Clustering of Neighborhoods in Florida, Ohio, and Texas

A. Introduction

A.1. Background

Real estate isn't what it used to be. Advancement in technology has brought about a growing digital competition between real estate agencies. Previously, real estate agents had customers hanging on their every word; after all, buying and selling a home was complicated. But today, real estate has become much more transparent, and consumers have access to ample information online. Market comparisons, crime statistics, school ratings, and much more information can be gotten online. To capture the attention of customers in this new digital space, successful real estate agents need to make their listings come alive online and embrace technology and new means of marketing themselves, including social media, blogging, search engine optimization, and more.

Although technology is making the real estate industry more competitive, it is also helping real estate agents by giving them faster and more efficient means of taking care of business so they can focus on providing unmatched customer service to their clients.

A.2. Problem Description

A Real Estate Agency is looking to improve its customer service experience by finding and suggesting neighborhoods that are similar to the customer's previous neighborhood as most customers usually prefer neighborhoods that have similarities with their previous homes for better adaptation. The basis for similarity measurement between neighborhoods for this problem is the type of venues (e.g. Cafe Shops, Supermarkets, Restaurants, Pharmacies, Gas Stations, etc.) within the neighborhood. The Agency has its operation base in Florida, Ohio, and Texas and wants to find all similar neighborhoods in the three states they operate.

A.3. Problem Statement

Grouping neighborhoods in Florida, Ohia, and Texas into clusters based on the similarity of venues within the neighborhoods

A.4. Data Description

For the Florida, Ohio and Texas neighborhood data, I web scrapped the information of the boroughs and respective neighborhoods for the three states from Wikipedia pages.

- Florida: https://en.wikipedia.org/wiki/List_of_municipalities_in_Florida
- Ohio: https://en.wikipedia.org/wiki/List_of_cities_in_Ohio
- Texas: https://en.wikipedia.org/wiki/List_of_cities_in_Texas

from the tables, I extracted only the borough and the neighborhood columns and then grouped the neighborhoods by their respective boroughs.

	Borough	Neighborhood
0	Allen	Delphos, Lima
1	Ashland	Ashland
2	Ashtabula	Ashtabula, Conneaut, Geneva
3	Athens	Athens, Nelsonville
4	Auglaize	Saint Marys, Wapakoneta
5	Belmont	Martins Ferry, Saint Clairsville
6	Butler	Fairfield, Hamilton, Middletown, Monroe, Oxfor...
7	Champaign	Urbana
8	Clark	New Carlisle, Springfield
9	Clermont	Loveland, Milford

B. Methodology

B.1 Tools and Libraries

In this project, the python modules used are;

Pandas: This is like the Microsoft Excel of python. it is a fast, powerful, flexible and easy to use tool for data analysis and manipulation.

Folium: Excellent library for visualizing geospatial data *ie. data that contain geographical (latitude, longitude, altitude) as part of its features.*

Geocoder: A simple and consistent geocoding library for getting geographical information of addresses. Works well with different geocoding providers like Google, Bing, ArcGis & many more.

Matplotlib: A comprehensive library for creating static, animated, and interactive visualizations in python.

Requests: Elegant and simple Http library for Python. The requests module allows you to send Http requests using Python.

Sklearn: Free machine learning library for Python. It features various algorithms like support vector machine, random forests, and k-neighbors, and it also supports python numerical and scientific libraries like NumPy and SciPy.

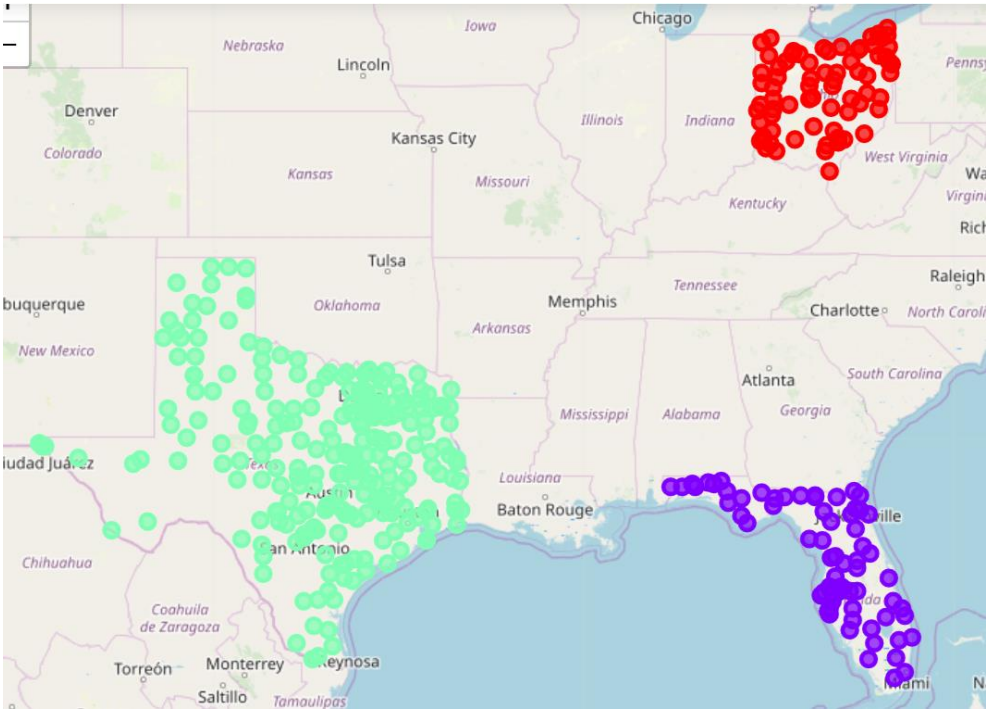
B.2 Data Preparation

To visualize the different neighborhoods and to utilize the Foursquare API to get the location data of the neighborhoods and explore its venues, I had to get the latitude and longitude of the

	Borough	Neighborhood	Latitude	Longitude
0	Allen	Delphos, Lima	40.771510	-84.105802
1	Ashland	Ashland	40.868650	-82.315500
2	Ashtabula	Ashtabula, Conneaut, Geneva	41.889260	-80.786730
3	Athens	Athens, Nelsonville	39.328500	-82.104440
4	Auglaize	Saint Marys, Wapakoneta	40.560902	-84.221740
5	Belmont	Martins Ferry, Saint Clairsville	40.029550	-81.038910
6	Butler	Fairfield, Hamilton, Middletown, Monroe, Oxfor...	40.587860	-82.423760
7	Champaign	Urbana	40.137620	-83.769511
8	Clark	New Carlisle, Springfield	40.013100	-84.724120
9	Clermont	Loveland, Milford	39.047998	-84.151803

different neighborhoods using any appropriate geocoding library (I used the geocoder library in this project with the ArcGIS API). The neighborhoods were geocoded by the geocoder and their respective latitudes and longitudes were extracted and added to the pandas dataframe.

Now that the location coordinates have been added, visualization of geographic details of the three states and their boroughs was possible using folium and circle markers were superimposed on the map using the latitude and longitude values of the different boroughs.



To explore the different boroughs and extract venues within them, I used the requests library to send Http requests to the Foursquare API.

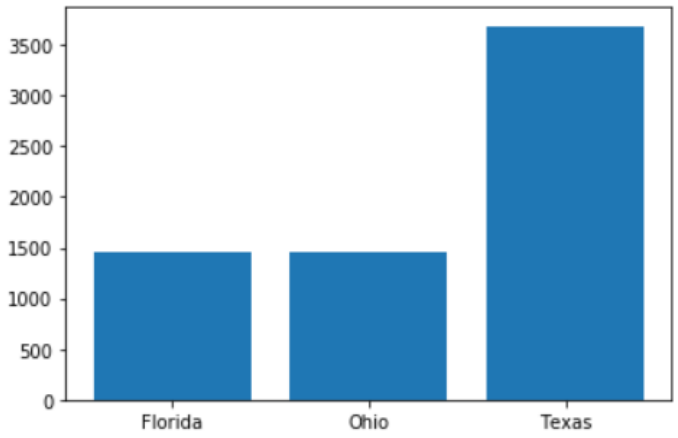
	Borough	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Alachua	Alachua, Archer, Gainesville, Hawthorne, High ...	29.79309	-82.49428	Conestogas Restaurant	29.792109	-82.495888	American Restaurant
1	Alachua	Alachua, Archer, Gainesville, Hawthorne, High ...	29.79309	-82.49428	New York Pizza Plus	29.795401	-82.502413	Pizza Place
2	Alachua	Alachua, Archer, Gainesville, Hawthorne, High ...	29.79309	-82.49428	Mi Apá Latin Café	29.798116	-82.502088	Latin American Restaurant
3	Alachua	Alachua, Archer, Gainesville, Hawthorne, High ...	29.79309	-82.49428	Walgreens	29.794937	-82.495102	Pharmacy
4	Alachua	Alachua, Archer, Gainesville, Hawthorne, High ...	29.79309	-82.49428	Bev's Better Burgers	29.792840	-82.495964	American Restaurant

For every venue in a borough, the venue name, category, latitude, and longitude was pulled from the Foursquare API.

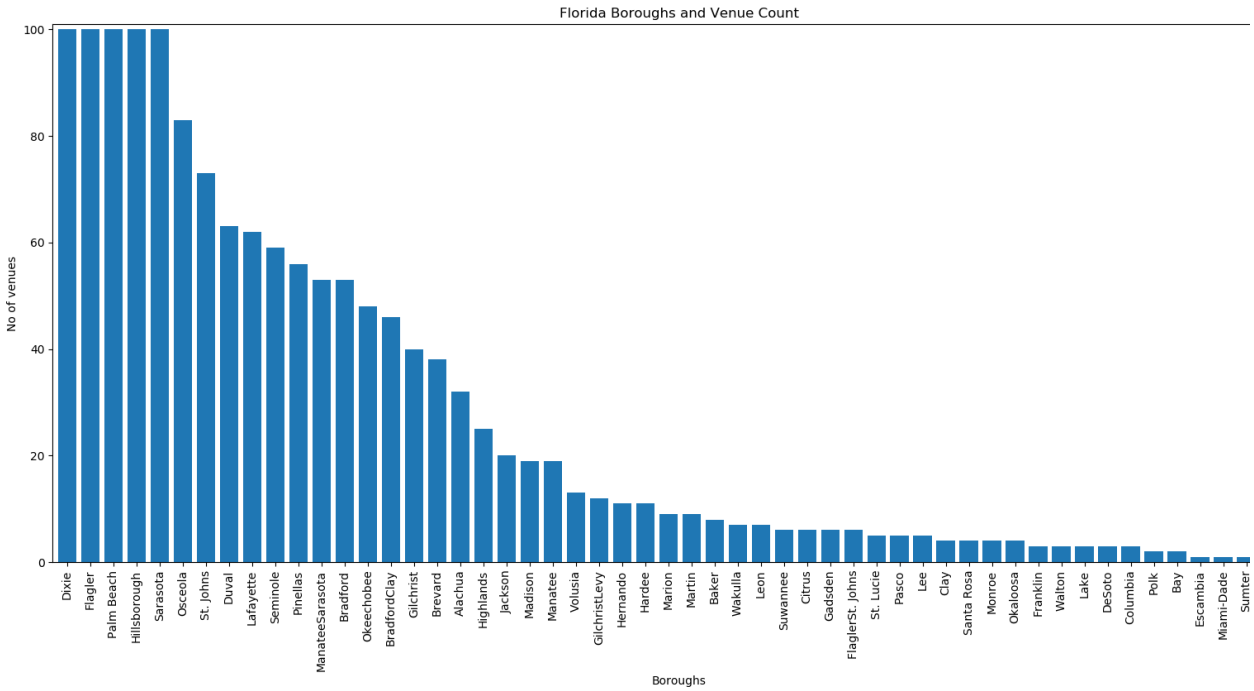
B.3 Exploratory Data Analysis

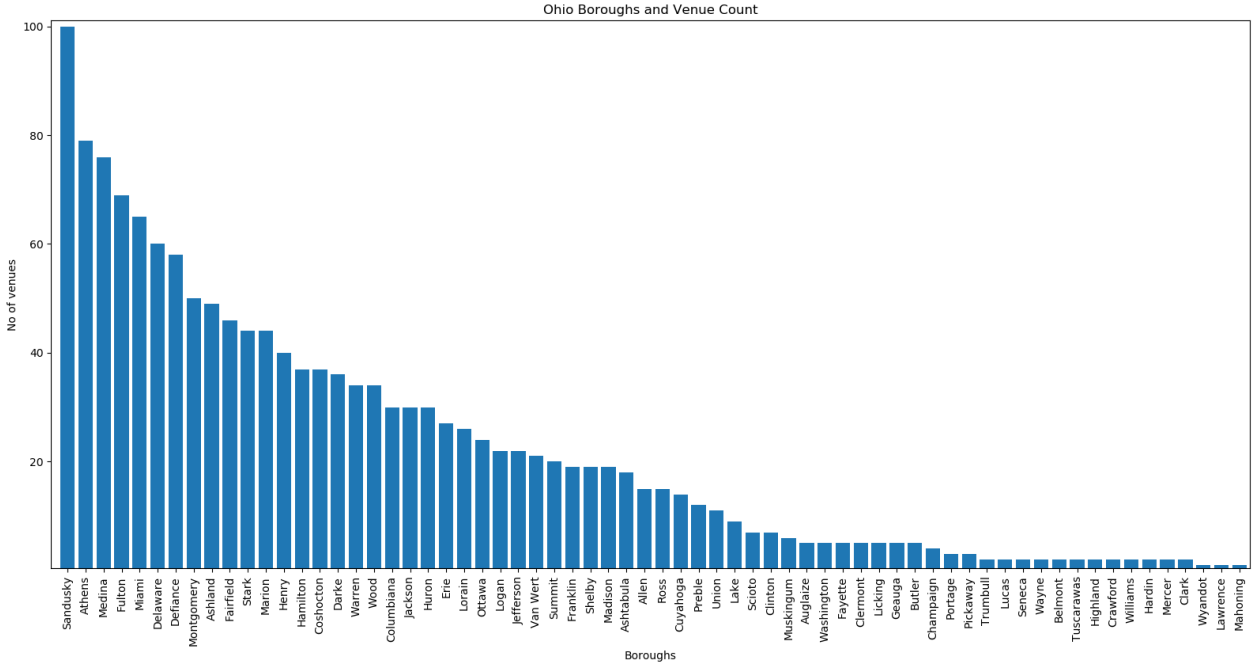
The total number of venues returned by the Foursquare API was 6595 with Texas having the greatest number of venues (3687).

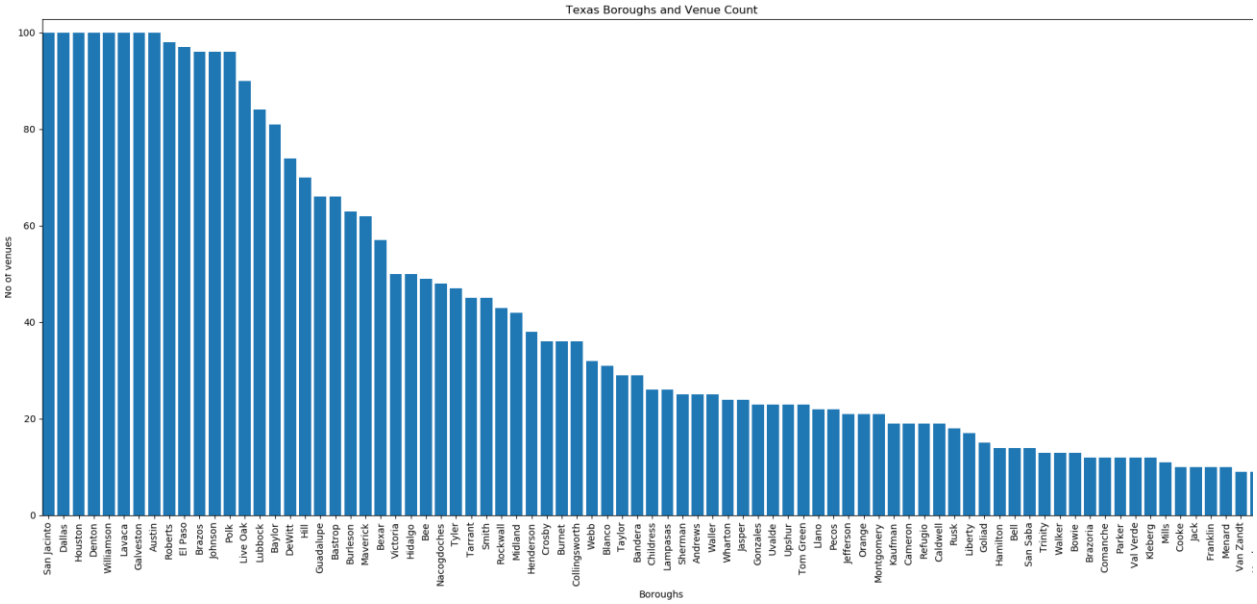
	State	No of Venues
0	Florida	1455
1	Ohio	1453
2	Texas	3687



Within each state, certain boroughs were noticed to have a high number of venues.




Ohio Boroughs and Venue Count
No of venues
Boroughs
Sandusky
Athens
Medina
Fulton
Miami
Delaware
Defiance
Montgomery
Ashland
Fairfield
Stark
Marion
Henry
Hamilton
Coshocton
Darke
Warren
Wood
Columbiana
Jackson
Huron
Erie
Lorain
Ottawa
Logan
Jefferson
Van Wert
Summit
Franklin
Shelby
Madison
Ashtabula
Allen
Ross
Cuyahoga
Preble
Union
Lake
Scioto
Clinton
Muskingum
Auglaize
Washington
Fayette
Clermont
Licking
Geauga
Butler
Champaign
Portage
Pickaway
Trumbull
Lucas
Seneca
Wayne
Belmont
Tuscarawas
Highland
Crawford
Williams
Hardin
Mercer
Clark
Wyandot
Lawrence
Mahoning


Texas Boroughs and Venue Count
No of venues
Boroughs
San Jacinto
Dallas
Houston
Denton
Williamson
Lavaca
Galveston
Austin
Roberts
El Paso
Brazos
Johnson
Polk
Live Oak
Lubbock
Baylor
DeWitt
Hill
Guadalupe
Bastrop
Burleson
Maverick
Bexar
Victoria
Hidalgo
Bee
Nacogdoches
Tyler
Tarrant
Smith
Rockwall
Midland
Henderson
Crosby
Burnet
Collingsworth
Webb
Blanco
Taylor
Bandera
Childress
Lampasas
Sherman
Andrews
Waller
Wharton
Jasper
Gonzales
Uvalde
Upshur
Tom Green
Llano
Pecos
Jefferson
Orange
Montgomery
Kaufman
Cameron
Refugio
Caldwell
Rusk
Liberty
Goliad
Hamilton
Bell
San Saba
Trinity
Walker
Bowie
Brazoria
Comanche
Parker
Val Verde
Kleberg
Mills
Cooke
Jack
Franklin
Menard
Van Zandt

I created a table to show the ten most common venue categories for each borough of the three different states.

	Borough	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	Alachua	Pizza Place	Mexican Restaurant	Discount Store	American Restaurant	Pharmacy	Fast Food Restaurant	BBQ Joint	Park	Clothing Store	Sandwich Place
1	Baker	Taco Place	Farmers Market	Garden Center	Park	Intersection	American Restaurant	Grocery Store	Farm	Food & Drink Shop	Food Court
2	Bay	Gift Shop	Lake	Zoo Exhibit	Fondue Restaurant	Frozen Yogurt Shop	Fried Chicken Joint	French Restaurant	Fountain	Forest	Football Stadium
3	Bradford	Pharmacy	Pizza Place	Intersection	Restaurant	Golf Course	Liquor Store	Bakery	Discount Store	Bagel Shop	Ice Cream Shop
4	BradfordClay	Fast Food Restaurant	Breakfast Spot	Pizza Place	Gym	Grocery Store	Sandwich Place	Accessories Store	Pharmacy	Big Box Store	Shipping Store

B.4 Clustering

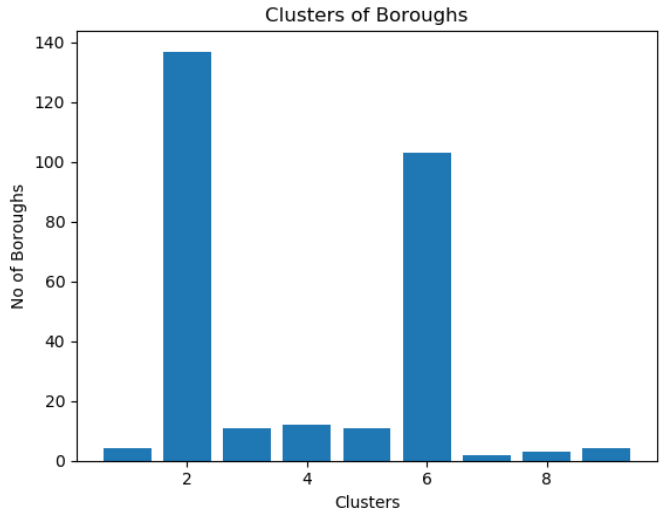
The clustering algorithm used was KMeans Clustering which is one of the most common unsupervised learning techniques. It is an iterative algorithm that tries to partition the dataset into *k* pre-defined distinct non-overlapping subgroups (clusters). Before applying the KMeans algorithm, the various venue categories in the venues dataframe were one hot encoded to convert categorical variables into numerical values suitable for the clustering algorithm.

	Borough	Neighborhoods	Latitude	Longitude	Accessories Store	Advertising Agency	Airport	Airport Service	Airport Terminal	American Restaurant	...	Library	Lighthouse	Mountain	M Sc
0	Alachua	Alachua, Archer, Gainesville, Hawthorne, High ...	29.793090	-82.494280	0.000000	0.03125	0.0	0.0	0.0	0.0625	...	0.0	0.0	0.0	
1	Baker	Glen St. Mary, Macclenny	30.797290	-86.682380	0.000000	0.00000	0.0	0.0	0.0	0.1250	...	0.0	0.0	0.0	
2	Bay	Callaway, Lynn Haven, Mexico Beach, Panama Cit...	30.278741	-85.615480	0.000000	0.00000	0.0	0.0	0.0	0.0000	...	0.0	0.0	0.0	
3	Bradford	Brooker, Hampton, Lawtey Nour Town, Starke	28.068500	-82.526900	0.000000	0.00000	0.0	0.0	0.0	0.0000	...	0.0	0.0	0.0	
4	BradfordClay	Keystone Heights	30.131051	-81.759014	0.021739	0.00000	0.0	0.0	0.0	0.0000	...	0.0	0.0	0.0	

After encoding the categorical variables, the KMeans algorithm was applied to cluster the boroughs into 10 clusters and the cluster labels for the different boroughs wad added to the dataframe.

	Borough	Neighborhoods	Latitude	Longitude	Cluster Label	Accessories Store	Advertising Agency	Airport	Airport Service
0	Alachua	Alachua, Archer, Gainesville, Hawthorne, High ...	29.793090	-82.494280	5	0.000000	0.03125	0.0	0.0
1	Baker	Glen St. Mary, Macclenny	30.797290	-86.682380	1	0.000000	0.00000	0.0	0.0
2	Bay	Callaway, Lynn Haven, Mexico Beach, Panama Cit...	30.278741	-85.615480	1	0.000000	0.00000	0.0	0.0
3	Bradford	Brooker, Hampton, Lawtey Nour Town, Starke	28.068500	-82.526900	1	0.000000	0.00000	0.0	0.0
4	BradfordClay	Keystone Heights	30.131051	-81.759014	5	0.021739	0.00000	0.0	0.0
5	Brevard	Cape Canaveral, Cocoa, Cocoa Beach, Grant-Valk...	28.293484	-80.732356	5	0.000000	0.00000	0.0	0.0
6	Citrus	Crystal River, Inverness	28.850779	-82.465949	3	0.000000	0.00000	0.0	0.0
7	Clay	Green Cove Springs, Orange Park, Penney Farms	29.983055	-81.857782	1	0.000000	0.00000	0.0	0.0
8	Columbia	Fort White, Lake City	30.069510	-82.697490	1	0.000000	0.00000	0.0	0.0
9	DeSoto	Arcadia	27.186258	-81.809501	2	0.000000	0.00000	0.0	0.0

C. Results

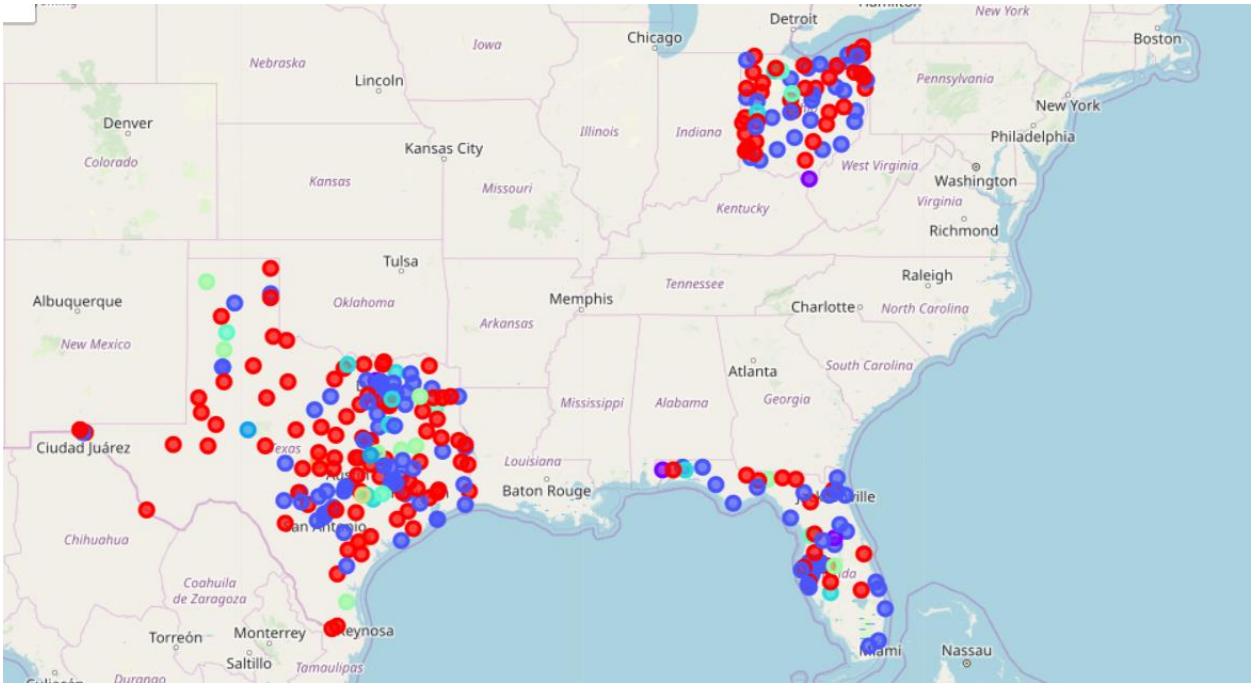


A bar chart of the different clusters and the number of boroughs within the clusters is shown above and it shows that cluster 2 and cluster 6 have the highest number of boroughs. To get a better understanding of the different clusters' venue characteristics, a dataframe was created to show the most common venue categories in the different clusters.

Cluster Label	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
1	Discount Store	Bar	Food	Sandwich Place	American Restaurant	Mexican Restaurant	Grocery Store	Burger Joint	Waterfront	Ice Cream Shop
2	American Restaurant	Convenience Store	Pizza Place	Bar	Gas Station	Campground	Mexican Restaurant	Sandwich Place	Hotel	History Museum
3	Farm	Construction & Landscaping	Stables	Massage Studio	Athletics & Sports	Sandwich Place	Furniture / Home Store	Forest	Convenience Store	Trail
4	Construction & Landscaping	Furniture / Home Store	Trail	Convenience Store	Steakhouse	Bar	Antique Shop	Food	Paper / Office Supplies Store	Restaurant
5	Park	Home Service	Mountain	Harbor / Marina	State / Provincial Park	Airport Service	River	Golf Course	Seafood Restaurant	Trail
6	Fast Food Restaurant	Discount Store	Mexican Restaurant	Sandwich Place	Pizza Place	Convenience Store	American Restaurant	Grocery Store	BBQ Joint	Gas Station
7	Ice Cream Shop	Waterfront	Hospital	Hot Spring	Hotel	IT Services	Indian Restaurant	Indie Movie Theater	Insurance Office	Intersection
8	Food	Furniture / Home Store	Waterfront	Irish Pub	Hot Spring	Hotel	IT Services	Ice Cream Shop	Indian Restaurant	Indie Movie Theater
9	Home Service	Cuban Restaurant	Paper / Office Supplies Store	Outdoor Supply Store	Plaza	Miscellaneous Shop	Insurance Office	Hot Dog Joint	Hot Spring	Hotel

At first look at the table above, it can be deduced that boroughs in clusters 3 and 4 are rural as the most common venues are farms, landscapes, stables, forests and trails which are very common for countryside regions. It can also be presumed that boroughs in clusters 7, 8 and 9 are urban.

Finally, a map was created to show the different clusters created and their boroughs.

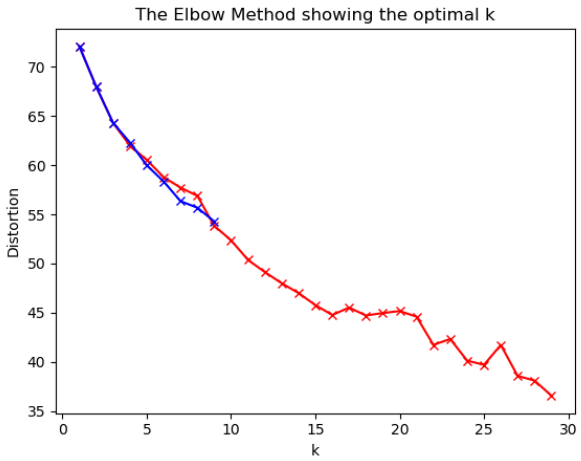


D. Discussion

KMeans algorithm was used in this clustering project because it is relatively simple to implement, scales to large data sets guarantees convergence, adapts easily to new examples and can generalize to clusters of different shapes and sizes, such as elliptical clusters.

KMeans also has its disadvantages as the value of *k* has to be chosen manually, the result is highly dependent on initial values, scaling with a large number of dimensions and the inability to ignore outliers.

The major problem faced in this project was having to choose the value of *k* manually. Although it can be solved by using the Elbow method which is a plot of the Loss vs the No of Clusters to find the optimal (*k*), the elbow method does not always give a clear optimal value of *k*.



from the plot above, it can be seen that there is no clear elbow showing the optimal value of *k*. The blue line gives a plot of *k* ranging from 1 to 10 and the red line gives a plot of *k* ranging from 1 to 30. it can be observed that there is no clear elbow point and thus *k* was randomly set to 10. The inability of the elbow method to give a clear optimal value of k is due to the problem of scaling in large dimensions. Here, we have a dataset with 360 different venue categories which means 360 features or dimensions and as the number of dimensions increases, a distance-based similarity measure converges to a constant value between any given examples.

Other clustering algorithms can be used in further study and the efficiency can be compared with that of the KMeans algorithm currently used in this project. Moreover, it is obvious that not every clustering method can yield high-quality results for this sort of problem.

E. Conclusion

In this study, I clustered neighborhoods within Florida, Ohio, and Texas based on the similarity of venues in the different neighborhoods, I extracted the location data for the boroughs and explored the venues within the different boroughs. I performed Exploratory Data Analysis and built a clustering model to cluster the neighborhoods. This model is very useful in helping real estate agents to suggest suitable locals for their clients and customers can easily identify neighborhoods of preference based on similarity to a familiar neighborhood.

F. References

- [1] Wikipedia
- [2] Foursquare API

Thanks for reading.

Feel free to give comments and criticize.

Jubilee Imhanzenobe.