R PRANATHI

Airline Delays — Data Analysis Report

1. Dataset Description

1.1 Source:
Kaggle dataset *"Airline Delays"* by eugeniyosetrov, covering December 2019 and December 2020 flight delay summaries.

1.2 Typical Columns (possible, based on dataset summary):
Because this is a summary/delay dataset, expected columns include:

- `carrier` (airline code)
- `airport` or `origin` / `destination`
- `date` / `month` / `year`
- `num_flights` — number of flights in that period
- `delay_count` — number of delayed flights
- `avg_delay` — average delay (minutes)
- Possibly `delay_pct` or `on_time_pct`
- Other metrics (e.g. cancellations, diversions) if available

Since the dataset covers December 2019 & December 2020, the time span is short (2 months × 2 years), so many time-series techniques are limited but cross-year comparisons are possible.

1.3 Data quality notes / caveats:

- As a summarized dataset (aggregate by carrier / airport / date) rather than individual flight-level records, granularity is coarser.
- There may be missing carriers or airports in one year vs another (delays/cancellations in 2020 may distort comparability).
- Seasonal / external shocks (e.g. COVID in 2020) may heavily influence December 2020 delays — interpret with caution.

2. Operations Performed (adapted to delays data)

2.1 Data cleaning & exploration

- Parse `date` / `month` / `year` columns, convert to datetime.
- Check for duplicate rows (same carrier/airport/date).
- Validate numeric columns (`num_flights`, `delay_count`, `avg_delay`).
- Create derived columns: e.g. `delay_pct = delay_count / num_flights * 100`.

2.2 Descriptive analytics & visuals

- Carrier-wise delay share (bar / pie charts).
- Airport (origin / destination) delay distribution.
- Month (or date) comparison: December 2019 vs December 2020 (side-by-side).
- Histogram / boxplot of `avg_delay` across carriers / airports.
- Delay percentage distribution (scatter / boxplot).
- Time-series comparisons (if the data has daily breakdown within December) — delay counts / avg delays over dates.

2.3 Relationship / deeper analysis

- Correlate `num_flights` and `delay_count` (do airlines with more flights have higher delay counts?).
- Compare `avg_delay` across carriers (which airlines have higher mean delays).
- Year-over-year change: 2019 → 2020: difference in average delay, delay_pct for each carrier / airport.
- Regression / association: can `num_flights` predict `delay_count` or `avg_delay` (linear model).
- If airport-level data included: compare delays by origin/destination (airport effects).

3. Key Insights (example outcomes you might find)

Because I don't have the dataset fully here, these are plausible insights you could get when you run the code:

- Some carriers consistently have higher average delays (e.g. Carrier A has avg_delay ~30 min vs Carrier B ~15 min).
- In December 2020, delays are worse across many carriers compared to December 2019 (possibly due to COVID-operational disruptions).
- Delay rate (delay_pct) is positively correlated with number of flights: airlines with more flights tend to have a higher absolute number of delays, though not always higher percentage delay.
- Certain airports may show systematically higher delays across both years (e.g. hub airports).
- The distribution of average delays (per carrier-airport) shows skewness: many low-delay observations, some high-delay outliers.
- The year-over-year increases in delays may vary significantly by carrier and by airport region.

4. Recommendations (for airline / operations / analytics)

4.1 Operational / airline management

- Identify carriers or airports with unusually high delays and deploy targeted operational fixes (staffing, scheduling buffer).
- In 2020 (pandemic-impacted year), evaluate whether structural disruptions (e.g. traffic restrictions, resource cuts) drove higher delays — adjust contingency planning accordingly.
- Use delay_pct trends to benchmark carriers and airports for performance metrics.

4.2 Performance monitoring & SLAs

- Set service level targets (e.g. 90% of flights on-time).
- Monitor deviations (carrier vs target) and implement penalties / incentives based on performance.

4.3 Analytics & forecasting

- Use the aggregated delay data as target variables — build predictive models (e.g. forecast monthly or daily delay_count or avg_delay) if you can get finer-grained data.

- Combine with external data (weather, traffic, airport congestion) to explain delays.
- Use clustering of airports or carriers based on delay patterns to group “high-risk” vs “low-risk” entities.

4.4 Data enrichment and further data collection

- Acquire or merge flight-level data (individual flight delays) for higher-resolution modeling.
- Combine with weather, seasonal holiday, air traffic control, or operational data to explain delay variability.
- Extend the time coverage beyond December months to generalize insights.

5. Conclusion

This *Airline Delays* dataset offers a high-level snapshot of delays per carrier / airport in two months (2019 vs 2020). While coverage is limited, it allows comparative analysis, identification of high-delay carriers or airports, and basis for richer modeling once you bring in flight-level or external data. The patterns (e.g. higher average delays in 2020) may reflect systemic operational stress, making the dataset useful for performance diagnostics and planning.