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

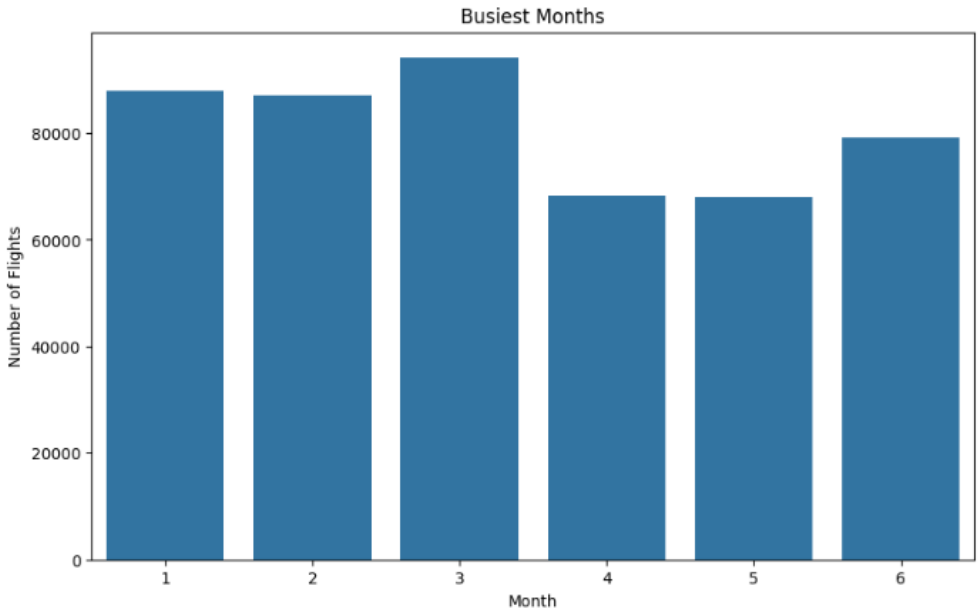
Exploratory Data Analysis (EDA) Report

1. Definition:

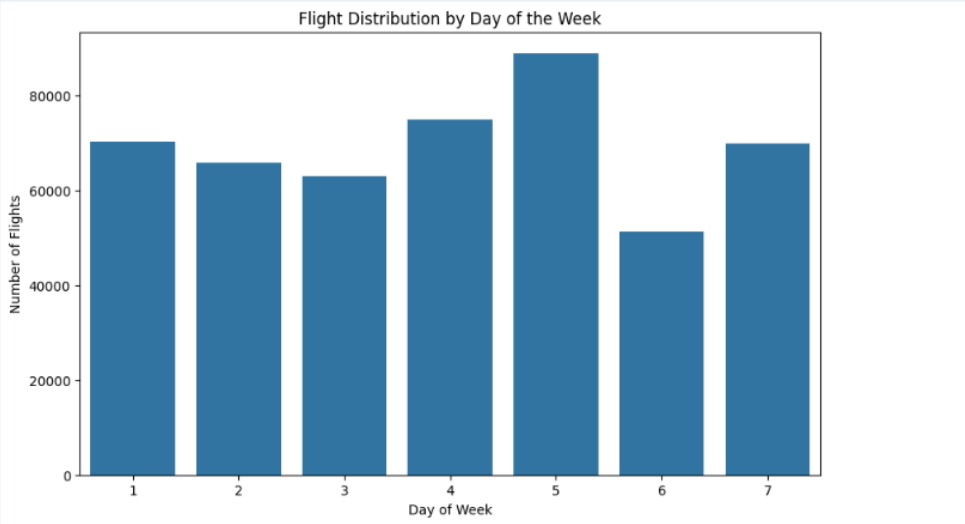
Univariate analysis examines a single variable to describe its distribution, central tendency, and spread using tools like histograms, mean, median, and standard deviation.
Bivariate analysis investigates the relationship between two variables, checking for association or correlation, often using scatter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

2. Analysis

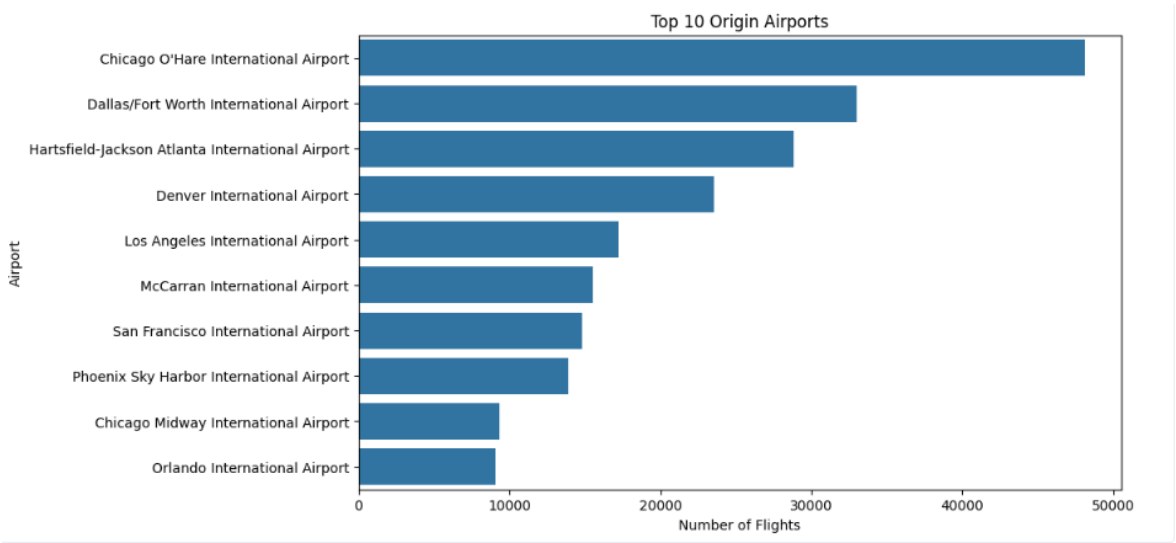
- **Top Airlines:** The top airlines are ***Southwest Airlines co , American airlines inc , American eagle airlines inc , united airlines inc , skywest airlines inc*** …. This is calculated by computing no of times a flight is booked..
- **Top Routes:** Analysis of the most frequent routes highlights 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.***

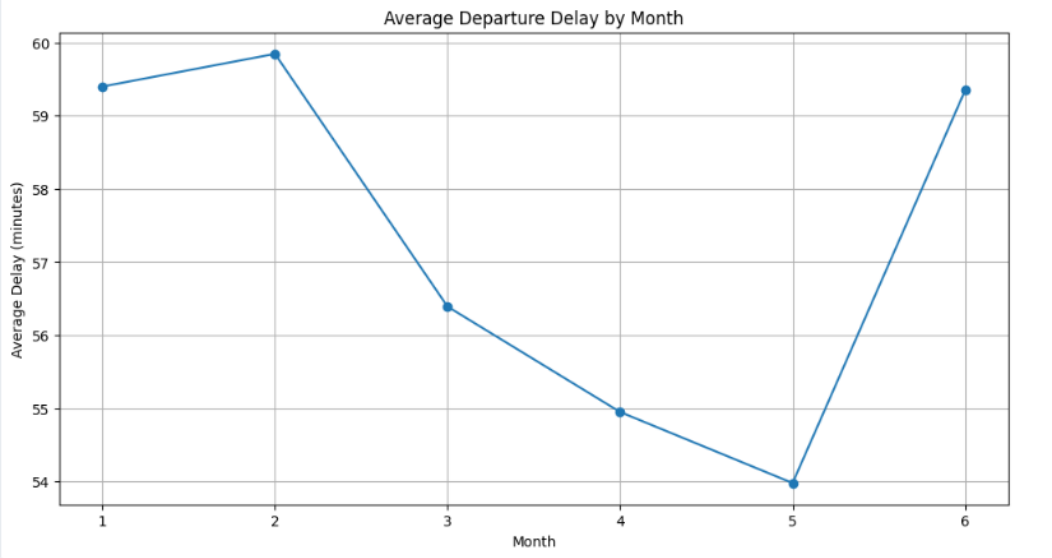


- **<u>Top Origin Airports:</u>**
  The busiest origin airports are typically large metropolitan or hub airports, handling a high proportion of total departures. These airports are likely critical nodes in the flight network. **The top origin aiports are : *Chicago O'Hare international airport and Dallas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

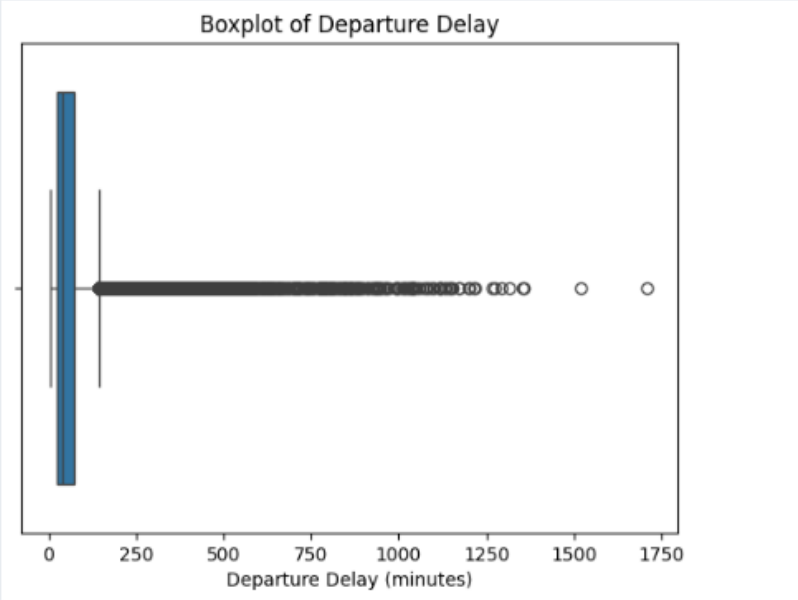


3. Delay Pattern Analysis

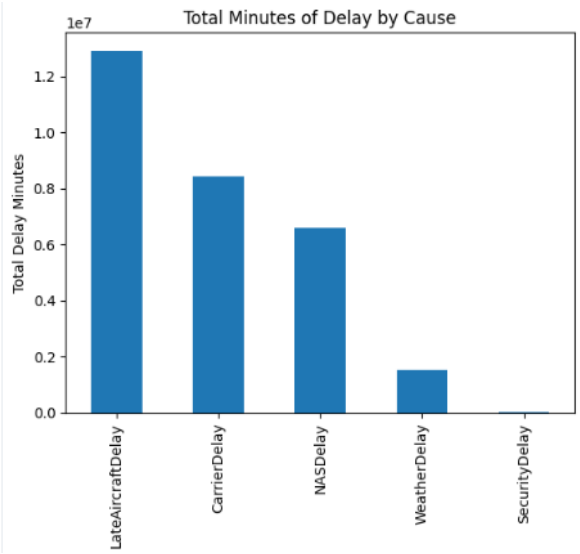
- **<u>Average Departure Delay by Month:</u>** The line plot of average departure delay reveals noticeable monthly fluctuations. Certain months exhibit higher average delays, potentially influenced by weather conditions or seasonal congestion**. Most delay is caused in *February* and least in *May.***



- **<u>Departure Delay Distribution:</u>** 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.

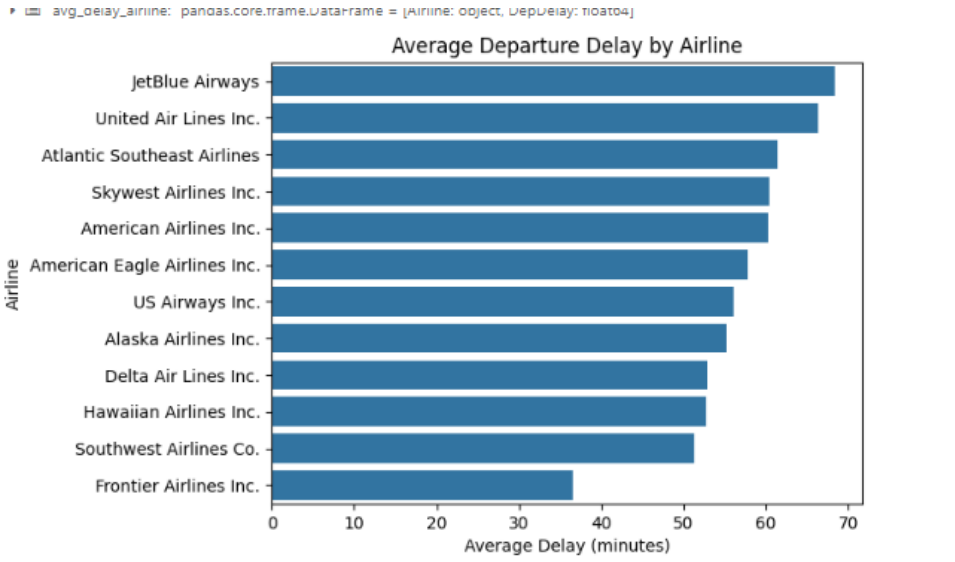


- **Delay Causes:** Summing across delay categories (`CarrierDelay`, `WeatherDelay`, `NASDelay`, `SecurityDelay`, and `LateAircraftDelay`) reveals that ***Late Aircraft Delay*** *and* ***Carrier Delay*** contribute the most to total delay minutes.
  This suggests that delays often propagate across the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

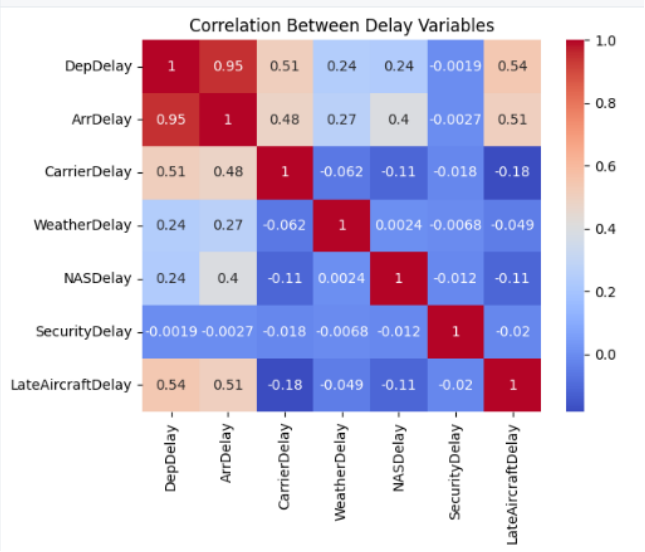


4. **<u>Key Insights and Observations</u>**

- 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