Fuel Economy in Vehicles

Brandon Cunningham

2024-04-24

Abstract

In this paper we use data from the Department of Energy to investigate various aspects of vehicles by plotting these factors and using a linear regression model attempt to assess their impact on the fuel economy of vehicles as measured by combined miles per gallon (MPG). Through this analysis it is found that there are many factors, including the transmission type of the vehicle, engine displacement, and more that all have significant impact on the combined MPG, and that the factor with the largest impact on combined MPG is the fuel type of the vehicle.

Data Preperation and Graphing

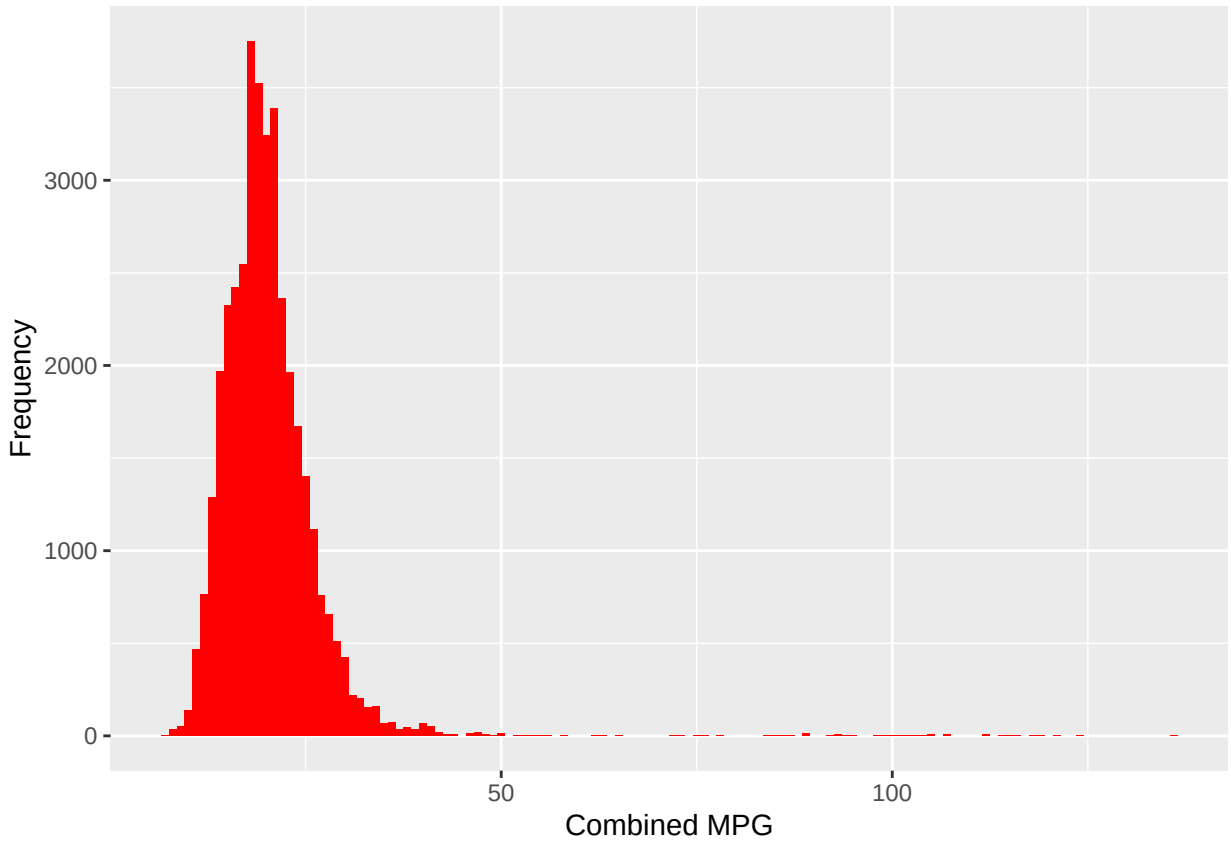
```
url <- 'https://raw.githubusercontent.com/btc2628/DATA607/main/fuel_data/fuel.csv'
fuel_raw <- readr::read_csv(url)
fuel_narrowed <- fuel_raw[, c('year', 'drive', 'transmission', 'engine_cylinders',
        'engine_displacement', 'turbocharger', 'fuel_type', 'combined_mpg_ft1')]
```

```
fuel_narrowed <- fuel_narrowed %>%
  mutate(transmission = str_extract(transmission, "^[^\\s\\(]+"))

fuel_narrowed <- fuel_narrowed %>%
  mutate(drive = case_when( grepl("4-Wheel Drive|All-Wheel Drive", drive) ~
                              "4-Wheel or All-Wheel Drive", TRUE ~ drive))
```

```
describe(fuel_narrowed$combined_mpg_ft1)
```

```
##    vars     n  mean   sd median trimmed  mad min max range skew kurtosis   se
## X1    1 38113 20.22 6.77     19   19.61 4.45   7 136   129 5.73    66.63 0.03
```

```
ggplot(fuel_narrowed, aes(x=combined_mpg_ft1)) +
   geom_bar(fill="red") +
  labs(x='Combined MPG', y='Frequency')
```

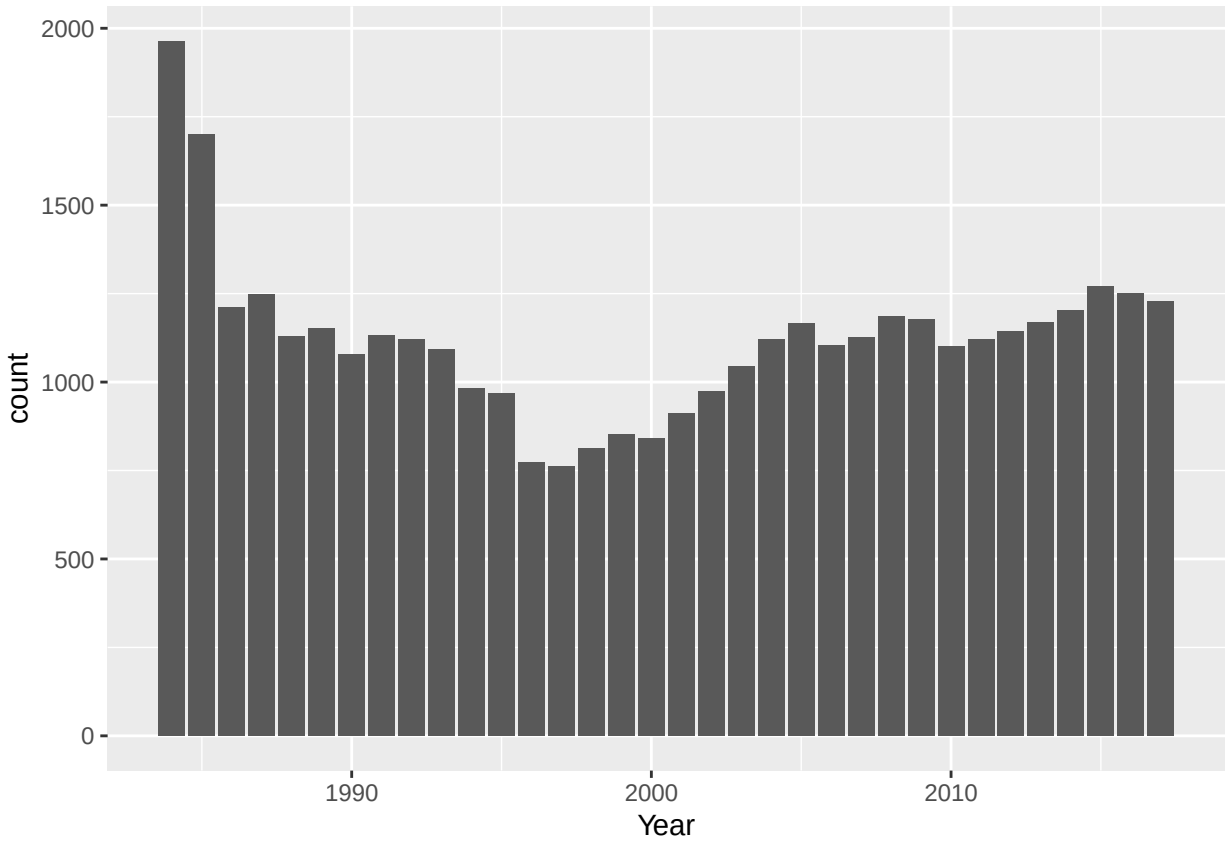


Here we get an overview of the distribution of fuel economies in the data set, it has a mean of 20.22mpg and median of 19mpg with a mainly normal distribution with a slight right skew and a few extreme outliers to the right.

```
describe(fuel_narrowed$year)
```

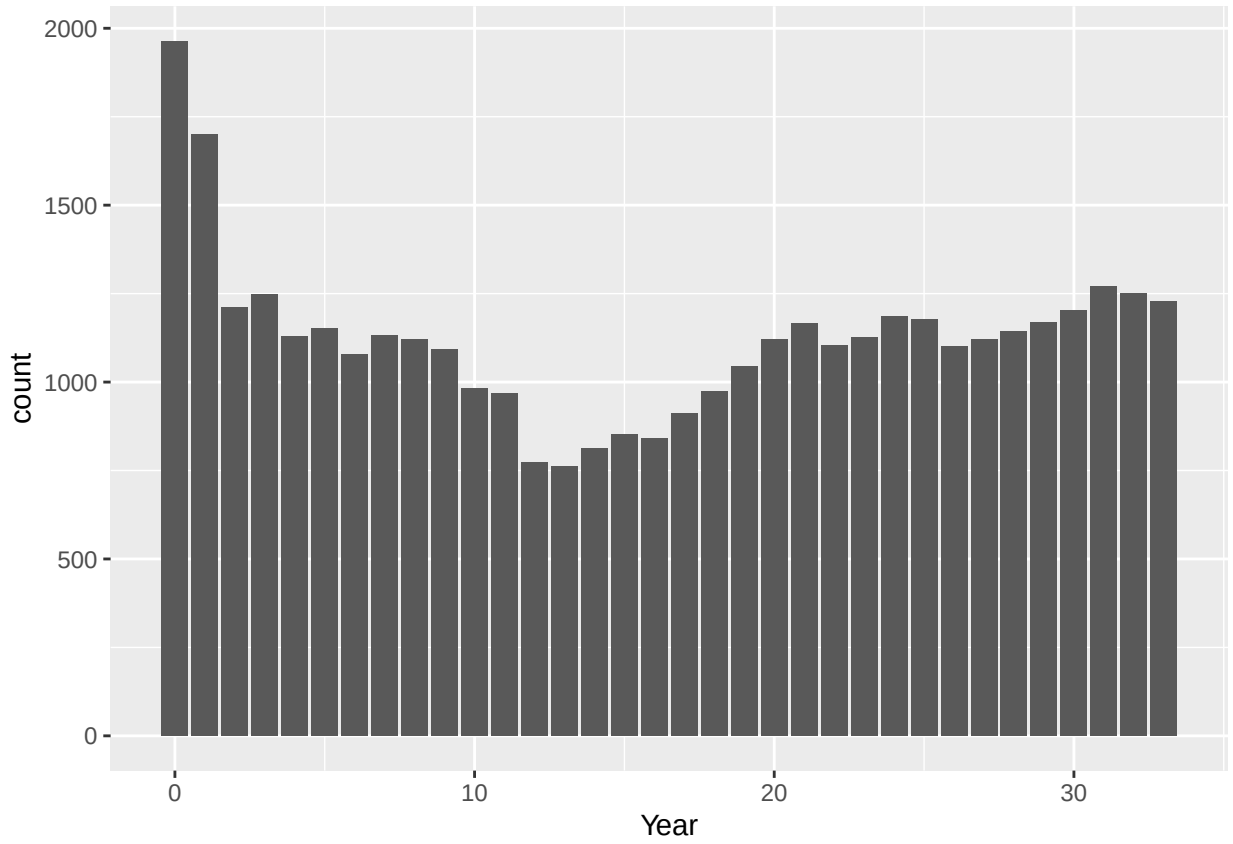
```
##    vars     n    mean    sd median trimmed   mad  min  max range  skew kurtosis
## X1    1 38113 2000.19 10.46   2001 2000.18 13.34 1984 2017    33 -0.02    -1.33
##      se
## X1 0.05
```

```
ggplot(fuel_narrowed, aes(x=year)) +
   geom_bar() +
  labs(x='Year')
```

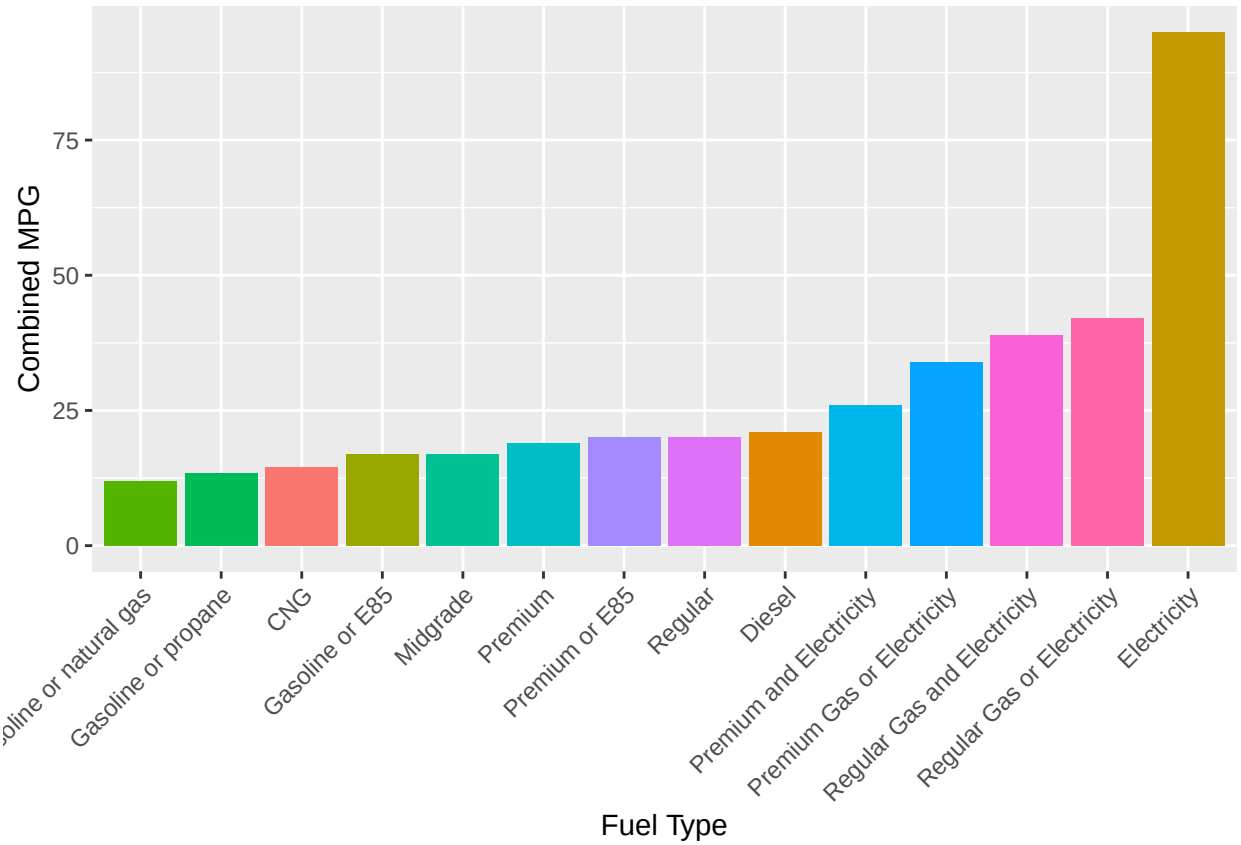


The data in this dataset has cars from year 1984 to 2017, with a relatively even distribution between the years. For the sake of prediction we are going to change year to be the number of years since 1984 so the year variable will start at 0.

```
fuel_narrowed$year <- fuel_narrowed$year -1984
ggplot(fuel_narrowed, aes(x=year)) +
   geom_bar() +
  labs(x='Year')
```

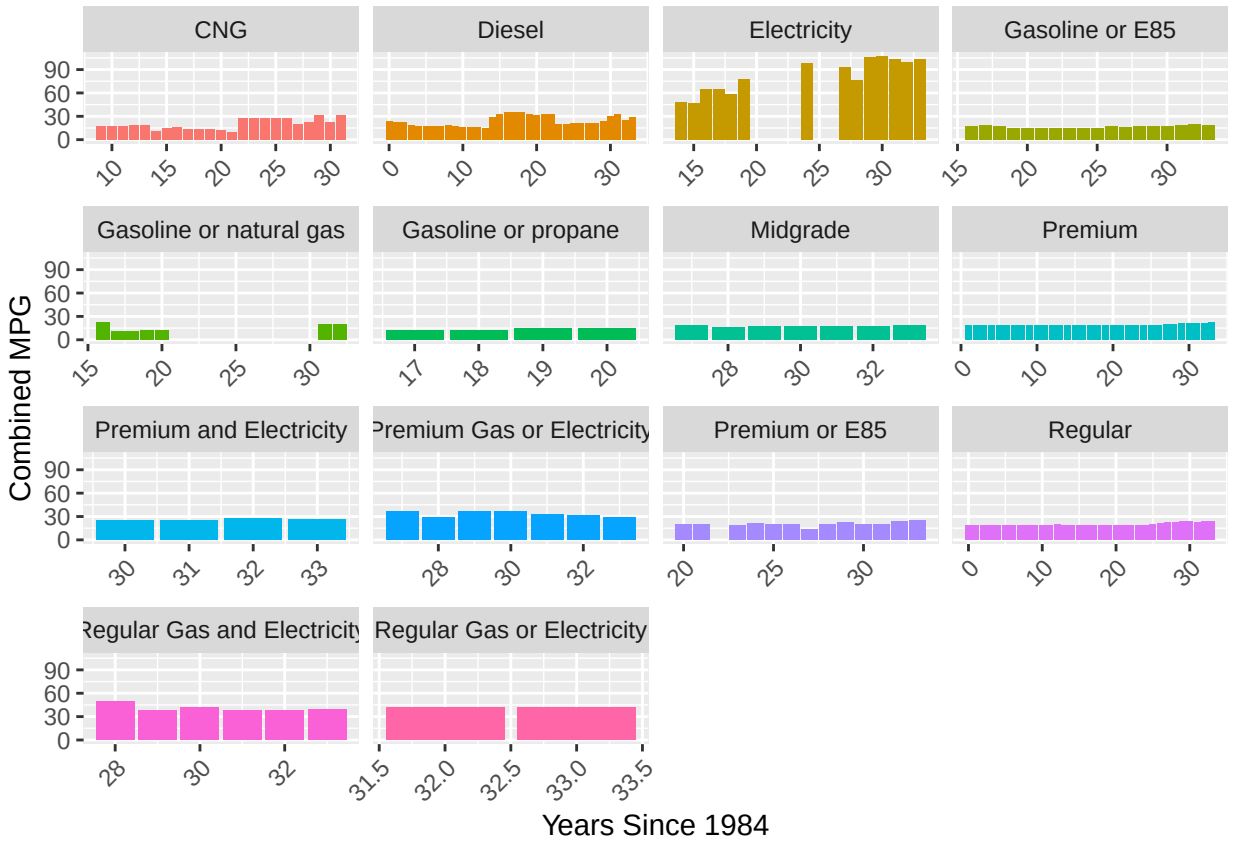



```
ggplot(fuel_narrowed, aes(x=reorder(fuel_type, combined_mpg_ft1, FUN=median),
                          y=combined_mpg_ft1, fill=fuel_type)) +
   geom_bar(stat="summary", fun="median", show.legend = FALSE) +
  labs(x='Fuel Type', y='Combined MPG') +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



In the above graph we can see that any car that uses electricity has a larger median combined MPG than every other fuel type, and we can see the ranking of the rest of the fuel types.

```
ggplot(fuel_narrowed, aes(x=year, y=combined_mpg_ft1, group=fuel_type, fill=fuel_type)) +
   geom_bar(stat="summary", fun="median", show.legend = FALSE) +
  facet_wrap(~fuel_type, scales = 'free_x') +
  labs(x='Years Since 1984', y='Combined MPG') +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

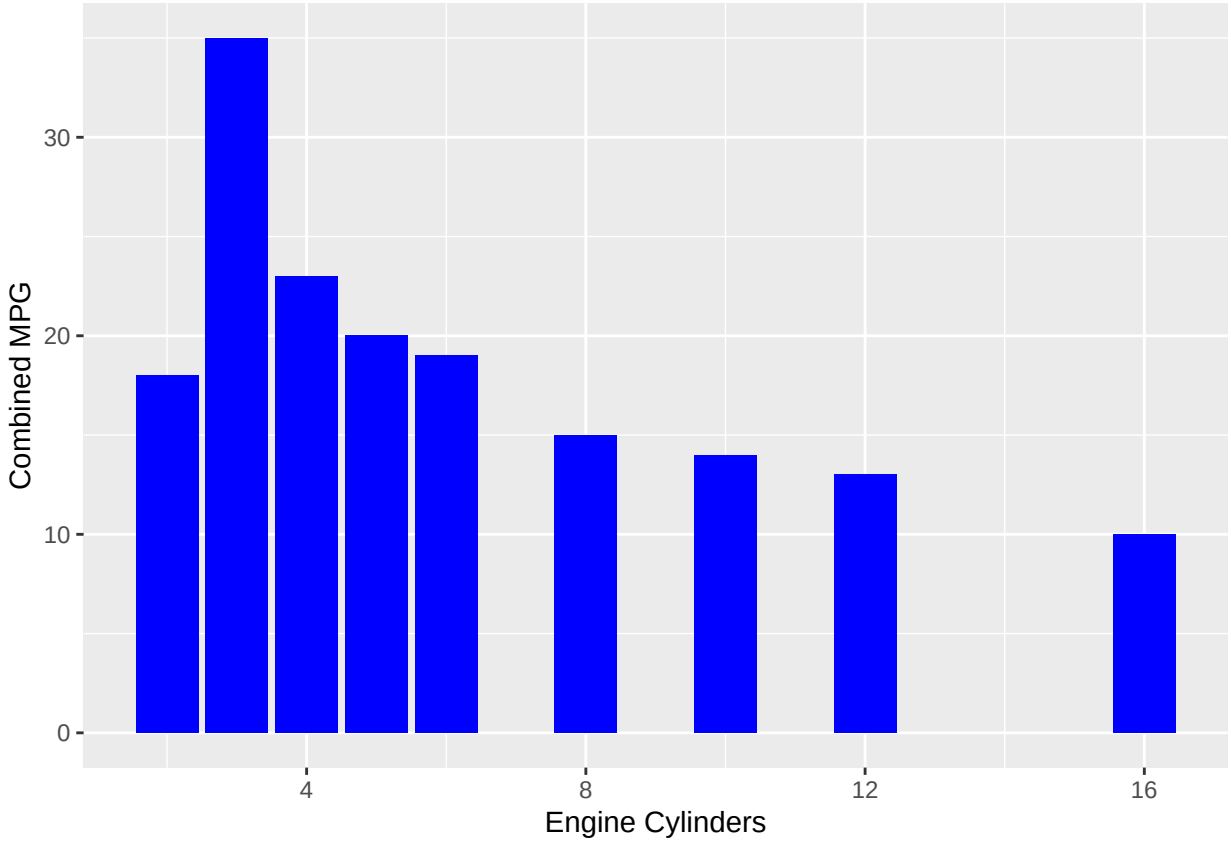


The graphs above show the average fuel mileage of cars per year over time, the graphs are separated into different fuel types as different fuels have different energy densities, or like in the case of electric cars, different ways of measuring their mpg equivalents.

```
describe(fuel_narrowed$engine_cylinders)
```

```
##    vars     n mean   sd median trimmed  mad min max range skew kurtosis   se
## X1    1 37977 5.74 1.75      6    5.59 2.97   2  16    14 0.85      0.9 0.01
```

```
ggplot(fuel_narrowed, aes(x=engine_cylinders, y=combined_mpg_ft1)) +
   geom_bar(stat="summary", fun="median", show.legend = FALSE, fill="blue") +
  labs(x='Engine Cylinders', y='Combined MPG')
```

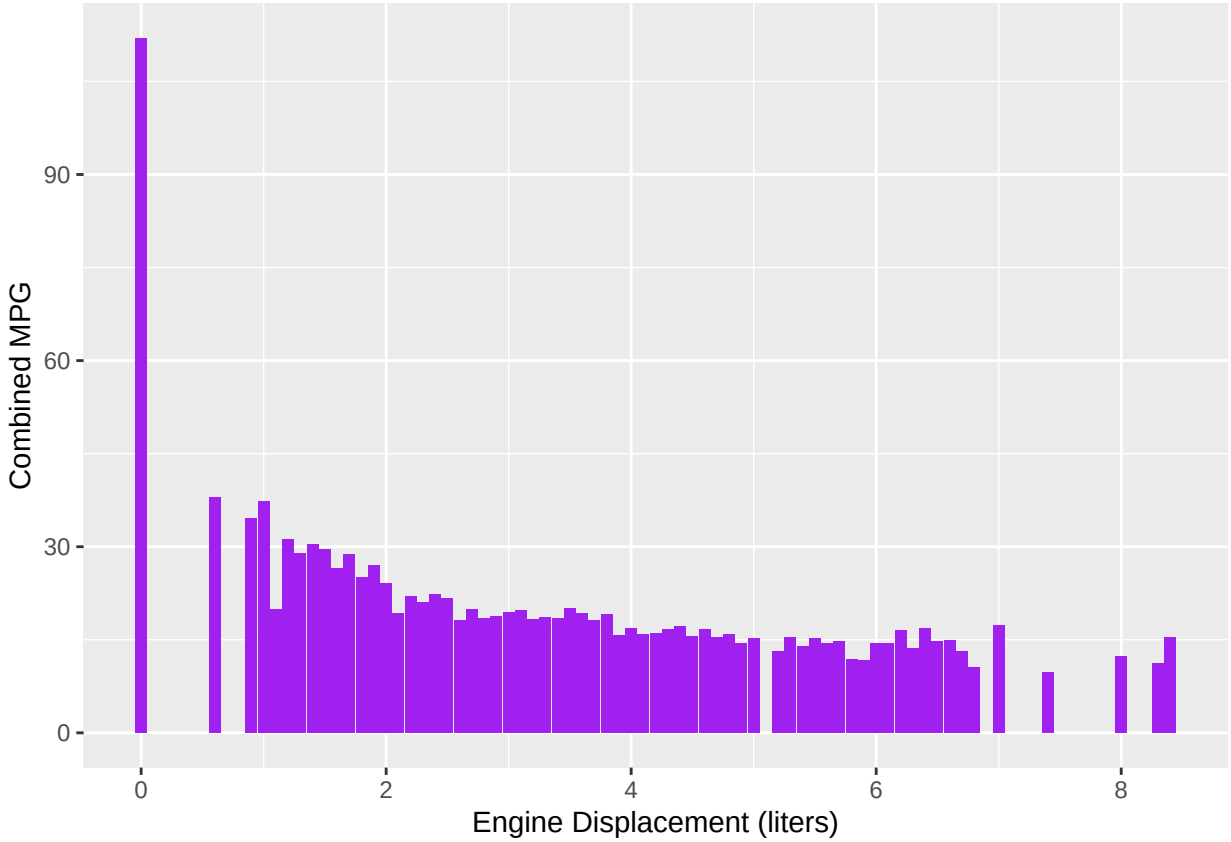


The graph above shows the relationship between an engines cylinders and its combined MPG, the relationship appears to be exponential decay with changes in MPG being more drastic as the displacement increases from a low value and leveling off over time, with the exception of silgle cylinder engines which appear to be less efficient than 2 cylinder engines.

```
describe(fuel_narrowed$engine_displacement)
```

```
##    vars     n mean   sd median trimmed  mad min max range skew kurtosis   se
## X1    1 37979 3.32 1.36      3    3.21 1.48   0 8.4   8.4 0.62    -0.57 0.01
```

```
ggplot(fuel_narrowed, aes(x=engine_displacement, y=combined_mpg_ft1)) +
   geom_bar(stat="summary", fun="mean", show.legend = FALSE, fill="purple") +
  labs(x='Engine Displacement (liters)', y='Combined MPG')
```



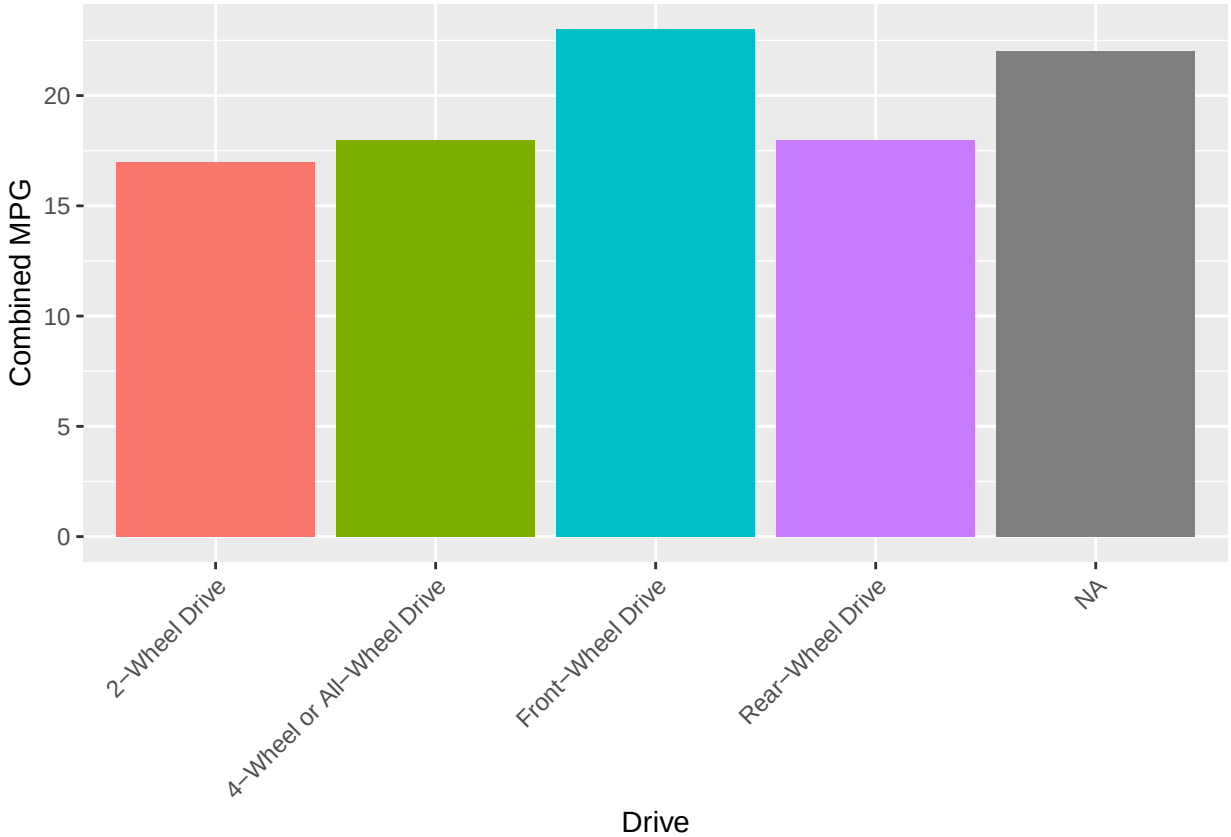
The graph above shows the relationship between an engines displacement in liters and its combined MPG, the relationship appears to be exponential decay with changes in MPG being more drastic as the displacement increases from a low value and leveling off over time.

```
categorical_cols <- c('drive', 'transmission', 'turbocharger', 'fuel_type')
categorical_summary <- list()

for (col in categorical_cols) {
  categorical_summary[[col]] <- table(fuel_narrowed[[col]], useNA = "ifany")
}

categorical_summary
```

```
## $drive
##
##              2-Wheel Drive 4-Wheel or All-Wheel Drive
##                        507                       10048
##          Front-Wheel Drive           Rear-Wheel Drive
##                      13351                       13018
##                       <NA>
##                       1189
##
## $transmission
##
##      Auto Automatic    Manual      <NA>
##       932     24745     12425        11
##
## $turbocharger
```

```
##
##  TRUE  <NA>
##  5239 32874
##
## $fuel_type
##
##                          CNG                       Diesel
##                           60                         1014
##                  Electricity             Gasoline or E85
##                          133                         1223
##     Gasoline or natural gas         Gasoline or propane
##                           20                            8
##                     Midgrade                      Premium
##                           77                        10133
##     Premium and Electricity  Premium Gas or Electricity
##                           25                           18
##              Premium or E85                      Regular
##                          122                        25258
## Regular Gas and Electricity  Regular Gas or Electricity
##                           20                            2
```

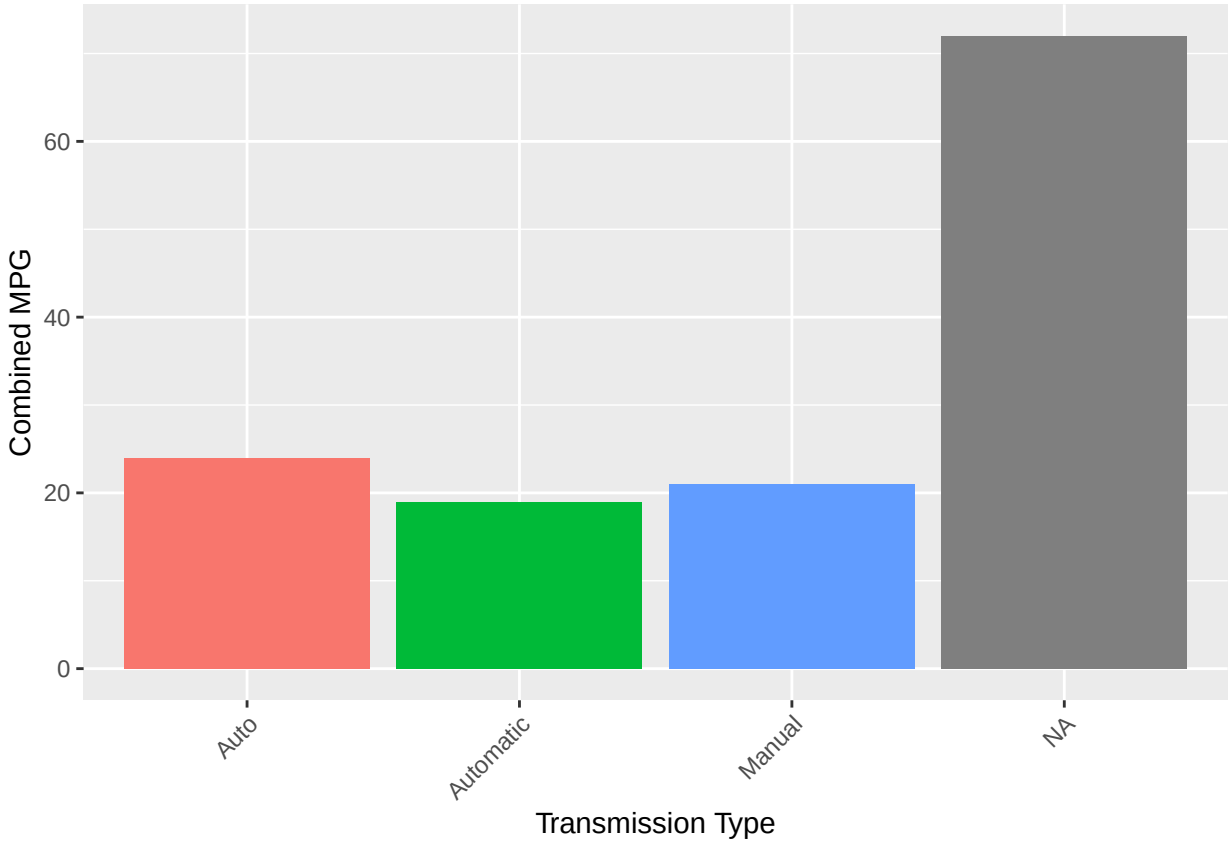
```
ggplot(fuel_narrowed, aes(x=drive, y=combined_mpg_ft1, fill=drive)) +
   geom_bar(stat="summary", fun="median", show.legend = FALSE) +
  labs(x='Drive', y='Combined MPG') +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



From this graph we can see that front wheel drive cars appear to have the highest combined MPG with cars without a drive named having around the same fuel millage as front wheel drive, and all other types having

around the same lower value.

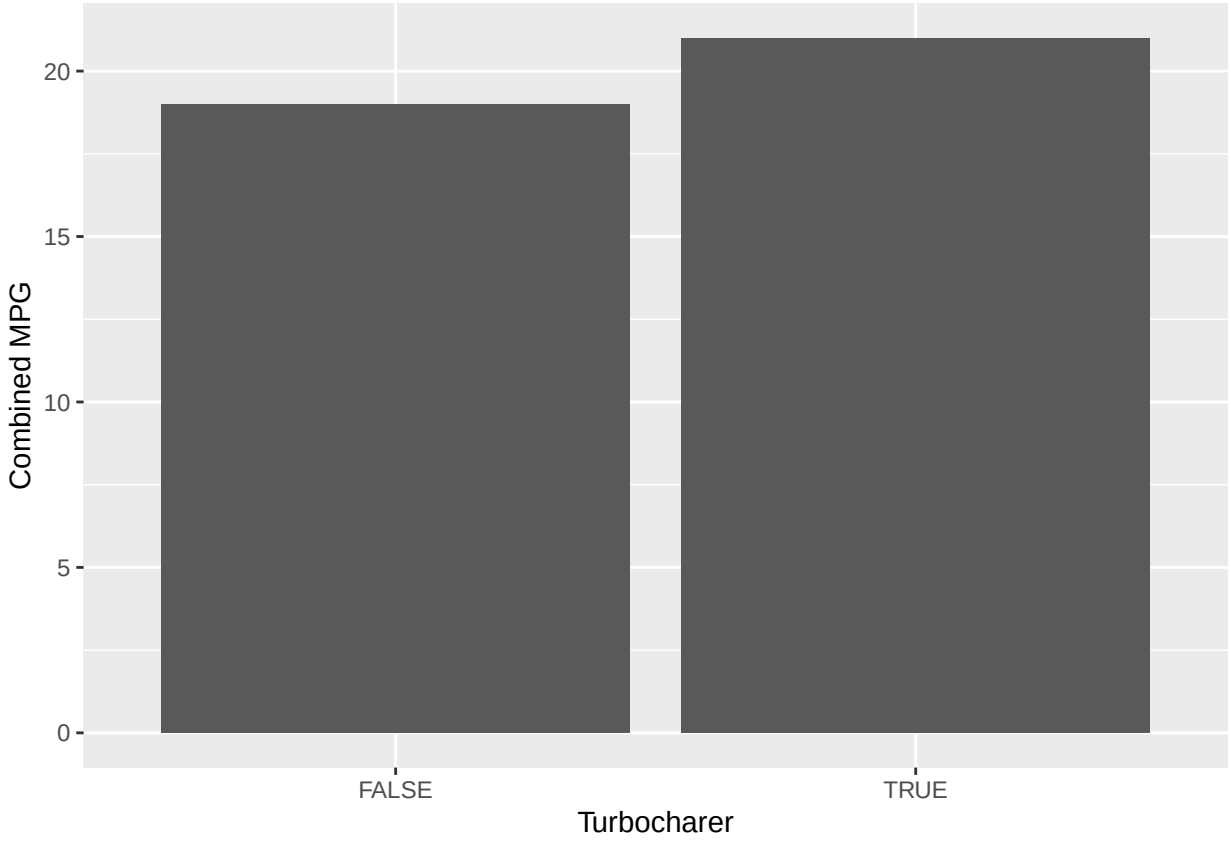
```
ggplot(fuel_narrowed, aes(x=transmission, y=combined_mpg_ft1, fill=transmission)) +
   geom_bar(stat="summary", fun="median", show.legend = FALSE) +
  labs(x='Transmission Type', y='Combined MPG') +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



This graph shows the relationship between transmission type and combined MPG, there does not appear to be much of a relationship between the two aside from that when transmission type is not given then there is a much higher combined MPG. However, after looking through the data this is because 9 of 11 of the NA's are for electric vehicles which is far more than would be expected.

```
fuel_narrowed$turbocharger[is.na(fuel_narrowed$turbocharger)] <- FALSE
```

```
ggplot(fuel_narrowed, aes(x=turbocharger, y=combined_mpg_ft1)) +
   geom_bar(stat="summary", fun="median", show.legend = FALSE) +
  labs(x='Turbocharer', y='Combined MPG')
```



This graph shows that cars with a turbocharger appear to have slightly higher combined fuel mileage than cars without a turbocharger.

Finding important columns

```
fuel_omitted <- na.omit(fuel_narrowed)
nrow(fuel_omitted)/nrow(fuel_narrowed)
```

```
## [1] 0.9653924
```

The first step is to make sure all of the data has no missing values. Since as we can see above over 95% of the rows has no missing values we can simply drop all rows where data is missing and use the remaining rows for analysis.

```
fuel_model <- lm(combined_mpg_ft1 ~ ., data = fuel_omitted)
summary(fuel_model)
```

```
##
## Call:
## lm(formula = combined_mpg_ft1 ~ ., data = fuel_omitted)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.5142 -1.6135 -0.2353  1.2248 28.7842
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       27.633277   0.378837  72.942  < 2e-16 ***
```

```
## year                                  0.145870   0.001578  92.451  < 2e-16 ***
## drive4-Wheel or All-Wheel Drive      -1.321503   0.124917 -10.579  < 2e-16 ***
## driveFront-Wheel Drive                1.903014   0.125177  15.203  < 2e-16 ***
## driveRear-Wheel Drive                -0.273194   0.123541  -2.211   0.0270 *
## transmissionAutomatic                -2.535514   0.090599 -27.986  < 2e-16 ***
## transmissionManual                   -1.842084   0.093446 -19.713  < 2e-16 ***
## engine_cylinders                     -0.399385   0.019360 -20.629  < 2e-16 ***
## engine_displacement                  -1.950753   0.025968 -75.120  < 2e-16 ***
## turbochargerTRUE                     -0.979241   0.045917 -21.327  < 2e-16 ***
## fuel_typeDiesel                       7.761830   0.353532  21.955  < 2e-16 ***
## fuel_typeGasoline or E85              0.177495   0.348904   0.509   0.6110
## fuel_typeGasoline or natural gas     -1.313229   0.680747  -1.929   0.0537 .
## fuel_typeGasoline or propane          0.230128   0.992405   0.232   0.8166
## fuel_typeMidgrade                     2.579960   0.454751   5.673 1.41e-08 ***
## fuel_typePremium                      0.563917   0.342067   1.649   0.0992 .
## fuel_typePremium and Electricity      3.999760   0.629442   6.354 2.12e-10 ***
## fuel_typePremium Gas or Electricity   5.855793   0.709652   8.252  < 2e-16 ***
## fuel_typePremium or E85               1.798713   0.416946   4.314 1.61e-05 ***
## fuel_typeRegular                      0.656037   0.341152   1.923   0.0545 .
## fuel_typeRegular Gas and Electricity 15.479837   0.681349  22.719  < 2e-16 ***
## fuel_typeRegular Gas or Electricity  14.782108   1.895152   7.800 6.36e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.636 on 36772 degrees of freedom
## Multiple R-squared:  0.7305, Adjusted R-squared:  0.7304
## F-statistic:  4747 on 21 and 36772 DF,  p-value: < 2.2e-16
```

```
reduced_fuel_model <- step(fuel_model, direction = "backward")
```

```
## Start:  AIC=71351.28
## combined_mpg_ft1 ~ year + drive + transmission + engine_cylinders +
##     engine_displacement + turbocharger + fuel_type
##
##                       Df Sum of Sq    RSS   AIC
## <none>                             255532 71351
## - engine_cylinders     1      2957 258490 71773
## - turbocharger         1      3161 258693 71802
## - transmission         2      8214 263746 72511
## - engine_displacement  1     39214 294747 76602
## - fuel_type           12     46001 301534 77418
## - drive                3     47895 303427 77666
## - year                 1     59395 314927 79039
```

```
reduced_fuel_model
```

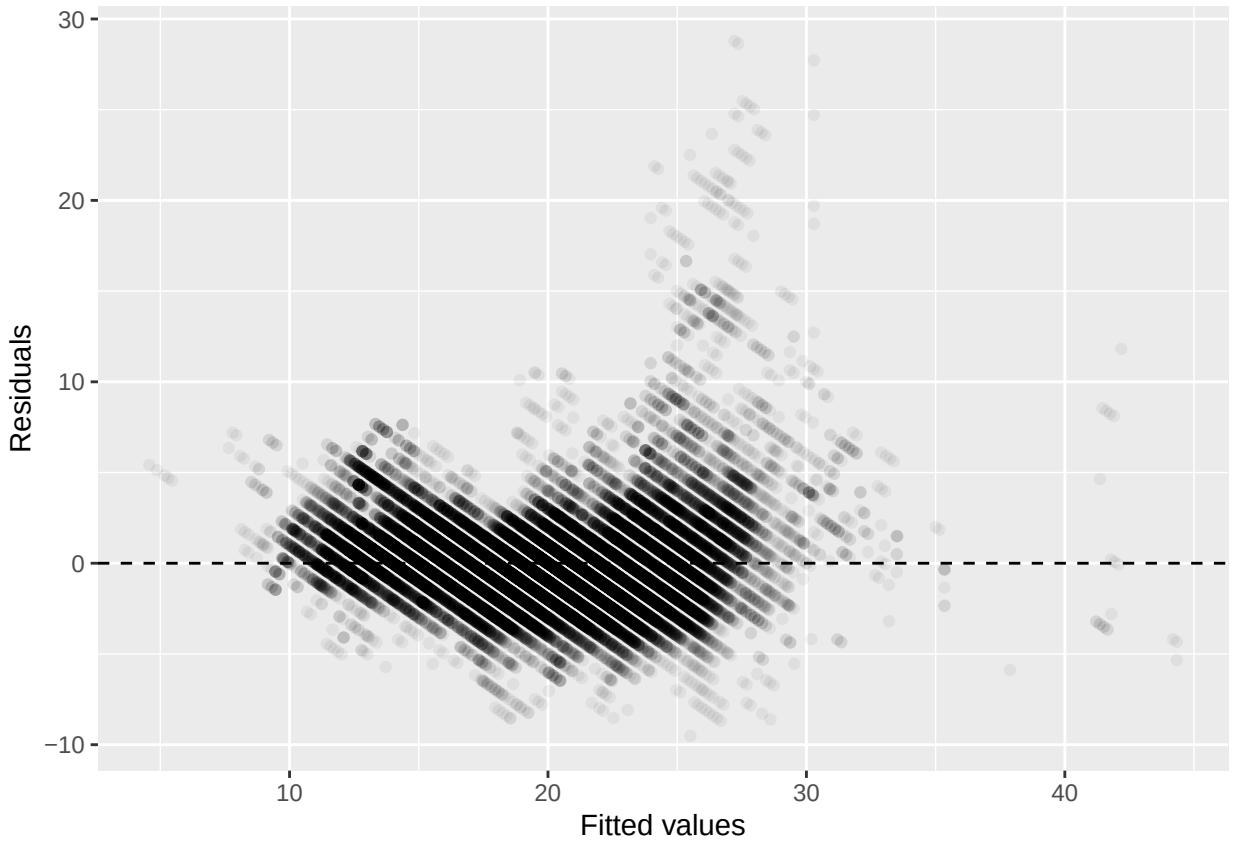
```
##
## Call:
## lm(formula = combined_mpg_ft1 ~ year + drive + transmission +
##     engine_cylinders + engine_displacement + turbocharger + fuel_type,
##     data = fuel_omitted)
##
## Coefficients:
##                           (Intercept)                                  year
##                               27.6333                                0.1459
```

```
##      drive4-Wheel or All-Wheel Drive                driveFront-Wheel Drive
##                              -1.3215                                1.9030
##                driveRear-Wheel Drive                transmissionAutomatic
##                              -0.2732                               -2.5355
##                   transmissionManual                     engine_cylinders
##                              -1.8421                               -0.3994
##                  engine_displacement                     turbochargerTRUE
##                              -1.9508                               -0.9792
##                      fuel_typeDiesel             fuel_typeGasoline or E85
##                               7.7618                                0.1775
##     fuel_typeGasoline or natural gas         fuel_typeGasoline or propane
##                              -1.3132                                0.2301
##                    fuel_typeMidgrade                     fuel_typePremium
##                               2.5800                                0.5639
##     fuel_typePremium and Electricity   fuel_typePremium Gas or Electricity
##                               3.9998                                5.8558
##              fuel_typePremium or E85                     fuel_typeRegular
##                               1.7987                                0.6560
## fuel_typeRegular Gas and Electricity   fuel_typeRegular Gas or Electricity
##                              15.4798                               14.7821
```

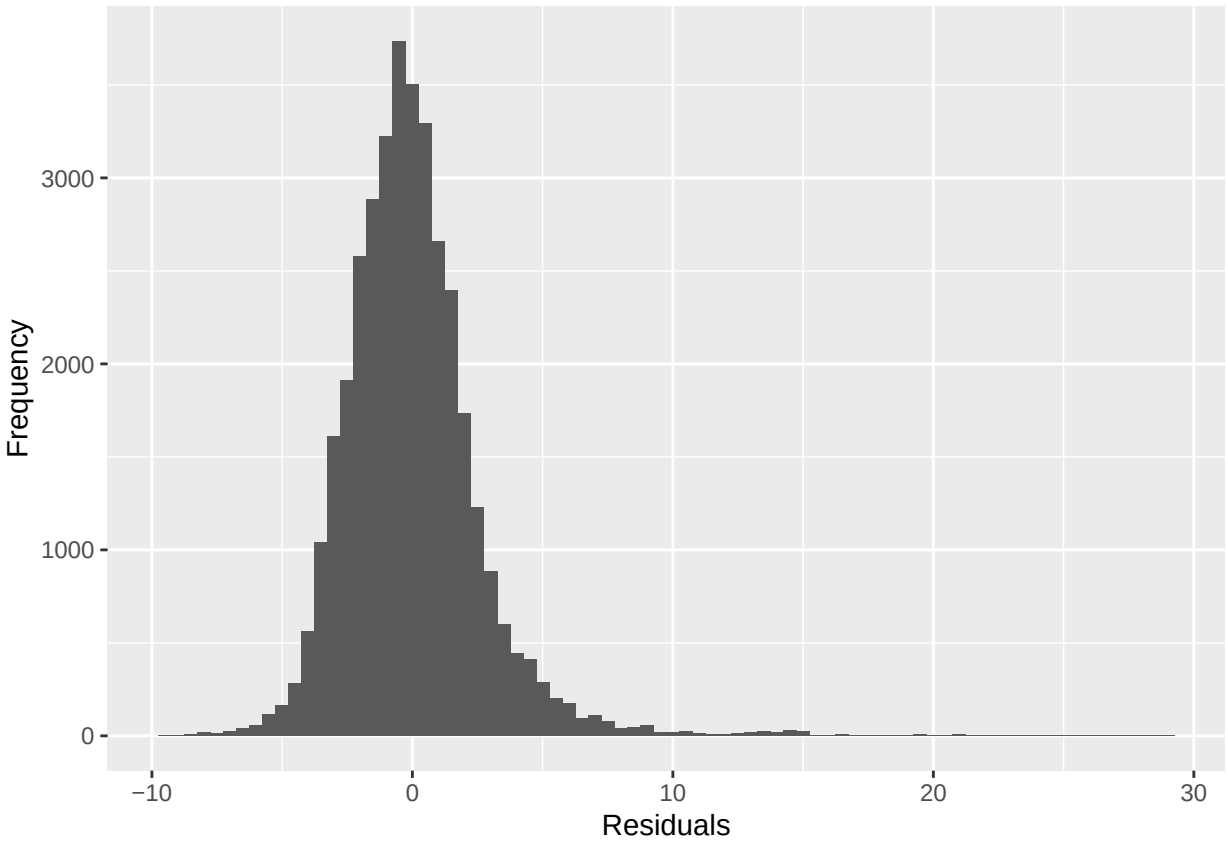
What we can see from the linear model summary and backwards steps is that there is not a single column that if removed would improve the performance of the linear model, therefore the model is good as is and does not need to be re-fitted to a new subset of columns. Then looking at the multiple and adjusted r-squared values which are essentially identical at 0.73 meaning that 73% of the variance can be explained by the predictors and as the two values are so similar, adding more predictors probably won't improve the accuracy.

Residuals

```
ggplot(data = fuel_model, aes(x = .fitted, y = .resid)) +
  geom_point(alpha=0.05) +
  geom_hline(yintercept = 0, linetype = "dashed") +
  xlab("Fitted values") +
  ylab("Residuals")
```




```
ggplot(data = fuel_model, aes(x = .resid)) +
  geom_histogram(binwidth = .5) +
  xlab("Residuals") +
  ylab("Frequency")
```



There are no clear patterns which stand out in the plot of residuals against fitted values, while there are some extreme outliers, with well over 30 thousand data points this is to be expected, and the vast majority of the points fall along the 0 line. For the histogram we can see an almost perfect normal distribution centered on 0 with a possible slight right skew from some outliers. Overall there do not appear to be any patterns in the residuals that would indicate that least squares regression is not appropriate.

Final Model

Predicted_MPG = 27.6333 + 0.1459 * year - 1.3215 * drive4wd + 1.9030 * drivefwd - 0.2732 * driverwd
- 2.5355 * transmissionAuto - 1.8421 * transmissionManual - 0.3994 * engine_cylinders
- 1.9508 * engine_displacement - 0.9792 * turbocharger
+ 7.7618 * fuel_typeDiesel + 0.1775 * fuel_typeGasolineOrE85 - 1.3132 * fuel_typeGasolineOrNaturalGas +
0.2301 * fuel_typeGasolineOrPropane + 2.5800 * fuel_typeMidgrade + 0.5639 * fuel_typePremium
+ 4.0000 * fuel_typePremiumAndElectricity + 5.8558 * fuel_typePremiumOrElectricity + 1.7987 *
fuel_typePremiumOrE85
+ 0.6560 * fuel_typeRegular + 15.4798 * fuel_typeRegularGasAndElectricity + 14.7821 * fuel_typeRegularGasOrElectricity

Conclusion

Through this analysis we have seen that almost all factors that go into the design of a vehicle can have some impact on the fuel economy and therefore environmental impact of the vehicle. Many of the factors above like fuel type, transmission type, engine cylinders, etc... are factors that manufacturers have full control over when designing the vehicle regardless of whether they're designing a truck, sedan, or any other vehicle type, so having a better understanding of what choices lead to better combined MPG can help these manufactures make better informed decisions on the impact of the vehicles they are designing. With all that in mind, this analysis has a few limitations, firstly this data only includes vehicles from 1984 to 2017. Another limitation is that this data set is based purely on US vehicles and US fuel economy standards, any vehicles sold exclusively

outside of the US would not be included, and other countries will have different methods of measuring fuel economy which may or may not produce the same results as the analysis based on the US Department of Energy's Combined MPG standard.