Exploring the Effects of Key Vehicle Features on Fuel Efficiency

Berke Derin Berktay, Maidah Shah, Paris Wang, Bonnie Yam

December 13, 2023

Abstract:

This study investigates key vehicle features affecting fuel efficiency using the Auto dataset comprising 392 vehicles. It focuses on the relationship between miles per gallon (mpg) and various vehicle characteristics such as weight, year, origin, and engine specifics. The research employs both linear and logistic regression analyses to discern patterns. Key findings indicate that *horsepower, weight, year,* and *origin* were statistically significant predictors of fuel efficiency; while *cylinders, weight,* and *year* were statistically significant predictors of high fuel efficiency. However, it is strongly recommended to interpret these findings in the context of the limitations discussed in the report. Overall, this research can be helpful in understanding fuel efficiency trends, aiding in environmental protection, and developing effective marketing strategies for vehicles.

1 Introduction:

Miles per gallon (mpg) describes how many miles a vehicle can travel per each gallon of fuel. It is often one of the main considerations that influences a customer's decision when purchasing a new vehicle. Drivers typically prefer a vehicle that is capable of travelling further distances per each gallon of fuel in order to save on fuel costs. Overall, the higher a vehicle's mpg, the better its fuel efficiency.

In addition to saving money, governments have set auto emission standards to reduce the release of carbon dioxide which contributes to global warming. As mpg and fuel efficiency increase, emissions decrease. In 1975, the Corporate Average Fuel Economy (CAFE) standards were enacted in the United States [1]. These standards called for manufacturers to double fuel efficiency to 27.5 mpg within 10 years [1]. The Japanese government also adopted the same standards [2]. Later on in 1988, the European Union started regulating vehicle emissions by calling for manufacturers to reduce emissions to 140 g CO_2/km, which is equivalent to approximately 40 mpg [1].

Motivated by global warming and cost saving concerns, this study will look to investigate which key features of vehicles contribute to better fuel efficiency, i.e., higher mpg. This is an important consideration for both the environment and for marketing purposes when selling vehicles to customers concerned with reducing emissions and increasing fuel savings. The study will use the Auto dataset [3], which contains information on 392 vehicles including the name, weight, year, origin, and other details relevant to the engine.

Table 1 displays the descriptive statistics for the variables selected from the dataset to be included in the study analysis.

Table 1. Descriptive Statistics for Variables in Regression Analyses (N = 392)

Variables	Mean	SD	Range
Independent Variables			
cylinders	N/A	N/A	3 - 8
displacement	194.4	104.6	68 - 455
horsepower	104.5	38.5	46 - 230
weight	2977.6	849.4	1613 - 5140

acceleration	15.5	2.8	8 - 24.8
year	N/A	N/A	70 - 82
origin	N/A	N/A	1 - 3
Dependent Variable			
mpg	23.4	7.8	9 - 46.6

Figure 1 displays the correlation between the selected variables. The left-most column displays the correlation between the numeric variables and mpg. As can be seen, *acceleration* has a weak positive correlation with *mpg*, whereas *displacement, horsepower,* and *weight* have a strong negative correlation with mpg.

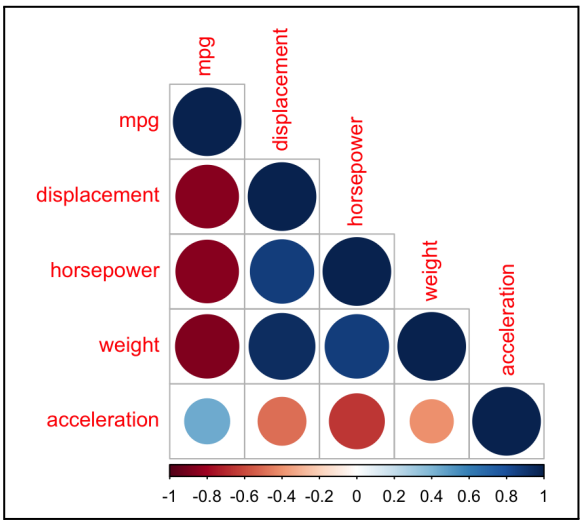


Figure 1. Correlation Matrix for Numeric Variables in Regression Analyses

2 Methods:

When preparing the data for analysis, it was decided to drop the *name* variable, as it was determined that the *year* and *origin* variables were sufficient to describe a vehicle's history. Additionally, it was found that the *cylinders, year,* and *origin* variables were discrete, and thus, were converted to categorical for the model cases where one-hot encoding was used. However, models without one-hot encoding used variables in their discrete forms. *Origin* 1, 2, and 3 represented The United States, Europe, and Japan, respectively, and these were already encoded in the dataset. There were initially 12 categories for *year* (i.e., 1971-1982), but the values were grouped into two decades, i.e., 70's and 80's, to reduce the number of coefficients in the regression models.

Descriptive statistics were prepared and a correlation plot was created to observe the individual correlation levels between *mpg* and the numerical features. It was then decided to perform a multiple linear regression as a starting point for the statistical analysis, as the test would provide a baseline understanding of the relationships between the variables. The multiple linear regression was conducted between the dependent variable, *mpg*, and the remaining features as independent variables. This satisfied the variable requirements of multiple linear regression, as the test sought to predict a continuous variable, *mpg*, using at least two predictors of nominal, ordinal, or interval level.

Given the insights drawn from the multiple linear regression and the scan of auto emissions standards as outlined in the introduction, it was decided to perform additional statistical tests to

determine which features could be used to predict whether a vehicle had high fuel efficiency. As such, feature engineering was used to create a new binary outcome variable, *HighMPG*, where 1 = 'high mpg' and 0 = 'low mpg'. The creation of this variable was informed by the CAFE auto emissions standards [1], which called for vehicle manufacturers to double fuel efficiency to 27.5 mpg. This was used as the threshold to categorize mpg > 27.5 as 'high mpg', and otherwise, as 'low mpg'.

Thereafter, a binary logistic regression was conducted between the dependent variable, *HighMPG*, and the remaining features as independent variables. This satisfied the variable requirements of the binary logistic regression, as the test sought to predict a binary outcome, *HighMPG*, using at least two binary or continuous predictors.

After the completion of the binary logistic regression, it was decided that cost and fuel efficiency should be observed in a more prioritized manner, as while the binary classification served well for basic compliance, it oversimplified the scenario. Environmental impact and fuel cost are not just about being above or below a threshold; they involve degrees of efficiency and impact. As such, a multinomial logistic regression was performed, as the test allowed a more nuanced categorization of fuel efficiency, for example, in the form of low, moderate, or high fuel efficiency. This categorization can also help consumers to make more informed decisions based on fuel efficiency levels beyond just the standardized thresholds.

For the multinomial logistic regression, feature engineering was used to create a new outcome variable, *mpg_cluster*, where each vehicle was assigned to either cluster 0, 1, or 2. Before creating this variable, K-means clustering was performed with the elbow method to determine the optimal K-value. As seen in Appendix 1, the elbow in question is 3. The cluster visualization in Appendix 2, shows a clear distinction between the mpg groupings. It is interesting to note that the model used cluster 1 as the lowest numerical value group and cluster 0 as the next lowest.

Furthermore, it should be noted that each of the three regression tests (multiple linear, binary logistic, and multinomial logistic) were performed both with and without one-hot encoding. Boxplots were also produced for each variable to observe outliers, and capping was performed to reduce the effect of outliers. This was done to increase the accuracy of the model for the majority of the dataset. Cross-validation was also performed to give a more accurate measure of a model's performance by using multiple training and validation sets. Based on the results of the models both with and without one-hot encoding, conclusions concerning whether or not to use one-hot encoding were made. After performing exploratory data analysis, multiple linear regression, binary logistic regression, and multinomial logistic regression with the help of feature engineering and machine learning techniques, the statistical results were ready for further analysis in the context of the research question in the study.

3 Results & Discussion:

A multiple linear regression was conducted on the outcome variable, *mpg*. The results of the regression are included in Appendix 3. At a significance level of $\alpha = 0.05$, it was found that out of the seven vehicle features being measured, *weight*, *year*, *origin* and *horsepower* were the only statistically significant predictors of *mpg*. *Weight*, *year*, and *origin* had extremely small p-values near 0, while *horsepower* had a p-value < 0.05. The coefficients of *horsepower* and *weight* were negative, whereas the coefficient for *year* was positive. In other words, more recently manufactured vehicles that weigh less and have lower horsepower, were more likely to have better fuel efficiency, holding *cylinders,*

displacement, and *acceleration* constant. The *origin* of the vehicle, i.e., The United States, Europe, or Japan, also influenced fuel efficiency. However, given the method by which *origin* was encoded in the dataset and model, it was not meaningful to interpret the coefficient. Rather, solely the p-value was used to determine that *origin* was a statistically significant predictor of fuel efficiency.

Additionally, the feature engineering technique, one-hot encoding, was used to determine which *year* and *origin* of a vehicle had greater influence in predicting fuel efficiency. As shown in Appendices 3 and 5, the adjusted R^2 of the models with and without one-hot encoding were 83.2% and 82.9%, respectively. Although this was not a significant difference, the model without one-hot encoding had a lower standard deviation by 1.5% as seen in Appendices 4 and 6. Therefore, it was decided to use the multiple linear regression model without one-hot encoding, and overall, determine that *horsepower, weight, year,* and *origin* were statistically significant predictors of fuel efficiency.

Furthermore, a binary logistic regression without one-hot encoding was conducted on the new outcome variable, *HighMPG.* The results of the regression are included in Appendix 7. At a significance level of $\alpha = 0.05$, it was found that out of the seven vehicle features being measured, *cylinders, weight,* and *year* had p-values < 0.05, and as such, were the only statistically significant predictors of *HighMPG*. The coefficients for *cylinders* and *year* were positive, whereas the coefficient for *weight* was negative. In other words, more recently manufactured vehicles that weigh less and have more cylinders are more likely to have high fuel efficiency, i.e., mpg > 27.5, holding all other variables constant.

It was also found that both binary logistic regression models with and without one-hot encoding had the same accuracies at 89.28% as seen in Appendices 8 and 10. Further details of the summary regression results for the binary logistic regression model with one-hot encoding can be found in Appendix 9. The model with one-hot encoding had a greater standard deviation by 1.27%. Thus, the binary logistic regression model without one-hot encoding was selected for inclusion in the analysis.

In Appendix 11, it can be seen that in the multinomial logistic regression model without one-hot encoding, each cluster had a different number of predictors for *mpg*. In cluster 1, *weight* had a p-value of 0.003, while *year* had a p-value near 0. In cluster 2, *cylinders, displacement, origin,* and *year* had p-values of 0.002, 0.047, 0.014, and near 0, respectively. The model without one-hot encoding had an accuracy of 84.75%, versus an 80.51% accuracy from the model with one-hot encoding, as seen in Appendices 11 and 12. However, given the context of the study where fuel efficiency was being categorized into two groups, either lower or higher than 27.5 mpg, it was decided to focus on the multiple linear and binary logistic regression models, rather than the cluster-based multinomial logistic regression model which recommended classifying fuel efficiency further than these two groups, i.e., into three clusters. While this test was helpful to understand which variables were statistically significant predictors of fuel efficiency in each of the three clusters, it was decided that these findings would be more meaningful to explore further along with supporting literature and automotive laws as next steps.

Along with comparing the accuracy and standard deviation scores of each model with and without one-hot encoding, another decision making factor for using models without one-hot encoding was to maintain a consistent and fair approach across all three regression tests. In terms of the similar accuracy scores for the models both with and without one-hot encoding, research found that this may be a result of using variables like *cylinders* and *year*, which although numerical, could be interpreted

as having an inherent rank order (i.e., 3 cylinders < 5 cylinders, or 1980 > 1973) [4]. This would make these variables less suitable for one-hot encoding, and thus, may not influence the accuracy score between the two models with and without one-hot encoding as expected.

4 Limitations and Conclusion:

This study looked to examine which key vehicle features affect fuel efficiency, and whether a vehicle is likely to have high fuel efficiency based on seven attributes: *cylinders, displacement, horsepower, weight, acceleration, year,* and *origin.*

A multiple linear regression was conducted and it was found that *horsepower, weight, year,* and *origin* were the only statistically significant predictors of fuel efficiency. Furthermore, a logistic regression was conducted on a new binary outcome, *HighMPG*, and it was found that *cylinders, weight,* and *year* were the only statistically significant predictors of high fuel efficiency, i.e., higher than 27.5 mpg.

There were also several limitations observed in the study. A limitation of the method included one-hot encoding all three discrete variables, *cylinders*, *year*, and *origin*, in the regression models. The values of the *cylinders* and *year* variables could be meaningful without one-hot encoding, whereas the *origin* variable could be interpreted as a nominal variable where assigning values to countries of origin did not imply a rank order. As such, it is recommended as a next step to rebuild these models with only one-hot encoding the *origin* variable to determine how this affects the models' accuracy, and whether such models should be used in place of models where all three discrete variables are one-hot encoded, or models with no one-hot encoding.

Another limitation of the study came from the use of secondary data. This restricted the factors that were able to be assessed in terms of influence on predicting fuel efficiency, as compared to data being collected specifically for the purpose of this study. There were also limited details on when, where, or how the data was collected, or which vehicles were included in the sample. Along with the small sample size of the dataset, these factors limited the generalizability of the study's findings to a larger population. As next steps, it is recommended to seek further clarification on the data collection process to help to better understand restrictions around the generalizability of the study. As well, determining if a larger sample was collected can help to enrich the analysis, and potentially, increase the generalizability of the study's findings to a larger population.

A third limitation of the study came from the removal of the *name* variable during data preprocessing. This variable contained information on the make and model of a vehicle. While it was initially determined that this feature would be difficult to categorize into similar classes and interpret in the context of a regression model, as a next step, it would be recommended to explore various approaches and extract meaningful information from this variable to observe trends between specific vehicle models and fuel efficiency.

Moreover, it is also recommended to review further research on existing studies and automotive laws to better understand the relationship between key vehicle features and fuel efficiency. This can help to assign meaning to the study's findings by determining how well the results support findings from other literature on predictors of high fuel efficiency. This can also help to understand uncertainties in the studies findings. One such uncertainty was the finding from the binary logistic regression model that showed that vehicles with more cylinders had better fuel efficiency. There are

many studies which contradict this finding, so it is recommended to further look into this. Another uncertainty was the finding from the elbow method which recommended comparing vehicles in three clusters, rather than the current two clusters informed by automotive laws which suggested a threshold value of 27.5 mpg for the *HighMPG* variable. Studies which classify fuel efficiency as low, medium, or high can be reviewed and used to inform the development of another regression model for comparison purposes.

Overall, while this research can be helpful in understanding fuel efficiency trends for environmental protection and marketing strategies, findings should be considered in the context of the limitations and next steps discussed in the report.

References

[1] Klier, T., & Linn, J. (2016). *Comparing US and EU Approaches to Regulating Automotive Emissions and Fuel Economy*. Retrieved from https://www.rff.org/publications/issue-briefs/comparing-us-and-eu-approaches-to-regulating-automotive-emissions-and-fuel-economy/

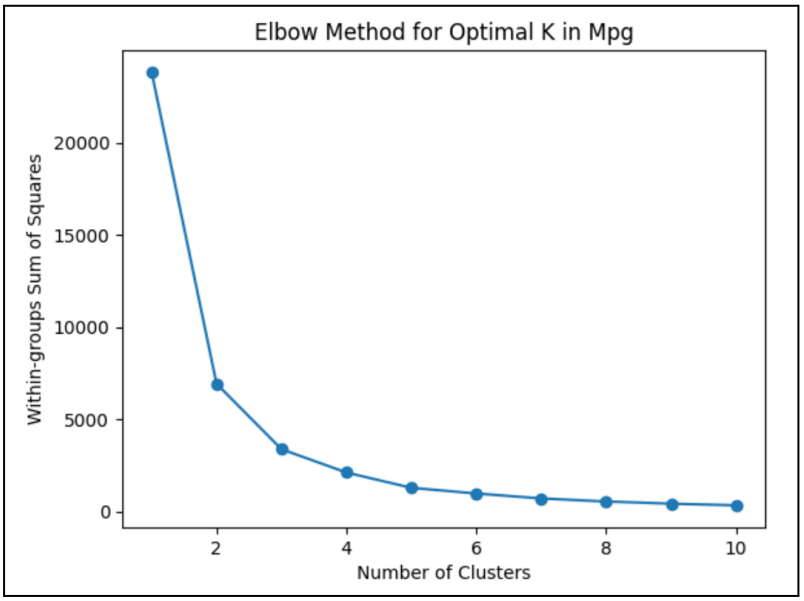
[2] Campbell, L. B., & Madrid-Crost, M. C. (1992). The competitive effects of U.S. and Japanese auto emission standards: are strong environmental regulations the reason Japanese cars sell themselves? *Canada-United States Law Journal*, *18*, 287-.

[3] "Auto: Auto data set in ISLR2: Introduction to statistical learning, Second edition," Auto: Auto Data Set in ISLR2: Introduction to Statistical Learning, Second Edition, https://rdrr.io/cran/ISLR2/man/Auto.html (accessed Dec. 2, 2023).

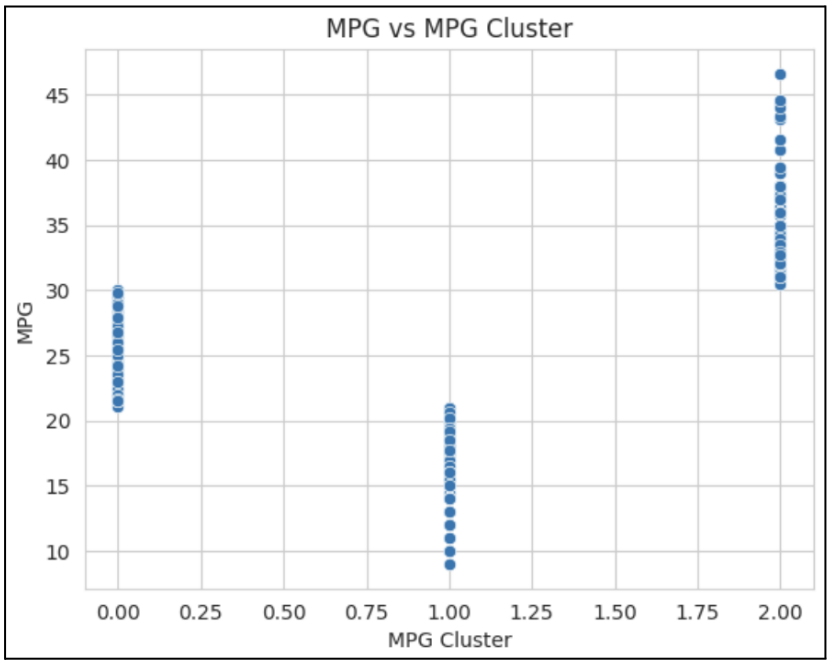
[4] V. Dey, "When to use one-hot encoding in deep learning?," Analytics India Magazine, https://analyticsindiamag.com/when-to-use-one-hot-encoding-in-deep-learning/ (accessed Dec. 11, 2023).

Appendix

Appendix 1: Elbow Method for Finding the Optimal Number of Clusters in Mpg



Appendix 2: K-means Cluster Visualization



Appendix 3: Summary Regression Results for Multiple Linear Regression Model Without One-hot Encoding

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    mpg   R-squared:                       0.834
Model:                            OLS   Adj. R-squared:                  0.829
Method:                 Least Squares   F-statistic:                     190.5
Date:                Sat, 02 Dec 2023   Prob (F-statistic):          9.83e-100
Time:                        19:06:03   Log-Likelihood:                -699.37
No. Observations:                 274   AIC:                             1415.
Df Residuals:                     266   BIC:                             1444.
Df Model:                           7
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        -11.7590      5.440     -2.162      0.032     -22.470      -1.048
cylinders     -0.1743      0.366     -0.476      0.634      -0.894       0.546
displacement   0.0136      0.009      1.507      0.133      -0.004       0.031
horsepower    -0.0372      0.018     -2.106      0.036      -0.072      -0.002
weight        -0.0058      0.001     -7.558      0.000      -0.007      -0.004
acceleration  -0.0713      0.127     -0.561      0.575      -0.322       0.179
year           0.7027      0.058     12.119      0.000       0.589       0.817
origin         1.4497      0.312      4.647      0.000       0.835       2.064
==============================================================================
Omnibus:                       32.550   Durbin-Watson:                   1.808
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               61.306
Skew:                           0.646   Prob(JB):                     4.87e-14
Kurtosis:                       4.923   Cond. No.                     8.90e+04
==============================================================================
```

Appendix 4: Cross-validation Accuracy and Standard Deviation for Multiple Linear Regression Model Without One-hot Encoding

```
# calculate adj r2
n = len(X_train)
p = len(X.columns)
adj_R2 = 1- ((1-R2) * (n-1)/(n-p-1))

# use adjusted R2 as cross validation score
print("Cross-validation results using adjusted R2: \n%0.3f accuracy with a standard deviati
```

```
Cross-validation results using adjusted R2:
0.817 accuracy with a standard deviation of 0.039
```

Appendix 5: Summary Regression Results for Multiple Linear Regression Model With One-hot Encoding

OLS Regression Results

Dep. Variable:	mpg	R-squared:	0.839
Model:	OLS	Adj. R-squared:	0.832
Method:	Least Squares	F-statistic:	123.8
Date:	Wed, 06 Dec 2023	Prob (F-statistic):	5.22e-97
Time:	04:53:10	Log-Likelihood:	-695.19
No. Observations:	274	AIC:	1414.
Df Residuals:	262	BIC:	1458.
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	23.2876	0.189	123.204	0.000	22.915	23.660
displacement	0.3358	0.989	0.340	0.734	-1.611	2.283
horsepower	-2.1828	0.611	-3.573	0.000	-3.386	-0.980
weight	-4.1786	0.640	-6.524	0.000	-5.440	-2.918
acceleration	-0.0361	0.317	-0.114	0.910	-0.661	0.589
cylinders_3	-0.6554	0.217	-3.016	0.003	-1.083	-0.228
cylinders_4	0.0195	0.337	0.058	0.954	-0.643	0.682
cylinders_5	0.0612	0.194	0.315	0.753	-0.322	0.444
cylinders_6	-0.6406	0.159	-4.041	0.000	-0.953	-0.328
cylinders_8	0.6881	0.382	1.800	0.073	-0.065	1.441
year_70	-1.1264	0.107	-10.569	0.000	-1.336	-0.917
year_80	1.1264	0.107	10.569	0.000	0.917	1.336
origin_1	-0.4258	0.159	-2.675	0.008	-0.739	-0.112
origin_2	-0.0028	0.169	-0.017	0.987	-0.335	0.329
origin_3	0.5001	0.156	3.206	0.002	0.193	0.807

Omnibus:	36.852	Durbin-Watson:	1.888
Prob(Omnibus):	0.000	Jarque-Bera (JB):	60.349
Skew:	0.781	Prob(JB):	7.86e-14
Kurtosis:	4.687	Cond. No.	1.82e+16

Appendix 6: Cross-validation Accuracy and Standard Deviation for Multiple Linear Regression Model With One-hot Encoding

```
# calculate adj r2
n = len(X_train)
p = len(X.columns)
adj_R2 = 1- ((1-R2) * (n-1)/(n-p-1))
print("Using R2: %0.3f accuracy with a standard deviation of %0.3f" % (R2.mean(), R2.std()))
print("Using Adjusted R2: %0.3f accuracy with a standard deviation of %0.3f" % (adj_R2.mean
```

```
Using R2: 0.815 accuracy with a standard deviation of 0.051
Using Adjusted R2: 0.805 accuracy with a standard deviation of 0.054
```

Appendix 7: Summary Regression Results for Binary Logistic Regression Model Without One-hot Encoding

Logit Regression Results

Dep. Variable:	HighMPG	No. Observations:	274
Model:	Logit	Df Residuals:	266
Method:	MLE	Df Model:	7
Date:	Sat, 02 Dec 2023	Pseudo R-squ.:	0.6818
Time:	18:52:23	Log-Likelihood:	-52.939
converged:	True	LL-Null:	-166.35
Covariance Type:	nonrobust	LLR p-value:	2.354e-45

	coef	std err	z	P>|z|	[0.025	0.975]
const	-4.5255	0.821	-5.512	0.000	-6.135	-2.916
cylinders	2.0819	0.949	2.193	0.028	0.221	3.943
displacement	-3.2279	2.078	-1.554	0.120	-7.300	0.844
horsepower	-1.9206	1.159	-1.657	0.097	-4.192	0.351
weight	-2.9962	1.411	-2.123	0.034	-5.762	-0.231
acceleration	-0.3518	0.450	-0.782	0.434	-1.234	0.530
year	1.7617	0.324	5.430	0.000	1.126	2.398
origin	0.2183	0.283	0.770	0.441	-0.337	0.774

Appendix 8: Cross-validation Accuracy and Standard Deviation for Binary Logistic Regression Model Without One-hot Encoding

```
# Print the average accuracy across all folds
print("Average Score for logistic regression without one-hot encoding: {}%({}%)".format(round(np.mean(accuracy),3),round(np.std(accuracy),3)))
```

```
Fold 1: Accuracy: 87.5%
Fold 2: Accuracy: 92.5%
Fold 3: Accuracy: 82.051%
Fold 4: Accuracy: 79.487%
Fold 5: Accuracy: 71.795%
Fold 6: Accuracy: 89.744%
Fold 7: Accuracy: 94.872%
Fold 8: Accuracy: 94.872%
Fold 9: Accuracy: 100.0%
Fold 10: Accuracy: 100.0%
Average Score for logistic regression without one-hot encoding: 89.282%(8.718%)
```

Appendix 9: Summary Regression Results for Binary Logistic Regression Model With One-hot Encoding

Logit Regression Results

Dep. Variable:	HighMPG	No. Observations:	274
Model:	Logit	Df Residuals:	262
Method:	MLE	Df Model:	11
Date:	Tue, 05 Dec 2023	Pseudo R-squ.:	0.6889
Time:	20:32:40	Log-Likelihood:	-51.750
converged:	False	LL-Null:	-166.35
Covariance Type:	nonrobust	LLR p-value:	6.248e-43

	coef	std err	z	P>|z|	[0.025	0.975]
const	-35.7414	7.02e+04	-0.001	1.000	-1.38e+05	1.38e+05
displacement	-5.6197	2.606	-2.156	0.031	-10.728	-0.511
horsepower	-2.9733	1.189	-2.500	0.012	-5.304	-0.642
weight	-1.5989	1.605	-0.996	0.319	-4.746	1.548
acceleration	-1.0239	0.470	-2.179	0.029	-1.945	-0.103
cylinders_4	27.5393	7.02e+04	0.000	1.000	-1.38e+05	1.38e+05
cylinders_5	3.5308	1.81e+06	1.95e-06	1.000	-3.55e+06	3.55e+06
cylinders_6	30.1780	7.02e+04	0.000	1.000	-1.38e+05	1.38e+05
cylinders_8	12.4474	5.48e+05	2.27e-05	1.000	-1.07e+06	1.07e+06
year_80	3.8522	0.804	4.794	0.000	2.277	5.427
origin_2	-0.3497	0.741	-0.472	0.637	-1.802	1.102
origin_3	0.1816	0.698	0.260	0.795	-1.187	1.550

Appendix 10: Cross-validation Accuracy and Standard Deviation for Binary Logistic Regression Model With One-hot Encoding

```
# Print the average accuracy across all folds
print("Average Score for logistic regression with one-hot encoding: {}%({}%)".format(round(np.mean(accuracy),3),round(np.std(accuracy),3)))
```

```
Fold 1: Accuracy: 87.5%
Fold 2: Accuracy: 92.5%
Fold 3: Accuracy: 84.615%
Fold 4: Accuracy: 74.359%
Fold 5: Accuracy: 69.231%
Fold 6: Accuracy: 97.436%
Fold 7: Accuracy: 92.308%
Fold 8: Accuracy: 94.872%
Fold 9: Accuracy: 100.0%
Fold 10: Accuracy: 100.0%
Average Score for logistic regression with one-hot encoding: 89.282%(9.983%)
```

Appendix 11: Summary Regression Results and Accuracy Score for Multinomial Logistic Regression Model Without One-hot Encoding

MNLogit Regression Results

Dep. Variable:	mpg_cluster	No. Observations:	274
Model:	MNLogit	Df Residuals:	258
Method:	MLE	Df Model:	14
Date:	Tue, 12 Dec 2023	Pseudo R-squ.:	0.6859
Time:	00:13:13	Log-Likelihood:	-89.940
converged:	True	LL-Null:	-286.38
Covariance Type:	nonrobust	LLR p-value:	4.015e-75

mpg_cluster=1	coef	std err	z	P>|z|	[0.025	0.975]
const	10.4969	7.494	1.401	0.161	-4.190	25.184
cylinders	0.9242	0.554	1.667	0.095	-0.162	2.011
displacement	-0.0170	0.015	-1.160	0.246	-0.046	0.012
horsepower	0.0370	0.029	1.293	0.196	-0.019	0.093
weight	0.0040	0.001	2.939	0.003	0.001	0.007
acceleration	0.0799	0.191	0.418	0.676	-0.295	0.455
year	-0.3697	0.097	-3.818	0.000	-0.559	-0.180
origin	-0.6535	0.521	-1.254	0.210	-1.675	0.368

mpg_cluster=2	coef	std err	z	P>|z|	[0.025	0.975]
const	-60.0181	14.191	-4.229	0.000	-87.833	-32.203
cylinders	2.5204	0.808	3.118	0.002	0.936	4.105
displacement	-0.0567	0.029	-1.987	0.047	-0.113	-0.001
horsepower	-0.0208	0.044	-0.474	0.635	-0.107	0.065
weight	-0.0040	0.002	-1.644	0.100	-0.009	0.001
acceleration	0.1219	0.224	0.544	0.586	-0.317	0.561
year	0.7997	0.171	4.671	0.000	0.464	1.135
origin	0.9282	0.379	2.448	0.014	0.185	1.671

This model got an accuracy of 84.75% on the testing set

Appendix 12: Summary Regression Results and Accuracy Score for Multinomial Logistic Regression Model With One-hot Encoding

```
                          MNLogit Regression Results
==============================================================================
Dep. Variable:            mpg_cluster   No. Observations:                  274
Model:                        MNLogit   Df Residuals:                      250
Method:                           MLE   Df Model:                           22
Date:                Tue, 12 Dec 2023   Pseudo R-squ.:                 0.6689
Time:                        00:13:14   Log-Likelihood:                -94.829
converged:                      False   LL-Null:                       -286.38
Covariance Type:            nonrobust   LLR p-value:                1.253e-67
==============================================================================
mpg_cluster=1       coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const            -7.1933        nan        nan        nan         nan         nan
displacement      0.0013      0.014      0.087      0.931      -0.027       0.030
horsepower        0.0575      0.031      1.884      0.060      -0.002       0.117
weight            0.0015      0.001      1.093      0.274      -0.001       0.004
acceleration      0.2400      0.202      1.186      0.236      -0.157       0.637
cylinders_3      -0.2561   1.97e+07   -1.3e-08      1.000   -3.86e+07    3.86e+07
cylinders_4      -2.6369   1.95e+07  -1.35e-07      1.000   -3.82e+07    3.82e+07
cylinders_5      -2.5106   1.95e+07  -1.29e-07      1.000   -3.82e+07    3.82e+07
cylinders_6      -0.3283   1.95e+07  -1.68e-08      1.000   -3.83e+07    3.83e+07
cylinders_8      -1.4614   1.95e+07  -7.51e-08      1.000   -3.81e+07    3.81e+07
year_70          -2.7293   1.26e+07  -2.17e-07      1.000   -2.47e+07    2.47e+07
year_80          -4.4641   1.26e+07  -3.55e-07      1.000   -2.46e+07    2.46e+07
origin_1         -2.3448   1.71e+07  -1.37e-07      1.000   -3.36e+07    3.36e+07
origin_2         -1.8926   1.68e+07  -1.12e-07      1.000    -3.3e+07     3.3e+07
origin_3         -2.9560   1.65e+07  -1.79e-07      1.000   -3.24e+07    3.24e+07
------------------------------------------------------------------------------
mpg_cluster=2       coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const             8.5491        nan        nan        nan         nan         nan
displacement     -0.0857      0.034     -2.518      0.012      -0.152      -0.019
horsepower       -0.0930      0.045     -2.081      0.037      -0.181      -0.005
weight           -0.0009      0.003     -0.331      0.741      -0.006       0.004
acceleration     -0.3175      0.217     -1.466      0.143      -0.742       0.107
cylinders_3     -17.2969        nan        nan        nan         nan         nan
cylinders_4       7.7332        nan        nan        nan         nan         nan
cylinders_5      -0.2897        nan        nan        nan         nan         nan
cylinders_6      13.2484        nan        nan        nan         nan         nan
cylinders_8       5.1541        nan        nan        nan         nan         nan
year_70           2.0080   1.32e+07   1.52e-07      1.000    -2.6e+07     2.6e+07
year_80           6.5411   1.32e+07   4.94e-07      1.000    -2.6e+07     2.6e+07
origin_1          2.6619   8.66e+06   3.08e-07      1.000    -1.7e+07     1.7e+07
origin_2          1.8856   8.66e+06   2.18e-07      1.000    -1.7e+07     1.7e+07
origin_3          4.0016   8.66e+06   4.62e-07      1.000    -1.7e+07     1.7e+07
==============================================================================
This model got an accuracy of 80.51% on the testing set
/usr/local/lib/python3.10/dist-packages/statsmodels/discrete/discrete_model.py:5471: RuntimeWarning: invalid value encountered in sqrt
  bse = np.sqrt(np.diag(self.cov_params()))
```