ggplot2

Exploratory Data Analysis

July 14, 2017

Let's look at the data `mpg` and it's factor variable `manufacturer`:

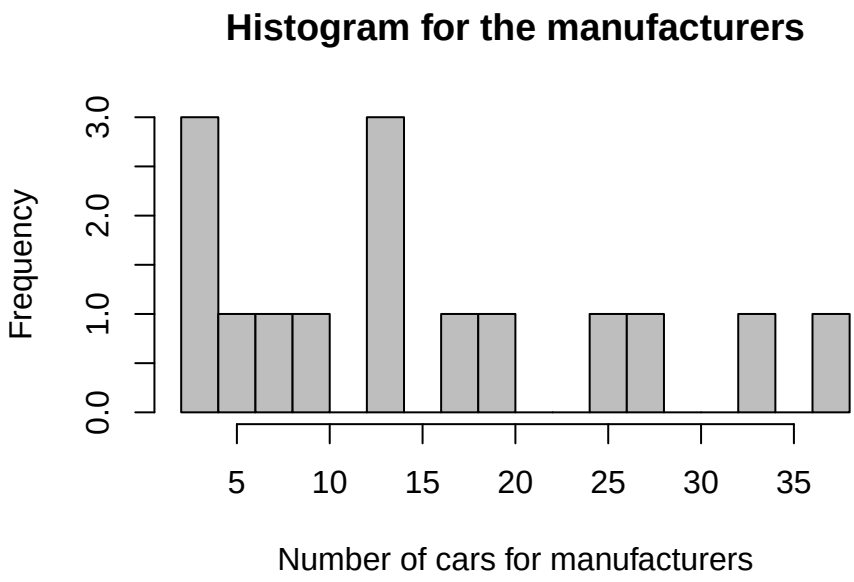
```
library(ggplot2)
str(mpg)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    234 obs. of  11 variables:
##  $ manufacturer: chr  "audi" "audi" "audi" "audi" ...
##  $ model       : chr  "a4" "a4" "a4" "a4" ...
##  $ displ       : num  1.8 1.8 2 2 2.8 2.8 3.1 1.8 1.8 2 ...
##  $ year        : int  1999 1999 2008 2008 1999 1999 2008 1999 1999 2008 ...
##  $ cyl         : int  4 4 4 4 6 6 6 4 4 4 ...
##  $ trans       : chr  "auto(l5)" "manual(m5)" "manual(m6)" "auto(av)" ...
##  $ drv         : chr  "f" "f" "f" "f" ...
##  $ cty         : int  18 21 20 21 16 18 18 18 16 20 ...
##  $ hwy         : int  29 29 31 30 26 26 27 26 25 28 ...
##  $ fl          : chr  "p" "p" "p" "p" ...
##  $ class       : chr  "compact" "compact" "compact" "compact" ...
```

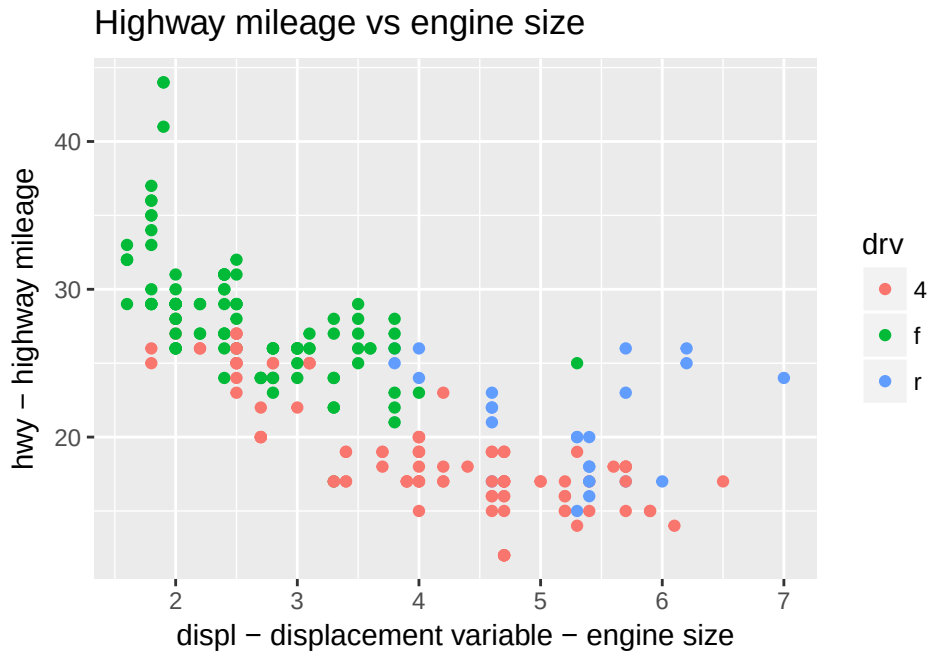
```
table(mpg$manufacturer)
```

```
##
##       audi  chevrolet      dodge       ford      honda    hyundai
##         18         19         37         25          9         14
##       jeep land rover    lincoln    mercury     nissan    pontiac
##          8          4          3          4         13          5
##     subaru     toyota volkswagen
##         14         34         27
```

```
hist(table(mpg$manufacturer),
     breaks = 20,
     col = 8,
     main = "Histogram for the manufacturers",
     xlab = "Number of cars for manufacturers")
```

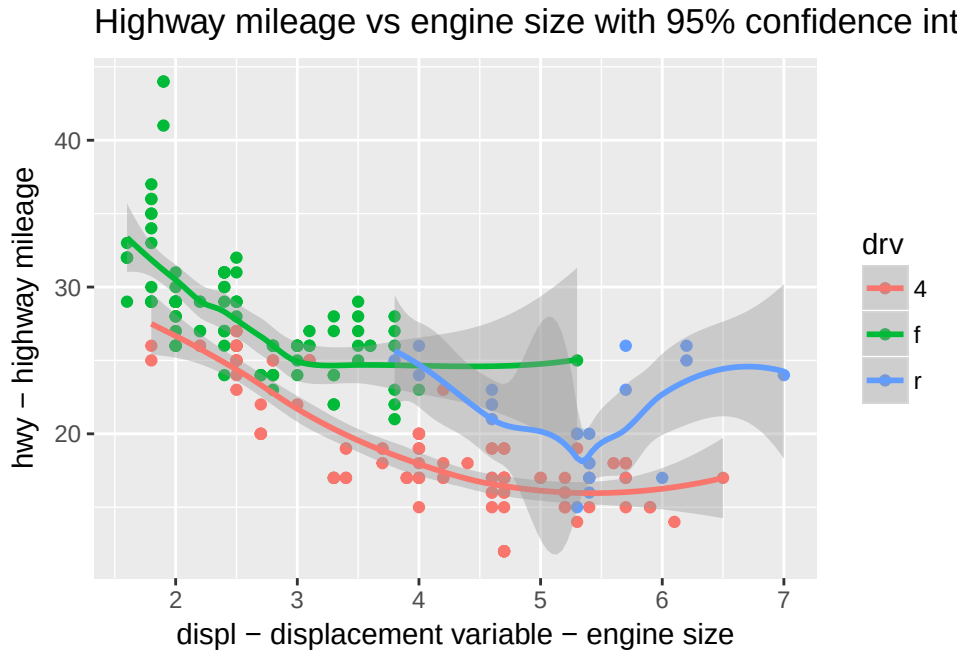



```
qplot(displ,hwy,data = mpg,color=drv,
      main = "Highway mileage vs engine size",
      xlab = "displ - displacement variable - engine size",
      ylab = "hwy - highway mileage"
        )
```



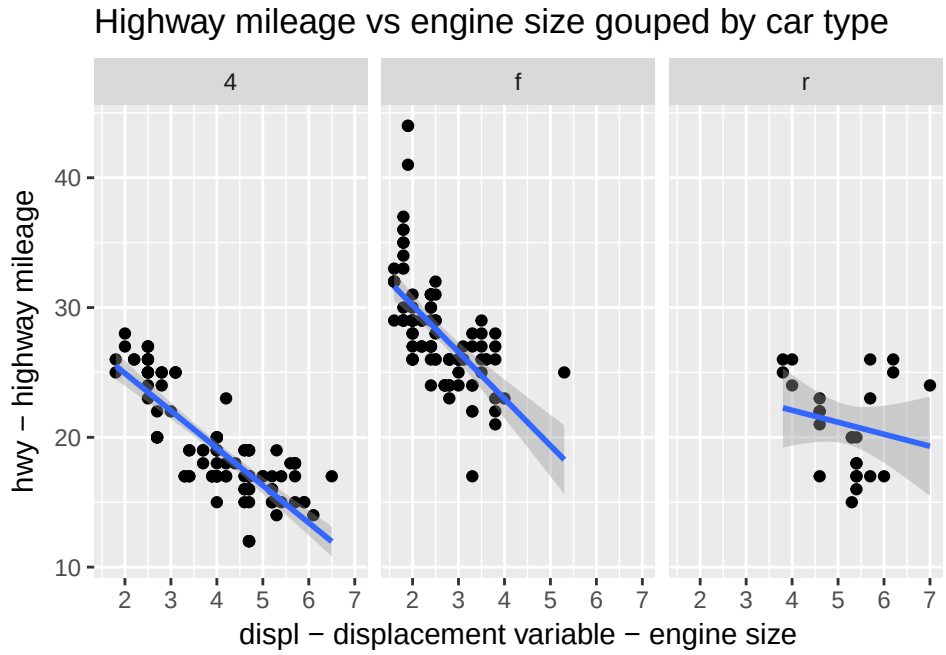

```
qplot(displ,hwy,data = mpg,color=drv,
      geom = c("point","smooth"),
      main = "Highway mileage vs engine size with 95% confidence int",
      xlab = "displ - displacement variable - engine size",
      ylab = "hwy - highway mileage"
```

```
)
```

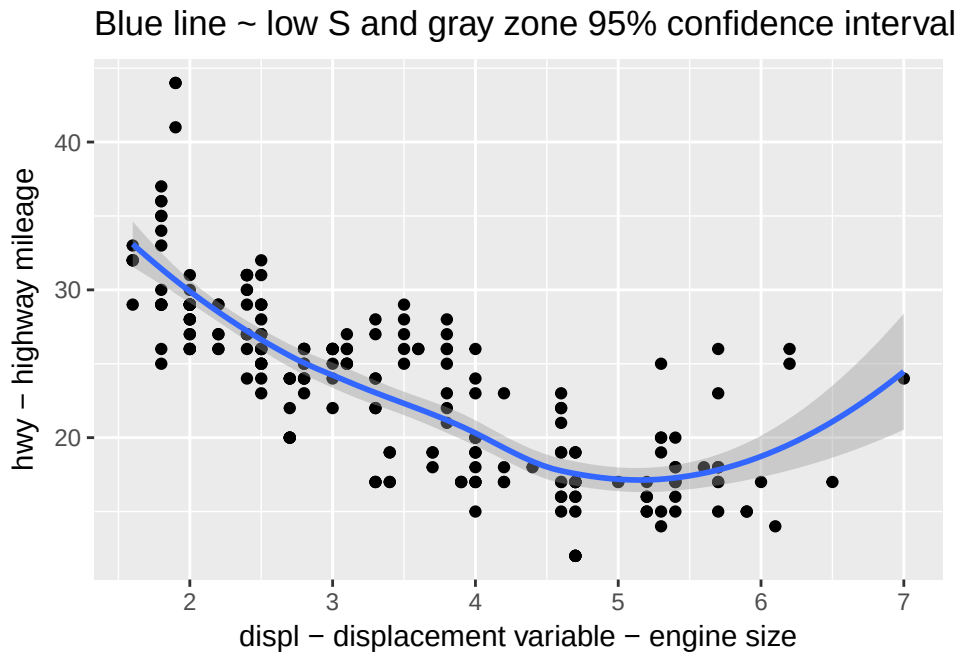


or smoothing the data

```
qplot(displ,hwy,data = mpg,
      facets = .~drv,
      #color=drv,
      #geom = c("point","smooth"),
      main = "Highway mileage vs engine size gouped by car type",
      xlab = "displ - displacement variable - engine size",
      ylab = "hwy - highway mileage") + geom_smooth(method = "lm")
```

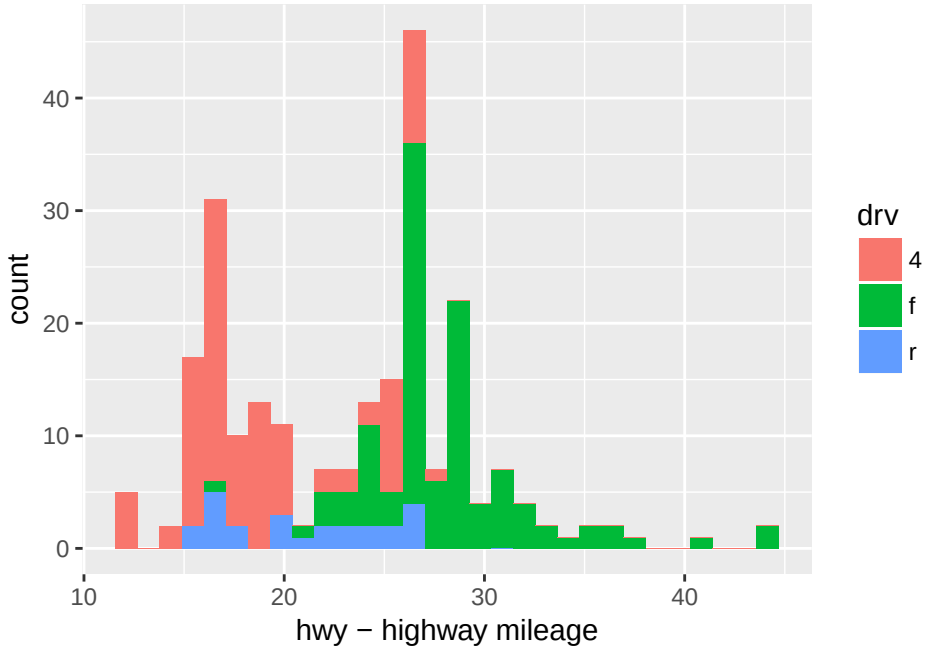


```
qplot(displ,hwy,data = mpg,
      geom = c("point","smooth"),
      xlab = "displ - displacement variable - engine size",
      ylab = "hwy - highway mileage",
      main = "Blue line ~ low S and gray zone 95% confidence interval"
        )
```



Now let's make some histograms:

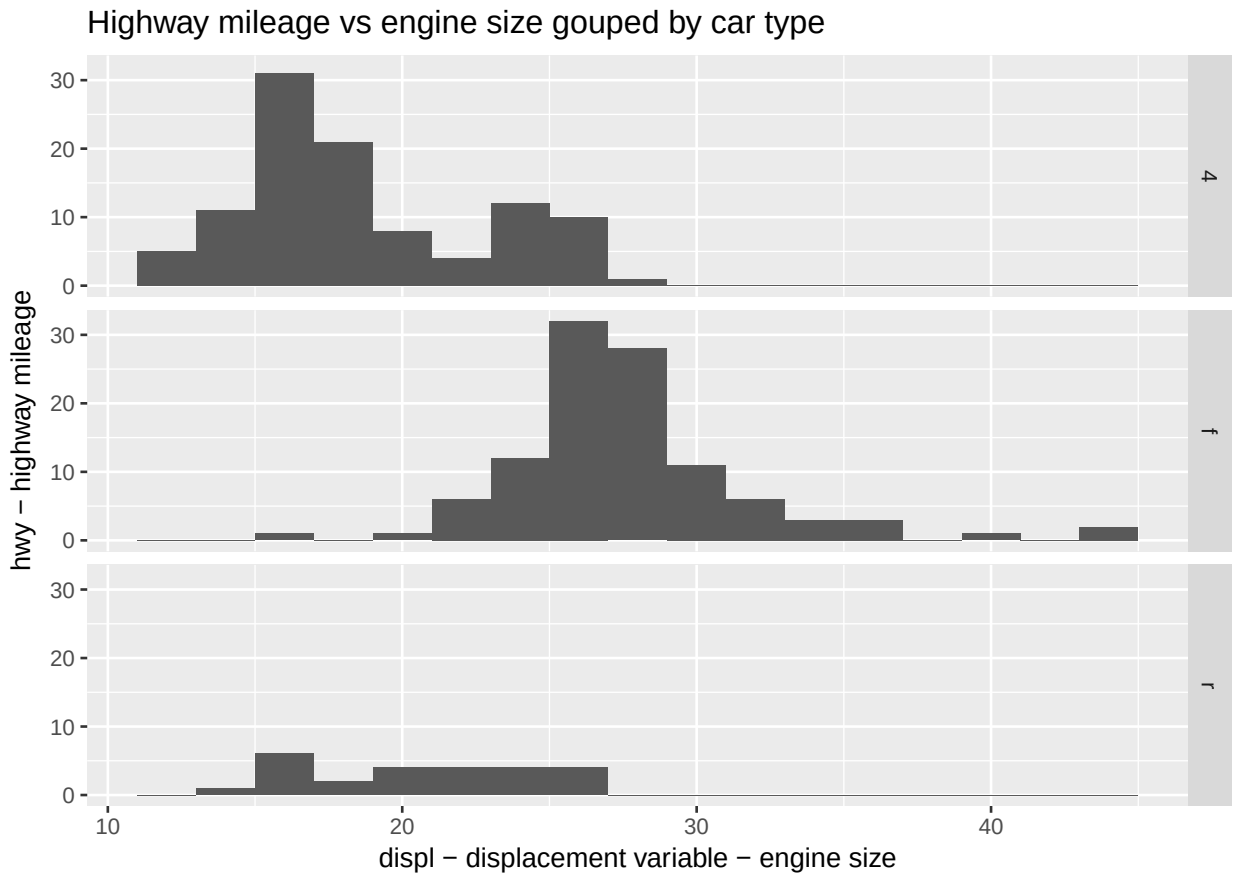
```
qplot(hwy,data = mpg,
      fill=drv,
      xlab = "hwy - highway mileage")
```



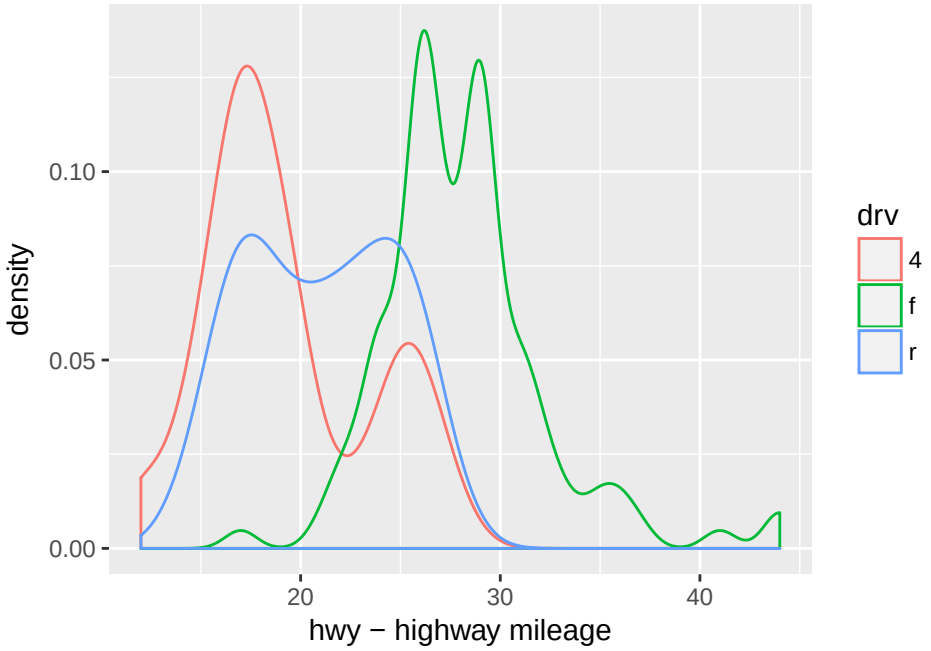

```
      main = "Histogram of highway mileage for diffrent car types"
```

or

```
qplot(hwy,data = mpg,
      facets = drv~.,
      binwidth=2,
      main = "Highway mileage vs engine size gouped by car type",
      xlab = "displ - displacement variable - engine size",
      ylab = "hwy - highway mileage"
        )
```

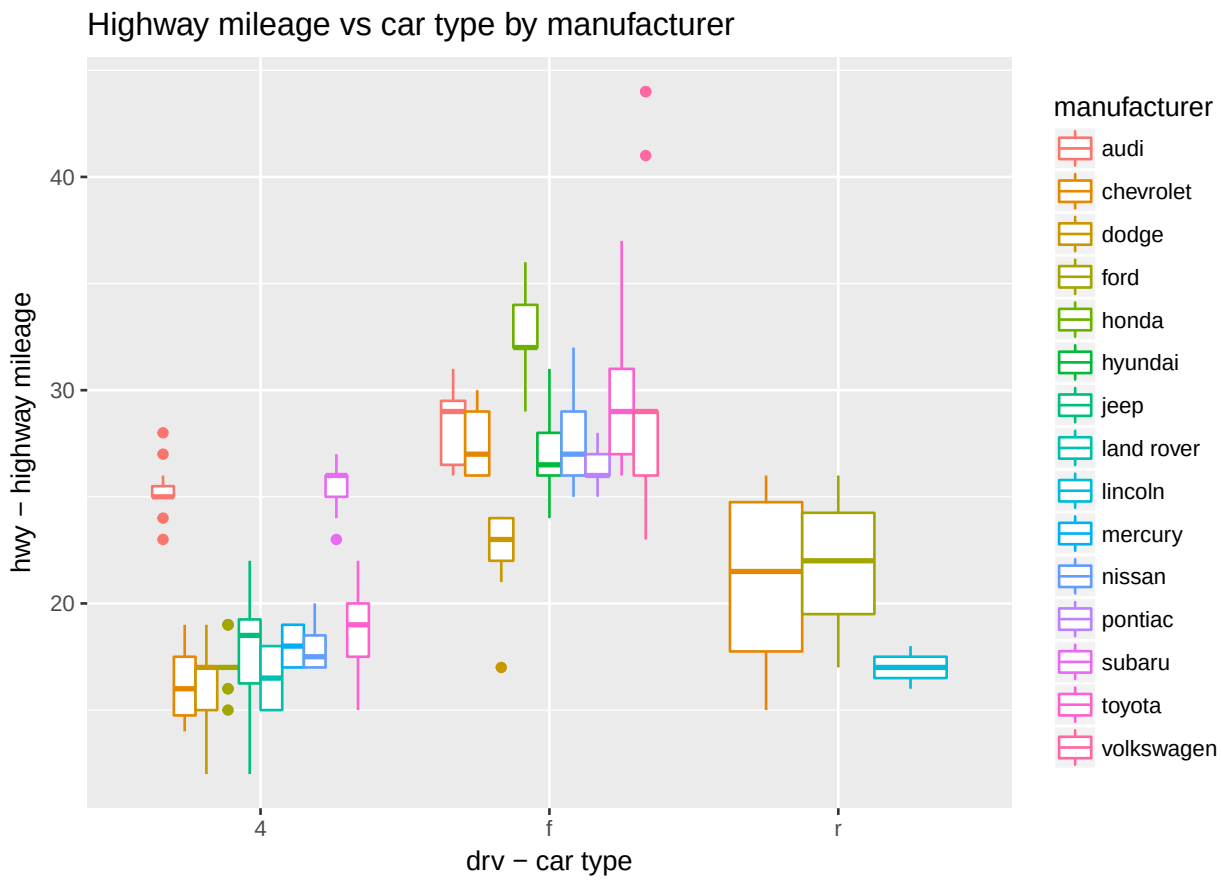



```
qplot(hwy,data = mpg,
      geom = "density",
      color= drv,
      xlab = "hwy - highway mileage")
```




```
      main = "Histogram of highway mileage for diffrent car types"
```

```
qplot(drv,hwy,data = mpg,
      geom = "boxplot",
      color=manufacturer,
      main = "Highway mileage vs car type by manufacturer",
     xlab = "drv - car type",
     ylab = "hwy - highway mileage"
      )
```


Highway mileage vs car type by manufacturer
hwy − highway mileage
40
30
20
4
f
r
drv − car type
manufacturer
audi
chevrolet
dodge
ford
honda
hyundai
jeep
land rover
lincoln
mercury
nissan
pontiac
subaru
toyota
volkswagen