# Title

# Automobiles Cars Dataset

## Table of contents:


1. Introduction
2. Aim
3. Business Questions
4. Data Exploration
5. Data Cleaning
6. Hypothesis Testing
7. Findings
8. Observations
9. Insights
10. Answers to the business questions
11. Conclusion


## Introduction

The automobile information dataset offers thorough information on a variety of vehicles. It contains details such as the maker, the model, the year, the engine performance, the fuel efficiency, the measurements, the safety features, and more.

## Aim

Understanding the variables that affect the fuel efficiency (mpg) of automobiles is the challenge we are confronting. Our goal is to pinpoint the factors that significantly affect fuel efficiency, such as the number of cylinders, displacement, horsepower, and weight.

## Business Questions

1. What changes have been made to the model year of various car names over time? Are there any special trends or patterns?

1. Does a vehicle's brand or name considerably affect whether it is from the USA, Japan, or Europe? Are some brands more closely linked to certain countries?

1. Is there a relationship between the model year of a car and its fuel economy (MPG)? Has MPG continually increased with newer model years?

1. Do vehicles from different nations (such as the USA, Japan, and Europe) differ in terms of fuel efficiency (MPG)? What effect does the origin have on MPG?

1. Does a vehicle's acceleration capability have an impact on its fuel economy (MPG)? Is fuel efficiency generally better or worse for cars with quicker acceleration?

# Data Exploration

Importing the dataset using read.csv and then displaying it using head(data).

In [1]:
```
getwd()
setwd('/Users/ishratshaikh/Downloads')
data <- read.csv("Automobile.csv")
head(data)
```

'/Users/ishratshaikh'

A data.frame: 6 × 9

| | name | mpg | cylinders | displacement | horsepower | weight | acceleration | model_year | origin |
|---|---|---|---|---|---|---|---|---|---|
| | <chr> | <dbl> | <int> | <dbl> | <int> | <int> | <dbl> | <int> | <chr> |
| **1** | chevrolet chevelle malibu | 18 | 8 | 307 | 130 | 3504 | 12.0 | 70 | usa |
| **2** | buick skylark 320 | 15 | 8 | 350 | 165 | 3693 | 11.5 | 70 | usa |
| **3** | plymouth satellite | 18 | 8 | 318 | 150 | 3436 | 11.0 | 70 | usa |
| **4** | amc rebel sst | 16 | 8 | 304 | 150 | 3433 | 12.0 | 70 | usa |
| **5** | ford torino | 17 | 8 | 302 | 140 | 3449 | 10.5 | 70 | usa |
| **6** | ford galaxie 500 | 15 | 8 | 429 | 198 | 4341 | 10.0 | 70 | usa |

Using str(data) to check the data types of the dataset columns. As we can see that there are only first few rows selected from the dataset and there are 398 observations.

In [2]:
```
str(data)
```

```
'data.frame':   398 obs. of  9 variables:
 $ name        : chr  "chevrolet chevelle malibu" "buick skylark 320" "plymouth satellit
e" "amc rebel sst" ...
 $ mpg         : num  18 15 18 16 17 15 14 14 14 15 ...
 $ cylinders   : int  8 8 8 8 8 8 8 8 8 8 ...
 $ displacement: num  307 350 318 304 302 429 454 440 455 390 ...
 $ horsepower  : int  130 165 150 150 140 198 220 215 225 190 ...
 $ weight      : int  3504 3693 3436 3433 3449 4341 4354 4312 4425 3850 ...
 $ acceleration: num  12 11.5 11 12 10.5 10 9 8.5 10 8.5 ...
 $ model_year  : int  70 70 70 70 70 70 70 70 70 70 ...
 $ origin      : chr  "usa" "usa" "usa" "usa" ...
```

Genetrating a summary of the dataset using summary() to get the basic statistics of the dataset.

In [3]:
```
summary(data)
```

```
     name                mpg          cylinders      displacement
 Length:398         Min.   : 9.00   Min.   :3.000   Min.   : 68.0
 Class :character   1st Qu.:17.50   1st Qu.:4.000   1st Qu.:104.2
 Mode  :character   Median :23.00   Median :4.000   Median :148.5
                    Mean   :23.51   Mean   :5.455   Mean   :193.4
                    3rd Qu.:29.00   3rd Qu.:8.000   3rd Qu.:262.0
                    Max.   :46.60   Max.   :8.000   Max.   :455.0

   horsepower        weight      acceleration     model_year
 Min.   : 46.0   Min.   :1613   Min.   : 8.00   Min.   :70.00
 1st Qu.: 75.0   1st Qu.:2224   1st Qu.:13.82   1st Qu.:73.00
 Median : 93.5   Median :2804   Median :15.50   Median :76.00
 Mean   :104.5   Mean   :2970   Mean   :15.57   Mean   :76.01
 3rd Qu.:126.0   3rd Qu.:3608   3rd Qu.:17.18   3rd Qu.:79.00
 Max.   :230.0   Max.   :5140   Max.   :24.80   Max.   :82.00
 NA's   :6
    origin
 Length:398
 Class :character
 Mode  :character
```

Using colSums(is.na()) to check if there are any missing values in the dataset and then printing the output.

Output: As we can see there are 6 missing values in the horsepower column.

In [4]:
```r
# Checking for missing values
missing_values <- colSums(is.na(data))
print(missing_values)
```

```
        name          mpg    cylinders displacement   horsepower       weight
           0            0            0            0            6            0
acceleration   model_year       origin
           0            0            0
```

Cheching for the sum of the missing values in the dataset.

In [5]:
```r
sum(is.na(data))
```

6

Creating a variable to store the unique count values from the origin column (data$origin) to create a frequency table. Next printing the frequencies for origin and visualising the frequency bar plot using barplot() with title "Frequency of Origin" x-axis as "Origin" and y-axis as "Frequency".

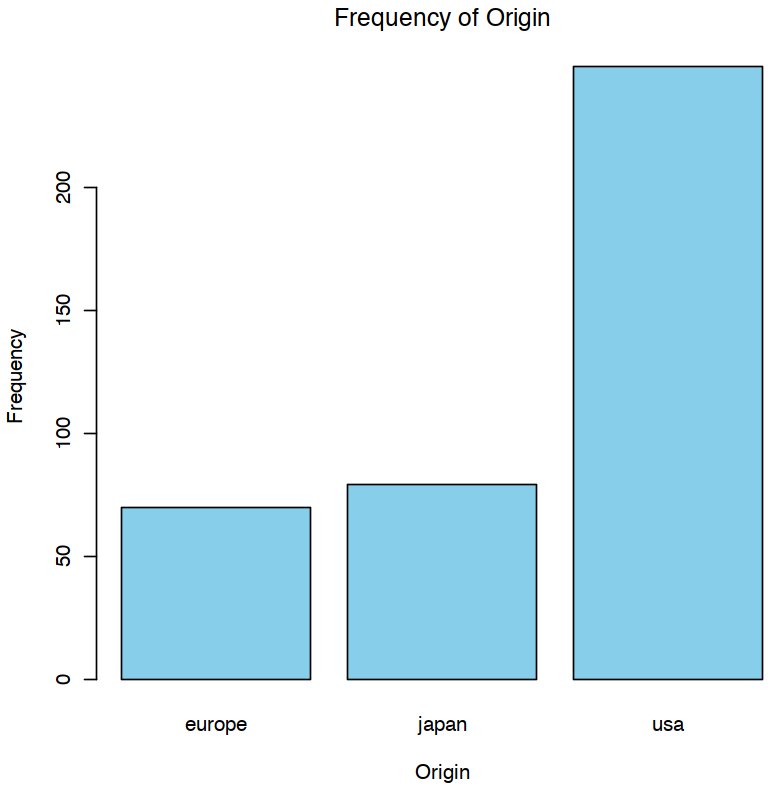
Output: As we can see that the frequency of USA is higher than other countries which means that USA have more cars as compared to other countries.

In [6]:
```r
# Creating a frequency table for 'origin'
origin_frequency <- table(data$origin)
print("Frequency Table for 'origin':")
print(origin_frequency)

# Creating a bar plot to visualize the frequencies
barplot(origin_frequency, main = "Frequency of Origin",
        xlab = "Origin", ylab = "Frequency", col = "skyblue")
```

```
[1] "Frequency Table for 'origin':"

europe  japan    usa
    70     79    249
```

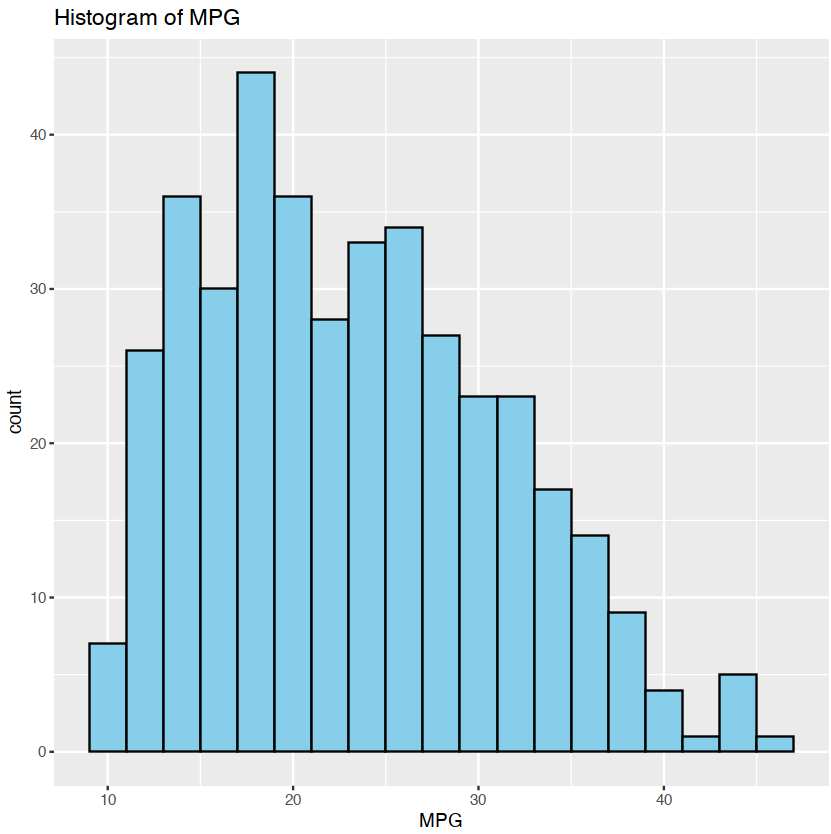



In [7]:
```
library(ggplot2)
```

Using the ggplot package to create a histogram to check the distribution of mpg. Using aes(x = mpg) to plot the chart with mpg on the x-axis. Next Adding the histogram layer using geom_histogram() and then specifying the width of the bars using binwidth. Now setting the title and x-axis of the chart using labs().

Output: As we can see that the result of highest value of MPG distribution is somewhere between 10 to 20 and the count is more than 40.

In [8]:
```
# Histogram of MPG (Miles Per Gallon)
ggplot(data, aes(x = mpg)) +
  geom_histogram(binwidth = 2, fill = "skyblue", color = "black") +
  labs(title = "Histogram of MPG", x = "MPG")
```




# Data Cleaning

Using data$horsepower[is.na(data$horsepower)] to directly store the mean of the horsepower (mean(data$horsepower, na.rm = TRUE)) with the missing values.

In [9]:
```
# Imputing missing 'horsepower' values with the mean
data$horsepower[is.na(data$horsepower)] <- mean(data$horsepower, na.rm = TRUE)
```

Checking for the null values in the dataset. Now we can see that the dataset is clean.

In [10]:
```
# Checking for missing values in the entire dataset
sum(is.na(data))
```

0

# Hypothesis Testing

## 1- Vehicle Name and Model Year Relationship

### Hypothesis:

Null Hypothesis (H0): The name or brand of a vehicle is not significantly related to its model year.

Alternative Hypothesis (H1): The name or brand of a vehicle is significantly associated with its model year.

### Statistical Analysis:

The link between the name of the vehicle (brand or model) and the model year was examined in this
an Analysis of Variance (ANOVA) test. A statistically significant correlation between the

vehicle name and the model year was aimed at. We compare means across many groups, hence ANOVA is applied.

### Result:

Df (Degrees of Freedom): The residuals have 93 degrees of freedom, compared to 304 for the Brand factor.

Sum Sq (Sum of Squares): The sum of the squared differences between the observed values and the mean value for each group (vehicle name), as well as the residuals. In this instance, the vehicle name's square sum is 4904.2, whereas the residuals' square sum is 523.7.

Mean Sq (Mean Squares): To find the mean squares, divide the sum of squares by the number of degrees of freedom. The mean square is 16.1324 for the vehicle name and 5.6314 for residuals.

F-value: The ratio between the mean squares for the vehicle name and the residuals is known as the F-statistic. The F value for the car name in this analysis is 2.8647.

Pr(>F): The p-value for the F-statistic is represented by this. The p-value here (8.667e-09) is extremely close to zero, indicating strong evidence against the null hypothesis.

In [11]:
```r
# Perform an Analysis of Variance (ANOVA) test to analyze the relationship between Vehic
model <- lm(model_year ~ name, data = data)
anova_result <- anova(model)

# Print the ANOVA table
print(anova_result)

# Check if the p-value is less than your chosen significance level (e.g., 0.05) to deter
if (anova_result$"Pr(>F)"[1] < 0.05) {
  cat("The null hypothesis is rejected. There is a significant relationship between Vehi
} else {
  cat("The null hypothesis is not rejected. There is no significant relationship between
}
```

```
Analysis of Variance Table

Response: model_year
           Df Sum Sq Mean Sq F value    Pr(>F)
name      304 4904.2 16.1324  2.8647 8.667e-09 ***
Residuals  93  523.7  5.6314
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
The null hypothesis is rejected. There is a significant relationship between Vehicle Nam
e and Model Year.
```

# 2 - Vehicle Name and Origin Relationship

### Hypothesis:

Null Hypothesis (H0): There is no significant relationship between a vehicle's name and its origin.

Alternative Hypothesis (H1): A vehicle's name is significantly related to its origin.

### Statistical Analysis:

Chi-squared test by Pearson Information: Contingency table listing the frequency of each name and origin combination. Chi-squared statistic (X-squared): 796 Degrees of Freedom (df): 608. P-value: 3.967e-07,

which is extremely near to zero.

### Result:

The Pearson's Chi-squared test's p-value is incredibly low (3.967e-07), showing significant evidence against the null hypothesis. The 'Vehicle Name' and 'Origin' variables appear to be significantly correlated, according to the low p-value.

### Conclusion

We reject the null hypothesis due to the Chi-squared test's results for independence. This indicates that there is a statistically significant association in the dataset between the names of the vehicles ("Vehicle Name") and the origins of those vehicles ("Origin"). In other words, a car's name or brand is linked to its country of manufacture. This discovery may be useful for understanding car branding and market positioning based on names and places of origin.

In [12]:
```r
# Load the necessary libraries
library(gmodels)

# Create a contingency table of 'name' and 'origin'
contingency_table <- table(data$name, data$origin)

# Perform the Chi-squared test for independence
chi_squared_test <- chisq.test(contingency_table)

# Print the Chi-squared test results
print(chi_squared_test)

# Check if the p-value is less than your chosen significance level (e.g., 0.05) to deter
if (chi_squared_test$p.value < 0.05) {
  cat("The null hypothesis is rejected. There is a significant relationship between Vehi
} else {
  cat("The null hypothesis is not rejected. There is no significant relationship between
}
```

```
Warning message in chisq.test(contingency_table):
“Chi-squared approximation may be incorrect”
        Pearson's Chi-squared test

data:  contingency_table
X-squared = 796, df = 608, p-value = 3.967e-07

The null hypothesis is rejected. There is a significant relationship between Vehicle Nam
e and Origin.
```

# 3 - Model Year and MPG Relationship

### Hypothesis:

Null Hypothesis (H0): The model year of a vehicle is not significantly related to its fuel efficiency (MPG).

Alternative Hypothesis (H1): The model year of a vehicle is significantly associated with its fuel efficiency (MPG).

### Statitical Analysis:

Residuals: Residuals are the variations between the measured MPG values and those that the regression . They shed light on how well the model fits the information.

Minimum: The model's MPG underestimation for some data points is indicated by the smallest residual, which is -12.024. 25% of the residuals fall below -5.451 in the first quartile (1Q). Median: The median residual is -0.390, which shows that the model is typically just a little bit off from the actual MPG figures. 75% of the residuals fall below 4.947 in the third quartile (3Q). Maximum: The greatest residual is 18.200, indicating that for some data points, the model overestimates MPG by this amount. Coefficients: Coefficients tell us how the independent variable (model_year) and the dependent variable (MPG) are related.

The projected MPG when the model_year is zero is represented by the intercept, which is -69.55560 (which in this case is not a useful interpretation). model_year: Model_year's coefficient is 1.22445. It means that, on average, for every year that the model_year increases, the MPG rises by 1.22445 units. Remaining standard deviation: This measures the typical gap between actual results and predictions. It is around 6.379 in this instance, showing a normal error in the model's MPG prediction.

Multiple R-squared: The R-squared value (0.3356) reveals how much of the variance in MPG the model is able to account for. In this instance, the'model_year' variable explains around 33.56% of the variation in MPG.

R-squared value is adjusted dependent on the number of predictors included in the model. Its value in this instance is 0.3339, which is a little less than R-squared.

F-statistic: The F-statistic evaluates the model's overall significance. It shows that the model is statistically significant with a value of 200 and degrees of freedom (1 and 396).

p-value: The exceptionally low p-value (2.2e-16) shows strong evidence that the null hypothesis is false. It implies that "Model Year" and "MPG" have a substantial relationship.

## Result:

The study returns an extremely low p-value, indicating that the link between "Model Year" and "MPG" is very significant.

## Conclusion

In conclusion, a car's "Model Year" has a statistically significant and favorable relationship with its fuel economy (MPG). The MPG typically rises in line with the 'Model Year'. According to this finding, automobiles with newer model years are more fuel-efficient, which is in line with the long-term trend of increasing fuel efficiency in the automotive industry.

In [13]:
```r
# Perform linear regression to analyze the relationship between Model Year and MPG
lm_model_1 <- lm(mpg ~ model_year, data = data)

# Summary of the regression model
summary(lm_model_1)

# Check if the p-value for model_year is less than your chosen significance level (e.g.,
if (summary(lm_model_1)$coefficients[2, "Pr(>|t|)"] < 0.05) {
  cat("The null hypothesis is rejected. There is a significant relationship between Mode
} else {
  cat("The null hypothesis is not rejected. There is no significant relationship between
}
```

```
Call:
lm(formula = mpg ~ model_year, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-12.024  -5.451  -0.390   4.947  18.200

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -69.55560    6.58911  -10.56   <2e-16 ***
model_year    1.22445    0.08659   14.14   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.379 on 396 degrees of freedom
Multiple R-squared:  0.3356,    Adjusted R-squared:  0.3339
F-statistic:   200 on 1 and 396 DF,  p-value: < 2.2e-16
The null hypothesis is rejected. There is a significant relationship between Model Year
and MPG.
```

# 4 - Origin and MPG Relationship

## Hypothesis:

Null Hypothesis (H0): The origin of a vehicle does not significantly affect its fuel efficiency (MPG).

Alternative Hypothesis (H1): The origin of a vehicle is significantly related to its fuel efficiency (MPG).

## The informarion provided by the ANOVA.

This ANOVA test defines whether the origin of a vehicle has a statistically significant impact on its fuel economy (MPG).

Degrees of Freedom (Df): Since there are three categories—the United States, Japan, and Europe—the origin factor has two degrees of freedom, or the number of categories minus one. There are 395 degrees of freedom for the residuals.

Sum of Squares (Sum Sq): The MPG variance attributable to the nation of origin is represented by the sum of squares (Sum Sq) of 8072.8 for the origin component. The residual sum of squares, which represents the unexplained variation, is 16179.8.

Mean Squares (Mean Sq): Mean squares are calculated by multiplying the total number of squares by the number of degrees of freedom in each square. 4036.4 is the mean square for the origin factor, while 41.0 is the mean square for the residuals.

F Value (F value): The F-statistic is equal to 98.542 when the mean square for the origin factor is divided by the mean square for the residuals.

p-value (Pr(>F)): A very small p-value, less than 2.2e-16, is associated with the F-statistic. Strong evidence is presented here to disprove the null hypothesis.

## Conclusion:

The research gives a very low p-value, indicating that the link between the nation of origin and the MPG of automobiles is very significant.

In conclusion, the country of origin has a big impact on MPG (miles per gallon). It is clear from rejecting the null hypothesis that the origin affects MPG. This suggests that vehicles from other nations typically have differing fuel efficiency, making it a key factor in MPG study and comparison.

In [14]:
```r
# Perform an Analysis of Variance (ANOVA) test to analyze the relationship between Origi
model <- lm(mpg ~ origin, data = data)
anova_result <- anova(model)

# Print the ANOVA table
print(anova_result)

# Check if the p-value is less than your chosen significance level (e.g., 0.05) to deter
if (anova_result$"Pr(>F)"[1] < 0.05) {
  cat("The null hypothesis is rejected. There is a significant relationship between Orig
} else {
  cat("The null hypothesis is not rejected. There is no significant relationship between
}
```

```
Analysis of Variance Table

Response: mpg
           Df  Sum Sq Mean Sq F value    Pr(>F)
origin      2  8072.8  4036.4  98.542 < 2.2e-16 ***
Residuals 395 16179.8    41.0
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
The null hypothesis is rejected. There is a significant relationship between Origin and
MPG.
```

# 5 - Acceleration and MPG Relationship

## Hypothesis:

Null Hypothesis (H0): Vehicle acceleration is not significantly related to its fuel efficiency (MPG).

Alternative Hypothesis (H1): Vehicle acceleration is significantly associated with its fuel efficiency (MPG).

## Statistical Analysis:

We are examining whether the relationship between acceleration and the fuel economy of automobiles (MPG) in this straightforward linear regression analysis.

## Result:

Degrees of Freedom (Df): The residuals in this study have a total of 396 degrees of freedom.

The residual standard error, which represents the standard deviation of the residuals, is 7.101.

Multiple R-squared: The coefficient of determination (R-squared) is 0.1766, which means that the predictor variable, acceleration, explains about 17.66% of the variation in MPG.

R-squared adjusted: The R-squared adjusted, which takes into account the number of predictors in the model, is 0.1746.

With 1 and 396 degrees of freedom for the model and residuals, respectively, the F-statistic has a value of 84.96.

p-value (Pr(>|t|)): The acceleration predictor variable's p-value, which is less than 2.2e-16, rejects the null hypothesis.

### Conclusion:

The research yields a very low p-value, indicating a very significant correlation between acceleration and vehicle fuel economy (MPG).

In conclusion, acceleration has a substantial impact on a car's fuel economy, and a positive coefficient of 1.1912 shows that as acceleration rises, MPG tends to rise as well. However, as evidenced by the low R-squared value of 17.66%, this association is only moderately strong. Even while acceleration has an impact on MPG, there are other factors that also have an impact on fuel efficiency.

In [15]:
```r
# Perform linear regression to analyze the relationship between Acceleration and MPG
lm_model_3 <- lm(mpg ~ acceleration, data = data)

# Summary of the regression model
summary(lm_model_3)

# Check if the p-value for acceleration is less than your chosen significance level (e.g
if (summary(lm_model_3)$coefficients[2, "Pr(>|t|)"] < 0.05) {
  cat("The null hypothesis is rejected. There is a significant relationship between Acce
} else {
  cat("The null hypothesis is not rejected. There is no significant relationship between
}
```

```
Call:
lm(formula = mpg ~ acceleration, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-18.007  -5.636  -1.242   4.758  23.192

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    4.9698     2.0432   2.432   0.0154 *
acceleration   1.1912     0.1292   9.217   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.101 on 396 degrees of freedom
Multiple R-squared:  0.1766,    Adjusted R-squared:  0.1746
F-statistic: 84.96 on 1 and 396 DF,  p-value: < 2.2e-16
The null hypothesis is rejected. There is a significant relationship between Acceleratio
n and MPG.
```

# Findings:

1. Vehicle Name and Model Year Relationship:

   a. **Finding:** The name or brand of the car and its model year have a strong connection.

1. Vehicle Name and Origin Relationship:

   b. **Finding:** There is a significant correlation between the model year of the car and its name or brand.

1. Model Year and MPG Relationship:

 A car's fuel economy (MPG) is significantly positively correlated with its model year.

1. Origin and MPG Relationship:

   d. **Finding:** A vehicle's fuel efficiency (MPG) is considerably influenced by its country of origin (e.g., the USA, Japan, or Europe).

1. Acceleration and MPG Relationship:

   e. **Finding:** Fuel efficiency (MPG) and vehicle acceleration have a significant positive correlation.

## Observations:

1. **Observation:** Different car names display differences in how their model years have evolved over time, showing a variety of developments in the automotive sector.
2. **Observation:** Various advances in the automobile industry can be seen in how different car names reflect variances in how their model years have changed through time.
3. **Observation:** Newer model years typically have greater MPGs, showing a long-term trend in the industry toward more fuel-efficient cars.
4. **Observation:** The MPG of vehicles from various sources varies, with some origins being associated with greater fuel efficiency than others.
5. **Observation:** MPG tends to climb with acceleration, indicating that faster-accelerating cars may also be more fuel-efficient.

## Insights:

1. Brand Evolution: Over time, automakers have changed the models of their vehicles, probably in response to shifting consumer tastes and market requirements.
2. Brand identity: Some car brands have close ties to particular nations, which may influence consumer views and purchasing behavior.
3. Trends in Fuel Efficiency: Environmental restrictions and consumer desire for sustainable automobiles are probably to blame for the automotive industry's steady trend of fuel efficiency improvement with newer model years.
4. Origin Variations: The manufacturing procedures rules, and technology improvements may all have an impact on the fuel efficiency of different countries' vehicles.
5. The relationship between vehicle acceleration and fuel economy suggests that manufacturers can design vehicles to maximize both acceleration and MPG.

## Answers to the business questions

1. The results of the investigation demonstrate a strong correlation between vehicle names and model years. Various model year modification may have been seen under different vehicle names. Specific

trends and patterns connected to specific vehicle names can be found with further data analysis.

1. The results of the investigation show that there is a strong correlation between the origins of vehicle names. Some brands or car names may be closely linked to particular nations of origin, indicating that customer preferences and brand identity influence origin decisions.

1. The investigation demonstrates a strong correlation between model years and fuel economy (MPG). A trend towards more fuel-efficient automobiles over time is suggested by the tendency of newer model years to show improved MPG.

1. The data shows that there is a substantial correlation between the country of origin of a car and its MPG. The fuel economy of cars from various nations varies, with some origins being linked to more fuel-efficient cars than others.

1. The investigation demonstrates a strong correlation between acceleration and fuel economy (MPG). Because fuel-efficient vehicles typically have better acceleration, manufacturers may be able to build fuel-efficient automobiles with increased acceleration.

## Conclusion:

In this examination of the "Automobiles Cars Dataset," some important connections between factors and fuel economy (MPG) were found. These discoveries have significant consequences for the automotive sector:

1. Consumer preferences and market competition are greatly influenced by vehicle brands and how they have changed over time.

2. The brand identity of a car is directly related to its place of manufacture, which affects how buyers see its performance and build quality.

3. The industry's efforts to manufacture sustainable vehicles are shown in a trend of newer model years to have higher fuel efficiency.

4. Vehicles from several nations display differences in fuel efficiency, highlighting possible areas for improvement.

5. Acceleration performance is linked to increased fuel economy, providing an opportunity for automakers to design vehicles with enhanced acceleration and fuel efficiency.

In order to satisfy changing consumer expectations and legal obligations, automakers can use these data to shape their product development, marketing, and positioning strategies.