**Final Project – Fundamentals of Machine Learning  
Group 7 – Analysing Fuel Sources**

**Group Members – Archana Gadicharla**

**Nikhila Reddy Vantari**

|  |  |
| --- | --- |
| **Archana Gadicharla** | Feature selection using correlation matrix and principal component analysis. Model building through K-Means clustering. Comparison of clustering models in R Studio, followed by a PowerPoint presentation and project report. |
| **Nikhila Reddy Vantari** | Exploratory Data Analysis: Analyzing the distribution of initial variables before and after feature selection. Preparing and exploring the initial data. Model Development: Building the DBSCAN model in R Studio. Creating a comprehensive project report and a PowerPoint presentation. |

**Contribution:**

# **FINAL PROJECT – ANALYSING FUEL SOURCES**

## **Executive Summary**

## This analysis evaluates a fuel dataset to uncover key insights using feature engineering and data preprocessing for predictive modeling. The process involves loading the data, addressing missing values, converting categorical data to numeric, and removing irrelevant columns. The findings underscore the importance of feature selection, data cleaning, and clustering to understand complex datasets. Missing values are handled via imputation or removal, categorical variables are converted for model compatibility, and irrelevant columns are discarded. The analysis employs basic stats, feature selection through correlation matrices, PCA, and clustering (K-means and DBSCAN), with K-means proving more effective for this dataset.

**Dataset Overview**

* No of rows: 38,113  
  Number of Columns: 81
* The dataset contains 81 columns representing various attributes related to vehicles and fuel efficiency. The first five rows of the dataset indicate that the data covers vehicles from different manufacturers, classes, and models.

**Notable Observations**

**Vehicle Information:** The columns year, make, and model describe each vehicle uniquely.

**Transmission Details**: transmission and transmission\_type provide information on the vehicle's transmission type.

**Fuel Economy:** Columns like composite\_city\_mpg, composite\_highway\_mpg, and composite\_combined\_mpg indicate fuel efficiency.

**Range and Consumption:** There are multiple columns covering different types of fuel consumption and vehicle range, such as range\_ft1, city\_range\_ft1, and highway\_range\_ft1.

**Goal**

The goal of this project is to analyze fuel consumption data using clustering techniques to identify key trends and extract valuable insights from the data.  
**Data Exploration:** Understand the dataset through summary statistics and visualization  
 Data Cleaning: Handle missing values and transform categorical variables into numerical features  
 **Feature Selection:** Perform Principal Component Analysis (PCA) to reduce dimensionality  
 **Model Building:** Apply machine learning algorithms to build predictive models and evaluate their performance.

**Problem Assertion**

The task involves analyzing a complex fuel dataset to uncover valuable insights and predict outcomes using machine learning techniques. This dataset contains a mix of numerical and categorical features, which require thorough preprocessing to address missing values and irrelevant features. The primary challenge is to accurately analyze fuel efficiency, vehicle characteristics, and their interrelationships while effectively managing the inherent complexities of the data.

**Data Preparation**

1. Data Cleaning

The dataset was cleaned to remove irrelevant columns and address missing values. Key steps included:

* -Removing unnecessary columns.
* -Handling missing values.
* -Converting categorical variables to numerical format.

2. Data Analysis

The cleaned data was analyzed using various visualization techniques such as histograms, box plots, and scatter plots to understand the distribution of key variables.

3. Feature Selection

Feature selection was performed using correlation matrices and variance thresholds to identify and retain the most informative features. Principal Component Analysis (PCA) was also applied to reduce dimensionality.

PCA Analysis  
A graph with text and numbers

Description automatically generated

A graph of the same graph

Description automatically generated with medium confidence

**Correlation matrice**

**A close-up of a graph

Description automatically generated**

**Findings & Observations**

Through the clustering analysis, distinct patterns were identified. For instance, the k-means clustering technique effectively grouped vehicles based on fuel consumption characteristics. The findings emphasize the importance of appropriate clustering techniques in analyzing complex datasets.

**Data Analysis**

1. K-means Clustering

K-means clustering was applied to the principal components of the dataset to identify distinct clusters. The silhouette method was used to determine the optimal number of clusters.

**A graph showing different colored shapes

Description automatically generated with medium confidence**

2. DBSCAN Clustering

DBSCAN clustering was employed to identify non-linear clusters within the data. However, the results revealed that the clusters formed were not relevant or meaningful.

**A graph with numbers and symbols

Description automatically generated**

**Conclusion:**The analysis of the fuel dataset underscores the importance of comprehensive data preprocessing, including feature engineering, data cleaning, and clustering techniques, to reveal meaningful insights. Through Principal Component Analysis (PCA) and K-means clustering, the study identifies distinct patterns in fuel consumption, emphasizing the value of appropriate data processing in handling complex datasets. While K-means clustering provided actionable insights into fuel efficiency, DBSCAN clustering could have been more effective due to the non-linear nature of the data. The study demonstrates that practical data analysis can significantly enhance understanding of fuel consumption and vehicle characteristics, aiding decision-making in the energy sector.

**Data Source:** <https://www.kaggle.com/datasets/tanishqdublish/vehcile-fuel-consumption>