# Data Analyst Hiring Challenge Report

## Objective:

The primary objective of this analysis is to thoroughly investigate the provided dataset and develop a robust machine learning model that accurately predicts the Effective State of Charge (SOC) of the battery. This report will also generate key performance indicators (KPIs) to evaluate the model's performance and the overall characteristics of the dataset.

## A. Data Analysis

### 1. Exploration of the Dataset:

Summary Statistics: A comprehensive analysis of the dataset reveals key statistics such as the mean, median, and standard deviation for each feature. This initial exploration helps to understand the central tendencies and variability within the data.

Correlation Analysis: By constructing a correlation matrix, we can identify the relationships between the featured columns (e.g., Fixed Battery Voltage, Portable Battery Voltage, etc.) and the target variable, Effective SOC. This analysis is crucial in determining which features may significantly influence the prediction of the target.

Visualizations: To further enhance our understanding, various visualizations, including histograms to depict distributions, scatter plots to show relationships, and box plots to highlight potential outliers, will be generated. These visual tools facilitate a clearer interpretation of the data and its characteristics.

### 2. Featured Columns:

The analysis focuses on the following features that are believed to have a strong relationship with Effective SOC:  
- Fixed Battery Voltage  
- Portable Battery Voltage  
- Portable Battery Current  
- Fixed Battery Current  
- Motor Status (On/Off)  
- BCM Battery Selected  
- Portable Battery Temperatures  
- Fixed Battery Temperatures

### 3. Target Column:

The target variable for this analysis is the Effective SOC, which reflects the current charge level of the battery in a percentage format.

## B. Machine Learning Model Development

### 1. Models Developed:

Multiple regression models were employed to predict Effective SOC, including:  
- Linear Regression: This model achieved an impressive R² value of 0.9998 and a mean squared error (MSE) of 0.0009, indicating an excellent fit to the data.  
- Decision Tree: With an R² value of 0.8816 and MSE of 0.4660, this model provided a reasonable understanding of the relationships but showed less accuracy than linear regression.  
- Random Forest: This ensemble method yielded an R² of 0.9719 and an MSE of 0.1106, demonstrating strong predictive power by averaging multiple decision trees.  
- Gradient Boosting: Achieving an R² of 0.9886 and MSE of 0.0447, this model showcased its ability to capture complex patterns within the data.  
- Support Vector Regressor: This model provided an R² of 0.9856 and an MSE of 0.0566, indicating effective performance with robust generalization capabilities.  
- K-Nearest Neighbors: Although this model resulted in a lower R² of 0.7669 and an MSE of 0.9174, it still offered insights into local relationships in the data.

### 2. Justification of Models:

The models were selected based on their performance metrics and suitability for the regression task. The linear regression model exhibited the highest R² value, indicating a strong linear relationship between the features and Effective SOC. Other models were included to compare and contrast their predictive capabilities, ensuring a comprehensive evaluation of different approaches.

## C. Key Performance Indicators (KPIs)

### 1. Definitions and Calculations:

Charge Cycle: This KPI will be calculated based on the number of times the battery is charged and discharged, providing insights into its longevity and usage patterns.  
Range: The effective range will be estimated by analyzing the voltage and current metrics, which are critical in understanding how far the battery can perform before needing a recharge.  
Battery Performance: This KPI will evaluate the efficiency of the battery by comparing the Effective SOC against other relevant factors, such as voltage and current levels.  
Additional KPIs: Other important metrics may include temperature effects on battery performance, degradation rates, and any anomalies that could indicate potential issues with battery health.

## D. Documentation and Presentation

### 1. Report Structure:

Introduction: The introduction section will clearly articulate the problem statement, the significance of the analysis, and the objectives we aim to achieve.  
Methodology: A detailed description of the analytical approach taken, including the preprocessing steps, feature selection, model training, and evaluation processes, will be outlined.  
Results: This section will present the performance metrics for each model alongside the calculated KPIs, showcasing how well each model performed in predicting Effective SOC.  
Conclusion: A summary of key findings, insights gained from the analysis, and recommendations for future work or improvements to the model will conclude the report.

## E. Deliverables

### 1. Outputs Required:

A fully functional machine learning model capable of predicting Effective SOC based on the input features.  
A comprehensive set of calculated KPIs, complete with explanations of their significance in assessing battery performance and reliability.  
A well-organized and clearly written report summarizing the methodology employed, results obtained, and insightful recommendations based on the analysis.