Predicting Hotel Ratings: The Data-Driven Approach

Presented by:

Lokesh Kumar

Krithika Natarajan

Parthiban P

Vigneshwaran A

Anand Mohan

Mentored by:

Nimesh Marfatia

PROJECT LIFECYCLE

Problem Framing

Predicting Hotel Ratings – The Data Driven Approach

Data Collection

Web scraping data from booking.com

Data Cleaning

Preprocess and clean the data to handle missing values, outliers, and ensure it's in a suitable format for analysis.

EDA & Feature Engineering

Explore the data to gain insights, identify patterns, and visualize relationships between variables.

Create new features or transform existing ones to improve model performance.

Modelling

Data Splitting, model selection, model training & model evaluation

Inference

Business Recommendation

PROBLEM STATEMENT

The Indian Tourism and Travel industry relies heavily on online booking websites for hotel information, primarily through user-generated numerical ratings and reviews. However, these numerical ratings often lack the depth to convey the complete picture of a hotel's quality.

This project aims to develop a comprehensive hotel rating system for Indian hotels by leveraging web scraping and data analysis techniques to consider both text reviews and numerical ratings.

The goal is to provide hotels with actionable insights for improving their services, ultimately leading to increased customer satisfaction, loyalty, and business success.

NAVIGATING THE INDIAN HOSPITALITY LANDSCAPE

01
Online Hotel Reservations Surge

02
Flourishing Indian Tourism Industry

03
Rising International Arrivals

04
Uncharted Research Territory

05
Empowering Hotels for Excellence

06
Unveiling the Roadmap to Success

RESEARCH OBJECTIVE

This project analyzes text reviews, user ratings, and hotel amenities to create an overall hotel rating. Utilizing web scraping and data analysis on booking.com data, we pinpoint top-rated amenities, areas for improvement, and factors impacting hotel success. Our study empowers hotels with actionable insights to boost customer satisfaction, loyalty, and business success.

Analyzing customer sentiments in the Indian hotel industry is like deciphering the heartbeat of hospitality, where each review is a note, and every positive sentiment is a step towards the crescendo of guest satisfaction

Stages of Data Collection

1st step

Extracting hotel listings from **booking.com**

•Utilized Instant Data Scraper Browser Extension

2nd step

Gathering hotel information

- Developed an automated Python script
- Utilized BeautifulSoup4 library
- Employed XPATH for extracting desired variables

3rd step

Extraction of User Review

- Developed automated Python script
- Utilized Selenium library
- Extracted user reviews for each hotel from the website

4th step

Sentiment Analysis

- Text Analysis using NLP and Sentiment Analysis using VADER

5th step

Data Consolidation

- 7072 Hotel Records
- 27 Variables
- 98304 User Reviews

Data Pre-processing

- Outliers are present in all numerical variables, but the overall distribution is satisfactory.

Dropped 10 non significant variable

Non-essential Variables

Duplicate and Null Value

Outlier Treatment

Transformation

Removed the following data points:

- 192 Duplicate Values
- Null Values in “Staff, Facilities, Cleanliness, Value for Money, Location, Polarity and Comfort”

- In the 'Number of User Ratings' variable, 1222 records were eliminated due to having fewer than 5 user ratings

- 7 New features are created by transforming the existing variables

Data Transformation & Feature Engineering

- **Accessibility:** Distance of hotel from city center classified as Far (1), Near (2) & Very Near(3).

- **Type of City:** Destinatia (Only Tourism friendly) & Excursa (Business & Tourism)

- **Travel Sustainable Property:** Badge given to hotels for the sustainable practices implemented, ranging from Levels 1-3.

- **Free Wi-Fi:** Combined the numerical Free Wi-Fi with the Boolean column retrieved from “Features”

- **Polarity:** Sentiment of online user text reviews given by users.

- **Features:** Bucketed salient features offered by the hotels into 21 combined Boolean features. While some features were natively retained, rest were formed by combining existing features.

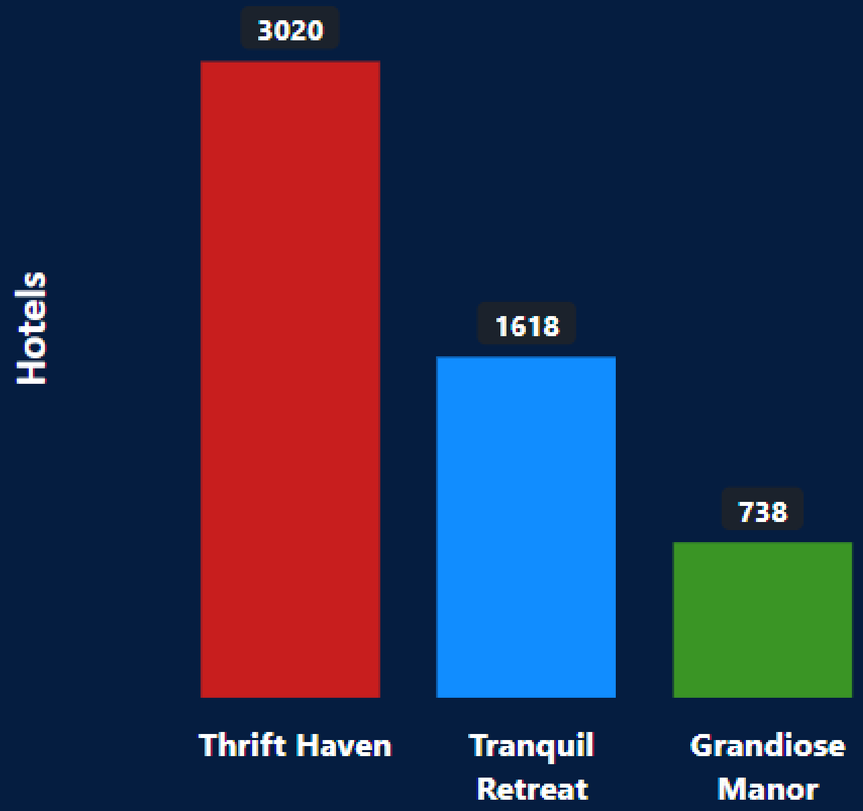
NEW FEATURES	EXISTING FEATURES
Customer Support	24-hour front desk, Tour desk, Wake-up Service
Smoking Zone	Non-smoking rooms, Designated Smoking area
Laundry	Laundry, Ironing Service, Dry Cleaning, Linen
Security	CCTV in common areas, 24-hour security, CCTV outside property, Safety deposit box, Security alarm, Smoke alarms
TV	TV, Flat Screen TV
Free Toiletries	Free Toiletries, Toilet Paper
Cab Service	Car Hire, Airport Shuttle
Furnished	Wardrobe or closet, Clothes rack, Desk, Telephone, Electric Kettle, Trouser Press
Special Treatment	Breakfast in the room, Express Check-in/ Check-Out

Data Transformation & Feature Engineering

Hotel Categorization:

- ✓ The hotels were categorized based on the Star Ratings and the Price of availability.
- ✓ This approach was taken up to arrive at targeted insights for the categories.
- ✓ Modelling was done separately for all three categories.

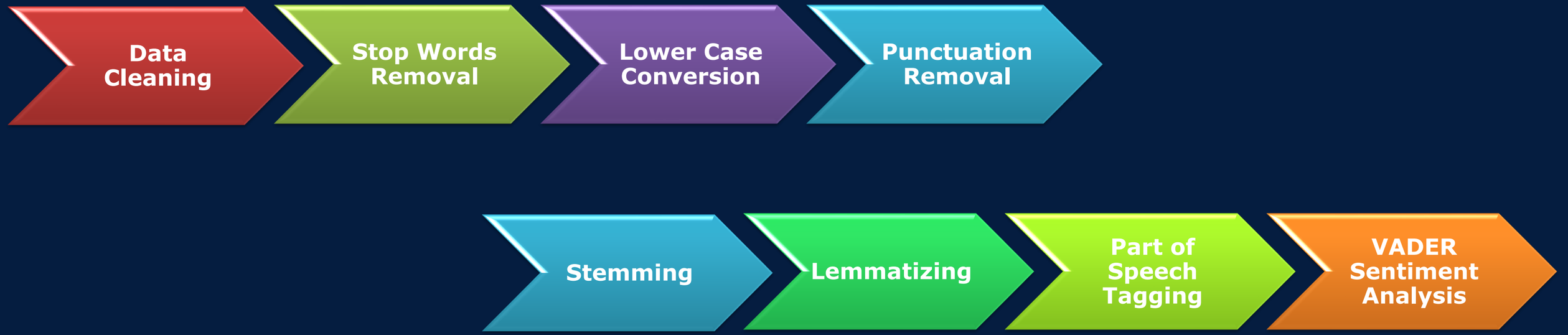
Category	Star Rating	Price (Rs)
Thrift Haven	1 & 2	< 2500
Tranquil Retreat	3	2500 to 5000
Grandiose Manor	4 & 5	> 5000



Text Analytics on User Reviews

- **Technique:** NLP with Sentiment Analysis
- **Library used for NLP:** NLP Tool Kit Library for Python
- **Library used for Sentiment Analysis:** VADER
- **Approach:** Bag of Words approach.

- **Reason for preferring VADER:**
  - ✓Extensively trained using social media & internet based text datasets.
  - ✓Has the ability to understand internet slang better than other libraries.

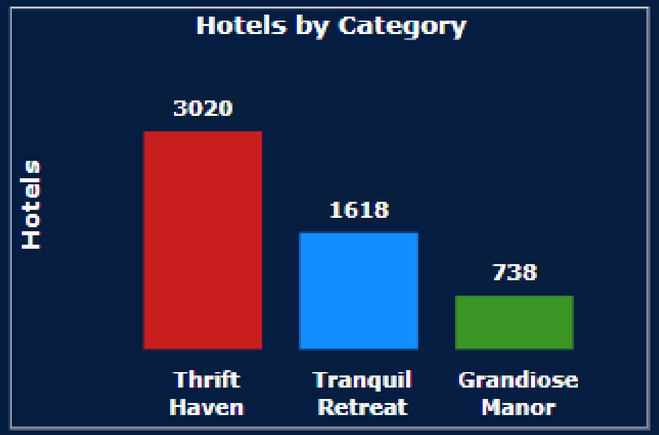


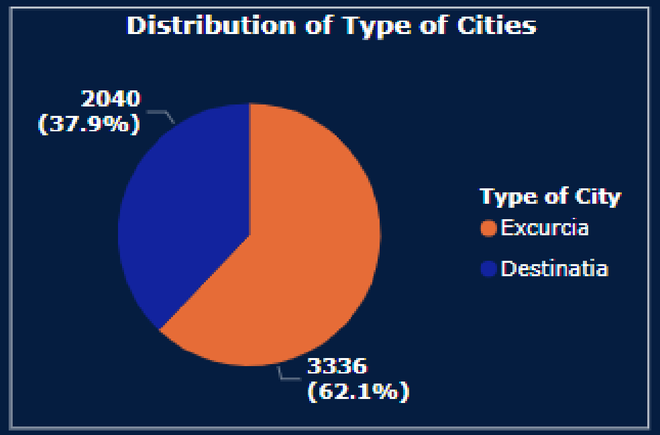
Text Analytics on User Reviews

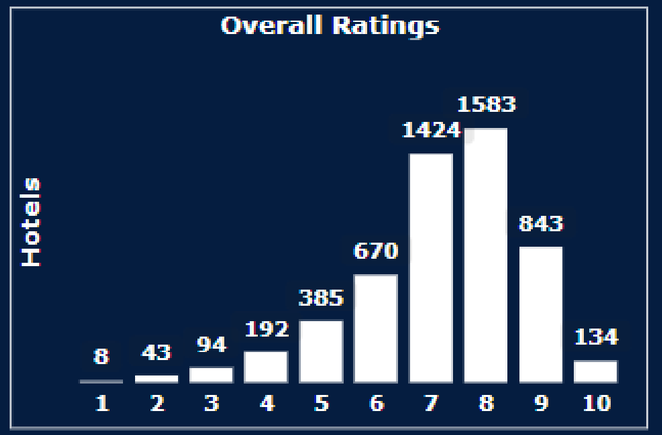
Steps involved in Sentiment Analysis:

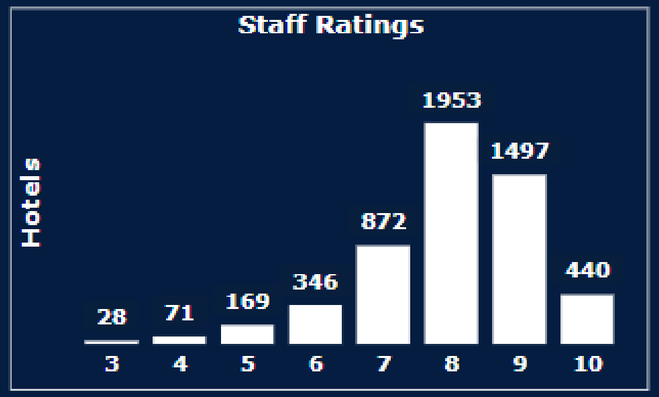
- ✓ VADER library was used
- ✓ VADER uses Sentiment Intensity Analyzer
- ✓ Takes the processed text as input and gives the Polarity as output
- ✓ Polarity score ranges from -1 to +1.
- ✓ Included the Polarity score in the original dataset.

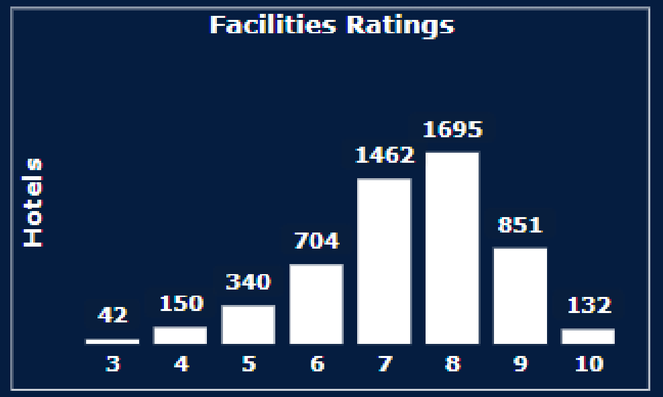
EXPLORATORY DATA ANALYSIS

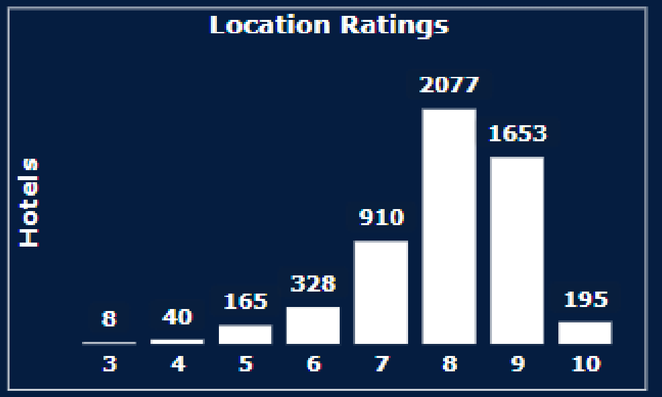


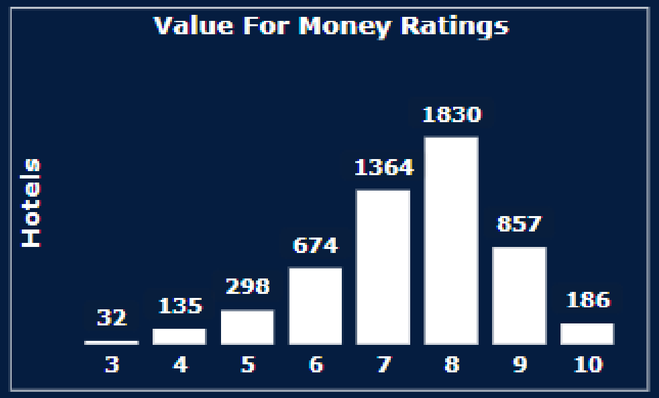


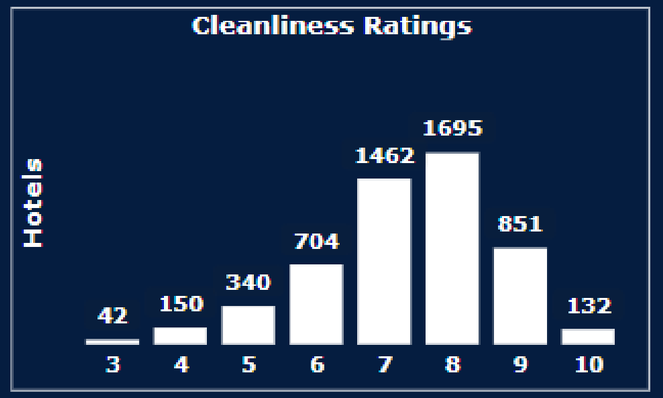


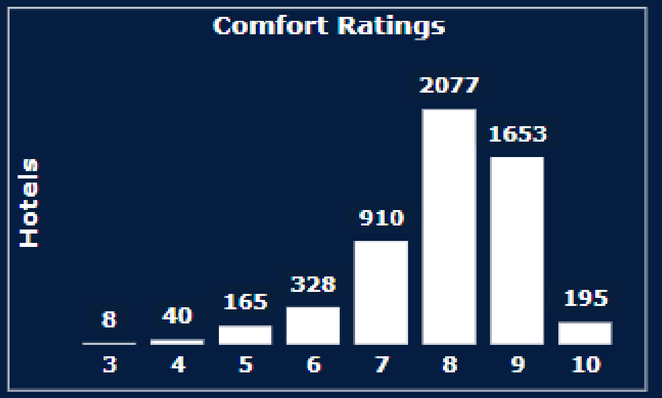




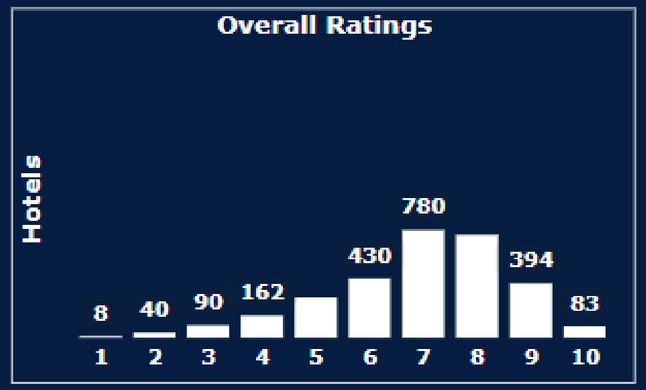


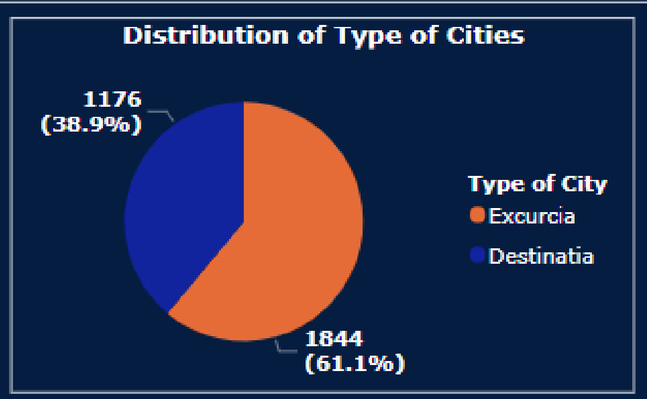


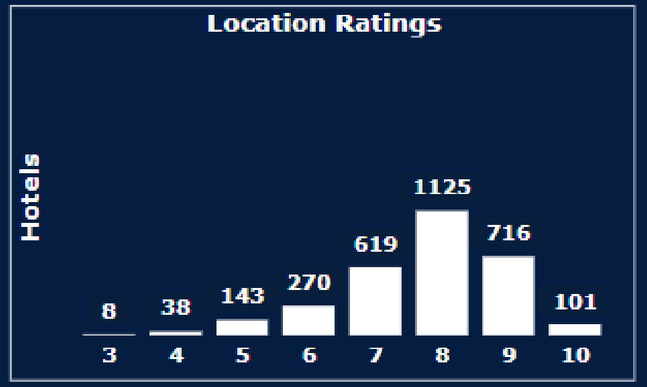


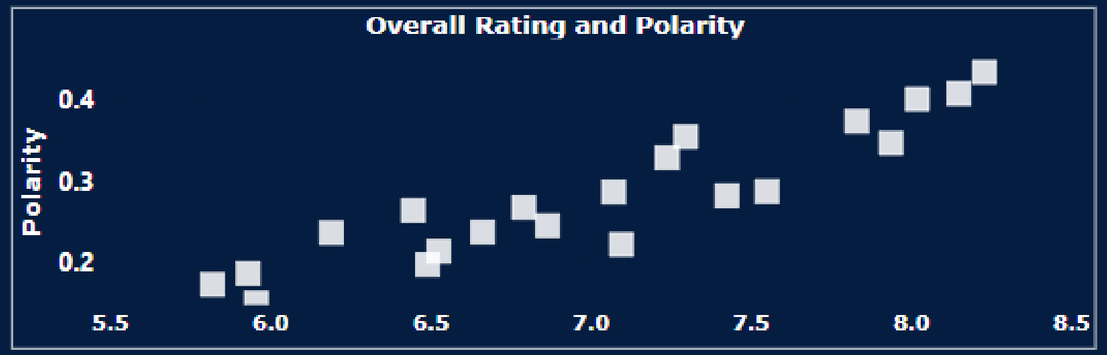


EDA - THRIFT HAVEN



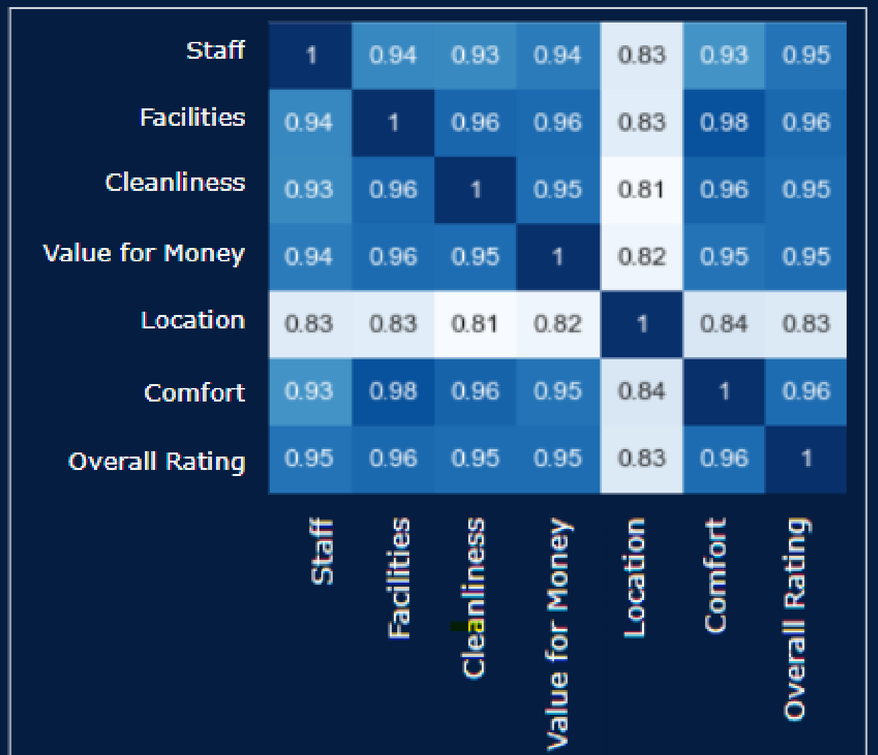




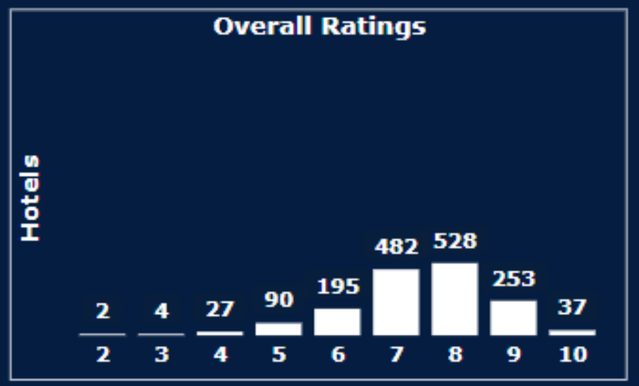


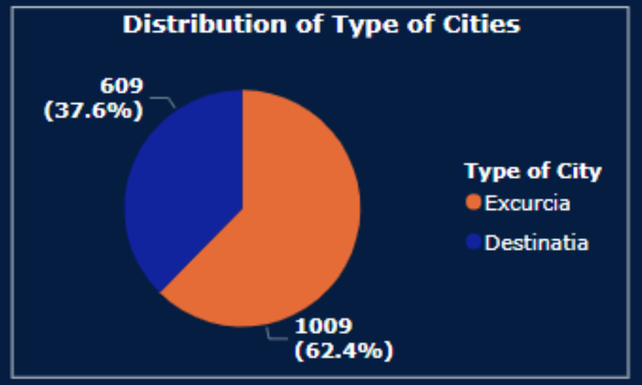
Prominent Features Offered

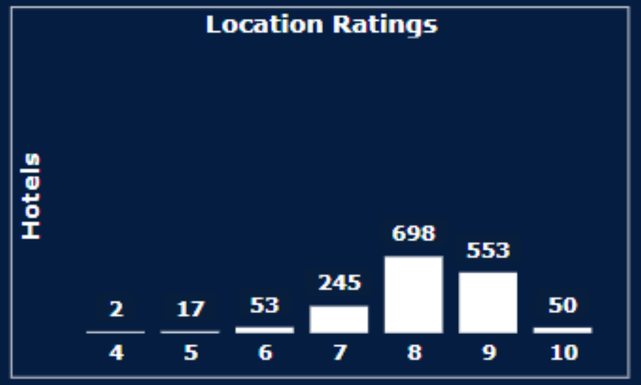
Prominent Features	Hotels Offering	Overall %	Rating > 7 (in %)
AC	2720	90.07	88.01
Wi-Fi	2238	74.11	83.68
Free Toiletries	1474	48.81	57.11
TV	1443	47.78	48.59
Free Breakfast	929	30.76	31.01
Total	8804	291.53	308.40

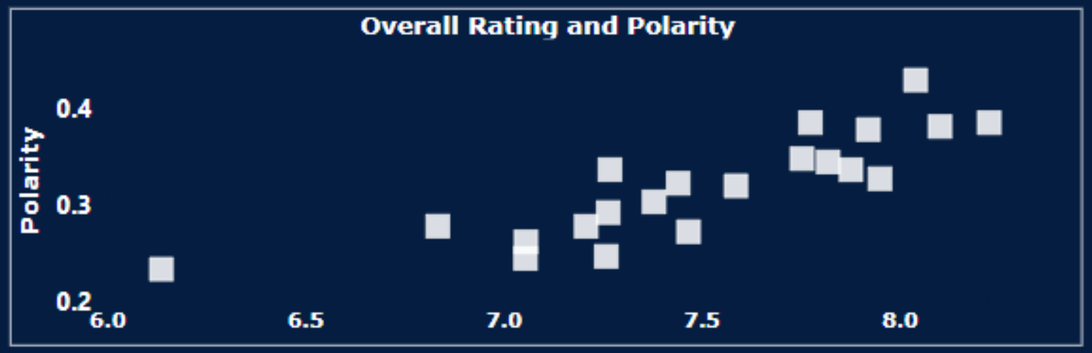


EDA - TRANQUIL RETREAT



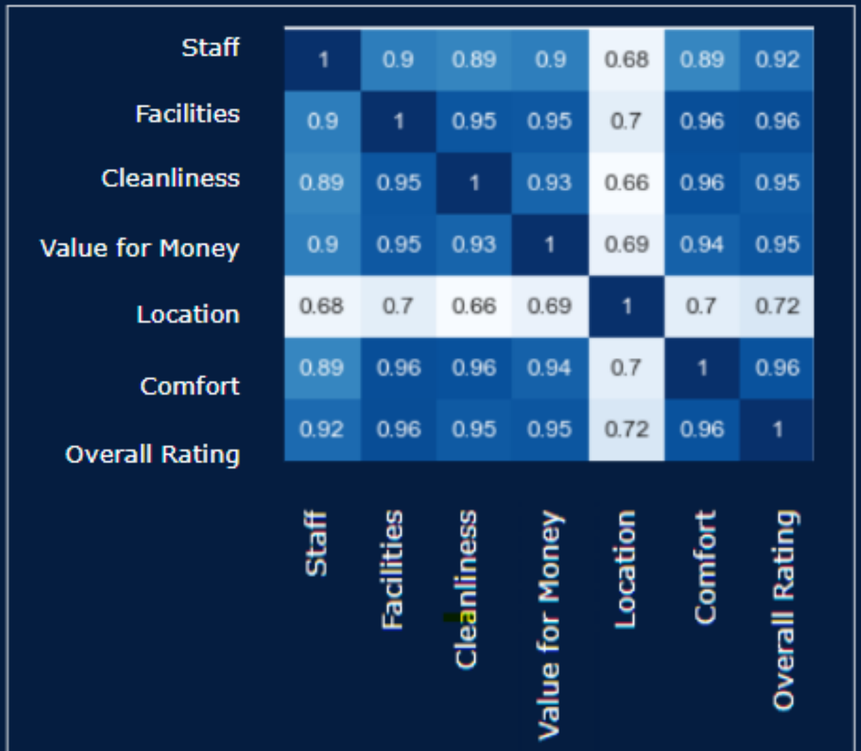






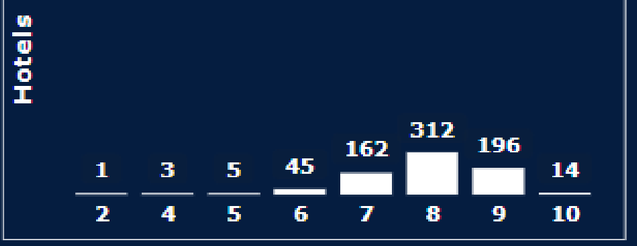
Prominent Features Offered

Prominent Features	Hotels Offering	Overall %	Rating > 7 (in %)
AC	1498	92.58	90.93
Wi-Fi	1418	87.64	89.26
Hindi	1022	63.16	62.31
TV	1020	63.04	59.63
Restaurant	787	48.64	48.89
Free Breakfast	622	38.44	35.00
Currency Exchange	562	34.73	33.70
Total	6929	428.23	419.72

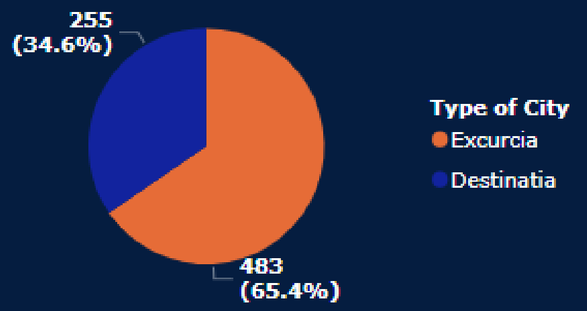


EDA - GRANDIOSE MANOR

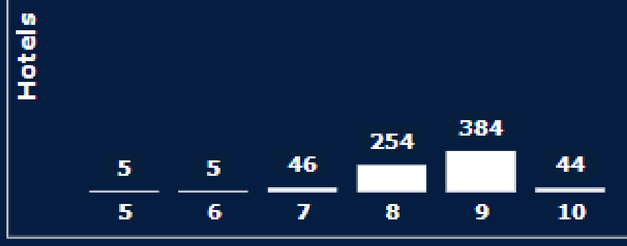
Overall Ratings



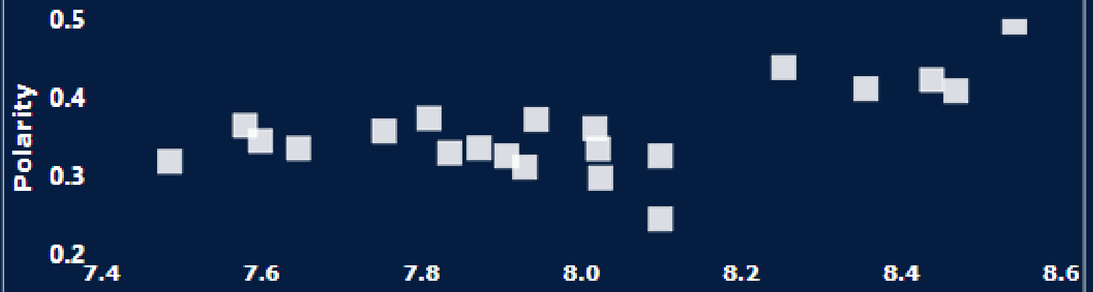
Distribution of Type of Cities



Location Ratings

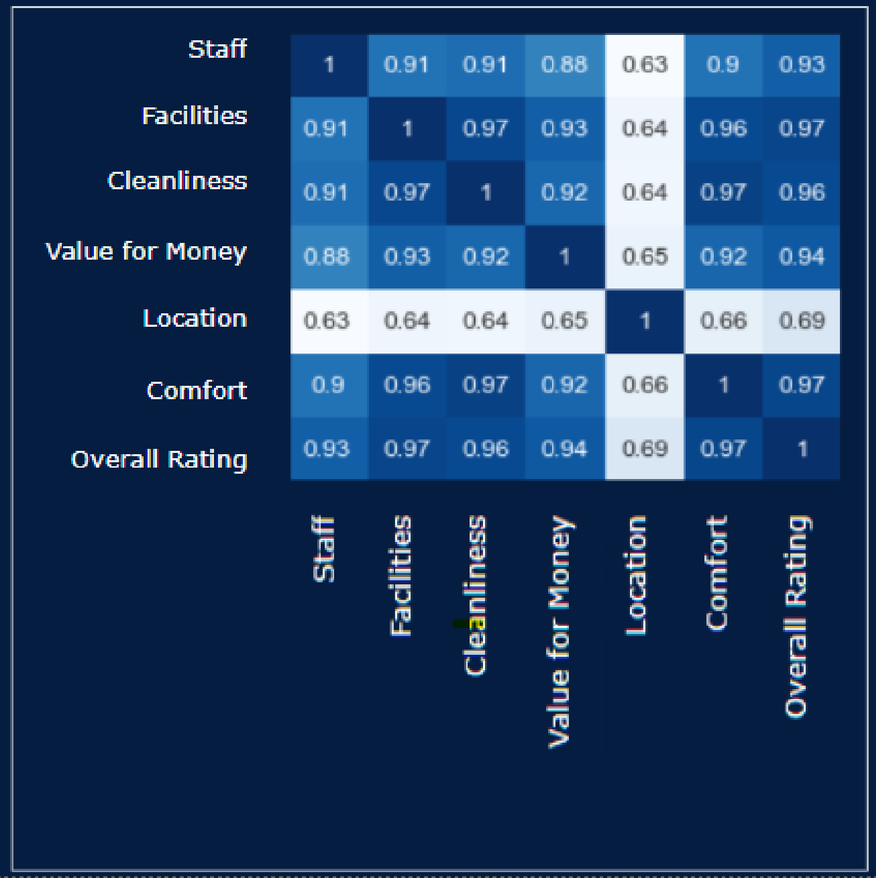


Overall Rating and Polarity



Prominent Features Offered

Prominent Features	Hotels Offering	Overall %	Rating > 7 (in %)
Wi-Fi	665	90.11	90.97
Laundry	385	52.17	47.26
Room Service	357	48.37	43.23
Restaurant	324	43.90	40.16
Lift	292	39.57	35.16
Room With A View	192	26.02	25.32
Total	2215	300.14	282.10



Modeling Approach

- Target Variable : Overall rating (continuous variable)
- Number of dependent Variables : 34
    - ❖ No of continuous variables: 07
    - ❖ No of binary variables: 27
- Since the target variable is a continuous variable, this is a regression problem and we will be using the following Algorithms for analysis.

1. Linear regression
2. Ridge regression
3. Lasso regression
4. Elastic net regression
5. Decision tree / CART
6. Random forest

Equation based models

Tree based models

Modeling Approach

➢ Approach taken in equation based models

Removing columns with VIF > 5	Removing columns with P value > 0.05	Train-test split : 75:25
Dealing with multi co-linearity	Dealing statistically unimportant variables	For checking model validity

➢ Approach taken in tree based models

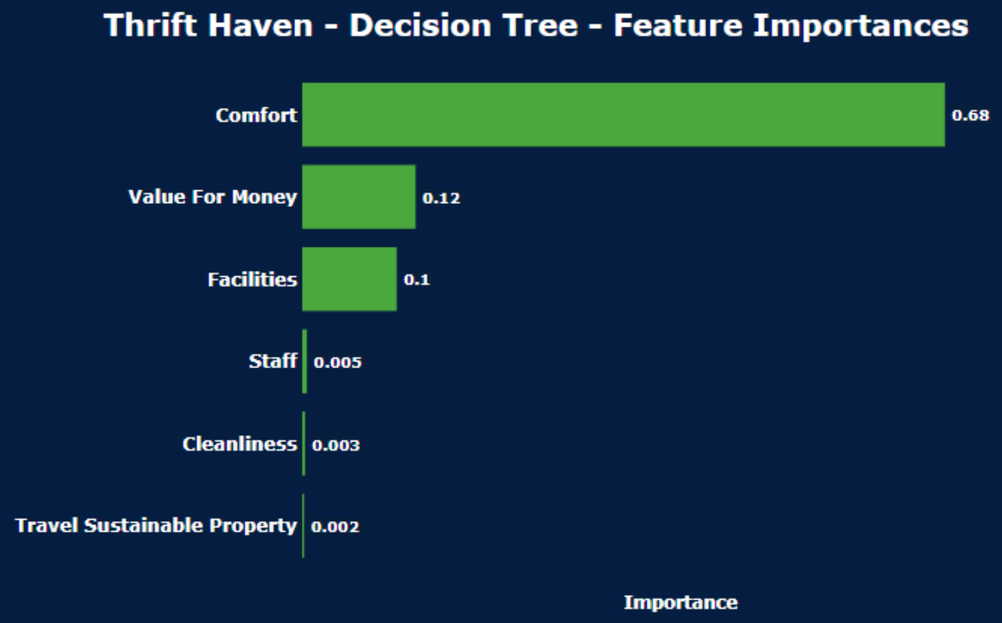
- ❖ CART : Grid search optimization
- ❖ Random forest : No pruning, since it uses bagging technique
- ❖ In both models Train-Test split is 75:25

Modeling Results

Thrift haven

Models	R^2		MAPE		RMSE	
	TRAIN	TEST	TRAIN	TEST	TRAIN	TEST
Linear Regression	0.791	0.797	0.100	0.102	0.764	0.765
Ridge	0.791	0.797	0.100	0.104	0.763	0.771
Lasso	0.700	0.694	0.123	0.130	0.915	0.948
Elastic Net	0.788	0.794	0.117	0.124	0.855	0.855
Decision Tree	0.973	0.972	0.030	0.031	0.273	0.273
Random Forest	0.976	0.880	0.037	0.070	0.264	0.570

- Decision tree, despite having high R^2, does not explain the importance of variables effectively.
- All the important columns are from Rating columns. Effect of features are very insignificant while using Decision tree
- In linear regression, since we are eliminating columns based on VIF, we get more weightage on Features too



Linear regression Equation:

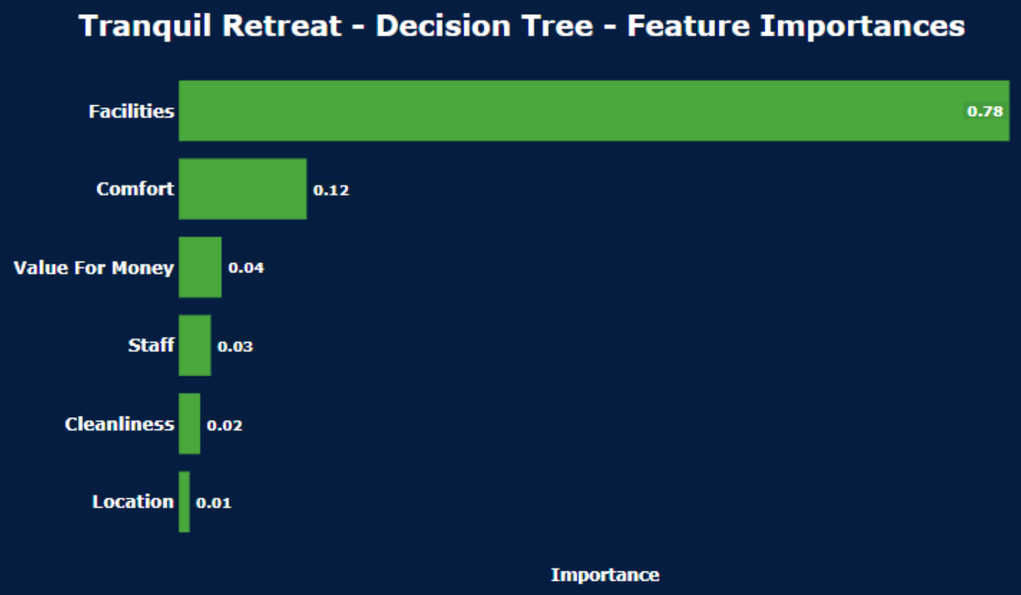
- 0.431 + 2.33 * Polarity + 0.722 * Location + 0.065 *Type of City Destinatia + 0.115 * Accessibility_2+ 0.159 * Free Wi-Fi + 0.599 * Air conditioning - 0.182* Smoking Zone + 0.197 * Free Breakfast + 0.147 * TV + 0.093 * Travel Sustainable Property 1 + 0.271 * Free toiletries

Modeling Results

Tranquil retreat

Models	R^2		MAPE		RMSE	
	TRAIN	TEST	TRAIN	TEST	TRAIN	TEST
Linear Regression	0.718	0.753	0.070	0.069	0.631	0.606
Ridge	0.718	0.754	0.070	0.070	0.630	0.612
Lasso	0.543	0.587	0.092	0.093	0.802	0.792
Elastic Net	0.709	0.752	0.079	0.080	0.693	0.693
Decision Tree	0.947	0.951	0.029	0.028	0.276	0.259
Random Forest	0.968	0.798	0.025	0.058	0.217	0.529

- In linear regression, by eliminating columns based on VIF, we assign more significance to features.
- Hence we are zeroing in on Linear Regression as best fit.



Linear regression Equation:

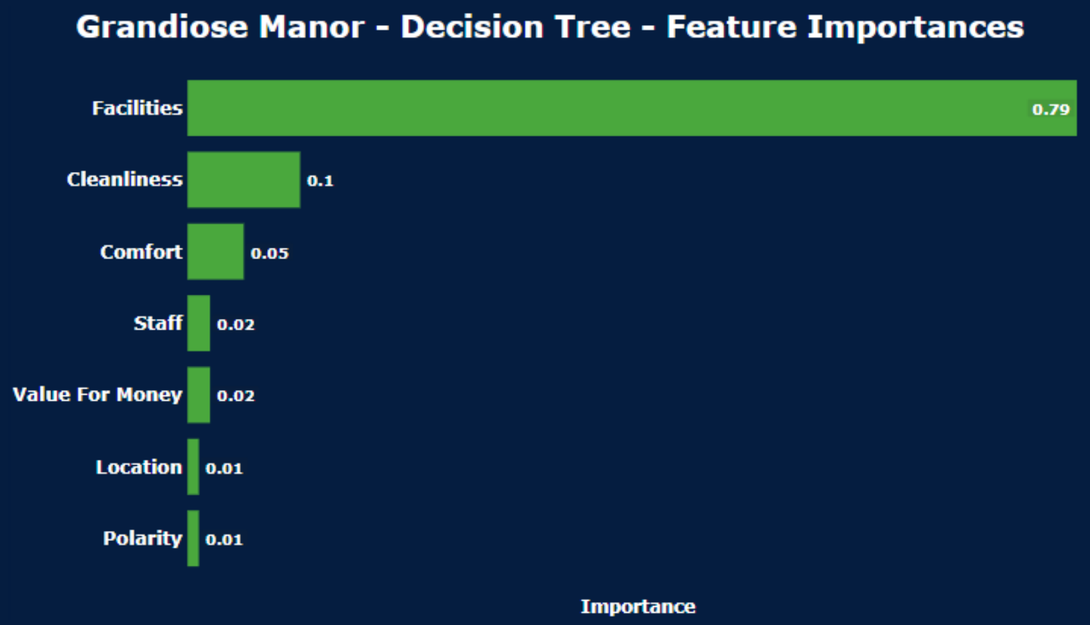
0.22 + 0.496 * Location + 2.092 * Polarity + 0.195 *Restaurant + 0.18 * Travel Sustainable Property 3 + 0.842 * Free Wi-Fi + 0.35 * TV + 0.254 * Type of City Destinatia - 0.266 * Smoking Zone + 0.307 * Free Breakfast + 0.408 * Currency exchange + 0.44 * Hindi + 0.24 * Air conditioning

Modeling Results

Grandiose manor

Models	R^2		MAPE		RMSE	
	TRAIN	TEST	TRAIN	TEST	TRAIN	TEST
Linear Regression	0.909	0.908	0.028	0.031	0.280	0.308
Ridge	0.909	0.908	0.028	0.032	0.279	0.312
Lasso	0.870	0.858	0.034	0.041	0.334	0.387
Elastic Net	0.902	0.898	0.050	0.059	0.580	0.580
Decision Tree	0.984	0.980	0.009	0.011	0.110	0.130
Random Forest	0.954	0.785	0.021	0.042	0.203	0.443

- In linear regression, VIF-based column elimination increases feature importance.
- Thus Linear Regression is the accurate model of choice.

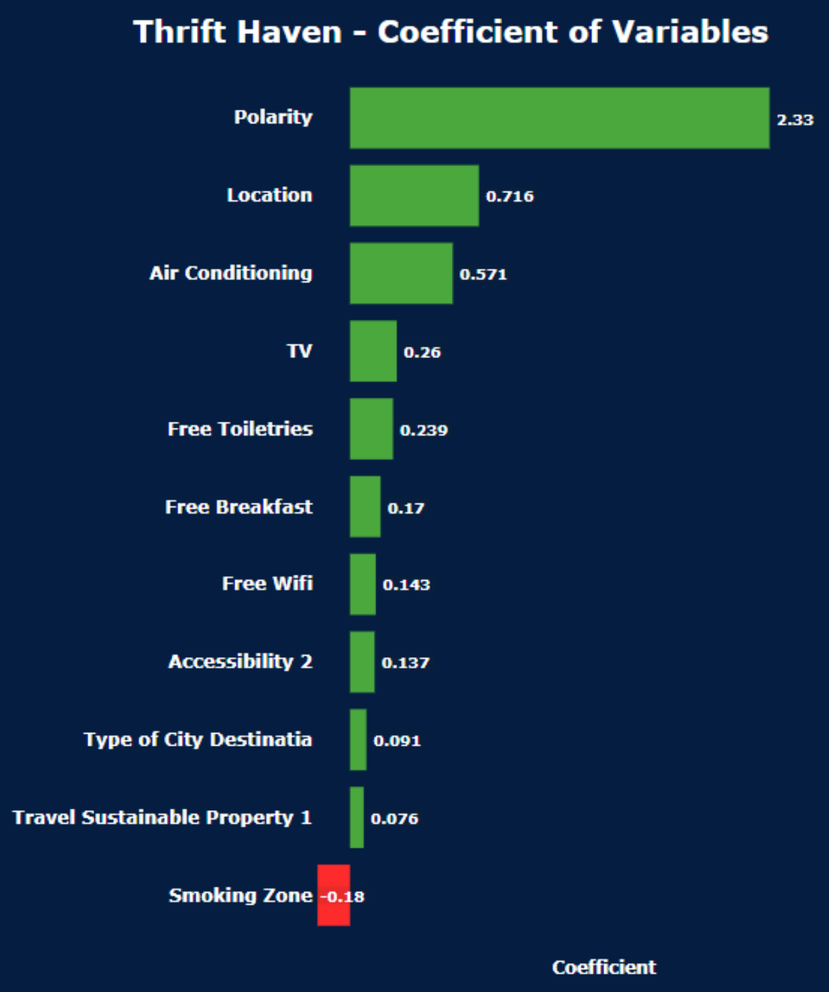


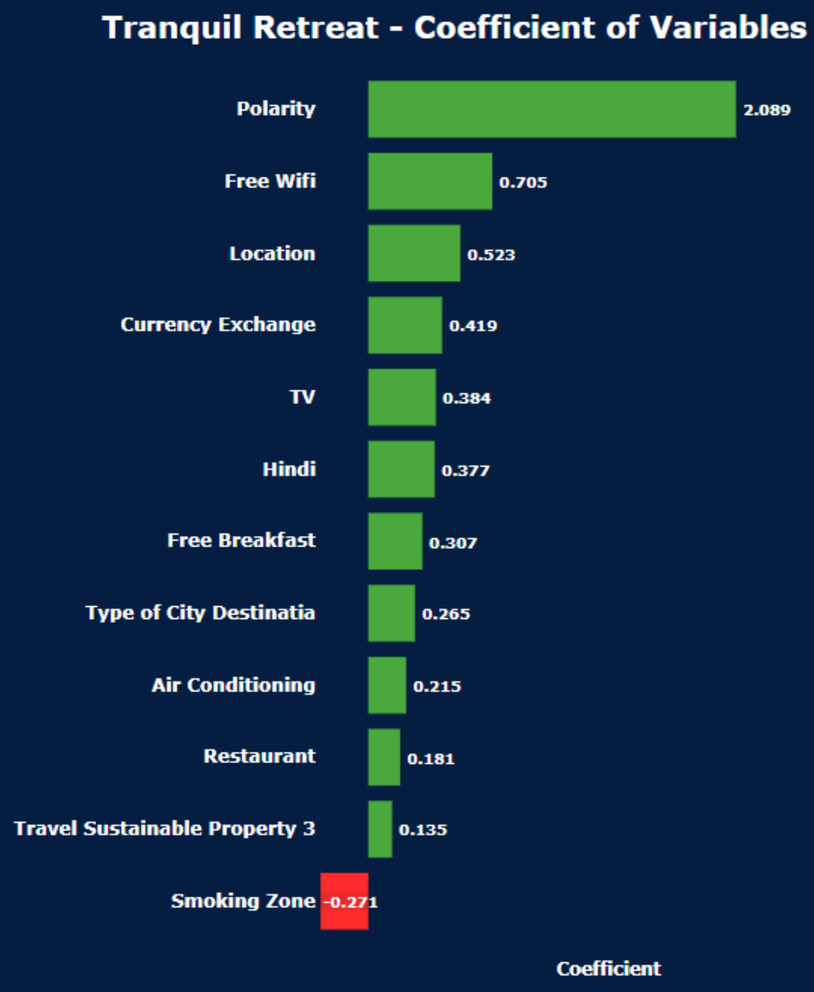
Linear regression Equation:

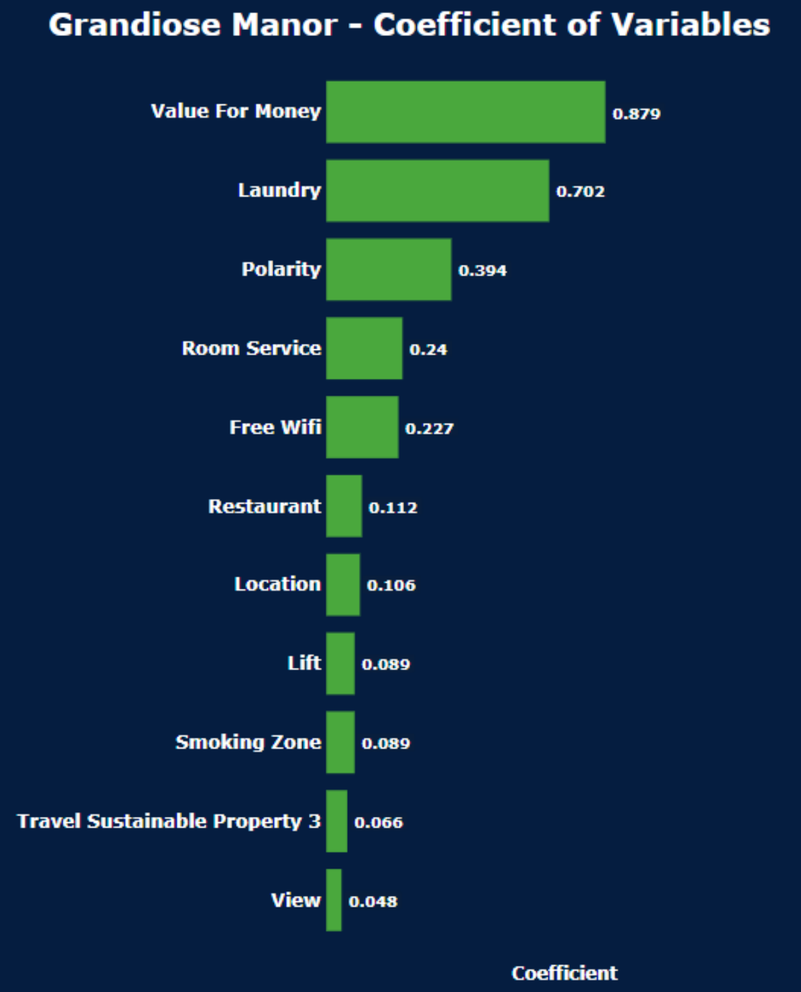
-1.417 + 0.876 * Value for Money + 0.108 * Location + 0.416 * Polarity + 0.079 * View + 0.176 * Room service + 0.192 * Free Wi-Fi + 0.075 * Lift + 0.843 * Laundry + 0.087 * Restaurant + 0.068 * Travel Sustainable Property 3 + 0.091 * Smoking Zone

Modeling Results – Linear Regression

Comparison based on category







INFERENCE

Through the analysis, it has become evident that customer expectations for amenities vary based on the preferred budget category of rooms, which are classified in this project as Thrift Haven (Standard), Tranquil Retreat (Deluxe), and Grandiose Manor (Luxury).

Thrift Haven

- The standout attributes within this category encompass air conditioning, Wi-Fi access, complimentary toiletries, television, and inclusive breakfast service.
- Analysis revealed that over 90% of the featured hotels offer air conditioning, with approximately 88% of them receiving ratings exceeding 7 out of 10. Similarly, accommodations with strong Wi-Fi, complimentary toiletries, television, and complimentary breakfast were found to receive ratings of 7 and higher at rates of 83%, 57%, 48%, and 31%, respectively.
- To enhance their offerings, hotels in this category should prioritize the provision of complimentary toiletries, improving television and entertainment options, and ensuring the quality of their complimentary breakfast services.

Tranquil Retreat

- In this category, we evaluated several key amenities, including air conditioning (AC), Wi-Fi access, Hindi-speaking staff, television (TV), on-site restaurant, complimentary breakfast, and currency exchange services.
- For this customer segment, hotels with good AC and Wi-Fi, both with ratings above 7, stood at 90% and 89%, respectively.
- Approximately 62% of hotels with Hindi-speaking staff received ratings of 7 and above, highlighting the comfort provided to tourists from diverse regions due to a commonly spoken language.
- Notably, TV, restaurant, complimentary breakfast, and currency exchange were highly preferred by customers in this category, with ratings above 7 standing at 59%, 48%, 35%, and 33%, respectively. These preferences set them apart from customers in the 'Thrift Haven' category of hotels.

Grandiose Manor

- Customers in the luxury hotel segment expect premium value, such as scenic views and private smoking areas. Basic amenities like AC are considered standard.
- For these hotels, the key focus should be on providing outstanding Wi-Fi, laundry service (with willingness to pay extra), prompt room service, in-house restaurants, reliable lift service, and rooms with a view.
- Amenities mentioned above have led to ratings exceeding 7, with Wi-Fi at 90%, laundry at 47%, room service at 43%, restaurant at 40%, lift service at 35%, and rooms with a view at 25%.
- In this segment, hotels must excel in every amenity, as customer reviews profoundly impact brand value and reputation, which are challenging to establish and even harder to maintain

Recommendations

Emphasize Amenities
Promote praised amenities in marketing plans

Harness User-Generated Content
Encourage guests to share their experiences on social media with a unique hashtag and repost for engagement and trust

Manage online reputation
Respond promptly to reviews, positive and negative, to demonstrate customer appreciation

Address Negative Feedback
Act on negative reviews to demonstrate your commitment to improvement

Promote Positive Reviews
Feature them prominently on your website and marketing materials

Train Staff for Excellence
Exceptional service boosts reviews and customer loyalty.

Leverage Testimonials
Use positive reviews as testimonials in emails and on social media.

Continuous Learning, Unlearning the know and Revisiting the strategy is the success mantra

THANK YOU