# Task 1 :

# Problem Statement: Airline Passenger Data Preprocessing

## Scenario Overview

An airline collects **passenger flight booking and check-in data** from three different sources:

1. **Online Booking System (JSON)** – Direct airline website & mobile app bookings.
2. **Third-Party Travel Agency (XML)** – Bookings made through external agencies.
3. **Airport Check-In System (JSON)** – Actual check-in records at the airport.

Each source contains **different sets of passengers**, meaning some passengers book directly, some through travel agencies, and others only appear at check-in. However, **no single source has all the data**, and merging them presents challenges like **missing fields, inconsistent formats, and duplicate records**.

The final **goal** is to create a **clean, structured dataset** by extracting relevant information from all sources, resolving inconsistencies, and removing duplicate entries.

## Dataset Information

### Online Booking Dataset (online_booking.json)

- Passengers who booked directly through the airline's website/app.
- Extra Fields: Loyalty status, payment method, special requests.
- Issues:
    - **passenger_name stored as first_name + last_name.**
    - **seat_number is missing for some passengers.**
    - **Phone numbers stored with various formats (e.g., +1-555-123-4567).**

### Third-Party Travel Agency Dataset (third_party_travel_agency.xml)

- Passengers who booked through external travel agencies.
- Extra Fields: Frequent flyer info, gate numbers, aircraft type.
- Issues:
    - **ticket_class is missing for some passengers.**
    - **departure_time and arrival_time stored in different formats.**

    - **Phone numbers stored with inconsistent formatting (e.g., (555) 123-4567).**

### Airport Check-In Dataset (airport_check_in_data.json)

- **Passengers who checked in at the airport.**
- **Extra Fields: Baggage weight, boarding pass issued.**
- **Issues:**
    - **ticket_price_usd is missing in some .**
    - **Phone numbers may have different formats from other sources.**

## Handling Missing Values

In the dataset, some fields are missing in different files, so we need to ensure they are properly handled. Here's how the script deals with them:

### Missing Ticket Class (Some Missing in Travel Agency File)

**Approach:**

- **If ticket_class is missing, fill it with "Unknown" .**

### Missing Seat Number (Some Missing in Online Booking)

**Approach:**

- **If seat_number is missing, fill it with "Unassigned".**

### Missing Ticket Price (Check-in File Missing Ticket Prices)

**Approach:**

- **Estimate ticket prices based on ticket class.**

### Missing Airport Information

**Approach:**

- **If departure_airport or arrival_airport is missing, fill with "Unknown".**

### Missing Payment Status

**Approach:**

- **If payment_status is missing, use "Completed" for online bookings and "Unknown" for check-in records.**

# Data Format Normalization

## Date and Time Format Standardization

**Approach:**

- **Convert all date/time strings to a standard UTC format (YYYY-MM-DD HH:MM:SS UTC).**
- **Ensure consistent timezone representation across all records.**
- **Handle various input formats (12-hour vs 24-hour time, different date formats).**

## Phone Number Normalization

**Approach:**

- **Strip all non-essential characters (parentheses, spaces) while preserving essential ones (+, -).**
- **Ensure all phone numbers follow a consistent format.**
- **Retain country codes where available.**
- **Example: Convert "(555) 123-4567" and "555.123.4567" to "555-123-4567".**

# Final Preprocessing & CSV Creation Strategy

1. **Extract relevant fields** from each dataset.
2. **Normalize formats** (date-time, phone numbers, ticket class).
3. **Drop duplicate records** based on **booking_id + passport_number**.
4. **When handling duplicates,** prioritize records that have actual values (not "Unassigned" or "Unknown") over those with placeholder values. If multiple records exist for the same passenger, preserve the most complete record with meaningful data.
5. **Store in a structured CSV format** with the following columns:

# Final CSV File Format

The consolidated output CSV file should contain the following columns:

1. **booking_id** - Unique identifier for the booking
2. **passenger_name** - Full name of passenger
3. **passport_number** - Passenger's passport ID
4. **email** - Contact email address
5. **phone_number** - Normalized contact phone number
6. **flight_number** - Airline flight identifier
7. **departure_airport** - Airport code for departure
8. **departure_time_utc** - Standardized departure time in UTC

9. **arrival_airport** - Airport code for arrival
10. **arrival_time_utc** - Standardized arrival time in UTC
11. **ticket_class** - Class of service (Economy, Business, First, Unknown)
12. **seat_number** - Assigned seat or "Unassigned"
13. **ticket_price_usd** - Ticket price in USD or 0.00 if unknown
14. **payment_status** - Status of payment (Completed, Pending, Unknown)

# Task 2: Mean, Median, and Mode Analysis for Different Datasets

## Problem Statement: Statistical Measure Performance Analysis

### Scenario Overview

In data analysis, central tendency measures (mean, median, and mode) each have strengths and weaknesses depending on the data distribution. This task explores three distinct datasets where each of these measures outperforms the others as a representative value.

You will create dummy datasets and analyze which statistical measure performs best for each scenario, providing a comprehensive explanation of why one measure outperforms the others in each case.

### Dataset Requirements

Create three different datasets:

1. **Income Distribution Dataset (Array 1)** - A dataset representing household incomes in a community.
    - Should contain at least 100 data points.
    - Should include some outliers representing high-income individuals.
2. **Product Rating Dataset (Array 2)** - A dataset representing customer ratings for a popular product.
    - Should contain at least 50 data points.
    - Should be discrete values (e.g., 1-5 star ratings).
    - Should have a clear mode (most common rating).
3. **Temperature Dataset (Array 3)** - A dataset representing daily high temperatures in a city for a month.

- Should contain 30-31 data points.
- Should follow an approximately normal distribution.
- Should include a few minor outliers (unusually hot or cold days).

## Analysis Requirements

For each dataset:

1. Calculate the mean, median, and mode.
2. Analyze which measure of central tendency provides the most representative value for the dataset.
3. Explain why this measure is more appropriate than the others in this specific context.
4. Describe a real-world scenario where using the wrong measure could lead to misinterpretation.

**Bonus Task: Medium Article and LinkedIn Post**

**Task Description**

Write a small article for Medium about the data preprocessing, file handling, and statistical techniques discussed in the previous tasks, and share your solution on LinkedIn