Capstone Project

Analysis of outbound flights from Anchorage airport for 2023 & 2024

The source code is accessible from GitHub.

Executive summary

This capstone project analyzed outbound flights from Anchorage International Airport (ANC) for 2023–2024 to understand causes of flight delays, cancellations, and diversions, and to assess whether these could be predicted using data analytics and machine learning.

Methods

Flight data from the U.S. Bureau of Transportation Statistics was combined with Meteostat weather data and airport geolocation data from OurAirports. After cleaning and merging, 23 relevant features were retained. A Random Forest model was trained to predict three outcomes—diverted, cancelled, and critically delayed flights—using SMOTE to rebalance the highly imbalanced dataset.

Key Findings

- Airline operations, not weather, were suspected to be the leading cause of major delays—especially late aircraft and carrier-related issues.
- Alaska Airlines had the lowest proportion of critical delays, while American Airlines showed higher delays and diversions.
- Seasonal and event-driven patterns were evident, with spikes during summer travel and severe weather events like snowstorms and volcanic ash.
- The machine learning model performed poorly after correcting for data leakage, indicating insufficient and imbalanced data for reliable prediction.

Conclusion and Recommendations

While data visualization revealed meaningful operational insights, machine learning predictions were limited by data scope and imbalance. Future work should:

- Expand datasets to include multi-year and inbound flight data.
- Incorporate richer operational and flight-path weather variables.
- Test alternative ML models (e.g., XGBoost, LightGBM).
- Overall, the project demonstrates the potential of data-driven aviation analytics while highlighting the need for more comprehensive data to achieve accurate delay prediction.

Table of Contents

Introduction

For this data analytics project, the goal was to explore and visualize flight delay information for a specific airport of departure. Anchorage, Alaska was chosen as a suitable option—it is a relatively small airport (potentially making it more interesting than a large hub, though this is subjective) and is located in a climate zone with significant weather variability, including frequent storms.

The original intent was to analyze departures from Edinburgh, UK; however, detailed public flight-level data was not available. Therefore, Anchorage was selected as a practical alternative.

The U.S. Bureau of Transportation Statistics provides publicly available flight records dating back to 1987, containing up to 109 different parameters per flight. This extensive dataset offers valuable opportunities for analysis that can benefit airlines, airports, and passengers when used effectively.

A secondary point of interest for this project was determining whether airlines have sufficient data (weather, primarily) in advance to inform passengers about major upcoming delays, cancellations, or diversions. The latter outcome, in particular, presents an interesting target variable for testing machine learning prediction models.

Methods

Data Sources

The first data source consisted of historic flight data, downloaded from the U.S. Bureau of Transportation Statistics for Alaska (filtered by geography) for 2023 and 2024, resulting in a total of 24 monthly files. Out of the 109 available columns, 23 were retained—those that appeared most relevant to flight delays and suitable for data visualization.

The second data source was Meteostat weather data. Initially, an attempt was made to extract the data using the Meteostat API; however, after multiple connection issues related to certificates and SSH, the Python library was used instead. The weather variables of interest included air temperature, relative humidity, total precipitation, snow depth, wind direction, average wind speed, wind peak gust, and sea-level air pressure, for both origin and destination airports. Snow depth and wind peak gust were later dropped due to unavailability for Anchorage.

During data preparation, it became apparent that there was no direct link between the two datasets. To bridge this, airport location data from OurAirports.com was incorporated. Based on IATA airport codes, latitude and longitude coordinates were added, allowing Meteostat weather data to be retrieved for specific locations, dates, and times.

Data Cleaning & Preparation

In addition to the data parameters described in the previous section, several data preparation steps were performed:

- Carrier names were replaced with their full descriptions instead of two-letter codes.
- Cancellation codes were expanded to include descriptive text.
- Delay values (carrier, weather, NAS, security, and late aircraft) were filled with zeros where missing.
- A total delay column was added, representing the sum of all individual delay components.
- A critical delay column was introduced, defined as any delay exceeding 60 minutes, or when a flight was cancelled or diverted.
- Flight dates were converted to datetime format.

For most visualizations, additional intermediate preparation steps such as grouping, averaging, and counting were performed. Full details of these transformations can be found in the data preparation section of the source code. Numerous plots were initially generated; however, those with limited analytical value or discussion potential were later excluded.

Merging the various data sources required several format conversions, primarily to align date and time fields. To improve efficiency, departure–destination pairs with associated datetimes were precomputed, reducing the number of Meteostat weather lookups required. The process was also parallelized across multiple cores for faster execution, and the resulting dataset was saved locally.

For the Random Forest prediction component, missing numerical values were replaced with the median of the respective column, while missing categorical values were replaced with the mode.

Random Forest Predictions

A Random Forest model was selected for this analysis, as it is well-suited to the characteristics of the dataset. It offers several advantages:

- Handles nonlinear relationships effectively
- Performs well with correlated features
- Tolerates a moderate level of noise and missing data
- Can work with imbalanced datasets and mixed data types
- Provides strong baseline performance with minimal tuning

The target variables for prediction were defined as:

- Diverted: yes/no
- Cancelled: yes/no
- Critical delay: yes/no

Because the dataset was highly imbalanced, it required rebalancing. The SMOTE (Synthetic Minority Oversampling Technique) method was used; however, in hindsight, this may not have been ideal due to the very small proportion of diversions, cancellations, and critical delays in the data.

Ideally, multiple models and rebalancing methods would have been tested, but this was not feasible within the project's time constraints. The machine learning component was included primarily for skill development and to provide a benchmark for making some conclusions.

Cross-validation was used to tune model parameters. A total of 20 iterations were performed (a practical compromise between thoroughness and efficiency), with scoring based on the F1 metric, which is most appropriate for imbalanced data. Cross-validation was set to 3 folds due to the large dataset size.

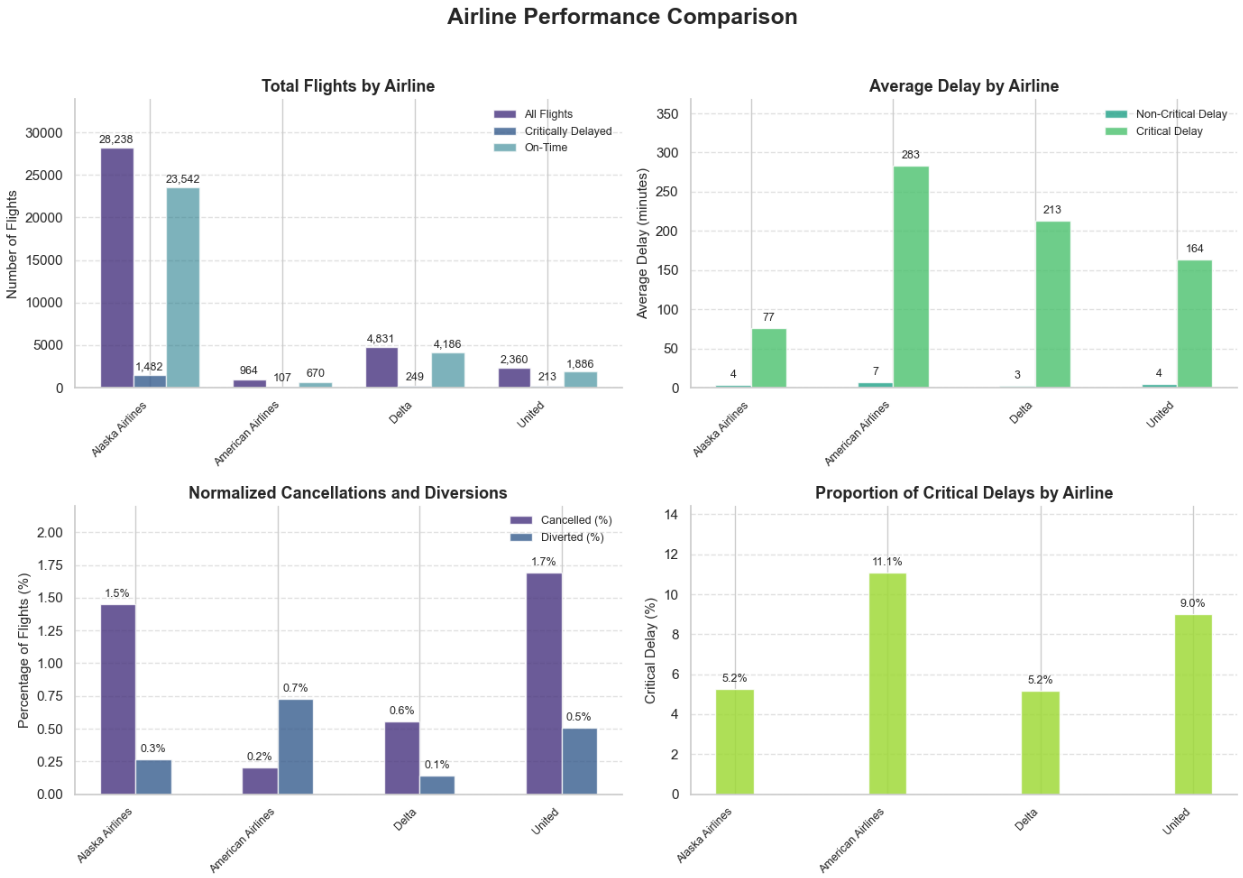
The hyperparameters tested included:

- n_estimators: [100, 150, 200, 300]
- max_depth: [10, 15, 20, 25, None]
- min_samples_split: [2, 4, 6, 8]
- min_samples_leaf: [1, 2, 3, 4]
- max_features: ['sqrt', 'log2', None]
- bootstrap: [True, False]

Results

Data Analysis

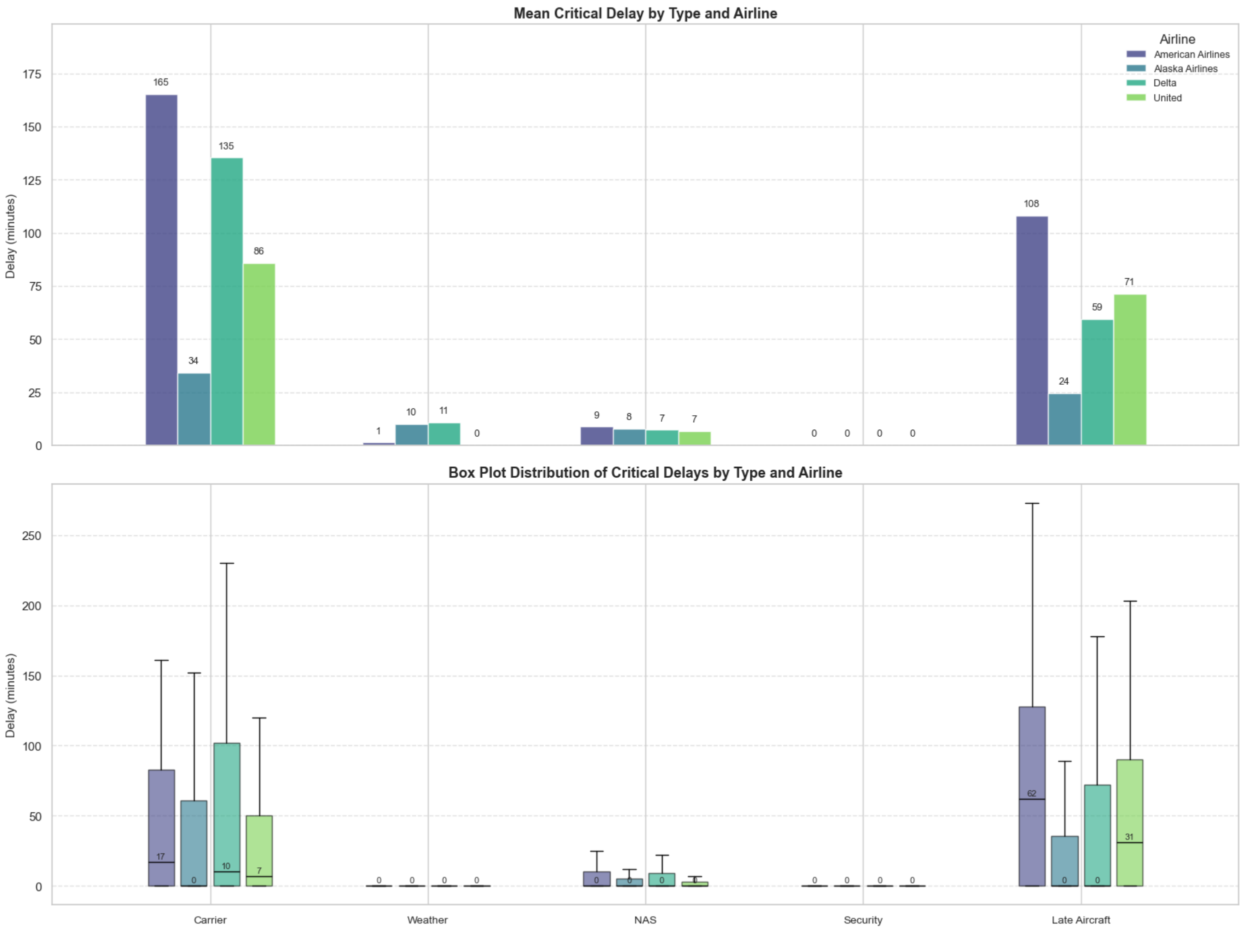
Airline Performance Comparison



From these observations, we can see that Alaska Airlines leads in total flight numbers by a wide margin. It also has the lowest number of critical delays, suggesting stronger operational performance. In contrast, American Airlines shows the opposite trend—possibly indicating that Alaska Airlines is better adapted to local weather conditions and/or has more efficient logistics in Anchorage.

However, when cancellations and delays are examined separately, Alaska Airlines no longer leads. Interestingly, American Airlines is the only carrier with a higher percentage of diversions than cancellations, which may reflect a different corporate approach—choosing to proceed with flights even when a diversion is likely. Overall, American Airlines has the highest proportion of critical delays, an issue that warrants further attention from management.

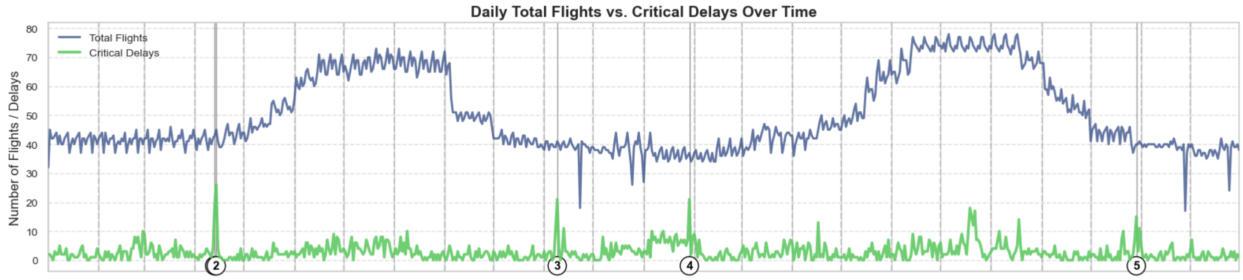
Critical Delay distributions

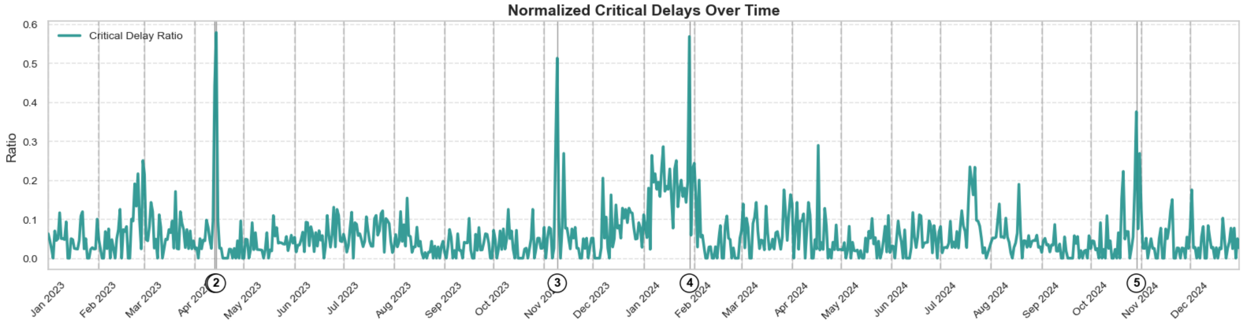


Looking further into the distribution, it is evident that the primary causes of delays, by a large margin, are those attributable to the airline itself—specifically, carrier delays and late aircraft. Delays attributed to the National Air System are relatively minor, and interestingly, security-related delays are nearly zero, which is a commendable achievement for Anchorage airport staff.

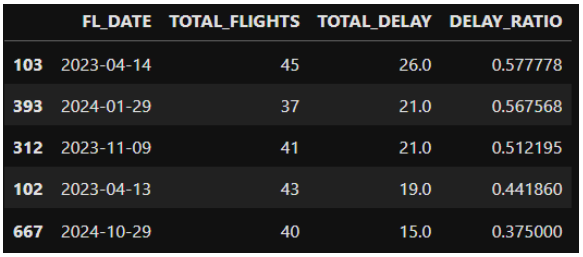
It was initially expected that weather-related delays would represent a larger proportion; however, they are lower than anticipated. This will be examined further in the analysis and incorporated into predictive modeling efforts. Once again, American Airlines leads in delay patterns, while Alaska Airlines performs best, suggesting that Alaska may have better-aligned logistics and operations at the local airport.

Delays over time


Daily Total Flights vs. Critical Delays Over Time


Normalized Critical Delays Over Time

From these plots, it is evident—as expected—that flights follow a clear seasonal pattern, with a higher number of travelers during the summer and noticeable dips in winter, particularly around Thanksgiving, Christmas, and New Year's Eve.



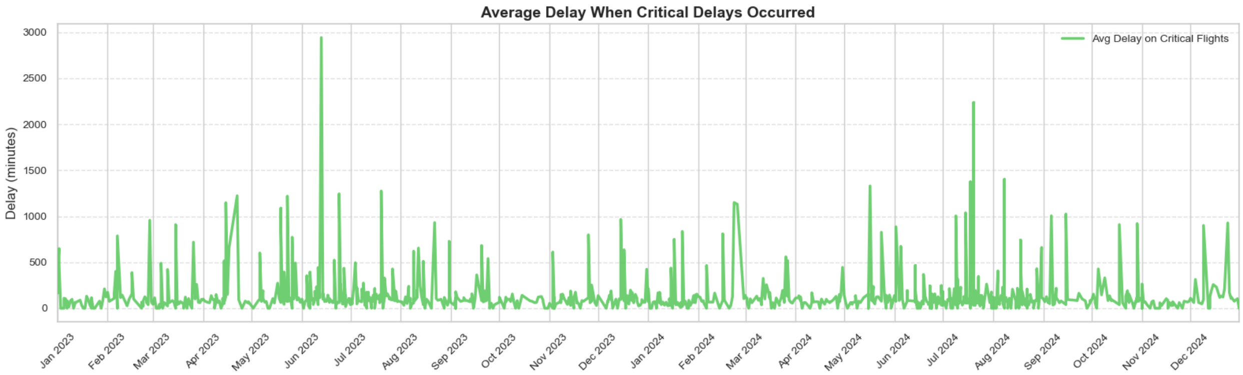
	FL_DATE	TOTAL_FLIGHTS	TOTAL_DELAY	DELAY_RATIO
103	2023-04-14	45	26.0	0.577778
393	2024-01-29	37	21.0	0.567568
312	2023-11-09	41	21.0	0.512195
102	2023-04-13	43	19.0	0.441860
667	2024-10-29	40	15.0	0.375000

When examining the normalized values, several distinct spikes appear. The top five examples were investigated, revealing the following related news events:

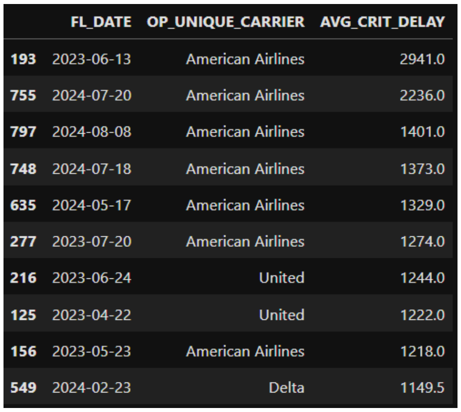
1. **2023-04-14** – "*Russia volcano disrupts Alaska flights for third day in a row*"
2. **2023-11-09** – "*Southcentral Alaska buried in more than a foot of snow from winter storm*"
3. **2024-01-29** – "*Alaska governor's annual speech to lawmakers delayed as high winds disrupt flights*"
4. **2024-10-29** – "*Season's first big snow fouls roads and shifts Anchorage schools to remote learning*"

It is therefore clear that weather-related events are the primary contributors to large-scale, single-day critical delays. However, some of these—such as volcanic eruptions—are inherently difficult to predict, unlike more common weather phenomena such as storms.

Average critical delay duration over time



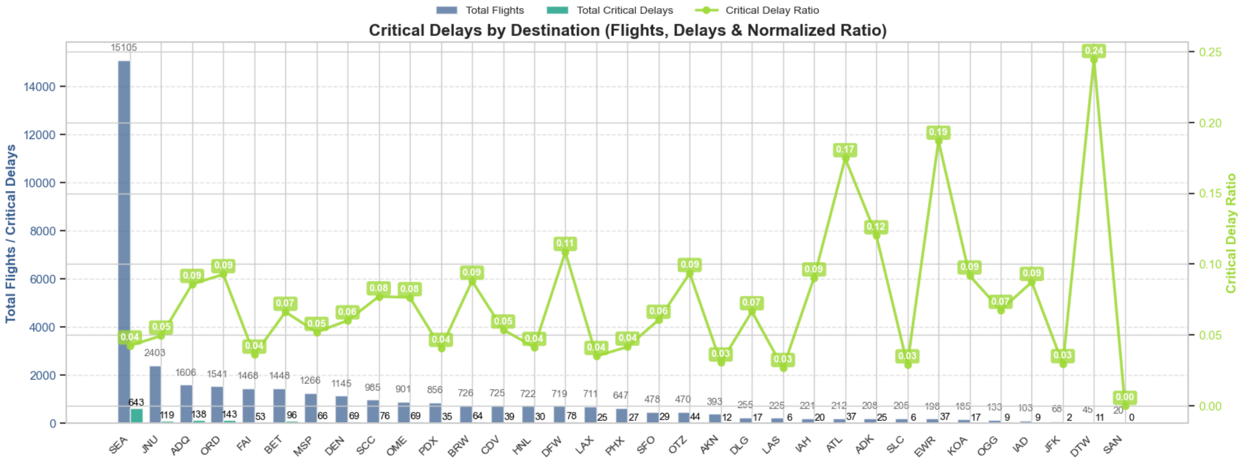
When examining average delay timelines for flights with critical delays, the spikes shift compared to the previous graph. This is primarily because cancellations and diversions do not have a recorded *TOTAL_DELAY* value, only a critical delay indicator.



	FL_DATE	OP_UNIQUE_CARRIER	AVG_CRIT_DELAY
193	2023-06-13	American Airlines	2941.0
755	2024-07-20	American Airlines	2236.0
797	2024-08-08	American Airlines	1401.0
748	2024-07-18	American Airlines	1373.0
635	2024-05-17	American Airlines	1329.0
277	2023-07-20	American Airlines	1274.0
216	2023-06-24	United	1244.0
125	2023-04-22	United	1222.0
156	2023-05-23	American Airlines	1218.0
549	2024-02-23	Delta	1149.5

A closer look at the top 10 delay cases reveals that American Airlines stands out as the clear outlier. Some of these cases are highly concerning — the longest recorded delay exceeds 49 hours without a cancellation. Such occurrences warrant serious review and procedural improvements. Combined with the fact that American Airlines operates the fewest outbound flights from Anchorage, these findings raise notable operational concerns.

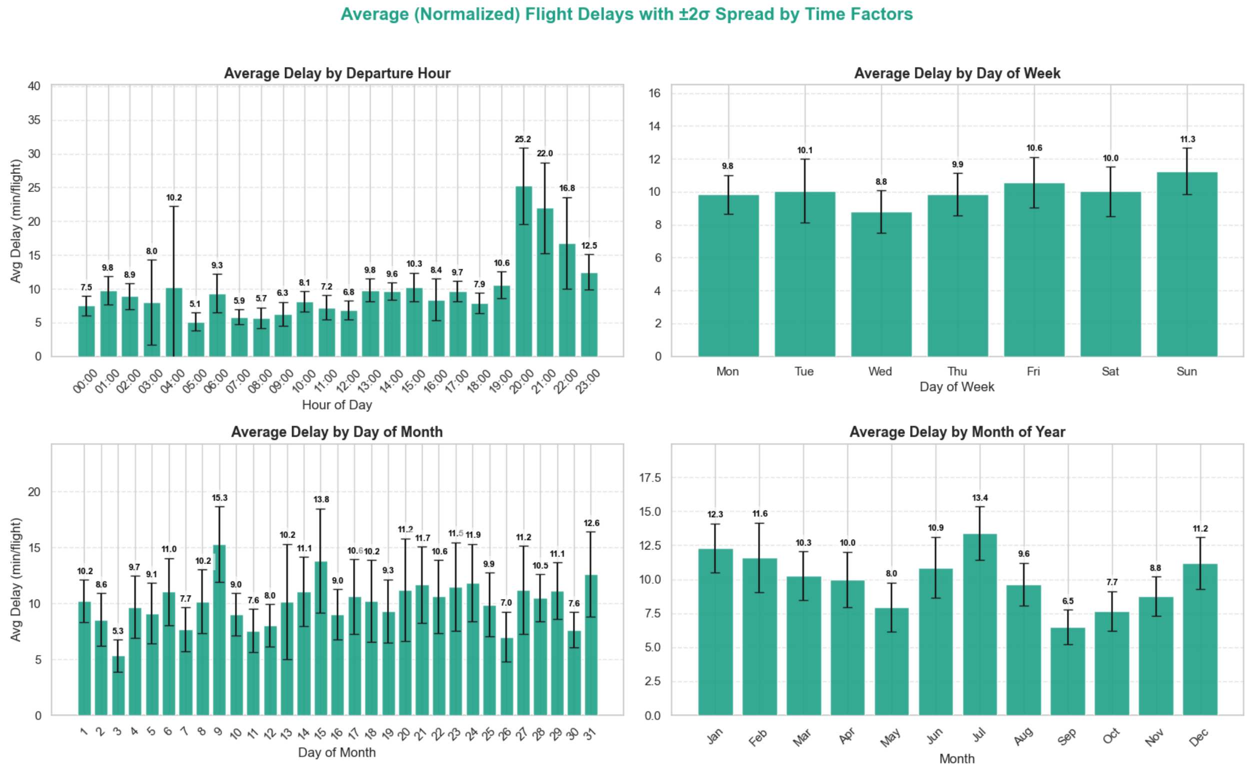
Critical delays by destination



Some destination airports show significant spikes in critical delay ratios — notably EWR (Newark Liberty), ATL (Hartsfield–Jackson Atlanta), and DTW (Detroit Metropolitan Wayne). A common factor among these is that they are large hub airports, where the spikes may be driven by arrival-side delays rather than departures from Anchorage. However, this cannot be confirmed without also analyzing inbound and outbound flights for those airports.

Unsurprisingly, the leading destination is Seattle, a major transportation hub located relatively close to Anchorage.

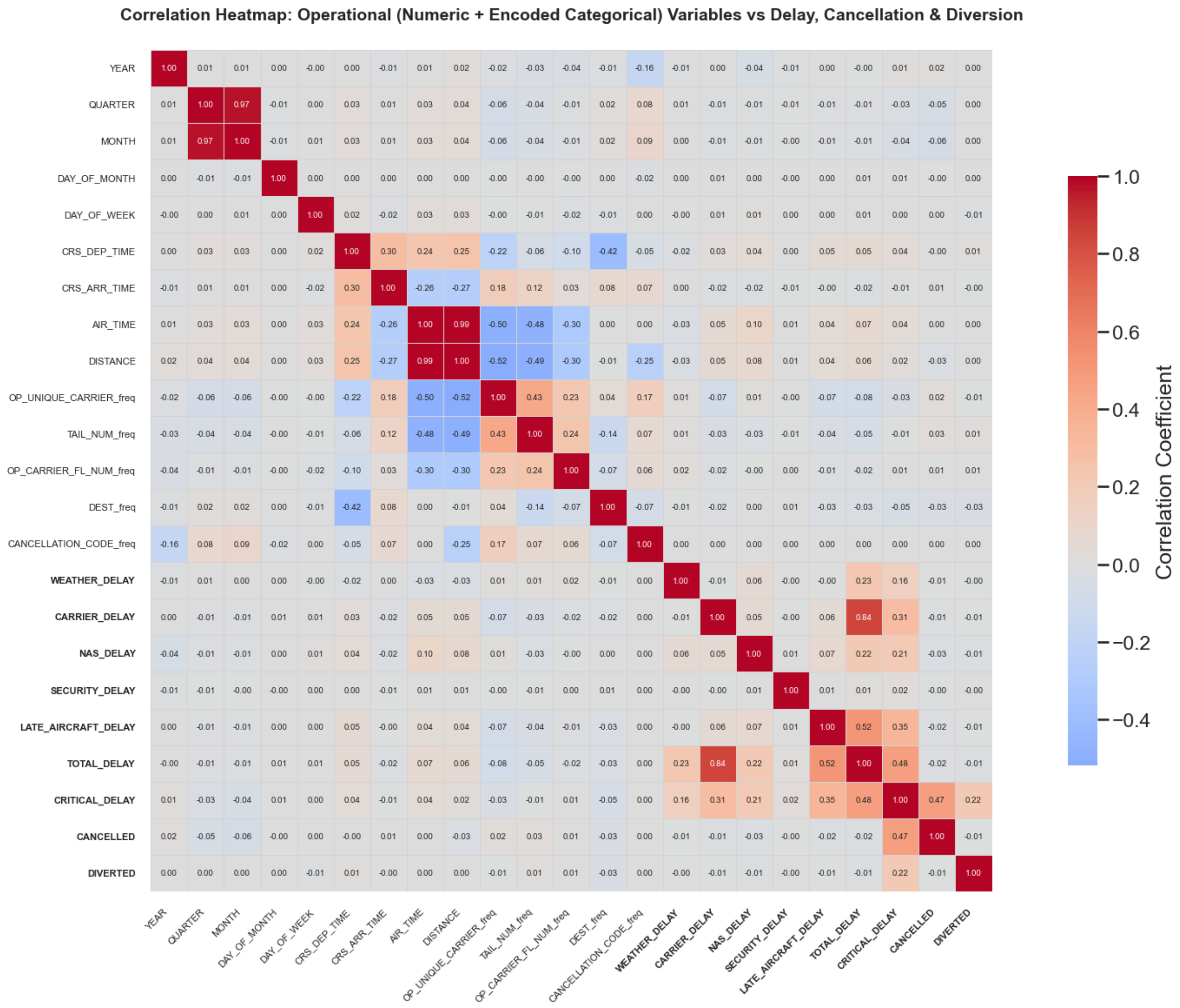
Average delays by departure timing



When examining the distribution of time factors against average delays, it is difficult to identify clear patterns or root causes without additional context. However, several observations stand out and may warrant further investigation by airline or airport staff:

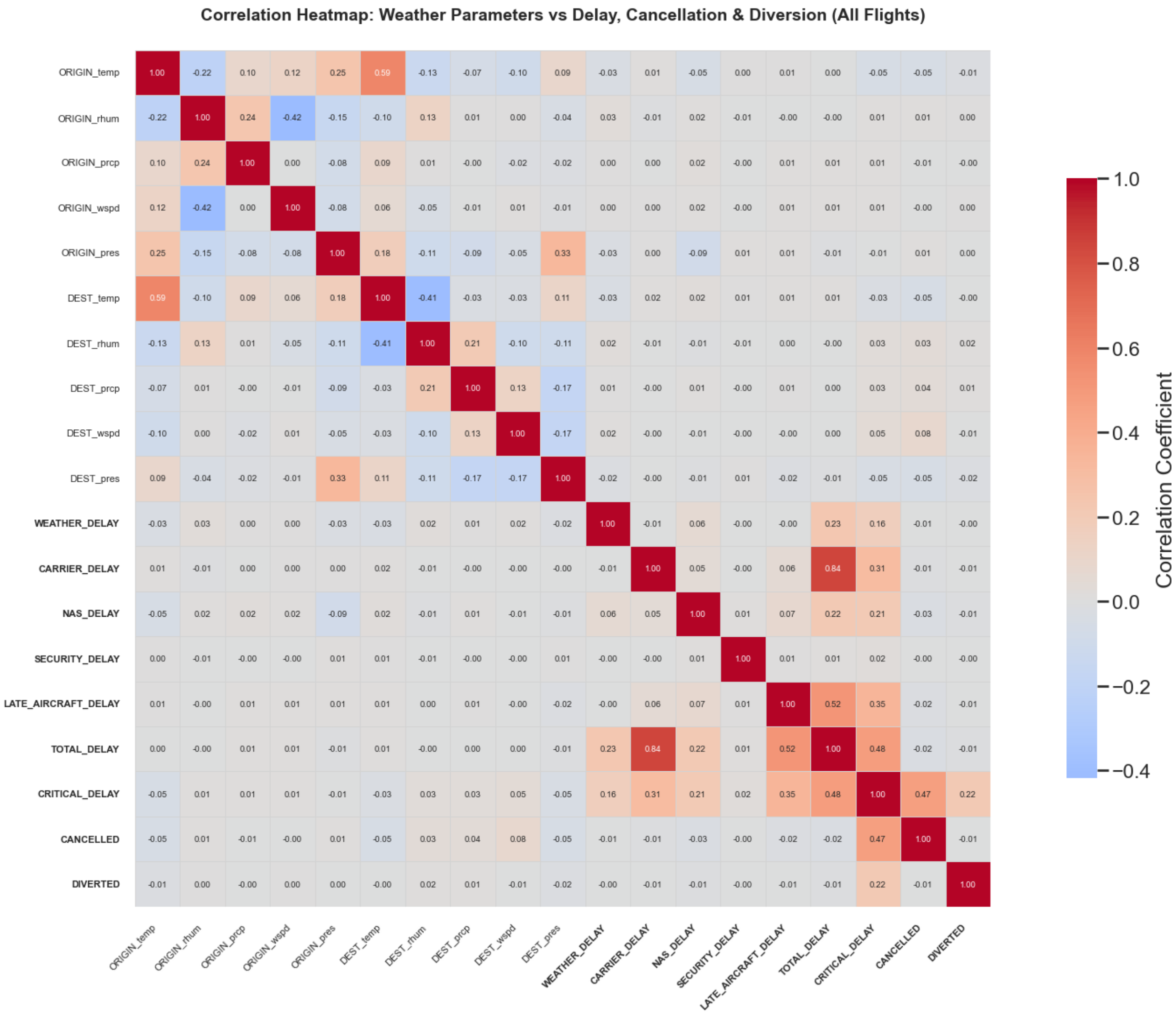
- A noticeable spike in average delays occurs for departures around 20:00.
- Day of the week shows very little variation in delay patterns.
- Day of the month displays slight fluctuations — potentially influenced by behavioral or operational factors.
- Monthly trends suggest that summer months tend to experience more delays (possibly due to higher passenger volumes), while winter months are affected by adverse weather conditions.

Non-weather heatmap



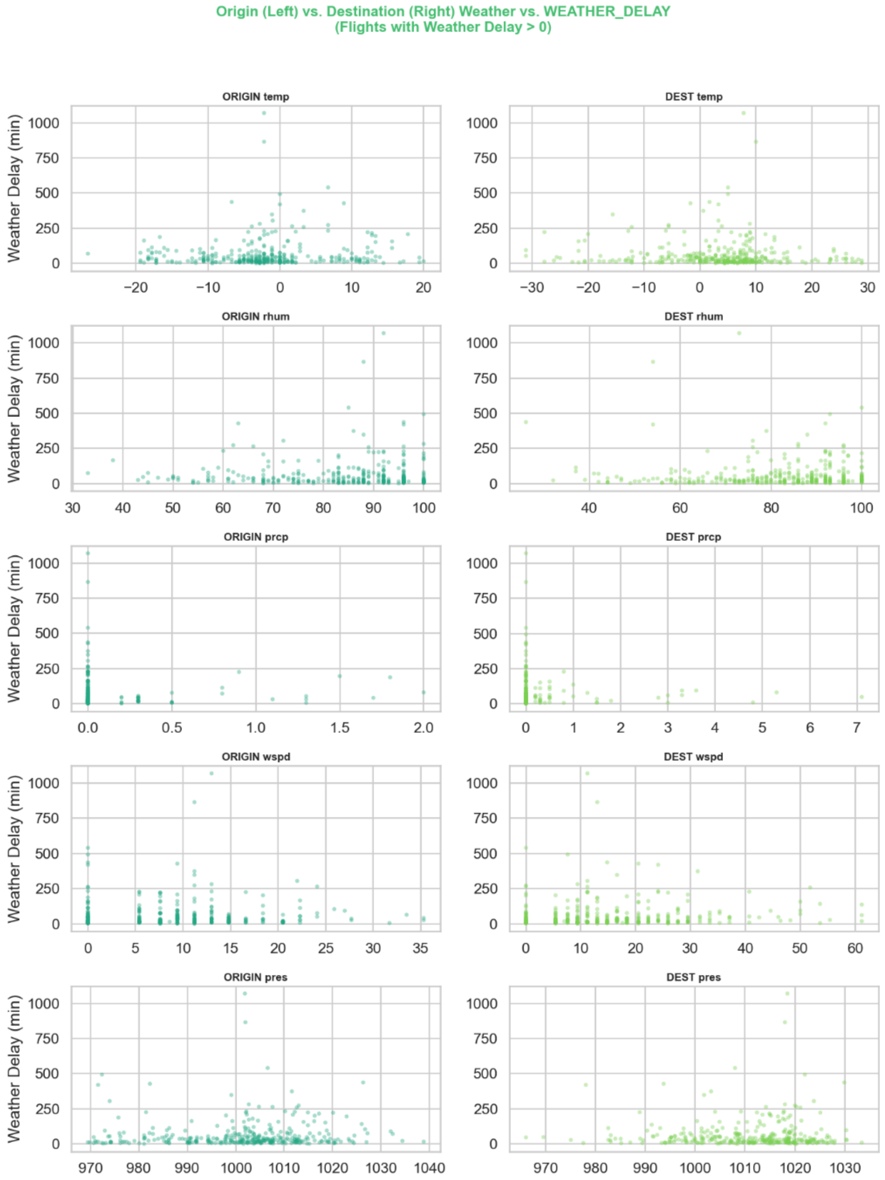
When observing the correlation coefficient heat map of non-weather parameters (in normal text) against the prediction parameters (in bold), very few meaningful conclusions can be drawn. While some parameters are correlated with each other — and the prediction variables also show inter-correlation — there is little correlation between the two groups. This highlights the complexity of flight analytics, where numerous interacting factors influence outcomes, and no single parameter serves as a definitive cause.

Weather heatmap



Similar observations can also be made for weather parameters, with a slight exception for destination windspeed that has 0.05 and 0.08 coefficients respectively for critical delays and cancellations. Albeit these are still quite low.

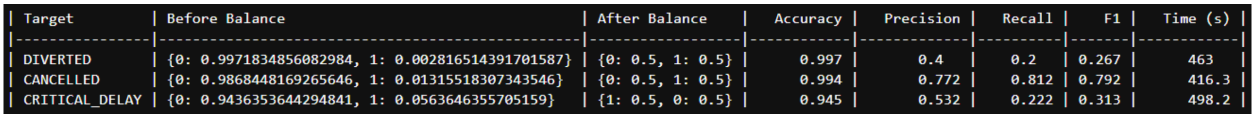
Weather delay for weather parameters



When examining non-zero weather delays against various origin and destination weather parameters, the data appears quite chaotic or random. It almost seems as though weather is not a strong predictor of weather-related delays — or perhaps the chosen weather parameters were not the most optimal. However, a subtle pattern emerges in that the origin and destination weather delay distributions are notably similar to one another.

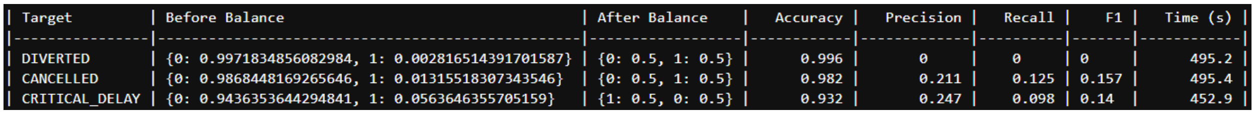
Random Forest Predictions

The initial Random Forest run produced promising F1 scores for predicting cancellations, but poor results for critical delays and diversions. While it was encouraging that at least one target variable performed well, further inspection of the feature importances revealed that *AIR_TIME* carried a disproportionately high weight of around 0.5. Upon closer examination, it became clear that *AIR_TIME* was zero for many cancelled flights, leading the model to learn an unrealistic rule: *0 airtime ⇒ flight likely cancelled*. Recognizing this flaw, *AIR_TIME* was removed, and the models were rerun.



```
| Target         | Before Balance                                | After Balance    |  Accuracy |  Precision |   Recall |    F1 |   Time (s) |
|----------------|-----------------------------------------------|------------------|-----------|------------|----------|-------|------------|
| DIVERTED       | {0: 0.9971834856082984, 1: 0.002816514391701587} | {0: 0.5, 1: 0.5} |     0.997 |        0.4 |      0.2 | 0.267 |        463 |
| CANCELLED      | {0: 0.9868448169265646, 1: 0.01315518307343546}  | {0: 0.5, 1: 0.5} |     0.994 |      0.772 |    0.812 | 0.792 |      416.3 |
| CRITICAL_DELAY | {0: 0.9436353644294841, 1: 0.0563646355705159}   | {1: 0.5, 0: 0.5} |     0.945 |      0.532 |    0.222 | 0.313 |      498.2 |
```

The second run, however, produced significantly lower F1 scores, which was disappointing. While the exercise was valuable from a learning perspective, it became evident that the current dataset is not well-suited for robust machine learning predictions. Even with additional models, tuning, and rebalancing, a dramatic improvement would be unlikely.



```
| Target         | Before Balance                                | After Balance    |  Accuracy |  Precision |   Recall |    F1 |   Time (s) |
|----------------|-----------------------------------------------|------------------|-----------|------------|----------|-------|------------|
| DIVERTED       | {0: 0.9971834856082984, 1: 0.002816514391701587} | {0: 0.5, 1: 0.5} |     0.996 |          0 |        0 | 0     |      495.2 |
| CANCELLED      | {0: 0.9868448169265646, 1: 0.01315518307343546}  | {0: 0.5, 1: 0.5} |     0.982 |      0.211 |    0.125 | 0.157 |      495.4 |
| CRITICAL_DELAY | {0: 0.9436353644294841, 1: 0.0563646355705159}   | {1: 0.5, 0: 0.5} |     0.932 |      0.247 |    0.098 | 0.14  |      452.9 |
```

This is primarily due to two reasons:

1. The main causes of long delays and cancellations are airline-related issues—such as late aircraft or operational factors—which were not included in this dataset. Incorporating prior flight data for specific aircraft registrations could have improved predictive accuracy.
2. Although the dataset appears sizable (approximately 37,000 records), it remains too small for such a complex and interconnected domain. Flight delays are influenced by numerous interdependent variables, and the limited scope of available data restricts meaningful model performance.

Conclusion

This project set out to explore and visualize flight delay data for departures from Anchorage, Alaska, and to assess whether machine learning methods could be used to predict critical flight outcomes such as cancellations, diversions, and significant delays. Through extensive data collection, cleaning, and integration of multiple datasets — including flight records from the Bureau of Transportation Statistics, weather data from Meteostat, and airport location data from OurAirports — a comprehensive analytical framework was established.

The analysis revealed several key insights. Operational factors under airline control — particularly late aircraft and carrier-related issues — were the dominant causes of critical delays, outweighing weather-related factors. This finding was somewhat unexpected given Anchorage's challenging and highly variable climate. Among the airlines analyzed, Alaska Airlines demonstrated the strongest performance, with the lowest proportion of critical delays, while American Airlines exhibited significantly higher delay durations and diversion rates, suggesting potential areas for operational improvement.

Seasonal and event-based analyses confirmed that major disruptions, such as volcanic ash clouds and severe snowstorms, caused short-term spikes in delays, emphasizing the importance of real-time monitoring and proactive operational planning.

The machine learning component, based on a Random Forest model, provided valuable learning experience but achieved limited predictive success. After correcting for data leakage issues (such as the inclusion of *airtime*), model performance declined noticeably. This underscores the challenge of predicting flight outcomes without richer contextual and operational data. The small sample size, high class imbalance, and exclusion of preceding flight information further constrained model accuracy.

In summary, while the exploratory data analysis delivered meaningful operational insights, the predictive modeling highlighted the limitations of the available dataset for developing robust machine learning applications.

Next steps

With additional data availability, several logical next steps are recommended:

- Enhance data inputs, where feasible:
    - Include anonymized passenger information (e.g., age, nationality, number of past flights) to explore behavioral impacts on delays, such as whether airlines tend to wait for late passengers.
    - Incorporate flight path weather data to capture adverse conditions en route.
    - Add previous flight data for each aircraft registration, as prior delays may predict future scheduling disruptions.
    - Integrate airline and airport staff schedules (where available), since staffing levels directly influence operational performance.
- Experiment with additional machine learning models (e.g., LightGBM, CatBoost, XGBoost, or neural networks such as TabNet and MLP) and conduct in-depth hyperparameter tuning. More advanced data balancing techniques should also be explored.
- Incorporate real-time external data sources, such as automated news monitoring tools, to detect events with potential flight disruption impacts — for example, volcanic eruptions, severe storms, earthquakes, protests, or security incidents.
- Reassess the 60-minute threshold for defining critical delays. This arbitrary cutoff could be adjusted to align with regulatory standards or specific operational needs, though doing so would impact all downstream analyses.