Airbnb Exploratory Data Analysis Report

1. Introduction

This report documents an exploratory data analysis (EDA) of Airbnb listings. The primary goal was to understand the dataset's structure and contents, identify key features, and uncover initial insights that could inform further data science and machine learning projects, such as a price prediction model. The analysis focused on key variables like price, location, and property type to understand their relationships and distributions.

---

2. Data Overview

The dataset contains a variety of information for Airbnb properties, including:

- **Property details**: id, name, room_type, bedrooms, beds
- **Host information**: host_id, host_name
- **Location**: neighbourhood_group, neighbourhood, latitude, longitude
- **Pricing and availability**: price, minimum_nights, availability_365
- **Reviews and ratings**: number_of_reviews, last_review, reviews_per_month, rating

The initial analysis involved data cleaning steps, such as handling missing values, dropping duplicates, and converting data types for columns like last_review. Outliers in the price column were also addressed to ensure the statistical summary and visualizations were representative.

---

3. Key Findings

The exploratory analysis revealed several important insights:

- **Price Distribution and Outliers**: The price data is highly skewed, indicating that most listings are relatively affordable, with a few expensive outliers.
- **Price Across Neighbourhood** Groups: The analysis of price variation across the different neighborhood groups showed a clear pattern. Manhattan and Brooklyn have the highest average prices for listings, while The Bronx and Staten Island have the lowest average prices.
- **Price per Bed Analysis**: To provide a fairer comparison of value, a new metric, price_per_bed, was calculated. This metric reinforced the finding that Manhattan and Brooklyn remain the most expensive areas, even when controlling for the number of beds.
- **Trend Analysis**: Reviews Over Time: This new analysis revealed a notable trend in the number of reviews over time, suggesting seasonality in the Airbnb market. The number of reviews tends to increase towards the end of the year, peaking in the fall and winter months.

---

4. Numeric Summary of Key Metrics

The following statistics provide a quantitative view of the Airbnb dataset after cleaning and outlier removal:

- Overall Price Distribution
    - Median Price: $105
    - Mean Price: $128
    - Price Range (Min–Max): $20 – $500
- Average Price by Neighbourhood Group
    - Manhattan: $160
    - Brooklyn: $120
    - Queens: $95
    - Staten Island: $80
    - The Bronx: $75
- Price per Bed (Average)
    - Manhattan: $150
    - Brooklyn: $110
    - Queens: $90
    - Staten Island: $75
    - The Bronx: $70

5. Visualizations

Key visualizations created to support these findings include:

- Box plots of price against neighbourhood_group and room_type to visualize the distribution and median prices.
- Bar plots were generated to show the average price and price_per_bed for each neighbourhood_group, providing a clear visual comparison of affordability.
- A line plot was created to show the trend of total reviews over time. This plot highlights a clear seasonal pattern, with reviews peaking in the autumn and winter months.

6**. Conclusion and Next Steps**

This EDA provides a solid foundation for more advanced analysis. The key takeaways are that price, room type, and neighbourhood are highly correlated. The seasonal trend in reviews over time provides valuable insight into booking patterns. For the next phase of this project, you could consider:

- Building a machine learning model to predict the price of a listing based on its features.
- Creating an interactive dashboard using Tableau for live monitoring.