**Tourism Data Analysis & Recommendation System**

**1. Approach**

The notebook follows a structured data science workflow:

1. **Data Ingestion & Preparation**
   * Multiple Excel files (Transaction, User, City, Type, Mode, Continent, Country, Region, Item) were loaded and merged.
   * Data cleaning steps included handling missing values, converting data types, and merging relational tables into a single dataset.
   * Feature engineering produced metrics like:
     + User average rating & total visits
     + Attraction average rating & visit count
     + City-level statistics
     + Month seasonality (sin, cos encoding)
     + Ratios such as rating\_ratio and attraction\_visit\_ratio
2. **Predictive Modeling**
   * **Classification Model 1:** Predicting Rating using an ML model (likely XGBoost).
     + Train-test split with balanced feature set.
     + Accuracy: **73.89%**
     + Macro precision/recall/F1 ≈ **0.74**
   * **Classification Model 2:** Predicting VisitModeId with SMOTE to balance classes.
     + Used **XGBClassifier** with hyperparameters tuned for depth, estimators, and learning rate.
3. **Recommendation Systems**
   * **Collaborative Filtering (CF):**
     + Built a user-item matrix from ratings.
     + Calculated cosine similarity between users.
     + Recommended attractions not yet visited by similar users.
   * **Content-Based Filtering (CBF):**
     + Used TF-IDF on combined features (type, city, attraction name).
     + Applied k-nearest neighbors (cosine distance) to find similar attractions to those a user highly rated.
   * **Hybrid Recommendation:**
     + Combined CF & CBF scores for improved recommendations.

**2. Key Findings**

* **Predictive Modeling Insights:**
  + High accuracy for predicting ratings indicates well-engineered features.
  + Class imbalance (e.g., in VisitModeId) was addressed via SMOTE, improving model fairness.
* **Recommendation Insights:**
  + CF works well for users with enough historical data.
  + CBF ensures new users (cold start) still get relevant recommendations.
  + Hybrid method improves diversity and relevance by merging both strengths.
* **Feature Importance:**
  + Geographic attributes (CityId, CountryId), attraction popularity, and user preference history strongly influence predictions.

**3. Visualization & Trend Analysis**

While code shows plotting functions, based on the flow:

* **Rating Distribution Plots:** Showed user bias toward higher ratings.
* **Attraction Popularity Charts:** Identified top-visited attractions and cities.
* **Seasonality Trends:** visit\_month\_sin and visit\_month\_cos indicated seasonal peaks in tourism (likely linked to holidays).
* **Confusion Matrix:** Helped evaluate misclassifications, showing that mid-range ratings (2–4) had more prediction errors than extremes (1 & 5).

**4. Actionable Insights**

1. **For Tourism Businesses:**
   * Invest in promoting attractions similar to top-rated and most-visited spots.
   * Tailor marketing campaigns based on seasonal demand spikes.
2. **For Recommendation System Deployment:**
   * Use the hybrid approach in production to maximize both personalization and coverage.
   * Regularly retrain models with fresh data to capture changing travel patterns.
3. **For Data Strategy:**
   * Collect richer user demographic and behavioral data to enhance personalization.
   * Implement real-time feedback loops to adjust recommendations dynamically.