

Understanding Flight Delays

Flight delays are a significant challenge for both the airline and passengers, causing disruptions, financial losses, and dissatisfaction.

Analyze historical flight data to uncover delay patterns, identify operational inefficiencies, and predict delays before they occur. By identifying delay patterns, predicting delays, and uncovering the factors that contribute most to delays.

These insights will help the airline make data-driven decisions to optimize scheduling, improve on-time performance, and enhance passenger satisfaction.

Key Finding

- Flight Delay Statistics – Insights from the Data
 - 61.15 % of flights that experienced a departure delay also arrived late.
 - 15.37 % of flights were delayed at departure but still managed to arrive on time.
 - 23.48 % of flights that departed on time ended up arriving late.
- So it's important to focus on decreasing number of flights delayed at departure airport.
- 1.38% of Flights Experienced Severe Delays (Over 3 Hours)
 - While this may seem like a small percentage, severe delays can result in significant financial losses, especially if the causes are attributable to airline-related issues such as operational inefficiencies, maintenance problems, or crew shortages.
 - These delays may trigger compensation obligations, increased passenger dissatisfaction, and disruption to the entire flight schedule, magnifying their impact across the network.
- 22.12% of flights departing from NYC airports experienced delays, with the average delay lasting approximately 65 minutes.
- Atlanta (ATL) is the top destination, receiving the highest number of flights 16,898 (light each 31 minutes)
- Out of 104 destination airports, 82 airports maintained a consistent schedule. The remaining 22 airports experienced irregular flight traffic.
- Hawaiian Airlines operated 342 flights without any cancellations.
- ExpressJet Airlines offers the widest coverage across the U.S. from New York City, serving 61 destinations. However, it also holds the highest cancellation rate among all carriers, with approximately 5.51% of its flights—2,817 out of 51,108—being canceled.
- Delayed and canceled flights Heatmap at 2023 shows
 - July experienced the highest delay rate at 30.19% due to weather condition [source](#), while September saw the lowest delay rate at 14.29%.
 - February had the highest cancellation rate at 15.28%, whereas November had the lowest cancellation rate at 2.82%.
 - High cancellation rate at February warrants further investigation, as no significant evidence of bad weather has been found to explain the spike.
- Based on your flight data, the GradientBoostingClassifier model can predict whether your flight will experience a departure delay, achieving an accuracy of 78.82%.

Departure

Data Overview

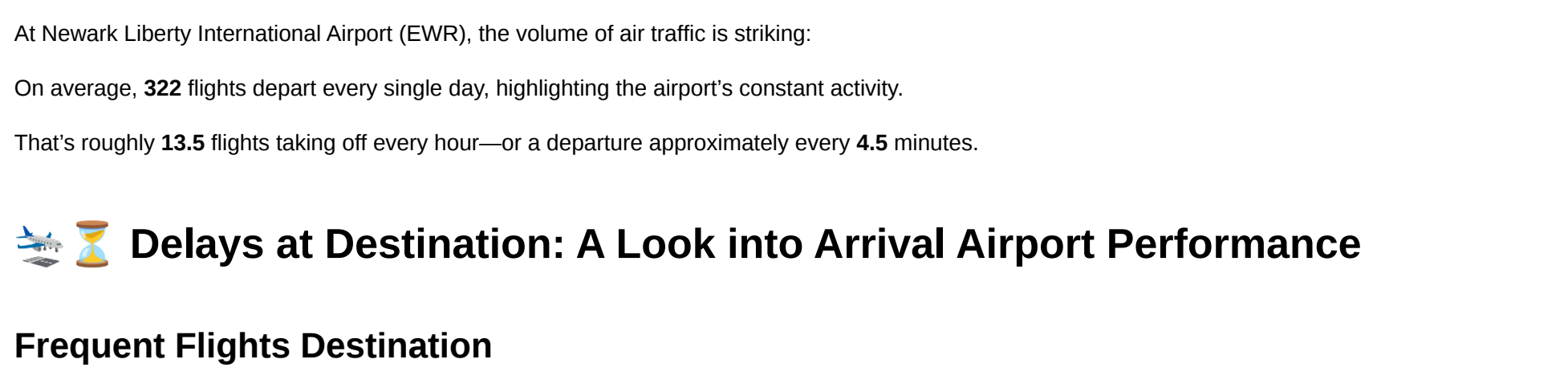
- We have three origin for our flights, these airports are serving New York City
 - JFK : John F. Kennedy Airport - New York (International)
 - EWV : Newark Liberty Airport - New Jersey (International)
 - LGA : LaGuardia Airport - New York (domestic)
- Flights are landing on 104 destination inside USA or its territories (like Puerto Rico and the U.S. Virgin Islands).
- In the United States, the definition of a flight delay varies depending on the context:
 - Federal Aviation Administration (FAA): Considers a flight delayed if it departs or arrives 15 minutes or more after its scheduled time.
 - U.S. Department of Transportation (DOT):
 - Defines a 'significant delay' as a delay of 3 hours or more for domestic flights and 6 hours or more for international flights. Such significant delays may entitle passengers to compensation or refunds.
 - Airlines are not required by law to compensate passengers for flight cancellations if the reason is out of their control.
 - However, if the cancellation is due to a reason within the airline's control (e.g., staffing, mechanical issues), most airlines offer compensation or rebooking options.
- Data cleaning
 - A total of 6,256 flights lacked departure time information, which likely indicates cancellations due to weather disruptions or other operational issues. These records were segregated for focused analysis on cancellations.
 - Additionally, 458 flights were missing arrival time, and 717 records lacked both air time and arrival delay. Since these represented a very small portion of the dataset, we removed them to maintain data integrity without significantly impacting the analysis.



Departure Airport Delay

Newark Liberty International Airport (EWR) emerged as the busiest among the three major New York airports, handling a total of 117,596 departure flights, out of which 20,175 flights were delayed. It was followed by John F. Kennedy International Airport (JFK), which recorded 109,416 total departure flights with 23,347 delays. LaGuardia Airport (LGA) had the lowest traffic volume, with 101,509 total departure flights and 19,792 delayed flights.

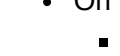
Delayed flights has on average about one hour delay



At Newark Liberty International Airport (EWR), the volume of air traffic is striking:

On average, 322 flights depart every single day, highlighting the airport's constant activity.

That's roughly 13.5 flights taking off every hour—or a departure approximately every 4.5 minutes.



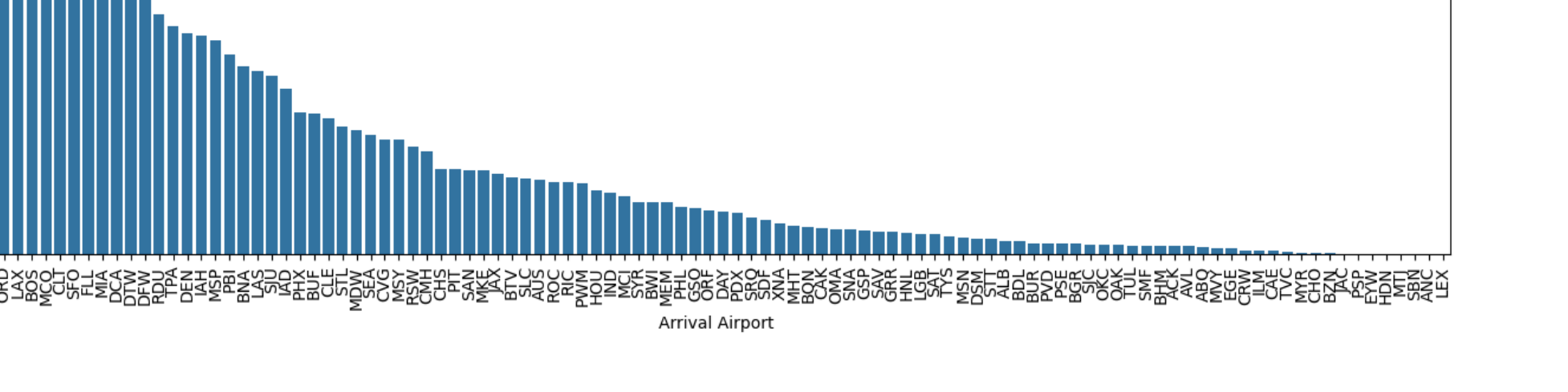
Delays at Destination: A Look into Arrival Airport Performance

Frequent Flights Destination

The data reveals flights from New York City airports cover a wide network of 104 unique destinations across the U.S.

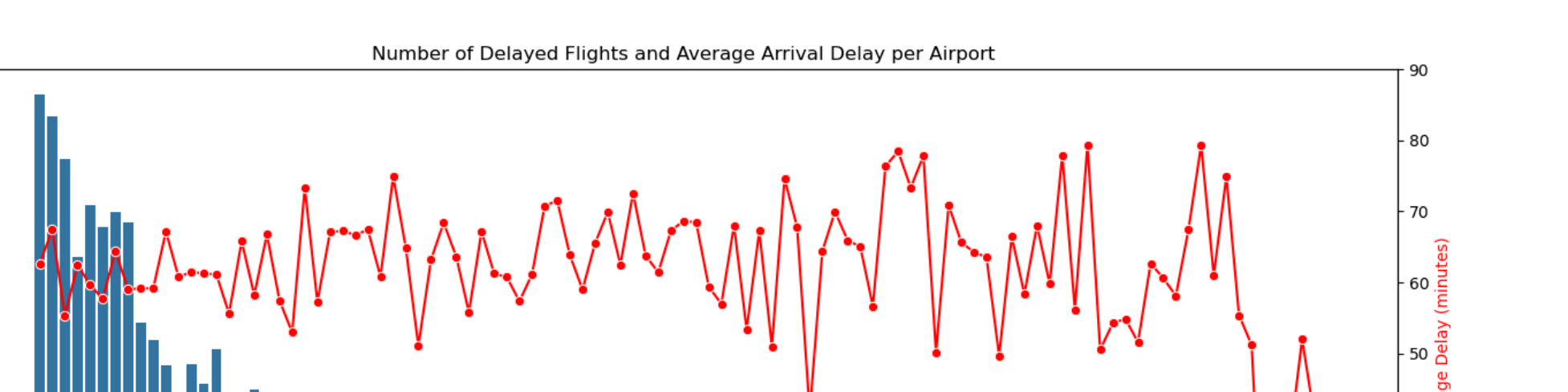
- Atlanta (ATL) is the top destination, receiving the highest number of flights at 16,898, followed closely by Chicago O'Hare (ORD) with 16,642, and Los Angeles (LAX) with 16,076 flights.
- On the other end of the spectrum, some destinations receive only a handful of flights annually:
 - Lexington (LEX) had just 1 flight.
 - Anchorage (ANC) received 8 flights.
 - South Bend (SBN) recorded 10 flights for the year.

This highlights both the scale and diversity of flight distribution from NYC airports—ranging from major hubs with constant traffic to niche routes serving specific regional demands.



Flights Arrival Delay

On average delay arrival time is the same regardless the count of flight that airport receive



Airports that recorded low flight counts from NYC, it's highly likely that these were non-regular, possibly chartered or private flights rather than part of a scheduled commercial service.

Some possible reasons:

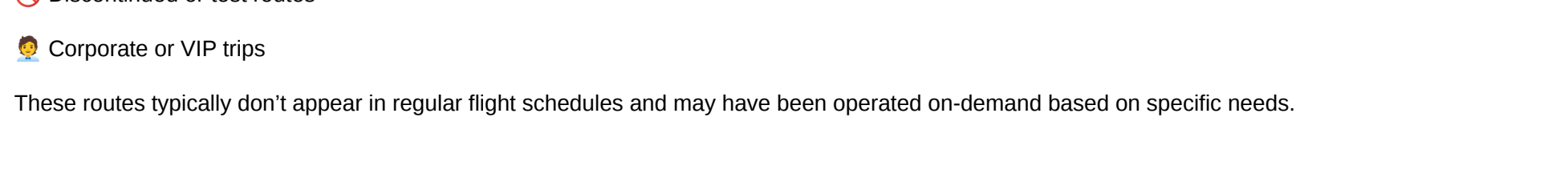
- Private business or government travel
- Seasonal or special events
- Discontinued or test routes
- Corporate or VIP trips

These routes typically don't appear in regular flight schedules and may have been operated on-demand based on specific needs.

Most Delayed Arrival Airports

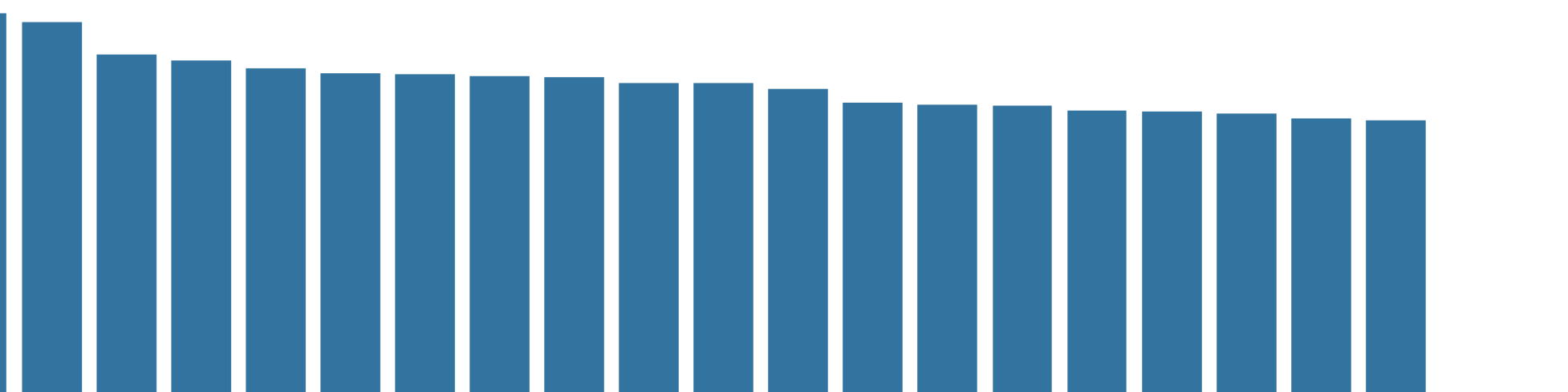
Tulsa International Airport (TUL) recorded the highest average arrival delay at 72.5 minutes.

However, this does not necessarily reflect poor performance at TUL itself. The data shows that the average delay for flights departing to TUL already begins delayed at the origin airport, with an average departure delay of 68.4 minutes. This suggests that the majority of the delay occurs before the flight even leaves, rather than being caused by operations at Tulsa International Airport.



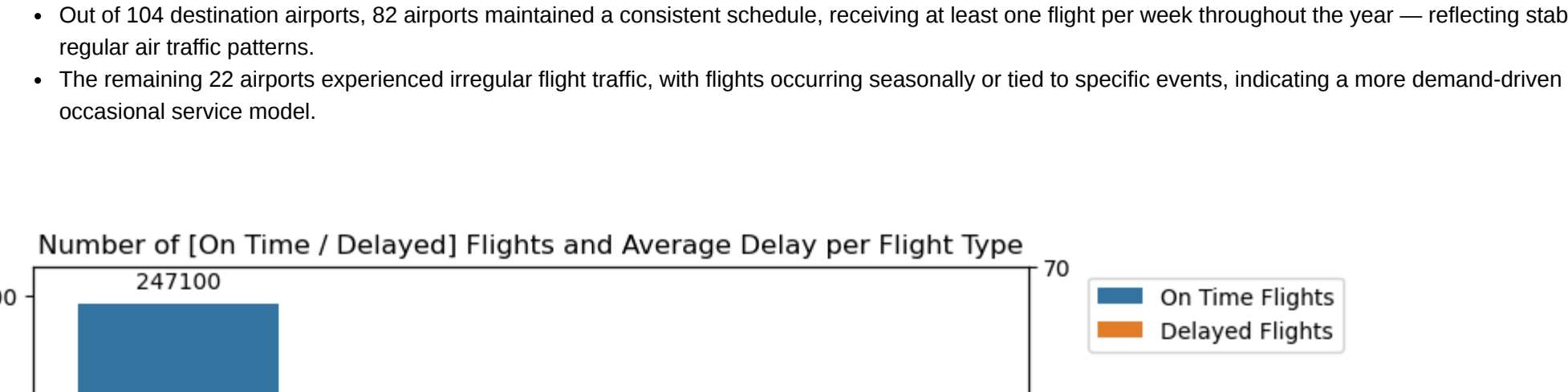
Scheduled vs. Irregular Flights: Uncovering Patterns in Air Traffic Operations

- Out of 104 destination airports, 82 airports maintained a consistent schedule, receiving at least one flight per week throughout the year—reflecting stable, regular air traffic patterns.
- The remaining 22 airports experienced irregular flight traffic, with flights occurring seasonally or tied to specific events, indicating a more demand-driven or occasional service.



Flight Route Performance — Top 50 Routes by Volume and Average Arrival Delay

- High Volume, Low Delay:
 - Routes like JFK → LAX (11,559) flight per year – 30 flight per day, 1 flight every 47 minutes.
 - Despite high traffic, (average) arrival times remain moderate, suggesting efficient operations on these busy routes.
- Low Volume, High Delay:
 - Routes towards the right end, such as EWR → STL and EWR → CVG have fewer flights but relatively higher delays.
 - This might indicate irregular scheduling air traffic control, or gate availability.



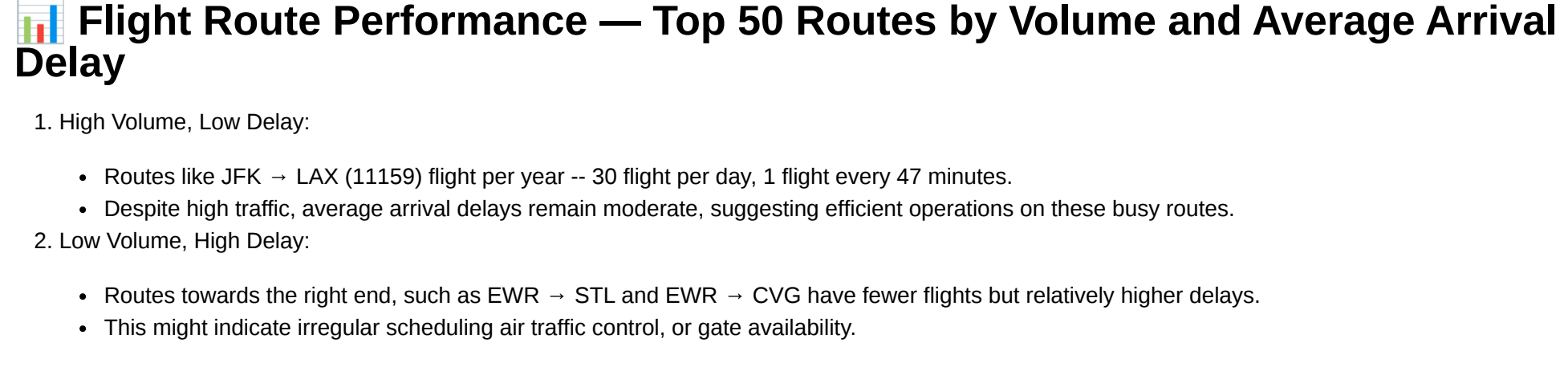
Airline Flight Volume and Average Delay

- United Airlines Inc. operates the highest number of flights among all carriers, maintaining an impressive average delay of under 5 minutes, indicating efficient operational performance.
- In contrast, ExpressJet Airlines Inc., despite being among the top 3 airlines in terms of flight volume, reports an average departure delay exceeding 15 minutes, highlighting potential scheduling or operational challenges.
- Smaller airlines exhibit a high degree of variation in delay times, suggesting that limited resources or less frequent scheduling may contribute to inconsistent performance.



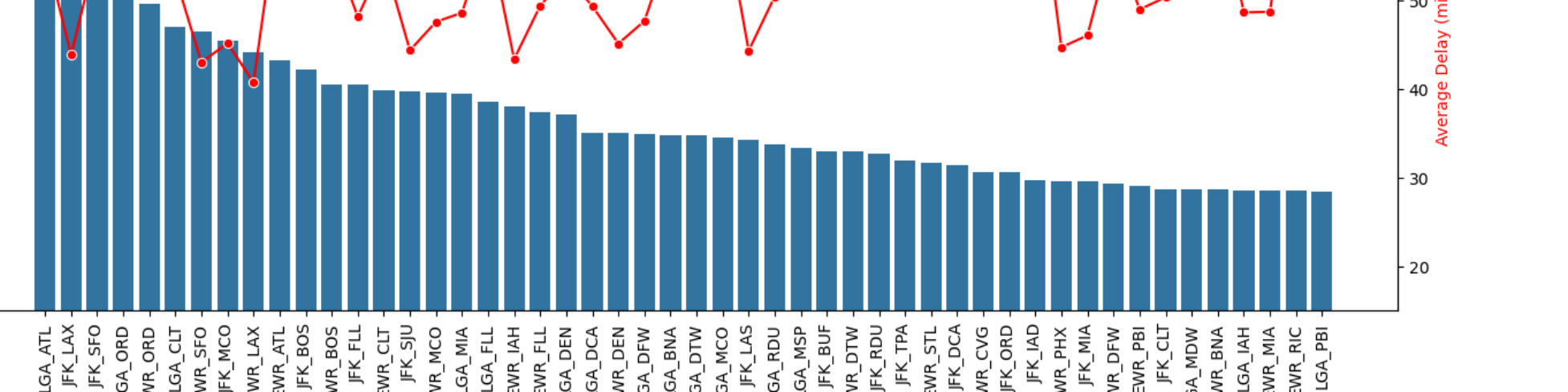
Airlines Destinations

- ExpressJet Airlines serves 61 destinations, outperforming United Airlines, which—despite operating a higher number of flights—covers only 47 destinations, 14 fewer than ExpressJet.
- Hawaiian Airlines, Alaska Airlines and Frontier Airlines are serving only one destination



Canceled Flights for Airlines

- ExpressJet Airlines recorded the highest number of canceled and delayed flights in 2023, with a relatively modest 25,037 flights and serving only 20 destinations.
- Endeavor Air ranked second in cancellations, with 1,234 canceled flights, despite operating a relatively modest 25,037 flights and serving only 20 destinations.
- Endeavor Air ranked third in with 1044 canceled flights, it serves a wider network of 48 destinations while operating a lower volume of 17,294 flights, indicating a broader reach with fewer total operations.



Comparing the Top 3 Airline carrier we noticed

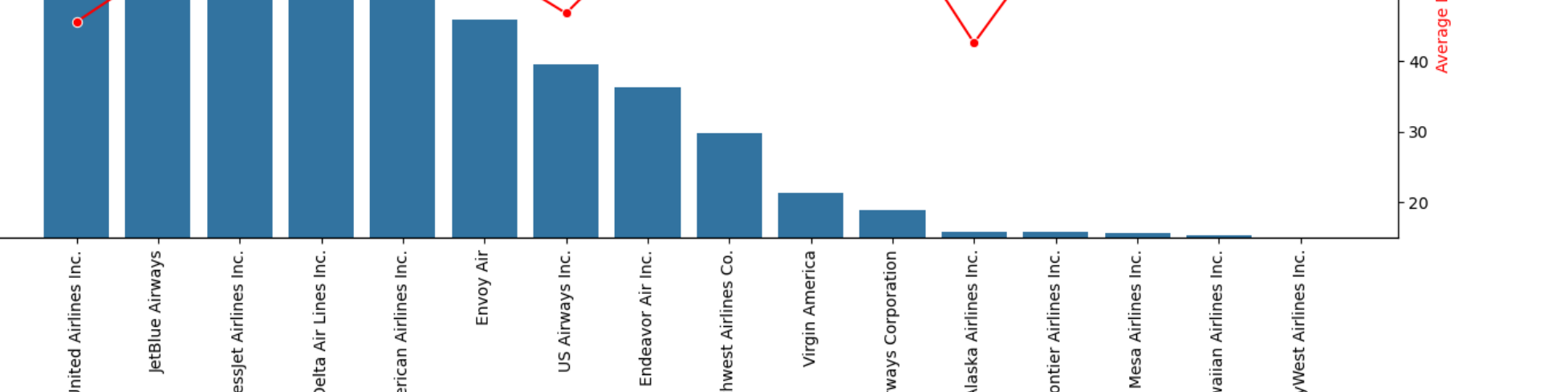
- ExpressJet Airlines Inc. stands out with the highest number of delayed and canceled flights, yet it also covers the most destinations, indicating a broad but possibly overextended operational network.
- United Airlines Inc. operates the largest volume of flights, but across a more limited number of routes. Interestingly, it maintains the lowest delay rate, highlighting efficient route management.
- JetBlue Airways follows a similar strategic approach to United, focusing on fewer cancellations and demonstrating strong performance with the lowest cancellation rate among the top carriers.

Overall Relationship

- More destinations don't guarantee fewer cancellations. ExpressJet serves the most destinations but also leads in cancellations (5.5% of its flights).
- Higher flight volume with fewer destinations (like United) may benefit from centralized operations and better reliability.
- Airlines with fewer destinations and flights (like Endeavor and Endeavor) can still face high cancellation rates, suggesting that operational quality, not just scale, influences reliability.

Monthly Departure Delay Trends Across NYC Airports at 2023

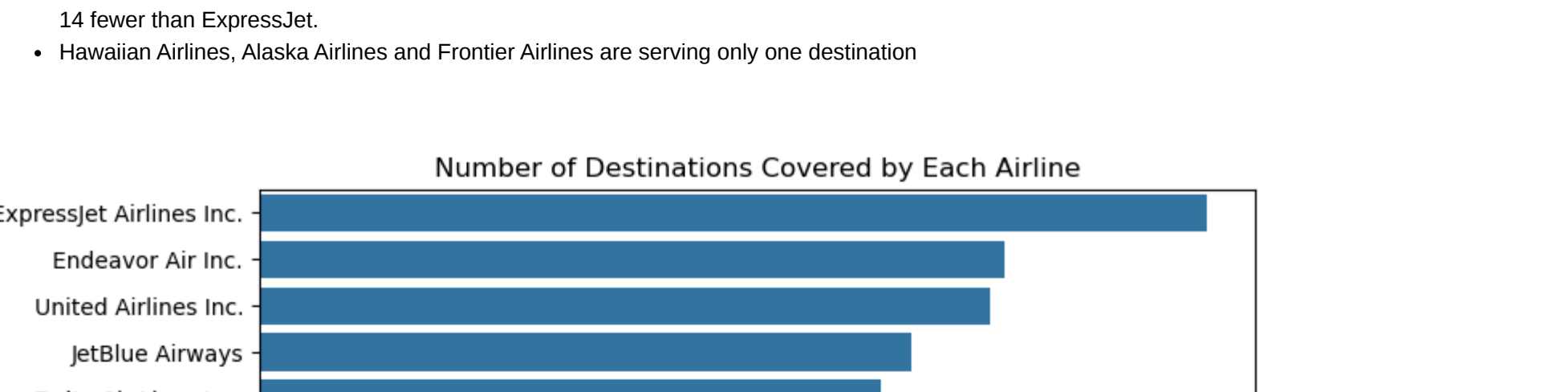
- All three airports—JFK, LGA, and EWR—show a noticeable spike in departure delays during the summer months (June to August), with JFK peaking in July at the highest delay.
- Departure delays drop significantly in September through November, hitting the lowest in October and November across all airports, indicating more stable operations in fall.



Tracking Daily Flight Delays: Heat Map of Delayed Departures

A closer look at daily flight delays throughout 2023—visualized via a heatmap—reveals key patterns. Darker cells indicate days with a high volume of delays, serving as clear signals for further investigation.

These peaks could point to underlying issues such as severe weather, operational issues, or seasonal travel surges that warrant deeper analysis.

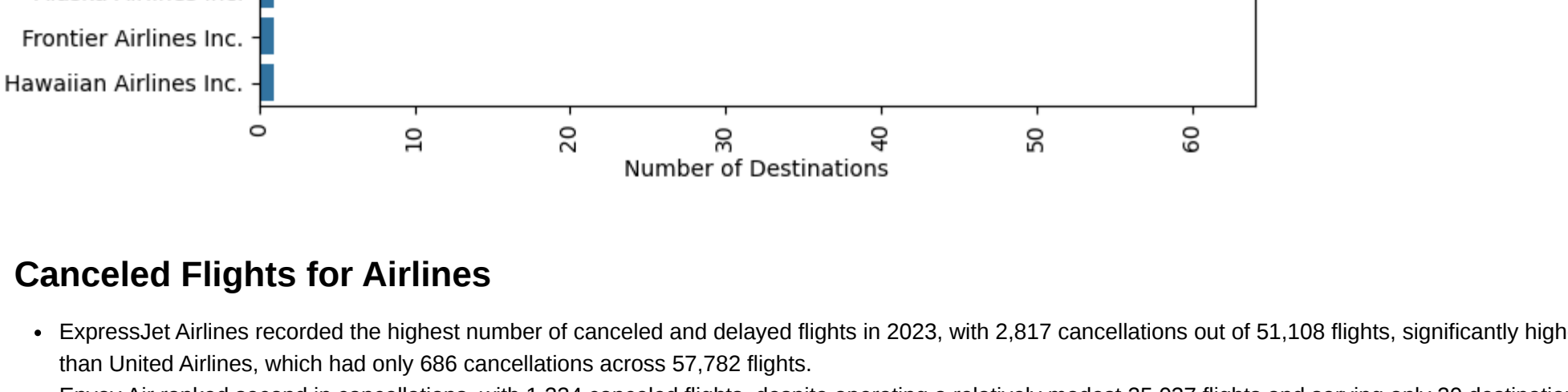


By quick search on web

- June experienced a surge in long delays, primarily due to frequent thunderstorms disrupting flight schedules.
- July saw extended delays caused by multiple rounds of heavy rainfall, impacting operations over several days.
- In September, 2 days delay appeared again, coinciding with intense rainfall events.
- October continued the trend, with persistent heavy rainfall contributing to elevated delay times.
- December was marked by storms and widespread rain, leading to a notable increase in delay lengths, particularly around the holiday travel period.

These findings highlight how seasonal weather patterns, especially in summer and winter, have a significant impact on flight punctuality.

Understanding Flight Cancellations and Delays: A Head-to-Head Comparison



By comparing the flight cancellation map with the delay map, a clear correlation emerges—days with a high volume of delayed flights often coincide with a notable increase in cancellations. This pattern suggests frequently escalates into flight cancellations.

However, there's a notable exception on February 8 and 9. During these two days:

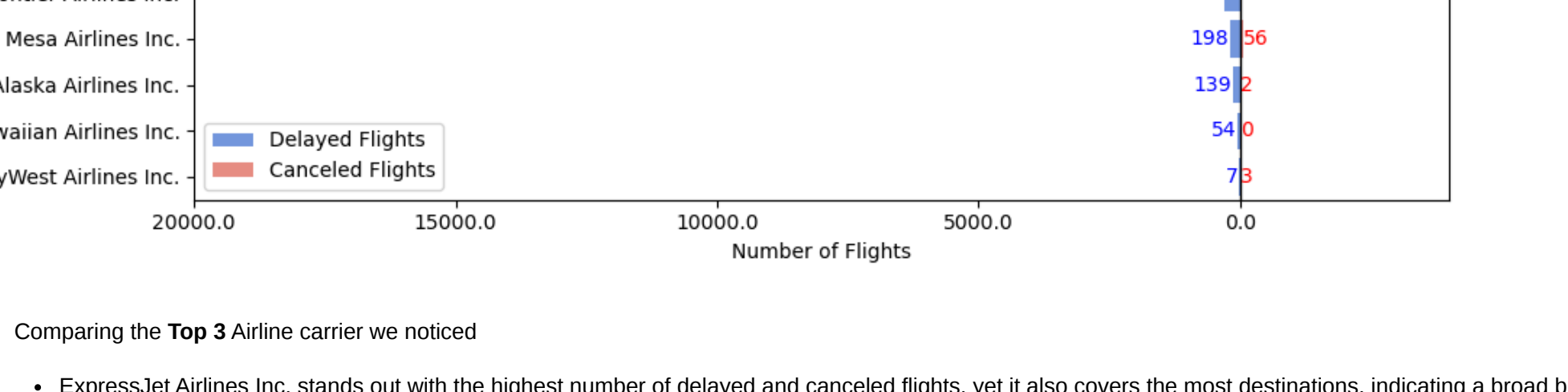
- The number of cancellations surged dramatically, despite the overall delay rates remaining within normal averages.
- This anomaly indicates that cancellations may have been driven by non-delay-related factors, such as weather warnings, staffing issues, or ground operations disruptions.

Investigating the Drop in Flight Volume – February 2023

Upon examining the flight data for February 2023, there's a notable drop in the total number of recorded flights from all departure airports on February 8 and 9.

Is the drop a result of incomplete or missing data entries, where flights were not logged or erroneously excluded? Or does it reflect a real operational impact, such as an increase in cancellations, weather-related disruptions, or air traffic control limitations?

Only by verifying both data integrity and external operational factors can we confirm whether this decline reflects a real-world disruption or a data reporting artifact.



Developing a Model to Forecast Flight Departure Delays

Based on 2023 flight data, we will build a predictive model to determine whether a flight will be delayed, using only ticket-related details.

To ensure the model reflects real-world performance, we will take the following preprocessing steps:

- Exclude actual flight timing data such as `dep_time`, `dep_delay`, `arr_time`, and `arr_delay` to avoid data leakage.
- Remove the `tailnum` column, as it provides limited predictive value and may introduce unnecessary noise.
- Engineer a new feature by extracting the week number of the year, which can help capture seasonal trends in delays.
- Apply label encoding to categorical columns, including `origin`, `dest`, `airline_name`, and `flight_type`, to make them suitable for machine learning algorithms.

Understanding feature importance in our model reveals that `day`, `air_time`, and `week_number` are the most influential variables in predicting flight delays



```
LogisticRegression() :
Training Accuracy : 0.586794980180827
Validation Accuracy : 0.5843676697188445
Accuracy : 0.775868692555955
Precision: 0.3037190082644258
Recall: 0.6148258534666697
F1 score: 0.01964190272581507
[15049 2371
14337 1471]
```

```
DecisionTreeClassifier() :
Training Accuracy : 1.0
Validation Accuracy : 0.6121847979767475
Accuracy : 0.726851932733646
Precision: 0.388065978583197
Recall: 0.61651753659202616
F1 score: 0.3970464277255141
[14689 9287]
[5586 5888]
```

```
RandomForestClassifier() :
Training Accuracy : 0.794792191140186
Validation Accuracy : 0.739039010434125
Accuracy : 0.7880769115518586
Precision: 0.6148258534666697
Recall: 0.09513946423638878
F1 score: 0.1571462930671638
[150217 7371]
```

The GradientBoostingClassifier achieved the highest performance, with an accuracy of 78.01% in correctly predicting whether a flight will be delayed or not.

Conclusion:

- The date of your travel can help predict your flight delay status, as it is often correlated with weather conditions.
- Be sure to check weather forecasts in advance of your planned flight to stay informed.
- Choose your airline carefully, and research the airline's reputation for reliability and punctuality.
- Review the refund policy to understand your options in case of cancellations or delays.