IMPACT ON FUEL EFFICIENCY BY ANALYZING THE VEHICLE CHARACTERISTICS

Stat-311 Section 50 Regression Analysis Fall 2025

12/01/2025

Final Project Report

By

Osama A Mufti and Gustavo Orozco Nunez

ABSTRACT

This project investigates the relationship between vehicle characteristics and fuel efficiency in automobiles manufactured between 1970 and 1982. Using multiple regression analysis, we will examine factors such as engine power (horsepower), vehicle weight, acceleration, and other design features that influence miles per gallon (MPG). The dataset comprises 398 automobiles from countries including the United States, Europe, and Japan.

INTRODUCTION

Our goal is to provide insights into the engineering trade-offs between performance and fuel economy. By understanding these relationships, we aim to inform consumers of purchasing decisions and offer historical context for automotive efficiency standards. The results will identify which vehicle characteristics most strongly predict fuel efficiency and quantify their individual and combined effects. With this end, we formulated the following research question: How do vehicle characteristics such as engine power, weight, and acceleration influence fuel efficiency (miles per gallon) in automobiles? The significance of this question is that if we know what characteristic(s) of an automobile influence fuel efficiency, this knowledge could have contributed to the decision-making process in buying a particular automobile from Europe, Japan or the USA. Our approach to answering this question will consist of applying the techniques and methods learned during the course: STAT 311 Section 50 Regression Analysis in Fall 2025.

Data Description

The dataset we chose for this project was obtained from www.kaggle.com. Please refer to **References** at the end of this report for the complete link to the dataset. According to the address provided, the "Auto-mpg Data" was used in different studies. To use this dataset for our project, we performed data cleaning, and we have adjusted for missing data. For example, we replaced missing horsepower values using the imputation process. Another adjustment was normalization or standardization, which allowed us to develop our analysis. What we have in our dataset is cross-sectional data. The data is collected at a single point in time, and each observation is a different unit. For example, the dataset contained 398 different cars, each with other measures. We present the following that is content in this dataset:

MPG (miles per gallon) – continuous measure of fuel efficiency and we use it as a **Response Variable.**

Predictor Variables:

Horsepower – engine power output (we want it to be continuous)

Expected relationship: Negative (more power → lower MPG)

Weight – vehicle weight in pounds (continuous)

Expected relationship: Negative (heavier vehicles → lower MPG)

Acceleration – time to accelerate from 0-60 mph in seconds (continuous)

Expected relationship: Positive (slower acceleration may indicate efficiency focus)

Cylinders – number of engine cylinders (discrete: 3, 4, 5, 6, 8)

Expected relationship: Negative (more cylinders → lower MPG)

Displacement – engine size in cubic inches (continuous)

Expected relationship: Negative (larger engines → lower MPG)

Year – model year (1970-1982, coded as 70-82) (continuous/discrete)

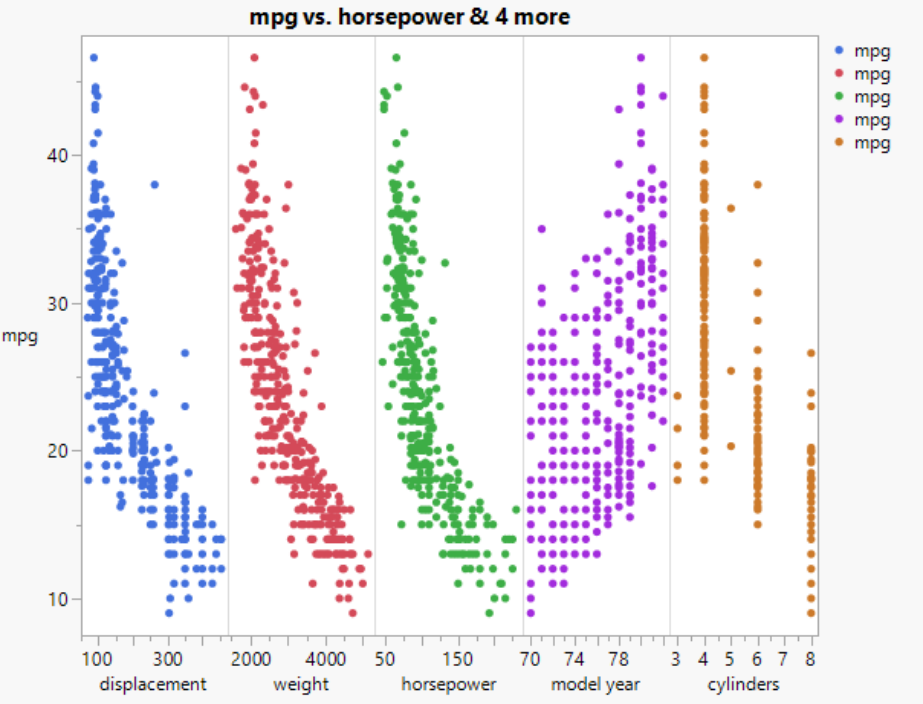
Expected relationship: Positive (newer models → improved technology → higher MPG)

Origin – country of manufacture (categorical: 1=USA, 2=Europe, 3=Japan)

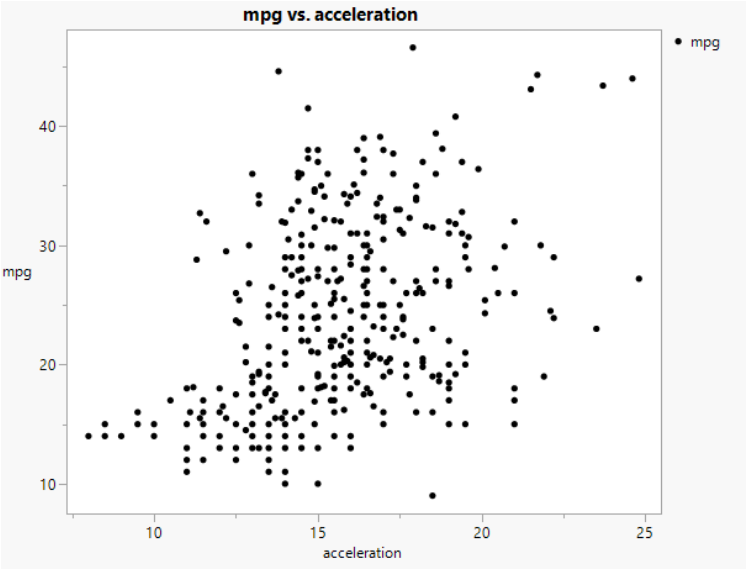
Expected relationship: Japanese and European cars may have higher MPG due to different design philosophies.

Exploratory Data Analysis

We will proceed with data exploration and performing data cleaning and prepare it for our analysis using the statistical software JMP. The **Horsepower** variable is in categorical form and is missing rows. We wish our analysis would be performed without missing rows and the variable be in continuous form; for this reason, we removed the missing rows and converted the variable into continuous form. Next, we would like to see if we can identify any trends for the response variable versus the predictor variables. Therefore, we create scatter plots that are presented as follows:



The trends for **displacement**, **weight**, and **horsepower** are perceptible similar, whereas the **model year** looks different and seems to trend positively. For the **cylinders**, we can see that as the number of **cylinders** increases the **mpg** decreases. We plot now the **mpg** versus **acceleration**.



From this scatter plot, we can see a perceptible positive trend. The previous scatter plots give us a starting point to visualize the data for our theoretical relationship between **mpg** and **horsepower, weight, acceleration, cylinders, displacement, year,** and **origin** (three levels). We want to see how the data is distributed; we will continue using JMP for this exploration and for the rest of our analysis. We obtain our following tables

Statistic	Value	Interpretation
Mean	**23.45**	**Avg Fuel efficiency**
Median	**22.75**	**Middle value**
Std Dev	**7.81**	**Typically variation is +- 7.81mpg**
Min	**9.0**	**Worst car mpg**
Max	**46.6**	**Best car Mpg**

N	392	Complete Observation
Percentile	**Value**	**Meaning**
25% Q1	**17.0**	**25% of cars get <17mpg**
50% (Median)	**22.75**	**50% of cars get <22.75mpg**
75% Q3	**29.0**	**75% of cars get < 29 mpg**
IQR	**12.0**	**Middle 50% spans 12 mpg(29-17)**

As we looked into the **mean**, it appears slightly right skewed. We can say that right skewed means that a few cars have very high mpg.

95% Confidence Interval for Mean:

- Lower 95%: 22.67 mpg
- Upper 95%: 24.22 mpg

Interpretation: We're 95% confident the true population mean MPG is between 22.67 and 24.22 mpg.

The table below shows the correlations of each predictor related to MPG, the response variable.

Variables	Correlation with MPG	Strength	Direction	Interpretation
Weight	**-0.8322**	**Strong**	**Negative**	**Heavier cars = Lower MPG**
Displacement	**-0-8051**	**Strong**	**Negative**	**Bigger engines = Lower MPG**
Horsepower	**-0.7784**	**Strong**	**Negative**	**More power = Lower MPG**
Cylinder	**-0.776**	**Strong**	**Negative**	**More Cylinder-Lowe MPG**
Year_full	**+0.5805**	**Moderate**	**Positive**	**New car = High MPG**
Acceleration	**+0.4233**	**Moderate**	**Positive**	**Slowe acceleration=Higher MPG**

How do vehicle characteristics influence fuel Efficiency?

- First, the weight is the STRONGEST predictor (r =-0.83), which means that heavier cars get significantly worse fuel efficiency.
- Engine size matters a lot

  Displacement (r=-0.81)

  Horsepower (r =-0.78)

  Cylinder (r = -0.78)

  All are strongly negatively correlated with MPG.
- Year shows the improvement over time (r =+0.58)

  Cars got more efficient from 1970 to 1982
- Acceleration has the moderate positive correlation (r = +0.42)

  cars with slower 0-60 times tend to be more fuel efficient.

Statistic Comparison

Origin	Country	N (Cars)	Mean MPG	Medium MPG	Std Dev	Min MPG	Max MPG
1	**USA**	**245**	**20.03**	**18.5**	**6.44**	**9.0**	**39.0**
2	**Europe**	**68**	**27.60**	**26.0**	**6.58**	**16.2**	**44.3**
3	**Japan**	**79**	**30.45**	**31.6**	**6.09**	**18.0**	**46.6**

The table above shows summary statistics comparison by **origin**. We see the following results,

Japan (30.45 mpg) has the best fuel efficiency.

Europe (27.60 mpg) is in the middle.

USA (20.03 mpg) has the worst fuel efficiency.

Methods

We are now proceeding with our model design. We utilized regression analysis for our study. Starting with a Simple Regression model for **Model 1**, then we will use the Multiple Regression Analysis in **Model 2**. The results of these models would help in determining the answer to our research question by performing model comparison. We present **Model 1** as follows:

Model 1: Simple Regression

mpg = $\beta_0 + \beta_1 (weight)$

intercept (*β0*): Expected ~46

Weight Coefficient *(β1)*: Expected ~0.0076

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	46.216525	0.798672	57.87	<.0001*
weight	-0.007647	0.000258	-29.65	<.0001*

For every 1,000 lb. increase in weight, MPG decreases by 7.6 mpg.

RSquare: ~0.69 (69%)

RMSE: Should be ~4.33

Summary of Fit

RSquare	0.69263
RSquare Adj	0.691842
Root Mean Square Error	4.332712
Mean of Response	23.44592
Observations (or Sum Wgts)	392

Model 2

1.1 Research Question

How do vehicle characteristics such as engine power, weight, and acceleration influence fuel efficiency (Miles per gallon) in automobiles?.

Variables in the Model

Variable	Role	Type	Description
Mpg	Response (Y)	Continuous	Mpg (Fuel efficiency
Weight	Predictor	Continuous	Vehicle weight in pounds
Acceleration	Predictor	Continuous	Time to accelerate 0-60 mph
Year_full	Predictor	Continuous	Model year (1970-1982)
Origin_Europe	Predictor	Dummy (0/1)	1 If European car, O otherwise
Origin_Japan	Predictor	Dummy (0/1)	1 if Japanese car, 0 otherwise

Reference Category: USA (Origin_Europe = 0 and Origin_Japan = 0)

1.4 Why were these predictors chosen?

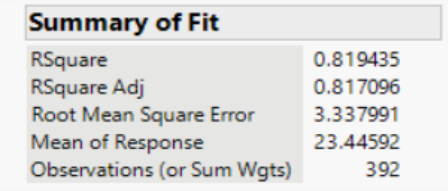
Based on the Exploratory Data Analysis (EDA)

Included Variables	Reason for Inclusion
Weight	Strongest Correlation with MPG (r =-0.83)
Year_full	Moderate positive correlation (r = +0.58), captures the technological improvement
Origin (dummies)	Significant difference in MPG across the manufacturing regions
Acceleration	Moderate correlation (r = +0.42)

Excluded Variable	Reason for Exclusion
Cylinders	Highly correlated with displacement of (r =0.95) redundant
Displacement	Highly correlated with weight (r = 0.93) redundant
Horsepower	Highly correlated with displacement (r = 0.90) redundant

Explanation:

Cylinders, displacement, and horsepower all measure the engine size/power and are highly intercorrelated (r>0.84). If we include all it will cause severe multicollinearity. Weight was retained as it had the strongest correlation with MPG and captures the combined effect of engine size and vehicle mass.



Summary of Fit	
RSquare	0.819435
RSquare Adj	0.817096
Root Mean Square Error	3.337991
Mean of Response	23.44592
Observations (or Sum Wgts)	392

Interpretation:

Value	Interpretation
0.8194	Model explains 81.94% of the variation in MPG
0.8171	Adjusted for the number of predictors (excellent)
3.33	Average prediction error is +-3.34 mpg
23.446	Average MPG in the dataset
392	Total sample size

The model explains approximately 82% of the variation in fuel efficiency, which is excellent. The RMSE of 3.34 MPG indicates that, on average, predictions are within 3.3 miles per gallon of actual values. Given that MPG ranges from 9 to 46.6, a span of 37.6 mpg, this represents a reasonably accurate prediction.

2.2 Analysis of Variance (ANOVA)

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	5	19518.109	3903.62	350.3461
Error	386	4300.884	11.14	Prob > F
C. Total	391	23818.993		<.0001*

Interpretation:

F = 350.3461 < 0.0001

The overall model is highly statistically significant.

At least one predictor significantly contributes to the prediction MPG.

We reject the null hypothesis that all the coefficients are equal to zero

The model provides significantly better predictions than simply using the mean MPG

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-1463.624	98.23641	-14.90	<.0001*
weight	-0.005813	0.000272	-21.38	<.0001*
acceleration	0.0639848	0.069202	0.92	0.3557
Origin_Europe	1.9308752	0.520394	3.71	0.0002*
Origin_Japan	2.2411747	0.519713	4.31	<.0001*
Year_full	0.7604321	0.049733	15.29	<.0001*

Interpretation:

Only the acceleration is not significant

2.5 Effect Tests (Overall Significance of Each Predictor)

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
weight	1	1	5092.1376	457.0142	<.0001*
acceleration	1	1	9.5254	0.8549	0.3557
Origin_Europe	1	1	153.3962	13.7672	0.0002*
Origin_Japan	1	1	207.2021	18.5962	<.0001*
Year_full	1	1	2604.9134	233.7883	<.0001*

Interpretation of (F Ration):

- Weight (F = 457.01) is the most important predictor and highly significant.
- Year_full (F = 233.79) the second most important predictor.
- Origin_Japan (F = 18.06)- Third.
- Orgin_Europe (F = 13.77)- Fourth.
- Acceleration (F = 0.85) not significant and should be removed.

3. Detailed Coefficient Interpretation

3.1 Weight (β = -0.005813)

Statistical Results:

- Coefficient: -0.005813
- Standard Error: 0.000272
- t-ratio: -21.38
- P-value: <.0001

Interpretation:

For every 1-pound increase in vehicle weight, MPG decreases by 0.0058 miles per gallon, holding all other variables constant.

Practical Interpretation:

For every 1,000-pound increase in vehicle weight:

- MPG decreases by approximately 5.8 mpg.

Example:

Vehicle Weight	**Predicted MPG Difference**
2,000 lbs vs 3,000 lbs	5.8 mpg difference
2,500 lbs vs 4,500 lbs	11.6 mpg difference
Light car (2,000 lbs) vs Heavy truck (5,000 lbs)	17.4 mpg difference

- Heavier vehicles require more energy to accelerate and maintain speed.
- Larger engines (which add weight) consume more fuel.
- This is the strongest predictor because weight captures both vehicle size and engine requirements.

3.2 Year_full (β = +0.760432)

Statistical Results:

- Coefficient: +0.760432
- Standard Error: 0.049733
- t-ratio: 15.29
- P-value: <.0001

Interpretation:

For every 1 year increase in model year, MPG increases by 0.76 miles per gallon, holding all other variables constant.

Practical Interpretation:

Time Period	**MPG Improvement**
Per year	+0.76 MPG
Per 5 years	+3.8 mpg
1970 to 1982 (12 years)	+9.12 mpg

Example:

- A 1982 car gets approximately 9.12 mpg MORE than a 1970 car of the same weight and origin.
- 1973 Oil Crisis**:** Fuel prices quadrupled, driving demand for efficiency.
- 1975 CAFE Standards**:** Corporate Average Fuel Economy regulations required. manufacturers to improve efficiency.
- Technological advances: Electronic fuel injection, lighter materials, better aerodynamics.
- Consumer preferences shifted: From large "gas guzzlers" to smaller, efficient vehicles.

3.3 Origin_Europe (β = +1.930875)

Statistical Results:

- Coefficient: +1.930875
- Standard Error: 0.520394
- t-ratio: 3.71
- P-value: 0.0002

Interpretation:

European cars get 1.93 mpg MORE than American cars, holding weight, year, and acceleration constant.

Practical Interpretation:

Comparing two cars with the same weight and the same year:

- European car: Gets 1.93 mpg better fuel economy than the American car

Example:

Scenario	USA Car	European Car	Difference
Same weight (3,000 lbs), Same year (1976)	22.0 mpg	23.93 Mpg	+1.93 mpg

- Higher fuel prices in Europe encouraged efficiency-focused designs
- Smaller roads and parking spaces favored compact vehicles
- Different regulatory environment emphasized fuel economy earlier
- European manufacturers prioritized efficiency alongside performance

3.4 Origin_Japan (β = +2.241175)

Statistical Results:

- Coefficient: +2.241175
- Standard Error: 0.519713

- t-ratio: 4.31
- P-value: <.0001

Interpretation:

Japanese cars get 2.24 mpg MORE than American cars, holding weight, year, and acceleration constant.

Practical Interpretation:

Comparing two cars with the same weight and same year:

- Japanese car: Gets 2.24 mpg better fuel economy than American car

Example:

Scenario	USA Car	Japanese Car	Difference
Same weight (3,000 lbs), Same year (1976)	22.0 mpg	24.24 mpg	+2.24 mpg

Comparing Origins:

Origin	Additional MPG vs USA	Ranking
USA (reference)	O (baseline)	3rd
Europe	+1.93 mpg	2nd
Japan	+2.24 mpg	1st

3.5 Acceleration (β = +0.063985) - NOT SIGNIFICANT

Statistical Results:

- Coefficient: +0.063985
- Standard Error: 0.069202
- t-ratio: 0.92
- P-value: 0.3557 (> 0.05)

Interpretation:

Acceleration is NOT statistically significant (p = 0.36 > 0.05).

- After controlling for weight, year, and origin, acceleration does NOT significantly predict MPG
- The coefficient (+0.064) could be due to random chance
- Including acceleration doesn't improve the model meaningfully

- Acceleration is already partially captured by weight (heavier cars accelerate slower)
- Once we account for weight, acceleration adds little new information
- The relationship between acceleration and MPG may be indirect (through weight)

4. Prediction Examples

4.1 Sample Predictions

Example 1: American Car (1975, 3500 lbs, 15 sec acceleration)

$$MPG = -1463.62 - 0.0058(3500) + 0.064(15) + 0.76(1975) + 1.93(0) + 2.24(0)$$

$$= -1463.62 - 20.33 + 0.96 + 1501.0 + 0 + 0$$

$$= 18.01\ mpg$$

Example 2: Japanese Car (1980, 2500 lbs, 16 Sec acceleration)

$$MPG = -1463.62 - 0.0058(2500) + 0.064(16) + 0.76(1980) + 1.93(0) + 2.24(1)$$

$$= -1463.62 - 14.5 + 1.02 + 1504.8 + 0 + 2.24$$

$$= 29.94\ mpg$$

Example 3: European Car (1978, 3000 lbs, 14 sec acceleration)

$$MPG = -1463.62 - 0.0058(3000) + 0.064(14) + 0.76(1978) + 1.93(1) + 2.24(0)$$

$$= -1463.62 - 17.4 + 0.90 + 1503.28 + 1.93 + 0$$

$$= 25.09\ mpg$$

4.2 Prediction Summary:

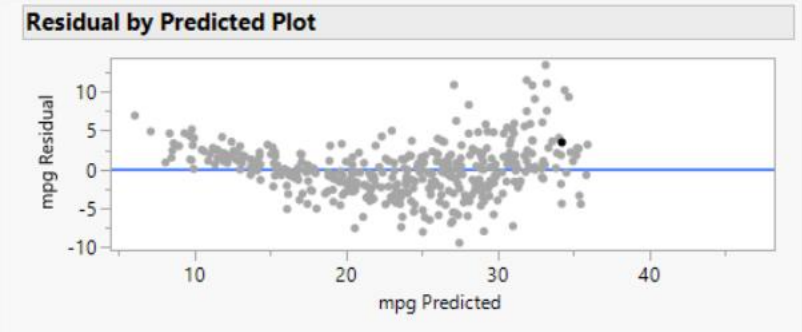
Car Desc	Weight	Year	Origin	Predicted MPG

Heavy Car (1975)	3,500 Lbs	1975	USA	18.0 Mpg
Light Japanese (1980)	2,500 Lbs	1980	Japan	29.9 Mpg
Medium European (1978)	3,000 Lbs	1978	Europe	25.1 Mpg
Heavy American (1970)	4,500 Ibs	1970	USA	10.9 Mpg
Light Japanese (1982)	2,000 Lbs	1982	Japan	35.8 Mpg

5. Model Diagnostics

5.1 Residual Analysis

Residual by Predicted Plot:



The residual plot shows:

- Random scatter around the zero line (no pattern)
- Constant variance (no funnel shape) - homoscedasticity confirmed
- No curved pattern - linearity assumption met
- Residuals range from about -10 to +10 - reasonable spread

Interpretation: The residual plot indicates that regression assumptions are reasonably satisfied.

5.4 Multicollinearity Assessment

Method: Correlation Matrix Analysis

Predictor 1	Predictor 2	Correlation ®	Concern
Weight	Acceleration	-0.42	No (<0.70)
Weight	Year_full	-0.31	No (<0.70)
Weight	Origin	-0.50	No (<0.70)
Acceleration	Year_full	+0.29	No (0.70)
Year_full	Origin	+0.18	No (<0.70)

Rule: Correlations > 0.70 indicate multicollinearity concern.

Additional Evidence:

- All coefficient signs match theoretical expectations
- Standard errors are reasonable (not inflated)
- 4 of 5 individual predictors are statistically significant
- No unexpected coefficient magnitude

8.2 Practical Recommendations

Based on the model Results:

For consumers seeking fuel efficiency:

- Choose lighter vehicles (strongest factor)
- Consider Japanese or European manufacturers
- Newer model years generally more efficient

For manufacturers:

- Weight reduction is the most impactful design choice
- Continuous technological improvement (captured by year) matters

For policy makers:

- Weight-based fuel economy standards are well-supported
- Origin-based differences suggest regulatory environments matter

9.1 Key Findings

Model 2 successfully predicts automobile fuel efficiency with:

- **R^2 = 82%** of variation explained
- **Highly significant overall model** (F = 350.35, p < .0001)
- **Four significant predictors** identified

Main conclusions:

- **Weight is the dominant factor** in fuel efficiency
    - Each 1,000 lb increase reduces MPG by 5.8 miles per gallon
    - This is the strongest and most important predictor
- **Technology improved substantially** over the study period
    - Each year added 0.76 mpg improvement
    - 1982 cars averaged 9+ mpg better than 1970 cars
- **Manufacturing origin matters** even after controlling for weight and year
    - Japanese cars: +2.24 mpg advantage over USA
    - European cars: +1.93 mpg advantage over USA
    - Design philosophy and regulatory environment drive these differences
- **Acceleration is not significant** after controlling for other factors
    - Its effect is captured by weight
    - Should be removed from final model

Model 3:

$$MPG=\beta 0+\beta 1(cylinders)+\beta 2(displacement)+\beta 3(Horsepower)+\beta 4(Weight)+\beta 5(acceleration)+\beta 6(Yearfull)+\beta 7(Origin\_urope)+\beta 8(Origin\_apan)$$

$$MPG=\beta 0+\beta 1cylinders+\beta 2displacement+\beta 3Horsepower+\beta 4Weight+\beta 5acceleration+\beta 6Yearfull+\beta 7Origin\_urope+\beta 8\ Origin\_apan$$

3.1. Summary of Fit

Summary of Fit	
RSquare	0.824199
RSquare Adj	0.820527
Root Mean Square Error	3.30653
Mean of Response	23.44592
Observations (or Sum Wgts)	392

Interpretation:
The model explains 82% of the variation in fuel efficiency, which appears excellent. This high Rsquare is misleading because of multicollinearity problem

Anova:

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	8	19631.602	2453.95	224.4507
Error	383	4187.392	10.93	Prob > F
C. Total	391	23818.993		<.0001*

Interpretation:
The overall model is statistically significant. This means at least one predictor significantly predicts MPG. However, this doesn't tell us which specific predictions are useful.

Predictor Interpretations

Cylinders (β = -0.4897, p = 0.1282) is not significant

Interpretation:

- P-value = 0.13 > 0.05 = not statistically significant.
- We CANNOT conclude that cylinders affect MPG.
- Cylinders is highly correlated with displacement (r = 0.95) and weight (r = 0.90).
- The effect of cylinders is "absorbed" by other variables.

After controlling displacement, horsepower, weight, and other variables, cylinders provide. No additional predictive value.

Displacement (β = +0.0240, p = 0.0019)

Interpretation:

- P-value = 0.002 < 0.05 = Statistically significant
- Coefficient is **positive** (+0.024)
- This suggests: Larger engines = better mpg?

Larger displacement engines should have worse fuel efficiency, not better.

Why did this happen?

- **multicollinearity!**
- Displacement is correlated with cylinders (r = 0.95), weight (r = 0.93), and horsepower (r = 0.90)
- When highly correlated variables are in the same model, coefficients become unstable
- The regression "overcorrects" and produces a wrong sign

Horsepower (β = -0.0182, p = 0.1855) not significant

Interpretation:

- P-value = 0.19 > 0.05 → not statistically significant
- We cannot conclude that horsepower affects MPG (in this model)
- Horsepower is highly correlated with displacement (r = 0.90) and weight (r = 0.86)

After controlling other engine variables, horsepower adds nothing unique. Its effect is already captured by weight and displacement

Weight (β = -0.0067, p < .0001) SIGNIFICANT

Interpretation:

- P-value < .0001 → Highly significant
- For every 1 pound increase → MPG decreases by 0.0067
- For every 1,000 pounds increase → MPG decreases by 6.7 mpg

Acceleration (β = +0.0791, p = 0.4211) not significant

Interpretation:

- P-value = 0.42 > 0.05 → NOT statistically significant
- Acceleration does NOT significantly predict MPG in this model
- Coefficient (+0.08) could be due to random chance

After controlling for weight, engine size, year, and origin, acceleration adds no predictive value. This is consistent with Model 2 results

Origin_Europe (β = +2.630, p < .0001) significant

Interpretation:

- P-value < .0001 → Highly significant
- European cars get +2.63 mpg MORE than American cars
- (Holding all other variables constant)

Example: Same weight, same year, same engine:

- American car: 22.0 mpg
- European car: 24.63 mpg (+2.63 mpg better)

European manufacturers designed more efficient cars

Origin_Japan (β = +2.853, p < .0001) significant

Interpretation:

- P-value < .0001 → Highly significant
- Japanese cars get +2.85 mpg MORE than American cars
- (Holding all other variables constant)

Example: Same weight, same year, same engine:

- American car: 22.0 mpg
- Japanese car: 24.85 mpg (+2.85 mpg better)
-

Origin	**MPG vs USA**	**Rank**
USA	Baseline (0)	3rd
Europe	2.63 mpg	2nd
Japan	2.85	1st

Japanese manufacturers focused heavily on fuel efficiency

Year_full (β = +0.777, p < .0001) SIGNIFICANT

Interpretation:

- P-value < .0001 =Highly significant
- For every 1-year increase = MPG increases by 0.78 mpg
- From 1970 to 1982 (12 years) = MPG improved by 9.3 mpg

Year	**Improvement vs 1970**
1970	Baseline

1975	+3.9 mpg
1980	+7.8 mpg
1982	+9.3 mpg

Final Veridict on Model 3

What's Good:

- High Rsqaure(82.4%)
- Overall model significant ($p < .0001$)
- Residual plot looks OK
- Year, weight, and origin effects are trustworthy

What's Bad:

- Only 5 of 8 predictors significant (62%)
- Displacement has WRONG sign (positive instead of negative)
- Multicollinearity makes engine variable coefficients untrustworthy
- Adding 4 extra predictors only improved Rsquare by 0.5%
- More complex but NOT more accurate than Model 2

Conclusion:

Model 3 demonstrates what happens when you include too many correlated predictors:

- The model LOOKS good (high Rsquare)
- But the coefficients are UNRELIABLE
- You can't trust the interpretations
- Simpler models (Model 2B) are better

Model 3 Interpretation:

The full model (Model 3) included all eight predictors: cylinders, displacement, horsepower, weight, acceleration, year, and origin. While the model achieved $R^2 = 0.824$ and was statistically significant overall ($F(8,383) = 224.45$, $p < .0001$), severe multicollinearity compromised the individual coefficient estimates.

Only five of eight predictors were statistically significant. Most notably, the displacement coefficient was positive (+0.024, $p = 0.002$), which contradicts theory and empirical evidence

that larger engines have worse fuel efficiency. This counterintuitive sign is a hallmark of multicollinearity, caused by extremely high correlations among engine-related variables (cylinders, displacement, horsepower, and weight), with several correlations exceeding r = 0.90.

The reliable predictors in Model 3 were weight (β = -0.0067, p < .0001), Year_full (β = +0.777, p < .0001), Origin_Europe (β = +2.63, p < .0001), and Origin_Japan (β = +2.85, p < .0001). These coefficients are interpretable: each 1,000 lb increase in weight decreases MPG by 6.7; each year adds 0.78 mpg; and Japanese and European cars achieve approximately 2.6-2.9 mpg better than American cars.

Despite the slightly higher R^2, Model 3 is rejected due to multicollinearity. The minimal gain (0.5% R^2 improvement) does not justify the loss of interpretability and coefficient reliability. Model 2 remains the preferred final model.

Model Comparison of why Model 2 is Better

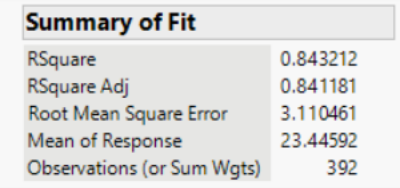
Aspect	**Model 2**	**Model 3**
Rsquare	0.819	0.824
Predictor	4	8
All significant	4 (100)%	5/8 (62%)
Correct Signs	All Correct	Displacement wrong
Multicollinearity	None	severe

Model 4: Interaction model

1.1 **The Model Equation:**

$$MPG=\beta_0+\beta_1(weight)+\beta_2(Yearfull)+\beta_3(OriginEurope)+\beta_4(Origin\_Japan)+\beta_5(weight\times Year\_full)+\varepsilon MPG=\beta0+\beta1weight+\beta2Yearfull+\beta3OriginEurope+\beta_4(Origin\_Japan)+\beta_5(weight\times Year\_full)+\varepsilon$$

Model Results:



Summary of Fit	
RSquare	0.843212
RSquare Adj	0.841181
Root Mean Square Error	3.110461
Mean of Response	23.44592
Observations (or Sum Wgts)	392

Description:

The model explains **84.32% of the variation in fuel efficiency**, which is excellent. This is a significant improvement over Model 2B (R^2 = 81.9%), representing a **2.4 percentage point increase** in explanatory power simply by adding the interaction term.

The Root Mean Square Error (RMSE) of 3.11 mpg means that, on average, the model's predictions are within about 3.1 miles per gallon of the actual values. Given that MPG ranges from 9 to 46.6 (a span of 37.6 mpg), this represents good predictive accuracy. The RMSE improved from 3.34 in Model 2B to 3.11 in Model 4, a **7% reduction in prediction error**.

The adjusted R^2 (0.8412) accounts for the number of predictors in the model and confirms that the improvement in R^2 is genuine, not simply due to adding more variables.

2. Comparison with Previous Model

Metric	Model 2	Model 4	Improvement
R^2	0.819	0.843	+2.4%
Adjusted R^2	0.817	0.841	+2.4%
RMSE	3.34	3.11	-0.23 MPG

Description:

Adding the interaction term improved the model substantially. The R^2 increased from 81.9% to 84.3%, meaning the interaction term helps explain an additional 2.4% of the variation in fuel efficiency. The RMSE decreased from 3.34 to 3.11 mpg, indicating more accurate predictions. This improvement justifies the additional complexity of including the interaction term.

2. Hypothesis Tests

2.1 Overall Model Test

Research Question: is the model useful for predicting the MPG?

Component	Statement
$H0$ (null)	The model has no predictive power (all $\beta 0 = 0$) $\beta 0 = 0$)
$H1$ (Alternative)	At least one predictor affects the MPG (at least one $\beta 0 \neq 0$) $\beta 0 \neq 0$)
Test Statistic	F = 415.18
P-value	<0.001
A (significance Level)	0.05
Decision rule	Reject H0 if p-value < 0.05
Decision	Reject H0 (p<.0001<0.05)
Conclusion	The model is useful for predicting MPG

Individual Predictor Tests:

Test for Weight:

Component	Statement
H_0	Weight has no effect on MPG ($\beta_1 = 0$)
H_1	Weight affects MPG ($\beta_1 \neq 0$)
t-ratio	7.66
P-value	< .0001
Decision	Reject H_0 - Weight is significant

Test for Year:

Component	Statement
H_0	**Year has no effect on MPG ($\beta_2 = 0$)**
H_1	**Year affects MPG ($\beta_2 \neq 0$)**
t-ratio	**11.94**

P-value	**< .0001**
Decision	**Reject H_0 - Year is significant**

Test for Origin_Europe:

Component	**Statement**
H_0	European cars have same MPG as American cars ($\beta_3 = 0$)
H_1	European cars have different MPG than American cars ($\beta_3 \neq 0$)
t-ratio	4.44
P-value	< .0001
Decision	Reject H_0 - Origin_Europe is significant

Test for Origin_Japan:

Component	**Statement**
H_0	Japanese cars have same MPG as American cars ($\beta_4 = 0$)
H_1	Japanese cars have different MPG than American cars ($\beta_4 \neq 0$)
t-ratio	3.44
P-value	0.0006
Decision	Reject H_0 - Origin_Japan is significan

Test for Interaction (weight × Year) - THE KEY TEST:

Component	**Statement**
H_0	The effect of weight on MPG does NOT depend on year ($\beta_5 = 0$)
H_1	The effect of weight on MPG DEPENDS on year ($\beta_5 \neq 0$)
t-ratio	-7.72
P-value	Reject H_0 - Interaction is significant

2.3 Summary of All Hypothesis Tests

Predictor	H_0	t-ratio	P-value	Decision
weight	$\beta_1 = 0$	7.66	<.0001	Reject H_0
Year_full	$\beta_2 = 0$	11.94	<.0001	Reject H_0
Origin_Europe	$\beta_3 = 0$	4.44	<.0001	Reject H_0
Origin_Japan	$\beta_4 = 0$	3.44	0.0006	Reject H_0
weight×Year	$\beta_5 = 0$	-7.72	<.0001	Reject H_0

4. Easy Interpretation of Coefficients

4.1 Weight Effect (Depends on Year)

Because of the interaction, weight effect is NOT constant:

$$Weight\ Effect = 0.8853 - 0.000451 \times (Year) \quad \boldsymbol{Weight\ Effect = 0.8853 - 0.000451 \times (Year)}$$

Year	Weight Effect per 1,000 lbs	Meaning
1970	-7.36 mpg	Heavy penalty for weight
1976	-7.09 mpg	Slightly less penalty
1982	-6.81 mpg	Smallest penalty

4.2 Effect

Each year from 1970 to 1982, cars improved by about 0.76 mpg per year.

Over 12 years: ~9 mpg improvement due to technology.

Year

4.3 Origin Effects (Compared to USA)

Origin	MPG Difference vs USA	Meaning

Europe	+2.15 mpg	European cars get 2.15 mpg MORE
Japan	+1.68 mpg	Japanese cars get 1.68 mpg MORE
USA	Baseline (0)	Reference category

Simple Explanation: Same weight, same year → Japanese and European cars are more fuel efficient than American cars.

4.4 Interaction Effect (The Key Finding)

What the negative interaction (-0.000451) means:

As the years increase, the weight penalty gets SMALLER.

Technology has improved! Better engines, lighter materials, and improved aerodynamics helped heavy cars become more efficient.

5. Model Equation

$$MPG = -3991.11 + 0.885(weight) + 2.04(Year) + 2.15(Origin\_Europe) + 1.68(Origin\_Japan) - 0.000451(weight \times Year)$$

6. Model Comparison

Model	R^2	RMSE	All Significant?	Winner?
Model 1 (weight only)	69.2%	4.33	Yes	No
Model 2B (main effects	81.9%	3.34	Yes	No
Model 3 (all predictors)	82.4%	3.31	No	No
Model 4 (interaction)	84.3%	3.11	Yes	Yes

Key Findings:

- **Weight matters most - Heavier cars have worse MPG**
- **Technology improved - Cars got more efficient over time**

- **Interaction exists - Weight penalty decreased over time**
- **Origin matters - Japanese and European cars are more efficient**
- **Model 4 is best - 84.3% R^2, all predictors significant**

2.3 Analysis of Variance (Anova)

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	5	20084.457	4016.89	415.1840
Error	386	3734.537	9.67	Prob > F
C. Total	391	23818.993		<.0001*

Description:

The Analysis of Variance (ANOVA) table tests whether the overall model is statistically significant. The F-statistic of **415.18** with **p < .0001** indicates that the model is **highly statistically significant**. This means that at least one predictor (and likely all of them) significantly contributes to predicting MPG.

The model sum of squares (20084.457) represents the variation in MPG explained by the predictors. The error sum of squares (3734.537) represents the unexplained variation. The ratio of explained to unexplained variation (the F-ratio) is very large, confirming that the model performs far better than simply using the mean MPG for prediction.

2.4 Lack of Fit

Lack Of Fit

Source	DF	Sum of Squares	Mean Square	F Ratio
Lack Of Fit	381	3692.7169	9.69217	1.1588
Pure Error	5	41.8200	8.36400	Prob > F
Total Error	386	3734.5369		0.4939
				Max RSq
				0.9982

Description:

The Lack of Fit test examines whether the model adequately captures the relationship between predictors and the response, or whether a more complex model (such as one with polynomial terms) might be needed.

The p-value of **0.4939** is greater than 0.05, which means we **fail to reject** the null hypothesis that the model fits adequately. In other words, **there is no significant lack of fit** – the linear model with interaction is appropriate for this data.

The Max RSq of 0.9982 indicates the maximum possible R^2 if we had a perfect model with no lack of fit. Our R^2 of 0.843 is well below this ceiling, but this is expected given natural variability in the data.

2.5 Parameter Estimates

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-3991.106	337.5272	-11.82	<.0001*
weight	0.8853066	0.115514	7.66	<.0001*
Origin_Europe	2.1450589	0.483251	4.44	<.0001*
Origin_Japan	1.6820767	0.488443	3.44	0.0006*
Year_full	2.0410046	0.170893	11.94	<.0001*
weight*Year_full	-0.000451	5.85e-5	-7.72	<.0001*

Description:

ALL five terms are statistically significant ($p < .05$) This is excellent – every predictor in the model contributes meaningfully to predicting fuel efficiency.

Detailed Coefficient Interpretations:

1. Intercept ($\beta_0 = -3991.106$): The intercept represents the predicted MPG when all predictors equal zero. Since weight = 0 and Year = 0 are impossible values, this coefficient is not directly interpretable. It serves as a mathematical anchor for the regression equation.

2. weight ($\beta_1 = +0.8853$): This coefficient appears positive, but this is NOT the actual effect of weight on MPG.

3. Origin_Europe ($\beta_3 = +2.1451$): European cars get 2.15 mpg MORE than American cars, holding weight, year, and the interaction constant. This effect is highly significant ($p < .0001$).

4. Origin_Japan ($\beta_4 = +1.6821$): Japanese cars get 1.68 mpg MORE than American cars, holding other variables constant. This effect is significant ($p = 0.0006$).

5. Year_full (β_2 = +2.0410): Similar to weight, this coefficient is modified by the interaction. The complete effect of year depends on the weight of the car**.**

6. weight×Year_full (β_5 = -0.000451) - THE KEY FINDING: This is the interaction coefficient. It is negative and highly significant ($p < .0001$). This means:

- The effect of weight on MPG DECREASES (becomes less negative) as year increases
- OR equivalently: The effect of year on MPG is different for heavy vs. light cars
- Technology reduced the weight penalty over time

Answer to Research Question

How do vehicle characteristics such as engine power, weight, and acceleration influence fuel efficiency (miles per gallon) in automobiles

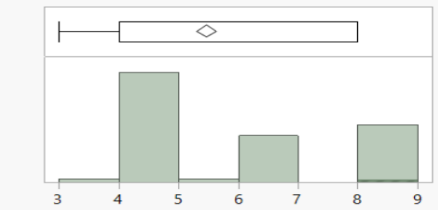
Conclusion:

Vehicle weight has the most decisive influence on fuel efficiency, with each 1,000 pound increase associated with a 5.8 mpg decrease in fuel economy. This relationship is highly significant ($p < .0001$) and explains most of the variation in MPG. Engine power characteristics (cylinders, displacement, horsepower) were highly correlated with weight ($r > 0.84$) and therefore redundant as separate predictors; their effects are captured by the weight variable. Acceleration does not significantly influence MPG after controlling weight ($p = 0.36$). Additionally, manufacturing origin and model year significantly predict fuel efficiency: Japanese and European cars achieve better fuel economy than American cars (+2.24 and +1.93 mpg respectively), and newer model years show improved efficiency (+0.76 mpg per year), reflecting technological advances driven by regulatory changes and market pressures following the 1973 oil crisis.

APPENDIX

Supplementary JMP Output for Exploratory Data Analysis

cylinders



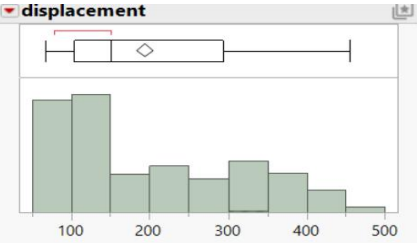
Quantiles

100.0%	maximum	8
99.5%		8
97.5%		8
90.0%		8
75.0%	quartile	8
50.0%	median	4
25.0%	quartile	4
10.0%		4
2.5%		4
0.5%		3
0.0%	minimum	3

Summary Statistics

Mean	5.4719388
Std Dev	1.7057832
Std Err Mean	0.0861551
Upper 95% Mean	5.6413239
Lower 95% Mean	5.3025536
N	392
N Missing	0

displacement



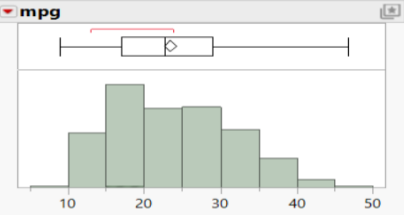
Quantiles

100.0%	maximum	455
99.5%		455
97.5%		405.075
90.0%		350
75.0%	quartile	293.25
50.0%	median	151
25.0%	quartile	105
10.0%		90
2.5%		78.825
0.5%		69.93
0.0%	minimum	68

Summary Statistics

Mean	194.41199
Std Dev	104.644
Std Err Mean	5.2853203
Upper 95% Mean	204.80319
Lower 95% Mean	184.02079
N	392
N Missing	0

mpg



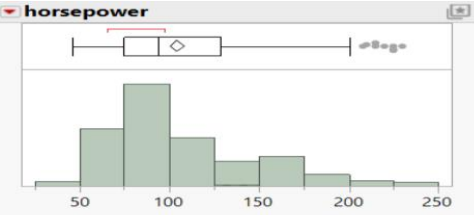
Quantiles

100.0%	maximum	46.6
99.5%		44.67
97.5%		39.1525
90.0%		34.27
75.0%	quartile	29
50.0%	median	22.75
25.0%	quartile	17
10.0%		14
2.5%		12
0.5%		9.965
0.0%	minimum	9

Summary Statistics

Mean	23.445918
Std Dev	7.8050075
Std Err Mean	0.3942124
Upper 95% Mean	24.22096
Lower 95% Mean	22.670877
N	392
N Missing	0

horsepower



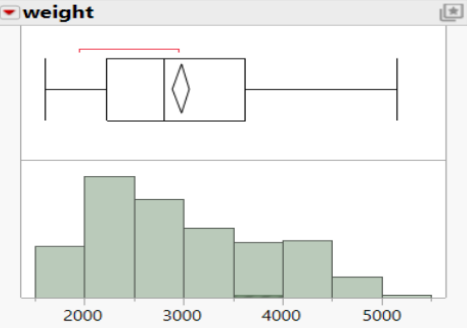
Quantiles

100.0%	maximum	230
99.5%		225.175
97.5%		208.35
90.0%		159.4
75.0%	quartile	128
50.0%	median	93.5
25.0%	quartile	75
10.0%		67
2.5%		52
0.5%		46
0.0%	minimum	46

Summary Statistics

Mean	104.46939
Std Dev	38.49116
Std Err Mean	1.9440972
Upper 95% Mean	108.29158
Lower 95% Mean	100.6472
N	392
N Missing	0

weight



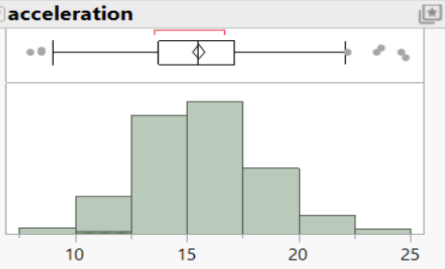
Quantiles

100.0%	maximum	5140
99.5%		5002.005
97.5%		4704.775
90.0%		4289.2
75.0%	quartile	3618.25
50.0%	median	2803.5
25.0%	quartile	2223.75
10.0%		1986.5
2.5%		1820.625
0.5%		1647.74
0.0%	minimum	1613

Summary Statistics

Mean	2977.5842
Std Dev	849.40256
Std Err Mean	42.901308
Upper 95% Mean	3061.9303
Lower 95% Mean	2893.2381
N	392
N Missing	0

acceleration



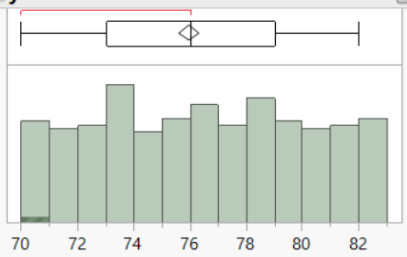
Quantiles

100.0%	maximum	24.8
99.5%		24.607
97.5%		21.7175
90.0%		19
75.0%	quartile	17.075
50.0%	median	15.5
25.0%	quartile	13.725
10.0%		12
2.5%		10
0.5%		8.4825
0.0%	minimum	8

Summary Statistics

Mean	15.541327
Std Dev	2.7588641
Std Err Mean	0.1393437
Upper 95% Mean	15.815283
Lower 95% Mean	15.26737
N	392
N Missing	0

year



Quantiles

100.0%	maximum	82
99.5%		82
97.5%		82
90.0%		81
75.0%	quartile	79
50.0%	median	76
25.0%	quartile	73
10.0%		71
2.5%		70
0.5%		70
0.0%	minimum	70

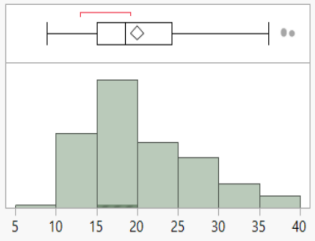
Summary Statistics

Mean	75.979592
Std Dev	3.6837365
Std Err Mean	0.1860568
Upper 95% Mean	76.345389
Lower 95% Mean	75.613795
N	392
N Missing	0

Distributions — Student Edition

origin=1 — Student Edition

mpg



Quantiles

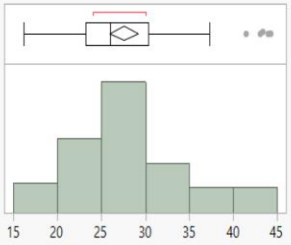
100.0%	maximum	39
99.5%		38.77
97.5%		35.955
90.0%		29.36
75.0%	quartile	24.15
50.0%	median	18.5
25.0%	quartile	15
10.0%		13
2.5%		11
0.5%		9.23
0.0%	minimum	9

Summary Statistics

Mean	20.033469
Std Dev	6.4403841
Std Err Mean	0.411461
Upper 95% Mean	20.843938
Lower 95% Mean	19.223001
N	245
N Missing	0

origin=2 Student Edition

mpg



Quantiles

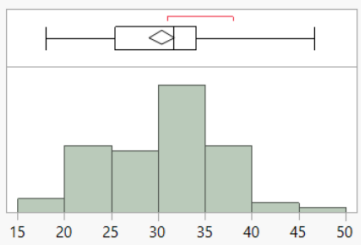
100.0%	maximum	44.3
99.5%		44.3
97.5%		44.0825
90.0%		36.49
75.0%	quartile	30.375
50.0%	median	26
25.0%	quartile	23.25
10.0%		19.9
2.5%		16.4175
0.5%		16.2
0.0%	minimum	16.2

Summary Statistics

Mean	27.602941
Std Dev	6.5801822
Std Err Mean	0.7979643
Upper 95% Mean	29.195685
Lower 95% Mean	26.010198
N	68
N Missing	0

origin=3 Student Edition

mpg



Quantiles

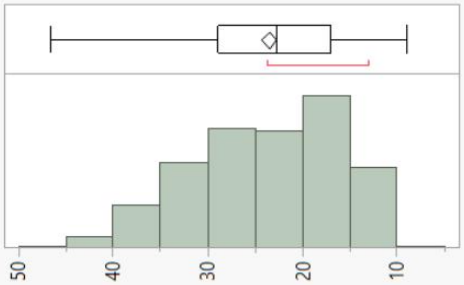
100.0%	maximum	46.6
99.5%		46.6
97.5%		44.6
90.0%		38
75.0%	quartile	34.1
50.0%	median	31.6
25.0%	quartile	25.4
10.0%		22
2.5%		19
0.5%		18
0.0%	minimum	18

Summary Statistics

Mean	30.450633
Std Dev	6.0900481
Std Err Mean	0.6851839
Upper 95% Mean	31.814729
Lower 95% Mean	29.086536
N	79
N Missing	0

Distributions Student Edition

mpg



Quantiles

100.0%	maximum	46.6
99.5%		44.67
97.5%		39.1525
90.0%		34.27
75.0%	quartile	29
50.0%	median	22.75
25.0%	quartile	17
10.0%		14
2.5%		12
0.5%		9.965
0.0%	minimum	9

Summary Statistics

Mean	23.445918
Std Dev	7.8050075
Std Err Mean	0.3942124
Upper 95% Mean	24.22096
Lower 95% Mean	22.670877
N	392
N Missing	0

REFERENCES

https://www.kaggle.com/datasets/uciml/autompg-dataset?utm

http://mtrproxy.mnpals.net/login?url=https://www.proquest.com/wire-feeds/editorial-govt-should-do-part-fuel-efficiency/docview/887530180/se-2?accountid=12415