Title: Monitoring Vehicle Health through Engine Control Unit (ECU) Data

Team Members: Raghav Agarwal <raghav.agarwal@gwmail.gwu.edu>; Sanjayram Raja Srinivasan <sanjayram.rajasrinivasan@gwmail.gwu.edu>;Shikha Kumari <shikha.kumari@gwu.edu>; Harshavardana Reddy Kolan <harshavardanareddy.kolan@gwu.edu>

Research Topic:

Our project involves the utilization of a comprehensive dataset encompassing a diverse range of engine types. To facilitate data analysis and enhance fuel efficiency, we initiate our methodology by segmenting the dataset into distinct clusters or dataframes via clustering techniques. Subsequently, our analysis focuses on identifying the cluster or dataframe exhibiting the maximum fuel mass consumption, thereby serving as our primary area of optimization.

Linear regression (LM) models are constructed for each cluster or dataframe, excluding the one with the highest fuel mass consumption, allowing for in-depth examination of the relationships between engine parameters and fuel consumption within each specific cluster. We then proceed to offer practical recommendations for improving fuel efficiency by suggesting both trimmable and non-trimmable parameters tailored to the cluster with the least favorable fuel mass consumption. Our modeling approach encompasses an exploration of interaction terms to augment model accuracy.

Furthermore, we conduct a comprehensive feature contribution analysis to pinpoint the parameters that significantly impact fuel mass consumption within each cluster, relative to their respective LM models. The ultimate goal of this undertaking is to devise an engine recommendation system, which leverages values derived from distinct LM models, enabling us to provide actionable guidance to drivers seeking to reduce fuel consumption and enhance fuel efficiency while on the road.

SMART Research Questions:

1. Can we develop data-driven recommendations to improve fuel efficiency in various engine clusters and provide drivers with actionable guidance to reduce fuel consumption on the road?
2. By conducting data clustering, creating linear regression models, and exploring interaction terms, can we realistically identify features contributing to fuel mass consumption and suggest practical parameters for optimization?
3. Is this project relevant to addressing the environmental and economic concerns related to fuel efficiency, and can it provide valuable insights for the automotive industry?
4. Is it possible to create separate linear regression models for each cluster/dataframe, and what challenges might arise in the process?

Very interesting. A classic linear regression set up, but with good motivation and context. Please also try a regression tree.

Source of the dataset: https://publikationen.bibliothek.kit.edu/1000085073

Number of observations in the dataset: (2693824 rows × 14 columns)

Github repository link: **Rag-hav385/Intro_to_data_science_project: Project. (github.com)**