DATA 400 Project Proposal 1

Predicting Airbnb Rental Prices

Research Question: Can Airbnb rental prices be effectively estimated using publicly available listing features, and what are the most influential factors that impact pricing across neighborhoods?

Motivation: Short-term rental platforms like Airbnb have transformed the urban housing market. My mini project aims to explore how different features such as amenities, location, and listing descriptions affect pricing, which can help hosts optimize their strategies, assist renters in identifying fair prices, and offer platforms deeper insights into market behavior.

Proposed Data Sources: I will scrape data directly from the Airbnb website, focusing on publicly accessible listing information. The scraped dataset will include features such as price, room type, number of reviews, availability, listing description length, and listed amenities. If time permits, I may also enhance the dataset by scraping or integrating additional contextual features such as neighborhood characteristics, walkability scores, or proximity to major landmarks using external sources like Google Maps or OpenStreetMap.

Methodology: I will begin the project with exploratory data analysis (EDA), including summary statistics and visualizations such as boxplots, heatmaps, and geospatial maps to explore how listing prices relate to features like the number of amenities, location, and listing title length. I will also identify outliers, missing values, and potential data biases. If time permits, I will build a simple regression model such as Linear Regression or a Decision Tree, or test more advanced models like Random Forest or XGBoost. Model performance will be evaluated using basic metrics including R square, MAE, and RMSE. Additional features may be incorporated if I choose this idea for my semester long project, such as proximity to tourist attractions (using coordinates or external POI data), seasonality patterns based on availability calendars, or sentiment analysis of listing descriptions and reviews if text data is available.

Challenges: Scraped data may vary in quality or structure. Due to time constraints, I may not be able to fully clean, or model all features. I also have ethical considerations such as ensuring that no personal information is used and avoiding any modeling that introduces socioeconomic bias (associating pricing with demographic features of neighborhoods).