Exploratory Data Analysis of Paris Airbnb Listings: Insights and Trends from the December 2023 Dataset*

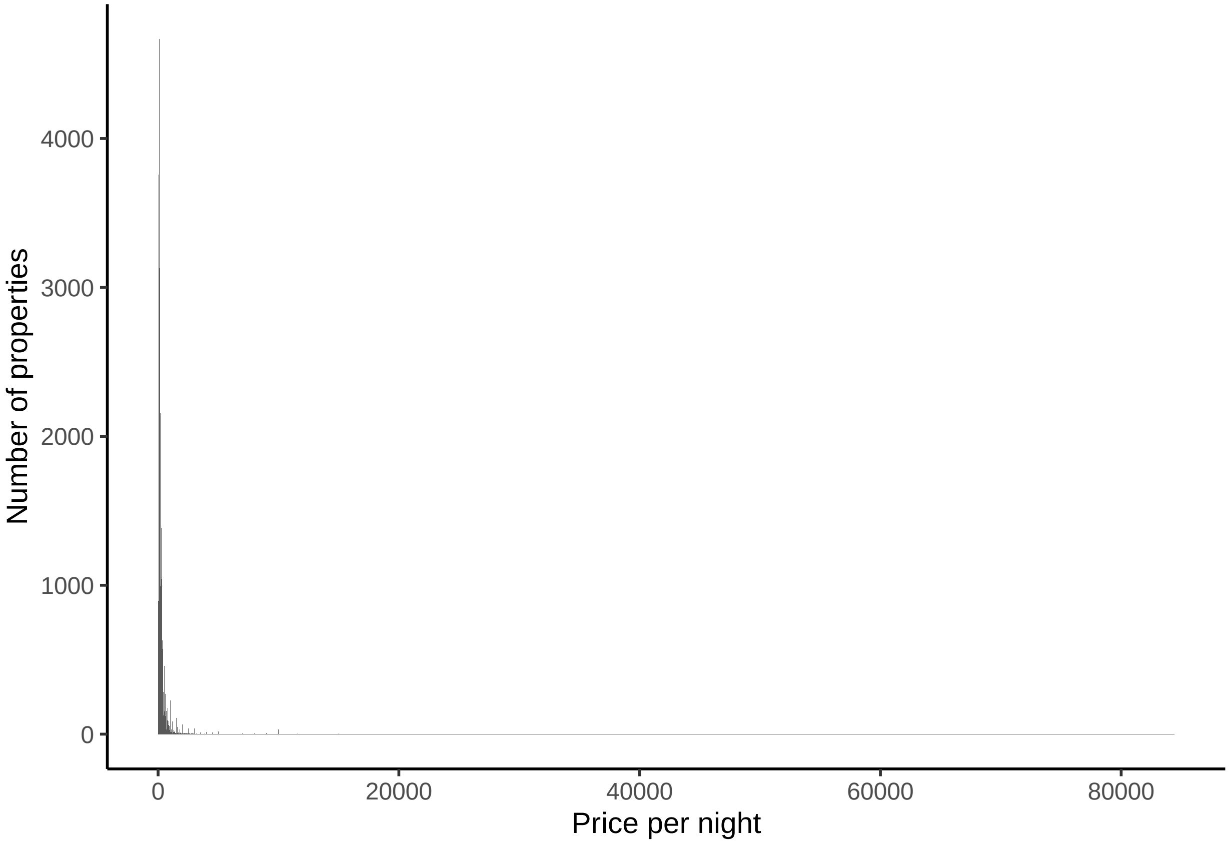
Yingqi Pang

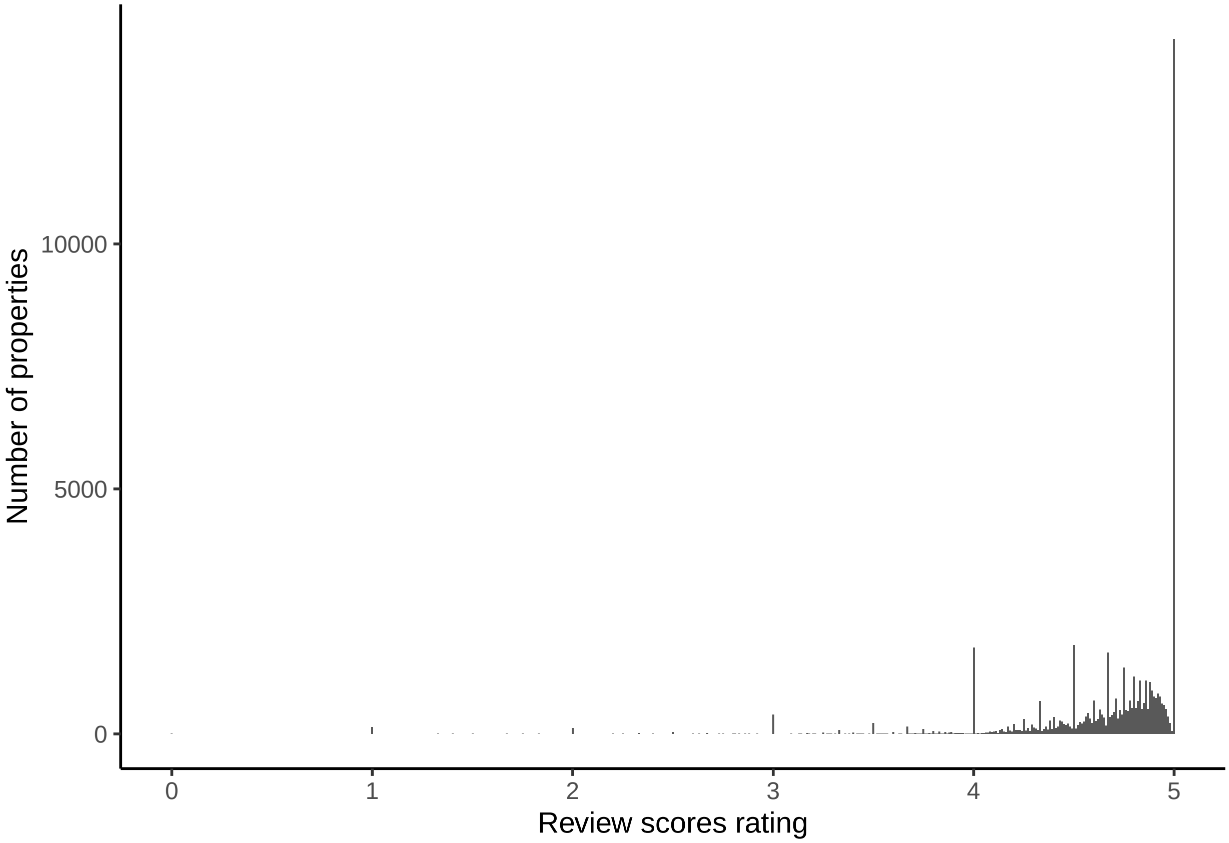
05 March 2024

Contents

*Code and data are available at: https://github.com/pangyin2/Airbnb-listings-in-Paris.git

Result


Number of properties
4000
3000
2000
1000
0
0
20000
40000
60000
80000
Price per night


Number of properties
10000
5000
0
0
1
2
3
4
5
Review scores rating

Discussion

It is an exploratory data analysis (EDA) on Airbnb listings in Paris, France, based on the dataset released on December 12, 2023. It illustrates a systematic approach to data analysis, starting from data acquisition to preprocessing, analysis, and visualization. Here are some key aspects covered:

1. **Data Acquisition**: It begins by downloading the Airbnb listings dataset for Paris directly from the Inside Airbnb website. This approach ensures that the analysis is based on the most recent data available, enhancing the relevance and timeliness of the insights derived.

2. **Data Preparation**: Once the dataset is downloaded, it's saved locally to facilitate easy access and to reduce dependency on external servers for future analyses. This step is crucial for reproducibility and efficiency in data analysis workflows.

3. **Variable Selection**: It selects a subset of variables from the dataset for analysis. This selection is based on the variables' potential to provide insights into the listings' characteristics, such as host details, property features, and review scores. Focusing on specific variables makes the analysis more manageable and targeted.

4. **Data Cleaning**: A significant portion of the script is dedicated to cleaning the data, particularly the price variable. The cleaning process involves converting the price from a character string to a numeric value, which is essential for any quantitative analysis. This step highlights the importance of preprocessing data to ensure its suitability for analysis.

5. **Exploratory Data Analysis (EDA)**: It uses the `ggplot2` package for data visualization, offering a glimpse into the distribution of prices among the listings and the relationship between different variables. EDA is a critical step in understanding the data's underlying patterns and informing further analysis.

6. **Handling Missing Data**: It addresses the challenge of missing data, particularly in variables like superhost status and host response times. It demonstrates strategies for filtering out or recoding missing values, which is crucial for maintaining the integrity of the analysis.

7. **Logistic Regression Analysis**: Towards the end, it applies logistic regression to explore the relationship between superhost status and other variables like response time and review scores. This step transitions the analysis from descriptive to inferential, allowing for hypotheses testing about the factors influencing superhost status.

8. **Data Saving**: Finally, it saves the processed and analyzed dataset in a Parquet file, a columnar storage file format, for efficient storage and access. This step ensures that the cleaned and analyzed data can be easily reused for future analyses or modeling.