Home Work Assignment 2

Pradeep Sahoo

6/14/2018

My Github repository for my assignments can be found at this URL: My Github

Exercises

Section 3.2.4: all exercises

1. Run ggplot(data = mpg). What do you see?

```
library(mdsr)
library(tidyverse)
```

```
data(mpg)
```

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


Count of Rows

```
count(mpg)
```

```
## # A tibble: 1 x 1
##       n
```

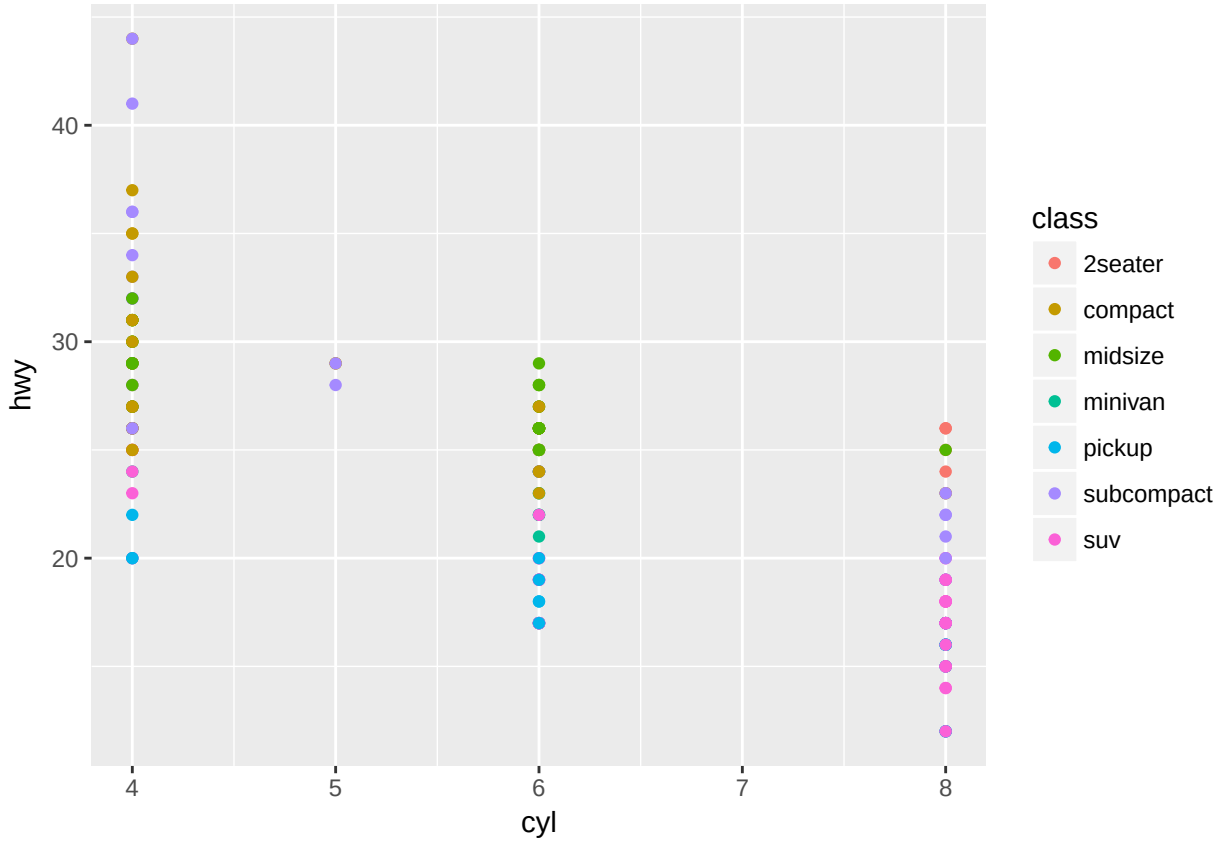
```
##   <int>
## 1   234
```

```
variable.names(mpg)
```

```
##  [1] "manufacturer" "model"        "displ"        "year"
##  [5] "cyl"          "trans"        "drv"          "cty"
##  [9] "hwy"          "fl"           "class"
```

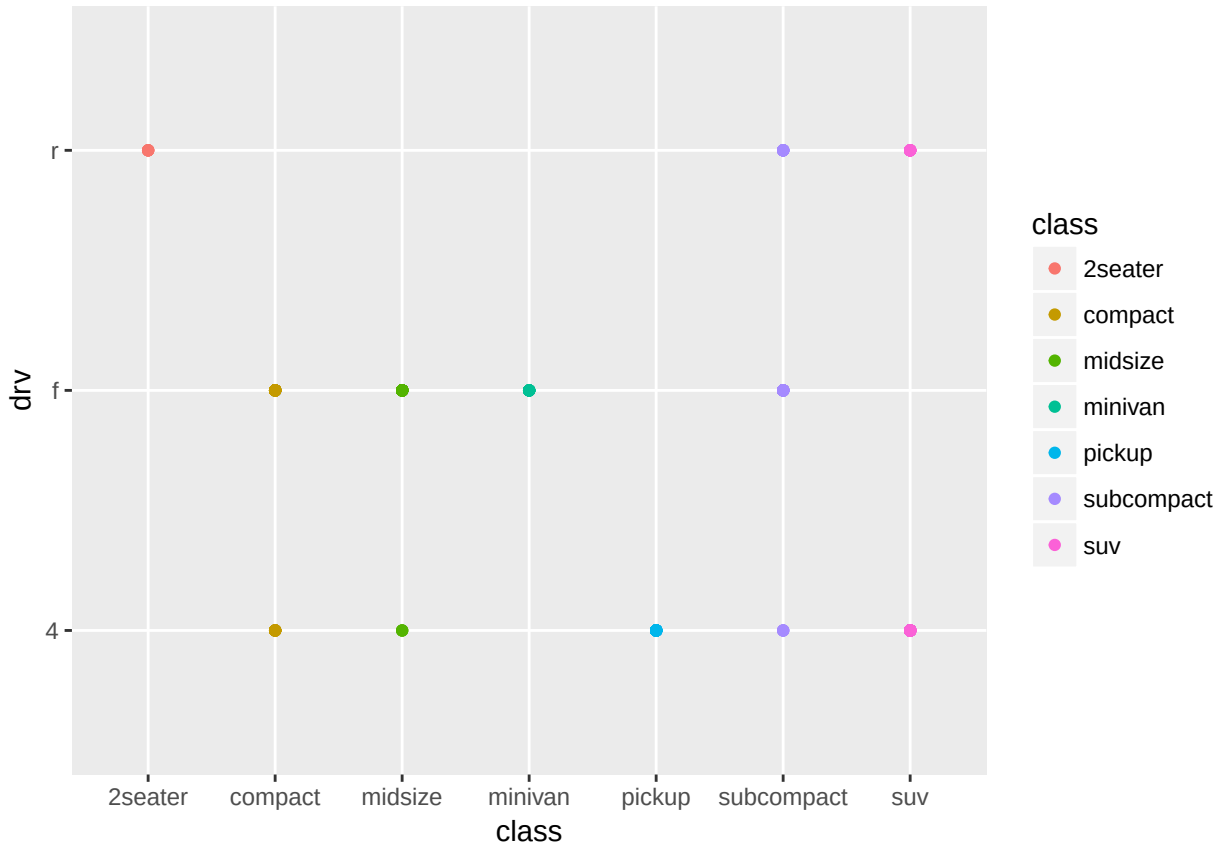
Make a scatterplot of hwy vs cyl.

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



What happens if you make a scatterplot of class vs drv? Why is the plot not useful?

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

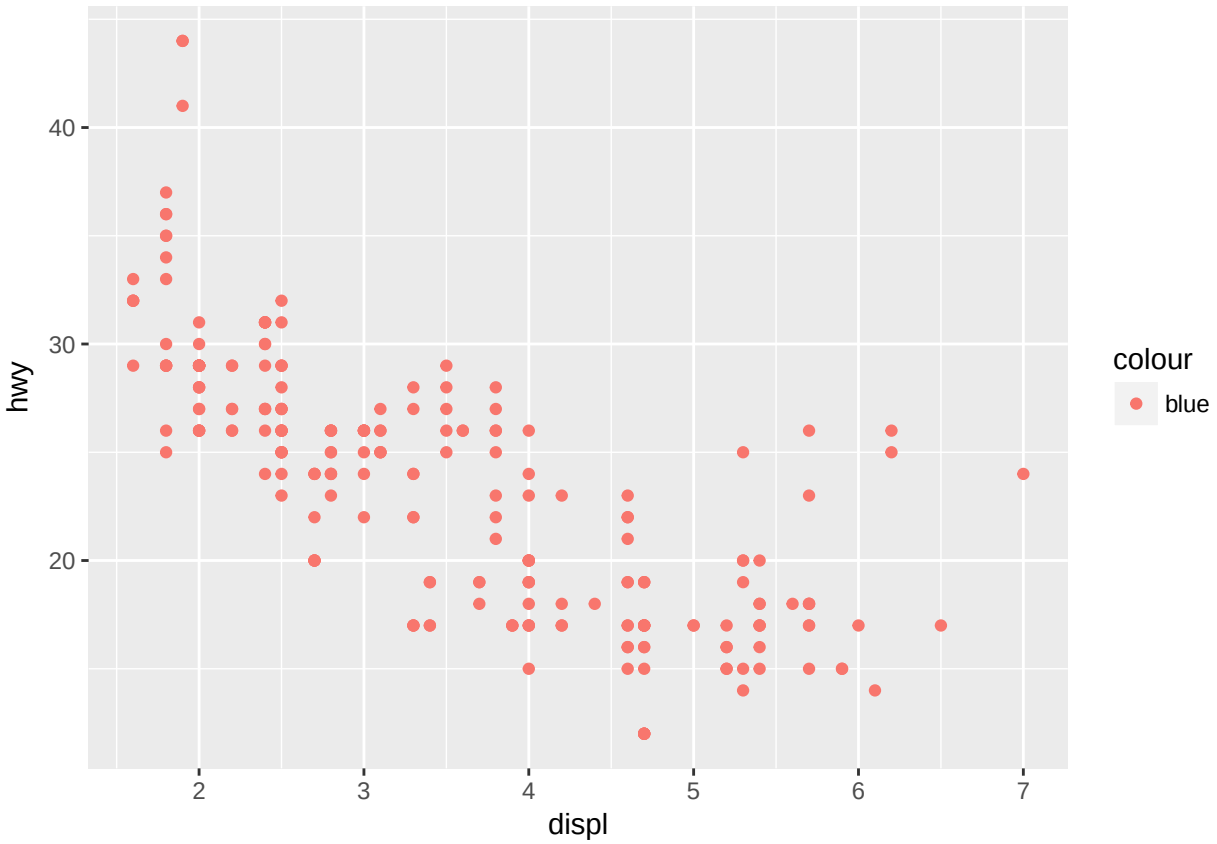


There is only 3 results for drv, which is not a ideal candidate for scatter plot.

Section 3.3.1: all exercises

What's gone wrong with this code? Why are the points not blue?

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



In this case the color is part of aesthetics, so any of the variables should be part of color so that the variables can be treated to group by.

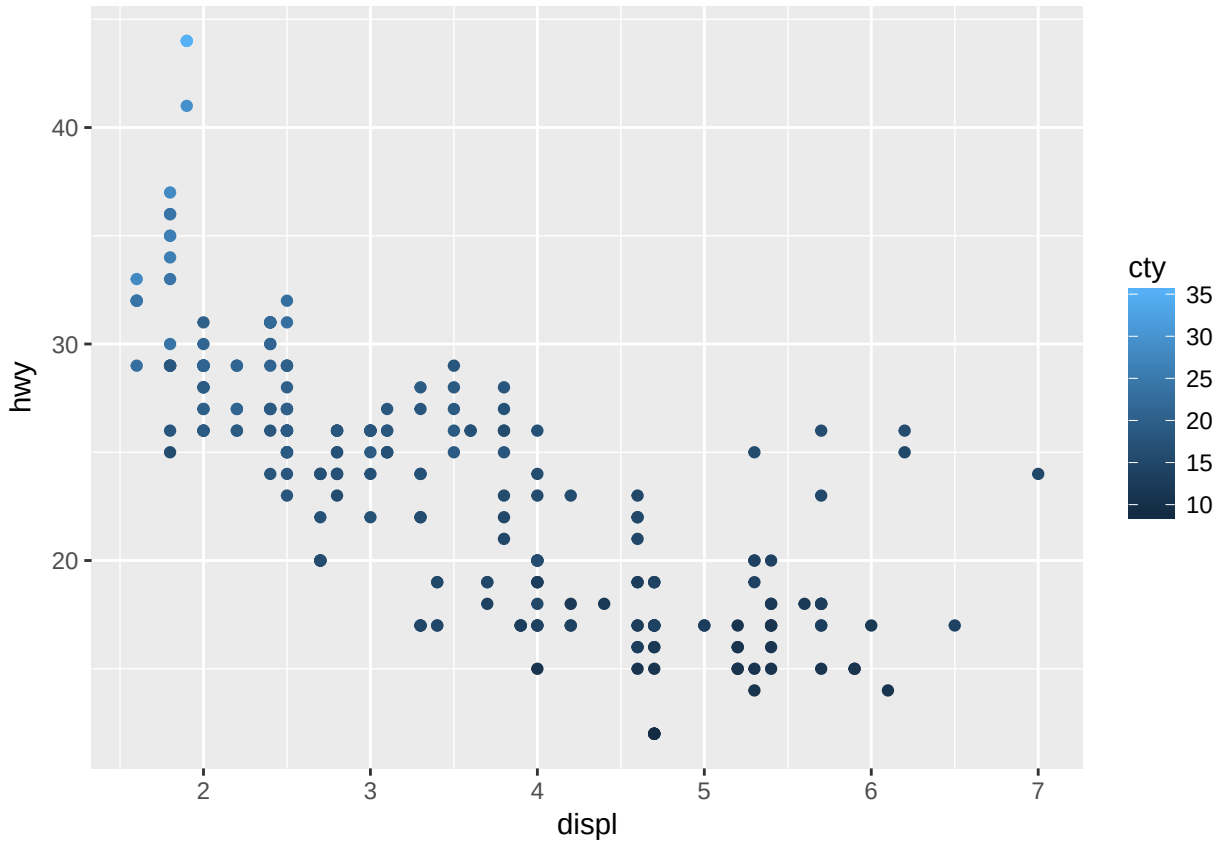
Which variables in mpg are categorical? Which variables are continuous? (Hint: type ?mpg to read the documentation for the dataset). How can you see this information when you run mpg?

Categorical variables : “manufacturer” “model” “year” “trans” “drv” “class”

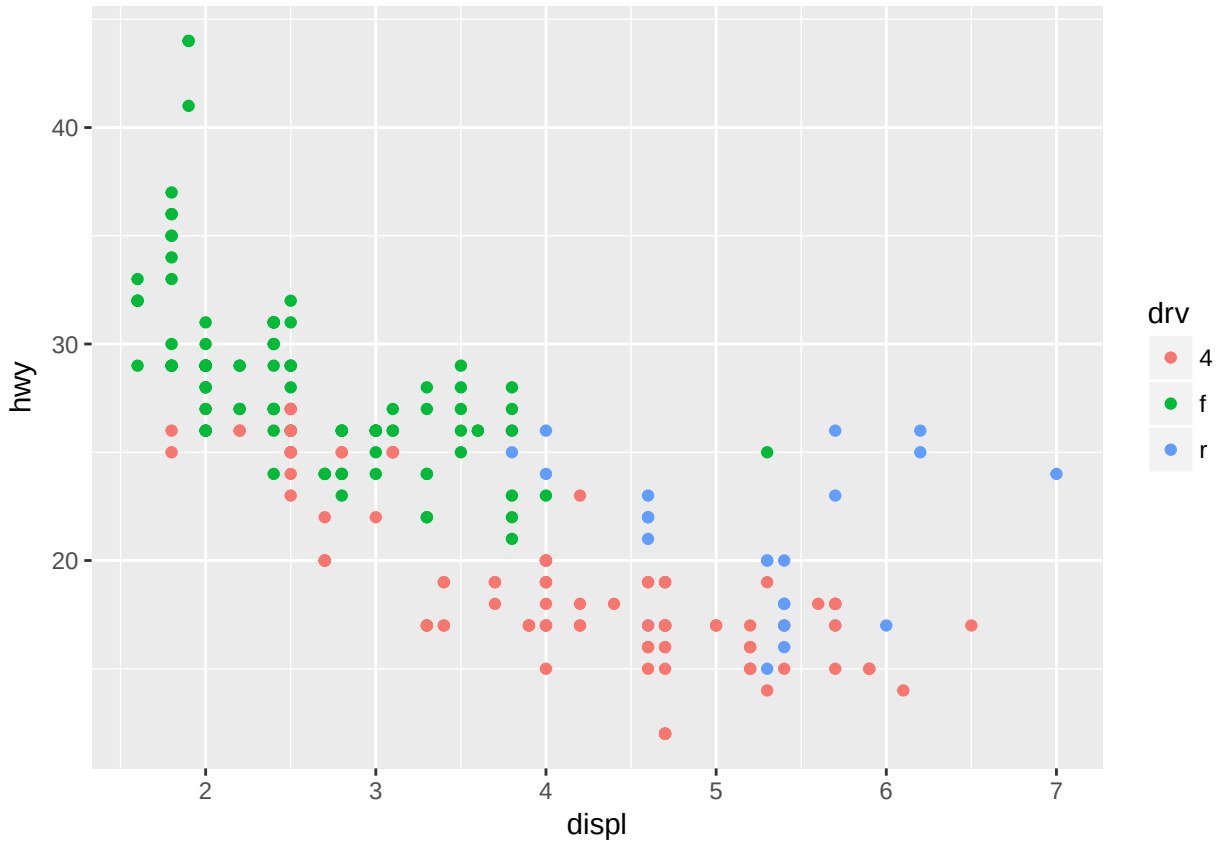
Continous Variable : “displ” “cyl” “cty” “hwy”

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

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

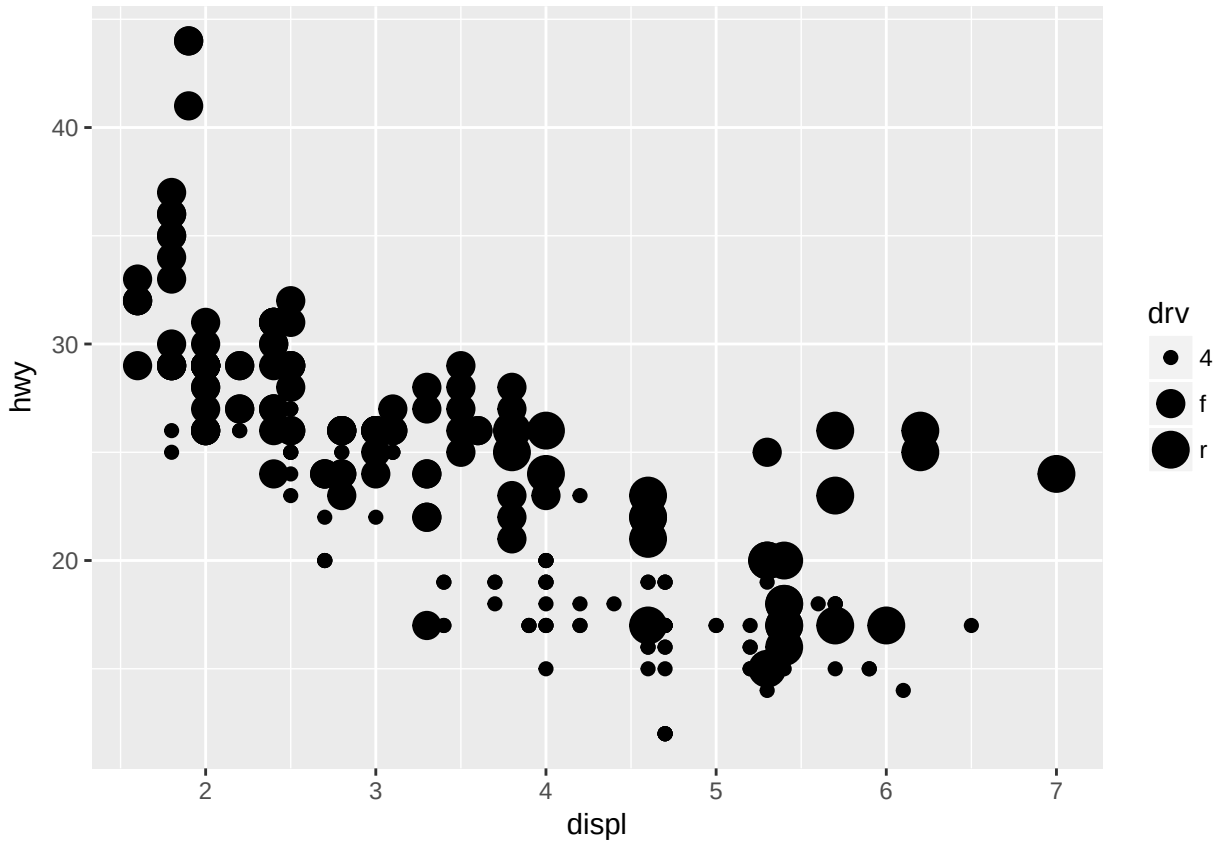



```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy, color = drv))
```

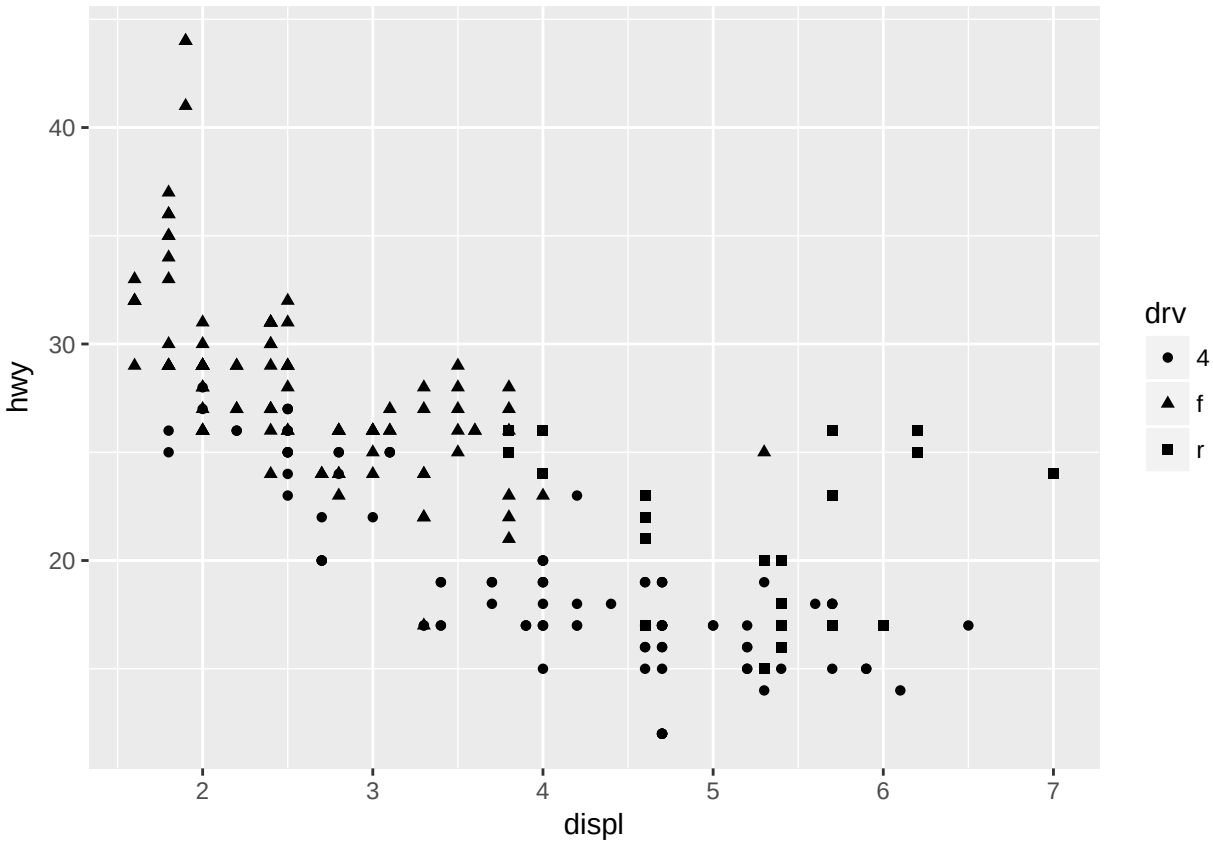



```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy, size = drv))
```

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




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

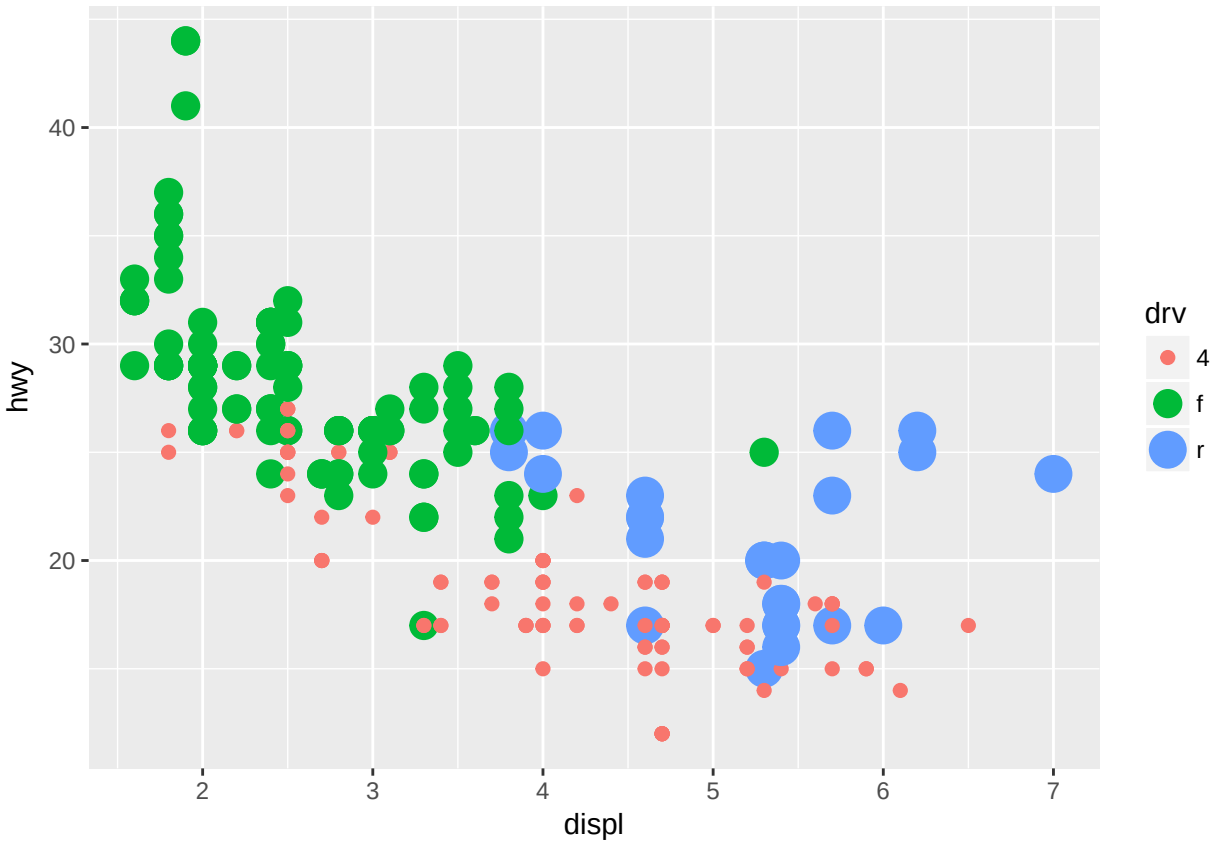


Categorical Variable used in Color, Shape, and Size gives clarity about the data. This helps in Analysis.

4. What happens if you map the same variable to multiple aesthetics?

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

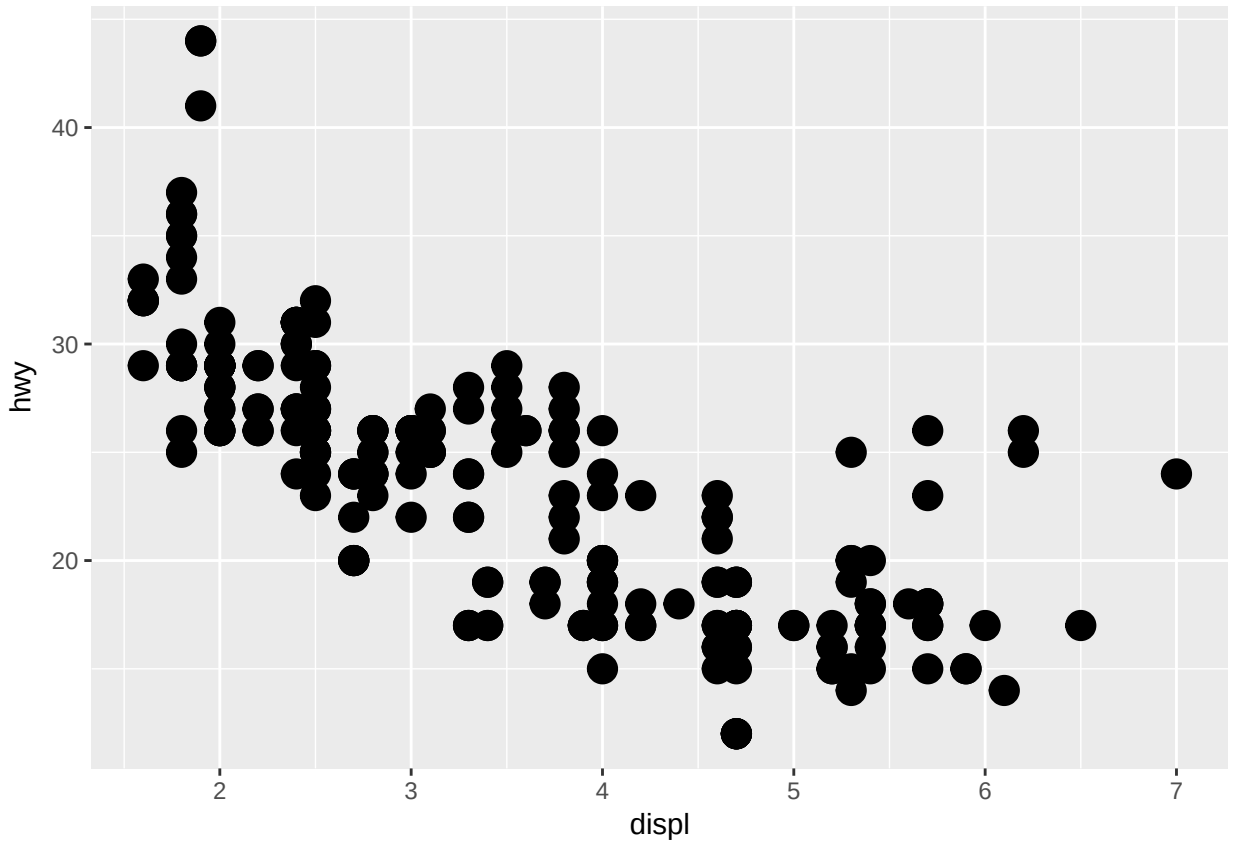
```
## Warning: Using size for a discrete variable is not advised.
```



Both the variables are shown in the legends. This behavior creates confusion.

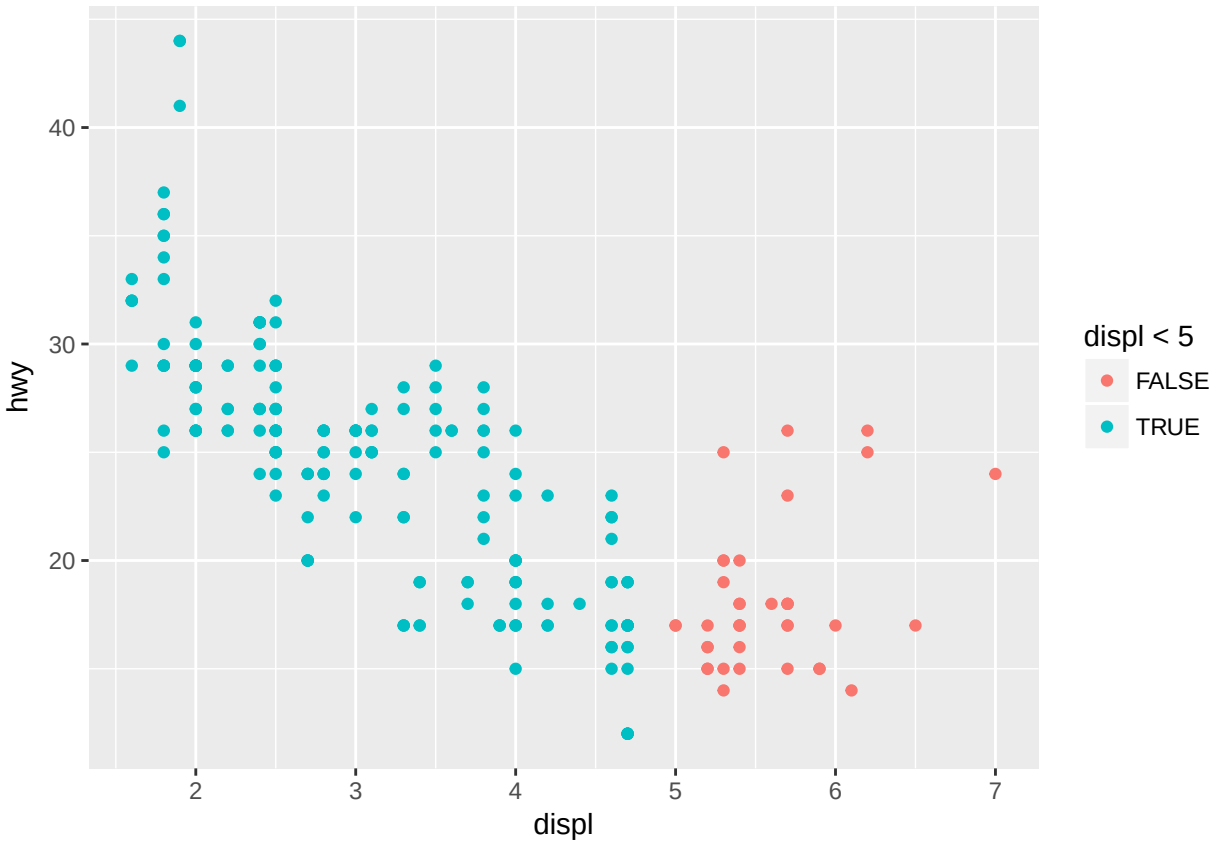
5. What does the stroke aesthetic do? What shapes does it work with? (Hint: use ?geom_point)

```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy), stroke = 3)
```



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

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

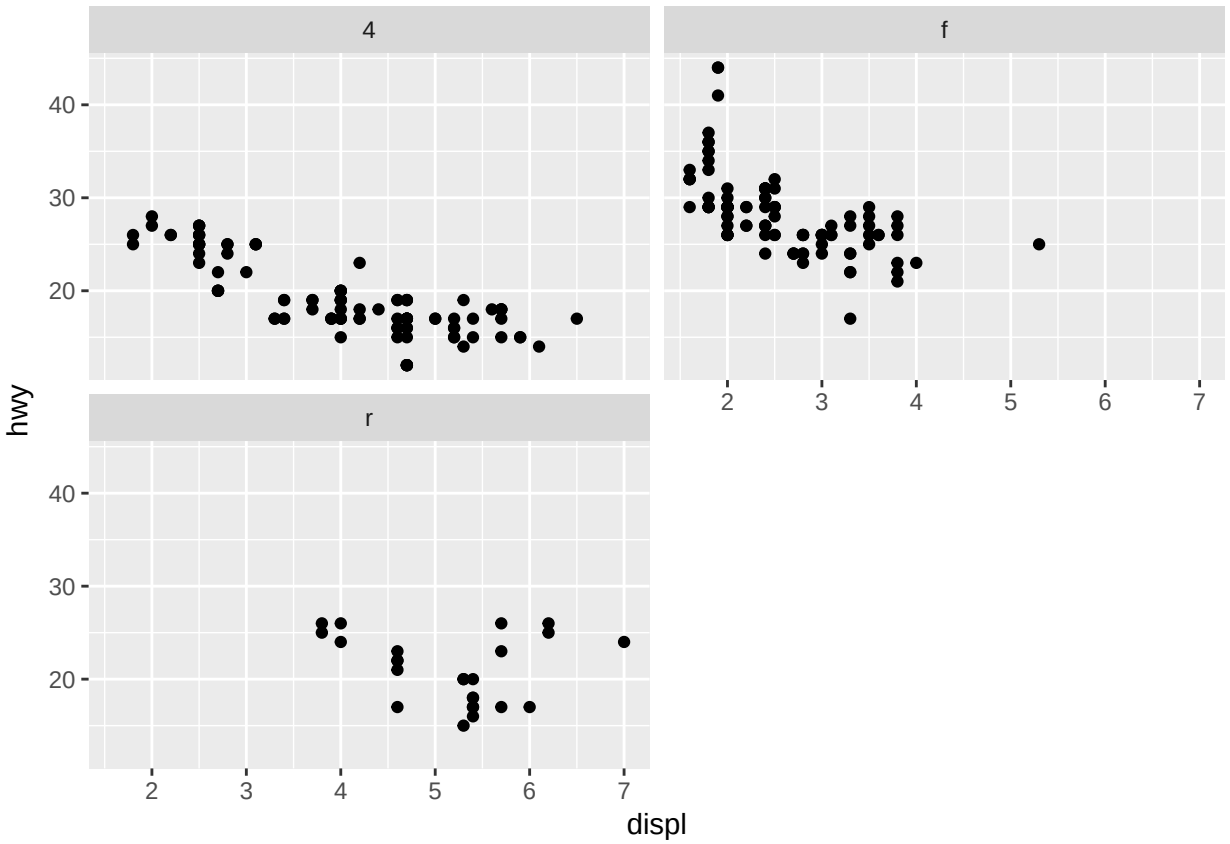


This will create a Temporary variable which is taking True or False based on the condition (displ < 5). That mean, if displ is < 5 then True else False.

Section 3.5.1: #4 and #5 only

4. Take the first faceted plot in this section:

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



Facet will help to analyze the data based on the Items of a Variable distinctly. Where as the ability to compare among the items is lost by using Facet. This is a disadvantage of Facets over Color. In case of Color this comparison or trend can be derived.

But the color aesthetics is unreadable or very clumsy when the data set is large.

5. Read ?facet_wrap. What does nrow do? What does ncol do? What other options control the layout of the individual panels? Why doesn't facet_grid() have nrow and ncol variables?

nrow : number of rows for the facet. ncol : number of columns for the facet.

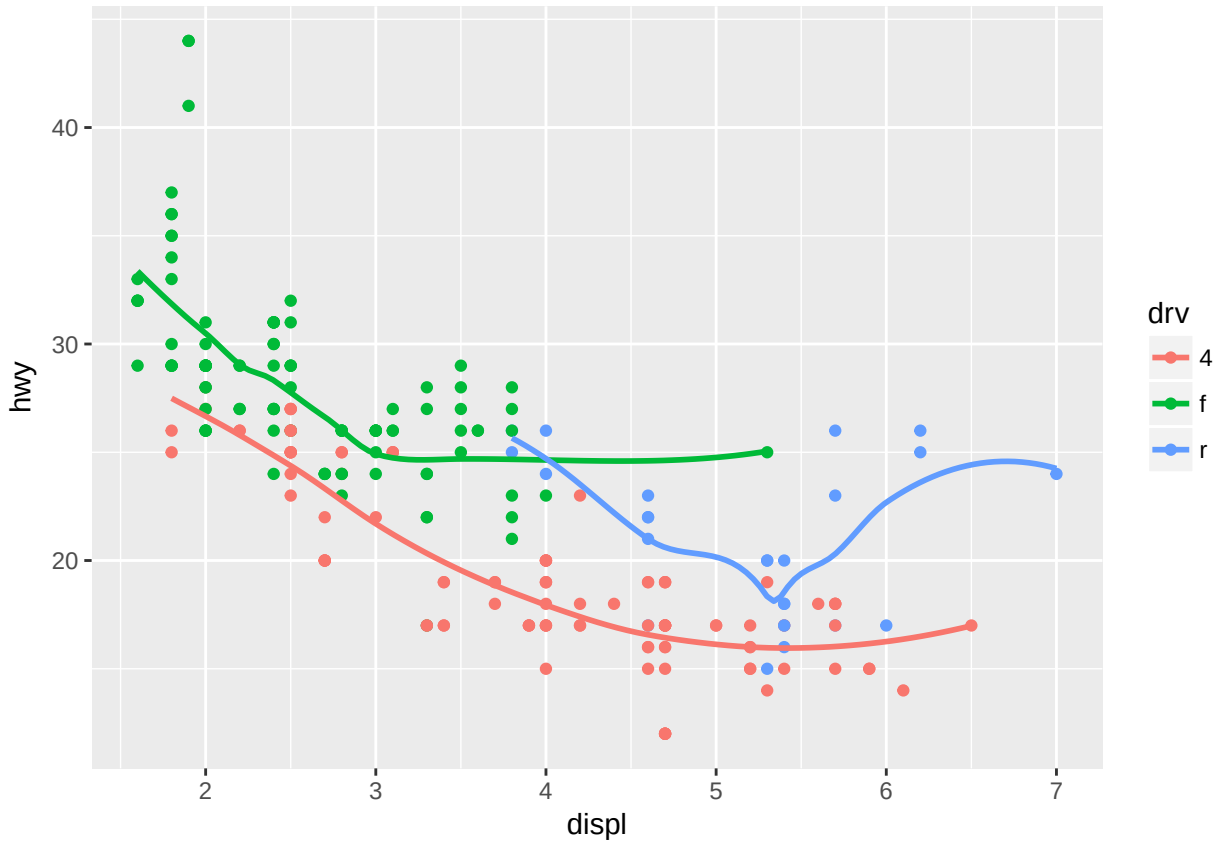
Section 3.6.1: #1-5. Extra Credit: Do #6

1. What geom would you use to draw a line chart? A boxplot? A histogram? An area chart?

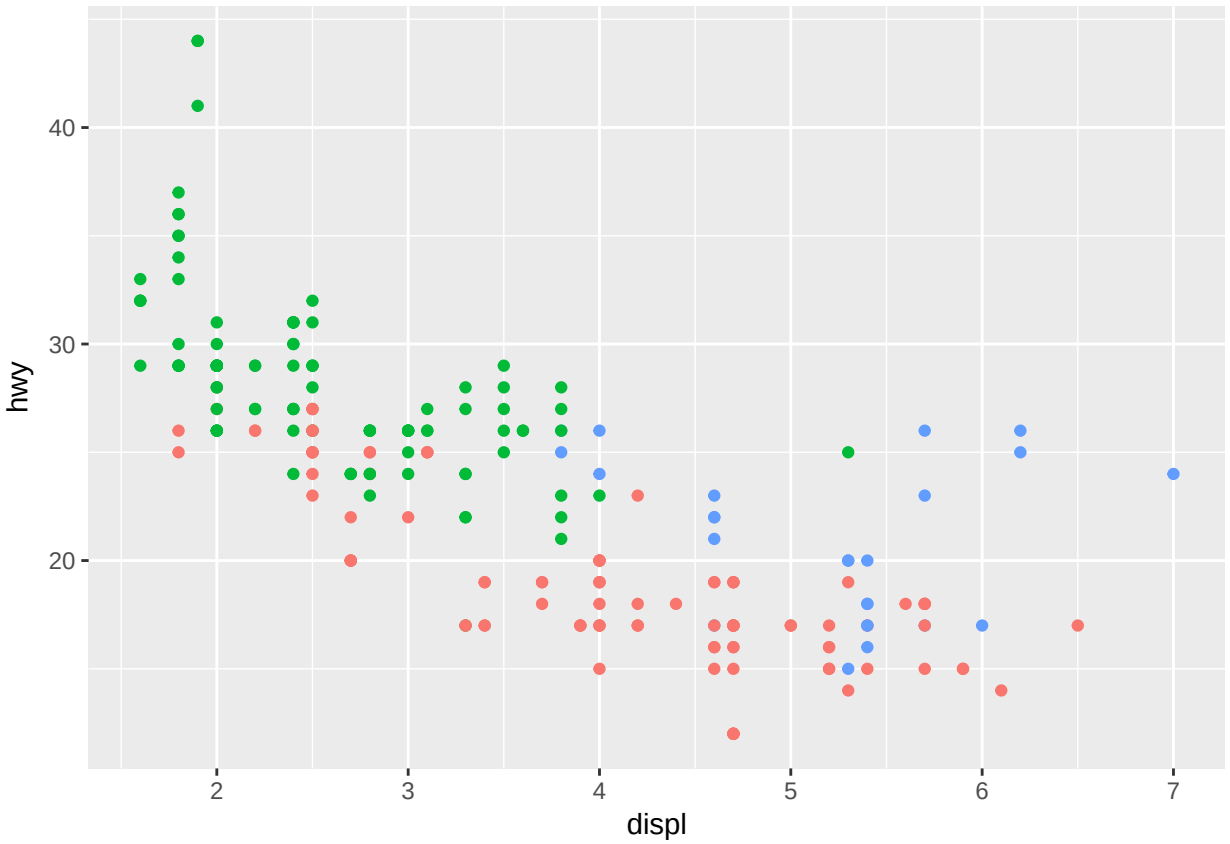
chart - geom_line A boxplot - geom_boxplot A histogram - geom_histogram area chart -geom_area

```
ggplot(data = mpg, mapping = aes(x = displ, y = hwy, color = drv)) +
  geom_point() +
  geom_smooth(se = FALSE)
```

```
## `geom_smooth()` using method = 'loess'
```




```
ggplot(data = mpg, mapping = aes(x = displ, y = hwy, color = drv))+geom_point(show.legend = FALSE)
```



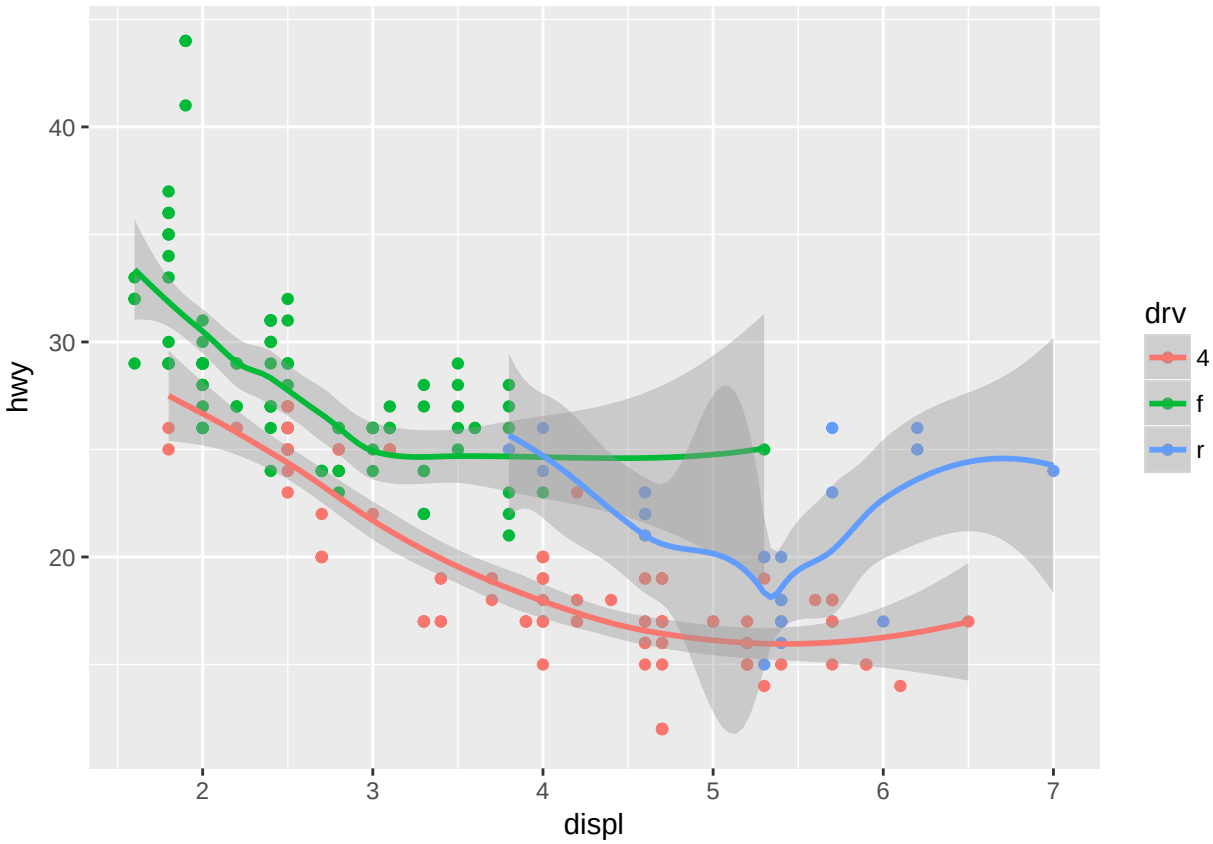
3. What does show.legend = FALSE do? What happens if you remove it? Why do you think I used it earlier in the chapter?

In the above example I have used show.legend = FALSE. This command removes the legend from the chart though the points are colored based on the Legend.

4. What does the se argument to geom_smooth() do? It helps in seeing patterns in when there is a lot of Data.

```
ggplot(data = mpg, mapping = aes(x = displ, y = hwy, color = drv)) +
  geom_point() +
  geom_smooth(se = TRUE)
```

```
## `geom_smooth()` using method = 'loess'
```



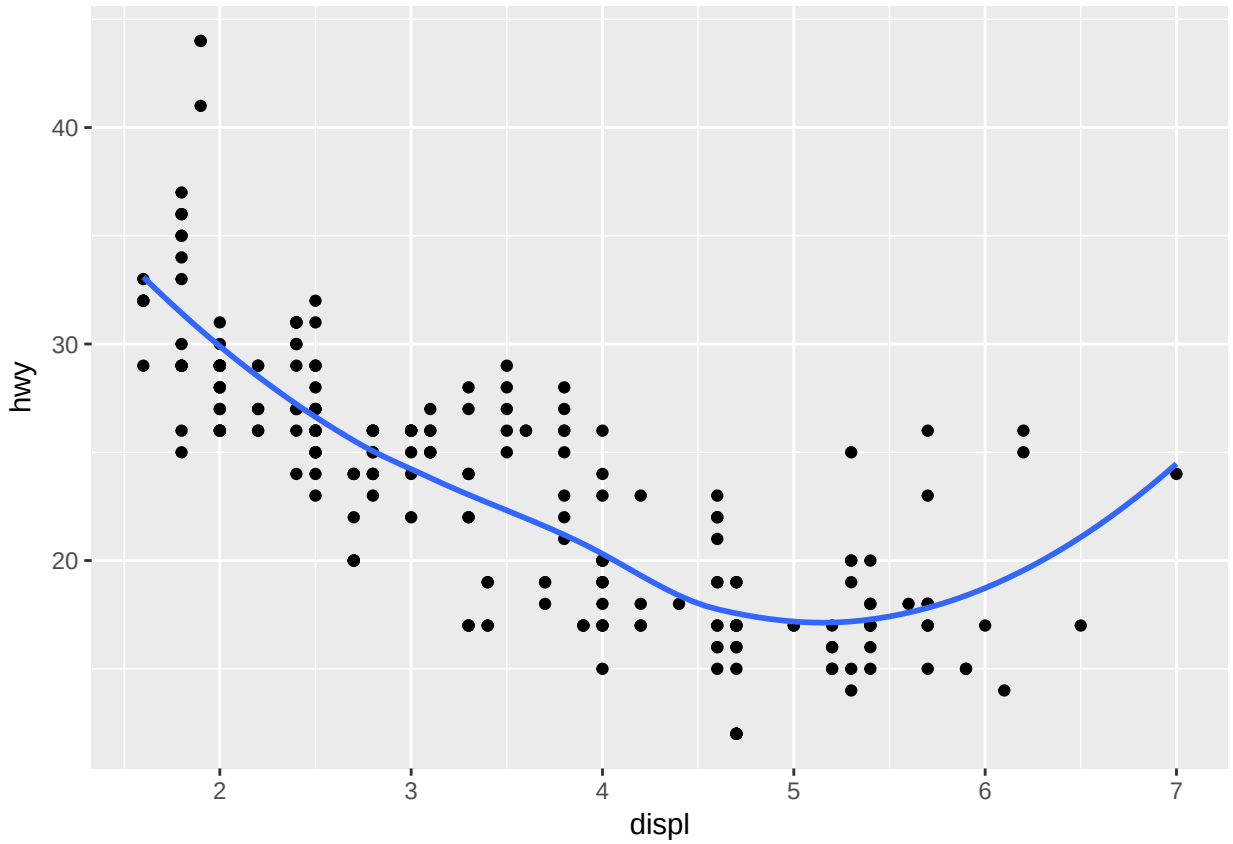
5. Will these two graphs look different? Why/why not?

There will be no difference between these 2 graphs as the mapping setting is used in ggplot() funtion is inherited by the following 2 other function.

6. Recreate the R code necessary to generate the following graphs.

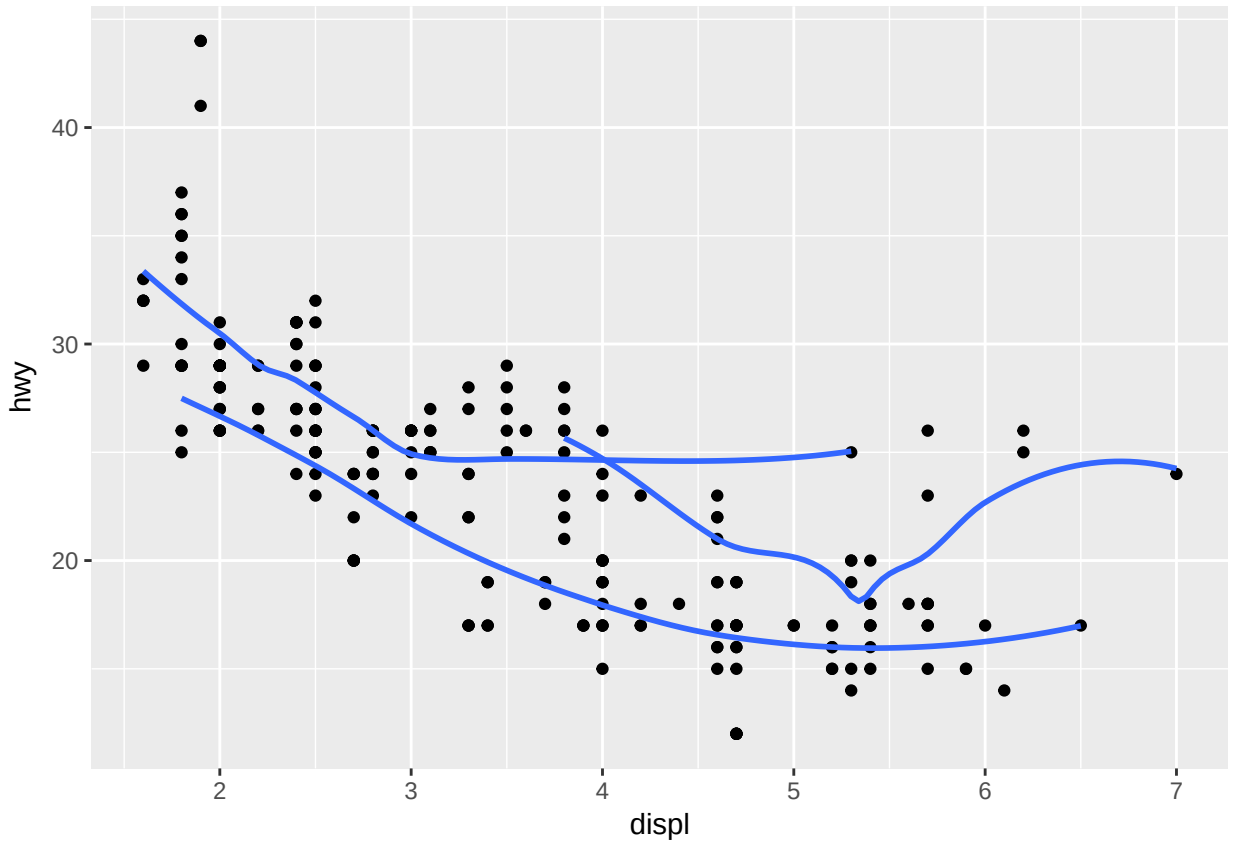
```
ggplot(data = mpg, mapping = aes(x = displ, y = hwy)) + geom_point() + geom_smooth(se = FALSE)
```

```
## `geom_smooth()` using method = 'loess'
```



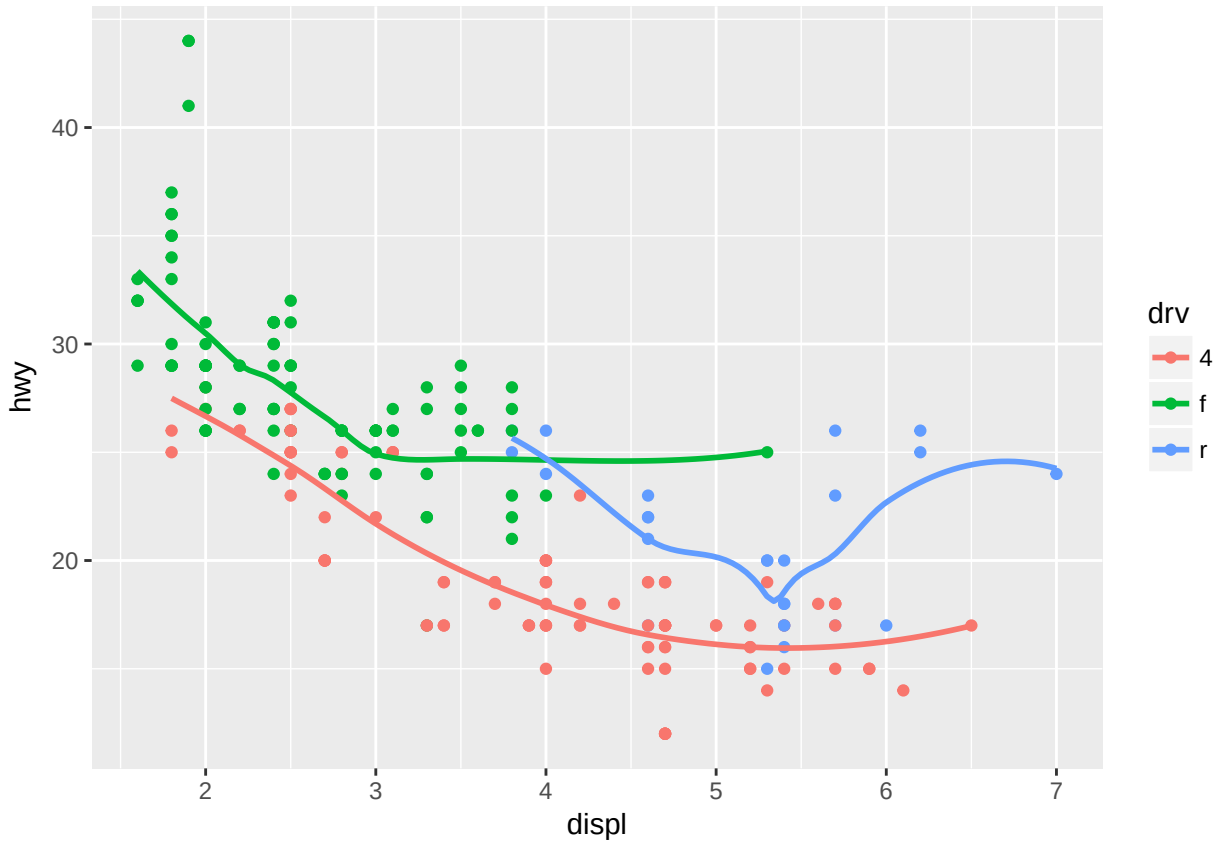

```
ggplot(data = mpg, mapping = aes(x = displ, y = hwy)) +
  geom_point() +
  geom_smooth(aes(group = drv),se = FALSE)
```

```
## `geom_smooth()` using method = 'loess'
```



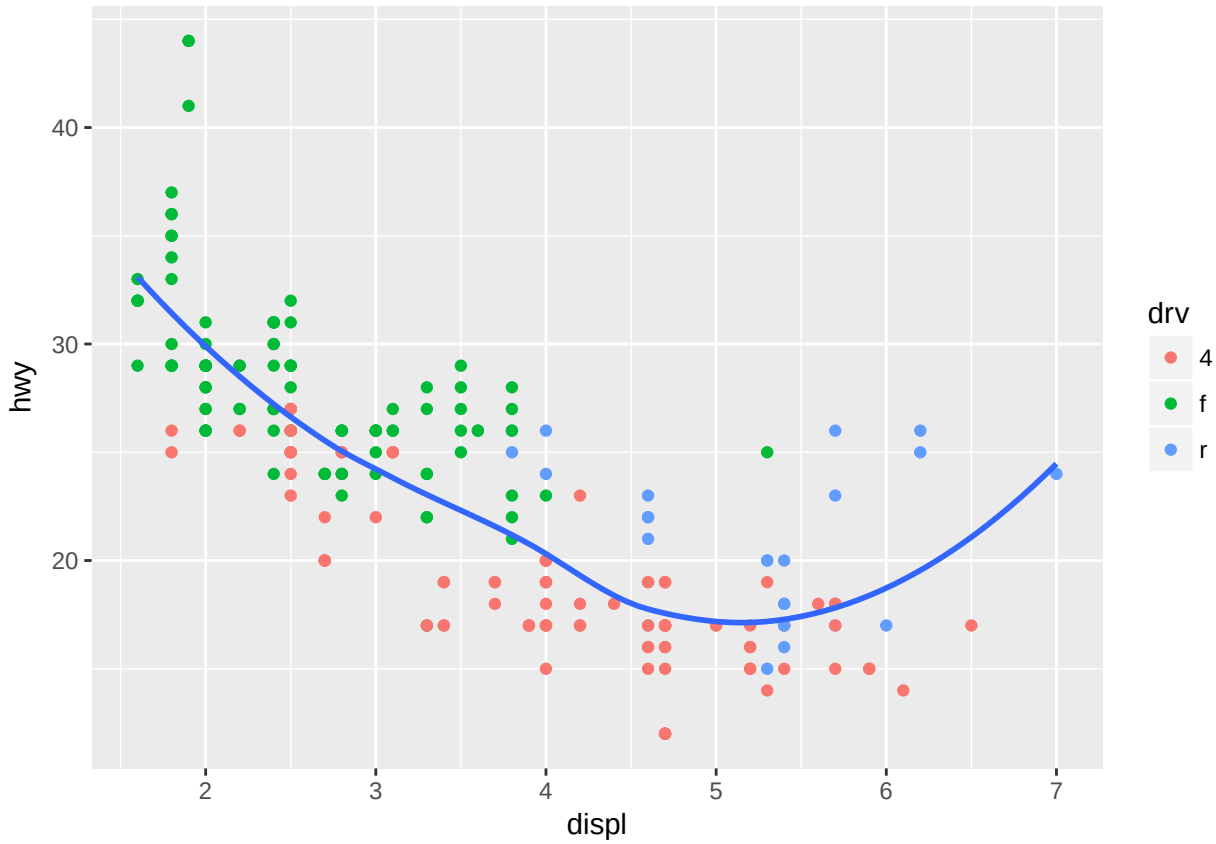

```
ggplot(data = mpg, mapping = aes(x = displ, y = hwy, color = drv)) +
  geom_point() +
  geom_smooth(aes(group = drv),se = FALSE)
```

```
## `geom_smooth()` using method = 'loess'
```

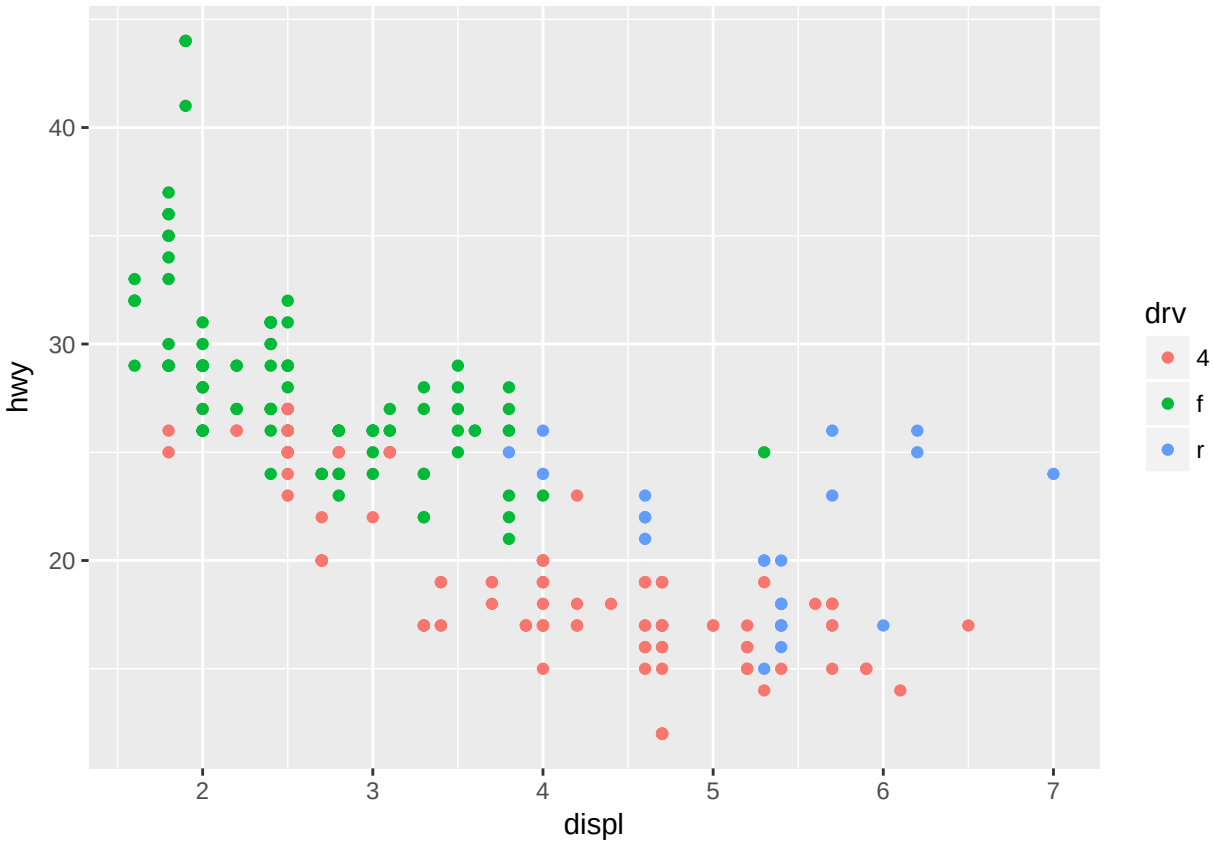



```
ggplot(data = mpg, mapping = aes(x = displ, y = hwy)) +
  geom_point(aes(color = drv)) +
  geom_smooth(se = FALSE)
```

```
## `geom_smooth()` using method = 'loess'
```




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

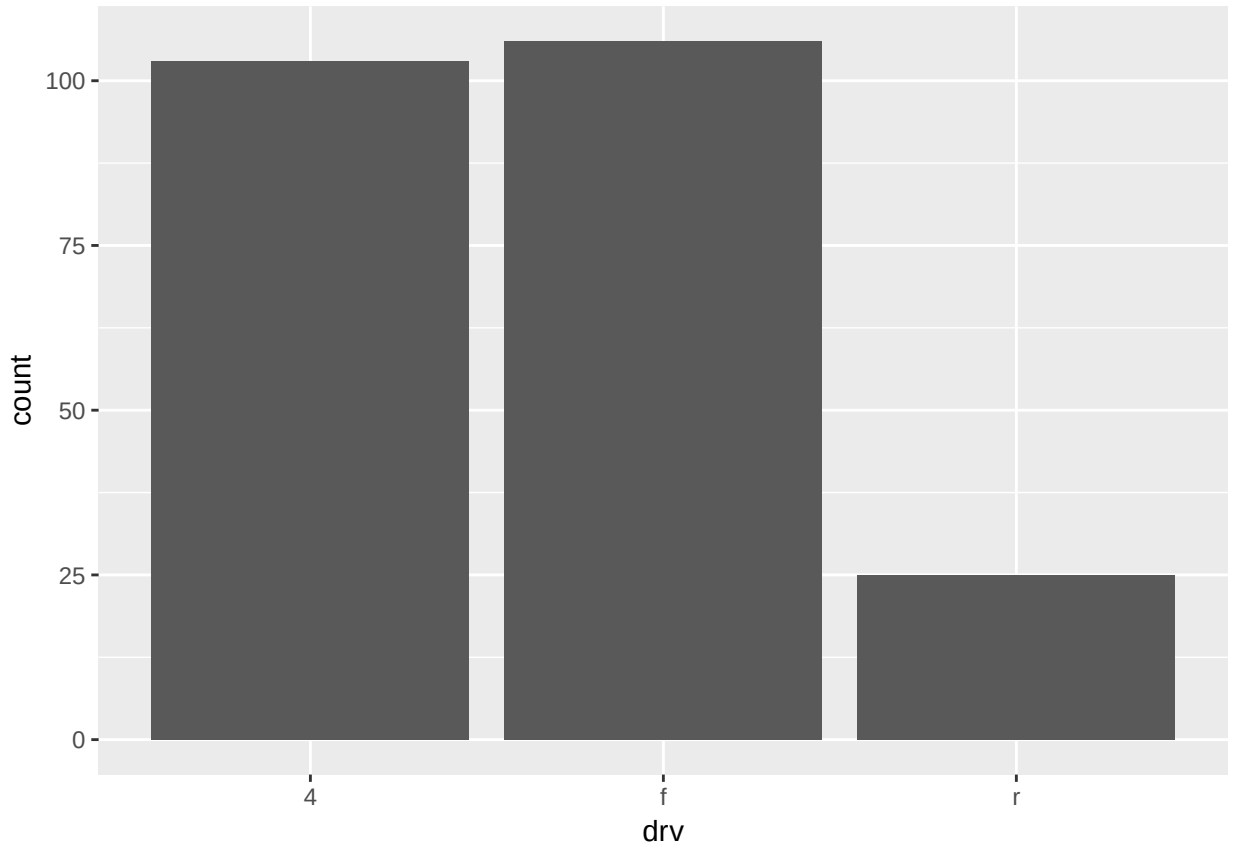


Section 3.7.1: #2 only

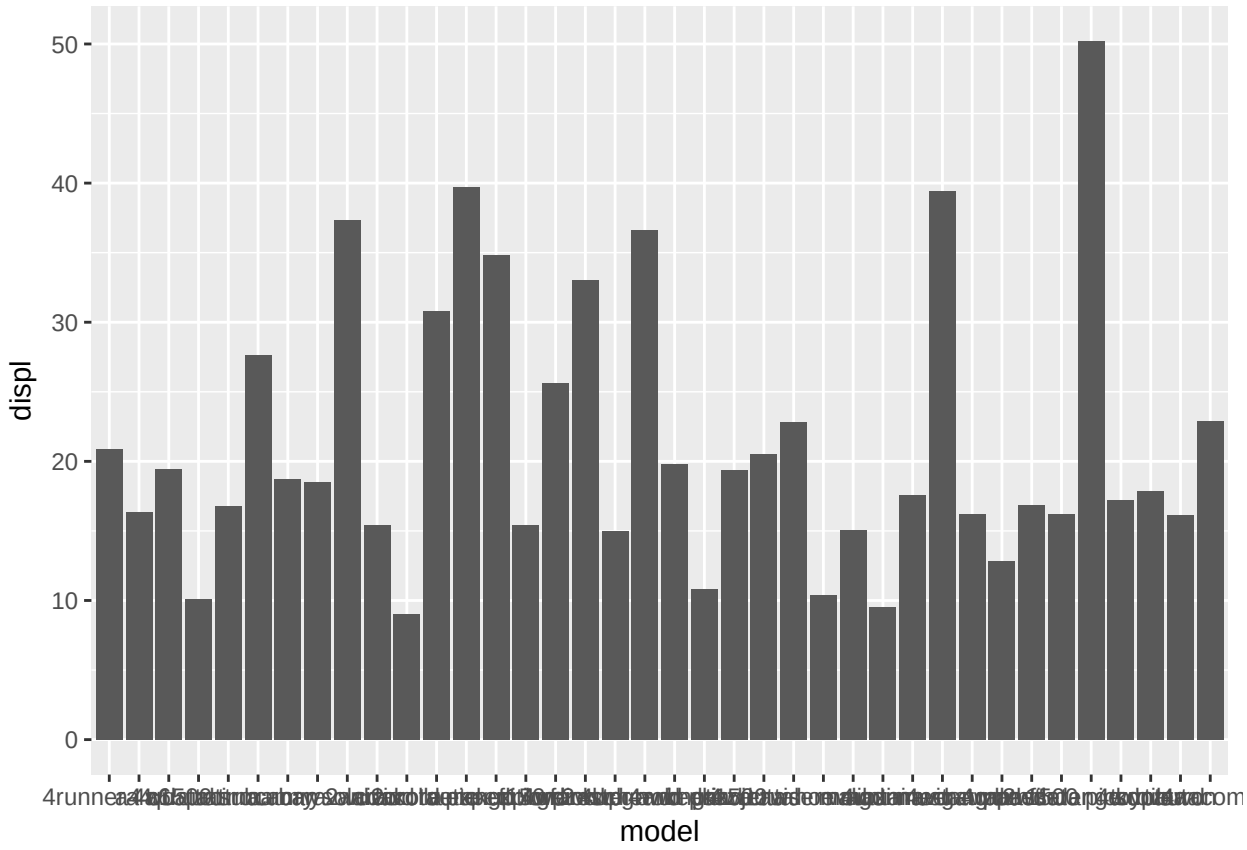
geom_bar makes the height of the bar proportional to the number of cases in each group. This uses the static_count. Passes only 1 variable.

geom_col represents the values in the data. This passes 2 variable.

```
ggplot(data = mpg) +
geom_bar( mapping = aes(x = drv))
```




```
ggplot(data = mpg) +
geom_col( mapping = aes(x = model, y = displ))
```


50
40
30
20
10
0
displ
4runner
model