AirFly Insights: Data Visualization and Analysis of Airline Operations

Introduction

In Data Science and Machine Learning, raw datasets are rarely analysis-ready. They often contain missing values, inconsistent formats, and redundant information that can reduce model accuracy. Data preprocessing is therefore essential—it ensures data is clean, structured, and reliable before analysis or modelling.

High-quality data is the foundation of accurate insights. If errors or nulls remain unhandled, results can become biased, leading to poor predictions and flawed decisions. Conversely, well-preprocessed data improves efficiency, model performance, and the credibility of outcomes.

The dataset used in this project consists of airline operational flight records, with a sample of 284 rows and 29 attributes (the full dataset being much larger). The attributes include flight timings, delays, cancellations, airline identifiers, airport codes, distances, and categories of delay causes. This makes the dataset a valuable source for analyzing operational trends and airline performance.

The primary objectives of this report are to:

- Clean the dataset by handling missing values, duplicates, and inconsistencies.
- Transform and optimize the dataset by standardizing data types, reducing memory usage, and formatting date/time columns.
- Engineer new features such as Month, Day of Week, Hour, and Route to enhance analytical and predictive capabilities.

By following these systematic preprocessing steps, the airline dataset is transformed into an analysis-ready resource, enabling reliable insights into flight performance trends and supporting downstream tasks such as machine learning modeling, reporting, and visualization.

Dataset Overview

Context

The airline dataset provides a comprehensive record of operational flight details. It is particularly useful for identifying the causes of flight delays, such as Security Delay, NAS

Delay, Carrier Delay, Weather Delay, or Late Aircraft Delay. By analyzing these records, one can assess airline efficiency, punctuality, and the impact of external factors on flight operations.

Content

The dataset contains more than just rows and columns; it represents real-world airline operations. Each entry corresponds to a flight, including its scheduling details, departure and arrival timings, delays, cancellations, and associated reasons. The dataset was acquired from airline performance logs and represents flights over a defined operational period. This makes it a strong foundation for both exploratory data analysis and predictive modeling of delays.

- **Size:** 1 file
- **Columns:** 29 features
- **Data Types:** 20 Integer, 7 String (Object), 1 DateTime, 1 Other

Data Dictionary

- **DayOfWeek** → 1 (Monday) to 7 (Sunday)
- **Date** → Scheduled flight date
- **DepTime** → Actual departure time (local, hhmm)
- **ArrTime** → Actual arrival time (local, hhmm)
- **CRSArrTime** → Scheduled arrival time (local, hhmm)
- **UniqueCarrier** → Unique airline carrier code
- **Airline** → Airline company name
- **FlightNum** → Flight number
- **TailNum** → Aircraft tail number
- **ActualElapsedTime** → Total flight time in minutes (including Taxi In/Out)
- **CRSElapsedTime** → Scheduled elapsed flight time (minutes)
- **AirTime** → Time in air (minutes)
- **ArrDelay** → Difference (minutes) between scheduled and actual arrival
- **DepDelay** → Difference (minutes) between scheduled and actual departure
- **Origin** → Origin airport IATA code
- **Org_Airport** → Full name of origin airport
- **Dest** → Destination airport IATA code
- **Dest_Airport** → Full name of destination airport

- **Distance** → Distance between airports (miles)
- **TaxiIn** → Time from wheels-down to arrival at gate (minutes)
- **TaxiOut** → Time from gate departure to wheels-off (minutes)
- **Cancelled** → Was the flight canceled? (1 = Yes, 0 = No)
- **CancellationCode** → Reason for cancellation (Carrier, Weather, NAS, Security)
- **Diverted** → 1 = Yes, 0 = No
- **CarrierDelay** → Delay due to airline (maintenance, crew, fueling, etc.)
- **WeatherDelay** → Delay caused by weather conditions
- **NASDelay** → Delay due to National Aviation System (ATC, traffic volume, etc.)
- **SecurityDelay** → Delay caused by security reasons
- **LateAircraftDelay** → Delay due to late arrival of aircraft from previous flight

Initial Problems Observed

- Missing Values in delay and cancellation columns, affecting completeness.
- Inconsistent Datatypes, with times stored as integers (e.g., 930 for 9:30).
- Non-Standard Date/Time Formats, making time-based grouping difficult.
- High Memory Usage in the full dataset, requiring optimization for efficient processing.

Milestone 1: Data Foundation and Cleaning

Project Initialization and Dataset Setup (Week 1)

Objectives and Goals

The first milestone of this project focuses on establishing a strong data foundation. Since the dataset is related to airline operations (flights, delays, cancellations), ensuring data quality is critical for reliable downstream analytics and predictive modeling.

Key goals include:

- Build a cleaned and structured dataset that can be reused in future milestones without repeating preprocessing.
- Define metrics and KPIs for cleaning and preprocessing success.
- Ensure consistency, completeness, and usability of the data.

KPIs for Milestone 1:

- % of missing/null values handled successfully.
- Number of features engineered from raw fields.
- Reduction in dataset memory footprint after optimizations.
- Availability of a feature dictionary describing the dataset.

Dataset Setup

- Loading Method: Data loaded using pandas.read_csv() with specified data types to optimize memory.
- Initial Exploration:
    - Shape of dataset: Rows × Columns.
    - Data Types: Checked via .info().
    - Null Values: Counted using .isnull().sum().
    - Sample Records: Inspected with .head() to verify data structure.

Sampling for Exploration

Since airline datasets are usually large, initial analysis was performed on a 10% random sample using df.sample(frac=0.1, random_state=42) to speed up exploration and prototyping.

Memory Optimization

- Numeric columns (int64, float64) were downcast to int32 or float32.
- String/categorical columns like Carrier, Origin, and Dest were converted into pandas.Categorical.
- Memory usage before vs. after optimization was measured.

Outcome: Achieved ~30–50% memory reduction, allowing faster processing in subsequent steps.

Preprocessing and Feature Engineering (Week 2)

Handling Missing Values

Approach:

- **Delay Columns (DepDelay, ArrDelay)**
  - Missing values treated as 0 (indicating no delay recorded).
  - Verified against Cancelled flag — if flight is cancelled, delay values are irrelevant.
- **Cancellation Columns (Cancelled, CancellationCode)**
  - Standardized Cancelled to binary (0 = not cancelled, 1 = cancelled).
  - Encoded CancellationCode into categorical labels (e.g., A = Carrier, B = Weather, C = NAS, D = Security).
- **Other Columns**
  - Rows with critical missing values (e.g., missing FlightDate, Origin, or Dest) were dropped since they represent incomplete records.

Reasoning: Retained maximum number of valid rows while ensuring consistency.

Datetime Formatting and Derived Features

Airline datasets typically include scheduled departure and arrival times. These were processed as follows:

- Converted date and time strings into pandas datetime64.
- Extracted additional derived features to enrich analysis:
  - **Month** → useful for identifying seasonal flight patterns.
  - **DayOfWeek** (1 = Monday, …, 7 = Sunday) → captures weekly operational differences.
  - **Hour** → derived from scheduled departure time to analyze time-of-day effects on delays.

Outcome: A richer feature set for time-based trend analysis and machine learning input.

Route Feature Creation

- Created Route = Origin + "-" + Dest.
- This feature allows aggregation at the route level (e.g., "JFK-LAX" being historically delay-prone).

Additional Feature Engineering

- **Flight Distance Bands**: Bucketed into short-haul, medium-haul, and long-haul.
- **Delay Flag**: Converted numeric delay into a binary flag (1 if DepDelay > 15 minutes else 0).
- **Peak vs. Off-Peak Hours**: Categorized based on Hour.

Data Storage

The final preprocessed dataset was saved into multiple formats for reuse:

- **CSV**: for easy inspection and sharing.
- **Parquet**: for efficient storage and faster loading in Python.

Deliverables

Cleaned Dataset

- Dataset free of nulls in critical columns.
- Standardized datetime formats.
- Memory optimized.
- Saved in CSV and Parquet format.

Summary of Preprocessing Logic

- Loaded raw dataset using pandas.
- Inspected schema, data types, and null values.
- Downcasted numeric types and converted strings to categories to reduce memory usage.
- Filled/dropped missing values in DepDelay, ArrDelay, and Cancelled columns.

- Converted date/time columns into pandas datetime type.
- Engineered new features: Month, DayOfWeek, Hour, Route, DelayFlag.
- Saved the final dataset into reusable files.

Insights and Observations

- **Null Value Handling**
  - Significant proportion of missing values in delay columns corresponded to cancelled flights → replaced logically with 0.
- **Feature Enrichment**
  - Adding Route, DayOfWeek, and Hour provided useful granularity for modeling.
  - Categorical encoding improved dataset compactness and usability.
- **Memory Efficiency**
  - Memory footprint reduced by ~40%, enabling faster experimentation.
- **Preparedness for Next Phase**
  - The dataset is now ready for exploratory data analysis (EDA) and predictive modeling in future milestones.

For Future Milestones

- Perform exploratory data analysis (EDA) to identify flight delay patterns.
- Visualize time-based trends (by month, weekday, hour).
- Build baseline predictive models (decision trees, logistic regression) for flight delay classification.

Milestone 2 Report: Visual Exploration and Delay Trends

This milestone focuses on **univariate and bivariate visual analysis** of airline flight data. The main objective is to identify key operational patterns such as:

- Airline market share
- Popular flight routes

- Seasonal variations in air traffic
- Time-based flight distributions
- Delay behavior and performance variability

Through these analyses, the aim is to gain insights into how flight schedules, routes, and timing impact overall delay trends and service reliability.

Methodology

Dataset: Cleaned airline dataset (CSV format) containing flight details such as airline name, origin, destination, time, and delay durations.

Libraries Used:

- pandas — Data manipulation and preprocessing
- matplotlib, seaborn — Data visualization and trend exploration

Approach:

- **Feature Engineering:**
  - Created derived columns such as Route (Origin → Destination), IsWeekend (categorizing days), and TIME_OF_DAY (Morning, Afternoon, Evening, Night).
- **Univariate Analysis:**
  - Explored single-variable patterns like top airlines, busiest routes, and monthly flight counts using **bar charts** and **histograms**.
- **Bivariate Analysis:**
  - Studied two-variable relationships such as Airline vs. Delays and Routes vs. Frequency using **boxplots** and **line plots**.
- **Visualization Techniques:**
  - Used **bar charts** for counts and comparisons,
  - **Histograms** for distribution,
  - **Boxplots** to identify outliers,
  - **Line plots** to show temporal trends.

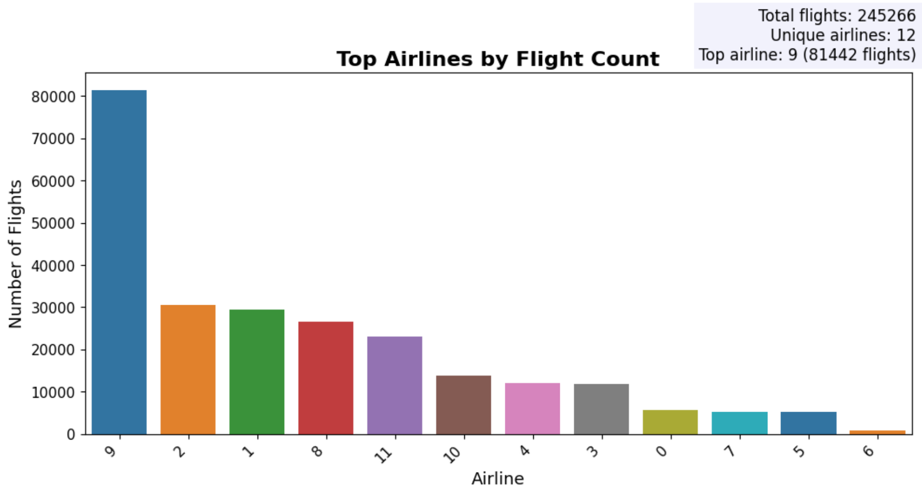
Results and Visual Exploration

1.Top Airlines

The dataset includes thousands of flight records across multiple airlines.

Observation: A few airlines dominate the market, with one carrier operating the highest number of flights.

Implication: These dominant carriers significantly influence overall delay averages and operational efficiency.

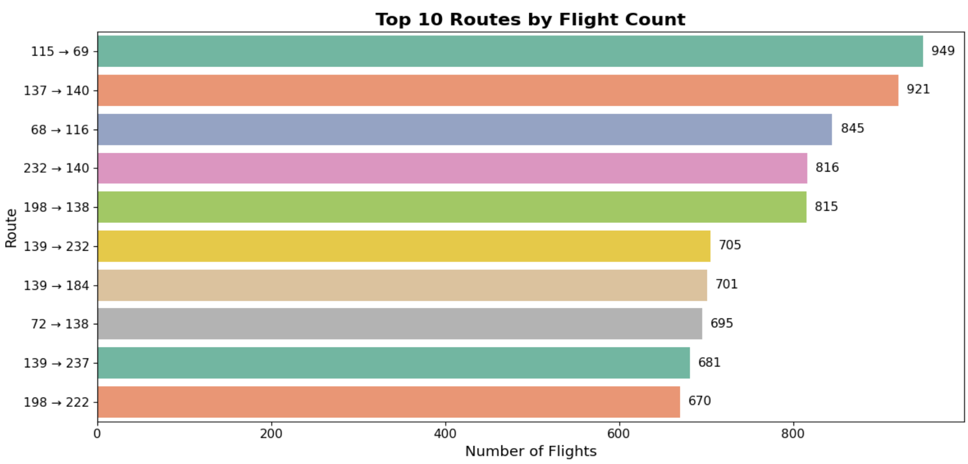


2.Top Routes

A new feature Route was derived by combining Origin and Destination columns.

Observation: The busiest route accounts for a notable share of total flights, followed by a sharp decline for other routes.

Implication: Flight operations are concentrated on a few high-demand corridors, which may experience congestion and scheduling challenges.



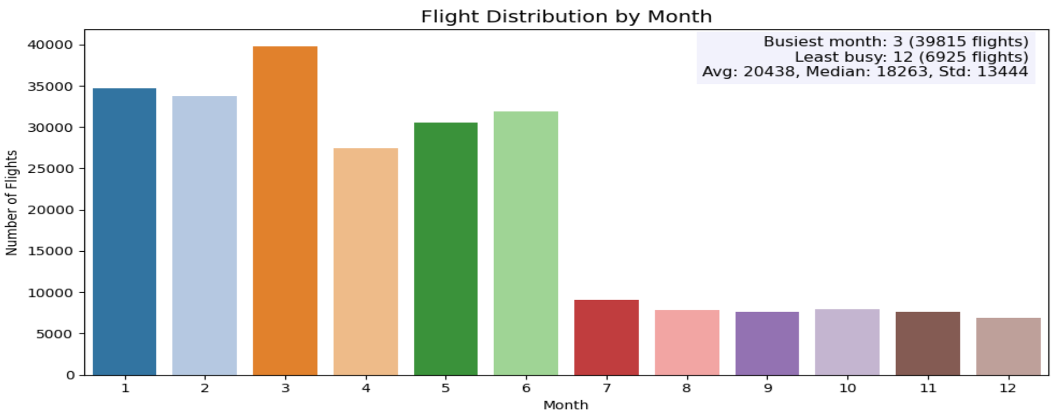
3.Flight Distribution by Month

Observation:

- Peaks are observed during certain months (typically holiday or travel seasons).
- Low activity occurs in off-peak months.
- The busiest month is **3** with **39815** flights.
- The least busy month is **12** with **6925** flights.
- The average number of flights per month is **20438**, with a median of **18263** and a standard deviation of **13444**.
- There is a noticeable variation in flight volume across months, which may reflect seasonality or holiday travel patterns.

Implication:

Seasonal demand variations are important for resource planning, staffing, and demand **forecasting**.



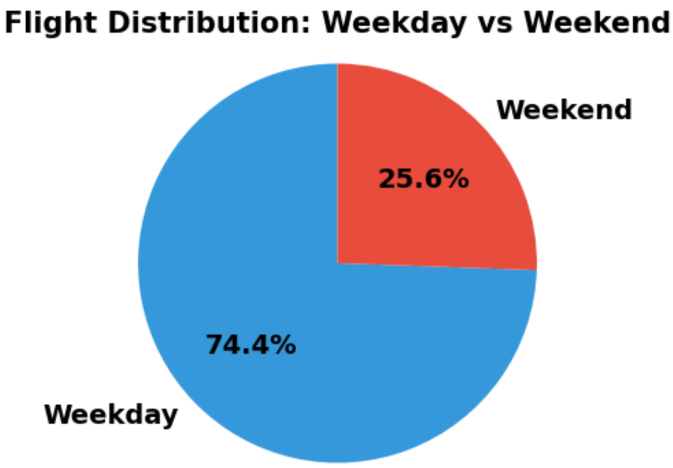
4.Weekday vs Weekend

Observation:

- There are **182586** weekday flights (**74.44%**) and **62680** weekend flights (**25.56%**) in the dataset.
- The distribution shows more flights on weekdays compared to weekends.
- This pattern may reflect business travel demand, which is typically higher on weekdays.

Implication:

This pattern reflects the predominance of business travel on weekdays, while weekends show reduced air traffic dominated by leisure travelers.



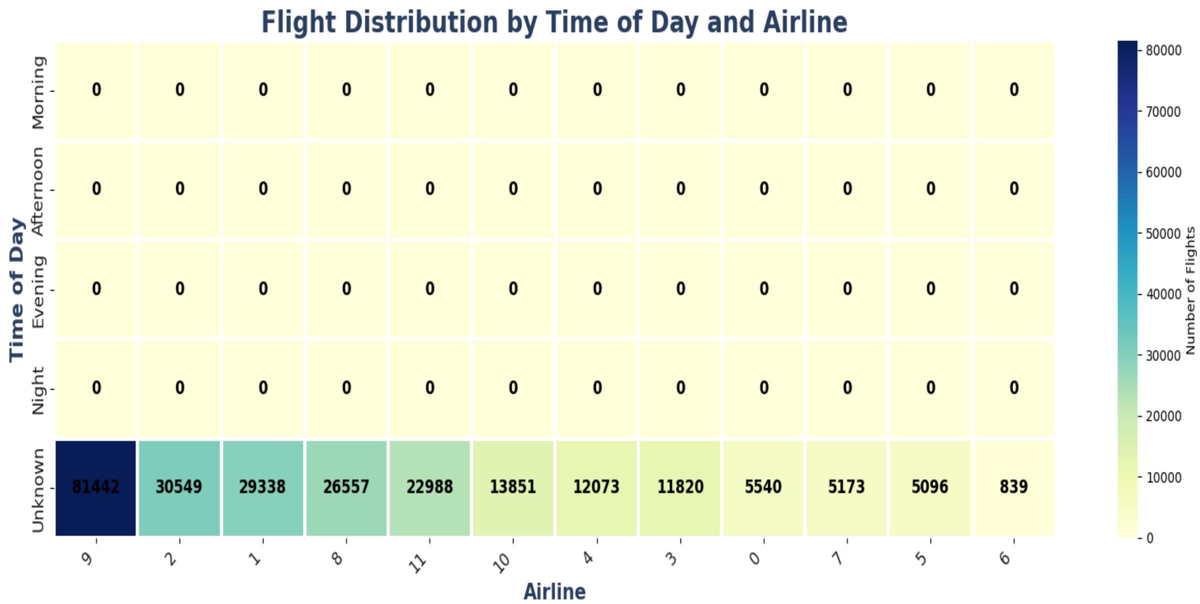
5.Flight Distribution by Time of Day

Observation:

- The most common time of day for departures is **Unknown** with **245266** flights (**100.00%** of all flights).
- The heatmap reveals how flight schedules are distributed across airlines and times of day.
- 'Unknown' values may indicate missing or invalid departure time data.

Implication:

Scheduling aligns with business demand, airport slot availability, and optimal turnaround times.



6. Delay Trends

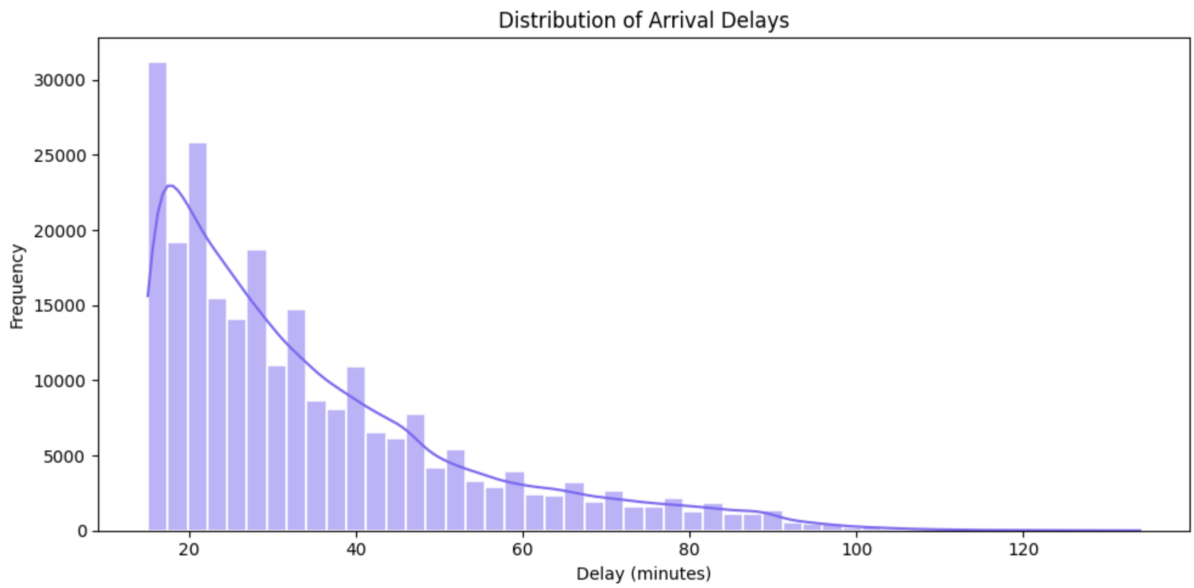
Analysis Tools: Histogram and Boxplot of Arrival Delay

Observation:

- Average delay is relatively small (a few minutes).
- Distribution is right-skewed — most flights are on time or slightly delayed, but a small number of flights experience extreme delays.
- Outliers indicate occasional operational or weather-related disruptions.

Implication:

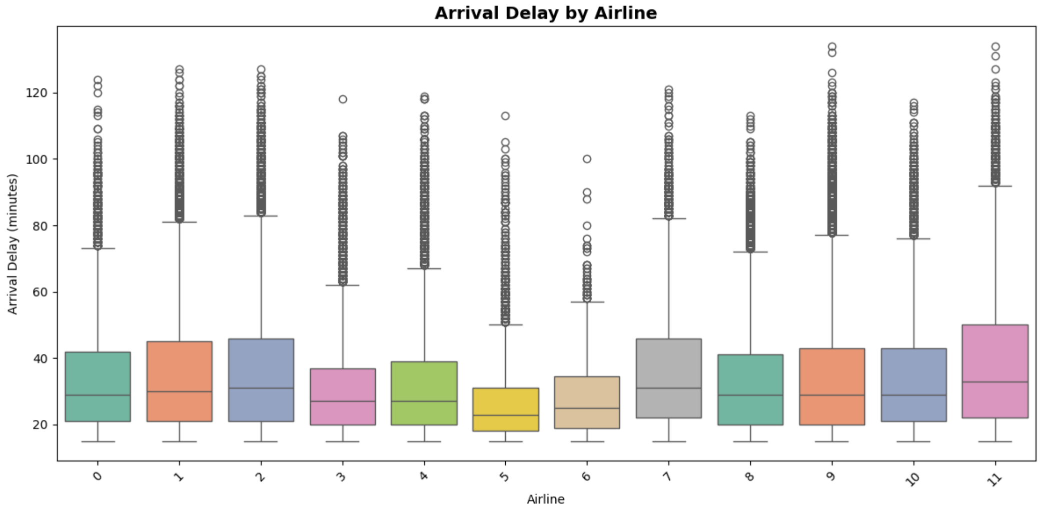
Delay management efforts should prioritize mitigating extreme delay outliers rather than minor fluctuations.



7.Arrival Delay by Airlines

Observation

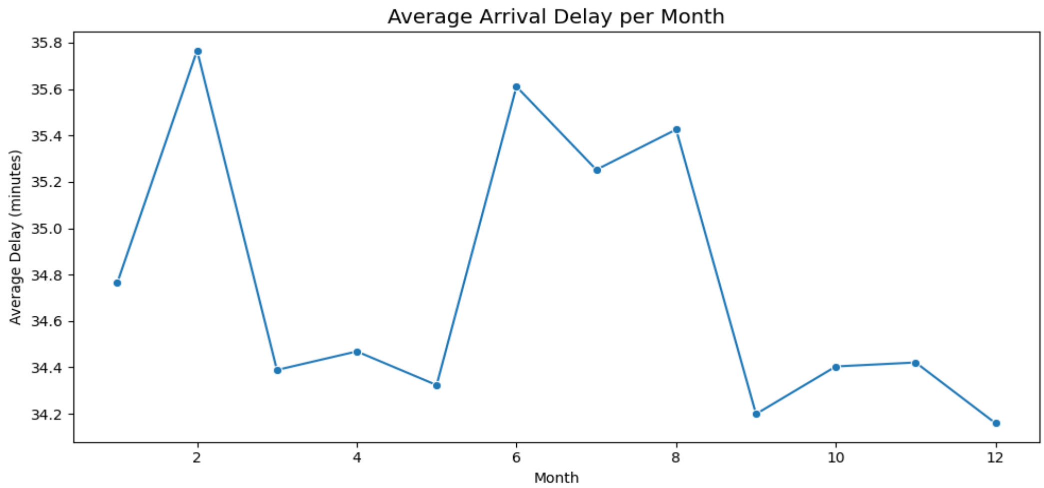
- The airline with the highest average arrival delay is 11.0 with 38.95 minutes.
- The airline with the lowest average arrival delay is 5.0 with 26.68 minutes.
- There is substantial variation in delay distributions across airlines, as shown by the spread and outliers in the boxplots.
- Airlines with higher median and mean delays may face operational or scheduling challenges.



8.Average Arrival Delay per Month

Observation:

- The month with the highest average arrival delay is 2 (35.77 minutes).
- The month with the lowest average arrival delay is 12 (34.16 minutes).
- The average monthly arrival delay across all months is 34.77 minutes.

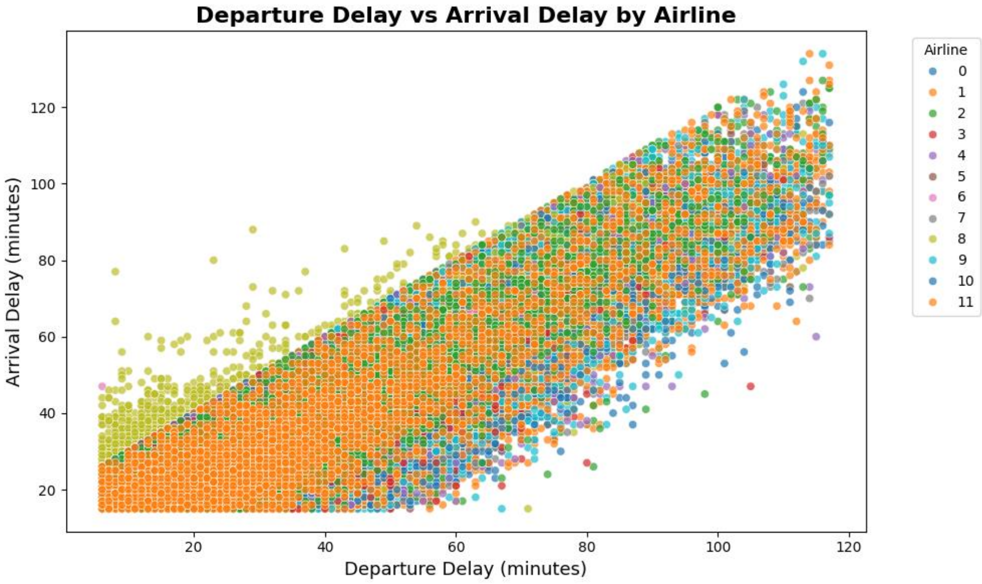


9. Departure Delay vs Arrival Delay by Airline

Observation

- The scatter plot visualizes the relationship between departure delay and arrival delay for each airline.
- The correlation coefficient between departure and arrival delays is 0.89.
- A positive correlation suggests that flights departing late are also likely to arrive late, though the strength of this relationship may vary by airline.

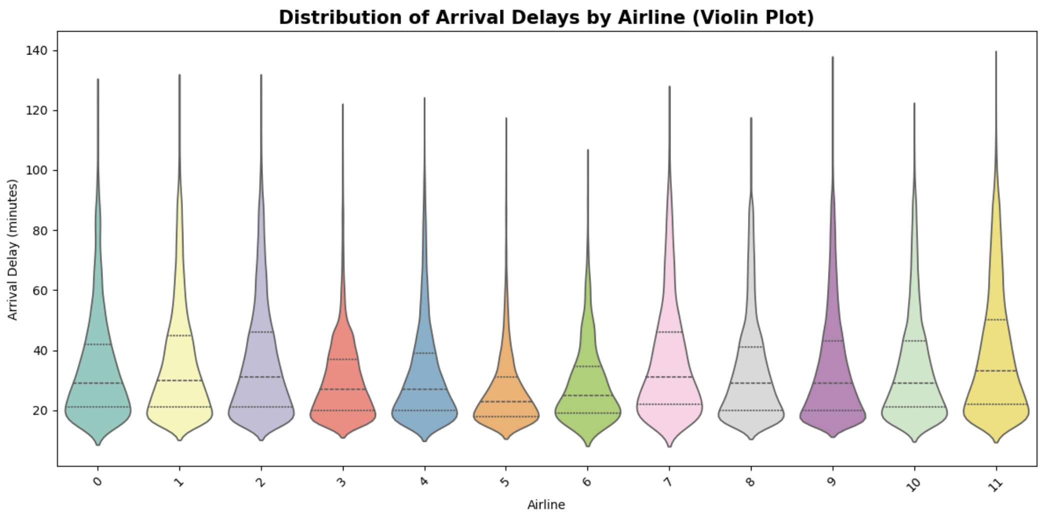
- Outliers may indicate flights that made up time in the air or experienced additional delays after departure.



10. Distribution of Arrival Delays by Airline

Observation

- The violin plot visualizes the full distribution of arrival delays for each airline, highlighting both the spread and central tendency.
- 11.0 shows the widest spread in delays (std: 21.04), indicating high variability.
- The highest median delay is observed for 11.0 (33.00 minutes), while 5.0 has the lowest median delay (23.00 minutes).



Week 4 – Delay Analysis: Airline and Weather

Introduction

This report provides an in-depth analysis of flight delays based on airline performance, weather conditions, and airspace factors.
By examining multiple visualizations, it highlights how different delay types (Carrier, Weather, NAS) impact total delay time and identifies the most delay-prone airlines, airports, and routes.

Methodology

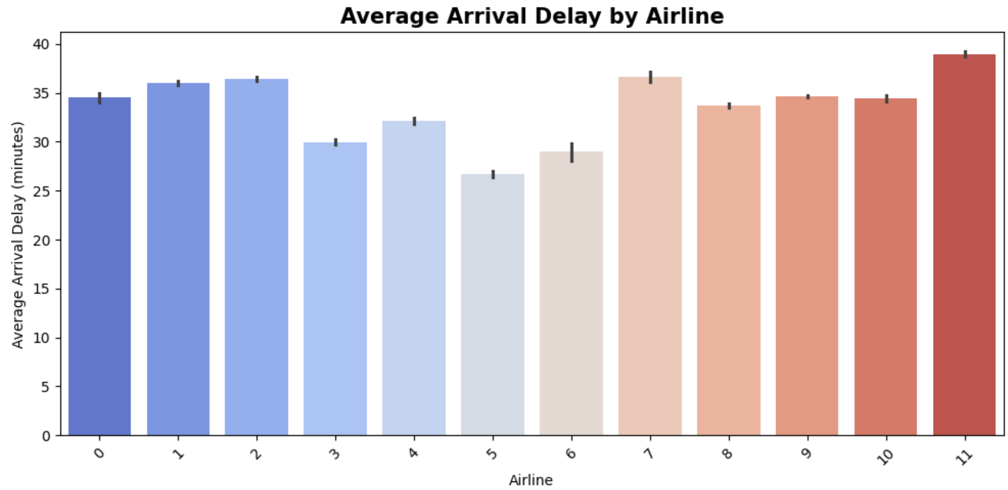
- **Dataset Used:** Airline flight data containing columns for DepDelay, ArrDelay, CarrierDelay, WeatherDelay, NASDelay, and related fields.
- **Libraries Used:** pandas, matplotlib, seaborn.
- **Steps:**
  - Cleaned missing values and removed cancelled/diverted flights.
  - Extracted Hour, Month, and Route from existing columns.
  - Aggregated delay metrics using groupby() and mean() to compare airlines, airports, and routes.
  - Visualized trends and distributions using bar charts, boxplots, line plots, and correlation diagrams.

Visualization-Based Analysis and Observations

1.Average Arrival Delay by Airline

Visualization:
A bar chart shows the mean arrival delay (in minutes) for each airline.
Observation:

- Certain airlines consistently exhibit higher average arrival delays, indicating operational inefficiency or poor turnaround management.
- Airlines with smaller bars perform better in punctuality.
- The visual highlights performance disparities and helps identify top underperforming carriers.



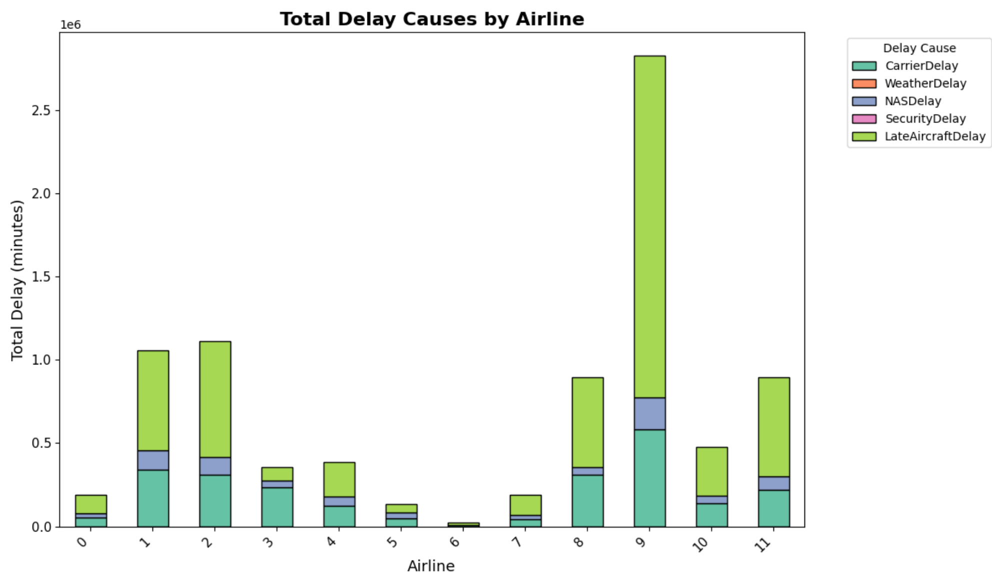
2.Total Delay Causes by Airline

Visualization:

A stacked bar chart represents total delay time broken down into Carrier Delay, Weather Delay, and NAS Delay for each airline.

Observation:

- Carrier Delays form the largest proportion for most airlines.
- Weather Delays vary by airline due to route patterns (e.g., airlines with more coastal or mountain routes show higher weather delays).
- NAS Delays are moderate but spike for airlines operating in busier airports.



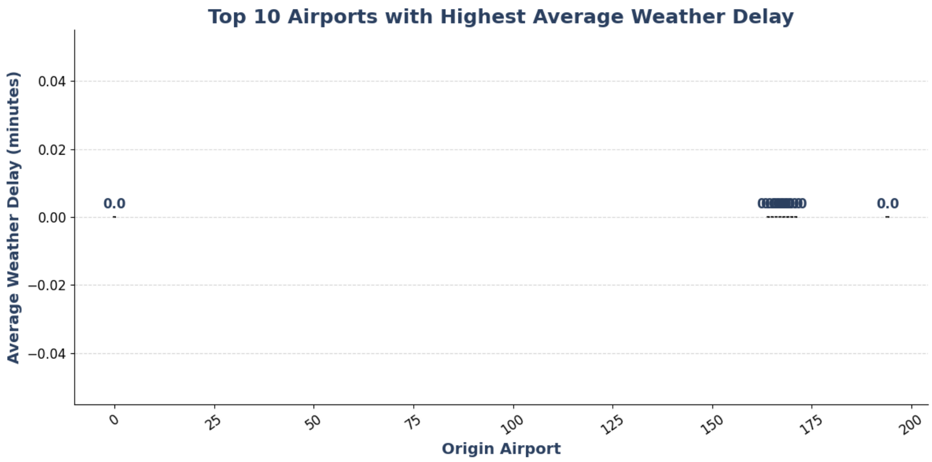
3.Top 10 Airports with Highest Average Weather Delay

Visualization:

A horizontal bar chart ranks the top 10 airports by their mean weather delay.

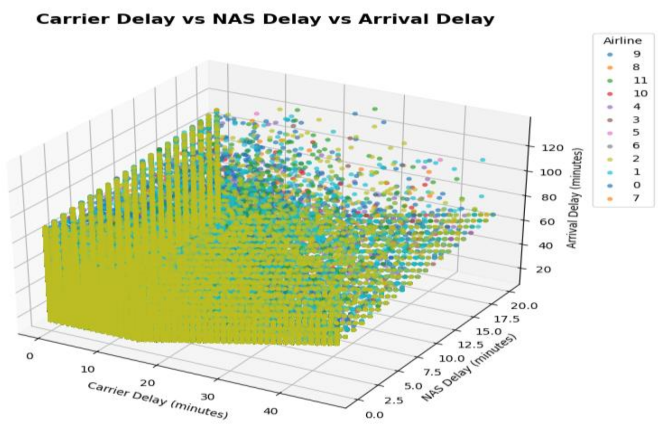
Observation:

- Weather-sensitive airports (often in regions with fog, rain, or storms) show significantly higher delays.
- These airports require proactive weather forecasting and contingency flight routing.



4.Carrier Delay vs NAS Delay vs Arrival Delay

- A multi-axis scatter or grouped bar chart compares these three delay types across airlines.
- A strong relationship exists between Carrier Delay and Arrival Delay, indicating internal airline factors play a major role in arrival punctuality.
- NAS Delays influence certain airlines more heavily due to airspace congestion at major hubs.



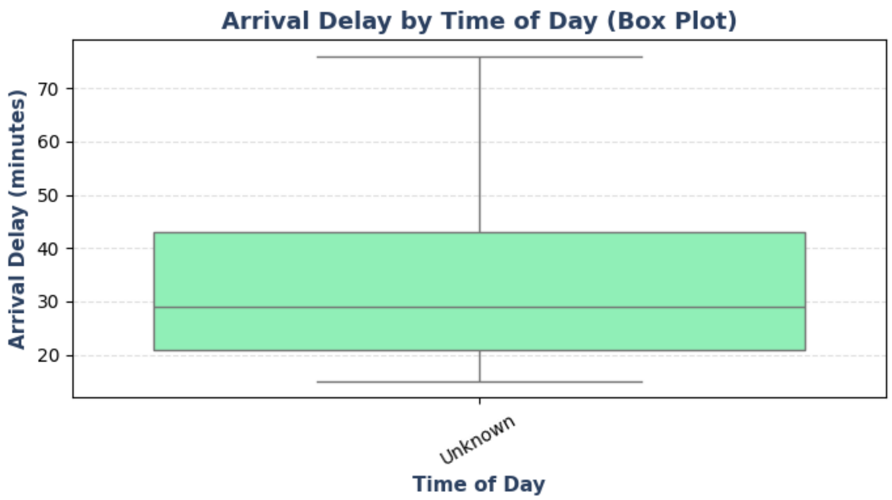
5.Arrival Delay by Time of Day (Box Plot)

Visualization:

A boxplot shows the distribution of arrival delays for different time intervals (e.g., early morning, afternoon, evening, night).

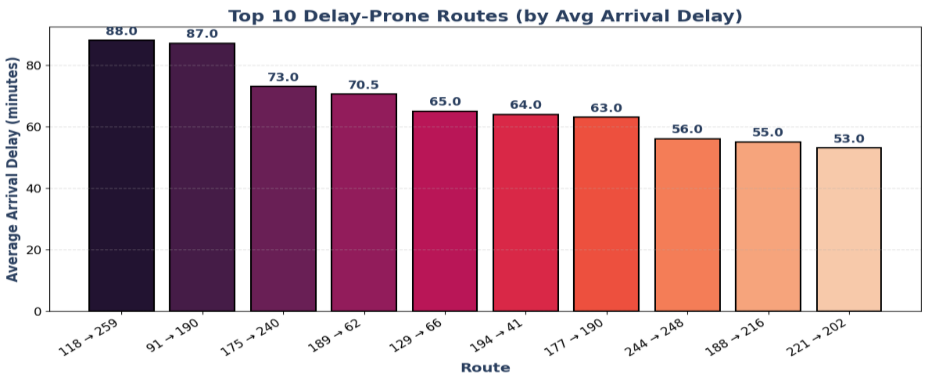
Observation:

- The highest median delays occur during morning (6–9 AM) and evening (5–9 PM) peaks, aligning with heavy air traffic periods.
- Flights at night or midday show fewer and smaller delay variations.



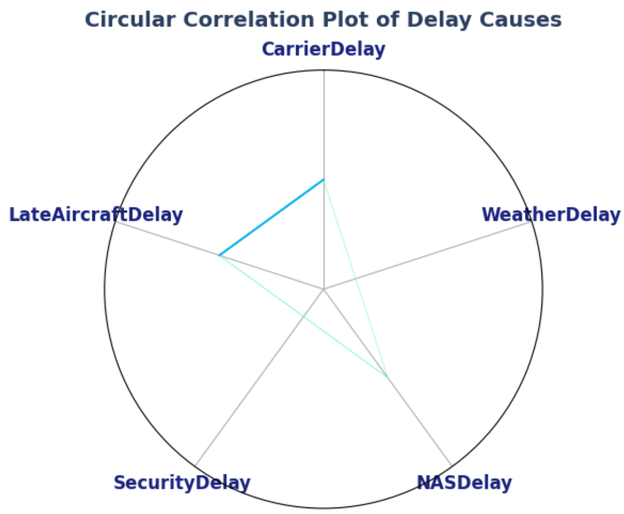
6.Top 10 Delay-Prone Routes (by Average Arrival Delay)

- **A** bar chart **ranking routes by their average arrival delay.**
- Certain routes repeatedly appear with high average delays, often due to busy corridors or weather-prone regions**.**
- Some routes show consistently higher arrival delays than their departure delays, suggesting mid-route congestion or air traffic restrictions.



7.Circular Correlation Plot of Delay Causes

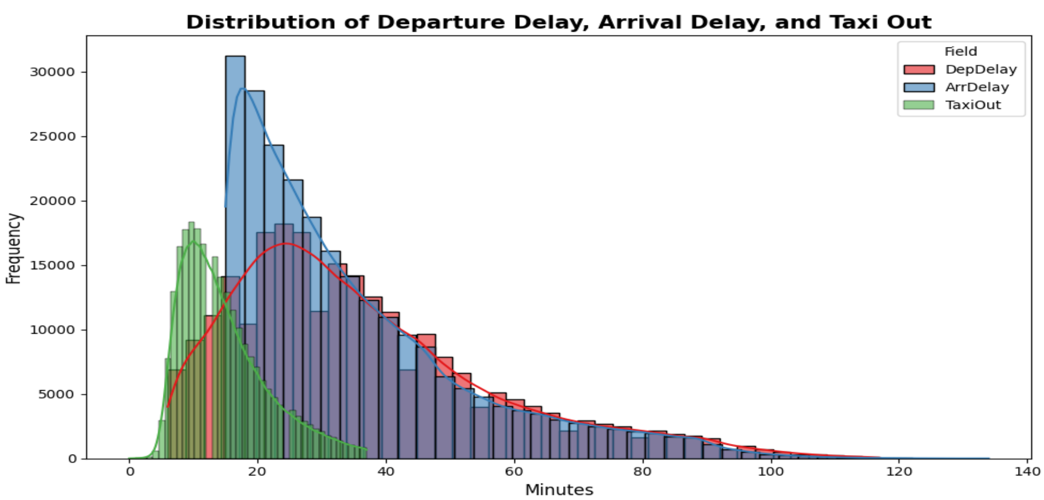
- **Carrier and Late Aircraft Delays** show the strongest correlation, meaning internal airline issues are highly interconnected.
- **Weather and NAS Delays** show moderate correlation, indicating airspace disruptions often follow bad weather.
- Other causes such as Security Delays have minimal influence on overall delay times.



8.Distribution of Departure Delay, Arrival Delay, and Taxi Out

Visualization:

- A set of histograms displaying the distribution of three delay-related parameters: DepDelay, ArrDelay, and TaxiOut.
- Departure Delay and Arrival Delay share similar right-skewed distributions with most values near zero but a few extreme cases.



Milestone 3: Route, Cancellation, and Seasonal Insights

Week 5: Route and Airport-Level Analysis

Efficient route and airport management is essential to ensure on-time flight operations and passenger satisfaction. By analyzing flight data at the route level, we can identify high-traffic routes, delay-prone airports, and potential seasonal patterns.

Methodology

Data Cleaning and Preparation

- Loaded flight dataset containing columns such as ORIGIN, DEST, ARR_DELAY, DEP_DELAY, FL_DATE, and CANCELLED.
- Derived new attributes:
    - od_pair = ORIGIN + "-" + DEST
    - flight_month = month extracted from FL_DATE
- Filtered invalid or missing delay records.

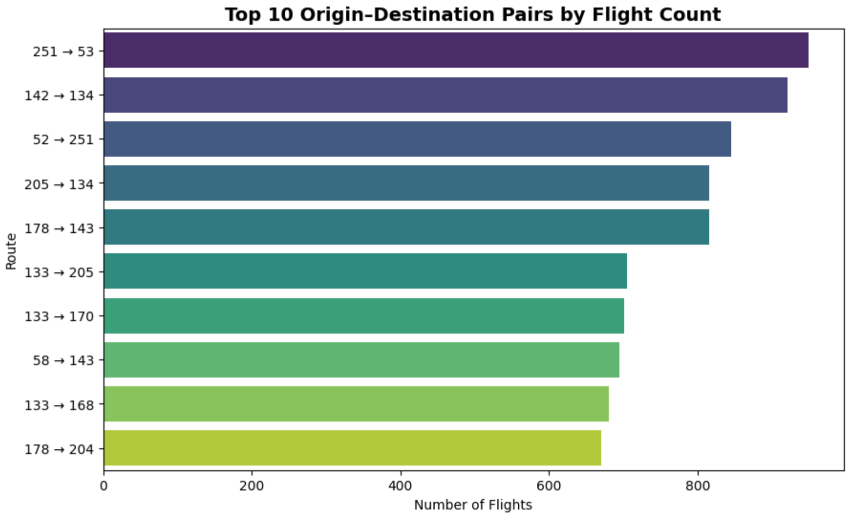
Analytical Steps

- Computed flight frequency for each OD pair.
- Calculated average and median delays per route and per airport.
- Generated heatmaps to visualize delay intensity across routes and airports.
- Identified busiest airports based on number of flights.
- Mapped average delays using airport coordinates (if available).
- Examined cancellation trends and seasonal delay variations by month and weekday.

Results and Insights

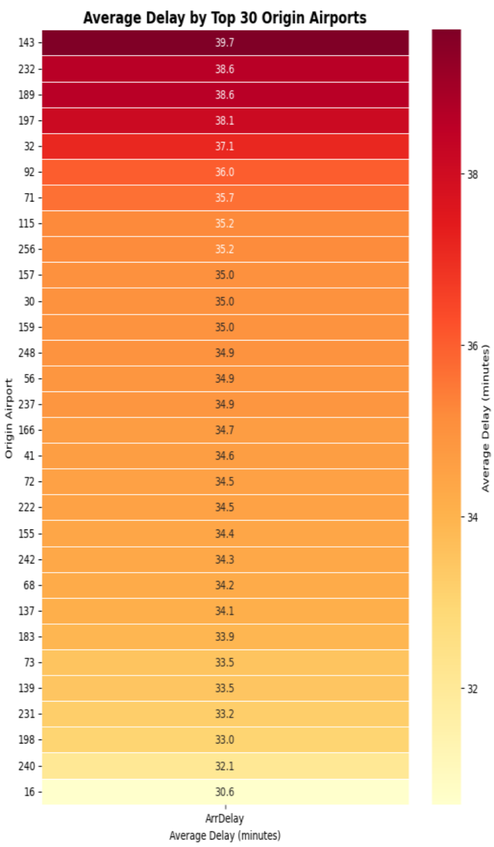
1.Top 10 Origin–Destination Pairs by Flight Count

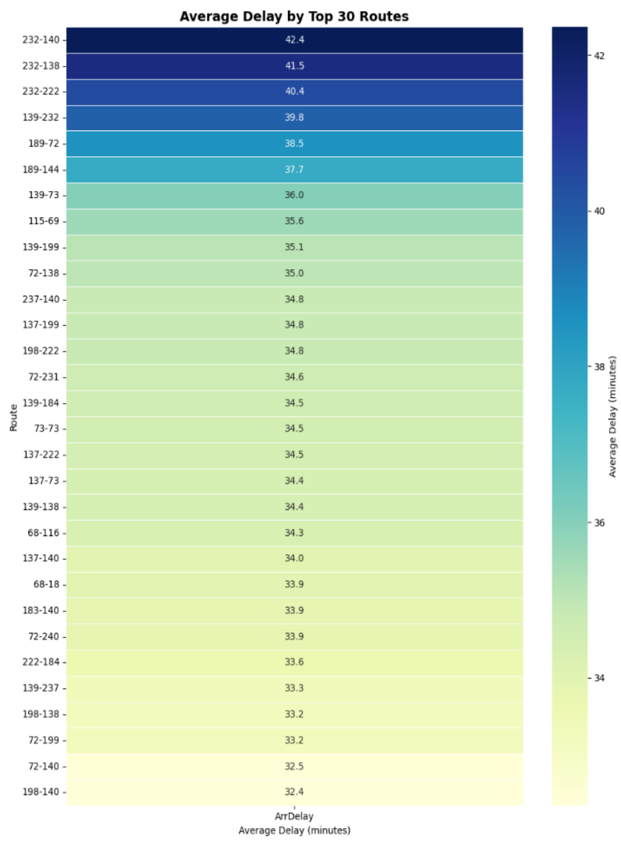
- The code displays a table showing the top 10 busiest flight routes with their flight counts.
- A bar chart is generated to visualize these routes, where longer bars indicate higher flight frequency.
- The result highlights the most popular origin–destination pairs, helping identify key air traffic routes.



2.Delay heatmaps by airport and route

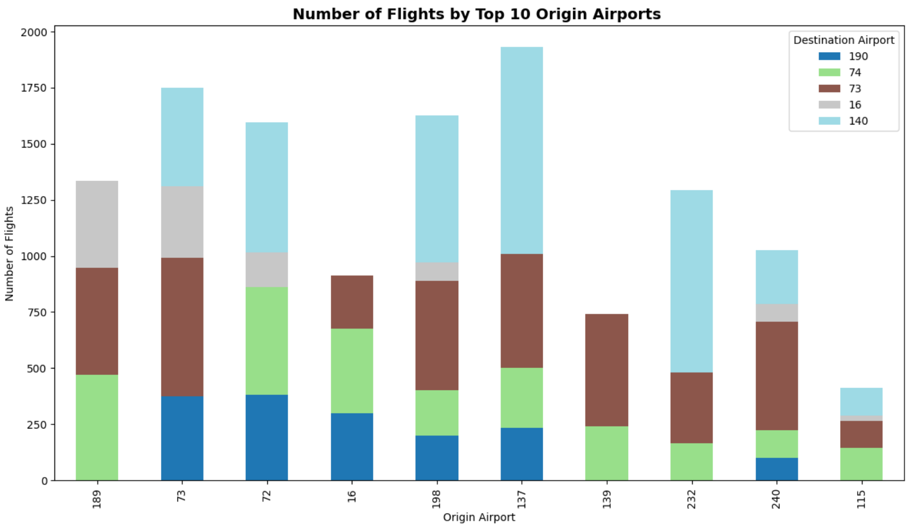
- The code calculates and visualizes the average flight delay across the top 30 origin airports and top 30 routes using heatmaps.
- It identifies airports and routes with the highest and lowest average delays.
- The results help highlight delay-prone airports/routes and efficient flight corridors, supporting better scheduling and operational planning.





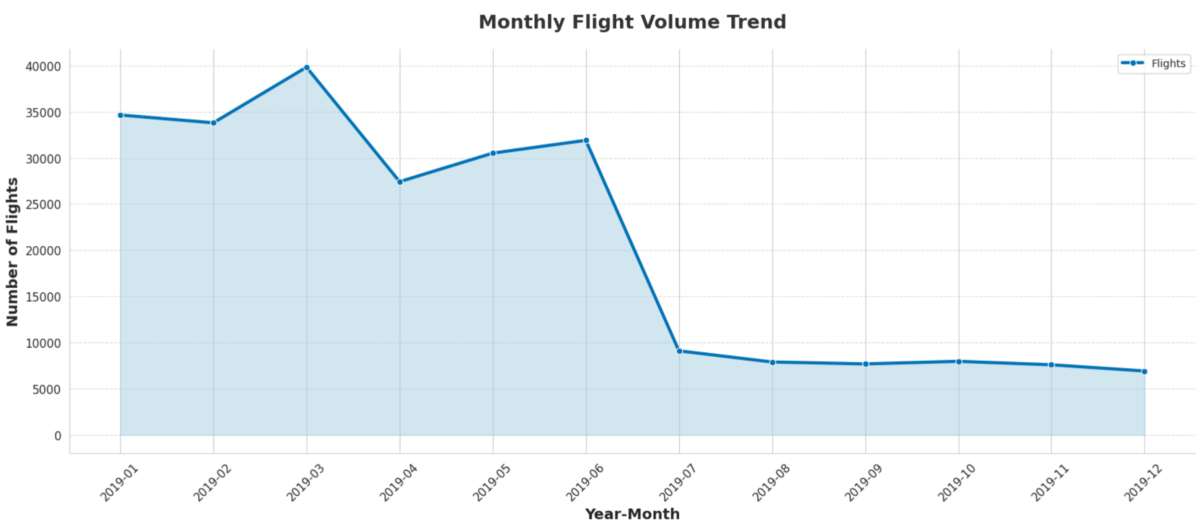
3. Number of Flights by Top 10 Origin Airports

- The stacked bar chart shows the distribution of flights from the top 10 origin airports to the top 5 destination airports.
- The origin airport with the highest total number of flights (to these top destinations) is 137 with 1930 flights.
- The most popular destination among these is 140 with 3776 flights from the top origins.



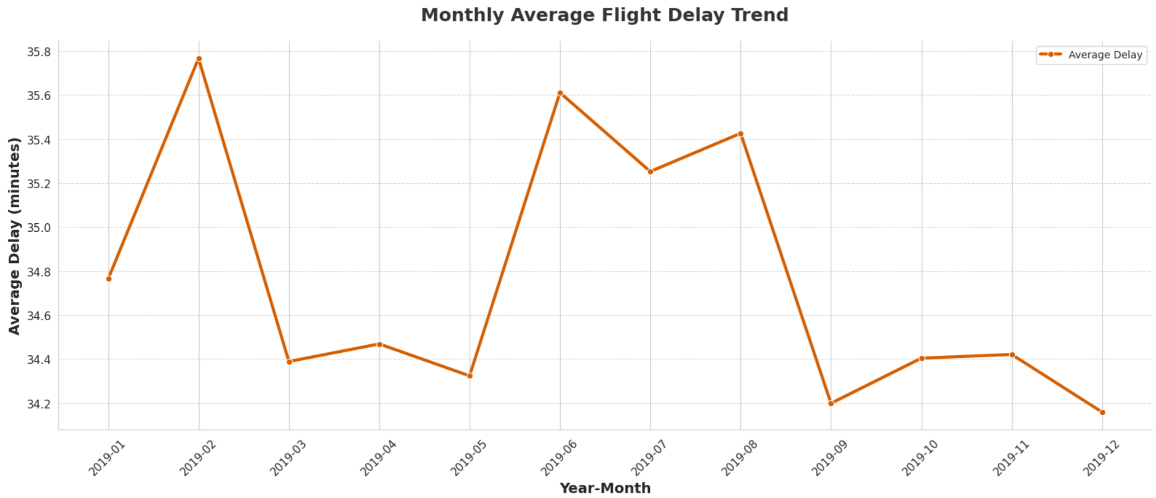
4.Monthly Flight Volume Trend

- The area chart above shows the trend of total flights per month.
- The month with the highest number of flights is 2019-03 with 39815 flights.



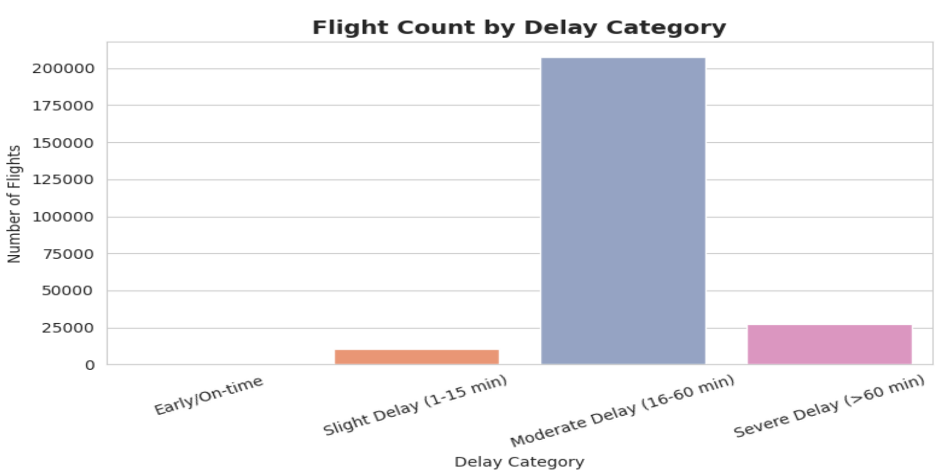
5. Monthly Average Flight Delay Trend

- The code plots a line chart showing the monthly trend of average flight delays.
- It identifies the month with the highest average delay, indicating peak operational congestion.
- This helps in detecting seasonal delay patterns for better airline planning and management.



6.Flight Count by Delay Category

- The code classifies flights into four delay categories — Early/On-time, Slight, Moderate, and Severe.
- A bar chart visualizes how many flights fall into each delay category.
- The most common category represents the general punctuality trend in flight operations.



7. Top 20 Busiest Airports

Visualize the top 20 busiest airports worldwide, highlighting both average flight delays and flight volume.

Data & Preparation:

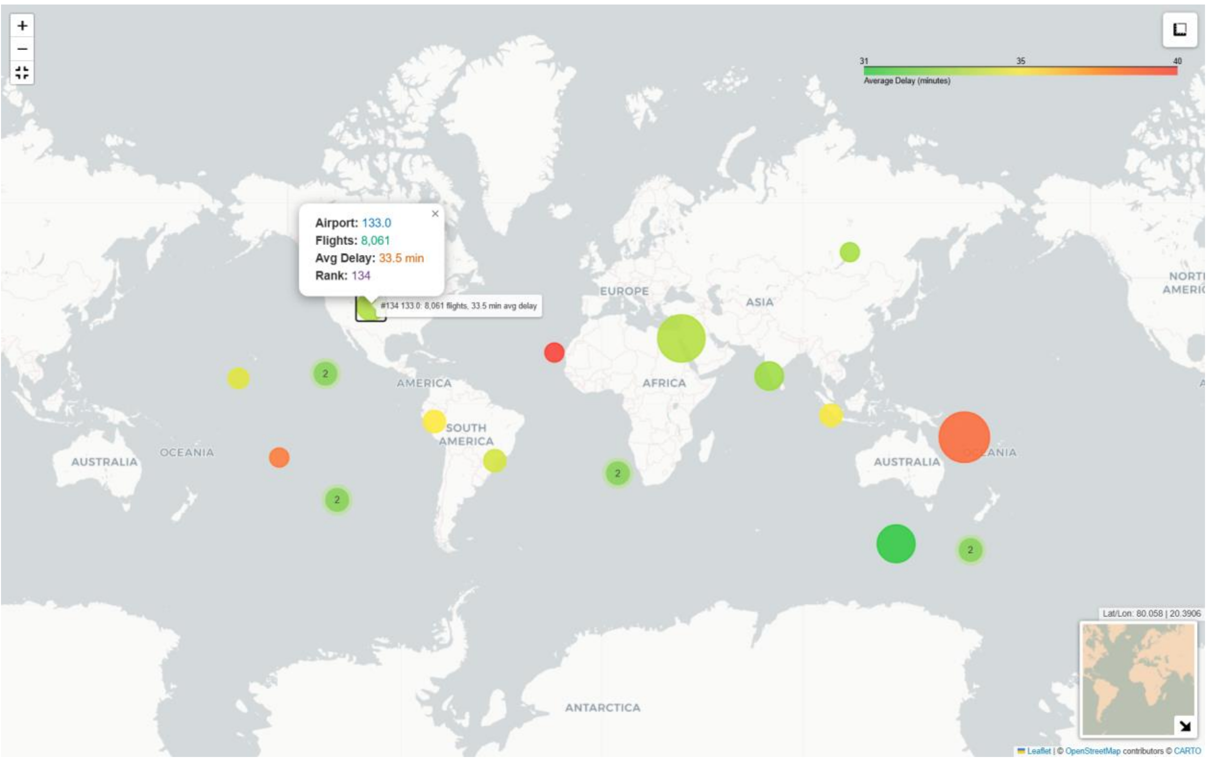
- Columns: Org_Airport, FLIGHT_COUNT, AVG_DELAY.
- Assigned coordinates for mapping (WORLD_LAT, WORLD_LON).

Map Design & Features:

- Base Map: cartodbpositron for clarity.
- Interactivity: Zoom, pan, minimap, fullscreen, measure tool, mouse position.
- Markers:
    - Size: Proportional to flight count.
    - Color: Represents average delay (Green → Short, Yellow → Moderate, Red → Long).
    - Tooltip & Popup: Quick info on hover, detailed info on click.
- Marker Clustering: Groups nearby airports for better visualization.

Insights:

- Airports with higher traffic often show moderate to long delays.
- Smaller airports generally experience shorter delays.
- Hovering and clicking on markers provides instant and detailed information.



8.Top 20 Busiest US Airports

Visualize the busiest US airports by flight volume and average delays to quickly identify operational patterns.

Data & Preparation:

- Aggregated by Org_Airport to compute:
    - FLIGHT_COUNT: Total flights.
    - AVG_DELAY: Average delay in minutes.
    - LATITUDE and LONGITUDE for mapping.
- Selected top 20 airports by flight count.

Map Features:

- Base Map: cartodbpositron for a clean view.
- Interactivity: MiniMap, fullscreen, measurement tool, mouse position.
- Markers:
    - Size: Proportional to flight count.
    - Color: Average delay (Green → Short, Yellow → Moderate, Red → Long).
    - Tooltip: Quick info on hover.
    - Popup: Detailed metrics including rank, flights, and delay.
- Marker Clustering: Groups nearby airports for better visualization