Investigation of Houston and New York

— A City Preview for City Migrant and Business Investor

1. Introduction

1.1 Background

Each year, roughly 40 million Americans, or about 14 percent of the U.S. population, move at least once [1]. Young people want to move to location with better job opportunity while elders may want to move to location with warm climate. But there are many aspects that will affect living experience in a new city. And people also have their own life style. The new city is not always as we expected. To help people to make moving decision, it is important to help them find out if the living experience matches their own life style, which means helping people "see" the new city before they actually sell their stuff, pack their furniture and drive thousands miles there.

1.2 Problem

People like to compare current living cities to the perspective one. The problem is how the city will affect the daily life. Of course, people's preference could include weather, population and future living expense, which can be found easily with google. Another question cannot be easily answered is how the city looks like in terms of density of food shops or other types of store your preferred and distribution of stores or other types of venues over the whole city. This question is actually affecting your daily life. For example, Italian food lover moved to a city with barely any Italian Restaurant and each of them is 20mils apart, which is the exact scenario I am helping people to avoid with city venue investigation in this report. With limited access data and available time, I redefine the problem to comparing venue distribution between two biggest cities in USA, New York and Houston.

1.3 Interest

There are millions of people moving every week, and millions small business starting every week. This report can help them with understanding the cities they are going to live and work in. It can save people's money by avoiding decision mistake.

[1]https://www.usatoday.com/story/money/economy/2018/07/05/cities-americans-growing-population-migration/35801343/

2. Data acquisition and cleaning

2.1 Data sources

Data from two cities, New York and Houston, is used in this report.

The neighborhood data for New York is downloaded from website https://geo.nyu.edu/catalog/nyu_2451_34572, which is json file and I extract all the neighborhoods, their latitude and longitude information from it. For Houston, the neighborhood name is scraped from Wikipedia page. The latitude and longitude are converted from neighborhood using geopy module in python.

The venue data of two cities is provided in json through API of Foursquare website. The venue data includes much information, such as venue name, location, category, menus, users' tips, hours and etc. In this study, I used venue name, category, latitude, longitude and neighborhood, which I selected from full json files. The data will be used in clustering cities' neighborhoods based on category of venues.

The neighborhood and venue data cover all the area in Houston and New York. For Houston, the number of neighborhoods is 114 and for New York, the number of neighborhoods is 306. The number of venues is over 10000 in Houston and over 30000 in New York.

2.2 Data cleaning

The data for New York is pretty clean. Here I only need do data cleaning on Houston data. The neighborhood name of Houston is from Wikipedia. But after conversion to latitude and longitude, I found geopy output some wrong values. There are several types of mistakes: 1. Out of Houston range; 2. Longitude is positive value instead of negative; 3. Some latitude and longitude values are string not float data. For those mistakes, I manually correct those mistakes.

2.2 Data Feature Group

The venues from Foursquare have about 300 categories, for example, for shops related to food, it has over 50 categories, such as Ice Cream Shop, Bubble Tea shop, Italian Restaurant and etc. While listing very specific category could be really handy for studying specific category, it is hard to use them to represent an overall impression of the city. So I create a few super-categories. Below is the table for relation of super-categories and categories.

Venue Super Category	Venue Category (Foursquare API)
Food and Drink	afghan restaurant,african restaurant,airport food court,american restaurant,arepa restaurant,argentinian restaurant,asian restaurant,australian restaurant,austrian restaurant,bbq joint,bagel shop,bakery,bar,beach bar,beer bar,beer garden,beer store,bistro,brazilian restaurant,breakfast spot,brewery,bubble tea shop,burger joint,burmese restaurant,burrito place,café,cajun / creole restaurant,candy store,cantonese restaurant,caribbean restaurant,cheese shop,chinese restaurant,chocolate shop,churrascaria,cocktail bar,coffee shop,colombian restaurant,comfort food restaurant,creperie,cuban restaurant,cupcake shop,deli / bodega,dessert shop,dim sum restaurant,diner,distillery,dive bar,donut shop,dumpling restaurant,eastern european restaurant,empanada restaurant,ethiopian restaurant,falafel restaurant,fast food restaurant,filipino restaurant,fish & chips shop,fondue restaurant,food,food & drink shop,food court,food truck,french restaurant,fried chicken joint,gaming cafe,gastropub,gay bar,german restaurant,gluten-free restaurant,greek restaurant,hawaiian restaurant,health food store,himalayan restaurant,hookah bar,hot dog joint,hotel bar,hotpot restaurant,hunan restaurant,ice cream shop,indian chinese restaurant,indian restaurant,indonesian restaurant,irish pub,israeli restaurant,italian restaurant,japanese restaurant,jewish

	restaurant,juice bar,karaoke bar,kofte place,korean restaurant,kosher restaurant,latin american restaurant,lebanese restaurant,mac & cheese joint,malay restaurant,mediterranean restaurant,mexican restaurant,middle eastern restaurant,modern european restaurant,modern greek restaurant,molecular gastronomy restaurant,mongolian restaurant,moroccan restaurant,new american restaurant,noodle house,paella restaurant,pakistani restaurant,persian restaurant,peruvian restaurant,pizza place,polish restaurant,portuguese restaurant,pub,public art,ramen restaurant,restaurant,russian restaurant,sake bar,salad place,salon / barbershop,sandwich place,seafood restaurant,shabu-shabu restaurant,shanghai restaurant,snack place,soba restaurant,soup place,south american restaurant,southern / soul food restaurant,spanish restaurant,sports bar,sri lankan restaurant,steakhouse,street food gathering,sushi restaurant,szechuan restaurant,taco place,taiwanese restaurant,tapas restaurant,tea room,tex-mex restaurant,thai restaurant,tibetan restaurant,tiki bar,turkish restaurant,vegetarian / vegan restaurant,venezuelan restaurant,veterinarian,vietnamese restaurant,whisky bar,wine bar,wine shop,wings joint,smoothie shop
Daily Essential	accessories store,animal shelter,automotive shop,bank,big box store,board shop,butcher,cemetery,church,comic shop,convenience store,discount store,dog run,dry cleaner,duty-free shop,electronics store,eye doctor,fabric shop,fish market,flea market,flower shop,frozen yogurt shop,fruit & vegetable store,furniture / home store,gift shop,gourmet shop,grocery store,hardware store,herbs & spices store,hobby shop,home service,kids store,kitchen supply store,laundry service,library,lingerie store,liquor store,market,mattress store,miscellaneous shop,mobile phone shop,neighborhood,optical shop,other repair shop,paper / office supplies store,pet service,pet store,pharmacy,pie shop,record shop,shipping store,shoe repair,shoe store,shop & service,shopping mall,smoke shop ,souvenir shop,supermarket,warehouse store
Fashion	antique shop,boutique,clothing store,cosmetics shop,department store,design studio,event space,general entertainment,government building,health & beauty service,jewelry store,massage studio,men's store,nail salon,residential building (apartment / condo),spa,supplement shop,tanning salon,tattoo parlor,thrift / vintage store,women's store
Education	college academic building,college rec center,high school,school, elementary school
Entertainment	aquarium,arcade,art gallery,art museum,arts & crafts store,bookstore,castle,circus,comedy club,concert hall,gun range,gun shop,historic site,history museum,indie movie theater,indie theater,jazz club,martial arts dojo,movie theater,museum,music store,music venue,nightclub,nightlife spot,opera house,other nightlife,performing arts venue,planetarium,science museum,social club,street art,theater,toy / game store,used bookstore,video game store,video store,zoo,zoo exhibit
Indoor Recreation	athletics & sports,bowling alley,boxing gym,climbing gym,cycle studio,dance studio,gym,gym / fitness center,gymnastics gym,indoor play area,motorcycle shop,pilates studio,recreation center,rock club,skating rink,sporting goods shop,sports club,tennis court,weight loss center,yoga studio
Outdoor Related	baseball field,baseball stadium,basketball court,basketball stadium,beach,bike shop,boat or ferry,botanical garden,bridge,campground,college baseball diamond,college basketball court,farm,farmers market,field,football stadium,fountain,garden,garden center,golf course,golf driving range,island,lake,mini golf,monument / landmark,national park,other great outdoors,outdoor sculpture,outdoors & recreation,paintball field,park,playground,plaza,pool,pool hall,racetrack,sculpture garden,shopping plaza,skate park,soccer field,soccer stadium,state / provincial park,surf spot,tennis stadium,theme park,theme park ride / attraction,toll plaza,track stadium,volleyball
Inner-City Transportation	bus station,bus stop,gas station,harbor / marina,pier,stationery store,trail,train station

Inter-City Transportation	airport lounge,airport service

3. Exploratory Data Analysis

3.1 Mapview of neighborhoods

City of Houston has only 113 neighborhoods inside about 600 square miles and New York City has 306 neighborhoods inside about 300 square miles. We are expecting the less dense neighborhood in Houston. The maps below show neighborhoods distribution of both cities.

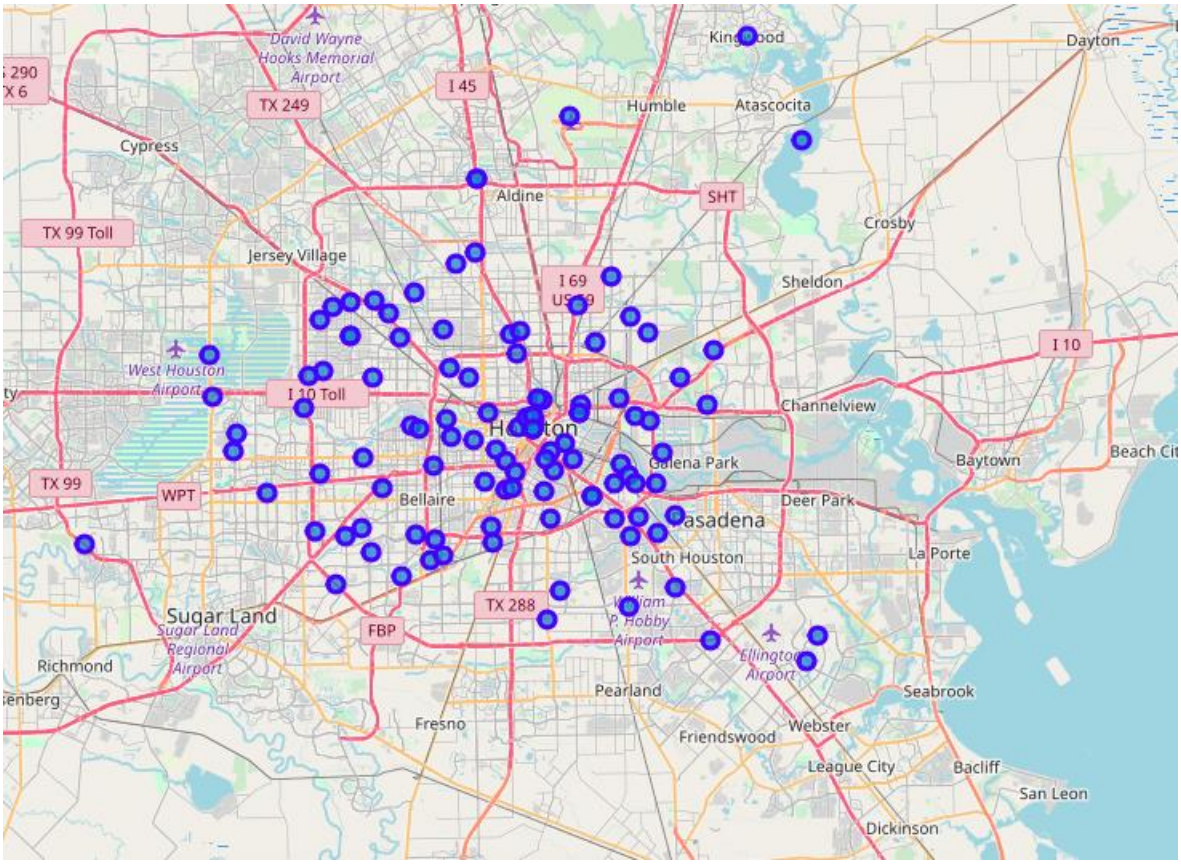

Figure 1: The neighborhood overview on Houston map

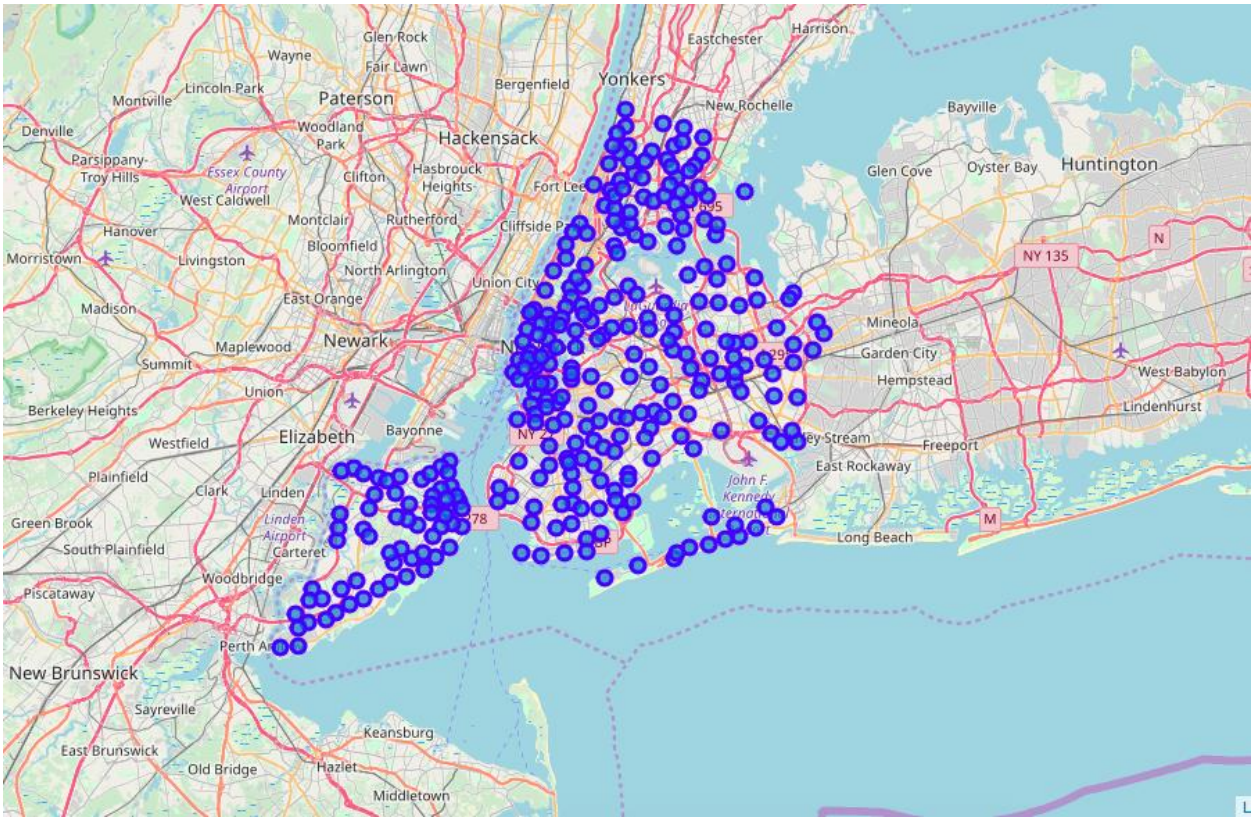

Figure 2: The neighborhood overview on New York city map

As we can see from the figures above, the New York city has much denser neighborhood than Houston.

3.2 Mapview of venues in neighborhoods

The population of New York City is about 8 millions while the population of Houston is about 2 millions. The venues cover all the locations that are related to people's everyday life, such as restaurant or supermarket. We are not surprised to see more venues in New York (over 30000) than in Houston (over10000). The maps below show venues distribution of both cities. There density of venues in New York are much higher than in Houston.

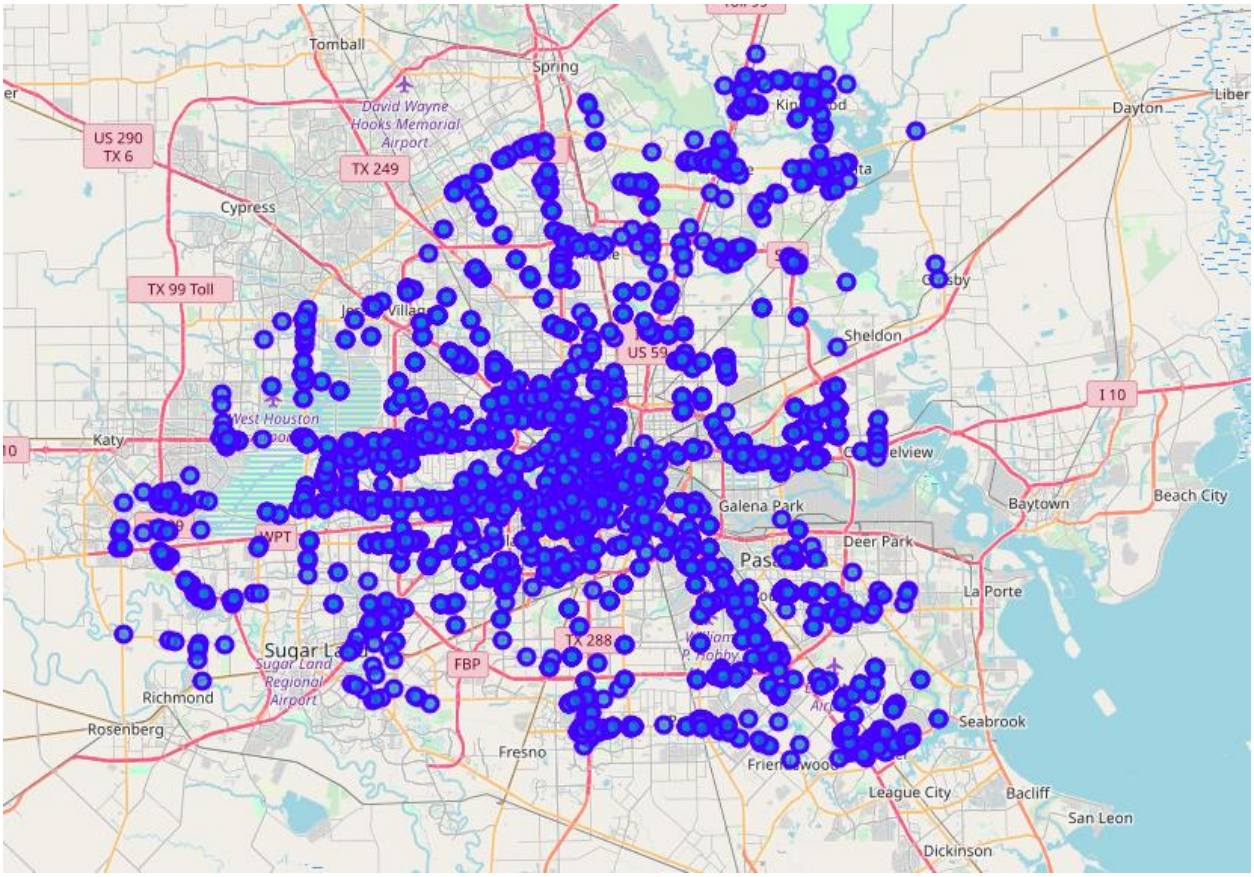

Figure 3: Venues locations on Houston map

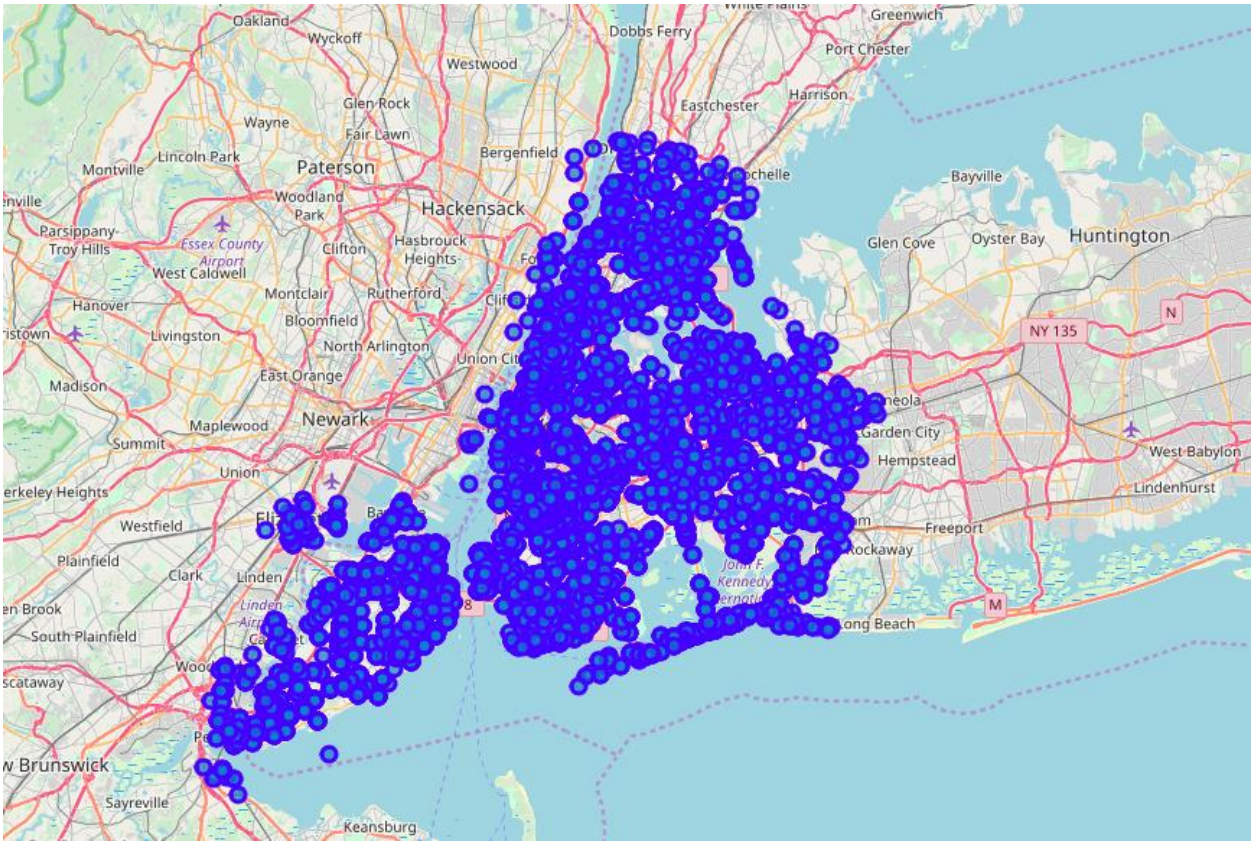

Figure 4: Venues location on New York map

3.3 Comparison of Percentage of Venues in each super category in Houston and New York City

Table 1 is the direct comparison of percentage of venues in each super category in Houston and New York City. Food and Drink category takes more than 60% of total venues, which is reasonable since people rely on food and both cities are highly populated. They both have Inter City Transportation, which is also basic venue type for a big international city. While other venue super categories have similar venue percentage, the Inner City Transportation in New York has 2 times venues of in Houston.

From the pie chart, we can visually see the percentage of venues more clearly. The two cities are very similar in terms of distribution of venue super category. We can guess that this is because they both are big and mature cities. Every aspect of people's lives is covered with enough venues.

Venue Super Category	Houston super category percentage	New York super category percentage
Daily Essential	10.28617	8.430337
Education	0.00886	0.091812
Entertainment	5.696819	7.089222
Fashion	2.587047	3.026527
Food Drink	68.47701	62.73732
Indoor Workout	3.18951	4.370922

Inner City Transportation	1.612475	0.888612
Inter City Transportation	0.159476	0.180346
Others	2.152919	2.331377

Table 1 Comparison of percentage of venues in each super category for Houston and New York City.

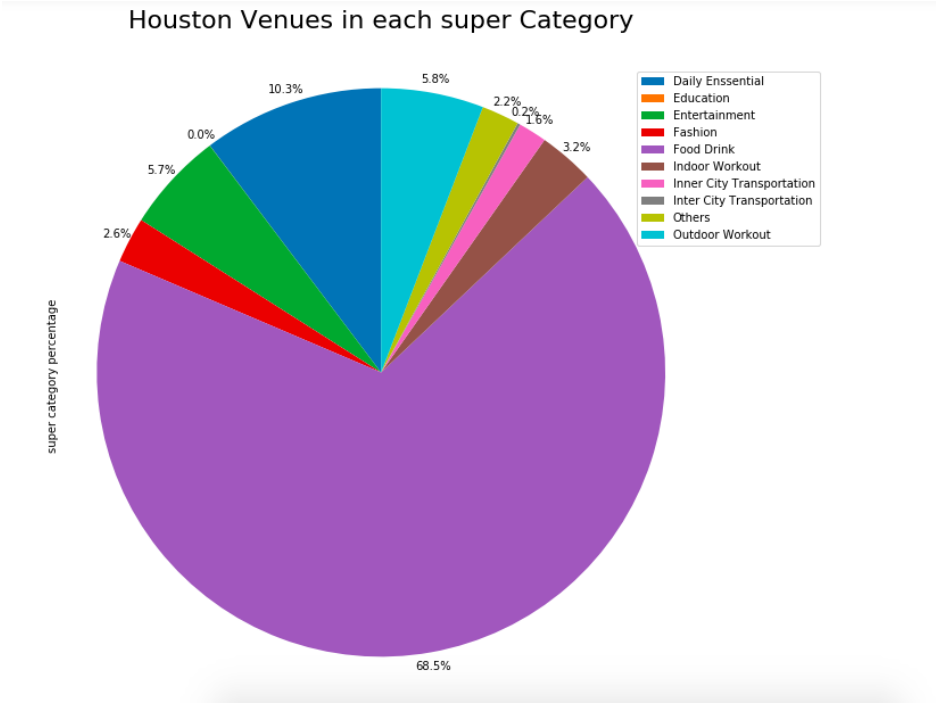

Figure 5: The percentage of venues in each super category for Houston

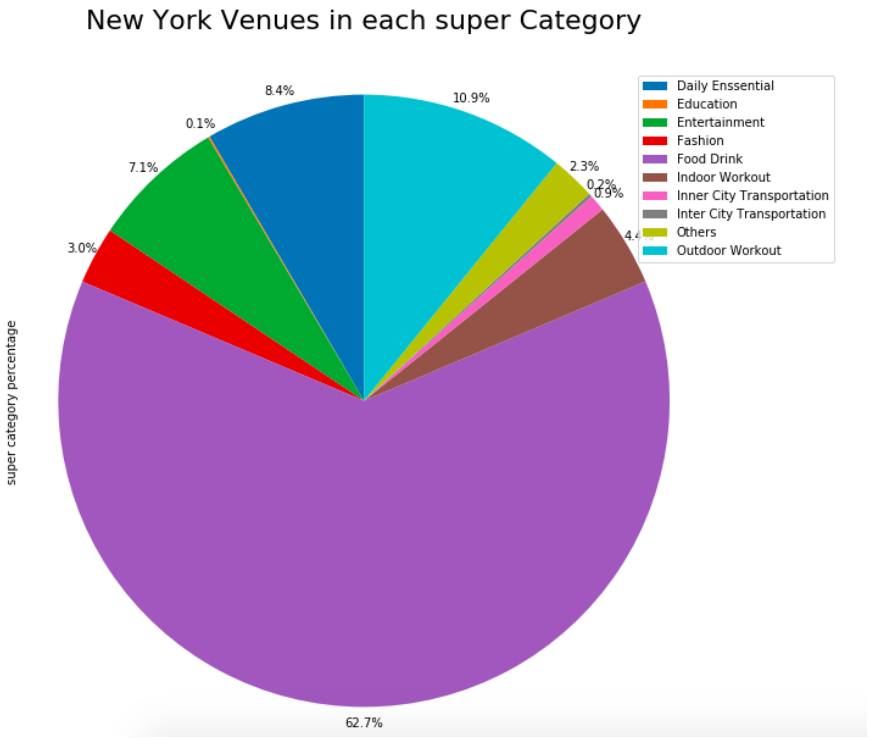

Figure 6: The percentage of venues in each super category for Houston

4. Methodologies

For each of neighborhood, it has its own percentage of venues in each super category, which is what we need if we are considering moving there. The problem here is it is hard to go through over 100 or 300 neighborhoods and check your favorite one. So the problem is to find similar neighborhoods and then we can use the properties of this set of similar neighborhood to decide if those neighborhoods are we are looking for.

The method used in this report is k-mean clustering algorithm. First is to find how many clusters for each city. So for each cities, I applied elbow method to find the optimum cluster numbers. The number for Houston is 4, after which the decrease of distortion slows down. For New York, it is a little to select since the distortion keeps pretty fast decrease ratio, but I still 4 as the cluster numbers for New York for the easy comparison of clustering results later. The following two graphs show the elbow plot for Houston and New York.

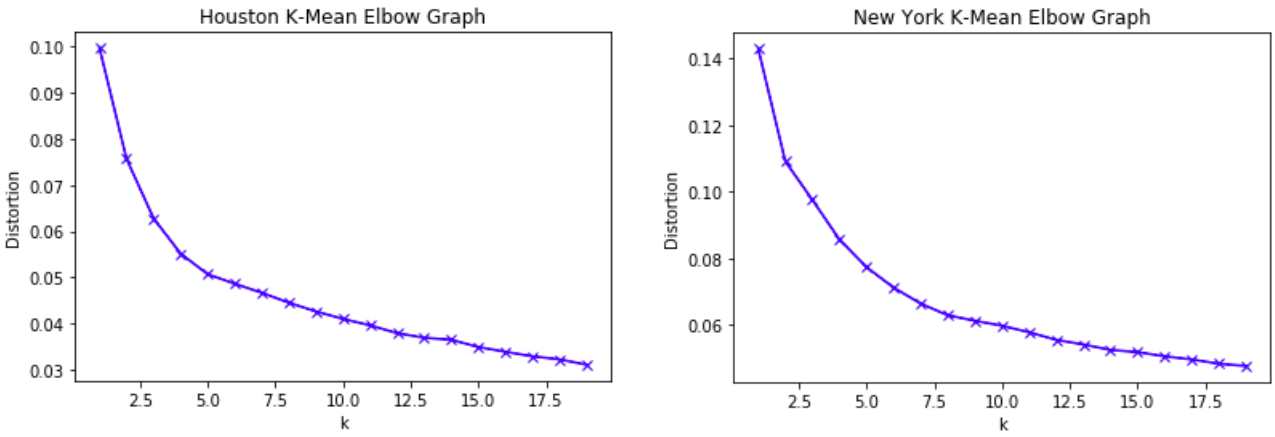

Figure 7: Elbow Plot for k-mean clustering method for Houston and New York city

5. Results

5.1 View neighborhood cluster in a pie chart

After clustering neighborhood, I sum all the venues in each super category inside each cluster together. We can see the major functionality of each clusters of neighborhood. From analysis pie chart of super category for two cities, the Food and Drink is the biggest category. Here we see the same result.

Let's take a look at Houston first. Beside of Food and Drink, the second biggest category in Cluster 1 and Cluster 2 is Daily Essential. We can guess from this information Cluster1 and Cluster2 are mainly residential area. Cluster 1 has super high percentage in Food and Drink and Cluster 2 has similar ratio of Food and Drink as in cities' pie chart. Cluster 2 is more mature and probably more expensive neighborhood. So I like to call Cluster 1 as Developing Residential Neighborhoods and Cluster 2 as Mature Residential Neighborhoods. Cluster 0 and Cluster 3 have 3 high ratio categories: Entertainment, Daily Essential and Outdoor Workout. I feel those two clusters are more vibrant and active. Cluster 0 has similar ratios between those three categories; I prefer to call it as Mix Residential and Entertainment Neighborhoods. Cluster 3 has little Daily Essential venues but has more ratio on Entertainment and Outdoor Workout, which makes me believe that this is probably those area mainly for fun. I will call this cluster as Entertainment Neighborhoods.

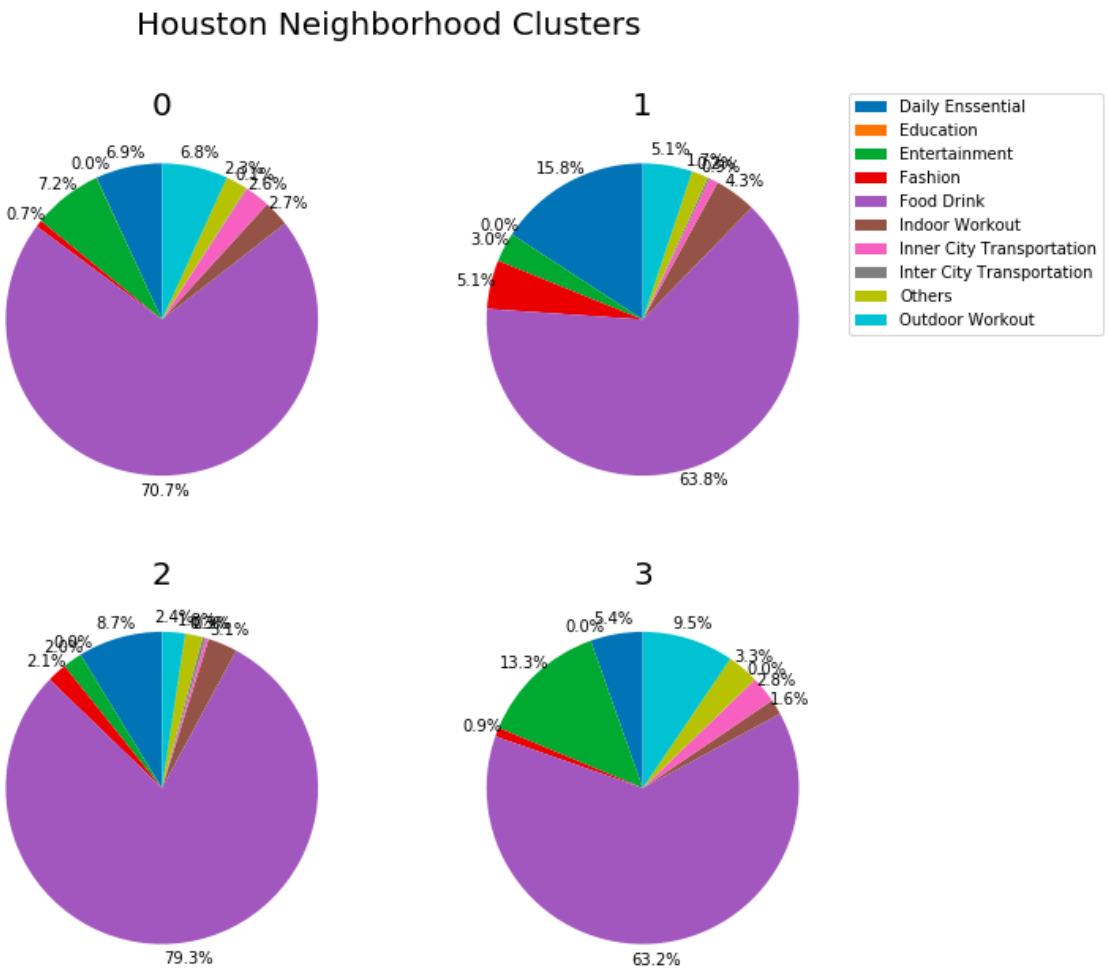

Figure 8: Pie charts for 4 neighborhood clusters in Houston.

For New York City, from pie chart, we can see ratios in each cluster are very diverse. Each cluster seems to have its own functionality. Cluster 1 has dramatically large amount of Outdoor Workout venues, I will call it Outdoor Neighborhoods. Cluster3 has vast amount of Entertainment venues, I will call it Entertainment Neighborhoods. Cluster 0 has well balanced ratio in each category and the ratio is similar to New York city's pie chart ratio, as the same in Houston's cluster 1, I will call this one as Mature Residential Neighborhoods. Cluster2 has a super high ratio in Food and Drink, and given the city is New York, I guess those area are for shopping and business, so I call it as Business Neighborhoods.

So I discussed the properties of each clusters and named then based on the ratio of venues in each category. In Houston, clusters from 0 to 4 are Mix Residential and Entertainment Neighborhood, Developing Residential Neighborhoods, Mature Residential Neighborhoods and Entertainment Neighborhoods

In New York City, clusters from 0 to 4 are Mature Residential Neighborhoods, Outdoor Neighborhoods, Business Neighborhoods and Entertainment Neighborhoods. Each name represents major character of its cluster but for more detail ratio, it is better to use pie chart.

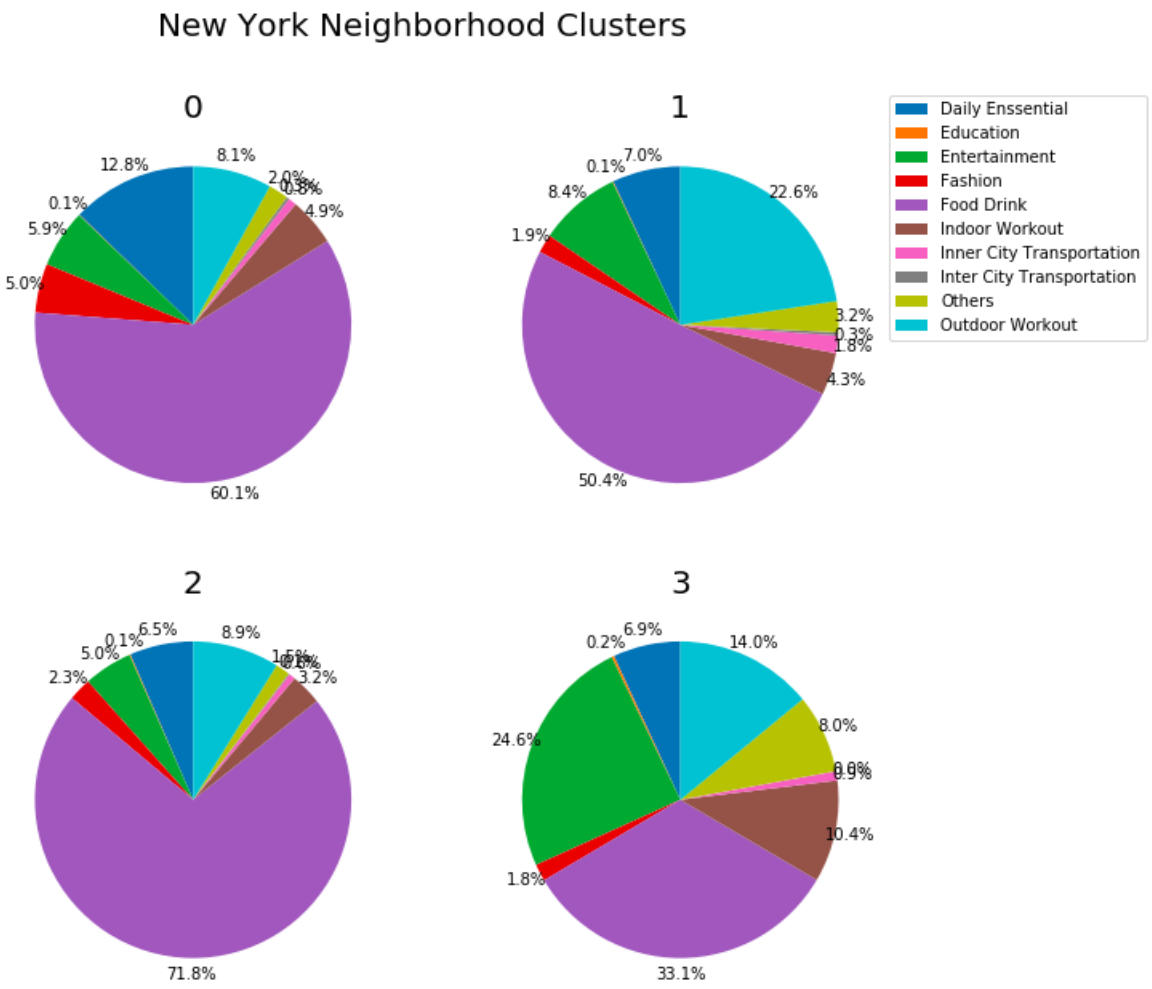

Figure 9: Pie charts for 4 neighborhood clusters in Houston.

5.2 View city cluster results in map view

The following two maps are clusters displayed on Houston (Figure10) and New York City map (Figure11).

The difference is huge. Four neighborhoods clusters spatially spread and are well separated in Houston. But they are concentrated and mix with neighborhoods in other clusters spatially in New York. So in this case, I can imagine, if I live in Residential Neighborhoods and want to find some entertainment, I need

to go to Entertainment Neighborhoods, which cannot be found around my home. In another word the functionality of neighborhoods has physical geographical boundary, which I didn't see in New York's clustering map.

In this case, it sound like it is more convenient to live in New York, because geographically, it has more diversity in terms of functionally of nearby neighborhood which can save some travel time for example and live in a more dynamic environment. But for people who like to have more boundary of living environment and most of time stay clear from some functionality, such as entertainment, Houston is their best choice.

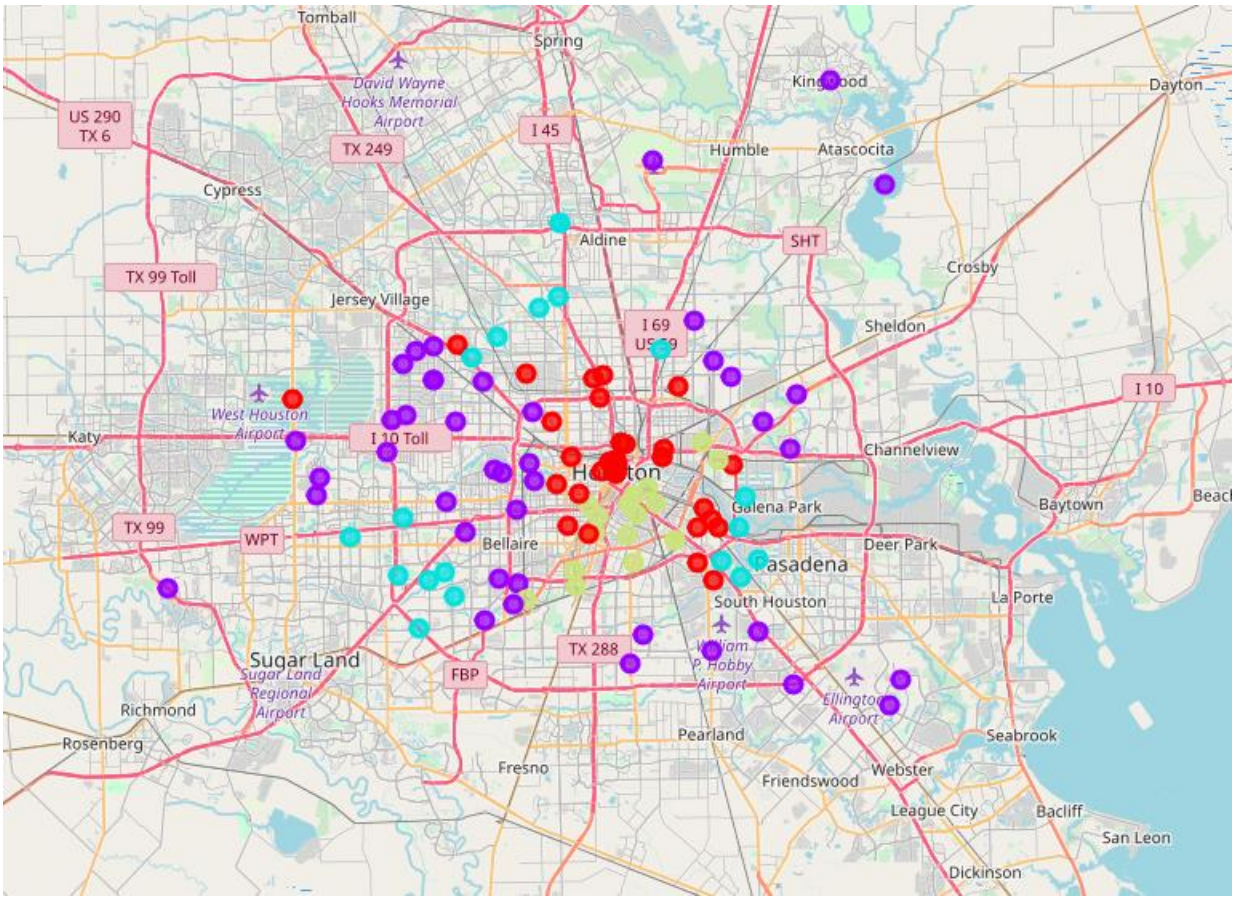

Figure 10: Neighborhood Clusters in Houston, red: Cluster 0; purple: Cluster 1; blue: Cluster 2;

yellow: Cluster 3.

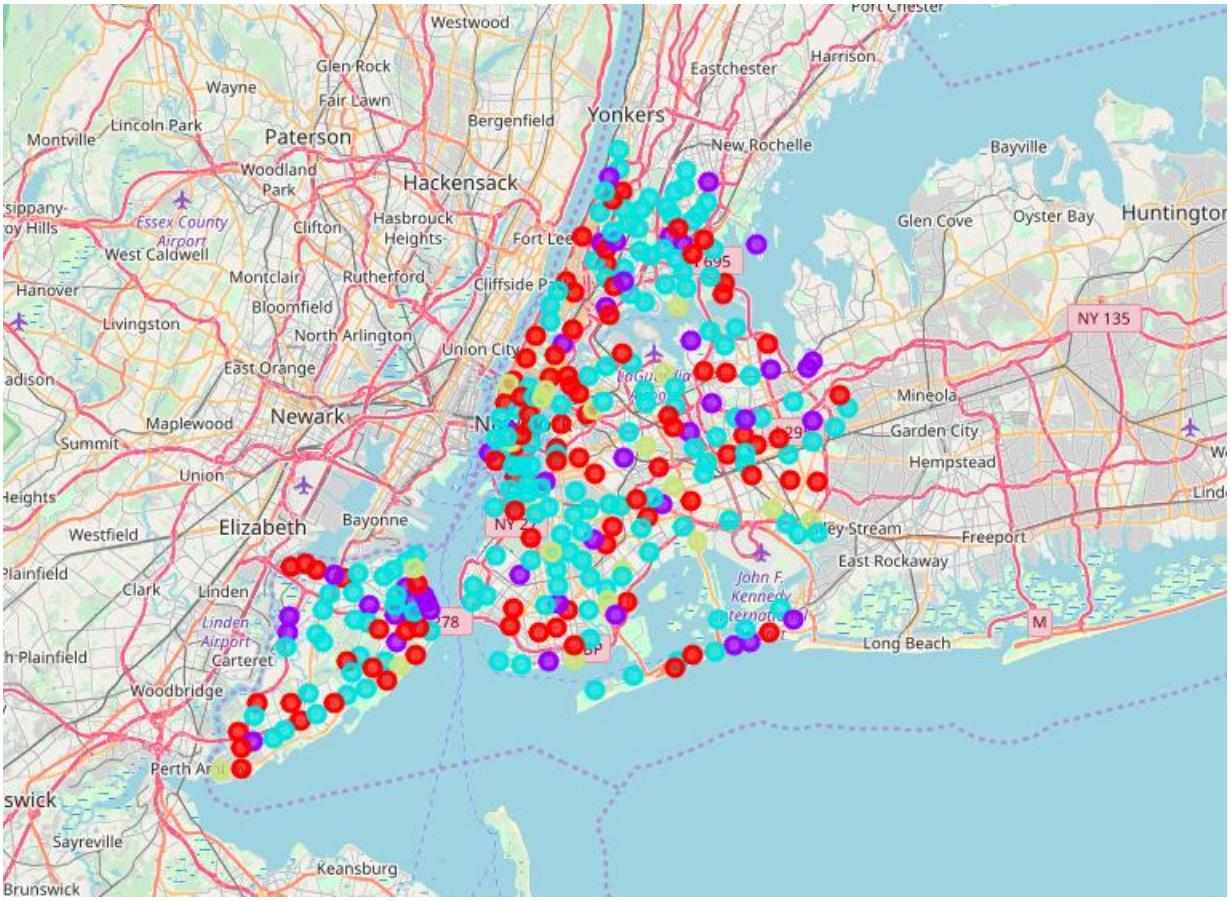

Figure 11: Neighborhood Clusters in New York, red: Cluster 0; purple: Cluster 1; blue: Cluster 2;

yellow: Cluster 3.

6. Discussion:

From above clusters results, we can clearly see the similarity and difference between Houston and New York. The venues are more intertwined and mixed in New York than in Houston, which is correlated with higher density of population and larger population in New York. Overall, the results are simply and easy to understand and are good representation for both cities. For this reason, I think k-mean is good choice here despite other clustering algorithm may give slight better distortion.

7. Conclusion:

In this report, to show an example of helping people “see” the city before they move, I compared two biggest cities in USA, Houston and New York. The venues information is used for visualizing city’s neighborhood functionality. Venue super category is defined to better represent overall impression of neighborhood. Using k-mean clustering method, I summarized neighborhoods in to 4 neighborhood

clusters for each city and used ratio of venues in each super category to discuss neighborhood properties. While the neighborhoods in different clusters are mixed and close to each other in New York, they are spatially separated in Houston. This will affect the people's daily life.

8. Future Work:

Venues data is only a small part of data. New York and Houston are only two cities. The future work will include more affecting index, such as weather, and also more cities in two the model.