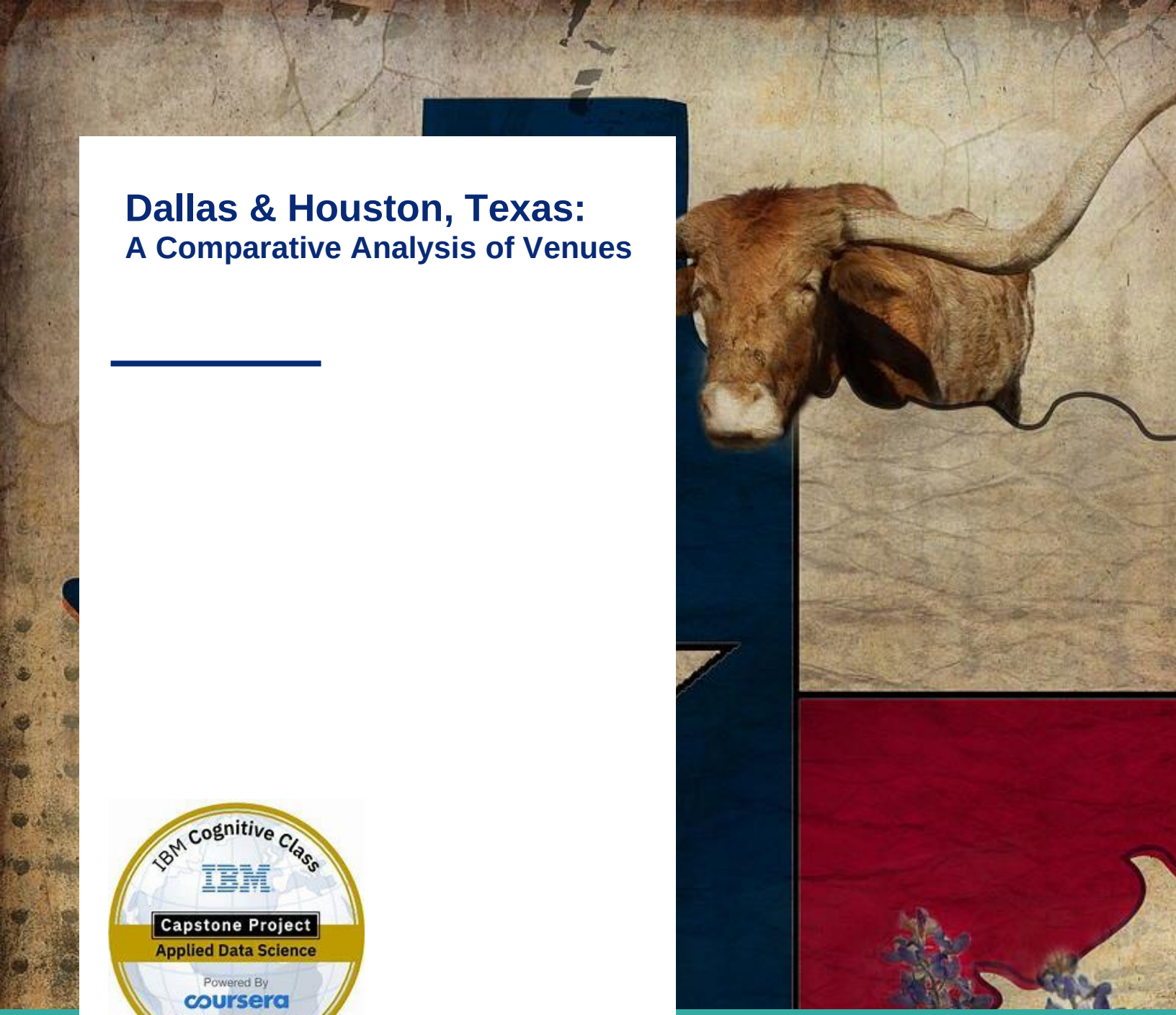


Dallas & Houston, Texas:
A Comparative Analysis of Venues

AUGUST 9

IBM Data Science Capstone Project
Authored by: Paul Bristow

Introduction

PROBLEM DESCRIPTION

Having lived in Texas for approximately 21 years, I wanted to better understand the differences between two of the largest and most populous cities in Texas. This report will compare and contrast the distribution of venues throughout the City of Dallas and the City of Houston, TX. Using various data gathering, analytic and visualization tools learned throughout the Data Science Certification Program from IBM through Coursera, this report will use quantitative analysis to provide a better understanding on the distribution of venues throughout the active zip codes associated with these two cities.

The intent of this document is not to dissuade or present a bias view, but to provide a sound analytical approach to examining and documenting the distribution of venues throughout Dallas and Houston. Should any variances be identified between Dallas and Houston, this report will suggest approaches that can be used for future analysis of those variances.

As a result of this report, readers considering relocating or visiting Dallas or Houston should be enabled to make better decisions based on their preferences associated to the venues and distribution of those throughout these two cities.

DATA

For this analysis, the data used will be retrieved from multiple sources, including:

- Gas Lamp Media Blog – file with all the US zip codes and their associated latitude, longitude, city, state, and county. (First 5 rows of represented in Figure 1)

	zip_code	latitude	longitude	city	state	county
0	501	40.922326	-72.637078	Holtsville	NY	Suffolk
1	544	40.922326	-72.637078	Holtsville	NY	Suffolk
2	601	18.165273	-66.722583	Adjuntas	PR	Adjuntas
3	602	18.393103	-67.180953	Aguada	PR	Aguada
4	603	18.455913	-67.145780	Aguadilla	PR	Aguadilla

http://docs.gaslamp.media/wp-content/uploads/2013/08/z

- World Population Review Site – file containing 2021 active Texas zip codes with their associated population

	Zip Code	City	County	Population
0	77449	Katy	Harris County	128294.0
1	77494	Katy	Fort Bend County	118291.0
2	79936	El Paso	El Paso County	111620.0
3	75034	Frisco	Collin County	108525.0
4	77084	Houston	Harris County	107673.0

https://worldpopulationreview.com/zips/texas

- Data will be retrieved through Foursquare API searching for “venues” throughout the active Dallas and Houston ZIP Code regions.

Data from the first two sources will be merged to extract only active Zip codes located in each of the cities. The data from this merged file will be used with the Foursquare API to

provide segmentation, clustering and visualization of similar regions throughout Houston and Dallas.

Using an unsupervised machine learning technique with the merged data and the venues gathered from the Foursquare API, K-Means clustering will provide segmentation and clustering data. Through georeferencing and Python Folium, as well as, Plotly's Python plotting library, maps and charts will be used to visualize regions and the distribution of venues throughout Dallas and Houston.

METHODOLOGY

The first step for processing was to have consistent sets of data established for the City of Dallas and Houston showing the active zip codes within those geographies. By using the Foursquare API, geospatial data was obtained for the zip codes and the various venues located within them. The data also included categorization of the venues to assist with determining clustering similarities across the active zip codes.

This resulted in the creation of two Python dataframes resembling below (first 5 rows):

Dallas Venue Data

	City	Zipcode	Zipcode Latitude	Zipcode Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Dallas	75201	32.781179	-96.790329	Metropolitan Cafe	32.781776	-96.792636	Diner
1	Dallas	75201	32.781179	-96.790329	Ruibal's Plants of Texas	32.778413	-96.790415	Garden Center
2	Dallas	75201	32.781179	-96.790329	Green Door Public House	32.778527	-96.792077	American Restaurant
3	Dallas	75201	32.781179	-96.790329	Dallas Farmers Market	32.777630	-96.789290	Farmers Market
4	Dallas	75201	32.781179	-96.790329	Palmieri Cafe	32.777678	-96.790263	Coffee Shop

Houston Venue Data

	City	Zipcode	Zipcode Latitude	Zipcode Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Houston	77002	29.807651	-95.391447	Lei Low	29.808812	-95.389321	Tiki Bar
1	Houston	77002	29.807651	-95.391447	Halbert Park	29.808657	-95.394615	Park
2	Houston	77002	29.807651	-95.391447	Bellissimo Ristorante	29.809147	-95.389702	Italian Restaurant
3	Houston	77002	29.807651	-95.391447	Prestigious Nails	29.803985	-95.393199	Cosmetics Shop
4	Houston	77002	29.807651	-95.391447	Puebla's Mexican Kitchen & Bakery	29.808029	-95.389307	Mexican Restaurant

Following above, the data will be examined to see the distribution of venue categories across both cities to compare and contrast the similarities and differences of those venue categories in Dallas and Houston.

Once that is completed, the data for both cities will be independently processed using machine a learning technique referred to as "One Hot Encoding". This technique will take the categorical data, "Venue Category" (see above table headings) and assign numerical values to them so they can be further used for Cluster Analysis to determine zip codes with similar distributions of venues.

Lastly, using "The Elbow Method", optimal number of similar clusters should be determined by city. The distribution of the clusters will then be highlighted using Folium, a powerful data visualization library in Python that was built primarily to help people visualize geospatial data.

RESULTS

The following table illustrates the data obtained through the zip code databases used and from processing the venue related data from the Foursquare API.

	Dallas	**Houston**
Population(M)	1.4	3.2
Active Zip Codes processed	47	96
Venues Uncovered	2473	4028
Unique Venue Categories to City	43	104
Total Unique Venue Categories	280	341
% of Top 10 Venue Categories	27.3	26.1

Notes:

- After reviewing the data, it appeared as if many of the Venues categorized as "Convenience Stores" where actually gas stations (i.e. Quick Trip, 7 Eleven, Race Track, etc). To ensure consistency in the data comparison between Dallas and Houston, these were combined as "Gas Stations".

- “Unique Venue Categories to City” means the number of categories in one city that aren’t in the other. Per the above table, there were 43 venue categories identified in Dallas that were not in Houston. This is based on the data extracted from the Foursquare API. Coding of those venue categories might not be consistent from one city to another. For example, one category that was found in the Houston data and not in the Dallas data was “Beer Store”. If someone was to use this data because they are considering opening a “Beer Store” in Dallas, they might consider looking at the details on the venues tied to “Beer Stores” in Houston before progressing further. Using this approach, they can determine if they represent similar venues, but, coded differently for Dallas (i.e. liquor store).

The top 10 venues by category, in Dallas, can be depicted as follows:

The Dallas data was examined to see the top 10 venues per zip code. A table highlighting those was produced resembling below (first 5 rows only)

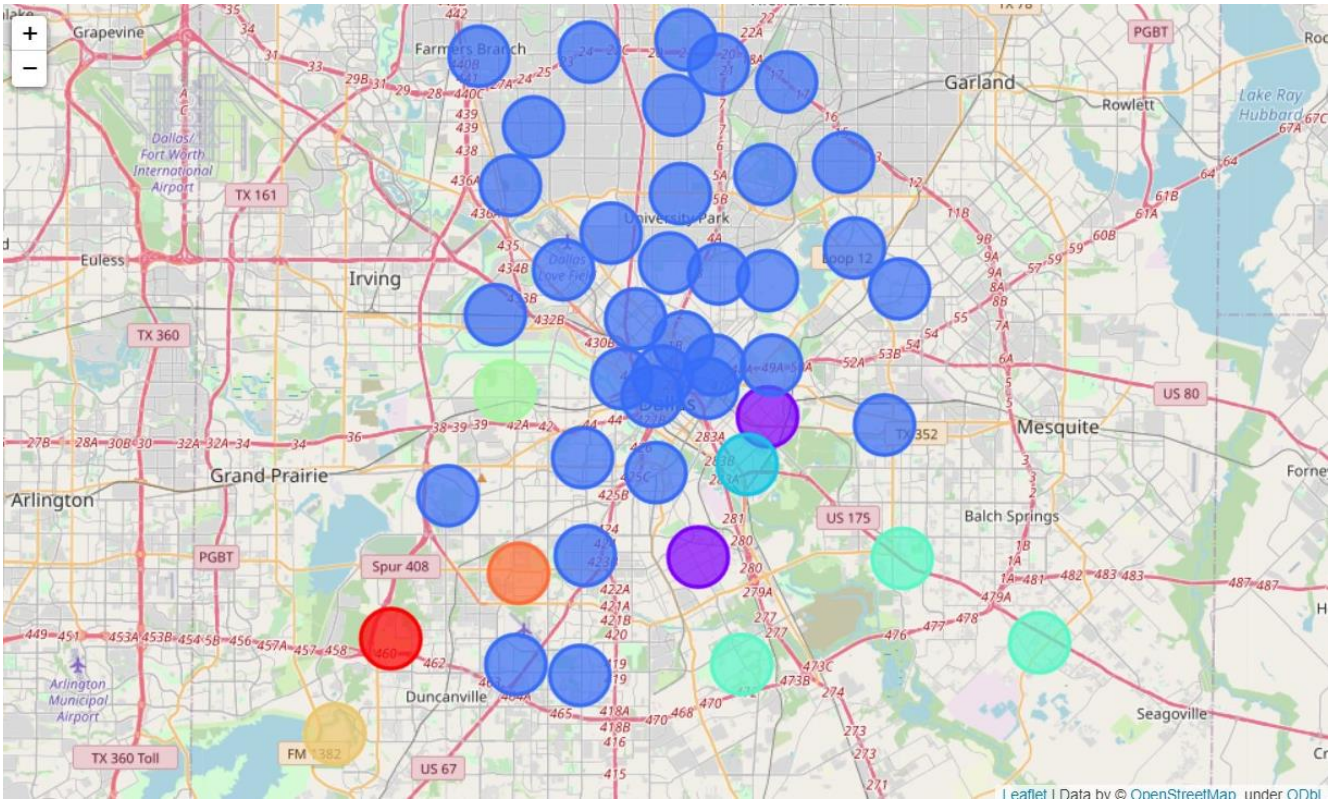
	City	Zipcode	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Dallas	75201	Hotel	Bar	American Restaurant	Cocktail Bar	Coffee Shop	Burger Joint	Gym	Rock Club	Nightclub	Taco Place
1	Dallas	75202	Skate Park	Restaurant	BBQ Joint	Coffee Shop	Sandwich Place	Steakhouse	Park	Pawn Shop	Bank	Seafood Restaurant
2	Dallas	75203	Light Rail Station	Clothing Store	Discount Store	Fried Chicken Joint	Gas Station	Park	Food	Pizza Place	Market	Scenic Lookout
3	Dallas	75204	Fast Food Restaurant	Pizza Place	Gas Station	Mexican Restaurant	Sandwich Place	Taco Place	Thai Restaurant	Fried Chicken Joint	Bank	Coffee Shop
4	Dallas	75205	Furniture / Home Store	Pizza Place	French Restaurant	Mexican Restaurant	Sushi Restaurant	Yoga Studio	Sandwich Place	Italian Restaurant	Spa	Gift Shop

Using One Hot Encoding, the Venue Categories in each active Dallas Zip Code were numericized to be used for K-means clustering. This resulted in the production of a table (47 rows x 215 columns) resembling this:

	City	Zipcode	American Restaurant	Antique Shop	Aquarium	Art Gallery	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	Automotive Shop	...	Video Store	Vietnamese Restaurant	Volleyball Court	Warehouse	Wine Bar	Wings Joint	Women's Store	Yoga Studio	Zoo
0	Dallas	75201	0.060000	0.010000	0.00	0.010000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.010000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
1	Dallas	75202	0.000000	0.000000	0.00	0.062500	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
2	Dallas	75203	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
3	Dallas	75204	0.000000	0.000000	0.00	0.000000	0.014286	0.014286	0.000000	0.000000	...	0.028571	0.028571	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
4	Dallas	75205	0.018519	0.000000	0.00	0.000000	0.000000	0.000000	0.018519	0.000000	...	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.037037	0.000000

Using “The Elbow Method” for cluster analysis on the data above, it was determined that the optimal number of clusters to highlight for Dallas was 8. The following map was produced by Folium to illustrate the distribution of these clusters throughout Dallas.

The following map was extracted using Folium with the common clusters of venues highlighted throughout Dallas.



Each Cluster has a list of the Zip Codes and Top 10 Venues associated to each Zip Code. These can be reviewed in Source code store on GitHub for this project. A table for each cluster was produced highlighting the top 10 venues per zip code in the clusters.

Most common venues by Dallas Zip Code for Cluster 1

	zip_code	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
32	75236	0	Intersection	Campground	Gas Station	Trail	Home Service	Fishing Spot	Eye Doctor	Fabric Shop	Farmers Market	Fast Food Restaurant

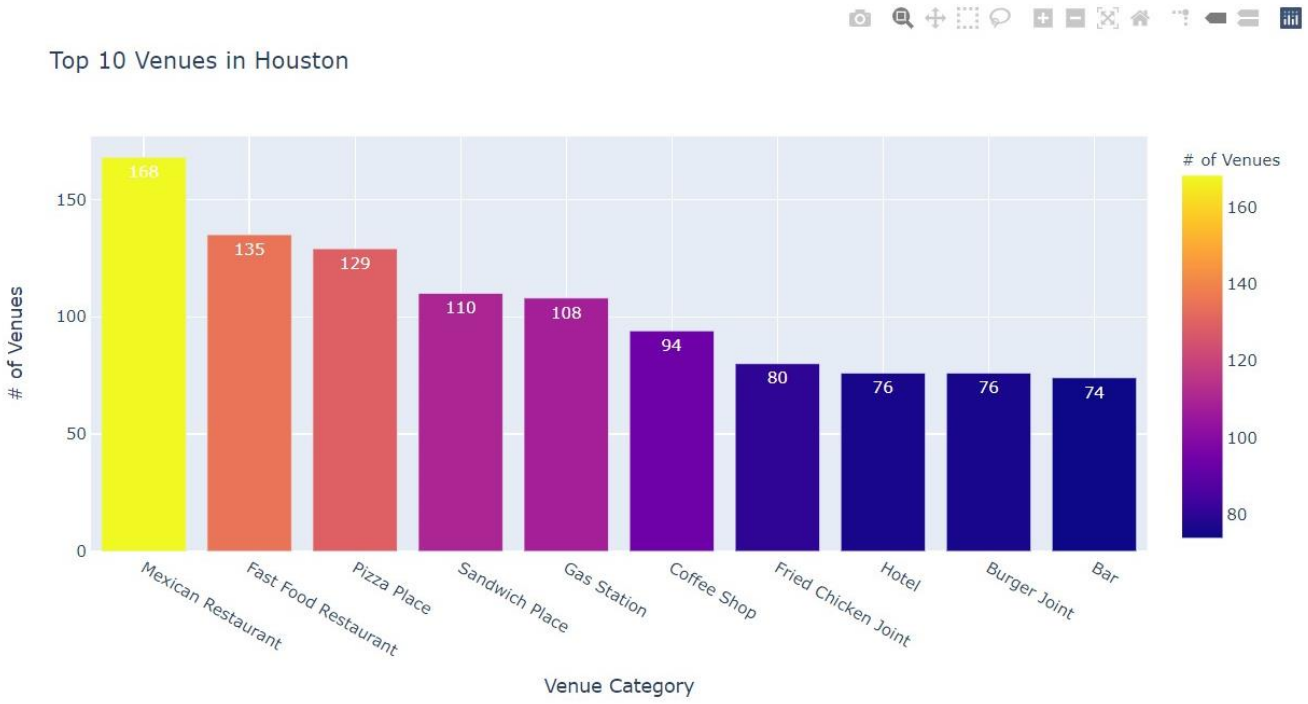
Most common venues by Dallas Zip Code for Cluster 2

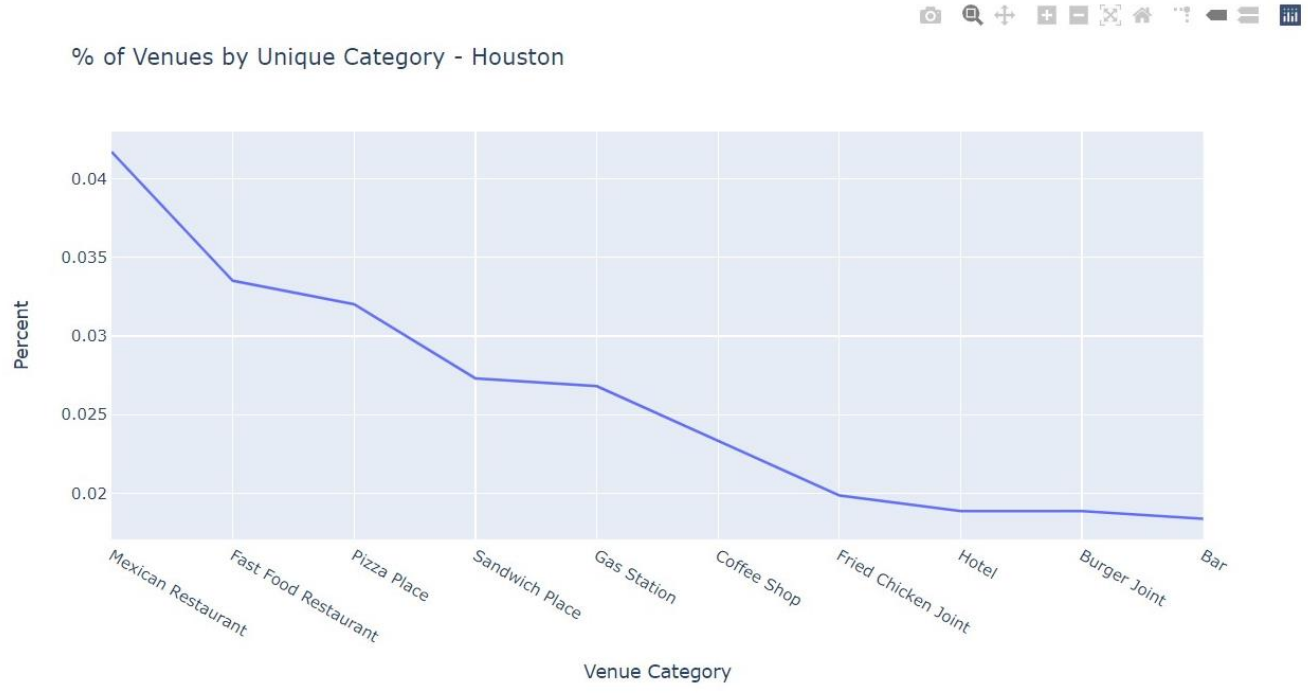
	zip_code	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
9	75210	1	Theme Park Ride / Attraction	Fried Chicken Joint	Recreation Center	Discount Store	Performing Arts Venue	Fast Food Restaurant	Opera House	General Entertainment	Mattress Store	Train Station
14	75216	1	Recreation Center	Discount Store	Fast Food Restaurant	Fried Chicken Joint	Dance Studio	Eye Doctor	French Restaurant	Football Stadium	Food Truck	Food Court

Most common venues by Dallas Zip Code for Cluster 3

	zip_code	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 M Comm Ve
0	75201	2	Hotel	Cocktail Bar	Steakhouse	American	New American	Coffee	Japanese	Restaurant	Bar	Italian

The top 10 venues by category, in Houston, can be depicted as follows:





The Houston data was examined to see the top 10 venues per zip code. A table highlighting those was produced resembling below (first 5 rows only)

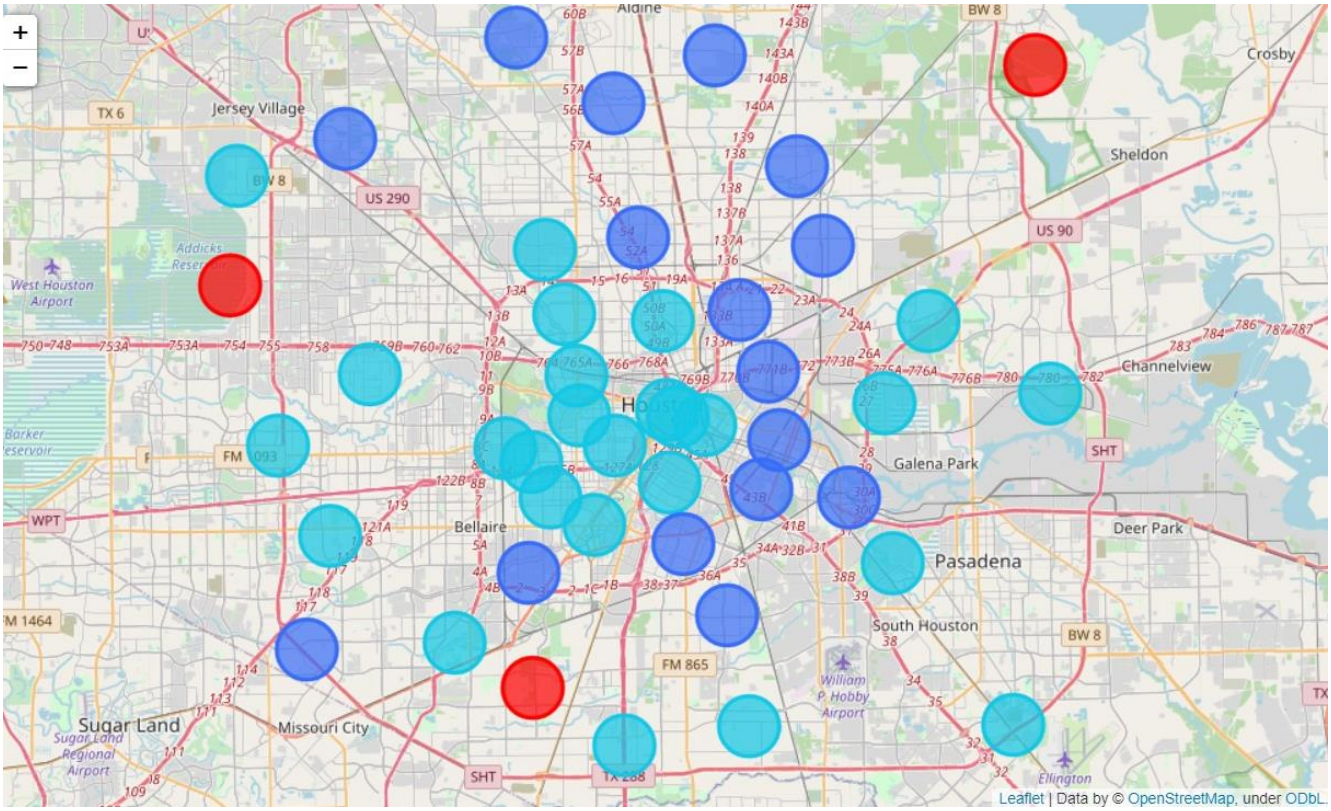
	City	Zipcode	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Houston	77002	Hotel	Mexican Restaurant	Park	Italian Restaurant	Burger Joint	Lounge	Theater	New American Restaurant	Baseball Stadium	Gym
1	Houston	77003	Mexican Restaurant	Bar	Hotel	Italian Restaurant	Food Truck	Vietnamese Restaurant	Sandwich Place	Brewery	Music Venue	American Restaurant
2	Houston	77004	Park	Fast Food Restaurant	Bar	Gas Station	Fried Chicken Joint	Lounge	Southern / Soul Food Restaurant	Breakfast Spot	Coffee Shop	Gym
3	Houston	77005	Ice Cream Shop	Coffee Shop	American Restaurant	Italian Restaurant	Mexican Restaurant	Bookstore	Bar	Food Truck	Burger Joint	Bakery
4	Houston	77006	Coffee Shop	Bar	Italian Restaurant	Mexican Restaurant	Mediterranean Restaurant	Beer Garden	Pizza Place	Cocktail Bar	Breakfast Spot	Gay Bar

Using One Hot Encoding, the Venue Categories in each active Houston Zip Code were numericized to be used for K-means clustering. This resulted in the production of a table resembling this:

	City	Zipcode	ATM	Accessories Store	Afghan Restaurant	African Restaurant	Airport	Airport Food Court	Airport Lounge	Airport Service	...	Well	Whisky Bar	Wine Bar	Wine Shop	Wings Joint	Wo
0	Houston	77002	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.01	0.010000	0.00	0.000000	0.0
1	Houston	77003	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.00	0.000000	0.00	0.000000	0.0
2	Houston	77004	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.00	0.015152	0.00	0.015152	0.0
3	Houston	77005	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.00	0.030000	0.01	0.000000	0.0
4	Houston	77006	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.01	0.020000	0.00	0.000000	0.0

Using “The Elbow Method” for cluster analysis on the data above, it was determined that the optimal number of clusters to highlight for Houston was 5.

The following map was produced by Folium to illustrate the distribution of these clusters throughout Dallas.



DISCUSSION

As reflected in the above data, it appears as if Houston has more unique venue categories than in comparison to Dallas. This is based on the current data available through the Foursquare API. As more venue information is populated within Foursquare, venue cluster dynamics may change for both cities.

80% of the top 10 Venue Categories were the same in both cities. In the remaining 20% of venues, “Parks” and “American Restaurants” contributed to those categories in Dallas and “Fried Chicken” and “Burger Joints” contributed to those in Houston.

Lastly, based on the One Hot encoding analysis conducted, Dallas had 8 similar clusters spread throughout the city, whereas, Houston had 5. This could mean that for Dallas residents to find specific venues, they may have to leave those geographic areas associated to the clusters to find what they are looking for. This may not be an issue given that Dallas

geography is significantly smaller than compared to Houston (386 sq miles for Dallas compared to 628 sq miles for Houston).

The diversity of venues appeared to be higher in the Houston segments compared to Dallas. Dallas had venue categories not present in Houston and vice-versa. Categories were obtained from Foursquare and the accuracy of those categories were dependent on how they were coded in Foursquare. For example, a "Burger Joint" might have been encoded as an "American Restaurant" or vice-versa.

CONCLUSION

Dallas and Houston are two of the largest cities in Texas. By looking at the venues in both cities, you can see the similarities and differences of these cities based on those venues. This can be used to assess venues that might be successful in Houston to see if there might be an opportunity to bringing them to Dallas or vice-versa.

Looking at the various clusters can help determine the relationship between venues. This can also be used to evaluate the potential for adding new venues based on a specific category.

The tools used in this project can be used for much more analysis to help with determining optimal placement of new venues should someone think of doing so. This could include correlating demographics analysis and per capita analysis along with the venue data extracted from the Foursquare API.

As stated in the introduction, this project was designed to provide insight into the distribution of venues throughout Dallas and Houston. While the report has provided cursory information, looking at the full source code output and further evaluating the data from Foursquare could be instrumental in helping gain further insight to how these two Texas cities live with the distributions of venues spread throughout them.