Amounts and Proportions

Session 4

PMAP 8921: Data Visualization with R
Andrew Young School of Policy Studies
May 2020

Plan for today

Reproducibility

Amounts

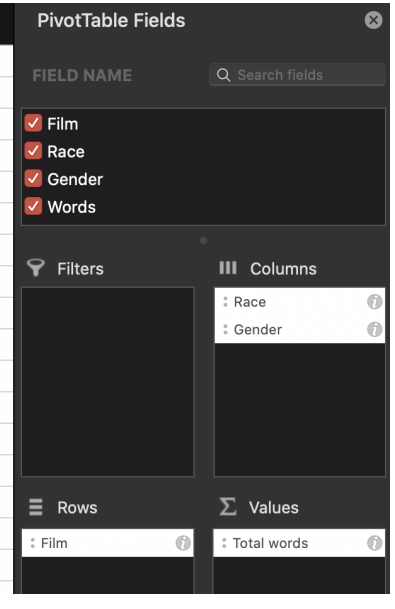
Proportions

Reproducibility

Why am I making you learn R?

Pivot Tables do the same thing!

Total words	Column Labels									
	Elf		Elf Total	Hobbit		Hobbit Total	Man		Man Total	Grand Total
Row Labels	Female	Male		Female	Male		Female	Male		
The Fellowship Of The Ring	1229	971	2200	14	3644	3658	0	1995	1995	7853
The Return Of The King	183	510	693	2	2673	2675	268	2459	2727	6095
The Two Towers	331	513	844	0	2463	2463	401	3589	3990	7297
Grand Total	**1743**	**1994**	**3737**	**16**	**8780**	**8796**	**669**	**8043**	**8712**	**21245**



Why am I making you learn R?

More powerful

Free and open source

Reproducibility

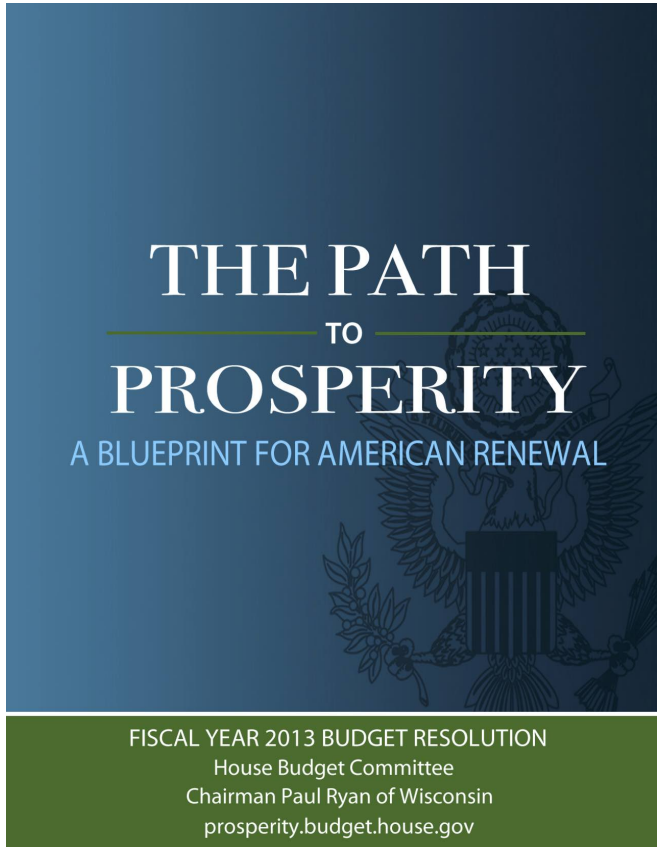
Austerity and Excel

Growth in a Time of Debt
Carmen M. Reinhart and Kenneth S. Rogoff
NBER Working Paper No. 15639
January 2010, Revised January 2010
JEL No. E2,E3,E6,F3,F4,N10

ABSTRACT

We study economic growth and inflation at different levels of government and external debt. Our analysis is based on new data on forty-four countries spanning about two hundred years. The dataset incorporates over 3,700 annual observations covering a wide range of political systems, institutions, exchange rate arrangements, and historic circumstances. Our main findings are: First, the relationship between government debt and real GDP growth is weak for debt/GDP ratios below a threshold of 90 percent of GDP. Above 90 percent, median growth rates fall by one percent, and average growth falls considerably more. We find that the threshold for public debt is similar in advanced and emerging economies. Second, emerging markets face lower thresholds for external debt (public and private)—which is usually denominated in a foreign currency. When external debt reaches 60 percent of GDP, annual growth declines by about two percent; for higher levels, growth rates are roughly cut in half. Third, there is no apparent contemporaneous link between inflation and public debt levels for the advanced countries as a group (some countries, such as the United States, have experienced higher inflation when debt/GDP is high). The story is entirely different for emerging markets, where inflation rises sharply as debt increases.

Debt:GDP ratio
90%+ → −0.1% growth



Paul Ryan's 2013 House budget resolution

Austerity and Excel

Thomas Herndon

Over time, another problem emerged: Other researchers, using seemingly comparable data on debt and growth, couldn't replicate the Reinhart-Rogoff results. They typically found some correlation between high debt and slow growth — but nothing that looked like a tipping point at 90 percent or, indeed, any particular level of debt.

Finally, Ms. Reinhart and Mr. Rogoff allowed researchers at the University of Massachusetts to look at their original spreadsheet — and the mystery of the irreproducible results was solved. First, they omitted some data; second, they used unusual and highly questionable statistical procedures; and finally, yes, they made an Excel coding error. Correct these oddities and errors, and you get what other researchers have found: some correlation between high debt and slow growth, with no indication of which is causing which, but no sign at all of that 90 percent "threshold."

From Paul Krugman, "The Excel Depression"

Austerity and Excel

Table 1. Real GDP Growth as the Level of Government Debt Varies:
Selected Advanced Economies, 1790-2009
(annual percent change)

		Central (Federal) government debt/ GDP			
Country	Period	Below 30 percent	30 to 60 percent	60 to 90 percent	90 percent and above
Australia	1902-2009	3.1	4.1	2.3	4.6
Austria	1880-2009	4.3	3.0	2.3	n.a.
Belgium	1835-2009	3.0	2.6	2.1	3.3
Canada	1925-2009	2.0	4.5	3.0	2.2
Denmark	1880-2009	3.1	1.7	2.4	n.a.
Finland	1913-2009	3.2	3.0	4.3	1.9
France	1880-2009	4.9	2.7	2.8	2.3
Germany	1880-2009	3.6	0.9	n.a.	n.a.
Greece	1884-2009	4.0	***0.3***	***4.8***	2.5
Ireland	1949-2009	4.4	4.5	4.0	2.4
Italy	1880-2009	***5.4***	***4.9***	1.9	0.7
Japan	1885-2009	4.9	3.7	3.9	0.7
Netherlands	1880-2009	4.0	2.8	2.4	2.0
New Zealand	1932-2009	2.5	2.9	3.9	**3.6**
Norway	1880-2009	2.9	4.4	n.a.	n.a.
Portugal	1851-2009	4.8	2.5	1.4	n.a.
Spain	1850-2009	***1.6***	3.3	***1.3***	2.2
Sweden	1880-2009	2.9	2.9	2.7	n.a.
United Kingdom	1830-2009	2.5	2.2	2.1	1.8
United States	1790-2009	4.0	3.4	3.3	***-1.8***
Average		**3.7**	**3.0**	**3.4**	**1.7**
Median		**3.9**	**3.1**	**2.8**	**1.9**
Number of observations = **2,317**		866	654	445	352

Debt:GDP ratio = 90%+ → 2.2% growth (!!)

Genes and Excel

Septin 2

Membrane-Associated Ring Finger (C3HC4) 1

2310009E13

	A	B
1	**Actual value**	**What Excel turns it into**
2	SEPT2	2-Sep
3	MARCH1	1-Mar
4	2310009E13	2.31E+19

20% of genetics papers between 2005–2015 (!!!)

General guidelines

Don't touch the raw data

If you do, explain what you did!

Use self-documenting, reproducible code

R Markdown!

Use open formats

Use .csv, not .xlsx

R Markdown in real life

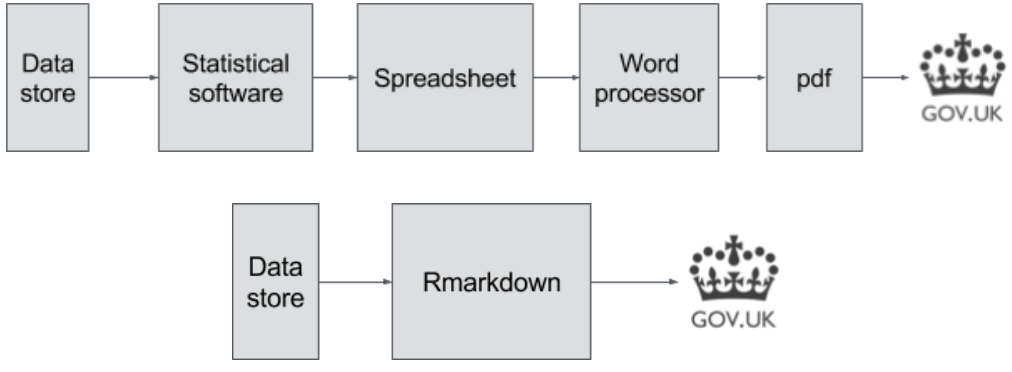
3.1.2 Data Visualization

We use ggplot2 as our main package to create ad-hoc exploratory graphics as well as polished-looking customized visualizations. When combined with tools to clean and transform data, ggplot2 allows analysts to quickly translate insights into high quality, compelling visualizations. In addition to the static graphics of `ggplot2`, we often make interactive visualizations or dashboards using R packages such as `plotly` (Sievert et al. 2017), `leaflet` (Cheng et al. 2017), `dygraphs` (Vanderkam et al. 2017), `DiagrammeR` (Sveidqvist et al. 2017), and `shiny` (Chang et al. 2017).

3.1.3 Reproducible Research

At Airbnb, all R analyses are documented in `rmarkdown`, where code and visualizations are combined within a single written report. Posts are carefully reviewed by experts in the content area and techniques used, both in terms of methodologies and code style, before publishing and sharing with the business partners. The peer review process is

Airbnb, ggplot, and rmarkdown

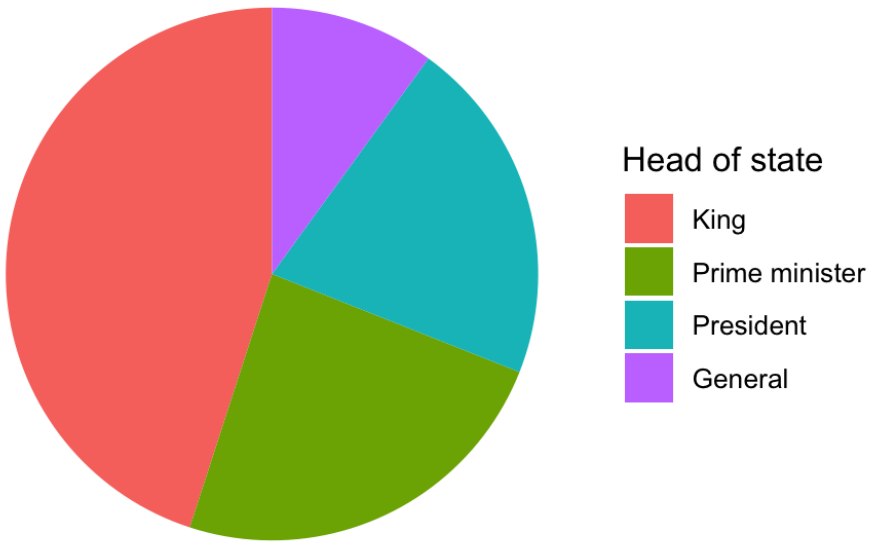


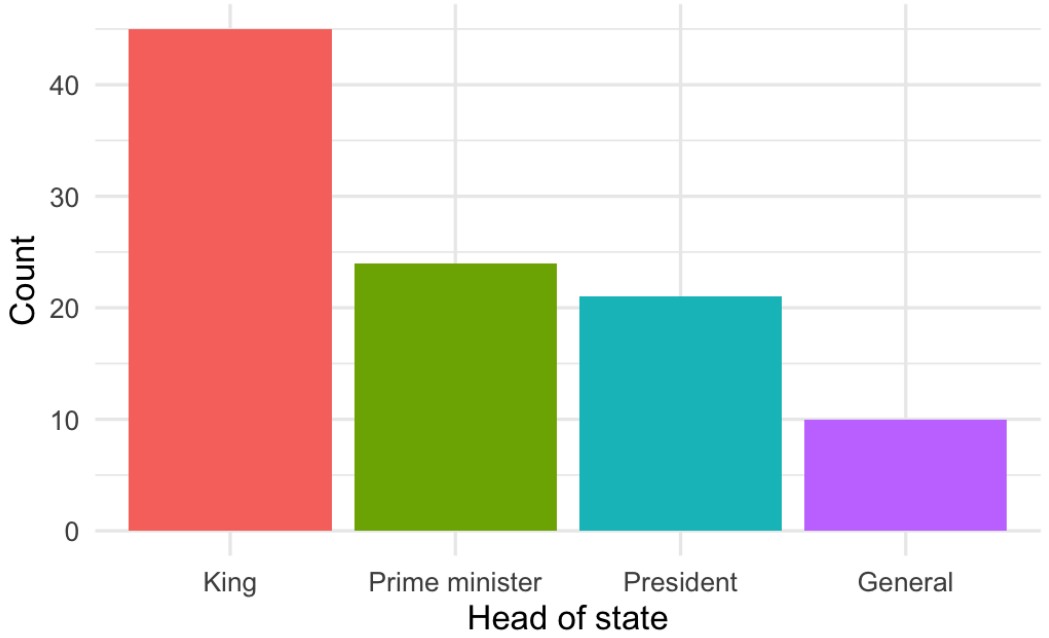
The UK's reproducible analysis pipeline

Amounts

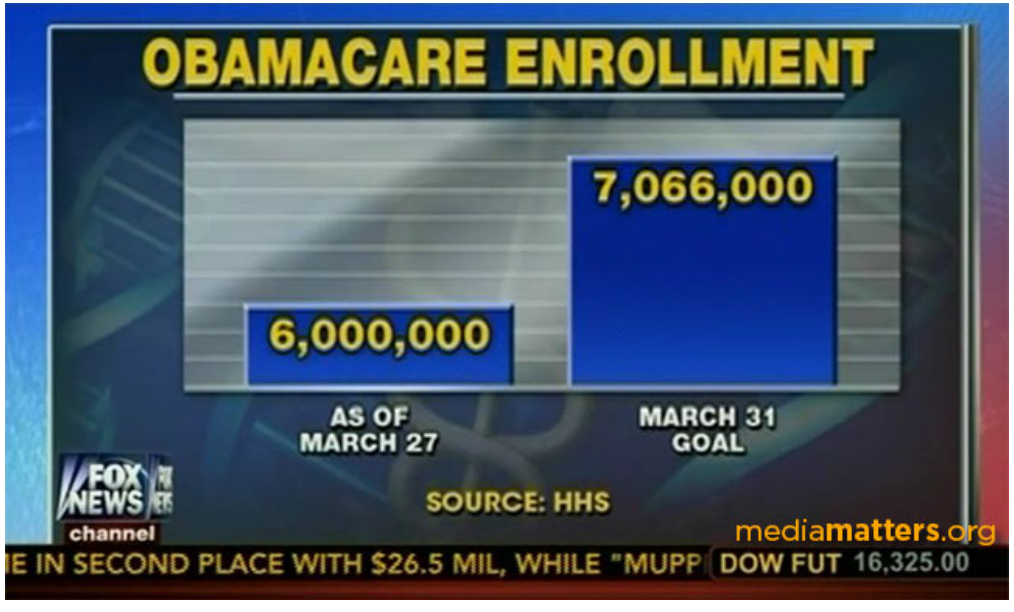
Yay bar plots!

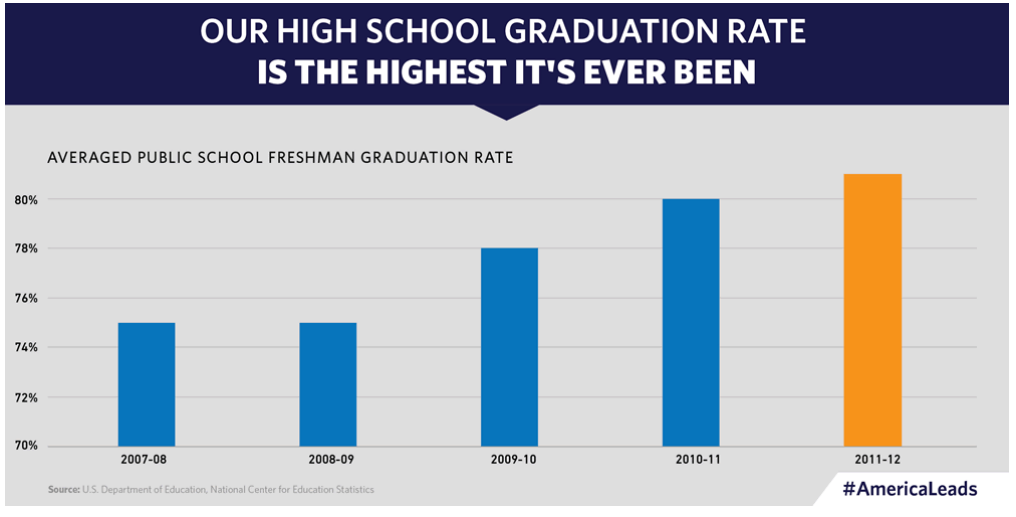
We are a lot better at visualizing line lengths than angles and areas





Oh no bar plots!





Start at zero

The entire line length matters, so don't truncate it!

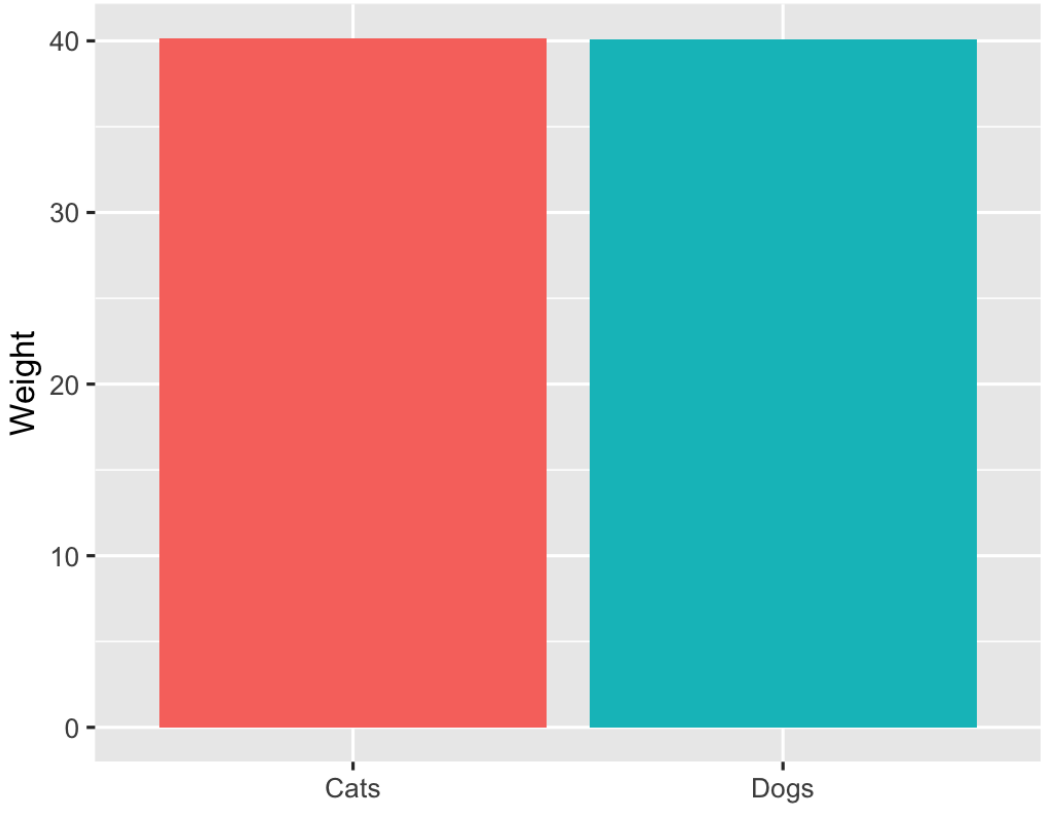
Always start at 0

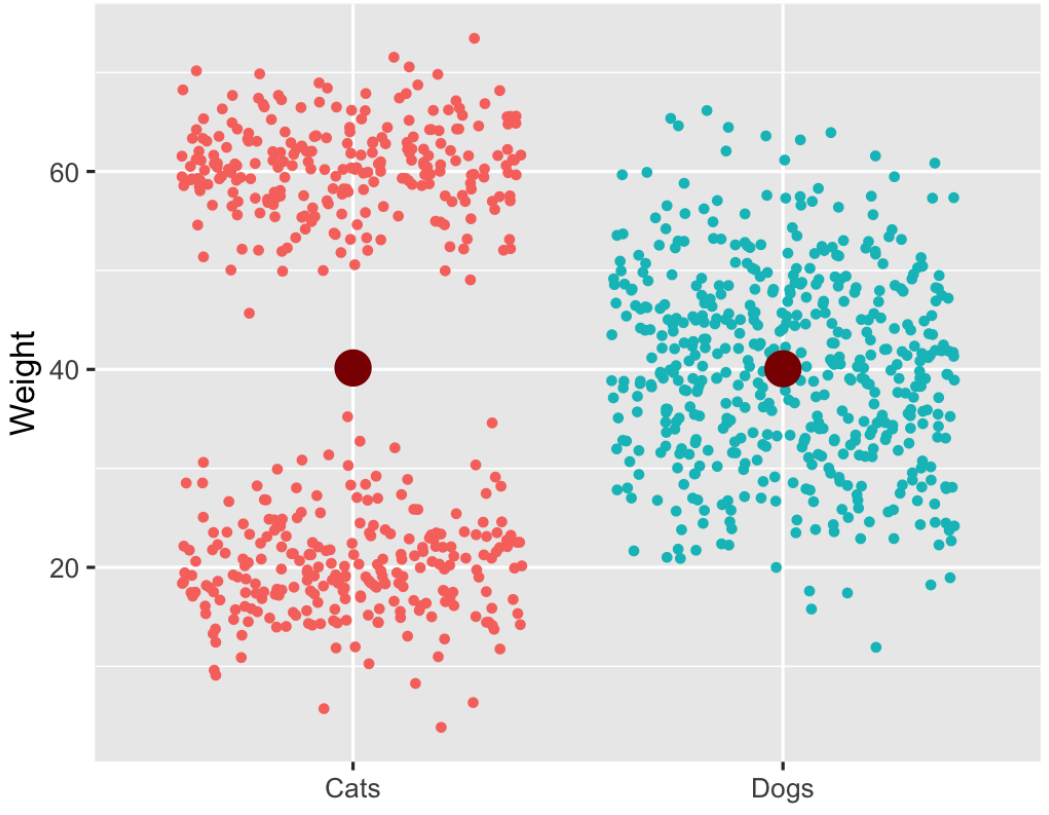
(Or don't use bars)

Bar plots and summary statistics

#barbarplots

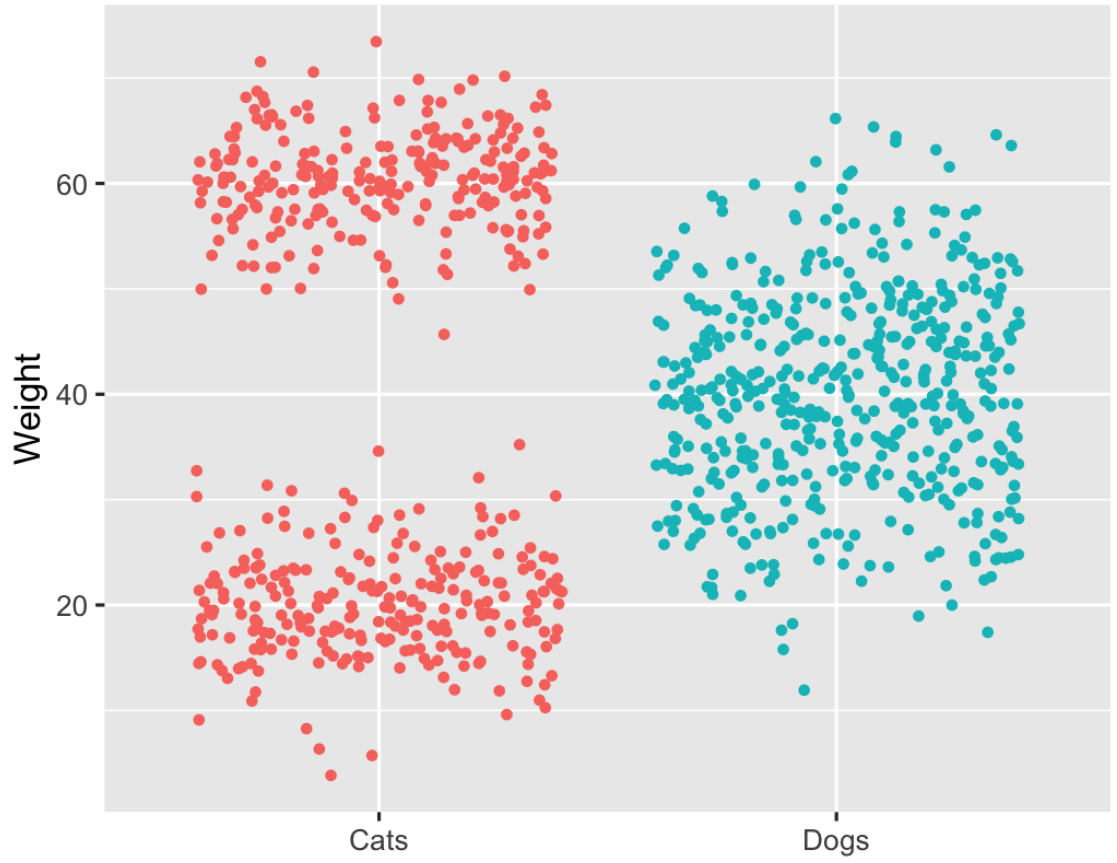
Bar plots and summary statistics





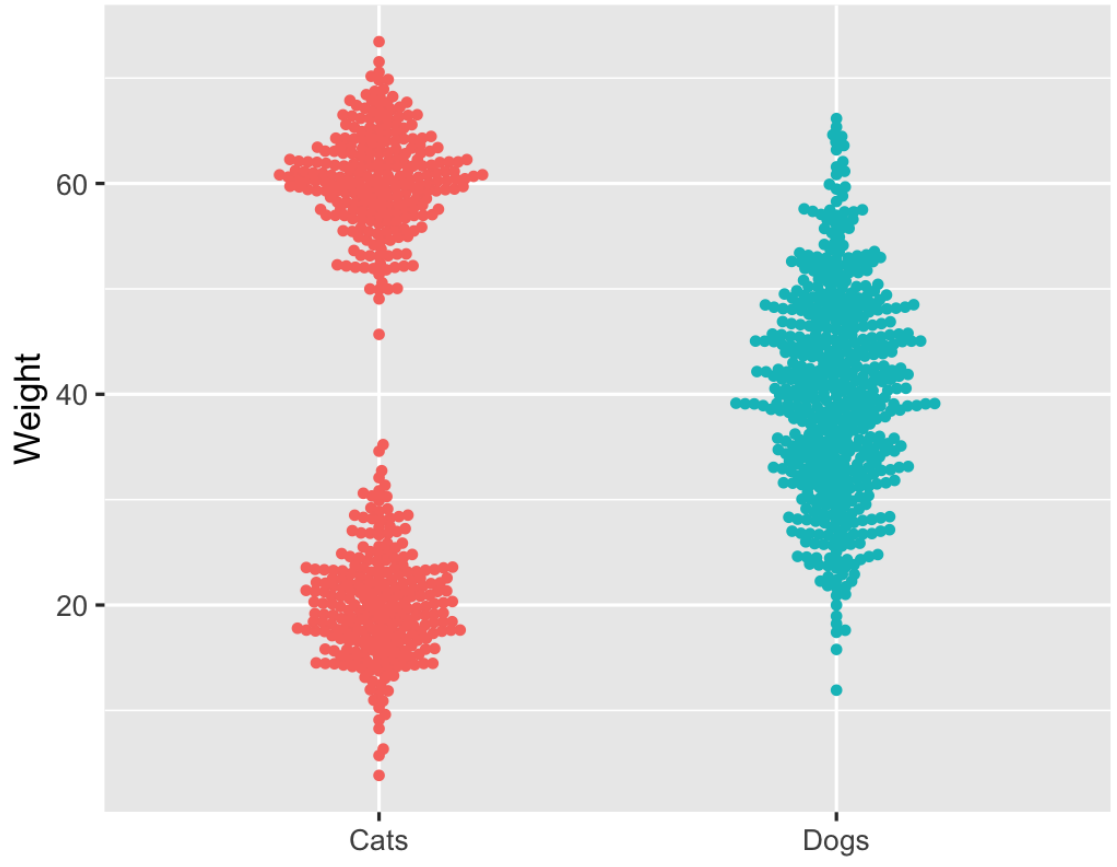
Show more data with strip plots

```
ggplot(animals,
       aes(x = animal_type,
           y = weight,
           color = animal_type)) +
  geom_point(position = position_jitter(heigh
             size = 1) +
  labs(x = NULL, y = "Weight") +
  guides(color = FALSE)
```



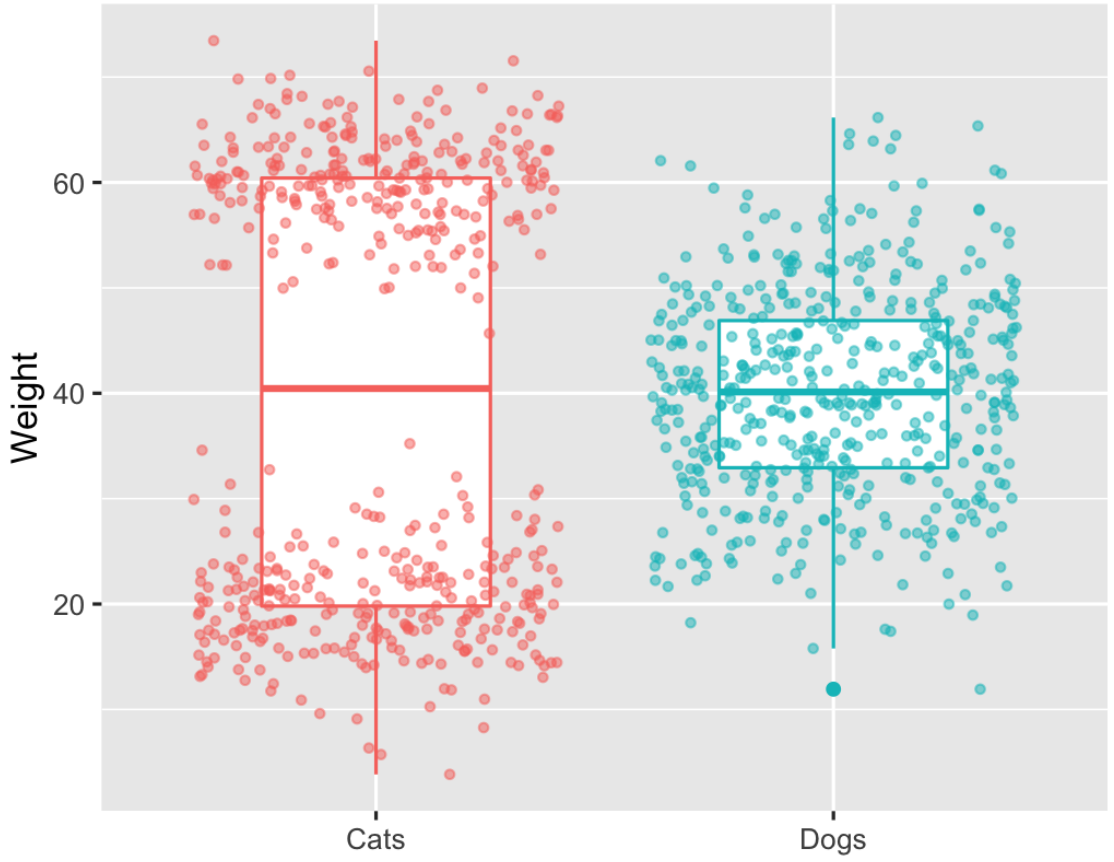
Show more data with beeswarm plots

```
library(ggbeeswarm)

ggplot(animals, aes(x = animal_type,
                    y = weight,
                    color = animal_type)) +
  geom_beeswarm(size = 1) +
  # Or try this too:
  # geom_quasirandom() +
  labs(x = NULL, y = "Weight") +
  guides(color = FALSE)
```



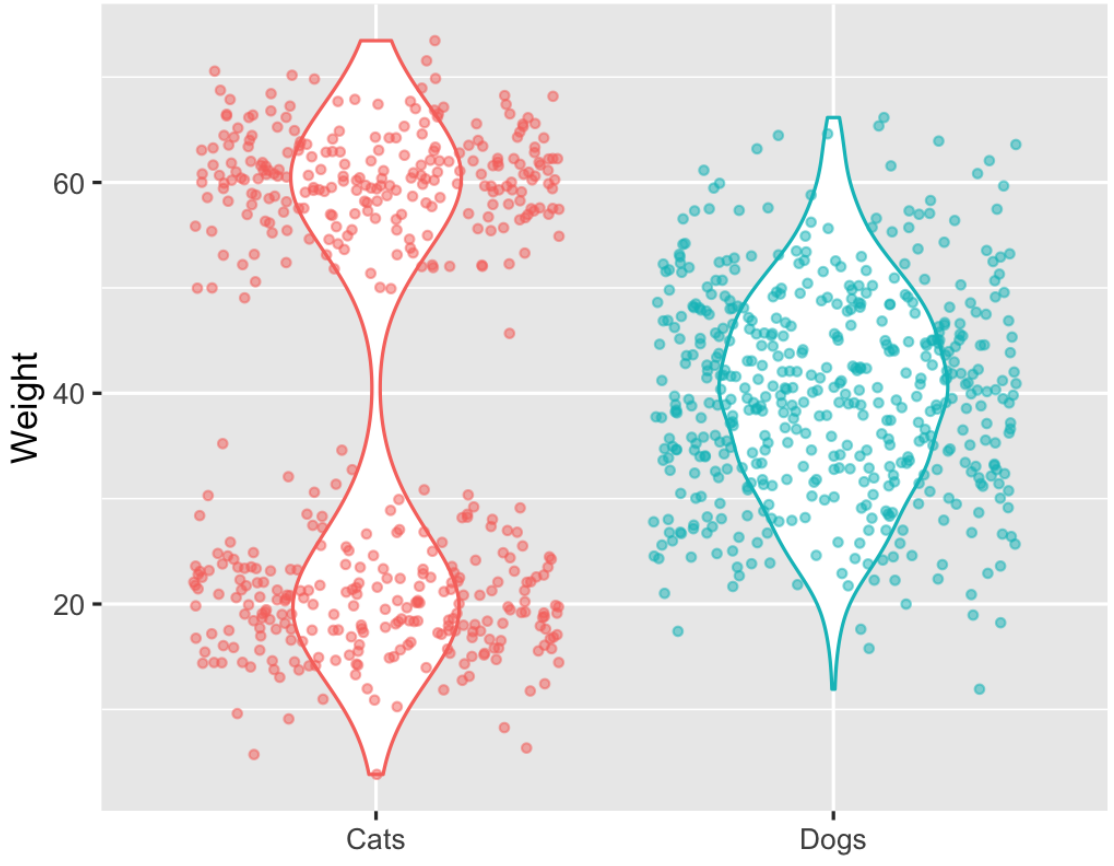
Combine boxplots with points

```
ggplot(animals, aes(x = animal_type,
                    y = weight,
                    color = animal_type)) +
  geom_boxplot(width = 0.5) +
  geom_point(position = position_jitter(heigh
             size = 1, alpha = 0.5) +
  labs(x = NULL, y = "Weight") +
  guides(color = FALSE)
```



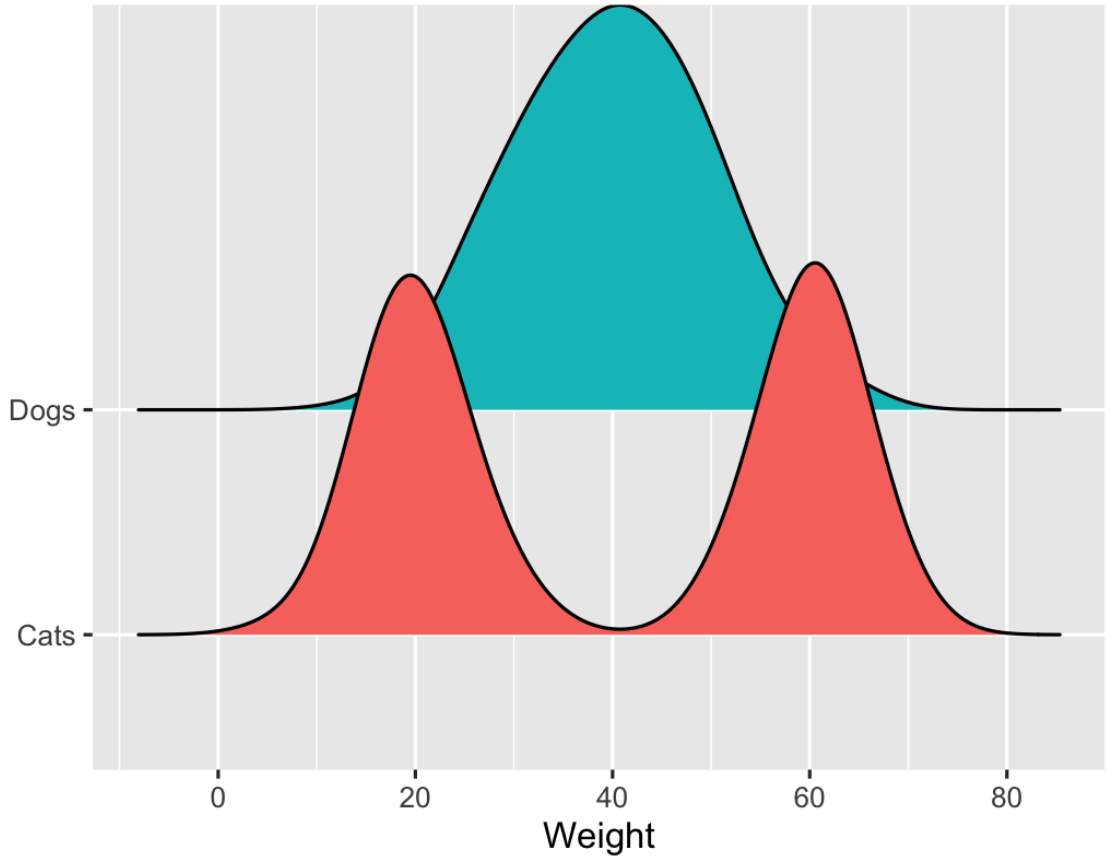
Combine violins with points

```
ggplot(animals, aes(x = animal_type,
                    y = weight,
                    color = animal_type)) +
  geom_violin(width = 0.5) +
  geom_point(position = position_jitter(heigh
             size = 1, alpha = 0.5) +
  labs(x = NULL, y = "Weight") +
  guides(color = FALSE)
```



Overlapping ridgeplots

```
library(ggridges)

ggplot(animals, aes(x = weight,
                    y = animal_type,
                    fill = animal_type)) +
  geom_density_ridges() +
  labs(x = "Weight", y = NULL) +
  guides(fill = FALSE)
```



General rules

Bar charts always start at zero

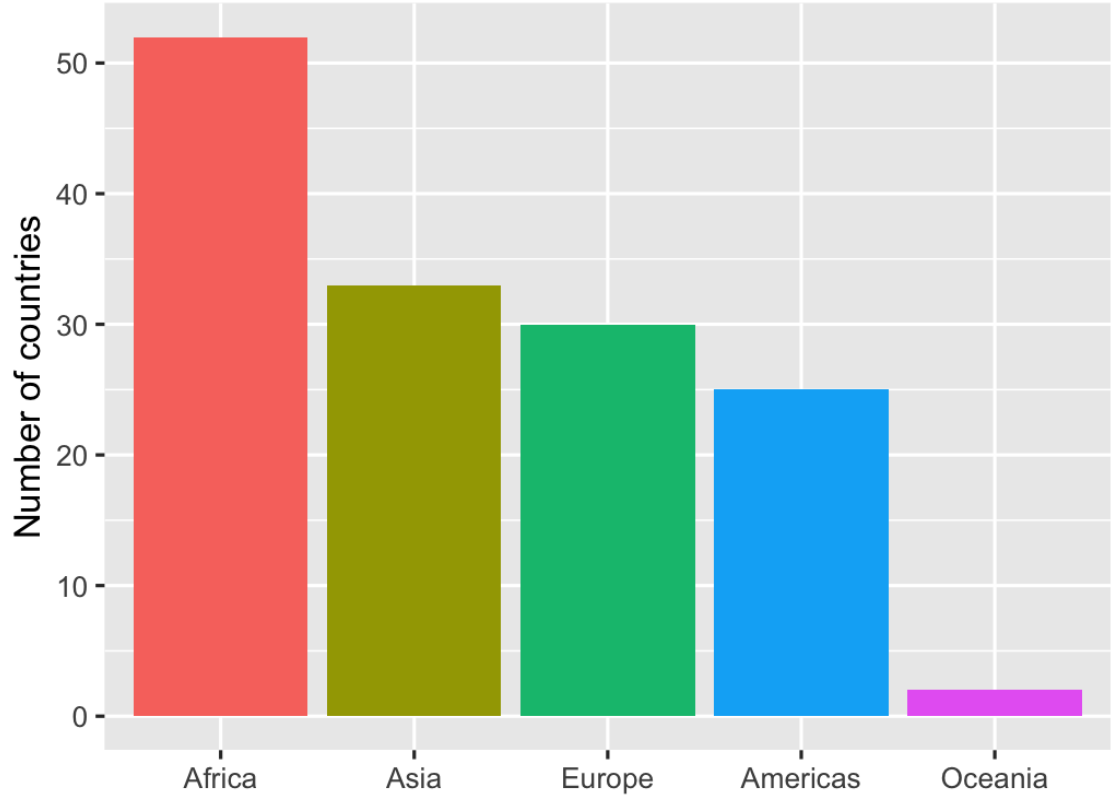
Don't use bars for summary statistics.
You throw away too much information.

The end of the bar is often all that matters

Lots of alternatives

We'll use a summarized version of the gapminder dataset as an example

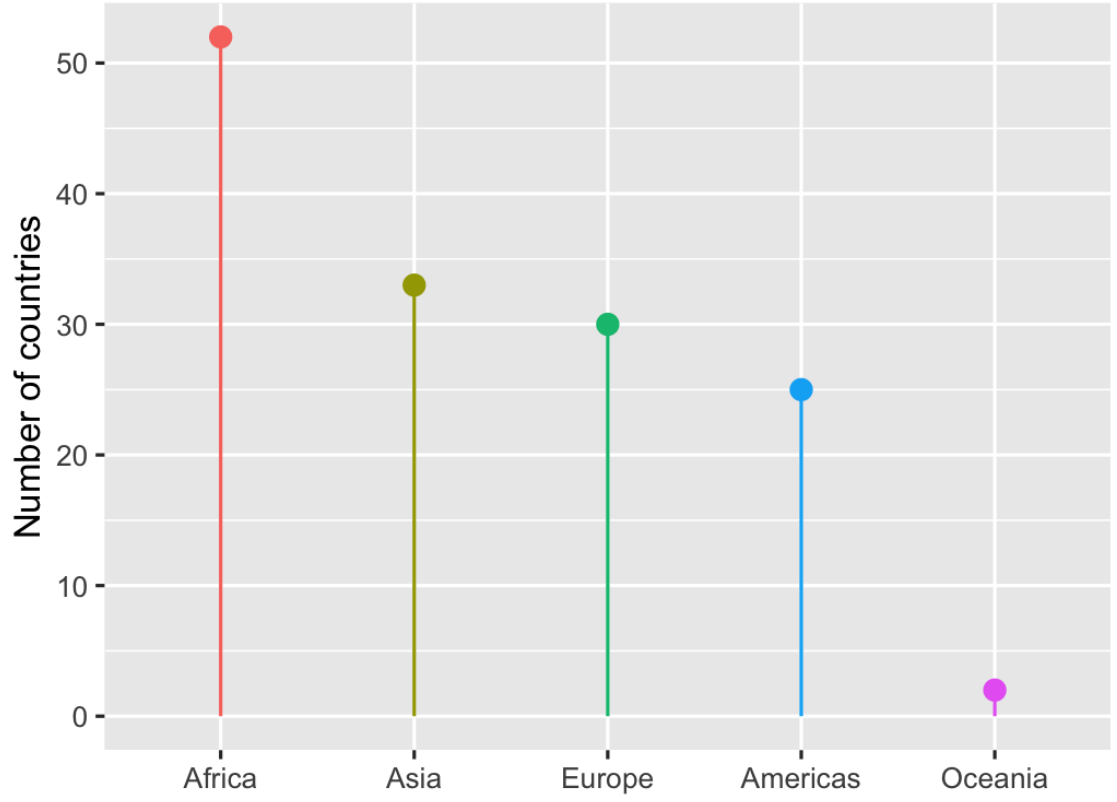
```
library(gapminder)
gapminder_continents <- gapminder %>%
  filter(year == 2007) %>%  # Only look at 2(
  count(continent) %>%  # Get a count of cont
  arrange(desc(n)) %>%   # Sort descendingly
  # Make continent into an ordered factor
  mutate(continent = fct_inorder(continent))

ggplot(gapminder_continents,
       aes(x = continent, y = n, fill = cont
  geom_col() +
  guides(fill = FALSE) +
  labs(x = NULL, y = "Number of countries")
```



Alternatives: Lollipop charts

Since the end of the bar is important, emphasize it the most

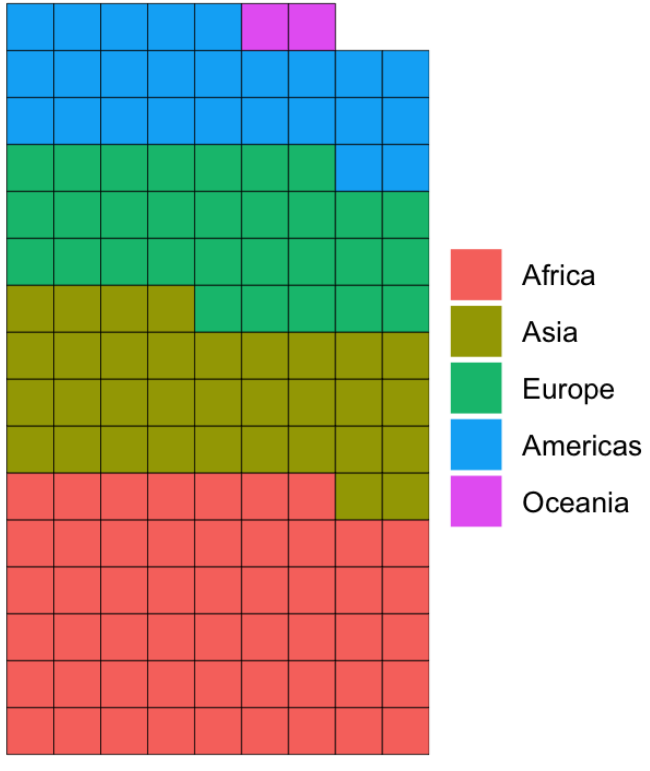
```
ggplot(gapminder_continents,
       aes(x = continent, y = n,
           color = continent)) +
  geom_pointrange(aes(ymin = 0, ymax = n)) +
  guides(color = FALSE) +
  labs(x = NULL, y = "Number of countries")
```



Alternatives: Waffle charts

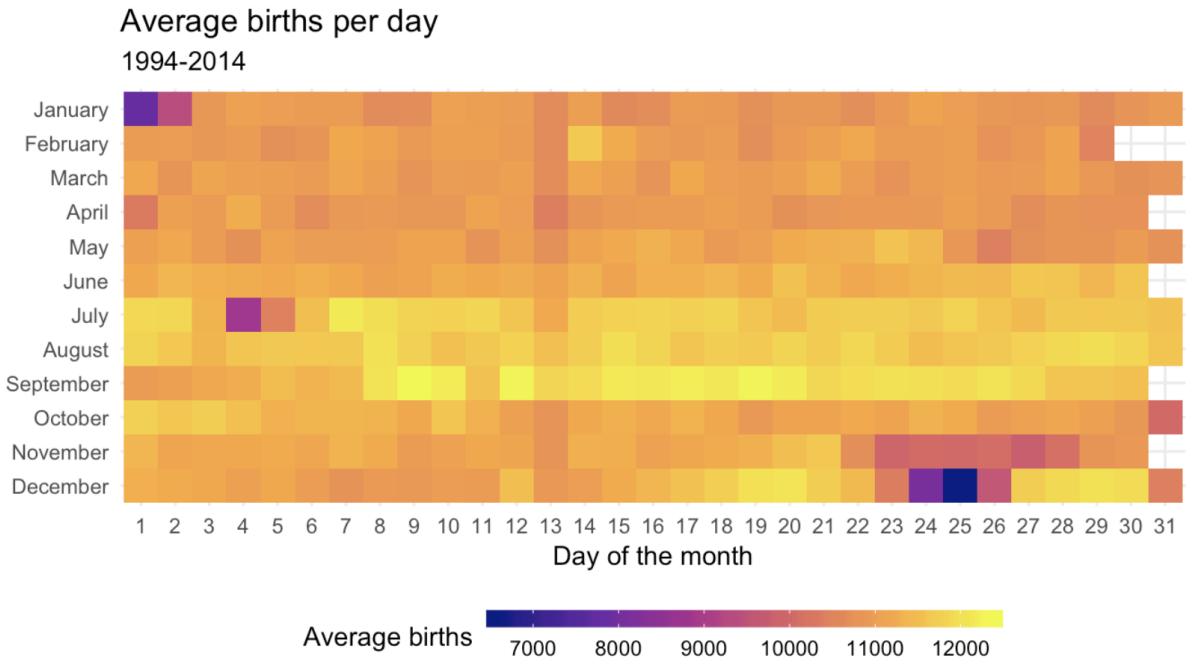
Show the individual observations as squares

```
# This has to be installed in a special way-
# Run this in your console:
# devtools::install_github("hrbrmstr/waffle")
library(waffle)

ggplot(gapminder_continents,
       aes(x = continent, y = n,
           fill = continent)) +
  geom_waffle(aes(values = n),  # geom_waffle
              n_rows = 9,  # It has lots of
              flip = TRUE) +
  labs(fill = NULL) +
  coord_equal() +  # Make all the squares sq
  theme_void()  # Use a completely empty the
```



Alternatives: Heatmaps

If exact counts are less important, try a heatmap with `geom_tile()`



Proportions

Why proportions?

Sometimes we want to compare values across a whole population instead of looking at raw counts

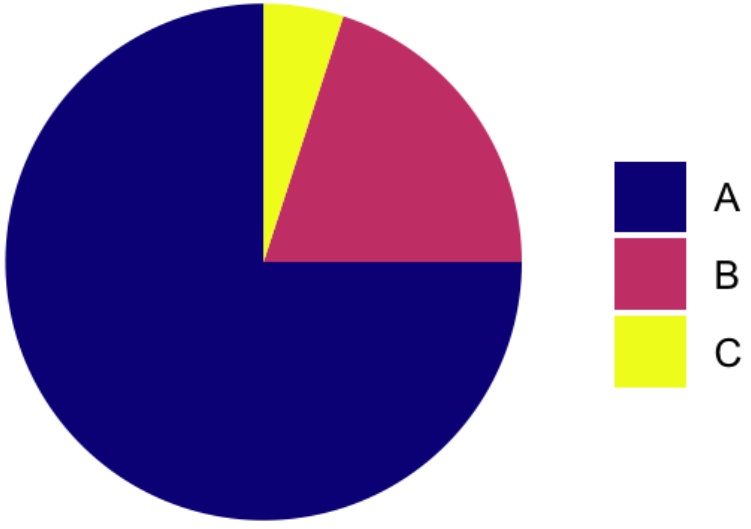
Only do this when it makes analytical sense!

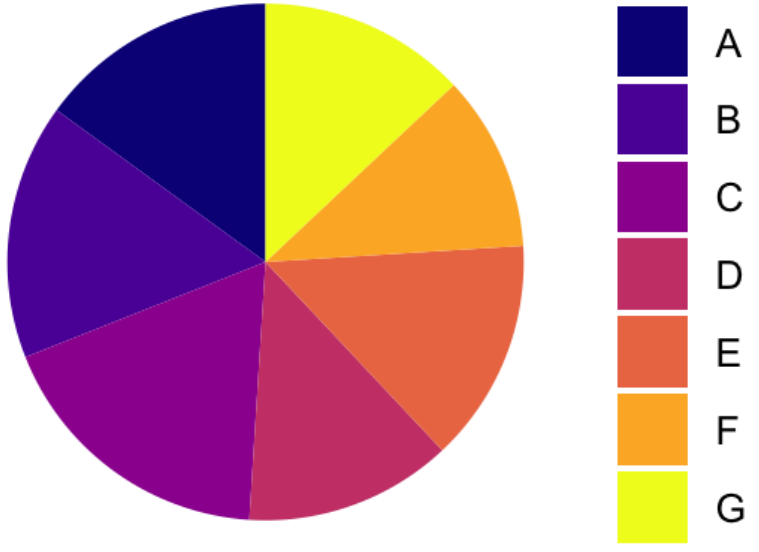
COVID-19 amounts vs. proportions

Pie charts

Perceptual issues with angle and fill space

Only okay(ish) if there are a few easily distinguishable categories





Alternatives

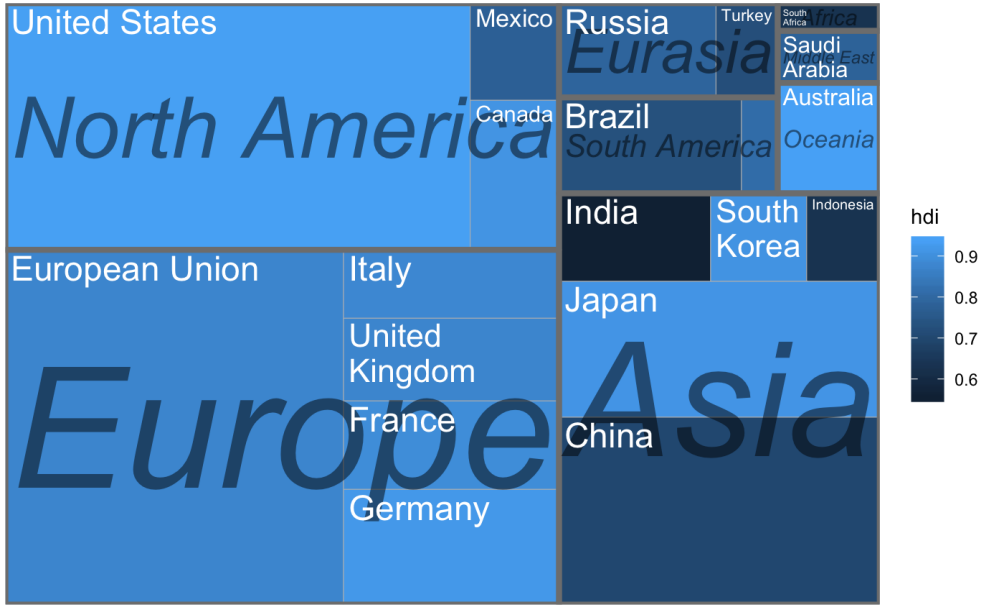
Bar plots

Any of the alternatives to bar plots

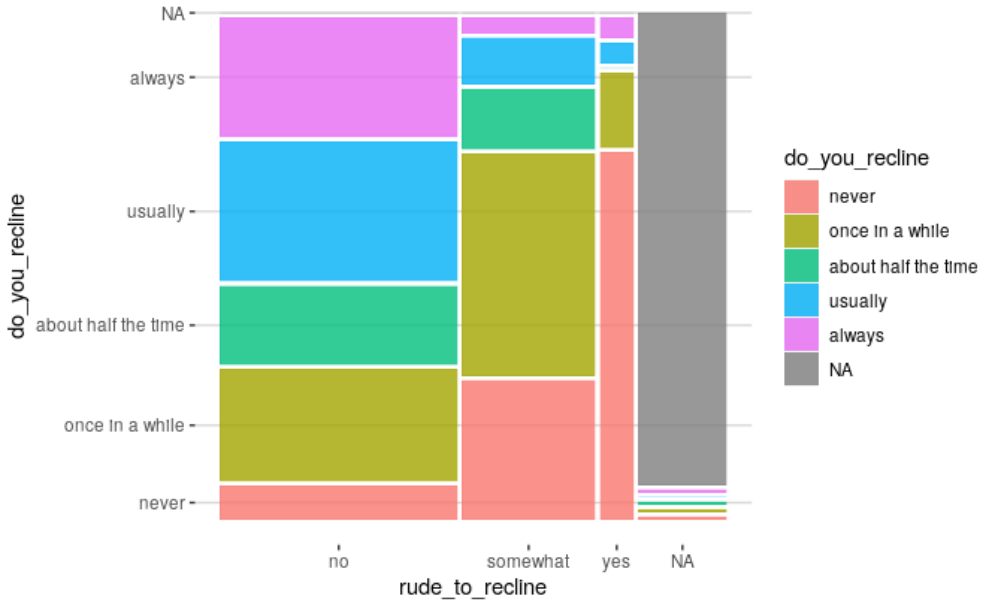
Treemaps and mosaic plots
(but these can still be really hard to interpret)

Treemaps and mosaic plots

Treemaps with the **treemapify** package



Mosaic plots with the **ggmosaic** package



Alternatives

Bar plots

Any of the alternatives to bar plots

Treemaps and mosaic plots
(but these can still be really hard to interpret)

Specialized figures like parliament plots

Parliament plots

Parliament plots with the **ggparliament** package