**AirFly Insights: Data Visualization and Analysis of Airline Operations**

**Data Cleaning**

**Introduction**

The airline dataset contained raw operational flight records with missing values, unoptimized data types, and inconsistent time formats. To make it analysis-ready, a step-by-step cleaning and transformation process was applied using Pandas.

The main objectives were:

* Explore dataset shape, schema, and memory usage.
* Optimize datatypes to reduce memory footprint.
* Handle missing values in critical columns.
* Standardize date and time formats.
* Create new derived features for analysis.
* Save the cleaned dataset for reuse.

**Data Exploration**

Before cleaning, the dataset was examined to understand its structure:

* Shape (rows, columns) was checked with .count() and len(df.columns).
* Schema information with df.info().
* Datatypes with df.dtypes.
* Missing values per column with df.isnull().sum().
* Memory usage with df.memory\_usage(deep=True).

This provided insights into the dataset size, missing data distribution, and memory efficiency.

**Memory Optimization**

To improve efficiency, a custom function optimize\_dtypes() was applied:

* Integers were downcasted from int64 → int32/int16.
* Floats were downcasted from float64 → float32.
* Object columns with low cardinality were converted into category.

Result: Memory usage reduced significantly, improving performance for large datasets.

**Handling Missing Values**

Missing values were replaced as follows:

* ArrivalDelay → 0
* DepartureDelay → 0
* Cancelled → 0

This ensured no nulls in critical operational columns.

**Standardizing Date and Time**

* **Date column** was converted to datetime.
* **DepTime column** (e.g., 930 → 09:30) was standardized using zero-padding and then converted into a proper time object with pd.to\_datetime(...).dt.time.

This allowed consistent time-based analysis.

**Feature Engineering**

Several derived features were created to support downstream analysis:

* **Month** → Extracted from Date.
* **DayOfWeekNum** → Extracted from Date (0 = Monday, 6 = Sunday).
* **DayName** → Full weekday name.
* **Hour** → Extracted from DepTime.
* **Route** → Concatenation of Origin and Dest (e.g., JFK-LAX).

These new variables make it easier to study trends over time, flight schedules, and route performance.

**Saving the Cleaned Dataset**

The final cleaned dataset was exported as a CSV file for reuse:

**Outcome**

* Dataset shape, schema, and memory usage analyzed.
* Memory usage reduced via datatype optimization.
* Missing values handled systematically.
* Dates and times standardized for consistency.
* New analytical features (Month, DayOfWeek, DayName, Hour, Route) created.
* Cleaned dataset exported successfully.