Airfly Insights Project

About Dataset

The dataset used in this project is the Flight Delay dataset, which contains details about scheduled and actual flight timings, airline carriers, origin and destination airports, flight numbers, distances, cancellations, and delay details. This dataset is essential for analyzing flight patterns, identifying causes of delays, and building predictive models to improve operational efficiency.

The airline delay dataset was analyzed to understand flight patterns, delays, and overall performance. The raw dataset contained 484,551 rows and 29 columns. After cleaning and feature engineering, the final dataset has 484,549 rows and 33 columns.

Key Performance Indicators (KPIs)

- Average delay time per airline, airport, and route
- Percentage of delayed flights
- On-time performance rate
- Delay distribution by time of day, day of week, and season
- Feature importance in delay prediction models

Data Cleaning Steps

- Converted 'Date' column into datetime format for proper time-series analysis.
- Extracted month from 'Date' for seasonal trend analysis.
- Checked top and bottom rows to inspect dataset structure using head() and tail().
- Inspected dataset size, shape, column names, and data types for consistency.
- Handled missing values: filled missing values in 'Org_Airport' and 'Dest_Airport' with 'Unknown'.
- Removed duplicate records based on 'UniqueCarrier', 'FlightNum', and 'AirNum' columns.
- Sampled 10% of dataset for exploratory analysis to optimize performance.
- Checked memory usage and optimized storage if required.
- Analyzed unique values in categorical features such as 'Org_Airport'.
- Computed minimum, maximum, and average distance for flights to understand range of operations.
- Analyzed flight distribution by day of week and calculated cancellation rate.

Metrics and Insights

1. Flight counts and cancellation rates were calculated to assess airline reliability.
2. Minimum, maximum, and average distances helped identify flight route characteristics.
3. Flight counts by day of week showed peak travel periods.
4. Cancellation rates provided insight into operational disruptions.
5. Airline frequency analysis highlighted major carriers in the dataset.
6. Sampling allowed efficient exploratory analysis without overwhelming computation.