# Regression Analysis for MPG Prediction

## 1. Reading and Exploring the Data

```
# Set working directory
setwd("C://Users//surajm//Desktop//DRocket//auto+mpg")

# Read data
mpgdata <- read.table("auto-mpg.data-original")

# Display head and tail of the data
head(mpgdata)
tail(mpgdata)

# Summary statistics of the data
summary(mpgdata)
```

### Output :

```
> head(mpgdata)
  V1 V2  V3  V4   V5   V6 V7 V8                        V9
1 18  8 307 130 3504 12.0 70  1 chevrolet chevelle malibu
2 15  8 350 165 3693 11.5 70  1         buick skylark 320
3 18  8 318 150 3436 11.0 70  1        plymouth satellite
4 16  8 304 150 3433 12.0 70  1             amc rebel sst
5 17  8 302 140 3449 10.5 70  1               ford torino
6 15  8 429 198 4341 10.0 70  1          ford galaxie 500
> tail(mpgdata)
```

```
    V1 V2  V3 V4   V5   V6 V7 V8               V9
401 27  4 151 90 2950 17.3 82  1 chevrolet camaro
402 27  4 140 86 2790 15.6 82  1  ford mustang gl
403 44  4  97 52 2130 24.6 82  2         vw pickup
404 32  4 135 84 2295 11.6 82  1    dodge rampage
405 28  4 120 79 2625 18.6 82  1      ford ranger
406 31  4 119 82 2720 19.4 82  1       chevy s-10
> # Summary statistics of the data
> summary(mpgdata)
       V1              V2              V3              V4               V5      
 Min.   : 9.00   Min.   :3.000   Min.   : 68.0   Min.   : 46.00   Min.   :1613  
 1st Qu.:17.50   1st Qu.:4.000   1st Qu.:105.0   1st Qu.: 75.75   1st Qu.:2226  
 Median :23.00   Median :4.000   Median :151.0   Median : 95.00   Median :2822  
 Mean   :23.51   Mean   :5.475   Mean   :194.8   Mean   :105.08   Mean   :2979  
 3rd Qu.:29.00   3rd Qu.:8.000   3rd Qu.:302.0   3rd Qu.:130.00   3rd Qu.:3618  
 Max.   :46.60   Max.   :8.000   Max.   :455.0   Max.   :230.00   Max.   :5140  
 NA's   :8                                       NA's   :6                      
       V6              V7              V8             V9           
 Min.   : 8.00   Min.   :70.00   Min.   :1.000   Length:406        
 1st Qu.:13.70   1st Qu.:73.00   1st Qu.:1.000   Class :character  
 Median :15.50   Median :76.00   Median :1.000   Mode  :character  
 Mean   :15.52   Mean   :75.92   Mean   :1.569                     
 3rd Qu.:17.18   3rd Qu.:79.00   3rd Qu.:2.000                     
 Max.   :24.80   Max.   :82.00   Max.   :3.000                     
```

- The code sets the working directory and reads the "auto-mpg.data-original" dataset.
- `head(mpgdata)` and `tail(mpgdata)` display the first and last few rows of the dataset.
- `summary(mpgdata)` provides summary statistics for each column.

## 2. Data Preprocessing and Initial Analysis

```
# Select relevant columns
mpgdata <- mpgdata[, 1:7]

# Summary statistics of the 'mpg' column
summary(mpgdata$mpg)

# Rename columns
colnames(mpgdata) <- c("mpg", "cylinders", "displacement", "horsepower", "weight", "acceleration", "model_year")

# Display column names
names(mpgdata)
```

**Output:**

```
> summary(mpgdata$mpg)
Length  Class   Mode 
     0   NULL   NULL 

> # Display column names
> names(mpgdata)
[1] "mpg"          "cylinders"    "displacement" "horsepower"   "weight"      
[6] "acceleration" "model_year"  
```

- The code selects the first 7 columns of the dataset and performs initial analysis on the 'mpg' column.
- Column names are renamed for better readability.

## 3. Linear Regression Model

```
# Create a subset of data with selected columns
mpgdata1 <- mpgdata[, c("mpg", "weight", "model_year")]

# Remove rows with missing values
mpgdata1 <- na.omit(mpgdata1)

# Build the linear regression model
myfirstmodel <- lm(mpg ~ ., data = mpgdata1)

# Display summary of the model
summary(myfirstmodel)
```

### Output:

```
> summary(myfirstmodel)

Call:
lm(formula = mpg ~ ., data = mpgdata1)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.8777 -2.3140 -0.1211  2.0591 14.3330 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.420e+01  3.968e+00  -3.578 0.000389 ***
weight      -6.664e-03  2.139e-04 -31.161  < 2e-16 ***
model_year   7.566e-01  4.898e-02  15.447  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.435 on 395 degrees of freedom
```

```
Multiple R-squared:  0.8079,	Adjusted R-squared:  0.8069 
F-statistic: 830.4 on 2 and 395 DF,  p-value: < 2.2e-16
```

- A subset of the data is created with 'mpg', 'weight', and 'model_year'.
- Rows with missing values are removed, and a linear regression model (`myfirstmodel`) is built using the 'lm()' function.
- The summary of the model is displayed, including coefficients and statistical measures.

## 4. Actual vs. Fitted Values and Evaluation

```
# Display the names of the coefficients
names(coef(myfirstmodel))

# Extract actual and fitted values
actual_fitted <- data.frame(actual = mpgdata1$mpg, fitted = myfirstmodel$fitted.value
s)

# Display the first 10 rows of actual and fitted values
head(actual_fitted, 10)

# Calculate Mean Absolute Percentage Error (MAPE)
mape <- mean(abs((mpgdata1$mpg - myfirstmodel$fitted.values) / mpgdata1$mpg) * 100)

# Display MAPE
mape
```

### Output:

```
> names(coef(myfirstmodel))
[1] "(Intercept)" "weight"      "model_year" 

> head(actual_fitted, 10)
   actual    fitted
1       18 15.411851
2       15 14.152377
3       18 15.864995
4       16 15.884987
5       17 15.778364
6       15  9.834182
7       14  9.747552
8       14 10.027435
9       14  9.274416
10      15 13.106148

> mape
```

```
[1] 12.09346

> # Display the length of 'mpg' and fitted values
> length(mpgdata$mpg)
[1] 406
> length(myfirstmodel$fitted.values)
[1] 398
> # Display the length of 'mpg' and fitted values
> length(mpgdata$mpg)
[1] 406
> length(myfirstmodel$fitted.values)
[1] 398
```

- The names of the coefficients in the regression model are displayed.
- Actual and fitted values are extracted and displayed for the first 10 rows.
- The Mean Absolute Percentage Error (MAPE) is calculated and presented as a measure of model performance.

## 5. Conclusion

- The linear regression model (`myfirstmodel`) suggests that both 'weight' and 'model_year' are significant predictors of 'mpg'.
- **The model's performance is evaluated using MAPE, indicating a 12.0934608179649 % average absolute percentage error.**