R for Data Science PSTAT 131

JEFFREY CHAN

12/13/2020

please use this command to install the library `install.packages("tidyverse")`

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.3     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.0
```

```
## -- Conflicts ------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

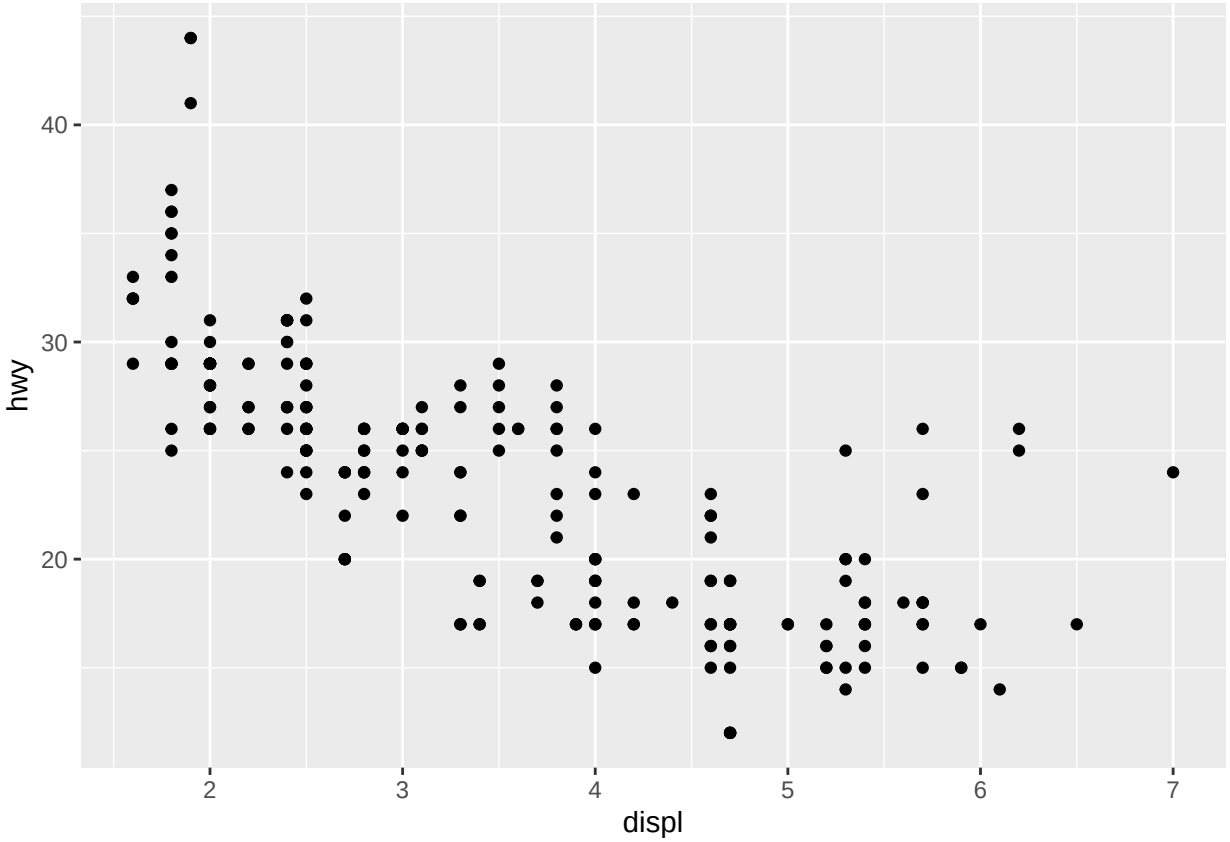
^ if you can see, there is some conflict from tidyverse library and dplyr library. So when we are using those functions, either from stats or dplyr, we need to specify the library name.

build in data frame / data called mpg Lets find out if do big engines use more fuel

```
str(mpg)
```

```
## tibble [234 x 11] (S3: tbl_df/tbl/data.frame)
##  $ manufacturer: chr [1:234] "audi" "audi" "audi" "audi" ...
##  $ model       : chr [1:234] "a4" "a4" "a4" "a4" ...
##  $ displ       : num [1:234] 1.8 1.8 2 2 2.8 2.8 3.1 1.8 1.8 2 ...
##  $ year        : int [1:234] 1999 1999 2008 2008 1999 1999 2008 1999 1999 2008 ...
##  $ cyl         : int [1:234] 4 4 4 4 6 6 6 4 4 4 ...
##  $ trans       : chr [1:234] "auto(l5)" "manual(m5)" "manual(m6)" "auto(av)" ...
##  $ drv         : chr [1:234] "f" "f" "f" "f" ...
##  $ cty         : int [1:234] 18 21 20 21 16 18 18 18 16 20 ...
##  $ hwy         : int [1:234] 29 29 31 30 26 26 27 26 25 28 ...
##  $ fl          : chr [1:234] "p" "p" "p" "p" ...
##  $ class       : chr [1:234] "compact" "compact" "compact" "compact" ...
```

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



In ggplot2, ggplot() creates the coordinate system, and then i can add layers on top of the coordinate. in deed, geom_point is my second layer. Also, ggplot will take the data set as parameter / input. geom_point creates the scatter dots. this goes in pair in side the geom_point(mappint = aes(x = , y =)) this will tell the graph what data on the x and y axies

Template `ggplot(data = <DATA>) +  <GEOM_FUNCTION>(mapping = aes(<MAPPINGS>))`

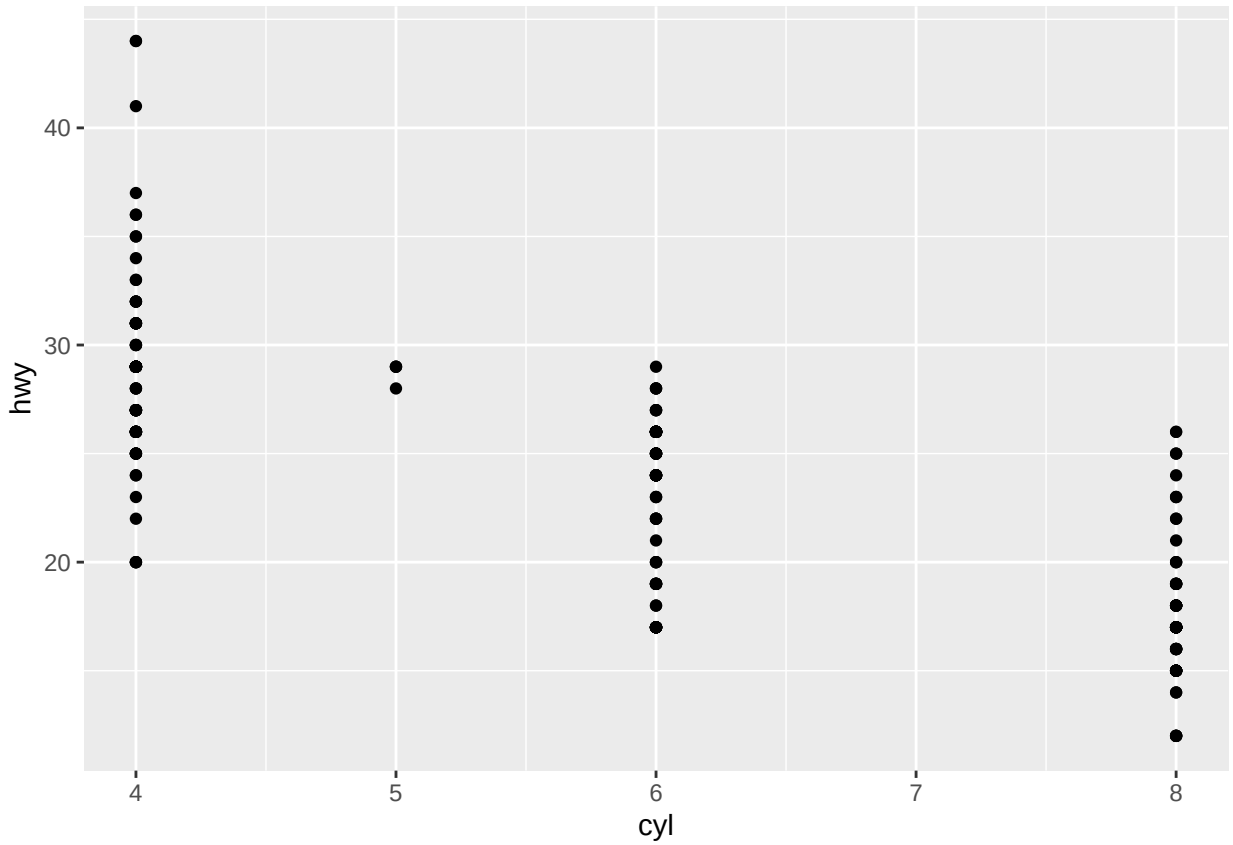
exercises

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

```
dim(mtcars)
```

```
## [1] 32 11
```

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



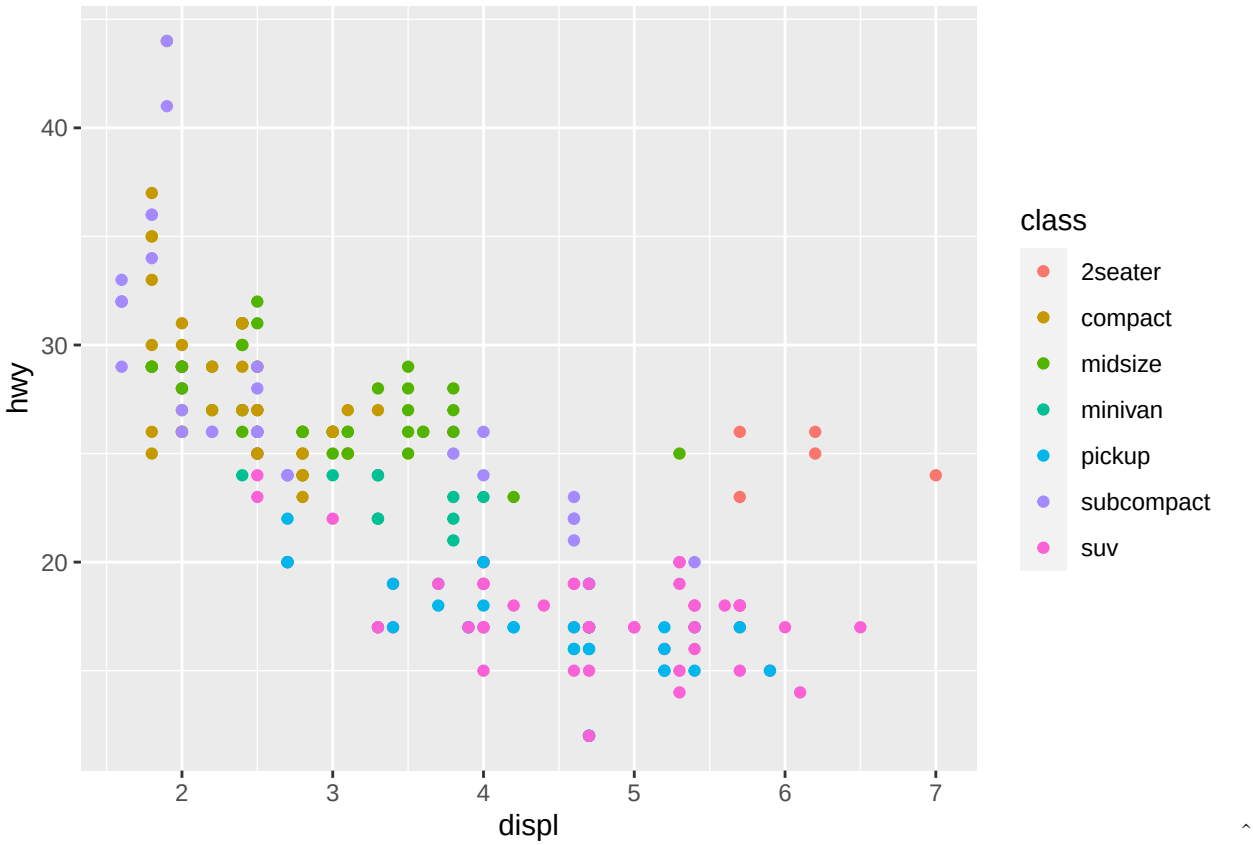
let jump back to mpg plot. we know that we plot hwy ~ displ or y = hwy, x = displ, but what if we want to know more insight like what type of vehicles are those? we can identify it with color Lets check what variables do they have in mpg

```
str(mpg)
```

```
## tibble [234 x 11] (S3: tbl_df/tbl/data.frame)
##  $ manufacturer: chr [1:234] "audi" "audi" "audi" "audi" ...
##  $ model       : chr [1:234] "a4" "a4" "a4" "a4" ...
##  $ displ       : num [1:234] 1.8 1.8 2 2 2.8 2.8 3.1 1.8 1.8 2 ...
##  $ year        : int [1:234] 1999 1999 2008 2008 1999 1999 2008 1999 1999 2008 ...
##  $ cyl         : int [1:234] 4 4 4 4 6 6 6 4 4 4 ...
##  $ trans       : chr [1:234] "auto(l5)" "manual(m5)" "manual(m6)" "auto(av)" ...
##  $ drv         : chr [1:234] "f" "f" "f" "f" ...
##  $ cty         : int [1:234] 18 21 20 21 16 18 18 18 16 20 ...
##  $ hwy         : int [1:234] 29 29 31 30 26 26 27 26 25 28 ...
##  $ fl          : chr [1:234] "p" "p" "p" "p" ...
##  $ class       : chr [1:234] "compact" "compact" "compact" "compact" ...
```

ˆ as we see we have class that illustrate what type of vehicle is it so we will use that info to plot a better graph

```
ggplot(mpg) +
  geom_point(aes(x = displ, y = hwy, color = class))
```

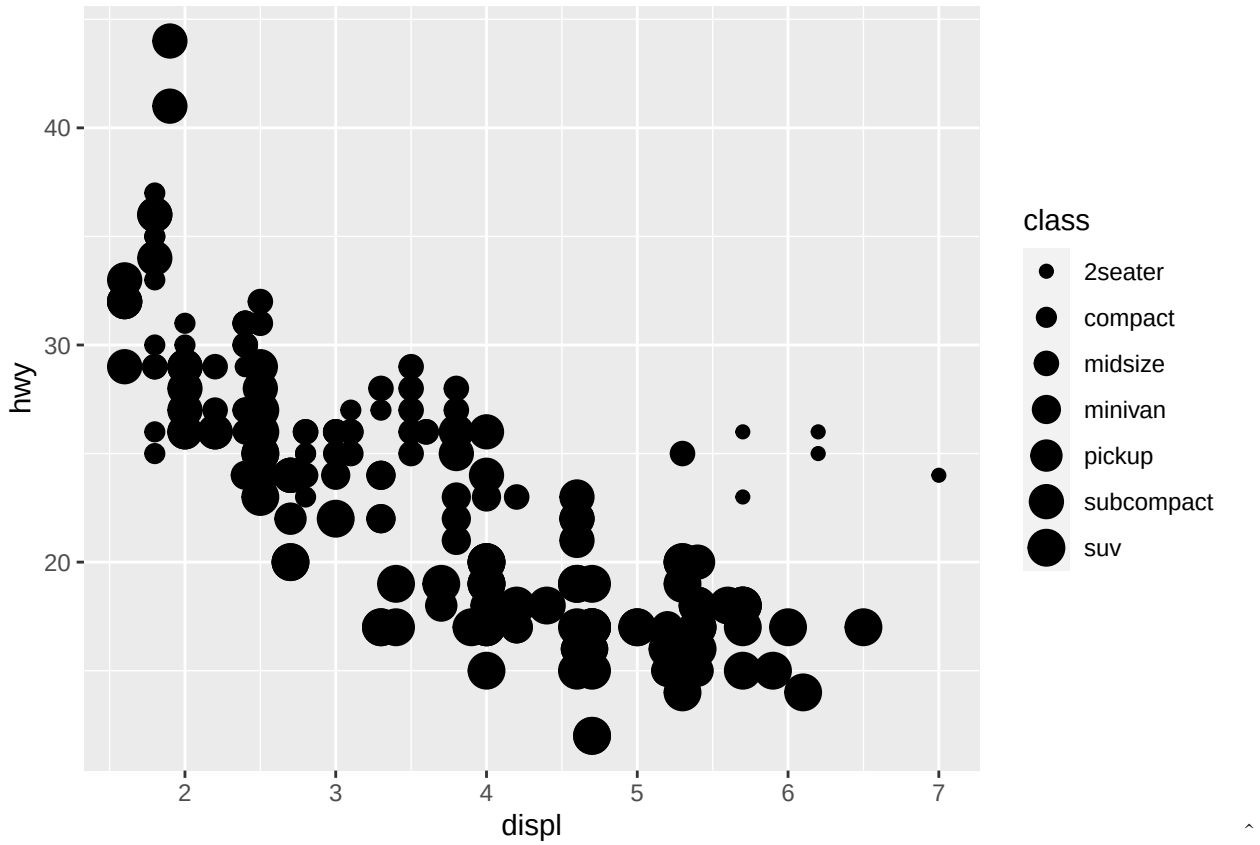


now we set different color by it's own type of vehicle type it is more information from and we can more a better conclusion. color function comes with automatically labelling the vehicle types at the side.

now we can plot the graph with the following

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

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

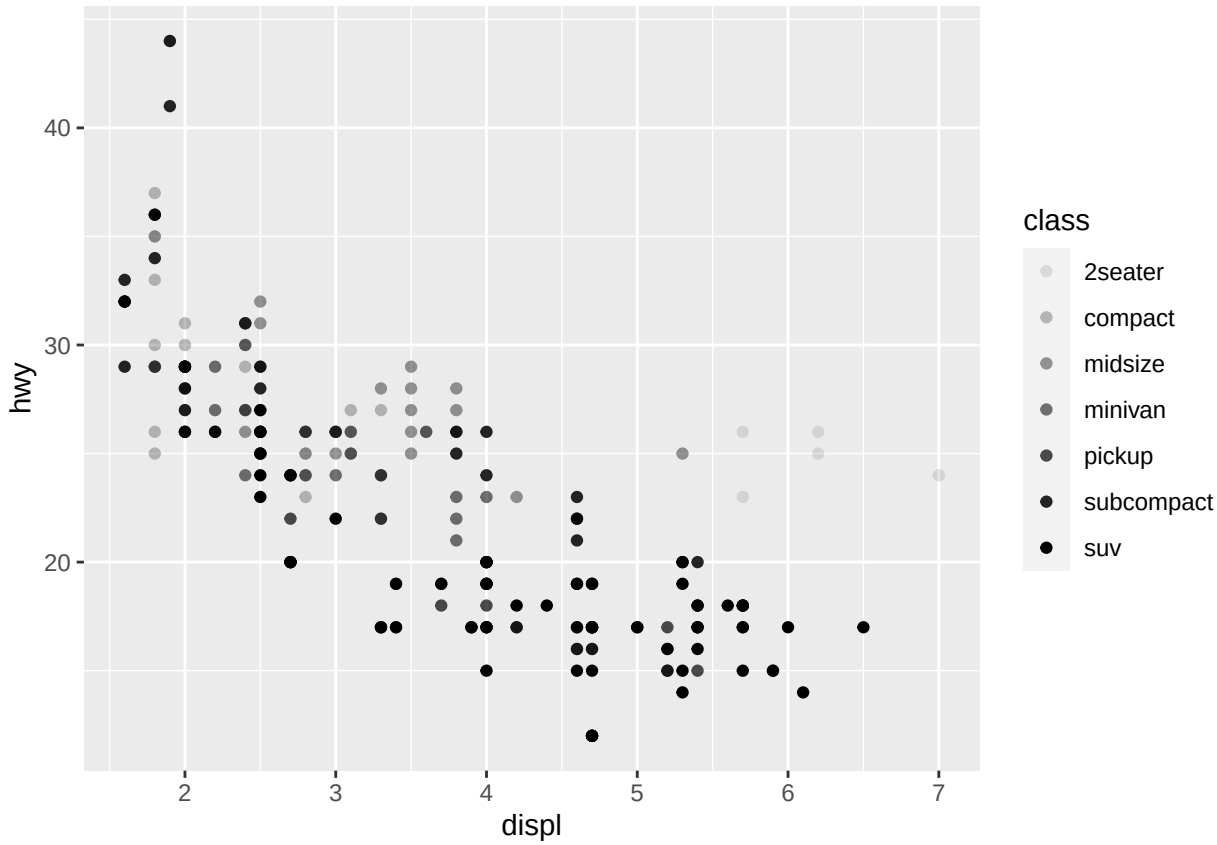


as you can see, we can a warning "using size for a discrete varaible is not advised." because mapping an unordered variable (class) to an ordered aesthetic (size) is not a good idea.

Alternative

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

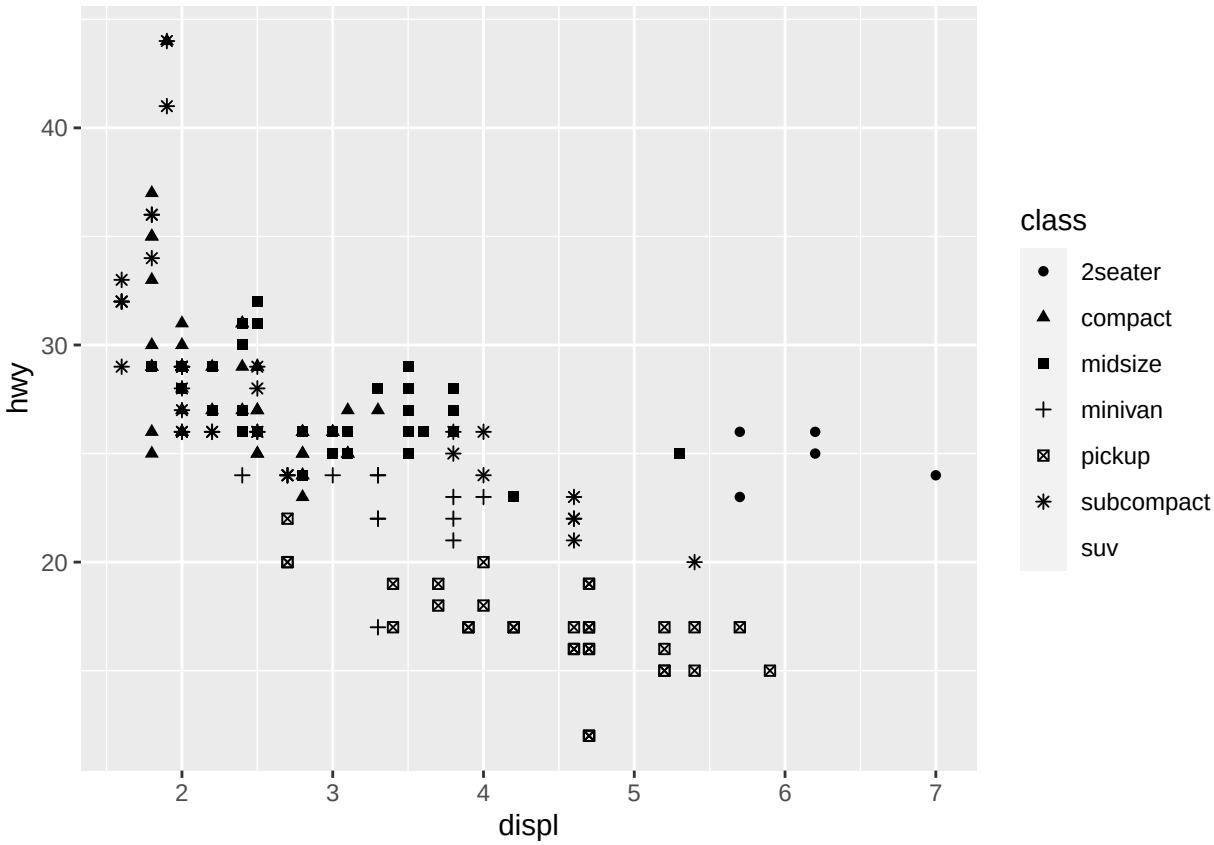
```
## Warning: Using alpha for a discrete variable is not advised.
```




```
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.
```

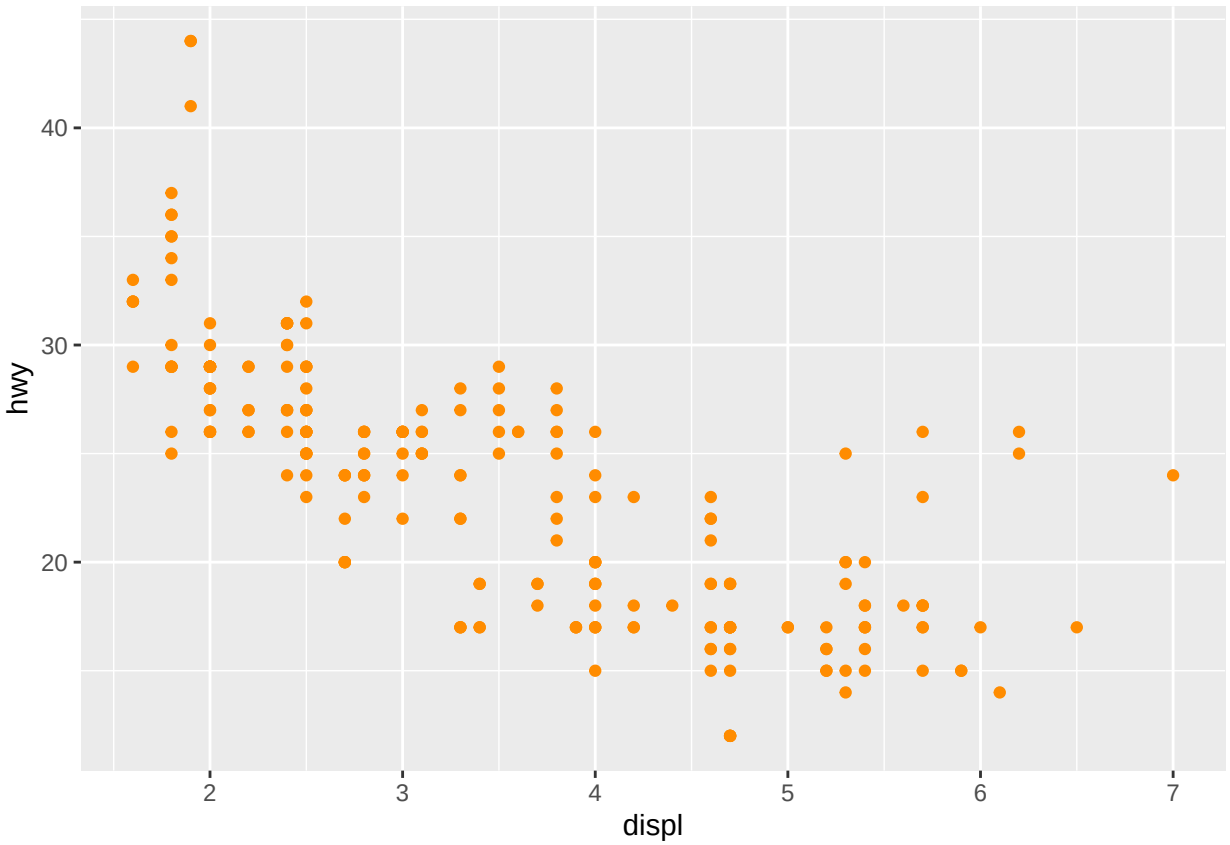
```
## Warning: Removed 62 rows containing missing values (geom_point).
```



By now you should see the warning if you use rmd to read this file. shape = class this function right here, in ggplot2, it only reserves 6 different symbols only.

we can also set the color manually like the following

```
ggplot(mpg) +
  geom_point(aes(x = displ, y = hwy), color = "darkorange")
```



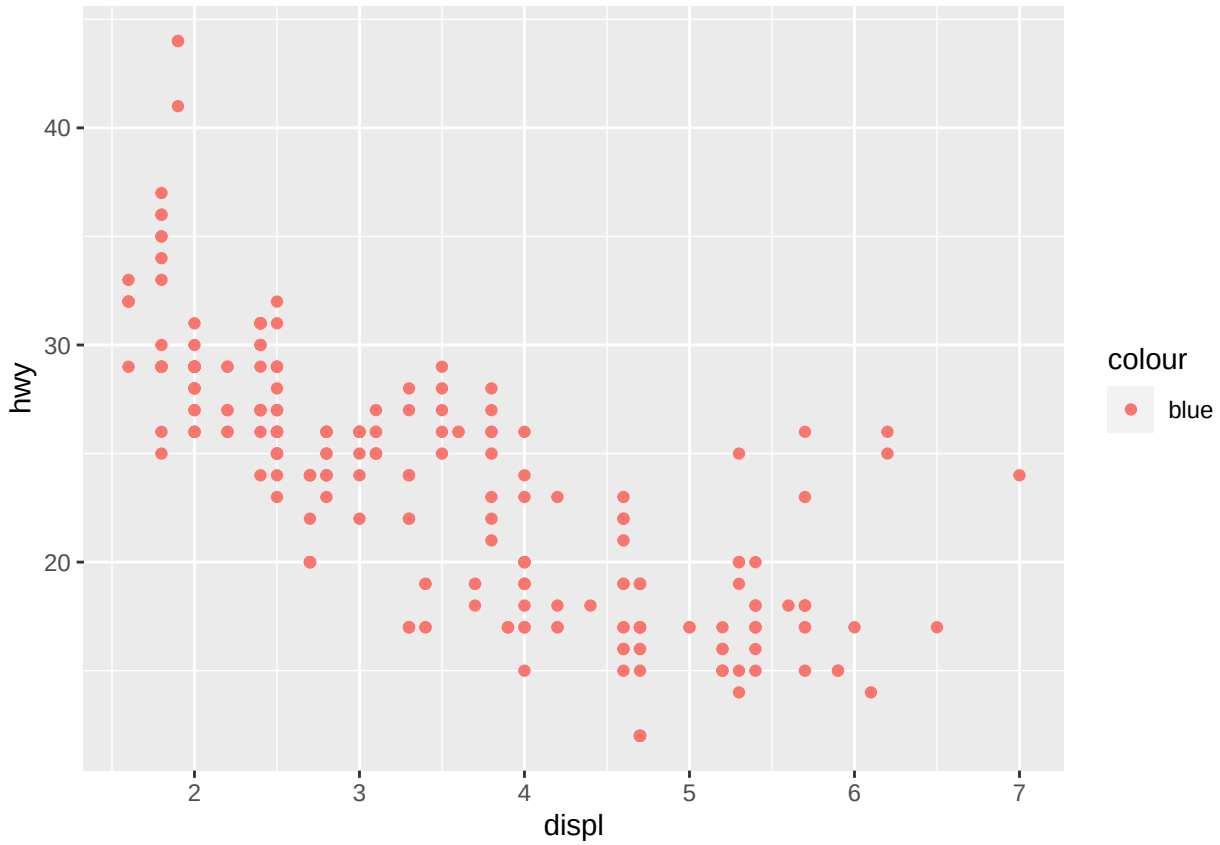
I hope you are paying attention to the code itself. As you can see, if we want to define the color, the argument is outside of aes. `ggplot(mpg) + geom_point(aes(x = displ, y = hwy), color = "darkorange")`

Compare with the previous one `ggplot(data = mpg) + geom_point(mapping = aes(x = displ, y = hwy, shape = class))` `ggplot(data = mpg) + geom_point(mapping = aes(x = displ, y = hwy, alpha = class))` `ggplot(mpg) + geom_point(aes(x = displ, y = hwy, size = class))` `ggplot(mpg) + geom_point(aes(x = displ, y = hwy, color = class))`

All the color function is INSIDE the aes function. Please take note for this step If you refere to the table / the book. 0-14 hollow shapes can be defined with color 15-19 solid shapes are filled with color 21-24 filled shapes and it can be defined a new boarder color

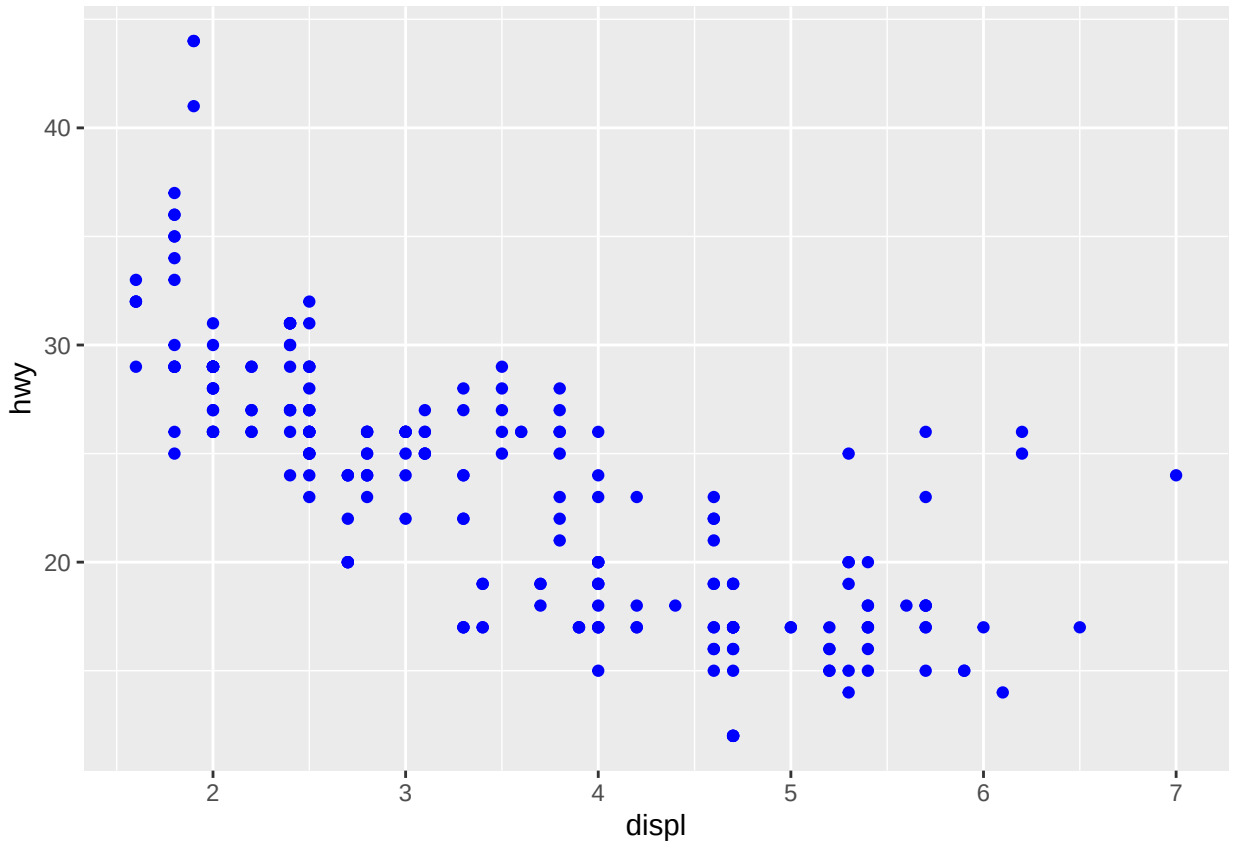
Q1 Why it is not blue?

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



because color = "blue" is inside the aes, move it outside of aes will work

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



Q2 Map a continuous variable to color, size, and shape. How do these aesthetics behave differently for categorical versus contin- uous variables?

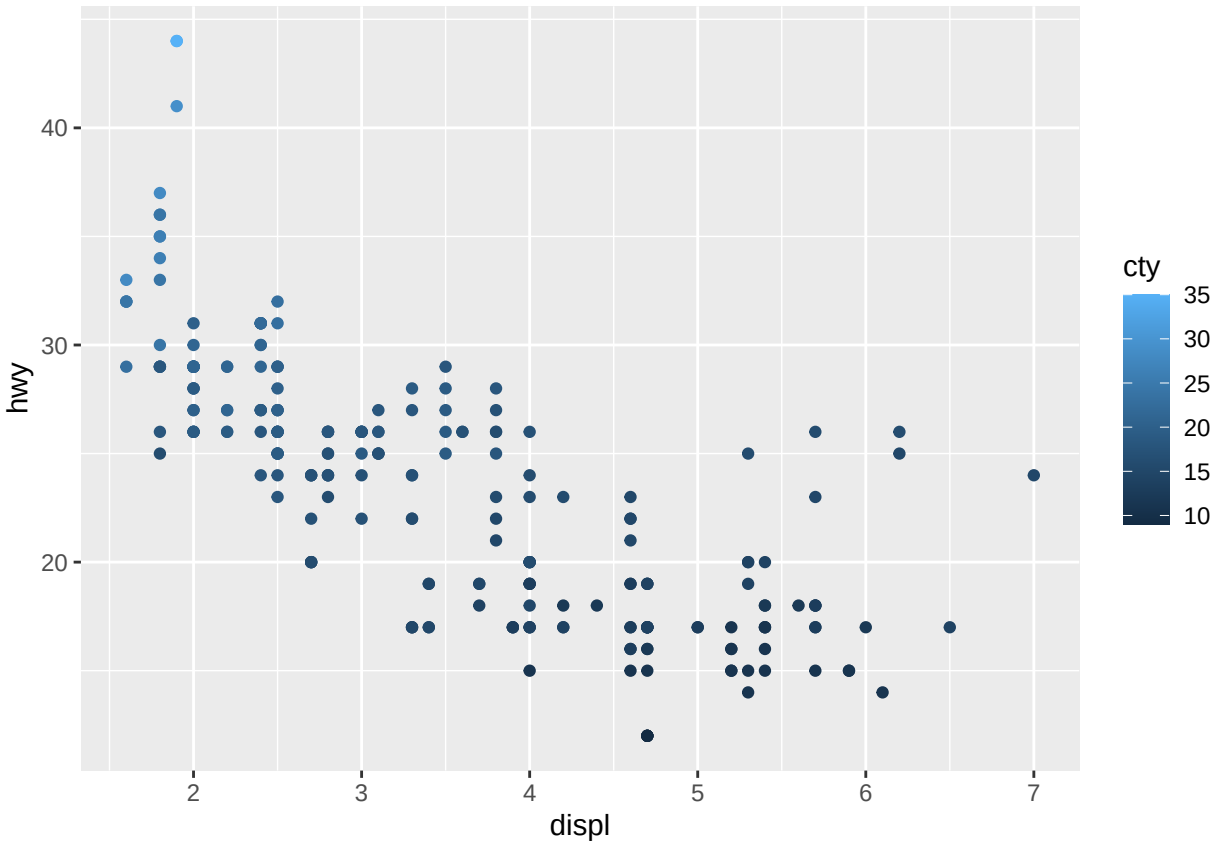
```
mpg
```

```
## # A tibble: 234 x 11
##    manufacturer model    displ  year   cyl trans   drv     cty   hwy fl    class
##    <chr>        <chr>    <dbl> <int> <int> <chr>   <chr> <int> <int> <chr> <chr>
##  1 audi         a4         1.8  1999     4 auto(l~ f        18    29 p     comp~
##  2 audi         a4         1.8  1999     4 manual~ f        21    29 p     comp~
##  3 audi         a4         2    2008     4 manual~ f        20    31 p     comp~
##  4 audi         a4         2    2008     4 auto(a~ f        21    30 p     comp~
##  5 audi         a4         2.8  1999     6 auto(l~ f        16    26 p     comp~
##  6 audi         a4         2.8  1999     6 manual~ f        18    26 p     comp~
##  7 audi         a4         3.1  2008     6 auto(a~ f        18    27 p     comp~
##  8 audi         a4 quat~   1.8  1999     4 manual~ 4        18    26 p     comp~
##  9 audi         a4 quat~   1.8  1999     4 auto(l~ 4        16    25 p     comp~
## 10 audi         a4 quat~   2    2008     4 manual~ 4        20    28 p     comp~
## # ... with 224 more rows
```

ˆ when you type mpg, you will have this table. quick important note, when you see under the variable name, then it is categorical variable.

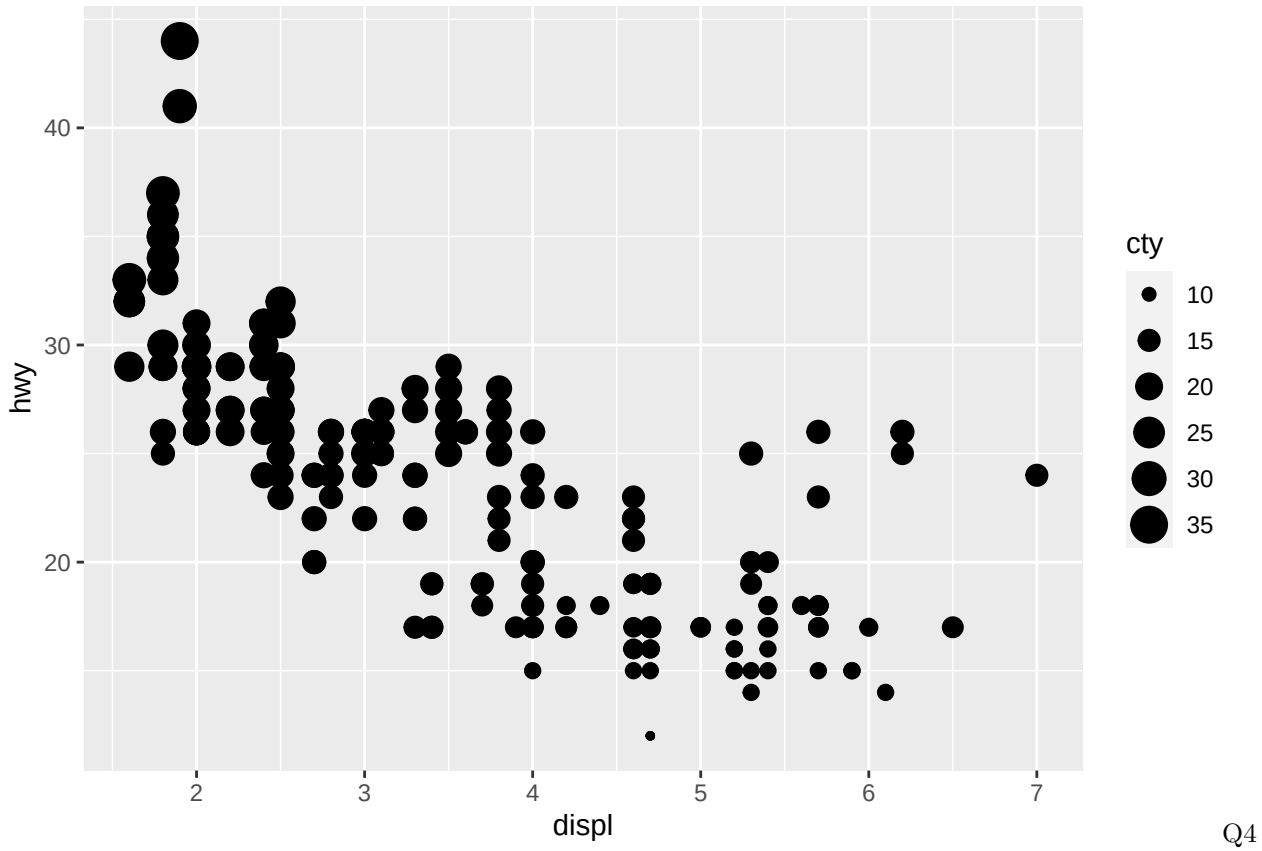
Q3 Map a continuous variable to color, size, and shape. How do these aesthetics behave differently for categorical versus contin- uous variables?

```
ggplot(mpg) +
  geom_point(aes(x=displ, y=hwy, color=cty))
```



instead of using discrete colors, continuous variable uses a scale that varies from light to dark.

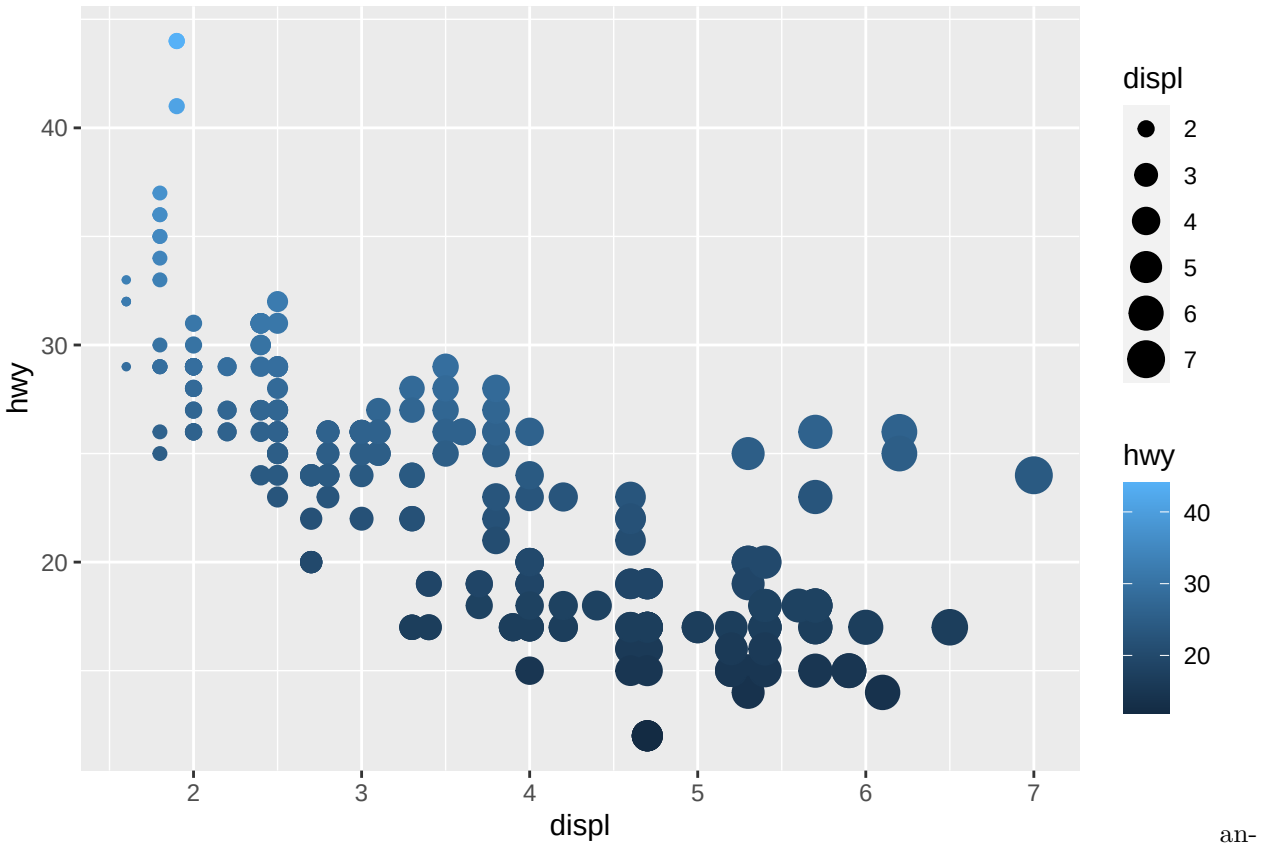
```
ggplot(mpg, aes(x=displ, y = hwy, size = cty)) +
  geom_point()
```



Q4

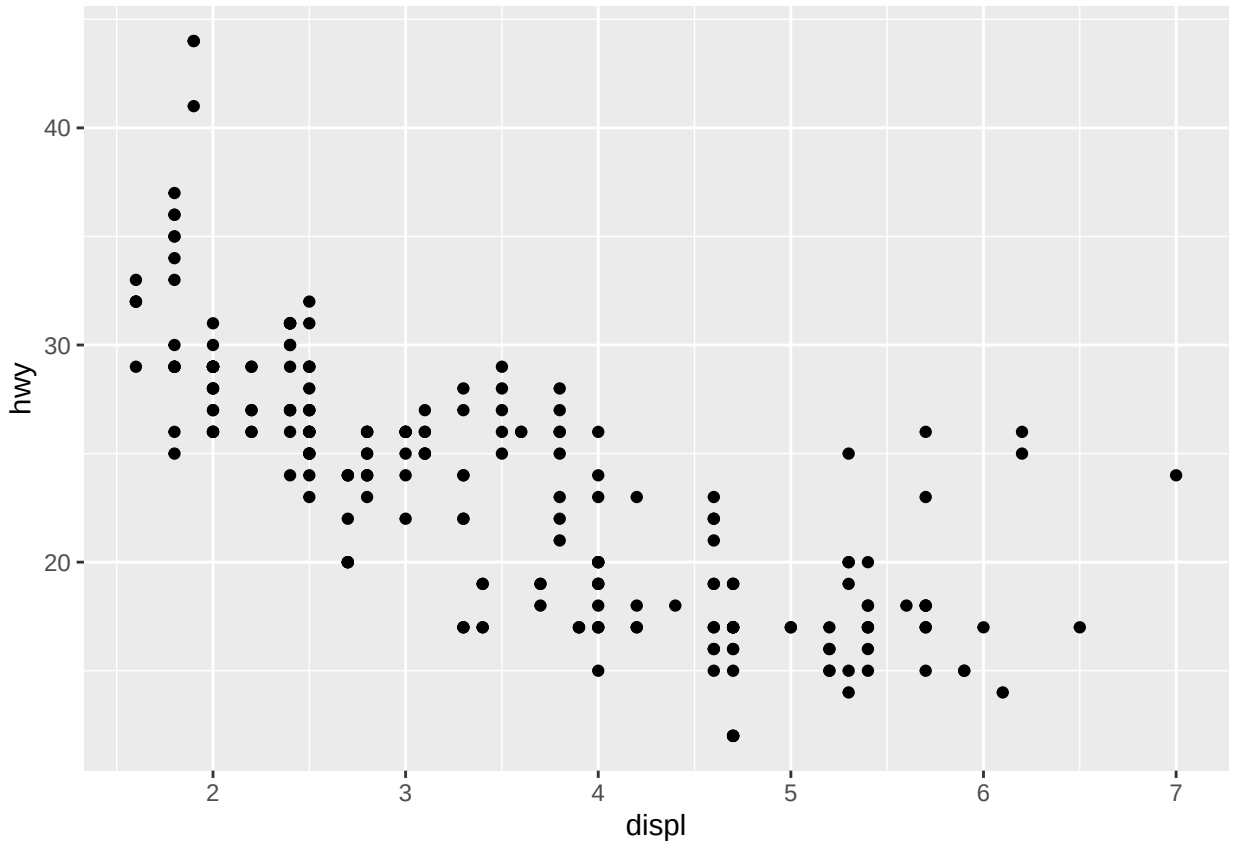
What happens if you map the same variable to multiple aesthetics? it just looks bad, because it is redundant

```
ggplot(mpg, aes(x = displ, y = hwy, colour = hwy, size = displ)) +
  geom_point()
```



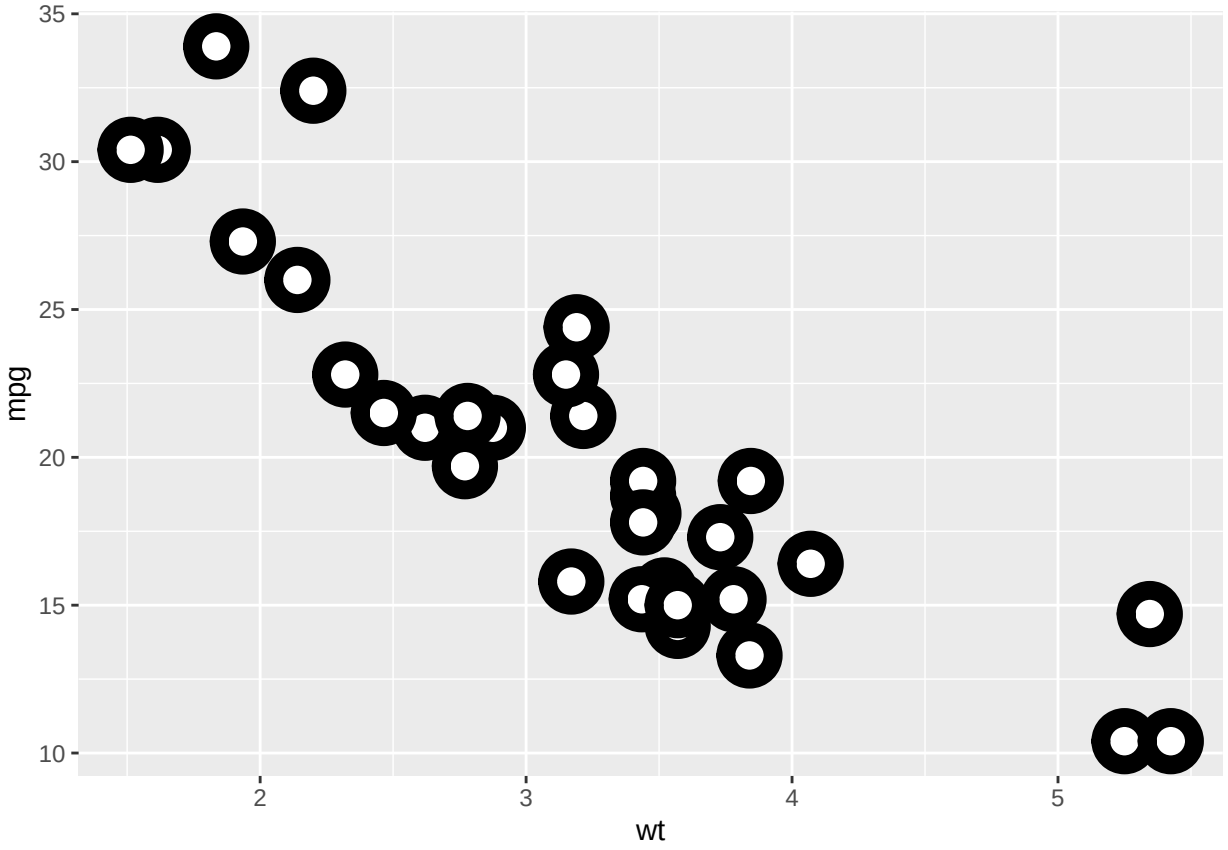
another way of plotting the ggplot

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

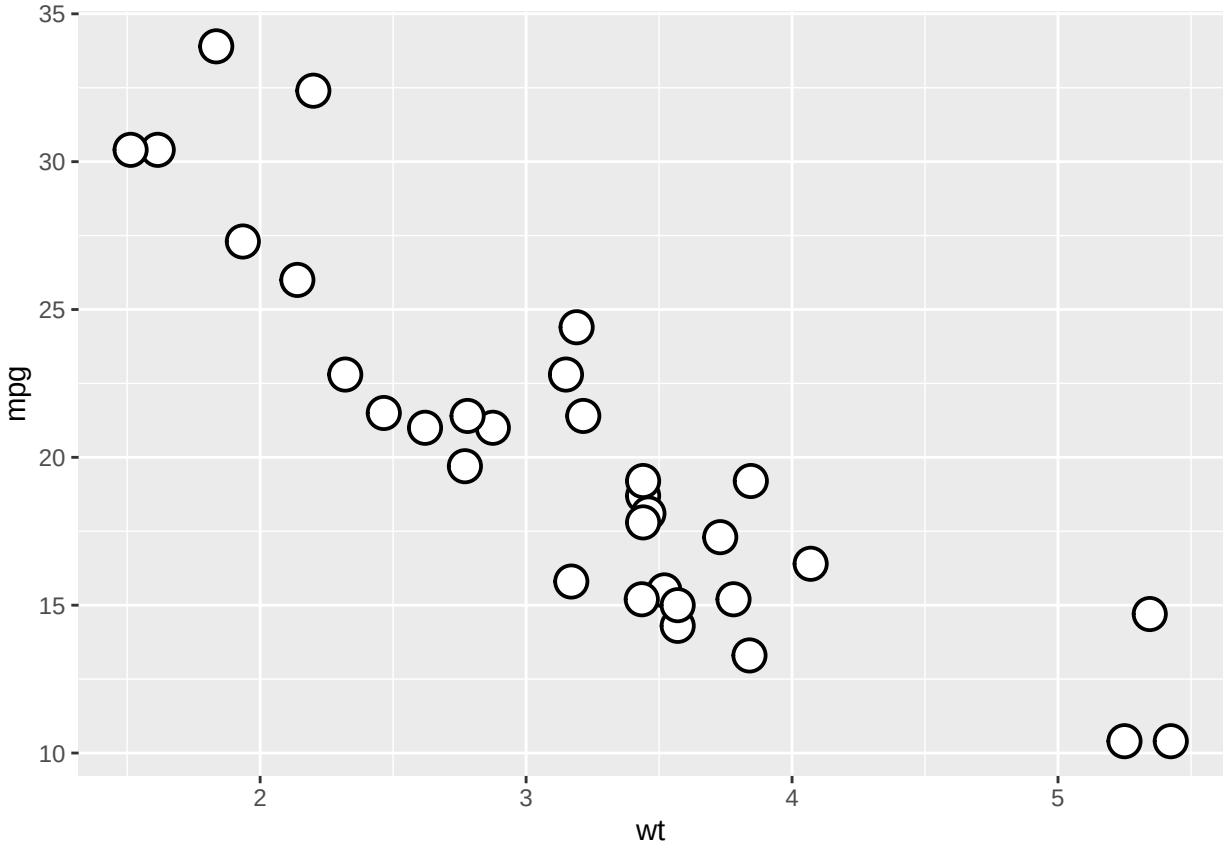


Q5 What does the stroke aesthetic do? What shapes does it work with? (Hint: use ?geom_point.)

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



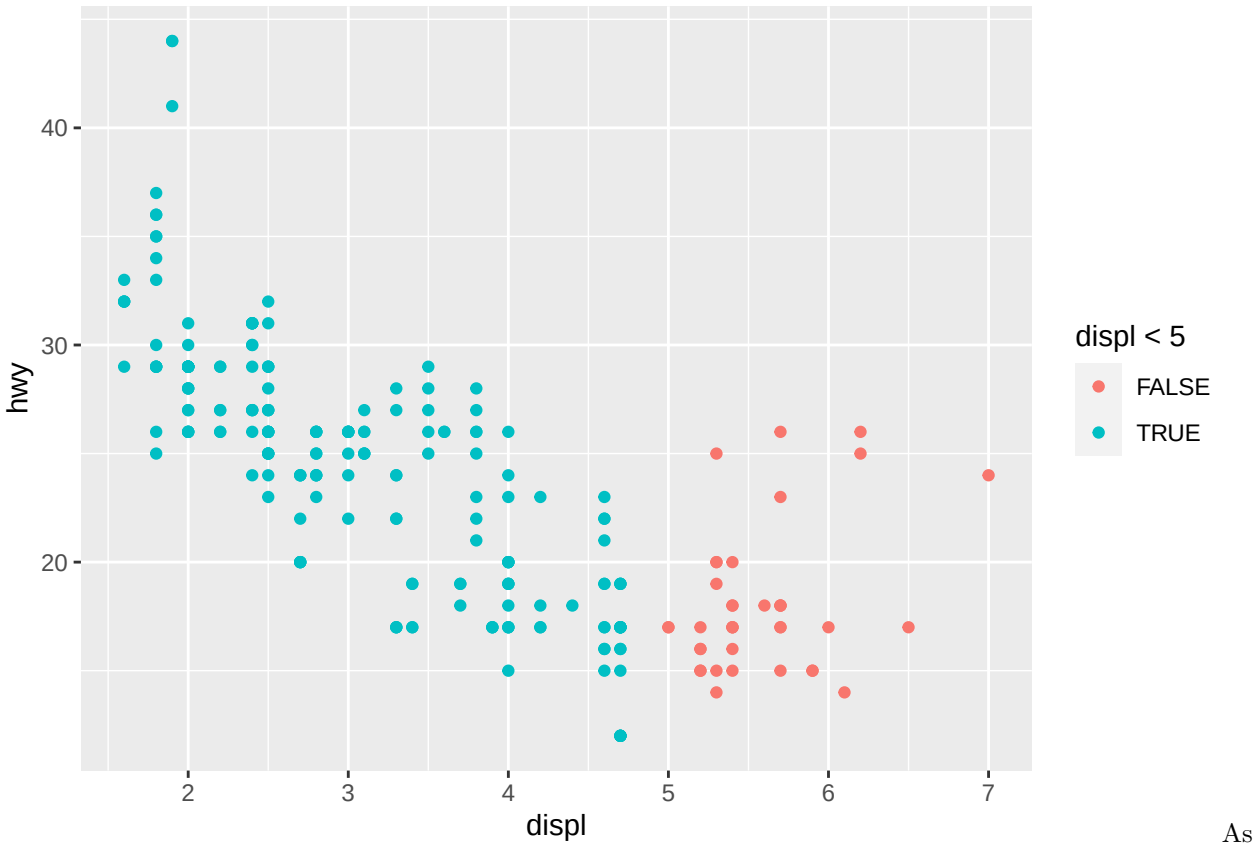

```
ggplot(mtcars, aes(wt,mpg)) +
  geom_point(shape = 21, color = "black", fill = "white", size = 5,
             stroke = 1)
```



As you can see stroke control the border size and stroke will only work for the shapes (21-25)

Q6 What happens if you map an aesthetic to something other than a variable name, like aes(color = displ < 5)?

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



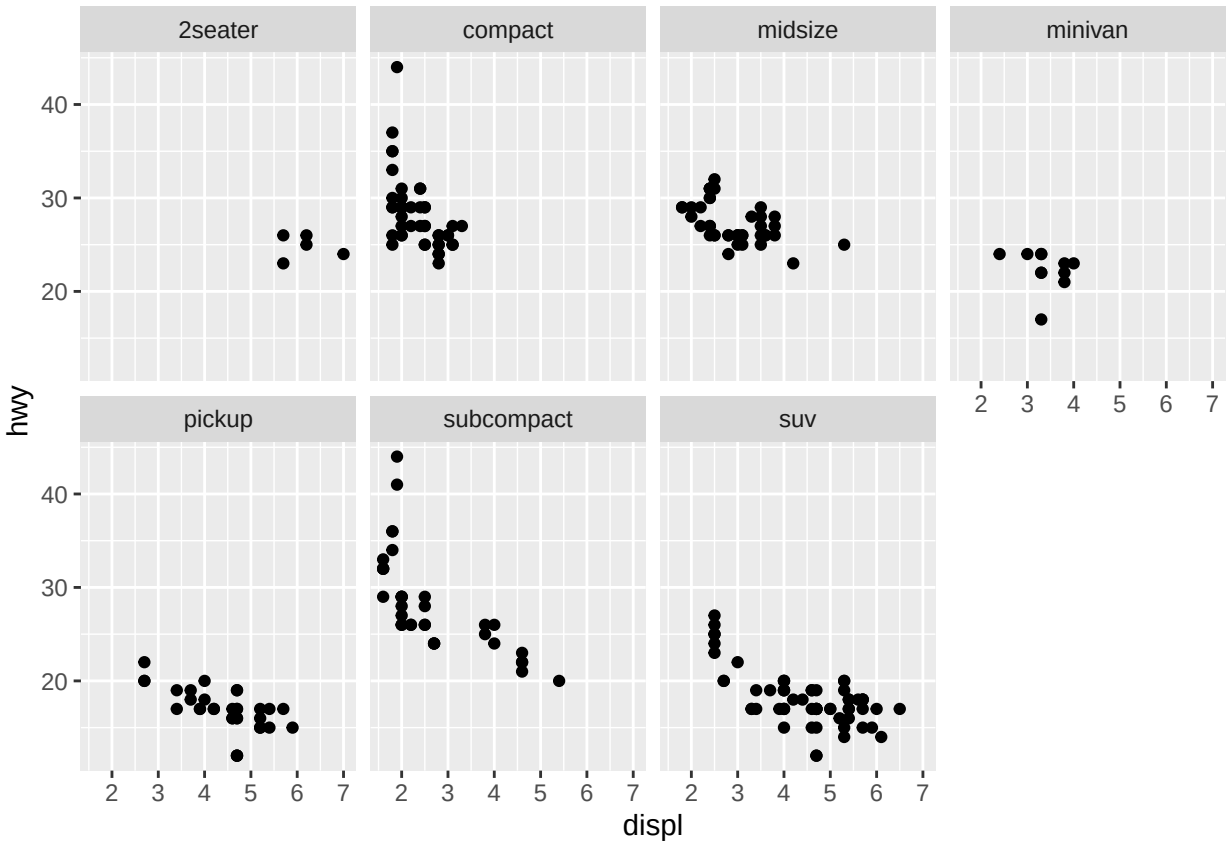
As you can see, if the displacement is >= 5 it will switch color from green to red.

```
mpg
```

```
## # A tibble: 234 x 11
##    manufacturer model    displ  year   cyl trans   drv     cty   hwy fl    class
##    <chr>        <chr>    <dbl> <int> <int> <chr>   <chr> <int> <int> <chr> <chr>
##  1 audi         a4         1.8  1999     4 auto(l~ f        18    29 p     comp~
##  2 audi         a4         1.8  1999     4 manual~ f        21    29 p     comp~
##  3 audi         a4         2    2008     4 manual~ f        20    31 p     comp~
##  4 audi         a4         2    2008     4 auto(a~ f        21    30 p     comp~
##  5 audi         a4         2.8  1999     6 auto(l~ f        16    26 p     comp~
##  6 audi         a4         2.8  1999     6 manual~ f        18    26 p     comp~
##  7 audi         a4         3.1  2008     6 auto(a~ f        18    27 p     comp~
##  8 audi         a4 quat~   1.8  1999     4 manual~ 4        18    26 p     comp~
##  9 audi         a4 quat~   1.8  1999     4 auto(l~ 4        16    25 p     comp~
## 10 audi         a4 quat~   2    2008     4 manual~ 4        20    28 p     comp~
## # ... with 224 more rows
```

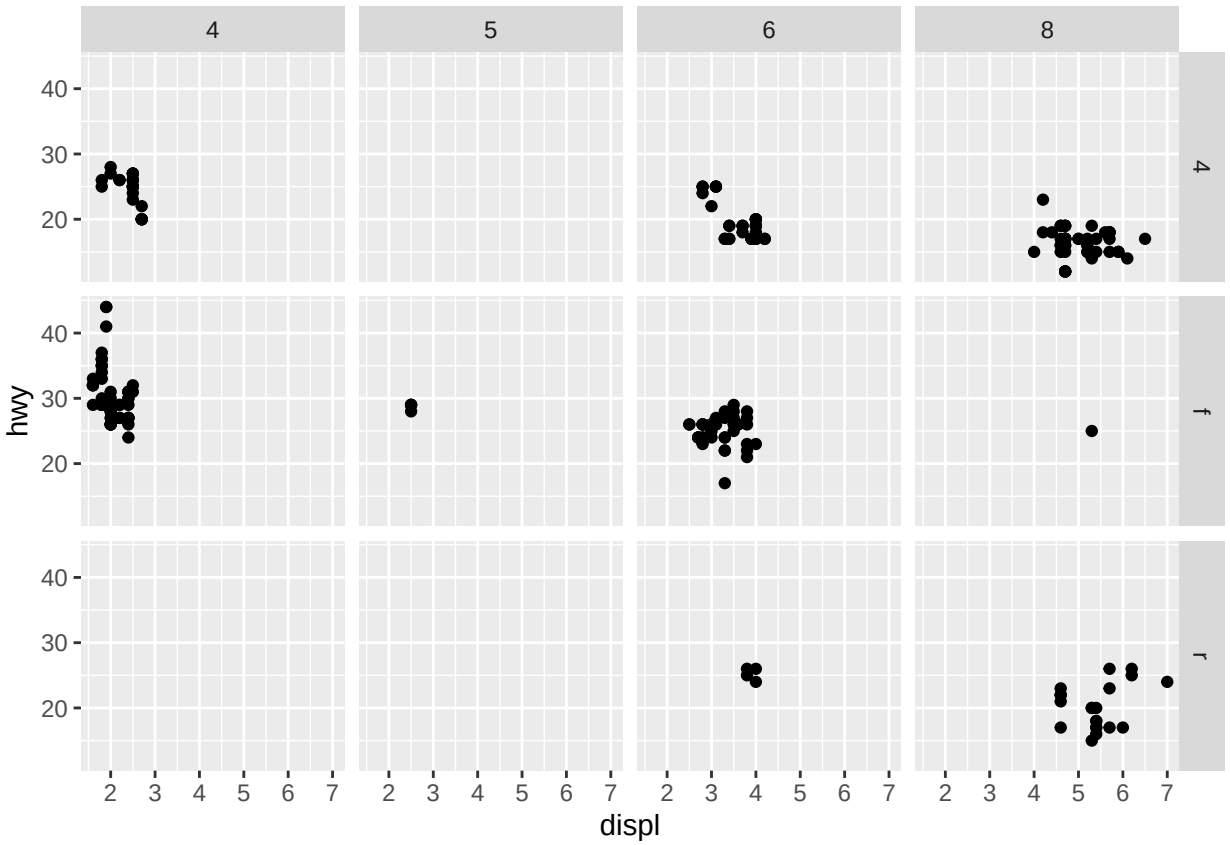
Facets this will plot the graph by it's own kind code: facet_wrap(y ~ x variable_name, nrow = #?) i can do it by itself (~ class)

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



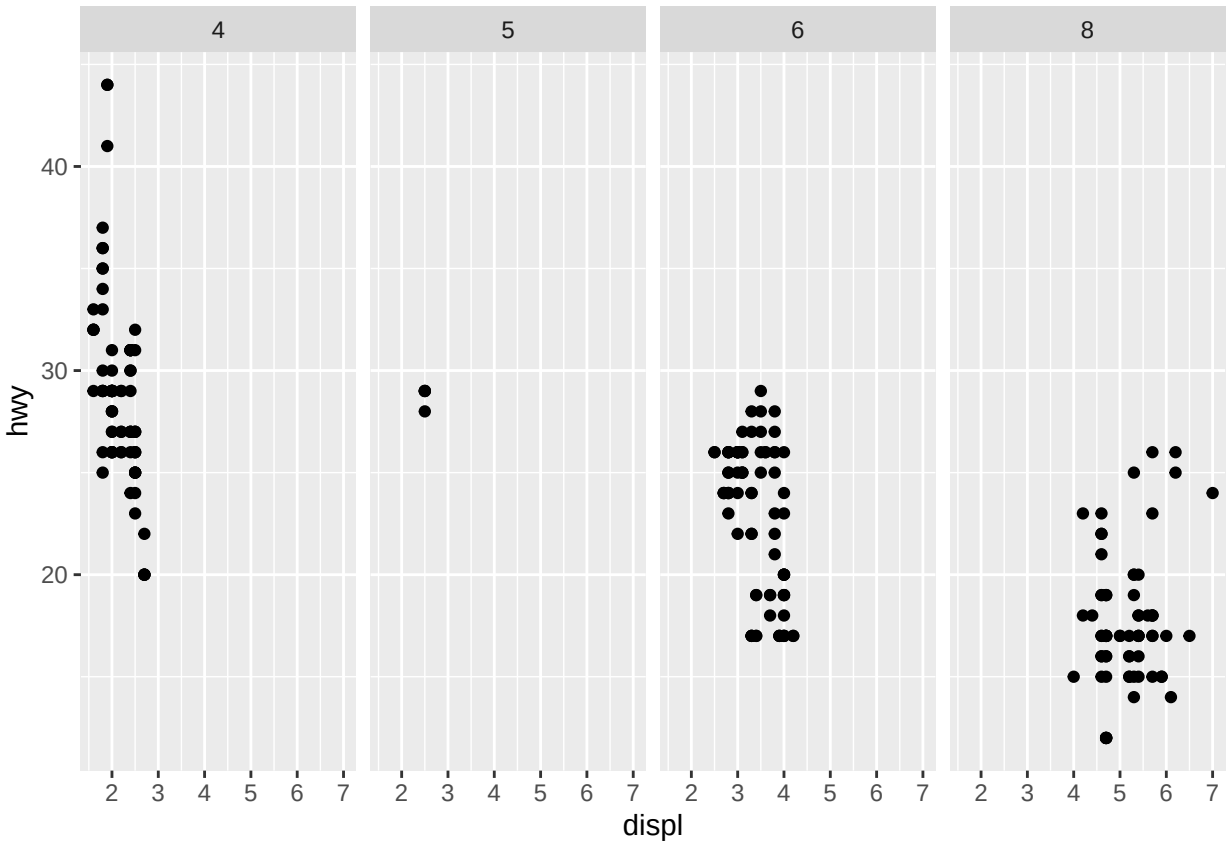
let's plot drv ~ cyl

```
ggplot(mpg) +
  geom_point(aes(displ,hwy)) +
  facet_grid(drv ~ cyl)
```



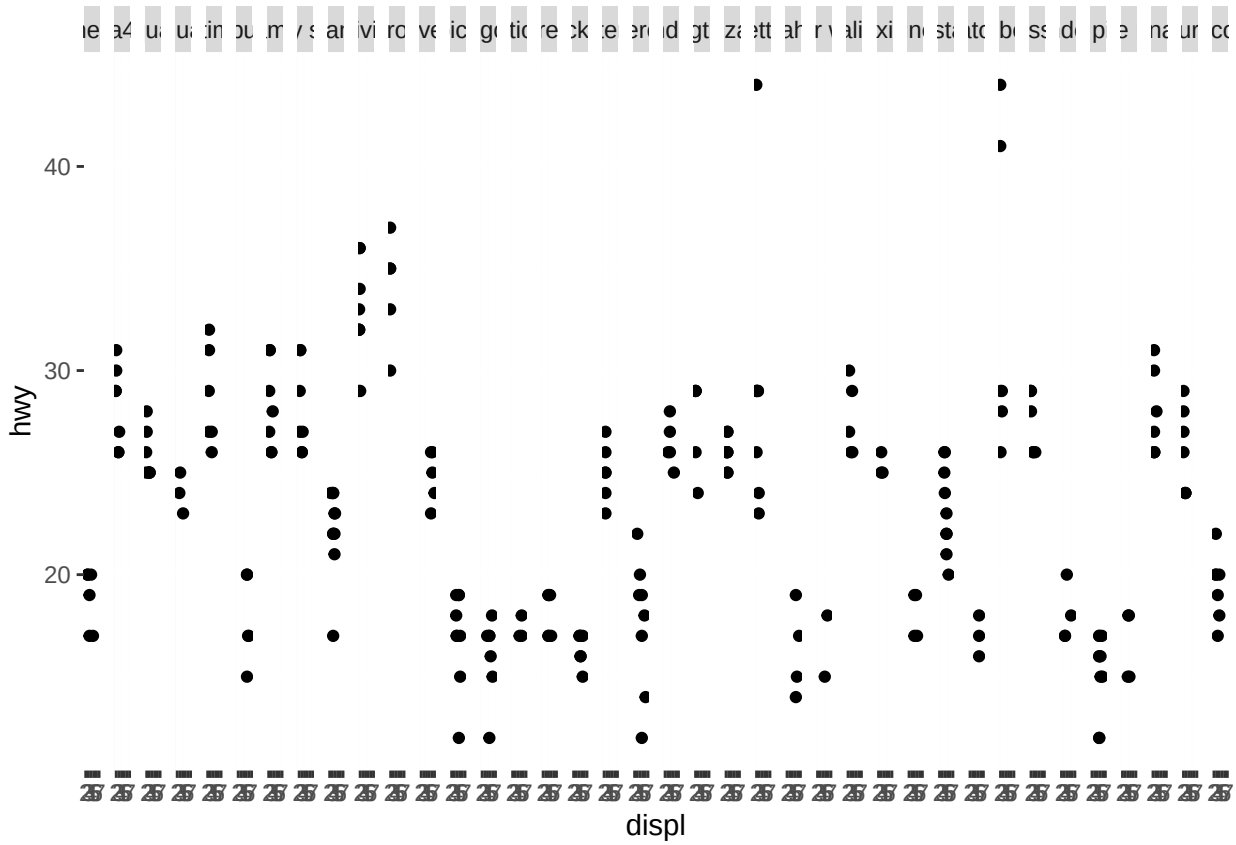
grid vs wrap is i have 2 more variables at the side. 4 5 6 8 on top is the cylinder, and 4 f r is the drive train. try to figure out how to read the variables from the data set with help function.

```
ggplot(mpg) +
  geom_point(aes(displ,hwy)) +
  facet_grid(. ~ cyl)
```



Q1 What happens if you facet on a continuous variable?

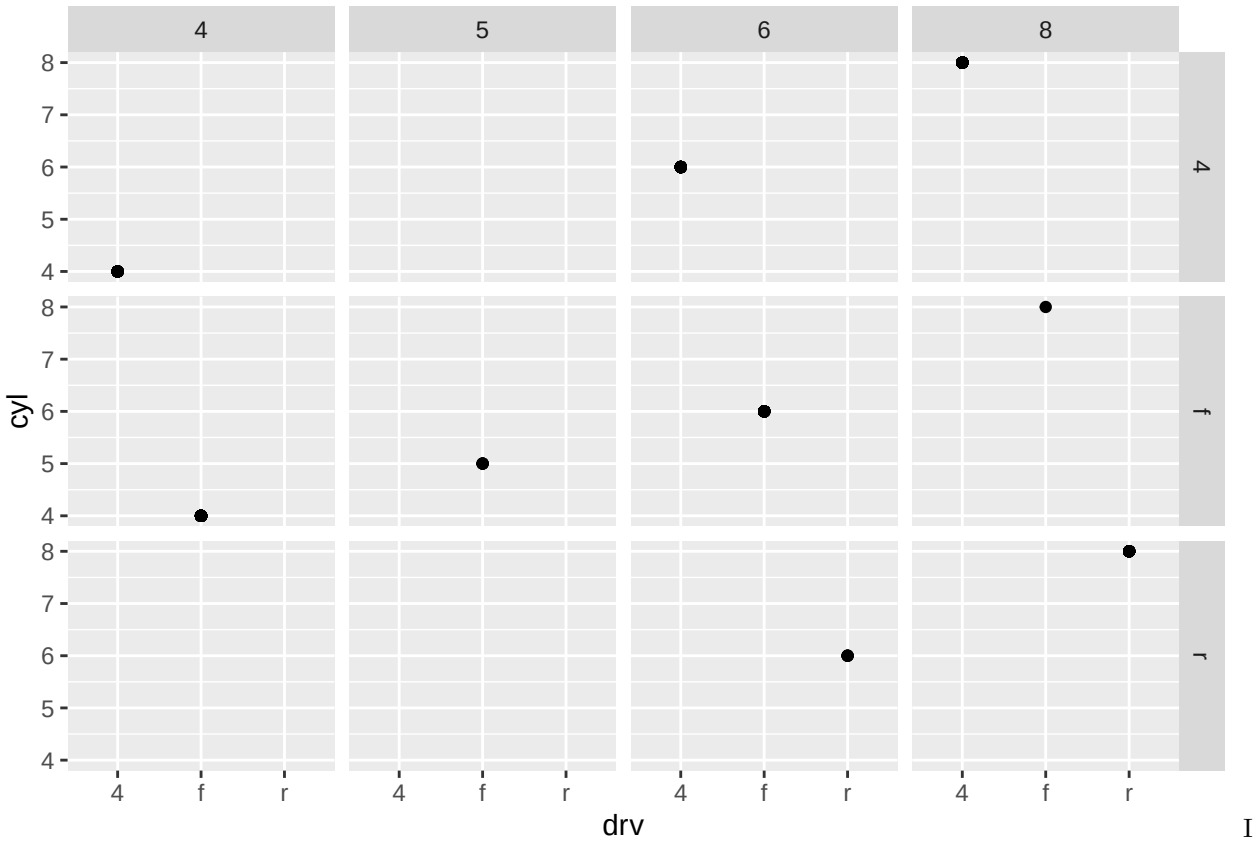
```
ggplot(mpg) +
  geom_point(aes(displ,hwy)) +
  facet_grid(. ~ model)
```



As you can see the graph above, if i facet a continuous variable, the plot will become unreadable.

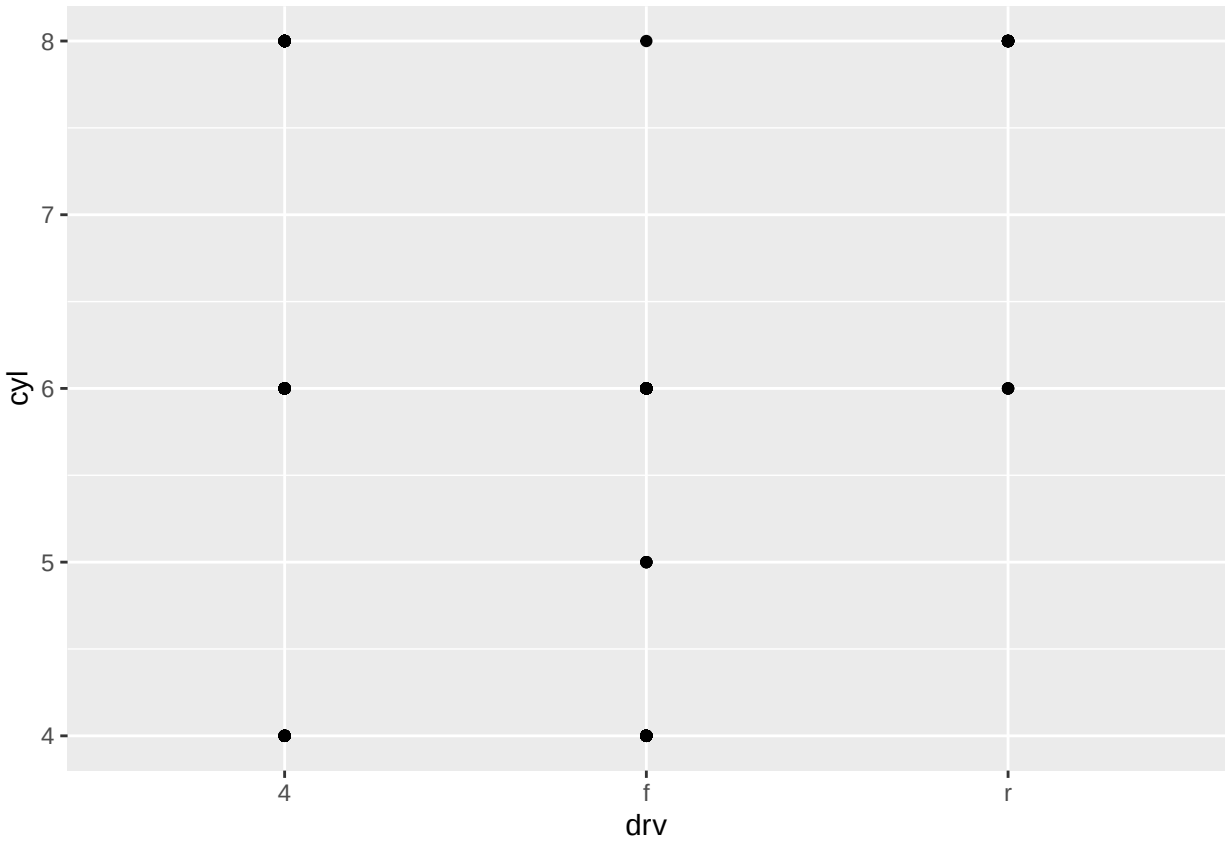
Q2 What do the empty cells in a plot with facet_grid(drv ~ cyl) mean? How do they relate to this plot?

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



have two empty plots at the lower left corners

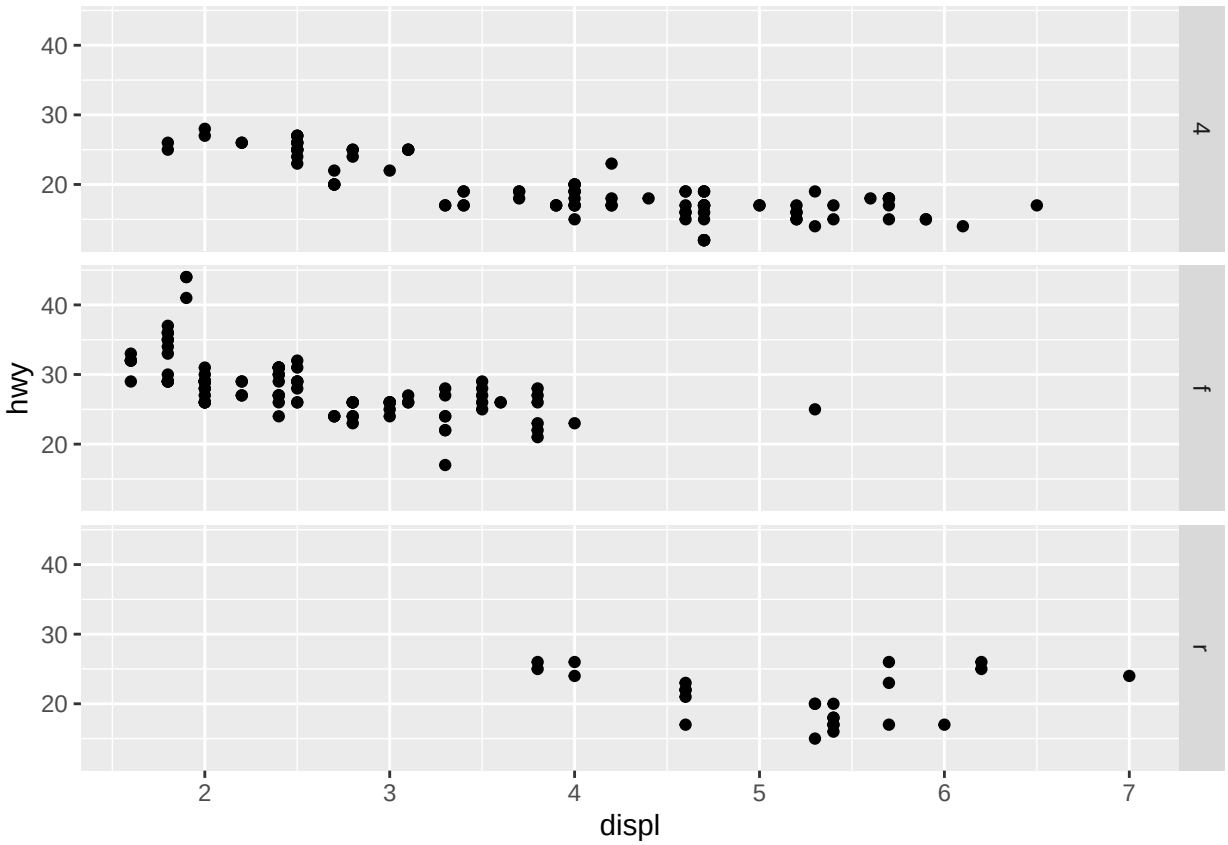
```
ggplot(mpg) +
  geom_point(aes(drv,cyl))
```



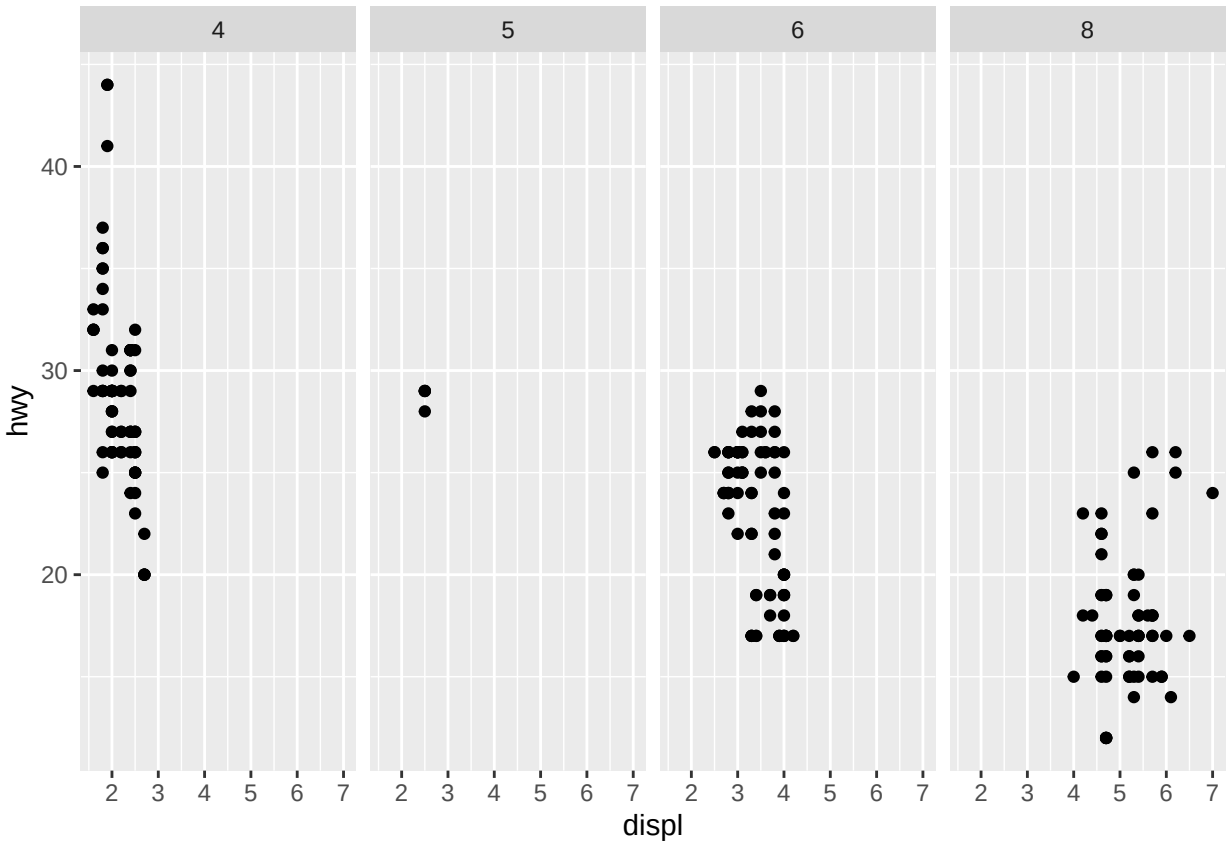
The empty cells (facets) in this plot are combinations of drv and cyl that have no observations. These are the same locations in the scatter plot of drv and cyl that have no points. (from the book but i still dont get it)

Q3 What plots does the following code make? What does . do?

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



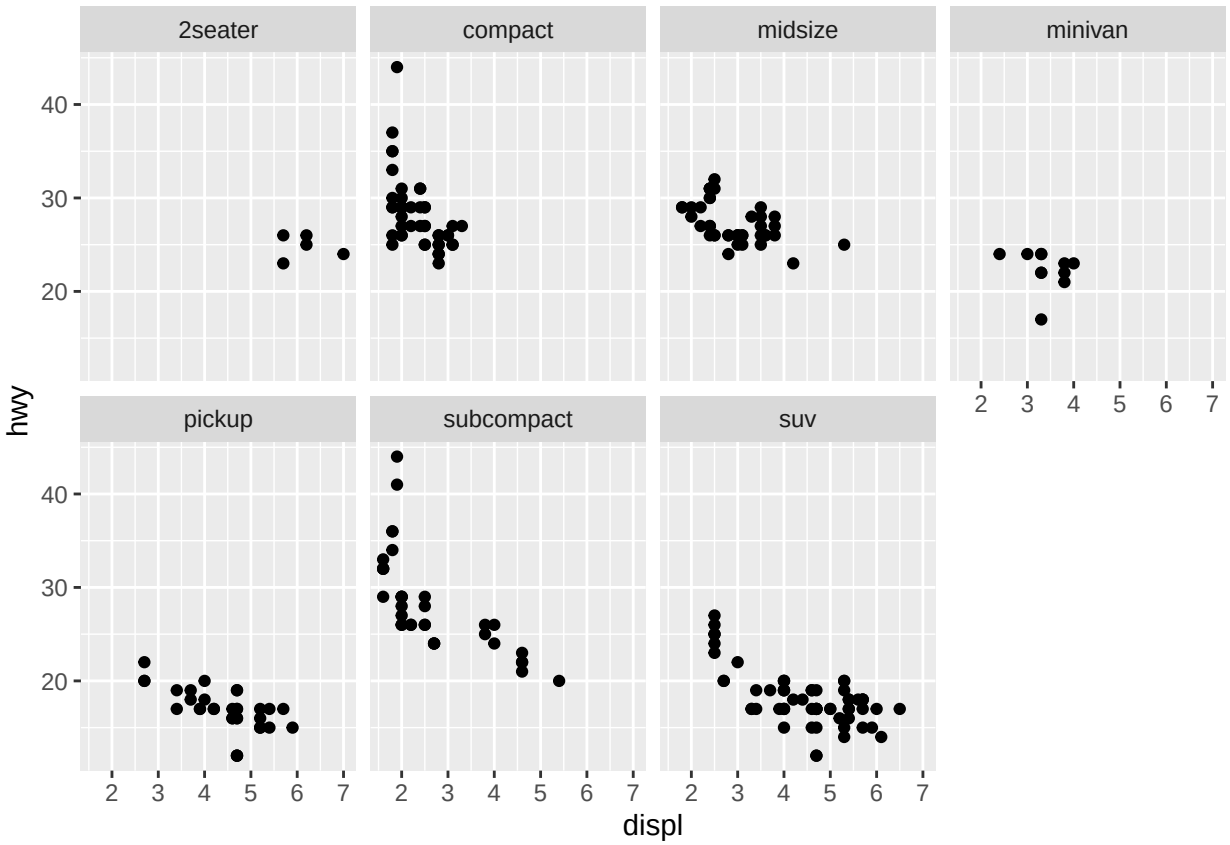

```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy)) +
  facet_grid(. ~ cyl)
```



Well, by the book the dot ignores the dimension when faceting. just remember 1 thing, facet_grid(y ~ x)

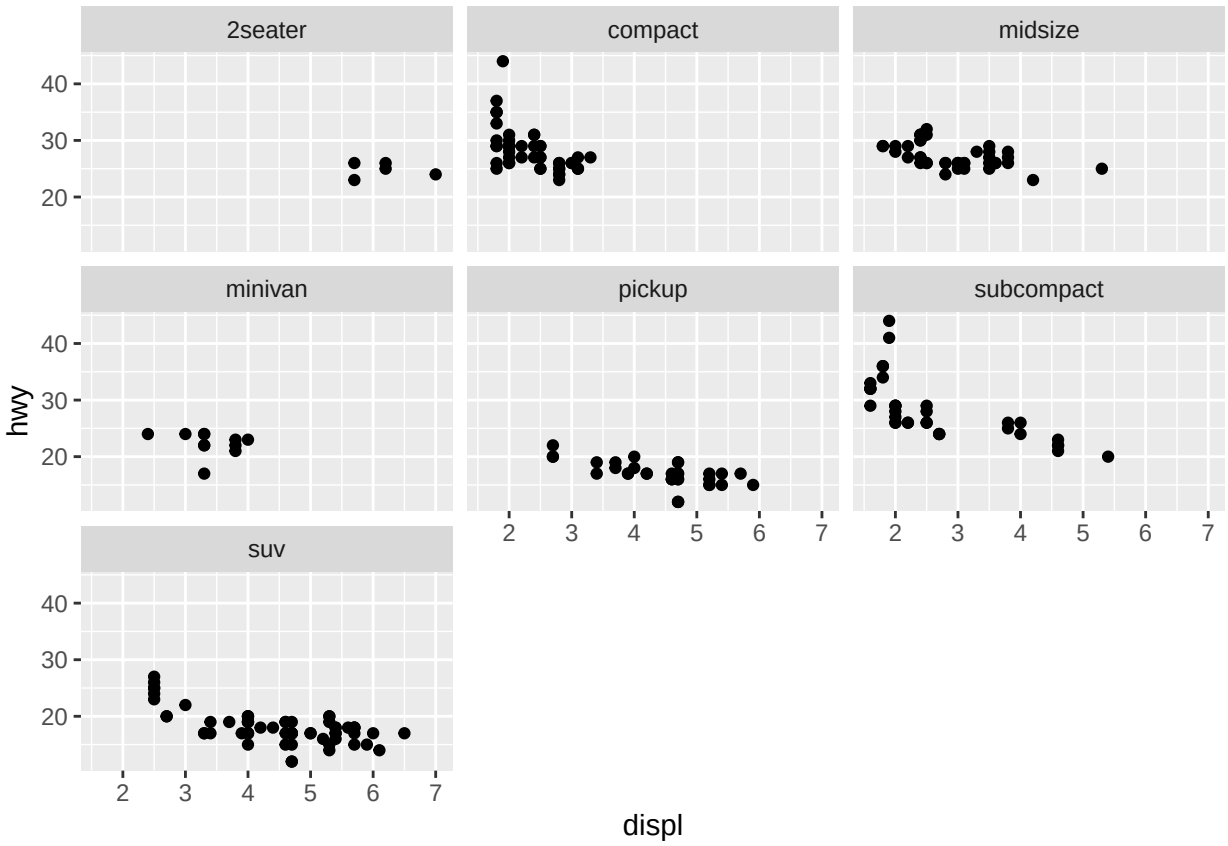
Q4 Take the first faceted plot in this section: ggplot(data = mpg) + geom_point(mapping = aes(x = displ, y = hwy)) + facet_wrap(~ class, nrow = 2) What are the advantages to using faceting instead of the color aesthetic? What are the disadvantages? How might the balance change if you had a larger dataset?

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



obviously, by using facet_wrap it is more readable.

```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy)) +
  facet_wrap(~ class, nrow = 3, ncol = 3)
```



We can define how many rows and cols, also for facet_grid, we do not have to identify how many cols and rows.

Q6 When using facet_grid() you should usually put the variable with more unique levels in the columns. Why? Since we usually have more spacing for cols than rows