# FlightAnalysis Project Documentation

## 1. Introduction

The FlightAnalysis Project aims to analyze airline performance, delays, and operational efficiency using SQL (MSSQL) for data processing and Power BI/Tableau for visualization. Git was used for version control throughout the project.  
  
The database includes three key tables—airlines, airports, and flights—with more than 5.8 million flight records in total.  
  
GitHub Repository: <https://github.com/PARADOXop/U.S.-Airline-Performance-Delay-Analysis>

## 2. Data Cleaning

* Assessed null values across all tables.
* Verified and recalculated departure\_delay\_min and arrival\_delay\_min.
* Identified approximately 14,000 rows with NULL values in flight\_number.
* Missing departure\_time and arrival\_time values were primarily due to cancellations or diversions.
* Delay type columns (air\_system\_delay, weather\_delay, etc.) did not always match the total recorded delay.
* A decision is pending on whether to drop rows with missing flight\_number.
* No significant missing values were found in critical fields after cleaning.

## 3. Data Transformation

* Converted HHMM string columns (scheduled and actual times) into DATETIME format for accurate time calculations.
* Created a new flight\_date column using the year, month, and day columns, then dropped those original columns.
* Defined delay categories for both arrival and departure delays: On-Time (≤0 min), Short (1–30 min), Medium (31–120 min), Long (>120 min).
* Created a readable cancellation category based on cancellation\_reason: A → Air System, B → Security, C → Airline, D → Weather.
* Added a flight status column for better readability: Cancelled, Diverted, or Completed.
* Dropped redundant HHMM columns after conversion.
* Recomputed departure\_delay\_min and arrival\_delay\_min as time differences between scheduled and actual times.
* Performed transformations in batches due to large data size (~5.8M rows).

## 4. Exploratory Data Analysis (EDA)

General Observations:

* Security and Air System issues account for roughly 82% of total cancellations.
* Airline, aircraft, and air-system factors cause most delays in completed flights.
* Around 10.6% of flights departed late but still arrived on time.

Airline Analysis:

* Top 3 Airlines by On-Time Arrivals: Hawaiian Airlines (89.5%), Alaska Airlines (87.66%), Delta Airlines (87%).
* Top 3 Airlines by On-Time Departures: Alaska Airlines (74.71%), Hawaiian Airlines (73.54%), SkyWest Airlines (70.44%).
* Top 3 Airlines with Highest Cancellation Rates: Alaska Airlines, American Airlines, Atlantic Eagle Airlines.

Airport Analysis: Insights at the airport level were limited in the initial EDA and will be explored further in the Power BI dashboard.

## 5. KPI and Dashboard Insights

System reliability remains high overall, with a cancellation rate of approximately 1.5% and minimal diversions. However, delays tend to peak during June and July, with smaller spikes in February and December. This indicates a need for additional resource planning during summer and winter months.  
  
Delays in departure almost always lead to delayed arrivals. Once departure delays exceed 30 minutes, arrival delays increase sharply, and cancellation risks rise. This underlines the importance of early gate-level interventions.  
  
Chicago and Atlanta airports report the highest cancellation rates, followed by DFW, Houston, Denver, LAX, and New York. These airports also experience longer taxi times, suggesting potential congestion at gates or on ramps.  
  
Seasonal traffic trends show increased flight volumes from March through August, with a significant drop afterward. Most delays are short, but long delays, though fewer, have a disproportionate operational impact.

## 6. Executive Takeaways

* Overall reliability is strong but vulnerable during peak summer months; reducing long delays can significantly cut cancellations.
* Focus on Chicago and Atlanta to reduce operational bottlenecks and coordinate efforts with security and ATC programs.
* Shift performance evaluation from absolute counts to rate-based metrics, emphasizing long-delay-to-cancellation conversion.
* Implement rotation health checks to identify aircraft with multiple tight turnarounds and insert buffer time proactively.
* Adopt risk-based boarding for flights at high risk of delay to protect against cascading schedule disruptions.
* Develop station-specific playbooks for weather-affected hubs like Chicago and Atlanta to optimize flight rescheduling.
* Introduce a taxi-time leaderboard to monitor and address rising taxi-in/out averages by airport.
* Run quarterly improvement sprints focusing on dominant cancellation causes such as security or air traffic issues.
* Protect key connections by allowing short, strategic delays in feeder flights during peak operational stress.