**Introduction to Data science (phase 01)**



**Submitted To:**

Dr. Fakhar Mustafa

**Submitted By:**

Ayesha Raza (SP23-BCS-028)

Fatima Mehboob (SP23-BCS -037)

Iman Akram (SP23-BCS -056)

**Department of Computer Science**

**Comsats University Islamabad**

**(Sahiwal campus)**

# Group 14: Tourism and Economic Growth

**Phase 1: Data Transformation and Preparation**

## 1. Dataset Description

This project investigates the relationship between international tourism arrivals and economic growth across multiple countries from 1995 to 2022. Two key data sources were integrated — the United Nations World Tourism Organization (UNWTO) for tourism arrivals and the World Bank for GDP data. The dataset includes 217 countries and over 4,900 observations, structured as a panel format (country–year). This enables both cross-country and time-series analysis to assess how tourism contributes to economic performance.

## 2. Data Importing

Tourism and GDP data were collected using the World Bank’s World Development Indicators (WDI) database. Tourism data includes international arrivals per country per year, while GDP data provides total GDP (current US$) and GDP growth (%). Both datasets were merged using country name, ISO3 code, and year to ensure uniformity and comparability.

## 3. Data Cleaning

Data cleaning ensured the accuracy and consistency of the dataset. Country names were standardized using ISO3 codes to resolve naming mismatches. Missing data for small or island nations were removed to avoid bias. Duplicates were checked using country-year identifiers, and no duplicates remained after cleaning. Extreme values from large economies were verified and retained as genuine observations.

## 4. Data Transformation

After cleaning, the tourism and GDP datasets were merged on a country–year basis to form a unified dataset. A new feature, Tourism-to-GDP Ratio, was created to measure tourism’s relative contribution to each country’s economy. The data were then reshaped into panel (long) format, with each row representing one country in one year. All variables were standardized and verified to ensure completeness. The dataset remained unbalanced, as not all countries had data for every year, which is typical in international economic datasets.

## 5. Challenges and Solutions

|  |  |  |
| --- | --- | --- |
| **Challenge** | **Description** | **Solution Implemented** |
| Inconsistent country naming | Mismatched country names between tourism and GDP datasets | Standardized using ISO3 codes |
| Missing observations | Some countries lacked complete records | Removed or interpolated missing values |
| Different time ranges | Datasets covered different periods | Merged using only common years (1995–2022) |
| Panel imbalance | Unequal data availability across countries | Handled as an unbalanced panel in analysis |

## 6. Summary

The final dataset combines global tourism and economic indicators for the years 1995–2022. It enables the exploration of how changes in tourism activity influence GDP growth across nations. The Tourism-to-GDP Ratio provides a meaningful measure of tourism’s role in national economies. This clean and transformed dataset serves as the foundation for Phase 2, where fixed and random effects models are applied to analyze the impact of tourism on economic growth.