**Impact of a number of automobile engine factors on the vehicle’s MPG**

**Ankita Ashok Jethva (1028996)**

**Venkatesh Shamala (1031744)**

**Saint Peter’s University, New Jersey, 07306**

**DS-510, Introduction to Data Science**

**Professor: Sameep Jain**

**November 19, 2023**

# **Abstract**

Analysis of automobiles mpg being affected by various engine factors

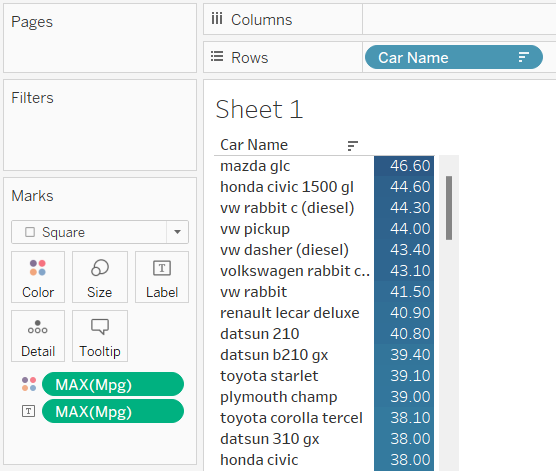
This project investigates the impact of various automobile engine factors on a vehicle's miles per gallon (mpg) using the 'auto-mpg.csv' dataset comprising information for 398 automobile models. The dataset includes details on cylinders, displacement, horsepower, weight, acceleration, model year, origin, car name, and mpg. Initial analysis and visualizations are conducted using Tableau Public, referencing materials available in week 9.

The study employs diverse visual plots and charts to interpret the dataset's insights. Simple linear regression and multiple linear regression techniques are applied to the first 300 samples to determine the relationship between mpg and independent variables. Key regression metrics—Multiple R-squared, Adjusted R-squared, and the complete Linear Regression equation—are documented and logged for all models explored.

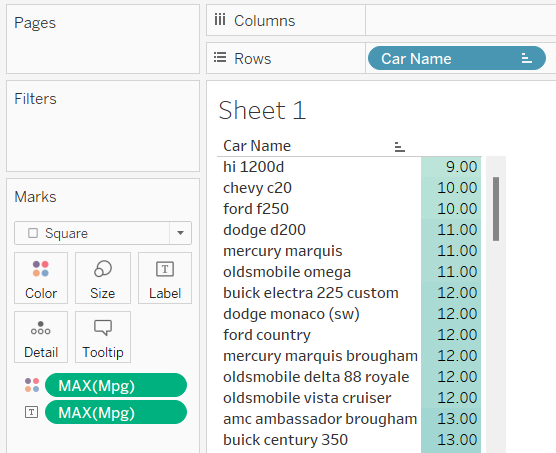
Using the remaining 98 samples, the best linear model(s) predict each automobile's mpg, comparing these predictions against the reported mpg. This comparative analysis is complemented by Residual Plots and Histograms to visually represent prediction accuracy.

The submission includes a comprehensive report outlining the research process, RStudio's code implementation, and visualizations in Word/PDF format generated through RStudio's 'Knit Document' or 'Compile Report' functions. The investigation aims to uncover the intricate relationship between automobile engine specifications and fuel efficiency, providing valuable insights for automotive industry stakeholders and enthusiasts.

# **Initial analysis and visualizations using tableau public**

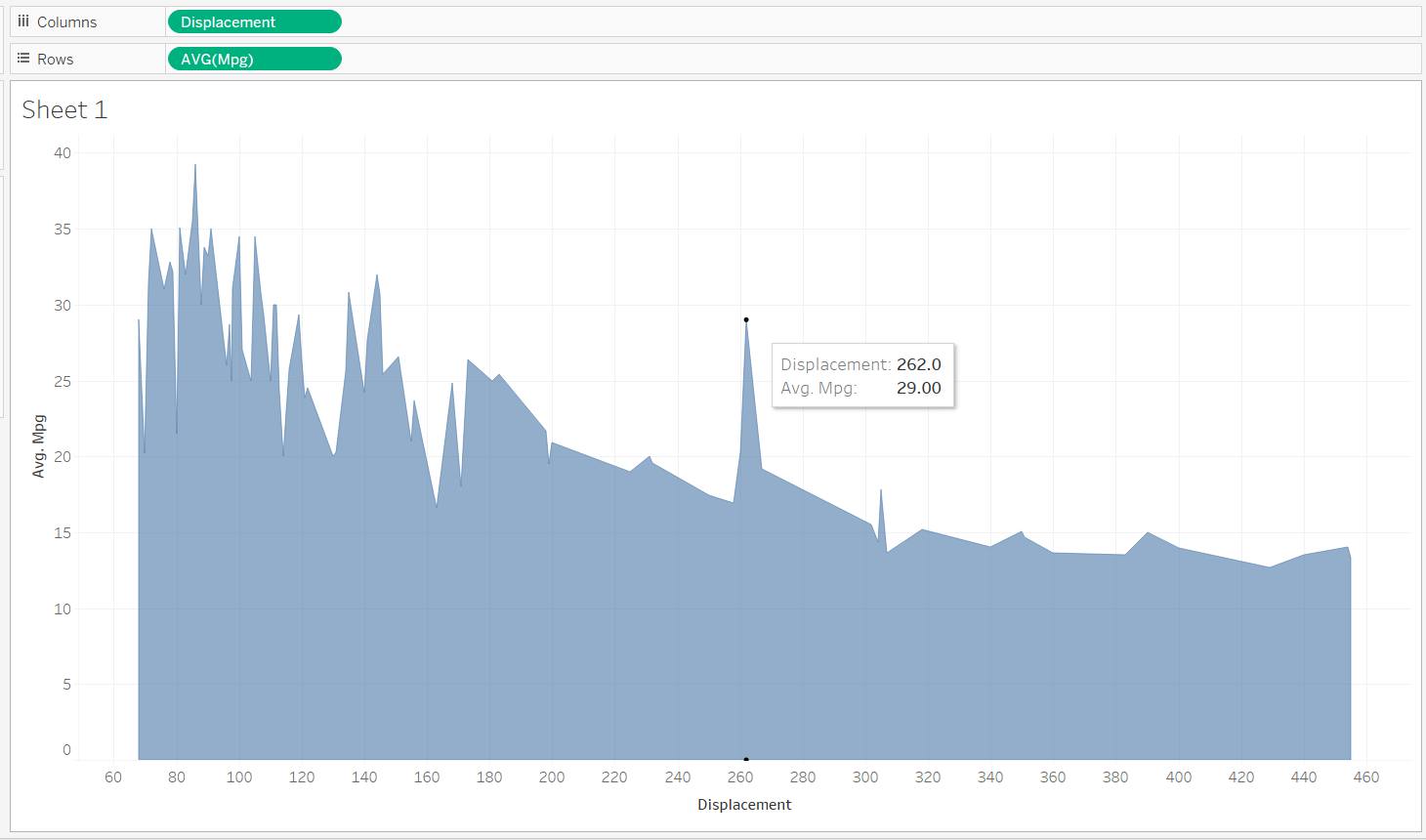


Mazda glc has the maximum mpg value among all cars which is about 46.60 miles per gallon of fuel



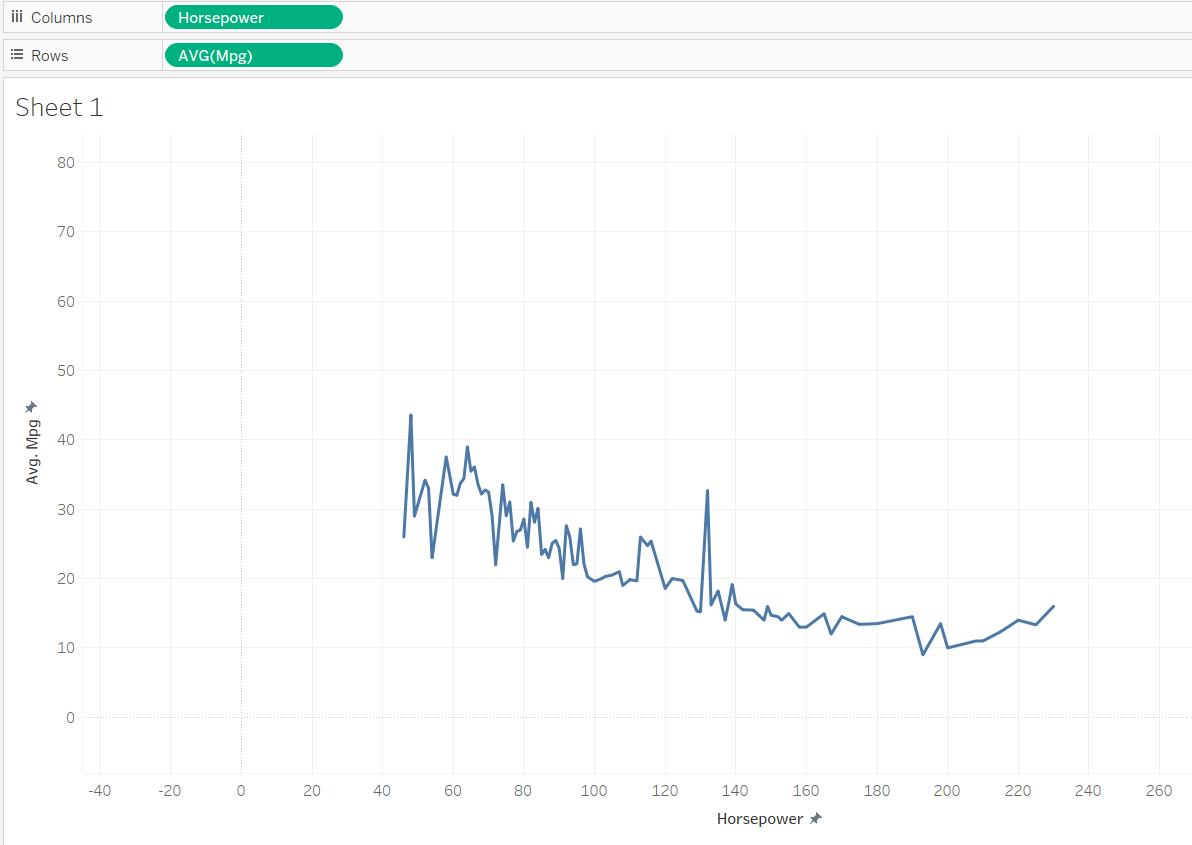
Hi 1200d has the minimum mpg value among all cars which is about 9 miles per gallon

## ***Plot Showing Displacement vs MPG***



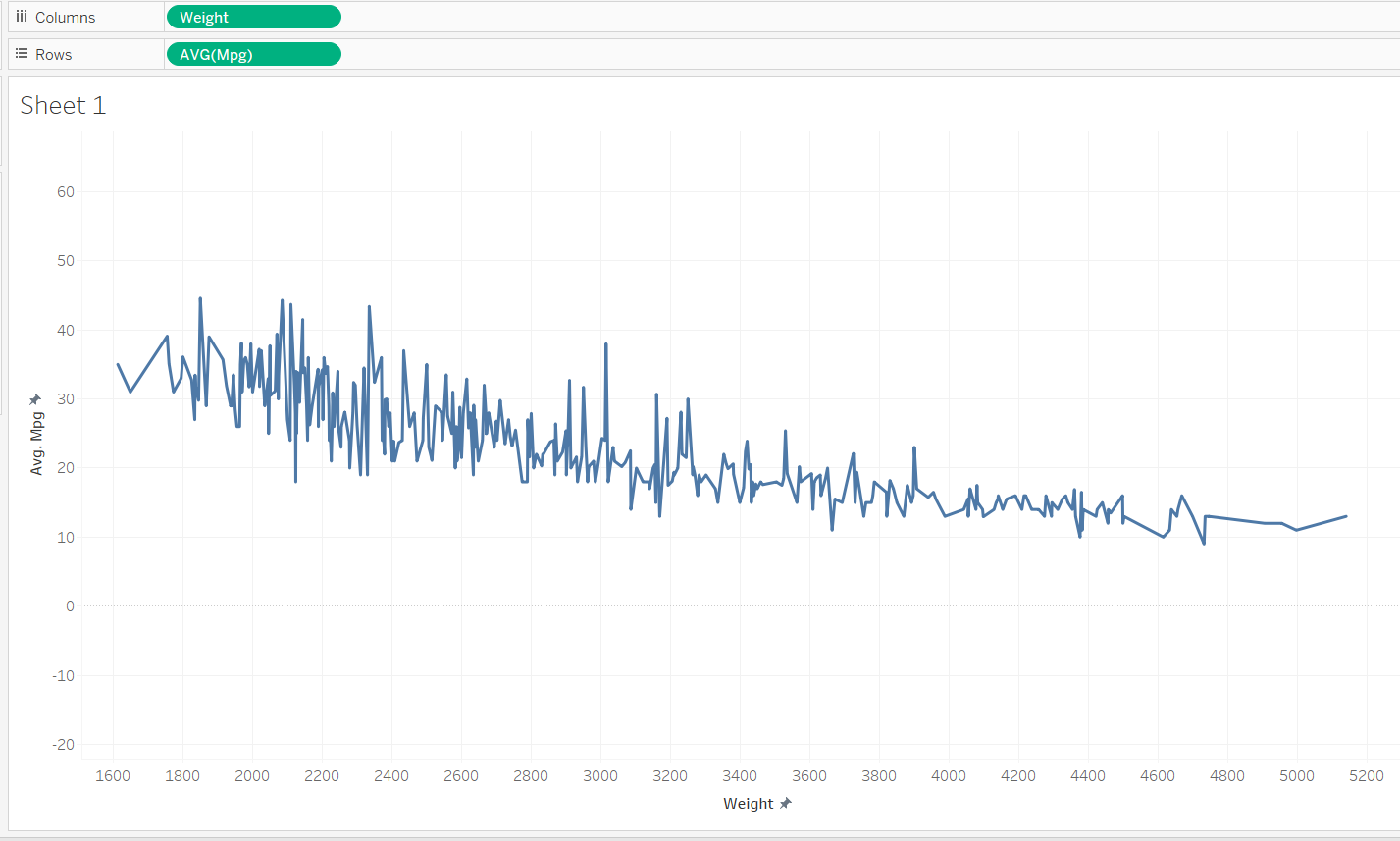
Displacement of the engine is inversely proportion to fuel economy. More the displacement more is the power and less is the efficiency. In high displacement engine the fuel required is more for engine to work so MPG is less.

## ***Plot Showing Horse Power vs MPG***



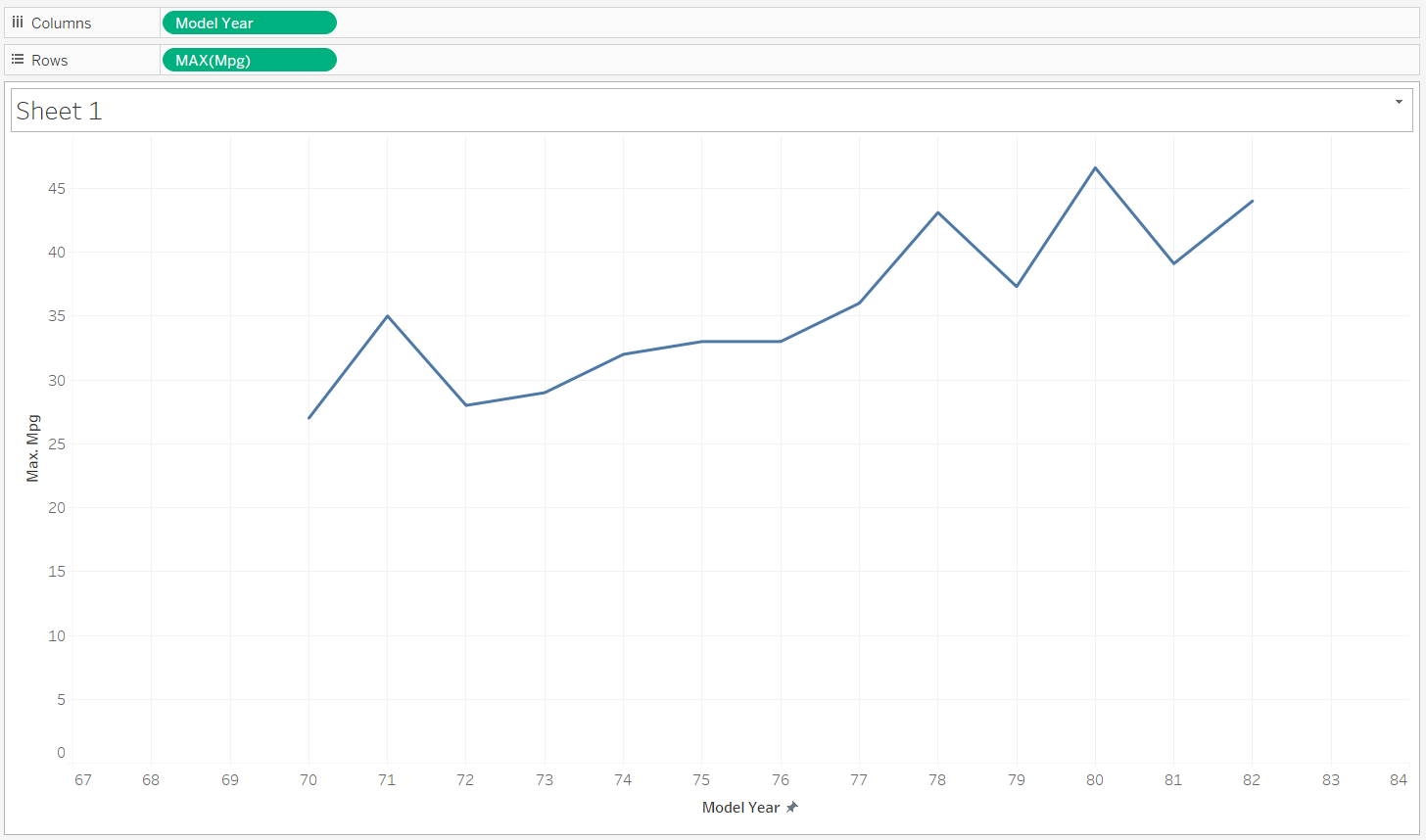
Horsepower is a major factor in a vehicle's fuel consumption. More power generally means higher fuel consumption. This means you can save money on fuel, and reduce your impact on the environment, by choosing a vehicle with no more horsepower than you need

## ***Plot showing Weight vs MPG***



This indicates that the lighter the car is, the higher or better, gas mileage it gets. So, the heavier the car, the worse the gas mileage it is.

## ***Plot showing Model year vs MPG***



This Graph shows that Automobile companies consistently prioritize the enhancement of fuel efficiency in their vehicles as a strategic goal

# **2) R Code to perform simple linear regression for first 300 samples**

# Select the CSV file interactively

file\_path <- file.choose()

# Read the first 300 samples from the selected CSV file

data <- read.csv(file\_path, nrows = 300)

# Assign weight and mpg to variables

x <- data$weight

y <- data$mpg

# Perform simple linear regression

model <- lm(y ~ x)

# Display regression summary

regression\_summary <- summary(model)

print(regression\_summary)

# Plot the data and regression line

plot(x, y, main = "Simple Linear Regression", xlab = "Weight", ylab = "MPG")

abline(model, col = "red")

# Extract coefficients for the regression equation

intercept <- coef(model)[1]

slope <- coef(model)[2]

# Display the regression equation on the plot

equation <- sprintf("Regression Equation: y = %.2f + %.2fx", intercept, slope)

text(min(x), max(y), pos = 4, labels = equation, col = "magenta")

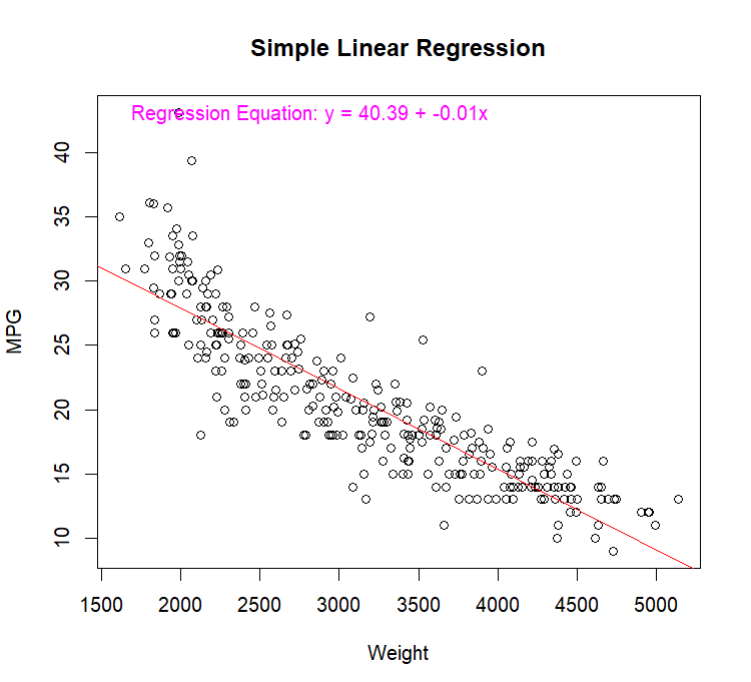
## ***Procedure***

The analysis aimed to establish a relationship between mpg and various automobile engine factors. After careful examination of the dataset, weight emerged as a parameter showcasing a discernible impact on mpg, demonstrating a consistent increase or decrease.

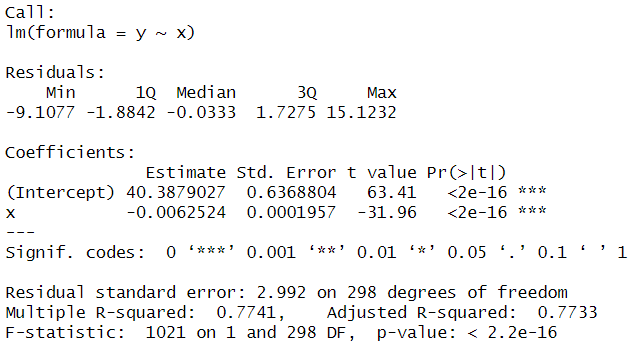
For the initial 300 samples, weight was selected as the independent variable for simple linear regression due to its observable effect on mpg. In RStudio, the above R code was executed to conduct the regression analysis

The objective was to derive insights into the impact of weight on mpg by executing a simple linear regression model. This involved fitting a linear relationship between weight and mpg, evaluating the significance and strength of this relationship through the regression summary, and visualizing the linear trend via a plot. The regression summary provides crucial information such as coefficients, p-values, R-squared, and the regression equation, offering a clear understanding of the relationship's nature and significance. Additionally, the plot showcases the regression line, accentuating the direction and magnitude of the relationship between weight and mpg.

## ***Scatterplot given by R studio***

****

## ***Printed Regression Summary***



## **Result:**

1. **Multiple R-squared = 0.7741**
2. **Adjusted R-squared = 0.7733**
3. **Complete Linear Regression equation is Y = 40.39 + -0.01x**

# **R code to perform Multiple Linear Regression for First 300 samples**

# Select the CSV file interactively

file\_path <- file.choose()

# Read the first 300 samples from the selected CSV file

data <- read.csv(file\_path, nrows = 300)

# Assign predictor variables (weight and displacement) and response variable (mpg)

predictor\_variables <- c("weight", "displacement")

response\_variable <- "mpg"

# Extract the specified variables from the data

model\_data <- data[, c(predictor\_variables, response\_variable)]

# Perform multiple linear regression

model <- lm(paste(response\_variable, "~", paste(predictor\_variables, collapse = " + ")), data = model\_data)

# Display regression summary

regression\_summary <- summary(model)

print(regression\_summary)

# Extract coefficients for the regression equation

coefficients <- coef(model)

# Display the linear regression equation

equation <- sprintf("Linear Regression Equation:\n%s = %.2f + %.2f\*weight + %.2f\*displacement", response\_variable, coefficients[1], coefficients[2], coefficients[3])

print(equation)

# Plot the data and regression line using all coefficients

plot(data$weight, data$mpg, main = "Multiple Linear Regression", xlab = "Weight", ylab = "MPG")

abline(coefficients[1], coefficients[2], col = "red", lty = 2)

# Add equation text to the plot at coordinates (3000, 400)

text(3500, 40, labels = equation, col = "blue", pos = 3)

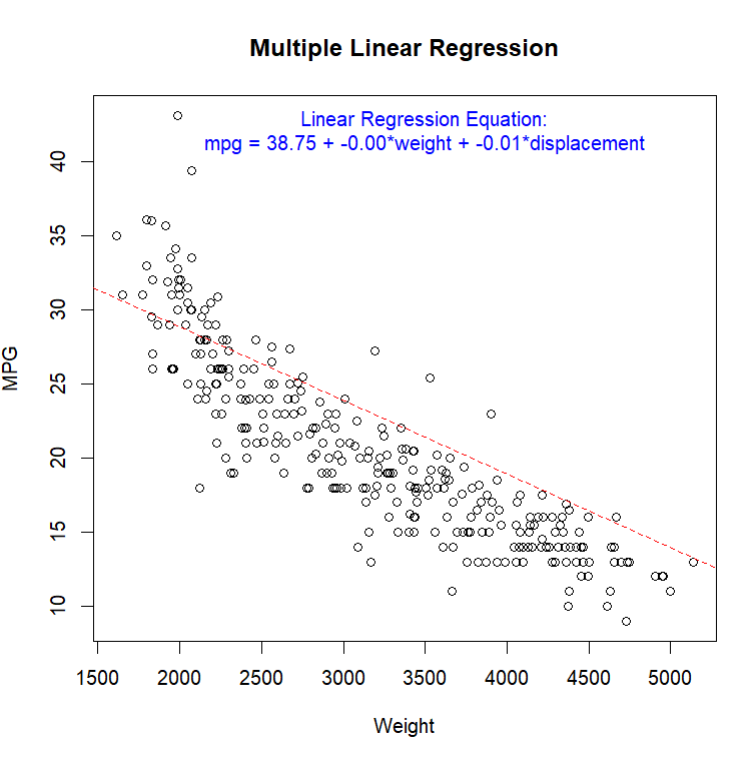
## ***Procedure***

For Multiple linear regression we considered weight and displacement emerged as a parameter showcasing a discernible impact on mpg, demonstrating a consistent increase or decrease.

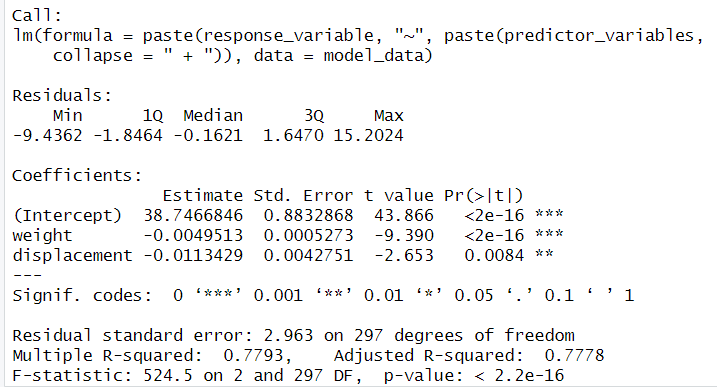
For the initial 300 samples, weight and displacement was selected as the independent variables for Multiple linear regression due to its observable effect on mpg. In RStudio, the above R code was executed to conduct the regression analysis

The regression summary provides crucial information such as coefficients, p-values, R-squared, and the regression equation, offering a clear understanding of the relationship's nature and significance. Additionally, the plot showcases the regression line, accentuating the direction and magnitude of the relationship between weight and mpg.

## ***Scatterplot given by R studio***

****

## ***Printed Regression summary***



## ***Result:***

1. **Multiple R-squared = 0.7793**
2. **Adjusted R-squared = 0.7778**
3. **Linear Regression equation is MPG = 38.75 + -0.00×weight + -0.01×displacement**

# **R code to perform Simple Linear Regression for remaining 98 samples**

# Select the CSV file interactively

file\_path <- file.choose()

# Read the data from the selected CSV file

data <- read.csv(file\_path)

# Assign weight and mpg to variables

x <- data$weight

y <- data$mpg

# Perform simple linear regression

model <- lm(y ~ x)

# Display regression summary

regression\_summary <- summary(model)

print(regression\_summary)

# Plot the data and regression line

plot(x, y, main = "Simple Linear Regression", xlab = "Weight", ylab = "MPG")

abline(model, col = "red")

# Extract coefficients for the regression equation

intercept <- coef(model)[1]

slope <- coef(model)[2]

# Display the regression equation on the plot

equation <- sprintf("Regression Equation: y = %.2f + %.2fx", intercept, slope)

text(min(x), max(y), pos = 4, labels = equation, col = "magenta")

# Calculate residuals

residuals <- residuals(model)

# Create a layout for residual plots

par(mfrow = c(2, 1))

# Residuals vs. Fitted plot

plot(model$fitted.values, residuals, main = "Residuals vs. Fitted", xlab = "Fitted Values", ylab = "Residuals", col = "blue")

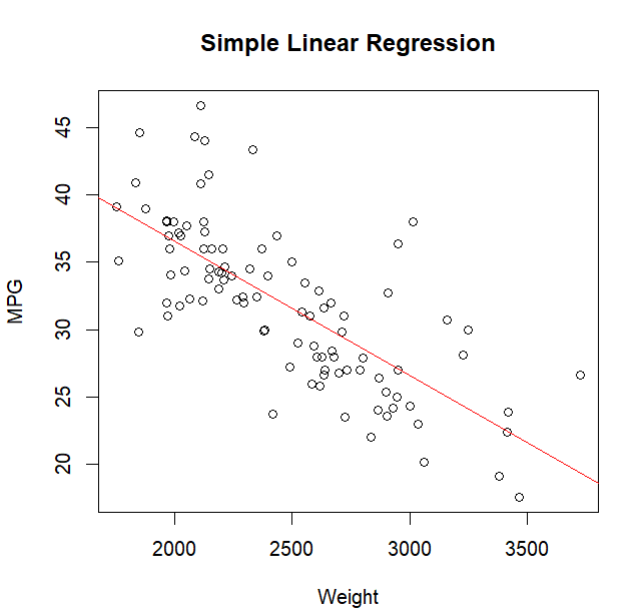
# Reset the layout

par(mfrow = c(1, 1))

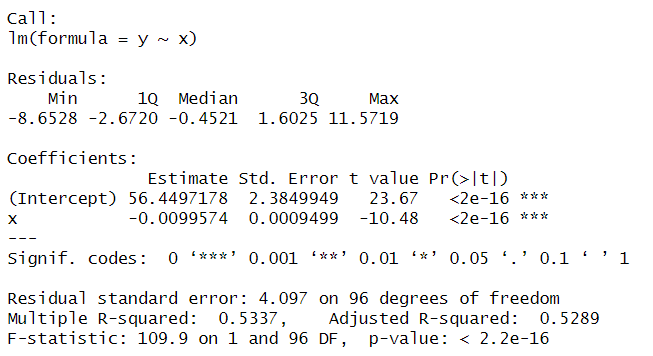
# Plot histogram of residuals

hist(residuals, main = "Histogram of Residuals", xlab = "Residuals", col = "lightblue", border = "black")

## ***Scatterplot given by R-studio for remaining 98 samples***

****

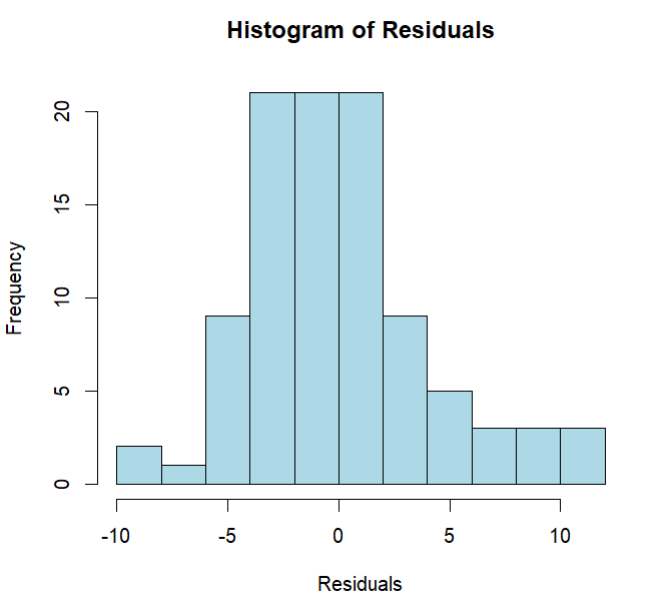
## ***Printed regression summary***

****

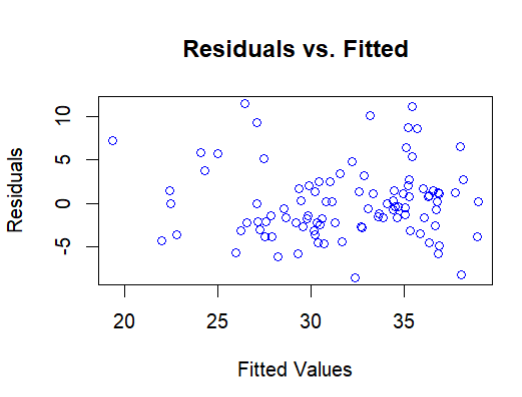
## ***Result:***

1. **Multiple R-squared = 0.5337**
2. **Adjusted R-squared = 0.5289**
3. **Linear Regression equation is y = 36.46+ -0.00x**

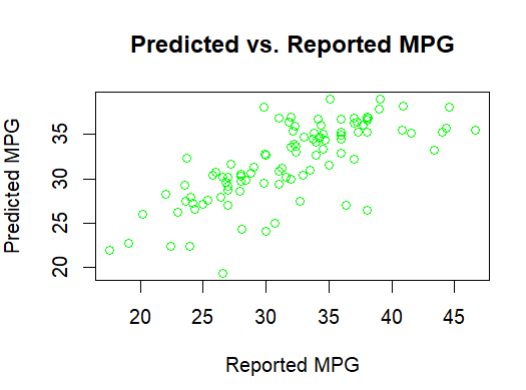
## ***Plot showing histogram of residual values***

****

## ***Plot showing Residuals vs actual fitted values***

****

## ***Plot showing Predicted values of MPG vs actual reported MPG***

****

## ***Conclusion***

Through rigorous analysis and regression modelling, this project provides insights into the significant engine factors impacting a vehicle's mpg. The conducted linear regression analyses yielded valuable relationships between mpg and specific independent variables, offering predictive capabilities for future automobile mpg predictions.

The comparison between predicted and actual mpg values for the remaining dataset demonstrates the effectiveness of the established models, with the residual plot and histogram aiding in the evaluation of model accuracy.

Overall, this project serves as a comprehensive exploration of the relationship between automobile engine factors and mpg, utilizing statistical techniques to create predictive models and validate their efficacy against real-world data.