

A solution to ggplot2-book

Raju Rimal

2017-05-19

Contents

Prerequisites	5
1 Getting Started	7
1.1 Fuel economy data	7
1.2 Key components	10
1.3 Colour, size, shape and other aesthetic attributes	12
1.4 Facetting	13

Prerequisites

This is a solution to the problems in ggplot2-book. In order to run all the solution, following packages need to be installed and loaded.

```
# devtools::install_github("hadley/tidyverse")
pkgs <- c("ggplot2", "dplyr", "pander")
for (pkg in pkgs) require(pkg, character.only = TRUE)
```


Chapter 1

Getting Started

1.1 Fuel economy data

1.1.1 Exercise 2.2.1

1. List five functions that you could use to get more information about the mpg dataset. `str`, `summary`
2. How can you find out what other datasets are included with `ggplot2`?

```
data_ggplot <- data(package = "ggplot2")
pander(data_ggplot$result[, -c(1:2)], justify = "rl",
        split.cells = 50, emphasize.verbatim.cols = 1,
        caption = "(\\#tab:ggplot2-data) Datasets in ggplot2 package")
```

Table 1.1: Datasets in ggplot2 package

Item	Title
diamonds	Prices of 50,000 round cut diamonds
economics	US economic time series
economics_long	US economic time series
faithfuld	2d density estimate of Old Faithful data
luv_colours	'colors()' in Luv space
midwest	Midwest demographics
mpg	Fuel economy data from 1999 and 2008 for 38 popular models of car
msleep	An updated and expanded version of the mammals sleep dataset
presidential	Terms of 11 presidents from Eisenhower to Obama
seals	Vector field of seal movements
txhousing	Housing sales in TX

3. Apart from the US, most countries use fuel consumption (fuel consumed over fixed distance) rather than fuel economy (distance travelled with fixed amount of fuel). How could you convert `cty` and `hwy` into the European standard of l/100km?

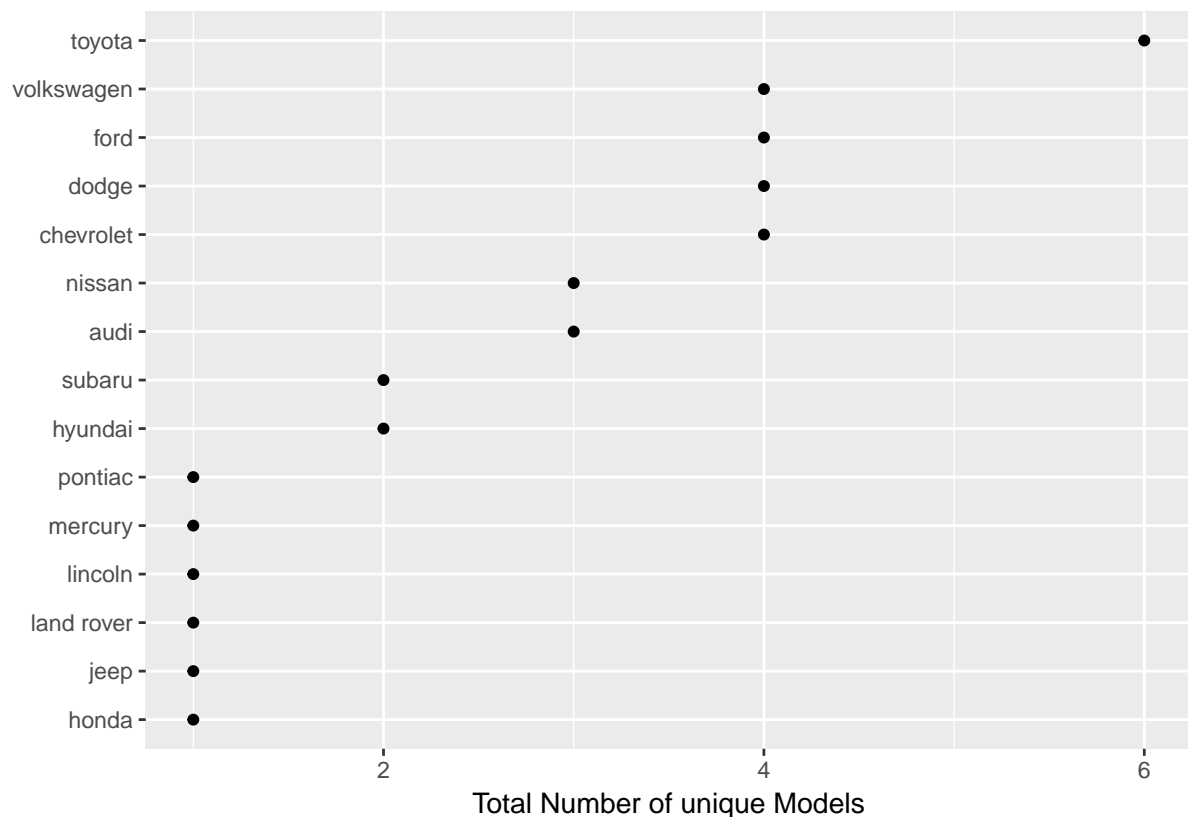
```
litre_per_km <- function(mile_per_gallon) {
  return(3.78541 / (1.60934 * mile_per_gallon))
}
```

```
mpg_eu <- mpg %>%
  mutate(cty = litre_per_km(cty) * 100,
         hwy = litre_per_km(hwy) * 100)
```

4. Which manufacturer has the most the models in this dataset? Which model has the most variations? Does your answer change if you remove the redundant specification of drive train (e.g. “pathfinder 4wd”, “a4 quattro”) from the model name?

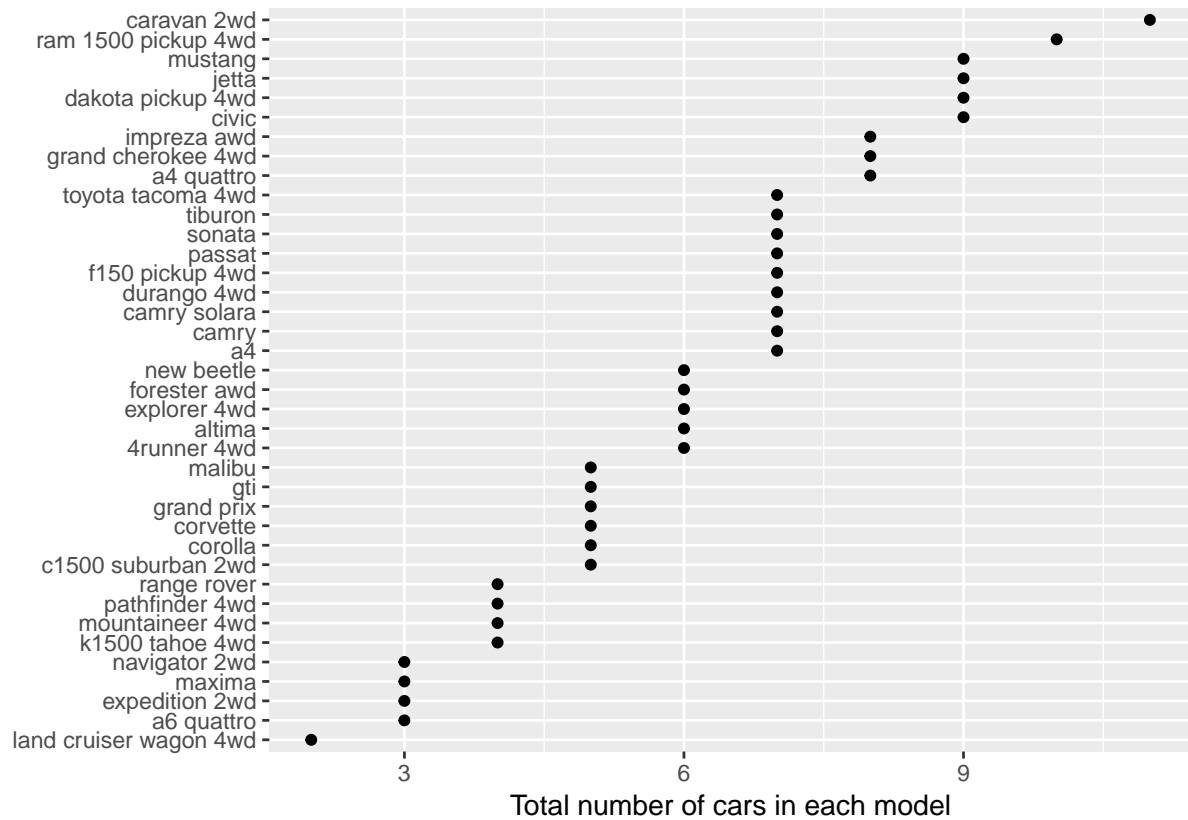
Manufacturer with Most models:

```
mpg %>%
  group_by(manufacturer) %>%
  summarize(model_count = length(unique(model))) %>%
  arrange(-model_count) %>%
  ggplot(aes(model_count, reorder(manufacturer, model_count))) +
  geom_point() +
  labs(x = "Total Number of unique Models", y = NULL)
```



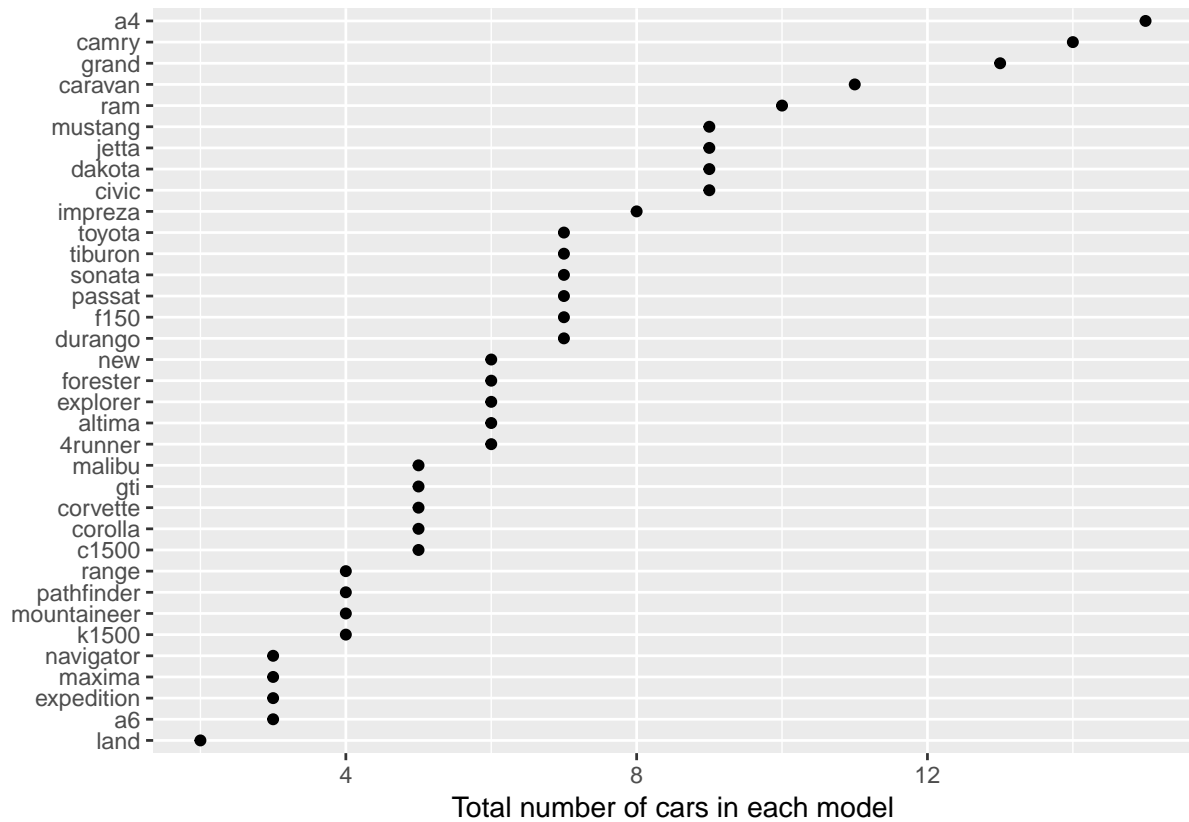
Model with most variations:

```
mpg %>%
  mutate(model_trim = gsub("[a-zA-Z0-9]+.*", "\\1", model)) %>%
  group_by(model) %>%
  summarize(variation = length(model)) %>%
  arrange(-variation) %>%
  ggplot(aes(variation, reorder(model, variation))) +
  geom_point() +
  labs(x = "Total number of cars in each model", y = NULL)
```

Variation after removing redundant specification:

```
mpg %>%
  mutate(model_trim = gsub("[a-zA-Z0-9]+.*", "\\1", model)) %>%
  group_by(model_trim) %>%
  summarize(variation = length(model)) %>%
  arrange(-variation) %>%
  ggplot(aes(variation, reorder(model_trim, variation))) +
  geom_point() +
  labs(x = "Total number of cars in each model", y = NULL)
```

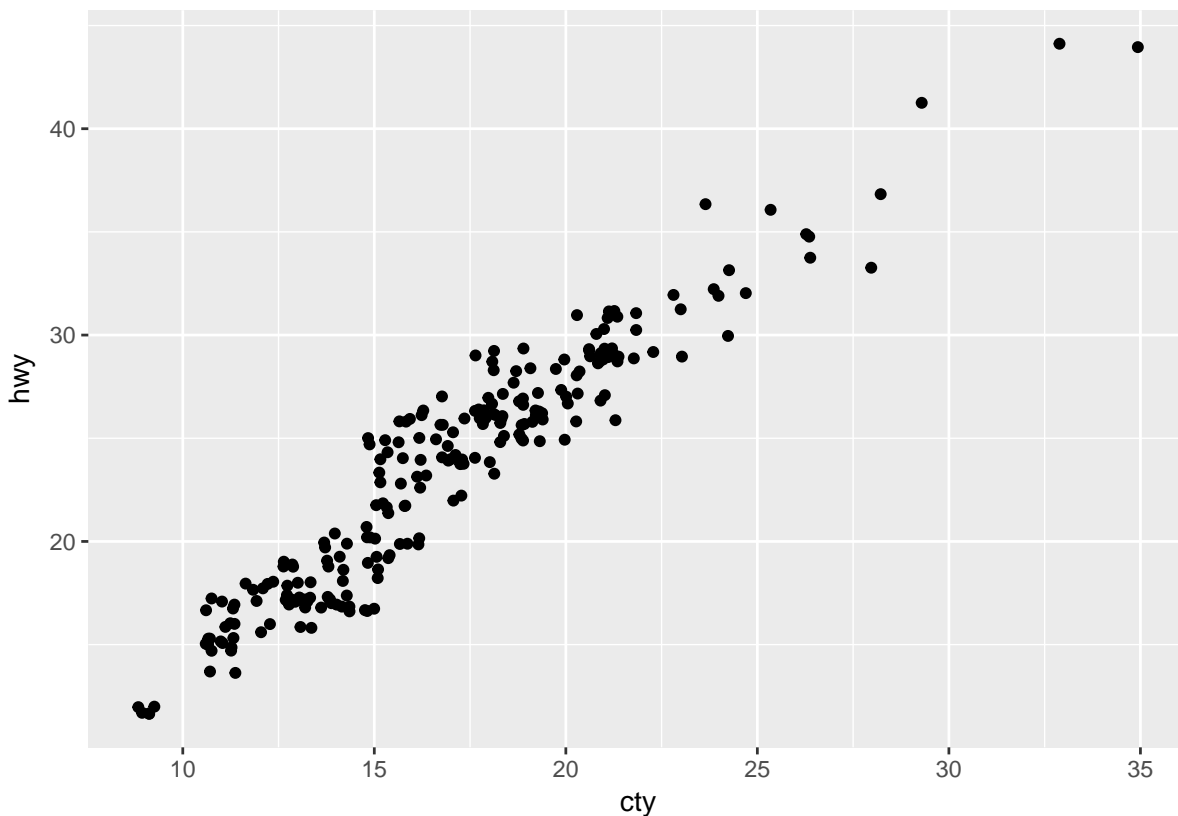


1.2 Key components

1.2.1 Exercises 2.3.1

1. How would you describe the relationship between `cty` and `hwy` ? Do you have any concerns about drawing conclusions from that plot?

```
ggplot(mpg, aes(cty, hwy)) + geom_point(position = "jitter")
```



It's a linear relationship.

2. What does `ggplot(mpg, aes(model, manufacturer)) + geom_point()` show? Is it useful? How could you modify the data to make it more informative?

Not very informative but shows how many models a manufacturer has. It is more visible if we use `geom_count` which uses the count summary statistics for each combination of `model` and `manufacturer`.

3. Describe the data, aesthetic mappings and layers used for each of the following plots. You'll need to guess a little because you haven't seen all the datasets and functions yet, but use your common sense! See if you can predict what the plot will look like before running the code.

- a. `ggplot(mpg, aes(cty, hwy)) + geom_point()`

Data is `mpg`, x and y axis are mapped to `cty` and `hwy` variables and a layer of `point` is added.

- b. `ggplot(diamonds, aes(carat, price)) + geom_point()`

Data is `diamonds`, x and y axis are mapped to `carat` and `price` variables and a layer of `point` is added.

- c. `ggplot(economics, aes(date, unemploy)) + geom_line()`

Data is `economics`, x and y axis are mapped to `date` and `unemploy` variables and a layer of `line` is added.

- d. `ggplot(mpg, aes(cty)) + geom_histogram()`

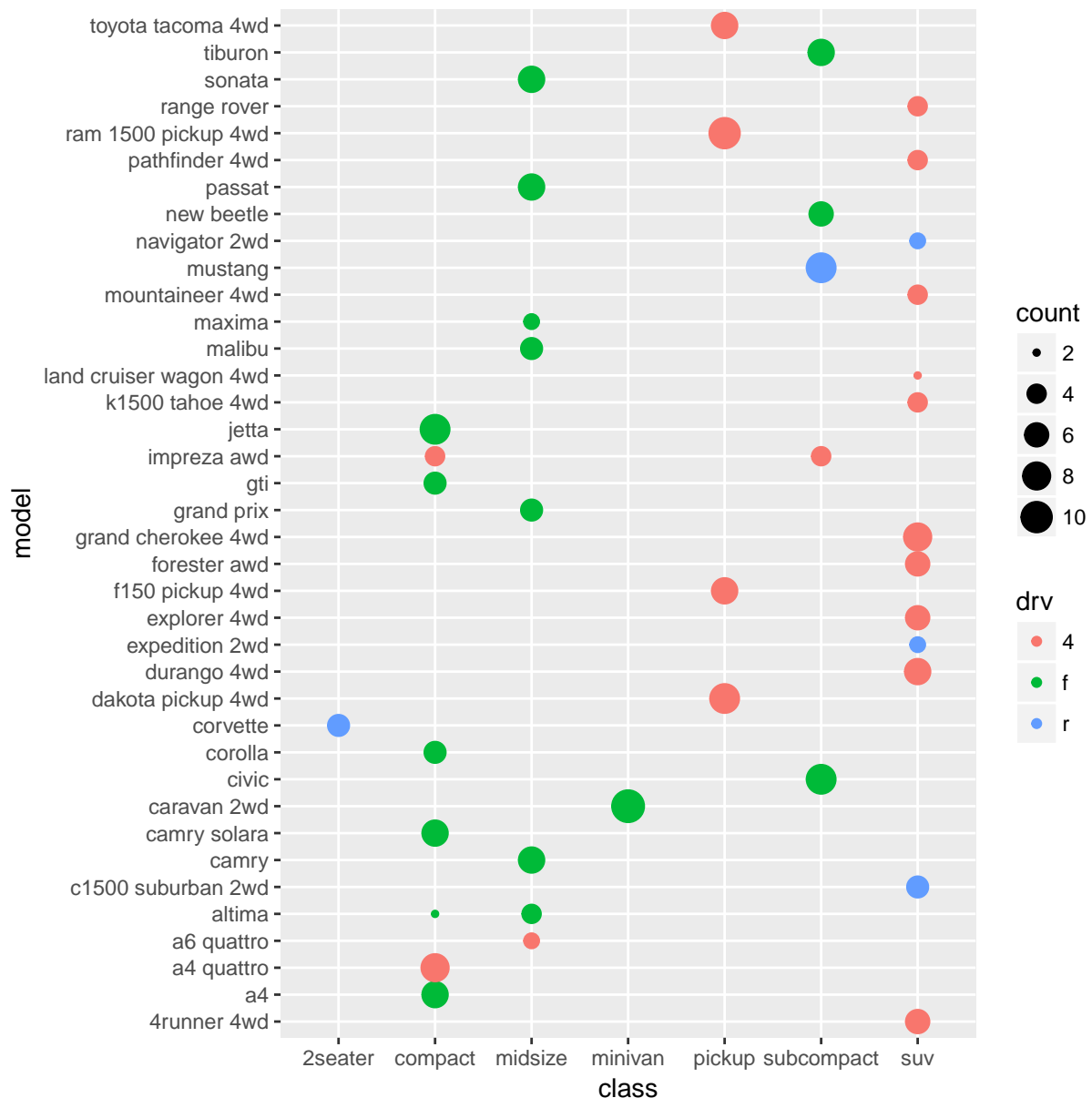
Data is `mpg`, x axis are mapped to `cty` and a layer of `histogram` is added which uses the default bins of 30.

1.3 Colour, size, shape and other aesthetic attributes

1.3.1 Exercises 2.4.1

1. Experiment with the colour, shape and size aesthetics. What happens when you map them to continuous values? What about categorical values? What happens when you use more than one aesthetic in a plot?
2. What happens if you map a continuous variable to shape? Why? What happens if you map trans to shape? Why?
3. How is drive train related to fuel economy? How is drive train related to engine size and class?

```
mpg %>%  
  group_by(model, class, drv) %>%  
  summarize(count = n()) %>%  
  ggplot(aes(class, model, color = drv, size = count)) +  
  geom_point()
```



Here, we see that suv and pickup has mostly 4 wheel drive while rest are front-wheeled drive.

1.4 Facetting

1.4.1 Exercises 2.5.1

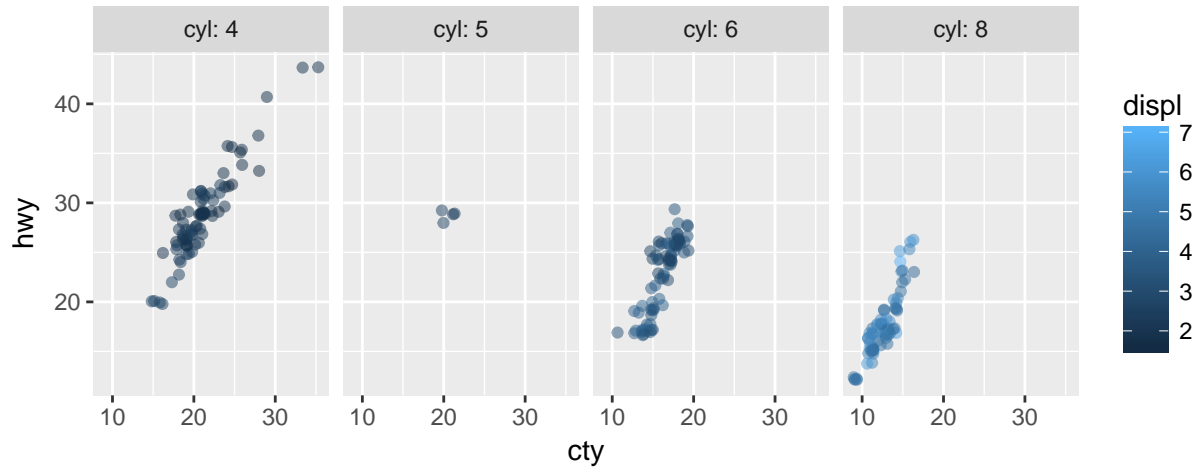
1. What happens if you try to facet by a continuous variable like `hwy` ? What about `cyl` ? What's the key difference?

When a continuous variables, like `hwy`, is used for facet, ggplot converts it into factor and creates facet from all unique value of that continuous variable. Here `hwy` has many unique values so we will get many facets for each of them while `cyl` has few discrete values and is useful to use for faceting.

2. Use facetting to explore the 3-way relationship between fuel economy, engine size, and number of cylinders. How does facetting by number of cylinders change your assesement of the relationship

between engine size and fuel economy?

```
ggplot(mpg, aes(cty, hwy, color = displ)) +  
  geom_point(alpha = 0.5, position = "jitter") +  
  facet_grid(.~cyl, labeller = label_both)
```



Here we can see that larger engine size has lower mileage in both city and highway. In addition, vehicle with large number of cylinder has larger engine size. Further there are very few vehicle having 5 cylinder.

3. Read the documentation for `facet_wrap()`. What arguments can you use to control how many rows and columns appear in the output?

The `nrow` and `ncol` arguments in `facet_wrap()` controls the number of rows and columns.

4. What does the `scales` argument to `facet_wrap()` do? When might you use it?

Here `scales` can take three values – `free`, `free_x` and `free_y`. `free_x` gives separate x-axis for each facet, `free_y` gives separate y-axis for each facet and similarly, `free` gives separate x and y axis for each facet.

Bibliography