Project Progress Report

What makes this airplane burn more fuel?

An analysis of Fuel Burn during Cruise Phase for Aircraft Tail 687

Author:

1. Duy Nguyen

GitHub Repository: https://github.com/dcnguyen060899/data_5100_final_project

Problem Statement

As we preprocessed (cleaning and creating new features) and narrowed down to 312 flights from the same 4-engine aircraft (1,882,573 rows × 32 columns, 4 Hz sampling during cruise phase), we're iteratively asking: Which cockpit controls, flight conditions, and aircraft performance metrics are most associated with instantaneous fuel burn (dependent variable – total fuel flow across all four engines), and by how much during cruise phase?

Why do we think this matters: Imagine you're a CEO at United Arlines or Boeing, fuel represents the largest operating cost for airlines and a primary source of aviation CO_2 emissions. The Boeing ecoDemonstrator program has demonstrated significant industry investment in operational efficiency, evaluating over 250 fuel-reducing technologies since 2012 with successes in flight optimization procedures that reduce fuel use, emissions, and community noise (Boeing Commercial Airplanes, 2024). We believe that with transparent, quantified relationships from flight data analysis provide a strong foundation for improving pilot technique, flight planning, and maintenance priorities, enabling cost reduction and emissions mitigation without new hardware investments.

These following questions would help us understand and guide us to break down our main research question by addressing nuances of our project statement:

1. Which engine and flight parameters have the strongest influence on fuel flow?
2. How much does fuel flow change with specific parameter adjustments?
3. Do altitude and speed effects work independently, or do they interact in complex ways?
4. How does aircraft weight modify the relationship between altitude/speed and fuel consumption?
5. Should optimal cruise strategy be static (same throughout flight) or adaptive (changing as fuel burns and aircraft weight decreases)?

Data Source

NASA DASHlink – Tail 687:

- **Original dataset:** 651 commercial flights recorded in 2012
- **Flights analyzed:** 312 flights that reached cruise altitude (above 25,000 feet)
- **Final analytical dataset:** 1,882,573 measurements representing approximately 130 hours of cruise flight time
- **Sampling rate:** 4 Hz (4 measurements per second)
- **Variables:** 32 parameters after preprocessing and feature engineering

Key Parameters:

- **Engine Performance Parameters:** Fuel Flow (FF_1, FF_2, FF_3, FF_4), Fan Speed (N1_1, N1_2, N1_3, N1_4), Core Speed (N2_1, N2_2, N2_3, N2_4), Exhaust Gas Temperature (EGT_1, EGT_2, EGT_3, EGT_4)
- **Flight Control & Envelope:** Altitude (ALT), Mach Number (MACH), True Airspeed (TAS), Angle of Attack (AOAC), Altitude Rate (ALTR), Wind Speed (WS), Wind Direction (WD), Track Angle (TRK)
- **Derived Features:** Total Fuel Flow (sum of FF_1 through FF_4), Average N1 (mean across all engines), Average N2, Average EGT, Headwind Component (calculated from wind speed, wind direction, and track angle), Cumulative Fuel Burned (proxy for aircraft weight reduction during flight)
- **Metadata:** Flight ID for tracking individual flights

Data Preprocessing Completed:
We successfully converted 312 .mat files from NASA DASHlink into a unified pandas DataFrame. Our preprocessing pipeline included:

- Filtering for cruise phase (altitude above 25,000 feet, stable altitude rate)
- Handling data quality (zero missing values confirmed in final dataset)
- Creating derived features for analysis
- Identifying and addressing multicollinearity (removed True Airspeed due to 0.935 correlation with Mach number)

We confirmed these parameters are sufficient to address our research questions because they capture the complete operational picture for fuel consumption. Flight conditions (altitude, Mach, wind) control for environmental factors, engine performance metrics (N1, N2, EGT) across all four engines indicate health and efficiency, and derived metrics (cumulative fuel burned) proxy for aircraft weight changes during flight. Together with fuel flow measurements across 312 flights, we can quantify which factors most influence fuel consumption during cruise phase.

Analytical Approach

Independent Variables (X) - Detailed Description

Based on our exploratory data analysis, we have identified five key independent variables for our regression models:

1. Altitude (ALT) - Measured in feet

- Range in dataset: 25,000 to 43,000 feet (cruise altitudes only)

- Expected Theoretical Relationship (for Comparison): Negative correlation with fuel flow (higher altitude = thinner air = more efficient)
- Operational relevance: Pilots can request altitude changes; key optimization parameter

2. Mach Number (MACH) - Dimensionless ratio of aircraft speed to speed of sound
- Range in dataset: 0.52 to 0.86
- Expected Theoretical Relationship (for Comparison): Positive correlation with fuel flow (the more thrust —> the faster —> the more fuel)
- Operational relevance: Speed directly controlled by pilots/autopilot

3. Average N1 (Engine Fan Speed) - Percentage of maximum fan speed
- Range in dataset: 85% to 99%
- Expected Theoretical Relationship (for Comparison): Strong positive correlation with fuel flow (because N1 is direct driver of thrust and fuel consumption)
- Operational relevance: Reflects power setting, controlled by throttle

4. Headwind Component - Derived from wind speed, wind direction, and aircraft track, measured in knots
- Range in dataset: -150 to +150 knots (negative = tailwind, positive = headwind)
- Expected Theoretical Relationship (for Comparison): Positive correlation with fuel flow (headwinds require more power)
- Operational relevance: Environmental factor, not directly controllable but informs routing decisions

5. Cumulative Fuel Burned - Derived metric, measured in thousands of pounds
- Range in dataset: 0 to 50+ thousand pounds burned
- Expected Theoretical Relationship (for Comparison): Negative correlation with fuel flow (lighter aircraft = more efficient)
- Operational relevance: Proxy for aircraft weight, changes throughout flight

Variable Selection Decision: We removed True Airspeed (TAS) from our model despite its 0.376 correlation with fuel flow because it exhibited severe multicollinearity with Mach number (correlation = 0.935). True airspeed and Mach number are mathematically related (TAS = Mach × speed of sound), so including both would inflate standard errors and make coefficient interpretation unreliable. We retained Mach as it is the standard operational metric pilots use.

Dependent Variable (y) - Detailed Description

Total Fuel Flow - Sum of fuel flow across all four engines (FF_1 + FF_2 + FF_3 + FF_4), measured in pounds per hour (lbs/hr)
- Range in dataset: 4,200 to 7,500 lbs/hr during cruise
- This is our target variable for optimization
- Represents instantaneous fuel consumption rate
- Lower values indicate better fuel efficiency

Exploratory Data Analysis

Stage 1: Univariate Analysis

We analyzed the distribution of all 32 variables in our dataset. Key findings:

- Zero missing values confirmed across 1.88 million measurements
- All variables fall within expected operational ranges based on aeronautical engineering standards
- We discovered "outliers" in speed (5.4% of data), which were thereby validated as legitimate slow-cruise fuel-saving operations, not data errors
- Engine performance parameters show consistent patterns across all four engines

Thus far, in the notebook, we conducted:

- **EDA** examined 32 variables
- **Bivariate** analysis identified 5 key predictors
- **Exploratory modeling** (supporting data analysis) uses 3 primary variables (altitude, mach, weight_category) because weight interaction proved most important

We will include **N1** and **headwind** to explore in the final interaction models.

Stage 2: Bivariate Analysis

We calculated correlations between all potential independent variables and our dependent variable (total fuel flow). The results were ranked by correlation strength and direction:

Variable	Correlation with Fuel Flow	Interpretation
Average N1 (Engine Power)	+0.693	Strongest predictor - direct driver of fuel consumption
Altitude	-0.412	Strong negative relationship - higher altitude reduces fuel burn
Mach Number	+0.376	Moderate positive - faster speeds increase fuel consumption
Cumulative Fuel Burned	-0.269	Negative relationship - lighter aircraft more efficient
Headwind Component	-0.168	Weak negative - counterintuitively, dataset shows modest headwind
True Airspeed	+0.376	Not included in final model due to multicollinearity with Mach (r=0.935)

Preliminary Insights from Bivariate Analysis

Key Finding 1: Engine Power Dominates

- Average N1 (engine fan speed) shows the strongest correlation (0.693) with fuel flow. This makes physical sense because engine power is the direct cause of fuel consumption. Thereby, our scatter plots reveal an approximately linear relationship with slope around 112 lbs/hr per 1% increase in N1.

Key Finding 2: Altitude Efficiency

- With higher cruise altitudes, we observed an association with significantly lower fuel consumption (correlation -0.412). The relationship appears non-linear, with diminishing returns above 38,000 feet. This aligns with aeronautical theory: thinner air at high altitude reduces drag.

Key Finding 3: Speed-Fuel Tradeoff

- Mach number shows us a moderate positive correlation (0.376) with fuel flow. That means, the faster cruise speeds, the more thrust required and thereby burn more fuel. However, to our surprise, the relationship is weaker than expected, suggesting that at optimal cruise altitudes, the aerodynamic efficiency partially compensates for increased speed.

Key Finding 4: Weight Reduction Effect

- The cumulative fuel burned (proxy for aircraft getting lighter as flight goes on) shows a negative correlation (-0.269) with fuel flow. As the aircraft burns fuel toward later cruise phase and becomes lighter throughout the flight, it requires less fuel to maintain altitude and speed. This effect is measurable and consistent across flights.

Key Finding 5: Wind Effects are Complex

- Headwind component shows weak correlation (-0.168). Initially, this counterintuitive finding (we expected positive correlation) can potentially warrants us a deeper investigation in modeling section. A possible explanations include confounding between wind and operational decisions:
    - Pilots adjust power settings to compensate for headwinds (confounds wind with fuel flow)
    - Pilots request altitude changes specifically to find favorable winds (confounds altitude with wind conditions)

Plan for Completion

Where We Are Now:

We have successfully completed data preprocessing, exploratory data analysis and some initial modeling. Our dataset of 312 flights (1.88 million cruise measurements) is cleaned with zero missing values. We have identified five key independent variables through correlation analysis and addressed multicollinearity. However, we have not yet run multivariate regression full and reduced models along with ANOVA tests.

Remaining Steps for Next 3 Weeks:

1. **Multivariate Regression Modeling (15-18 person-hours, ~1-2 Weeks)**
    - Test interaction models with stratified term: Headwind vs. Tailwind condition (4-5 hours)

- Test interaction models with stratified term: Low, Medium, and High altitudes (4-5 hours)
- Run diagnostics: residuals, ANOVA, assumption checks (4-5 hours)
- Interpret coefficients in operational terms with domain validation (3-4 hours)

2. **Final Presentation and Documentation (12-14 person-hours, Week 3)**
   - Synthesize findings into executive summary and recommendations (4 hours)
   - Create 15-20 slide presentation with narrative arc (4 hours)
   - Practice presentation and refine (3 hours)
   - Polish GitHub repository and export final notebook (2 hours)

Uncertainties and Impediments:

Technical: Interaction effects may be complex to interpret. Mitigation: Start with visualizations before statistical tests; consult instructor's per request if needed.

Timing: In the next two weeks, our modeling work will overlap with our DATA 5300 midterms. Mitigation: plan to push ourselves to finish multivariate regression modeling section within one weeks instead of two.

Interpretation: Regression coefficients must translate to actionable operations guidance. Mitigation: Focus on interpretable linear models; frame findings as associations rather than causal claims; domain expert reviews all recommendations. Our plan is to build understanding incrementally through iterative modeling, starting with simple relationships and progressively adding complexity to reveal deeper insights.

Overall confidence: We are highly confident with where we going and where we currently at because we have a strong data foundation with a clear methodology, effective team collaboration, and realistic timeline.

Citations

- Boeing Commercial Airplanes. (2024). The Boeing ecoDemonstrator Program [Backgrounder]. Boeing. https://www.boeing.com/content/dam/boeing/boeingdotcom/principles/environment/pdf/BKG-ecoDemonstrator_2024_Oct.pdf
- Boeing. (2024). ecoDemonstrator Program. Boeing Sustainability. https://www.boeing.com/sustainability/ecodemonstrator
- Matthews, B. (2012). Flight Data for Tail 687. NASA DASHlink (C3). https://c3.ndc.nasa.gov/dashlink/resources/664/

Updated Task:

- Interaction analysis and regression modeling, contributes to all stages of the data science project cycle, and responsible for technical documentation by maintaining GitHub repository workflow.
- Baseline regression models and visualizations, and contributes to all stages of the data science cycles.
- Stratified analysis, provides aeronautical domain expertise validation, designs final presentation, and contributes to all stages of the data science project cycle.