### **REPORT**

### 

### **ON**

### 

### **TOURISM EXPERIENCE ANALYTICS:**

### 

### **CLASSIFICATION,PREDICTION**

### 

### **AND**

### 

### **RECOMMENDATION SYSTEM**

### 

### 

### 

### 

### 

### **By:**

### **Gayathri S K**

**Introduction**

In today's data-driven tourism industry, platforms and agencies aim to enhance traveler experiences by providing personalized recommendations, predicting user satisfaction, and anticipating visitor behavior. This project, **Tourism Experience Analytics**, focuses on analyzing user preferences, travel patterns, and attraction features to achieve three primary goals:

* **Regression:** Predicting user ratings for attractions.
* **Classification:** Identifying the likely visit mode (e.g., Family, Business).
* **Recommendation:** Suggesting tourist attractions tailored to user profiles.

### **💼 Business Use Cases:**

1. **Personalized Recommendations**: Suggest attractions based on historical preferences and demographics.
2. **Tourism Analytics**: Analyze trends in popular regions and attraction types to inform strategy.
3. **Customer Segmentation**: Classify users for targeted marketing.
4. **Customer Retention**: Enhance loyalty with data-driven suggest

**Dataset Overview**

| **Column Name** | **Description** |
| --- | --- |
| UserId | Unique identifier of the user |
| VisitYear, VisitMonth | Time of visit to the attraction |
| AttractionId | Unique ID of the tourist attraction |
| Rating | Rating given by the user (1 to 5 scale) |
| AttractionTypeId | Type of attraction (e.g., beaches, ruins, parks) |
| Attraction | Name of the attraction |
| Continent, Region, CityName, Country | User's geographic location |
| VisitMode | Purpose of visit (Family, Business, Friends, etc.) |
| AttractionType | Description of the attraction category |
| User\_Visit\_Count | Number of visits made by the user |
| Avg\_User\_Rating | User’s average rating across attractions |
| Attraction\_Popularity | Average rating of the attraction across all visitors |

### 

### 

### 

### 

### 

### **Methodology**

### **1. Data Import and Integration**

Nine data tables were imported:

* **transaction, user, item, type, visit mode, city, country, region, continent**

Merged step-by-step using pandas:

* Step A: user + continent on ContinentId
* Step B: Result A + region on RegionId
* Step C: Result B + city on CityId
* Step D: Result C + country on CountryId
* Step E: transaction + item on AttractionId
* Step F: Result E + Result D on UserId
* Step G: Join with visit\_mode by matching VisitMode and VisitModeId
* Final merge with type on AttractionTypeId

### **2. Data Cleaning**

* Removed insignificant columns:  
   'TransactionId', 'AttractionCityId', 'ContinentId', 'RegionId', 'CountryId', 'CityId', 'VisitModeId'
* Standardized and validated:  
  + Null values treated
  + Fixed 3,722 duplicate records
  + Ensured rating consistency

### **3. Feature Engineering**

Created meaningful new variables:

* User\_Visit\_Count: Number of visits made by the user
* Avg\_User\_Rating: User's average rating given
* Attraction\_Popularity: Average rating per attraction

### **4. Exploratory Data Analysis (EDA)**

#### **A. Data Quality Checks:**

* Structure, nulls, duplicates, unique values, rating inconsistencies

#### **B. Visual Analysis:**

* User Demographics (continent, region, city)
* Mode of Visit Distribution
* Attraction Popularity and Type
* Rating Behavior Patterns
* Visit Trends by Year/Month

### **5. Predictive Modeling**

#### **A. Regression (Rating Prediction)**

* **Input Features**: User demographics, visit time, visit mode, attraction type, popularity
* **Models Tried**: Linear Regression, Decision Tree, Random Forest, XGBoost, CatBoost
* **Best Model**: **CatBoost** (highest accuracy)

#### **B. Classification (Visit Mode Prediction)**

* **Input Features**: Continent, Country, User\_Visit\_Count, Attraction\_Popularity
* **Models Tried**: Random Forest, XGBoost, CatBoost
* **Best Model**: **Random Forest Classifier**

#### **C. Attraction Recommendation**

* **Content-Based**: Matches attractions to user history using type, location, popularity
* **Collaborative Filtering**: Uses user-user similarity for recommendations

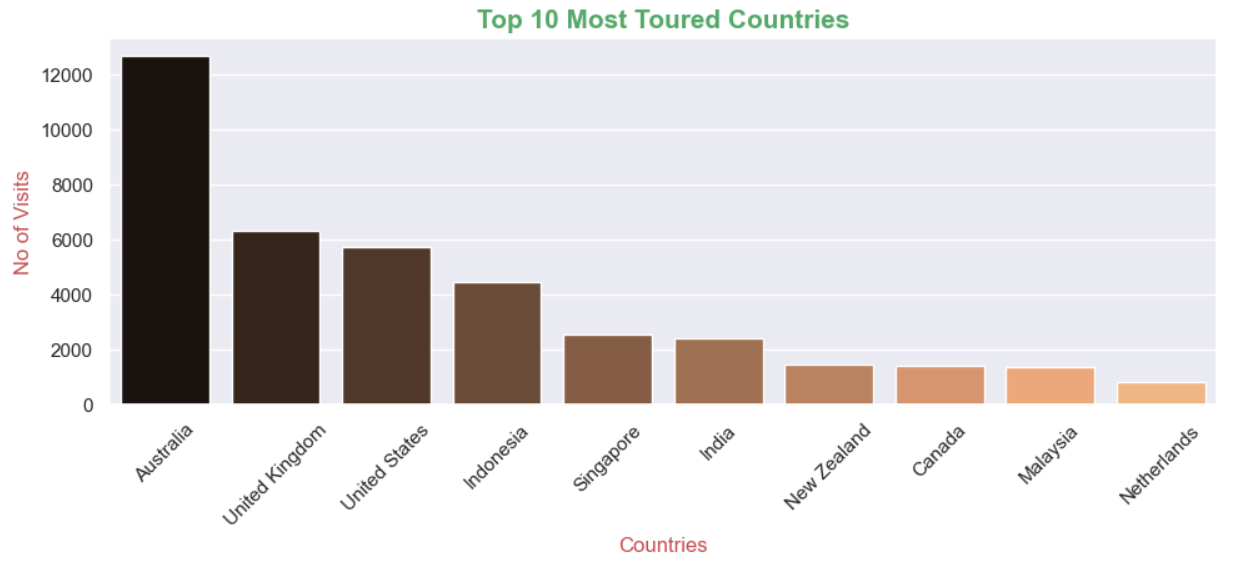
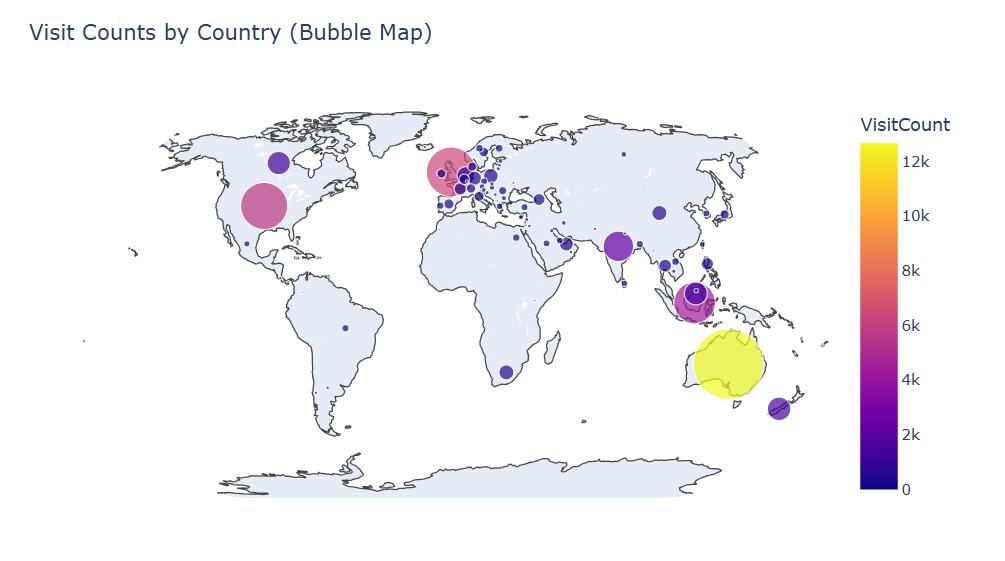
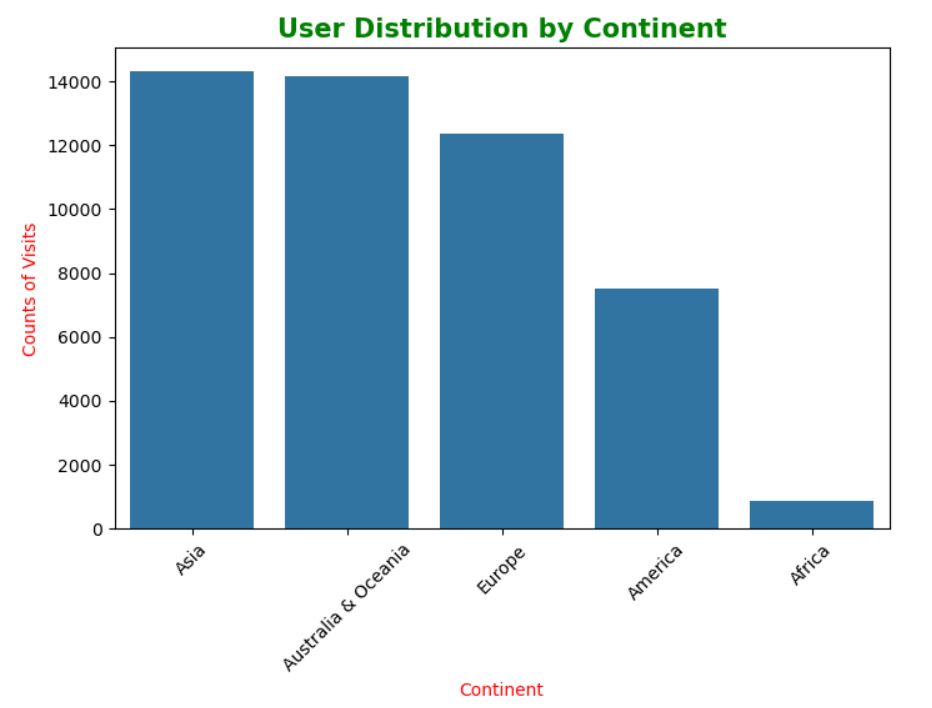
### **6. Deployment with Streamlit**

Developed an interactive **Streamlit web app** with:

* Visit mode predictor
* Attraction rating predictor
* Personalized recommendations

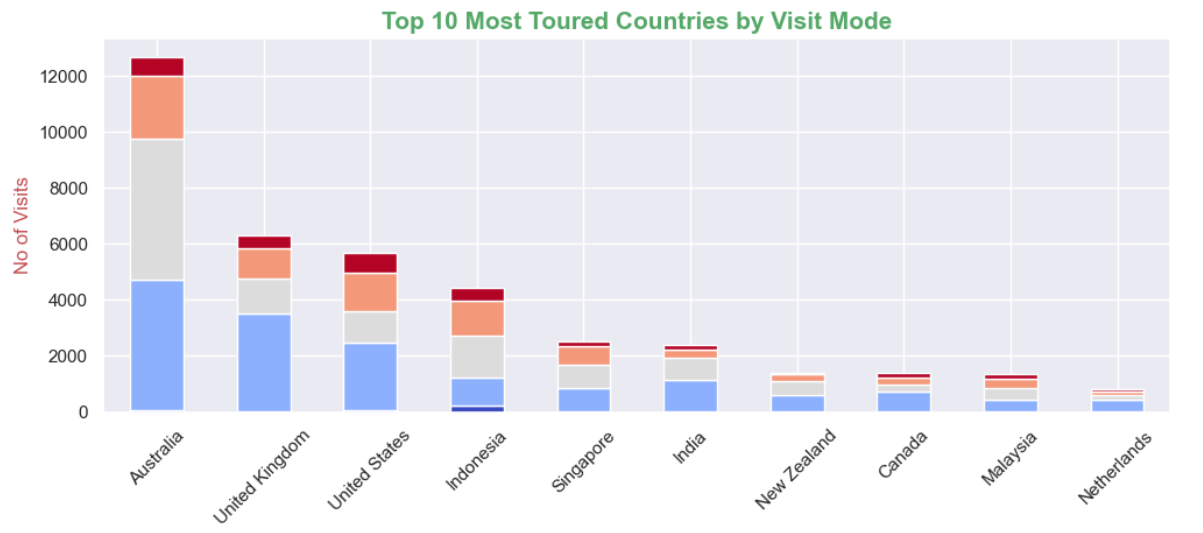
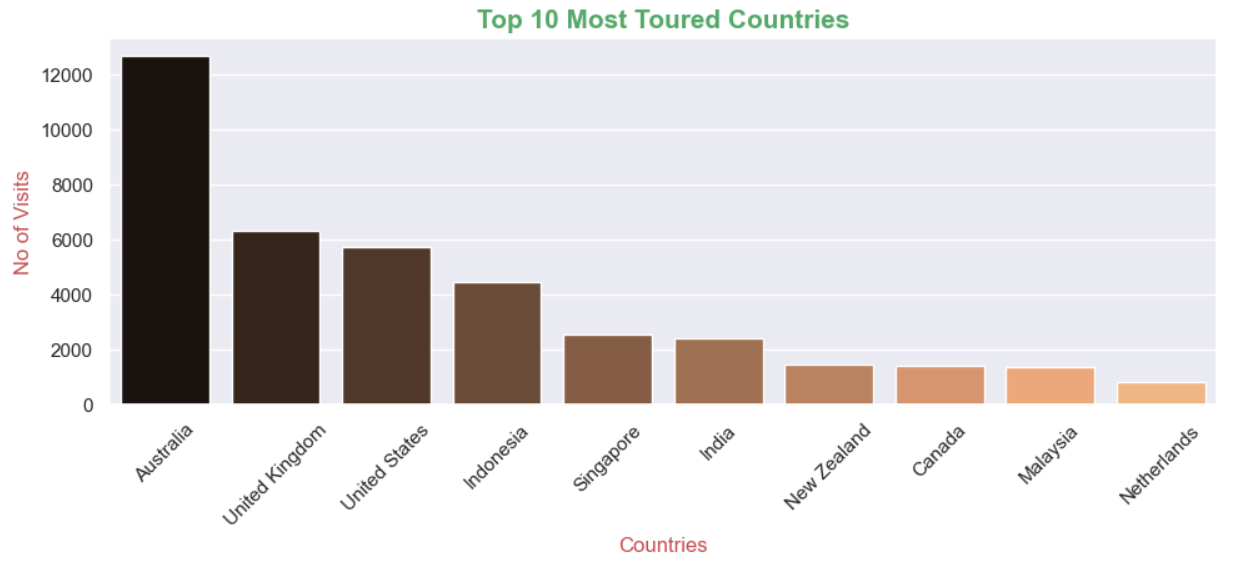
**Exploratory Data Analysis Insights**

### **🧑‍💼User Demographics Analysis – Key Insights**



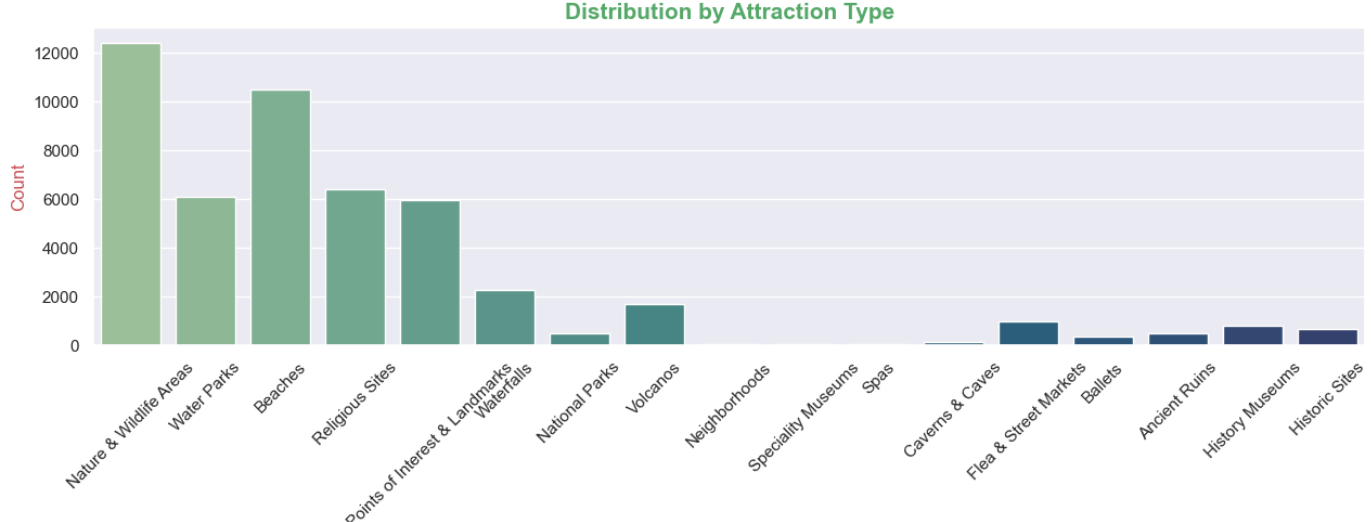
1. **Asia and Australia Are Top Tourism Continents** Asia and Australia attract the highest number of tourists, reflecting their rich cultural heritage and natural diversity. Their global appeal is driven by unique landscapes, vibrant traditions, and well-developed tourism infrastructure.
2. **Australia Leads as the Most Visited Country** Australia emerges as the top destination, likely due to its iconic attractions like the Great Barrier Reef, diverse wildlife, and traveler-friendly policies. The UK and US follow, favored for their cultural significance and established tourism sectors.
3. **Highly Rated Cities Reflect Tourism Excellence** Cities such as **Worrigee**, **Curepipe**, and **Kingscliff** received some of the highest ratings, indicating strong local efforts and effective tourism management that enhance visitor satisfaction.

### **🧳Mode of Visit Analysis – Key Insights**



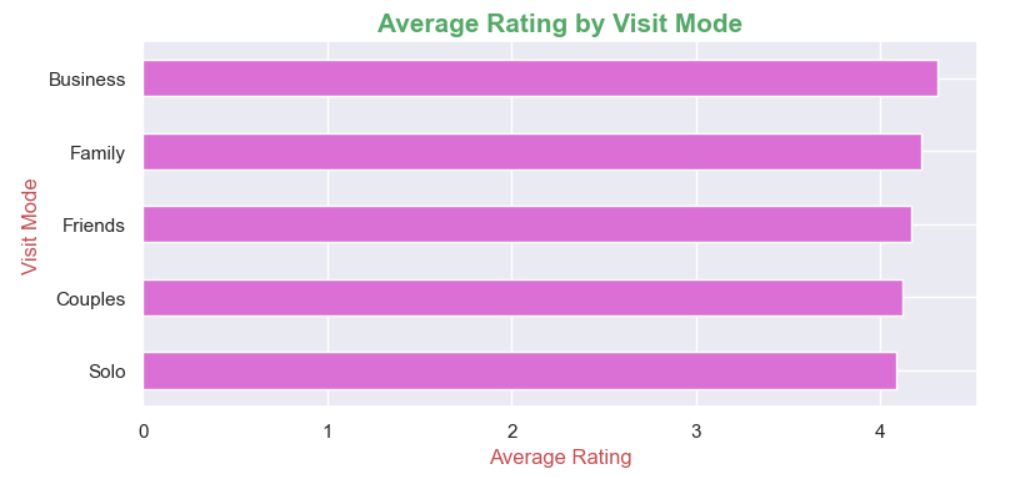
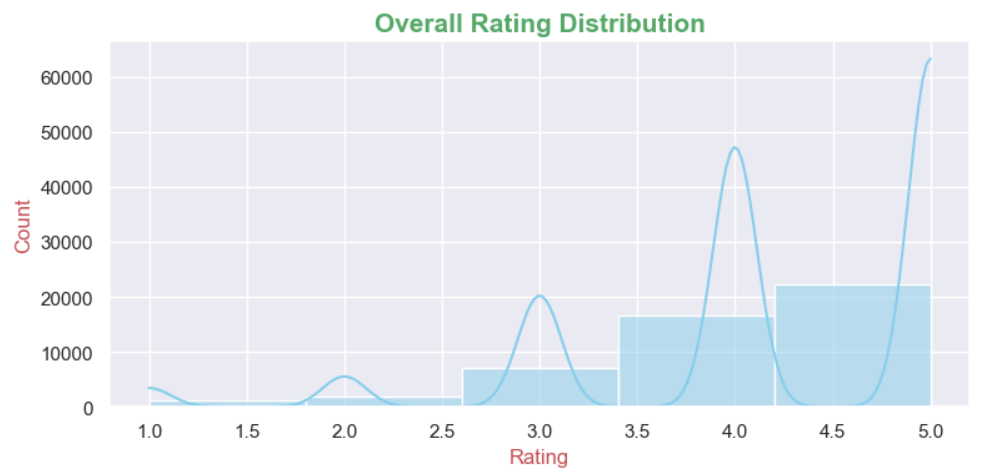
1. **Tour Purpose Varies by Country** Countries show different patterns of tourism. **Australia** attracts both **Couples** and **Families**, while **Indonesia** sees more **Business travelers**, likely due to its economic activity.
2. **Couples Prefer Western Destinations** **Couples** favor countries like the **US**, **UK**, and **Australia**, suggesting a preference for privacy, comfort, and well-developed tourism services.
3. **Business Travel Is Limited** **Business tourism** is relatively low overall, with **Indonesia** leading, indicating focused work-related travel with limited sightseeing.

### **🏖️ Attraction Analysis – Key Insights**



1. **Asian Cities Lead in Attraction Diversity** Cities like **Jakarta**, **Singapore**, and **Bali** have the **highest number of unique attraction types**, showcasing Asia’s broad appeal through nature, spirituality, and modern leisure experiences.
2. **Nature & Cultural Attractions Are Most Preferred** Tourists consistently favor **wildlife sanctuaries, beaches, waterparks, temples**, and **nature reserves**. This reflects a global interest in **scenic, relaxing, and culturally immersive destinations**.
3. **Top Attractions Dominate Tourist Flow** Attractions such as **Sacred Monkey Forest Sanctuary (12,368 visits)**, **Waterbom Bali**, and **Tegalalang Rice Terrace** top the charts in visit volume, indicating strong brand recognition and visitor satisfaction.
4. **Cities as Tourism Hubs** Popular cities tend to offer a variety of attractions catering to different visitor interests, making them strategic tourism hubs. This enhances their global visibility and repeat visits.

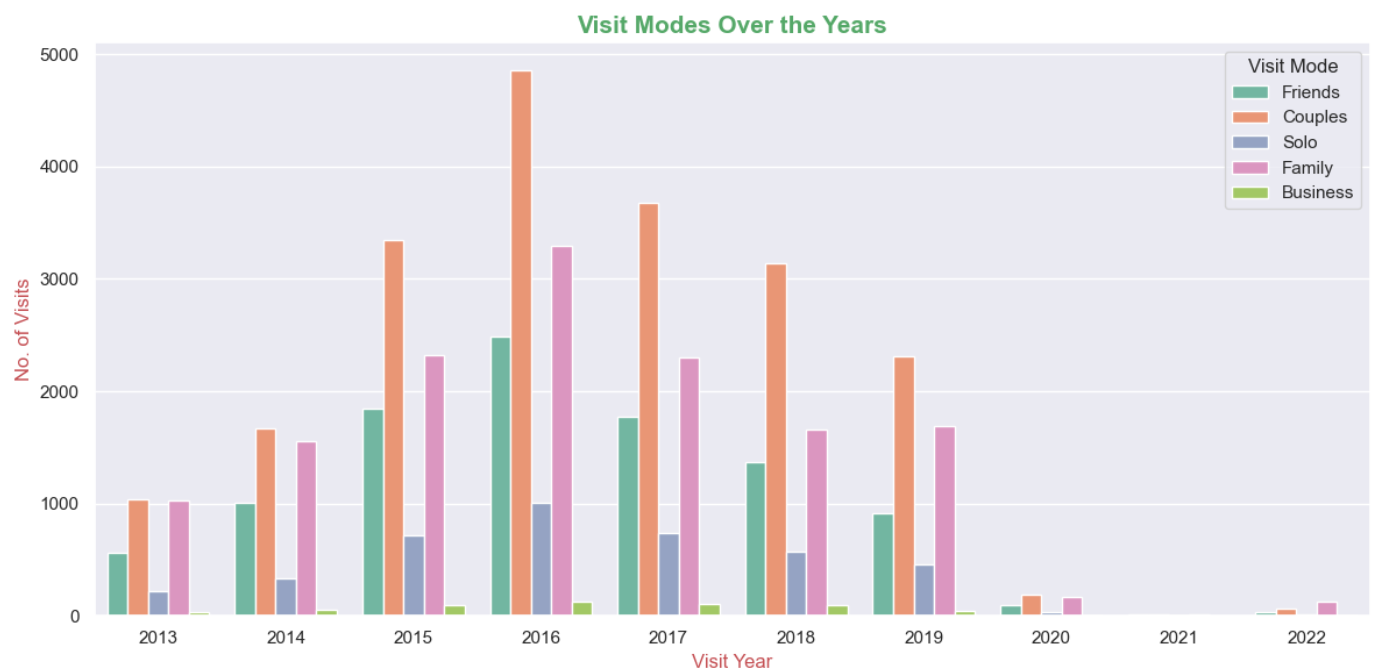
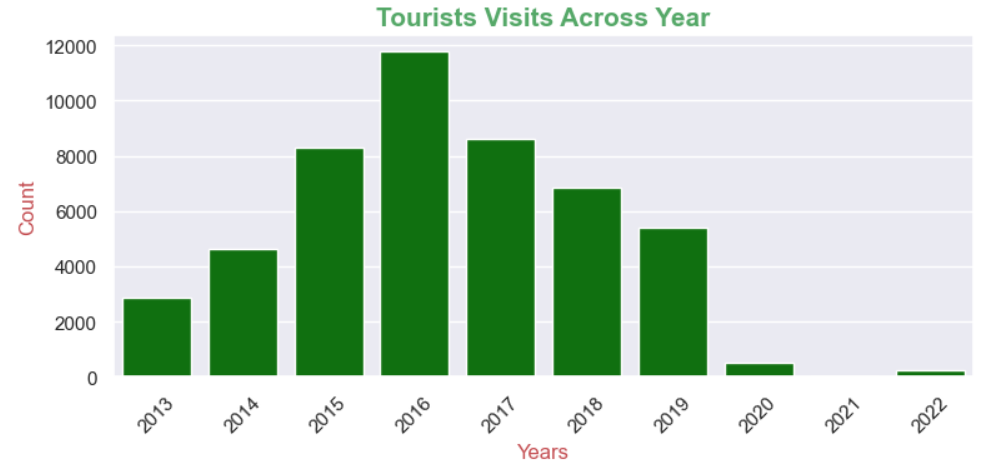
### **⭐ Rating Behavior – Key Insights**





1. **High Satisfaction Among Tourists** The rating distribution is skewed towards the higher end, with **5 being the most common rating**. This suggests strong satisfaction with aspects like hospitality, sightseeing experiences, and government support in maintaining tourist destinations.
2. **Business Travelers Rate Higher** Among different visit modes, **business travelers tend to give higher average ratings**, possibly due to structured experiences or well-supported facilities.
3. **Low-Rated Countries** Countries like **Cuba, Monaco, and Uruguay** receive relatively lower average ratings. This highlights potential gaps in tourism infrastructure or service quality, signaling areas where **local and government efforts could be improved** to boost tourist satisfaction and economic returns.

### **⭐ Visit Trends– Key Insights**



1. **Yearly Trend:**

Tourist visits show fluctuations over the years, with noticeable dips (e.g., 2020–2021 likely due to COVID-19) and recovery by 2022.

1. **Visit Modes:**Family and Friends are the dominant visit modes, indicating group travel preferences.

Business visits remain consistently low, suggesting the destination is primarily leisure-focused.

1. **Key Observation:**Post-pandemic recovery is evident, with 2022 visits nearing pre-2020 levels, reflecting restored travel confidence.

**Model Building & Evaluation**

### **1. Regression Model: Predicting Tourist Attraction Ratings**

**Objective**:  
 To predict the rating a user might assign to a tourist attraction based on their historical behavior, demographics, and attraction features.

**Encoding Techniques**:

* **Low-cardinality categorical features** (VisitMode, Continent, Region): Encoded using **Label Encoding**
* **High-cardinality features** (Country, CityName, AttractionType, Attraction): Encoded using **Target Encoding**

**Model Performance (R² Score)**:

| **Model** | **R² Score** |
| --- | --- |
| Linear Regression | 0.071 |
| Decision Tree | -0.012 |
| Random Forest | 0.062 |
| Tuned Random Forest | 0.088 |
| XGBoost Regressor | 0.061 |
| **CatBoost Regressor** | **0.1341** (Best) |

Despite CatBoost performing better than other models, **all regression models yielded low R² scores**, indicating weak predictive power for continuous rating prediction.

**Alternative Approach**:  
 To address low R², the problem was reframed as **rating classification**. A **CatBoost Classifier** was implemented to predict discrete rating classes instead of continuous values.

* **CatBoost Classifier Accuracy**: **47.73%** This model was selected for the final implementation in place of regression.

### **2. Classification Model: Predicting Mode of Visit**

**Objective**:  
 To classify the likely mode of visit (e.g., Business, Family, Friends, Couples) based on user profile and attraction data.

**Encoding Applied**:

* **Label Encoding**: Continent, Country

**Model Performance (Accuracy)**:

| **Model** | **Accuracy** |
| --- | --- |
| **Random Forest** | **47.89%** |
| XGBoost Classifier | 47.89% |
| CatBoost Classifier | 46.99% |

**Class Imbalance Handling**:

* **SMOTE** was applied to address class imbalance. However, it introduced noise and **reduced accuracy to 36%**. Therefore, SMOTE was discarded.

**Final Model Selected**: **Random Forest Classifier** (best accuracy, robust generalization)

### **3. Recommendation System: Suggesting Tourist Attractions**

**Objective**:  
 To recommend personalized attractions based on user behavior and attraction characteristics.

#### **A. Collaborative Filtering**

* Approach: Based on **user-user similarities**
* Implementation: **k-Nearest Neighbors (kNN)** used to avoid memory issues with cosine similarity
* Input: User's past ratings
* Output: Attractions liked by similar users

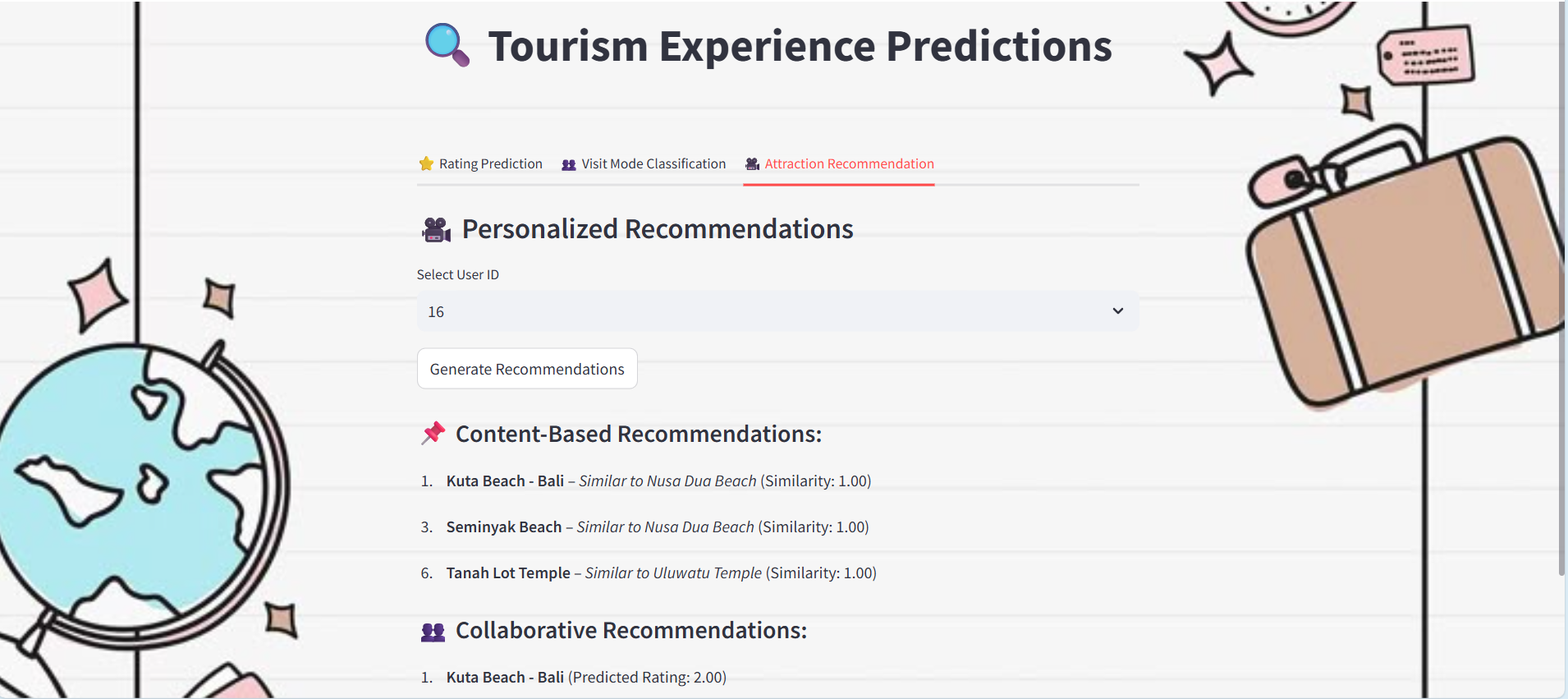
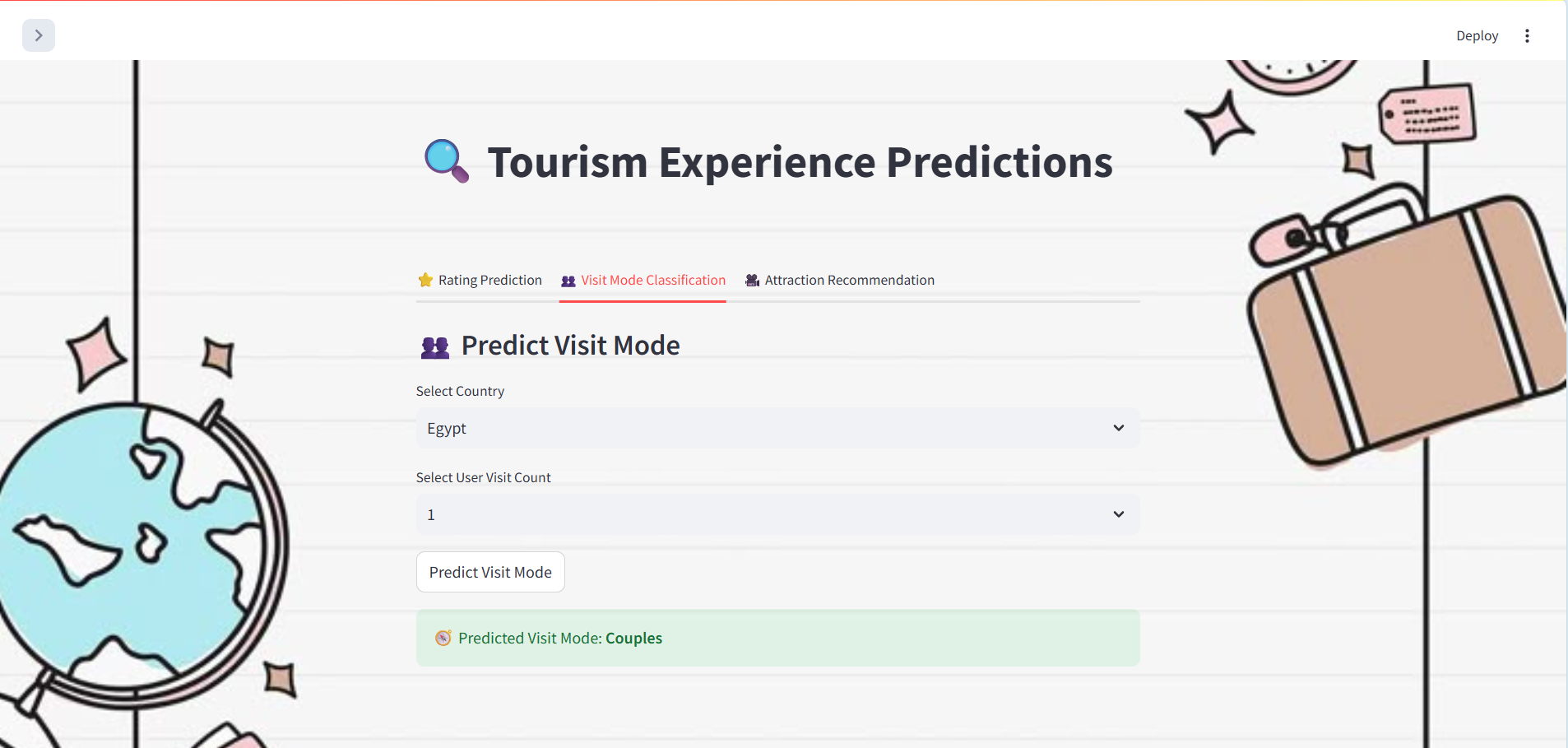
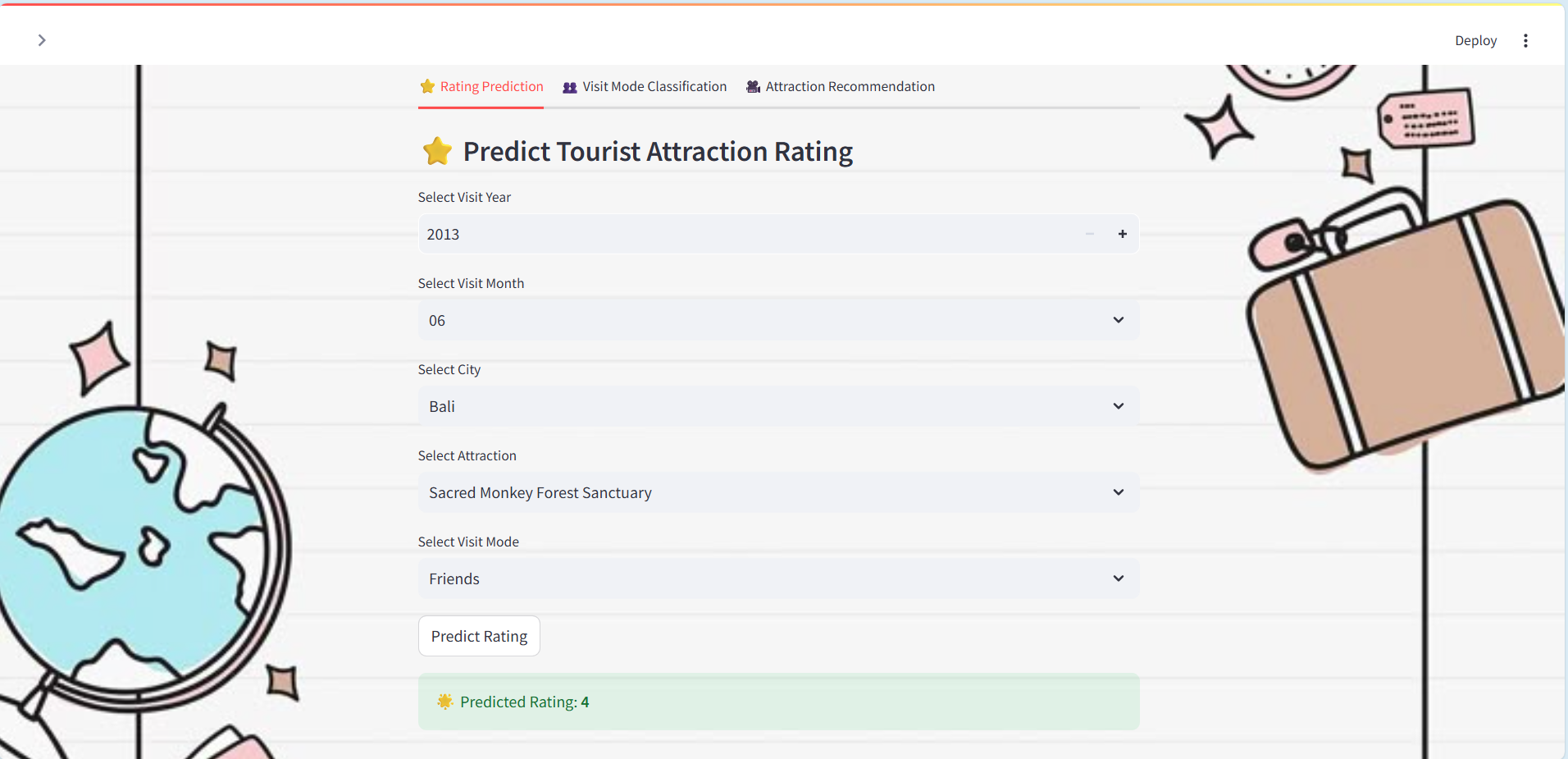
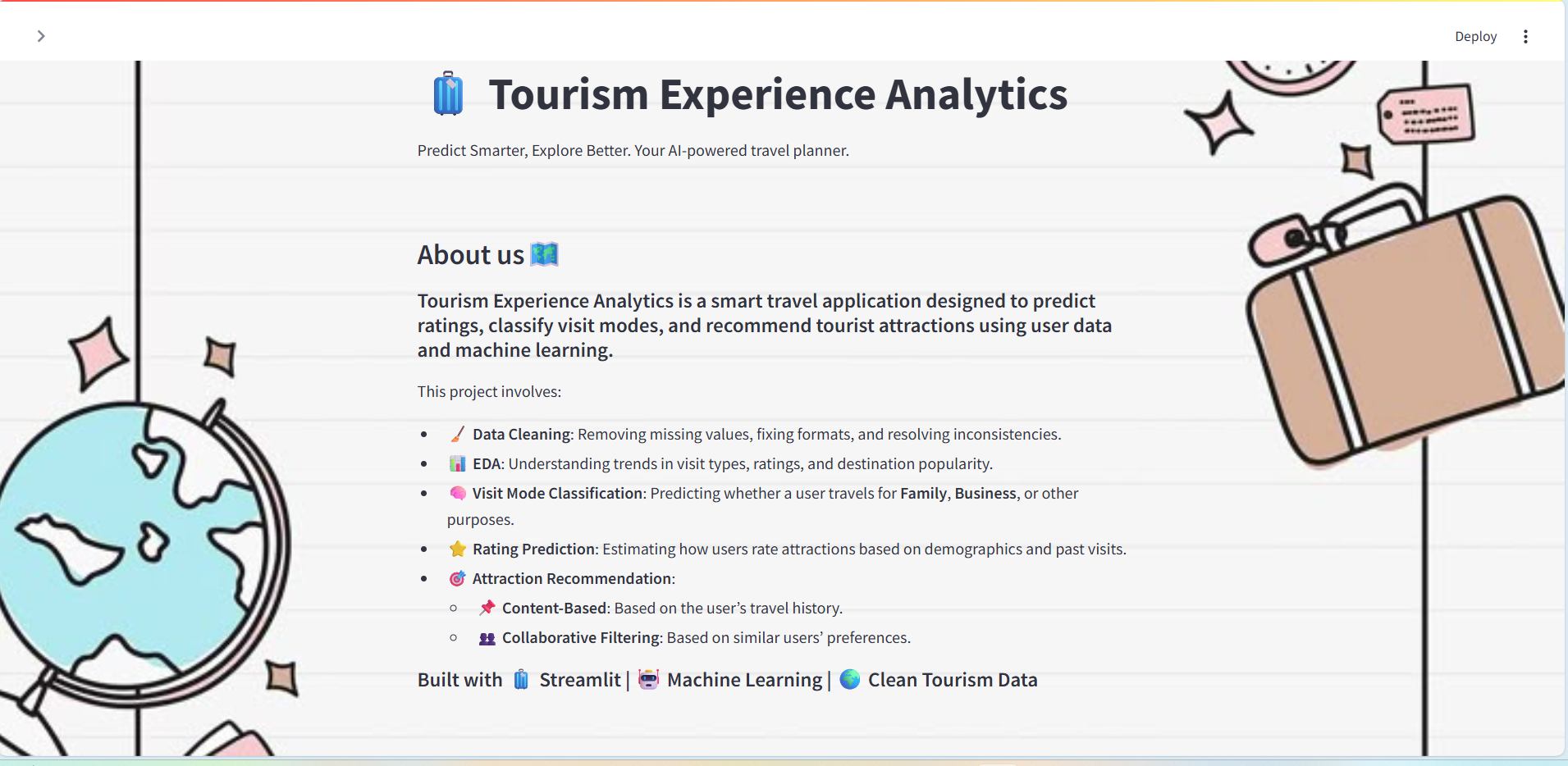
#### **B. Content-Based Filtering**

* Approach: Recommend attractions **similar to those already visited**
* Features Used: AttractionType, CityName, Continent, Country, Attraction\_Popularity
* Method: Similarity computed using feature vectors for attraction profiles

### **✅ Summary of Model Selection**

| **Task** | **Best Model Used** | **Key Metric** | **Value** |
| --- | --- | --- | --- |
| **Rating Prediction** | CatBoost Classifier | Accuracy | 47.73% |
| **Visit Mode Classification** | Random Forest Classifier | Accuracy | 47.89% |
| **Recommendation** | kNN (Collaborative) + Content-Based | N/A (qualitative) | Personalized suggestions |

**User Interface-Streamlit App Output**

****