park	9541	32.856710	-85.923739	State / Provincial Park
8	Alexander City	32.93884	-85.95295	Jim Bob's	1090	32.929910	-85.948142	American Restaurant
9	Alexander City	32.93884	-85.95295	Subway	2193	32.919853	-85.959221	Sandwich Place
10	Alexander City	32.93884	-85.95295	Dairy Queen Grill & Chill	1726	32.933872	-85.970458	Ice Cream Shop

With this data, we can now check how many venues were returned for each neighbouring city of the selected state. (Alabama here) and examine how many unique categories can be curated from all the returned venues. (In this case 218). Then, we will sort the DataFrame and assign a venue number to each unique venue for plotting purposes & also to see which venue category is present how much in any particular state. We will also define a colour scheme for different venue categories to be plotted on the map. Following visual was obtained after that:

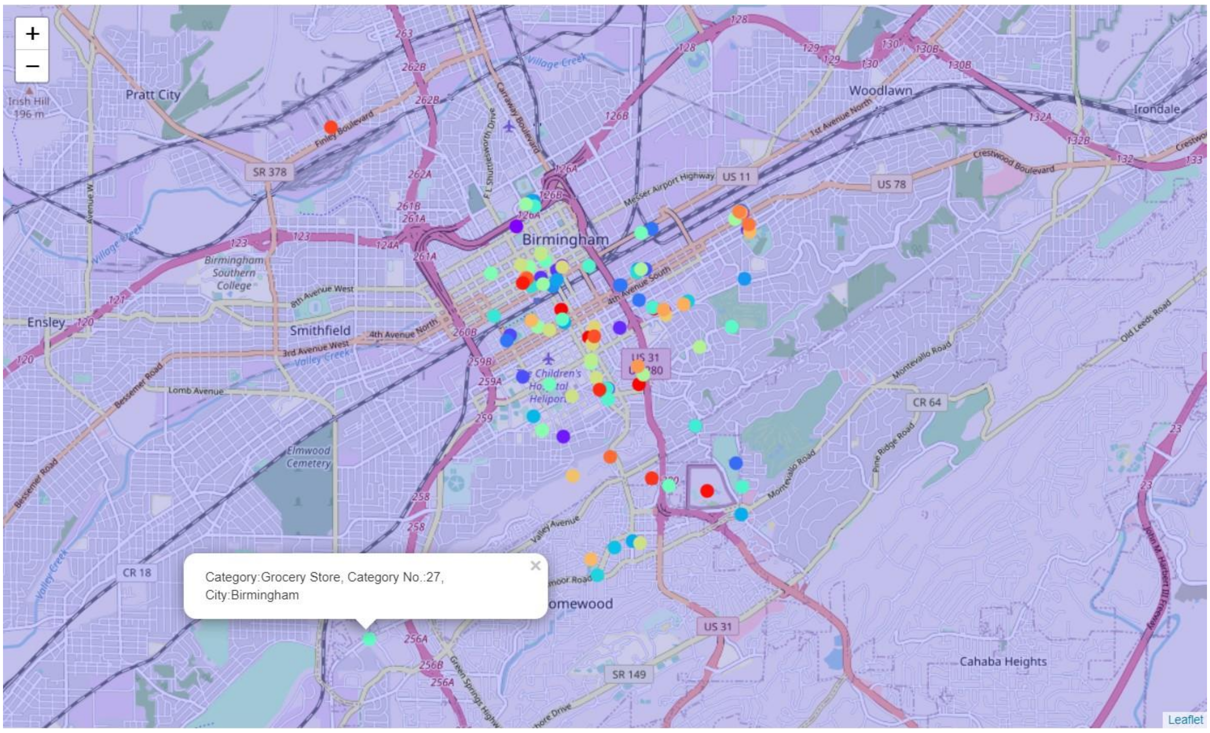

+
−
Roswell
Sandy Springs
Athens
Atlanta
Peachtree City
Talladega National Forest (Talladega Division)
Talladega National Forest (Oakmulgee Division)
Category:Airport, Category No.:1, City:Demopolis
Meridian
Macon
Warner Robins
Albany
Tifton
De Soto National Forest
Hattiesburg
Leaflet

Here we can see that each venue category has different number & colour. (You can distinguish the points in zoomed-in map)

We can also view this illustration for a separate city just by repeating the above methods for a specific city i.e., examining the unique venue categories, sorting and then forming a DataFrame with unique venue number. Here I took an example of 'Birmingham'. Following visual will be obtained:

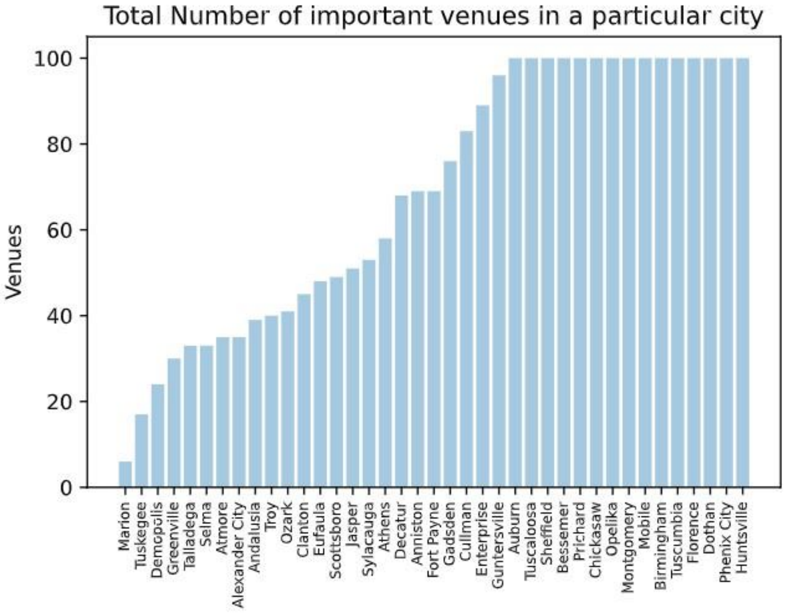


	Neighborhood	Latitude	Longitude	VenueName	Distance	VenueLatitude	VenueLongitude	VenueCategory	VenueNumber
66	Birmingham	33.52068	-86.81176	Newk's Express Cafe	1910	33.507419	-86.798701	American Restaurant	0
46	Birmingham	33.52068	-86.81176	Galley & Garden	2924	33.501173	-86.790663	American Restaurant	0
50	Birmingham	33.52068	-86.81176	Jack Brown's Beer & Burger Joint	2410	33.511413	-86.788280	American Restaurant	0
22	Birmingham	33.52068	-86.81176	Pies & Pints	1326	33.511170	-86.803148	American Restaurant	0
0	Birmingham	33.52068	-86.81176	Birmingham Museum Of Art	220	33.522311	-86.810409	Art Museum	1



Birmingham, Alabama

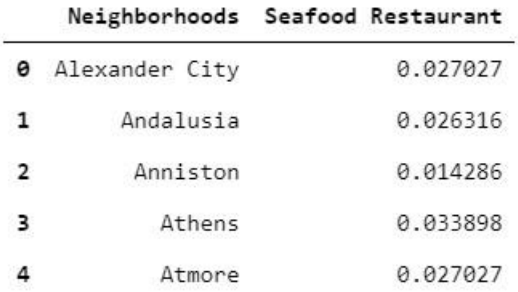
We can also check the number of venues returned for each city and plot them to see which city have more popular venues and by how much.



Here, we can see that how some popular cities reached the limit of venues(100). On the other hand; Marion, Tuskegee, Demopolis, Greenville, Talladega, Selma, Atmore, Alexander City, Andalusia are below 40 venues in our given coordinates with Latitude and Longitude.

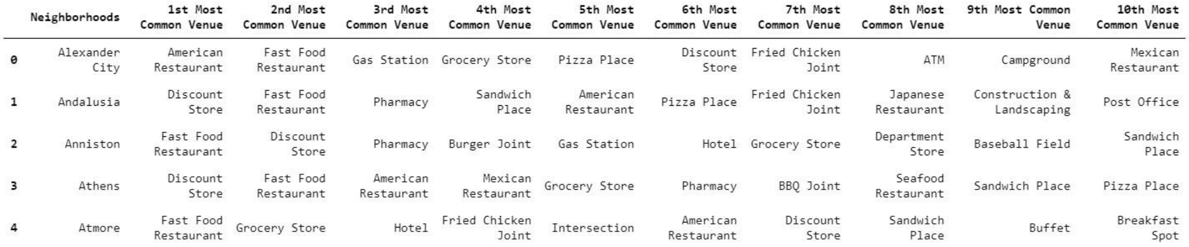
Analysis

Now, we will analyse each neighbouring city by grouping the rows by neighbouring city and taking the mean of the frequency of occurrence of each venue category. By doing so, we are also preparing the data for use in clustering. Also, here we can check percentage existence of any particular category of venue in all the cities of the selected state. For example, Seafood Restaurants.



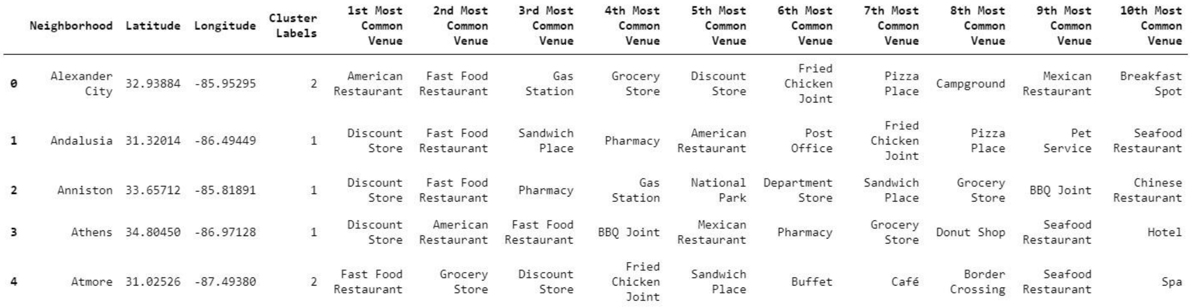
	Neighborhoods	Seafood Restaurant
0	Alexander City	0.027027
1	Andalusia	0.026316
2	Anniston	0.014286
3	Athens	0.033898
4	Atmore	0.027027

Now, lets create a table which shows list of top 10 venue categories for each neighbouring city.



	Neighborhoods	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Alexander City	American Restaurant	Fast Food Restaurant	Gas Station	Grocery Store	Pizza Place	Discount Store	Fried Chicken Joint	ATM	Campground	Mexican Restaurant
1	Andalusia	Discount Store	Fast Food Restaurant	Pharmacy	Sandwich Place	American Restaurant	Pizza Place	Fried Chicken Joint	Japanese Restaurant	Construction & Landscaping	Post Office
2	Anniston	Fast Food Restaurant	Discount Store	Pharmacy	Burger Joint	Gas Station	Hotel	Grocery Store	Department Store	Baseball Field	Sandwich Place
3	Athens	Discount Store	Fast Food Restaurant	American Restaurant	Mexican Restaurant	Grocery Store	Pharmacy	BBQ Joint	Seafood Restaurant	Sandwich Place	Pizza Place
4	Atmore	Fast Food Restaurant	Grocery Store	Hotel	Fried Chicken Joint	Intersection	American Restaurant	Discount Store	Sandwich Place	Buffet	Breakfast Spot

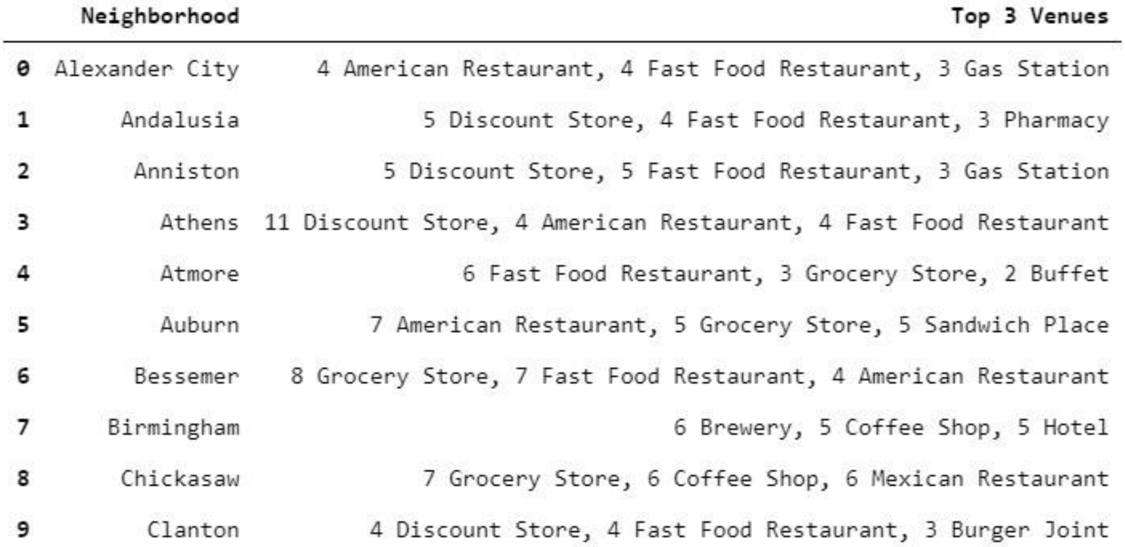
Lastly, we will perform clustering of the neighbouring cities with similar common venue categories by using **K-means algorithm**. It is one of the simplest and popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project. We will cluster the neighbouring cities into 5 clusters based on their similarities. The results will allow us to identify which city is similar to which city. Based on this, it will help us to answer the question “In any of the US states which city would be best to open which business?” according to the presence or absence of a certain category of a venue.



	Neighborhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Alexander City	32.93884	-85.95295	2	American Restaurant	Fast Food Restaurant	Gas Station	Grocery Store	Discount Store	Fried Chicken Joint	Pizza Place	Campground	Mexican Restaurant	Breakfast Spot
1	Andalusia	31.32014	-86.49449	1	Discount Store	Fast Food Restaurant	Sandwich Place	Pharmacy	American Restaurant	Post Office	Fried Chicken Joint	Pizza Place	Pet Service	Seafood Restaurant
2	Anniston	33.65712	-85.81891	1	Discount Store	Fast Food Restaurant	Pharmacy	Gas Station	National Park	Department Store	Sandwich Place	Grocery Store	BBQ Joint	Chinese Restaurant
3	Athens	34.80450	-86.97128	1	Discount Store	American Restaurant	Fast Food Restaurant	BBQ Joint	Mexican Restaurant	Pharmacy	Grocery Store	Donut Shop	Seafood Restaurant	Hotel
4	Atmore	31.02526	-87.49380	2	Fast Food Restaurant	Grocery Store	Discount Store	Fried Chicken Joint	Sandwich Place	Buffet	Café	Border Crossing	Seafood Restaurant	Spa

Here is my merged table with cluster labels for each neighbouring city.

One of my aim was also to show the number of top 3 venues information for each city for more deep study purposes like how many of a particular top venue is present in that particular city thus proving, that venue category is profitable there and is it too many to open more of that or can more of that business be opened there. Thus, I grouped each city by the number of top 3 venues and I combined those information in Top 3 venues column.

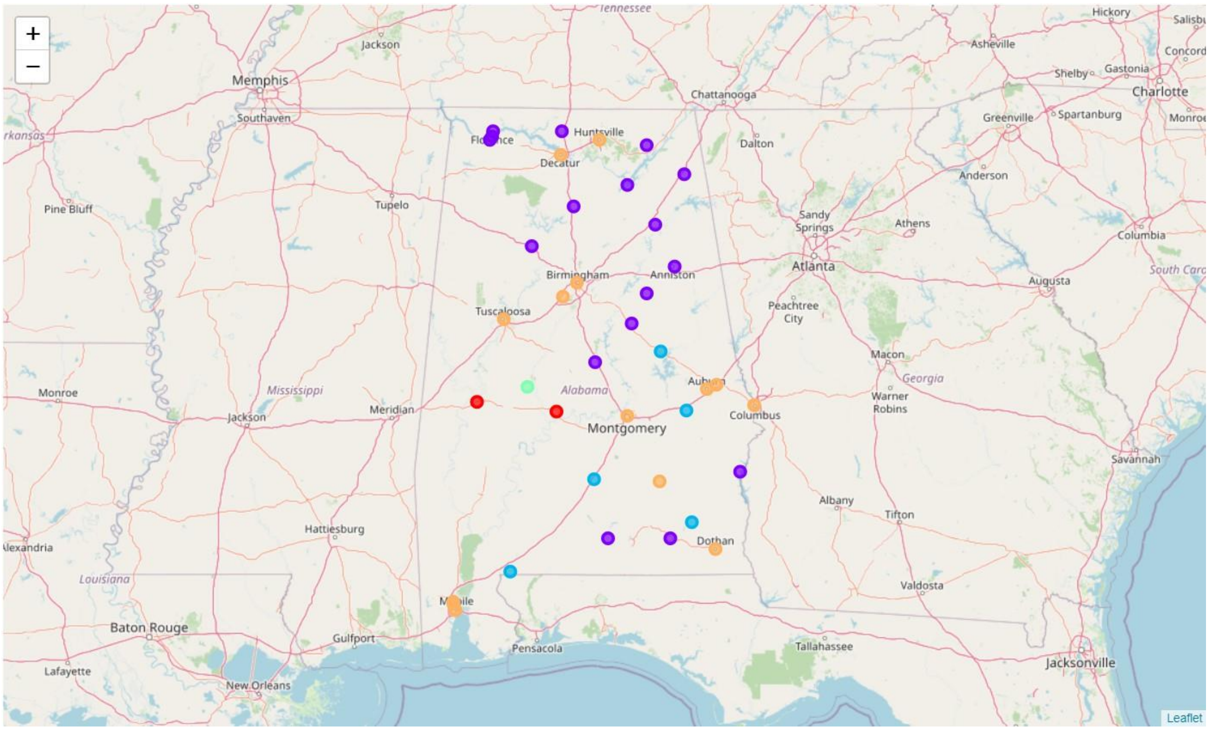


	Neighborhood	Top 3 Venues
0	Alexander City	4 American Restaurant, 4 Fast Food Restaurant, 3 Gas Station
1	Andalusia	5 Discount Store, 4 Fast Food Restaurant, 3 Pharmacy
2	Anniston	5 Discount Store, 5 Fast Food Restaurant, 3 Gas Station
3	Athens	11 Discount Store, 4 American Restaurant, 4 Fast Food Restaurant
4	Atmore	6 Fast Food Restaurant, 3 Grocery Store, 2 Buffet
5	Auburn	7 American Restaurant, 5 Grocery Store, 5 Sandwich Place
6	Bessemer	8 Grocery Store, 7 Fast Food Restaurant, 4 American Restaurant
7	Birmingham	6 Brewery, 5 Coffee Shop, 5 Hotel
8	Chickasaw	7 Grocery Store, 6 Coffee Shop, 6 Mexican Restaurant
9	Clanton	4 Discount Store, 4 Fast Food Restaurant, 3 Burger Joint

Cities and their number of Top 3 Venues

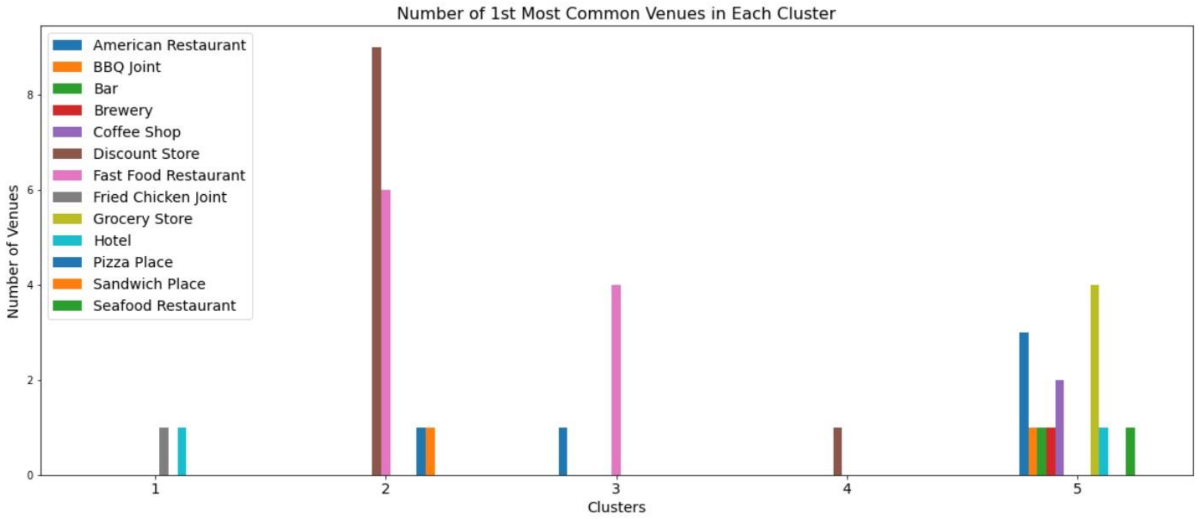
Results

The results from the k-means clustering shows that we can categorize the cities into 5 clusters having similar venue categories. The results of the clustering are visualized in the map below with cluster 1 in red colour, cluster 2 in purple colour, cluster 3 in light blue colour, cluster 4 in mint green colour and cluster 5 in yellow colour.

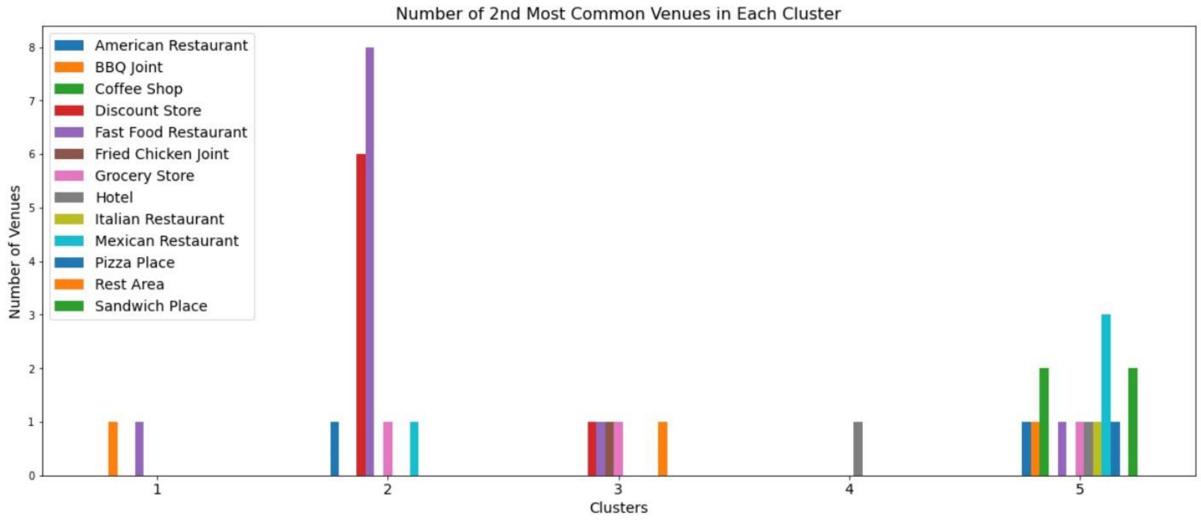


K-Means Clustering of similar cities

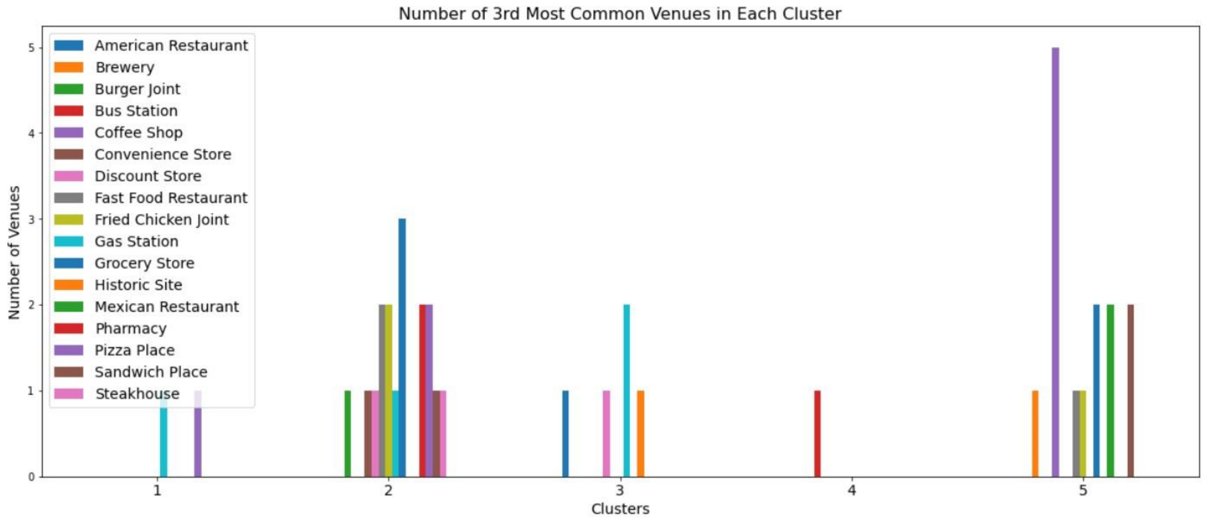
We can also estimate the number of 1st Most Common Venue, 2nd Most Common Venue & 3rd Most Common Venue in each cluster. Thus, we can create a bar chart which may help us to find which venue is concentrated in which cluster.



Graph 1



Graph 2



Graph 3

Discussion

As we get our observations from the map in Results section, most of the neighbouring cities are concentrated into cluster 2 & 5. On the other hand, cluster 1 & 3 have only two cities each whereas cluster 4 is quite unique with only 1 city. We can also observe here from the graphs that cluster 5 is missing or have very low ‘Discount Store’ & ‘Fast Food Restaurant’. This represents a great opportunity to open or invest in these categories of venues in cluster 5 cities or vice-versa we can observe that cluster 2 is missing on or have low number of ‘Coffee Shops’ as compared from cluster 5 thus creating also a great opportunity to open or invest in that category. From another perspective, the results also shows that the oversupply of ‘Discount Store’ & ‘Fast Food Restaurant’ are present in the cluster 2 cities. Property developers with unique selling propositions will stand out moreover they will face no or less competition if they go forward with the above-mentioned opportunity. Therefore, this project recommends property developers to capitalize on these findings to open or invest in any business category in a particular city or area. Moreover, property developers or investors should be wary if they want to open ‘Discount Store’ in cluster 1 cities. This can also benefit people wanting to move to a new place and if they have a particular need for a venue like say ‘Discount Store’ for an economic friendly stay, then they can move to cluster 2 cities. Also, we can compare all the cities of the state as how they are similar or dissimilar. Thus, the battle between them.

Conclusion

In this study, I analysed the similarities & dissimilarities between the cities of a particular state. I identified the different categories of venues and sorted them. Then by means of graphical representation & k-means clustering, I categorized different cities into cluster groups. The model will help many stakeholders & business people to predict the profit for opening a particular business in a particular area or also, if they should open that particular business or not. People are turning to big cities to start a business or work. They can achieve better outcomes through their access to the platforms where such information is provided. And, not only for investors but also city managers can manage the city more regularly by using similar data analysis types or platforms.