AIRBNB DATA ANALYSIS

SUMMARY

This project performs an exploratory data analysis (EDA) of Airbnb listings in New York City. The analysis focuses on understanding price trends, host behaviors, and neighborhood popularity to derive insights from the hospitality market.

Project Objectives

The primary goal is to clean and analyze a large dataset (over 100,000 entries) to identify key factors influencing Airbnb listings. Specifically, the notebook aims to:

- Cleanse the data by handling missing values and incorrect data types.
- Analyze pricing and service fee structures.
- Identify popular neighborhoods and room types.
- Visualize trends, such as reviews over time.

Data Cleaning & Preprocessing

The project involves significant data preparation steps to ensure analysis accuracy:

- **Handling Missing Values:** The analyst identified thousands of missing values in columns like `last review` (15,893) and `house_rules` (52,131).
- **Imputation:** Missing `reviews per month` were filled with 0, and missing `last review` dates were filled with the minimum date in the dataset.
- **Row Removal:** Listings missing critical information like the listing `NAME` or `host name` were dropped.
- **Feature Removal:** Highly sparse columns like `license` and `house_rules` were removed entirely to simplify the model.
- **Type Conversion:** The `last review` column was converted to a standard datetime format for chronological analysis.

Exploratory Data Analysis (EDA)

The analysis covers several dimensions of the Airbnb market:

- **Price Analysis:** The notebook explores the distribution of listing prices and associated service fees.
- **Host Verification:** It examines the count of verified vs. unconfirmed host identities.
- **Geographic Trends:** Analysis is broken down by `neighbourhood group` (Boroughs like Manhattan, Brooklyn, etc.) and specific `neighbourhoods`.

- **Listing Attributes:** The project reviews room types, construction years, and minimum night requirements to understand the variety of inventory.

Key Insights (Based on Initial Cells)

- **Data Scale:** The initial dataset contains 102,599 records across 26 different features.
- **Service Fees:** Listings typically include a `service fee` that appears to be roughly 20% of the price (e.g., $193 fee for a $966 price).
- **Review Activity:** There is a wide range of review activity, with some listings having over 1,000 reviews while others have none.
- **Time Series:** The project utilizes a line plot to track the "Number of Reviews Over Time," indicating a focus on growth and seasonal trends in the Airbnb market.

1. Data Profiling & Structural Integrity

- **Dataset Overview:** You successfully identified that the dataset contains **102,599 records** across **26 features**.
- **Technical Audit:** You used `.info()` and `.describe()` to identify mixed data types (e.g., column 25) and statistical outliers, such as the minimum nights being -1223 and maximum availability being 3677.

2. Comprehensive Data Cleaning

- **Missing Value Management:** You quantified a high volume of missing data, notably in `last review` (15,893) and `house_rules` (52,131).
- **Strategic Imputation:** You intelligently filled `reviews per month` with **0** and `last review` with the **minimum date** found in the dataset to maintain data continuity.
- **Noise Reduction:** You dropped the `license` and `house_rules` columns, which were too sparse to provide analytical value, and removed rows missing critical identifiers like `NAME` or `host name`.

3. Feature Engineering & Visualization

- **Data Transformation:** You converted the `last review` column into a `datetime` format, enabling chronological analysis.
- **Temporal Analysis:** You successfully plotted the "Number of Reviews Over Time" by grouping reviews by month, which helps in identifying seasonal trends in the NYC market.

Recommendations for Improvement

To take this project to a professional or portfolio-ready level, consider the following enhancements:

1. Fix Logical Data Inconsistencies

- **Filter Outliers:** Your summary statistics show **negative values for minimum nights** (-1223) and **availability exceeding 365 days** (3677).
    - *Recommendation:* Filter your dataframe to only include logical ranges (e.g., `df = df[(df['minimum nights'] > 0) & (df['availability 365'] <= 365)]`).

2. Numerical Feature Cleaning

- **Currency Conversion:** Columns like `price` and `service fee` are currently "object" types because of the "$" sign.
    - *Recommendation:* Strip the "$" and "," characters and convert these to floats so you can calculate the average price per neighborhood or borough.

3. Advanced Visualizations

- **Geospatial Mapping:** Since you have `lat` and `long` coordinates, you can create a heat map.
    - *Recommendation:* Use a library like `Folium` or `Plotly` to map listings across New York City boroughs to visualize price density.
- **Categorical Insights:** Create a bar chart comparing the average price across different `neighbourhood groups` (Manhattan vs. Brooklyn, etc.).

4. Correlation Analysis

- **Heatmap:** Use `sns.heatmap(df.corr())` to see if there is a relationship between `price`, `number of reviews`, and `review rate number`. This would help determine if higher-rated homes actually command higher prices.