Analysis of Airbnb Dataset

by Visualization and Modeling for Short-term Rental Hosts to Improve Business Strategies

Supported by 2020 NYU Shanghai Deans' Undergraduate Research Fund

Han, Xinyao
Major: Data Science
Email: xh1082@nyu.edu

Ke, Xu
Major: Data Science
Email: kx417@nyu.edu

Li, Yuxuan
Major: Finance
Email: yl5668@nyu.edu

Chen, Guodong
Research Mentor
Email: gc1947@nyu.edu

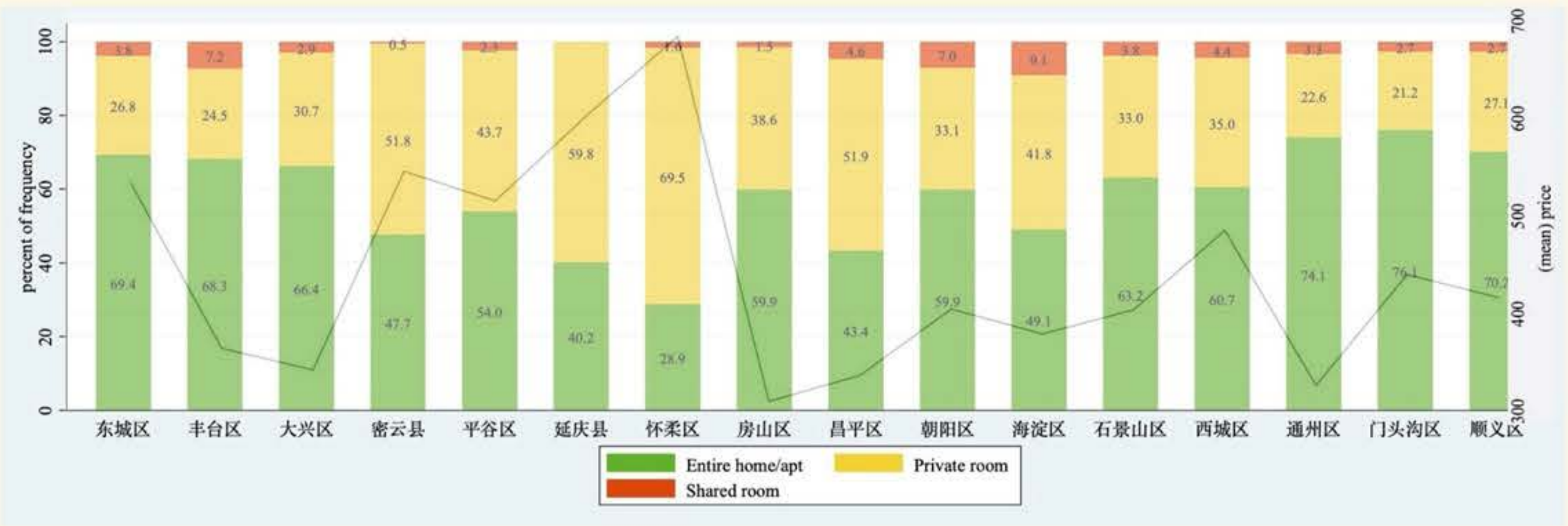
Background

- Sharing economy has sprung up in the 21st century and Airbnb is a typical business practice of the sharing economy model, which reached a valuation of $35 billion with two million people staying each night by October 2019.
- However, due to COVID 19, the Airbnb market falls into disruption and depression. This research aims at providing valuable suggestions on how to develop proper pricing and marketing strategies in the post-pandemic market.

Data and Methodology

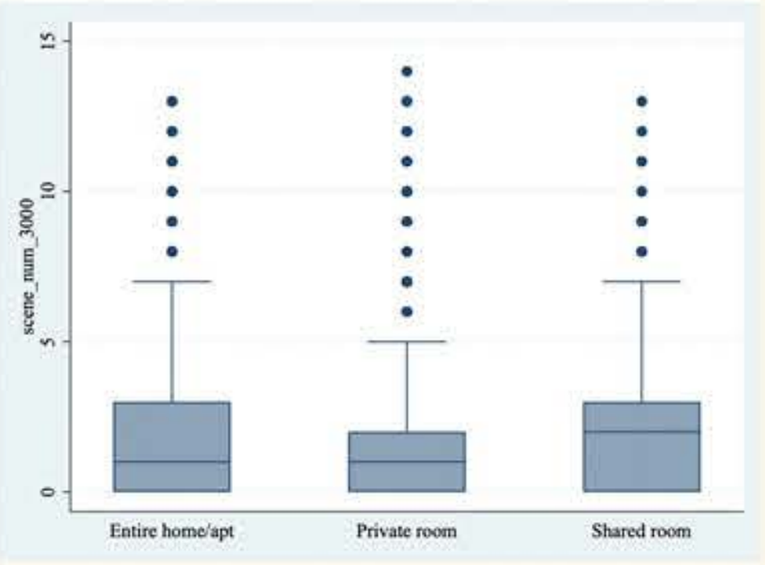
- 28456 listings in Beijing from Insideairbnb; metro station locations from Wikipedia; Tourist attraction and business center list from Beijing Municipal official website.
- Price heat map(ArcGIS); Principal Component Analysis for amenities (Stata); Multiple Linear, Polynomial Linear and Elastic Net Linear Regression for modeling (Python); Term Frequency-Inverse Document Frequency for textual analysis (R and Python).

Descriptive Analysis



The Distribution of Room Type by District and Mean Price of Each District

- Price is more correlated with district, not the room type.
- While Huairou district has the least entire home, it has the highest average price.



Box Graph of Number of Tourist Attractions by Room Type

- Compared with entire home and private room, there are more tourist attractions around shared room within 3km.
- This sheds lights on why shared room demonstrates higher price elasticity when interacted with number of tourist attractions in MLR.

Principal Component Analysis of Amenities

- After breaking each entry into separate words of amenities by python, PCA is employed to analyze their correlation. Top two linear combinations are chosen for regression analysis, which are renamed as **convenience** and **comforts**.
- In order to test the model, we classified all amenities again based on criteria offered by Airbnb and common sense, including family-friendly, kitchen, bedroom, bathroom, security, holiday, and so on. By substituting these manually categorized variables into the model, we got similar results, which justifies the PCA model.
- Huairou district has the highest score on amenities, which explains its high price.

NYU 上海 SHANGHAI 纽约大学

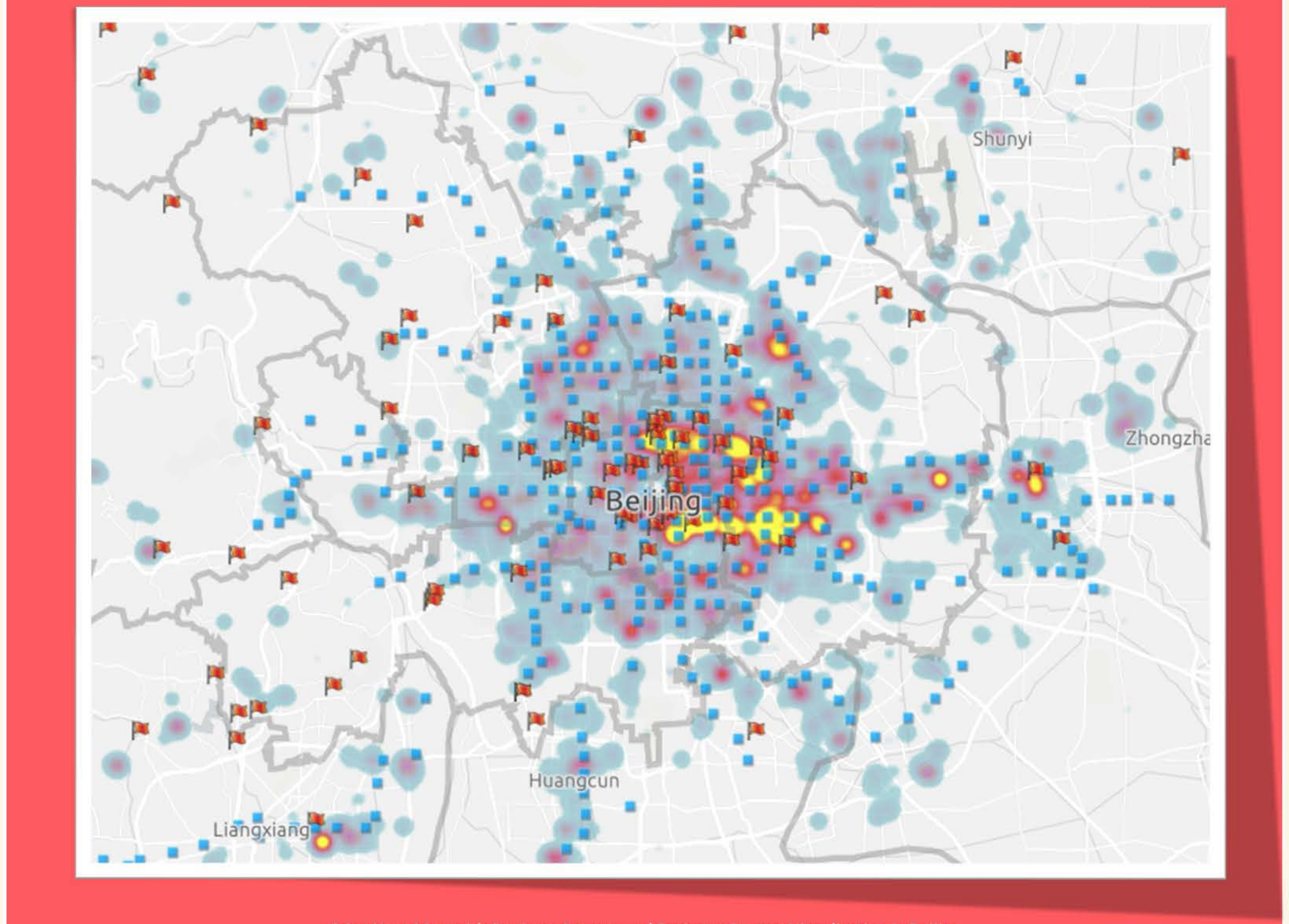
airbnb

Influential Factors of Price

- Number of guests and bathrooms
- Number of metro stations, tourist attractions, and business centers
- Room type and amenities

Key Terms for Marketing

- Enthusiasm of the Host
- Quality of Service and Management

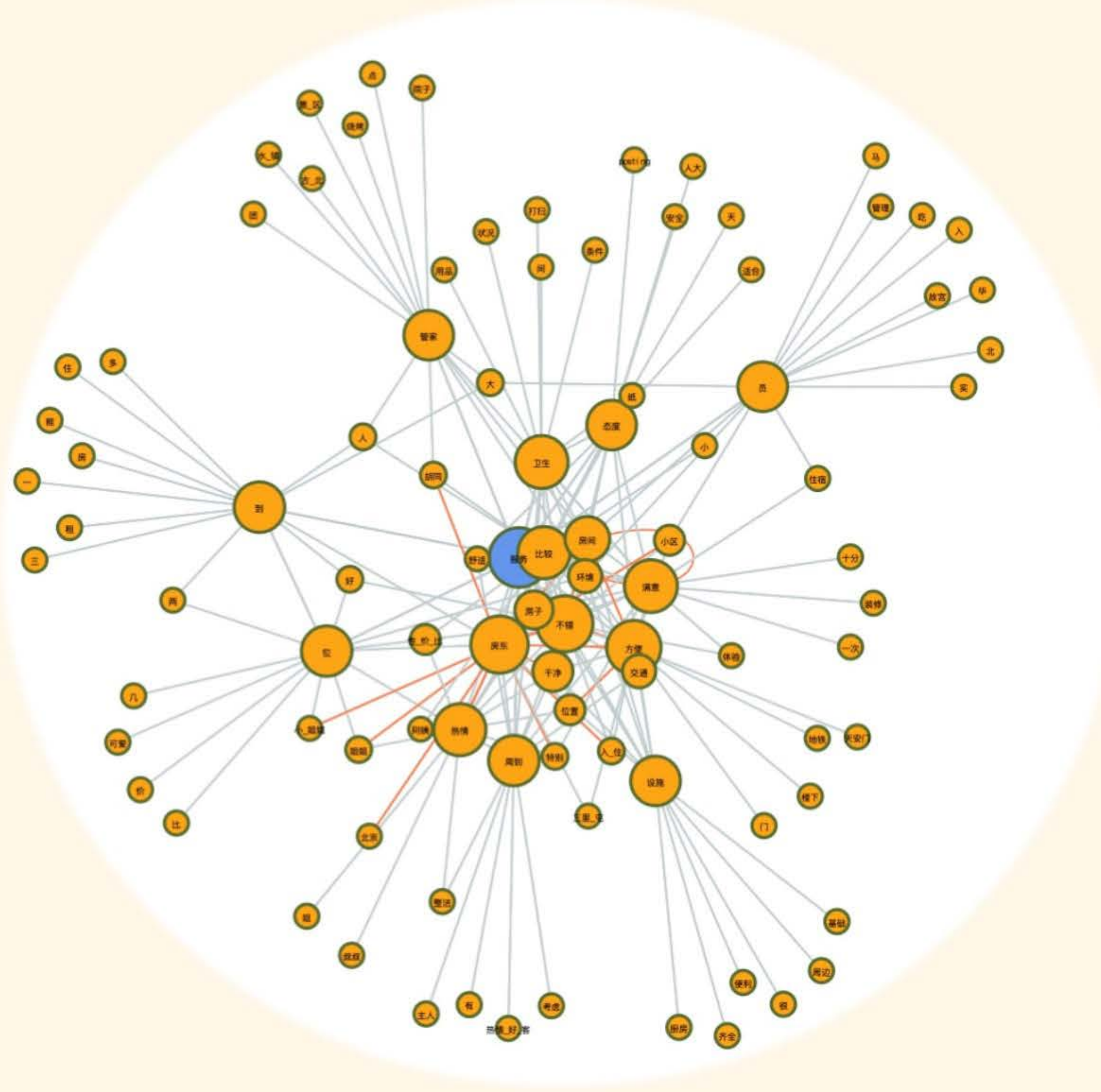


Price Heat Map with Stations, Scenes, and Business Centers Visualization in Beijing

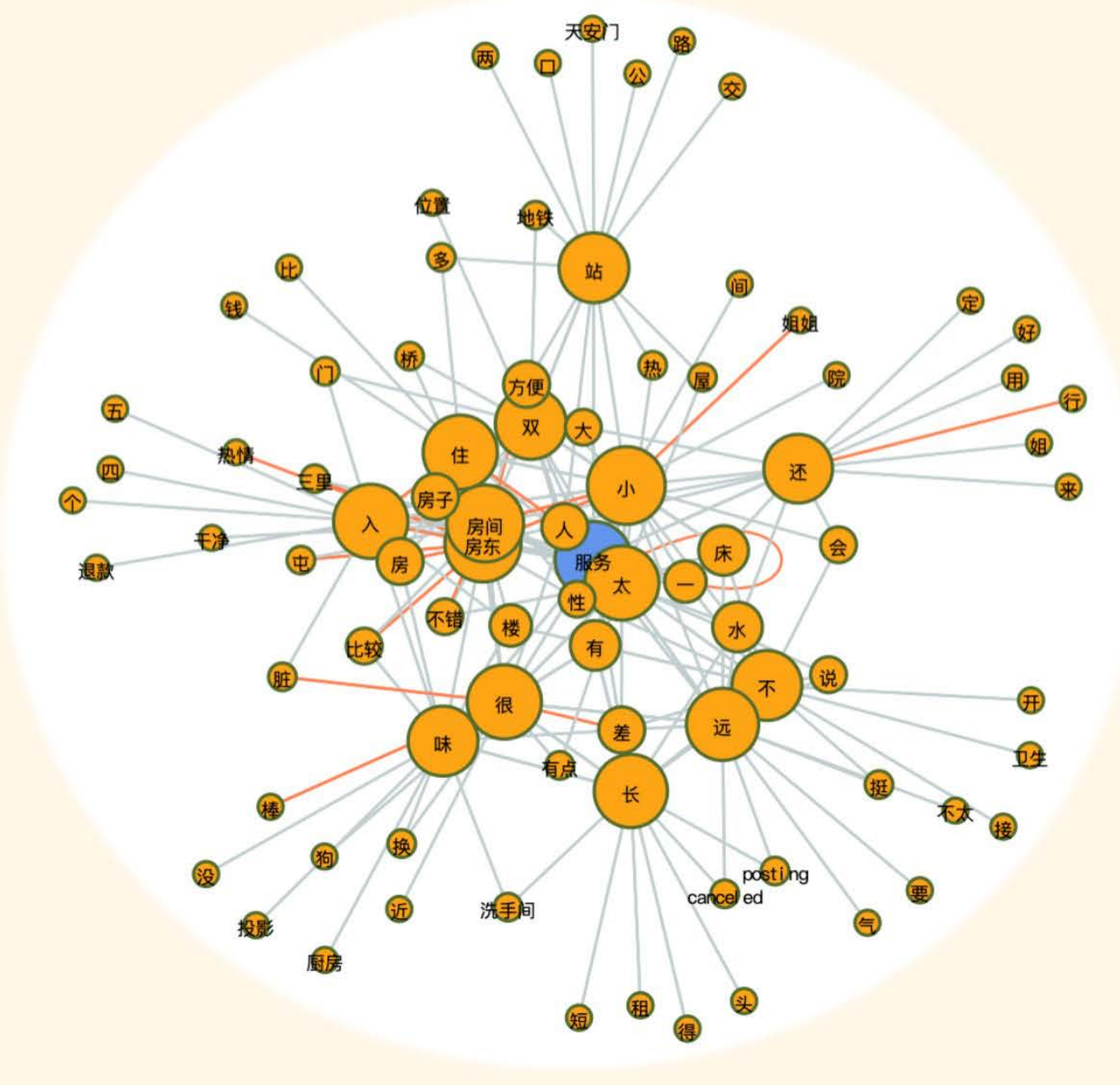
Price from low to high | 🚩 A-level Tourist Attractions | ■ Metro Stations

Textual Analysis of Reviews

- Use Python NLP package jieba to separate Chinese sentences into words.
- Calculate Term Frequency-Inverse Document Frequency for each word.
- select top 5 words with the highest TF-IDF value.
- Use R to depict the co-occurrence network.



Co-occurrence Network of "Service" among Reviews for Listings with High Rating



Co-occurrence Network of "Service" among Reviews for Listings with Low Rating