Business Report for AIRBNB

Submitted By: Navin Kumar V

Course: Data science and analytics

Instructor: Dinesh Kumar

Date:

1. Executive Summary:

This report examines an Airbnb dataset of 60,209 listings, concentrating on important factors such room type, number of beds, price, and review scores. The purpose is to identify pricing patterns, factors that influence listing ratings, and insights into booking preferences.

It can be noted that,

- The dataset is dominated by entire home or apartment listings.
- Many entries lack review scores.

2. Introduction:

Problem Statement:

Understanding what influences Airbnb listing prices and ratings can help hosts optimize their offers and increase revenue.

Objective:

This analysis aims to explore key insights into room types, pricing patterns, review scores, and booking policies by doing various visual representations between different variables and finding patterns or factors affecting the variables.

Dataset Description:

The dataset consists of 60,209 Airbnb listings with attributes like

- id (int) – Unique room id
- room type (string) – Describes the type of room
- accommodates (int) – Number of people per room
- bathroom (int) – Number of bathrooms in the room
- cancellation policy (string) – Severity of cancellation policy (Strict, moderate, flexible)
- cleaning_fee (string) – Whether there is a fee for cleaning (True/False)

- instant_bookable (string) – Whether the room can be booked instantly (T/F)
- review_scores_rating (int) – rating provided by the users about the room out of 100
- bedrooms (int) – Number of bedrooms available in the room
- beds (int) – Number of beds provided
- log_price (float) – Log transformed price values

Missing values are present in several fields, particularly review scores.

3. Data Preparation:

Data Cleaning:

Handling Missing values:

- Review_scores_rating column has the most number of missing values of nearly 22% (13,601). We'll fill the missing values with the mean value.
- The rest of the columns room_type, accommodates, bathrooms, cancellation_policy, cleaning_fee, bedrooms, beds all have missing values less than 0.3% and hence they are dropped.

Duplicate Values:

- There are no duplicates in the dataset

4. Exploratory Data Analysis (EDA):

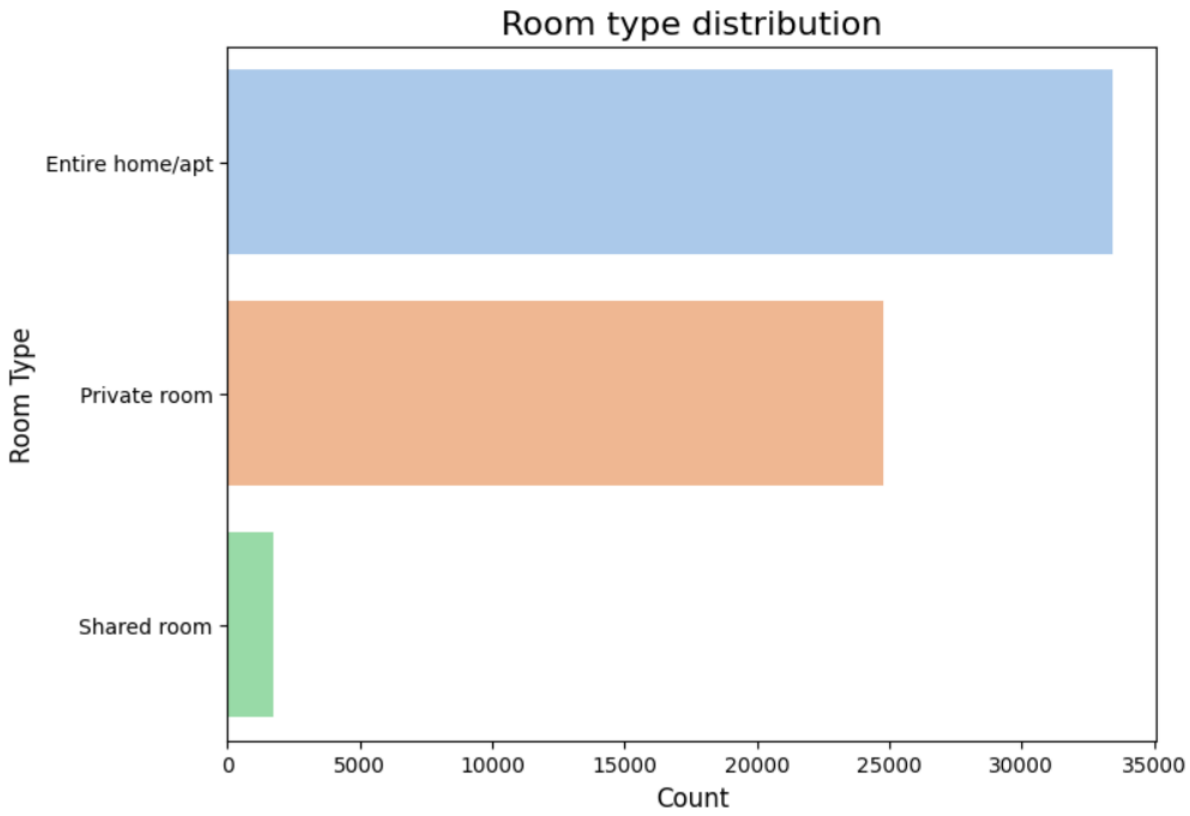
The 10 essential steps of EDA are performed:

- Data import
- Data Inspection
- Data dimension
- Data information
- Data summarization (Statistical)
- Data types
- Data Columns
- Check missing values

- Check duplicates
- Data distribution

Univariate Analysis:

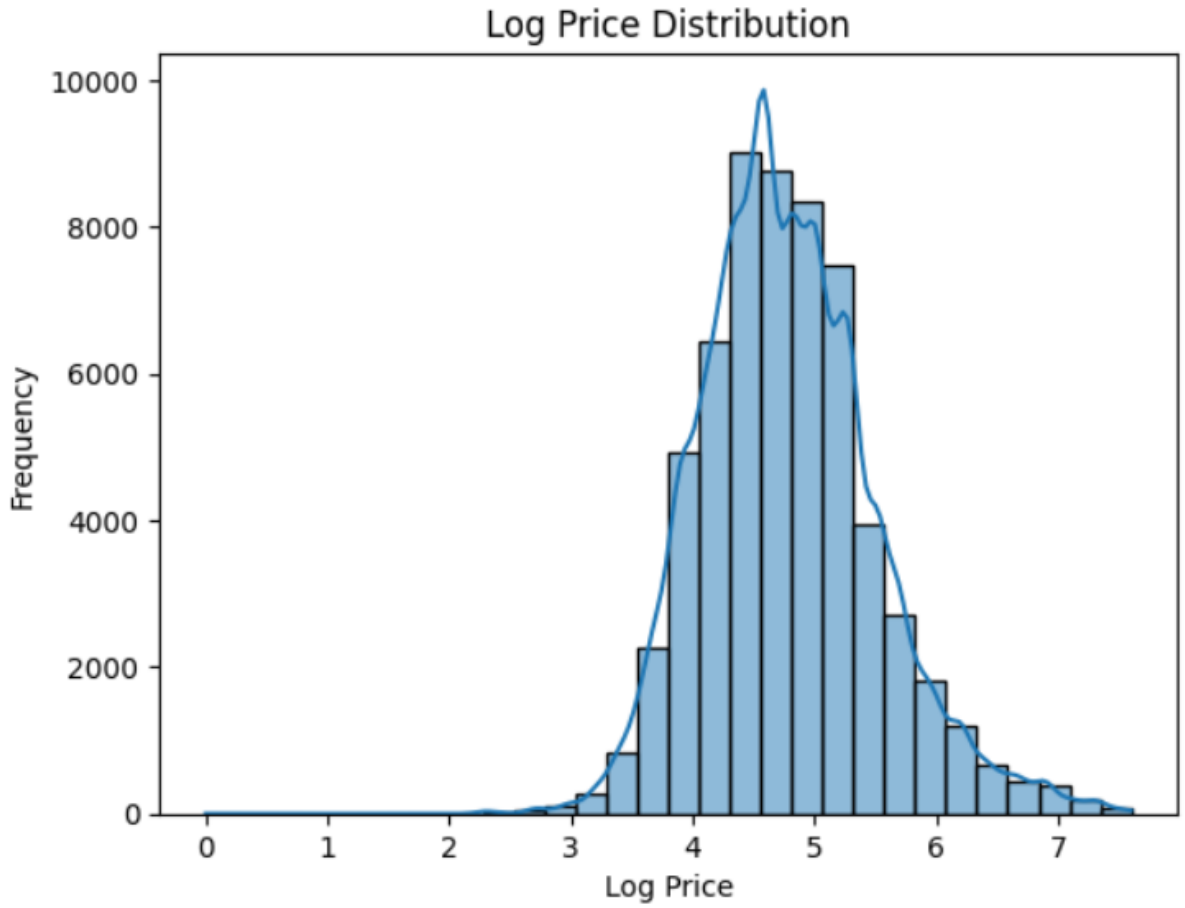
- **The Room type distribution is visualised using Count Plot.**



Inference:

- Most listings are likely Entire Home/Apt.
- Shared room is available only in small numbers.

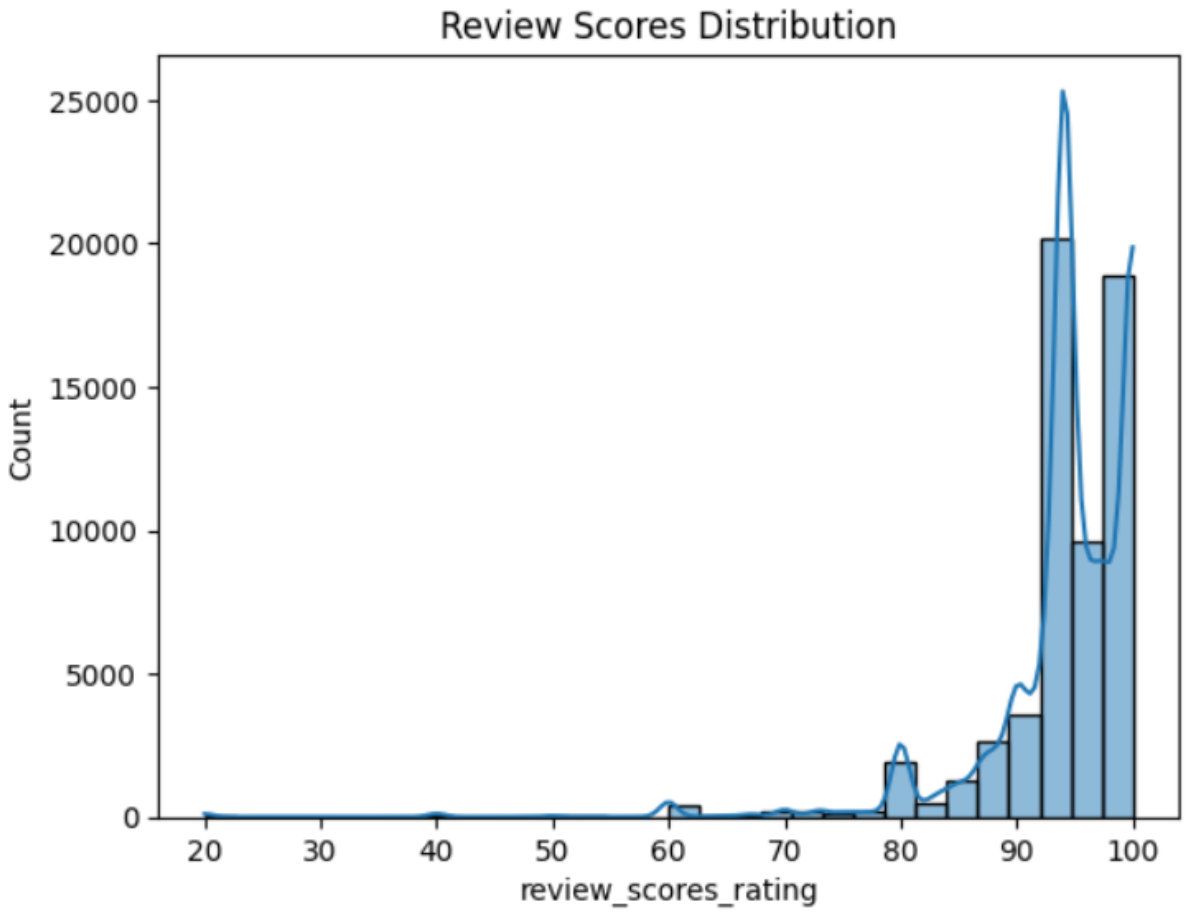
- **Price Distribution is visualized using Histogram**



Inference:

- The histogram follows a bell-shaped curve, indicating a normal distribution of log-transformed prices.
- The highest frequency of listings falls around log price 4.5 to 5.5.
- Some listings have log price values above 6 and 7, suggesting high-end or luxury properties.
- Very few listings have a log price below 3, meaning budget listings are rare in the dataset.

- **Review scores distribution using Histogram**

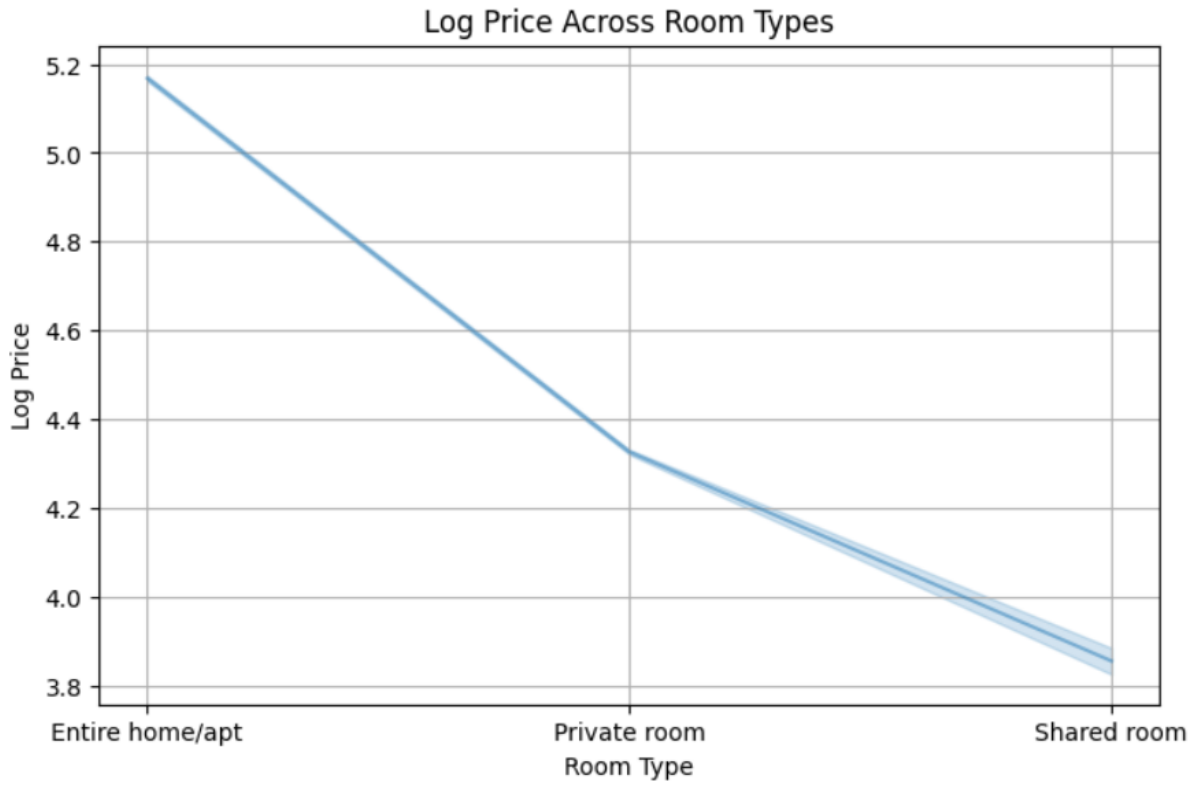


Inference:

- We see a left-skewed distribution, indicating that most of the guests liked the rooms and gave high ratings.
- Rooms with less rating indicate that the customers were dissatisfied with the rooms.

Bi-Variate / Multi-Variate Analysis:

- **Line chart of Price vs Room Type**



Inference:

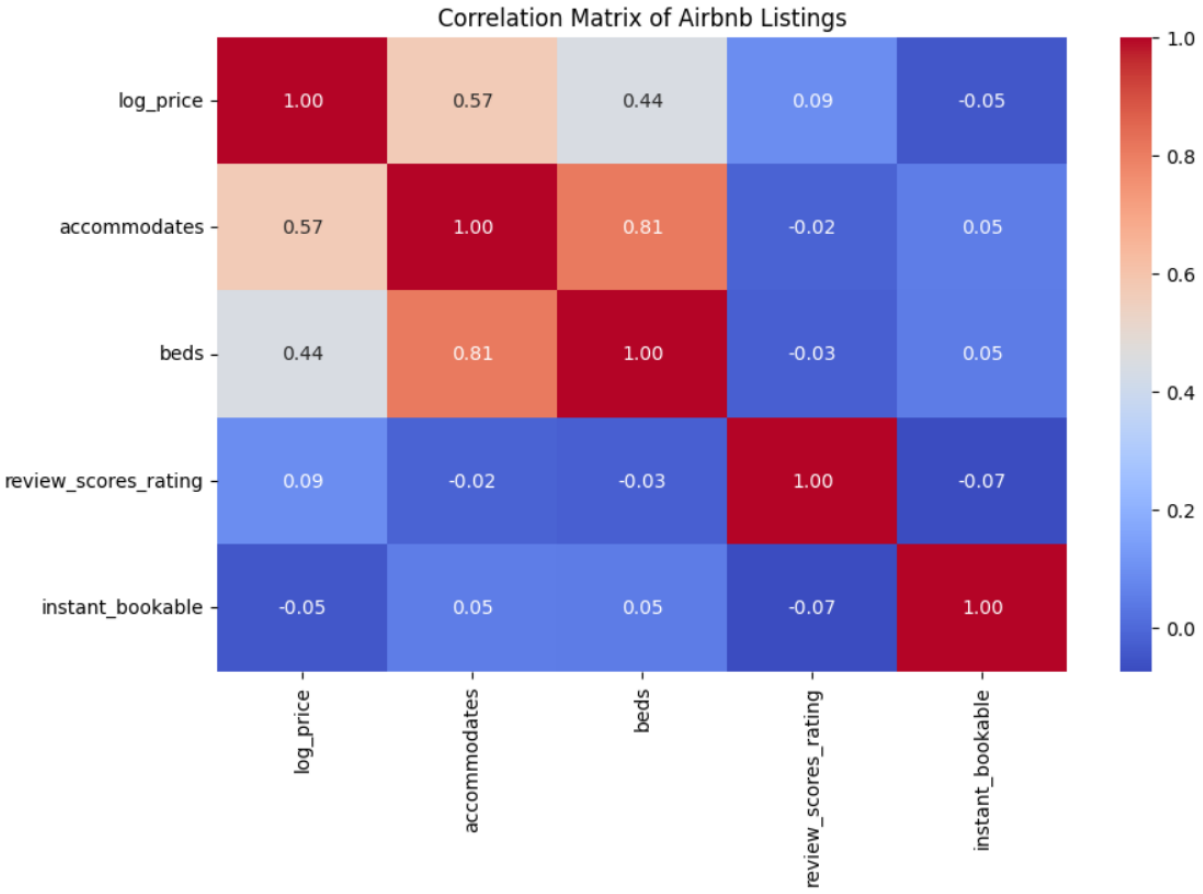
- Entire homes/apartments have the highest log price, followed by private rooms, and shared rooms have the lowest.
- This suggests that the level of privacy and exclusivity significantly influences price.

- **Scatter Plot of Price vs Review score rating**

Inference:

- High ratings are common, but they don't necessarily mean a high price.
- Some expensive properties maintain high ratings, but price alone doesn't guarantee better reviews.
- The presence of clustered values indicates filling with mean of review score.

- **Heat Map – Correlation Matrix**



Values close to +1 → Strong positive correlation

Values close to -1 → Strong negative correlation

Values close to 0 → No significant correlation

Inference:

Log Price vs. Accommodates (Moderate to High Positive Correlation)

- Listings that accommodate more guests tend to have higher prices.

Log Price vs. Number of Beds (Moderate Positive Correlation)

- More beds generally lead to higher prices, but this correlation might be weaker than accommodates.

Log Price vs. Review Scores Rating (Weak Positive Correlation)

- Higher-rated properties tend to charge slightly more.

Accommodates vs. Beds (High Positive Correlation)

- Naturally, listings with more beds also accommodate more guests.

Instant Bookable vs. Review Scores (Weak or No Correlation)

- If instant booking is not strongly correlated with review scores, this suggests that customers do not necessarily prefer instant booking properties over non-instant ones.

5. Results and Interpretation

Key Findings

- **Room Type Distribution:**
  - The dataset is dominated by Entire Home/Apt listings.
  - Shared rooms are rare, suggesting lower demand for shared spaces.
- **Price Distribution:**
  - The log price follows a normal distribution, with most listings in the range of 4.5 to 5.5.
  - Few listings have exceptionally high prices, indicating luxury properties.
  - Budget listings (log price below 3) are uncommon.
- **Review Scores Distribution:**
  - The left-skewed distribution indicates that most guests give high ratings.
  - Rooms with lower ratings suggest customer dissatisfaction in some cases.
- **Price vs. Room Type:**
  - Entire homes/apartments have the highest log price, followed by private rooms, and then shared rooms.
  - This confirms that privacy and exclusivity drive pricing.

- **Price vs. Review Scores:**
    - Higher review scores do not necessarily mean higher prices.
    - Some expensive properties maintain high ratings, but price alone doesn't guarantee good reviews.
- **Correlation Matrix (Heatmap) Analysis:**
    - Log Price vs. Accommodates (Moderate-High Positive Correlation): Listings that accommodate more guests tend to be priced higher.
    - Log Price vs. Beds (Moderate Positive Correlation): More beds generally lead to higher prices, though slightly weaker than accommodates.
    - Log Price vs. Review Scores (Weak Positive Correlation): Higher-rated properties charge slightly more, but not significantly.
    - Instant Bookable vs. Review Scores (Weak or No Correlation): Customers do not necessarily prefer instant bookable properties over non-instant ones.

6. Recommendations:

1. Pricing Strategy for Hosts:

- Entire home/apartments can charge a premium due to higher demand.
- Hosts should optimize pricing based on room type and accommodates to maximize revenue.

2. Improving Review Scores:

- While price and ratings are weakly correlated, maintaining a high review score can still increase demand.
- Hosts should focus on guest experience and cleanliness to enhance ratings.

3. Targeting Budget Travelers:

- Since low-cost listings are rare, there is an opportunity to introduce more budget-friendly accommodations.

4. Leveraging Instant Booking:

- Since instant booking has no strong correlation with review scores, hosts may use it as a competitive advantage to increase bookings without affecting ratings.

5. **Luxury Property Positioning:**
    - High-end properties with log price above 6-7 need better marketing as they cater to a niche market.

7. Conclusion

Summary of Analysis

- The analysis explored pricing trends, room type distribution, review scores, and booking preferences among 60,209 Airbnb listings.
- Privacy and capacity significantly impact pricing, while review scores have a weaker effect.
- The data suggests opportunities for optimizing pricing strategies, improving guest experiences, and targeting specific market segments.

Limitations

- **Missing Data:** The review_scores_rating column had 22% missing values, which were filled using the mean, potentially reducing data accuracy.
- **Lack of Location Data:** The dataset does not analyze how location influences price and reviews, which is a crucial factor in Airbnb bookings.
- **Possible Data Bias:** The dataset may not be representative of all Airbnb listings globally, affecting generalizability.

Next Steps

- Include Geographic Data Analysis to explore how neighborhood and location impact pricing and reviews.
- Time-Series Analysis to determine if prices fluctuate seasonally or based on demand patterns.
- Sentiment Analysis on Guest Reviews to extract qualitative insights about what guests like or dislike about listings.
- Advanced Predictive Modeling (e.g., machine learning) to predict optimal pricing strategies for hosts.