R for Data Science: Chapter 3

Nick Sun

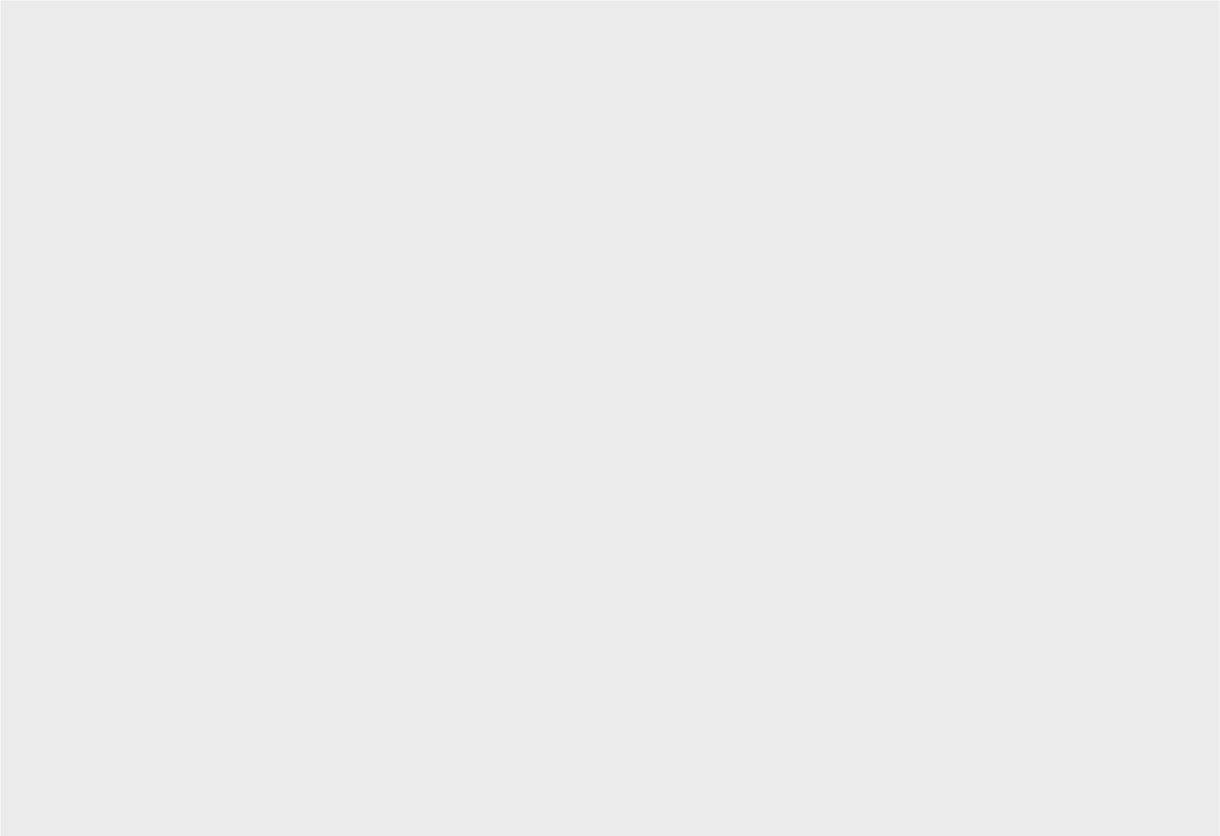
April 8, 2019

3.2.4

```
## 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" ...
```

R Markdown

```
ggplot(data =mpg)
```



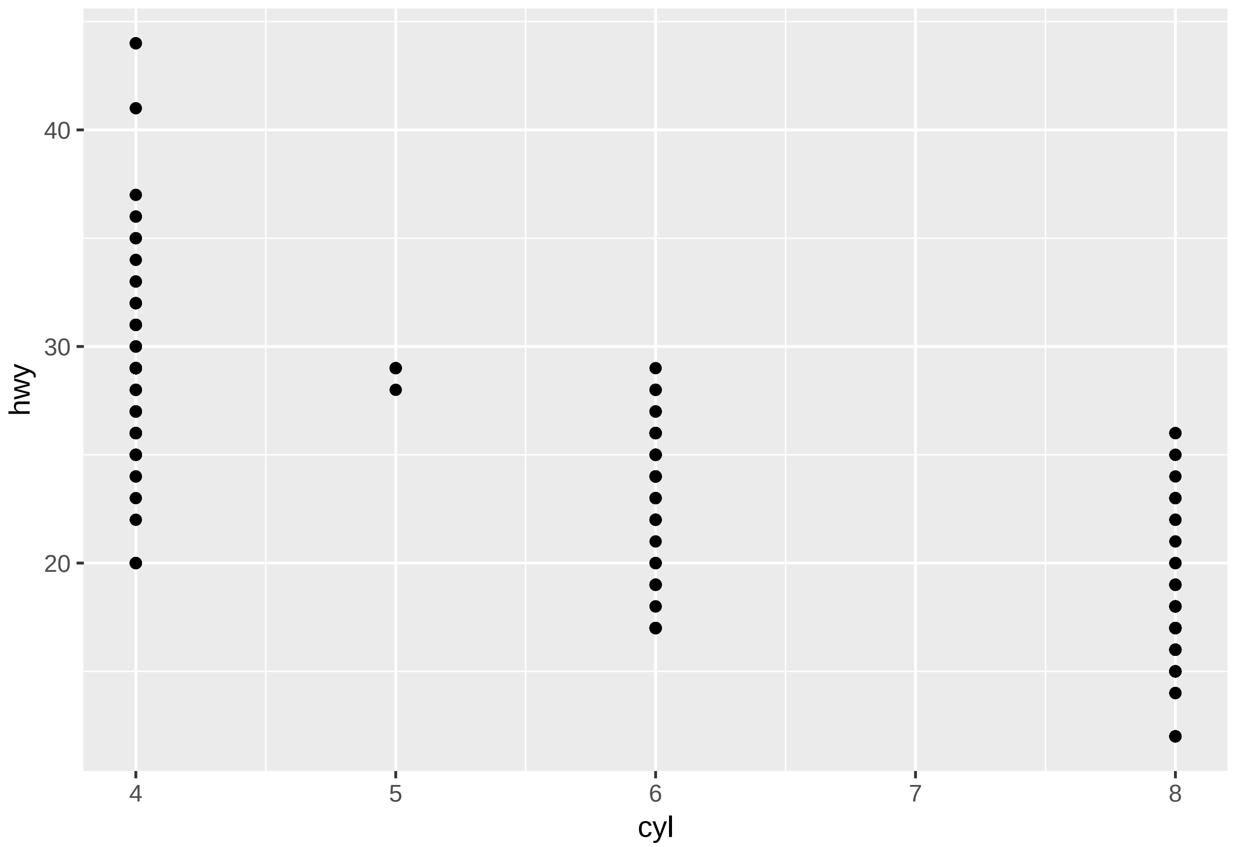
I SEE NOTHING (of value that is).

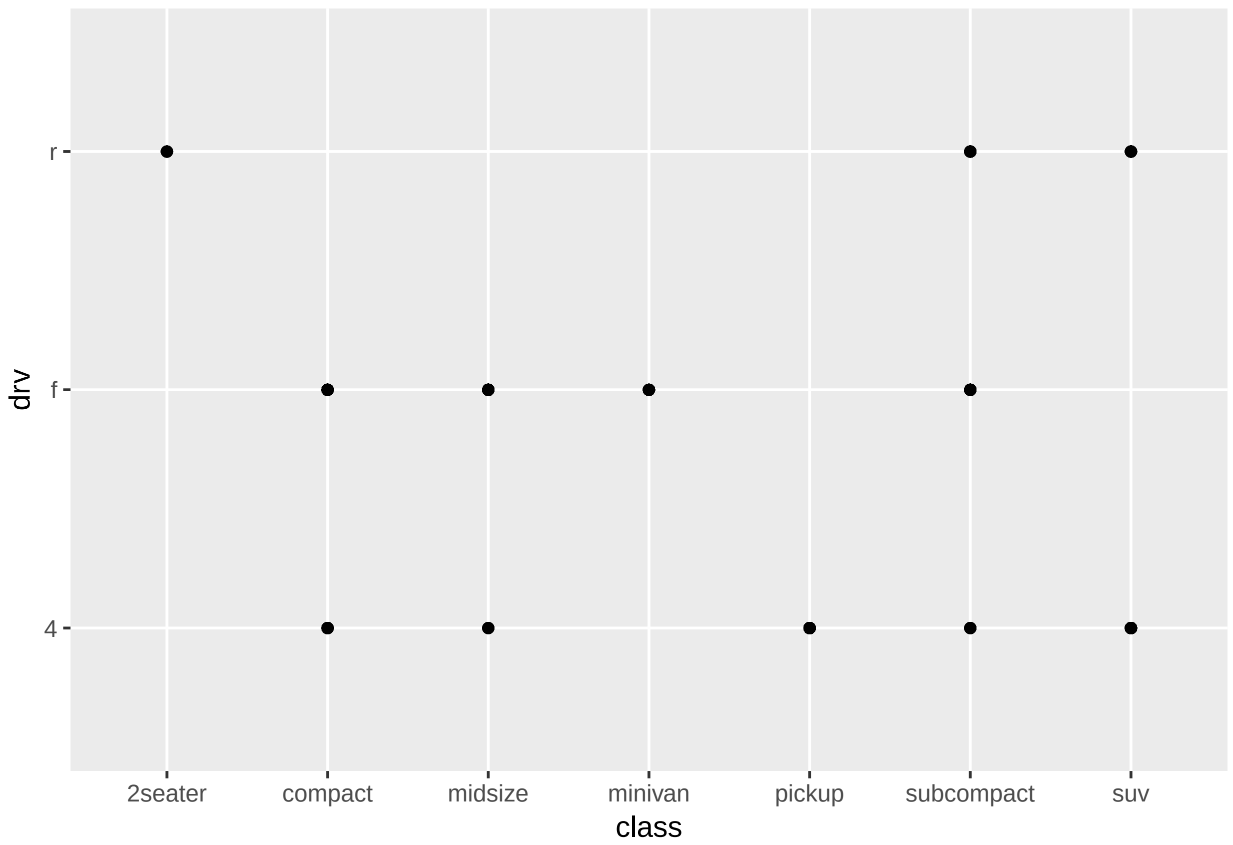
```r
dim(mpg)
```

```
## [1] 234  11
```

THERE ARE 234 ROWS AND 11 COLUMNS.

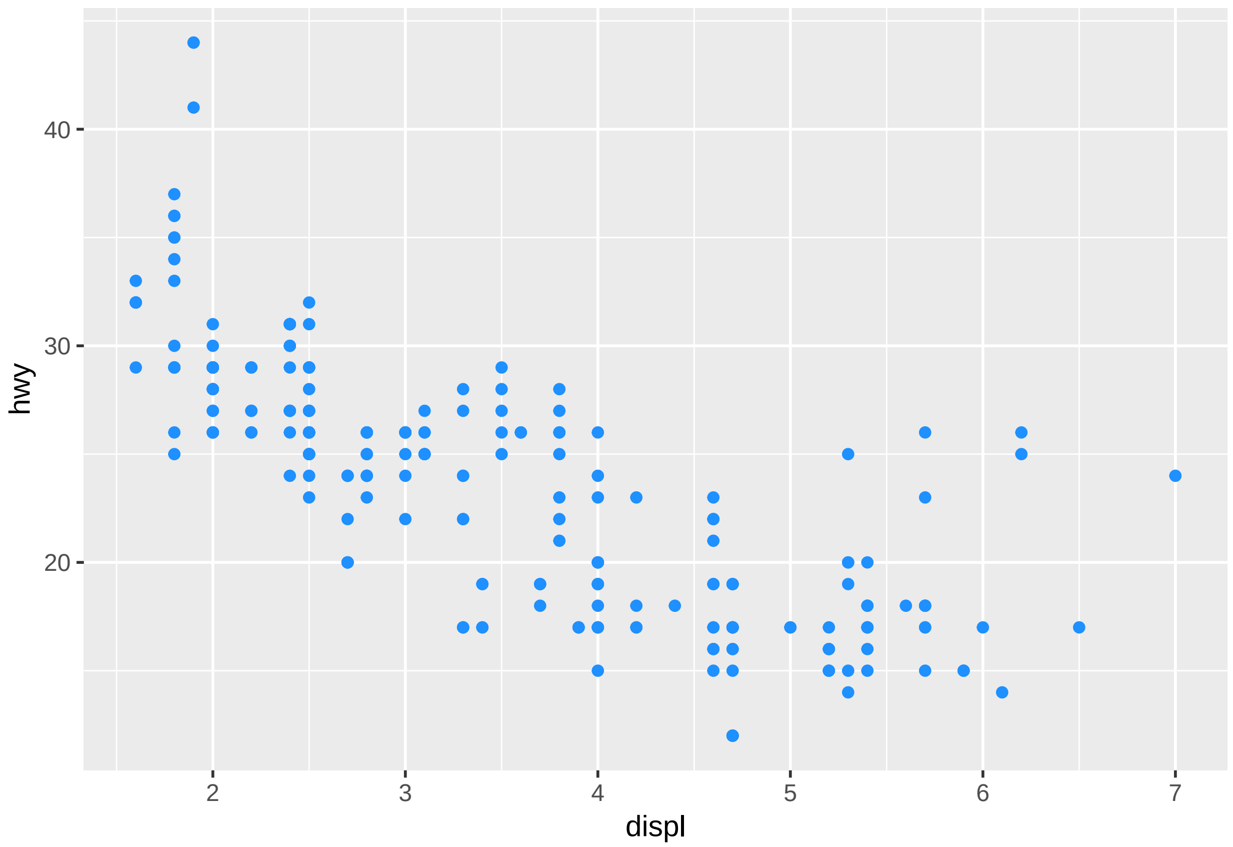
This is a categorical variable. **f** stands for front-wheel drive, **r** stands for rear wheel drive, and **4** stands for a four wheel drive car.


40
30
20
hwy
4
5
6
7
8
cyl



This scatterplot really doesn't tell us anything. Almost all the different classes of cars have some cars that fall into one of the drive categories.

3.3.1

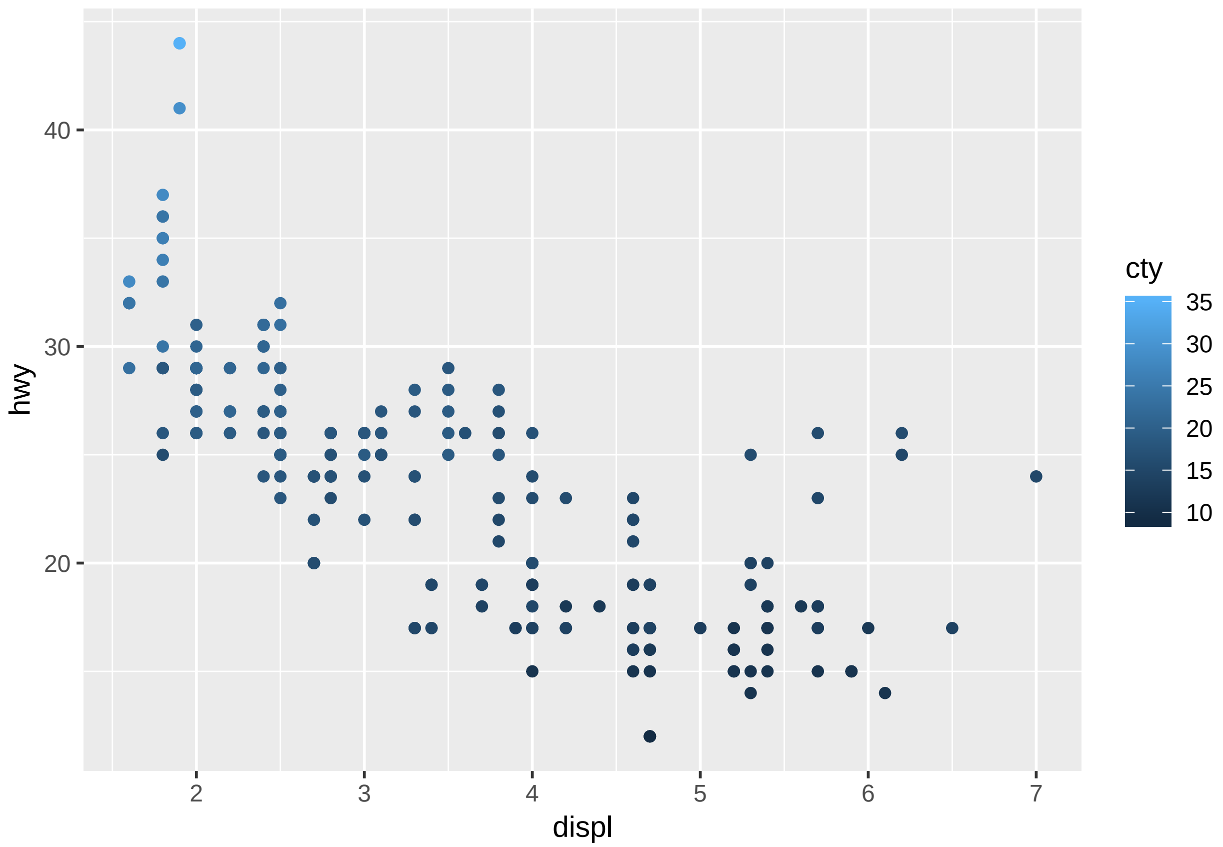


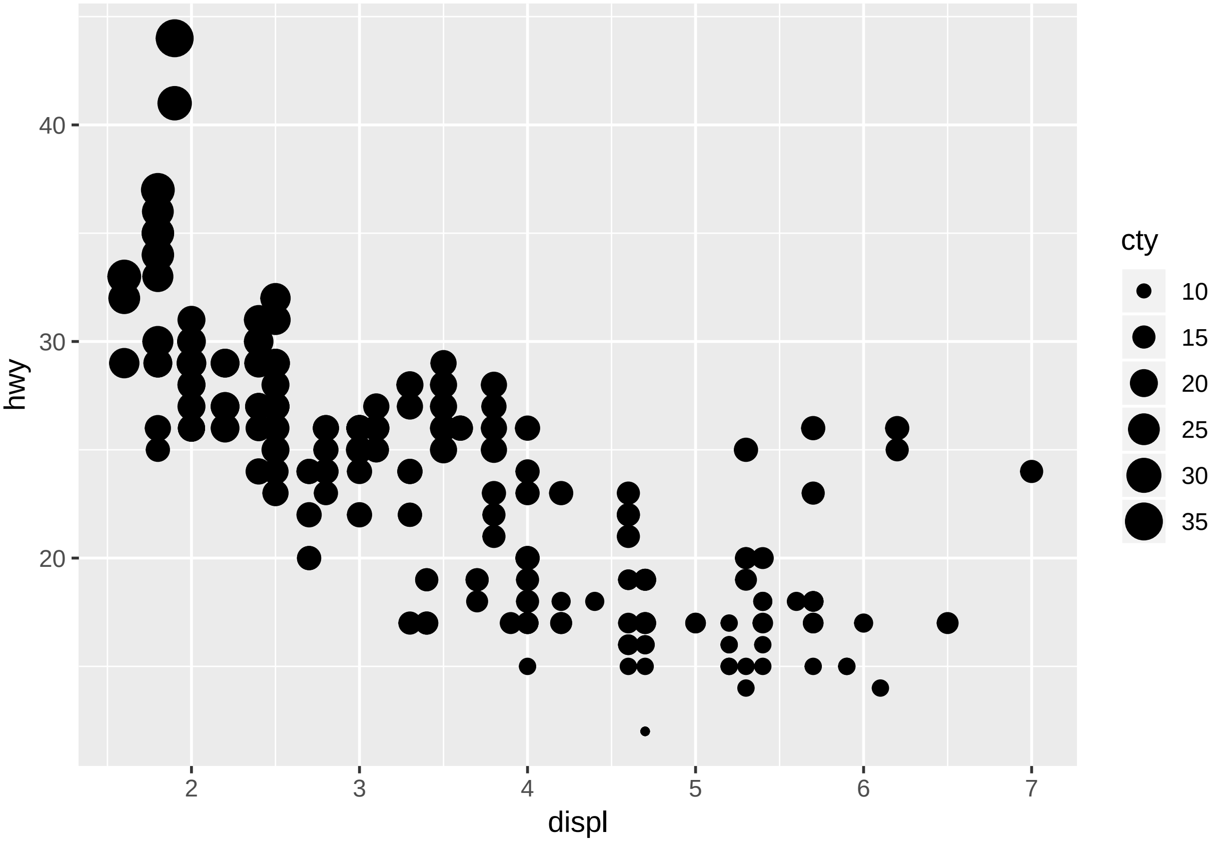
The color aesthetic belongs in the `geom_point()` layer, not inside the `aes()` mapping layer.

```
## 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" ...
```

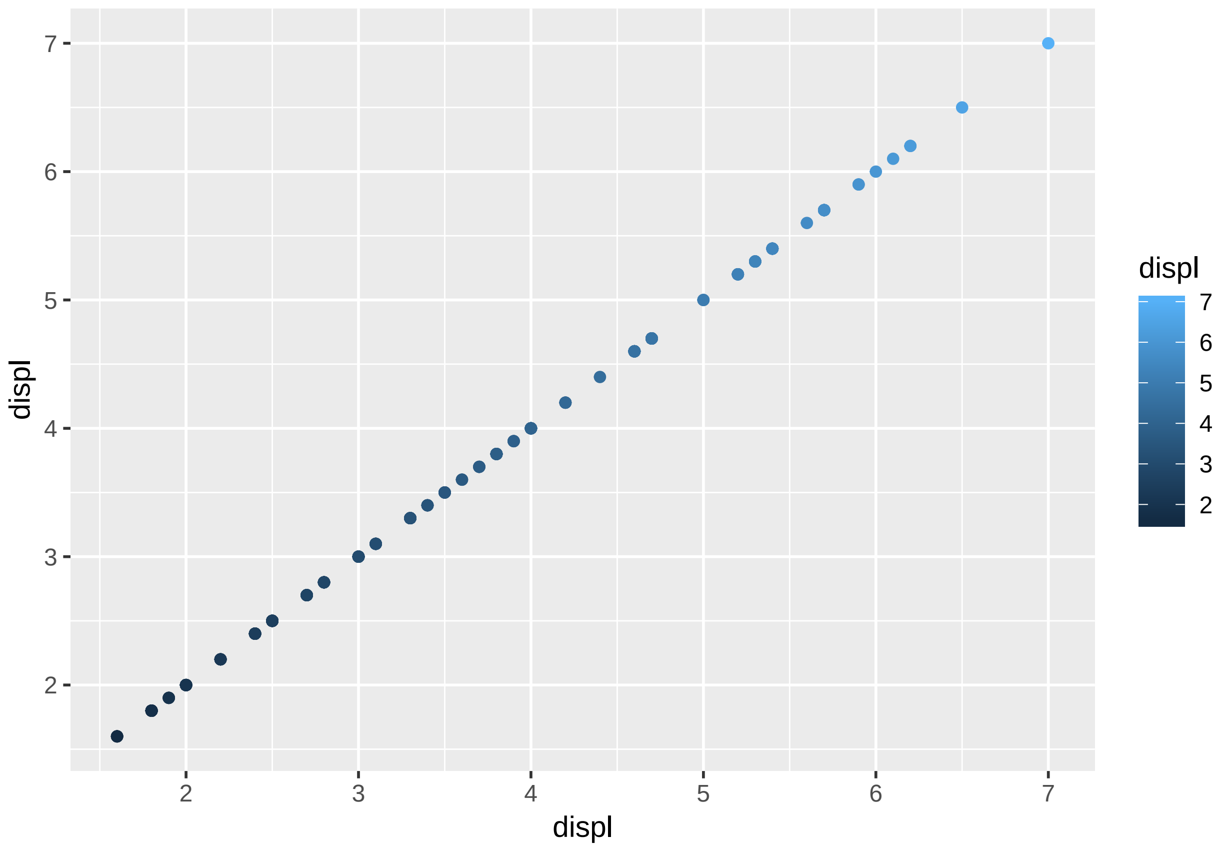
```
## starting httpd help server ... done
```

Manufacturer, model, cyl, trans, drv, fl, and class are all categorical variables. The quantitative variables are displ, cty, hwy. I use the `str()` function which automatically displays the data type of each column in the tibble.

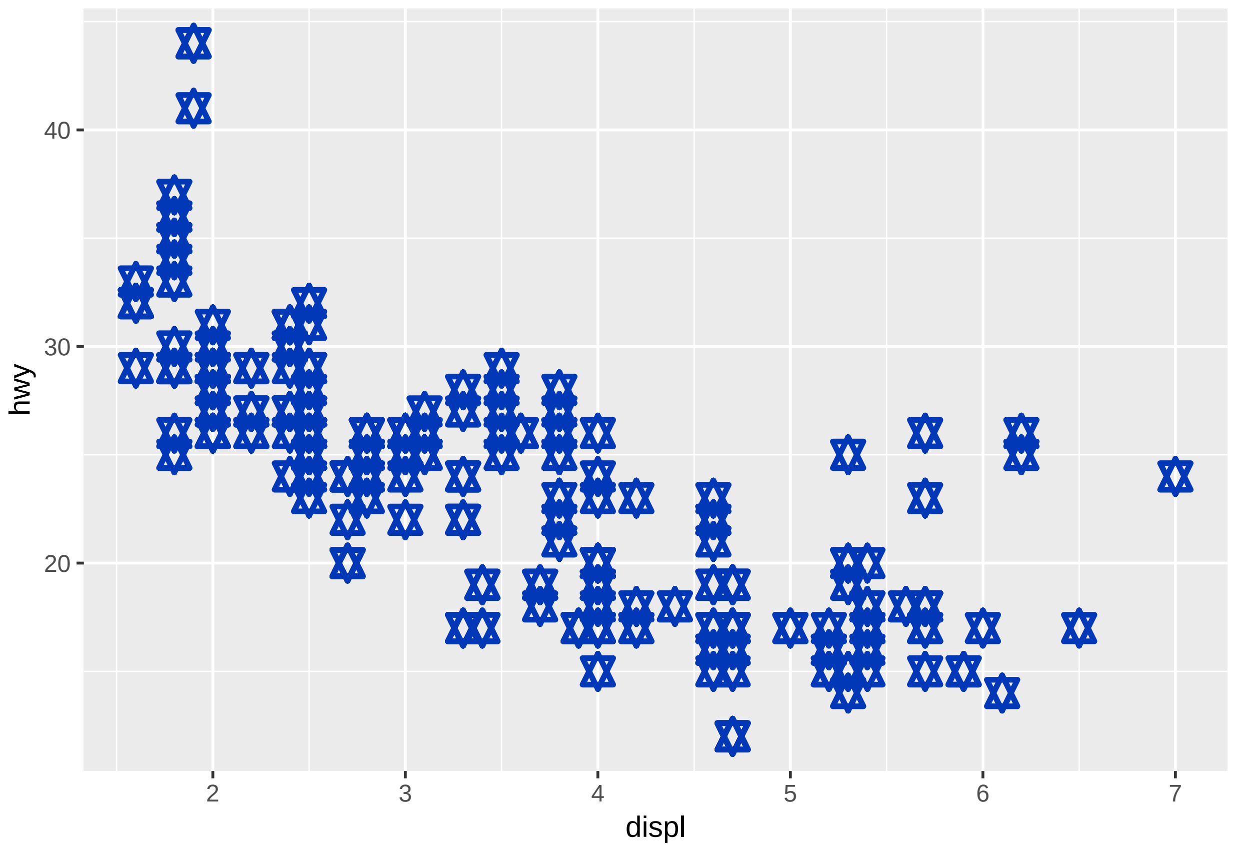

hwy
40
30
20
2
3
4
5
6
7
displ
cty
35
30
25
20
15
10



Color mapped to a continuous variable creates a color gradient for each of the points. Likewise, the size variable creates a size gradient - bigger points correspond to larger values of displ. A continuous variable cannot be mapped to `shape`.

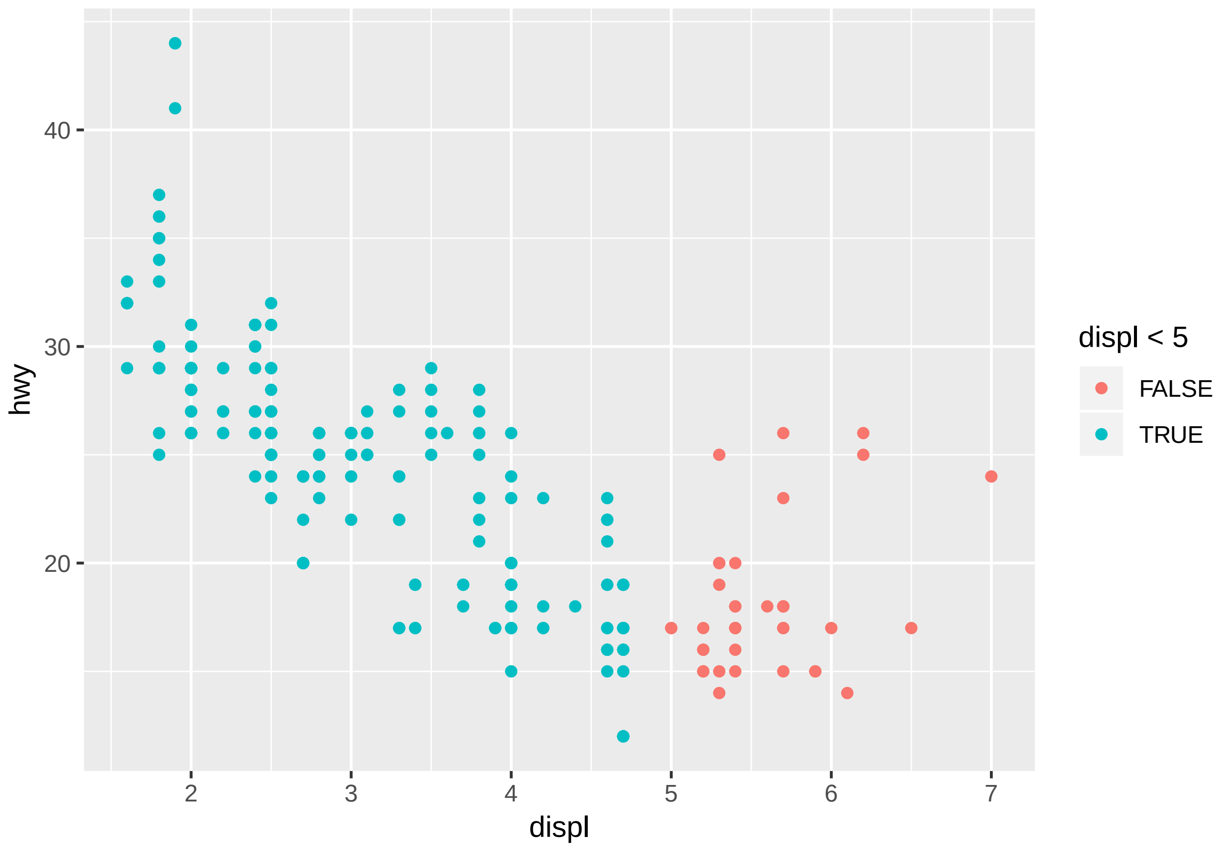


You can do this, but it creates a pretty useless plot in my opinion.



The `stroke` aesthetic appears to change the width of the borders of each of the points.

```
ggplot(data = mpg, aes(x = displ, y = hwy, color = displ < 5)) +
  geom_point()
```



It binarizes the continuous variable! Pretty neat.