Lab - Getting Started with ggplot2

Alireza Mostafizi

12 April 2018

3.1 Introduction

In this lab, I worked through *Sections 3.1 through 3.4 of R for Data Science*, documented the process, and answered to the exercises.

3.1.1 Prerequisites

First we load the `tidyverse` and `ggplot2` libraries.

```
library(tidyverse, ggplot2)
```

```
## -- Attaching packages ----------------------------------- tidyverse 1.2.1 --
## v ggplot2 2.2.1     v purrr   0.2.4
## v tibble  1.4.2     v dplyr   0.7.4
## v tidyr   0.8.0     v stringr 1.3.0
## v readr   1.1.1     v forcats 0.3.0
## -- Conflicts -------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

3.2 First steps

There are a few questions mentioned in this section as following:

- Do cars with big engines use more fuel than cars with small engines?
- What does the relationship between engine size and fuel efficiency look like? Is it positive? Negative? Linear? Nonlinear?

3.2.1 The `mpg` dataframe

Here is how `mpg` dataframe looks like:

```
mpg
```

```
## # A tibble: 234 x 11
##    manufacturer model    displ  year   cyl trans   drv     cty   hwy fl
##    <chr>        <chr>    <dbl> <int> <int> <chr>   <chr> <int> <int> <chr>
##  1 audi         a4        1.80  1999     4 auto(l~ f        18    29 p
##  2 audi         a4        1.80  1999     4 manual~ f        21    29 p
##  3 audi         a4        2.00  2008     4 manual~ f        20    31 p
##  4 audi         a4        2.00  2008     4 auto(a~ f        21    30 p
##  5 audi         a4        2.80  1999     6 auto(l~ f        16    26 p
##  6 audi         a4        2.80  1999     6 manual~ f        18    26 p
##  7 audi         a4        3.10  2008     6 auto(a~ f        18    27 p
##  8 audi         a4 quat~  1.80  1999     4 manual~ 4        18    26 p
```

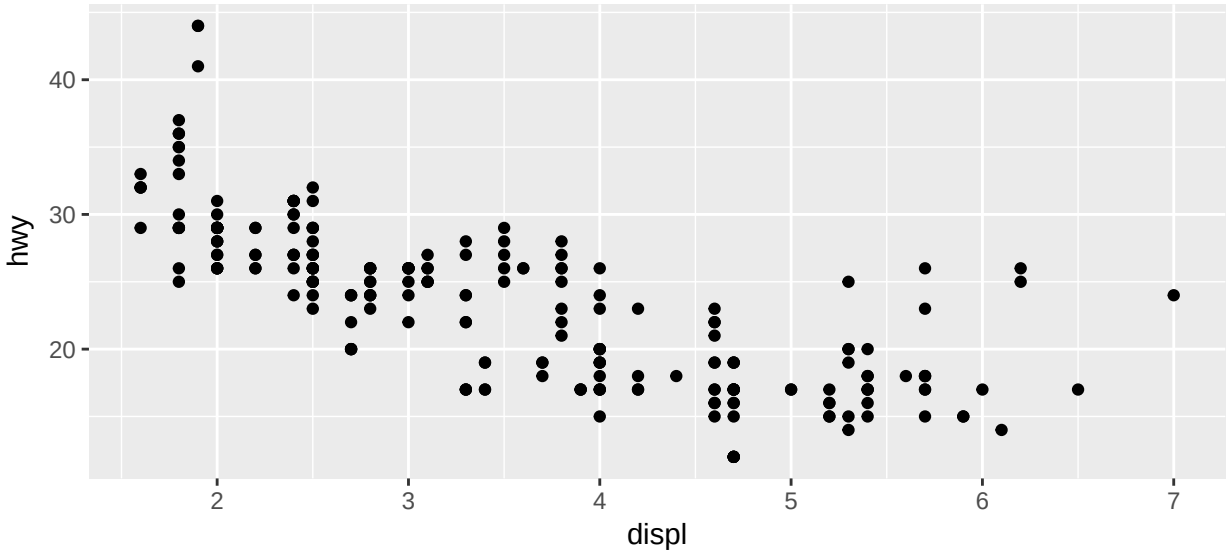
```
##  9 audi         a4 quat~  1.80  1999     4 auto(l~ 4        16    25 p
## 10 audi         a4 quat~  2.00  2008     4 manual~ 4        20    28 p
## # ... with 224 more rows, and 1 more variable: class <chr>
```

And here some information about the dataset:

3.2.2 Creating a ggplot

Let's start answering the questions by inverstigating the relationship between the engine size (`displ`) and the fuel efficiency or mpg (`hwy`) of 38 popular models of cars from 1999 to 2008.

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



It can be seen that that the cars with larger engines have lower mpg, and thus, lower fuel efficiency. At this point, it can be stated that the relationship looks non-linear.

3.2.3 A graphing template

The code for the plot generated above can be generalized to the following format.

```
ggplot(data = <DATA>) +
  <GEOM_FUNCTION>(mapping = aes(<MAPPINGS>))
```

3.2.4 Excersie

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

This code generates an empty graph.

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

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

```
nrow(mpg)
```

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

```
ncol(mpg)
```

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

The data has 234 rows and 11 columns.

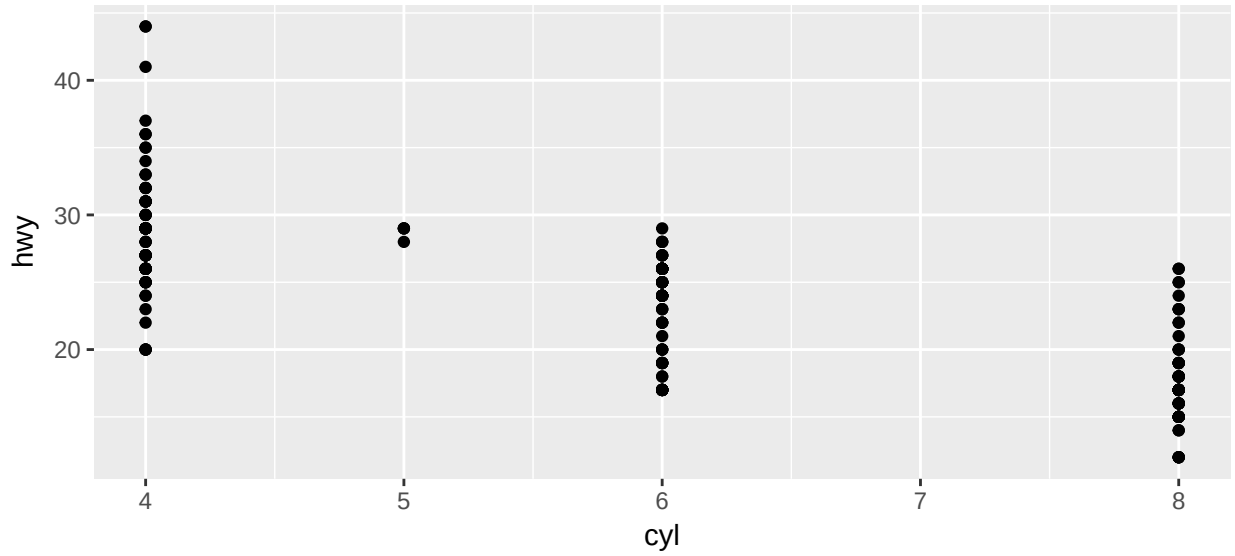
3. What does the `drv` variable describe? Read the help for `?mpg` to find out.

```
?mpg
```

Looking at the documentation of the data, `drv` shows if the car is rear-wheel drive (r) or front-wheel drive (f) or four wheel drive (4).

4. Make a scatterplot of `hwy` vs `cyl`.

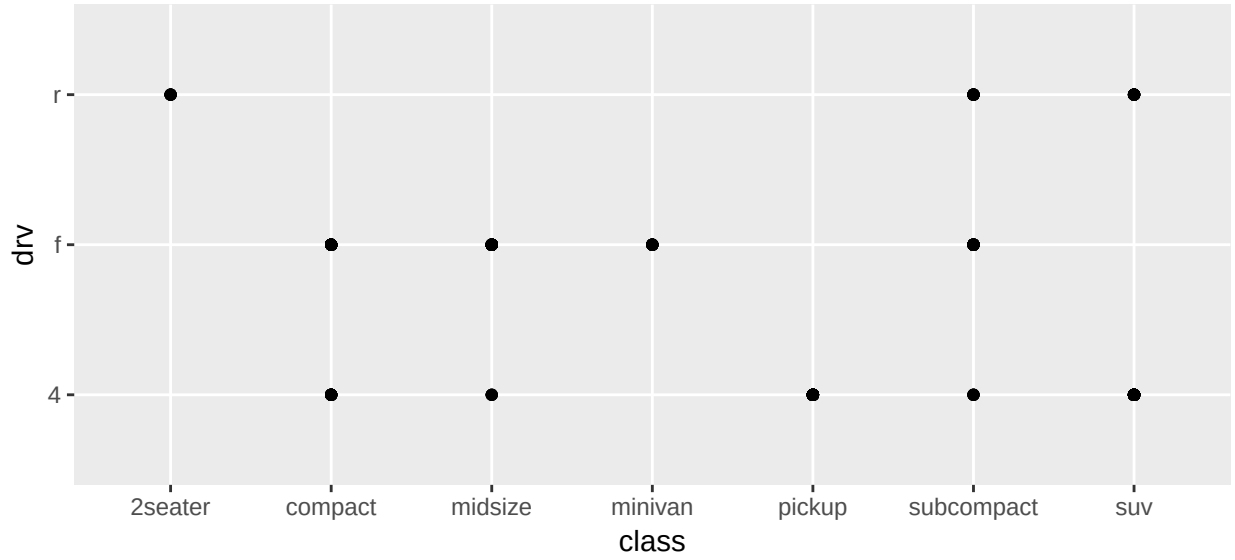
```
ggplot(data = mpg) +
  geom_point(mapping =  aes(cyl, hwy))
```



It apprers that there is a negative relationship between number of cylanders and fuel efficiency.

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(class, drv))
```



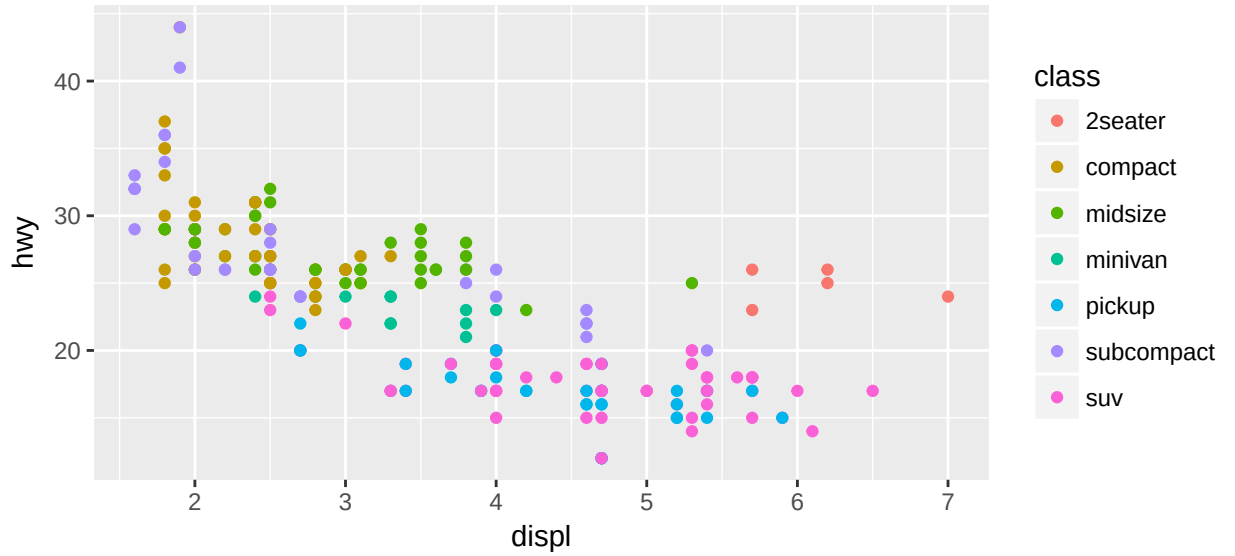
This graph somewhat shows the coverage of the dataset in terms of the class and powered axle. It is not very helpful for a few reasons:

- Multiple datapoints may overlap, and thus, it's impossible to tell if there is only one datapoint at each intersection or many.
- Although it might suggest that certain classes have certain specifics in terms of powered axles, but these results may not be conclusive.
- There is no specific order that either axis could be organized with.

3.3 Aesthetic mappings

In this section I explored different *aesthetics*, e.g. size, shape and color, to visalize the `mpg` dataset.

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



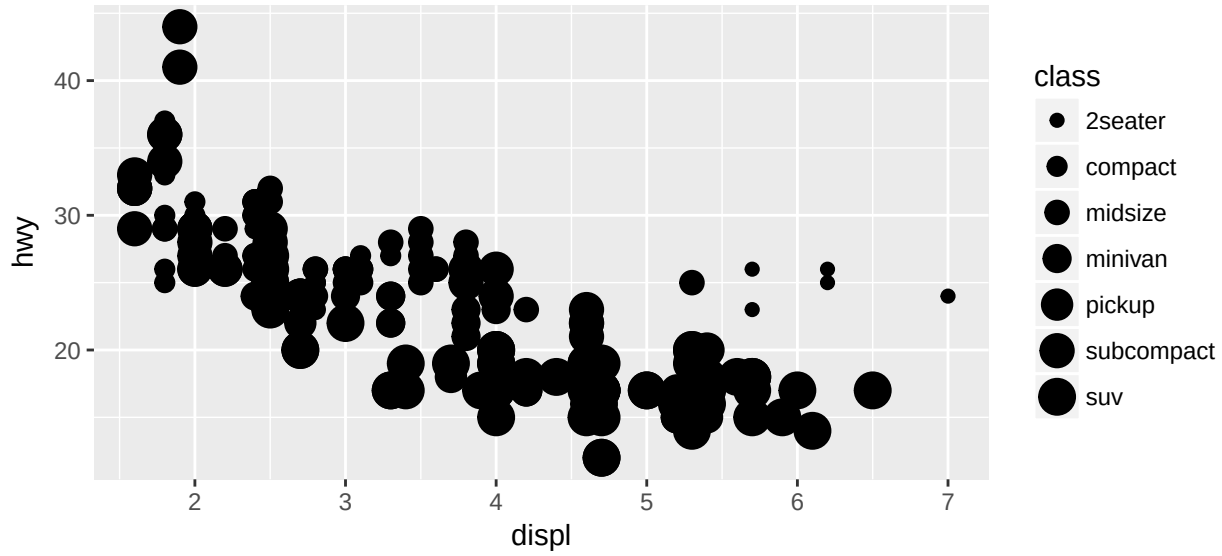
For instance the above graph categorizes the relationship between engine size and the fuel efficiency for

different classes of vehicles. And with this, we are able to explain that the cars with large engine and high fuel efficiency are mostly the 2seaters.

Similarly, instead of color, we can work with size, shape, and alpha, shown respectively below from top to buttom.

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

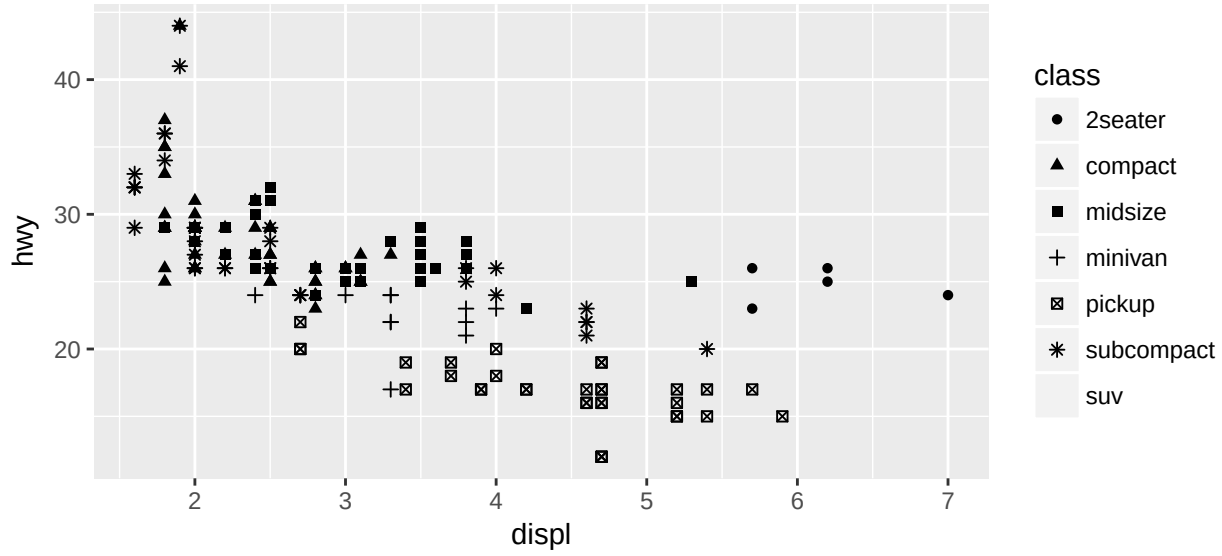
```
## Warning: Using size for a discrete variable is not advised.
```



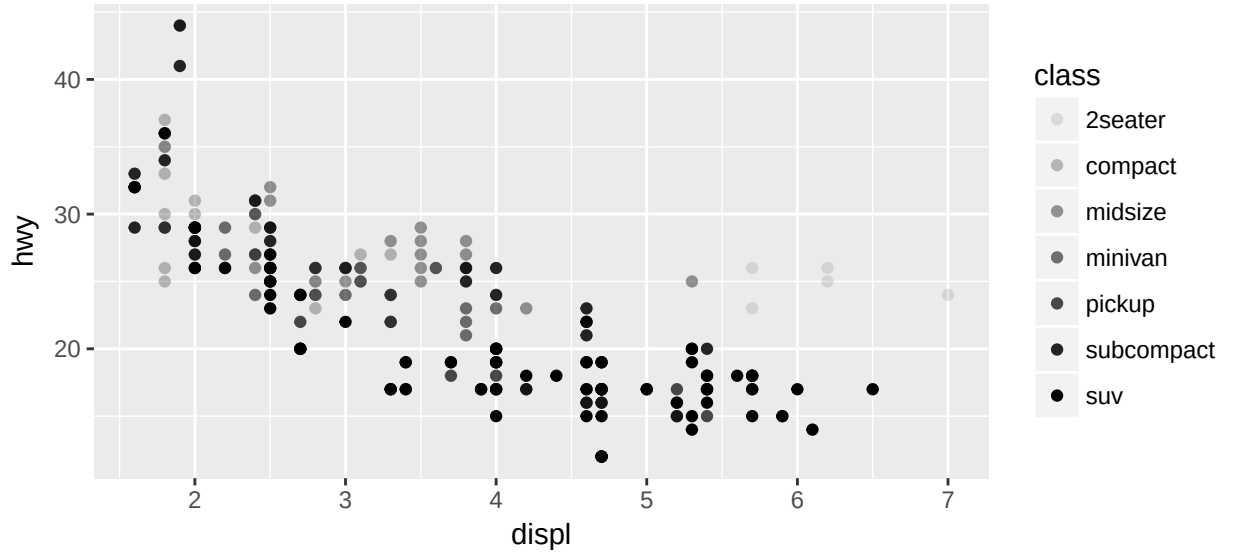

```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy, shape = class))
```

```
## 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).
```



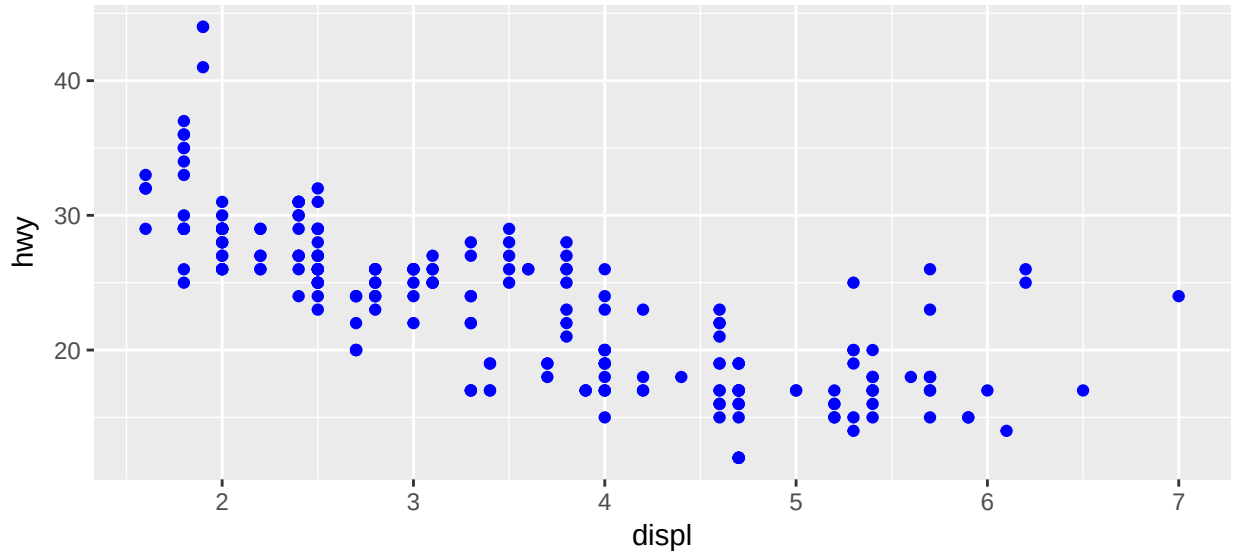
```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy, alpha = class))
```



And of course, there will be a warning for using ordered aesthetic for unordered variable. In addition, ggplot2 suggests only 6 different shapes in each plot.

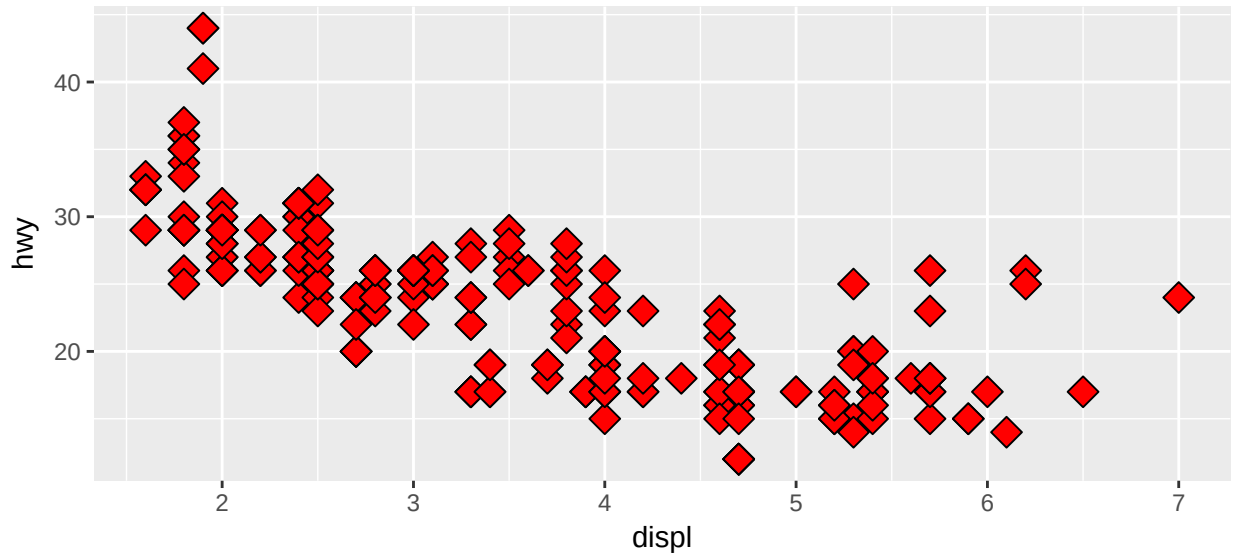
The aesthetic can also be controlled manually, but outside of the `aes()`.

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



or with a different color, size, and shape.

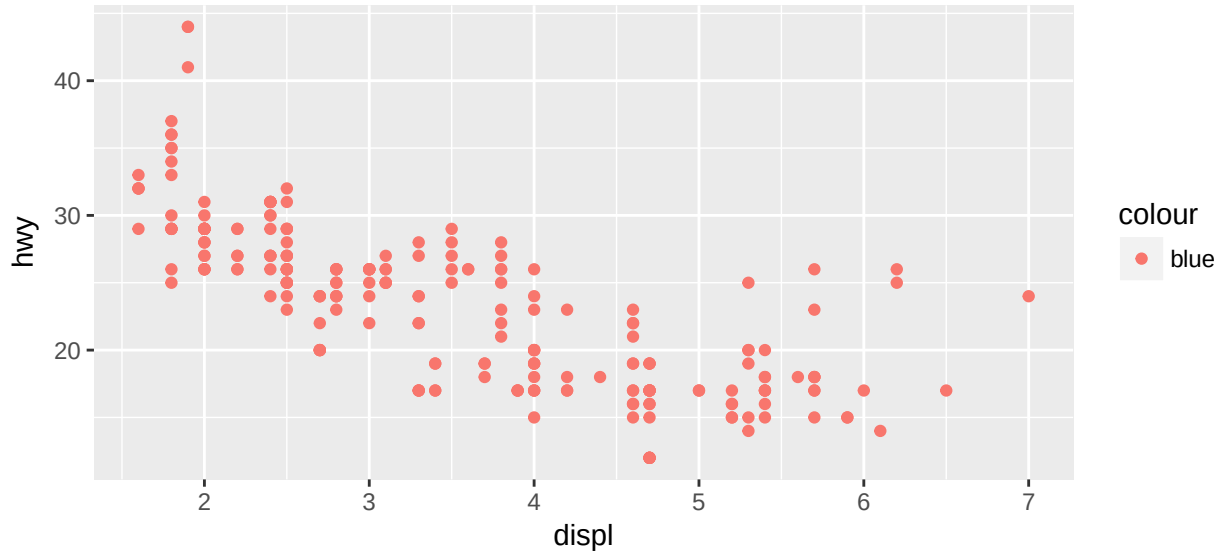
```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy), color = "black", fill="red", size = 4, shape = 23)
```



3.3.1 Exercises

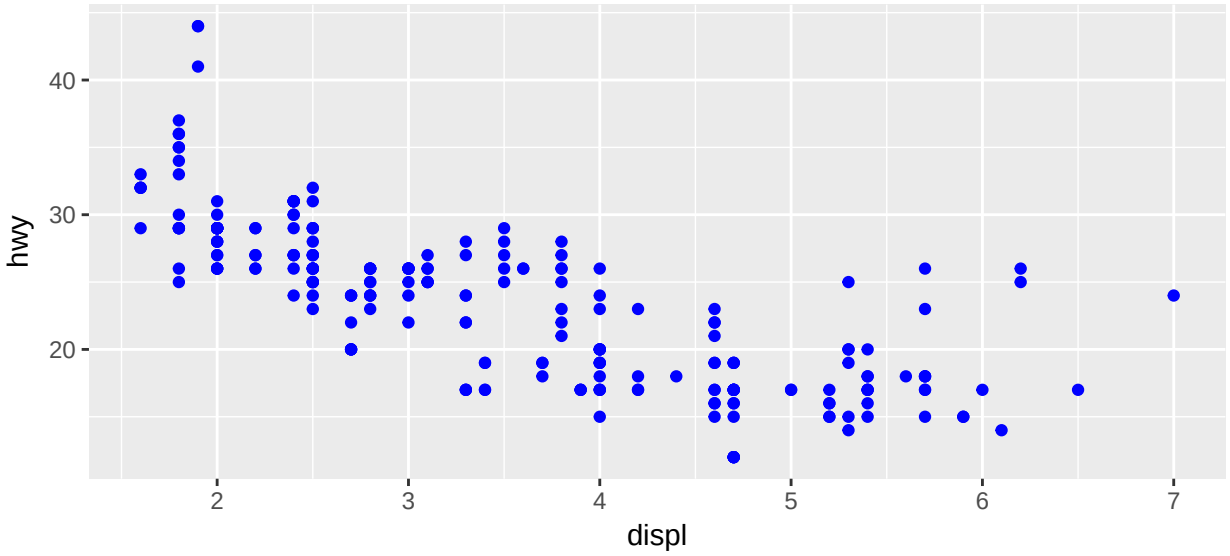
1. What's gone wrong with this code? Why are the points not blue?

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



The problem is that the `color` is defined inside the `aes()`. It has to be changed to the following.

```
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`?

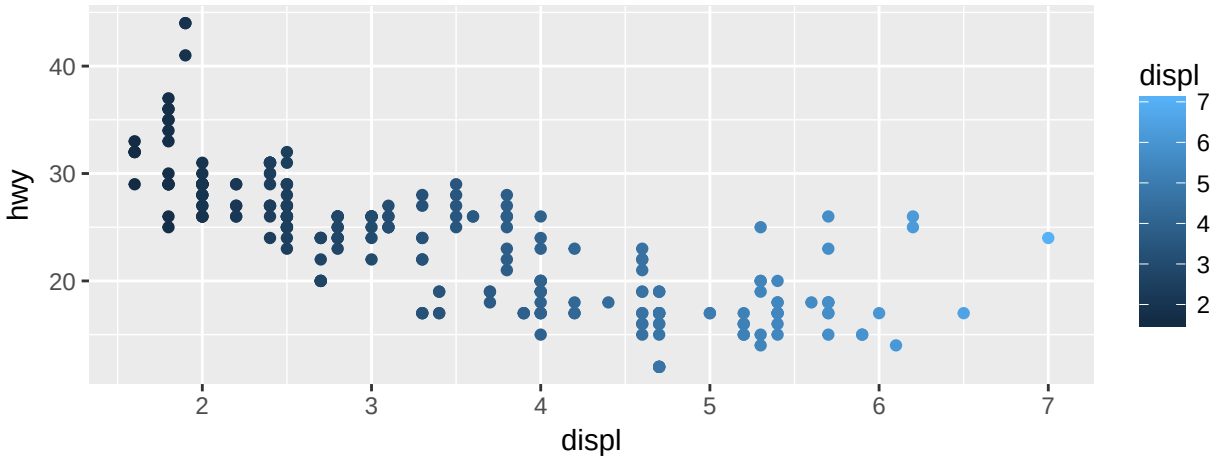
- Categorical varaibles: manufacturer, model, year, cyl, trsn, drv, fl, and class
- Continious variables: displ, cty, and hwy

Basically, if the variable is string, it's certainly categorical. But if it is intiger, it could be either categorical or continous. `?mpg` and `str(mpg)` help to differenciate categorical and continous variables.

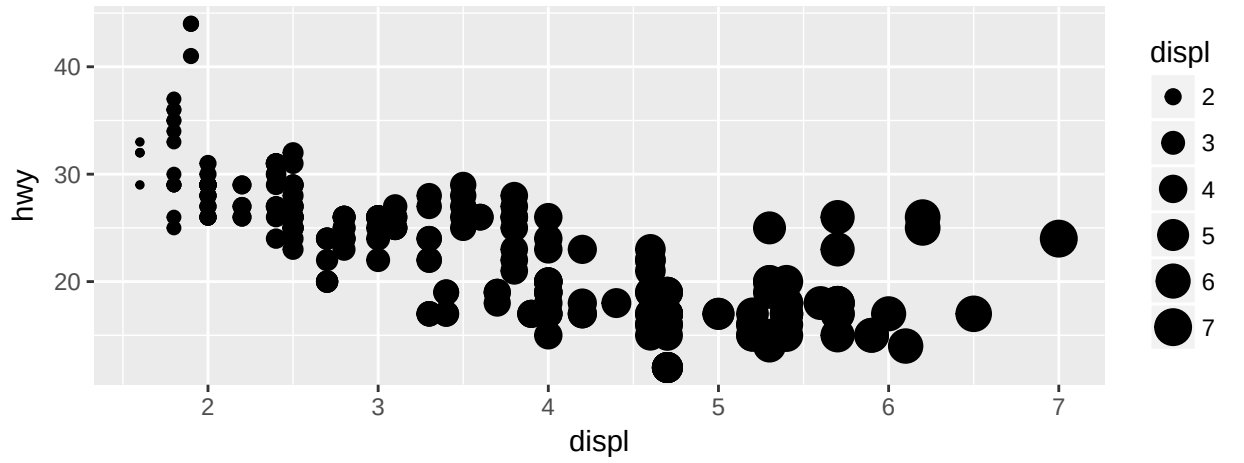
3. Map a continuous variable to `color`, `size`, and `shape`. How do these aesthetics behave differently for categorical vs. continuous variables?

I mapped `displ` to these `color`, `size`, and `shape` repectively.

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




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




```
## Throws an error!
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy, shape = displ))
```

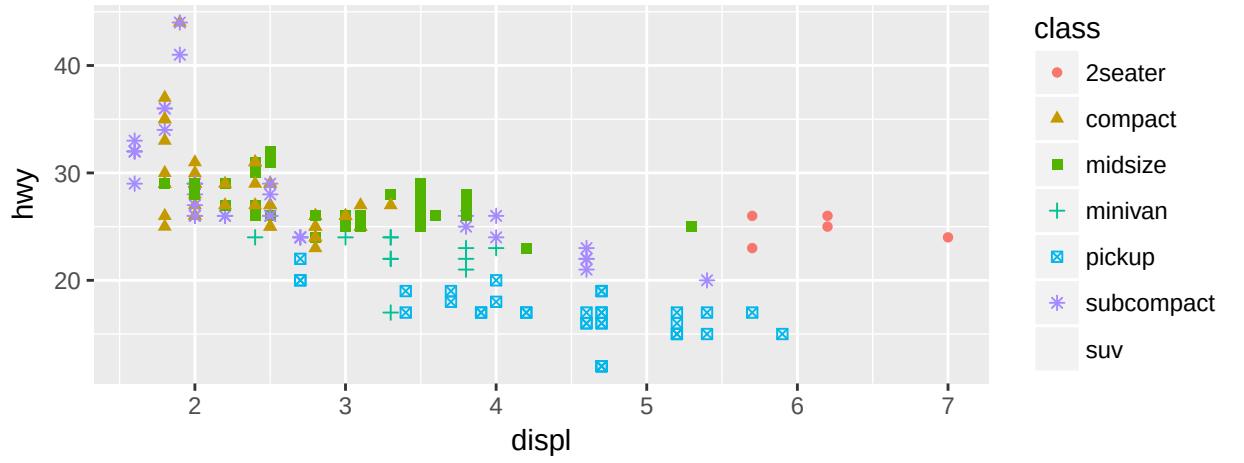
Mapping a continous variable to shape throws an error as shape is an unordered aesthetic. But it can be mapped to color or size. When mapped to color, it creates a color bar where each value is mapped to a certain shade of black and blue. And similarly, when mapped to size, the size of data point corresponds to the value of the mapped variable.

On the other hand, as we saw before, a categorical variable (e.g. `class`) could be mapped to either shape, color, size, or even alpha without any issue. However, as mentioned before, mapping an unordered variable to an ordered aethetic like size is not advised.

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

It just creates redundacny. In other words, you can read the value of single variable in two different ways. For instance, the class of vehicle can be read from both color and the shape of the data point the figure below.

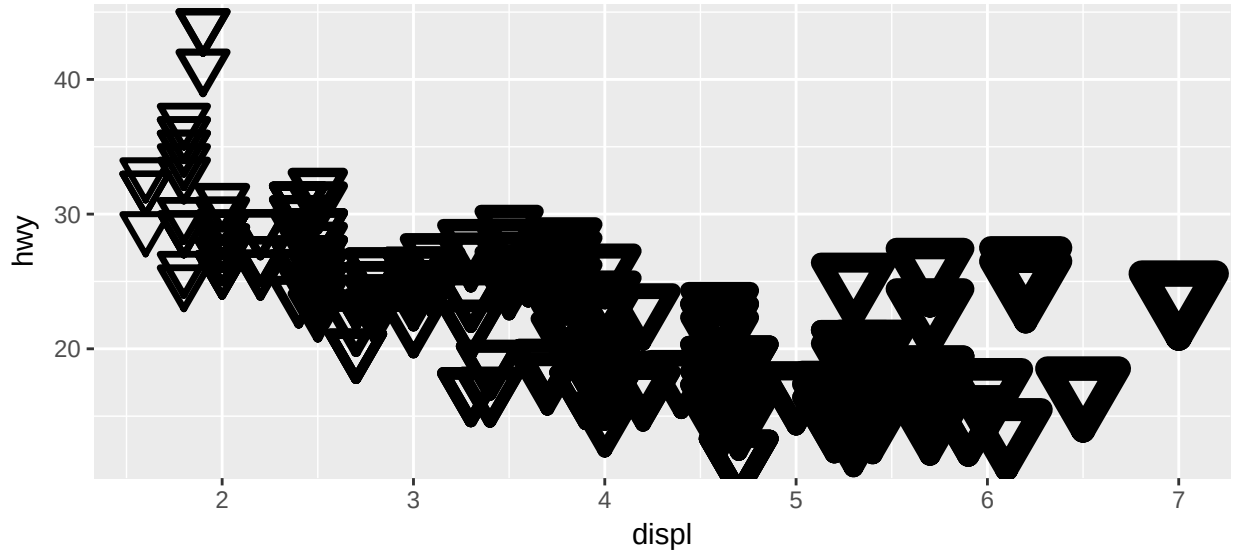
```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy, shape = class, color = class))
```



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

It changes the width of the border for the shapes that have a border.

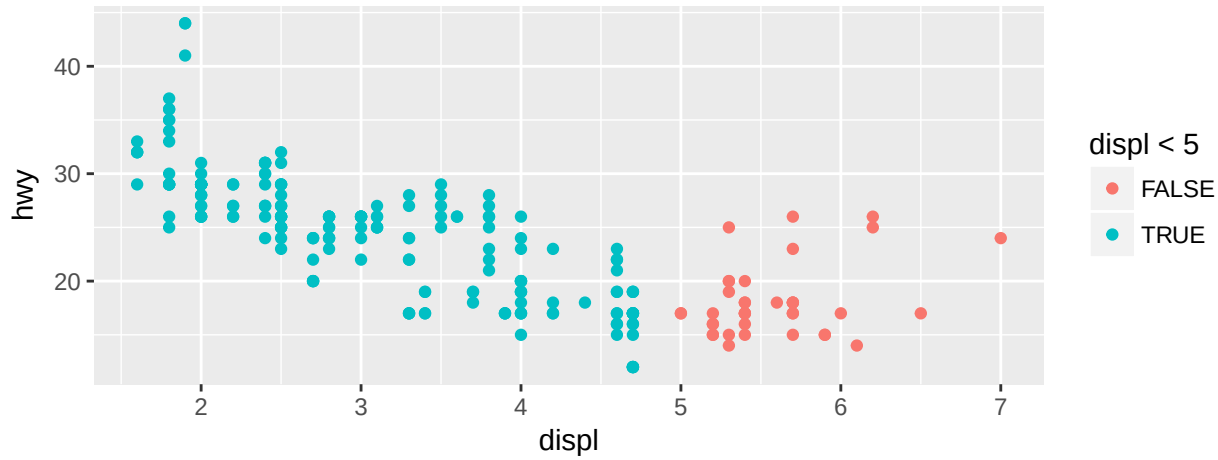
```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy, stroke = displ), shape = 6, size = 5)
```



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

It brakes the data in two different colors for value greater and less than the input threshold, in this case `displ <5` and `displ > 5`.

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



3.4 Common problems

Just make sure not to put the + sign at the beginning of the line. This will throw an error!

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