Data-Driven Analysis and Forecasting of the Los Angeles Airbnb Market

MA678 Midterm Project Proposal

Fengyuan Shen (Vincent)

1. Personal Statement:

With a dedicated academic path in the MSSP at Boston University, my goal is to merge the rigorous discipline of statistics with the vibrant possibilities of data science in the tech industry. My commitment lies in transforming complex datasets into actionable insights that drive product innovation and strategic decisions. The Airbnb dataset provides me an opportunity to enhance my analytical skills, equipping me to extract actionable insights and contribute meaningfully to the tech industry as a data scientist.

2. Literature Review:

In recent years, the rise and development of Airbnb has had a profound impact on urban housing markets and community dynamics. Studies have shown that the distribution of Airbnb listings is influenced by a number of factors *[1]*, including geographic location, socioeconomic status, and transportation accessibility.

According to a comprehensive review by Guttentag *[2][3]* of 132 peer-reviewed journal articles, research on Airbnb has been categorized into six primary domains: the preferences and motivations of Airbnb guests, the incentives and objectives of hosts, the impact of Airbnb on local destinations, regulatory aspects, its influence on the tourism sector, and the operational dynamics of the company itself.

Therefore, based on the previous analysis on Airbnb, I come up with following research objectives.

3. Research Objectives:

(1) *Do data cleaning, process missing values and outliers, and classify descriptive text content to form operable variables.*

(2) *Evaluate the Airbnb market in Los Angeles, such as calculating the distribution of housing in different areas, prices, housing characteristics, etc.*

(3) *Study how factors such as the number of bedrooms, user ratings, and the geographic location of the home affect the price of the home, and use multilevel regression analysis to predict the price of the home with different characteristics.*

(4) *If time permits, I would also like to do sentiment analysis on user reviews and incorporate this factor into the house price prediction model.*

4. Datasets:

The datasets come from Inside Airbnb: http://insideairbnb.com/get-the-data
(1) LA Listings Data: contains listings for Airbnb listings in the Los Angeles area with 75 distinct features.
http://data.insideairbnb.com/united-states/ca/los-angeles/2023-09-03/data/listings.csv.gz
(2) LA Calendar Data: an overview of available listings for the coming year.
http://data.insideairbnb.com/united-states/ca/los-angeles/2023-09-03/data/calendar.csv.gz
(3) LA Review Data: covers the user's review information.
http://data.insideairbnb.com/united-states/ca/los-angeles/2023-09-03/data/reviews.csv.gz

5. Proposed Timeline of Work:

(1) Data Cleaning: 11.9 – 11.20; **(2) EDA:** 11.21 – 11.24;
(3) Modeling and Validation: 11.25 – 12.5; **(4) Write Up:** 12.5 – 12.7

6. Reference:

[1] *Deboosere, R., Kerrigan, D. J., Wachsmuth, D., & El-Geneidy, A. (2019). Location, location and professionalization: a multilevel hedonic analysis of Airbnb listing prices and revenue. Regional Studies, Regional Science, 6(1), 143-156.*
[2] *Jiao, J., & Bai, S. (2020). An empirical analysis of Airbnb listings in forty American cities. Cities, 99, 102618.*
[3] *Guttentag, D. (2019). Progress on Airbnb: A literature review. Journal of Hospitality and Tourism Technology, 10(4)*