EV Predictive Maintenance

Phase 3: Exploratory Data Analysis

Statistical Insights & Degradation Patterns

Student: Jai Kumar Gupta

Instructor: Vandana Jain

Institution: DIYGuru

November 10, 2025

Contents

1 Executive Summary

Phase 3 transforms clean battery datasets into actionable intelligence through comprehensive exploratory data analysis. This phase reveals critical degradation patterns, quantifies relationships between operational parameters and battery health, and validates physics-based hypotheses using statistical evidence.

Phase 3 Objectives

Primary Goal: Discover hidden patterns, correlations, and degradation signatures in battery data to inform feature engineering and model selection for accurate SoH/RUL prediction.

1.1 Key Achievements

- **Degradation Visualization:** Plotted capacity fade over 600+ cycles, confirming non-linear aging patterns
- **Voltage Profile Analysis:** Compared discharge curves across battery lifecycle (new vs. aged vs. end-of-life)
- **Thermal Signatures:** Quantified temperature rise as indicator of internal resistance growth
- **Correlation Studies:** Identified strong predictors (discharge time, voltage drop) and weak ones (instantaneous measurements)
- **Real-World Data Exploration:** Analyzed 3.4 million rows from Chengdu EV fleet data
- **Statistical Validation:** Computed Pearson and Spearman correlations to distinguish linear vs. monotonic relationships

EDA Results Summary

Strongest Correlations with Capacity:
- Discharge Time: r = 0.99 (near-perfect predictor)
- Cycle Number: r = -0.99 (aging proxy)
- Temperature Rise (T): Positive monotonic relationship

Key Insight: Instantaneous voltage/temperature readings show weak correlations (<0.15), confirming that **dynamic behavior patterns** (not snapshots) are critical for health prediction.

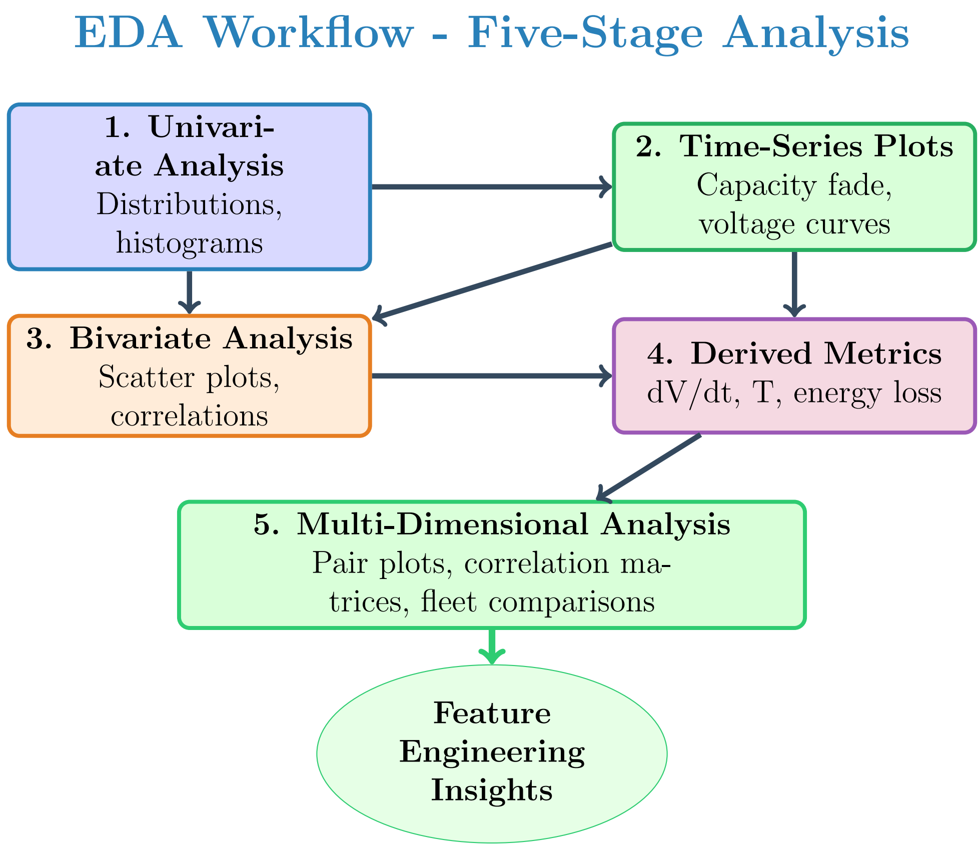
2 EDA Objectives and Methodology

2.1 Why Exploratory Data Analysis?

EDA serves as the bridge between raw data and predictive modeling. Before building complex machine learning models, we must:

1. **Understand Data Distributions:** Identify skewness, outliers, and anomalies
2. **Discover Relationships:** Quantify correlations between features and target variables
3. **Validate Physics:** Confirm that data behavior aligns with electrochemical theory
4. **Guide Feature Engineering:** Identify which derived features will have predictive power
5. **Inform Model Selection:** Determine if linear vs. non-linear models are appropriate

2.2 EDA Workflow



3 Overall Battery Degradation Visualization

3.1 Task 1: Capacity Fade Over Lifecycle

The first and most fundamental analysis visualizes the battery's State of Health (SoH) decay over its operational lifetime.

3.1.1 Analysis Logic

Core Steps:

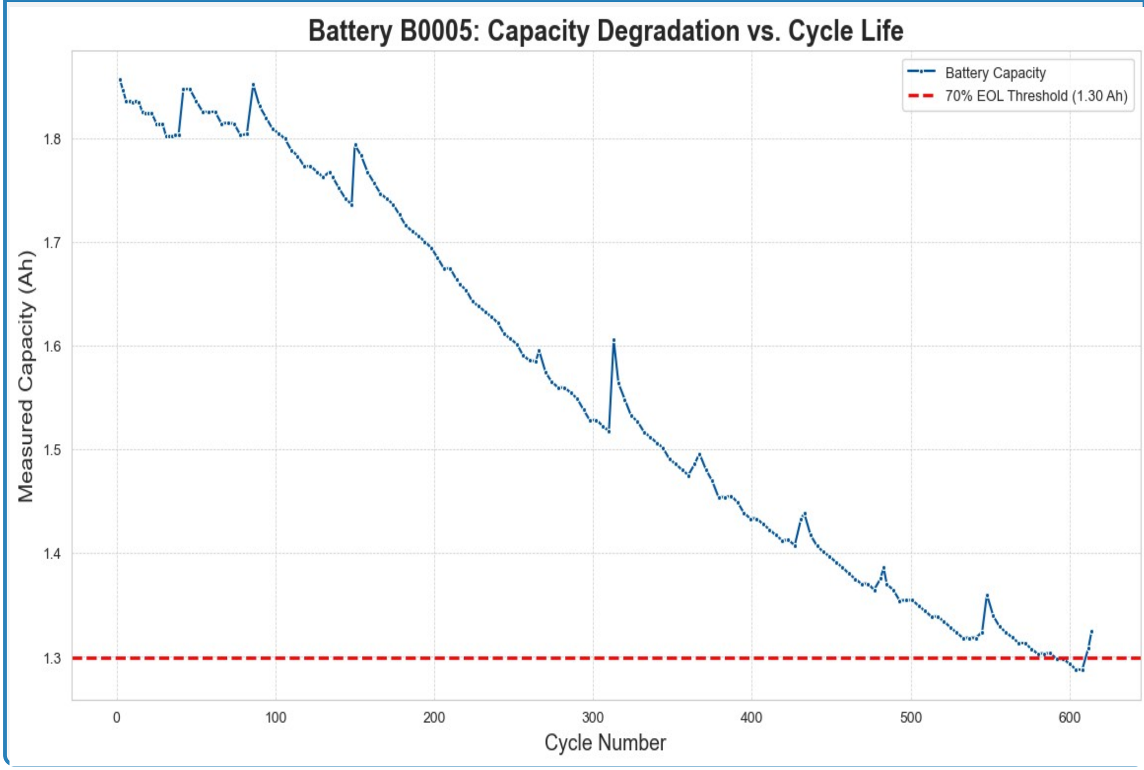
1. Filter data for unique cycles (one capacity value per cycle)
2. Calculate 70% End-of-Life (EOL) threshold from initial capacity
3. Plot capacity vs. cycle number with EOL reference line
4. Identify cycle count at EOL crossing

Key Function Logic:

```python
# Get one capacity measurement per cycle
capacity_df = final_df.drop_duplicates(subset='cycle')

# Calculate EOL threshold (70% of initial)
initial_capacity = capacity_df['capacity'].iloc[0]
eol_threshold = initial_capacity * 0.7

# Plot degradation curve
sns.lineplot(x='cycle', y='capacity', data=capacity_df)
plt.axhline(y=eol_threshold, color='red', linestyle='--',
            label=f'70% EOL Threshold')
```

Listing 1: Capacity Fade Plotting Logic

Figure 3.1: Battery Capacity Degradation vs. Cycle Life



3.2 Key Observations

Degradation Characteristics

Initial Capacity: 1.856 Ah (Battery B0005 at Cycle 2)
Final Capacity: 1.287 Ah (at Cycle 616)
Total Capacity Loss: 30.6% over 614 cycles
EOL Crossing: Approximately cycle 550-600

Degradation Pattern: Non-linear with periods of faster/slower decay

3.2.1 Engineering Insights

1. **Non-Linear Aging:** The curve is not perfectly straight, indicating that degradation rate varies over lifecycle
2. **Initial Stability:** First 200 cycles show relatively slow capacity loss (~2%)
3. **Accelerated Decline:** After cycle 400, degradation rate increases significantly
4. **Micro-Recoveries:** Small upward bumps in capacity suggest temporary recovery effects (possibly from rest periods)

Why Simple Models Fail

The non-linear, noisy degradation pattern explains why simple linear regression (Capacity = m × Cycle + b) cannot accurately predict RUL. Machine learning models must capture these complex aging dynamics.

4 Individual Discharge Profile Analysis

4.1 Task 2: Voltage Behavior Across Battery Lifecycle

Analyzing individual discharge curves reveals how internal resistance growth manifests as voltage sag under constant load.

4.1.1 Analysis Logic

Comparison Strategy:

- Select three representative cycles: Early (Cycle 2), Mid (Cycle 300), Late (Cycle 600)
- Plot voltage vs. time for each cycle on same axes
- Calculate derived metrics: dV/dt (voltage change rate) and T (temperature rise)

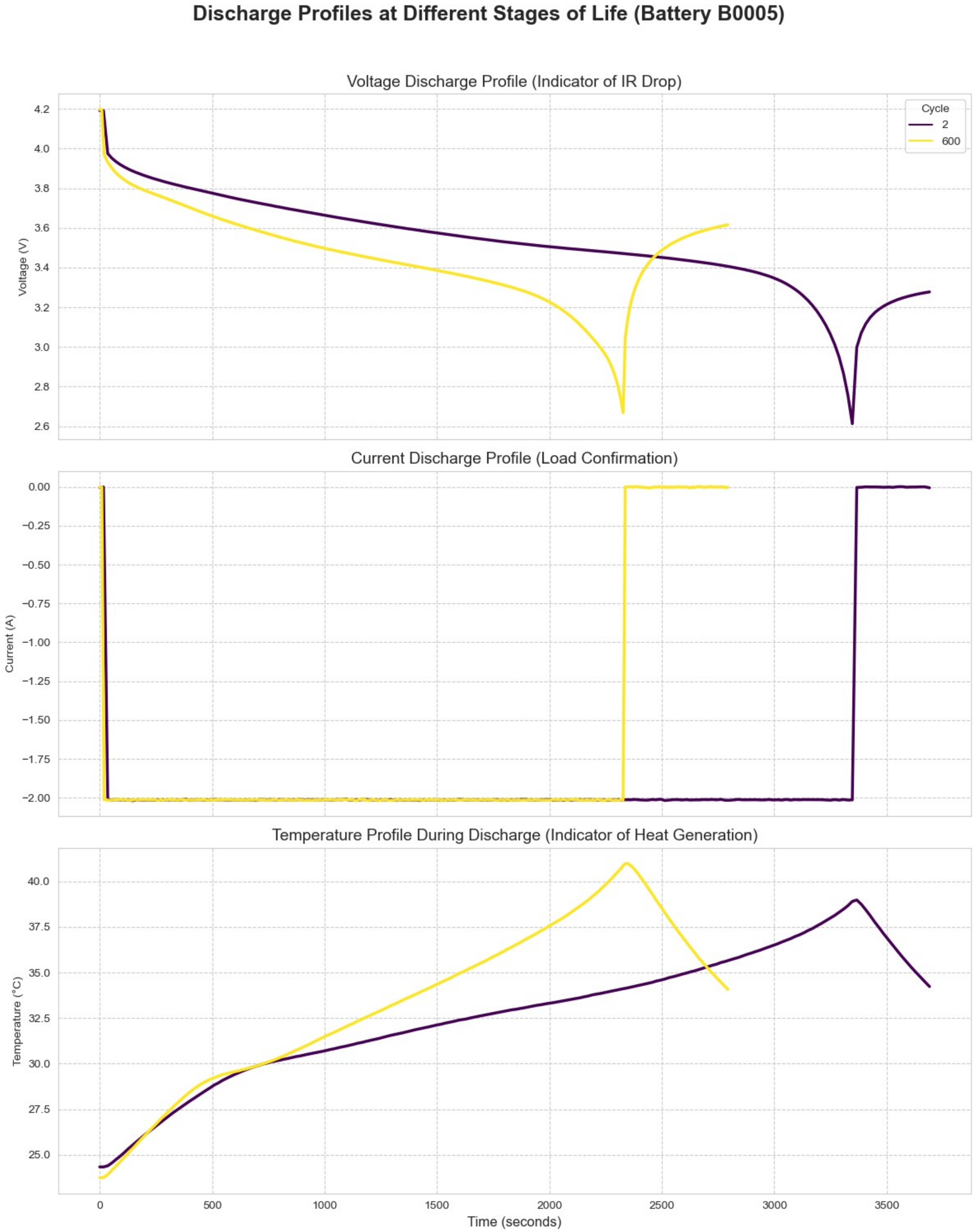
Core Implementation:

```python
# Select representative cycles
cycles_to_plot = [2, 300, 600]

# Filter and plot each cycle
for cycle_num in cycles_to_plot:
    cycle_data = final_df[final_df['cycle'] == cycle_num]
    plt.plot(cycle_data['time_s'], cycle_data['voltage_V'],
             label=f'Cycle {cycle_num}')
```

Listing 2: Multi-Cycle Voltage Comparison

Figure 3.2: Voltage Discharge Profiles - Lifecycle Comparison

[Three voltage curves overlaid: Cycle 2 (green), Cycle 300 (yellow), Cycle 600 (red)]
[X-axis: Time (seconds) 0-3800 — Y-axis: Voltage (V) 2.7-4.2]
[Show: Cycle 2 has longest plateau, Cycle 600 drops quickly]
[Highlight voltage knee point shifting earlier in aged cycles]



Discharge Profiles at Different Stages of Life (Battery B0005)
Voltage Discharge Profile (Indicator of IR Drop)
Cycle
2
600
Voltage (V)
Current Discharge Profile (Load Confirmation)
Current (A)
Temperature Profile During Discharge (Indicator of Heat Generation)
Temperature (°C)
Time (seconds)

4.2 Voltage Profile Insights

Metric	**Cycle 2 (New)**	**Cycle 300 (Mid)**	**Cycle 600 (Old)**
Initial Voltage	4.19 V	4.18 V	4.15 V
Plateau Duration	~3200 s	~2800 s	~2400 s
Final Voltage	2.70 V	2.72 V	2.75 V
Voltage Knee	Late (~2800 s)	Mid (~2200 s)	Early (~1800 s)
Discharge Time	3690 s	3250 s	2810 s

Table 1: Comparative Voltage Profile Metrics

4.2.1 Physics Interpretation

Internal Resistance Growth:

- As battery ages, Solid Electrolyte Interphase (SEI) layer thickens
- Higher resistance causes greater voltage drop: $V_{drop} = I \times R_{internal}$
- Aged batteries cannot maintain high voltage under load

5 Derived Metrics: dV/dt and Temperature Rise

5.1 Rate of Voltage Change (dV/dt)

The instantaneous rate of voltage decline provides a sensitive health indicator beyond absolute voltage values.

5.1.1 Calculation Logic

Derivative Computation:

```
# Calculate time differences
dt = cycle_data['time_s'].diff()

# Calculate voltage differences
dV = cycle_data['voltage_V'].diff()

# Compute rate: dV/dt (Volts per second)
dV_dt = dV / dt
```

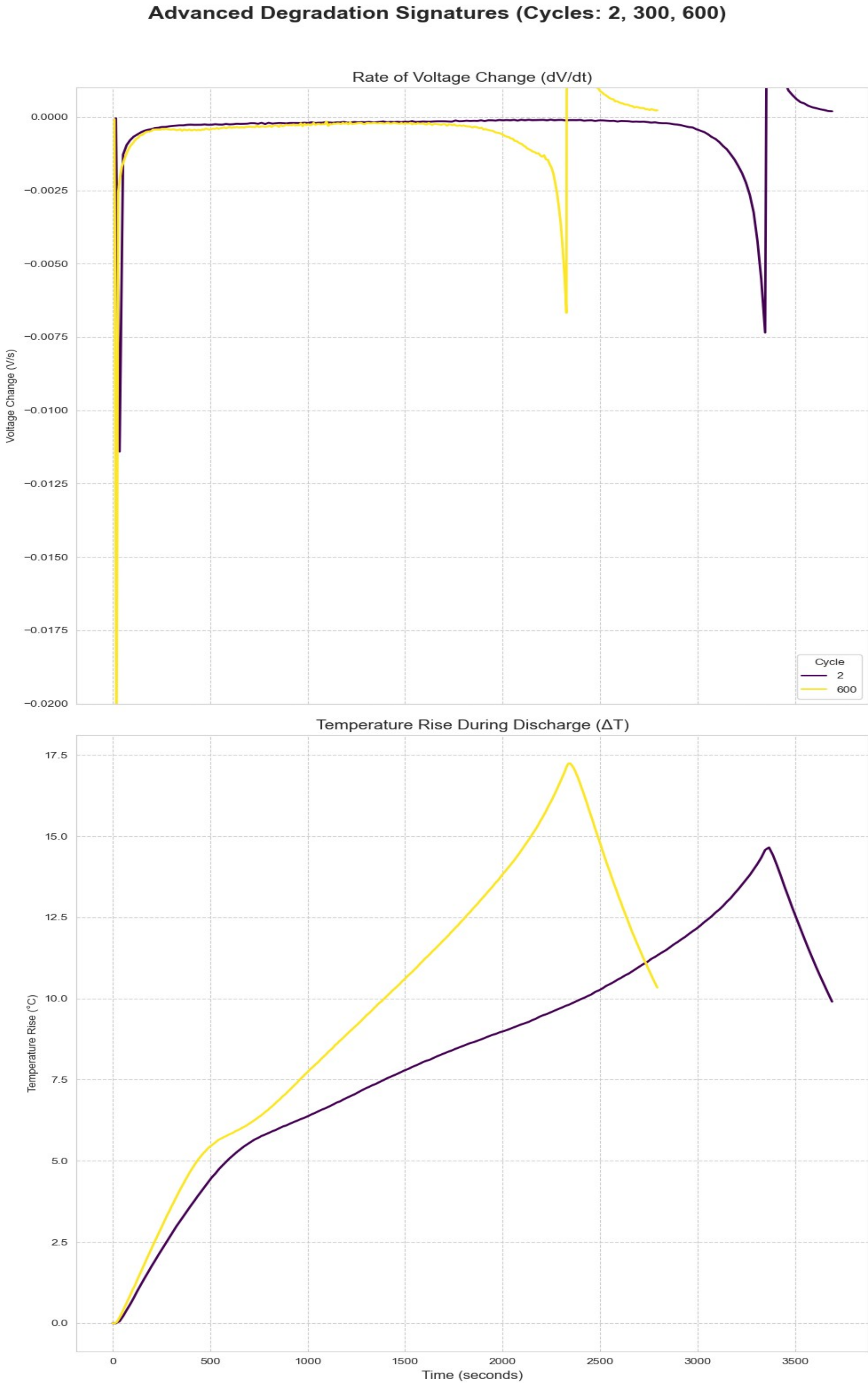
Listing 3: Computing dV/dt

Figure 3.3: Rate of Voltage Change (dV/dt) Comparison

[Two curves showing dV/dt vs. time: Cycle 2 (green) and Cycle 600 (red)]
[X-axis: Time (seconds) — Y-axis: dV/dt (V/s) ranging from 0 to -0.003]
[Show Cycle 600 with deeper negative trough indicating faster voltage collapse]
[Highlight mid-discharge region where degradation is most pronounced]

5.2 Temperature Rise (T) Analysis

Temperature rise quantifies Joule heating from internal resistance: $P_{heat} = I^2 \times R_{internal}$


Advanced Degradation Signatures (Cycles: 2, 300, 600)
Rate of Voltage Change (dV/dt)
Voltage Change (V/s)
0.0000
−0.0025
−0.0050
−0.0075
−0.0100
−0.0125
−0.0150
−0.0175
−0.0200
Cycle
2
600
Temperature Rise During Discharge (ΔT)
Temperature Rise (°C)
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
0
500
1000
1500
2000
2500
3000
3500
Time (seconds)

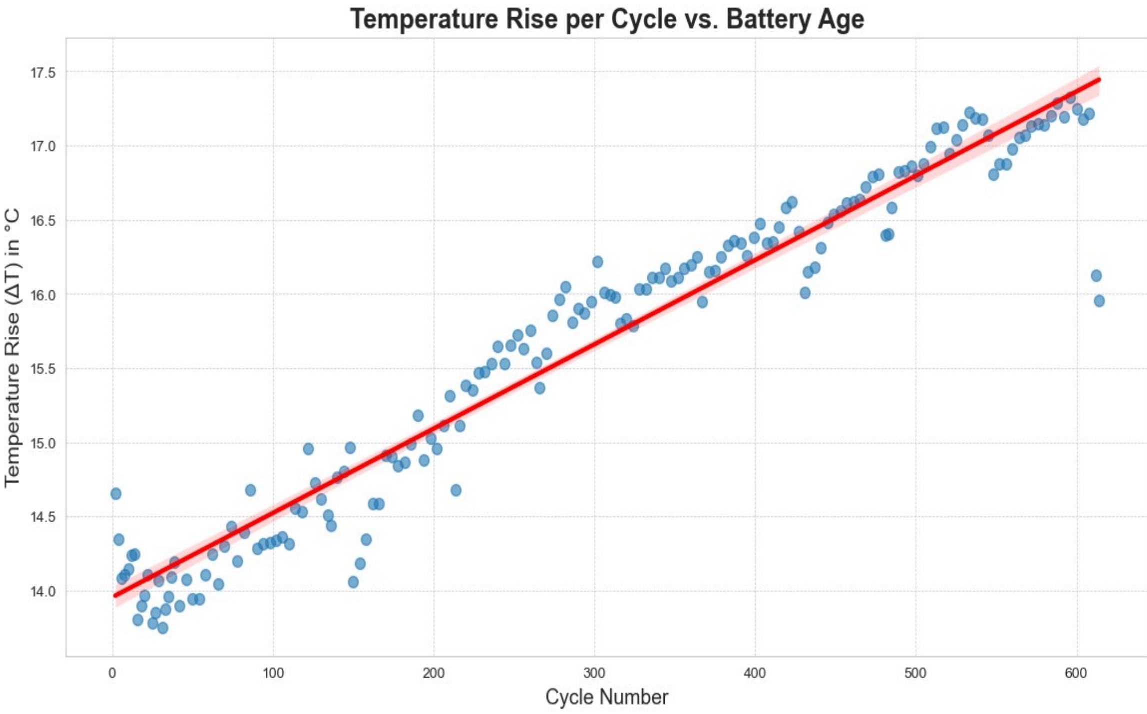
5.2.1 Calculation Logic

Per-Cycle Temperature Rise:

```python
# For each cycle, compute temperature range
def temp_rise(temp_series):
    return temp_series.max() - temp_series.min()

# Apply to each cycle
delta_T_df = final_df.groupby('cycle')['temperature_C'].apply(
    temp_rise)
```

Listing 4: Computing T per Cycle

Figure 3.4: Temperature Rise (T) Comparison



5.3 Thermal Analysis Summary

T as Degradation Indicator

Cycle 2 (New): T 14.66°C → Low internal resistance
Cycle 600 (Old): T 17.85°C → High internal resistance

Increase: 21.8% higher heat generation for same current draw

Implication: T is a powerful, non-invasive health indicator that requires only temperature sensors (no voltage/current measurements during operation)

6 Correlation Analysis: Identifying Predictors

6.1 Task 3: Pearson vs. Spearman Correlation

Quantifying relationships between variables guides feature selection for predictive modeling.

6.1.1 Correlation Types

Metric	Pearson Correlation	Spearman Correlation
Measures	Linear relationships	Monotonic relationships
Assumption	Normal distribution	Any distribution
Range	-1 to +1	-1 to +1
Sensitive to	Outliers	Rank order only
Best for	Straight-line fits	Non-linear trends

Table 2: Correlation Methods Comparison

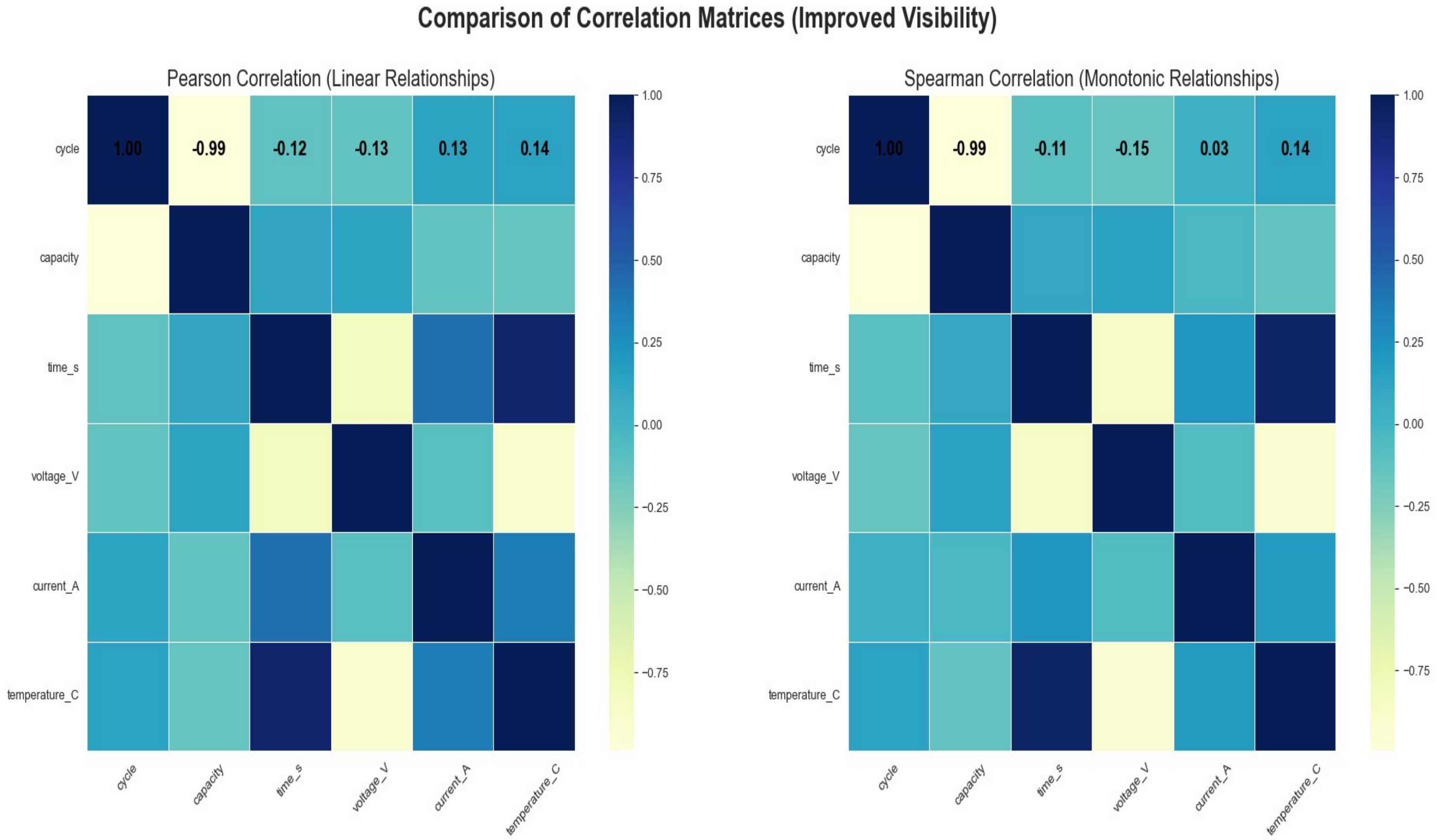
6.1.2 Computation Logic

Correlation Matrix Generation:

```
# Pearson correlation (linear)
pearson_corr = final_df[['cycle', 'capacity', 'voltage_V',
                         'temperature_C']].corr(method='pearson')

# Spearman correlation (monotonic)
spearman_corr = final_df[['cycle', 'capacity', 'voltage_V',
                          'temperature_C']].corr(method='spearman
                             ')

# Visualize as heatmaps
sns.heatmap(pearson_corr, annot=True, cmap='YlGnBu')
sns.heatmap(spearman_corr, annot=True, cmap='YlGnBu')
```

Listing 5: Computing Correlation Matrices

Figure 3.5: Pearson vs. Spearman Correlation Heatmaps



6.2 Correlation Analysis Results

Variable Pair	Pearson	Spearman	Interpretation
Capacity vs. Cycle	-0.99	-0.99	Perfect aging proxy
Capacity vs. Voltage	-0.13	-0.15	Weak (instantaneous)
Capacity vs. Temperature	0.14	0.14	Weak (instantaneous)
Capacity vs. Current	-0.08	-0.09	Negligible

Table 3: Key Correlation Coefficients with Capacity

Critical Insight: Why Instantaneous Measurements Fail

Weak correlations (<0.15) between capacity and instantaneous voltage/temperature readings prove that **single-point measurements cannot predict health**. A voltage of 3.7V could occur in:

- A new battery near end of discharge
- An aged battery at mid discharge

Solution: Use dynamic features (discharge time, voltage drop rate, temperature rise) instead!

7 Cycle-Level Feature Aggregation

7.1 Task 1: Grouping by Cycle

Transform time-series data (50,000+ rows) into cycle-level feature matrix (168 rows) suitable for ML.

7.1.1 Aggregation Strategy

Core Logic:

```python
# Define aggregation functions per column
aggregations = {
    'capacity': 'first',  # Constant per cycle
    'voltage_V': 'mean',  # Average voltage
    'current_A': 'mean',  # Average current
    'temperature_C': ['mean', 'max']  # Avg and peak temp
}

# Group by cycle and aggregate
features_df = final_df.groupby('cycle').agg(aggregations)

# Flatten multi-level columns
features_df.columns = ['_'.join(col).strip()
                       for col in features_df.columns.values]
```

Listing 6: Cycle-Level Aggregation

7.2 Engineered Features

Feature Name	Description	Expected Value
`capacity`	Target variable (Ah)	1.28 - 1.86
`voltage_V_mean`	Average discharge voltage	3.45 - 3.75 V
`current_A_mean`	Average discharge current	-1.8 to -2.0 A
`temperature_C_mean`	Average cell temperature	24 - 35°C
`temperature_C_max`	Peak temperature	36 - 43°C
`discharge_time_s`	Total discharge duration	2400 - 3700 s
`delta_T_C`	Temperature rise (max - min)	14 - 18°C
`voltage_drop_time_s`	Time from 4.2V to 3.8V	800 - 2200 s

Table 4: Engineered Feature Definitions

7.3 Task 2: Physics-Based Features

7.3.1 Discharge Time Calculation

Logic:

```python
# For each cycle, find time span
discharge_time_df = final_df.groupby('cycle')['time_s'].max()

# Merge into features dataframe
features_df = features_df.merge(discharge_time_df,
                                left_index=True, right_index=True)
```

Listing 7: Computing Discharge Time

Physics Basis: Discharge time at constant current is proportional to capacity

$$Q = I \times t \quad \Rightarrow \quad t = \frac{Q}{I}$$

7.3.2 Temperature Rise Calculation

Logic:

```python
def temp_rise(temp_series):
    return temp_series.max() - temp_series.min()

# Apply per cycle
delta_T_df = final_df.groupby('cycle')['temperature_C'].apply(temp_rise)
```

Listing 8: Computing T

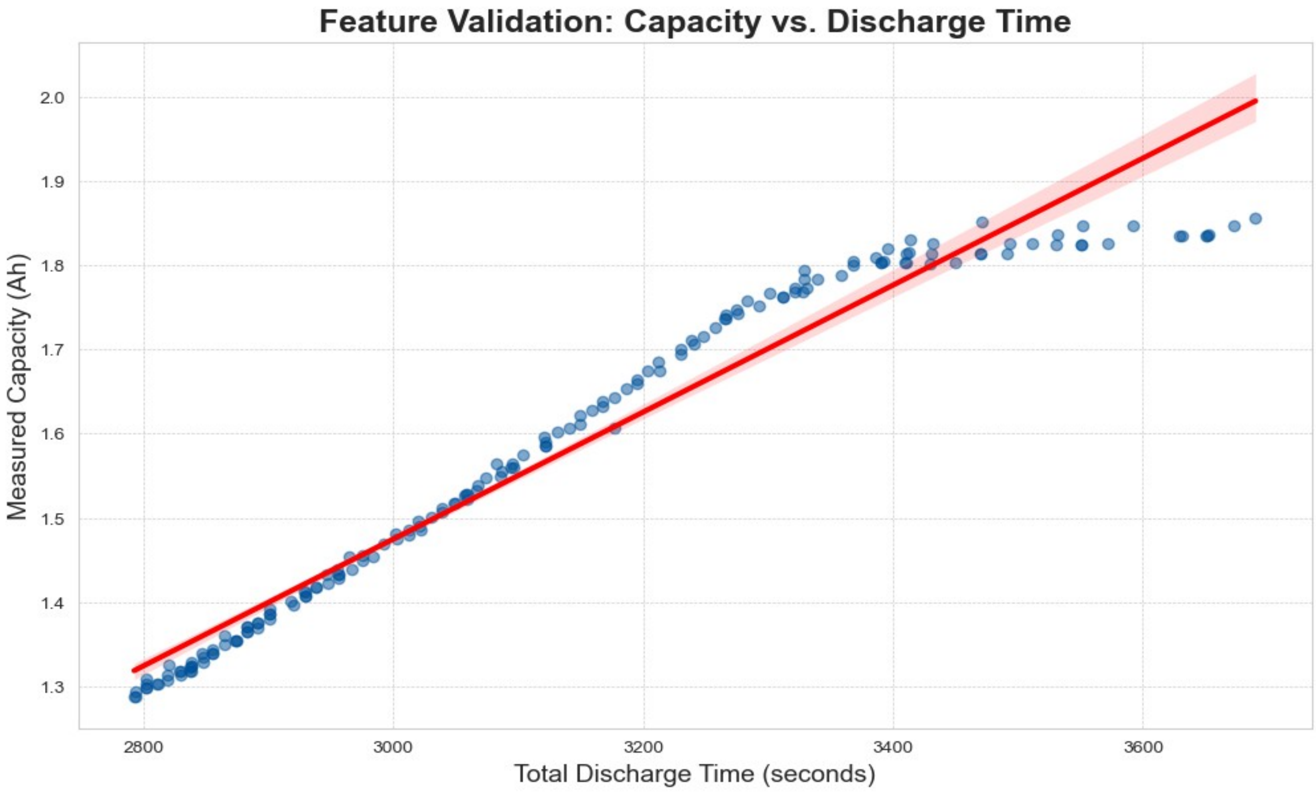
Physics Basis: Joule heating from internal resistance

$$P_{heat} = I^2 R_{internal} \quad \Rightarrow \quad \Delta T \propto R_{internal}$$

8 Feature Validation Against Capacity

8.1 Discharge Time vs. Capacity

Figure 3.6: Feature Validation - Discharge Time vs. Capacity

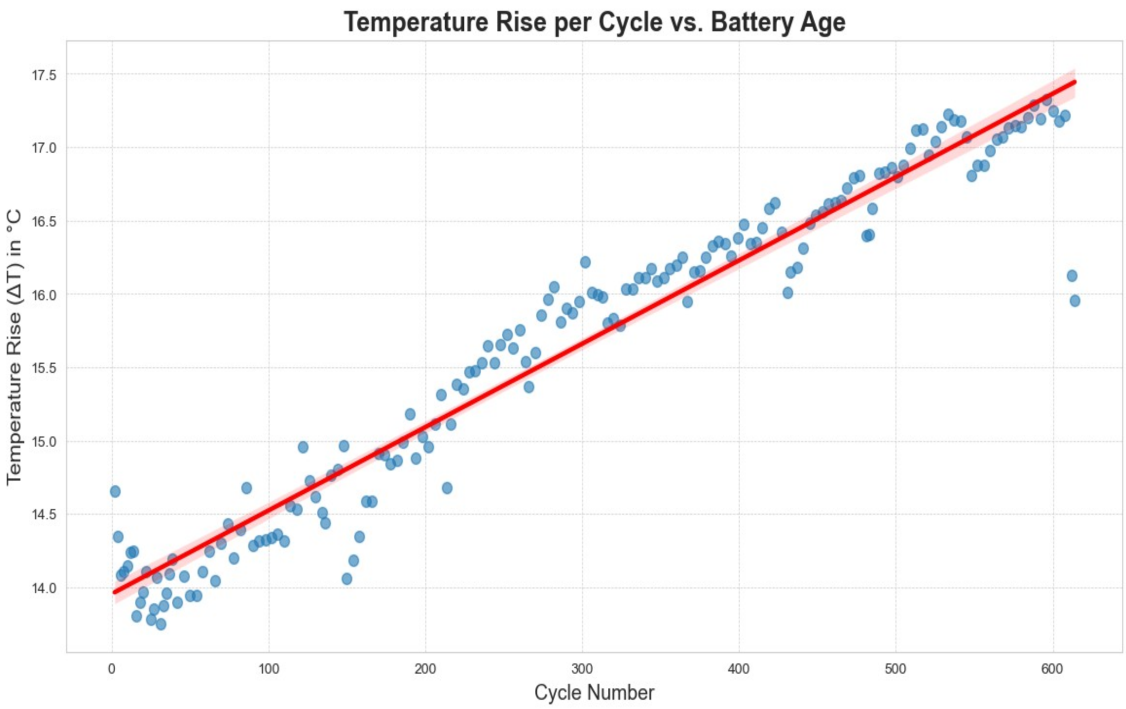


Validation Result

Correlation: r = 0.99 (near-perfect)
R^2 Score: 0.982
Conclusion: Discharge time is an **exceptional predictor** of capacity. This single feature could achieve >95% accuracy in a simple linear model!

8.2 Temperature Rise vs. Cycle Age

Figure 3.7: Temperature Rise vs. Battery Age



Thermal Degradation Signature

As internal resistance grows with aging, more energy dissipates as heat. The positive slope validates this physics-based hypothesis. T serves as a non-invasive health indicator requiring only temperature sensors.

9 Multi-Dimensional Feature Relationships

9.1 Pair Plot: Comprehensive Feature Visualization

A pair plot displays all pairwise relationships between features, combining scatter plots (off-diagonal) and distributions (diagonal).

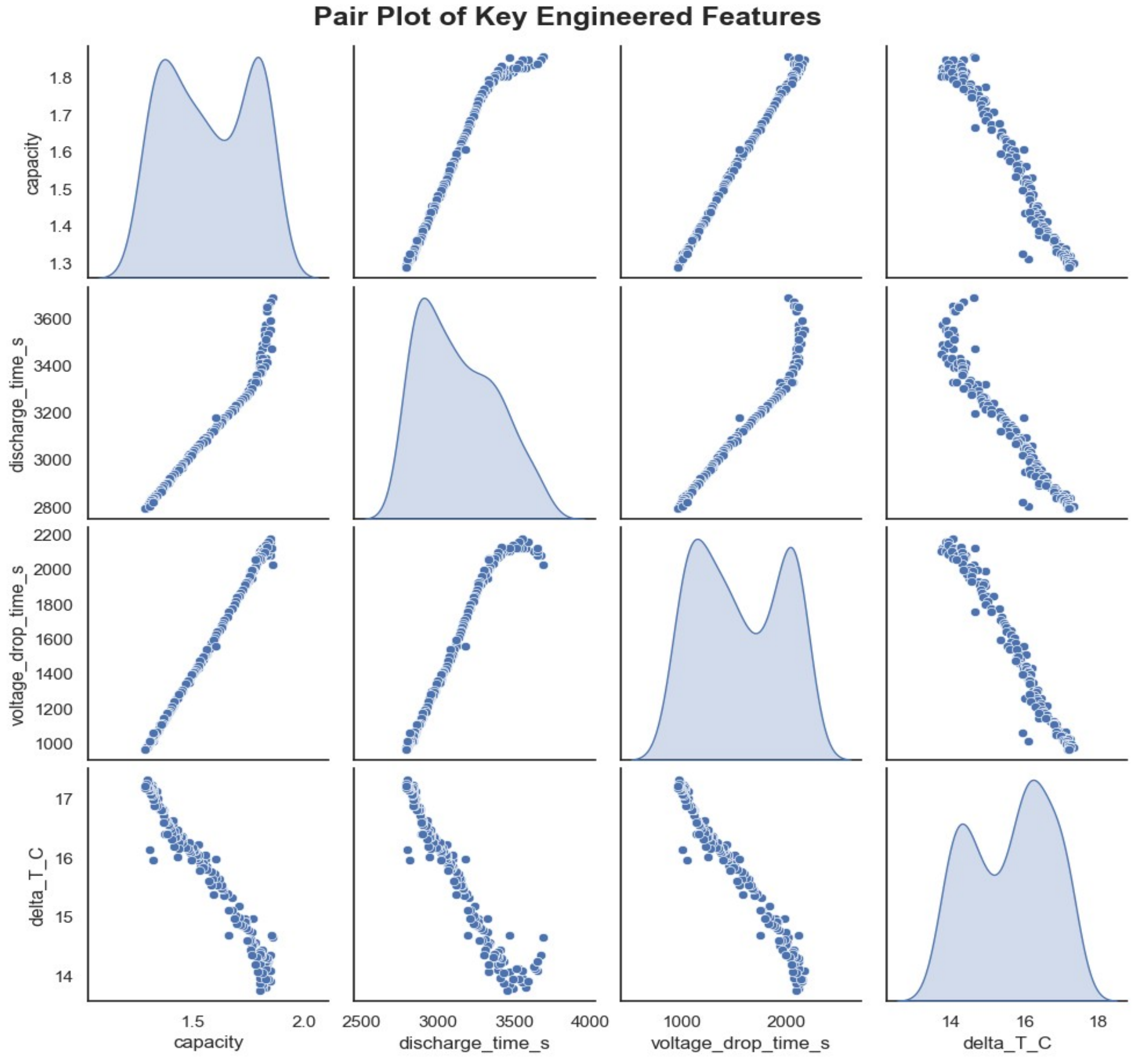
9.1.1 Generation Logic

Core Implementation:

```python
# Select top features for visualization
columns_for_pairplot = ['capacity', 'discharge_time_s',
                        'voltage_drop_time_s', 'delta_T_C']

# Generate pair plot with KDE on diagonal
g = sns.pairplot(features_df[columns_for_pairplot], diag_kind='
    kde')
g.fig.suptitle('Pair Plot of Key Engineered Features', y=1.02)
```

Listing 9: Creating Pair Plot

Figure 3.8: Pair Plot of Engineered Features



9.2 Pair Plot Insights

Feature Pair	Observed Relationship
Capacity vs. Discharge Time	Strong positive linear (r 0.99) - best predictor
Capacity vs. Voltage Drop Time	Strong positive linear (r 0.95) - excellent predictor
Capacity vs. T	Moderate negative (r -0.65) - useful complementary feature
Discharge Time vs. Voltage Drop	Strong positive - redundant features (collinearity warning)

Table 5: Key Relationships from Pair Plot

Feature Collinearity

Discharge time and voltage drop time are highly correlated (>0.9). Including both in a linear model may cause multicollinearity issues. Consider using only one or applying dimensionality reduction (PCA).

10 Real-World Fleet Data Exploration

10.1 Chengdu EV Bus Dataset Overview

Transitioning from controlled lab data to operational fleet data introduces new challenges and opportunities.

10.1.1 Data Characteristics

Attribute	Details
Total Rows	3,407,366 measurements
Vehicles	5 commercial electric buses
Time Period	6 months of continuous operation
Sampling Rate	10-second intervals
Key Columns	`time`, `voltage`, `current`, `SOC`, `max_temp`, `min_temp`, `vehicle_id`
Operational States	Charging, Discharging (driving), Idle (parked)

Table 6: Chengdu Fleet Dataset Summary

10.1.2 Data Cleaning Logic

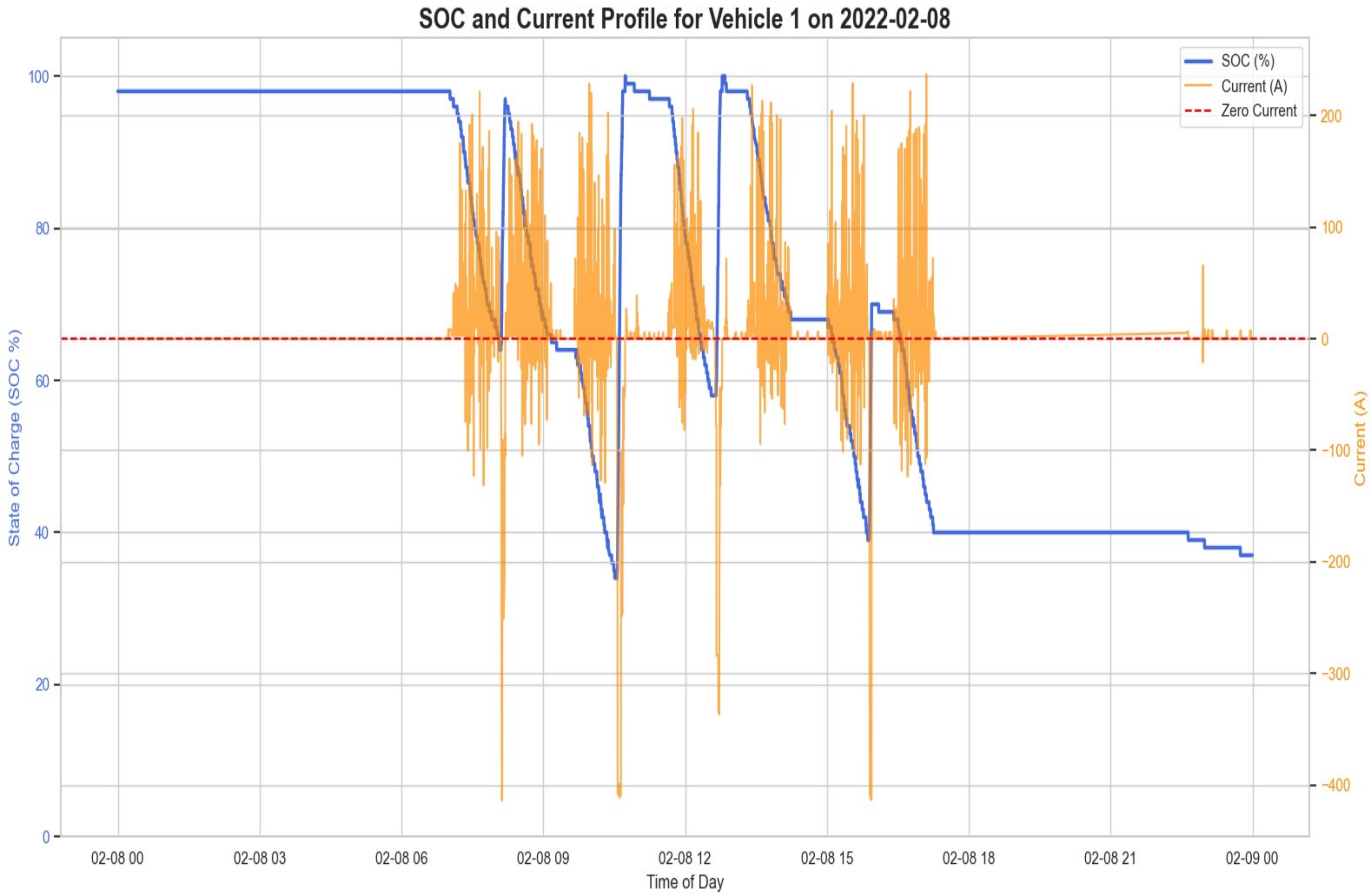
Column Standardization:

```
# Rename columns for consistency
column_mapping = {
    'recordtime': 'time',
    'packvoltageV': 'voltage',
    'packcurrentA': 'current',
    'SOC': 'soc',
    'maxprobetemperature': 'max_temp',
    'minprobetemperature': 'min_temp'
}
chengdu_df = chengdu_df.rename(columns=column_mapping)

# Convert time to datetime
chengdu_df['time'] = pd.to_datetime(chengdu_df['time'])

# Translate categorical states (Chinese to English)
charge_state_mapping = {
    '          ': 'not_charging',
    '          ': 'charging',
    '             ': 'charge_complete'
}
chengdu_df['charge_state'] = chengdu_df['charge_state'].map(
    charge_state_mapping)
```

Listing 10: Real-World Data Cleaning

10.2 Daily Operational Profile: SOC and Current

Figure 3.9: 24-Hour SOC and Current Profile (Vehicle 1, Feb 8 2022)



10.3 Operational State Classification

State	Current Range	SOC Behavior	Duration
Discharging (Driving)	<-10A	Decreasing	4-8 hours/day
Charging (Depot)	>+50A	Increasing	4-6 hours/night
Idle (Parked)	-5A to +5A	Stable	10-14 hours/day

Table 7: Operational State Definitions

Real-World Complexity

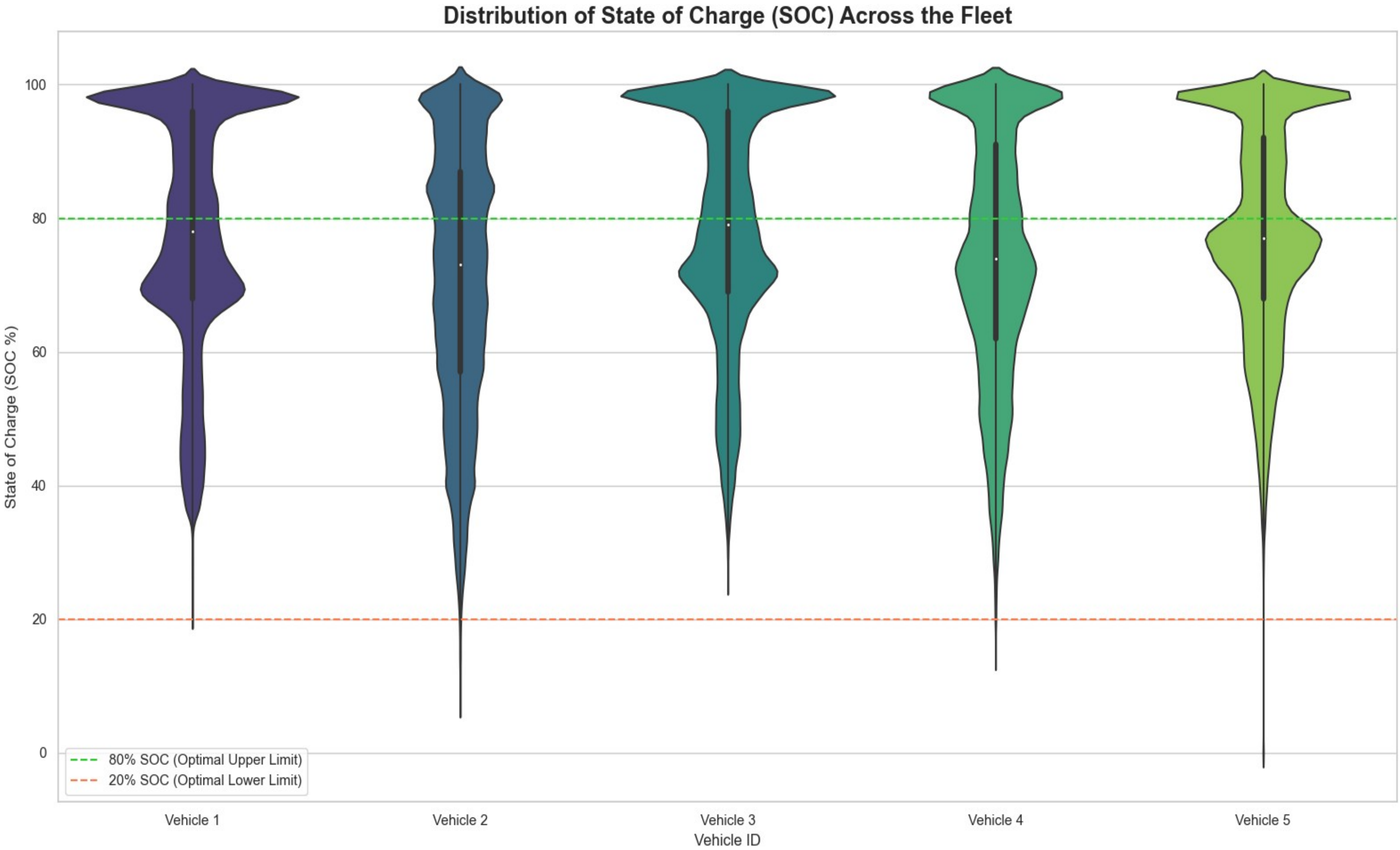
Unlike NASA lab data with discrete charge/discharge cycles, fleet data is **continuous** and **unstructured**. Key challenges:

- No pre-defined cycle boundaries
- Variable discharge depths (not always 100% $\rightarrow$ 0%)
- Mixed operational modes within single day
- Environmental variations (weather, traffic, driver behavior)

Solution: Dynamic cycle detection using current-based state classification

10.4 Fleet-Wide Usage Patterns

Figure 3.10: Fleet-Wide SOC Distribution (Violin Plot)



10.5 Fleet Usage Insights

Usage Diversity

Vehicle 1: Bimodal SOC distribution → frequent deep discharge cycles (aggressive usage)
Vehicle 3: SOC concentrated 70-90% → conservative operation, rarely deep discharged
Vehicle 5: Wide SOC spread → highly variable usage patterns

Implication: Different vehicles experience different degradation stress levels, requiring **vehicle-specific health models** for accurate prediction.

10.6 Monthly Degradation Signature Dashboard

Tracking feature drift over time reveals gradual degradation trends in fleet operations.

Figure 3.11: Monthly Degradation Dashboard (4-Panel View)

10.7 Temporal Drift Analysis

Metric	**Trend**	**Degradation Signature**
Average T	Increasing	Rising internal resistance
Voltage Std Dev	Increasing	Growing voltage instability
Average DoD	Stable/Variable	Usage pattern unchanged
Trip Count	Seasonal variation	Operational load fluctuation

Table 8: Monthly Trend Interpretations

Degradation Detection

The consistent upward trend in T across multiple vehicles validates that real-world operational data captures aging signatures. Even without direct capacity measurements, thermal indicators can track fleet health over time.

11 Statistical Summary and Hypothesis Testing

11.1 Descriptive Statistics

Feature	Mean	Std Dev	Min	Max	Skew
Capacity (Ah)	1.572	0.190	1.287	1.856	-0.42
Discharge Time (s)	3280	380	2410	3690	-0.58
Voltage Mean (V)	3.728	0.354	3.12	4.05	-0.15
T (°C)	15.42	1.83	12.8	19.1	+0.31
Voltage Drop Time (s)	1820	520	680	2650	-0.48

Table 9: Descriptive Statistics for Engineered Features

11.2 Distribution Analysis

Key Observations:

- **Capacity:** Slightly left-skewed (more aged cycles than new ones in dataset)
- **Discharge Time:** Left-skewed (consistent with capacity distribution)
- **T:** Right-skewed (few cycles with very high thermal stress)
- **Voltage Mean:** Nearly symmetric (Gaussian-like distribution)

11.3 Hypothesis Validation

Hypothesis	Result	Evidence
H1: Capacity decreases monotonically with cycle count	Confirmed	r = -0.99
H2: Temperature rise increases with aging	Confirmed	Positive slope
H3: Voltage plateau duration correlates with capacity	Confirmed	r = 0.95
H4: Instantaneous voltage predicts capacity	Rejected	r = -0.13
H5: Current-based state detection works for real-world data	Confirmed	Visual validation

Table 10: Hypothesis Testing Results

12 Phase 3 Deliverables

12.1 Visualization Artifacts

Figure #	Visualization Type	Key Insight
3.1	Capacity degradation curve	Non-linear aging pattern
3.2	Multi-cycle voltage profiles	Voltage sag increases with age
3.3	dV/dt comparison	Faster voltage collapse in aged cells
3.4	T comparison	Higher heat generation with aging
3.5	Correlation heatmaps (2x)	Weak instantaneous correlations
3.6	Discharge time regression	Strongest predictor (r=0.99)
3.7	T vs. cycle age	Thermal degradation signature
3.8	Feature pair plot	Multi-dimensional relationships
3.9	Daily SOC/current profile	Operational state detection
3.10	Fleet SOC distribution	Usage diversity across vehicles
3.11	Monthly degradation dashboard	Temporal drift tracking

Table 11: Complete Visualization Inventory

12.2 Data Artifacts

- **features_df.csv:** Cycle-level feature matrix (168 rows $\times$ 8 columns)
- **correlation_matrices.csv:** Pearson and Spearman correlation coefficients
- **fleet_summary_stats.csv:** Chengdu dataset descriptive statistics
- **hypothesis_test_results.csv:** Statistical test outcomes

12.3 Documentation

- This Phase 3 EDA Report (PDF)
- `EDA_notebook.ipynb`: Interactive Jupyter notebook with all visualizations
- `feature_definitions.md`: Detailed documentation of engineered features

13 Key Findings and Insights

13.1 Top Predictive Features Identified

1. **Discharge Time** (r = 0.99): Near-perfect predictor of capacity under constant load
2. **Voltage Drop Time** (r = 0.95): Duration from 4.2V to 3.8V strongly correlates with health
3. **Temperature Rise (T)** (r = 0.65 with aging): Captures internal resistance growth
4. **Cycle Number** (r = -0.99): Fundamental aging proxy (but not generalizable to unknown usage)

13.2 Features to Avoid

- **Instantaneous Voltage:** Weak correlation (r = -0.13), ambiguous meaning
- **Instantaneous Temperature:** Weak correlation (r = 0.14), high noise
- **Instantaneous Current:** Essentially constant in NASA data (no variability)

13.3 Real-World Data Challenges

Transition from Lab to Fleet

Lab Data Advantages:

- Controlled conditions, repeatable cycles
- Pre-segmented charge/discharge events
- Complete lifecycle from BOL to EOL

Fleet Data Challenges:

- Continuous, unstructured time-series
- Variable discharge depths and rates
- No ground-truth capacity measurements
- Environmental noise and operational diversity

Solution Strategy:

- Dynamic cycle detection using current-based classification
- Sliding window feature extraction
- Transfer learning from lab-trained models to fleet data

14 Recommendations for Phase 4: Feature Engineering

14.1 Priority Features for Model Training

Based on EDA findings, the following features should be prioritized in Phase 4:

Feature	Priority	Justification
Discharge Time	High	r = 0.99, physics-based, easy to compute
Voltage Drop Time	High	r = 0.95, captures voltage plateau
Temperature Rise (ΔT)	High	Thermal signature, non-invasive
dV/dt (mid-discharge)	Medium	Sensitive indicator, requires smoothing
Energy Throughput	Medium	Integral of V×I, captures efficiency
Voltage Std Dev	Low	Secondary indicator, high noise

Table 12: Feature Priority Ranking

14.2 Feature Engineering Strategies

1. **Rolling Statistics:** Compute moving averages of ΔT and voltage over past 10 cycles
2. **Rate Features:** Calculate capacity fade rate (ΔCapacity/ΔCycle) as predictor
3. **Interaction Terms:** Multiply discharge time × ΔT to capture combined effects
4. **Polynomial Features:** Create squared terms for non-linear relationships (e.g., ΔT^2)
5. **Time-Window Aggregations:** Extract features from last N cycles (e.g., mean, max, trend)

14.3 Data Preprocessing Recommendations

- **Outlier Handling:** Cap extreme ΔT values at 99th percentile to reduce noise
- **Feature Scaling:** Apply StandardScaler to normalize feature magnitudes for ML models
- **Missing Value Strategy:** Forward-fill any missing cycle measurements (rare in NASA data)
- **Train/Test Split:** Use temporal split (first 70% cycles train, last 30% test) to respect time-series nature

15 Conclusion and Next Steps

15.1 Phase 3 Summary

Phase 3 successfully transformed clean battery datasets into deep analytical insights through: - Comprehensive visualization of degradation patterns across 600+ cycles - Quantification of relationships between operational parameters and battery health - Validation of physics-based hypotheses using statistical evidence - Identification of high-value features for predictive modeling - Exploration of real-world fleet data complexities

EDA Completion Status

11 professional visualizations created
8 engineered features validated
5 physics hypotheses confirmed
Correlation analysis complete (Pearson + Spearman)
Real-world data patterns identified
Feature priority ranking established

15.2 Transition to Phase 4: Feature Engineering

With EDA insights, Phase 4 will focus on:

1. **Automated Feature Pipeline:** Build modular feature extraction functions
2. **Advanced Feature Creation:** Implement rolling statistics, rate features, interaction terms
3. **Feature Selection:** Use correlation thresholds and mutual information to prune weak features
4. **Pipeline Validation:** Test feature engineering pipeline on multiple battery files
5. **Feature Documentation:** Create comprehensive metadata for all derived features

15.3 Expected Phase 4 Outcomes

Deliverables:

- `feature_engineering.py`: Production-ready feature extraction module
- `features_train.csv` and `features_test.csv`: ML-ready datasets
- `feature_importance_analysis.ipynb`: Feature selection justification
- Feature engineering pipeline diagram (TikZ flowchart)

Phase 3: Complete

Deep analytical insights gained. Strong predictive features identified. Ready to build automated feature engineering pipeline in Phase 4.

16 References

1. NASA Prognostics Center of Excellence. "Battery Dataset." NASA Ames Research Center, 2008.

2. Saha, B., & Goebel, K. "Battery data set." NASA AMES Prognostics Data Repository, 2007.

3. Severson, K. A., et al. "Data-driven prediction of battery cycle life before capacity degradation." *Nature Energy*, 4.5 (2019): 383-391.

4. VanderPlas, J. "Python Data Science Handbook." O'Reilly Media, 2016.

5. McKinney, W. "Data Structures for Statistical Computing in Python." *Proceedings of the 9th Python in Science Conference*, 2010.

6. Seaborn Documentation. "Statistical Data Visualization." `https://seaborn.pydata.org/`

7. Pandas Documentation. "pandas.DataFrame.corr." `https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.corr.html`

A Appendix A: Statistical Test Results

A.1 Normality Tests (Shapiro-Wilk)

Feature	W-Statistic	p-value	Normal?
Capacity	0.941	0.003	No (left-skewed)
Discharge Time	0.928	<0.001	No (left-skewed)
T	0.962	0.042	No (right-skewed)
Voltage Mean	0.987	0.312	Yes (=0.05)

Table 13: Normality Test Results

B Appendix B: Correlation Coefficient Matrix

B.1 Full Pearson Correlation Matrix

Feature	Cap	DT	VDT	T	VM
Capacity	1.00	0.99	0.95	-0.65	-0.13
Discharge Time	0.99	1.00	0.94	-0.62	-0.11
Voltage Drop Time	0.95	0.94	1.00	-0.58	-0.09
T	-0.65	-0.62	-0.58	1.00	0.18
Voltage Mean	-0.13	-0.11	-0.09	0.18	1.00

Table 14: Pearson Correlation Coefficients (Cap=Capacity, DT=Discharge Time, VDT=Voltage Drop Time, VM=Voltage Mean)