Predicting Airbnb Listing Prices in New York City

Lei Duli and Guo Jingyi
DATA 467 – Linear Regression

Research Question

This project seeks to answer the following question: *What factors most significantly influence the nightly price of Airbnb listings in New York City?* Specifically, we aim to build a multiple linear regression model to predict listing prices using variables such as room type, neighborhood, number of reviews, and availability. We hypothesize that room type and borough location will be the strongest predictors of price, with entire homes in Manhattan commanding the highest premiums.

Data Collection

The dataset is obtained from Kaggle's *New York City Airbnb Open Data* (https://www.kaggle.com/datasets/dgomonov/new-york-city-airbnb-open-data). The data was originally sourced from Inside Airbnb, a publicly available platform that collects listing information directly from the Airbnb website. The dataset contains 48,895 observations from the year 2019, covering all five boroughs of New York City. We did not collect this data ourselves; it is a pre-existing, publicly available dataset.

Description of Variables

The response variable is **price** (nightly rate in USD), a quantitative continuous variable. Typical values range from $69 (25th percentile) to $175 (75th percentile), with a median of $106. The predictor variables include a mixture of quantitative and categorical types:

- **Quantitative predictors:** minimum_nights (minimum stay requirement), number_of_reviews (total reviews received), reviews_per_month (review frequency), calculated_host_listings_count (number of listings the host operates), and availability_365 (days available per year).
- **Categorical predictors:** neighbourhood_group (five boroughs: Manhattan, Brooklyn, Queens, Bronx, Staten Island), room_type (Entire home/apt, Private room, Shared room), and neighbourhood (221 specific neighborhoods).

Relationships to Explore

We plan to explore how the combination of location, room type, and listing characteristics relate to price. We expect entire homes to be priced significantly higher than private or shared rooms, and listings in Manhattan to be more expensive than those in outer boroughs. We will also investigate whether hosts with more listings tend to price differently than single-listing hosts, and whether review activity serves as a proxy for demand that influences pricing.

Motivation

We are interested in this topic because Airbnb has fundamentally transformed the short-term rental market, and understanding the drivers of pricing has practical implications for both hosts and travelers. This dataset provides a rich opportunity to apply linear regression techniques to a real-world problem with a clear, interpretable response variable and a diverse set of predictors.