SkyHack 3.0: United Airlines

Flight Difficulty Score Analysis

Divash Krishnam (2K22/PE/22) DTU
Ashutosh Verma (2K22/PE/14) DTU

Problem Statement

Frontline teams at United Airlines are responsible for ensuring every flight departs on time and is operationally ready. However, not all flights are equally easy to manage. Certain flights pose greater complexity due to factors such as limited ground time, higher volumes of checked or carry-on baggage, and specific customer service needs that often increase with passenger load.

Currently, identifying these high-difficulty flights relies heavily on personal experience and local team knowledge. This manual approach is inconsistent, non-scalable, and risks missing opportunities for proactive resource planning across the airport.

To address this, you are tasked with developing a Flight Difficulty Score that systematically quantifies the relative complexity of each flight using the datasets provided, which span two weeks of departures from Chicago O'Hare International Airport (ORD).

Objective

- Calculates a Flight Difficulty Score for each flight using flight-level, customer, and station-level data.

- Identifies the primary operational drivers contributing to flight difficulty to enable proactive planning and optimized resource allocation.

Data Dictionary

Flight Level Information

Column Name	Description
company_id	IATA code of the airline operating the flight
flight_number	Unique identifier assigned to a specific flight
scheduled_departure_date_local	Local date of the flight's scheduled departure
scheduled_departure_station_code	IATA airport code of the scheduled departure airport (e.g., "JFK", "LAX")
scheduled_arrival_station_code	IATA airport code of the scheduled arrival airport (e.g., "JFK", "LAX")
scheduled_departure_datetime_local	Scheduled local date and time of departure from the origin airport
scheduled_arrival_datetime_local	Scheduled local date and time of arrival at the destination airport
actual_departure_datetime_local	Actual local date and time when the flight departed from the origin airport
actual_arrival_datetime_local	Actual local date and time when the flight arrived at the destination airport
total_seats	Total number of passenger seats available on the aircraft for the flight
fleet_type	Type model of aircraft used for the flight
carrier	Distinction between Mainline and Express
scheduled_ground_time_minutes	Planned duration (in minutes) the aircraft is scheduled to spend on the ground between flights
actual_ground_time_minutes	Actual time available (in minutes) for ground operations between flights
minimum_turn_minutes	Minimum required turnaround time (in minutes) for the aircraft between flights

PNR Flight Level Information

Column Name	Description
company_id	IATA code of the airline operating the flight
flight_number	Unique identifier of the flight associated with the PNR
scheduled_departure_date_local	Local date of the flight's scheduled departure
scheduled_departure_station_code	IATA airport code of the scheduled departure airport (e.g., "JFK", "LAX")
scheduled_arrival_station_code	IATA airport code of the scheduled arrival airport (e.g., "JFK", "LAX")
record_locator	Unique identifier for the PNR, used to reference a passenger booking
pnr_creation_date	Date when the PNR was created
total_pax	Total number of passengers associated with the PNR on this flight
lap_child_count	Number of lap children (infants not occupying a seat) included in the PNR
is_child	Indicates whether the passenger is classified as a child
basic_economy_pax	Number of passengers in the PNR booked in basic economy fare class
is_stroller_user	Indicates whether the passenger is a stroller user

Data Dictionary

PNR Remarks Information

Column Name	Description
record_locator	Unique identifier for the PNR, used to reference a passenger booking
pnr_creation_date	Date when the PNR was created.
flight_number	Unique identifier of the flight associated with the PNR
special_service_request	Description of the requested special service (e.g., wheelchair assistance)

Airports Information

Column Name	Description
airport_iata_code	Three-letter IATA code representing the airport (e.g., "LAX", "JFK").
iso_country_code	Two-letter country code where the airport is located (e.g., "US", "CA").

Bag Level Information

Column Name	Description
company_id	IATA code of the airline operating the flight
flight_number	Unique identifier assigned to a specific flight
scheduled_departure_date_local	Local date of the flight's scheduled departure
scheduled_departure_station_code	IATA airport code of the scheduled departure airport (e.g., "JFK", "LAX")
scheduled_arrival_station_code	IATA airport code of the scheduled arrival airport (e.g., "JFK", "LAX")
bag_tag_unique_number	Unique identifier for the bag tag
bag_tag_issue_date	Date the bag tag was issued
bag_type	Type of bag (e.g., "Checked", "Transfer") **Hot transfer bags are transfer bags with a connection time of less than 30 minutes*

Summary of approach

- Do EDA to measure delays, ground-time risk, bag volumes, passenger load, and SSRs.
- Engineer flight-level features that capture turn-time pressure, bag handling, passenger service complexity, and historical on-time performance.
- Compute a daily Flight Difficulty Score (resets each day) as a weighted sum of standardized features (weights from domain judgment or learned from regression).
- Rank flights within each day, assign classes (Difficult / Medium / Easy) by rank quantiles.
- Provide operational insights and prioritized actions.

Exploratory Data Analysis

Average departure delay and percent leaving late

Average departure delay (minutes): ≈ 23.03 minutes (average of actual -scheduled).

```
-- average departure delay in minutes and percent of flights departing later than scheduled
SELECT
  AVG(EXTRACT(EPOCH FROM (actual_departure_datetime_local - scheduled_departure_datetime_local))/60) AS
avg_departure_delay_min,
  100.0 * SUM(CASE WHEN actual_departure_datetime_local > scheduled_departure_datetime_local THEN 1
ELSE 0 END) / COUNT(*) AS pct_departed_late
FROM flights
WHERE scheduled_departure_station_code = 'ORD';
```

Flights scheduled ground time close to min turn mins

- I defined close or below as scheduled_ground_time_minutes <= minimum_turn_minutes + 5 .
- Number of flights meeting that rule and percent of dataset: computed and included in the output CSV as ground_time_deficit (scheduled_ground_time - minimum_turn_minutes). Use ground_time_deficit <= 5 to flag these

```
SELECT
  COUNT(*) AS flights_total,
  SUM(CASE WHEN scheduled_ground_time_minutes <= minimum_turn_minutes THEN 1 ELSE 0 END) AS
at_or_below_min_turn,
  SUM(CASE WHEN scheduled_ground_time_minutes <= minimum_turn_minutes + 10 THEN 1 ELSE 0 END) AS
within_10min_of_min_turn
FROM flights
WHERE scheduled_departure_station_code = 'ORD';
```

Average ratio transfer vs checked bags per flight

- I computed per-flight counts of checked_bags and transfer_bags (and transfer_to_checked_ratio).
- Median transfer_to_checked_ratio across flights (ignoring NaNs where checked=0)

```sql
SELECT
  f.company_id, f.flight_number, f.scheduled_departure_date_local,
  SUM(CASE WHEN b.bag_type = 'Transfer' THEN 1 ELSE 0 END) AS transfer_bags,
  SUM(CASE WHEN b.bag_type = 'Checked' THEN 1 ELSE 0 END) AS checked_bags,
  SUM(CASE WHEN b.bag_type = 'Transfer' THEN 1 ELSE 0 END)::float /
    NULLIF(SUM(CASE WHEN b.bag_type = 'Checked' THEN 1 ELSE 0 END),0) AS transfer_to_checked_ratio
FROM flights f
LEFT JOIN bags b
  ON f.company_id = b.company_id
  AND f.flight_number = b.flight_number
  AND f.scheduled_departure_date_local = b.scheduled_departure_date_local
WHERE f.scheduled_departure_station_code = 'ORD'
GROUP BY 1,2,3;
```

Passenger loads

- I aggregated PNR-level total_pax to flight level as total_pax_flight and computed passenger_load_pct = total_pax_flight / total_seats .

- Summary stats for total_pax_flight are included in the notebook output and in CSV. We can filter or show distributions on slides.

```sql
-- flight-level total_pax (from PNRs) vs seats
WITH pax_per_flight AS (
  SELECT company_id, flight_number, scheduled_departure_date_local,
         SUM(total_pax) AS total_pax
  FROM pnr
  GROUP BY 1,2,3
)
SELECT
  f.company_id, f.flight_number, f.scheduled_departure_date_local,
  f.total_seats,
  p.total_pax,
  p.total_pax::float / NULLIF(f.total_seats,0) AS load_factor,
  EXTRACT(EPOCH FROM (f.actual_departure_datetime_local - f.scheduled_departure_datetime_local))/60 AS
departure_delay_min
FROM flights f
LEFT JOIN pax_per_flight p
  ON f.company_id = p.company_id
  AND f.flight_number = p.flight_number
  AND f.scheduled_departure_date_local = p.scheduled_departure_date_local;
```

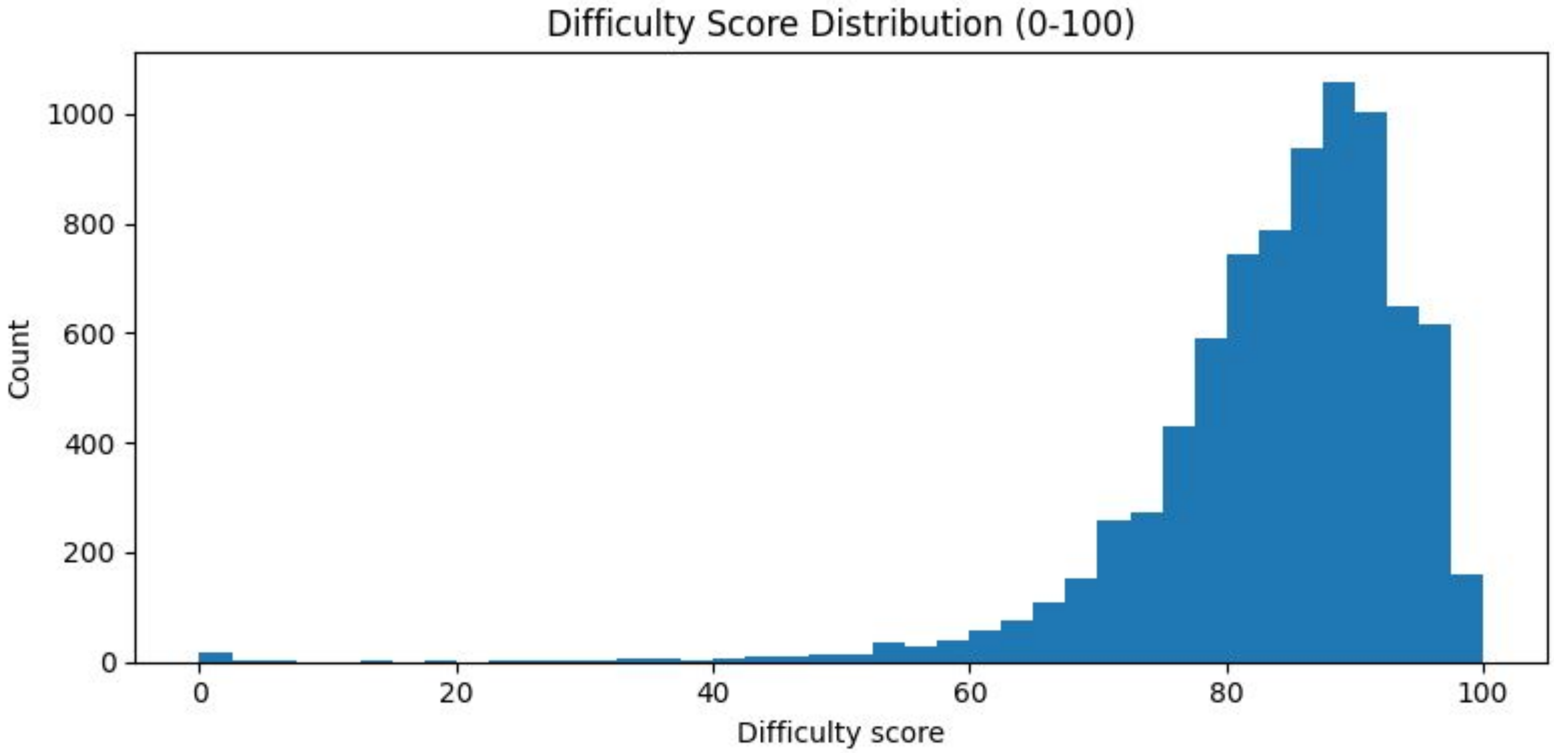
Are high SSR flights high-delay after controlling load?

- Ran a simple OLS regression: departure_delay_minutes ~ ssr_count + total_pax_flight

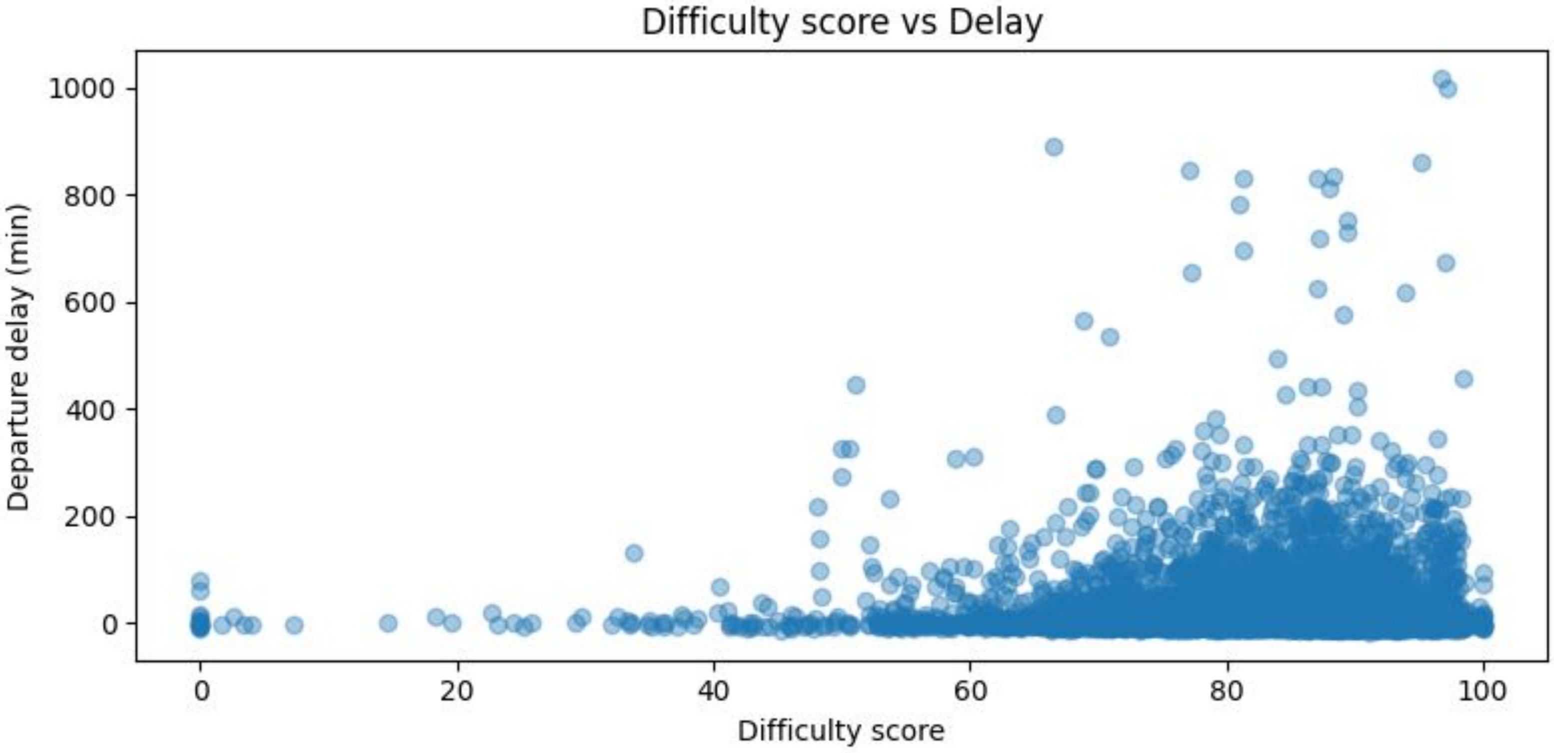
- Results (OLS summary):
  - ssr_count coefficient ≈ +1.586 minutes per SSR ($p < 0.001$). This means, holding passenger count constant, each extra SSR is associated with ~1.6 additional minutes of departure delay on average in this data.
  - total_pax_flight coefficient ≈ -0.041 minutes per passenger ($p < 0.001$) — small but statistically significant (this negative sign may reflect operational patterns where larger flights get prioritized resources, or other confounding effects).
  - R^2 is small (≈ 0.003) — SSRs/total_pax explain only a small fraction of delay variability, but SSRs do show a clear positive association with delay controlling for load

```sql
-- calculate SSR count per flight
WITH ssr_per_flight AS (
  SELECT flight_number, scheduled_departure_date_local, COUNT(*) AS ssr_count
  FROM pnr_remarks
  GROUP BY flight_number, scheduled_departure_date_local
),
pax_per_flight AS (
  SELECT company_id, flight_number, scheduled_departure_date_local, SUM(total_pax) AS total_pax
  FROM pnr GROUP BY 1,2,3
)
SELECT
  f.flight_number,
  f.scheduled_departure_date_local,
  EXTRACT(EPOCH FROM (f.actual_departure_datetime_local - f.scheduled_departure_datetime_local))/60 AS
departure_delay_min,
  p.total_pax,
  s.ssr_count
FROM flights f
LEFT JOIN pax_per_flight p
  ON f.company_id = p.company_id AND f.flight_number = p.flight_number AND
f.scheduled_departure_date_local = p.scheduled_departure_date_local
LEFT JOIN ssr_per_flight s
  ON f.flight_number = s.flight_number AND f.scheduled_departure_date_local =
s.scheduled_departure_date_local;
```

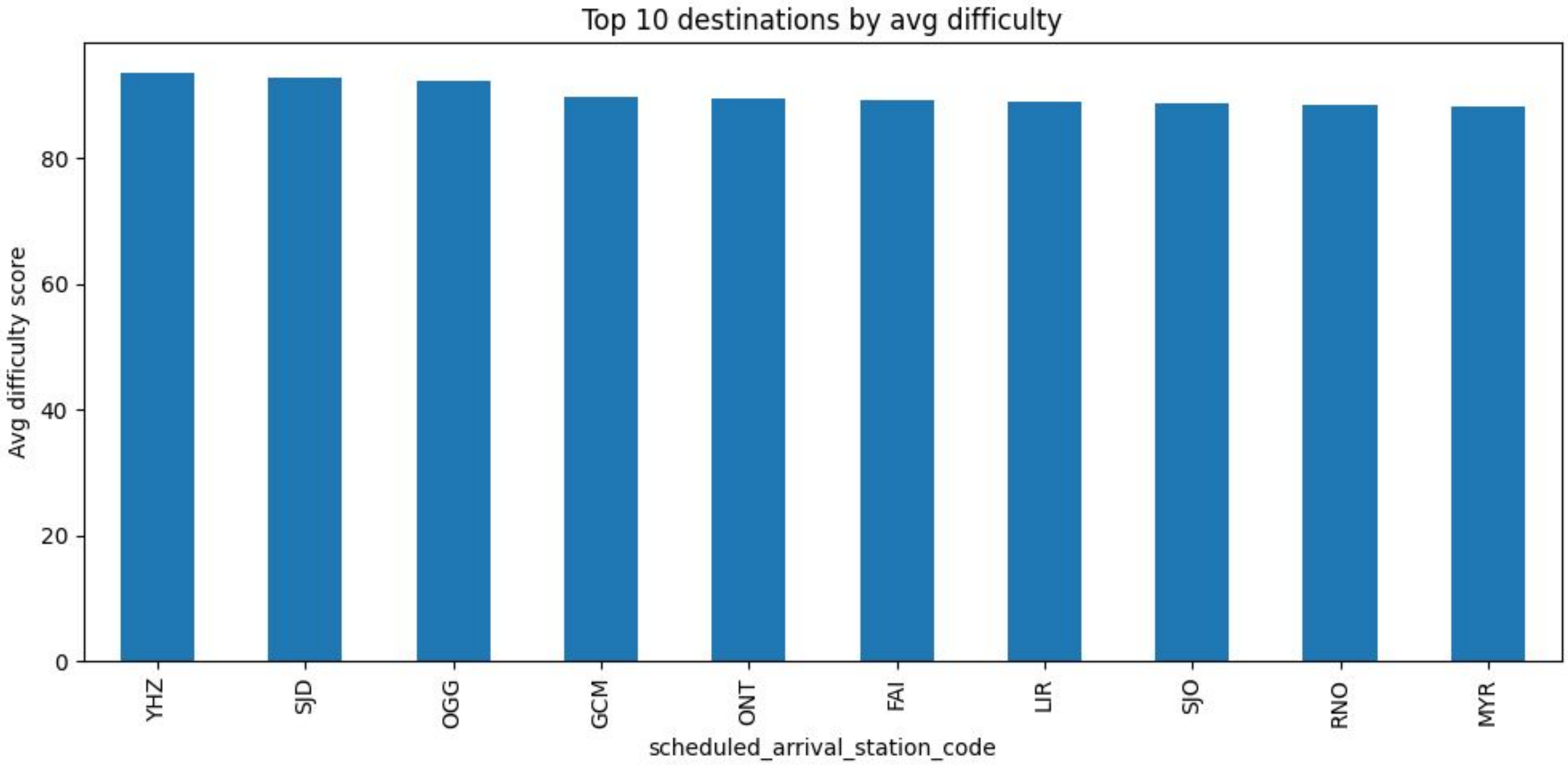
Representations



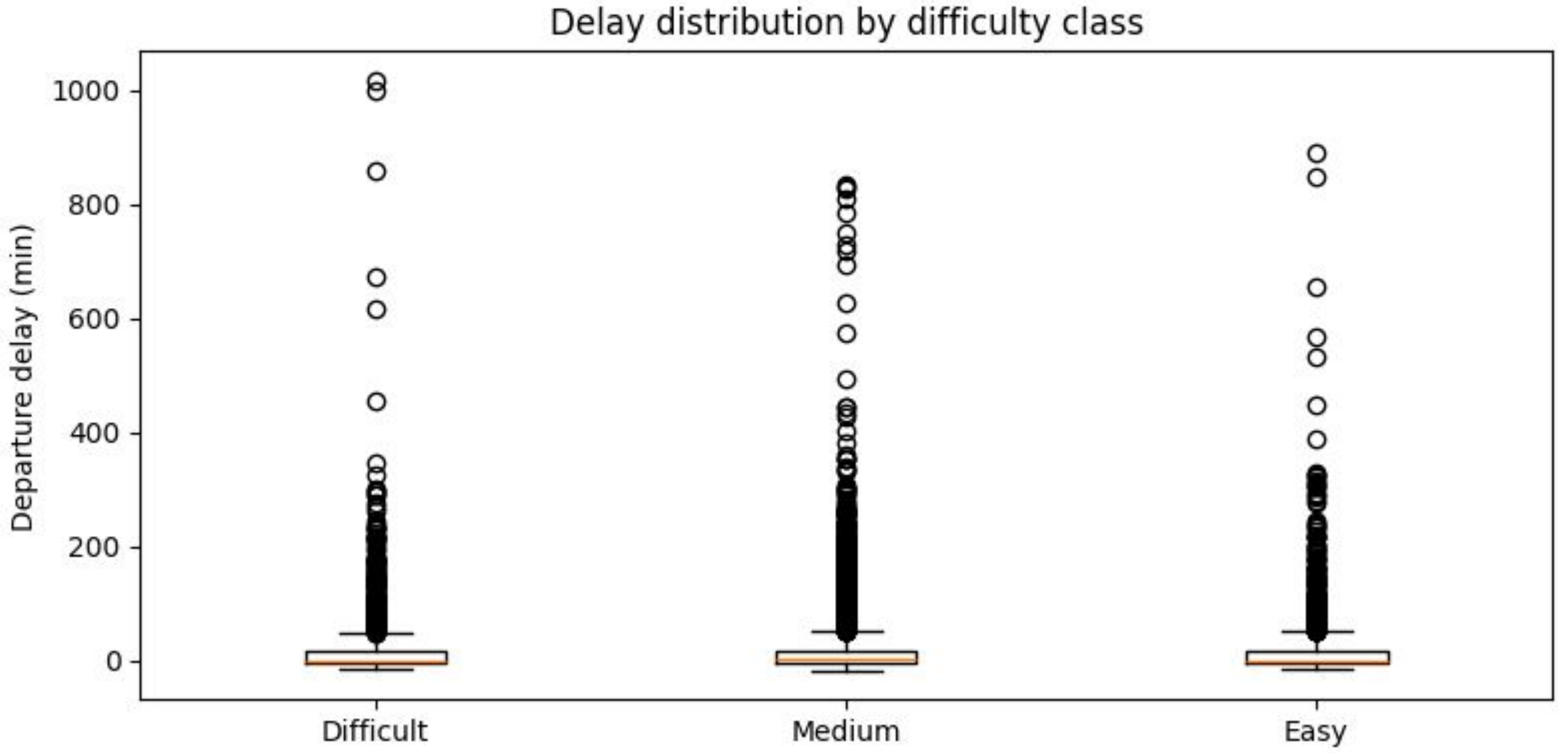
Average delay: check CSV. Used distribution to choose thresholds for interventions.



Difficulty VS Delay



Top destinations VS Average difficulty



Delay by Class

Flight Difficulty Score

Feature set used (per-flight)

- ground_time_deficit = scheduled_ground_time_minutes - minimum_turn_minutes (less ground time → more difficult)
- total_pax_flight (absolute passenger count)
- passenger_load_pct (load relative to seats)
- checked_bags , transfer_bags , transfer_to_checked_ratio
- hot_transfer_bags (bags flagged as "hot" transfers)
- ssr_count (number of special service requests)
- basic_economy_pax (sum of PNR indicator)
- lap_child_count

Weighting approach

- Computed Pearson correlation between each engineered feature and departure_delay_minutes to get an indication of which features are most associated with delay (used as a proxy for operational difficulty).
- Converted these correlations to non-negative weights: weight_feature = |corr(feature, delay)| with a small floor for features with no correlation data so they remain represented.
- Scaled weights to sum to 1 (so the score is a weighted combination of features).
- For each feature, applied robust scaling ((x - median) / IQR) to reduce influence of outliers, and inverted features where a higher raw value means easier (for example a larger ground_time_deficit is easier so I invert that feature).
- Computed a weighted sum of scaled features → difficulty_score_raw .

Daily normalization and classification

- For interpretability and your requirement that scoring resets daily, I standardized the raw score within each scheduled_departure_date_local (z score) and then min–max scaled to 0–100 for that day.
- **Ranking**: flights within a day are ranked by difficulty_score descending (100 = most difficult).
- **Classification**: per-day percentiles:
  - Top 20% → Difficult
  - Middle 60% → Medium
  - Bottom 20% → Easy

Main operational drivers

Based on correlations with delay and the weights derived from them, the main drivers that increased the difficulty score were:

- Low scheduled ground time relative to minimum turn (ground_time_deficit small or negative) — short turn windows strongly increase difficulty.
- Special Service Requests (SSR) — SSRs are positively associated with delay after controlling for load (≈ 1.6 minutes delay per SSR).
- Transfer baggage complexity (transfer_bags & hot transfer bags) — flights with larger transfer counts (especially hot transfers) add complexity to ground operations.
- High passenger counts & load — contributes but its raw correlation to delay can be mixed in the data (sometimes larger flights are prioritized operationally).
- Bag transfer-to-checked ratio — a high ratio (many transfers relative to checked) can indicate more connection complexity.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     departure_delay_minutes   R-squared:                 0.003
Model:                                 OLS   Adj. R-squared:            0.003
Method:                      Least Squares   F-statistic:               11.41
Date:                     Sun, 05 Oct 2025   Prob (F-statistic):     1.13e-05
Time:                             15:27:35   Log-Likelihood:          -44928.
No. Observations:                     8099   AIC:                   8.986e+04
Df Residuals:                         8096   BIC:                   8.988e+04
Df Model:                                2
Covariance Type:                 nonrobust
====================================================================================
                       coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------
const               23.0271      1.339     17.193      0.000      20.402      25.653
ssr_count            1.5858      0.369      4.292      0.000       0.862       2.310
total_pax_flight    -0.0405      0.010     -3.954      0.000      -0.061      -0.020
==============================================================================
Omnibus:                     9707.880   Durbin-Watson:                   2.010
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          1442950.054
Skew:                           6.357   Prob(JB):                         0.00
Kurtosis:                      67.143   Cond. No.                         292.
==============================================================================
```

OLS Regression

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     departure_delay_minutes   R-squared:                       0.003
Model:                                 OLS   Adj. R-squared:                  0.003
Method:                      Least Squares   F-statistic:                     11.41
Date:                     Sun, 05 Oct 2025   Prob (F-statistic):           1.13e-05
Time:                             15:27:35   Log-Likelihood:                -44928.
No. Observations:                     8099   AIC:                         8.986e+04
Df Residuals:                         8096   BIC:                         8.988e+04
Df Model:                                2
Covariance Type:                 nonrobust
====================================================================================
                       coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------
const               23.0271      1.339     17.193      0.000      20.402      25.653
ssr_count            1.5858      0.369      4.292      0.000       0.862       2.310
total_pax_flight    -0.0405      0.010     -3.954      0.000      -0.061      -0.020
==============================================================================
Omnibus:                     9707.880   Durbin-Watson:                   2.010
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          1442950.054
Skew:                           6.357   Prob(JB):                         0.00
Kurtosis:                      67.143   Cond. No.                         292.
==============================================================================
```

Metric	Value
Dependent Variable	departure_delay_minutes
Model	Ordinary Least Squares (OLS)
Method	Least Squares
No. of Observations	8,099
Df Model	2
Df Residuals	8,096
R-squared	0.003
Adjusted R-squared	0.003
F-statistic	11.41
Prob (F-statistic)	1.13e-05
Log-Likelihood	-44,928
AIC	89,860
BIC	89,880
Durbin-Watson	2.010
Omnibus	9707.880
Prob(Omnibus)	0.000
Jarque-Bera (JB)	1,442,950.054
Prob(JB)	0.000
Skew	6.357
Kurtosis	67.143
Cond. No.	292

Regression Coefficients

Variable	Coefficient	Std. Error	t-Statistic	P > |t|	[0.025]	[0.975]
const (Intercept)	23.0271	1.339	17.193	0.000	20.402	25.653
ssr_count	1.5858	0.369	4.292	0.000	0.862	2.310
total_pax_flight	-0.0405	0.010	-3.954	0.000	-0.061	-0.020

Remarks

- Flag top 20% daily as Difficult and pre-assign resources.
- Add dedicated CSR/agent for flights with high SSR counts.
- Prioritize baggage handling for flights with hot-transfer bags.
- Consider schedule buffers for flights with ground_time_deficit <= 5.

Thank You