Data Visualization on Airlines Data

Data Quality

This report summarizes the results of data quality checks performed on the Flights, Tickets, and Airports datasets. The goal is to identify potential issues such as missing values, invalid entries, and duplicates that may affect downstream analysis or modeling.

Flights Dataset

- Missing values in key columns: DEP_DELAY (50.3K), ARR_DELAY (55.9K), OCCUPANCY (310), DISTANCE (2.7K)
- Extreme departure delays: 14485 flights with delays less than -15 minutes
- Invalid distances: 230 flights with negative values
- Occupancy: No invalid values found

Tickets Dataset

- Missing and invalid fares: 17412 tickets with $0 fare
- Duplicates: 1,090,367 duplicate records based on route and trip type

Airports Dataset

- Missing IATA codes: 46,187
- Duplicate IATA codes: 46,306
- Airport type spread: Mostly small_airport and heliport, only 614 are large_airport

Key Recommendations

- Impute or remove missing and extreme delay values in the flight's dataset.
- Investigate and fix negative distances and zero ticket prices.
- Deduplicate ticket records and validate IATA_CODE uniqueness in the airport data.
- Consider filtering out non-commercial airport types (like heliport, closed) if focusing on passenger flight routes.

Data Cleaning

The clean_all_data() function prepares and cleans the Flights, Tickets, and Airports datasets for analysis. It ensures consistency, removes invalid or missing records, and aligns all three datasets using valid U.S. airport codes and logical data filters. It focuses on:

- Filtering to domestic, medium and large airports
- Removing cancelled, duplicate, and incomplete flights
- Cleaning ticket price and passenger values
- Standardizing all datasets with a common 'ROUTE' key for merging and comparison

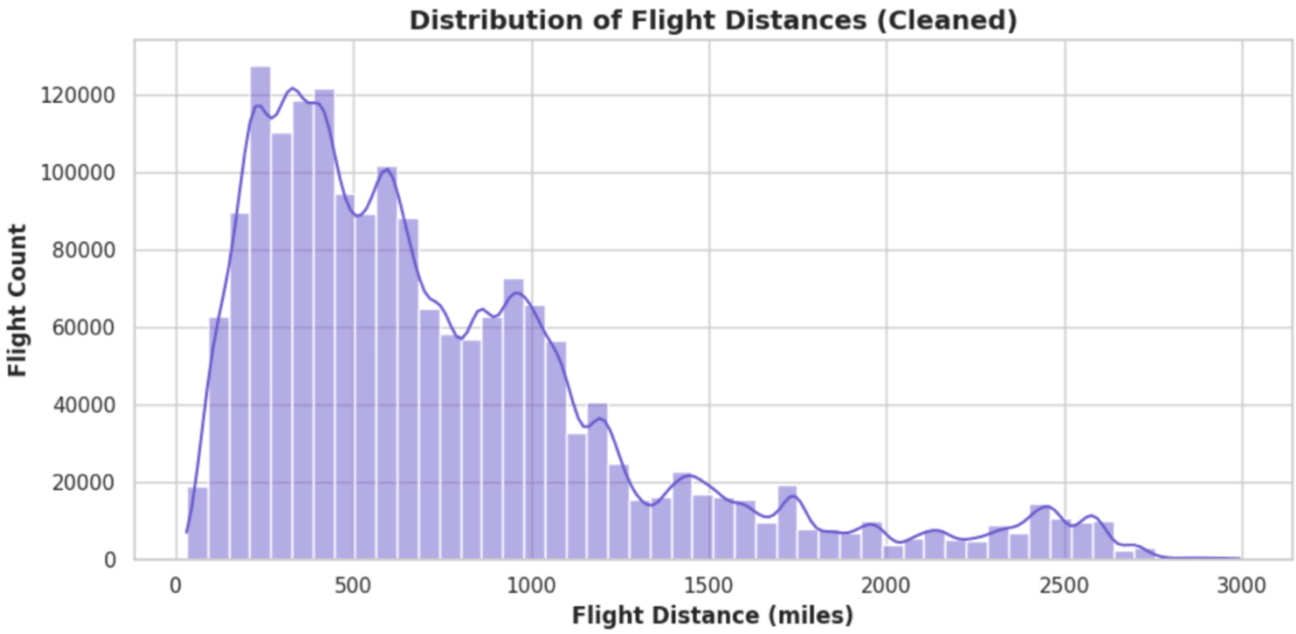
Summary

- All DISTANCE, OCCUPANCY, and delay values cleaned and capped
- Includes derived columns like ROUTE, WEEKDAY, and FLIGHT_DAY
- Excludes zero or missing prices and passengers
- Valid ROUTE column added for merging
- Only US-based, medium or large airports retained
- Non-commercial types and missing codes removed

Exploratory Data Analysis (EDA)

Performed some exploratory Data Analysis on the Airlines data and EDA is the process of analyzing and summarizing datasets using visual and statistical techniques to better understand their structure, patterns, relationships, and potential anomalies before applying advanced modeling or making decisions.

Graph 1: Distribution of Flight Distances (Histogram)



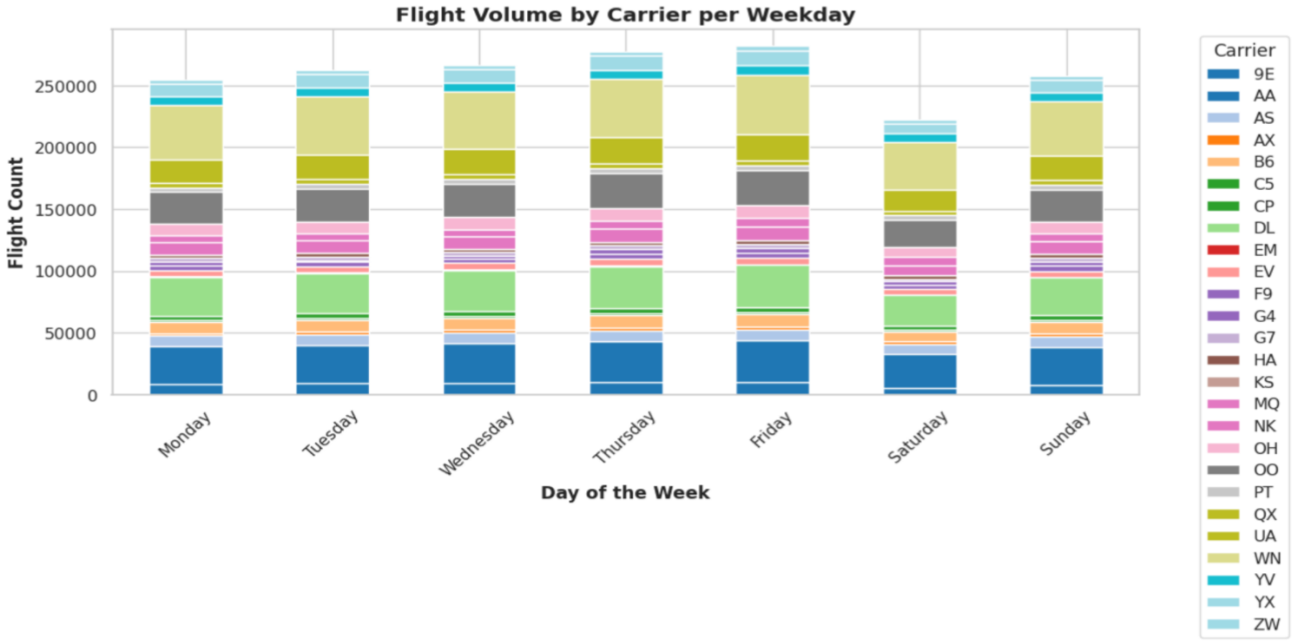
Data Interpretation

- Most flights fall between 200 to 1200 miles, indicating strong demand for short- to medium-haul routes.
- A sharp peak around 300–400 miles suggest highly frequent regional connections.
- Flight frequency steadily declines as distance increases, with relatively fewer long-haul flights beyond 1500 miles.
- This distribution confirms the operational focus on domestic and mid-range routes, ideal for optimizing fleet utilization and cost efficiency.

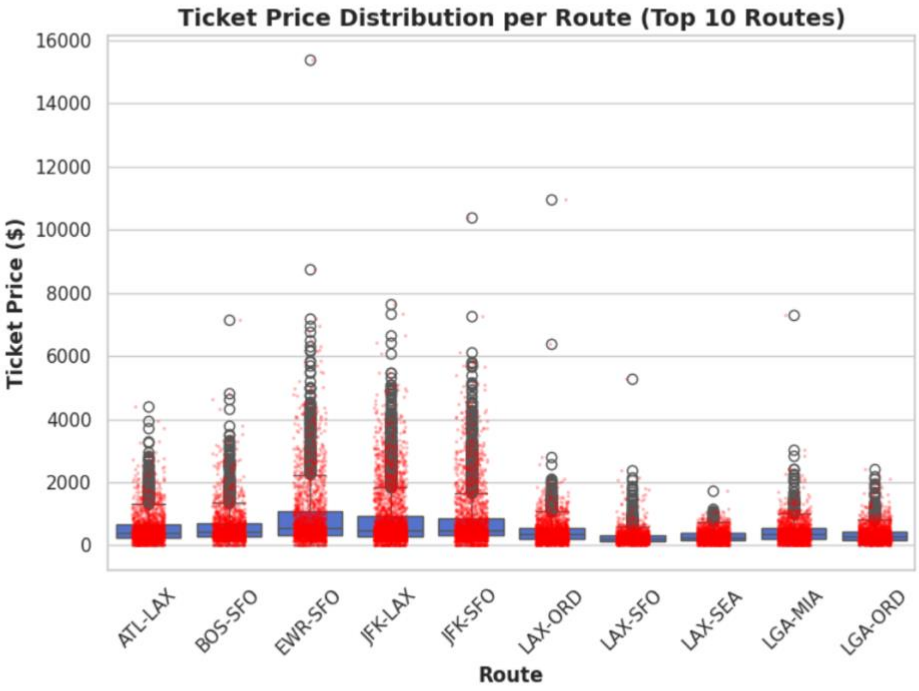
Graph 2: Flight Volume by Carrier per Weekday

Data Interpretation

- Flight activity is highest on weekdays, especially Monday to Friday, aligning with typical business travel patterns.
- Major carriers operate consistently across all days, indicating a broad route network and regular demand.
- Smaller or regional airlines show variable presence, often focusing on specific days or routes.
- Weekend flight volumes are noticeably lower, likely reflecting reduced business travel and selective leisure routes.
- The chart helps identify which airlines are dominant on specific days, aiding scheduling and partnership strategies.



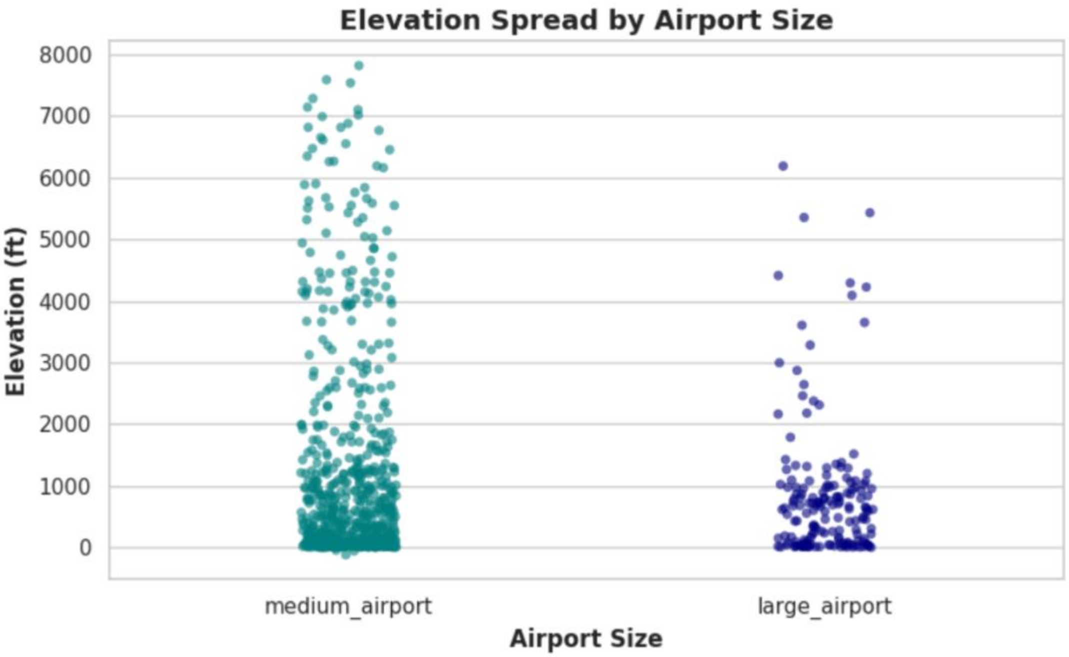
Graph 3: Ticket Price Distribution per Route (Box Plot)



Data Interpretation

- This graph reveals how ticket prices vary across the busiest domestic routes. Several routes, such as JFK-LAX and EWR-SFO, show a wide price range with many high-value outliers, indicating dynamic pricing or premium cabin sales.
- Routes like LAX-SEA and LAX-ORD have lower and more consistent pricing, suggesting tighter fare bands and fewer outliers.
- The dense red dots show that most tickets are clustered below $1000, with only a small portion priced at extremes.
- This distribution helps identify pricing volatility and revenue potential by route useful for yield management and route planning.

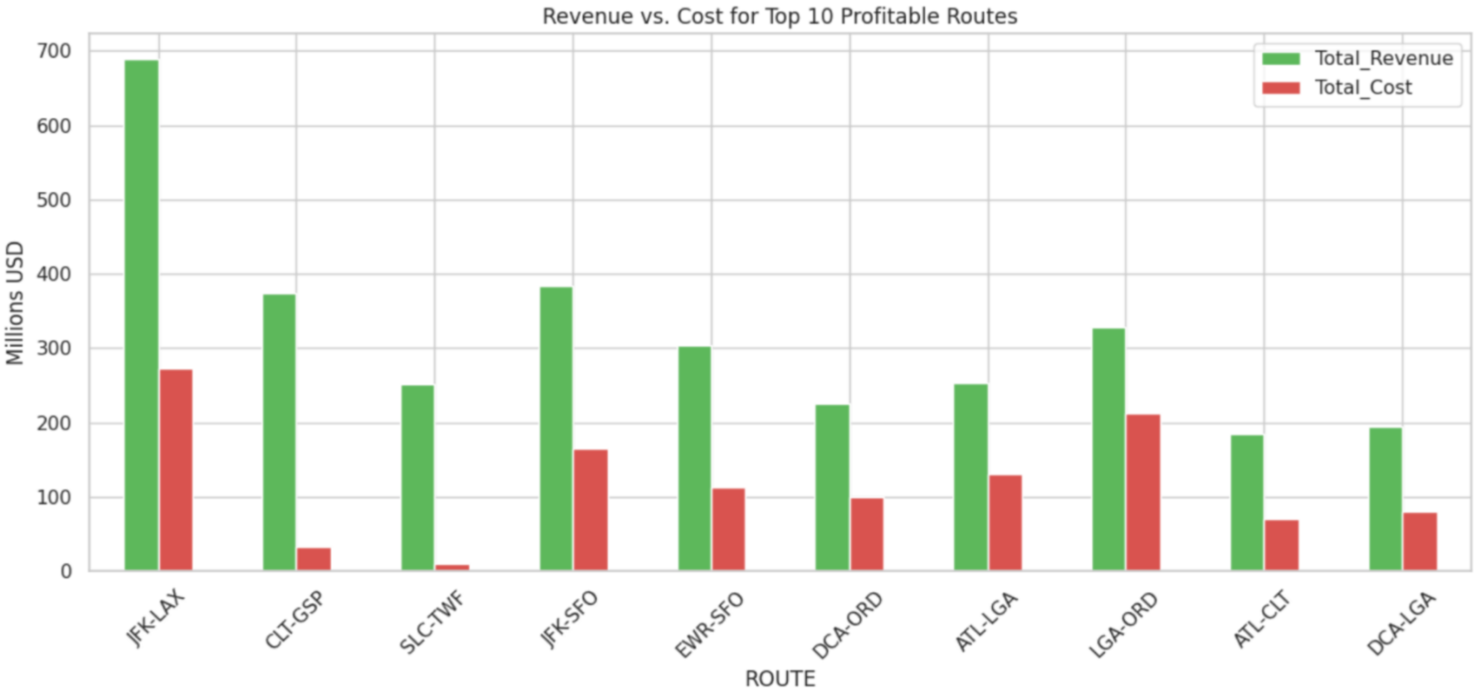
Graph 4: Elevation Spread by Airport Size (Strip Plot or Jittered Scatter Plot)



Data Interpretation

- Medium airports show a wide elevation range, with several located at high altitudes (>5000 ft), reflecting geographic diversity across the U.S.
- Large airports are mostly concentrated at lower elevations (<1500 ft), likely due to proximity to major cities and favorable operating conditions.
- Useful for assessing infrastructure resilience, aircraft performance needs, and strategic airport development across elevation zones.

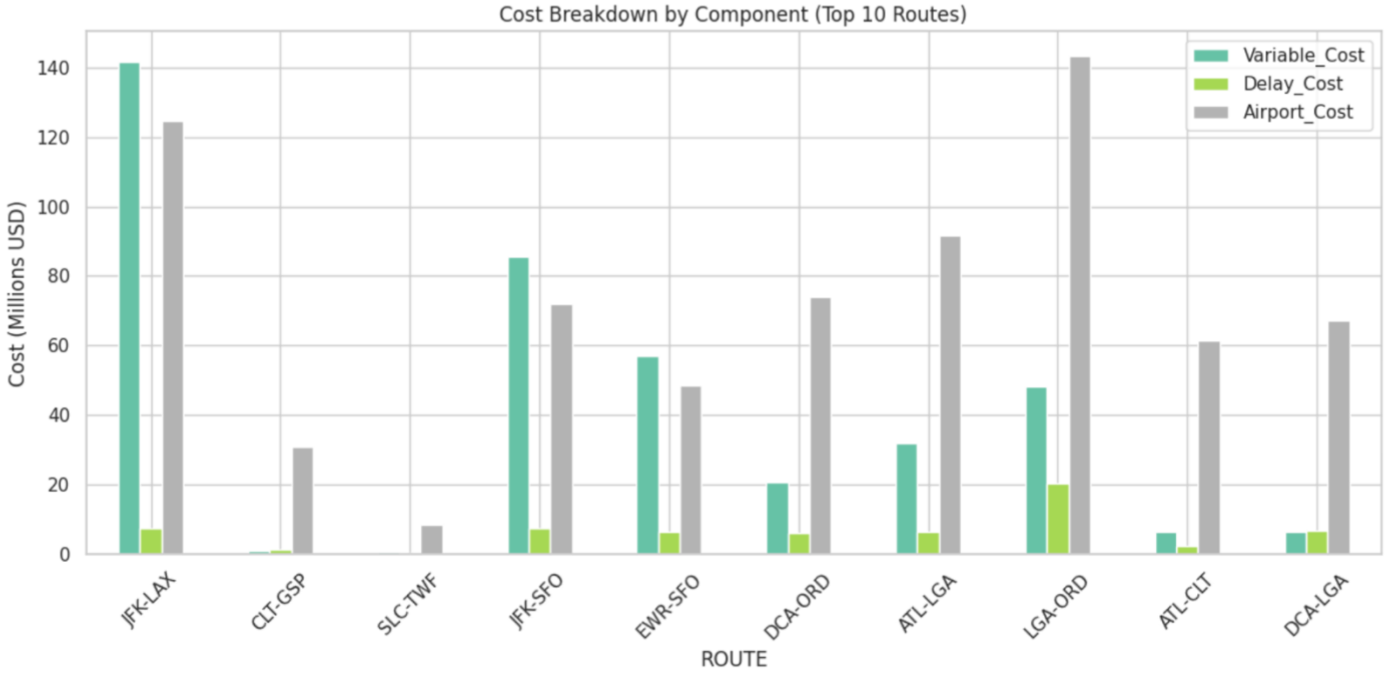
Graph 5: Revenue vs Cost per Route (Stacked Bar Chart)



Data Interpretation

To visually compare revenue and cost side-by-side for each of the top 10 most profitable round trip routes. This helps you understand: Which routes generate the most revenue, Which routes are cost-heavy and how much profit margin exists for each route.

Graph 6 : Breakdown of Cost Components (Line Plot)



Data Interpretation

To help understand which types of costs dominate across the top 10 most profitable round trip routes.

- If Variable Cost dominates - Optimize aircraft efficiency, route length, or fuel use.
- If Delay Cost dominates - Target operational efficiency or airport congestion issues.
- If Airport Cost dominates - Rethink airport choices or negotiate fees

Reference:

- All the graphs are referenced through attached ipynb file