

Exploratory Data Analysis & Visualization

MOHAMMAD LAKHI
2512430
ACCA7015

Agenda

- INTRODUCTION TO EDA
- DATASET OVERVIEW
- DATA CLEANING
- UNIVARIATE ANALYSIS
- MULTI-VARIATE ANALYSIS
- EXECUTIVE SUMMARY
- CONCLUSION
- RECOMMENDATIONS

Introduction to EDA

- ▶ Exploratory Data Analysis is the phase where we understand the data rather than just processing the data [1].
- ▶ EDA is used to identify key patterns, irregularities and relationship between data. It combines visual inspection and statistical analysis, to get more effective enquiries and refined prediction [1][2].
- ▶ Similarly, EDA is recommended before any formal analysis to identify data quality or any missing values which can affect later results [3].

Dataset Overview

```
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
```

- ▶ Dataset: AB_NYC_2019.csv
- ▶ Source: Kaggle
- ▶ Contains 48,895 entries and 16 variables.
- ▶ Mix of numerical values and geospatial features.

Data Cleaning

	0
id	0
name	16
host_id	0
host_name	21
neighbourhood_group	0
neighbourhood	0
latitude	0
longitude	0
room_type	0
price	0
minimum_nights	0
number_of_reviews	0
last_review	10052
reviews_per_month	10052
calculated_host_listings_count	0
availability_365	0

dtype: int64

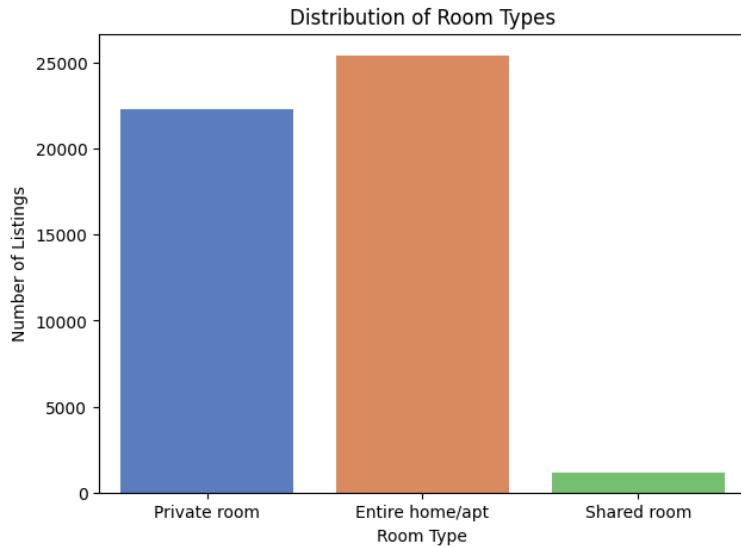
- ▶ Using '**isna().sum()**' to find the missing values.
- ▶ Missing values:
 - 'name' - 16
 - 'host_name' - 21
 - 'last_review' - 10052
 - 'reviews_per_month' - 10052

Data Cleaning

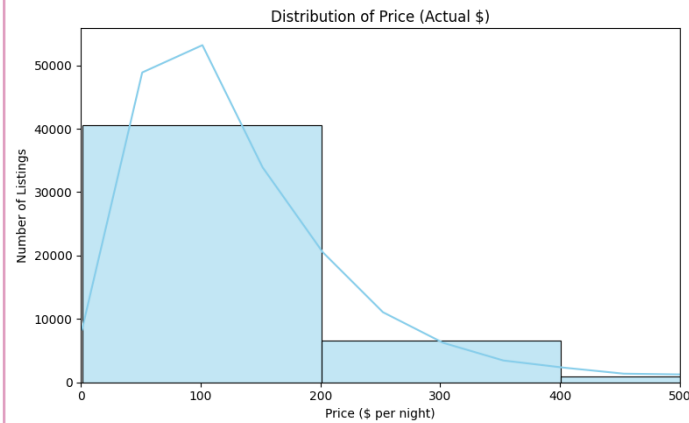
	0
host_id	0
neighbourhood_group	0
neighbourhood	0
latitude	0
longitude	0
room_type	0
price	0
minimum_nights	0
number_of_reviews	0
reviews_per_month	0
calculated_host_listings_count	0
availability_365	0
dtype: int64	

- ▶ Removed 'id', 'name', 'host_name' and 'last_review'.
- ▶ Removed duplicate and invalid records.
- ▶ For 'reviews_per_month', missing values filled with '0'.
- ▶ Final dataset – 48,885 rows * 12 columns.

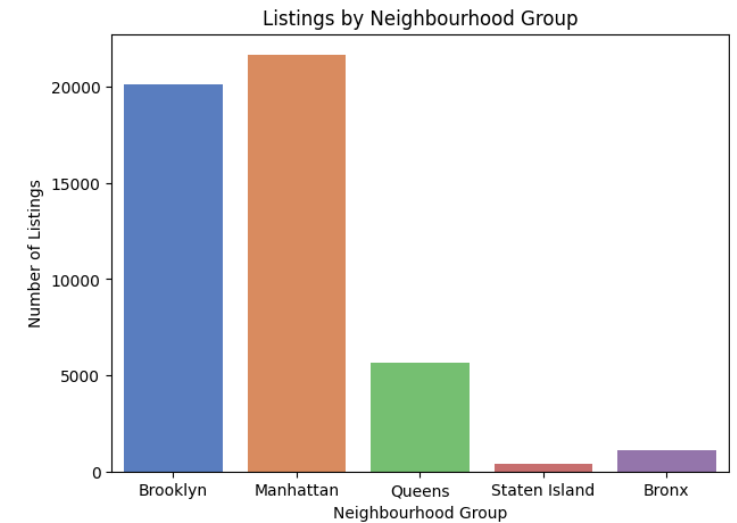
Univariate Analysis



- This bar chart indicates that people prefer more private rooms and home type rather than shared rooms.

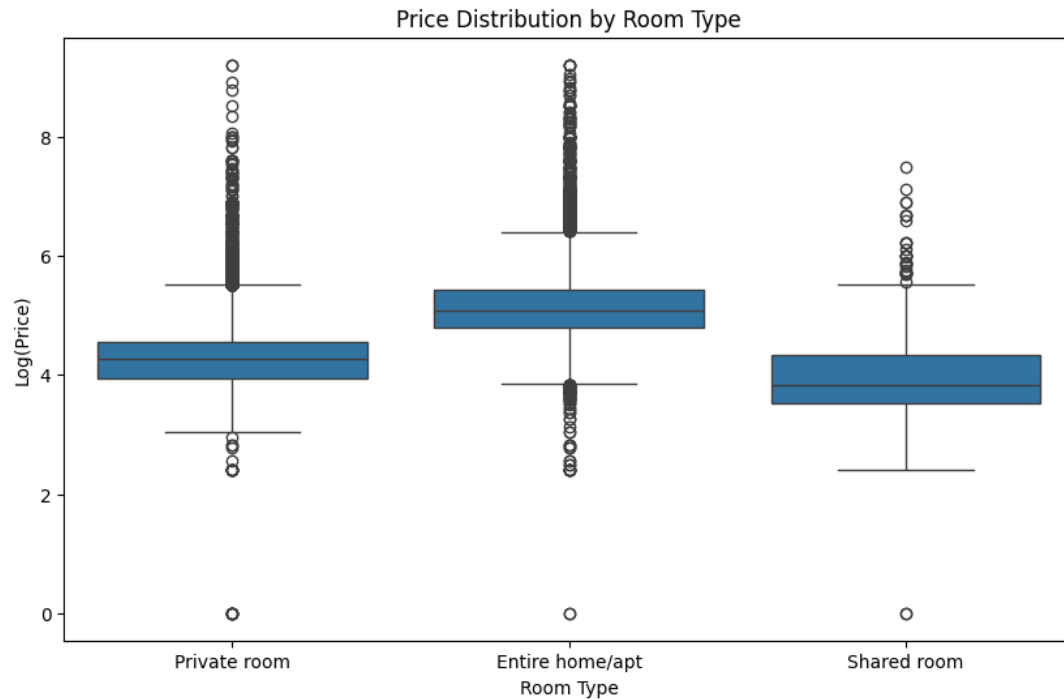


- The majority of the listings are priced less than \$200 per night, showing the affordability for travellers.



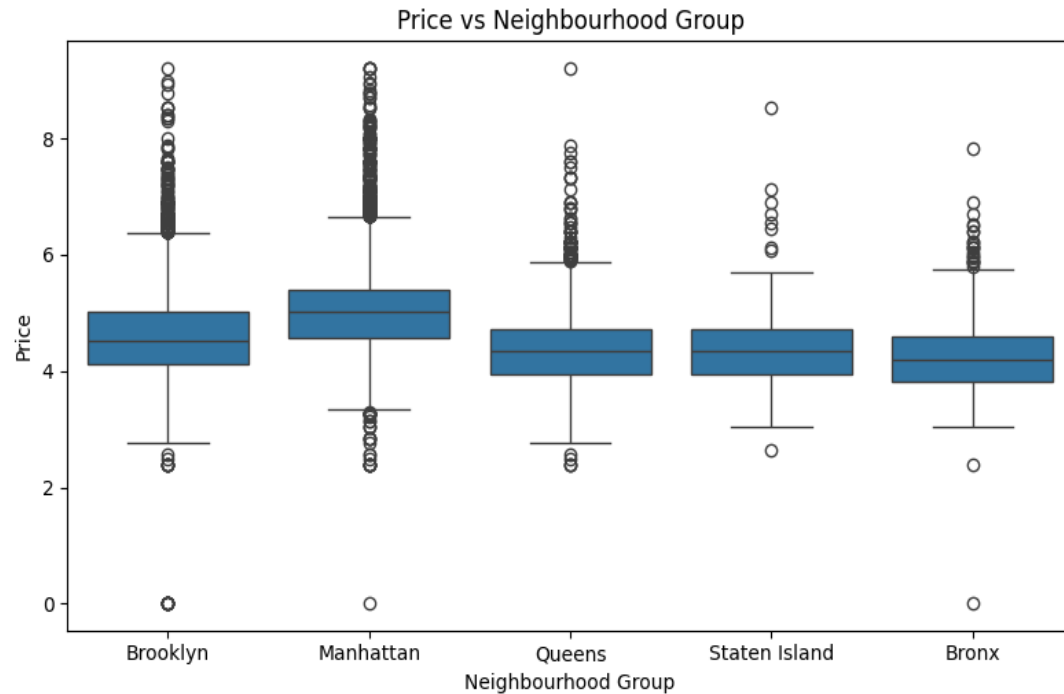
- In this dataset, the listings are mainly located in Manhattan and Brooklyn, indicating their popularity with tourists.

Multi-Variate Analysis



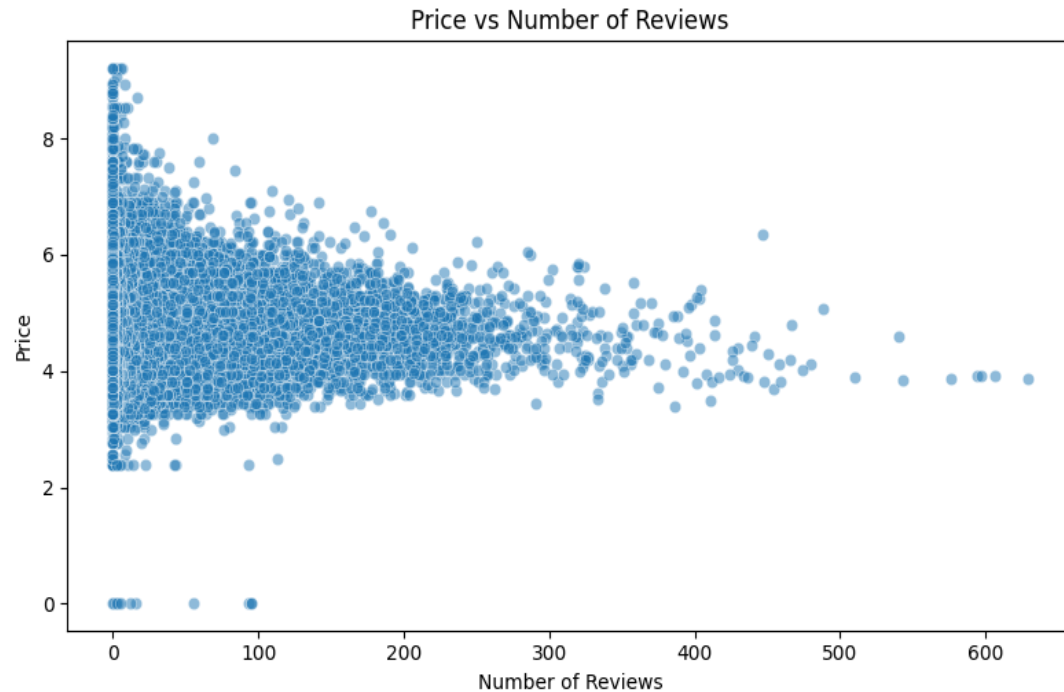
- ▶ This box plot shows how room type affects the price rate.
- ▶ The average cost of Entire home/apt is the highest among room types.
- ▶ Private and Shared rooms are more cost effective.
- ▶ This difference proves that room type has major impact on the price distribution [4].

Multi-Variate Analysis



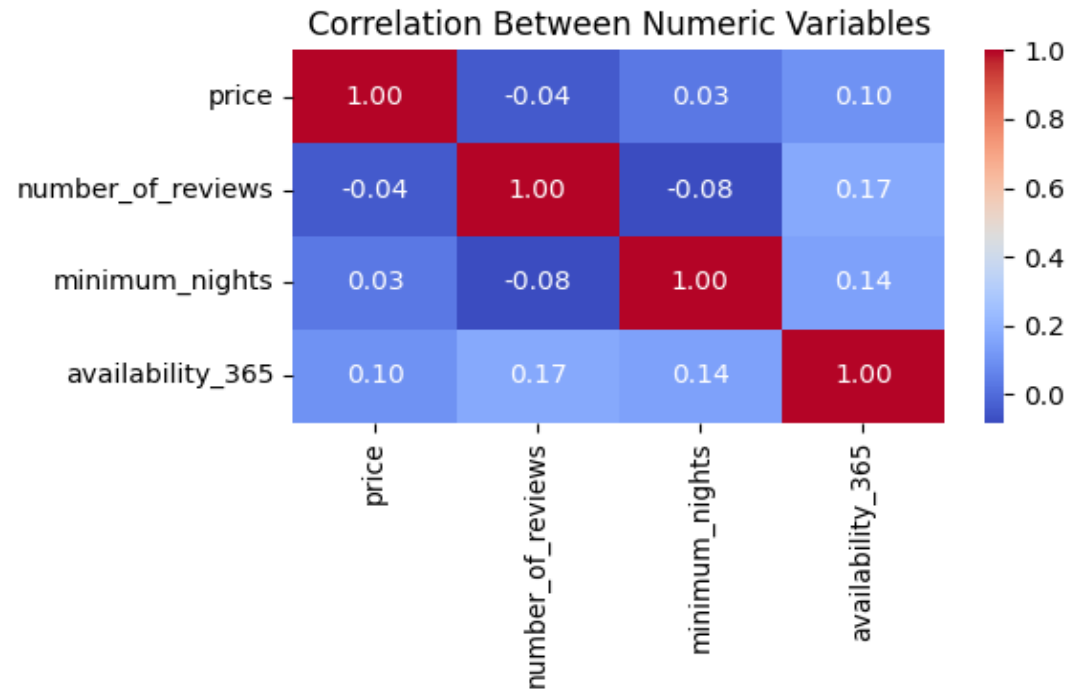
- ▶ Price changes significantly between New York City's areas.
- ▶ Manhattan and Brooklyn have the highest pricing among other cities.
- ▶ This box plot shows the demands and tourism in the major cities.

Multi-Variate Analysis



- ▶ This scatter plot represents rooms having more review affects their pricing.
- ▶ It overall suggests that affordable options have a higher booking frequency, compared to high priced rooms having less reviews [5].

Multi-Variate Analysis



- ▶ This Correlation heatmap illustrates how the price has the relation with majority of numeric variables.
- ▶ Whereas, minimum night has the strongest relation with the pricing.
- ▶ Availability of rooms has the lowest correlation with every other values [4].

Executive Summary

- ▶ This study explored Airbnb listings of major cities of New York, to understand the major factors that affects pricing and distribution.
- ▶ Data cleaning process removes missing and unacceptable values, which increases the accuracy and reliability of the insights.
- ▶ Univariate and Multi-Variate Analysis showed that room-type and location are the major factors for price distribution.
- ▶ Queens, Bronx and Staten Island has the most affordable listings, which attracts more reviews and booking of rooms.

Conclusion

- ▶ The pricing in New York cities are strongly influenced by location and room types.
- ▶ Manhattan and Brooklyn listing has the highest prices as they are located at central of New York.
- ▶ Room types is a simple variable, describes entire homes are more costly than private and shared rooms.
- ▶ This analysis shows Queens and Bronx have the most affordable pricing, which leads to higher booking rates and more reviews.

Recommendations

- ▶ Optimize pricing strategy: In order to compete, Queens and Bronx should highlight affordability and positive reviews for better attraction.
- ▶ Target market classification: Using Location and room type data, pricing tool can be developed to assist hosts in balancing demand and costing.
- ▶ Improve guest experience: Customer satisfaction is the major key factor for attracting more booking, It can be done by improving quality which will lead to good reviews.
- ▶ Utilize visual insights: Identify areas that lacks, and improve them by analysing their performance across the areas.

References

- ▶ [1] Phanwadee Sinthong and M. P. Carey, "Exploratory Data Analysis with Database-backed Dataframes: A Case Study on Airbnb Data," Dec. 2021, doi: <https://doi.org/10.1109/bigdata52589.2021.9671603>.
- ▶ [2] V. D. Poian *et al.*, "Exploratory data analysis (EDA) machine learning approaches for ocean world analog mass spectrometry," vol. 10, May 2023, doi: <https://doi.org/10.3389/fspas.2023.1134141>.
- ▶ [3] A. M. Sharifnia, D. E. Kpormegbey, D. K. Thapa, and M. Cleary, "A Primer of Data Cleaning in Quantitative Research: Handling Missing Values and Outliers.," *PubMed*, Mar. 2025, doi: <https://doi.org/10.1111/jan.16908>.
- ▶ [4] V. Toader, A. L. Negrușă, O. R. Bode, and R. V. Rus, "Analysis of price determinants in the case of Airbnb listings," *Economic Research-Ekonomska Istraživanja*, vol. 35, no. 1, pp. 1–17, Aug. 2021, doi: <https://doi.org/10.1080/1331677x.2021.1962380>.
- ▶ [5] A. Sarkar, J. B. Pick, and S. Jabeen, "Spatiotemporal Patterns and Socioeconomic Influences on Host Participation in Short-Term Rental Markets: Airbnb in San Francisco," *ISPRS international journal of geo-information*, vol. 13, no. 1, pp. 4–4, Dec. 2023, doi: <https://doi.org/10.3390/ijgi13010004>.