COMPSCIX 415.2 Homework 2

Michelle Gomez

6/18/2018

3.2.4 Exercises:

1.

I see a blank graph with no axis labels.

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


2.

There are 234 rows and 11 columns in mpg.

```
dim(mpg)
```

```
## [1] 234  11
```

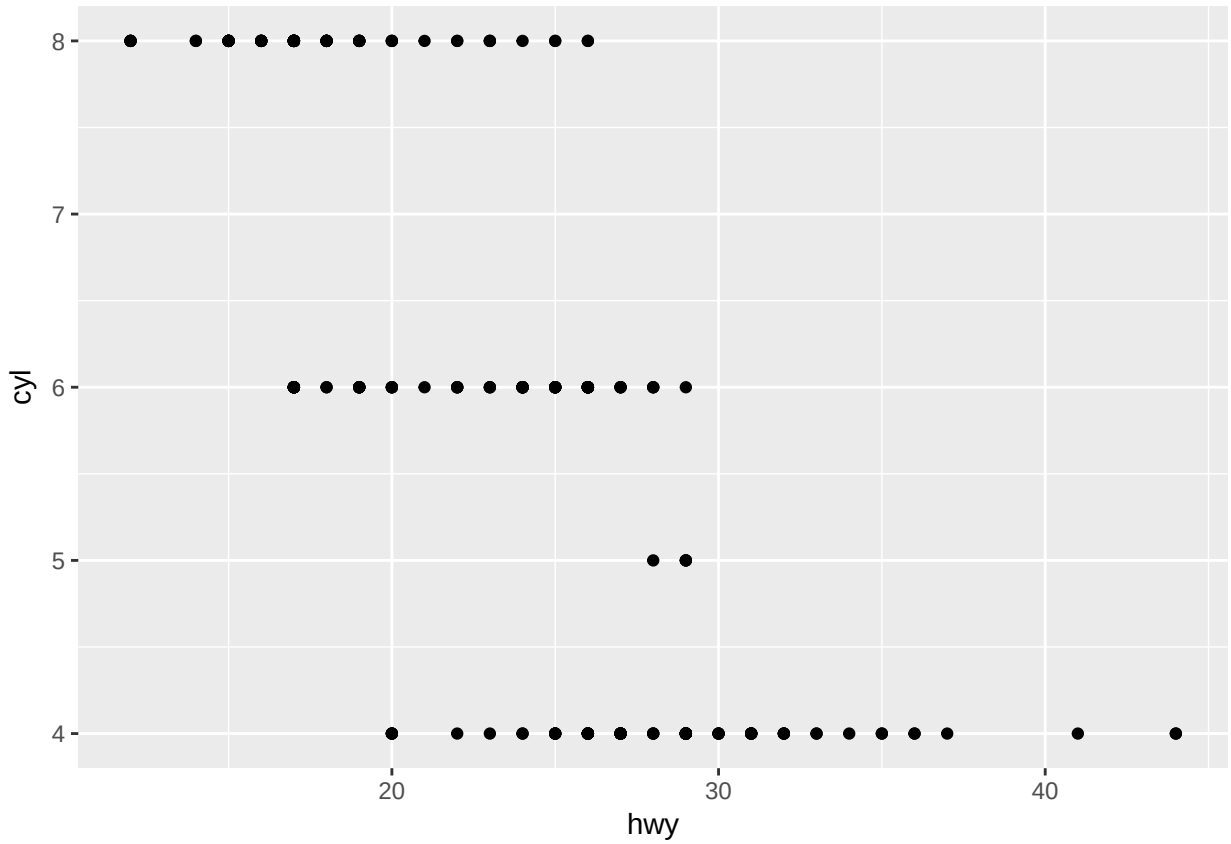
3.

The drv variable describes f = front-wheel drive, r = rear wheel drive, 4 = 4wd.

```
?mpg
```

4.

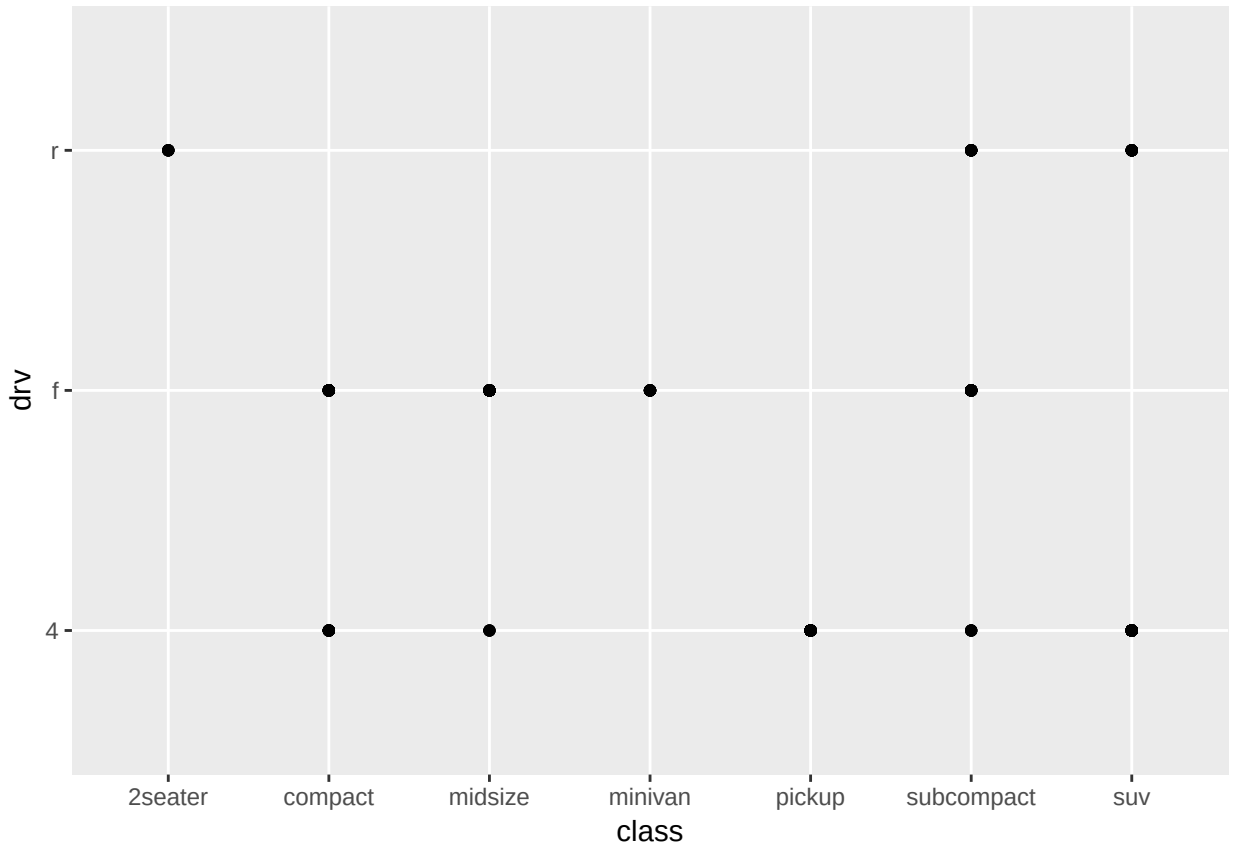
```
mpg %>% ggplot(aes(x = hwy, y = cyl)) +
geom_point()
```



5.

This graph is not useful because we have two categorical values that are not continuous graphed against each other and so the scatterplot does not illustrate a relationship across the x axis.

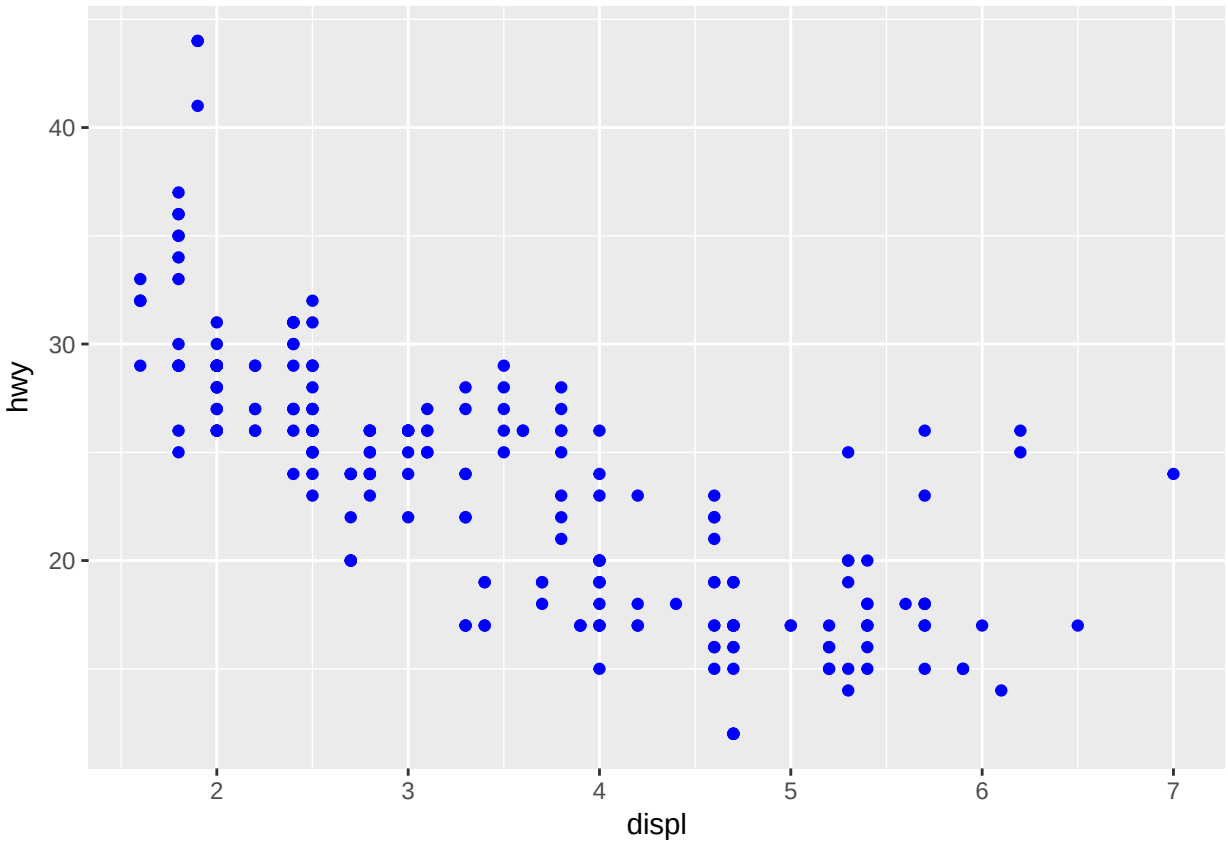
```
mpg %>% ggplot(aes(x = class, y = drv)) +
geom_point()
```



3.3.1 Exercises

1.

```
#ggplot(data = mpg) +
#geom_point(mapping = aes(x = displ, y = hwy, color = "orange"))
#color is set manually outside of aes()
ggplot(data = mpg) +
geom_point(mapping = aes(x = displ, y = hwy), color = "blue")
```



2.Which variables in mpg are categorical? Which variables are continuous? How can you see this information when you run mpg ?

Using the supply function, we can see the class of each variable– numeric, integer, or character. According to the results, categorical variables include all characters including: manufacturer, model, trans, drv, fl, class. There are some variables that are expressed as integers but the only continuous variable seems to be "displ".

```
sapply(mpg, class)
```

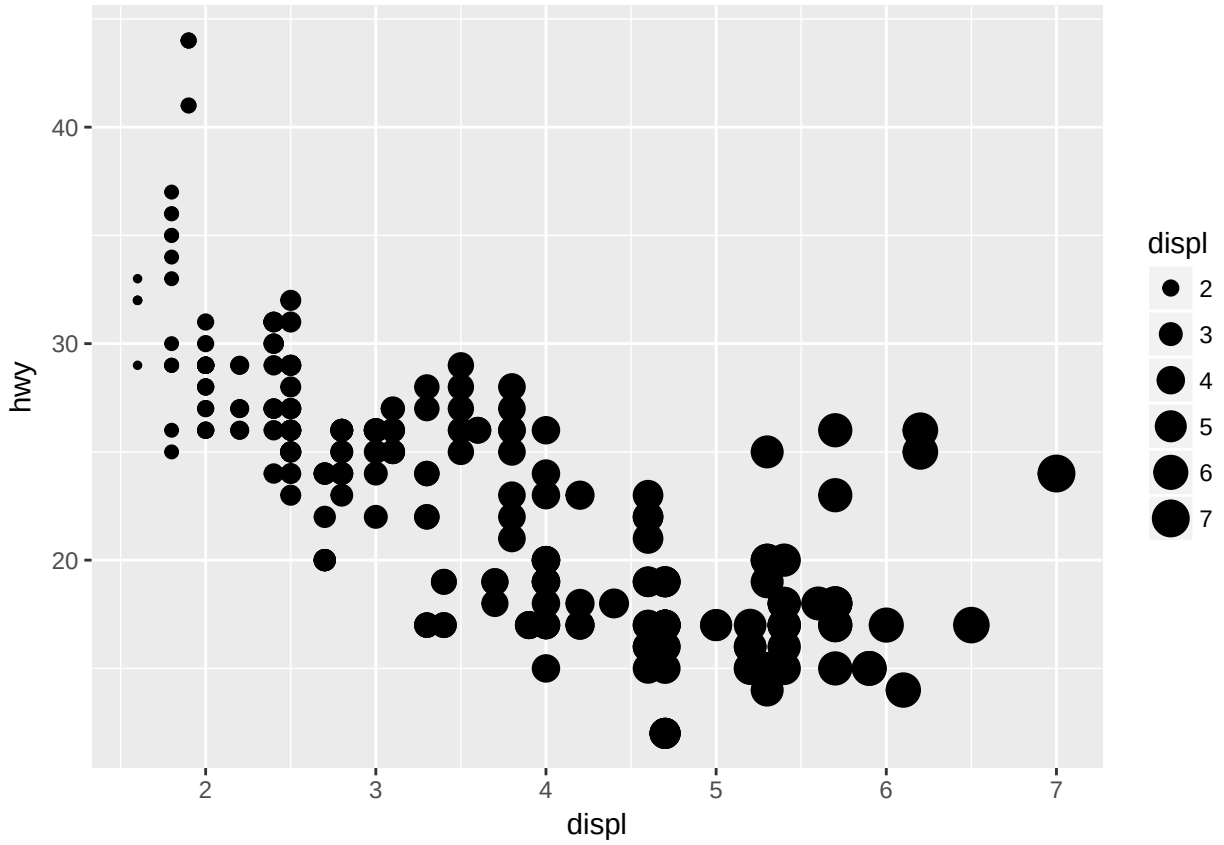
```
## manufacturer        model        displ         year          cyl
##  "character"  "character"    "numeric"    "integer"    "integer"
##        trans          drv          cty          hwy           fl
##  "character"  "character"    "integer"    "integer"  "character"
##        class
##  "character"
```

3.

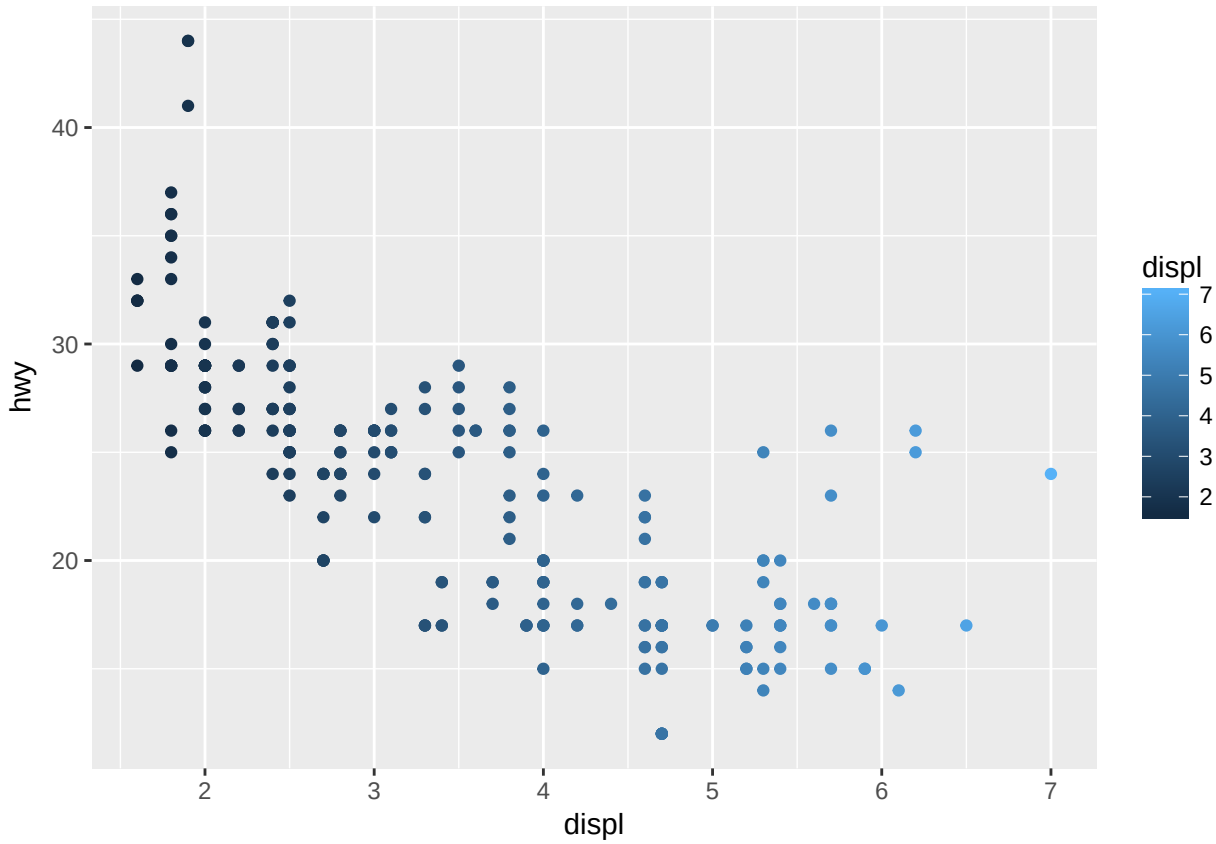
ggplot cannot map aes 'shape' to a continuous variable but can map 'size' and 'color'. While ggplot succeeds in mapping all aesthetics previously described to a categorical variable, using size for a categorical variable is not advised.
Additionally, the aesthetics are mapped in a spectrum of color and size for continuous variables, but remain static for categorical variables.

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

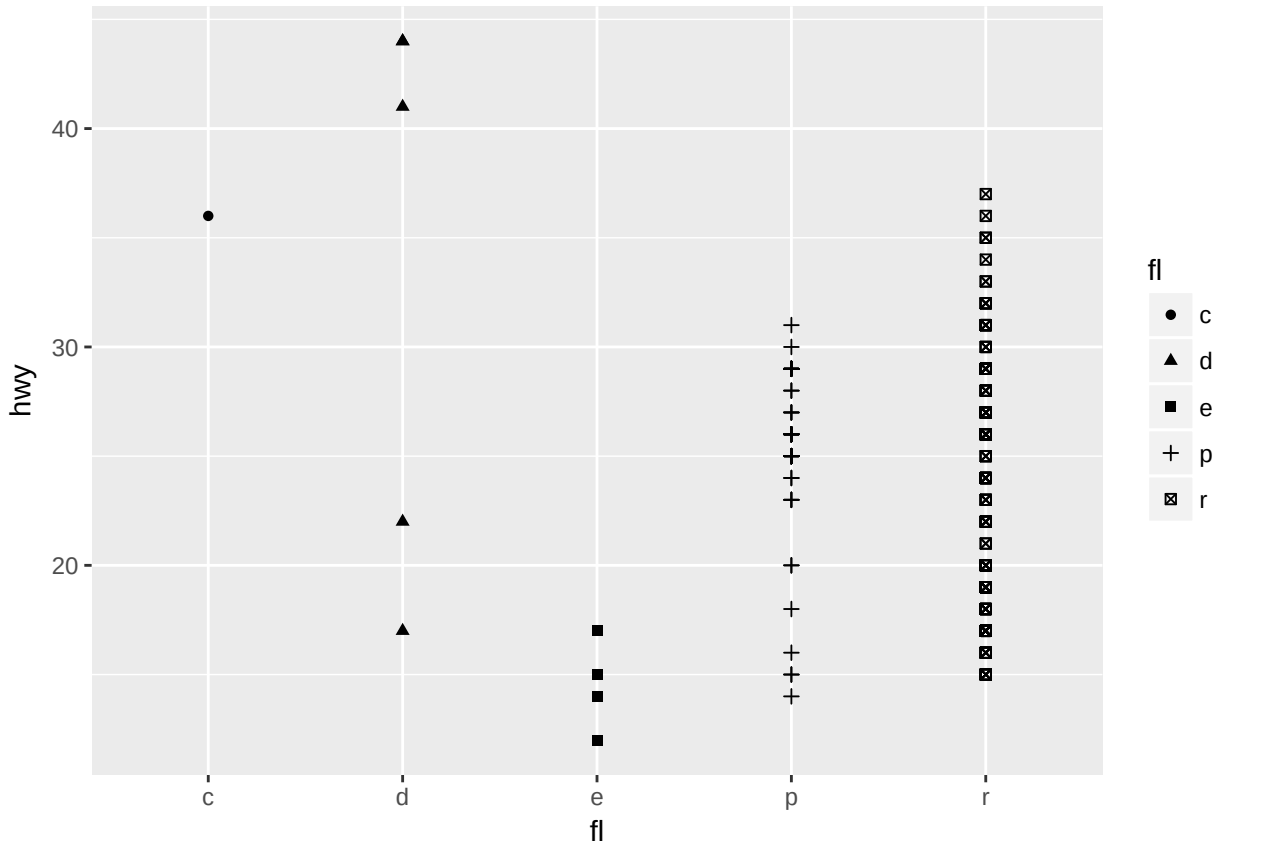
```
# ERROR message: A continuous variable cannot be mapped to a shape.

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




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

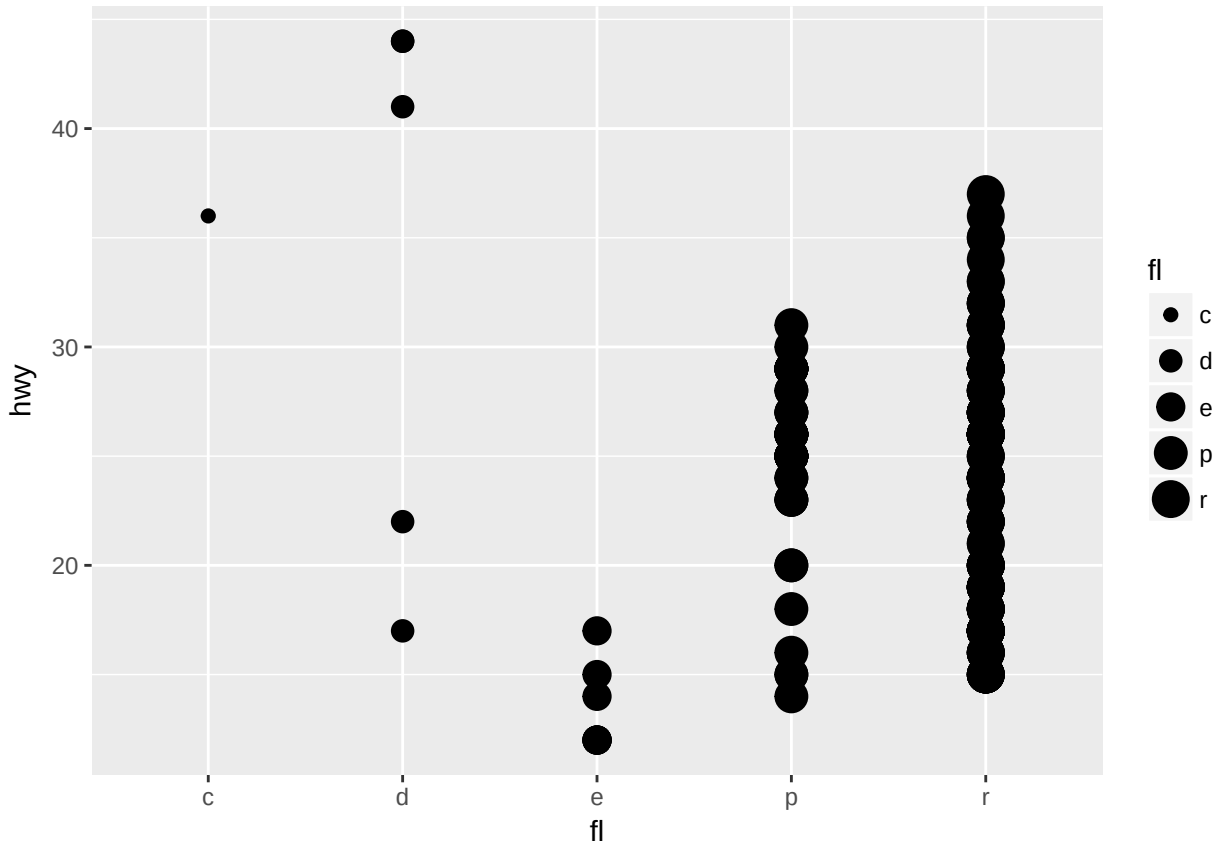



```
ggplot(data = mpg) +
geom_point(aes(x = fl, y = hwy, shape = fl))
```

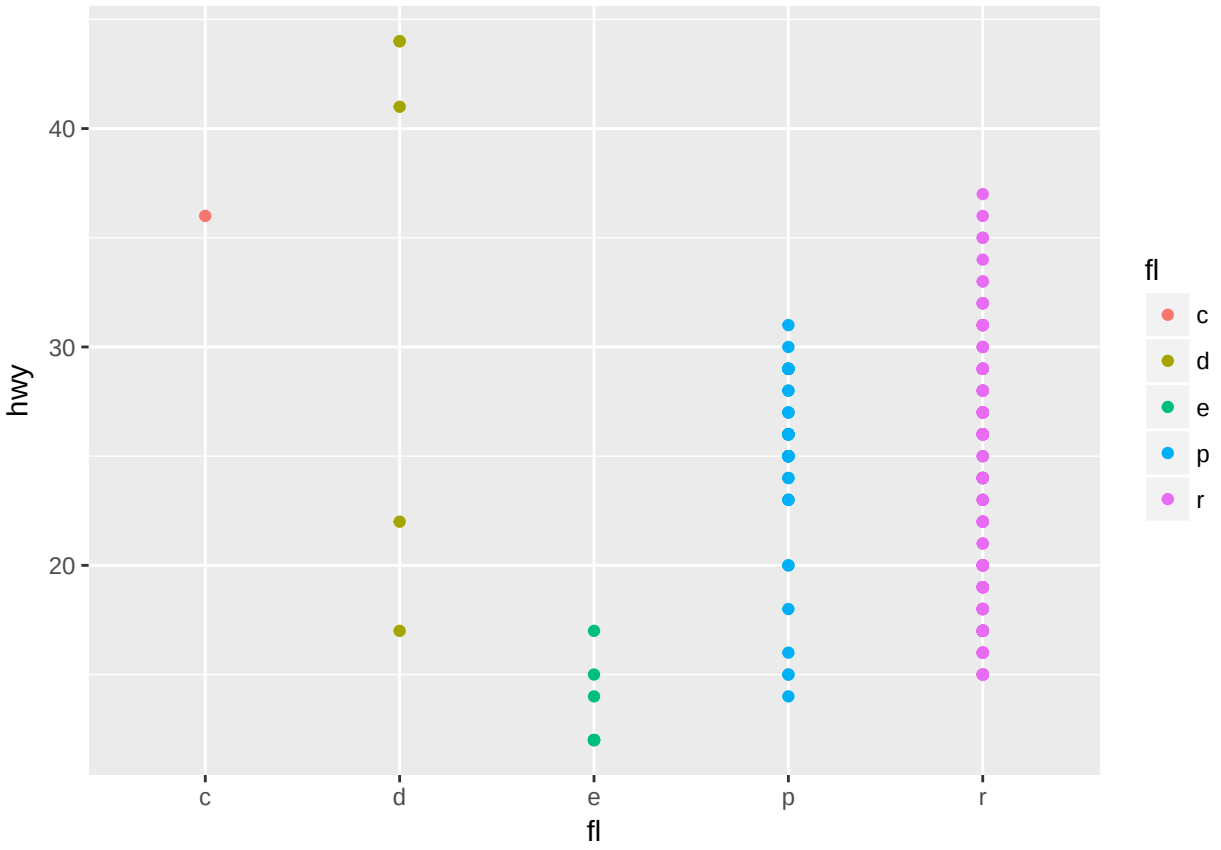



```
ggplot(data = mpg) +
geom_point(aes(x = fl, y = hwy, size = fl))
```

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




```
#message: using size for a discrete variable is not advised
ggplot(data = mpg) +
geom_point(aes(x = fl, y = hwy, color = fl))
```

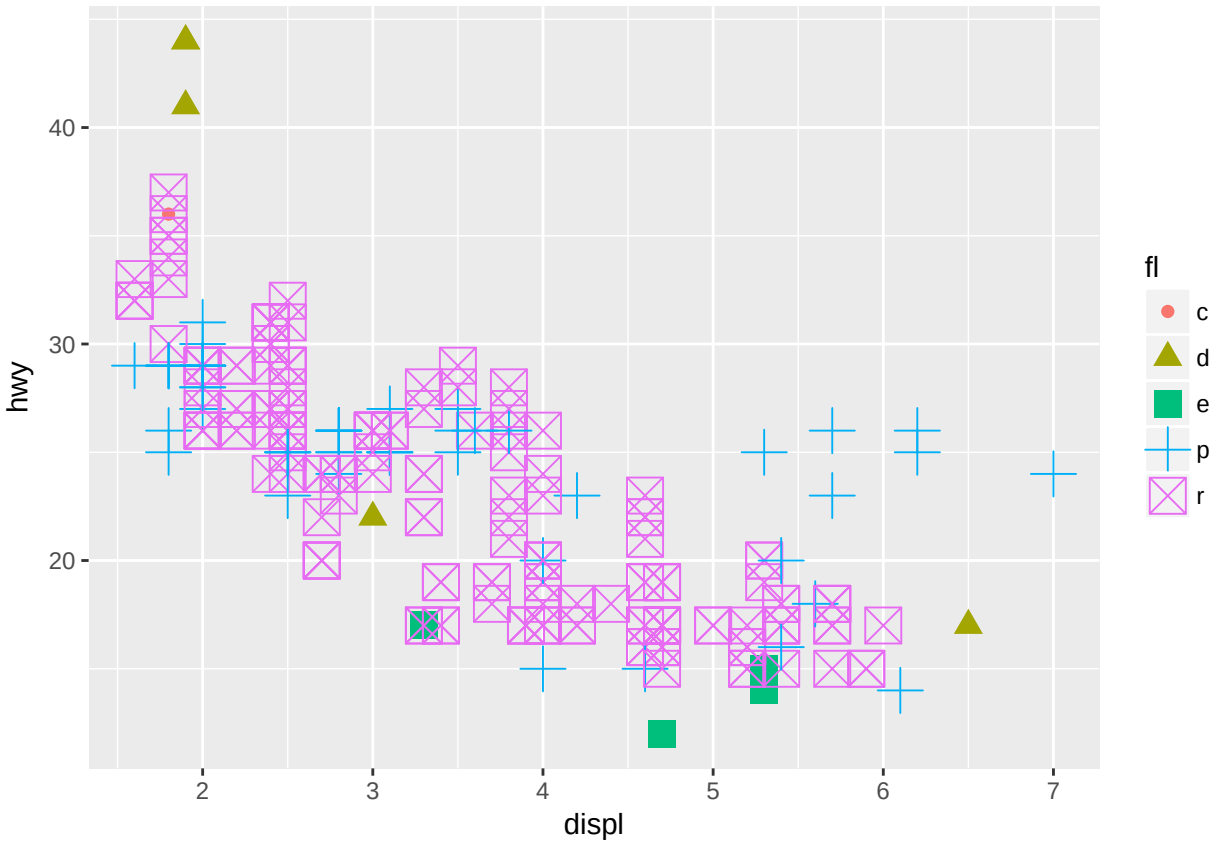


4.

When 'fl' is mapped to multiple aesthetics, these compound so that each one has a unique shape, size, and color.

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

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



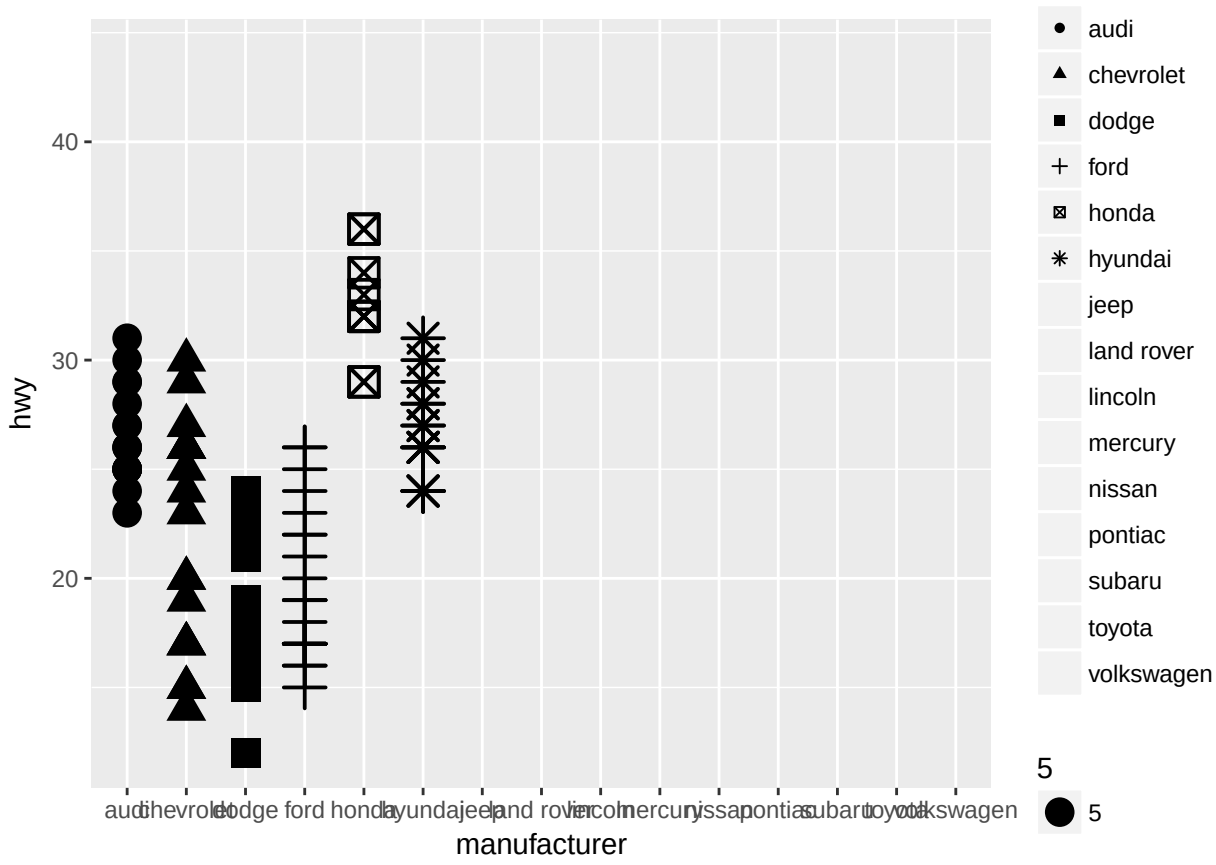
5.

stroke aesthetic changes the border or thickness of a shape.

```
ggplot(data = mpg) +
geom_point(aes(x = manufacturer, y = hwy, shape = manufacturer, size = 5, stroke = 1))
```

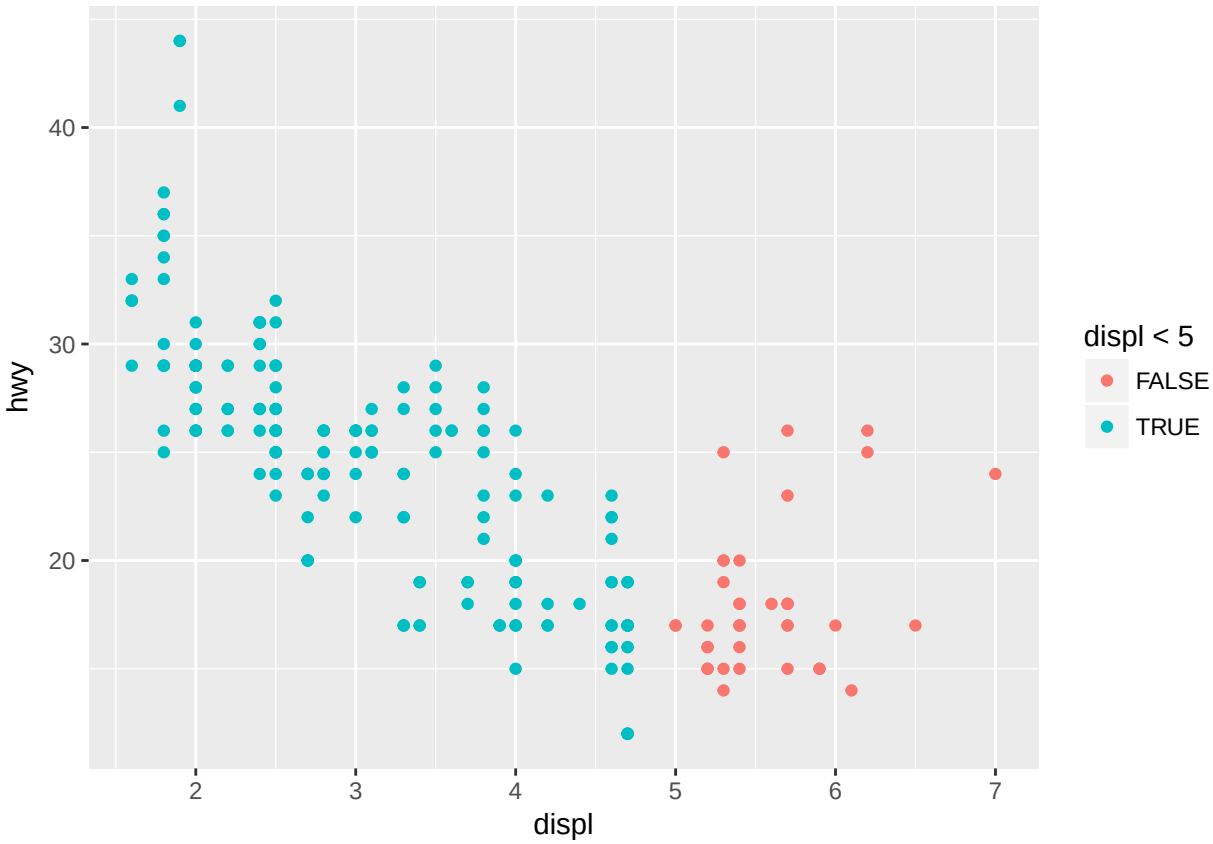
```
## Warning: The shape palette can deal with a maximum of 6 discrete values
## because more than 6 becomes difficult to discriminate; you have
## 15. Consider specifying shapes manually if you must have them.
```

```
## Warning: Removed 112 rows containing missing values (geom_point).
```



###6.
If you map an aesthetic to something other than a variable name, you create a true/false binary as seen in this function.

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

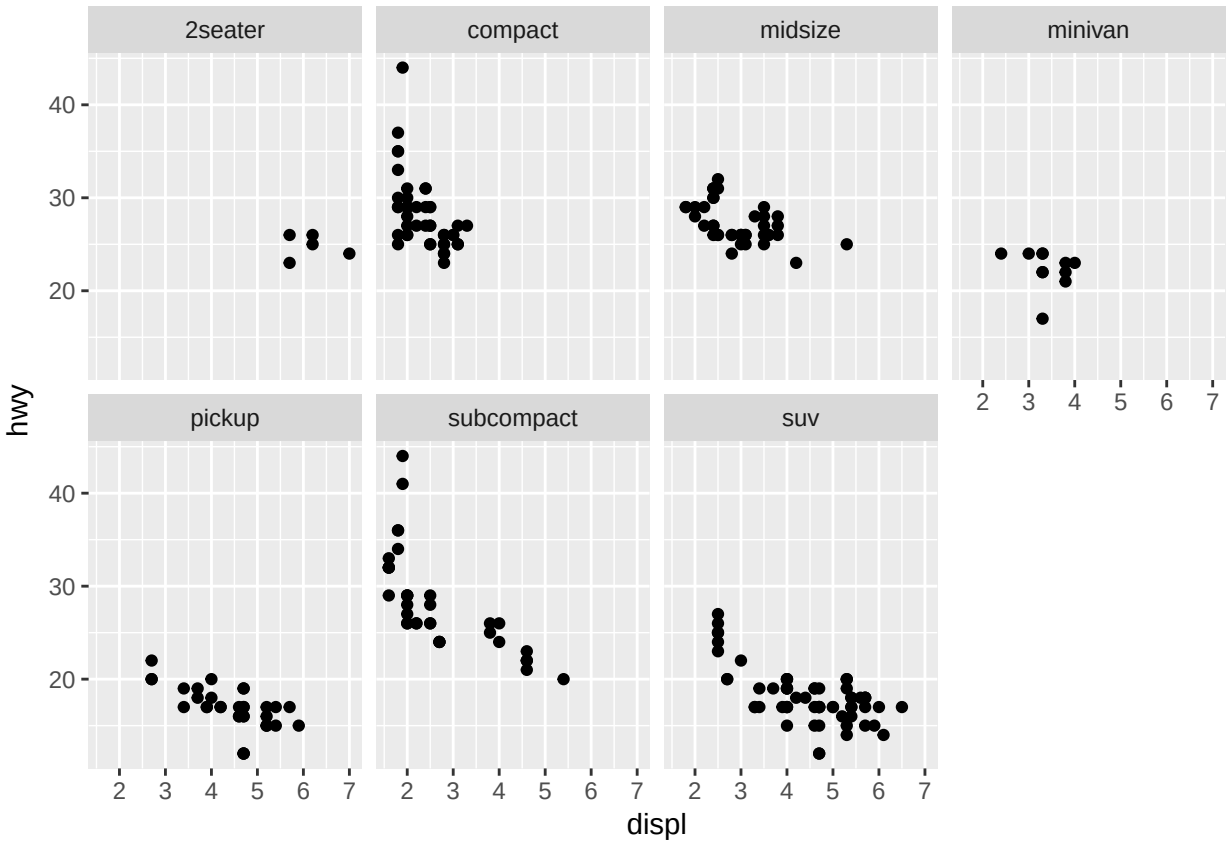


3.5.1 Exercises:

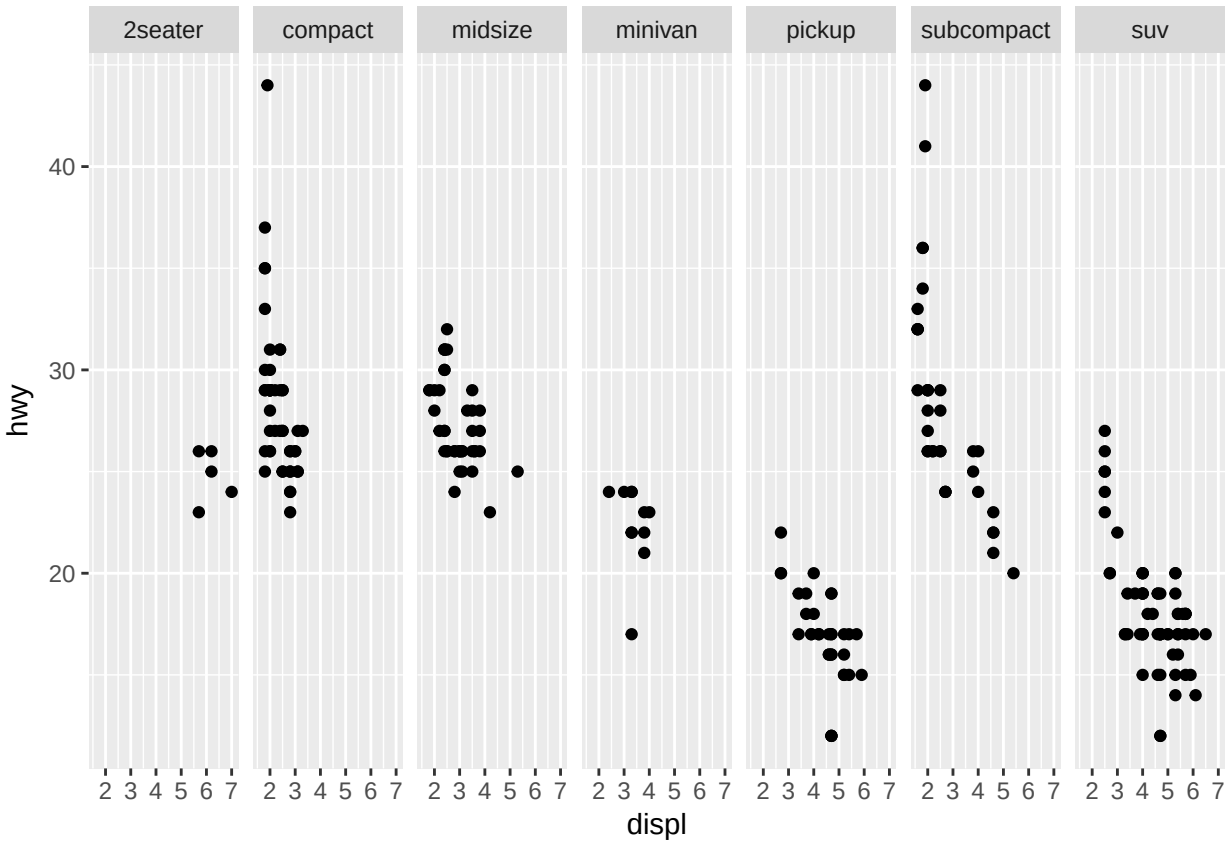
4.

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?
Faceting allows us to isolate the categorical/discrete variables into small plots side by side for comparison while a color aesthetic may hinder the ability to quickly compare relationships within each 'class', in this case. A disadvantage is that you can't quickly identify what subgroup has an overall max and min.
The balance might change if I had a larger data set because color aesthetic will cloud any identification of trends due to too many points in one plot, while faceting can clean this up.

```
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)) +
facet_grid(~ class)
```



###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 lets you set the number of rows you want the facets to appear in while ncol lets you set the number of columns.
Other options to control layout include: switch (swith y and x axis), strip.position (det labels on one side), scales (fixed or free scales), shrink (scales shrink to statistics output), etc.
facet_grid() automatically sets the facets into columns and there are no rows because they all share the same scale.

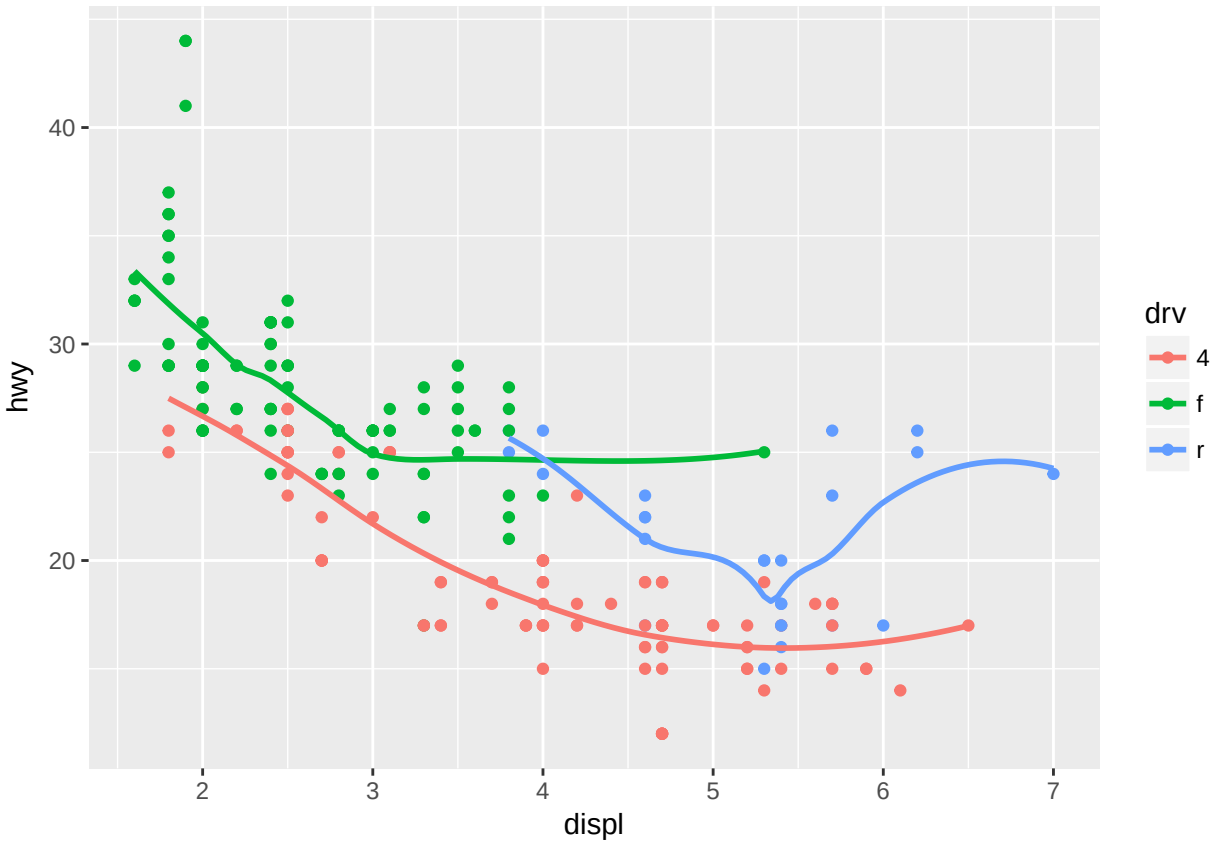
3.6.1 Exercises:

1.

What geom would you use to draw a line chart? A boxplot? A histogram? An area chart? To draw a line chart use geom_line, geom_step, or geom_path.
To draw a boxplot use geom_boxplot.
To draw a histogram use geom_histogram. To draw an area chart use geom_area. ###2.
I predicted that the graph would have one representation of smooth but instead it has a line for each drv without the variance shading typical of geom_smooth.

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

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

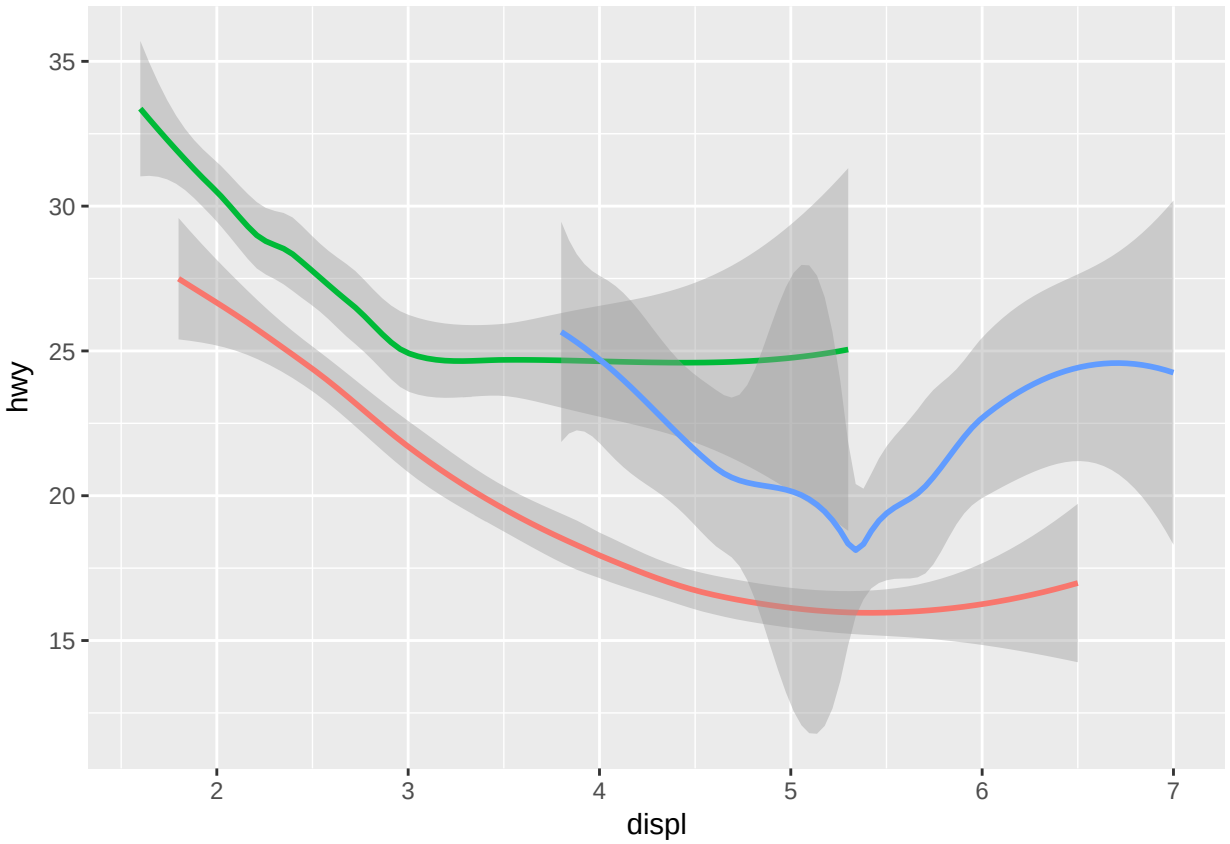


3.

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 removes the auto-legend feature on the right side of the graph.

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

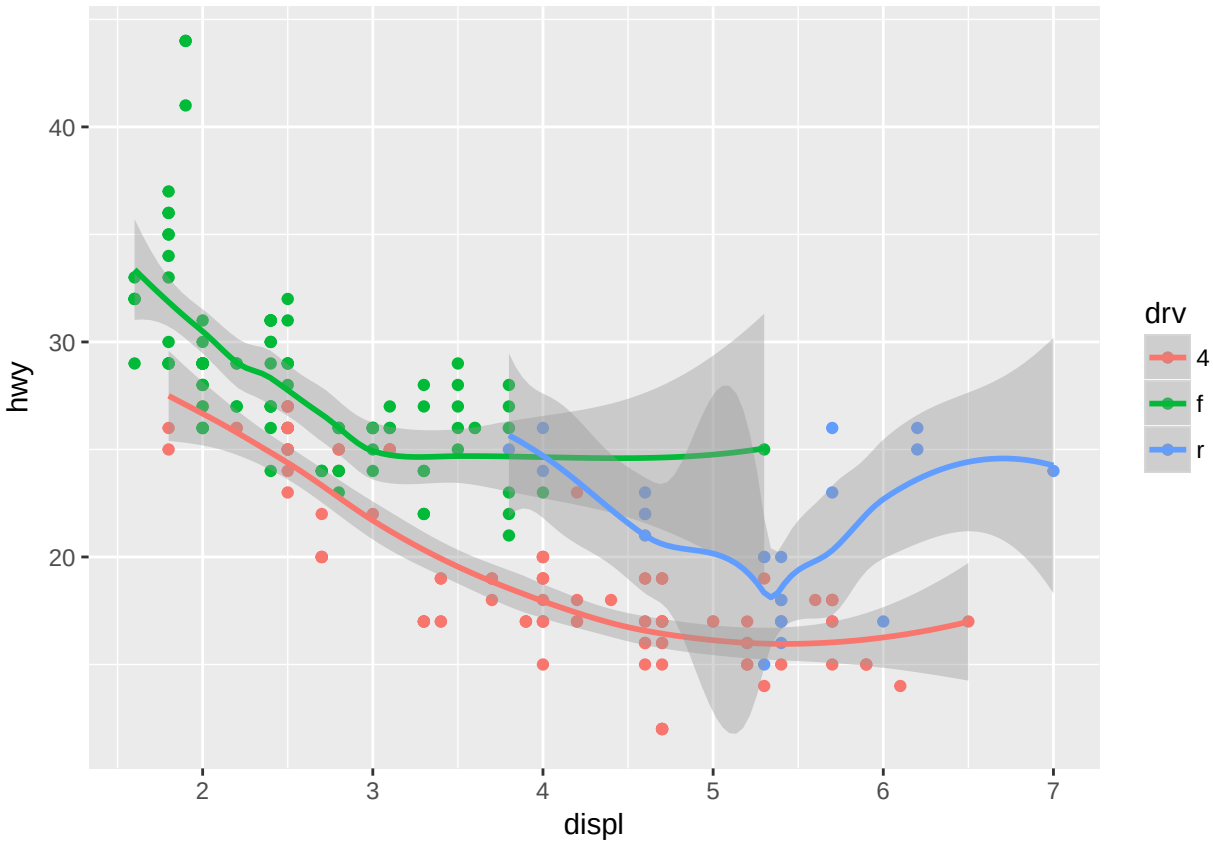
```
## `geom_smooth()` using method = 'loess'
```



###4. What does the se argument to geom_smooth() do?
I think the argument 'se' stands for standard error and when you set it to false in geom_smooth, you remove the standard error from this mapping function

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

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

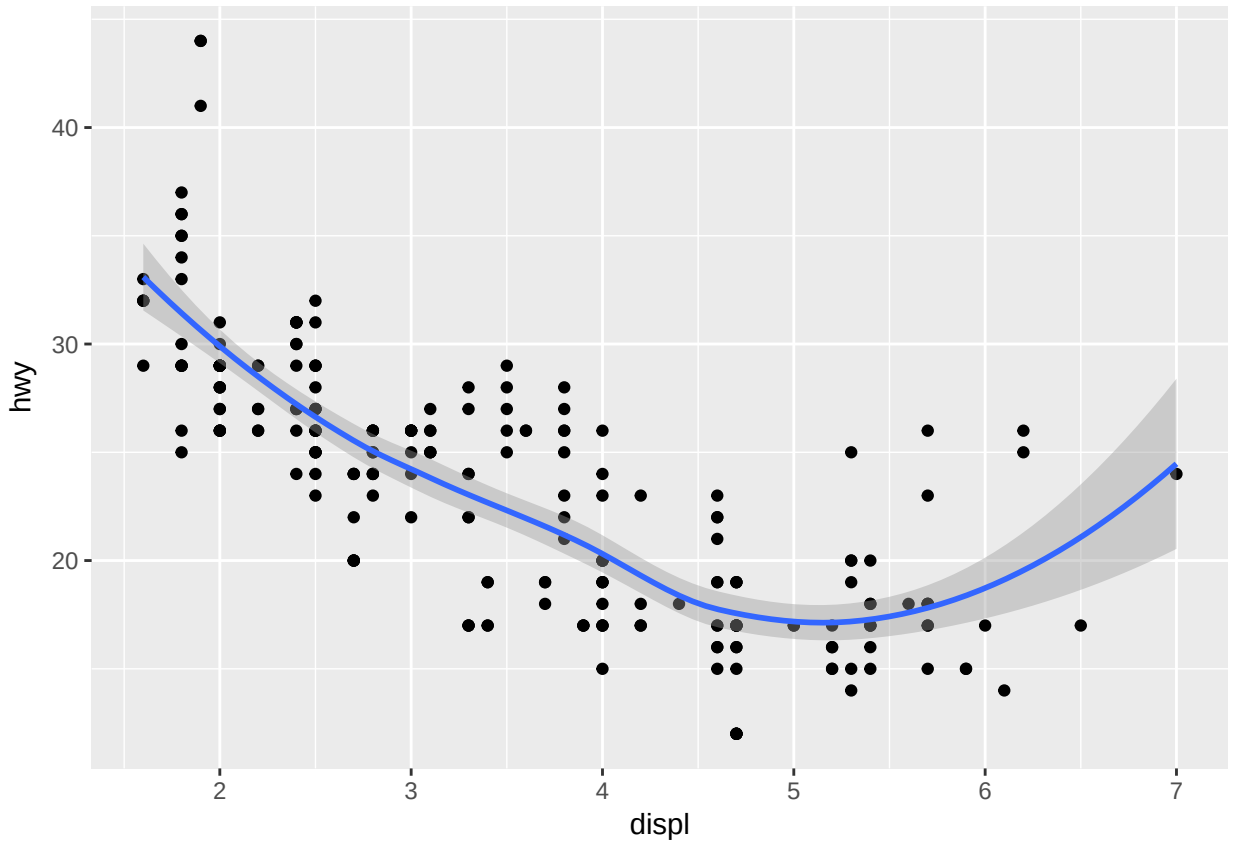


5.

These graphs look the same because they are both using the same data and aesthetics for geom_point and geom_smooth. The first code is less redundant.

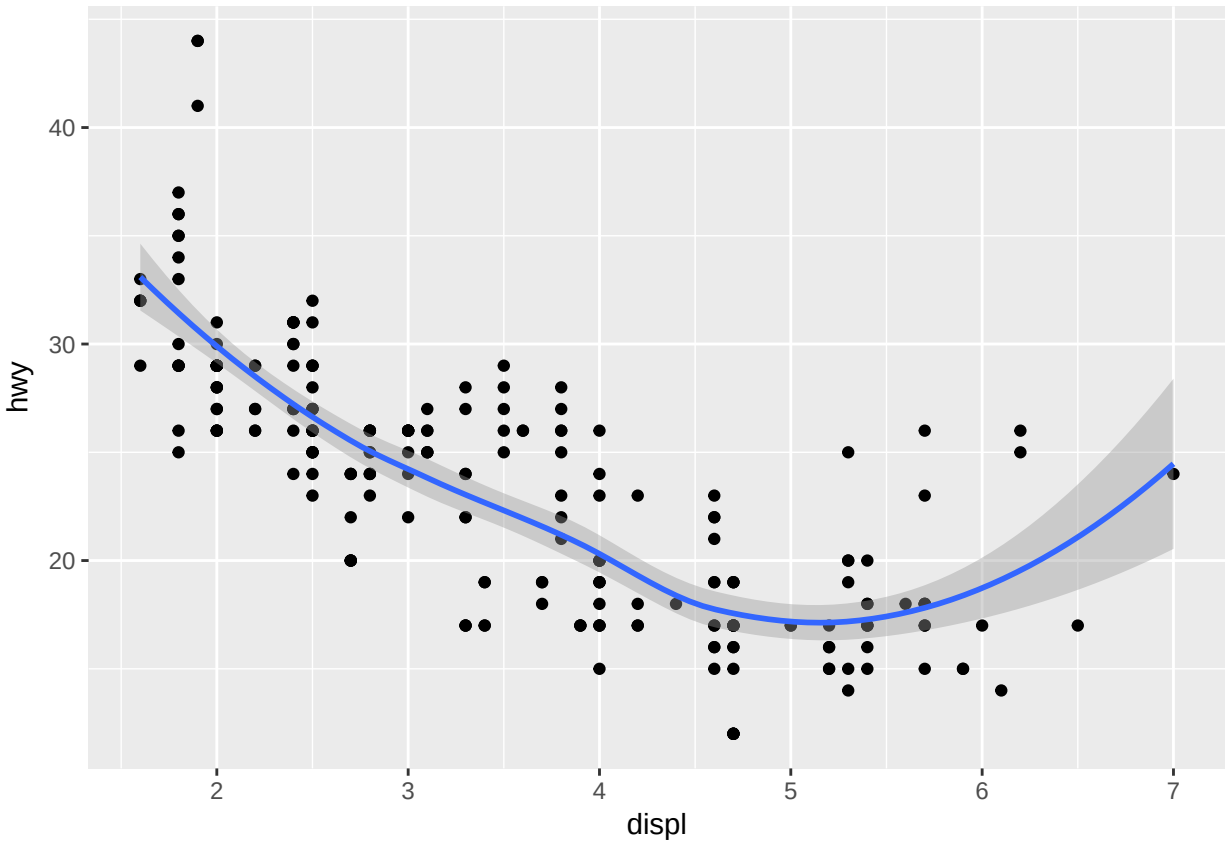
```
ggplot(data = mpg, mapping = aes(x = displ, y = hwy)) +
geom_point() +
geom_smooth()
```

```
## `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'
```



3.7.1 Exercises:

1.

What is the default geom associated with stat_summary() ? How could you rewrite the previous plot to use that geom function instead of the stat function?
According to RMarkdown help, stat_summary() is used "to override the default connection between geom_histogram/geom_freqpoly and stat_bin".

```
#ggplot(data = diamonds) +
#geom_pointrange( mapping = aes(x = cut, y = depth, ymin = fit-se,
#ymax= fit+se))
```