Predicting Flight Delays Using 2024 US Flight Data

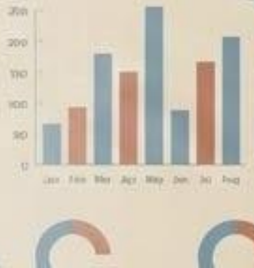
Dataset Stats

7,079,081 flights

348 airports

15 airlines

Full year **2024**



The Problem:

7+ million US flights in 2024

Flight delays cost airlines $1000+ per incident, passengers $100+ in disruption

Question: Can we predict delays BEFORE departure using only pre-flight information?

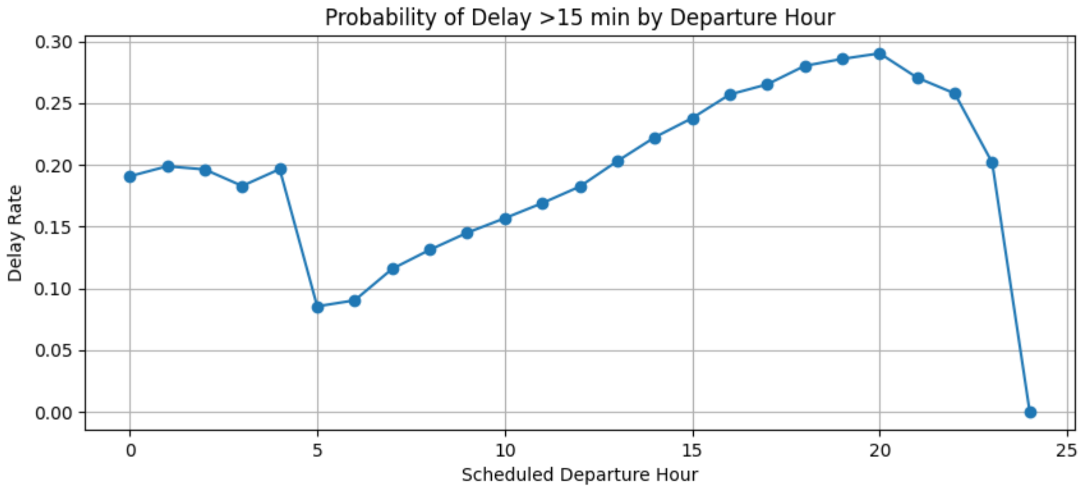
Data Overview

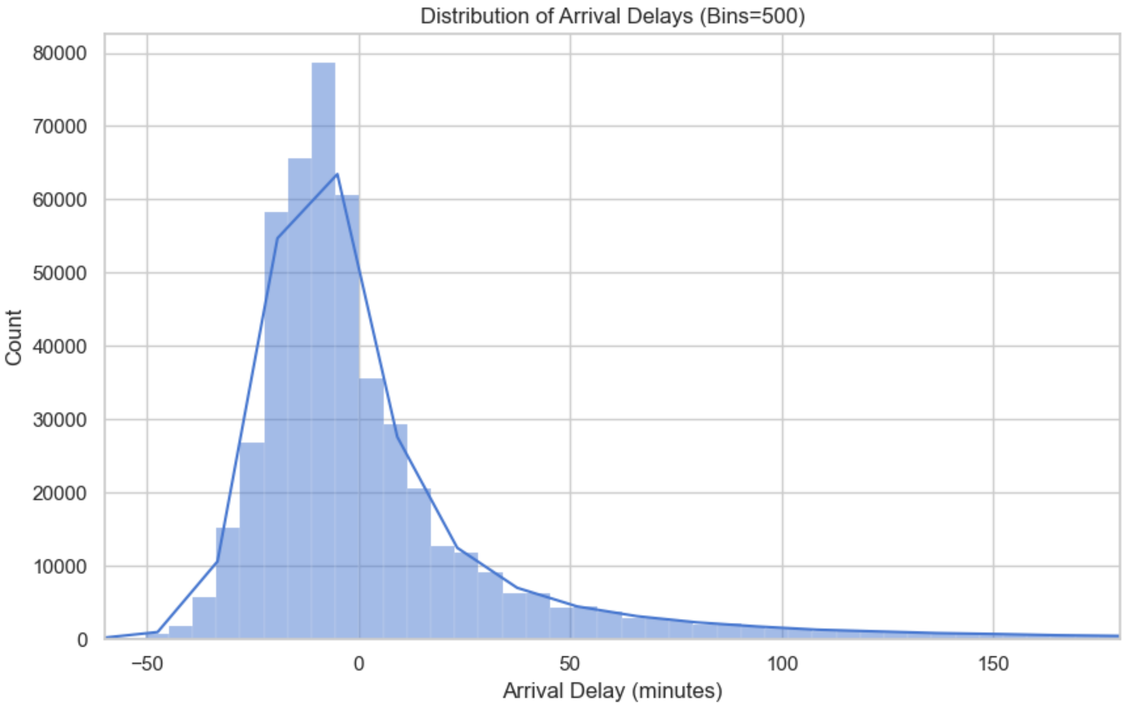
Dataset Highlights:

- **35 features:** timing, delays, airports, carriers, weather impacts
- **Challenge:** only 20% of flights delayed >15 min (imbalanced classes)

Key Insight:

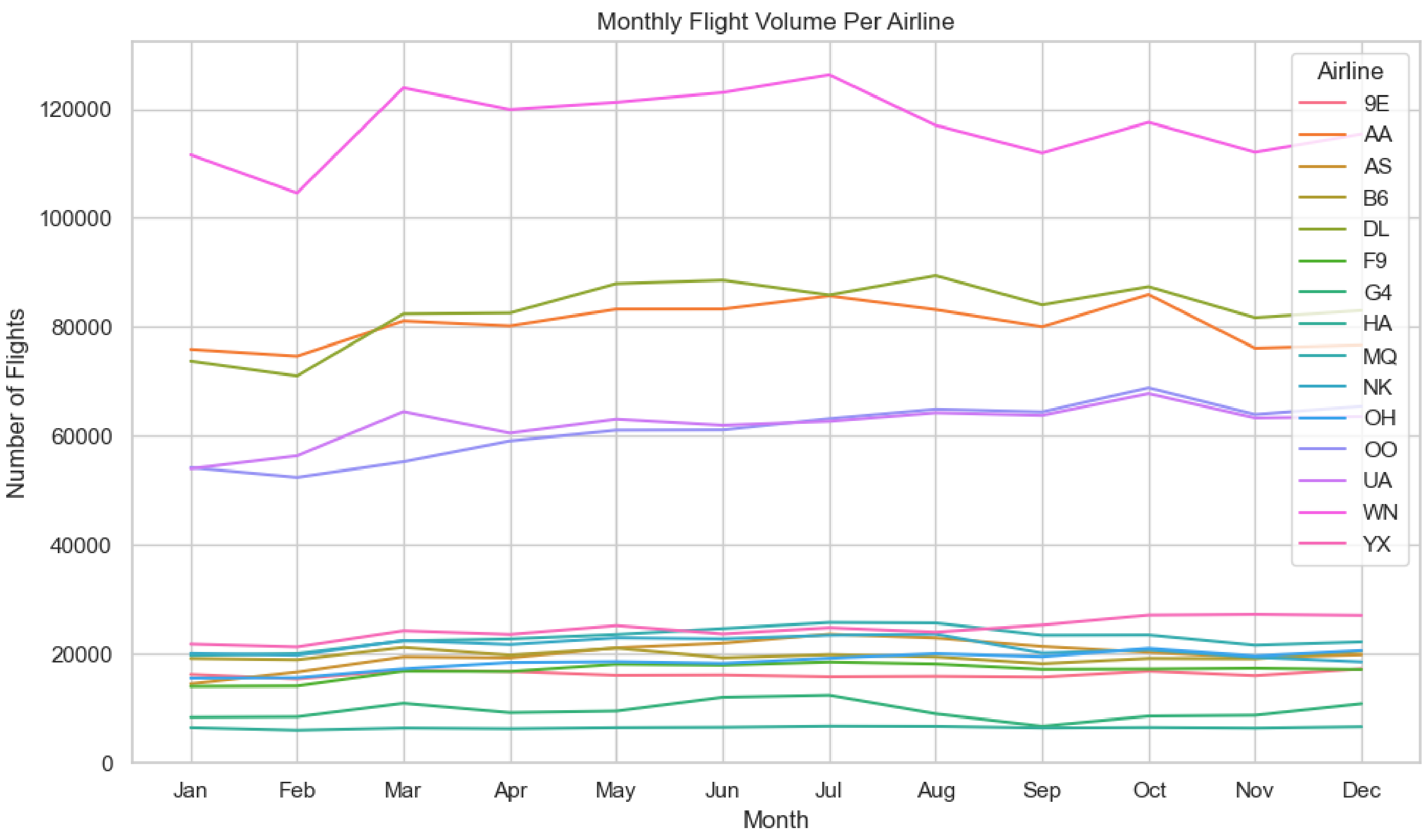
Most flights are on time, but when delays happen they cascade through the system

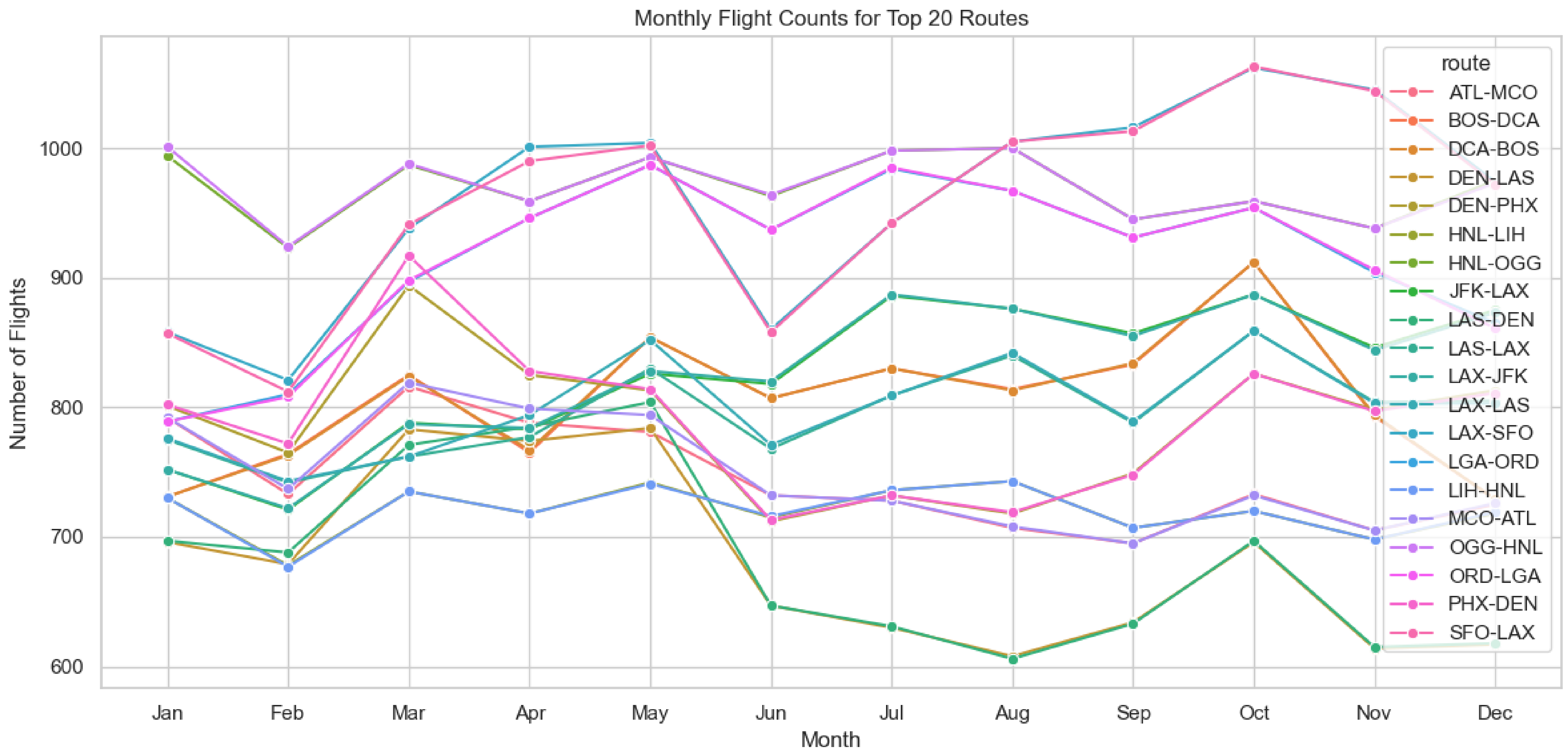




General Data

- Major carriers dominate overall traffic
- July shows the highest combined activity, showing summer travel demand
- Smaller regional airlines maintain stable low volume per month with minimal fluctuation



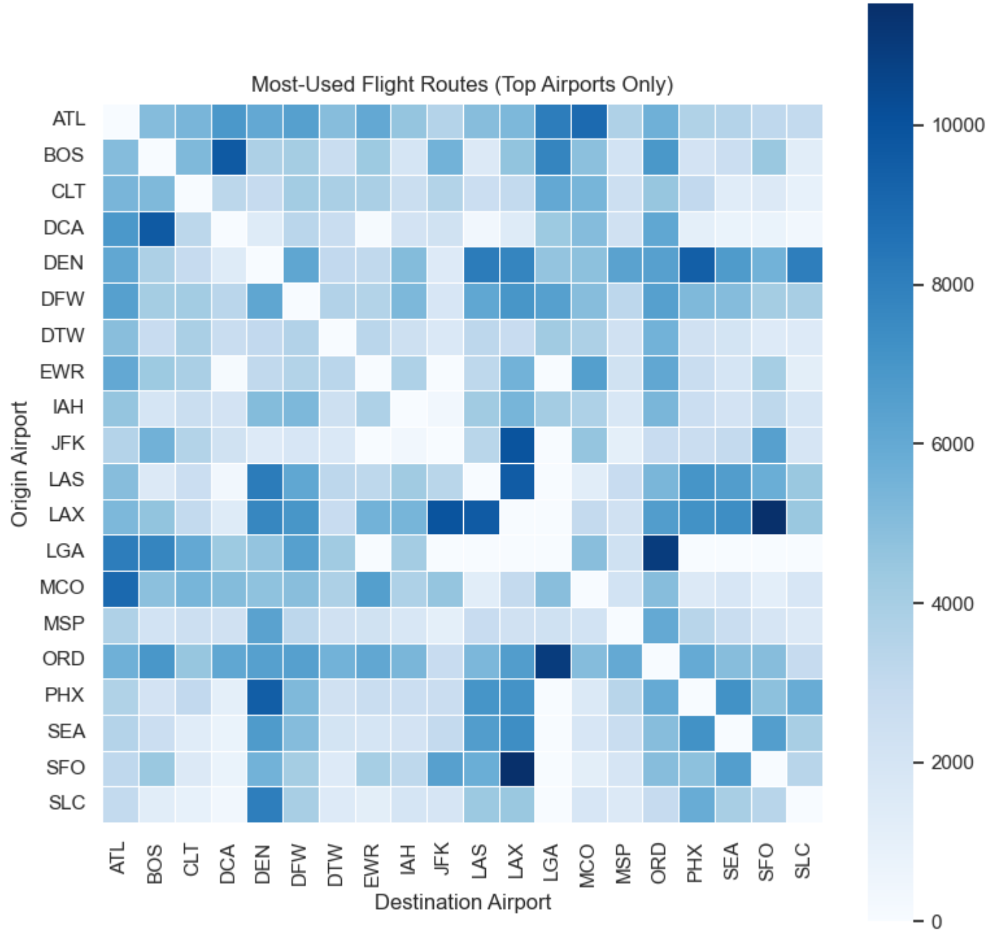


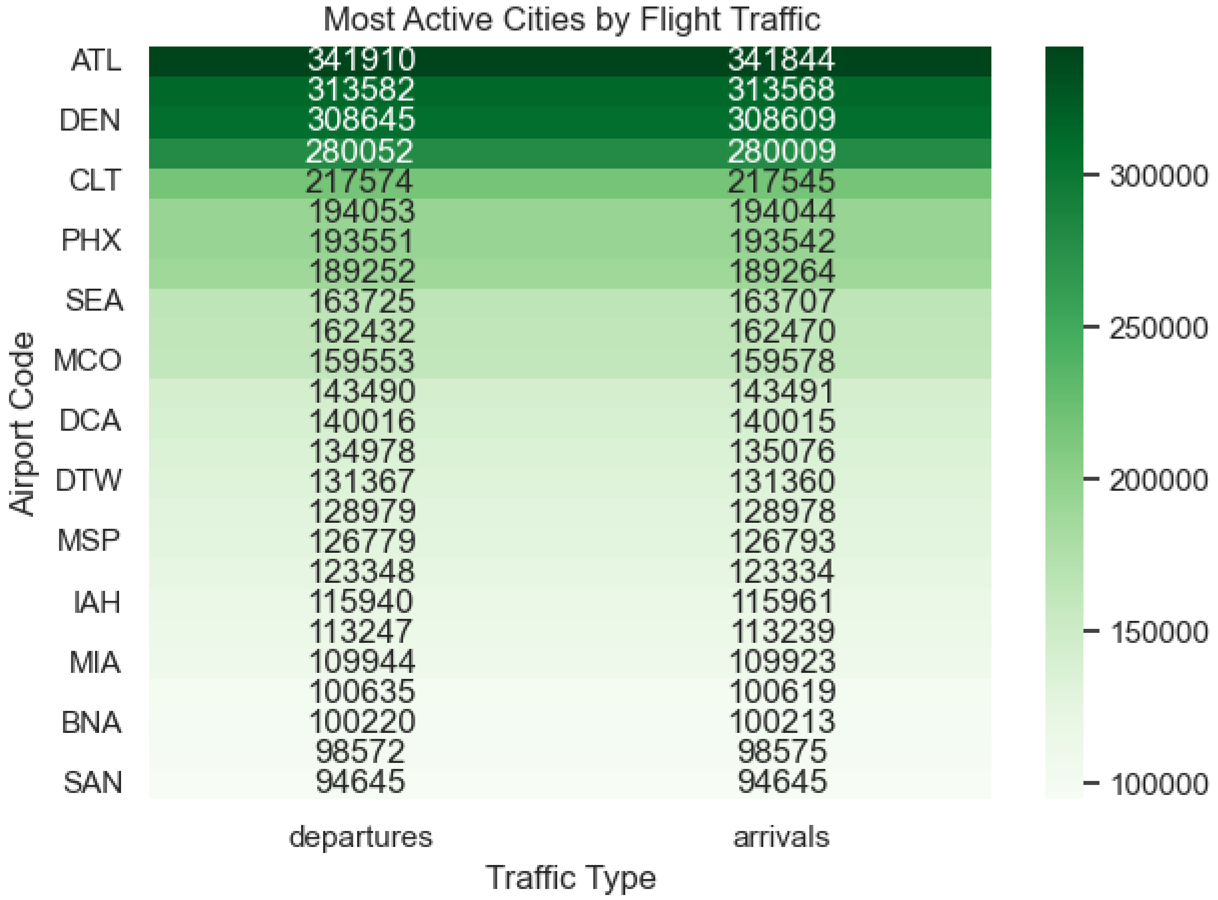
Southern destinations surge in winter demand:

- **ATL–MCO** *(Atlanta ↔ Orlando, Florida)*
- **MCO–ATL** *(Orlando ↔ Atlanta)*
- **PHX–DEN** *(Phoenix, Arizona ↔ Denver, Colorado)*

Northeast business routes peak outside of winter

- **BOS–DCA** *(Boston ↔ Washington, D.C.)*
- **DCA–BOS** *(Washington, D.C. ↔ Boston)*
- **LGA–ORD** *(New York LaGuardia ↔ Chicago O'Hare)*





Major hub airports dominate U.S. air traffic activity

- **ATL** – Atlanta, GA
- **DEN** – Denver, CO
- **ORD** – Chicago O’Hare, IL
- **CLT** – Charlotte, NC

High-density route corridors cluster around:

- **Los Angeles** – LAX (CA)
- **New York** – JFK & LGA (NY)
- **Chicago** – ORD (IL)
- **Atlanta** – ATL (GA)

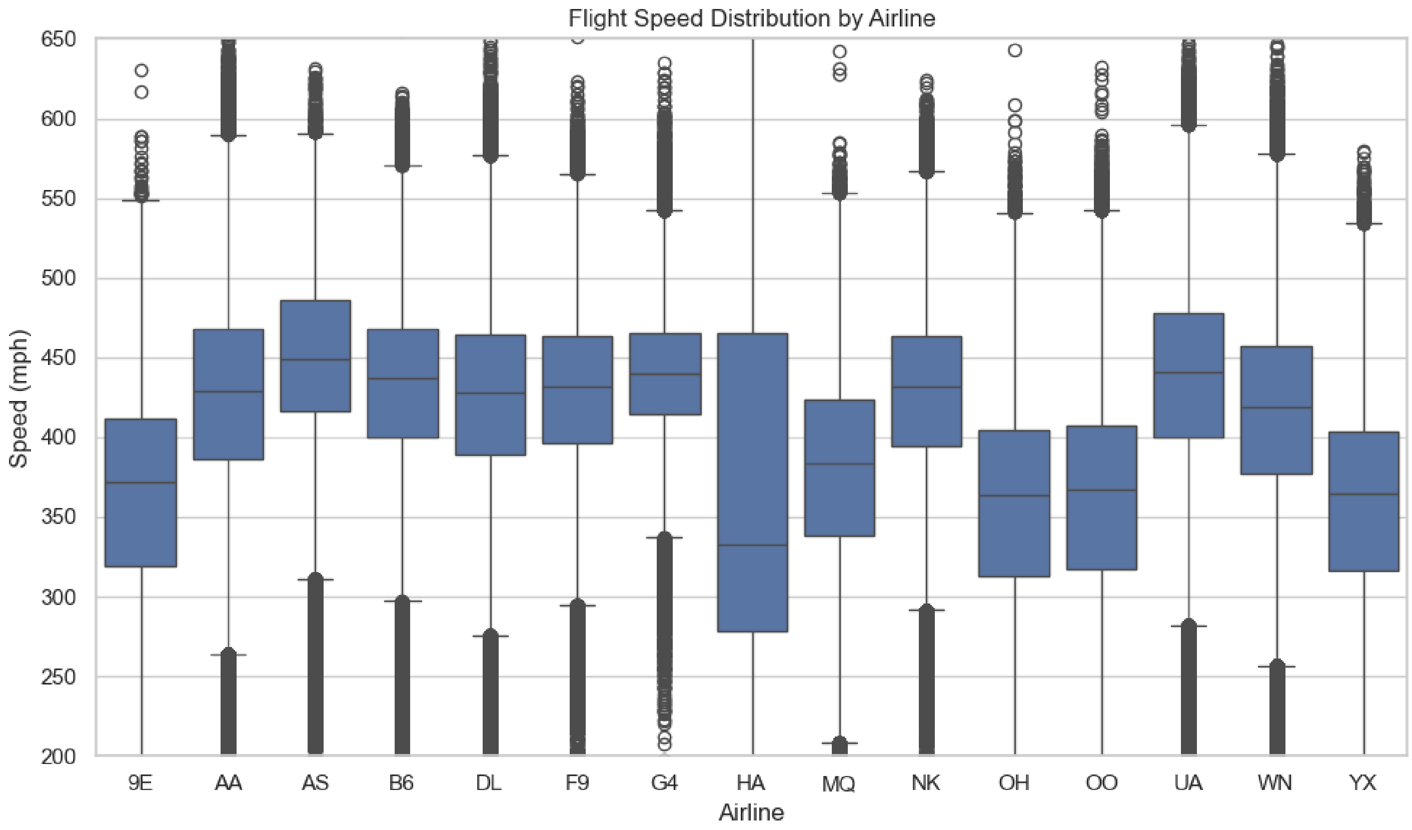
Airliner Speed

Legacy/mainline airlines show higher median speeds due to longer average routes and jet-only fleets:

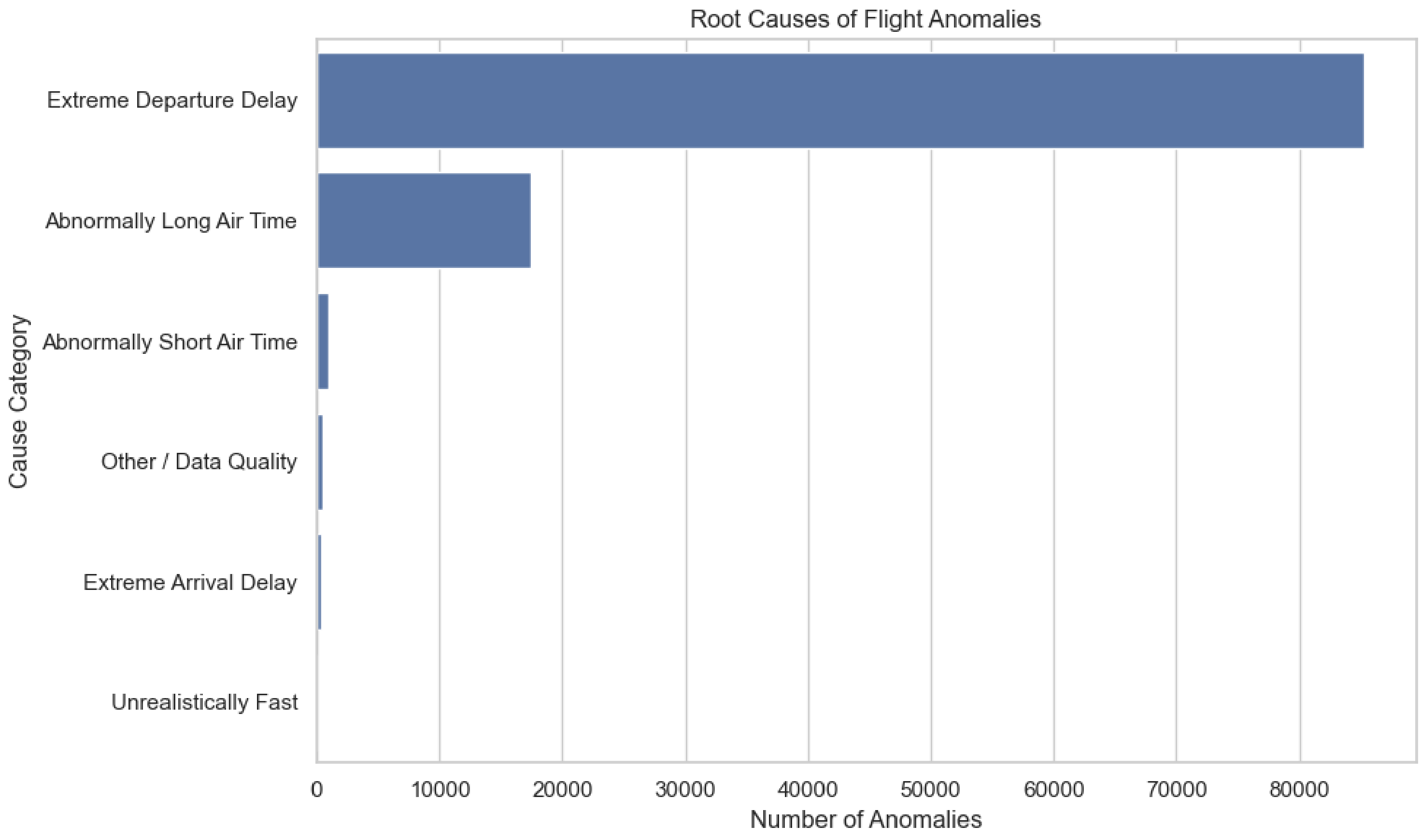
- **UA** – United Airlines
- **AA** – American Airlines
- **DL** – Delta Air Lines
- **AS** – Alaska Airlines

Regional carriers trend slower, with lower medians caused by short-haul routes and turboprop/regional jet operations:

- **9E** – Endeavor Air
- **HA** – Hawaiian Airlines (inter-island turboprops)
- **MQ** – Envoy Air
- **YX** – Republic Airways

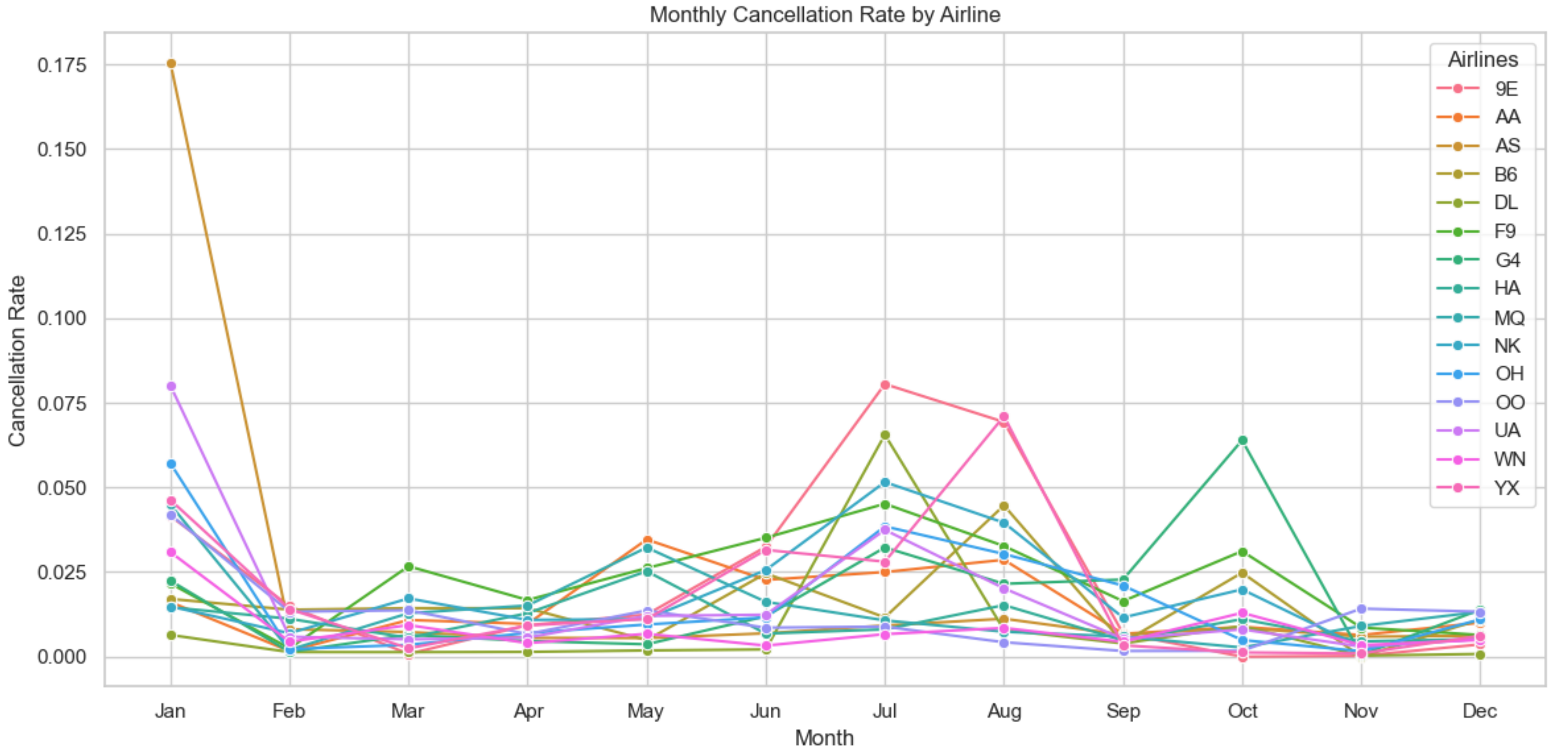


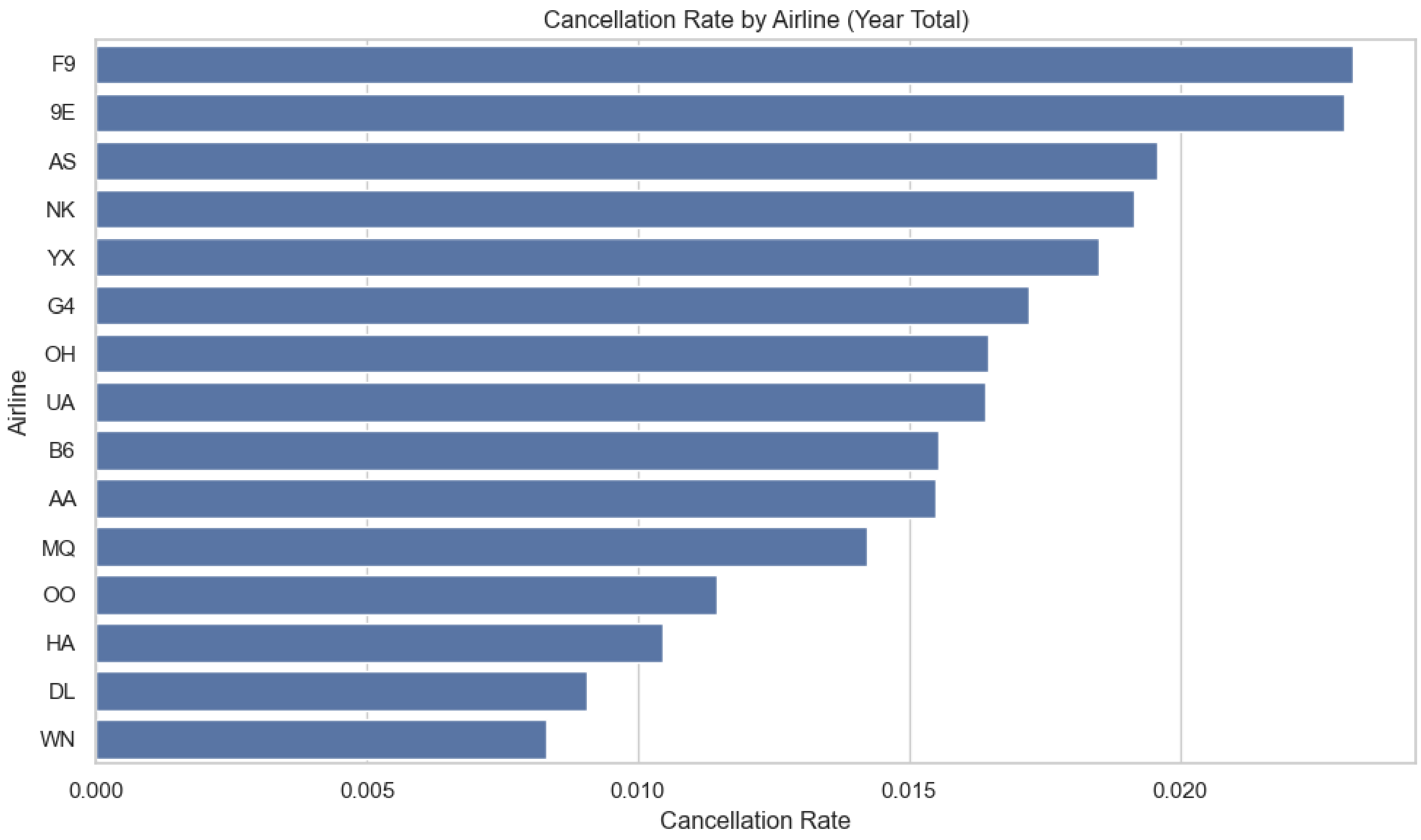
Flight Anomalies



Cancellations

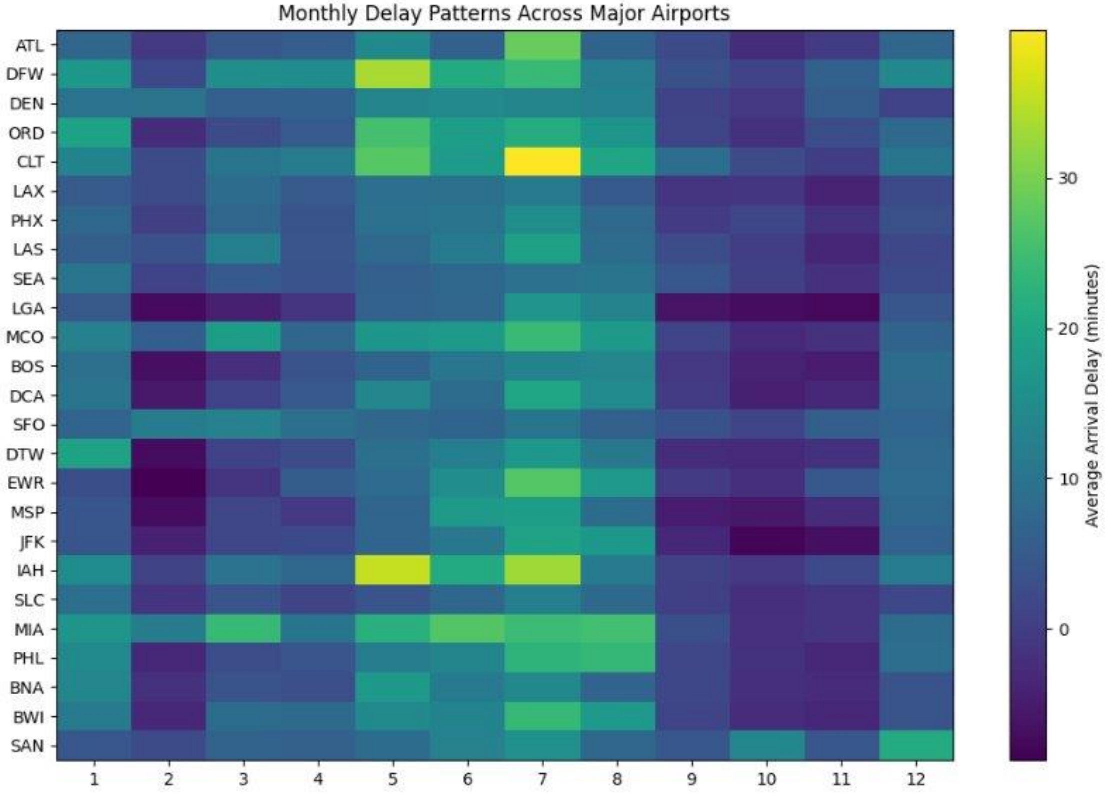
- Cancellation rates stay low most months, typically below 2% across airlines
- Major disruptions appear as sharp spikes rather than steady increases
- Winter (January) and peak summer (July–August) show the highest volatility
- Operational stress windows:
    - Winter storms (weather-driven cancellations)
    - Crew shortages + congestion (mid-summer surges)
- Large-network airlines show lower volatility compared to smaller or regional operators

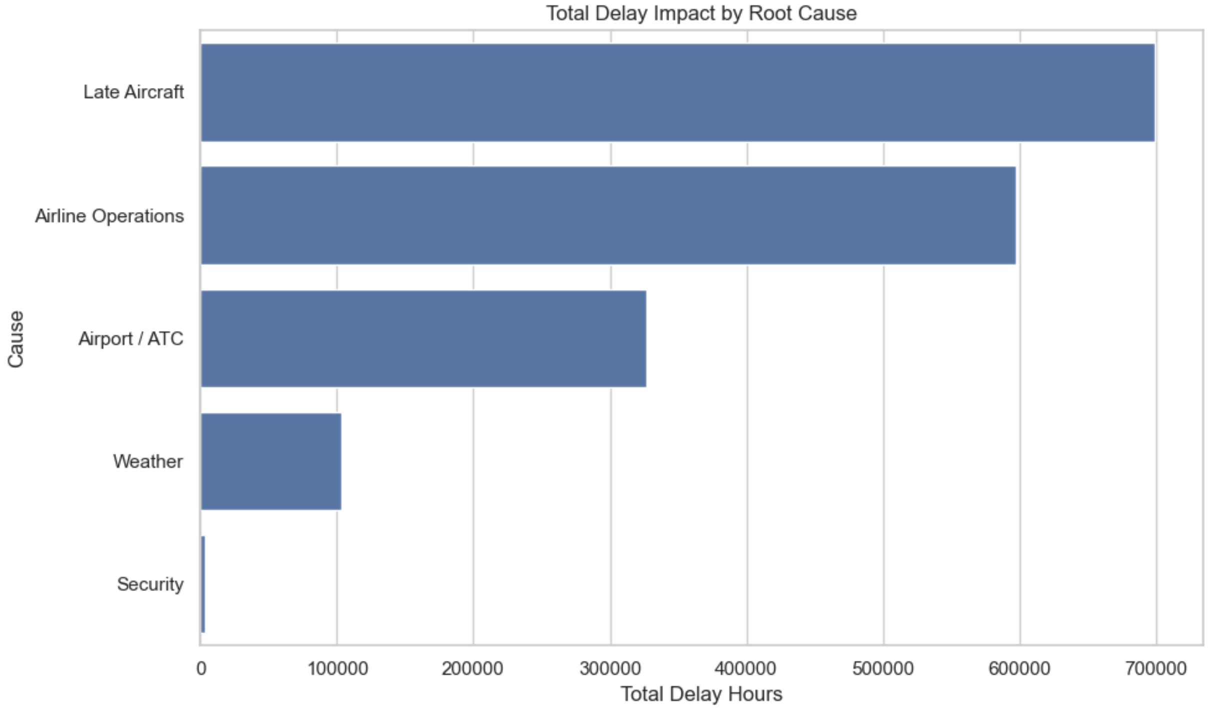




Highest cancellation rates (least reliable):	**F9 – Frontier Airlines**
	9E – Endeavor Air
	AS – Alaska Airlines
	NK – Spirit Airlines
Mid-tier reliability:	**UA – United Airlines**
	AA – American Airlines
	B6 – JetBlue Airways
	G4 – Allegiant Air
	YX – Republic Airways
	OH – PSA Airlines
Lowest cancellation rates (most reliable):	**WN – Southwest Airlines**
	DL – Delta Air Lines
	HA – Hawaiian Airlines

- Regional and ultra-low-cost carriers show higher cancellation rates. Driven by:
  - Smaller fleets (less rerouting flexibility)
  - Tighter crew availability
  - Weather sensitivity on short-haul routes
- Major network carriers show the lowest cancellation rates:
  - WN – Southwest Airlines
  - DL – Delta Air Lines
  - HA – Hawaiian Airlines



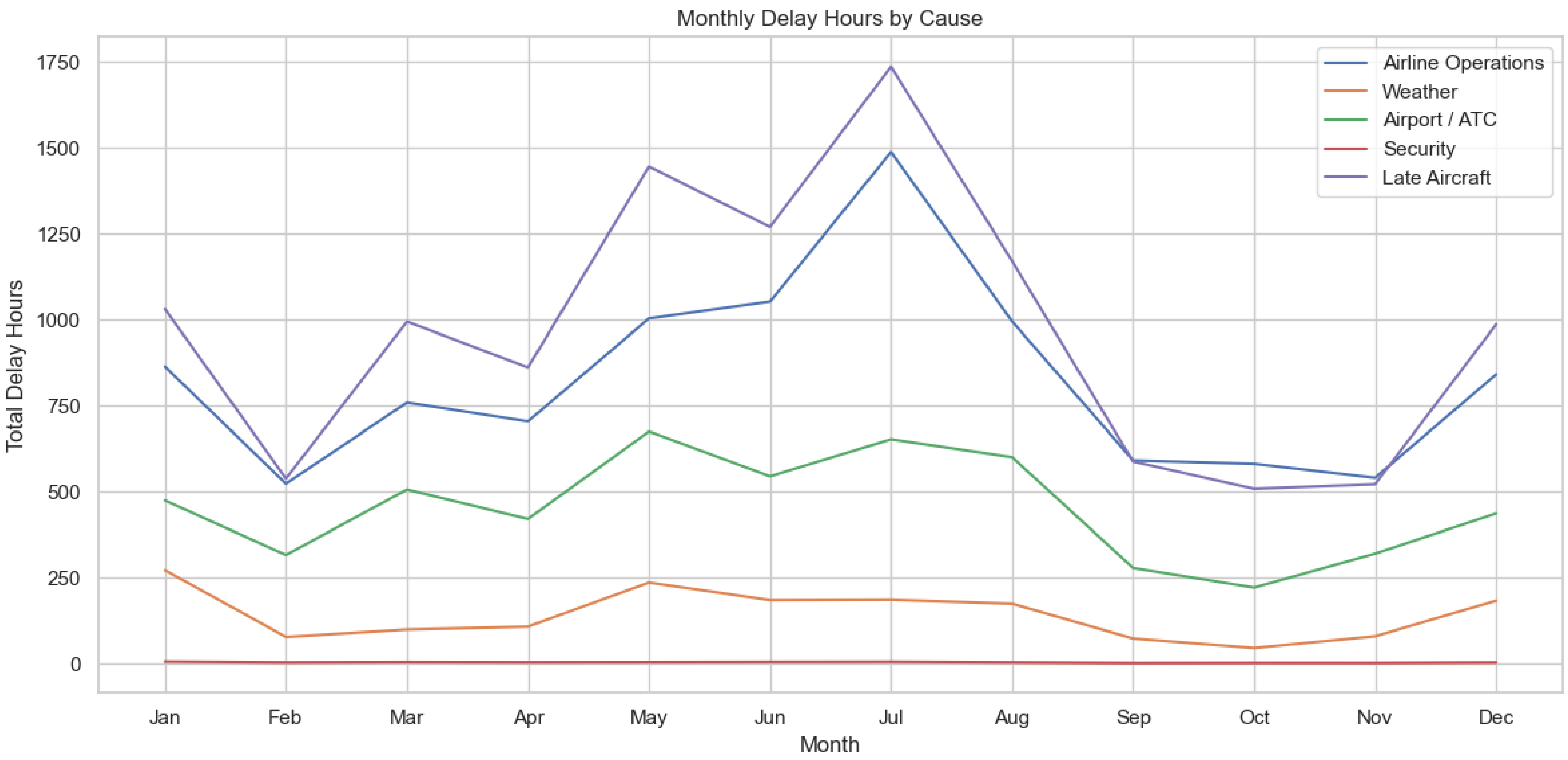


Across nearly all hubs on the heatmap, delay intensity climbs sharply from May → August

June and July show the highest concentration of yellow bands

- Driven by:
    - Capacity limits
    - High utilization schedules
    - Congested hubs

- Late aircraft arrival accounts for the largest share of accumulated delay hours
- Operational issues (crew availability, maintenance scheduling, gate assignments) rank second
- Airport congestion & air traffic control (ATC) contributes less, but still significant
- Weather, while highly disruptive when it occurs, contributes far less than total yearly delay time
- Security delays relatively minimal

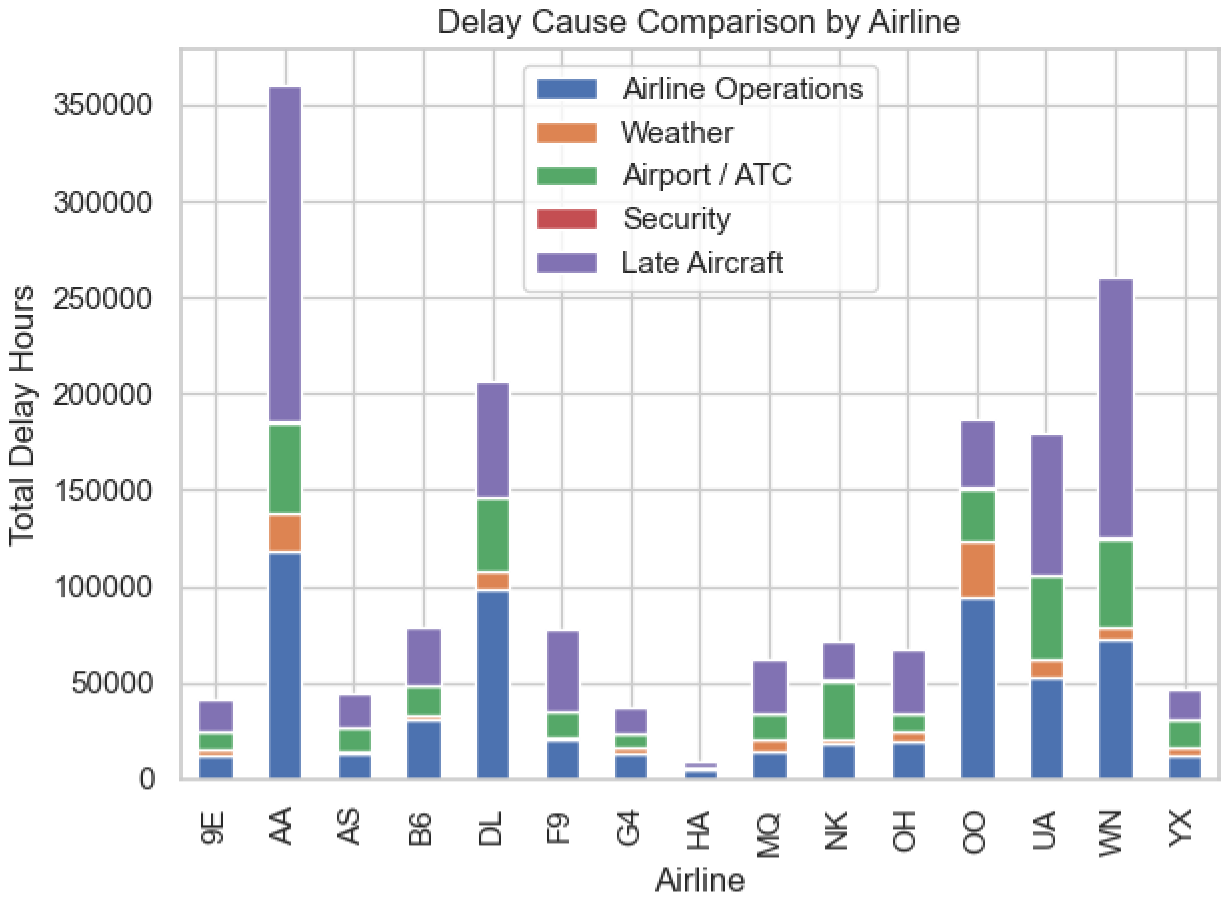


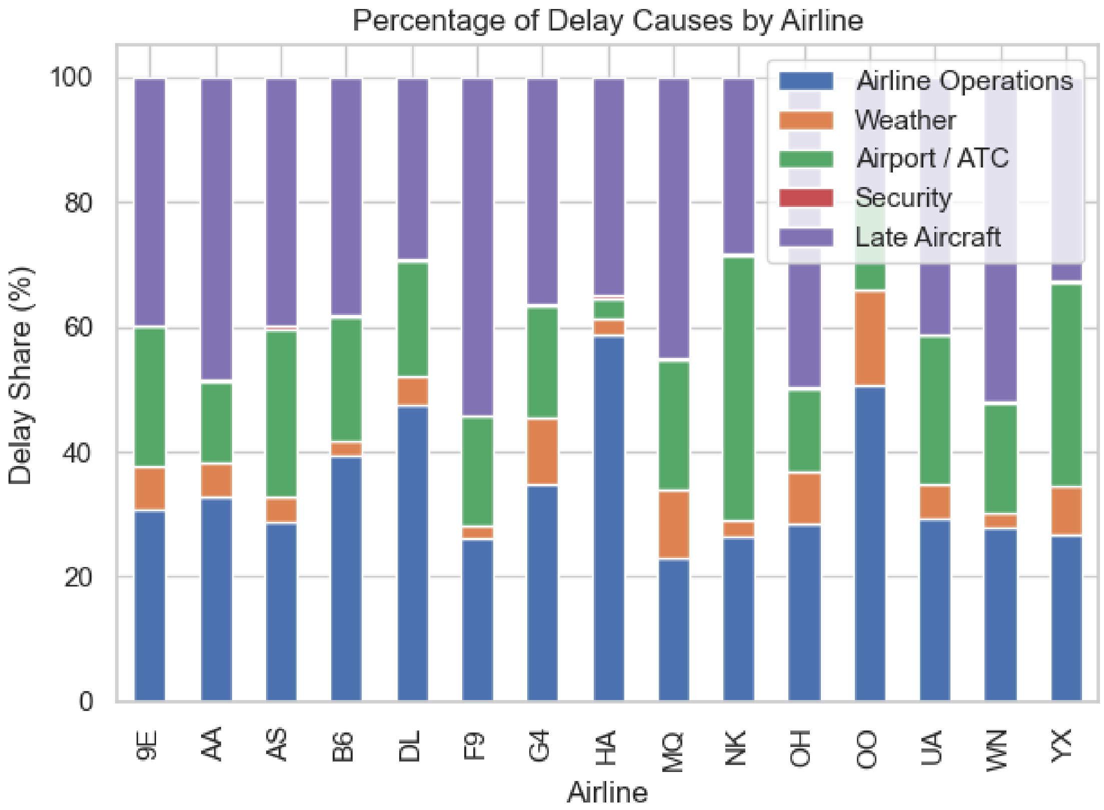
Monthly Delay Hours by Cause
Airline Operations
Weather
Airport / ATC
Security
Late Aircraft
1750
1500
1250
1000
750
500
250
0
Total Delay Hours
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Month

Major Network Airlines
(AA, DL, UA, WN)

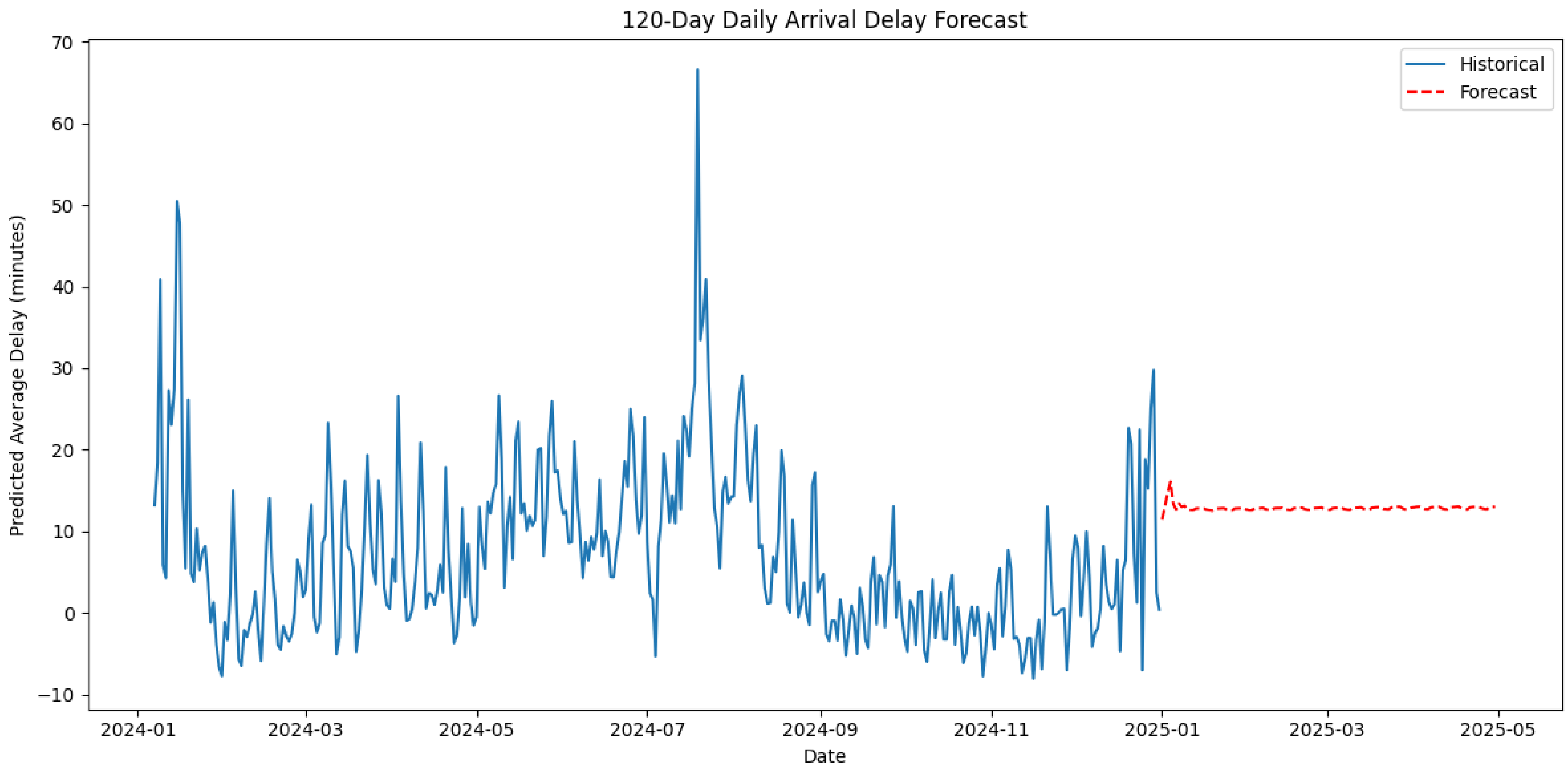
- More diversified causes
- Lower cancellation rates but large absolute delay hours due to higher total operations volume

- Late aircraft arrivals drive most delays
- Airline operational issues are the second-largest cause.
- Hub congestion and weather add seasonal spikes, with ATC delays clustering during morning & evening bank times and weather causing short, intense disruptions mainly in winter storms and summer thunderstorms.



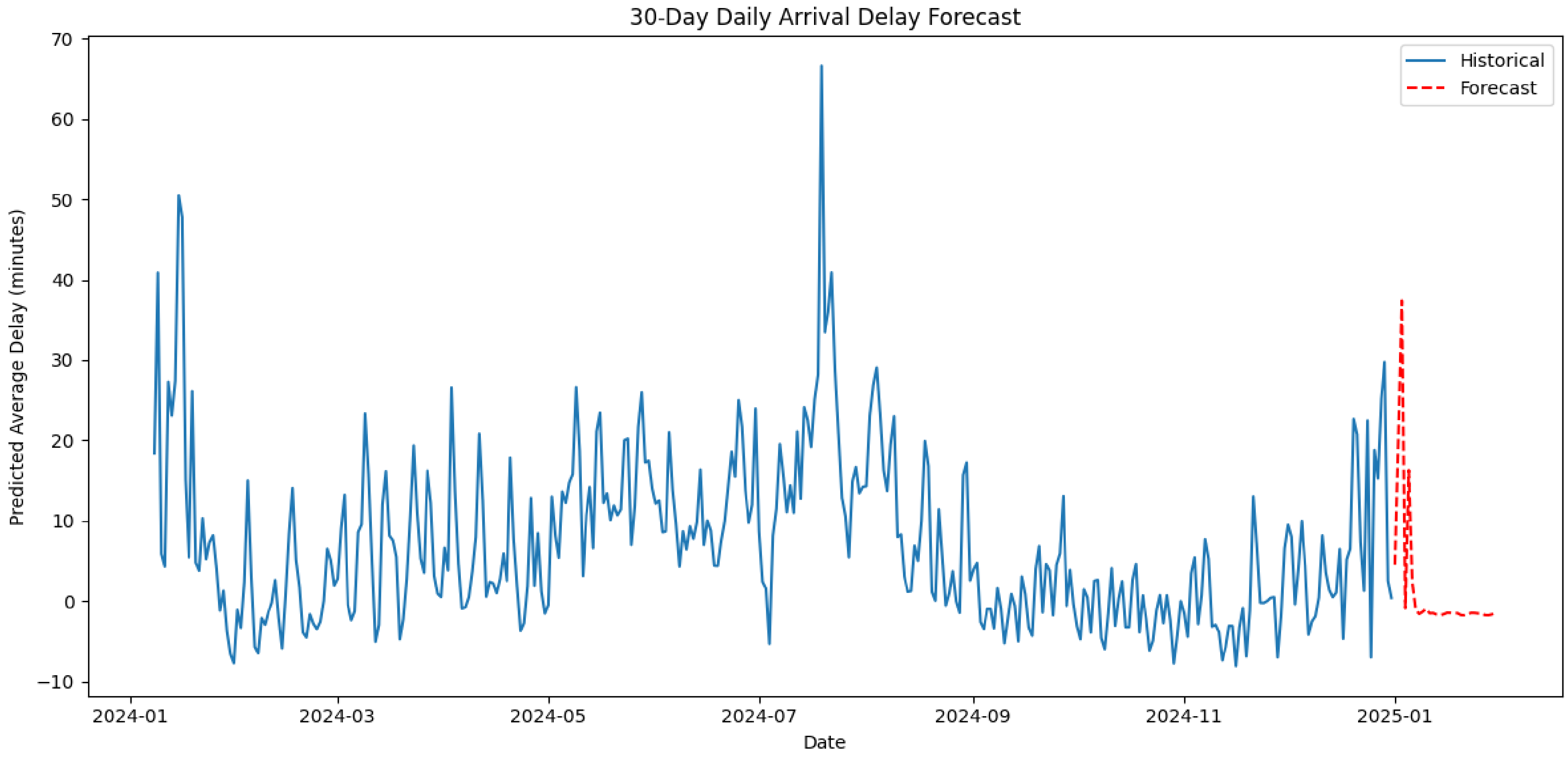


Future Delays: Random Forest With Rolling Window Smoothing



- Lacks dynamic input variability
- Without future weather, operation schedules, or airline volume the projection flattens
- Randomness in history comes from daily variability which is missing from the future inputs

Future Delays: Random Forest With Autoregressive Feature Modeling



- Keeps each day's information separate
- Allows model to learn whether delays spike after calm days, whether congestion cascades, and whether weekly cyclic effects exist.
- Mimics the natural disruptions that may occur

Machine Learning Approach

Models Tested

1. **Logistic Regression** (baseline) → ROC-AUC: 0.65
2. **Random Forest** → ROC-AUC: 0.66
3. **XGBoost (basic)** → ROC-AUC: 0.69
4. **XGBoost (tuned)** → ROC-AUC: 0.68, F1: 0.40

Code:

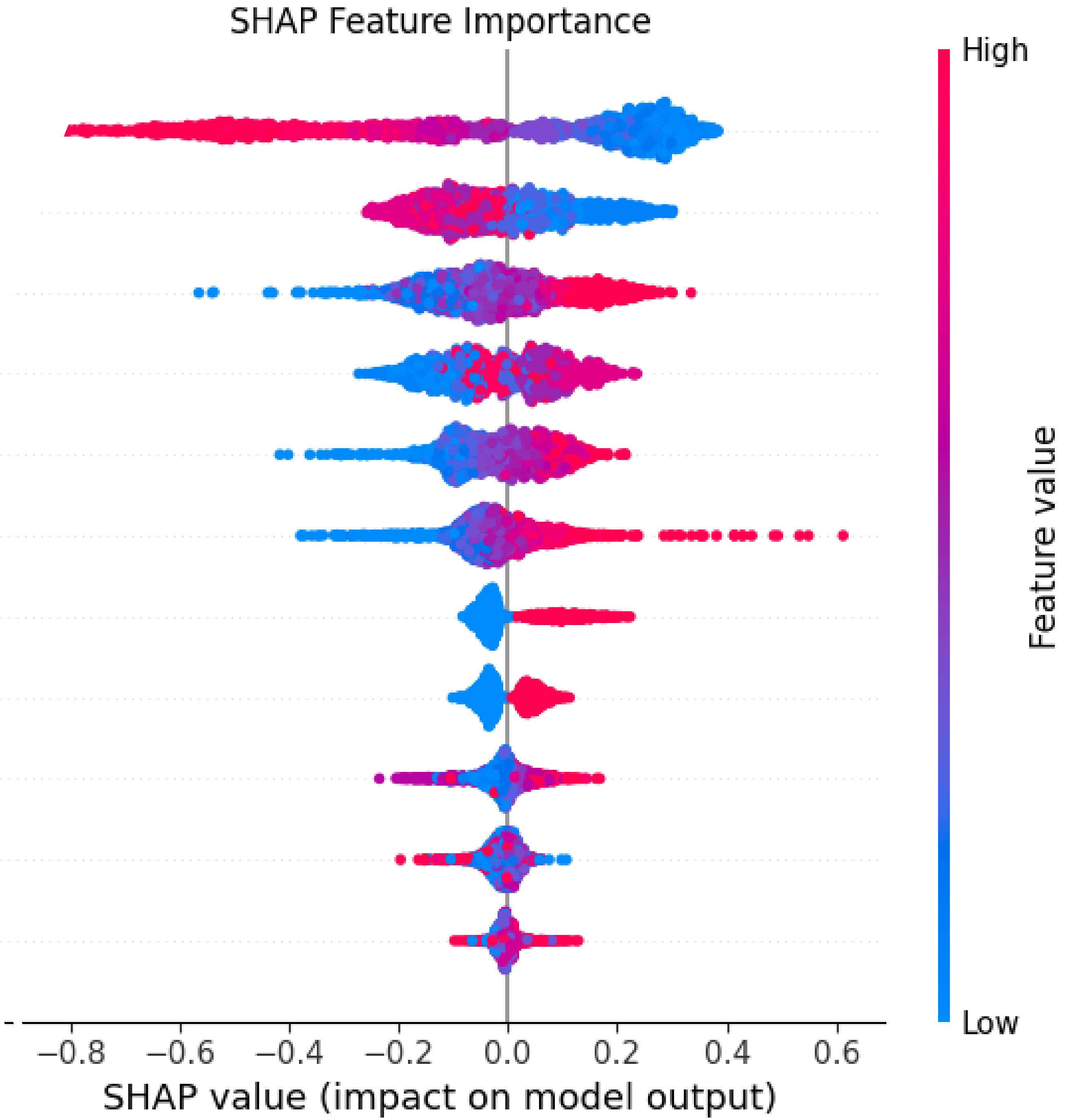
```
# Hyperparameter tuning with RandomizedSearchCV
param_dist = {
  'n_estimators': [100, 200, 300],
  'max_depth': [5, 7, 9, 11],
  'learning_rate': [0.01, 0.05, 0.1],
  'scale_pos_weight': [3, 5, 7] # Handle class imbalance
}

search = RandomizedSearchCV(
   XGBClassifier(),
   param_dist,
   cv=3,
   scoring='roc_auc', n_iter=20
)
```

Key Point: Tuning improved F1 from 0.07 to 0.40; important for practical use

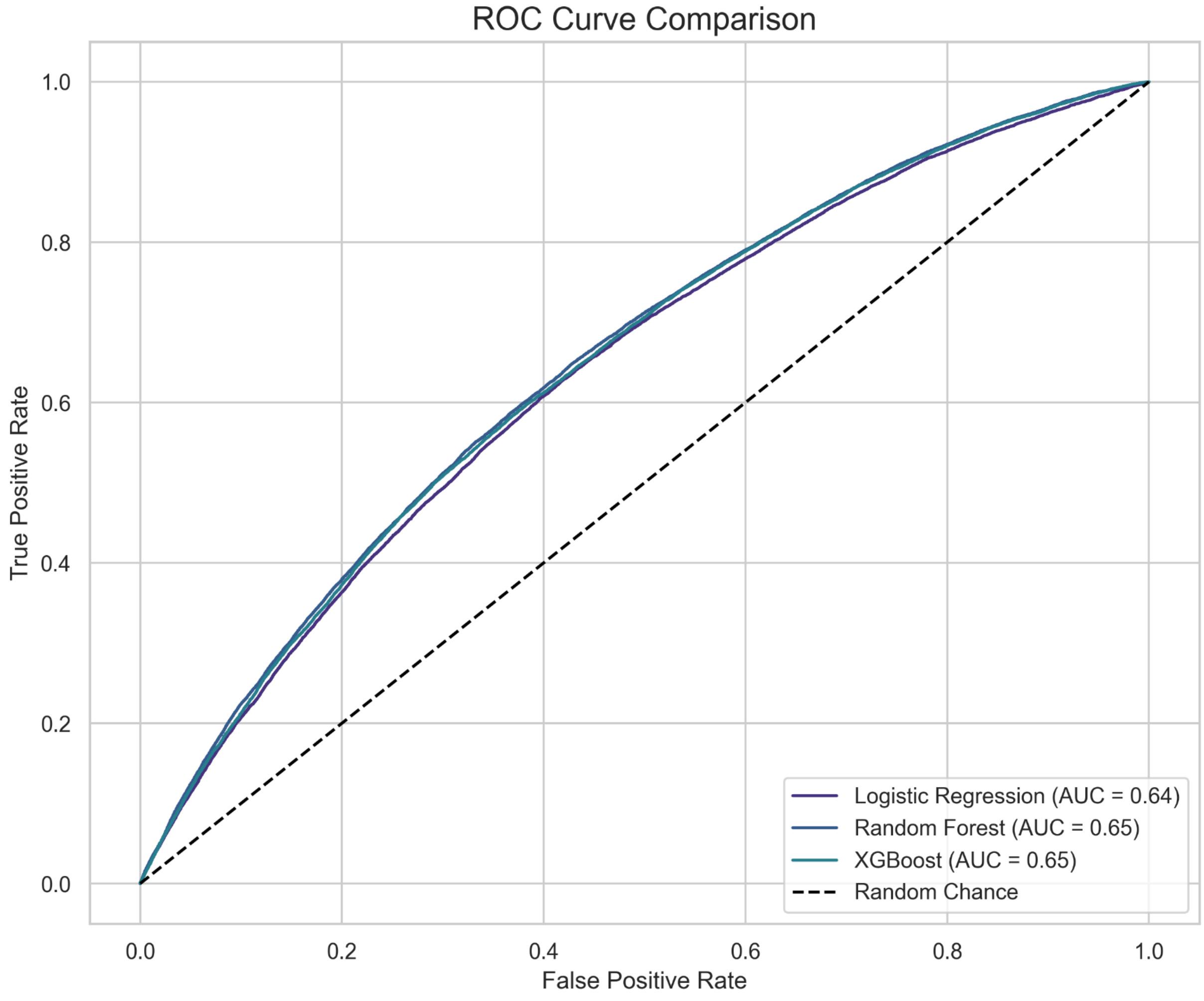
Feature Importance – What Drives Delays?

Top Predictive Features:

1. **Departure Hour** - Evening flights 3x more likely delayed
2. **Origin Airport** - Hub congestion matters most
3. **Destination Airport** - Endpoint capacity constraints
4. **Carrier** - Some airlines have systematic issues
5. **Distance Bucket** - Medium-haul (800-1500 mi) highest risk



Model Performance & Business Value



Performance Breakdown:

- **Precision:** 29% (of predicted delays, 29% are correct)
- **Recall:** 65% (catch 2/3 of actual delays)
- **ROC-AUC:** 0.68 (decent discriminative ability)

Business Cost Analysis:

False Negative (missed delay):	$1,000 cost
False Positive (unnecessary alert):	$100 cost

Model saves: $8M+ per year vs no prediction

Trade-off Explanation: "We tuned for recall - better to warn passengers and be wrong sometimes than miss a delay"