**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).

***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.

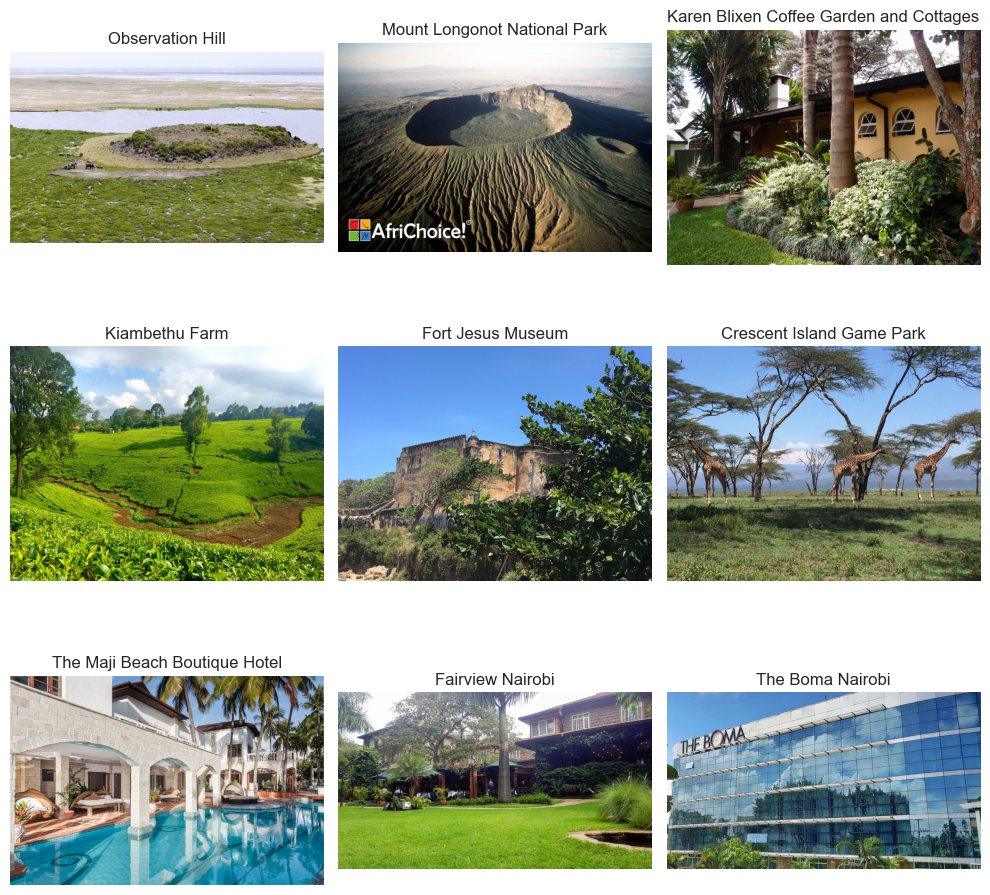
Feature Engineering

We created a new column weighted\_sentiment

***EDA***

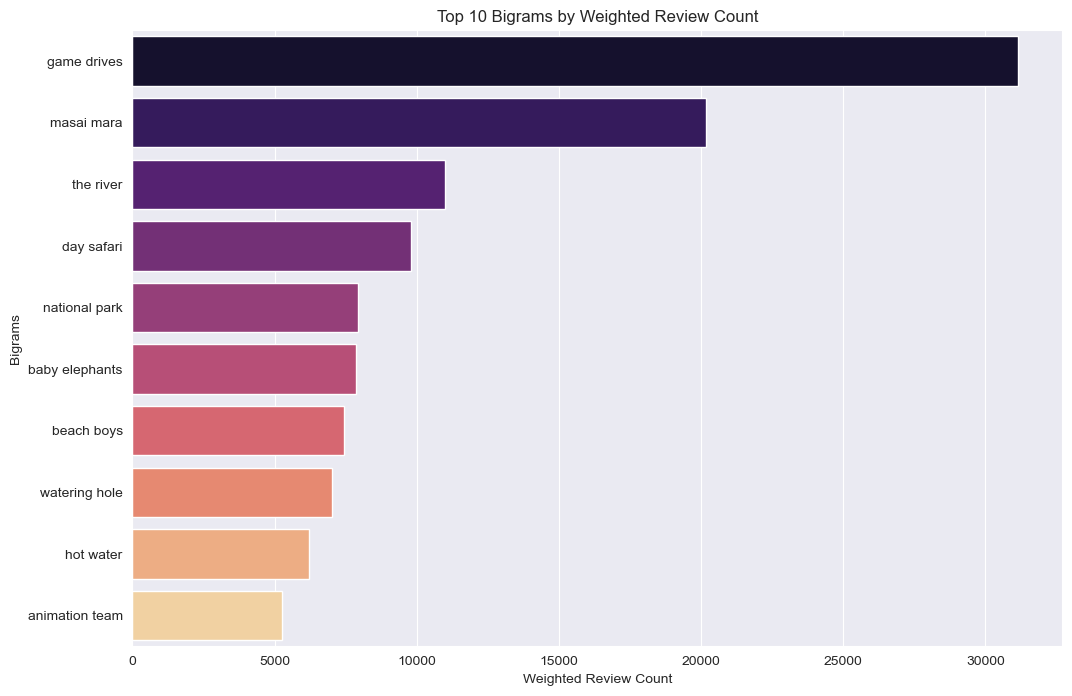
***Visual of the Destinations***

Destinations



#### *****Review Bigrams Visual*****

#### A visualization of the most common bigrams in the reviews, weighted by the number of reviews, can be useful for the tour recommendation system to find the most popular words used in the reviews.



Extract the main bigram that describes the destinations

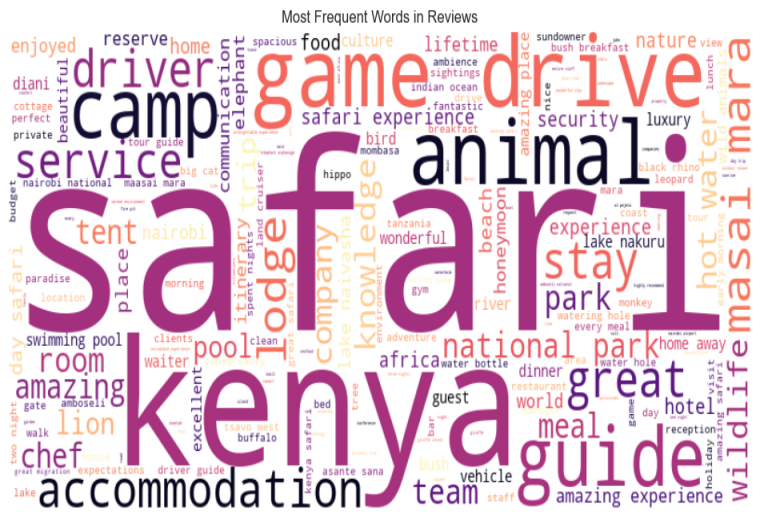
#### *****Word Cloud Visual of Reviews*****

****Safari and Wildlife:** safari, wildlife, animals, game drive,** and specific animals such as lion, elephant, and giraffe.

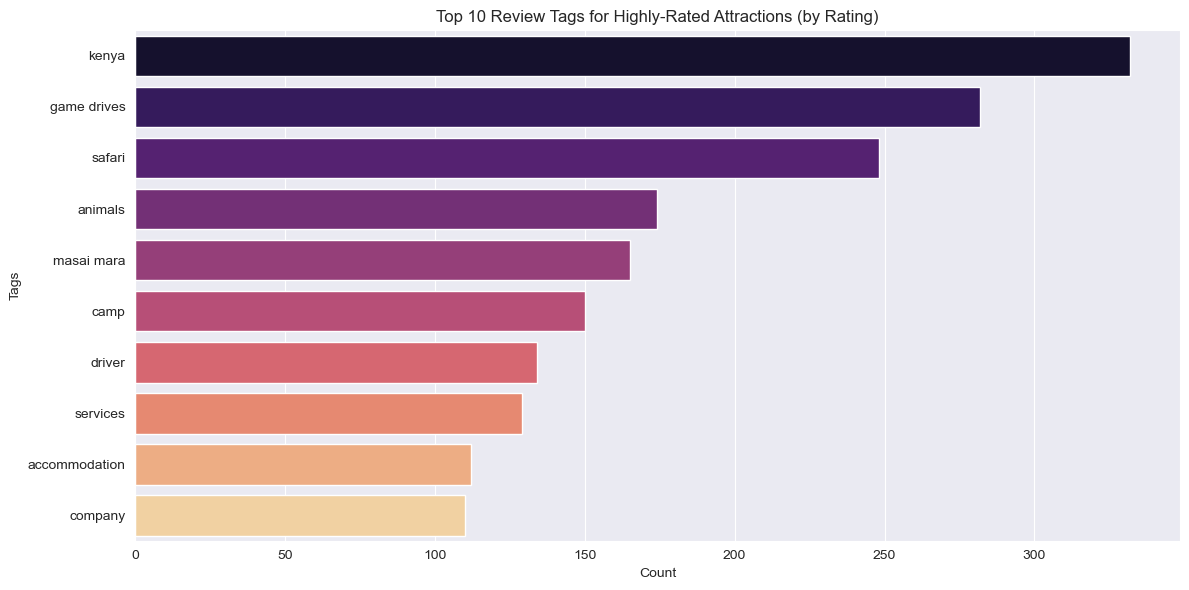
**Accommodation and Service:** Key terms include **lodge,** **camp**, **hotel**, **service**, and **staff**.

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.**

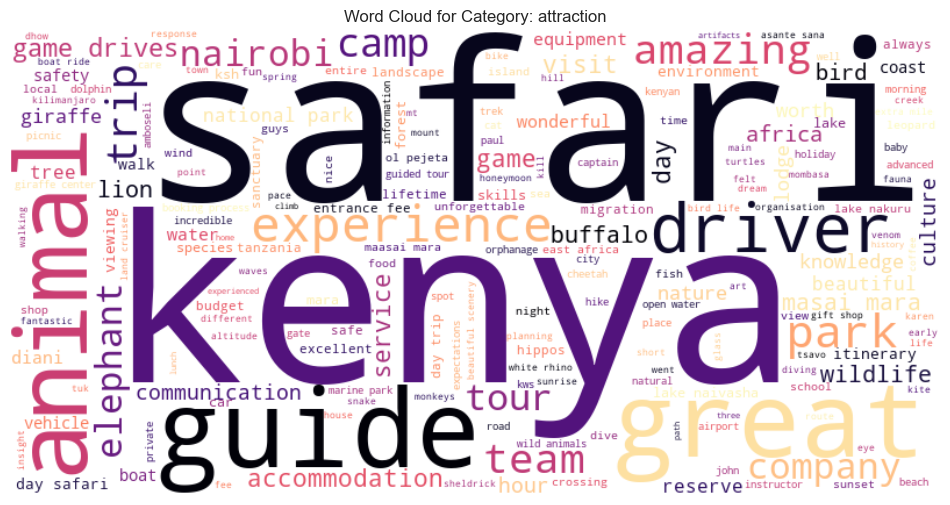


### *****What are the most common review tags associated with highly-rated attractions ?*****



most common review tags associated with highly-rated attractions

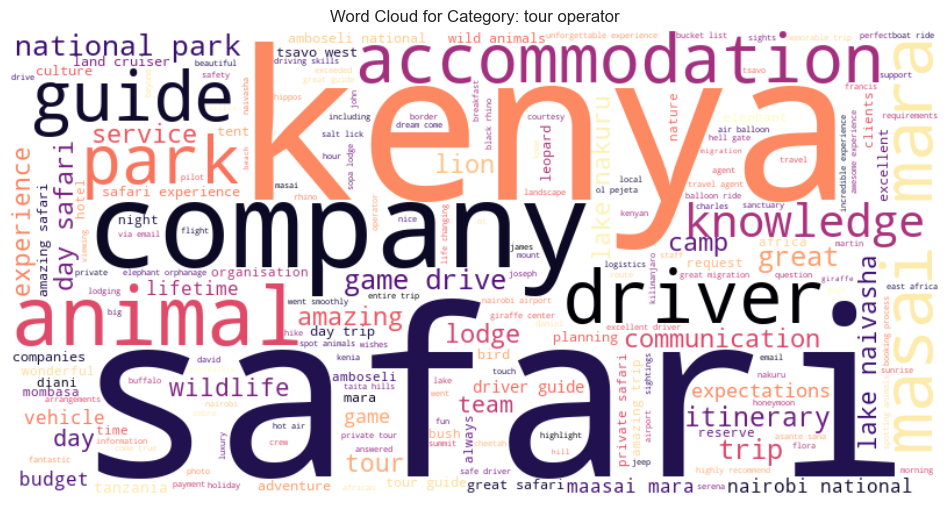
**Most frequently used words in customer reviews for each specific category**



**Category: Attraction**

**Dominant Words:** Safari, Kenya, Guide, Park, Driver, Animal, Experience, Great, National Park, Company





#### ***Word Cloud Insights***

##### Category: **Attraction**

****Dominant Words:**** Safari, Kenya, Guide, Park, Driver, Animal, Experience, Great, National Park, Company

**Key Insights:**

* Safari is the primary attraction, with Kenya being a strong associated term.
* Guided experiences and park visits are crucial aspects.
* The presence of "driver" and "company" suggests organized tours are popular.
* Positive sentiment is indicated by words like "experience" and "great."

**Category: "Hotel"**

**Dominant Words:** Camp, Game, Stay, Food, Animal, Lodge, Pool, Room Service, Hot Water, Tent

**Key Insights:**

* This category focuses on accommodation and facilities within a safari setting.
* Camping and lodges are popular options.
* Amenities like food, pool, and hot water are valued.
* The word "animal" suggests proximity to wildlife is important.

**Category: "Tour Operator"**

**Dominant Words:** Company, Safari, Guide, Team, Knowledge, Masai Mara, Itinerary, Communication, Driver, Animal

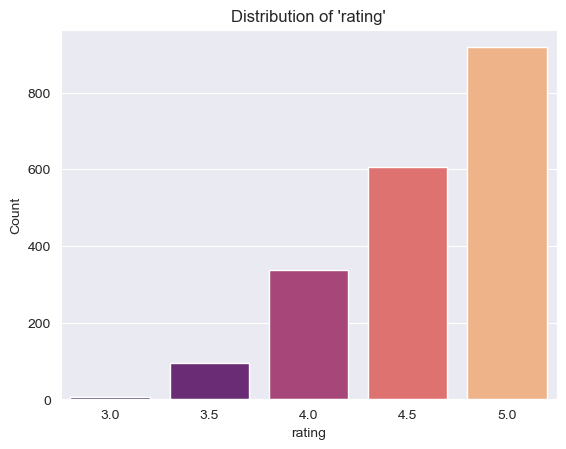
**Key Insights:**

* This category emphasizes the role of the operator in providing a safari experience.
* Expertise and knowledge are important attributes of the operator.
* The presence of "Masai Mara" and "itinerary" suggests specific destinations and planned trips.
* Good communication and a strong team are valued.

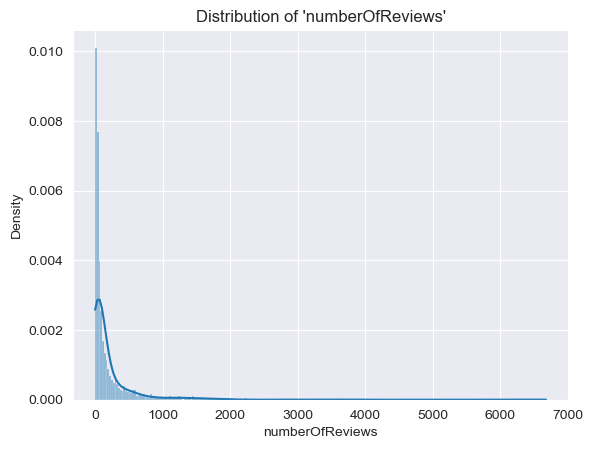
**Overall Insights:**

* Safari in Kenya is the core attraction.
* Guided tours and organized experiences are preferred.
* Accommodation options range from camping to lodges with amenities.
* Operators play a crucial role in providing a successful safari experience.
* Positive sentiment towards the overall experience is evident.

***Univariate Analysis***

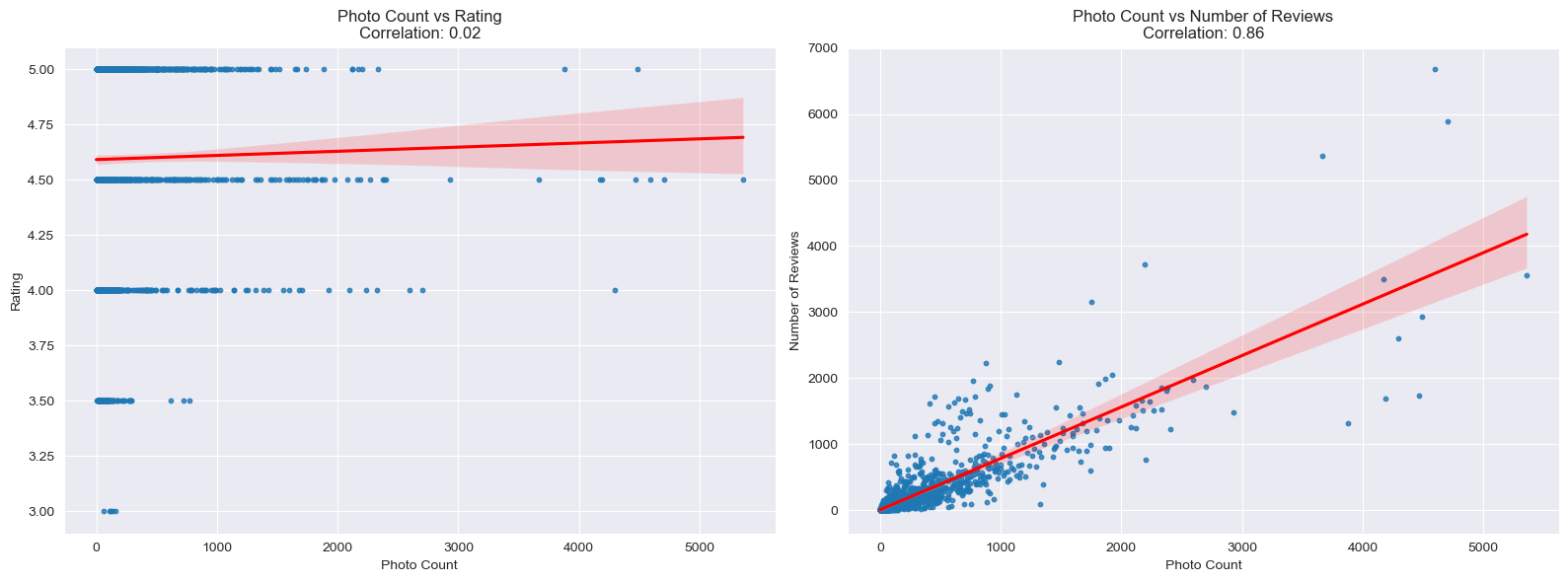


**Rating Distribution:** The distribution is heavily skewed towards higher ratings, with a sharp peak around 4.5 and a long tail towards lower ratings. This suggests that while most reviews are positive, there's still a significant portion with lower ratings.



**Number of Reviews Distribution:** The distribution is skewed, with most businesses having a lower number of reviews and a few outliers with a large number of reviews, indicating a few businesses dominate in visibility.

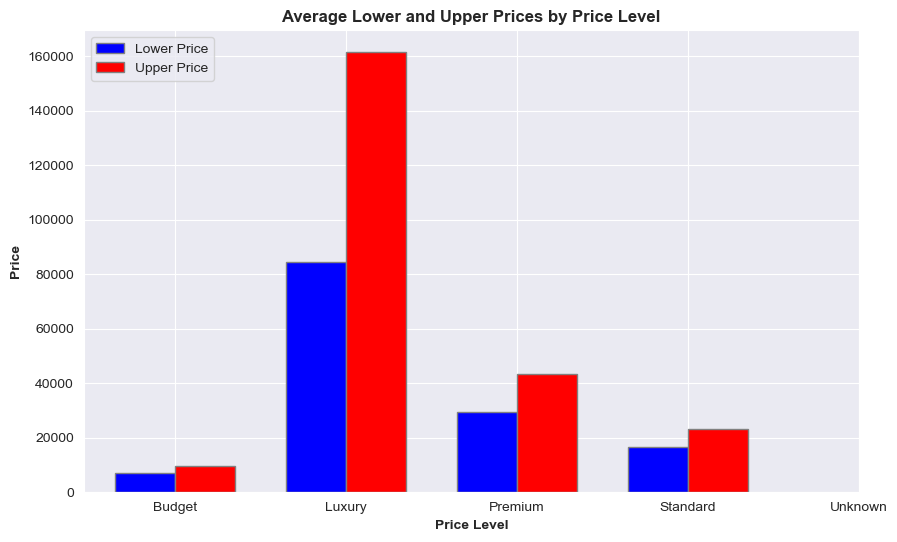
***Correlation Analysis***



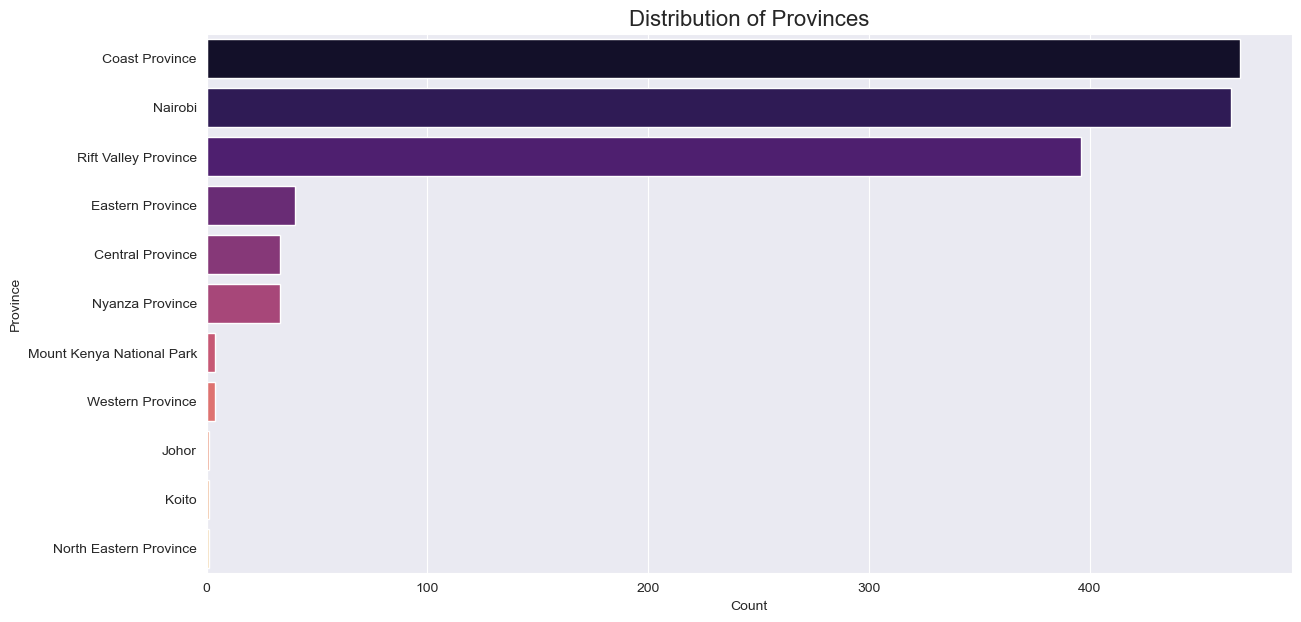
Correlation Insights:

* Photo Count and Number of Reviews: Have a strong positive correlation (0.86), suggesting that restaurants with more reviews tend to have more photos.
* Photo Count and Rating: The quality of the photos, rather than the quantity, likely has a greater impact on the rating. Focus on uploading high-quality, relevant images to showcase the listing effectively.
* Photo Count and Reviews: More photos can attract attention and encourage users to engage with the listing, leading to more reviews.

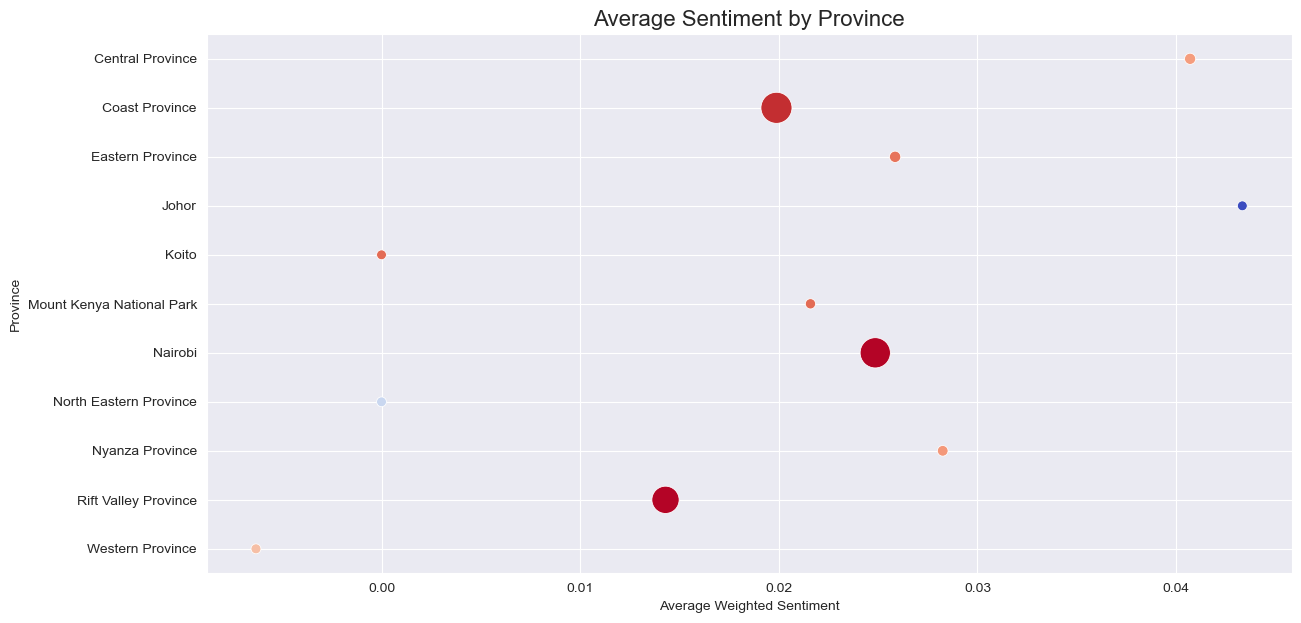
***How do the lower and upper price ranges vary across different price levels***

****

***Which provinces have the most listings***



***How does the average sentiment vary across different provinces?***

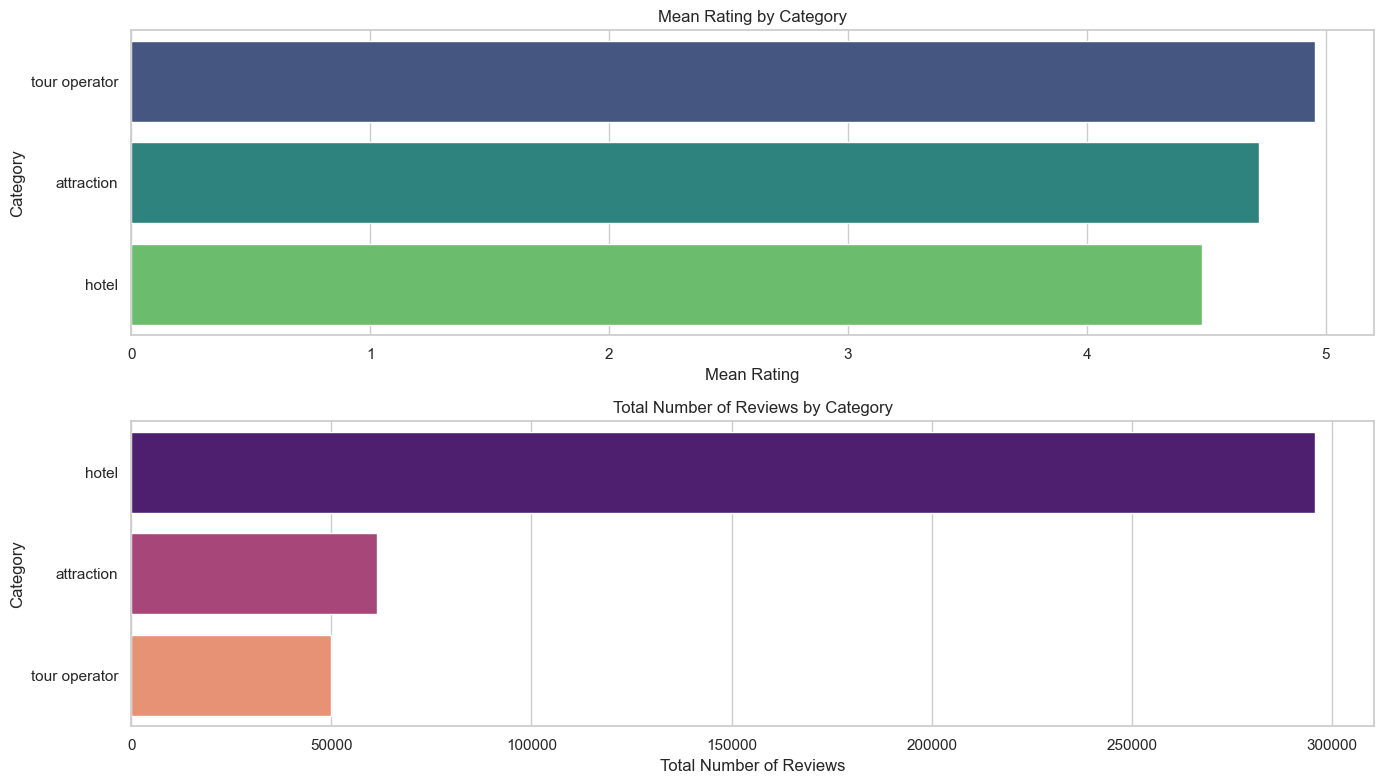


* **High Sentiment:** A few provinces stand out with notably higher sentiment scores:
  + Nairobi
  + Rift Valley Province
  + Coast Province

This could indicate factors contributing to positive sentiment

* **Low Sentiment:** Several provinces exhibit lower sentiment scores, with some clustering near the 0.00 mark:
  + Central Province
  + Western Province
  + North Eastern Province

***How do average ratings and total number of reviews vary across different categories?***



**Tour Operators Shine**

* Tour operators have the highest average rating, indicating strong customer satisfaction with their services.
* However, they have the lowest number of reviews, which could suggest less exposure or a smaller tourists base compared to other categories.

**Hotels Dominate Reviews**

* Hotels receive the most reviews, making them the most reviewed category.
* This could be due to a larger number of hotels or a higher tendency for tourists to leave reviews for hotel experiences.

**Attraction Satisfaction**

* Attractions fall in the middle for both average rating and total number of reviews.
* This suggests a balance between tourists satisfaction and review volume.

### ****What is the average price range for different categories?****

### 

***EDA Summary***

The attraction & tour operator categories lack price data and this presents a challenge. We will handle the issue by creating separate models for each category.

We will build 3 different recommendation models: one for each category. Since attractions & most tour operators don’t have price data, we will exclude priceLevel from the models and focus on other features like rating, reviews and sentiments.

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

***Modeling***

## *****Preprocessing*****

## **Encode the location, province and category columns so as to fit them into the models.**

* Split the data into categories

The data is split so as to work with the different requirements. This is to avoid dropping rows that don't have price values which would be helpful for the hotel category.

* Drop null values

***Hotel predictor***

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.

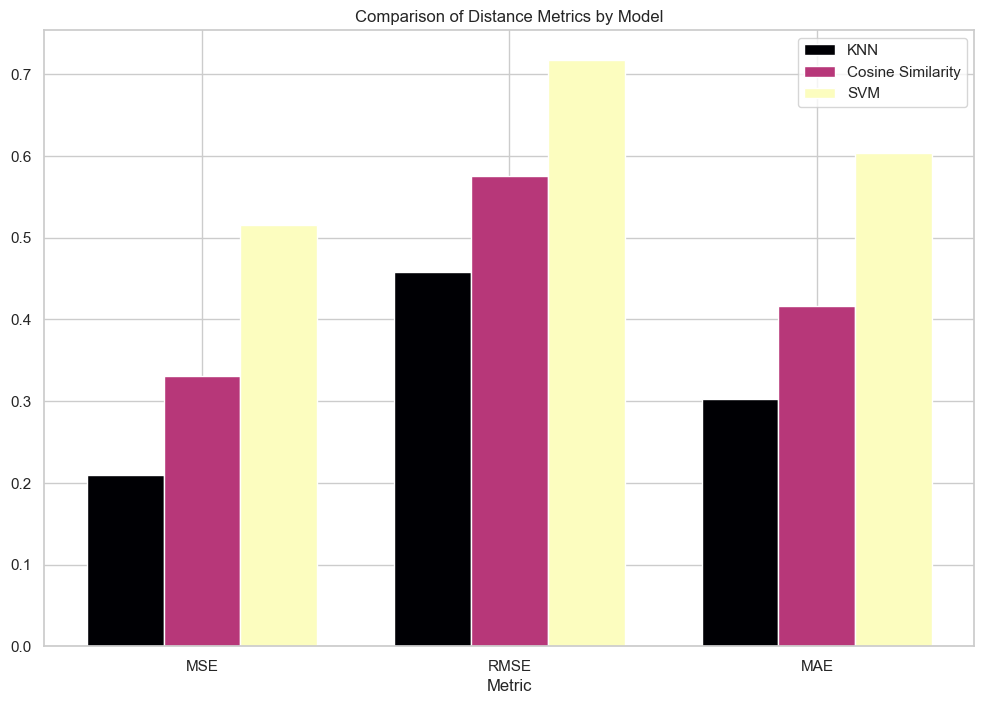
***Tour Operator predictor***

***Models used***

#### Tuned KNN

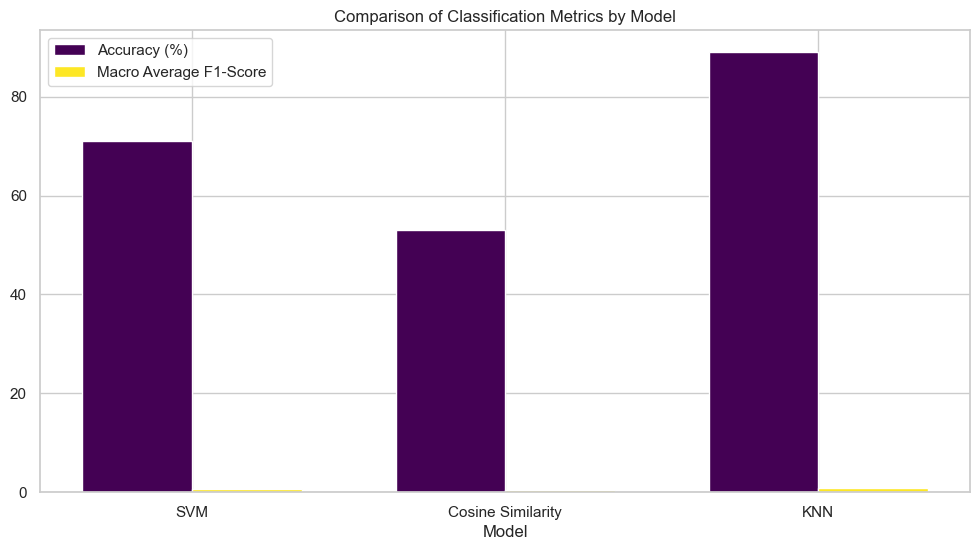
1. Cosine similarity

***Model Evaluation***  
We'll create a bar plot to compare the MSE, RMSE, and MAE for the KNN, Cosine Similarity, and SVM models.



#### ***Bar Plot for Classification Metrics***

We'll create a bar plot to compare the Accuracy and Macro Average F1-Score for the KNN, Cosine Similarity, and SVM models.



The KNN model performed best with an accuracy of 89% and a Macro Average F1-Score 77%

The Cosine Similarity model is the second best model with an Accuracy of 71% and a Macro Average F1-Score 70%

**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.

**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.

***Recommendations***

1. Model Tuning: Continue fine-tuning the KNN model to further reduce errors and improve classification metrics, particularly for the "Similar" class. This could involve adjusting the number of neighbors or using weighted distances.
2. Hybrid Model Approach: Consider combining the strengths of both KNN and Cosine Similarity models in a hybrid approach, where one model is used for initial filtering and the other for fine-tuning recommendations.
3. Develop and integrate advanced recommendation algorithms that curate personalized travel itineraries based on individual preferences, historical travel data, and real-time user inputs. This will enhance the relevance and appeal of suggested destinations and activities.
4. Provide real-time updates on local events, weather conditions, and special offers relevant to the traveler’s current or upcoming location. This can be achieved through push notifications or in-app alerts, ensuring that travelers have the most up-to-date information.
5. Consider expanding the recommendation system to include data and insights from other countries. This will allow for a broader range of travel options and cross-country promotional opportunities.