

Data Visualization

A Practical Introduction

Nith Kosal

Future Forum

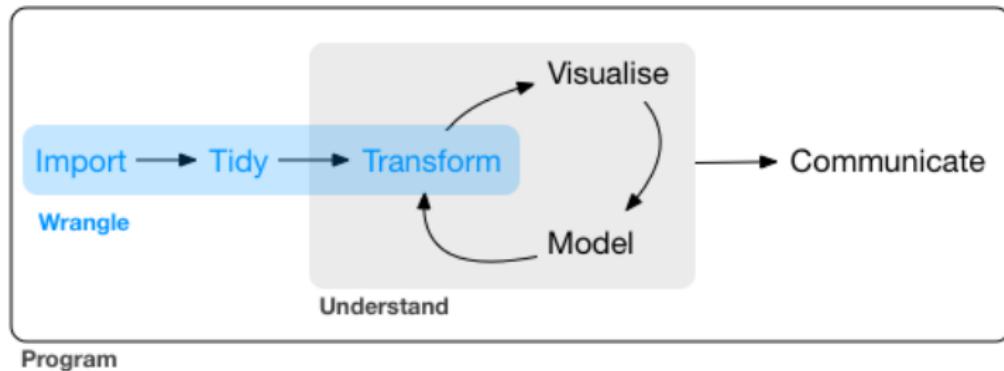
2021 YRP Seminar, May 13, 2021

Goal of Today's Seminar

- ▶ Know the types of graphs we should avoid
- ▶ Learn about RStudio IDE (integrated development environment)
- ▶ Data visualization in R
- ▶ Data import and transformation
- ▶ Data export
- ▶ Discussion about your questions or any concerns about R or datasets

Remember: It will not work in a flow and structure, because we practice in the real example as a researcher have a dataset and applied it for a specific research study.

The Thing to know when you work with data



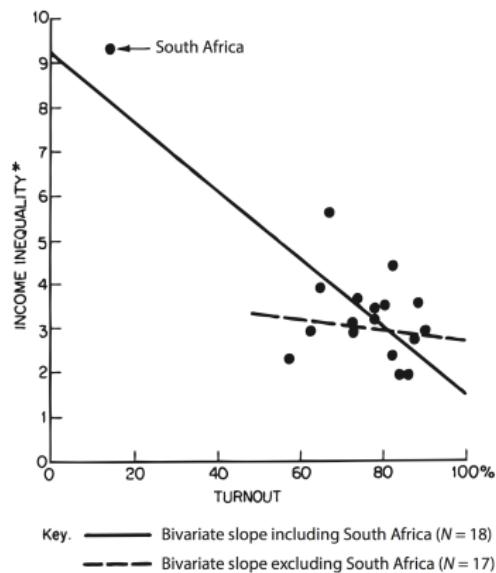
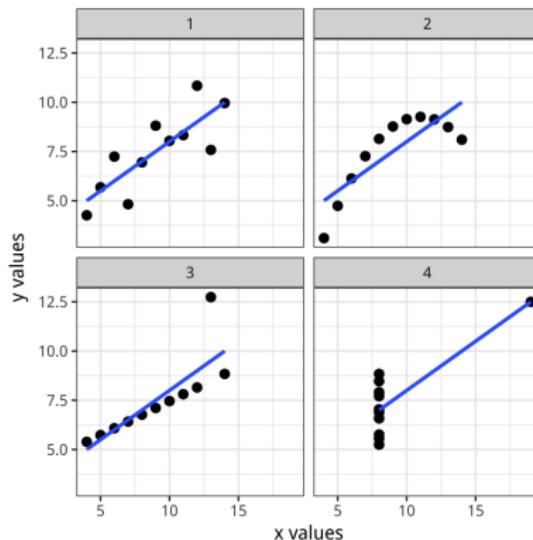
Source: Grolemund and Wickham (2016)

- ▶ Tidying your data means storing it in a consistent form that matches the semantics of the dataset with the way it is stored.
- ▶ Together, tidying and transforming, we called *wrangling*

What does data visualization mean?

- ▶ A reporting tool: the representation of information in the form of a chart, diagram, picture, etc.
- ▶ It is a part of *hypothesis generation or data exploration*.
 - A precise mathematical model to generate falsifiable predictions.
 - Only use an observation once to confirm a hypothesis.
- ▶ Modelling is an important part of the exploratory process.

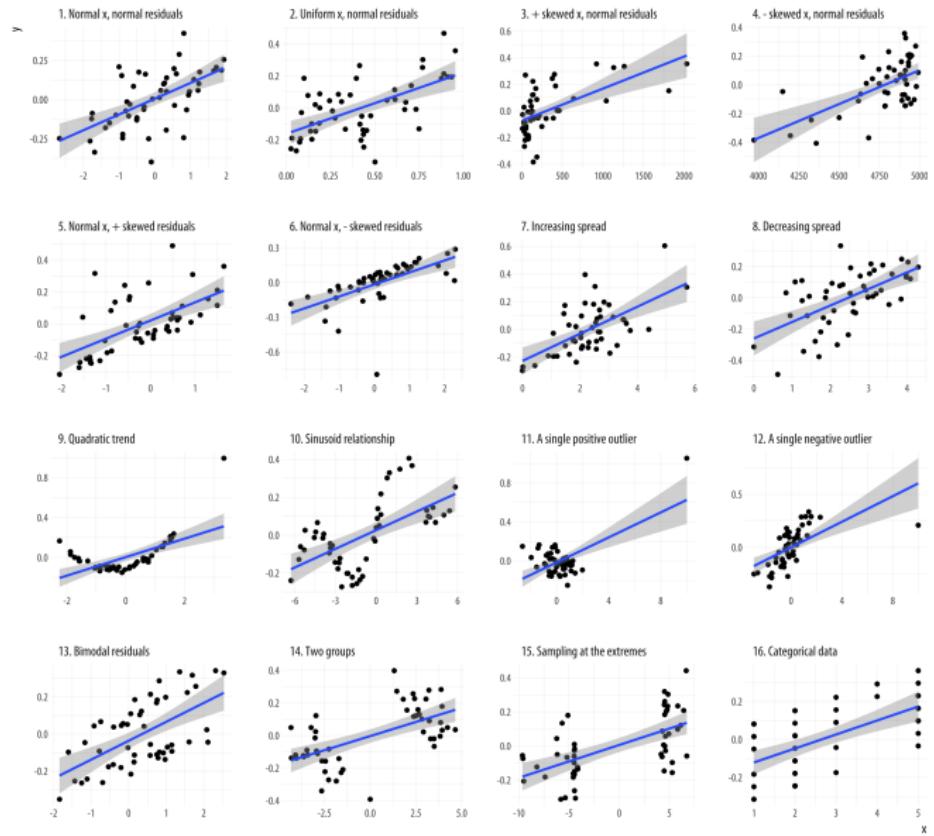
Now, We Look at Some Data Visualizations



Source: Chatterjee and Firat (2007), Jackman (1980)

- ▶ Scatterplots are the workhorse of data visualization in social science.
- ▶ A scatterplot shows the relationship between two quantities, such as height and weight, age and income, or time and unemployment.

Within each panel, the correlation between the x and y variables is set to be 0.6, a pretty good degree of association. But the actual distribution of points is created by a different process in each case.

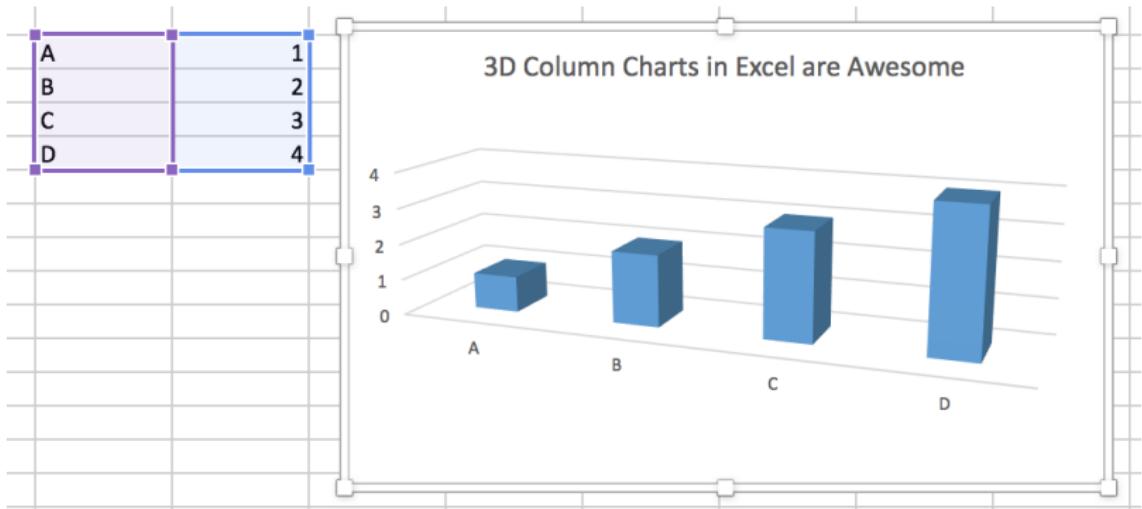


Source: Jan Vanhove (2016)



Source: Kieran Healy (2018)

The bars are hard to read and compare.



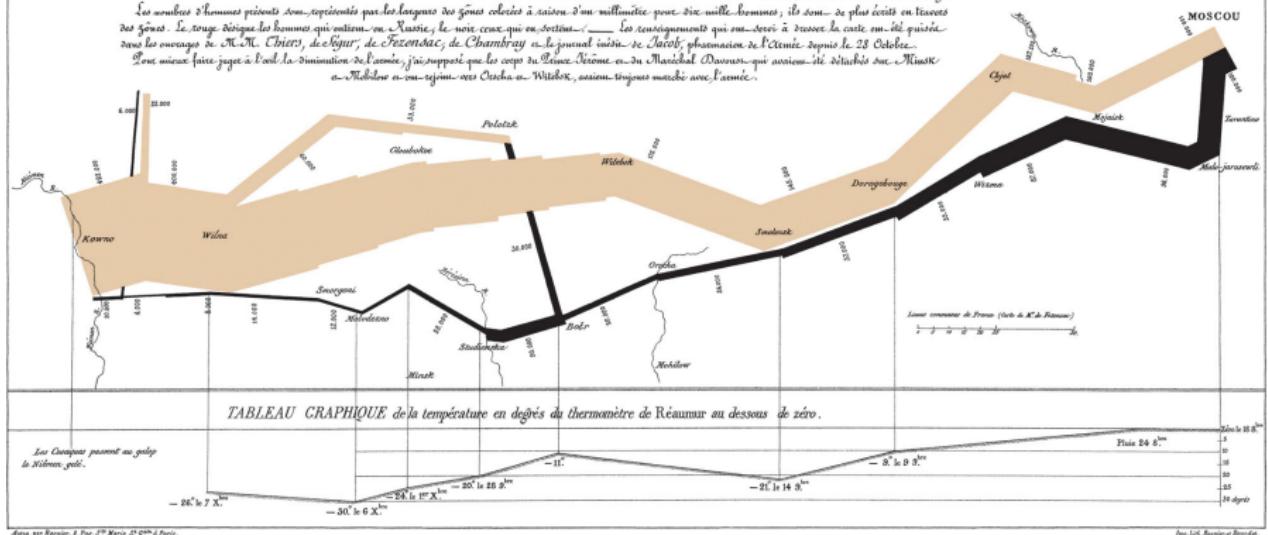
- ▶ Scatterplots are the workhorse of data visualization in social science.
- ▶ A scatterplot shows the relationship between two quantities, such as height and weight, age and income, or time and unemployment.

Carte Figurative des pertes successives en hommes de l'Armée Française dans la campagne de Russie 1812-1813.

Ouvrue par M. Minard, Ingénieur Général des Ponts et Chaussées et établie à Paris, le 20 Novembre 1869.

Les nombreuses tâches présentées sont... représentées par les longues des gîtes colonisés à saison d'un millionne pour dix mille hommes; ils sont de plus courts en lieux des gîtes. Le temps détruit les hommes qui entrent en Russie, le vent crée qui va sortir... Les renseignements qui me servent à dresser la carte ont été quitté dans les ouvrages de M. Chier, de Chézy, de Sébastien, de Chambry et de Sotov, pharmacien de l'Armée depuis le 28 Octobre.

Long mieux faire que à l'ouest, la diminution de l'armée j'ai rapporté que les corps au Peine, Neman et du Mecklembourg-Pomeranie qui avaient été détruits sur le Niémen et le Névez et qui se sont avec Orléans et Wilek, assuré toujours marche avec l'armée.



Avec ses Repères, à Paris, 2^e édition, 25 octobre 1869.

Imp. L. H. Richepin et Cie.

Source: Charles Joseph Minard — Minard's visualization of Napoleon's retreat from Moscow.

Edward Tufte's comment: “Graphical excellence is the well-designed presentation of interesting data—a matter of substance, of statistics, and of design, it consists of complex ideas communicated with clarity, precision, and efficiency.”

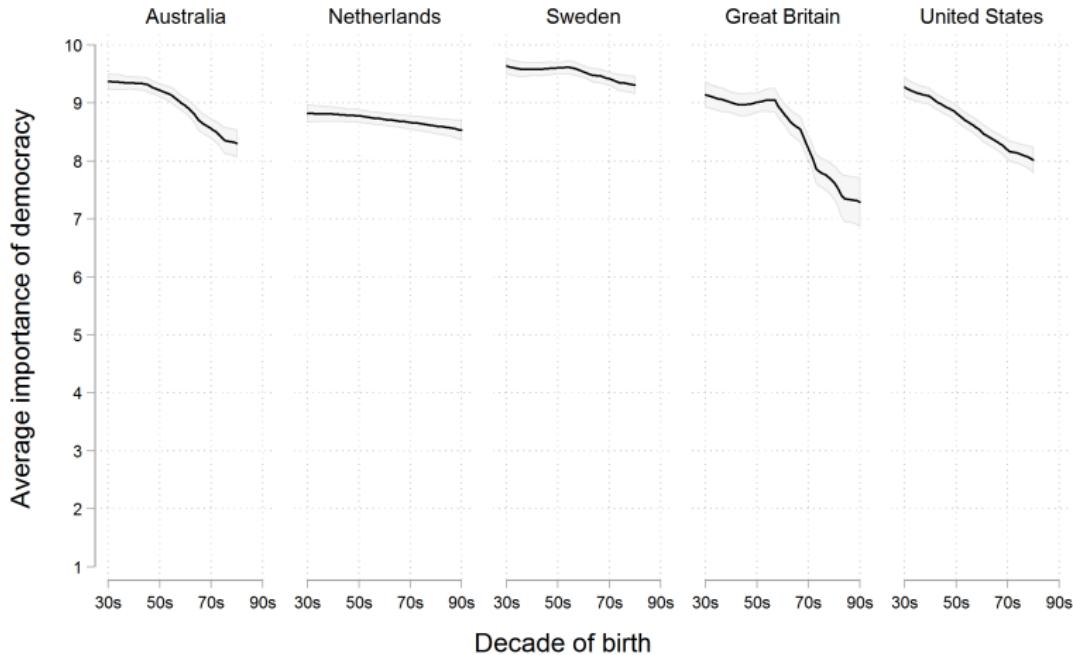
MONSTROUS COSTS

Total House and Senate campaign expenditures,
in millions



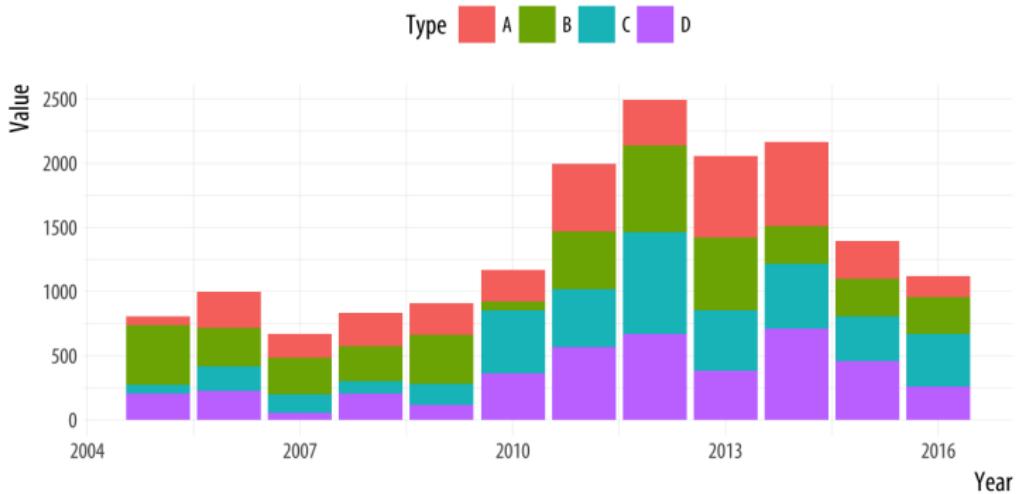
Source: 'Monstrous Costs' by Nigel Holmes

- ▶ Scatterplots are the workhorse of data visualization in social science.
- ▶ A scatterplot shows the relationship between two quantities, such as height and weight, age and income, or time and unemployment.



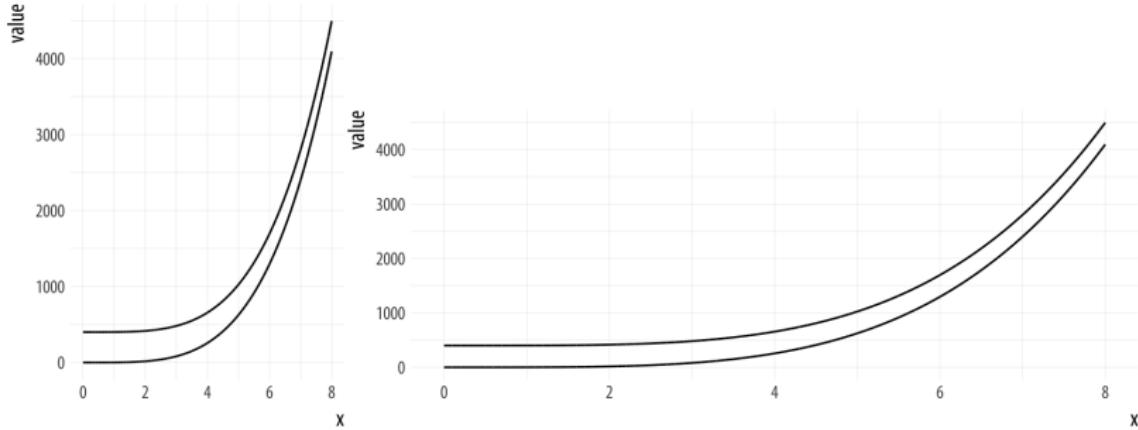
Graph by Erik Voeten, based on WVS 5

- ▶ Scatterplots are the workhorse of data visualization in social science.
- ▶ A scatterplot shows the relationship between two quantities, such as height and weight, age and income, or time and unemployment.



Source: 'Monstrous Costs' by Nigel Holmes

- ▶ Scatterplots are the workhorse of data visualization in social science.
- ▶ A scatterplot shows the relationship between two quantities, such as height and weight, age and income, or time and unemployment.



Source: 'Monstrous Costs' by Nigel Holmes

- ▶ Scatterplots are the workhorse of data visualization in social science.
- ▶ A scatterplot shows the relationship between two quantities, such as height and weight, age and income, or time and unemployment.

Variable Types in the Statistical Software

- ① int stands for integers
- ② dbl stands for doubles, or we call real numbers
- ③ chr stands for character vectors, and strings
- ④ dttm stands for date-times (a date + a time)

At the same time, there are three other common types of variables

- ① lgl stands for logical, vectors that contain only TRUE or FALSE
- ② fctr stands for factors, which R uses to represent categorical variables with fixed possible values.
- ③ date stands for dates.