Data Collection and Preprocessing Phase

Date	15 March 2024
Team ID	SWTID1720078183
Project Title	Predictive Modeling for Fleet Fuel Management using Machine Learning
Maximum Marks	6 Marks

Data Exploration and Preprocessing

Identifies data sources, assesses quality issues like missing values and duplicates, and implements resolution plans to ensure accurate and reliable analysis.

Section	Description
Data Overview	**Basic Statistics**: • 'speed': Mean = 41.93, Std = 13.60, Min = 14, Max = 90 • 'temp_outside': Mean = 11.36, Std = 6.99, Min = -5, Max = 31 • Other columns like AC, rain, sun are binary indicators with statistics indicating their frequencies. **Dimensions**: 388 rows and 12 columns. **Structure**: The dataset includes both numerical and categorical data types, with some columns having missing values: • 5 columns are int64 (e.g., speed, temp_outside). • 7 columns are object (e.g., distance, consume).
Univariate Analysis	The preprocessing notebook performs basic operations like filling missing values for 'temp_inside' and 'temp_outside' with their mean values, but doesn't explicitly provide detailed univariate analysis like histograms or descriptive statistics for each variable.
Bivariate Analysis	The preprocessing notebook doesn't include specific cells dedicated to bivariate analysis (e.g., correlation matrices or scatter plots). The focus appears to be on preprocessing and model preparation rather than exploratory data analysis.

Multivariate Analysis	The preprocessing notebook includes steps to create polynomial features and fit various regression models, indicating multivariate analysis. For instance: • Creating polynomial features: PolynomialFeatures(degree=2, include_bias=False) • Applying different regression models such as 'HistGradientBoostingRegressor' and 'RandomForestRegressor'.
Outliers and Anomalies	The preprocessing notebook does not explicitly address outliers and anomalies. It focuses on feature engineering and model fitting without detailing outlier detection or treatment methods.

Data Preprocessing Code Screenshots

Loading Data	`import pandas as pd` `data = pd.read_csv('path_to_dataset.csv')`
Handling Missing Data	`data.fillna(data.mean(), inplace=True)`

Data Transformation	`from sklearn.preprocessing import StandardScaler`
	`scaler = StandardScaler()`
	`scaled_data = scaler.fit_transform(data)`
	
	`# Normalize the data`
	`scaler = StandardScaler()`
	`X_scaled = scaler.fit_transform(X)`

Feature Engineering	`# Features and target variable`
	`X = data.drop('consume', axis=1)`
	`y = data['consume']`

Save Processed Data	`data.to_csv('fuelConsumption.csv', index=False)`