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

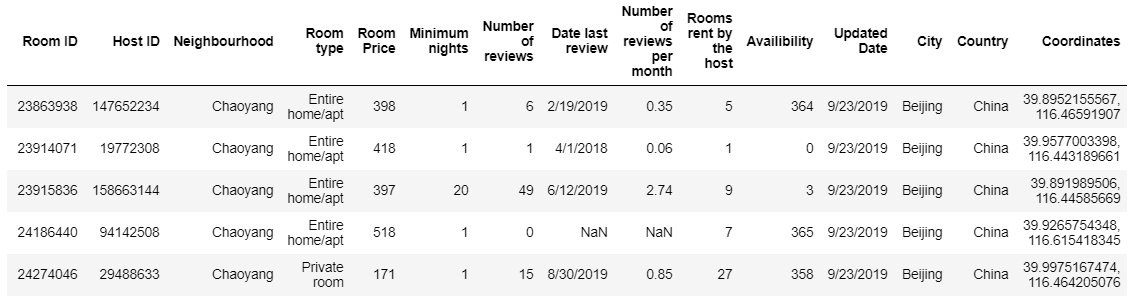
There are more than 11,000 Airbnb rooms listed in data set published by "public.opendatasoft.com". For potential investors who want to start an Airbnb business in Beijing, it would beneficial to conduct an analysis based on location, and see if there is any correlation between location features and Airbnb 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.

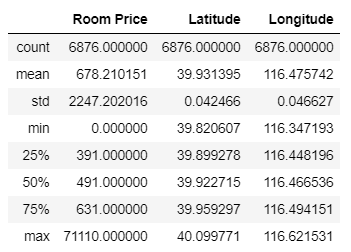
## Data Source

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.





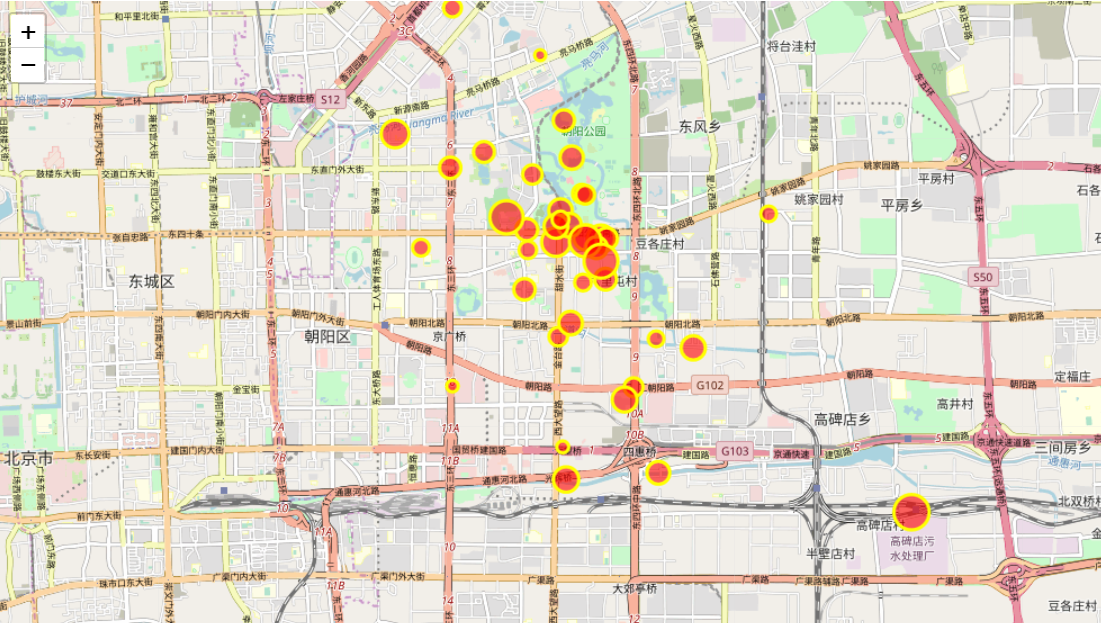
I choose 3 key columns which are price, latitude and longitude. See right table of data description: price range is big enough for analysis

## 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:



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

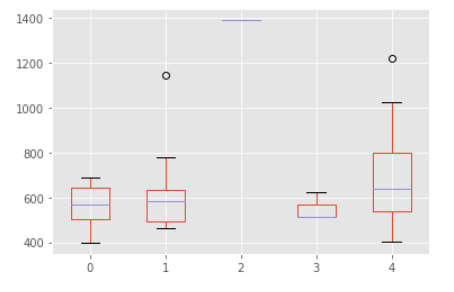
Coffee Shop 4

Coworking Space 1

Gym 1

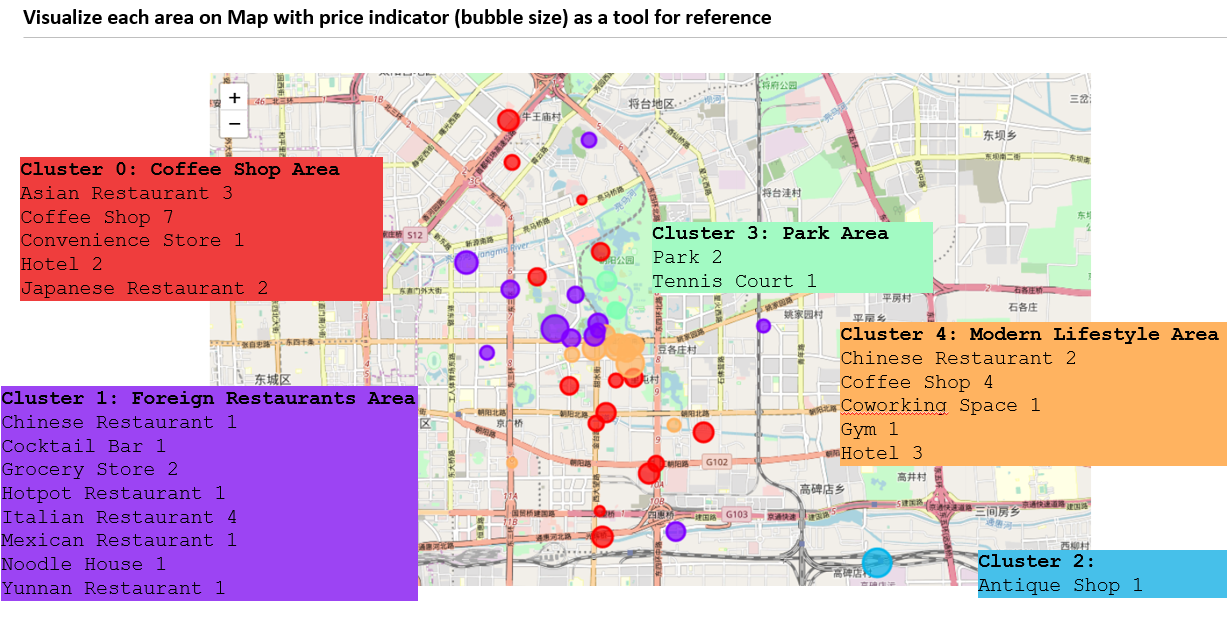
Hotel 3

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:



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.