Stat. 450 Section 1 or 2: Homework 1

Prof. Eric A. Suess

So how should you complete your homework for this class?

- First thing to do is type all of your information about the problems you do in the text part of your R Notebook.
- Second thing to do is type all of your R code into R chunks that can be run.
- If you load the tidyverse in an R Notebook chunk, be sure to include the "message = FALSE" in the {r}, so {r message = FALSE}.
- Last thing is to spell check your R Notebook. Edit > Check Spelling... or hit the F7 key.

Homework 1

```
Read: Chapter 1, 2, 3
Download an install the current version of R and RStudio.
Do 3.2.4 Exercises 1, 2, 3, 4, 5
Do 3.3.1 Exercises 1, 2, 3, 4, 6
Do 3.5.1 Exercises 1, 2, 4
```

3.2.4 Exercises

1.

We see nothing. Well actually we see the first layer of a ggplot2 plot.

```r
library(tidyverse)

ggplot(data = mpg)
```

2.

By viewing the mpg dataframe we see there are 234 rows and 11 columns.

```
mpg
```

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

3.

The variable drv has levels: f = front-wheel drive, r = rear wheel drive, 4 = 4wd

```r
help(mpg)  # opens the help file

?mpg  # another way to open the help file.

str(mpg)  # traditional way to look at the variables in a dataframe
```

```
## 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
glimpse(mpg)  # the way to look at the variables in a tibble
```

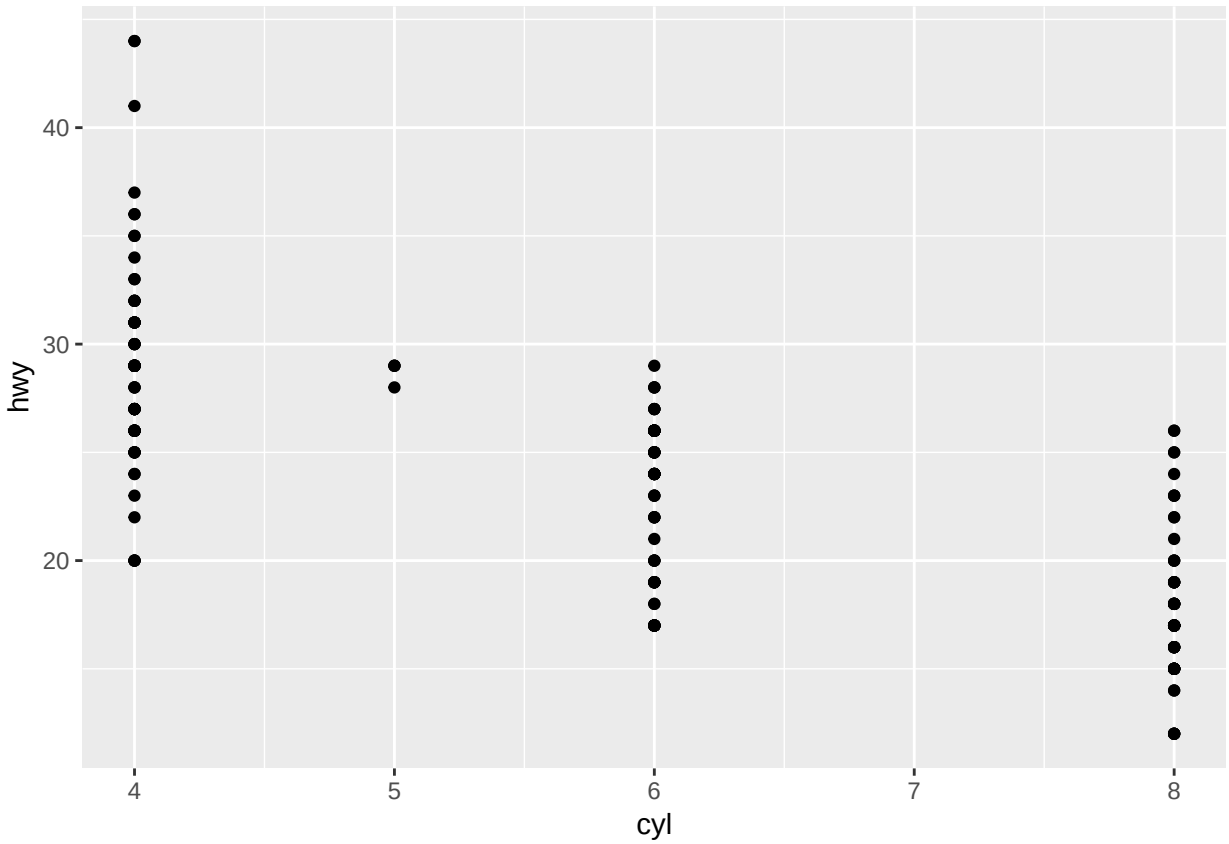
```
## Observations: 234
## Variables: 11
## $ manufacturer <chr> "audi", "audi", "audi", "audi", "audi", "audi", "...
## $ model        <chr> "a4", "a4", "a4", "a4", "a4", "a4", "a4", "a4 qua...
## $ displ        <dbl> 1.8, 1.8, 2.0, 2.0, 2.8, 2.8, 3.1, 1.8, 1.8, 2.0,...
## $ year         <int> 1999, 1999, 2008, 2008, 1999, 1999, 2008, 1999, 1...
## $ cyl          <int> 4, 4, 4, 4, 6, 6, 6, 4, 4, 4, 4, 6, 6, 6, 6, 6, 6...
## $ trans        <chr> "auto(l5)", "manual(m5)", "manual(m6)", "auto(av)...
## $ drv          <chr> "f", "f", "f", "f", "f", "f", "f", "4", "4", "4",...
## $ cty          <int> 18, 21, 20, 21, 16, 18, 18, 18, 16, 20, 19, 15, 1...
## $ hwy          <int> 29, 29, 31, 30, 26, 26, 27, 26, 25, 28, 27, 25, 2...
## $ fl           <chr> "p", "p", "p", "p", "p", "p", "p", "p", "p", "p",...
## $ class        <chr> "compact", "compact", "compact", "compact", "comp...
```

```r
View(mpg)  # opens the data in a spreadsheet in RStudio
```

4.

Scatterplot of y = hwy versus x = cyl. The average highway miles per gallon goes down as the number of cylinders increases.

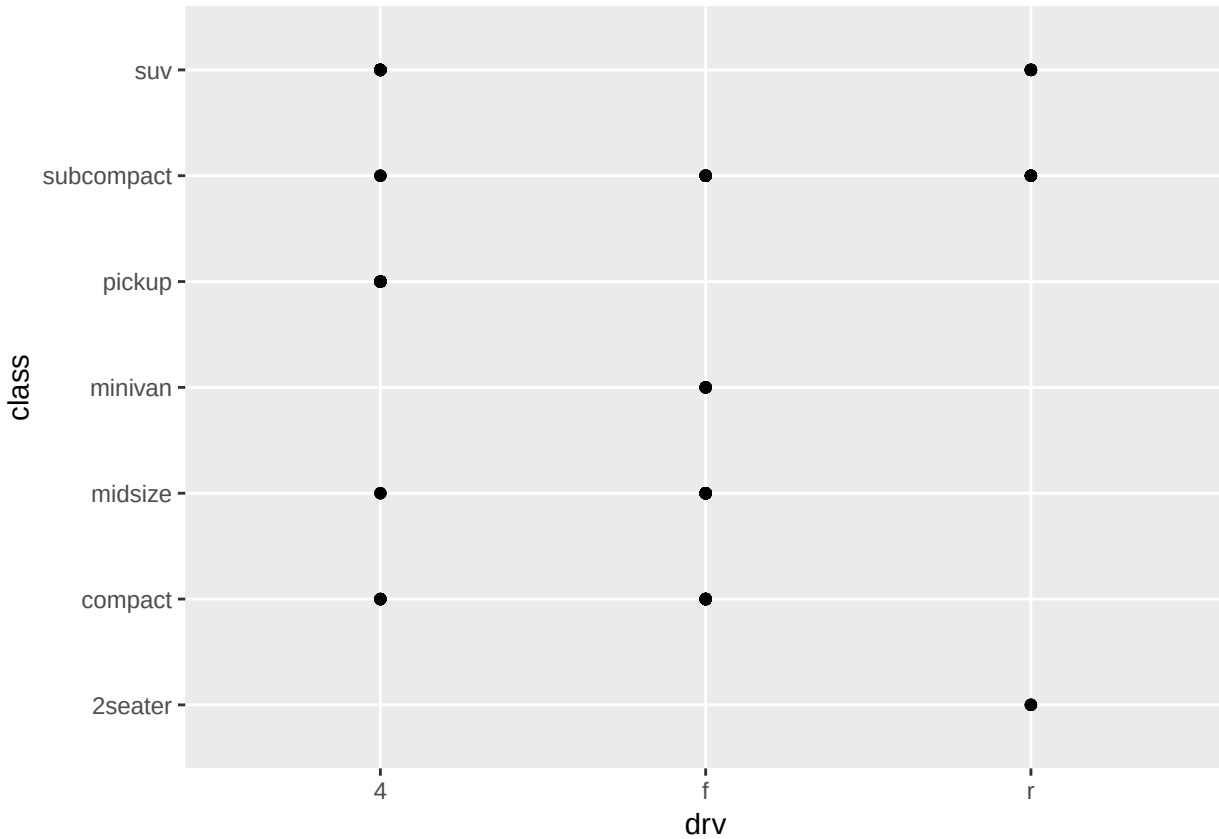
```
ggplot(mpg, aes(y = hwy, x = cyl)) +
  geom_point()
```



5.

This is not useful because there are many observations on each point in the plot. Plotting categorical variables in a scatterplot is not useful. It would be better to make a contingency table.

```
ggplot(mpg, aes(y = class, x = drv)) +
  geom_point()
```




```
count(mpg, drv, class)
```

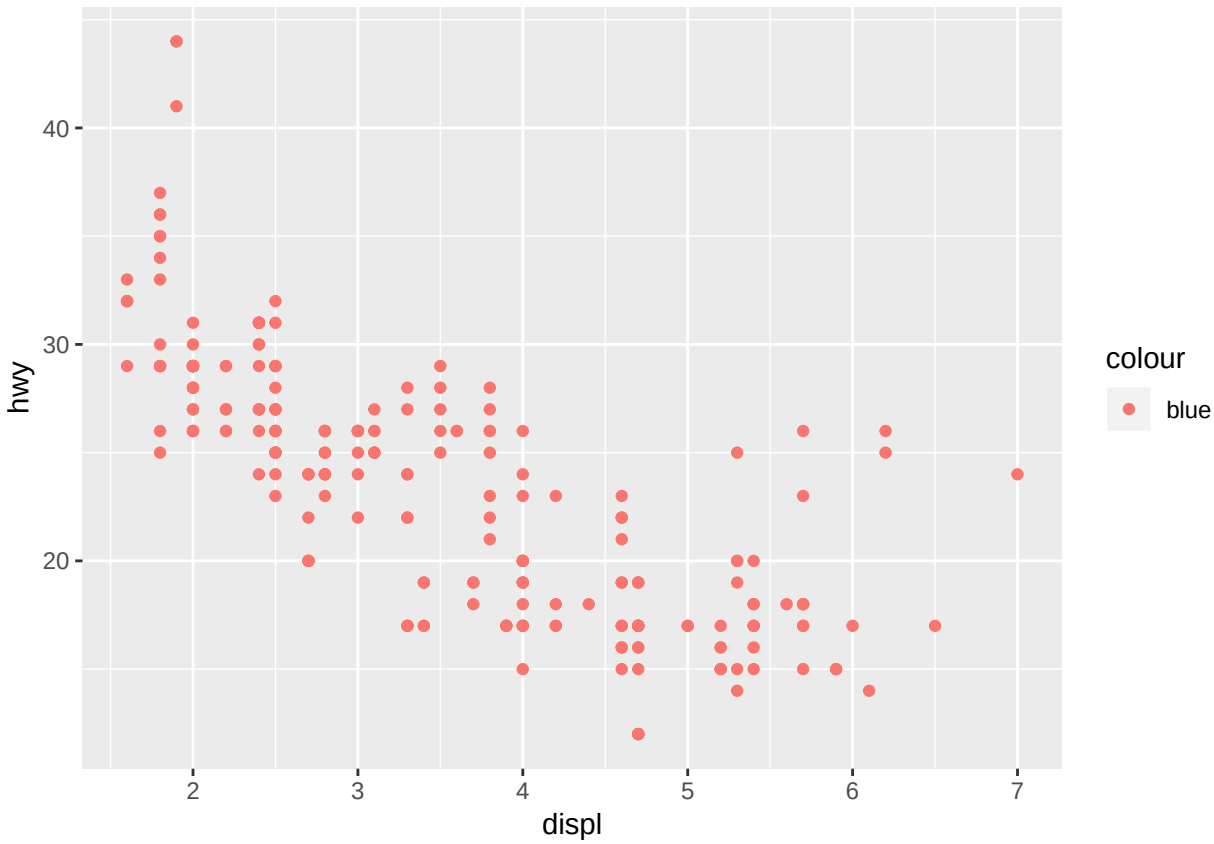
```
## # A tibble: 12 x 3
##    drv   class          n
##    <chr> <chr>      <int>
##  1 4     compact       12
##  2 4     midsize        3
##  3 4     pickup        33
##  4 4     subcompact     4
##  5 4     suv           51
##  6 f     compact       35
##  7 f     midsize       38
##  8 f     minivan       11
##  9 f     subcompact    22
## 10 r     2seater        5
## 11 r     subcompact     9
## 12 r     suv           11
```

3.3.1 Exercises

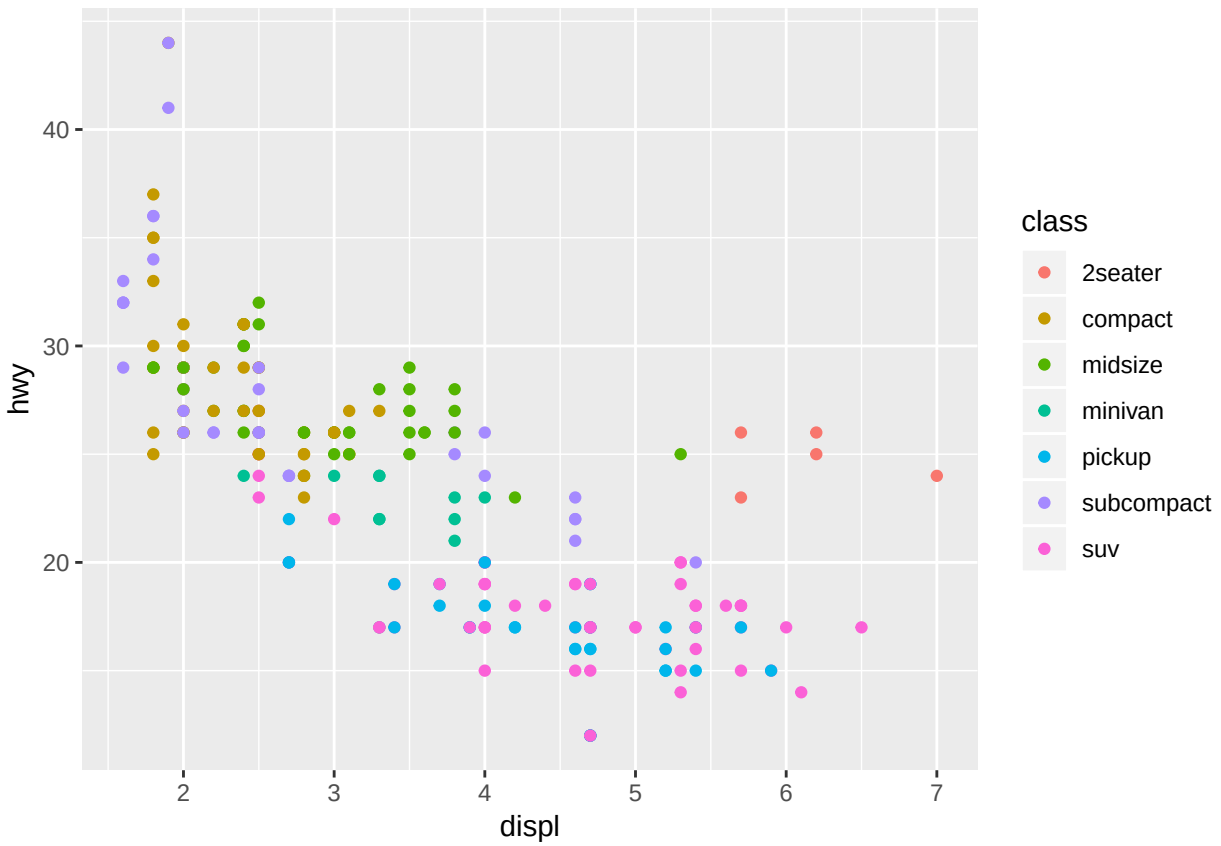
1.

If color is in the aes as a mapping it would need a variable from the dataframe to give the plot different colors. For example, putting in drv as the color. Alternatively, to change all of the points to blue, the color needs to be outside of the aes.

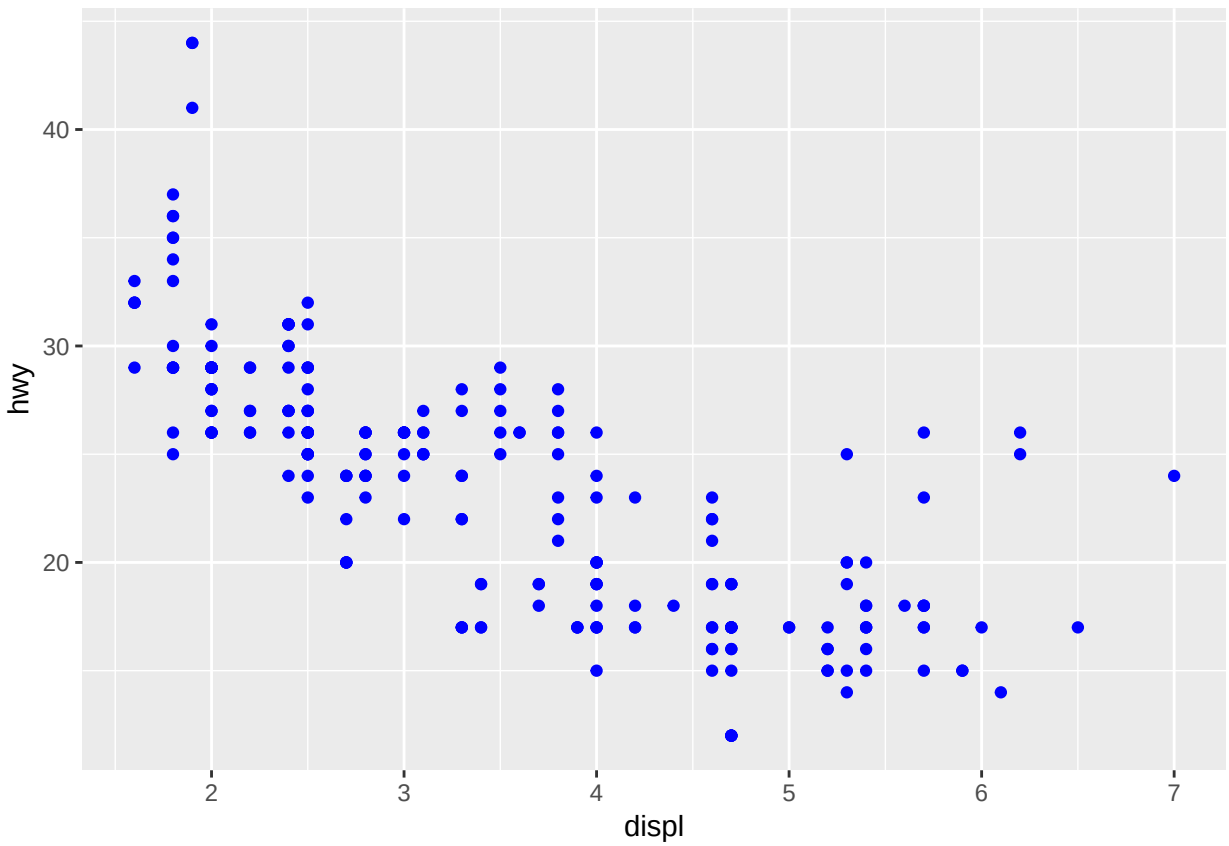
```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy, color = "blue"))
```




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




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


40
30
20
hwy
2
3
4
5
6
7
displ

2.

The categorical variables are the ones with under the variable names. The continuous variables are the ones with under them.

```
mpg
```

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

```
?mpg

glimpse(mpg)
```

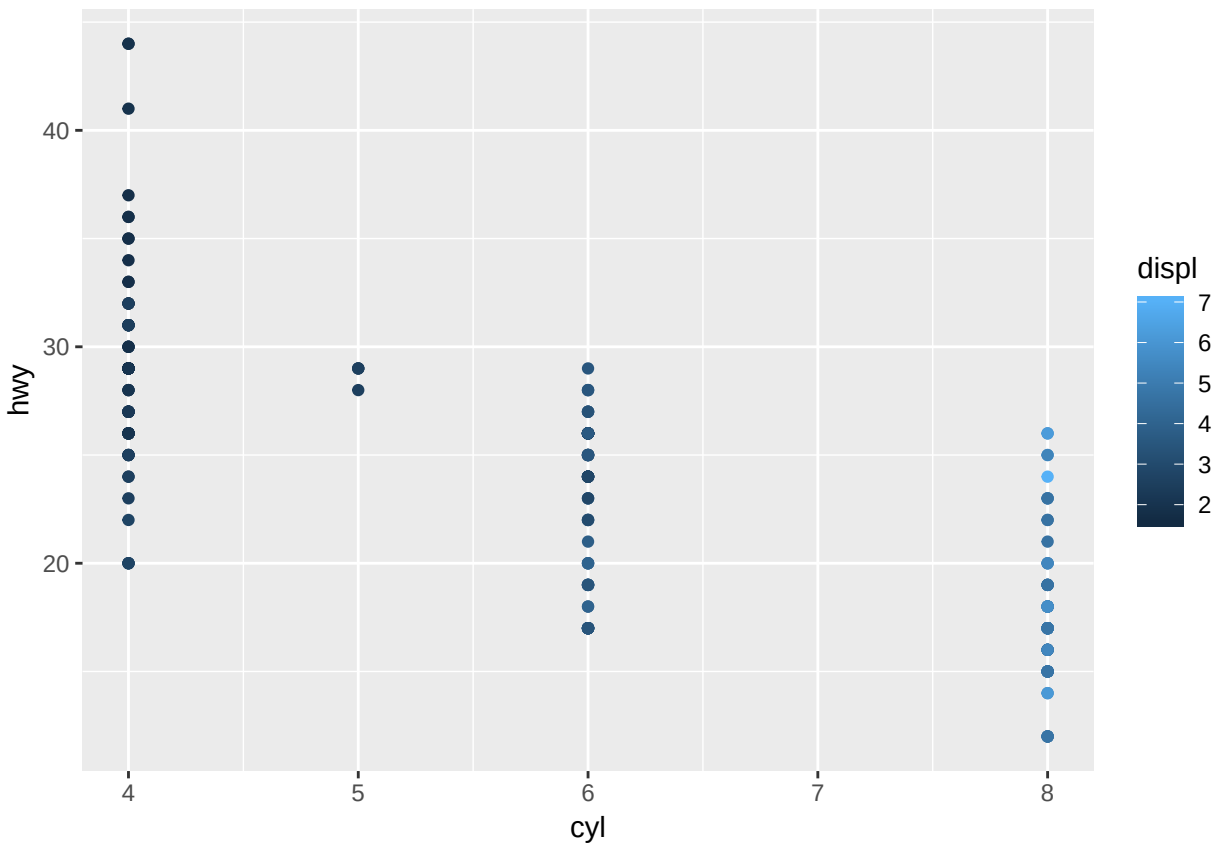
```
## Observations: 234
## Variables: 11
## $ manufacturer <chr> "audi", "audi", "audi", "audi", "audi", "audi", "...
## $ model        <chr> "a4", "a4", "a4", "a4", "a4", "a4", "a4", "a4 qua...
## $ displ        <dbl> 1.8, 1.8, 2.0, 2.0, 2.8, 2.8, 3.1, 1.8, 1.8, 2.0,...
## $ year         <int> 1999, 1999, 2008, 2008, 1999, 1999, 2008, 1999, 1...
## $ cyl          <int> 4, 4, 4, 4, 6, 6, 6, 4, 4, 4, 4, 6, 6, 6, 6, 6, 6...
## $ trans        <chr> "auto(l5)", "manual(m5)", "manual(m6)", "auto(av)...
## $ drv          <chr> "f", "f", "f", "f", "f", "f", "f", "4", "4", "4",...
## $ cty          <int> 18, 21, 20, 21, 16, 18, 18, 18, 16, 20, 19, 15, 1...
## $ hwy          <int> 29, 29, 31, 30, 26, 26, 27, 26, 25, 28, 27, 25, 2...
## $ fl           <chr> "p", "p", "p", "p", "p", "p", "p", "p", "p", "p",...
## $ class        <chr> "compact", "compact", "compact", "compact", "comp...
```

3.

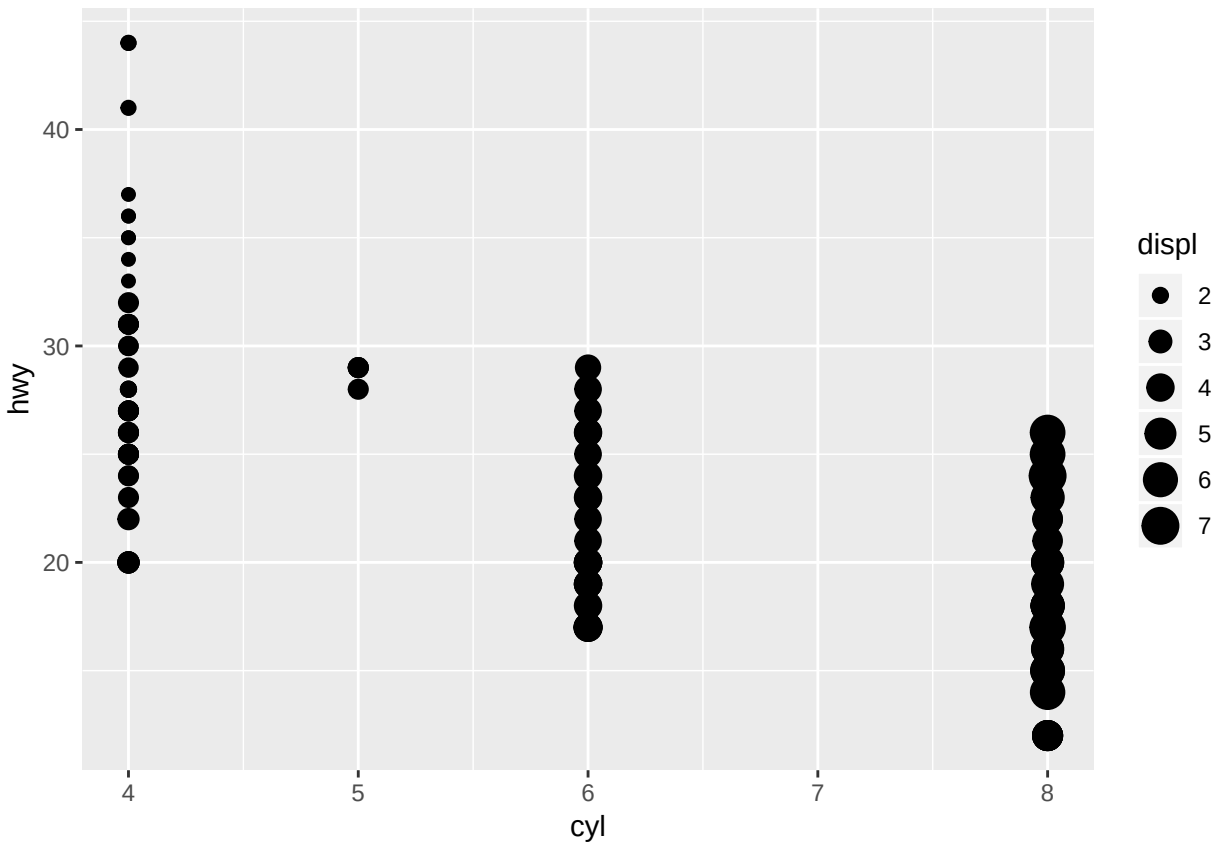
Map a continuous variable to color, size, and shape. How do these aesthetics behave differently for categorical vs. continuous variables?

Brighter colors are used for higher values of color. Bigger shapes are used for higher values of size. A continuous variable cannot be used for shape.

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

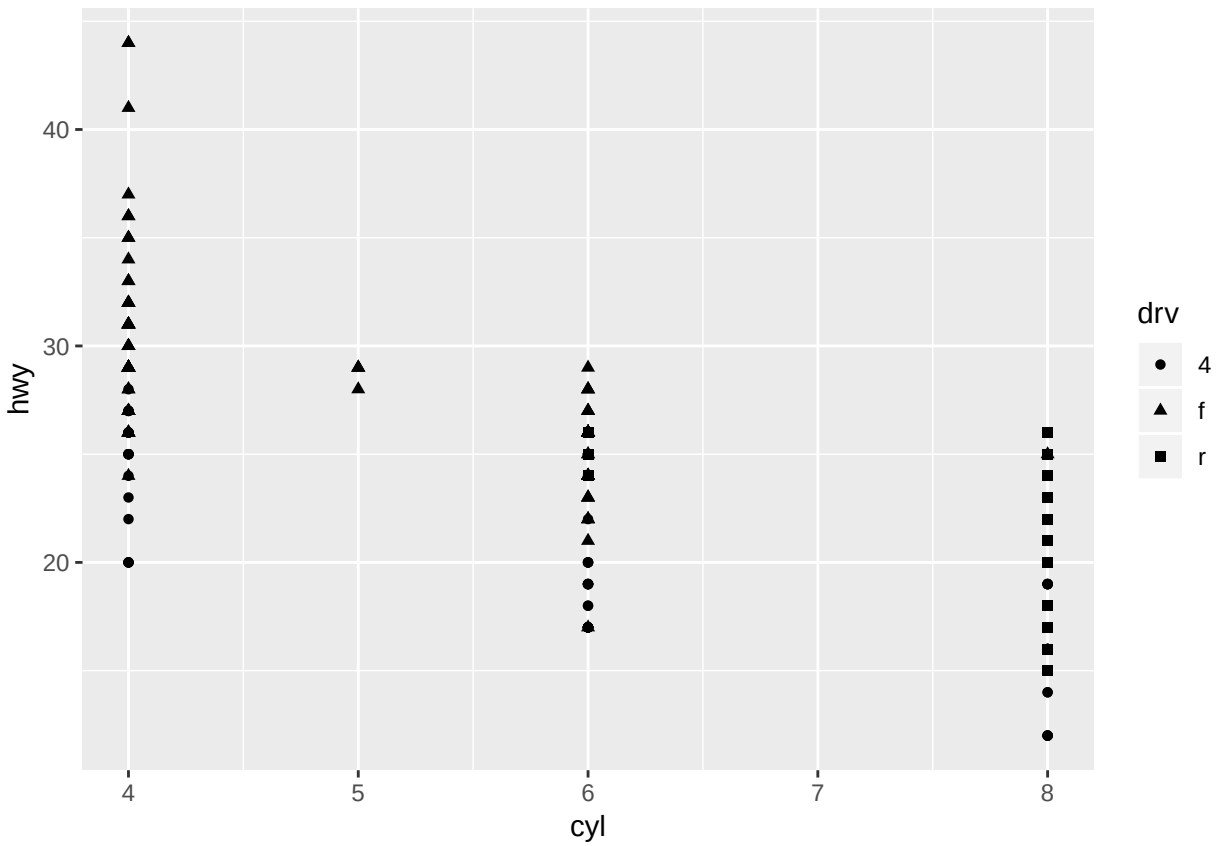



```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = cyl, y = hwy, size = displ))
```




```
# ggplot(data = mpg) +
#   geom_point(mapping = aes(x = cyl, y = hwy, shape = displ))  # gives a error

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

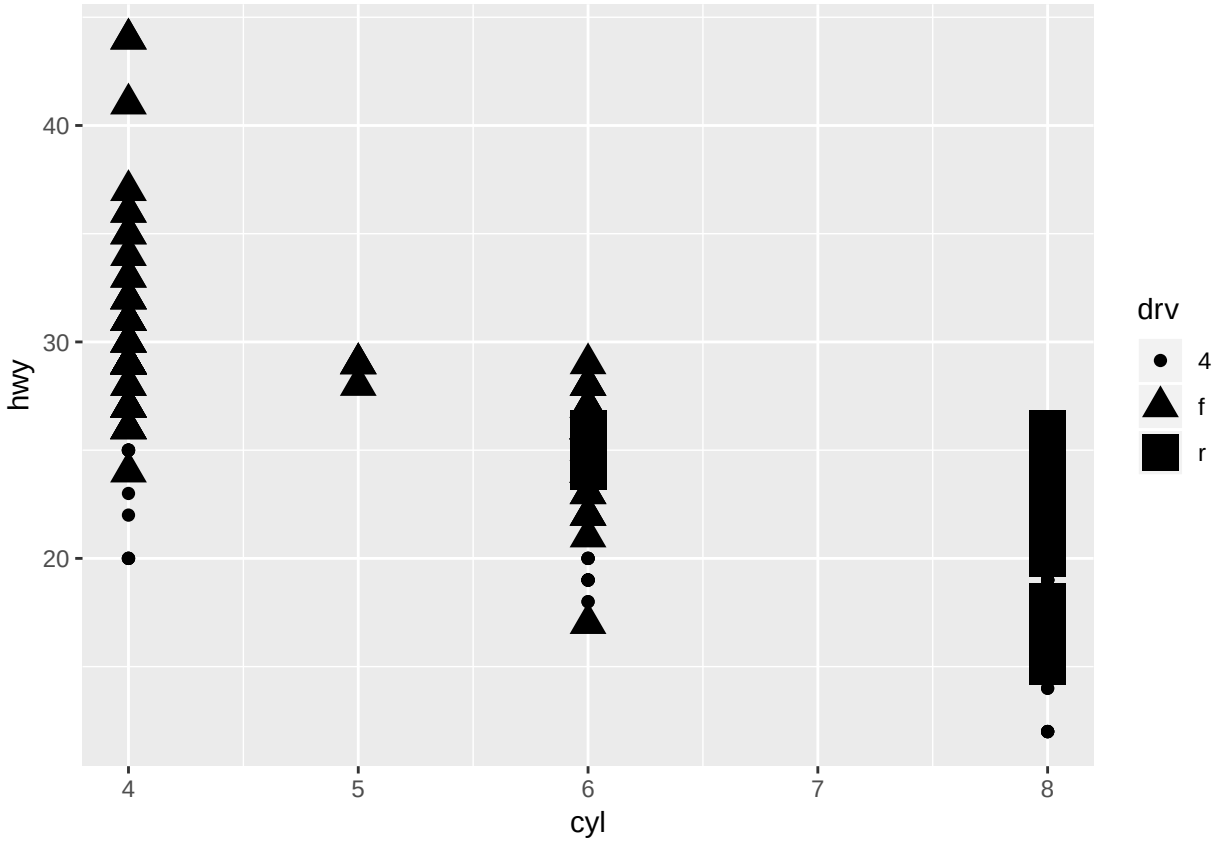

40
30
20
hwy
4
5
6
7
8
cyl
drv
4
f
r

4.

Can use two mappings for the same variable. This is not good practice!

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

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

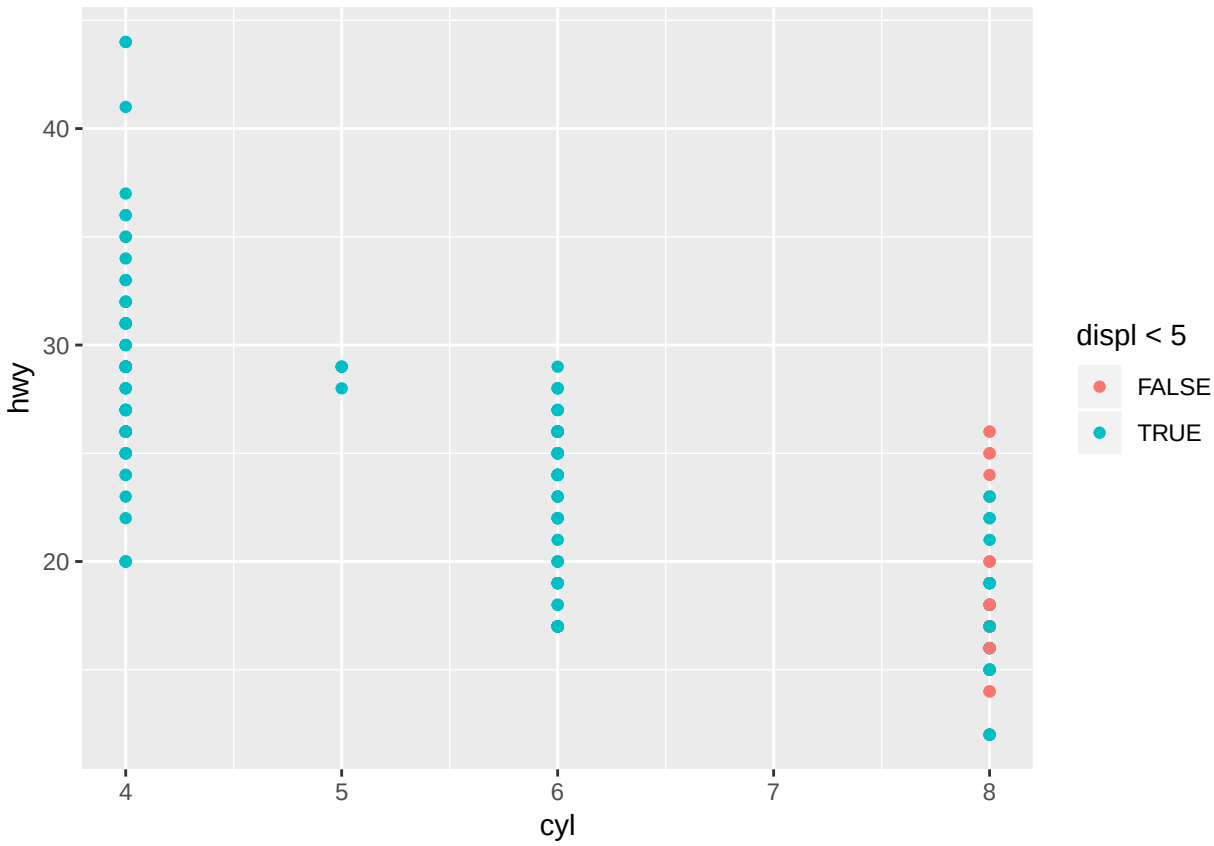


6.

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

The color changes for the TRUE and FALSE values of the inequality.

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

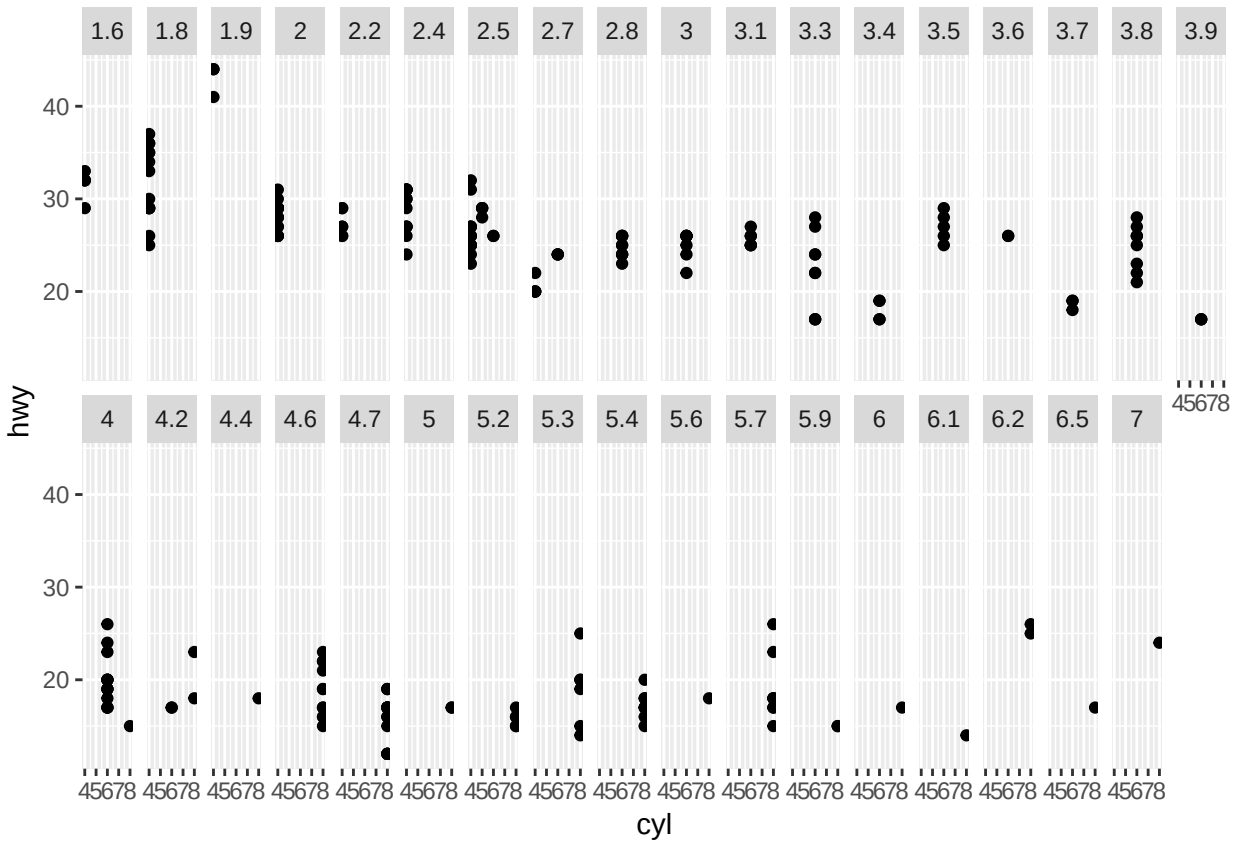


3.5.1 Exercises

1.

Is a continuous variable is used, each value of the variable is used. So potentially many many plots will be made. This may not be useful. Faceting should be done with a categorical variable.

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

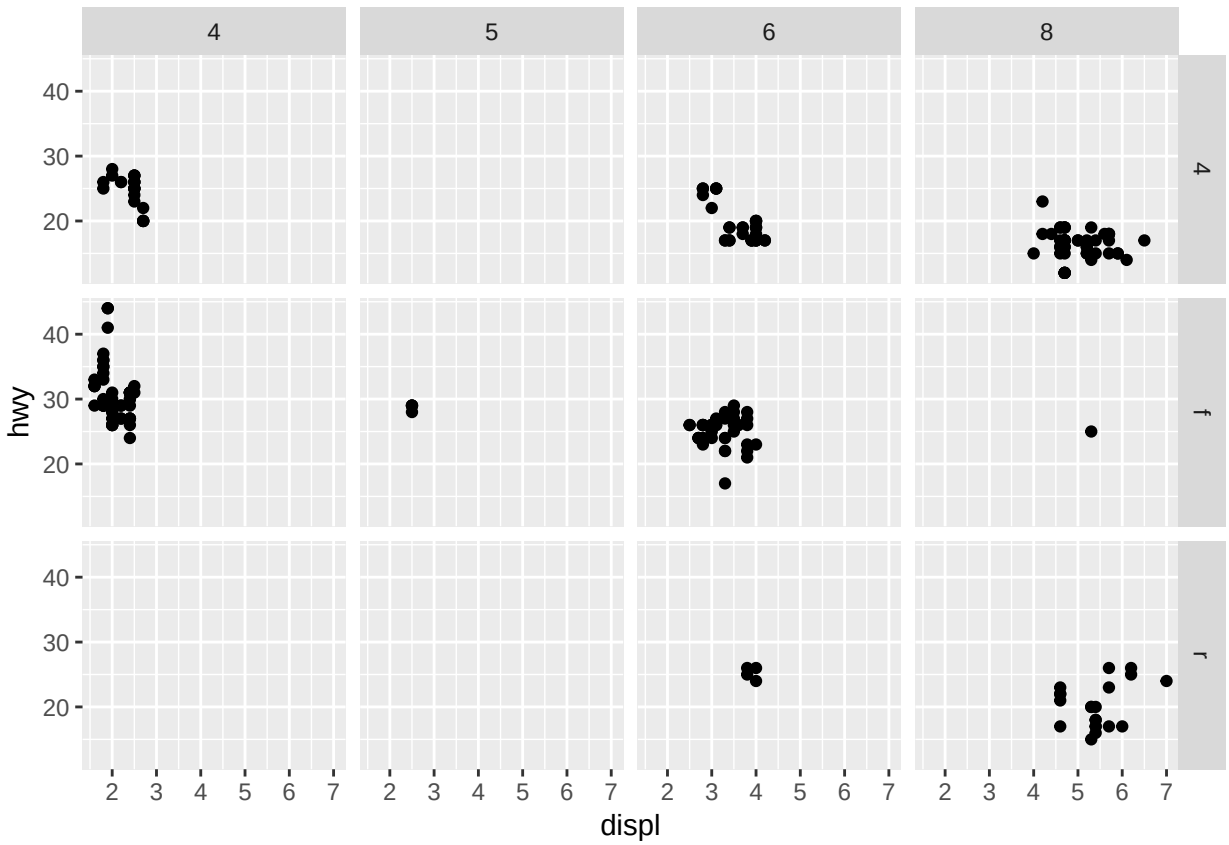


2.

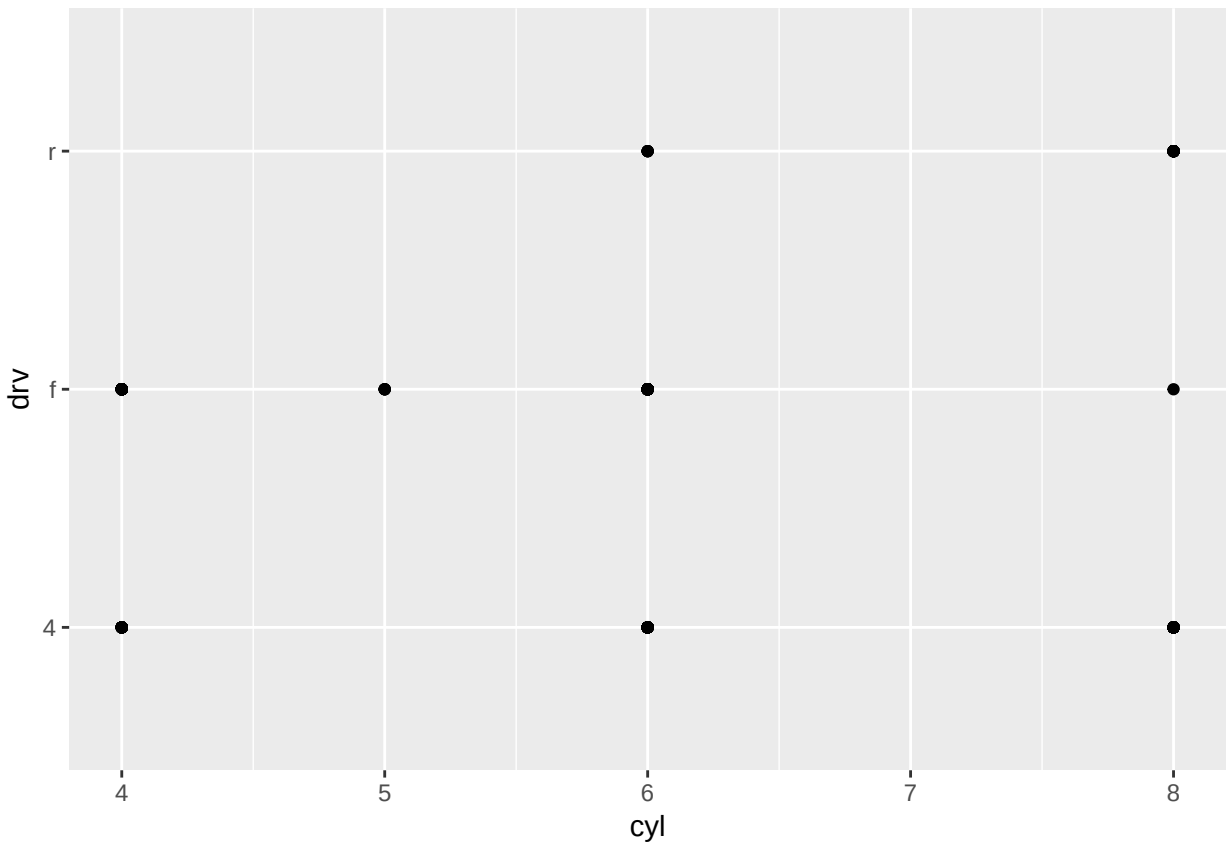
The missing cells in the plot means there is no data available for that combination of values of the variables.

Switch x and y in the second plot to see the relationship.

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




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

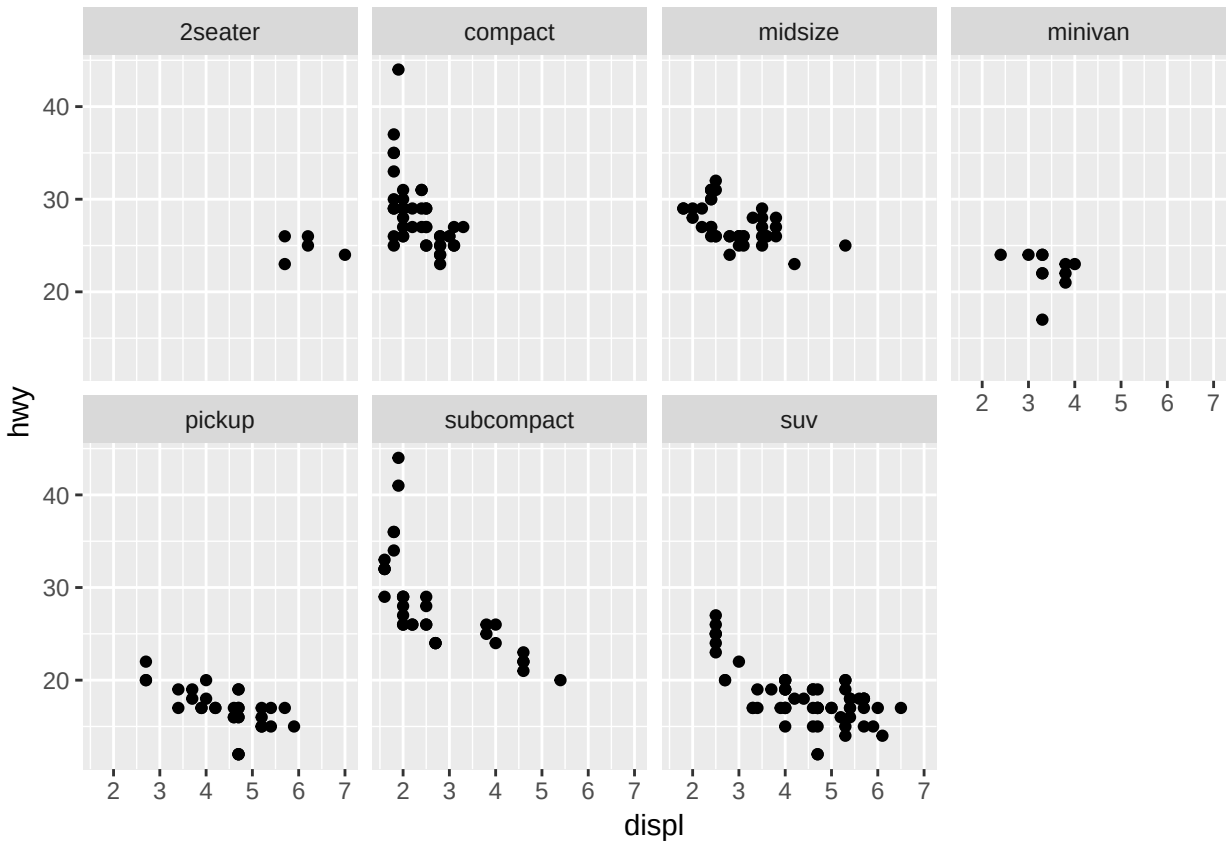

r
f
4
drv
4
5
6
7
8
cyl

4.

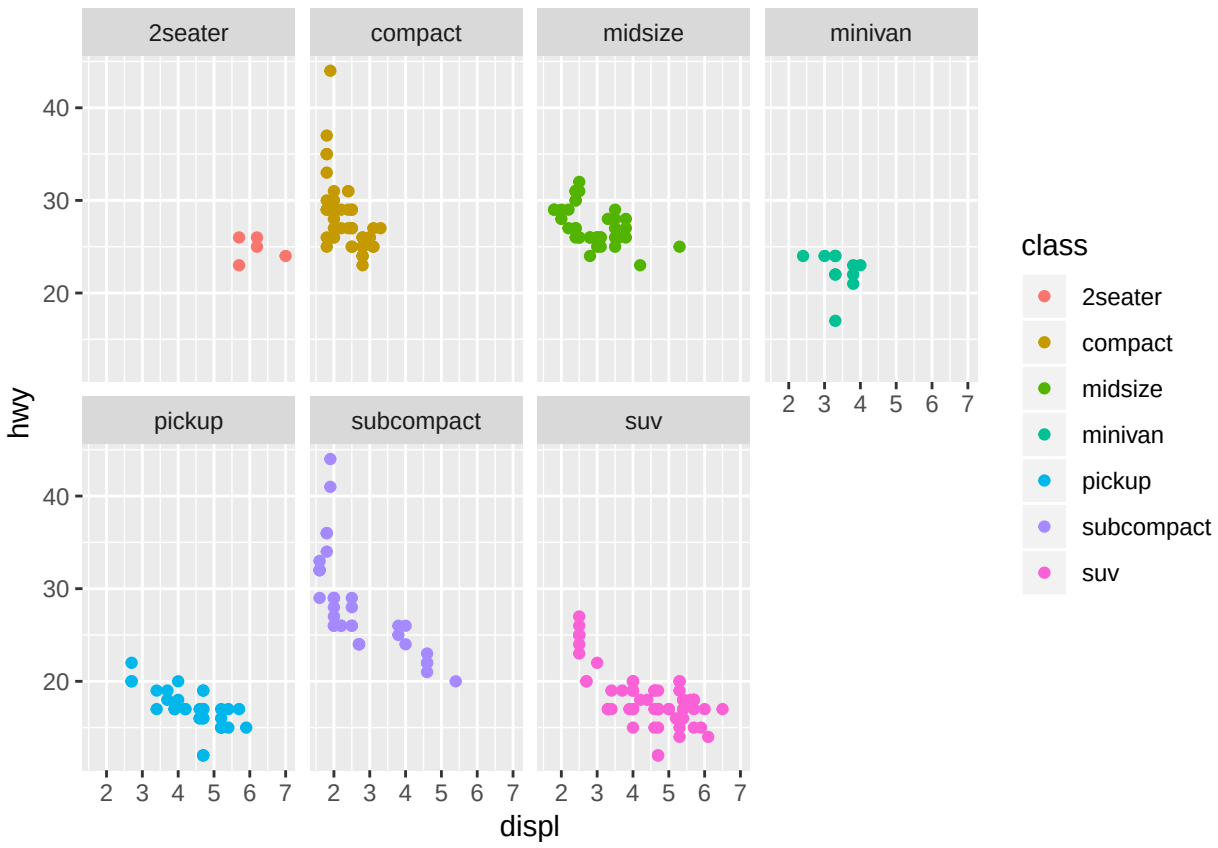
Compare to 3.3.1 Exercise 1.

Faceting make is easier to see where the data is relative to the other variable used for color.

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




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


2seater
compact
midsize
minivan
pickup
subcompact
suv
hwy
40
30
20
2 3 4 5 6 7
displ
class
2seater
compact
midsize
minivan
pickup
subcompact
suv