Clustering neighborhoods in San Francisco and Houston

Arnold Jiadong Yu

1. Introduction

1.1 Background

With the development of technologies, travel became affordable and easy to access. Especially in the industry of information technologies, most work can be done remotely by only one computer. Therefore, more and more people start to travel around and experience various cultures of cities and countries while working remotely. However, a lot of researches need to be done before moving to next location to ensure best experience. As a result, it is advantageous for individuals to compare neighborhoods of various cities. This will not only save up a lot of time, but also give an initial idea of how the neighborhoods are formed. For example, this can give a person guide if he or she wants to move to Pairs for a month.

1.2 Problem

A person has enough experience of the current city, which neighborhood he likes or doesn't like. Now he wants to move to another city. Before he moves, he needs to find out how similar or dissimilar between neighborhoods in both cities.

1.3 Interest

Individuals who like traveling will definitely like this idea. Travel companies can offer intense travel plans based on this idea such as housing, transportation, and etc.

1.4 Challenges

A lot of data is needed to perform a rich and detailed comparison between two cities such as venues, real estate, population density, population variety, transportations, food variety, others' tips, recommendations, and etc. Unfortunately, it is impossible for me to obtain all information. Therefore, this project preforms a basic comparison using venues of each neighborhood.

2. Data Description

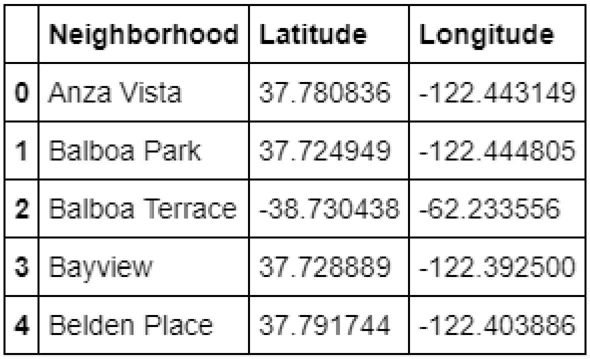
2.1 Data Source

Neighborhood data of two cities can be obtained from Wikipedia.

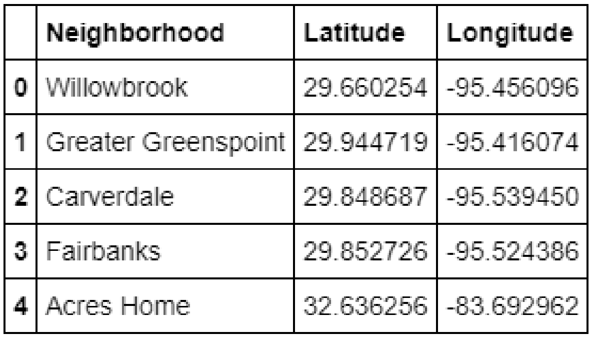
Venues of each neighborhood can be obtained using Foursquare API.

2.2 Data Cleaning

Neighborhood data need to be extracted from webpages using BeautifulSoup and put into a desirable format. There are total 123 neighborhoods extracted from San Francisco Neighborhood data and 88 neighborhoods extracted from Houston Neighborhood data. Afterwards, latitude and longitude are extracted of each neighborhood by using geopy library. Venues of each neighborhood are collected using its latitude and longitude. A dataframe is built using all venues of both cities respect to neighborhoods. Since not all latitude and longitude of each neighborhood can be extracted. We end up with a total number of 132 neighborhoods with its latitude and longitude. Below are example of latitude and longitude of neighborhoods in both cities.



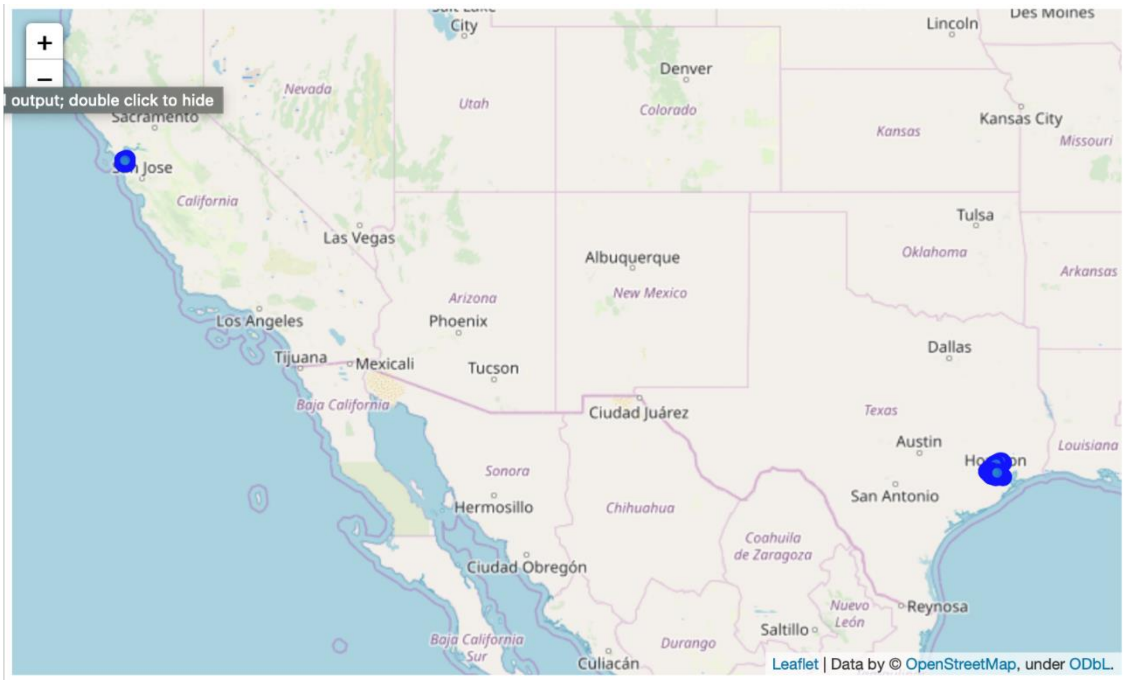
	Neighborhood	Latitude	Longitude
0	Anza Vista	37.780836	-122.443149
1	Balboa Park	37.724949	-122.444805
2	Balboa Terrace	-38.730438	-62.233556
3	Bayview	37.728889	-122.392500
4	Belden Place	37.791744	-122.403886



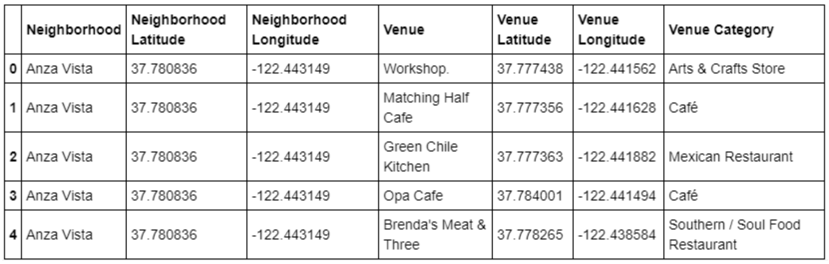
	Neighborhood	Latitude	Longitude
0	Willowbrook	29.660254	-95.456096
1	Greater Greenspoint	29.944719	-95.416074
2	Carverdale	29.848687	-95.539450
3	Fairbanks	29.852726	-95.524386
4	Acres Home	32.636256	-83.692962

2.3 Feature Selection

A filter is applied to each latitude and longitude to ensure that all the information which extracted is relevant to San Francisco and Houston. Then a map is built to ensure that the filter is successfully applied.

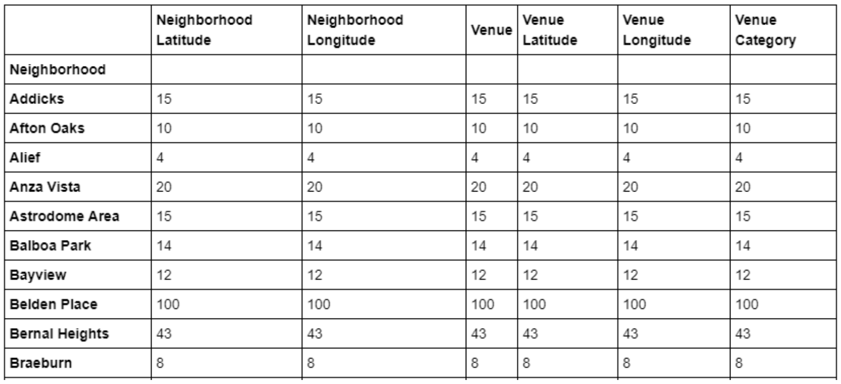


Then venues of each neighborhood are extracted using Foursquare API using latitude and longitude of each neighborhood with a radius 500 meter and limit 100 venues constraints. There are a total number of 4498 venues of all neighborhoods.



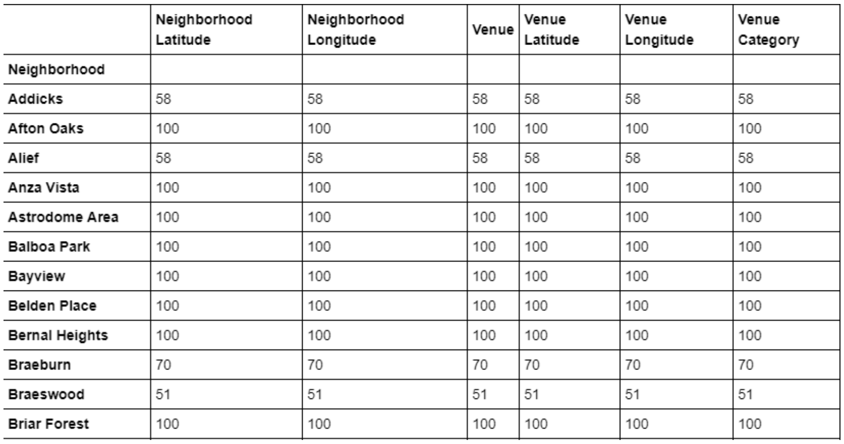
	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Anza Vista	37.780836	-122.443149	Workshop.	37.777438	-122.441562	Arts & Crafts Store
1	Anza Vista	37.780836	-122.443149	Matching Half Cafe	37.777356	-122.441628	Café
2	Anza Vista	37.780836	-122.443149	Green Chile Kitchen	37.777363	-122.441882	Mexican Restaurant
3	Anza Vista	37.780836	-122.443149	Opa Cafe	37.784001	-122.441494	Café
4	Anza Vista	37.780836	-122.443149	Brenda's Meat & Three	37.778265	-122.438584	Southern / Soul Food Restaurant

A detailed count is performed to check number of venues in each neighborhood.



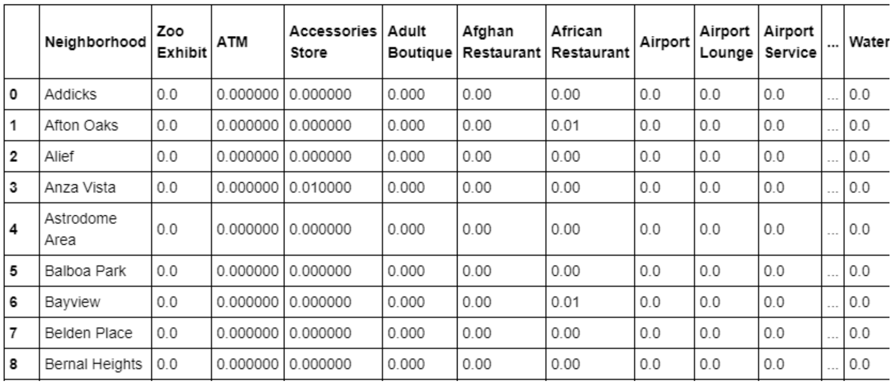
	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Neighborhood						
Addicks	15	15	15	15	15	15
Afton Oaks	10	10	10	10	10	10
Alief	4	4	4	4	4	4
Anza Vista	20	20	20	20	20	20
Astrodome Area	15	15	15	15	15	15
Balboa Park	14	14	14	14	14	14
Bayview	12	12	12	12	12	12
Belden Place	100	100	100	100	100	100
Bernal Heights	43	43	43	43	43	43
Braeburn	8	8	8	8	8	8

Since each neighborhood have a different number of venues, it is strongly biased. As a result, we need to change the parameter limit and radius to have a close number of venues in each neighborhood. Limit 100 and Radius 2000 is used. There are 10901 venues extracted and 391 unique venues.

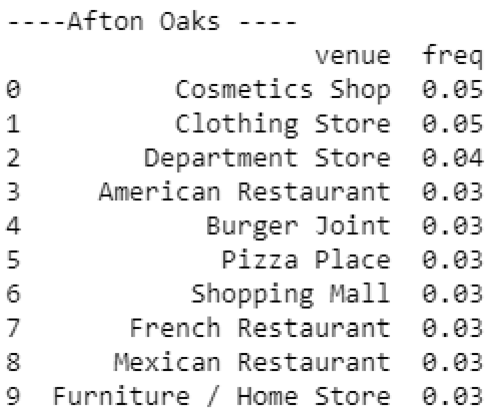


	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Neighborhood						
Addicks	58	58	58	58	58	58
Afton Oaks	100	100	100	100	100	100
Alief	58	58	58	58	58	58
Anza Vista	100	100	100	100	100	100
Astrodome Area	100	100	100	100	100	100
Balboa Park	100	100	100	100	100	100
Bayview	100	100	100	100	100	100
Belden Place	100	100	100	100	100	100
Bernal Heights	100	100	100	100	100	100
Braeburn	70	70	70	70	70	70
Braeswood	51	51	51	51	51	51
Briar Forest	100	100	100	100	100	100

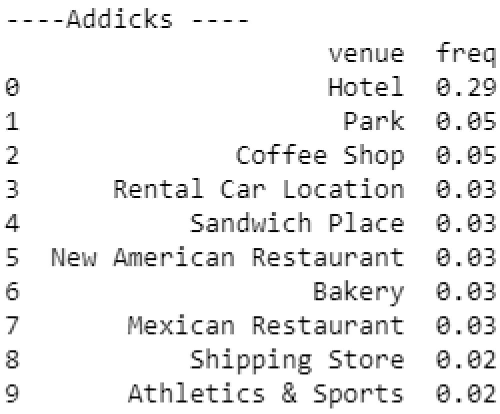
Each column is normalized afterwards based on each column. The top 10 frequencies are also displayed.



	Neighborhood	Zoo Exhibit	ATM	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	Airport	Airport Lounge	Airport Service	...	Water
0	Addicks	0.0	0.000000	0.000000	0.000	0.00	0.00	0.0	0.0	0.0	...	0.0
1	Afton Oaks	0.0	0.000000	0.000000	0.000	0.00	0.01	0.0	0.0	0.0	...	0.0
2	Alief	0.0	0.000000	0.000000	0.000	0.00	0.00	0.0	0.0	0.0	...	0.0
3	Anza Vista	0.0	0.000000	0.010000	0.000	0.00	0.00	0.0	0.0	0.0	...	0.0
4	Astrodome Area	0.0	0.000000	0.000000	0.000	0.00	0.00	0.0	0.0	0.0	...	0.0
5	Balboa Park	0.0	0.000000	0.000000	0.000	0.00	0.00	0.0	0.0	0.0	...	0.0
6	Bayview	0.0	0.000000	0.000000	0.000	0.00	0.01	0.0	0.0	0.0	...	0.0
7	Belden Place	0.0	0.000000	0.000000	0.000	0.00	0.00	0.0	0.0	0.0	...	0.0
8	Bernal Heights	0.0	0.000000	0.000000	0.000	0.00	0.00	0.0	0.0	0.0	...	0.0



```
----Afton Oaks ----
                     venue  freq
0            Cosmetics Shop  0.05
1            Clothing Store  0.05
2          Department Store  0.04
3       American Restaurant  0.03
4              Burger Joint  0.03
5              Pizza Place   0.03
6             Shopping Mall  0.03
7         French Restaurant  0.03
8        Mexican Restaurant  0.03
9  Furniture / Home Store    0.03
```

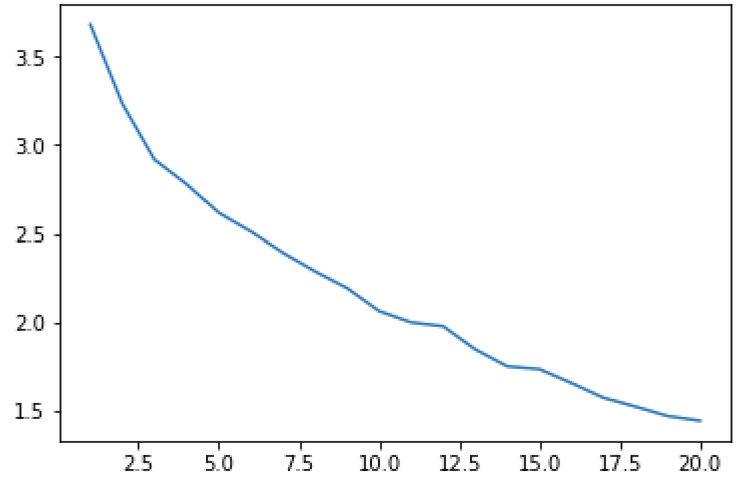


```
----Addicks ----
                      venue  freq
0                     Hotel  0.29
1                      Park  0.05
2               Coffee Shop  0.05
3       Rental Car Location  0.03
4            Sandwich Place  0.03
5  New American Restaurant   0.03
6                    Bakery  0.03
7        Mexican Restaurant  0.03
8            Shipping Store  0.02
9        Athletics & Sports  0.02
```

3. Methodology

1.1 Choose k

The Elbow method is used to choose the suitable k. A graph of k from 1 to 20 is plotted.

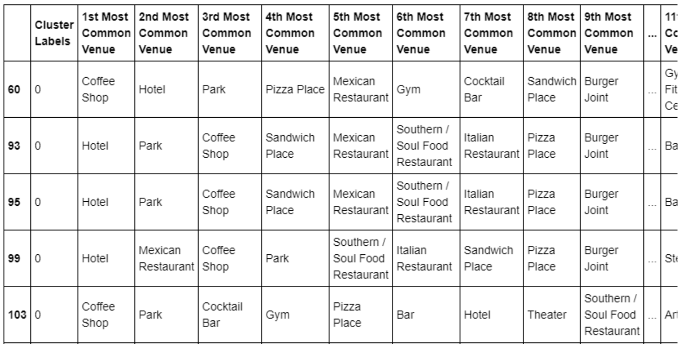


1.2 Kmean

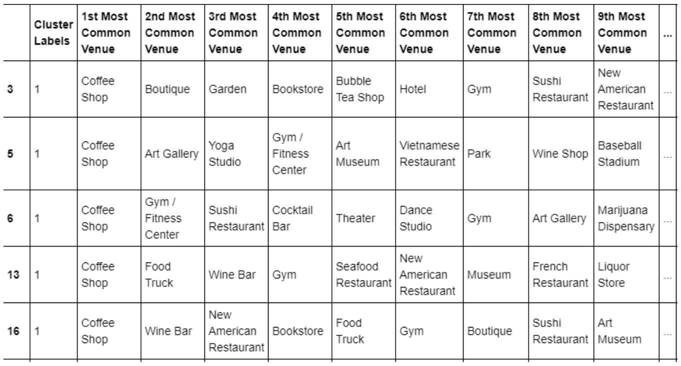
Kmean is used to cluster with k = 14.

4. Results

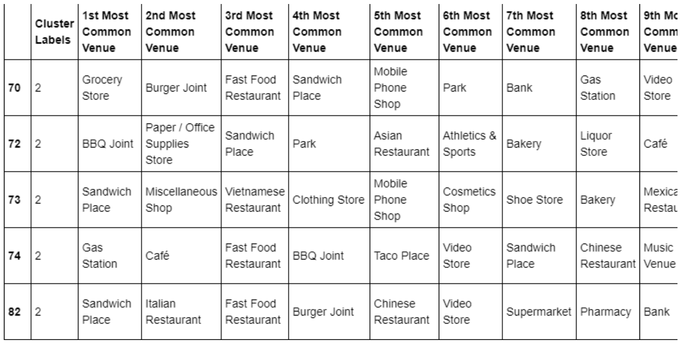
All results are printed based on their cluster. Below are clusters of 0 to 2. All clusters can be referenced to notebook on GitHub.



	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	...	11t Co Ve
60	0	Coffee Shop	Hotel	Park	Pizza Place	Mexican Restaurant	Gym	Cocktail Bar	Sandwich Place	Burger Joint	...	Gy Fit Ce
93	0	Hotel	Park	Coffee Shop	Sandwich Place	Mexican Restaurant	Southern / Soul Food Restaurant	Italian Restaurant	Pizza Place	Burger Joint	...	Ba
95	0	Hotel	Park	Coffee Shop	Sandwich Place	Mexican Restaurant	Southern / Soul Food Restaurant	Italian Restaurant	Pizza Place	Burger Joint	...	Ba
99	0	Hotel	Mexican Restaurant	Coffee Shop	Park	Southern / Soul Food Restaurant	Italian Restaurant	Sandwich Place	Pizza Place	Burger Joint	...	St
103	0	Coffee Shop	Park	Cocktail Bar	Gym	Pizza Place	Bar	Hotel	Theater	Southern / Soul Food Restaurant	...	Ar



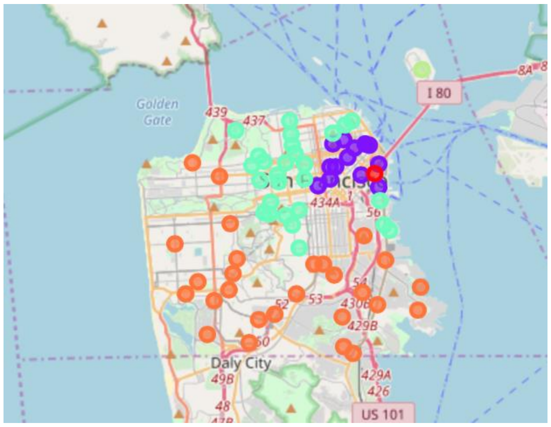
	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	...
3	1	Coffee Shop	Boutique	Garden	Bookstore	Bubble Tea Shop	Hotel	Gym	Sushi Restaurant	New American Restaurant	...
5	1	Coffee Shop	Art Gallery	Yoga Studio	Gym / Fitness Center	Art Museum	Vietnamese Restaurant	Park	Wine Shop	Baseball Stadium	...
6	1	Coffee Shop	Gym / Fitness Center	Sushi Restaurant	Cocktail Bar	Theater	Dance Studio	Gym	Art Gallery	Marijuana Dispensary	...
13	1	Coffee Shop	Food Truck	Wine Bar	Gym	Seafood Restaurant	New American Restaurant	Museum	French Restaurant	Liquor Store	...
16	1	Coffee Shop	Wine Bar	New American Restaurant	Bookstore	Food Truck	Gym	Boutique	Sushi Restaurant	Art Museum	...

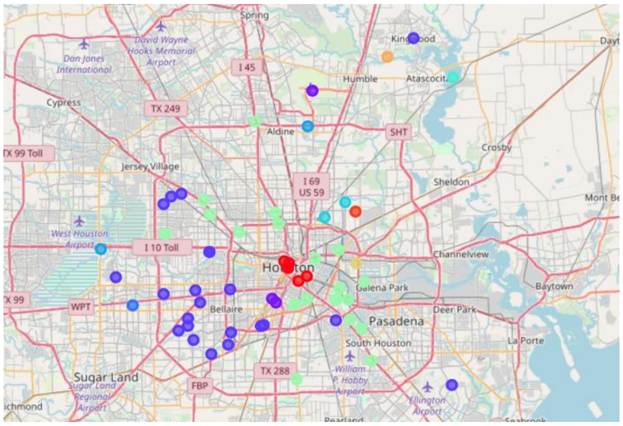


	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Mo Comm Venue
70	2	Grocery Store	Burger Joint	Fast Food Restaurant	Sandwich Place	Mobile Phone Shop	Park	Bank	Gas Station	Video Store
72	2	BBQ Joint	Paper / Office Supplies Store	Sandwich Place	Park	Asian Restaurant	Athletics & Sports	Bakery	Liquor Store	Café
73	2	Sandwich Place	Miscellaneous Shop	Vietnamese Restaurant	Clothing Store	Mobile Phone Shop	Cosmetics Shop	Shoe Store	Bakery	Mexica Restau
74	2	Gas Station	Café	Fast Food Restaurant	BBQ Joint	Taco Place	Video Store	Sandwich Place	Chinese Restaurant	Music Venue
82	2	Sandwich Place	Italian Restaurant	Fast Food Restaurant	Burger Joint	Chinese Restaurant	Video Store	Supermarket	Pharmacy	Bank

5. Conclusion

As a result, the map showed clusters in both cities and can offer an idea of how similar the neighborhoods are.





6. Future Direction

A better cluster can be done by using more detailed data of each neighborhood such as occurrence of crimes.

7. Reference

1. Beautiful Soup Documentation. https://beautiful-soup-4.readthedocs.io/en/latest/#kinds-of-objects

2. geopy Documentation. https://pypi.org/project/geopy/

3. Foursquare API Documentation. https://developer.foursquare.com/docs/api/endpoints

4. List of neighborhoods in San Francisco. https://en.wikipedia.org/wiki/List_of_neighborhoods_in_San_Francisco

5. List of neighborhoods in Houston. https://en.wikipedia.org/wiki/List_of_Houston_neighborhoods