Part 1.1 — Feature Engineering & Data Preparation

Question

“Explain the data processing steps to select the best features and prepare them for the learning phase.”

What the data looks like

Pickled Pandas DataFrames stored in `question_1/`:

- `signals_fuel_flow.pkl` (lb/s)
- `signals_altitude.pkl` (ft)
- `signals_wind.pkl` (knots)
- `signals_vitesse.pkl` (km/h)

Each file has:

- **Index:** time vector (can differ by flight)
- **Columns:** one column per flight (e.g., column `44` is flight #44)

Approach (what the script does)

1. **Load signals** for all flights and align by each flight’s own time index.
2. **Per-flight feature extraction** from each time series (robust summary statistics + simple dynamics).
3. **Quality checks**
    - Ensure the same set of flights exists across all four signals.
    - Verify no missing values after extraction.
4. **Normalization (optional output)**
    - Save both the **raw feature matrix** and a **normalized** version (e.g., standard scaling) for modeling.
5. **Persist artifacts** to disk for downstream steps.

Engineered features (21)

Computed per flight:

- **Fuel flow (ff_*)**
  - `ff_mean`, `ff_std`, `ff_max`, `ff_min`
  - `ff_rate_change` — mean absolute Δ fuel flow / Δ time
- **Altitude (alt_*)**
  - `alt_mean`, `alt_std`, `alt_max`, `alt_min`, `alt_range` (=max–min)
  - `alt_climb_rate` — median climb/descent rate
- **Wind (wind_*)**
  - `wind_mean`, `wind_std`, `wind_max`, `wind_min`
- **Speed (speed_*)**
  - `speed_mean`, `speed_std`, `speed_max`, `speed_min`, `speed_change` (mean |Δspeed|/Δt)
- **Flight duration**
  - `duration` — span of the time index for that flight

These capture level (mean), variability (std/range), extremes (min/max), and simple dynamics (rates/changes) across fuel, altitude, wind, and speed.

Key checks printed by the script

- Number of flights loaded
- Progress logs (every ~20 flights)
- Number of features extracted and their names
- Missing-value check
- Feature summary statistics (sanity scan)

Output:

```
(/Users/Ayush/Downloads/SFA_Intern_AI_technical_test/Solution/venv) Ayush@MacBook-Pro-38 Answer_1 % python3 step_1_feature_engineering.py
Loading flight signals...
Loaded 100 flights

Extracting features from time series...
  Processing flight 0/100...
  Processing flight 20/100...
  Processing flight 40/100...
  Processing flight 60/100...
  Processing flight 80/100...

Extracted 21 features for 100 flights

Feature columns:
['ff_mean', 'ff_std', 'ff_max', 'ff_min', 'ff_rate_change', 'alt_mean', 'alt_std', 'alt_max', 'alt_min', 'alt_range', 'alt_climb_rate', 'wind
_mean', 'wind_std', 'wind_max', 'wind_min', 'speed_mean', 'speed_std', 'speed_max', 'speed_min', 'speed_change', 'duration']

No missing values - good to go!

Feature statistics:
           ff_mean       ff_std       ff_max       ff_min  ...     speed_max     speed_min  speed_change    duration
count   100.000000   100.000000   100.000000   100.000000  ...  1.000000e+02    100.000000    100.000000  100.000000
mean     19.238197    16.693448    80.503693    12.150055  ...  1.587060e+04     32.590000      1.787876   21.070000
std     168.264141   113.391990   548.500394   129.827493  ...  1.499121e+05   1610.573159     17.486782    2.944795
min      -3.477813     0.000047     0.000168   -80.000000  ...  5.700000e+02 -15852.000000    -50.666667   15.000000
25%       0.000331     0.000357     0.001061     0.000021  ...  8.550000e+02     95.000000    -10.138889   19.000000
50%       0.000717     0.000961     0.002391     0.000043  ...  8.550000e+02    190.000000      0.000000   21.000000
75%       0.001468     0.002538     0.006627     0.000065  ...  9.500000e+02    285.000000     13.234428   23.000000
max    1670.681818  1042.572030  5000.000000  1295.000000  ...  1.500000e+06    895.000000     42.222222   29.000000

[8 rows x 21 columns]

Saved features to ../../question_1/processed_features.pkl
Saved normalized features to ../../question_1/processed_features_normalized.pkl
```

- `../../question_1/processed_features.pkl` — raw per-flight feature table (shape: `n_flights × 21`)
- `../../question_1/processed_features_normalized.pkl` — normalized version of the same table

How to run:

```
# From the project root (adjust path if needed)
  - python3 step_1_feature_engineering.py
```

Expected runtime: a few seconds on 100 flights.

Part 1.2 — Fuel Flow Model @ 8000 ft (Why not wind?)

Question

“Build a fuel flow model as a function of the effective speed range of data, at a constant altitude of 8000 ft. Explain why *not* to include wind speed.”

Approach

1. **Load & align signals** (fuel flow, altitude, wind, speed) for all flights; keep only timestamps where all signals exist.
2. **Sanitize**
   - Drop non-physical values (≤ 0).
   - Constrain to a plausible speed envelope **[200, 950] km/h** to remove outliers.
3. **Fix altitude**
   - Filter to **8000 ft ± 500 ft** to approximate constant-altitude conditions.
4. **Define effective speed range**
   - Use **P10–P90** within the 8 kft band to avoid tails (here it equals the full in-band range).
5. **Reduce autocorrelation**
   - Aggregate per flight (**median** of fuel, speed, wind) to get one point per flight.
6. **Modeling**
   - Fit and compare:
     - Linear: `fuel ~ speed`
     - Linear + wind: `fuel ~ speed + wind`
     - Robust (Huber): `fuel ~ speed`
     - Quadratic: `fuel ~ speed + speed²`
   - Evaluate with in-sample R^2/RMSE and plot fit + residuals.
7. **“Why not wind?” test**
   - Inspect correlations and ΔR^2 after adding wind.

Data & filters (from run)

- Raw points combined: **2105** across **100** flights
- After cleaning: **1686** points
- At **8000 ft ± 500**: **145** points, **36** flights
- Effective speed range (P10–P90 at 8 kft): **285–950 km/h** (same as in-band range)

Findings

- **Correlation (flight-level medians)**
    - Speed vs fuel: **+1.000**
    - Wind vs fuel: **+0.072**
- **Models**
    - Linear (speed): **R^2 = 0.9992**, RMSE ≈ 0
    - Linear (speed + wind): **ΔR^2 = +0.0000** → **negligible**
    - Robust (Huber): mirrors linear
    - Quadratic: **no gain** ($\Delta R^2 = 0$)
- **Visualization**: saved as `../../question_1/fuel_flow_analysis.png`

Why not wind?

- **Physics:** Wind changes **ground speed**, not **indicated airspeed** or **engine thrust**, so it should not drive fuel burn at fixed altitude/speed.
- **Data:** Very weak correlation with fuel, and **no improvement** in model fit when added.

> **Conclusion:** At ~8 kft, **fuel ≈ f(speed)** is well-captured by a **simple linear model**; wind adds no predictive value.

Outputs

```
(/Users/Ayush/Downloads/SFA_Intern_AI_technical_test/Solution/venv) Ayush@MacBook-Pro-38 Answer_1 % python3 step_2_fuel_flow_model.py
Combined 2105 data points from 100 flights

STEP 1: Sanitizing data
============================================================
Removed 3 non-physical values (<=0)
Clipped to plausible speed range [200-950 km/h]: removed 416 outliers
Clean dataset: 1686 points remaining

STEP 2: Filtering for 8000 ft altitude band
============================================================
Points at 8000 ± 500 ft: 145
  Flights represented: 36

STEP 3: Defining effective speed range
============================================================
Full speed range at 8k ft: 285 - 950 km/h
Effective range (P10-P90): 285 - 950 km/h
Points in effective range: 145

STEP 4: Aggregating by flight
============================================================
Aggregated to 36 flight-level observations
This reduces autocorrelation from time-series data

STEP 5: Correlation Analysis
============================================================
Correlations with fuel_flow:
  Speed:    +1.000
  Wind:     +0.072
  Altitude: +nan

STEP 6: Building Fuel Flow Models
============================================================

Model 1: Linear Regression (Speed only)
  fuel_flow = -0.0000 + 0.000000 * speed
  R² = 0.9992
  RMSE = 0.0000 lb/s

Model 2: Linear Regression (Speed + Wind)
  fuel_flow = 0.0000 + 0.000000 * speed + -0.000000 * wind
  R² = 0.9992
  ΔR² = +0.0000  ✗ Negligible improvement

Model 3: Robust Regression (Huber, Speed only)
  fuel_flow = -0.0000 + 0.000000 * speed
  R² = 0.9992

Model 4: Quadratic Regression (Speed + Speed²)
  fuel_flow = -0.0000 + 0.000000 * speed + -0.00000000 * speed²
  R² = 0.9992
  ΔR² = +0.0000  ✗ Linear is sufficient

STEP 7: Creating visualization...
Saved plot to ../../question_1/fuel_flow_analysis.png

============================================================
SUMMARY: Why NOT wind speed?
============================================================
1. Correlation: Speed=+1.000 vs Wind=+0.072
2. Model improvement with wind: ΔR² = +0.0000 (negligible)
3. Physics: Fuel burn depends on engine thrust (airspeed),
   not wind, which only affects ground speed

Best model: Linear with R² = 0.9992
============================================================
(/Users/Ayush/Downloads/SFA_Intern_AI_technical_test/Solution/venv) Ayush@MacBook-Pro-38 Answer_1 %
```

- **Plot:** `../../question_1/fuel_flow_analysis.png` (scatter + linear/quadratic fits and residuals)

How to run

```
# From your project root (adjust path if needed)
  - python3 step_2_fuel_flow_model.py
```

Part 1.3 — Fuel Flow vs Altitude @ Constant Speed (Why not wind?)

Question

“Build a fuel flow model as a function of altitude (0–15,000 ft) at a constant speed of 665 km/h. Explain why *not* to include wind speed.”

Approach

1. **Load & clean** all four signals (fuel, altitude, wind, speed) and keep only timestamps where every signal is present.
2. **Speed banding** to approximate constant speed: **665 ± 25 km/h**.
3. **Coverage check**: confirm available altitude span within the band (data provides **1000–10,000 ft**, not the full 0–15,000 ft).
4. **Variance decomposition** (between vs within flights) on fuel flow to understand where the signal comes from.
5. **Two modeling views**
   - **Raw (between-flight effects)**: `fuel ~ altitude` using flight-level data.
   - **Within-flight (demeaned)**: remove flight baselines, then test `Δfuel ~ Δaltitude`.
6. **Wind check**: add wind to the raw model and measure **ΔR^2**.
7. **Diagnostics & visualization**: produce a single figure summarizing coverage, variance, and model behavior.

Data & coverage (from run)

- Clean dataset: **1691** points from **99** flights
- At **665 ± 25 km/h**: **184** points, **44** flights
- Altitude coverage in this band: **1000–10,000 ft** (not 0–15k)
- Points per flight: mean **4.2**, median **4** (limited within-flight variation)

Findings (as printed)

- **Fuel units** are small (0.0001–0.01 lb/s) → dataset is simulated/scaled for the exercise.
- **Variance decomposition**
  - Between-flight variance: **~97.8%**
  - Within-flight variance: **~2.2%**
  - Interpretation: different aircraft/config baselines dominate the variation.

- **Signal tests**
  - **Model A (raw)** `fuel ~ altitude`: **CV $R^2 \approx -0.618$** → not predictive.
  - **Model B (demeaned)** `Δfuel ~ Δaltitude`: **CV $R^2 \approx -28.227$** → no within-flight altitude effect.
  - **Random (permuted) baseline**: CV R^2 around **−0.237**, similar to A, confirming lack of learnable signal.
- **Wind**
  - Correlation with fuel: **+0.175** (weak).
  - Adding wind to raw model: **CV $R^2 = -0.811$**; **$\Delta R^2 = -0.193$** hurts, not helps.
- **Physics rationale**
  - At constant **airspeed**, thrust requirement (and hence fuel flow) is largely unchanged by **wind**, which affects **ground speed**, not **engine power**.
  - With speed held constant and cruise-heavy segments, altitude changes are minimal and do not drive a reliable change in fuel flow.

Conclusion:
At constant speed (665 km/h), there is **no meaningful within-flight relationship** between altitude and fuel flow in this dataset; observed "effects" in raw models come from **between-flight baselines**. **Wind** does not improve prediction and should be excluded.

OUTPUT

```
(/Users/Ayush/Downloads/SFA_Intern_AI_technical_test/Solution/venv) Ayush@MacBook-Pro-38 Answer_1 % python3 step_3_diagnostic_analysis.py
======================================================================
DIAGNOSTIC ANALYSIS: Fuel Flow vs Altitude at Constant Speed
======================================================================

Clean dataset: 1691 points from 99 flights

======================================================================
DATA AT 665 ± 25 km/h
======================================================================
Total points: 184
Flights: 44
Altitude range: 1000 - 10000 ft
Note: Data only covers 1000-10000 ft, NOT 0-15k ft

Points per flight: mean=4.2, median=4
Flights with <3 points: 12 of 44

======================================================================
DIAGNOSTIC 1: Fuel Flow Scale Analysis
======================================================================

Fuel flow statistics:
count    184.000000
mean       0.001035
std        0.002061
min        0.000072
25%        0.000114
50%        0.000290
75%        0.000814
max        0.010628
Name: fuel_flow, dtype: float64

Observation: Values are in the 0.0001-0.01 lb/s range
Expected cruise fuel flow: 2-10 lb/s for commercial jets
Conclusion: This is simulated/scaled data, NOT realistic units

======================================================================
DIAGNOSTIC 2: Variance Decomposition
======================================================================

Total variance:          0.00000425
Between-flight variance:  0.00000415 (97.8%)
Within-flight variance:   0.00000009 (2.2%)

Conclusion: 98% of variance is BETWEEN flights
            Only 2% varies WITHIN flights

This means each flight has a different baseline fuel flow (aircraft/config)
but altitude has minimal effect within each flight at constant speed.

======================================================================
DIAGNOSTIC 3: Within-Flight Altitude Effect
======================================================================

Flights with ≥3 points: 8
Altitude range within flights: mean=3750 ft, median=3000 ft

Within-flight correlations (altitude vs fuel_flow):
  Mean: -1.000
  Median: -1.000
  Std: 0.000
  Range: [-1.000, -1.000]

WARNING: 8/8 flights have near-perfect correlations
  This suggests flights have very few altitude points (likely cruising)
  Correlations from 2-3 points are unreliable
```

```
(/Users/Ayush/Downloads/SFA_Intern_AI_technical_test/Solution/venv) Ayush@MacBook-Pro-38 Answer_1 % python3 step_3_diagnostic_analysis.py
Conclusion: At constant speed, flights maintain relatively constant altitude
            (cruise phase), providing insufficient altitude variation for modeling.

======================================================================
DIAGNOSTIC 4: Signal Detection Test
======================================================================

Effective range: 2000 - 10000 ft
Points: 172 from 41 flights

Model A: RAW fuel flow ~ altitude
  Grouped CV R² (5-fold): -0.6182 ± 1.4031
  Interpretation: No predictive power

Model B: DEMEANED fuel flow ~ altitude (within-flight effect)
  Grouped CV R² (5-fold): -28.2268 ± 35.8585
  Interpretation: After removing flight offsets, no altitude effect remains

Model C: RANDOM (permuted) baseline
  Grouped CV R² (5-fold): -0.2367 ± 0.3682

Key Insight:
  Raw model (A) captures between-flight differences (R²=-0.618)
  Demeaned model (B) shows no within-flight altitude effect (R²=-28.227)
  Conclusion: Altitude 'effect' is actually just different aircraft/configs

======================================================================
DIAGNOSTIC 5: Why NOT Wind Speed?
======================================================================

Correlations with fuel_flow:
  Altitude: -0.369
  Wind:     +0.175

Adding wind to raw model:
  CV R² with wind: -0.8111
  ΔR² = -0.1929

Reasons to exclude wind:
  1. Raw wind magnitude ≠ headwind component (need track angle)
  2. ΔR² = -0.1929 (slight change)
  3. Physics: At constant airspeed, wind doesn't change thrust requirement
  4. Wind affects ground speed, not indicated airspeed or engine power

======================================================================
Creating visualizations...
======================================================================
Saved: ../../question_1/diagnostic_report.png

======================================================================
FINAL CONCLUSION
======================================================================

Question: Model fuel flow as function of altitude (0-15k ft) at 665 km/h

Findings:
  1. Data coverage: Only 2000-10000 ft available (not 0-15k)
  2. 98% of variance is between flights (different aircraft/configs)
  3. Within flights: only 2% variance with altitude
  4. Raw model CV R²: -0.618 (captures between-flight differences)
  5. Demeaned model CV R²: -28.227 (no within-flight effect)
  6. Wind adds no value (ΔR² = -0.1929)

Conclusion:
  At constant airspeed (665 km/h), altitude has NO meaningful
  within-flight effect on fuel flow. The data represents cruise conditions
  where altitude varies minimally. The weak correlation in raw data reflects
```

Artifacts

- **Figure:** `../../question_1/diagnostic_report.png`
- **Note:** All visualizations for this part (and earlier parts) are saved in the project's **`question_1`** folder.

How to run

```
# From the project root (adjust path if needed)
  -  python3 step_3_diagnostic_analysis.py
```

Part 1.4 — Bonus: 2D Fuel Flow Model (Speed + Altitude)

Question

“Build a fuel flow model as a function of speed and altitude on the effective range of data.”

Approach

1. **Load & clean** all four signals (fuel, altitude, wind, speed); keep rows where all are present.
2. **Determine effective range (P5–P95)** to avoid tail artifacts:
   - Altitude: 1000–10,000 ft
   - Speed: 190–950 km/h
   - Fuel flow: 0.04–7.46 lb/s (after scale factor in script)
3. **Diagnose structure before modeling**
   - Altitude–speed correlation (moderate: +0.061) → some independent variation.
   - Within-flight coefficient of variation (alt ~0.49, speed ~0.40) → enough dynamics for within-flight modeling.
   - **Variance decomposition (log1p fuel)** to separate between-flight vs within-flight signal (ICC = 0.692).
4. **Model families (GroupKFold by flight)**
   - **Between-flight** models on flight-level medians:
     - Linear [alt, speed]
     - Linear + interactions [alt, speed, alt×speed]
     - Physics features: dynamic pressure `q = 0.5·ρ·V²` and `1/q`
   - **Within-flight** models on demeaned data:
     - Linear [Δalt, Δspeed]
     - Physics [Δq, Δ(1/q)]
5. **Evaluate** with R^2, RMSE, MAE (cross-validated, grouped by flight) and produce a consolidated diagnostic figure.

Data & coverage (from run)

- After cleaning: **1,878** observations from **100** flights
- Effective ranges: **Altitude 1k–10k ft**, **Speed 190–950 km/h**
- Altitude–Speed correlation: **+0.061** (moderate)

Findings

- **Variance split:** SS_between = 69.2%, SS_within = 30.8% → **ICC = 0.692** (between-flight offsets dominate).
- **Between-flight models** (flight-level medians)
    - Linear [alt, speed]: **R^2 = −1.133 ± 0.957**
    - With interactions: **R^2 = −0.689 ± 0.412**
    - Physics [q, 1/q]: **R^2 = −0.666 ± 0.416**
      → Cross-aircraft predictability is poor; offsets overwhelm.
- **Within-flight models** (demeaned)
    - Linear [Δalt, Δspeed]: **R^2 = +0.493 ± 0.151**
    - Physics [Δq, Δ(1/q)]: **R^2 = +0.048 ± 0.033**
      → A simple linear model captures within-flight effects best on this dataset.

Interpretation

- Between flights: No reliable cross-aircraft prediction (aircraft/config/weight differences dominate).
- Within flights: Altitude/speed effects are **detectable** and moderately predictive with a linear model.

Artifacts

- **Figure:** `../../question_1/bonus_2d_comprehensive_diagnostics.png`
- **Note:** All visualizations for this part (and others) are saved under the project's **`question_1`** folder.

Output:

```
● (/Users/Ayush/Downloads/SFA_Intern_AI_technical_test/Solution/venv) Ayush@MacBook-Pro-38 Answer_1 % python3 step_4_bonus_2d_model.py
================================================================================
BONUS: 2D Fuel Flow Model (Speed + Altitude)
================================================================================

[1/6] Loading data...
  Flights with complete signals: 100
  After cleaning: 1,878 observations from 100 flights

[2/6] Determining effective range and 2D coverage…
  Effective range (P5–P95):
    Altitude:     1000 –    10000 ft
    Speed:         190 –       950 km/h
    Fuel flow:      0.04 –        7.46 lb/s

  Altitude–Speed correlation: +0.061 (moderate)
  Within-flight median CV – altitude: 0.4919, speed: 0.4027

[3/6] Variance decomposition (log1p fuel)…
  SS total:   2279.8132
  SS between: 1577.2474 (69.2% of total)
  SS within:  702.5659 (30.8% of total)
  ICC (between / total): 0.692

[4/6] Building models…
  Cross-validation folds – between: 5
  Between Linear [alt, speed]       | R²= -1.133±0.957 | RMSE=0.8517 | MAE=0.3857
  Between Interactions [alt, speed, alt*speed] | R²= -0.689±0.412 | RMSE=0.8183 | MAE=0.3769
  Between Physics [q, 1/q]          | R²= -0.666±0.416 | RMSE=0.8156 | MAE=0.3788
  Cross-validation folds – within:  5
  Within Linear [Δalt, Δspeed]      | R²= +0.493±0.151 | RMSE=0.4332 | MAE=0.2526
  Within Physics [Δq, Δ(1/q)]       | R²= +0.048±0.033 | RMSE=0.5865 | MAE=0.3623

[5/6] Model performance summary…
  Best between-flight R²: -0.666
  Best within-flight  R²: +0.493
  Between-flight physics model largest absolute error (log1p units): 12.79 on flight_id=12.0

[6/6] Creating diagnostic figure…
/Users/Ayush/Downloads/SFA_Intern_AI_technical_test/Solution/Answer_1/step_4_bonus_2d_model.py:424: UserWarning: This figure includes Axes th
at are not compatible with tight_layout, so results might be incorrect.
  plt.tight_layout()

  Saved diagnostic figure: ../../question_1/bonus_2d_comprehensive_diagnostics.png

================================================================================
FINAL SUMMARY
================================================================================
Data overview:
  Observations: 1,878
  Flights:      100
  Effective altitude range: 1000–10000 ft
  Effective speed range:    190–950 km/h

Key findings:
  ICC (between/total SS): 0.692
  Between-flight best R²: -0.666
  Within-flight  best R²: +0.493

Interpretation:
  Between flights: No reliable cross-aircraft predictability; offsets dominate.
  Within flights: Altitude/speed effects are detectable.

Caveats:
  Results assume FUEL_SCALE=1000.0 is appropriate.
  Speed may be ground speed if not explicitly TAS/Mach.
  ISA density is assumed; real atmosphere may differ.
  Models are valid within the effective range only.
```

zsh Ans…
Python flas…
zsh Answer_1

How to run

```
# From the project root (adjust path if needed)
  - python3 step_4_bonus_2d_model.py
```