IBM Data Science Capstone Project

Beijing Chaoyang District Airbnb Data Analysis

Lorin Zhao

Abstract

The analysis based on Beijing Chaoyang District data from “public.opendatasoft.com”, and API provided by Foursquare, find clustering patterns for neighborhoods. It also provide reference for investors and customers to choose their best Airbnb location.

Keywords: Data Science, clustering, Airbnb data

IBM Data Science Capstone Project

Background

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 Source

I acquired Airbnb listing data from "public.opendatasoft.com"， which contains a list of Airbnb assets and its location, price, number of rent information. In addition, I used API provided by "Foursquare" to find common venues around Airbnb assets to portrait its location features.

Prepare 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
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
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
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
24186440	94142508	Chaoyang	Entire home/apt	518	1	0	NaN	NaN	7	365	9/23/2019	Beijing	China	39.9265754348, 116.615418345
24274046	29488633	Chaoyang	Private room	171	1	15	8/30/2019	0.85	27	358	9/23/2019	Beijing	China	39.9975167474, 116.464205076

	Room Price	Coordinates	Latitude	Longitude
0	398	39.8952155567, 116.46591907	39.8952155567	116.46591907
1	418	39.9577003398, 116.443189661	39.9577003398	116.443189661
2	397	39.891989506, 116.44585669	39.891989506	116.44585669
3	518	39.9265754348, 116.615418345	39.9265754348	116.615418345
6	455	39.9331540173, 116.45231878	39.9331540173	116.45231878
...	...	...	...	...
11820	391	39.9394216027, 116.447158718	39.9394216027	116.447158718
11823	292	39.915874248, 116.499888534	39.915874248	116.499888534
11826	525	39.9001321372, 116.470426807	39.9001321372	116.470426807
11827	647	39.9324756528, 116.467495116	39.9324756528	116.467495116
11829	801	39.9133870212, 116.47546383	39.9133870212	116.47546383

	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

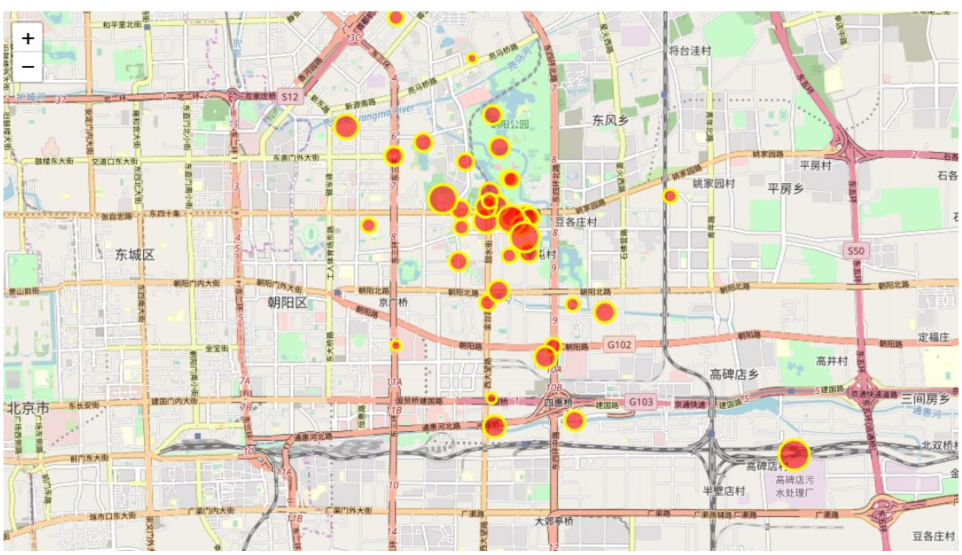
I choose 3 key columns which are price, latitude and longitude. See right table of data description: price range is big enough for analysis

DBSCAN (10,000+ Rooms to 40 Neighborhoods)

There are more than 10,000 records, which is hard to show on map. And for rooms too close with each other, there won't be much difference in terms of location features. So I 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

Below is results of 40 neighborhoods on map, created by folium:



Add Location Features and Segment based on features

Next step, I added most common venues using Foursqure API, process the data to easily show top 10 most common venues for each neighborhood:

	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	-1.0	Hotel	Shopping Mall	Grocery Store	Fast Food Restaurant	New American Restaurant	Park	Coffee Shop	Chinese Restaurant	Gym	Bagel Shop
1	0.0	Coffee Shop	Hotel	Fast Food Restaurant	New American Restaurant	Sandwich Place	Chinese Restaurant	Park	Bagel Shop	Grocery Store	Flower Shop
2	1.0	Italian Restaurant	Hot Spring	Vietnamese Restaurant	Grocery Store	Bakery	Bar	Japanese Restaurant	Park	Sandwich Place	Yunnan Restaurant
3	2.0	Asian Restaurant	Chinese Restaurant	Yoga Studio	Bookstore	Coffee Shop	Café	Farmers Market	Pizza Place	Flower Shop	Fast Food Restaurant
4	3.0	Gym	Grocery Store	Park	Coffee Shop	Bagel Shop	Fast Food Restaurant	New American Restaurant	Chinese Restaurant	Dim Sum Restaurant	Department Store

And then, I ran a K-Nearest algorithm to cluster neighborhoods into Clusters: 5 Clusters found based on K Nearest Algorithm:

Based on the frequency of most common venue, I named each cluster based on their characteristics:

Cluster 0: Coffee Shop Area

Asian Restaurant 3

Coffee Shop 7

Convenience Store 1

Hotel 2

Japanese Restaurant 2

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

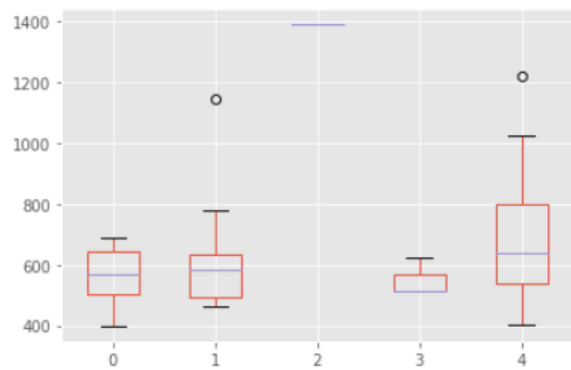
Coffee Shop 4

Coworking Space 1

Gym 1

Hotel 3

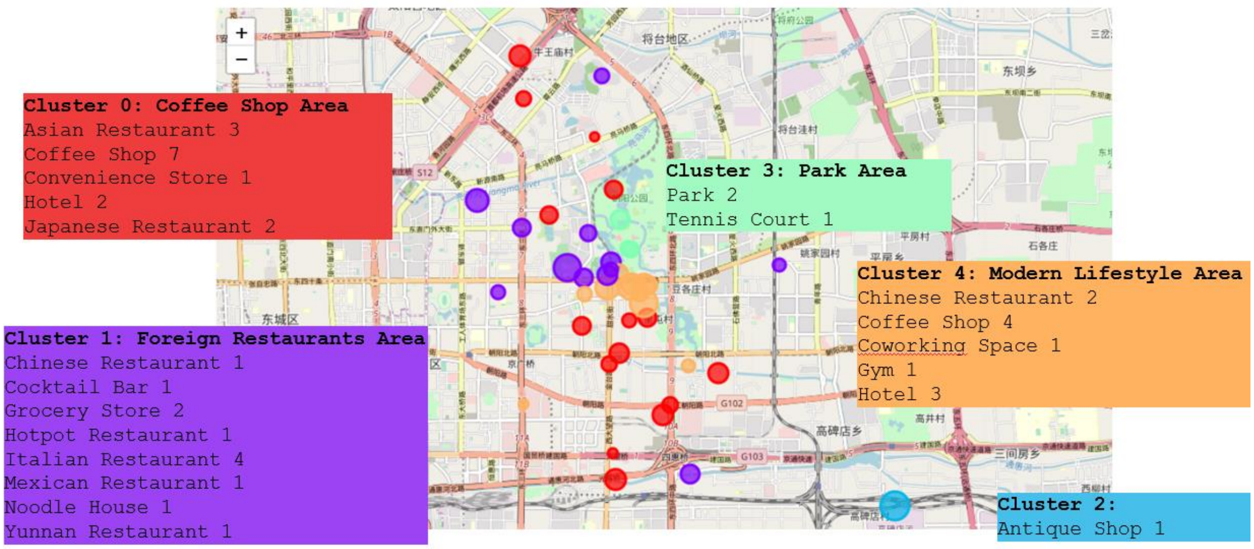
To evaluate if different venue has different price performance, I drew a box plot. Although it's not very significant, the chart shows Antique shop area and Modern Lifestyle area out performs in price:



Visualization

Finally, as one of the objectives, I did a visualization to show different cluster and their price performance. This could be beneficial for both investors and customers:

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



References

Google Map

public.opendatasoft.com

Foursquare

Footnotes

This analysis is not possible without technics and skills taught in IBM Data Science course on Coursera. It can be further improved in the future, to have more detailed analysis with more technics such as polynomial regression can be used to predict price given the location features.