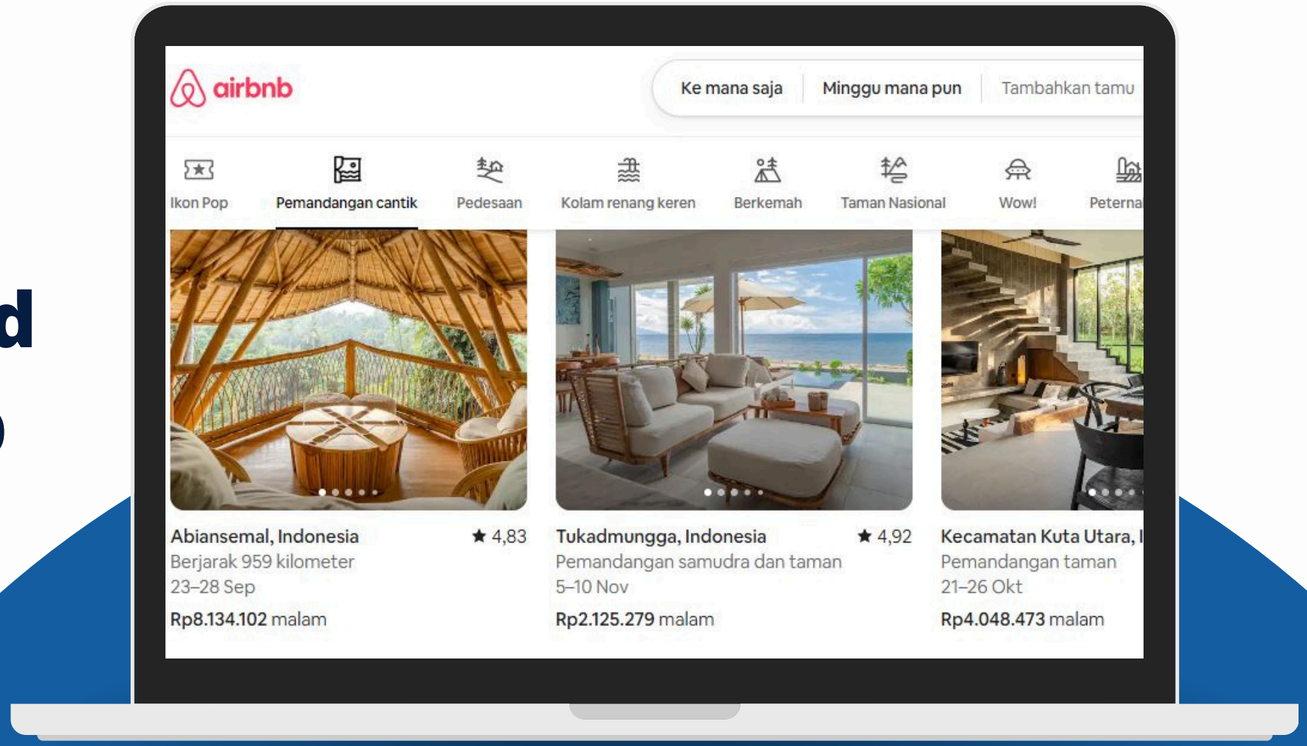
AirBnB Mini Project

Recommendation System Based on Weighted Rating, Content-Based Filtering, and Clustering Model for Airbnb

Jevon Tama Sianipar



Background

In 2024, global tourism has nearly returned to pre-pandemic levels, with international tourist arrivals reaching 97% of those figures and a 20% increase from early 2023 (UNWTO).

Airbnb saw significant growth, with **150 million users and a 55% increase in bookings from 2023 to 2024, totaling 300 million bookings**. This highlights the need for effective recommendation systems to enhance user experience.

Advanced algorithms and user data can help Airbnb offer personalized recommendations, improving customer satisfaction, engagement, and loyalty.

Objectives

Valuable Insight

Identify root problems and derive insights as a basis for business actions and the development of a recommendation system

Recommendation System

Find the best recommendation system to suggest the highest-rated listings and provide recommendations based on the user's booking history

Business Implementation

Develop actionable implementations based on the results of the recommendation system and data analysis

Data Highlight

This project uses a dataset that was directly scraped from the Airbnb website and subsequently uploaded to Kaggle. The dataset contains **23 columns and 12,805 rows.**

The data has **10 numerical features and 13 categorical features.** The feature details are as follows:

Column Name	Description
Unnamed	The index number of each row
ID	Unique identifier for each listing
name	Name of the Airbnb listing
rating	Average rating of the listing
reviews	Number of reviews received
host_name	Name of the host
host_id	Unique identifier for the host
address	Location of the listing (city, region, country)
features	Summary of features (number of guests, bedrooms, beds, bathrooms)
amenities	List of amenities provided
safety_rules	Safety rules of the listings
house_rules	House rules of the listings
price	Price per month
country	Country where the listing is located
bathrooms	Number of bathrooms
beds	Number of beds
guests	Number of guests the listing can accommodate
toilets	Number of toilets
bedrooms	Number of bedrooms
studios	Number of studio units
checkin	Check-in time
checkout	Check-out time
Img_links	Image of listing in a url format
Features	All listing features (total guests, total beds, total bedroom, bathroom,etc)

Data Preprocesing

01

Missing Value and duplicates

19% of the values in the 'checkout' column and 6.25% in the 'checkin' column are missing. These columns contain categorical values indicating check-in and check-out times. Missing values are filled with 'Not specified' to indicate the time is not specified. No duplicates were found in the data.

02

Drop features

Drop the 'Unnamed: 0', 'img_links', 'features', and 'address' columns. The 'features' column is redundant as its details are represented by 'guests', 'beds', 'toilets', and 'bathrooms'. 'Country' will represent the location. 'Unnamed: 0' contains no information, and 'img_links' are unnecessary for this project.

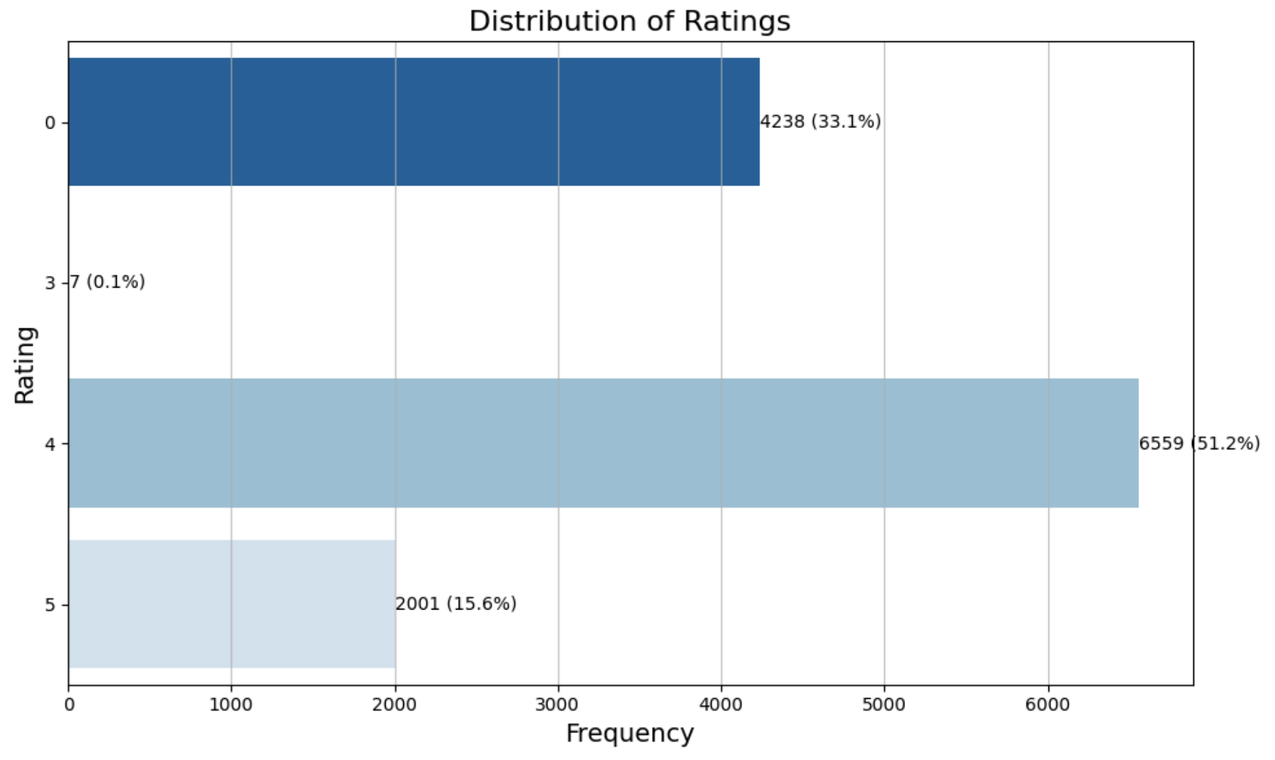
03

Create average daily price

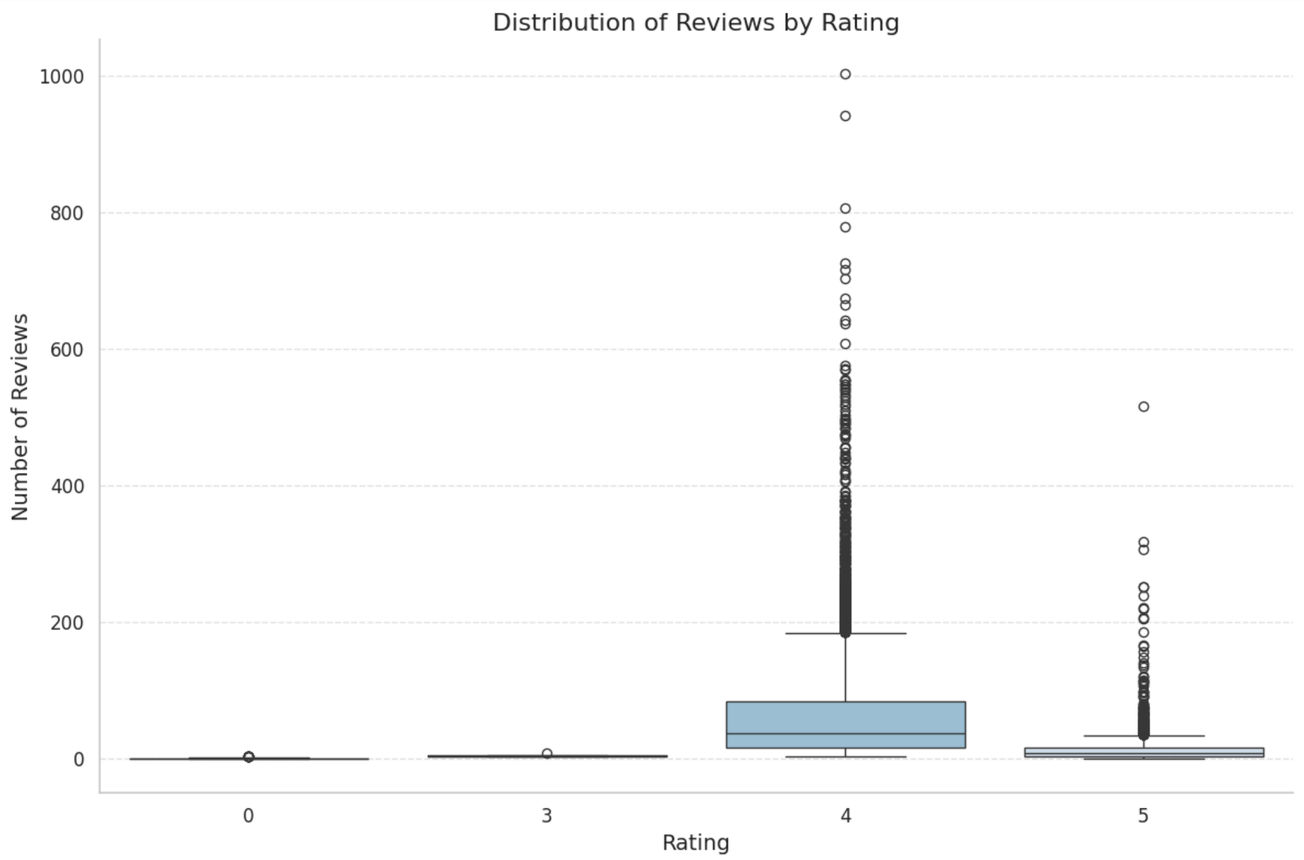
The listing price is currently shown by month, not by day. Convert the 'price' column into a 'price_fix' column by dividing the price by 30 to approximate the daily listing price per US dollar.

Exploratory Data Analysis

Distribution of Ratings and Review Counts by Rating

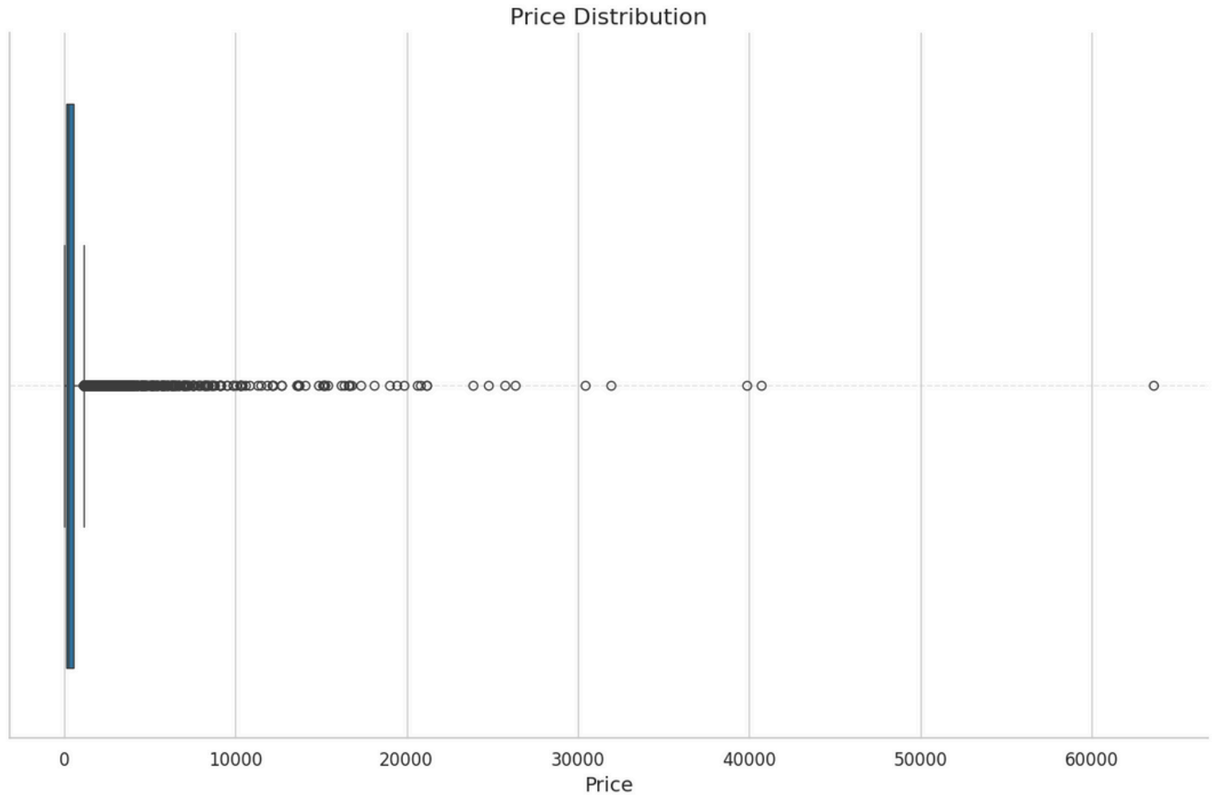


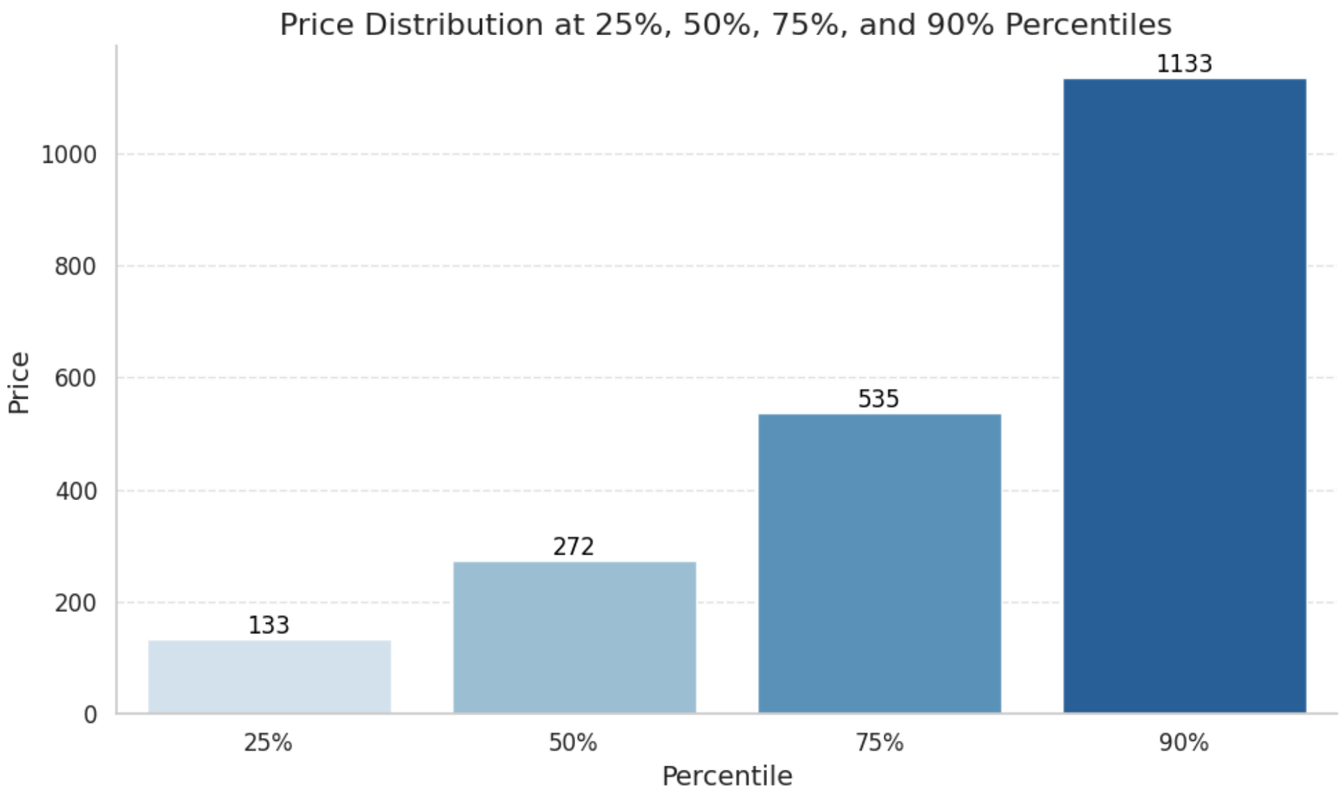
The rating distribution indicates that the majority of listings have a rating of 4, comprising 51.2% of the total. However, a significant portion of listings, 33.1%, remain unrated, which presents a potential issue.



Based on the analysis, the number of rated listings aligns with the volume of reviews, with listings rated 4 having the highest review count, reaching up to 1,000. However, unrated listings show almost no reviews, suggesting little to no bookings through the Airbnb platform for these listings. This issue should be addressed by the recommendation system, and these listings require targeted marketing campaigns.

Price Distribution

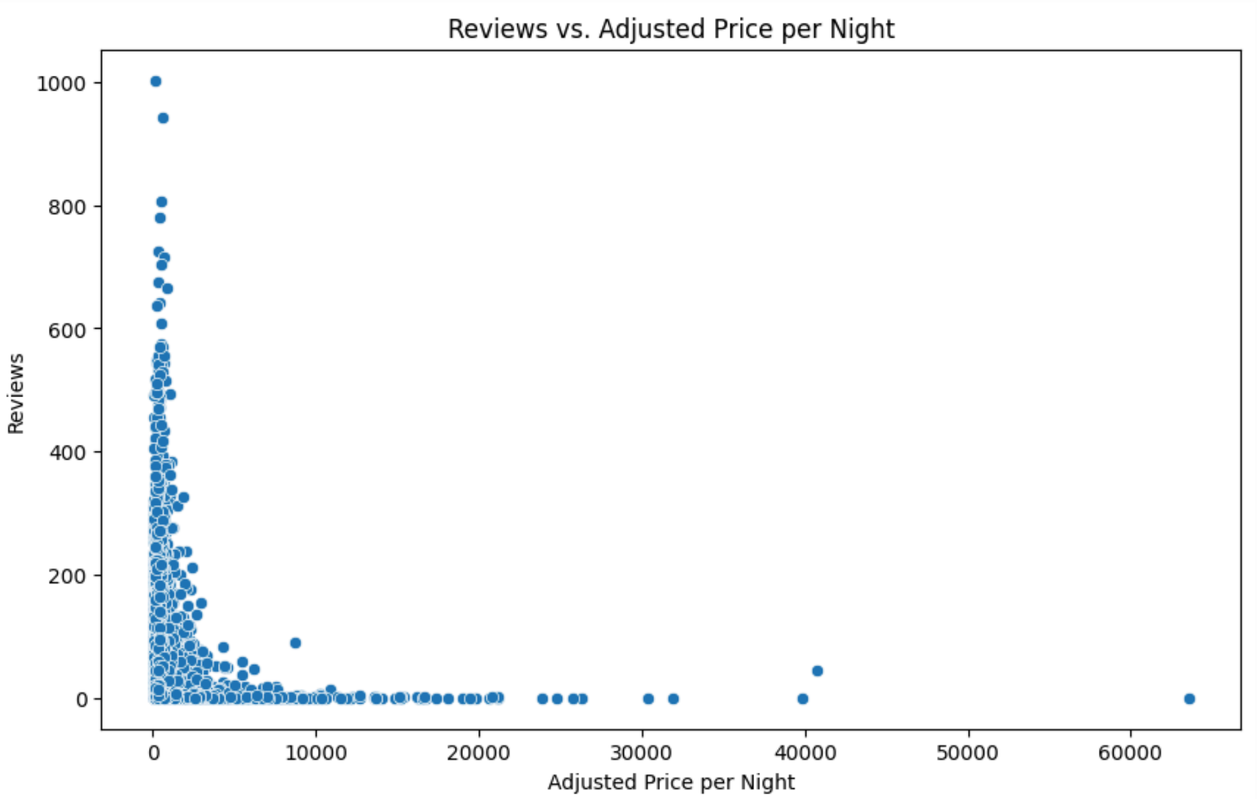


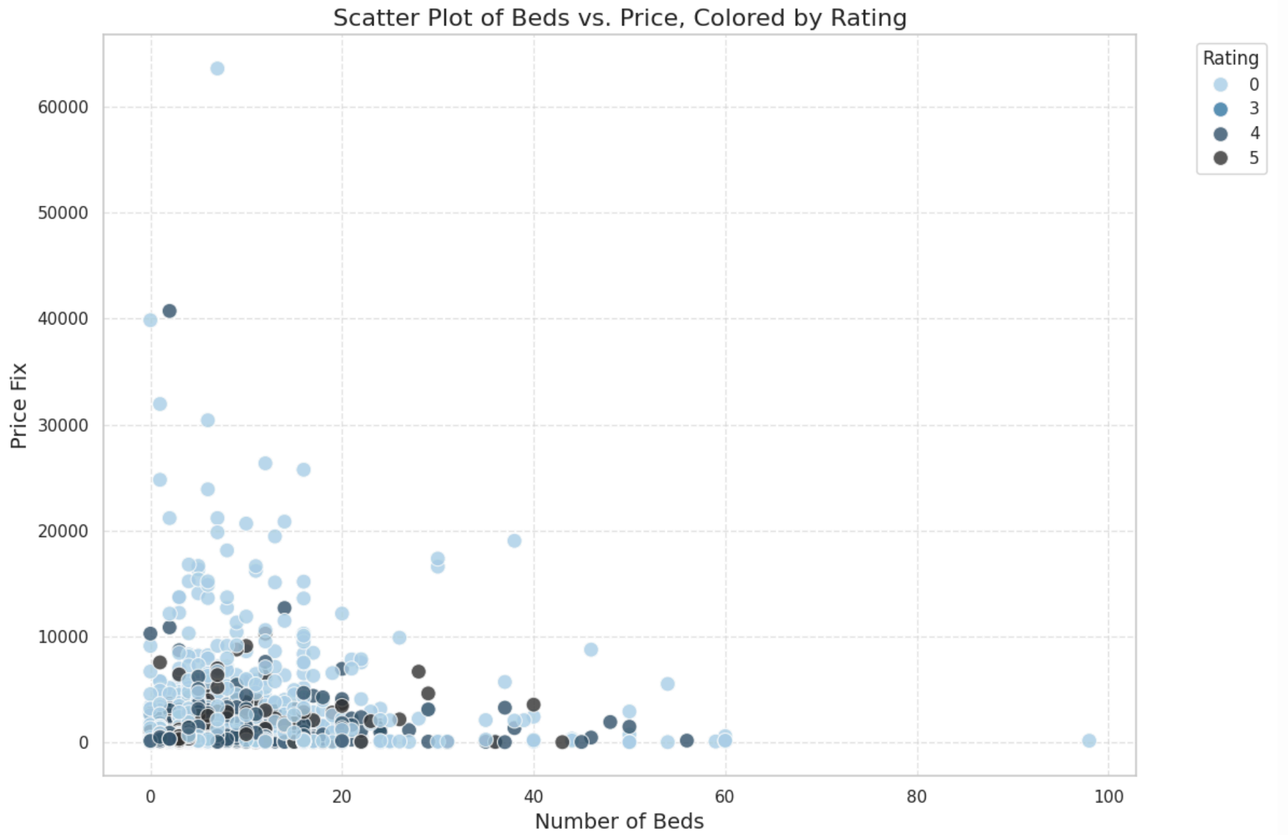


Airbnb listing prices range from as low as $30 to as high as $60,000, indicating a diverse price distribution. This diversity highlights the platform's potential to attract various market segments.

Upon detailed analysis, it is observed that 10% of the listings are classified as premium, with prices exceeding 1000 USD. Conversely, 90% of the listings fall within the price range of 100 to 1133 USD.

Analysis on Impact of Listing Prices



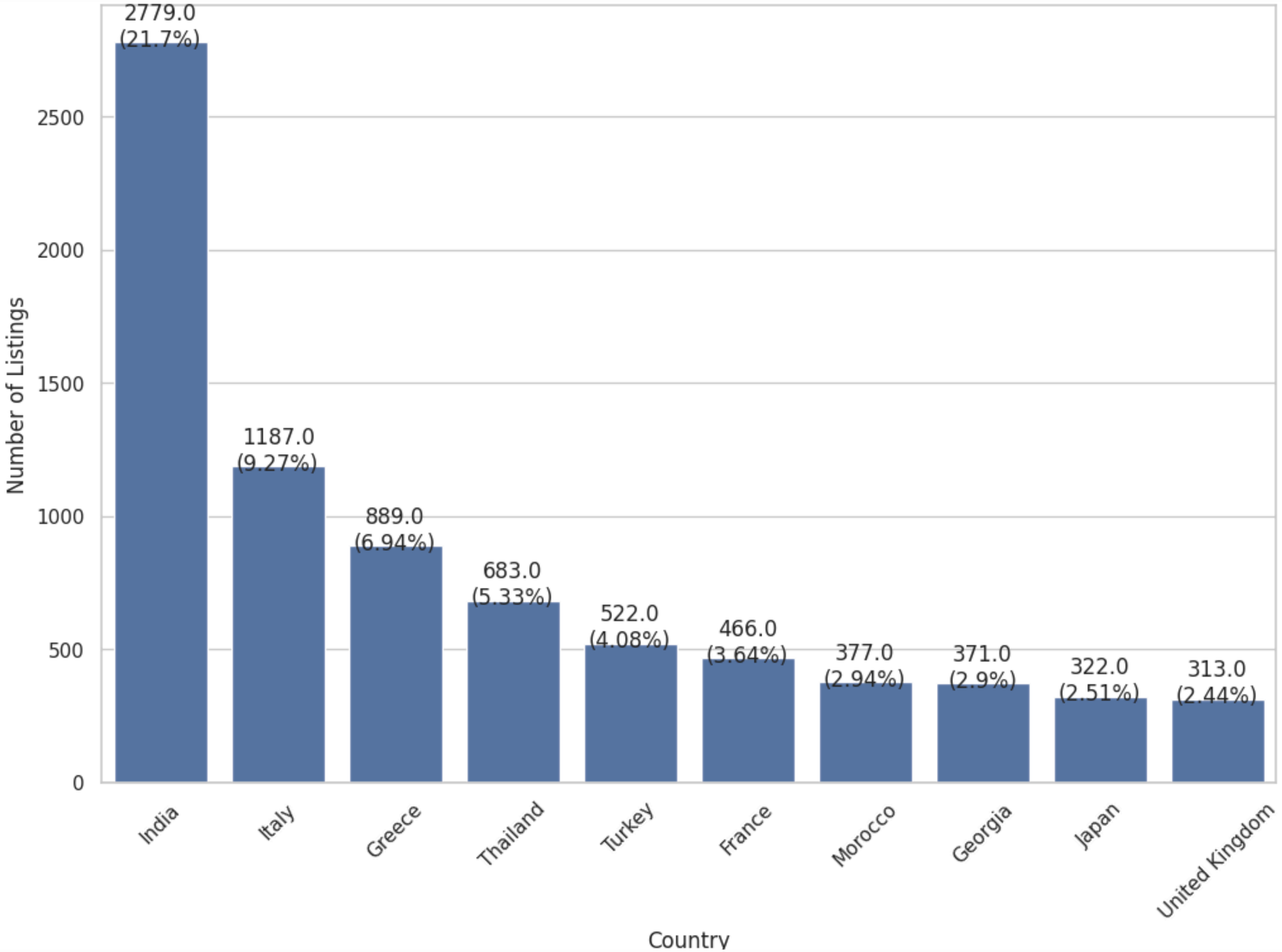


The results show no direct impact of price on the total number of reviews. However, most premium-priced listings (above 5,000 USD) still lack reviews, indicating that Airbnb users tend to avoid booking premium-priced listings.

Price per night is not directly influenced by the number of beds, which can also indicate the area or capacity of the listing. High-priced listings often have fewer beds than lower-priced ones, possibly leading to their unrated status.

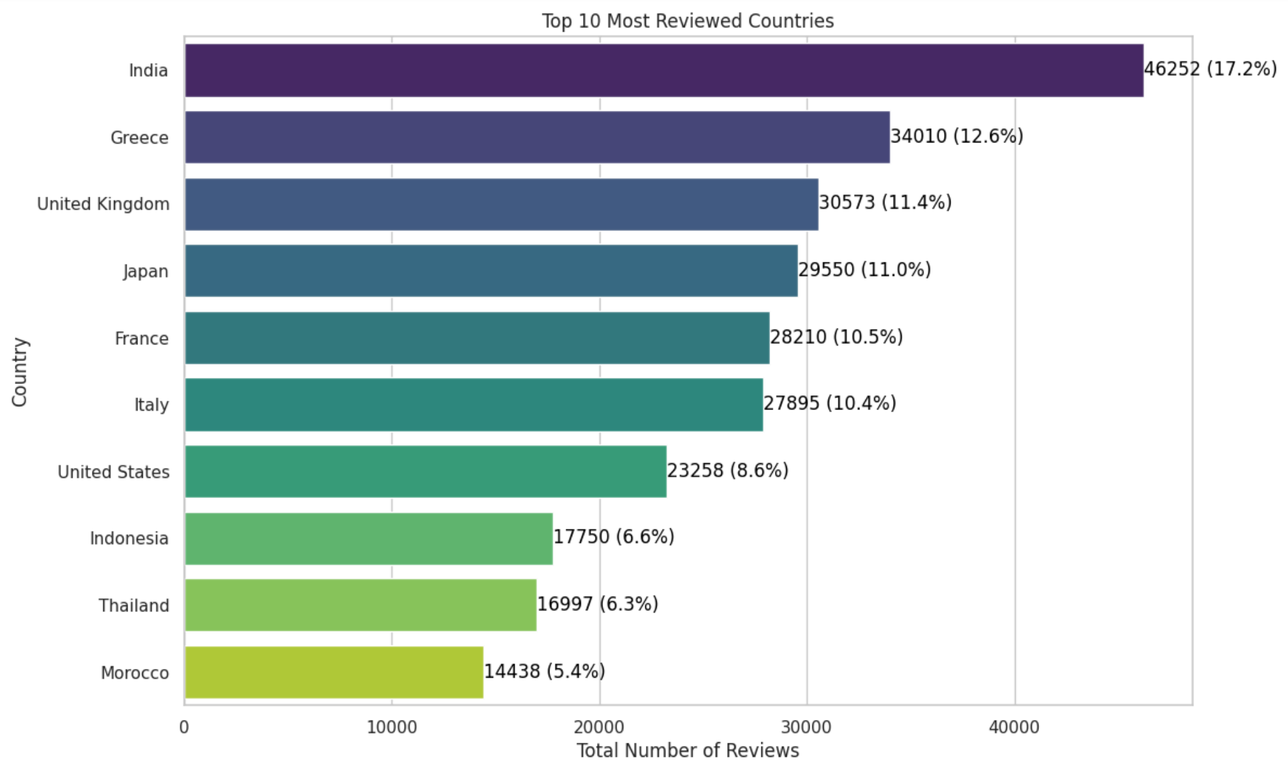
We recommend guiding hosts to price listings based on area and capacity and developing a model to determine listing prices using relevant features.

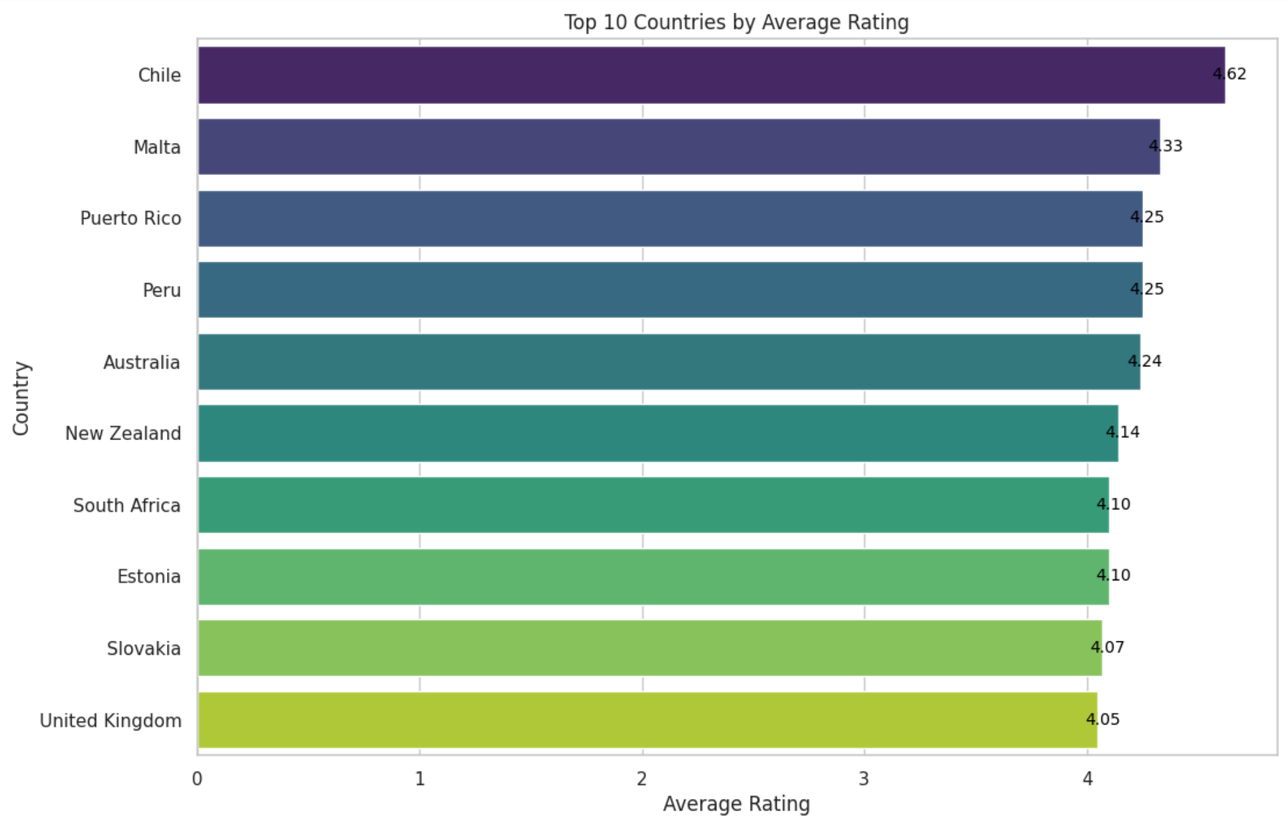
Top 10 countries with highest listing count



India is the country with the highest number of listing which reach 2779 listing , **which contribute to 21.7% of the total listing in AirBnb Platform. followed by italy 9.27% and greec 6.94%.** it show that india is already a high prospect market for AirBnB while other countries in the top 10 listing can be a potential high prospect market.

Top 10 countries with highest listing Reviews and Average Rating

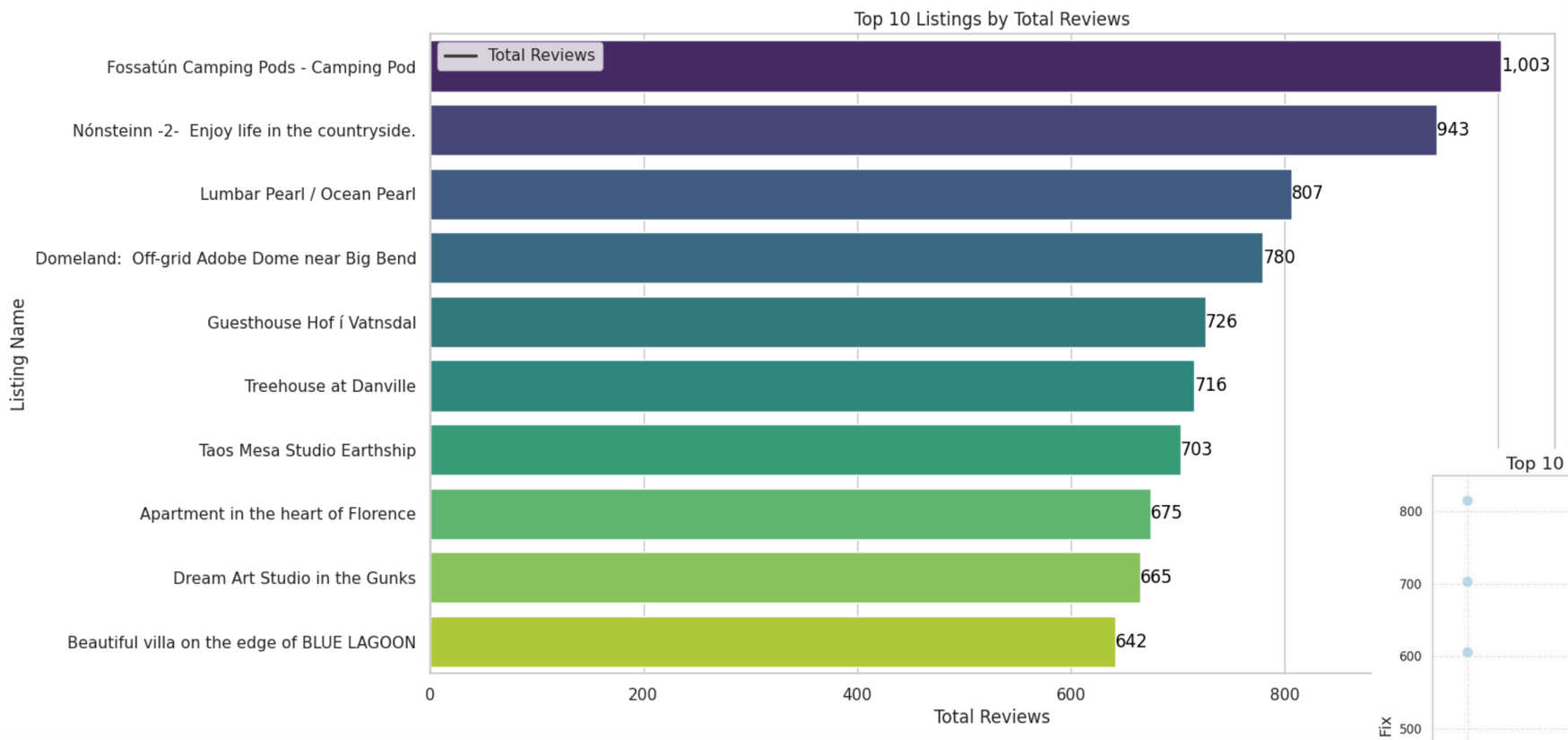


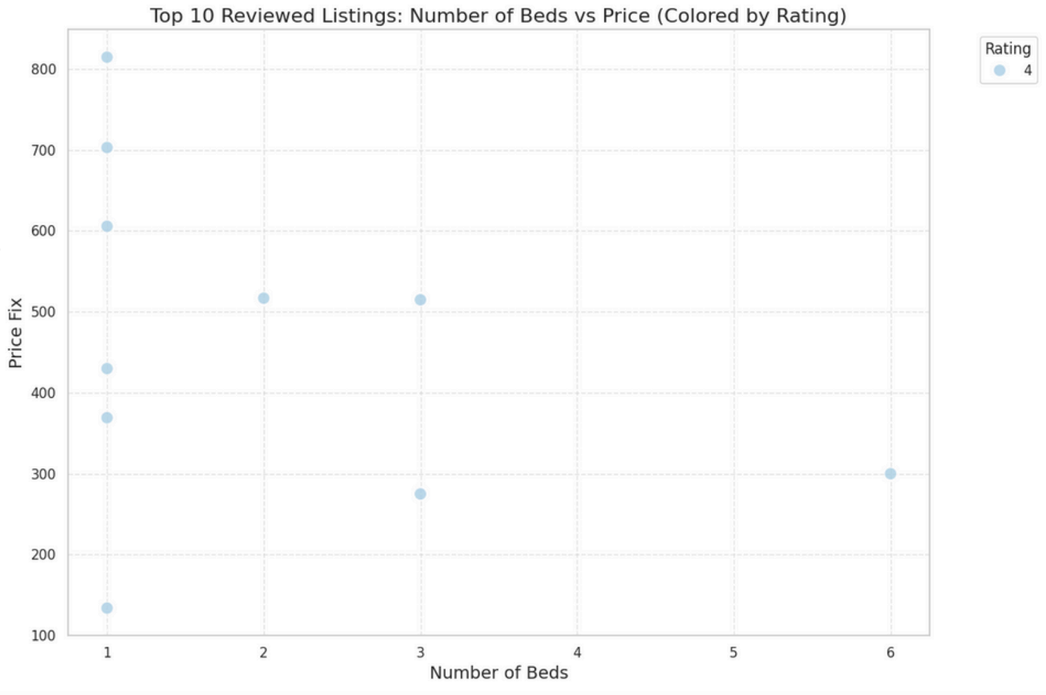


The total number of reviews aligns with the total number of listings for several countries. This indicates that the volume of listings in these countries correlates with user activity on Airbnb, as evidenced by the total number of reviews

Conversely, the average rating does not align with the total number of reviews. The top 10 countries with the highest average ratings differ from those with the highest total reviews. This suggests a need to identify features that influence high ratings in these countries and apply these insights to enhance customer experience in countries with high review volumes.

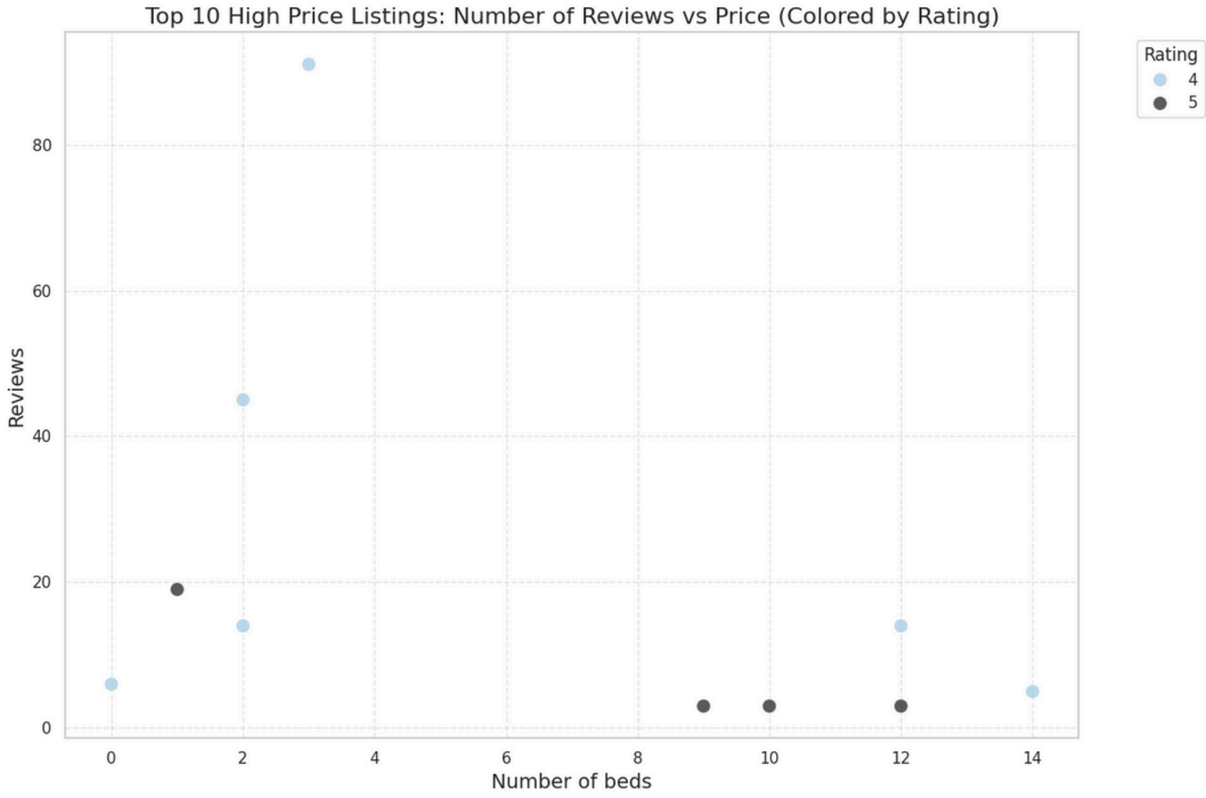
Top 10 Reviewed Listing Analysis





Analysis of the top 10 most-reviewed listings indicates that Airbnb users generally prefer smaller listings, with capacities ranging from 1 to 3 beds. Most top 10 listings have 1 bed, and only one listing has 6 beds. Additionally, these listings fall within a price range of 100 to 600 USD.

Top 10 High Price Listing Analysis



Analysis of the top 10 premium listings shows that higher prices correlate with larger areas and greater capacity, with listings featuring up to 14 beds.

Reviews range from 10 to 90, with ratings between 4 and 5, indicating a segmented market for premium listings. Despite previous analysis suggesting that unrated listings are often high-priced, these features can serve as a guideline for enhancing premium user experiences.

Recommendation System

(Before Data Engineering)

Weighted Rating Recommendation System

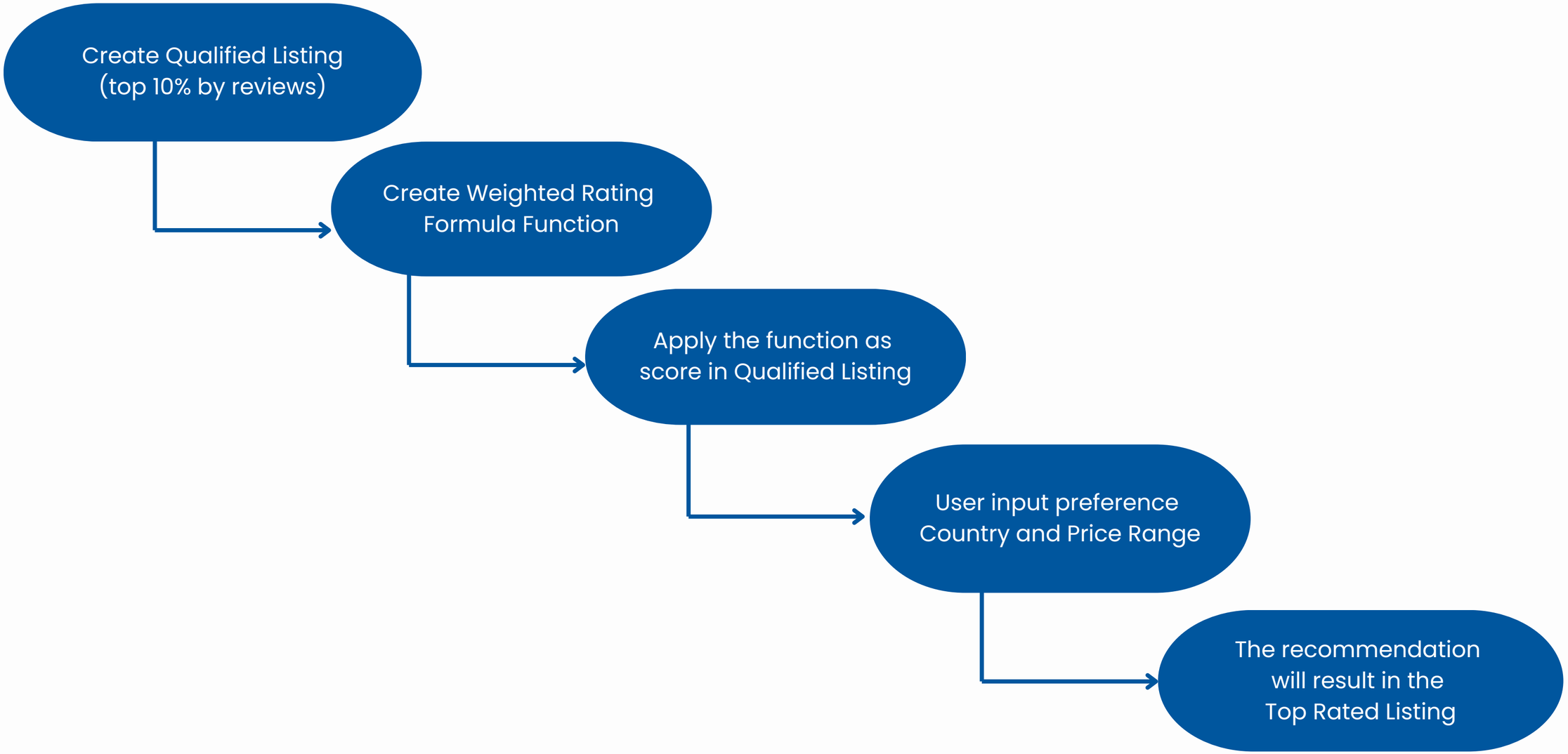
Goal: To recommend top-rated listings based on the user's preferred price and country.

$$W = \frac{Rv + Cm}{v + m}$$

C = mean rating across all listings
m = minimum number of reviews required to be listed
v = number of review
r = rating

The user will choose their preferred location and price range on the platform, and the platform will recommend high-rated listings determined by a score based on a weighted rating system.

Weighted Rating Recommendation System



Recommendation Result

	name	reviews	rating	price_fix
2,130	Beautiful villa on the edge of BLUE LAGOON	642	4	429.8
1,427	HIDEOUT LIGHTROOM - Eco Bamboo Home	308	4	358.93
2,208	Brand New Renovated 2 BR Villa in Seminyak Center	250	4	385.67
2,191	Villa Tulla in Central Canggu	240	4	340.5
2,223	Villa Quincy in Canggu	219	4	340.5
2,210	1BR private villa with private exotic pool	216	4	446.07
12,717	NEW romantic experience in canggu	183	4	386.07
2,268	BEAUTIFUL HOME - NEW POOL & RICE FIELD VIEWS!	180	4	357.63

Using the recommendation system simulation with Streamlit, for example, the user inputs "Indonesia" as the country, with a minimum price of 300 and a maximum price of 500.

The recommendation system will suggest listings with high reviews and ratings in Indonesia.

For the full Streamlit documentation, you can check it here

Content Based Filtering Recommendation System

Goal: To recommend listings based on the user's history of amenities and country

The user will be recommended listings based on similar countries and amenities from their booking history. The similarity will be determined by cosine similarity.

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}},$$

Cosine similarity is a measure of similarity between two non-zero vectors in an inner product space that calculates the cosine of the angle between them.

Content Based Filtering Recommendation System

Country and amenities features are combined into a single text feature

Use CountVectorizer to create a count_matrix from the combined feature

The count_matrix will be used with cosine similarity to measure similarity.

Input the user's booking history into the recommendation function based on cosine similarity

The recommendation will result in the Listing with similar amenities and country

Recommendation Result

User Booking History

amenities	count
wifi,freeparkingonpremises,petsallowed,tvwiths...	1
wifi,pool,tv,airconditioning	1
bayview,oceanview,kitchen,wifi,dedicatedworksp...	1
seaview,beachaccess–beachfront,kitchen,wifi,fr...	1
gardenview,beachaccess–beachfront,kitchen,free...	1
wifi,dedicatedworkspace,washingmachine,aircond...	1

country	count
india	2
thailand	2
egypt	1
australia	1
guadeloupe	1
estonia	1
taiwan	1
greece	1

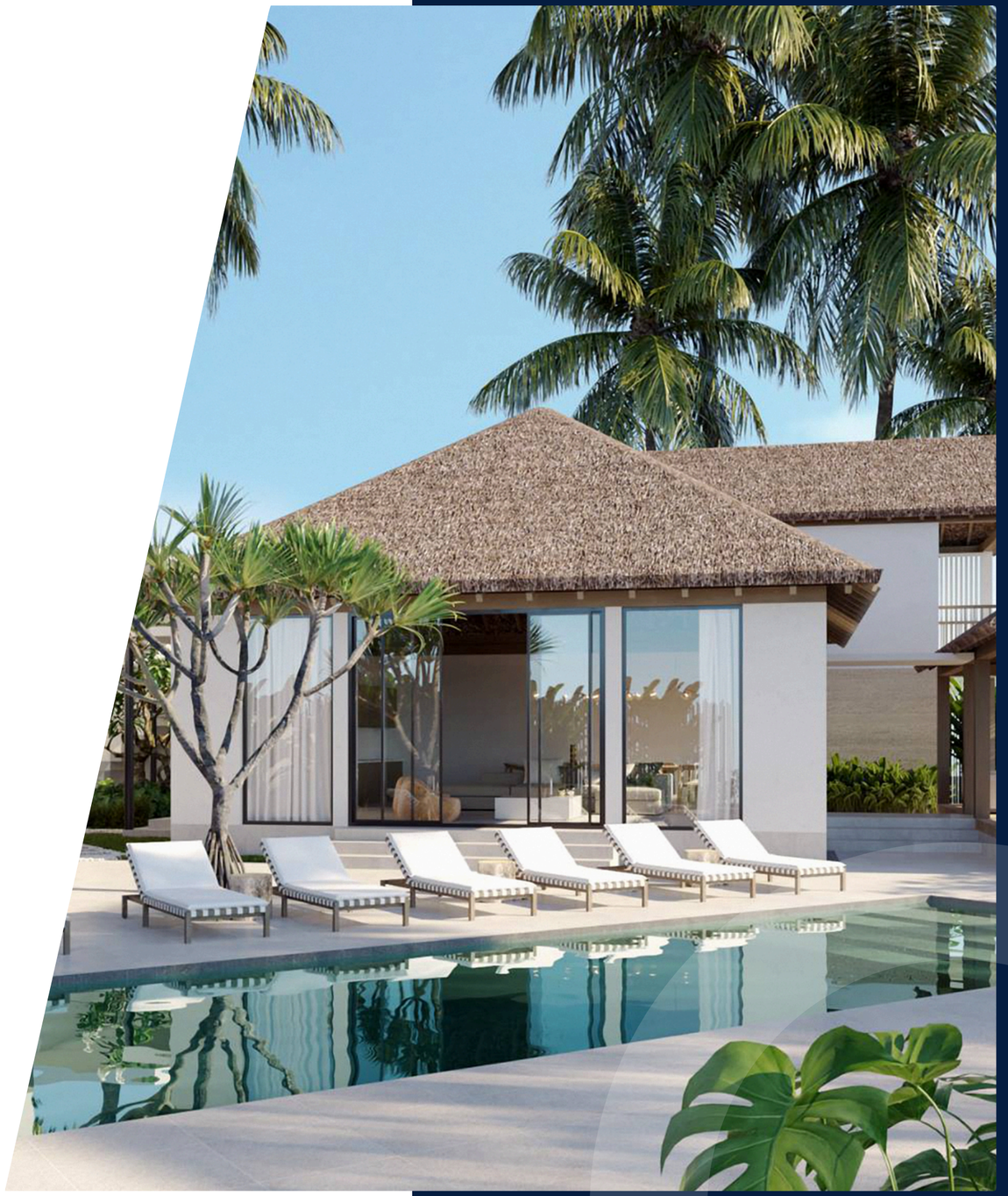
The user history of listing amenities and country

Recommendation Result

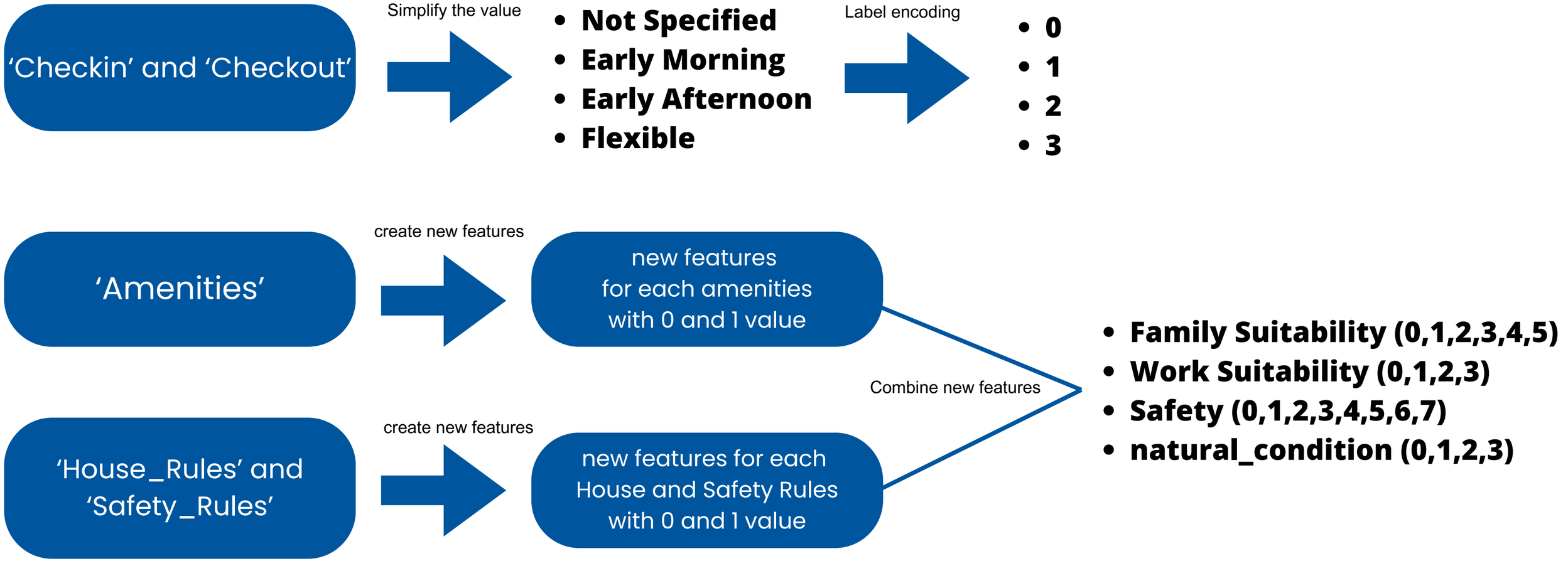
rating	amenities	country
5	mountainview,kitchen,wifi,petsallowed,hdtv,bat...	india
5	gardenview,valleyview,kitchen,wifi–35mbps,dedi...	india
4	kitchen,wifi,dedicatedworkspace,freeparkingonp...	india
5	mountainview,resortview,lakeaccess,wifi–47mbps...	india
4	waterfront,freeparkingonpremises,sharedpool,in...	india
5	freeparkingonpremises,cot,breakfast,unavailabl...	india
5	lakeaccess,freeon-streetparking,petsallowed,ga...	india
4	mountainview,kitchen,wifi,dedicatedworkspace,f...	india

Recommendation system suggests listings with similar amenities to the user's previous bookings. Since the user has booked in India twice, the recommendation prioritizes listings in India.

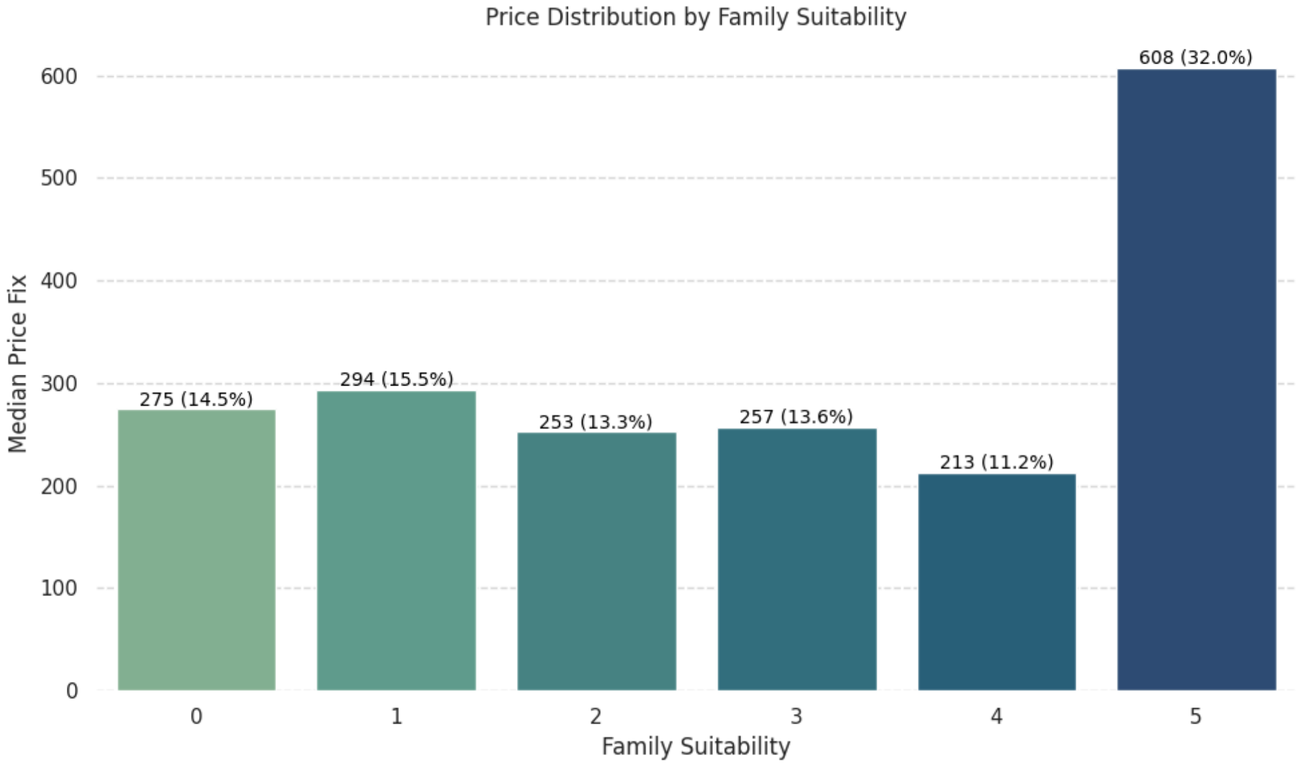
Feature Engineering



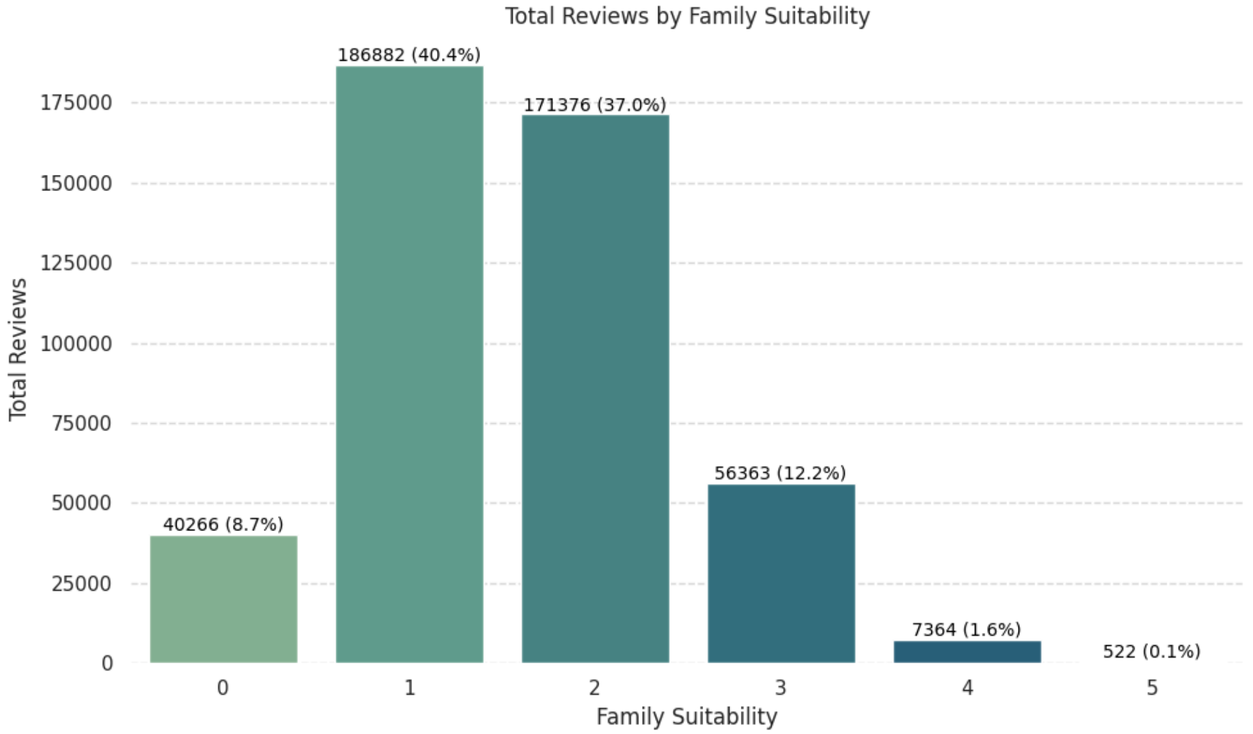
Feature Engineering for Clustering



Price and Reviews Distribution by Family Suitability

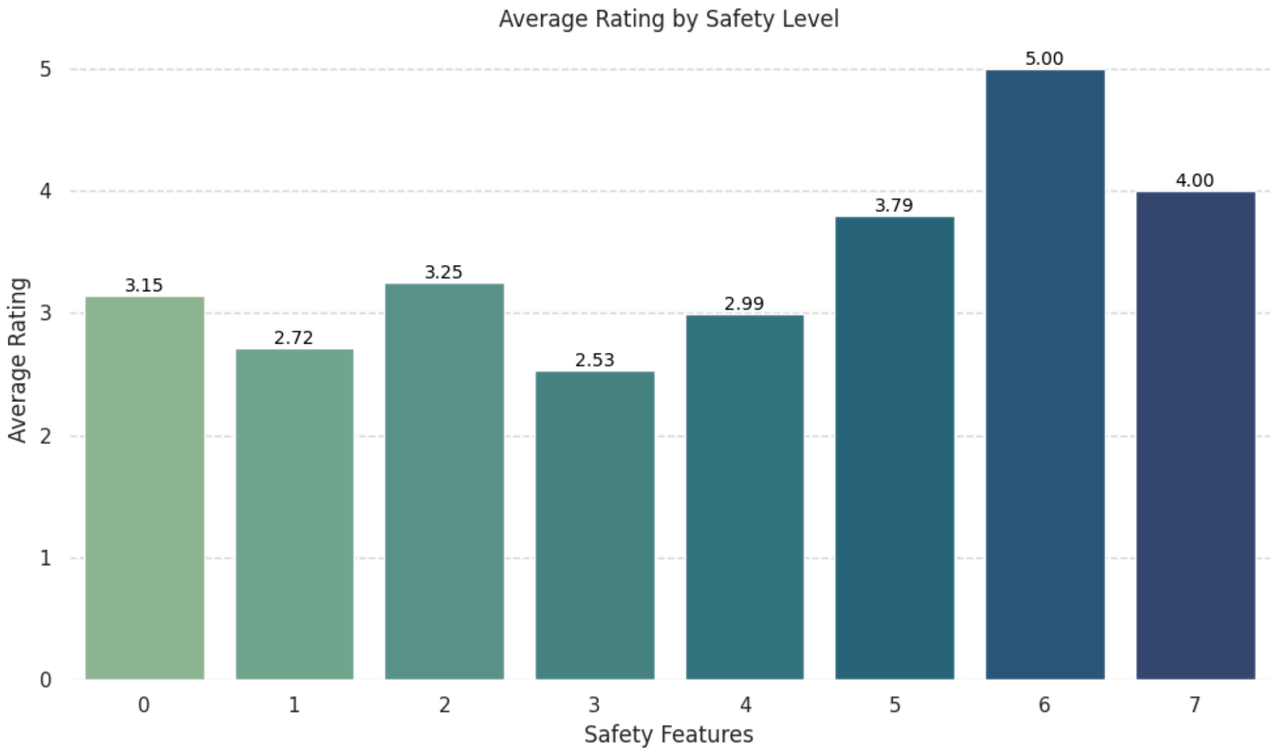


Higher amenities and safety features that contribute to the family suitability index lead to increased listing prices. Listings with a family suitability index of 5 have the highest average price, at $608, while those with an index of 0 to 4 average around $200.

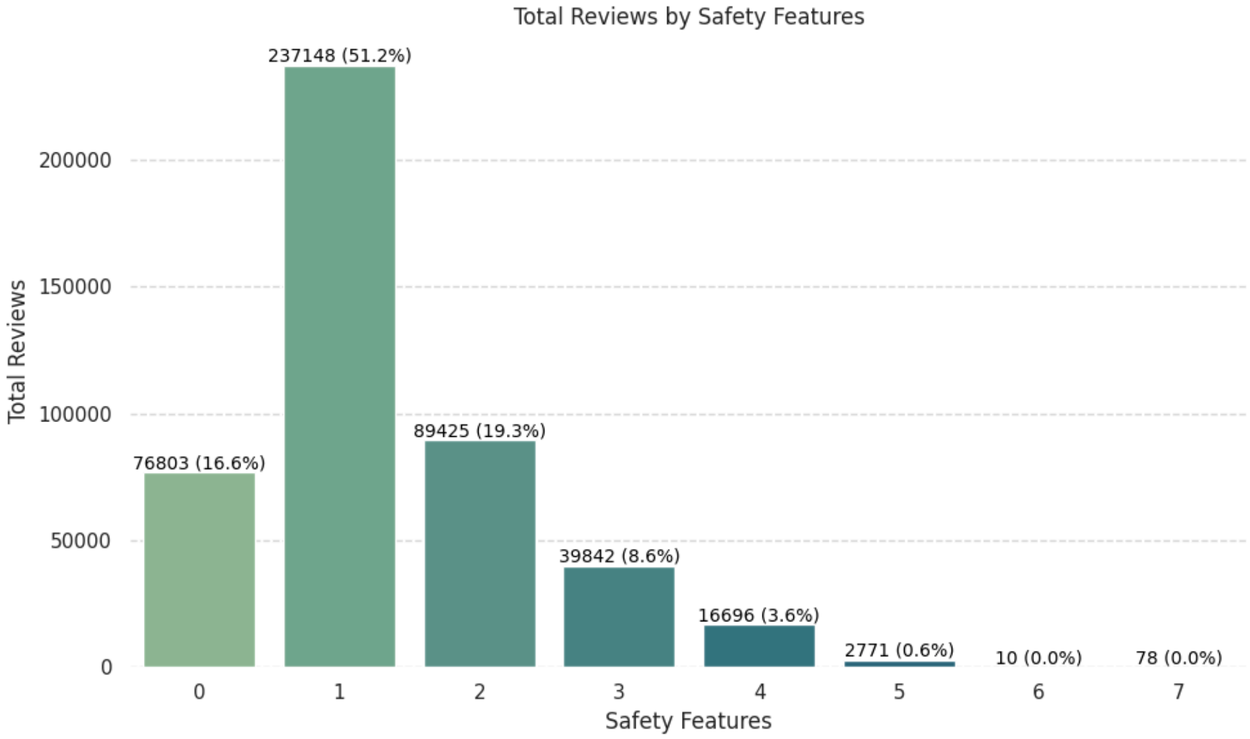


Meanwhile, listings with low family suitability have higher total reviews. Listings with a family suitability index of 1 or 2 have the highest total reviews, indicating that most users on the Airbnb platform are likely individuals, small families, or groups of friends.

Rate and Reviews Distribution by Safety Features



A higher safety index, which includes various amenities and safety features, leads to higher average ratings. Listings with safety index levels 6 and 7 have average ratings of 4 and 5.



Meanwhile, listings with lower safety indexes have higher total reviews, indicating that users on the Airbnb platform may not prioritize safety-related amenities and features. Listings with a safety index of 1 have the highest total reviews, followed by indexes of 2, 0, and 3.

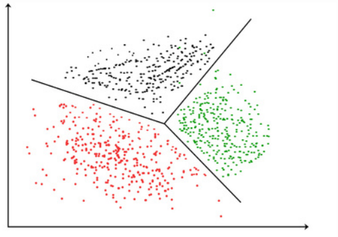
K means Clustering Based Recommendation System

K means Clustering Based Recommendation System

Goal: recommend listing for user based on cluster of listing created by model.

User Booking History

Recommendation function based top listing in each cluster



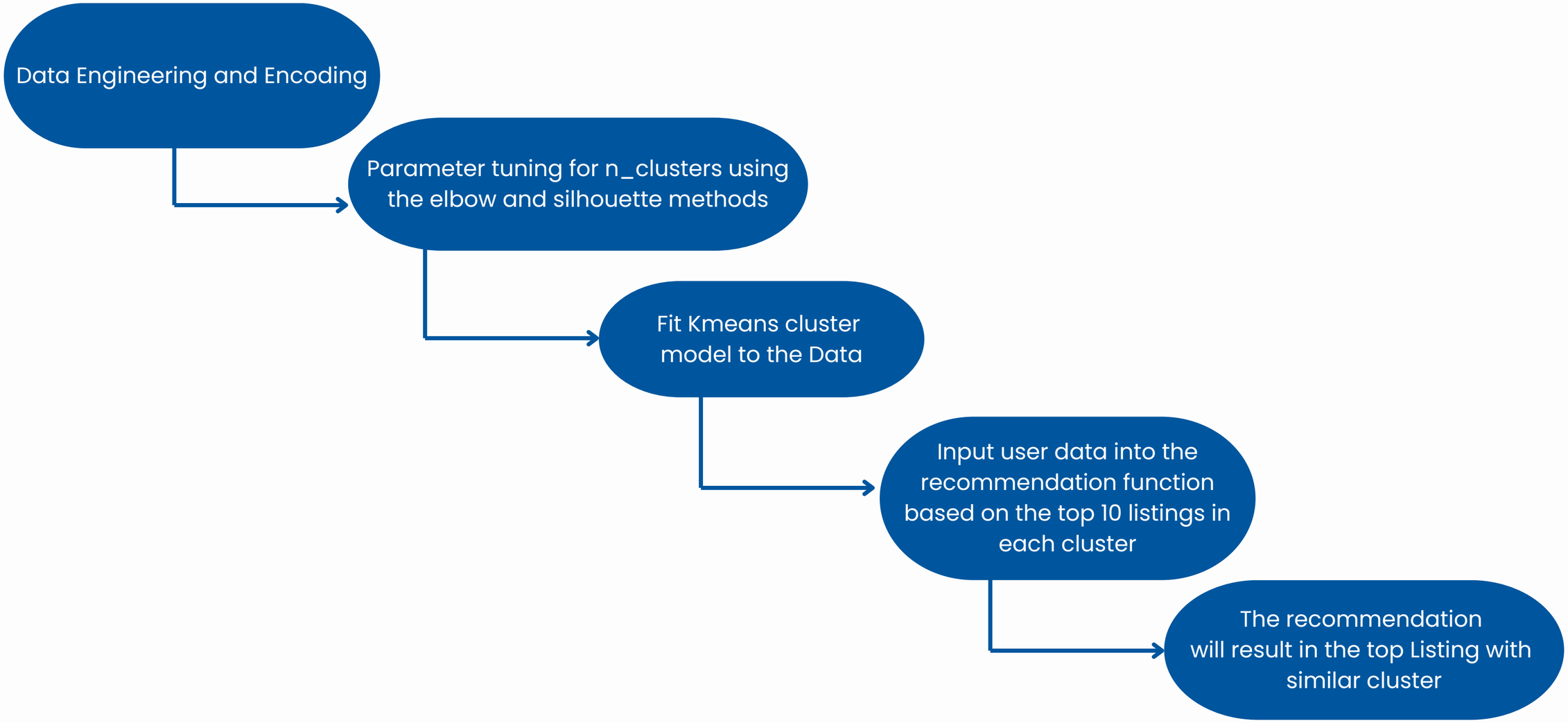
Recommendation Result

The user will be recommended top listings with the same cluster from listing in their booking history

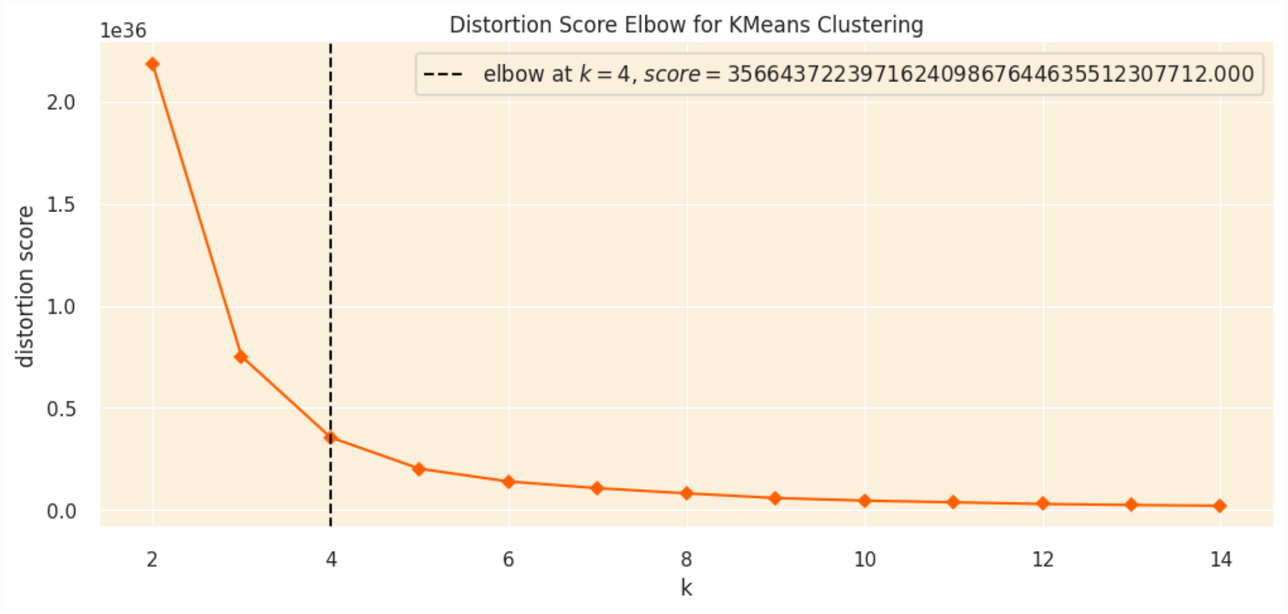
K-means clustering is an unsupervised machine learning algorithm used to partition a dataset into K distinct, non-overlapping subsets (clusters) based on feature similarity.

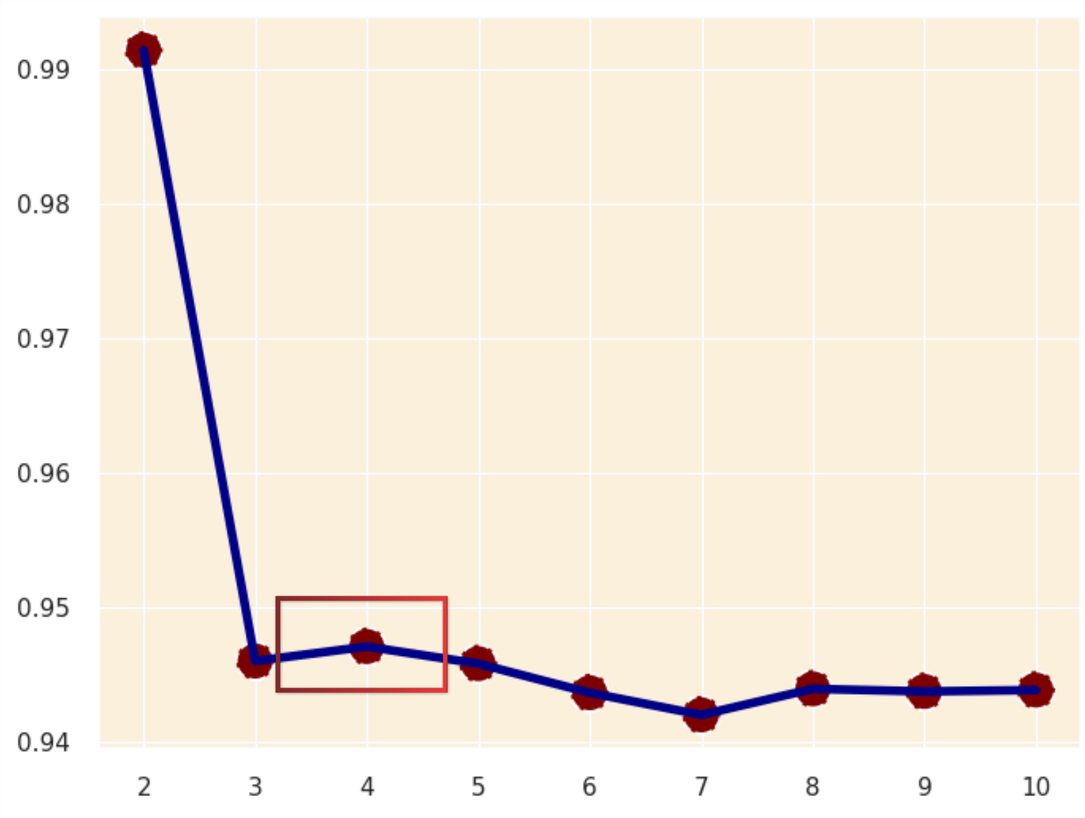
It aims to minimize the variance within each cluster by iteratively assigning data points to the nearest cluster centroid and updating the centroids based on the mean of the assigned points.

K means Clustering Recommendation System



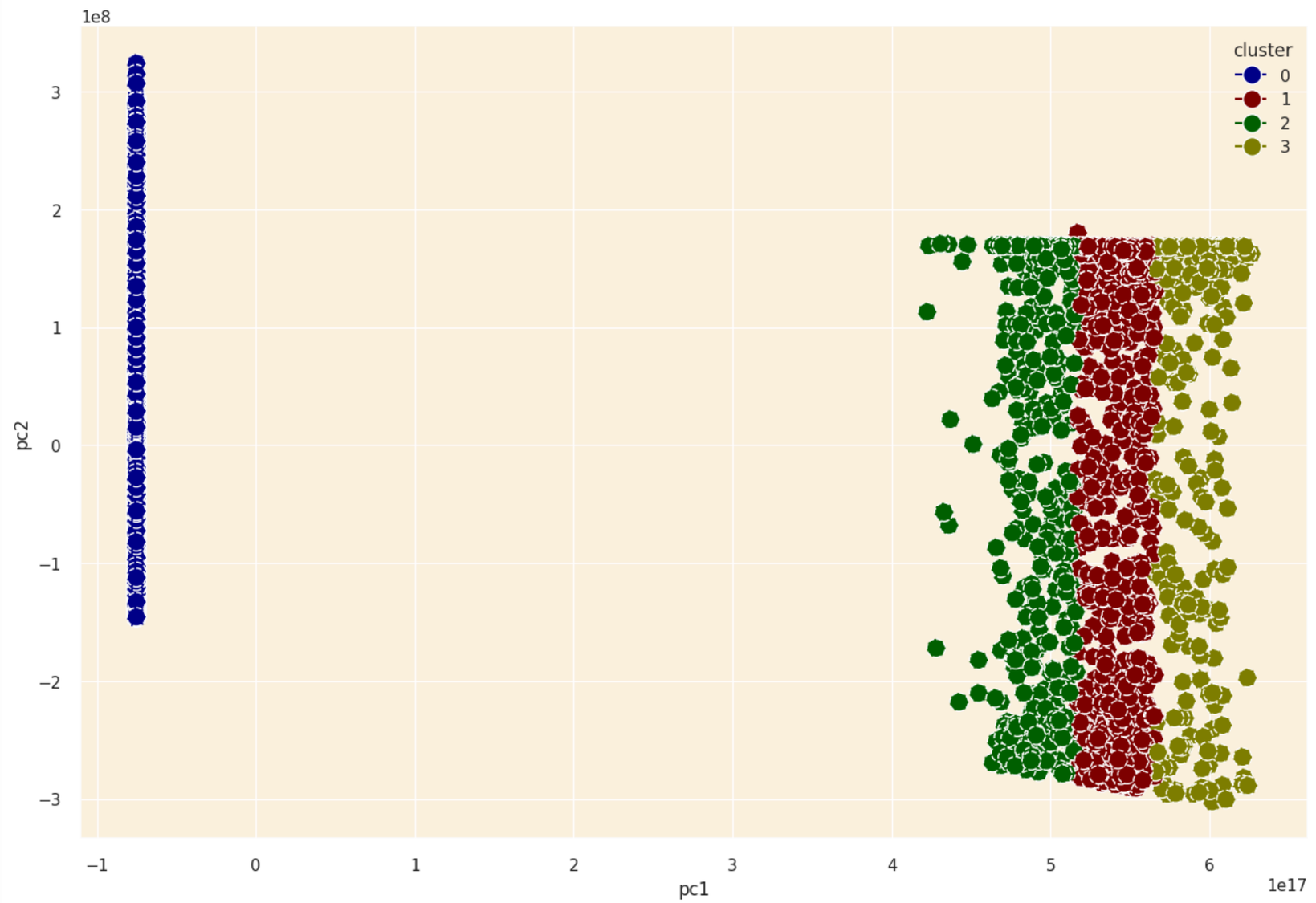
Parameter Tuning (Elbow and Silhouette Method)





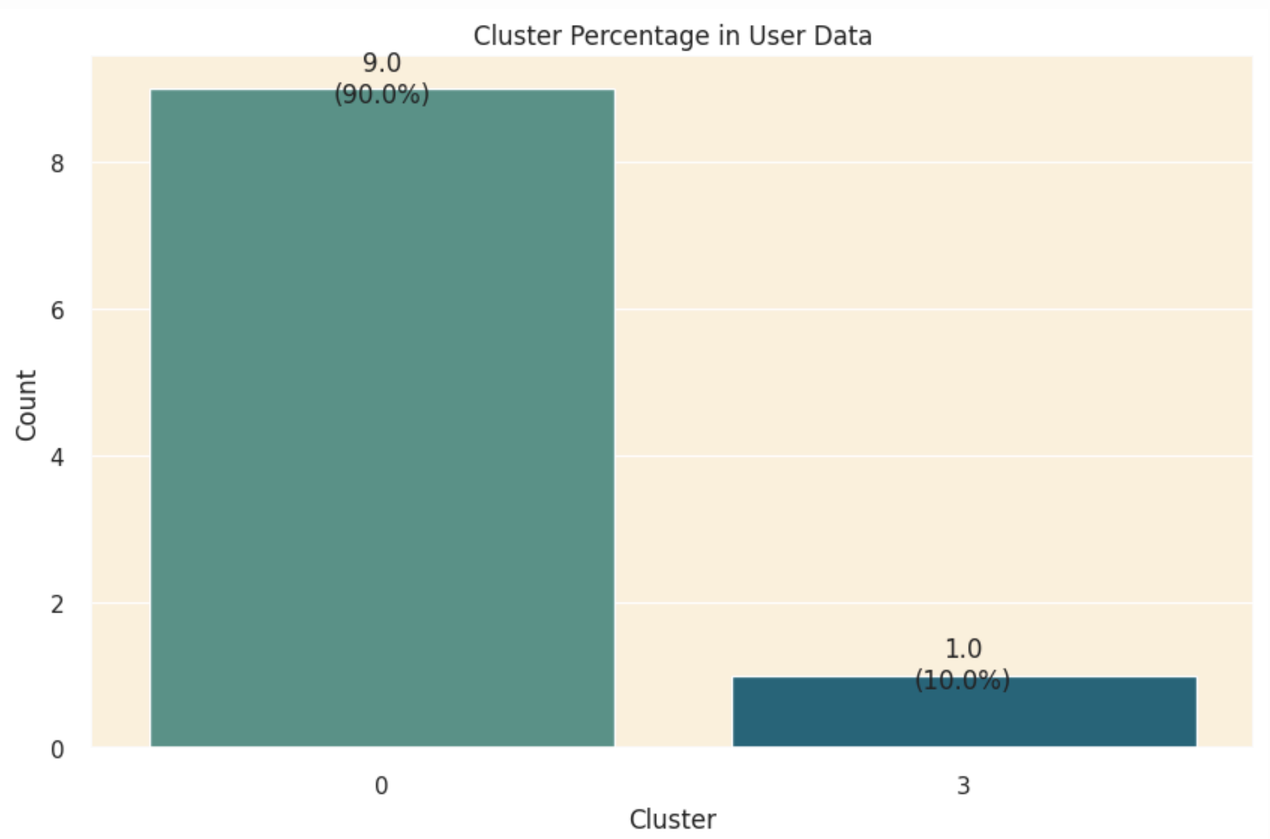
Both the elbow and silhouette results indicate that the suitable number of clusters is 4.

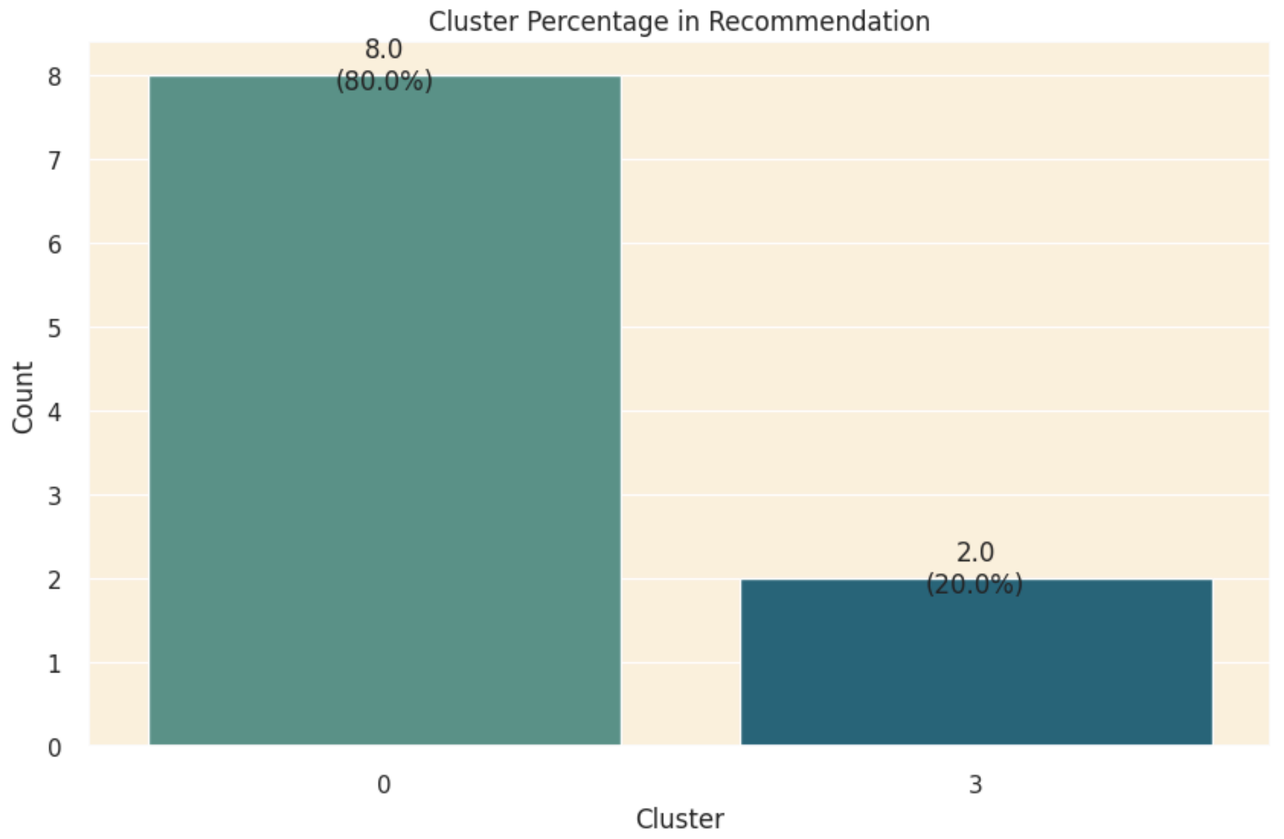
Clustering Result



The results from the clustering analysis reveal that Cluster 0 differs significantly from the other clusters, while Clusters 1, 2, and 3 exhibit more similarity in various features.

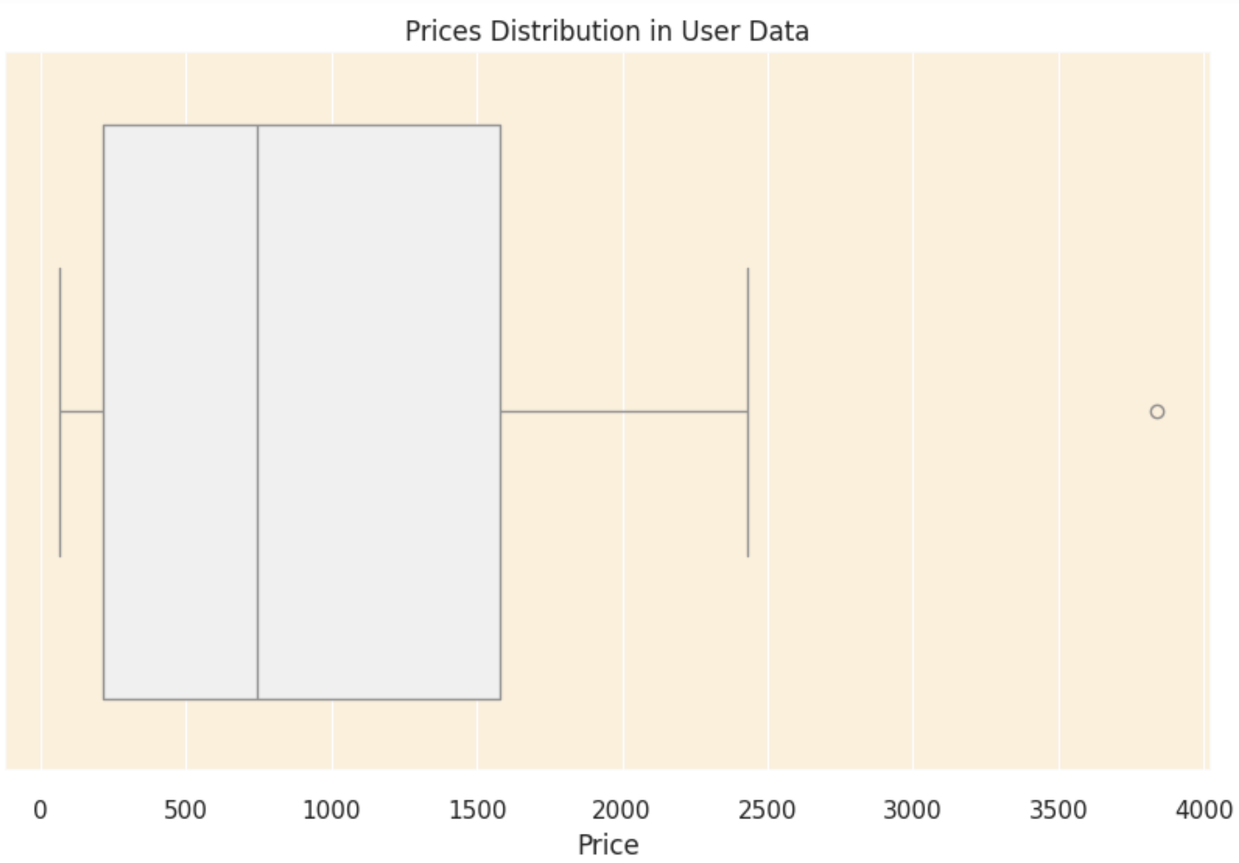
Recommendation system result





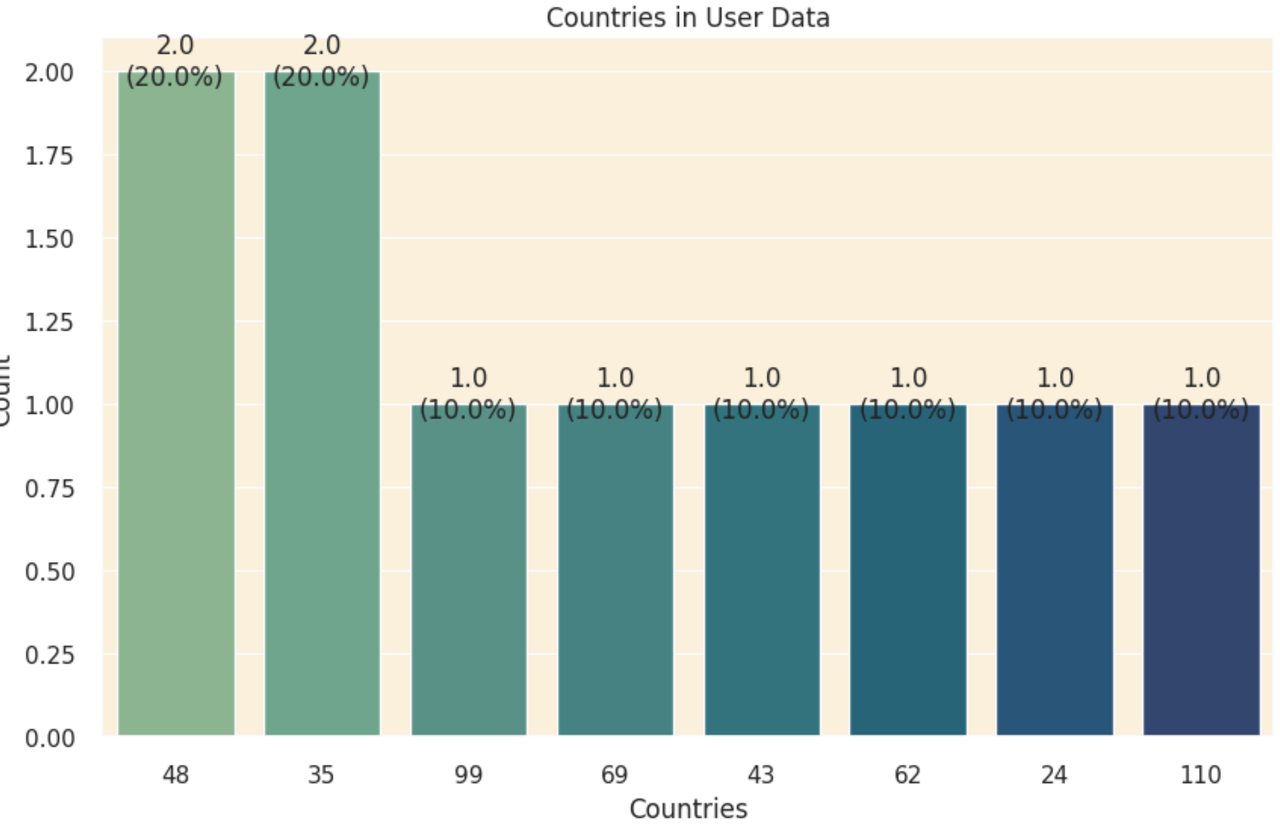
First, we evaluate the percentage of listings in the user's booking history compared to the recommendation results.
The results show that the user has 90% of their booking history in listings from cluster 0 and 10% in cluster 3. The recommendations provide 80% of listings from cluster 0 and 20% from cluster 3. This indicates that the recommendations are accurate, as they align closely with the user's booking history, with only a minor difference in percentages from the amount of recommended listing cluster.

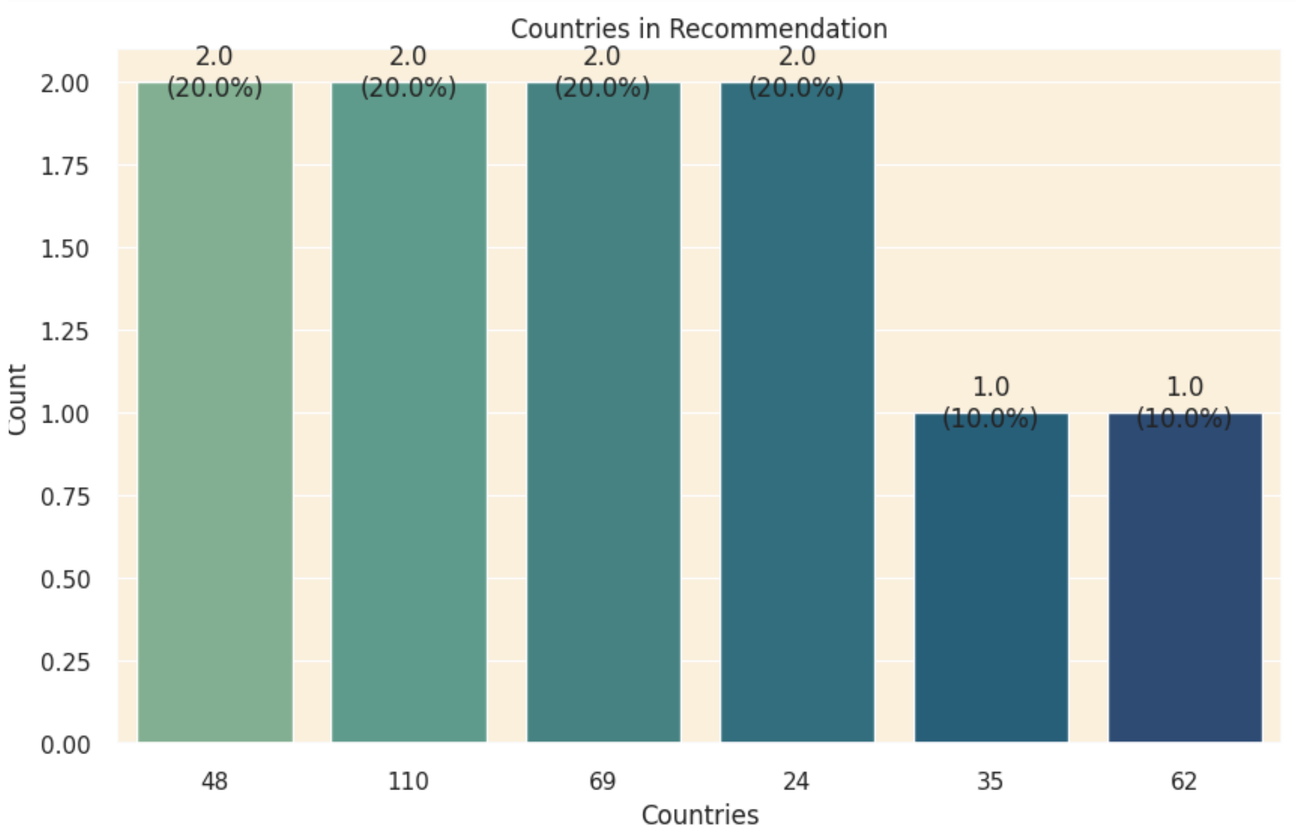
Recommendation system result



Next, we evaluated the price range of listings recommended by the system by comparing it with the price range in the user's booking history. The user's booking history shows a price range from $100 to $2,500, with one listing priced around $3,900. The recommendations present a price range from $100 to $2,500, indicating that the system accurately reflects the user’s price range preferences, although it does not account for the outlier listing priced at $3,900.

Recommendation system result





Lastly, we compared the listing countries in the user's booking history with those in the recommendation results. Since the data is encoded, countries are represented by numbers. The countries represented by numbers 48, 35, 69, 24, 110, and 62, which appear in the user's booking history, also appear in the recommendations. This suggests that the recommendation system accurately aligns with the user's past location preferences, specifically regarding the countries of previous bookings.

CONCLUSION

01 A Weighted Rating recommendation system is highly useful for providing general recommendations for listings with good reviews and ratings based on user preferences. In this project, recommendations are made based on the targeted country and price range.

02 A recommendation system based on a clustering model is highly effective, especially for recommendations based on user booking history. This is because the recommended listings are chosen with better specifications, focusing on features that align with the user's booking history. As a result, the variation in the recommendations is closely related to the user's past bookings.

ACTIONS

01 Deploying a recommendation system using Weighted Rating and K-Means cluster models on the Airbnb platform can greatly enhance the user experience. The Weighted Rating model provides general recommendations based on high ratings and positive reviews, making it ideal for new users. The K-Means cluster model offers personalized recommendations based on booking history, ensuring active users remain engaged with the platform.

02 Creating a "Share Listings" feature allows users to share recommended listings with friends or family via social media. Additionally, providing a recommendation feedback option enables users to evaluate the accuracy of the recommendations, ensuring they align with their preferences. This feedback can be used for ongoing improvements.

THANK YOU!

Notebook:

Jevon Tama Sianipar _ AirBnB Recommendation System

Data Source:

AirBnB dataset (last update mei 2024)