Faculty of Media Engineering and Technology
German University in Cairo
Dr. Nourhan Ehab

# Milestone 1: International Hotel Booking Analytics

## Introduction

This project implements a comprehensive machine learning pipeline to predict the country group of hotel reviews based on user demographics, hotel characteristics, and review scores. The application addresses a real-world business problem: understanding regional preferences and patterns in hotel bookings to improve marketing strategies, recommendation systems, and customer targeting. Hotel booking data, when analyzed effectively, provides invaluable insights into consumer behavior and the impact of various promotions. By leveraging analytics, hotels can optimize pricing strategies, personalize guest experiences, and forecast demand with greater accuracy, ultimately leading to increased revenue and customer satisfaction.

## Dataset:

1. **Hotels Dataset (25 hotels)**

The *hotels.csv* file serves as the inventory of hotels in the dataset, providing essential context for user reviews. Each row represents a unique property with key identifiers such as *hotel_id* and *hotel_name*, alongside geographical details like *city*, *country*, *lat*, and *lon*. These attributes are indispensable for answering the city-level question.

A hotel's baseline quality is represented by fields such as *star_rating*, *cleanliness_base*, *comfort_base*, and *facilities_base*. These base metrics can be contrasted with customer review scores to evaluate where hotels meet or deviate from expectations.

2. **Reviews Dataset (50,000 reviews)**

The *reviews.csv* file is the heart of the dataset, serving as the central log for all customer feedback. Each record in this table represents a unique review, providing a multi-faceted view of a customer's experience. Beyond a single overall score, the table breaks down satisfaction into several key dimensions.

The file contains unique identifiers like *review_id*, *user_id*, and *hotel_id*, allowing seamless joining with the other tables to connect reviews with specific customers and hotels.
The scoring system, with columns including *score_overall*, *score_cleanliness*, *score_comfort*, *score_facilities*, *score_location*, *score_staff*, and *score_value_for_money*.

3. **Users Dataset (2,000 users)**
The *users.csv* file provides a list of unique customers, offering essential demographic insights with columns *user_id*, *country*, *age*, *gender*, and *traveller_type* [*Solo, Business, Family, Couple*].

Dataset Source: [International Hotel Booking Analytics | Kaggle]

---

## Project Objectives

1. **Data Cleaning:**

   Remove unnecessary columns and handle null values/duplicates, if any.

2. **Data-Engineering Questions**

   From the given dataset, you should analyze the given CSV file to answer and **visualize** the reasoning for the following data engineering questions:
   - Which city is best for each traveler type? For each traveler type, recommend the best city based on the given reviews.
   - What are the top 3 countries with the best value-for-money score per traveler's age group?

3. **Predictive Modeling Task**
   Develop a statistical ML model or shallow FFNN to predict the country groups in the new column 'country_group', given the following features:
   - Score-Based Features from the hotel's info and the users' reviews
   - Features about the user, including their age group, type, and gender

- Quality-Based Features represent the overall score and value for money based on the user's review with respect to the hotel's information.

The model's output will be one of the predefined country groups listed below. This is a multi-class classification problem. The developed models should be evaluated using *accuracy, precision, recall*, and *F1-score*.

4. **Model Explainability**

To gain deeper insights into the model's behaviour, apply explainable AI techniques (SHAP, LIME) to interpret predictions, ensuring transparency about the most influential features.

| Country Group | Countries |
|---|---|
| North_America | United States and Canada |
| Western_Europe | Germany, France, the United Kingdom, the Netherlands, Spain, and Italy |
| Eastern_Europe | Russia |
| East_Asia | China, Japan, and South Korea |
| Southeast_Asia | Thailand and Singapore |
| Middle_East | United Arab Emirates and Turkey |
| Africa | Egypt, Nigeria, and South Africa |
| Oceania | Australia and New Zealand |
| South_America | Brazil and Argentina |
| South_Asia | India |
| North_America_Mexico | Mexico |

Faculty of Media Engineering and Technology
German University in Cairo
Dr. Nourhan Ehab

## Project Deliverables

1. A refined Jupyter Notebook documenting all steps and providing a reproducible workflow; Adding justifications and explanations to the design and results.
2. A cleaned dataset with the country_group column added, and modified columns after the feature engineering step (if any).
3. An analytical text report answering and visualizing the data engineering questions. The report should also include the features used in the predictive model and why each feature was selected.
4. Predictive model (statistical ML or shallow FFNN) for predicting country_group.
5. XAI outputs (SHAP plots and LIME explanations) illustrate the contribution of different features to predictions.
6. An inference function that takes raw input and returns the model prediction.

## Submission and Deadline

Please submit your GitHub repository, fulfilling the specified requirements, using the following form: https://forms.gle/GyFaidvUY2BoKCby6 by **October 22nd at 11:59 pm**

Ensure your repository remains private until the deadline. After the deadline, you will be required to make it public or add the course account (csen903w25-sys) as a collaborator for milestone grading.