Telling stories with charts – Mastering ggplot2

Sarid Research Services

Materials by Adi Sarid https://adisarid.github.io and http://www.sarid-ins.com

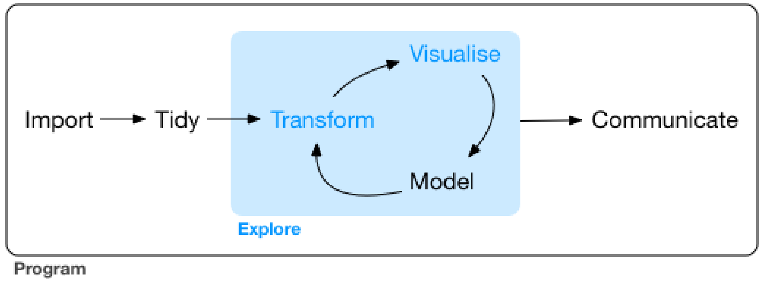
Reminder from the previous unit

- We talked about RStudio and got to know the environment
- We cloned the repository, opened a new project, opened a new markdown
- Understood the difference between scripts and RMarkdown
- We learned the base-r syntax, including loops, and functions
- We created an example function which computes the Fibonnacci series, and we did that in two methods: recursion and a loop
- We talked about debugging

Today we will be discussing visualizations

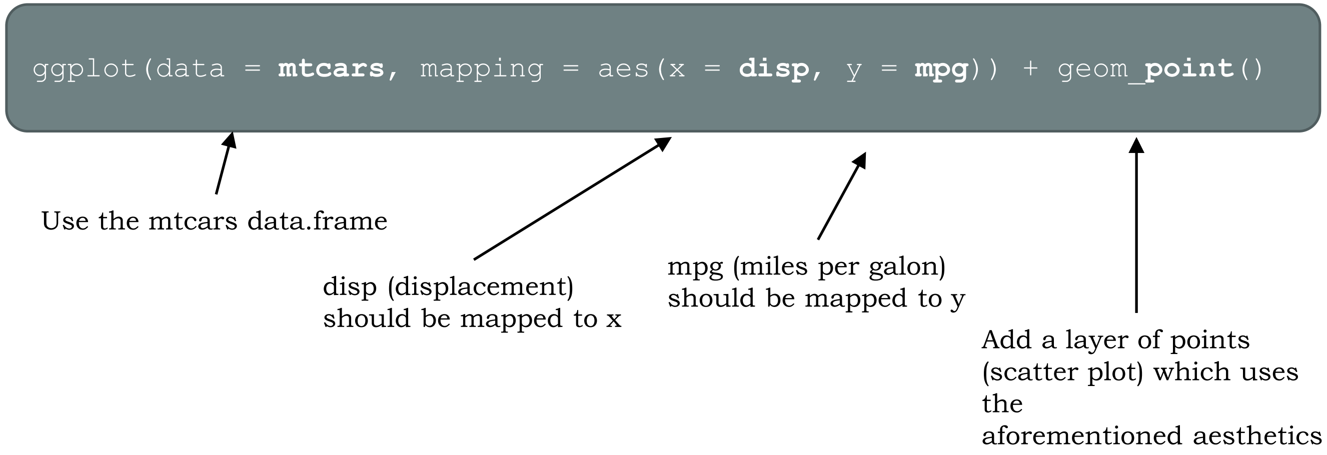
Why start with visualizations?

- Getting you "up to speed" with data exploration, the crucial triangle of the workflow, in which "Visualize" is a key part
- Visualizations help our understanding but are also a key part in communicating
- With charts you can generate leads for in-depth exploration
- Sometimes to generate a plot we have to use some transformations, so you will see some transformations as well
- Why ggplot2? a very advanced and flexible interface, which is also based on a sound theory "the grammar of graphics"



Spot the “aesthetics”

- Each ggplot is based on “aesthetics”, the different elements which are data-dependent and are “mapping” data elements into chart elements (like a function). I.e., how the data influences the chart (e.g., fill, color, axis, etc.)
- How many “aesthetics” can you spot in the following graph?
- What story does this chart tells you?



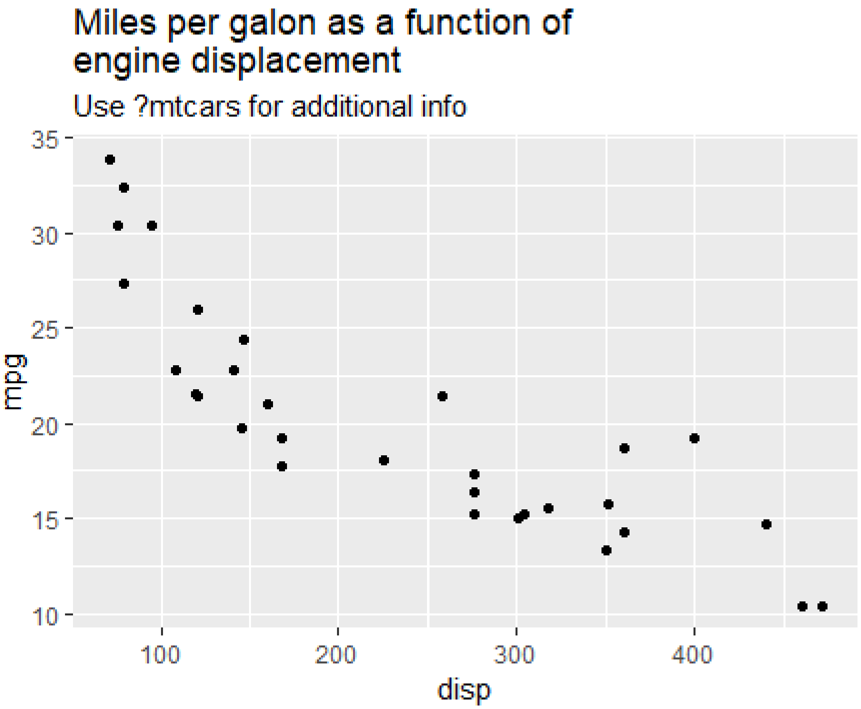
```
ggplot(data = mtcars, mapping = aes(x = disp, y = mpg)) + geom_point()
```

Use the mtcars data.frame

disp (displacement)
should be mapped to x

mpg (miles per galon)
should be mapped to y

Add a layer of points
(scatter plot) which uses
the
aforementioned aesthetics



If specified “in order” the “data =” and “mapping =” can be dropped:

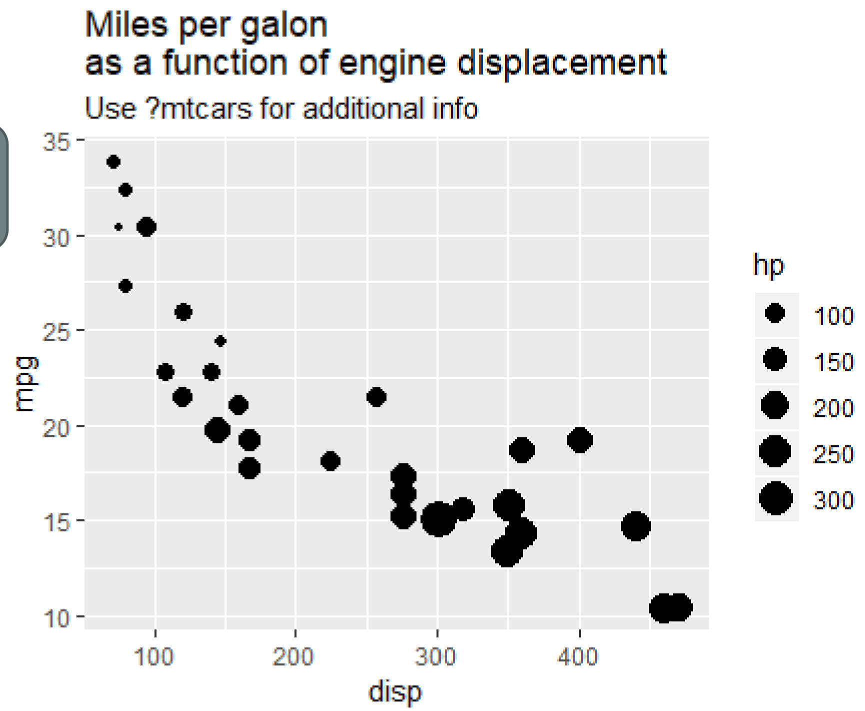
```
ggplot(mtcars, aes(x = disp, y = mpg)) + geom_point()
```

This is generally true for every function’s arguments

Let’s complicate things Spot the aesthetics (2)

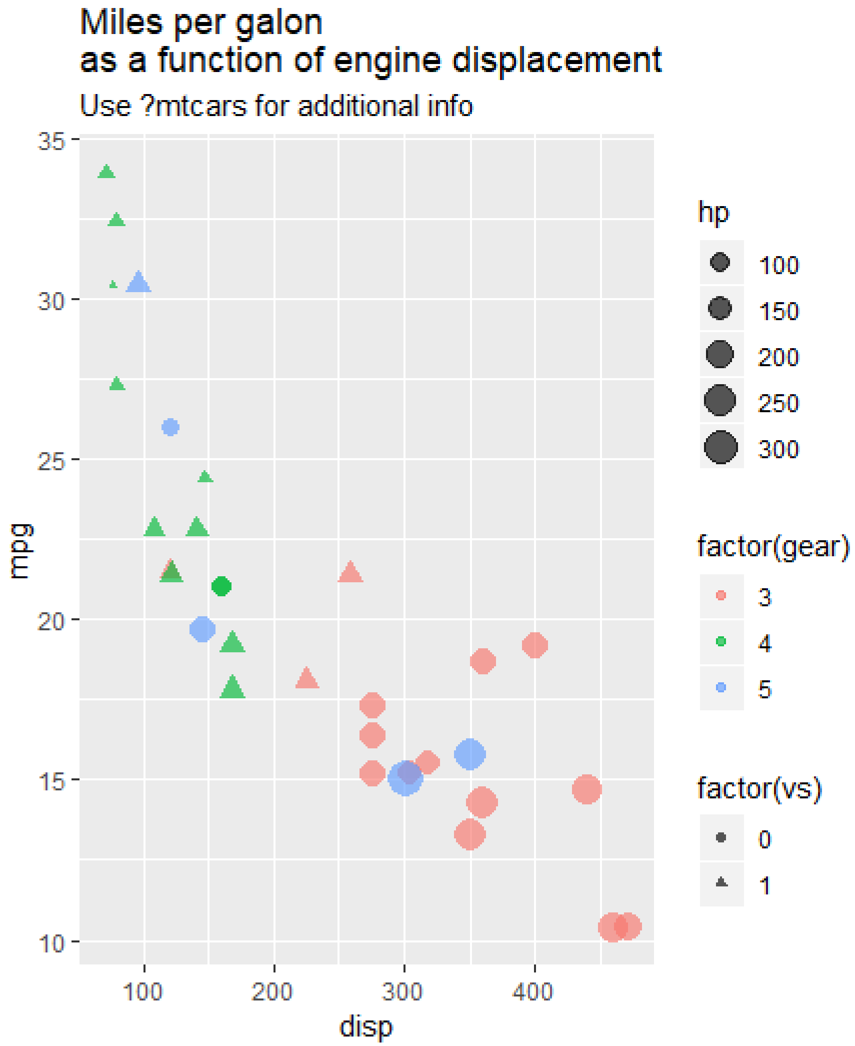
- What have I added?
- What can you deduce from this chart, that you couldn’t from the previous one?

```
ggplot(data = mtcars, mapping = aes(x = disp, y = mpg)) +
  geom_point(aes(size = hp))
```



Even further

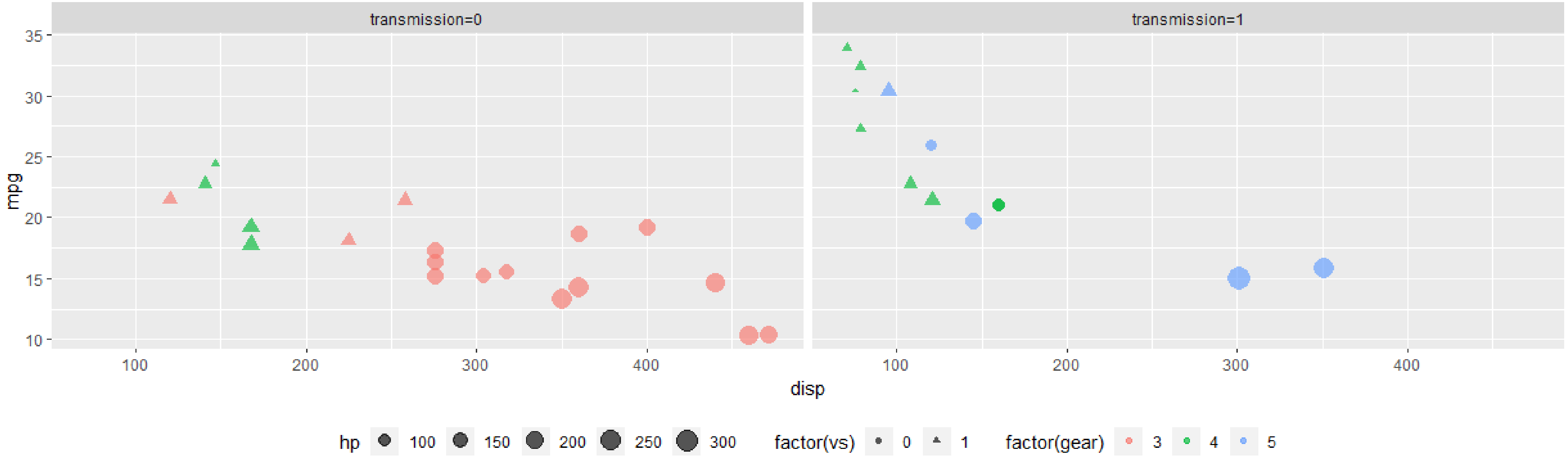
- I added *vs*: 0=V-shaped engine, 1=strait
- What can you deduce from this chart, that you couldn't from the previous one?



Materials by Adi Sarid https://adisarid.github.io and http://www.sarid-ins.com

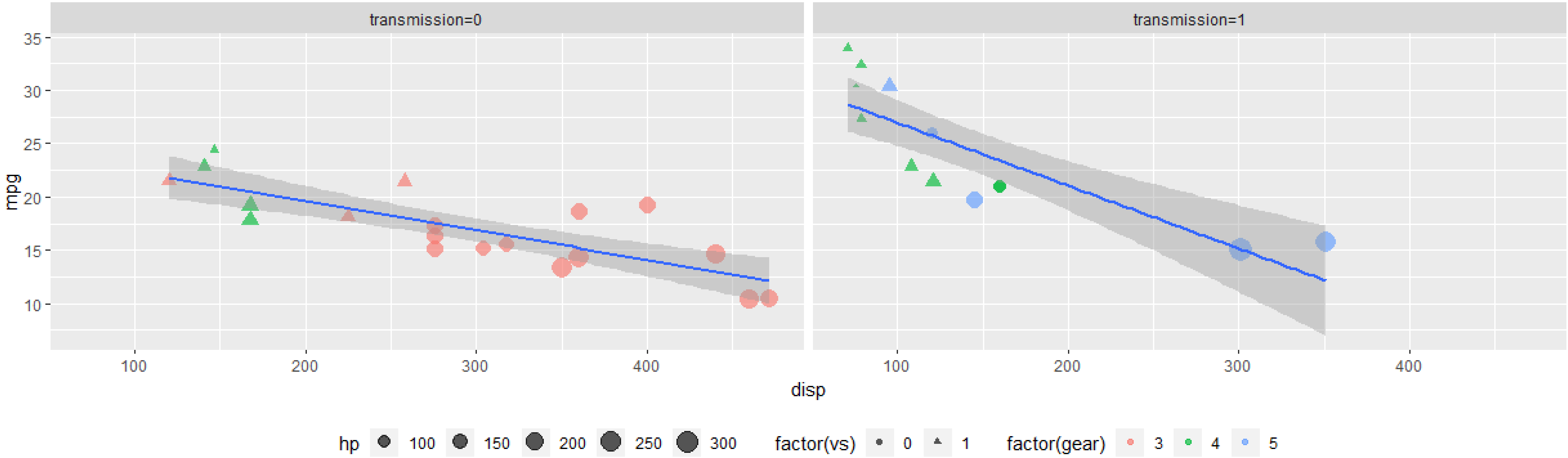
Facets

- We can “split” the chart (or look at different levels), using “facets”
  - For example, split by the transmission type (1 = manual, 0 = automatic)
  - Try to analyze the graph, what makes cars more efficient (=higher mpg)



Stats

- We can add various “statistics helpers”, such as smoothing (linear regression, lowess, polynoms, etc.)

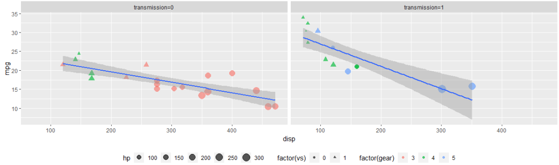


Materials by Adi Sarid https://adisarid.github.io and http://www.sarid-ins.com

Warning!

- **Let’s stop here.**
- ggplot2 has a lot of flexibility,
  but does that mean we should “push it”?
- 6 dimensions (mpg, disp, hp, vs, gear, am) means
- 15 (6 choose 2) 2-vars relationships
- 20 (6 choose 3) 3-vars relationships….
- Not really helpful: our short-term memory can process up to 7±2 “items” (some say even less)
- Too complex chart simply get lost in translation, and here comes our true challenge

https://dribbble.com/shots/1231599-Data-Monster-2x



1. Match the chart’s complexity to your audience
2. Generate charts that drive understanding and insights

Back to the “drawing board”

Open up the ggplot2 cheat sheet and look at the left side of the first page

- A ggplot is comprised of:
  - Mappings (“aes()”) which control how variables are mapped to properties
    - Can be global or local
  - Geoms (geom_*) which control the graphic expression of the mappings
    - Such as geom_point for scatterplots, geom_line, geom_histogram, geom_boxplot,… there are ~50 different geom_* functions
- Additional features include
  - Stats which add “statistical dressing” to the chart
    - Such as smoothing, density, ecdf,… there are ~30 different stat_* functions
  - Coordinates for controlling axis (~10 different coord_* functions)
  - Facets (splits the graph)
  - Scale (scale_*_* controls properties of the aesthetics, such as colors, axis labels, etc.)
  - Theme (theme()) which gives us control on all other elements of the graph

Telling stories with charts

- We will have the chance to practice all these elements and technical aspects in exercises, but first, how do you build the "right" chart?
- These set of questions will help guide you:
  - How many variables are involved?
  - What are the properties of each variable?
    - Continuous (numeric) / Discrete (factor) / Ordinal (ordered factor) / Date / Logical
  - Consider what are appropriate mappings
    - By axis / color / shape / fill / size / other
- Fine tuning: titles, axis titles, size, legend
  - Are you missing important data?
  - Are you creating any distortions due to axis or scales?

tidyverse prerequisites

- Throughout the exercise you might also have the chance to use the following tidyverse (dplyr) functions:
  - mutate() – create a new variable
  - glimpse() – show the first few values of each vector
  - filter() – filter the data according to a specific condition
  - count() – count the number of observations per each combination
  - group_by() – group the dataset by a specific set of variables
  - summarise() – conduct summarizing operations (like mean or sum) according to the dataset's grouping
  - %>% pipe operator

- Lets demonstrate these over a live R session, via the *mtcars* dataset.

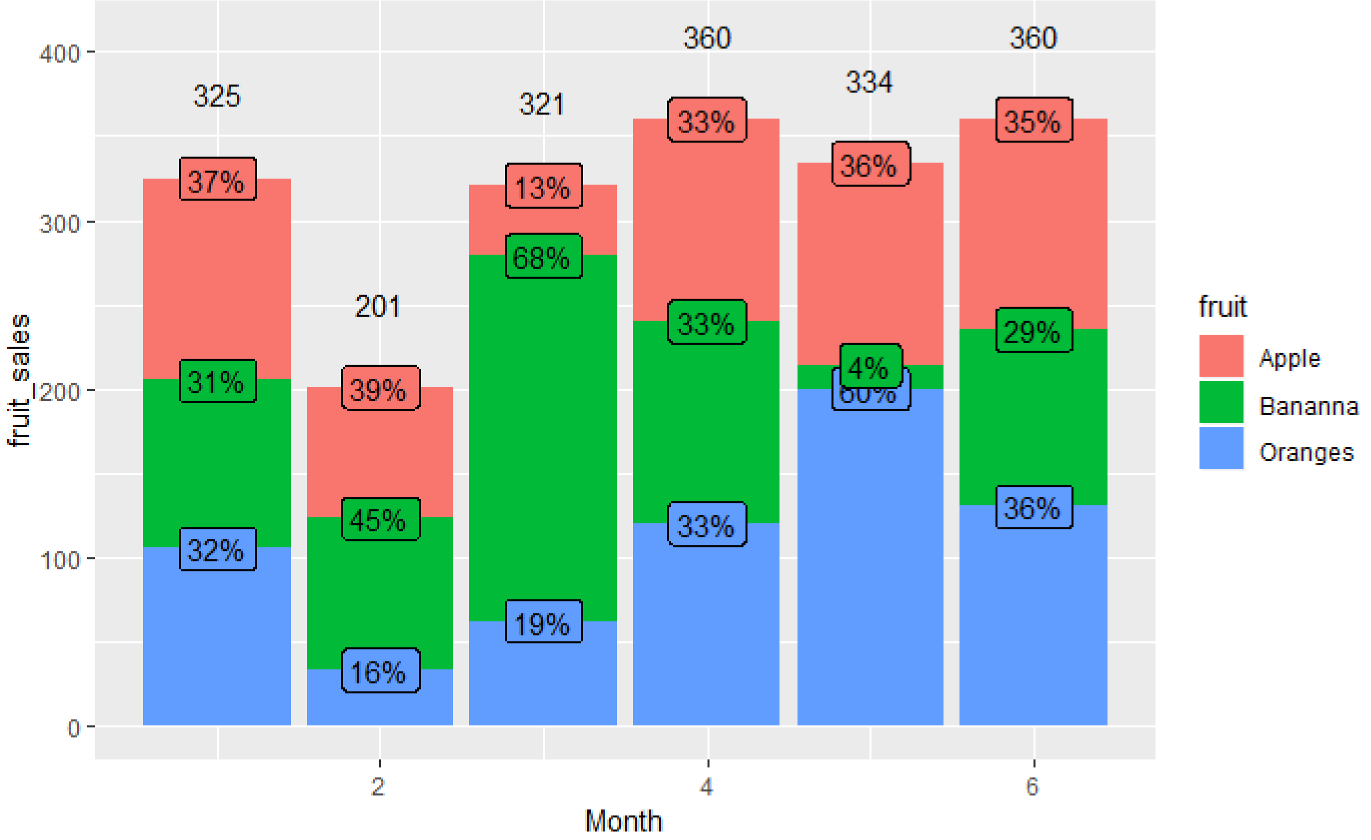
Exercise

- From the exercise folder open 02-Plotting.Rmd and start
  - “Exercise 1: the *google play* dataset”
- Before starting, if you want a stable copy of your work, it is recommended you save it in a separate location (to not run it over when you git pull in the future)
- After we solve the exercise together (or if you finished early on), continue to exercise 1.5 (related)

Mini exercise – how would you…(1)?

- **Use the ggplot2 cheat sheet**
- Answer in groups of 2-3
- What are the **three** geoms required to produce this chart?
- What are the aesthetic mappings?

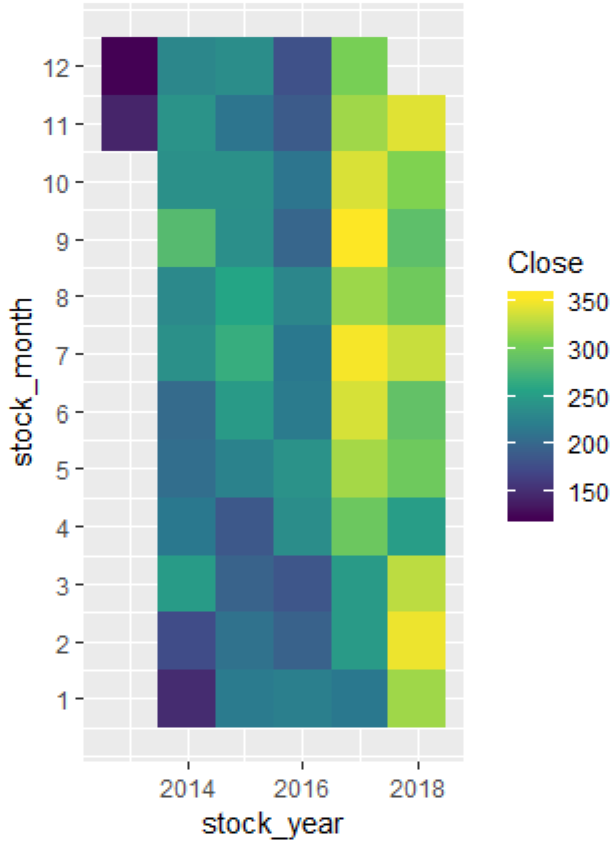
4 minutes



Mini exercise – how would you…(2)?

- TSLA (Tesla) stock closing price
- **Use the ggplot2 cheat sheet**
- Answer in groups of 2-3
- What is the **one** geom required to produce this chart?
- What are the aesthetic mappings?

4 minutes



Exercise

- In 02-Plotting.Rmd, continue to exercise 2.
- In 02-Plotting.Rmd, continue to exercise 3.
- In 02-Plotting.Rmd, continue to exercise 4.