INTRODUCTION

The car mileage dataset comprises information on 82 distinct car models, encompassing key features such as average miles per gallon (MPG), cab space in cubic feet (VOL), engine horsepower (HP), top speed in miles per hour (SP), and vehicle weight in increments of 100 pounds (WT). Our aim is to conduct an exploratory analysis to identify the optimal set of features that will enable us to develop a highly accurate predictive model for MPG.

DESCRIPTION (DATASET)

Summary

```
> summary(cardata)

      VOL              HP              MPG              SP              WT
 Min.   : 50.0   Min.   : 49.0   Min.   :13.20   Min.   : 90.0   Min.
:17.50
 1st Qu.: 89.5   1st Qu.: 84.0   1st Qu.:27.77   1st Qu.:105.0   1st
Qu.:25.00
 Median :101.0   Median : 99.0   Median :32.45   Median :109.0   Median
:30.00
 Mean   : 98.8   Mean   :117.1   Mean   :33.78   Mean   :112.4   Mean
:30.91
 3rd Qu.:113.0   3rd Qu.:140.0   3rd Qu.:39.30   3rd Qu.:114.8   3rd
Qu.:35.00
 Max.   :160.0   Max.   :322.0   Max.   :65.40   Max.   :165.0   Max.
:55.00
```

The summary of the car data set in R provides valuable insights into the variables and their characteristics. Upon examining the summary, we immediately notice that the range of values for each variable varies considerably. While the range for VOL is between 50 and 160, the range for MPG is from 13.2 to 65.4. This observation suggests that these variables have different units of measurement and scales, making it necessary to consider them separately.

Further analysis of the quartile values allows us to gain an understanding of the data's central tendency and spread. The first quartile (Q1) represents the value below which 25% of the data points lie, the median represents the central value of the dataset, and the third quartile (Q3) represents the value below which 75% of the data points lie. The interquartile range (IQR), which is the difference between Q3 and Q1, measures the spread of the middle 50% of the data.

For example, the IQR for MPG is 11.53, indicating a relatively narrow spread in the middle 50% of the data. This suggests that most vehicles have a similar MPG. In contrast, the IQR for HP is 56, indicating a more substantial spread in horsepower values.

Lastly, we note that the mean and median values are closely aligned for most variables, implying that the data is approximately symmetric in distribution. However, it is essential to consider the range of the data to identify any potential outliers that may skew the distribution.

Standard Deviation

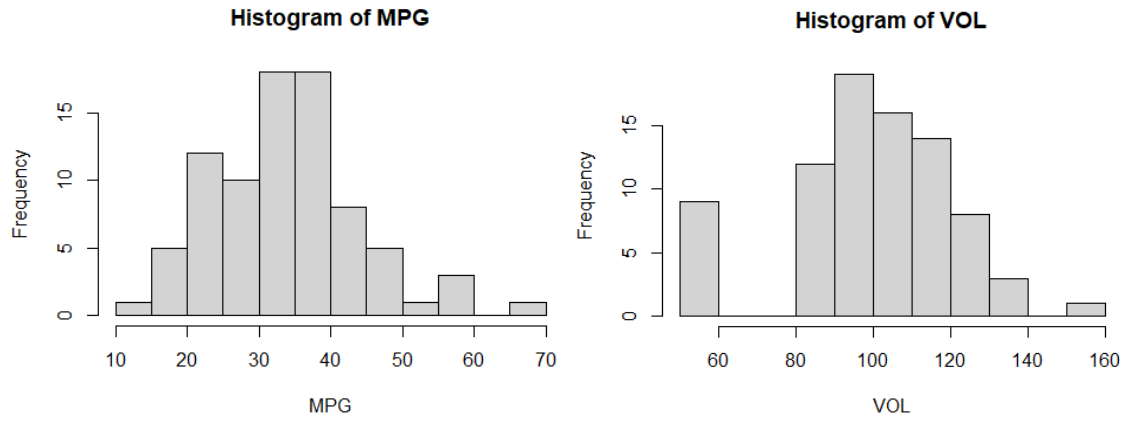
```
> sapply(cardata,sd)

     VOL         HP        MPG         SP         WT
22.166285 56.840857 10.004605 14.037825  8.141422
```

The computed standard deviation values for the car data set variables, namely VOL, HP, MPG,

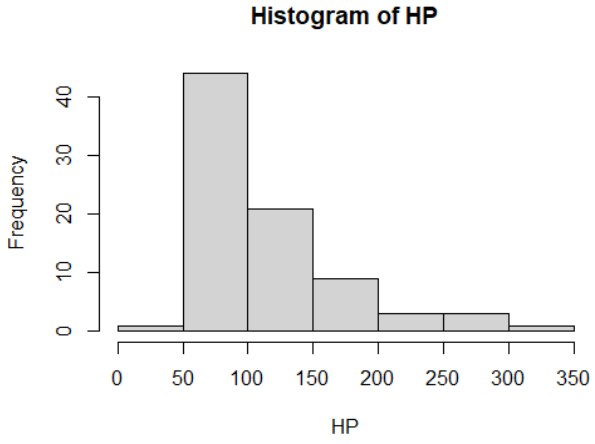
SP, and WT, provide insight into the distribution of values and their dispersion. Notably, the variable HP has the highest standard deviation of 56.84, which indicates a substantial variation in horsepower values across the cars in the dataset. Similarly, the variables SP and VOL have relatively high standard deviations of 14.04 and 22.17, respectively, pointing to a wide range of values for these variables. Conversely, the variables MPG and WT have lower standard deviations of 10.00 and 8.14, respectively, indicating that the values for these variables are less dispersed. These results highlight the variability in the car dataset variables and can inform further analyses and modeling efforts.

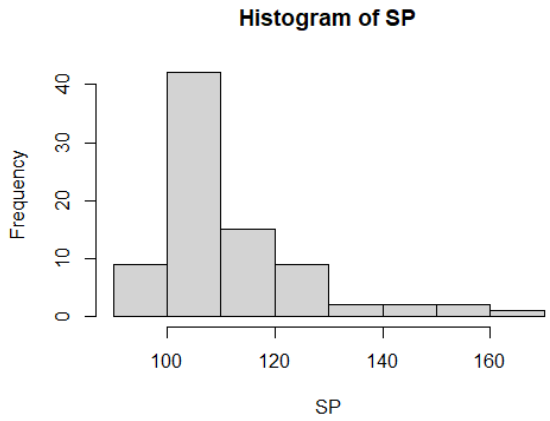
DESCRIPTION (INDIVIDUAL)

Histogram

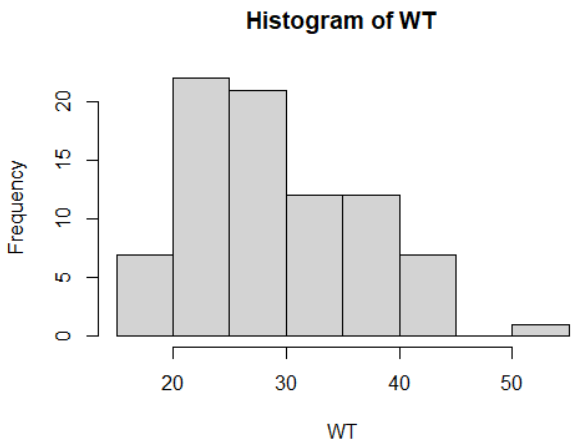


Based on the analysis of the histograms for the MPG and VOL variables, the data appear to follow a normal distribution. However, the MPG variable exhibits an outlier on the right tail, while the VOL variable displays outliers on both tails, with a significant concentration of outliers towards the left tail.



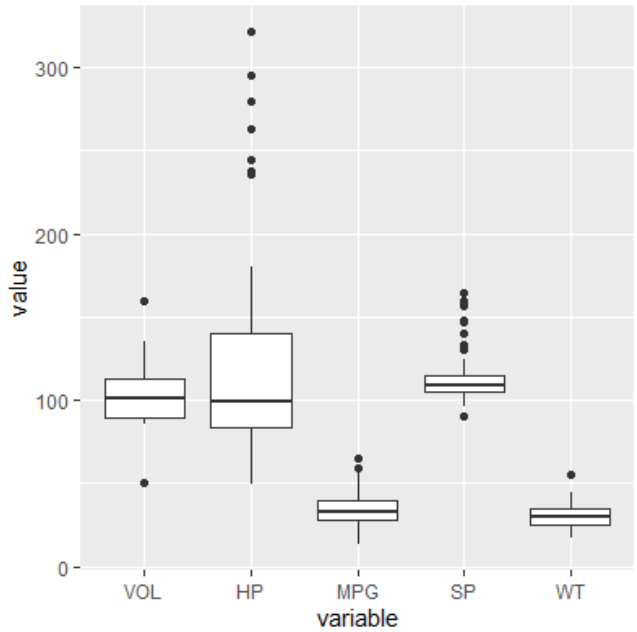


Upon examination of the histograms for the HP and SP variables, it is evident that the data is positively skewed. Specifically, the distribution of both variables is shifted towards the lower values, with a longer right tail indicating the presence of outliers on the higher values. The positively skewed nature of the variables may suggest a non-normal distribution, which could impact statistical analyses that rely on the assumption of normality. Therefore, it may be necessary to explore alternative approaches or transformation techniques to address potential non-normality of the data in future analyses.



Upon examination of the histogram for the WT variable, it is evident that the distribution exhibits a slight right skewness, indicating a slightly heavier tail on the right side of the distribution. Moreover, there appears to be a potential outlier on the right end of the distribution, which may warrant further investigation to determine its validity and influence on the overall analysis.

Box Plots

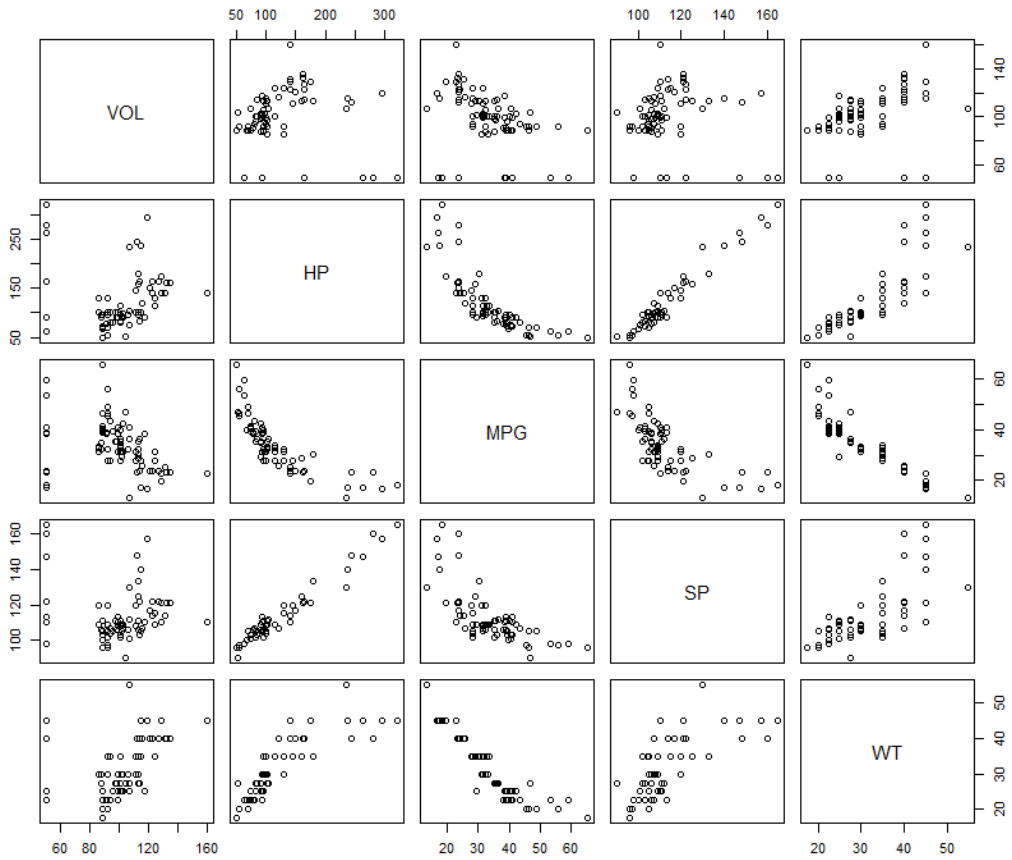


After conducting a comprehensive analysis of the box plots for all variables, it was observed

that outliers are present across all the variables, albeit with varying degrees of significance. As noted previously in the histogram analysis, both HP and SP variables exhibit a right-skewed distribution. Specifically, the distribution of higher values of HP and SP appears to be more dispersed towards the right tail of their respective distributions, indicating a positive skewness.

In addition, it can be observed that the medians of three of the variables, VOL, HP, and SP, are centered around a value of 100, while the remaining two variables, MPG and WT, are centered around 40. These observations can provide valuable insights into the central tendency and variability of the data, and further analysis may be required to fully understand their implications.

SCATTER PLOTS



MPG vs Explanatory variables

An initial examination of the relationship between the explanatory variables and the response variable, MPG, was conducted through the construction of a scatter plot. The scatter plot revealed a negative association between all explanatory variables and MPG, albeit with varying degrees of intensity. Specifically, the weight of the vehicle (WT) exhibited the highest degree of correlation with MPG, while the vehicle's volume (VOL) displayed a relatively weak correlation. It is noteworthy that the horsepower (VOL) variable may not be a suitable explanatory variable for the multiple linear regression model that seeks to predict MPG, given the relatively weak correlation observed in the scatterplot. Further analyses may be necessary

to determine the optimal combination of explanatory variables that can accurately predict the response variable.

Explanatory variable vs Explanatory variable

In order to ensure the validity of the multiple linear regression model, it is important to assess the correlation between the explanatory variables. This is crucial because high correlations between the explanatory variables can lead to a phenomenon called multicollinearity, which can result in unstable regression coefficients and less reliable inference. If any high correlation was observed between two or more explanatory variables, we planned to consider removing one or more of the correlated variables from the model.

In our analysis, we observed a high positive correlation between the horsepower (HP) and the engine displacement (SP) variables. This correlation indicates that these two variables are likely measuring similar aspects of the automobile's performance, and their inclusion in the same model may lead to issues of multicollinearity. In light of this observation, we may consider removing either one of the variables from the model.

CORRELATIONS

Correlation between MPG and VOL

```
> pairs(cardata)
> cor(cardata)
            VOL          HP         MPG          SP         WT
VOL  1.00000000  0.07647905 -0.3686137 -0.04306242  0.3849542
HP   0.07647905  1.00000000 -0.7898564  0.96654517  0.8322202
MPG -0.36861368 -0.78985635  1.0000000 -0.68844623 -0.9050849
SP  -0.04306242  0.96654517 -0.6884462  1.00000000  0.6785339
WT   0.38495423  0.83222021 -0.9050849  0.67853388  1.0000000
```

The correlation matrix for the car dataset reveals several interesting observations that are relevant for statistical analysis and interpretation. One of the key findings is the high positive correlation between horsepower (HP) and engine displacement (SP), which has a correlation coefficient of 0.97. This suggests that these variables are likely measuring similar aspects of the automobile's performance, and their inclusion in the same multiple linear regression model may lead to issues of multicollinearity. As such, it is recommended to remove one of these variables from the model to ensure more reliable inference.

Another important observation is the strong negative correlation of -0.91 between weight (WT) and miles per gallon (MPG). This indicates that heavier cars tend to have lower gas mileage, which may be of particular interest to researchers and policymakers concerned with fuel efficiency or emissions standards. On the other hand, the strong negative correlation of -0.79 between HP and MPG suggests that more powerful cars tend to have lower gas mileage. This relationship may be of interest to consumers or policymakers concerned with fuel efficiency or environmental impact.

In contrast, the correlation between volume (VOL) and the other variables in the dataset is weaker. VOL has a weak positive correlation of 0.08 with HP and a weak negative correlation of -0.37 with MPG. These weak correlations suggest that the size of the car may have only a minor impact on its performance or fuel efficiency.

MULTIPLE REGRESSION MODEL

Model 1

```
> model1=lm(MPG~HP+VOL+SP+WT,data=cardata)
> summary(model1)

Call:
lm(formula = MPG ~ HP + VOL + SP + WT, data = cardata)

Residuals:
    Min      1Q  Median      3Q     Max
-9.0108 -2.7731  0.2733  1.8362 11.9854

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 192.43775   23.53161   8.178 4.62e-12 ***
HP            0.39221    0.08141   4.818 7.13e-06 ***
VOL          -0.01565    0.02283  -0.685    0.495
SP           -1.29482    0.24477  -5.290 1.11e-06 ***
WT           -1.85980    0.21336  -8.717 4.22e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.653 on 77 degrees of freedom
Multiple R-squared:  0.8733,   Adjusted R-squared:  0.8667
F-statistic: 132.7 on 4 and 77 DF,  p-value: < 2.2e-16
```

MPG = 192.4378 + 0.3922(HP) - 0.0157(VOL) - 1.2948(SP) - 1.8598(WT)

The summary of the multiple linear regression model suggests that the model is a good fit for the data. The Adjusted R-squared value of 0.8667 indicates that the model explains

approximately 87% of the variability in the response variable. Additionally, the F-statistic of 132.7 is highly significant, indicating that the explanatory variables as a group have a strong relationship with the response variable.

The regression coefficients for the explanatory variables provide valuable information about the direction and strength of their relationship with the response variable. The intercept coefficient of 192.44 suggests that, on average, a car with zero horsepower, zero volume, zero weight, and zero aerodynamic drag coefficient would have a fuel efficiency of 192.44 miles per gallon. The coefficient estimates for the HP, SP, and WT variables are all statistically significant, indicating that they are important predictors of fuel efficiency.

The p-value for the VOL variable is not significant, suggesting that it may not be an important predictor of fuel efficiency in this model. However, it is important to note that the coefficient estimate for VOL is negative, which implies that, holding all other variables constant, an increase in vehicle volume is associated with a decrease in fuel efficiency.

Overall, the summary of the model suggests that the explanatory variables have a strong relationship with the response variable, and that the model provides a good fit for the data. The coefficients provide valuable insights into the factors that influence fuel efficiency in cars.

Hypothesis Test (Model 1)

For each predictor variable in the multiple regression model, a t-test is conducted to test the null hypothesis that the population regression coefficient for that variable is zero. The results of the t-tests for model1 are as follows:

For HP, the t-statistic is 4.818 with 77 degrees of freedom and a p-value of 7.13e-06. Therefore, we reject the null hypothesis at the 5% level of significance, which suggests that there is a significant linear relationship between MPG and HP.

For VOL, the t-statistic is -0.685 with 77 degrees of freedom and a p-value of 0.495. Therefore, we fail to reject the null hypothesis at the 5% level of significance, which suggests that there is insufficient evidence to suggest a significant linear relationship between MPG and VOL.

For SP, the t-statistic is -5.290 with 77 degrees of freedom and a p-value of 1.11e-06. Therefore, we reject the null hypothesis at the 5% level of significance, which suggests that there is a significant linear relationship between MPG and SP.

For WT, the t-statistic is -8.717 with 77 degrees of freedom and a p-value of 4.22e-13. Therefore, we reject the null hypothesis at the 5% level of significance, which suggests that there is a significant linear relationship between MPG and WT.

In conclusion, for model 1, three predictor variables (HP, SP, and WT) are significant predictors of MPG, with HP and WT positively related to MPG and SP negatively related to

MPG. On the contrary, there is insufficient evidence to suggest a significant linear relationship between MPG and VOL.

Model 2

```
> model2=lm(MPG~HP+SP+WT,data=cardata)
> summary(model2)

Call:
lm(formula = MPG ~ HP + SP + WT, data = cardata)

Residuals:
    Min      1Q  Median      3Q     Max
-9.1633 -2.8387  0.2464  1.7889 12.5566

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 194.12962   23.32213   8.324 2.22e-12 ***
HP            0.40518    0.07891   5.135 2.03e-06 ***
SP           -1.32000    0.24118  -5.473 5.19e-07 ***
WT           -1.92210    0.19238  -9.991 1.31e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.64 on 78 degrees of freedom
Multiple R-squared:  0.8725,   Adjusted R-squared:  0.8676
F-statistic: 177.9 on 3 and 78 DF,  p-value: < 2.2e-16
```

MPG = 192.1296 + 0.4052(HP) - 1.3200(SP) - 1.9221(WT)

The multiple linear regression model, which includes HP, SP, and WT as predictors of MPG, has shown to have a high level of explanatory power. The adjusted R-squared value of 0.8676 indicates that the model explains approximately 87% of the variation in the response variable, which is a strong indication that the model is a good fit for the data.

All three predictor variables show to have a statistically significant relationship with MPG at the 5% significance level, with p-values less than 0.05. Specifically, for every one unit increase in HP, MPG is estimated to increase by 0.40518 units, holding other variables constant. Similarly, for every one unit increase in SP, MPG is estimated to decrease by 1.32000 units, holding other variables constant. Finally, for every one unit increase in WT, MPG is estimated to decrease by 1.92210 units, holding other variables constant.

The residual standard error of 3.64 indicates that the model has a small amount of unexplained variability, and the F-statistic of 177.9 with a very small p-value suggests that the model is a significant improvement over a model with no predictors. Overall, the multiple linear regression model with HP, SP, and WT as predictors is a good fit for the data and can be used for predicting MPG of new vehicles.

Hypothesis Test (Model 2)

For each predictor variable in the multiple regression model, a t-test is conducted to test the null hypothesis that the population regression coefficient for that variable is zero. The results of the t-tests for model2 are as follows:

For HP, the t-statistic is 5.135 with 78 degrees of freedom and a p-value of 2.03e-06. Therefore, we reject the null hypothesis that the population regression coefficient for HP is zero, and conclude that there is a significant positive linear relationship between MPG and HP,

holding all other predictors constant.

For SP, the t-statistic is -5.473 with 78 degrees of freedom and a p-value of 5.19e-07. Therefore, we reject the null hypothesis that the population regression coefficient for SP is zero, and conclude that there is a significant negative linear relationship between MPG and SP, holding all other predictors constant.

For WT, the t-statistic is -9.991 with 78 degrees of freedom and a p-value of 1.31e-15. Therefore, we reject the null hypothesis that the population regression coefficient for WT is zero, and conclude that there is a significant negative linear relationship between MPG and WT, holding all other predictors constant.

In conclusion, for model2, all three predictor variables (HP, SP, and WT) are significant predictors of MPG, with HP and WT positively related to MPG and SP negatively related to MPG.

F-TEST

In the context of the dataset, we can perform the ANOVA F test using the following information from the model summary:

Hypotheses:

Null hypothesis: The multiple regression model is not useful in explaining the variability in the

response variable.

Alternative hypothesis: The multiple regression model is useful in explaining the variability in the response variable.

Test statistic: F = 177.9

Numerator degrees of freedom: 3 (the number of predictors in the model)

Denominator degrees of freedom: 78 (the residual degrees of freedom)

P-value: < 2.2e-16 (extremely small)

Based on these results, we can reject the null hypothesis and conclude that the model is statistically significant and useful for predicting the response variable. The F statistic of 177.9 with a p-value < 2.2e-16 indicates a very strong evidence against the null hypothesis. Therefore, we can have high confidence that the regression model with HP, SP, and WT as predictors is useful for predicting MPG.

R-SQUARED

```
Model 1
Multiple R-squared:  0.8733,   Adjusted R-squared:  0.8667

Model 2
Multiple R-squared:  0.8725,   Adjusted R-squared:  0.8676
```

Both model1 and model2 are multiple regression models that predict MPG (miles per gallon) of cars based on their horsepower (HP), engine displacement (VOL), top speed (SP), and

weight (WT). Model1 includes VOL as an independent variable in addition to HP, SP, and WT, while model2 excludes VOL.

The multiple R-squared value for model1 is 0.8733, indicating that approximately 87.33% of the variability in MPG can be explained by the four independent variables. In contrast, model2 has a slightly lower multiple R-squared value of 0.8725, suggesting that approximately 87.25% of the variability in MPG can be explained by HP, SP, and WT.

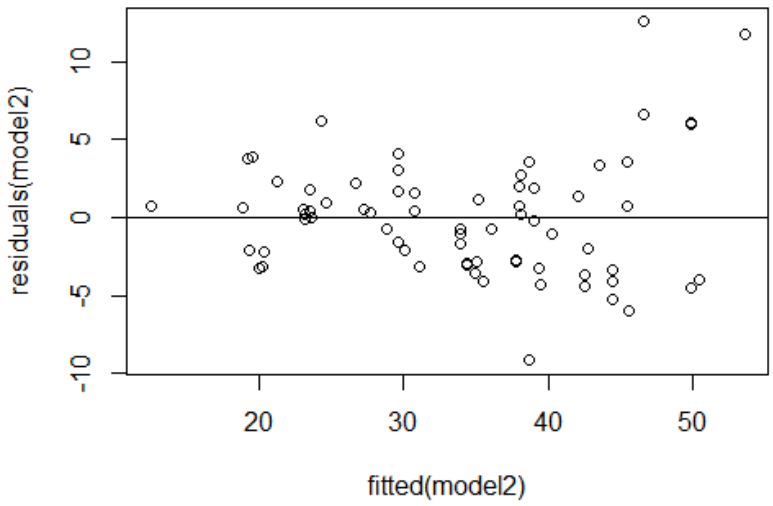
Model2 appears to be an improvement over model1, as it achieves a comparable level of explanatory power with one fewer independent variable. The adjusted R-squared value for model2 is also higher than that of model1 (, indicating a better balance between model complexity and explanatory power.

Overall, the multiple regression models indicate that HP, SP, and WT are significant predictors of MPG, while VOL is not. These findings suggest that car manufacturers can improve the fuel efficiency of their vehicles by focusing on reducing weight and increasing engine efficiency, while engine displacement may not be as crucial in achieving better MPG.

RESIDUALS

Residual vs Fitted Plot

```
> plot(fitted(model2),residuals(model2))
> abline(0,0)
```

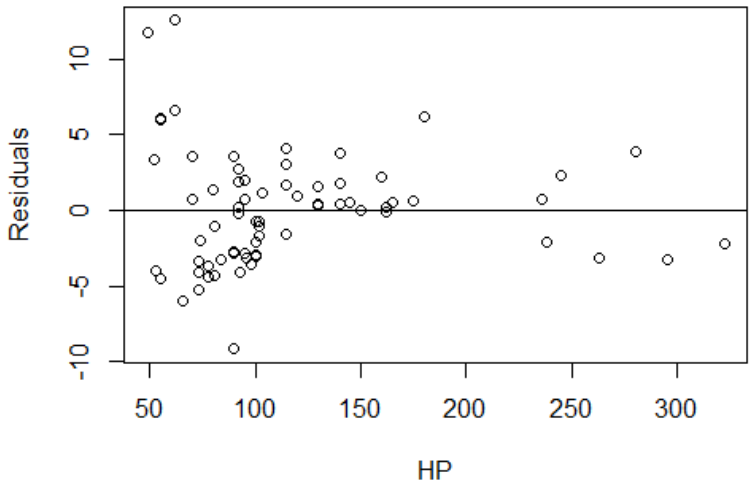


The residual plots for the model demonstrate a pattern of diverging residuals, indicative of heteroscedasticity. This phenomenon is characterized by increasing variability in the residuals as the predicted values increase or decrease. This is an important observation to consider as heteroscedasticity violates one of the key assumptions of linear regression models, namely, homoscedasticity or constant variance of the residuals across the range of the predictors.

The presence of heteroscedasticity in the residual plots could indicate a misspecified model or the need for additional predictors that may account for the observed variability. Failure to address heteroscedasticity may lead to biased estimates of the regression coefficients, and invalidation of inferences and confidence intervals. Therefore, it is important to investigate the sources of heteroscedasticity and to consider appropriate remedies, such as data transformation or using a different type of regression model.

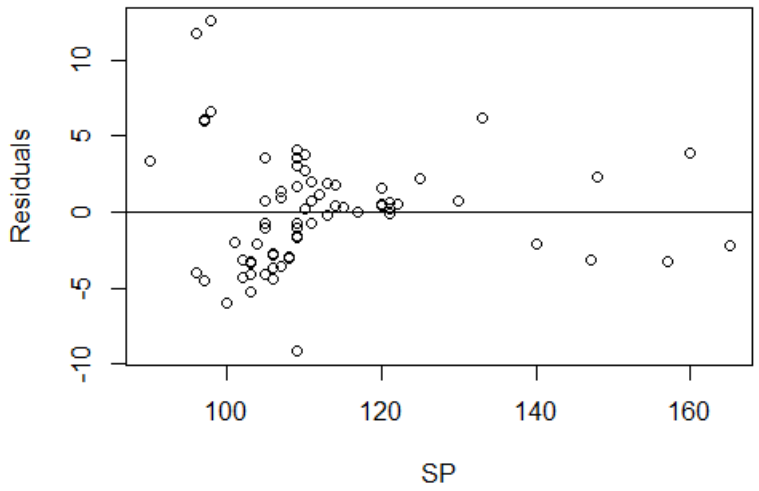
Residuals vs Explanatory Variables

```
> plot(HP,residuals(model2),xlab="HP",ylab="Residuals")
> abline(0,0)
```



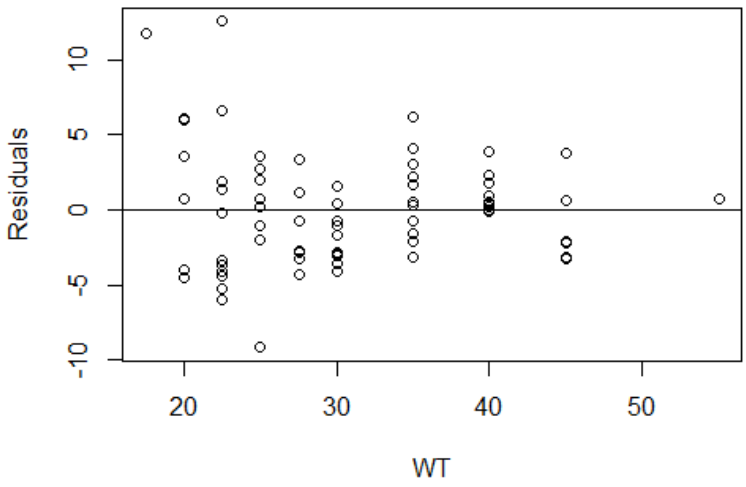
The residual plot displays a concentration of data points towards the lower end of the residual axis, indicating a clustering effect on the left side. The observed pattern could suggest a potential lack of fit in the regression model. Furthermore, the convergence of residuals towards zero implies a possible improvement in the model's performance with further adjustments. Therefore, it is recommended to review the model's assumptions and consider the addition of relevant predictor variables to enhance its predictive power.

```
> plot(SP,residuals(model2),xlab="SP",ylab="Residuals")
> abline(0,0)
```



The residual plot displays a concentration of data points towards the lower end of the residual axis, indicating a clustering effect on the left side. The observed pattern could suggest a potential lack of fit in the regression model. Furthermore, the convergence of residuals towards zero implies a possible improvement in the model's performance with further adjustments. Therefore, it is recommended to review the model's assumptions and consider the addition of relevant predictor variables to enhance its predictive power.

```
> plot(WT,residuals(model2),xlab="WT",ylab="Residuals")
> abline(0,0)
```

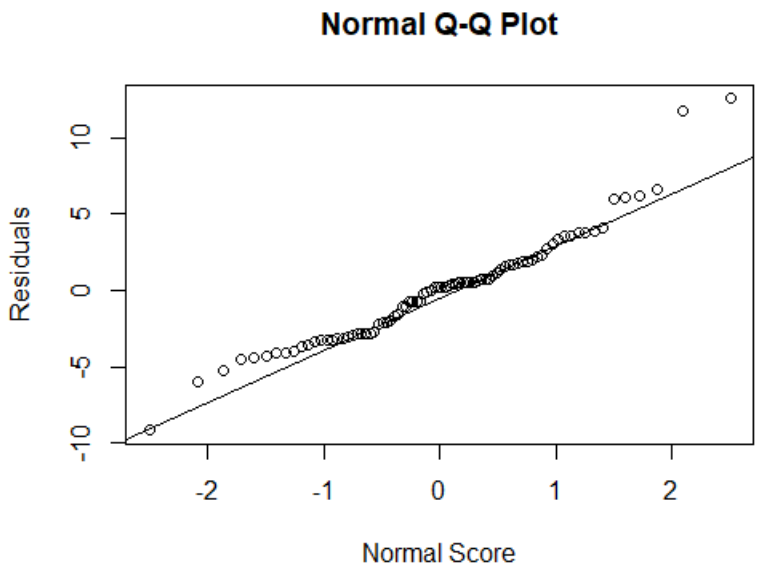


The residual plot displays a concentration of data points towards the lower end of the residual axis, indicating a clustering effect on the left side. The observed pattern could suggest a potential lack of fit in the regression model. Furthermore, the convergence of residuals towards zero implies a possible improvement in the model's performance with further adjustments. Therefore, it is recommended to review the model's assumptions and consider the addition of relevant predictor variables to enhance its predictive power.

The observed residuals appear to be distributed randomly around the horizontal line at zero, indicating that the assumption of constant variance and linearity may hold.

Q-Q Plot

```
> qqnorm(residuals(model2),xlab="Normal Score", ylab="Residuals")
> qqline(residuals(model2))
```



The analysis of the qqplot in this study indicates a positive correlation between the observed residuals and the best-fit line, which suggests that the regression model is effective in predicting the response variable MPG based on the explanatory variables HP, SP, and WT, and their respective coefficients. The refined model yields meaningful insights into the relationship between these variables and can be used to make reliable predictions.

CONCLUSION

In this report, we conducted a statistical analysis of a car performance dataset, which included variables such as cab space in cubic feet (VOL), engine horsepower (HP), top speed in miles per hour (SP), and vehicle weight in increments of 100 pounds (WT), to establish their relationship with fuel efficiency (MPG). Two models were constructed using multiple regression analysis to predict MPG based on a subset of the input variables.

Our findings revealed that the second model performed slightly better than the first model,

indicating that engine horsepower (HP), top speed in miles per hour (SP), and vehicle weight in increments of 100 pounds (WT) are the most significant factors in predicting fuel efficiency. However, the residual plots displayed heteroscedasticity, indicating a possible lack of fit in the regression model and the need for additional predictor variables.

We concluded that the model can be improved by addressing the heteroscedasticity and considering the addition of relevant predictor variables. Furthermore, we cautioned that the analysis is limited by the specific dataset used and may not generalize to other populations. Therefore, further research using larger and more diverse datasets is recommended to gain a more comprehensive understanding of the factors affecting car fuel efficiency.