

Unveiling Trends and Patterns in Automotive Fuel Efficiency



Tushar Vishwanath

Graduate Student and Assistant in Mechanical Engineering,

School for Engineering of Matter, Transport and Energy

Arizona State University

e-mail: tvishwan@asu.edu

United States Environmental Protection Agency

The EPA upholds national environmental standards through assessments, research, and collaborations, fostering comprehensive protection and sustainable practices throughout the United States.

EPA does its emission regulation using data collected that they store in their fuel economy database (www.fueleconomy.gov).

Fuel economy is one of the major targets that must be accounted for controlling emission as well as usage of fuel.

According to a census conducted by NADA(National Automobile Dealers Association) there are 280 million cars being used in the US out of which round 256 million are gasoline based.





United States
Environmental Protection
Agency

The average fuel economy for cars in the US is around 24-25 miles per gallon. Increasing the economy by even 1 mile/gallon on an average and considering an average car travels around 29 miles a day [federal highway administration (FHWA)] we can save 3.65 million gallons of fuel every day.

References :

- ☐ www.highways.dot.gov
- ☐ www.fueleconomy.gov
- ☐ www.epa.gov
- ☐ www.nada.org
- ☐ www.eia.gov



Research Question

- 1) Which Parameter has the most significant impact on fuel economy
- 2) How does the change in the most significant parameter affect the fuel economy

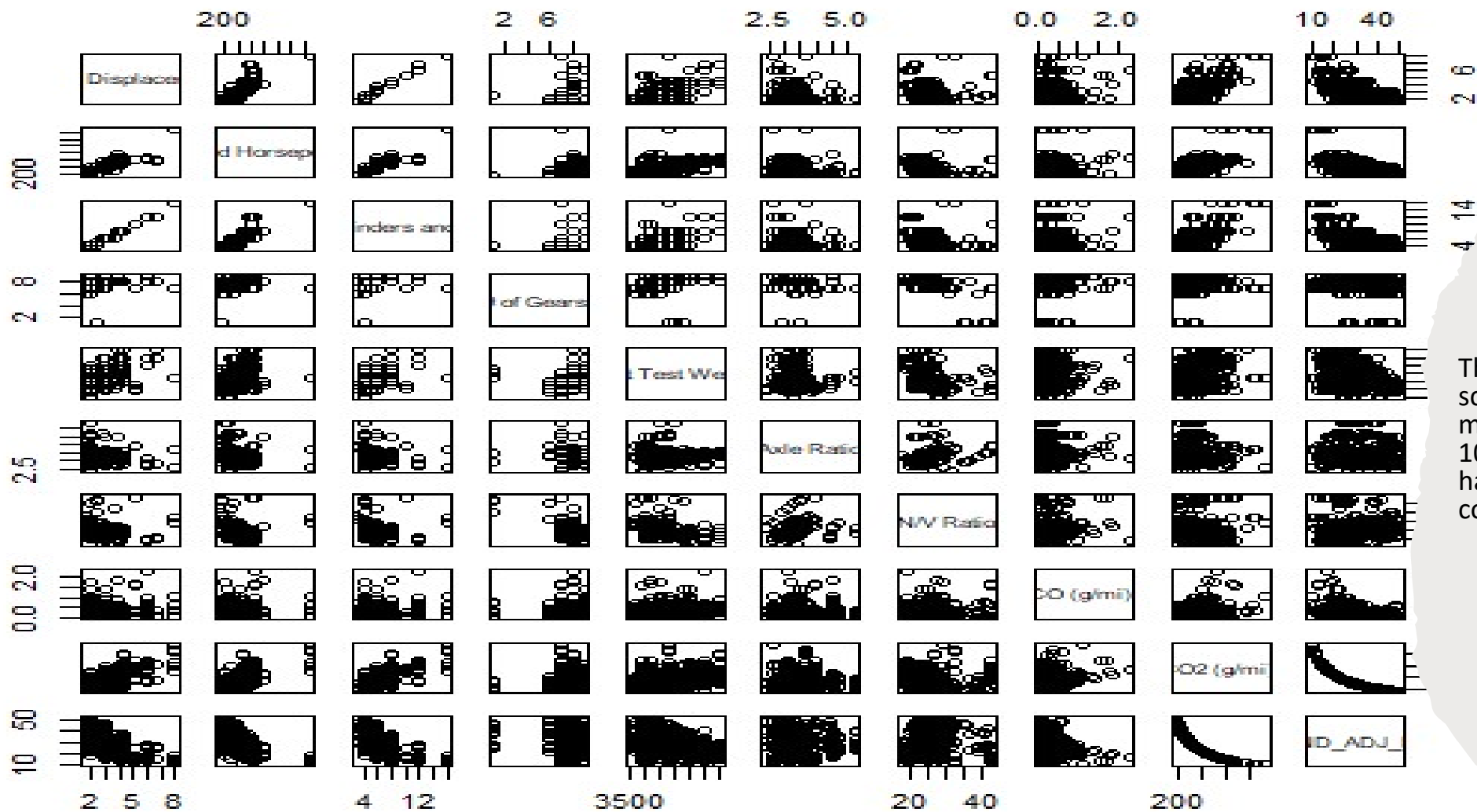
```
# Filter 'car' dataset for 'Tier 2 Cert Gasoline' and 'All wheel Drive', selecting numeric columns
filtered_data <- car %>%
  filter(`Test Fuel Type Cd` == "61" & `Drive System Description` == "All wheel Drive") %>%
  select(all_of(columns_to_keep))
```

Data		
car	4467 obs. of 67 variables	
filtered_data	771 obs. of 12 variables	

Data Table

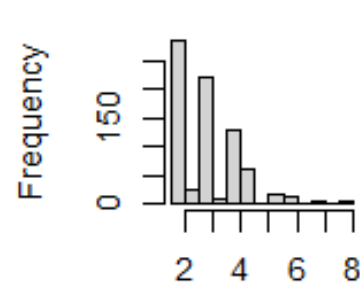
Variable	Type	Units	Typical range
Vehicle Displacement	Categorical	litres(L)	1 to 9
Rated Horsepower	Continuous	horsepower(1HP = 746 Watts)	50 to 1600
No of cylinders and Rotors	Categorical	NA	2 to 16
No of gears	Categorical	NA	2 to 10
Equivalent test Weight	Continuous	pounds	1000 to 6000
Axle ratio	Continuous	revolutions/turn	1 to 10
N/V ratio	Continuous	RPM/MPH	2 to 5
Test Fuel Type	Categorical	NA	61 and 62
Mileage	Continuous	MPG	9 to 30
CO	Continuous	g/mi	almost negligible
CO2	Continuous	g/mi	100 to 1000

Pairs(mydata)

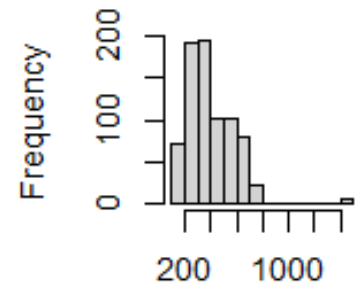


The scatterplots matrices of the 10 variables we have considered

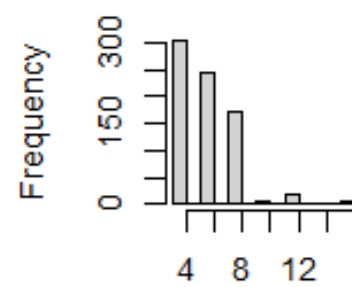
Histograms



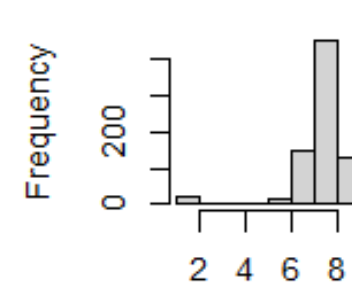
Test Veh Displacement (L)



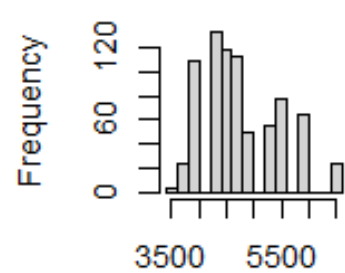
Rated Horsepower



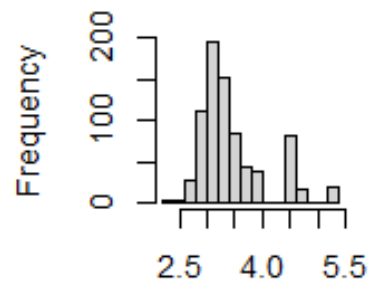
of Cylinders and Rotors



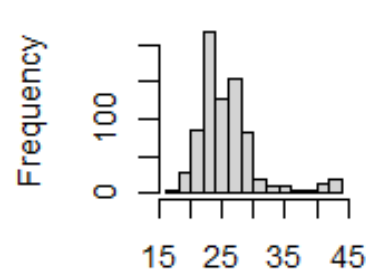
of Gears



Equivalent Test Weight (lbs)



Axle Ratio



N/V Ratio

Histograms of
considered factors
that affect the
vehicle mileage

Model 1 (Full Model)

After analysis of the linear model, we eliminated terms with higher P values. Our linear model contains 7 factors now:

Response: RND_ADJ_FE						
	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Test Veh Displacement (L)	1	17133.7	17133.7	1503.222	< 2.2e-16	***
Rated Horsepower	1	1594.8	1594.8	139.921	< 2.2e-16	***
Equivalent Test Weight (lbs.)	1	658.4	658.4	57.765	9.651e-14	***
Axle Ratio	1	1372.0	1372.0	120.373	2.091e-15	***
THC (g/mi)	1	658.4	658.4	57.765	3.127e-14	***
CO (g/mi)	1	4448.3	4448.3	390.267	< 2.2e-16	***
CO2 (g/mi)	1	21466.2	21466.2	1883.341	< 2.2e-16	***

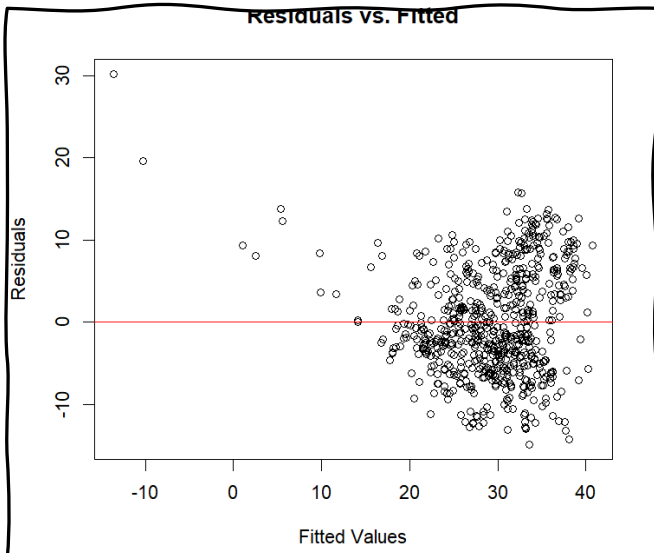
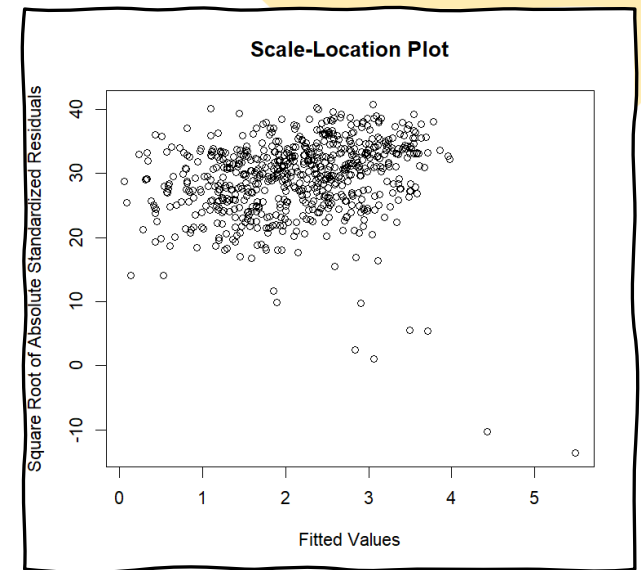
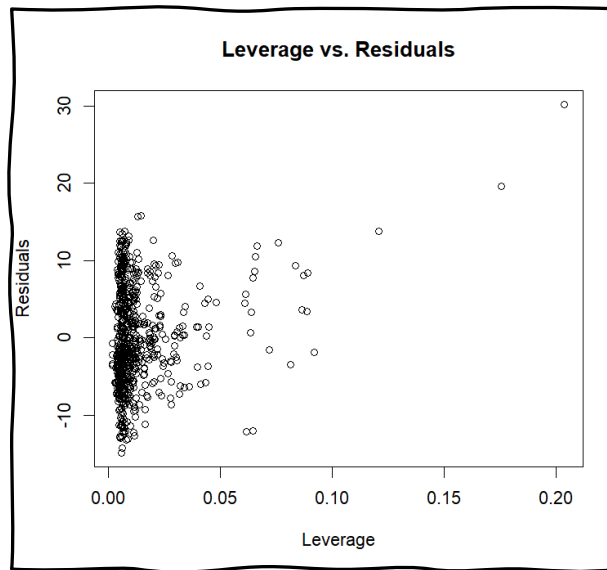
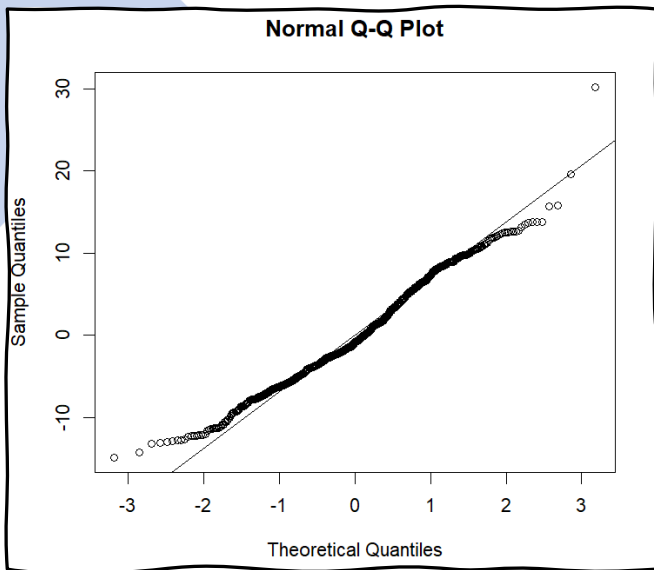
The linear model after rejecting the factors with high P values.

The rejected factors include :

- No of gears
- N/V ratio
- No of cylinders and rotors

For simplification of the model, we are also rejecting the factors with a relatively high P value, they include :

- Equivalent Test weight
- Axle ratio
- THC



Vif(Full Model)

Test Veh Displacement (L)	48.787564	# of Cylinders and Rotors	54.084406
# of Gears	1.225479	Rated Horsepower	5.753412
Equivalent Test weight (lbs.)	1.746058	Axle Ratio	1.168730
THC (g/mi)	1.550065	CO (g/mi)	1.557416
CO2 (g/mi)	2.238667		

Model 2(Reduced Model 1)

#without (Test Veh Displacement)

```
reduced_model_1 <- lm(RND_ADJ_FE ~ `Rated Horsepower` + `CO (g/mi)` + `CO2 (g/mi)`, data = mydata)
```

For our reduced model 1 we consider an interaction of all factors except Test Vehicle Displacement

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	53.1312462	0.4136579	128.442	<2e-16	***
`Rated Horsepower`	0.0023565	0.0009526	2.474	0.0136	*
`CO (g/mi)`	-1.3950756	0.5724858	-2.437	0.0151	*
`CO2 (g/mi)`	-0.0734626	0.0016321	-45.011	<2e-16	***

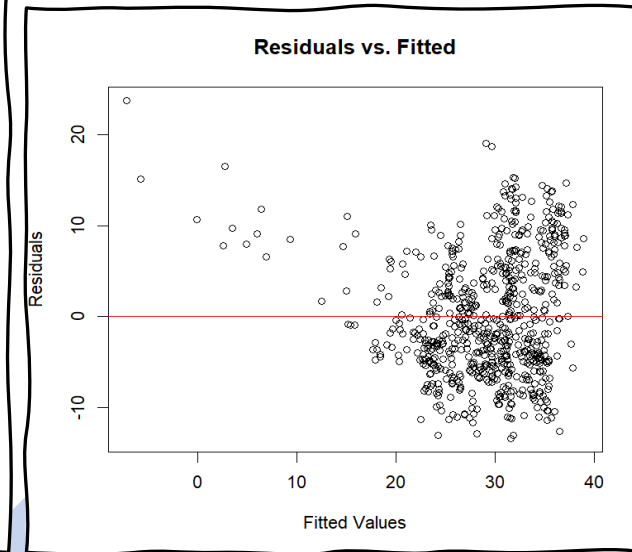
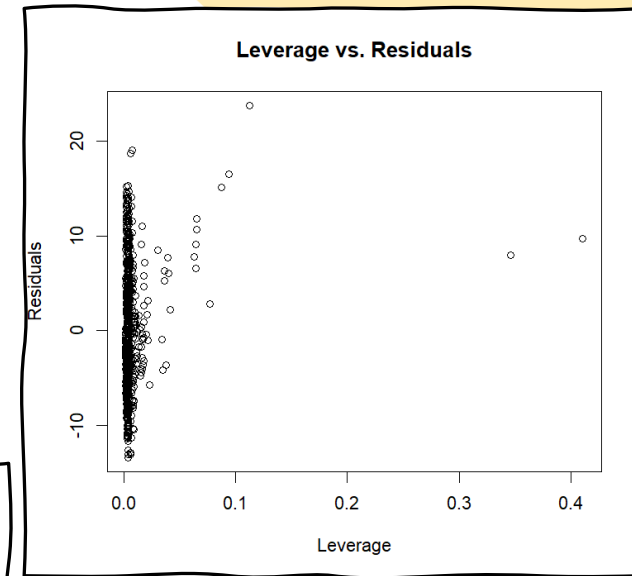
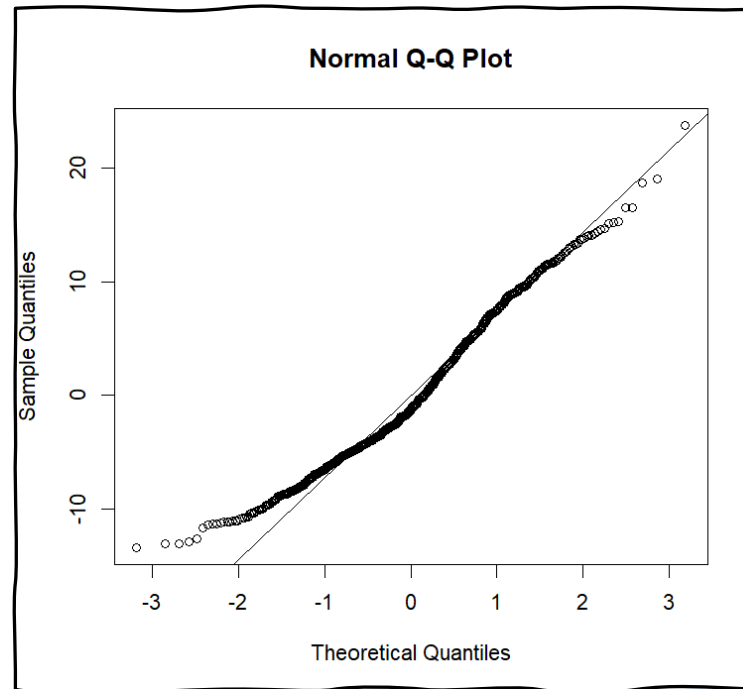
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.454 on 695 degrees of freedom
(72 observations deleted due to missingness)

Multiple R-squared: 0.8497, Adjusted R-squared: 0.8491

F-statistic: 1310 on 3 and 695 DF, p-value: < 2.2e-16

- The graphs for Reduced Model 1 show that most of the points in the residuals are concentrated in the same area, with some scattered outliers.
- The Q-Q plot closely aligns with a straight line, suggesting that the residuals follow a normal distribution.



Model 3(Reduced Model 2)

#without (CO)

```
reduced_model_3 <- lm(RND_ADJ_FE ~ `Test Veh Displacement (L)` + `Rated  
Horsepower` + `CO2 (g/mi)`, data = mydata)
```

For our reduced model 2 we
consider an interaction of all factors
except Carbon Monoxide Emissions.

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	53.137876	0.431920	123.027	<2e-16 ***
`Test Veh Displacement (L)`	0.283237	0.243647	1.162	0.245
`Rated Horsepower`	0.001560	0.001526	1.022	0.307
`CO2 (g/mi)`	-0.076046	0.001530	-49.710	<2e-16 ***

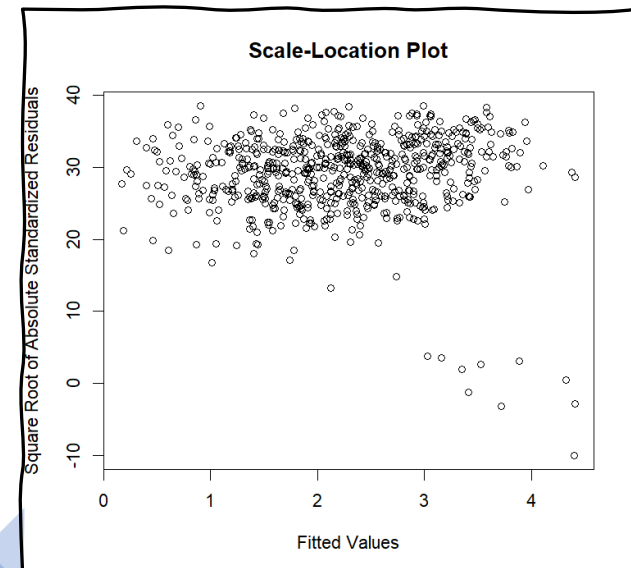
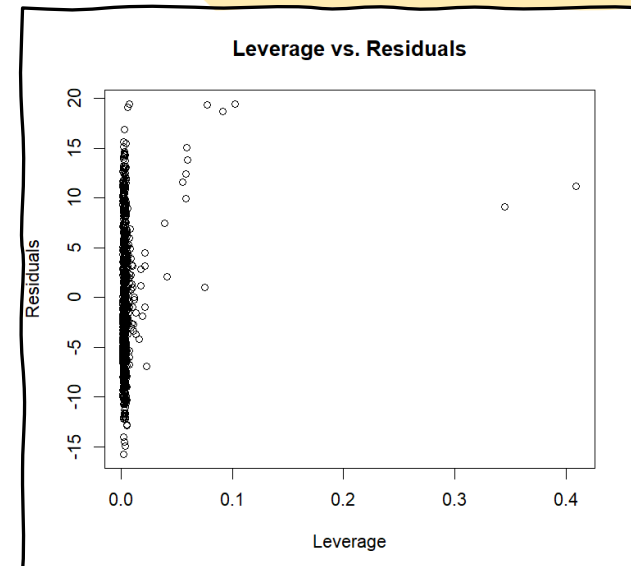
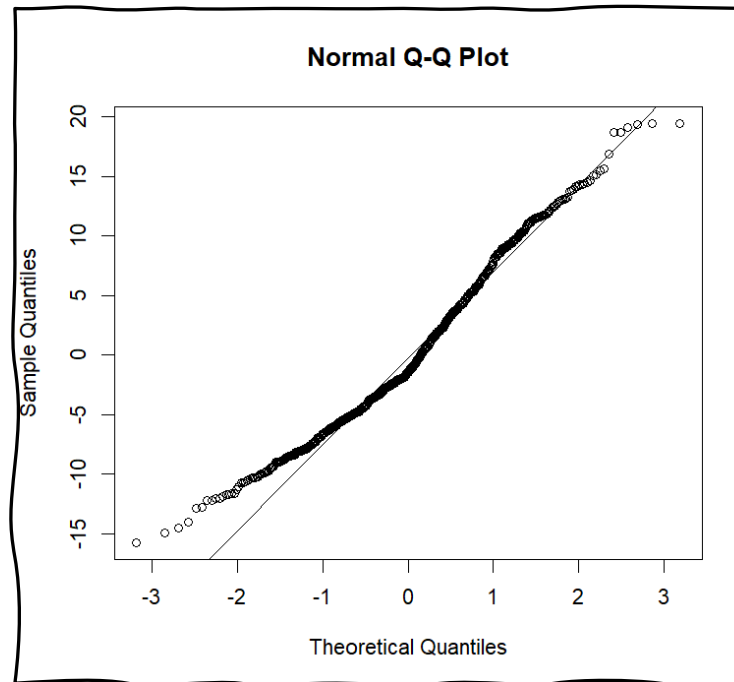
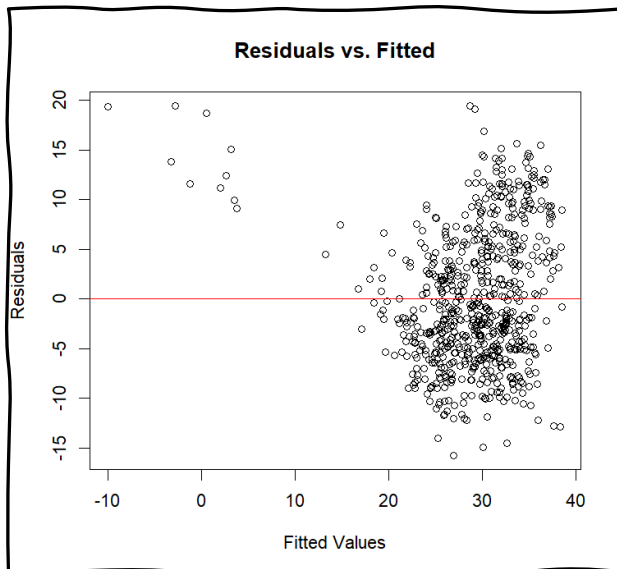
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.468 on 709 degrees of freedom
(58 observations deleted due to missingness)

Multiple R-squared: 0.849, Adjusted R-squared: 0.8484

F-statistic: 1329 on 3 and 709 DF, p-value: < 2.2e-16

- The graphs of Reduced Model 2 show that the points in the residuals are more confined to a line, with very few scattered points throughout the plot.
- The Q-Q plot aligns with the straight line, but there is a slight deviation at the lower end.



Model 4(Reduced Model 3)

#without (Rated Horsepower)

```
reduced_model_2 <- lm(RND_ADJ_FE ~ `Test Veh Displacement (L)` + `CO (g/mi)`  
+ `CO2 (g/mi)`, data = mydata)
```

For our reduced model 3 we consider an interaction of all factors except Rated Horsepower Emissions.

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	52.957610	0.428754	123.515	<2e-16	***
`Test Veh Displacement (L)`	0.377778	0.154208	2.450	0.0145	*
`CO (g/mi)`	-1.331825	0.579572	-2.298	0.0219	*
`CO2 (g/mi)`	-0.073635	0.001686	-43.676	<2e-16	***

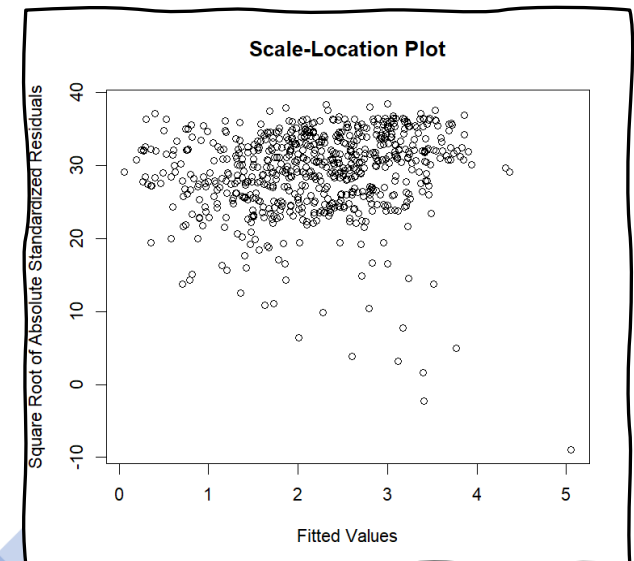
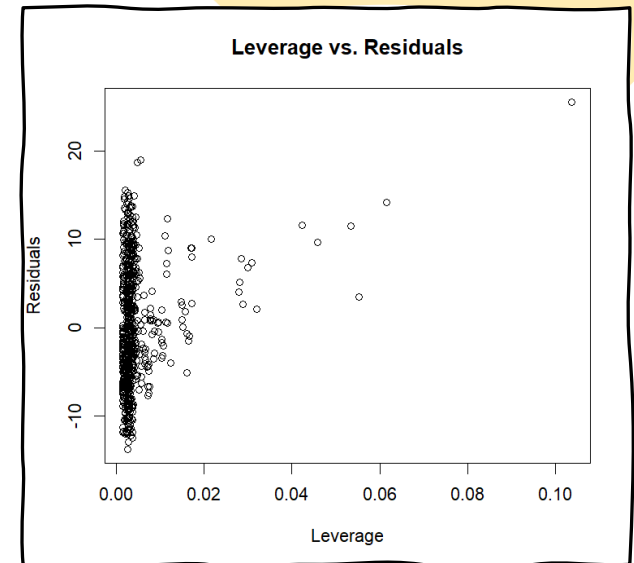
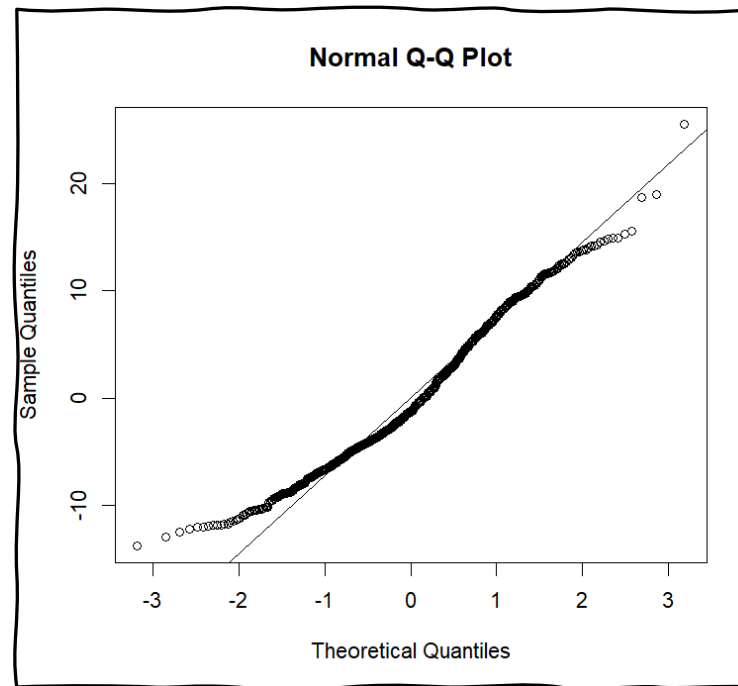
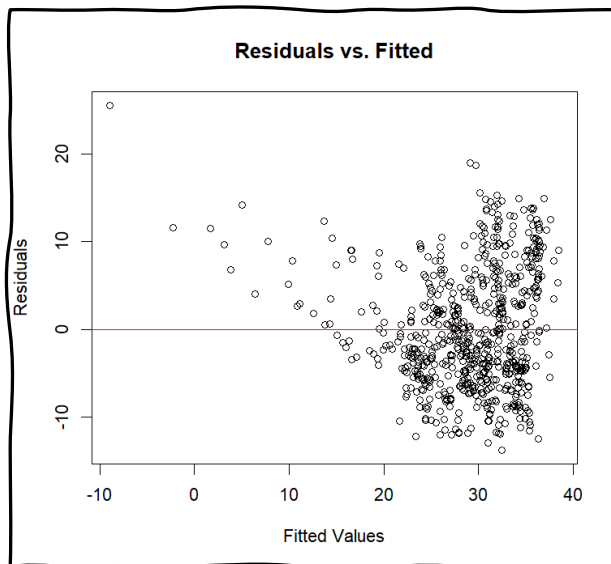
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.455 on 695 degrees of freedom
(72 observations deleted due to missingness)

Multiple R-squared: 0.8497, Adjusted R-squared: 0.8491

F-statistic: 1310 on 3 and 695 DF, p-value: < 2.2e-16

- The plots of Reduced Model 3 show that the points in the residuals are slightly scattered, with some possible outliers.
- The Q-Q plot mostly aligns with the straight line, but there is a slight deviation at the upper and lower ends.



Model 5(Reduced Model 4)

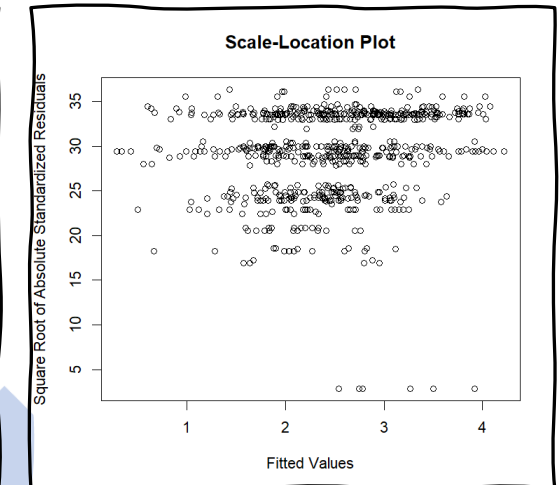
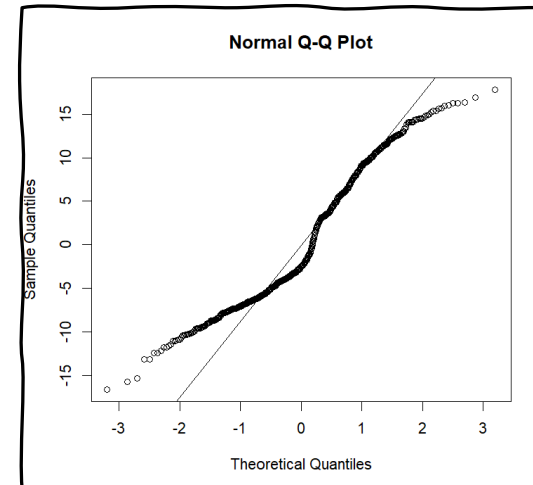
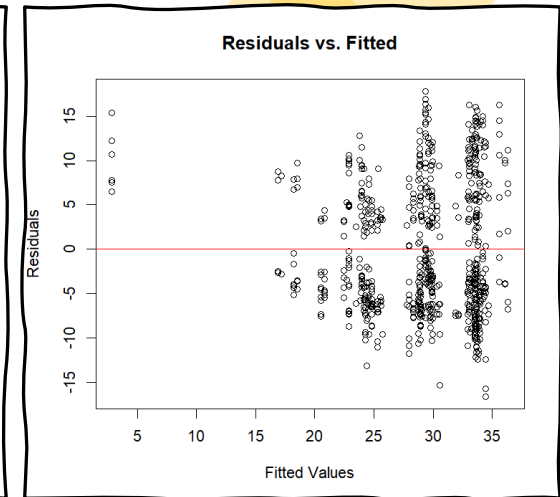
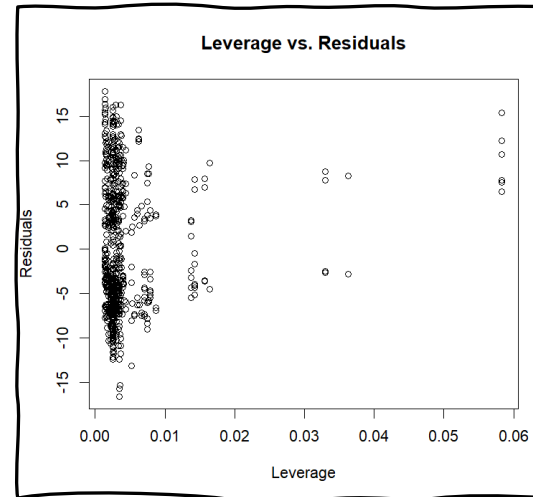
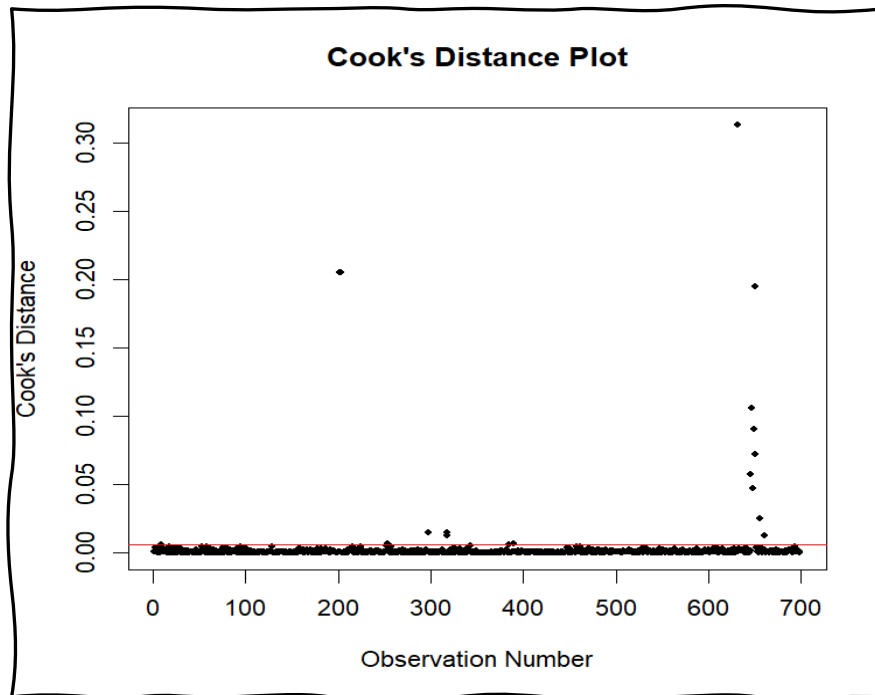
```
#without (CO_2)
reduced_model_4 <- lm(RND_ADJ_FE ~ `Test Veh Displacement (L)` + `Rated
Horsepower` + `CO (g/mi)`, data = mydata)
```

For our reduced model 4 we consider an interaction of all factors except Carbon Dioxide emissions.

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    44.587172   0.746394   59.737 < 2e-16 ***
`Test Veh Displacement (L)` -2.942339   0.454855  -6.469 1.86e-10 ***
`Rated Horsepower`    -0.009128   0.002902  -3.145 0.00173 **
`CO (g/mi)`      -12.406237   0.999217 -12.416 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

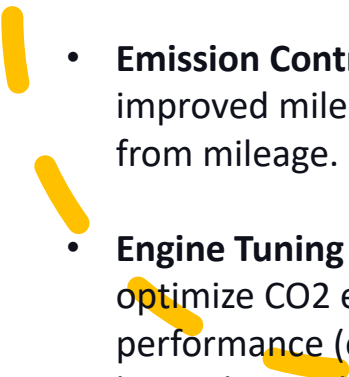
Residual standard error: 6.638 on 695 degrees of freedom
(72 observations deleted due to missingness)
Multiple R-squared:  0.4451,    Adjusted R-squared:  0.4427
F-statistic: 185.8 on 3 and 695 DF,  p-value: < 2.2e-16
```


- In this model, the Q-Q plot deviates the most from the straight line.
- The least R-squared value signifies that the model explains only a small proportion of the variability in the data.
- This shows that the CO2 is the most significant factor



Discussion/Challenges faced



- ❑ Though our model looks pretty straightforward we considered and performed many iterations to filter out data and to select the factors that affect our model. We still feel that we could have run different conditions to obtain a better model. Even though our conclusion may be solid, thinking through the automotive aspect creates a question on why important factors such as number of cylinders have lesser impact on the fuel economy.
 - ❑ Though the data is vast, there are certain parameters that are not included which may affect the relationship between CO2 emissions and fuel efficiency :
 - **Vehicle Weight and Aerodynamics:** Enhancing aerodynamics can enhance mileage efficiency without proportionately impacting CO2 emissions due to their non-linear relationship.
 - **Emission Control Systems:** Engine technologies designed for emission reduction may not always correlate with improved mileage. These systems can impact engine performance differently, affecting CO2 emissions distinct from mileage.
 - **Engine Tuning and Performance Modifications:** Modifications aimed at enhancing mileage might not necessarily optimize CO2 emissions, and vice versa. Tuning or performance modifications in vehicles like tuner cars for high performance (e.g., Lamborghini Huracan) vs. untuned models like a standard Honda Civic create divergent data loops due to their distinct design objectives.
- 



Results/Conclusion

- From the analysis of the full model and reduced model we can conclude that the Model5 (Reduced model 4) is the least optimized model with the R squared value of 44.5% and with its QQ plot being the one that deviates from the linear the most.
- Since model 5 is generated without CO2 this implies that CO2 might be a significant predictor influencing the overall performance of the model.
- Other variables might not fully compensate for CO2's absence, indicating its unique contribution to the model. Factors like Test Veh Displacement (L), Rated Horsepower, and CO (g/mi) demonstrate some significance, but their combined influence might not fully replace CO2's impact.
- The data shows CO2 is crucial in how the model works but understanding it better might need more study or different ways of looking at things. However, we're focused on different parts of the research, so we're not exploring this further right now.

