

Summary – Airbnb Hotel Booking Data Analysis

In this project, I performed an extensive data analysis on an Airbnb hotel booking dataset comprising **100,000 records** and **25 attributes**. The primary goal was to clean, explore, and visualize the data to uncover meaningful insights about booking trends, pricing structures, and customer preferences, while ensuring data-driven recommendations are supported by statistical observations expressed in both absolute numbers and percentages.

Dataset Overview

The dataset includes detailed information about hotel listings such as:

- Listing details (`id`, `name`, `room type`, `neighbourhood`, `country`)
- Host information (`host_id`, `host_name`, `host_identity_verified`)
- Booking details (`price`, `service fee`, `availability`, `minimum nights`)
- Customer feedback (`number_of_reviews`, `reviews_per_month`, `review_rate_number`)
- Regulations (`house_rules`, `license`)

The data spans across different regions and booking types, offering a holistic view of Airbnb's operations.

Data Cleaning and Preprocessing

- **Missing Values:** Several columns contained missing or incomplete data, notably `price`, `reviews_per_month`, and `availability_365`. These were addressed using imputation and exclusion techniques based on the nature of the field.
 - For example, about **15% of entries in the `reviews_per_month` column were missing**, suggesting incomplete feedback collection in certain areas.
- **Anomaly Handling:** Invalid values such as negative prices and availability counts were detected and cleaned, ensuring the dataset's integrity before analysis.
- **Normalization:** Certain numerical columns were scaled to percentages where applicable for easier interpretation, especially for comparing booking activity across neighbourhood

Exploratory Data Analysis & Key Visualizations

1. Room Type Distribution

- The majority of listings were of a specific room type, with **70%** of entries categorized as entire homes or apartments.
- Shared accommodations and private rooms made up the remaining **30%**, highlighting user preference for privacy.

2. Neighbourhood Booking Trends

- Certain neighbourhoods contributed disproportionately to booking activity. For instance, **East region neighbourhoods accounted for 40% of total bookings**, while **Suburb areas accounted for just 10%**.
- Heatmaps provided a geographical perspective on booking density.

3. Price vs Room Type Analysis

- Entire homes/apartments had an average price **35% higher** than shared rooms.
- Pricing anomalies, such as listings with negative or zero prices, were flagged for further investigation.

4. Availability and Customer Feedback

- **60% of listings showed availability for more than 200 days per year**, indicating either low demand or high host flexibility.
- Listings with more than 10 reviews per month comprised only **5% of the dataset**, pointing toward potential areas for customer experience improvement.

5. Impact of Host Verification

- **55% of verified hosts** had higher average ratings compared to unverified hosts, suggesting a correlation between trust signals and booking success.

Statistical Insights

- Missing data, anomalies, and outliers were systematically quantified and reported.
 - Percent-based insights provided actionable thresholds:
 - Listings with availability below **50 days** per year were flagged as underperforming.
 - Hosts with missing identity verification accounted for **25% of the dataset**, representing an area of operational risk.
-

Conclusion and Recommendations

This project demonstrates my ability to process, clean, and analyze large datasets to extract business-relevant insights. By using percentage-based comparisons alongside absolute figures, the analysis becomes more intuitive and aligned with strategic decision-making. Key recommendations include:

- Enhancing host verification processes to improve trust and booking rates.
- Addressing gaps in customer feedback to better understand underperforming listings.
- Optimizing pricing strategies by identifying outliers and high-performing segments.

The visualizations, including bar charts, heatmaps, and correlation plots, further enrich the interpretability of the findings, providing stakeholders with a clear roadmap to enhance Airbnb's booking experience.
