DSA210 Final Project Report – Nuray Ağca - 32574

Airbnb Listings and Annual Revenue Analysis

---

1. Introduction and Motivation

Airbnb is one of the most widely used platforms for short-term accommodation, allowing hosts to generate income through property rentals. Despite operating under the same platform, Airbnb listings vary greatly in terms of annual revenue. These differences may be influenced by listing characteristics such as property type, host status, user ratings, pricing, and availability.

The motivation of this project is to analyze which Airbnb listing features are associated with higher annual revenue and to understand how these relationships vary across different cities. By examining both statistical relationships and predictive modeling results, this project aims to provide insights into factors that contribute to higher-performing Airbnb listings.

---

2. Data Source and Collection

The dataset used in this project is a publicly available Airbnb listings dataset obtained from Kaggle:

https://www.kaggle.com/datasets/airbnb/airbnb-listings-data

The dataset includes detailed information about Airbnb listings, such as:

- Listing characteristics (property type, room type, amenities)
- Host attributes (Superhost status)
- Pricing and performance metrics (average daily rate, occupancy rate, annual revenue)
- User ratings and review counts
- Geographic information

For this project, CSV files from the following cities were combined into a single dataset:
Toronto, London, New York City, San Francisco, Miami, Tokyo, Sydney, Los Angeles, and Dubai.

All data used is publicly available and does not include personal or sensitive information.

---

3. Data Cleaning and Preparation

Several preprocessing steps were applied to prepare the dataset for analysis:

- Column names were standardized and converted to snake_case.
- Data types were corrected for numerical, categorical, boolean, and date variables.
- Duplicate records were removed.
- Listings with missing or invalid annual revenue values were excluded.
- Boolean variables such as Superhost status were standardized.
- A subset of relevant variables was selected to focus the analysis.

After cleaning, the final dataset contained **145,825 listings** with **36 variables**.

---

4. Feature Engineering and Enrichment

To improve interpretability and support later analysis, additional features were created:

- **Amenities count** to represent listing richness.
- **City-level median revenue**, used for relative comparisons.
- **Revenue-related ratios**, such as revenue per booking and revenue per available day.

These derived features allowed revenue comparisons across cities with different market scales and demand levels.

---

5. Exploratory Data Analysis (EDA)

Exploratory analysis revealed several key patterns:

- Annual revenue is highly right-skewed, with a small number of listings earning extremely high income.
- Median annual revenue differs substantially across cities, with cities such as Sydney and Tokyo showing higher median revenues than others.
- Entire home listings tend to generate higher revenue compared to private room listings.

Visualizations were used to examine revenue distributions and city-level comparisons.

---

6. Hypothesis Testing

Three hypotheses were tested using non-parametric methods due to non-normal revenue distributions:

H1: Superhost Effect

H□: There is no difference in annual revenue between Superhost and non-Superhost listings.
H□: Superhost listings generate higher annual revenue.

A Mann–Whitney U test showed that Superhost listings have significantly higher median annual revenue than non-Superhost listings ($p < 0.05$).

H2: Property Type Effect

H□: Entire home and private room listings have equal annual revenue.
H□: Entire home listings generate higher annual revenue.

The Mann–Whitney U test indicated that entire home listings earn significantly more revenue than private room listings ($p < 0.05$).

H3: Rating and Revenue Relationship

H□: There is no relationship between user ratings and annual revenue.
H□: User ratings are related to annual revenue.

Spearman correlation analysis showed a statistically significant but weak positive relationship between overall rating and annual revenue.

---

7. Machine Learning Analysis

A supervised learning approach was applied to predict annual Airbnb revenue.

- The target variable (annual revenue) was log-transformed to address skewness.
- The dataset was split into training and test sets.
- A preprocessing pipeline handled missing values and categorical variables.
- A Linear Regression model was trained and evaluated.

The model achieved a strong R^2 score, indicating that listing characteristics explain a substantial portion of the variation in annual revenue. Despite its simplicity, the model provided interpretable results consistent with the statistical findings.

---

8. Key Findings

- Superhost status is associated with higher annual revenue.
- Entire home listings generate more revenue than private room listings.
- User ratings are statistically related to revenue, although the effect size is small.
- Revenue patterns vary significantly across cities.
- Listing features collectively provide meaningful predictive power for revenue estimation.

---

9. Limitations and Future Work

This project has several limitations:

- Only a linear regression model was used due to computational constraints.
- External datasets such as tourism demand or economic indicators were not included.
- Seasonal or time-based revenue variations were not analyzed.

Future work could explore more complex models, incorporate external data sources, or analyze temporal trends in revenue.

---

10. Ethical Considerations and AI Usage

All data used in this project is publicly available and ethically appropriate.
AI tools such as ChatGPT were used solely for writing assistance and improving clarity. All analysis, coding, and interpretation decisions were made by the author.