1. Problem Description:

The goal of the project is to analyze a dataset of vehicles and understand the effect of factors affecting the fuel efficiency, in other words, to understand how each factor like weight, engine size etc. affect the fuel efficiency.

The primary objective was to build a machine learning model that accurately predicts the fuel efficiency based on its known attributes.

2. Methodology:

Exploratory Data Analysis (EDA):

- **Bivariate Analysis**: We used scatter plots to examine the relationship between fuel efficiency and key numerical features.
- **Categorical Analysis**: We used box plots to see how categorical features like engine_config, release_year, and manufacture_region impacted fuel_efficiency.

Data Preprocessing:

- **Outlier Detection:** Used quartile method and box plots to determine outliers and eliminated them.

- **One-Hot Encoding:** Converted categorical features (engine_config, manufacture_region) into numerical features.

Model Building and Evaluation:

- **Model Selection:**

Evaluated using three algorithms:

1. Linear Regression (as a baseline)
2. Random Forest Regressor
3. Gradient Boosting Regressor

- **Evaluation Metrics**:

  Models were compared based on R-squared (R^2), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE) to determine the best performer.

3. Key Results and Insights:

The Exploratory Data Analysis (EDA) revealed several key insights:

- **Strong Negative Correlations**: fuel_efficiency has a strong negative relationship with vehicle_mass and engine_volume. As vehicles get heavier or have larger engines, their fuel efficiency drops significantly.
- **Engine Configuration**: engine_config (number of cylinders) is a major predictor. 4-cylinder engines were, on average, the most efficient, with efficiency dropping steadily as the cylinder count increased to 6 and 8.
- **Time Trend**: There is a clear positive trend in efficiency by release_year. Cars manufactured later in the dataset (approaching the early 1980s) are noticeably more fuel-efficient, likely reflecting responses to oil crises and new regulations.
- **Region of Origin**: manufacture_region is a distinct factor. Region 3 (likely Japan) produced vehicles with the highest median efficiency, while Region 1 (likely USA) had the lowest.

4. Results:

Model	**R-squared (R^2)**	**MAE**	**RMSE**
	Higher is Better	**Lower is Better**	**Lower is Better**
RandomForestRegressor	**0.904**	**1.86**	**2.38**
LinearRegression	0.888	1.95	2.57
GradientBoostingRegressor	0.886	1.92	2.60

Based on these results, the **RandomForestRegressor is the clear winner** and the best-performing model.