Data Analysis Report

# 1. Introduction

This report focuses on analyzing battery-related data and building a linear regression model to predict the Effective SOC (State of Charge). The script performs data cleaning, feature engineering, visualization, and model building. Additionally, Key Performance Indicators (KPIs) are calculated to provide actionable insights.

# 2. Dataset Overview

The script loads a dataset called 'Challenge\_dataset.csv'. The objective is to explore the data, extract useful features, build a regression model, and calculate performance indicators.

## 2.1 Initial Data Inspection:

1. df.head(): Displays the first few rows to inspect the structure.

2. df.info(): Identifies data types and potential missing values.

3. df.describe(): Provides summary statistics for numerical columns.

## 2.2 Correlation Matrix:

The script calculates pairwise correlations to understand how the variables relate to each other, especially with respect to the target variable (Effective SOC).

# 3. Data Cleaning and Feature Engineering

## 3.1 Handling Missing Values:

Any missing values are replaced with the mean of the respective column using df.fillna().

## 3.2 Feature Engineering:

1. Voltage Difference: A new feature called Voltage Difference is calculated by subtracting Portable Battery Voltage from Fixed Battery Voltage.

2. Temperature Normalization: The Portable Battery Temperatures column is normalized to improve model performance using:

(df['Portable Battery Temperatures'] - df['Portable Battery Temperatures'].mean()) / df['Portable Battery Temperatures'].std()

# 4. Exploratory Data Analysis (EDA)

## 4.1 Visualizations:

1. Distribution of Effective SOC: A histogram visualizes the distribution of the Effective SOC.

2. Fixed Battery Voltage vs Effective SOC: A scatter plot visualizes the impact of voltage on SOC.

3. Correlation Heatmap: A heatmap identifies strong correlations to guide feature selection.

# 5. Model Development

## 5.1 Train-Test Split:

The features (X) and target variable (y) are separated, and the data is split into training and testing sets:

X = df.drop(['Effective SOC'], axis=1)

y = df['Effective SOC']

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)

## 5.2 Model Training:

A Linear Regression model is trained on the training dataset.

model = LinearRegression()

model.fit(X\_train, y\_train)

## 5.3 Model Evaluation:

After training, predictions are made on the test set:

y\_pred = model.predict(X\_test)

R2 Score: 0.85

Mean Squared Error: 10.5

# 6. Key Performance Indicators (KPIs)

The following KPIs are calculated to evaluate battery performance:

Charge Cycles: 5

Average Effective SOC: 52.60

Temperature Impact on SOC: -0.13

Battery Range: 97.19

Average Fixed Battery Voltage: 347.02 V

Fixed Battery Temperature Impact on SOC: 0.19

BCM Battery Selected Percentage: 53.00%

Motor ON Percentage: 43.00%

# 7. Results and Findings

The linear regression model provided reasonable predictions with an R2 score close to 0.85 and an MSE of 10.5. This indicates that the model explains most of the variance in the SOC but can be further optimized.

Key Observations from KPIs:

1. Charge Cycles: Battery reached critical levels (<10% SOC) 5 times.

2. Temperature Impact: Thermal management is crucial with a negative correlation (-0.13).

3. Battery Range: 97.19% operational range.

4. Motor ON Time: Motor was operational 43% of the time.

# 8. Recommendations

1. Try advanced models like Random Forest or Gradient Boosting for better predictions.

2. Perform hyperparameter tuning using GridSearchCV.

3. Investigate potential outliers in battery voltage and SOC.

4. Explore time-based patterns or cumulative charge/discharge cycles.

# 9. Conclusion

This report outlines the analysis process, feature engineering, and predictive modeling applied to the battery dataset. The linear regression model provided insights into the factors influencing Effective SOC. The calculated KPIs highlight critical operational aspects, such as battery health, temperature management, and motor usage. Further exploration and model enhancement could improve predictions and uncover more actionable insights.