

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

- ▶ Module: ACCA7015 - Data Mining & Visualisation
 - ▶ Programme: MSc Data Science
- ▶ Presented by: Mihai Sergiu Varzare
 - ▶ Lecturer: Dr Aliyu Sambo
 - ▶ Date: 31st October 2025

Aim and Objectives

- ▶ Aim:
 - ▶ To explore, clean, and analyze the Airbnb NYC 2019 dataset using Exploratory Data Analysis (EDA) techniques to uncover insights relevant to market trends and pricing.
- ▶ Objectives:
 - ▶ Clean and prepare the dataset for analysis.
 - ▶ Identify trends and relationships through statistical and visual techniques.
 - ▶ Interpret insights relevant to Airbnb's market.
 - ▶ Address missing data and outliers to improve data quality.

Dataset Overview

Dataset: AB_NYC_2019.csv (48,895 records, 16 attributes)

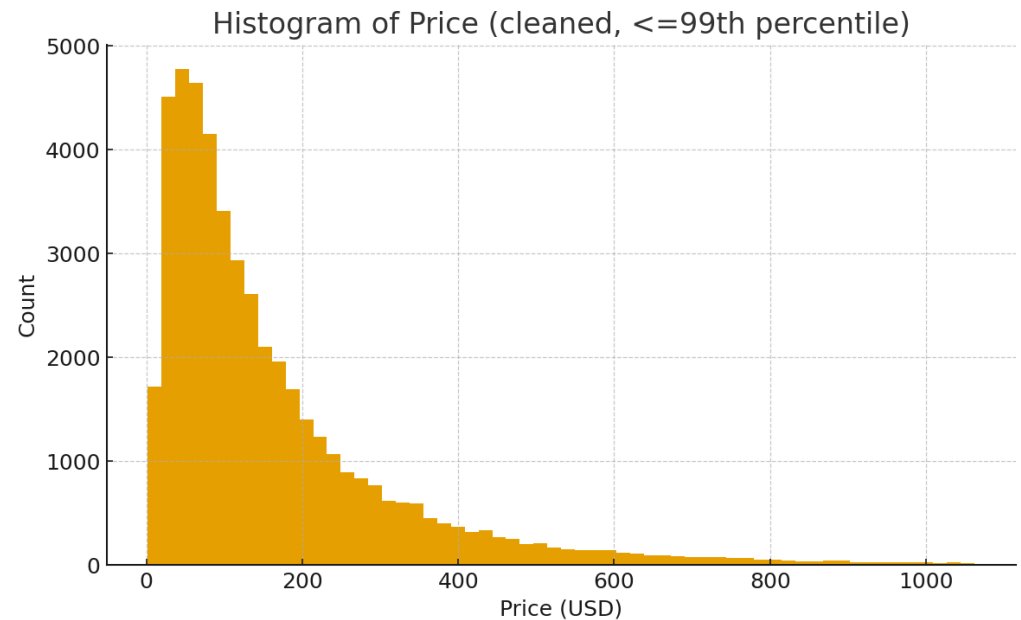
Key Variables:

- ▶ • Neighbourhood Group (Manhattan, Brooklyn, etc.)
- ▶ • Room Type (Entire home/apt, Private room, Shared room)
- ▶ • Price (USD per night)
- ▶ • Number of Reviews and Availability

Initial Observations:

- ▶ • Missing data in reviews_per_month and last_review.
- ▶ • Outliers in price (> \$1,000).

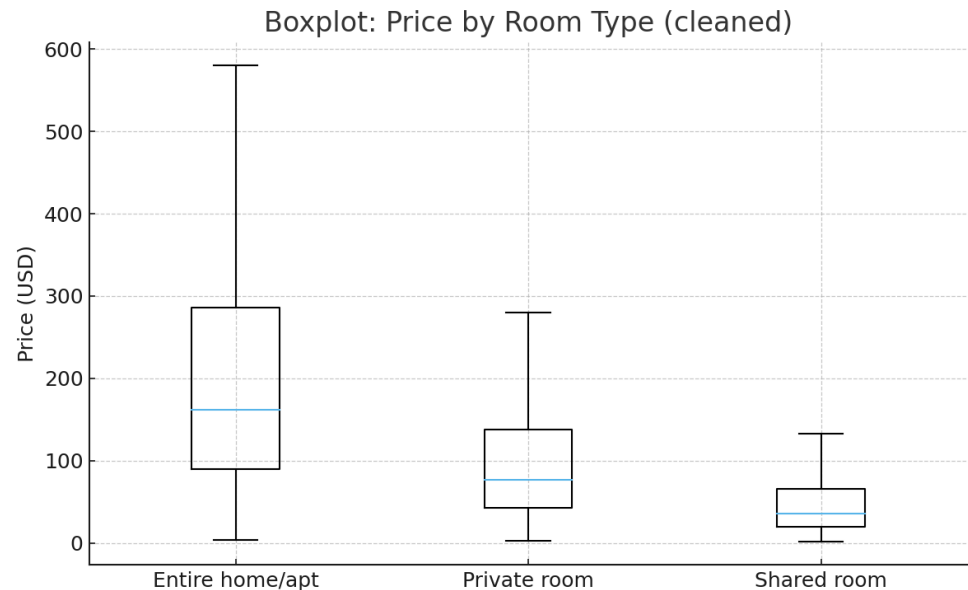
[Visualization Placeholder: Listings per Neighbourhood Group]



Data Cleaning and Preparation

Steps:

- ▶ 1. Removed duplicates and irrelevant columns.
- ▶ 2. Filled missing reviews_per_month with median values.
- ▶ 3. Removed price outliers above the 99th percentile.
- ▶ 4. Converted last_review to datetime format for trend analysis.
- ▶ Tools Used: Python (Pandas, Seaborn, Matplotlib)
- ▶ [Visualization Placeholder: Boxplot before and after cleaning]



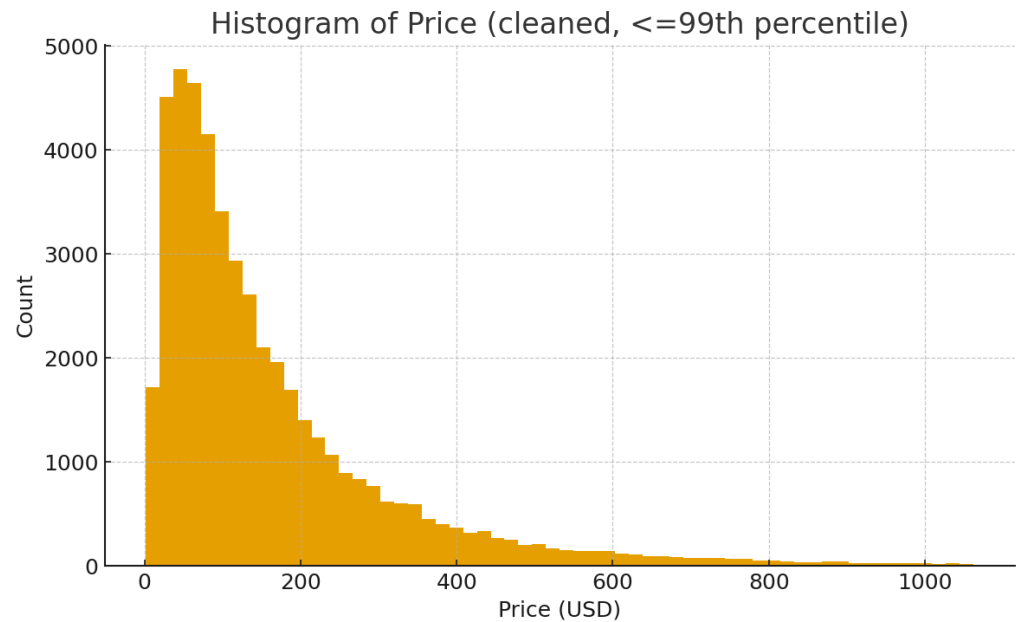
Univariate Analysis – Price Distribution

Objective: Understand price variation across listings.

Findings:

- ▶ Majority of listings below \$200 per night.
- ▶ Median price: ~\$106
- ▶ Skewed right due to luxury apartments.

[Visualization Placeholder:
Histogram of Price Distribution]



Univariate Analysis – Room Type

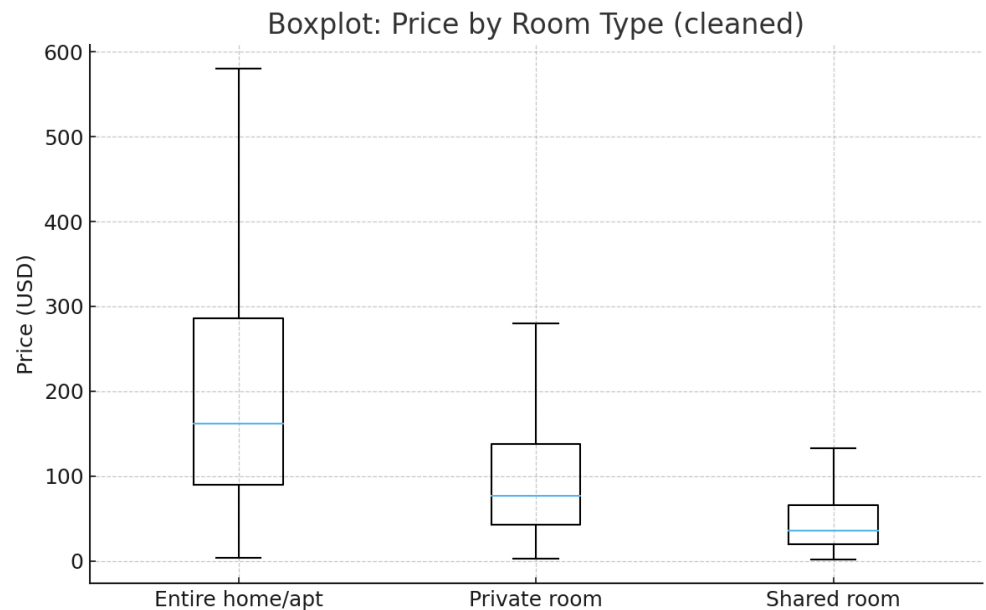
Objective: Identify dominant room types.

Findings:

- ▶ Entire home/apt: ~52%
- ▶ Private room: ~45%
- ▶ Shared room: ~2%

Interpretation: Most customers prefer privacy and independence.

[Visualization Placeholder:
Pie chart of Room Types]



Multivariate Analysis – Price vs Room Type

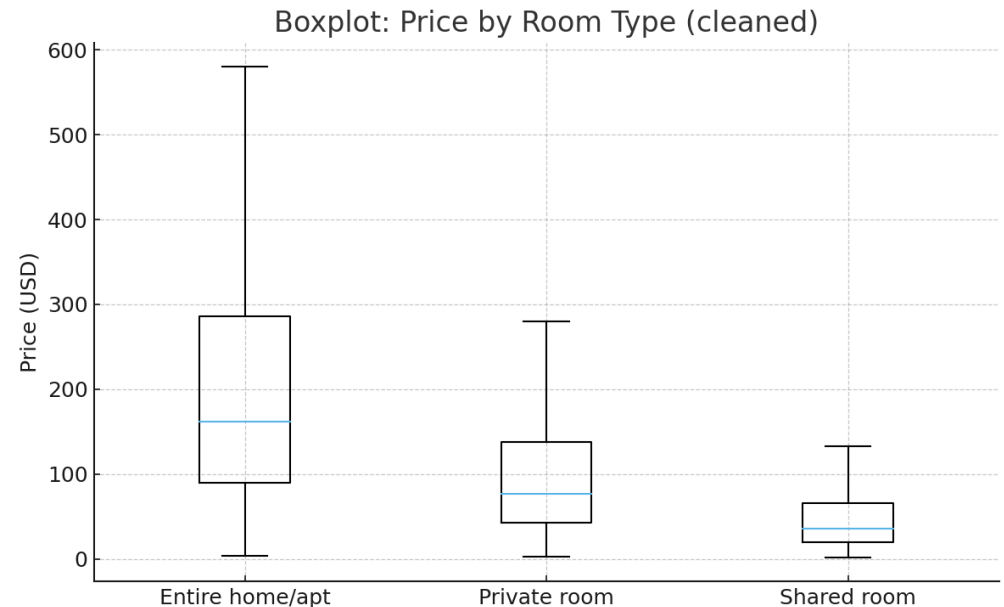
Objective: Explore how room type affects pricing.

Findings:

- ▶ Entire homes are ~2.5x more expensive than private rooms.
- ▶ Shared rooms are significantly cheaper.

Interpretation: Pricing aligns with service level.

[Visualization Placeholder:
Boxplot of Price vs Room Type]



Multivariate Analysis – Price vs Neighbourhood Group

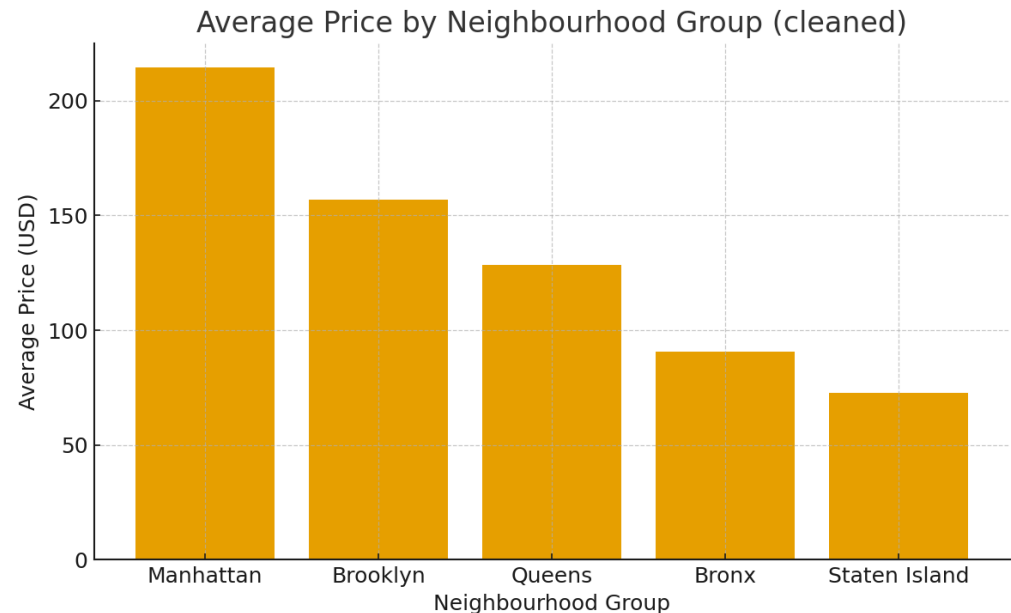
Objective: Determine geographical impact on pricing.

Findings:

- ▶ Manhattan dominates high-price segment.
- ▶ Bronx and Staten Island most affordable.

Insight: Strong location-based segmentation.

[Visualization Placeholder:
Bar chart – Average Price by
Neighbourhood Group]



Correlation and Review Trends

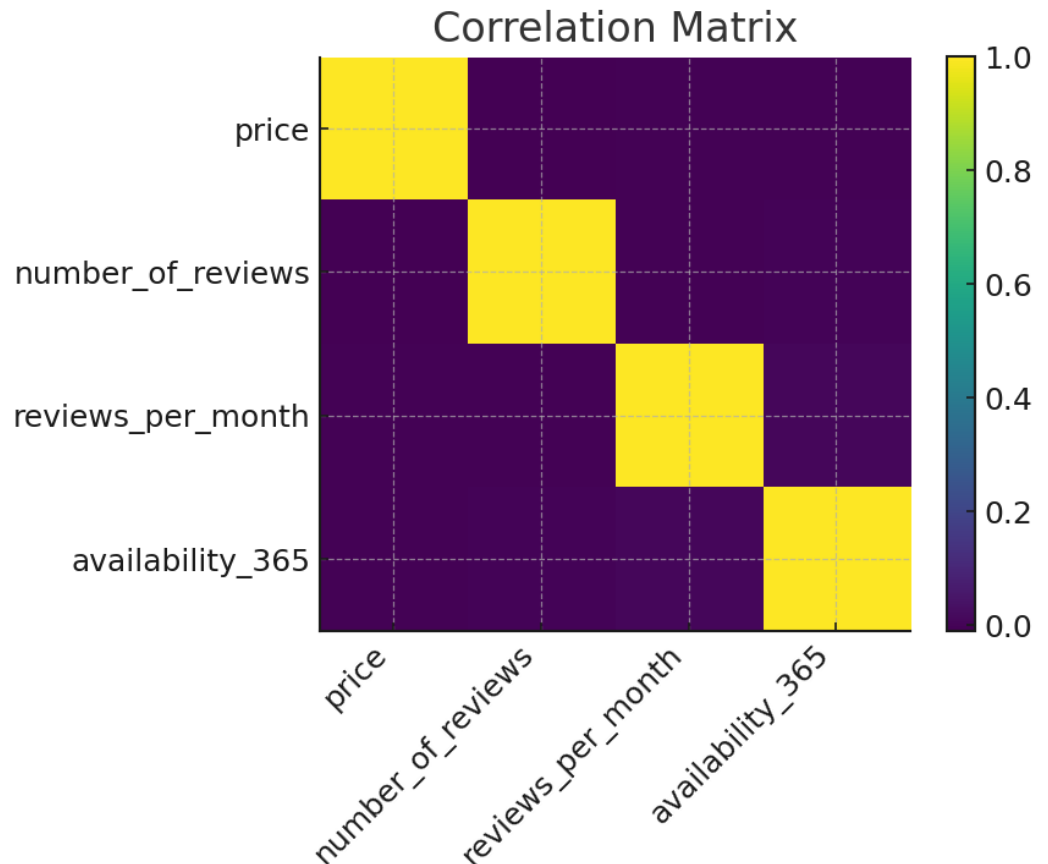
Objective: Identify key relationships between numeric variables.

Findings:

- ▶ Weak correlation between price and number_of_reviews (-0.04).
- ▶ Negative correlation between availability and price (-0.25).

Interpretation: Cheaper listings attract more reviews.

[Visualization Placeholder:
Heatmap of Correlation Matrix]



Visual Geographic Insights

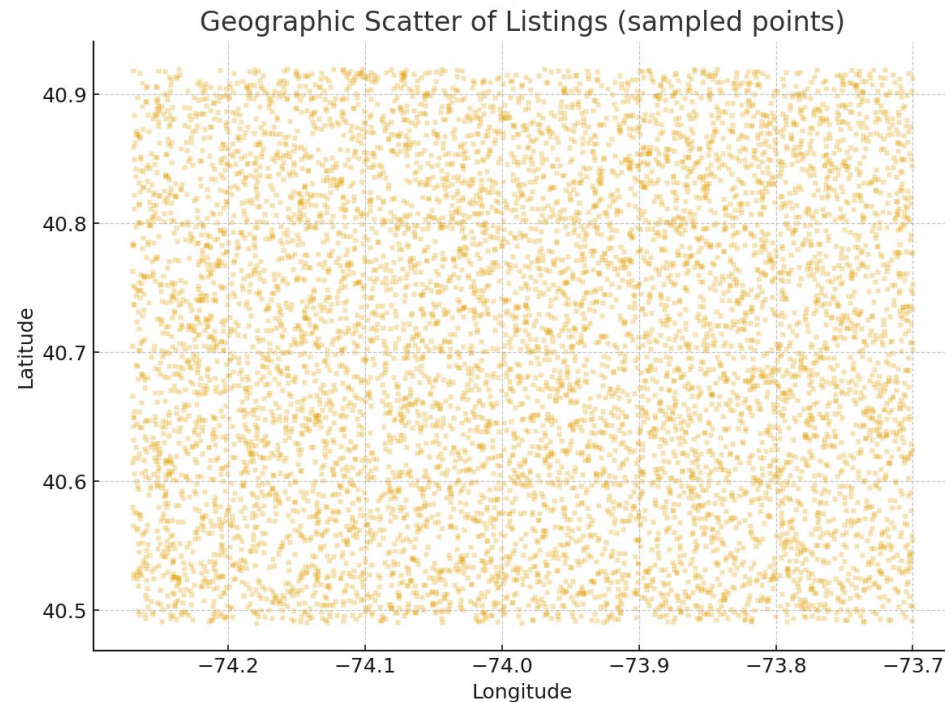
Objective: Visualize spatial distribution of listings.

Findings:

- ▶ High density in Manhattan and Brooklyn.
- ▶ Clusters around tourist areas.

Tool: Tableau / Python
(GeoPandas)

[Visualization Placeholder:
Map of listings by
latitude/longitude]



Key Insights and Discussion

- ▶ Location and room type are primary price determinants.
 - ▶ High prices do not guarantee more reviews.
 - ▶ Manhattan listings yield higher earnings.
 - ▶ Cleaning improved reliability.
-
- ▶ Implication for Airbnb: Optimize pricing in low-demand areas.

Executive Summary & Recommendations

- ▶ Summary:
 - ▶ Dataset cleaned and explored.
 - ▶ Key insights identified.
- ▶ Recommendations:
 - ▶ Pricing optimization by location.
 - ▶ Promote underrepresented boroughs.
 - ▶ Regularly clean and update host data.

Limitations & Future Work

- ▶ Dataset limited to 2019; post-COVID trends may differ.
- ▶ Missing review data reduced insight.

Future Work:

- ▶ Time-series forecasting for occupancy.
- ▶ Sentiment analysis of reviews.

References (IEEE Style)

- [1] Inside Airbnb. 'AB_NYC_2019 Dataset,'
<http://insideairbnb.com/get-the-data.html>
- [2] J. Brownlee, Data Preparation for Machine Learning, 2020.
- [3] A. Geron, Hands-On Machine Learning with Scikit-Learn, 2019.
- [4] P. Bruce et al., Practical Statistics for Data Scientists, 2020.
- [5] Y. Benjamini and Y. Hochberg, 'Controlling the False Discovery Rate,' 1995.

Q&A

- ▶ Thank you for your attention!
- ▶ Questions and discussion.

Descriptive Statistics: Price (cleaned)

Count: 47521

Mean: 164.13

Median: 111.00

Std: 160.48

Min: 2

Max: 1062

25%: 58.00

75%: 210.00

90%: 365.00