

Chapter 02: First steps

Stan Piotrowski

September 13 2021

Contents

Setup	1
Fuel economy data exercises	1
Aesthetic attributes exercises	3
Faceting exercises	5
Plot geoms exercises	5

Setup

This chapter focuses on getting familiar with the basic recipes to create graphics using ggplot2.

```
# Load libraries
library(tidyverse)
library(kableExtra)
library(patchwork)

# Set ggplot2 theme
my_theme <- theme(
  panel.grid = element_blank(),
  panel.background = element_rect(fill = "white", color = "black")
)
```

Fuel economy data exercises

- 1) *List five functions that you could use to get more information about the `mpg` dataset.*

If you wanted more general information about the `mpg` dataset (e.g., descriptions of the underlying data, or where to find more detailed information or the source), you could use `?`, `??`, or `help()`. If you wanted to get a quick summary of the data and see the distribution of each variable, you could use the `summary()` function.

There are a few additional functions that you could use, some of which are described at the following page: <https://www.r-project.org/help.html>. There is no such vignette for the `mpg` dataset, but `browseVignettes()` or `vignette()` can be used to find tutorials for selected packages. Additionally, if you really don't know the name of the package you are interested in and don't want to consult Google, you could use the `apropos()` function to identify the object or function in the R environment using regular expression pattern matching.

- 2) *How can you find out what other datasets are included with `ggplot2`?*

To find which datasets are included in `ggplot2` (or any other package you are interested in), use the command `data(package = "<package_name>")`. For example, using `data(package = "ggplot2")`, we can see that there are datasets not only on fuel economy (`mpg`), but others on diamond prices and characteristics (`diamonds`), seal movements (`seals`), and housing sales in Texas (`txhousing`).

Table 1: Fuel economy (distance traveled with one US gallon) and fuel consumption (liters consumed per 100 kilometers).

manufacturer	model	year	cty	hwy	cty_lpk	hwy_lpk
audi	a4	1999	18	29	13.07	8.11
audi	a4	1999	21	29	11.20	8.11
audi	a4	2008	20	31	11.76	7.59
audi	a4	2008	21	30	11.20	7.84
audi	a4	1999	16	26	14.70	9.05
audi	a4	1999	18	26	13.07	9.05

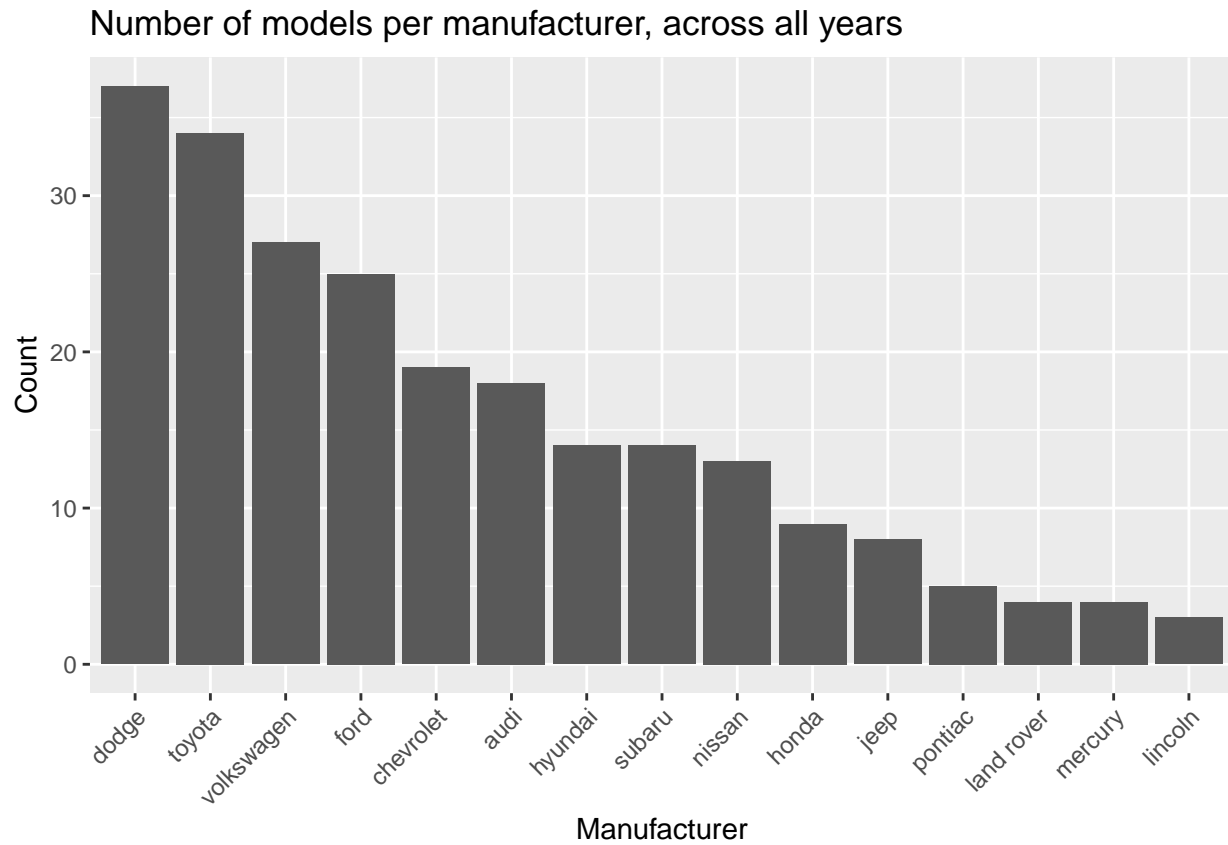
- 3) *Apart from the US, most countries use fuel consumption (fuel consumed over fixed distance) rather than fuel economy (distance traveled with a fixed amount of fuel). How could you convert the `cty` and `hwy` into the European standard of l/100km?*

To convert the fuel economy variables `cty` and `mpg` into fuel consumption by the European standard of l/100km, we can simply create a new variable `lpkm` by dividing the mpg estimate by the conversion factor 235.21.

```
# Convert miles per gallon to liters per 100 kilometers
mpg %>%
  mutate(cty_lpk = round(235.21 / cty, 2),
         hwy_lpk = round(235.21 / hwy, 2)) %>%
  dplyr::select(manufacturer, model, year, cty, hwy, cty_lpk, hwy_lpk) %>%
  head() %>%
  kbl(caption = "Fuel economy (distance traveled with one US gallon) and fuel consumption (liters consumed per 100 kilometers)",
      kable_classic_2(full_width = FALSE))
```

- 4) *Which manufacturer has the most 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?*

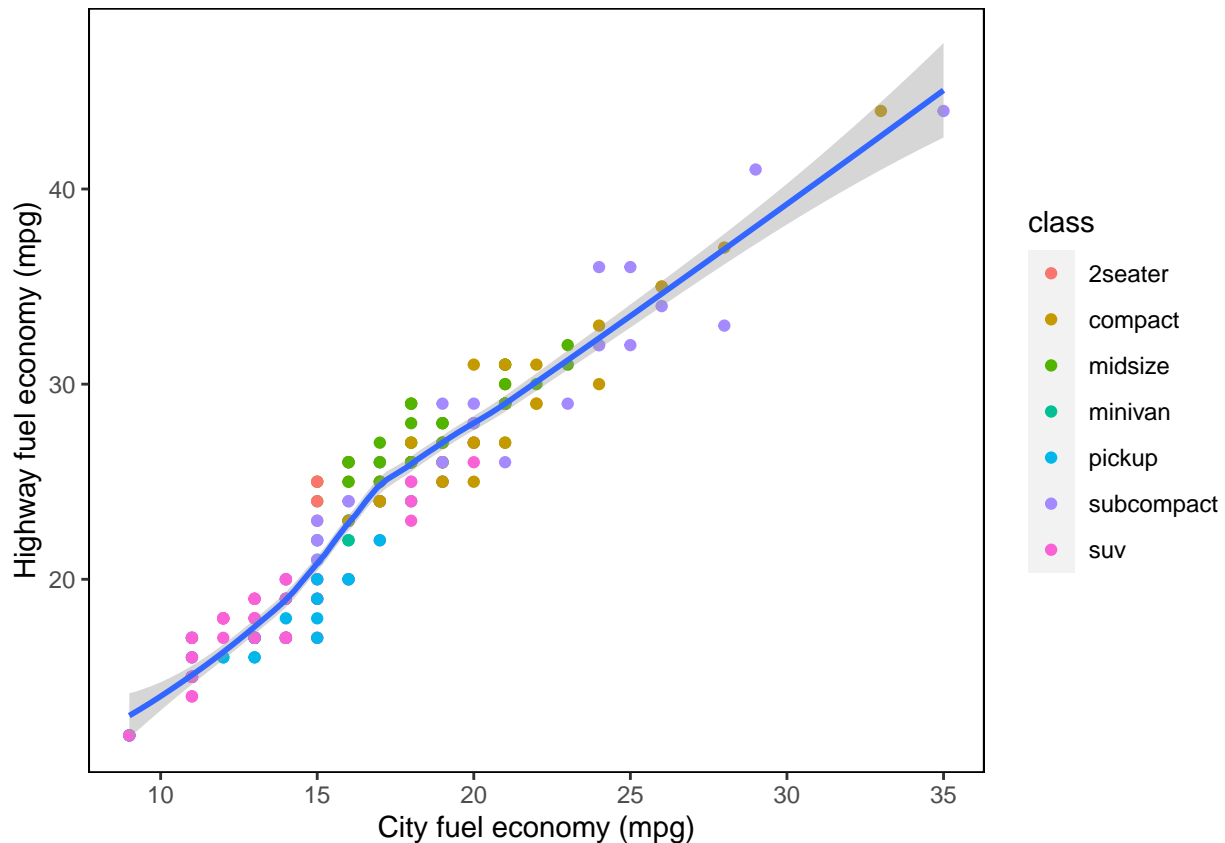
Dodge has the most models in the dataset (37; see visualization below), and the Dodge caravan 2wd has the most variations (11). Interestingly, the number doesn’t change if we remove the redundant information (e.g., “quattro” or “4wd”), likely because manufacturers don’t want to use the same model name as competitors.



Aesthetic attributes exercises

- 1) *How would you describe the relationship between `cty` and `hwy`? Do you have any concerns about drawing conclusions for that plot?*

First, let's generate the plot of the two variables.



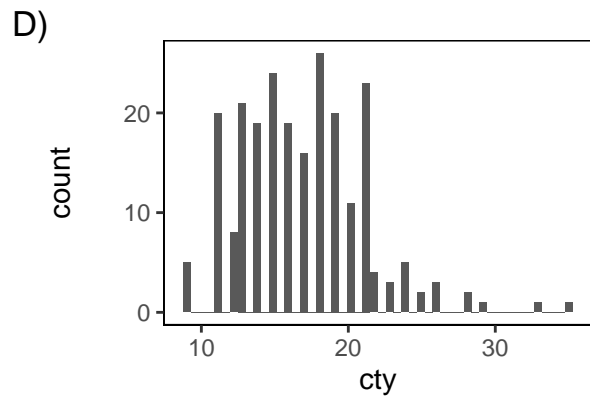
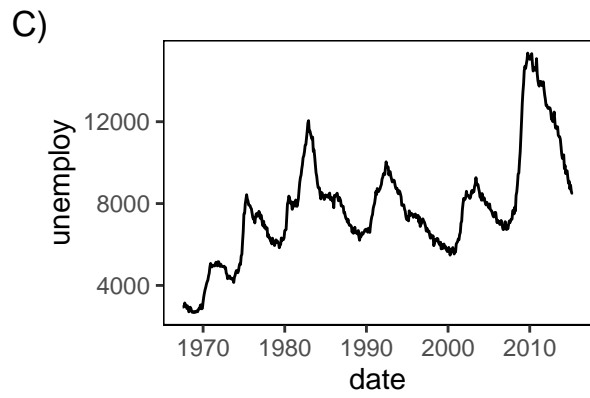
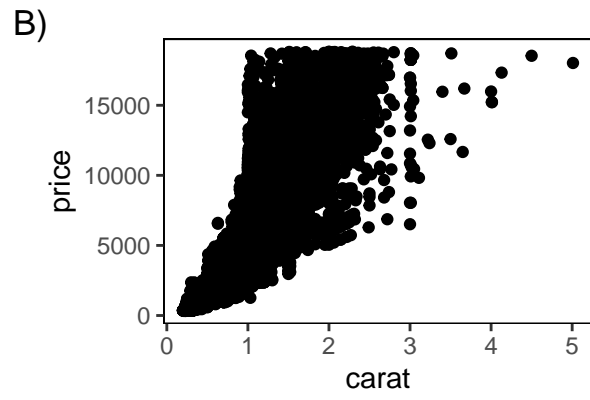
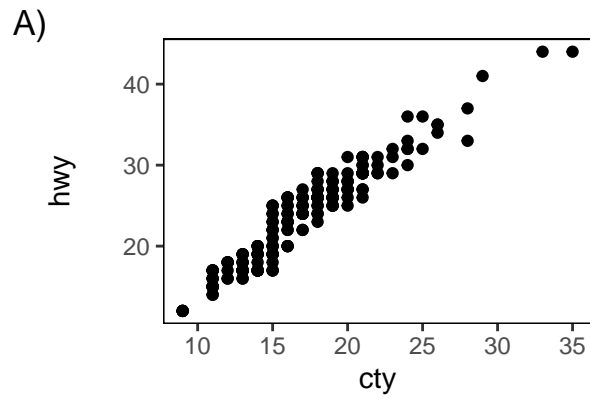
From the plot above, we can see that in general, the highway fuel economy increases as the city fuel economy increases. However, by mapping `class` to the color aesthetic, we can see that there are differences within each vehicle class that likely have to do with other features of the vehicles, like the year, perhaps.

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?

In general, the plot produced from the code `ggplot(mpg, aes(model, manufacturer)) + geom_point()` is not very useful because it only shows us the manufacturer for each car model in the dataset in the same way we would view the data in a table. To modify the data and make it more informative for visualization, we could create a plot to show the total number of vehicles for each manufacturer using a bar chart, for example.

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 see 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.

The first plot is identical to one of the plots we generated previously, except there is no variable mapped to the color aesthetic. The second plot shows the distribution diamond prices as a function of the number of carats. The third plot is a time series line plot of the number of unemployed Americans over time, starting in 1970 and extending until April 2015. Finally, the fourth plot shows the distribution of city fuel economy (mpg) for all vehicle models and manufacturers.



Faceting exercises

Plot geoms exercises