**Tourism Experience Analytics: Classification, Prediction, and Recommendation System**

**1. Executive Summary**

This report presents a comprehensive analysis of tourism data, encompassing geographic insights, time-based trends, attraction ratings, user behaviour, attraction categories, and the performance of predictive and recommendation models. Key findings highlight the dominance of nature and relaxation-focused attractions, the impact of the COVID-19 pandemic on visit trends, and the potential for strategic interventions to optimize visitor experiences and marketing efforts.

**2. Geographic Insights**

Geographic analysis reveals several key patterns:

* **Top Visited Attractions:** Cities such as Singapore, Melbourne, London, and Perth demonstrate high visit volumes. Specific attractions like Waterbom Bali and Merapi Volcano are frequently visited, suggesting their strong appeal across various locations. This indicates both popularity and diversity in attractions across cities.
* **Country-wise Attraction Count:** Indonesia, Singapore, Thailand, and the US lead in the number of available attractions, with Southeast Asia being a prominent tourism hotspot. A high attraction count correlates with active tourism ecosystems.
* **Bottom Countries by Attraction Count:** Countries like Azerbaijan, Burkina Faso, and the Marshall Islands have the fewest listed attractions. These regions may be underexplored or under-documented, representing opportunities for growth or data enrichment.
* **Attraction Density:** North America, Europe, and Southeast Asia exhibit high attraction density, while parts of Africa, Central Asia, and smaller island nations show low density. This highlights regional imbalances in tourist infrastructure or data.
* **Cities with Most Attractions:** Jakarta, Singapore, Kuala Lumpur, and Bali stand out with over 25 attractions, indicating concentrated tourism efforts in urban and capital regions, especially across Southeast Asia.
* **Cities with Fewest Attractions:** Cities like Aachen, Miri, and Misina list only 1 or 1.5 attractions, suggesting they are either less frequented, not fully mapped, or represent emerging/niche tourism markets.

**3. Time-Based Analysis**

* **Monthly Visit Patterns:** July and August are peak months for tourist visits, driven by the summer travel season, school holidays, and favorable weather. February and November show lower visit counts, indicating off-peak periods. This suggests that travel campaigns and resource allocation can be adjusted seasonally.
* **Yearly Visit Trends:**
  + **Growth Period (2013–2016):** A sharp increase in visits, peaking around 2016, likely due to expanded infrastructure, increased promotion, or global travel growth.
  + **Decline Phase (2017–2019):** A gradual decline possibly signals market saturation, economic shifts, or external events.
  + **Sharp Drop (2020):** A clear plummet in visits around 2020 indicates the significant impact of the COVID-19 pandemic.
  + **Slow Recovery (2021–2022):** A modest recovery post-pandemic suggests lingering caution, travel restrictions, or evolving traveler priorities.

Strategic recovery initiatives and benchmarking against historical peaks are crucial for future goal-setting.

**4. Attraction Ratings Analysis**

* **Top-Rated Attractions:** Sacred Monkey Forest Sanctuary and Tegenungan Waterfall have the highest number of 5+ average ratings and substantial user engagement, indicating consistently exceptional visitor experiences. Balekambang Beach, despite a high rating, has fewer total ratings, suggesting it might be a hidden gem.
* **Average Rating by Country:** Countries such as Afghanistan, Uzbekistan, Libya, Bolivia, and Sudan have the highest average attraction ratings (near 5), possibly due to a smaller but highly satisfied visitor base. Conversely, Yemen, Moldova, and Barbados have the lowest average ratings, potentially indicating a need for quality improvement or different user expectations.
* **Strategic Implications:**
  + High-rated but less-visited attractions (e.g., Balekambang Beach) can be promoted in "offbeat destinations" campaigns.
  + Countries with low average ratings could benefit from improving infrastructure or visitor services.
  + Countries with few reviews but high ratings may require better visibility or marketing support.

**5. User Behaviour Analysis**

* **Yearly Rating Trend:** The average rating remained stable from 2013 to 2017 (~4.15–4.2) but significantly jumped post-2018, peaking in 2021 at over 4.5, possibly reflecting improvements in tourism services or changing user perception.
* **Visit Frequency by Visit Mode per Region:** Australia has the highest number of visits, particularly for Family and Business purposes. Southern Europe and Southeast Asia also attract high visits, especially for Couples and Family. Solo and Friends visit modes are comparatively rare.
* **Rating Variance per User:** Most users exhibit low rating variance (0–1), indicating consistent evaluation behaviour. A few users show high variance, potentially representing outliers or diverse experiences.
* **Average Rating per User:** The majority of users give high average ratings (4–5), indicating overall satisfaction. Only a small portion consistently rates below 3.
* **Distribution of Visit Counts per User:** Most users visit attractions only once or twice, suggesting a large base of casual or one-time tourists.
* **Repeat Visits (User-Attraction Level):** Merapi Volcano and Malioboro Road are the most frequently revisited attractions. One user notably visited Merapi Volcano 49 times, signalling strong individual attraction, possibly due to specialized interest.
* **Most Revisited Attractions:** Sacred Monkey Forest Sanctuary leads in repeat visits (~3000+), followed by Merapi Volcano, Waterbom Bali, and Malioboro Road, which are highly engaging locations. These can be prioritized in recommendations for returning visitors or loyalty programs.

**6. Attraction Categories Analysis**

* **Distribution of Attraction Types:** Nature & Wildlife Areas (25%) and Beaches (20.6%) dominate all attractions. Religious Sites, Water Parks, and Points of Interest & Landmarks also have significant shares (~11-12%). This indicates a strong focus on nature and relaxation in the tourism landscape.
* **Average Rating by Attraction Type:** Water Parks, Spas, Caverns & Caves, and National Parks have the highest ratings (above 4.5). Beaches and Historic Sites have the lowest average ratings. This suggests that while nature and beaches are common, user satisfaction is highest for Water Parks and Spas, and Historic Sites and Beaches may require improvements or manage mixed user expectations.
* **Popularity of Attraction Types:** Nature & Wildlife Areas and Beaches have the highest visit counts (~12K+). Spas, Specialty Museums, and Neighborhoods are least visited. This highlights that attractions tied to natural beauty and cultural/religious significance draw more tourists, while high-rated types like Spas and Caverns are underutilized despite strong user satisfaction.

**7. Strategic Insights and Recommendations**

| Aspect | High | Medium | Low |
| --- | --- | --- | --- |
| **Popularity** | Nature, Beaches, Religious Sites | Water Parks, Volcanoes | Spas, Museums |
| **User Rating** | Water Parks, Spas, Caverns | Volcanoes, Museums | Historic Sites, Beaches |
| **Distribution** | Nature, Beaches | Water Parks, Landmarks | Ballets, Ruins |

Based on the analysis, the following recommendations are made:

* **Promote highly rated but under-visited attractions:** Focus marketing efforts on attractions like Spas and Caverns to increase their visibility and visitation.
* **Enhance tourist experience at beaches and historic sites:** Implement initiatives to improve the quality of services and amenities at these locations to boost visitor satisfaction.
* **Balance investment:** Strategically allocate resources between popular natural attractions and quality niche experiences to diversify tourism offerings.

**8. Predictive and Recommendation Systems**

* **Rating Prediction with Regression (CatBoost):** The model achieved a Mean Squared Error (MSE) of 0.23 and an R² Score of 0.75, indicating solid performance and explaining 75% of the variance in user rating prediction.
* **Visit Mode Classification (CatBoost):** The model achieved an accuracy of 45.25%. It showed good precision and recall for classes 2 and 3 but poor performance on minority classes (1 and 5), indicating a strong class imbalance.
* **Recommendation Systems:**
  + **Content-Based Filtering (CBF):** Recommends attractions with similar attributes (AttractionType, City, Country, Rating) using TF-IDF and Cosine Similarity.
  + **Collaborative Filtering (CF):**
    - **Item-Based:** Utilizes the SVD algorithm to extract latent factors and compute cosine similarity between attractions, recommending top N similar attractions.
    - **User-Based:** Recommends top-rated unvisited attractions for a specific user (e.g., user\_id = 70333) by training an SVD model, identifying seen/unseen attractions, predicting ratings, and selecting the top 5 with the highest predicted ratings.
  + **Hybrid Recommender System:** Combines Collaborative Filtering (SVD) and Content-Based Filtering (TF-IDF cosine similarity). It preprocesses content, trains the CF model, and then combines content match with CF prediction using a weighted hybrid score to recommend top N attractions.

This report provides a holistic view of the tourism landscape, offering actionable insights for strategic planning and personalized visitor experiences.