Flight Delay and Cancellation Analysis (2019-2023)

Bishal Rauniyar Devansh Trivedi Manish Thapa

1 GitHub

https://github.com/Thapamanish/flight-delay-prediction

2 Abstract

Delays and cancellations of flights have been an ongoing source of cost and frustration for aviation in the U.S. In this project, we analyze a large-scale dataset comprising about 3 million domestic flights operated between January 2019 and August 2023. It covers 18 airlines and approximately 380 origin and destination airports.

The flow is consciously database-first: all preprocessing, feature engineering, and exploratory statistics reside as SQL queries running on DuckDB, with Python used mostly for visualization and modeling. We undertake an elaborate exploratory data analysis (EDA) to see how delays sit by carrier, time, airport, and cause. Thereafter, two classification scenarios are developed using logistic regression and Random Forest models. The pre-flight model, i.e., based only on information known at booking time, and the real-time model that also uses actual departure delay.

Pre-flight, the Random Forest gets an AUC of 0.653. Modest, but it gives actionable risk estimates. Real-time gets an AUC of 0.956 with 91% precision and 77% recall. This proves that departure delay is the dominant driver of arrival delay. Feature-importance analysis shares with us that about 84% of the predictive power of this model comes from departure delay. This sits well in consistency with our EDA finding that late aircraft cascading is the primary delay cause.

3 Introduction

Millions of passengers are affected by flight delays and cancellations, creating huge operational and financial challenges to the airlines and airports. These small disruptions typically snowball into a tightly coupled schedule system, increasing impacts from weather events, congestion, and maintenance issues. Therefore, knowing where, when, and why delays happen is fundamental toward any effort aimed at improved reliability and enhanced passenger experience.

1. We've describe patterns of flight delays and cancellations across airlines, airports, and time periods, including the impact of the COVID-19 pandemic.
2. Quantify the contribution of different delay causes (late aircraft, carrier, weather, National Airspace System (NAS), security).

3. Have build and evaluate classification models to predict whether a flight will arrive late, under two scenarios:

   3.1. Pre-Flight Prediction: only booking-time information is available.

   3.2. Real-Time Prediction: actual departure delay is also known.

We do not use pandas as our preferred engine for analysis, but rather apply a database-centric design. DuckDB offers in-process SQL analytics scaling up to millions of rows, maintaining the same expressiveness and reproducibility that SQL provides.

4 Data and Infrastructure

4.1 Dataset Overview

- **Source:** U.S. Department of Transportation, Bureau of Transportation Statistics (On-Time Performance data, accessed via Kaggle).
- **Time period:** January 2019 - August 2023
- **Size:** 3,000,000 sampled flight records, 32 attributes
- **Scope:** Domestic flights within the United States

Key attributes

- **Flight details:** flight date, airline, flight number, origin and destination airports
- **Operational status:** completed, delayed, cancelled, or diverted
- **Delay metrics:** arrival and departure delay in minutes, plus cause-specific delay minutes (carrier, weather, NAS, security, late aircraft)
- **Cancellation reasons:** categorical codes for the primary cancellation cause
- **Time fields:** scheduled and actual departure/arrival timestamps

Simple descriptive statistics from the unprocessed table indicate that these 3 million flights cover 18 airlines and approximately 380 unique origin as well as destination airports. The mean value for great-circle distance is about 809 miles while the average arrival delay for all flights (including those arriving early) stands at 4.26 minutes.

4.2 Database-First Workflow with DuckDB

All core data operations are implemented in SQL using DuckDB:

1. Raw CSV files are loaded into a DuckDB table flights_raw.
2. Data quality checks and summary statistics are computed via SQL.
3. A cleaned table flights_clean is created with:
   - filtered and validated flights,

- derived date parts (year, month, day of week),
- distance and duration fields,
- and a binary label IS_DELAYED indicating arrival delay ≥ 15 minutes.

4. A separate table delay_reasons aggregates cause-specific delay minutes (carrier, weather, NAS, security, late aircraft).

Python acts mostly as a client: it runs SQL through DuckDB, gets results converted to DataFrames, and visualizations outputted by Matplotlib/Seaborn. This is the best way to reflect real production data-engineering patterns while also keeping the notebook tight and reproducible.

5 Data Quality and Preprocessing

5.1 Missing Data Assessment

We start by checking the level of missingness in some important columns (such as arrival delay, delay causes, cancellation codes, and airline identifiers).

- Cause-specific delay fields (DELAY_DUE_CARRIER, DELAY_DUE_WEATHER, DELAY_DUE_NAS) and CANCELLATION_CODE have high missingness (≈80-97%), because they are populated only when a flight is delayed or cancelled.
- Core fields such as airline and flight date have no missing values.

This pattern is expected given the data collection process, we therefore treat cause fields as conditionally observed rather than globally missing.

5.2 Dataset Summary and Cleaning

Using SQL, we compute:

- total number of flights,
- number of unique airlines and airports,
- average distance and arrival delay,
- counts of cancelled and diverted flights.

We then create a cleaned dataset flights_clean with the following steps:

1. **Filter to valid records**
   - Keep domestic, non-null flights with valid time and distance fields.
2. **Create derived columns**
   - Year, month, and day-of-week from flight date
   - Distance bands (e.g., <500 miles, 500-1000, 1000-1500, >1500)
   - IS_DELAYED = 1 if arrival delay ≥ 15 minutes, else 0.
3. **Prepare for classification**
   - Categorical features (airline, origin, destination, time buckets) are kept as codes to be one-hot encoded later in the modeling notebook.

6 Exploratory Data Analysis

6.1 Airline and Destination Composition

We first examine which airlines and airports dominate the sample.

- A bar chart of the top 15 airlines by flight count shows that Southwest, Delta, American, SkyWest, and United account for a large share of total flights, with Southwest alone operating nearly one-fifth of all flights in the sample.
- A scatter/bubble chart of top destination airports (by total flights and delay rate) highlights major hubs such as ATL, DEN, LAX, and ORD. Some airports combine high traffic with above-average delay rates, making them critical nodes for network reliability.

Interpretation: Delays at a small number of large carriers and hub airports have disproportionate impact, because they affect a large fraction of travelers and aircraft rotations.

6.2 Day-of-Week Patterns

Using SQL, we compute for each day of the week:

- total flights,
- number of delayed flights,
- delay rate (% of flights delayed),
- average delay minutes among all flights.

Results show:

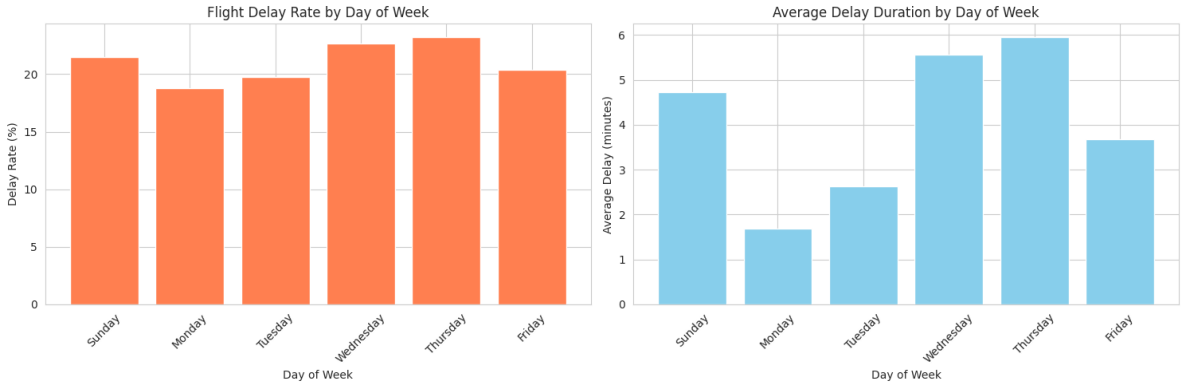

Figure 1: Delay rate and average delay by day of the week.

- Delay rates range from about 19-23% across weekdays.
- Wednesday and Thursday exhibit the highest delay rates (∼22-23%) and the longest average delay durations (∼5.5-6 minutes).
- Monday has relatively lower delay rates and shorter average delays.

Interpretation: Mid-week congestion and tightly packed schedules may contribute to higher delays, while Mondays benefit from "reset" schedules and lower residual cascading.

6.3 Annual and Monthly Trends (COVID-19 Impact)

Yearly aggregation reveals:

- In 2019, about 22% of flights were delayed, with total delay minutes around 3.9 million.
- In 2020, during the peak of COVID-19 restrictions, traffic collapsed (≈450K flights vs 740K in 2019), delay rates dropped to about 12%, and net arrival delays became negative on average (flights tended to arrive early).
- By 2022-2023, traffic and delays not only recovered but exceeded 2019 levels, with total delay minutes peaking at roughly 4.6 million minutes in 2022 and high delay rates continuing into 2023.

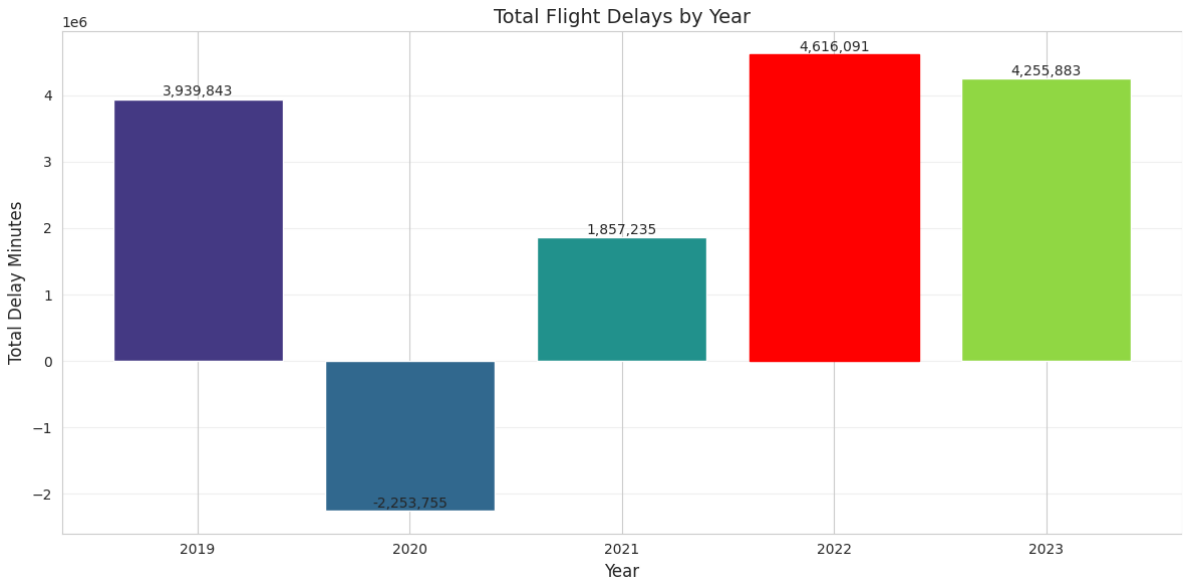

Figure 2: Total annual delay minutes, highlighting the 2020 dip and the 2022 peak.

Monthly trend plots also illustrate the strong summer peak and relatively quiet winter months, with 2020 being an exceptionally low-delay year.
Interpretation: The pandemic dramatically reduced both volume and delays in 2020. As demand rebounded faster than infrastructure and staffing in 2022-2023, congestion and delays increased beyond pre-pandemic levels.

6.4 Delay Causes

Aggregating cause-specific delay minutes across the dataset, we obtain the breakdown shown in Figure [Delay Causes]:

- Late Aircraft: 13.6 million minutes (37.7% of all delay minutes)
- Carrier-related: 13.2 million minutes (36.7%)
- NAS (airspace/ATC): 7.0 million minutes (19.5%)

- Weather: 2.1 million minutes (5.9%)
- Security: 0.08 million minutes (0.2%)

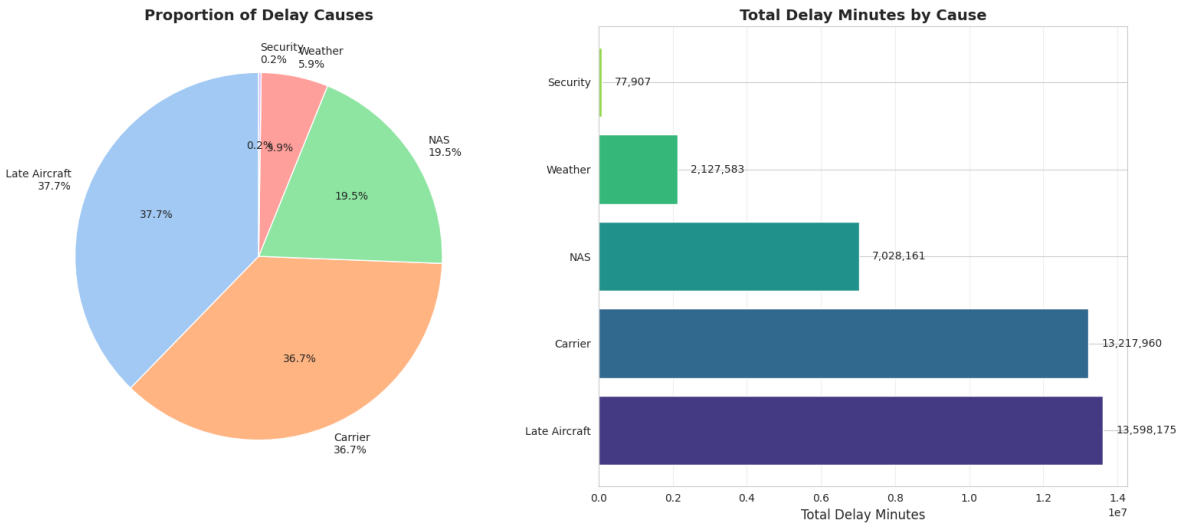

Figure 3: Visualizes the share of each delay cause, with late aircraft and carrier-related delays dominating.

Carrier and late-aircraft delays plummeted in 2020, rose again in 2021, reached a peak in 2022, and slightly improved in 2023.These were the two large components of total delay. Other smaller components, NAS and weather, followed similar albeit smaller magnitude patterns.
Interpretation: Late aircraft and carrier operational issues summed up almost three-fourths of the total delay minutes. Since late aircraft delays come from previous flights, it serves as a very good indicator of cascading disruptions—which will later on prove to be the most important feature in our predictive models.

6.5 Routes and Distance

Advanced SQL queries identify:

- The most problematic routes by delay rate are the top 20. Major hubs or weather-sensitive airports have many of them.
- There is a bar chart of delay rate by distance band. Medium to long haul flights have slightly higher delay rates than very short flights. The differences though, are modest.

Interpretation: Some origin-destination pairs just keep being unreliable, so adding some extra buffer time or adjusting the schedule might really help.

7 Predictive Modeling

The EDA shows strong patterns about delays in association with late aircraft departure or late aircraft arrival. To measure how predictable delays are, we set up two classification scenarios in a different notebook for modeling inside Google Colab.

7.1 Problem Definition and Features

We define a binary label:

$$Y = 1 \text{ if arrival delay } \geq 15 \text{ minutes, and } Y = 0 \text{ otherwise.}$$

Two feature sets are considered:

1. **Scenario A - Pre-Flight Prediction**
    - Only information available at or before booking time:
    - airline,
    - origin and destination airports,
    - scheduled departure time (binned into hour-of-day),
    - day of week,
    - month / season,
    - distance band.
2. **Scenario B - Real-Time Prediction**
    - All features from Scenario A, plus actual departure delay in minutes (or a binned version, e.g., 0-5, 5-15, >15 minutes).

Data is split into training and test sets (e.g., 80/20). Categorical variables are one-hot encoded, continuous variables are standardized where appropriate.

7.2 Models and Evaluation Metrics

We train two model families in scikit-learn:

- **Logistic Regression** – a simple linear baseline for comparison.
- **Random Forest** – our main non-linear model, with separate pre-flight and real-time versions.

We evaluate models using:

- Area Under the ROC Curve (AUC),
- Accuracy,
- Precision,
- Recall, and
- F1-score.

7.3 Scenario A - Pre-Flight Prediction

The booking-time features only result in an AUC of 0.653 from the Random Forest on the held-out test set, which is approximately 30% better than random guessing (AUC 0.5). It carries a moderate overall accuracy and F1-score because most delays are driven by factors that occur during the day-of-operations, information not available at the time of booking.

Logistic regression performs slightly worse than Random Forest, thereby confirming that there do exist some nonlinear interactions between airline, route, and schedule. However, it should be noted that flights from both these models present probabilistic risk scores which can readily be used in ranking flights by their expected delay risk.

The pre-flight Random Forest achieves a ROC AUC of approximately 0.65, which confirms that using only pre-flight information provides limited ability to separate delayed flights from on-time flights.

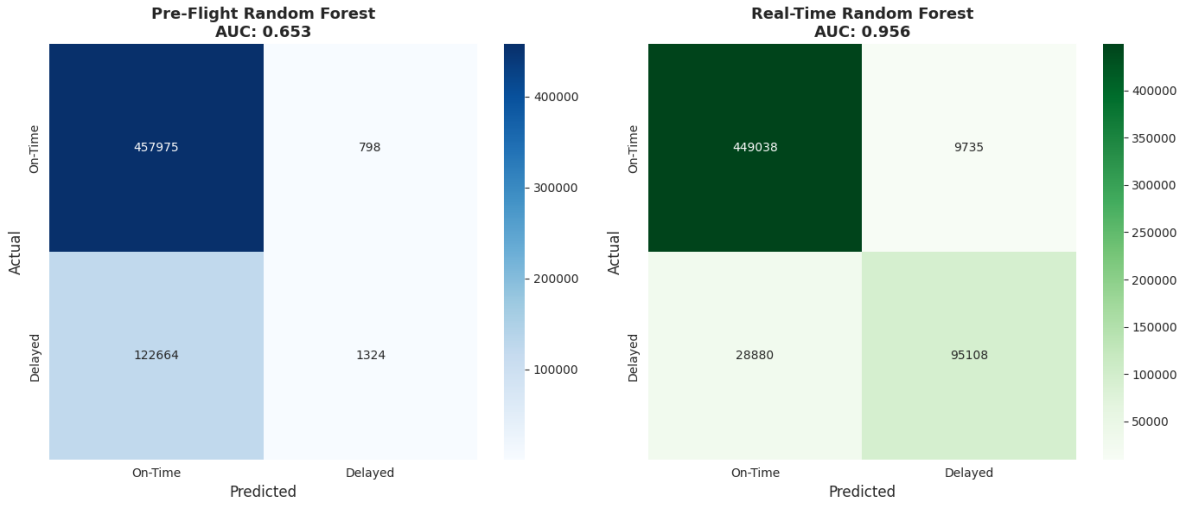

Figure 4: Confusion matrices for the pre-flight (left) and real-time (right) Random Forest models. The titles show the ROC AUC for each setting, illustrating the large gain in discriminative power once real-time departure delay is available.

As shown in Figure 4 above, the real-time model correctly captures most delayed flights while keeping a low false-positive rate.

Interpretation: Even when fed with little information, if it can beat random odds at distinguishing between low-risk and high-risk flights, then perhaps this is the data that passengers need to select less risky itineraries, or airlines want to audit a route when a model continually flags it with high predicted delay probability.

7.4 Scenario B - Real-Time Prediction

When we add departure delay as a feature, model performance improves dramatically:

- The Random Forest reaches an AUC of 0.956.

- On the test set it achieves 91% precision and 77% recall for predicting delayed arrivals.

A confusion-matrix plot shows that the model has been able to pick up the overwhelming majority of delayed flights with relatively low false alarm rates.

In contrast, the real-time Random Forest reaches a ROC AUC of about 0.96, indicating that once departure delay is observed the model can almost perfectly distinguish delayed vs. on-time arrivals. This large jump from roughly 0.65 to 0.96 AUC highlights how powerful real-time information is compared to purely pre-flight features.

Interpretation: Once a flight is late departing, knowing how late it was at departure pretty much fully determines whether it will arrive late. The about 30-point AUC improvement from adding this single feature quantifies cascading effects of late departures on arrival delays and validates our EDA finding that late aircraft delay is the dominant cause.

7.5 Feature Importance

A feature-importance analysis of the Random Forest in Scenario B shows that:

- Departure delay alone accounts for about 84% of the model's total importance score.
- The remaining signal is distributed across airline, route, time-of-day, and day-of-week features.

This proves that operational condition at departure massively inspires arrival delays: if a flight leaves the gate in way late, then most probably it will also arrive late.
Operational implication: Interventions that may prevent or reduce early-day departure delays- that may include ground-time buffers, and improved turnaround processes as well as the prioritization of on-time departures for certain aircraft- would have multiplicative benefits by preventing delay cascades throughout the network.

8 Discussion

Our analysis leads to several key observations:

1. **Concentration of impact**
   A relatively small number of airlines, airports, and routes account for a large share of total flight volume and delays. Improvements targeted at these nodes could yield outsized benefits.
2. **Pandemic dynamics**
   The COVID-19 period of 2020 had half the flights that were seen in 2019, and delays were sharply reduced, average arrival times even becoming slightly late on net. As demand returned, not only did delays return but they came in greater than pre-pandemic levels indicating that staffing as well as infrastructure scale back lagged passenger demand.
3. **Cascading delays dominate**
   It is the late aircraft and carrier issues that together account for about seventy-five percent of the total minutes of delay, the late aircraft alone contributes 37.7% percent. This tallies with the modeling result which has departure delay as by far the most important feature in predicting late arrivals.

4. **Predictability depends on information horizon**
   At booking time (Scenario A), delays are only partially predictable, an AUC of 0.653 indicates useful but limited discrimination. In real time (Scenario B), once departure delay is observed, arrival delay becomes highly predictable (AUC 0.956), emphasizing the importance of real-time monitoring and dynamic decision-making.

9 Conclusion and Future Work

9.1 Conclusion

In this project we:

- Assembled and cleaned a large sample of 3 million U.S. domestic flights from 2019-2023 using a database-first DuckDB and SQL pipeline.
- Performed extensive exploratory data analysis to understand patterns of delays by airline, airport, time, and cause, including the impact of COVID-19.
- Quantified how late aircraft and carrier delays dominate total delay minutes and how certain routes and distance bands are particularly delay-prone.
- Built two classification scenarios (pre-flight and real-time) using logistic regression and Random Forest models, demonstrating that:
  - booking-time information provides modest predictive power, and
  - adding departure delay yields near-perfect discrimination of delayed arrivals.

Overall, the results highlight both structural patterns (airlines, routes, causes) and highly local, operational factors (departure delays) that drive arrival performance.

9.2 Future Work

Several extensions are natural:

- Incorporate external data sources, such as detailed weather observations or air-traffic control restrictions, to better explain variability not captured by schedule and carrier features.
- Explore more advanced models (e.g., gradient boosting, XGBoost, or calibrated probability models) and compare them systematically with logistic regression and Random Forest.
- Build a dashboard or web app that exposes real-time risk scores for flights, using the Scenario B model as a backend.
- Perform what-if simulations quantifying how reducing early-day departure delays or adding buffer time on specific routes would impact downstream delays.