Comparison of two cities: Houston and Chicago

Yuan Zhang

1. Introduction

1.1 Background

My husband and I lived in Houston for almost 7 years. My husband has recently got a job offer in Chicago, and we plan to move to Chicago by the end of September. We are familiar with the climate, food, museums, sports and all the fun in Houston, which can be regarded as our second hometown in the world. On the other hand, we have never been to Chicago and the only impression for Chicago is a global, multicultural, and cosmopolitan city.

1.2 Problem

My goal is to perform a comparison of the two cities to see how similar or dissimilar they are. Such techniques allow users to identify similar neighborhoods among cities based on amenities or services being offered locally, and thus can help in understanding the local area activities.

1.3 Interest

There will be other job seekers like us would be interested in the difference between two cities, this will help them to make a better decision when they move to a new city. There will be some company would like to expand their business, they would benefit from such analysis. Any people who live in these two cities can also refresh their ideas by checking one city analysis. Such analysis tool is very extensible to other cities comparison, which can give us a basic idea when we want to compare two cities in the world.

2. Data Preparation

2.1 City neighborhoods data from Wikipedia

City data was extracted from the respective Wikipedia pages using Request and Beautifulsoup libraries in Python. There is only neighborhoods data in Wikipedia page, we need to first extract the neighborhoods data and get the coordinate from geopy.

2.2 Weather and climate data from https://www.timeanddate.com/weather/usa/

We can search for different city crime rate and types from this website and use Beautifulsoup to scrape the data. The research here is limited to the most recent whole year.

2.3 City crime data is from https://www.neighborhoodscout.com/il/

We can search for different city crime rate and types from this website and use Beautifulsoup to scrape the data. The research here is limited to the most recent whole year.

2.4 Foursquare location data

Foursquare is a social location service that allows users to explore the world around them. We took advantage of Foursquare data to extract venue based information for all of the neighborhoods under within two cities. The call to the API returns a JSON file and we need to turn that into pandas data-frame.

3. Methodology

3.1 Web scraping

The incredible amount of data on internet is a rich resource for any field of research or personal interest. The Python libraries requests, and Beautiful Soup are powerful tools to get useful data from the website. Please see the codes below for extracting neighborhood data from Wikipedia.

```
url_houston = "https://en.wikipedia.org/wiki/List_of_neighborhoods_in_Houston"
req_houston = requests.get(url_houston)
```

```
soup_houston = BeautifulSoup(req_houston.content)
```

By analyzing the html format and using the python loop function, we can put the data from website to a pandas data frame. Of course, some simple data cleaning is needed. And by using the geopy package, we will be able to find each neighborhood latitude and longitude. The dataframes generated for Houston and Chicago are shown as below.

	Neighborhoods	Latitude	Longitude
0	Willowbrook, Houston, TX	29.660254	-95.456096
1	Greater Greenspoint, Houston, TX	29.944719	-95.416074
2	Carverdale, Houston, TX	29.848687	-95.539450
3	Fairbanks / Northwest Crossing, Houston, TX	29.852726	-95.524386
4	Acres Home, Houston, TX	29.854948	-95.446884
5	Hidden Valley, Houston, TX	29.892914	-95.418131
6	Westbranch, Houston, TX	29.839541	-95.551522
7	Addicks / Park Ten, Houston, TX	29.788404	-95.644854
8	Spring Branch West, Houston, TX	29.801410	-95.547879
9	Langwood, Houston, TX	29.826282	-95.482795

	Neighborhoods	Latitude	Longitude
0	Albany Park, Chicago, IL	41.970329	-87.715992
1	Altgeld Gardens, Chicago, IL	41.655259	-87.609584
2	Andersonville, Chicago, IL	41.977139	-87.669273
3	Archer Heights, Chicago, IL	41.811422	-87.726165
4	Armour Square, Chicago, IL	41.840231	-87.632986
5	Ashburn, Chicago, IL	41.747533	-87.711163
6	Auburn Gresham, Chicago, IL	41.743167	-87.653728
7	Avalon Park, Chicago, IL	41.745035	-87.588658
8	Avondale, Chicago, IL	41.938921	-87.711168
9	Avondale Gardens, Chicago, IL	41.938921	-87.711168

3.2 Foursquare API

After registering for in the foursquare developer tab, we can easily use Foursquare API to get the information we need. The call to the API returns a JSON file and we need to turn that into a data-frame. The venue name and category are the information we need. To have a fair result for later analysis, we usually sort the neighborhoods by the number of venues they have. I decided to pick 100 most popular spots in each neighborhood within a radius of 500m.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Willowbrook, Houston, TX	29.660254	-95.456096	Ultimate Radiant Barrier & Insulation	29.656731	-95.455377	Furniture / Home Store
1	Willowbrook, Houston, TX	29.660254	-95.456096	Drum Teacher	29.658871	-95.460107	Performing Arts Venue
2	Greater Greenspoint, Houston, TX	29.944719	-95.416074	Carnival @ Greenspoint	29.944084	-95.414857	Theme Park
3	Greater Greenspoint, Houston, TX	29.944719	-95.416074	GNC	29.945597	-95.411736	Supplement Shop
4	Greater Greenspoint, Houston, TX	29.944719	-95.416074	Finish Line	29.946231	-95.411861	Shoe Store

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Albany Park, Chicago, IL	41.970329	-87.715992	Nighthawk	41.967974	-87.713415	Cocktail Bar
1	Albany Park, Chicago, IL	41.970329	-87.715992	Chicago Produce	41.970553	-87.716327	Grocery Store
2	Albany Park, Chicago, IL	41.970329	-87.715992	Hot Dog Station	41.967880	-87.713404	Hot Dog Joint
3	Albany Park, Chicago, IL	41.970329	-87.715992	Peking Mandarin Restaurant	41.968292	-87.715783	Chinese Restaurant
4	Albany Park, Chicago, IL	41.970329	-87.715992	Great Sea Chinese Restaurant	41.968496	-87.710678	Chinese Restaurant

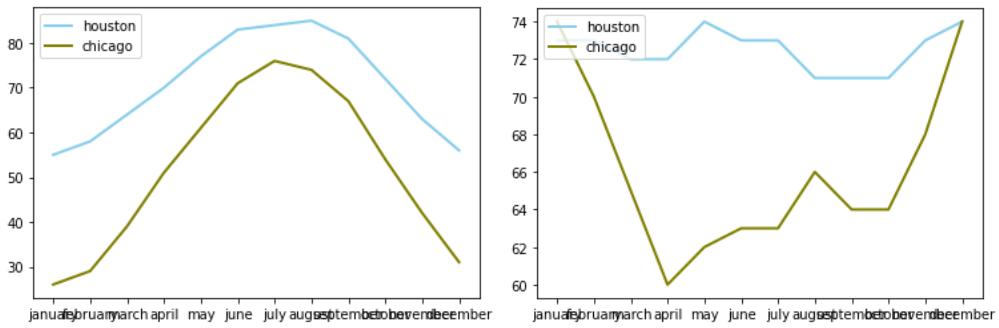
3.3 K-means

I performed a clustering analysis using the '*k-means*' algorithm in order to categorize similar neighborhoods into clusters based on the similarities provided by the venue categories. To perform such analysis, we need to first calculate the frequency of each venue category in the neighborhood lists and then cluster the neighborhoods by the frequency. We may need to test the proper number of clusters to have a meaningful analysis.

4. Exploratory Data Analysis and Discussion

4.1 Weather data

Not everyone can handle bitterly cold weather. On the other hand, not everyone prefers year-round sunshine. We want to make sure the climate is reasonable before we move to Chicago. I scraped some history data from **https://www.timeanddate.com/weather/usa/**, which allows me to choose different cities and make comparison between them. The temperature and humidity data are shown as below.



We can conclude that Houston is overall warmer than Chicago, the temperature difference in Houston among 4 seasons is much smaller than Chicago. I already got used to the warm weather in Houston since I have been here for many years, but I do not worry about the cold weather in Chicago, since I was born in the North China, which has similar climate to New York. As for the right figure, it shows the humidity comparison. It is very obvious that Houston is very humid through the whole year, while Chicago shows a much lower humidity, especially in Summer. I am happy about this change, since I do not really like humidity in Houston, which is too much for me. It always feels uncomfortably sweaty and buggy during the long summer in Houston. All in all, I would say I am optimistic about the climate change from Houston to Chicago.

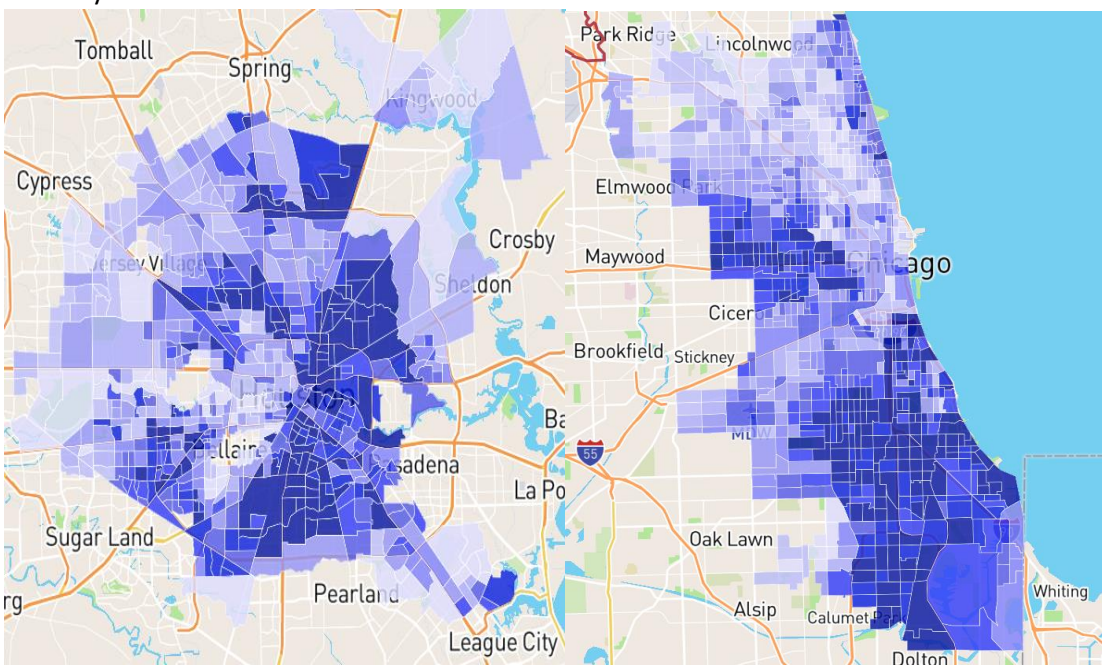
4.2 Crime data

Unfortunately, crime in cities is unavoidable. No matter where you move, there will likely be some amount of crime. The question is: are crime rates high enough that you should be concerned? Here I investigate a city's crime rates prior to moving. The data is got from

https://www.neighborhoodscout.com/il/.

City	Murder	Rape	Robbery	Assault
Houston	0.12	0.57	3.97	6.35
Chicago	0.18	0.66	2.96	5.68

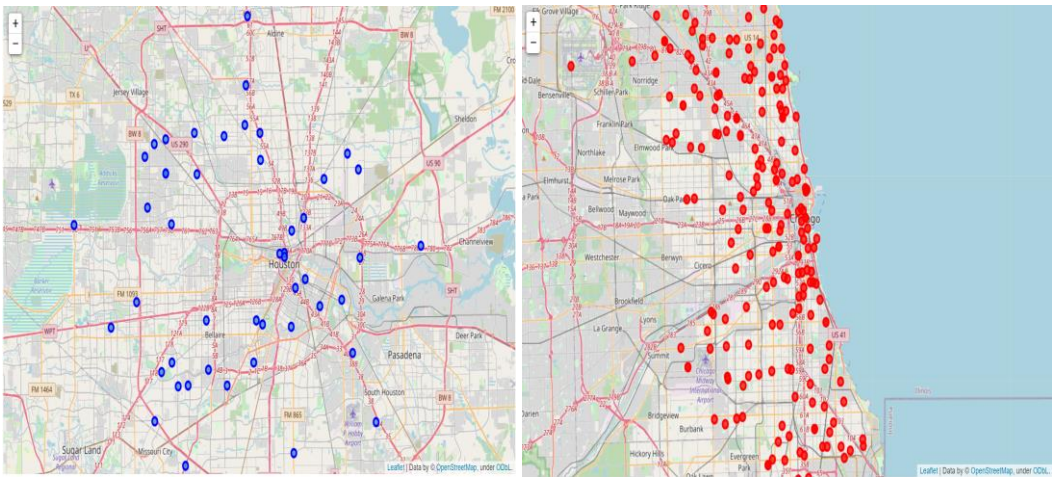
We can clearly see that Houston has less murder and rape, while it suffers more from Robbery and Assault.



This figure shows that Chicago and Houston similar level of crime rate and North and West part of two cities are relatively safe. So we should choose the apartments in the safe region when we move there.

4.3 Neighborhood data

There are 57 neighborhoods in Houston and 233 neighborhoods in Chicago. Houston has 2.31 million population while Chicago has 2.71 million population. Here I show the maps with neighborhood center being highlighted. We can see Chicago is denser while Houston is more spread-out. The neighborhood distribution map is shown below.



4.4 Venue Data

We obtained the data from Foursquare. The top five venues in Houston downtown are concert hall, park, theatre, performing arts venue, and coffee shop. I am not surprised by this list. When I was a student in Rice University, I went to the concert hall which is famous for Houston symphony every weekend with my student coupon. For the people who love music, please do not miss it. And surely there are some cute parks located in the downtown area. The one I like most is the Discovery Green. I hit this park every week to collect pokemon when the game is popular.

```
----Downtown, Houston, TX----
                   venue  freq
0           Concert Hall  0.09
1                   Park  0.09
2                Theater  0.09
3  Performing Arts Venue  0.06
4             Coffee Shop  0.06
```

Then I check how Chicago downtown venues look like. I am a little surprised that the top five venues are Circus, American restaurant, Harbor/Marina, Train Station and Hill. We can clearly see the downtown of Chicago is more friendly to Tourists. Watching fantastic shows in Circus, sightseeing the harbor and trying some traditional American food are all typical tourists' activities. I am looking forward to moving to Chicago and hit downtown area now.

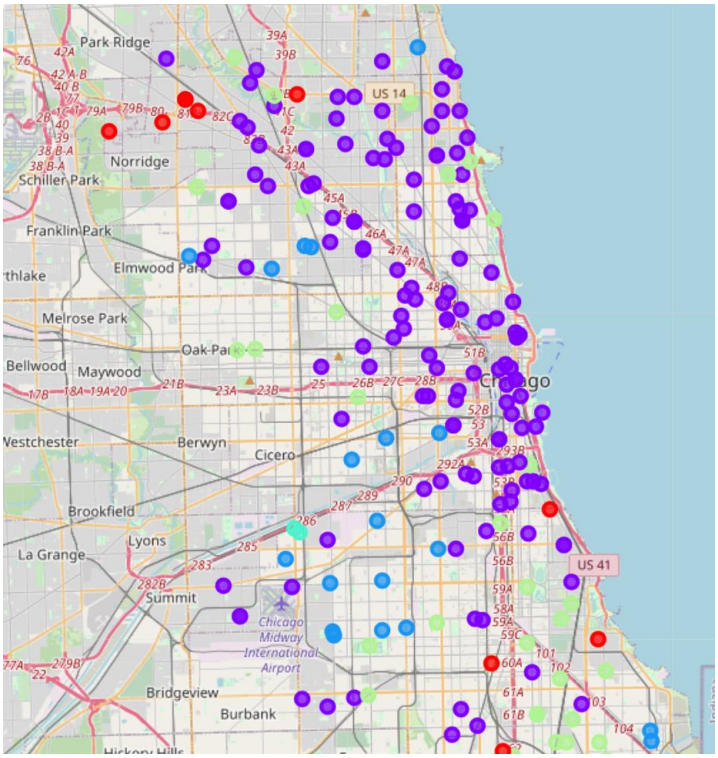
```
----River North, Chicago, IL----
                   venue  freq
0                 Circus  0.11
1  American Restaurant  0.11
2      Harbor / Marina  0.11
3        Train Station  0.11
4                 Hill  0.11
```

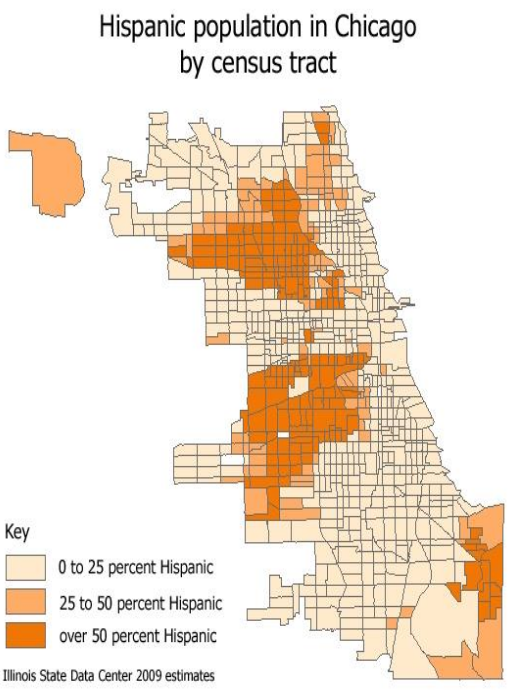
Before we did the K-means analysis, we sort the frequency of each venues for all the neighborhoods in Houston and Chicago. The tables are shown below.

	Neighborhood	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	Acres Home, Houston, TX	Rental Service	Home Service	Garden Center	Electronics Store	Food	Flower Shop	Financial or Legal Service	Fast Food Restaurant	Farmers Market	Eye Doctor
1	Addicks / Park Ten, Houston, TX	Hotel	Indian Restaurant	Steakhouse	Coffee Shop	Sushi Restaurant	New American Restaurant	Motel	Residential Building (Apartment / Condo)	Motorcycle Shop	Shipping Store
2	Alief, Houston, TX	Pool	Football Stadium	Playground	Shoe Repair	Zoo Exhibit	Electronics Store	Flower Shop	Financial or Legal Service	Fast Food Restaurant	Farmers Market
3	Astrodome Area, Houston, TX	Bar	Football Stadium	BBQ Joint	General Entertainment	Sports Bar	Gift Shop	Speakeasy	Coffee Shop	Light Rail Station	Theme Park Ride / Attraction
4	Braeburn, Houston, TX	Bar	Pharmacy	Food Service	Tennis Court	Food	Food Truck	Bike Trail	Donut Shop	Dry Cleaner	Duty-free Shop

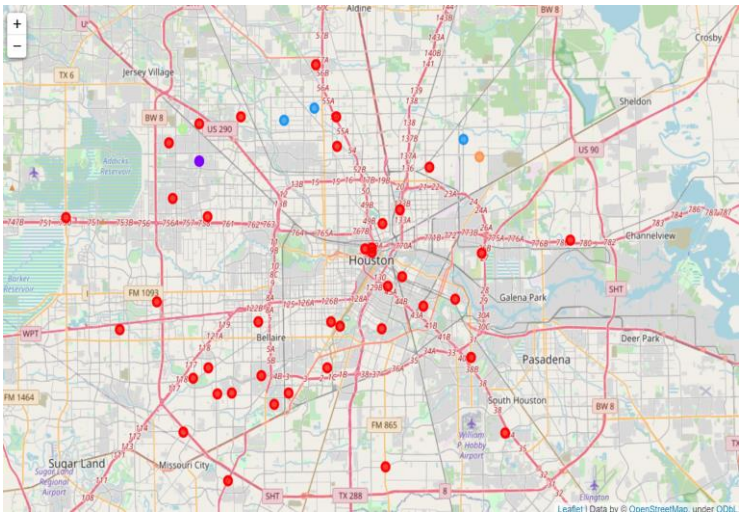
	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Albany Park, Chicago, IL	Mexican Restaurant	Chinese Restaurant	Grocery Store	Mobile Phone Shop	Accessories Store	Bus Station	Sandwich Place	Train Station	Thrift / Vintage Store	Karaoke Bar
1	Altgeld Gardens, Chicago, IL	Grocery Store	Park	Zoo Exhibit	Eastern European Restaurant	Elementary School	English Restaurant	Escape Room	Ethiopian Restaurant	Event Space	Exhibit
2	Andersonville, Chicago, IL	Italian Restaurant	Coffee Shop	Lounge	Breakfast Spot	Pet Store	Bar	Pizza Place	Beer Bar	Flower Shop	Miscellaneous Shop
3	Archer Heights, Chicago, IL	Mexican Restaurant	Mobile Phone Shop	Grocery Store	Big Box Store	Gas Station	Bank	Bar	Gym / Fitness Center	Sandwich Place	Coffee Shop
4	Armour Square, Chicago, IL	Chinese Restaurant	Ice Cream Shop	Italian Restaurant	Gas Station	Sandwich Place	Asian Restaurant	Grocery Store	Cosmetics Shop	Hot Dog Joint	Currency Exchange

Let us first check how Chicago looks like with Kmeans analysis. It takes some time to decide the proper K value, we used 6 at last have a reasonable result. We can see lots of the neighborhoods belong to the same category (purple dots), the most common venues of such places are sandwich place, restaurants, hotels and coffee shops. As we go from the center to the outside of Chicago, we see more blue and red dots shows up. For blue region, the most common venue is the Mexican restaurant, and I am somewhat surprised by this result. With the immigration of thousands of Mexicans into Texas cities, cultures were bound to clash and mix, so there are lots of Mexican restaurants in Houston, and the flavor" Tex-mex" is very popular. I have not thought there are so many Mexican restaurants in Chicago. I also compare this result with the Latino population distribution map, and I see some similarity. The blue dots region can overlay with high density Latino population, especially the northwest and southwest wings. As for the red cluster, we can see lots of parks show up. Usually, the parks will be more in suburb area.





Then we can investigate how Houston behaves. This machine learning analysis of the venue-based data revealed most of the neighborhood in Houston can be grouped together into one cluster. The most common venues were always stadiums, cafes, bars, hotels or restaurants.



We can use the venue data to compare the cities is a more comprehensive way where one can also explore different levels of information. With more delicate analysis, we can answer questions like which city has more restaurants? Is a particular neighborhood more commercial or residential?

5. Conclusion

With this comparison, we can conclude that Chicago is a modern, lively and populated city, just like Houston. It has more venue categories, more restaurants, hotels, and bars, so it looks like more commercial than Houston. It should be friendly to tourists and young people. It has similar level of crime rate as Houston, so I would never suggest people walk outside during night. And honestly, I do not know which one is tougher? Houston's summer or Chicago's winter?

The capstone project provided a medium understanding in depth about how real life data science projects work. All steps from understanding the business problem, data understanding to data preparation, and model building were discussed in detail here. I understand there is still many drawbacks in current analysis, but such work has a great potential to solve the problem in real life scenarios.

No matter how Chicago compares with Houston, we probably already decided to move to this new city. I have more confidence that I will like Chicago in the future now.