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.