



# AIRFLY INSIGHTS

Data Visualization and Analysis of Airline Operations

Presented by  
Krithika M

## Overview



- Project Statement
- Data Acquisition and Understanding
- Data Cleaning and Feature Engineering
- Univariate and Bivariate Analysis
- Delay Cause Analysis
- Route and Airport-Level Exploration
- Cancellation and Seasonal Trends
- Final Dashboard
- Challenges
- Future Scope
- Conclusion



## Project Statement

The objective of this project is to analyze large-scale airline flight data to uncover operational trends, delay patterns, and cancellation reasons using data visualization techniques. It involves cleaning the data, creating meaningful features, and exploring patterns across airlines, airports, and seasons.

The goal is to help understand **airline and airport-level performance** and contribute to **actionable insights** using visual analysis

## Data Acquisition and Understanding



### Dataset Source (Kaggle)

- Chosen for its large, reliable, and well-organized flight data.
- Contains 484,552 rows x 30 columns,

### Why This Dataset?

- Comprehensive coverage of U.S. domestic flights.
- Includes detailed delay categories and cancellation reasons.
- Suitable for trend analysis, operational insights, and visualization.

### Initial Data Exploration

- **Defined goals, KPIs, and workflow** - Set the project direction and identified key metrics to focus on.
- **Loaded data using pandas** - Imported the dataset into Python for easy handling and analysis.
- **Checked schema, data types, size, and null values** - Reviewed the structure, column types, dataset size, and missing data.
- **Performed sampling and memory optimization** - Reduced memory load and created smaller samples for faster processing.



## Data Cleaning and Feature Engineering

- **Handled nulls** – Cleaned missing values in delay and cancellation columns.
- **Created features** – Added Month, Day of Week, Hour, and Route for better analysis.
- **Formatted datetime** – Converted time columns into proper datetime format.
- **Saved processed data** – Stored the cleaned dataset for quicker future use.

```
# Handle Nulls in Delay and Cancellation Columns
```

```
import numpy as np
```

```
# Fill delay columns with 0
```

```
delay_cols = ['ArrDelay', 'DepDelay', 'CarrierDelay', 'WeatherDelay',  
             'NASDelay', 'SecurityDelay', 'LateAircraftDelay']  
df[delay_cols] = df[delay_cols].fillna(0)
```

```
# Standardize cancellation columns
```

```
df['Cancelled'] = df['Cancelled'].map({'Y': 1, 'N': 0})  
df['Diverted'] = df['Diverted'].fillna(0)  
df['CancellationCode'] = df['CancellationCode'].fillna('None')
```

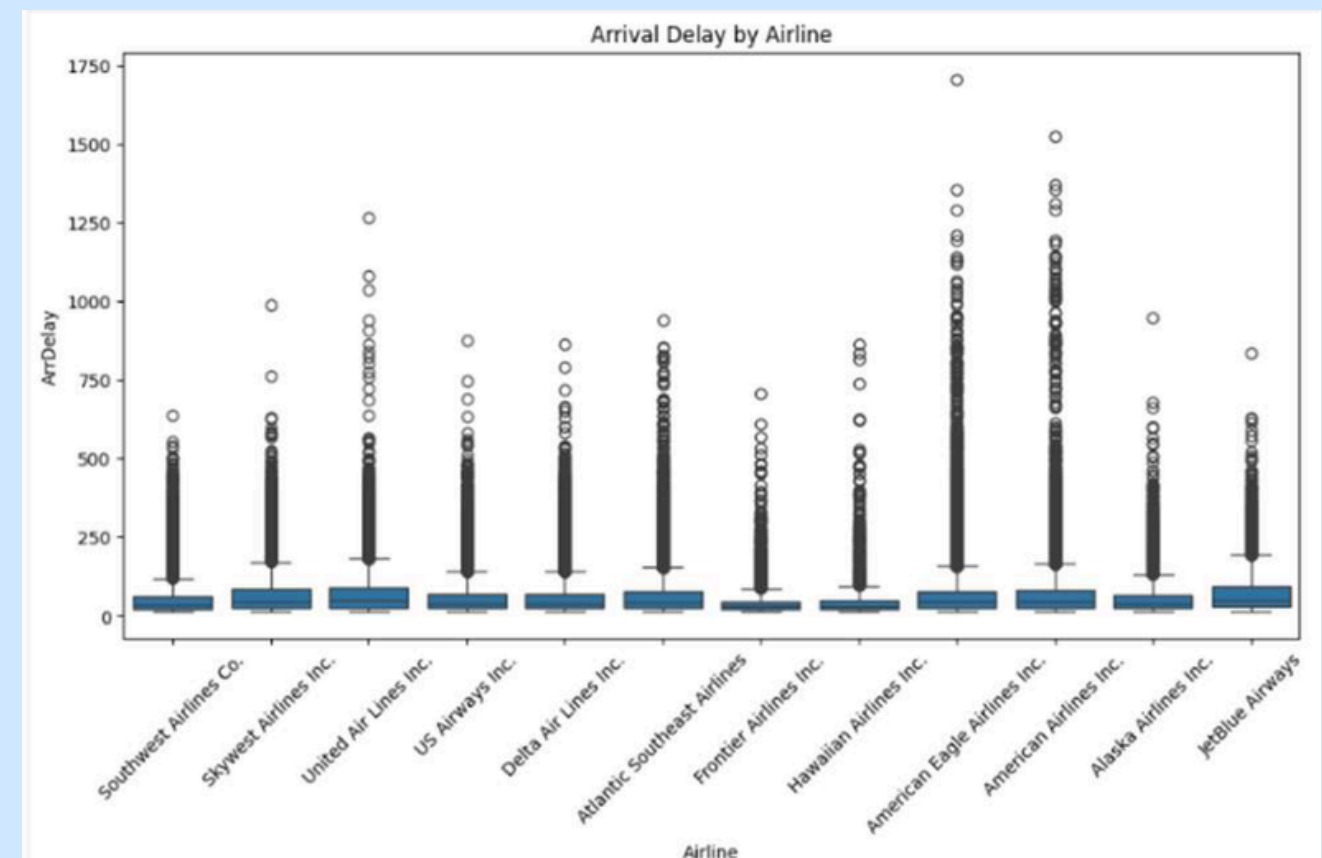
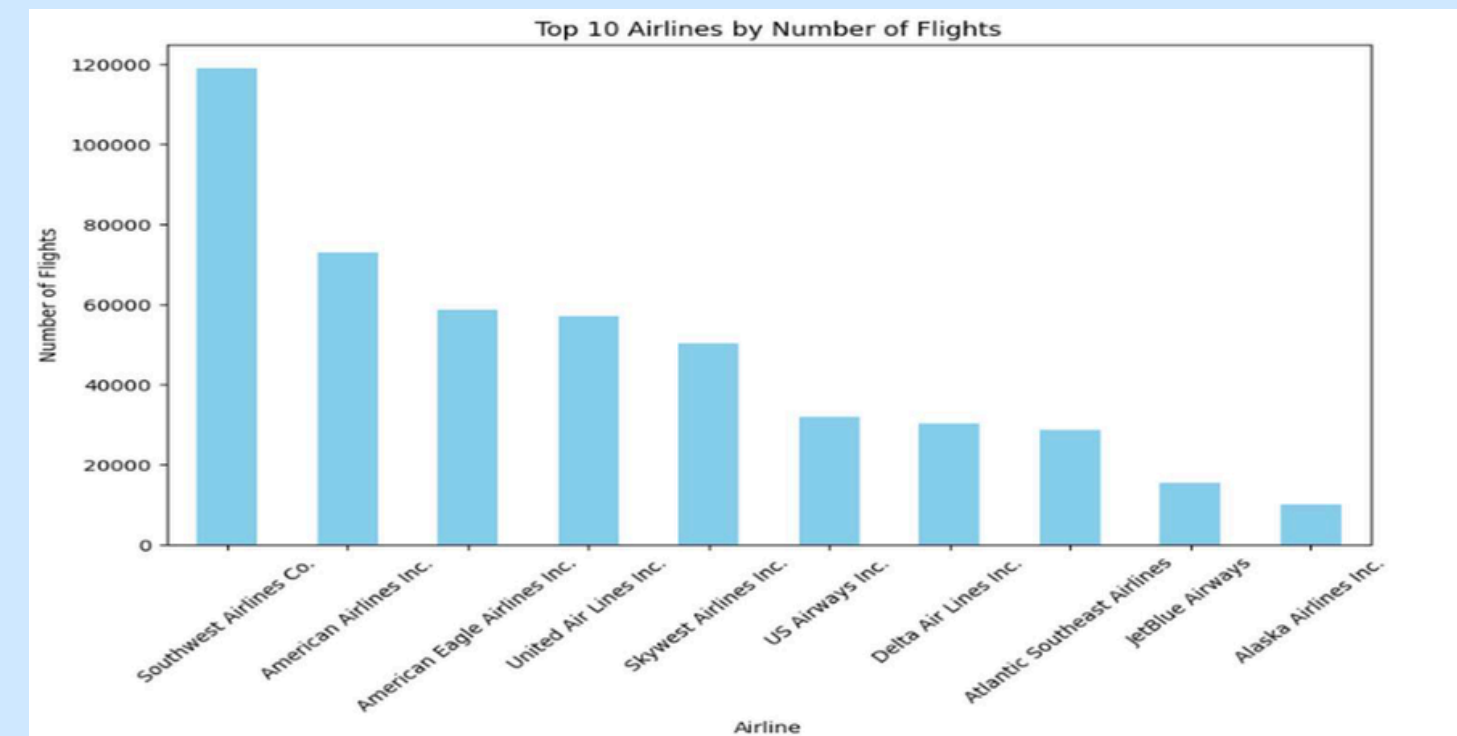
	Date	DepTime	ArrTime	CRSArrTime
0	2019-01-03	18:29:00	19:59:00	19:25:00
1	2019-01-03	19:37:00	20:37:00	19:40:00
2	2019-01-03	16:44:00	18:45:00	17:25:00
3	2019-01-03	14:52:00	16:40:00	16:25:00
4	2019-01-03	13:23:00	15:26:00	15:10:00

	Date	Month	DayOfWeekNum	DepHour	Route
0	2019-01-03	1	3	18.0	IND-BWI
1	2019-01-03	1	3	19.0	IND-LAS
2	2019-01-03	1	3	16.0	IND-MCO
3	2019-01-03	1	3	14.0	IND-PHX
4	2019-01-03	1	3	13.0	IND-TPA

# Univariate and Bivariate Analysis

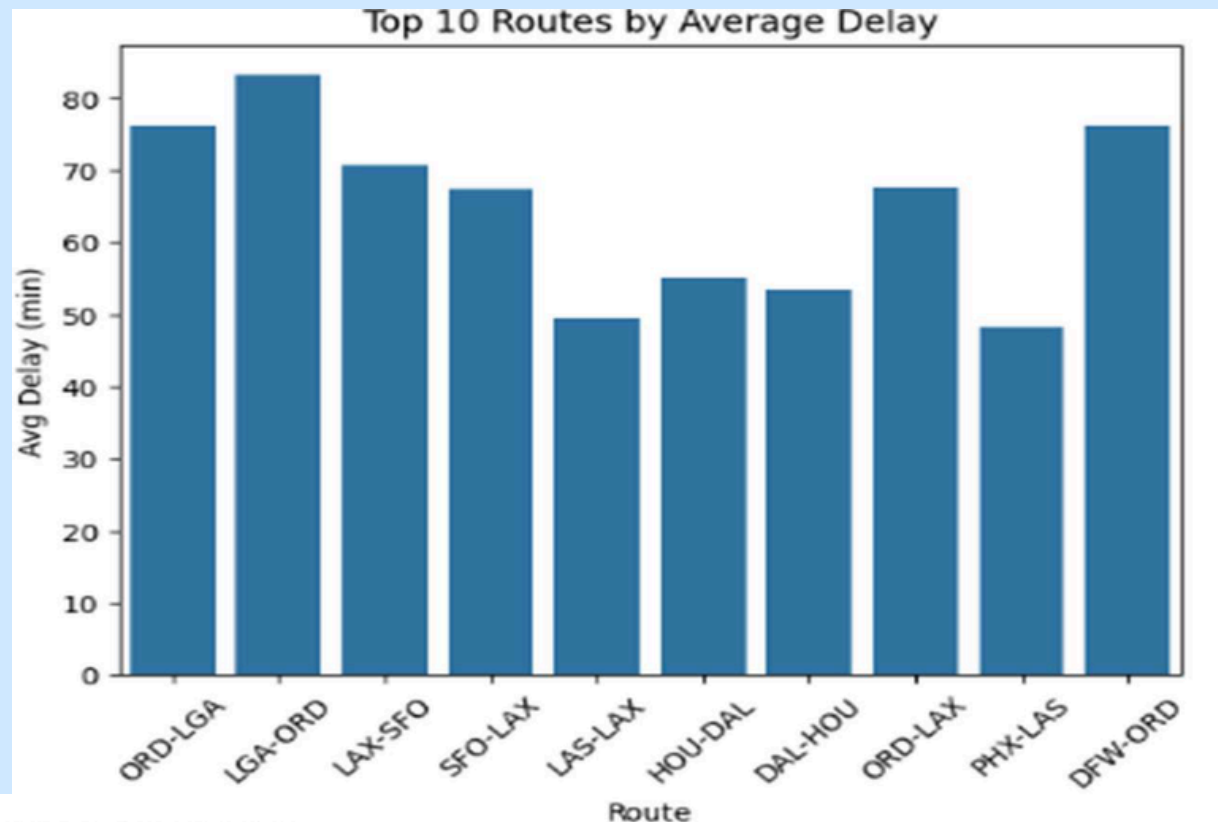


- **Univariate Analysis** – Examined individual columns like airlines, months, departure hours, and delay types to understand basic distributions and patterns.
- **Bivariate Analysis** – Compared two variables (e.g., Airline vs Delay, Airport vs Cancellations, Time vs Delay) to identify relationships and trends.
- **Visualizations Used** – Bar charts, line charts, histograms, and boxplots to clearly show patterns.
- **Outcome** – Identified peak delay times, top delay-prone airlines, busy routes, and seasonal flight patterns.





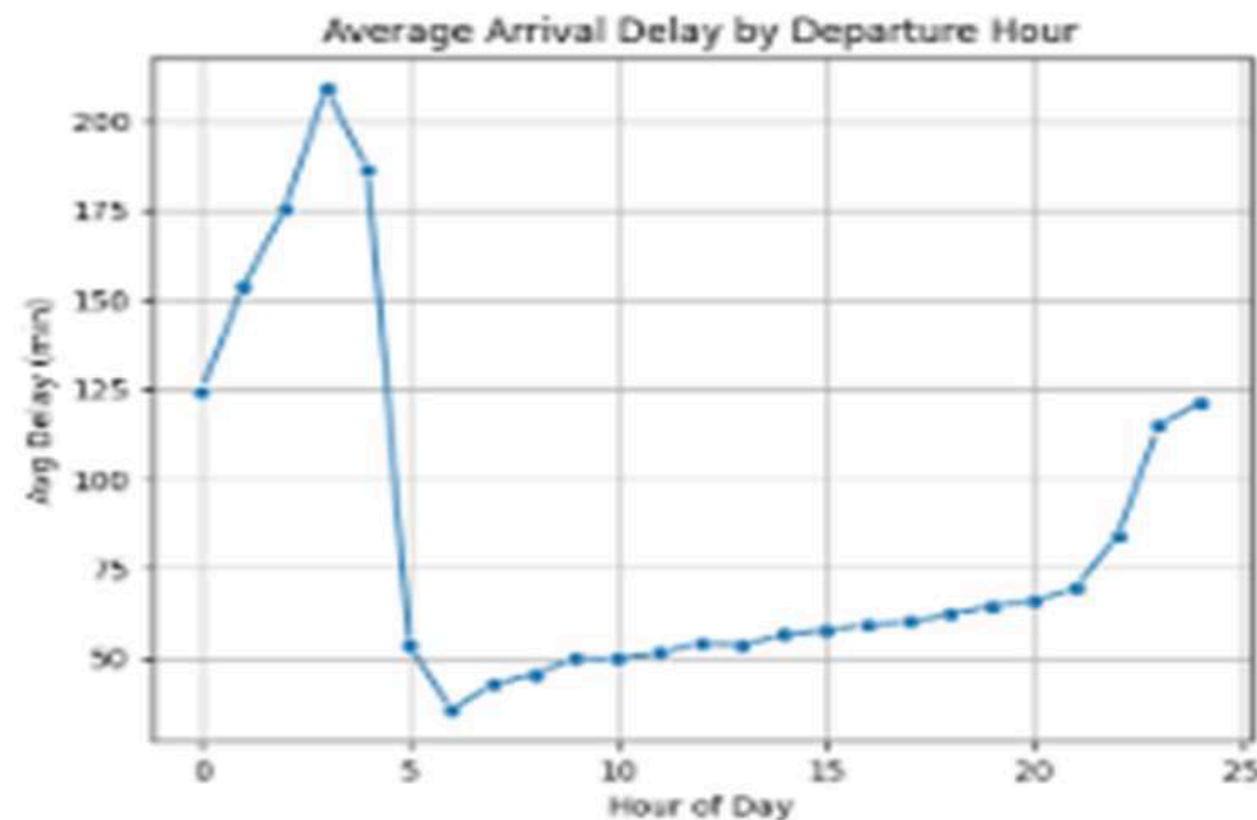
# Delay Cause Analysis



Studying the reasons and patterns behind flight delays to understand which factors impact on-time performance.

What's Done –

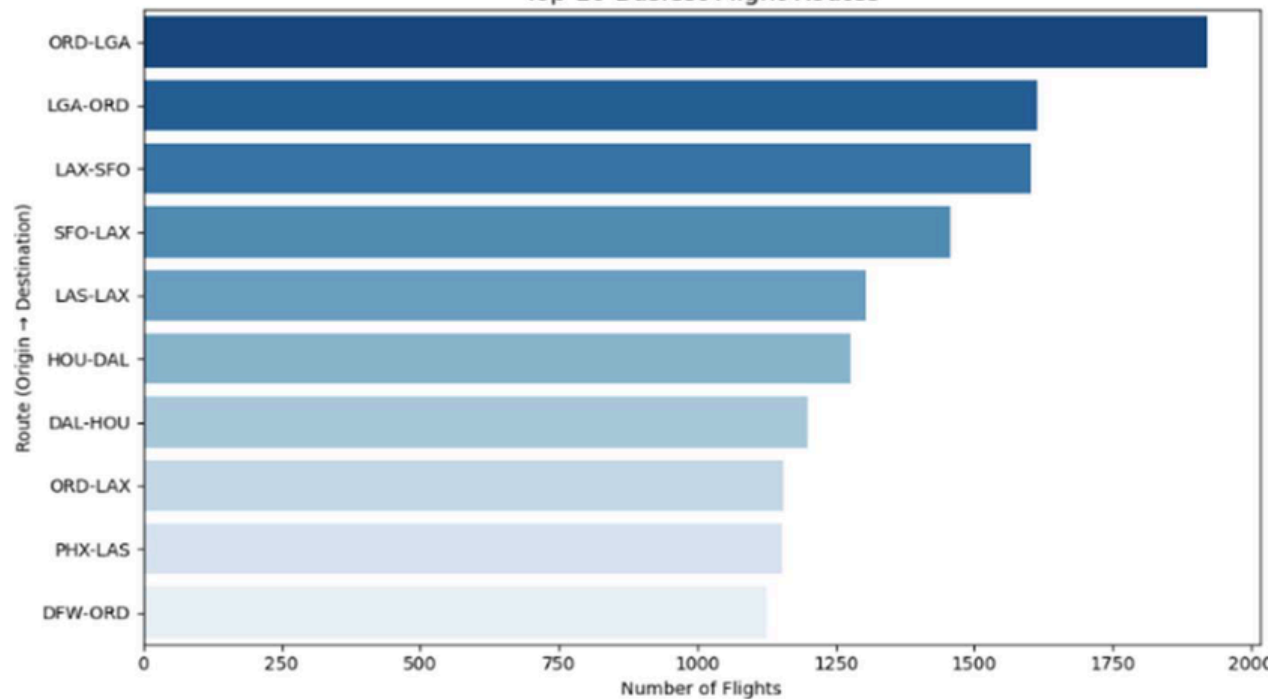
- Analyzed major **delay categories**: Carrier, Weather, NAS, and Late Aircraft.
- Compared delay durations across airlines, airports, and **time of day**.
- Identified which **airlines** and **routes** experience the **highest delays**.
- Used visualizations to highlight peak delay hours and major delay contributors.



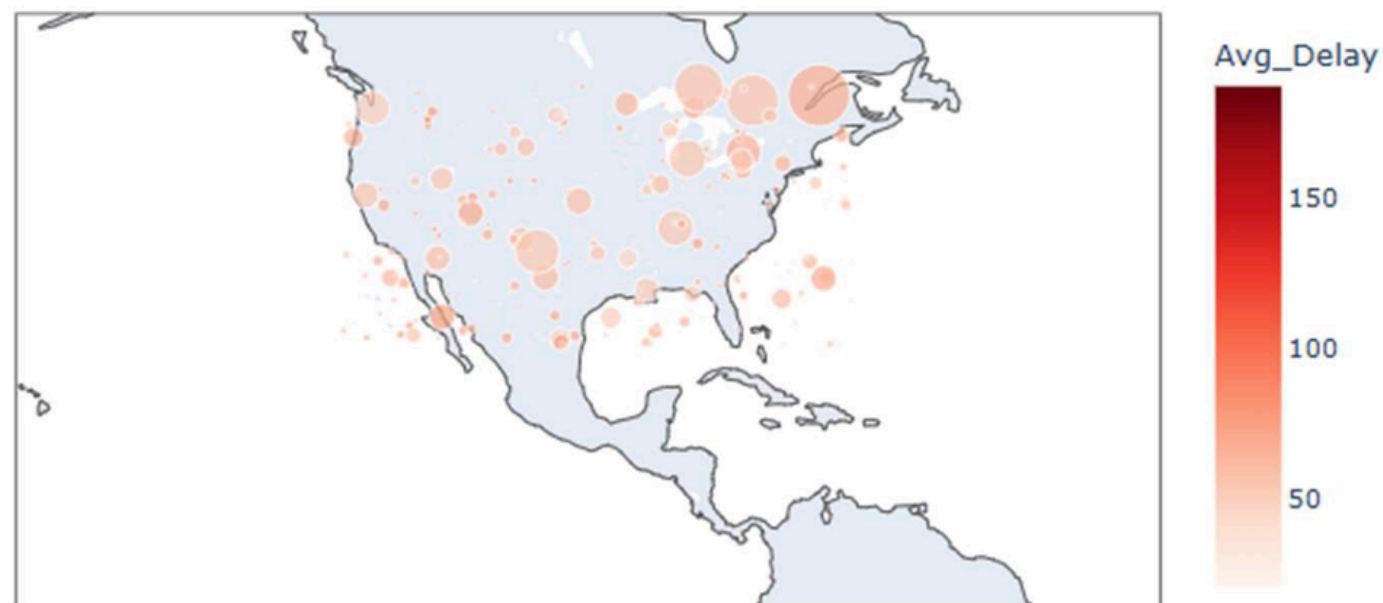
# Route and Airport-Level Exploration



Top 10 Busiest Flight Routes



Busiest Airports and Average Delays (Simulated Coordinates)



Studying how different routes and airports perform in terms of traffic, delays, and cancellations.

What's Done –

- Identified top origin–destination routes based on flight volume.
- Analyzed **airport-wise delays** to find the most **delay-prone** and **busiest airports**.
- Created route-level visuals to compare average delays across major flight paths.
- Checked how airport congestion and route distance impact **flight performance**.



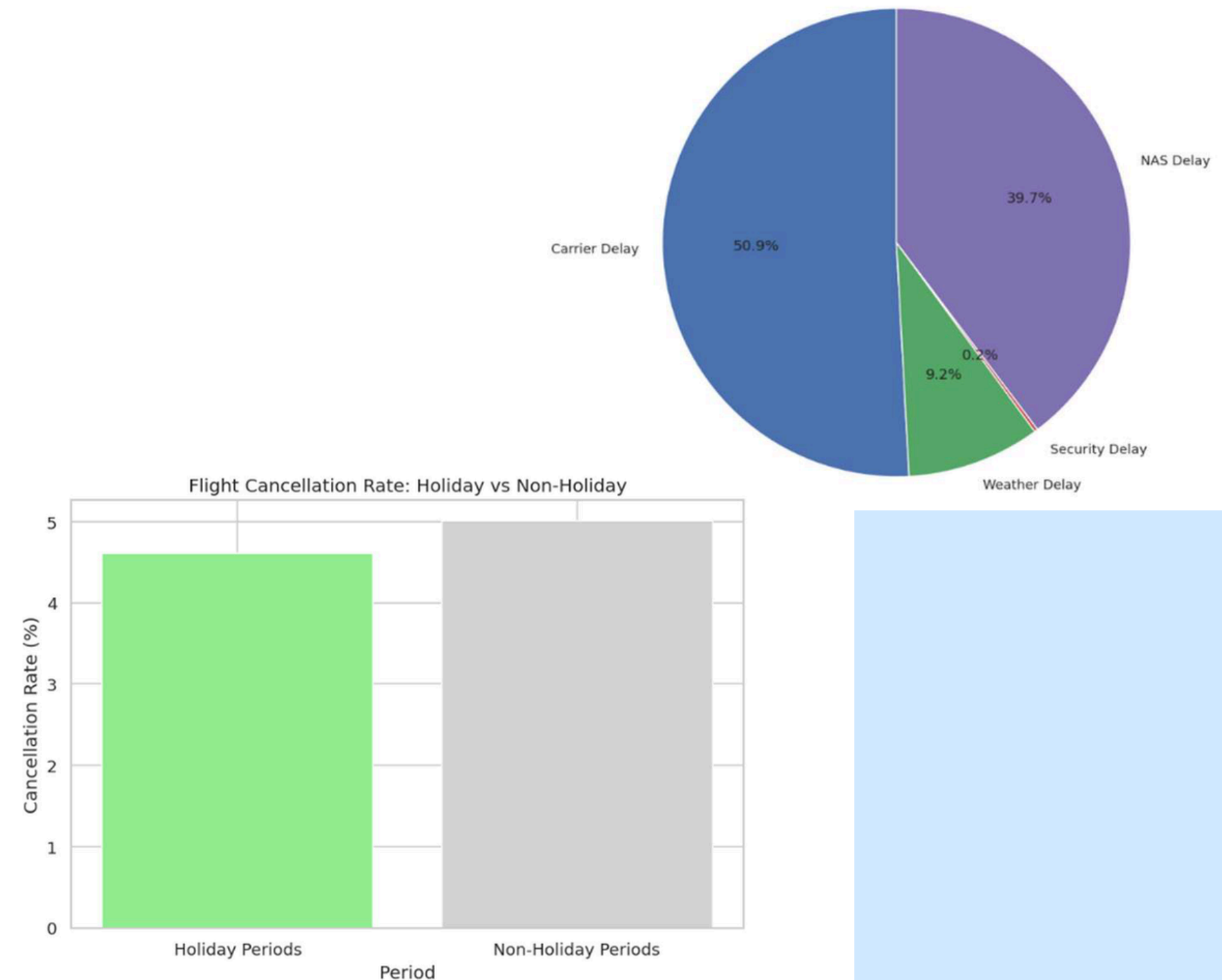


# Cancellation and Seasonal Trends

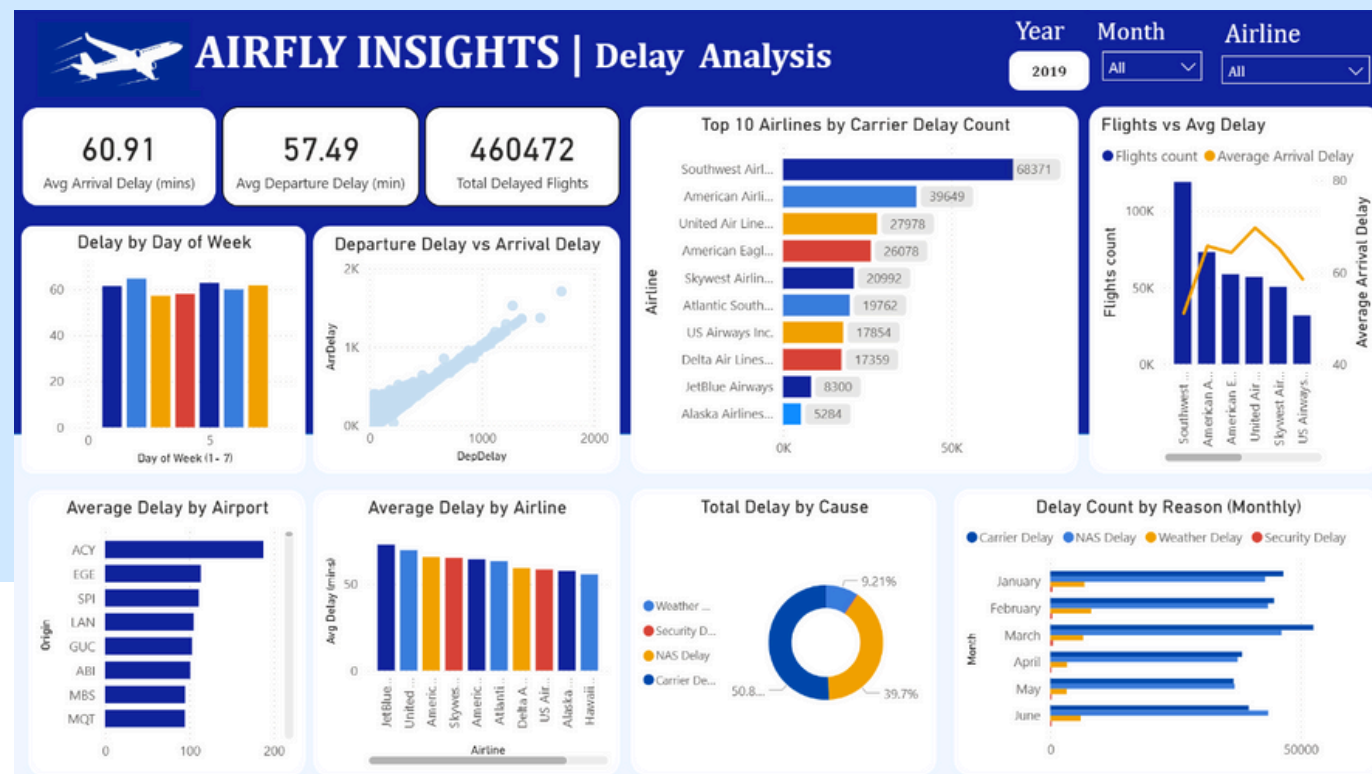
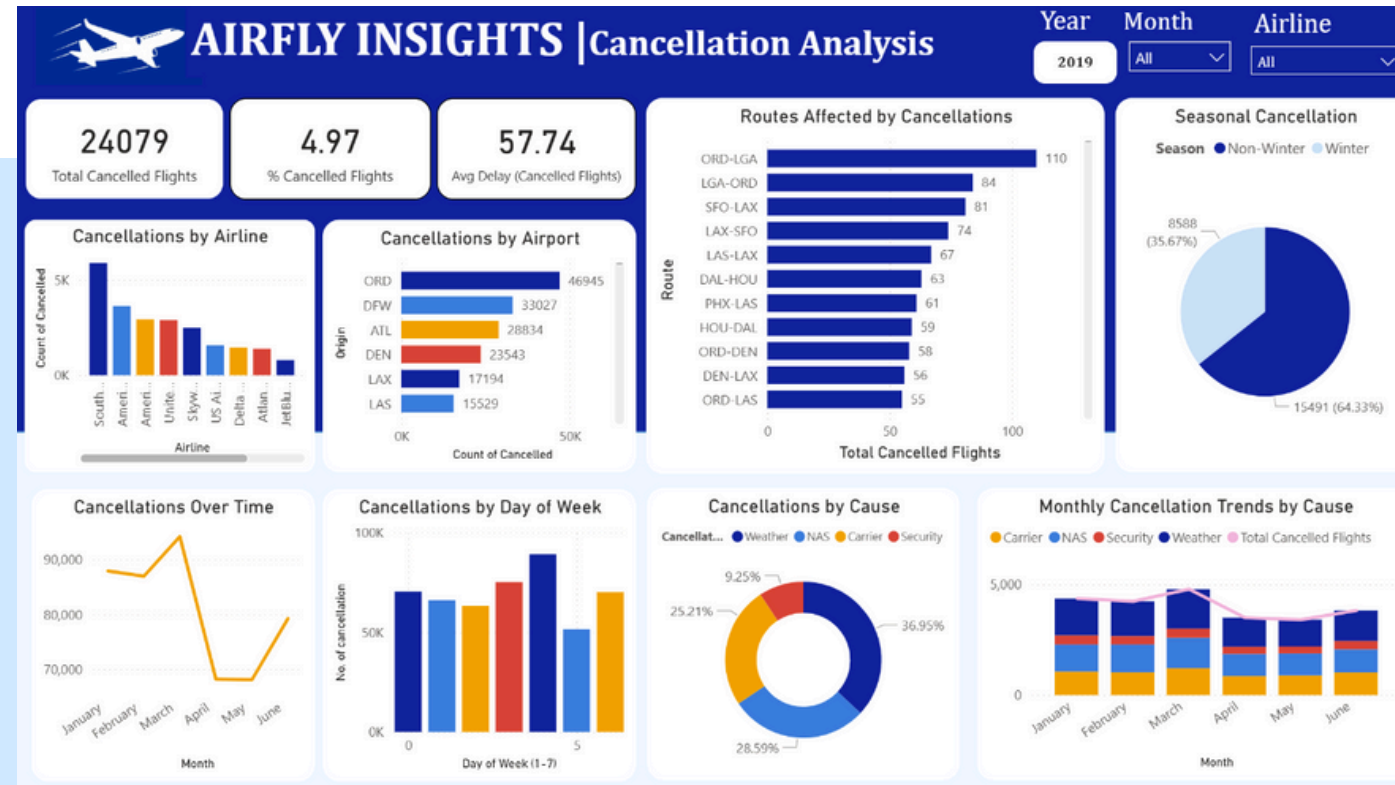
Understanding why flights get cancelled and how cancellation patterns change across months, airlines, and airports.

What's Done –

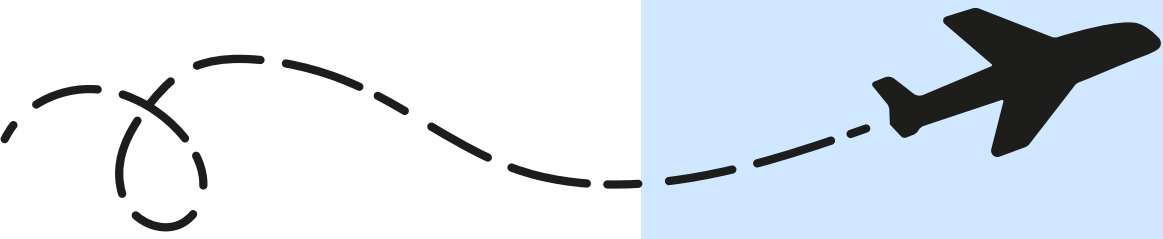
- Analyzed monthly and seasonal cancellation patterns to see when **cancellations peak**.
- Studied **cancellation reasons** such as Carrier, Weather, NAS, and Security.
- **Mapped cancellation codes** (A, B, C, D, N) to their correct reasons and assigned them properly for analysis.
- Compared cancellations across **airlines** and **routes** to find which ones face the most disruptions.



# Final Dashboard



- **Import Data:** Load the cleaned dataset into Power BI.
- **Clean & Transform:** Use Power Query to fix data types and create new fields (Month, Route, Delay Type).
- **Model Data:** Build relationships and calculated columns if needed.
- **Create Visuals:** Add charts like bar, line, map, and pie for key metrics.
- **Add KPIs:** Show total flights, % cancelled, and average delay.
- **Apply Filters:** Use slicers for Airline, Airport, Month, and Cause.
- **Design Layout:** Arrange visuals neatly with clear titles and colors.
- **Publish & Share:** Upload to Power BI Service for interactive viewing.



## Challenges

- **Large Dataset Size** – Handling a big dataset caused slow processing and required memory optimization.
- **Missing and Inconsistent Data** – Many delay and cancellation columns had null or incorrect values that needed careful cleaning
- **Complex Date & Time Formatting** – Converting multiple time columns into proper datetime formats was challenging.
- **Mapping Cancellation Codes** – Cancellation codes (A, B, C, D, N) had to be correctly mapped to their reasons.

## Future Scope



- **Build Predictive Models** – Use machine learning to predict flight delays and cancellations in advance.
- **Real-Time Dashboard** – Develop a live Power BI or Streamlit dashboard using updated flight data.
- **Integrate Weather & Traffic Data** – Combine external datasets to improve accuracy of delay analysis.
- **Route Optimization Insights** – Suggest the best-performing routes and airports based on historical trends.



## Conclusion

- Cleaned and analyzed large airline datasets to understand operational patterns.
- Identified key delay causes, cancellation trends, and high-risk routes.
- Visual insights revealed peak delay times, seasonal trends, and airline performance.
- The findings support better decision-making and future real-time improvements.



Thank You!

