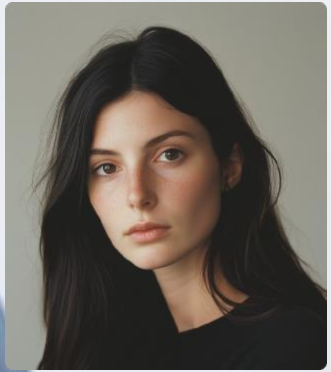
Analyzing & Predicting Airline Delays

A Comprehensive Data Science Approach

MSIS 502 Final Project

Meet the *TEAM*

Mithun Manjunath



Norah Alomaim

Fardeen Meeran

Ashish Dubey

Surbhi Meena

Agenda

Next

Introduction

About next slides
About next slides
About next slides

The Cost of Flight *Delays*

Flight delays cost billions

U.S. flight delays cost airlines and passengers

- Fuel Burn During Taxiing
- Crew Overtime & Scheduling
- Aircraft Maintenance
- Aircraft rescheduling

$33 billion

Passengers bear hidden costs:

Costs travelers

- Lost Productivity
- Rebooking
- Time Stress
- Overnight Stay

$70 per day

Economic efficiency is impacted:

Reduce overall economic productivity in the transportation sector.

- Disrupted Supply chain
- Business Travel Inefficiency
- Productivity Loss
- Airline Inefficiency

$11.5 million

Drivers of Delay and Cost *Variation*

X.

Carrier Delays

Text if any

X.

Seasonal Delay Pattern

Text if any

X.

NAS and weather

Text if any

X.

Airline - Specific Trends

Text if any

X.

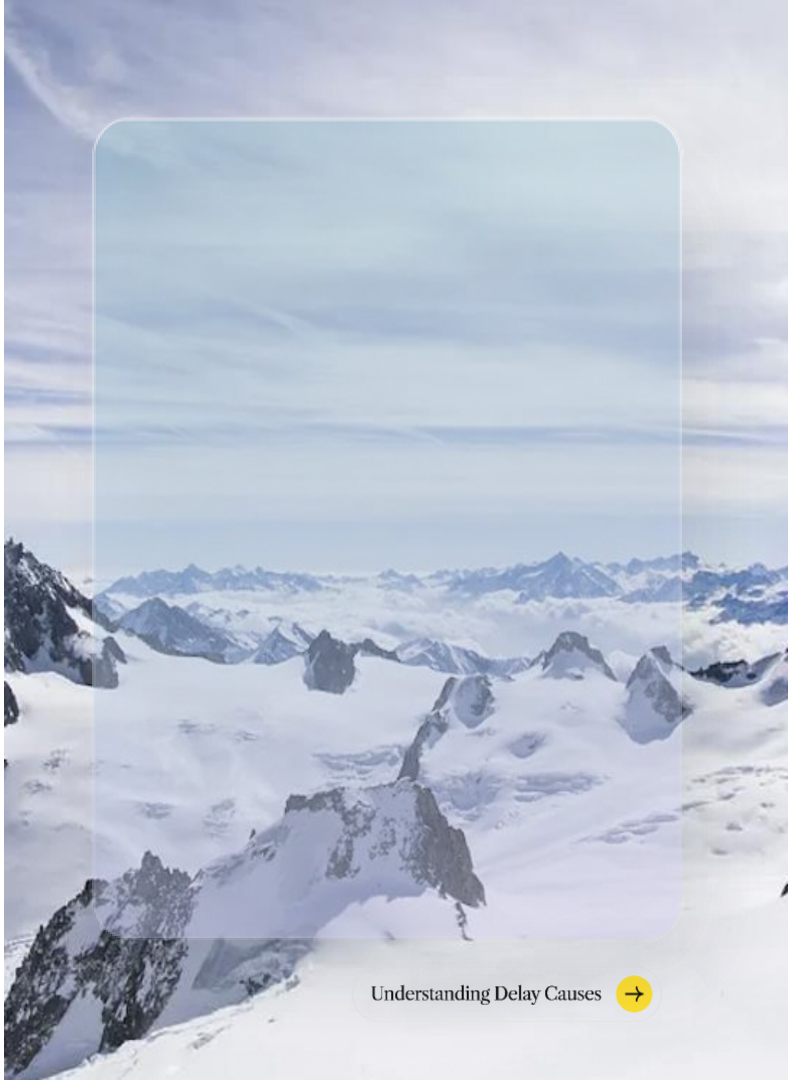
Late Aircraft Propagation

Text if any

X.

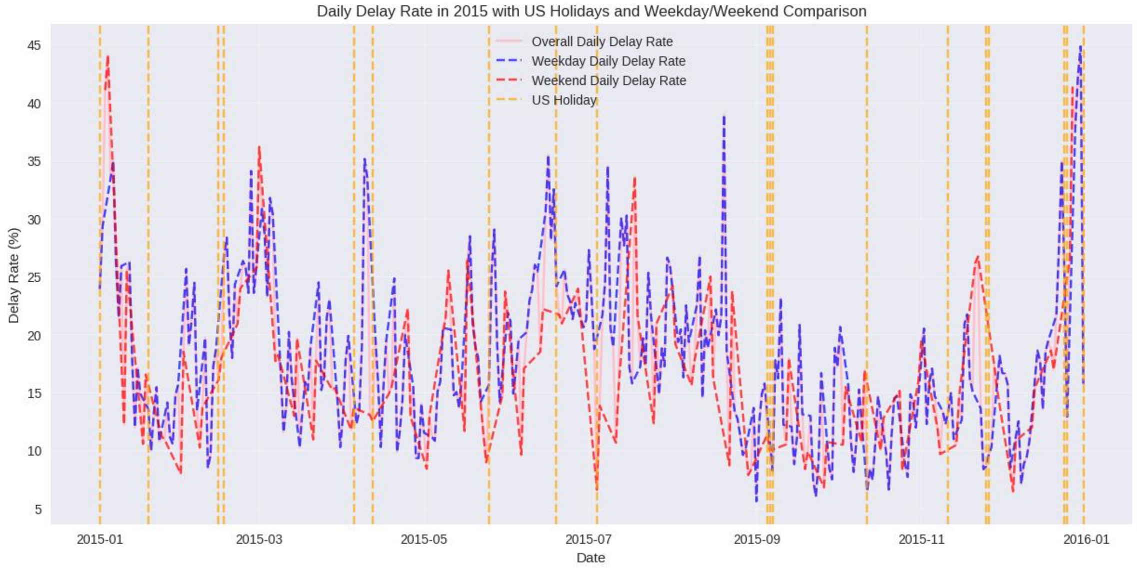
Regional Disparities

Text if any



DAILY DELAY RATE IN 2015

The big Picture



Option A >45%

Option B 43%

Option C 45%

Option D 42% (Security)

Description:

Project Goal & *Overview*

Data Analysis

Analyze 2015 US flight data (millions of records) to uncover delay trends and patterns.

Delay Prediction

Predict delays for individual flights in their current state using data-driven models.

Operational Insight

Goat is to recommend strategies to reduce delays and improve airline performance

Business Impact

Address $4.57B+ annual delay costs by improving efficiency and customer satisfaction.

Drivers of Delay and Cost *Variation*

01

Data Overview

Raw data (5.8M flights,
322 airports, 14 airlines

02

Missing Value Analysis

NaN values analysis :
Features with Integer vs String

01

Data Distribution

Data spread : Overview, feature-wise spread, month-wise distribution and outlier check

01

Key column Completeness

AIRLINE,
ORIGIN_AIRPORT,
DEPARTURE_TIME,
DEPARTURE_DELAY,
DESTINATION_AIRPORT,

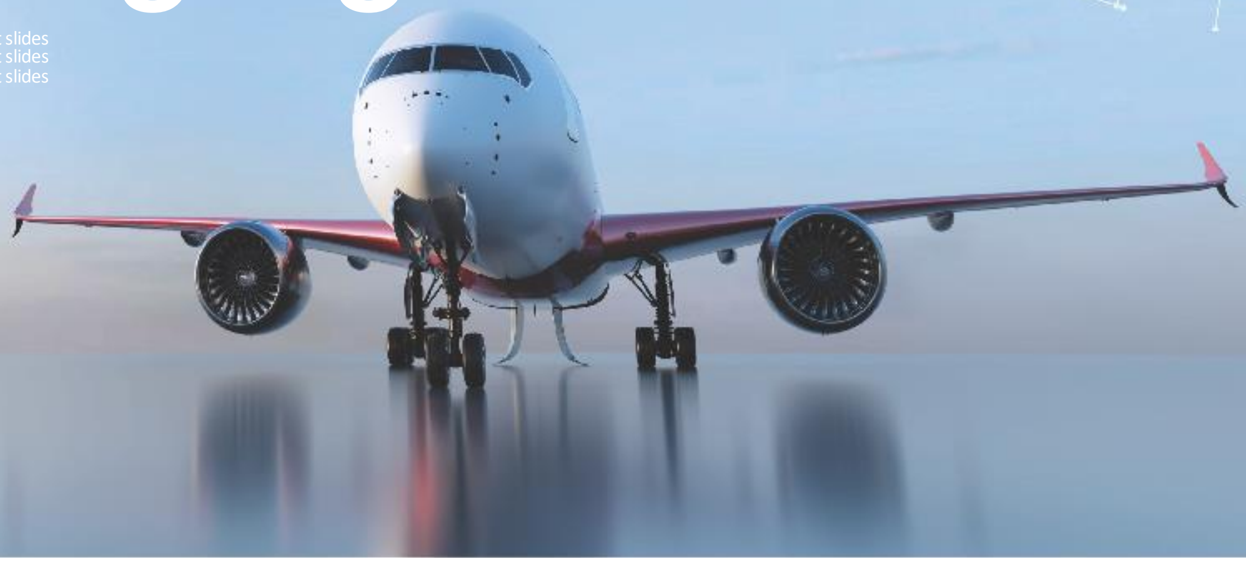
01

Negative Delay Values

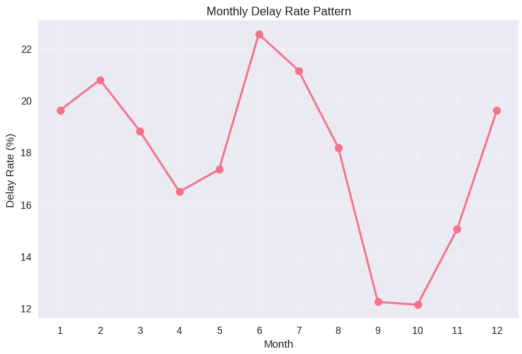
Identification and handling of negative values in delay columns

Identifying Signals

About next slides
About next slides
About next slides

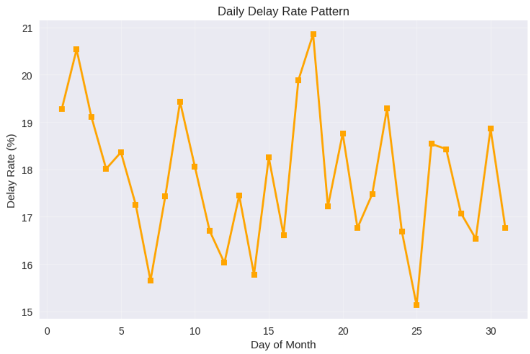


Visualization : *Temporal*



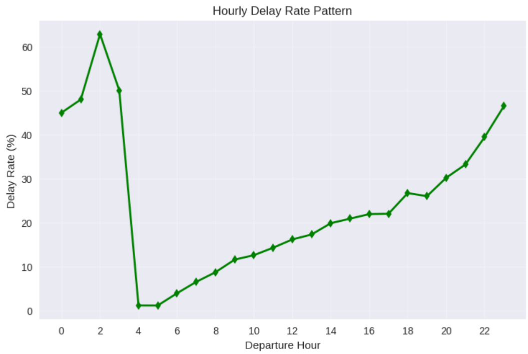
Option A >45%

Highlight: Time is an important dimension



Option A >45%

Highlight:



Option A >45%

Highlight:

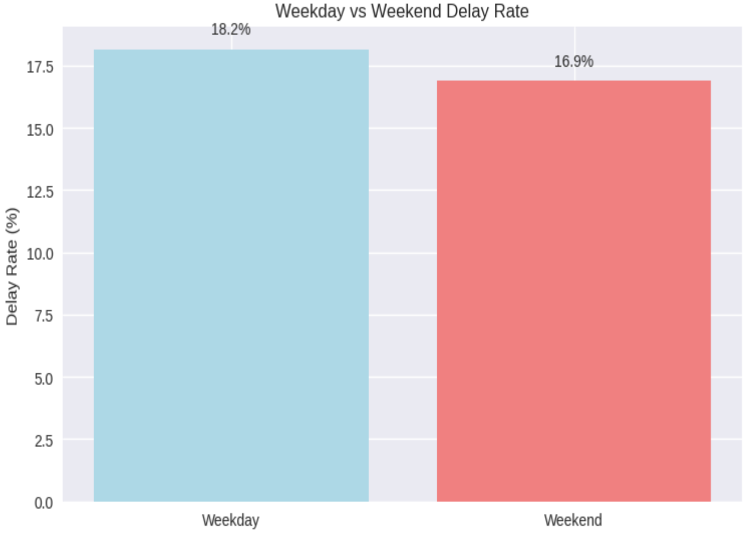
1. Highlight: Time is an important dimension

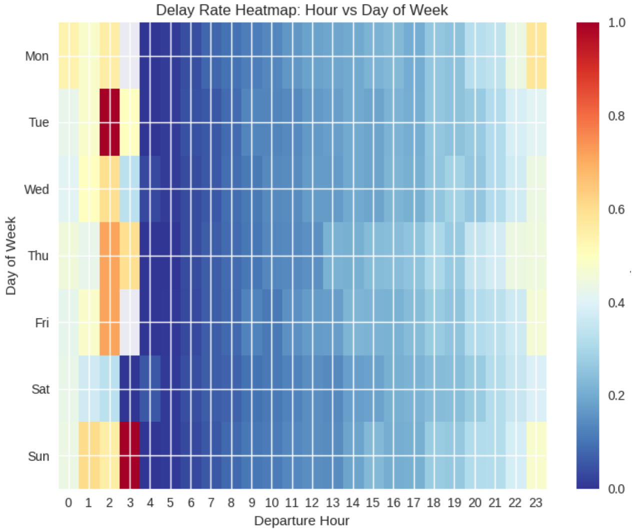
2. High variability & seasonality in delay rates as a function of time makes forecasting a challenge

3. Why are Delay rates high in June & July?

DAILY DELAY RATE IN 2015

The big Picture





Option A >45%

Option B 43%

Option C 45%

Option D 42% (Security)

Description:

Description:

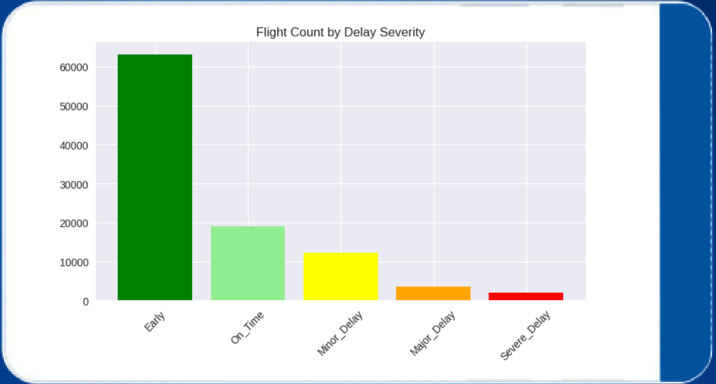
1. Highlight: Time is an important dimension

2. High variability & seasonality in delay rates as a function of time makes forecasting a challenge

3. Why are Delay rates high in June & July?

The Cost of Flight *Delays*

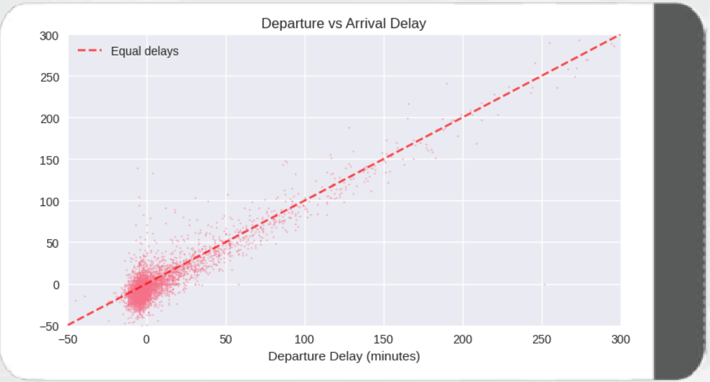
Flight Count by Delay Severity



Highlight: Time is an important dimension

On_Time
On_Time
Early
Minor_Delay
Major_Delay

Departure vs Arrival Delay

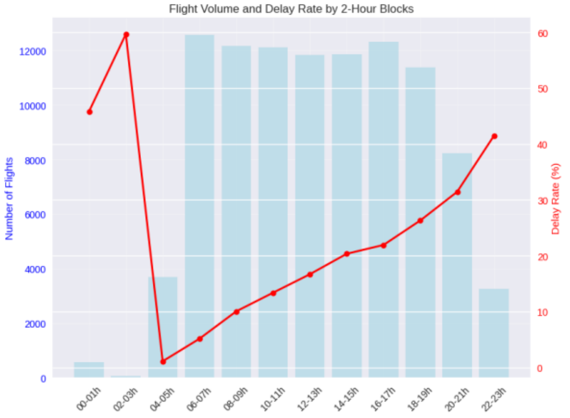


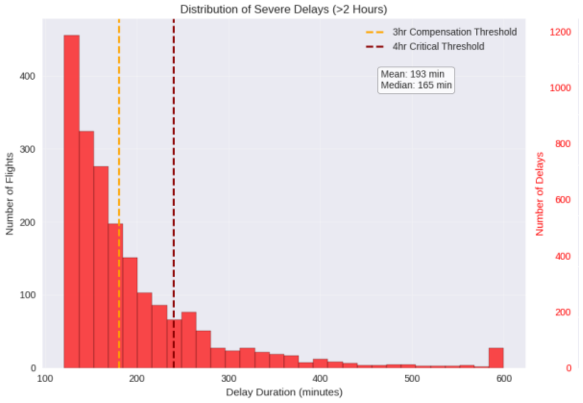
Highlight: Time is an important dimension

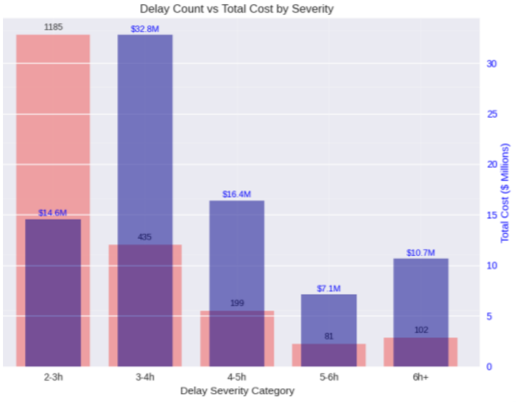

Visualization
Journey
About next slides
About next slides
About next slides

Visualizations

Days Correlations and Impact of Severe Delays







Option A >45%

Option B 43%

Option C 45%

Option D (Security) 42%

Description: **Frequency :** Early on we were unsure of cascading effects. The first plot shows that even after frequency of flights drop the delays are increasing. So cascading effect is more likely.

Description:

Description:

Key Points

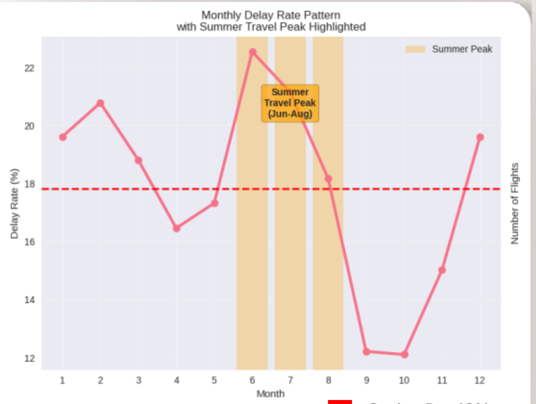
1. No of flights with severe delays and their cost.
2. Cost increase is exponential with delay severity. (Ratio of block to pink blocks)
3. if the effects are cascading into subsequent flights then the costs incurred compound as well.

Seasonality

Is it a factor for delays?

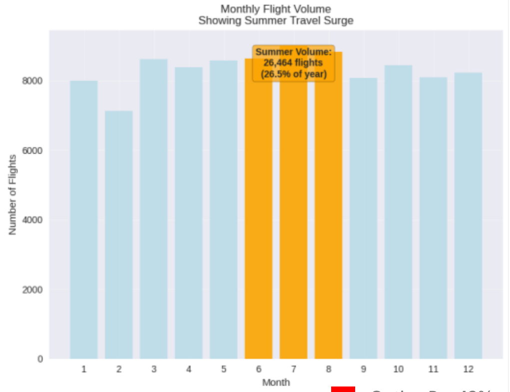
Book before it's gone

Seasonality : *relational patterns*



Option A >45% Option B 43%

Highlight: Time is an important dimension



Option A >45% Option B 43%

Highlight:

Option A >45% Option B 43%

Highlight:

1. Monthly : Summer peak in delays have a correlation with empirical observation of large crowds & limited capacity.

2. Weekly :
Weekend & weekday have similar trends with delays

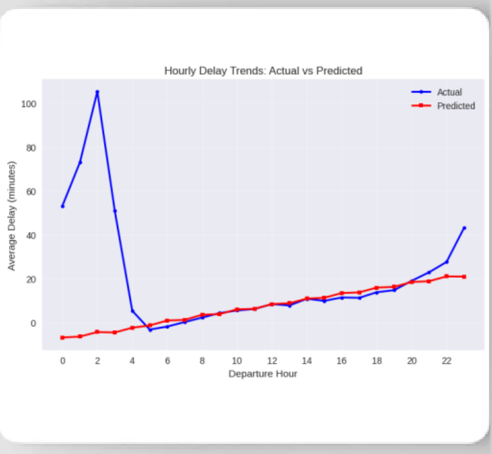
Predictive modeling

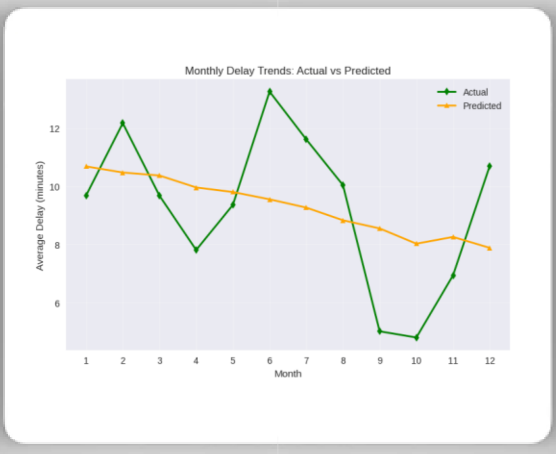
About next slides
About next slides
About next slides

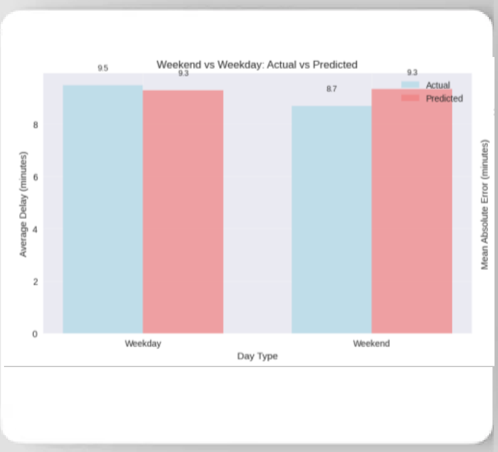
Prediction with *Linear Regression*

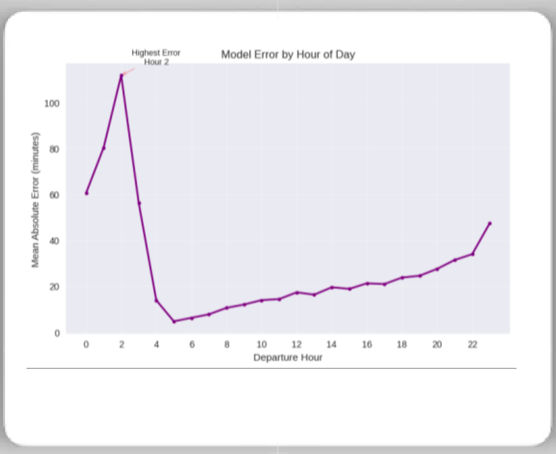
Description:

```
TRAINING LINEAR REGRESSION MODEL
========================================
Training set size: 80,000
Test set size: 20,000

MODEL PERFORMANCE METRICS
------------------------------
Training Set:
  RMSE: 37.09 minutes
  R²: 0.0286
  MAE: 17.81 minutes

Test Set:
  RMSE: 35.96 minutes
  R²: 0.0261
  MAE: 17.61 minutes

Model Interpretation:
  Model has limited predictive power (R² = 0.0261)
  On average, predictions are off by 17.6 minutes
```



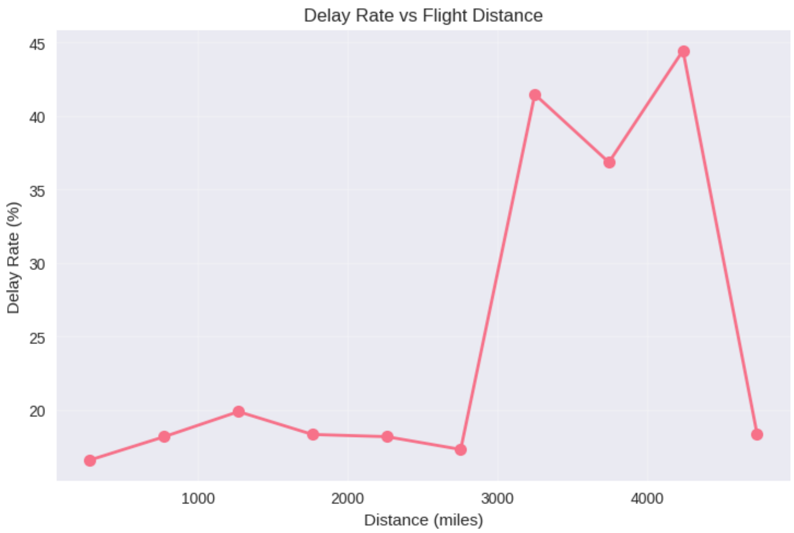




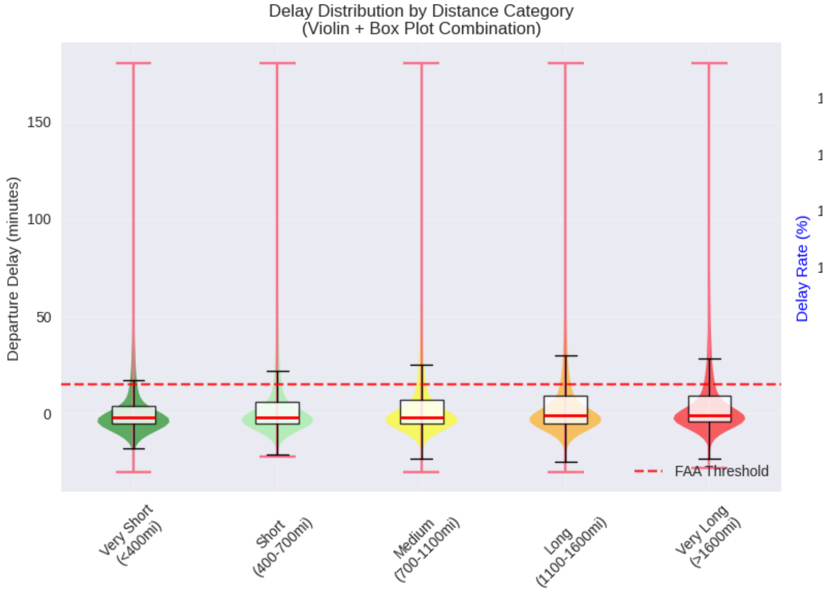


DAILY DELAY RATE IN 2015

Distance as a feature: An Enigma



Description:



Description:

1. Highlight: Long tail of values a concern. Hence removing it. Also removed are similar temporal features (Peak hour, Is_weekend & Day) to avoid overfitting.

Prediction with *Gradient Regression Model & Model & Random Forest Regression Model Model*

```
TRAINING GRADIENT BOOSTING REGRESSION MODEL
==================================================

GRADIENT BOOSTING MODEL PERFORMANCE METRICS
----------------------------------------
Training Set:
  RMSE: 37.48 minutes
  R²: 0.0079
  MAE: 14.21 minutes

Test Set:
  RMSE: 36.28 minutes
  R²: 0.0087
  MAE: 14.07 minutes

Model Interpretation (Gradient Boosting):
  Gradient Boosting Model explains 0.9% of variance in delays
  On average, predictions are off by 14.1 minutes
```

Highlights:

```
TRAINING RANDOM FOREST REGRESSION MODEL
==================================================

RANDOM FOREST MODEL PERFORMANCE METRICS
----------------------------------------
Training Set:
  RMSE: 29.33 minutes
  R²: 0.3924
  MAE: 13.75 minutes

Test Set:
  RMSE: 37.48 minutes
  R²: -0.0579
  MAE: 17.64 minutes

Model Interpretation (Random Forest):
  Random Forest Model has limited predictive power (R² = -0.0579)
  On average, predictions are off by 17.6 minutes
```

Highlights:

Conclusion

Add text here if any:

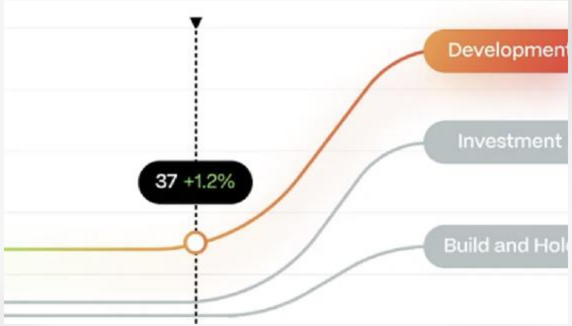
- Cascading effects with delays need action to reduce compounded costs and sustained negative customer sentiment
- Daily variability is a fundamental problem
- Monthly Seasonality mandates increasing staffing
- Key features for modeling: Month, Departure hour, 2 hour blocks (for distribution) and Origin Airport
- Learning rate plays a key role in prediction given the massive variability

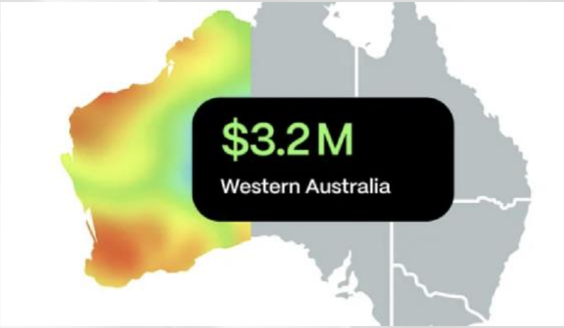
Appendix

Appendix *List*

1. Text here

Drivers of Delay and Cost Variation





Airline Operating Costs

Airline Operating Costs Direct costs include fuel burn during taxiing, crew overtime, maintenance scheduling issues, and gate occupancyy.

Average airline cost per delayed minute = $ 74.24+

Passenger Time Valuation

DOT and FAA benchmark $47.10 per hour as the value of of passenger time, adjusted across trip purpose and demographics.

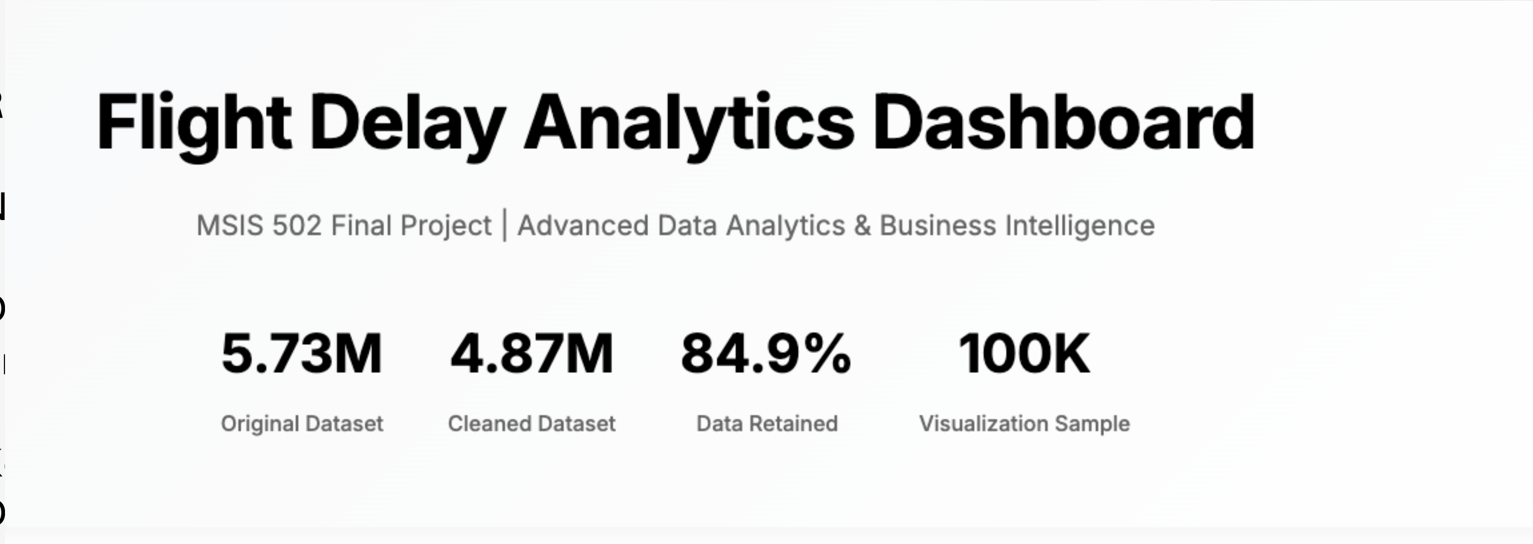
passenger delay cost per hours = $47.10 per hour , compounding over millions of flights.

Broader Economic Impact

Average delay cost for 850M passengers/year = $6.8B annually in lost time, expenses, and disruptions.

Data Exploration

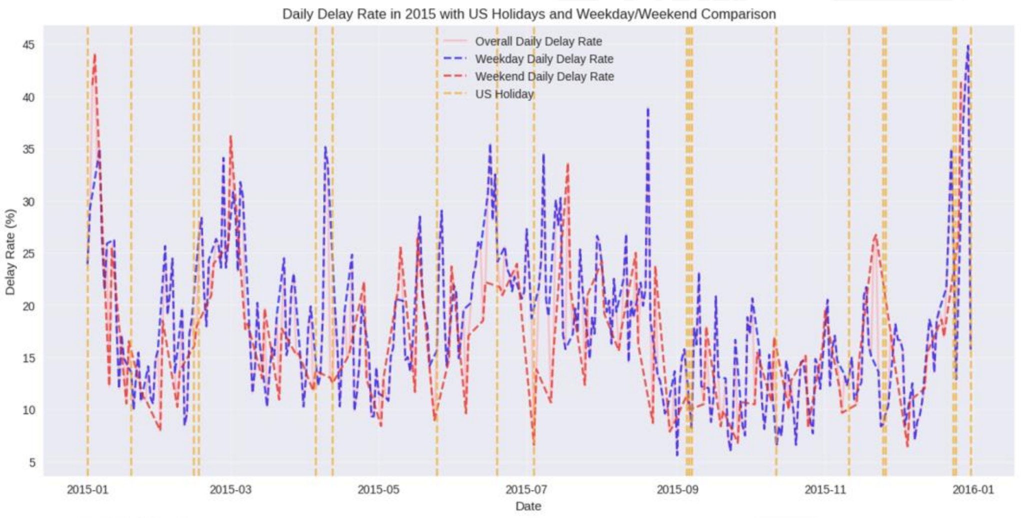
Project Goal & Overview



- R
- N
- D
  a
- K
  D
- Negative values in Delay columns

DATA Exploration

Predict delay for a given flight in its Current state and Recommended Operational improvements.



00%

2025

Analyze 2015 US flight data (millions of records) to identify pattern in delays

00%

Business context

High delay costs (>$4.57B/year); goal is to predict delays and their ramification on costs and bad customer experience.

00%

Flight Operations

provide insights on key areas of improvement

Data Cleaning Strategies & Tools Used

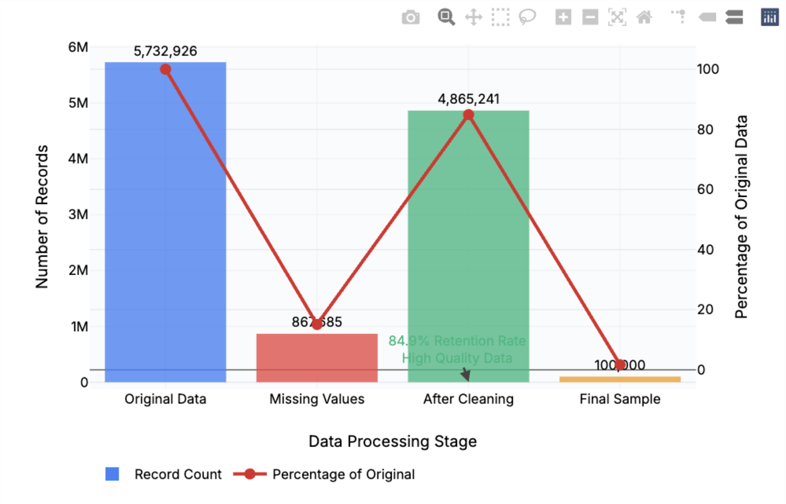
STEP 1.

Missing Depa
Delays

Paste the ref screensh

- Cancelled flights: 9,64
- Operated flights for a
- critical delay colum

Data Quality Assessment Dashboard



Quality Metrics Summary

Data Retention Rate:	Missing Value Treatment:	Stratified Sample Size:	Monthly Balance:
84.9%	**100% Addressed**	**100K Records**	**Achieved**

Data Cleaning Strategies & Tools Used

STEP 3.

Feature Eng

Numeric Fea

Paste the ref scre

- Time categories cr
- Delay categories c
- Route features cre
- Scaled features: ['I SCHEDULED_DEPA

Data Preparation & Quality Assessment

Comprehensive data cleaning and preprocessing workflow, transforming raw flight data into analysis-ready datasets through systematic quality assessment, missing value treatment, and strategic sampling methodologies.

> *"Quality data is the foundation of actionable insights. Our systematic approach to data preparation ensures robust analysis while maintaining representativeness across all temporal and operational dimensions."*
>
> ***— Data Engineering Excellence***

Original Dataset Quality

5.73M total flight records spanning 12 months with significant data quality challenges including missing values in critical delay columns and temporal distribution imbalances requiring strategic intervention.

Missing Value Strategy

Implemented sophisticated imputation: **0-fill for delays** (indicating on-time performance), **row removal for critical identifiers**, and **median imputation for distance** to maintain analytical integrity.

Temporal Rebalancing

Addressed **January overrepresentation** through stratified sampling, ensuring equal monthly representation while preserving within-month variance and operational patterns.

Feature Engineering

P

H

01

02

03.

🔬 Hypothesis 1: Peak Hour Concentration

Theory: Delays concentrate during peak operational hours. **Evidence:** Hours 17-20 show 40%+ delay rates vs. 15% average. **Action:** Redistribute 15% of peak flights to off-peak windows for 20% delay reduction.

🔬 Hypothesis 2: Hub Complexity Impact

Theory: Larger airports have disproportionate delay rates. **Evidence:** ATL and ORD show 21%+ delays vs. 14-16% at smaller airports. **Action:** Deploy dedicated delay mitigation teams at top 5 hubs.

🔬 Hypothesis 3: Distance-Delay Correlation

Theory: Longer flights are more delay-prone. **Evidence:** >1600mi flights show 22.7% delays vs. 14.2% for <400mi. **Action:** Enhanced pre-flight checks for long-haul operations.

🔬 Hypothesis 4: Cost Concentration

Theory: Severe delays drive disproportionate costs. **Evidence:** 2% of flights (>2hrs) represent 40% of delay costs. **Action:** Early intervention systems to prevent escalation.

🔬 Hypothesis 5: Seasonal Volume Impact

Theory: Summer travel creates systemic delays. **Evidence:** June-August show 25% volume increase with 15% delay rate spike. **Action:** Seasonal capacity planning and staffing adjustments.

🔬 Hypothesis 6: Airline Efficiency Variance

Theory: Operational practices drive performance differences. **Evidence:** Best airlines (14.8%) vs. worst (22.1%) show 50% performance gap. **Action:** Best practice sharing and benchmarking programs.

Optimize Peak Hour Scheduling

Hours 19, 18, and 20 show highest delay rates. Redistribute flights to off-peak times for 15-20% delay reduction and millions in cost savings through strategic schedule optimization.

Weekend Operations Strategy

Develop separate operational protocols for weekend vs weekday flights to optimize resource allocation and improve customer experience through tailored service delivery models.

Airport-Specific Interventions

Target improvement programs at worst-performing airports to achieve 10-15% on-time performance improvement at critical hubs through focused operational excellence initiatives.

Severe Delay Prevention

Implement early warning systems to prevent minor delays from becoming severe. Average cost per severe delay: $40,720, representing

d holidays

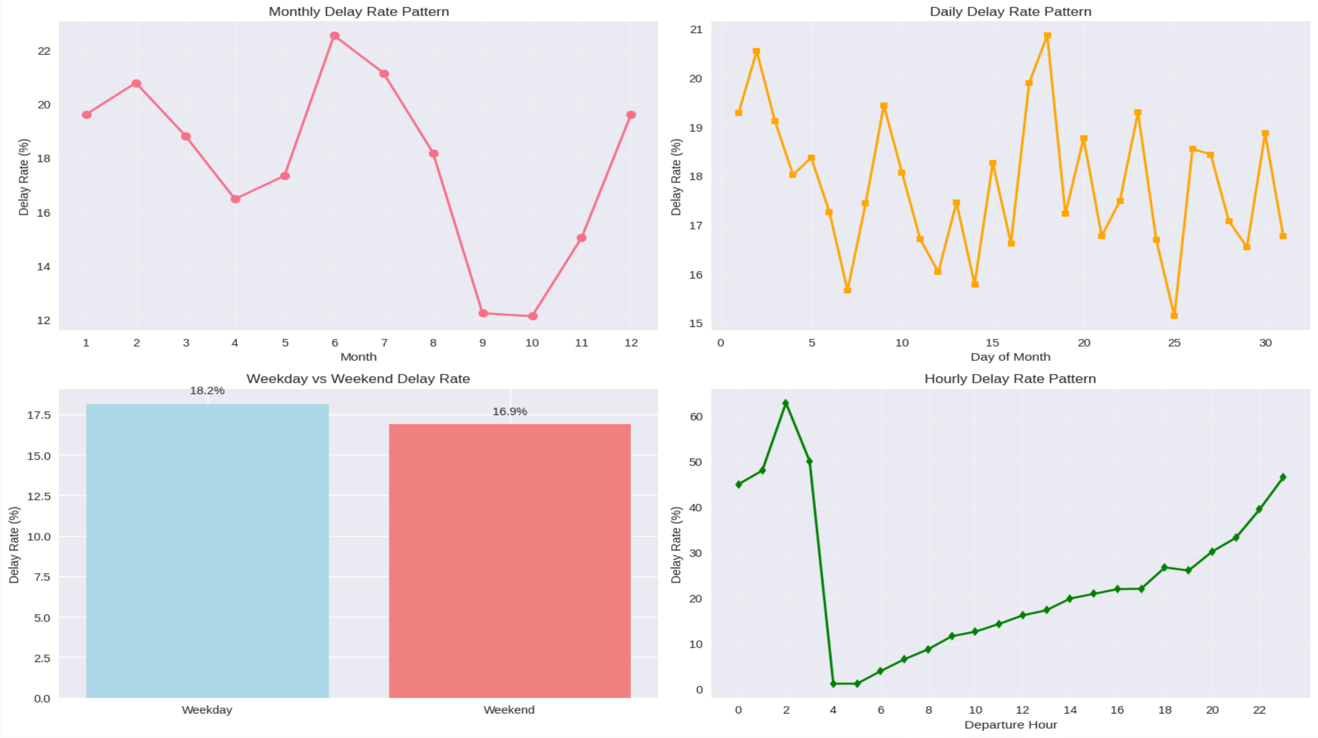
delay exposure

ct

delay probabilities

Purpose of the *Dataset*

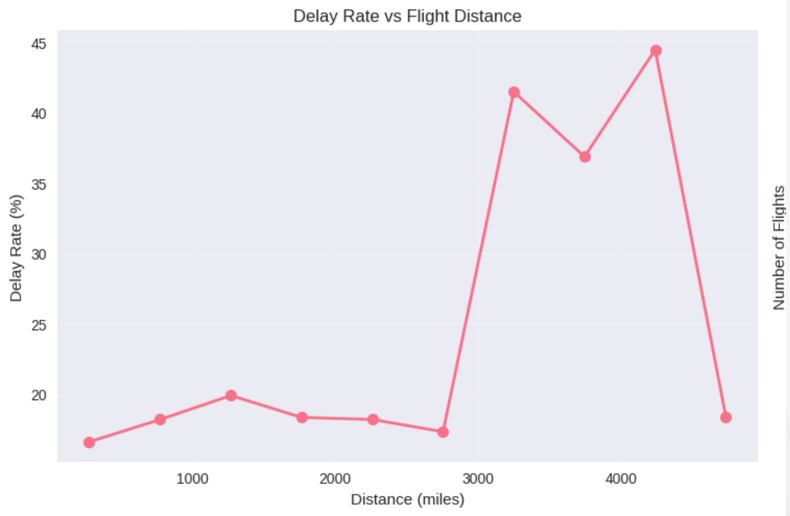
Hypotheses & Deep Dive Questions



- Time is an important dimension
- High variability & seasonality in delay rates as a function of time makes forecasting a challenge
- Why are Delay rates high in June & July?
- Does weekday scheduling work differently from weekends?
- Delay rates increase linearly from 4 AM. Is the effect cascading? Is the graveyard shift the leveler?

Purpose of the *Dataset*

Hypotheses & Deep Dive Questions

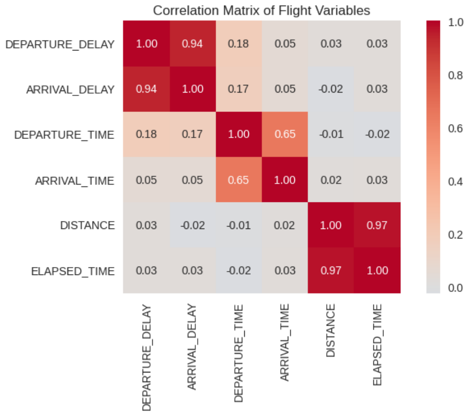


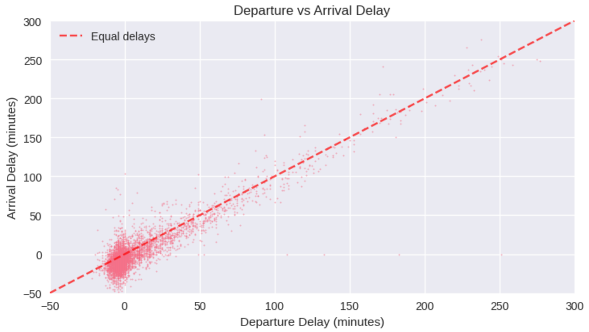
- There is a massive jump in delay rate for long distance flights !
- Was there a skew from outliers?
- Were there issues with connecting flights for flights with layover?

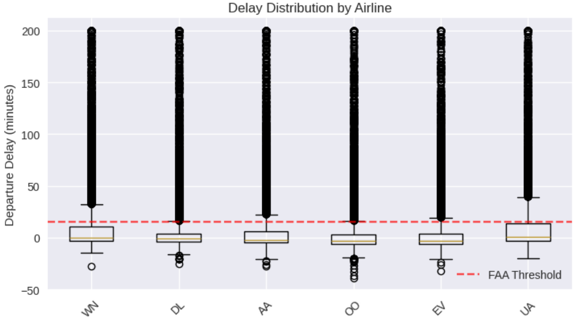
Purpose of the *Dataset*

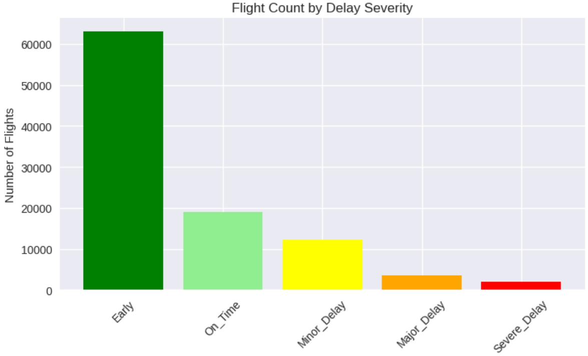
Hypotheses & Deep Dive Questions

Breaking down delays









While the arrival and departure delays are correlated, it is not a 100%, so departure delays have more contributors. And cascading delays for ground staff may be a big factor.

Minor - 15 min
Major - 60 min
Severe - 2 hr and more

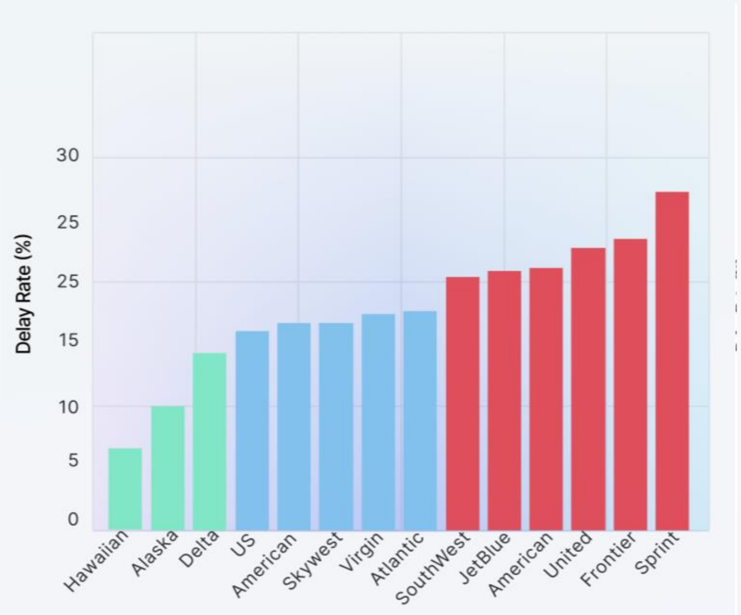
What are the cost of these delays?

Visualization Journeys

Tell the audience what you expect to happen…

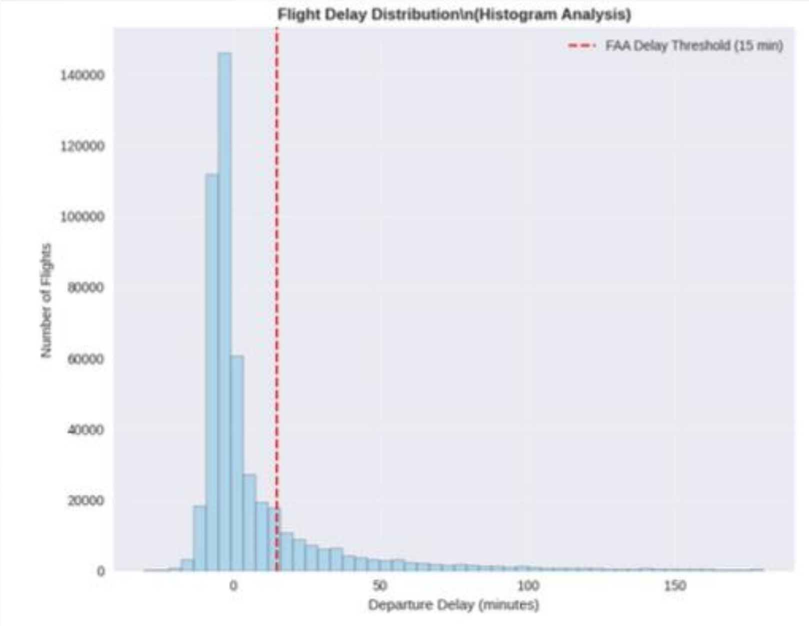
Visualization during *Processing & Preparation – sheet 1*

Airline Performance Comparison (Bar Chart Analysis)



Description:

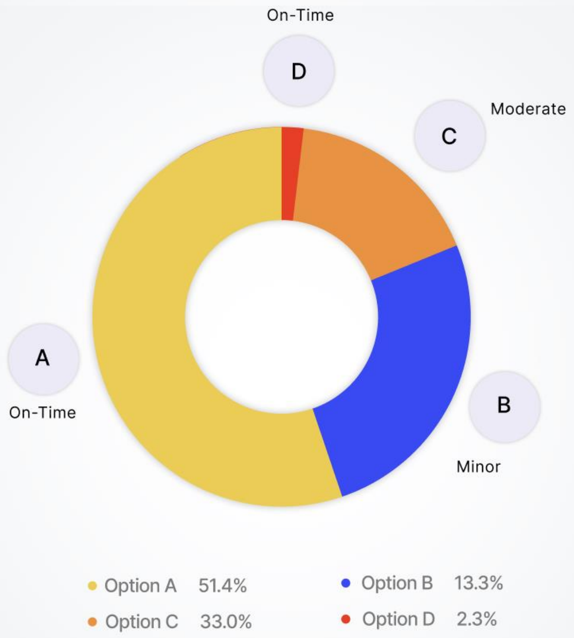
Flight Delay Distribution (Histogram Analysis)



Description:

Visualization during *Processing & Preparation- – sheet 3*
sheet 3

Flight Delay Severity Distribution



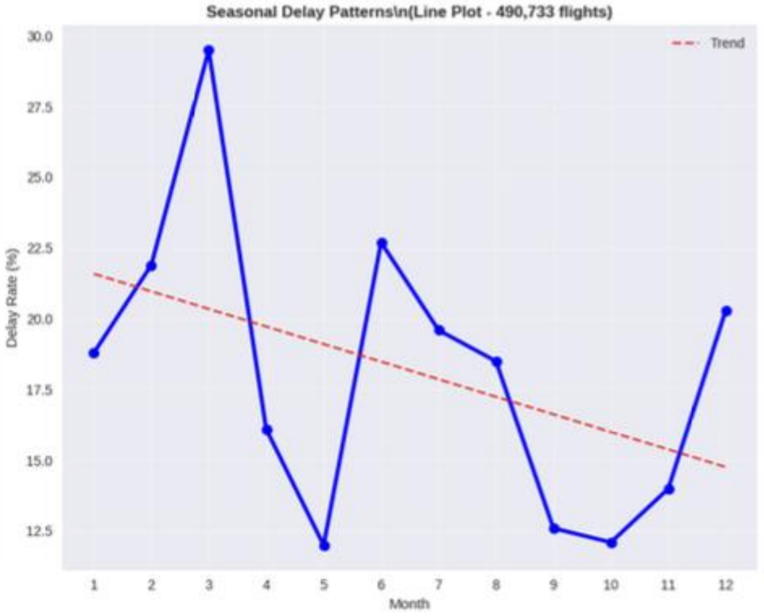
Description:

Top 10 Busiest Airport in Delay Response

Description:

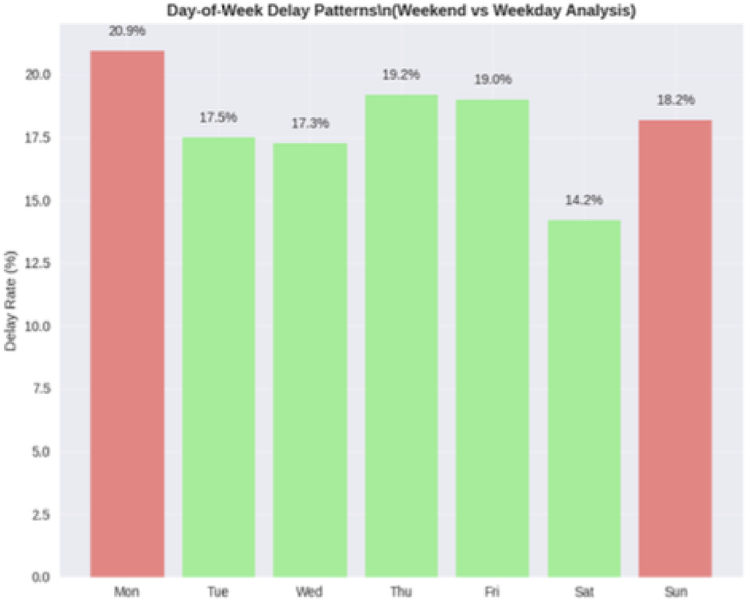
Base Visualization from *Raw Data– sheet 1*

Seasonal Delay Pattern Plot – 490,733 flights



Description:

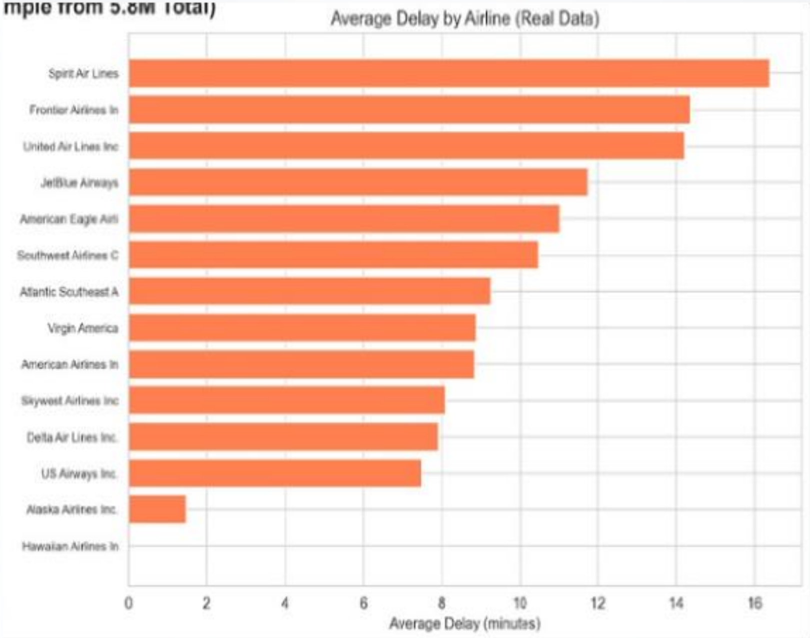
Day-of-Week Delay Patterns\n(Weekend vs Weekday Analysis



Description:

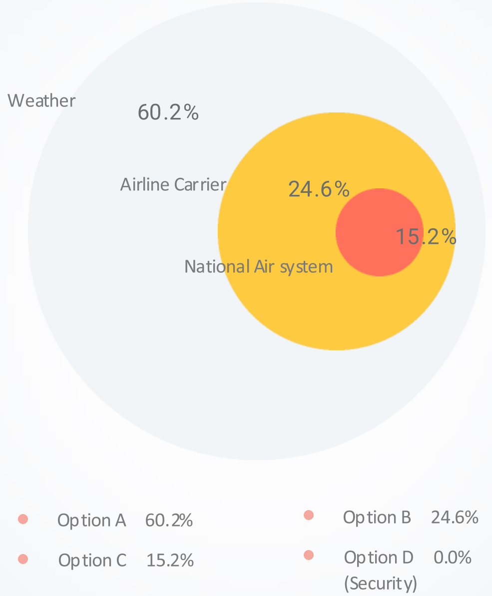
Base Visualization from *Raw Data– sheet 2*

Average Delay by Airline (Real Data)



Description:

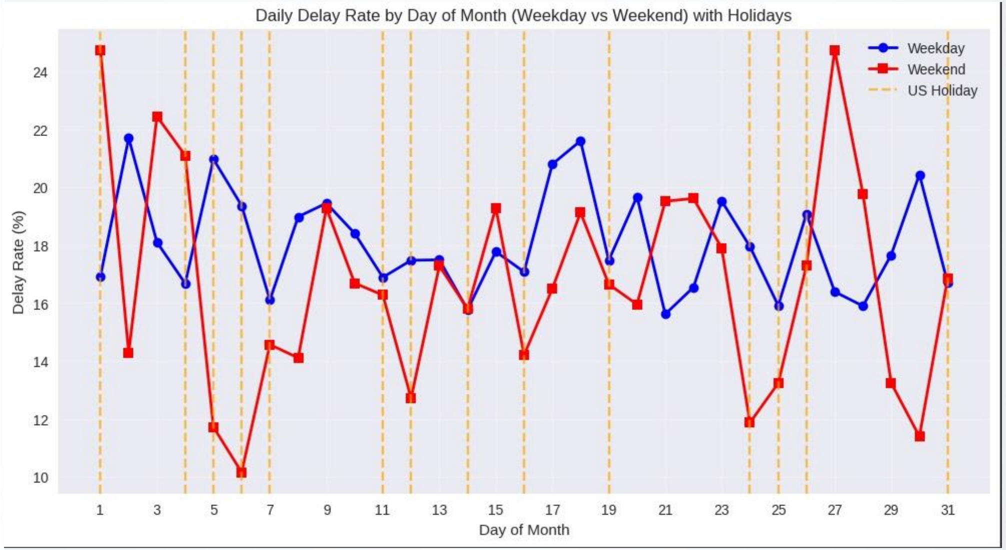
Flight Cancellation Reason (Pie Chart - 9,261 flights) flights)



Description:

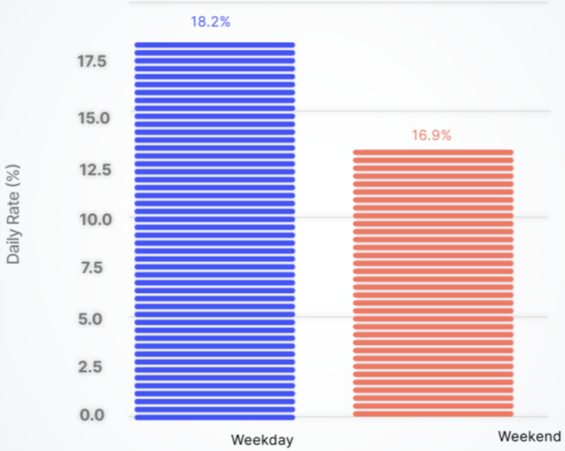
Base Visualization from *Raw Data– sheet 3*

Daily Delay Rate by Day of Month (Weekend vs Weekday) with Holidays



Description:

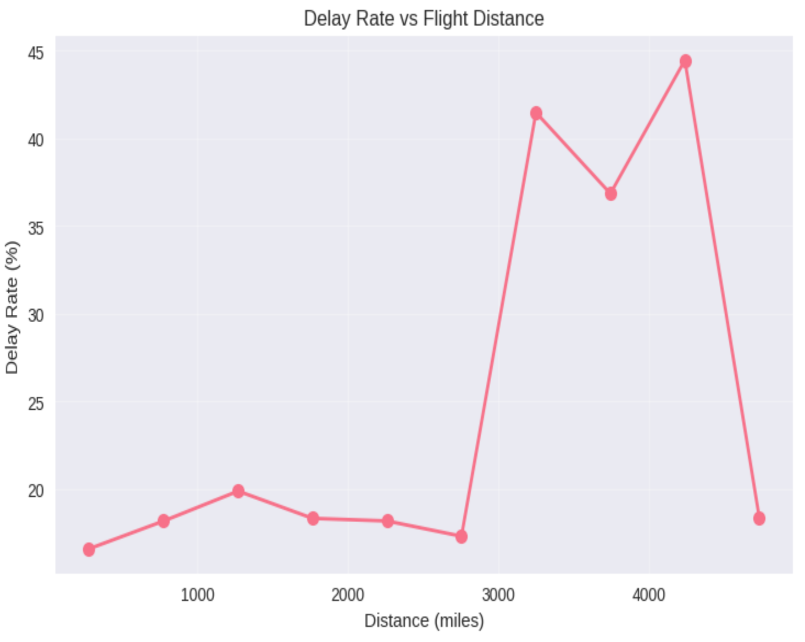
Weekday vs Weekend Delay Rate



Description:

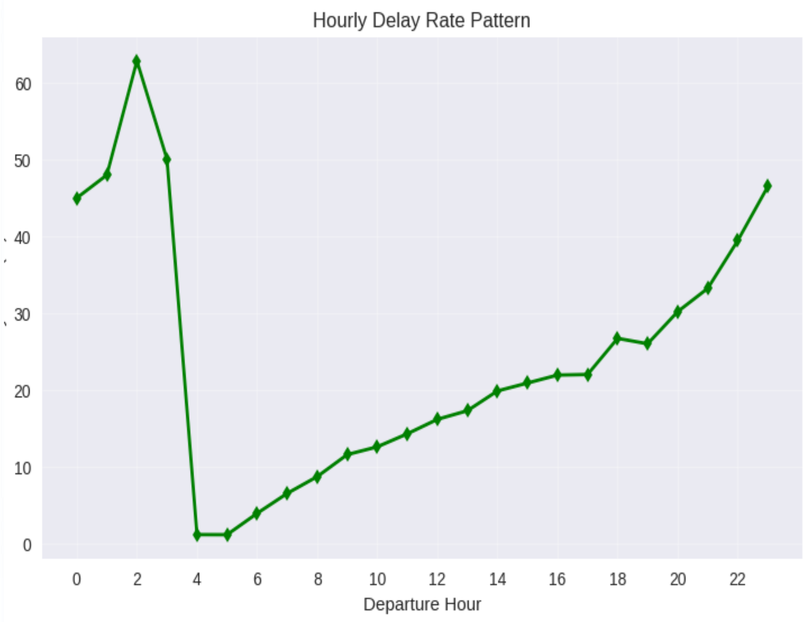
Base Visualization from *Raw Data– sheet 4*

Delay Rate vs Flight Distance



Tests the hypothesis that holidays and weekends influence delays.

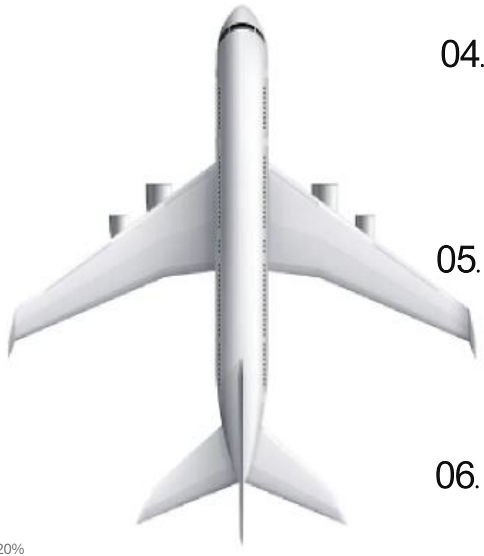
Hourly Delay Rate Pattern



Description:

Purpose of the *Dataset*

Hypotheses & Deep Dive Questions



01. Impact of Time of Day on Delays?

When Are Delays Most Frequent?

Identify whether mornings, afternoons, or evenings have the highest or lowest delay rates.

02. Long Delay Time Windows

Pinpoint High- Risk Hours

Discover Specific time blocks where most long delays (>2 hours) occur

03. Cost Saving from Delay Reduction

Quantifying Delay Reduction Benefits

Measure cost and efficiency gains from cutting long delays (>2 hours) by 20%

04. Delay Patterns by Day Type

Weekday vs. Weekend Delays

Compare delay likelihood and severity across weekends, and holidays

05. Best Days to Operate

Optimize Flight Scheduling

Recommend the lowest-risk days or time blocks for minimal delay exposure

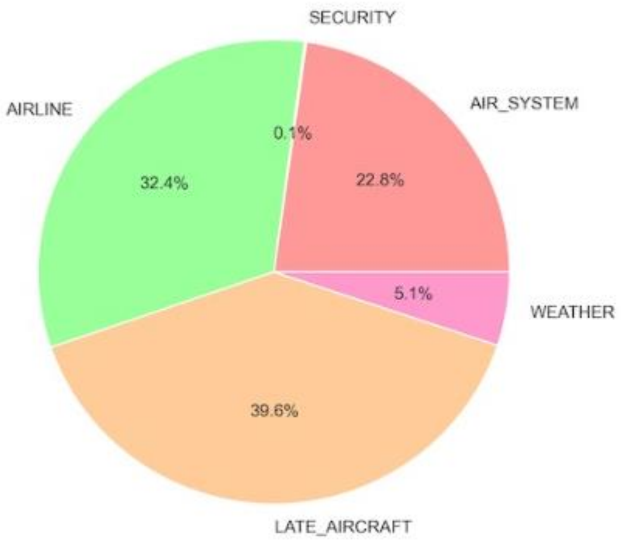
06. Combined Time & Day Impact

Delay Risk by Time & Time

Understand how day type and time of day interact to a affect delay probabilities

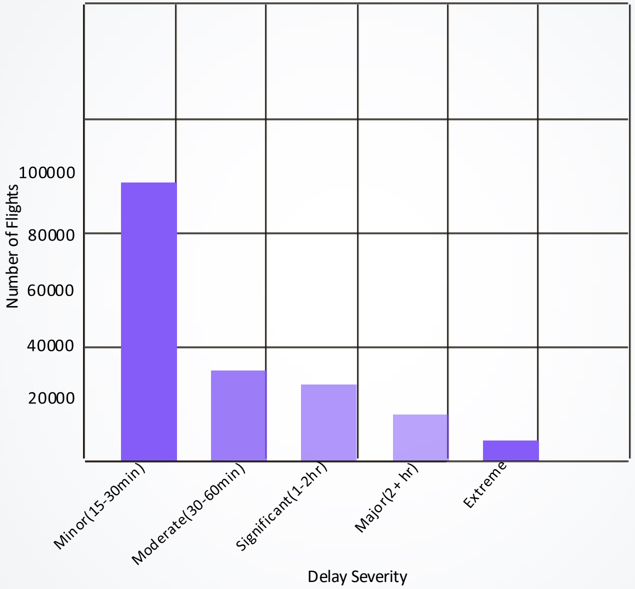
Exploratory Analysis – Sheet 1

Delay Distribution by Cause (Minutes pf Delays in (%)



The majority of delays are due to late aircraft (39.6%) and airline related issues (32.4%), followed by air traffic system (NAS) delays at (22.8%). Weather and security contribute minimally by <(6%).

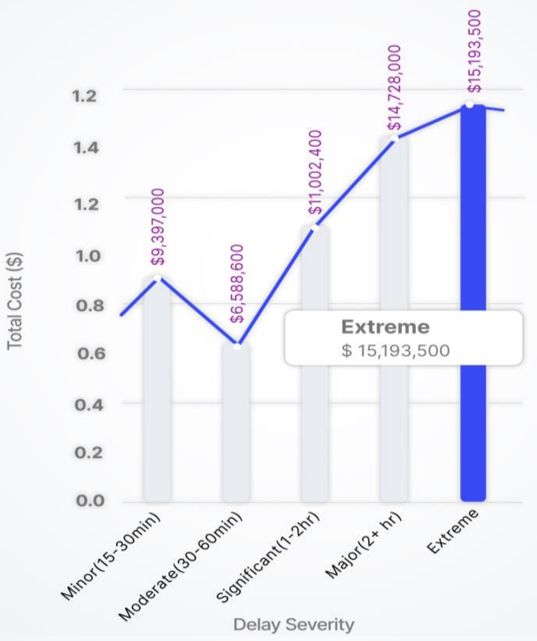
Flight Delays by Severity Level (Number of Flights)



Most delays are minor (<15 min), but thousands of flights still experience significant to extreme delays. However, these less frequent long delays cause major disruptions for both airlines and passengers.

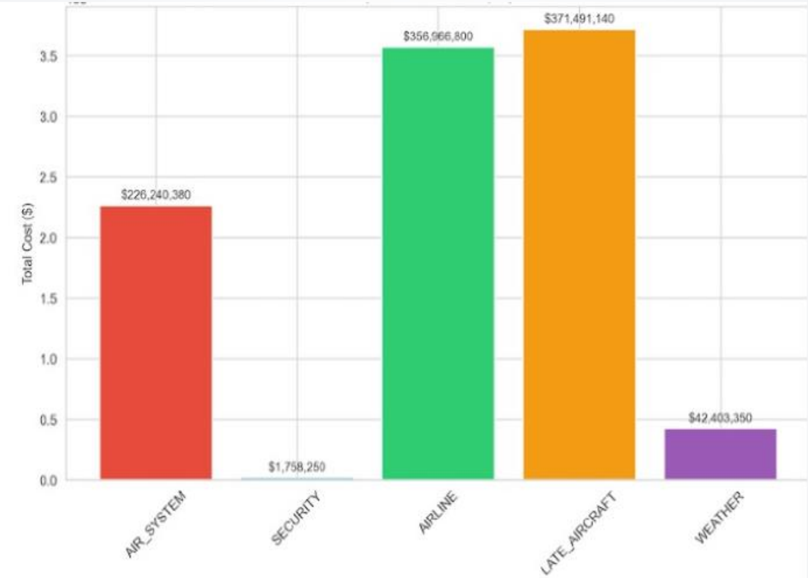
Exploratory Analysis – Sheet 2

Cost Impact by Delay Severity (Total Financial Impact)



Description:

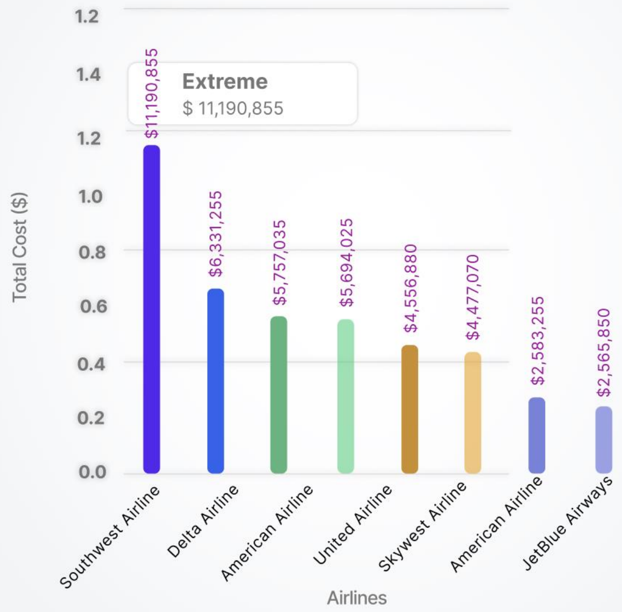
Financial Impact by Delay Type (Total Cost in sample)



Late aircraft delays result in the highest financial impact ($371M), followed closely by airline related delays ($356M). Air traffic system (NAS) delays cost over $226M, while weather and security contribute minimally.

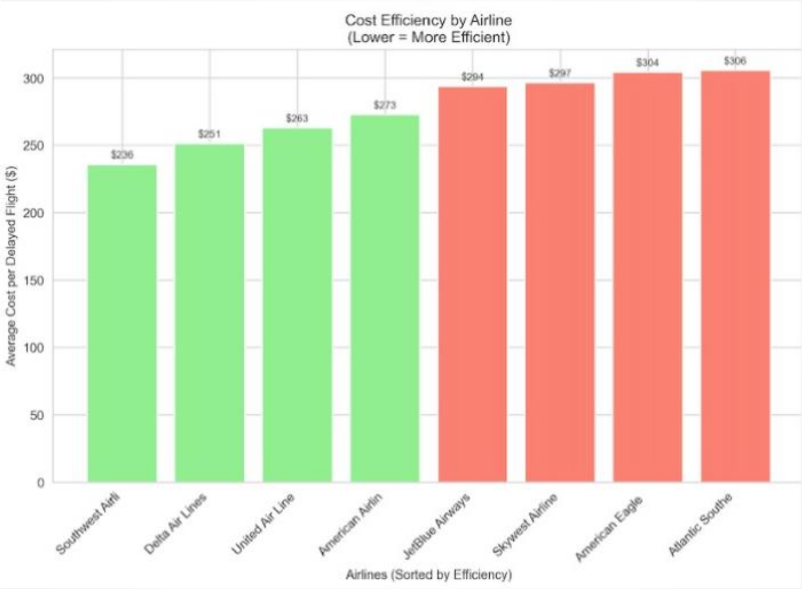
Exploratory Analysis – Sheet 3

Total Delay Costs by Airline (Sample Period))



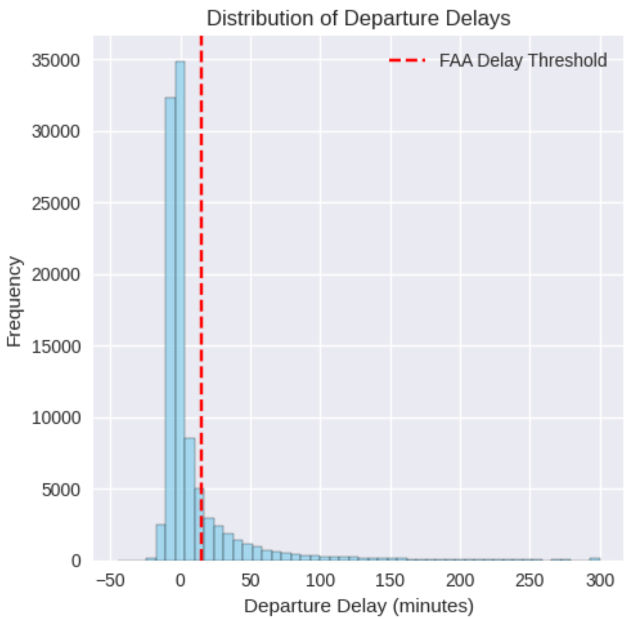
Description:

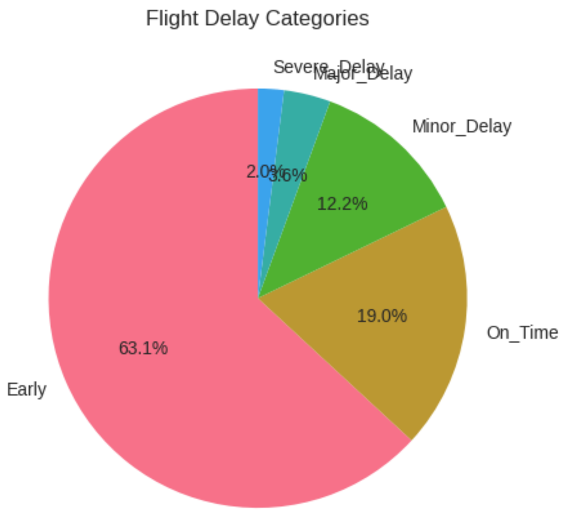
Cost Efficiency by Airline (Lower = more efficient)

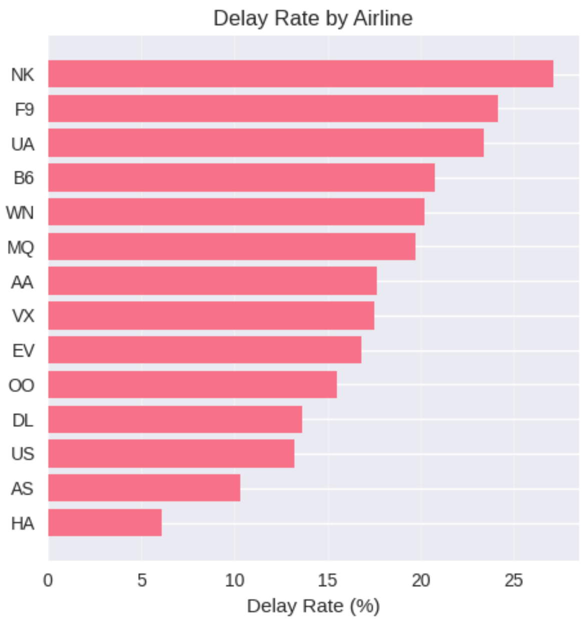


Description:

Overall Pattern of Flight Departure Delays





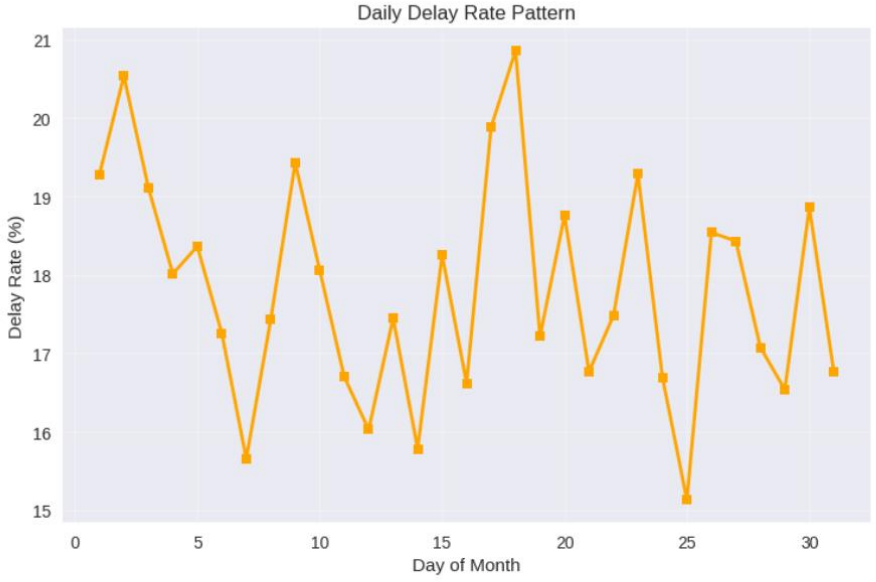


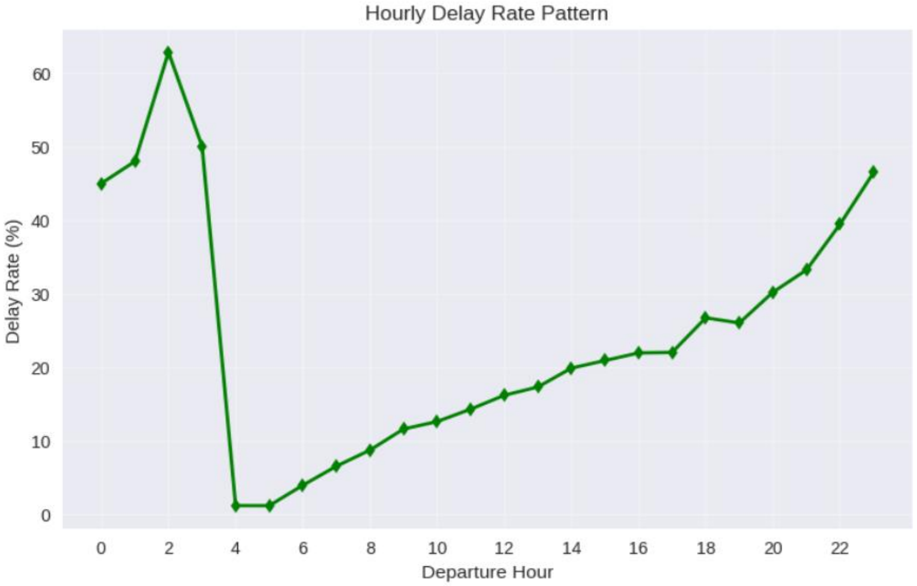
Most flights depart on time or with minor delays under 15 minutes. However, a long tail on the right shows that some flights experience major and even severe delays beyond the FAA's 15-minute threshold.

63% of flights depart early, 19% are on time, and the rest face delays of varying severity. While major and severe delays are rare, they are still operationally significant.

Related to performance, NK (Spirit Airlines), show high delay rates (25%), followed by F9 (Frontier) and UA (United) with rates above 20%. In contrast, HA (Hawaiian) and AS (Alaska) demonstrate better performance with delay rates below 10%. These performance differences will be explored further in later sections.

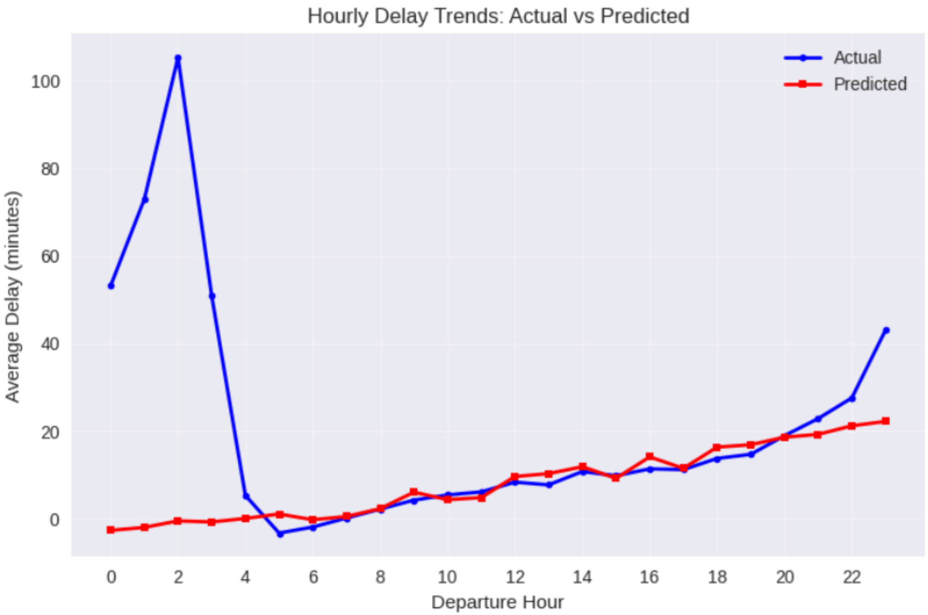
Daily and Hourly Delay Patterns

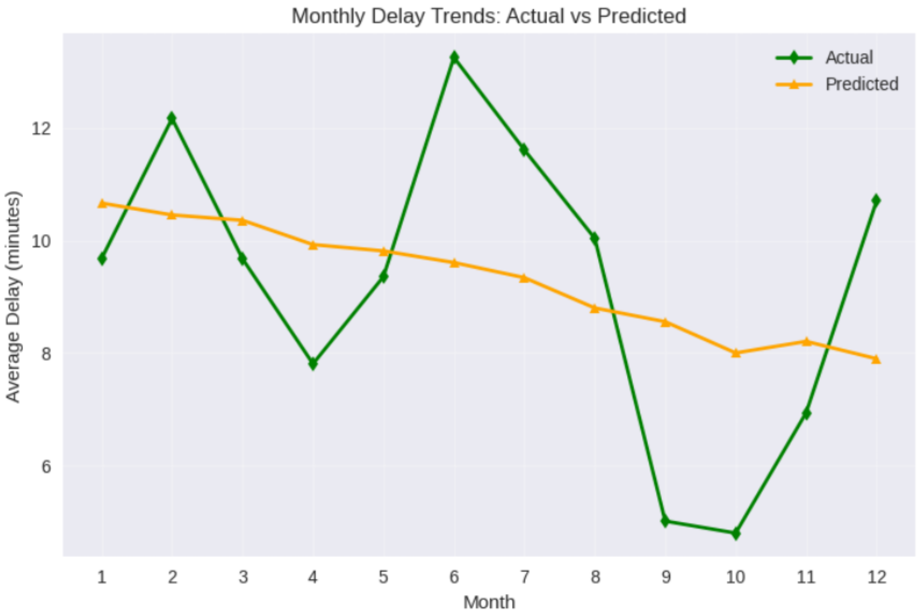


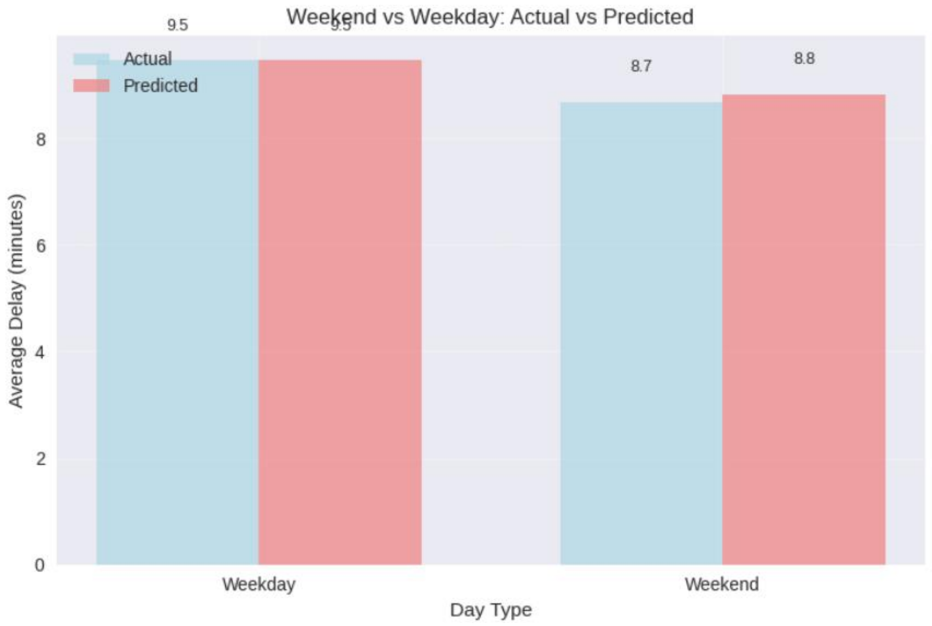


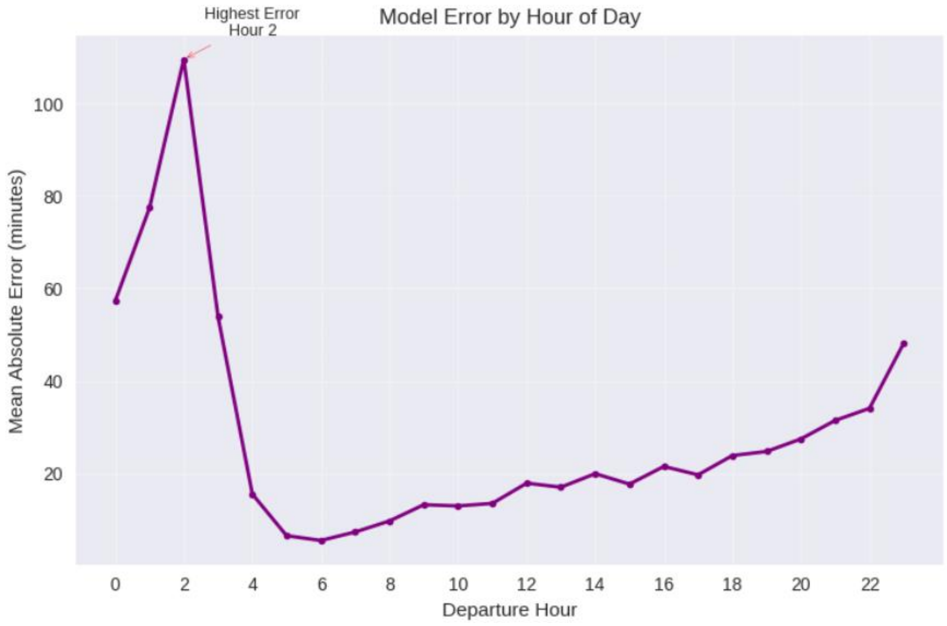
The delay rate varies throughout the month without a consistent trend. Peaks on certain days may reflect external factors such as weather conditions or operational bottlenecks.

Late night and early morning flights (12 AM–3 AM) show the highest delay rates (up to 63%).
After 4 AM, delays rates drop and stay lower through the day then start rising again in the evening.

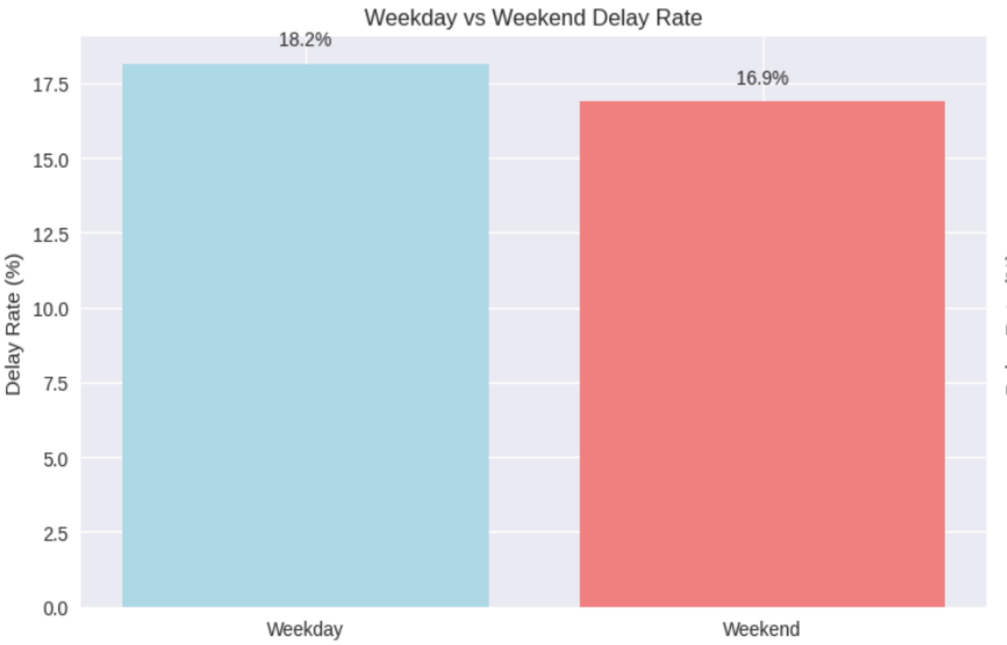

Hourly Delay Trends: Actual vs Predicted
Actual
Predicted
Average Delay (minutes)
100
80
60
40
20
0
0
2
4
6
8
10
12
14
16
18
20
22
Departure Hour


Monthly Delay Trends: Actual vs Predicted
Actual
Predicted
Average Delay (minutes)
12
10
8
6
1
2
3
4
5
6
7
8
9
10
11
12
Month


Weekend vs Weekday: Actual vs Predicted
Actual
Predicted
9.5
9.5
8.7
8.8
Average Delay (minutes)
0
2
4
6
8
Weekday
Weekend
Day Type


Model Error by Hour of Day
Highest Error
Hour 2
Mean Absolute Error (minutes)
20
40
60
80
100
0
2
4
6
8
10
12
14
16
18
20
22
Departure Hour

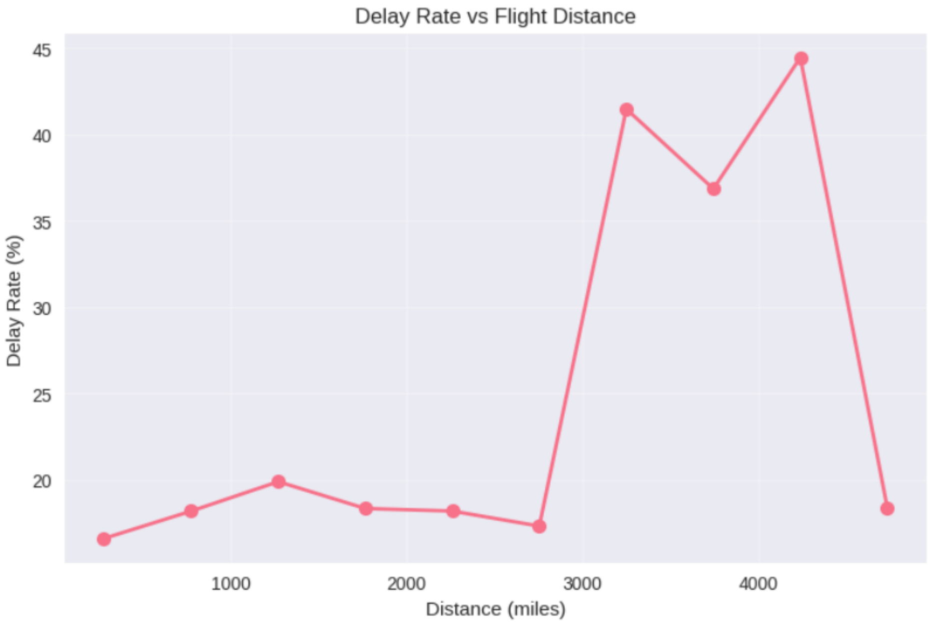
Delay Patterns by Month and Week



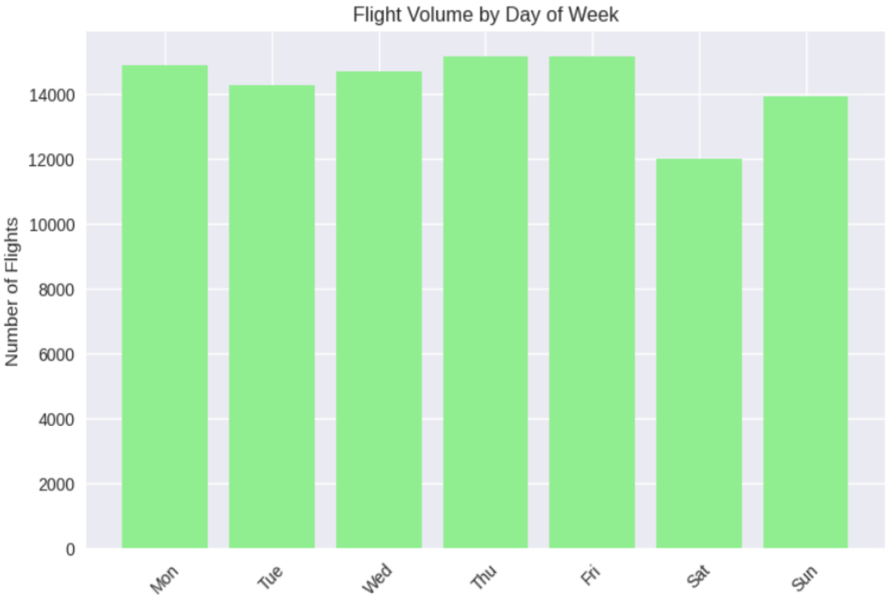
June and July see the highest delay rates, exceeding (22%) likely due to summer travel season. However, September and October show the lowest delay rates (12%).

Weekdays have a slightly higher delay rate than weekends. The difference isn't huge, but it suggests that delays might be a bit more common during the workweek, possibly due to heavier flight schedules.

Flight Distance and Volume Patterns

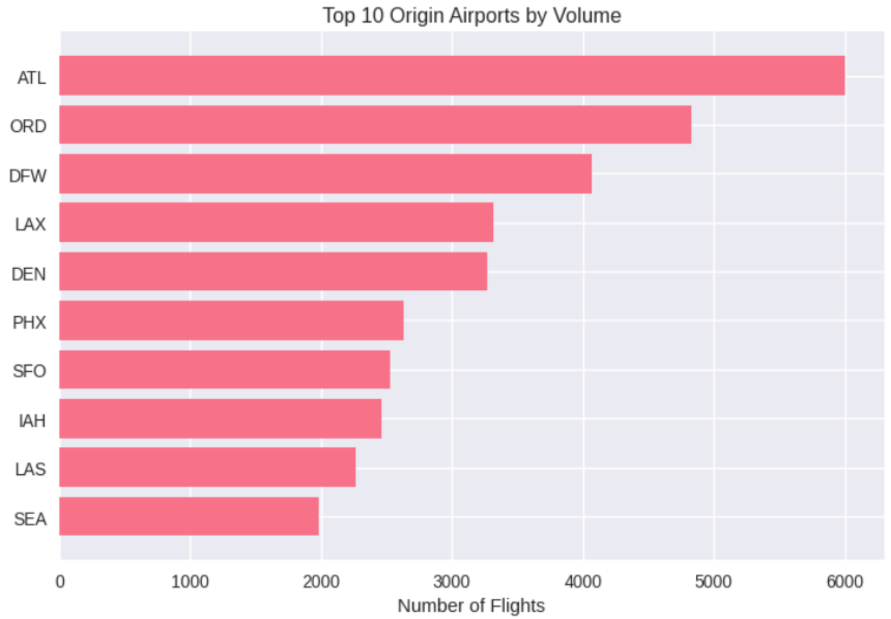


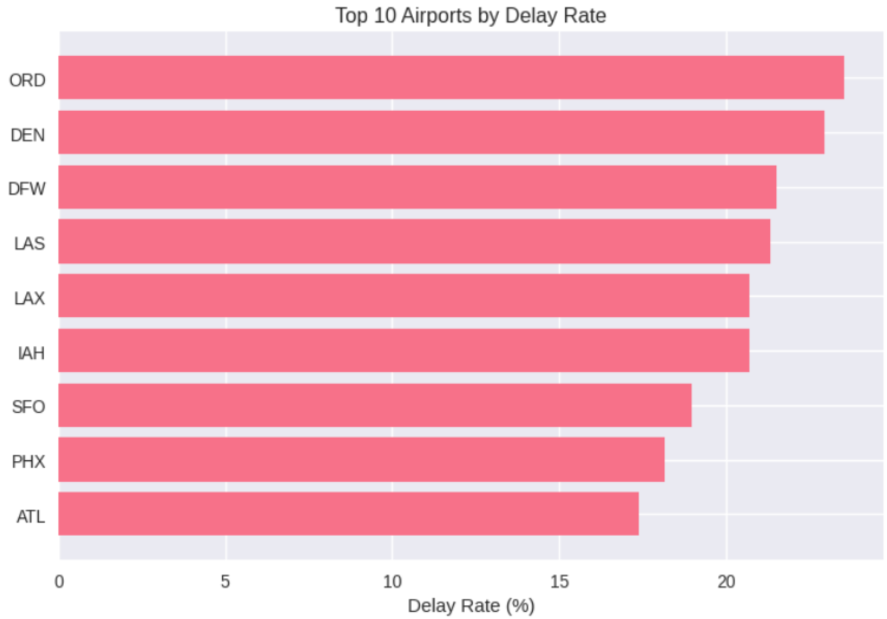
For flights under 3,000 miles, delay rates remain fairly stable. After 3,000 miles, there is a sharp increase in delays, reaching above 40%. For flights over 4,000 miles, delay rates drop again possibly because very long flights have more generous scheduling buffers which reducing the chance of arriving late.



Flight volumes remain consistently high from Monday through Friday, dip on Saturday and increase again on Sunday, reflecting common weekly travel patterns.

Top Airports by Flight Volume and Delay Rate

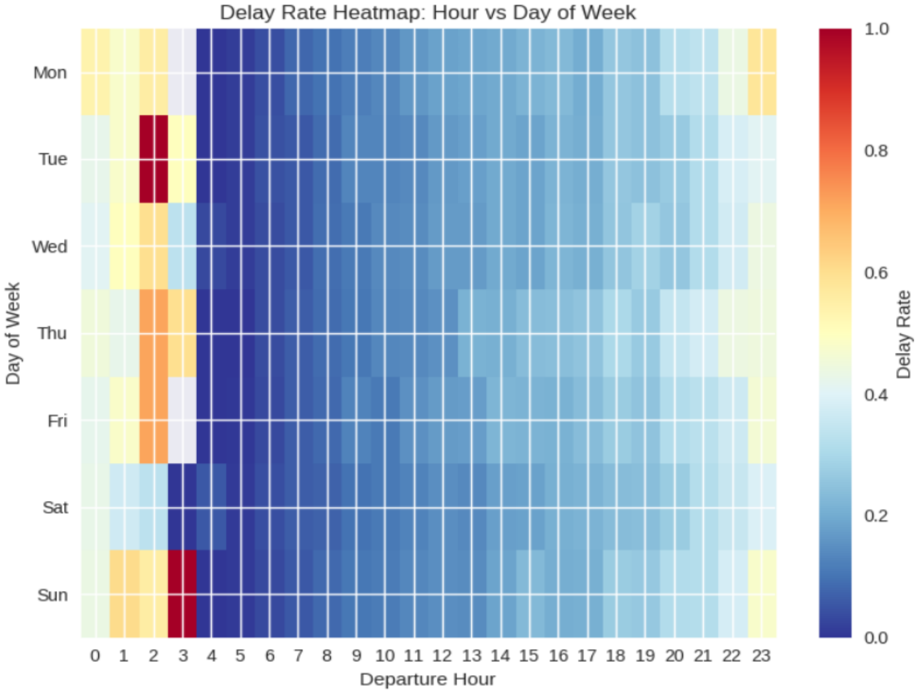


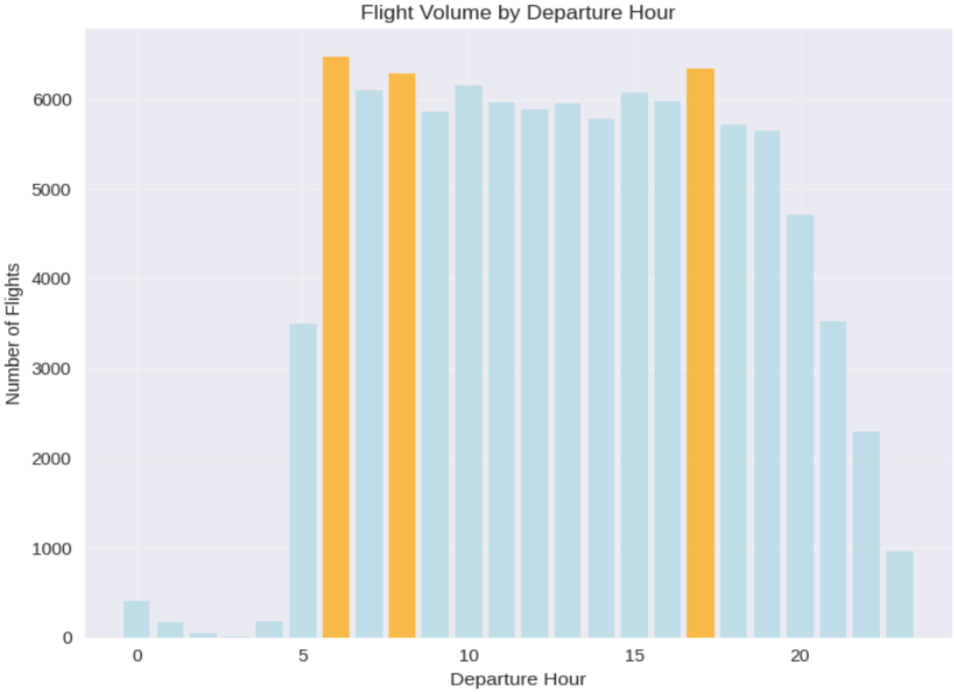


Atlanta (ATL) has the most departing flights by far, followed by Chicago O'Hare (ORD) and Dallas/Fort Worth (DFW). Other major airports like Los Angeles (LAX), Denver (DEN), and Phoenix (PHX) also handle large numbers of flights, reflecting their role as primary connection points in the national network.

Chicago O'Hare (ORD) records the highest delay rate among major airports exceeding 23%. Denver (DEN) and Dallas/Fort Worth (DFW) follow closely. Interestingly, some of the busiest hubs such as Atlanta (ATL) have relatively lower delay rates compared to their volume, indicating more efficient operations despite high traffic.

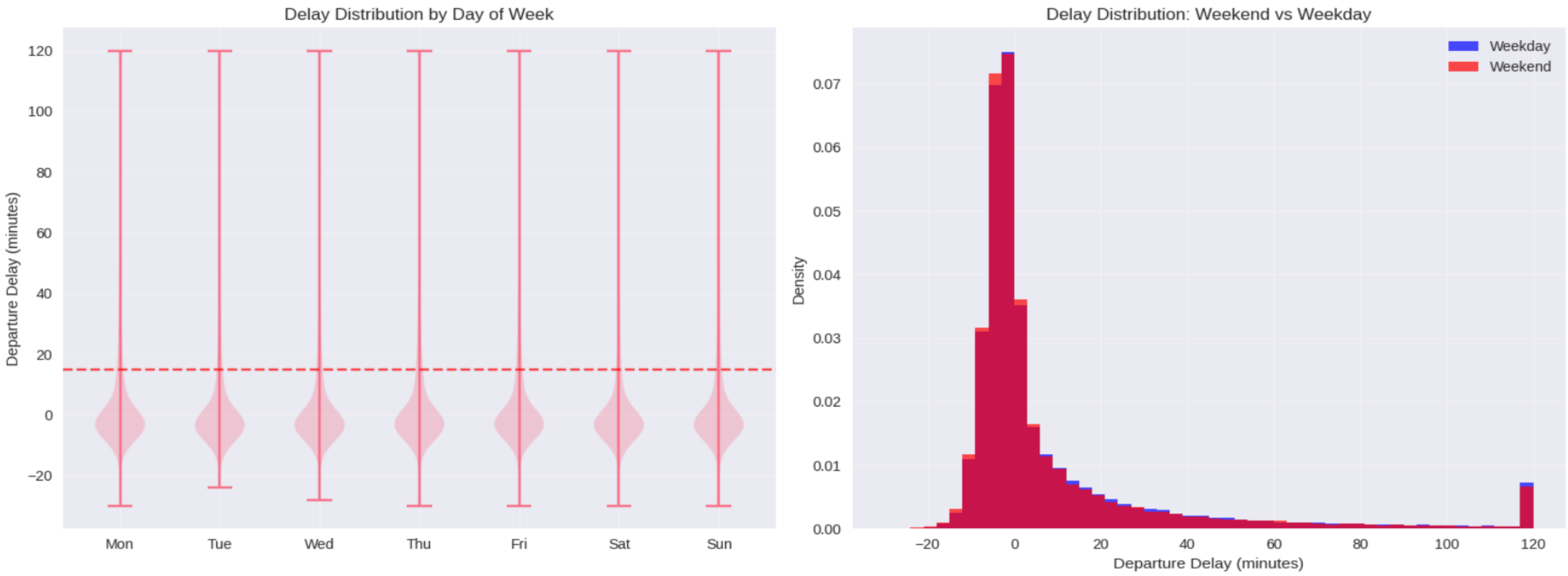
Flight Delays and Volumes by Time of Day



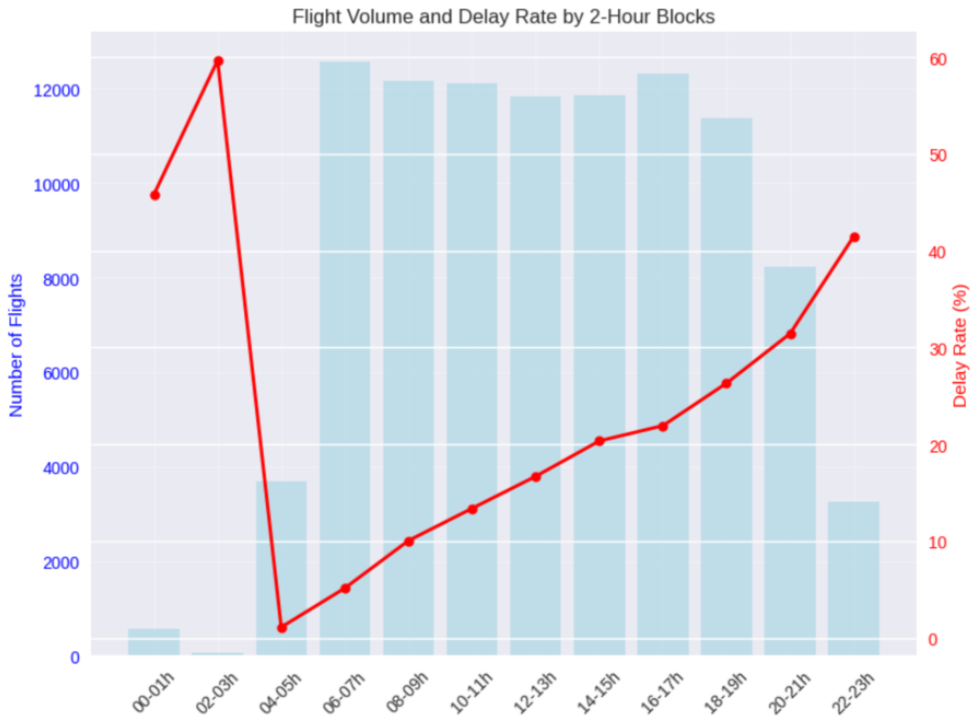


Delay rates are highest in the early morning hours between 2 AM and 4 AM, especially on Tuesdays and Sundays. This may be due to low flight volume, overnight operational constraints or reduced staffing. Midday to early evening hours generally show lower delay rates with a gradual increase again late at night.

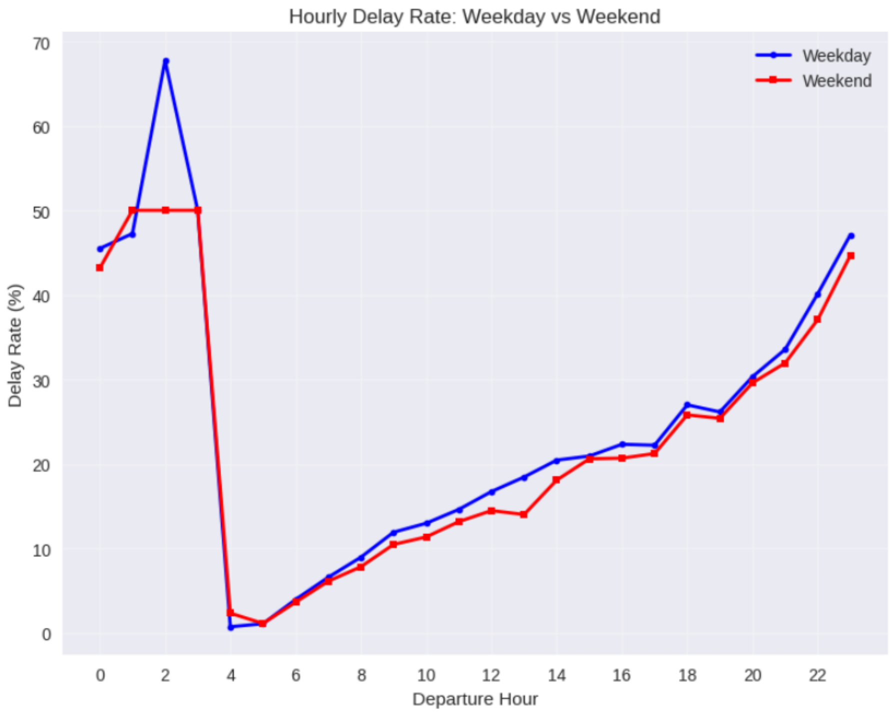
Flight volumes peak between 6 AM and 8 AM and again from 5 PM to 6 PM, reflecting common travel demand patterns for morning departures and evening returns. Volumes drop sharply after 9 PM and remain minimal overnight.



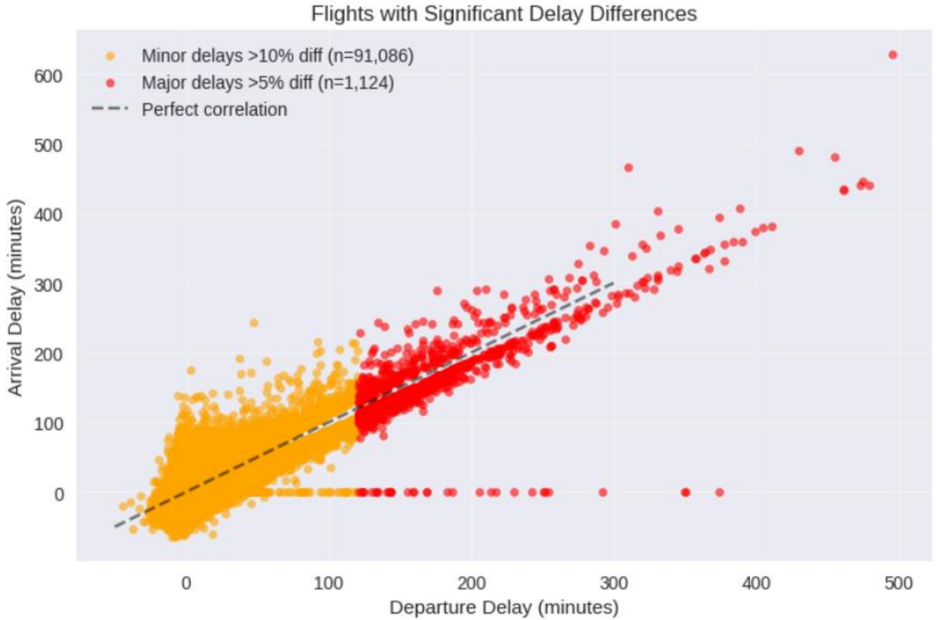
Delay Distribution by Day of Week
Departure Delay (minutes)
120
100
80
60
40
20
0
−20
Mon
Tue
Wed
Thu
Fri
Sat
Sun
Delay Distribution: Weekend vs Weekday
Weekday
Weekend
Density
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
−20
0
20
40
60
80
100
120
Departure Delay (minutes)

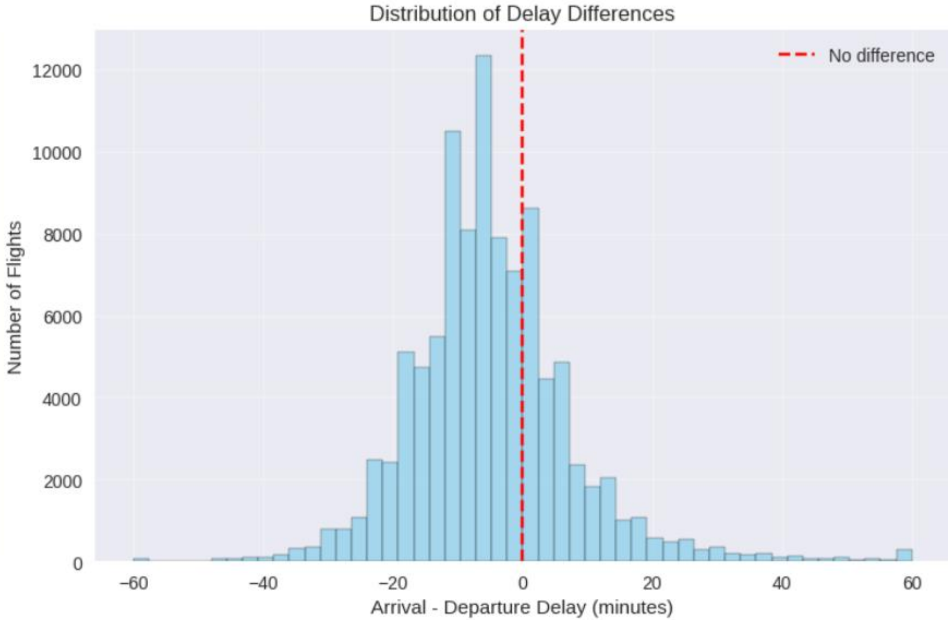


Delay rates peak between 2–3 AM despite low flight counts, stay low during the morning rush, then rise again late at night.



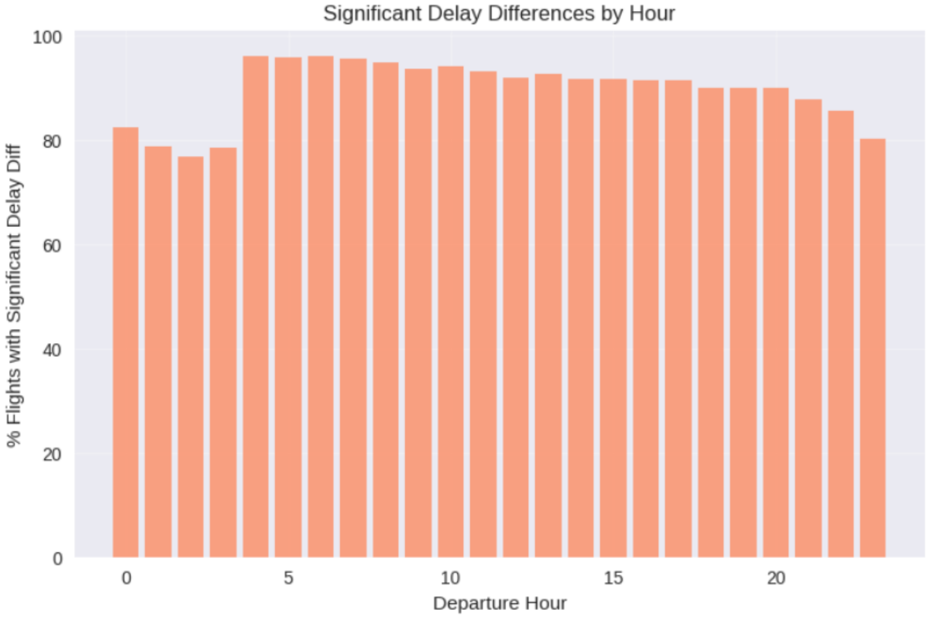
Both spike at 2–3 AM, with the spike higher on weekdays; weekends tend to have slightly lower delays during evening hours.

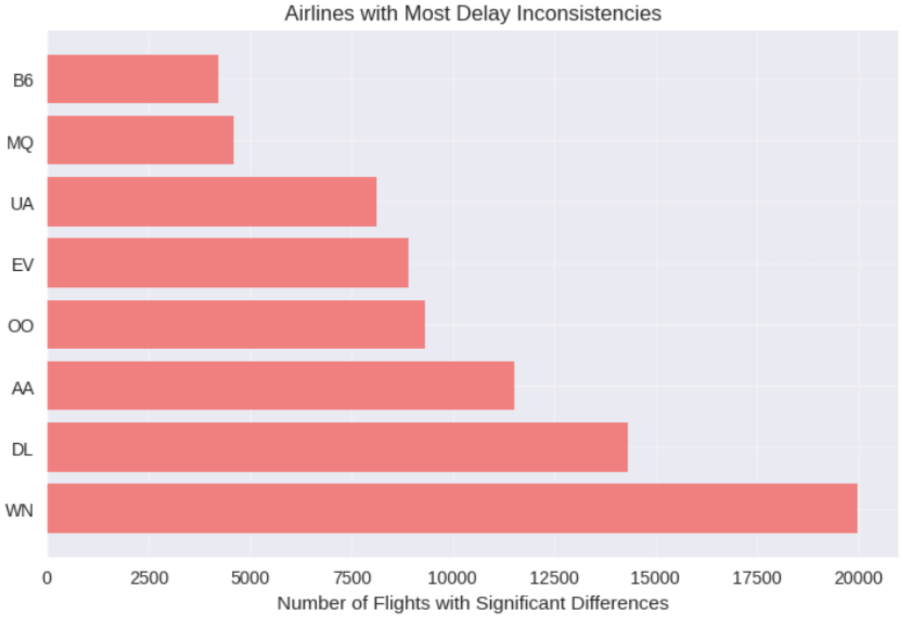


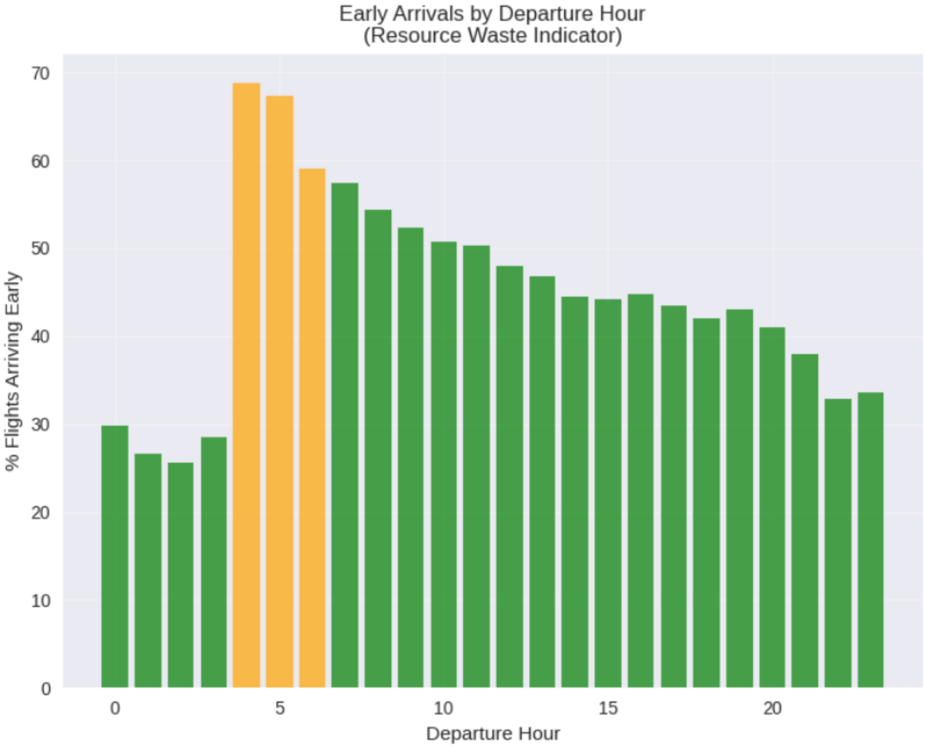


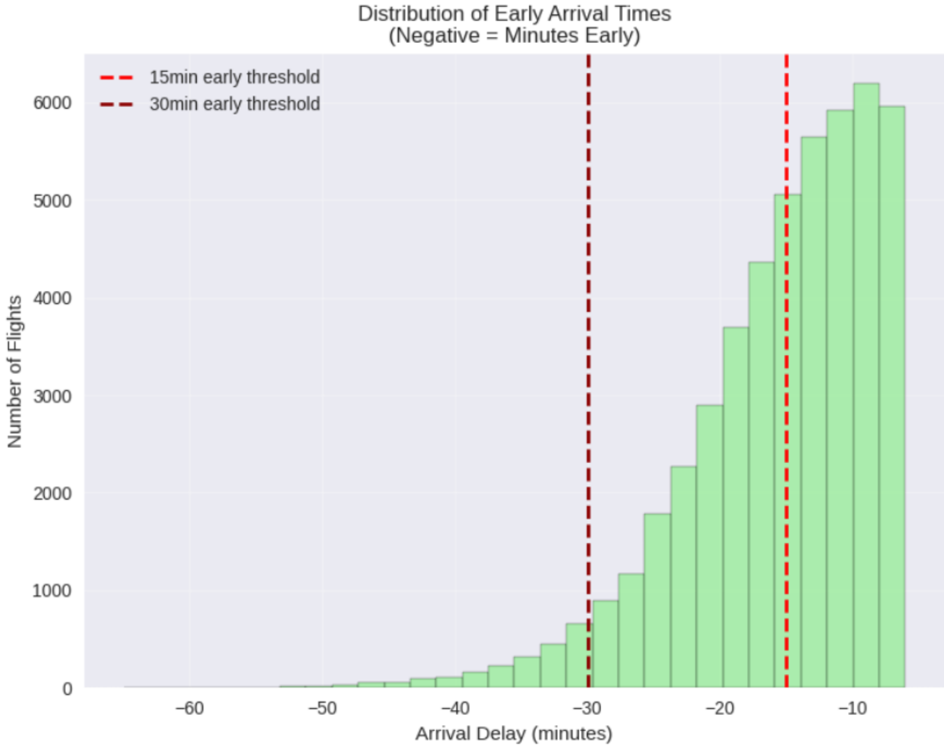
Significant Delay Differences by Hour

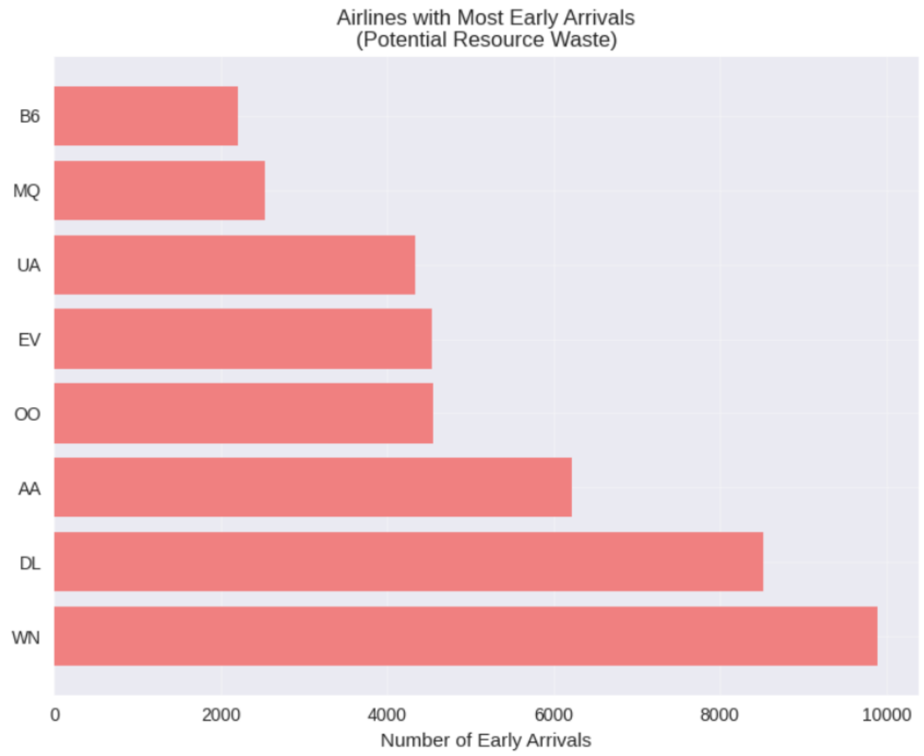
Airlines with Most Delay Inconsistencies

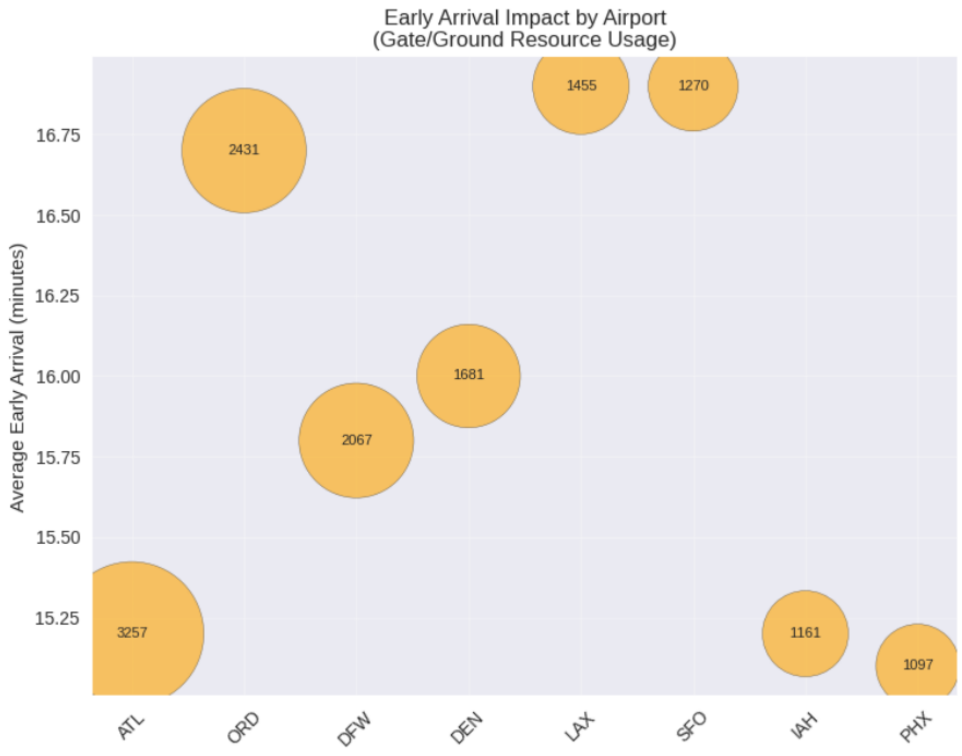

Significant Delay Differences by Hour
% Flights with Significant Delay Diff
Departure Hour
0
5
10
15
20
100
80
60
40
20


Airlines with Most Delay Inconsistencies
B6
MQ
UA
EV
OO
AA
DL
WN
Number of Flights with Significant Differences
0
2500
5000
7500
10000
12500
15000
17500
20000


Early Arrivals by Departure Hour
(Resource Waste Indicator)
% Flights Arriving Early
Departure Hour
0
10
20
30
40
50
60
70
0
5
10
15
20

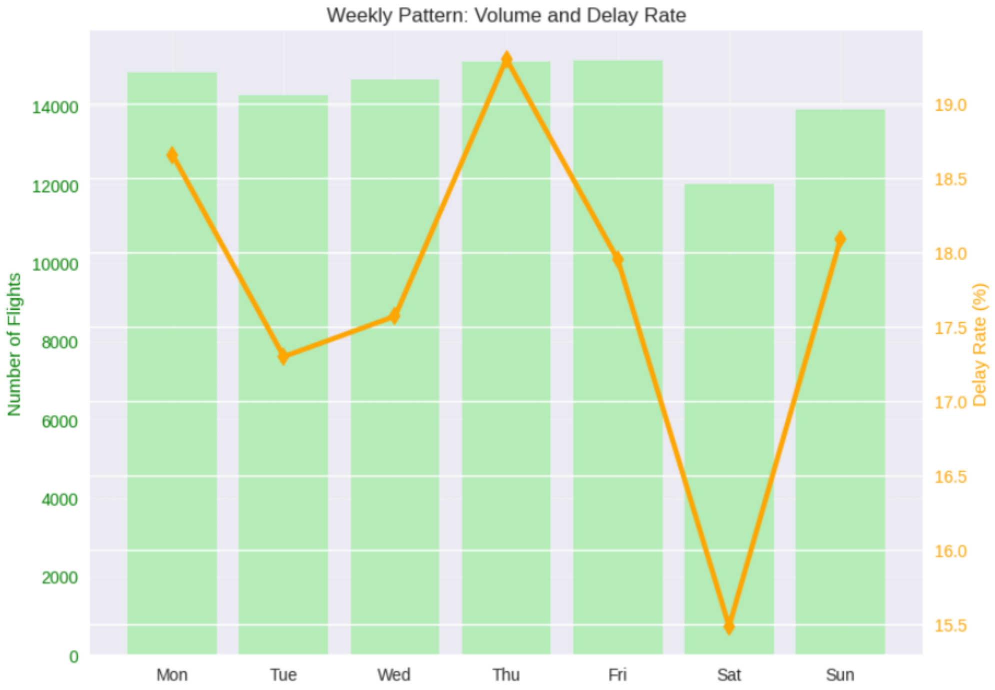

Distribution of Early Arrival Times
(Negative = Minutes Early)
15min early threshold
30min early threshold
Number of Flights
Arrival Delay (minutes)
0
1000
2000
3000
4000
5000
6000
−60
−50
−40
−30
−20
−10

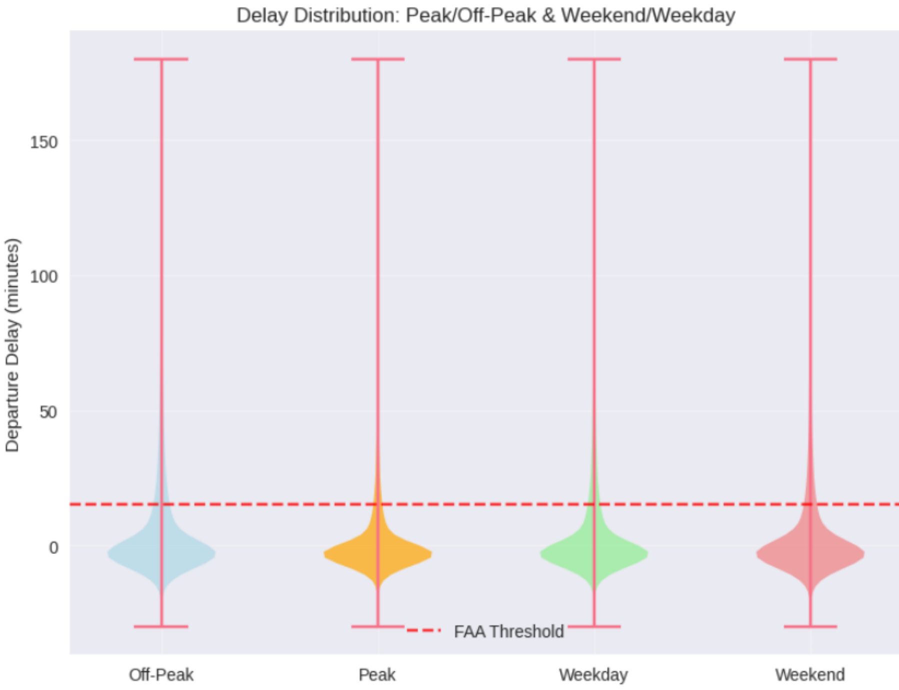


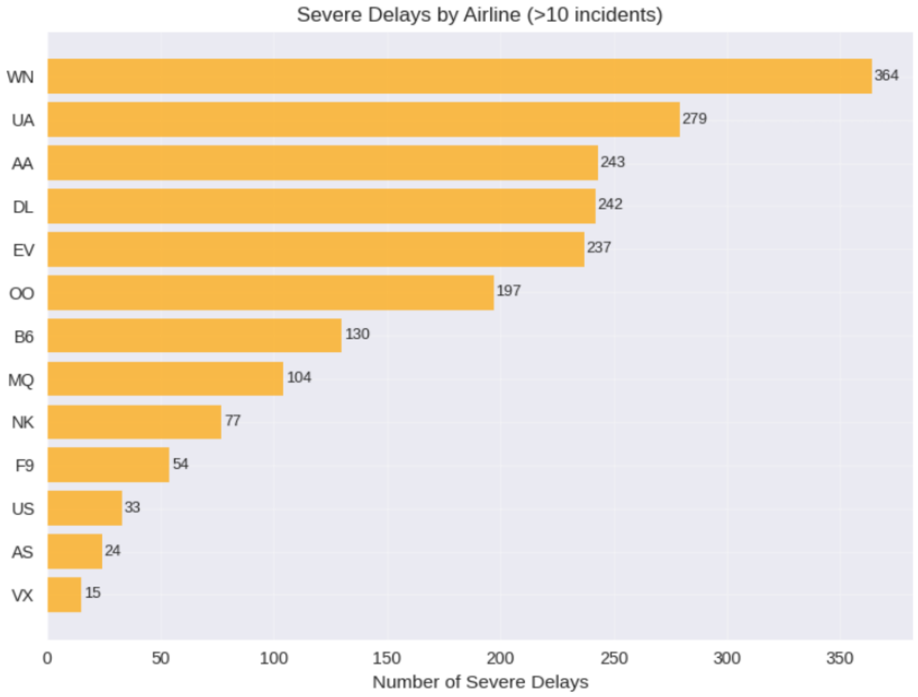


WN (Southwest) has the highest number of early arrivals (nearly 10,000), followed by DL (Delta) and AA (American Airlines). These patterns could indicate potential resource inefficiencies, as gates and ground crews may not be ready before scheduled times.

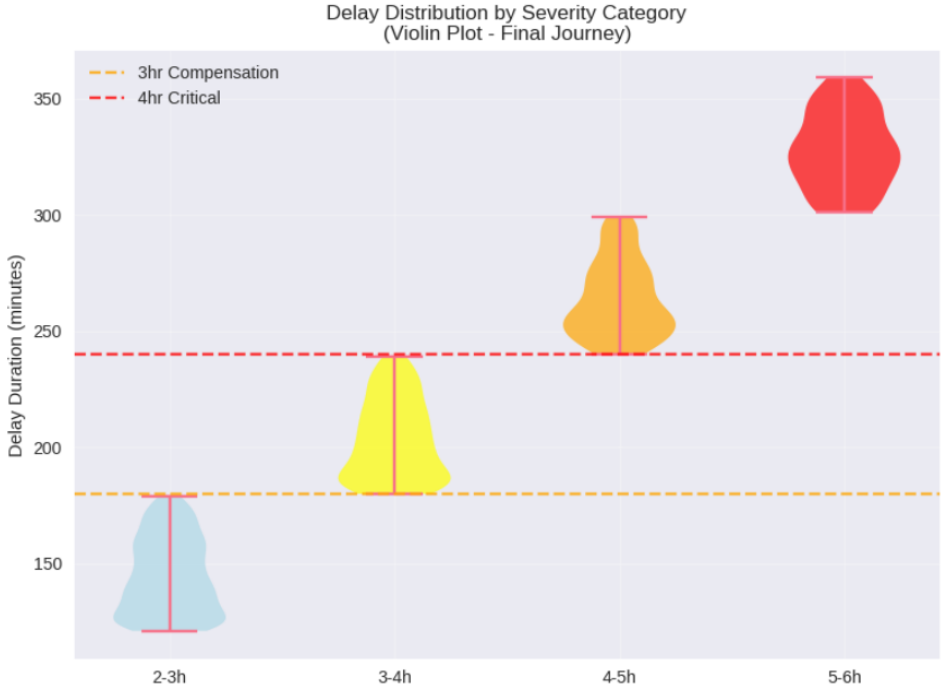
ATL experiences the largest total number of early arrivals (3,257), but ORD (Chicago O'Hare) shows the highest average early arrival time (~16.7 minutes). Airports like LAX and SFO have fewer early arrivals but still average over 16.7 minutes early.


Weekly Pattern: Volume and Delay Rate
Number of Flights
Delay Rate (%)
0
2000
4000
6000
8000
10000
12000
14000
15.5
16.0
16.5
17.0
17.5
18.0
18.5
19.0
Mon
Tue
Wed
Thu
Fri
Sat
Sun

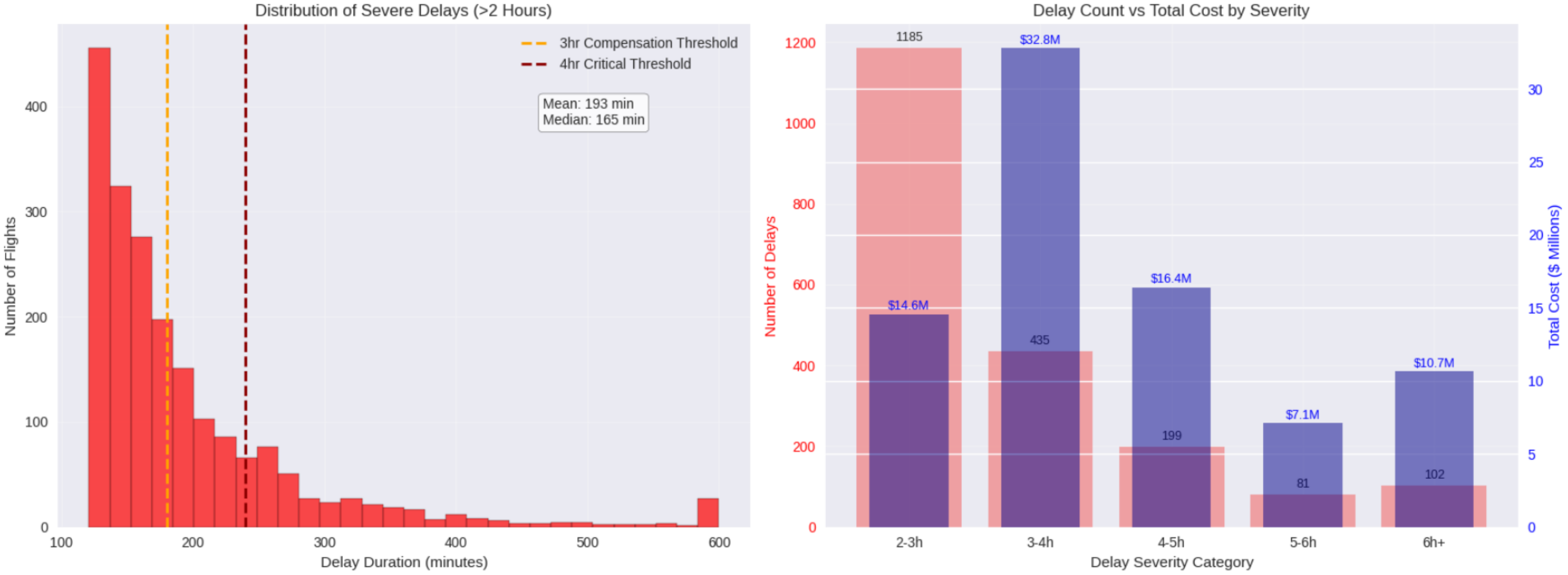

Delay Distribution: Peak/Off-Peak & Weekend/Weekday
Departure Delay (minutes)
0
50
100
150
FAA Threshold
Off-Peak
Peak
Weekday
Weekend

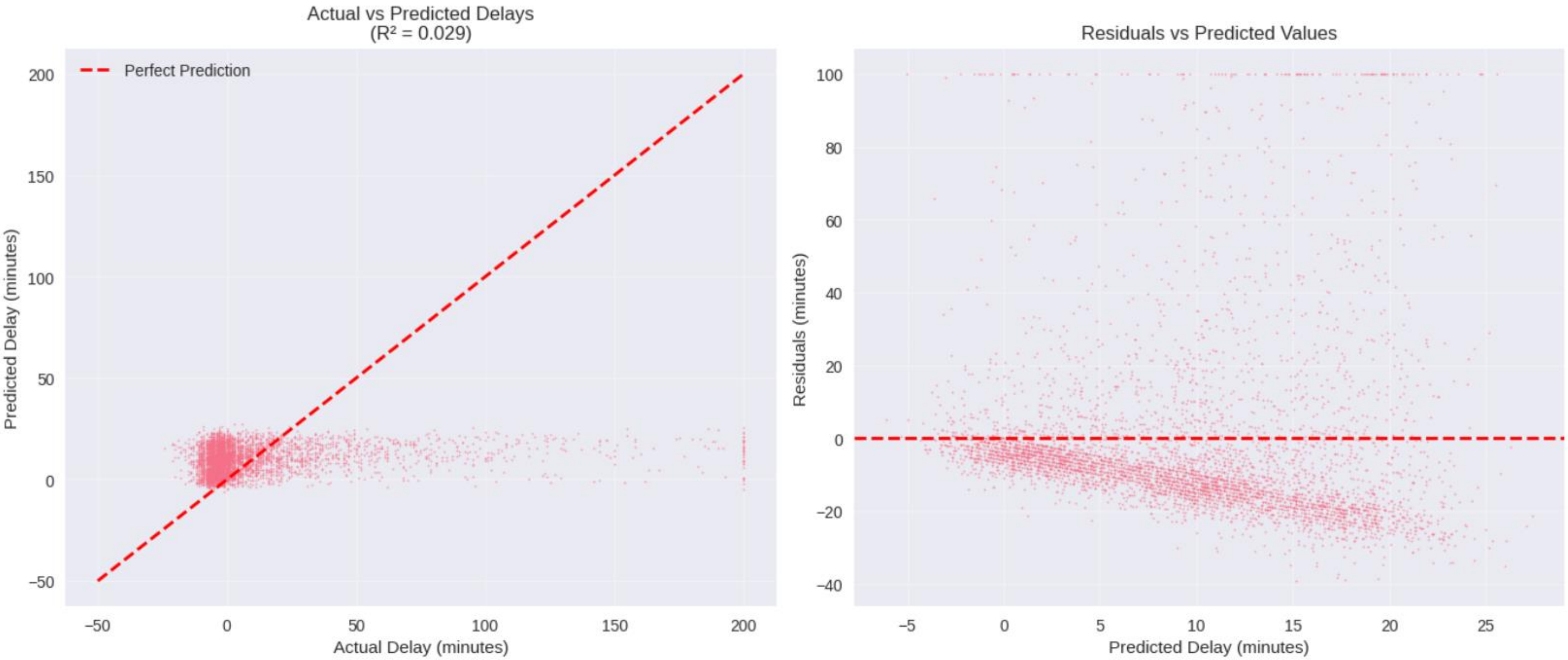


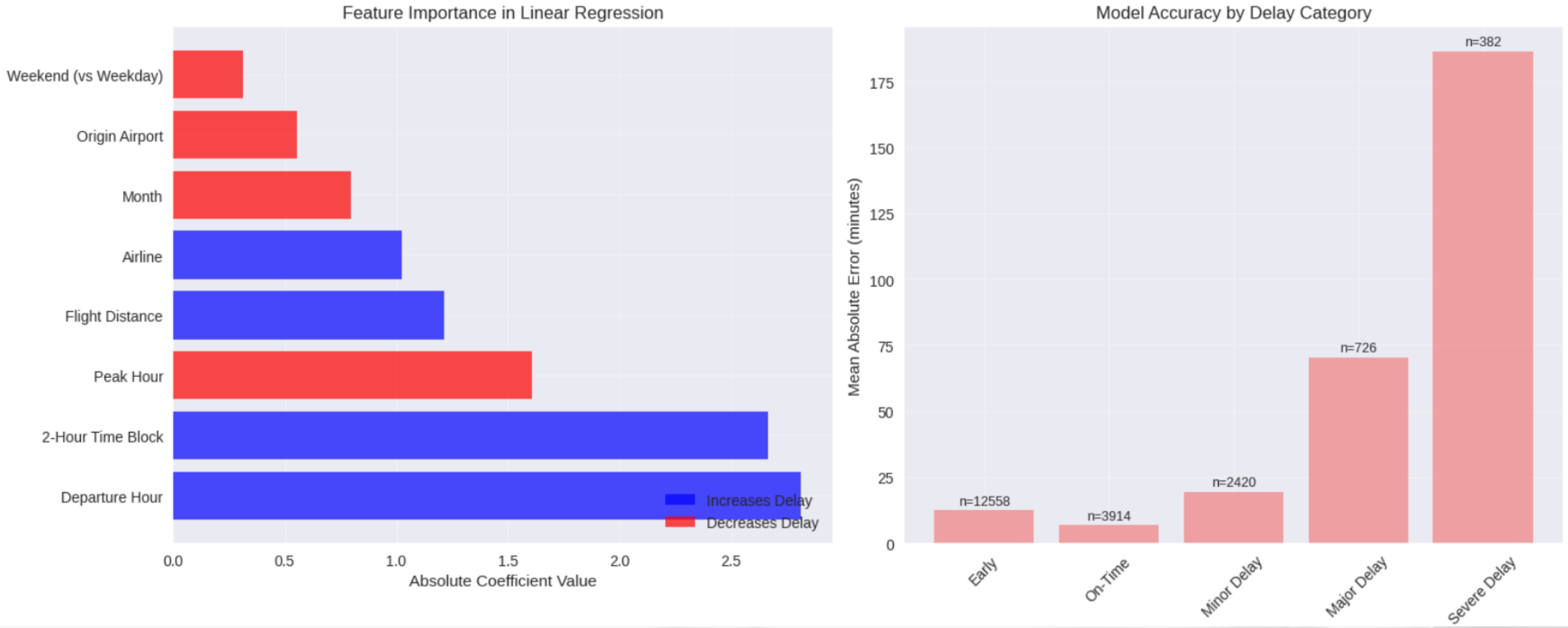
Severe Delays by Airline (>10 incidents)
WN
UA
AA
DL
EV
OO
B6
MQ
NK
F9
US
AS
VX
364
279
243
242
237
197
130
104
77
54
33
24
15
0
50
100
150
200
250
300
350
Number of Severe Delays



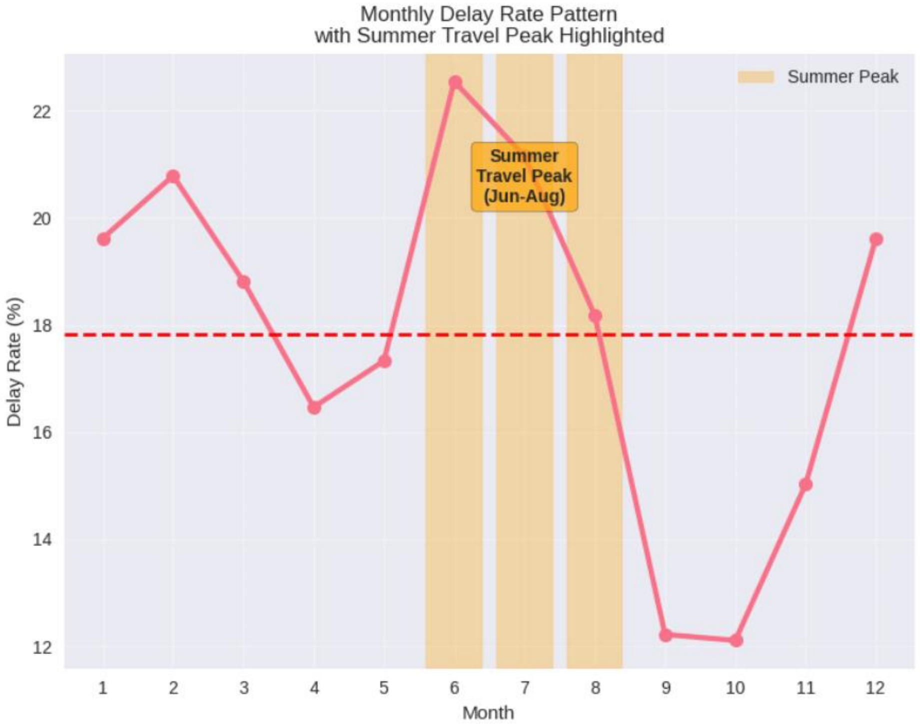
Delay Distribution by Severity Category
(Violin Plot - Final Journey)
3hr Compensation
4hr Critical
Delay Duration (minutes)
350
300
250
200
150
2-3h
3-4h
4-5h
5-6h

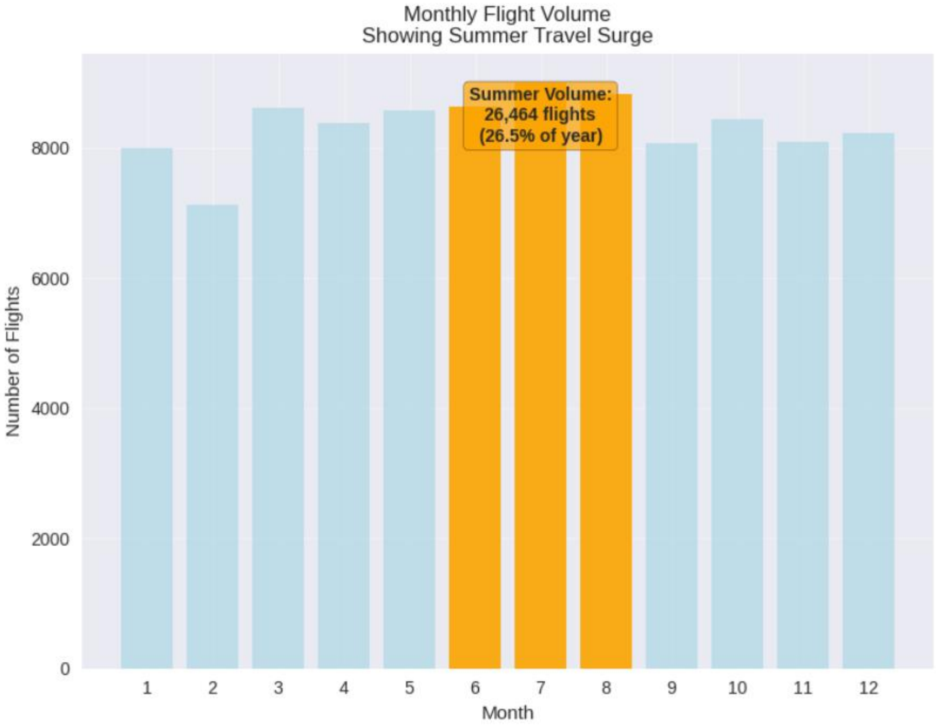

Distribution of Severe Delays (>2 Hours)
3hr Compensation Threshold
4hr Critical Threshold
Mean: 193 min
Median: 165 min
Number of Flights
Delay Duration (minutes)
Delay Count vs Total Cost by Severity
1185
$32.8M
$16.4M
$14.6M
435
199
$7.1M
81
$10.7M
102
Number of Delays
Total Cost ($ Millions)
2-3h
3-4h
4-5h
5-6h
6h+
Delay Severity Category


Actual vs Predicted Delays
(R² = 0.029)
Perfect Prediction
Predicted Delay (minutes)
Actual Delay (minutes)
200
150
100
50
0
−50
−50
0
50
100
150
200
Residuals vs Predicted Values
Residuals (minutes)
Predicted Delay (minutes)
100
80
60
40
20
0
−20
−40
−5
0
5
10
15
20
25



Feature Importance in Linear Regression
Weekend (vs Weekday)
Origin Airport
Month
Airline
Flight Distance
Peak Hour
2-Hour Time Block
Departure Hour
Increases Delay
Decreases Delay
0.0
0.5
1.0
1.5
2.0
2.5
Absolute Coefficient Value
Model Accuracy by Delay Category
Mean Absolute Error (minutes)
0
25
50
75
100
125
150
175
n=12558
n=3914
n=2420
n=726
n=382
Early
On-Time
Minor Delay
Major Delay
Severe Delay





Delay rates peak in June–August, with July exceeding 22%, well above the 18% annual average. September and October record the lowest delays, well below this average.

Flight volumes are highest in summer months, totaling 26,464 flights (26.5% of the year's total) from June to August. This seasonal increase aligns with peak travel demand, which likely contributes to the elevated delay rates seen in the same period.

Airline Performance Comparison

rline Performance Analysis

s dual-axis visualization reveals the relationship between operational volume and delay performance. Notably, higher volume airlines don't necessarily correlate with higher ay rates, suggesting that operational efficiency is achievable at scale with proper systems and processes.

Cost per Delay Minute

$74.24+

Average airline cost per delay minuteAverage airline cost per delay minute

Aggregate Industry Cost

$47.10+

passenger delay cost at $47.10 per hour, compounding over millions of flights.

Passenger Impacted

850M

passengers/year imoacted

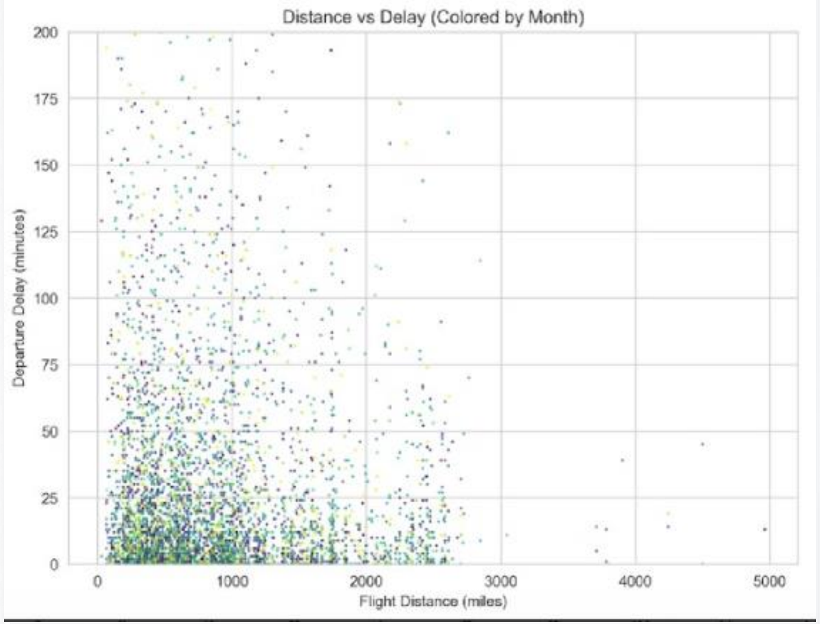
Passenger Burden

$6.8B

For 850M passengers/year, average delay delay cost is $6.8B annually in lost time, time, expenses, and disruptions.

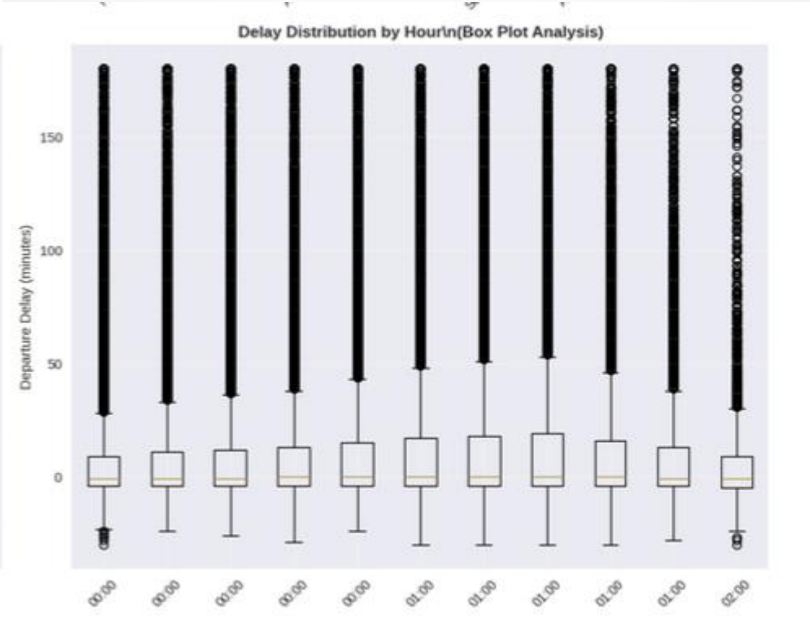
Visualization during *Processing & Preparation – sheet 2*

Distance vs Delay (Colored by Month)

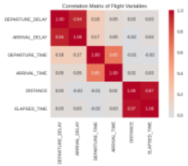


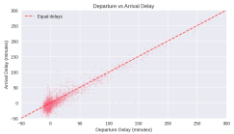
Description:

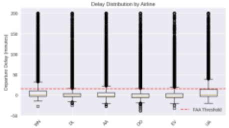
Delay Distribution by Hour (Box Plot Analysis) Analysis)

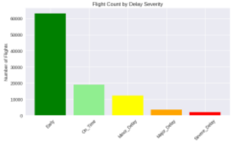


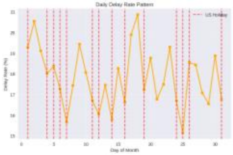
Description:


Correlation Matrix of Flight Variables
DEPARTURE_DELAY
ARRIVAL_DELAY
DEPARTURE_TIME
ARRIVAL_TIME
DISTANCE
ELAPSED_TIME


Departure vs Arrival Delay
Equal delays
Arrival Delay (minutes)
Departure Delay (minutes)


Delay Distribution by Airline
Departure Delay (minutes)
FAA Threshold
WN
DL
AA
OO
EV
UA


Flight Count by Delay Severity
Number of Flights


Daily Delay Rate Pattern
US Holiday
Delay Rate (%)
Day of Month