

# MY472 – Week 8: Data Visualisation

Friedrich Geiecke

MY472: Data for Data Scientists

16 November 2020

# Course

1. Introduction to data
2. The shape of data
3. Introduction to scraping
4. Advanced scraping
5. Working with APIs
6. (Reading week)
7. Textual data
8. Data visualisation
9. Creating and managing databases
10. Interacting with online databases
11. Cloud computing

# Outline

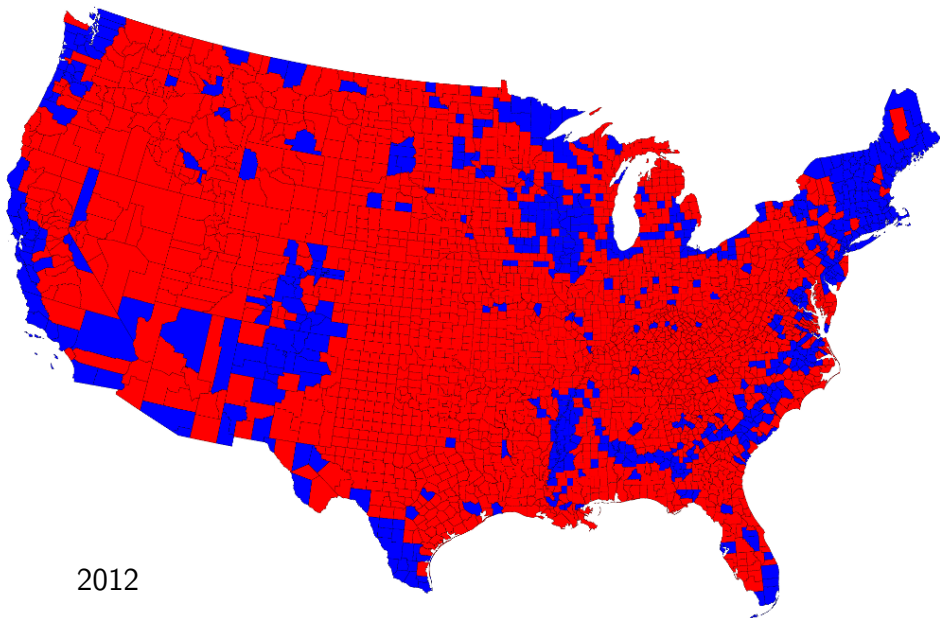
1. Introduction
2. Some principles of data visualisation
3. ggplot2
4. Coding session

# Outline

1. Introduction
2. Some principles of data visualisation
3. ggplot2
4. Coding session

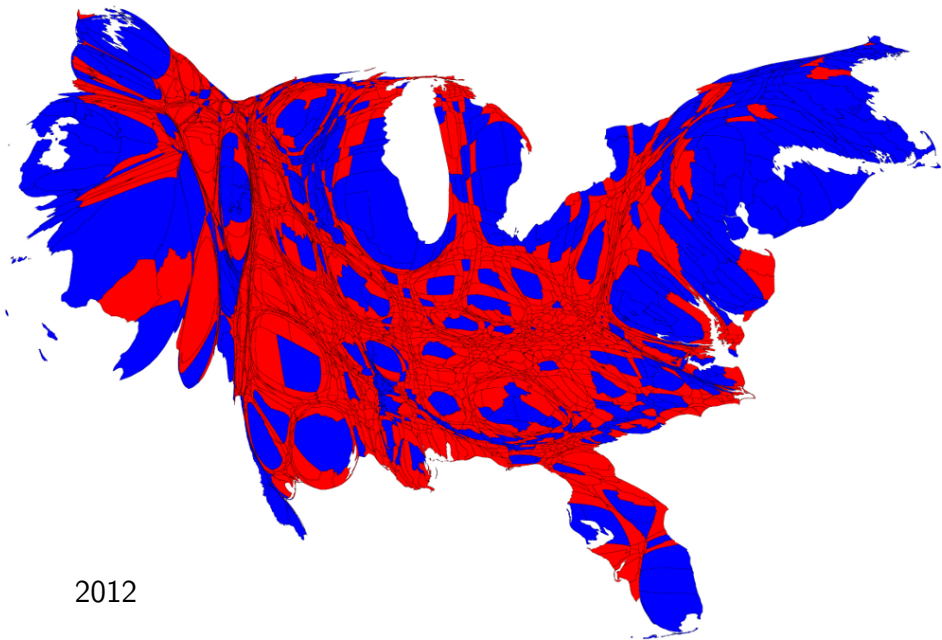
## Why visualisation can be helpful: Anscombe examples

01-anscombe.Rmd



2012

**Source:** Mark Newman (Michigan)

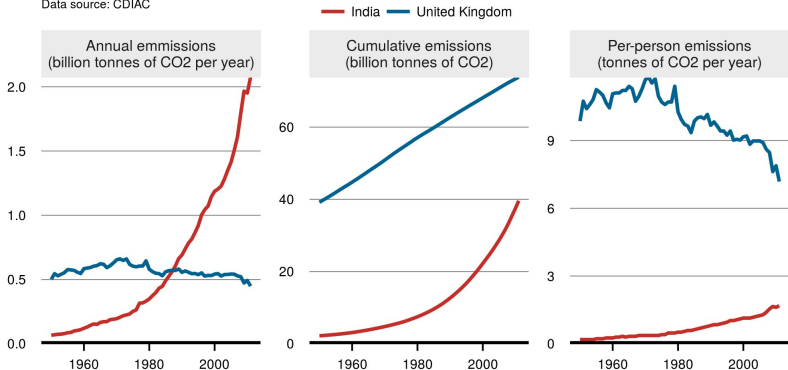


2012

**Source:** Mark Newman (Michigan)

### Three ways to compare the carbon emissions of India and United Kingdom

Data source: CDIAC



Note: figures cover energy and cement related activities  
Figure by [robert.wilson@strath.ac.uk](mailto:robert.wilson@strath.ac.uk)

Source: New York Times



# Outline

1. Introduction
2. Some principles of data visualisation
3. ggplot2
4. Coding session

# Principles by Edward Tufte

- ▶ Show the data
- ▶ Avoid distorting what the data have to say
- ▶ Allow viewer to compare
- ▶ Serve a clear purpose: description, exploration, tabulation or decoration
- ▶ Be closely integrated with the statistical and verbal descriptions of the dataset
- ▶ Graphics can reveal data (e.g. Anscombe Quartet)

# General guidelines

- ▶ Maximize data-to-ink ratio
- ▶ Avoid misleading decisions
  - ▶ Y axis starts at 0
  - ▶ Comparison of areas is hard
  - ▶ Use comparable units
  - ▶ Erase chart junk
- ▶ Use text to inform and contextualise. Add annotations
- ▶ Appropriate use of scales (x/y axes, color, size, shape...)
- ▶ Use small multiples to facilitate comparisons
- ▶ Always cite your sources

# Outline

1. Introduction
2. Some principles of data visualisation
3. [ggplot2](#)
4. Coding session

# What is the grammar of graphics?

## **The grammar of graphics.**

*A statistical graph is a mapping from data to aesthetic attributes (color, shape, size) of geometric objects (points, lines, bars). The plot may also contain statistical transformations of the data and is drawn on a specific coordinate system. Faceting can be used to generate the same plot for different subsets of the data. It is the combination of these independent components that make up a graphic.*

**Hadley Wickham**, *ggplot2*, page 3

# Data visualisation with ggplot2

## Why **ggplot2**?

- ▶ Based on “Grammar of Graphics” (Wilkinson, 1999)
  - consistent, modular, and very flexible
- ▶ Sensible defaults for quick exploratory plots
- ▶ But also easy to customize, extend
- ▶ Excellent online resources

# Grammar



Source: Thomas Lin Pedersen [[link](#)]

# Grammar

- data** Data to visualise, for ggplot2 in a 'tidy' format
- (aesthetic) mapping** Mapping variables in the data to components of the graphic such as axes
- stats** Statistical transformations of the data, e.g. binning or averaging
- scales** Translation/mapping of e.g. categorical variables such as political party to shapes or colours
- geom** Geometric objects that are drawn to represent the data: bars, lines, points, etc.
- facets** Breaking up the data into subsets, to be displayed independently on a grid
- coordinates** Coordinate system; provides axes and gridlines to make it possible to read the graph
- theme** Parts that do not follow from the data: Background colours, fonts, etc.



## Online resources

- ▶ Main documentation page: <https://ggplot2.tidyverse.org/>
- ▶ Book by Hadley Wickham, Danielle Navarro, and Thomas Lin Pedersen: <https://ggplot2-book.org/>
- ▶ R Graph gallery for ggplot2  
<https://www.r-graph-gallery.com/ggplot2-package.html>
- ▶ Two recent video workshops by Thomas Lin Pedersen, [video 1](#), [video 2](#), and the repo with associated [exercises](#)
- ▶ StackOverflow, tag: ggplot2  
<https://stackoverflow.com/questions/tagged/ggplot2>

# Outline

1. Introduction
2. Some principles of data visualisation
3. ggplot2
4. Coding session

## Coding session

02-ggplot2-basics.Rmd

03-scales-axes-legends.Rmd