Project Report: Tourism Experience Analytics & Recommendation System

1. Executive Summary

The Tourism Experience Analytics project was developed to provide travel agencies and platforms with data-driven tools to enhance customer satisfaction and engagement. By leveraging a complex relational database of tourist transactions, demographics, and attraction features, three distinct machine learning objectives were achieved:

- **Regression:** Predicting the exact rating a user will give an attraction.
- **Classification:** Anticipating a user's travel mode (e.g., Solo, Couples, Family).
- **Recommendation:** Supplying highly personalized attraction suggestions using a hybrid filtering engine.

The project culminated in the successful deployment of an interactive, cloud-hosted Streamlit dashboard.

2. Data Architecture & Preprocessing

To prepare the data for machine learning, extensive data engineering was required to unify and clean the raw inputs.

- **Data Consolidation:** Successfully joined 9 distinct relational tables (Transactions, Users, Cities, Countries, Regions, Continents, Attractions, Attraction Types, and Visit Modes) into a single, comprehensive analytical matrix.
- **Missing Value Imputation:** Handled null values in crucial categorical columns by standardizing them with an 'Unknown' placeholder and

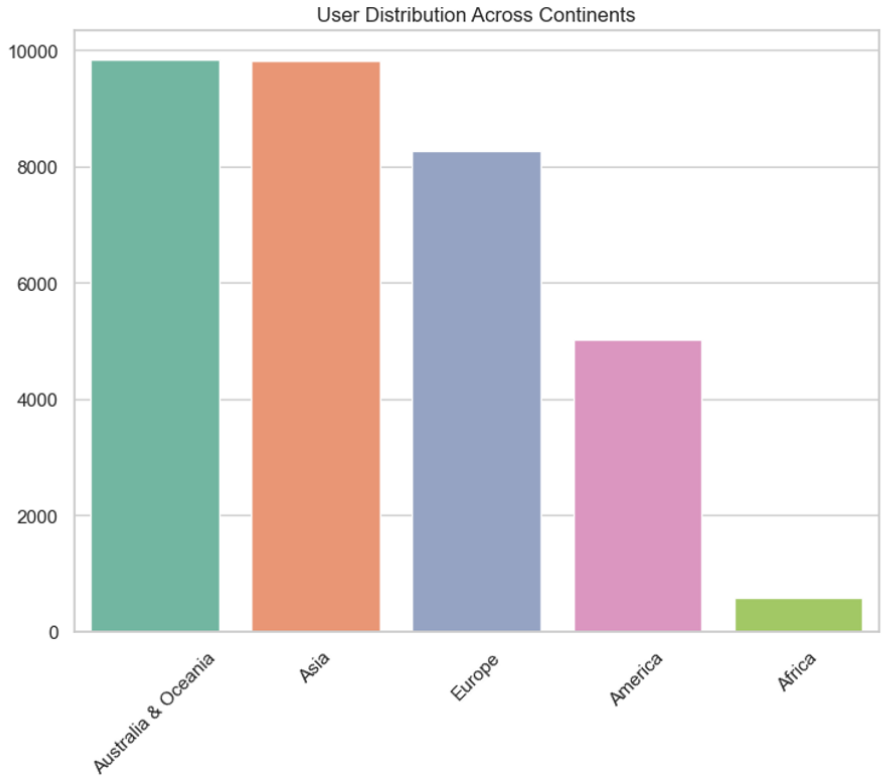
applying uniform .title() text formatting to ensure categorical consistency.

- **Outlier Eradication:** Enforced strict numerical boundaries on the target variable, filtering the dataset to only include valid attraction ratings on a strict 1.0 to 5.0 scale.
- **Feature Engineering:** Encoded categorical strings into numeric arrays using LabelEncoder to prepare the data for LightGBM and regression models, and scaled numeric inputs using StandardScaler.

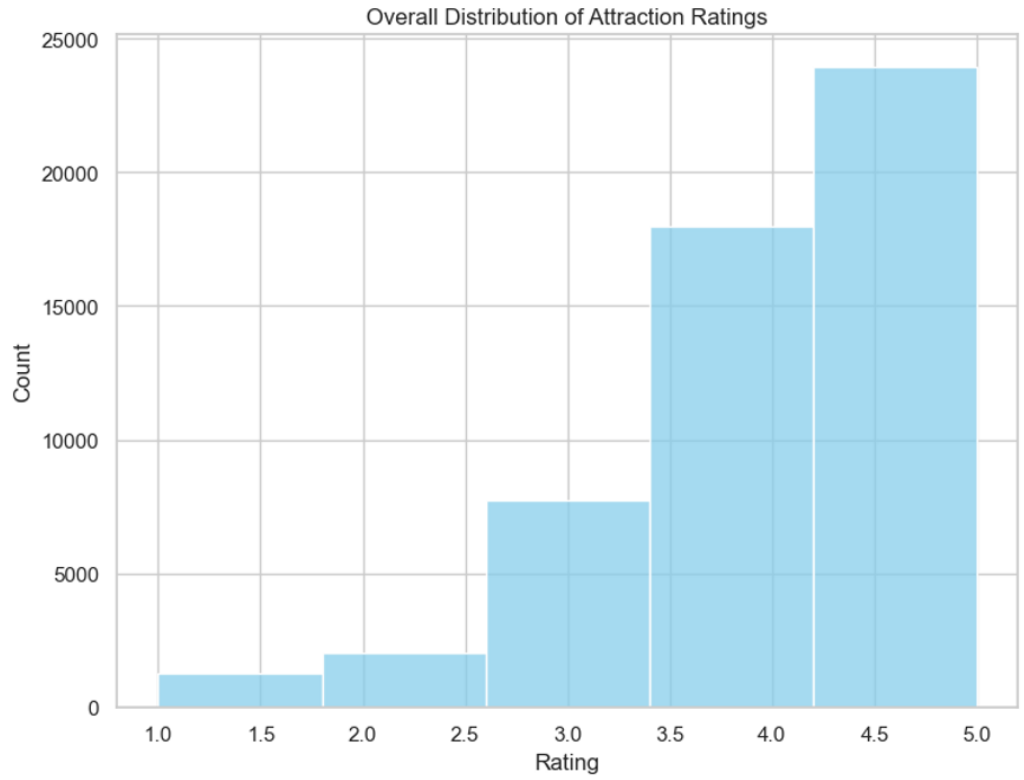
3. Exploratory Data Analysis (EDA) & Visualization

Before modeling, extensive data visualizations were generated in the Jupyter Notebook to plot trends, identify outliers, and spot patterns in global tourism behavior.

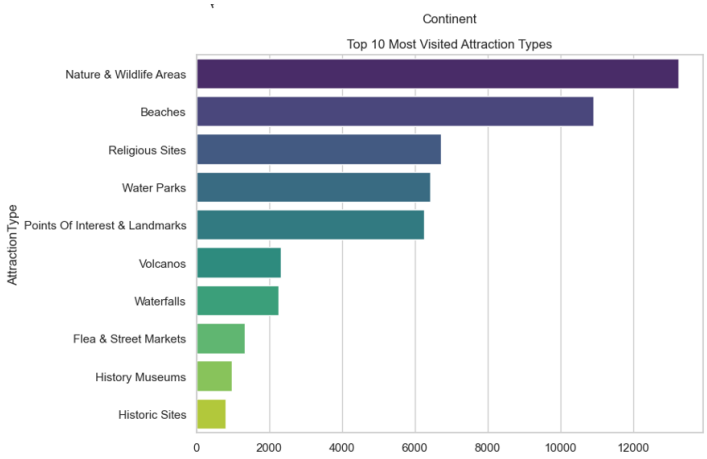
- **Geographic Distribution:** Visualized user distribution across continents, countries, and regions to identify the primary origins of the tourist base.

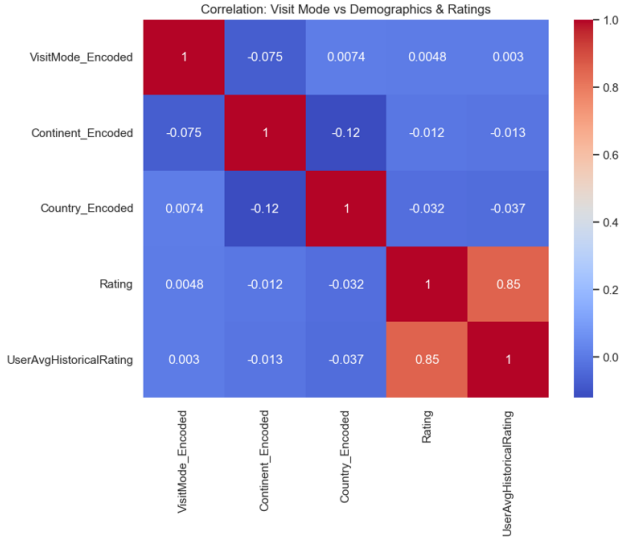


- **Rating Distributions:** Analyzed the distribution of ratings across different attractions, confirming the dataset was successfully cleaned of outliers outside the 1-5 scale.



- **Attraction Popularity:** Explored attraction types based on user visitation and ratings, identifying categories like Beaches and Ancient Ruins as top performers.
- **Demographic Correlations:** Investigated the correlation between VisitMode and user demographics (identifying patterns such as certain regions heavily favoring "Family" travel versus "Solo" travel).





---

4. Predictive Modeling & Evaluation

Three distinct machine learning pipelines were built, trained, and evaluated to solve the core business objectives.

Objective 1: Rating Prediction (Regression)

Designed to estimate user satisfaction to help agencies proactively manage expectations.

- **Linear Regression:** Achieved a Mean Squared Error (MSE) of 0.2525 and an R^2 of 0.7319.
- **Random Forest Regressor:** Achieved an MSE of 0.3176 and an R^2 of 0.6627.
- **Conclusion:** The baseline Linear Regression model outperformed the advanced ensemble method, indicating a strong, highly linear relationship between the engineered features and user satisfaction.

Objective 2: Visit Mode Prediction (Classification)

Designed to predict how a user is traveling to enable targeted marketing campaigns.

- **Logistic Regression:** Achieved a baseline accuracy of 43.18%.
- **LightGBM:** Selected as the final model, achieving an overall accuracy of 49.70%.
- **Conclusion:** LightGBM successfully navigated the highly imbalanced classes. It demonstrated excellent recall (81%) for identifying "Couples" and strong precision (78%) for identifying niche segments like "Business" travelers.

Objective 3: Personalized Suggestions (Recommendation Engine)

Designed to increase customer retention by suggesting undiscovered attractions.

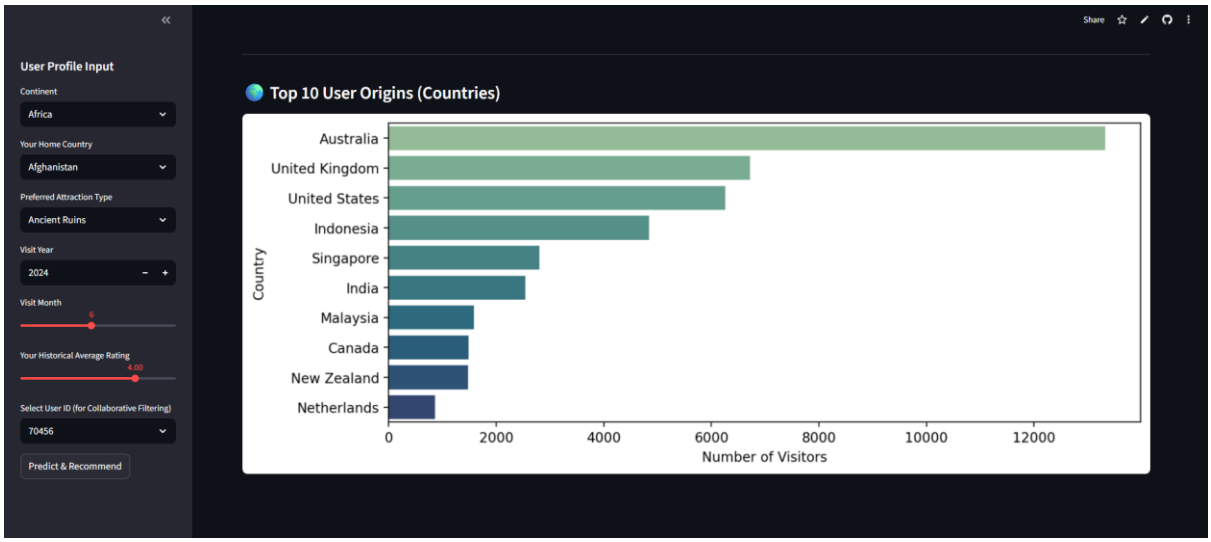
- **Collaborative Filtering:** Built using Singular Value Decomposition (SVD) via the scikit-surprise library.
- **Performance:** Achieved a Root Mean Square Error (RMSE) of 0.9244.
- **Conclusion:** The low RMSE proves the model's ability to accurately map latent user-item matrices and identify hidden historical patterns to predict future affinity.

5. Application Deployment

To make the machine learning models accessible to non-technical stakeholders, an interactive web application was built and deployed to the Streamlit Community Cloud.

- **User Interface:** Features a dynamic sidebar for demographic inputs (Continent, Country, Historical Rating).

- **Hybrid Recommendation Logic:** Combines Content-Based filtering (narrowing the global dataset strictly by the user's preferred attraction category) with Collaborative Filtering (using the SVD model to rank the filtered list and output the personalized "Top 5").
- **Live Analytics Dashboard:** Automatically renders three key visual insights: Top 10 Most Visited Attraction Types, User Distribution by Travel Mode, and Top 10 User Origins.



- **Accessibility:** The models and preprocessors were serialized using joblib for rapid, real-time inference in the cloud environment.

6. Business Insights & Actionable Takeaways

The deployment of these analytical models provides travel platforms with several immediate strategic advantages:

- **Targeted Promotional Campaigns:** By utilizing the LightGBM classification model, businesses can dynamically offer family-bundle discounts to users predicted to travel as a "Family," or romantic getaway packages to predicted "Couples," vastly improving marketing ROI.
- **Proactive Experience Management:** The linear regression model allows agencies to flag users who are statistically likely to leave a low rating for a specific attraction, allowing for proactive customer service interventions.
- **Increased Customer Retention:** The hybrid recommendation engine acts as an automated, highly personalized travel agent, keeping users engaged with the platform by continuously surfacing highly relevant attractions they are mathematically likely to enjoy.
- **Market Trend Identification:** The global tourism analytics dashboard allows stakeholders to continuously monitor shifting popularity in attraction types and regional user growth.