# Airbus Data Management & Governance challenge

This document outlines a comprehensive approach to solving the Airbus Data Management & Governance Challenge.

## Defining Success (Qualitative and Quantitative) :

- Qualitative Indicators:
    - Diversity of books: A broad distribution of genres and authors' nationalities
    - Accessibility and usability: Ease of access to books (physical and digital)
    - User satisfaction: Positive feedback from teachers, students, and parents via surveys

- Quantitative Indicators (KPIs):
    - Total number of borrowings per month (about 200 to 400 books per month)
    - Percentage of genre diversity (at least 10% per genre)
    - Representation of authors' nationalities (about 30 different nationalities)
    - Library catalog renewal rate (adding about 500 new titles per year).
    - Number of active users (about 200 regular users per month would be a good start).

## Data Sources and Enrichment :

To create a robust dataset, i integrated multiple data sources and performed enrichment using modern techniques :

- Data Sources : I combined five public Kaggle datasets, including books, ratings, tags, to-read lists, and prices.
- External enrichment : Additional data was sourced from the Wikipedia API to classify authors by nationality.
- Genre and Age Classification: Genres and age groups were inferred based on parsed tags and metadata.

## **Data Cleaning and Preprocessing :**

- Removed  books with incomplete metadata (missing titles , authors, language).
- Removed duplicates based on isbn title and authors matching.
- Retained only English language books to meet requirements.
- Removed extreme values to ensure balanced selections.

## **Scoring Logic :**

I assigned to each book a final score based on the following key performance indicators (KPIs) :

- Popularity : based on number of ratings.
- Diversity : based on genres and authors nationalities.
- Price efficiency : Books offering the best value for cost.

The final score was calculated as a weighted average.

## **Final Selection :**

After scoring, i applied a quota-based selection strategy to ensure diversity :

- per genre : to prevents overrepresentation
- per nationality : to ensure diverse cultural representation
- per age group : books were grouped into Children, Young Adult and Adult categories

## **Data Visualization and Dashboard :**

I built a Streamlit dashboard with Plotly visualizations to enhance usability :

- Interactive Tabs: Users can toggle between the full dataset and the final selection.
- Genre & Nationality Distribution
- Score & Price Visualizations
- Score vs. Year of Publication: A scatter plot visualizing how book scores vary by publication year, helping identify trends in quality over time.
- Downloadable CSV Outputs: Enables easy data export .

## Results Summary :

- Total books analyzed : ~ 9800
- Final books selected : 5000
- Simulated budget for books ~ 20 000 €

## Estimating the total budget :

| Expense Category | Estimated cost |
|---|---|
| physical books | 23 000 € |
| Infrastructure (shelves, furniture) | 10 000 € |
| Staff / Volunteers | 8 000 € |
| Digital book licenses | 15 000 € |
| Digital management system | 5 000 € |
| Other expenses (maintenance, internet, software) | 5 000 € |
| **Total Estimated Cost** | 66 000 € |

## Challenges and Learning Experience :

This challenge was both demanding and enriching. One of the first difficulties was finding a comprehensive dataset that met the selection criteria. Many datasets lacked key metadata, requiring extensive preprocessing and enrichment using external APIs, such as the Wikipedia API for author nationality classification and more. Another issue was the API integration itself retrieving data was time-consuming. Additionally, balancing genre, nationality, and age group quotas while maintaining a high-quality selection proved to be a complex multi-objective optimization problem. Despite these efforts, I am not fully satisfied with the final dataset. The dataset itself could have been improved with more extensive sourcing and richer metadata. Overall, this experience deepened my expertise in data cleaning, feature engineering, and visualization while reinforcing the importance of governance in data-driven decision-making.

## Apprenticeship Expectations :

I am looking forward to starting my two-year apprenticeship in September 2025 as part of my Master's program in Data Science and Engineering. My main expectations for this experience are to apply my theoretical knowledge to real-world data challenges, enhance my skills in data management, machine learning, and data governance, and gain hands-on experience with large-scale data processing and analytics. I aim to progressively take on more responsibilities, contributing to data-driven decision-making and improving my technical and professional skills in a dynamic and challenging environment.