Texas Derby – Austin, Dallas, or Houston?

IBM Data Science Final Capstone Project

Yuhao Sun

2020.4

Table of Contents

Introduction / Scope

Austin, Dallas, and Houston are the three major cities in the great state of Texas. In the past decades, these three cities have experienced rapid population growth amid Texas's progressing economy. **Austin**, being the capital of the Lonestar State with an eclectic mix of history, culture, and gorgeous sights, has also emerged as a technology hub. Austin's Silicone Hill is considered a substitute for the Bay Area's Silicon Valley. **Dallas**, together with its rapidly growing twin city of Fort Worth and other satellite cities in the vicinity, stands as a multi-versed financial center at the Southwest. **Houston** has always been the energy center in America. It's engineering, petrochemical, and other industries have empowered this city with a robust lifestyle.

Texas has been consistently ranked **number on**e for the inflow of migrants from other states across America due to employment opportunities and relatively low living costs. While most of the people inflow will settle down in one of the three cities abovementioned, the cities themselves have a lot of things to offer and explore. This analysis intends to show which areas (by Zip Code) of a town resemble those of the others **(Austin, Dallas, Houston).**

This report could be helpful for different use cases:

- **People moving from one city to the other** often would like to live in a particular type of neighborhood. This comparison can help those to filter for areas similar (or even different, if you are up for something new) to what you are used to.
- **People moving from other states to Texas** often wonder where to settle down. This comparison can help those to filter for neighborhoods similar or different from their original came from, and introduce them into local fun and popular places
- **Companies expanding within one of the cities** might want to look for a similar type of neighborhoods, as they are targeting a specific user group. The comparison can offer the first indication.
- **Companies expanding from one city to the other** might also try to find a neighborhood to settle in first. They can use their experience from the original city and look for a fitting (e.g., similar) neighborhood in the second one.
- **Which city among the three is the most convenient in terms of living facilities?**

Data Collection and Explanation

Two different kinds of data are used for this comparison.

1. **City zip code area and corresponding geographical data**: to analyze the cities on a meaningful level, they need to be divided into different areas, e.g., neighborhoods, boroughs, or simply by **zip codes**. Luckily the zipcode data is readily available on the internet, and one can be found here.

   This data (uszip.csv), when unzipping, includes 33099 rows of almost **all zip codes across the USA**. It also consists of latitude and longitude coordinates for each zip code, which will be useful later when it comes to geo-location analysis using FourSquare API. Other relevant information, such as corresponding city, county, population, etc. are also included in the dataset. Some cleaning and filtering of data need to be done first in order to **limit the analysis to Texas**, and more specifically, to the **three cities mentioned**.

   When we talk about a city, we often refer to the **metropolis** together with its adjacent areas. For example, when we talk about Dallas, it doesn't only include the city of Dallas, but also include the city of Fort Worth, Plano, Frisco, Denton, etc.. The same principle applies to Austin (The greater Austin Area) and Houston (The greater Houston Area). One way to achieve this is to **include county data**. For example, the city of Dallas is in Dallas County, but the whole Dallas-Forth Worth (DFW) area will consist of 4 counties, i.e., Dallas, Collin, Tarrant, and Denton. So **county data** and **zipcode data** will be used for geo-location.

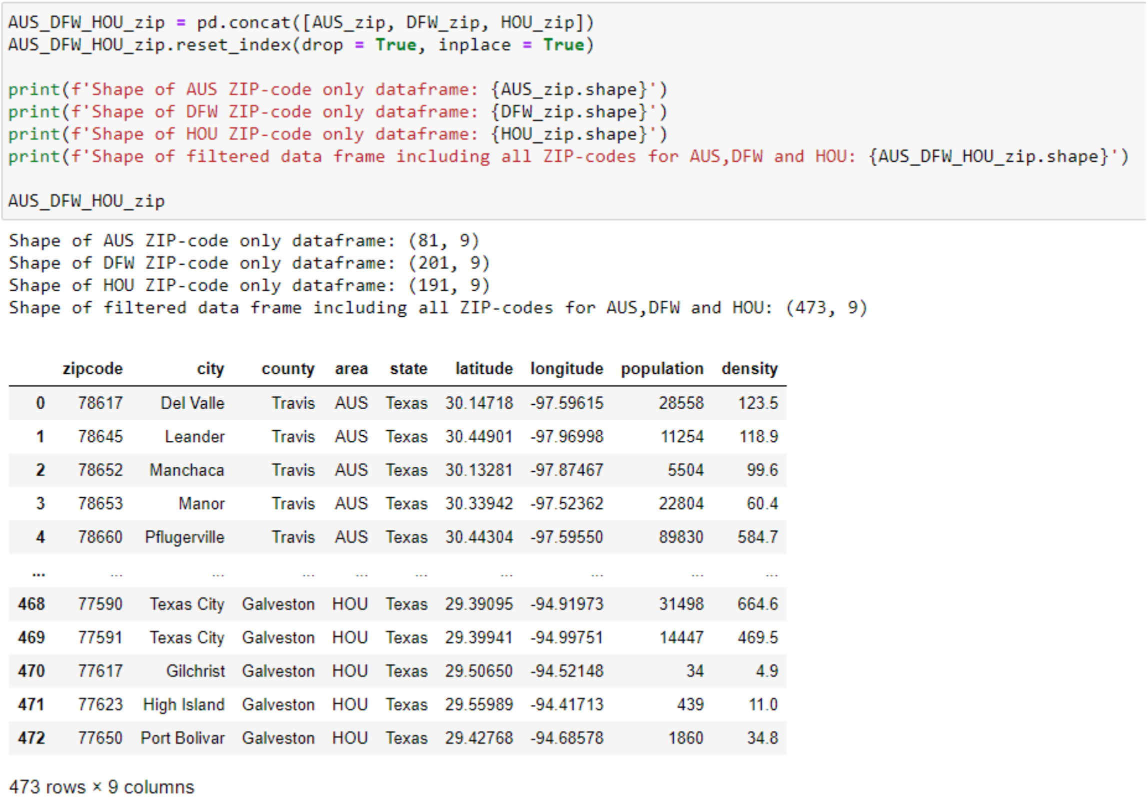
2. **Venue data from Foursquare API**: The first 500 venues per ZIP-code in **Austin (AUS), Dallas (DFW), and Houston (HOU)** are scraped from the **FourSquare API**, in order to cluster according to the zip and county boundaries. Because the zipcode data also contains latitudes and longitudes, the **coordinates** are used to match the venue data and extract information about the Venue name, category, locations using the Foursquare API.

   Other standard applications and analyses of the datasets are referred to various IBM Coursera Course Labs.

Methodology

Basic data processing.

As mentioned before, we will limit the scope of analysis to the three city areas, and each area will consist of four counties. And then, we will gradually reduce redundant information and produced an informational DataFrame containing all the zipcodes and coordinate information.



```
AUS_DFW_HOU_zip = pd.concat([AUS_zip, DFW_zip, HOU_zip])
AUS_DFW_HOU_zip.reset_index(drop = True, inplace = True)

print(f'Shape of AUS ZIP-code only dataframe: {AUS_zip.shape}')
print(f'Shape of DFW ZIP-code only dataframe: {DFW_zip.shape}')
print(f'Shape of HOU ZIP-code only dataframe: {HOU_zip.shape}')
print(f'Shape of filtered data frame including all ZIP-codes for AUS,DFW and HOU: {AUS_DFW_HOU_zip.shape}')

AUS_DFW_HOU_zip
```

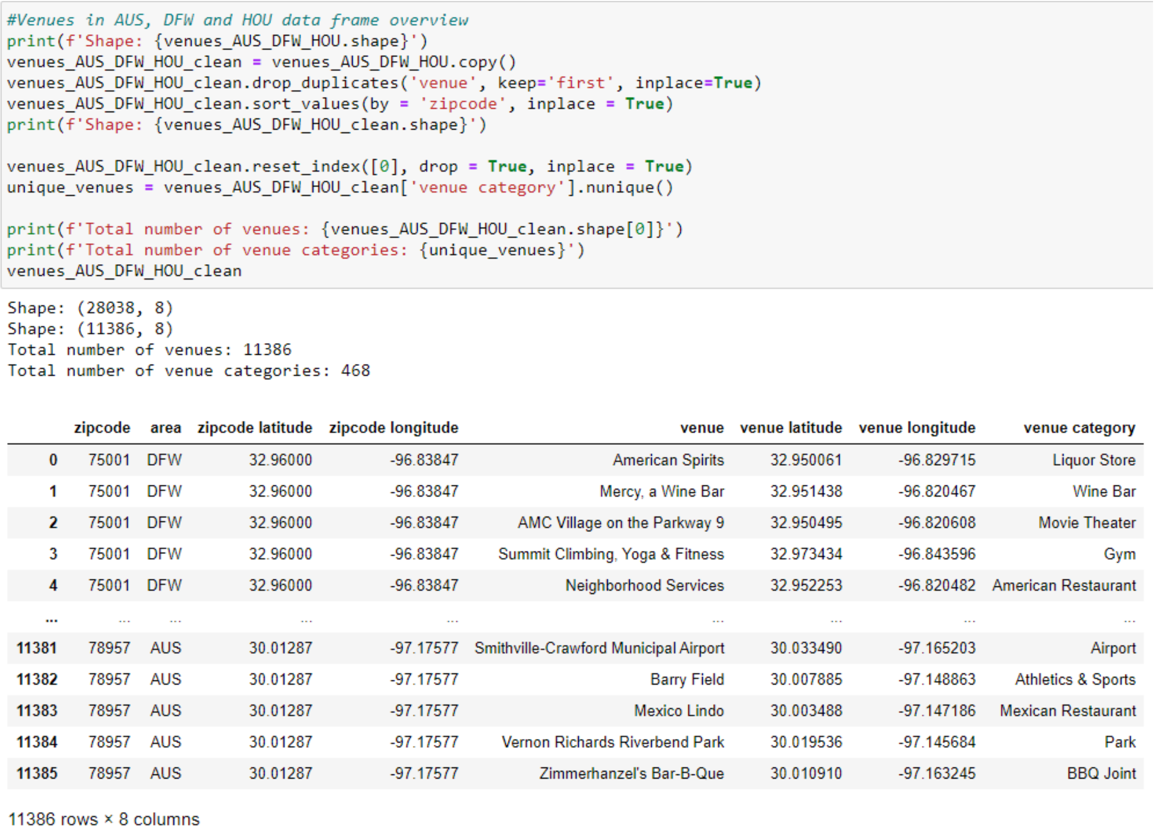
```
Shape of AUS ZIP-code only dataframe: (81, 9)
Shape of DFW ZIP-code only dataframe: (201, 9)
Shape of HOU ZIP-code only dataframe: (191, 9)
Shape of filtered data frame including all ZIP-codes for AUS,DFW and HOU: (473, 9)
```

	zipcode	city	county	area	state	latitude	longitude	population	density
0	78617	Del Valle	Travis	AUS	Texas	30.14718	-97.59615	28558	123.5
1	78645	Leander	Travis	AUS	Texas	30.44901	-97.96998	11254	118.9
2	78652	Manchaca	Travis	AUS	Texas	30.13281	-97.87467	5504	99.6
3	78653	Manor	Travis	AUS	Texas	30.33942	-97.52362	22804	60.4
4	78660	Pflugerville	Travis	AUS	Texas	30.44304	-97.59550	89830	584.7
...	...	...	...	...	...	...	...	...	...
468	77590	Texas City	Galveston	HOU	Texas	29.39095	-94.91973	31498	664.6
469	77591	Texas City	Galveston	HOU	Texas	29.39941	-94.99751	14447	469.5
470	77617	Gilchrist	Galveston	HOU	Texas	29.50650	-94.52148	34	4.9
471	77623	High Island	Galveston	HOU	Texas	29.55989	-94.41713	439	11.0
472	77650	Port Bolivar	Galveston	HOU	Texas	29.42768	-94.68578	1860	34.8

473 rows × 9 columns

Data Scraping using FourSquare API

The FourSquare API is set up and run to scrape venues per zipcode with associated geo-coordinates. It will retrieve the first 500 venues per zipcodes in AUS, DFW, and HOU areas. A total of 11386 unique venues area extracted from the API.



```python
#Venues in AUS, DFW and HOU data frame overview
print(f'Shape: {venues_AUS_DFW_HOU.shape}')
venues_AUS_DFW_HOU_clean = venues_AUS_DFW_HOU.copy()
venues_AUS_DFW_HOU_clean.drop_duplicates('venue', keep='first', inplace=True)
venues_AUS_DFW_HOU_clean.sort_values(by = 'zipcode', inplace = True)
print(f'Shape: {venues_AUS_DFW_HOU_clean.shape}')

venues_AUS_DFW_HOU_clean.reset_index([0], drop = True, inplace = True)
unique_venues = venues_AUS_DFW_HOU_clean['venue category'].nunique()

print(f'Total number of venues: {venues_AUS_DFW_HOU_clean.shape[0]}')
print(f'Total number of venue categories: {unique_venues}')
venues_AUS_DFW_HOU_clean
```

```
Shape: (28038, 8)
Shape: (11386, 8)
Total number of venues: 11386
Total number of venue categories: 468
```

	zipcode	area	zipcode latitude	zipcode longitude	venue	venue latitude	venue longitude	venue category
0	75001	DFW	32.96000	-96.83847	American Spirits	32.950061	-96.829715	Liquor Store
1	75001	DFW	32.96000	-96.83847	Mercy, a Wine Bar	32.951438	-96.820467	Wine Bar
2	75001	DFW	32.96000	-96.83847	AMC Village on the Parkway 9	32.950495	-96.820608	Movie Theater
3	75001	DFW	32.96000	-96.83847	Summit Climbing, Yoga & Fitness	32.973434	-96.843596	Gym
4	75001	DFW	32.96000	-96.83847	Neighborhood Services	32.952253	-96.820482	American Restaurant
...	...	...	...	...	...	...	...	...
11381	78957	AUS	30.01287	-97.17577	Smithville-Crawford Municipal Airport	30.033490	-97.165203	Airport
11382	78957	AUS	30.01287	-97.17577	Barry Field	30.007885	-97.148863	Athletics & Sports
11383	78957	AUS	30.01287	-97.17577	Mexico Lindo	30.003488	-97.147186	Mexican Restaurant
11384	78957	AUS	30.01287	-97.17577	Vernon Richards Riverbend Park	30.019536	-97.145684	Park
11385	78957	AUS	30.01287	-97.17577	Zimmerhanzel's Bar-B-Que	30.010910	-97.163245	BBQ Joint

11386 rows × 8 columns

Exploratory Data Analysis

Before clustering, we compare the venues by exploratory data analysis. We will first examine the most common venues that appeared in the datasets for the three cities to gain an overall taste of the three cities. This step is used to determine which city is most convenient in terms of most common venues that appeared in the datasets.

Clustering – K Means

The final data frame is used for the analysis of the ZIP-codes: clustering in order to find similarity in areas among the three cities.

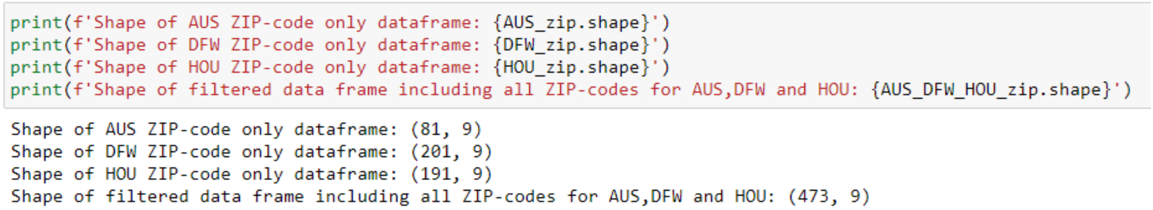
Clustering is used for the segmentation of the different ZIP-codes and is one of the unsupervised machine learning methods. Each cluster is a group of objects (i.e., ZIP-code) that are similar to other objects in the cluster, and dissimilar to data points in other clusters.

For this analysis, k-Means clustering is used. k-Means is a type of partition-based clustering in order to partition the database into groups of individuals with similar characteristics. It divides data into non-overlapping subsets (clusters) without any cluster-internal structure. k-Means tries to minimize intra-cluster distances (e.g., Euclidean or other methods for measuring distance) and maximize inter-cluster distances. It is an iterative algorithm, but the results depend on the initial defined number of clusters. In turn, this means that results (i.e., clusters) are guaranteed, but may not be optimum. Therefore, the algorithm will be run several times with different amounts of initially defined clusters. The algorithm returns inertia, or cost, which can be recognized as a measure of how internally coherent clusters are.

Data Visualization with Folium

Finally, we will show the clustering results using the Folium library to map each zip code area on a map to show the similarities/differences among the three cities.

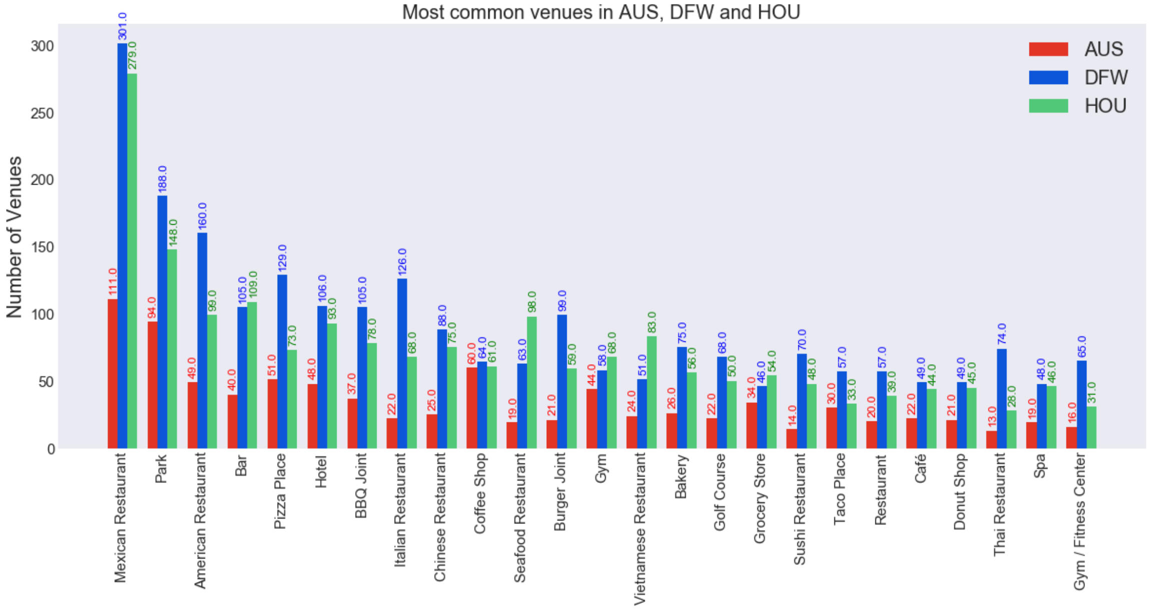
Results and Discussion



```
print(f'Shape of AUS ZIP-code only dataframe: {AUS_zip.shape}')
print(f'Shape of DFW ZIP-code only dataframe: {DFW_zip.shape}')
print(f'Shape of HOU ZIP-code only dataframe: {HOU_zip.shape}')
print(f'Shape of filtered data frame including all ZIP-codes for AUS,DFW and HOU: {AUS_DFW_HOU_zip.shape}')
```

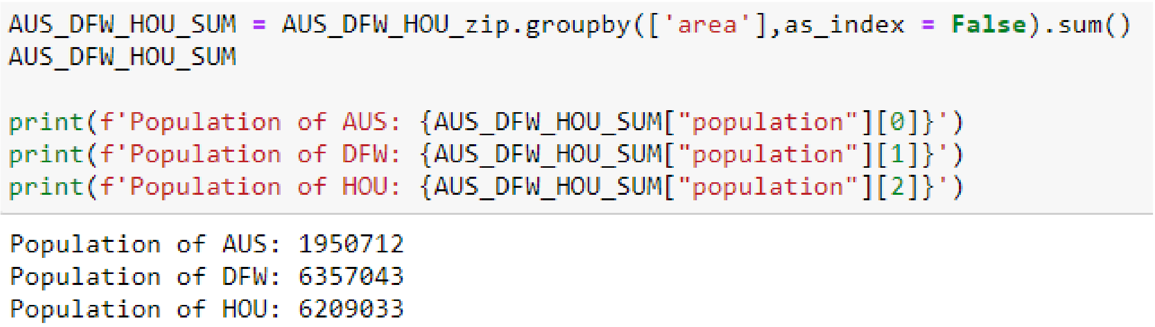
```
Shape of AUS ZIP-code only dataframe: (81, 9)
Shape of DFW ZIP-code only dataframe: (201, 9)
Shape of HOU ZIP-code only dataframe: (191, 9)
Shape of filtered data frame including all ZIP-codes for AUS,DFW and HOU: (473, 9)
```

DFW and HOU are about the same size, while AUS is less than half the size of the two mega-cities. (from the previous calculation, we see that the dataset contains **201 zip codes for DFW, 191 for HOU, and only 81 for AUS**). Therefore, in terms of the absolute number of venues returned, AUS is the least, while DFW is slightly better than HOU overall.



- **Mexican restaurants are the most common venue type**. This is expected as Texas shares borders with Mexico, and a significant portion of the population is of Mexican/Latino/Hispanic ethnicity. So the Mexican cuisine here in all three cities must be authentic.
- The three cities have roughly the **same number of coffee shops**, which comes as the 10th most popular venue. Considering the size and population of AUS, this shows AUS people enjoy a higher number of coffee shops. In general, **AUS people enjoy a more bourgeois lifestyle**.

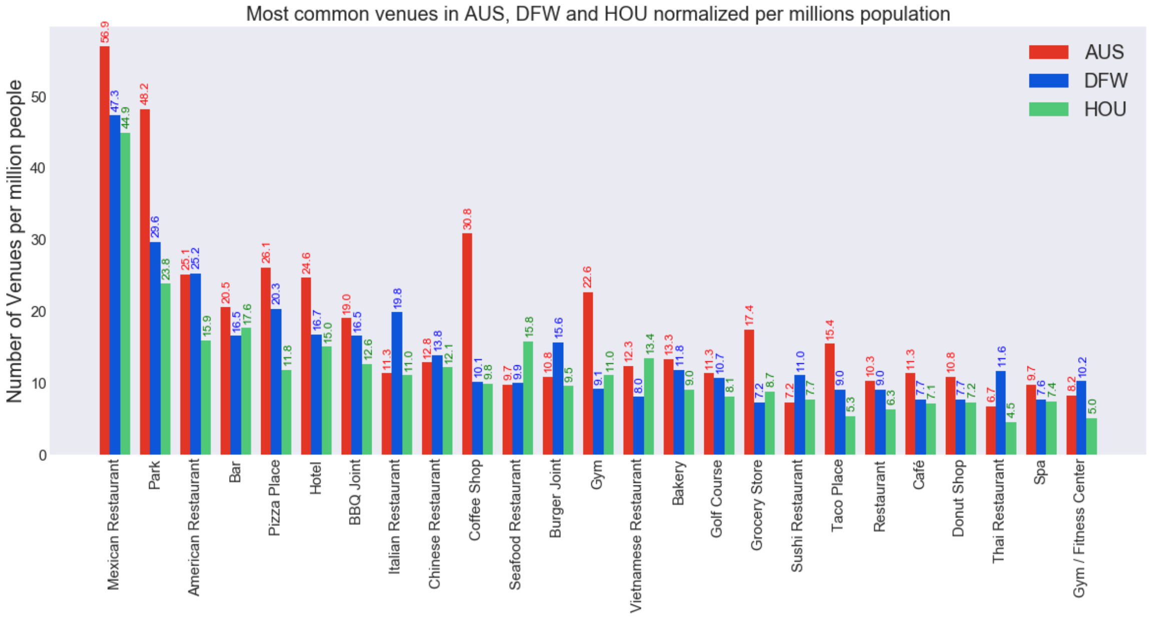
For a better comparison, we should look at those numbers when **normalized with the number of inhabitants of all three cities.**



```
AUS_DFW_HOU_SUM = AUS_DFW_HOU_zip.groupby(['area'],as_index = False).sum()
AUS_DFW_HOU_SUM

print(f'Population of AUS: {AUS_DFW_HOU_SUM["population"][0]}')
print(f'Population of DFW: {AUS_DFW_HOU_SUM["population"][1]}')
print(f'Population of HOU: {AUS_DFW_HOU_SUM["population"][2]}')
```

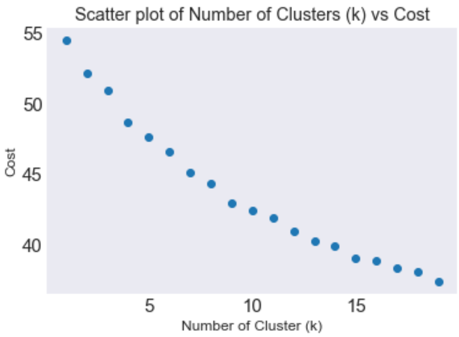
```
Population of AUS: 1950712
Population of DFW: 6357043
Population of HOU: 6209033
```



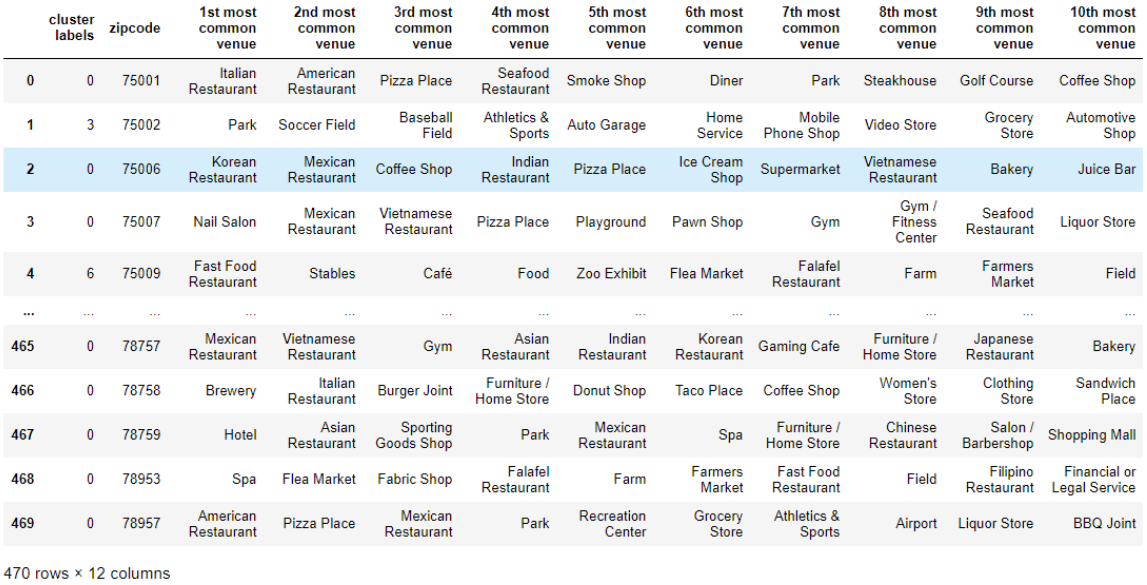
Normalized data shows that, in general, **AUS is the most convenient city** amongst the three. It almost topped every category.

- **Mexican restaurants** are still the most popular venue for all three cities.
- AUS people enjoy a disproportionally large number of parks, coffee shops, gyms, grocery stores, and Taco place, which align with **Austinites' hipster and bourgeois living style**.
- In general, **DFW is slightly better than HOU** in terms of the most common venues.

For clustering practices, we used the K-means method. The problem with k-Means: increasing k always reduces inertia/cost. The value of cost as a function of k is plotted, and an 'elbow' point is determined where the rate of decrease sharply shifts. This is selected as the right k for clustering (elbow method).



Therefore, we choose k=8 as the number of clusters. And to extract the top 10 most common venues that appeared in each zip code area.

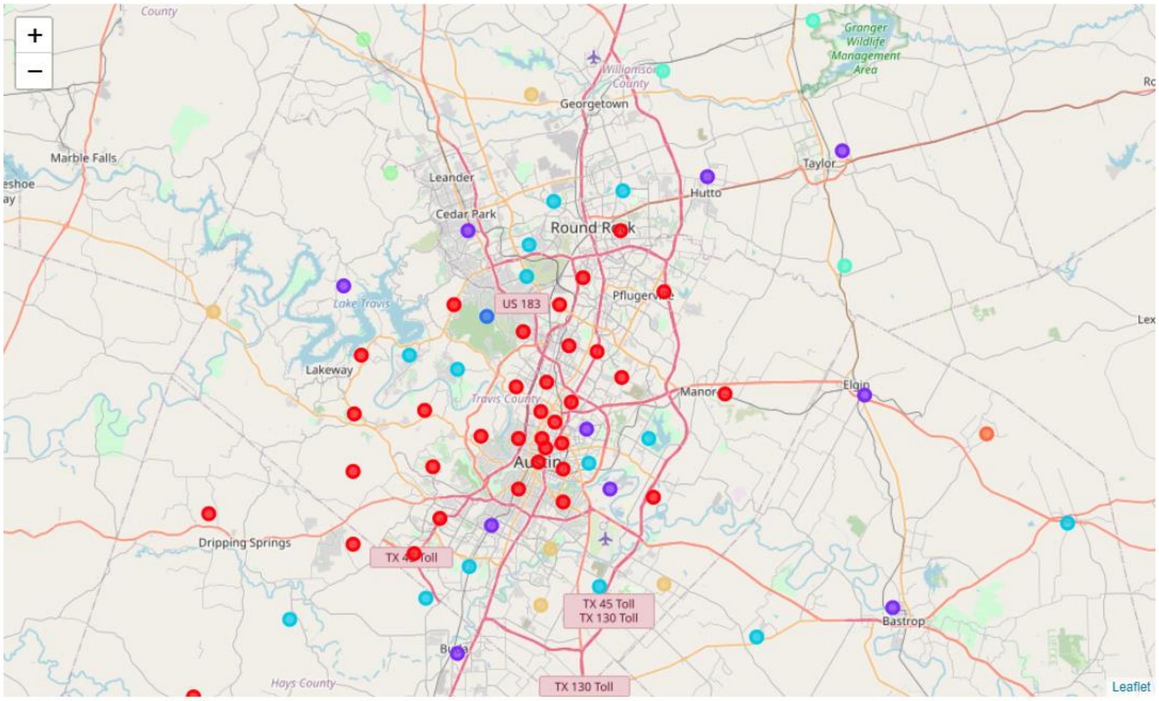


	cluster labels	zipcode	1st most common venue	2nd most common venue	3rd most common venue	4th most common venue	5th most common venue	6th most common venue	7th most common venue	8th most common venue	9th most common venue	10th most common venue
0	0	75001	Italian Restaurant	American Restaurant	Pizza Place	Seafood Restaurant	Smoke Shop	Diner	Park	Steakhouse	Golf Course	Coffee Shop
1	3	75002	Park	Soccer Field	Baseball Field	Athletics & Sports	Auto Garage	Home Service	Mobile Phone Shop	Video Store	Grocery Store	Automotive Shop
2	0	75006	Korean Restaurant	Mexican Restaurant	Coffee Shop	Indian Restaurant	Pizza Place	Ice Cream Shop	Supermarket	Vietnamese Restaurant	Bakery	Juice Bar
3	0	75007	Nail Salon	Mexican Restaurant	Vietnamese Restaurant	Pizza Place	Playground	Pawn Shop	Gym	Gym / Fitness Center	Seafood Restaurant	Liquor Store
4	6	75009	Fast Food Restaurant	Stables	Café	Food	Zoo Exhibit	Flea Market	Falafel Restaurant	Farm	Farmers Market	Field
...	...	...	...	...	...	...	...	...	...	...	...	...
465	0	78757	Mexican Restaurant	Vietnamese Restaurant	Gym	Asian Restaurant	Indian Restaurant	Korean Restaurant	Gaming Cafe	Furniture / Home Store	Japanese Restaurant	Bakery
466	0	78758	Brewery	Italian Restaurant	Burger Joint	Furniture / Home Store	Donut Shop	Taco Place	Coffee Shop	Women's Store	Clothing Store	Sandwich Place
467	0	78759	Hotel	Asian Restaurant	Sporting Goods Shop	Park	Mexican Restaurant	Spa	Furniture / Home Store	Chinese Restaurant	Salon / Barbershop	Shopping Mall
468	0	78953	Spa	Flea Market	Fabric Shop	Falafel Restaurant	Farm	Farmers Market	Fast Food Restaurant	Field	Filipino Restaurant	Financial or Legal Service
469	0	78957	American Restaurant	Pizza Place	Mexican Restaurant	Park	Recreation Center	Grocery Store	Athletics & Sports	Airport	Liquor Store	BBQ Joint

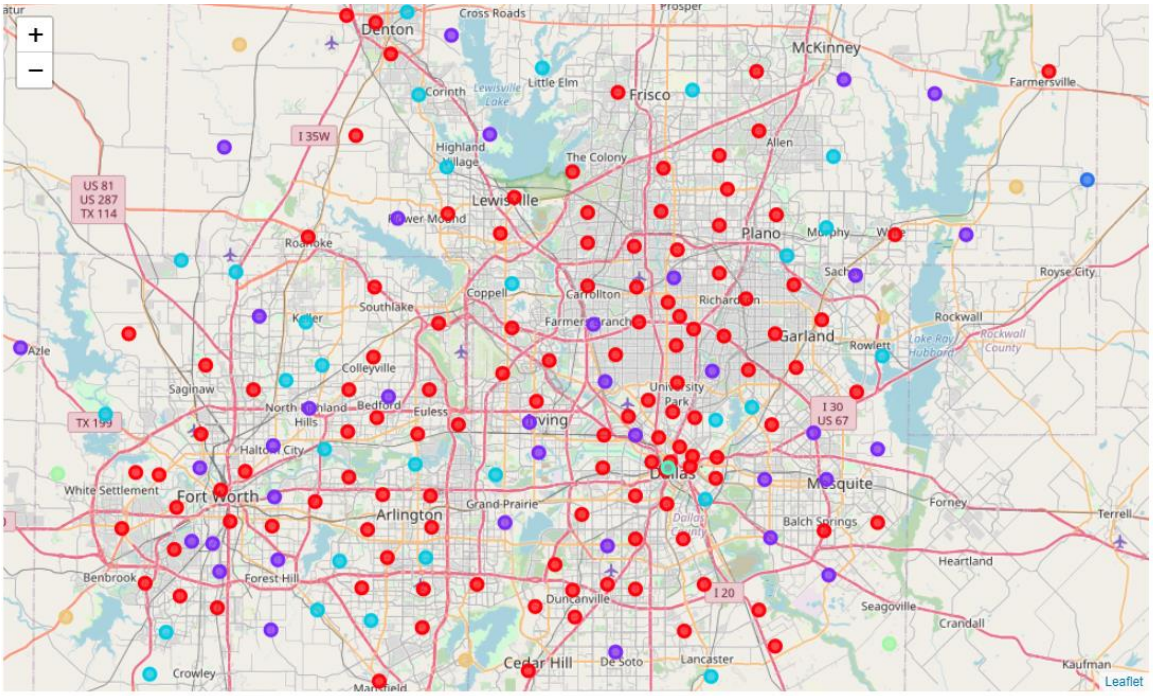
470 rows × 12 columns

Then we use the Folium library to map the clusters in each city.

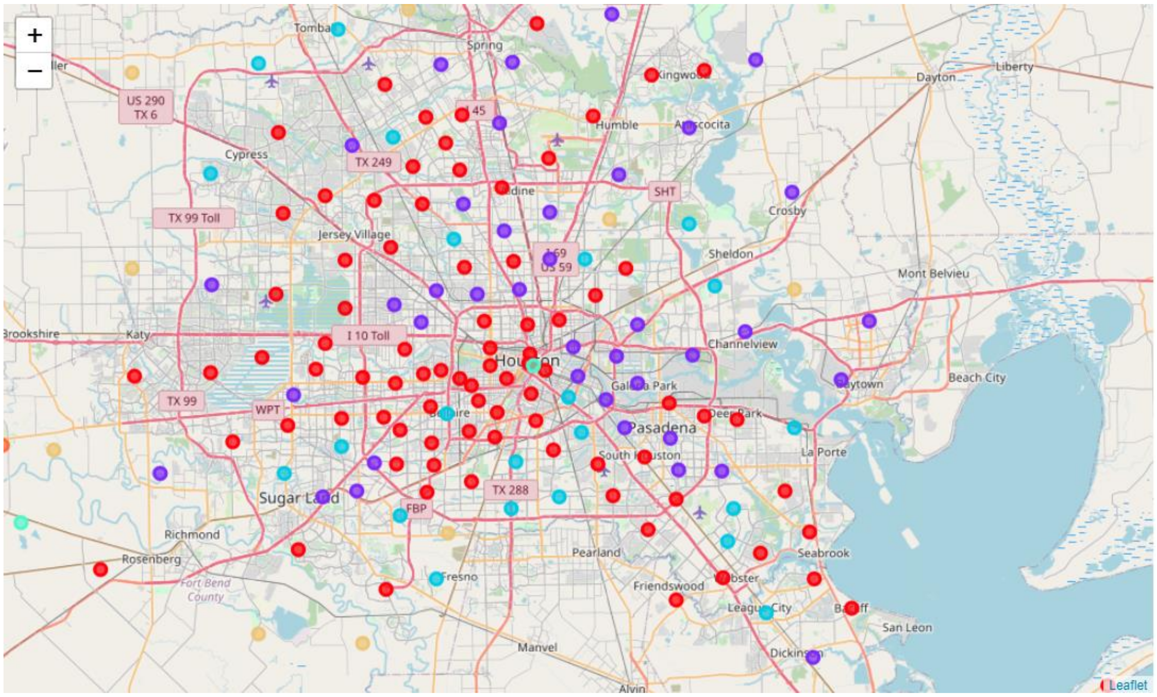
Austin clustering.



DFW clustering:



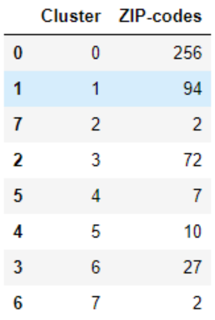
Houston clustering:



As can be seen, certain clusters are more common, while some are unique or rare. This is expected, as some zipcode areas share a common cause, e.g., housing. These neighborhoods fall into the same cluster as they share a similar venue structure. They are close to the city center.

At the same time, zipcode areas located at the center of all three cities share similar trends too. These are more venue crowded neighborhoods.

We also find that about half of the zip code areas falls into the same cluster, meaning in such areas, living facilities are relatively close, so the living standard should be very convenient alike.



	Cluster	ZIP-codes
0	0	256
1	1	94
7	2	2
2	3	72
5	4	7
4	5	10
3	6	27
6	7	2

Conclusion:

From the results and analysis, we can conclude:

1. **Austin** is the most convenient city among the three in terms of the number of most common venues in the data sets, normalized by the cities population.
2. For the two major metropolises**, DFW is slightly better than HOU**.
3. Texas people love **Mexican food**, as the number of Mexican restaurants is the most in all three cities.
4. In general, the three cities have **very similar zipcode areas/clusterings**; one should expect **roughly the same level** of living standard in any of the cities.

Thanks a lot for reading, all feedback, ideas, and comments are much appreciated.