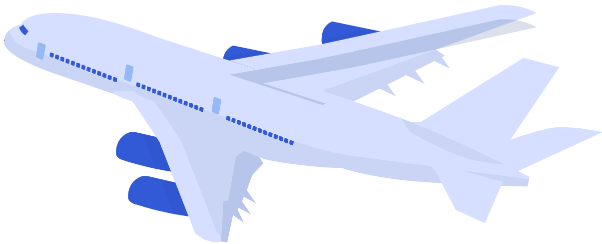
Group 15 - Data Dyno

Weather-Driven Flight Delay Predictions



Analyzing how weather impacts delays and how predictive models can improve performance

Geon Kim, George Ezzat, Jutipong Puntuleng, Steven Gourgy, Melissa MacNab

Table of contents

1 Topic & ML Task

2 Data Collection

3 Data Integration

4 Data Cleaning

5 Data Transformation

6 Data Visualization & ETL

7 Modelling & Evaluation

8 Conclusion

Topic & ML Task

- Flight delays influenced by weather: temperature, precipitation, wind
- Daily climate strongly impacts airport operations
- Integrating flight records + weather data allows us to study how weather drives delay severity

The predictive task involves building a multi-class classifier that predicts the delay severity of a flight based on combined flight and weather attributes

On-Time
≤ 0 min

Minor
1 - 15 min

Moderate
16 – 60 min

Severe
> 60 min

Data Collection

Datasets

U.S. Flight Delay Dataset

30M+ flights, airports, carriers, distances, delay minutes

Global Daily Climate Dataset

27M+ records of temperature, precipitation, wind

Files

Flight_Delay.parquet

detailed U.S. flight delay data

features_added.parquet

flight dataset with extra attributes

daily_weather.parquet

temperature, wind, precipitation per day

cities.csv

city name, lat/long lookup

countries.csv

country-level weather station info

Data Collection

Data Storage

Data stored as CSV and loaded into Panda DataFrames

Environment

Processed locally in Python 3.12 (VS Code / Jupyter Kernel)

Processing Pipeline

Cleaned, integrated, and transformed for downstream modeling

Data Integration

Join Keys
Standardized Date (YYYY-MM-DD) and City Name / IATA Code

City Name Cleaning
Normalized cities: lowercase, removed state codes, trimmed extra spaces

Airport → Weather Mapping
Linked each airport to its nearest weather station using lookup tables

Unit Harmonization
Aligned units across datasets (°F→°C, normalized wind speed, consistent precipitation units)

Integrated Schema
Integrated full pipeline: Flights → Airports → WeatherStations → DailyWeather

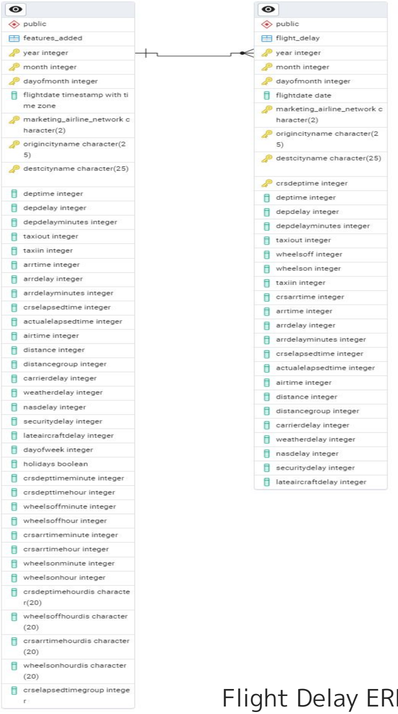
Data Integration

Key Mappings

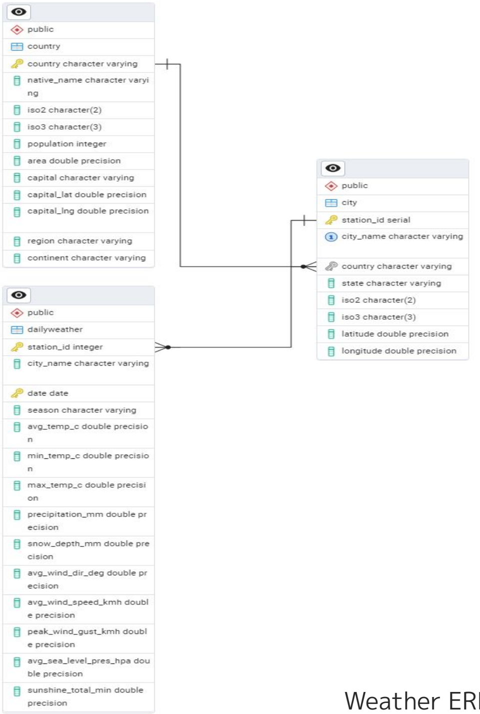
Flight → Airport
(city name / IATA code)

Airport → WeatherStation
(lookup)

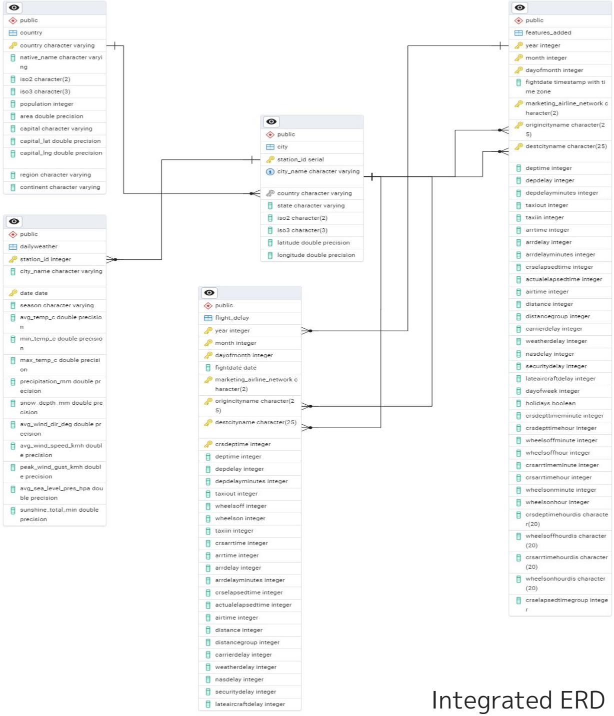
WeatherStation → Daily Weather
(station_id + date)



Flight Delay ERD



Weather ERD



Integrated ERD

Data Cleaning

Null Values

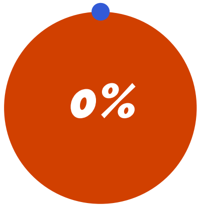
Causes

City/date mismatches
Incomplete station coverage

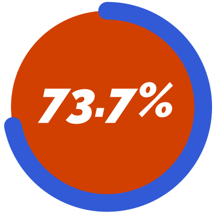
Imputation

Median Imputation
Has_weather_data flag

Flight Data



Weather Data



Features Null Count → Median

avg_temp_c: 22,297,290 → 16.6 °C
precipitation_mm: 22,215,998 → 0.0mm
avg_wind_speed_kmh: 22,332,980 → 11.9km/h

Data Cleaning

Outliers

Detection	Causes	Retention Policy
IQR	Extreme delays Heavy weather are real events	Outliers represent relevant events Create indicator variables (is_severe_delay & is_heavy_precipitation)

843,792 flights flagged as severe delays (>120 min)

flights flagged for heavy precipitation (>25 mm) **193,348**

Data Transformation

Scaling

StandardScaler & MinMax scaling for continuous features

Ensure consistent range for numerical features

Encoding

One-hot encoding for categorical variables (month, day of week, carrier, airport)

Feature Engineering

Derived features for improved meaning

E.g. is_weekend, is_severe_delay, is_heavy_precipitation

Data Visualization & ETL

ETL Pipeline

- Extract: load raw flight & weather data
- Transform: clean tables, engineer features, merge weather & flights
- Load: final dataset for modeling & visualization

Visualization

- Delay distributions
- Weather trends
- Weather vs delay relationships

Dashboard

- Delay severity counts
- Weather breakdown
- Key contributing delay features

Modelling & Evaluation

Model

Random Forest Classifier

Works well with nonlinear and mixed features

Training

80/20 split with stratification

Trained model on cleaned & transformed features

Features

Distance, month, day of week, carrier, airport

Temperature, precipitation, wind speed, heavy precipitation flag

Metrics

Accuracy

Macro F1 Score

ROC-AUC

Modelling & Evaluation

Before vs After (DC & DT)

Missing Values
22.3M → 0

Weather Coverage
26.33% → 100%

Features
13 → 36

Modeling Constraints

Class Imbalance
70%

Severe Delay Outliers
843 792

Model Performance

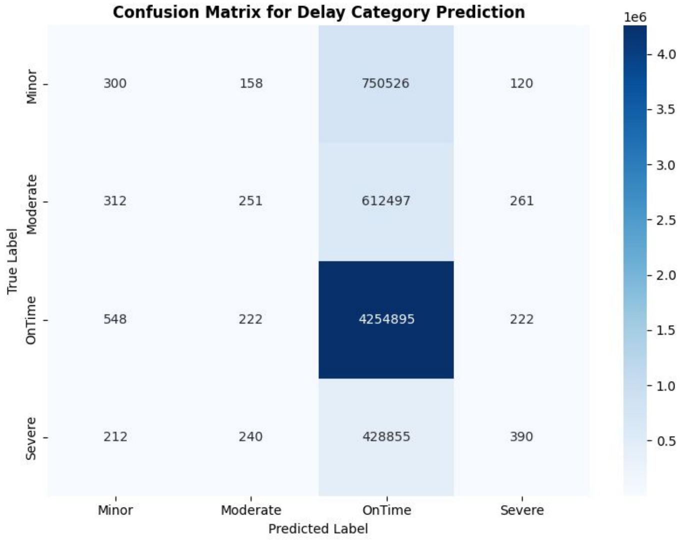
Accuracy
70.34%

F1-Score
0.5813

Per-Class F1

OnTime
0.83

Delays
0.00



Conclusion

- Weather features significantly impact flight delay patterns
- Cleaning, imputation, and feature engineering improved dataset quality
- Random Forest delivered solid performance given extreme imbalance
- Dashboard and visualizations help interpret delay factors

Demo time!