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 of 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