# DV AIRFLY INSIGHTS

## WEEK 1

### About the dataset:

- The dataset has **484,551 rows** and **29 columns**.
- There are **null values**: (`Dest_Airport`: 1,479 missing),(`Org_Airport`: 1,177 missing)
- There are 2 duplicate rows

### KPI's

- **Average Arrival Delay (AAL)** : 60.91 minutes
- **Average Departure Delay :** 57.49 minutes
- **On-time arrival performance :** 0% (no flight as arrival delay<0)
- **Cancellation rate :** 0% (no cancelled flights)
- **Diversion Rate : 0%** (there are no diverted flights)
- **Average weather delay :** 3.15 minutes , **Average carrier delay :** 17.41 minutes , **Late Aircraft Delay:** 26.65 , **National Aviation System (NAS) Delay :** 13.60 , **Security Delay:** 0.08
- The primary drivers of the observed delays are **Late Aircraft Delay** and **Carrier Delay**

### Cleaning Process

1. Import libraries
2. Load dataset
3. Check summary , datatypes , shape
4. Check for null values and replaced with mode value
5. Checked for duplicates and removed them
6. Converting Date format
7. Creating day , month , year , route columns

# WEEK 2

## 1. Handle nulls in delay and cancellation columns

- The **delay-related columns** (`DepDelay`, `ArrDelay`, etc.) and the **cancellation column** (`Cancelled`) are checked for missing values.
- Handled them by replacing with mode values
- This ensures that calculations like average delay or cancellation rates don't break because of `NaN`.

## 2. Create derived features: Month, Day of Week, Hour, Route

- From the datetime columns, extracted new **categorical and numerical features**:
  - **Day,Month,Year** → from Date column.
  - **Route** → a new feature created by combining `Origin` and `Dest` (e.g., `"JFK-LAX"`).
- These derived features are useful for **pattern analysis** (like busiest month, delays by day, etc.).

## 3. Format datetime columns

- Columns like `FlightDate`, `DepTime`, `ArrTime` are converted into **datetime .**

## 4. Duplicates removal

- Use the function duplicated() to find the rows that are duplicated . It returns the row number the True(if duplicated), False(if not)
- Duplicated().any() returns if there exists any duplicates.
- We can view the duplicate rows by using `data[data.duplicated(keep=False)]`

## Insights

1. Missing values in Org_Airport and Dest_Airport were filled with mode → No nulls left in these key categorical columns.
2. Duplicate rows were detected and removed → dataset integrity improved.
3. New columns created : day,month,hour,route

# WEEK 3

## Exploratory Data Analysis (EDA) Report

### 1. Definition:

Univariate analysis examines a single variable to describe its distribution, central tendency, and spread using tools like histograms, mean, median, and standard deviation.
Bivariate analysis investigates the relationship between two variables, checking for association or correlation, often using scatter plots, correlation coefficients, or cross-tabulations.

### 2. Analysis

- **Top Airlines:** The top airlines are ***Southwest Airlines co , American airlines inc , American eagle airlines inc , united airlines inc , skywest airlines inc*** …. This is calculated by computing no of times a flight is booked..
- **Top Routes:** Analysis of the most frequent routes highlights key city pairs with high air traffic density. These routes often correspond to business or hub-related travel corridors. **The top routes are *ord-Lga , lga-ord , lax-sfo , sfo-lax…..***
- **Busiest Months:** The month-wise flight count shows seasonal variation, with peaks during certain months. This could correspond to holiday seasons or favorable travel periods, leading to increased demand. ***The top two busiest moths are January and march.***

- **Flights by Day of Week:**
  The distribution of flights across days indicates consistent operations, though mid-week days generally experience slightly lower volumes compared to weekends or Fridays. ***Thursday and Friday are the busiest days of a week.***

- **Top Origin Airports:**
  The busiest origin airports are typically large metropolitan or hub airports, handling a high proportion of total departures. These airports are likely critical nodes in the flight network. **The top origin aiports are : *Chicago O'Hare international airport and Dallas worth international airport***

### 3. Delay Pattern Analysis

- **Average Departure Delay by Month:** The line plot of average departure delay reveals noticeable monthly fluctuations. Certain months exhibit higher average delays, potentially influenced by weather conditions or seasonal congestion**. Most delay is caused in *February* and least in *May.***

- **Departure Delay Distribution:** The boxplot indicate that most flights depart on time or within a short delay window. However, there are significant outliers representing severe delays. The distribution is right-skewed, with a small number of flights experiencing very high delays.

- **Delay Causes:** Summing across delay categories (`CarrierDelay`, `WeatherDelay`, `NASDelay`, `SecurityDelay`, and `LateAircraftDelay`) reveals that ***Late Aircraft Delay*** *and* ***Carrier Delay*** contribute the most to total delay minutes.
  This suggests that delays often propagate across the network due to late incoming aircraft or internal airline issues rather than external factors like weather or security.

## 4. **Key Insights and Observations**

- Airlines with higher operational volumes tend to have greater total delays, but not necessarily the highest average delays, indicating efficiency differences among carriers. ***JetBlue airways and united airlines*** **have most delays.**
- Delay patterns vary seasonally, implying external influences such as weather or demand surges.
- A stacked bar chart shows the airlines and proportionate delays due to various reasons

# WEEK 4

## Flight Delay Analysis Report

The goal of this analysis is to explore and visualize various causes of flight delays across different airlines. The dataset used (`cleaned_dataset.csv`) includes attributes like airline, types of delay (Carrier, Weather, NAS, Security, Late Aircraft), and total delay times.

### Analysis

**1.Average Delay Causes by Airline:** A grouped bar chart showing average delay (in minutes) per airline for each delay cause.This plot helps identify which airlines experience more frequent or severe delays.Typically, Late Aircraft Delay and Carrier Delay are the most significant contributors across most airlines.Some airlines show particularly high weather-related delays, possibly due to operational regions.

**2. Average Carrier Delay by Airline:** A bar plot showing the average carrier delay (minutes) for each airline.Carrier delays are directly related to operational inefficiencies (like crew or maintenance issues).Airlines with higher averages may have internal process inefficiencies or scheduling challenges. The blue color palette indicates intensity visually, allowing quick comparison.

**3. Correlation Between Delay Causes:** A correlation heatmap showing interdependence among various delay causes.Strong correlations suggest that when one delay cause increases, another might as well.Example: Late Aircraft Delay is often strongly correlated with NAS Delay or Carrier Delay, meaning one delay tends to trigger others.Weak or negative correlations indicate independent causes (like Security Delays, which tend to be isolated).

**4. Distribution of Delays by Airline (Boxplot):** Boxplots show how delays are distributed (spread, median, outliers) for each cause.Helps visualize variability and outliers.**Weather Delays** typically show wider spread due to unpredictable conditions.**Security Delays** have lower median and fewer outliers, indicating stability.

**5. Average Delay by Hour of Day:** This bar chart visualizes the average delay duration (in minutes) for different types of delays—Carrier, Weather, NAS, Security, and Late Aircraft—across each hour of the day (0–24 hours)**.** The early morning hours (1 AM to 4 AM) show the highest delay values, particularly for Late Aircraft Delay and Carrier Delay**.** Late Aircraft Delays spike dramatically around 2 AM to 3 AM, crossing 120 minutes on average, indicating that overnight operations and aircraft turnaround issues significantly affect flights scheduled at these times.

**6. Average Delay by Day of Week:** This grouped bar chart presents the average delay duration (in minutes) for each delay cause across the days of the week (Monday–Sunday**)**. The highest overall delays are observed on Friday and Saturday, while Wednesday and Sunday show relatively lower averages.

**7. Average Contribution of Delay Types:** This pie chart displays the percentage contribution of each delay type to the overall average flight delay. The Late Aircraft Delay segment dominates the chart, contributing nearly half of the total delay time**.** Carrier Delays also form a significant portion, suggesting internal airline operations play a critical role.

## Delay Trend Insights

- **Carrier and Late Aircraft Delays** dominate, indicating airline-controlled issues.

- **Weather Delays** are less predictable but regionally influenced.
- **NAS (National Airspace System) Delays** reflect air traffic control or congestion-related inefficiencies.
- **Security Delays** are minimal but crucial for passenger safety.

# WEEK 5

## Objective

The purpose of this analysis is to explore **how flight delays vary across different flight routes and airports**. By examining route-level and airport-level data, we aim to identify high-traffic routes, the correlation between various delay causes, and the overall performance of top airports.

## Analysis

1. **Top 10 Flight Routes by Frequency:** A bar chart displaying the Top 10 most frequent flight routes (e.g., IND-BWI, IND-LAS, IND-MCO, etc.).The highest flight frequencies are observed on popular domestic routes such as IND-BWI, IND-LAS, and IND-MCO, indicating heavy traffic between these airport pairs.High route frequency often correlates with higher overall delays, as increased air traffic leads to congestion, slot management issues, and turnaround delays.
2. **Delay Cause Correlations by Route:** A grid of heatmaps (3×3) showing the correlation between delay types (Carrier, Weather, NAS, Security, Late Aircraft) across the top 9 routes.On most routes, Carrier Delay and Late Aircraft Delay show strong positive correlations (0.6–0.8), indicating that one often triggers the other.
3. **Delay Correlations by Airport:** Another set of heatmaps showing delay cause correlations for the top 9 origin airports.Airports such as ATL, LAX, and ORD exhibit strong Carrier–Late Aircraft correlations, indicating heavy traffic and tight scheduling pressures.Weather Delay correlations vary geographically—airports in regions with frequent storms or snow (like DEN or ORD) show higher Weather Delay correlations.
4. **Busiest Airports and Average Delays:** This bar chart displays the average delay (in minutes) across the busiest airports in the U.S.The top airports analyzed include Chicago O'Hare (ORD), Dallas/Fort Worth (DFW), Hartsfield–Jackson Atlanta (ATL), Denver (DEN), Los Angeles (LAX), and others.Chicago O'Hare International Airport (ORD) records the highest average delay, nearing 70 minutes, indicating congestion and frequent operational delays.
5. **Average Delay by Day of the Week and Top 10 Airports**: The heatmap visualizes the average delay for the top 10 airports across days of the week, with color intensity representing delay duration.ORD (Chicago O'Hare) and SFO (San Francisco) exhibit consistently high delays throughout the week, visible through darker color gradients**.**
6. **Mean Delay Breakdown for Major Airports**: This stacked bar chart represents the mean delay composition for 15 major airports, categorized by delay type — Carrier Delay, Weather Delay, NAS Delay, Security Delay, and Late Aircraft Delay.Late Aircraft Delays (purple) are the dominant cause across almost all airports, suggesting systemic schedule propagation — once a delay starts, it tends to affect subsequent flights**.**
7. **Airport Performance: Volume vs. Average Delay:** This scatter plot visualizes the relationship between number of flights (volume) and average delay (minutes) across airports. There is no strong linear correlation between flight volume and average delay, suggesting that operational efficiency rather than volume alone drives delay performance.

## Overall Insights

- **ORD** and **JFK** emerge as the most delay-prone airports due to high operational demand and late aircraft propagation.

- **PHX**, **ATL**, and **DFW** show better control over average delays, demonstrating optimized flight scheduling and resource management.
- **Late Aircraft** and **Carrier Delays** remain the two biggest contributors to total delay time.
- Weekday trends indicate that **delays peak midweek**, aligning with business travel demand surges.
- The route and airport-level analysis helps pinpoint **which airports require focused operational interventions** and **policy-level changes** to improve punctuality.

# WEEK 6

## Monthly and Daily Cancellation Volume

The **monthly cancellation trends** graph (image_e689ff.png) shows the total number of cancellations per month for the first six months (Month 1 through Month 6).

- **Highest Volume:** Month 3 recorded the highest volume of cancellations, slightly above 45,000.
- **Lowest Volume:** Months 4 and 5 had the lowest cancellation volumes, both slightly below 35,000.
- **Overall Trend:** The overall volume is high in the first three months (above 40,000) and then decreases before rising again in Month 6 (to approximately 40,000).

The **daily cancellation trends** graph (image_e68a1e.png) illustrates the daily fluctuation of cancellations within a typical month.

- **High Peaks:** Cancellation volumes often peak around the beginning of the month (Days 1, 3, 5) and around Day 21, with volumes exceeding 10,000 on Days 1 and 3.
- **Low Troughs:** Cancellations tend to decrease toward the end of the month, with the lowest volumes occurring on Days 30 and 31 (around 6,000).

## Monthly Cancellation Rate

The **Monthly Cancellation Trend** line graph (image_e68d68.png) shows the cancellation *rate* (cancellations as a proportion of total flights) over the six months.

- **Increasing Rate:** The cancellation rate generally shows an **upward trend** over the six-month period, rising from approximately **0.497** in Month 1 to approximately **0.5015** in Month 6.
- **Fluctuation:** There is a noticeable dip in the rate in Month 4 (to about 0.4985) before it resumes an increase through Month 5 and Month 6.

## Cancellation Causes and Seasonal Impact

The **delay causes for cancellation** graph (image_e68a3c.png) indicates the primary factors contributing to flight cancellations. The causes are categorized by their associated delay time:

- **Dominant Cause: Late Aircraft Delay** is the overwhelmingly dominant cause, with associated delay time exceeding 6 million units.
- **Secondary Causes: Carrier Delay** (over 4 million units) and **NAS Delay** (National Airspace System Delay, over 3 million units) are the next most significant factors.
- **Minor Causes: Weather Delay** and **Security Delay** contribute the least to cancellation delays, with Security Delay being negligible (below 100,000 units).

The **holiday cancellation analysis** graph (image_e68d46.png) compares cancellation volumes between winter and non-winter seasons.

- **Seasonal Difference:** Cancellations in the **non-winter** season are approximately **double** those in the **winter** season (2.0 vs. 1.0 on the normalized y-axis). This

is a counter-intuitive finding, suggesting non-weather or non-winter factors drive higher cancellation numbers.

## Airline-Specific Cancellation Rates

The **Cancellation Rates by Month and Airline** heatmap (image_e68d25.png) shows the cancellation rate for 11 airlines across the 6-month period, with darker red indicating a higher rate.

- **Highest Rate: Hawaiian Airlines Inc.** consistently shows some of the highest cancellation rates, peaking dramatically at **0.58** in Month 6.
- **Other High Rates:** Other airlines with noticeable high-rate months include **Frontier Airlines Inc.** (0.53 in Months 2 and 3) and **JetBlue Airways** (0.52 in Months 1 and 5).
- **Lower Rates: United Air Lines Inc.** and **American Airlines Inc.** generally exhibit rates at or below the average of 0.50, suggesting relatively stable performance.

## Summary of Key Findings

1. **Volume vs. Rate:** While cancellation *volume* fluctuated, the overall cancellation *rate* trended upward over the six months.
2. **Primary Driver:** The most significant contributing factor to delays resulting in cancellations is **Late Aircraft Delay**, far surpassing Weather Delay.
3. **Seasonal Anomaly:** Cancellations were **twice as high** in the non-winter season compared to the winter season.
4. **Airline Performance: Hawaiian Airlines Inc.** demonstrated the highest cancellation rate, particularly in Month 6.

# WEEK 7

## 1. Overall Delay Distribution and Volume

### Proportion of Delay Causes

The analysis shows the total delay time is heavily concentrated in a few categories:

- **Top Cause: Carrier Delay** accounts for the largest proportion of total delay time, at **33.87%** (33 Million units).
- **Secondary Cause: Late Aircraft Delay** is the second largest contributor, making up **22.07%** (22 Million units).
- **NAS Delay** (National Airspace System Delay) is also significant at **14.88%** (13 Million units).
- **Minor Causes:** Weather Delay (7.78%) and Security Delay (0.71%) are minor contributors to the overall delay total.

### Total Delay by Airline

The total accumulated delay time varies significantly among airlines:

- **Highest Total Delay: Southwest Airlines Co.** has the highest total delay time, approaching **20 Million** units.
- **Other High Delay Airlines: American Airlines Inc.** and **United Air Lines Inc.** also show high total delay times (around 15 Million units).
- **Lowest Total Delay: US Airways Inc.** has the lowest total delay time among the airlines shown.

## 2. Time-Based Delay Trends

### *Average Delay Over Time (Months)*

The total delay exhibits significant monthly fluctuation:

- **Peak Delay:** Total delay peaks in Month 4 at approximately **4.3 Million** units.
- **Lowest Delay:** The lowest total delay occurs in Month 9, at approximately **1.7 Million** units.
- **General Trend:** The trend suggests higher delays in the spring/early summer months (e.g., Month 4) and a noticeable drop as the year progresses (e.g., Months 9, 10).

### Total Delay by Day of Week

Delays are not evenly distributed across the week:

- **Highest Delay: Day 5** (likely Friday) records the highest total delay, exceeding **17.5 Million** units.
- **Lower Delays:** Days 6 (Saturday) and 7 (Sunday) have the lowest total delays, with Day 6 being the lowest at approximately **10 Million** units.
- **Implication:** Travel on Day 5 appears to be significantly more prone to delays than weekend travel.

## 3. Airline-Specific Delay Metrics

### *Key Performance Indicators*

Across the entire dataset:

- **Total Flights: 485,000** flights.
- **Average Arrival Delay (ArrDelay): 60.91** minutes.
- **Average Departure Delay (DepDelay): 57.50** minutes.

### Specific Delays for Top 10 Airlines

Analysis of specific delay types for top airlines shows high values for Carrier and NAS delays:

- **Carrier Delay: American Airlines Inc.** and **Alaska Airlines Inc.** have the highest sums of Carrier Delay.
- **NAS Delay (ArrDelay): Delta Air Lines Inc.** has the highest sum of NAS Delay (labeled as Sum of ArrDelay in the table, but context suggests this may be a different variable or a visualization label error, given the extremely high value of 1,797,817). Assuming this refers to overall arrival delay time, it is the highest.

The busiest routes are:

- **ORD-LGA:** The top route by count of flights (around 1,920).
- **LGA-ORD, LAX-SFO, SFO-LAX:** Also highly frequent routes.

## 4. Behavioral

### *Distance vs. Delay*

- **Relationship:** Longer flights (higher Sum of Distance) do not necessarily correlate with the highest Sum of Arrival Delay.
- **Outliers: Delta Air Lines Inc.** shows a high distance with a moderate delay, while **Hawaiian Airlines Inc.** shows a high delay despite a moderate distance.

### TaxiOut vs. TaxiIn Total Delay

- **Correlation:** A strong positive correlation exists between Taxi-Out time and Total Delay.
- **Outliers: Hawaiian Airlines Inc.** shows the highest Taxi-Out time (approaching 1.5M units) and the highest Taxi-In time (around 0.6M units), suggesting potential issues with ground operations contributing significantly to overall delays.