R for Data Science Exercises

Pete Lawson

Table of contents

Background

This `Quarto` book collects all of my completed exercises and notes from Wickham and Grolemund (2016).

Wickham, Hadley, and Garrett Grolemund. 2016. *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. First edition. Sebastopol, CA: O'Reilly.

1 Explore

---

title: “Explore - R for Data Science” author: “Pete Lawson” date: “2022-07-27” output: html_document

---

Exercises for **Explore** section of the R for Data Science textbook.

1.1 Data Visualization

1.1.1 First Steps

1.1.1.1 3.2.4 Exercises

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

empty gray background

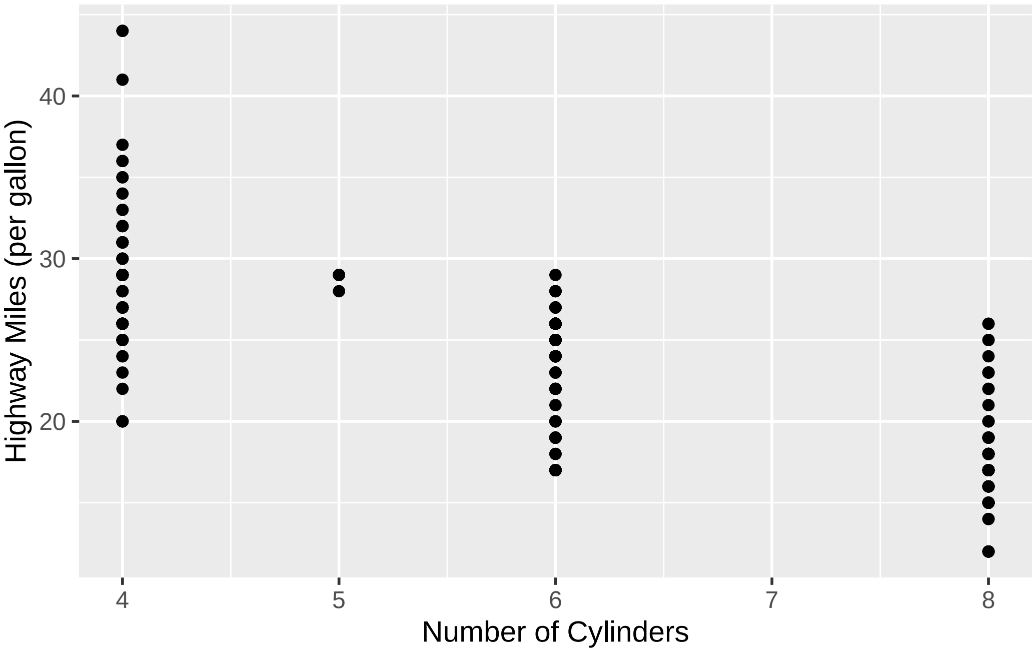
2. How many rows are in mpg? How many columns?

There are 234 rows and 11 cols.

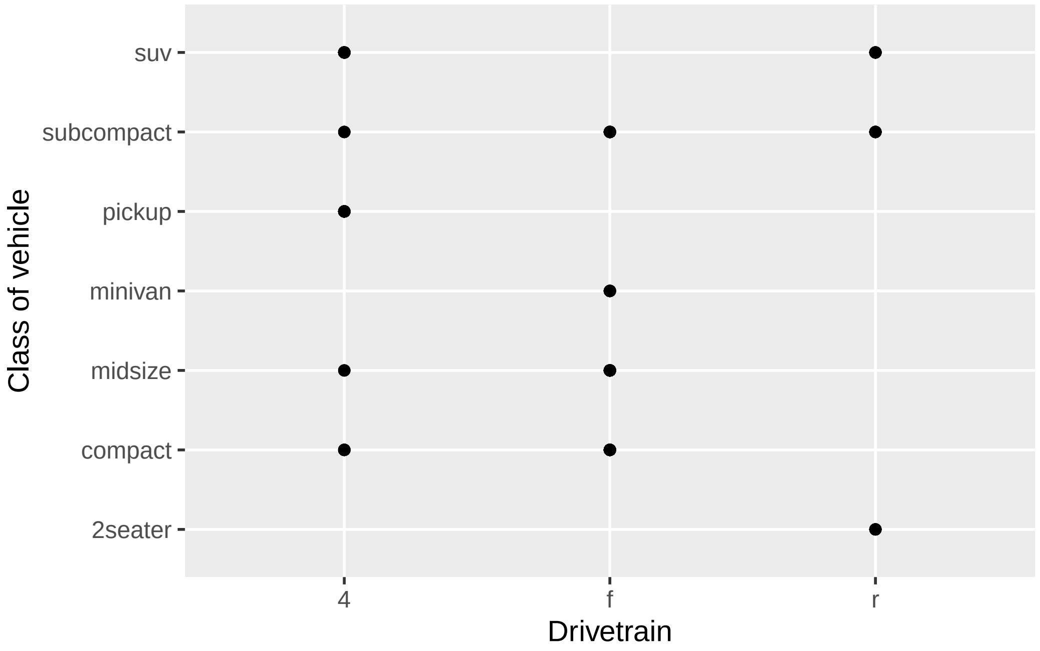
3. What does the drv variable describe? Read the help for ?mpg to find out.

the type of drive train, where f = front-wheel drive, r = rear wheel drive, 4 = 4wd

4. Make a scatterplot of hwy vs cyl.



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



Not useful, both variables are categorical, no continuous relationship exists between variables.

1.1.1.2 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"))
```

By assigning color inside aes, color is mapped to a variable, since only a factor of len 1 is present ("blue"), it is mapped to the first color in the default palette, red.

2. 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: Manufacturer, model, year, cyl, trans, drv, fl, class; *Continuous*: disp, cty, hwy

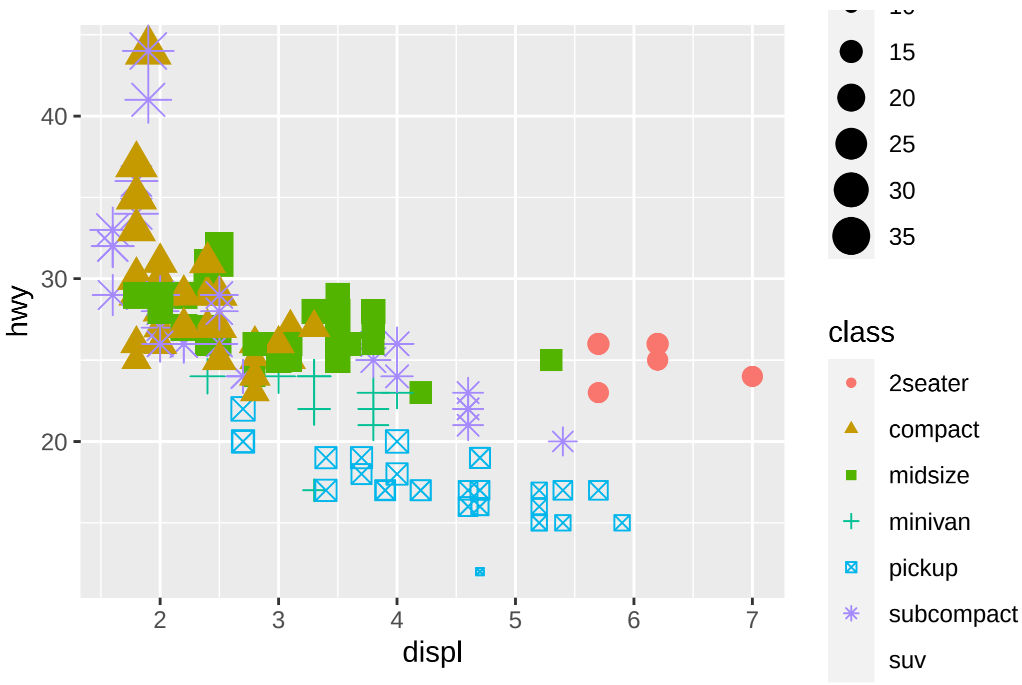
Described under colname, inferred by type, for ex. chr is a character, and thus a categorical label. Year, although an int, is in this case categorical. Cyl is also an int but is categorical; there is no 4.5 cylinder car (unless it is having a serious mechanical failure).

3. 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 = class, size=cty, 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).
```



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

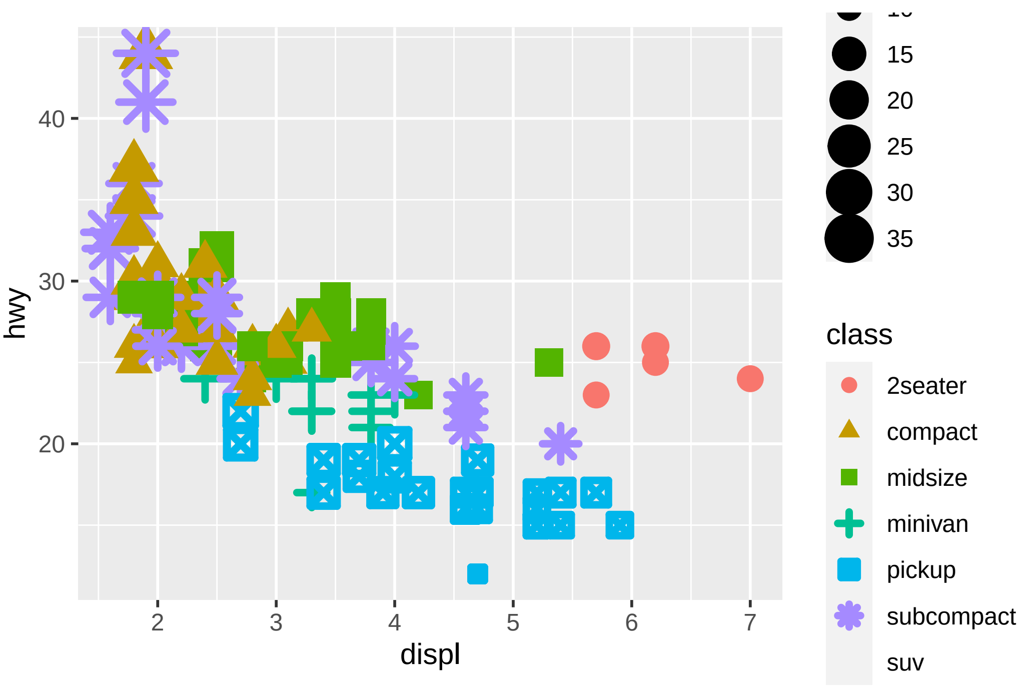
The variables legends are combined.

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, color = class, size=cty, shape = class),
             stroke = 2)
```

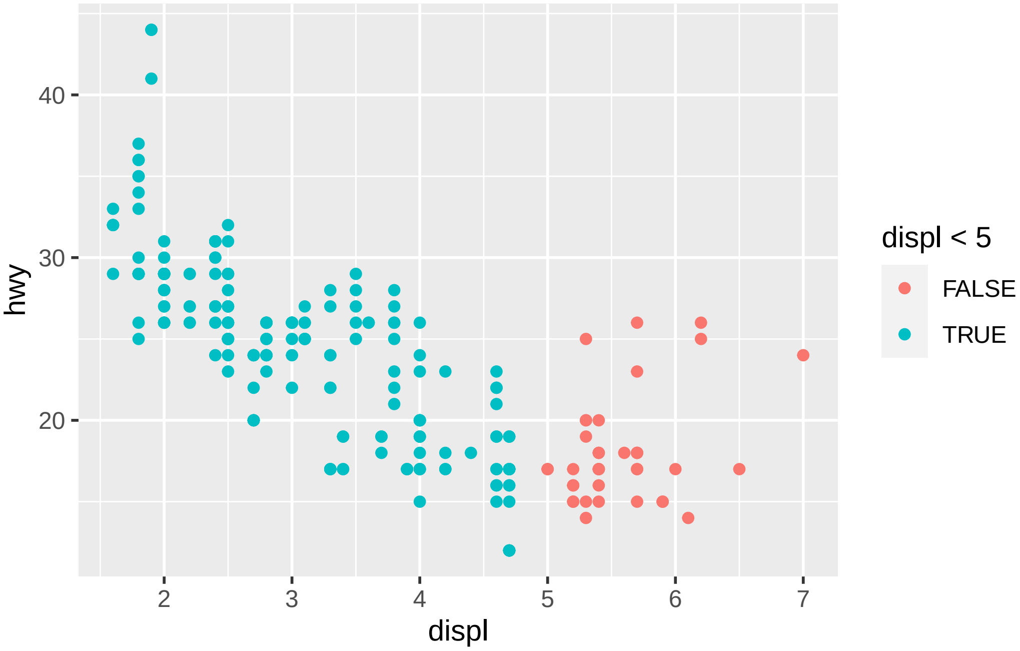
```
Warning: The shape palette can deal with a maximum of 6 discrete values because
more than 6 becomes difficult to discriminate; you have 7. Consider
specifying shapes manually if you must have them.
```

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



Changes the thickness of line border on points.

6. What happens if you map an aesthetic to something other than a variable name, like aes(colour = displ < 5)? Note, you'll also need to specify x and y.



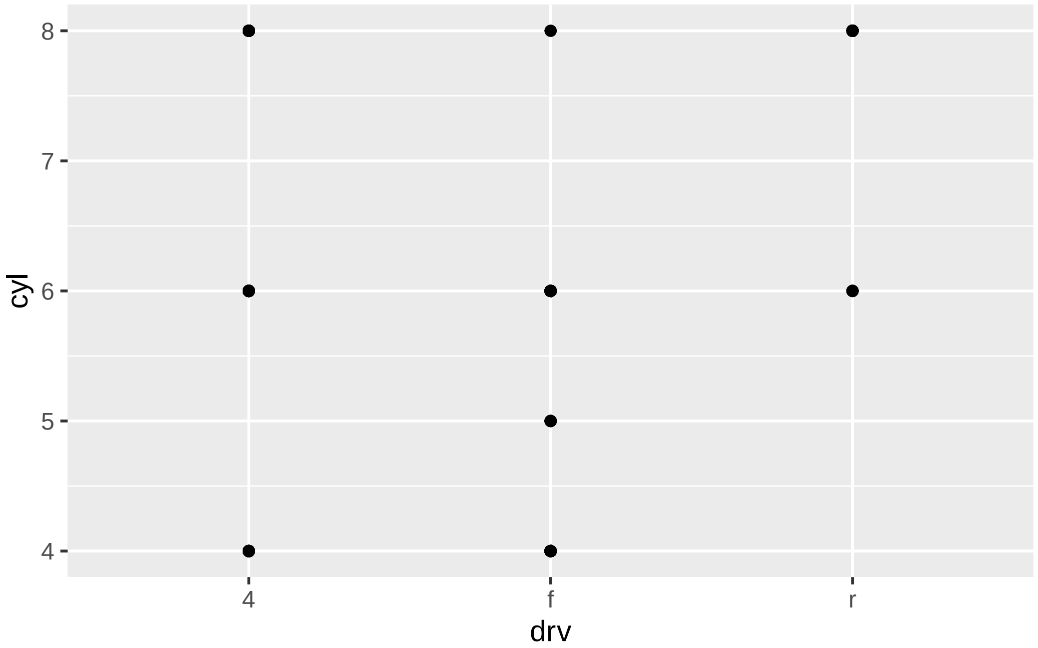
Colors points meeting that condition in one color, points not meeting that condition in another. FALSE is first alphabetically, so receives first color in palette (red) and TRUE receives blue.

1.1.1.3 3.5.1 Exercises

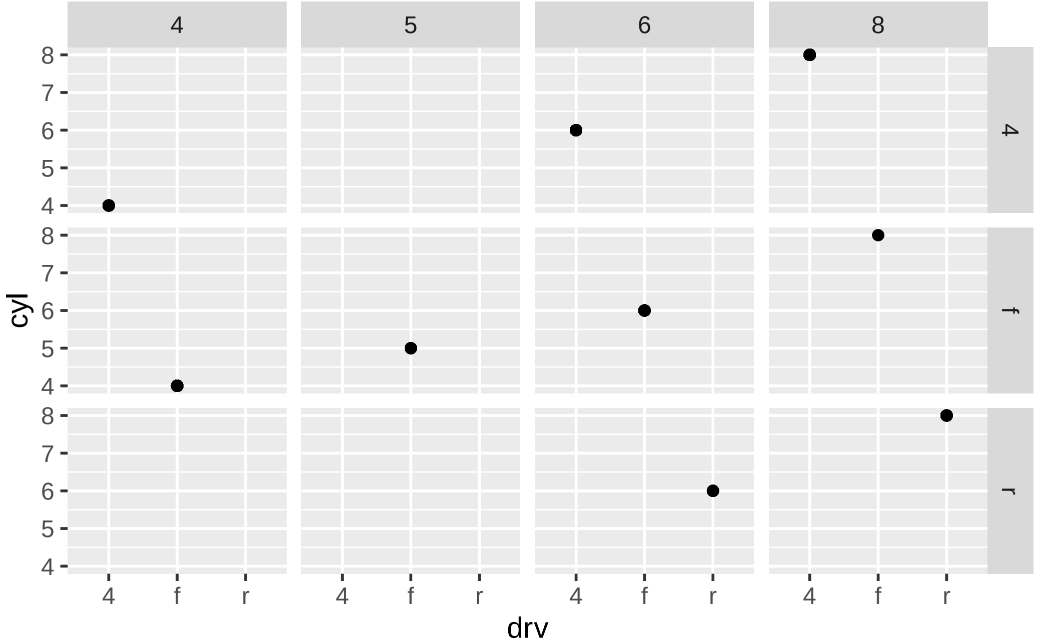
1. What happens if you facet on a continuous variable?

A facet is created for every occurrence of a unique continuous variable.

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

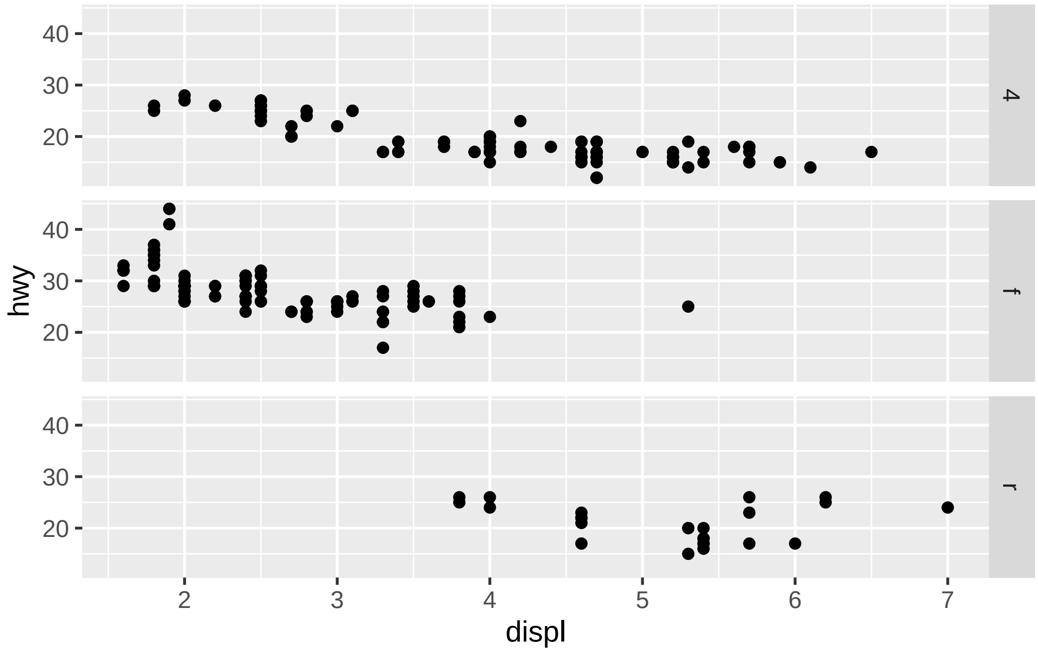


Empty cells (shown below) indicate that no data is present for that particular pair of variables, that is, there are no 4-wheel drive 5 cylinder vehicles, nor 4 or 5 cylinder rear wheel drive vehicles. This is reflected in the above plot, where no point is present for those variable combinations combinations.

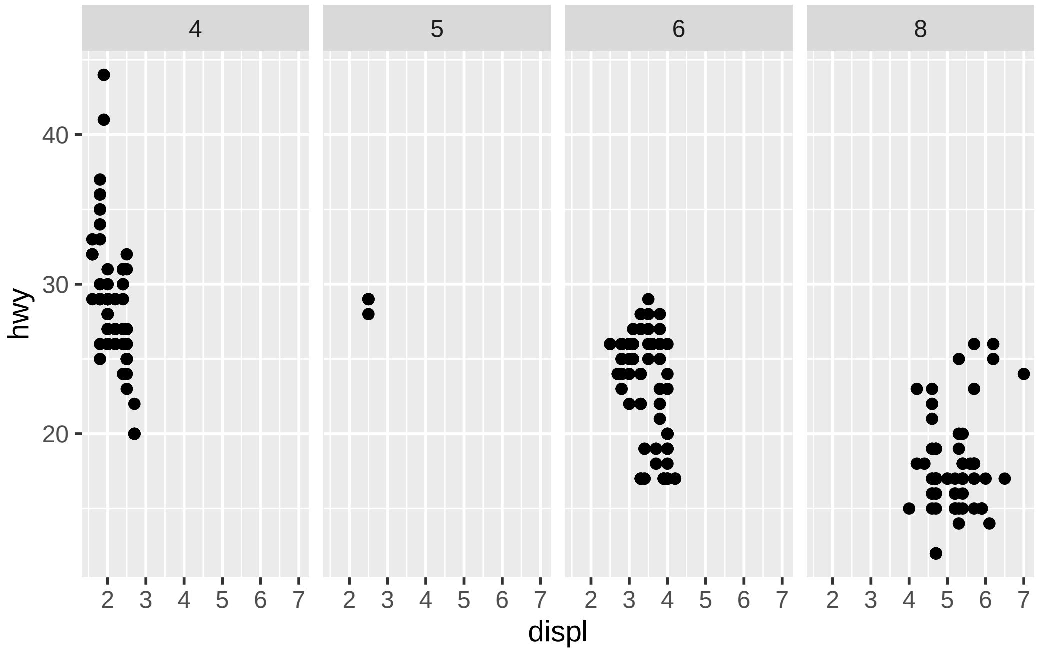


3. What plots does the following code make? What does . do?

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



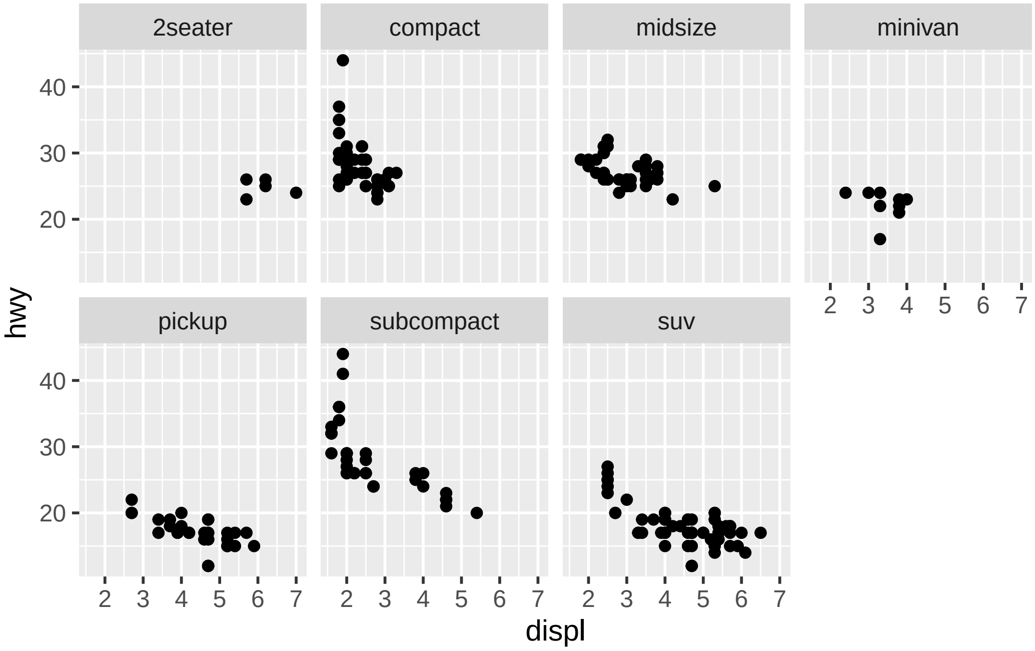
```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy)) +
  facet_grid(. ~ cyl)
```



Facet grid uses a row by column ordering. In the first plot, it will show facets by drive (f, r, 4) by row. The second plot will show cylinder count by column. The . is a dummy variable indicating no data. The code will not run without the dummy variable.

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?

Groups differentiated by color become difficult to discern when using more than 6 or more colors. Additionally, poor color choices can result in an inability to discern differences by group for those with colorblindness. Facets provide an unambiguous way of separating data by a categorical variable. It can make a visual comparison more difficult, which can be resolved by keeping each variable aligned along a common orientation (row vs column). With a larger number of variables, facets may become cumbersome.

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 arguments?
6. When using facet_grid() you should usually put the variable with more unique levels in the columns. Why?