**SAFARIHUB – TOUR RECOMMENDER SYSTEM REPORT**



***Business Understanding***

***Overview***

Tourism is a thriving industry in Kenya, and travelers often face the challenge of choosing the right destinations for their trips. Our project aims to address this problem by creating a recommendation system that assists users in discovering personalized tourist destinations in the country.

***Problem Statement***

Travelers often struggle to choose the most suitable tourist destinations for their trips. With an overwhelming number of options available, personalized recommendations are crucial. Our project aims to address this challenge by creating a recommendation system that suggests relevant destinations in Kenya based on user preferences and historical interactions.

***Stakeholders***

1. **Travelers**: They seek relevant recommendations based on their preferences, interests, and historical interactions.
2. **Tourism Agencies**: These organizations can enhance user experiences by providing tailored suggestions.
3. **Local Businesses**: Recommendations can drive footfall to local attractions, restaurants, and accommodations.

***Objectives*:**

* "Build a collaborative filtering model to recommend destinations."
* "Reduce cold-start problem by incorporating content-based features."
* "Model Recall score ≥ 80%"
* "Model Accuracy ≥ 80%"

***Proposed Solution and Metrics of Success***

We propose building a hybrid recommendation system that combines collaborative filtering and content-based approaches. Success metrics include accuracy, recall and precision scores.

***Challenges***

1. **Data Quality and Diversity**:
   * Presence of missing values, outliers, or inaccuracies.
   * Ensuring diverse and representative data across different types of destinations (e.g., cities, beaches, historical sites) is essential.
2. **Cold-Start Problem**:
   * New users with limited interaction history pose a challenge. How do we
   * Balancing collaborative filtering (based on user behavior) with content-based filtering (based on destination features) is critical.
3. **Scalability and Real-Time Recommendations**:
   * As the user base grows, the system must handle increased computational demands.
   * Providing real-time recommendations during user interactions requires efficient algorithms.
4. **User Engagement and Interpretability**:
   * Recommendations should align with user interests to keep them engaged.
   * Ensuring transparency and interpretability of the recommendation process is important.

***Conclusion***

Our project has significant implications for travelers, tourism agencies, and local businesses. By solving this problem, we contribute to enhancing travel experiences and promoting local economies.

***Data Understanding***

***Data Sources and Relevance***

* The dataset was scraped using the **APIFY Tripadvisor Scraper**.
* It contains information about tourist destinations, including their names, categories, ratings, review counts, images, and other relevant features.
* The data's relevance lies in its ability to help us recommend destinations to travelers based on their preferences and historical interactions.

***Dataset Overview***

* The dataset consists of **2567 entries** (rows).
* Key columns include:
  + **Name**: The name of the destination.
  + **Category**: The type of destination (e.g., city, beach, historical site).
  + **Rating**: The average user rating (ranging from 1.0 to 5.0).
  + **Number of Reviews**: The count of user reviews.
  + **Image**: URLs to images representing the destinations.
  + **Photo Count**: The number of photos associated with each destination.
  + **Price Range**: Information about the cost level (if available).
  + **Review Tags**: Descriptive tags associated with reviews.
  + **Photos**: Additional photo URLs.
  + **Price Level**: Indication of price range (if available).

***Justification for Feature Inclusion***

* **Name**, **Category**, and **Rating**: Essential for personalized recommendations.
* **Number of Reviews**: Reflects popularity and user engagement.
* **Image** and **Photo Count**: Enhance user experience.
* **Price Range** and **Price Level**: Useful for budget-conscious travelers.
* **Review Tags**: Provides insights into user preferences.

***Data Limitations***

* **Missing Values**: Some entries lack ratings, images, or price information.
* **Limited Price Data**: Only 1487 entries have price-related details.
* **Data Quality**: Ensure data quality and handle missing values appropriately.

***Columns***

* name
* category
* rating
* numberOfReviews
* image
* photoCount
* priceRange
* reviewTags
* photos
* priceLevel

***priceLevel column***

1. Luxury: The most expensive category, offering premium services and facilities.
2. Premium: Mid-range in price, providing high-quality services and accommodations.
3. Standard: Affordable options with good services and facilities.
4. Budget: The most economical choice, offering basic services and accommodations.

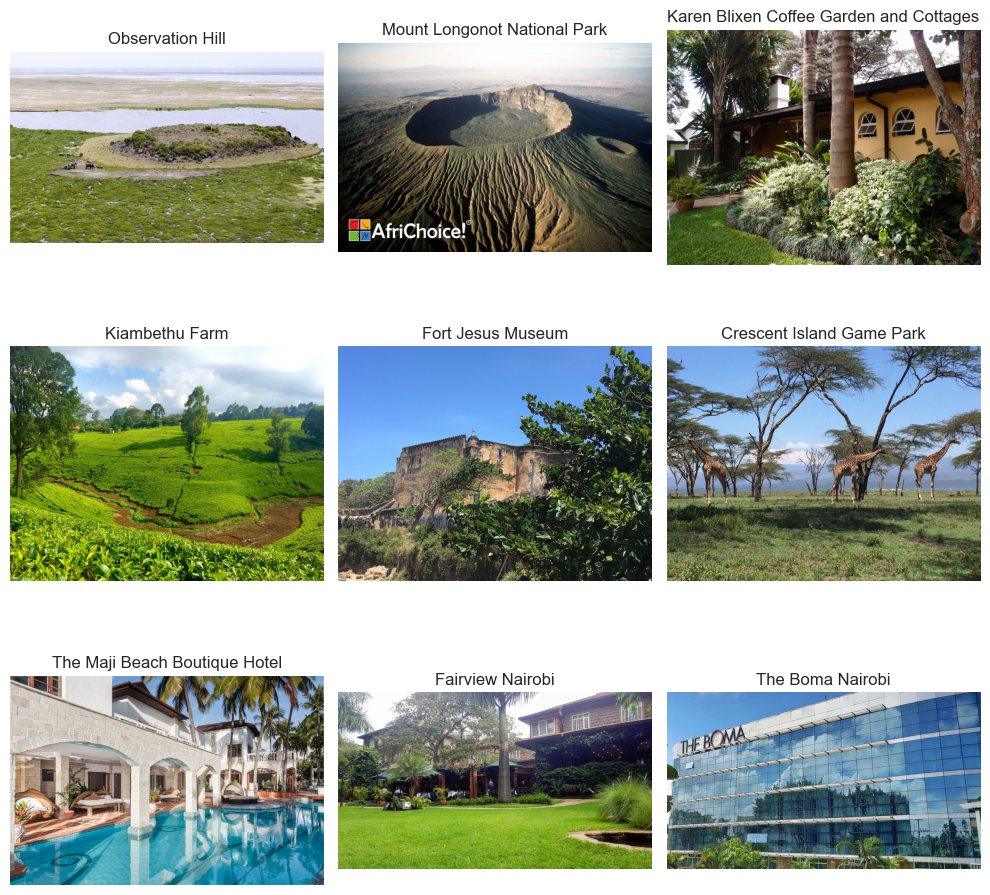
Feature Engineering

We created a new column weighted\_sentiment

***EDA***

***Visual of the Destinations***

Destinations



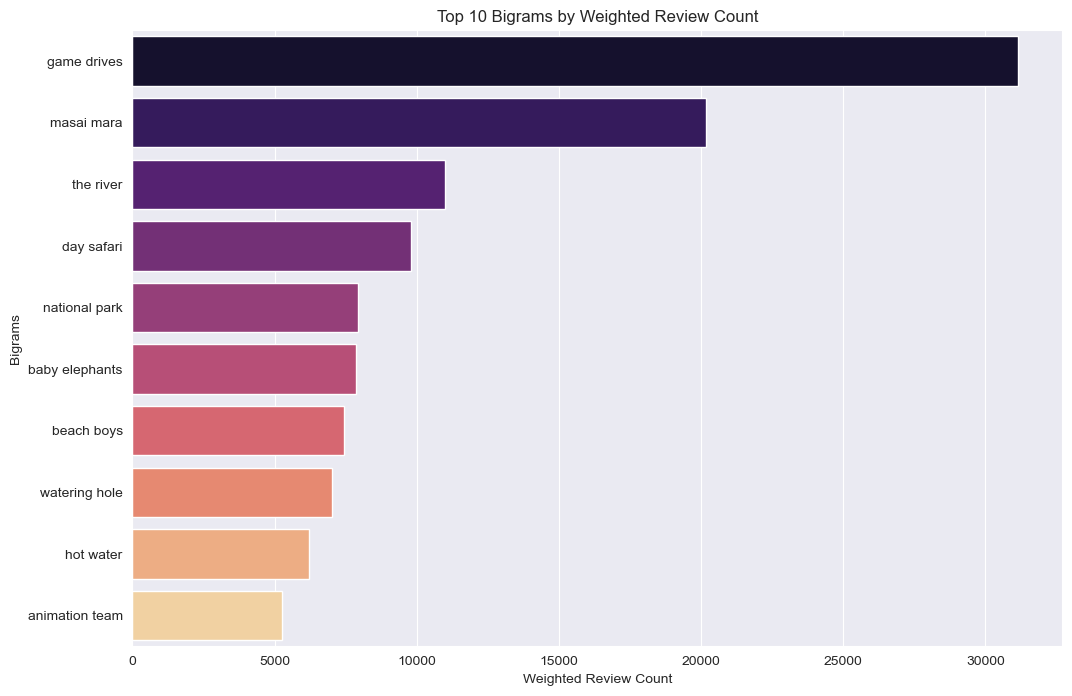
**Summary of Key Themes**

* Safari and Wildlife: The dominant theme is wildlife experiences, highlighted by terms like "safari," "wildlife," "animals," "game drive," and specific animals such as lion, elephant, and giraffe. This suggests a strong emphasis on safari and wildlife activities.
* Accommodation and Service: Key terms include "lodge," "camp," "hotel," "service," and "staff," reflecting a focus on guest accommodation and overall service quality.
* Location: Geographic terms such as "Masai Mara," "Lake Naivasha," "Kenya," and "Nairobi" underscore the regional focus of the reviews.
* Positive Sentiment: Words like "amazing," "great," "beautiful," and "wonderful" point to an overall positive sentiment towards the reviewed experiences.

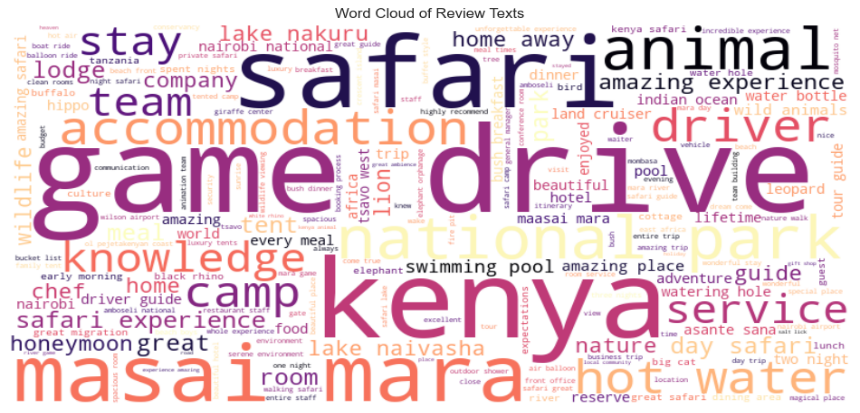
**Overall Sentiment**

* **Positive Sentiment:** Words like "amazing," "great," "beautiful," and "enjoyable" are prevalent across multiple word clouds, suggesting overall positive sentiment towards the locations.
* **Nature and Wildlife:** Words associated with nature, wildlife, and outdoor activities (e.g., "park," "safari," "wildlife," "forest") are prominent, indicating a strong focus on natural experiences.
* **Historical and Cultural Aspects:** Words like “slave,” “trade,” and “colonialism” may point to historical or cultural themes.The presence of “interesting,” “history,” and “culture” suggests educational or informative content.

**VISUAL 1**



Extract the main bigram that describes the destinations



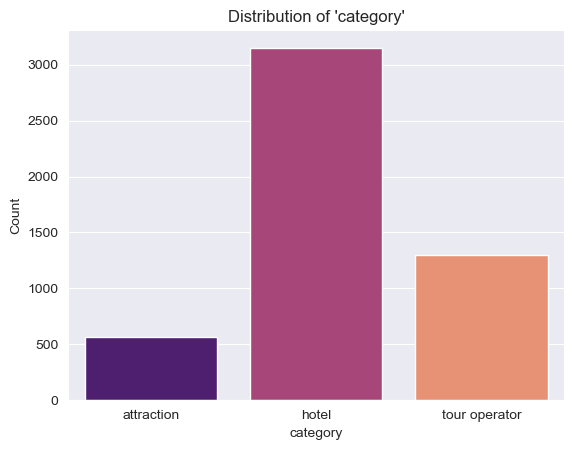
VISUAL 2

**Accommodation and Service:** Key terms include **lodge,** **camp**, **hotel**, **service**, and **staff**.

Location: Geographic terms such as **Masai Mara, Lake Naivasha, Kenya, and Nairobi** underscore the regional focus of the reviews.

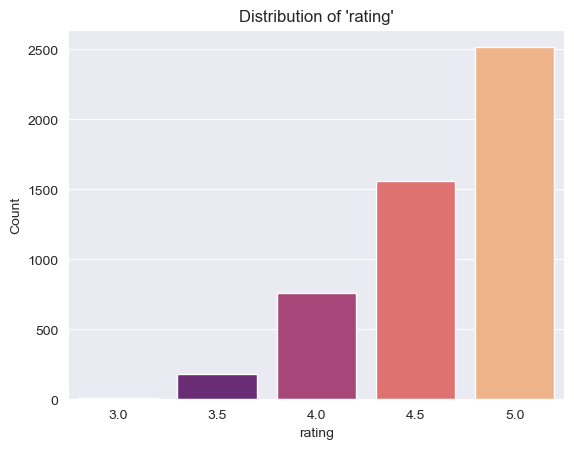
Positive Sentiment: Words like **amazing, great, beautiful, and wonderful.**

VISUAL 3



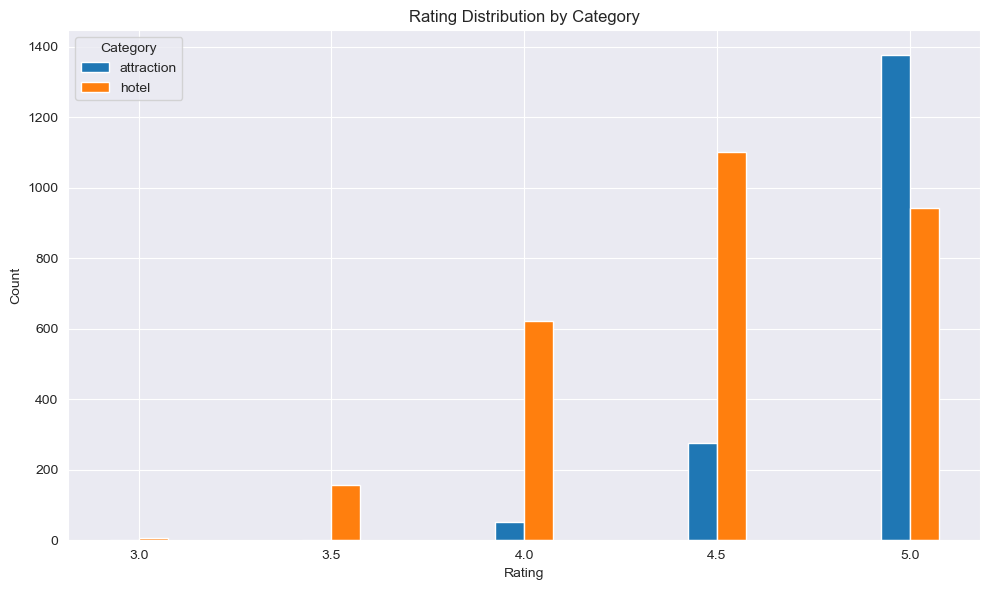
The most dominant category is the Hotel Category

VISUAL 4



Most tourist destinations are rated at 5.0

VISUAL 5



• Attractions generally received higher ratings, with a larger number of ratings at 4.5 and 5.0 compared to hotels. This suggests that customers are more satisfied with attractions.

• Bimodal Distribution for Hotels: Hotel ratings show a bimodal distribution with peaks around 3.5 and 4.5. This indicates a split in customer experiences, with some customers being highly satisfied while others are less so.

**Summary**

The attraction category lacks price data and this presents a challenge. We will handle the issue by creating separate models for each category.

We will build two different recommendation models: one for attractions and one for hotels. Since attractions don’t have price data, we will exclude priceLevel from the model for attractions and focus on other features like rating and reviews.

This approach will allow tailoring the recommendation model specifically to the available data for each category.

***Modeling***

Models Used:

**1. Model 1 – KNN**

**Tuning:**

* **CountVectorizer**
* **Hstack**
* **StandardScalar**

**Results**

The MSE of 0.610 indicates that the squared differences between predicted and actual ratings are relatively small. The RMSE of 0.781 suggests that, on average, our predictions are about 0.781 units away from the actual ratings. MAE of 0.711 shows that, on average, our predictions deviate by 0.711 units from the true ratings, indicating overall good model performance.

**2. Model 2 - Cosine Similarity**

**Distance Metrics:**

**Average Mean Squared Error (MSE): 0.2101912092719671**

**Average Root Mean Squared Error (RMSE): 0.45846614844715317**

**Average Mean Absolute Error (MAE): 0.3022205805532966**

**Recommended destinations**

**Kenyatta International Conference Center, Modan Rent A Car, Bamburi Beach,** **Fort Jesus Museum,** Mara River

**3.Model 3 - SVM**

### 4.Tuned KNN

**Distance Metrics:**

**Average Mean Squared Error (MSE): 0.3309639568619343**

**Average Root Mean Squared Error (RMSE): 0.5752946695928395**

**Average Mean Absolute Error (MAE): 0.41681208879029213**

**Overall Summary**

* **Both KNN & Cosine Similarity models have their strengths and weaknesses. The KNN model seems to perform better in terms of precision, while the cosine similarity model has a lower RMSE.**
* **SVM performed worse in all aspects.**
* The SVM model is performing well in predicting "Not Similar" attractions but is failing to identify "Similar" ones.
* The KNN model exhibits lower error metrics (MSE, RMSE, MAE) compared to the Cosine Similarity model. This suggests that the KNN model is better at minimizing errors when predicting similarity between hotels.
* Both models show similar performance in terms of classification, particularly for the "Similar" class. However, the KNN model slightly outperforms in F1-score for the "Similar" class, making it more reliable in identifying similar hotels.
* The recommendations provided by the KNN model seem more varied in price level and are closer in terms of distance (e.g., distances range between 0.800 and 0.898).
* The recommendations by the Cosine Similarity model seem to favor hotels with slightly different price levels and generally lower similarity scores. The distance metrics are not as consistent.
* The KNN model offers recommendations that are more consistent in terms of proximity to the target hotel, which is an essential factor in making relevant recommendations.

The KNN model appears to be the more suitable option for recommending hotels.

**Conclusion**

**The KNN model stands out as the most reliable option for providing recommendations, especially in terms of consistency and lower error metrics. It also balances precision and recall well, making it more suitable for real-world applications. The SVM model may excel in specific scenarios, particularly where classification of "Not Similar" entities is more critical. However, it underperforms in recommending similar items compared to the KNN model.**

**Model Selection and Performance:**

**KNN Model: The K-Nearest Neighbors (KNN) model appears to be the most reliable across different recommendation tasks. It provides lower error metrics (MSE, RMSE, MAE) and achieves higher precision, recall, and F1-scores, particularly in identifying "Similar" items. This model is well-suited for tasks where minimizing prediction errors and ensuring the accuracy of similarity detection are critical.**

**Cosine Similarity Model: While the Cosine Similarity model also performs reasonably well, it has slightly higher error metrics compared to the KNN model. Its recommendations tend to have lower similarity scores, suggesting that it may not capture the similarity between items as effectively as KNN.**

**SVM Model: The Support Vector Machine (SVM) model consistently underperforms compared to KNN and Cosine Similarity, especially in predicting "Similar" items. This model might not be suitable for the recommendation tasks in this context.**

**Hotels: The KNN model is particularly effective for hotel recommendations, providing more consistent proximity-based recommendations that are varied in price level and distance.**

**Attractions: For attractions, both KNN and Cosine Similarity models have their strengths. KNN tends to provide slightly better precision and F1-scores, making it more reliable for recommending similar attractions.**

**Tour Operators: KNN also outperforms in recommending similar tour operators, offering varied yet relevant options. It provides more consistent and relevant recommendations based on proximity to the target operator's location.**

**The lower MSE, RMSE, and MAE values for the KNN model indicate that it is better at minimizing prediction errors. This is crucial for tasks that require precise similarity recommendations.**

**The higher precision, recall, and F1-scores for the "Similar" class across different recommendation tasks suggest that the KNN model is better at correctly identifying similar items, reducing false positives and negatives.**