DV AIRFLY INSIGHTS

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

- % Missing values reduced
- Duplicate records removed
- % of dataset retained after cleaning
- Preprocessing time reduced
- Delay Columns : ArrDelay,DepDelay,CarrierDelay,WeatherDelay,NASDelay,SecurityDelay,LateAircraftDelay
- Cancellation columns : Cancelled , CancellationCode

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

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 .**

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