Airbnb Data Analysis Project

Project Title

Airbnb Data Analysis Project

Introduction

This project focuses on analyzing a large dataset of Airbnb listings primarily from New York City. The goal is to uncover trends and insights about listing characteristics, pricing, neighborhood popularity, and temporal patterns of user engagement through reviews. The dataset includes diverse features such as listing details, host information, prices, review metrics, and geographical attributes. This analysis helps understand the Airbnb market dynamics and provides valuable insights into pricing, room types, neighborhood impact, and host activities.

Objectives

- To clean and preprocess Airbnb data for accurate analysis.
- To understand the distribution of listing prices and room types.
- To analyze how Airbnb listings are distributed across different neighborhoods.
- To explore the relationship between price and room type.
- To observe trends in reviews over time to understand user engagement and listing popularity.
- To derive actionable insights about Airbnb marketplace characteristics in New York.

Data Description

The dataset contains 102,599 Airbnb listings with 26 columns capturing various attributes including:

- **Listing Details:** id, name, neighborhood group, neighborhood, latitude, longitude, country, room type, construction year, price, service fee, minimum nights.
- **Host Information:** host id, host name, host identity verification status, calculated host listings count.

- **Review Metrics:** number of reviews, last review date, reviews per month, review rate number.
- **Other Attributes:** instant bookable status, cancellation policy, availability days, house rules, license status (mostly missing).

Several columns contained missing values or inconsistent data types that were handled during preprocessing.

Methodology

- **Data Cleaning:**
    - Loaded data and identified columns with missing values and mixed data types.
    - Converted key columns like 'last review' to datetime.
    - Filled missing review-related fields with appropriate values (e.g., zero for reviews per month).
    - Dropped rows missing critical data such as listing name or host name.
    - Removed irrelevant columns like 'house_rules' and 'license' due to excessive missingness.
    - Corrected data types, particularly removing dollar signs for price and service fee columns and converting them to float.
    - Removed duplicate records for uniqueness.
- **Exploratory Data Analysis (EDA):**
    - Used statistical summaries to understand central tendencies and dispersions.
    - Visualized distributions of prices and room types using histograms and count plots.
    - Examined spatial distribution of listings across neighborhood groups.
    - Created box plots to investigate price variations across different room types.
    - Analyzed temporal patterns in review counts to detect trends in user engagement over time.

Insights

- Listing prices span a wide range with no dominant concentration in any specific price bracket, indicating diverse pricing strategies.
- The majority of listings are either "Entire home/apt" or "Private room", with fewer listings classified as "Shared room" or "Hotel room".
- Most Airbnb listings are concentrated in Manhattan and Brooklyn, highlighting their popularity and potential demand; Queens, Bronx, and Staten Island have fewer listings.
- Price differs notably by room type: shared rooms tend to be the least expensive, while private rooms and entire homes/apartments show greater variation and higher prices.
- Reviews over time exhibit distinct fluctuations, suggesting changes in market activity or external influences impacting review frequency. Peaks and dips may coincide with seasonality, policy changes, or other events affecting Airbnb usage.

Visualizations Summary

- **Price Distribution:** Histogram with KDE line showing an even spread of listing prices.
- **Room Type Distribution:** Count plot revealing dominance of entire homes and private rooms.
- **Neighborhood Group Distribution:** Count plot highlighting higher listing numbers in Manhattan and Brooklyn.
- **Price vs. Room Type:** Box plot showing price variation by accommodation type and illustrating affordability differences.
- **Reviews Over Time:** Line plot indicating trends and periodic changes in review frequency, useful for detecting seasonality or external impacts on listing popularity.

Conclusion

The Airbnb Data Analysis Project provides comprehensive insights into the New York City Airbnb market through data-driven exploration. Data cleaning and preprocessing ensured quality and reliability of analysis. The study reveals pricing diversity, room type popularity, and neighborhood concentration trends that can inform hosts, guests, and platform managers. Visualization of review trends over time adds temporal context to user engagement patterns. This foundational analysis can be extended with predictive modeling or geospatial analysis for deeper market understanding and strategic decision-making.