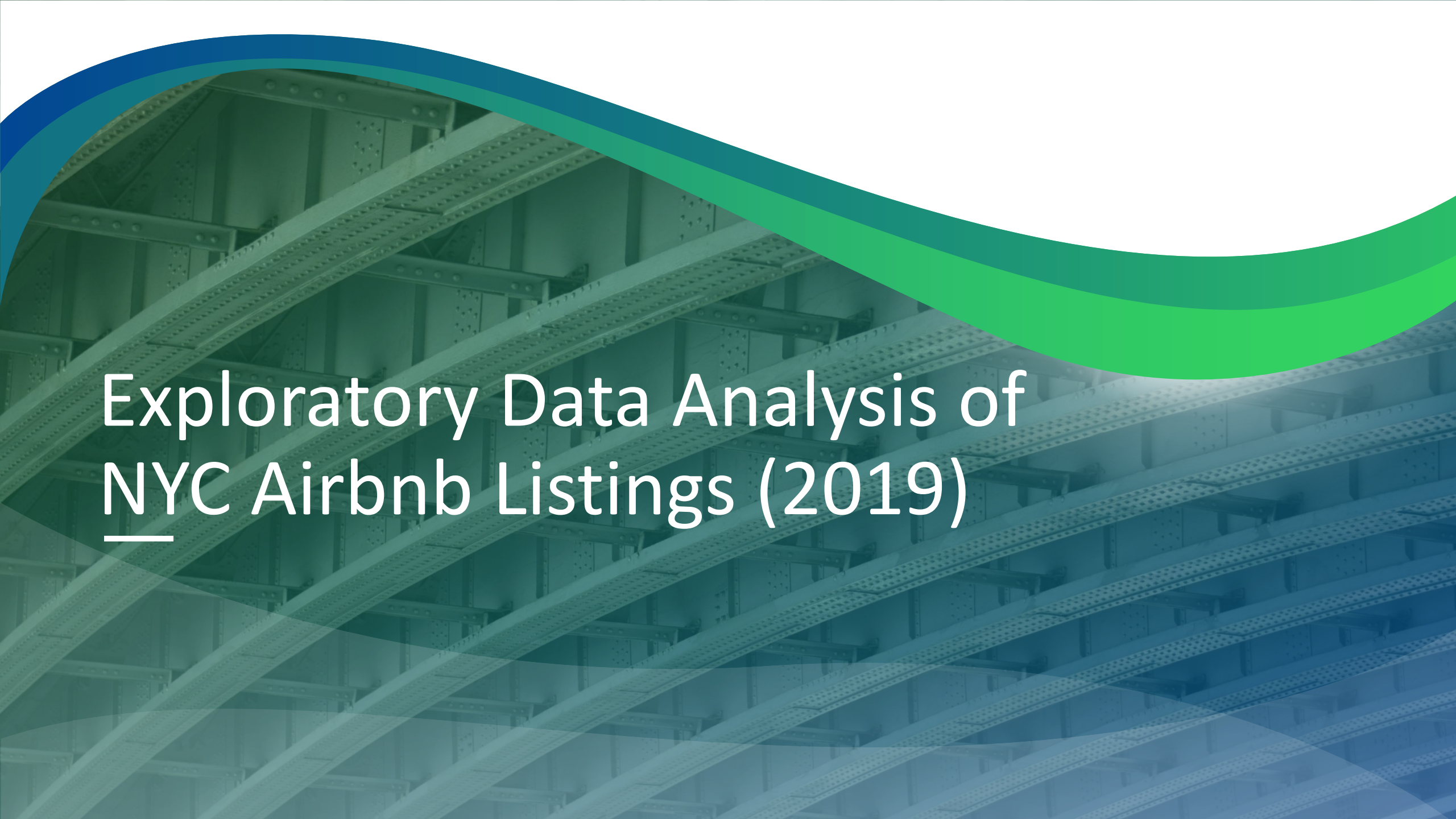




Exploratory Data Analysis of NYC Airbnb Listings (2019)

HELLO!

I am Khawja Redwanul Islam. In this presentation, I will be giving you a summary of the Airbnb dataset from 2019.



Agenda



Data Overview



Univariate Analysis



Executive Summary



Data Cleaning & Preparation



Multivariate Analysis



References



Introduction to Exploratory Data Analysis (EDA)

In this presentation, an Exploratory Data Analysis (EDA) approach is applied to the Airbnb NYC 2019 dataset to uncover key trends and market dynamics. EDA is a critical first step in the data mining process [1], focusing on identifying patterns and spotting anomalies with the help of summary statistics and graphical representations [2]. While this analysis focuses on public data, care must be taken in data mining to address privacy-preserving techniques when handling sensitive information [3]. The findings from this process reveal important market insights and provide actionable recommendations based on a data-driven case study of New York City [4].



Data Overview

Dataset: This analysis uses the AB_NYC_2019.csv dataset, a real-world collection of Airbnb listings from 2019, which provides a rich context for exploring rental patterns [4].

Initial State: The dataset contains approximately 48,000 listings and 16 features, including location, price, room type, and review metrics.

Objective: To clean and prepare the dataset, uncover key trends using visual and statistical techniques, and present findings relevant to the Airbnb market.

Initial Data Snapshot: The dataset's initial structure is shown at the right side of the slide, highlighting the different data types and their non-null counts.

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 48895 entries, 0 to 48894
Data columns (total 16 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   id                                    48895 non-null  int64
1   name                                48879 non-null  object
2   host_id                             48895 non-null  int64
3   host_name                           48874 non-null  object
4   neighbourhood_group                 48895 non-null  object
5   neighbourhood                       48895 non-null  object
6   latitude                           48895 non-null  float64
7   longitude                          48895 non-null  float64
8   room_type                          48895 non-null  object
9   price                              48895 non-null  int64
10  minimum_nights                     48895 non-null  int64
11  number_of_reviews                  48895 non-null  int64
12  last_review                        38843 non-null  object
13  reviews_per_month                  38843 non-null  float64
14  calculated_host_listings_count     48895 non-null  int64
15  availability_365                   48895 non-null  int64
dtypes: float64(3), int64(7), object(6)
memory usage: 6.0+ MB
```

Practical Data Handling: Cleaning & Preprocessing

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 48895 entries, 0 to 48894
Data columns (total 12 columns):
#   Column              Non-Null Count  Dtype
---  -
0   host_id              48895 non-null  int64
1   neighbourhood_group  48895 non-null  object
2   neighbourhood        48895 non-null  object
3   latitude             48895 non-null  float64
4   longitude            48895 non-null  float64
5   room_type            48895 non-null  object
6   price               48895 non-null  int64
7   minimum_nights       48895 non-null  int64
8   number_of_reviews    48895 non-null  int64
9   reviews_per_month    48895 non-null  float64
10  calculated_host_listings_count  48895 non-null  int64
11  availability_365      48895 non-null  int64
dtypes: float64(3), int64(6), object(3)
memory usage: 4.5+ MB
```

Objective: To identify and address issues such as missing data and outliers to improve the data quality for accurate analysis.

Key Cleaning Steps: Imputation: Missing monthly review values were set to 0.

Column & Row Removal: Irrelevant columns were dropped, and rows with missing last review dates were also removed to ensure data integrity.

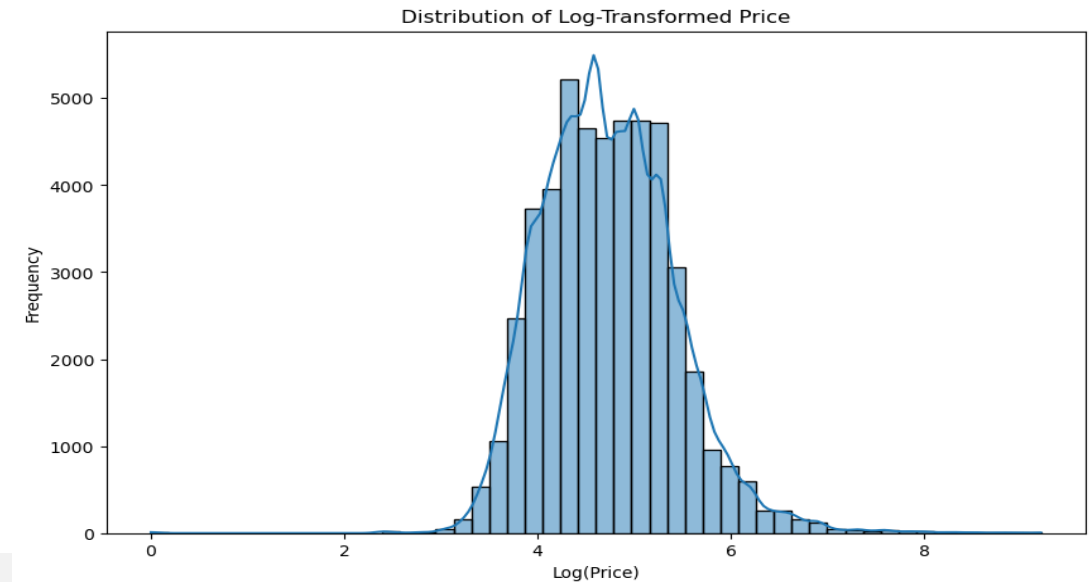
Outlier Handling: A log transformation was applied to the price column to normalise its distribution. The performance and scalability of such cleaning tools are key considerations in large datasets [5].

```
df.isna().sum()
```

```
host_id              0
neighbourhood_group  0
neighbourhood        0
latitude             0
longitude            0
room_type            0
price                0
minimum_nights       0
number_of_reviews    0
reviews_per_month    0
calculated_host_listings_count  0
availability_365     0
dtype: int64
```

EDA – Univariate Analysis

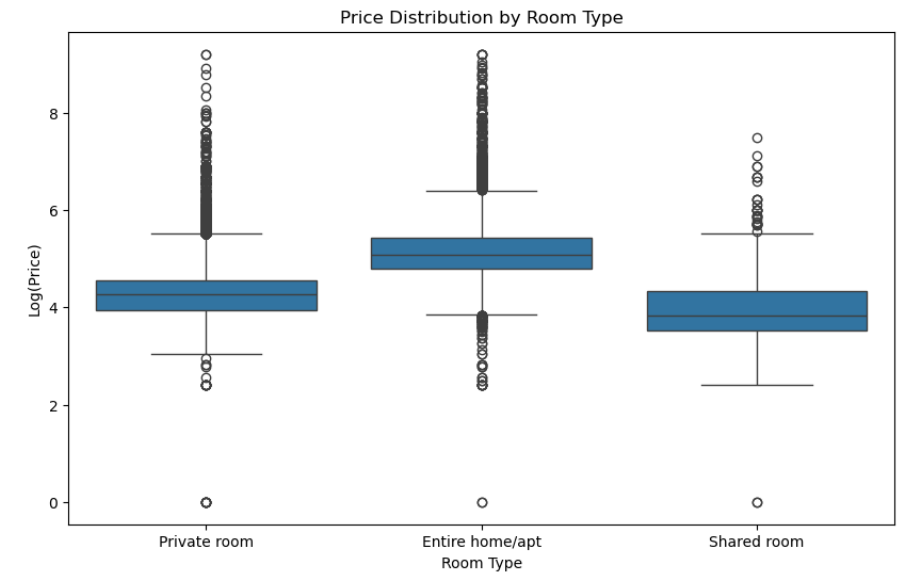
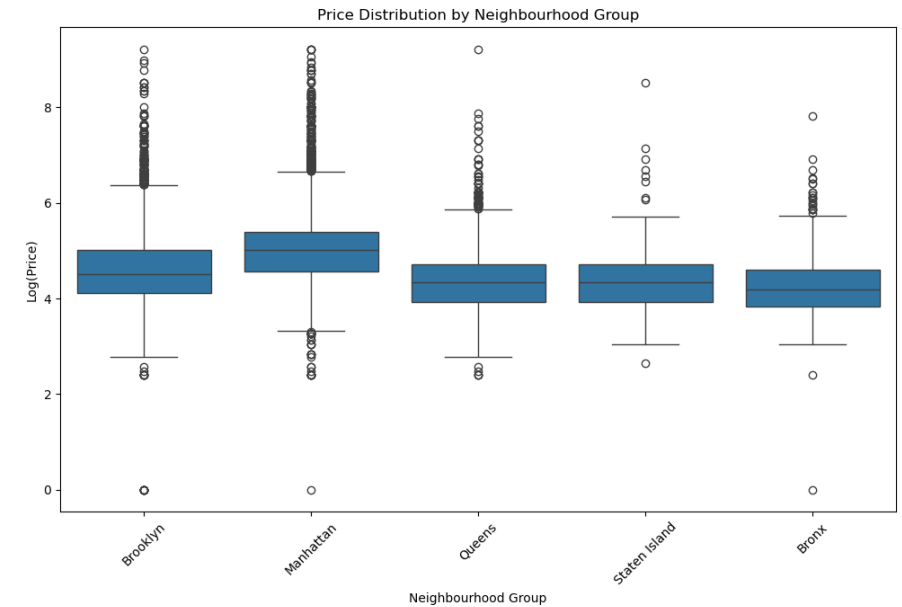
Price Distribution: The initial price histogram revealed a strong right-skew due to outliers. The log-transformed distribution provides a much clearer, more normalised view. This is a common EDA technique used to understand variable distributions [2].



EDA – Univariate Analysis

Geographic Distribution: Manhattan and Brooklyn dominate the market, together accounting for over 85% of all listings.

Room Type Frequency: The market is almost evenly divided between 'Entire home/apt' and 'Private room' listings.



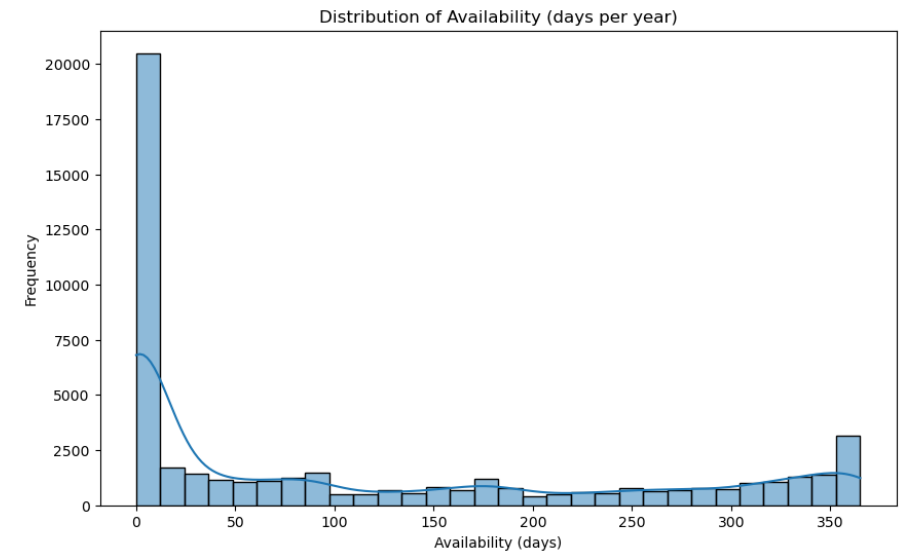
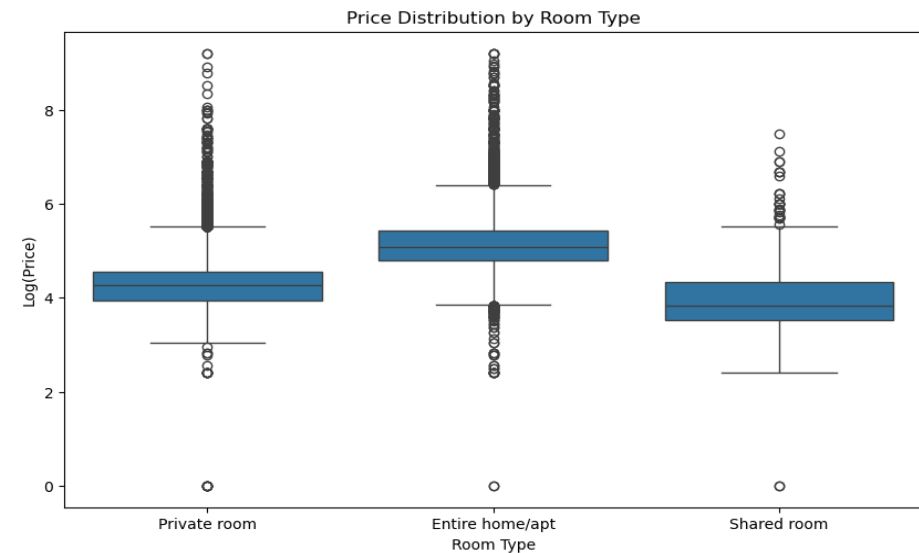
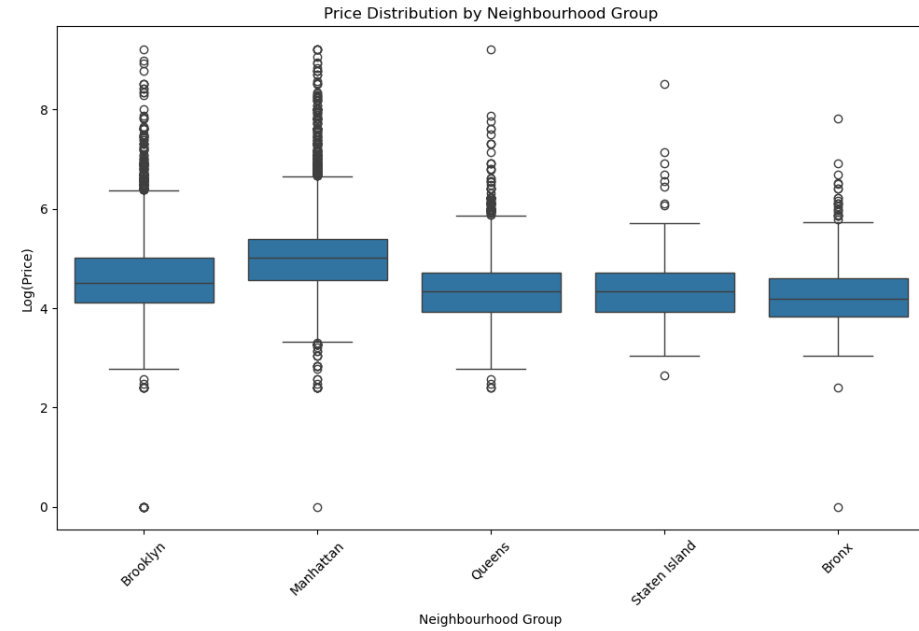
EDA – Multivariate Analysis



Price by Neighbourhood Group: A socio-economic analysis confirms that location is a primary driver of price [6]. Manhattan shows the highest price variability, while the Bronx is the most affordable.

Price by Room Type: 'Entire home/apt' listings are priced significantly higher than 'Private room' options.

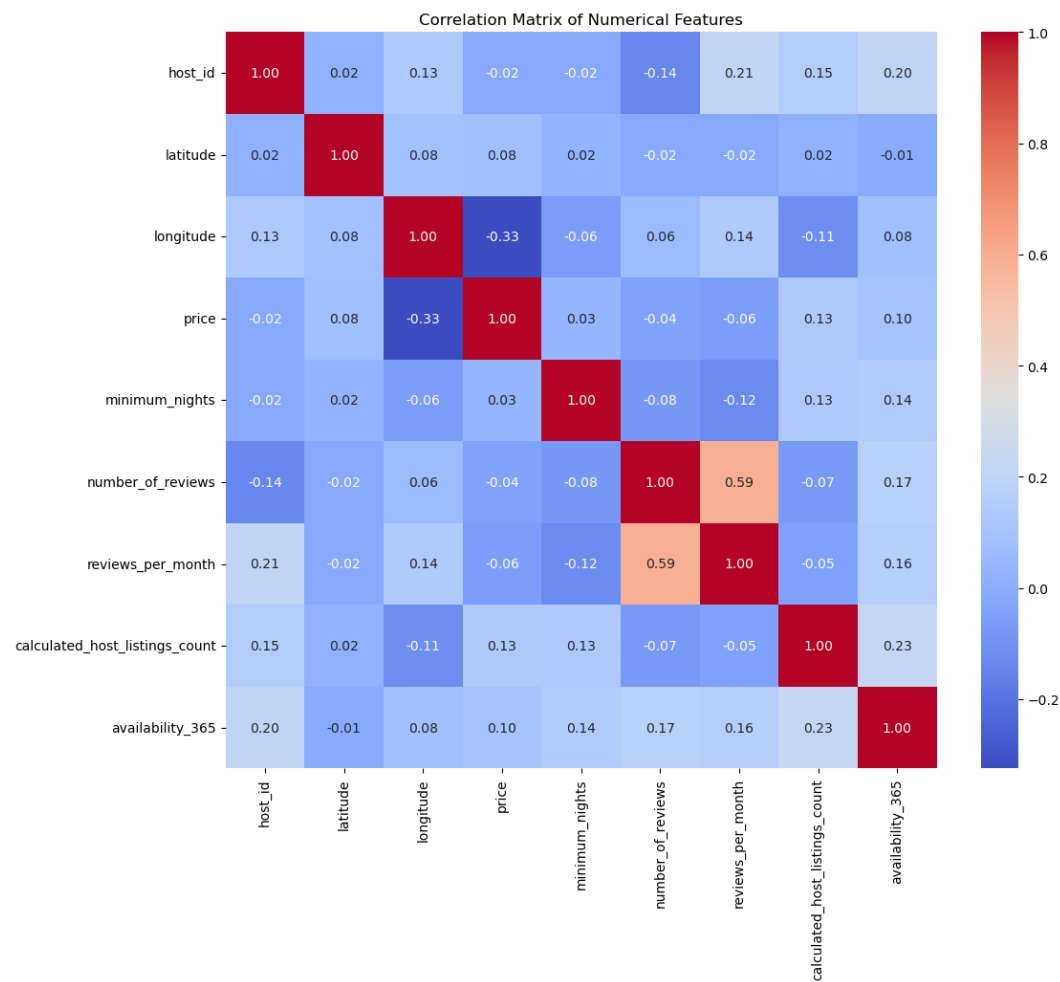
Host Activity: A small number of hosts manage a large portion of the listings, with the top host having over 300 properties. This insight helps in understanding user behaviour within the dataset [7].



EDA – Multivariate Analysis: Correlation & Geospatial

Correlation Insights:

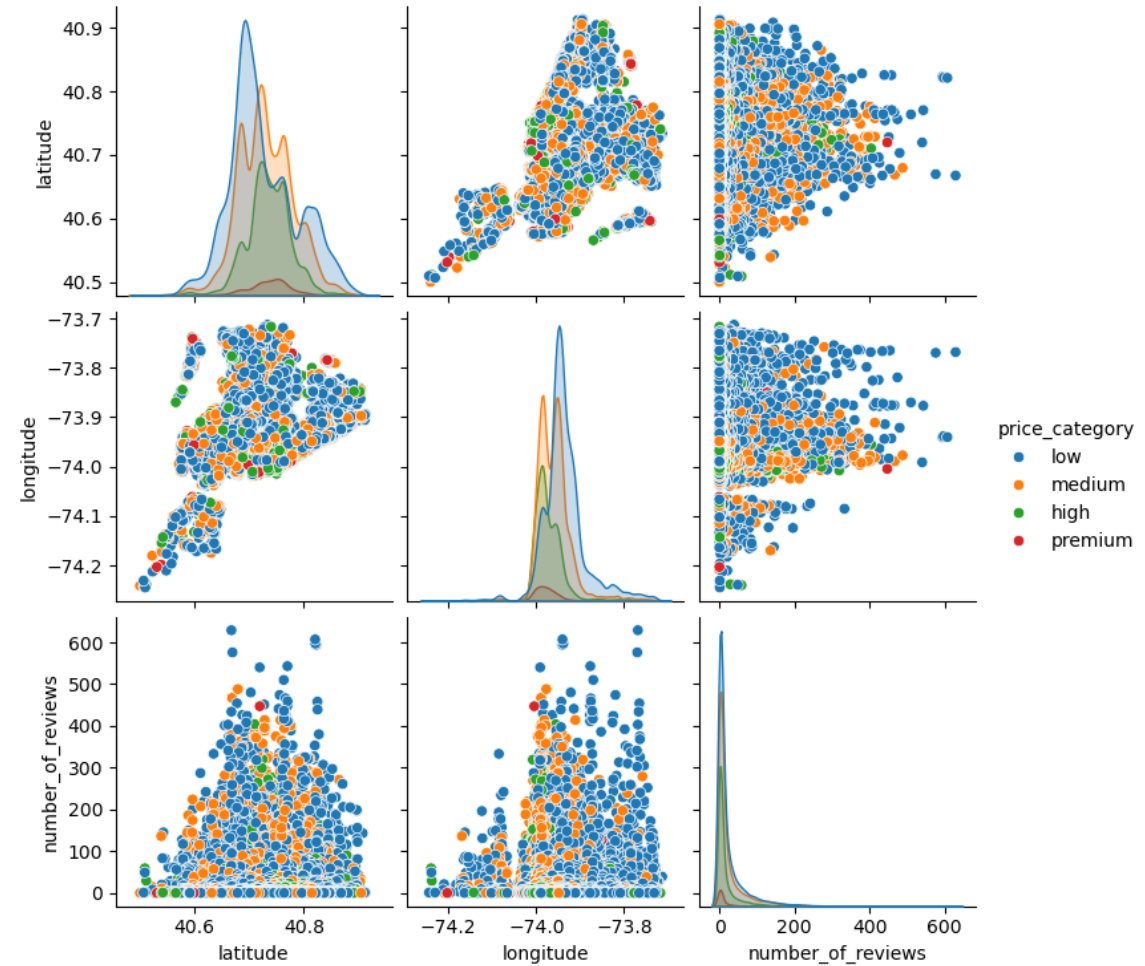
The correlation matrix reveals that most variables have weak linear correlations, which suggests that pricing is influenced by complex, non-linear factors that may require advanced testing [8].



EDA – Multivariate Analysis: Correlation & Geospatial

Geospatial Analysis:

The scatter plot of listings confirms that the most expensive properties are concentrated in central Manhattan. Such spatiotemporal analysis is crucial for understanding urban housing data [9].



Executive Summary



Conclusion:

This EDA successfully cleaned the dataset and revealed that location and room type are the most significant drivers of price, a common finding in EDA practice [1]. The analysis provides a foundation for data-driven Airbnb pricing strategies in metropolitan contexts [4].



Actionable Insights & Recommendations:

Hosts: Should benchmark pricing against similar properties in the same borough and room type.

Future Work: Further analysis could explore seasonal trends and host ratings to refine pricing strategies, which is a key component of visual analytics for urban housing data [10].

References

- [1] J. Zhang, K. Wang, and Y. Xie, "Exploratory Data Analysis in Practice: Trends, Challenges, and Future Directions," IEEE Access, vol. 11, pp. 107851–107868, 2023.
- [2] M. Raparathi et al., "Exploratory Data Analysis Techniques – A Comprehensive Review," AJMLRA, vol. 4, no. 1, pp. 22–35, 2024.
- [3] U. H. W. A. Hewage, R. Sinha, and M. A. Naeem, "Privacy-preserving data (stream) mining techniques and their impact on data mining accuracy: a systematic literature review," The Artificial intelligence review, vol. 56, no. 9, pp. 10427–10464, 2023.
- [4] M. Gutiérrez, A. del Valle, and J. M. García, "Data-Driven Insights into Airbnb Listings: A Case Study of New York City," Journal of Urban Informatics, vol. 5, no. 1, pp. 45–60, 2021.
- [5] P. Martins et al., "Performance and Scalability of Data Cleaning and Preprocessing Tools: A Benchmark on Large Real-World Datasets," Data, vol. 10, no. 5, pp. 1–20, 2025.
- [6] G. Dudas, "A Socio-Economic Analysis of Airbnb in New York City," Regional Statistics, vol. 7, no. 1, pp. 45–60, 2017.
- [7] G. Lv et al., "Understanding the Users and Videos by Mining a Novel Danmu Dataset," IEEE Trans. Big Data, vol. 8, no. 2, pp. 535–551, 2022.
- [8] S. Mohammadi, "Neural network for univariate and multivariate nonlinearity tests," Statistical Analysis and Data Mining, vol. 13, no. 1, pp. 50–70, 2020.
- [9] A. Alkhodair, M. Alrubaian, and A. Alamri, "A Spatiotemporal Analysis of Airbnb Listings: Insights from New York City," IEEE Trans. Computational Social Systems, vol. 9, no. 5, pp. 1290–1302, 2022.
- [10] L. Huang, H. Chen, and J. Wu, "Visual Analytics for Urban Housing Data: A Case Study on Airbnb," IEEE Computer Graphics and Applications, vol. 44, no. 2, pp. 88–97, 2024.

Thank You, Everyone

