<table>
<tr><td colspan="3" rowspan="2">Data Science Canvas</td><td>Project:</td><td colspan="4">Driver Behavior Classification & Fuel Consumption Prediction</td></tr>
<tr><td>Team:</td><td colspan="4">Riddhi Ranjan DE
Sarveshwaran N</td></tr>
<tr><td colspan="4">Problem Statement</td><td colspan="2">Execution & Evaluation</td><td colspan="2">Data Collection & Preparation</td></tr>
<tr><td>Business Case & Value Added
Analyze telematics data to detect unsafe driving and predict fuel usage. This improves fleet safety and reduces operational fuel costs.</td><td rowspan="2">Model Selection
Use clustering and gradient-boosted models to detect aggressive driving and predict fuel use. These methods handle unlabeled, non-linear telematics data effectively.

Aggressive Driving: K-Means (k=4) for pseudo-labels XGBoost Classifier for deployment
Fuel Prediction: XGBoost & HistGradientBoosting Regressors.
These handle non-linear behavior, large datasets, and unlabeled data.</td><td>Model Requirements
Models must avoid driver leakage, handle non-linear patterns, and use scaled numeric inputs. They must generalize to unseen drivers for valid performance.</td><td rowspan="2">Skills
Python (Pandas, NumPy, Scikit-learn).
ML: clustering, boosting, evaluation metrics.
Feature engineering & EDA.
Domain understanding of telematics (speed, rpm, braking).
Visualization (Matplotlib/Seaborn).</td><td>Model Evaluation
Evaluate accuracy, F1, and ROC-AUC for aggressive driving, and MAE for fuel prediction. Real-time monitoring is required for safety alerts.</td><td rowspan="2">Data Storytelling
The target audience needs clear visuals and actionable insights. Use simple cluster explanations and plots to highlight unsafe driving and inefficient trips. Visualize PCA, scatter plots, percentile thresholds. Provide practical insights: unsafe drivers, fuel-inefficient trips.</td><td>Data Selection & Cleansing
Clustering/classification used 6 behavior features; fuel prediction used 10 engineered features.
Timestamp removed, no missing data, and outliers kept as meaningful extremes.</td><td>Data Collection
Dataset: Driver Behavior & Route Anomaly (DBRA24) from Kaggle.
120,000 trips × 26 features collected from telematics sensors:
speed, acceleration, steering, rpm, route deviation, weather, road type.
Additional data can be gathered from vehicle sensors, OBD-II, and terrain APIs. It must be accurate, time-aligned, and consistently recorded per trip.</td></tr>
<tr><td>Data Landscape
You need trip-level telemetry, environmental context, and fuel data, which are already available. Additional vehicle specifications and terrain data would improve accuracy.</td><td>Software & Libraries
Python + Jupyter Notebook
Pandas, NumPy, Matplotlib, Seaborn
Scikit-Learn, XGBoost, HistGradientBoosting
(Optional future work: PyTorch for AE/Deep models)</td><td>Data Integration
All data is already in a single CSV, so migration isn’t needed. It should simply be loaded into a unified Python/pandas environment for processing.</td><td>Explorative Data Analysis
EDA showed meaningful outliers reflecting real extreme behaviors and clear driver-specific patterns. Descriptive statistics and distributions were created to assess data quality.
Confirmed 1 trip = 1 row (no time series).
Driver-level profiling identified behavioral variability.
Correlation heatmaps showed weak correlations → need multi-feature modeling.
Identified skewed features (rpm, acceleration variation).
Percentile thresholds computed (p50–p99).</td></tr>
</table>