Flight Delay Prediction Based on Weather Condition Analysis

Project Phase 1

Concordia University
COMP 333
Data Analytics
Dr. Essam Mansour

Team Name: DataDyno
Group 15

Steven Gourgy (40213440)
George Ezzat (40245502)
Geon Kim (40264507)
Melissa MacNab (40192264)
Jutipong Puntuleng (40080233)

Sunday, 2nd November 2025

Table of Contents

A. The main topic and the definition of the data science task

Background

Flight delays remain a major challenge in the aviation industry, leading to financial losses, logistical complications, and passenger dissatisfaction. According to reports from the U.S. Bureau of Transportation Statistics, weather-related factors account for a significant share of flight disruptions.

With the growing availability of aviation and meteorological datasets, data science offers a means to better understand and predict these delays. This project integrates U.S. flight delay and global weather data to analyze how factors such as temperature, precipitation, humidity, and wind speed influence delay patterns. The insights gained from this work could support air traffic management, airline scheduling, and customer communication systems to improve overall service reliability.

Predictive Task

The core data science task is classification. Given weather conditions and flight-specific features such as origin city, data, and distance, the objective is to predict the delay severity of a given flight. Classification labels will include:

- On-time: ≤ 0 minutes
- Minor delay: 1-15 minutes
- Moderate delay: 16-60 minutes
- Severe delay: > 60 minutes

B. Summary of the collected datasets

ER Diagram

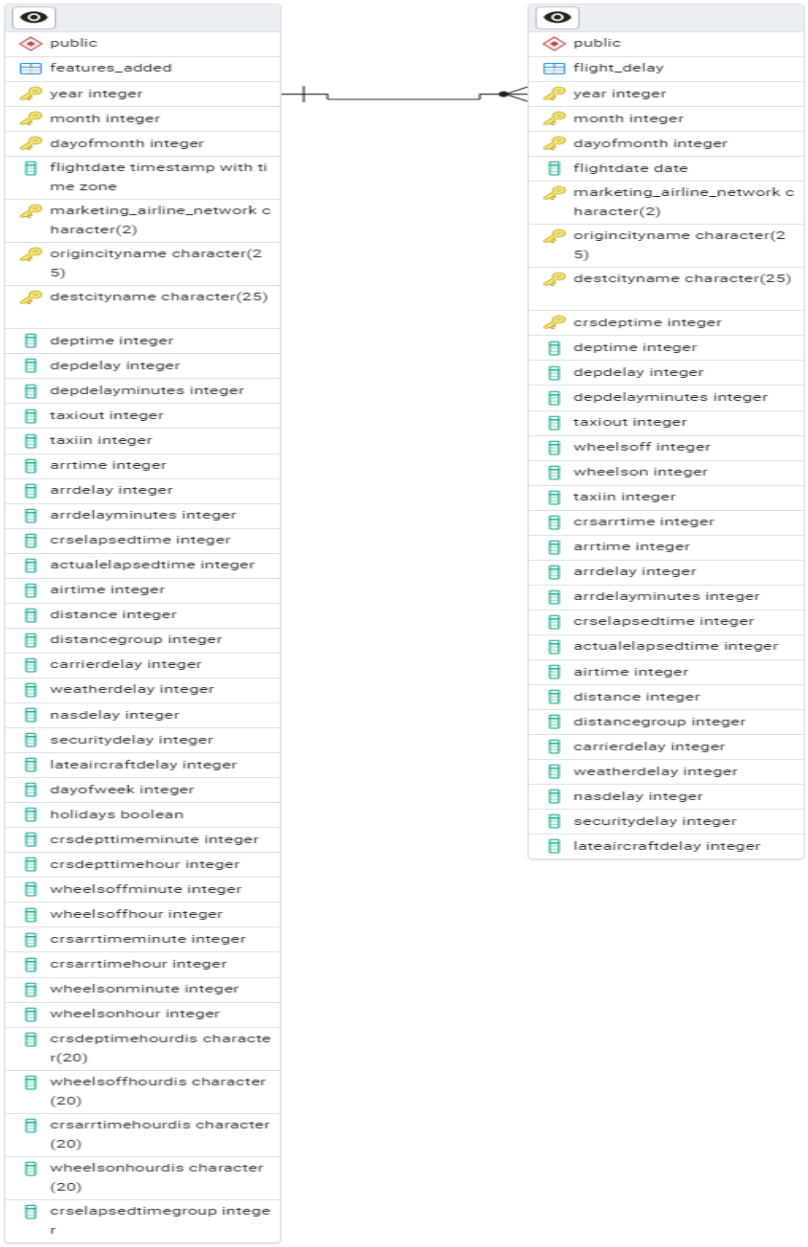

Figure 1: Flight_Delay Entities and relationships

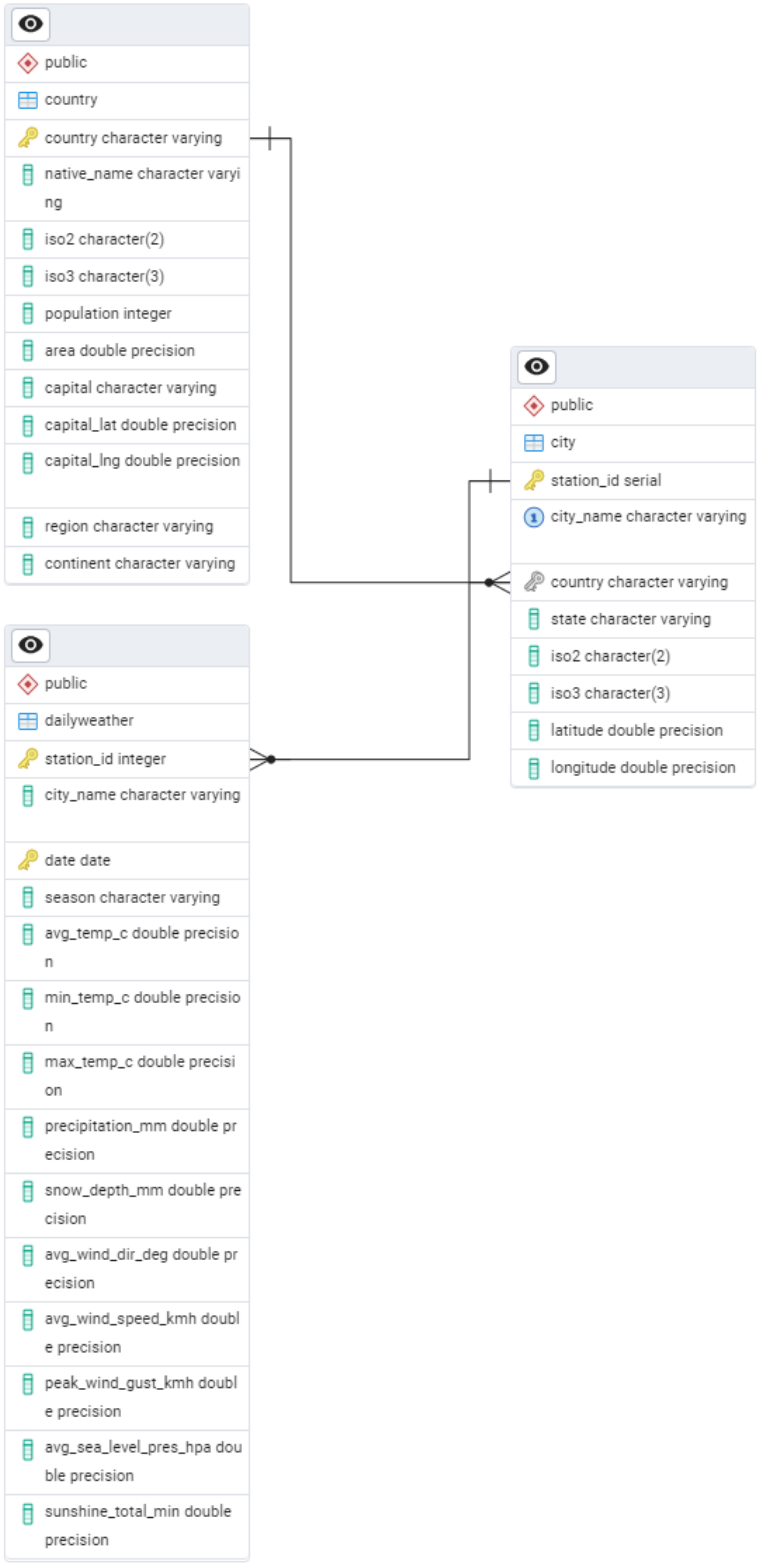

Figure 2: Weather Prediction

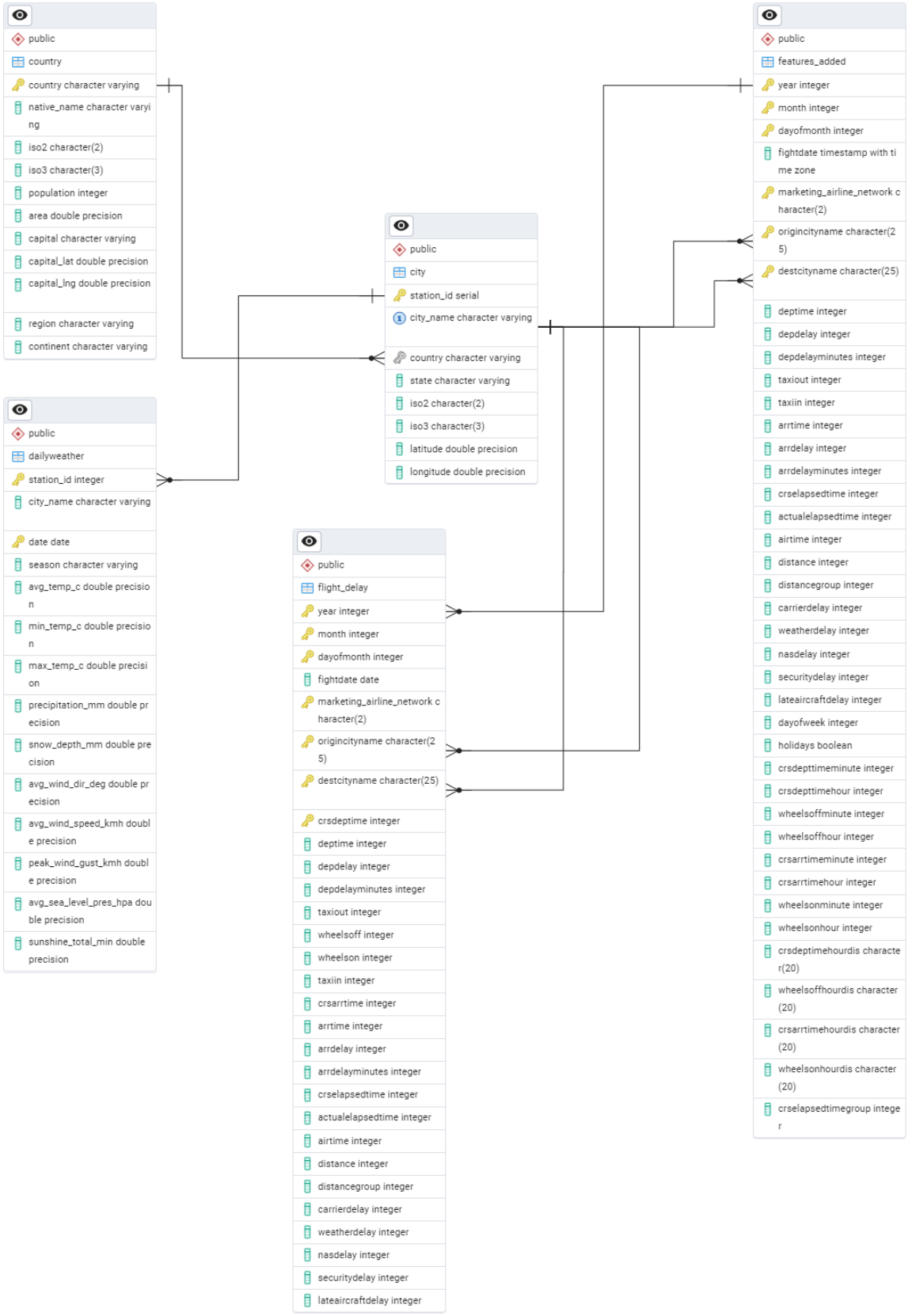

Figure 3: Flight Delay Prediction Based on Weather Condition

Data Sources

Dataset	Description	Source	Link
Weather Dataset	Global daily climate data with temperature, humidity, precipitation, etc.	Kaggle	https://www.kaggle.com/datasets/guillemservera/global-daily-climate-data
Flight Delay Dataset	US flight delay and arrival performance dataset with enriched features	Kaggle	https://www.kaggle.com/datasets/arvindnagaonkar/flight-delay?select=Flight_Delay.parquet

Statistics

Flight Delay Dataset

File	#Tables	#Rows	# Columns	Size (MB)
features_added.parquet	1	30,132,631	40	799.624 MB
Flight_Delay.parquet	1	30,132,672	29	754.104 MB

Total dataset size: 1,553.728 MB

Weather Dataset

File	#Tables	#Rows	# Columns	Size (MB)
daily_weather.parquet	1	27,635,763	14	28.4 MB
cities.csv	1	1,245	8	0.082 MB
countries.csv	1	213	11	0.020 MB

Total dataset size: 28.502 MB

C. Data Integration

1. Loading the Data

- **Tools Used**:
  Used kagglehub and openpyxl to access and load large datasets efficiently.
- **Files Loaded**:
  Loaded weather data (cities.csv, countries.csv, daily_weather.parquet) and flight delay data (Flight_Delay.parquet) with only relevant columns to optimize memory usage.

2. Data Cleaning for Integration

- **City Name Normalization:**
  - Cleaned city names in both datasets for matching by stripping whitespace, converting to lowercase, and removing state abbreviations.
  - flight_df['OriginCityName_clean']and daily_weather_df['city_name_clean'] created for accurate matching.
- **Datetime Conversion:**
  - Converted flight and weather dates to pandas datetime format for accurate merge operations.

3. Merging Datasets

- **Merge Logic:**
  - Merged flight data with weather data based on cleaned origin city name and flight date (left_on=['OriginCityName_clean', 'FlightDate'], right_on=['city_name_clean', 'date']).
  - Left join used to retain all flights, even those without weather match.
- **Post-Merge Statistics:**
  - Original flight data: 30,132,672 records
  - After merge: 30,250,045 records (due to possible duplicate city-date combinations)
  - Successful weather matches: 7,952,755 (~26.77% coverage)

4. Examples of Data Integration Cases

Example 1: Flights WITH Weather Data

Flight Date	Origin City	Departure Delay(min)	Avg Temp (°C)	Precipitation (mm)	Avg Wind Speed (km/h)
2018-01-07	Providence, RI	0.0	-14.6	0.0	15.1
2018-01-02	Des Moines, IA	27.0	-20.5	0.0	19.8
2018-01-06	Des Moines, IA	10.0	-15.7	0.0	15.8

Example 2: Flights WITHOUT Weather Data

Flight Date	Origin City	Departure Delay (min)	Avg Temp (°C)	Precipitation (mm)
2018-01-15	Newark, NJ	43.0	NaN	NaN
2018-01-16	Newark, NJ	81.0	NaN	NaN
2018-01-17	Newark, NJ	1.0	NaN	NaN

Example 3: High Precipitation Flights (Precipitation > 100mm)

Flight Date	Origin City	Departure Delay (mm)	Weather Delay (min)	Precipitation (mm)	Avg Wind Speed (km/h)
2018-10-11	Columbia, SC	31.0	0.0	113.0	28.1
2018-10-11	Columbia, SC	83.0	83.0	113.0	28.1
2018-10-11	Columbia, SC	192.0	192.0	113.0	28.1

City Coverage

Top Coverage (100%):

- Des Moines, IA
- Austin, TX
- Harrisburg, PA
- Santa Rosa, CA
- Tallahassee, FL

Bottom Coverage (0%):

- Butte, MT
- Cedar City, UT
- Allentown/Bethlehem/Easton, PA
- Cedar Rapids/Iowa City, IA

5. Functional Dependency Detection & Examples

Dependency 1: FlightDate + OriginCity → Weather Conditions

Given a specific flight date and origin city, the weather conditions (temperature, precipitation, wind speed) are uniquely determined.

Flight Date	Origin City	Avg Temp (°C)	Precipitation (mm)	Avg Wind Speed (km/h)
2018-01-01	Albany	-17.3	0.0	8.6
2018-01-01	Atlanta	-4.1	0.0	21.2
2018-01-01	Austin	-2.3	0.0	12.2
2018-01-01	Boise	-3.3	0.0	4.3
2018-01-01	Boston	-14.7	0.0	26.3

Dependency 2: OriginCity + DestinationCity → Distance

The pair of origin and destination city determines a fixed distance.

Verification: Each route has 1 unique distance value (should be 1)

Origin City	Destination City	Distance (km)	Unique Values	Flight Count
Aberdeen, SD	Minneapolis, MN	257.0	1	3368
Abilene, TX	Dallas/Fort Worth, TX	158.0	1	8100
Abilene, TX	Houston, TX	307.0	1	543
Adak Island, AK	Anchorage, AK	1192.0	1	338
Adak Island, AK	Cold Bay, AK	616.0	1	94

Dependency 3: Weather Conditions → Weather Delay

Higher precipitation is generally correlated with higher weather delay.

Precipitation Level	Avg Weather Delay (min)	Avg Departure Delay (min)	Flight Count
None/Light (0-1 mm)	0.78	15.14	645,152
Light (1-5 mm)	1.25	17.36	691,022
Moderate (5-10 mm)	1.46	18.70	370,212
Heavy (10-100 mm)	2.88	23.47	604,649

Correlation Matrix

Feature	Departure Delay (min)	Weather Delay (min)
Weather Delay	0.31	1.00
Arrival Delay (min)	0.98	0.31
Precipitation (mm)	0.06	0.05

D. Data Preparation

1. Handling Null Data

1.1 Missing Value Analysis

Our initial analysis revealed significant missing values in the merged dataset, particularly in weather-related features:

Feature	Missing Count	Missing Percentage
Avg Temp (°C)	22,297,290	73.71%
Precipitation (mm)	22,215,998	73.44%
Avg Wind Speed (km/h)	22,332,980	73.83%
Departure Time	0	0.00%

1.2 Null Handling Strategies

Strategy 1: Weather Data Nulls (73.7% missing)

Issue: Approximately 73.7% of flights (22,297,290 records) lack corresponding weather data due to city name matching failures between the flight and weather datasets.

Solution:

- Created a binary indicator variable *has_weather_data* (1 = has weather, 0 = no weather)

- Filled missing weather values with median values:
    - avg_temp_c: 16.6°C
    - precipitation_mm: 0.0mm
    - avg_wind_speed_kmh: 11.9 km/h

Justification:

- Median imputation is robust to outliers and provides reasonable estimates
- The indicator variable preserves information about data quality
- This approach retains all flight records rather than losing 73.7% of data

Results:

- Flights with original weather data: 7,952,755 (26.3%)
- Flights with imputed weather data: 22,297,290 (73.7%)

Example 1: Flight WITH Weather Data (Original)

Flight Date	City	Avg Temp (°C)	Precipitation (mm)	Has Weather Data
2018-01-07	Providence, RI	-14.6	0.0	1
2018-01-15	Newark, NJ	16.6 (median)	0.0 (median)	0

Example 2: Flight WITHOUT Weather Data (Imputed with Median)

Flight Date	City	Avg Temp (°C)	Precipitation (mm)	Has Weather Data
2018-01-15	Newark, NJ	16.6 (median)	0.0 (median)	0

Strategy 2: Cancelled Flights

Issue:

Flights with null DepTime values represent cancelled flights that never departed.

Solution:

Analysis revealed 0 cancelled flights in the dataset.

Impact:

No records were removed (0% of dataset).

2. Outlier Detection

2.1 Methodology

We employed the Interquartile Range (IQR) method to identify outliers:

- Lower Bound = Q1 - 1.5 × IQR
- Upper Bound = Q3 + 1.5 × IQR

2.2 Outlier Detection Results

Feature	Q1	Q3	IQR	Outliers Count	Outliers (%)
Departure Delay (min)	0.0	3.0	3.0	6,388,993	21.1%
Arrival Delay (min)	0.0	0.0	0.0	6,501,114	21.5%
Weather Delay (min)	0.0	0.0	0.0	386,195	1.3%
Precipitation (mm)	0.0	0.0	0.0	2,314,882	7.7%
Avg Wind Speed (km/h)	11.9	11.9	0.0	7,750,871	25.6%

2.3 Outlier Handling Decision

Decision: RETAIN all outliers in the dataset.

Justification:

1. **Real Events:** Extreme delays (>2 hours) represent actual operational events, not measurement errors
2. **Relevance:** Severe weather conditions and their impact on delays are central to our analysis
3. **Validity:** Weather measurements (high precipitation, extreme temperatures) are legitimate data points
4. **Analytical Value:** Extreme cases provide important information for understanding weather-delay relationships

Alternative Approach:

Instead of removing outliers, we created indicator variables:

- **is_severe_delay**: Flags delays exceeding 120 minutes (843,792 cases, 2.8%)
- **is_heavy_precipitation**: Flags precipitation exceeding 25mm (193,348 cases, 0.6%)

This allows models to treat extreme cases specially without losing the information.

2.4 Examples of Outliers

Example 1 - Severe Delays

Flight Date	City	Departure Delay (min)	Precipitation (mm)	Weather Delay (min)
2018-01-12	Greer, SC	520	0.0	0

Example 2 - Heavy Precipitation

Flight Date	City	Precipitation (mm)	Departure Delay (min)	Weather Delay (min)
2018-01-08	Sacramento, CA	64.8	22	N/A

Example 3 - Extreme Temperature

Flight Date	City	Avg Temp (°C)	Departure Delay (min)	Weather Delay (min)
2018-07-25	Phoenix, AZ	40.6	77	N/A

3. Data Transformation

3.1 Temporal Feature Engineering

***Purpose*:** Capture seasonal patterns and day-of-week variations in flight operations and weather conditions.

Features Created:

- **Month (1-12):** Captures seasonal weather patterns (winter storms, summer thunderstorms)
- **Day Of Week (0-6):** Captures weekly operational patterns (0=Monday, 6=Sunday)
- **IsWeekend (0/1):** Binary indicator for weekend vs weekday operations

Justification:

- Flight delays exhibit strong seasonal patterns due to weather variations
- Operational patterns differ between weekdays and weekends
- These features are commonly significant in delay prediction models

Distribution:

- Weekend flights: 8,278,029 (~27.4% of total)
- Weekday flights: 21,972,016 (~72.6% of total)

3.2 Delay Categorization

***Purpose*:** Create interpretable severity levels for delay analysis and classification tasks.

Categories Defined:

- **On-Time:** ≤ 0 minutes (early or on-time departures)
- **Minor:** 1-15 minutes delay
- **Moderate:** 16-60 minutes delay
- **Severe:** > 60 minutes delay

Distribution:

- **On-Time:** 21,279,437 (70.3%)
- **Minor:** 3,755,520 (12.4%)
- **Moderate:** 3,066,604 (10.1%)
- **Severe:** 2,148,484 (7.1%)

Justification:

- Industry-standard thresholds for delay severity
- Useful for both classification models and stakeholder communication
- Enables analysis of weather impact across different delay severities

3.3 Feature Scaling (Normalization)

***Purpose*:** Standardize feature scales for machine learning algorithms sensitive to feature magnitude.

Method: StandardScaler (z-score normalization)

- **Formula:** $z = (x - \mu) / \sigma$
- **Result:** mean = 0, standard deviation = 1

Features Scaled:

- **Distance**
- **Average Temperature (°C)**
- **Precipitation (mm)**
- **Average Wind Speed (km/h)**

Justification:

- Many ML algorithms (e.g., logistic regression, neural networks, SVM) perform better with normalized features
- Prevents features with large magnitudes from dominating the model
- Maintains the distribution shape while standardizing scale

Example:

Feature	Original Value	Scaled Value
Precipitation (mm)	0.0	-0.15
Average Temperature (°C)	16.6	0.04

4. Summary

4.1 Data Preparation Statistics

Final Dataset:

- Total records: 30,250,045
- Total features: 36
- Flights with matched weather data: 7,952,755 (26.3%)
- Records with complete key features: 30,250,045 (100%)

4.2 Features Created for Modeling

Indicator Variables:

1. **has_weather_data** - Weather data availability flag
2. **is_severe_delay** - Severe delay indicator (> 120 min)
3. **is_heavy_precipitation** - Heavy rain indicator (> 25mm)

Temporal Features:

1. **Month** - Month of the year (1-12)
2. **DayOfWeek** - Day of the week (0-6)
3. **IsWeekend** - Weekend indicator (0/1)

Categorical Features:

1. **DelayCategory** - Delay severity classification (OnTime/Minor/Moderate/Severe)

Scaled Features:

1. **Distance (Scaled)**
2. **Average Temperature (Scaled)**
3. **Precipitation (Scaled)**
4. **Average Wind Speed (Scaled)**

4.3 Data Quality Assessment

Completeness:

- All key features now have no missing values
- Weather data gaps addressed with imputation and flagging

Consistency:

- Outliers retained but flagged for special treatment

- All features properly scaled and encoded

Readiness:

- Dataset is prepared for exploratory data analysis
- Features are ready for machine learning model development
- Temporal and categorical features enable comprehensive analysis

4.4 Output

The prepared dataset was saved as **merged_data_prepared.csv** with all transformations applied and ready for modelling and analysis phases.

E. Report a set of tools or systems you plan to use

To ensure a comprehensive workflow, our team will leverage a combination of data processing, analysis, storage, and collaboration tools.

Data Storage and Management

- ***CSV Files and Pandas Dataframes*:**
  Used for initial data collection, exploration, and transformation. CVS provides a simple exchange format, while Pandas enables efficient in-memory data manipulation.
- ***MySQL Database*:**
  Used for structured storage and integration of our multiple datasets. MySQL's relational model supports the enforcement of relationships between flight and weather data. This makes it easier to query and manage the integrated data.

Development and Collaboration Environments

- ***Google Collab*:**
  A cloud-based Python environment that supports shared notebooks, facilitating team collaboration and reproducibility without local setup issues.
- ***GitHub*:**
  Serves as our central platform for version control, code management, and collaborative development. It allows our team members to track progress, review code, and manage issues.

Analysis and Modelling Tools

- ***Python*:**
  The primary language for our analysis and modelling.
- ***Jupyter Notebooks*:**
  Used to document analysis steps, visualize results, and maintain a clear workflow.