Airline Data Case Study Metadata

Author – Soumya Bhandari

1. Project Purpose

The objective of this case study is to analyze Q1 2019 U.S. domestic airline data to identify the top 5 round-trip routes for investment. The analysis includes route profitability, operational performance, break-even modeling, and KPI tracking. The codebase supports a data-driven investment decision using reproducible Python functions and visual analytics.

2. Input Datasets

- **Flights.csv:** Flight-level operational data including delays, distance, occupancy
- **Tickets.csv**: Ticket fares and passenger estimates for each route
- **Airport_Codes.csv**: Airport metadata including IATA codes and size classification

3. Functional Modules

- **load_data()** – Loads CSVs into pandas DataFrames.
- **data_quality_checks()** – Performs null, duplicate, outlier, and data type validation across datasets.
- **clean_all_data()** – Filters Q1 data, removes canceled/invalid flights, handles airport types and distances.
- **calculate_busiest_routes()** – Identifies top 10 busiest round-trip routes by flight count.
- **calculate_profitability()** – Estimates profit based on ticket, baggage revenue and cost structure.
- **recommend_top_investment_routes()** – Ranks routes using a weighted scoring system based on multiple KPIs.
- **calculate_break_even_roundtrips()** – Computes the number of trips to recover a $90M aircraft cost.

5. Metadata: New Columns created during the analysis

	Key / Column Name	Description
	ROUTE	A directional route string in format ORIGIN-DEST
	ROUNDTRIP_ROUTE	A sorted combination of ORIGIN and DEST to represent bidirectional round trips
	Passengers	Estimated number of passengers on a flight (OCCUPANCY * 200)
	Ticket_Revenue	Revenue from ticket sales (Passengers * Ticket Price)
	Baggage_Revenue	Revenue from baggage fees (Passengers * 50% * $35)
	Total_Revenue	Total revenue from a flight = Ticket + Baggage revenue
	Variable_Cost	Cost based on distance (DISTANCE * [$8 fuel + $1.18 maintenance)
	Airport_Cost	Fixed airport usage costs based on airport size (Medium=$5K, Large=$10K per leg)
	Delay_Cost	Penalty for delays beyond 15 minutes ($75 per extra minute of DEP or ARR delay)
	Total_Cost	Sum of all costs: Variable + Airport + Delay
	Profit	Net income per flight: Total_Revenue - Total_Cost
	ON_TIME_DEP	Binary flag (1/0) showing if a flight departed within 15 minutes of schedule.
	ON_TIME_ARR	Binary flag (1/0) showing if a flight arrived within 15 minutes of schedule
	BreakEven_RoundTrips	Number of round trips required to recover $90M aircraft cost (90M / Profit per trip)
	score	Composite score used for route recommendation, weighted by KPIs
	TOTAL_ROUNDTRIP_FLIGHTS	Total number of round-trip flights on the route (one round trip = two one-way flights).
	AVG_DELAY_COST	Average cost per flight incurred due to departure and arrival delays.
	AVG_REVENUE_TO_COST_RATIO	Average ratio of revenue to cost per flight. Indicates how efficiently revenue is being generated.
	ON_TIME_DEP_PCT	Percentage of flights on the route that departed on time.
	ON_TIME_ARR_PCT	Percentage of flights on the route that arrived on time.
	COMBINED_ON_TIME_PCT	Average of departure and arrival punctuality rates, reflecting overall on-time reliability.

5. Outputs & Visuals

- Data Quality checks - Code
- Data Cleaning code – Code + EDA
- Top 10 Busiest Routes – Table + Horizontal bar plot
- Top 10 Most Profitable Routes – Table + Bar and Pie plots (also has 2 other graphs on the cost and revenue)
- Investment Recommendations – Table + Bubble chart
- Break-Even Analysis – Table + Horizontal bar plot
- KPI recommendation for dashboard

6. Business Assumptions

- Each plane services one round trip route
- Aircraft cost = $90 million (one-time fixed investment)
- Max capacity = 200 passengers per flight; 50% check bags
- Delay cost = $75/minute after first 15 minutes
- Baggage fee = $35 per bag
- Routes > 3000 miles are excluded (domestic filter)

7. Completion Summary

- Data Load & Quality Checks
- Data Cleaning & Processing
- Exploratory Data Analysis on Clean Data
- Busiest Route Identification
- Revenue & Cost Modeling
- Scoring-based Recommendations
- Break-even Profitability Analysis
- KPI Recommendations