

The Data Detective: Uncovering the Single Most Important Story in AirwaysBharath's Flight Data

Client: AirwaysBharath



Date: September 21, 2025

Prepared & Presented By: Gajarajan V Y






Table of Contents

1. Executive Summary
2. Analysis Workflow
3. Top 5 Key Insights
 - 3.1 Insight 1
 - 3.2 Insight 2
 - 3.3 Insight 3
 - 3.4 Insight 4
 - 3.5 Insight 5
4. Additional NUM-NUM, NUM-CAT, CAT-CAT, Multivariate Analysis Findings
5. Recommendations & Next Steps
6. Appendix
 - 6.1 Full Data Tables – Variables Understanding
 - 6.2 All Feature-Engineered Variables Understanding

1 Executive Summary

This project, titled "The Data Detective: Uncovering the Single Most Important Story in AirwaysBharath's Flight Data," analyzes a comprehensive airline dataset to derive actionable insights for improving profitability, customer retention, and operational efficiency. Prepared by Gajarajan V Y on September 21, 2025, the analysis employed advanced techniques including data cleaning with MICE (LightGBM) (Multiple Imputation by Chained Equations) imputation, NLP for sentiment scoring, statistical hypothesis tests like Kruskal–Wallis Test, Mann–Whitney U Test, Chi-Square Test for Independence (Contingency Test), statistical modeling via linear regression and Cramér's V, and multivariate methods like PCA and MCA.

Starting with 44 variables, the analysis applied rigorous **bivariate** statistical testing to uncover 25 significant numerical-numerical relationships, 40 numerical-categorical, 14 categorical-categorical, and 4 time-metric pairs. A subsequent **multivariate analysis** revealed 54 statistically significant patterns. From this **exhaustive process**, the **top five insights** were selected for their strength, relevance, and interpretability across both bivariate and multivariate dimensions.

Lever	Impact
 Customer Retention Risk	+255% profit per 1% risk increase, -223% costs
 High-Tier Customers	40% higher revenue, 12% cost savings
 Customer Sentiment	+39.56% conversion gains, boosted revenue
 Influential Merchants/Regions	32.82% higher route dynamics
 Flight Operations/Pricing	29.52% complexity cut, 22.12% pricing agility, 12.01% booking lift

Key Insights:

1. **Retention Risk Management:** Higher retention risk correlates with +255% profitability and -223% costs, offering high-reward opportunities if mitigated through targeted loyalty programs.
2. **Customer Tier Optimization:** High-tier customers drive ~40% more revenue and ~12% cost savings compared to medium-tier, while low-tier segments reduce margins; focus on upselling premiums.
3. **Sentiment Enhancement:** A 1% sentiment improvement yields up to +39.56% in conversion and revenue, emphasizing real-time analytics for feedback response.
4. **Route and Merchant Targeting:** Influential merchants and regions boost route dynamics by ~32.82%, recommending capacity expansion on high-value routes like MEL–LHR and HKG–FRA.
5. **Operational Streamlining:** Simplifying operations reduces complexity by ~29.52%, enhances pricing agility by ~22.12%, and increases bookings by ~12.01%, via layover reductions and dynamic pricing.

Additional Findings:

- Numerical-numerical elasticities highlight strong profitability lifts from fares and retention.
- Numerical-categorical swings show customer segments impacting revenue by $\pm 40\text{-}48\%$.
- Categorical-categorical associations are mostly independent, except weak airport-region links.
- Multivariate pillars focus on sentiment, routes, and operations for high-leverage gains.

Recommendations & Next Steps:

Implement three strategic initiatives: (1) Pilot retention for high-risk segments; (2) Upsell premiums and expand flagship routes; (3) Streamline operations with dynamic pricing. A 9-month rollout plan includes setup, pilots, and scaling, monitored by metrics like retention uplift, sentiment scores, load factors, and booking dynamics.

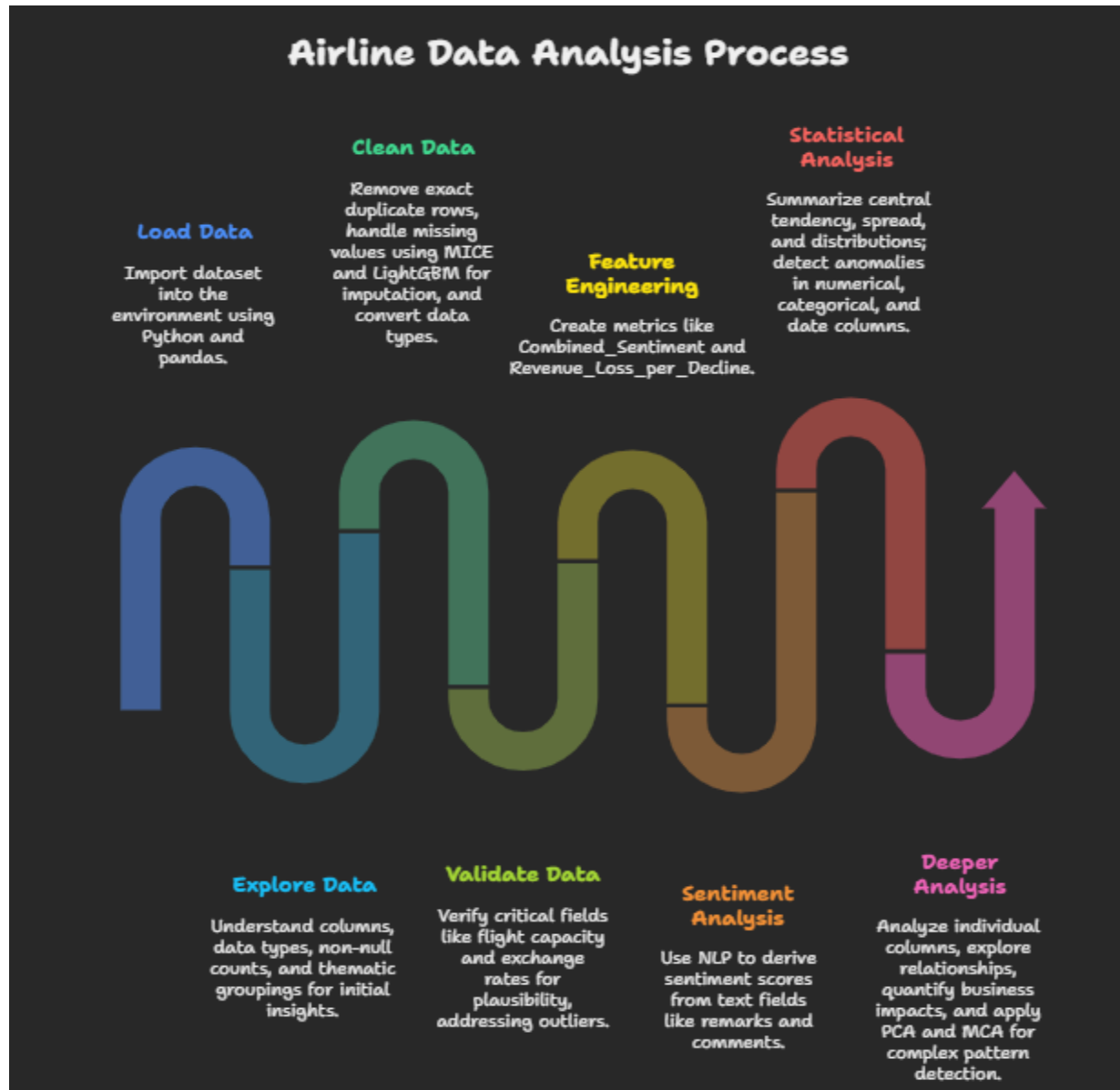
The appendix details all variables and feature-engineered metrics, providing a foundation for data-driven decisions to enhance AirwaysBharath's competitiveness and financial performance.

2 Analysis Workflow

The project follows a streamlined process to analyze airline data, ensuring robust data handling and actionable insights. It leverages advanced techniques like MICE (Multiple Imputation by Chained Equations), NLP (Natural Language Processing), LightGBM, and parallel computing to address data complexities efficiently.

Key Steps

1. **Load Data:** Import dataset into the environment using Python and pandas.
2. **Explore Data:** Understand columns, data types, non-null counts, and thematic groupings for initial insights.
3. **Clean Data:** Remove exact duplicate rows, handle missing values using MICE and LightGBM for imputation, and convert data types (e.g., dates).
4. **Validate Data:** Verify critical fields like flight capacity (load factors vs. seat counts) and exchange rates for plausibility, addressing outliers.
5. **Feature Engineering:** Create metrics like Combined_Sentiment (via NLP on remarks/comments) and Revenue_Loss_per_Decline.
6. **Sentiment Analysis:** Use NLP to derive sentiment scores from text fields like remarks and comments.
7. **Statistical Analysis:** Summarize central tendency, spread, and distributions; detect anomalies in numerical, categorical, and date columns.
8. **Deeper Analysis:**
 - **Univariate:** Analyze individual columns (numerical, categorical, dates, booleans) for patterns.
 - **Bivariate:** Explore relationships (numerical-numerical, numerical-categorical) using non-parametric tests and parallel computing.
 - **Impact:** Quantify business impacts of significant relationships using Linear regression (NUM-NUM, NUM-CAT) and Cramér's V (CAT-CAT).
 - **Multivariate:** Apply PCA (Principal Component Analysis) and MCA (Multiple Correspondence Analysis) for complex pattern detection.



Notes

- **Tools:** Python (pandas for data manipulation, scikit-learn for MICE, LightGBM for imputation, NLTK/VADER for NLP, statsmodels for regression, PCA/MCA for multivariate analysis).
- **Efficiency:** Parallel computing for bivariate and impact analyses to handle large-scale combinations.
- **AI Support:** GROK3 and Copilot for code assistance, summarization; NapkinAI for business appeal visualization.(Note: usual graphs created in seaborn are in .ipynb file)
- **Challenges:** Addressed RAM crashes by switching from Theil–Sen to Linear regression, handled data anomalies via validation and MICE/LightGBM imputation.

3 Top 5 Key Insights

Insight 1: Retention Risk Can Boost Profits — If Managed Well [NUM-NUM ANALYSIS]

- **What:** Higher customer retention risk is linked to significantly higher profits and lower costs, presenting a high-risk, high-reward opportunity in customer management.
- **Proof:** A 1% increase in retention risk correlates with ~255% higher profitability and ~223% lower costs (directional, based on OLS analysis where assumptions weren't fully met).
- **Impact (SO WHAT):** Targeting high-risk segments can dramatically boost profits and reduce costs, but unmanaged risks may lead to long-term churn and loss of market share.
- **Next (NOW WHAT):** Pilot loyalty and retention programs for high-risk, high-return customers to capture gains while mitigating churn, then scale based on ROI.

Insight 2: Profitability Rises with Customer Tier [NUM-CAT ANALYSIS]

- **What:** High-tier customers generate substantially more revenue and incur lower costs compared to Medium-tier, while Low-tier customers yield less revenue and higher costs.
- **Proof:** Relative to Medium-tier: High-tier shows ~40% higher fares, revenue, prices, and taxes, with ~12% lower costs; Low-tier has ~48% lower fares and revenue, with ~20% higher costs.
- **Impact (SO WHAT):** Prioritizing High-tier customers can increase revenue by ~40% and cut costs by ~12%, whereas Low-tier segments may erode overall margins.
- **Next (NOW WHAT):** Invest in upselling, premium services, and loyalty programs for High-tier customers; deprioritize or optimize costs for Low-tier. Test via A/B pilots to quantify impact.

Insight 3: Elevate Customer Sentiment [MULTI VARIANT ANALYSIS]

- **What:** Improving customer sentiment drives substantial uplifts in conversion rates, revenue, and overall satisfaction through direct and indirect effects.
- **Proof:** A 1% rise in sentiment and conversion leads to 39.56% higher sentiment dynamics, resulting in 19.31% indirect impact and 17.05% direct impact.
- **Impact (SO WHAT):** This can deliver up to 39.56% boosts in conversion, revenue, and satisfaction, enhancing customer loyalty and competitive edge.
- **Next (NOW WHAT):** Deploy real-time sentiment analytics with instant alerts to monitor and respond to feedback dynamically.

Insight 4: Target High-Value Routes & Merchants [MULTI VARIANT ANALYSIS]

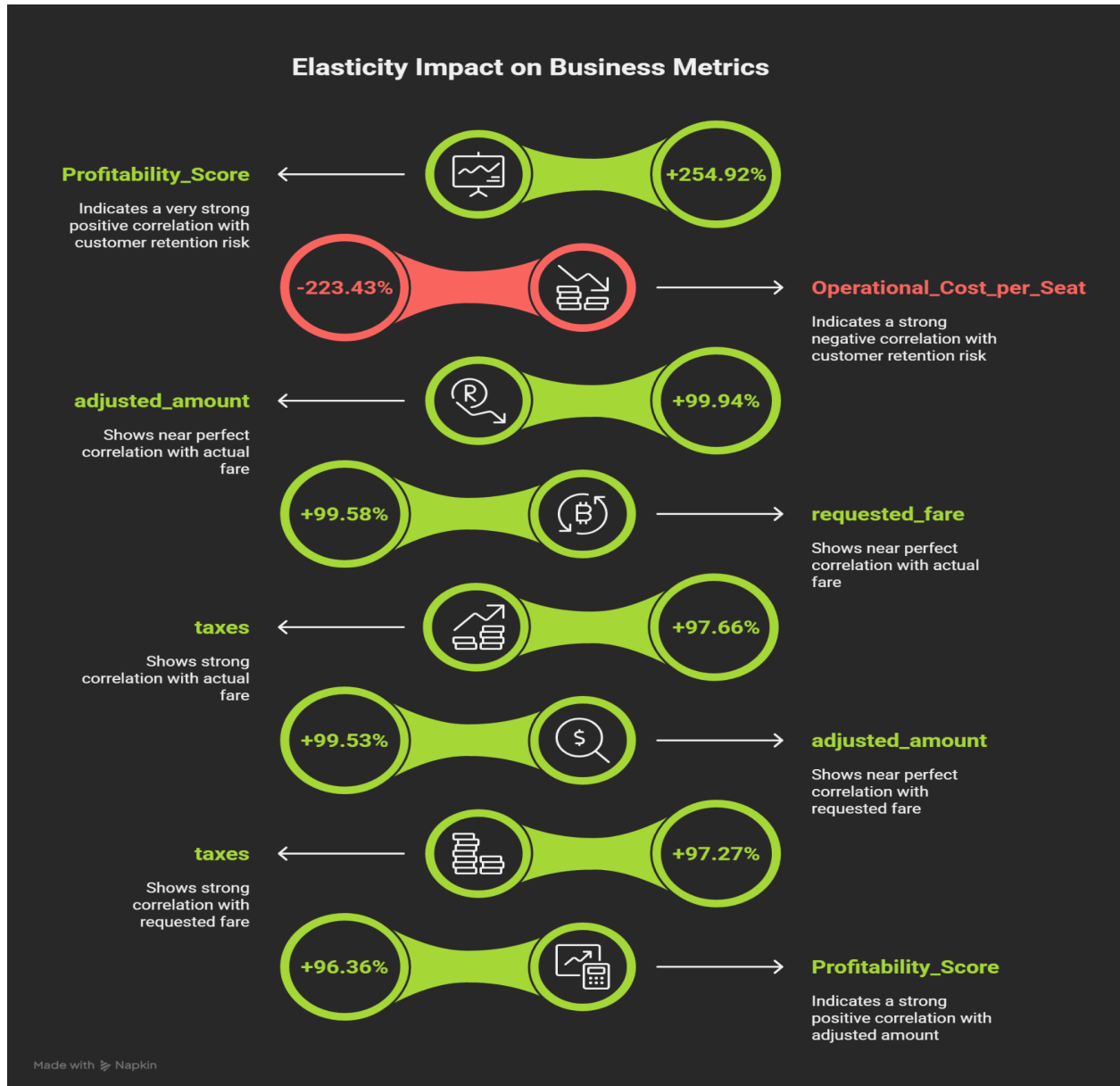
- **What:** Focusing on influential merchants and regions amplifies demand on key routes, creating cascading benefits in global route dynamics and customer preferences.
- **Proof:** A 1% increase in merchant and regional influence yields 32.82% higher global route dynamics; 1% rise in international travel patterns boosts merchant influence by 25.86%; 1% in global route dynamics increases customer location and route preferences by 21.42%.
- **Impact (SO WHAT):** Captures 32.82% more demand on flagship routes, driving significant revenue growth and market expansion.
- **Next (NOW WHAT):** Expand capacity on high-value routes (e.g., MEL–LHR, HKG–FRA) to capitalize on demand.

Insight 5: Simplify Operations & Optimize Pricing [MULTI VARIANT ANALYSIS]

- **What:** Streamlining flight operations and sentiment management reduces route complexity, while optimizing pricing and discounts enhances revenue and booking dynamics.
- **Proof:** A 1% increase in flight ops and sentiment reduces route complexity by 29.52%; 1% rise in sentiment and currency impact boosts pricing variability by 22.12%; 1% in discounts lifts revenue and booking dynamics by 12.01%.
- **Impact (SO WHAT):** Achieves 29.52% reduction in complexity, 22.12% greater pricing agility, and 12.01% uplift in bookings, improving efficiency and profitability.
- **Next (NOW WHAT):** Eliminate unnecessary layovers on high-demand routes to enhance efficiency and customer sentiment.

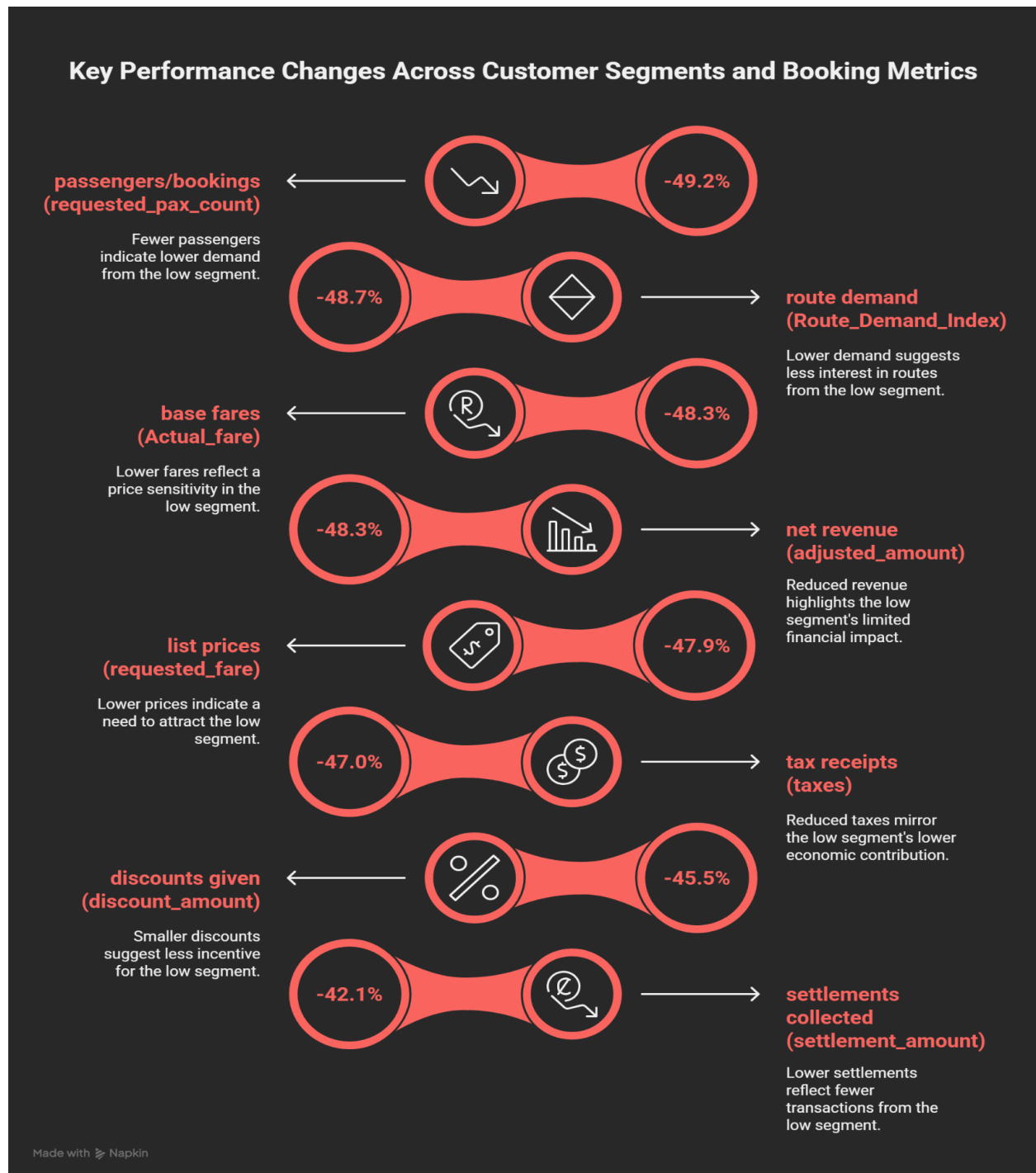
4 Additional Findings

4.1 NUM-NUM Analysis



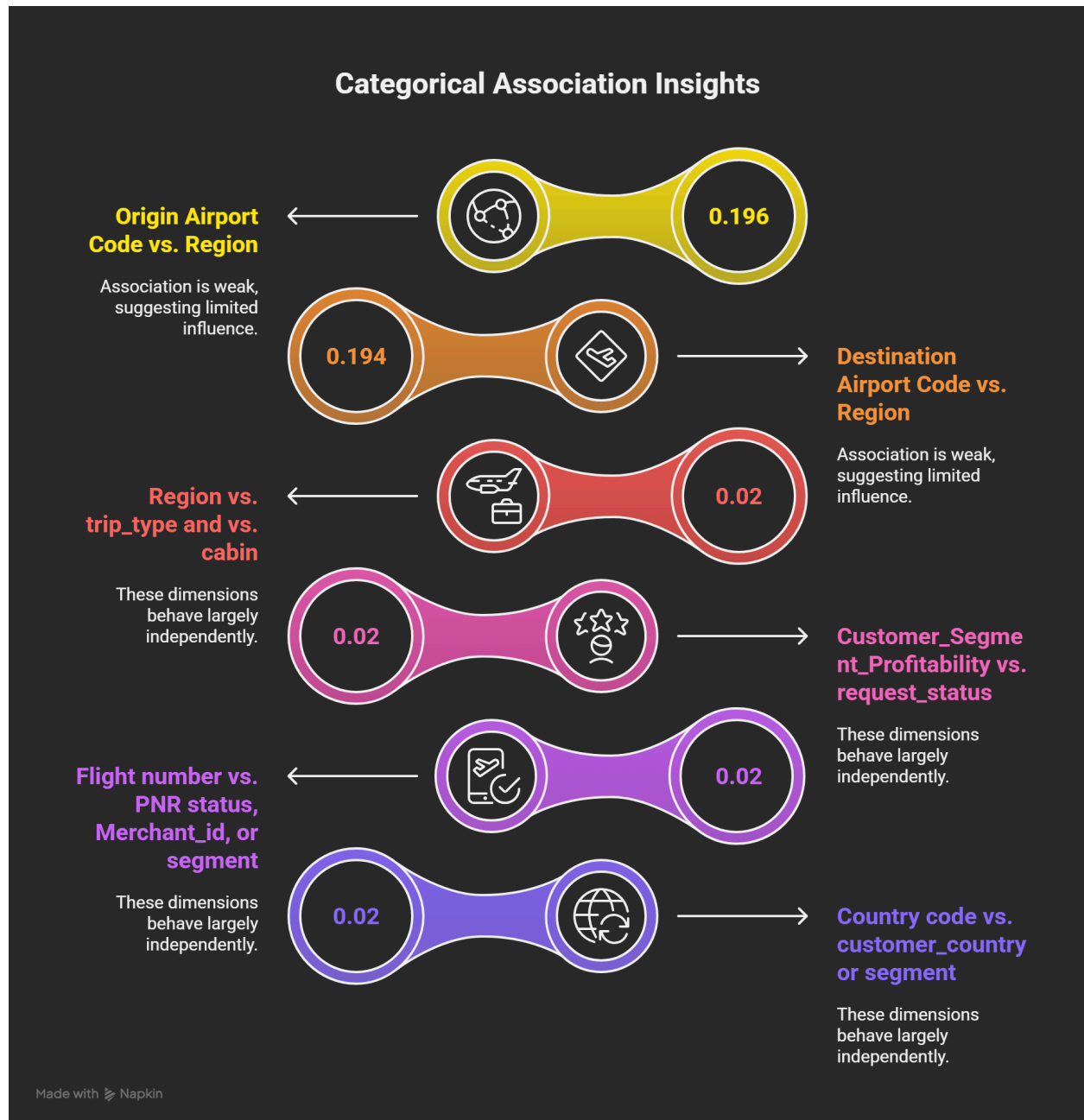
Log-log elasticities show that a 1% increase in retention risk yields +254.92% in profitability and -223.43% in costs; 1% hikes in fares or revenues generate ~97–100% lifts across related metrics, while tax and pricing shifts further boost profit and revenue losses modestly dampen conversion.

4.2 NUM-CAT Analysis



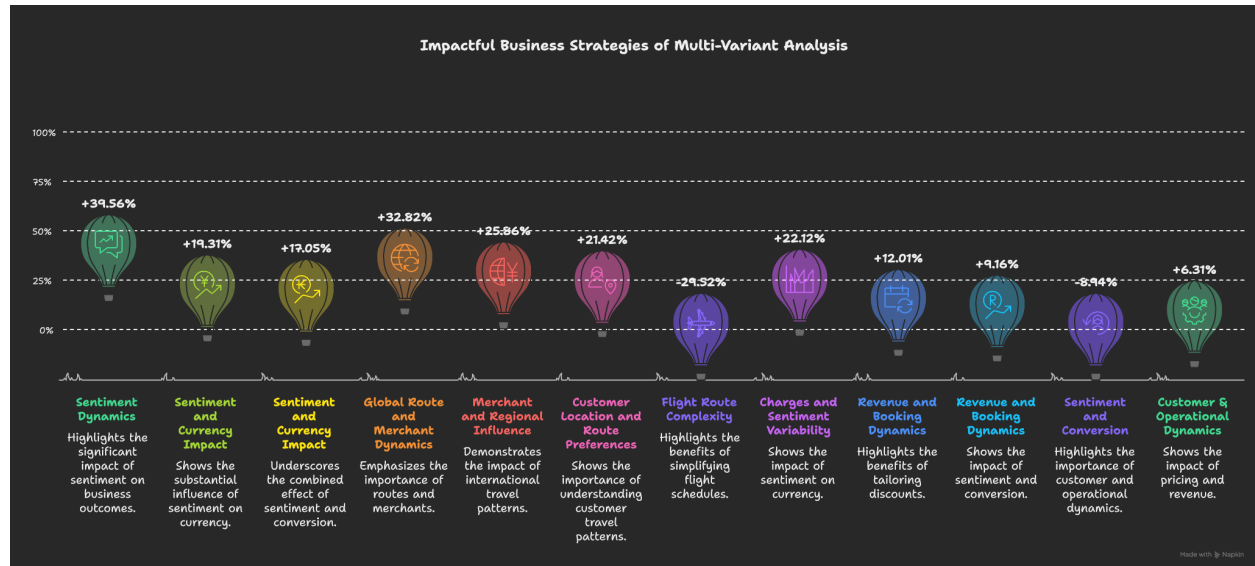
Customer segments drive $\pm 40\text{--}48\%$ swings in fares, revenue, and costs (High up $\sim 40\%$ /down $\sim 12\%$, Low down $\sim 48\%$ /up $\sim 20\%$); ticketed bookings boost conversion by 46%, missing remarks cut sentiment by $\sim 10\text{--}13\%$, and city-level factors shift passenger volumes and costs by up to $\sim 11.6\%$.

4.3 CAT-CAT Analysis



Only airport–region pairs show weak associations (Cramér's $V \approx 0.196/0.194$); all other categorical pairs are negligible ($V < 0.02$), confirming broad independence across dimensions.

4.4 MULTI-VARIANT Analysis



Multi-variant analysis reveals three high-leverage pillars—customer sentiment, route/merchant dynamics, and streamlined operations/pricing—with elasticities up to 39.56% and 32.82%, plus quick-win levers under 10% for incremental gains.

5 Recommendations & Next Steps

Strategic Initiatives

1. Pilot Loyalty & Retention for High-Risk Segments

- Maps to Insight 1
- Roll out targeted retention offers, real-time sentiment alerts, and agent empathy training to high-risk, high-value customers

2. Upsell Premium & Expand Flagship Routes

- Maps to Insights 2 & 4
- Launch personalized upsell campaigns and loyalty tiers for High-segment guests
- Increase capacity on MEL–LHR and HKG–FRA and deepen merchant partnerships.

3. Streamline Operations & Implement Dynamic Pricing

- Maps to Insights 3 & 5
- Eliminate low-utility layovers and simplify flight schedules
- Deploy AI-driven discounting and ancillary-fee optimization to boost booking dynamics

High-Level Rollout Plan

Phase	Timeline	Activities
Setup & Data Prep	Months 1–2	Integrate customer, route, sentiment, ops data; configure analytics pipelines
Pilot Initiative 1	Months 3–4	Launch retention and sentiment program; monitor feedback loops
Pilot Initiative 2	Months 5–6	Activate upsell campaigns; adjust route capacity; track load factors

Phase	Timeline	Activities
Pilot Initiative 3	Months 7–8	Optimize schedules; deploy dynamic pricing; A/B test discount thresholds
Review & Scale	Month 9	Assess pilots; refine tactics; plan enterprise rollout

Success Metrics to Monitor

- Retention rate uplift and churn reduction among high-risk segments
- Incremental profitability per segment and cost savings per seat
- Sentiment score improvements and conversion rate increases
- Load factor and net revenue gains on target routes
- Flight schedule complexity index and average layovers per itinerary
- Booking dynamics lift and pricing variability index

6 Appendix

6.1 Full Data Tables – Variables Understanding

Identifiers and Basic Trip Info

Column	Type	Non-null Count	Description
ticket_id	object	197,194	Unique ticket identifier
pnr	object	164,893	Passenger Name Record
origin_airport_code	object	197,218	Departure airport code
destination_airport_code	object	197,225	Arrival airport code
Region	object	197,193	Geographic region
trip_type	object	200,000	Trip type
stops	int64	200,000	Number of stops
cabin	object	200,000	Cabin class

Column	Type	Non-null Count	Description
requested_flight_number	object	197,184	Flight number
requested_pax_count	int64	200,000	Passenger count
group_category	object	200,000	Group booking category

Request and Merchant Details

Column	Type	Non-null Count	Description
request_type	object	200,000	Type of request
Merchant_id	object	200,000	Merchant identifier
Merchant_name	object	200,000	Merchant name
Merchant_type	object	200,000	Merchant type

Financial and Currency Info

Column	Type	Non-null Count	Description
currency	object	200,000	Transaction currency
exchange_rate	float64	170,000	Currency exchange rate
requested_fare	float64	200,000	Requested fare amount
Actual_fare	float64	200,000	Actual fare paid
adjusted_amount	float64	200,000	Adjusted total amount
discount_amount	float64	200,000	Discount applied
taxes	float64	200,000	Tax amount

Column	Type	Non-null Count	Description
other_Charges	float64	200,000	Miscellaneous charges
Special_Request_Charges	float64	200,000	Special request fees
settlement_amount	float64	200,000	Final settlement
flight_Capacity	float64	200,000	Flight capacity

Aircraft and Location Info

Column	Type	Non-null Count	Description
aircraft_type	object	197,166	Aircraft model
country_code	object	196,192	Country code
customer_country	object	200,000	Customer country
customer_city	object	197,214	Customer city

Dates and Timestamps

Column	Type	Non-null Count	Description
request_date	object	193,024	Date of request
departure_date	object	192,979	Departure date
arrival_date	object	193,054	Arrival date
offer_expiry_date	object	193,097	Offer expiry date
airline_update_date	object	193,127	Airline update date
created_at	object	197,165	Record creation time
updated_at	object	193,051	Last update time

Statutes and Remarks

Column	Type	Non-null Count	Description
request_status	object	200,000	Request status
agent_response_statuses	object	200,000	Agent response status
pnr_status	object	200,000	PNR status
payment_status	object	133,046	Payment status
reason_for_decline	object	45,169	Decline reason
user_remarks	object	59,931	User comments
airline_remarks	object	74,639	Airline comments

6.2 All Feature-Engineered Variables Understanding

Variable	Type	Description
Combined_Sentiment	float64	Weighted average of user, airline, and decline sentiment scores when remarks are present
Revenue_Loss_per_Decline	float64	Estimated revenue lost for non-honoured requests, adjusted by Combined_Sentiment
Dynamic_Pricing_Score	float64	Ratio of actual fare to mean fare (by region, cabin, stops), modulated by capacity and sentiment

Customer_Retention_Risk	float64	Risk score combining sentiment, request/PNR status, and passenger count relative to max pax
Operational_Cost_per_Seat	float64	Per-passenger cost including taxes and fees, adjusted upward by airline sentiment presence
Conversion_Flag	int64	Binary indicator (1 if ticketed and paid, 0 otherwise)
Booking_Lead_Time_Efficiency	float64	Normalized booking lead time weighted by conversion flag and user sentiment
Route_Demand_Index	float64	Demand measure based on pax count, capacity, ticket status, stops, and airline sentiment
Profitability_Score	float64	Total revenue per booking ($\text{Actual_fare} \times \text{requested_pax_count}$)
Customer_Segment_Profitability	category	Low/Medium/High segment label based on Profitability_Score percentiles and sentiment adjustments