Correlation Reversal Manipulation Revealed by Benford's Law and Random Forest

Poj Netsiri
Immatriculation number: 12402153
e12402153@student.tuwien.ac.at

Main Supervisors
Assoz.Prof. PD Dr. Jakob Müllner, Wirtschaftsuniversität Wien
Asst.Prof. Dr. Harald Puhr, Universität Innsbruck
Co-Supervisor
Assoz.Prof. Dr. Nysret Musliu, Technische Universität Wien

Domain-Specific Lectures: 194.068 Foundations of International Business 23 May 2025

Table of Contents

Abstract

Unethical data manipulation involves the deliberate alteration of data to support a predetermined hypothesis, often leading to misleading conclusions. In this study, a dataset of 32 datapoints initially reported a statistically significant but counterintuitive positive correlation ($r = +0.5$, $p < 0.001$), suggesting that cars with higher horsepower achieve better fuel efficiency (MPG). Despite the apparent statistical validity, Benford's Law revealed a substantial deviation from the expected digit frequency distribution ($\chi^2 > 30.58$, $p < 0.0001$), strongly indicating potential manipulation.

To investigate further, both Benford's Law and a Random Forest algorithm were employed to classify the data as manipulated or unmanipulated. Each method independently flagged anomalies and contributed to reconstructing an approximation of the original dataset. The reconstructed data revealed a reversal in correlation direction ($r = -0.57$, $p < 0.001$), not only aligning with established automotive principles—which associate higher horsepower with lower fuel efficiency—but also consistent with the direction of the known original data prior to manipulation, even if not matching its exact values.

While each method offers distinct advantages—Benford's Law for unsupervised anomaly detection and Random Forest for regression when trained on clean data—this study demonstrates that a hybrid approach is especially effective. With just 32 samples, a size common in scientific and engineering research, the combined methods successfully detected manipulation and restored the true correlation direction. These findings highlight the value of integrating statistical and machine learning techniques, even in small-sample scenarios, to uphold data integrity and scientific validity.

1 Introduction

Introduction

Unethical data practices pose a serious threat to the integrity of scientific research, allowing manipulated outcomes to be presented as legitimate findings, often misleading stakeholders. Such misconduct may arise from the pursuit of academic recognition, competition for funding, or institutional pressure. Common forms of manipulation include data falsification—where data are selectively altered or fabricated to support a hypothesis—and p-hacking, which involves testing multiple hypotheses until statistically significant results emerge.

These practices distort the scientific record, misinform policy decisions, and undermine public confidence in research. Because standard peer review processes often fail to detect subtle manipulations, the application of data-driven detection methods is essential for preserving research credibility.

This study investigates a given dataset consisting of only 32 datapoints—a sample size that, despite its modest scale, is typical in many scientific and engineering studies due to practical constraints or focused experimental designs. The dataset reports a statistically significant positive correlation between vehicle horsepower and fuel efficiency (miles per gallon). This counterintuitive claim contradicts well-established automotive principles, which associate higher horsepower with lower fuel efficiency.

To assess the integrity of the obtained data, Benford's Law [2] is applied to detect anomalies in the digit distribution, while a Random Forest algorithm [4] is used to classify manipulated versus unmanipulated values. Both methods are further employed to reconstruct an approximation of the original dataset. The reconstructed data reveal a reversal in the direction of correlation, not only aligning with established automotive principles—which associate higher horsepower with lower fuel efficiency—but also consistent with the direction of the known original data prior to manipulation, even if not matching its exact values.

This study demonstrates that Benford's Law and the Random Forest algorithm can be used not only to detect data manipulation but also to reveal correlation direction reversal, even within small datasets.

2 Datasets

The datasets include 11 features across 32 datapoints (Appendix), as detailed below.

2.1 Original Data

The original dataset, published in *the 1974 Motor Trend US magazine*, contains fuel consumption data along with 11 features for 32 automobiles from the 1973–74 model years. This dataset has become a widely used benchmark in the data science community for regression and exploratory analysis tasks. The full data is presented in Appendix A. Initial analysis reveals a strong negative Pearson correlation [1] ($r = -0.78$, $p < 0.001$) between horsepower and fuel efficiency.

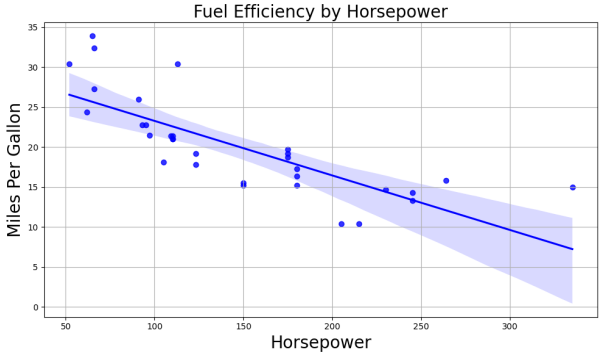

Figure 1: Negative correlation between horsepower and fuel efficiency. As horsepower increases, miles per gallon (MPG) tends to decrease, indicating that higher-performance vehicles typically consume more fuel.

2.2 Manipulated Data

The original dataset was manipulated using an unknown method by a fellow student (Tuvshin Selenge). The modified data is presented in Appendix B. Analysis shows a moderate positive Pearson correlation [1] ($r = +0.5, p < 0.001$).

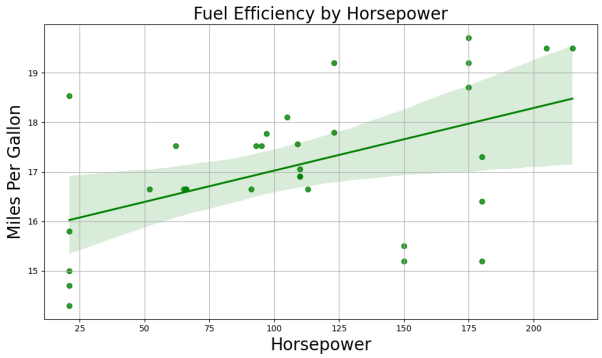

Figure 2: Positive correlation between horsepower and fuel efficiency. As horsepower increases, miles per gallon (MPG) tends to increase, indicating that higher-performance vehicles typically consume less fuel.

2.3 New Manipulated Data

The original data was manipulated using an unknown method by the author. This additional dataset is necessary to evaluate the performance of the pre-trained Random Forest classifier,

as the original (Section 2.1) and manipulated (Section 2.2) datasets were already used for supervised training. The data is presented in Appendix C. Analysis indicates a strong positive Pearson correlation [1] ($r = +0.89, p < 0.001$).

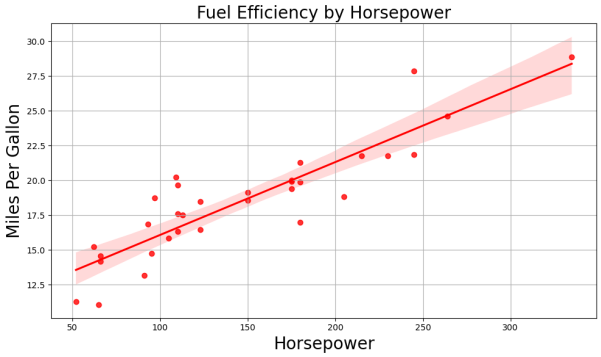

Figure 3: Positive correlation between horsepower and fuel efficiency. As horsepower increases, miles per gallon (MPG) tends to increase, indicating that higher-performance vehicles typically consume less fuel.

3 Methodology

3.1 Classification

3.1.1 Benford's Law

Benford's Law, also known as the *First-Digit Law*, describes the expected distribution of leading digits in many naturally occurring numerical datasets [2, 3]. According to this law, smaller digits appear as the first digit more frequently than larger ones.

Mathematical Expression

The probability $P(d)$ that a number has the first significant digit d, where $d \in \{1, 2, \ldots, 9\}$, is given by:

$$P(d) = \log_{10}\left(1 + \frac{1}{d}\right) \tag{1}$$

This results in the following expected distribution [2, 3, 7]:

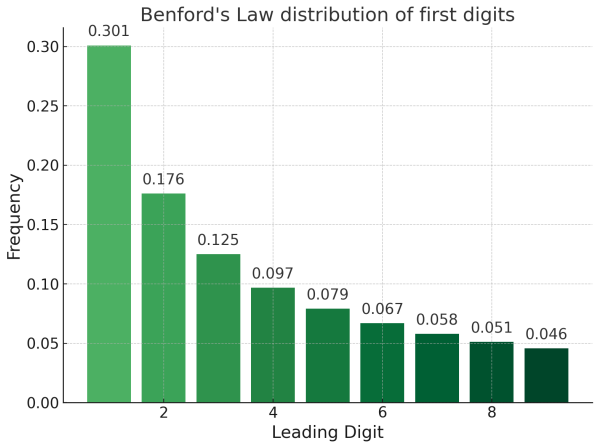

Figure 4: Benford's Law distribution of first digits. The distribution shows the expected frequency of each leading digit in naturally occurring datasets, with the digit 1 appearing about 30% of the time.

Benford's Law is a recognized tool in forensic data analysis for detecting anomalies and potential fraud. Deviations from the expected digit distribution can signal manipulation, often assessed using statistical methods like the χ^2 test. It is accepted in legal contexts, including use by the IRS and forensic auditors, and has met the Daubert standard for admissibility of scientific evidence in U.S. courts. Benford's Law is frequently applied alongside other techniques in fraud investigations [7, 8].

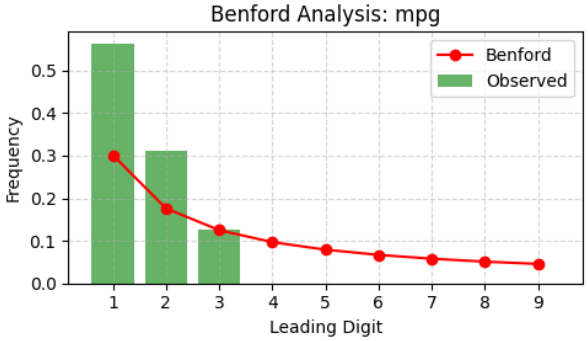

Figure 5: Benford's digit distribution of the unmanipulated original dataset (Appendix A) shows a χ^2 value of 20.0 with a p-value of 0.01. Given the threshold of $\chi^2 > 30.53$ and $p < 0.0001$ for detecting anomalies, this distribution is considered acceptable.

Figure 5 presents a Benford analysis of the unmanipulated original dataset, focusing on the leading digit distribution of miles per gallon (mpg) values. The observed frequencies (green bars) closely align with the expected Benford distribution (red line). The χ^2 test yields a value of 20.0 with a p-value of 0.01. Since this falls below the anomaly detection threshold of $\chi^2 > 30.53$ and $p < 0.0001$, the dataset is considered to conform acceptably to Benford's Law.

3.1.2 Random Forest Classifier

Random Forest is a machine learning algorithm that constructs a collection of decision trees during training and outputs the mode for classification [4, 5]. It is widely used in supervised learning tasks due to its high accuracy, robustness, and ability to handle various data types.

Mathematical Overview

Let $\{T_1(x), T_2(x), \ldots, T_K(x)\}$ be K decision trees trained on different bootstrapped samples of the training data. For a classification task, the Random Forest prediction $\hat{y}$ for input x is given by:

$$\hat{y} = \text{mode}\left(T_1(x), T_2(x), \ldots, T_K(x)\right) \tag{2}$$

Each tree is trained using a random subset of the features, which introduces diversity and reduces correlation among the trees.

Random Forest is highly effective for classification tasks, especially with high-dimensional data, mixed feature types, and nonlinear decision boundaries. It has been widely applied in areas such as fraud detection. While it can perform well on small datasets—particularly when informative features are present, hyperparameters are well-tuned, and cross-validation is used to avoid overfitting—simpler models like logistic regression may sometimes outperform it due to lower variance and greater interpretability [4, 5].

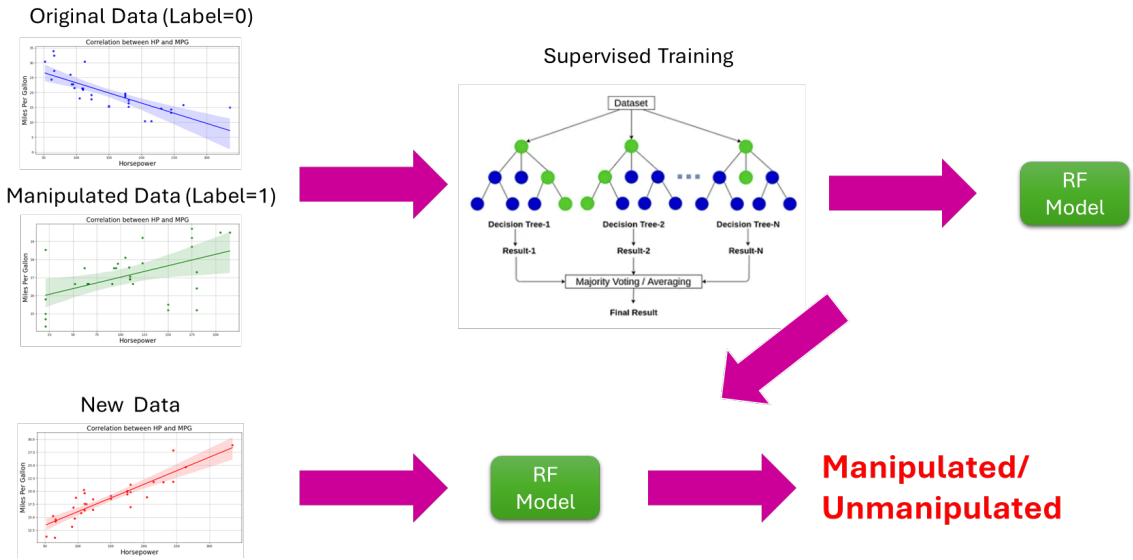

Figure 6: The original data (label = 0) and manipulated data (label = 1) are used to train a Random Forest model for classifying whether data points are manipulated. The pre-trained model is then applied to classify a new manipulated dataset.

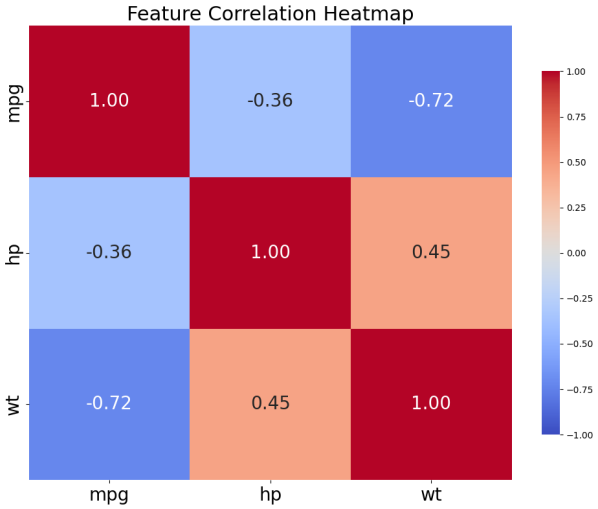

Figure 7: Heatmap showing the correlation between key features: miles per gallon (mpg), horsepower (hp), and weight (wt). features with correlation coefficients below 0.02 were excluded due to minimal contribution.

This study employs a Random Forest classifier to detect data manipulation, using the `scikit-learn` library [9]. As illustrated in Figure 6, the model was trained on a labeled data set in which the original (unmanipulated) records (Section 2.1) were labeled 0 and the manipulated records (Section 2.2) 1. The model was fine-tuned with the following hyperparameters: `n_estimators` = 40, `max_depth` = 5, `min_samples_split` = 20, `min_samples_leaf` = 3, `max_features` = ‘‘`sqrt`’’, `bootstrap` = `True`, and `random_state` = 42.

After training, the model was used to classify a newly manipulated dataset (Section 2.3) as either manipulated or unmanipulated. To improve model accuracy and minimize noise, feature selection was performed based on feature cross-correlation analysis (Figure 7). features showing meaningful correlations—specifically miles per gallon (mpg), horsepower (hp), and weight (wt)—were retained, while features with correlation coefficients below 0.02 were excluded from the analysis.

3.2 Prediction

3.2.1 Benford's Law

To better understand the intrinsic behavior of the original data, a stochastic simulation approach [6] is employed to reconstruct an approximation of the unmanipulated dataset from its manipulated version. Guided by the Benford profile derived in (Appendix A) Section 3.1.1, the method samples leading digits according to their expected logarithmic distribution, applies a uniformly distributed exponent to scale the values, and introduces normally distributed noise to simulate natural variability. This process generates a synthetic dataset that more closely reflects the structure of clean data. As shown in Figure 8, the reconstructed data is then analyzed to determine whether the observed correlations are consistent with expectations. A significant reversal in the direction of correlation between the manipulated and simulated data serves as a strong indicator of tampering.

From Equation (1):

$$
\begin{aligned}
d &\sim P(d) = \log_{10}\left(1 + \frac{1}{d}\right), \quad d \in \{1, \ldots, 9\} \\
e &\sim \mathcal{U}(1.3, 1.5) \\
\varepsilon &\sim \mathcal{N}(0, \sigma^2) \\
x &= d \cdot 10^e + \varepsilon
\end{aligned}
$$

- $e \sim \mathcal{U}(1.3, 1.5)$: The exponent e is sampled from a *uniform distribution* over the interval $[1.3, 1.5]$. This controls the *scale* of the synthesized value by varying the power of 10 applied to the leading digit.
- $\varepsilon \sim \mathcal{N}(0, \sigma^2)$: The noise term ε is drawn from a *normal distribution* with mean 0 and variance σ^2. This introduces slight randomness to simulate natural variability and avoid overly deterministic output.

Then the reconstructed data set is:

$$
X = \{x_i = d_i \cdot 10^{e_i} + \varepsilon_i \mid \texttt{min\_val} \leq x_i \leq \texttt{max\_val},\ i = 1, \ldots, n\} \tag{3}
$$

Where:

- $n = \texttt{n\_samples}$ `(32)`
- $\sigma = \texttt{noise\_std}$ `(0.1)`
- `min_val(15)`, `max_val(40)` are range constraints for valid MPG values

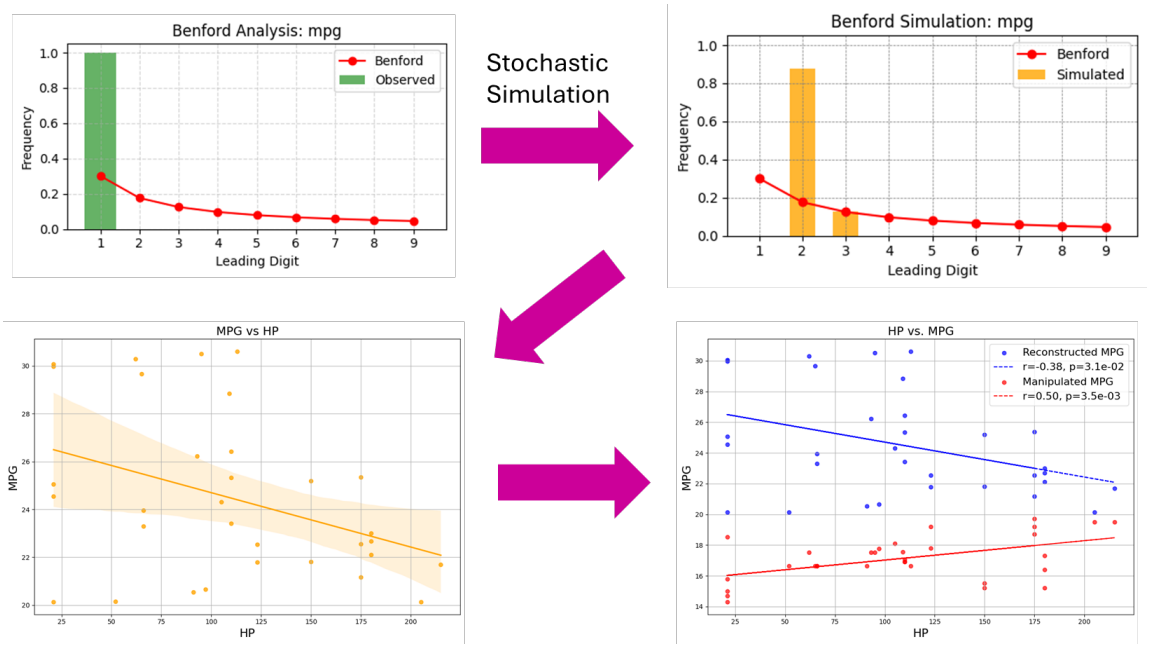

Figure 8: Benford's Law analysis is used to generate a digit distribution profile, which guides the stochastic simulation to estimate the unmanipulated dataset. The direction of the Pearson correlation in the estimated data is then analyzed to detect a reversal in correlation.

3.2.2 Random Forest Regressor

Random Forest is an ensemble learning algorithm that constructs multiple decision trees and combines their outputs to improve predictive accuracy and control overfitting. In regression tasks, each tree predicts a numerical value, and the final output is the average of all tree predictions [4, 5].

Mathematically, for a given input $\mathbf{x}$, the Random Forest regression prediction $\hat{y}$ is defined as:

$$\hat{y} = \frac{1}{T} \sum_{t=1}^{T} h_t(\mathbf{x}) \tag{4}$$

where T is the total number of decision trees in the forest, and $h_t(\mathbf{x})$ is the prediction of the t-th tree.

Originally developed for classification tasks, Random Forests are equally effective for regression problems due to their ability to model non-linear relationships and capture complex interactions between features. They are widely used in fields such as finance, bioinformatics, and engineering.

In the context of this study, the Random Forest Regressor is used to reconstruct an approximation of the original, unmanipulated dataset using the `scikit-learn` library [9]. By learning the underlying structure of clean data, the model estimates values that reflect the true correlation, thus supporting the detection of manipulation and reversal in data trends.

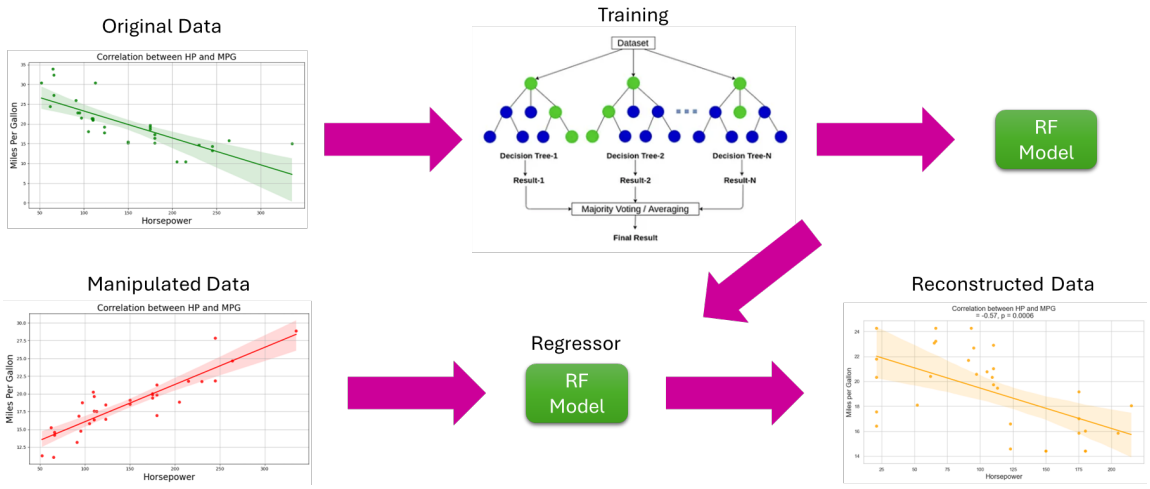

Figure 9: A Random Forest regressor is trained on the original data to learn the characteristics of the unmanipulated dataset. The manipulated data is then input into the pre-trained model to generate an estimated unmanipulated version. The Pearson correlation direction in the predicted data is analyzed to detect any reversal in correlation.

Figure 9 illustrates the process of reconstructing manipulated data using a Random Forest regressor. The model was first trained using the `scikit-learn` library [9] on the original, unmanipulated dataset to learn the underlying relationship between key features—specifically the negative correlation between horsepower (HP) and miles per gallon (MPG).

Once trained, the regressor was used to estimate an unmanipulated version of the manipulated dataset. The model was fine-tuned using the following hyperparameters: `n_estimators` = 300, `max_depth` = 4, `min_samples_leaf` = 5, and `random_state` = 2. The predicted (reconstructed) data is then analyzed to detect any reversal in correlation direction.

The results confirm that the regressor effectively captured the original relationship, restoring the expected negative correlation. This finding suggests that Random Forest regression is

effective not only in modeling complex relationships but also in approximating original patterns in data suspected of manipulation.

4 Result

4.1 Classification

4.1.1 Benford's Law

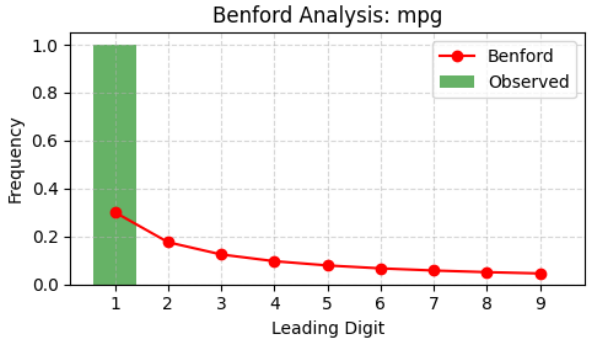

Figure 10: Benford's Law digit distribution of the manipulated `mpg` data. The observed distribution (in green) deviates notably from the expected Benford profile (in red), with several leading digits absent—indicating potential signs of manipulation.

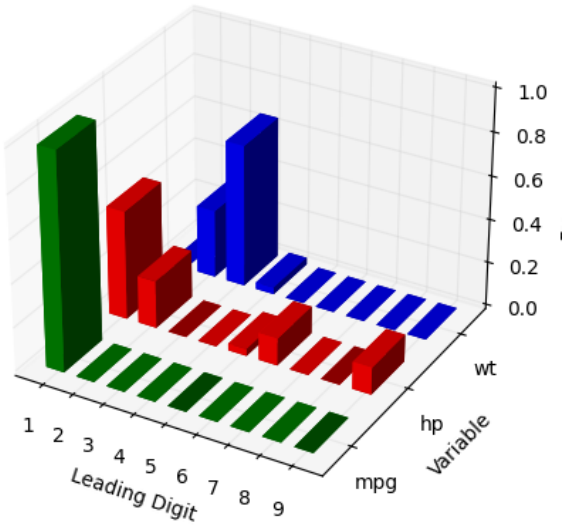

Figure 11: 3D visualization of Benford's Law distribution across three key features: miles per gallon (mpg) in green, horsepower (hp) in red, and weight (wt) in blue. The plot reveals significant deviations from the expected Benford profile, with several leading digits underrepresented or missing entirely.

The results demonstrate clear distinctions between manipulated and unmanipulated datasets based on conformity to Benford's Law. As shown in Figure 10, the digit distribution of the manipulated *mpg* data (green bars) deviates significantly from the expected Benford distribution (red line), with several leading digits missing or underrepresented—suggesting potential data tampering. This finding is further reinforced by the 3D visualization in Figure 11, which shows

Table 1: Benford's Law test results. Anomalies are flagged using thresholds of $\chi^2 > 30.58$ and $p < 0.0001$.

Dataset	χ^2	**p-value**	**Classification**
Manipulated	74.300	6.81×10^{-13}	Anomaly
Unmanipulated	23.383	0.00291	Acceptable

leading digit distributions across three major features: *mpg*, *hp*, and *wt* selected in Section 3.1.2. All three exhibit noticeable departures from the Benford profile, particularly *wt*, which shows a pronounced imbalance in digit occurrence. The profile of the expected Benford distribution (red line), obtained from this analysis, will later be used in Section 4.2.1 as input to the stochastic simulation process to estimate portions of the original data.

Statistical evidence is summarized in Table 1, where the manipulated dataset produces a χ^2 value of 74.3 and a p-value of 6.81×10^{-13}, well beyond the anomaly threshold ($\chi^2 > 30.58$, $p < 0.0001$). In contrast, the unmanipulated dataset yields a χ^2 value of 23.383 and a p-value of 0.00291, classifying it as acceptable. These results confirm that Benford's Law is effective in flagging suspicious irregularities in manipulated numerical data.

4.1.2 Random Forest Classifier

Table 2: Random Forest Test Results Using Thresholds of 50%

Dataset	**Prediction**	**Classification**
Manipulated	71.9 %	Anomaly
Unmanipulated	6.2%	Acceptable

Table 3: Performance Evaluation of Random Forest Classifier

Class	**Precision**	**Recall**	**F1-score**	**Support**	**Description**
0	0.83	1.00	0.91	5	Unmanipulated data: perfect recall, slightly lower precision.
1	1.00	0.88	0.93	8	Manipulated data: perfect precision, but slightly lower recall.
Accuracy			**0.92**	13	Overall correct classification rate.
Macro avg	0.92	0.94	0.92	13	Unweighted average across both classes.
Weighted avg	0.94	0.92	0.92	13	Weighted average based on class support.

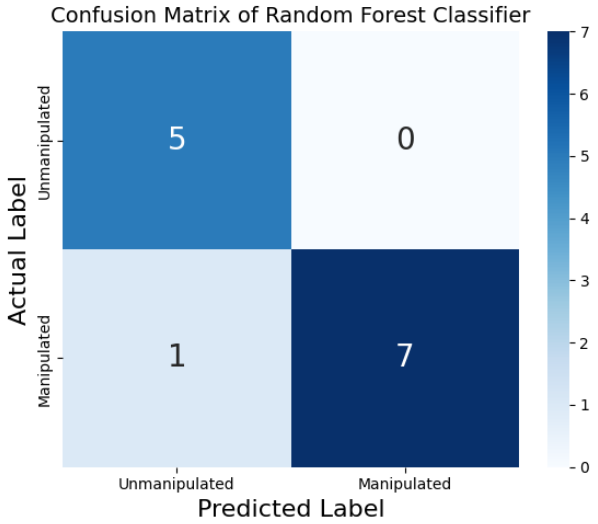

Figure 12: Confusion Matrix of Random Forest Classifier. The classifier correctly identified 5 unmanipulated and 7 manipulated instances, with only 1 misclassified manipulated case and no false positives for unmanipulated data. This results in an accuracy of 92.3%, precision of 1.00 for unmanipulated and 0.88 for manipulated data, and an overall F1-score of 0.91, demonstrating strong performance in distinguishing between manipulated and unmanipulated records.

The Random Forest model was evaluated using a 50% threshold for classification confidence, as summarized in Table 2. The manipulated dataset yielded a prediction rate of 71.9%, clearly surpassing the 50% threshold and resulting in an "Anomaly" classification. Conversely, the unmanipulated dataset had a low prediction rate of 6.2%, falling below the threshold and thus considered "Acceptable."

Table 3 presents the detailed performance metrics of the classifier using the `scikit-learn` library [9]. For class 0 (unmanipulated), the model achieved perfect recall (1.00) and a precision of 0.83. For class 1 (manipulated), the classifier reached perfect precision (1.00) and a recall of 0.88. The overall accuracy of the model was 92%, with a macro-averaged F1-score of 0.92.

These results are visually supported by the confusion matrix in Figure 12. The classifier correctly identified all 5 unmanipulated cases and 7 out of 8 manipulated cases, with only one false negative and no false positives. This performance corresponds to a precision of 1.00 for unmanipulated data and 0.88 for manipulated data, reinforcing the classifier's strong ability to distinguish between manipulated and authentic records.

4.2 Prediction

4.2.1 Benford's Law

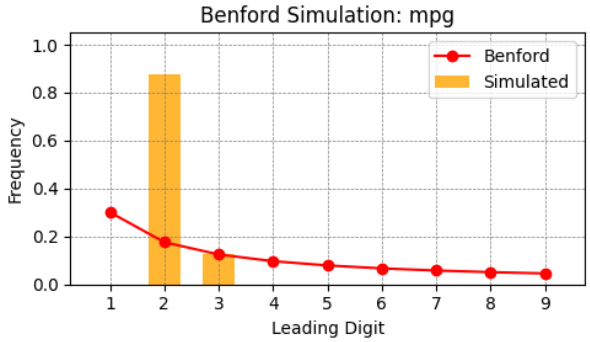

Figure 13: Using the Benford's Law digit distribution profile of the manipulated `mpg` data from the previous section, stochastic simulation was applied to estimate portions of the original data. The resulting distribution reveals more complete leading digit representation, showing improved compliance with Benford's Law.

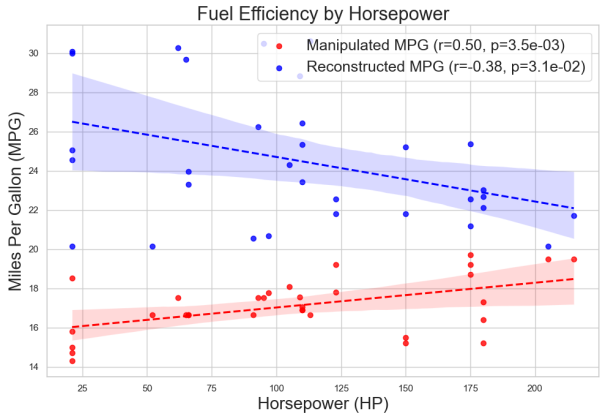

Figure 14: The manipulated data (red) shows a positive Pearson correlation between horsepower and miles per gallon (r = +0.50), whereas the reconstructed data (blue), generated using stochastic simulation guided by Benford's Law, reveals a negative correlation (r = -0.38). This indicates that the manipulation altered the data in a way that reversed the original direction of the correlation.

As shown in Figure 13, a stochastic simulation was conducted using Benford's Law to reconstruct portions of the manipulated `mpg` dataset. The resulting simulated distribution (orange bars) exhibits a more complete and balanced representation of leading digits, aligning more closely with the theoretical Benford profile (red line). This suggests improved compliance and plausibility compared to the original manipulated data.

Figure 14 further evaluates the effect of reconstruction by comparing the relationships between horsepower and miles per gallon. The manipulated data (red) displays a misleading positive Pearson correlation ($r = +0.50$), while the reconstructed data (blue), generated via the Benford-based simulation, shows a negative correlation ($r = -0.38$), which is more consistent with expected automotive behavior. This reversal indicates that the manipulation significantly distorted the original relationship, and that the simulation helped recover a more realistic pattern.

4.2.2 Random Forest Regressor

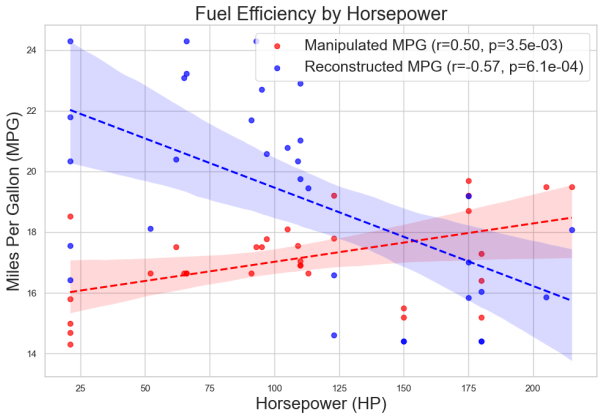

Figure 15: The manipulated data (red) exhibits a positive Pearson correlation between horsepower and miles per gallon (r = +0.50), while the reconstructed data (blue), derived through stochastic simulation based on Benford's Law, reveals a negative correlation (r = –0.57). This indicates that the manipulation altered the data in a way that reversed the original direction of the correlation.

Table 4: Evaluation Metrics for the Random Forest Regression Model

Metric	**Value**	**Description**
R^2 Score	0.6643	Proportion of variance in the target variable explained by the model; values closer to 1 indicate a better fit.
Mean Squared Error (MSE)	4.1079	Average squared difference between predicted and actual values; lower values reflect better predictive accuracy.
Mean Absolute Error (MAE)	1.7576	Average absolute difference between predicted and actual values; less sensitive to outliers than MSE.

Figure 15 compares the correlation between horsepower and miles per gallon in the manipulated versus reconstructed datasets. The manipulated data (red) shows a positive Pearson correlation ($r = +0.50$), while the reconstructed data (blue), generated using a Random Forest regressor trained on Benford-compliant features, reveals a strong negative correlation ($r = -0.57$). This reversal suggests that the manipulation distorted the natural relationship between the features and that the regression model successfully recovered a more realistic pattern consistent with domain expectations.

Table 4 summarizes the evaluation metrics for the Random Forest regression model using the `scikit-learn` library [9]. The model achieved an R^2 score of 0.6643, indicating that it explains approximately 66.4% of the variance in the target variable. The Mean Squared Error (MSE) of 4.1079 and Mean Absolute Error (MAE) of 1.7576 reflect reasonably accurate predictions, with lower error magnitudes. These results demonstrate that the model is effective in approximating the underlying structure of the original, unmanipulated data.

5 Discussion

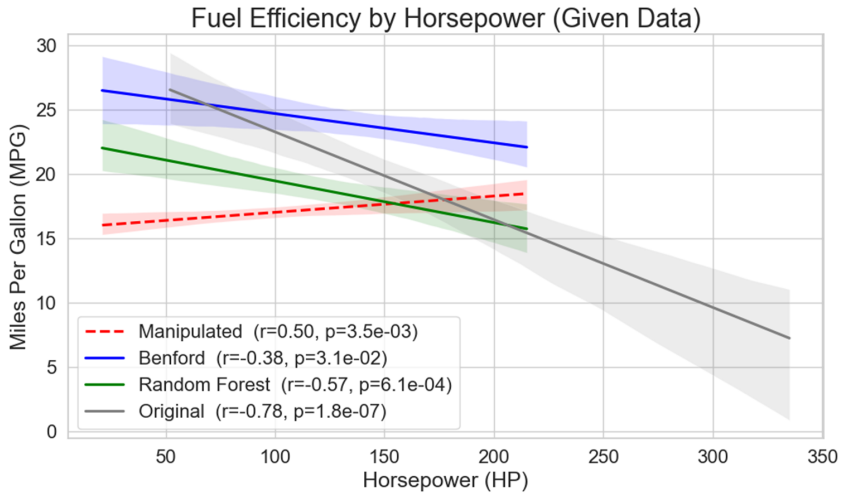

Figure 16: Regression trends for given and original datasets: the original data trend is shown in gray, the manipulated data (given data) trend is represented by a red dashed line, the Benford-reconstructed data trend appears as a solid blue line, and the Random Forest-reconstructed data trend is displayed in green

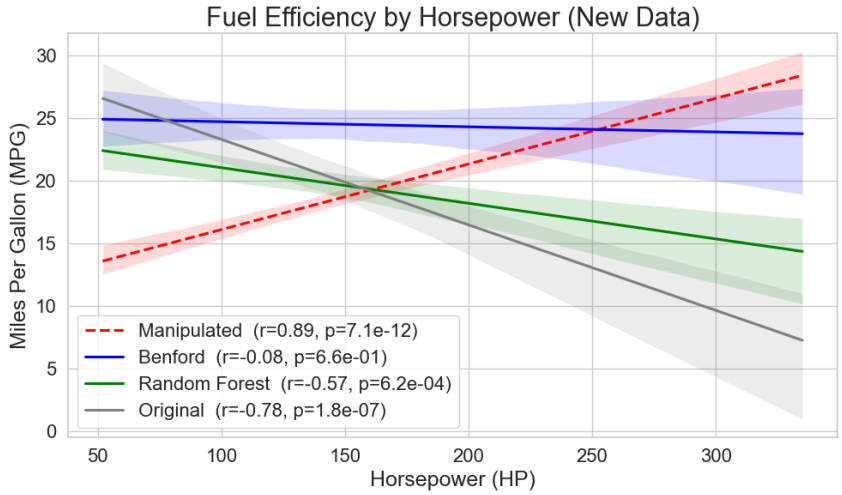

Figure 17: Regression trends for new and original datasets: the original data trend is shown in gray, the manipulated data (new manipulated data) trend is represented by a red dashed line, the Benford-reconstructed data trend appears as a solid blue line, and the Random Forest-reconstructed data trend is displayed in green

Figures 16 and 17 illustrate regression trends across original, manipulated, and reconstructed datasets. In both cases, the original data (gray line) reveals a strong negative correlation between horsepower and miles per gallon (MPG), consistent with known automotive principles. In contrast, the manipulated datasets exhibit misleading positive correlations—moderate in the given data ($r = +0.50$) and even stronger in the newly manipulated version ($r = +0.89$)—indicating substantial distortion of the underlying trend.

The Benford-reconstructed data (blue line), generated through stochastic simulation based on digit frequencies, recovers the correct trend direction but with reduced strength, particularly in the new dataset where the correlation is nearly flat ($r = -0.08$). In comparison, the Random

Forest-reconstructed data (green line) more accurately reflects the original structure, recovering a consistent and meaningful negative correlation ($r = -0.57$) in both cases.

These results support a hybrid approach: Benford's Law acts as a robust, unsupervised detector for identifying anomalies based on digit patterns, requiring no training data. Conversely, the Random Forest regressor requires access to labeled clean data but excels at reconstructing the quantitative structure of the original dataset. Together, these methods form a complementary framework for detecting data manipulation and identifying correlation direction reversal.

Limitations: This study is based on a small dataset ($n = 32$), a size commonly used in scientific and engineering research, particularly in controlled experiments or when data collection is costly. However, small sample sizes may limit generalizability, and Random Forest models can overfit if not properly tuned.

Future Work: Applying this methodology to larger, real-world datasets in domains such as epidemiology, finance, or public policy could further evaluate its robustness and practical scalability.

Overall, these findings reinforce the value of combining statistical and machine learning techniques to ensure data integrity and uncover manipulation in research data.

References

[1] Pearson, K. (1895). Note on Regression and Inheritance in the Case of Two Parents. *Proceedings of the Royal Society of London*, 58, 240–242.

[2] Benford, F. (1938). The Law of Anomalous Numbers. *Proceedings of the American Philosophical Society*, 78(4), 551–572.

[3] Hill, T. P. (1995). A Statistical Derivation of the Significant-Digit Law. *Statistical Science*, 10(4), 354–363.

[4] Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32.

[5] Liaw, A., & Wiener, M. (2002). Classification and Regression by Random Forest. *R News*, 2(3), 18–22.

[6] Robert, C. P., & Casella, G. (2004). *Monte Carlo Statistical Methods* (2nd ed.). Springer.

[7] Nigrini, M. J. (2012). *Benford's Law: Applications for Forensic Accounting, Auditing, and Fraud Detection*. Wiley.

[8] Zhou, W., & Tam, K. H. (2018). Detecting Data Manipulation with Benford's Law: A Review of Methodologies. *Forensic Science International*, 289, 45–58.

[9] Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow* (2nd ed.). O'Reilly Media.

Appendix

A Original Data

Table 5: Original Car Dataset

Model	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21	6	160	110	3.9	2.62	16.46	0	1	4	4
Mazda RX4 Wag	21	6	160	110	3.9	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108	93	3.85	2.32	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360	175	3.15	3.44	17.02	0	0	3	2
Valiant	18.1	6	225	105	2.76	3.46	20.22	1	0	3	1
Duster 360	14.3	8	360	245	3.21	3.57	15.84	0	0	3	4
Merc 240D	24.4	4	146.7	62	3.69	3.19	20	1	0	4	2
Merc 230	22.8	4	140.8	95	3.92	3.15	22.9	1	0	4	2
Merc 280	19.2	6	167.6	123	3.92	3.44	18.3	1	0	4	4
Merc 280C	17.8	6	167.6	123	3.92	3.44	18.9	1	0	4	4
Merc 450SE	16.4	8	275.8	180	3.07	4.07	17.4	0	0	3	3
Merc 450SL	17.3	8	275.8	180	3.07	3.73	17.6	0	0	3	3
Merc 450SLC	15.2	8	275.8	180	3.07	3.78	18	0	0	3	3
Cadillac Fleetwood	10.4	8	472	205	2.93	5.25	17.98	0	0	3	4
Lincoln Continental	10.4	8	460	215	3	5.424	17.82	0	0	3	4
Chrysler Imperial	14.7	8	440	230	3.23	5.345	17.42	0	0	3	4
Fiat 128	32.4	4	78.7	66	4.08	2.2	19.47	1	1	4	1
Honda Civic	30.4	4	75.7	52	4.93	1.615	18.52	1	1	4	2
Toyota Corolla	33.9	4	71.1	65	4.22	1.835	19.9	1	1	4	1
Toyota Corona	21.5	4	120.1	97	3.7	2.465	20.01	1	0	3	1
Dodge Challenger	15.5	8	318	150	2.76	3.52	16.87	0	0	3	2
AMC Javelin	15.2	8	304	150	3.15	3.435	17.3	0	0	3	2
Camaro Z28	13.3	8	350	245	3.73	3.84	15.41	0	0	3	4
Pontiac Firebird	19.2	8	400	175	3.08	3.845	17.05	0	0	3	2
Fiat X1-9	27.3	4	79	66	4.08	1.935	18.9	1	1	4	1
Porsche 914-2	26	4	120.3	91	4.43	2.14	16.7	0	1	5	2
Lotus Europa	30.4	4	95.1	113	3.77	1.513	16.9	1	1	5	2
Ford Pantera L	15.8	8	351	264	4.22	3.17	14.5	0	1	5	4
Ferrari Dino	19.7	6	145	175	3.62	2.77	15.5	0	1	5	6
Maserati Bora	15	8	301	335	3.54	3.57	14.6	0	1	5	8
Volvo 142E	21.4	4	121	109	4.11	2.78	18.6	1	1	4	2

Data source: `datasource:https://rpubs.com/hmiller/mtcars-example`

B Manipulated Data

Table 6: Manipulated Car Dataset

Model	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	16.91	6	160	110	3.9	2.62	16.46	0	1	4	4
Mazda RX4 Wag	16.91	6	160	110	3.9	2.875	17.02	0	1	4	4
Datsun 710	17.52	4	108	93	3.85	2.32	18.61	1	1	4	1
Hornet 4 Drive	17.06	6	258	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360	175	3.15	3.44	17.02	0	0	3	2
Valiant	18.1	6	225	105	2.76	3.46	20.22	1	0	3	1
Duster 360	14.3	8	360	21	3.21	3.57	15.84	0	0	3	4
Merc 240D	17.52	4	146.7	62	3.69	3.19	20	1	0	4	2
Merc 230	17.52	4	140.8	95	3.92	3.15	22.9	1	0	4	2
Merc 280	19.2	6	167.6	123	3.92	3.44	18.3	1	0	4	4
Merc 280C	17.8	6	167.6	123	3.92	3.44	18.9	1	0	4	4
Merc 450SE	16.4	8	275.8	180	3.07	4.07	17.4	0	0	3	3
Merc 450SL	17.3	8	275.8	180	3.07	3.73	17.6	0	0	3	3
Merc 450SLC	15.2	8	275.8	180	3.07	3.78	18	0	0	3	3
Cadillac Fleetwood	19.5	8	472	205	2.93	2.411	17.98	0	0	3	4
Lincoln Continental	19.5	8	460	215	3	2.411	17.82	0	0	3	4
Chrysler Imperial	14.7	8	440	21	3.23	2.411	17.42	0	0	3	4
Honda Civic	16.65	4	75.7	52	4.93	3.31	18.52	1	1	4	2
Toyota Corolla	16.65	4	71.1	65	4.22	3.31	19.9	1	1	4	1
Toyota Corona	17.77	4	120.1	97	3.7	2.465	20.01	1	0	3	1
Dodge Challenger	15.5	8	318	150	2.76	3.52	16.87	0	0	3	2
AMC Javelin	15.2	8	304	150	3.15	3.435	17.3	0	0	3	2
Camaro Z28	18.53	8	350	21	3.73	3.84	15.41	0	0	3	4
Pontiac Firebird	19.2	8	400	175	3.08	3.845	17.05	0	0	3	2
Fiat X1-9	16.65	4	79	66	4.08	3.31	18.9	1	1	4	1
Porsche 914-2	16.65	4	120.3	91	4.43	3.31	16.7	0	1	5	2
Lotus Europa	16.65	4	95.1	113	3.77	3.31	16.9	1	1	5	2
Ford Pantera L	15.8	8	351	21	4.22	3.17	14.5	0	1	5	4
Ferrari Dino	19.7	6	145	175	3.62	2.77	15.5	0	1	5	6
Maserati Bora	15	8	301	21	3.54	3.57	14.6	0	1	5	8
Volvo 142E	17.56	4	121	109	4.11	2.78	18.6	1	1	4	2

Data source: `https://github.com/pnetsiri/Interdisciplinary/blob/main/mtcars_manipulated.csv`

C New Manipulated Data

Table 7: New Manipulated Car Dataset

Model	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	17.593	6	160	110	3.9	2.62	16.46	0	1	4	4
Mazda RX4 Wag	16.323	6	160	110	3.9	2.875	17.02	0	1	4	4
Datsun 710	16.875	4	108	93	3.85	2.32	18.61	1	1	4	1
Hornet 4 Drive	19.646	6	258	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	20.032	8	360	175	3.15	3.44	17.02	0	0	3	2
Valiant	15.832	6	225	105	2.76	3.46	20.22	1	0	3	1
Duster 360	27.858	8	360	245	3.21	3.57	15.84	0	0	3	4
Merc 240D	15.255	4	146.7	62	3.69	3.19	20	1	0	4	2
Merc 230	14.761	4	140.8	95	3.92	3.15	22.9	1	0	4	2
Merc 280	18.465	6	167.6	123	3.92	3.44	18.3	1	0	4	4
Merc 280C	16.453	6	167.6	123	3.92	3.44	18.9	1	0	4	4
Merc 450SE	19.869	8	275.8	180	3.07	4.07	17.4	0	0	3	3
Merc 450SL	21.284	8	275.8	180	3.07	3.73	17.6	0	0	3	3
Merc 450SLC	16.973	8	275.8	180	3.07	3.78	18	0	0	3	3
Cadillac Fleetwood	18.85	8	472	205	2.93	2.411	17.98	0	0	3	4
Lincoln Continental	21.775	8	460	215	3	2.411	17.82	0	0	3	4
Chrysler Imperial	21.774	8	440	230	3.23	2.411	17.42	0	0	3	4
Fiat 128	14.588	4	78.7	66	4.08	2.2	19.47	1	1	4	1
Honda Civic	11.304	4	75.7	52	4.93	3.31	18.52	1	1	4	2
Toyota Corolla	11.075	4	71.1	65	4.22	3.31	19.9	1	1	4	1
Toyota Corona	18.751	4	120.1	97	3.7	2.465	20.01	1	0	3	1
Dodge Challenger	18.548	8	318	150	2.76	3.52	16.87	0	0	3	2
AMC Javelin	19.135	8	304	150	3.15	3.435	17.3	0	0	3	2
Camaro Z28	21.851	8	350	245	3.73	3.84	15.41	0	0	3	4
Pontiac Firebird	19.411	8	400	175	3.08	3.845	17.05	0	0	3	2
Fiat X1-9	14.182	4	79	66	4.08	3.31	18.9	1	1	4	1
Porsche 914-2	13.158	4	120.3	91	4.43	3.31	16.7	0	1	5	2
Lotus Europa	17.531	4	95.1	113	3.77	3.31	16.9	1	1	5	2
Ford Pantera L	24.639	8	351	264	4.22	3.17	14.5	0	1	5	4
Ferrari Dino	19.917	6	145	175	3.62	2.77	15.5	0	1	5	6
Maserati Bora	28.897	8	301	335	3.54	3.57	14.6	0	1	5	8
Volvo 142E	20.245	4	121	109	4.11	2.78	18.6	1	1	4	2

Data source: `https://github.com/pnetsiri/Interdisciplinary/blob/main/My_mpg_manipulated.csv`

Abbreviation

The abbreviation below provide definitions for key automotive features used in the dataset.

- `mpg` – Miles/(US) gallon
- `cyl` – Number of cylinders
- `disp` – Displacement (cu.in.)
- `hp` – Gross horsepower
- `drat` – Rear axle ratio
- `wt` – Weight (1000 lbs)
- `qsec` – 1/4 mile time
- `vs` – Engine (0 = V-shaped, 1 = straight)
- `am` – Transmission (0 = auto, 1 = manual)
- `gear` – Number of forward gears
- `carb` – Number of carburetors

Code Availability

The code and data used in this project are publicly available on GitHub at the following repository:

`https://github.com/pnetsiri/Interdisciplinary/`

Figure 18: QR code linking to the GitHub repository containing all code, data, and materials for this project

Users are welcome to explore, reproduce, and extend the work. Contributions and feedback are encouraged.