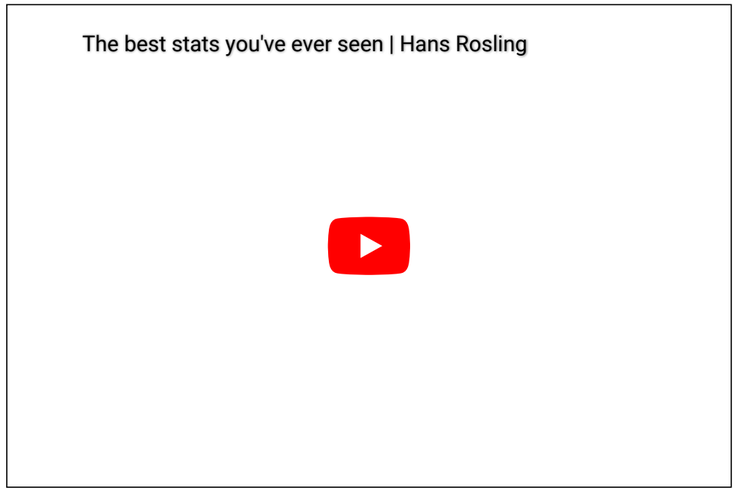
# Data visualization

# Purpose

1. Get you excited about storytelling with data
2. Show some tips and tricks to make your maps and charts pop

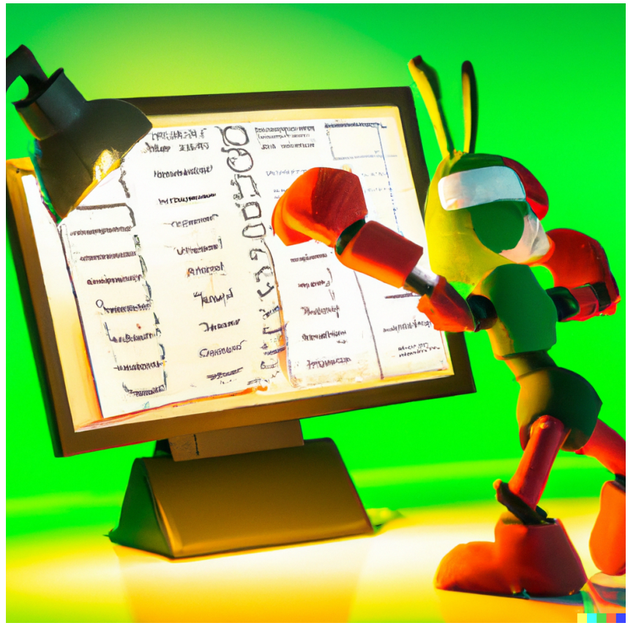



Hans Rosling's *the best stats you've ever seen*

This is not about the software. Rather, it is about the theory behind communicating well with data

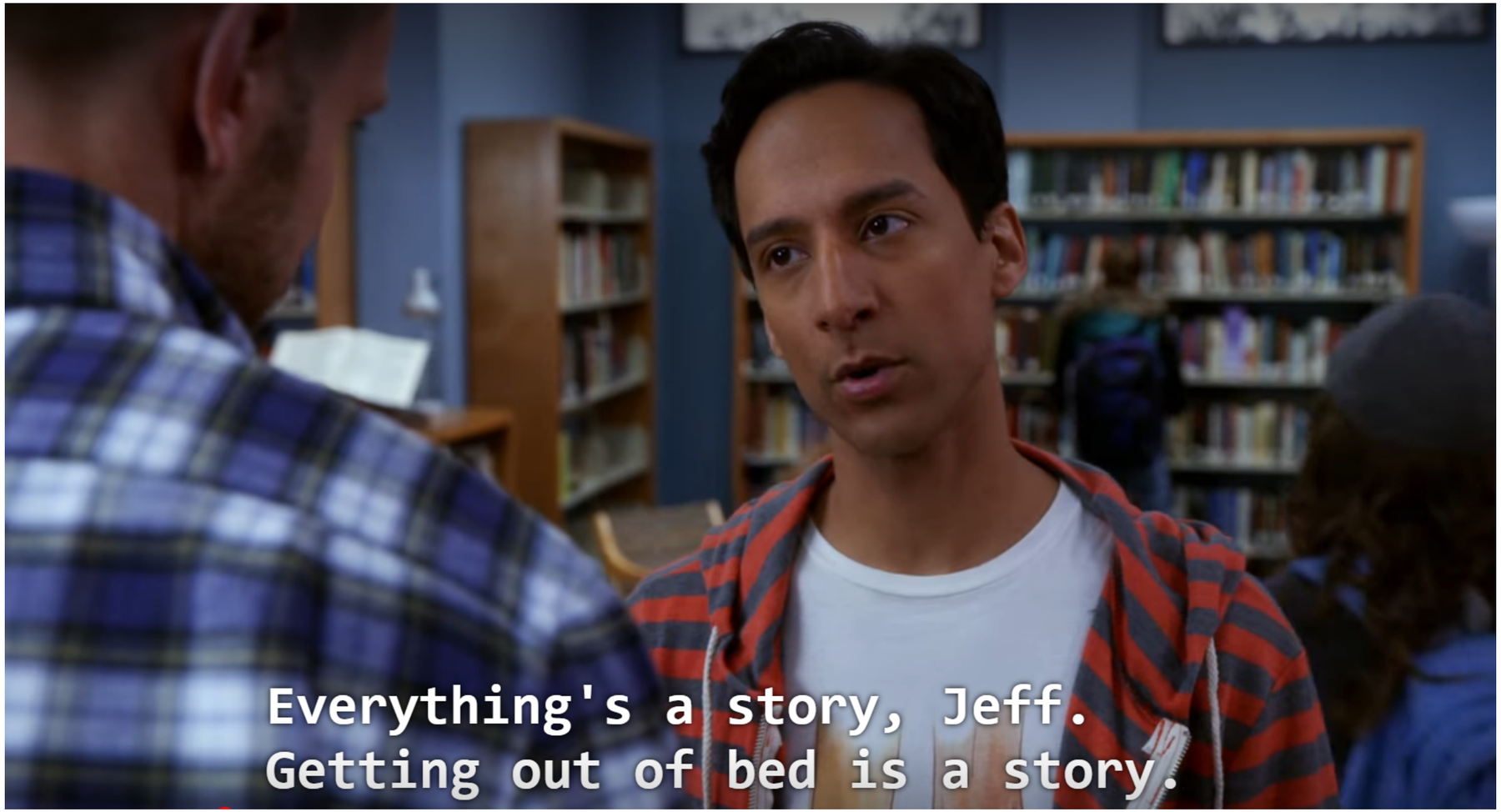
# Structure

1. Improving your maps
2. Overcoming Excel
3. Telling a story with data
4. Reproducing figures for publication

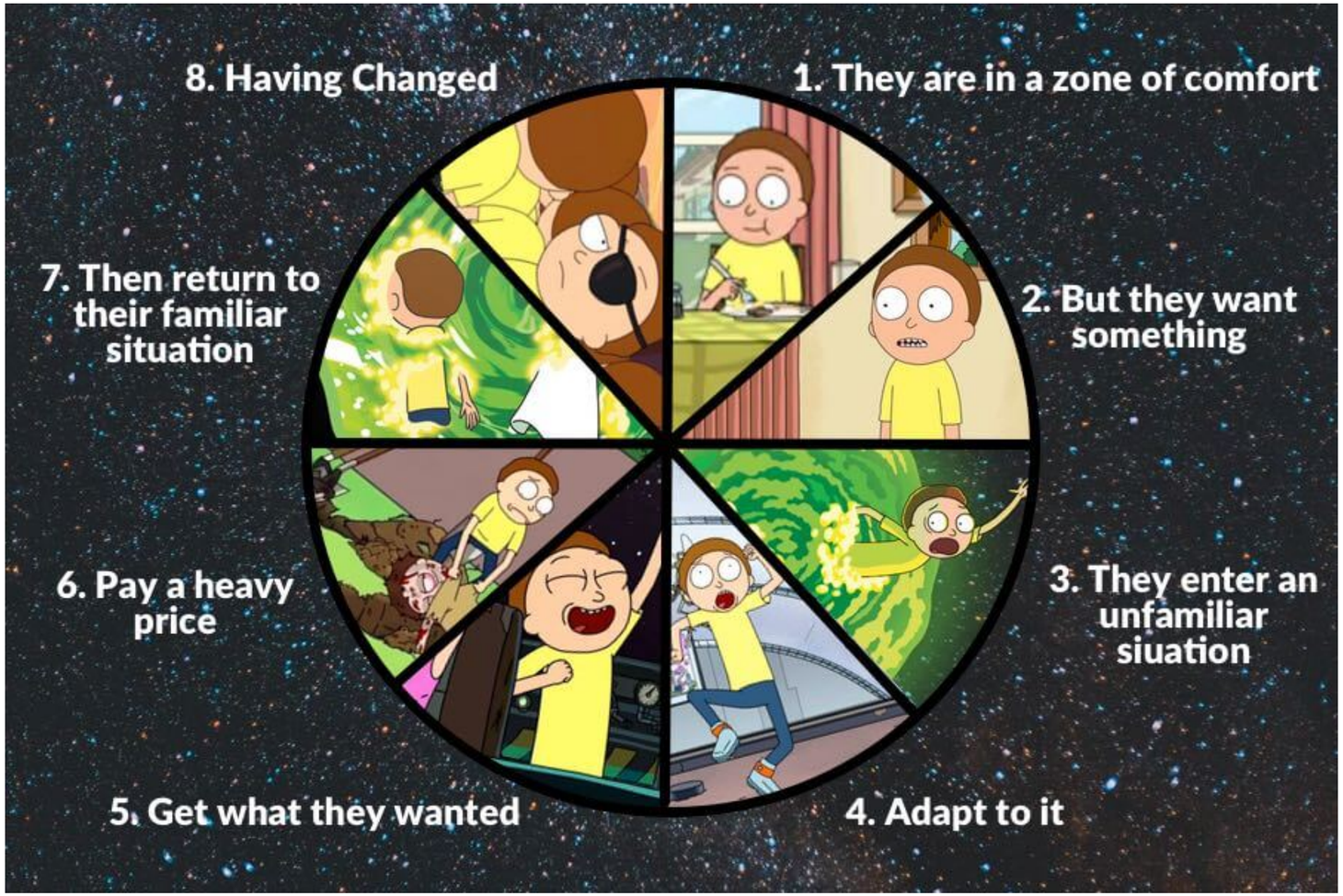


A cartoon in the style of Aardman Animations depicting an animatronic computer fighting with a spreadsheet, vibrant green lighting

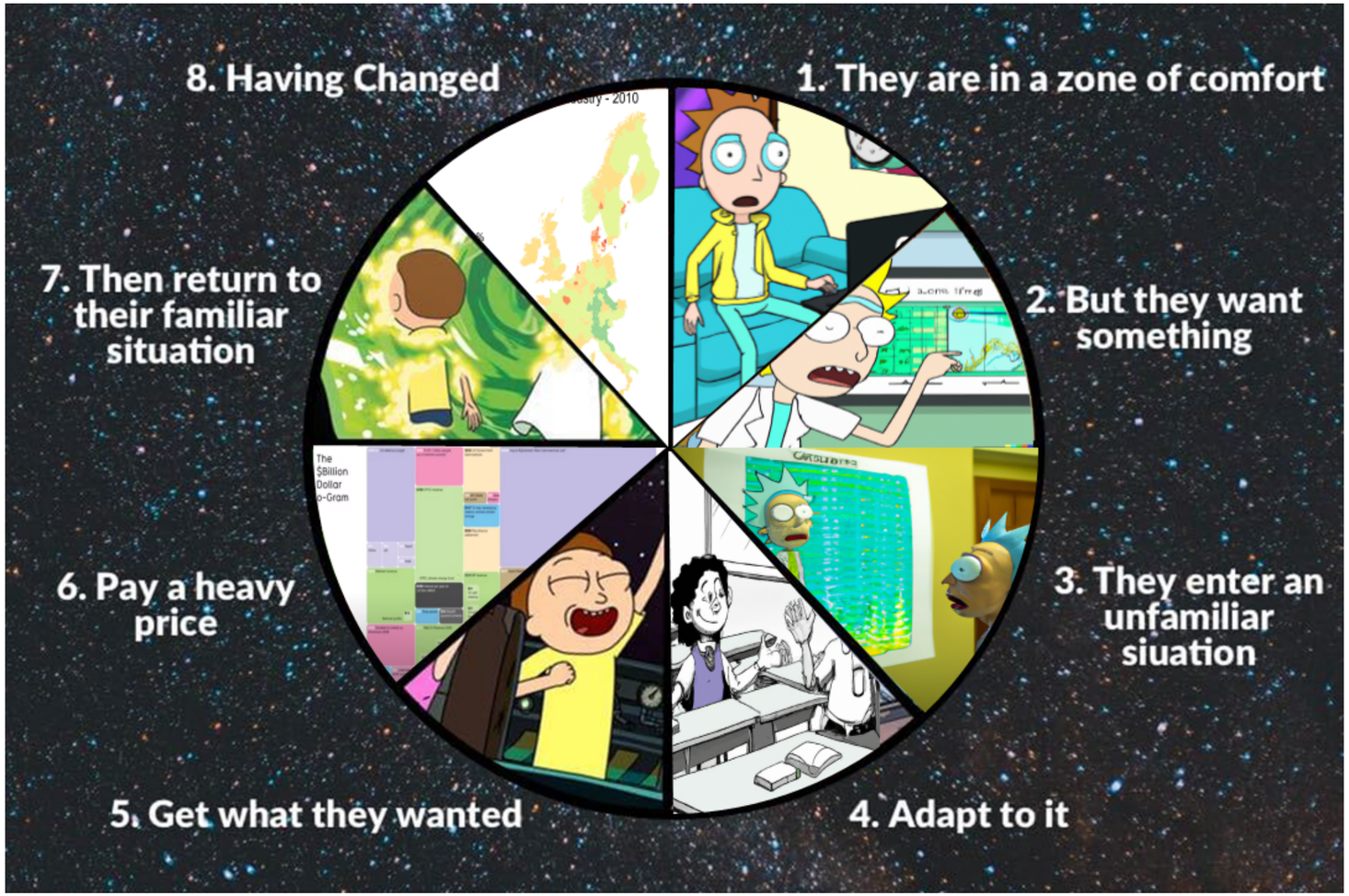
# Everything is a story



# Dan Harmon’s Story Circle



# Our Story Circle



# Improving your maps

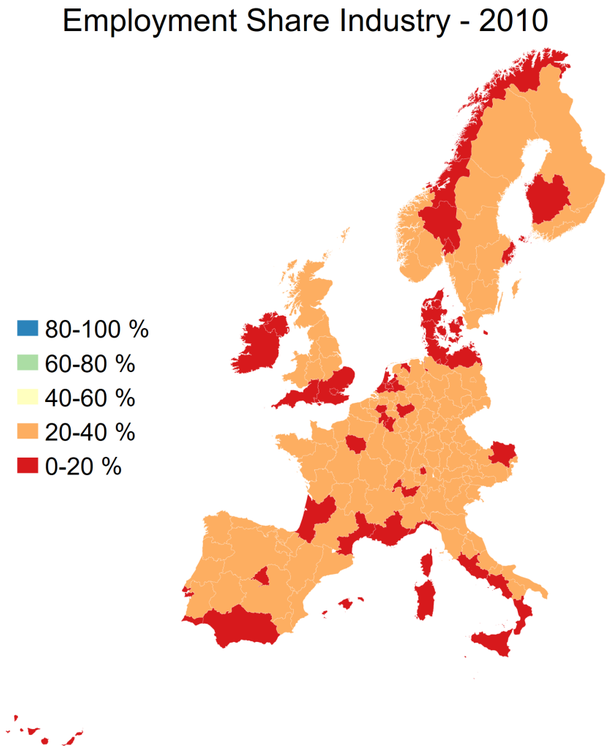
## Legend breaks

# Recap from Lab 1 exercies

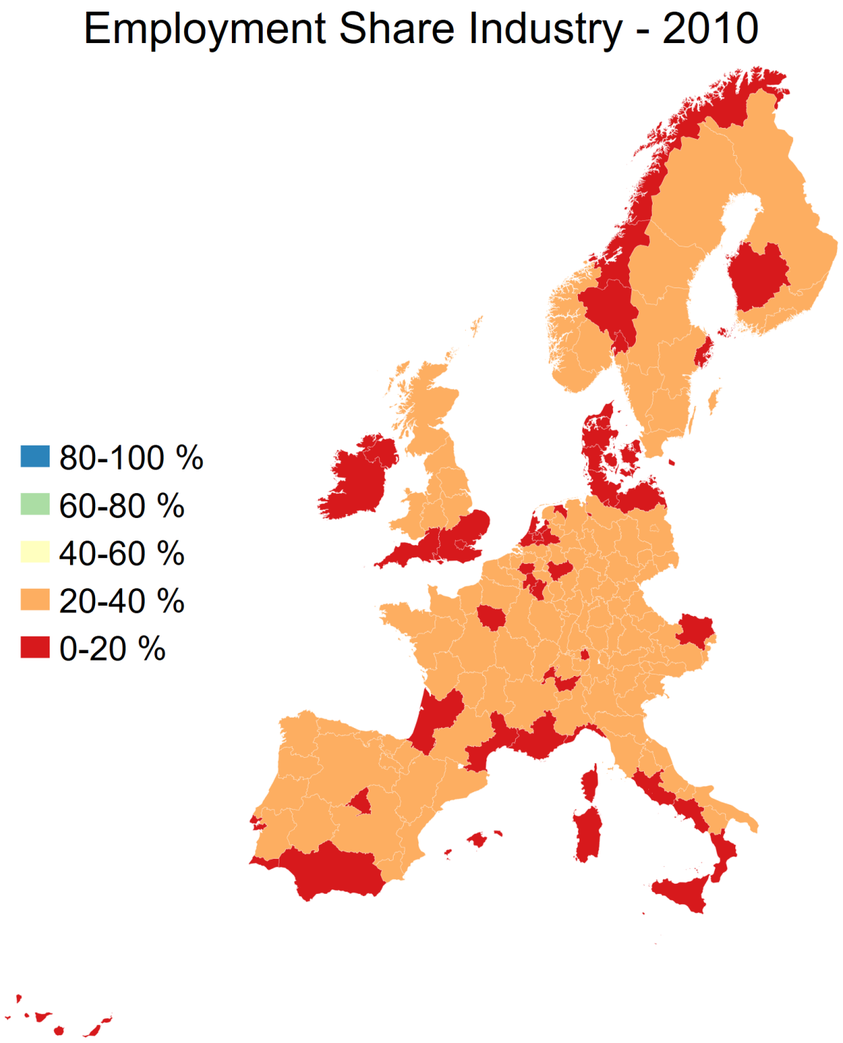
> Make a map of the share of employment in industry in the year 2010 across the whole dataset

# Recap from Lab 1 exercies

## What is wrong with this map? 🗺️



# Recap from Lab 1 exercies



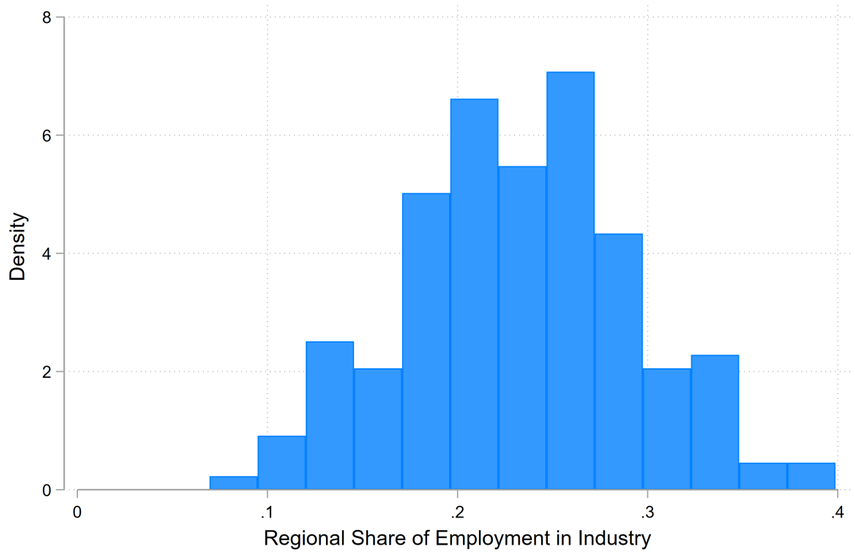

Discuss with your neighbour:

- What do we like?
- What is confusing?

```
spmap employment_share_industry using "nutscoord.dta"
  if year == 2010,
  id(_ID) fcolor(Spectral) legstyle(2)
    title("Employment Share Industry - 2010", size(larg
    osize(0.02 ..) ocolor(white ..)
    clmethod(custom) clbreaks(0 (0.2) 1)
    legend(pos(9) size(medium) rowgap(1.5)
    label(6 "80-100 %") label(5 "60-80 %")
    label(4 "40-60 %") label(3 "20-40 %") label(2 "0-20
    label(1 "No Data"))
    ndfcolor(gray) ndocolor(white ..) ndsize(0.02 ..)
```

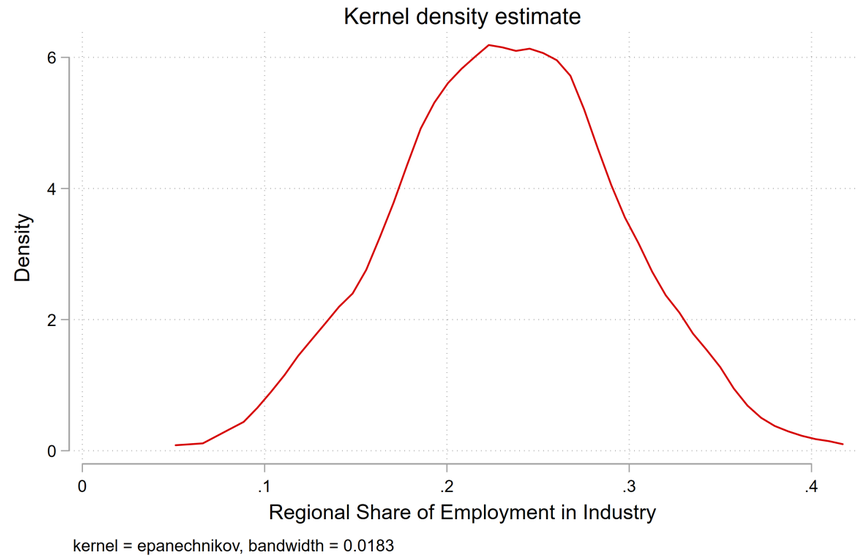
01:00

# Let’s plot the disribution of the data



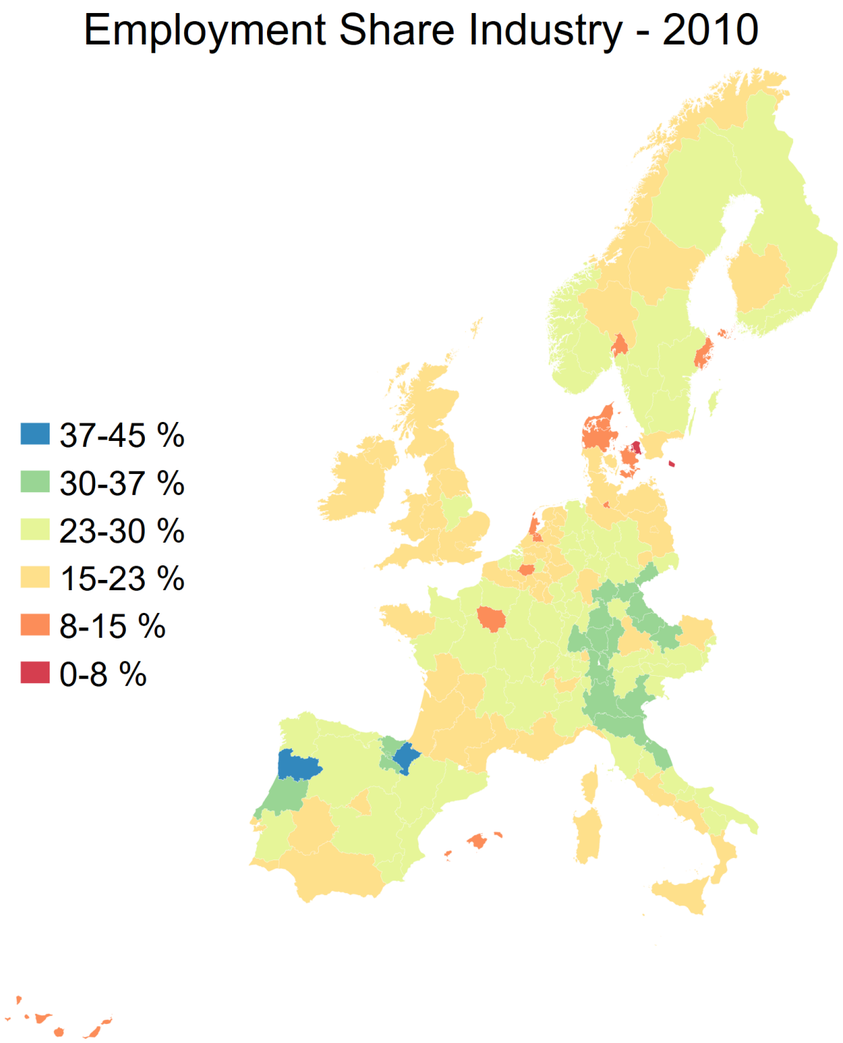

```
histogram employment_share_industry if year == 2010,
  color(midblue)
```

# Let’s plot the disribution of the data




```
1  kdensity employment_share_industry if year == 2010
```

# Now let’s make breaks based on this information



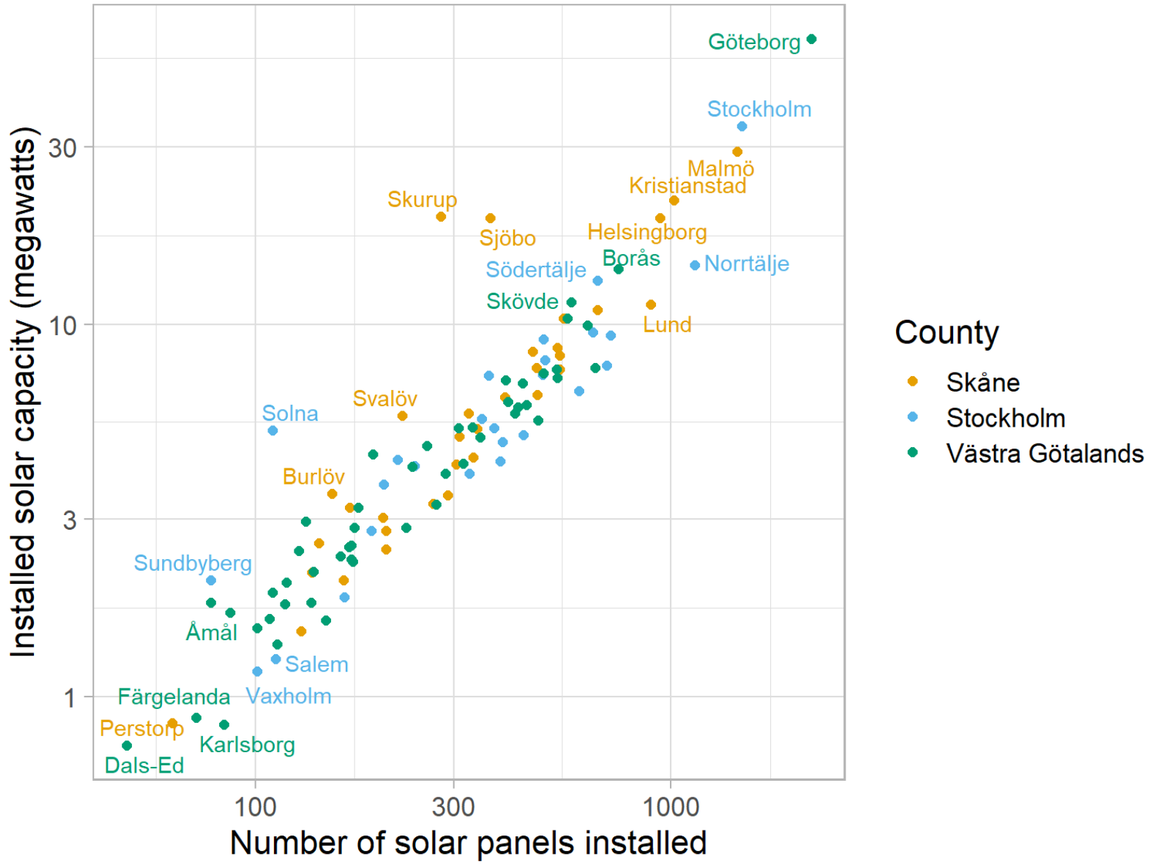

```
spmap employment_share_industry using "nutscoord.dta"
  if year == 2010, id(_ID) fcolor(Spectral) legstyle(2)
    title("Employment Share Industry - 2010", size(larg
    osize(0.02 ..) ocolor(white ..)
    clmethod(custom) clbreaks(0 (0.075) 0.5)
    legend(pos(9) size(medium) rowgap(1.5)
    label(7 "37-45 %") label(6 "30-37 %")
    label(5 "23-30 %") label(4 "15-23 %")
    label(3 "8-15 %") label(2 "0-8 %")
    label(1 "No Data"))
    ndfcolor(gray) ndocolor(white ..) ndsize(0.02 ..)
```

# Colour scales

Uses of color in data visualization

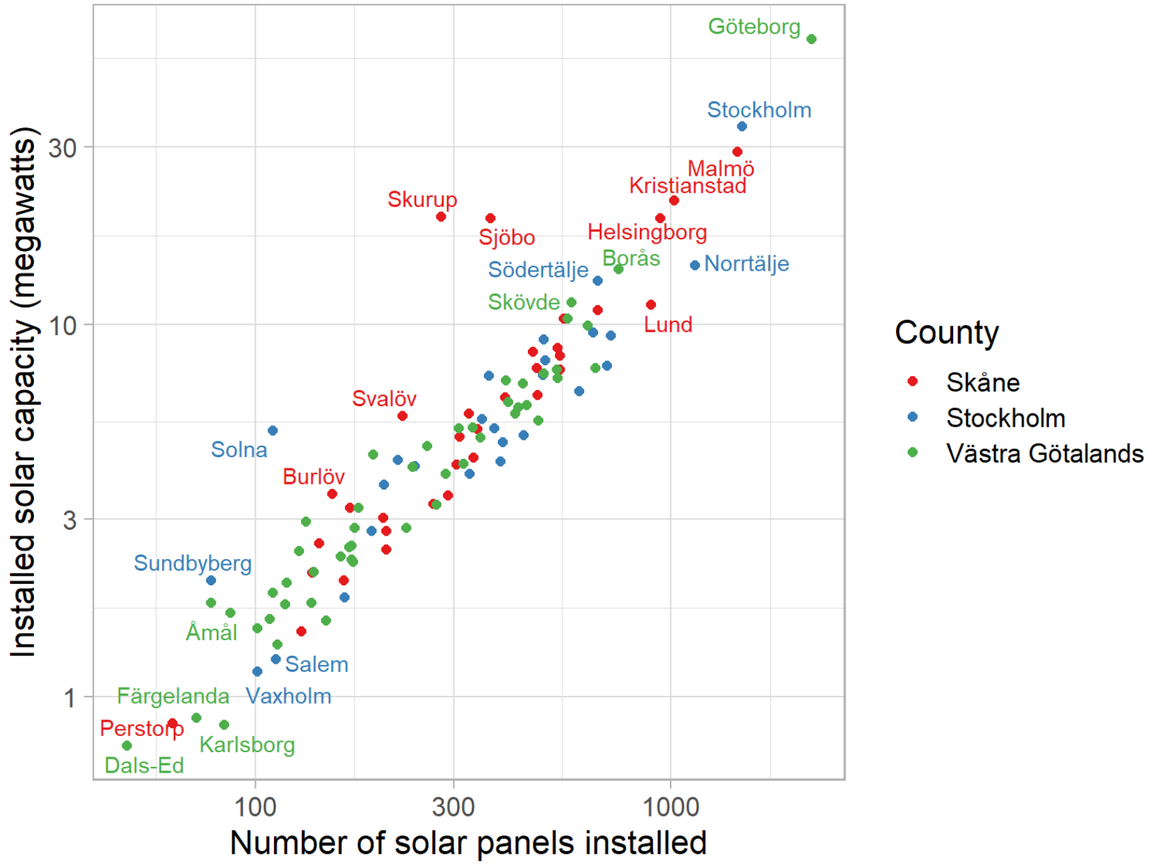
1. Distinguish categories (qualitative)

# Qualitative scale example



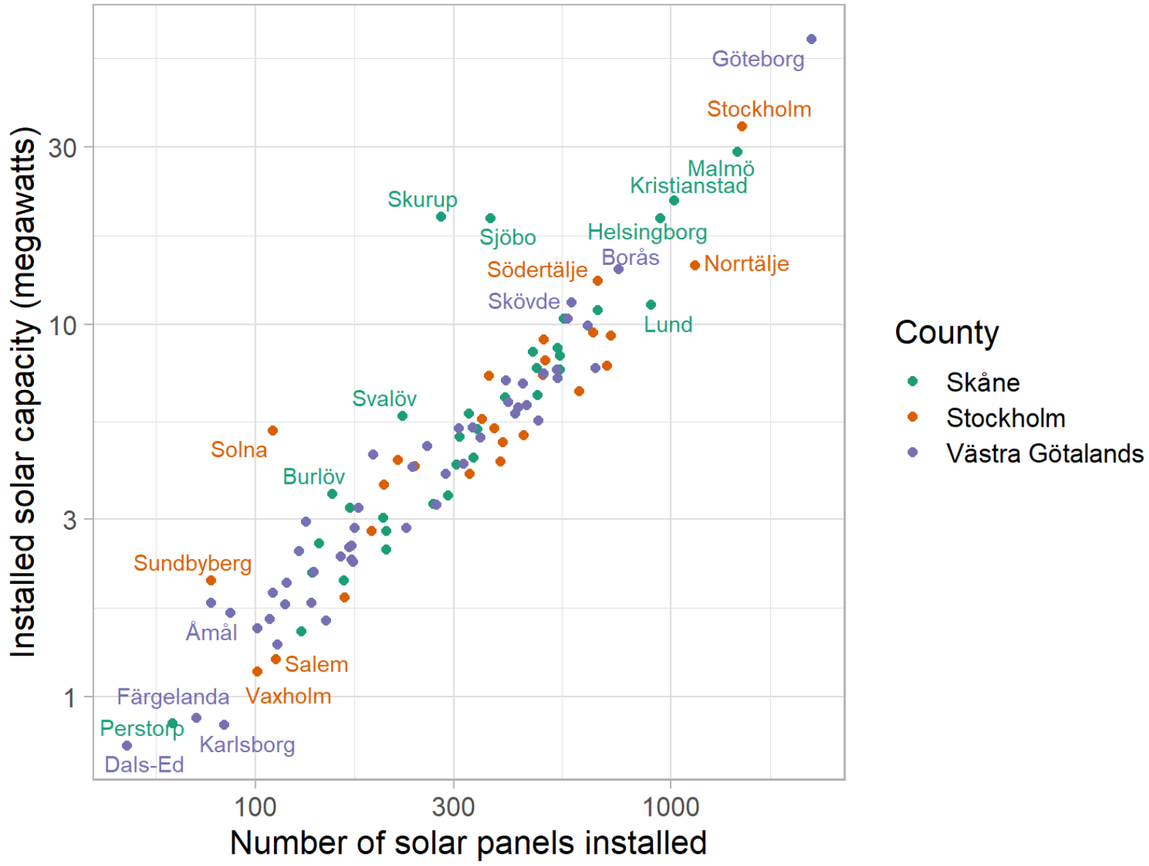

Palette name: Okabe-Ito

# Qualitative scale example




Palette name: Brewer Set1

# Qualitative scale example




Palette name: Brewer Dark2

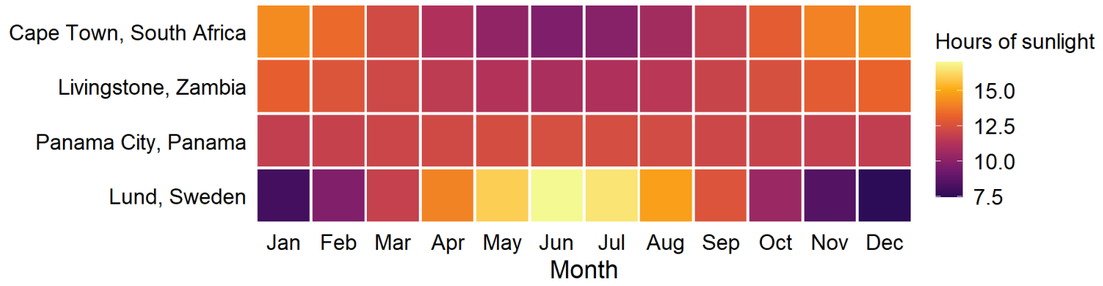
# Uses of color in data visualization

1. Distinguish categories (qualitative)

---

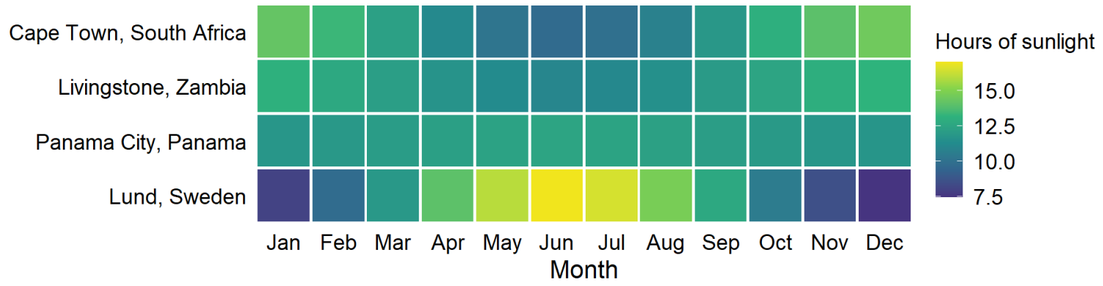
2. Represent numeric values (sequential)

# Sequential scale example




Palette name: inferno

# Sequential scale example




Palette name: viridis

# Uses of color in data visualization

1. Distinguish categories (qualitative)

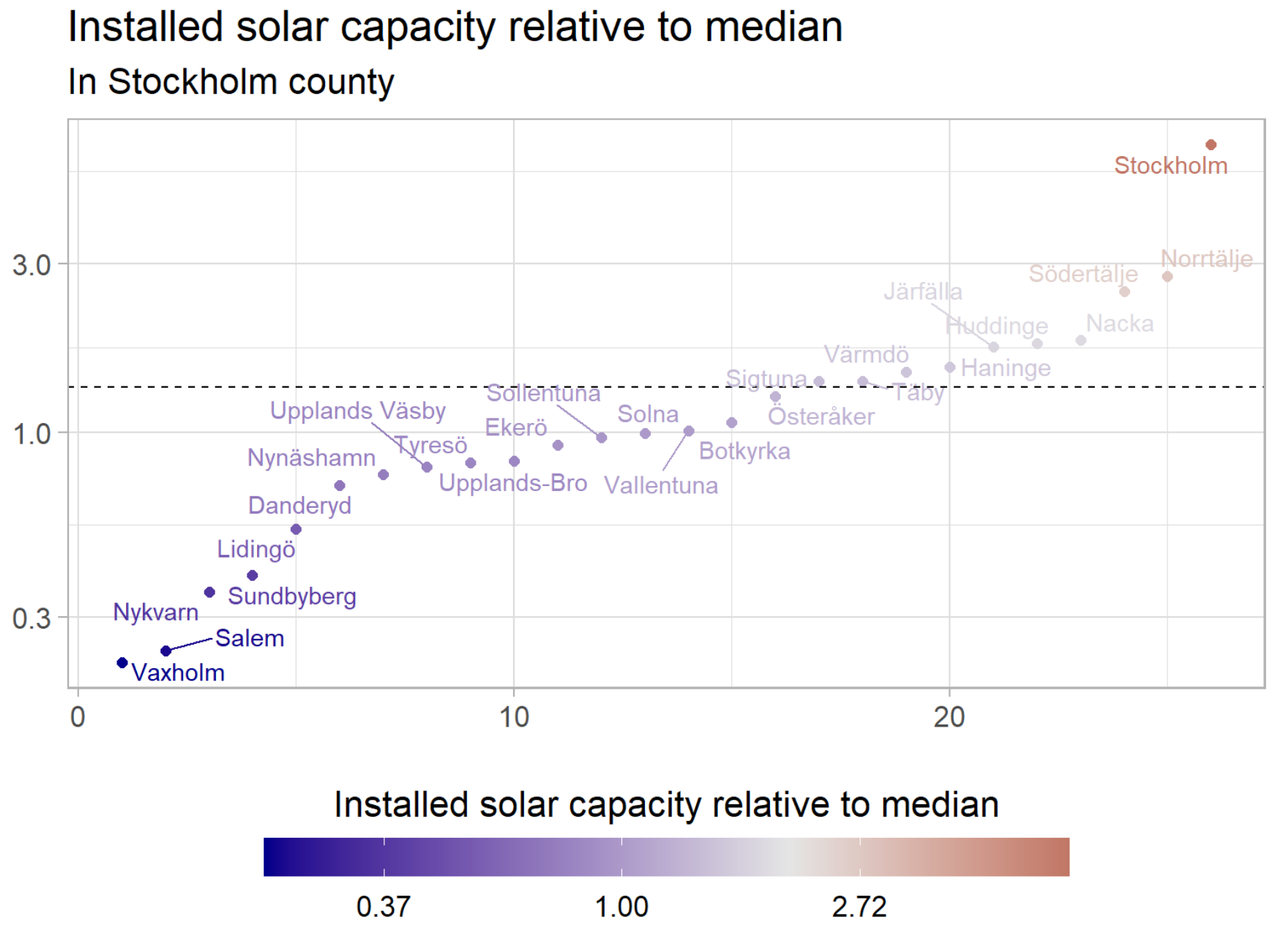
---

2. Represent numeric values (sequential)

---

3. Represent numeric values (diverging)

# Diverging scale example



# Uses of color in data visualization

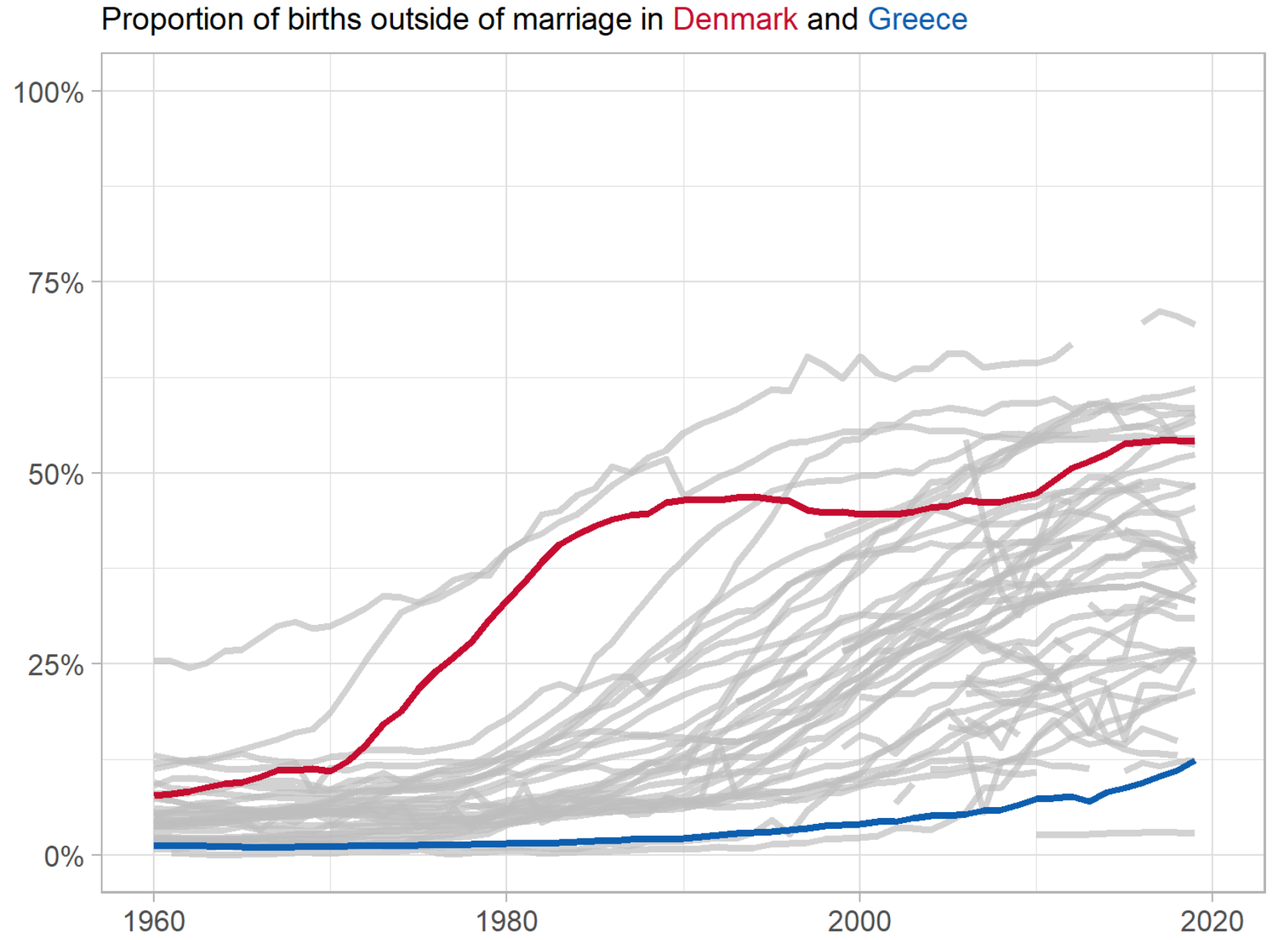
1. Distinguish categories (qualitative)

2. Represent numeric values (sequential)

3. Represent numeric values (diverging)

4. Highlight

# Highlight example



# Using density plots to set your legend breaks

Dataset: Solar panels in Sweden

**Installed solar capacity in Sweden**

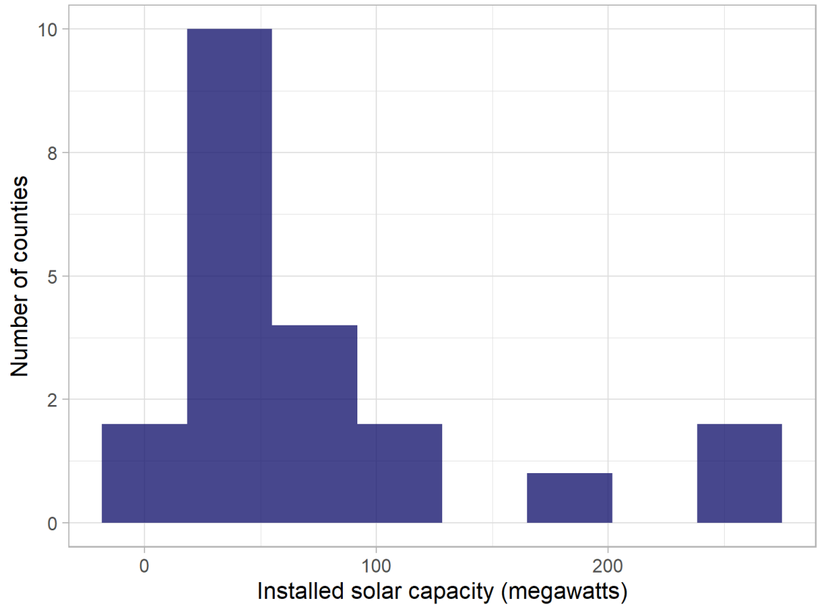
Year: 2021

| Swedish county | Installed solar capacity (megawatts) |
|---|---|
| Västra Götalands län | 266.21 |
| Skåne län | 256.25 |
| Stockholms län | 182.25 |
| Östergötlands län | 106.81 |
| Hallands län | 94.31 |
| Jönköpings län | 88.53 |
| Södermanlands län | 79.71 |
| Uppsala län | 79.11 |
| Kalmar län | 59.01 |
| Västmanlands län | 49.45 |

Source: Energimyndigheten

# How to decide on values for the bins?

Use a histogram or a density plot to see where the weight of the distribution is.

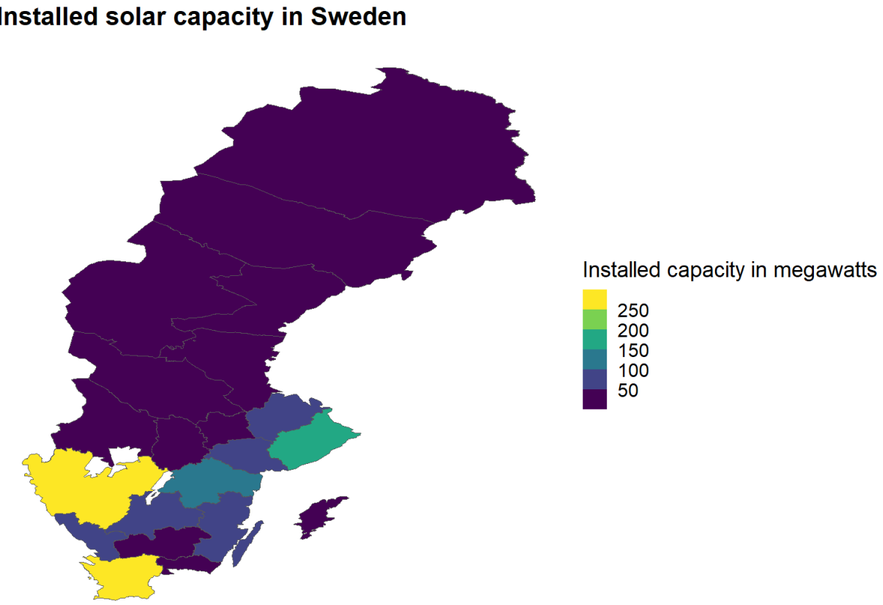



```
Reading layer `sverige-lan-counties-of-sweden' from data source
  `C:\Users\User\Documents\Recon\EOSE09-site\lectures\lecture-1-resources\maps\sverige-lan-counties-of-sweden.shp'
  using driver `ESRI Shapefile'
Simple feature collection with 21 features and 2 fields
Geometry type: MULTIPOLYGON
Dimension:     XY
Bounding box:  xmin: 10.95051 ymin: 55.32758 xmax: 24.17761 ymax: 69.05997
Geodetic CRS:  WGS 84
```

# Map with appropriate breaks

Ask your neighbour:

1. what kind of palette is this?
2. Is it appropriate to use with this data?



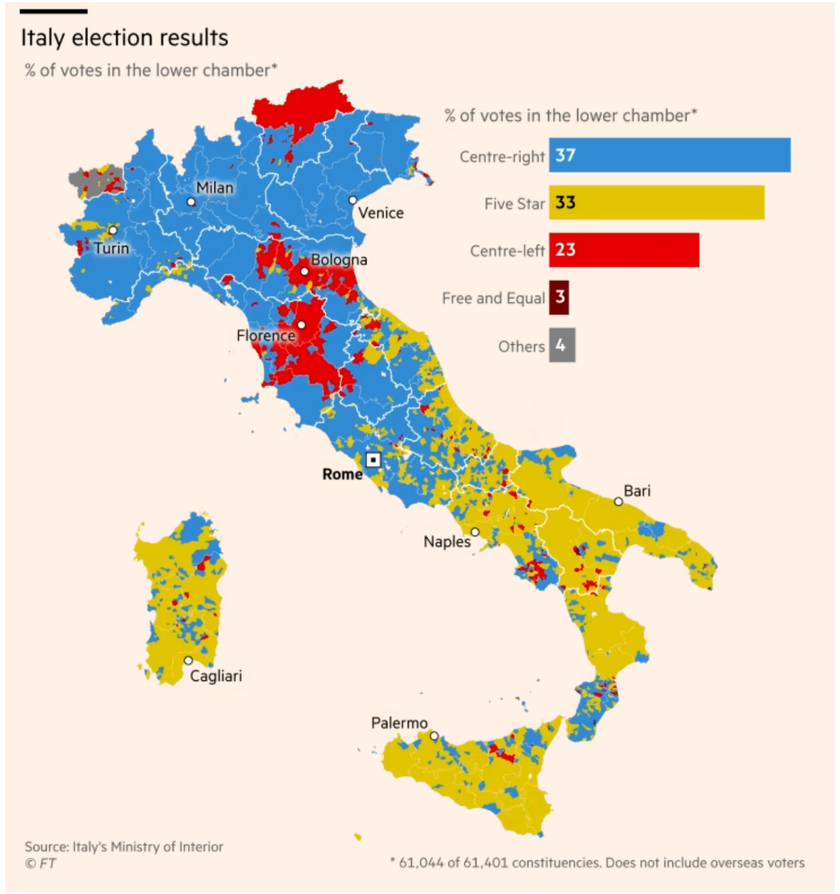

01:00

# Improving your maps

## Great Choropleths

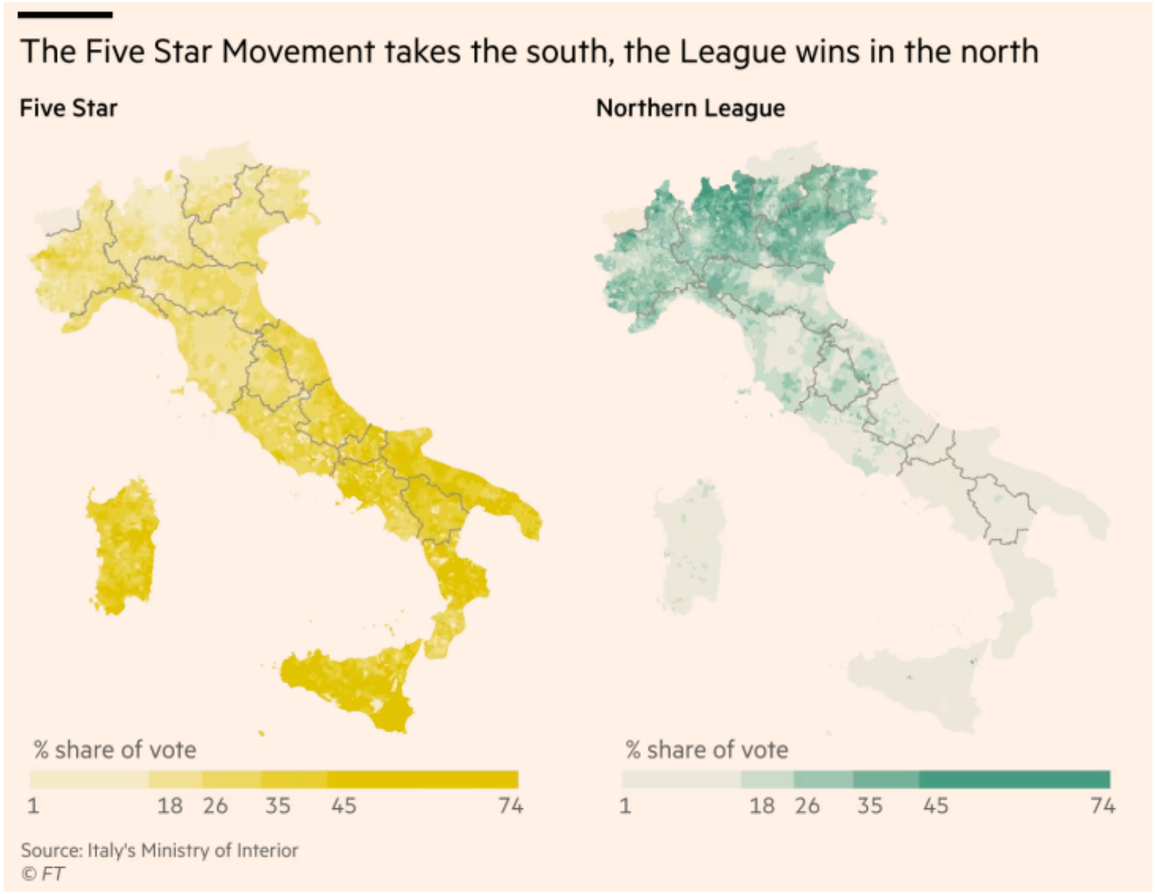
# Examples of great maps

## Financial Times analysis of Italian election results



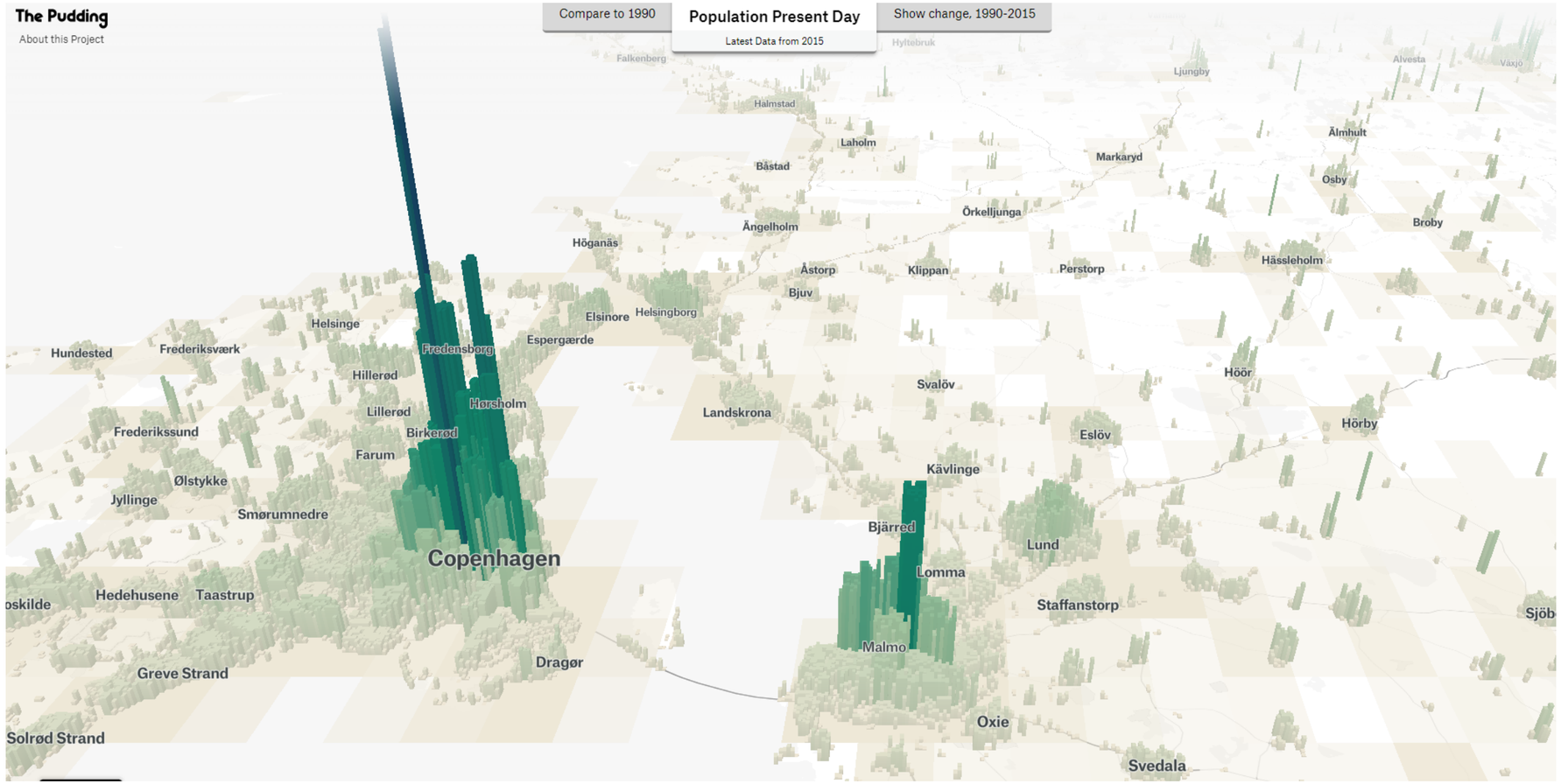
# Examples of great maps

## Financial Times analysis of Italian election results



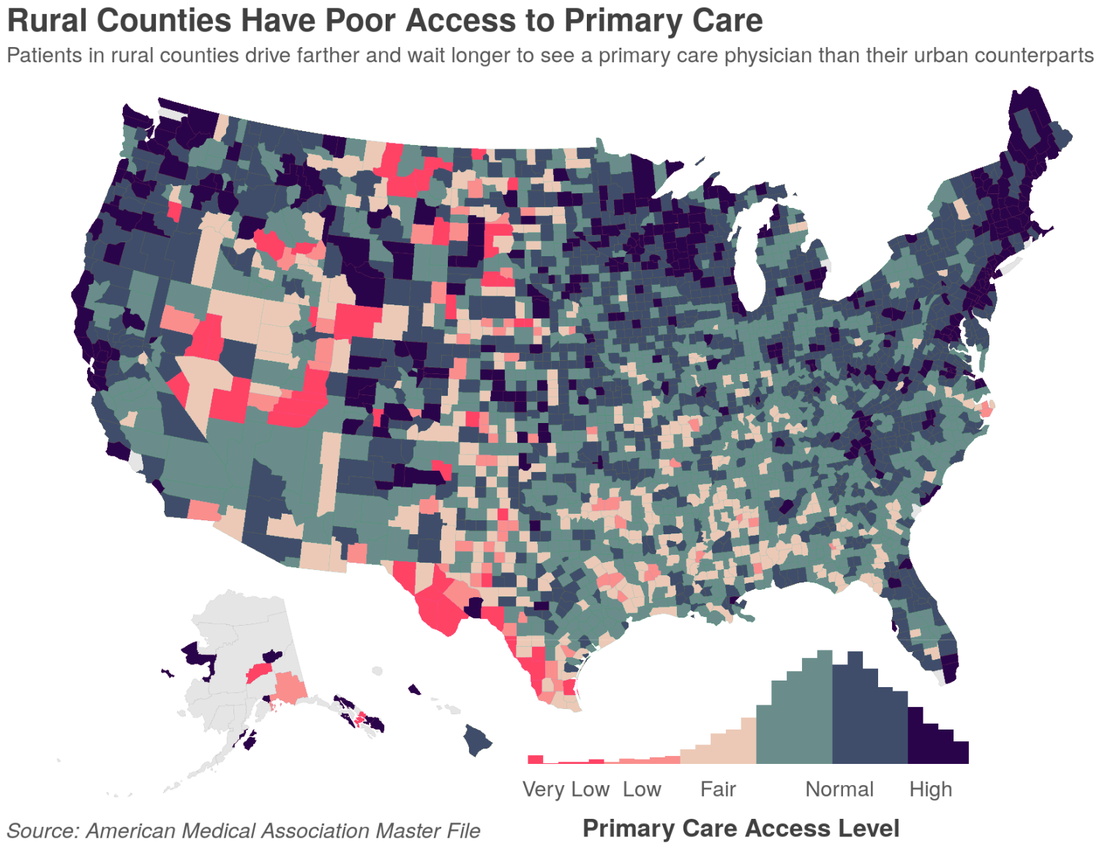
# Examples of great maps

## Human Terrain from The Pudding



# Examples of great maps

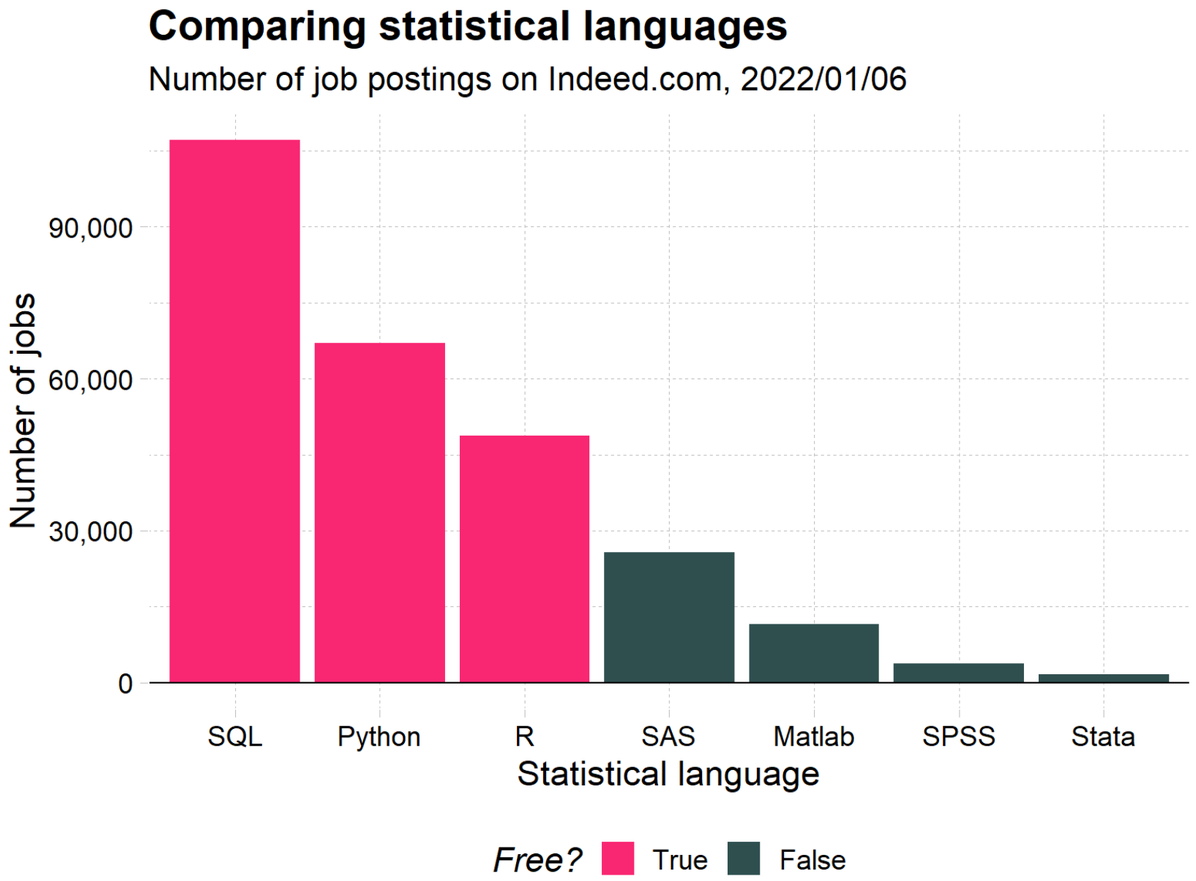
## The Coming Crisis: Exploring the U.S. Physician Shortage by Daniel Snow



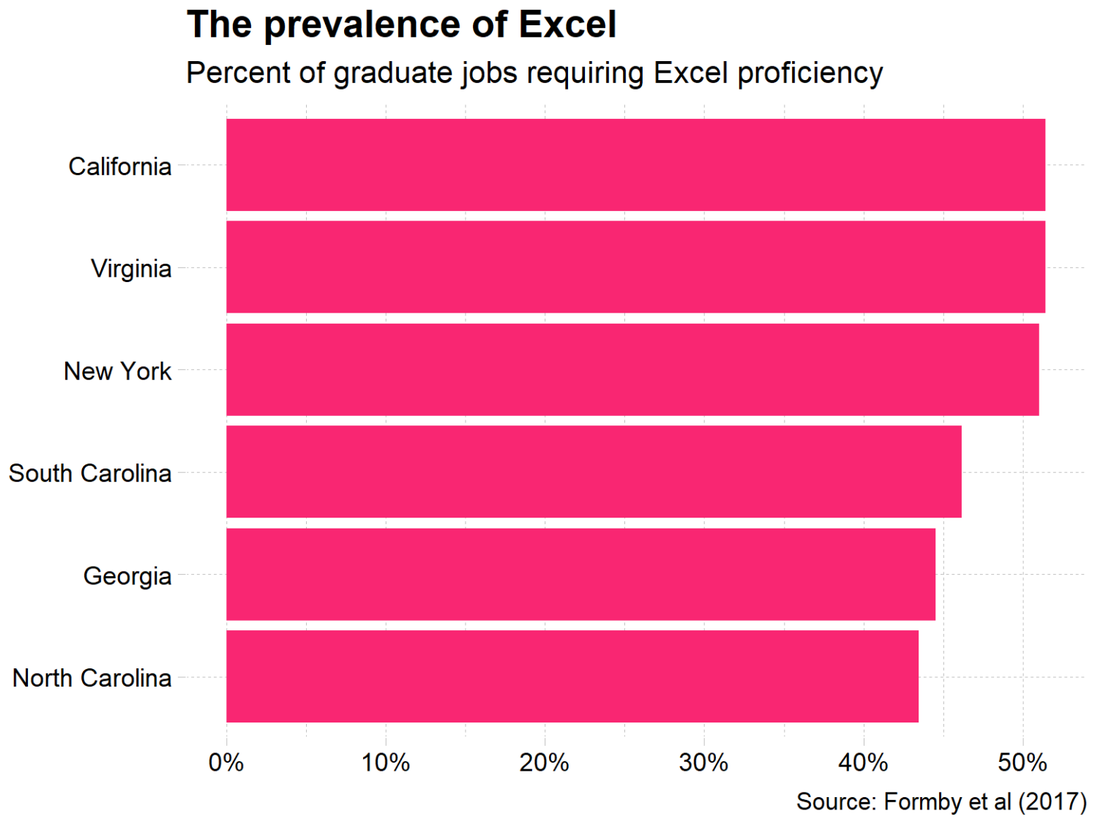
# Overcoming Excel

## Motivation

# Overcoming Excel



# Overcoming Excel




*Formby et al* (2017) Microsoft Excel: Is It An Important Job Skill for College Graduates?

# Overcoming Excel

## Takeaways:

1. You will likely use Excel in the future 📊
2. Excel's default plots and tables can be improved upon 📈
3. Simple rules can help you make your message clear 💎

# Overcoming Excel

## Charts

# Overcoming Excel: Column plot

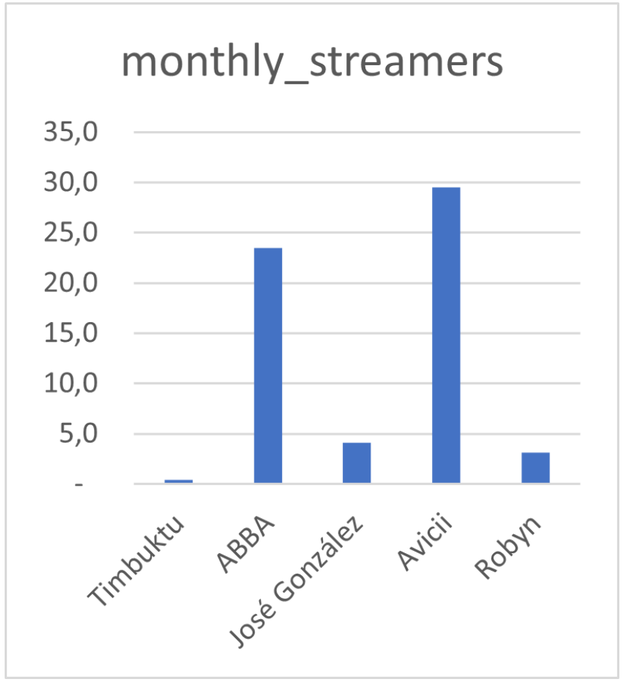
- We often encounter datasets containing simple amounts 👊
- Here is some data on a sample of Swedish musical artists 🎶
- I put this data into Excel, and asked for a recommended chart 📊

**Swedish musical artists**

| Rank | Artist | Monthly listeners (m) |
|---|---|---|
| 1 | Avicii | 29.47 |
| 2 | ABBA | 23.48 |
| 3 | José González | 4.07 |
| 4 | Robyn | 3.11 |
| 5 | Timbuktu | 0.38 |

Datasource: Spotify charts Nov 2022

# Your turn



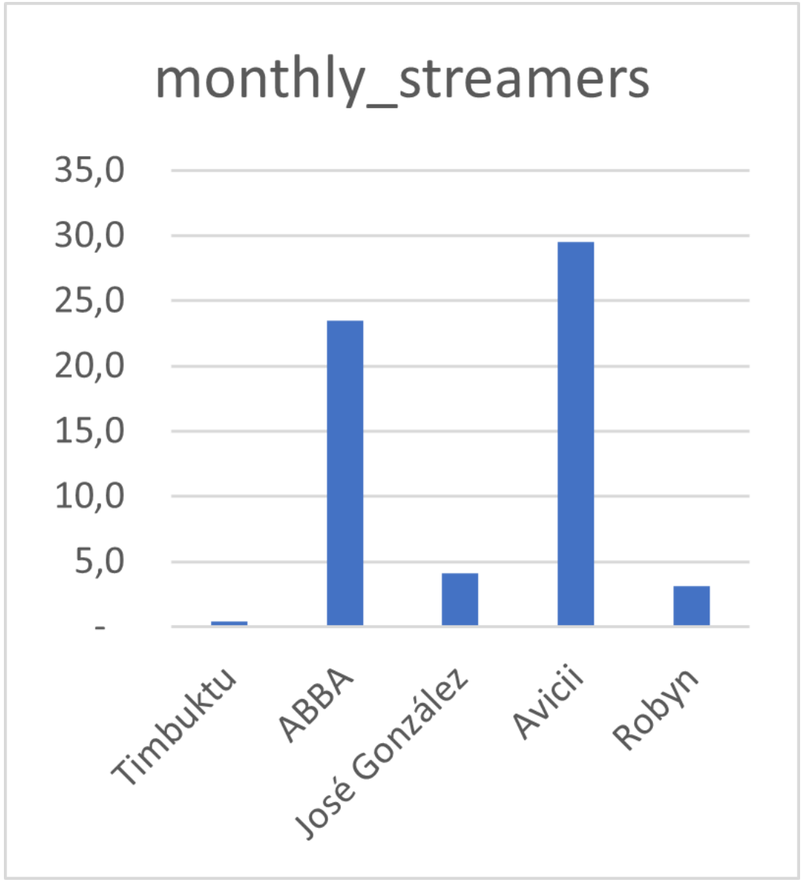

Discuss with your neighbour:

- What do we like?
- What is confusing?

02:30

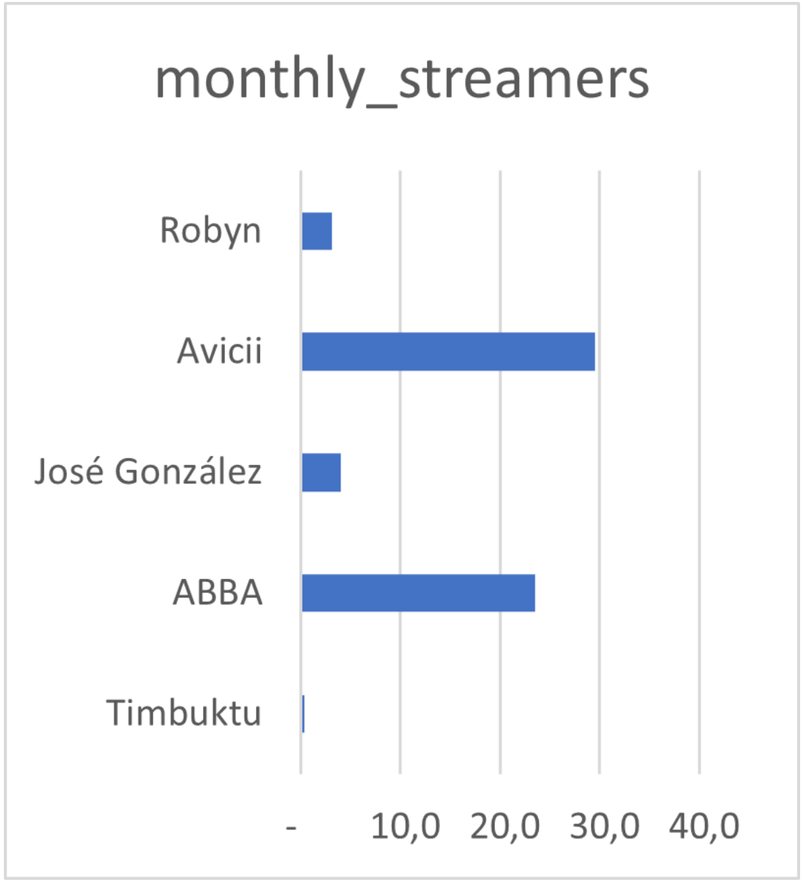
# Tip 1: Avoid rotated axis labels

## Ugly 🤢



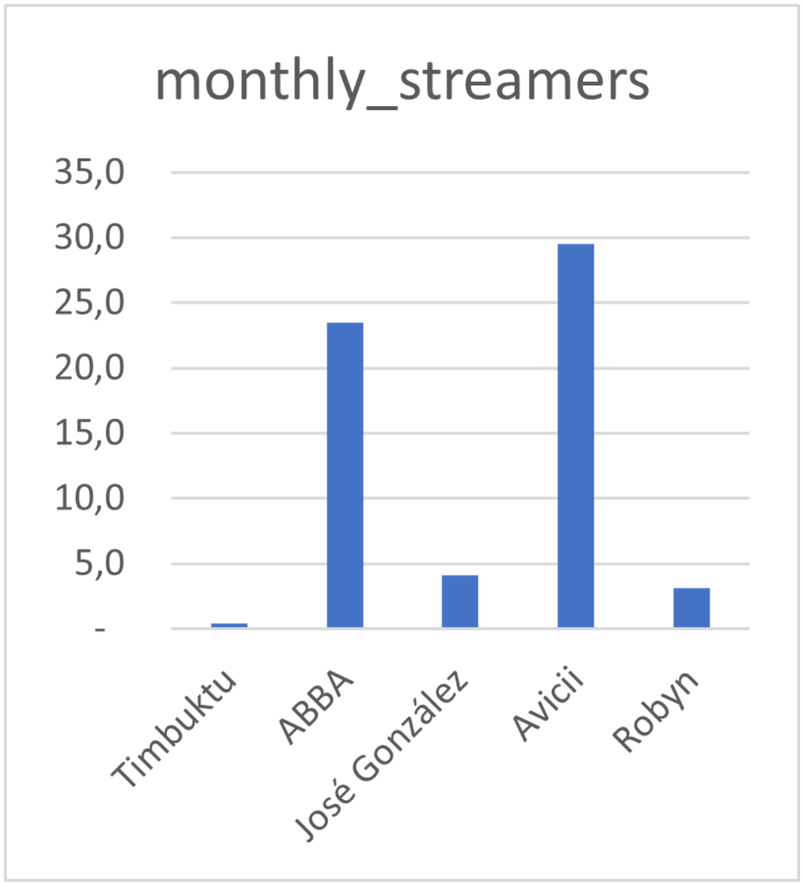
# Tip 1: Avoid rotated axis labels

## Flip axes so that the text is easier to read 👓



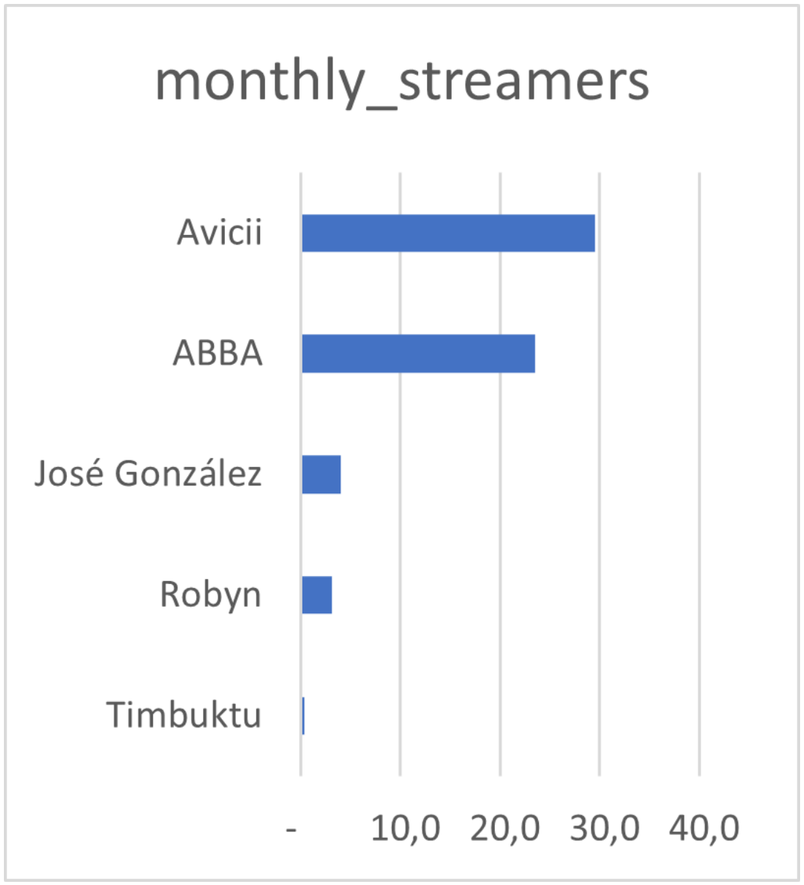
# Tip 2: Pay attention to the order of the bars

## Bad 👎



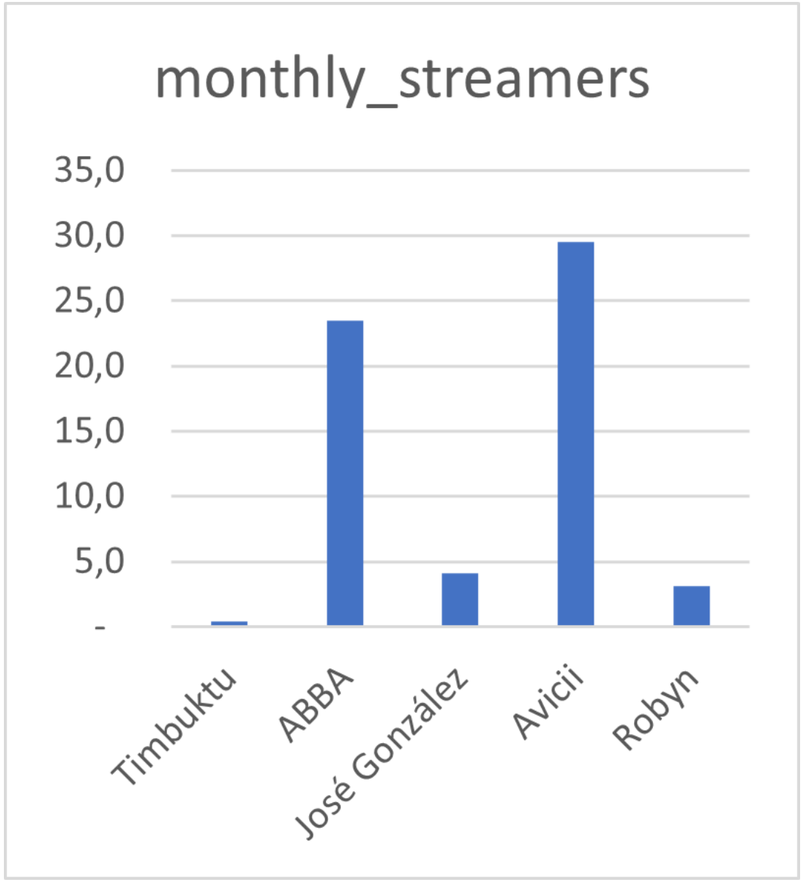
# Tip 2: Pay attention to the order of the bars

## It is clear that José González recieves more streams than Robyn



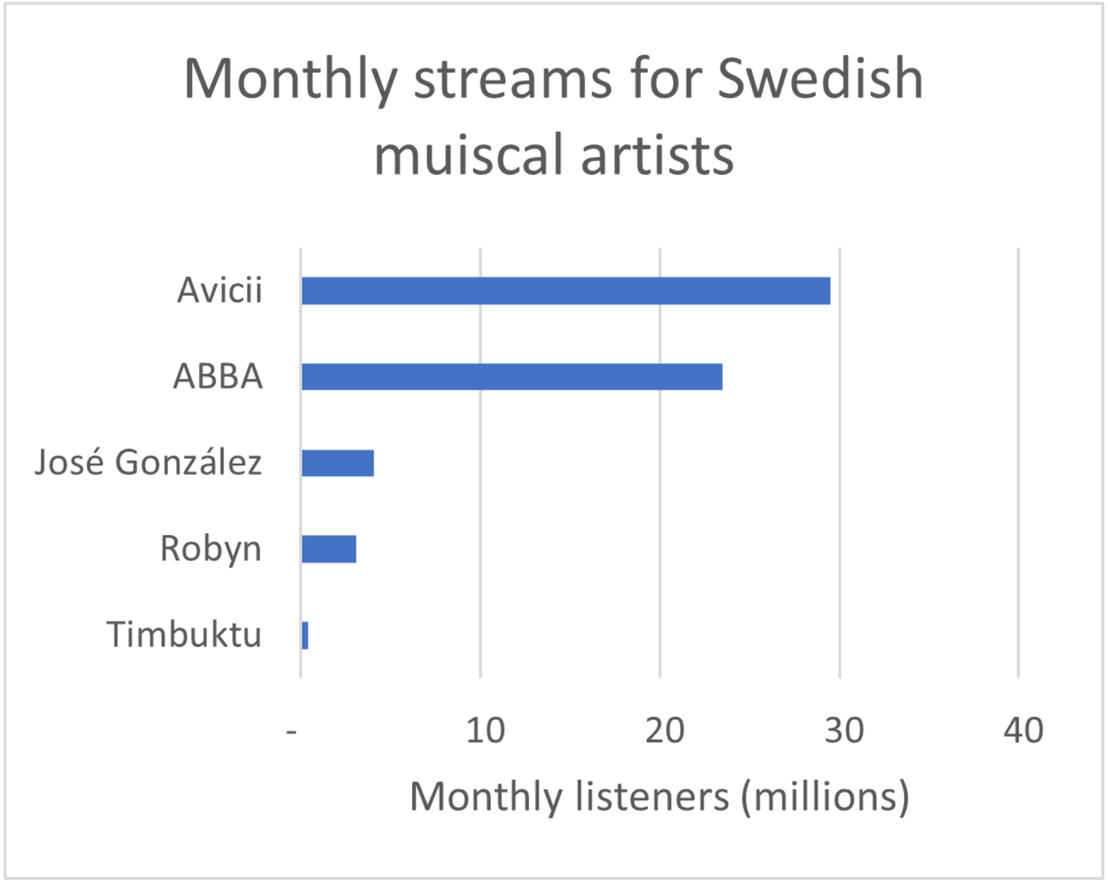
# Tip 3: Consider your titles, labels and axes

## Uninformative ❗



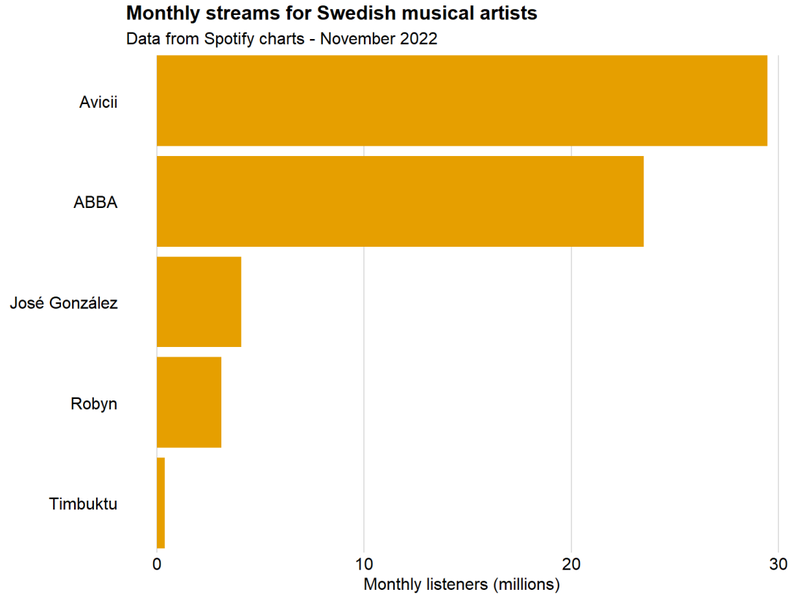
# Tip 3: Consider your titles, labels and axes

## Note the title, x-axis title, x-axis labels 📙



# Tip 3: Consider your titles, labels and axes

## Titles and captions have different application areas



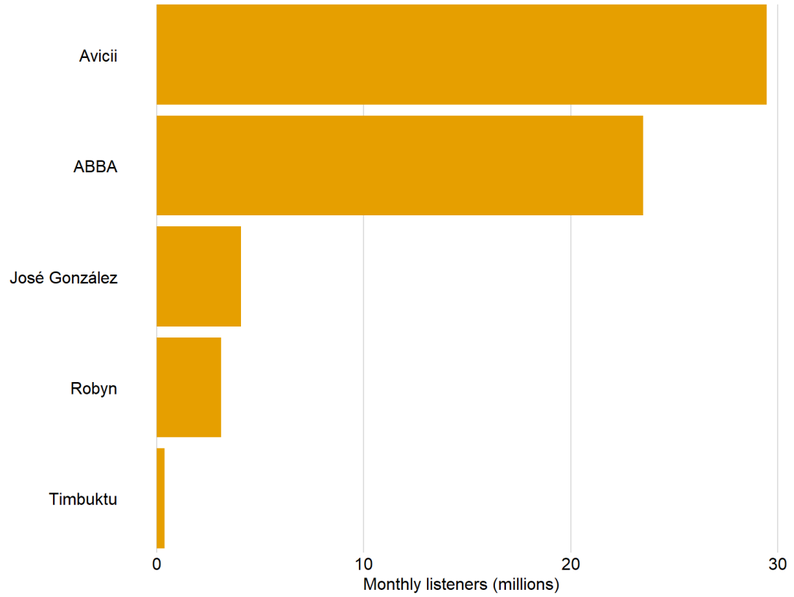
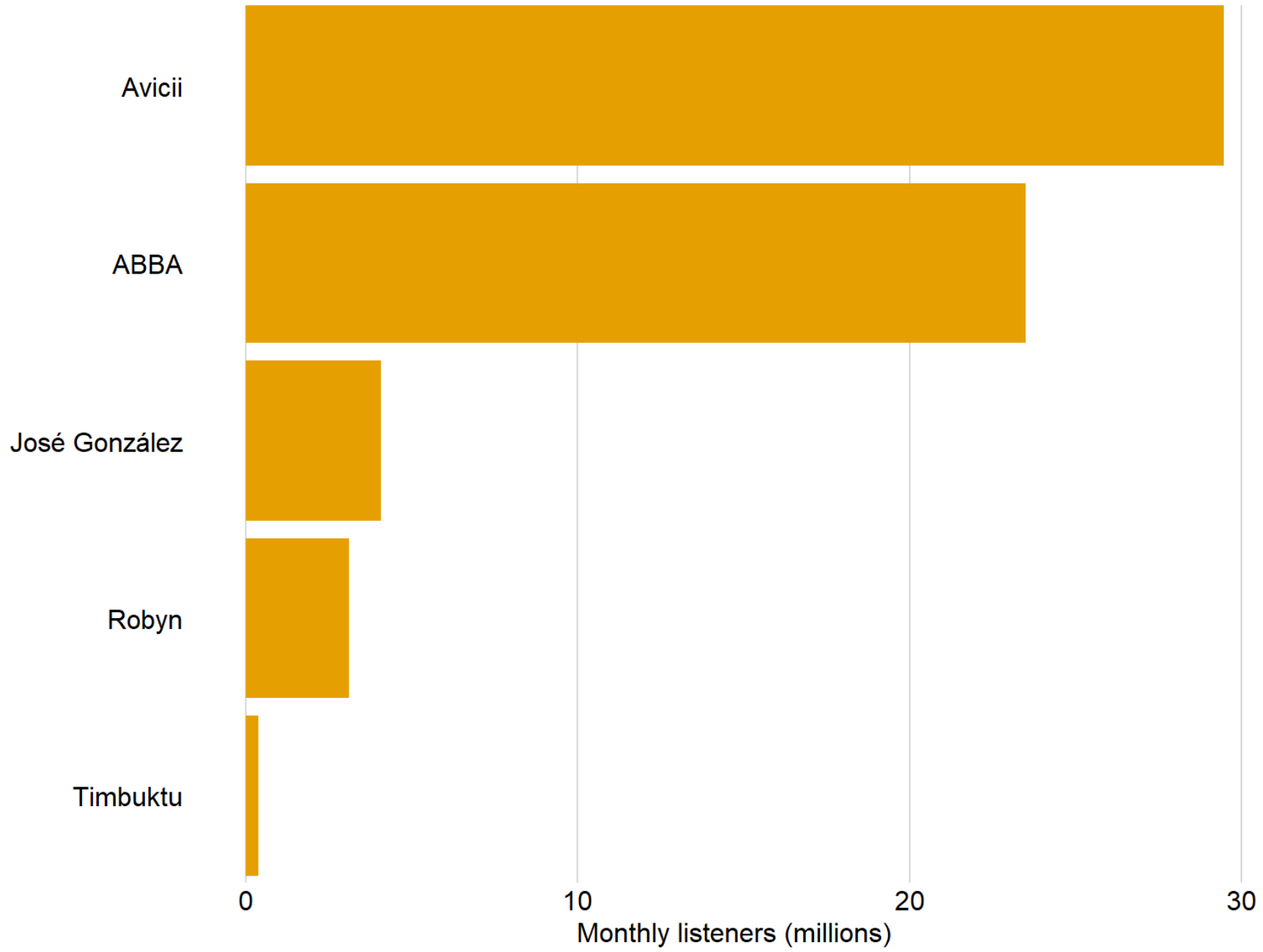
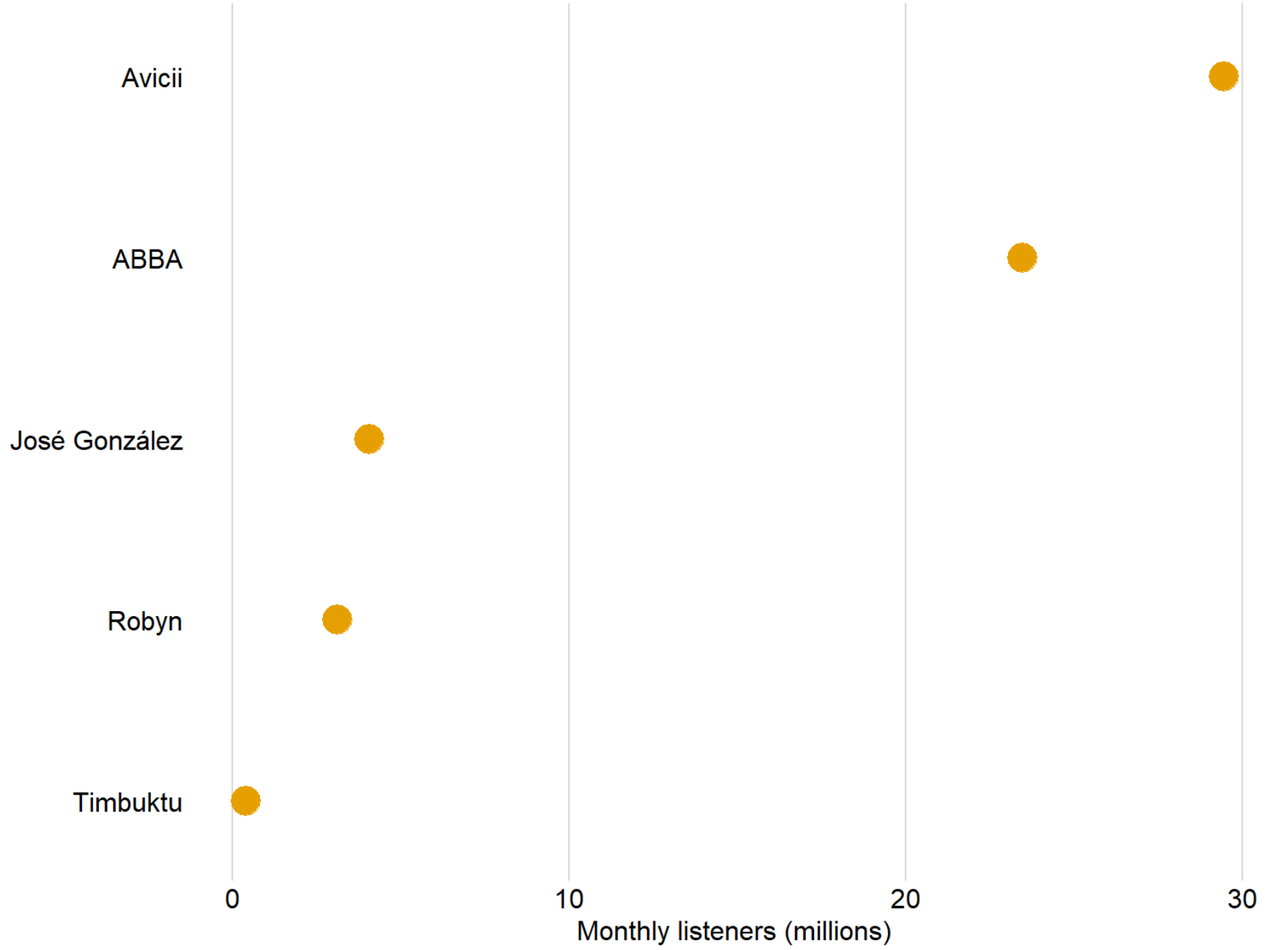




Figure 1: Monthly streams for Swedish musical artists. Data sources: Spotify charts in November 2022

# We can use dots instead of bars

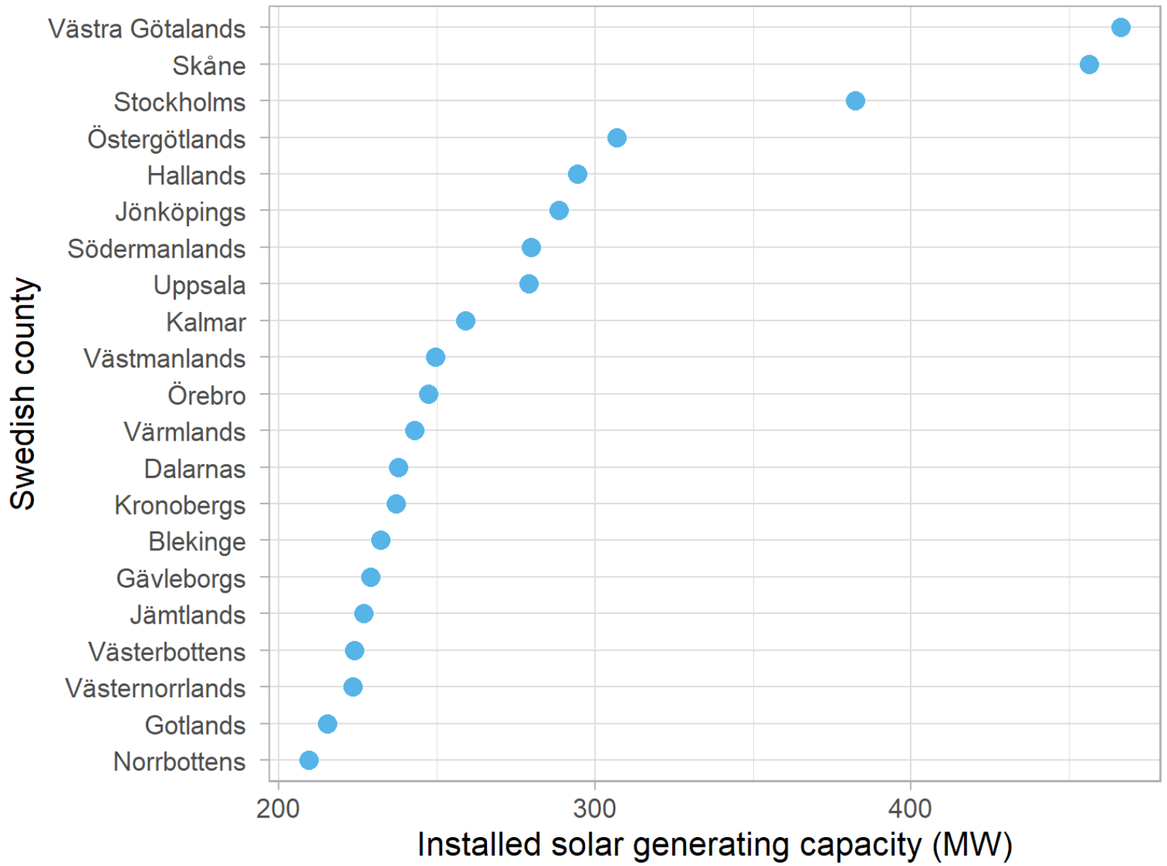


# We can use dots instead of bars

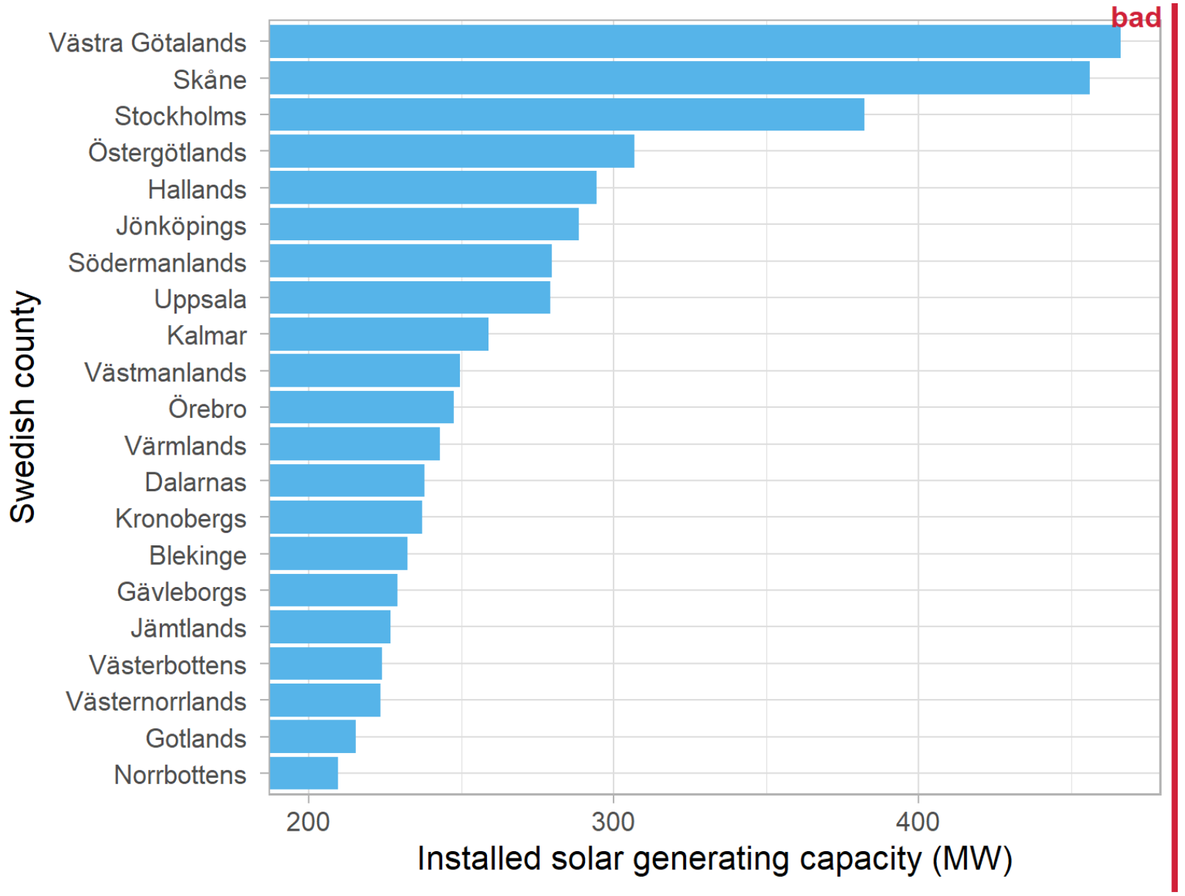


# Dots are preferable if we want to truncate the axes

Dataset: Solar panels in Sweden

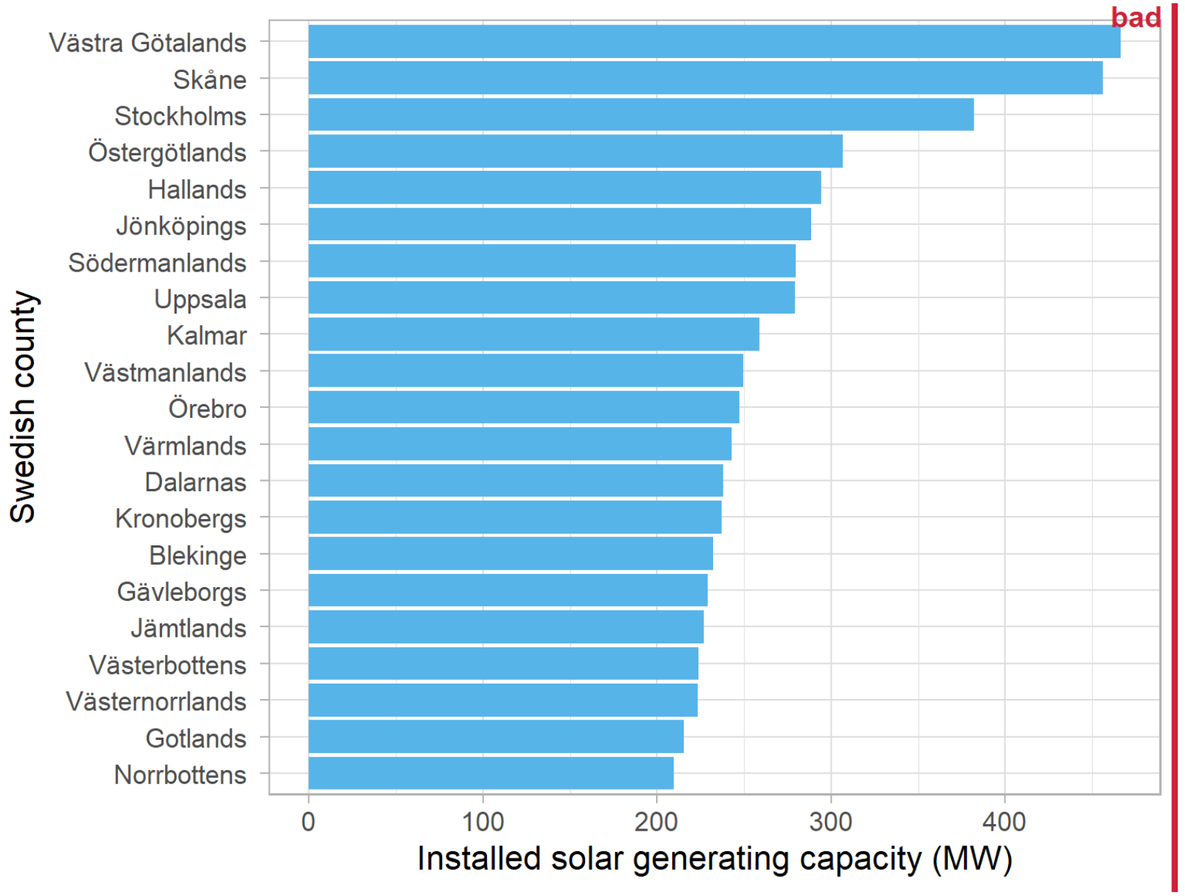


# Dots are preferable if we want to truncate the axes



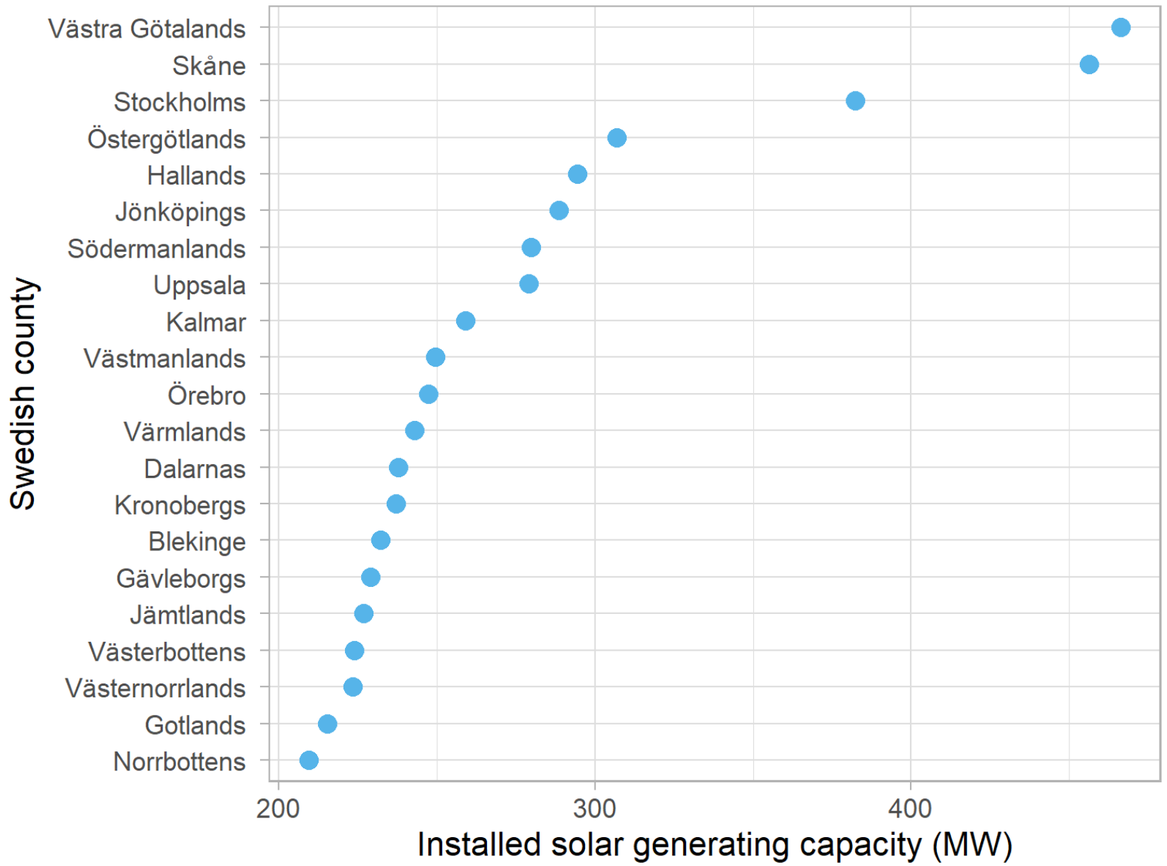

Bar lengths do not accurately represent the data values

# Dots are preferable if we want to truncate the axes




Key features of the data are obscured

# Dots are preferable if we want to truncate the axes



# Overcoming Excel

## Tables

# Overcoming Excel: Tables

- We often encounter datasets containing simple amounts 👎
- Here is some data on a sample of Swedish musical artists 🎶
- I put this data into Excel, and asked it to insert a table 💾

**Swedish musical artists**

| Rank | Artist | Monthly listeners (m) |
|---|---|---|
| 1 | Avicii | 29.47 |
| 2 | ABBA | 23.48 |
| 3 | José González | 4.07 |
| 4 | Robyn | 3.11 |
| 5 | Timbuktu | 0.38 |

Datasource: Spotify charts Nov 2022

# Your turn again

| rank | artist | monthly_streamers |
|---|---|---|
| 2 | ABBA | 23,5 |
| 1 | Avicii | 29,5 |
| 3 | José Gonzál | 4,07 |
| 4 | Robyn | 3,11 |
| 5 | Timbuktu | 0,383 |

Discuss with your neighbour:

- What do we like?
- What is confusing?

02:30

## Key rules for table layout

| Number | Rule |
|---|---|
| 1 | Do not use vertical lines. |
| 2 | Do not use heavy horizontal lines between data rows. (Horizontal lines as separator between the title row and the first data row or as frame for the entire table are fine.) |
| 3 | Text columns should be left aligned. |
| 4 | Number columns should be right aligned and should use the same number of decimal digits throughout. |
| 5 | Columns containing single characters are centred. |
| 6 | The header fields are aligned with their data, i.e., the heading for a text column will be left aligned and the heading for a number column will be right aligned. |

Source: Claus Wilke's *Fundamentals of Data Visualization*

# Let’s apply these rules

| Key rules for table layout | |
|---|---|
| Number | Rule |
| 1 | Do not use vertical lines. |
| 2 | Do not use heavy horizontal lines between data rows. (Horizontal lines as separator between the title row and the first data row or as frame for the entire table are fine.) |
| 3 | Text columns should be left aligned. |
| 4 | Number columns should be right aligned and should use the same number of decimal digits throughout. |
| 5 | Columns containing single characters are centred. |
| 6 | The header fields are aligned with their data, i.e., the heading for a text column will be left aligned and the heading for a number column will be right aligned. |
| Source: Claus Wilke's *Fundamentals of Data Visualization* | |

| rank | artist | monthly_streamers |
|---|---|---|
| 2 | ABBA | 23,5 |
| 1 | Avicii | 29,5 |
| 3 | José González | 4,07 |
| 4 | Robyn | 3,11 |
| 5 | Timbuktu | 0,383 |

Table A

| rank | artist | monthly_streamers |
|---|---|---|
| 2 | ABBA | 23,5 |
| 1 | Avicii | 29,5 |
| 3 | José González | 4,07 |
| 4 | Robyn | 3,11 |
| 5 | Timbuktu | 0,383 |

Table B

01:30

# Let’s apply these rules

| Key rules for table layout | |
|---|---|
| Number | Rule |
| 1 | Do not use vertical lines. |
| 2 | Do not use heavy horizontal lines between data rows. (Horizontal lines as separator between the title row and the first data row or as frame for the entire table are fine.) |
| 3 | Text columns should be left aligned. |
| 4 | Number columns should be right aligned and should use the same number of decimal digits throughout. |
| 5 | Columns containing single characters are centred. |
| 6 | The header fields are aligned with their data, i.e., the heading for a text column will be left aligned and the heading for a number column will be right aligned. |
| Source: Claus Wilke's *Fundamentals of Data Visualization* | |

| Rank | Artist | Monthly streamers (m) |
|---|---|---|
| 2 | ABBA | 23,5 |
| 1 | Avicii | 29,5 |
| 3 | José González | 4,1 |
| 4 | Robyn | 3,1 |
| 5 | Timbuktu | 0,4 |

Table C

| Rank | Artist | Monthly streamers (m) |
|---|---|---|
| 1 | Avicii | 29,5 |
| 2 | ABBA | 23,5 |
| 3 | José González | 4,1 |
| 4 | Robyn | 3,1 |
| 5 | Timbuktu | 0,4 |

Table D

01:30

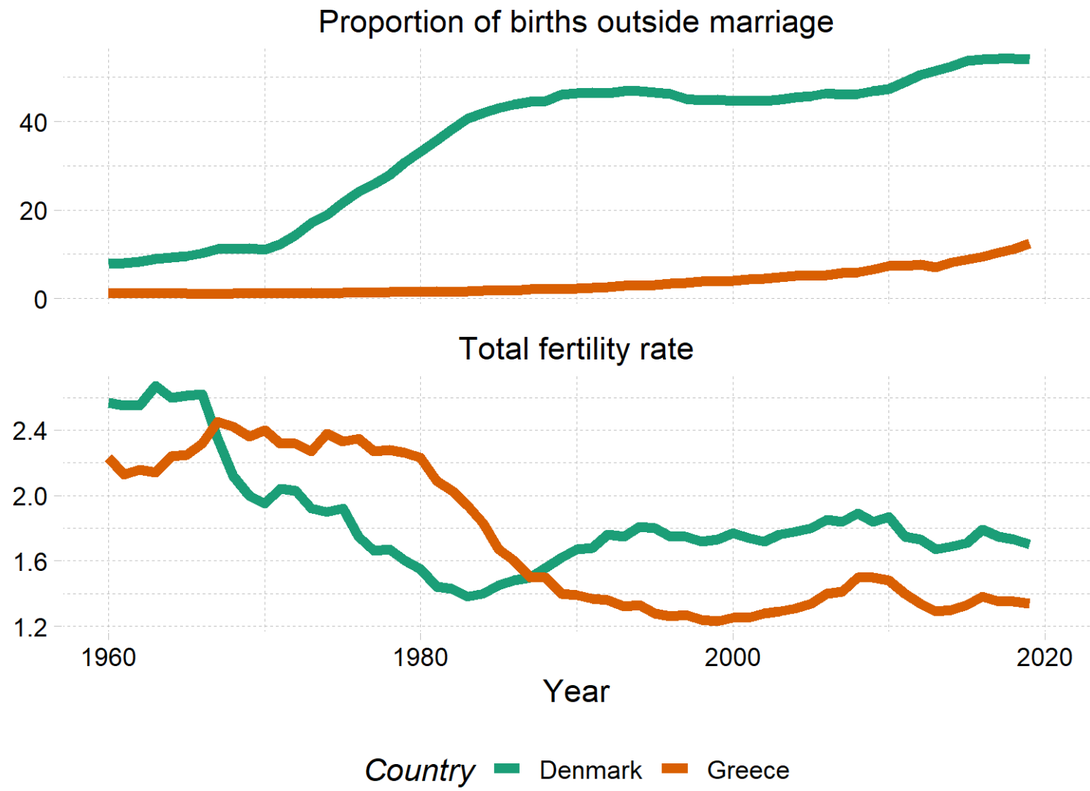
# Storytelling with data

## Related time series

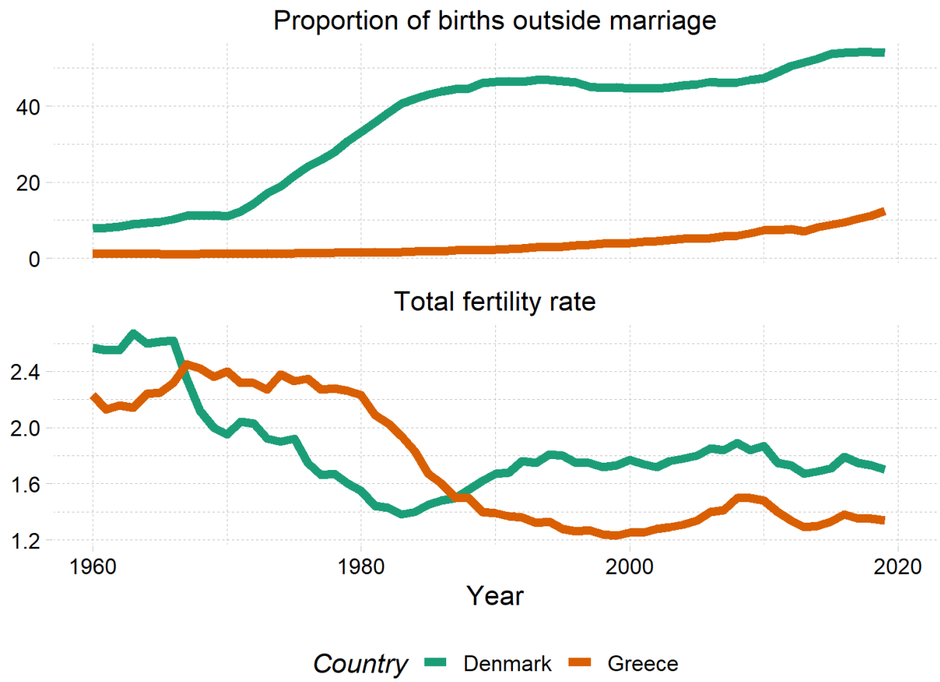
# Plotting related time series

Dataset: Fertility and births outside of marriage in Denmark and Greece.

Default choice for plotting is two line plots



# Plotting related time series



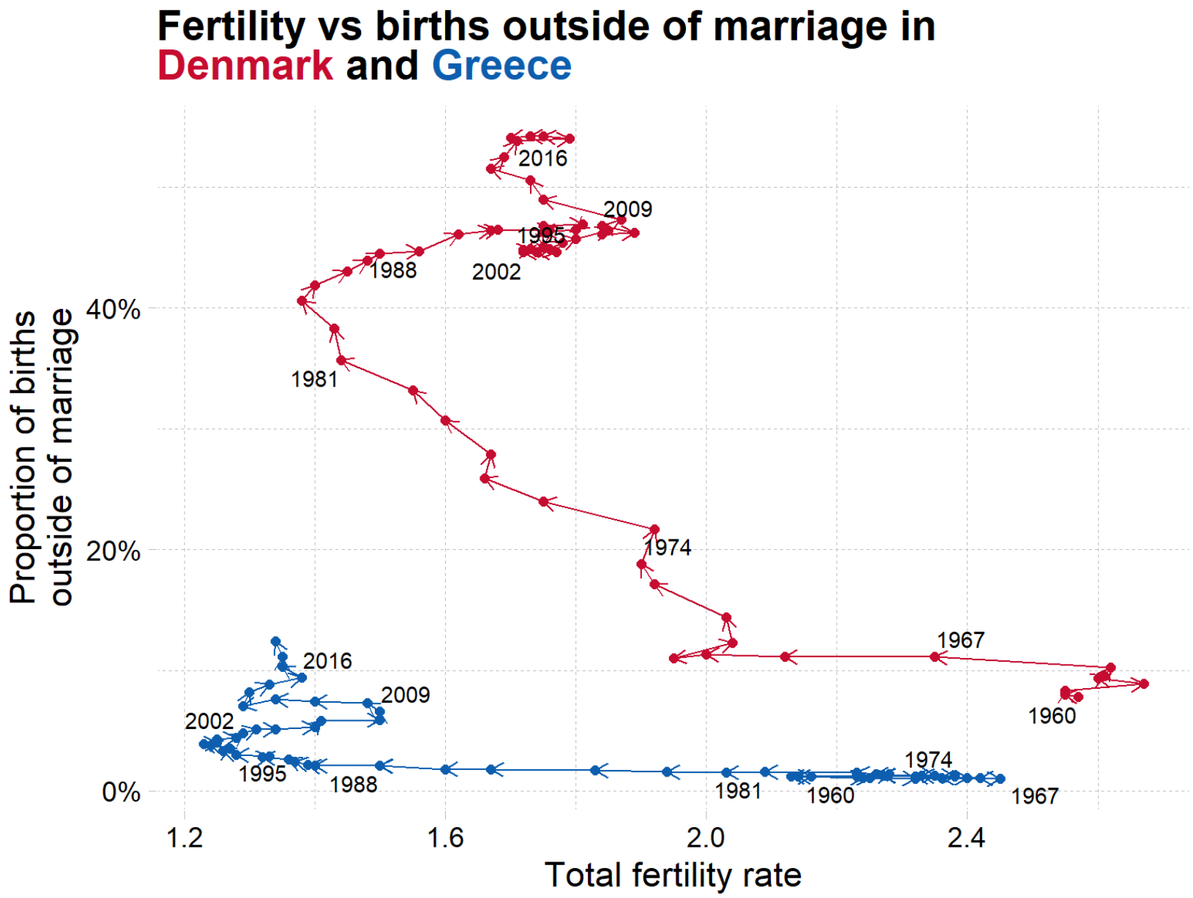

## Pros 👍

- Familiar

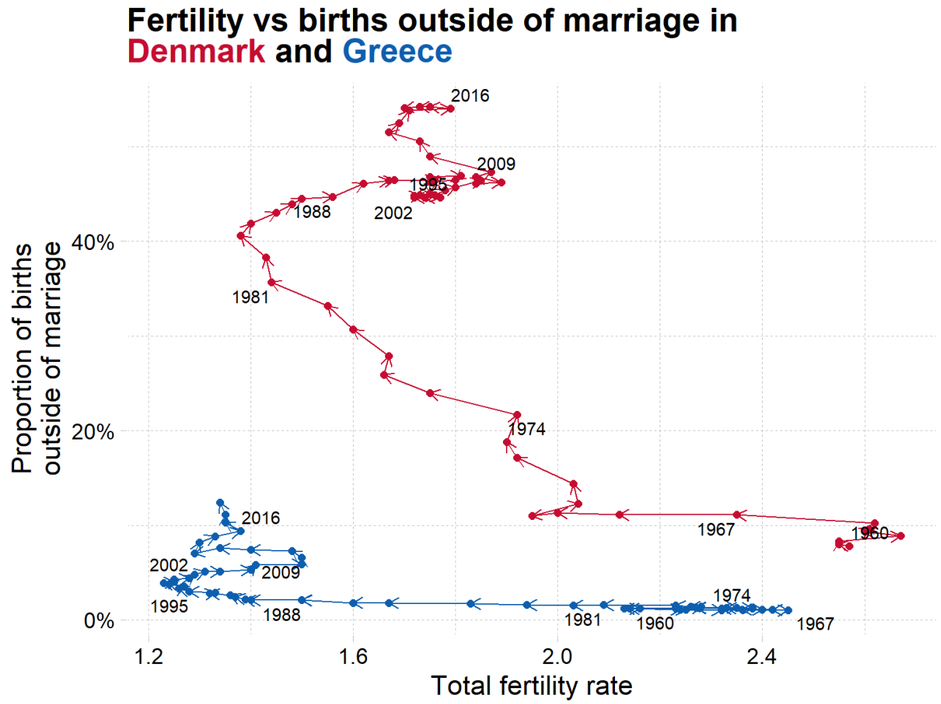
## Cons 👎

- Hard to keep track of each series
- Difficult to compare movements across short periods

# An alternative: time on a third axis

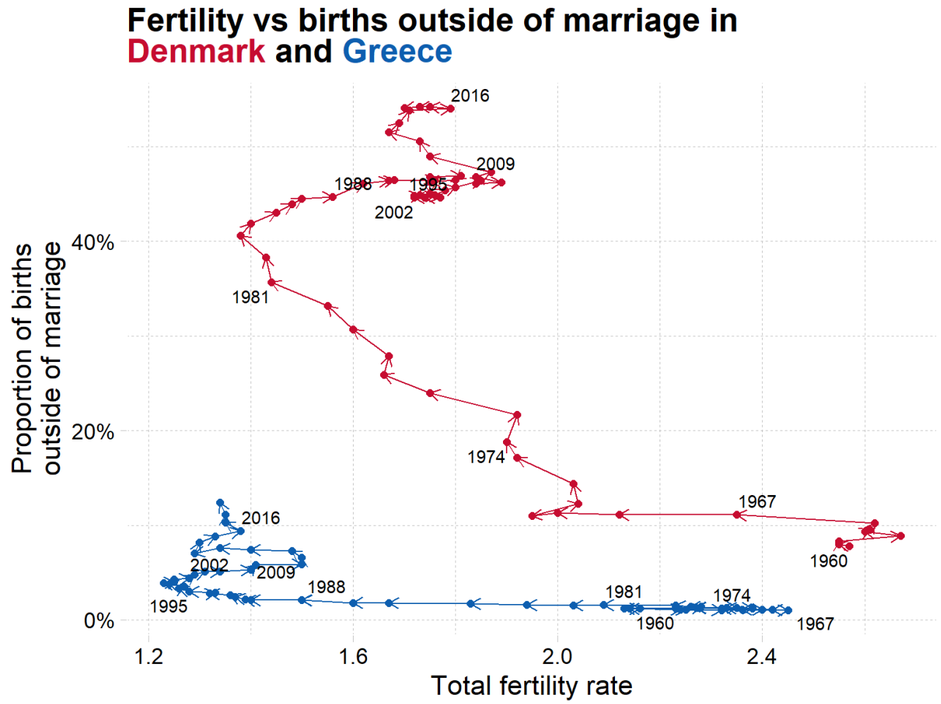


# What have we learned?




- Both countries saw a large drop in fertility from the 1960s until the 1980s
- In Denmark, after 1970 we see an increase in the share of children born outside of marriage
- In contrast, Greek families have relatively few children outside of marriage.
- After 1990, Danish fertility increased from 1.3 to 1.8, while Greek fertility remained at ‘lowest-low’ levels, below replacement.

# What have we changed?




- Indicators on the x- and y-axis and then show time with text labels
- Legend is replaced with colour coded title
- Colours have meaning (main colour of country flag)
- Percentage labels on the y-axis

# Storytelling with data

## Giving context

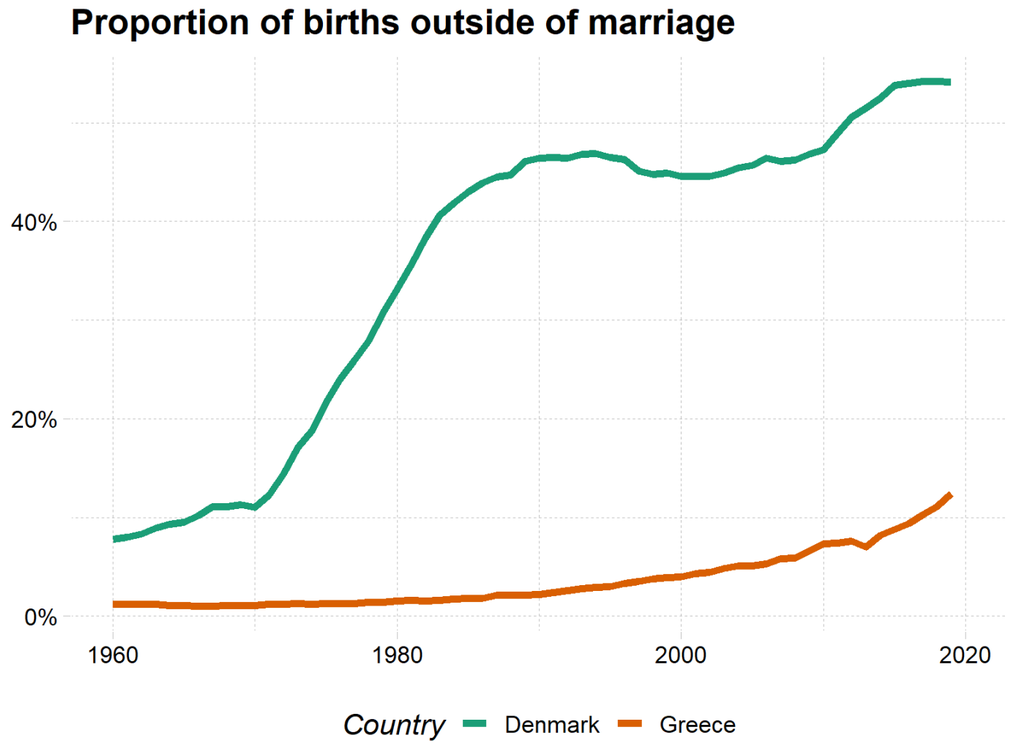
# Giving context

Sometimes we may want to show a particular series of data in its correct context.

For instance, in our line graph above which showed the evolution of the share of births outside of marriage in **Denmark and Greece**, we might want to know if these two represent the **extremes** within Europe.

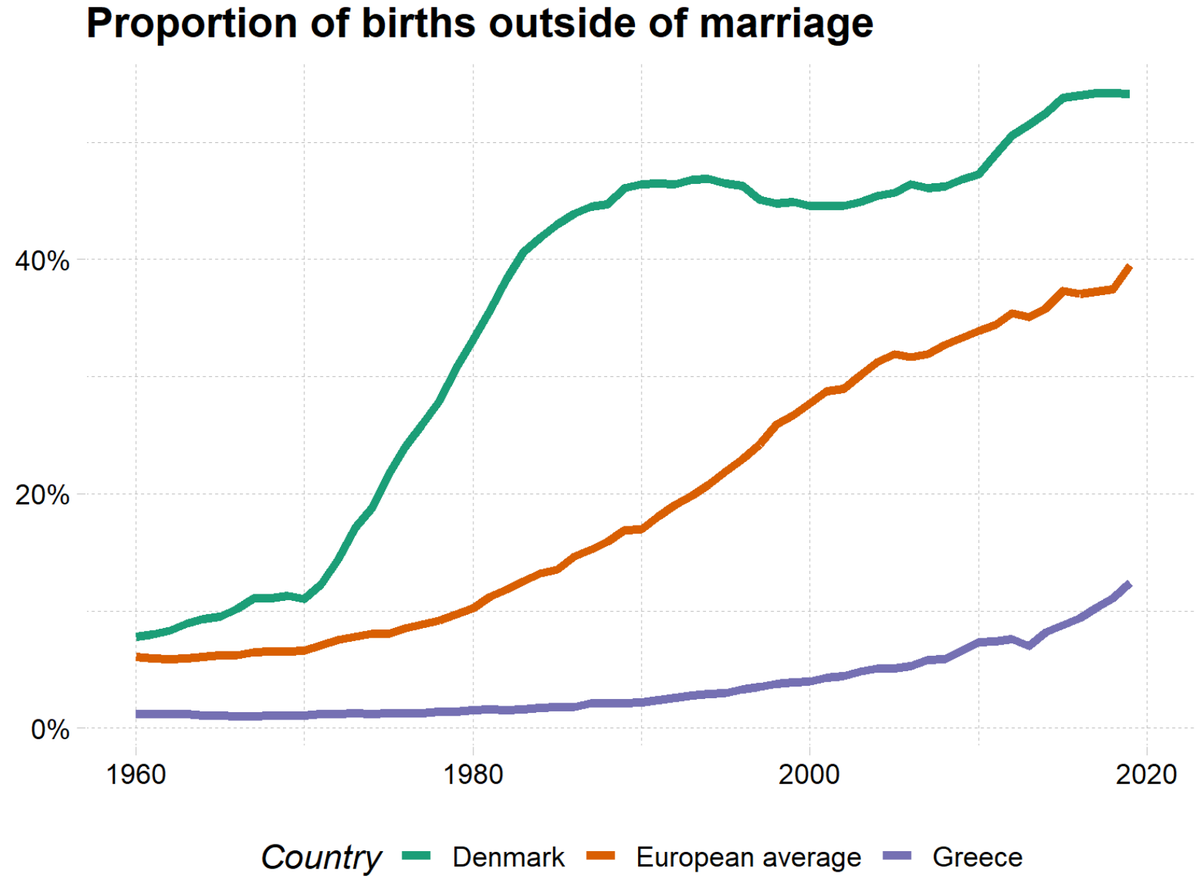
# Giving context

Do **Denmark and Greece** represent the **extremes** of the share of children born outside of marriage in Europe?

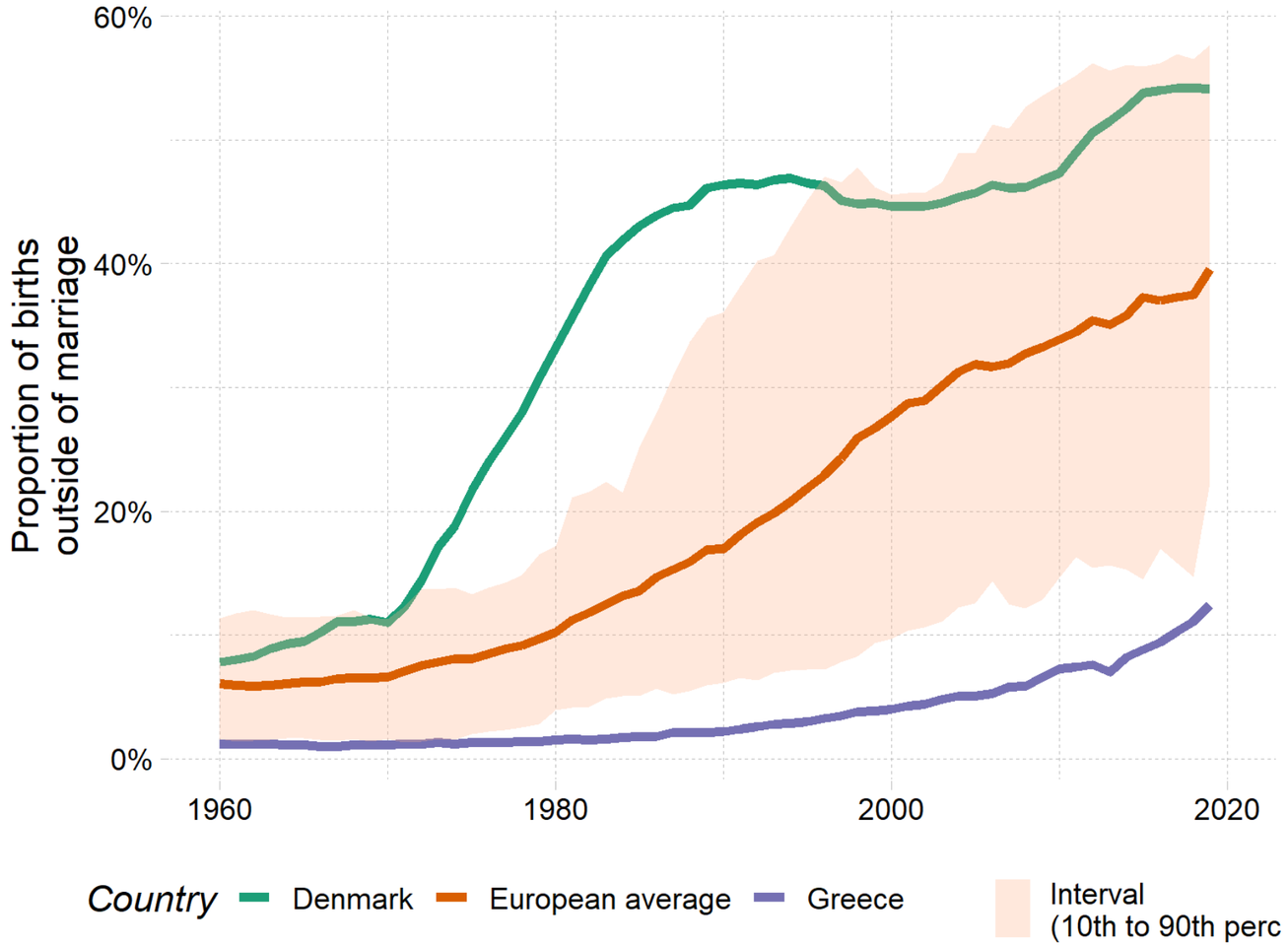


# Giving context with an average

One way to do this would be to show an average for Europe

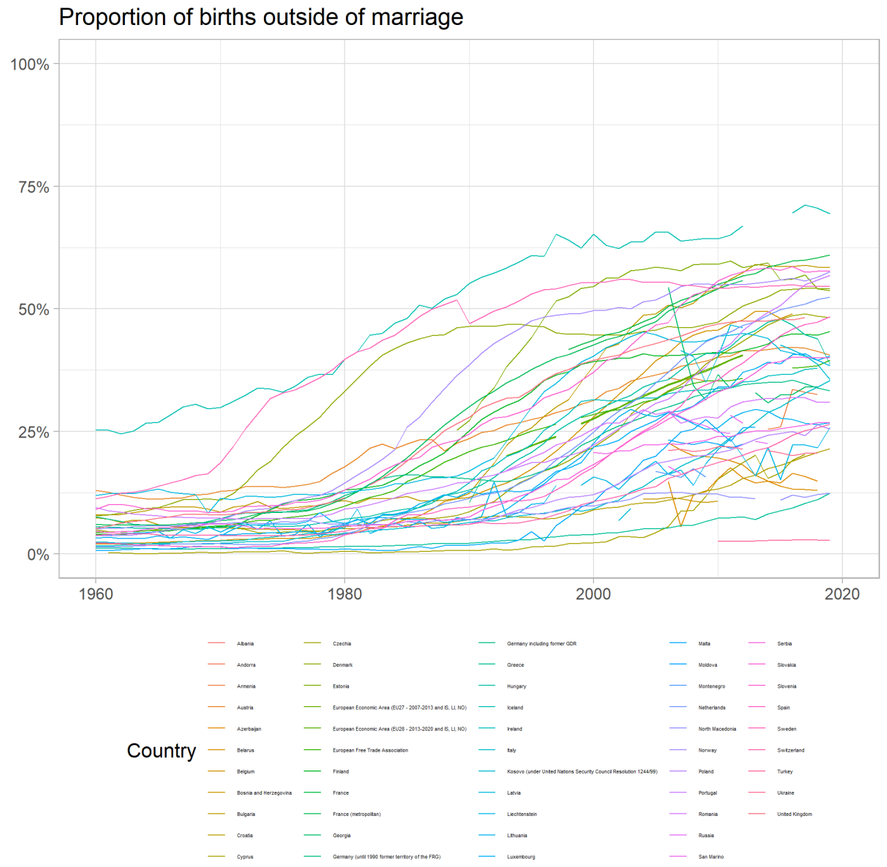


# Giving context with an interval ribbon



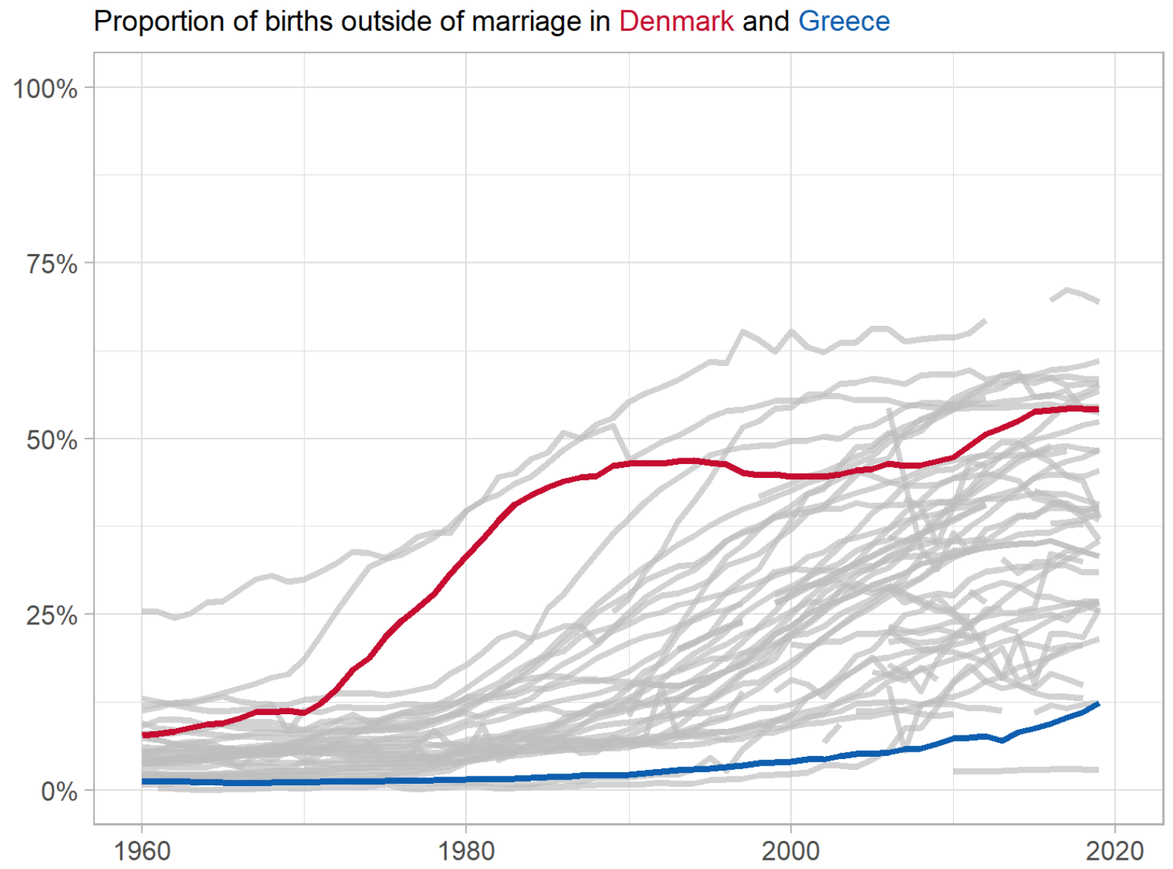
# Giving context with all of the data

This is silly

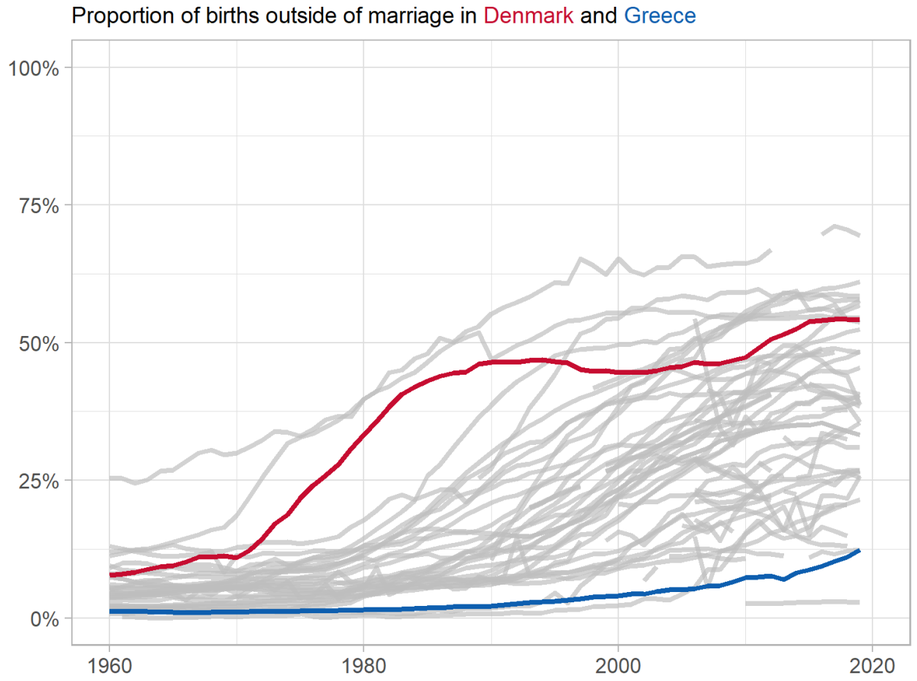


# Giving context with all of the data

Here we highlight the series we are interested in and draw in the remaining series in grey



# What have we changed?



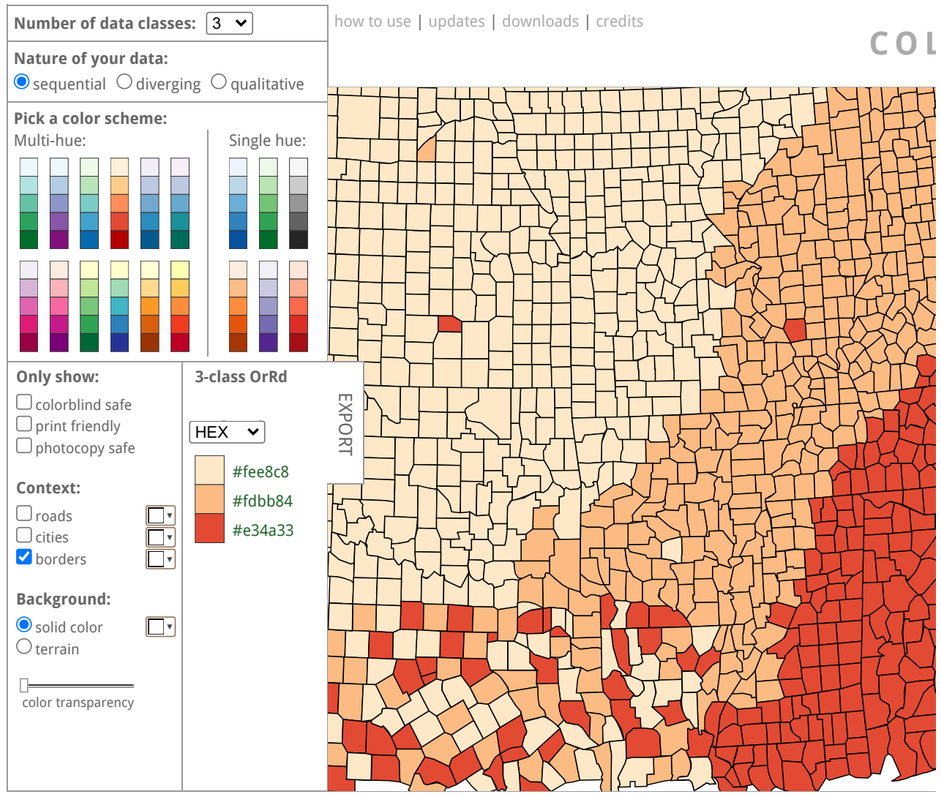

- Shows each of the series
- We can see that Denmark is a leader in the beginning, but is caught up by other nations
- Does not hide outliers
- Makes clear the trends in your countries of interest

# Storytelling with data

## Tips for polished figures

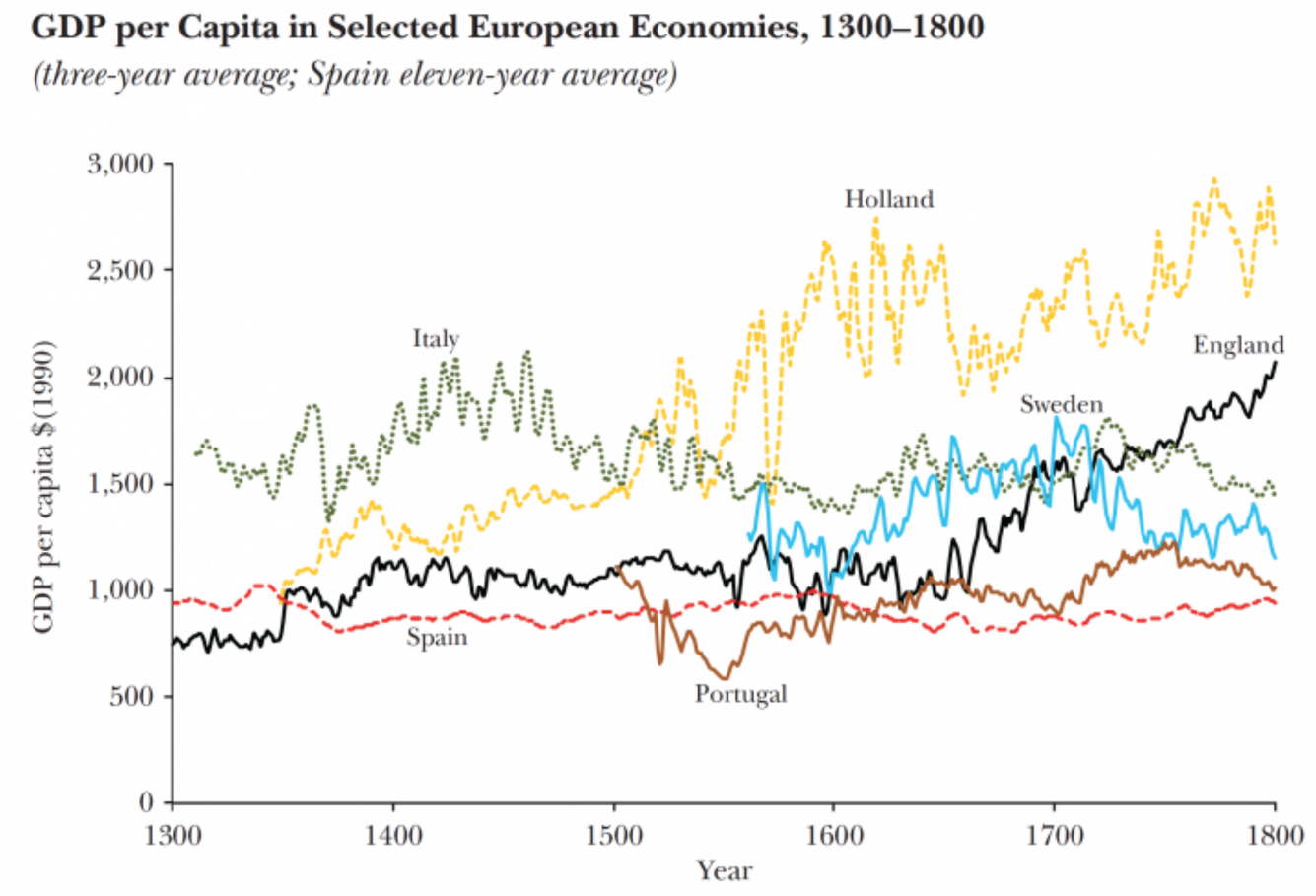
# Tips for polishing your figures

## Where to get great colours from for your plots:



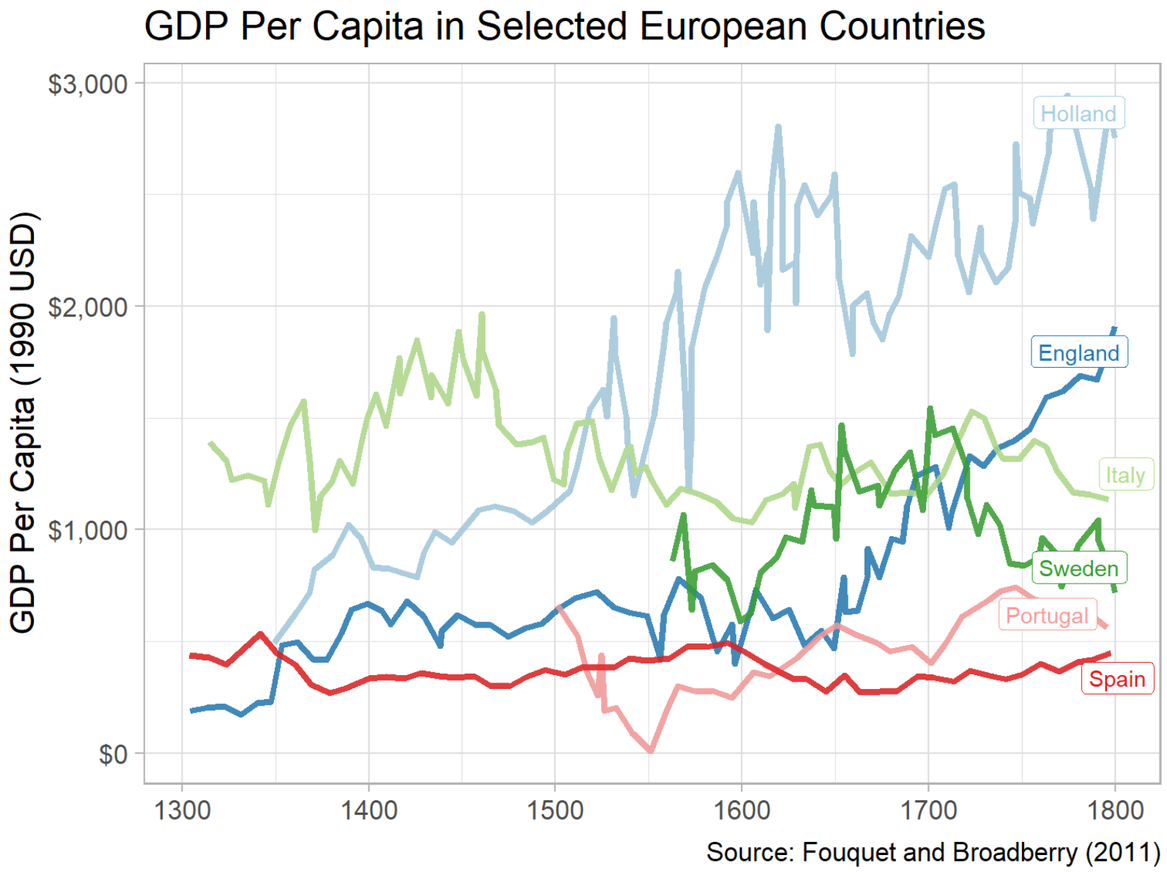

```
help spmap # Look for the palettes under fcolor
```

# Recreating published figures

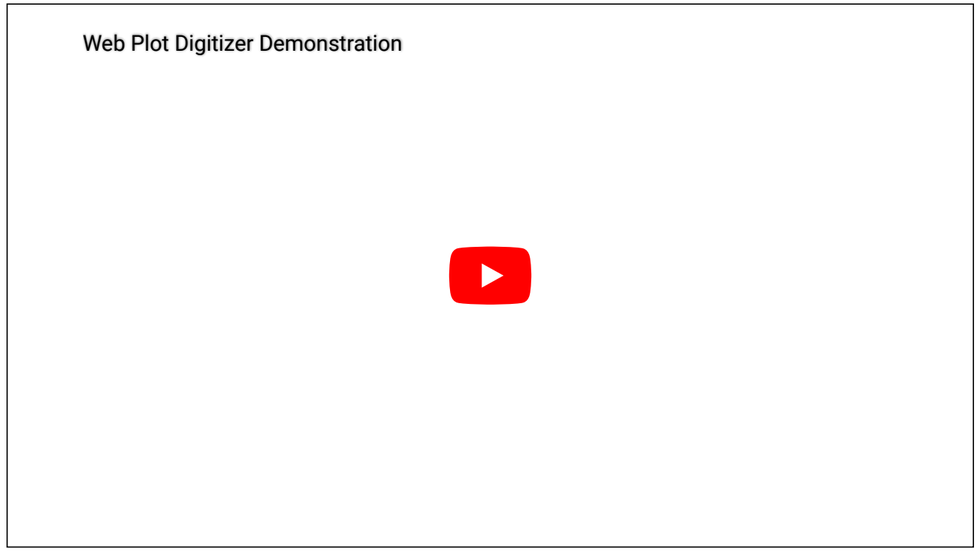



A FT chart published without the underlying data

# Recreating published figures



# Recreating published figures



# You pay a heavy price

# Thank you