FACULTY OF ENGINEERING AND BASIC SCIENCES
ACADEMIC PROGRAM: DATA ENGINEERING AND ARTIFICIAL INTELLIGENCE

COURSE: ETL (G01)
LAB-5. Data Quality Dimensions

1. Objectives

1. Identify data quality issues in the dataset according to the six main dimensions of data quality (completeness, accuracy, consistency, uniqueness, validity, timeliness).
2. Generate a **quality assessment report** describing the problems detected for each column.
3. Define **at least 6 data quality policies** to ensure the reliability of the dataset for analysis.
4. Propose **solutions and cleaning strategies** (e.g., imputation, normalization, deduplication, validation rules).
5. Apply some cleaning operations in Python using Pandas to improve data quality.

2. Context

Airplane crash records are critical for aviation safety analysis. The dataset *Airplane Crashes and Fatalities* (https://www.kaggle.com/datasets/thedevastator/airplane-crashes-and-fatalities/data) contains information about accidents from 1908 to recent years, including date, location, operator, aircraft type, fatalities, and other variables.
However, as with most real-world datasets, issues such as missing values, inconsistencies, duplicates, and invalid formats are present. If not addressed, these issues may lead to incorrect conclusions in safety reports and predictive analytics.

3. Business Objectives

a. Enhance Aviation Safety Analysis

- Objective: Ensure accident and fatality counts are correct so that aviation authorities can reliably identify the most dangerous aircraft types and operators.

b. Improve Historical Reliability for Trend Analysis

- Objective: Guarantee that all crashes have valid dates and locations recorded to enable consistent time-series analysis of accident frequency and geographic distribution.

c. Support Consistent Regulatory Reporting

- Objective: Standardize formats for dates, aircraft types, and operator names to ensure data comparability across multiple reports, agencies, and years.

d. Avoid Duplication in Safety Statistics

- Objective: Detect and remove duplicate crash records so that the total number of accidents and fatalities is not artificially inflated in official reports.

e. Enable Accurate Risk Models for Modern Aviation

- Objective: Prioritize recent crash records with up-to-date details, ensuring predictive risk models and safety recommendations reflect current conditions in aviation, not outdated patterns.

4. Tasks

a. Map Data Quality Dimensions vs Business Impact (taking into account the business objectives)

b. EDA

c. Exploratory Quality Check

- For each column, report:
  - Data type
  - Missing values
  - Number of unique values
  - Invalid formats (e.g., incorrect dates, wrong location formats, negative or impossible values)
  - Duplicates
  - Examples of problematic records

d. Quality Report

- Create a structured report summarizing issues per quality dimension.

e. Data Quality Policies Proposal

- Write 6 clear **data quality policies** (e.g., "All date fields must follow YYYY-MM-DD format", "Fatalities cannot exceed total passengers", "Operator names must be standardized").

f. Cleaning Actions

- Propose at least **two possible solutions per problem** (e.g., dropping vs. imputing null values, regex normalization for dates, merging duplicates).
- Implement a few cleaning operations with Pandas.

5. **Final Deliverables**
   a. A short-written **report (PDF)** containing:
      i. The Mapping of Data Quality Dimensions vs Business Impact.
      ii. The results of Exploratory Quality Checks.
      iii. The Quality Report.

iv. The Data Quality Policies Proposal.
v. The Cleaning Actions Proposal and Justification.

b. A **Python notebook/script** with exploration and quality checks and example cleaning solutions.