Exploring Data

2023-07-07

Load the necessary packages

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.2     v purrr   1.0.1
## v tibble  3.2.1     v dplyr   1.1.2
## v tidyr   1.3.0     v stringr 1.5.0
## v readr   2.1.3     v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(nycflights13)
library(gapminder)
library(Lahman)
```

Load the datasets into the environment

```
data("flights")
data("gapminder")
data("mpg")
```

Importing the dataset into the environment using 'dput()'

```
# generates R code to recreate the data
dput(mtcars)
```

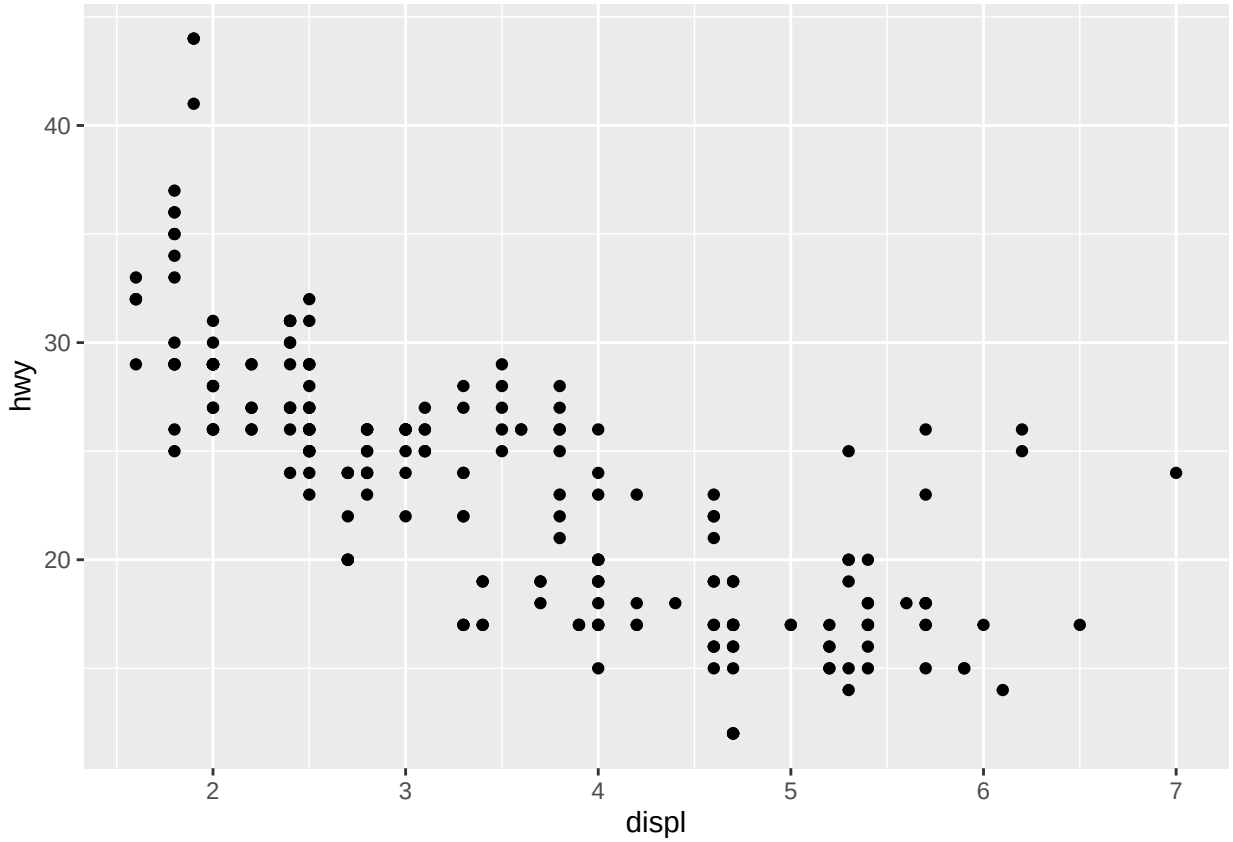
```
## structure(list(mpg = c(21, 21, 22.8, 21.4, 18.7, 18.1, 14.3,
## 24.4, 22.8, 19.2, 17.8, 16.4, 17.3, 15.2, 10.4, 10.4, 14.7, 32.4,
## 30.4, 33.9, 21.5, 15.5, 15.2, 13.3, 19.2, 27.3, 26, 30.4, 15.8,
## 19.7, 15, 21.4), cyl = c(6, 6, 4, 6, 8, 6, 8, 4, 4, 6, 6, 8,
## 8, 8, 8, 8, 8, 4, 4, 4, 4, 8, 8, 8, 8, 4, 4, 4, 8, 6, 8, 4),
##     disp = c(160, 160, 108, 258, 360, 225, 360, 146.7, 140.8,
##     167.6, 167.6, 275.8, 275.8, 275.8, 472, 460, 440, 78.7, 75.7,
##     71.1, 120.1, 318, 304, 350, 400, 79, 120.3, 95.1, 351, 145,
##     301, 121), hp = c(110, 110, 93, 110, 175, 105, 245, 62, 95,
##     123, 123, 180, 180, 180, 205, 215, 230, 66, 52, 65, 97, 150,
##     150, 245, 175, 66, 91, 113, 264, 175, 335, 109), drat = c(3.9,
##     3.9, 3.85, 3.08, 3.15, 2.76, 3.21, 3.69, 3.92, 3.92, 3.92,
##     3.07, 3.07, 3.07, 2.93, 3, 3.23, 4.08, 4.93, 4.22, 3.7, 2.76,
##     3.15, 3.73, 3.08, 4.08, 4.43, 3.77, 4.22, 3.62, 3.54, 4.11
##     ), wt = c(2.62, 2.875, 2.32, 3.215, 3.44, 3.46, 3.57, 3.19,
##     3.15, 3.44, 3.44, 4.07, 3.73, 3.78, 5.25, 5.424, 5.345, 2.2,
##     1.615, 1.835, 2.465, 3.52, 3.435, 3.84, 3.845, 1.935, 2.14,
##     1.513, 3.17, 2.77, 3.57, 2.78), qsec = c(16.46, 17.02, 18.61,
##     19.44, 17.02, 20.22, 15.84, 20, 22.9, 18.3, 18.9, 17.4, 17.6,
##     18, 17.98, 17.82, 17.42, 19.47, 18.52, 19.9, 20.01, 16.87,
```

```
##     17.3, 15.41, 17.05, 18.9, 16.7, 16.9, 14.5, 15.5, 14.6, 18.6
##     ), vs = c(0, 0, 1, 1, 0, 1, 0, 1, 1, 1, 1, 0, 0, 0, 0, 0,
##     0, 1, 1, 1, 1, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 1), am = c(1,
##     1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1,
##     0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1), gear = c(4, 4, 4, 3,
##     3, 3, 3, 4, 4, 4, 4, 3, 3, 3, 3, 3, 3, 4, 4, 4, 3, 3, 3,
##     3, 3, 4, 5, 5, 5, 5, 5, 4), carb = c(4, 4, 1, 1, 2, 1, 4,
##     2, 2, 4, 4, 3, 3, 3, 4, 4, 4, 1, 2, 1, 1, 2, 2, 4, 2, 1,
##     2, 2, 4, 6, 8, 2)), row.names = c("Mazda RX4", "Mazda RX4 Wag",
## "Datsun 710", "Hornet 4 Drive", "Hornet Sportabout", "Valiant",
## "Duster 360", "Merc 240D", "Merc 230", "Merc 280", "Merc 280C",
## "Merc 450SE", "Merc 450SL", "Merc 450SLC", "Cadillac Fleetwood",
## "Lincoln Continental", "Chrysler Imperial", "Fiat 128", "Honda Civic",
## "Toyota Corolla", "Toyota Corona", "Dodge Challenger", "AMC Javelin",
## "Camaro Z28", "Pontiac Firebird", "Fiat X1-9", "Porsche 914-2",
## "Lotus Europa", "Ford Pantera L", "Ferrari Dino", "Maserati Bora",
## "Volvo 142E"), class = "data.frame")
```

```
# copy the R code into 'mtcars <-'
mtcars <- structure(list(mpg = c(21, 21, 22.8, 21.4, 18.7, 18.1, 14.3,
                      24.4, 22.8, 19.2, 17.8, 16.4, 17.3, 15.2, 10.4, 10.4, 14.7, 32.4,
                      30.4, 33.9, 21.5, 15.5, 15.2, 13.3, 19.2, 27.3, 26, 30.4, 15.8,
                      19.7, 15, 21.4), cyl = c(6, 6, 4, 6, 8, 6, 8, 4, 4, 6, 6, 8,
                                               8, 8, 8, 8, 8, 4, 4, 4, 4, 8, 8, 8, 8, 4, 4, 4, 8, 6, 8
              disp = c(160, 160, 108, 258, 360, 225, 360, 146.7, 140.8,
                       167.6, 167.6, 275.8, 275.8, 275.8, 472, 460, 440, 78.7, 75.7,
                       71.1, 120.1, 318, 304, 350, 400, 79, 120.3, 95.1, 351, 145,
                       301, 121), hp = c(110, 110, 93, 110, 175, 105, 245, 62, 95,
                                         123, 123, 180, 180, 180, 205, 215, 230, 66, 52, 65, 97, 150,
                                         150, 245, 175, 66, 91, 113, 264, 175, 335, 109), drat = c(3.9
                                                                                                   3.9
                                                                                                   3.0
                                                                                                   3.1
                                         ), wt = c(2.62, 2.875, 2.32, 3.215, 3.44, 3.46, 3.57, 3.19,
                                                   3.15, 3.44, 3.44, 4.07, 3.73, 3.78, 5.25, 5.424, 5.
                                                   1.615, 1.835, 2.465, 3.52, 3.435, 3.84, 3.845, 1.93
                                                   1.513, 3.17, 2.77, 3.57, 2.78), qsec = c(16.46, 17.
                                                                                            19.44, 17.
                                                                                            18, 17.98,
                                                                                            17.3, 15.4
                                                   ), vs = c(0, 0, 1, 1, 0, 1, 0, 1, 1, 1, 1, 0, 0, 0,
                                                             0, 1, 1, 1, 1, 0, 0, 0, 0, 1, 0, 1, 0, 0,
```

Create a ggplot

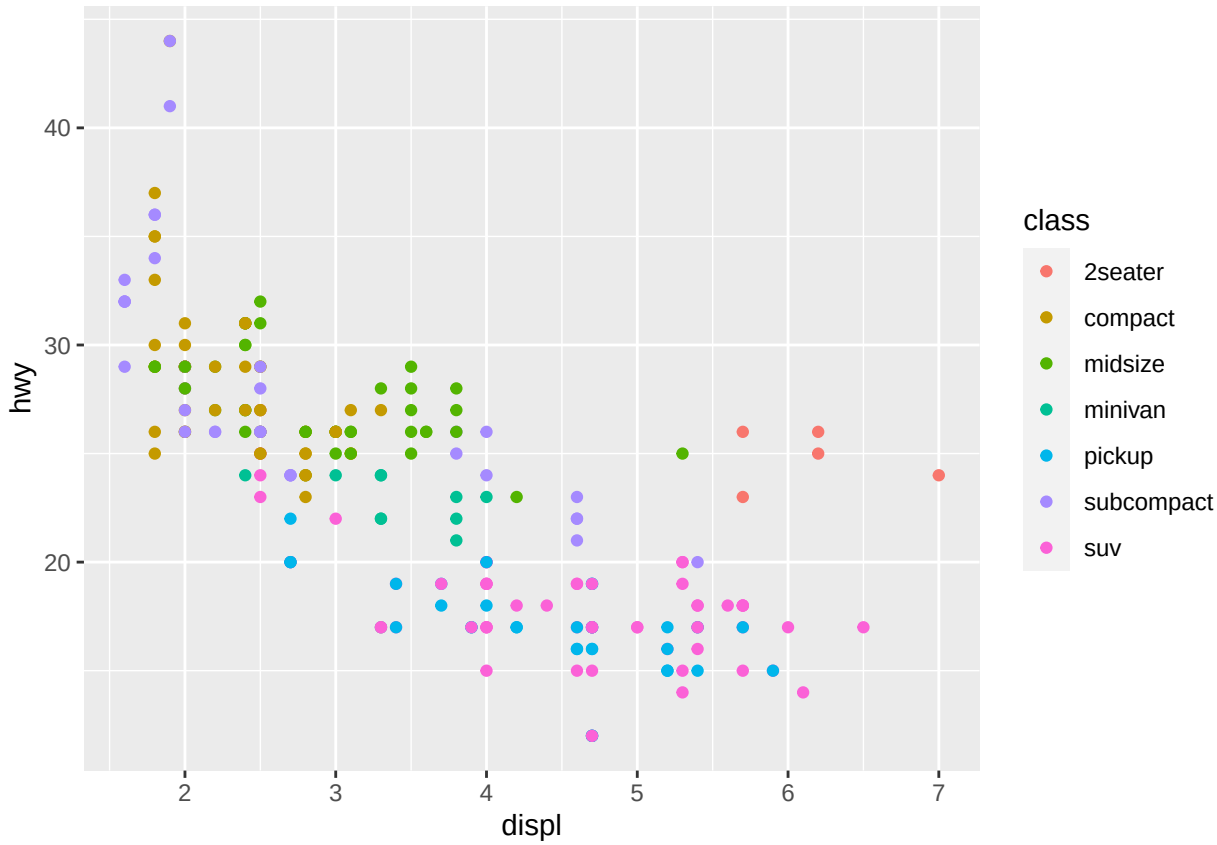
```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y= hwy))
```




```
# Negative relationship between engine size and fuel efficiency, some points do not follow the trend, c
```

Incorporate 'class' by mapping it to an aesthetic

```
ggplot(data = mpg)+
  geom_point(mapping = aes(x = displ, y = hwy, colour = class))
```

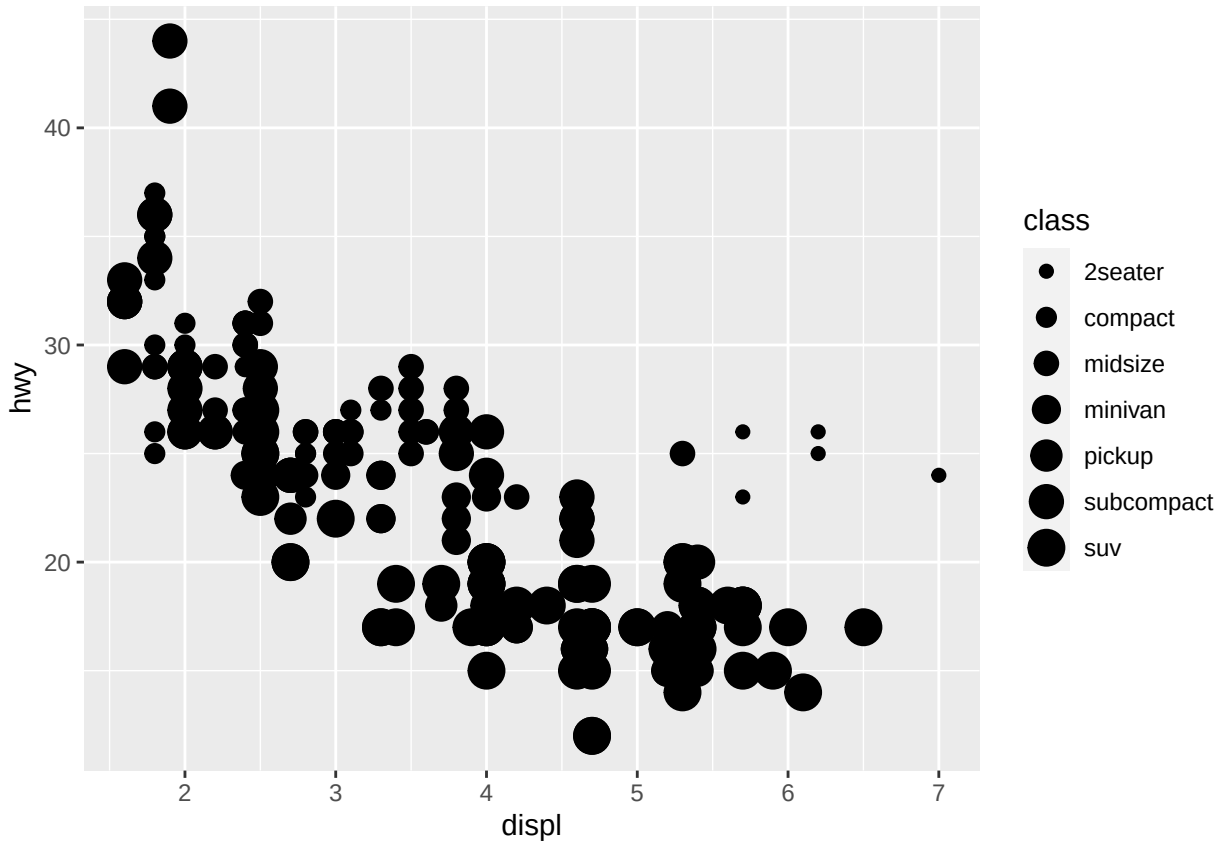



```
# The points relate to 2-seaters, these cars have larger engines but smaller bodies which improves thei
```

Map 'class' to 'size aesthetic'

```
ggplot(data = mpg)+
  geom_point(mapping = aes(x = displ, y = hwy, size = class))
```

```
## Warning: Using size for a discrete variable is not advised.
```

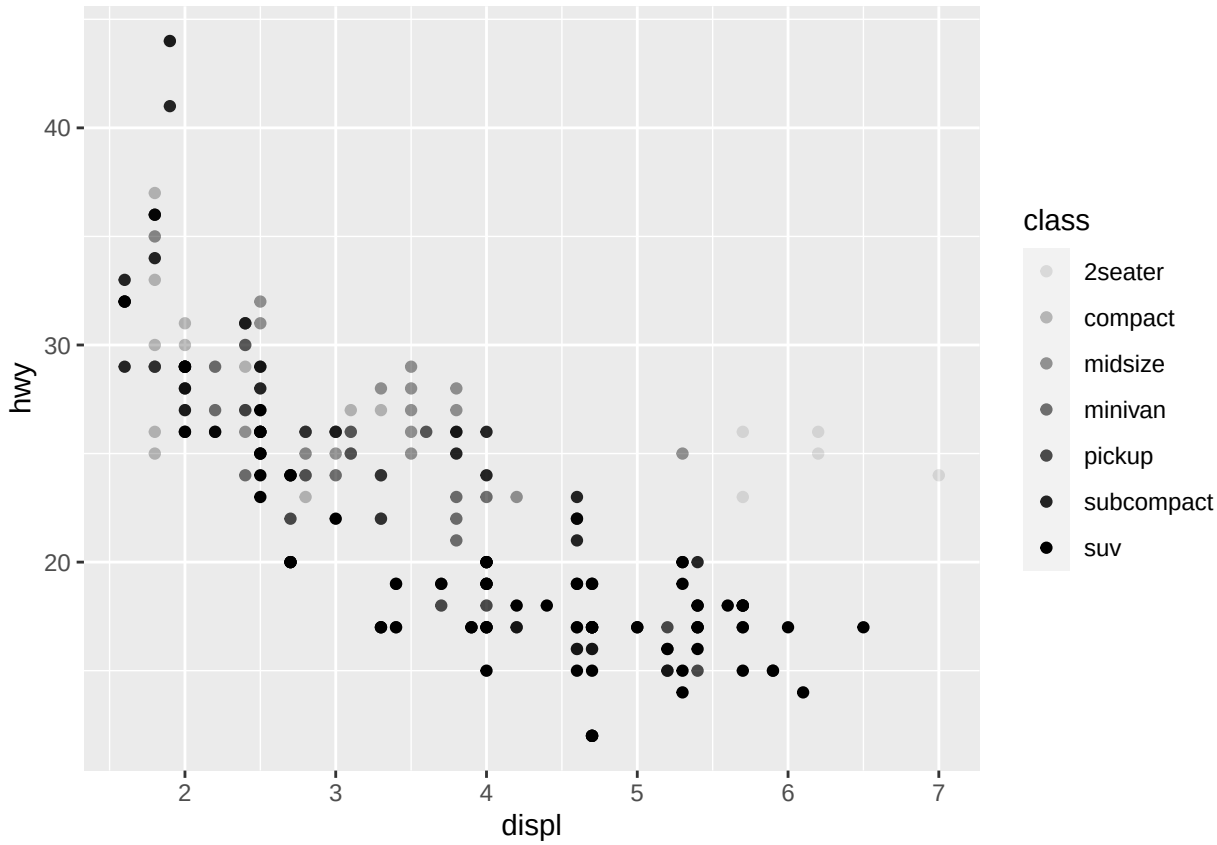



```
# Size of each point corresponds to classification
# Mapping an unordered variable (class) to an ordered aesthetic (size) is not a good idea, hence the wa
```

Map 'class' to 'alpha aesthetic'

```
ggplot(data = mpg)+
  geom_point(mapping = aes(x = displ, y = hwy, alpha = class))
```

```
## Warning: Using alpha for a discrete variable is not advised.
```

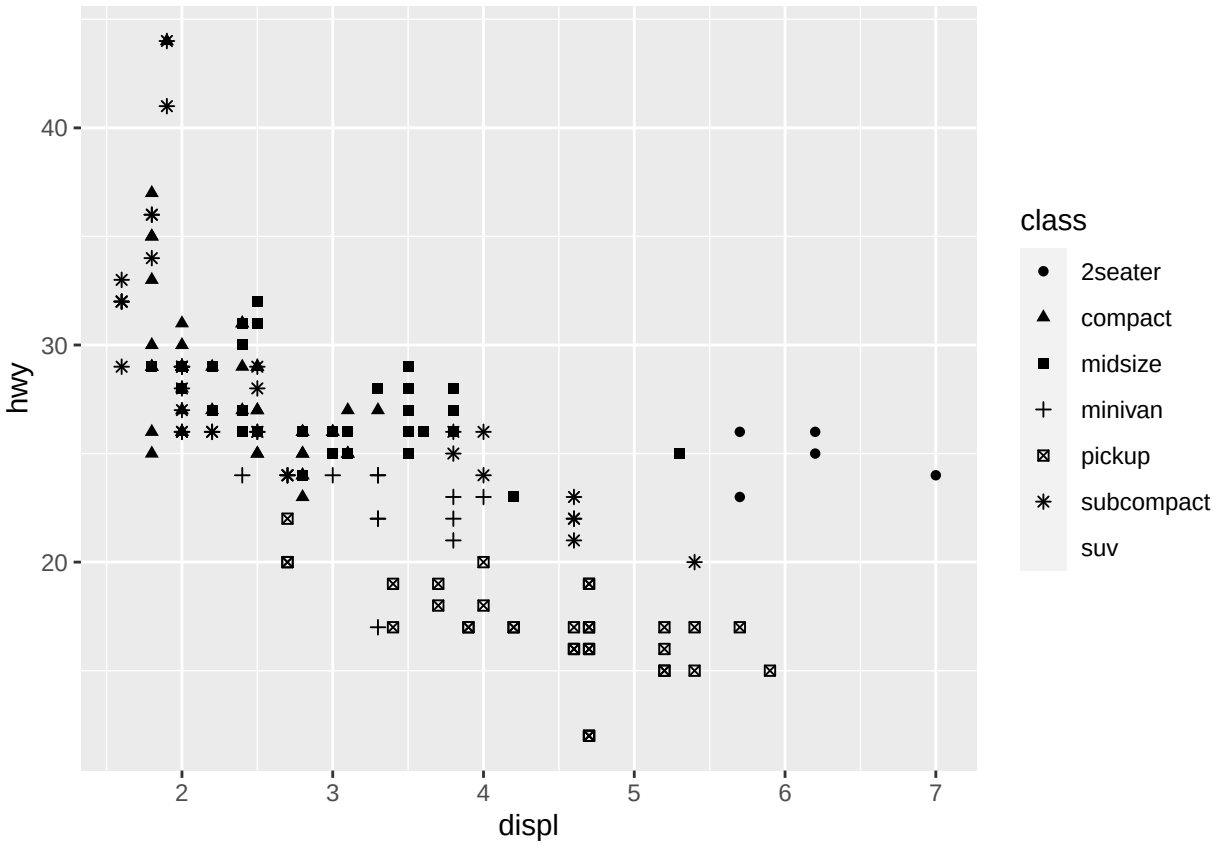


Map 'class' to 'shape aesthetic'

```
ggplot(data = mpg)+
  geom_point(mapping = aes(x = displ, y = hwy, shape = class))
```

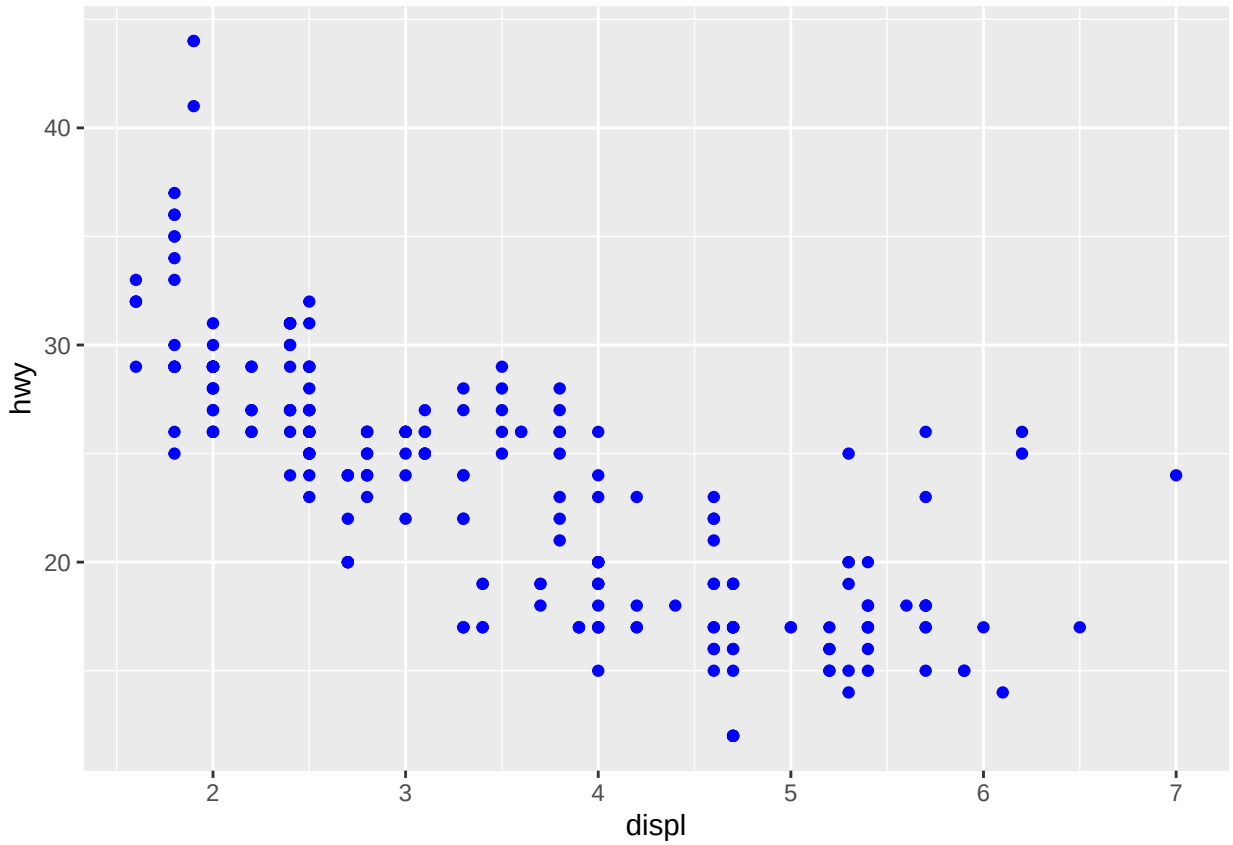
```
## Warning: The shape palette can deal with a maximum of 6 discrete values because
## more than 6 becomes difficult to discriminate; you have 7. Consider
## specifying shapes manually if you must have them.
```

```
## Warning: Removed 62 rows containing missing values (`geom_point()`).
```



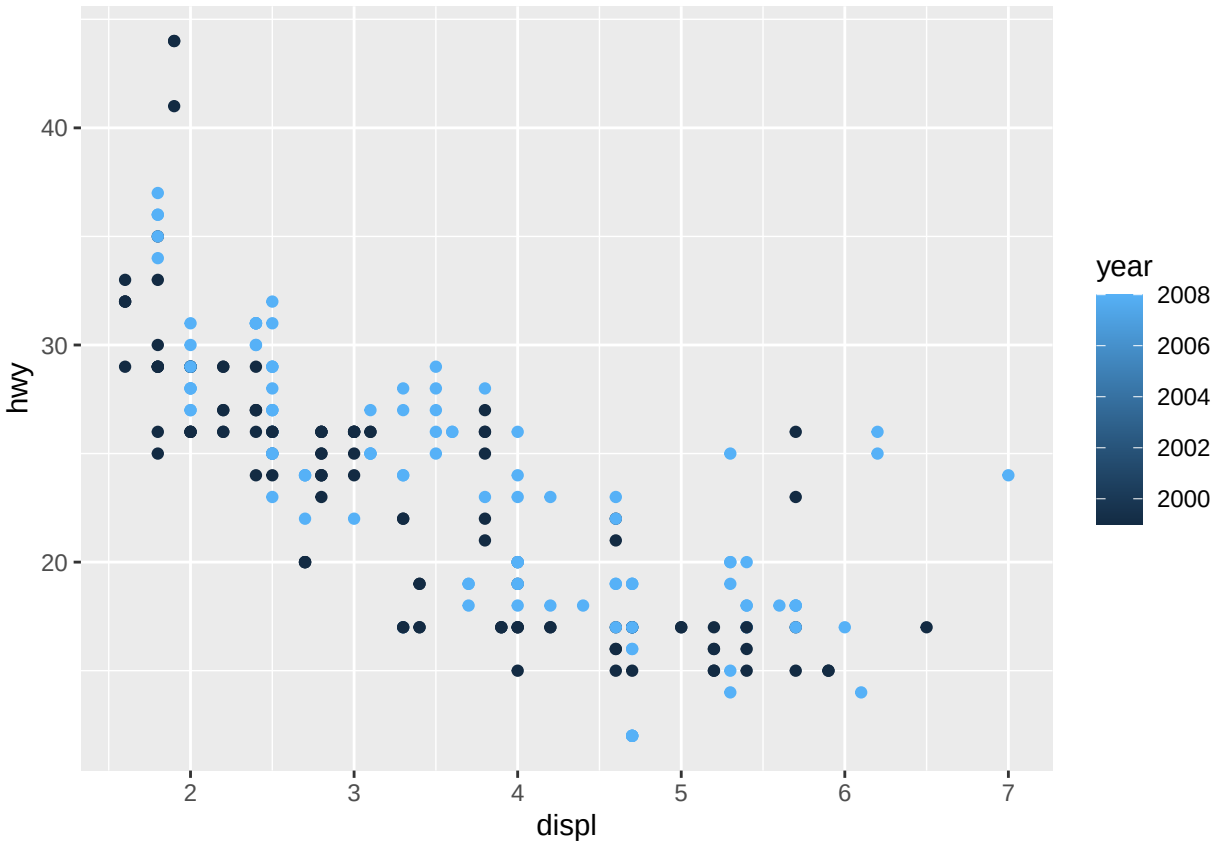
Make all the points blue

```
ggplot(data = mpg) +
  geom_point(mapping = aes (x = displ, y = hwy), colour = "blue")
```



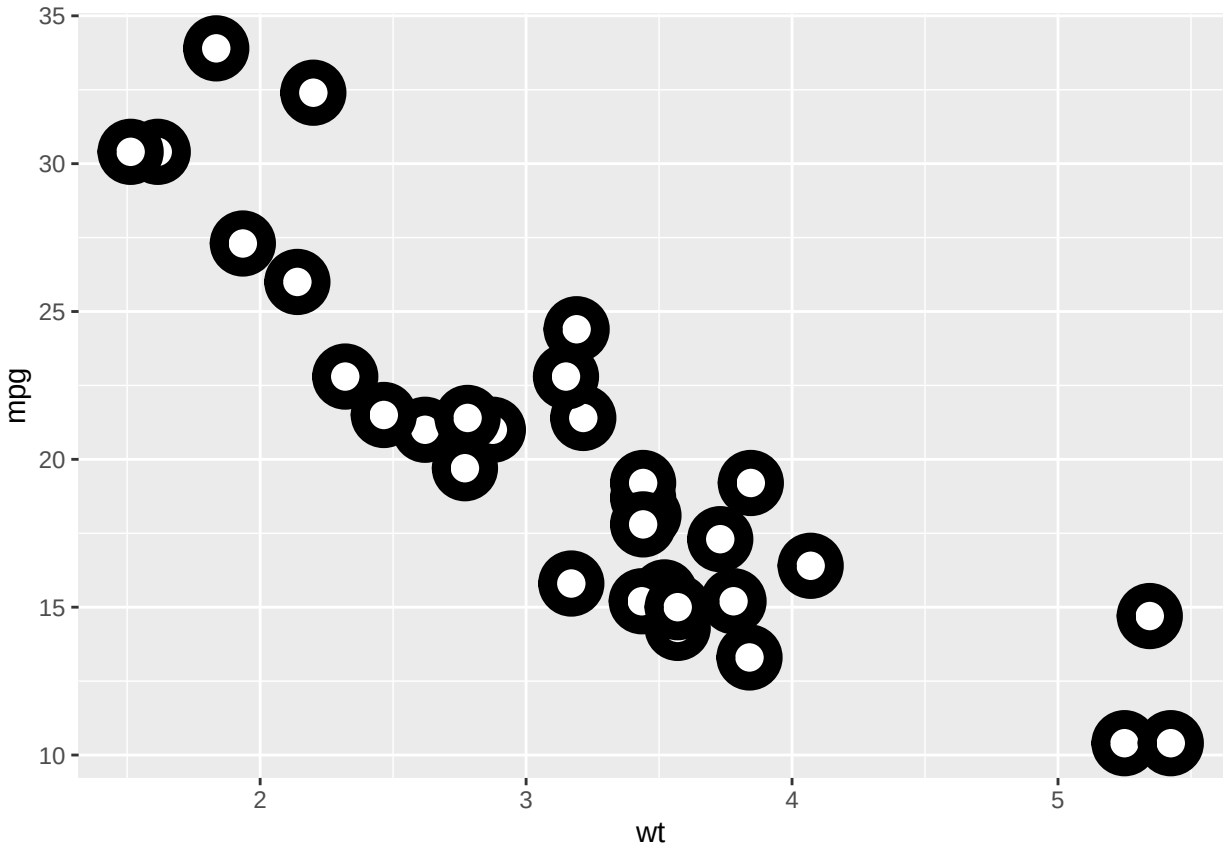
Mapping a continuous variable to colour

```
ggplot(data = mpg)+
  geom_point(mapping = aes(x = displ, y = hwy, colour = year))
```



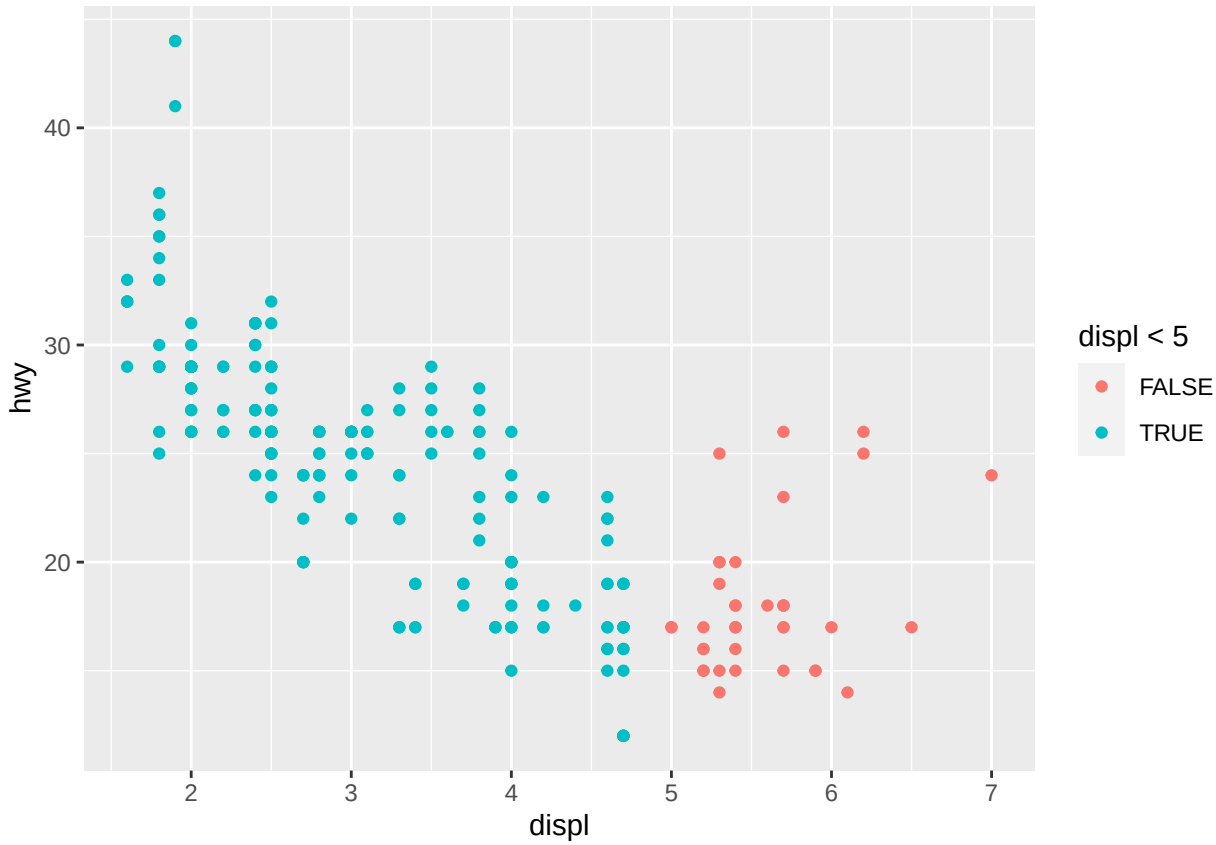
Stroke aesthetic - fills the outside of shape

```
ggplot(data = mtcars, aes(x = wt, y = mpg)) +
  geom_point(shape = 21, colour = "black", fill = "white", size = 5, stroke = 5)
```



Making the aesthetic something other than a variable name

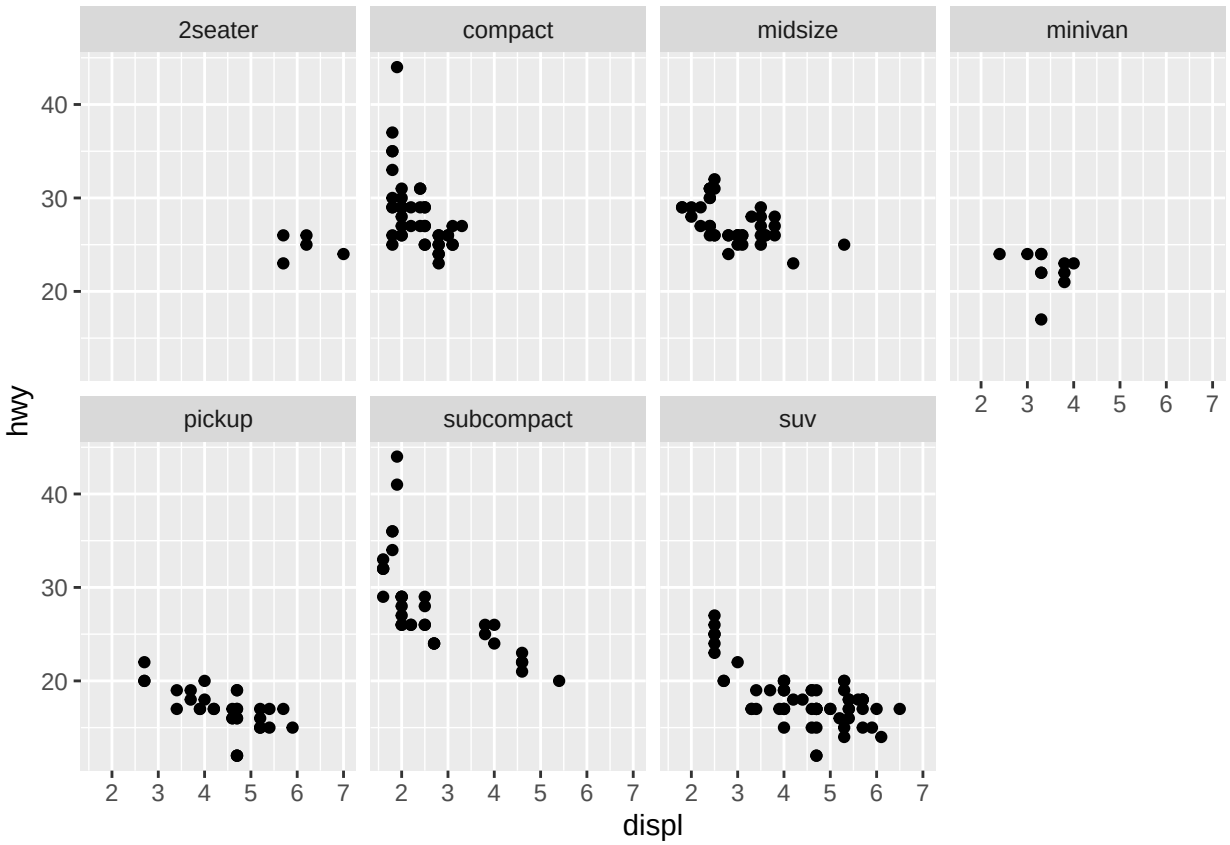
```
ggplot(data = mpg)+
  geom_point(mapping = aes(x = displ, y = hwy, colour = displ < 5))
```




```
# sets colour based on the specified engine size, below and above 5.
```

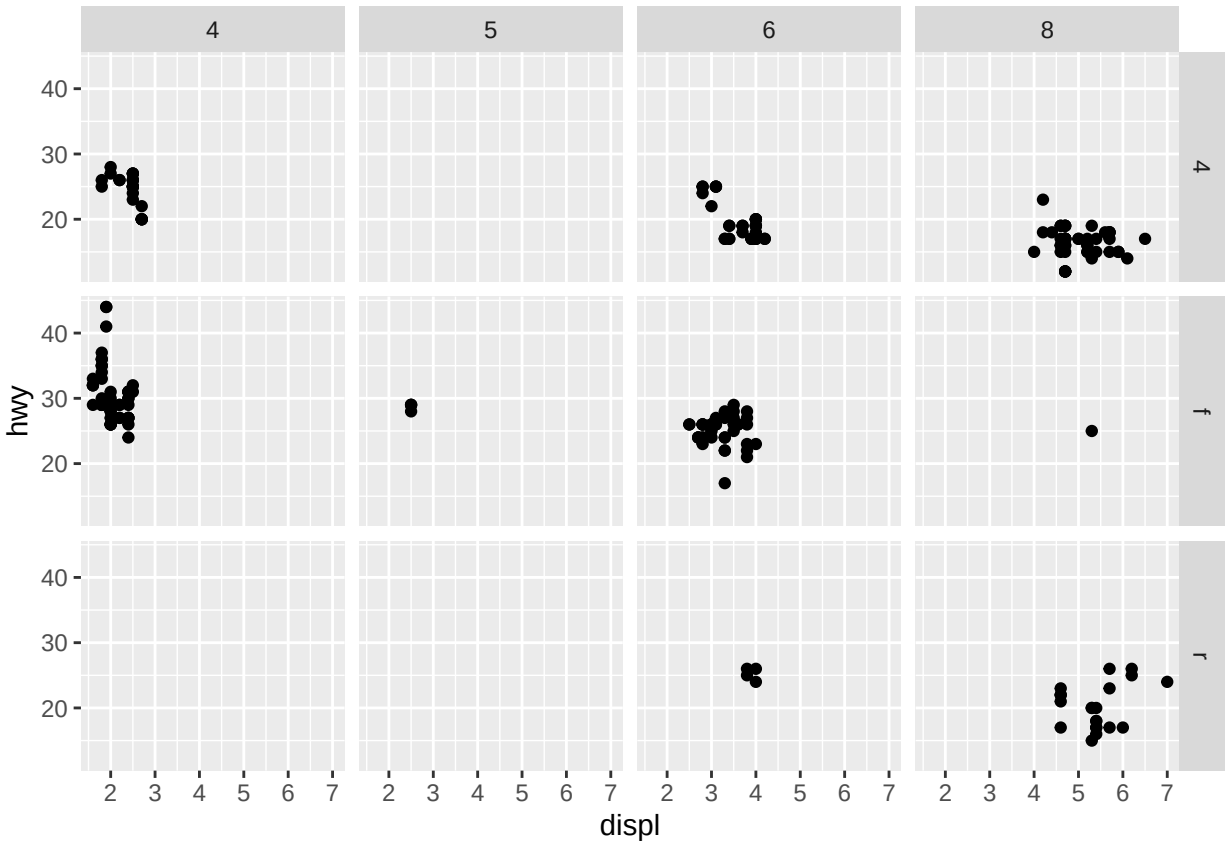
Faceting - 1 variable

```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy)) +
  facet_wrap(~ class, nrow = 2)
```



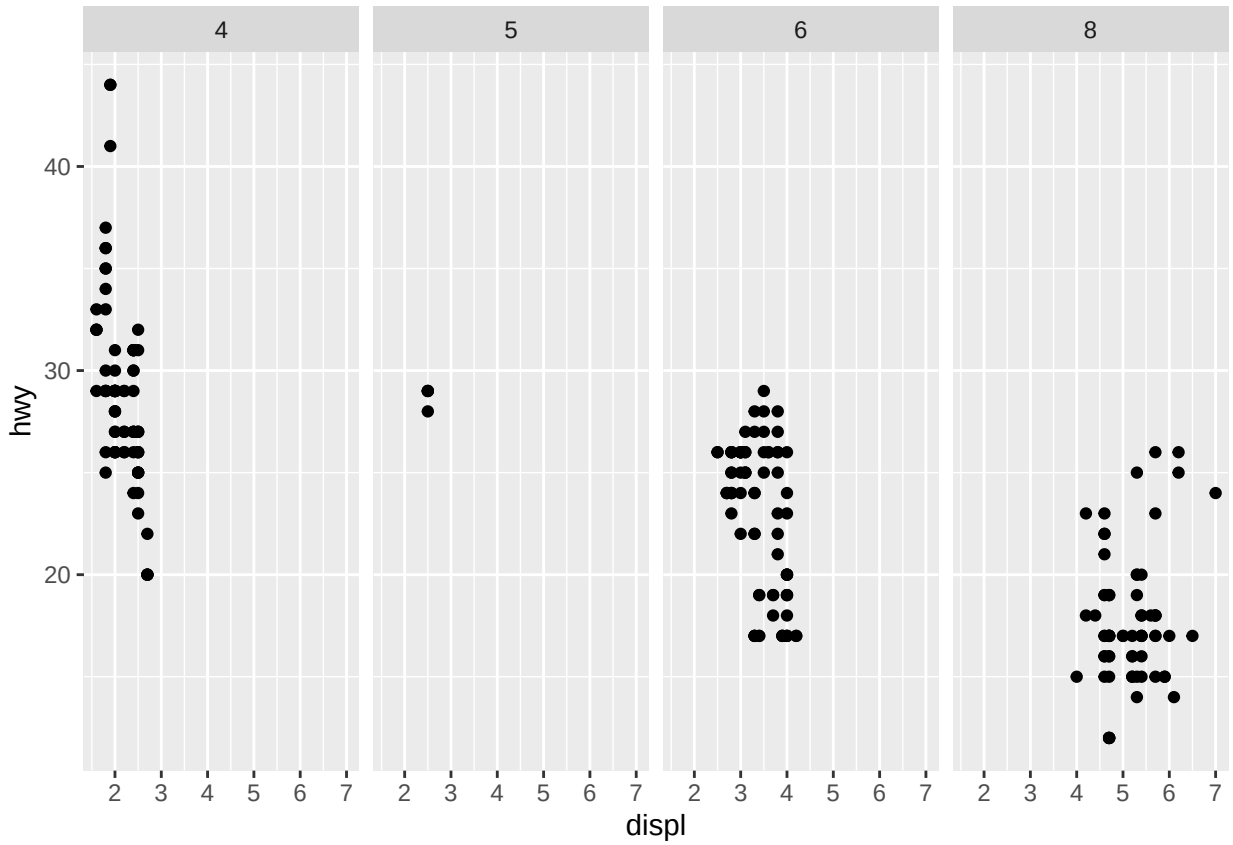
Faceting - 2 variables

```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy)) +
  facet_grid(drv ~ cyl) # use facet_grid() instead of facet_wrap()
```

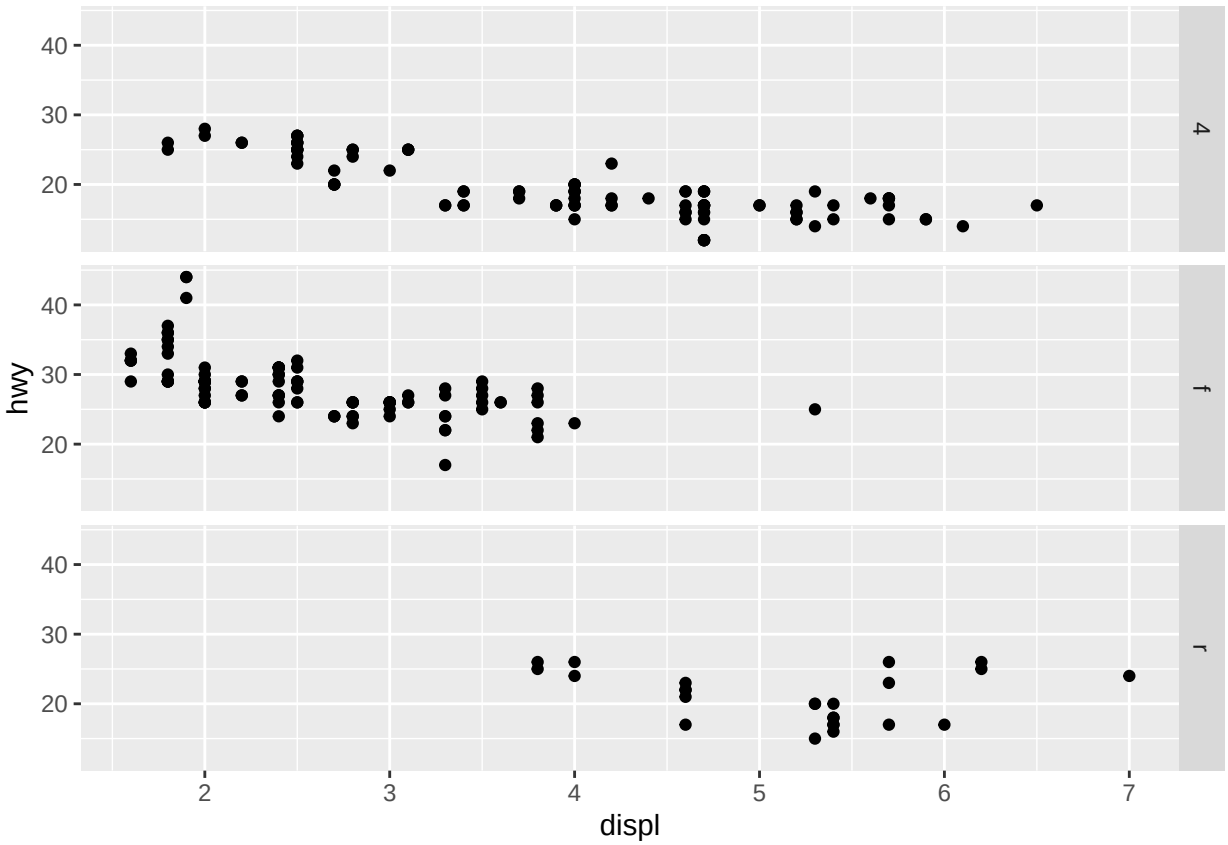


Faceting - not by rows or columns

```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy)) +
  facet_grid(. ~ cyl) # use the . ~ variable to not facet by either rows or columns
```

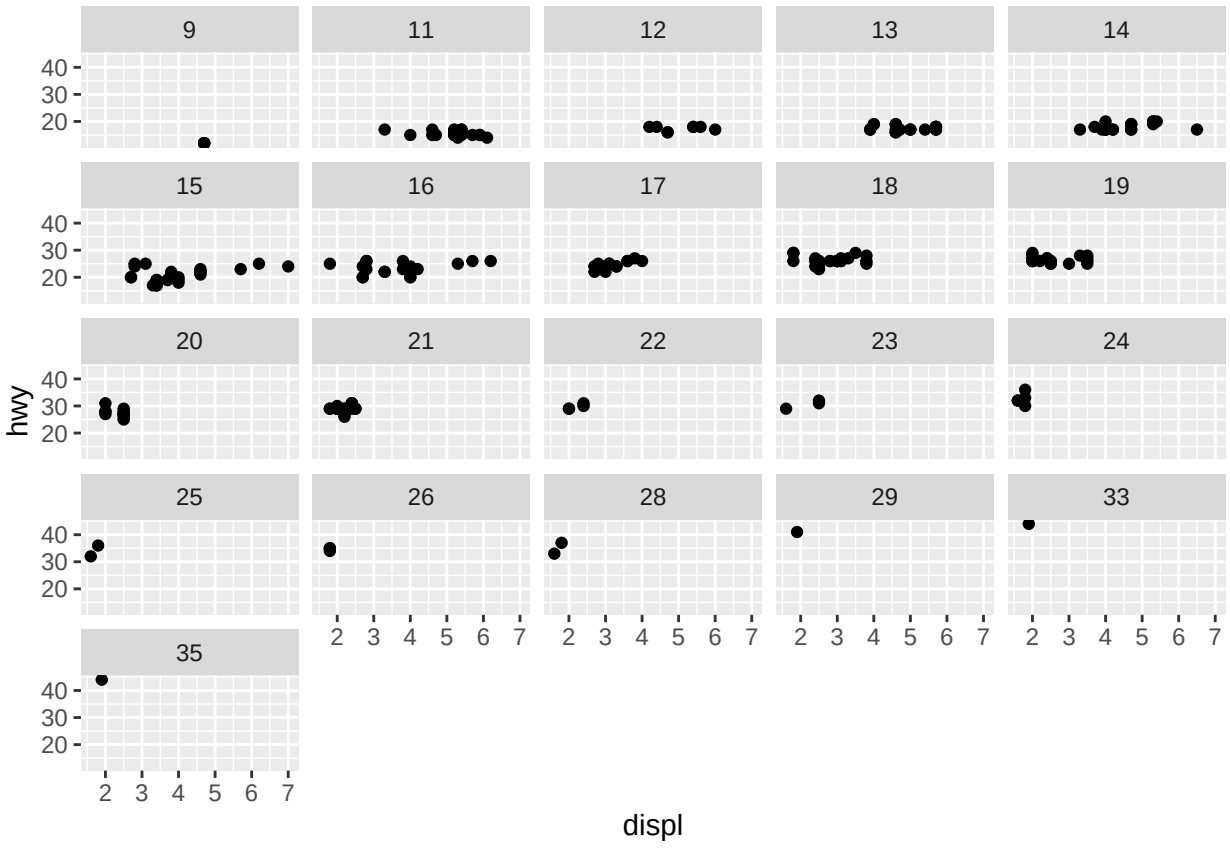



```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy)) +
  facet_grid(drv ~ .)
```



Facet a continuous variable

```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy)) +
  facet_wrap(~cty)
```


9
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
28
29
33
35
hwy
displ
40
30
20
2
3
4
5
6
7