# AIRBNB NYC 2019 DATASET - EDA & INSIGHTS ANALYSIS

Student name: Muhammad Haroon
Student ID: 2434025

**Course: ACCA7015 - Data Mining & Visualisation**

Prifysgol Cymru
Y Drindod Dewi Sant
University of Wales
Trinity Saint David

# TABLE OF CONTENTS


- Project Overview & Objectives
- The Data at a Glance
- Data Cleaning & Preprocessing
- Univariate Analysis - Key Distributions
- Statistical Summary
- Multivariate Analysis
- Advanced Analysis - Clustering and Segmentation
- Geographical Analysis
- Summary and Conclusion
- References

# PROJECT OVERVIEW & OBJECTIVES

- Data Mining and Visualization project using real life data set.
- Dataset: AB_NYC_2019.csv - This dataset consists of listing data of New York City Airbnb.

**Purpose**

- The goal of this objective is to recognize patterns, trends and information with regards to the NYC Airbnb market through a logical analysis of data.
- Research Question(s):How well and generally structured is the data?
- How are the listing prices distributed in NYC?
- What are the most popular types and boroughs of rooms?
- What are the peculiarities of an average listing?

# THE DATA AT A GLANCE

- Initial Data Overview
- Source: AB_NYC_2019.csv
- Size: 48,895 listings (rows) with 16 attributes (columns).
- Memory Usage: ~21.3 MB

**What's in the Dataset? Key Columns Include:**

- id, name, host_id, host_name
- neighbourhood_group (Borough), neighbourhood
- latitude, longitude
- room_type (Entire home, Private room, Shared room)
- price, minimum_nights
- number of reviews, number of reviews per month and last review.
- availability_365, calculated_host_listings_count



# DATA CLEANING & PREPROCESSING

Before data analysis the data shall be cleaned;



1. Identifying Missing Values:
   - **Action:** Systematically check for NaN or null values.
   - **Our Finding:**
     - name (16 missing)
     - host_name (21 missing)
     - last_review (10,052 missing)
     - reviews_per_month (10,052 missing)

2. Implementing Cleansing Strategies:
   - **For** last_review **&** reviews_per_month**:**
     - **Strategy:** Missing = No Reviews
     - **Action:** Fill last_review with 'No reviews', reviews_per_month with 0
   - **For** name **&** host_name**:**
     - **Strategy:** Preserve record with placeholder
     - **Action:** Fill with 'No name provided', 'Unknown host'

# DATA CLEANING & PREPROCESSING

3. Handling Outliers & Invalid Data:

- **Problem:** Prices ranged from $0 to $10,000 (unrealistic extremes)
- **Solution:** Apply domain knowledge to filter
- **Action:** Keep only listings priced between **$10 and $1,000**
- **Result:** Dataset reduced from 48,895 to **48,645 clean records**

```
============================================================
DATA CLEANING AND PREPROCESSING
============================================================

1. MISSING VALUES ANALYSIS:
                  Missing Count  Missing Percentage
name                         16            0.032723
host_name                    21            0.042949
last_review               10052           20.558339
reviews_per_month         10052           20.558339

2. DATA QUALITY ISSUES ADDRESSED:
   • Filled missing last_review with 'No reviews'
   • Filled missing reviews_per_month with 0
   • Filled missing name with 'No name provided'
   • Filled missing host_name with 'Unknown host'

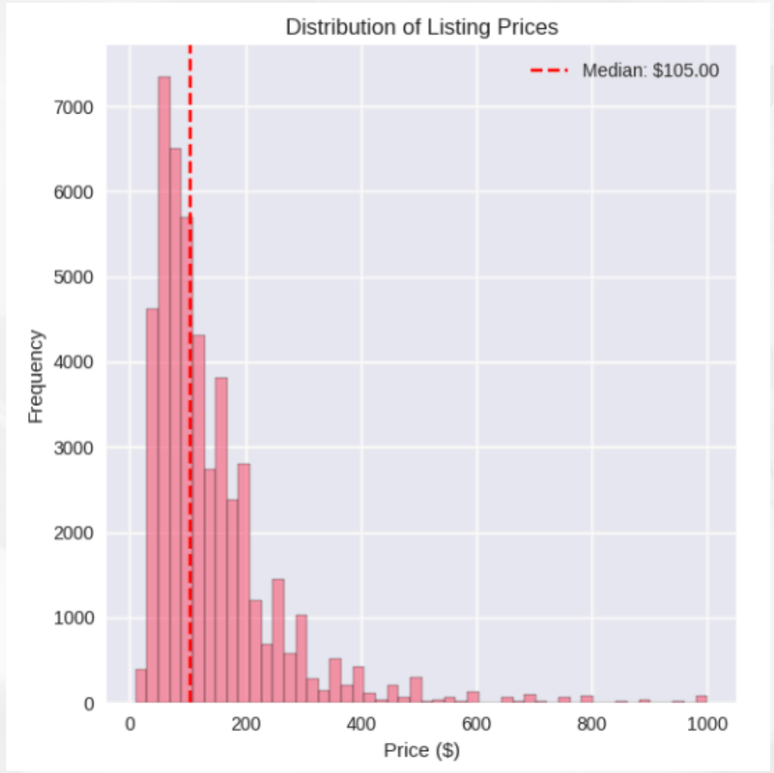
3. DUPLICATE RECORDS: 0

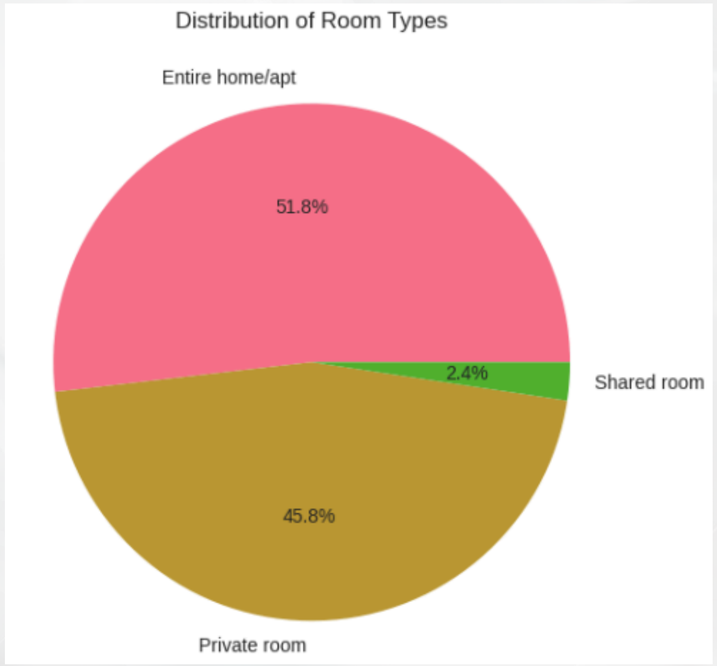
4. OUTLIER ANALYSIS - PRICE:
   • Original price statistics:
     Min: $0.00
     Max: $10000.00
     Mean: $152.72
     Median: $106.00
   • After removing outliers (<$10 and >$1000):
     Records remaining: 48,645
     New price range: $10.00 - $1000.00
```

# UNIVARIATE ANALYSIS - KEY DISTRIBUTIONS

## Distribution of Listing Prices

- **Price Distribution:** Heavily right-skewed.
- **Median Price:** ~$106 (after cleaning).
- **Insight:** Most listings are affordable, with a long tail of luxury properties.



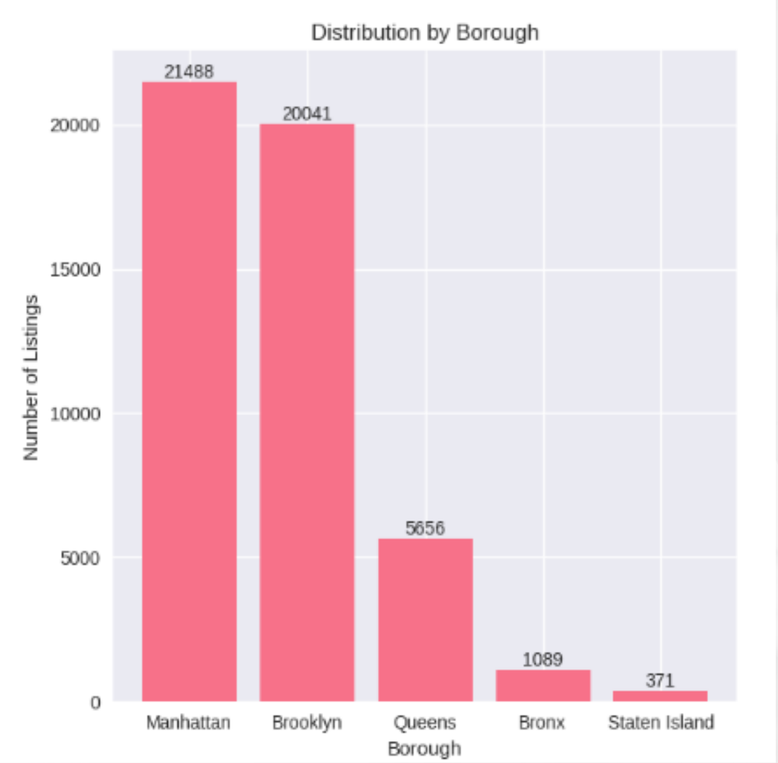

## Distribution of Room Types

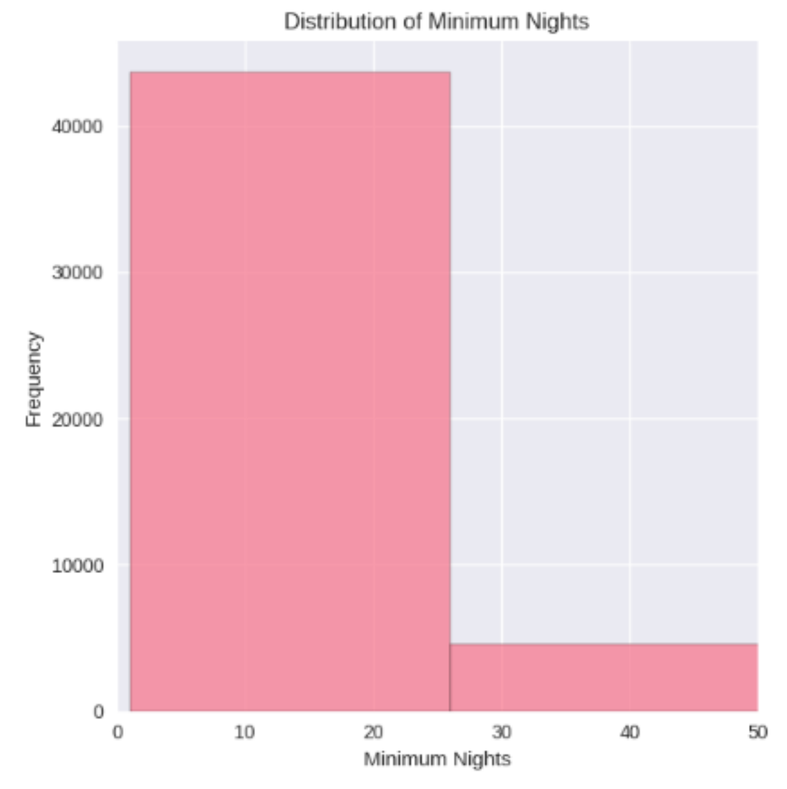
- **Entire home/apartment:** ~52%
- **Private room:** ~46%
- **Shared room:** ~2%
- **Insight:** The market is dominated by entire homes and private rooms.

# UNIVARIATE ANALYSIS - KEY DISTRIBUTIONS

**Distribution by Borough**

- Manhattan & Brooklyn dominate the market, containing the vast majority of listings.
- Queens, Bronx, and Staten Island have significantly fewer offerings.




**Distribution of Minimum Nights**

- Most common requirement is 1-7 nights.
- A small number of listings require very long stays (e.g., 30+ nights).

# STATISTICAL SUMMARY

- **Numerical Summary of Key Metrics (After Cleaning)**

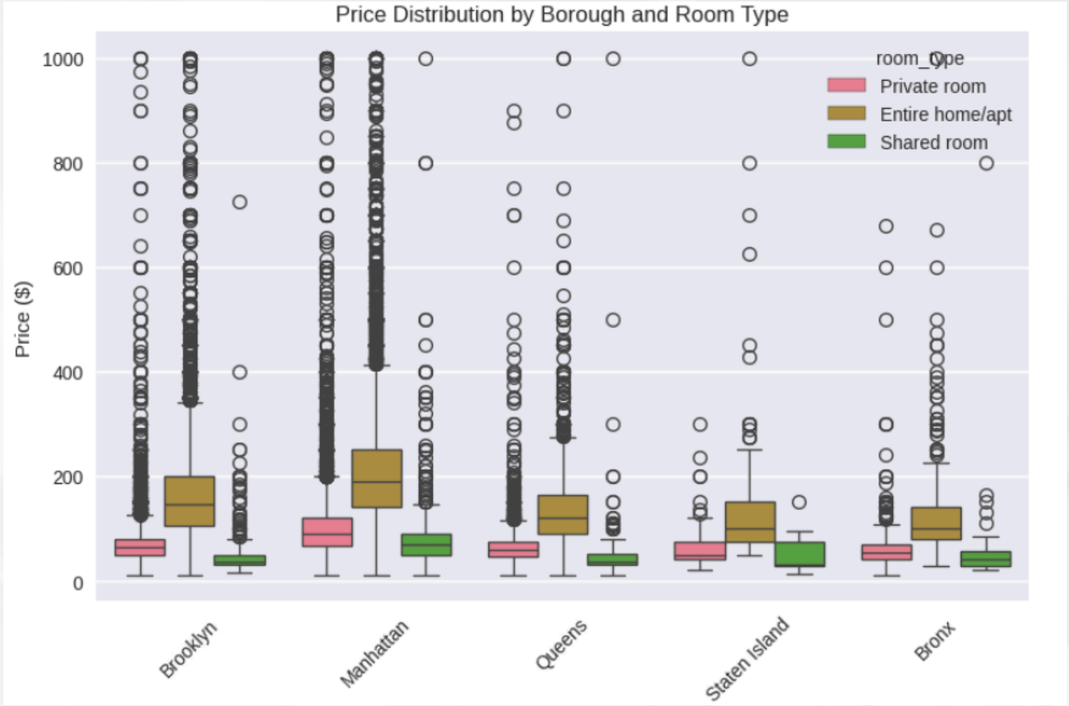
```
3. STATISTICAL SUMMARY:
              price  minimum_nights  number_of_reviews  availability_365
count  48645.000000    48645.000000       48645.000000      48645.000000
mean     141.311789        6.973461          23.354836        112.442944
std      116.731330       20.267263          44.631191        131.434828
min       10.000000        1.000000           0.000000          0.000000
25%       69.000000        1.000000           1.000000          0.000000
50%      105.000000        3.000000           5.000000         44.000000
75%      175.000000        5.000000          24.000000        226.000000
max     1000.000000     1250.000000         629.000000        365.000000
```

The high standard deviations indicate significant variability in the market.

# MULTIVARIATE ANALYSIS

**Price Distribution by Borough and Room Type**

- Univariate: Studying the distribution of price alone.
- Bivariate: Studying how price changes with neighbourhood_group.
- Multivariate: Studying how price is simultaneously influenced by neighbourhood_group, room_type, number_of_reviews, and availability_365.







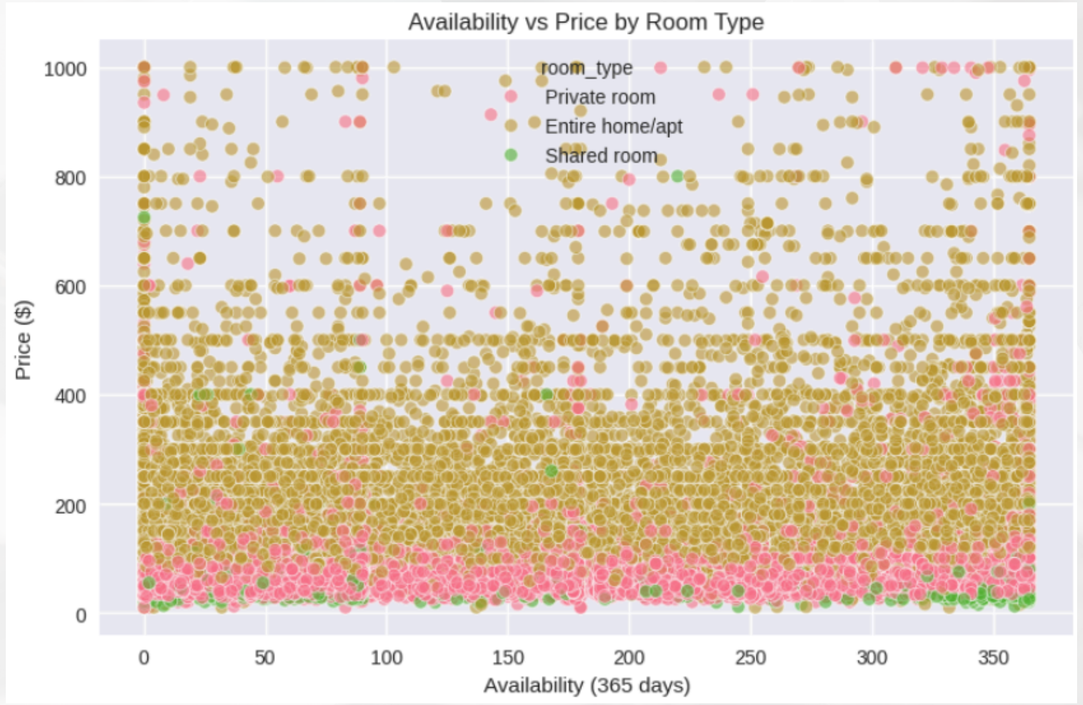
**Price Vs Number of Reviews**

- It shows a dense concentration of data points at low prices and a low number of reviews, with the number of reviews generally decreasing as the price increases. The majority of items have prices under $200 and fewer than 200 reviews.

# MULTIVARIATE ANALYSIS

**Availability vs Price by Room Type**
visualizes the relationship between Availability (365 days) (on the x-axis) and Price ($ on the y-axis), categorized by three room types:

- **Entire home/apt (Brown/Gold):** These are the most frequent type, spread across all availability and price points, including the highest prices (up to $1000).
- **Private room (Pink):** These are also very frequent but are overwhelmingly concentrated at lower price points (mostly below $200) across all availability levels.
- **Shared room (Green):** These are the least frequent and are exclusively found at the lowest price points (near $0).



# MULTIVARIATE ANALYSIS

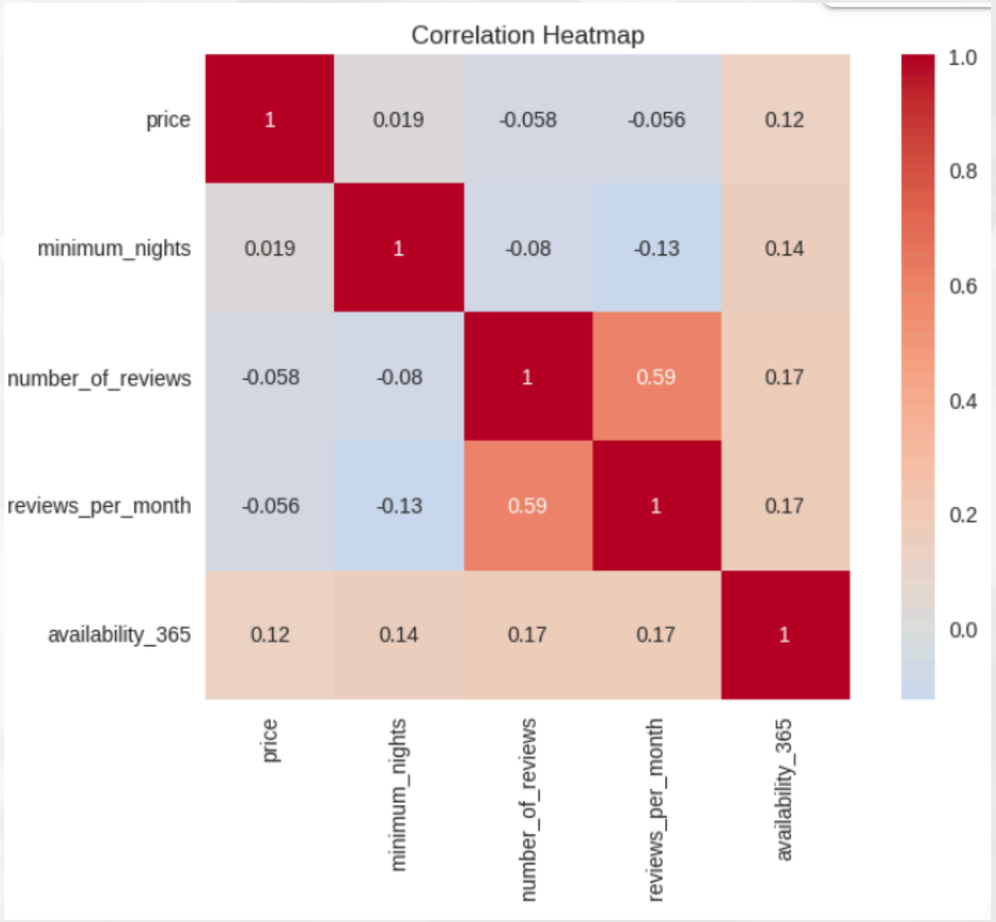
**Correlation Heatmap**

This Correlation Heatmap showing the linear relationship between five variables: price, minimum_nights, number_of_reviews, reviews_per_month, and availability_365.

The key relationships indicated by the color intensity and numerical values are:

- Strong Positive Correlation:
  - Number of reviews and reviews_per_month (0.59) are strongly correlated, which is expected.
- Weak/Negligible Correlations (The rest):
  - Price has a very weak positive correlation with availability_365 (0.12) and nearly no correlation with the other variables.
  - Minimum_nights has very weak correlations with all other variables.
- Weak Negative Correlation:
  - Number_of_reviews and minimum_nights (-0.08) have a very weak negative relationship.

Overall, the heatmap suggests that price and minimum_nights are largely independent of the other listed features.

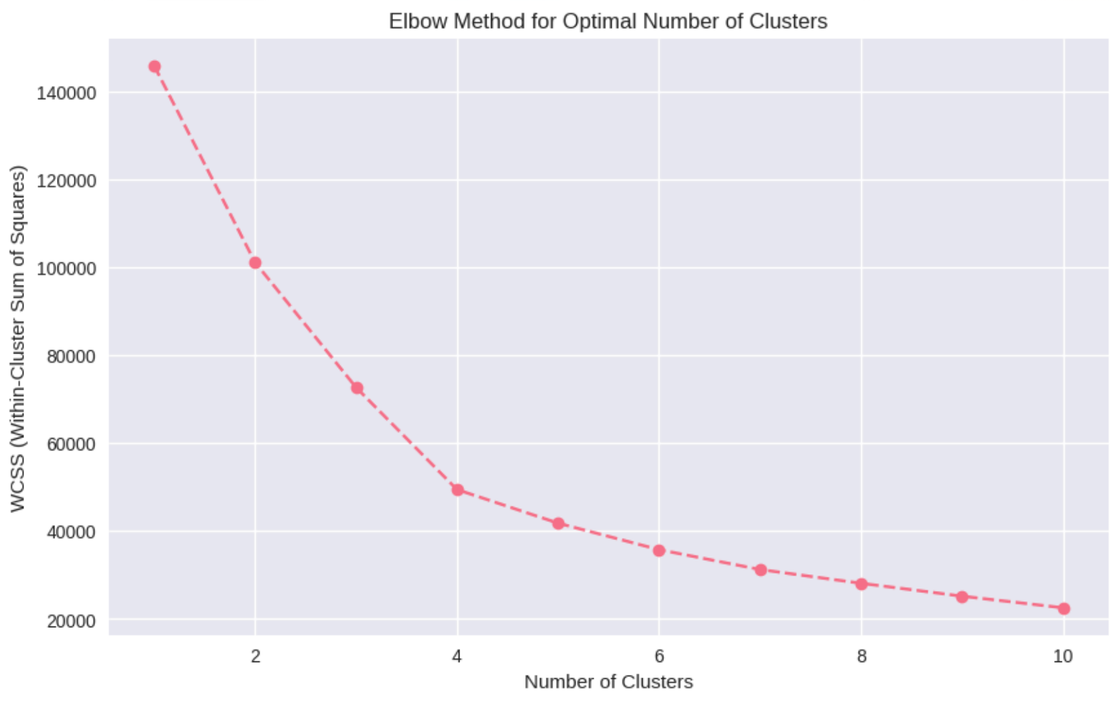


# ADVANCED ANALYSIS - CLUSTERING AND SEGMENTATION

**Elbow Method for Optimal Number of Clusters**

Used to determine the ideal number of clusters for K-Means clustering.

- The plot shows WCSS (Within-Cluster Sum of Squares) on the y-axis decreasing as the Number of Clusters (K) on the x-axis increases.
- The "elbow" point, where the rate of decrease in WCSS significantly slows down, appears to be at K = 3 or possibly K = 4.
- This elbow suggests that 3 or 4 is the optimal number of clusters for the underlying data, as adding more clusters beyond this point provides diminishing returns in reducing the WCSS.



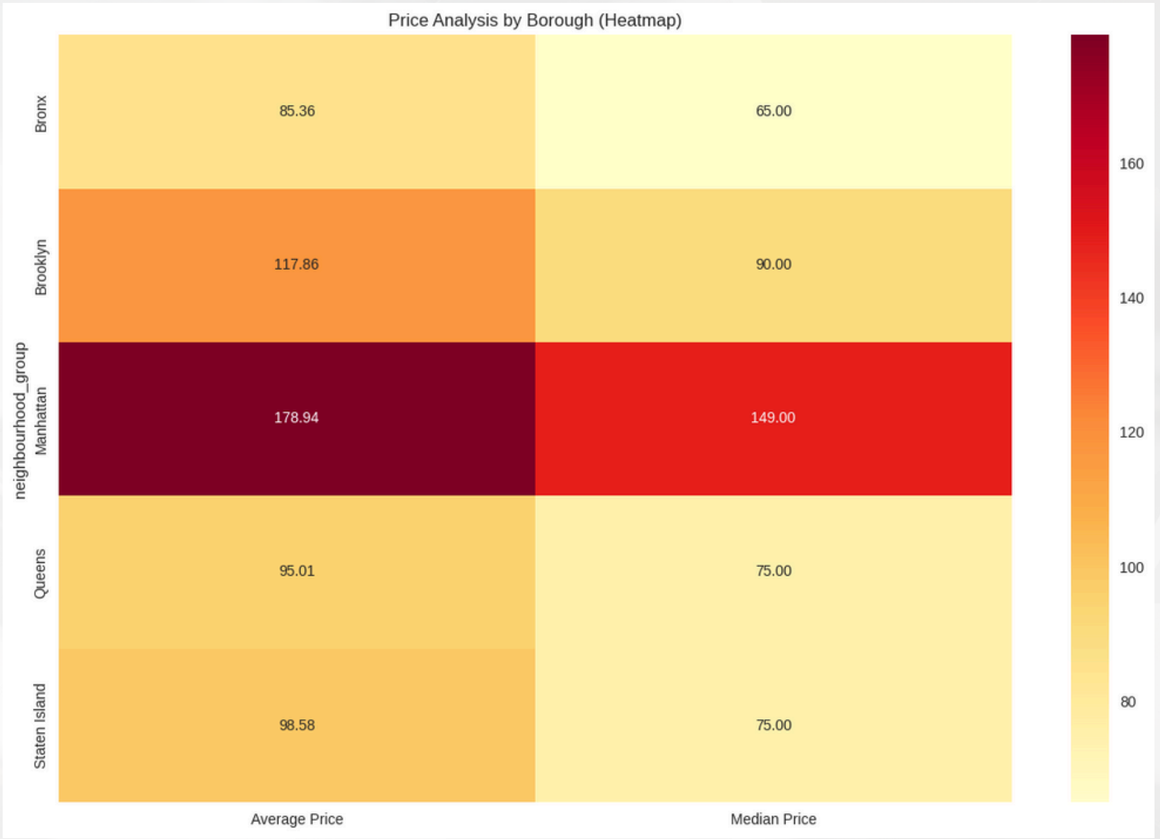

```
5. CLUSTERING ANALYSIS:
   • Optimal number of clusters: 4 (based on elbow method)
   • Cluster distribution:
cluster
0    29043
1     3082
2     3623
3    12897
Name: count, dtype: int64
```

# GEOGRAPHICAL ANALYSIS

**Heatmap - Price Analysis by Borough**

It is a comparison on the Average Price and Median Price in five boroughs (neighborhood groups) that are Manhattan, Brooklyn, Queens, Bronx, and Staten Island.
The data is visualized in the form of a heatmap: darker red means higher price.

- The darkest red color indicates the highest average and median price in Manhattan at an average of 178.94 and a median of 149.00.
- Brooklyn has the second-best average price of 117.86 and the median value of 90.00.
- Bronx boasts the lowest average price of the counties ( $85.36) and is equal to the median prices of Queens and Staten Island in the median price ( $65.00/75.00).
- As can be clearly seen in the chart, the prices are considerably higher in Manhattan than in all other boroughs.




7. GEOGRAPHICAL INSIGHTS:

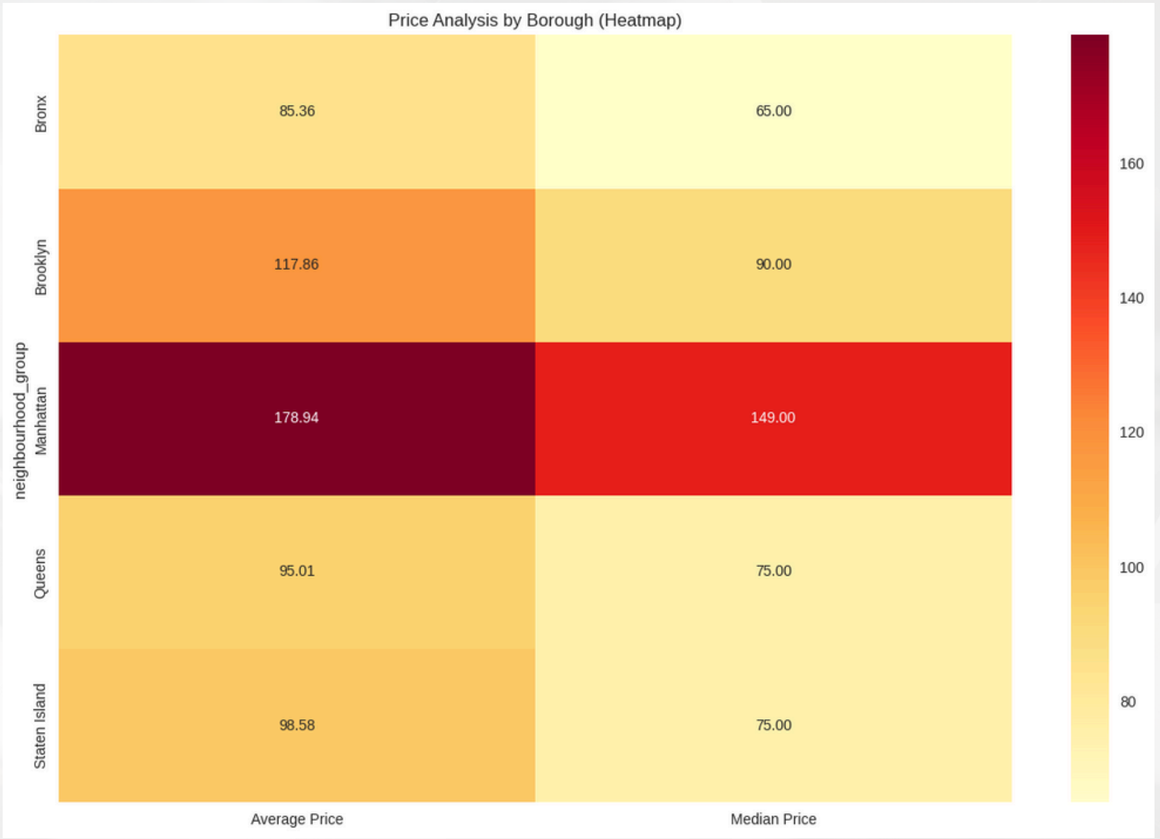
| neighbourhood_group | Average Price | Median Price | Listing Count |
|---|---|---|---|
| Bronx | 85.36 | 65.0 | 1089 |
| Brooklyn | 117.86 | 90.0 | 20041 |
| Manhattan | 178.94 | 149.0 | 21488 |
| Queens | 95.01 | 75.0 | 5656 |
| Staten Island | 98.58 | 75.0 | 371 |

# GEOGRAPHICAL ANALYSIS

**Heatmap - Price Analysis by Borough**

It makes comparisons of the Average Price and Median Price in five boroughs (neighborhood groups), Manhattan, Brooklyn, Queens, Bronx, and Staten Island.

- Visualizing the data in the heatmap with the use of color intensity (the darker the red color, the higher the price):
- The darkest red is Manhattan with the highest average price of (178.94) and the median price of (149.00).
- Brooklyn ranks number two in the average price and median price (117.86 and 90.00, respectively).
- Bronx averagely the lowest price of (85.36) and is tied with Queens and Staten Island in the median price (65.00/75.00).
- As it is well depicted in the chart, Manhattan has a significantly higher price than all other boroughs.




7. GEOGRAPHICAL INSIGHTS:

| neighbourhood_group | Average Price | Median Price | Listing Count |
|---|---|---|---|
| Bronx | 85.36 | 65.0 | 1089 |
| Brooklyn | 117.86 | 90.0 | 20041 |
| Manhattan | 178.94 | 149.0 | 21488 |
| Queens | 95.01 | 75.0 | 5656 |
| Staten Island | 98.58 | 75.0 | 371 |

# SUMMARY AND CONCLUSION

The given project has performed an in-depth Exploratory Data Analysis (EDA) of the NYC Airbnb 2019 data in order to reveal major tendencies and market insights.

**Key Findings:**

- Data Quality: Removed missing values, outliers, and whittled down 48895 listing to 48645 clean records.
- Price Insights: The market is skewed to the right and the median price is approximately in the range of $106. The prices of Manhattan listings are the highest (Avg: ~$179), whereas Bronx is the cheapest (Avg: ~$85).
- Listing Preferences: Entire homes/apartments (52) and private rooms (46) are the most popular with shared rooms being hard to find (2).
- Geographic Distribution: The enormous majority of listings are located in Manhattan and Brooklyn.
- Correlations: There were weak correlations among the variables such as price and availability among other variables such as reviews.
- Clustering: The Elbow Method identified 4 best customer/listing clusters to use the targeted strategy.
- Conclusion: The analysis provides practical recommendations to hosts, travellers, and policymakers and this includes strategies on pricing, demand patterns, and geographical patterns of Airbnb market in NYC.

# REFERENCES

[1] J. W. van Lith and J. Vanschoren, “From Strings to Data Science: A Practical Framework for Automated String Handling,” arXiv preprint arXiv:2111.01868, 2021.
[2] M. Kang and J. Tian, “Machine learning: Data pre-processing,” in Prognostics and Health Management of Electronics: Fundamentals, Machine Learning, and the Internet of Things, pp. 111–130, 2018.
[3] S. A. Macskassy and F. Provost, “Classification in networked data: A toolkit and a univariate case study,” Journal of Machine Learning Research, vol. 8, no. 5, 2007.
[4] J. Zhang, “Multivariate analysis and machine learning in cerebral palsy research,” Frontiers in Neurology, vol. 8, p. 715, 2017.
[5] A. Behura, “The cluster analysis and feature selection: Perspective of machine learning and image processing,” in Data Analytics in Bioinformatics: A Machine Learning Perspective, pp. 249–280, 2021.
[6] A. Lavallin and J. A. Downs, “Machine learning in geography – Past, present, and future,” Geography Compass, vol. 15, no. 5, p. e12563, 2021.

# THANK YOU