**Phase-2 Submission Template**

**Title:** Identify Patterns in Tourism Data to Enhance Travel Industry Decision-Making

**Student Name:** SAMUEL I  
**Register Number:** 411823243020  
**Institution:** RRASE College of Engineering  
**Department:** AI & DS  
**Date of Submission:** 08.04.2025  
**GitHub Repository Link:** [https://github.com/gikulkannan/Kannan-kannan]

**1. Problem Statement**

Tourism is a key economic driver for many regions. However, the travel industry often struggles with dynamic demand, seasonal fluctuations, and changing tourist preferences. The problem addressed in this project is to identify meaningful patterns and trends in tourism data to assist stakeholders—such as travel agencies, hotels, and government bodies—in making informed decisions. This analysis falls under exploratory and descriptive analytics, aiming to uncover insights such as peak travel times, popular destinations, and traveller demographics that influence travel behaviour.

**2. Project Objectives**

* To identify seasonal and regional patterns in tourist arrivals.
* To uncover correlations between factors like weather, festivals, or economic trends and tourism volume.
* To segment tourists based on demographics and preferences.
* To provide visual dashboards and actionable insights for industry use.

**Key Deliverables:**

* Pattern and trend analysis visualizations.
* Regional and temporal segmentation reports.
* Recommendations for targeted marketing and resource planning.

**3. Flowchart of the Project Workflow**

Minx

Copyedit

Data Collection → Data Cleaning → Data Pre-processing → Exploratory Data Analysis → Visualization & Insights → Reporting

**4. Data Description**

* **Dataset Name & Source:** Global Tourism Dataset from Cagle and World Tourism Organization (UNWTO)
* **Data Type:** Structured
* **Rows & Columns:** ~50,000 rows and 20+ attributes
* **Nature:** Static dataset collected over multiple years
* **Key Attributes:** Country, Month, Tourist Arrivals, Expenditure, Age Group, Purpose of Visit, Region, Stay Duration

**5. Data Pre-processing**

* **Missing Values:** Imputed using mean/mode for numerical/categorical data.
* **Duplicates:** Removed using pandas .drop duplicates ().
* **Formatting:** Standardized date formats, region names.
* **Encoding:** Label encoding for region and purpose of visit.
* **Outliers:** Detected using IQR method and removed extreme outliers.
* **Transformations:** Log transformation on highly skewed fields like expenditure.

**6. Exploratory Data Analysis (EDA)**

* **Univariate Analysis:**
  + Tourist arrivals distribution across months
  + Histogram of stay durations
* **Bivariate/Multivariate Analysis:**
  + Heat map of correlation between expenditure, stay, and country GDP
  + Pair plot of age group vs purpose vs spending
* **Key Insights:**
  + Peak tourist seasons observed in summer and winter holidays
  + Business travellers spend more per day compared to leisure travellers
  + European and Southeast Asian regions show highest repeat visitor rates

**7. Tools and Technologies Used**

* **Programming Language:** Python
* **IDE/Notebook:** Google Ecolab
* **Libraries:** pandas, humpy, matplotlib, seaborne, portly, sickest-learn
* **Optional Tools:** pandas-profiling for automated EDA

**9. Team Members and Roles**

**Kannan:** **Documentation & Presentation:** Writing reports and preparing final project submission

**Balaji:** **Data Collection & Cleaning:** Collecting datasets from sources and preparing for analysis

**Samuel:** **EDA & Feature Engineering:** Identifying patterns and creating meaningful features

**Zunaith:** **Model Development & Evaluation:** Training and testing predictive models