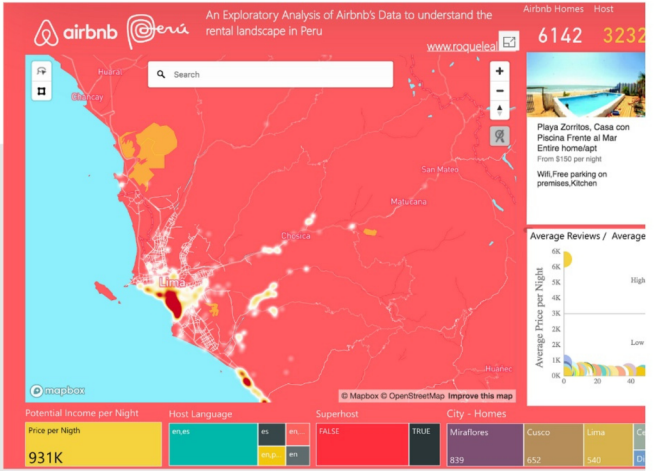
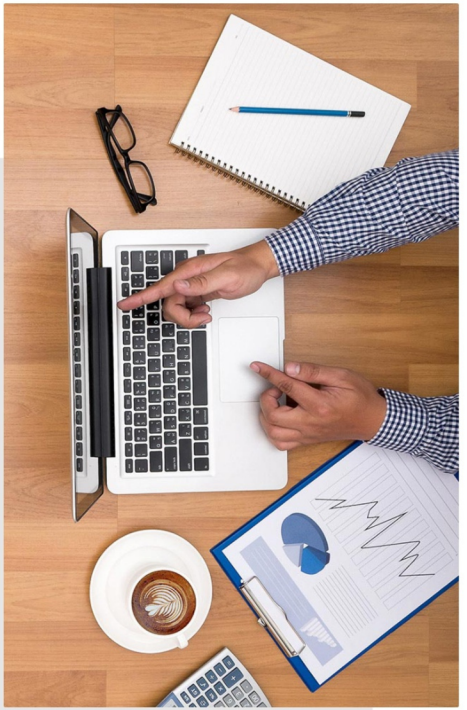
Understanding Airbnb Market Segments with K-Means Clustering



Microsoft Power BI

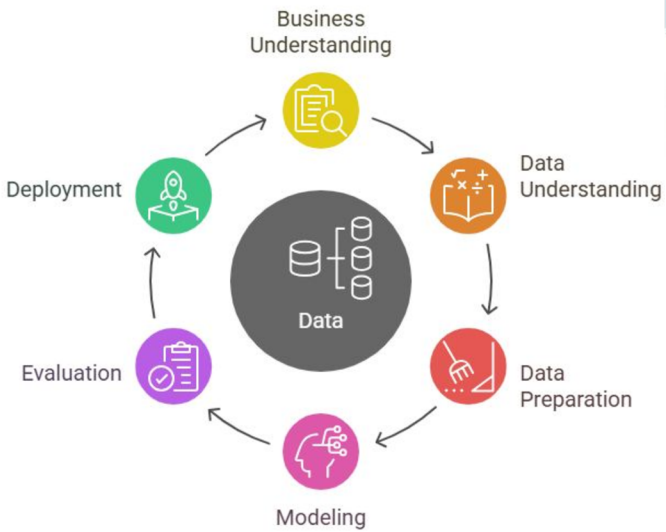
Shalihin
www.linkedin.com/in/shalihin2205

Table of Content



Methodology

CRISP-DM methodology was applied as a framework



Business Understanding

Background

Airbnb.com is a global platform for short-term rental accommodations that connects hosts with guests looking for a variety of accommodation options.

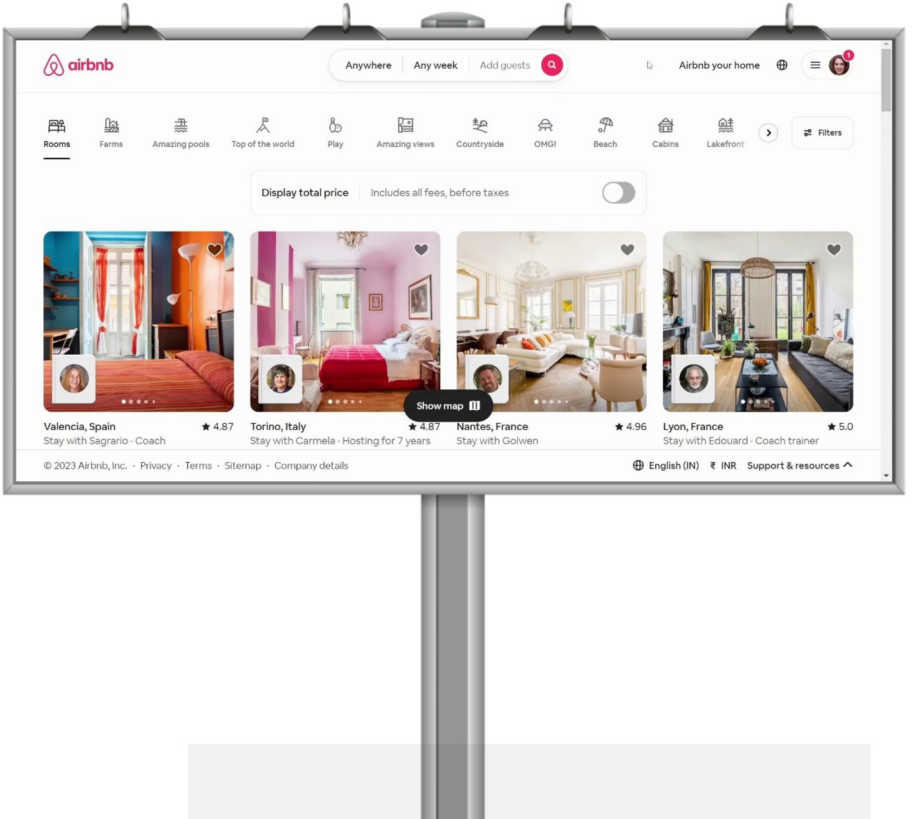
With increasing competition and the need for personalized service, it is important to understand the types or segments of listings available on their platform.

Objectives

Segment Airbnb listings in New York City using an unsupervised machine learning approach, with a focus on clustering and allowing users to access real-time insights.

Goals

Develop an end-to-end workflow that includes data preprocessing, K-Means clustering modeling, and visualizing results through an interactive web application using Streamlit.



Data Understanding

Data Description

Kolom	Tipe Data	Deskripsi
id	Integer	ID unik untuk setiap listing (penginapan).
name	String	Nama listing / penginapan.
host_id	Integer	ID unik dari host (pemilik listing).
host_name	String	Nama host / pemilik listing.
neighbourhood_group	String	Wilayah administratif besar (contoh: Manhattan, Brooklyn, dst.).
neighbourhood	String	Lingkungan atau wilayah kecil tempat listing berada.
latitude	Float	Koordinat lintang (latitude) dari lokasi listing.
longitude	Float	Koordinat bujur (longitude) dari lokasi listing.
room_type	String	Jenis kamar: `Entire home/apt`, `Private room`, atau `Shared room`.
price	Integer	Harga sewa listing per malam (dalam satuan mata uang: USD).
minimum_nights	Integer	Minimum jumlah malam untuk menyewa listing tersebut.
number_of_reviews	Integer	Total jumlah ulasan yang diberikan untuk listing ini.
last_review	String	Tanggal ulasan terakhir.
reviews_per_month	Float	Rata-rata ulasan per bulan untuk listing tersebut.
calculated_host_listings_count	Integer	Jumlah listing yang dimiliki host tersebut.
availability_365	Integer	Jumlah hari listing tersedia dalam 1 tahun (maksimal 365).

Data Understanding

48,895

Total Entries

Total number of entries in the dataset.

16

Features

- 10 numeric and 6 categorical Features
- Last_review should be changed to datetime

20.56%

Large Missing Value

last_review & reviews_per_month Features

0.3%

Low Missing Value

host_name & name features

Zero

Duplicates

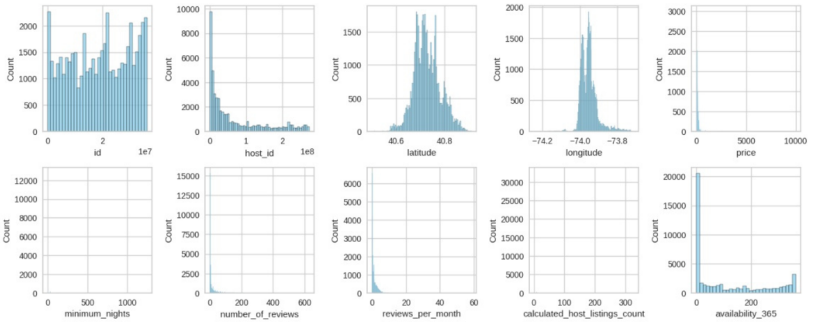
No Duplicates

Data Understanding

Numerical Statistical Summary

- Majority of features are positively skewed distribution.
- The features price, number_of_reviews, reviews_per_month, availability_365 have a value of 0.
- Longtitude has a value of -74.24, because it is a point coordinate location.

	id	host_id	latitude	longitude	price	minimum_nights	number_of_reviews	reviews_per_month	calculated_host_listings_count	availability_365
count	48895.000000	48895.000000	48895.000000	48895.000000	48895.000000	48895.000000	48895.000000	38843.000000	48895.000000	48895.000000
mean	19017143.236180	67620010.646610	40.728949	-73.952170	152.720687	7.029962	23.274466	1.373221	7.143982	112.781327
std	10983108.385610	78610967.032667	0.054530	0.046157	240.154170	20.510550	44.550582	1.680442	32.952519	131.622289
min	2539.000000	2438.000000	40.499790	-74.244420	0.000000	1.000000	0.000000	0.010000	1.000000	0.000000
25%	9471945.000000	7822033.000000	40.690100	-73.983070	69.000000	1.000000	1.000000	0.190000	1.000000	0.000000
50%	19677284.000000	30793816.000000	40.723070	-73.955680	106.000000	3.000000	5.000000	0.720000	1.000000	45.000000
75%	29152178.500000	107434423.000000	40.763115	-73.936275	175.000000	5.000000	24.000000	2.020000	2.000000	227.000000
max	36487245.000000	274321313.000000	40.913060	-73.712990	10000.000000	1250.000000	629.000000	58.500000	327.000000	365.000000

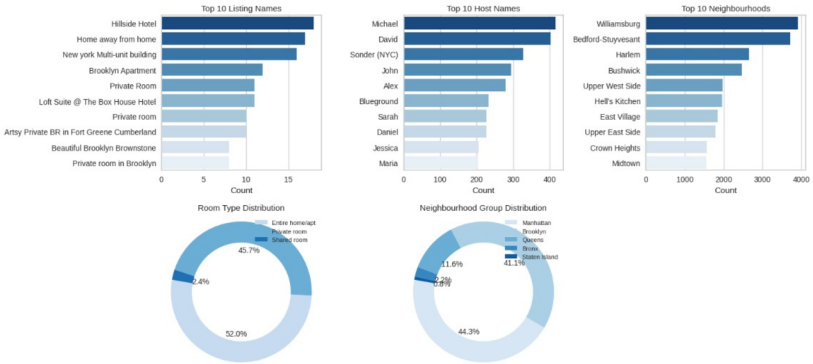


Data Understanding

Categorical Distribution Summary

- Categorical column shows information about unique, highest, and frequency values of categories in each feature.
- Categorical column data distribution imbalanced

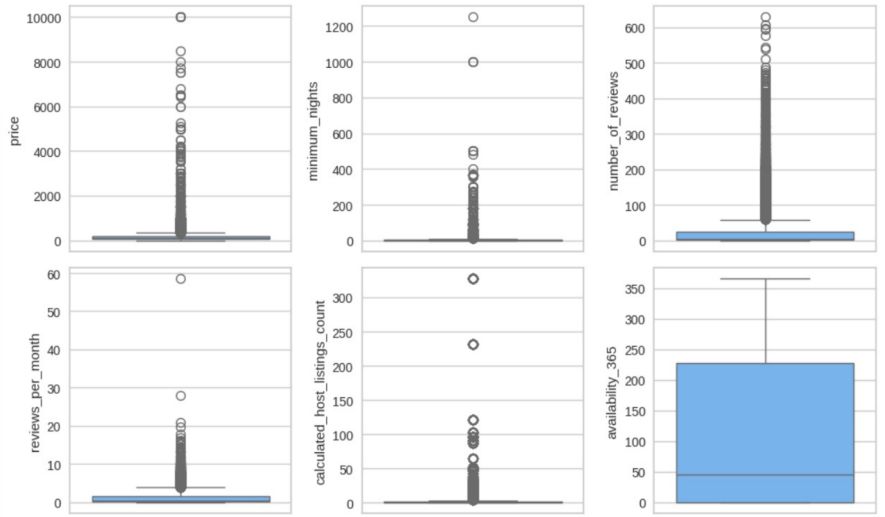
	count	unique	top	freq
name	48879	47905	Hillside Hotel	18
host_name	48874	11452	Michael	417
neighbourhood_group	48895	5	Manhattan	21661
neighbourhood	48895	221	Williamsburg	3920
room_type	48895	3	Entire home/apt	25409



Data Understanding

Outlier Summary Using Boxplot

- Most features are concentrated in the lower range and have extreme outliers.
- availability_365 shows a fairly even distribution.



Data Preparation

Handling Duplicate

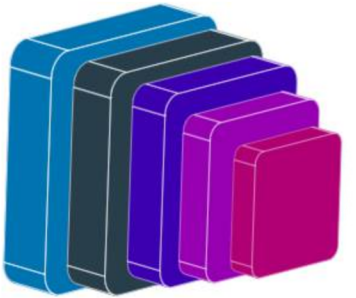
No Duplicates

Handling Missing Value

Remove missing values on name and host_name .
Drop last_review feature has 20% missing values.
reviews_per_month filled with a value of 0 = No Reviews

Handling Outlier

Capping to limit minimum and maximum values to outliers



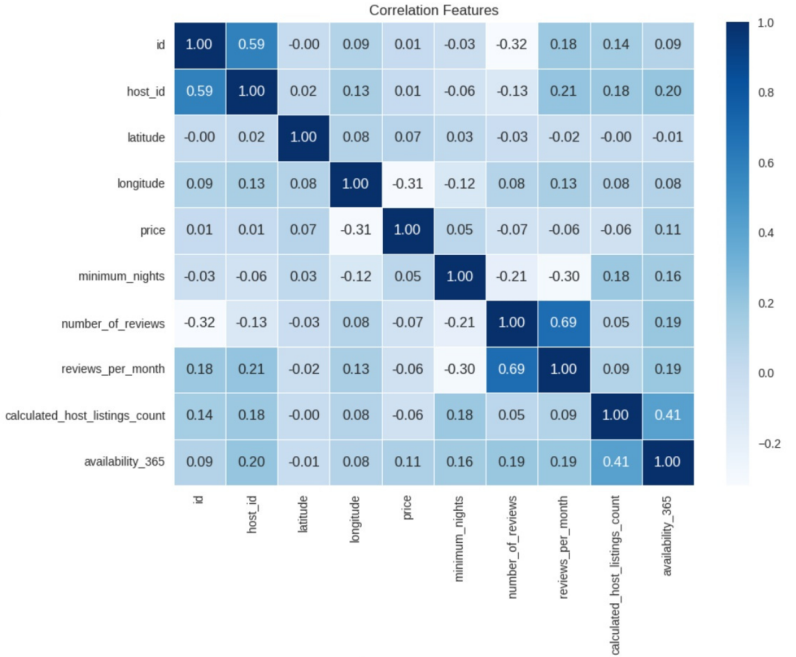
Feature Selection

Select **price, minimum_night, number_of_reviews, reviews_per_month, calculated_host_listings_count, availability_365**, as the primary feature

Standardization

Standardize data with Standart Scaler

Data Preparation



Moderate Correlation

0.69

number_of_review

0.69

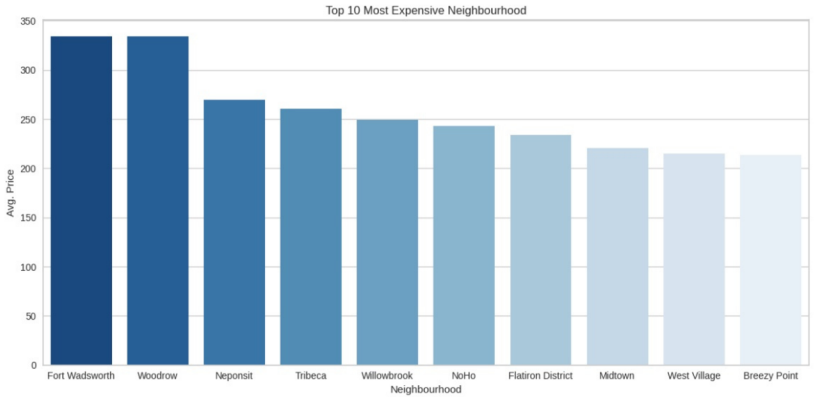
review_per_month

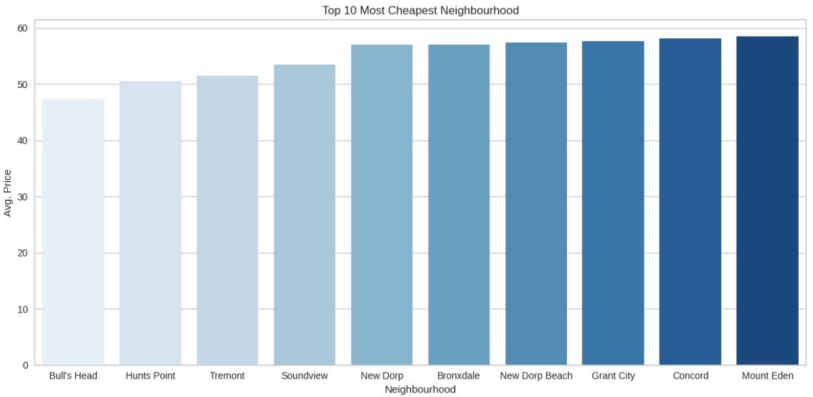
Dataset Airbnb

Exploratory Dataset Airbnb

Top 10 Expensive Neighbourhood VS Top 10 Cheapest Neighbourhood

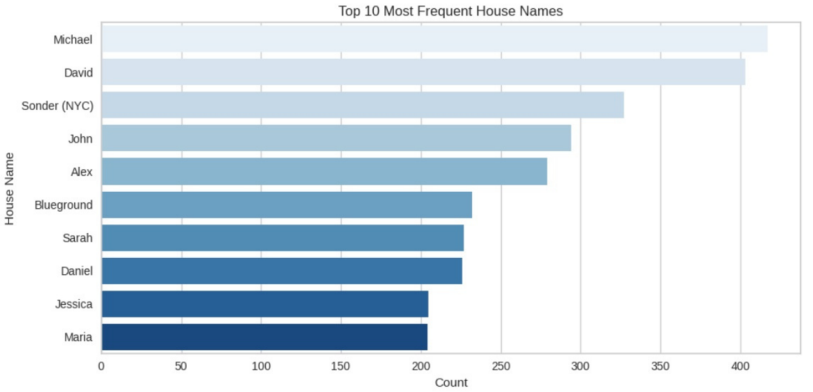
- ▲ **Fort Wadsworth, Woodrow, and Neponsit are the top 3 most expensive neighborhoods.**
- ▼ **Bull's Head, Hunts Point, and Tremont are the bottom 3 cheapest neighborhoods.**





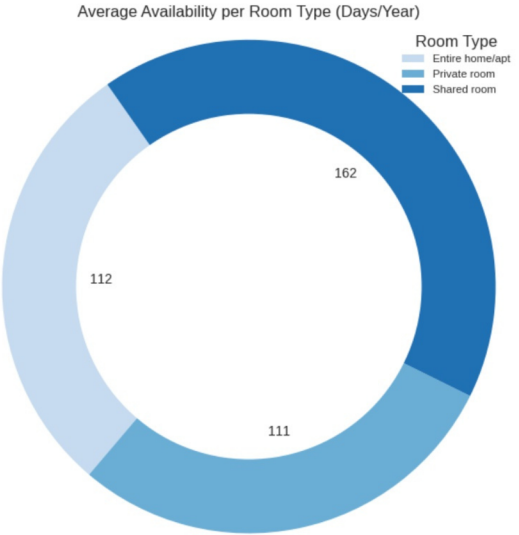
Exploratory Dataset Airbnb

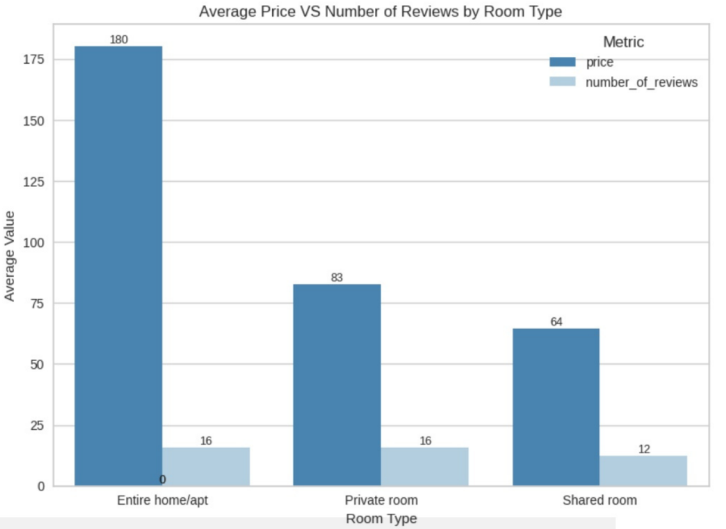
Michael with the most homeowners at 417 listings



Exploratory Dataset Airbnb

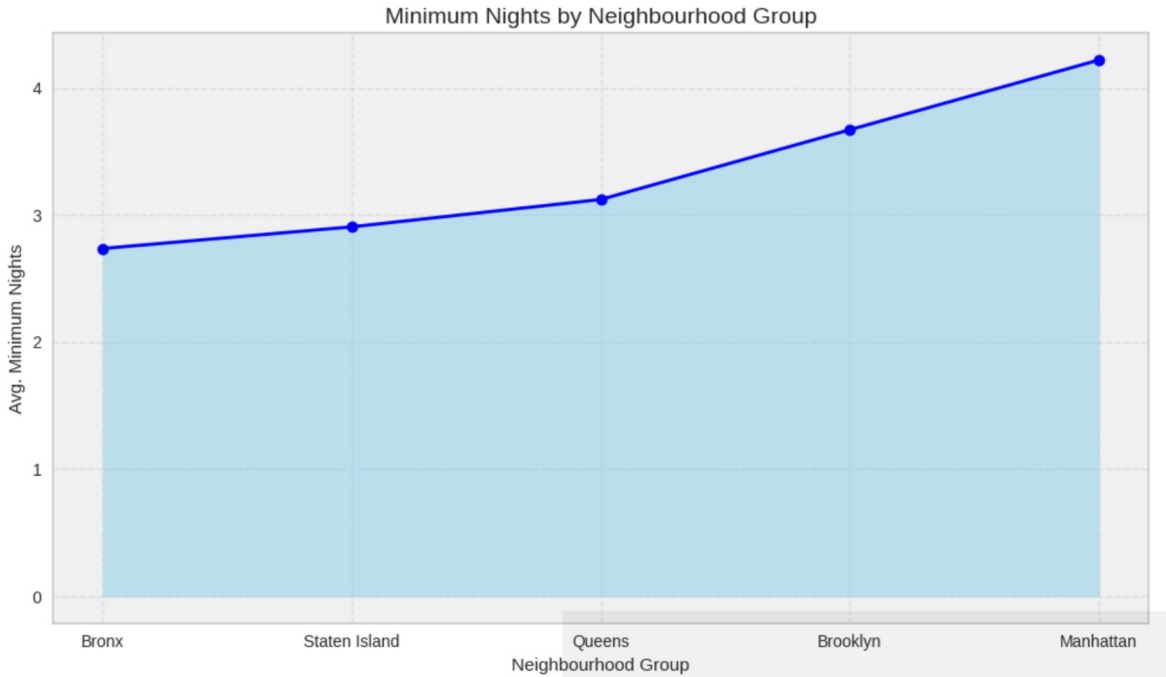
- Shared rooms have the highest average availability, at 162 days per year.
- Private rooms are the most frequently chosen by users and offer a good balance between price and reviews.





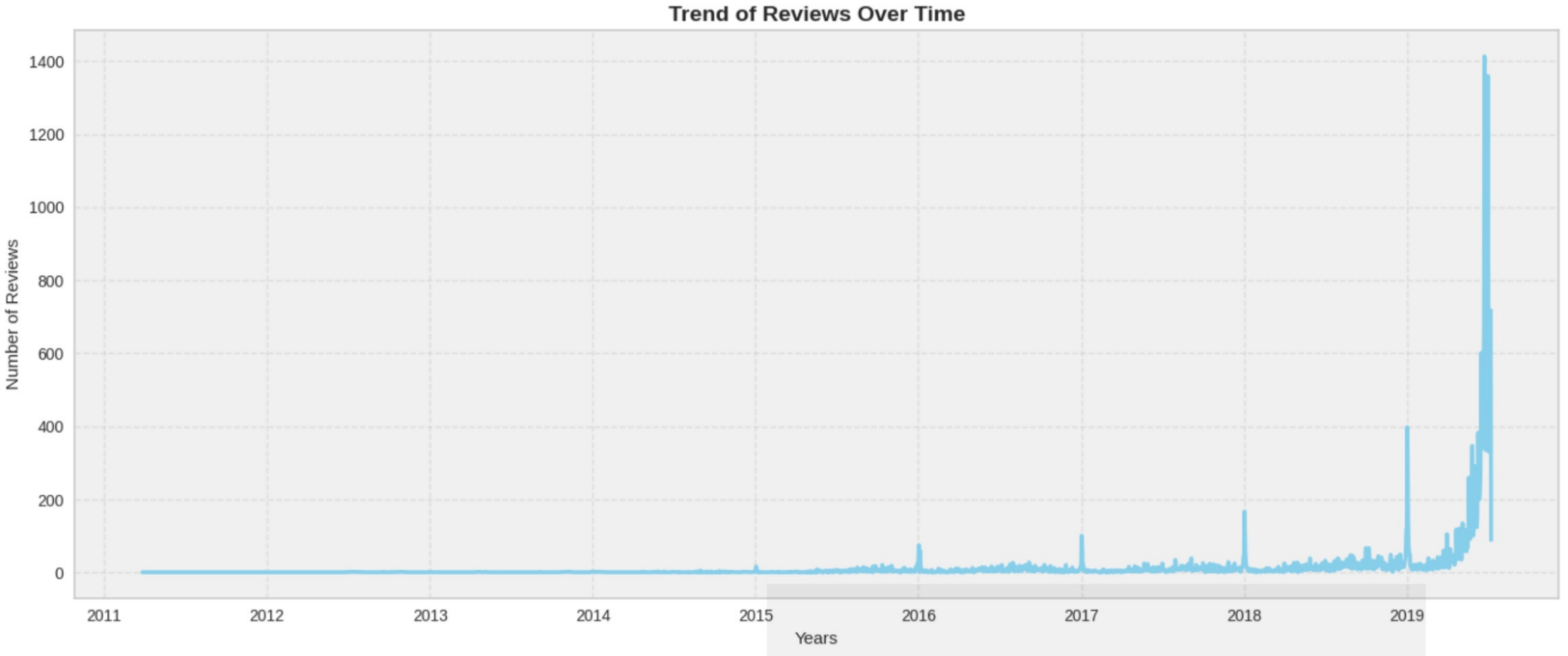
Exploratory Dataset Airbnb

Manhattan has the highest average minimum stay duration, which is 4 nights.



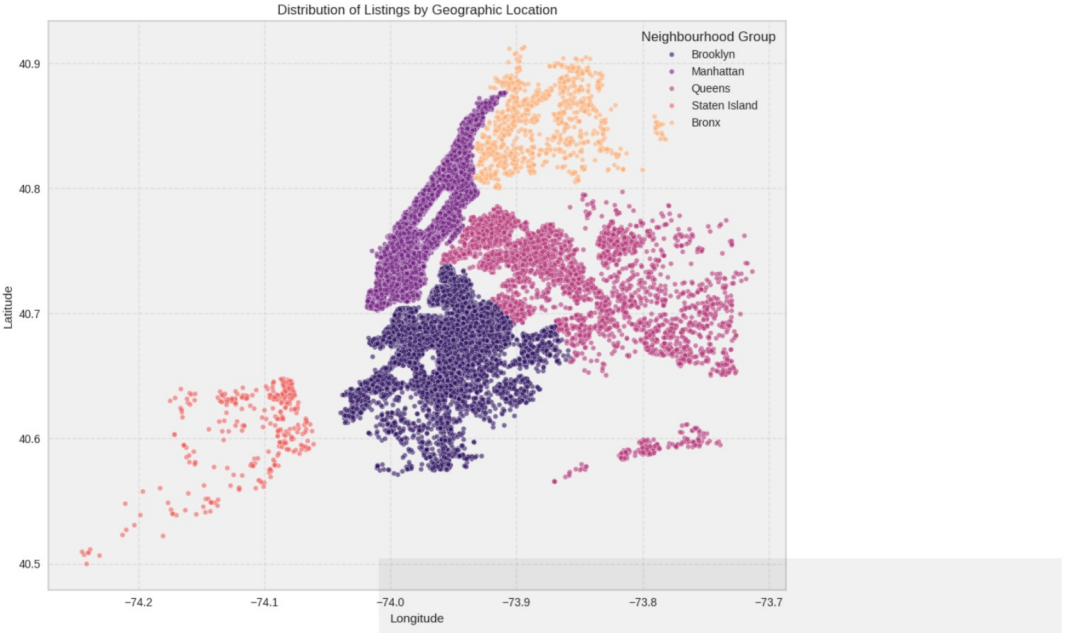
Exploratory Dataset Airbnb

Listings were last reviewed the most in 2019



Exploratory Dataset Airbnb

- The majority of listings are concentrated in Manhattan and Brooklyn, indicating these two boroughs as key hubs for Airbnb activity.
- Queens shows a fairly wide distribution of listings, though less dense than Manhattan and Brooklyn.
- Staten Island and The Bronx have fewer listings with a more sparse spread across the area.



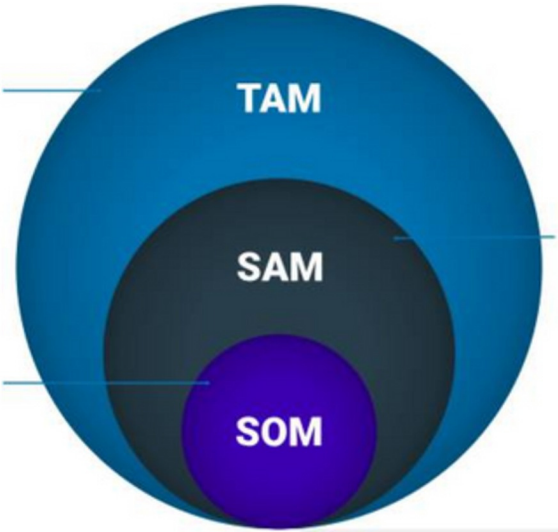
Modeling & Evaluation

Using K-Means for Data Clustering

3 & 4

Optimal Candidate Clusters

Number of clusters found by Elbow Method



0.32

Silhouette Score

Best cluster quality indicator is cluster 3

2 Component

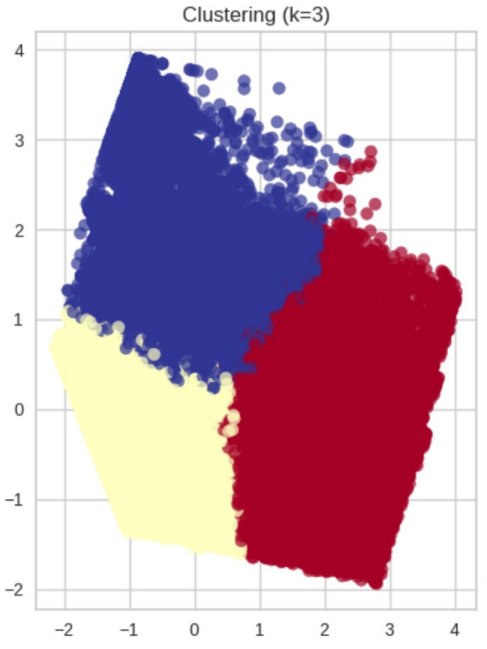
PCA Dimension Reduction

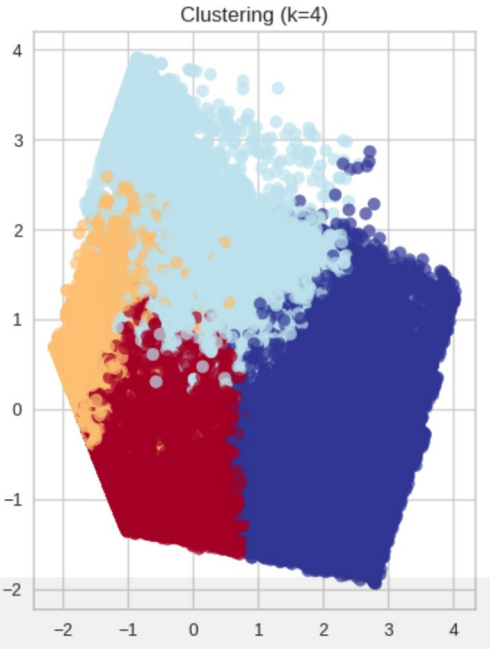
Visualize Clustering Results in 2-D space

Modeling & Evaluation

Cluster 3 VS Cluster 4

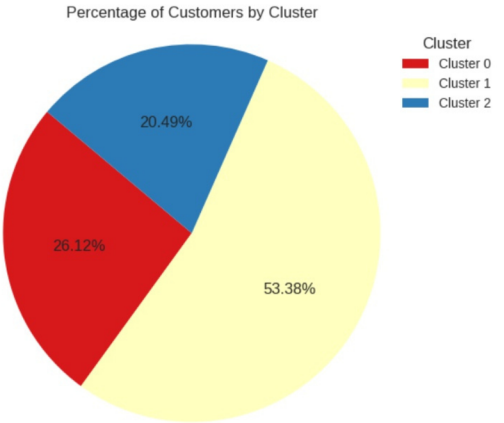
Cluster 4 = Overlapping

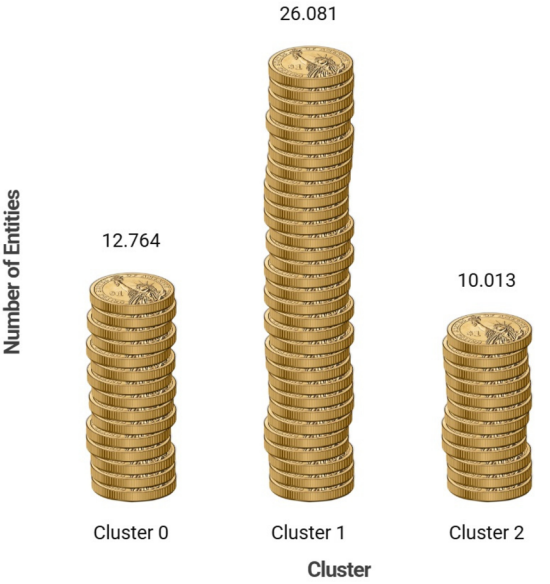




Interpretation & Recommendation

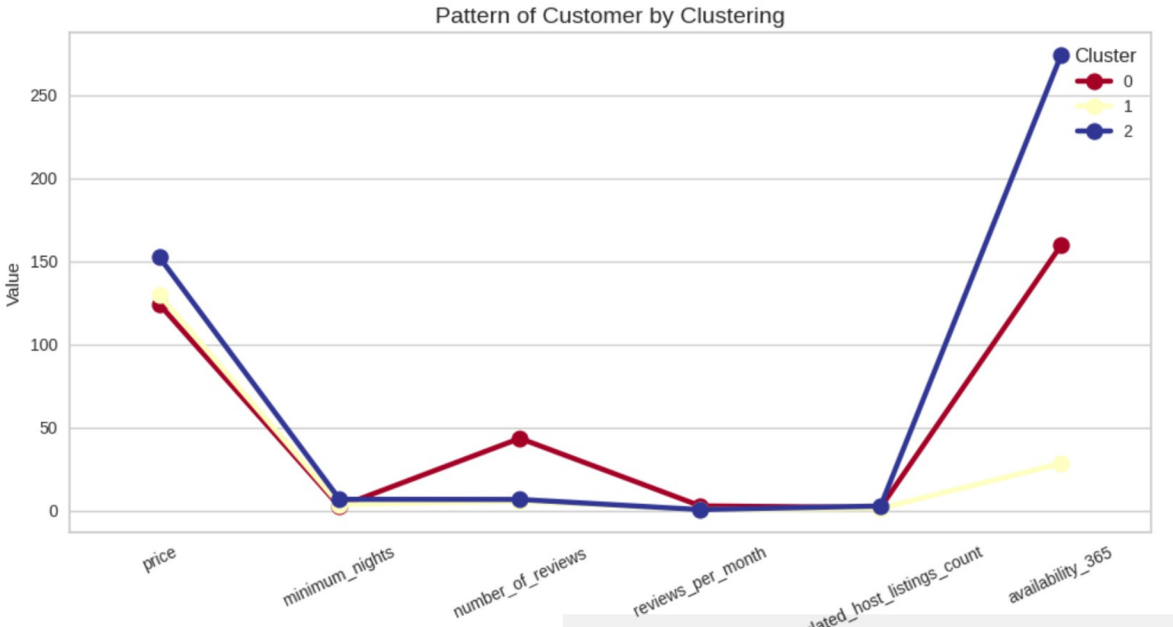
Cluster Sizes





Interpretation & Recommendation

Cluster Pattern based on Average of each feature per cluster



Interpretation & Recommendation

Interpretation

Cluster 🔴 is Populer

This cluster represents properties with affordable prices but high occupancy rates. Listings in this cluster perform very well in terms of number of reviews and rental frequency, indicating high appeal among budget travelers.

01

Cluster 🔵 is Premium

This cluster consists of exclusive property listings with higher rental rates and high availability. Targeting the upper-end market segment with a preference for comfort and quality of service.

03

Cluster 🟡 is Dormant

02

This cluster reflects listings with low rental activity. This is likely due to lack of visibility, less competitive pricing, or less strategic location.

Interpretation & Recommendation

Recommendation

01 **Cluster 🔴 - Populer**

Focus on optimizing pricing strategies and improving service quality to maintain high occupancy rates and strengthen customer loyalty.

02 **Cluster 🟡 - Dormant**

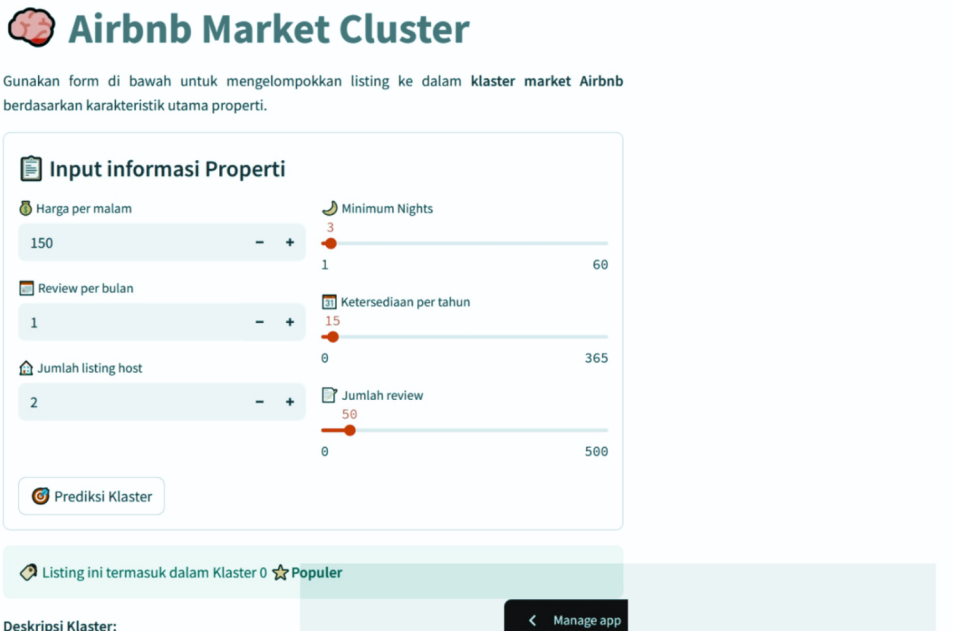
It is necessary to increase listing visibility through digital marketing strategies, adjust prices to be more competitive, and improve content quality (photos, descriptions, reviews) to attract guests.

03 **Cluster 🔵 - Premium**

Targeting exclusive and long-stay traveler customer segments. It is recommended to intensify personalized promotions and strengthen brand positioning to increase engagement and conversion.

Deployment Model

Link : https://airbnb-market-segments-with-k-means-clustering.streamlit.app/



Airbnb Listing Segmentation Project Conclusion

Airbnb listing segmentation successfully

The project successfully grouped Airbnb listings in New York City using the cluster method.

Data-driven recommendations prepared

Analysis resulted in strategic recommendations to improve listing performance and strengthen competitiveness in the market.

Real-time insights

Implemented a cluster-based machine learning model into an interactive web dashboard, allowing users to access real-time insights and make more informed business decisions.

Thanks

" Data bukan sekadar angka—ia adalah kompas yang menunjukkan ke mana arah kesuksesan harus dituju. Dengan memahami pola dari setiap kelompok pelanggan melalui clustering, kita tidak hanya memetakan perbedaan, tetapi membuka peluang untuk bertindak lebih tepat, lebih cepat, dan lebih personal".