AIRLINE DATA CHALLENGE - CapitalOne

Rinku Kalsi

1. Introduction

Objective:
Analyze the provided Airline, Tickets, and Flights datasets to identify the top 5 round-trip routes that are both highly profitable and operationally efficient. Perform a breakeven analysis on the selected routes and recommend key performance indicators (KPIs) that can be used to track and measure their ongoing success.

2. Tools & Software

1. Google Colab (Python, Pandas, NumPy, Matplotlib)
2. Tableau Public
3. Github (data source): https://github.com/CapitalOneRecruiting/DA-Airline-Data-Challenge/tree/main

3. Data Loading & Preprocessing

Datasets Used:

1. Airline
2. Tickets
3. Flights

Initial filtering based on instructions in problem statement

1. **Tickets:** Retained only roundtrips
2. **Airports:** Included only medium and large U.S. airports
3. **Flights:** Excluded cancelled flights

Duplicates & Nulls:

1. Removed duplicates from *Tickets* and *Flights* datasets, as these df had duplicate data.

2. Addressed null values across all datasets:

Airports*:*

Dropped ELEVATION_FT, CONTINENT, MUNICIPALITY, IATA_CODE (rows with nulls)

Tickets*:*

Cleaned PASSENGERS and ITIN_FARE (handled as strings with $, converted to float)

Flights*:*

Cleaned TAIL_NUM, DEP_DELAY, ARR_DELAY, AIR_TIME, DISTANCE, OCCUPANCY_RATE

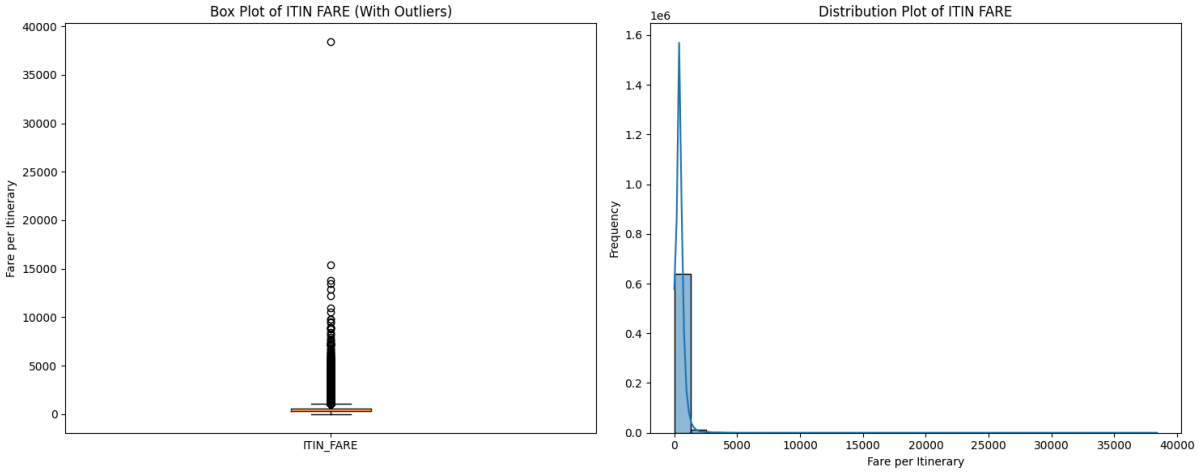
4. Dataset-Specific Processing

Airports Dataset

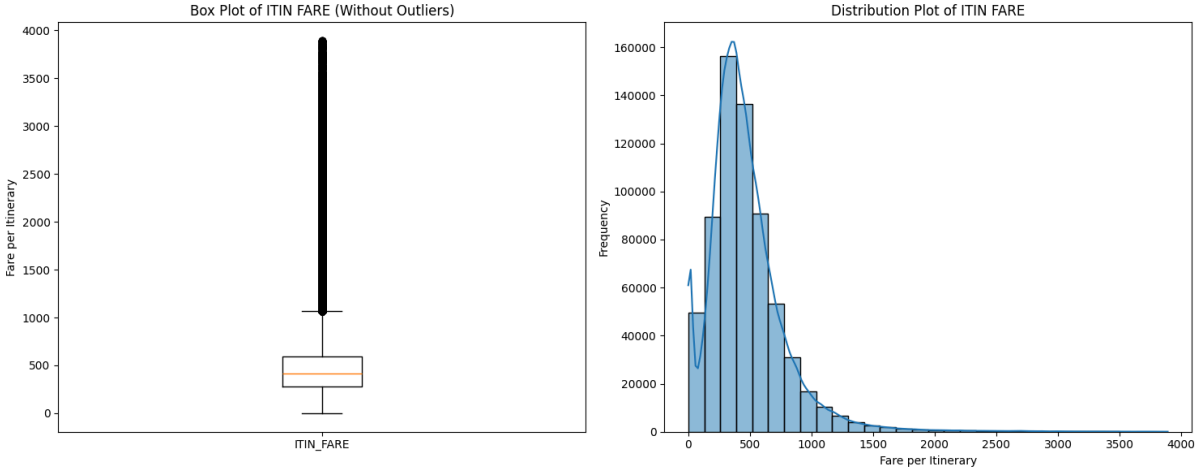
1. Dropped: CONTINENT, ISO_COUNTRY, ELEVATION_FT
2. Extracted Latitude and Longitude from COORDINATES for visualization.
3. Dropped rows with null IATA_CODE

Tickets Dataset

1. Ensured data pertains to Q1 2019
2. Created ROUTE (standardized roundtrips like A-B = B-A)
3. Cleaned and converted ITIN_FARE from string to float
4. Imputed null PASSENGERS and ITIN_FARE using group-wise average by ROUTE
5. Removed rows with non-positive fare or passenger counts
6. Identified and removed outliers using IQR on ITIN_FARE
7. Created WEIGHTED_AVG_FARE for each route



Plotted box plot and histogram of ITIN_FARE to identify outliers, further used IQR (Inter quartile range) to filter out the outliers.

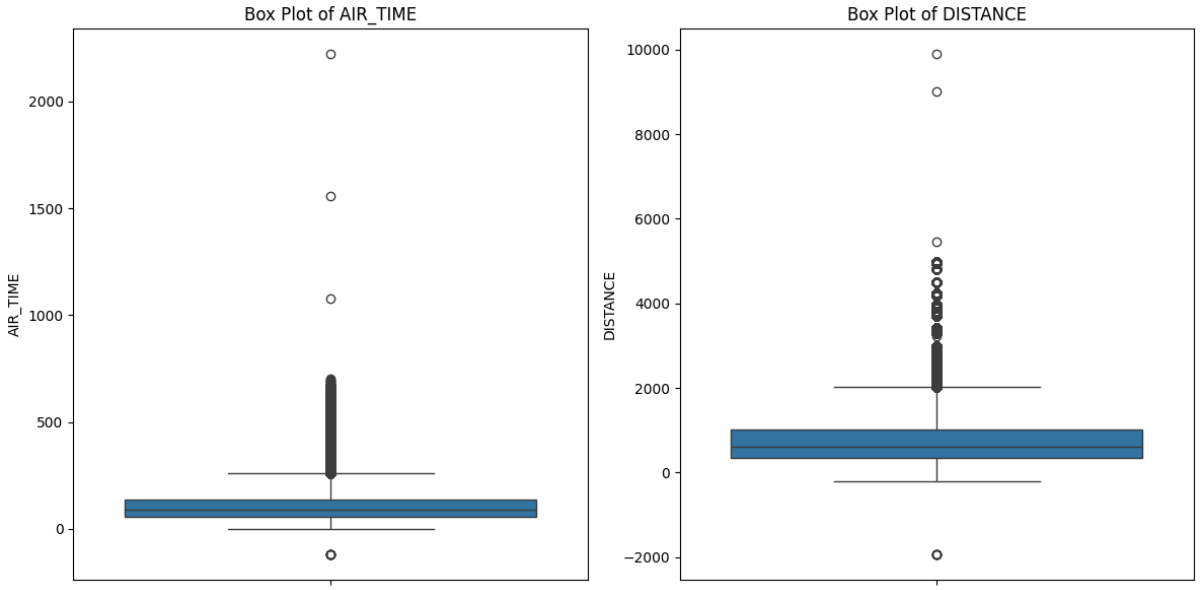


1. Using IQR to filter rows improved the distribution of the dataset.
2. Grouped the data by ROUTE to calculate the total WEIGHTED_FARE and total number of passengers.
3. Computed the **weighted average fare** per route to account for passenger volume, offering a more accurate ticket price estimate than a simple average.
4. Since each row represents a roundtrip, the weighted average fare was divided by two to reflect the **one-way fare** per trip.

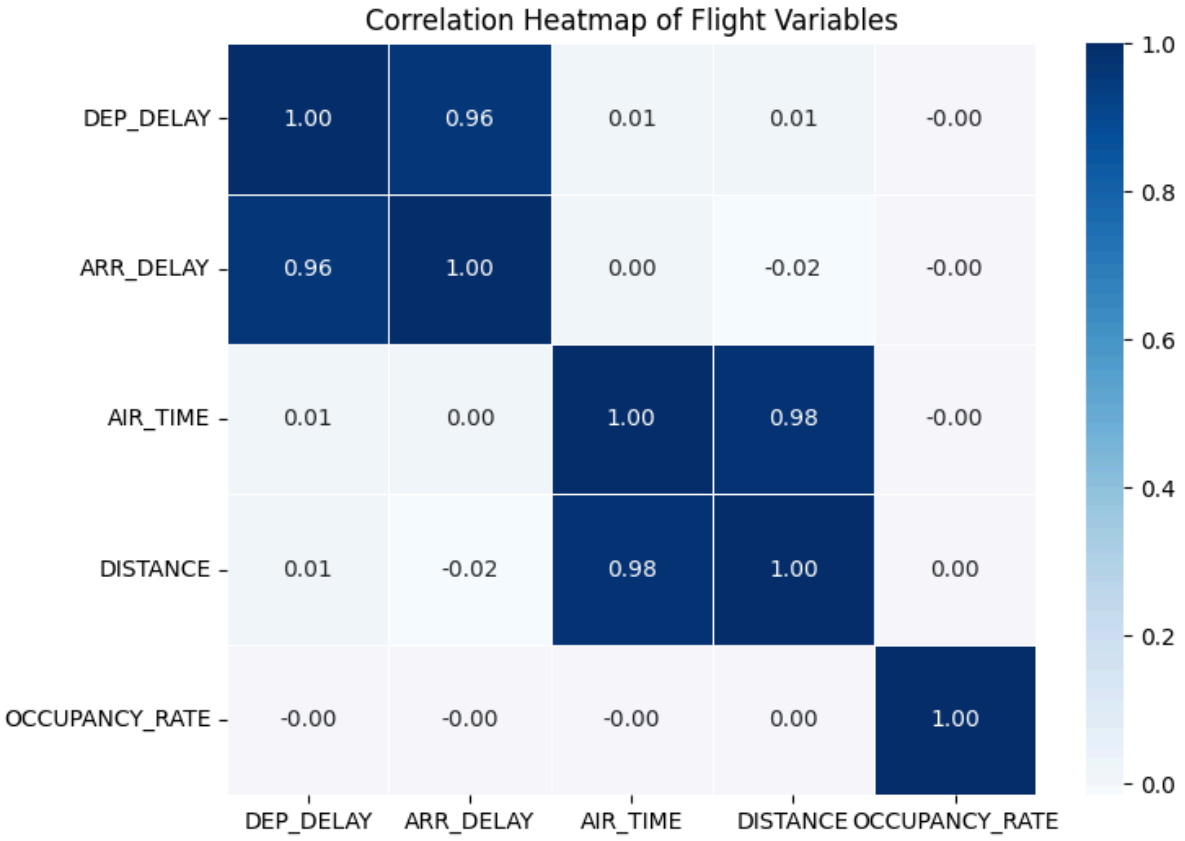
Flights Dataset

1. Standardized date formats to yyyy-mm-dd
2. Converted string values like “two”, “hundred”, and symbols ($, *) using word2number
3. Removed rows with negative or 0 values for AIR_TIME and DISTANCE
4. Analyzed correlations using heatmap
   Strong correlations: ARR_DELAY ↔ DEP_DELAY, AIR_TIME ↔ DISTANCE
5. Used KNN Imputer for missing values
6. Created ROUTE column for downstream joins

Plotted a box plot of AIR_TIME and DISTANCE to identify any outliers,



box plot shows the data in both columns has negative values



Plotted correlation Matrix

7. Looked at how AIR_TIME, DISTANCE, ARR_DELAY, DEP_DELAY, and OCCUPANCY_RATE are related by creating a heatmap.
8. The heatmap showed that arrival and departure delays are closely related, and so are air time and distance.
9. Occupancy rate didn't show much connection to the others.
10. Used these strong relationships to fill in missing values using a KNN imputer, which fills gaps based on similar rows.
11. After handling the missing data, added a ROUTE column to help merge this dataset with the others.

5. Data Merging (Munging)

To work with complete and consistent data, I used **INNER joins** across all datasets. This ensured only valid and matching records were included.

1. First, merged the **Airports** and **Flights** datasets using IATA_CODE from Airports and ORIGIN from Flights to get the **origin airport name and coordinates**.

2. Then, performed another join using IATA_CODE and DESTINATION to fetch the **destination airport details** (name, latitude, and longitude).

3. After that, **dropped unnecessary columns** like 'MUNICIPALITY', 'IATA_CODE', and 'COORDINATES' to keep the data clean.

4. Finally, joined this Flights-Airports dataset with the **Tickets** dataset using the standardized ROUTE column I had created earlier.

6. Data Analysis & Insights

Busiest Routes:

1. **Step 1:** Counted the number of flights operating on each **ROUTE** (combination of origin and destination airports).

2. **Step 2:** Merged this data with airport metadata to include geographic and airport-level information for both origin and destination.

3. **Step 3:** Extracted the **Top 10 routes** based on flight frequency, identifying the busiest air corridors in the dataset.

Most Profitable Routes:

Assumptions Used for Cost & Revenue Estimation:

1. Operational Costs:

- **Airport Fees (per leg):**

a. *Medium airports:* $2,500

b. *Large airports:* $5,000
(Assuming roundtrip costs are split equally per leg)

- **Operations & Maintenance (O&M):**

a. *Fuel, crew, etc.:* $8 per mile

b. *Insurance, depreciation, etc.:* $1.18 per mile

- **Delay Costs:**

a. First 15 minutes of delay are free

b. Beyond that: $75 per additional minute

2. Revenue Components:

- **Ticket Fare:** Calculated using the **weighted average fare** based on passenger count

- **Baggage Revenue:**

a. Assumed 50% of passengers carry **1 bag**

b. $35 per checked bag

- **Passenger Estimate:**

a. Total passengers = Occupancy Rate × 200 seats

Steps for Profitability Analysis:

1. **Calculated Total Revenue:**

a. TICKET_REVENUE + BAGGAGE_REVENUE

2. **Calculated Total Costs:**

a. OPERATIONAL_COST + O&M_COST + DELAY_COST

3. **Derived Net Profit:**

a. NET_PROFIT = TOTAL_REVENUE - TOTAL_COSTS

4. **Aggregated by ROUTE** to calculate total profit per route.

5. **Sorted routes** by descending net profit to identify the **Top 10 most profitable routes**.

Recommended KPIs to Track the Success of New Routes:

1. **Profit per Mile**
Formula: NET_PROFIT / DISTANCE

a. Measures how efficiently a route generates profit relative to the distance traveled.

b. Useful for comparing profitability between short-haul and long-haul routes.

2. **Profit per Roundtrip**
Formula: NET_PROFIT / NUMBER_OF_ROUNDTRIPS

a. Indicates the average profit earned from each complete roundtrip.

b. Helps assess whether increasing flight frequency improves overall profitability.

3. **Delay Cost per Trip**
Formula: (ARR_DELAY_COST + DEP_DELAY_COST) / NUMBER_OF_ROUNDTRIPS

a. Evaluates the average cost incurred due to delays on a per-trip basis.

b. Highlights operational inefficiencies and helps pinpoint routes with recurring delay-related costs.

4. **Projected Lifetime Profit**
Formula: PROFIT_PER_ROUNDTRIP × BREAKEVEN_NUMBER_OF_TRIPS

a. Estimates the total potential profit if the current performance continues until the route reaches breakeven.

b. Helps forecast long-term financial returns and track actual performance post-launch.

5. **Profit-to-Cost Ratio**
Formula: NET_PROFIT / TOTAL_COST

a. Shows how much profit is generated for every dollar spent.

b. A higher ratio reflects stronger financial viability and operational efficiency.

KPIs to Track Performance:

KPI	Formula	Purpose
Profit per Mile	NET_PROFIT / DISTANCE	Compares profitability across route lengths
Profit per Roundtrip	NET_PROFIT / NUMBER_OF_ROUNDTRIPS	Evaluates per-trip profitability
Delay Cost per Trip	(ARR_DELAY_COST + DEP_DELAY_COST) / NUMBER_OF_ROUNDTRIPS	Highlights operational bottlenecks
Projected Lifetime Profit	PROFIT_PER_ROUNDTRIP × BREAKEVEN_TRIPS	Forecasts route revenue potential
Profit to Cost Ratio	NET_PROFIT / TOTAL_COST	Measures overall financial efficiency

7. Route Recommendations

1. In this step, used the merged dataset to calculate the **number of roundtrips** for each route by dividing the total **number of flights** by 2.
2. To ensure meaningful insights, I filtered out routes with fewer than 100 roundtrips, this threshold helps focus only on routes with enough data to evaluate their profitability reliably.
3. Next, I created two key metrics to guide route recommendations:
4. **Profit-to-Cost Ratio**: Net Profit divided by Total Cost, which highlights how efficiently each route generates value.
5. **Delay Cost per Trip**: Calculated as the sum of arrival and departure delay costs divided by the number of roundtrips, to capture the operational burden on each route.

6. These two metrics work together to strike a balance between **financial performance** and **service quality**, a balance that reflects the company's motto, ***"On time, for you."***
7. Since the two metrics are on different scales, normalized them to a 0–1 range.
8. Further, calculated a **Profitability Score** by assigning 70% weight to the Profit-to-Cost Ratio and 30% to the inverse of Delay Cost per Trip.
9. Finally, sorted the routes based on this score and selected the **top 5** as the most promising recommendations.

8. Break Even Analysis & Profit

Profit per Roundtrip

Formula: NET_PROFIT / NUMBER_OF_ROUNDTRIPS

1. This KPI indicates how much net profit is earned from a single roundtrip on a given route.

2. It helps evaluate the financial return per complete operational cycle and is key to understanding the impact of increasing trip frequency.

Breakeven Analysis (Based on $90M Initial Investment)

Formula: BREAKEVEN_TRIPS = 90,000,000 / PROFIT_PER_ROUNDTRIP

1. Assuming an upfront investment of $90 million, this calculation estimates the number of roundtrips required for the investment to be recovered.
2. It provides a clear target for how many trips a route must complete at current profitability levels to reach breakeven.
3. This insight supports long-term route planning and helps assess whether projected trip volumes are sufficient to justify the investment.

9. Data Visualization

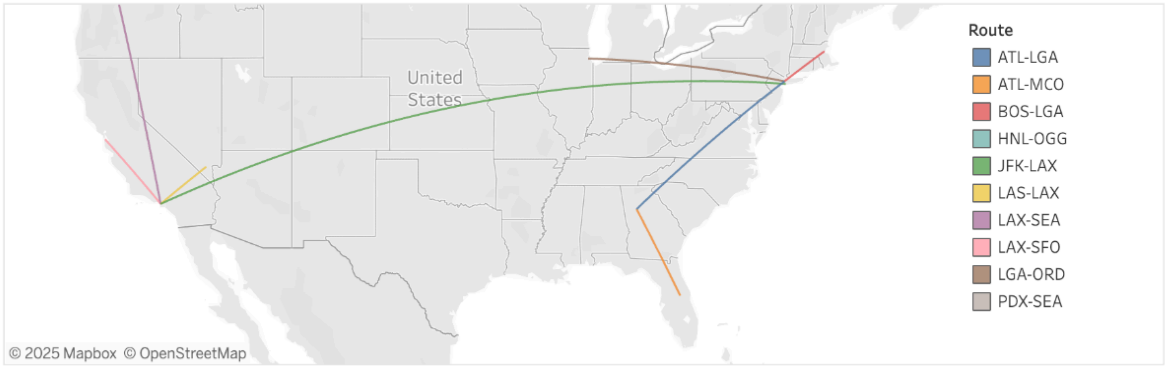
Built a Tableau dashboard to illustrate:

1. Busiest routes (Interactive Map and graphical summary)
2. Most profitable routes (Interactive Map and graphical summary)
3. Top 5 recommended routes (Interactive Map and graphical summary)

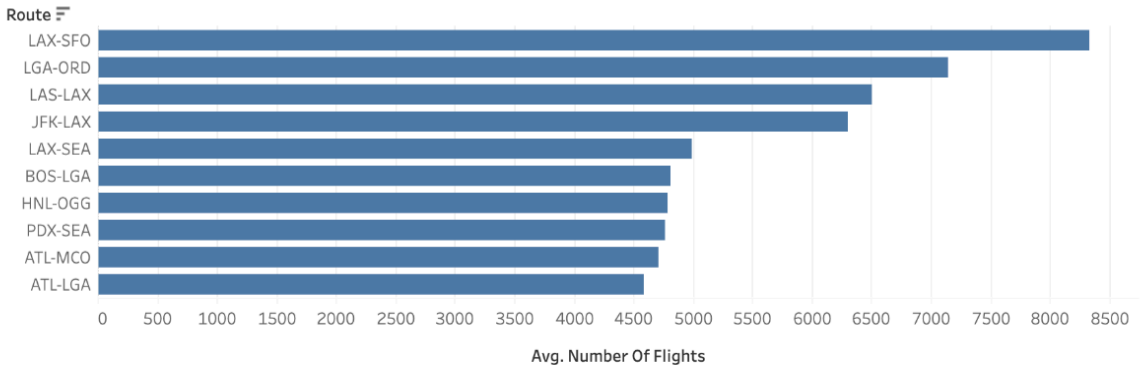
TABLEAU PUBLIC : DATA VISUALIZATION LINK

Dashboard 1 - Based on Operation and Busy factor

Top 10 busiest routes in Q1, 2019

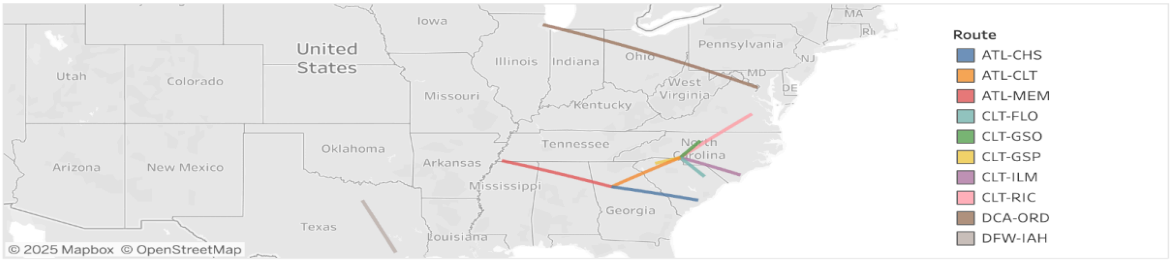


Busiest Routes by Number of Flights

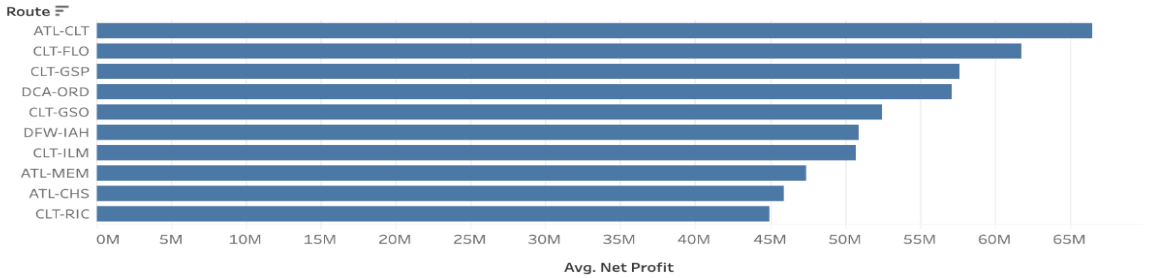


Dashboard 2 - Profitability

Top 10 most profitable routes in Q1, 2019

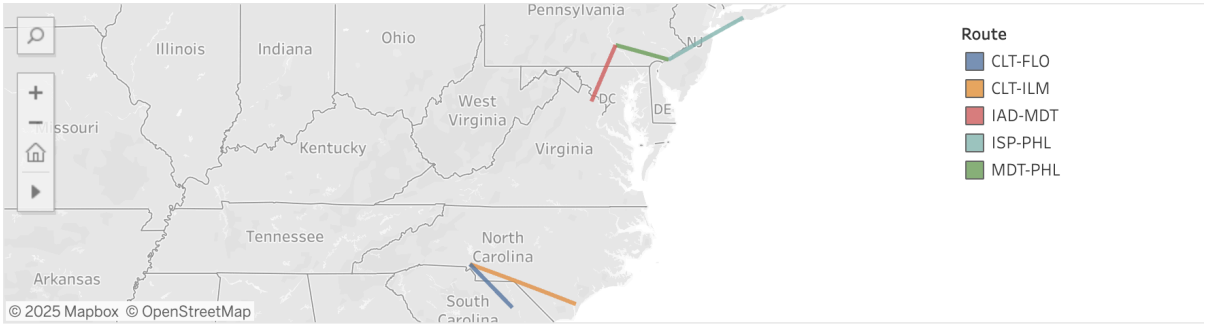


Profitable Routes



View on Tableau Public

Recommended Routes Map



Recommendations based on Profit cost ratio and Profit per roundtrip

Route	Delay Cost Per Trip	Net Profit	Profit Cost Ratio	Profit Per Roundtrip	Total Cost
CLT-FLO	1,779	61,798,963	13	245,234	4,686,395
CLT-ILM	1,450	50,710,538	3	69,277	14,536,609
IAD-MDT	2,595	19,834,534	4	72,921	5,263,766
ISP-PHL	2,084	12,964,406	4	78,572	3,221,315
MDT-PHL	2,006	41,285,571	6	103,994	7,363,469

10. If I Had More Time, Resources, or Data...

1. Analyze seasonality in ticket prices to optimize fare strategies across the year.
2. Incorporate weather data to assess its impact on delays and operational costs.
3. Evaluate multi-stop routes to explore additional revenue opportunities.
4. Segment passengers by fare class to understand revenue distribution and optimize pricing.
5. Simulate demand scenarios using historical trends or market conditions.
6. Build real-time dashboards to monitor KPIs and enable faster decision-making.

11. Metadata Summary

Column	Description
Latitude	Latitude from airport coordinates
Longitude	Longitude from airport coordinates

ROUTE	Combination of ORIGIN and DESTINATION codes
WEIGHTED_FARE	ITIN_FARE × PASSENGERS
WEIGHTED_AVG_FARE	Weighted average fare per passenger per route
OPERATIONAL_COST	Fixed cost based on airport size/type
N_PASSENGERS	Estimated passengers per flight (Occupancy Rate × 200)
BAGGAGE_REVENUE	Revenue from baggage fees per trip
TICKET_REVENUE	Revenue from ticket sales per trip
TOTAL_REVENUE	BAGGAGE_REVENUE + TICKET_REVENUE
DEP_DELAY_COST / ARR_DELAY_COST	Delay cost beyond 15 minutes at $75/min
OM_COST	Cost for fuel, crew, maintenance, insurance, depreciation
TOTAL_COST	Sum of OM cost, delay cost, and operational cost
NET_PROFIT	TOTAL_REVENUE - TOTAL_COST
NUMBER_OF_ROUNDTRIPS	Total flights divided by 2
PROFIT_COST_RATIO	Profit earned per dollar of cost
DELAY_COST_PER_TRIP	Average delay cost per roundtrip
PROFITABILITY_SCORE	Combined normalized score of profit and delays
PROFIT_PER_ROUNDTRIP	NET_PROFIT / NUMBER_OF_ROUNDTRIPS
BREAKEVEN_NUMBER_OF_TRIPS	Trips needed to recover $90M investment

12. New Columns Generated during analysis:

1. **ROUTE**: Combined ORIGIN and DESTINATION codes
2. **OPERATIONAL_COST**: Based on airport type (medium: $2.5K, large: $5K per leg)

3. **N_PASSENGERS**: Estimated as Occupancy Rate × 200
4. **BAGGAGE_REVENUE**: 50% × N_PASSENGERS × $35
5. **TICKET_REVENUE**: N_PASSENGERS × WEIGHTED_AVG_FARE
6. **TOTAL_REVENUE**: Sum of ticket and baggage revenue
7. **DEP_DELAY_COST & ARR_DELAY_COST**: $75/min if delay > 15 minutes
8. **OM_COST**: DISTANCE × (8 + 1.18) for fuel, crew, etc.
9. **TOTAL_COST**: Sum of OM, delay, and operational cost
10. **NET_PROFIT**: TOTAL_REVENUE - TOTAL_COST
11. **NUMBER_OF_ROUNDTRIPS**: Total flights divided by 2
12. **PROFIT_COST_RATIO**: NET_PROFIT / TOTAL_COST
13. **DELAY_COST_PER_TRIP**: Total delay cost per roundtrip
14. **PROFITABILITY_SCORE**: Normalized score combining profit and delay cost
15. **PROFIT_PER_ROUNDTRIP**: NET_PROFIT / NUMBER_OF_ROUNDTRIPS
16. **BREAKEVEN_NUMBER_OF_TRIPS**: 90M / PROFIT_PER_ROUNDTRIP