

Exploratory Data Analysis (EDA) of Airbnb Listings in New York City (2019)



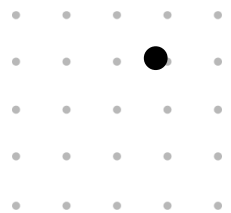
Dataset: AB_NYC_2019.csv
Module: L7 - Data Mining & Visualisation
Date: 31st october 2025

Presented by Shaidhul Hasan
2442113



Agenda

- **Introduction**
- **Aim & Objectives**
- **Data Overview**
- **Understanding the Dataset**
- **Univariate Analysis**
- **Multivariate Analysis**
- **Data Cleaning & Preparation**
- **Insights and Recommendations**
- **Executive Summary**
- **References**

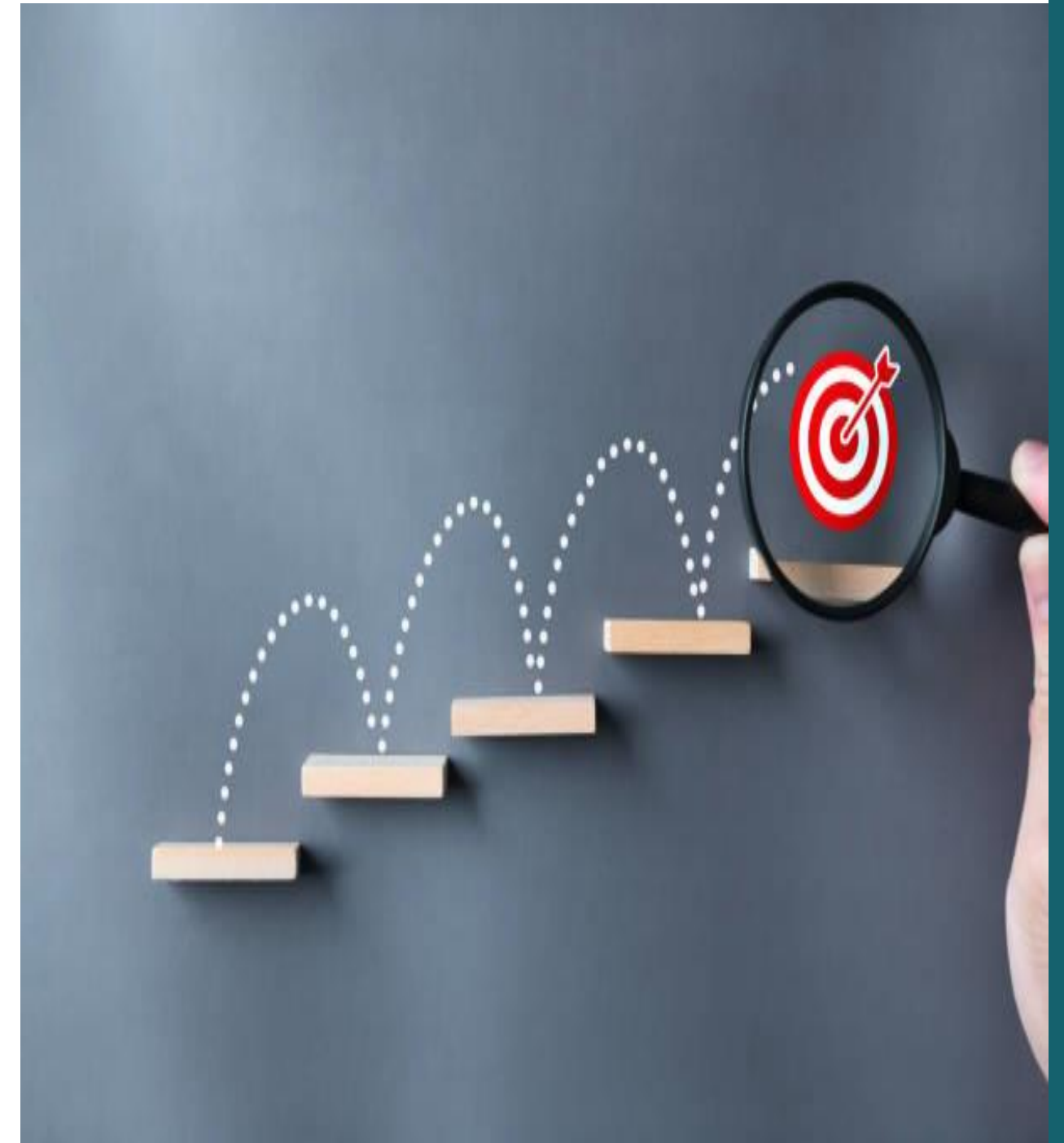


Introduction

Exploratory Data Analysis (EDA), a key phase in data science, focusses on using statistical summaries and visualisation to comprehend information. It allows analysts to find patterns, find irregularities, and find correlations between data prior to formal modelling [1].

The 2019 dataset for Airbnb New York City, which includes about 49,000 listings spread across five main boroughs, is subjected to EDA in this study. The goal is to learn more about how elements like room type, availability, and neighbourhood location impact listing performance and pricing . This procedure identifies trends in the Airbnb industry to help hosts and platform administrators make data-driven business decisions. Data cleansing, transformation, and visualisation are done with Python-based analytical tools including Pandas, NumPy, Matplotlib, and Seaborn [3]. A careful approach to handling missing and inconsistent data ensures reliability in the results [2]. Ultimately, this analysis highlights how exploratory data techniques provide a foundation for understanding complex datasets and optimizing decision-making in digital marketplaces.

The primary concern is to achieve an insight that would enable hosts, customers, and Airbnb to better their business strategies.



Aim & Objectives

Aim

To have an exploration, cleaning and analysis of Airbnb NYC 2019 data with the help of Exploratory Data Analysis (EDA).

Objectives

- To clean and prepare the data ready to analyse it.
- To identify trends, patterns and relationships in the data.
- To better understand with the help of graphs and statistics.
- To deal with missing data properly.
- To provide valuable recommendations on what to do in decision-making at Airbnb.



Data Overview

- The ABNYC2019.csv dataset has been used in this analysis.
- It contains 48,895 rows and 16 columns.



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

	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

```
df.info()
```

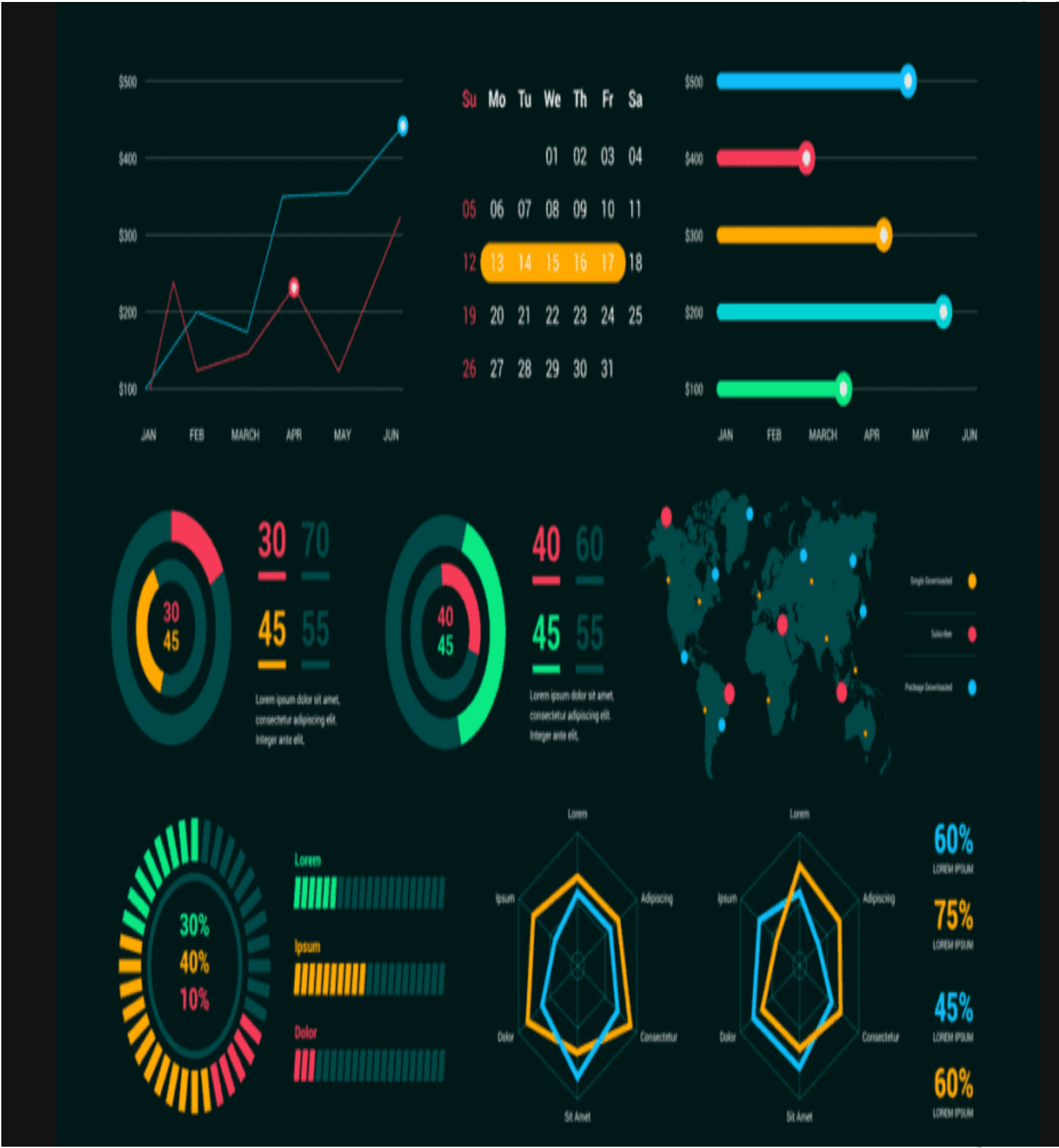
```
<class 'pandas.core.frame.DataFrame'>  
RangeIndex: 48895 entries, 0 to 48894  
Data columns (total 13 columns):  
#   Column                                     Non-Null Count  Dtype  
---  ---                                     -  
0   name                                     48879 non-null  object  
1   host_id                                48895 non-null  int64  
2   neighbourhood_group                     48895 non-null  object  
3   neighbourhood                           48895 non-null  object  
4   latitude                                48895 non-null  float64  
5   longitude                               48895 non-null  float64  
6   room_type                              48895 non-null  object  
7   price                                   48895 non-null  int64  
8   minimum_nights                         48895 non-null  int64  
9   number_of_reviews                      48895 non-null  int64  
10  reviews_per_month                      38843 non-null  float64  
11  calculated_host_listings_count          48895 non-null  int64  
12  availability_365                        48895 non-null  int64  
dtypes: float64(3), int64(6), object(4)  
memory usage: 4.8+ MB
```

Data Overview

Some of the main columns are:

Feature Name	Description
id	Unique ID number for each listing
name	Name of the listing
host_id	Unique ID number for the host
host_name	Name of the host
neighbourhood_group	Main area of the listing (Manhattan, Brooklyn, Queens, Bronx, Staten Island)
neighbourhood	Smaller local area inside each neighbourhood group
latitude	Latitude coordinate of the listing
longitude	Longitude coordinate of the listing
room_type	Type of room offered (Entire home, Private room, Shared room)
price	Price per night in USD
minimum_nights	Minimum number of nights required to book the listing
number_of_reviews	Total number of reviews received
last_review	Date of the most recent review
reviews_per_month	Average number of reviews per month
calculated_host_listings_count	Number of listings owned by the same host
availability_365	Number of available days in a year

The dataset provides a complete picture of Airbnb distribution in the city of New York and identifies the data influencing listing price and popularity.



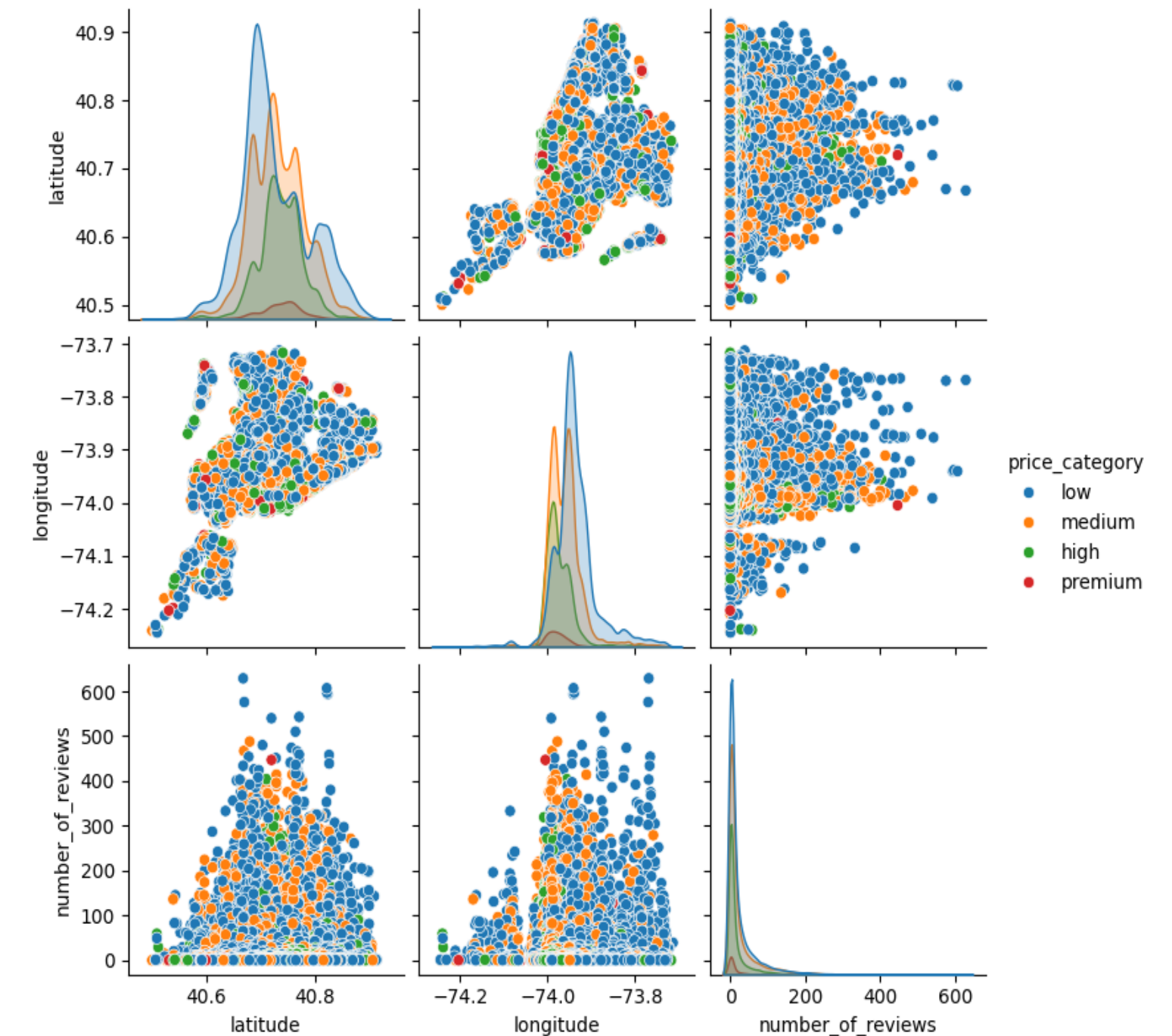
Understanding Dataset

Geographical insight

A scatter diagram of the latitude and longitude illustrates the geographic distribution of listings.

- The denser regions with Airbnb are Manhattan and Brooklyn.
- Central Manhattan and Brooklyn are the areas with the most expensive prices.
- Bronx and Staten Island have fewer listings and lower prices.

These insights demonstrate that the Airbnb business is focused on the regions where there exists business and tourism activity [4] .



Understanding Dataset



Data Quality

The quality of analysis can be influenced by the missing data [5].

Measures were implemented to control the two problems.

Missing Data:

- There were reviews per month with missing values that were substituted with 0.
- Missed last_review values were left blank as they have no impact on price or location.

Dealing with missing values contributed to better results quality [6].

Data Cleaning & Preparation

Data should be clean and consistent before it begins to be analysed.

It was followed by the following steps:

Eliminating unnecessary columns: We did not need some columns, such as id and name, and hostname, to get the result, and therefore they were omitted.

0

Handling missing data:

The value of reviews_per_month was missing, and this was filled with zero.

Last_review was translated into a date format [7].

1

Elimination of duplicates: Duplicate records were cut in order to eliminate one.

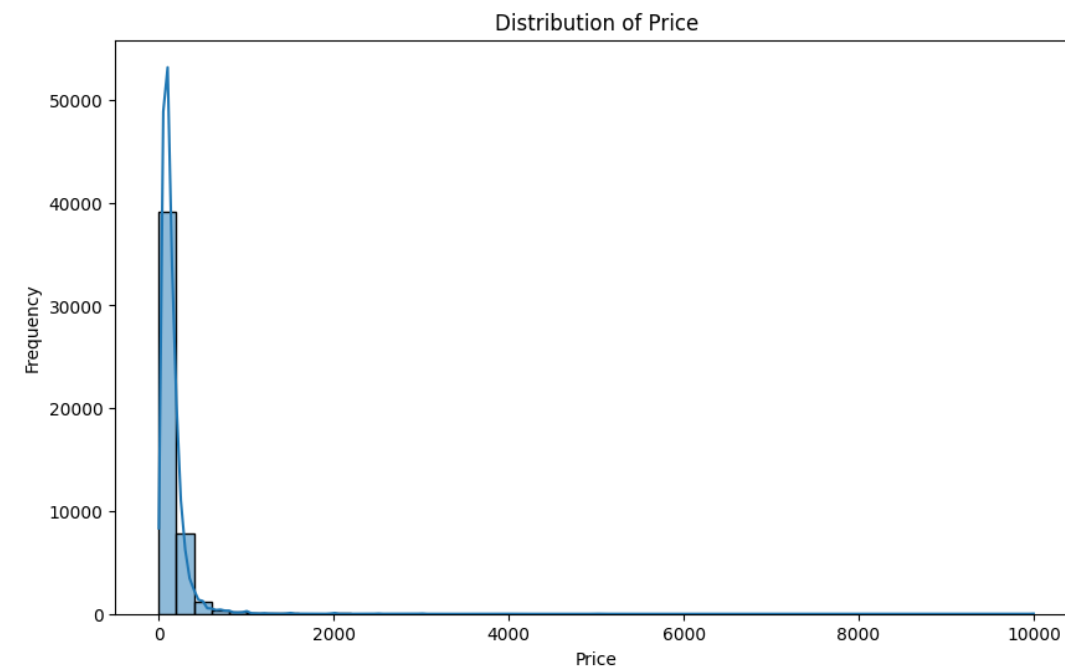
2

Data type correction: All the columns were verified for correct data types.

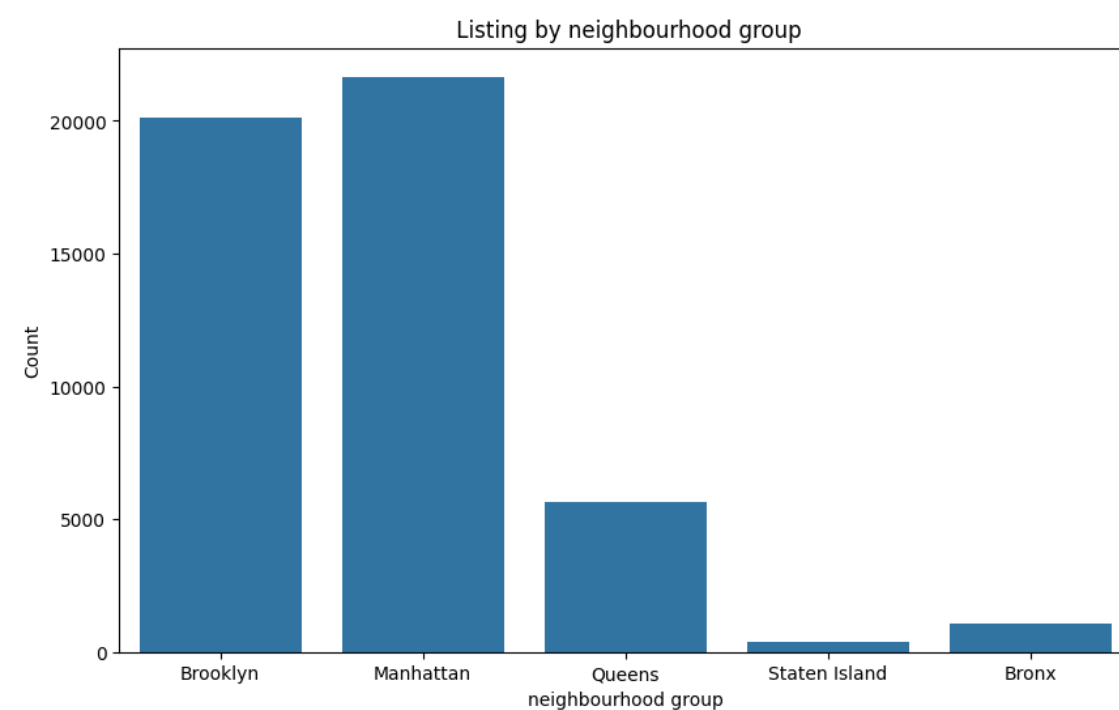
34

Consistency checking: Checking was made against values to ensure that features were of the expected formats.

```
plt.figure(figsize=(10, 6))
sns.histplot(df['price'], bins=50, kde=True)
plt.title('Distribution of Price')
plt.xlabel('Price')
plt.ylabel('Frequency')
plt.show()
```



```
plt.figure(figsize=(10, 6))
sns.countplot(data=df, x='neighbourhood_group')
plt.title('Listing by neighbourhood group')
plt.xlabel('neighbourhood group')
plt.ylabel('Count')
plt.show()
```



Univariate Analysis

Univariate analysis refers to the analysis of a variable one after another.

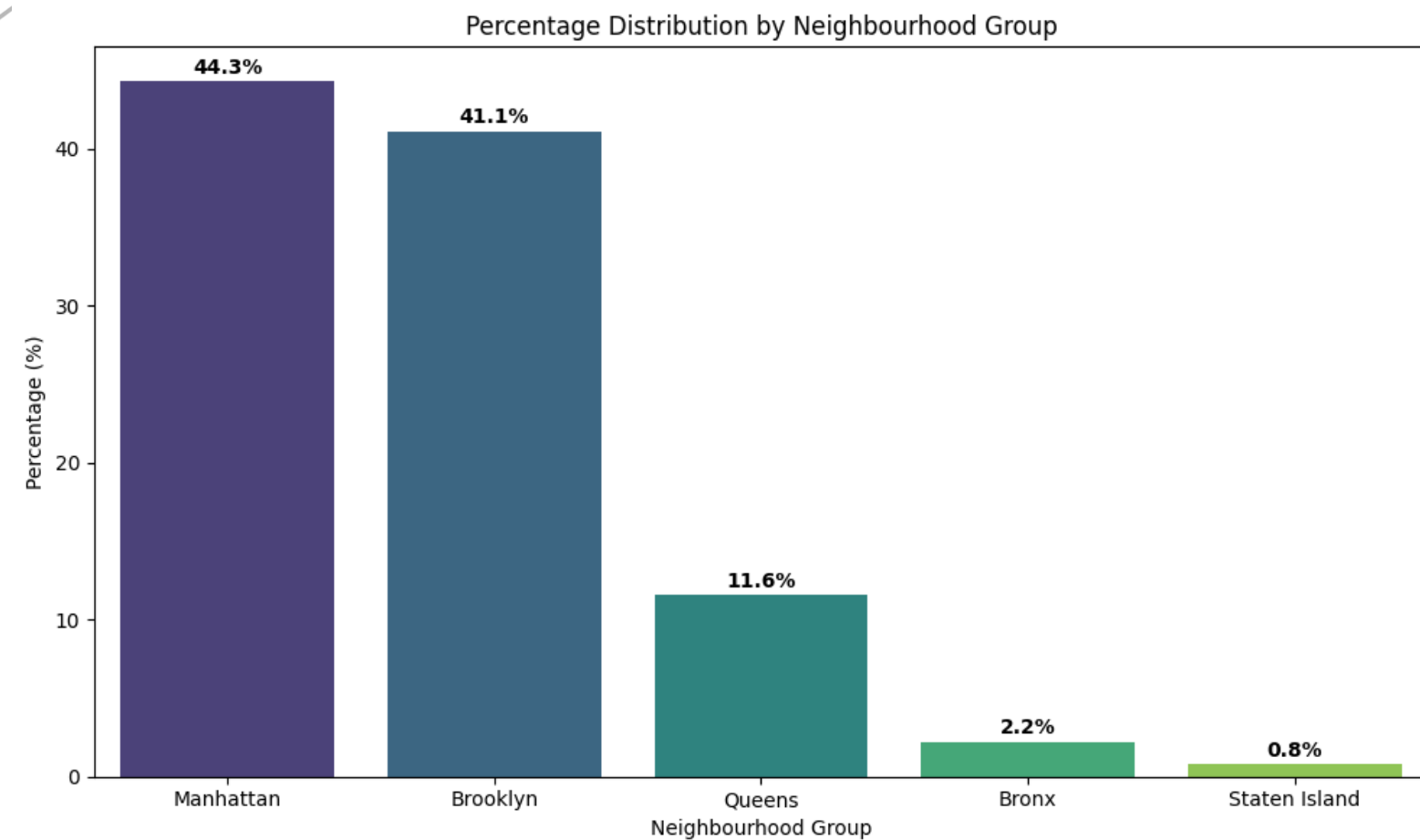
It assists in knowing the basic attributes and distribution of each column [8]

Categorical Variables

Neighbourhood Group:

The dataset indicates that there are five primary neighbourhood groups, namely Manhattan, Brooklyn, Queens, Bronx, and Staten Island. The greatest number of listings is in Manhattan and Brooklyn.

The listings are lower in Queens and the Bronx, and the lowest in Staten Island.



```
# Calculate percentage distribution
neigh_percent = df['neighbourhood_group'].value_counts(normalize=True) * 100
neigh_df = neigh_percent.reset_index()
neigh_df.columns = ['neighbourhood_group', 'percentage']

# Plot
plt.figure(figsize=(10, 6))
sns.barplot(
    data=neigh_df,
    x='neighbourhood_group',
    y='percentage',
    hue='neighbourhood_group',
    palette='viridis',
    legend=False
)

plt.title('Percentage Distribution by Neighbourhood Group')
plt.xlabel('Neighbourhood Group')
plt.ylabel('Percentage (%)')

# Add percentage labels
for i, v in enumerate(neigh_df['percentage']):
    plt.text(i, v + 0.5, f'{v:.1f}%', ha='center', fontweight='bold')

plt.tight_layout()
plt.show()
```

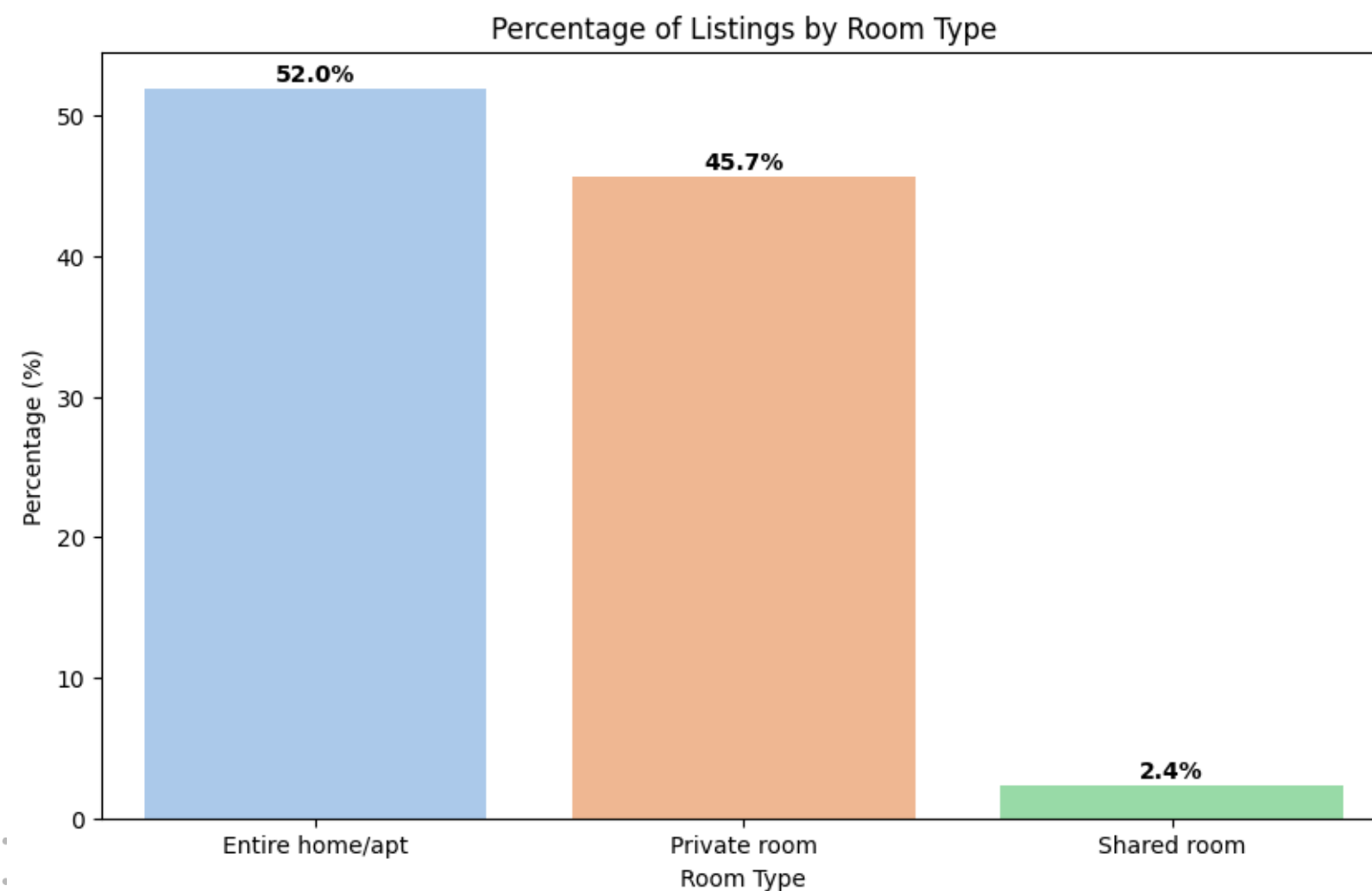
Neighbourhood Group	Percentage of Listings (%)
Manhattan	42.7
Brooklyn	42.4
Queens	11.8
The Bronx	2.26
Staten Island	0.811

```
# Calculate percentage
room_percent = df['room_type'].value_counts(normalize=True) * 100
room_df = room_percent.reset_index()
room_df.columns = ['room_type', 'percentage']

plt.figure(figsize=(10, 6))
sns.barplot(data=room_df, x='room_type', y='percentage', hue='room_type', legend=False, palette='pastel')
plt.title('Percentage of Listings by Room Type')
plt.xlabel('Room Type')
plt.ylabel('Percentage (%)')

# Add percentage labels
for i, v in enumerate(room_df['percentage']):
    plt.text(i, v + 0.5, f'{v:.1f}%', ha='center', fontweight='bold')

plt.show()
```



Univariate Analysis

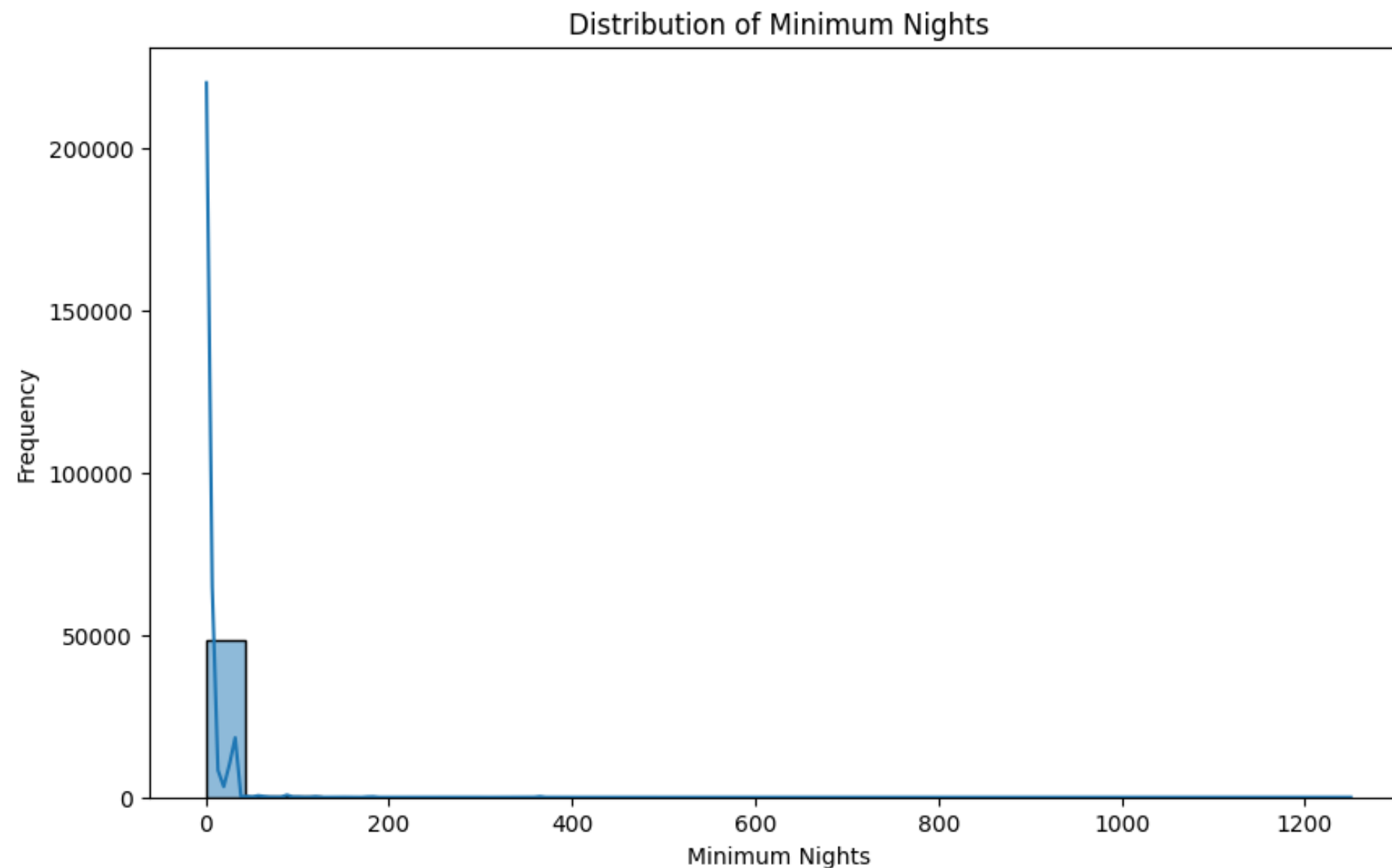
Room Type:

Entire home/apartment is the most popular type of room.

The second most common is the private rooms.

The least common are shared rooms.

That indicates that the users of Airbnb mostly choose full apartments or houses to be secluded.



Univariate Analysis

Numerical Variables

Price:

The prices are very low and very high.

The majority of the listings cost less than 300 per night. The high price few gives the long tail at the right side of the histogram (right-skewed).

Boxplots have numerous outliers in the price variable.

Minimum Nights:

There are numerous bookings with a minimum of 1-5 nights.

Minimum stays are set at high by some hosts, however.

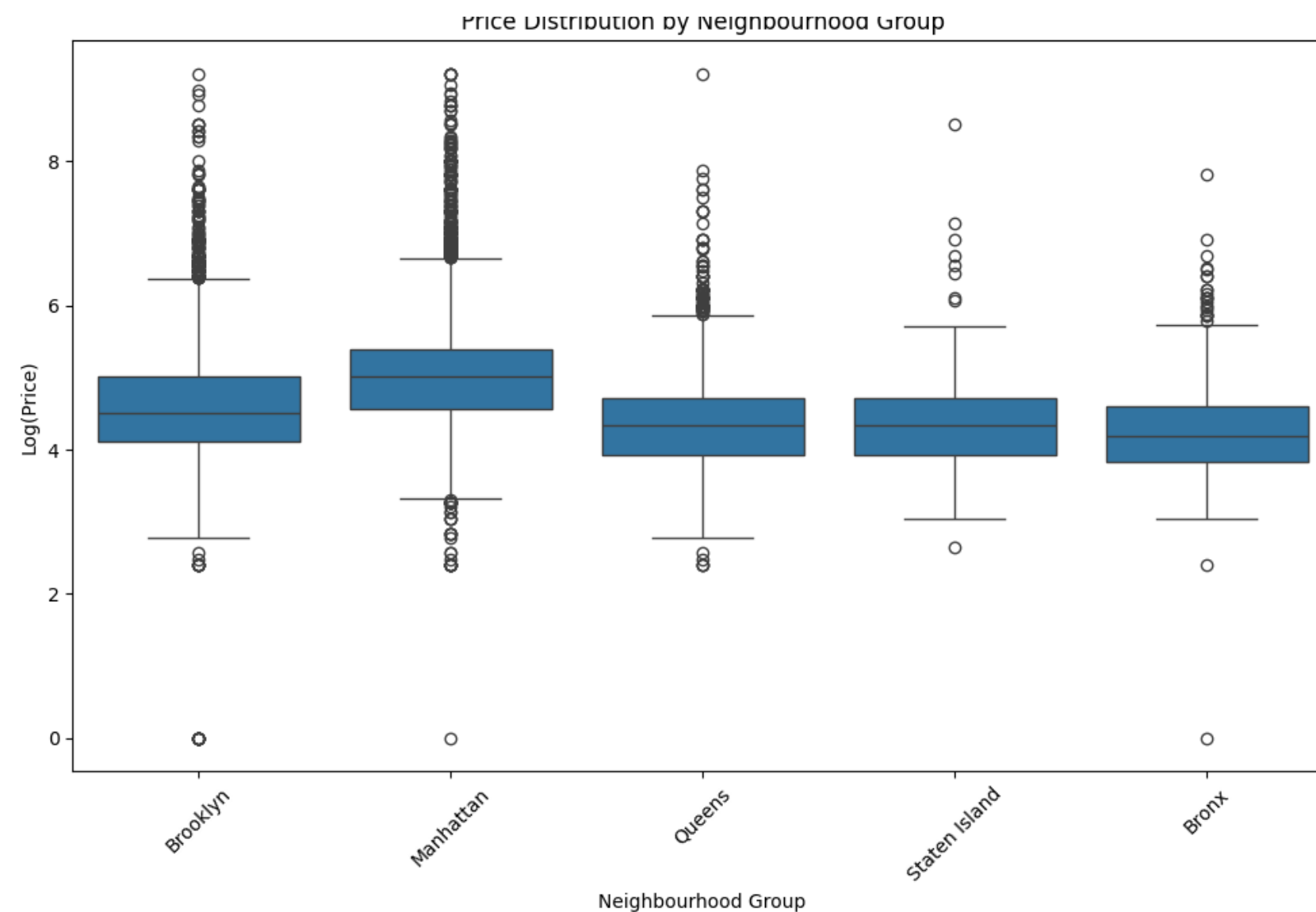
Reviews and availability:

The majority of the listings contain fewer than 100 reviews.

Availability changes; however, most of the listings offer over 150 days annually.

The following patterns can be used to comprehend the key traits of the Airbnb listings in New York.

```
# Relationship between 'neighbourhood_group' and 'price'
plt.figure(figsize=(12, 7))
sns.boxplot(x='neighbourhood_group', y='price', data=df)
plt.title('Price Distribution by Neighbourhood Group')
plt.xlabel('Neighbourhood Group')
plt.ylabel('Log(Price)')
plt.xticks(rotation=45)
plt.show()
```



Multivariate Analysis

The multivariate analysis assists in seeking the association among two or more variables [9]

Various comparisons were to be done in this project.

Price and Neighbourhood Group

The average price is in Manhattan.

Brooklyn is also second, with average prices.

Queens, Bronx and Staten Island are less expensive.

This reveals that place has a significant influence on price.

Room Type and Price

The priciest ones are complete houses.

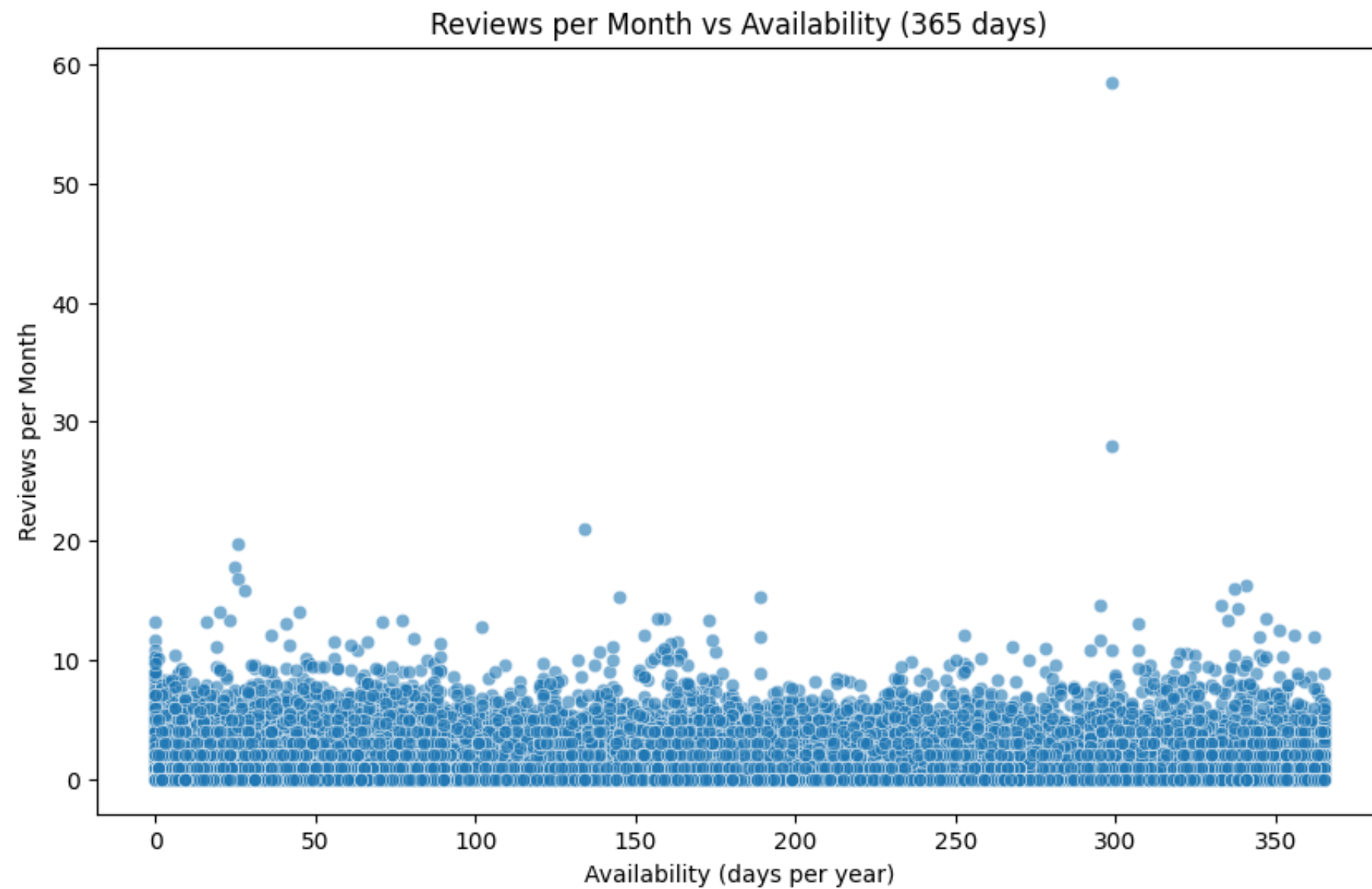
Private rooms cost much less.

The lowest-cost rooms are shared rooms.

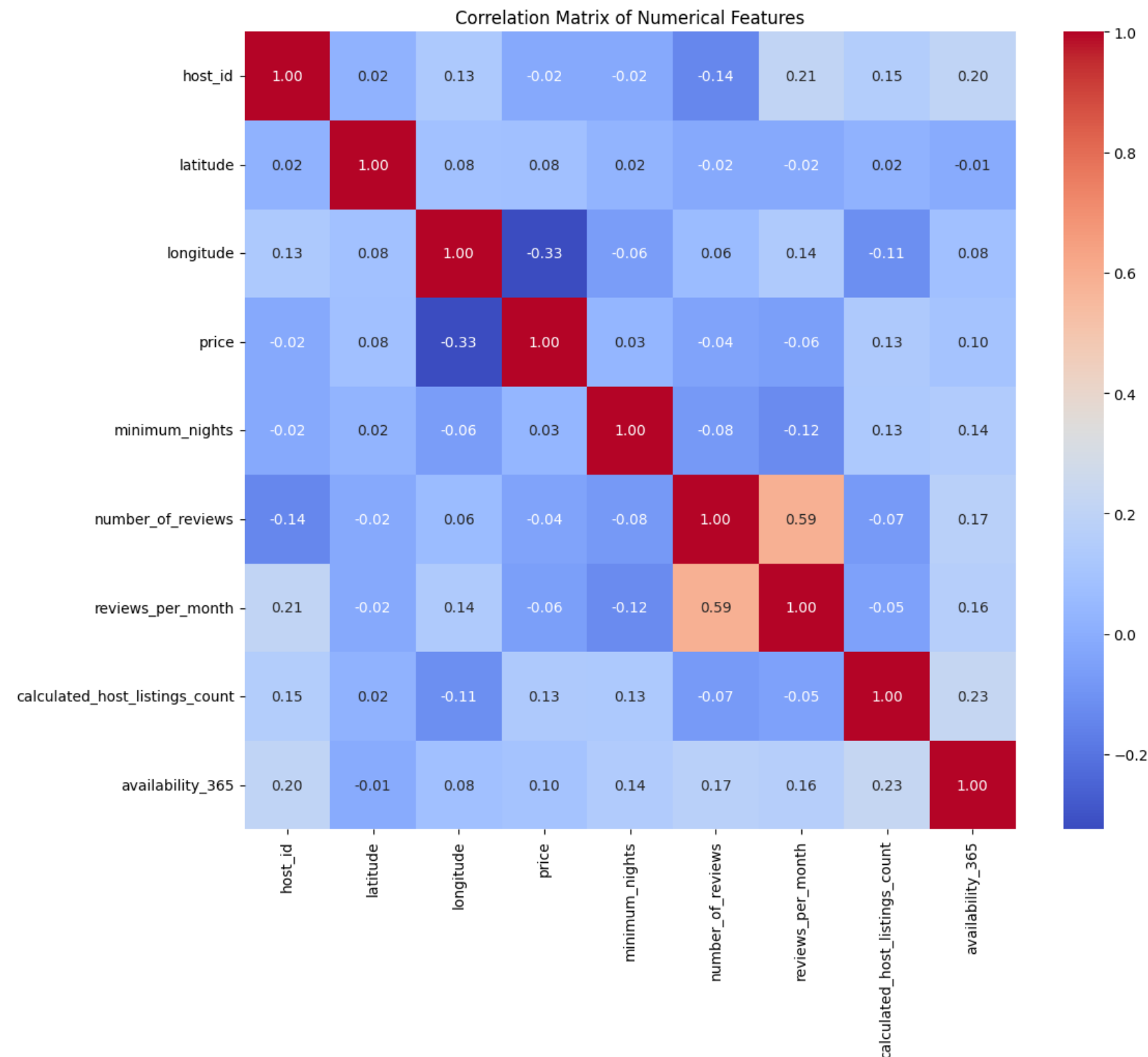
This trend can be found in every neighbourhood.

Multivariate Analysis

```
plt.figure(figsize=(10, 6))
sns.scatterplot(data=df, x='availability_365', y='reviews_per_month', alpha=0.6)
plt.title('Reviews per Month vs Availability (365 days)')
plt.xlabel('Availability (days per year)')
plt.ylabel('Reviews per Month')
plt.show()
```



```
# Correlation matrix for numerical features
plt.figure(figsize=(12, 10))
sns.heatmap(df.corr(numeric_only=True), annot=True, cmap='coolwarm', fmt=".2f")
plt.title('Correlation Matrix of Numerical Features')
plt.show()
```



Correlation Matrix

Correlation between Numerical Variables

The correlation heat map was made to observe any relationship between numerical columns.

number_of_reviews and reviews_per_month are positively related by a moderate association.

There is no significant correlation between price and reviews or availability, and hence other variables influence price more significantly.



```
features = ['reviews_per_month',
            'availability_365']

# Count missing values BEFORE cleaning
missing_before = df[features].isnull().sum()

# Cleaning: drop rows with missing values in any of these features
df_cleaned = df.dropna(subset=features)

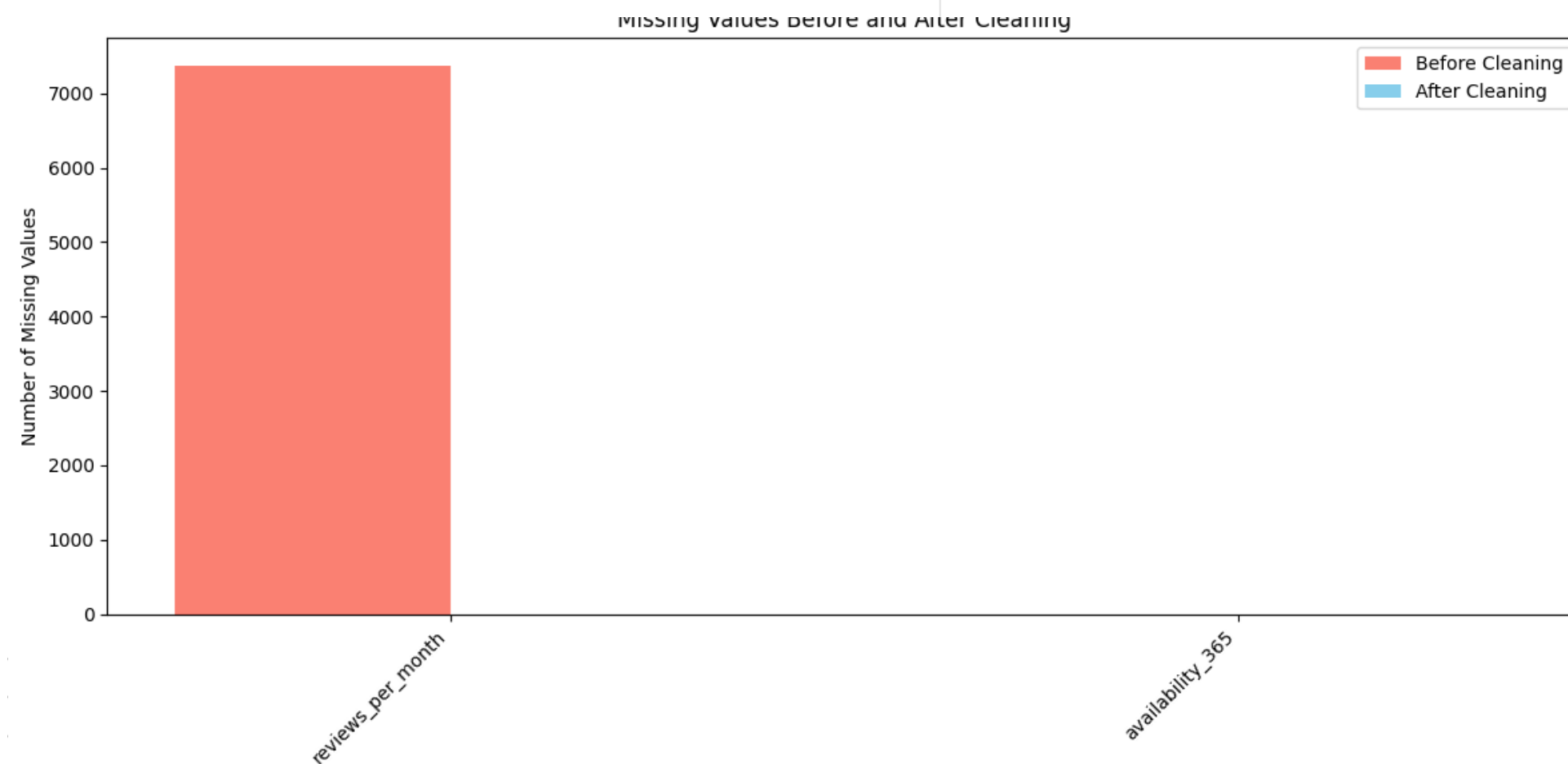
# Count missing values AFTER cleaning
missing_after = df_cleaned[features].isnull().sum()

# Prepare dataframe for plotting
missing_df = pd.DataFrame({
    'Feature': features,
    'Before Cleaning': missing_before.values,
    'After Cleaning': missing_after.values
})

# Plot side-by-side bar chart
x = range(len(features))
width = 0.35

fig, ax = plt.subplots(figsize=(12,6))
ax.bar(x, missing_df['Before Cleaning'], width=width, label='Before Cleaning', color='salmon')
ax.bar([i + width for i in x], missing_df['After Cleaning'], width=width, label='After Cleaning', color='skyblue')

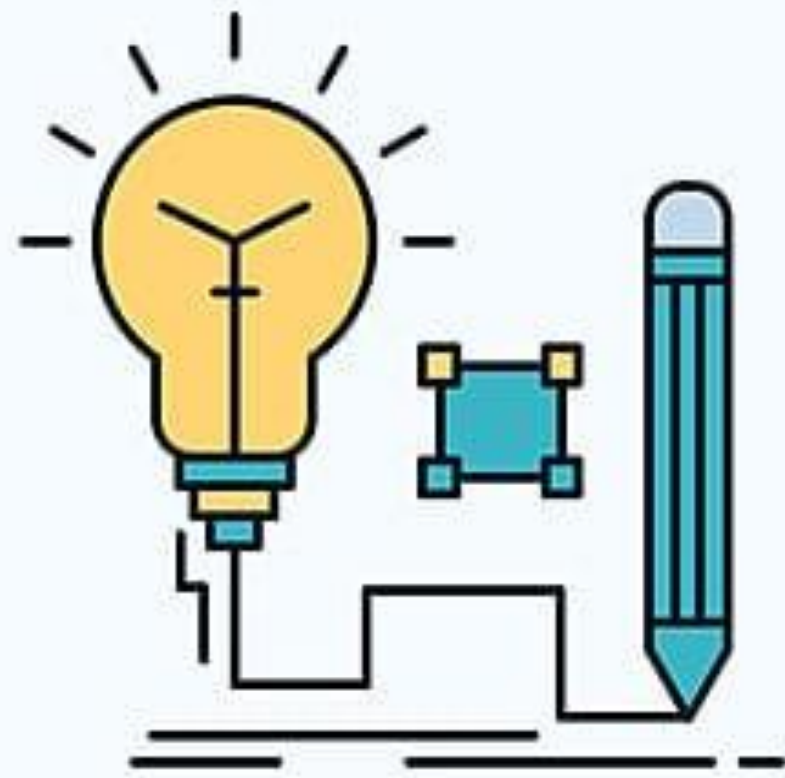
ax.set_xticks([i + width/2 for i in x])
ax.set_xticklabels(features, rotation=45, ha='right')
ax.set_ylabel("Number of Missing Values")
ax.set_title("Missing Values Before and After Cleaning")
ax.legend()
plt.tight_layout()
plt.show()
```



Missing Values Before and After Cleaning

Reviews_per month had missing values before cleaning but no missing values after cleaning.

Key Insight



The key results of the analysis are:

Location Matters:

- The highest amount of listings and prices is in Manhattan and Brooklyn.
- Staten Island and the Bronx are far less expensive.

Room Type Affects Price:

- House Dwellings are more costly.
- Shared rooms are the cheapest.

Availability and Reviews:

- Other listings have numerous days' available and hence low demand.
- A large number of reviews indicate good and inexpensive listings.

Price Range:

- Most prices are between \$50 and \$300.
- There have been a few high-end listings that exceeded 1000.

This information demonstrates that the market of Airbnb market in NYC is primarily central and tourist.

Recommendation



RECOMMENDATIONS

Based on the findings, the following recommendations are made:

For Hosts:

- Make competitive pricing based on location and room type.
- Enhance feedback by doing more to please customers and taking notes.
- Change the minimum nights to appeal to short-stay clients.

For Airbnb Platform:

- Demand has to be distributed by encouraging less popular regions such as the Bronx and Staten Island.
- Location-based pricing provides tools to assist hosts in setting fair prices.
- The number of missing data points can be encouraged to be filled in by hosts to enhance the quality of the listing.

For Future Analysis:

- Add seasonal information to verify the fluctuation of prices.
- Add host response rate and review scores to be analysed further.
- Estimating optimal prices uses clustering models or prediction models [10].

Limitations and Future Work



The process of EDA made the New York City Airbnb listings make sense.

It had shown that the kind of room and the neighbourhood are the leading factors which have an impact on the prices.

Most of the accommodation lies within Manhattan and Brooklyn which are the preferred destination of visiting tourists.

Information quality control, removing duplications and visualisation of the association offered a wonderful insight of how Airbnb works in NYC.

The study proves the importance of the cleaning up and visualisation of data in real decision making.

It can help Airbnb and hosts to plan better and acquire more customers and balance demand on the city level.

Executive Summary

This project analysed and examined Airbnb data from New York City in 2019 to examine and analyse it through EDA. The data were cleaned, visualised and analysed to perceive trends in price, location, and the room type. Manhattan and Brooklyn were found to have the most listings and the highest prices during the analysis. Another useful indicator is room type since the whole house will cost more than a personal or shared room. The methods of addressing missing increased the quality of the dataset. The visualisations also helped to see how Airbnb was distributed in the city and the price variance. The results can be beneficial to the hosts and airbnb to make better business choices.



References



1. Biswal, A. (2025) *What is exploratory data analysis? Steps and market analysis*, Simplilearn.com.
2. Malamahadevan (2025) *Exploratory Data Analysis (EDA) using Python*, Analytics Vidhya.
3. Kabir, M.A., Ahmed, F., Islam, M.M. and Ahmed, M.R. (2024) ‘Python for Data Analytics: A Systematic literature review of tools, techniques, and applications,’ *Academic Journal on Science, Technology, Engineering & Mathematics Education.*, 4(04), pp. 134–154.
4. Edwards, K. (2024) The Airbnb phenomenon: Analyzing its pioneering business Strategy - Keegan Edwards, Keegan Edwards. Available at: <https://keeganedwards.com/the-airbnb-phenomenon-analyzing-its-pioneering-business-strategy/>.
5. Alam, S., Ayub, M.S., Arora, S. and Khan, M.A. (2023) ‘An investigation of the imputation techniques for missing values in ordinal data enhancing clustering and classification analysis validity,’ *Decision Analytics Journal*, 9, p. 100341. Available at: <https://doi.org/10.1016/j.dajour.2023.100341>.
6. Ren, L., Wang, T., Seklouli, A.S., Zhang, H. and Bouras, A. (2023) “A review on missing values for main challenges and methods,” *Information Systems*, 119, p. 102268. Available at: <https://doi.org/10.1016/j.is.2023.102268>.
7. Nasima (2025) *Effective strategies for handling missing values in data analysis*, Analytics Vidhya.
8. Tessler, M. (2022) ‘Univariate analysis: variance, variables, data, and measurement,’ in *SpringerBriefs in sociology*, pp. 19–50.
9. Stevens, E. (2023) ‘An Introduction to Multivariate Analysis [With examples],’ *CareerFoundry*, 11 May.
10. Sáenz, J.V., Quiroga, F.M. and Bariviera, A.F. (2023) “Data vs. information: Using clustering techniques to enhance stock returns forecasting,” *International Review of Financial Analysis*, 88, p. 102657. Available at: <https://doi.org/10.1016/j.irfa.2023.102657>.



Thank
you!

