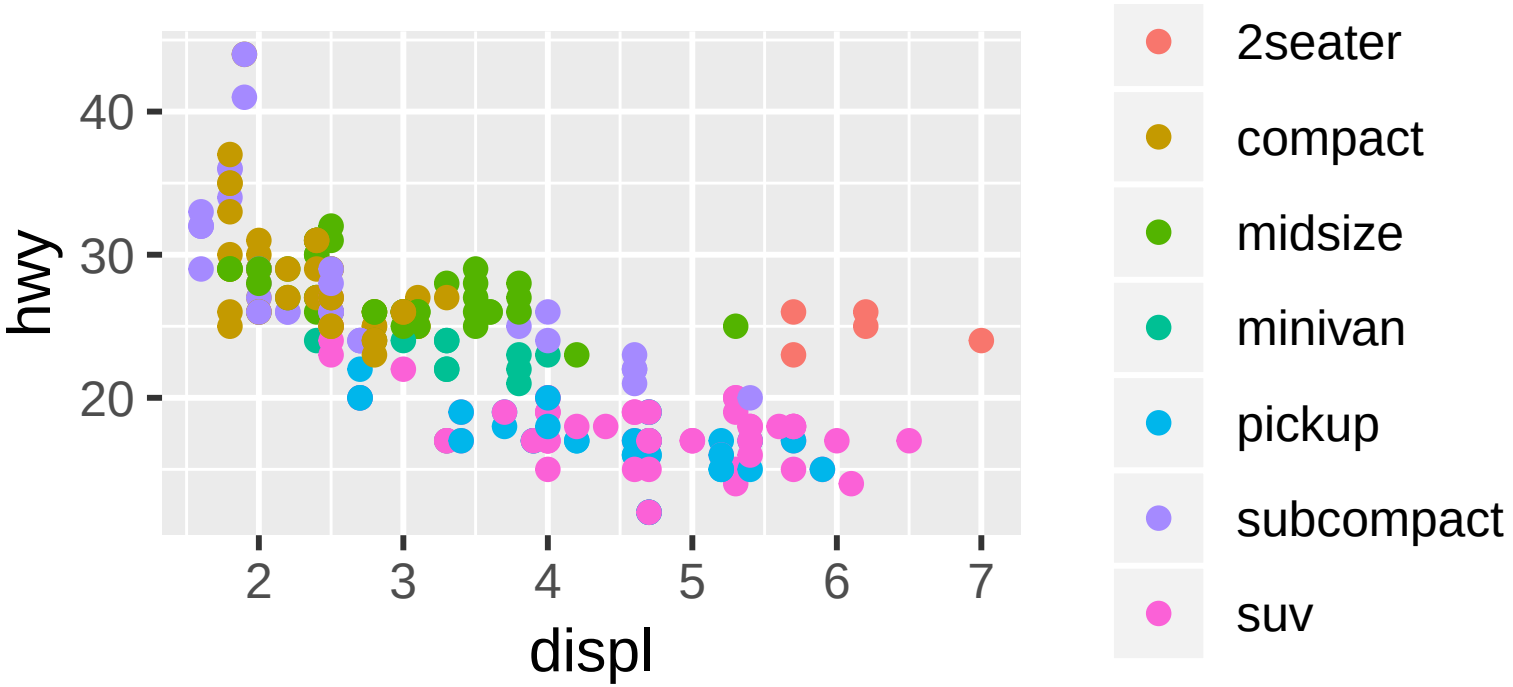
6. Exploring Data with ggplot2 (1)

CT1100 - J. Duggan

Data Exploration

"Data exploration is the art of looking at your data, rapidly generating hypotheses, quickly testing them, then repeating again and again and again." (Wickham and Grolemund 2017).



Data visualisation with ggplot2

"The simple graph has brought more information to the data analyst's mind that any other device." – John Tukey

```
d <- ggplot2::mpg  # get a copy of mpg
glimpse(d)  # show structure and some data
```

```
## Observations: 234
## Variables: 11
## $ manufacturer <chr> "audi", "audi", "audi", "audi", "audi"
## $ model        <chr> "a4", "a4", "a4", "a4", "a4", "a4", "a
## $ displ        <dbl> 1.8, 1.8, 2.0, 2.0, 2.8, 2.8, 3.1, 1.8
## $ year         <int> 1999, 1999, 2008, 2008, 1999, 1999, 20
## $ cyl          <int> 4, 4, 4, 4, 6, 6, 6, 4, 4, 4, 4, 6, 6,
## $ trans        <chr> "auto(l5)", "manual(m5)", "manual(m6)"
## $ drv          <chr> "f", "f", "f", "f", "f", "f", "f", "4"
## $ cty          <int> 18, 21, 20, 21, 16, 18, 18, 18, 16, 20
## $ hwy          <int> 29, 29, 31, 30, 26, 26, 27, 26, 25, 28
```

Fuel Economy Data Set (ggplot2::mpg)

This dataset contains a subset of the fuel economy data that the EPA makes available on http://fueleconomy.gov. It contains only models which had a new release every year between 1999 and 2008 - this was used as a proxy for the popularity of the car.

manufacturer	car manufacturer	**drv**	drive type
model	model name	**cty**	city miles per gallon
displ	engine disp (l)	**hwy**	highway miles per gallon
year	year of make	**fl**	fuel type
model	model name	**cty**	city miles per gallon
cyl	number of cylinders	**class**	"type" of car
trans	type of transm.		

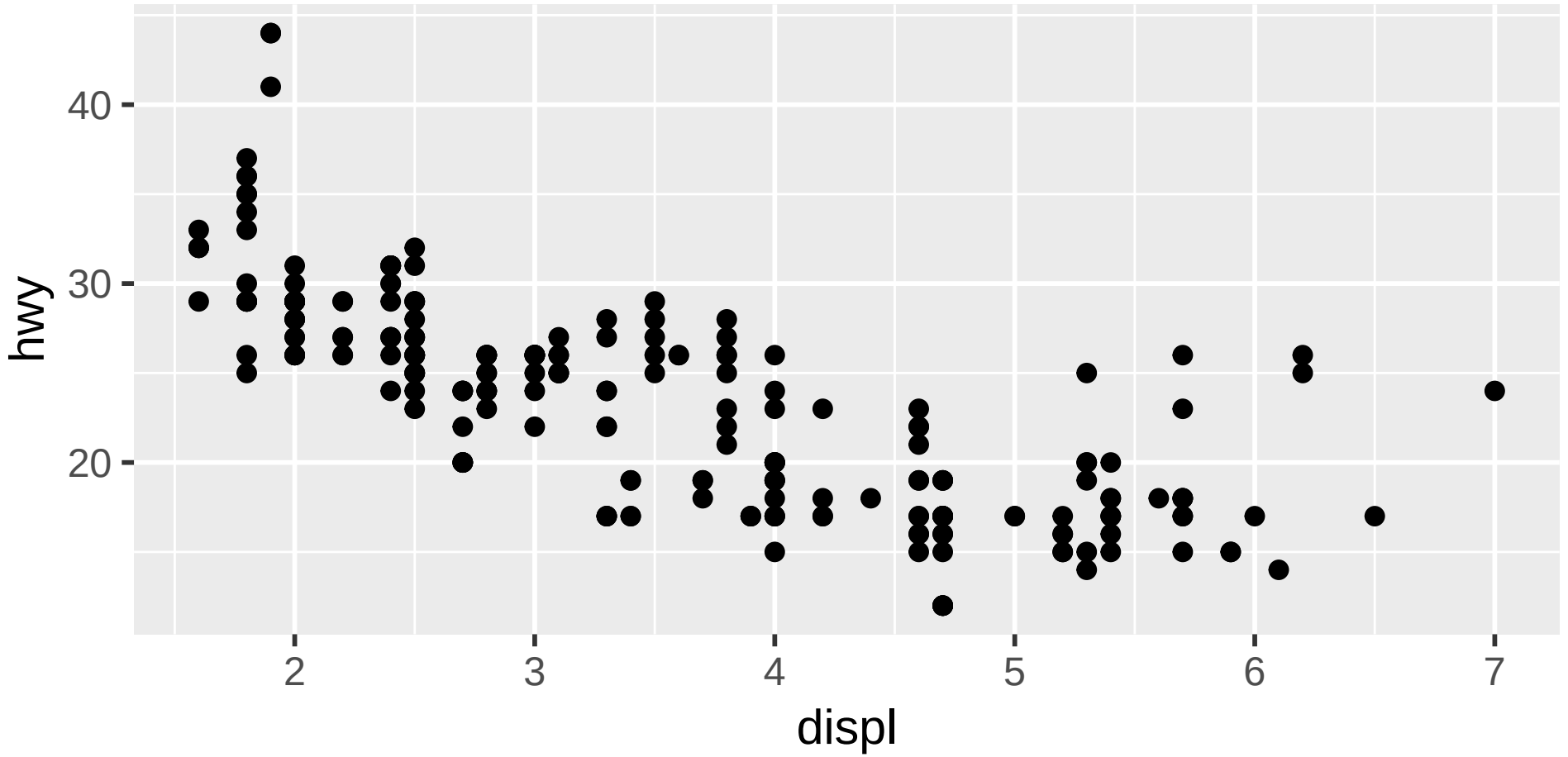
Exploring Data

Generate a first graph to help answer the following question

- Do cars with big engines use more fuel than cars with small engines
- What might the relationship between engine size and fuel efficiency look like?
  - Positive or negative?
  - Linear or non-linear?
- Variable (scatter plot)
  - **displ**, a car engine size in litres (x)
  - **hwy**, a car's fuel efficiency on highway - (y)
- ggplot2: layered approach
  - ggplot(data=*tibble_name*) + geom_point(mapping=aes(x=*col1*,y=*col2*))

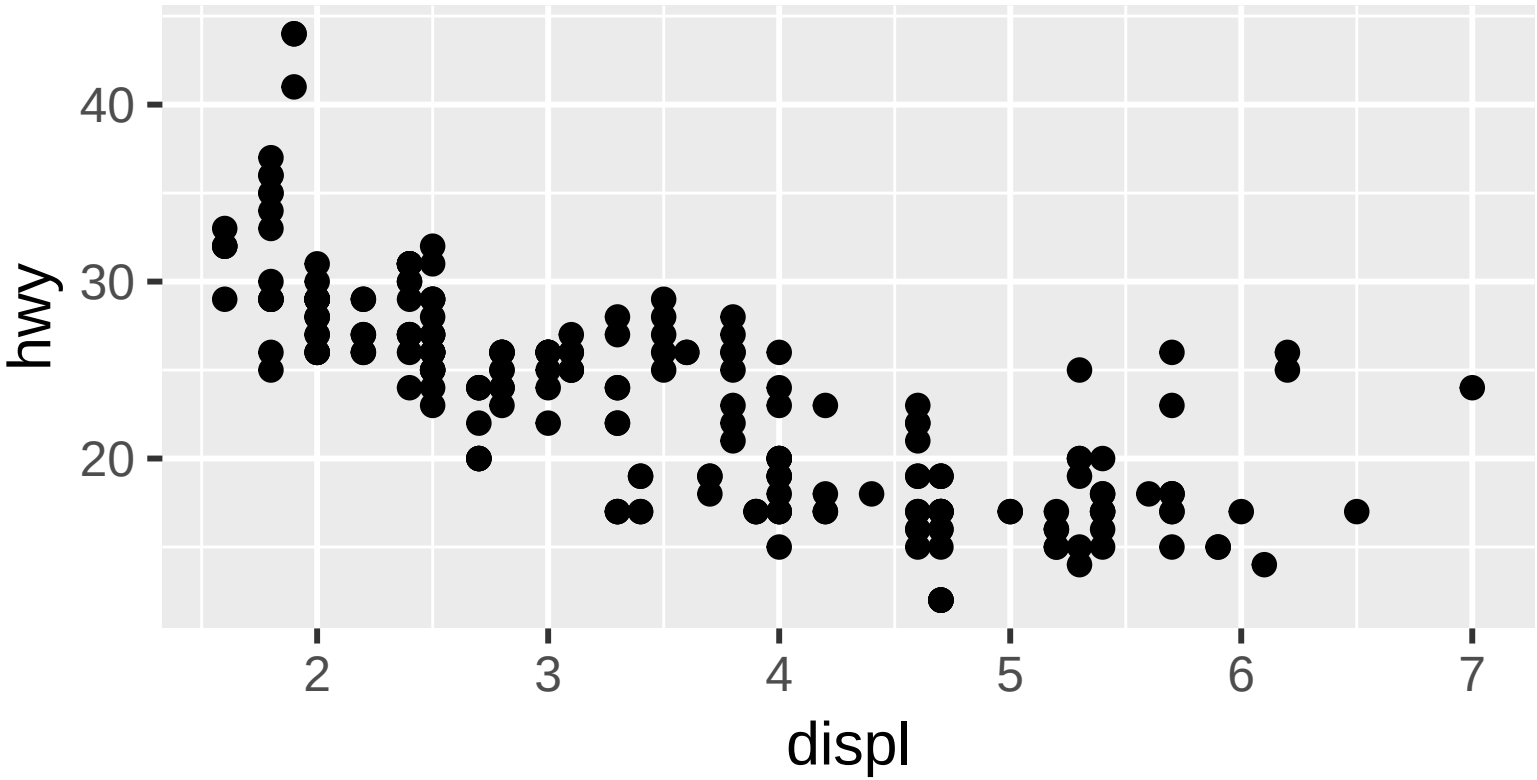
Plotting with ggplot2

```
ggplot(data = d) +    # specify the source tibble
  geom_point(mapping=aes(x=displ,   # map x, y vars
                         y=hwy))
```



Interpreting the plot

- The plot shows a negative relationship between engine size (displ) and fuel efficiency (hwy)
- Cars with big engines use more fuel
- Does this confirm or refute your hypothesis about fuel efficiency and engine size?



Aesthetic Mappings

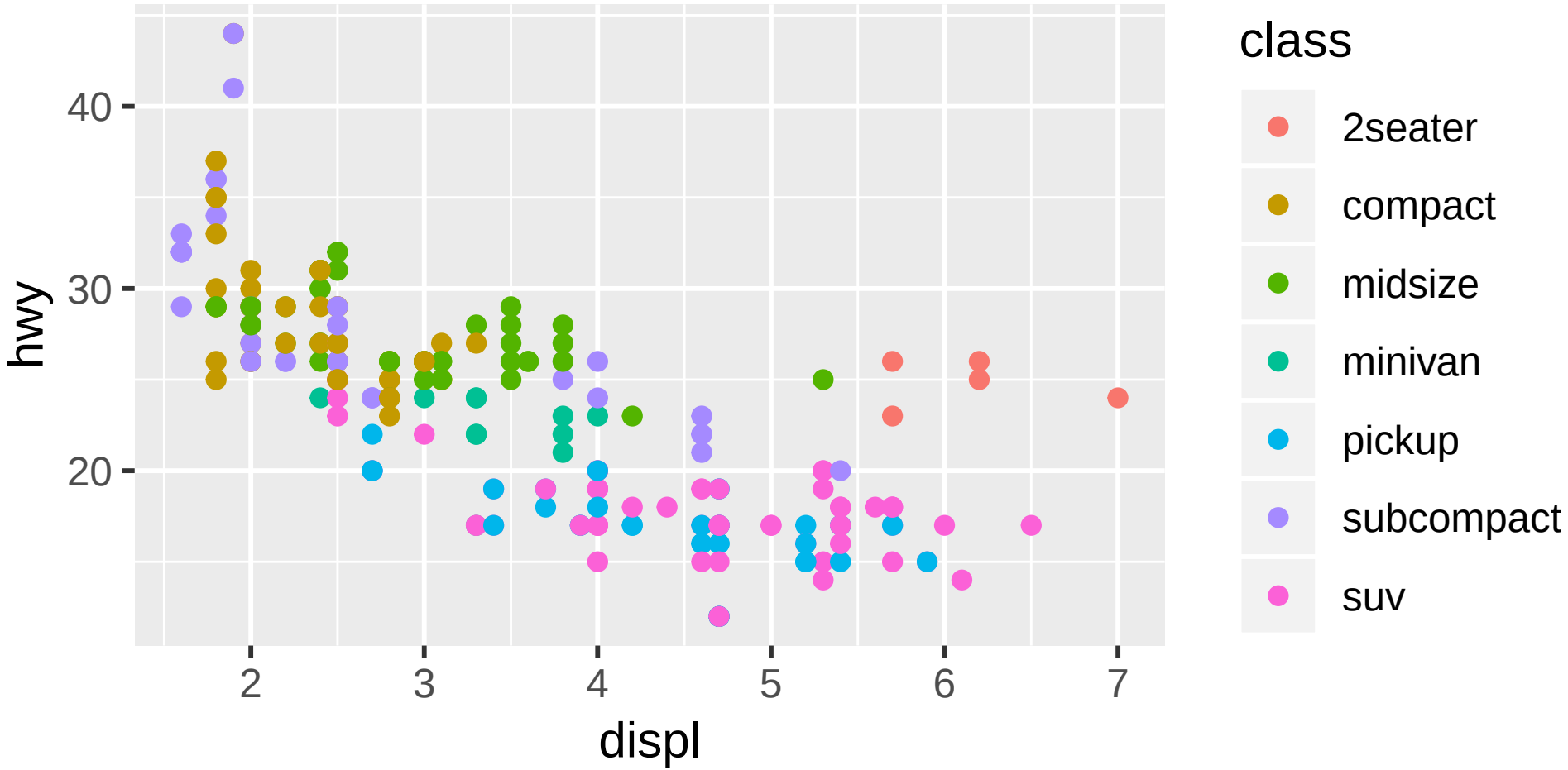
- A third variable can be added to a 2-D plot by mapping it to an aesthetic.
- An aesthetic is a visual property of the plot's objects.
- An aesthetic's level could be colour, size or shape

```
unique(d$class)
```

```
## [1] "compact"    "midsize"    "suv"        "2seater"    "mi
## [6] "pickup"     "subcompact"
```

In ggplot2 - Adding the third variable

```
ggplot(data=d)+
  geom_point(aes(x=displ,y=hwy,colour=class))
```

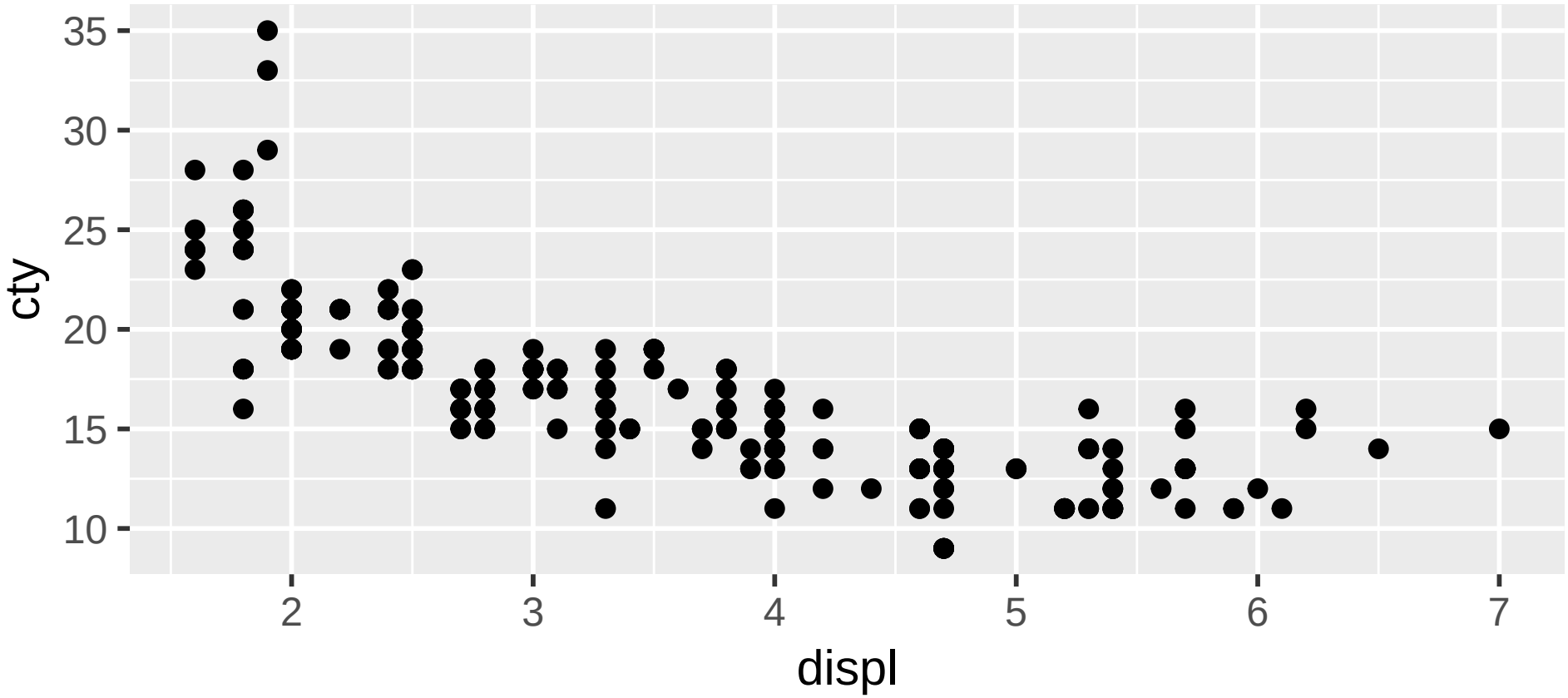


Exploring Data Relationships

Input (X)	Output (Y)	Hypothesis?	Reason
Displacement	City MPG	Negative?	Bigger cars,less efficient
Highway MPG	City MPG	Positive?	Should be closely related
Cylinders	Highways MPG	Negative?	More cylinders, less eff.
Cylinders	Displacement	Negative?	More cylinders, bigger eng.

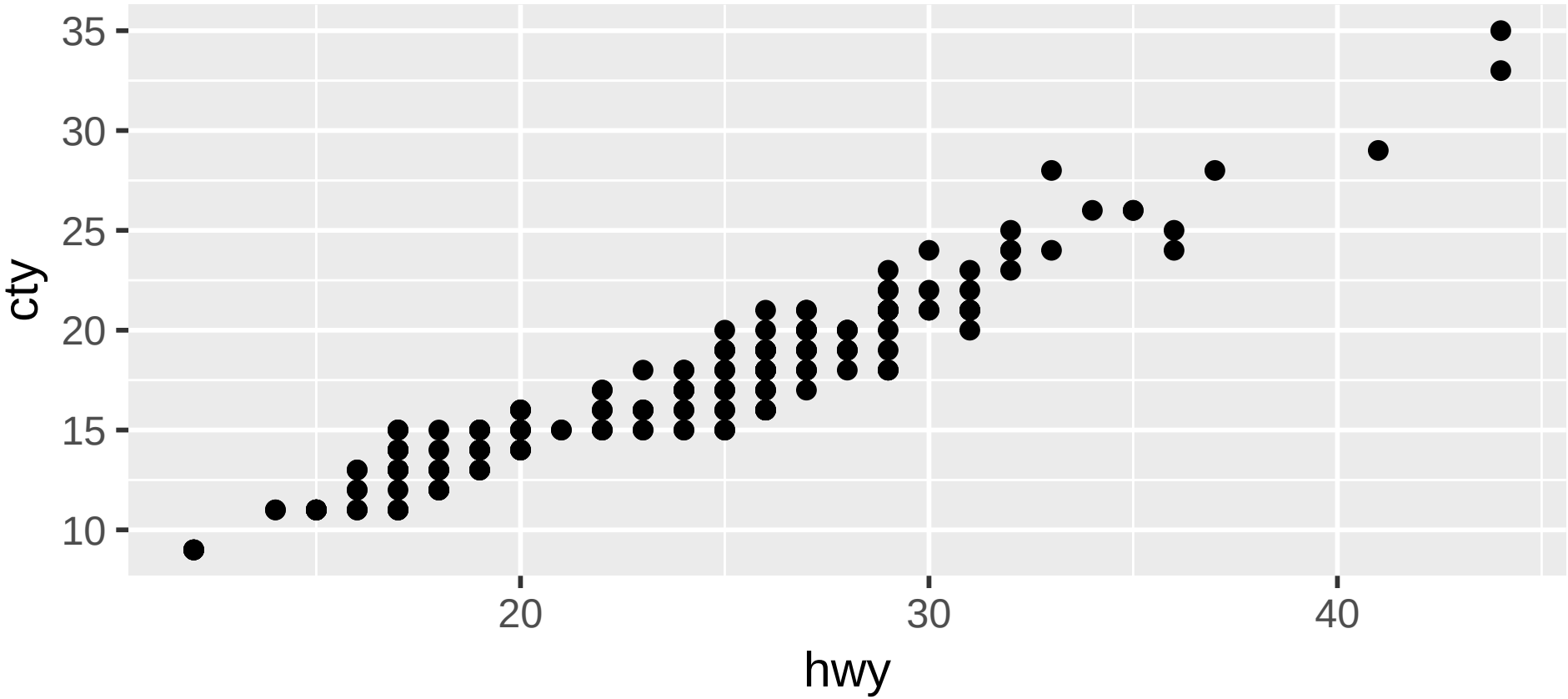
x=displ, y=cty

```
ggplot(data = mpg) +    # specify the source tibble
  geom_point(mapping=aes(x=displ,   # map x, y vars
                         y=cty))
```



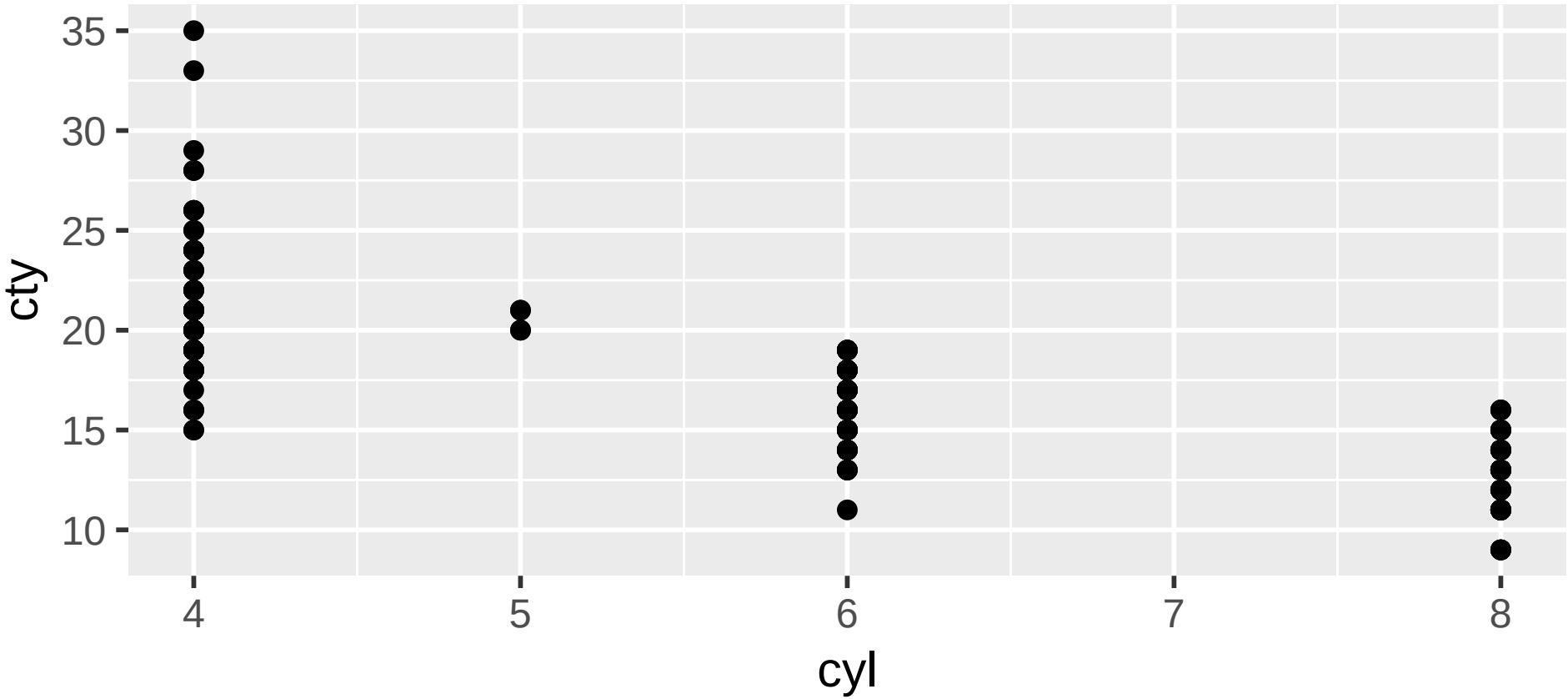
x=hwy, y=cty

```
ggplot(data = mpg) +    # specify the source tibble
  geom_point(mapping=aes(x=hwy,   # map x, y vars
                         y=cty))
```



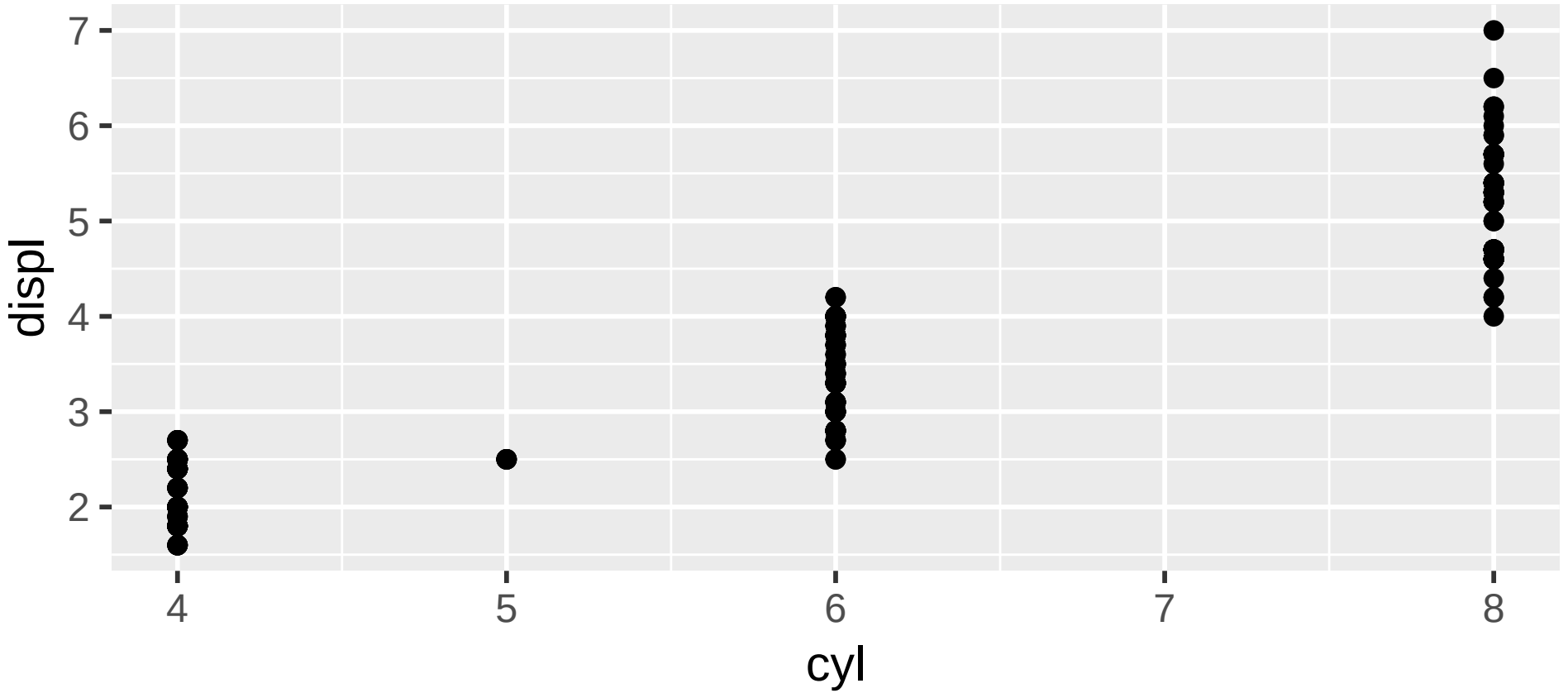
x=cyl, y=cty

```
ggplot(data = mpg) +    # specify the source tibble
  geom_point(mapping=aes(x=cyl,   # map x, y vars
                         y=cty))
```



x=cyl, y=displ

```
ggplot(data = mpg) +    # specify the source tibble
  geom_point(mapping=aes(x=cyl,  # map x, y vars
                         y=displ))
```



Challenge 6.1

- Redraw the graphs, and colour by **car class**
- Vary the size of the point by using the number of cylinders

Input (X)	Output (Y)	Hypothesis?	Reason
Displacement	City MPG	Negative?	Bigger cars,less efficient
Highway MPG	City MPG	Positive?	Should be closely related
Cylinders	Highways MPG	Negative?	More cylinders, less eff.
Cylinders	Displacement	Negative?	More cylinders, bigger eng.

Summary

- "The simple graph has brought more information to the data analyst's mind that any other device." – John Tukey]
- "Data exploration is the art of looking at your data, rapidly generating hypotheses, quickly testing them, then repeating again and again and again." (Wickham and Grolemund 2017).
- ggplot2 provides a layered approach to building charts