Venue Comparison in the Dallas-Fort Worth Metroplex

IBM Applied Data Science Capstone

Leonardo Torquato

Introduction

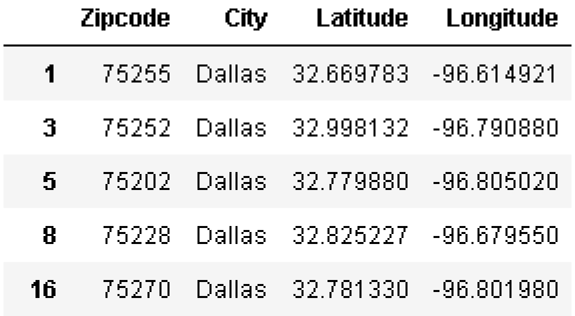
The Dallas-Fort Worth Metroplex is one of the largest metropolitan areas and one of the fastest-growing in the United States. It is largely composed of the two major cities, Dallas and Fort Worth, among other smaller cities in the surrounding region. Although being adjacent, the two cities share distinct cultures and stereotypes. Dallas for being an emerging tech region and Fort Worth being known as "where the West begins". Hosting a wide range of industries various emerging businesses may want to expand locally. Being key to their development would be selecting a location where it would be successful. An indication can be given by observing the location of present businesses and studying how they are positioned among other venues so that it can be compared to its success. Therefore, it is of interest to discover areas with similarities in its range of venues within these two cities.

Data

For this analysis, the data was retrieved from two sources. The first set of data was retrieved from Opendatasoft's available data on Dallas and Fort Worth ZIP Codes regions and associated coordinates as a CSV file. The second set of data was retrieved through Foursquare API with a search function for "venues" in Dallas and Fort Worth in the ZIP Code regions. These two elements of data allowed for segmentation, clustering, and visualization of similar regions in the two cities.

The data from the CSVs for both Dallas and Fort Worth were cleaned to contain only the required data for analysis. Columns including 'state', 'timezone', 'daylight saving time flag', and 'geopoint' were all dropped from the data frame. The columns entitled 'zip' were renamed to 'ZIP Code'. In the data frames were multiple instances of the same coordinates for multiple different ZIP codes. Those ZIP Codes were dropped from the data frame. A result of 'Lake Dallas' which is not actually in Dallas or Fort Worth was found in the data and dropped.

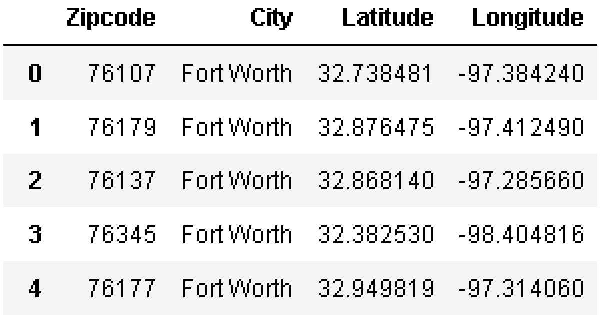
Figure 1 shows a segment of the cleaned data frame for the Dallas data. The final result included 52 ZIP Code regions.



	Zipcode	City	Latitude	Longitude
1	75255	Dallas	32.669783	-96.614921
3	75252	Dallas	32.998132	-96.790880
5	75202	Dallas	32.779880	-96.805020
8	75228	Dallas	32.825227	-96.679550
16	75270	Dallas	32.781330	-96.801980

Figure 1: Dallas Data Frame Sample

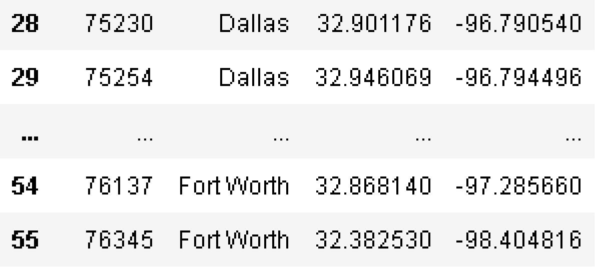
Figure 2 shows a segment of the cleaned data frame for the Fort Worth data. The final result included 32 ZIP Code regions.



	Zipcode	City	Latitude	Longitude
0	76107	Fort Worth	32.738481	-97.384240
1	76179	Fort Worth	32.876475	-97.412490
2	76137	Fort Worth	32.868140	-97.285660
3	76345	Fort Worth	32.382530	-98.404816
4	76177	Fort Worth	32.949819	-97.314060

Figure 2: Fort Worth Data Frame Sample

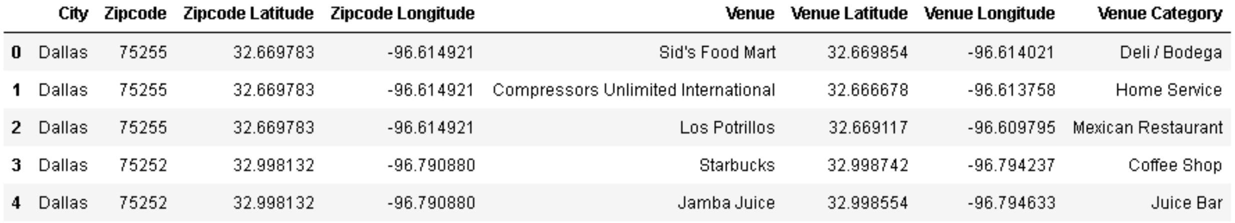
Next, the two previous data frames were concatenated into one data frame with a final total of 84 ZIP Code regions.



28	75230	Dallas	32.901176	-96.790540
29	75254	Dallas	32.946069	-96.794496
...	...	...	...	...
54	76137	Fort Worth	32.868140	-97.285660
55	76345	Fort Worth	32.382530	-98.404816

Figure 3: Concatenated Data Frame Sample

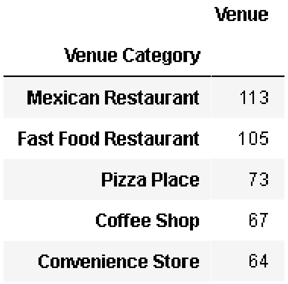
The next set of data was retrieved through a Foursquare API explore function in JSON format. A limit of 150 venues was set which would end up being more than available. A radius of 1000m was set which would provide a good range around each coordinate will keeping each area distinct from each other. The data given included the venue, the venue coordinates, and the venue category. Figure 4 shows a sample of the resultant data frame with venue information within those ZIP Codes. There was a total of 2383 venues found within all ZIP Code regions with 269 unique categories.



	City	Zipcode	Zipcode Latitude	Zipcode Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Dallas	75255	32.669783	-96.614921	Sid's Food Mart	32.669854	-96.614021	Deli / Bodega
1	Dallas	75255	32.669783	-96.614921	Compressors Unlimited International	32.666678	-96.613758	Home Service
2	Dallas	75255	32.669783	-96.614921	Los Potrillos	32.669117	-96.609795	Mexican Restaurant
3	Dallas	75252	32.998132	-96.790880	Starbucks	32.998742	-96.794237	Coffee Shop
4	Dallas	75252	32.998132	-96.790880	Jamba Juice	32.998554	-96.794633	Juice Bar

Figure 4: DFW Venues Sample

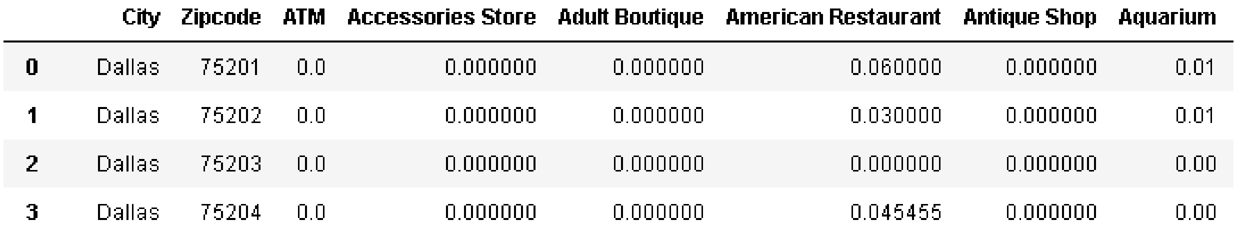
Next, the data set was sorted by Venue Category for supplementary data on the count of each venue type as seen in Figure 5.



	Venue
Venue Category	
Mexican Restaurant	113
Fast Food Restaurant	105
Pizza Place	73
Coffee Shop	67
Convenience Store	64

Figure 5: Venue Count Sample

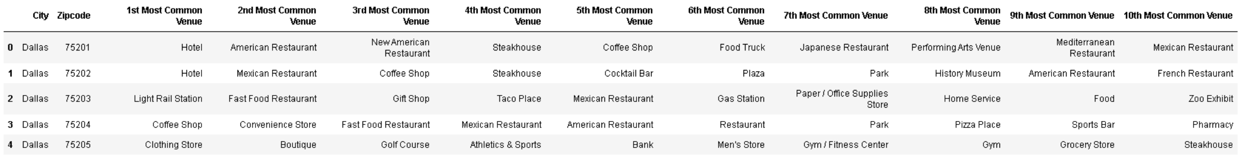
Further, one hot encoding was applied to the data which transferred the categorical features into numerical values allowing for better communication with the computer once the clustering algorithm was applied. Continuing, the rows were grouped by ZIP code and the venue categories are given a frequency of occurrence value as displayed in Figure 6.



	City	Zipcode	ATM	Accessories Store	Adult Boutique	American Restaurant	Antique Shop	Aquarium
0	Dallas	75201	0.0	0.000000	0.000000	0.060000	0.000000	0.01
1	Dallas	75202	0.0	0.000000	0.000000	0.030000	0.000000	0.01
2	Dallas	75203	0.0	0.000000	0.000000	0.000000	0.000000	0.00
3	Dallas	75204	0.0	0.000000	0.000000	0.045455	0.000000	0.00

Figure 6: Frequency of Venue Occurrence Sample

Figure 7 shows a final data frame where the top 10 most common venues for each ZIP Code region is presented.



	City	Zipcode	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	Dallas	75201	Hotel	American Restaurant	New American Restaurant	Steakhouse	Coffee Shop	Food Truck	Japanese Restaurant	Performing Arts Venue	Mediterranean Restaurant	Mexican Restaurant
1	Dallas	75202	Hotel	Mexican Restaurant	Coffee Shop	Steakhouse	Cocktail Bar	Plaza	Park	History Museum	American Restaurant	French Restaurant
2	Dallas	75203	Light Rail Station	Fast Food Restaurant	Gift Shop	Taco Place	Mexican Restaurant	Gas Station	Paper / Office Supplies Store	Home Service	Food	Zoo Exhibit
3	Dallas	75204	Coffee Shop	Convenience Store	Fast Food Restaurant	Mexican Restaurant	American Restaurant	Restaurant	Park	Pizza Place	Sports Bar	Pharmacy
4	Dallas	75205	Clothing Store	Boutique	Golf Course	Athletics & Sports	Bank	Men's Store	Gym / Fitness Center	Gym	Grocery Store	Steakhouse

Figure 7: Top Common Venues Sample

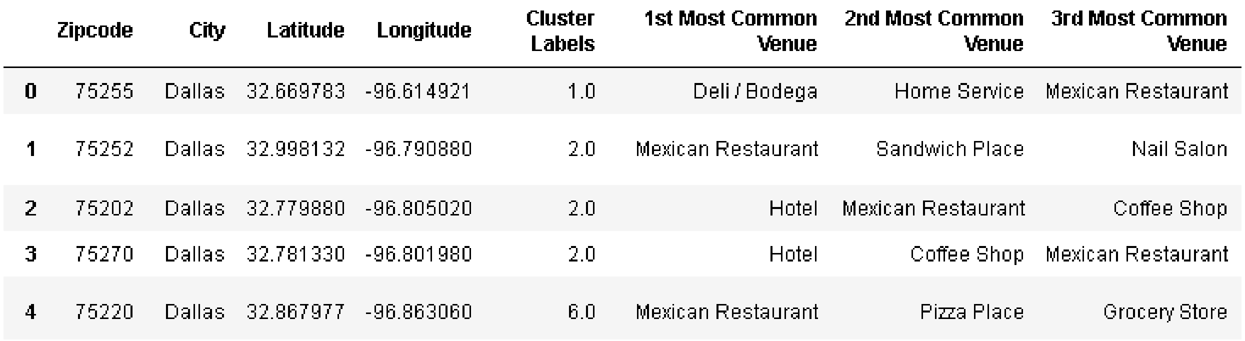
Through those steps, the data was prepared for the k-means clustering algorithm.

Methodology

Once the data was acquired, cleaned, and prepared it was then analyzed using a k-means clustering algorithm. k-means clustering is an unsupervised learning method that uses an iterative algorithm to partition the data into distinct clusters with similar traits.

For this study, 8 cluster groups were determined to be appropriate. Meaning that there would be 8 distinct cluster groups found out of the 84 ZIP Code regions. Before running the k-means clustering, the city and ZIP Code columns were identified to be dropped from the analysis so that the results could reflect on purely the common venue data.

Once the k-mean algorithm finished the cluster labels were then introduced to the data frame and the city, ZIP Code, and coordinates rejoined producing a table with all values as seen in Figure 8.



	Zipcode	City	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
0	75255	Dallas	32.669783	-96.614921	1.0	Deli / Bodega	Home Service	Mexican Restaurant
1	75252	Dallas	32.998132	-96.790880	2.0	Mexican Restaurant	Sandwich Place	Nail Salon
2	75202	Dallas	32.779880	-96.805020	2.0	Hotel	Mexican Restaurant	Coffee Shop
3	75270	Dallas	32.781330	-96.801980	2.0	Hotel	Coffee Shop	Mexican Restaurant
4	75220	Dallas	32.867977	-96.863060	6.0	Mexican Restaurant	Pizza Place	Grocery Store

Figure 8: Most Common Venues with Cluster Groups Sample

Results

To present the 8 different clusters the coordinates of the two cities of Dallas and Fort Worth were acquired using geocoding. Next, a folium map of both Dallas and Fort Worth was created showing the various clusters identified by unique colors. The Dallas map can be seen in Figure 9 and the Fort Worth map can be seen in Figure 10.

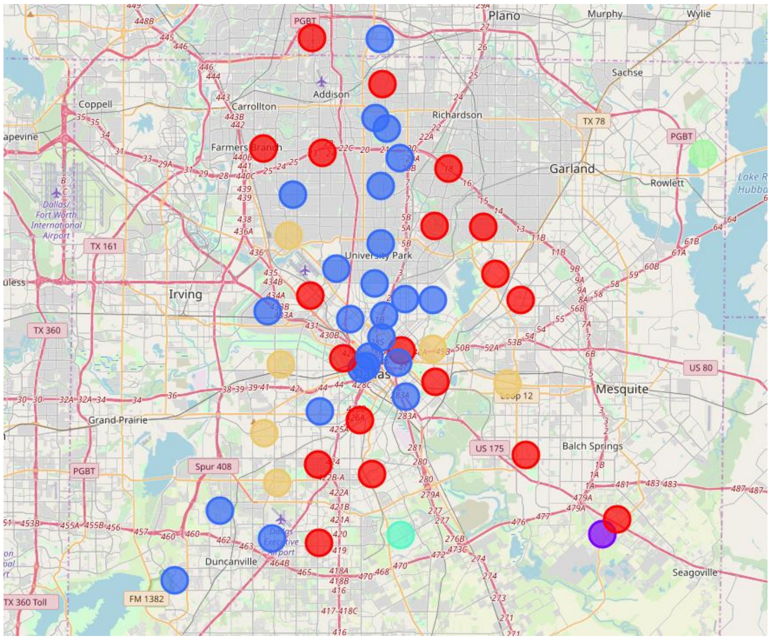


Figure 9: Dallas Clusters

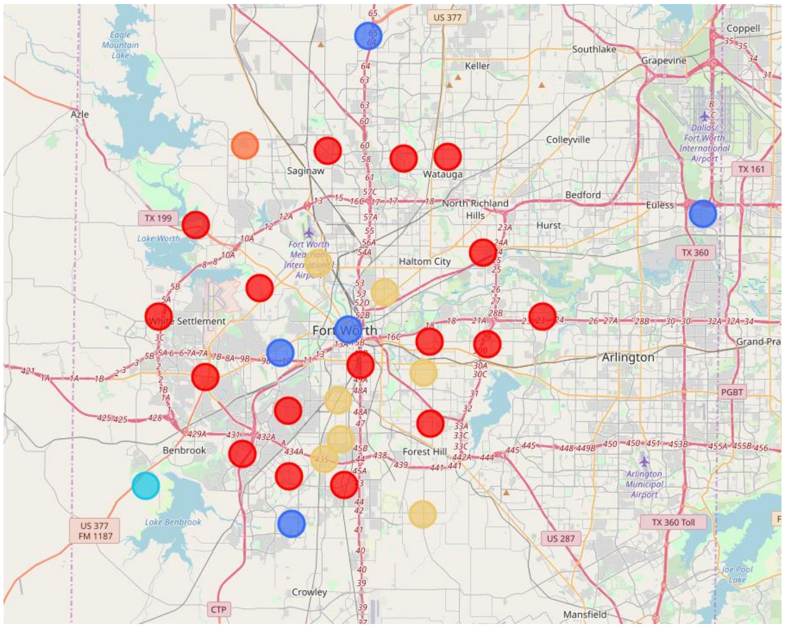


Figure 10: Fort Worth Clusters

Discussion

The largest cluster contained 36 similar ZIP Code regions. This was nearly equal with 19 clusters in Dallas and 17 clusters in Fort Worth. The next cluster contained 29 ZIP Codes regions with a significantly larger amount in Dallas at 24 and only 5 in Fort Worth. The next cluster contained 13 ZIP Codes with a close split of 6 in Dallas and 7 in Fort Worth. 5 clusters contained a single ZIP Code region, forming its separate cluster with less similarity than other clusters. Dallas was found to have 3 unique locations and Fort Worth 2.

Before the data was cleaned it was found that multiple ZIP Codes shared the same coordinates. This is likely due to error from the source. Acquiring the data for associated coordinates with ZIP Codes is not widely available and it was determined that a sufficient analysis could be completed while dropping those ZIP Code regions.

The analysis was conducted with only the venues available on Foursquare and there may be present venues that have not yet been recorded. An increasingly accurate model can be produced as more venues are defined.

Conclusion

The Dallas- Fort Worth Metroplex as one the largest and fastest-growing areas in the United States is the host of ongoing business expansion ventures. Essential to their success is location choice. For further study, each similar venue can be compared to its cluster traits to determine successful relationships. By identifying 8 venue clusters in the two cities of Dallas and Fort Worth one can observe various trends in positioning in different ZIP Code regions, exhibiting similar as well as unique clusters.