Data Curation – Global Airline Operations

1. Data Sourcing

Dataset Name	Description	Source	Size	Acquisition Method
Global Airline Dataset	Contains data on airlines, delays, flight durations, distances, passenger traffic, IATA codes, etc., over multiple years.	https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction	~2.6 MB	Manual download from Kaggle
Weather Data (optional)	Hourly weather data for major airports, potentially impacting delays.	NOAA / Open-Meteo API	To estimate	API or CSV depending on availability
Airport Info (optional)	Additional info on airports such as location, altitude, timezone.	https://ourairports.com/data/	< 1 MB	CSV, direct download

2. Dataset Profiling

Structure Discovery

- The main dataset contains approximately 130,000 rows and 25 columns.

- Columns are well-typed: numerical (duration, distance), categorical (airline, class, satisfaction).

- Some extreme values noticed in delay fields (e.g., >1000 minutes).

Content Discovery

- Data is well structured but includes some missing values, such as in 'Arrival Delay'.

- Minor typos/variants observed in categorical values (e.g., 'Eco' vs 'Economy').

Relationship Discovery

- Correlation found between distance and flight duration.

- Arrival delays seem correlated with customer satisfaction.

3. Data Wrangling

- Cleaning: Missing values were replaced with the median (numerical fields) or rows with extensive nulls were dropped.
- Standardization: Normalized category names (e.g., class types unified).
- Feature Engineering: Created new variables such as 'total_delay = Departure Delay + Arrival Delay'.
- Final Format: Cleaned dataset contains ~125,000 rows and 27 columns after enrichment.
- Format: CSV, UTF-8 encoded, ~2.8 MB file size.

4. Data Table Schema (excerpt)

Field Name	Type	Description
Airline	STRING	Name of the airline carrier
Date	DATETIME	Date of the flight
CarrierDelay	FLOAT	Delay caused by carrier (minutes)
WeatherDelay	FLOAT	Delay caused by weather
NASDelay	FLOAT	Delay caused by NAS
SecurityDelay	FLOAT	Delay caused by security
LateAircraftDelay	FLOAT	Delay from previous flight

TotalDelay	FLOAT	Sum of all delay causes
Month	INTEGER	Extracted month from Date
Year	INTEGER	Extracted year from Date
AirportCode	STRING	Code of departure airport
GDP	FLOAT	GDP of the country of departure airport (USD)
Weather_Condition	STRING	Summary of weather conditions
IsHoliday	BOOLEAN	Indicates if the flight happened on a holiday
IsWeekend	BOOLEAN	Indicates if the flight happened on a weekend