**Case Study**

**Auditing & Optimizing International Tourism Impact in Tanzania**

### **1. Business Problem**

Objective:  
 To identify patterns, anomalies, and insights from tourism data to improve travel services, monitor economic impact, and detect irregularities in travel-related spending and reporting in Tanzania.

**Business Use Case:** Tourism is a key economic driver in Tanzania. Understanding visitor profiles, preferences, spending patterns, and travel arrangements helps in policy-making, marketing, and financial audits. The aim is to equip auditors with analytical techniques to evaluate data integrity and detect areas requiring closer scrutiny or improvement.

### **2. Data Description**

**Dataset Overview:**

* **Rows:** 4,809 individual travel records
* **Columns:** 25 attributes detailing tourist demographics, travel behavior, arrangements, and expenses

**Key Variables:**

* **ID :** Serial number of the record (index)
* **country :** Tourist's home country
* **continent**: Continent of origin (Europe, Asia, etc.)
* **age\_group:** Tourist’s age category (e.g., 25–44, 45–64)
* **travel\_with:** Travel Companion type (e.g., Alone, Friends/Relatives, Spouse)
* **total\_female:** Number of females in the travel group
* **total\_male:** Number of males in the travel group
* **Gender:** Combined gender identity (e.g., Male, Female, Male & Female)
* **purpose:** Primary purpose of travel (e.g., Leisure and Holidays, Business)
* **main\_activity**: Main tourist activity in Tanzania (e.g., Wildlife Safari, Beach Holidays)
* **info\_source:** Source from which tourist learned about the destination
* **tour\_arrangement:** Mode of travel planning (e.g., Self, Travel Agent)
* **package\_transport\_int:** Whether international transport was included in the package (Yes/No)
* **package\_accomodation:** Whethe**r** accommodation was included in the package
* **package\_food:** Whether food/meals were part of the package
* **package\_transport\_tz:** Whether local transport within Tanzania was included
* **package\_sightseeing:** Whether sightseeing services were included
* **package\_guided\_tour:** Whether a guided tour was included
* **package\_insurance:** Whethe**r** insurance was included in the travel package
* **night\_mainland:** Number of nights stayed on mainland Tanzania
* **night\_zanzibar:** Number of nights stayed in Zanzibar
* **payment\_mode:** Mode of payment (e.g., Cash, Credit Card)
* **first\_trip\_tz:** Whether it was the tourist's first trip to Tanzania (Yes/No)
* **most\_impressing:** Tourist’s subjective impression (e.g., Wildlife, People, Culture)
* **total\_cost:** Total cost incurred by the tourist (in Tanzanian Shillings)

### **3. Instructions for Steps & Process**

#### **Basic Statistics**

* Use describe() for numerical summary (mean, median, min, max)

**Check for the following data inconsistencies:**

* **Incorrect Data Types** – Ensure all variables are appropriately typed
* **Inconsistent Categorization / Labels**
* **Mixed Units or Formats** – Standardize units across records
* **Mixing of Lowercase & Uppercase** – Standardize text entries

#### **Data Preparation**

* Handle missing data (e.g., travel\_with,most\_impressing)
* Check for Duplicate records and handle them
* Address inconsistencies in the data
* Convert data types where necessary (e.g., total\_female,total\_male to integers)
* Create new variables if useful (e.g.,total\_nights = night\_mainland + night\_zanzibar)

#### **Data Visualization (Guiding Questions)**

* Which regions, travel intentions, and age brackets are most associated with tourism patterns?
* How does the expenditure differ depending on the origin or chosen activities of travelers?
* What are the underlying relationships among the numerical features in the dataset?
* What proportion of travelers utilized each component of their travel package?
* How are travelers spread across different age brackets and regions?

**Generate data table from the dataset:**

* Which combinations of travel package components (transport, sightseeing, insurance, etc.) are most frequently chosen?
* Which travel companion types are most common across different purposes of visit?
* What are the average numbers of nights stayed on mainland vs. Zanzibar, grouped by purpose?

### **Interpretation & Insights**

Use the visualizations you’ve created to explore and interpret key findings from the dataset. Your goal is to generate meaningful insights that support decision-making in Tourism industry.

You are encouraged to:

* Observe and comment on patterns, anomalies, or trends evident in the data.
* Reflect on how historical comparisons can reveal shifts or early warning signals.

**Note:** Avoid simply answering predefined questions. Instead, approach your analysis with an open mind and document your observations based on what the data reveals.