IBM Data Science Capstone Project Report

Airbnb Analysis Beijing Chaoyang District

Shuren Qu

Backgrounds

- Objective

There are more than 11,000 Airbnb rooms listed in data set published by "public.opendatasoft.com". For potential investors who want to start an Airbnb business in Beijing, it would beneficial to conduct an analysis based on location and see if there is any correlation between location features and Airbnb monetization. And for both investors and travelers, they will be benefited from a visualization color coded each Airbnb asset with its segmentation associated with its location features.

- Data Used

Airbnb listing data from "public.opendatasoft.com“

API provided by "Foursquare" to find common venues

Data Process (1)

- Raw Data

There are many fields which may not useful to our analysis, so we will drop them, but remaining: "Coordinates", to find location features through Foursquare API calls; "Room price", as indication ability of monetization; And "Room type" as we want to only use the most common room type for analysis, as room type itself is a independent variable which price may depend on.

	Room ID	Host ID	Neighbourhood	Room type	Room Price	Minimum nights	Number of reviews	Date last review	Number of reviews per month	Rooms rent by the host	Availibility	Updated Date	City	Country	Coordinates
0	23863938	147652234	Chaoyang	Entire home/apt	398	1	6	2/19/2019	0.35	5	364	9/23/2019	Beijing	China	39.8952155567, 116.46591907
1	23914071	19772308	Chaoyang	Entire home/apt	418	1	1	4/1/2018	0.06	1	0	9/23/2019	Beijing	China	39.9577003398, 116.443189661
2	23915836	158663144	Chaoyang	Entire home/apt	397	20	49	6/12/2019	2.74	9	3	9/23/2019	Beijing	China	39.891989506, 116.44585669
3	24186440	94142508	Chaoyang	Entire home/apt	518	1	0	NaN	NaN	7	365	9/23/2019	Beijing	China	39.9265754348, 116.615418345
4	24274046	29488633	Chaoyang	Private room	171	1	15	8/30/2019	0.85	27	358	9/23/2019	Beijing	China	39.9975167474, 116.464205076

Data Process (2)

- Location Base Data

3 key columns are selected, which are price, latitude and longitude

See table below of data description: price range is big enough for analysis

	Room type	Room Price	Coordinates
0	Entire home/apt	398	39.8952155567, 116.46591907
1	Entire home/apt	418	39.9577003398, 116.443189661
2	Entire home/apt	397	39.891989506, 116.44585669
3	Entire home/apt	518	39.9265754348, 116.615418345
4	Private room	171	39.9975167474, 116.464205076
...	...	...	...
11825	Private room	697	39.9859502087, 116.435503117
11826	Entire home/apt	525	39.9001321372, 116.470426807
11827	Entire home/apt	647	39.9324756528, 116.467495116
11828	Private room	199	39.8875319023, 116.466949285
11829	Entire home/apt	801	39.9133870212, 116.47546383

11830 rows × 3 columns

	Room Price	Latitude	Longitude
count	6876.000000	6876.000000	6876.000000
mean	678.210151	39.931395	116.475742
std	2247.202016	0.042466	0.046627
min	0.000000	39.820607	116.347193
25%	391.000000	39.899278	116.448196
50%	491.000000	39.922715	116.466536
75%	631.000000	39.959297	116.494151
max	71110.000000	40.099771	116.621531

Data Process (3)

- DBSCAN (10,000+ Rooms to 40 Neighborhoods)

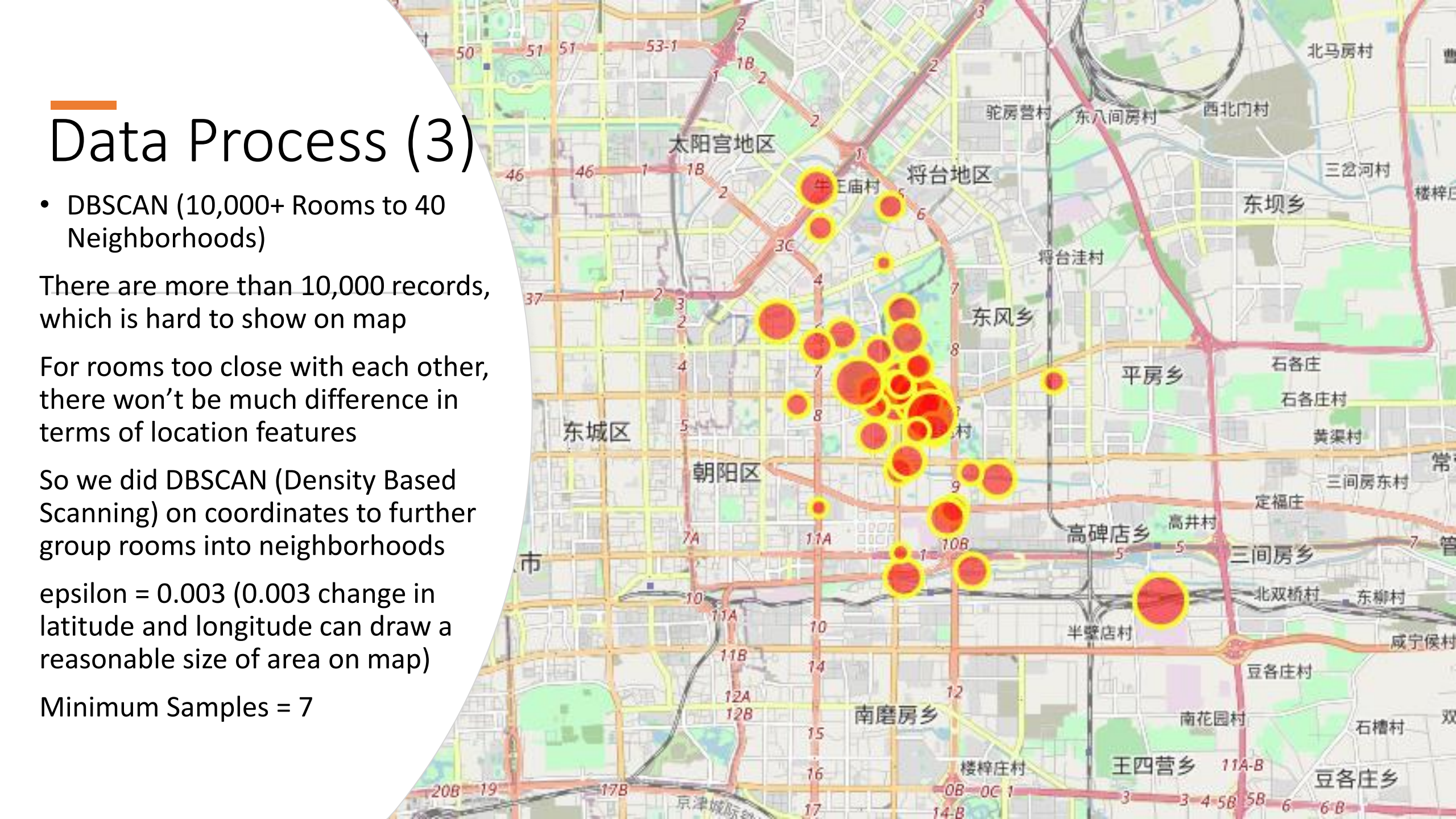
There are more than 10,000 records, which is hard to show on map

For rooms too close with each other, there won't be much difference in terms of location features

So we did DBSCAN (Density Based Scanning) on coordinates to further group rooms into neighborhoods

epsilon = 0.003 (0.003 change in latitude and longitude can draw a reasonable size of area on map)

Minimum Samples = 7



Data Segmentation

- Add Location Features and Segment based on features

Add most common venues using Foursqure API and runed K Nearest score to cluster neighborhoods into Clusters

5 Clusters found based on K Nearest Algorithm

	Neighborhood	Antique Shop	Asian Restaurant	Athletics & Sports	BBQ Joint	Bagel Shop	Bakery	Bar	Beijing Restaurant	Bookstore	...	Supermarket	Sushi Restaurant	Szechuan Restaurant	Taiwanese Restaurant	Tennis Court	Thai Restaurant	Vietnamese Restaurant	Xinjiang Restaurant	Yoga Studio	Yunnan Restaurant
0	-1.0	0.0	0.00	0.0	0.0	0.100	0.0	0.0	0.0	0.0	...	0.00	0.00	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0
1	0.0	0.0	0.00	0.0	0.0	0.100	0.0	0.0	0.0	0.0	...	0.00	0.00	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0
2	1.0	0.0	0.00	0.0	0.0	0.000	0.1	0.1	0.0	0.0	...	0.00	0.00	0.0	0.0	0.000000	0.0	0.1	0.0	0.0	0.0
3	2.0	0.0	0.20	0.0	0.0	0.000	0.0	0.0	0.0	0.1	...	0.00	0.00	0.0	0.0	0.000000	0.0	0.0	0.0	0.1	0.0
4	3.0	0.0	0.00	0.0	0.0	0.125	0.0	0.0	0.0	0.0	...	0.00	0.00	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0
5	4.0	0.0	0.00	0.0	0.0	0.100	0.0	0.0	0.0	0.0	...	0.00	0.00	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0
6	5.0	0.0	0.00	0.0	0.0	0.000	0.1	0.1	0.0	0.0	...	0.00	0.00	0.0	0.0	0.000000	0.0	0.1	0.0	0.0	0.0
7	6.0	0.0	0.00	0.0	0.0	0.100	0.0	0.0	0.0	0.1	...	0.00	0.00	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0
8	7.0	0.0	0.00	0.0	0.0	0.100	0.0	0.1	0.0	0.0	...	0.00	0.00	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0
9	8.0	0.0	0.00	0.0	0.0	0.000	0.0	0.0	0.0	0.0	...	0.00	0.00	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0
10	9.0	0.0	0.00	0.0	0.0	0.100	0.0	0.0	0.0	0.0	...	0.00	0.00	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0
11	10.0	0.0	0.00	0.0	0.0	0.100	0.1	0.0	0.0	0.0	...	0.00	0.00	0.0	0.0	0.000000	0.0	0.1	0.0	0.0	0.0
12	11.0	1.0	0.00	0.0	0.0	0.000	0.0	0.0	0.0	0.0	...	0.00	0.00	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0
13	12.0	0.0	0.00	0.0	0.0	0.000	0.0	0.0	0.0	0.0	...	0.00	0.00	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0
14	13.0	0.0	0.00	0.0	0.0	0.000	0.0	0.0	0.0	0.0	...	0.00	0.00	0.0	0.0	0.400000	0.0	0.0	0.0	0.0	0.0
15	14.0	0.0	0.00	0.0	0.0	0.000	0.0	0.0	0.0	0.0	...	0.00	0.00	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0
16	15.0	0.0	0.00	0.0	0.0	0.000	0.0	0.0	0.0	0.0	...	0.00	0.00	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0
17	16.0	0.0	0.00	0.0	0.0	0.000	0.0	0.0	0.0	0.0	...	0.00	0.00	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0
18	17.0	0.0	0.00	0.0	0.0	0.100	0.0	0.0	0.0	0.0	...	0.00	0.00	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0
19	18.0	0.0	0.00	0.0	0.0	0.000	0.0	0.0	0.0	0.0	...	0.10	0.00	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0
20	19.0	0.0	0.00	0.0	0.0	0.100	0.1	0.1	0.0	0.0	...	0.00	0.00	0.0	0.0	0.000000	0.0	0.1	0.0	0.0	0.0
21	20.0	0.0	0.00	0.1	0.0	0.000	0.1	0.1	0.0	0.0	...	0.00	0.00	0.0	0.0	0.000000	0.0	0.1	0.0	0.0	0.0
22	21.0	0.0	0.20	0.0	0.0	0.000	0.0	0.0	0.0	0.0	...	0.00	0.00	0.0	0.0	0.000000	0.0	0.0	0.1	0.0	0.0
23	22.0	0.0	0.20	0.0	0.0	0.000	0.0	0.0	0.0	0.1	...	0.00	0.00	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0
24	23.0	0.0	0.00	0.0	0.0	0.000	0.0	0.0	0.0	0.0	...	0.25	0.00	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0
25	24.0	0.0	0.00	0.0	0.0	0.000	0.0	0.0	0.0	0.0	...	0.00	0.00	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0
26	25.0	0.0	0.00	0.0	0.0	0.000	0.0	0.0	0.0	0.0	...	0.00	0.00	0.0	0.0	0.333333	0.0	0.0	0.0	0.0	0.0
27	26.0	0.0	0.30	0.0	0.0	0.000	0.0	0.0	0.0	0.0	...	0.00	0.00	0.0	0.0	0.000000	0.0	0.0	0.1	0.0	0.0
28	27.0	0.0	0.00	0.0	0.0	0.000	0.0	0.0	0.0	0.0	...	0.00	0.00	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0
29	28.0	0.0	0.00	0.0	0.0	0.000	0.0	0.0	0.0	0.0	...	0.00	0.25	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0

Data Segmentation Differentiations

- Naming each Cluster and draw box plot to examine price differentiations

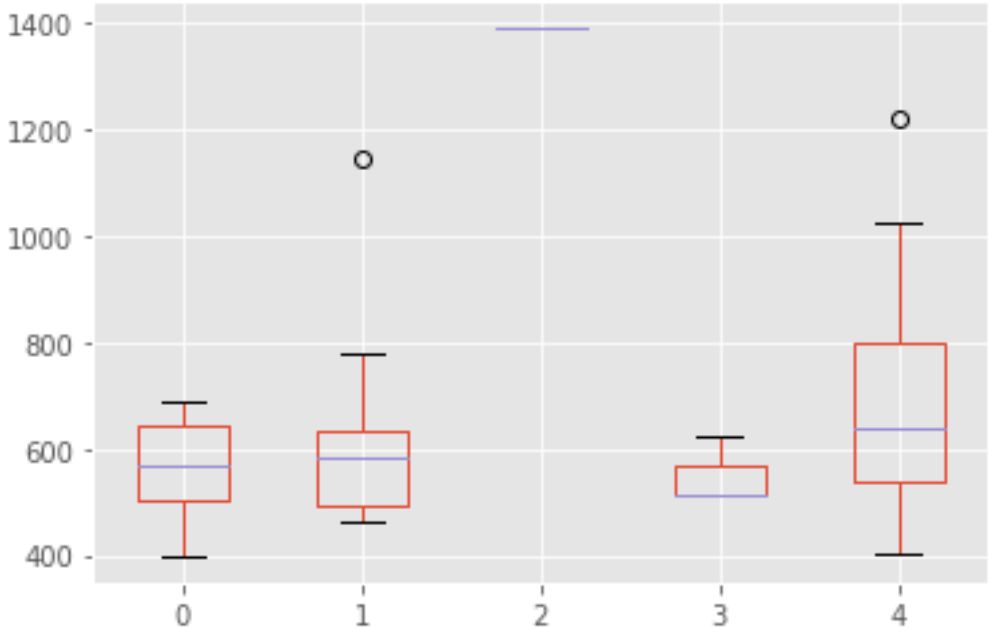
Cluster 0: Coffee Shop Area
Asian Restaurant 3
Coffee Shop 7

Cluster 1: Foreign Restaurants Area
Chinese Restaurant 1
Cocktail Bar 1
Grocery Store 2
Hotpot Restaurant 1
Italian Restaurant 4
Mexican Restaurant 1
Noodle House 1
Yunnan Restaurant 1

Cluster 2:
Antique Shop 1

Cluster 3: Park Area
Park 2
Tennis Court 1

Cluster 4: Modern Lifestyle Chinese Restaurant 2
Coffee Shop 4
Coworking Space 1
Gym 1
Hotel 3



Visualization

- Visualize each area on Map with price indicator (bubble size) as a tool for reference