Week4_AirFly_Insights

Objectives

- Perform advanced analyses on the flight dataset beyond basic EDA.

- Explore delay causes and their contributions by carrier and month.

- Investigate route-level insights including cancellations, delay averages, and variability.

- Analyze temporal patterns such as seasonality, weekday vs weekend performance, and autocorrelation.

- Study aircraft (tail-level) performance, turnaround times, and delay propagation.

- Identify airport hubs and connectivity patterns.

- Apply clustering, PCA, and basic predictive models for deeper feature understanding.

Tasks Completed

Delay Cause Analysis

- Identified delay cause columns and filled missing values.

- Computed average contributions by delay type (carrier, weather, NAS, security, late aircraft).

- Plotted stacked bars of delay causes across carriers and months.

Route-Level Insights

- Computed top routes by average delay and cancellation rate.

- Compared routes in a scatter of avg delay vs cancel rate.

Seasonal & Temporal Patterns

- Analyzed delays by quarter, weekday vs weekend, and hour of day.

- Built heatmap of average delays by weekday-hour combination.

- Conducted seasonal decomposition of daily flight counts and delays.

Airport-Level Analysis

- Ranked busiest airports by flights.

- Computed average delays and cancellation rates at origin airports.

- Analyzed airport-level delay vs cancellation correlations.

Aircraft & Turnaround

- Estimated turnaround times from actual arrival to next departure by tail number.

- Checked delay propagation between inbound and outbound flights of same tail.

- Identified tails with highest delay fraction or cancellations.

Advanced Analysis

- Clustered routes based on average delay and distance (k-means).

- Applied PCA to numeric features (sample).

- Constructed airport connectivity metrics (in-degree, out-degree) to detect hubs.

- Performed autocorrelation of daily delays to identify weekly patterns.

Model-Based Insights

- Ran quick linear regression and decision tree models to gauge feature importance for ARR_DELAY.

- Performed outlier detection with groupwise z-scores.

Exports & Reporting

- Exported route statistics summary as CSV.

- Built summary Word document with advanced insights.

Key Findings

- Delay causes vary across carriers: some carriers exhibit more carrier-related delays, others more weather-driven.

- Route-level analysis shows certain high-traffic routes also suffer high cancellation rates.

- Weekend vs weekday delays show slight differences, with weekends often worse in median values.

- Hourly patterns reveal late evening flights often experience higher delays.

- Turnaround and tail analysis confirm propagation: late inbound often results in delayed outbound flights.

- Connectivity analysis highlights key hubs with extensive in/out-degree links.

- PCA and clustering suggest distance and delay combine to group routes into distinct clusters.

- Autocorrelation in delays suggests weekly cycles, consistent with operational schedules.

Challenges Faced

- Not all datasets contained complete delay cause columns; missing values required careful handling.

- Turnaround computation depended on having consistent ACTUAL_DEP/ACTUAL_ARR times; missing values reduced sample size.

- Clustering and PCA required filtering routes with sufficient data to avoid instability.

- Seasonal decomposition and ACF analyses required resampling, which led to sparse series on smaller subsets.

Learnings

- Delay propagation is a real operational issue; aircraft arriving late often propagate the delay to next flights.

- Combining operational features (tail, flight number, route, time) with performance outcomes enables richer analysis.

- Connectivity metrics help in identifying true hubs beyond simple flight counts.

- Clustering and PCA, while exploratory, provide a different view of route typologies (short-haul on-time vs long-haul delay-prone).

- Autocorrelation analyses provide evidence for cyclical delay patterns, useful for forecasting and scheduling adjustments.