COMPSCIX 415.2 Homework 2

Ganesh Saravanan

6/18/2018

My Github repository for my assignments can be found at this URL:https://github.com/gsaravanan1/rstudiodemo.git

```
library(mdsr)
library(tidyverse)
library(ggplot2)
```

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

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


ANSWER: Nothing

2. How many rows are in mpg? How many columns?

```
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...
```

Rows : 234 , Columns :11

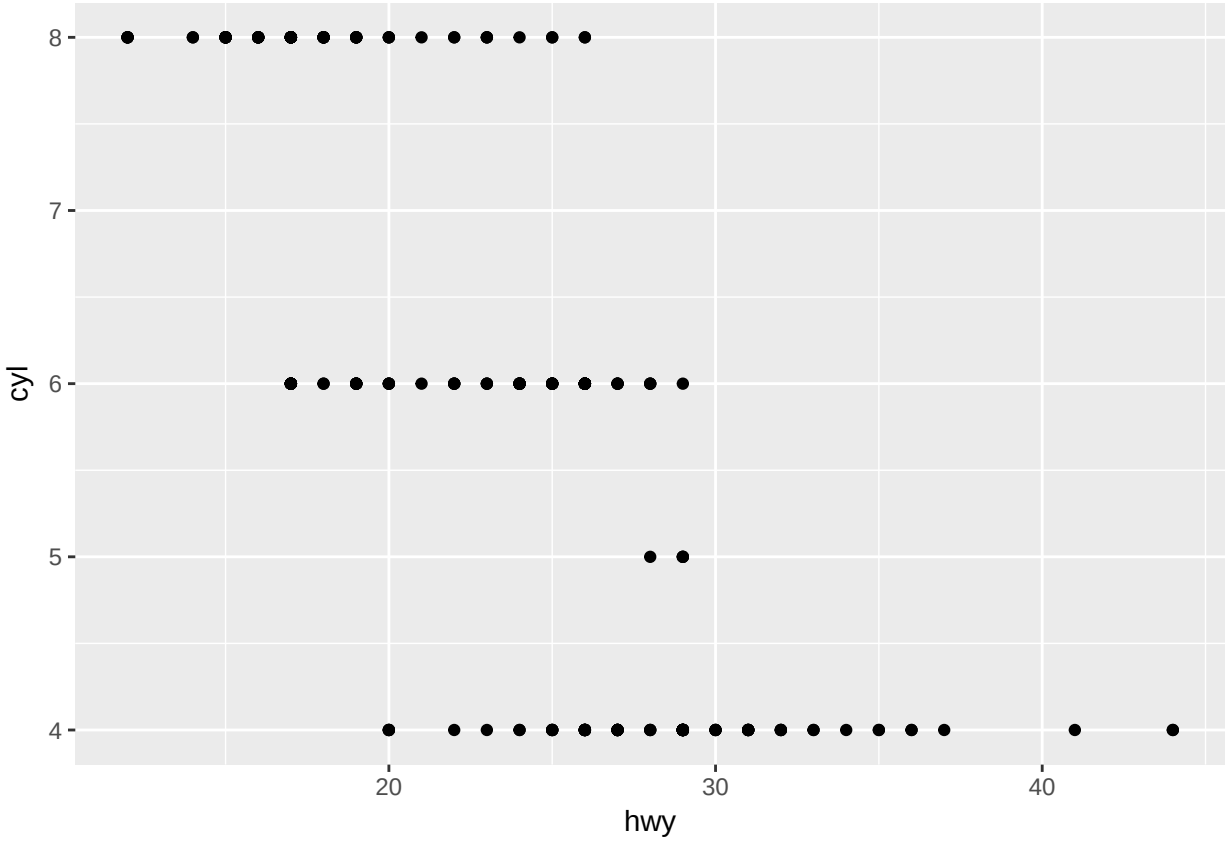
3. What does the drv variable describe? Read the help for ?mpg to find out.

```
?mpg
```

ANSWER: drv variable describes the vehicle wheel drive. values: f = front-wheel drive, r = rear wheel drive, 4 = 4wd

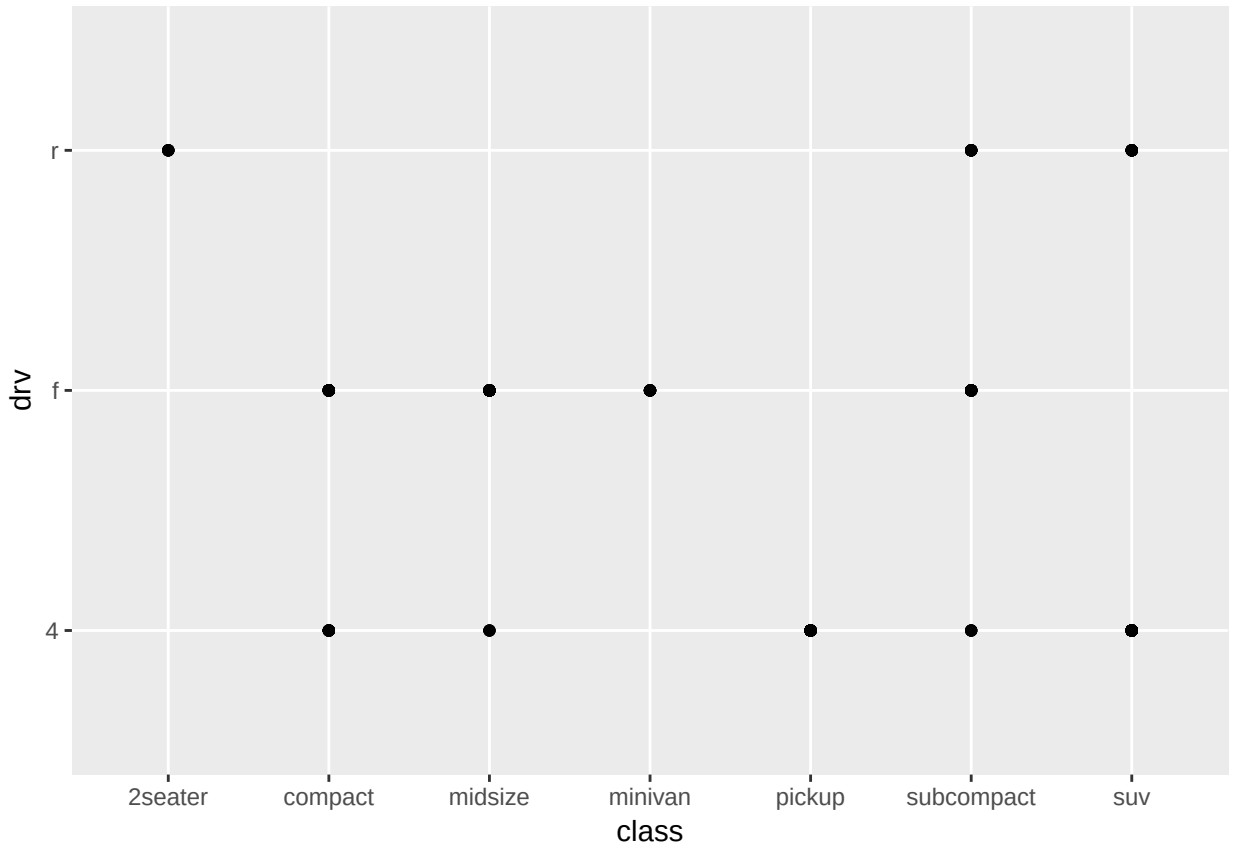
4. Make a scatterplot of hwy vs cyl.

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



5. 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))
```

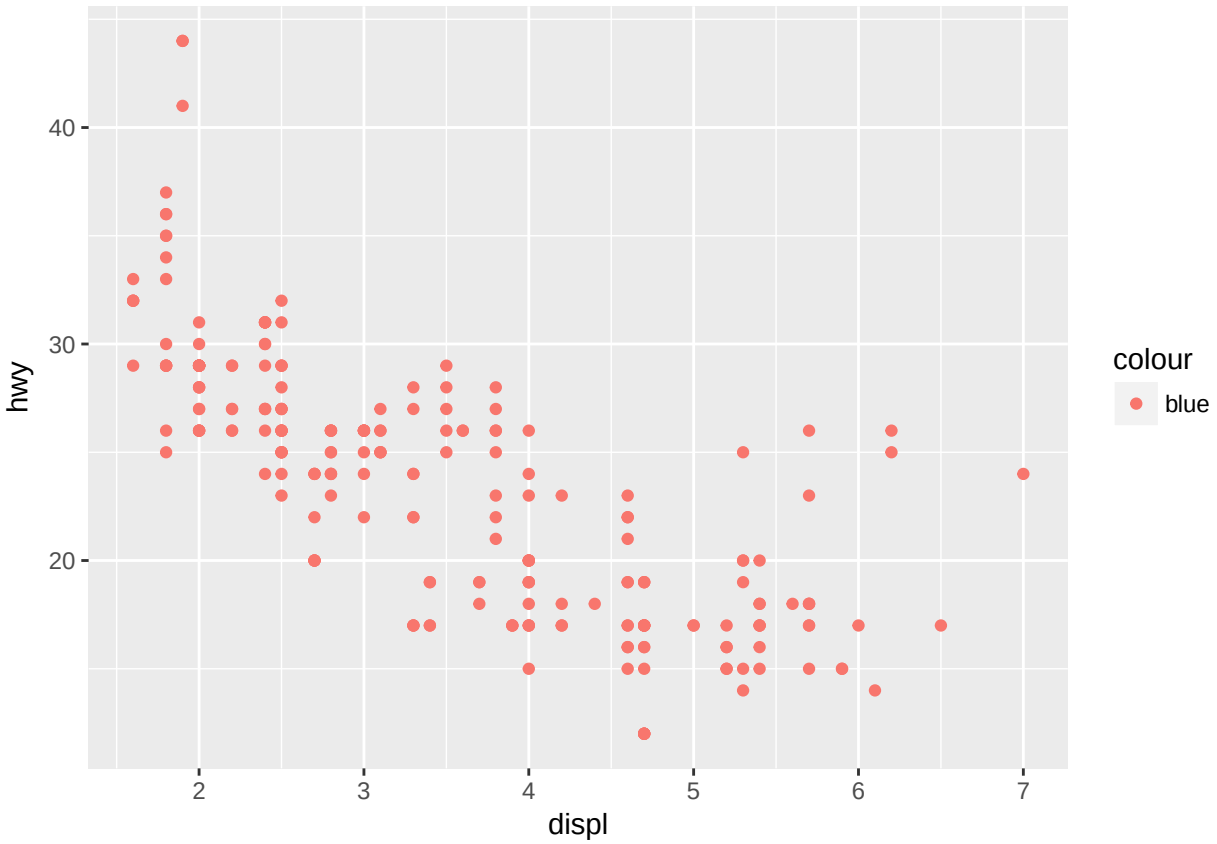


ANSWER : The plot is not useful, because ,both the fields are categorical.

3.3.1 Exercises

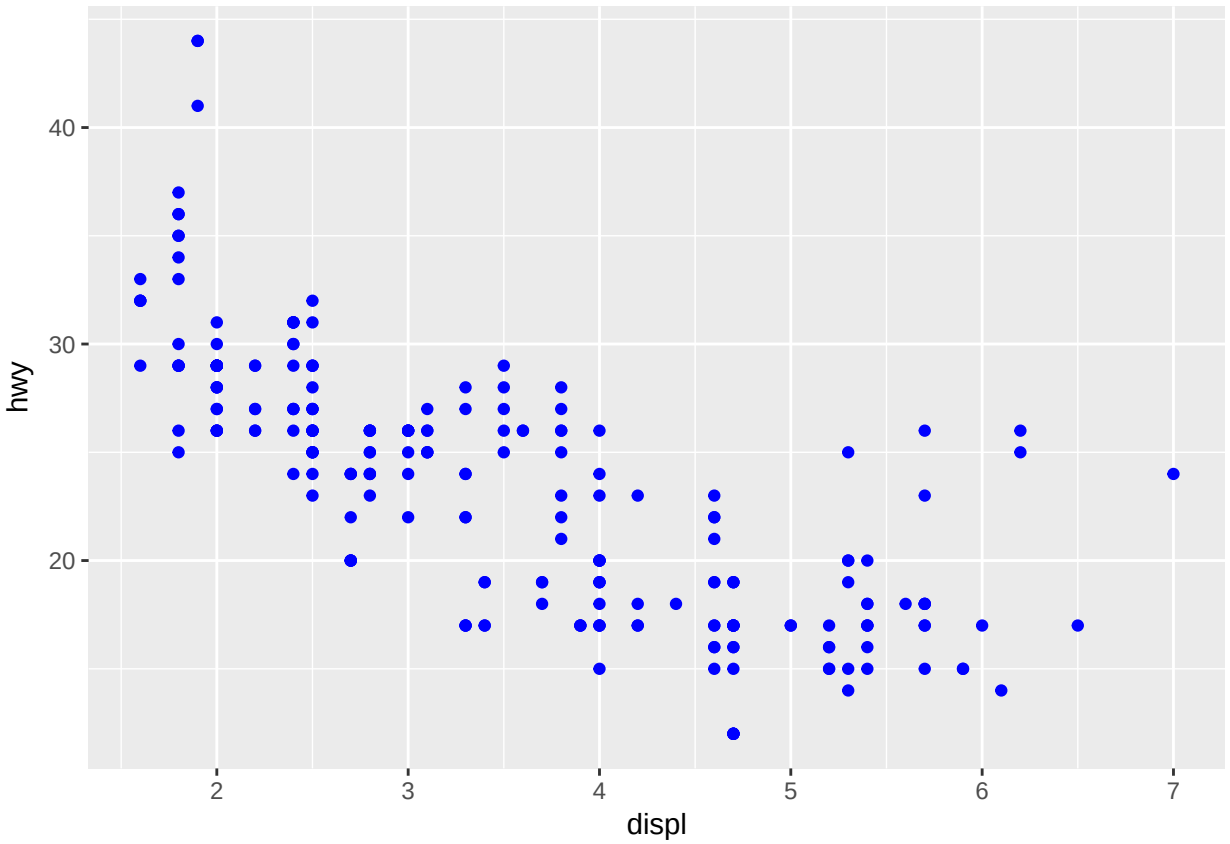
1. What's gone wrong with this code? Why are the points not blue?

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



Answer: color argument is to set outside of aes() function.

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

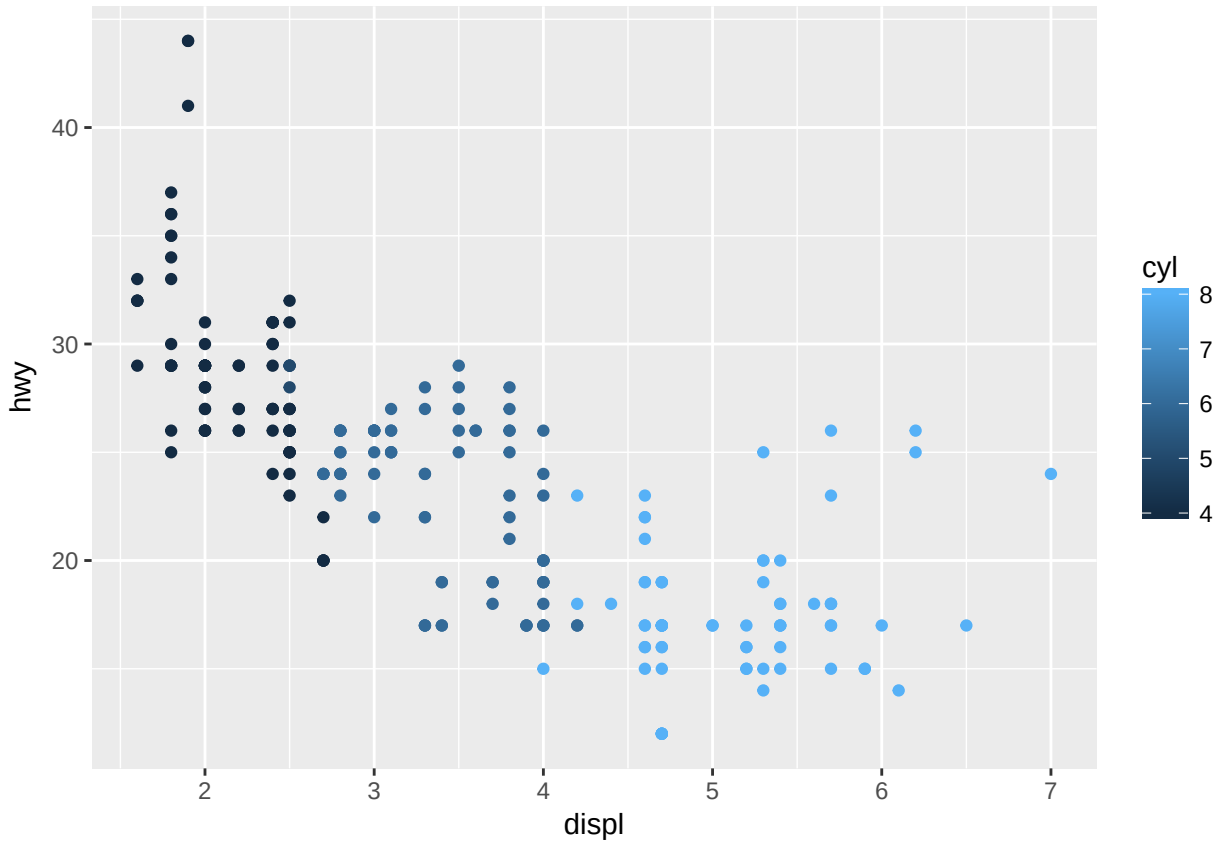


2. Which variables in mpg are categorical? Which variables are continuous? Categorical: manufacturer,model,trans,drv, fl, class Continuous: displ, year,cyl,cty, hwy

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

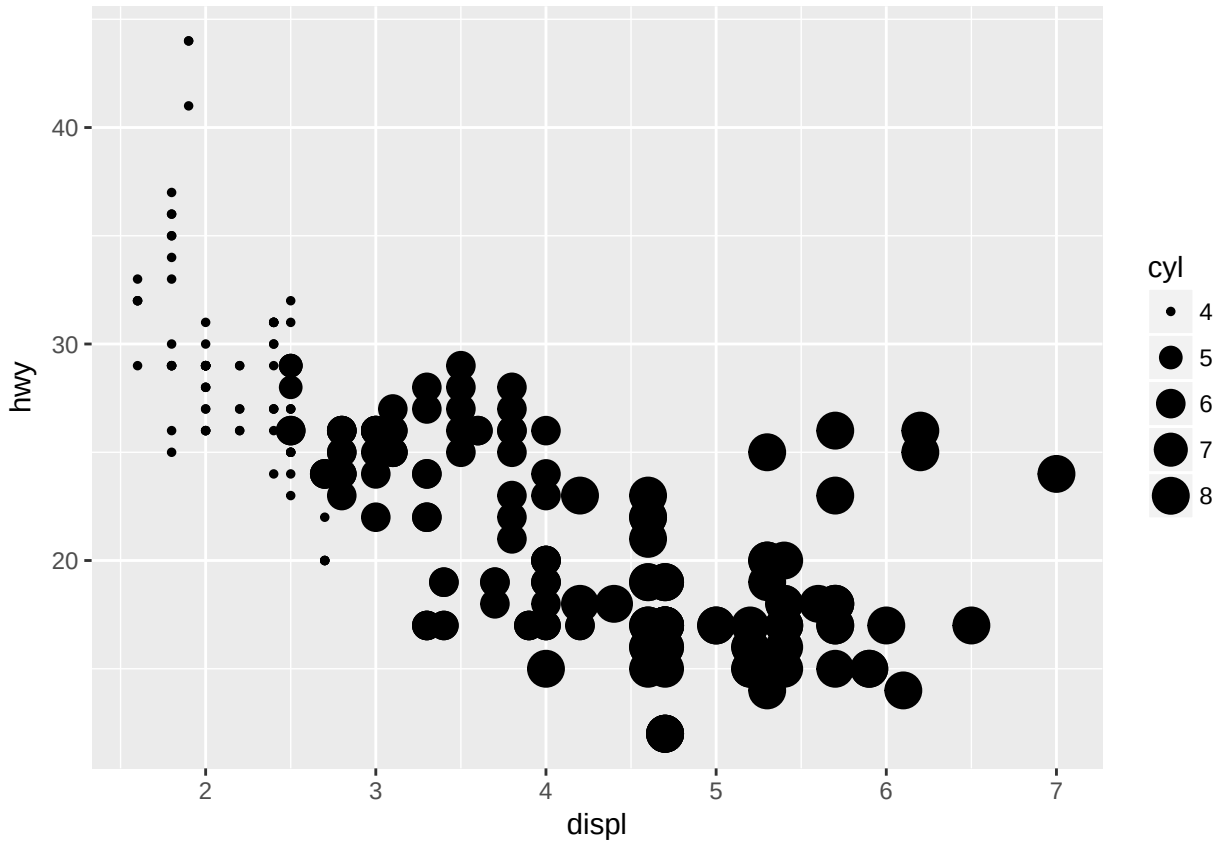
COLOR:

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



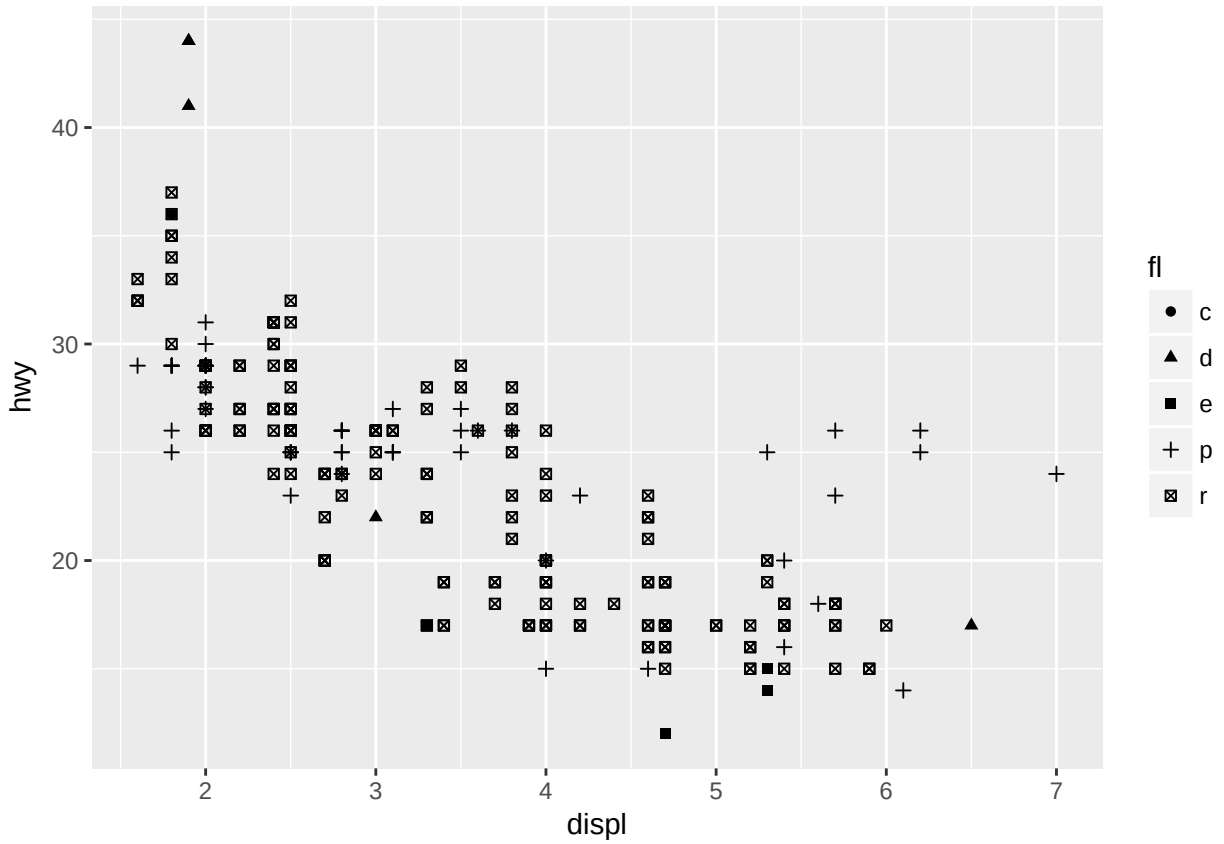
SIZE:

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



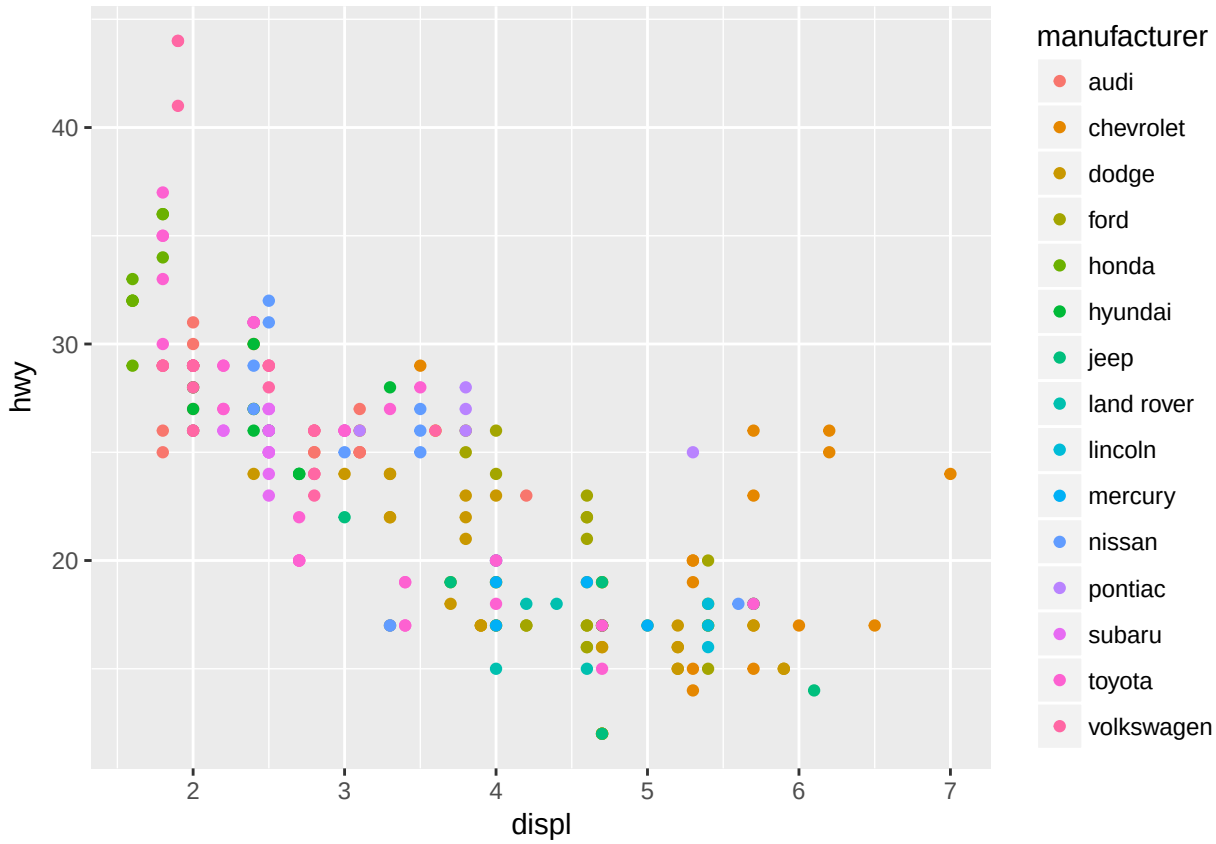
SHAPE:

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



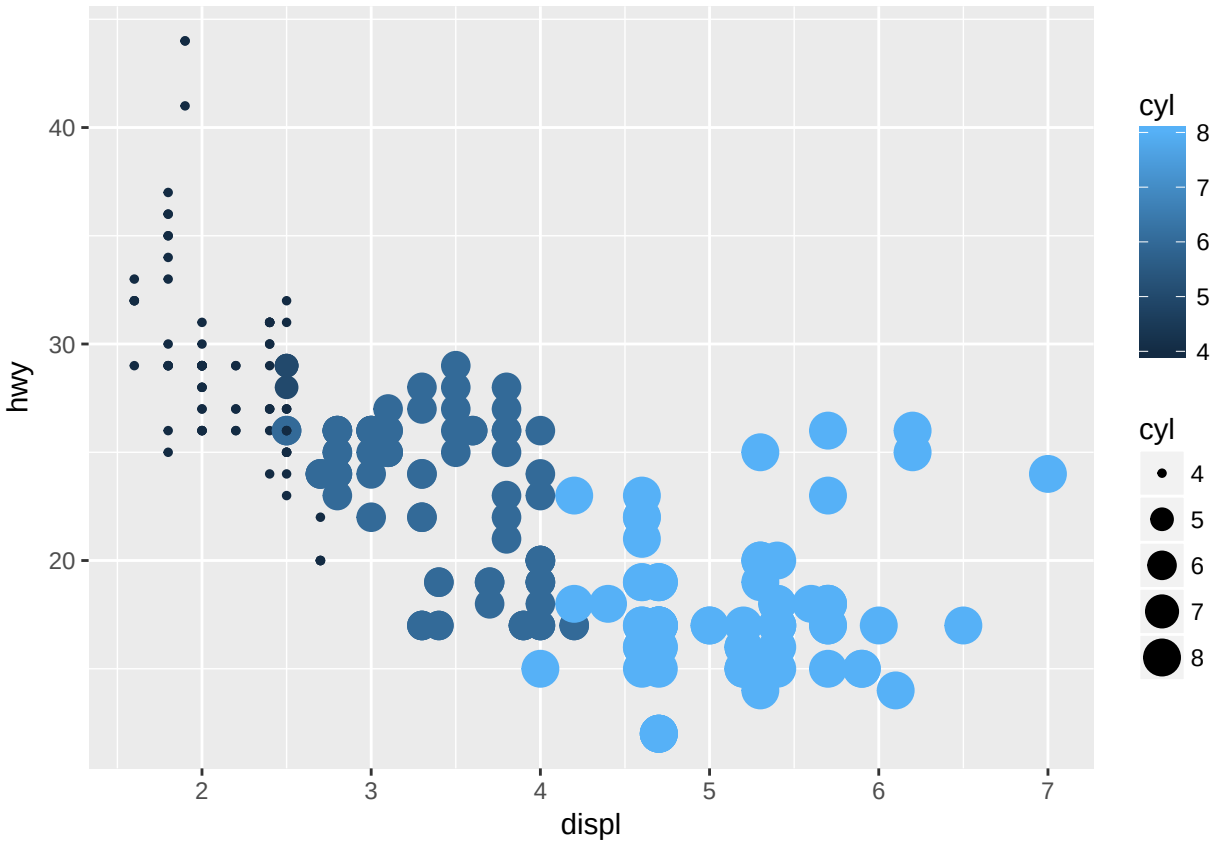
categorical Varianble : manufacturer represents each vaule with a unique color unike the continuous value with one color just the variation on the spectrum.

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



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

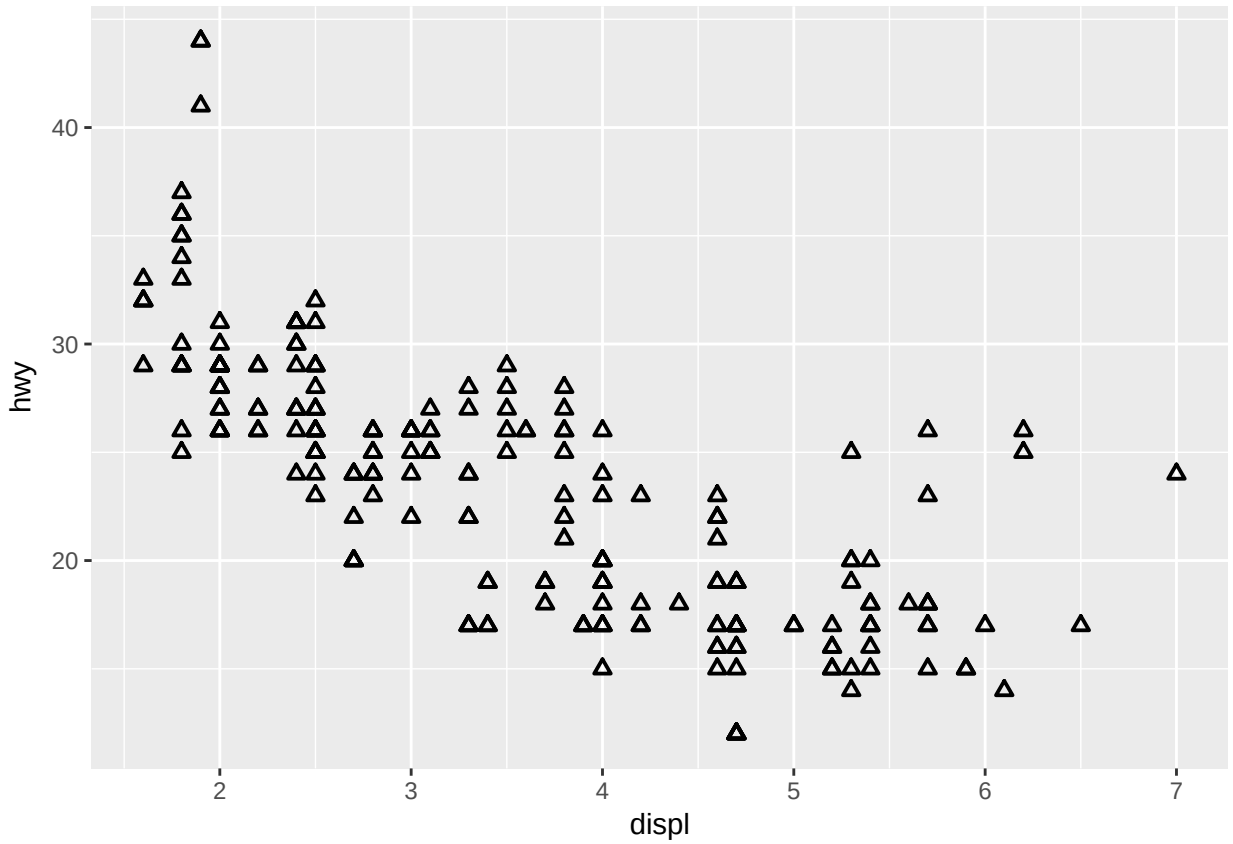
```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy, color = cyl, size=cyl))
```



Seperated by two legends.

4. What does the stroke aesthetic do? What shapes does it work with?

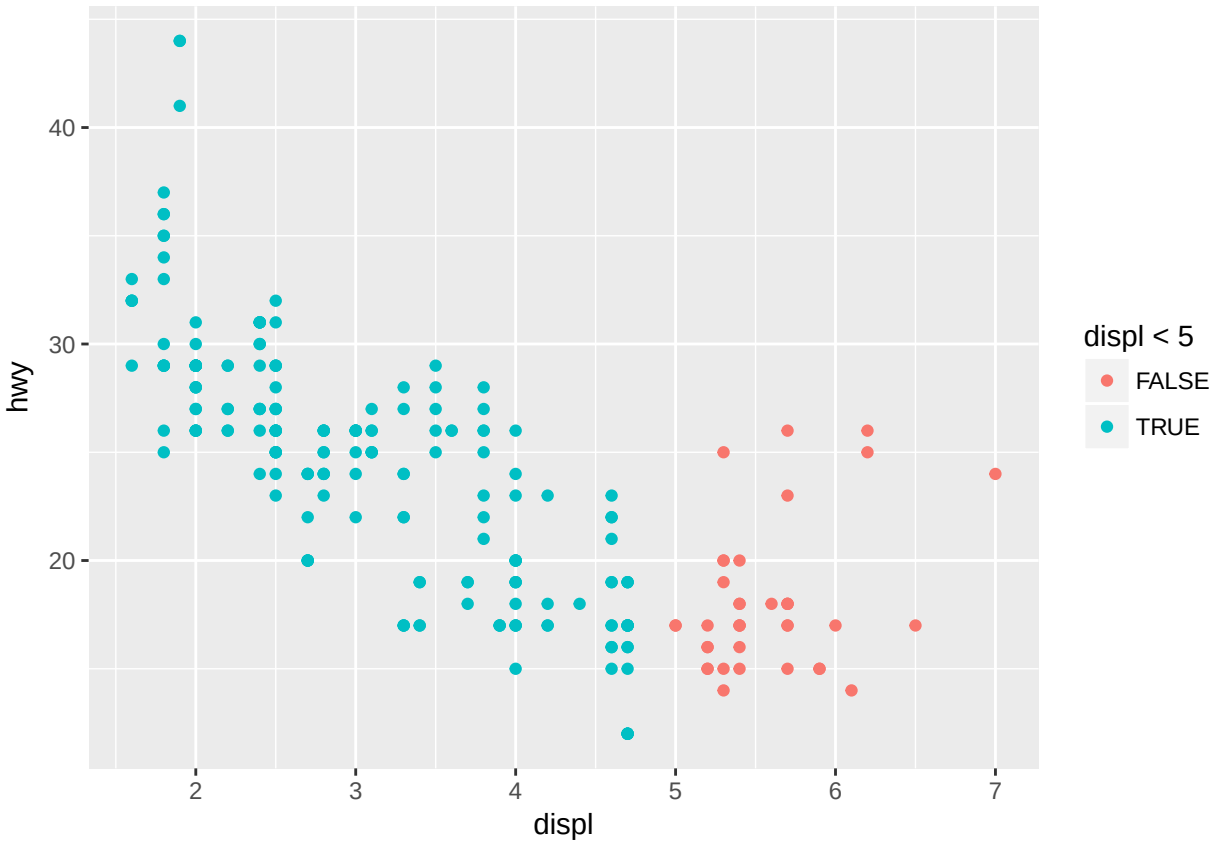
```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy), stroke = 1, shape = 24)
```



Stroke goes in hand with shape variable. It determines the thickness of the shape.

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))
```

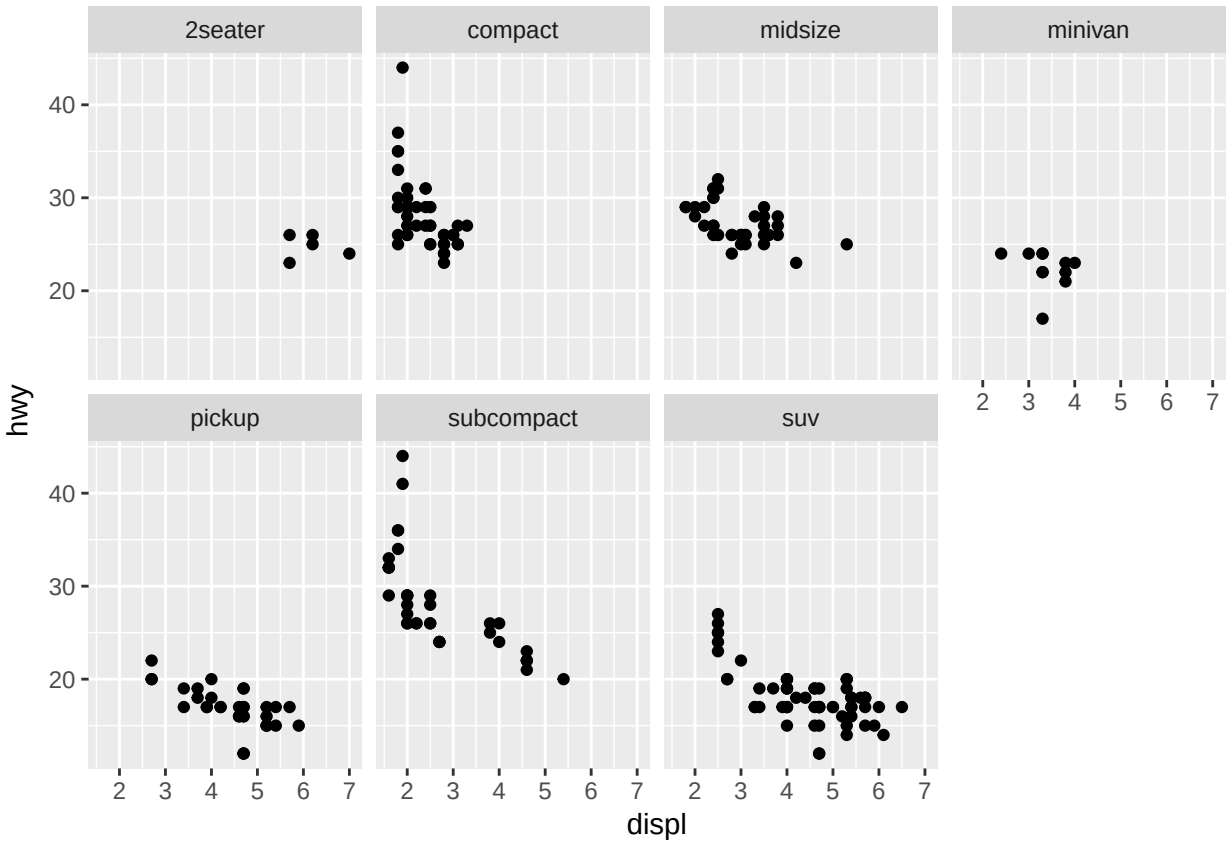


If the variable displ <5, the temp variable is has a flag - TRUE, else the flag - FALSE.

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(~ class, nrow = 2)
```



What are the advantages to using faceting instead of the colour aesthetic? What are the disadvantages? How might the balance change if you had a larger dataset?

The advantage is facet_wrap variable splits data into seperate VIZ's to see the trend in each facet. The disadvantage is the complexity in correlating the facets. With a larger data set, the data points may overlap finding hard to interpret.

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 argument?

nrow - # of rows a facet plot can have ncol - # of columns a facet plot can have shrink - If TRUE, will shrink scales to fit output of statistics, not raw data. If FALSE, will be range of raw data before statistical summary. facet_wrap is an extension to facet_grip with nrow, ncol options and is generally a better use of screen space than facet_grid because most displays are roughly rectangular.

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

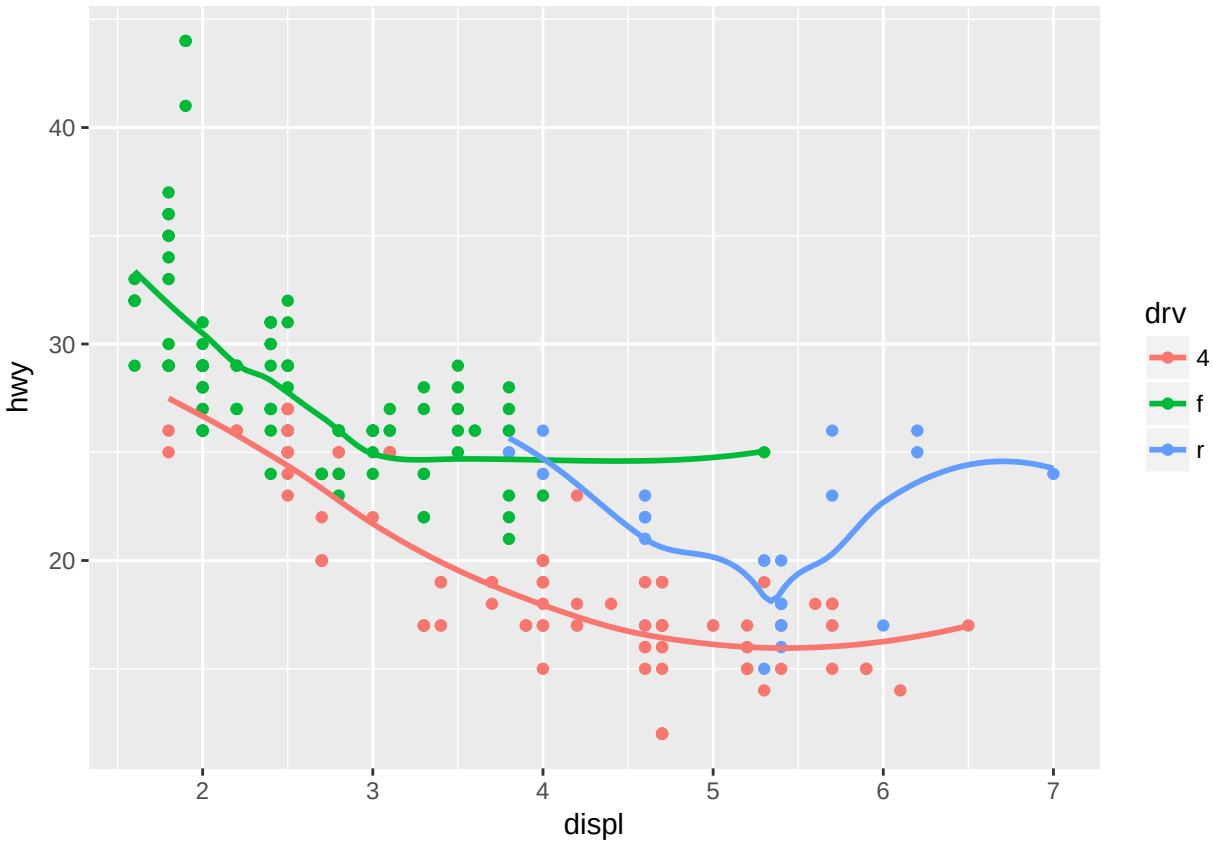
1. What geom would you use to draw a line chart? A boxplot? A histogram? An area chart?

Line chart - geom_line() Boxplot - geom_boxplot() Histogram - geom_histogram() Area chart - geom_area()

2. Run this code in your head and predict what the output will look like. Then, run the code in R and check your predictions.

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

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



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

show.legend = FALSE option removes the legend. Data is still plotted, but the key is removed.

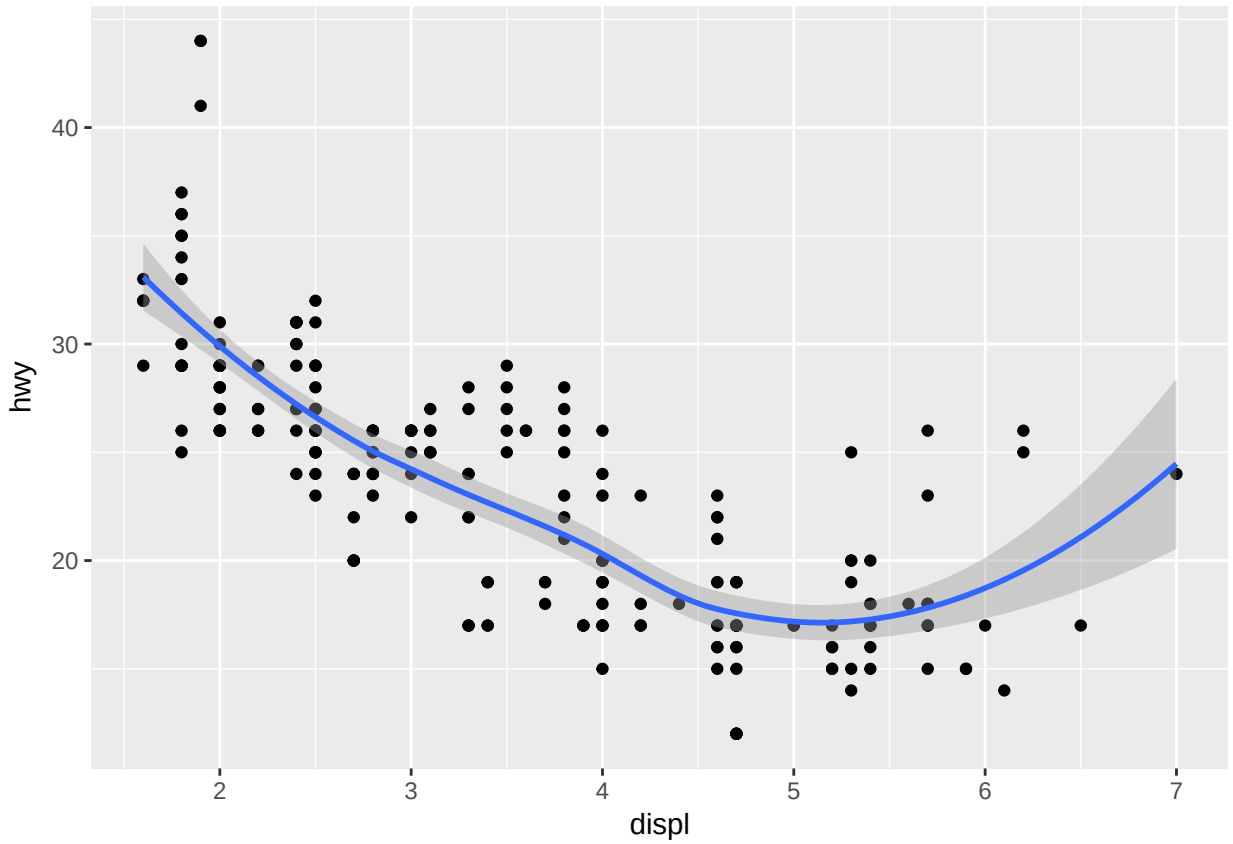
4. What does the se argument to geom_smooth() do?

se - display confidence interval around smooth? (TRUE by default)

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

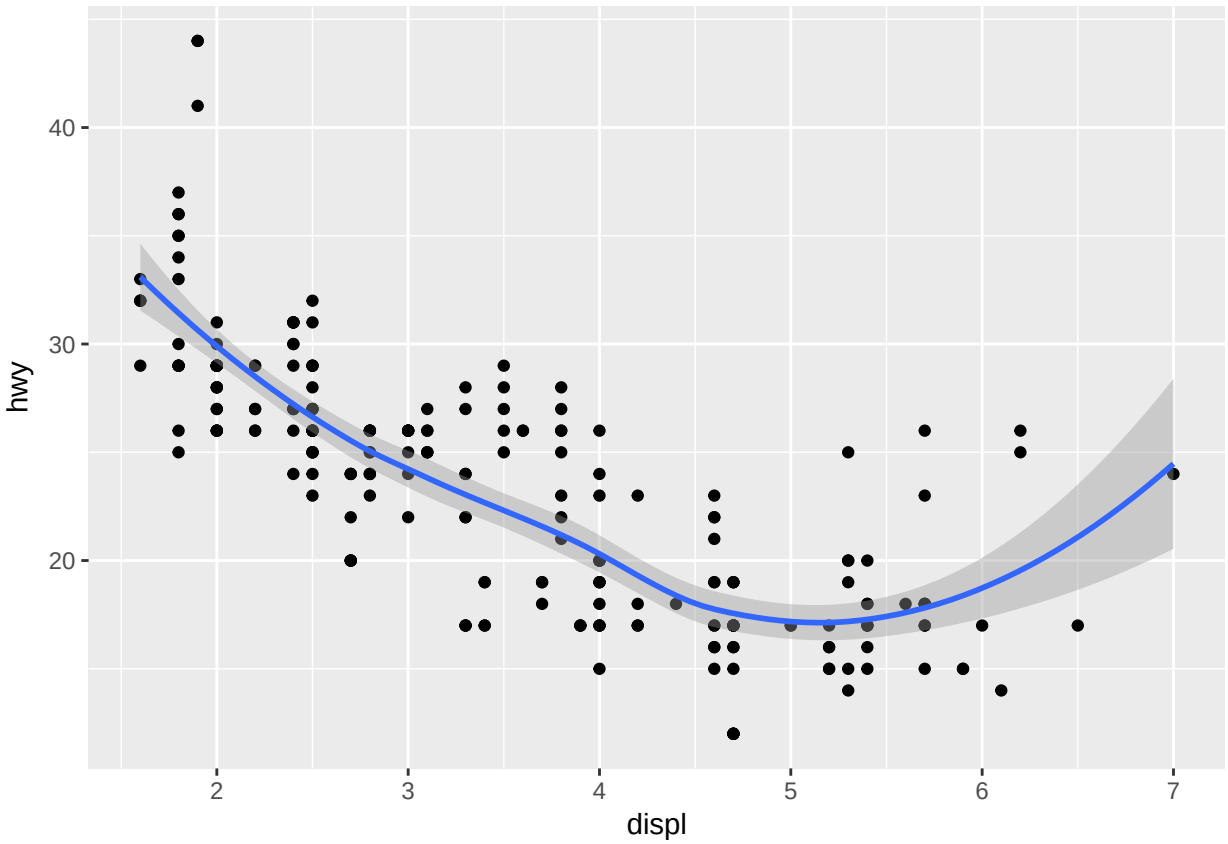
```
ggplot(data = mpg, mapping = aes(x = displ, y = hwy)) +
  geom_point() +
  geom_smooth()
```

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




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

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

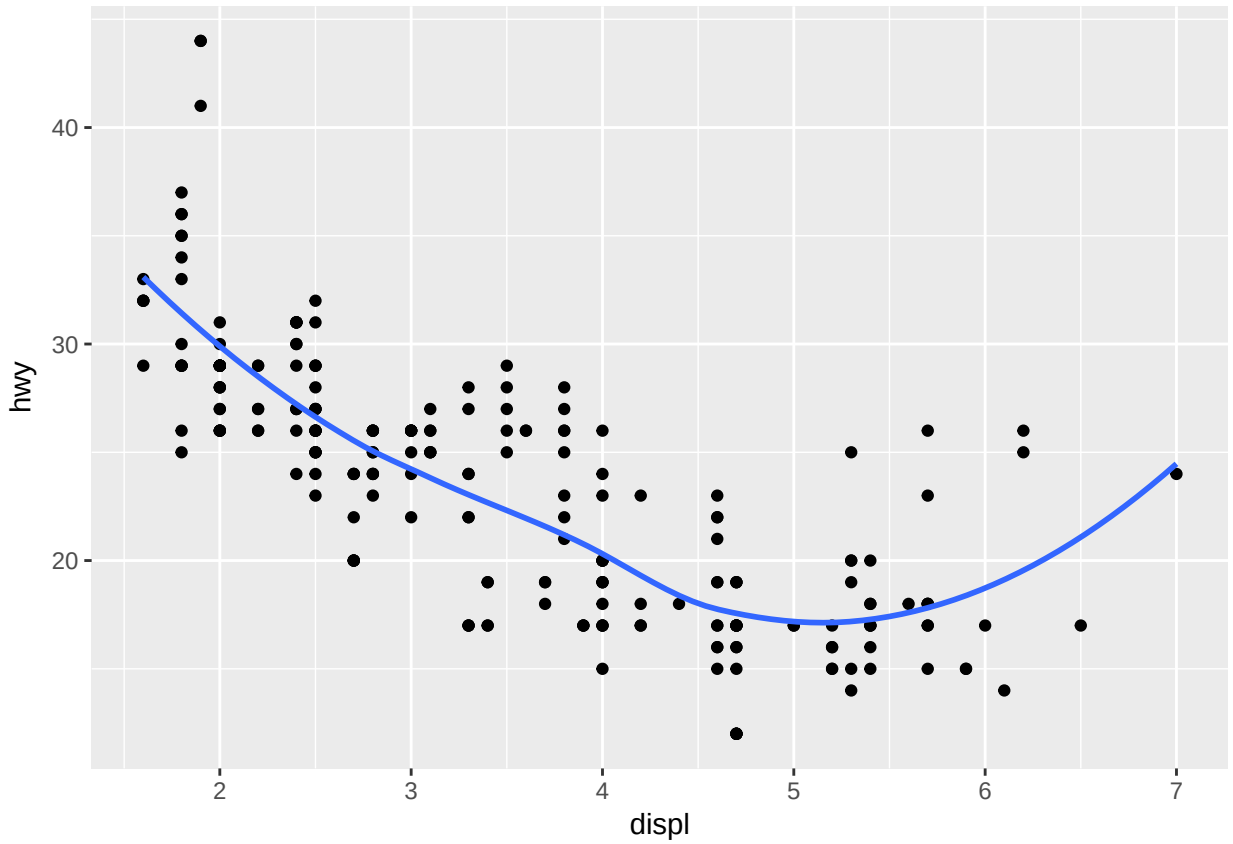


NO. The settings are the same. Difference is in the ggplot() functions syntax.

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

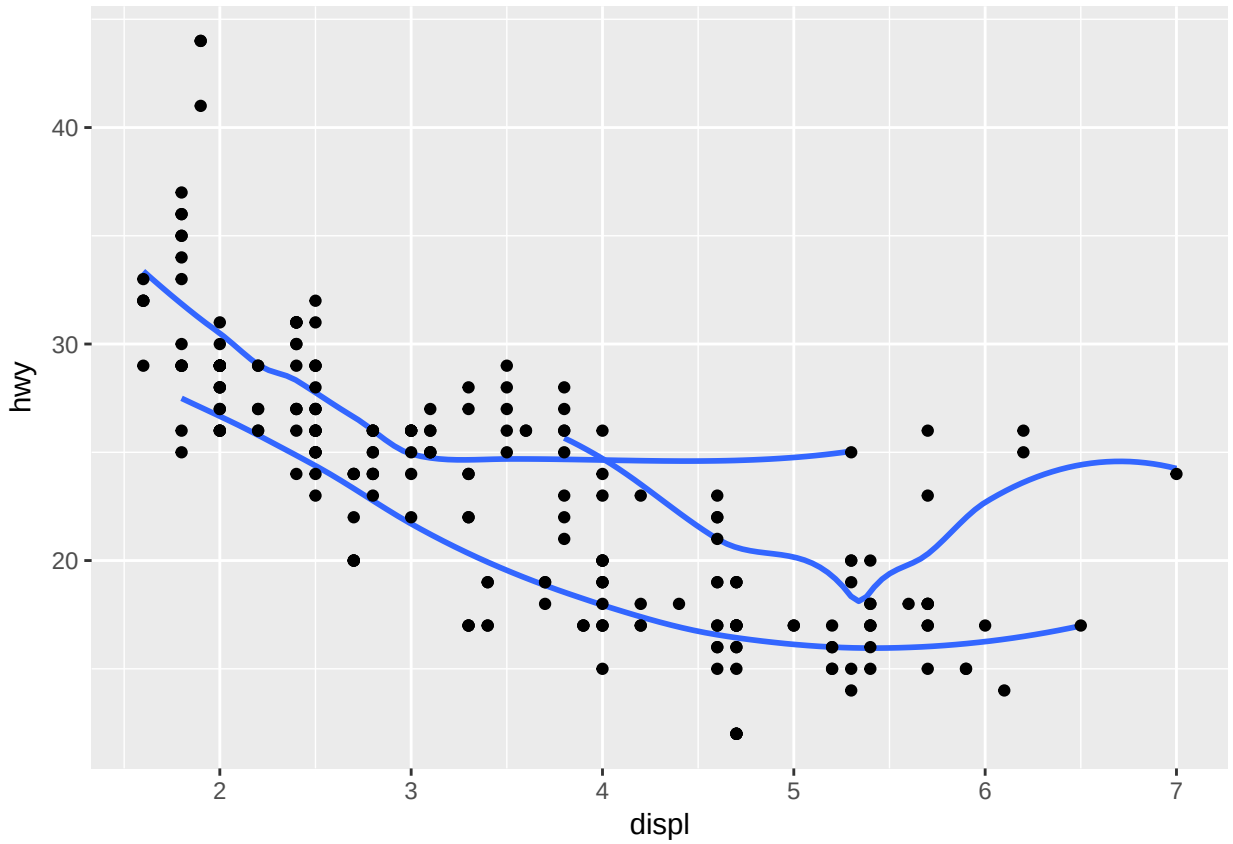
```
ggplot(data = mpg, mapping = aes(x = displ, y = hwy)) +
  geom_point() +
  geom_smooth(se = FALSE)
```

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



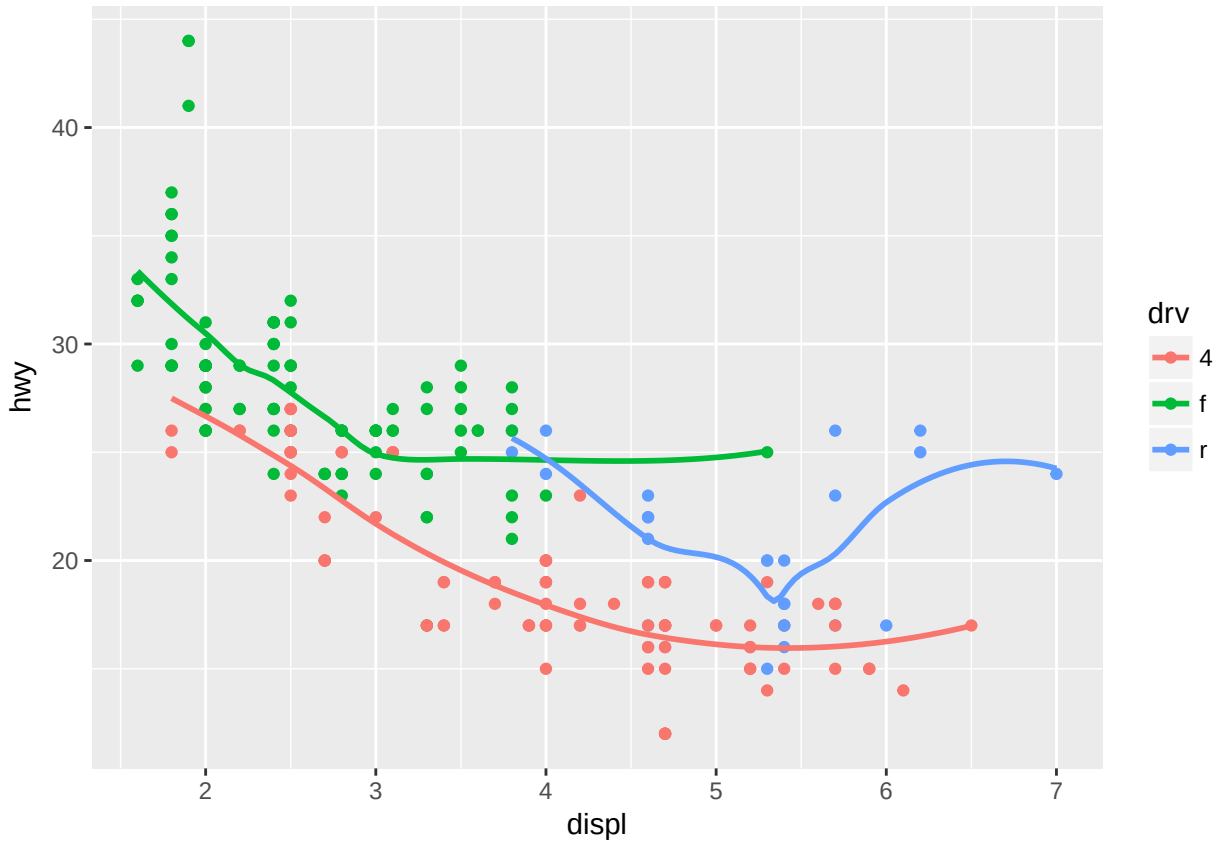

```
ggplot(data = mpg, mapping = aes(x = displ, y = hwy)) +
  geom_smooth(aes(group = drv), se = FALSE) +
  geom_point()
```

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



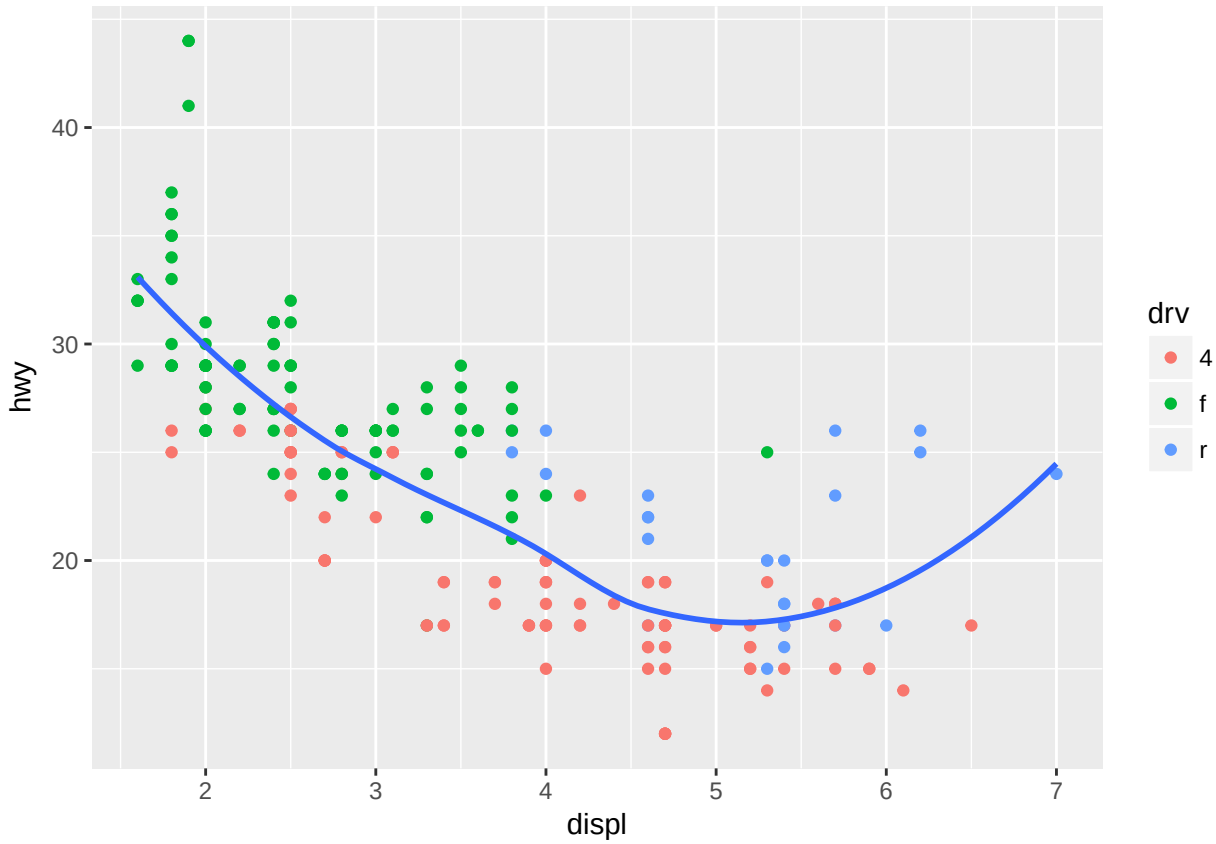

```
ggplot(data = mpg, mapping = aes(x = displ, y = hwy, color = drv)) +
  geom_point() +
  geom_smooth(se = FALSE)
```

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



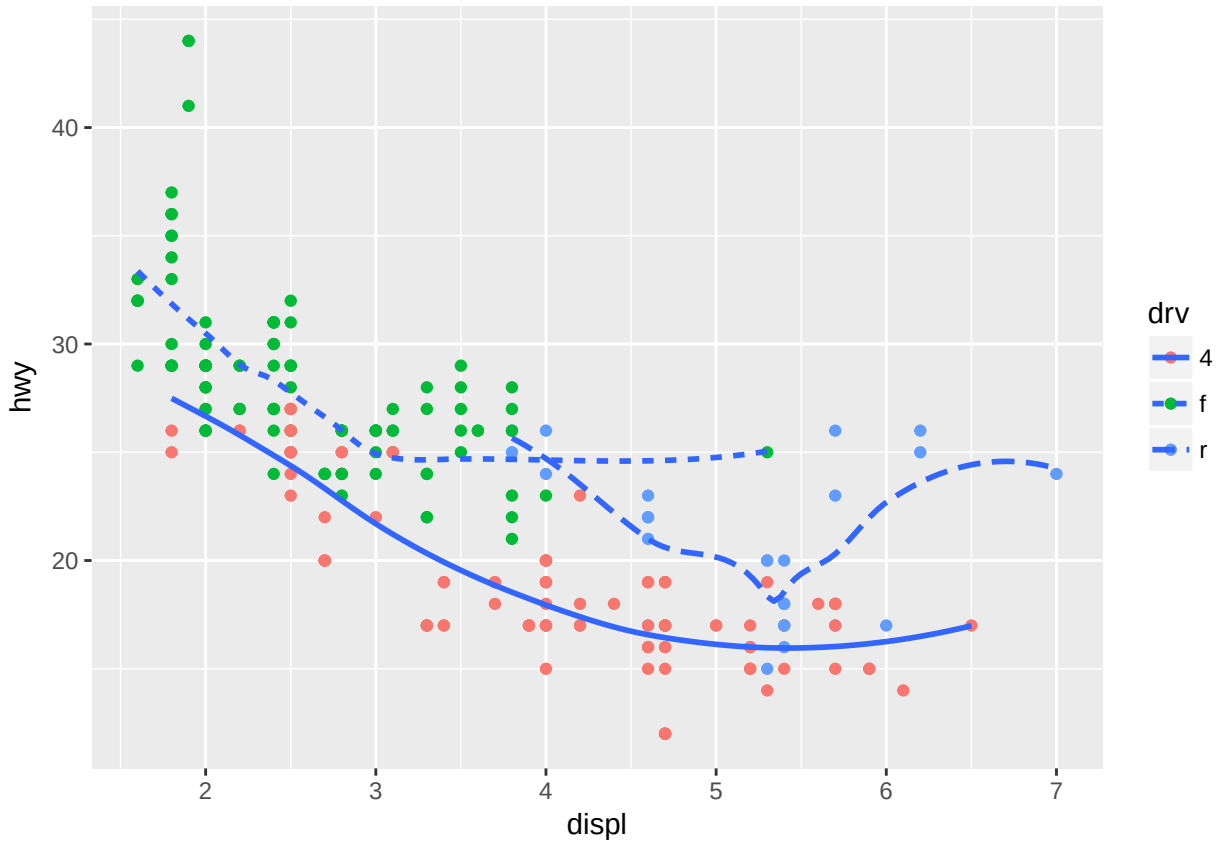

```
ggplot(data = mpg, mapping = aes(x = displ, y = hwy)) +
  geom_point(aes(color = drv)) +
  geom_smooth(se = FALSE)
```

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

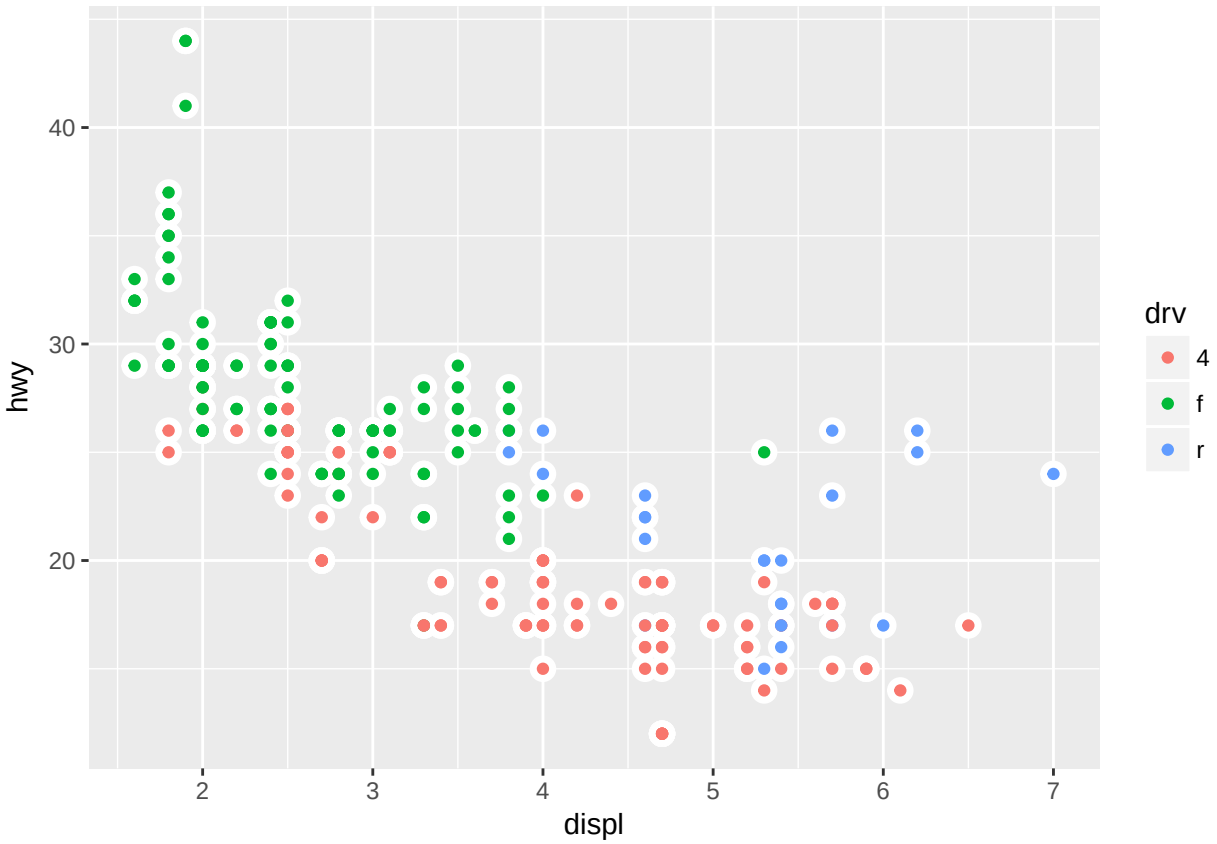



```
ggplot(data = mpg, mapping = aes(x = displ, y = hwy)) +
  geom_point(aes(color = drv)) +
  geom_smooth(aes(linetype = drv), se = FALSE)
```

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




```
ggplot(data = mpg, mapping = aes(x = displ, y = hwy)) +
  geom_point(size = 4, colour = "white") +
  geom_point(aes(colour = drv))
```



Section 3.7.1: #2 only

2. What does geom_col() do? How is it different to geom_bar()? geom_col makes the heights of the bars to represent values in the data. unlike geom_bar, which makes the height of the bar proportional to the number of cases in each group (or if the weight aethetic is supplied, the sum of the weights).

Final:

Look at the data graphics at the following link: What is a Data Scientist.

What works? All the views fit together well to tell a single story. The views flow well from one to the next.The single most important step to creating a great visualization is to have a point.The author has set the purpose well.

What doesn't work? I call for moiré vibration, heavy grids and self-promoting graphs that are used to demonstrate the graphic ability of the designer rather than display the data.

What would you have done differently? The data density of a graph is the proportion of the total size of the graph that is dedicated displaying data.I prefers high data density graphs.I want to maximize data density and the size of the data matrix within reason