Statistic Project

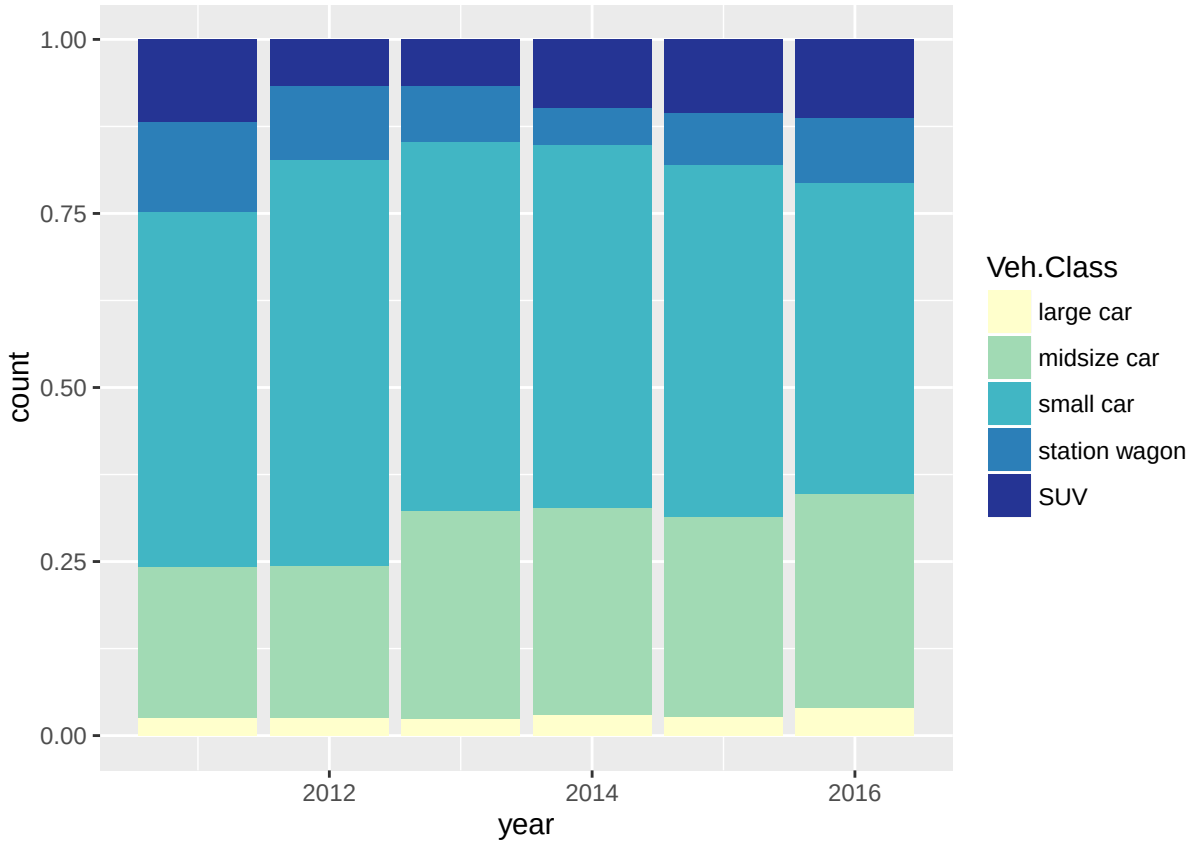
Descriptive Statistics of Variables

```
summary(data1)
```

```
##                Model          Displ             Cyl        Drive
##  MAZDA 3         :  38   Min.   : 2.00   Min.   :1.000   2WD:1435
##  VOLKSWAGEN Jetta:  37   1st Qu.:10.00   1st Qu.:3.000   4WD: 179
##  FORD Focus      :  22   Median :12.00   Median :3.000
##  KIA Forte       :  22   Mean   :11.81   Mean   :3.084
##  NISSAN Versa    :  22   3rd Qu.:12.00   3rd Qu.:3.000
##  CHEVROLET Cruze :  21   Max.   :20.00   Max.   :6.000
##  (Other)         :1452
##          Veh.Class    SmartWay         year      Transmission_number
##  large car    : 46   Elite:  35   Min.   :2011   Min.   :1.000
##  midsize car  :446   Yes  :1579   1st Qu.:2012   1st Qu.:3.000
##  small car    :833                Median :2014   Median :4.000
##  station wagon:137                Mean   :2014   Mean   :3.578
##  SUV          :152                3rd Qu.:2015   3rd Qu.:4.000
##                                   Max.   :2016   Max.   :7.000
##
##  Transmission_type              Fuel       Hwy.MPG.mean   City.MPG.mean
##  Man     :510      Diesel         :  72   Min.   :21.00   Min.   :20.00
##  SemiAuto:421      Ethanol/Gas    :  45   1st Qu.:33.00   1st Qu.:24.00
##  CVT     :213      Gas/Electricity:  59   Median :35.00   Median :26.00
##  Auto    :203      Gasoline       :1438   Mean   :35.37   Mean   :27.62
##  SCV     :108                             3rd Qu.:38.00   3rd Qu.:29.00
##  AutoMan : 87                             Max.   :53.00   Max.   :58.00
##  (Other) : 72
##   Cmb.MPG.mean  Greenhouse.Gas.Score.mean Air.Pollution.Score.mean
##  Min.   :20.0   Min.   : 6.000            Min.   :2.000
##  1st Qu.:27.0   1st Qu.: 7.000            1st Qu.:3.000
##  Median :29.0   Median : 7.000            Median :3.000
##  Mean   :30.5   Mean   : 7.444            Mean   :3.777
##  3rd Qu.:32.0   3rd Qu.: 8.000            3rd Qu.:5.000
##  Max.   :56.0   Max.   :10.000            Max.   :7.000
##
```

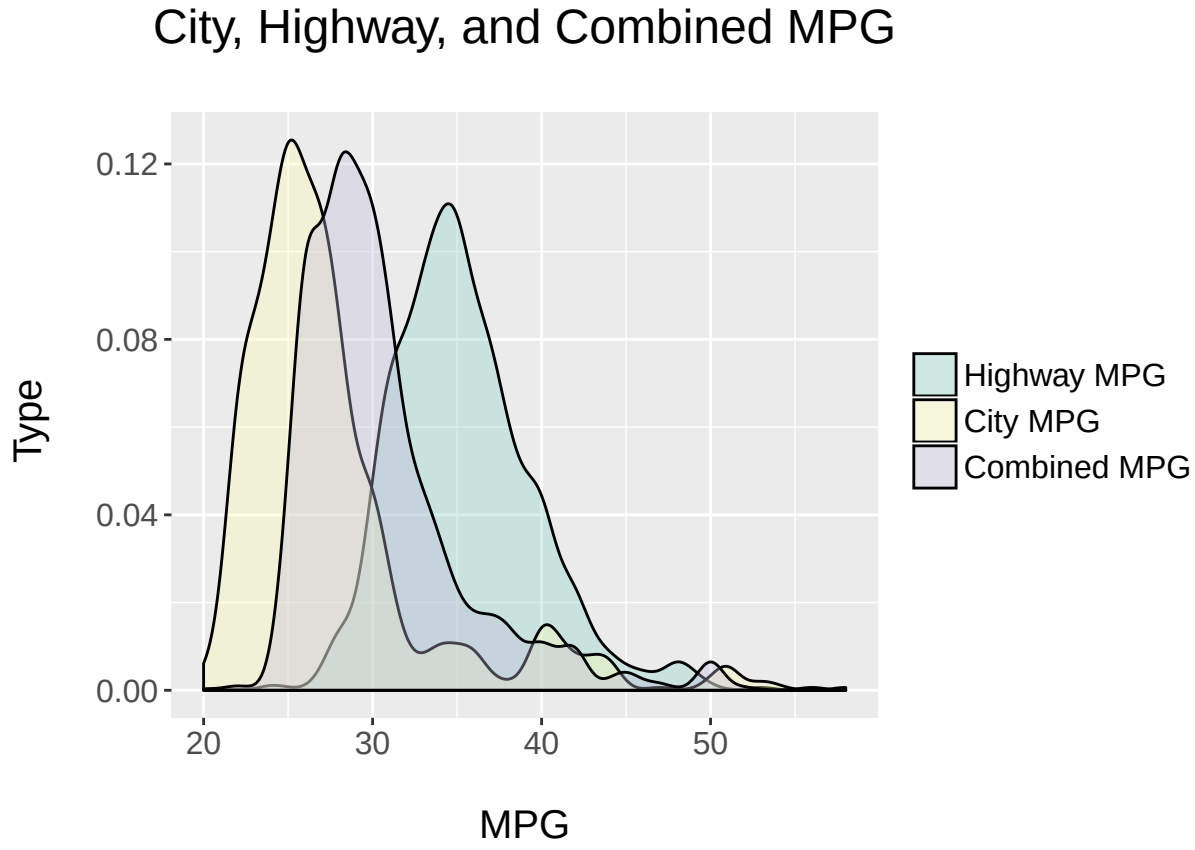
Vehicle Market Overview

```
library(ggplot2)
attach(data1)
veh.class_by_year <- table(Veh.Class, year)
ggplot(data1,aes(x = year,fill = Veh.Class),geom="text") +
  geom_bar(position = "fill" ) +
  scale_fill_brewer(palette="YlGnBu")
```

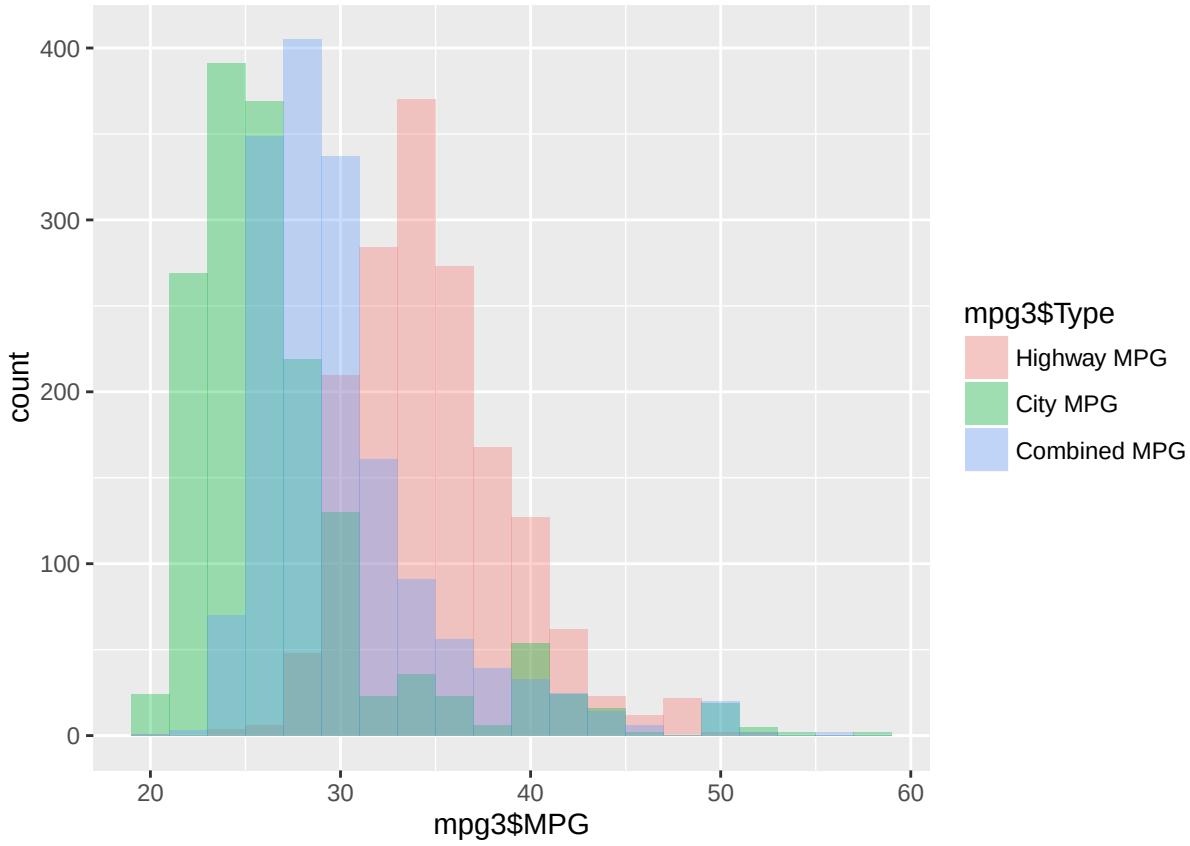


Graphs for interval variables (MPG)

```
highway <- data.frame(data1$Hwy.MPG.mean, 'Highway MPG')
city <- data.frame(data1$City.MPG.mean, 'City MPG')
combined <- data.frame(data1$Cmb.MPG.mean, 'Combined MPG')
name <- c('MPG', 'Type')
colnames(highway) <- name
colnames(city) <- name
colnames(combined) <- name
mpg3 <- rbind(highway,city,combined)
# density plot
ggplot(mpg3, aes(mpg3$MPG, fill = mpg3$Type)) +
  geom_density(alpha = 0.35) +
  labs(title="City, Highway, and Combined MPG\n", x="\nMPG", y="Type \n") +
  theme(text = element_text(size=15)) +
  scale_fill_brewer(palette="Set3", guide = guide_legend(title = NULL))
```

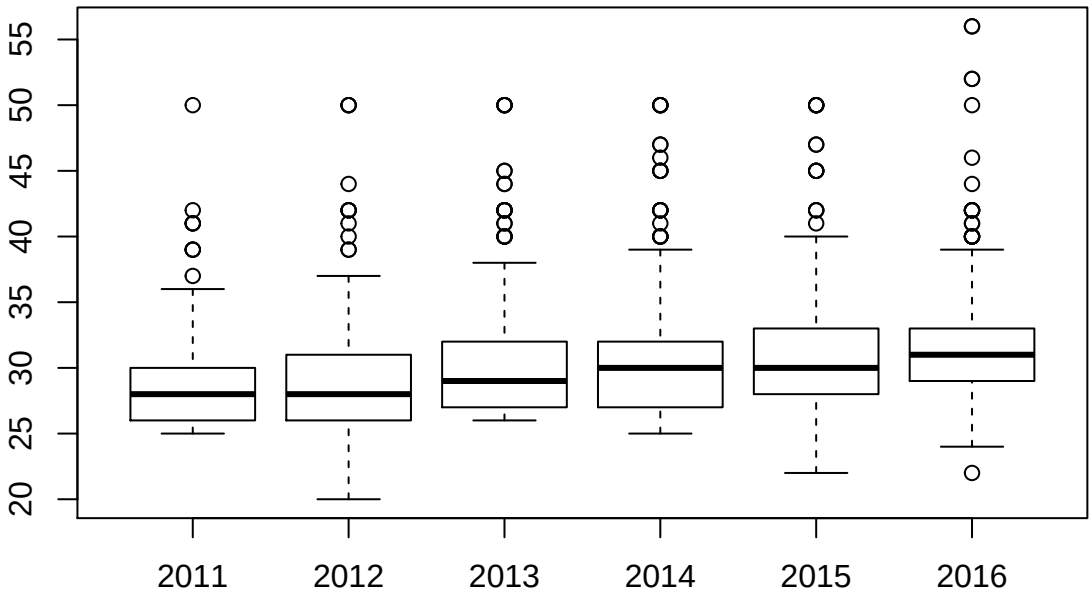



```
# histogra
ggplot(mpg3, aes(x = mpg3$MPG, fill = mpg3$Type)) +
  geom_histogram(alpha=0.35, position="identity",bins = 20)
```

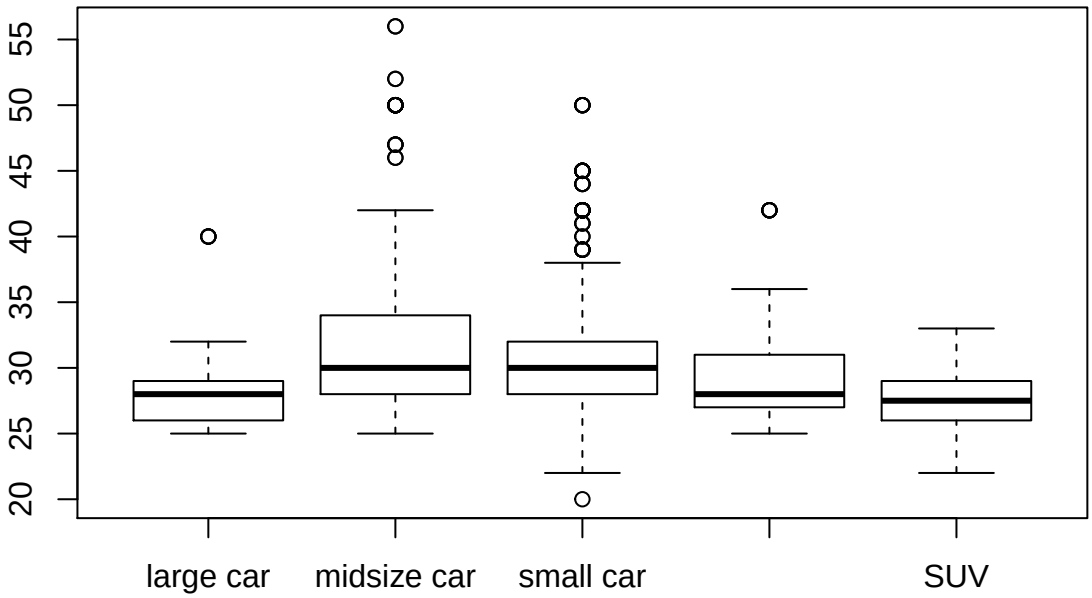


Quickly review MPG vs. other indexes

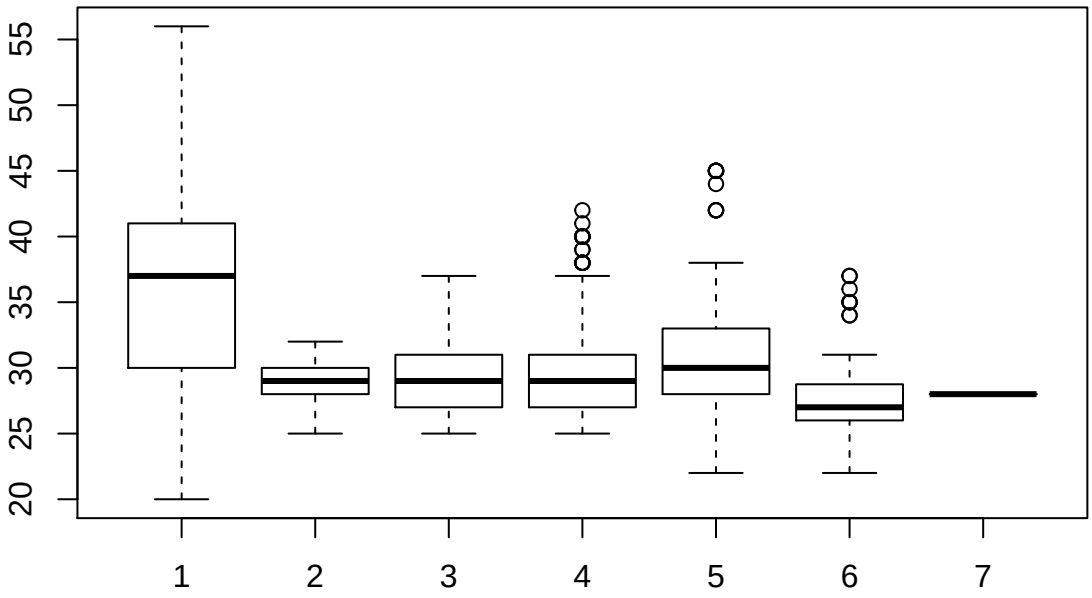
```
boxplot(Cmb.MPG.mean~year)
```



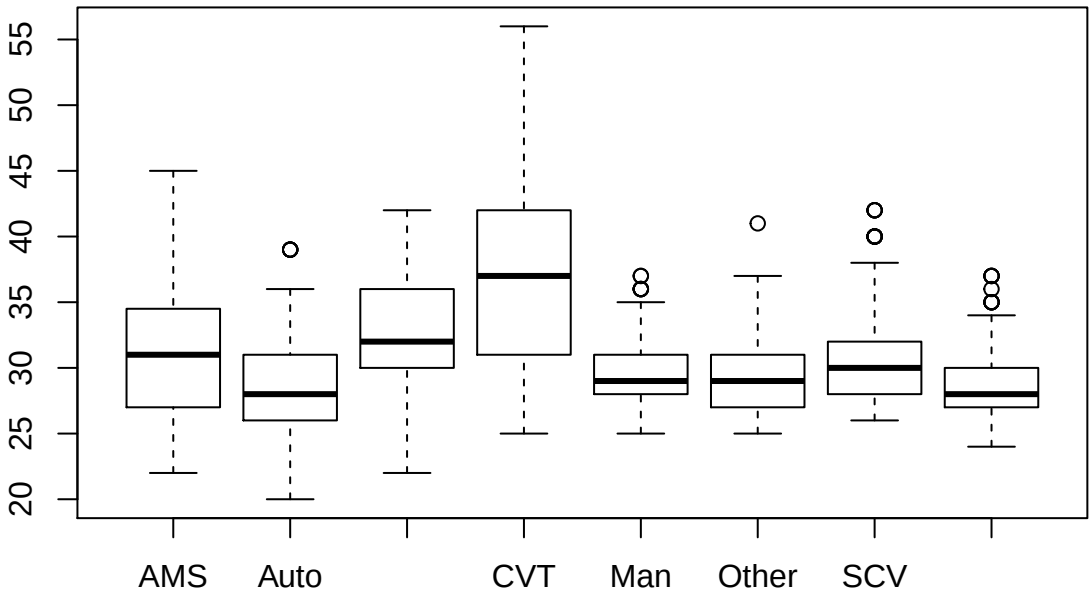

```
boxplot(Cmb.MPG.mean~Veh.Class)
```



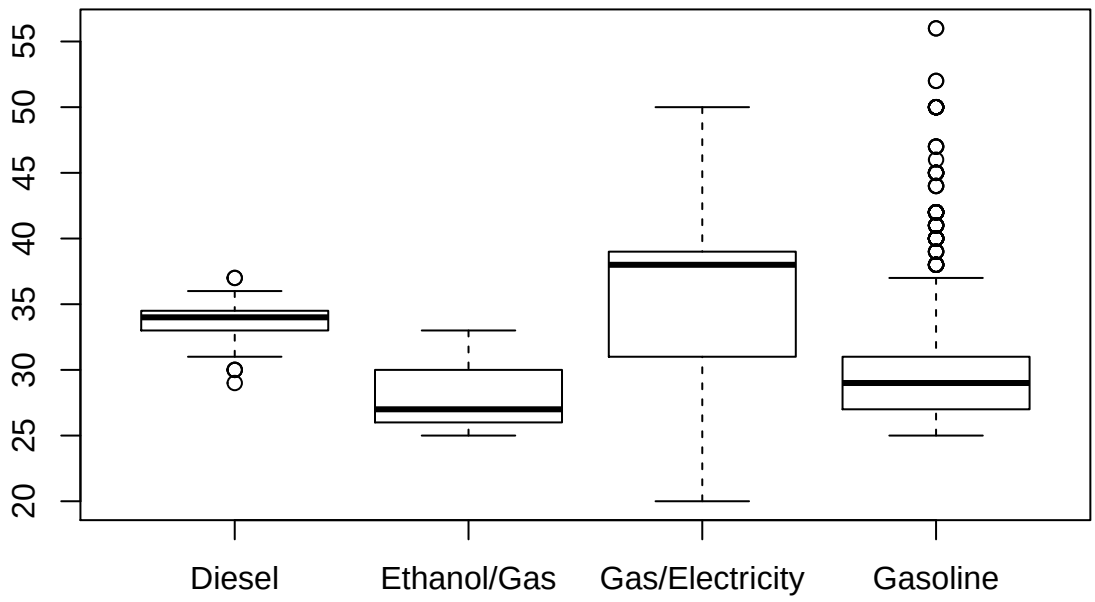

```
boxplot(Cmb.MPG.mean~Transmission_number)
```




```
boxplot(Cmb.MPG.mean~Transmission_type)
```

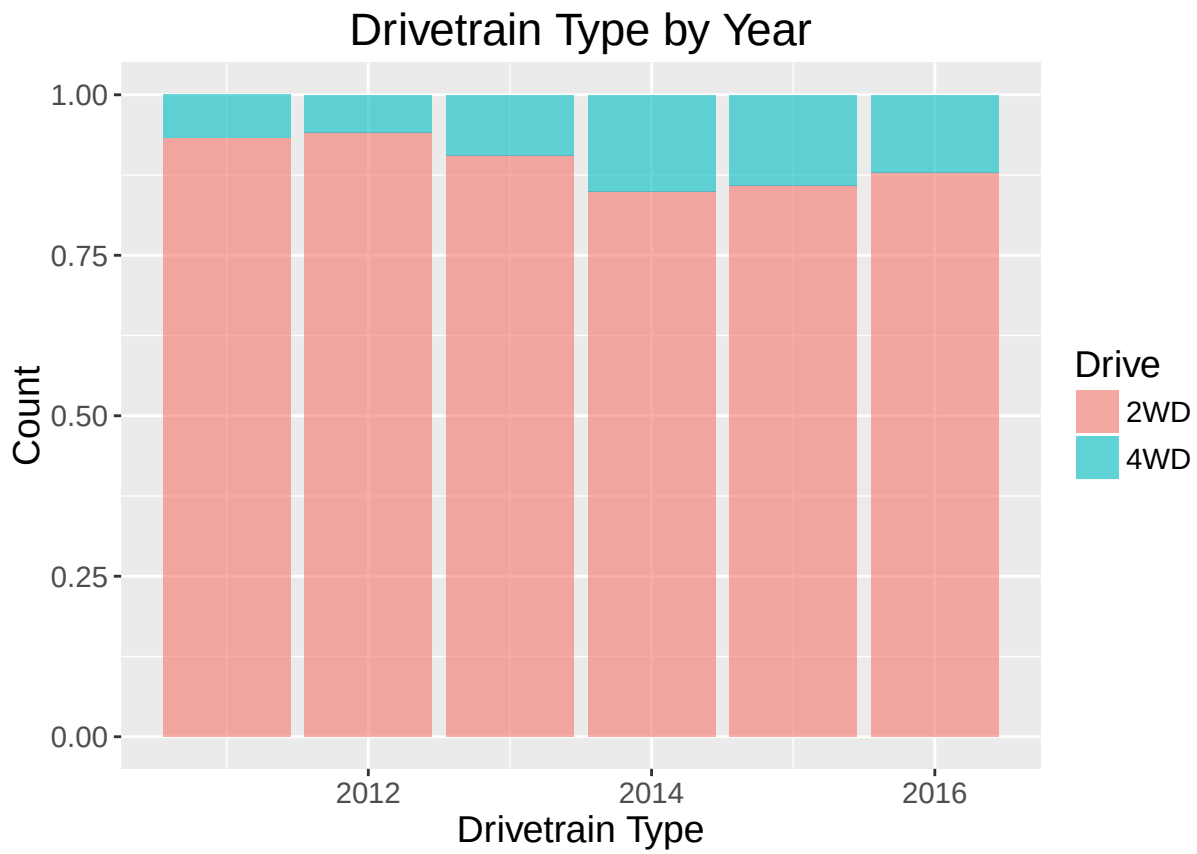



```
boxplot(Cmb.MPG.mean~Fuel)
```

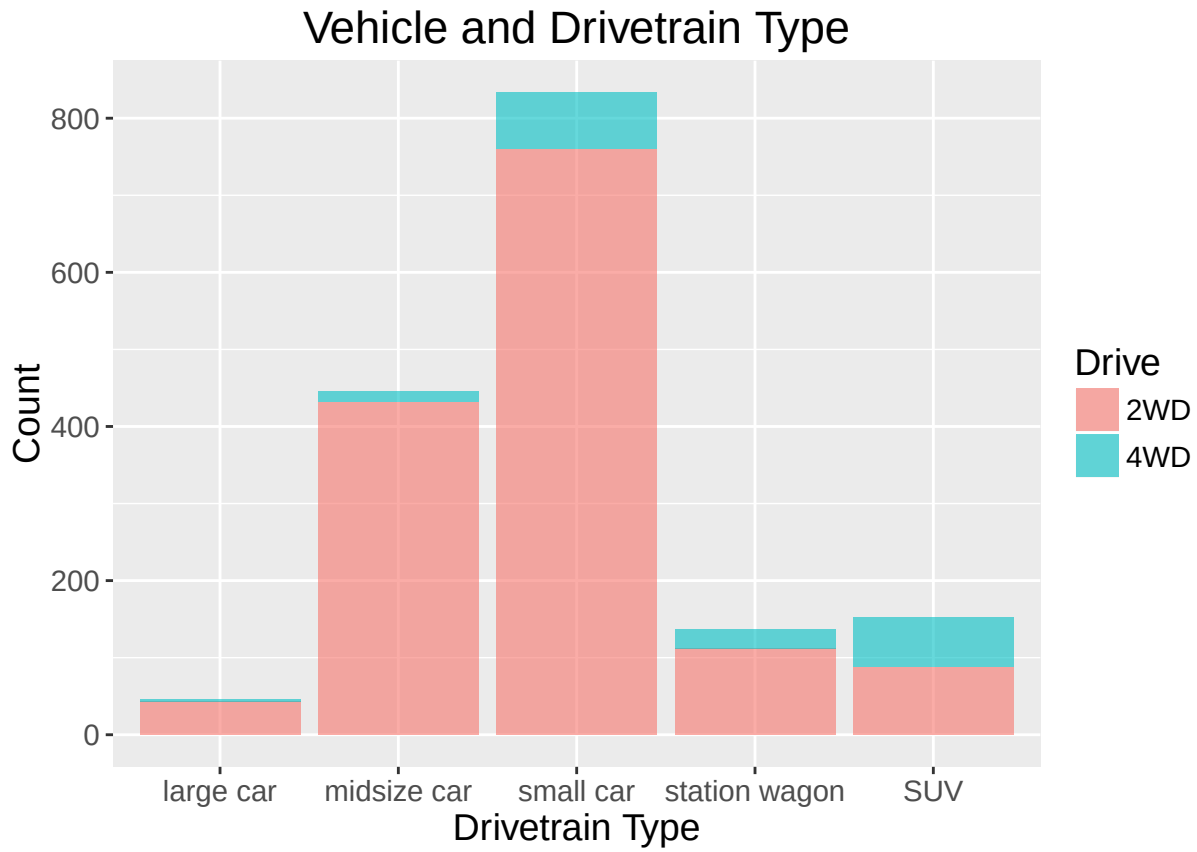


Vehicle vs. Drivetrain Type

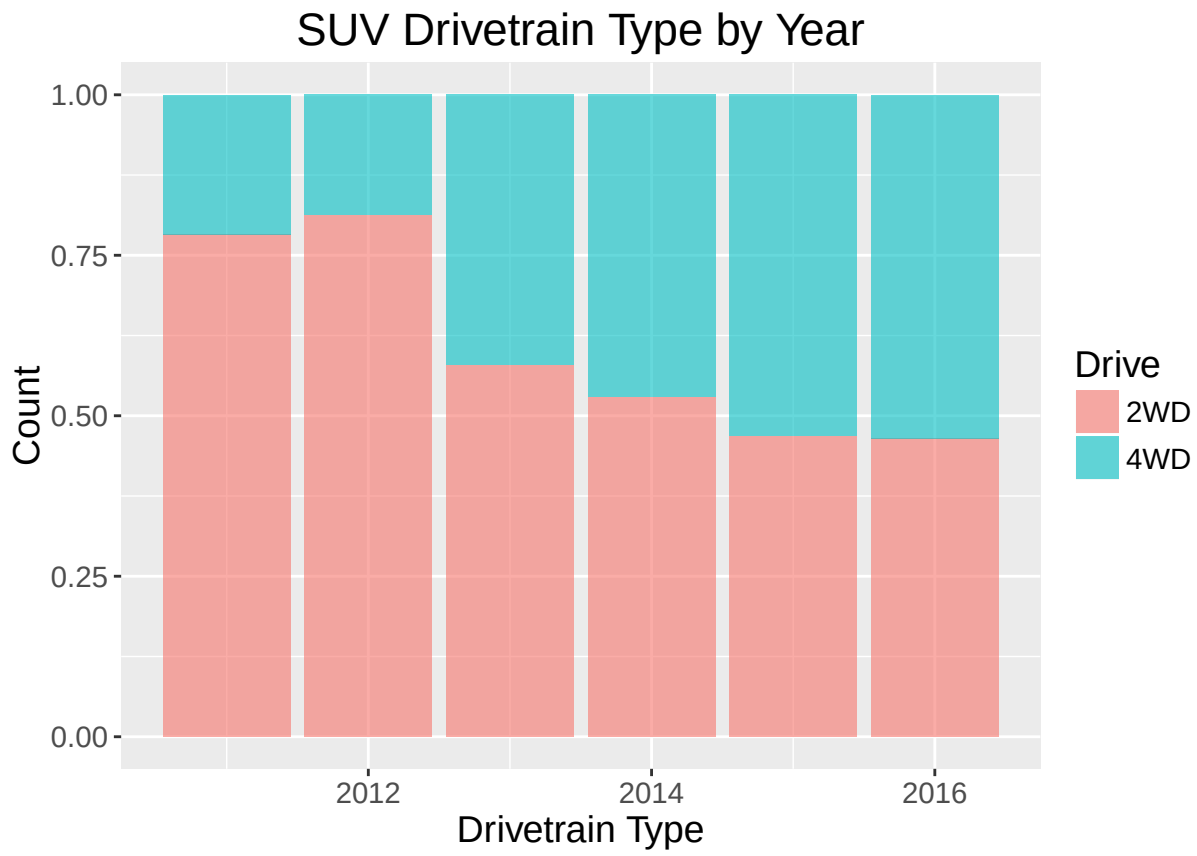
```
ggplot(data1, aes(x = year, fill = Drive)) +
  geom_bar(alpha=0.6, position="fill",  stat = "count" ) +
  labs(title="Drivetrain Type by Year", x="Drivetrain Type", y="Count") +
  theme(text = element_text(size=14))
```




```
ggplot(data1, aes(x = Veh.Class, fill = Drive)) +
  geom_bar(alpha=0.6, position="stack",  stat = "count" ) +
  labs(title="Vehicle and Drivetrain Type", x="Drivetrain Type", y="Count") +
  theme(text = element_text(size=14))
```



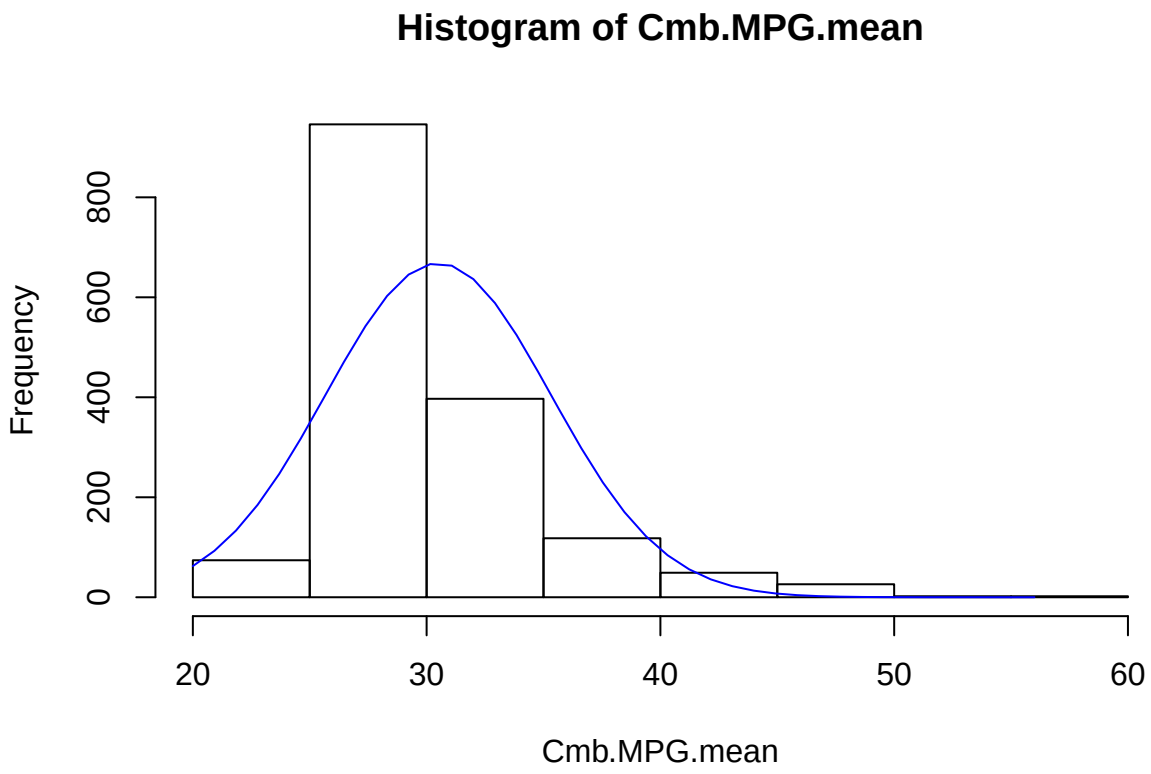

```
data1_SUV <- data1[data1$Veh.Class == 'SUV', ]
ggplot(data1_SUV, aes(x = year, fill = Drive)) +
  geom_bar(alpha=0.6, position="fill",  stat = "count" ) +
  labs(title="SUV Drivetrain Type by Year", x="Drivetrain Type", y="Count") +
  theme(text = element_text(size=14))
```



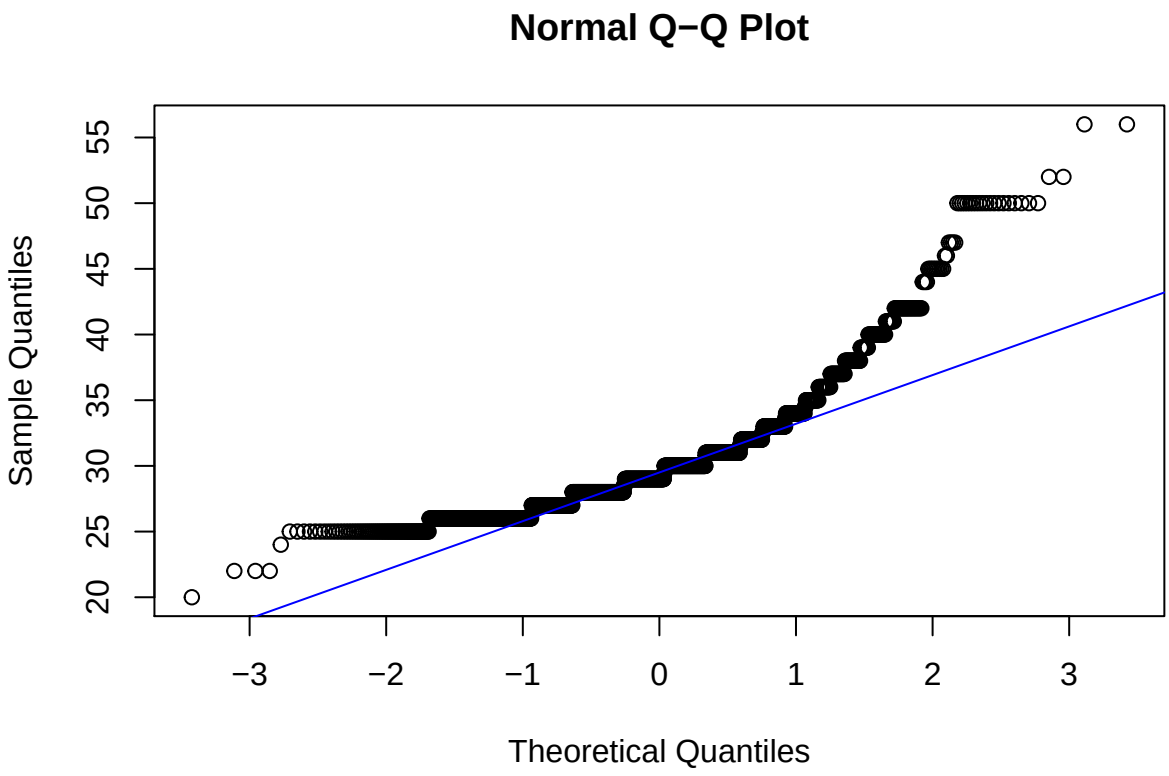
Checking normality of MPG

```
# Add normal curve
h <- hist(Cmb.MPG.mean)

x <- data1$Cmb.MPG.mean
xfit <- seq(min(x), max(x), length = 40)
yfit <- dnorm(xfit, mean = mean(x), sd = sd(x))
yfit <- yfit*diff(h$mids[1:2])*length(x)
lines(xfit, yfit, col="blue")
```




```
  # probability plot
qqnorm(x)
qqline(x, col = "blue")
```




```
  # Goodness of fit test of H0: normal
shapiro.test(Cmb.MPG.mean)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Cmb.MPG.mean
## W = 0.83542, p-value < 2.2e-16
```

Chi-square test for vehicle class vs. year

```
chisq.test(year, Veh.Class)
```

```
##
##  Pearson's Chi-squared test
##
## data:  year and Veh.Class
## X-squared = 30.614, df = 20, p-value = 0.0605
```

```
round((chisq.test(year, Veh.Class)$residual),2)
```

```
##       Veh.Class
## year   large car midsize car small car station wagon   SUV
##   2011     -0.23       -1.59     -0.11          2.10  1.11
##   2012     -0.29       -1.67      1.42          1.09 -1.34
##   2013     -0.39        0.70      0.32         -0.24 -1.51
```

```
##   2014      0.05        0.79      0.15         -2.09  0.26
##   2015     -0.23        0.41     -0.27         -0.57  0.61
##   2016      1.10        0.90     -1.50          0.42  0.96
```

ANOVA & Tukey test of performance by year

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

```
##               Df Sum Sq Mean Sq F value  Pr(>F)
## 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
```

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

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = data1$Cmb.MPG.mean ~ data1$year)
##
## $`data1$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
```

ANOVA & Tukey test of performance by class

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

```
##                  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
```

```
TukeyHSD(performance_class, 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
## midsize car-large car      3.4821018  1.51809287  5.44611068 0.0000139
## small car-large car        1.8872332 -0.03364699  3.80811330 0.0569143
## station wagon-large car    0.8145033 -1.34668986  2.97569653 0.8418909
## SUV-large car             -1.0244565 -3.15867625  1.10976321 0.6845017
## small car-midsize car     -1.5948686 -2.33900509 -0.85073215 0.0000001
## station wagon-midsize car -2.6675984 -3.90643442 -1.42876246 0.0000000
## SUV-midsize car           -4.5065583 -5.69771422 -3.31540237 0.0000000
## station wagon-small car   -1.0727298 -2.24198776  0.09652812 0.0898837
## SUV-small car             -2.9116897 -4.03030578 -1.79307357 0.0000000
## SUV-station wagon         -1.8389599 -3.33304067 -0.34487904 0.0070968
```

ANOVA & Tukey test of performance by transmission type

```
data1$Transmission_type <- factor(data1$Transmission_type)
performance_trans_type <- aov(data1$Cmb.MPG.mean~data1$Transmission_type)
summary(performance_trans_type)
```

```
##                           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
```

```
TukeyHSD(performance_trans_type, 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
## Auto-AMS          -3.423060867  -5.1933504 -1.65277136 0.0000001
## 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
## 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
## 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
```

ANOVA & Tukey test of performance by number of transmission

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

```
##                              Df Sum Sq Mean Sq F value Pr(>F)
## 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
```

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

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = data1$Cmb.MPG.mean ~ data1$Transmission_number)
##
## $`data1$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
```

ANOVA & Tukey test of performance by fuel type

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

```
##               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
```

```
TukeyHSD(performance_fuel, 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
```

ANOVA & Tukey test of performance by SmartWay

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

```
##                  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
```

```
TukeyHSD(performance_SmartWay, 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
```

ANOVA & Tukey test of performance by cylinder

```
data1$Cyl <- factor(data1$Cyl)
performance_cyl <- aov(data1$Cmb.MPG.mean~data1$Cyl)
summary(performance_cyl)
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## data1$Cyl      5   1914   382.7   17.32 <2e-16 ***
## Residuals   1608  35533    22.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = data1$Cmb.MPG.mean ~ data1$Cyl)
##
## $`data1$Cyl`
##           diff         lwr        upr     p adj
## 2-1  -4.344828 -10.8394664  2.1498112 0.3970520
## 3-1  -8.409257 -14.4174918 -2.4010217 0.0009609
## 4-1 -13.000000 -19.5299169 -6.4700831 0.0000002
## 5-1 -10.916667 -17.1195507 -4.7137827 0.0000085
## 6-1 -17.000000 -31.6923129 -2.3076871 0.0125993
## 3-2  -4.064429  -6.5792920 -1.5495663 0.0000633
## 4-2  -8.655172 -12.2420180 -5.0683268 0.0000000
## 5-2  -6.571839  -9.5216555 -3.6220227 0.0000000
## 6-2 -12.655172 -26.2966425  0.9862977 0.0868932
## 4-3  -4.590743  -7.1953576 -1.9861289 0.0000081
## 5-3  -2.507410  -4.1260426 -0.8887772 0.0001531
## 6-3  -8.590743 -22.0074590  4.8259725 0.4485987
## 5-4   2.083333  -0.9433641  5.1100308 0.3637428
## 6-4  -4.000000 -17.6583010  9.6583010 0.9608871
## 6-5  -6.083333 -19.5883377  7.4216710 0.7932566
```

Linear regression to test performance on Displ

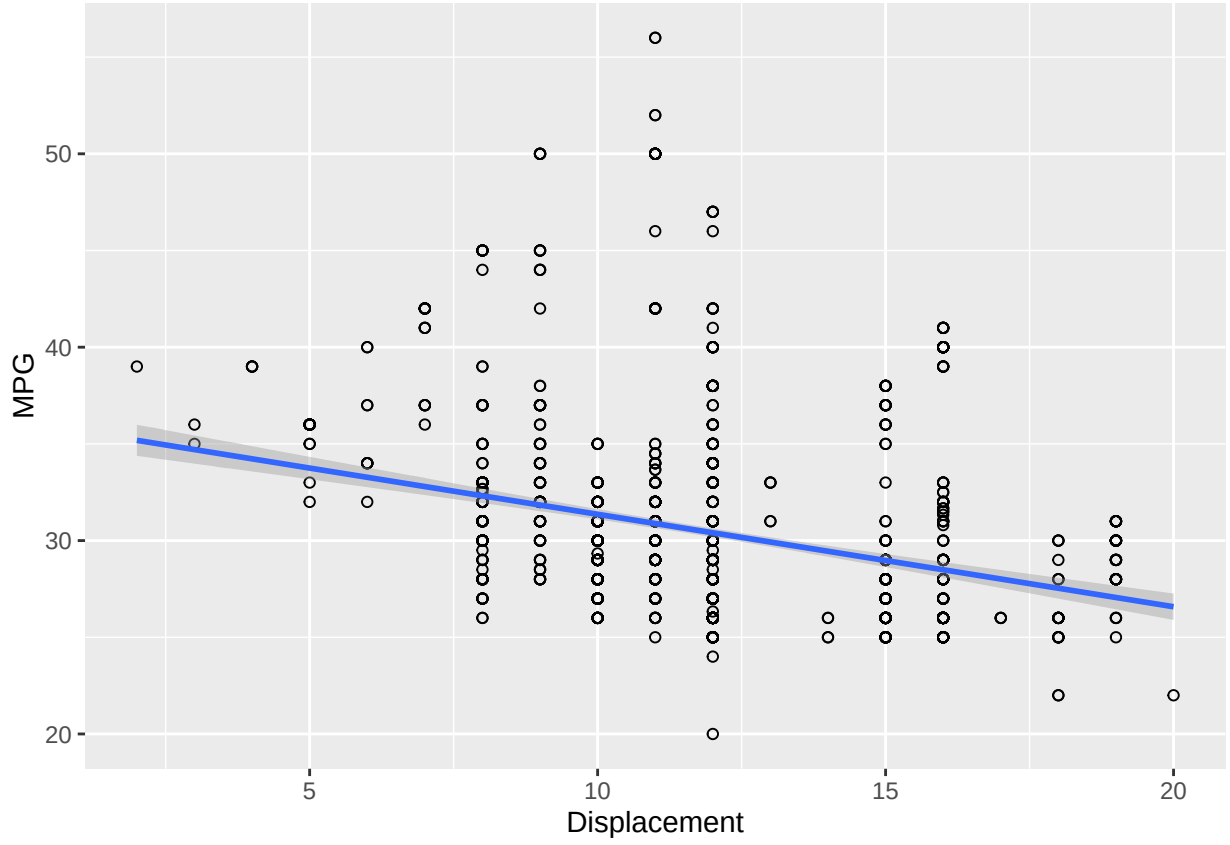
```
linefit_d = lm(data1$Cmb.MPG.mean ~ data1$Displ)
summary(linefit_d)
```

```
##
## Call:
## lm(formula = data1$Cmb.MPG.mean ~ data1$Displ)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -10.404  -2.882  -1.360   1.118  25.118
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 36.14021    0.48662   74.27   <2e-16 ***
## data1$Displ -0.47804    0.04005  -11.94   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.62 on 1612 degrees of freedom
## Multiple R-squared:  0.08122,    Adjusted R-squared:  0.08065
## F-statistic: 142.5 on 1 and 1612 DF,  p-value: < 2.2e-16
```

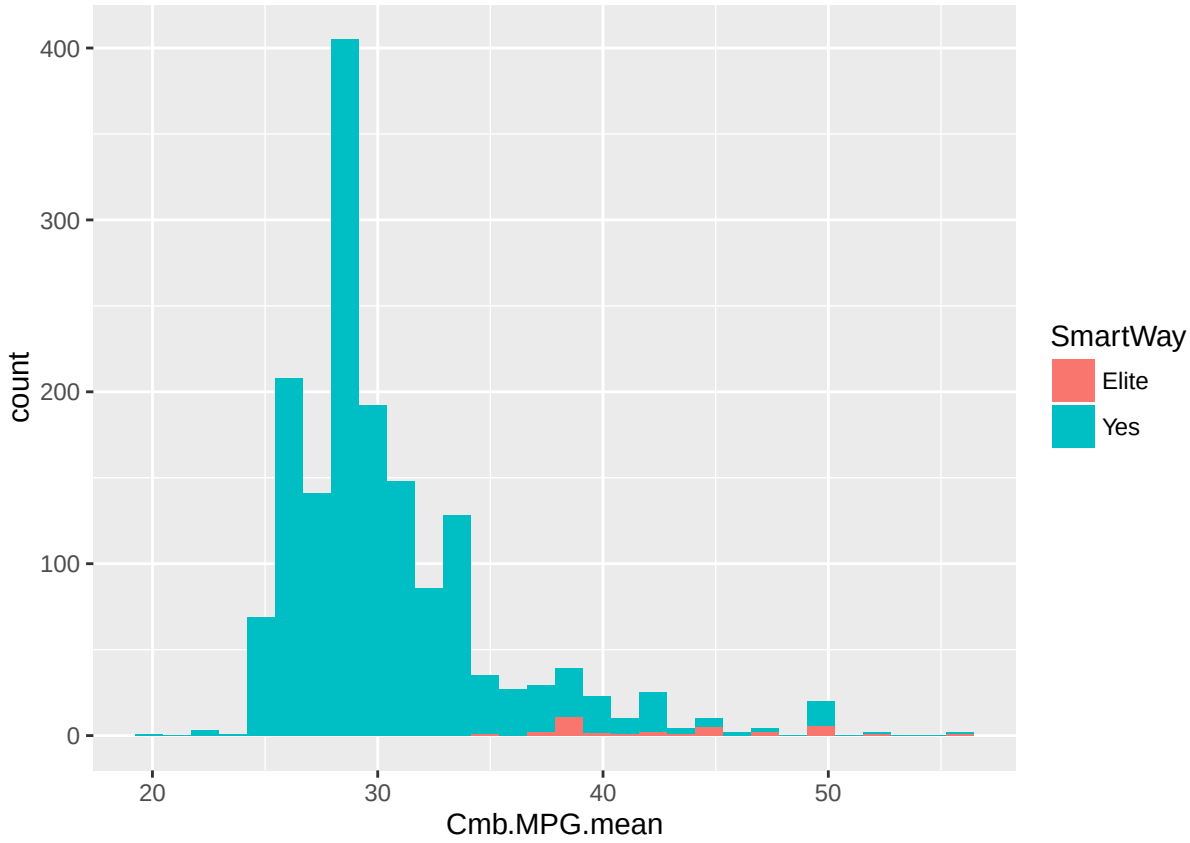
Visualize linear regression of performance vs. Displ

```
dat <- data.frame(Displacement = data1$Displ, MPG = data1$Cmb.MPG.mean)
ggplot(dat, aes(x=Displacement, y=MPG)) + geom_point(shape=1) + geom_smooth(method=lm)
```



Performance vs. SmartWay

```
qplot(Cmb.MPG.mean, data=data1, fill = SmartWay, bins = 30)
```


400
300
200
100
0
count
20
30
40
50
Cmb.MPG.mean
SmartWay
Elite
Yes