New York City Hotels Data Analysis

Problem Statement

With the rapid growth of short-term rental platforms like Airbnb, understanding the factors that influence listing performance has become crucial for both hosts and platform stakeholders. However, the vast amount of data generated by listings—such as pricing, availability, location, reviews, and host attributes—remains underutilized without proper analysis.

This project aims to address the following key problems:

- What are the major factors affecting the price of an Airbnb listing?
- How does location (neighbourhood and neighbourhood group) impact availability and demand?
- Is there a correlation between the number of reviews and host activity or listing quality?
- How do room types and amenities influence booking trends?

By analysing these variables, the project seeks to uncover patterns and relationships that can guide pricing strategies, improve customer satisfaction, and enhance operational decision-making on the Airbnb platform.

Introduction

In recent years, platforms like Airbnb have revolutionized the short-term rental market, providing travellers with a wide range of accommodation options and hosts with new income opportunities. This project focuses on analysing a dataset of Airbnb listings, with the objective of uncovering key insights into host behaviour, property availability, pricing strategies, and customer engagement.

Using data that includes listing details such as location, room type, price, availability, number of reviews, and host information, this analysis aims to answer important business and operational questions. For example: What factors influence the price of a listing? How does availability vary across neighbourhoods? What is the relationship between host activity and guest satisfaction?

By exploring these dimensions, the project not only helps in understanding the dynamics of Airbnb's rental ecosystem but also provides valuable insights for hosts, travellers, and platform stakeholders to make informed decisions. The ultimate goal is to transform raw data into actionable intelligence using techniques such as data cleaning, visualization, and statistical analysis.

Description about the data

Overview of the data

id: A unique identifier for each listing.
name: The title or name of the hotel listed by the host.
host_id: A unique identifier assigned to the host who listed the hotel.
host_name: The name of the person who hosts the hotel.
neighbourhood_group: The broader area or region where the property is located, such as a city district or zone.
neighbourhood: The specific neighborhood within the broader region where the property is located.
latitude: The latitude coordinate of the hotel's location, used for mapping and geographic analysis.
longitude: The longitude coordinate of the hotel's location, used for mapping and geographic analysis.
room_type: The type of room offered in the listing, such as "Entire home/apt," "Private room," or "Shared room."
price: The cost per night for staying at the property in the local currency, which is in dollars.
minimum_nights: The minimum number of nights a guest is required to book the property for a stay.
number_of_reviews: The total number of reviews the hotel has received from guests.
last_review: The date of the most recent review received by the listing.
reviews_per_month: The average number of reviews the hotel receives per month.
calculated_host_listings_count: The total number of active listings the host has on the platform.
availability_365: The number of days the hotel is available for booking within a year (365 days).
number_of_reviews_ltm: The number of reviews received over the last 12 months, also known as "last twelve months" (LTM) reviews.
license: The property's registration or license number, if required by local authorities.
rating: The overall rating of the hotel, often based on guest reviews.
bedrooms: The number of bedrooms available in the hotel.
beds: The number of beds available in the hotel.
baths: The number of bathrooms available in the hotel.

Data Cleaning and Transformation

1. **Remove Duplicates**
    - Check for and remove any duplicate rows based on the id or combination of all columns to ensure data integrity.
2. **Handle Missing Values**
    - Identify and deal with missing or null values in important columns like:
        - name (fill with "No name" or drop if critical)
        - last_review (fill with "No Review" or NaN)
        - reviews_per_month (fill with 0 where number_of_reviews is 0)
        - license, baths, bedrooms, beds (fill with "Not specified" or appropriate default values)
3. **Convert Data Types**
    - Ensure each column has the correct data type:
        - price, minimum_nights, number_of_reviews, etc. → integer or float
        - last_review → convert to datetime format
        - reviews_per_month → float
        - availability_365, number_of_reviews_ltm → integer
4. **Clean Special Characters from Text**
    - Remove special characters (like ¬∑ ,òÖ) from the name field for readability and analysis.
5. **Handle Inconsistent or Invalid Data**
    - Check for unrealistic values:
        - price <= 0 or extremely high → flag or remove
        - minimum_nights <= 0 or abnormally high (e.g., 365+) → correct or remove
        - availability_365 > 365 → correct
        - bedrooms, beds, baths should be numeric where applicable
6. **Standardize Categorical Columns**
    - Ensure consistent formatting in:
        - room_type (e.g., "Private room", not "private Room")

        - neighbourhood_group, neighbourhood, license → strip spaces, standardize capitalization

7. **Create Additional Useful Columns (Optional)**
    - Add derived columns such as:
        - review_year or days_since_last_review
        - price_per_bedroom if both price and bedrooms are available
8. **Binary Encoding for License**
    - The license column was converted into a binary format:
        - 1 for listings **with a license**.
        - 0 for listings **without a license**.
    - This transformation simplifies analysis and modeling.
9. **Outlier Detection**
    - Outliers were detected in the following columns:
        - price, beds, baths, bedrooms, and bathrooms.
    - These outliers can skew statistical measures and were visualized to assess their impact on the data.
10. **Handling Zero Ratings**
    - Zero entries in the rating column were identified.
    - Steps:
        - Counted the number of 0 ratings.
        - Calculated the **mean of the rating** column, excluding zero values.
        - Replaced 0s with the **mean rating**.
        - Rounded the ratings to **two decimal places** for consistency.
11. **Exporting Cleaned Data**
    - The cleaned dataset was saved as **Hotels_data1.csv**.
12. **Statistical Queries**
    - Counted hotels with perfect **5.00 rating**.
    - Identified hotel(s) with:
        - The **highest number of reviews**.
        - The **maximum price** for monthly accommodation.

Data Visualization and Insights

1. **Scatter Plot: Price vs. Number of Reviews**
    - Showed an **inverse relationship — lower-priced listings tend to receive more reviews**, suggesting affordability attracts more guests.
2. **Minimum Nights Analysis**
    - Retrieved **unique values** of minimum_nights.
    - Identified the **maximum value** and the **hotel name** associated with it.
    - Sorted and extracted the **top 10% listings** based on minimum stay requirement.
    - This highlights long-term stay listings which could be commercial rentals.
3. **Joint Plot: Price vs. Availability**
    - Used sns.jointplot to visualize the relationship between **price** and **availability_365**, color-coded by rating.
    - Helped understand the density of listings and their price-availability distribution.
4. **Histogram of Price Distribution (Log-Log Scale)**
    - Helped to:
        - **Identify price clusters** and **outliers**.
        - Understand if the market is dominated by budget or luxury accommodations.
        - Assess **data spread** using log-scale.
5. **Histograms for Column-Wise Distribution**
    - Individual histograms for key features (e.g., beds, price, bedrooms) were plotted to:
        - Understand **distribution patterns**.
        - Detect **skewness** and irregularities in the data.

Categorical Data Exploration

1. **License Distribution**
    - Retrieved **unique values** and plotted a **countplot** for licensed vs unlicensed hotels.
2. **Neighbourhood Group Analysis**
    - Extracted unique values.
    - Countplot revealed **distribution of listings** across different neighbourhood groups.
3. **Room Type Analysis**
    - Countplot showed the **frequency of room types** (e.g., private room, entire home).
    - Line plot displayed **average price trends** per room type and number of bedrooms.
4. **Box Plot: Availability vs Neighbourhood Group**
    - Boxplot illustrated the **spread of availability** across neighbourhoods.

Rating-Based Insights

1. **5-Star Listings**
    - Extracted all rows with a perfect **5.0 rating** into a new dataframe (df_rating).
    - Countplot for ratings helped visualize **distribution of listing quality**.

Advanced Visualizations using Plotly

1. **Bar Graph: Most Popular Neighbourhood**
    - Used plotly to create an interactive bar chart highlighting **neighbourhoods with the most listings**.
2. **Scatter + Histogram Combination Plot**

- Combined visualization helped analyze **price vs rating** distribution along with their histograms.

3. **Hotel Location Map**
    - Scatter plot using longitude and latitude to display **hotel locations**, color-coded by availability_365.
4. **Line Chart: Room Type vs Rating**
    - Line chart showed **how ratings vary across different room types**, providing insight into user preferences.
5. **Scatter Plots**
    - **Last Review vs Rating**
    - **Last Review vs Room Type**
    - Helped understand **recency of activity and guest satisfaction**.

Heatmap for Feature Correlation (Kendall Method)

- The **heatmap** visualized **correlation coefficients** between numerical features.
- Key insights:
    - Strong correlation between:
        - minimum_nights and license
        - reviews_per_month and number_of_reviews_ltm
        - beds and bedrooms
        - number_of_reviews and number_of_reviews_ltm
- Helped detect **redundancies**, **dependencies**, and **potential predictors** for modeling.

Learning Outcomes

By completing this data analysis project on Airbnb listings, the following key learnings were achieved:

1. **Data Cleaning & Preprocessing**
    - Understood the importance of handling missing values, encoding categorical variables, and removing outliers to improve data quality.
2. **Data Transformation Techniques**
    - Learned how to convert textual and boolean fields (e.g., license status) into numeric formats for ease of analysis.
3. **Exploratory Data Analysis (EDA)**
    - Gained experience in visualizing data using **matplotlib**, **seaborn**, and **plotly** to uncover patterns and relationships.
4. **Statistical Insights Extraction**
    - Developed the ability to compute meaningful metrics such as average ratings, maximum reviews, and price distributions to draw business conclusions.
5. **Correlation Analysis**
    - Understood how features interact using **correlation matrices** and **heatmaps**, helping to detect redundancies and associations in the dataset.
6. **Advanced Visualization**
    - Learned how to use joint plots, histograms, boxplots, and geographic plots to identify trends and customer behavior.
7. **Real-World Application**
    - Applied data science concepts to a real-world dataset, gaining practical experience in data-driven decision-making.

Future Scope

This project lays the foundation for more advanced work and offers multiple future directions:

1. **Predictive Modeling**
    - Build machine learning models to predict **hotel price**, **rating**, or **occupancy rate** based on features like room type, location, and number of reviews.
2. **Time Series Analysis**
    - Analyze **seasonal trends** using the last_review column to determine peak and off-peak periods in different neighborhoods.
3. **Geospatial Analysis**
    - Integrate with **GIS data** to offer more advanced location-based insights, like proximity to landmarks or transport hubs.
4. **Customer Sentiment Analysis**
    - Incorporate **review text data** (if available) to perform **natural language processing (NLP)** for extracting guest sentiment.
5. **Host Behavior Profiling**
    - Study hosts based on their listings, reviews, availability, and ratings to identify **top-performing hosts**.
6. **Regulatory Impact Analysis**
    - Examine how licensing and minimum nights policies affect listing performance across neighborhoods.

‘

Conclusion

This project provided a comprehensive understanding of how data analysis can uncover valuable insights in the hospitality industry. By cleaning, transforming, and analyzing the Airbnb dataset, we identified patterns in pricing, guest behavior, and availability that can benefit hosts, guests, and the platform itself.

Through effective use of visual tools and statistical methods, this analysis revealed:

- The influence of price on guest reviews,
- Patterns in room types and availability across neighborhoods,
- The impact of licensing and minimum stay policies on host behavior.

These findings not only help improve decision-making for current operations but also lay the groundwork for building intelligent systems that can predict, recommend, and optimize Airbnb experiences in the future.