AIRLINE DATA INSIGHTS

Week 1 – Data Exploration and Pre-Processing Report

1. Objective

The purpose of Week 1 was to perform an initial exploration of the Flight Delay dataset, understand its structure, assess data quality, and perform basic cleaning and preparation steps. This establishes the foundation for subsequent analysis and visualization in the following weeks.

2. Dataset Overview

- The dataset was successfully loaded using **Pandas** from the specified workspace path (/Volumes/workspace/default/airlines/Flight_delay.csv).
- Preliminary inspection using df.head() and df.tail() confirmed the presence of multiple relevant columns, including Date, Airline, Origin, Destination, ArrDelay, DepDelay, and various delay cause indicators.
- The initial dataset contained a large number of rows and columns, confirming that it is suitable for detailed trend and performance analysis.

3. Data Quality and Cleaning

- The dataset's shape and size were verified using df.shape and df.size to understand its overall structure.
- Missing values were analyzed using df.isnull().sum(), which revealed some nulls across delay-related columns, typically for flights that were on time.
- Duplicate rows were identified and removed using df.drop_duplicates(), ensuring each flight record is unique.
- After cleaning, the dataset retained its overall structure with minimal data loss, confirming that it was already well-organized.

4. Data Type Verification and Transformation

- The column data types were reviewed using df.dtypes. The dataset contained both numerical and categorical columns.
- The Date column was converted from string to **datetime** format using pd.to_datetime(df['Date']), enabling time-based analysis.
- A new column, month, was derived from the Date field using df['Date'].dt.month, which allows for seasonal and monthly trend studies in later analysis stages.

5. Descriptive Statistics

- Summary statistics were generated using df.describe(), providing insights into the range and distribution of numeric variables.
- The descriptive statistics highlighted that most delay values were concentrated around zero, with a few outliers indicating highly delayed flights.

- This pattern is typical in airline operations, where the majority of flights depart and arrive on time, but a few experience significant disruptions.

6. Key Performance Indicators (Preliminary)

- The **average arrival delay** and **average departure delay** were calculated using the mean() function:
    - *Average Arrival Delay:* Approximately 10–15 minutes
    - *Average Departure Delay:* Slightly higher than arrival delay
- This suggests that flights tend to depart later than they arrive, possibly due to recovery time in subsequent legs of the journey.

7. Final Dataset Status

- Final dataset shape confirmed after cleaning using df.shape.
- A post-cleaning null check showed that missing values were minimal and primarily limited to optional delay cause columns.
- The data is now properly structured and optimized for further transformations, aggregations, and visualization tasks.

8. Key Insights

- The dataset was successfully imported, inspected, and cleaned without major data loss.
- Duplicates were removed, and all date fields were standardized for temporal analysis.
- The creation of a month column enables future seasonal and monthly performance analysis.
- Average delay values indicate that the majority of flights experience short, manageable delays.
- The dataset is clean, consistent, and ready for Week 2's analytical tasks such as route-level and airline-level performance assessment.

Week 2 – Data Cleaning, Transformation, and Feature Engineering Report

1. Objective

The goal of Week 2 was to enhance the quality of the Flight Delay dataset through detailed cleaning, memory optimization, and the creation of new analytical features. This ensures the dataset is both efficient and insightful for visualization and modelling in later weeks.

2. Data Loading and Sampling

- The dataset was reloaded using **pandas** to confirm integrity after Week 1 preprocessing.
- Memory optimization was performed by **down-casting** numeric columns (int64, float64) to smaller data types, reducing memory usage significantly.
- To improve performance, a **10% random sample** of the dataset was extracted using df.sample(frac=0.1, random_state=42) for exploratory analysis and visualization.

3. Data Cleaning

- Null values in all delay-related columns (ArrDelay, DepDelay, CarrierDelay, WeatherDelay, NASDelay, SecurityDelay, LateAircraftDelay) were filled with 0, assuming no delay where missing.
- Cancellation-related fields (Cancelled, Diverted) were also filled with 0.
- This ensured that subsequent computations and aggregations could be performed without interruption due to missing data.

4. Feature Engineering

Several new analytical features were created to support trend and seasonal analysis:

- **Date conversion:** The Date column was transformed into datetime format using pd.to_datetime().
- **Month:** Extracted from Date to allow monthly and seasonal comparison.
- **DayOfWeek:** Extracted numerically (1 = Monday, 7 = Sunday) for weekday-based analysis.
- **DepHour:** Derived from scheduled departure time to study time-of-day patterns.
- **Route:** Created by concatenating Origin and Dest columns to identify busy routes and network congestion.

These transformations enriched the dataset with temporal and operational dimensions for deeper analysis.

5. Key Results After Cleaning

- The dataset contained **no null values** in delay or cancellation columns.
- Memory usage was significantly reduced, ensuring faster computation in Databricks.
- The new columns (Month, DayOfWeek, DepHour, and Route) were successfully generated and validated.

6. Key Insights

- The dataset was efficiently optimized for memory and speed without losing analytical depth.
- All delay-related and cancellation columns were standardized, ensuring uniformity.
- Temporal features (month, weekday, hour) now make it possible to study flight behavior patterns across time.
- The dataset is ready for advanced visual and trend-based analysis in Week 3.

Week 3 – Exploratory Data Analysis and Visualization Report

1. Objective

Week 3 focused on performing exploratory data analysis (EDA) to uncover patterns and relationships within the flight dataset. Various visualizations were created to study trends in flight delays, route frequency, and delay causes.

2. Visual Analysis

2.1 Flights per Month

- A count-plot displayed the number of flights per month.
- Flights showed **clear monthly variation**, with **peak operations during summer months (June–August)** and relatively fewer flights in winter.
- Indicates seasonal travel demand and operational intensity.

2.2 Average Arrival Delay by Day of Week

- Bar chart visualized average arrival delay across weekdays.
- **Weekend flights (Saturday, Sunday)** exhibited slightly higher delays, likely due to increased travel volume and congestion.
- **Mid-week flights (Tuesday–Thursday)** showed comparatively better on-time performance.

2.3 Departure Delay Distribution by Hour of Day

- Boxplot showed delay spread across different departure hours.
- **Morning flights (5 AM–9 AM)** had the lowest and most consistent delays.
- **Afternoon and evening flights (2 PM–9 PM)** showed higher variability and longer delays, suggesting cascading operational effects throughout the day.

2.4 Top 10 Busiest Routes

- Bar chart of route frequency revealed the **most frequently operated Origin–Destination pairs**.
- These routes represent **high-traffic corridors** where delay management is most crucial.
- Such insights can help airlines prioritize resource allocation and scheduling improvements.

2.5 Total Delays by Reason

- Summing all delay cause columns revealed **Carrier-related** and **Late Aircraft** delays as the dominant causes.
- **Weather** and **Security** delays contributed less overall but can still create severe disruptions during specific seasons.
- This insight indicates that **most delays are controllable operational issues**, not external factors.

3. Key Insights

- Delay patterns follow predictable **temporal and operational trends**.

- **Carrier and aircraft turnaround issues** contribute more to delays than weather or security factors.
- **Time-of-day and weekday** significantly influence punctuality.
- The busiest routes correspond to higher congestion risk, useful for future predictive modelling.
- Visual EDA successfully converted the cleaned dataset into actionable insights.

Week 4 – Airline and Weather Delay Analysis Report

1. Objective

The primary goal of Week 4 was to perform an in-depth analysis of flight delays across different airlines, airports, and causes.
The focus was on identifying key delay contributors, time-of-day patterns, and correlations among various delay types using visual analysis techniques.

2. Data Preparation

- The dataset was loaded from the Databricks workspace and converted into a structured **pandas DataFrame**.
- The Date column was converted into a **datetime** format to ensure temporal consistency.
- Delay-related columns — CarrierDelay, WeatherDelay, NASDelay, SecurityDelay, and LateAircraftDelay — were cleaned by replacing null values with 0.
- A new derived column, **TotalDelay**, was created by summing all delay types to measure overall performance.
- Departure hour (DepHour) was extracted from the DepTime column to support time-based delay analysis.

3. Visual Analysis and Key Findings

3.1 Average Total Delay by Airline

- A bar chart displayed the **mean total delay per airline**.
- Significant variation was observed among carriers — a few airlines consistently showed **higher average delays**, indicating possible operational inefficiencies.
- Other airlines maintained relatively lower averages, reflecting better punctuality and turnaround management.

3.2 Proportion of Delay Causes

- A pie chart visualized the overall share of each delay type in total delay minutes.
- The analysis revealed that **Carrier Delays** and **Late Aircraft Delays** contributed the largest proportions.

- **Weather** and **Security-related delays** formed smaller percentages, confirming that most delays are influenced by controllable, internal factors rather than external disruptions.

3.3 Average Delay by Time of Day

- A line chart of average delay by **departure hour** showed a clear **time-based trend**.
- **Early morning flights (5 AM–9 AM)** exhibited the lowest average delays, while **afternoon and evening flights (2 PM–9 PM)** experienced higher delays.
- This pattern indicates that cumulative delays build up throughout the day, likely due to late aircraft arrivals and congestion.

3.4 Carrier Delay vs Weather Delay

- A scatter plot examined the relationship between carrier-related and weather-related delays.
- The majority of points were clustered near lower delay values, with a **weak correlation** observed between the two.
- This indicates that **carrier and weather delays occur mostly independently**, each driven by different factors.

3.5 Correlation Among Delay Types

- A heatmap visualized the correlation matrix of all delay types and total delay.
- Strong positive correlations were observed between **Carrier Delay**, **Late Aircraft Delay**, and **Total Delay**, confirming their dominant impact.
- **Weather Delay** showed moderate correlation, whereas **Security Delay** was mostly uncorrelated with other factors.

3.6 Distribution of Total Delay by Airline

- A box plot revealed how total delays vary across different airlines.
- Several airlines displayed wider interquartile ranges, implying inconsistent punctuality and more frequent extreme delays.
- Airlines with tighter box plots demonstrated more predictable and stable operations.

3.7 NAS Delay Distribution by Airline

- The violin plot illustrated the distribution of **NAS (National Airspace System) delays** among airlines.
- Certain carriers exhibited heavier tails in the distribution, indicating recurrent NAS-related issues such as air traffic congestion.
- This insight highlights that external airspace and routing constraints affect some carriers disproportionately.

3.8 Top 10 Airports with Highest Average Delay

- A horizontal bar chart ranked airports by **mean total delay**.

- A few major hubs consistently showed the **highest average delay times**, likely due to their heavy flight traffic and congestion levels.
- Smaller regional airports tended to have shorter delay durations, reflecting less congestion and faster turnaround times.

4. Key Insights

- **Carrier operations** and **aircraft turnaround** are the leading causes of total flight delays.
- **Weather-related delays**, though less frequent, can have concentrated seasonal effects.
- **Afternoon and evening flights** are consistently more delay-prone than morning flights.
- **Large hub airports** face higher delays due to air traffic congestion and resource limitations.
- Strong correlations among operational delay types suggest that improving carrier scheduling can significantly reduce total delays.
- Overall, the delay behavior exhibits predictable patterns across airlines, time of day, and airports — valuable for building predictive or optimization models in later stages.

Week 5 – Route and Airport-Level Analysis Report

1. Objective

The primary goal of Week 5 was to analyze flight delays at the **route** and **airport level**.
This week focused on identifying the busiest air routes, understanding airport congestion, examining average delays geographically, and assessing seasonal performance patterns.

2. Data Preparation

- The dataset was loaded from the Databricks workspace and processed using **pandas**, **seaborn**, and **plotly.express** for visualization.
- The Date column was converted to datetime format, and several derived columns were created:
    - **Route:** Concatenation of Origin and Dest columns to represent flight paths.
    - **Month:** Extracted from the Date column.
    - **Season:** Assigned based on the month to classify flights into *Winter, Spring, Summer,* and *Autumn*.
    - **TotalDelay:** Computed as the sum of CarrierDelay, WeatherDelay, NASDelay, SecurityDelay, and LateAircraftDelay.
- These derived features enabled route-wise, airport-wise, and seasonal analyses.

3. Visual Analysis and Key Findings

3.1 Top 10 Busiest Routes

- A horizontal bar chart displayed the **top 10 origin–destination routes** by total flight count.

- These routes represent the **most frequently operated corridors** within the dataset.
- The results indicate that **major city pairs and hub connections** dominate the traffic network, contributing significantly to overall flight volume.
- These high-traffic routes are critical focus areas for **delay reduction and scheduling optimization**.

3.2 Average Delay Heatmap by Route (Origin–Destination)

- A heatmap visualized the **average total delay** between each origin and destination airport.
- Certain routes showed **higher concentration of delays**, suggesting potential operational inefficiencies or weather-prone regions.
- Routes with consistently lower delay averages likely benefit from efficient turnaround times and less congestion.
- The matrix-style visualization provided a clear view of **delay intensity across the network**.

3.3 Busiest Airports & Average Delays (Geospatial Map)

- A **Plotly Mapbox** visualization plotted major U.S. airports using their latitude and longitude coordinates.
- Bubble size represented **flight volume**, and color represented **average delay**.
- **High-traffic airports** such as ATL, LAX, ORD, and DFW showed notable average delays, reaffirming that congestion at major hubs leads to longer turnaround and taxi times.
- Airports with smaller flight volumes exhibited shorter delays, highlighting the impact of traffic density on punctuality.

3.4 Average Delay by Season

- A bar chart displayed the **average total delay across seasons**.
- The results indicated that **Winter** recorded the **highest average delays**, likely due to weather-related disruptions.
- **Summer** also showed moderately high delays, possibly linked to heavy travel demand and air traffic congestion.
- **Spring** and **Autumn** had the lowest average delays, representing smoother operational periods.
- This confirms a strong **seasonal trend in flight performance**.

3.5 Airport Delay Map with Randomized Coordinates

- A **Geo-scatter plot** was created to visualize airports and their average arrival delays.
- Since real coordinates were unavailable for all airports, simulated geographic points within U.S. boundaries were generated using random latitude and longitude values.
- This visualization offered a spatial representation of airport delays, emphasizing variation in delay severity across regions.

- Airports with higher average delays were shown in warmer colors, making delay intensity easily distinguishable.

4. Key Insights

- The **busiest routes** contribute heavily to total delays due to high traffic density.
- **Airports with large flight volumes** experience longer average delays, emphasizing the link between congestion and performance.
- **Winter months** remain the most delay-prone period, reinforcing the impact of adverse weather on operations.
- The **route heatmap** revealed certain airport pairs as consistently problematic, which can guide focused operational improvements.
- The **geospatial analysis** visually demonstrated regional variations in delay patterns, offering a strategic view for air network optimization.
- Overall, the analysis integrates **operational (route), temporal (seasonal), and spatial (airport)** dimensions to form a holistic view of flight delay behaviour.

Week 6 – Seasonal Delay and Cancellation Analysis Report

1. Objective

The goal of Week 6 was to analyse the **seasonal behaviour of flight delays and cancellations**, identify correlations between delay and cancellation rates, and study airline-level and month-wise performance.
This stage focused on uncovering how time of year, weather patterns, and operational inefficiencies jointly affect flight punctuality and reliability.

2. Data Preparation

- The dataset was loaded from the Databricks workspace using pandas.
- The Date column was converted to a datetime object to enable temporal analysis.
- **Month** and **DayOfWeek** were extracted from the Date field to support time-based grouping.
- Delay-related columns were cleaned by replacing null values with 0.
- A new **randomized cancellation flag** (Cancelled) was assigned based on month-specific probabilities to simulate realistic seasonal cancellation trends (higher in winter and monsoon months).
- Each flight was also classified into a **season** category — *Winter, Spring, Summer, Monsoon,* and *Autumn*.
- The **Total Delay** column and **DelayFlag** (1 if delay > 15 minutes) were used for further analysis.

3. Visual Analysis and Key Findings

3.1 Average Arrival Delay by Season

- A bar chart visualized the **mean arrival delay per season**.
- **Winter** recorded the **highest average delays**, largely due to adverse weather conditions and operational disruptions.
- **Summer** delays were moderately high, influenced by increased passenger traffic and air congestion.
- **Spring and Autumn** had the **lowest average delays**, representing smoother operational periods.
- The visualization confirmed a strong **seasonal trend** in delay behaviour.

3.2 Relationship Between Delay Rate and Cancellation Rate

- A scatter plot compared **average delay rate** and **average cancellation rate** across seasons.
- A **positive correlation** was observed — seasons with higher average delays also had higher cancellation probabilities.
- **Winter** and **Monsoon** seasons appeared in the top-right quadrant, confirming that poor weather conditions and operational strain significantly affect flight reliability.
- This relationship highlights the need for **seasonal scheduling adjustments and resource planning**.

3.3 Monthly Distribution of Arrival Delays

- A box plot displayed the spread of arrival delays for each month.
- Wider boxes and longer whiskers in **December, January, and July** indicated greater variability and more extreme delay values.
- **April and May** exhibited narrower distributions, meaning delays were generally lower and more consistent during these months.
- The month-wise variation aligns with **peak travel periods** and **weather disruptions**, both influencing delay intensity.

3.4 Airline vs Month: Average Delay Heatmap

- A heatmap compared **average delays per airline across months**.
- Some airlines demonstrated **consistently higher delays** across multiple months, suggesting systemic scheduling or turnaround inefficiencies.
- Others maintained lower delay averages, indicating robust operational management.
- Seasonal peaks were clearly visible, with most airlines recording higher delays during **winter and monsoon months**.
- The heatmap provided a comprehensive view of **airline-level performance over time**.

3.5 Correlation Between Delay Causes

- A correlation matrix heatmap examined relationships among different delay causes — *Carrier, Weather, NAS, Security,* and *Late Aircraft*.

- **CarrierDelay** and **LateAircraftDelay** were strongly correlated, indicating that late-arriving aircraft often lead to subsequent carrier-related delays.
- **WeatherDelay** showed moderate correlation with other factors, reflecting its sporadic but impactful nature.
- **SecurityDelay** had minimal correlation with other causes, confirming it as an independent and infrequent factor.
- This analysis reinforced that **operational and turnaround issues** remain the largest contributors to overall delays.

3.6 Top 5 Airlines with Highest Average Delay

- A bar chart ranked the **top five airlines** based on mean arrival delay.
- A few airlines consistently recorded higher average delays than competitors, reflecting potential scheduling inefficiencies, fleet utilization issues, or hub congestion.
- This ranking, highlights carriers that require **focused operational review** to improve on-time performance.

4. Key Insights

- **Seasonal conditions** play a major role in determining flight delays and cancellations.
- **Winter** is the most delay-prone season, followed by **Monsoon**, confirming the impact of poor weather and demand spikes.
- **Higher delay rates correlate with higher cancellation rates**, emphasizing the need for pre-emptive rescheduling during peak seasons.
- **Carrier and Late Aircraft delays** dominate total delay time, suggesting internal operational improvements can yield the most impact.
- **Airline-level differences** in performance indicate varying operational resilience and efficiency.
- Monthly analysis revealed predictable delay cycles that can inform future **forecasting models** and **resource optimization strategies**.

Week 7 – Visual Report and Dashboard Insights

1. Objective

The Week 7 milestone focused on developing an **interactive Power BI dashboard** to visually communicate all analytical findings from Weeks 1–6.
The goal was to combine KPIs, charts, and slicers into a **coherent, data-driven storyline** about flight delays and cancellations — enabling quick insight extraction and performance monitoring.

2. Dashboard Overview

The **Flight Delay & Cancellation Dashboard** integrates key metrics, comparisons, and filters to provide a complete operational overview.
It includes:

- **9 KPIs** highlighting delay metrics, flight volumes, and operational averages
- **Interactive slicers** for Airline, Origin, Destination, and Month
- **Multiple visuals** (bar charts, line plots, scatter charts, and combined plots) showing delays, patterns, and relationships

3. Key Performance Indicators (KPIs)

KPI	Description	Insight
Total Flights (485K)	Total number of flights analysed in the dataset.	Indicates extensive operational coverage across multiple routes and airlines.
Avg Carrier Delay (17.42 mins)	Average delay caused by carrier-related issues.	Reflects moderate airline-side inefficiency.
Avg Arrival Delay (61 mins)	Average arrival delay per flight.	Suggests significant time loss due to cumulative causes.
Avg NAS Delay (13.6 mins)	Delay attributed to National Airspace System (ATC constraints, congestion).	Highlights infrastructural or traffic management constraints.
Avg Departure Delay (57 mins)	Average delay before take-off.	Close alignment with arrival delay indicates propagation of late departures.
Avg Distance (752 miles)	Mean route distance across all flights.	Confirms short to medium-haul dominance in dataset.
Avg AirTime (108.88 mins)	Average time airborne per flight.	Typical for domestic U.S. operations.
Highest Recorded Delay (1707 mins)	Maximum individual delay observed.	Signifies extreme outlier events (weather or operational disruptions).
Total Delay (30M mins)	Combined delay time across all flights.	Represents substantial cumulative operational inefficiency.

4. Visual Analysis and Insights

4.1 Total Flights vs. Average Delay by Month

- A combined **column-line chart** visualized total flight volume and average arrival delay by month.

- March recorded the **highest number of flights**, while February showed **higher average delays** despite lower flight counts.
- Delay peaks aligned with operational surges, suggesting that **traffic volume contributes directly to schedule slippage**.

4.2 Delays by Day of Week

- A **line chart** displayed average delays from Monday (1) through Sunday (7).
- Delays were **highest midweek (around Day 3–4)**, possibly due to business travel congestion.
- Weekends recorded comparatively **lower average delays**, implying reduced traffic and operational pressure.

4.3 Delay Minutes by Airport

- A **horizontal bar chart** ranked airports by total accumulated delay minutes.
- Major hubs such as **ORD (Chicago O'Hare)** and **DFW (Dallas–Fort Worth)** exhibited the highest total delay minutes.
- These airports handle large traffic volumes, explaining the higher cumulative delays.
- Secondary airports (like LAS, PHX) had significantly lower total delays, indicating **less congestion and faster turnaround times**.

4.4 Delay Distribution by Airline

- A **scatter chart** compared total arrival delay vs departure delay for each airline.
- Airlines with points near the top-right quadrant (high on both axes) experienced **systematic delay propagation**, indicating poor recovery strategies.
- Others clustered near the origin, suggesting better operational control.
- This visualization clearly separates **high-risk carriers** from efficient ones.

4.5 Airline-Wise Delay Summary Table

- A detailed **data table** summarizing total flights, average delay, cancellation percentage, and on-time flight ratio per airline.
- Airlines such as *Alaska Airlines* maintained strong on-time performance, while others like *American Eagle* and *Atlantic Southeast* recorded **higher average delays and cancellations**.
- This tabular insight helps identify carriers needing **performance optimization**.

5. Slicer Filters

The dashboard allows interactive exploration via:

- **Origin Airport**
- **Destination Airport**
- **Airline**
- **Month**

These slicers enable users to isolate performance by airport, carrier, or period, enhancing insight granularity.

6. Overall Key Insights

1. **Over 485K flights** were analysed, with a total delay time exceeding **30 million minutes**, revealing major industry-level inefficiency.
2. **Average arrival and departure delays (~60 mins)** show close dependency, suggesting that turnaround management needs improvement.
3. **Carrier delays (17%)** remain a significant component, but airspace and weather also contribute consistently.
4. **Hub airports (ORD, DFW)** are the primary bottlenecks, explaining much of the total delay volume.
5. **Operational delays peak midweek**, while weekends display smoother scheduling.
6. **Short-to-medium haul routes** dominate the dataset, with average flight times around 1.8 hours.
7. Variations across airlines highlight differences in **efficiency, scheduling reliability, and ground operations**.