Statistics Project

Aayush Agrawal, Wenrui Liu, Byron Tang, and Russell Nour

Friday, August 19, 2016

Data Import and Manipulation

Data were downloaded by year from the US DOE, with each year of data as its own dataset in a separate comma delimited (CSV) file. The datasets for each year were appended to each other using a script in the R statistical computing language. Only those variables that were potentially relevant to the analysis were kept. In addition, electric and hydrogen-fueled vehicles were removed from the data. The fuel efficiency of these types of vehicles are not measured in miles per gallon, so in order to ensure an "apples to apples" comparison of cars by miles per gallon, only vehicles that use gasoline as fuel were included in the analysis. Similarly, for hybrid vehicles only the miles per gallon ratings for the gasoline/diesel operation of the engine were kept. Miles per gallon ratings were averaged for each vehicle make and model. The cleaned data enables a better understanding of trends in the vehicle market over the years. Different MPG performances across distinct car features could also be identified through descriptive and inferential statistics.

```r
#Initializing Library
library(readxl)
library(data.table)
library(ggplot2)
library(corrplot)
#Taking files with xlsx format
temp = list.files(pattern="*.xlsx")

#Collating
data <- NULL
data_colnames <-  colnames(read_excel(temp[1])[,1:17])
for (i in 1:length(temp)) {
  x <-  read_excel(temp[i])[,1:17]
  attr(x, "rownames") <- NULL
  colnames(x) <- data_colnames
  x[,'file'] <- temp[i]
  data <- rbind(data,x)
}

#Removing Duplicates
data <- data[!is.na(data[,'Cmb MPG']),]

#Cleansing data
data$year <- substring(text = data$file,first = 30,33)
data$Transmission_type <- sapply(strsplit(data$Trans,split = "-"), function(x) (x[1]))
data$Transmission_number <- gsub("^.*?-","",data$Trans)

#Removing unncessary columns
data <- data[,-c(4,7,8,9,10,18)]

#Writing clean data
write.csv(data,'Collated.csv',row.names = FALSE)

#Reading
```

```r
data <- read.csv('Collated.csv')

#Cleaning Data
data$SmartWay[data$SmartWay == 'yes'] = 'Yes'
data$Fuel[data$Fuel == 'Gasoline/Electricity'] = 'Gas/Electricity'
data$Fuel[data$Fuel == 'Gasoline/Electricty'] = 'Gas/Electricity'
data$Fuel[data$Fuel == 'Electricity/Gas'] = 'Gas/Electricity'
data$Veh.Class[data$Veh.Class == 'small SUV' | data$Veh.Class == 'standard SUV' ] = 'SUV'

data$Model <- as.character(data$Model)
data$Displ <- as.numeric(data$Displ)
data$Cyl <- as.numeric(data$Cyl)
data$'Air.Pollution.Score' <- as.numeric(data$'Air.Pollution.Score')

#Converting Transmission CVT to 1
data[data$Transmission_number == 'CVT','Transmission_number'] <- 1
data$Transmission_number <- as.numeric(data$Transmission_number)

# Removing Hydrogen and electric based cars
data <- data[!(data$Fuel == 'Hydrogen' | data$Fuel == 'Electricity' | data$Fuel == "CNG"),]
data$Fuel <- factor(data$Fuel, levels = c("Diesel", "Ethanol/Gas", "Gas/Electricity","Gasoline"))
data$Veh.Class <- factor(data$Veh.Class, levels = c("large car","midsize car","small car","station wagon

#For City MPG
data$'City.MPG' <- as.character(data$'City.MPG')
data[data$Fuel == 'Ethanol/Gas',]$"City.MPG" <- gsub("^.*?/","",data[data$Fuel == 'Ethanol/Gas',]$"City
data[data$Fuel == 'Electricity/Gas',]$"City.MPG" <- gsub("^.*?/","",data[data$Fuel == 'Electricity/Gas'
data[data$Fuel == 'Gas/Electricity',]$"City.MPG" <- sapply(strsplit(data[data$Fuel == 'Gas/Electricity'
data[data$Fuel == 'Gasoline/Electricity',]$"City.MPG" <- sapply(strsplit(data[data$Fuel == 'Gasoline/El

#For Hwy.MPG
data$'Hwy.MPG' <- as.character(data$'Hwy.MPG')
data[data$Fuel == 'Ethanol/Gas',]$"Hwy.MPG" <- gsub("^.*?/","",data[data$Fuel == 'Ethanol/Gas',]$"Hwy.MP
data[data$Fuel == 'Electricity/Gas',]$"Hwy.MPG" <- gsub("^.*?/","",data[data$Fuel == 'Electricity/Gas',]
data[data$Fuel == 'Gas/Electricity',]$"Hwy.MPG" <- sapply(strsplit(data[data$Fuel == 'Gas/Electricity',]
data[data$Fuel == 'Gasoline/Electricity',]$"Hwy.MPG" <- sapply(strsplit(data[data$Fuel == 'Gasoline/Elec

#ForCmb.MPG
data$'Cmb.MPG' <- as.character(data$'Cmb.MPG')
data[data$Fuel == 'Ethanol/Gas',]$"Cmb.MPG" <- gsub("^.*?/","",data[data$Fuel == 'Ethanol/Gas',]$"Cmb.MP
data[data$Fuel == 'Electricity/Gas',]$"Cmb.MPG" <- gsub("^.*?/","",data[data$Fuel == 'Electricity/Gas',]
data[data$Fuel == 'Gas/Electricity',]$"Cmb.MPG" <- sapply(strsplit(data[data$Fuel == 'Gas/Electricity',]
data[data$Fuel == 'Gasoline/Electricity',]$"Cmb.MPG" <- sapply(strsplit(data[data$Fuel == 'Gasoline/Elec

#For Greenhouse.Gas.Score
data$'Greenhouse.Gas.Score' <- as.character(data$'Greenhouse.Gas.Score')
data[data$Fuel == 'Ethanol/Gas',]$"Greenhouse.Gas.Score" <- gsub("^.*?/","",data[data$Fuel == 'Ethanol/G
data[data$Fuel == 'Electricity/Gas',]$"Greenhouse.Gas.Score" <- gsub("^.*?/","",data[data$Fuel == 'Elect
data[data$Fuel == 'Gas/Electricity',]$"Greenhouse.Gas.Score" <- sapply(strsplit(data[data$Fuel == 'Gas/E
data[data$Fuel == 'Gasoline/Electricity',]$"Greenhouse.Gas.Score" <- sapply(strsplit(data[data$Fuel ==
data$"Greenhouse.Gas.Score" <- as.numeric(data$"Greenhouse.Gas.Score")

#Chaging type of variables
```

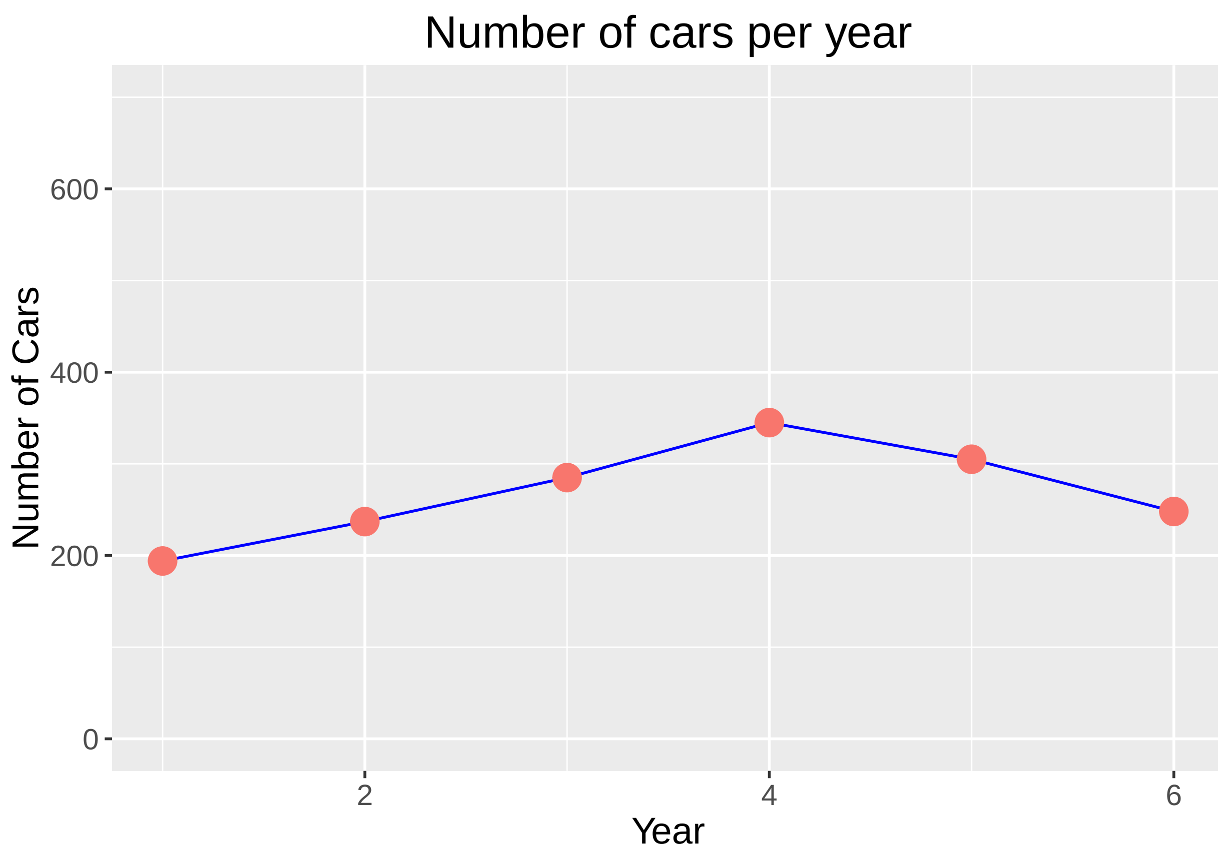
```
data$"City.MPG" <- as.numeric(data$"City.MPG")
data$"Hwy.MPG" <- as.numeric(data$"Hwy.MPG")
data$"Cmb.MPG" <- as.numeric(data$"Cmb.MPG")

#Grouping to remove duplicates
DT <- data.table(data)
DT1 <- DT[,.(Hwy.MPG.mean = mean(Hwy.MPG),City.MPG.mean = mean(City.MPG), Cmb.MPG.mean = mean(Cmb.MPG),

#Converting in a dataframe
data1 <- as.data.frame(DT1)

#Outlier treatment : Removing Chevrolet Volt
data1 <- data1[!data1$Cmb.MPG>90,]

#Releveling some factors for better beta coefficients interpretability
data1$Veh.Class <- relevel(data1$Veh.Class,"small car")
data1$Transmission_type <- relevel(data1$Transmission_type,"Auto")
```

Analyses

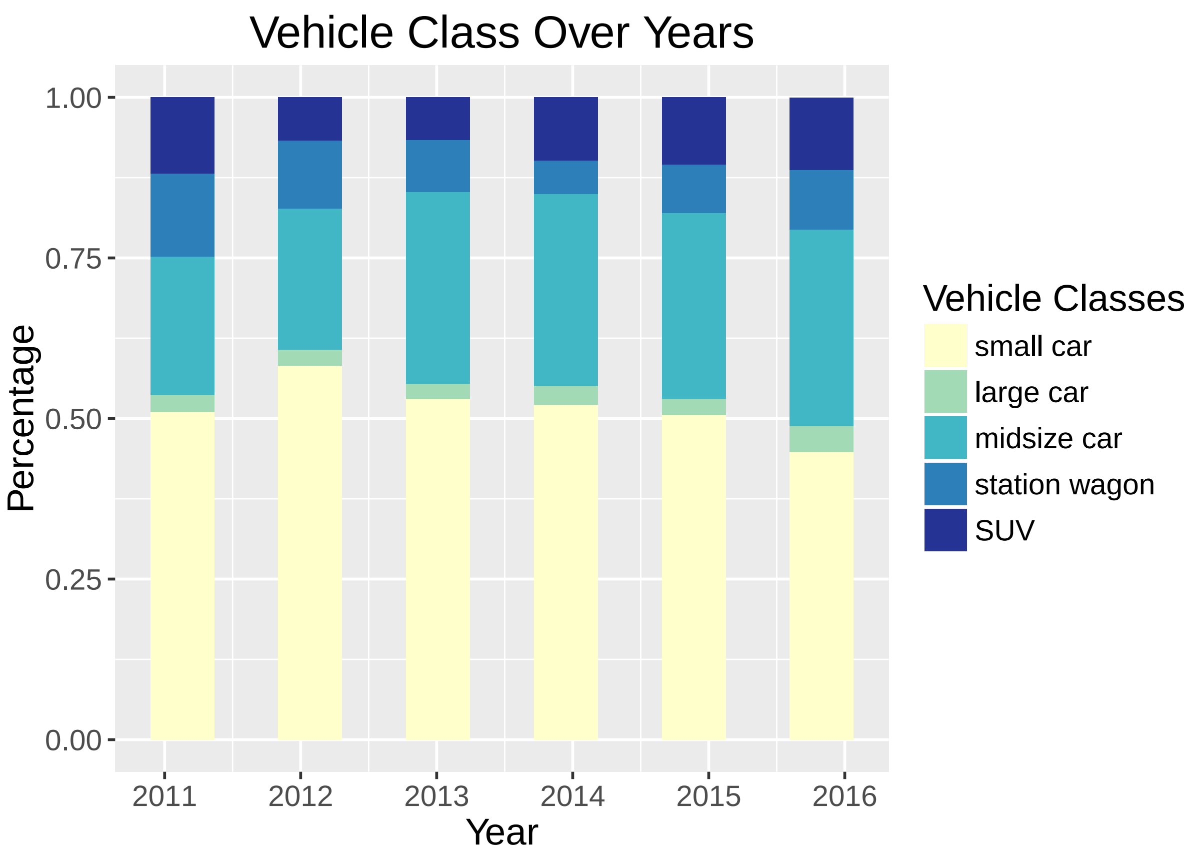
Vehicle Market Overview - 2011 to 2016

New vehicle launches grew steadily in the past years but peaked in year 2014. Small sedans are consistently the largest share in the market. However, the share of medium size sedans continues to grow and take market share from small sedans.

```
x = data.frame(table(data1$year))
colnames(x) <- c("year","N_cars")
x$year <- as.numeric(x$year)
ggplot(x,aes(x=year, y = N_cars)) + geom_line(data = x,aes(x = year, y = N_cars),colour = "blue") + xlab
```

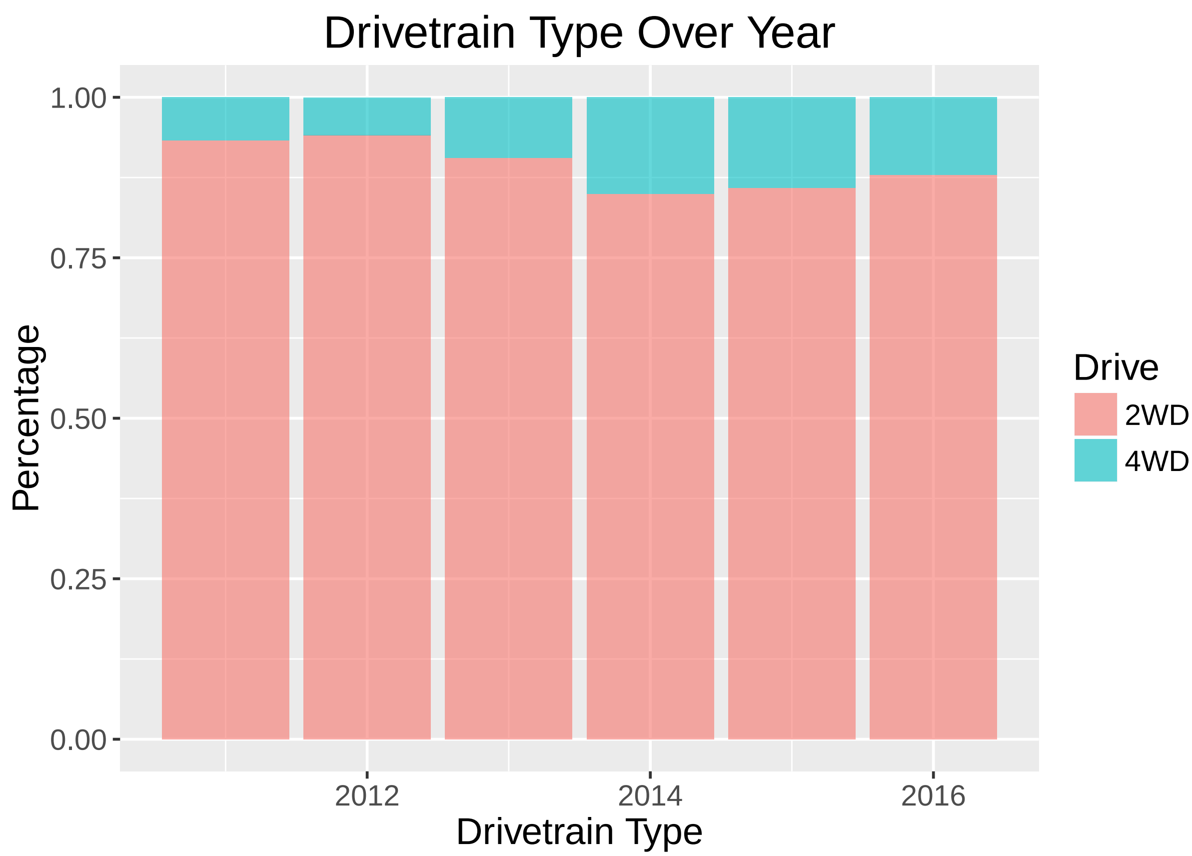



```
ggplot(data1,aes(x = year,fill = Veh.Class),geom="text") + geom_histogram(position = "fill", binwidth =
```

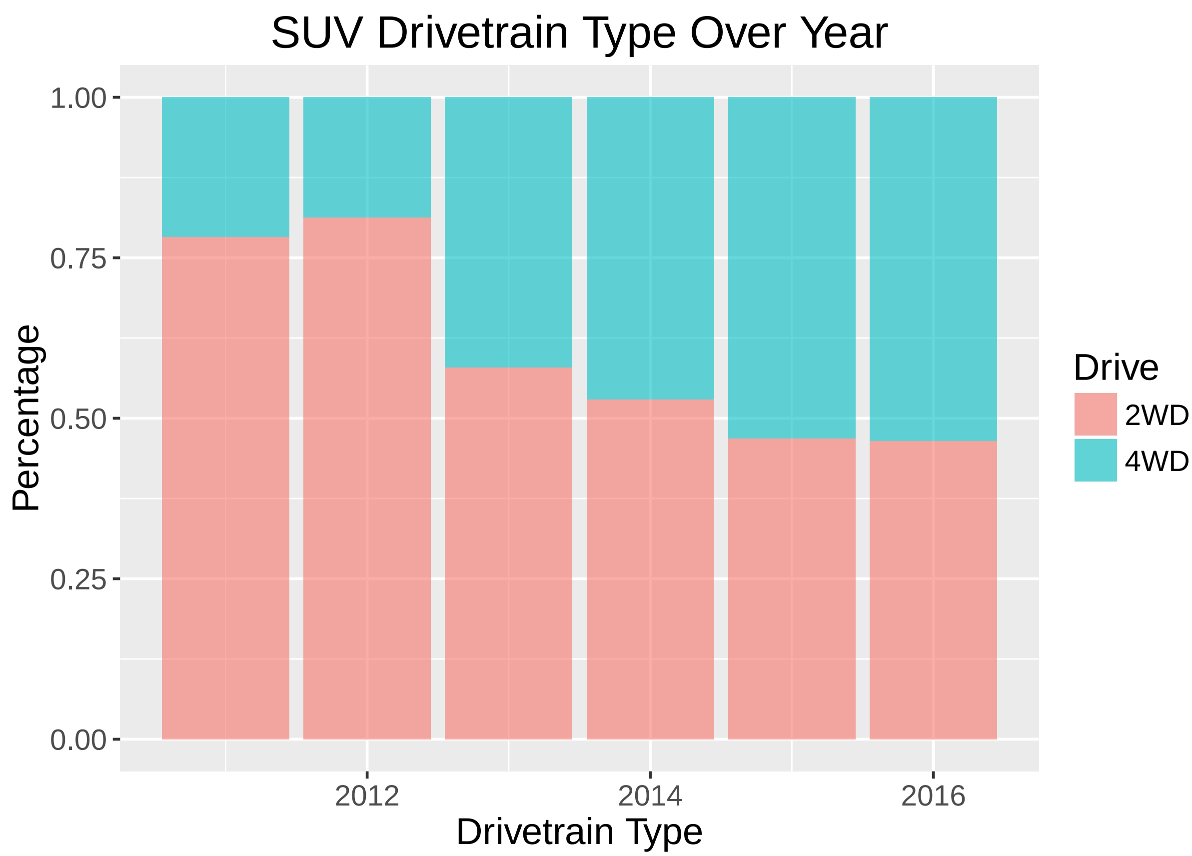


Two-wheel drive is the most common drivetrain type, but more and more four-wheel drive vehicles are coming to the market in recent years. It is a trend that manufacturers are adopting four-wheel drive in their SUVs to meet the needs for various road conditions.

```
ggplot(data1, aes(x = year, fill = Drive))+geom_bar(alpha=0.6, position="fill",  stat = "count" ) + labs
```



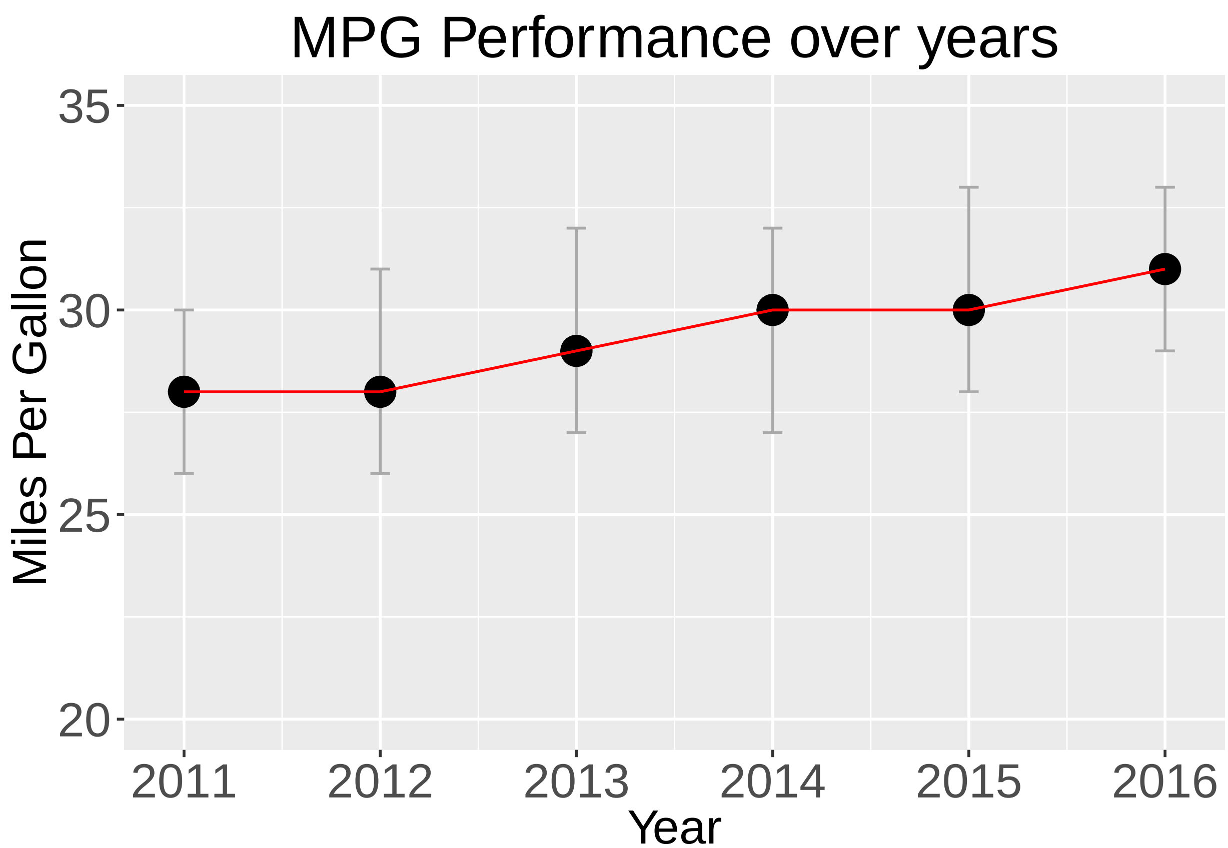

```
data1_SUV <- data1[data1$Veh.Class == 'SUV', ]
ggplot(data1_SUV, aes(x = year, fill = Drive))+geom_bar(alpha=0.6, position="fill",  stat = "count" ) +
```



Miles per Gallon

The median combined miles per gallon increases slightly every year. The difference is not statistically significant from year to year, but median combined MPG in 2015 and 2016 were significantly higher when compared to 2011. The chart shows the trend that combined MPG increases by year.

```
x = DT1[,.(Cmb.MPG.median = median(Cmb.MPG.mean), Cmb.MPG.25 = quantile(Cmb.MPG.mean,0.25),Cmb.MPG.75 =
ggplot(x, aes(x=year, y=Cmb.MPG.median)) +  expand_limits(y=c(20,35)) +
  geom_errorbar(aes(ymin=Cmb.MPG.25, ymax=Cmb.MPG.75), colour="darkgrey", width=.1) +
  geom_point(size = 5,colour = "black")+ geom_line(colour = "red") +
  labs(title="MPG Performance over years", x="Year", y="Miles Per Gallon") +theme(text = element_text(s
```



Anova test -

```
fit <- aov(data1$Cmb.MPG.mean ~ factor(data1$year))
summary(fit)
```

```
##                       Df Sum Sq Mean Sq F value  Pr(>F)
## factor(data1$year)     5   1367  273.39   12.18 1.3e-11 ***
## Residuals           1608  36080   22.44
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Tukey HSD test at 95% confidence interval-

```
TukeyHSD(fit, conf.level = .95)
```

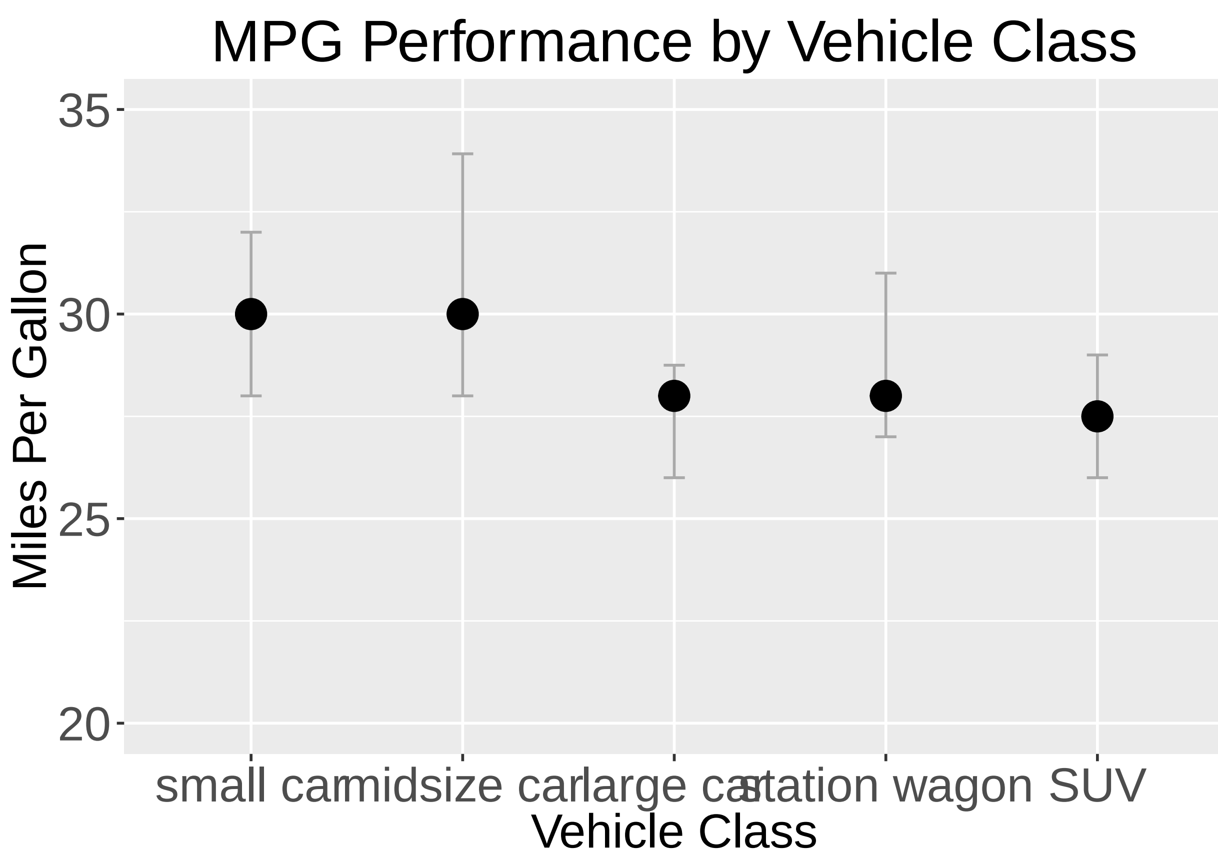
```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = data1$Cmb.MPG.mean ~ factor(data1$year))
##
## $`factor(data1$year)`
##                diff          lwr      upr     p adj
## 2012-2011  0.4812012 -0.827313286 1.789716 0.9010468
## 2013-2011  1.7511545  0.493216091 3.009093 0.0010489
## 2014-2011  1.6161990  0.403371914 2.829026 0.0020604
## 2015-2011  2.1831840  0.942062227 3.424306 0.0000086
## 2016-2011  2.9668898  1.671503302 4.262276 0.0000000
## 2013-2012  1.2699534  0.081851791 2.458055 0.0281268
```

```
## 2014-2012  1.1349979 -0.005232747 2.275228 0.0518667
## 2015-2012  1.7019829  0.531701032 2.872265 0.0005012
## 2016-2012  2.4856886  1.258006914 3.713370 0.0000001
## 2014-2013 -0.1349555 -1.216771055 0.946860 0.9992527
## 2015-2013  0.4320295 -0.681414984 1.545474 0.8786064
## 2016-2013  1.2157352  0.042107794 2.389363 0.0372718
## 2015-2014  0.5669850 -0.495229243 1.629199 0.6494658
## 2016-2014  1.3506907  0.225549923 2.475832 0.0082625
## 2016-2015  0.7837057 -0.371878842 1.939290 0.3810355
```

Miles per Gallon by Vehicle Class

Small and medium size cars have the highest median of combined MPG ratings. For both small and midsize cars, the median of combined MPG rating is significantly better than other types of cars, including large cars, station wagons, and SUVs.

```
x = DT1[,.(Cmb.MPG.median = median(Cmb.MPG.mean), Cmb.MPG.25 = quantile(Cmb.MPG.mean,0.25),Cmb.MPG.75 =
x$Veh.Class = ordered(x$Veh.Class, levels = c("small car","midsize car","large car","station wagon","SU
ggplot(x, aes(x=Veh.Class, y=Cmb.MPG.median)) +  expand_limits(y=c(20,35)) +
  geom_errorbar(aes(ymin=Cmb.MPG.25, ymax=Cmb.MPG.75), colour="darkgrey", width=.1) +
  geom_point(size = 5,colour = "black")+
  labs(title="MPG Performance by Vehicle Class", x="Vehicle Class", y="Miles Per Gallon") +theme(text =
```



Anova -

```
performance_year <- aov(data1$Cmb.MPG.mean~data1$Veh.Class)
summary(performance_year)
```

```
##                   Df Sum Sq Mean Sq F value Pr(>F)
## data1$Veh.Class    4   2744   686.0    31.8 <2e-16 ***
## Residuals       1609  34703    21.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Tukey HSD -

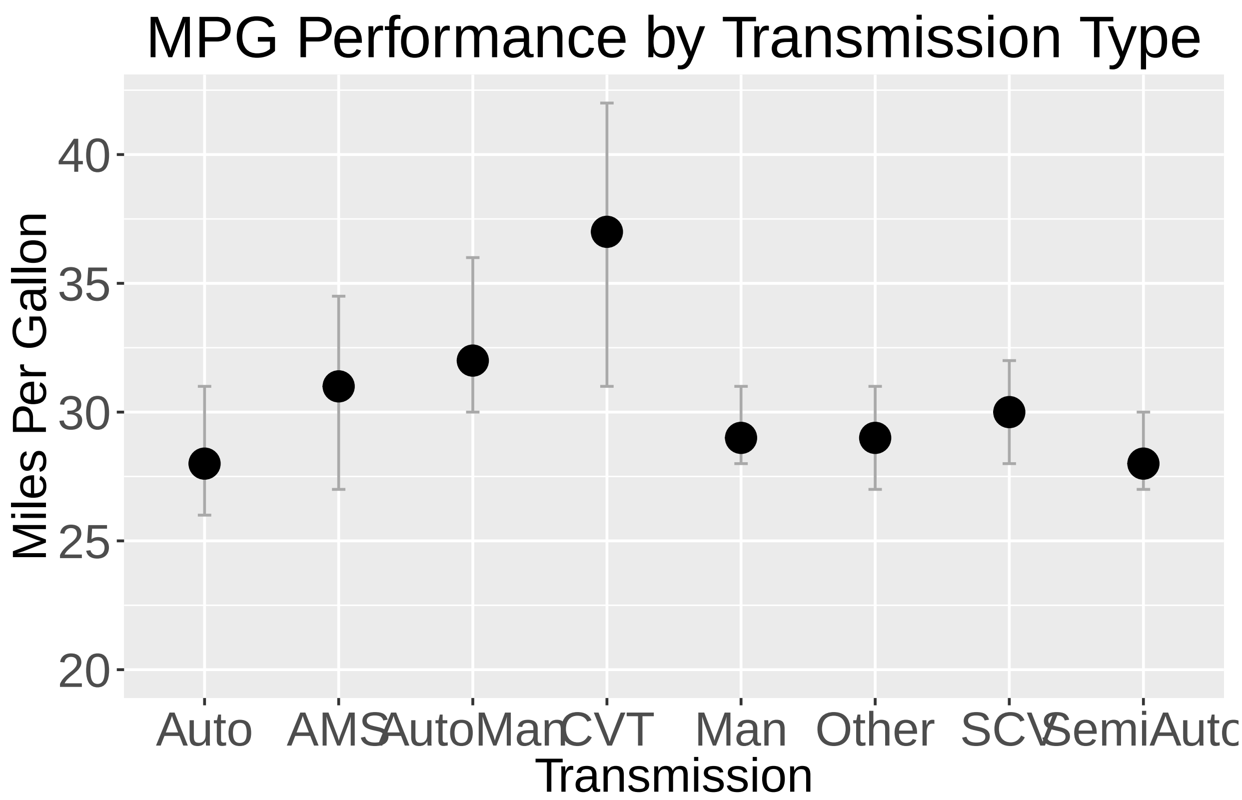
```
TukeyHSD(performance_year, conf.level = 0.95)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = data1$Cmb.MPG.mean ~ data1$Veh.Class)
##
## $`data1$Veh.Class`
##                                  diff        lwr         upr     p adj
## large car-small car       -1.8872332 -3.8081133  0.03364699 0.0569143
## midsize car-small car      1.5948686  0.8507322  2.33900509 0.0000001
## station wagon-small car   -1.0727298 -2.2419878  0.09652812 0.0898837
## SUV-small car             -2.9116897 -4.0303058 -1.79307357 0.0000000
## midsize car-large car      3.4821018  1.5180929  5.44611068 0.0000139
## station wagon-large car    0.8145033 -1.3466899  2.97569653 0.8418909
## SUV-large car             -1.0244565 -3.1586763  1.10976321 0.6845017
## station wagon-midsize car -2.6675984 -3.9064344 -1.42876246 0.0000000
## SUV-midsize car           -4.5065583 -5.6977142 -3.31540237 0.0000000
## SUV-station wagon         -1.8389599 -3.3330407 -0.34487904 0.0070968
```

Miles per Gallon by Transmission Type

Among all transmission types, continuously variable transmission (CVT) is the most efficient with a significantly higher MPG, getting on average an additional 5 to 8 miles per gallon than other transmission types.

```
x = DT1[,.(Cmb.MPG.median = median(Cmb.MPG.mean), Cmb.MPG.25 = quantile(Cmb.MPG.mean,0.25),Cmb.MPG.75 =
x$Transmission_type <- relevel(x$Transmission_type,"Auto")
ggplot(x, aes(x=Transmission_type, y=Cmb.MPG.median)) +  expand_limits(y=c(20,35)) +
  geom_errorbar(aes(ymin=Cmb.MPG.25, ymax=Cmb.MPG.75), colour="darkgrey", width=.1) +
  geom_point(size = 5,colour = "black")+
  labs(title="MPG Performance by Transmission Type", x="Transmission\n", y="Miles Per Gallon") +theme(t
```



Anova-

```r
performance_year <- aov(data1$Cmb.MPG.mean~data1$Transmission_type)
summary(performance_year)
```

```
##                           Df Sum Sq Mean Sq F value Pr(>F)    
## data1$Transmission_type    7  12467  1781.0   114.5 <2e-16 ***
## Residuals               1606  24980    15.6                   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Tukey HSD-

```r
TukeyHSD(performance_year, conf.level = 0.95)
```

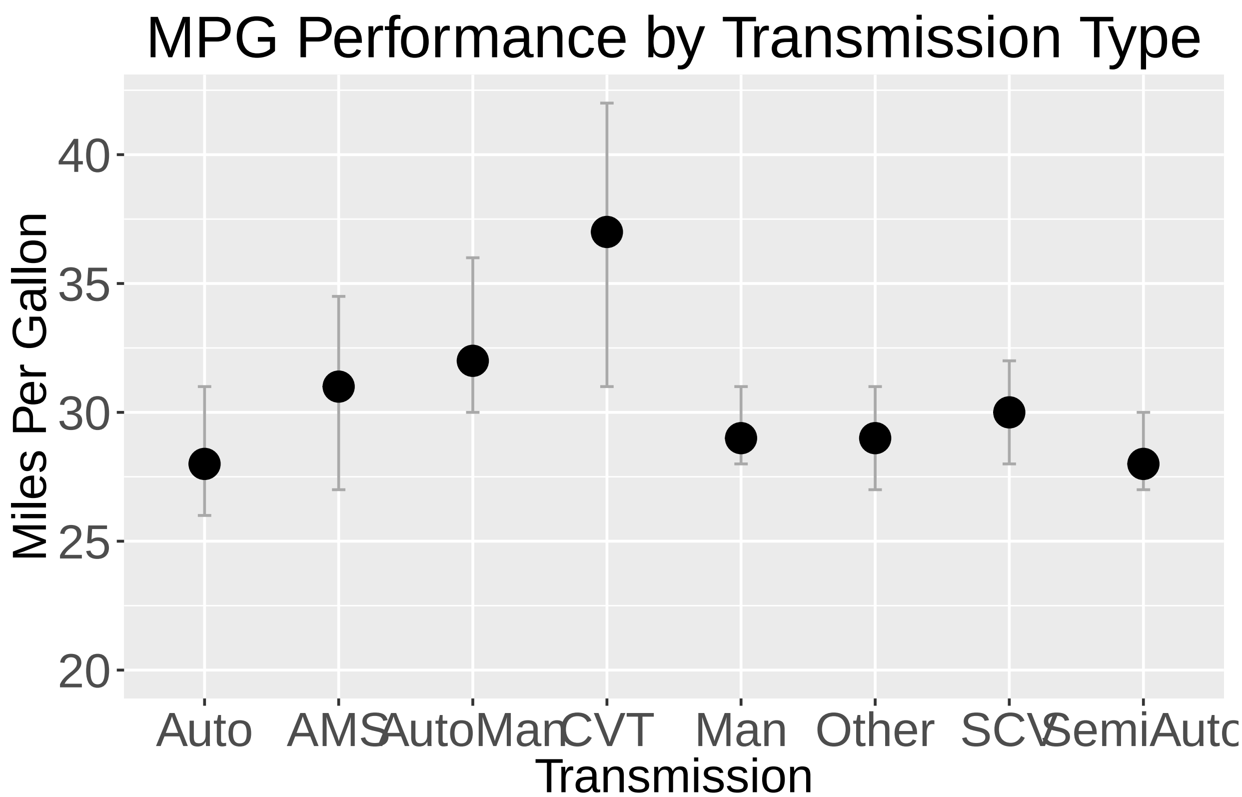
```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = data1$Cmb.MPG.mean ~ data1$Transmission_type)
##
## $`data1$Transmission_type`
##                          diff         lwr         upr     p adj
## AMS-Auto          3.423060867   1.6527714  5.19335037 0.0000001
## AutoMan-Auto      3.940886700   2.4071210  5.47465242 0.0000000
## CVT-Auto          8.423591357   7.2495686  9.59761412 0.0000000
## Man-Auto          0.811057986  -0.1822395  1.80435546 0.2055166
## Other-Auto        1.235316408  -2.1890076  4.65964045 0.9579966
## SCV-Auto          2.528051450   1.1024833  3.95361956 0.0000023
```

```
## SemiAuto-Auto     0.001514106  -1.0212392  1.02426744 1.0000000
## AutoMan-AMS       0.517825833  -1.5008123  2.53646395 0.9941936
## CVT-AMS           5.000530490   3.2396239  6.76143708 0.0000000
## Man-AMS          -2.612002880  -4.2579380 -0.96606776 0.0000437
## Other-AMS        -2.187744459  -5.8549583  1.47946943 0.6129687
## SCV-AMS          -0.895009416  -2.8327161  1.04269722 0.8566773
## SemiAuto-AMS     -3.421546761  -5.0854238 -1.75766968 0.0000000
## CVT-AutoMan       4.482704657   2.9597784  6.00563091 0.0000000
## Man-AutoMan      -3.129828713  -4.5182144 -1.74144303 0.0000000
## Other-AutoMan    -2.705570292  -6.2646390  0.85349841 0.2901802
## SCV-AutoMan      -1.412835249  -3.1371378  0.31146732 0.2017475
## SemiAuto-AutoMan -3.939372594  -5.3489822 -2.52976297 0.0000000
## Man-CVT          -7.612533370  -8.5890102 -6.63605656 0.0000000
## Other-CVT        -7.188274949 -10.6077577 -3.76879221 0.0000000
## SCV-CVT          -5.895539906  -7.3094393 -4.48164050 0.0000000
## SemiAuto-CVT     -8.422077251  -9.4285023 -7.41565217 0.0000000
## Other-Man         0.424258421  -2.9374633  3.78598010 0.9999438
## SCV-Man           1.716993464   0.4491520  2.98483491 0.0010799
## SemiAuto-Man     -0.809543881  -1.5977075 -0.02138027 0.0391550
## SCV-Other         1.292735043  -2.2210631  4.80653316 0.9533389
## SemiAuto-Other   -1.233802302  -4.6043449  2.13674025 0.9545752
## SemiAuto-SCV     -2.526537345  -3.8175859 -1.23548876 0.0000001
```

Miles per Gallon by Number of Transmissions

A single transmission delivers the best MPG performance. Vehicles with one transmission get 5 to 9 miles more miles per gallon than those with more transmissions. Variation also exists between different numbers of transmissions but are much less obvious.

```
x = DT1[,.(Cmb.MPG.median = median(Cmb.MPG.mean), Cmb.MPG.25 = quantile(Cmb.MPG.mean,0.25),Cmb.MPG.75 =
x$Transmission_type <- relevel(x$Transmission_type,"Auto")
ggplot(x, aes(x=Transmission_type, y=Cmb.MPG.median)) +  expand_limits(y=c(20,35)) +
  geom_errorbar(aes(ymin=x$Cmb.MPG.25, ymax=x$Cmb.MPG.75), colour="darkgrey", width=.1) +
  geom_point(size = 5,colour = "black")+
  labs(title="MPG Performance by Transmission Type", x="Transmission\n", y="Miles Per Gallon") +theme(t
```



Anova-

```
performance_year <- aov(data1$Cmb.MPG.mean~factor(data1$Transmission_number))
summary(performance_year)
```

```
##                                      Df Sum Sq Mean Sq F value Pr(>F)
## factor(data1$Transmission_number)     6  10794  1799.1   108.5 <2e-16 ***
## Residuals                          1607  26652    16.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Tukey HSD-

```
TukeyHSD(performance_year, conf.level = 0.95)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = data1$Cmb.MPG.mean ~ factor(data1$Transmission_number))
##
## $`factor(data1$Transmission_number)`
##            diff         lwr         upr     p adj
## 2-1 -7.53299629  -9.5304201 -5.53557249 0.0000000
## 3-1 -7.34911355  -8.5023882 -6.19583886 0.0000000
## 4-1 -7.15392075  -8.0383277 -6.26951384 0.0000000
## 5-1 -5.17055196  -6.6422940 -3.69880989 0.0000000
## 6-1 -8.74823370 -10.2090524 -7.28741503 0.0000000
## 7-1 -8.64927536 -14.0839313 -3.21461942 0.0000583
```

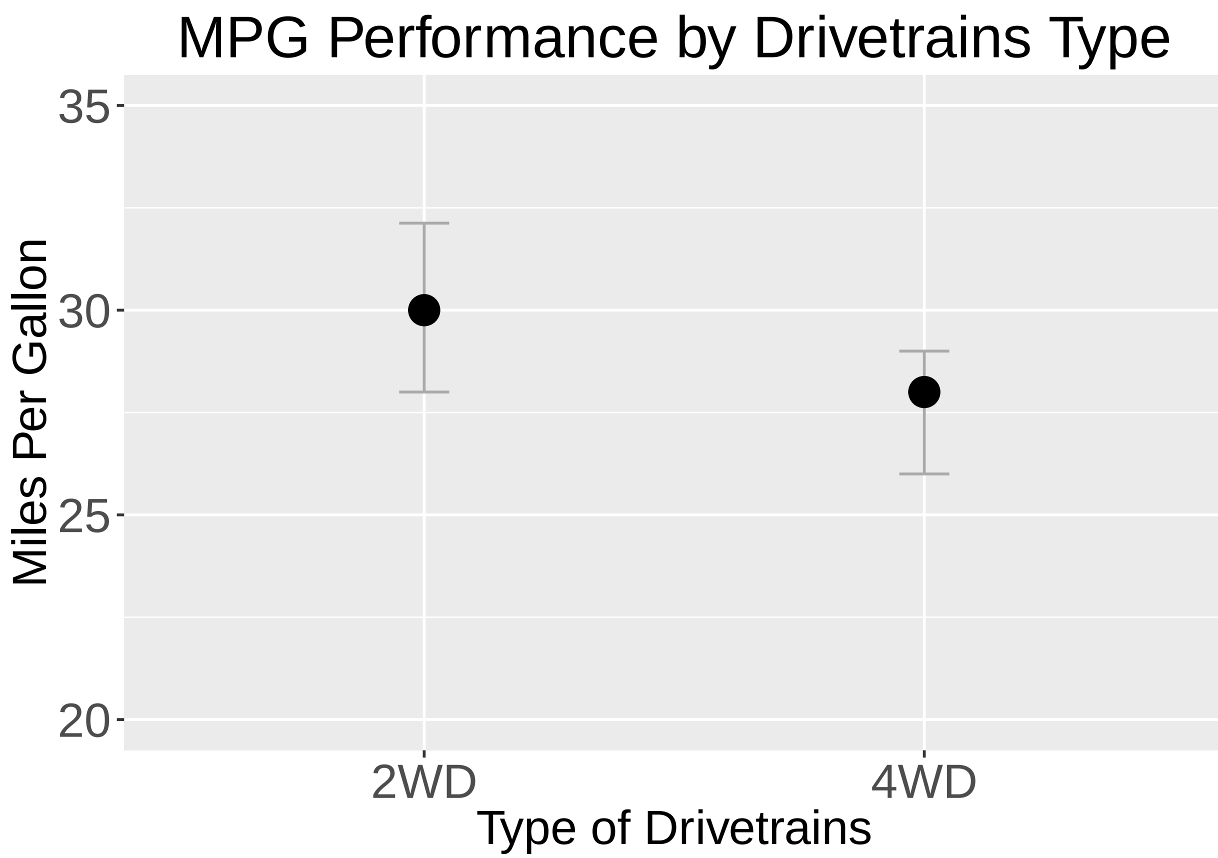
```
## 3-2  0.18388274  -1.8317847  2.19955018 0.9999688
## 4-2  0.37907554  -1.4957707  2.25392175 0.9969139
## 5-2  2.36244433   0.1490990  4.57578964 0.0275845
## 6-2 -1.21523740  -3.4213344  0.99085958 0.6653529
## 7-2 -1.11627907  -6.7968035  4.56424537 0.9973663
## 4-3  0.19519280  -0.7296796  1.12006519 0.9960877
## 5-3  2.17856159   0.6821532  3.67496998 0.0003662
## 6-3 -1.39912015  -2.8847865  0.08654621 0.0803168
## 7-3 -1.30016181  -6.7415494  4.14122575 0.9923062
## 5-4  1.98336879   0.6828415  3.28389605 0.0001464
## 6-4 -1.59431294  -2.8824657 -0.30616015 0.0049648
## 7-4 -1.49535461  -6.8861645  3.89545532 0.9830800
## 6-5 -3.57768174  -5.3221549 -1.83320854 0.0000000
## 7-5 -3.47872340  -8.9963933  2.03894646 0.5066492
## 7-6  0.09895833  -5.4158080  5.61372462 1.0000000
```

Miles per Gallon by Drivetrain Type

Vehicles with four wheel drive are less efficient than those with two wheel drive.

```r
x = DT1[,.(Cmb.MPG.median = median(Cmb.MPG.mean), Cmb.MPG.25 = quantile(Cmb.MPG.mean,0.25),Cmb.MPG.75 =
library(ggplot2)
ggplot(x, aes(x=Drive, y=Cmb.MPG.median)) +  expand_limits(y=c(20,35)) +
  geom_errorbar(aes(ymin=Cmb.MPG.25, ymax=Cmb.MPG.75), colour="darkgrey", width=.1) +
  geom_point(size = 5,colour = "black")+ geom_line(colour = "red") +
  labs(title="MPG Performance by Drivetrains Type", x="Type of Drivetrains", y="Miles Per Gallon") +ther
```

```
## geom_path: Each group consists of only one observation. Do you need to
## adjust the group aesthetic?
```



Anova-

```
performance_year <- aov(data1$Cmb.MPG.mean~factor(data1$Drive))
summary(performance_year)
```

```
##                       Df Sum Sq Mean Sq F value   Pr(>F)    
## factor(data1$Drive)    1   1461  1461.4   65.46 1.15e-15 ***
## Residuals           1612  35986    22.3                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Tukey HSD-

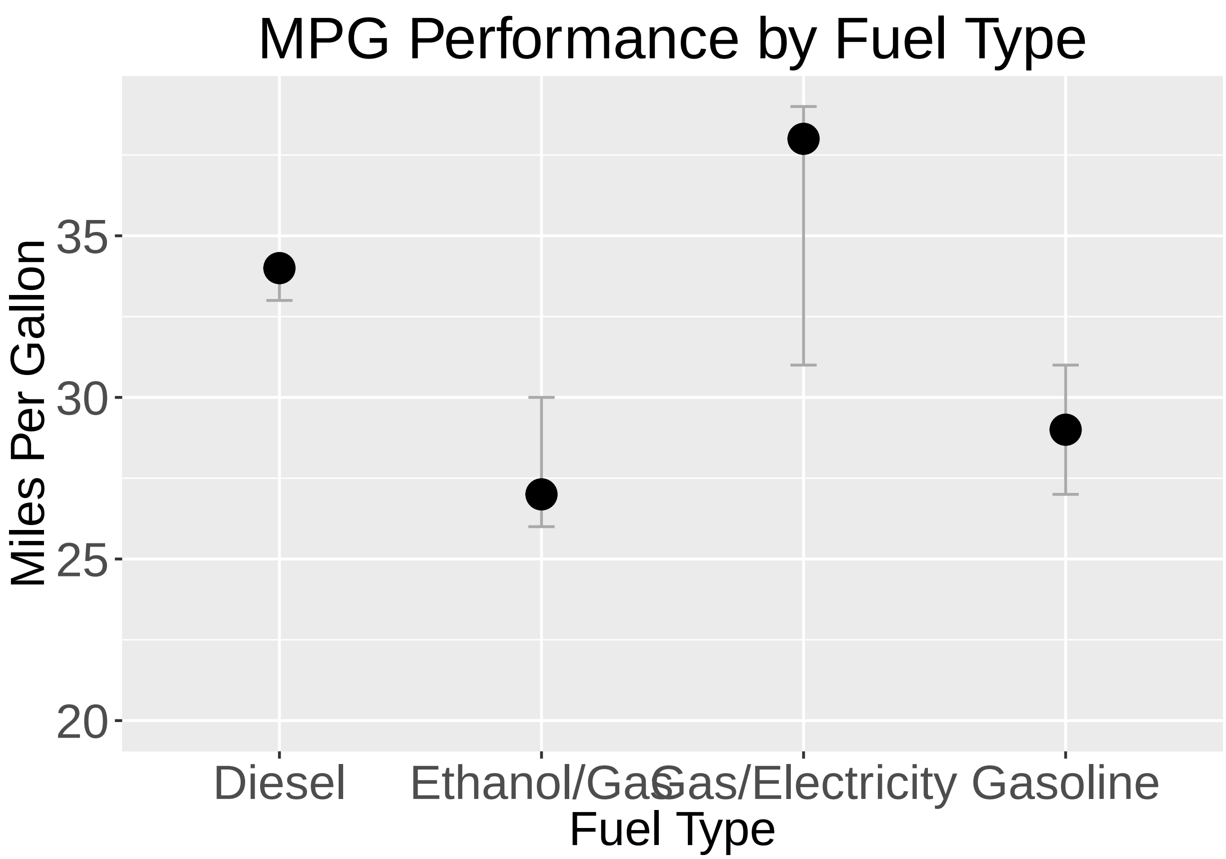
```
TukeyHSD(performance_year, conf.level = 0.95)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = data1$Cmb.MPG.mean ~ factor(data1$Drive))
##
## $`factor(data1$Drive)`
##              diff       lwr       upr p adj
## 4WD-2WD -3.03024 -3.764847 -2.295633     0
```

Miles per Gallon by Fuel Type

Although only the fuel efficiency during the gasoline operation of an engine was considered in this study, vehicles that are able to use both gasoline and electricity generally have higher MPG ratings (when using just gasoline) than cars with other fuel types. Diesel vehicles have higher miles per gallon ratings than gasoline vehicles, while vehicles that use a gasoline and ethanol mix have the lowest miles per gallon ratings.

```
x = DT1[,.(Cmb.MPG.median = median(Cmb.MPG.mean), Cmb.MPG.25 = quantile(Cmb.MPG.mean,0.25),Cmb.MPG.75 =
ggplot(x, aes(x=Fuel, y=Cmb.MPG.median)) +  expand_limits(y=c(20,35)) +
  geom_errorbar(aes(ymin=Cmb.MPG.25, ymax=Cmb.MPG.75), colour="darkgrey", width=.1) +
  geom_point(size = 5,colour = "black")+
  labs(title="MPG Performance by Fuel Type", x="Fuel Type", y="Miles Per Gallon") +theme(text = element_
```



Anova-

```
performance_year <- aov(data1$Cmb.MPG.mean~data1$Fuel)
summary(performance_year)
```

```
##                 Df Sum Sq Mean Sq F value Pr(>F)
## data1$Fuel       3   2984   994.8   46.47 <2e-16 ***
## Residuals     1610  34463    21.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Tukey HSD-

```
TukeyHSD(performance_year, conf.level = 0.95)
```

```
##   Tukey multiple comparisons of means
```

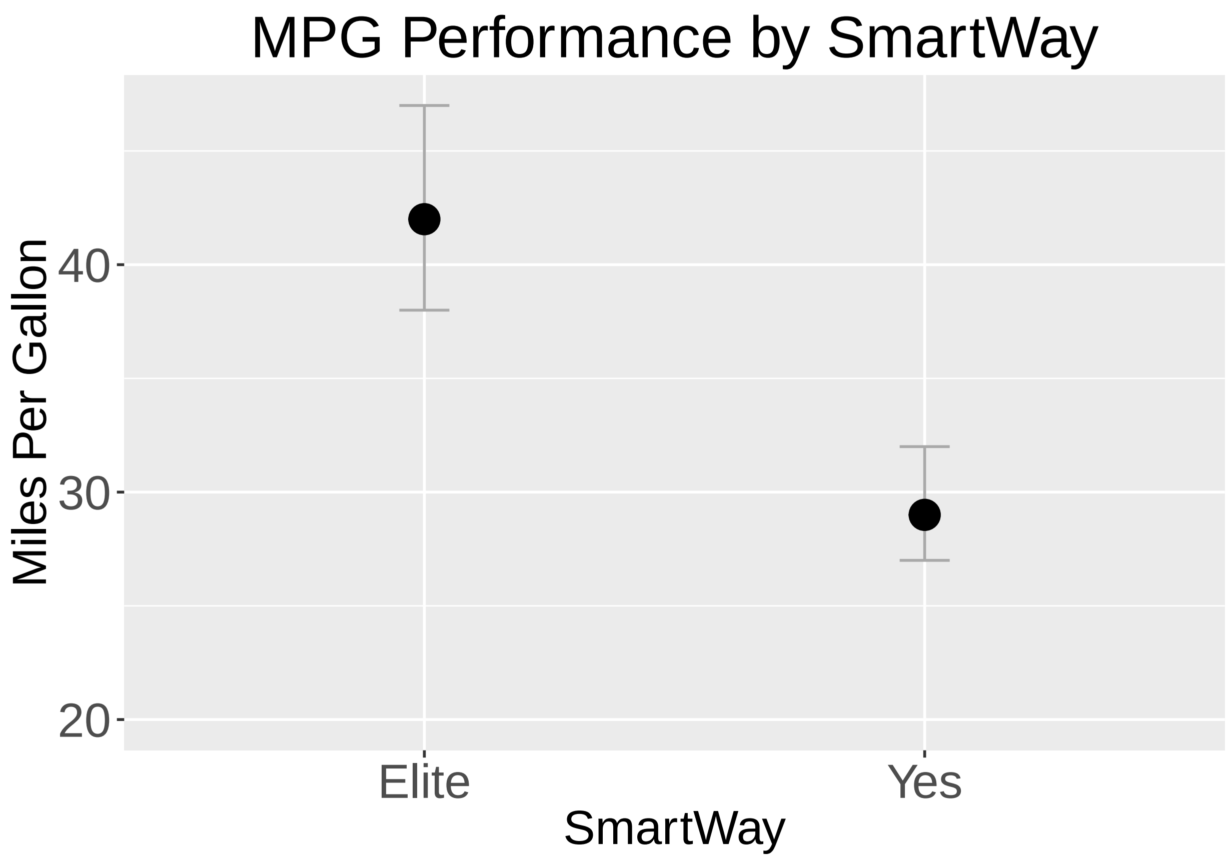
```
##     95% family-wise confidence level
##
## Fit: aov(formula = data1$Cmb.MPG.mean ~ data1$Fuel)
##
## $`data1$Fuel`
##                                   diff        lwr       upr     p adj
## Ethanol/Gas-Diesel           -5.811111 -8.0721325 -3.550090 0.0000000
## Gas/Electricity-Diesel        2.294727  0.2052979  4.384156 0.0247320
## Gasoline-Diesel              -3.533557 -4.9704590 -2.096656 0.0000000
## Gas/Electricity-Ethanol/Gas  8.105838  5.7509588 10.460717 0.0000000
## Gasoline-Ethanol/Gas          2.277554  0.4763239  4.078783 0.0064253
## Gasoline-Gas/Electricity     -5.828284 -7.4087668 -4.247802 0.0000000
```

Miles per Gallon and SmartWay

SmartWay Elite certification not only guarantees lower emissions but also signals a significant improvement in fuel efficiency. Elite certificated vehicles on average get 12.5 more miles per gallon than those that are not certified.

```
x = DT1[,.(Cmb.MPG.median = median(Cmb.MPG.mean), Cmb.MPG.25 = quantile(Cmb.MPG.mean,0.25),Cmb.MPG.75 =
library(ggplot2)
ggplot(x, aes(x=SmartWay, y=Cmb.MPG.median)) +  expand_limits(y=c(20,35)) +
  geom_errorbar(aes(ymin=Cmb.MPG.25, ymax=Cmb.MPG.75), colour="darkgrey", width=.1) +
  geom_point(size = 5,colour = "black")+ geom_line(colour = "red") +
  labs(title="MPG Performance by SmartWay", x="SmartWay", y="Miles Per Gallon") +theme(text = element_te
```

```
## geom_path: Each group consists of only one observation. Do you need to
## adjust the group aesthetic?
```



Anova-

```
performance_year <- aov(data1$Cmb.MPG.mean~data1$SmartWay)
summary(performance_year)
```

```
##                  Df Sum Sq Mean Sq F value Pr(>F)
## data1$SmartWay    1   5619    5619   284.6 <2e-16 ***
## Residuals      1612  31828      20
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Tukey HSD-

```
TukeyHSD(performance_year, conf.level = 0.95)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = data1$Cmb.MPG.mean ~ data1$SmartWay)
##
## $`data1$SmartWay`
##                diff       lwr       upr p adj
## Yes-Elite -12.81052 -14.29995 -11.32109     0
```

Predictive Statistics

The data was used to estimate the combined MPG for any vehicle given its attributes. A linear regression model was fit to the data. Predictors that were significant in the model include type of fuel used, engine displacement, vehicle class, year of the car, type of drivetrains, type of transmission, number of transmissions, and the SmartWay certification. The model explains about 60% of the variation in MPG. Considering the complexity behind fuel efficiency, this model is able to provide reasonable estimates. Interpreting the coefficients of the model reveals the following information: -Vehicles manufactured in recent years have higher MPG ratings -Larger engine displacement results in a lower MPG -Diesel vehicles get more miles per gallon than gasoline-powered vehicles -SUVs and station wagons on average get fewer miles per gallon than smaller vehicles -Four wheel drive vehicles tend to have lower MPGs than two wheel drive vehicles -Every vehicle in the data has a SmartWay designation, but some are marked as SmartWay Elite, the highest industry benchmark. On average, SmartWay Elite vehicles get more miles per gallon holding all other factors the same. Consumers concerned with fuel efficiency should look for SmartWay Elite certification.

In addition, some findings were counterintuitive: - Medium sized sedans are the most fuel efficient type of sedan, and large sedans are slightly more efficient than small sedans - It is generally believed that manual transmission is more fuel efficient than automatic transmissions, but our analysis showed no significant advantage. The most efficient transmission type is continuously variable transmission (CVT) which are often used in hybrid vehicles. - An increase in the number of transmissions leads to a lower MPG

Linear Model -

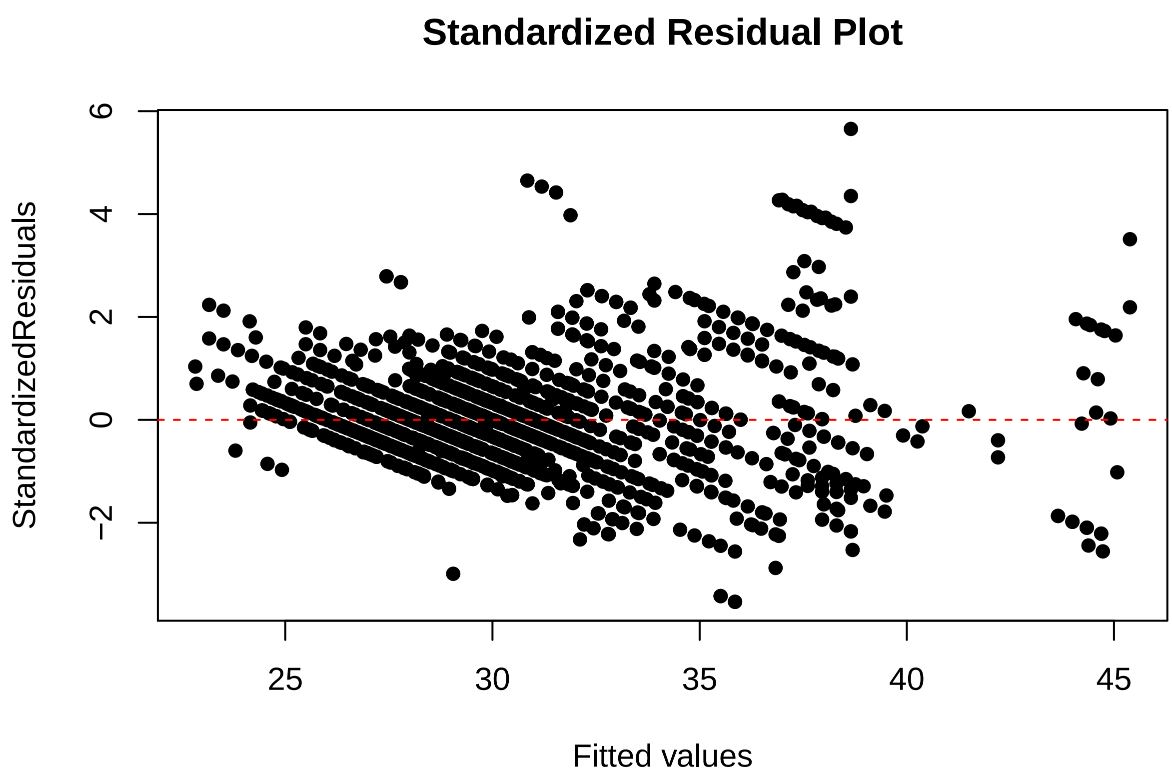
```r
linefit_model <- lm(Cmb.MPG.mean~ Fuel +Displ +Veh.Class+ Drive  + SmartWay  + year + Transmission_numbe
summary(linefit_model)
```

```
##
## Call:
## lm(formula = Cmb.MPG.mean ~ Fuel + Displ + Veh.Class + Drive +
##     SmartWay + year + Transmission_number + Transmission_type,
##     data = data1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.8541  -1.6747  -0.2647   1.2047  17.3489
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)             -652.15149  106.53766  -6.121 1.17e-09 ***
## FuelEthanol/Gas           -5.96909    0.61271  -9.742  < 2e-16 ***
## FuelGas/Electricity       -5.66837    0.59622  -9.507  < 2e-16 ***
## FuelGasoline              -5.40254    0.39729 -13.598  < 2e-16 ***
## Displ                     -0.42848    0.03152 -13.593  < 2e-16 ***
## Veh.Classlarge car        -0.04509    0.48447  -0.093  0.92585
## Veh.Classmidsize car       0.97853    0.19926   4.911 1.00e-06 ***
## Veh.Classstation wagon    -1.40628    0.29124  -4.829 1.51e-06 ***
## Veh.ClassSUV              -2.02771    0.31392  -6.459 1.39e-10 ***
## Drive4WD                  -2.33181    0.26926  -8.660  < 2e-16 ***
## SmartWayYes               -6.73490    0.58676 -11.478  < 2e-16 ***
## year                       0.34749    0.05294   6.564 7.06e-11 ***
## Transmission_number       -0.40742    0.11644  -3.499  0.00048 ***
## Transmission_typeAMS       1.09120    0.49330   2.212  0.02711 *
## Transmission_typeAutoMan   3.44301    0.40802   8.438  < 2e-16 ***
## Transmission_typeCVT       6.53533    0.44032  14.842  < 2e-16 ***
```

```
## Transmission_typeMan         0.18928    0.26441   0.716  0.47419
## Transmission_typeOther       2.67494    0.91739   2.916  0.00360 **
## Transmission_typeSCV         4.58299    0.40428  11.336  < 2e-16 ***
## Transmission_typeSemiAuto    0.10898    0.28440   0.383  0.70162
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.081 on 1594 degrees of freedom
## Multiple R-squared:  0.596,  Adjusted R-squared:  0.5912
## F-statistic: 123.8 on 19 and 1594 DF,  p-value: < 2.2e-16
```

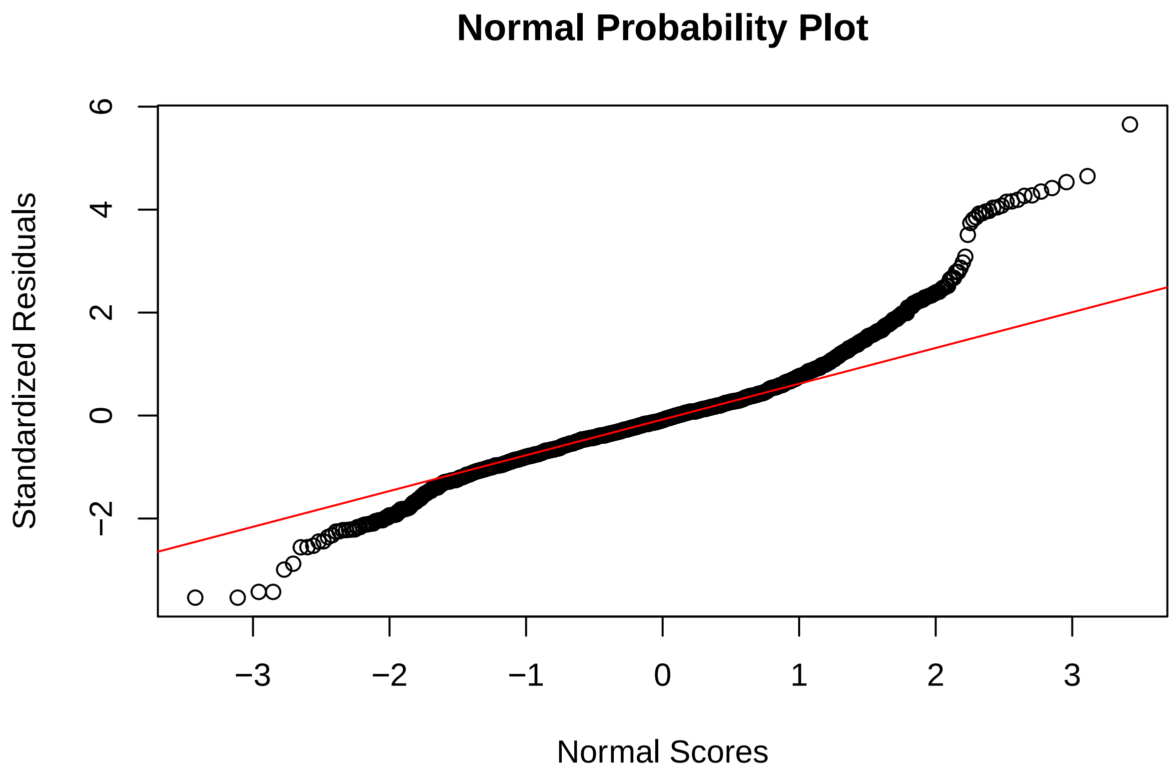
Standard Residual -

```
linefit_model.stres <- rstandard(linefit_model)
plot(linefit_model$fitted.values, linefit_model.stres, pch = 16, main = "Standardized Residual Plot", xl
abline(0,0, lty=2, col="red")
```



QQ Plot -

```
##QQ - Plot
qqnorm(linefit_model.stres, main = "Normal Probability Plot", xlab = "Normal Scores", ylab = "Standardi
qqline(linefit_model.stres, col = "red")
```



Shaphiro Test

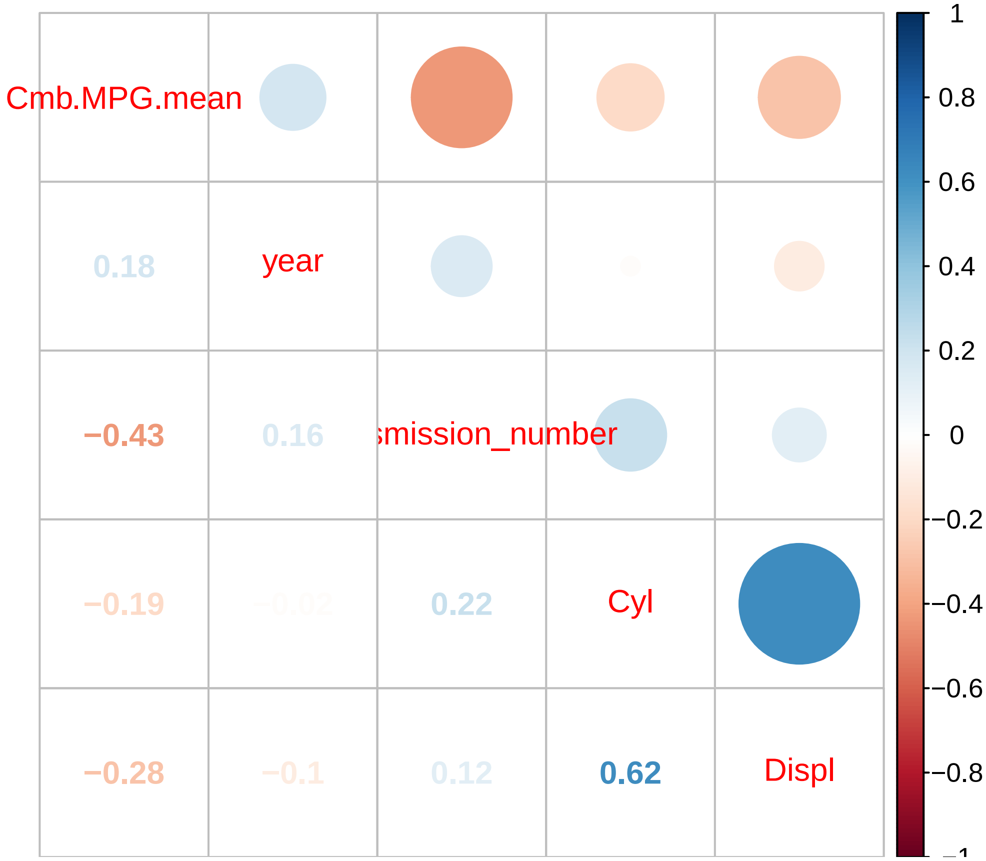
```
shapiro.test(linefit_model.stres)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  linefit_model.stres
## W = 0.9329, p-value < 2.2e-16
```

Correlation Plot -

We checked the collinearity between our interval predictors: year, number of transmissions, number of cylinders, and number of engine displacement, and we find the coefficient between number of cylinders and number of engine displacement is 0.62, which shows some evidence of collinearity. The cylinder variable has a positive coefficient, which means the MPG will increase as number of cylinders increase, and this is counter-intuitive. We decided to remove the cylinder variable from our regression. The new model's R square and adjusted R square dropped a bit, but the coefficients are more interpretable, so we prefered the model without cylinder variable.

```
M <- cor(data1[,c("Cmb.MPG.mean","Displ","Cyl","Transmission_number","year")])
corrplot.mixed(M,order = "AOE")
```


Cmb.MPG.mean
year
smission_number
Cyl
Displ
0.18
−0.43
0.16
−0.19
−0.02
0.22
−0.28
−0.1
0.12
0.62
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8