A Design-Based Statistical Investigation into the Determinants of Vehicle Fuel Consumption Efficiency

Group members: Jiahui Yan, Feng Gu, Ella, Yuzhuo Ye, Taiwo

1 Introduction and Research Question

Fuel efficiency remains one of the most important performance indicators in the automotive industry, influencing both environmental impact and consumer cost. Previous studies [1, 2] have shown that some mechanical factors, like engine size and horsepower can affect fuel consumption.

With the right combination of these factors, improvement on fuel efficiency has been proven achieved, and it promises the potential for developing more fuel-efficient vehicles by exploring optimal configurations on these factors and their interactions. Out of this background, this project aims to examine how a series of car factors affect the target variable—fuel efficiency.

2 Research Objectives

This project aims to systematically investigate the factors influencing automotive fuel efficiency through experimental design methodology. The primary objective is to examine how key automotive characteristics individually and collectively affect miles per gallon (MPG) performance. Specifically, we seek to:

1. Determine the main effects of each factor on fuel efficiency
2. Identify potential interaction effects among these factors to understand how combinations of characteristics influence fuel performance
3. Apply appropriate statistical analysis including assumption testing and effect size estimation

3 Data Description

The data are from the Auto MPG Dataset from UCI Machine Learning Repository [5].

The Auto MPG Dataset from UCI Machine Learning Repository contains 398 automobiles manufactured between 1970-1982. The dataset includes nine variables: the dependent variable MPG and eight predictors including cylinders, displacement, horsepower, weight, acceleration, model year, origin, and car name, [1] the variable meanings

[1] Here the car name is a long string with brand and model information.

are provided in Table 1 and their distributions are shown in Figure 1.

Table 1: Summary of Variables in the Auto MPG Dataset

Variable	**Description**	**Units / Type**
mpg*(target)	Miles per gallon (fuel efficiency)	miles/gallon
cylinders	Number of engine cylinders	$\mathbb{N}$
displacement	Engine displacement	in^3(cubic inches)
horsepower	Engine power output	hp
weight	Vehicle weight	lb
acceleration	Time to accelerate 0–60 mph	s
model year	Year of manufacture (two-digit)	$\mathbb{N}$
origin	Country/region of origin	$\{1, 2, 3\}$
car name	Car make and model	string

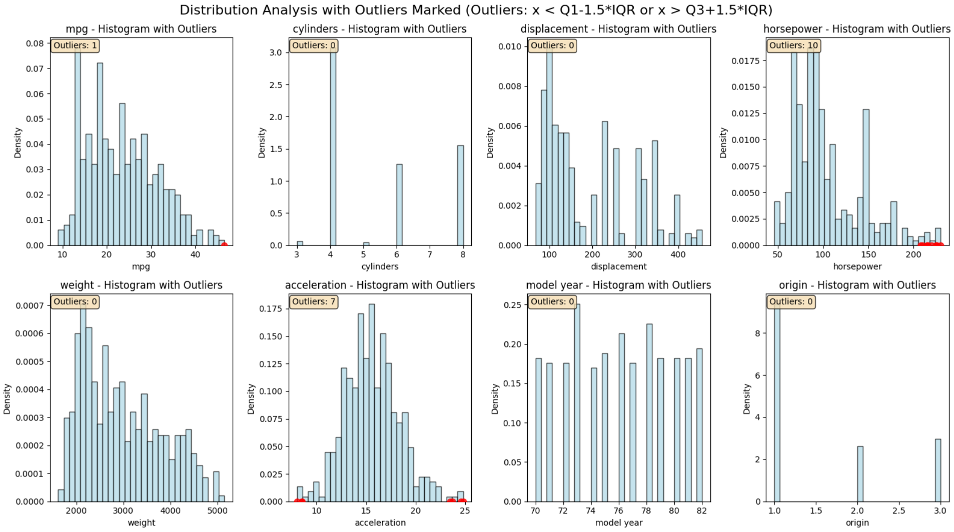

Figure 1: Distributions of Numerical Variables in the Auto MPG Dataset

4 Brief Literature Review

Studies in automotive engineering and energy policy emphasize that vehicle weight and engine design strongly influence fuel efficiency.

- [1] found that a 10% reduction in weight improves fuel economy by up to 8%.
- [2] reported that higher cylinder counts and engine power typically reduce mpg due to increased energy requirements.
- [3] observed that newer model years integrate improved engine management and aerodynamics, enhancing fuel efficiency despite higher horsepower.

While these studies analyzed individual factors, few examined their interaction effects through factorial experiments. This project addresses that gap by using a 2^4 factorial design to quantify both individual and combined influences of mechanical and temporal factors on mpg.

5 Hypotheses and Statistical Models

The following hypotheses will be tested to examine the relationships between automotive characteristics and fuel efficiency at $\alpha = 0.05$ significance level:

5.1 Main Effect Hypotheses

Table 2: Main Effect Hypotheses for Automotive Fuel Efficiency Analysis

Variable	**Null Hypothesis (H_0)**	**Alternative Hypothesis (H_1)**
Cylinders	Number of cylinders has no significant effect on MPG	Number of cylinders has a significant effect on MPG
Displacement	Engine displacement has no significant effect on MPG	Engine displacement has a significant effect on MPG
Horsepower	Engine horsepower has no significant effect on MPG	Engine horsepower has a significant effect on MPG
Weight	Vehicle weight has no significant effect on MPG	Vehicle weight has a significant effect on MPG
Acceleration	Acceleration time has no significant effect on MPG	Acceleration time has a significant effect on MPG
Model Year	Model year has no significant effect on MPG	Model year has a significant effect on MPG
Origin	Country/region of origin has no significant effect on MPG	Country/region of origin has a significant effect on MPG

5.2 Statistical Model Formulations and Hypotheses for Key Relationships

ANOVA Model for Categorical Factors:

$$\text{MPG}_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk} \tag{1}$$

where i represents cylinder groups, j represents origin groups, and $(\alpha\beta)_{ij}$ represents interaction effects. The error term is tentatively assumed to follow $\epsilon_{ijk} \sim N(0, \sigma^2)$ for observations within each group.

Multiple Linear Regression Model:

$$\begin{aligned}\text{MPG} = \beta_0 &+ \beta_1(\text{Cylinders}) + \beta_2(\text{Displacement}) + \beta_3(\text{Horsepower}) \\ &+ \beta_4(\text{Weight}) + \beta_5(\text{Acceleration}) + \beta_6(\text{Model Year}) \\ &+ \beta_7(\text{Origin}_2) + \beta_8(\text{Origin}_3) + \epsilon \end{aligned} \tag{2}$$

where $\epsilon \sim N(0, \sigma^2)$ and Origin is dummy-coded with USA as reference category.

Table 3: Specific Directional Hypotheses

Relationship	**Null Hypothesis**(H_0)	**Alternative Hypothesis**(H_1)
Cylinders-MPG	$\beta_{\text{cylinders}} = 0$	$\beta_{\text{cylinders}} < 0$ (More cylinders decrease MPG)
Weight-MPG	$\beta_{\text{weight}} = 0$	$\beta_{\text{weight}} < 0$ (Higher weight decreases MPG)
Horsepower-MPG	$\beta_{\text{horsepower}} = 0$	$\beta_{\text{horsepower}} < 0$ (Higher horsepower decreases MPG)
Model Year-MPG	$\beta_{\text{year}} = 0$	$\beta_{\text{year}} > 0$ (Newer cars have higher MPG)
Origin Effect	$\mu_{\text{USA}} = \mu_{\text{Europe}} = \mu_{\text{Japan}}$	At least one origin mean differs significantly

6 Overview of Experimental Design

6.1 Design Type

A 2^4 factorial design (no replication) will be used. This design evaluates four independent factors, each at two levels:

Factor	**Levels**	**Type**
Cylinders	4 vs. 8	Discrete
Horsepower	Low vs. High (median split)	Continuous (categorized)
Weight	Low vs. High (median split)	Continuous (categorized)
Model Year	Old vs. New (pre-1980 vs. post-1980)	Categorical

Dependent Variable: *Miles per Gallon (mpg)* — continuous measure of fuel efficiency.

6.2 Experimental Units

Each car in the dataset represents one experimental unit. Cars are assigned into treatment combinations according to the four factors. With a 2^4 factorial design, there are 16 treatment combinations (e.g., 4-cylinder, low horsepower, low weight, old model year).

6.3 Randomization and Control

- **Randomization:** Cars are randomly selected from the dataset for each treatment combination to minimize selection bias.
- **Control Variables:** Transmission type and engine displacement will be monitored to reduce confounding effects.

7 Statistical Analysis Plan

7.1 Data Preparation

Data from the car dataset will be preprocessed in R:

- Handle missing values using mean or median imputation.
- Categorize continuous predictors (horsepower and weight) into "low" and "high" based on median splits.
- Code categorical variables (e.g., cylinders = {4, 8}; model_year = {old, new}).

7.2 Analysis Method

A four-factor ANOVA (2^4 design) will be conducted to test main and interaction effects.

If assumptions are not met:

- Apply transformation (e.g., log(mpg)).
- Use multiple regression as an alternative model.

7.3 Statistical Tools

- Normality test: Shapiro–Wilk test
- Variance homogeneity: Levene's test
- Effect size: Partial η^2 using `effectsize` package
- Visualization: Interaction plots and residual diagnostics (ggplot2, interactions packages)

Main Effects: Cars with fewer cylinders, lower horsepower, lighter weight, and newer models are expected to have higher mpg.

Interaction Effects:

- The negative effect of weight may be stronger for older cars.
- The effect of horsepower may depend on the number of cylinders.

These outcomes could provide insights into energy-efficient automotive design and environmental sustainability initiatives.

8 Preliminary Analysis: One-Way ANOVA on Car Origin and Fuel Efficiency

As a preliminary investigation, we conducted a one-way ANOVA to examine whether car origin (USA, Europe, Japan) significantly affects fuel efficiency, and this analysis serves as a foundational step to more complex multi-factor designs.

Experimental Design: The one-way ANOVA employed a completely randomized design with car origin as the single factor at three levels: Origin 1 (USA, n=249), Origin 2 (Europe, n=70), and Origin 3 (Japan, n=79). Each car represents an independent experimental unit, with MPG as the continuous response variable.

Hypotheses Tested:

- H_0: $\mu_{\text{USA}} = \mu_{\text{Europe}} = \mu_{\text{Japan}}$ (all origin groups have equal mean MPG)
- H_1: At least one origin group has a significantly different mean MPG

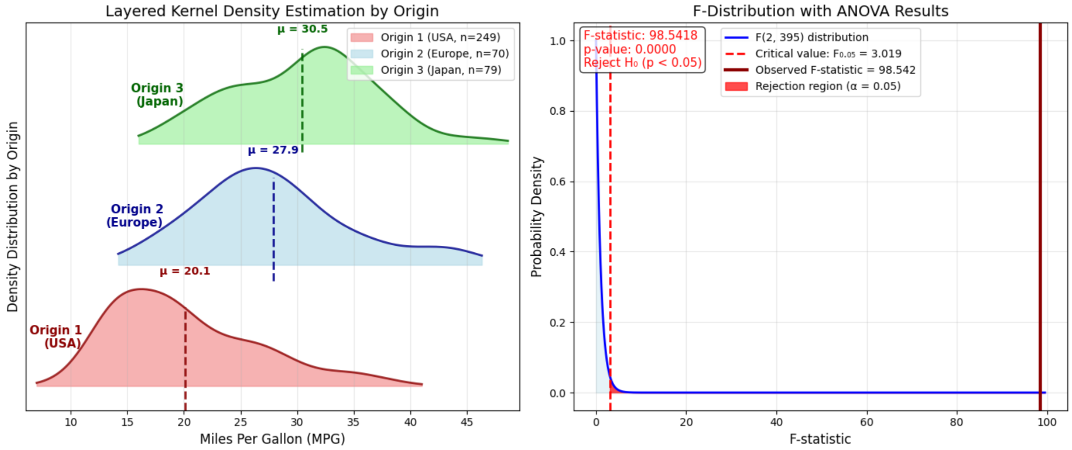

Figure 2: Preliminary ANOVA Results: Effect of Car Origin on Fuel Efficiency (MPG)

The ANOVA revealed a highly significant effect of car origin on fuel efficiency ($F_{(2,395)} = 98.54$, $p < 0.001$, $\eta^2 = 0.333$). Japanese cars showed highest efficiency (M = 30.45), followed by European (M = 27.89) and American cars (M = 20.08), leading to rejection of the null hypothesis.

This preliminary analysis confirms significant regional differences in automotive fuel efficiency, supporting the inclusion of origin as a key factor in subsequent multi-factorial designs.

Table 4: One-Way ANOVA Results: Effect of Car Origin on Fuel Efficiency (MPG)

ANOVA Summary	**Value**
Null Hypothesis (H_0)	All group mpg means are equal
Alternative Hypothesis (H_1)	At least one group mpg mean differs
Significance level (α)	0.050
Total sample size (N)	398
Number of groups (k)	3
Degrees of freedom (Between groups)	2
Degrees of freedom (Within groups)	395
Critical F-value	3.0186
Observed F-statistic	98.5418
p-value	< 0.001
Statistical Decision	**Reject** H_0
Effect Size (η^2)	0.3329

9 References

References

[1] Ahmad, N., et al. (2020). *Weight reduction and fuel efficiency in automotive design: An integrated assessment.* Transportation Research Part D: Transport and Environment, 86, 102446.

[2] Greene, D. L., & Welch, T. (2017). *Impact of engine size and vehicle weight on fuel economy: Policy implications.* Energy Policy, 108, 273–282.

[3] Li, F., & Zhao, J. (2022). *Advances in automotive efficiency: The role of model year innovation.* Applied Energy, 325, 119823.

[4] U.S. Department of Energy. (2021). *Fuel Economy Trends Report.* Retrieved from `https://www.energy.gov`

[5] U.S. Department of Energy. (2021). *Auto MPG Dataset.* UCI Machine Learning Repository. Retrieved from `https://www.kaggle.com/datasets/uciml/autompg-dataset`

10 Appendix

10.1 Figures

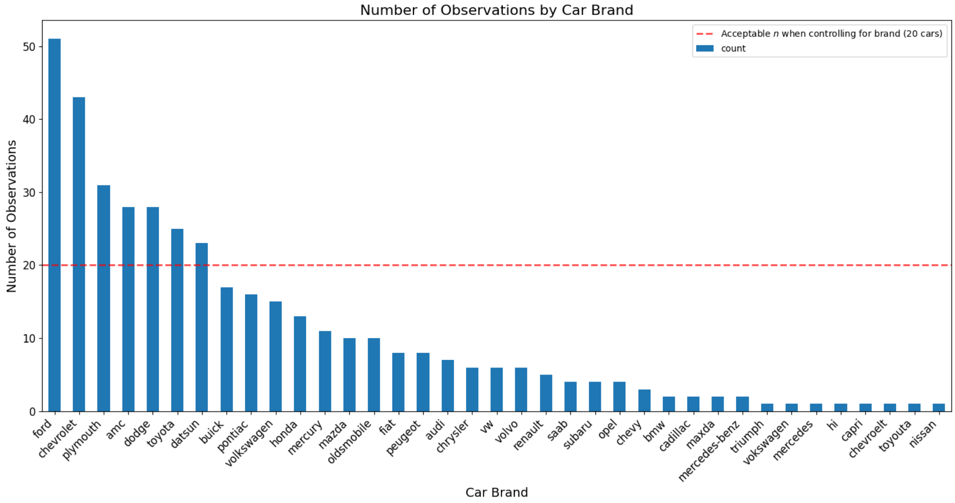

Figure 3: Distribution of samples by Car Brand

Table 5: Descriptive Statistics of Fuel Efficiency (MPG) by Car Origin

Origin	**Sample Size (n)**	**Mean MPG**	**Standard Deviation**
Origin 1 (USA)	249	20.08	6.40
Origin 2 (Europe)	70	27.89	6.72
Origin 3 (Japan)	79	30.45	6.09
Total	**398**	**23.51**	**7.82**

10.2 Tables

10.3 Ethical Considerations

Although the dataset is non-human and publicly available, ethical integrity is maintained by:

- **Data Source:** `https://www.kaggle.com/datasets/uciml/autompg-dataset`
- **Privacy and Compliance:** Ensuring no personally identifiable information is used.
- **Reproducibility:** Providing Python code and random seeds for full replication.
- **Fair Representation:** Avoiding selective reporting or p-hacking.
- **Academic Integrity:** Properly citing all sources and acknowledging dataset creators.