DATA VISUALIZATION

for historical sociolinguistics

RIK VOSTERS
Vrije Universiteit Brussel (VUB)

HiSoN summer school – Metochi, July 2022

www.historicalsociolinguistics.be

Intrigue - James Ensor, 1890 - Royal Museum of Fine Arts Antwerp

WELCOME

Introductions

Who are you?

What do you work on?

What sort of data do you work with?

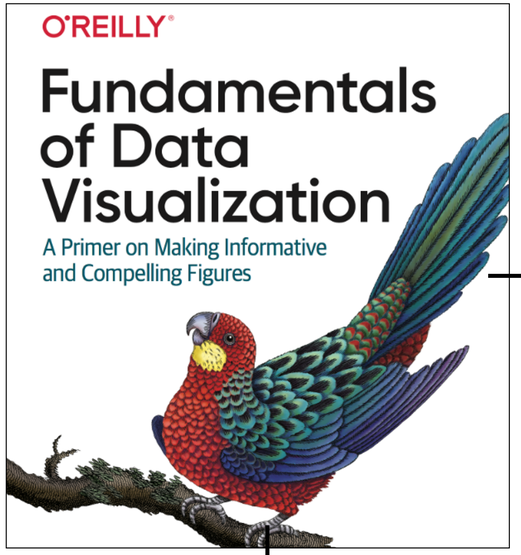
Experience with statistics, data viz, R?

Something about yourself not related to your work?

This course

- Visualizations, not statistics
- User, not expert
- No focus on software (but… R!)
- Data sets
- Covering basics (not: geospatial data, animated visualizations, ML, …)
- Hands-on, but not too hands-on
- Slides + code to reproduce examples → Dropbox
- Feel free to interrupt and ask questions

Main sources (1)



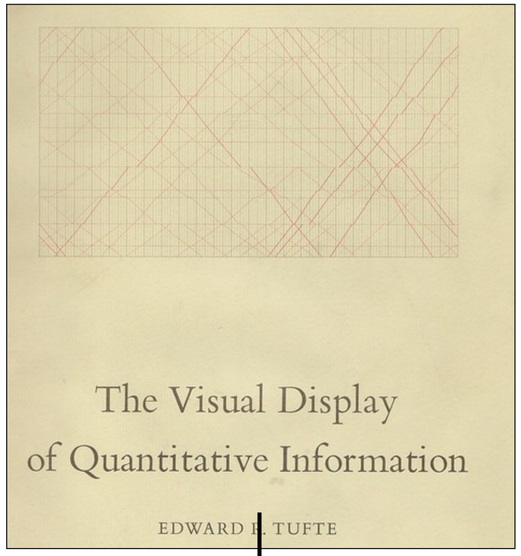
Main sourcebook for the course (and source for plots of non-linguistic examples!)

Wilke (2019)

Fundamentals of Data Visualization

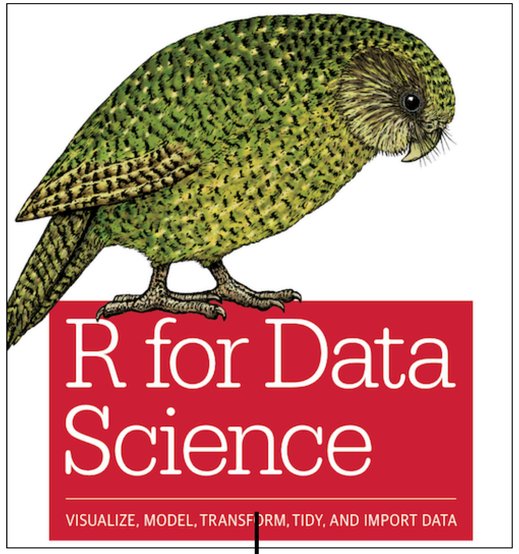
https://clauswilke.com/dataviz/introduction.html

Main sources (2)



Tufte (1983, 2001)

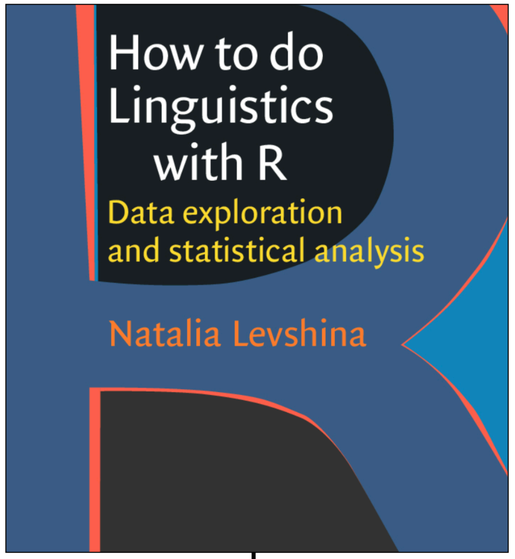
The Visual Display of Quantitative Information



Wickham & Grolemund(2016)

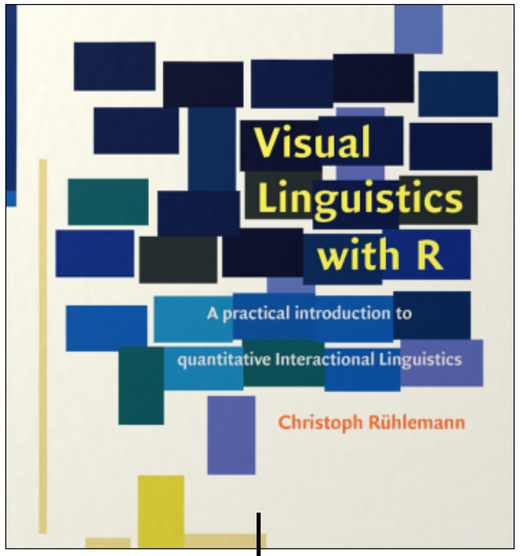
R for Data Science

https://r4ds.had.co.nz/



Levshina (2015)

How to do Linguistics with R



Rühlemann (2020)

Visual Linguistics with R

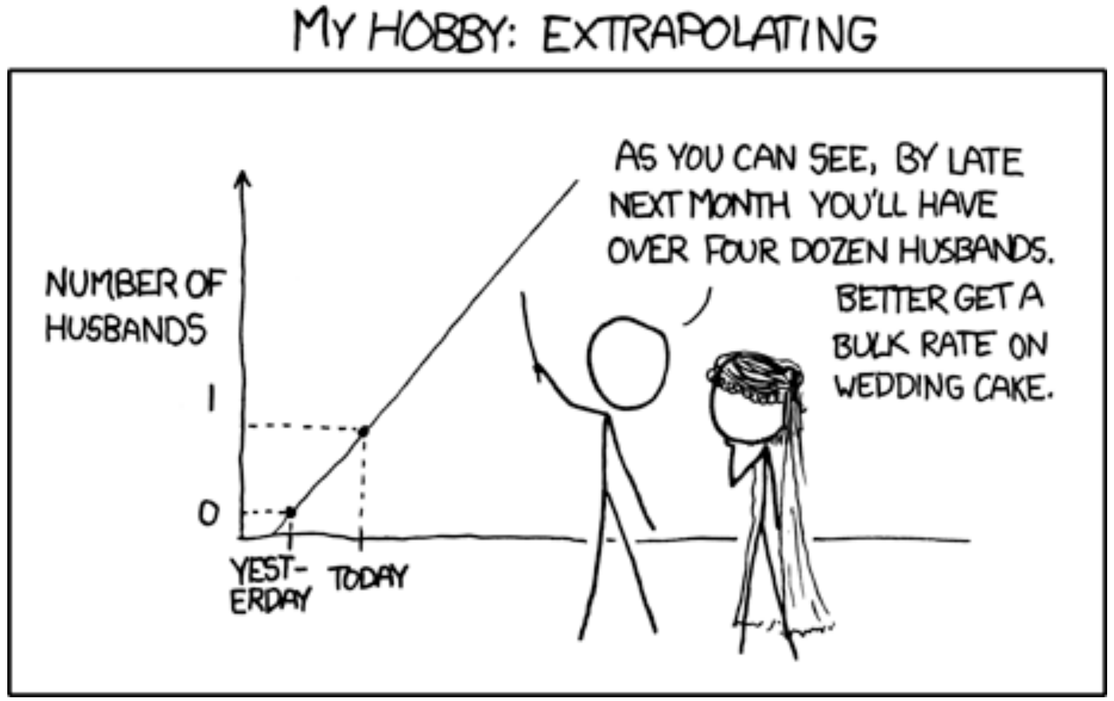
WELCOME

Overview

- Introduction
- Good, bad, and wrong visualizations
- From data to aesthetics
- A brief note on software
- Using color
- Two design principles
- Data storytelling
- Group work

This course

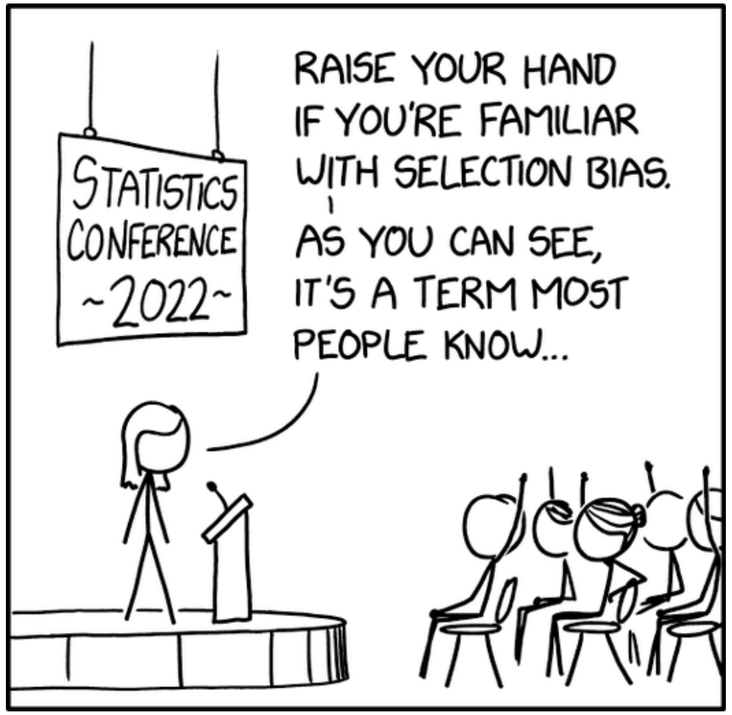
- Focus on quantitative research:
  - description – explanation
  - prediction – control



(c) Randall Munroe - xkcd.com

This course

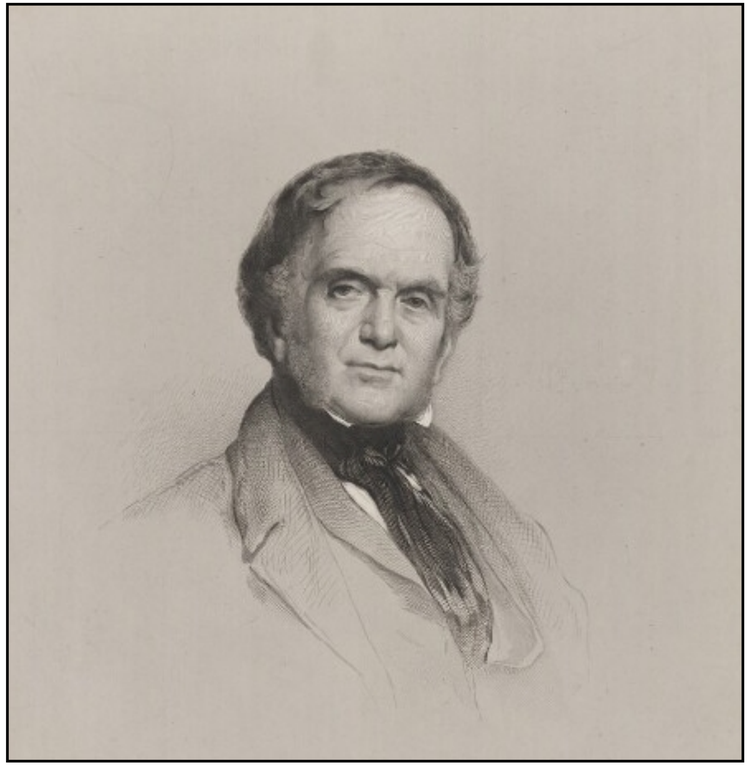
- Focus on quantitative research:
  - description – explanation
  - prediction – control
- Requirements:
  - objectivity — reliability — validity
- Statistics versus visualizations
- Limitations statistics ("lies, damned lies, and statistics")
  → good visualizations as a starting point

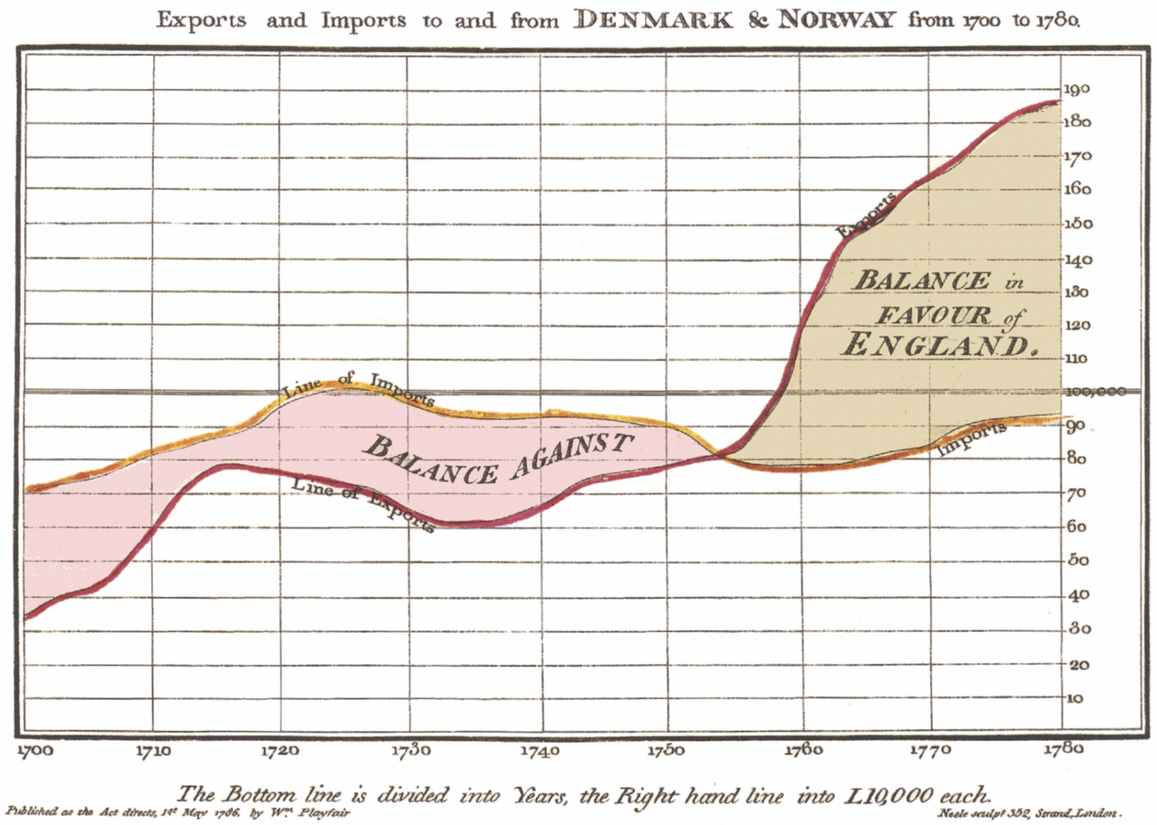


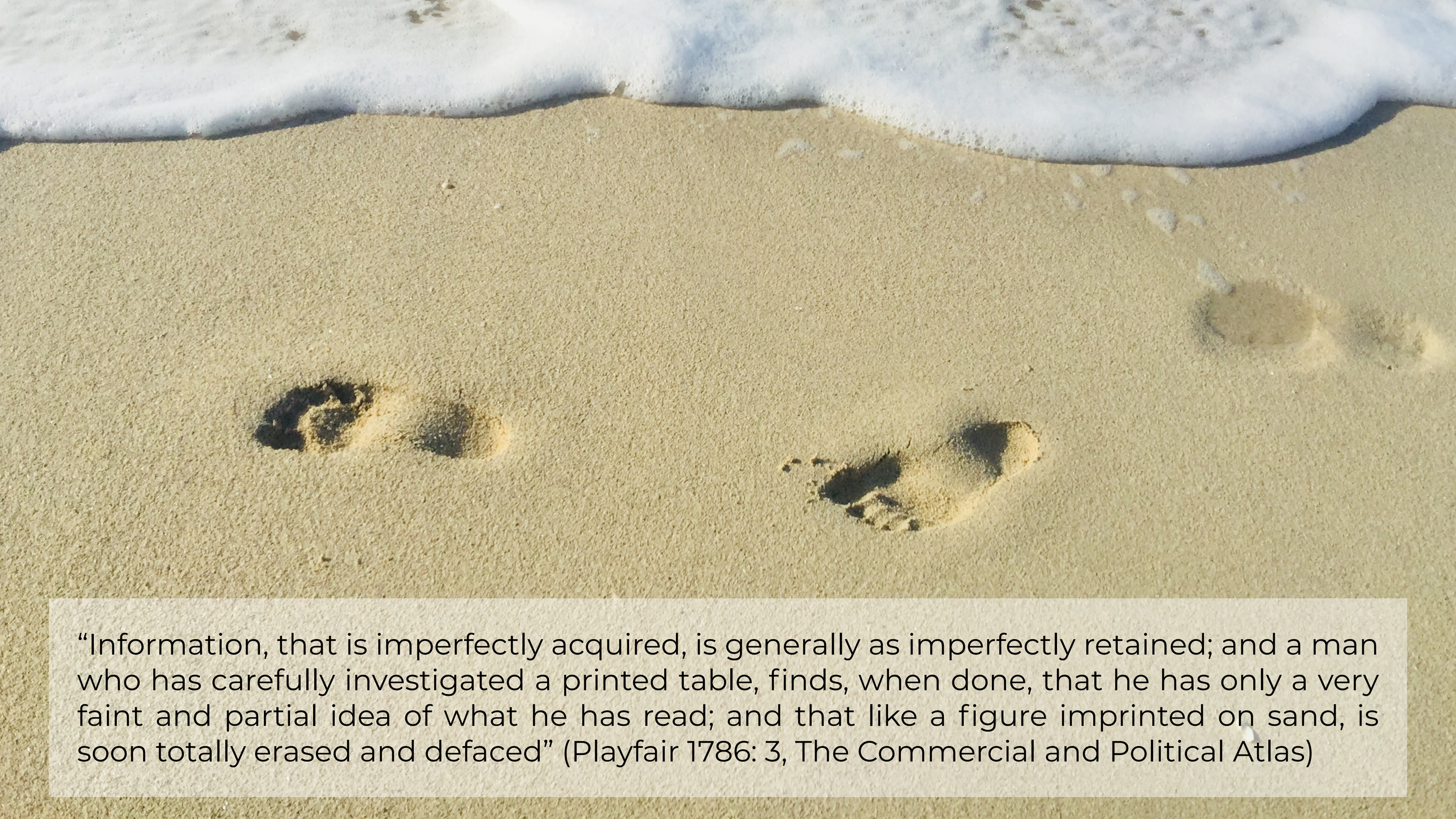
(c) Randall Munroe - xkcd.com

Pioneers

William Playfair (1759-1823)
Scottish engineer, political economist, and pioneer in modern graphical design







“Information, that is imperfectly acquired, is generally as imperfectly retained; and a man who has carefully investigated a printed table, finds, when done, that he has only a very faint and partial idea of what he has read; and that like a figure imprinted on sand, is soon totally erased and defaced” (Playfair 1786: 3, The Commercial and Political Atlas)

Pioneers

Charles Minard (1781-1870), French civil engineer and another pioneer in modern graphical design (designer of ‘flow maps’)

Carte figurative des pertes successives en hommes de l'Armée Française dans la campagne de Russie 1812–1813

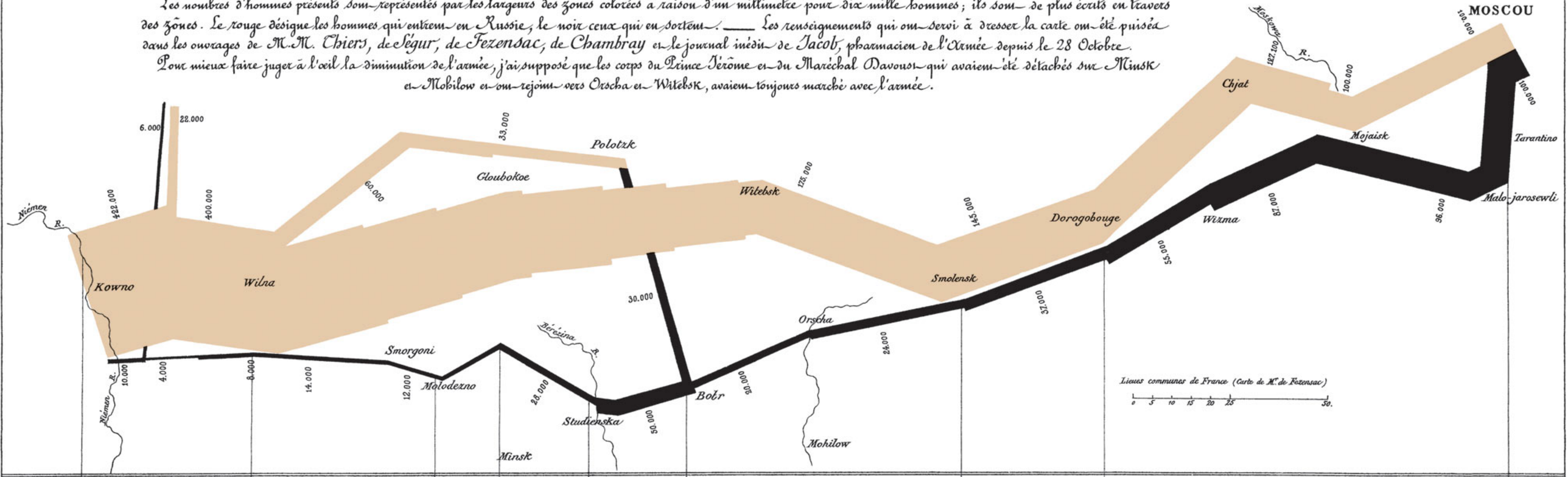
Carte Figurative des pertes successives en hommes de l'Armée Française dans la campagne de Russie 1812–1813.

Dressée par M. Minard, Inspecteur Général des Ponts et Chaussées en retraite.

Paris, le 20 Novembre 1869.

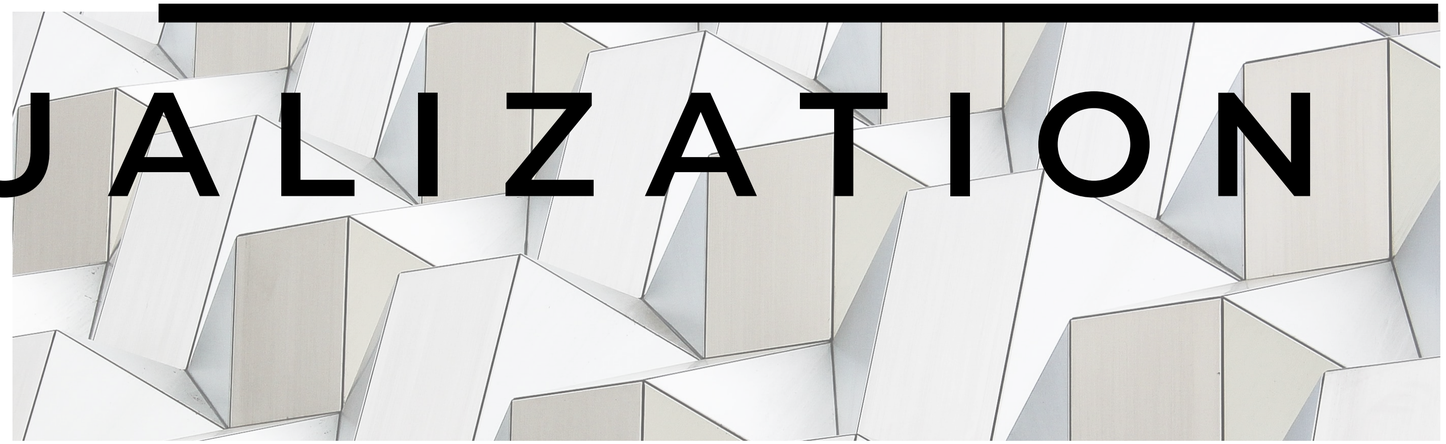
Les nombres d'hommes présents sont représentés par les largeurs des zônes colorées à raison d'un millimètre pour dix mille hommes; ils sont de plus écrits en travers des zônes. Le rouge désigne les hommes qui entrent en Russie, le noir ceux qui en sortent. —— Les renseignements qui ont servi à dresser la carte ont été puisés dans les ouvrages de M.M. Thiers, de Ségur, de Fezensac, de Chambray et le journal inédit de Jacob, pharmacien de l'Armée depuis le 28 Octobre.

Pour mieux faire juger à l'œil la diminution de l'armée, j'ai supposé que les corps du Prince Jérôme et du Maréchal Davoust qui avaient été détachés sur Minsk et Mohilow et ont rejoint vers Orscha et Witebsk, avaient toujours marché avec l'armée.



Autog. par Regnier, 8. Pas. Ste Marie St Gain à Paris.

DATA VISUALIZATION



IL BUONO, IL BRUTTO, IL CATTIVO

IL BUONO, IL BRUTTO, IL CATTIVO

Introduction

The Good

Following the principles of effective data visualization

The Bad

Perception-based problems; misleading, unclear, confusing, ...

The Ugly

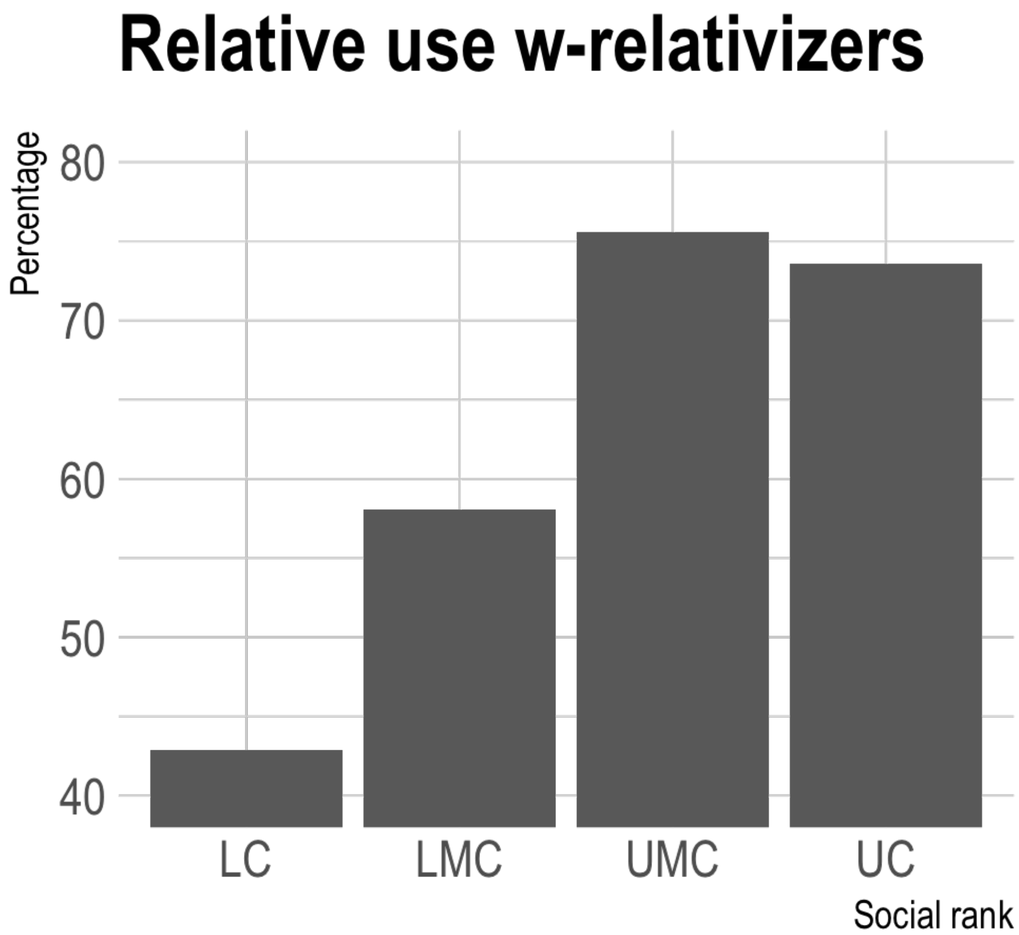
Aesthetic problems, but not incorrect

The Wrong

Mathematical problems; objectively incorrect

The Bad

Perception-based problems;
misleading, unclear, confusing, ...



Data: Rutten & Van der Wal (2014: 298)

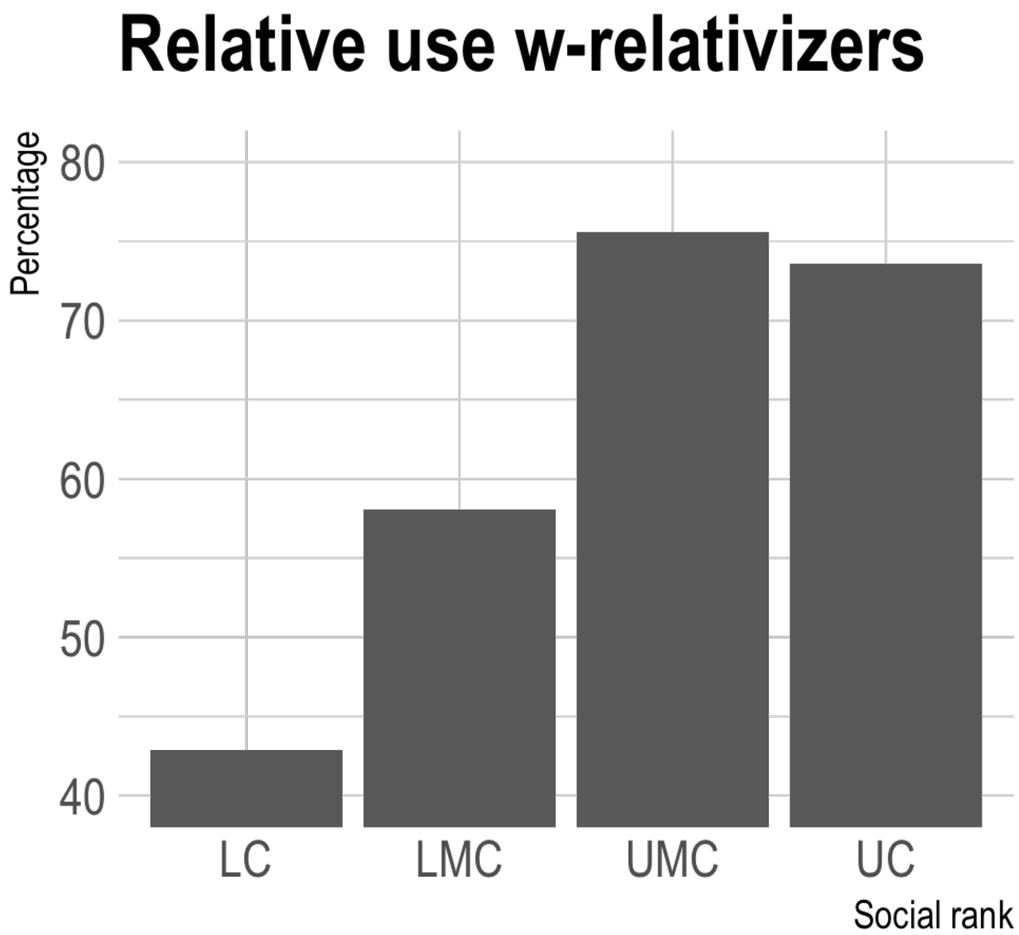
not authors' visualization!

The Bad

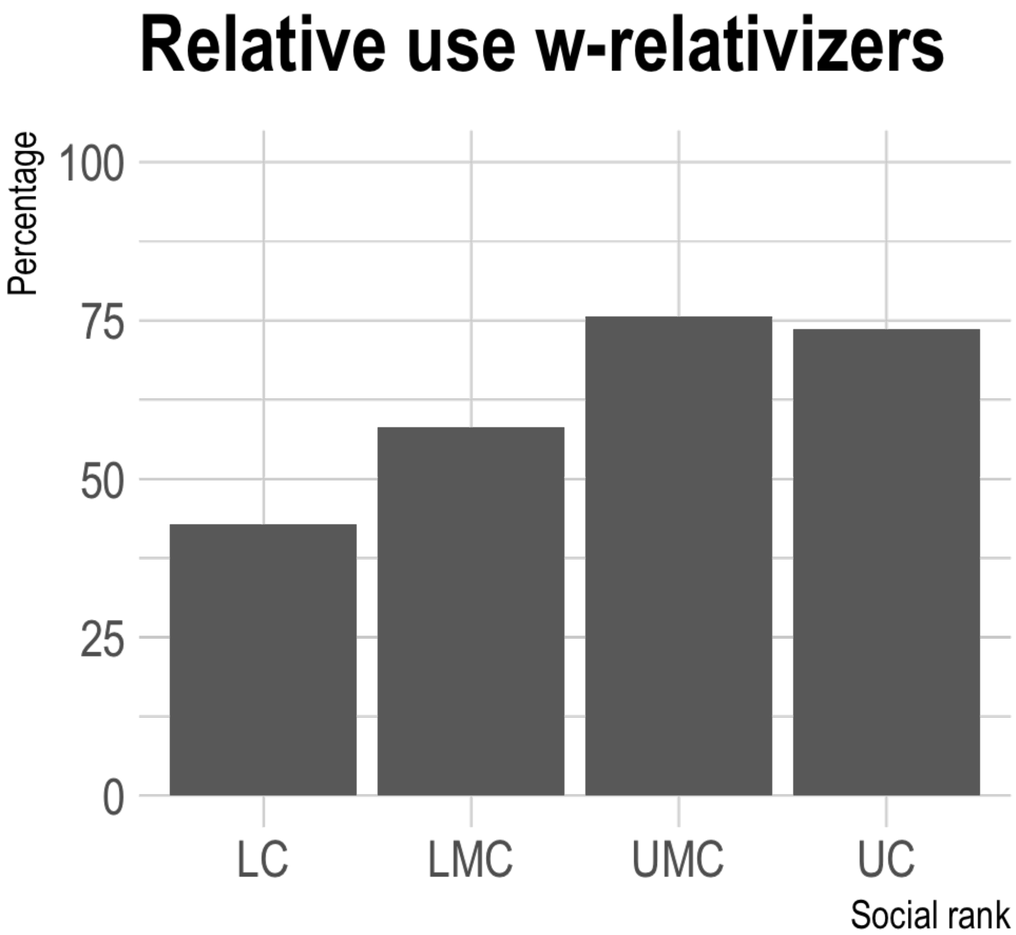
Perception-based problems; misleading, unclear, confusing, ...

The Good

Following the principles of effective data visualization



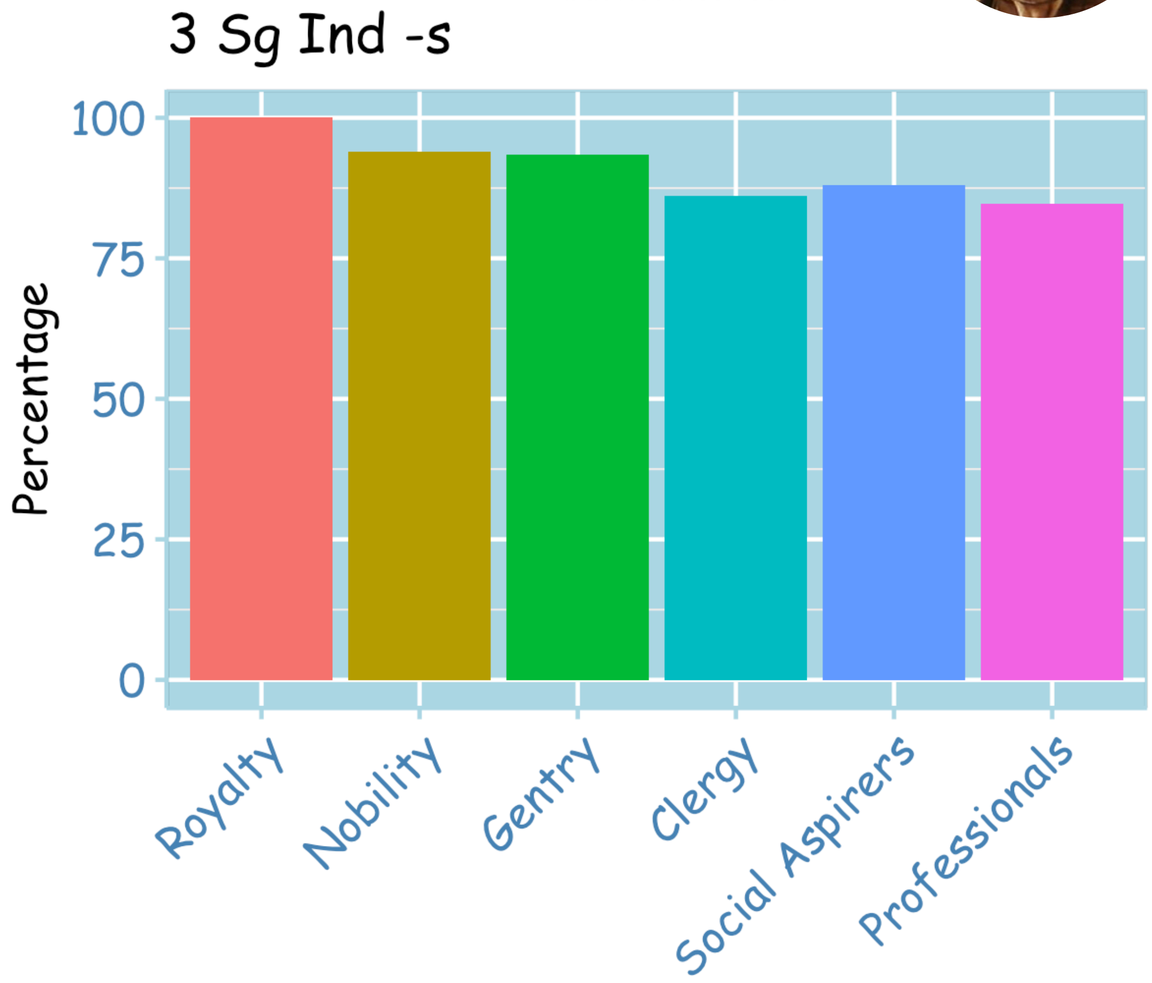
Data: Rutten & Van der Wal (2014: 298)



Data: Rutten & Van der Wal (2014: 298)

The Ugly

Aesthetic problems,
but not incorrect



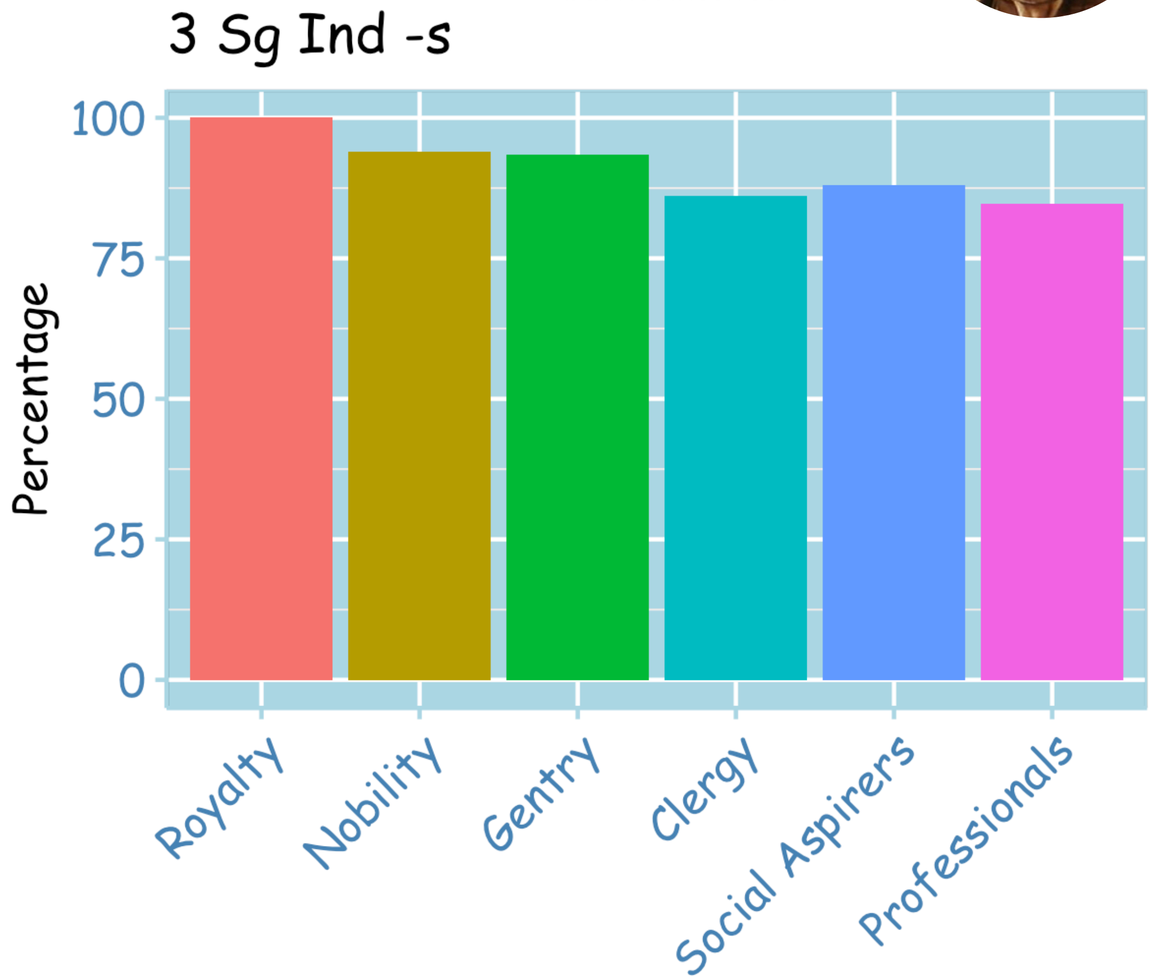
Data: Nevalainen & Raumolin-Brunberg (2016: 140); male informants, 1640–1681

The Ugly

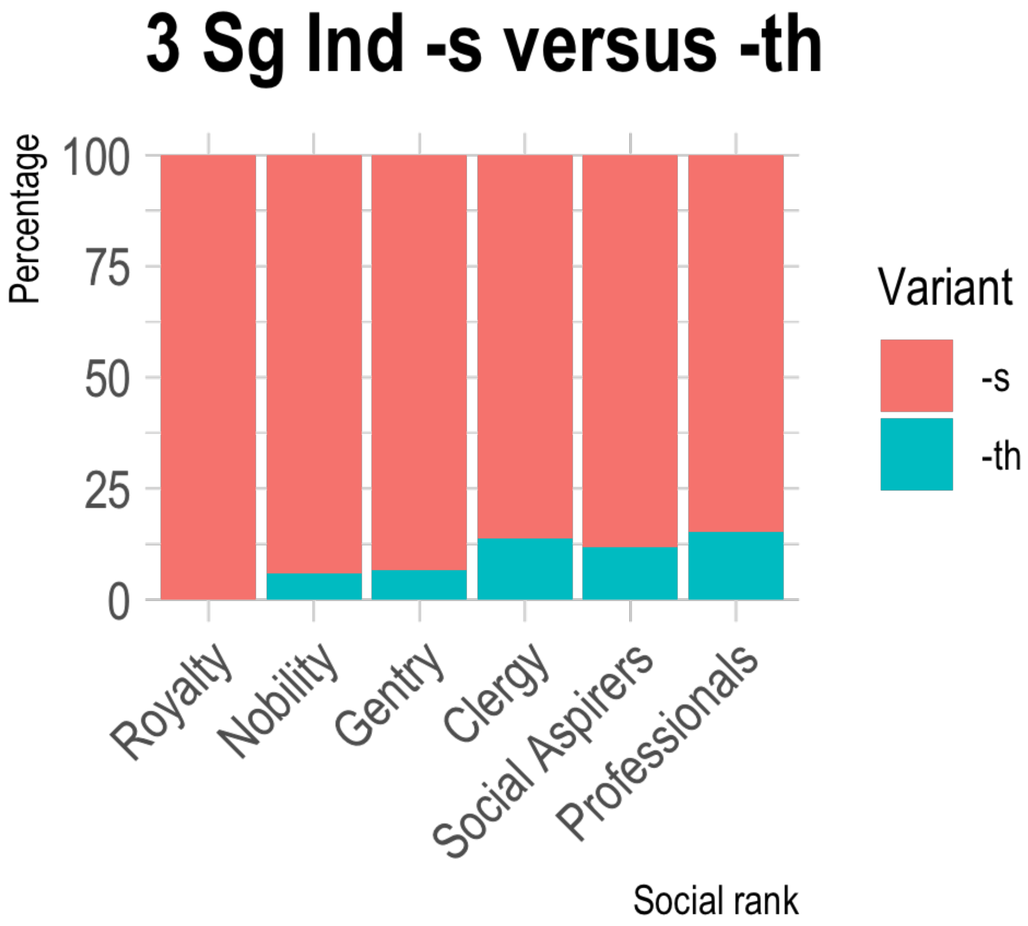
Aesthetic problems,
but not incorrect

The Good

Following the principles of
effective data visualization



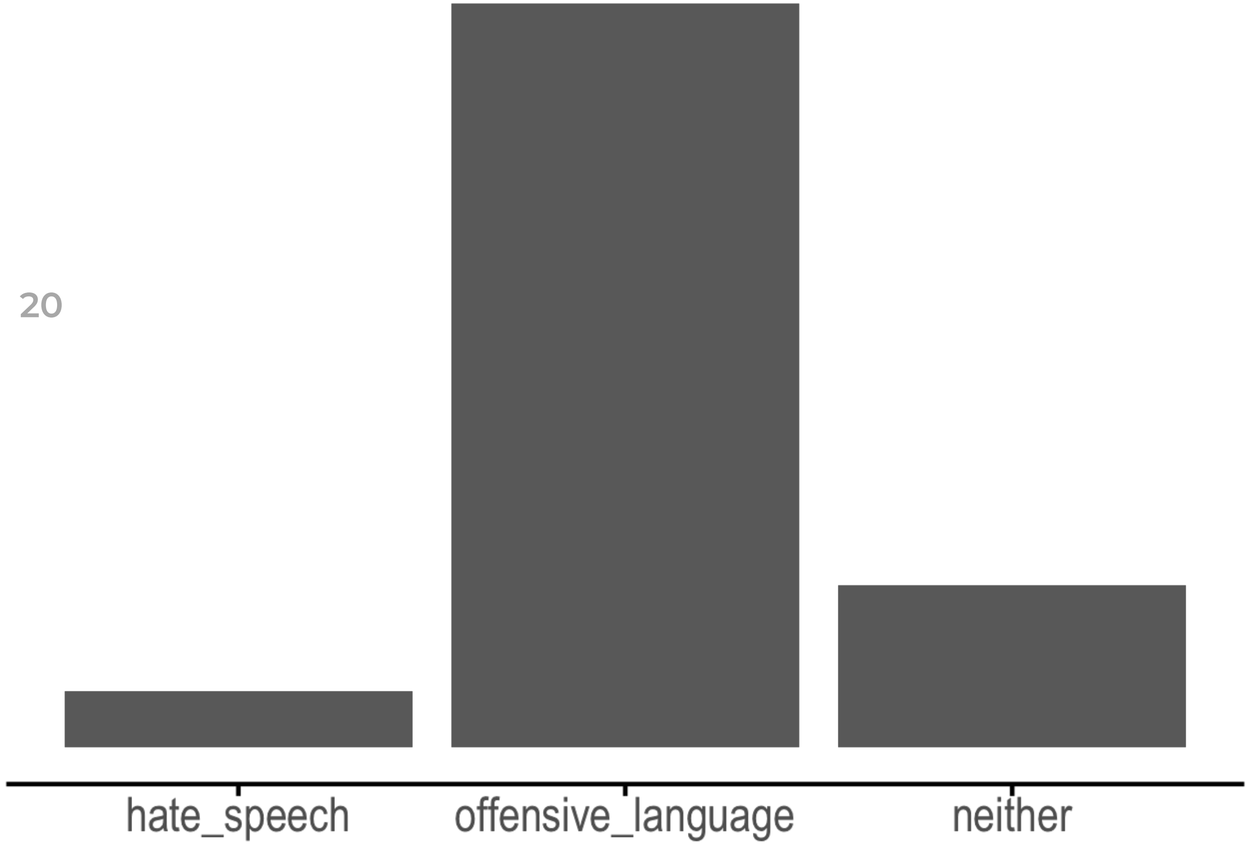
Data: Nevalainen & Raumolin-Brunberg
(2016: 140); male informants, 1640–1681



Data: Nevalainen & Raumolin-Brunberg
(2016: 140); male informants, 1640–1681

The Wrong

Mathematical problems;
objectively incorrect



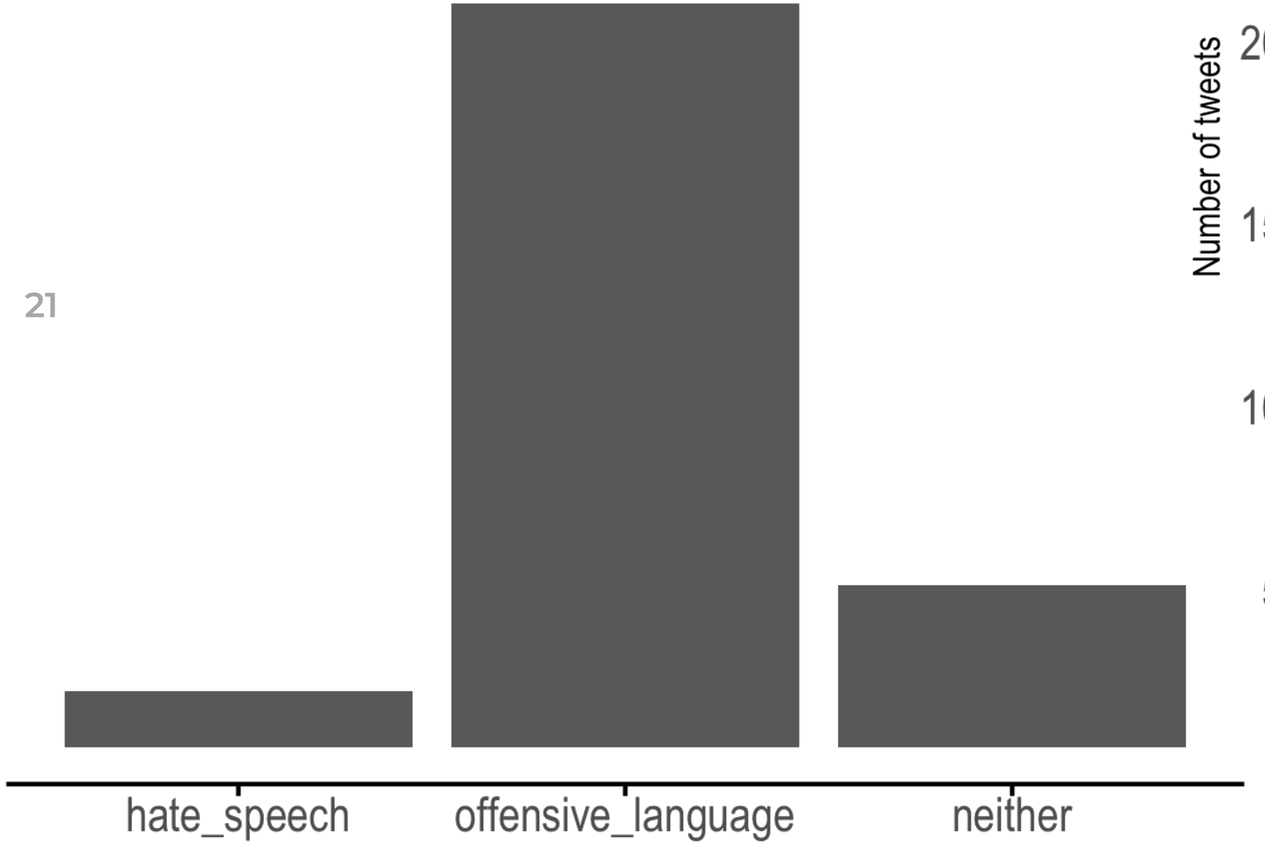
Data: www.kaggle.com

The Wrong

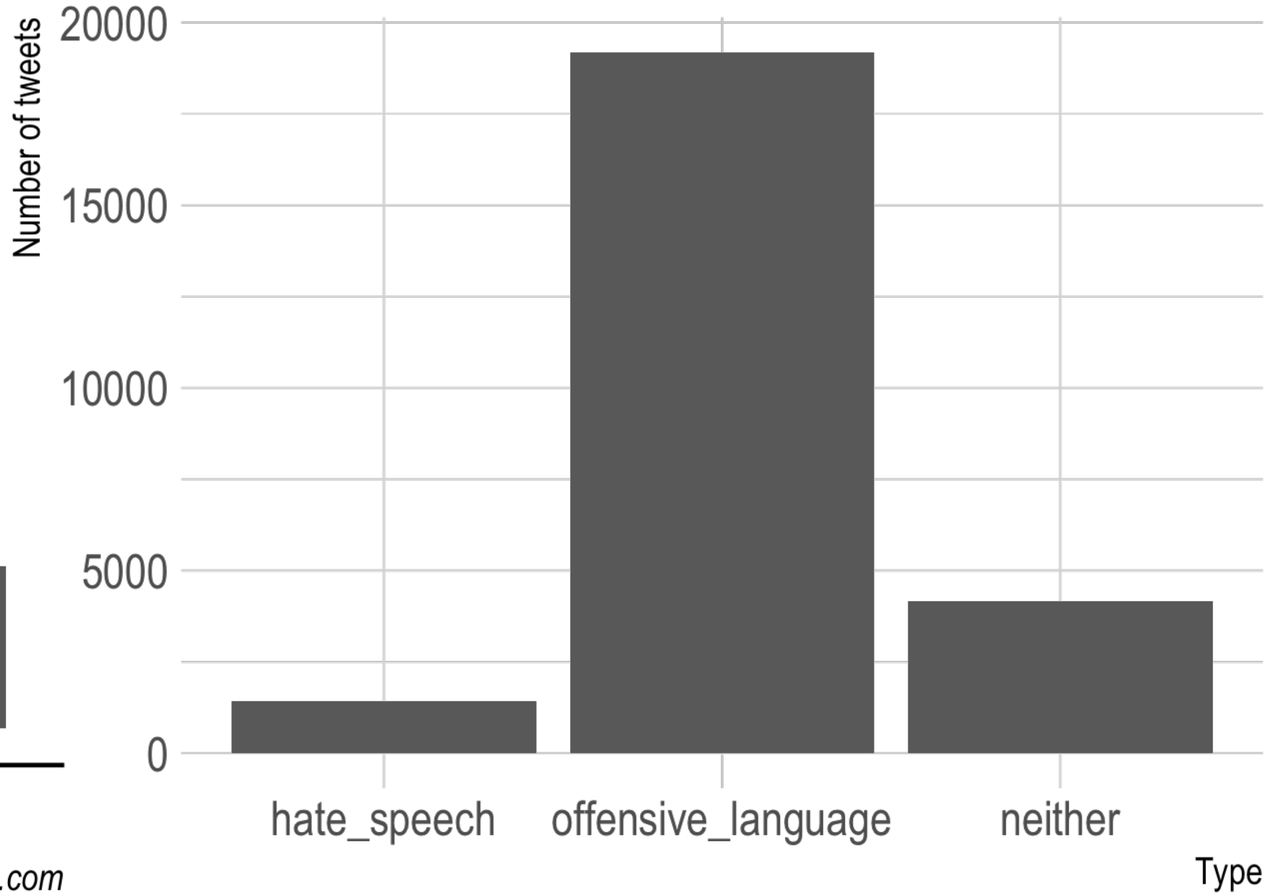
Mathematical problems; objectively incorrect

The Good

Following the principles of effective data visualization

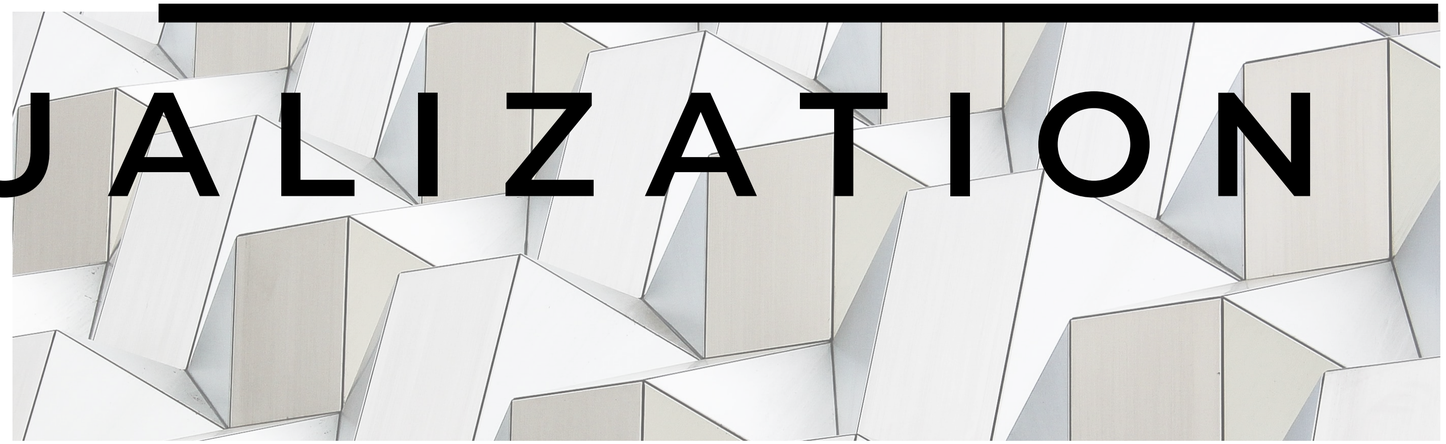


Data: www.kaggle.com



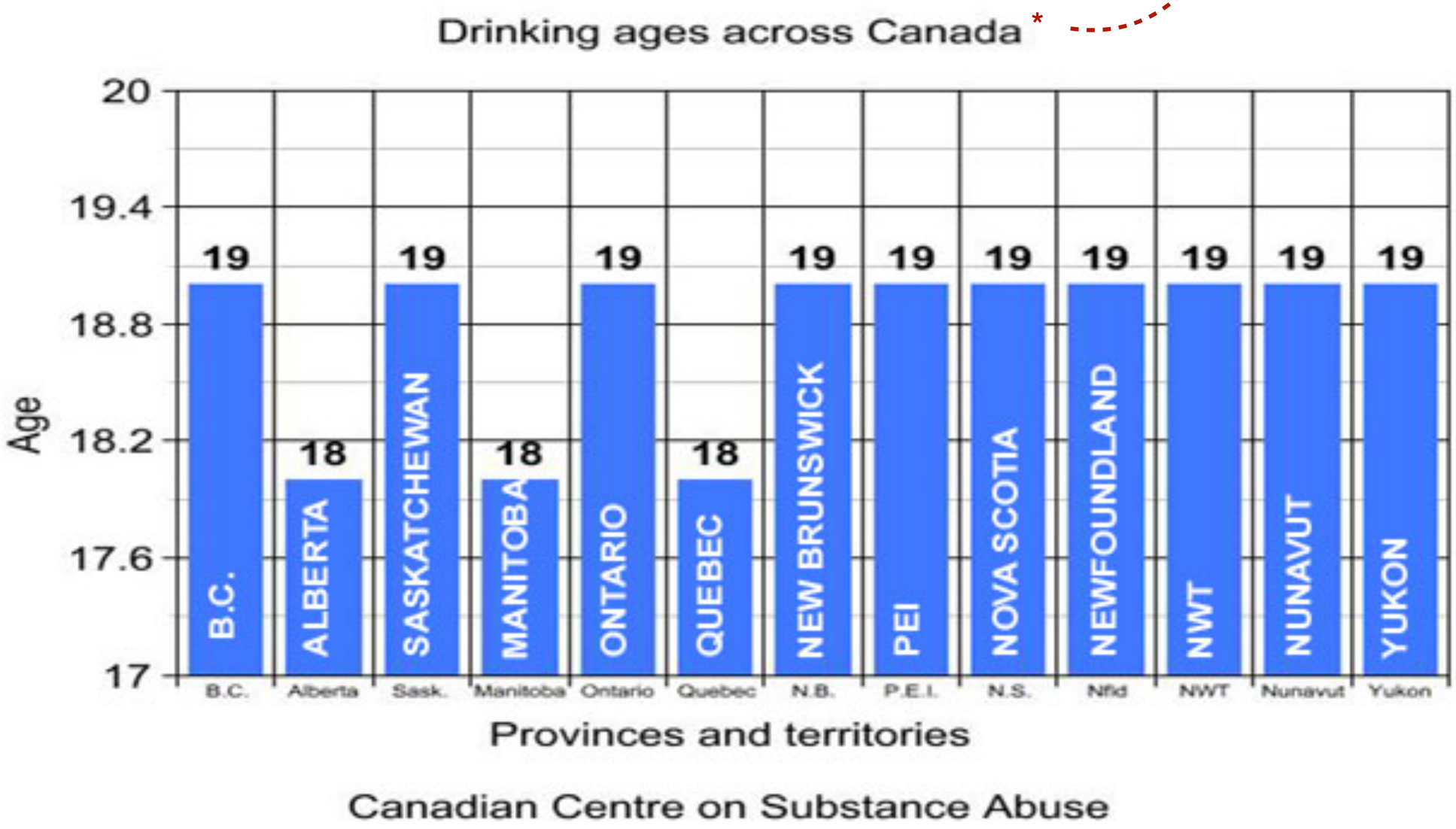
Data: www.kaggle.com

DATA VISUALIZATION



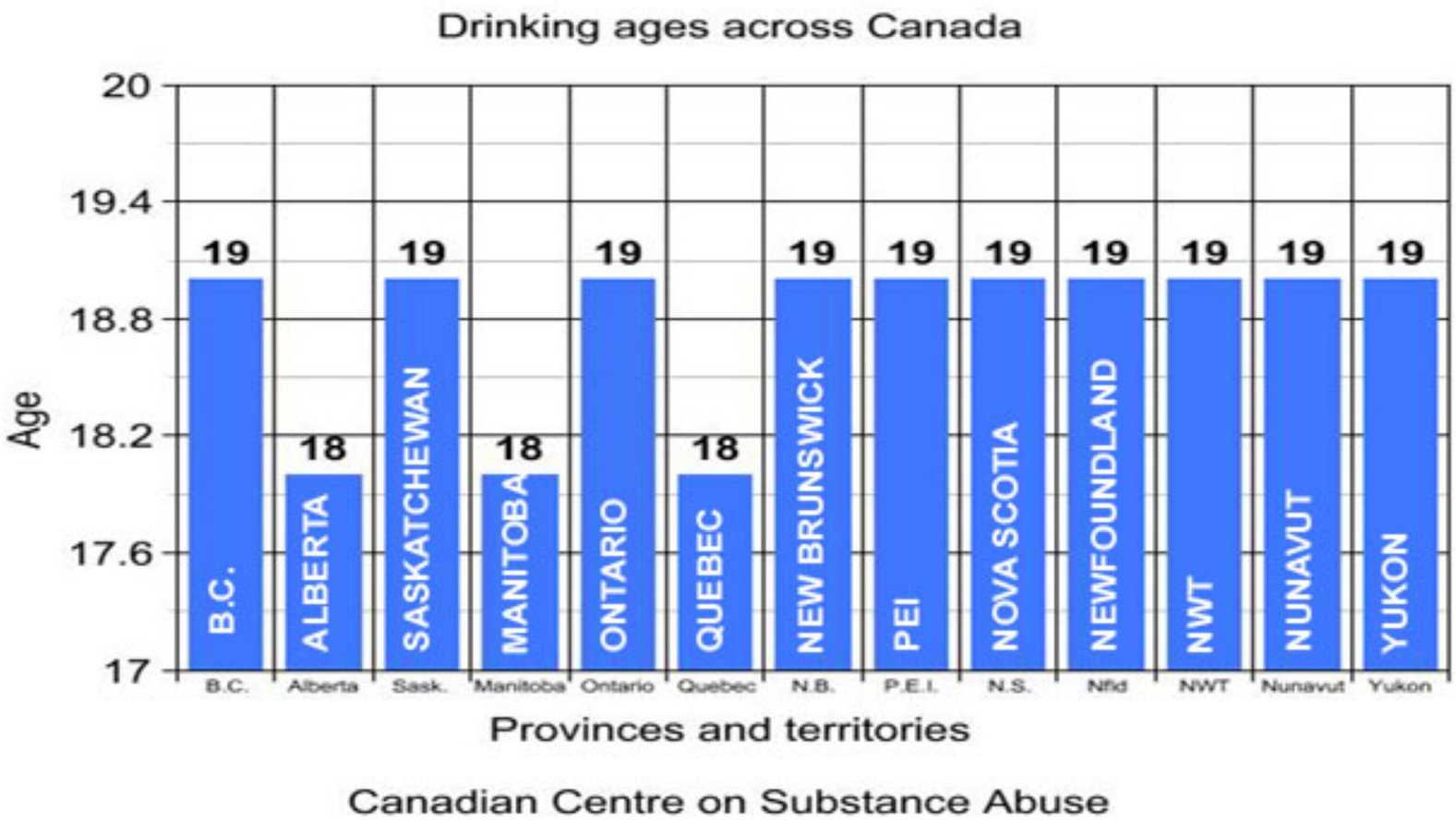
SOME PRACTICE

* (Probably a fabricated visualization…)



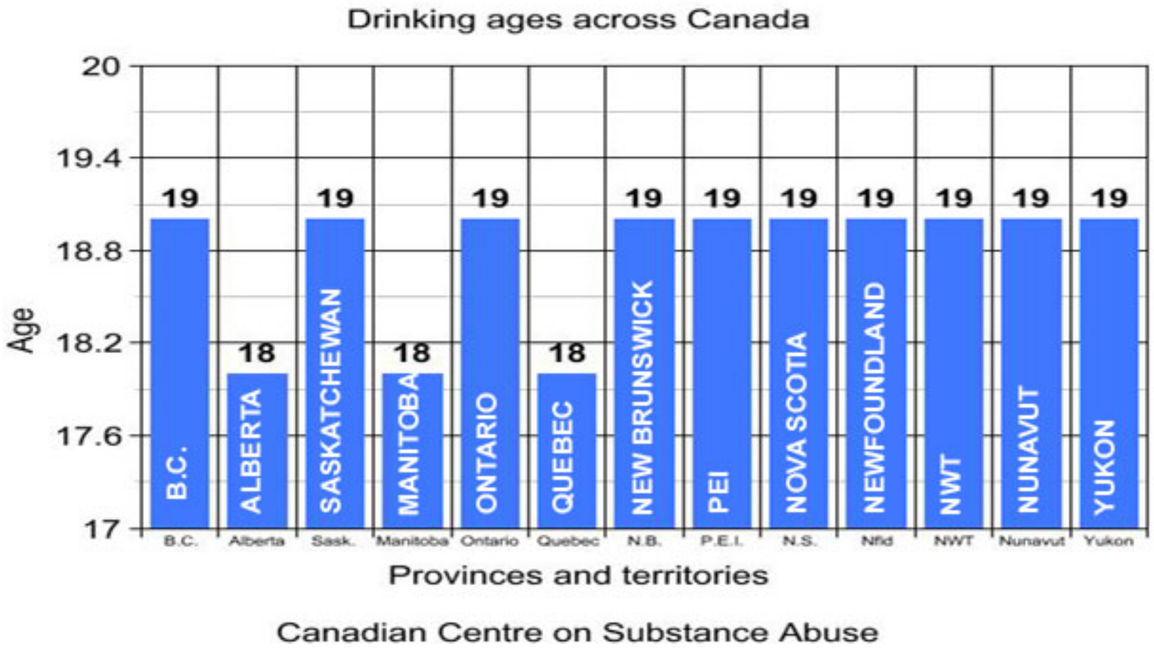
The Bad

Perception-based problems;
misleading, unclear, confusing, ...



The Bad

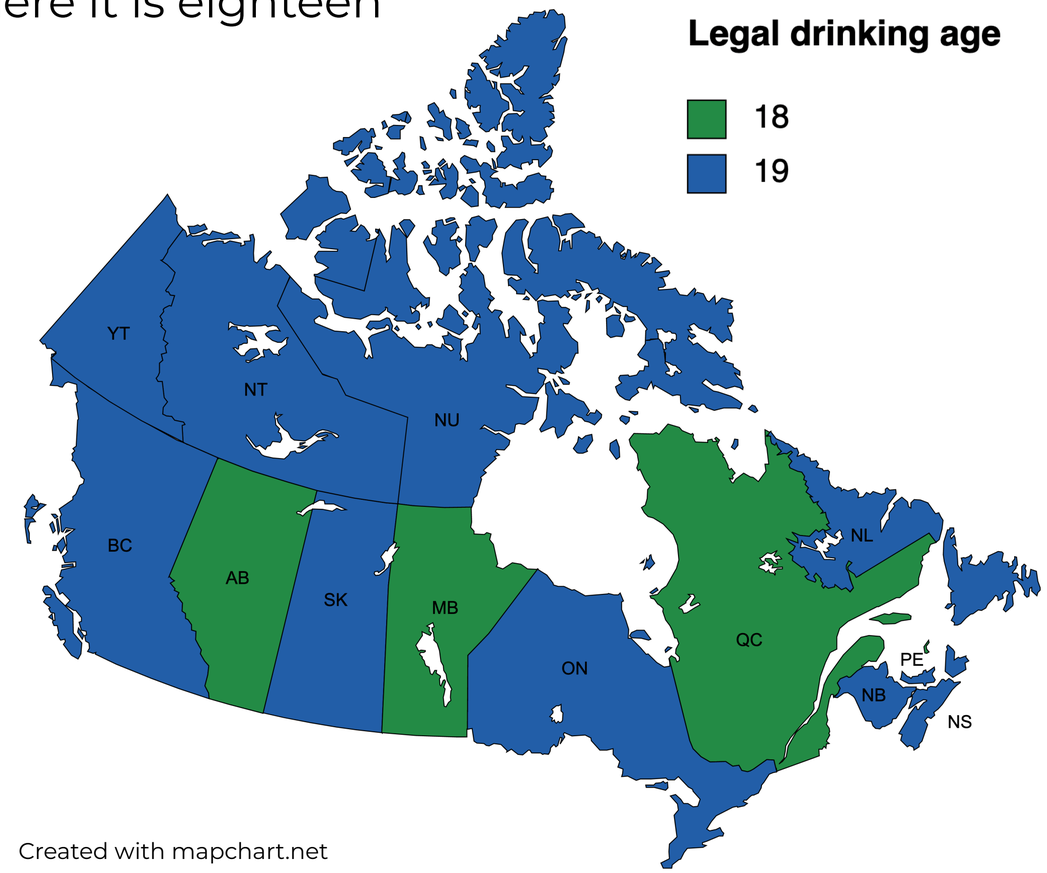
Perception-based problems; misleading, unclear, confusing, ...



The Good

Following the principles of effective data visualization

"The legal drinking age in Canada is nineteen, except for in Alberta, Manitoba and Quebec, where it is eighteen"



Created with mapchart.net

JOB LOSS BY QUARTER
15 MIL
13.5 MIL
9 MIL
7 MIL
DEC '07
SEPT '08
MARCH '09
JUNE '10
FOX NEWS
FOX NEWS
.COM
SOURCE: BLS
AMERICA'S NEWSROOM
N FAIRFAX, VA... BYRD WAS ADMITTED TO THE H(
S&P 3.08

The Wrong

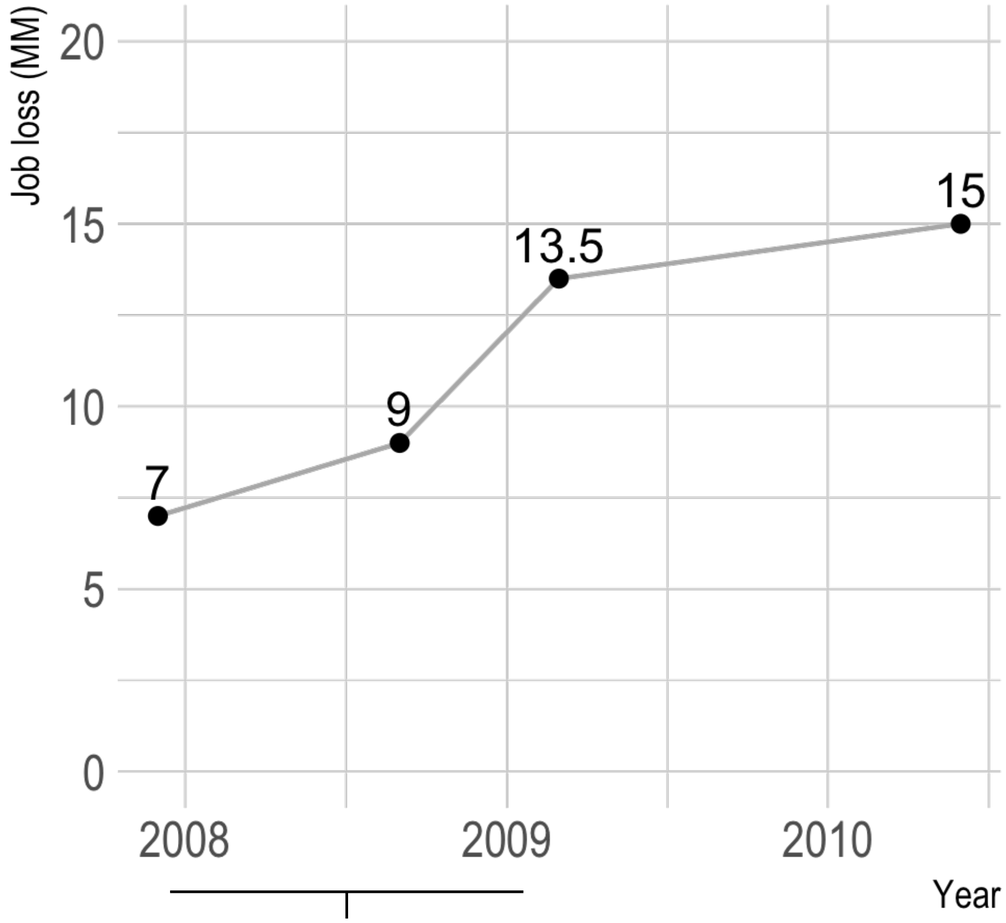
Mathematical problems;
objectively incorrect

The Wrong

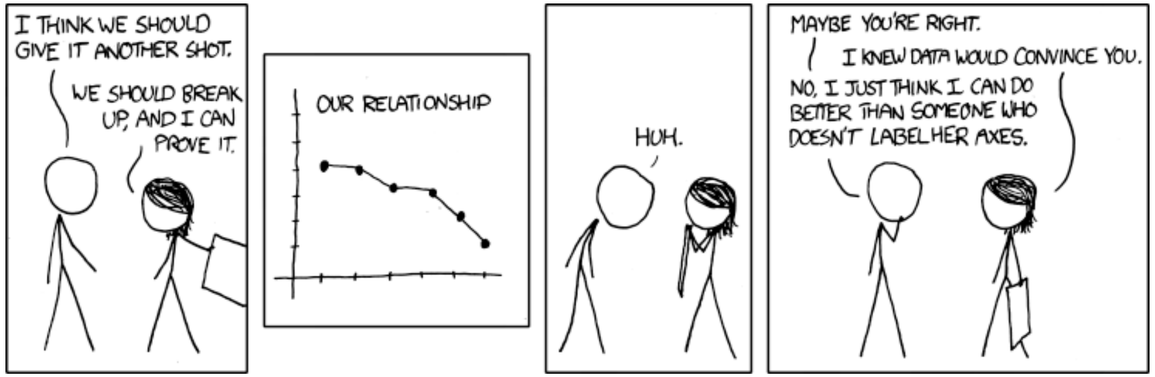
Mathematical problems; objectively incorrect

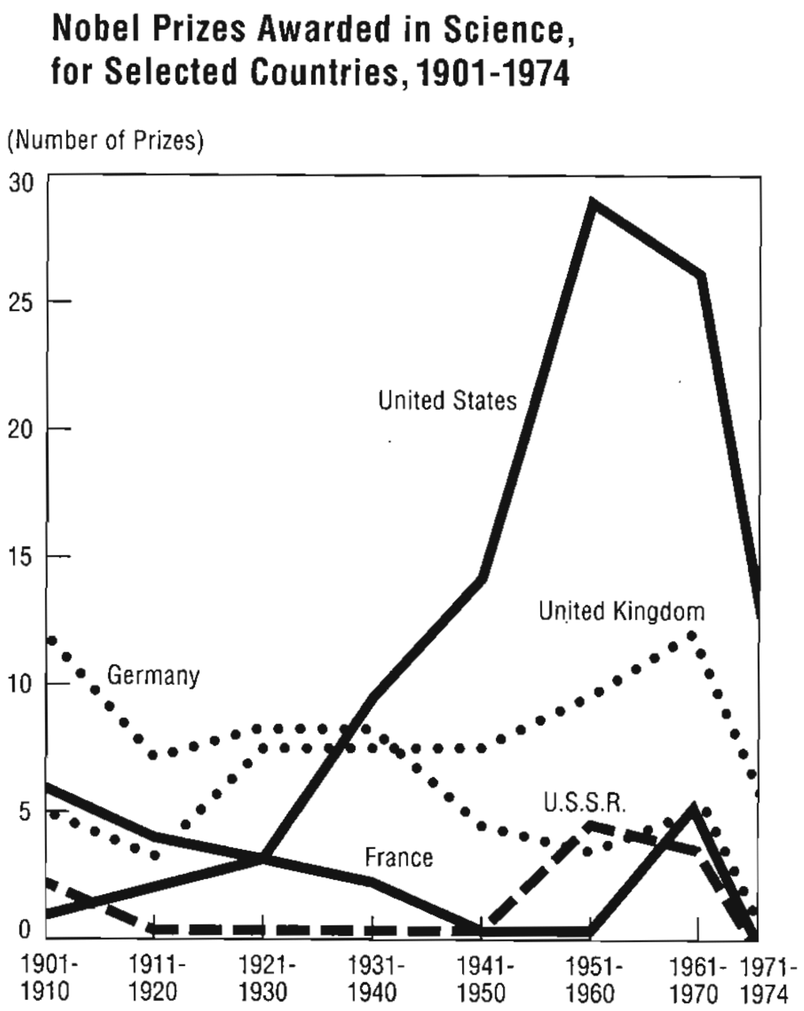
The Good

Following the principles of effective data visualization



2008-2009 Global Financial Crisis



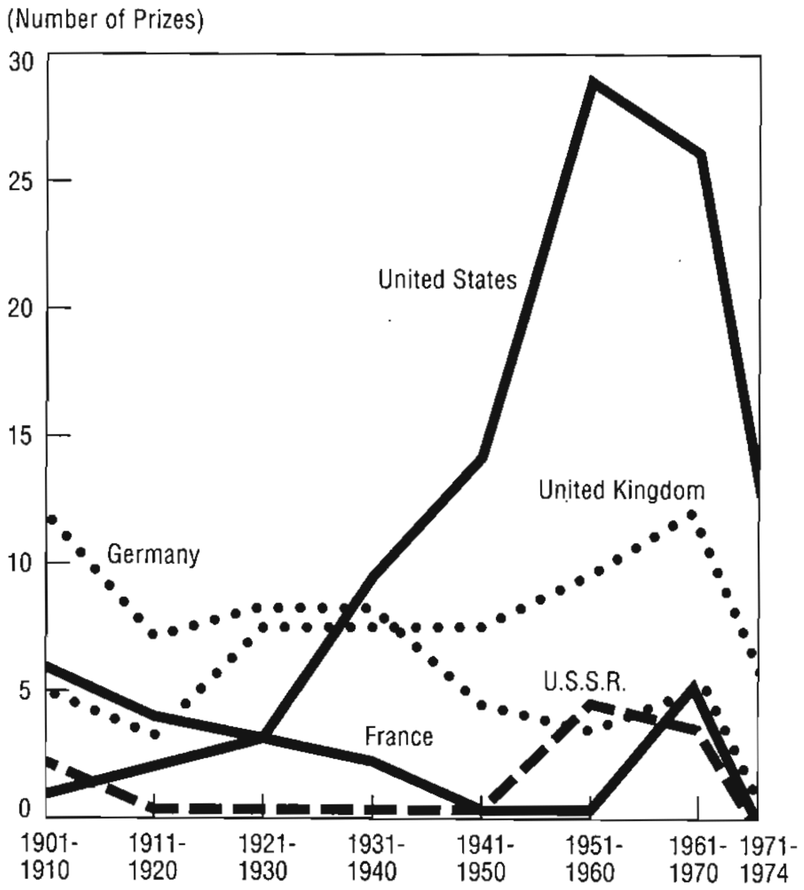


National Science Foundation, Science Indicators (1974: 15), as reproduced in Tufte (2001: 60)

The Bad

Perception-based problems;
misleading, unclear, confusing, ...

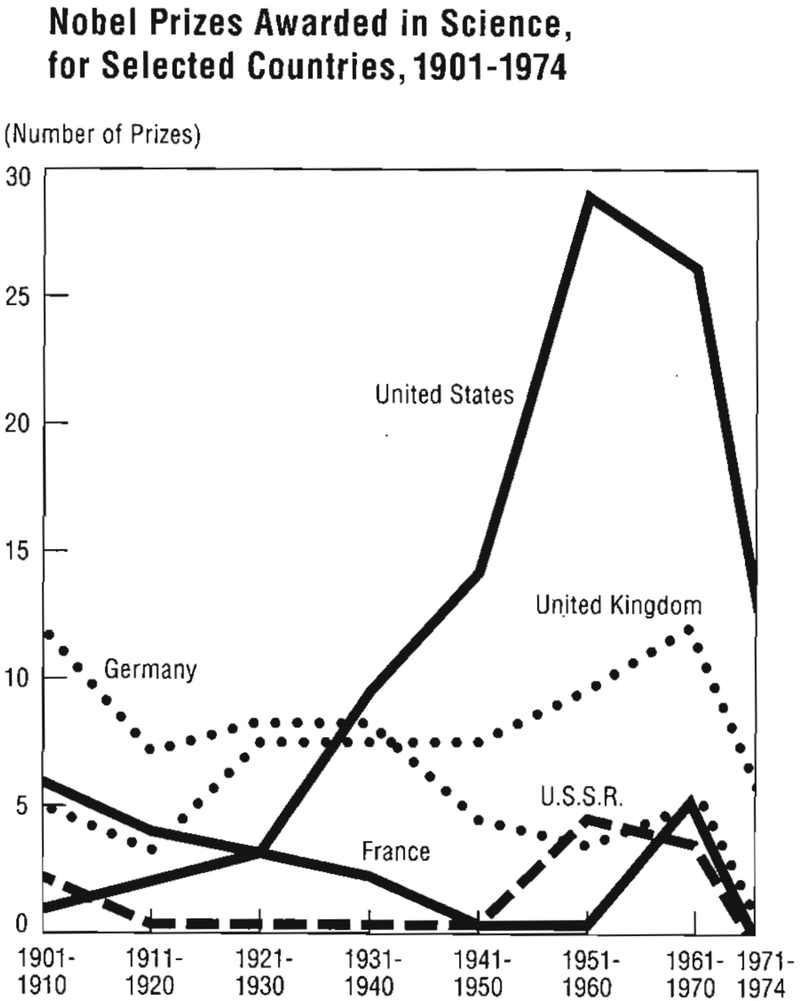
Nobel Prizes Awarded in Science, for Selected Countries, 1901-1974



National Science Foundation, Science Indicators (1974: 15), as reproduced in Tufte (2001: 60)

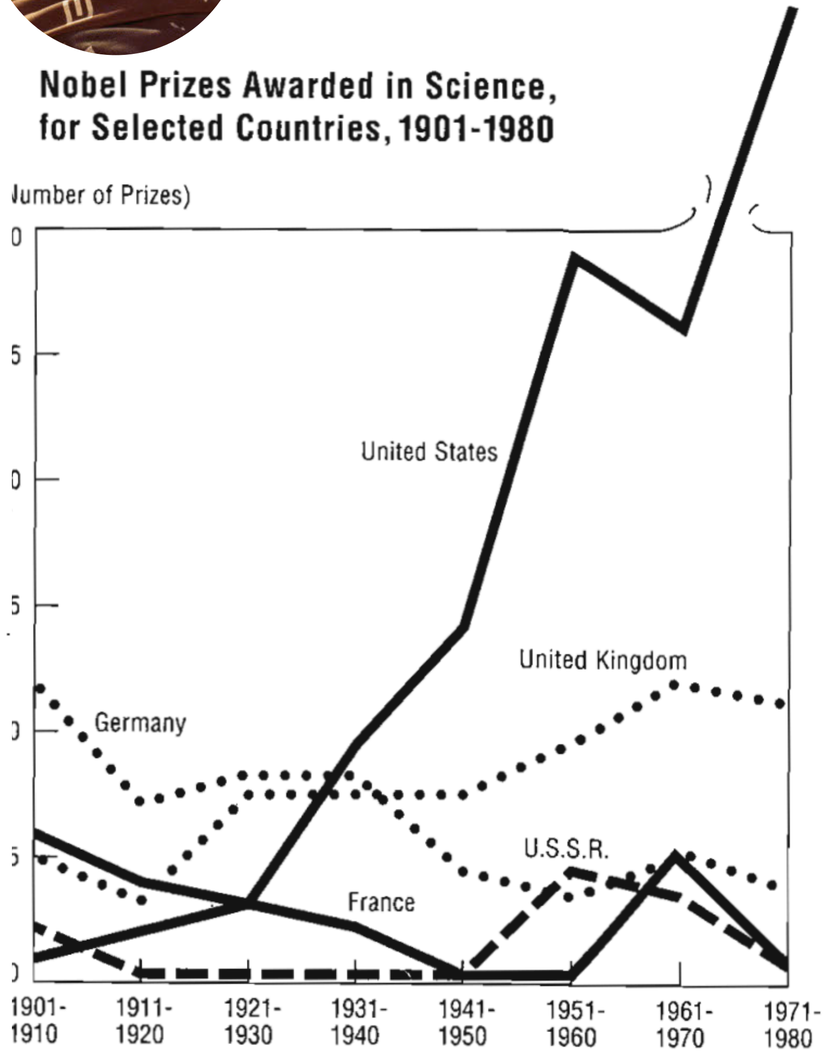
The Bad

Perception-based problems; misleading, unclear, confusing, ...



The Good *(more or less)*

Following the principles of effective data visualization

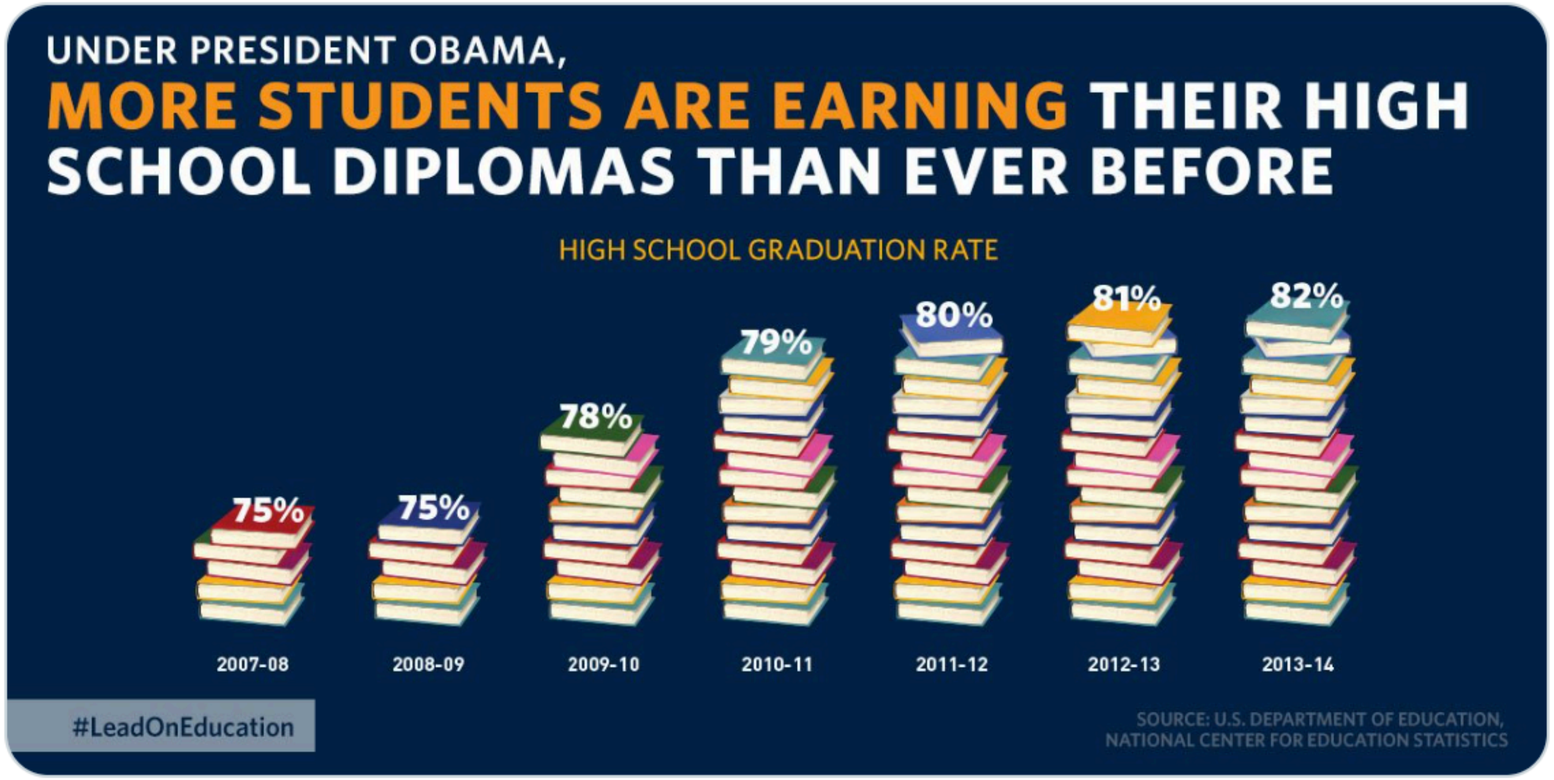


National Science Foundation, Science Indicators (1974: 15), as reproduced in Tufte (2001: 60)

White House Archived
@ObamaWhiteHouse

Good news: America's high school graduation rate has increased to an all-time high.🎓 wapo.st/1m40Mei

Tweet vertalen



7:11 p.m. · 16 dec. 2015 · TweetDeck

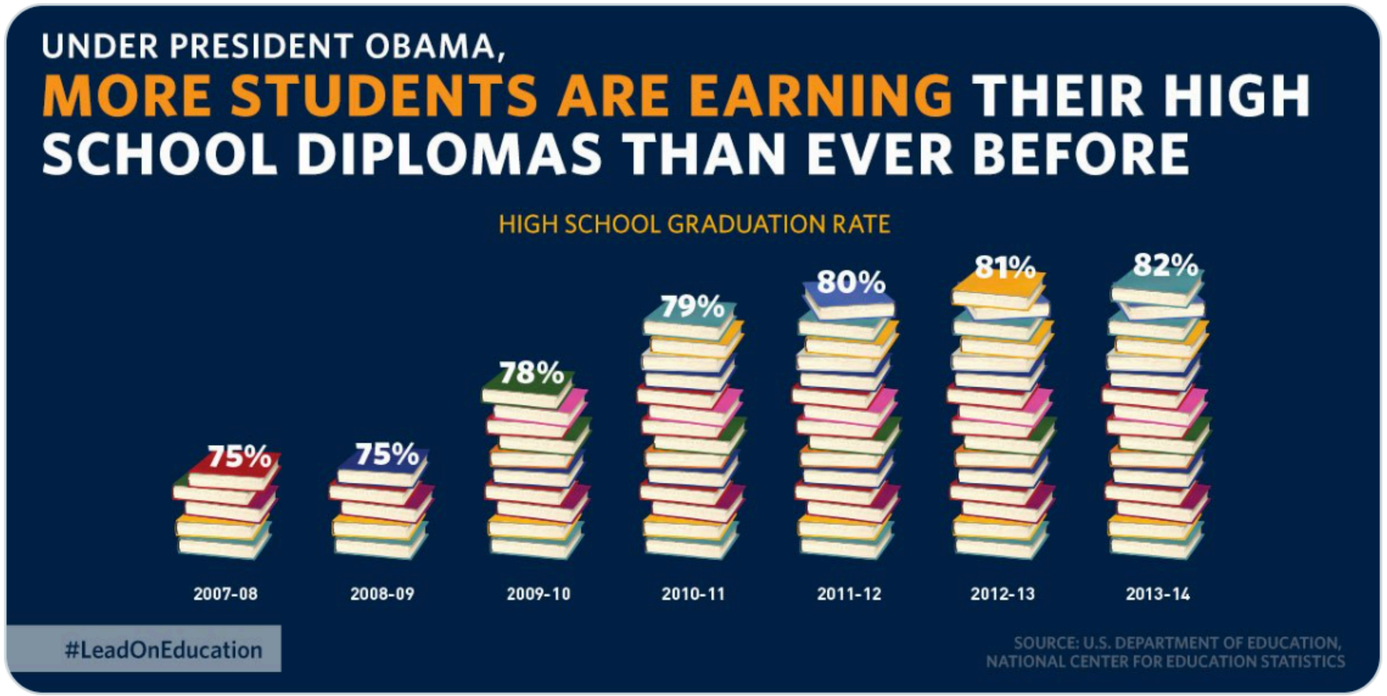
The Bad

Perception-based problems; misleading, unclear, confusing, ...

White House Archived
@ObamaWhiteHouse

Good news: America's high school graduation rate has increased to an all-time high. 🎓 wapo.st/1m40Mei

Tweet vertalen



The Bad

Perception-based problems; misleading, unclear, confusing, ...

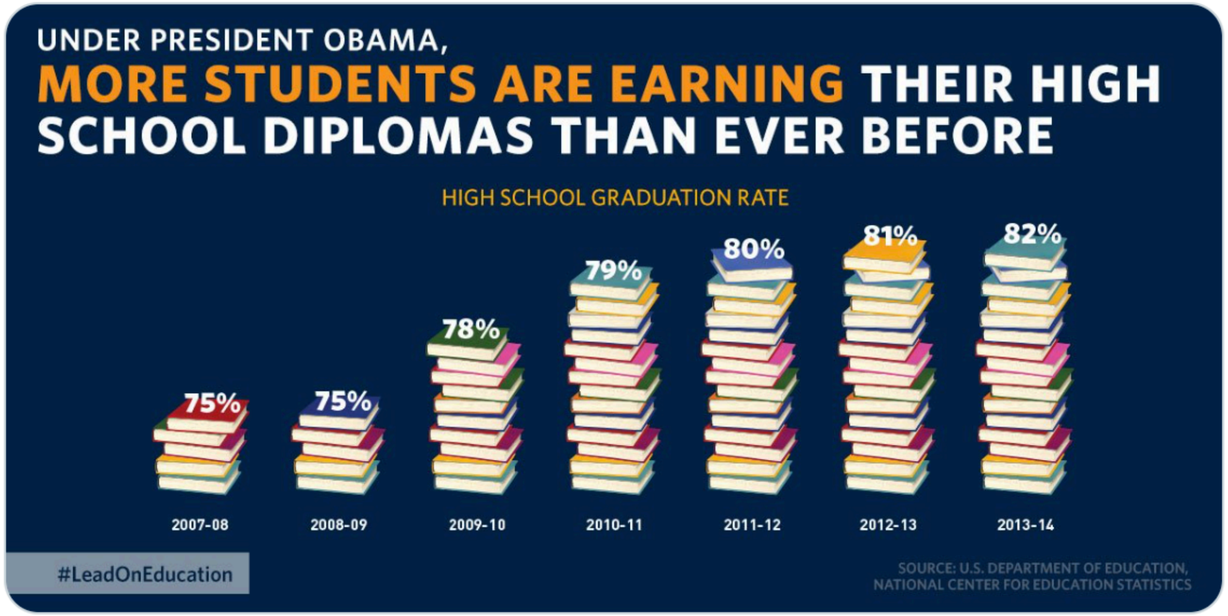
The Good

Following the principles of effective data visualization

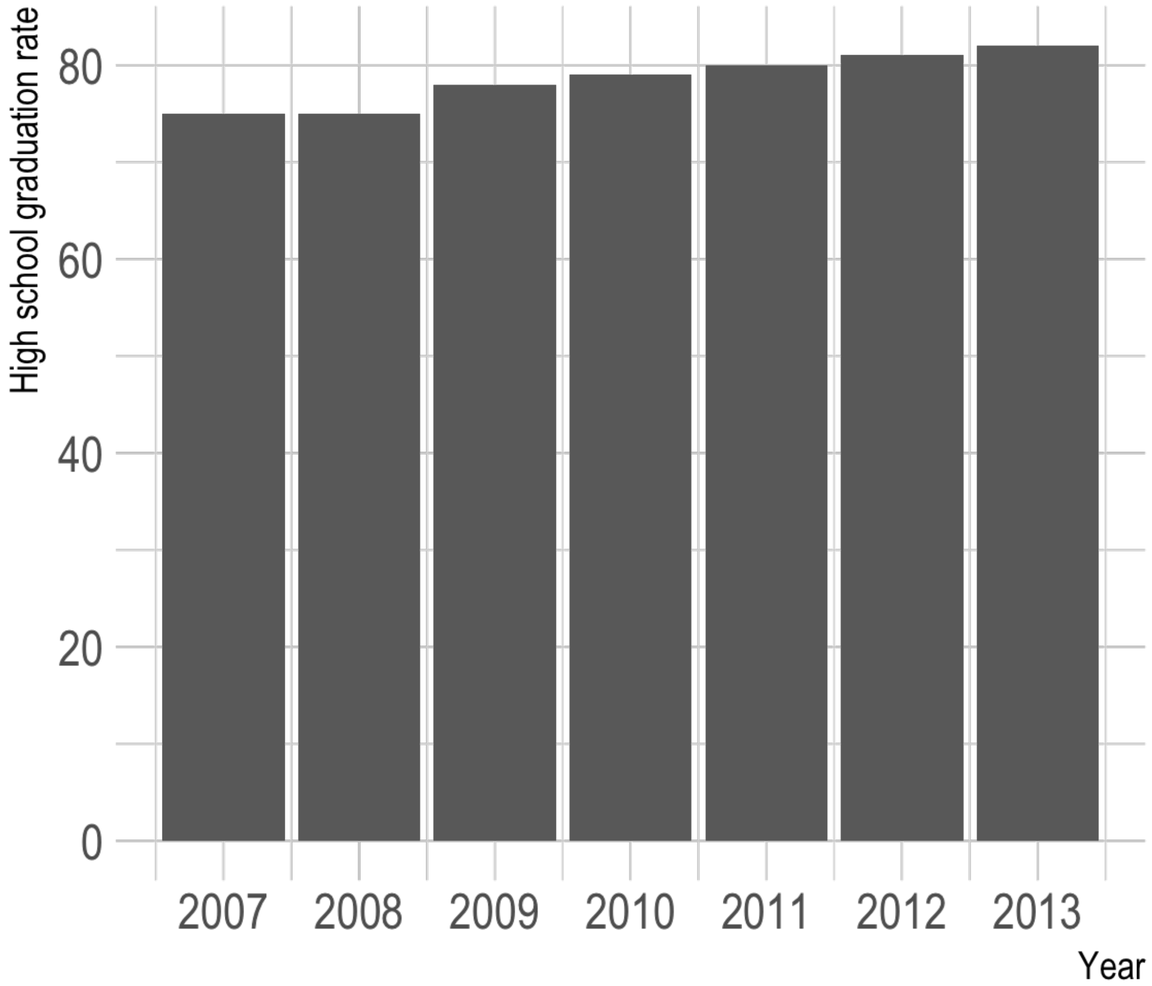
White House Archived
@ObamaWhiteHouse

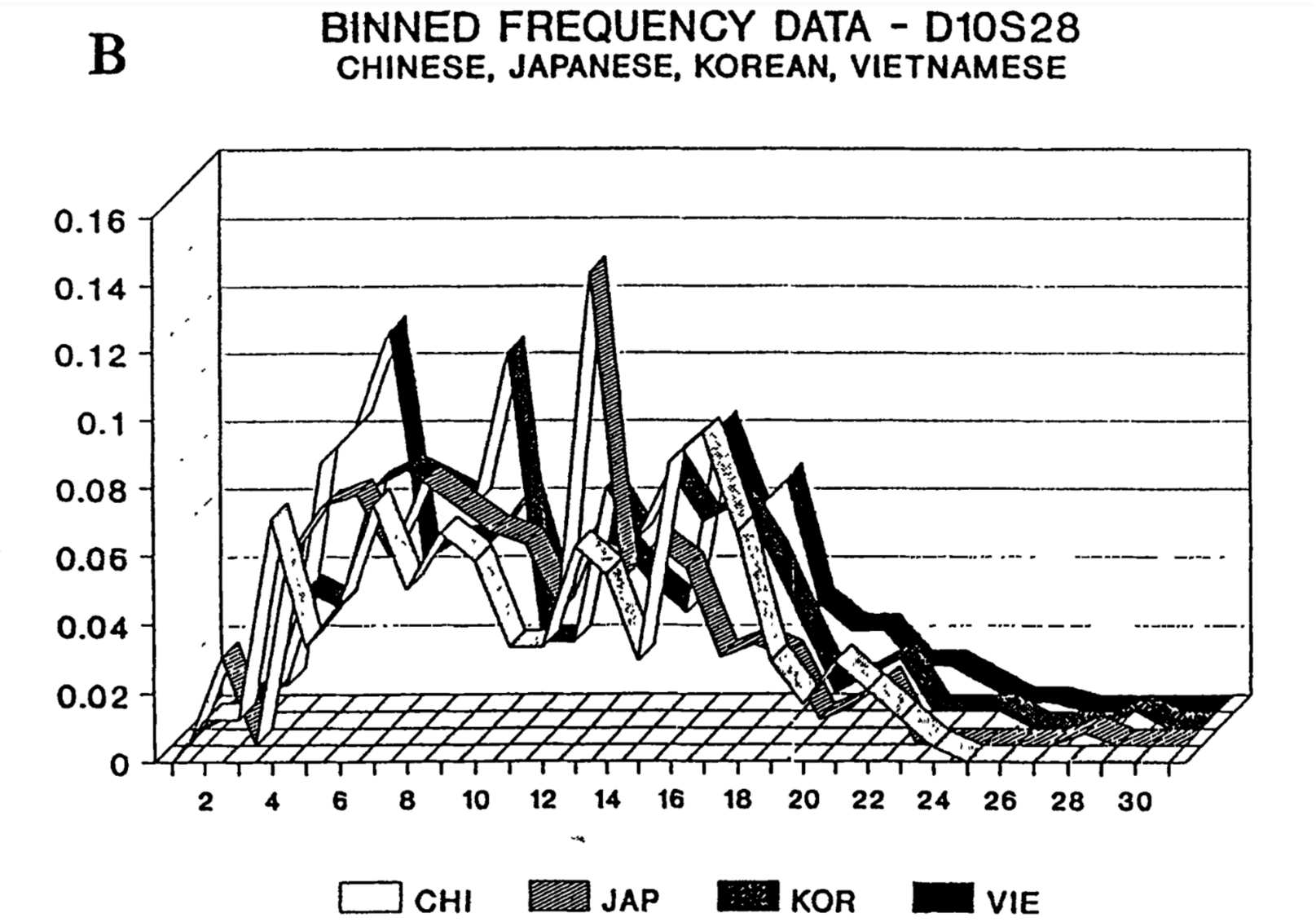
Good news: America's high school graduation rate has increased to an all-time high. 🎓 wapo.st/1m4OMei

Tweet vertalen



7:11 p.m. · 16 dec. 2015 · TweetDeck

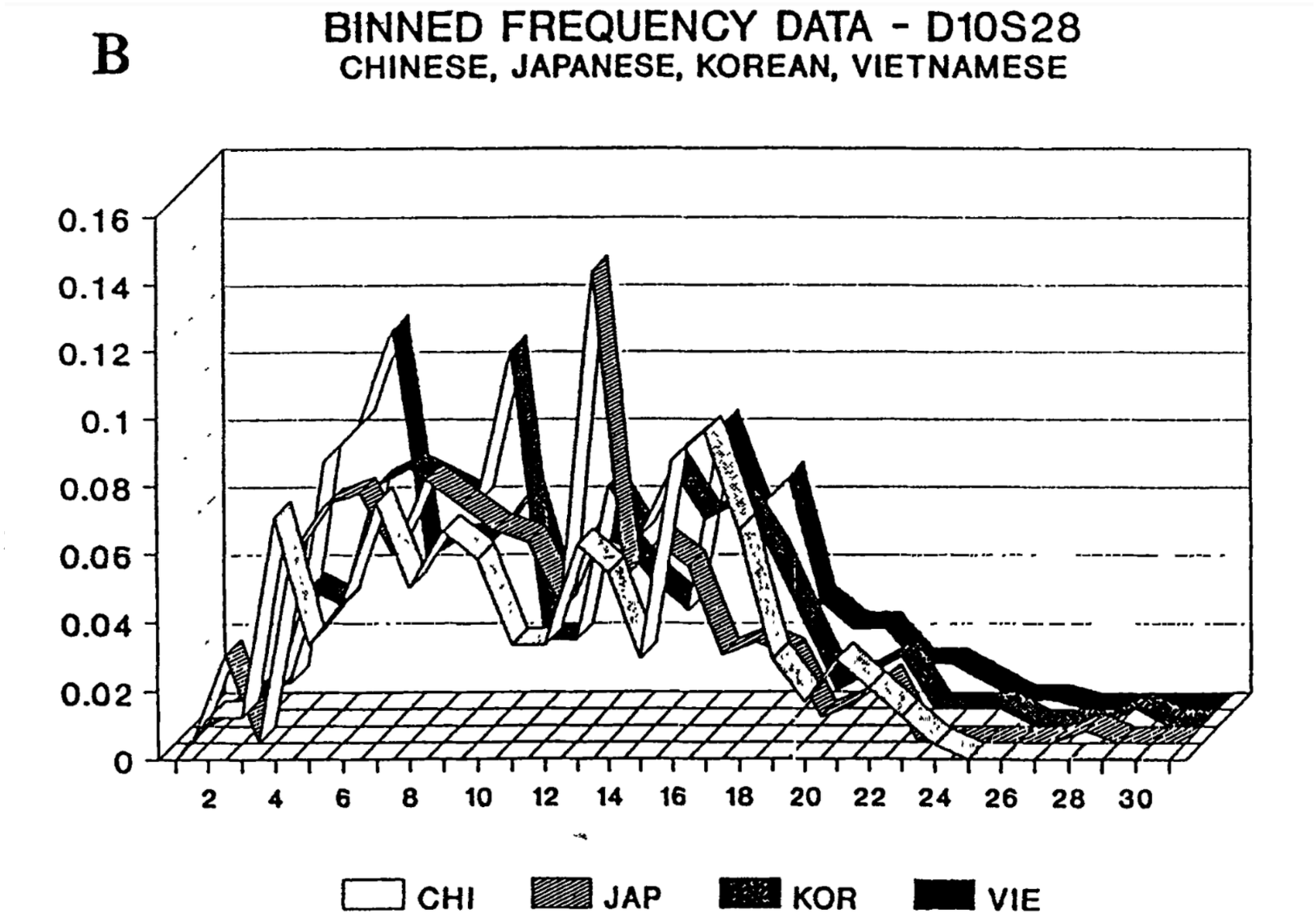




Roeder, K. (1994). DNA fingerprinting: a review of the controversy. Statistical Science, 222-247.

https://www.biostat.wisc.edu/~kbroman/topten_worstgraphs/

FIG. 4. *Fixed bin distribution (histogram) for two loci and four Asian subpopulations (used with permission from John Hartmann): the boundaries of the* 30 *bins (vertical axis) are determined by the FBI; these bins are not of equal length. Sample sizes (numbers of individuals) for Chinese, Japanese, Korean and Vietnamese are* 103, 125, 93 *and* 215 *for* D4S139 *and* 120, 137, 100 *and* 193 *for* D10S28. *The horizontal axis is the bin number; bins are not of equal length.*



Roeder, K. (1994). DNA fingerprinting: a review of the controversy. Statistical Science, 222-247.

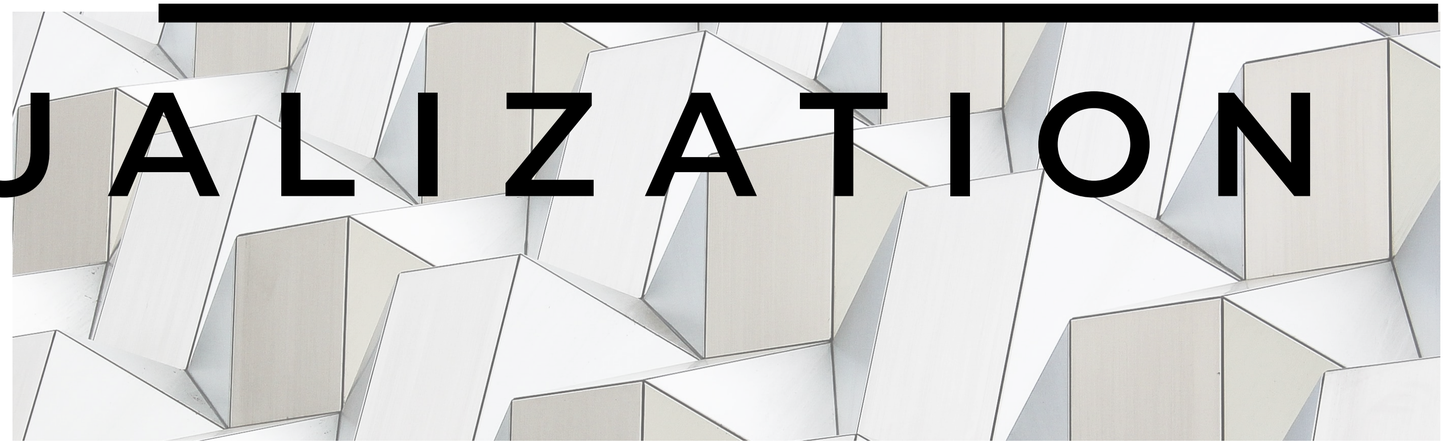
https://www.biostat.wisc.edu/~kbroman/topten_worstgraphs/

The Ugly

Aesthetic problems,
but not incorrect

FIG. 4. *Fixed bin distribution (histogram) for two loci and four Asian subpopulations (used with permission from John Hartmann): the boundaries of the* 30 *bins (vertical axis) are determined by the FBI; these bins are not of equal length. Sample sizes (numbers of individuals) for Chinese, Japanese, Korean and Vietnamese are* 103, 125, 93 *and* 215 *for* D4S139 *and* 120, 137, 100 *and* 193 *for* D10S28. *The horizontal axis is the bin number; bins are not of equal length.*

DATA VISUALIZATION

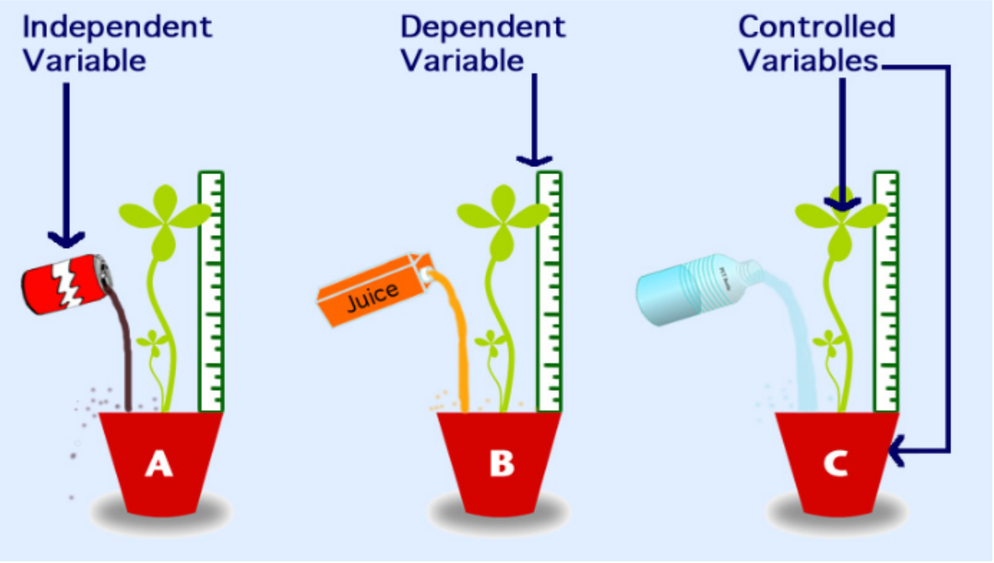


FROM DATA TO AESTHETICS

Data and variables

– Quantitative versus qualitative

– Variables: dependent v. independent



Types of data

– QT: **numerical: continuous**	scale: continuous
– example:	*1.0, 1.1, 1.2, 1.3, 1.4, 1.5, 1.6, 1.7, ...*
– QT: **numerical: discrete**	scale: discrete
– example:	*16C, 17C, 18C, 19C*
– QL: **categorical: ordinal**	scale: discrete
– example:	*LC, LMC, UMC, UC*
– QL: **categorical: nominal**	scale: discrete
– example:	*male, female*
– text	no scale (or: discrete)
– date/time	scale: continuous or discrete

Data and basic operations

- QT: numerical: continuous **??**
  - example: *1.0, 1.1, 1.2, 1.3, 1.4, 1.5, 1.6, 1.7, ...*
- QT: numerical: discrete **??**
  - example: century: *16, 17, 18, 19*
- QL: categorical: ordinal **??**
  - example: *LC, LMC, UMC, UC*
- QL: categorical: nominal **??**
  - example: *male, female*

Data and basic operations

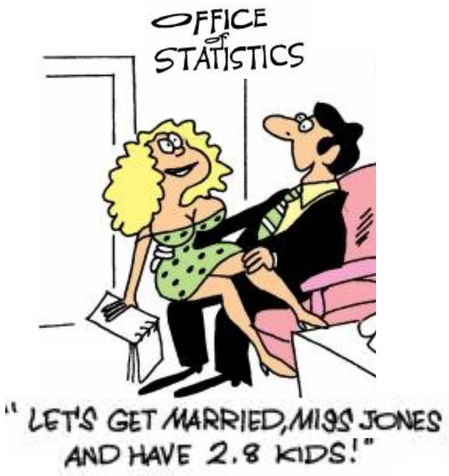
- QT: numerical: continuous — **??**
  - example: *1.0, 1.1, 1.2, 1.3, 1.4, 1.5, 1.6, 1.7, ...*
- QT: numerical: discrete — **??**
  - example: century: *16, 17, 18, 19*
- QL: categorical: ordinal — **??**
  - example: *LC, LMC, UMC, UC*
- QL: categorical: nominal — **=, ≠, #, %, mode**
  - example: *male, female*

Data and basic operations

- QT: numerical: continuous — **??**
  - example: — *1.0, 1.1, 1.2, 1.3, 1.4, 1.5, 1.6, 1.7, …*
- QT: numerical: discrete — **??**
  - example: — century: *16, 17, 18, 19*
- QL: categorical: ordinal — **<, >, median (+ IQR)**
  - example: — *LC, LMC, UMC, UC*
- QL: categorical: nominal — **=, ≠, #, %, mode**
  - example: — *male, female*

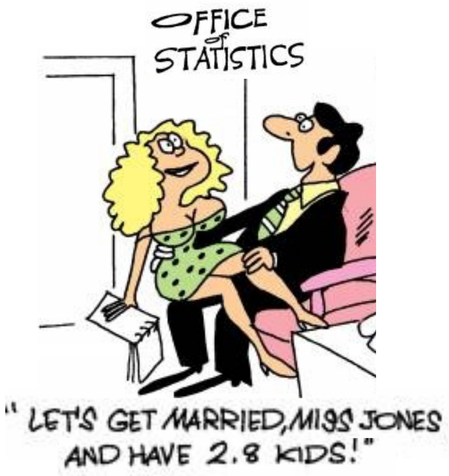
Data and basic operations

- QT: numerical: continuous — **??**
  - example: *1.0, 1.1, 1.2, 1.3, 1.4, 1.5, 1.6, 1.7, …*
- QT: numerical: discrete — **+, –, *, /,** *(mean (+ std. dev.))*
  - example: century: *16, 17, 18, 19*
- QL: categorical: ordinal — **<, >, median (+ IQR)**
  - example: *LC, LMC, UMC, UC*
- QL: categorical: nominal — **=, ≠, #, %, mode**
  - example: *male, female*



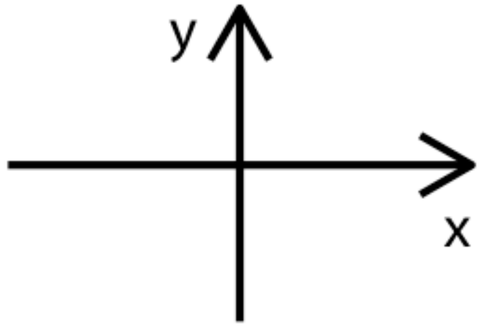
Data and basic operations

- QT: numerical: continuous — **+, –, *, /, mean (+ std. dev.)**
  - example: — *1.0, 1.1, 1.2, 1.3, 1.4, 1.5, 1.6, 1.7, …*
- QT: numerical: discrete — **+, –, *, /,** *(mean (+ std. dev.))*
  - example: — century: *16, 17, 18, 19*
- QL: categorical: ordinal — **<, >, median (+ IQR)**
  - example: — *LC, LMC, UMC, UC*
- QL: categorical: nominal — **=, ≠, #, %, mode**
  - example: — *male, female*

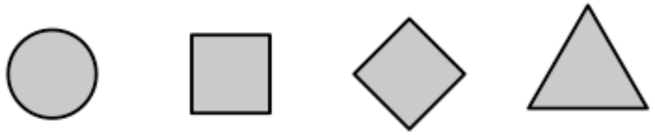


Aesthetics

position



shape

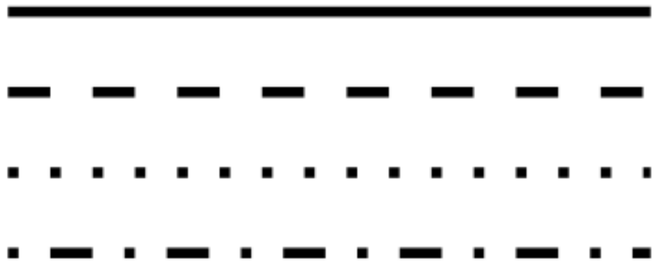


size

color

line width

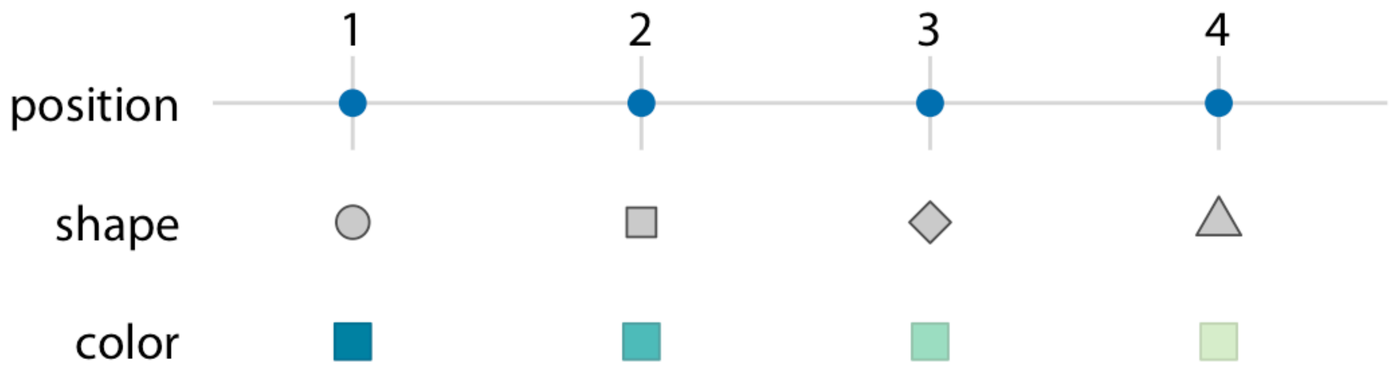
line type



(Wilke 2019)

Mapping data onto aesthetics

- fixed, intuitive and unambiguous **scales** help us map **data values** to **aesthetics** on a **one-to-one** basis *
- e.g. axes, legend, ...



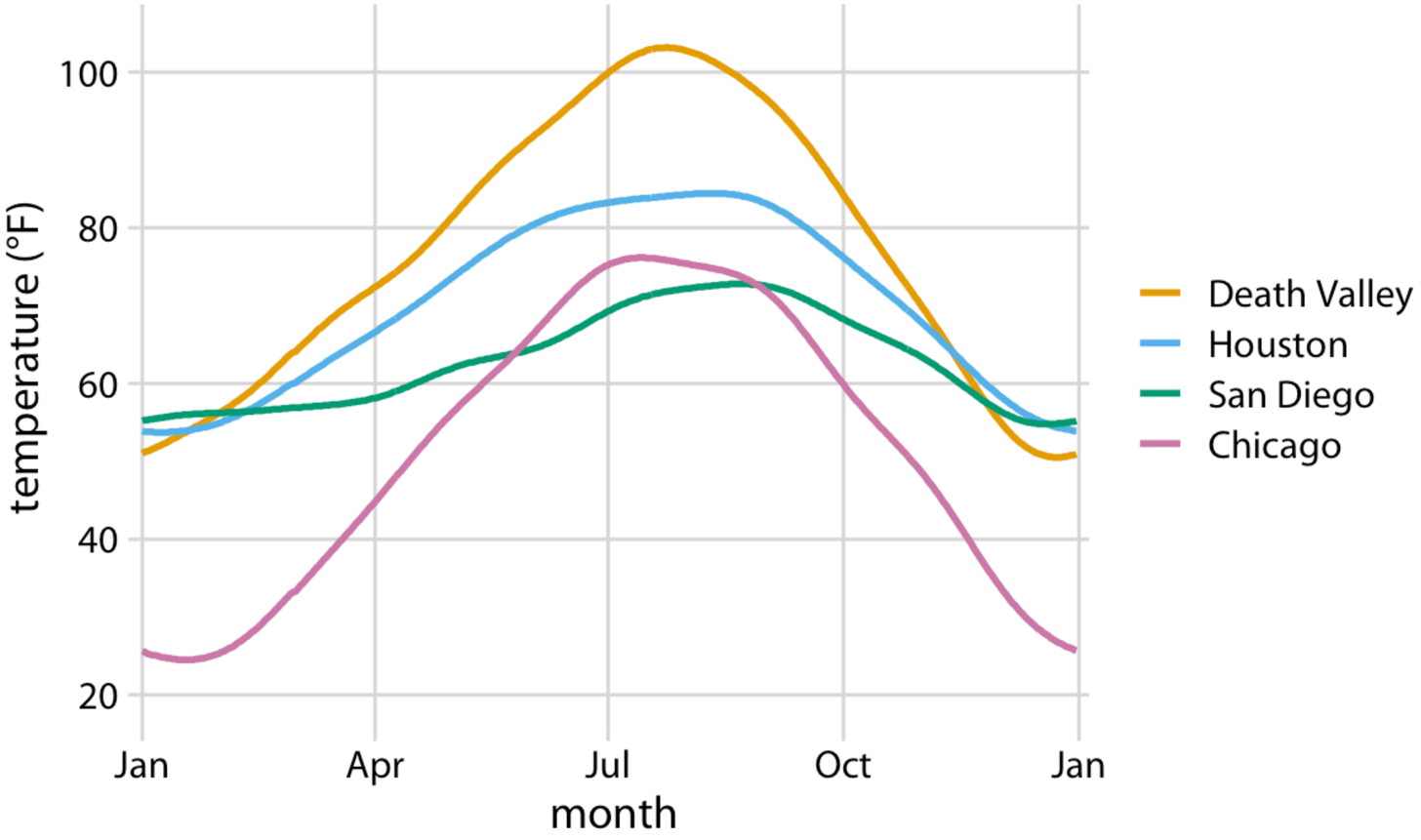
(Wilke 2019)

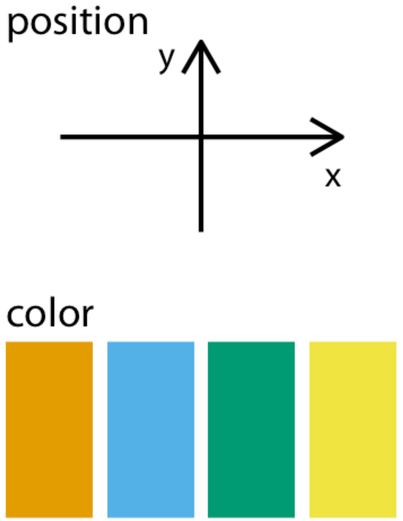
(*) **Tufte** (2001: 77):

"The number of **information-carrying** (variable) **dimensions depicted** should not exceed the number of **dimensions in the data**"

Mapping data onto aesthetics

– **example**: time, temperature, location





mapping:

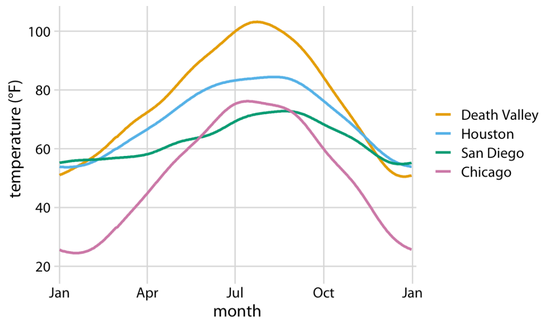
position x = time
position y = temp.
color = location

LINE CHART

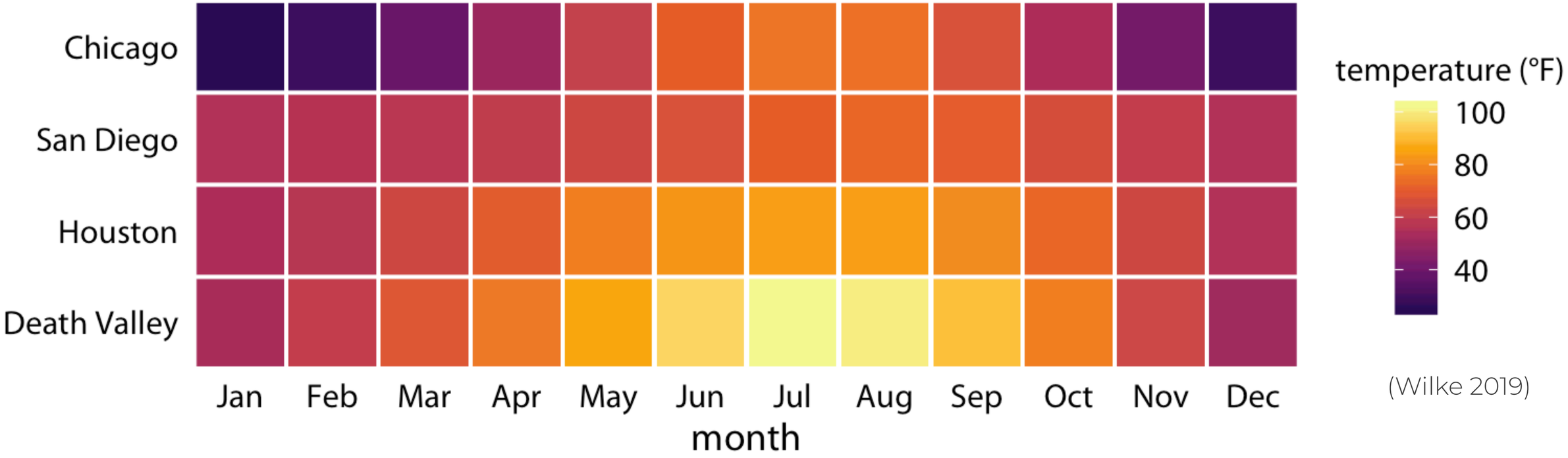
(Wilke 2019)

Mapping data onto aesthetics

– **example**: time, temperature, location



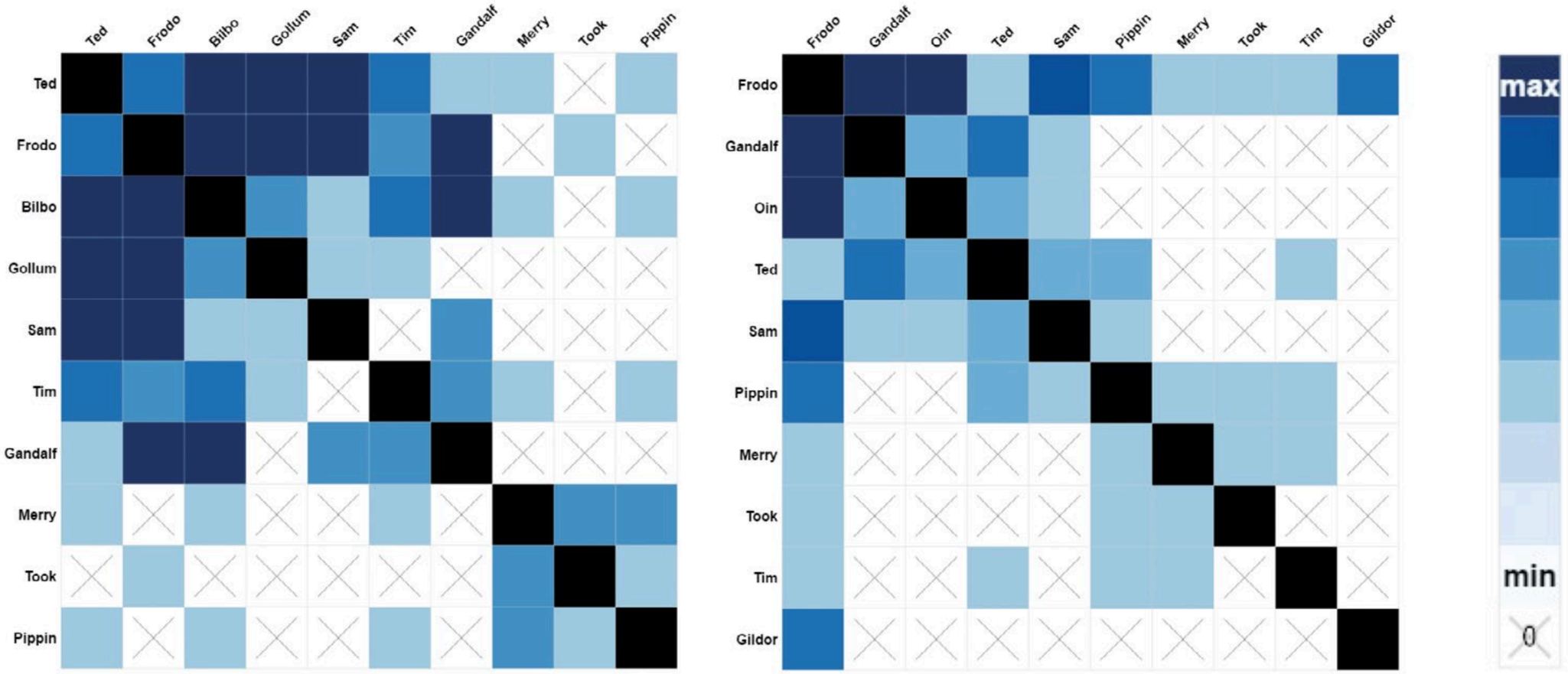
Heat map



(Wilke 2019)

mapping: position x = time | position y = location | color = temperature

Aside: heat maps and related plots

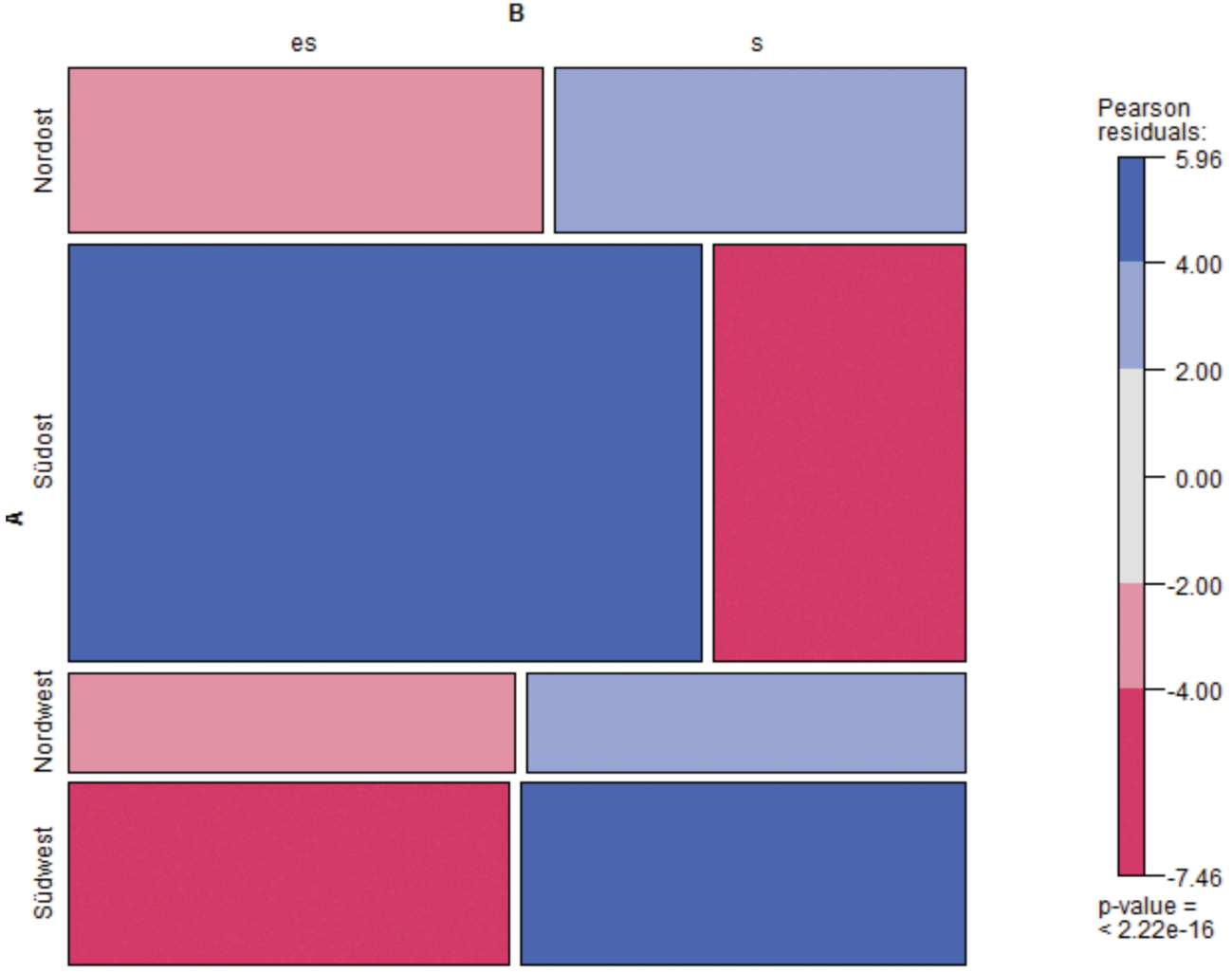


Visualization of the top ten character co-occurrences in third and fourth chapter of Lord of the Rings – The Fellowship of the Ring

(Flekova, Stoffel, Gurevych & Kiem 2018: 210)

ASSOCIATION PLOT

Example: Wolfer & Hansen-Morath (2018: 240):
Visualisierung sprachlicher Daten mit R

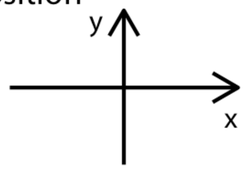


Mosaic plot for Gentive allomorphs -es and -s
(data: Fürbacher 2015)

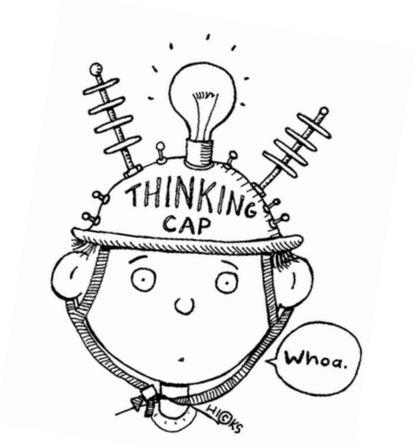
Mosaic plot

Aesthetics?

position

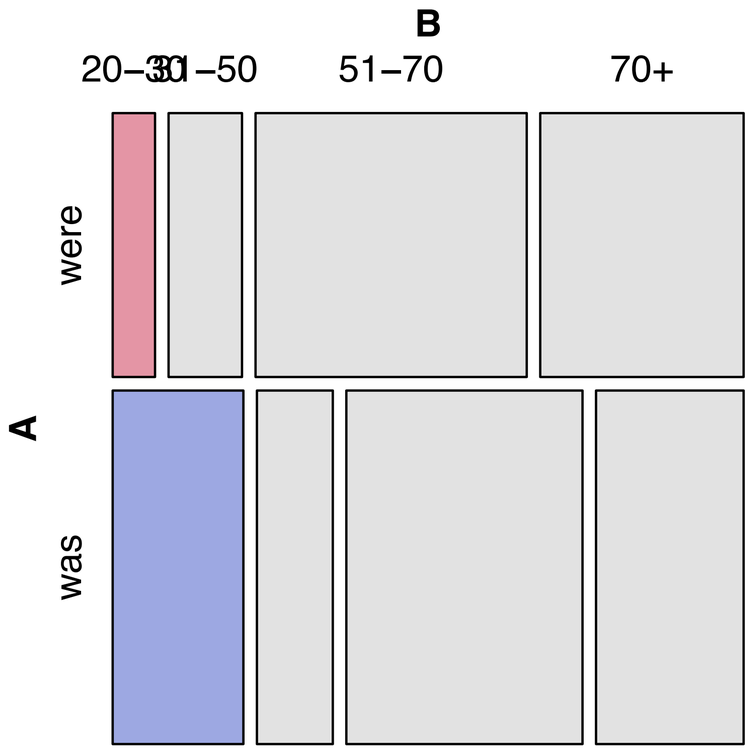


color

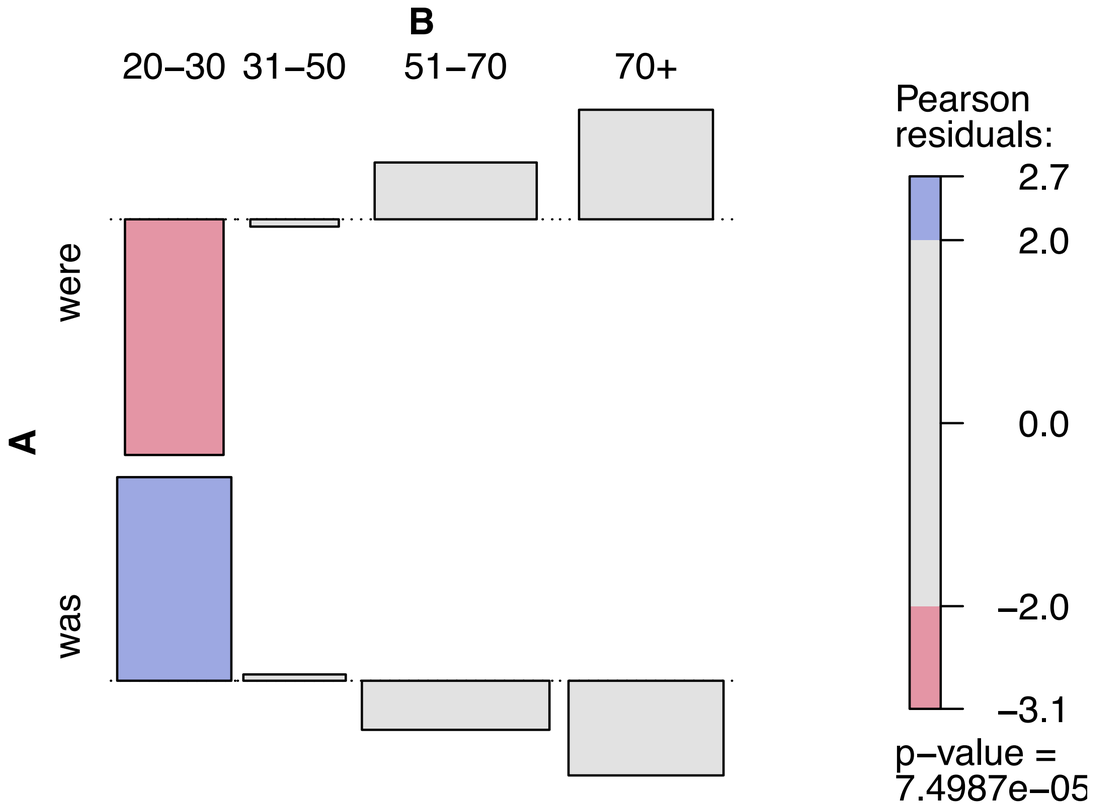


Mapping data onto aesthetics

Was/were variation per age group



Was/were variation per age group

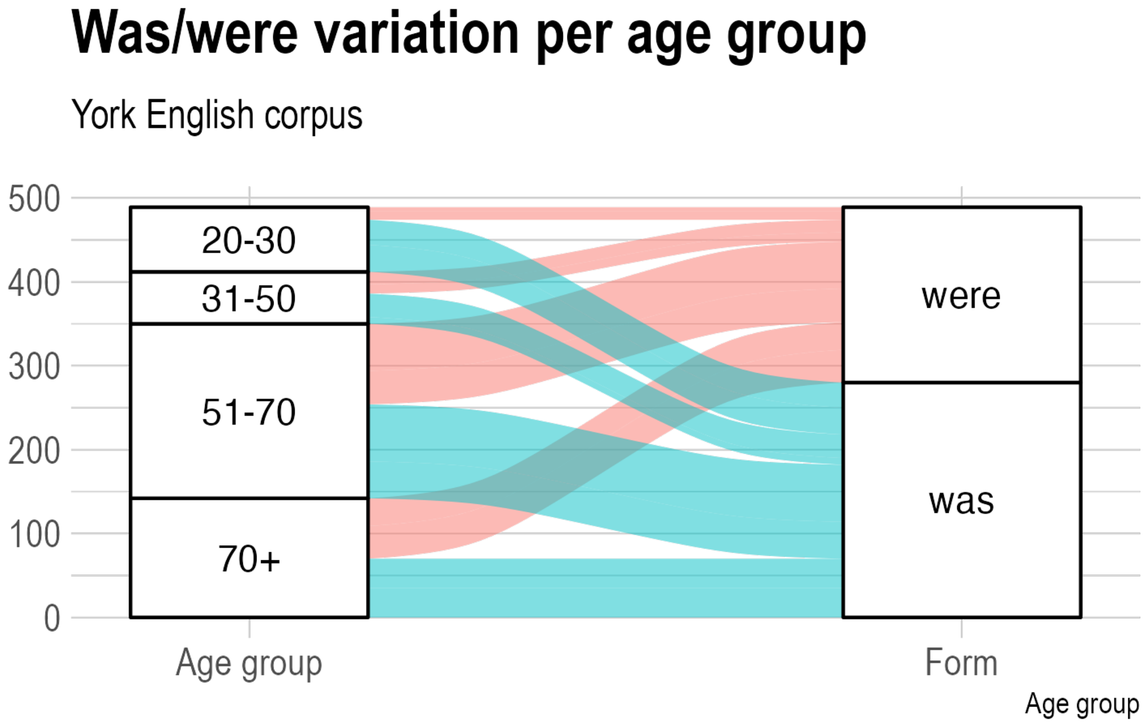


(Data: Tagliamonte & Baayen 2012)

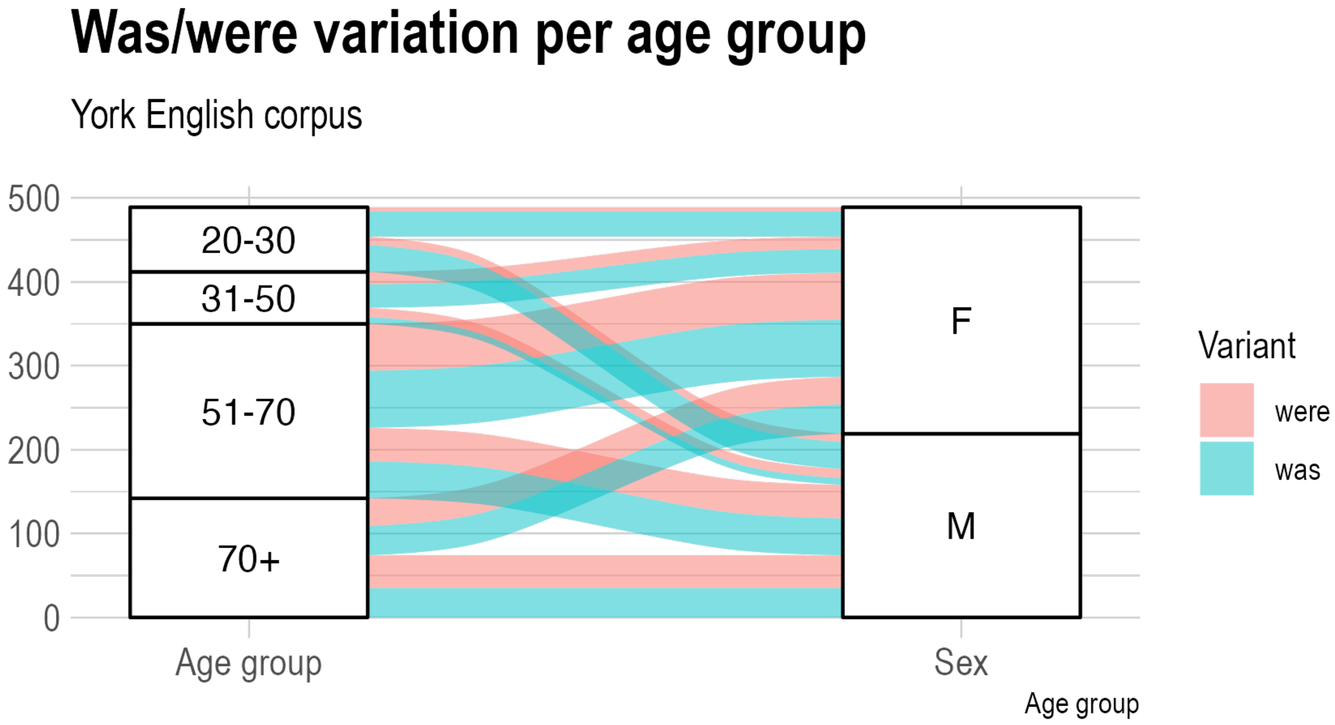
(Data: Tagliamonte & Baayen 2012)

Mapping data onto aesthetics

Sankey diagrams / alluvial plots

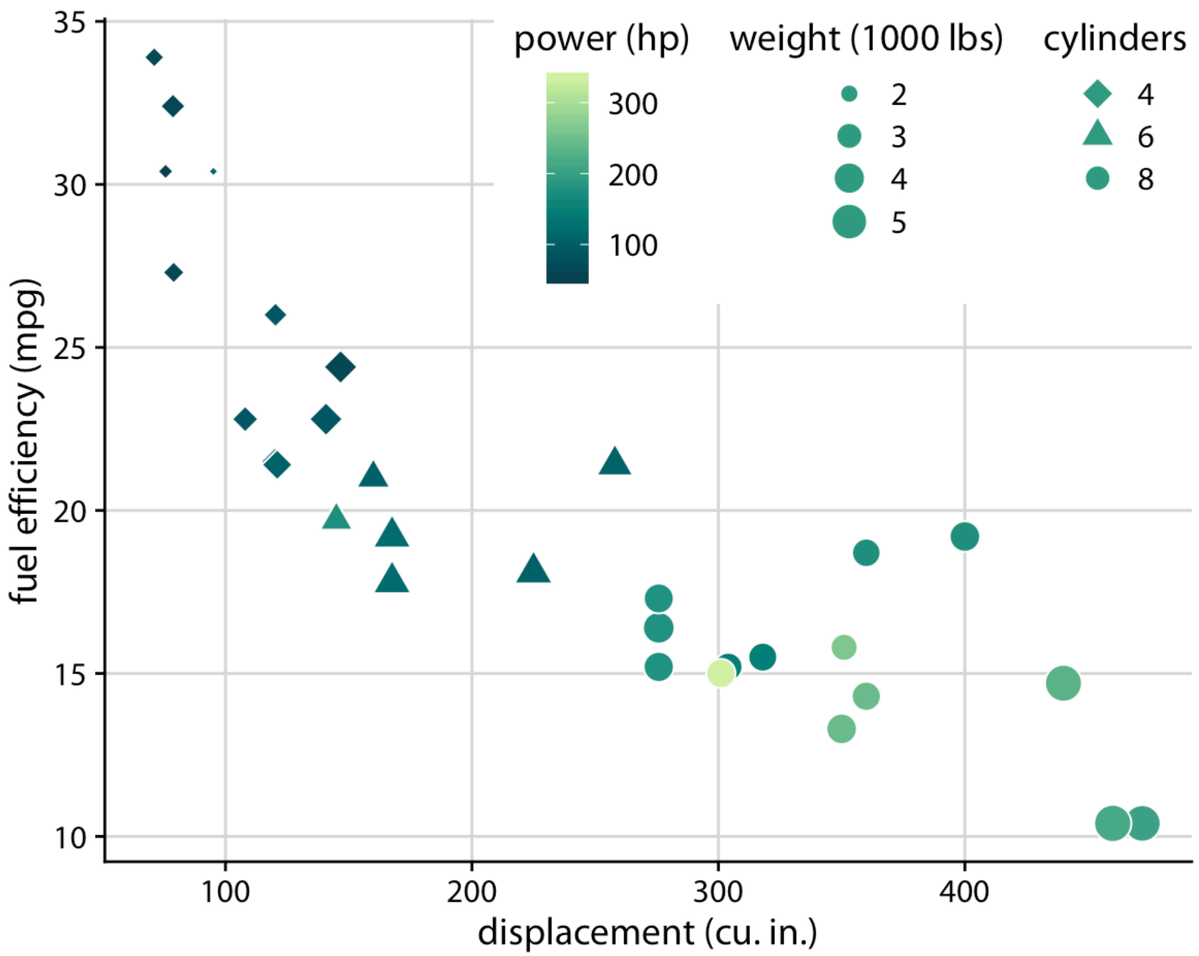


(Data: Tagliamonte & Baayen 2012)

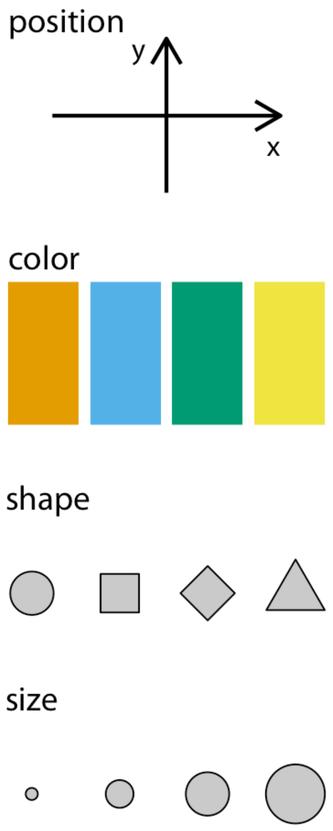


(Data: Tagliamonte & Baayen 2012)

Mapping data onto aesthetics

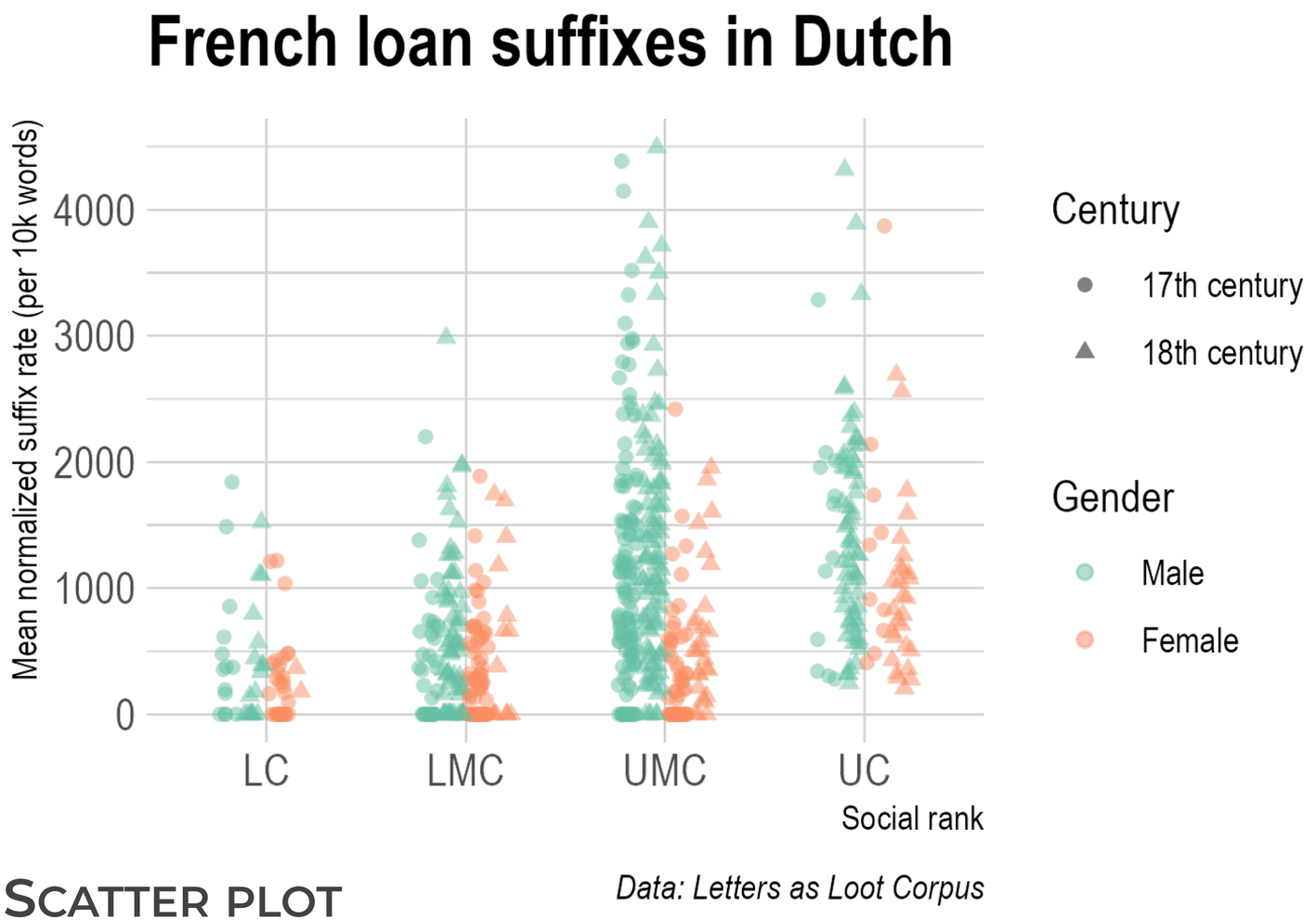


Five aesthetics:



(Wilke 2019)

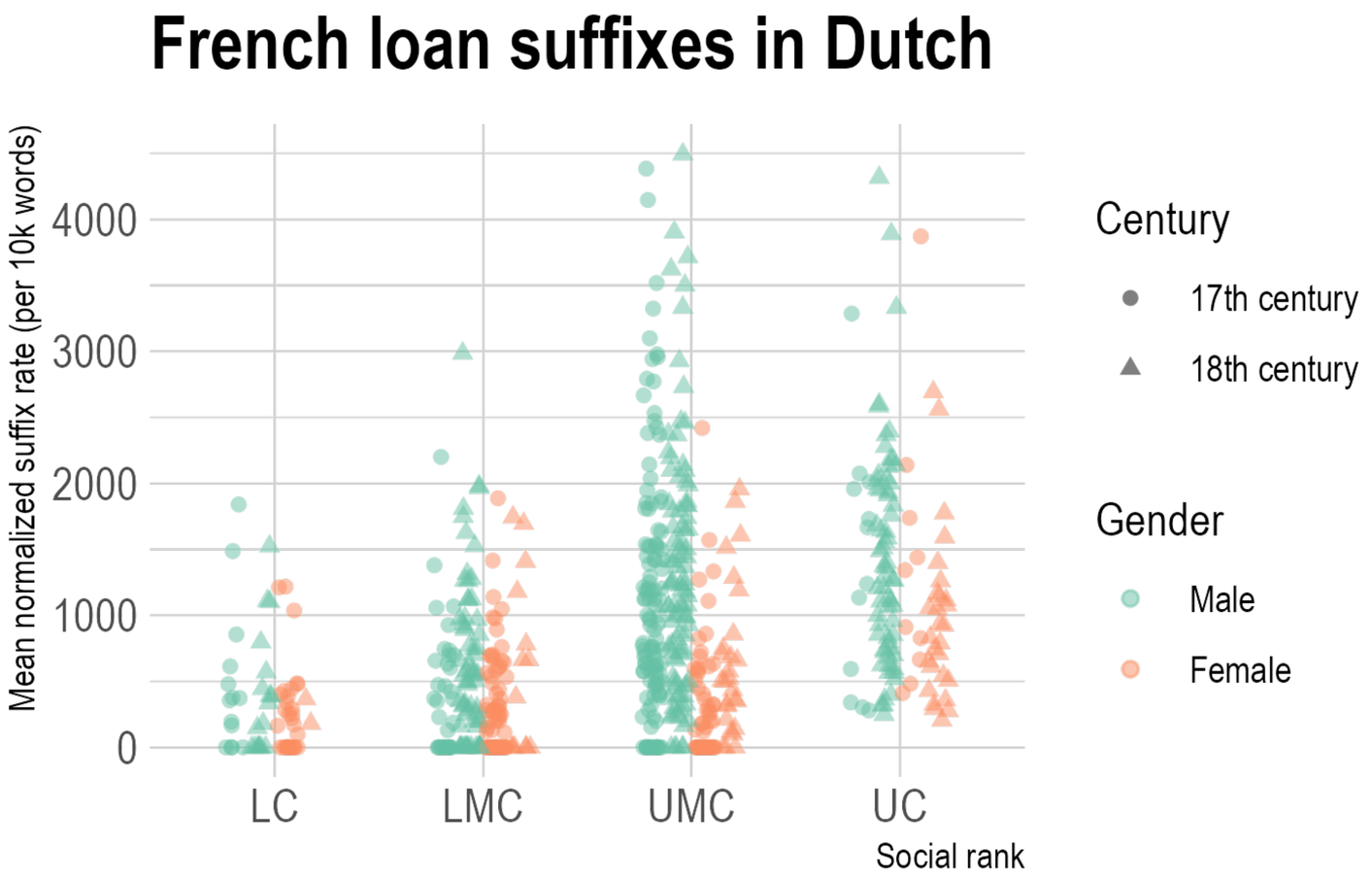
Mapping data onto aesthetics



SCATTER PLOT

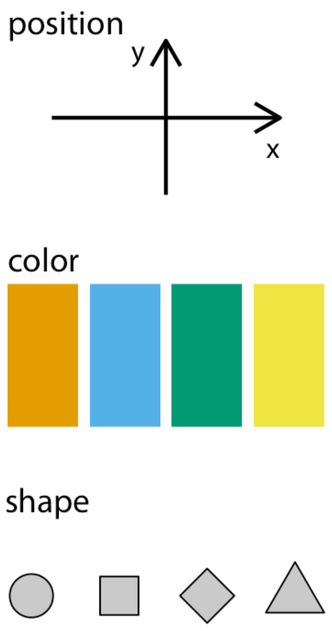
Aesthetics?

Mapping data onto aesthetics



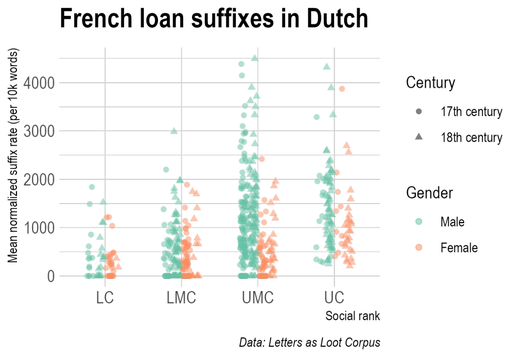
Data: Letters as Loot Corpus

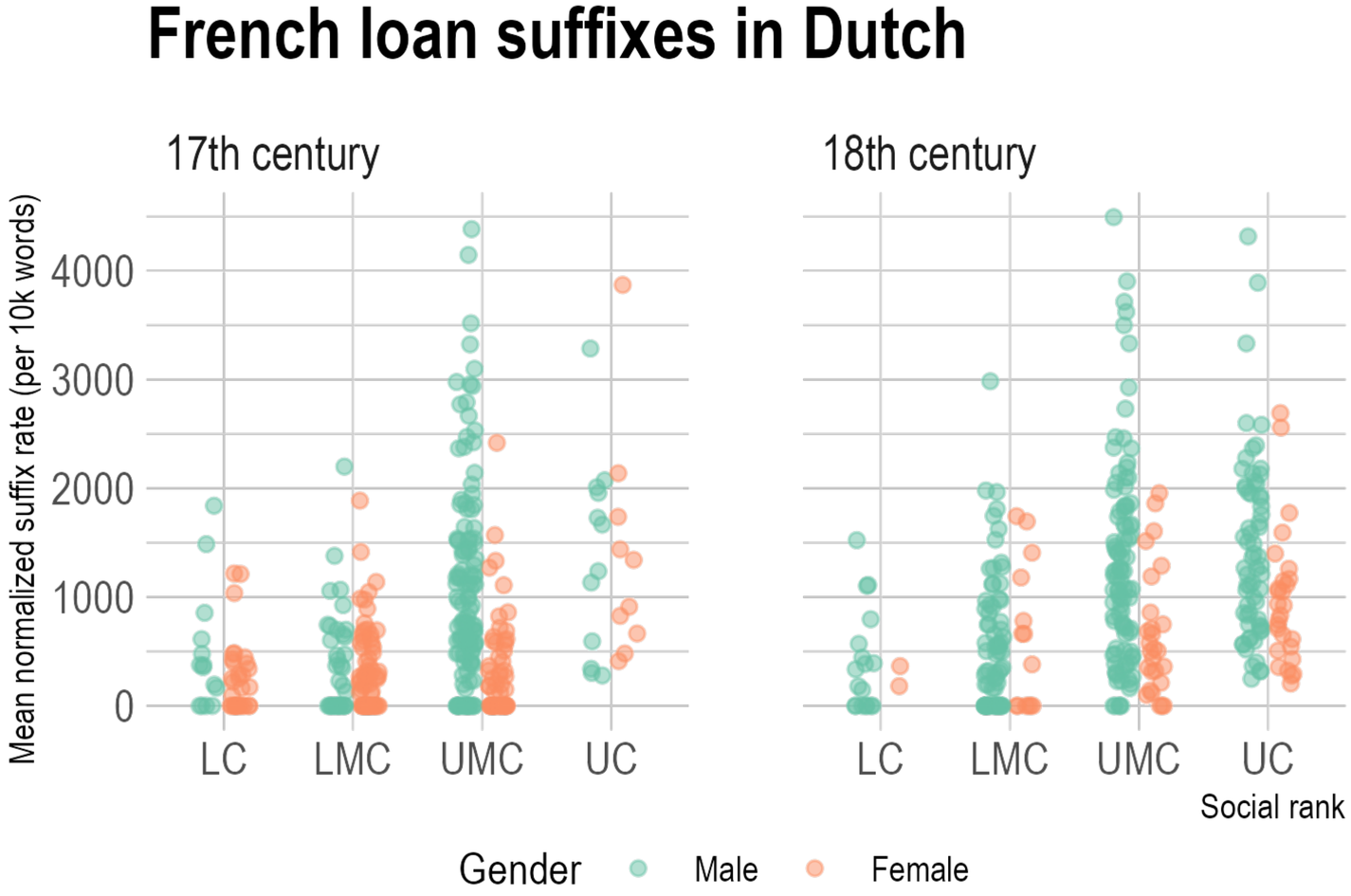
Aesthetics:



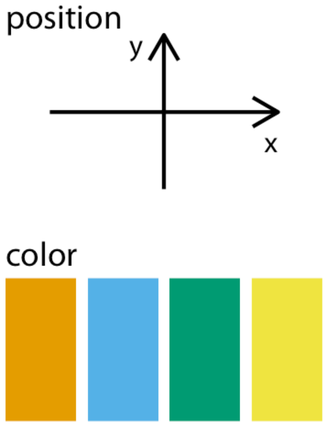
SCATTER PLOT

Mapping data onto aesthetics



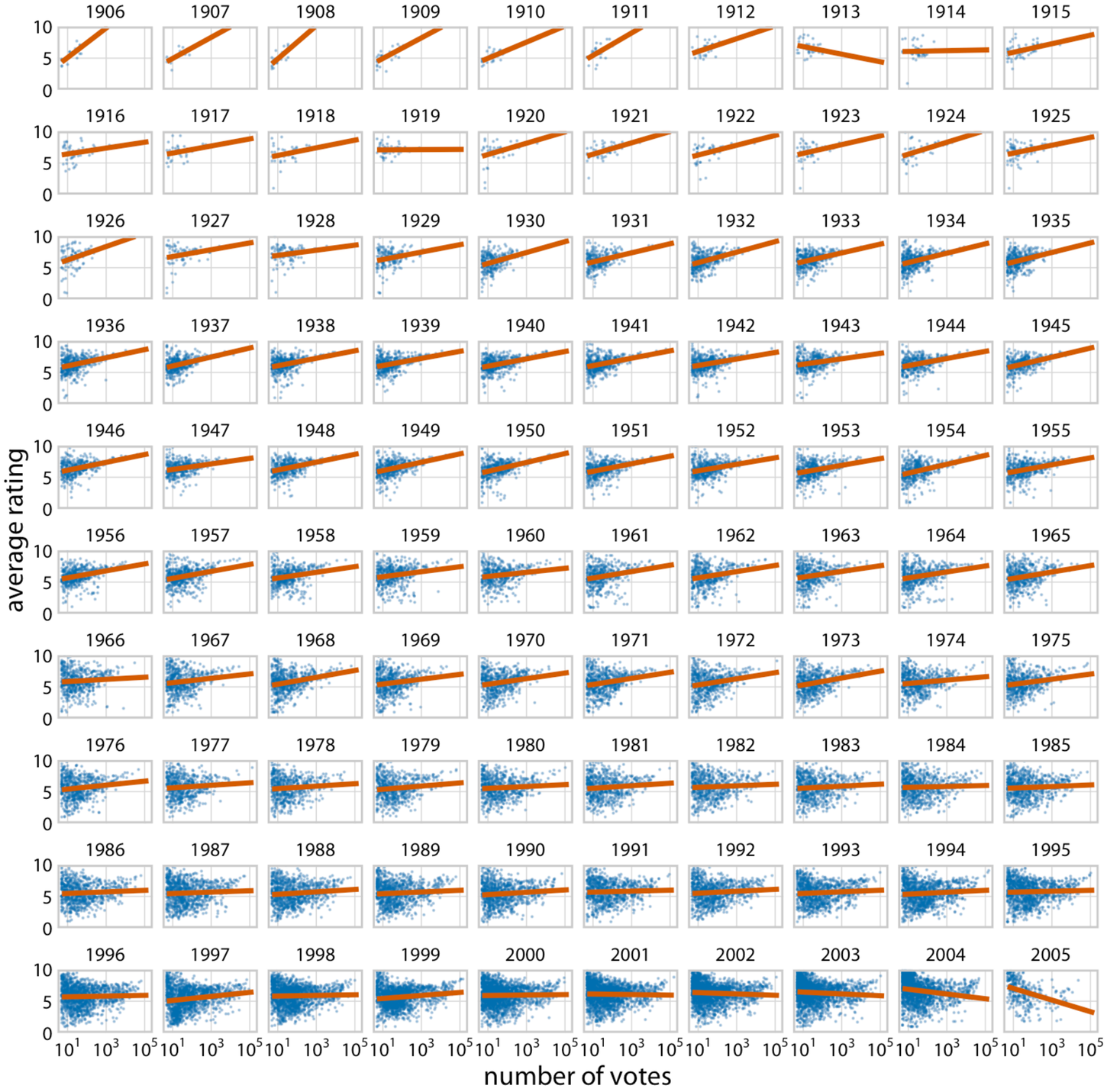


Aesthetics (4 → 3):



SCATTER PLOT

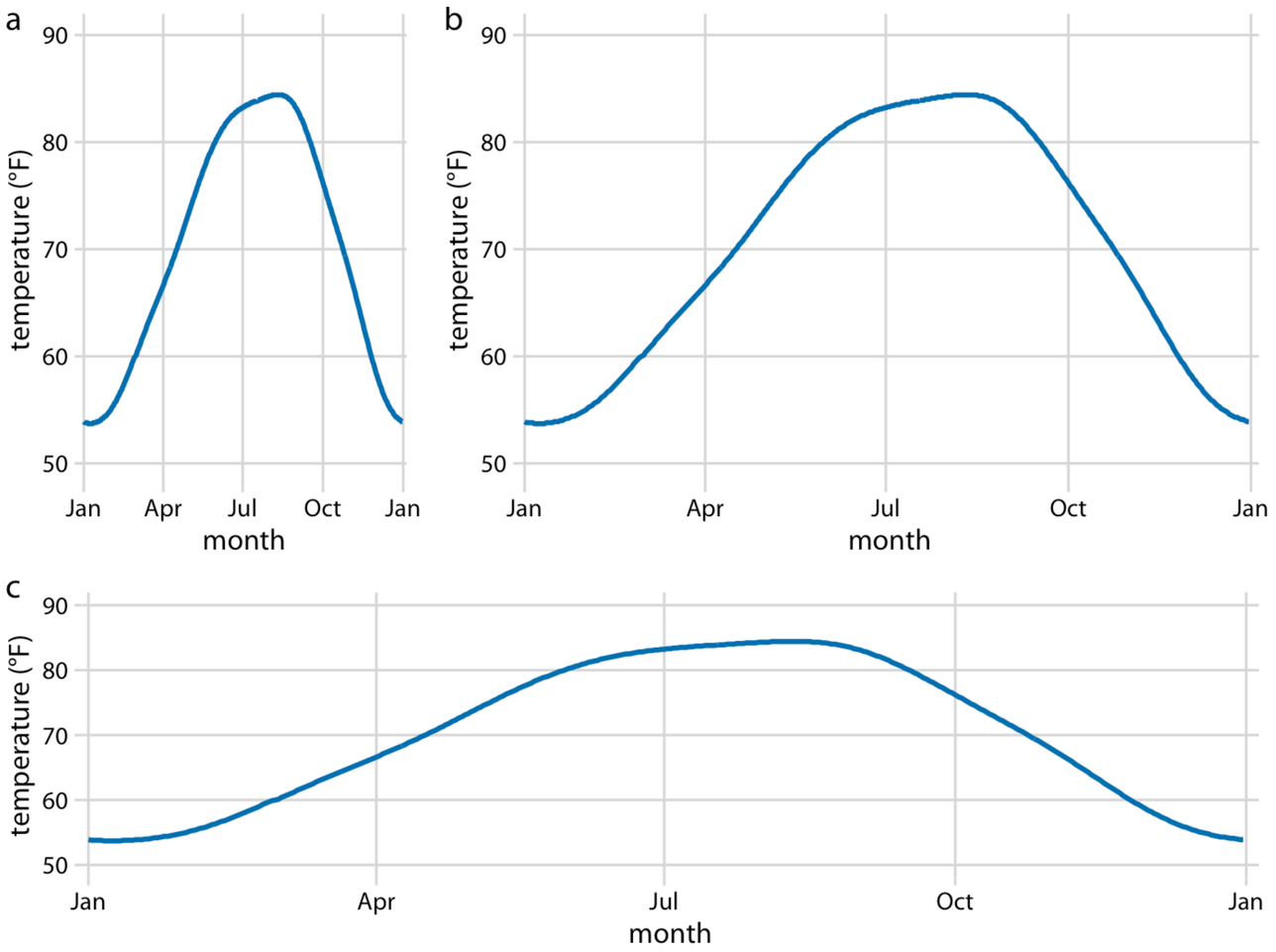
Data: Letters as Loot Corpus



(Wilke 2019; data: Internet Movie Database)

Coordinate systems

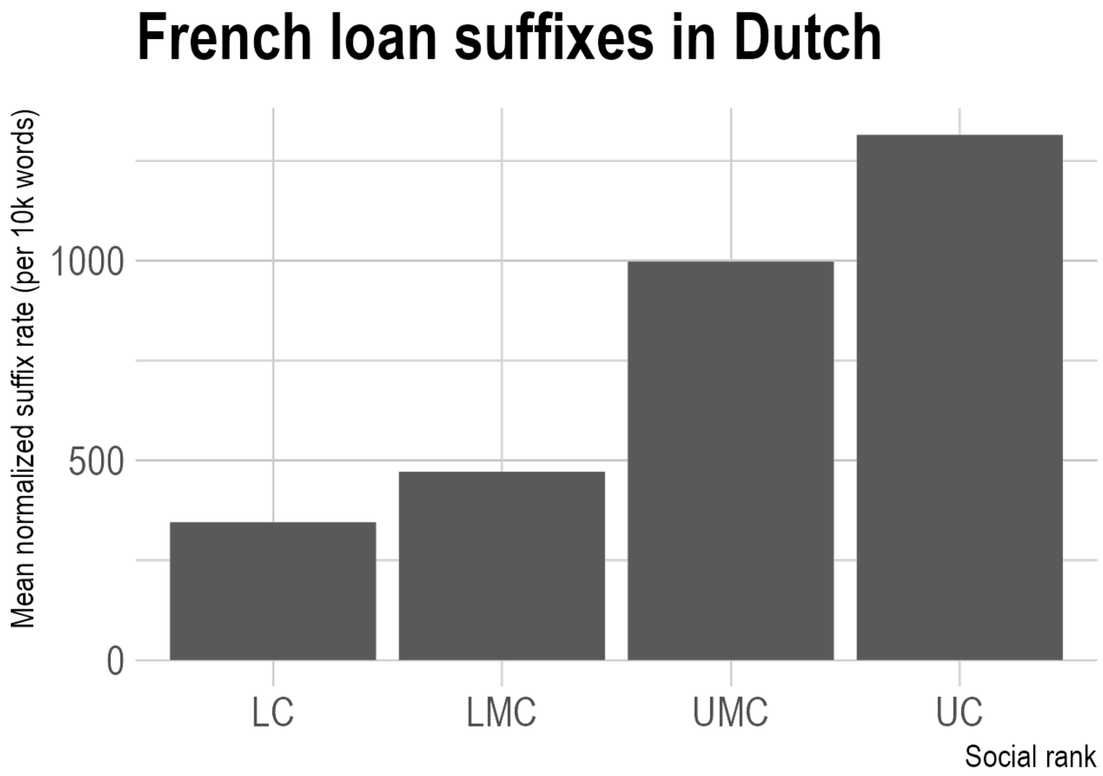
- cartesian coordinate system: horizontal **x** and vertical **y** axis
  - variable aspect ratio:
    ~ story to tell
  - example:
    Daily temperature normals
    for Houston, TX



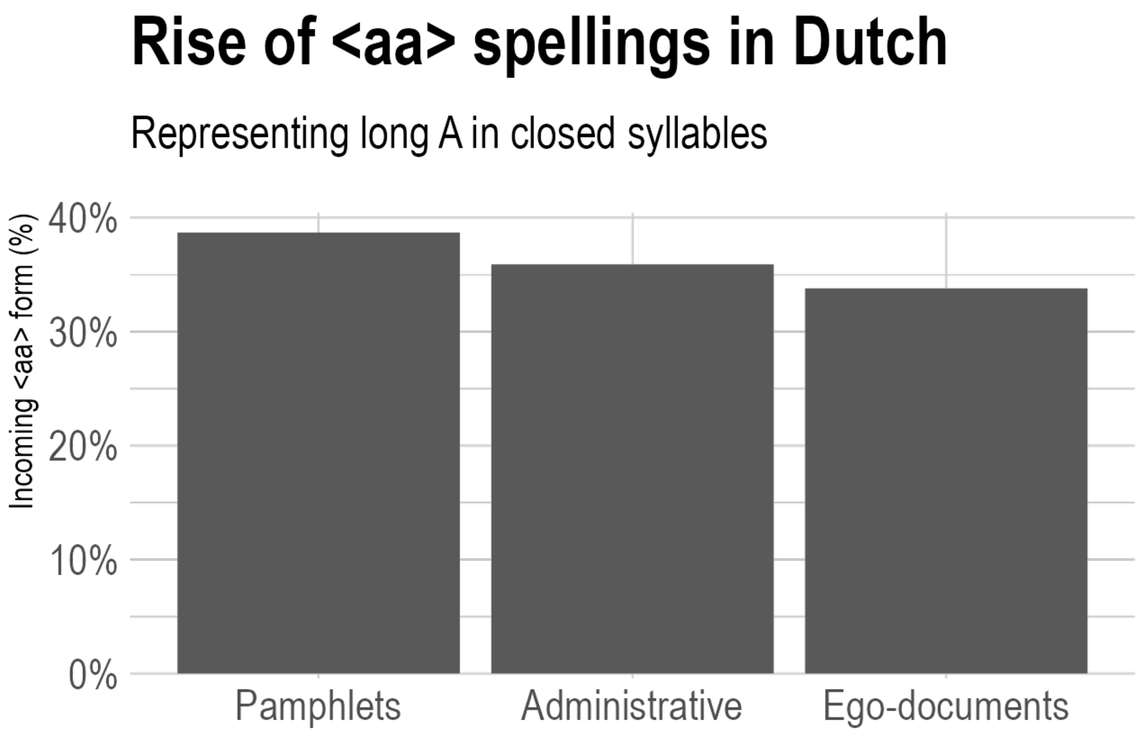
(Wilke 2019; data: NOAA)

Coordinate systems

- cartesian coordinate system: horizontal **x** and vertical **y** axis
  - barplots: x-axis = categorical
  - order: category-based or value-based



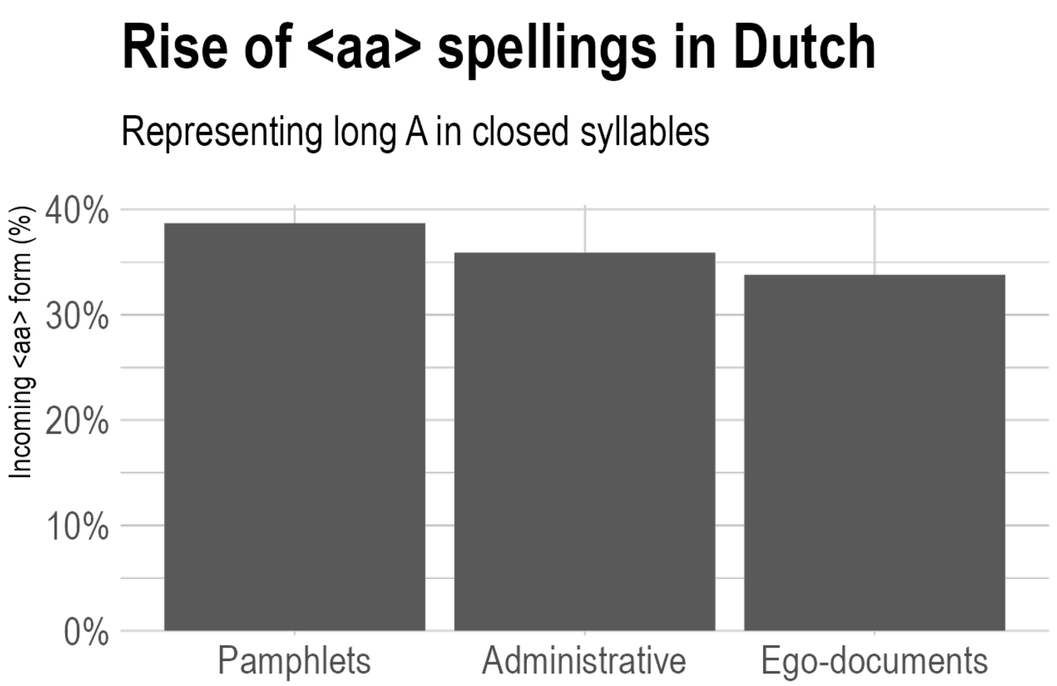
Data: Letters as Loot Corpus



Data: Van de Voorde (2021)

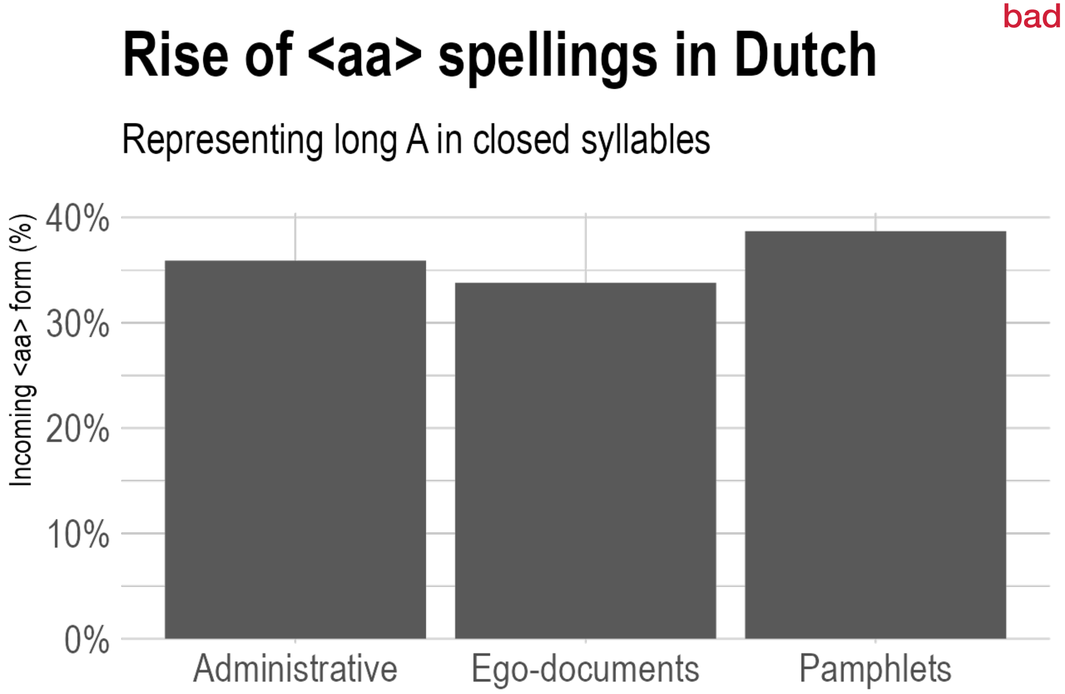
Coordinate systems

- cartesian coordinate system: horizontal **x** and vertical **y** axis
  - barplots: x-axis = categorical
  - order: category-based or value-based



Data: Van de Voorde (2021)

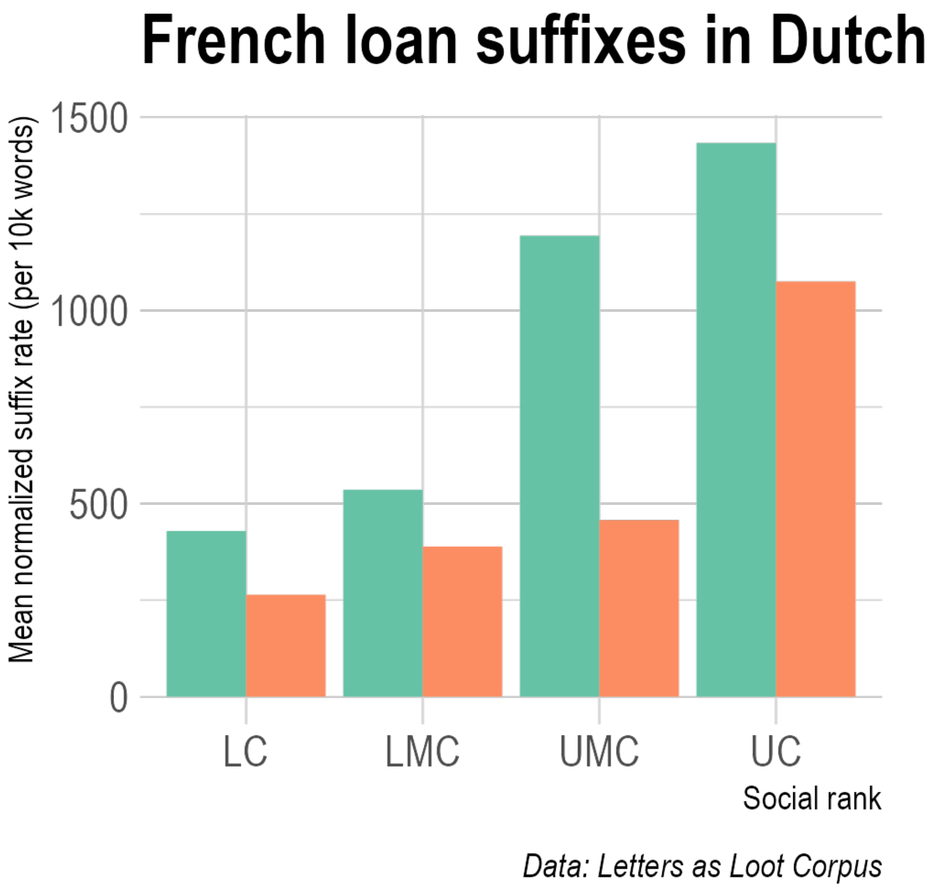
bad

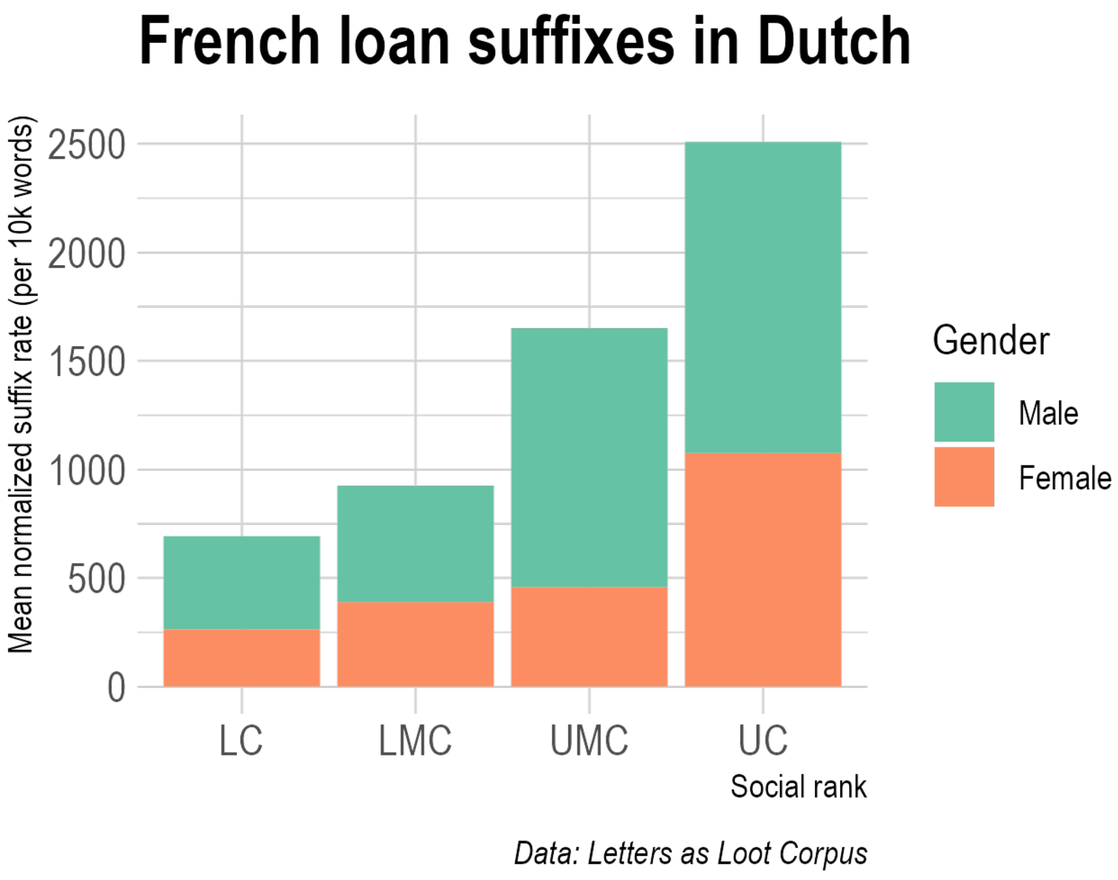


Data: Van de Voorde (2021)

Coordinate systems

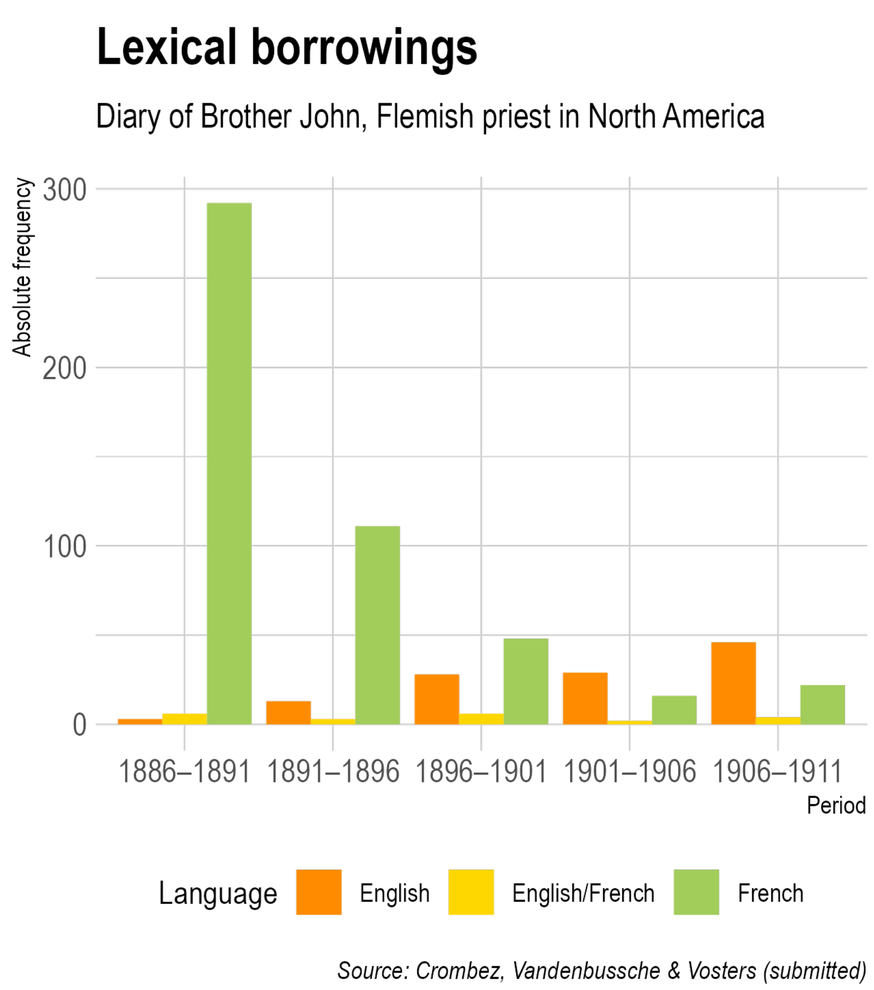
– **aside 1:** normal, dodged and stacked (percent) barplots

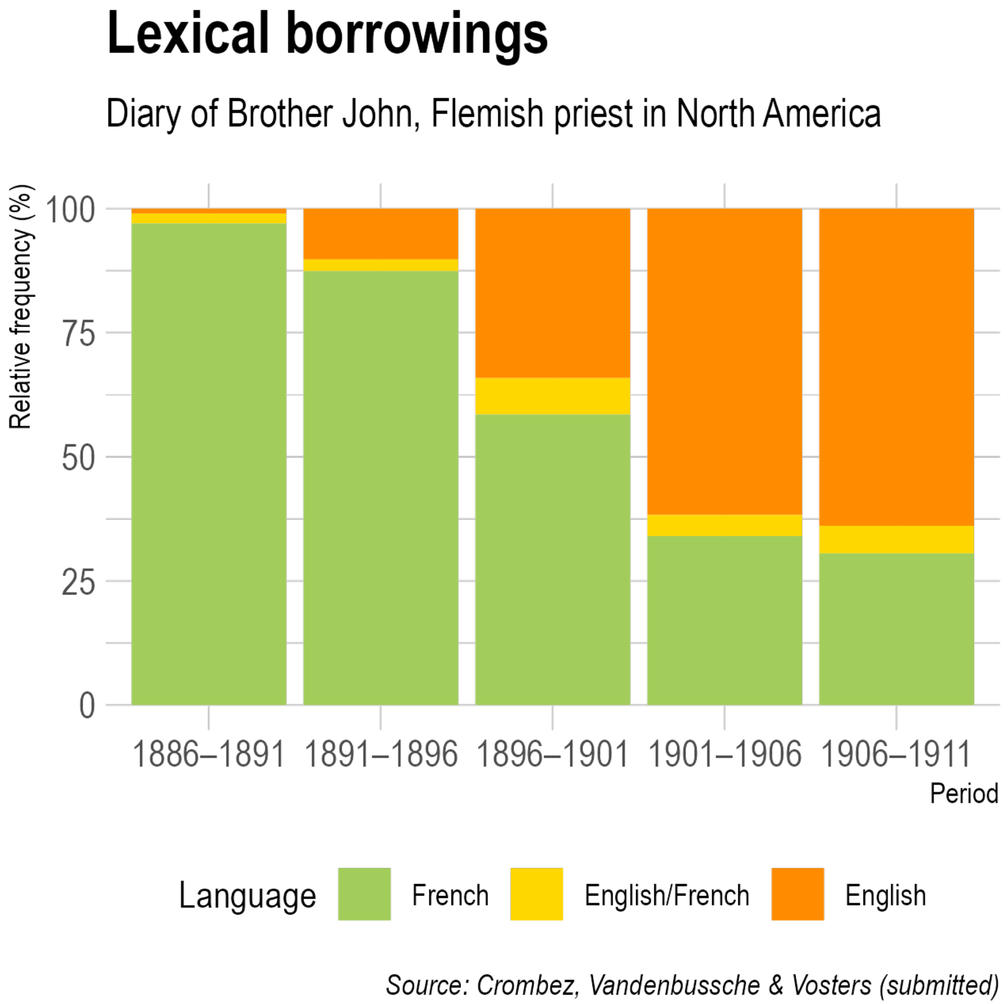




Coordinate systems

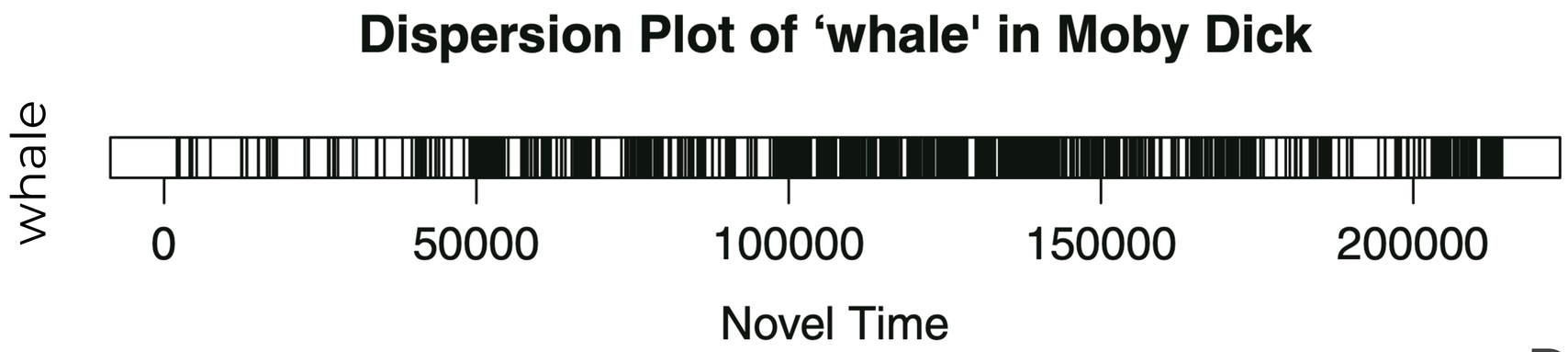
– **aside 1:** normal, dodged and stacked (percent) barplots



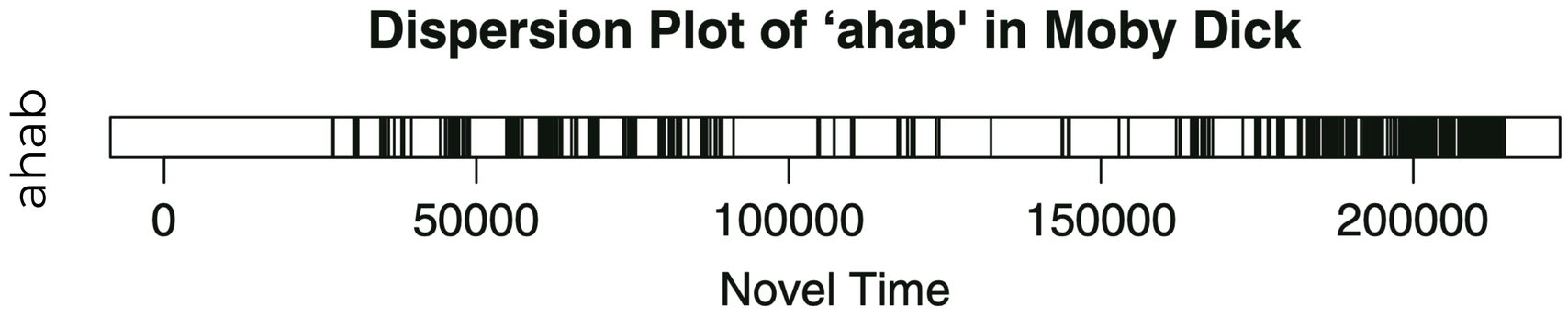


– **aside 2:** representing position

Example 1 : Jockers (2014: 32) – Text Analysis with R for Students of Literature



DISPERSION PLOT



Dispersion plots for “whale” and “ahab” in Moby Dick

Example 2: Flekova, Stoffel, Gurevych & Kiem (2018: 197):
Content-based Analysis and Visualization of Story Complexity

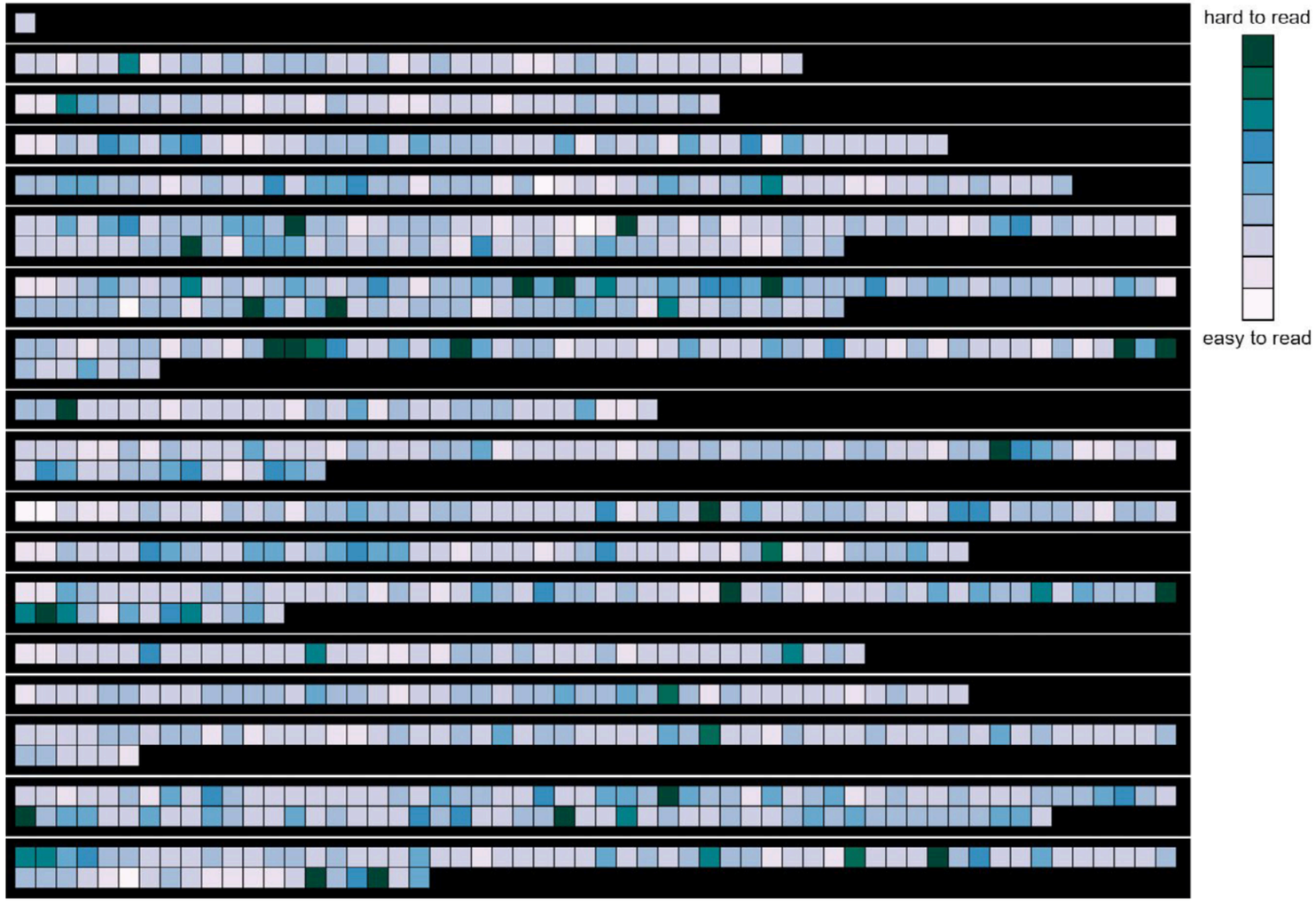


Figure 2: Flesch Reading Ease score per paragraph of *Harry Potter and the Sorcerer's Stone*.

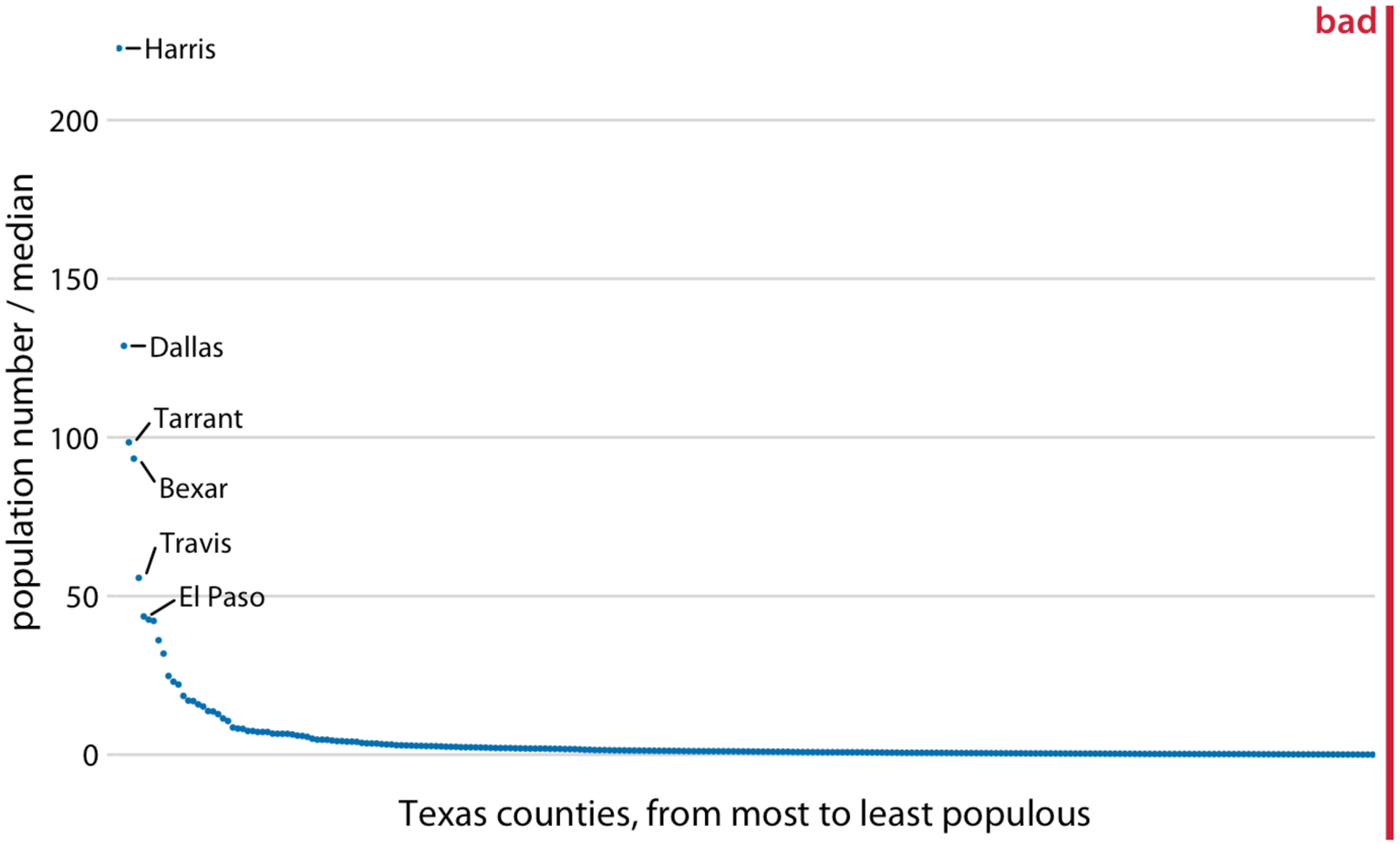
Coordinate systems

- non-linear axes
  - **transformations** of your data (~ currency)
  - logarithmic scale (or: **log scale**):
    - ‘linear in multiplication’: **+1** step on scale = **multiplied by z** (often z = **10**)
  - advisable for **ratio** data (‘natural zero’) and data obtained by **multiplication** or **division** (as multiplication on log scale is like addition on linear scale)
    - linear scale: overemphasizes ratios > 1 and obscures ratios < 1
  - **example**: population size of Texan towns divided by median

(Wilke 2019: data source: NOAA)

Coordinate systems

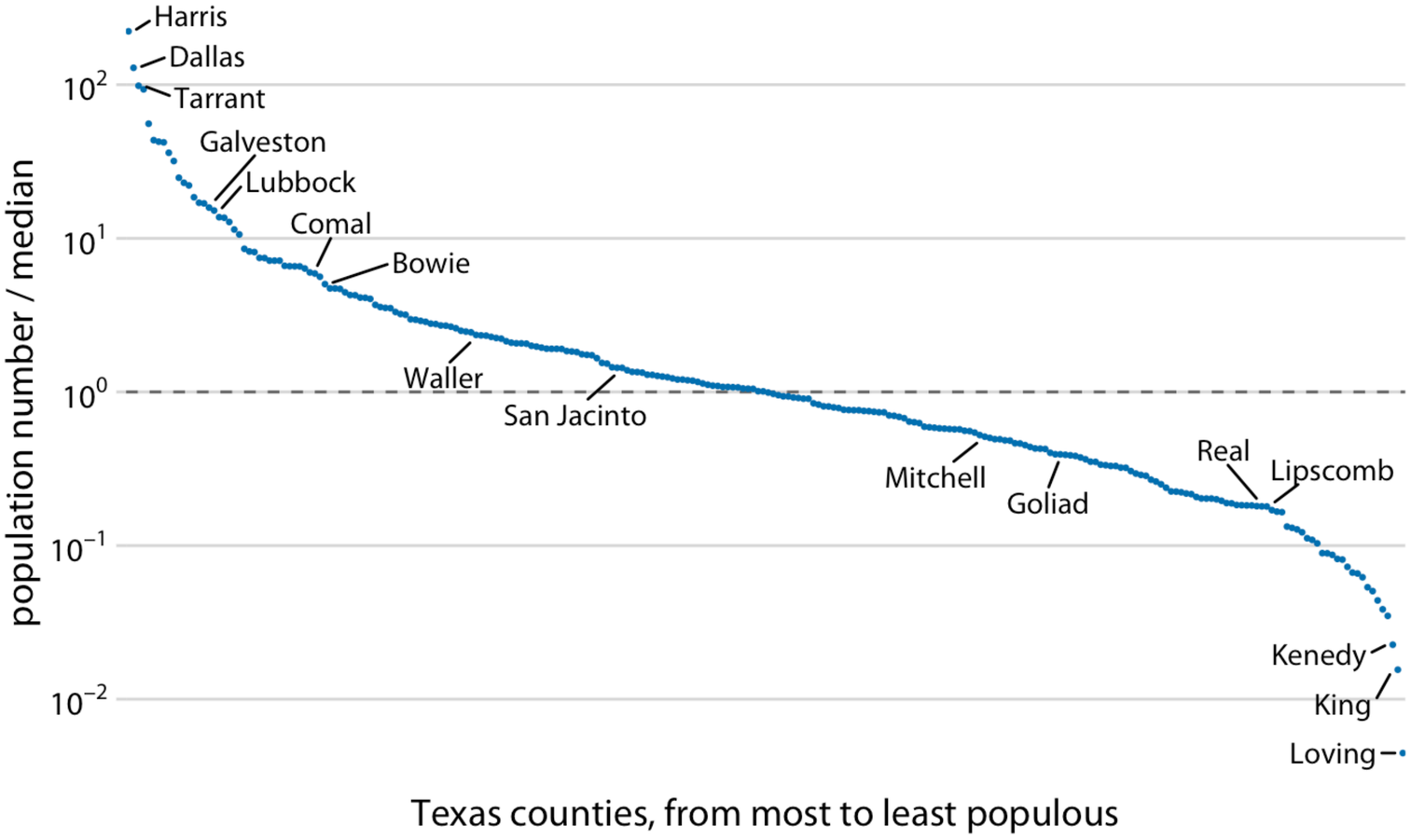
– **example**: population size of Texan towns divided by median



(Wilke 2019; data: 2010 Decennial U.S. Census)

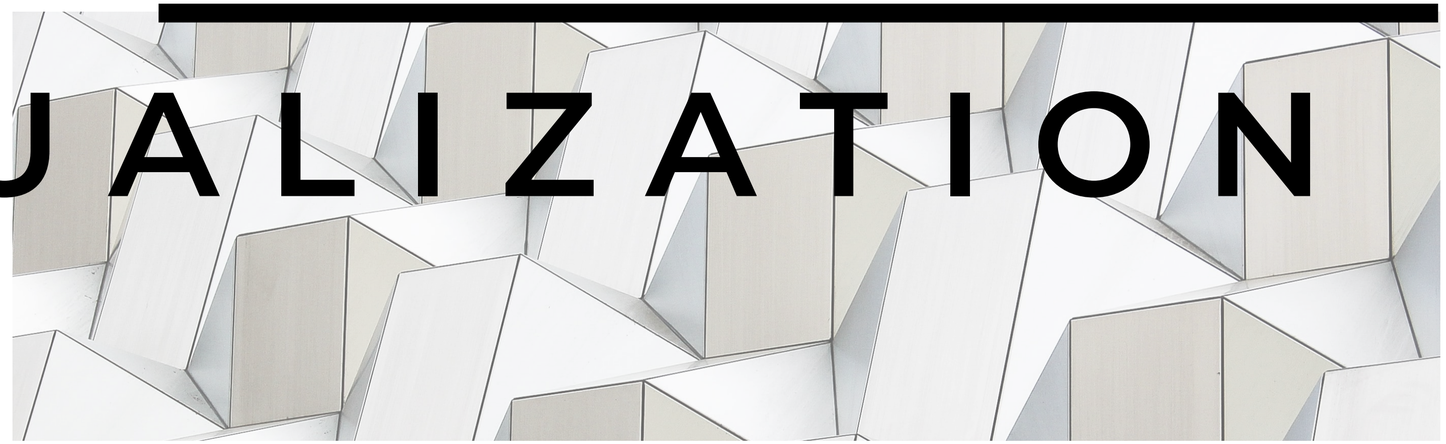
Coordinate systems

– **example**: population size of Texan towns divided by median



(Wilke 2019; data: 2010 Decennial U.S. Census)

DATA VISUALIZATION



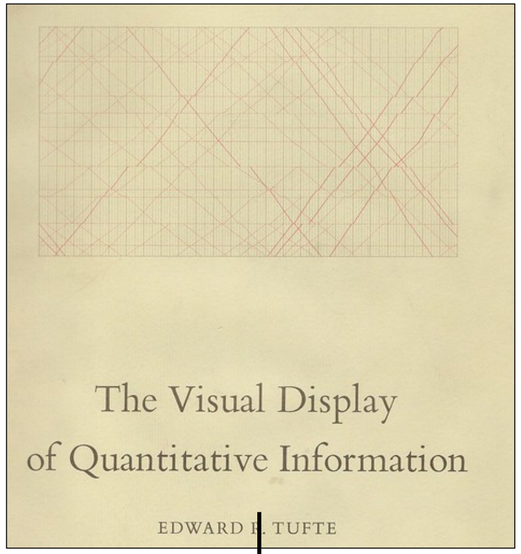
TWO DESIGN PRINCIPLES

Introduction

Two principles, after Tufte (1983)

1. Principle of the maximal data to ink ratio
2. Principle of proportional ink

Note: ink = bars, lines, text, shaded areas, etc.



Tufte (1983, 2001)

The Visual Display of Quantitative Information

⇒ introduce various types of plots, linked to these two principles

Data to ink

1. Principle of the maximal data to ink ratio

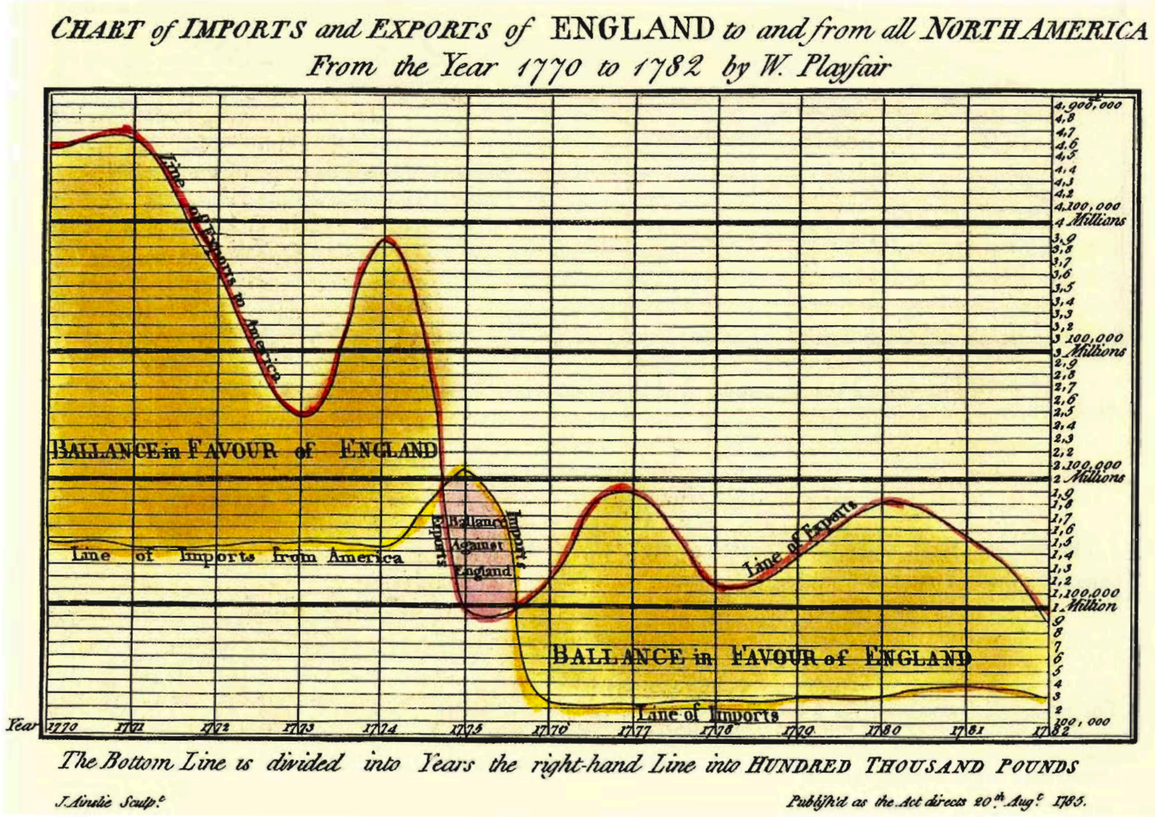
   Tufte (2001: 93):

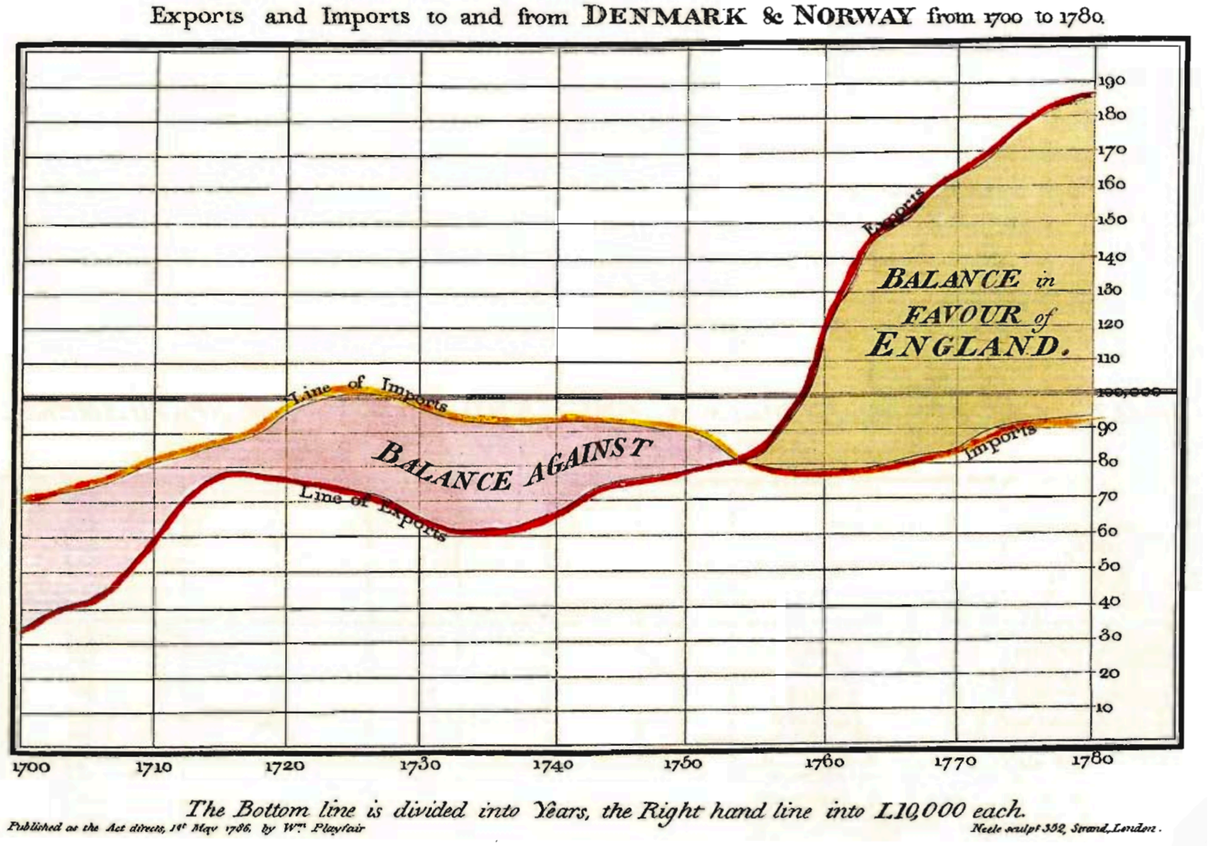
$$\text{Data-ink ratio} = \frac{\text{data-ink}}{\text{total ink used to print the graphic}}$$

   “Above all else, show the data” (Tufte 2001: 105)

Data to ink

1. Principle of the maximal data to ink ratio





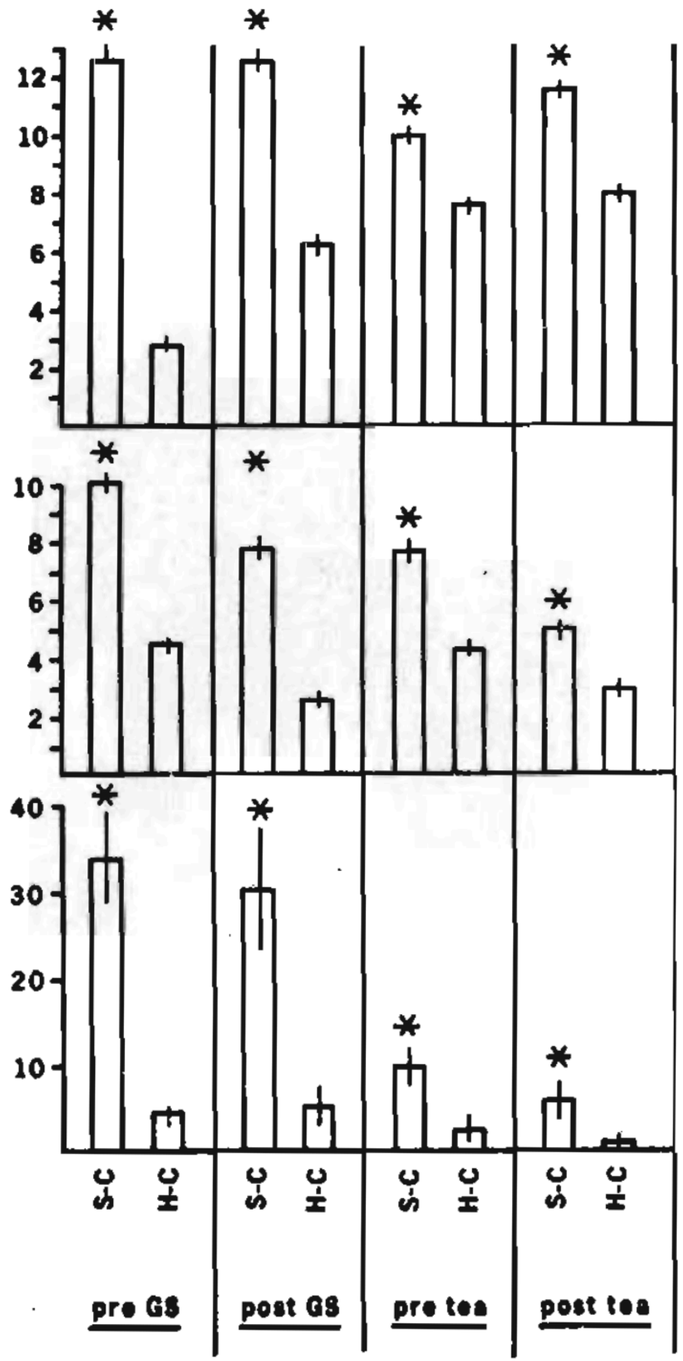
Playfair, The Commercial and Political Atlas (discussed in Tufte 2001: 91-92)

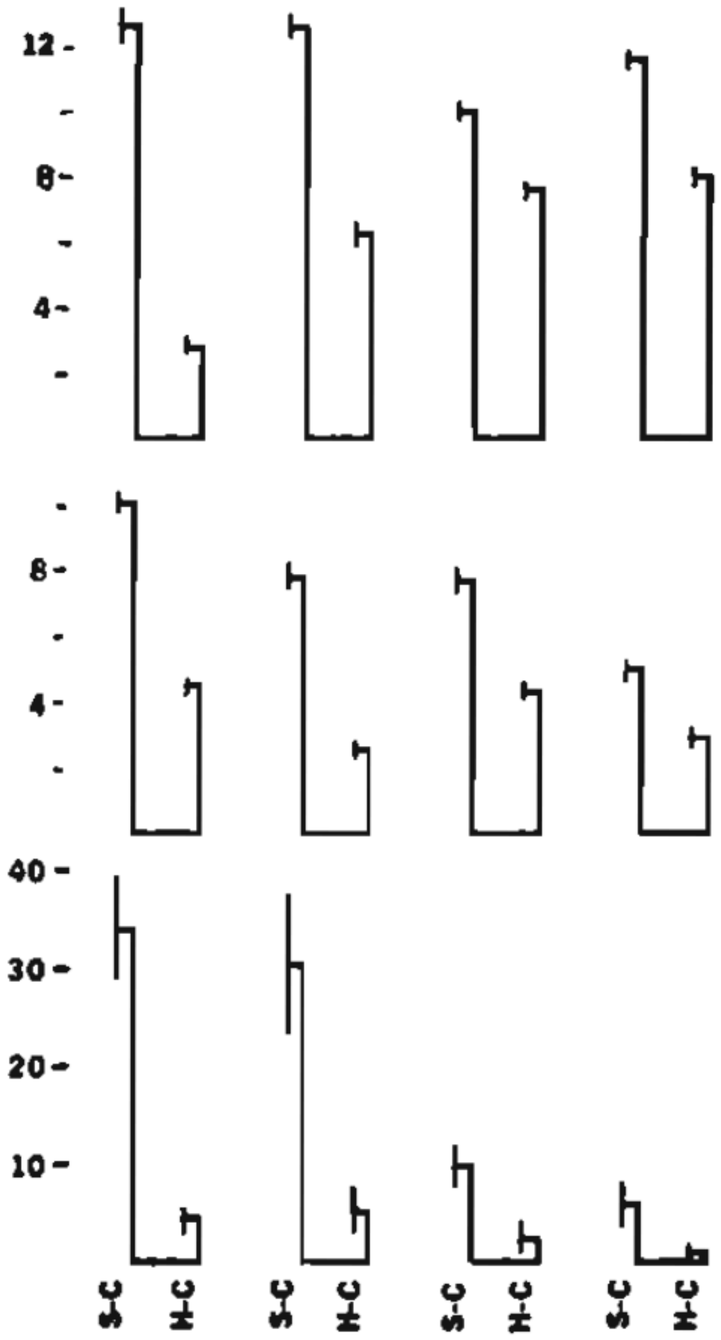
Data to ink

1. Principle of the maximal data to ink ratio

Sometimes leading to overly minimalistic designs
("Nothing can be erased without losing information”)

~ low acceptance (Inbar, Tractinsky & Meyer 2007)





James T. Kuznicki and N. Bruce McCutcheon, "Cross-Enhancement of the Sour Taste on Single Human Taste Papillae," *Journal of Experimental Psychology: General*, 108 (1979), 76.

(Tufte 2001: 100)

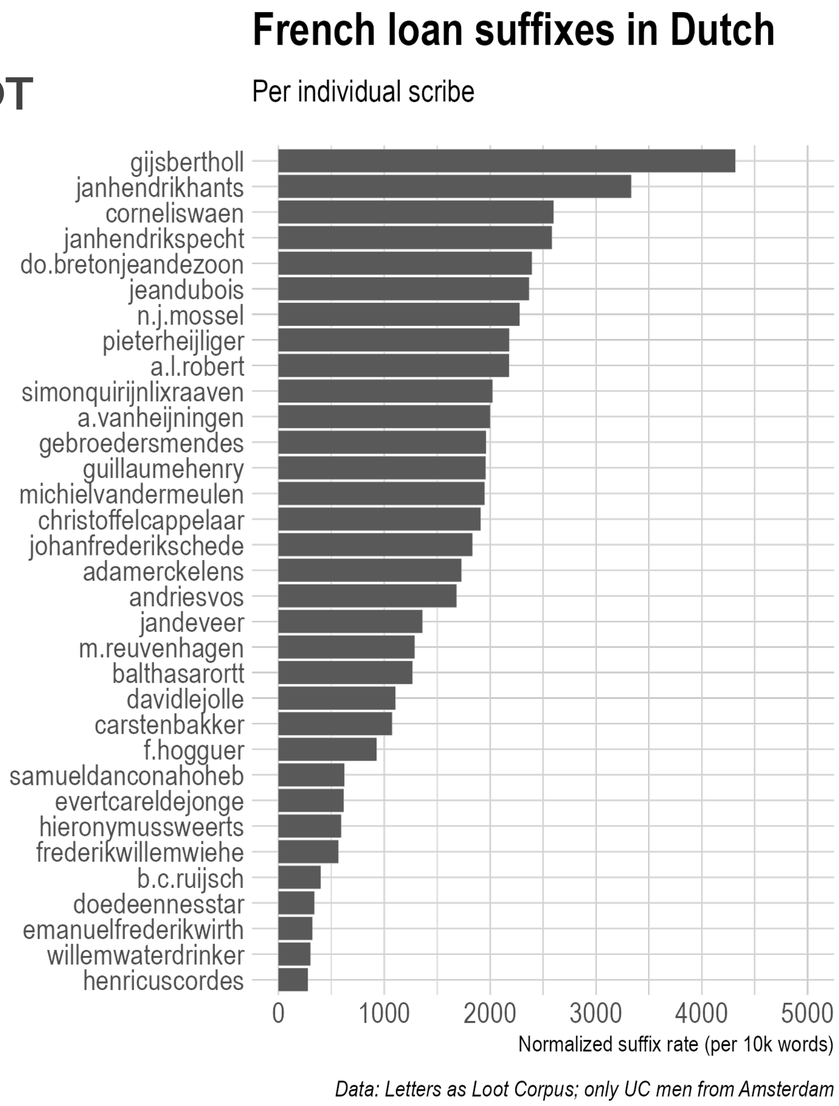
Data to ink

1. Principle of the maximal data to ink ratio

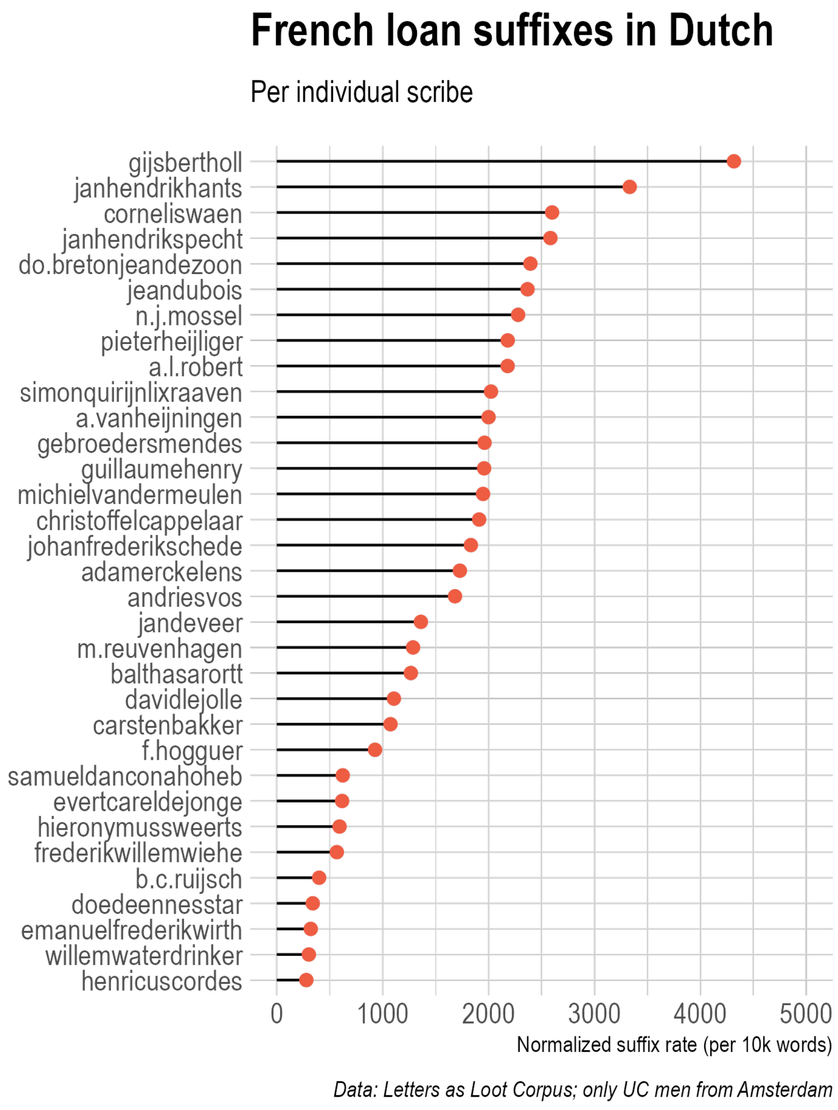
However, consider barplot versus alternatives...

Data to ink

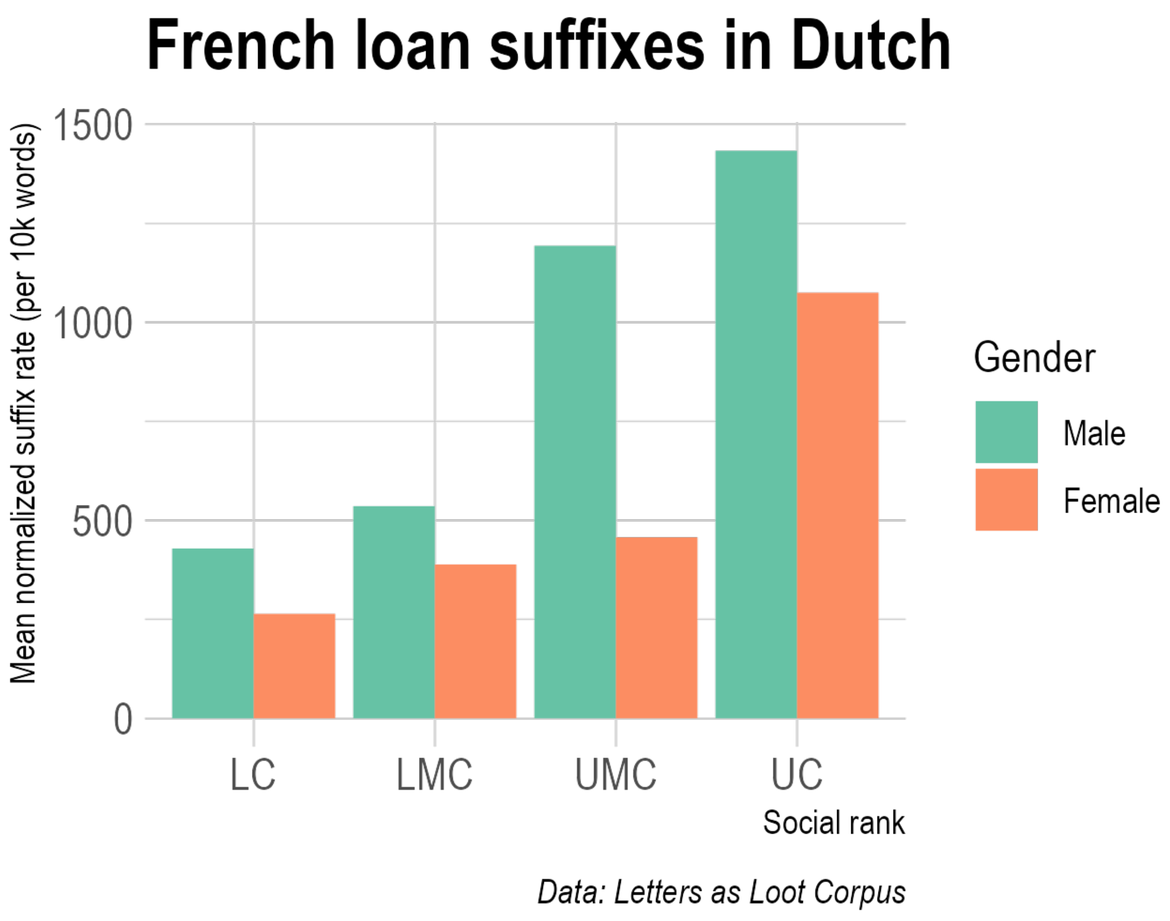
Barplot



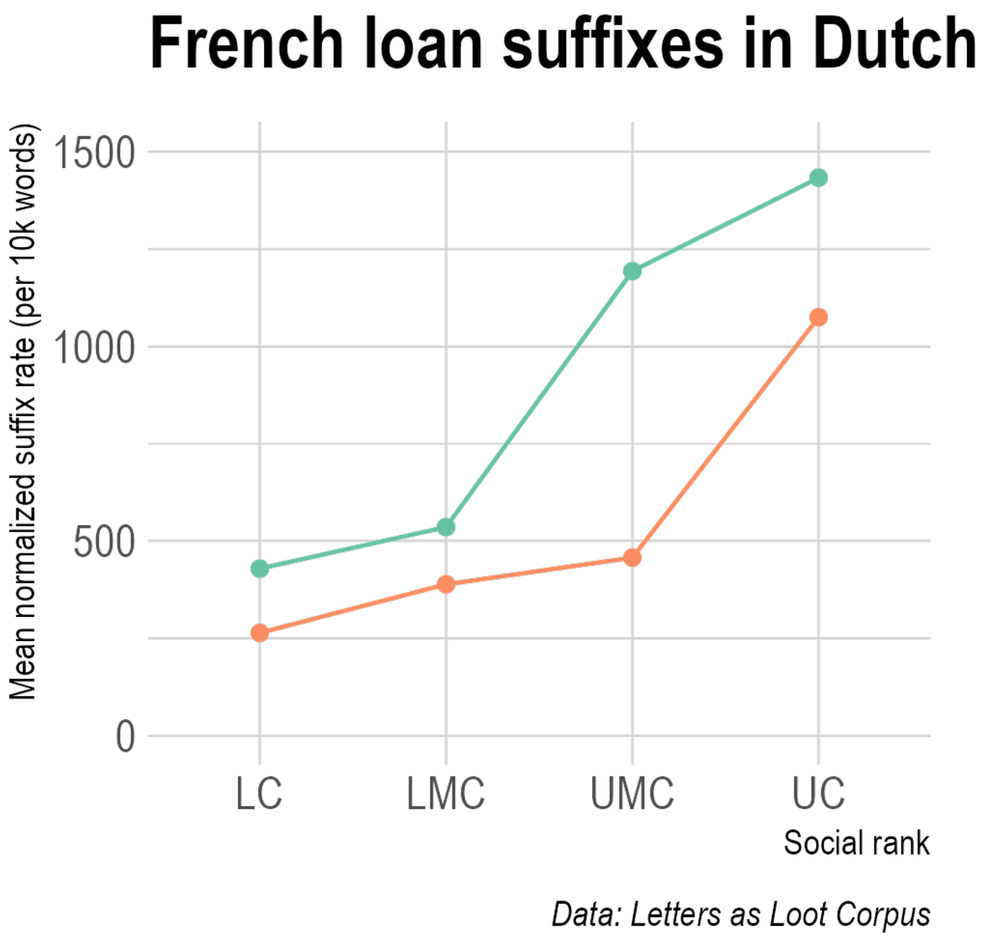
Dot plot



Data to ink

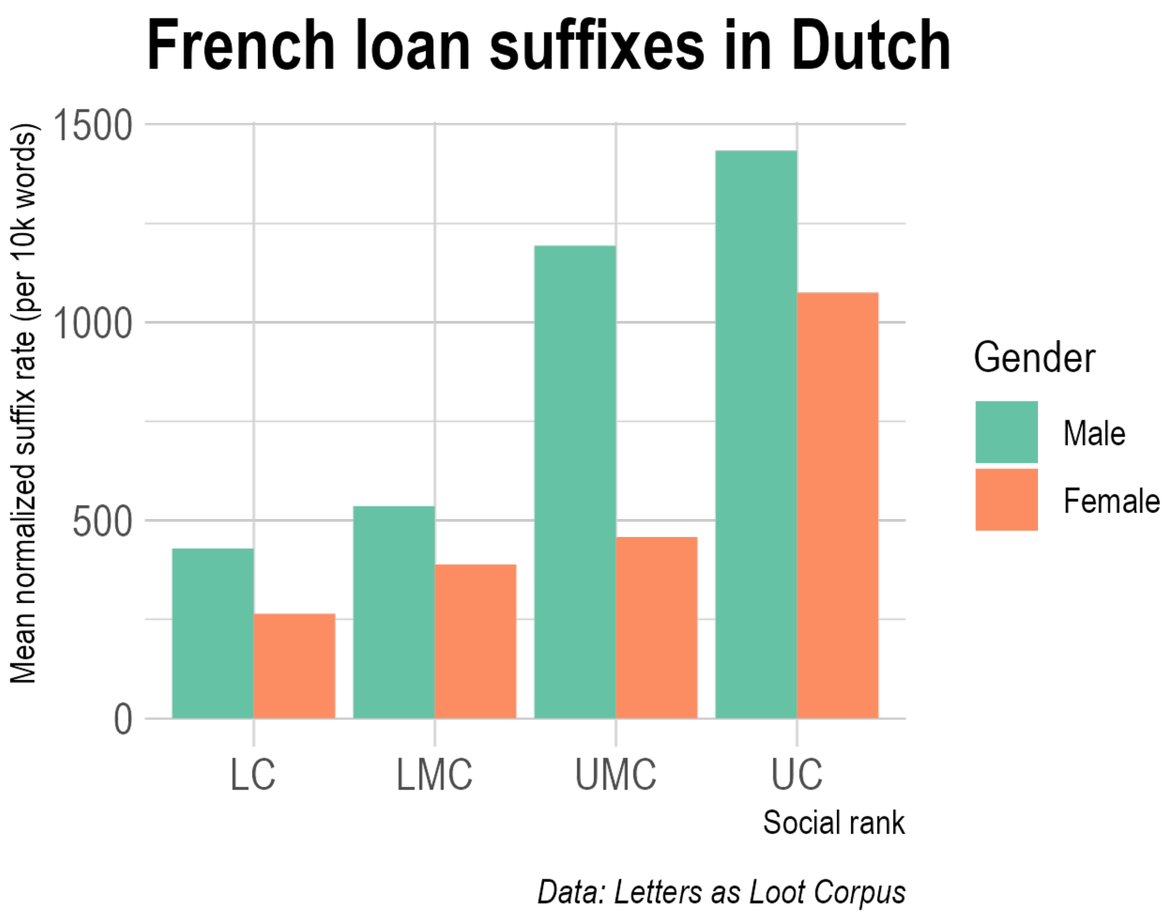


BARPLOT

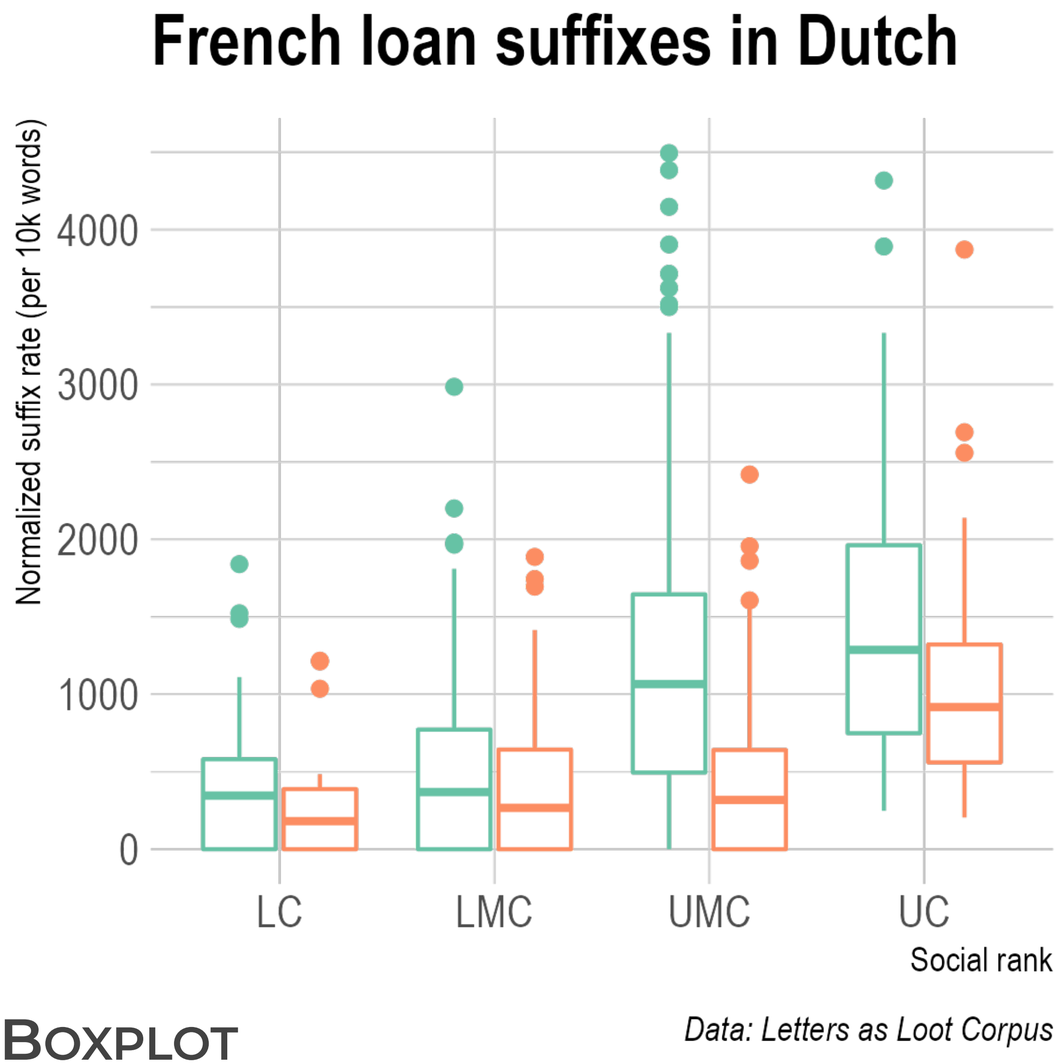


LINEPLOT

Data to ink



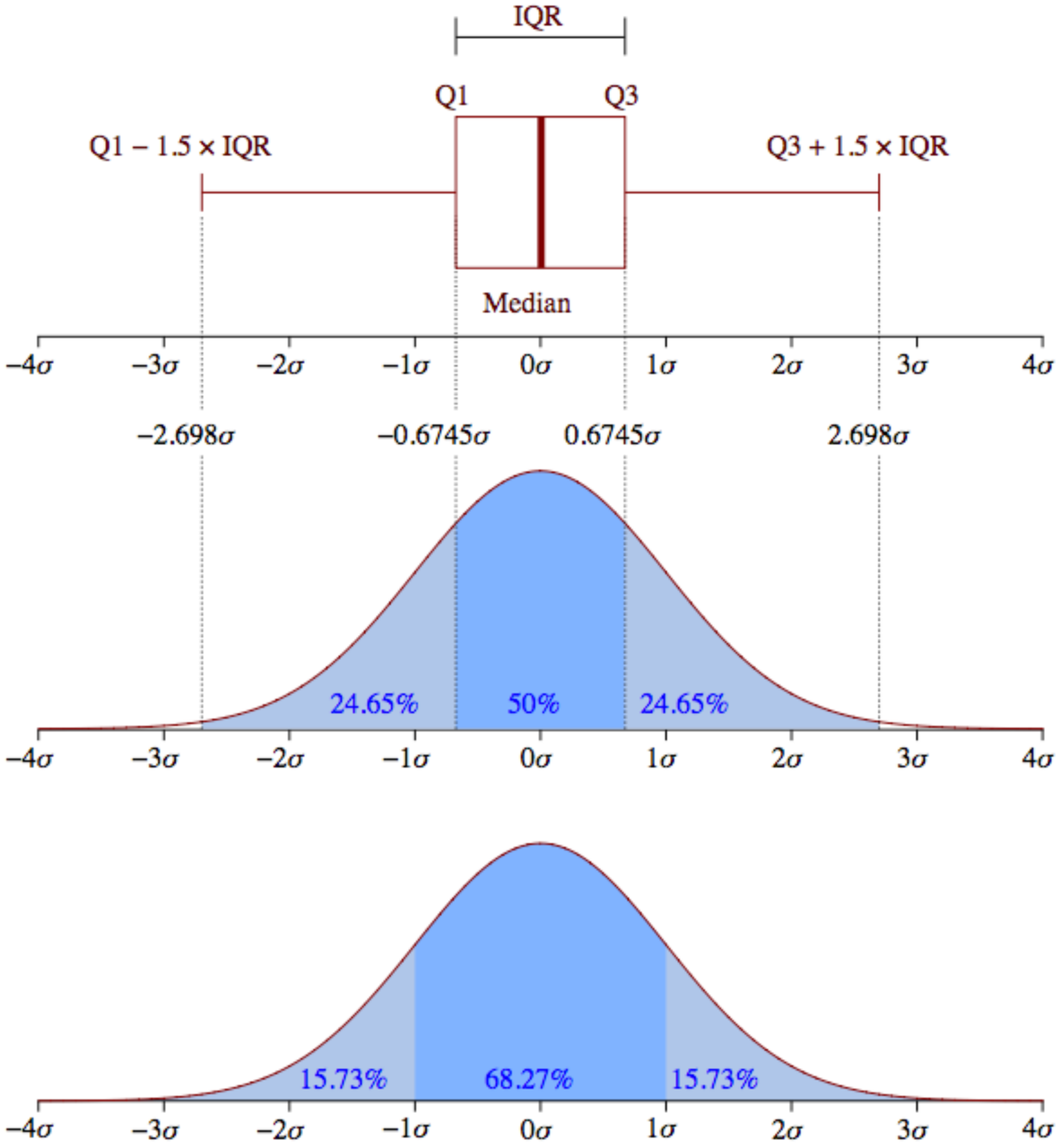
BARPLOT



BOXPLOT

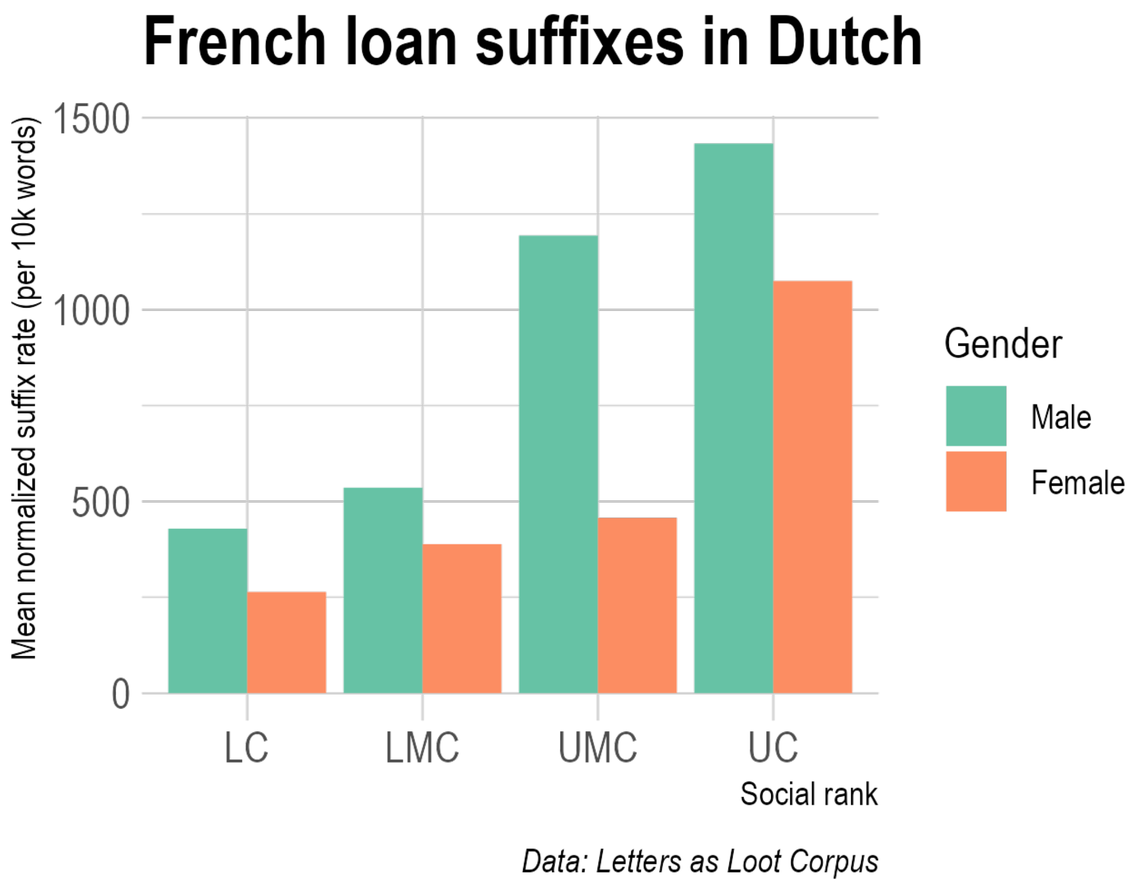
Boxplot

Density plot

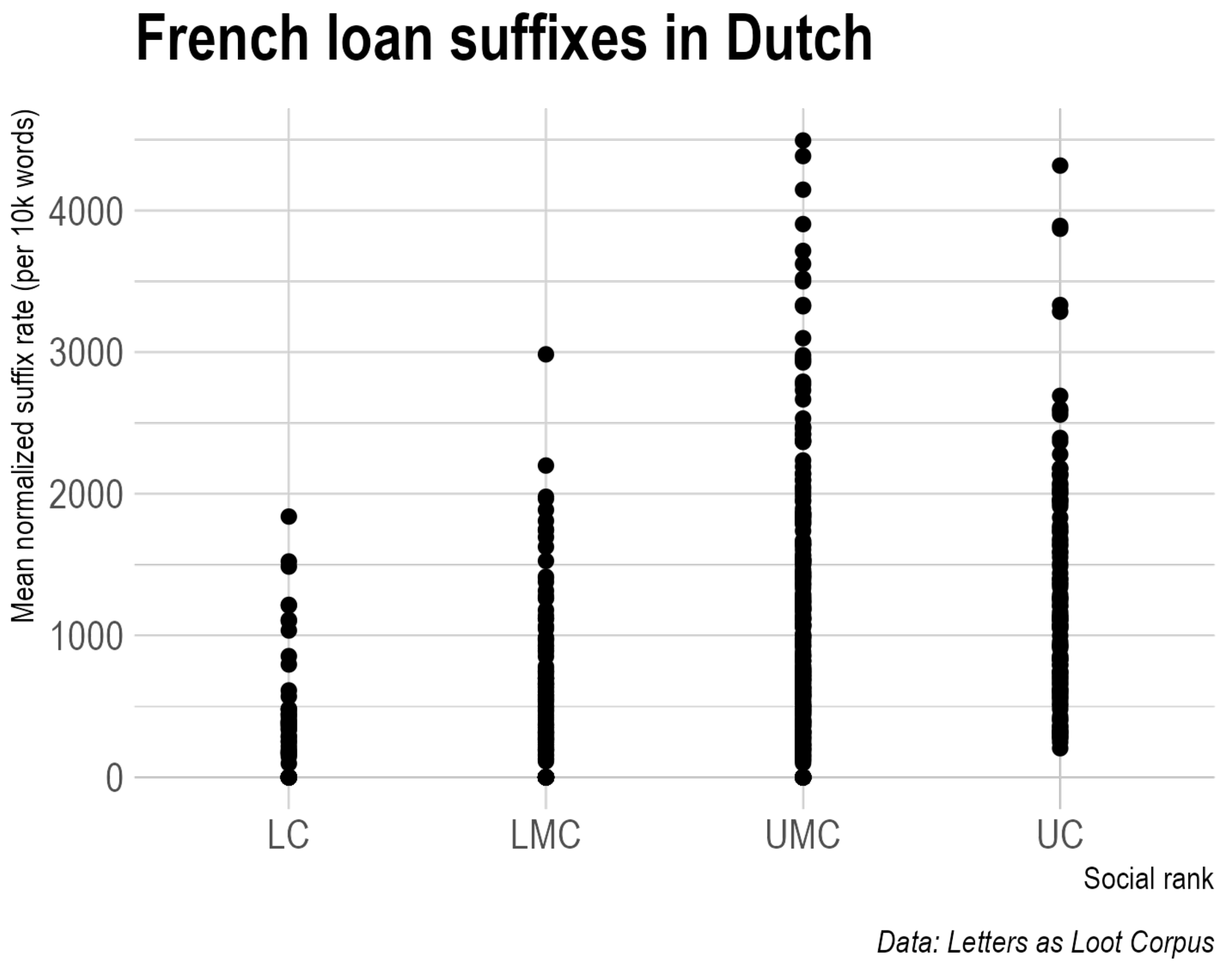


Wikipedia - Boxplot and a probability density function (pdf) of a Normal N(0,1σ2) Population

Data to ink

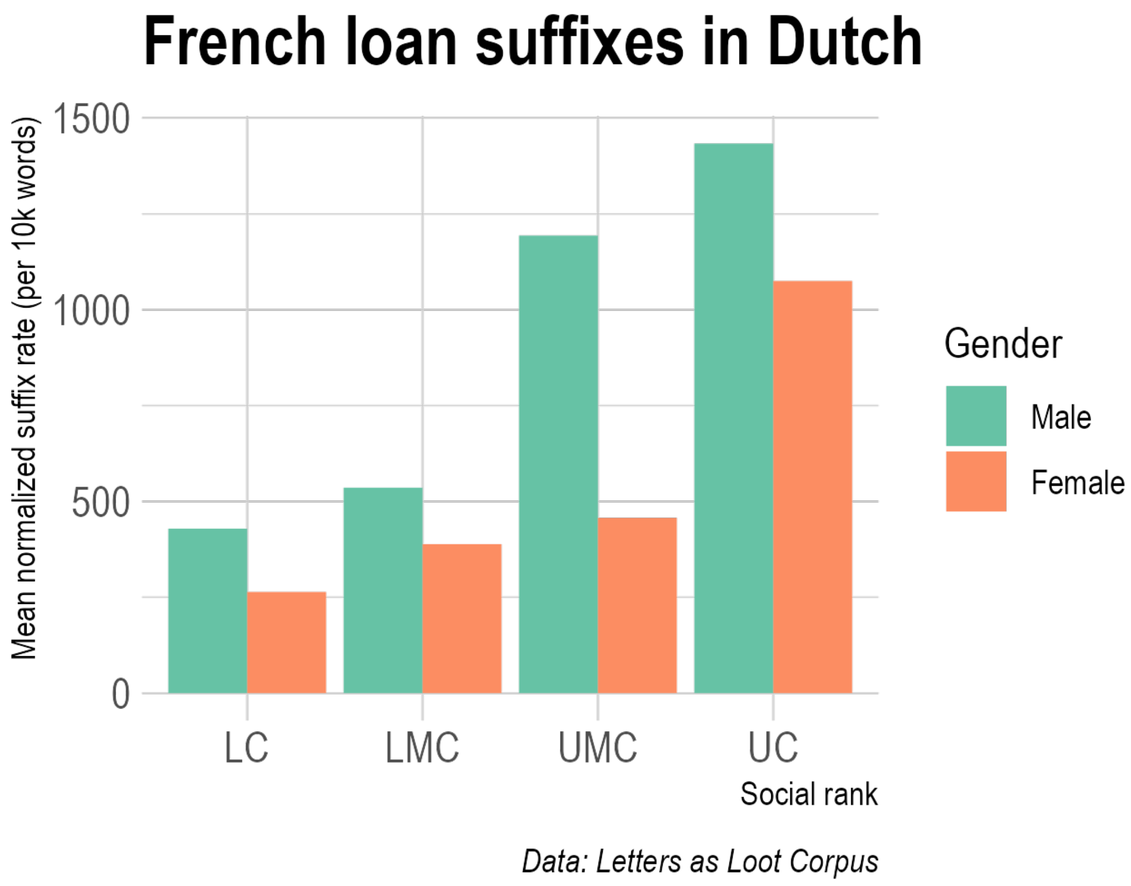


BARPLOT

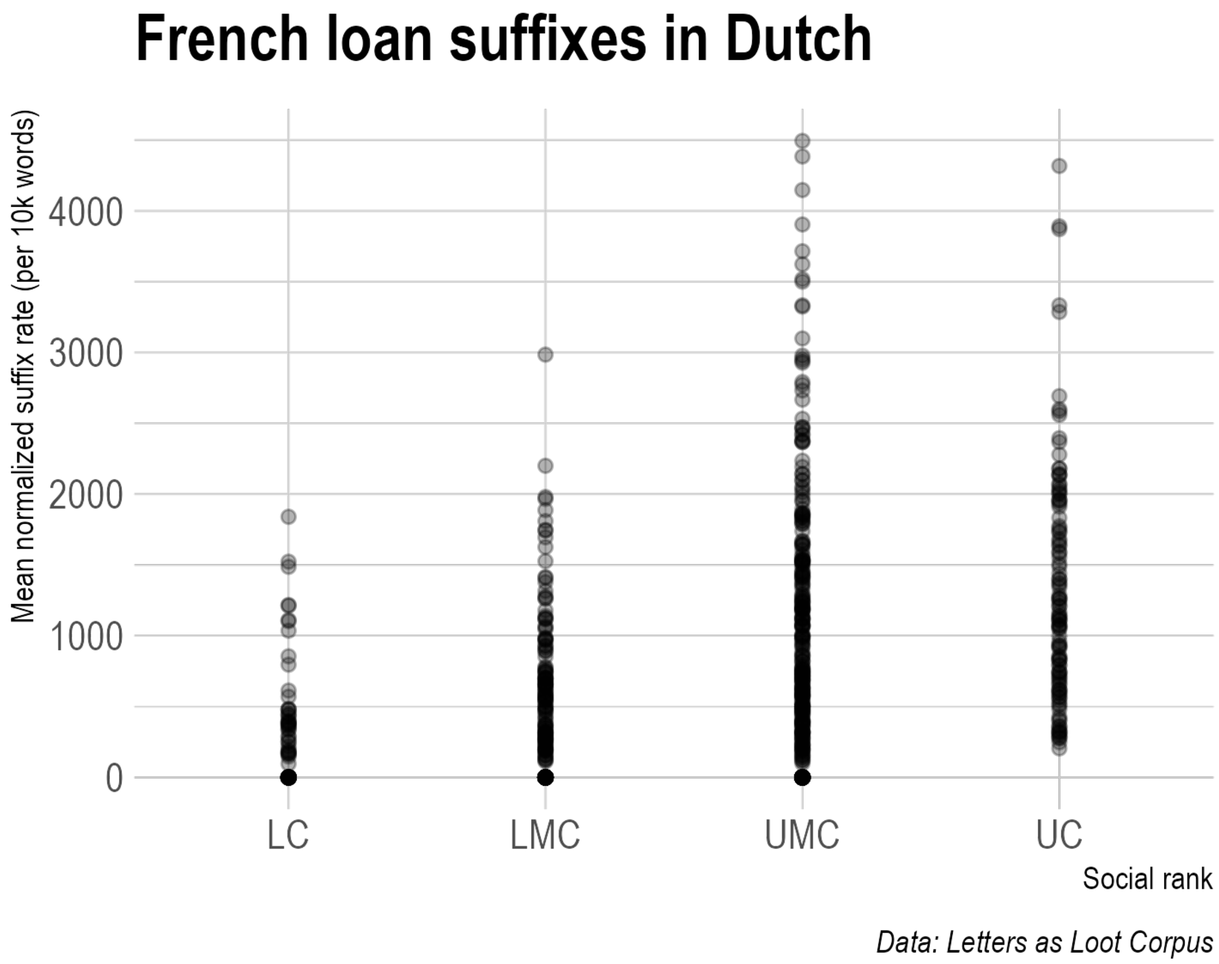


SCATTERPLOT

Data to ink

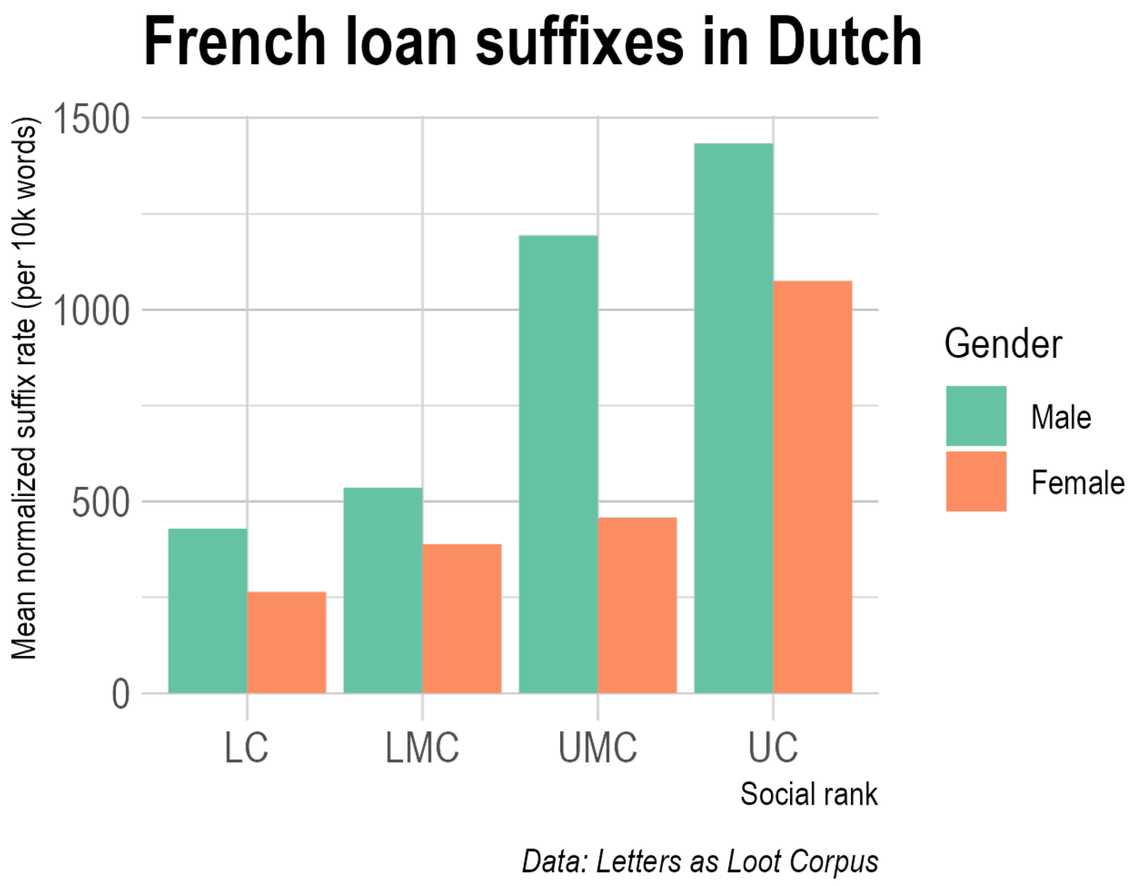


BARPLOT

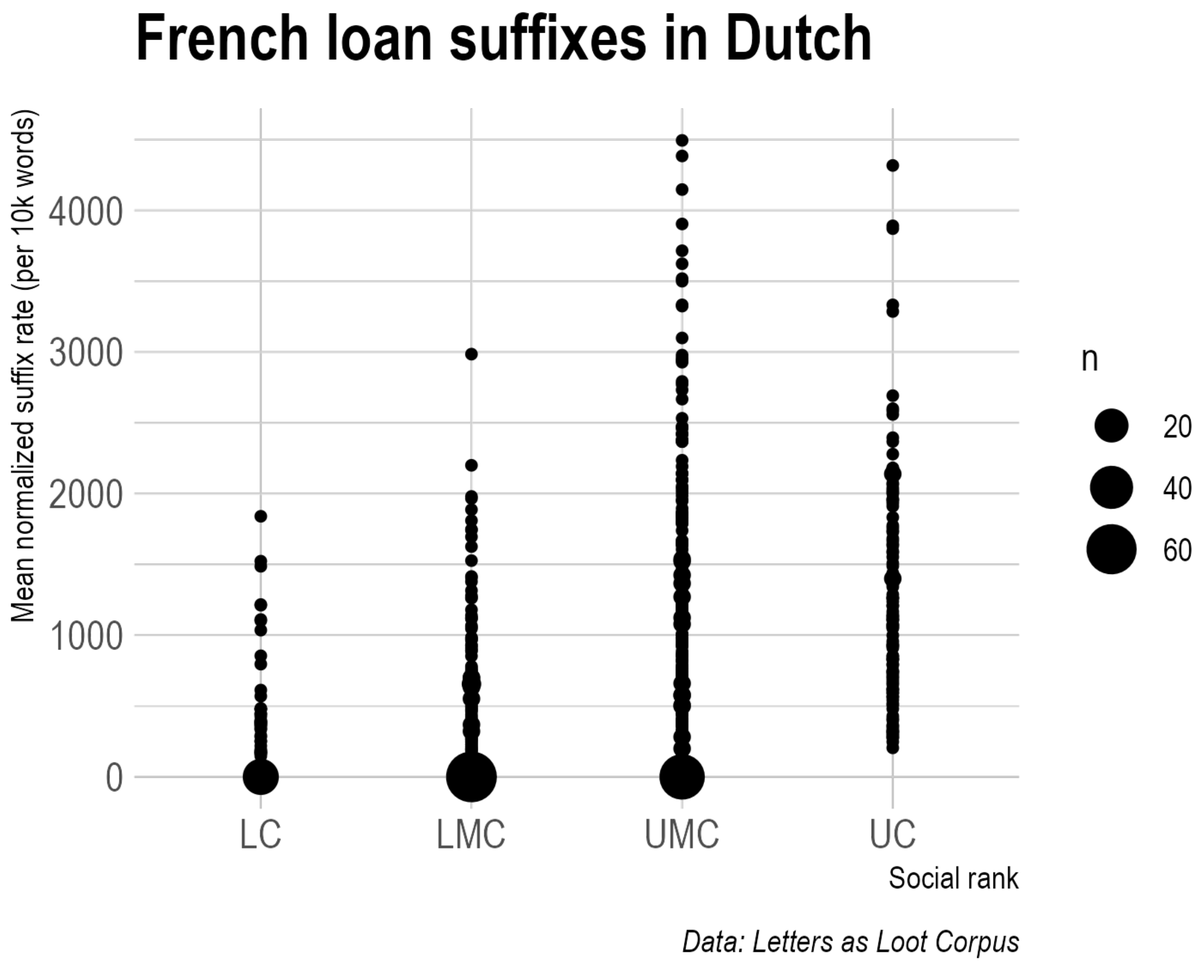


SCATTERPLOT

Data to ink

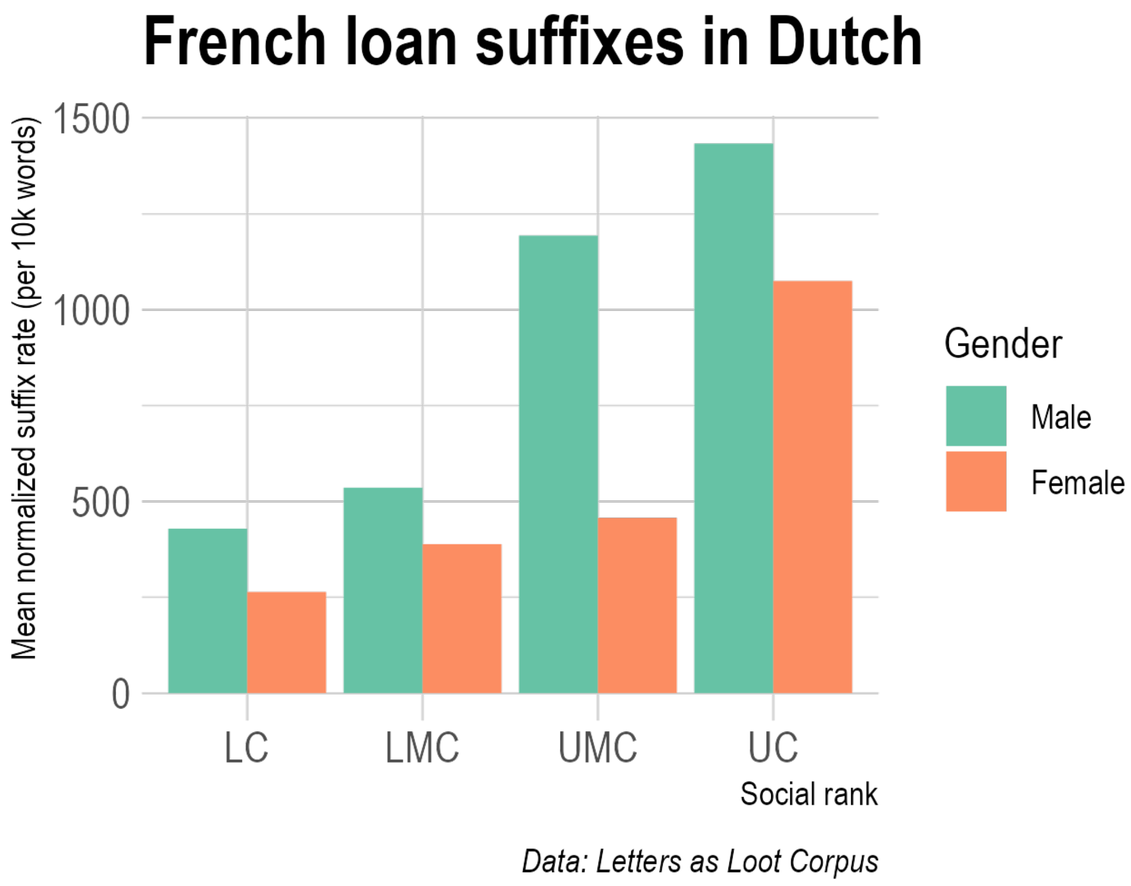


BARPLOT

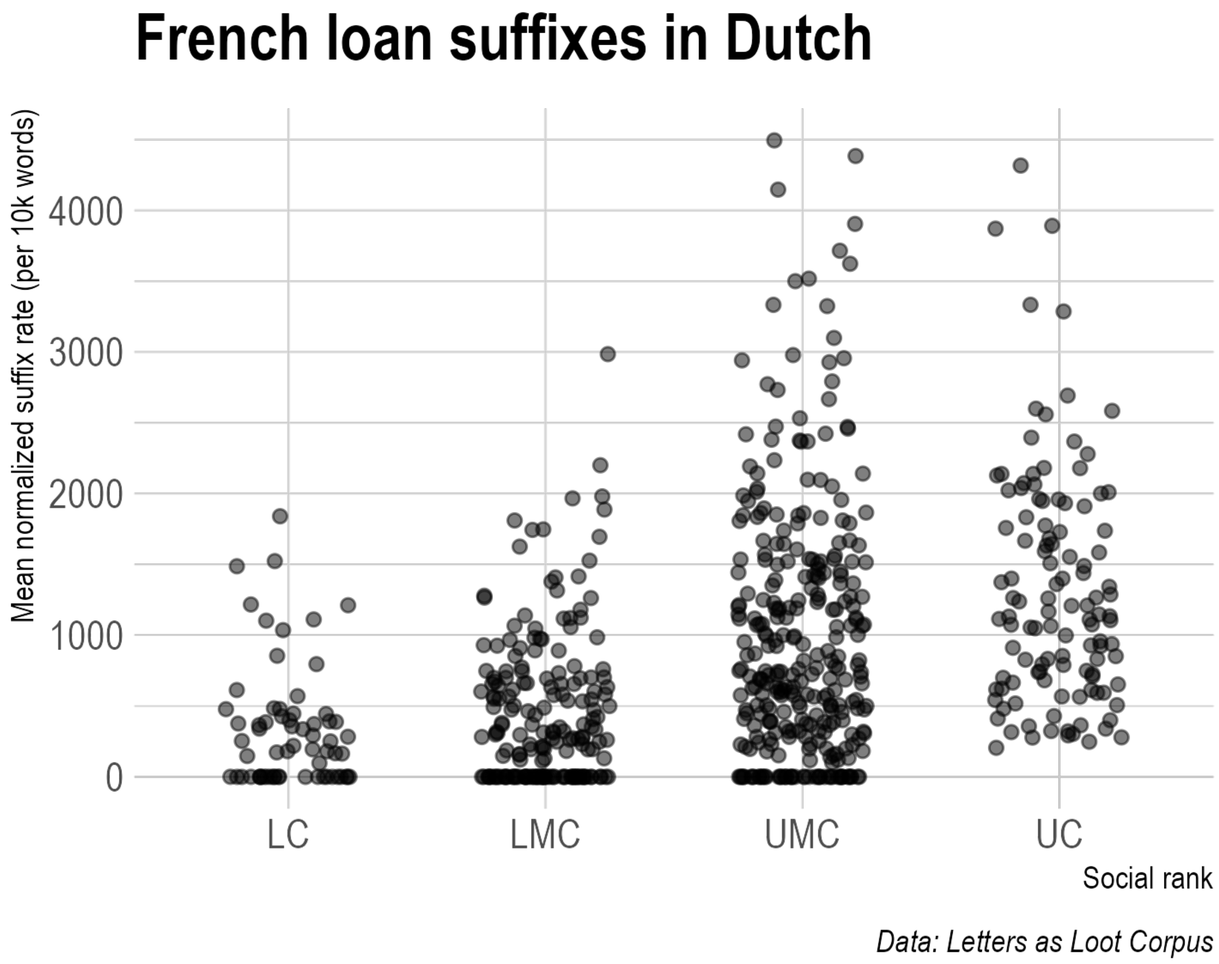


COUNT PLOT

Data to ink

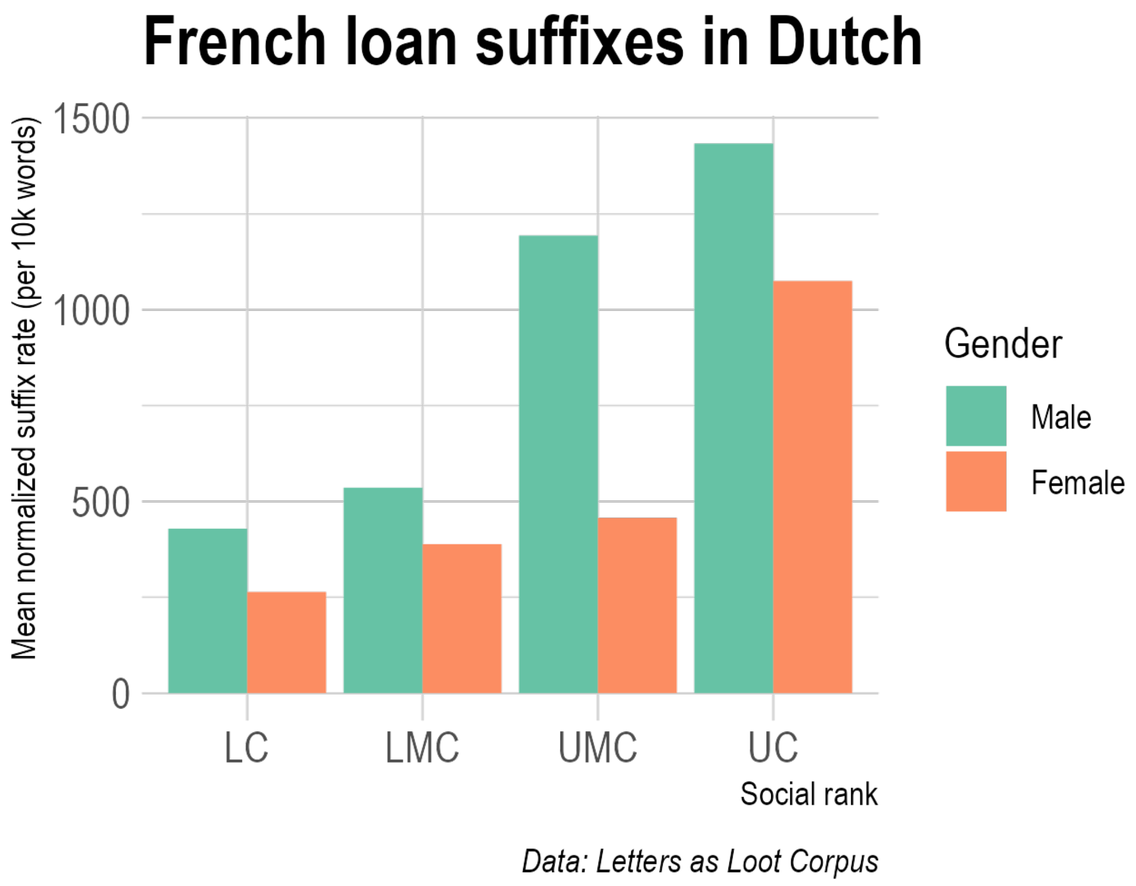


BARPLOT

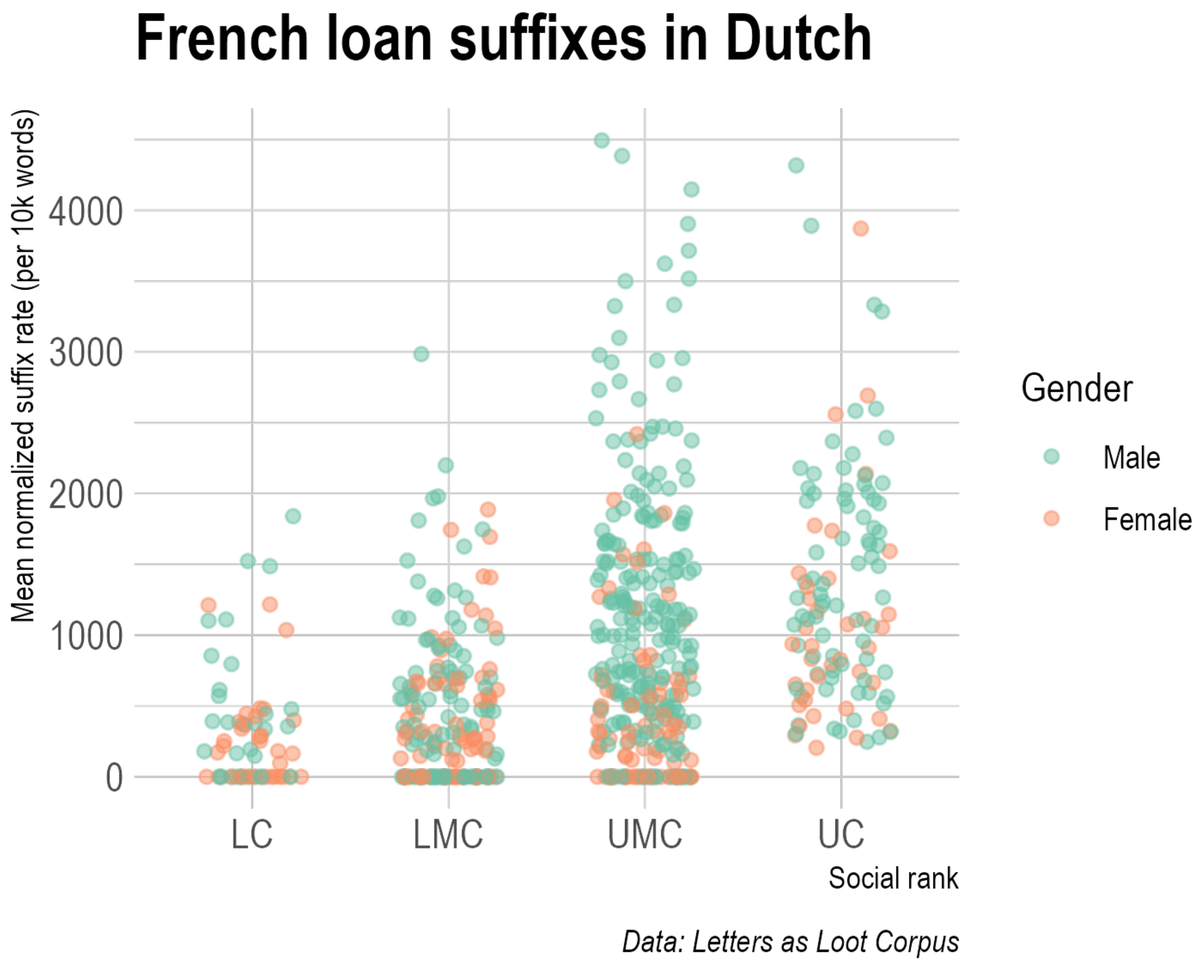


SCATTERPLOT

Data to ink

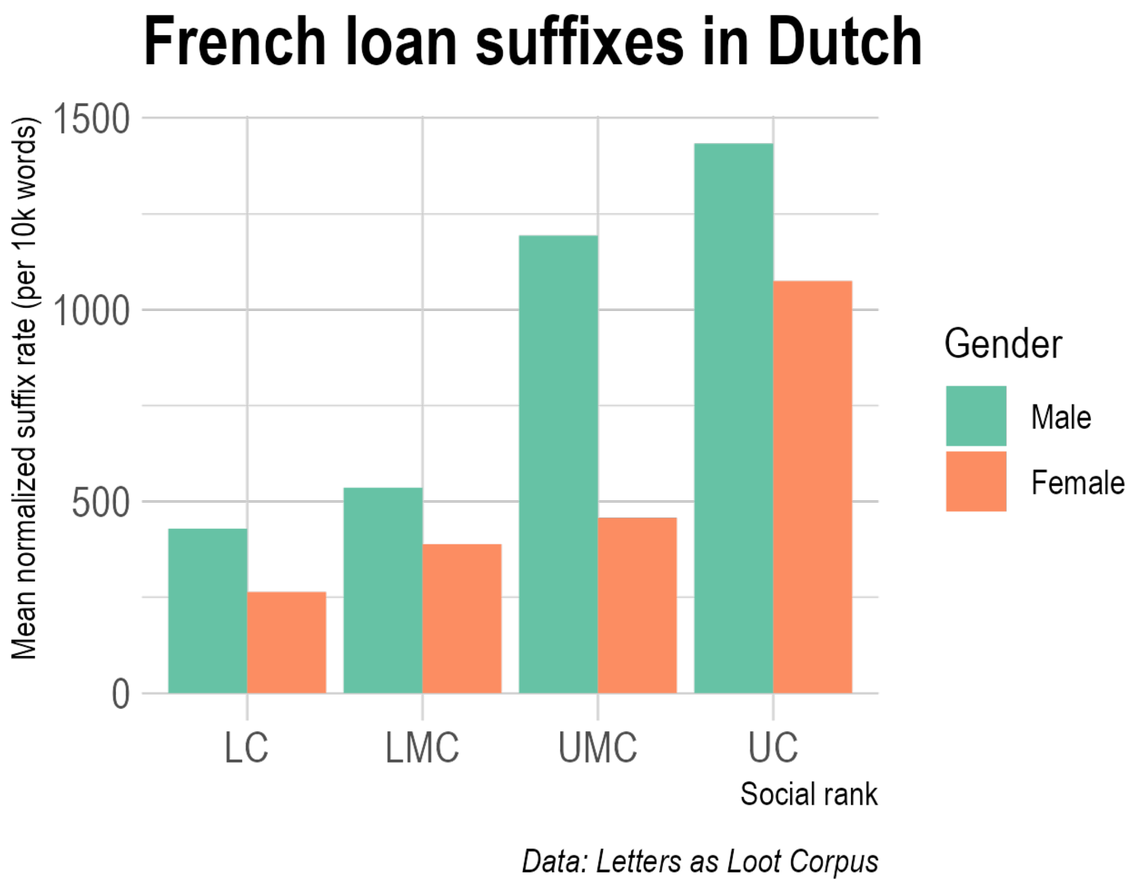


BARPLOT

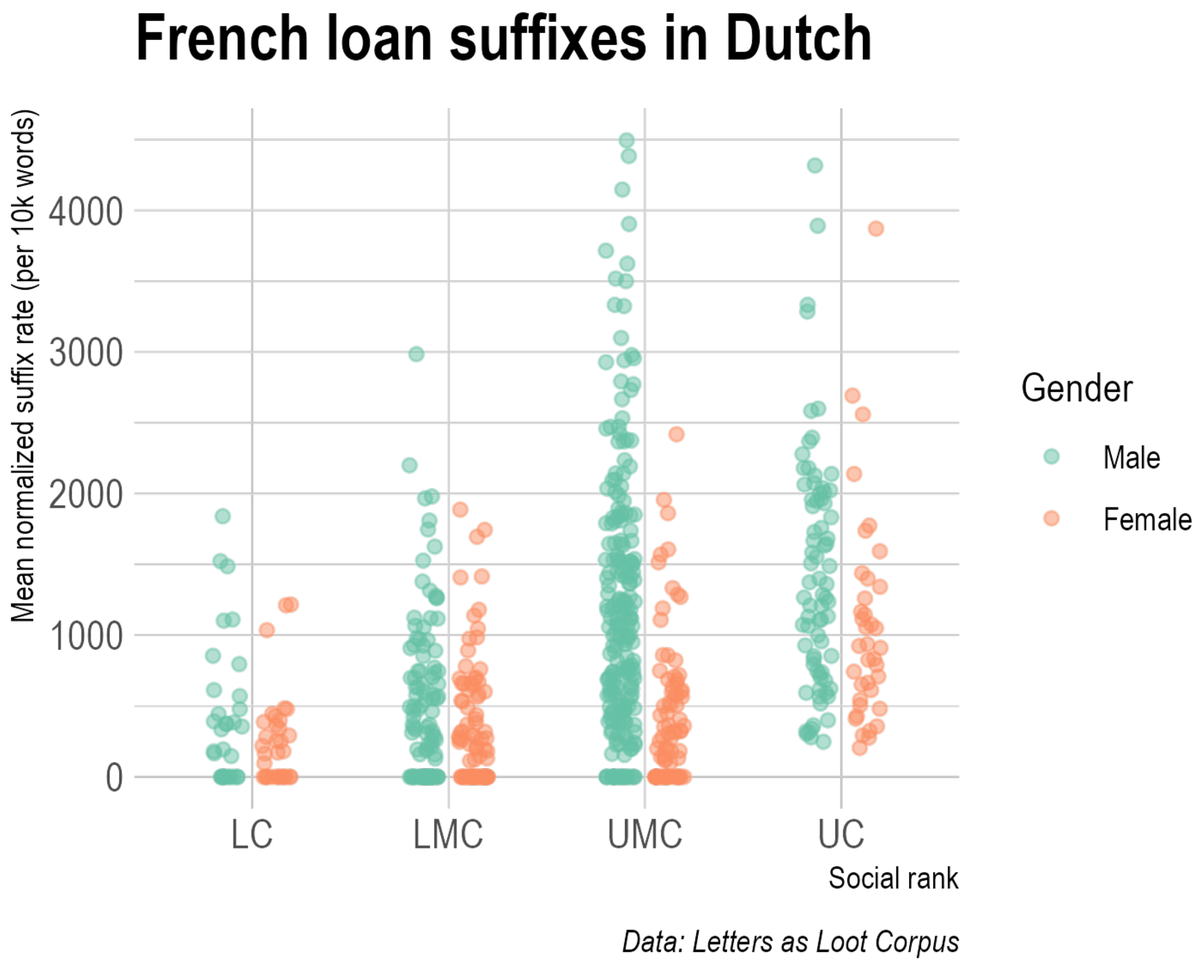


SCATTERPLOT

Data to ink

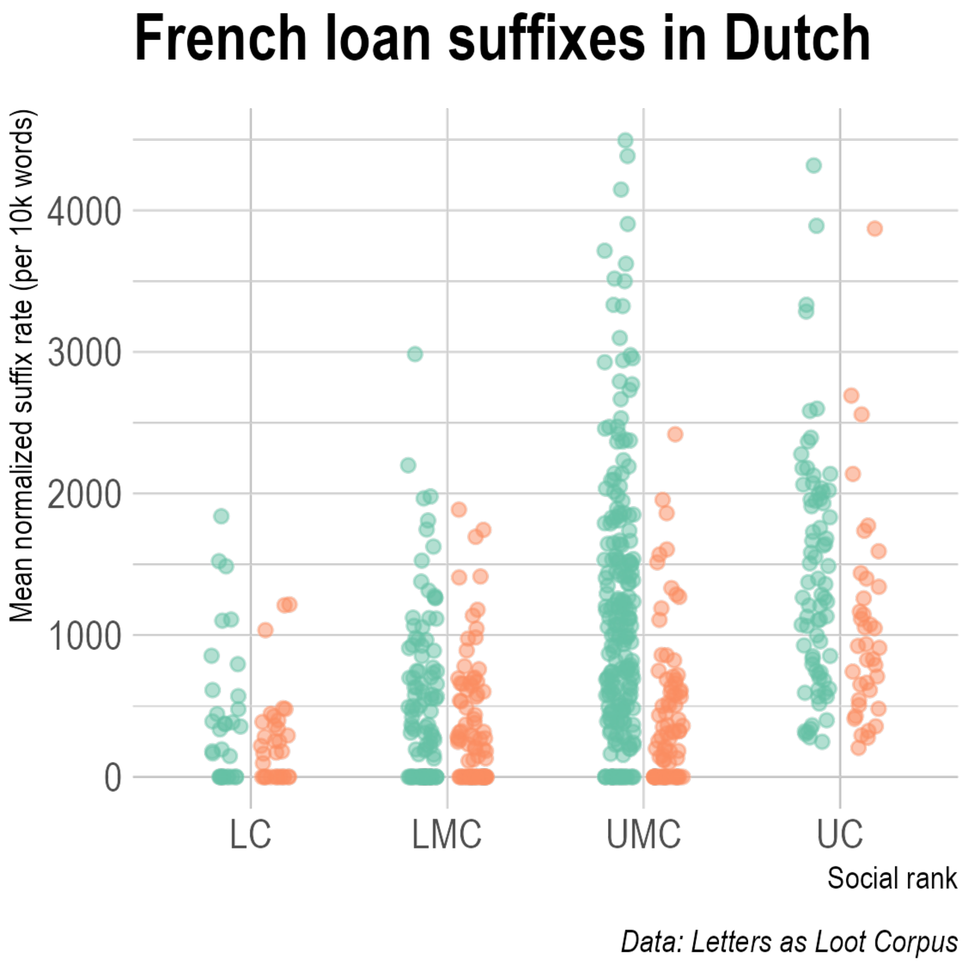


BARPLOT

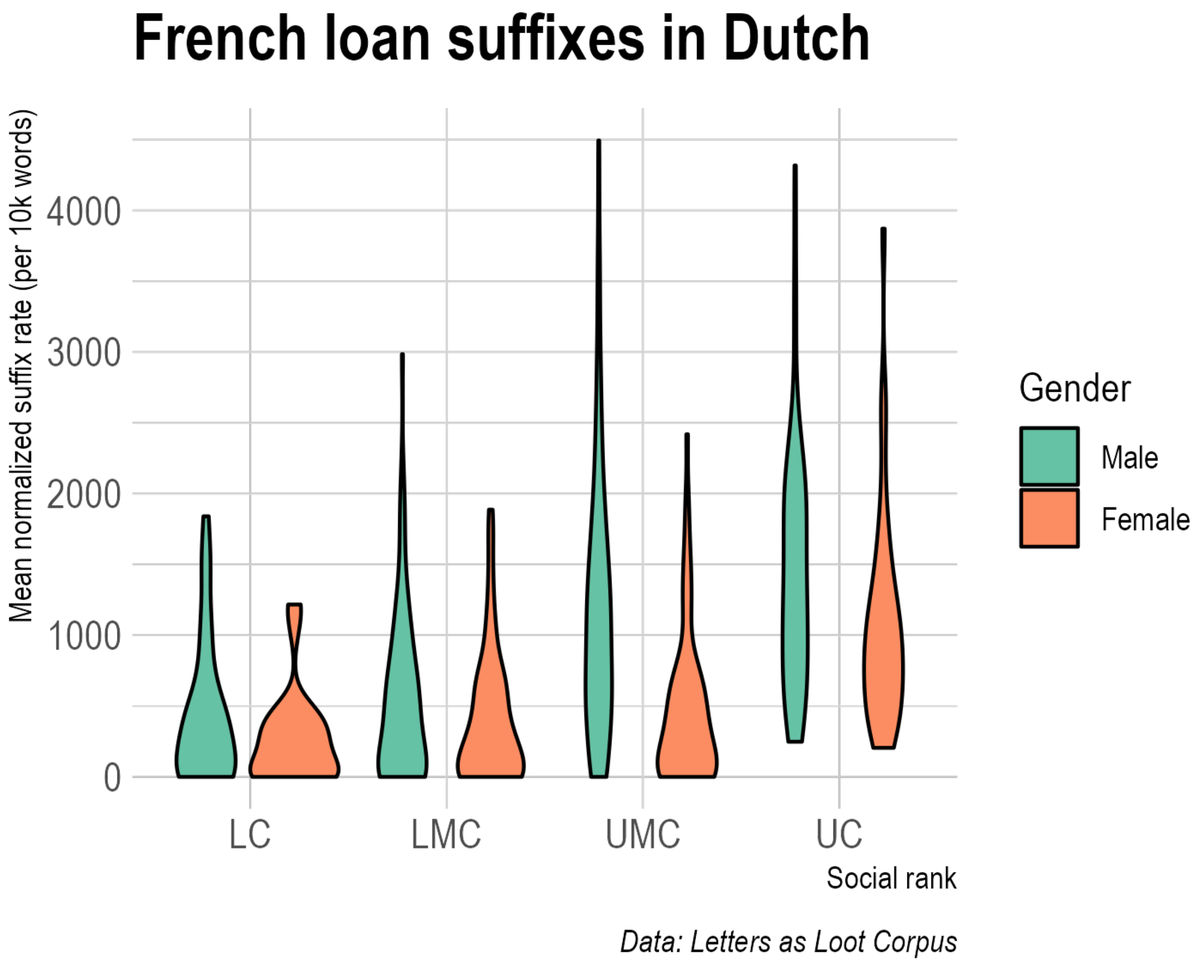


DODGED SCATTERPLOT

Data to ink

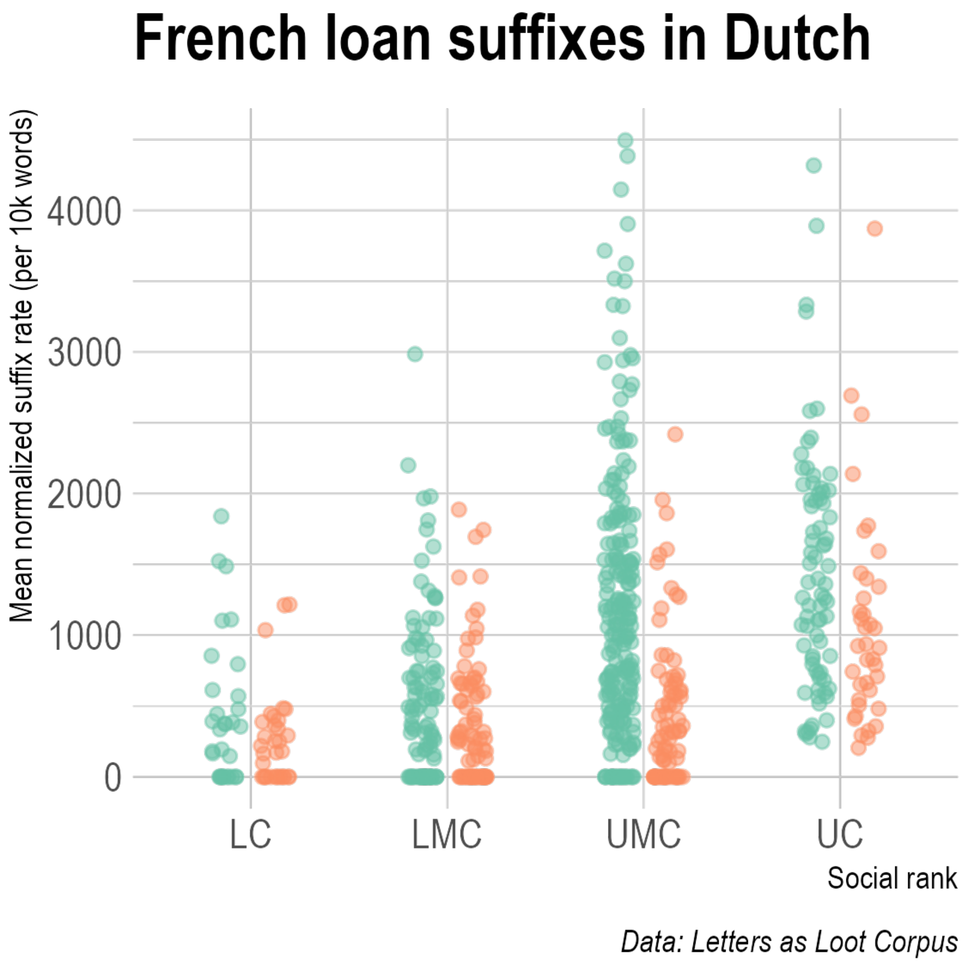


DODGED SCATTERPLOT

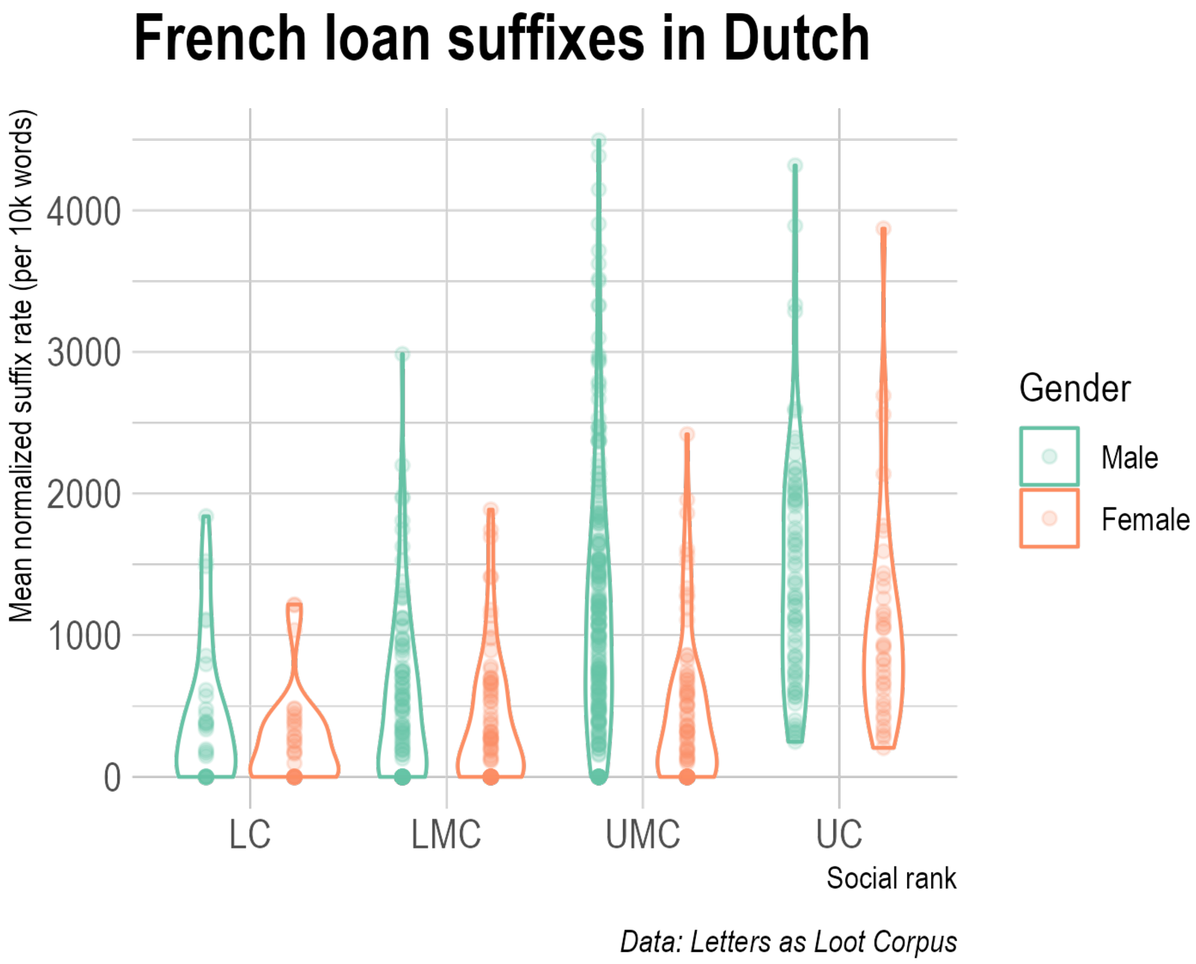


VIOLIN PLOT

Data to ink

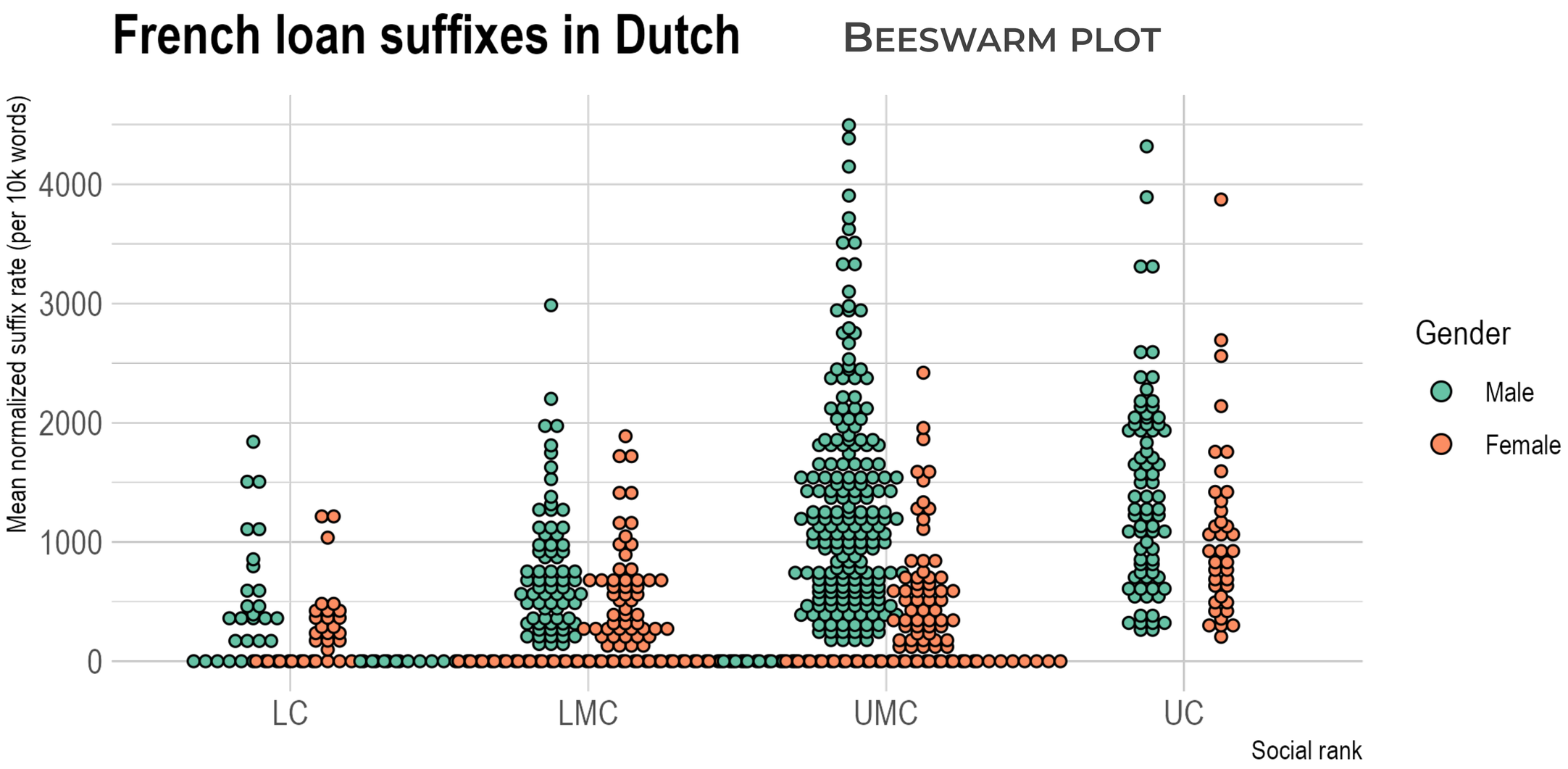


DODGED SCATTERPLOT



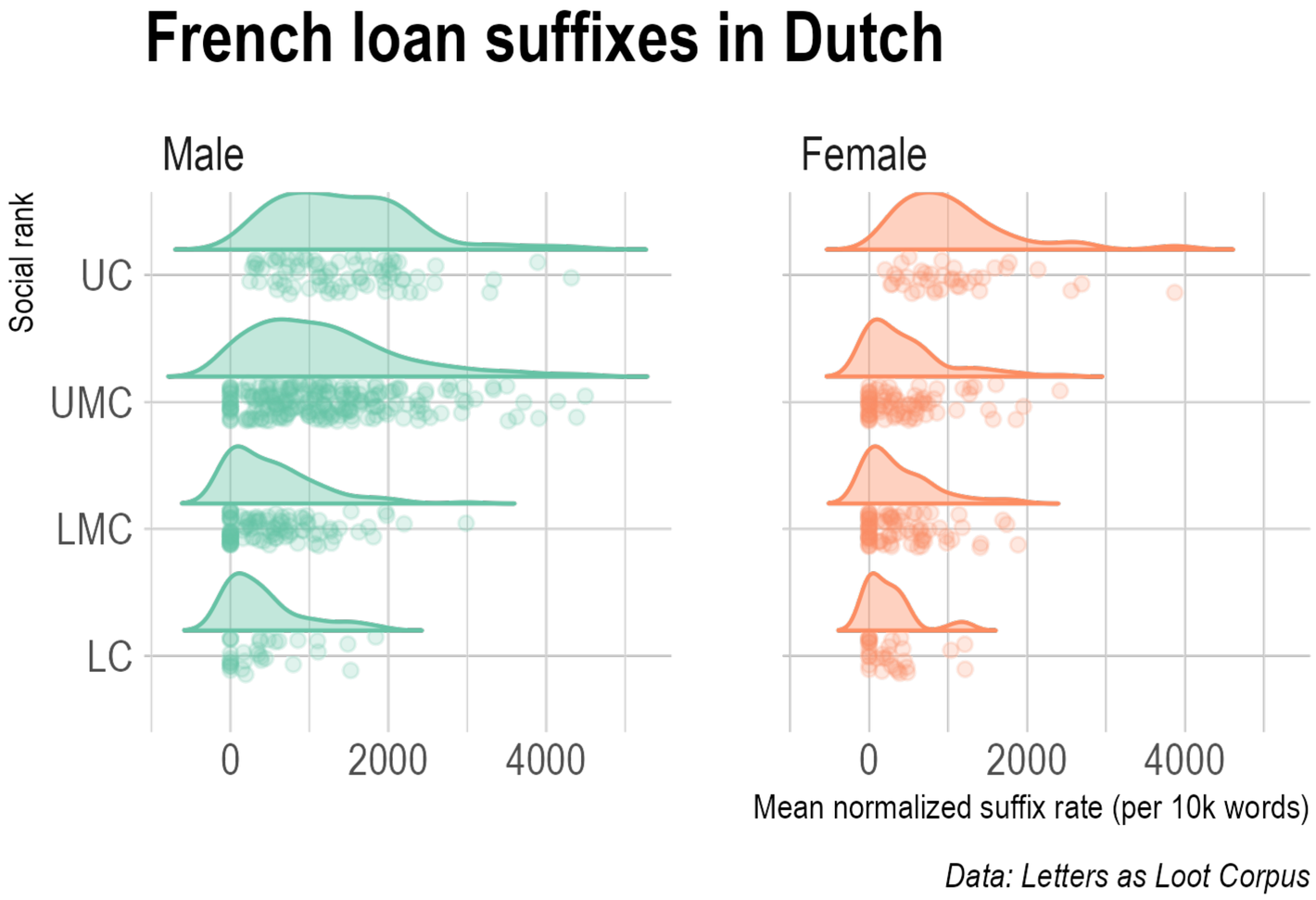
VIOLIN-SCATTER PLOT

Data to ink

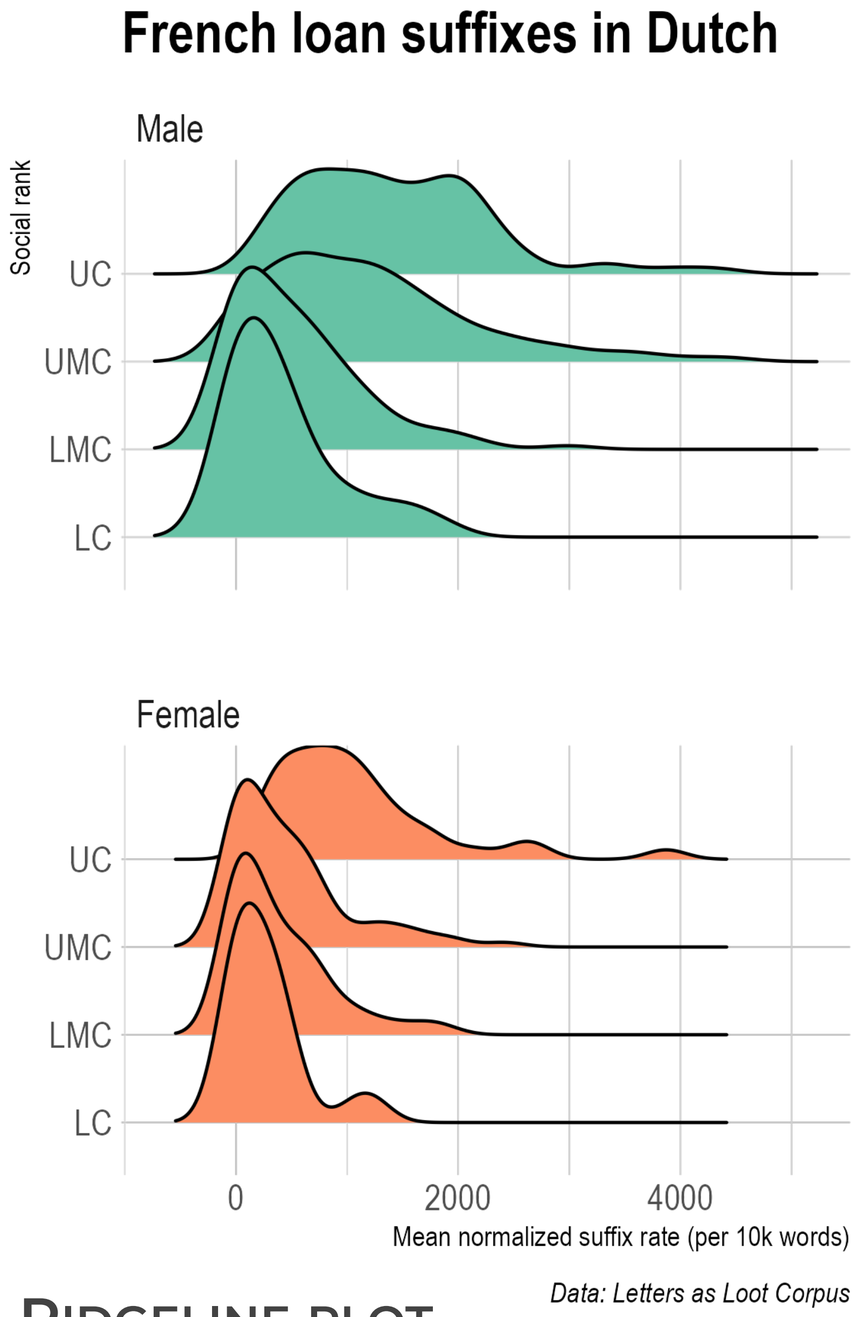


BEESWARM PLOT

Data to ink



RAINCLOUD PLOT



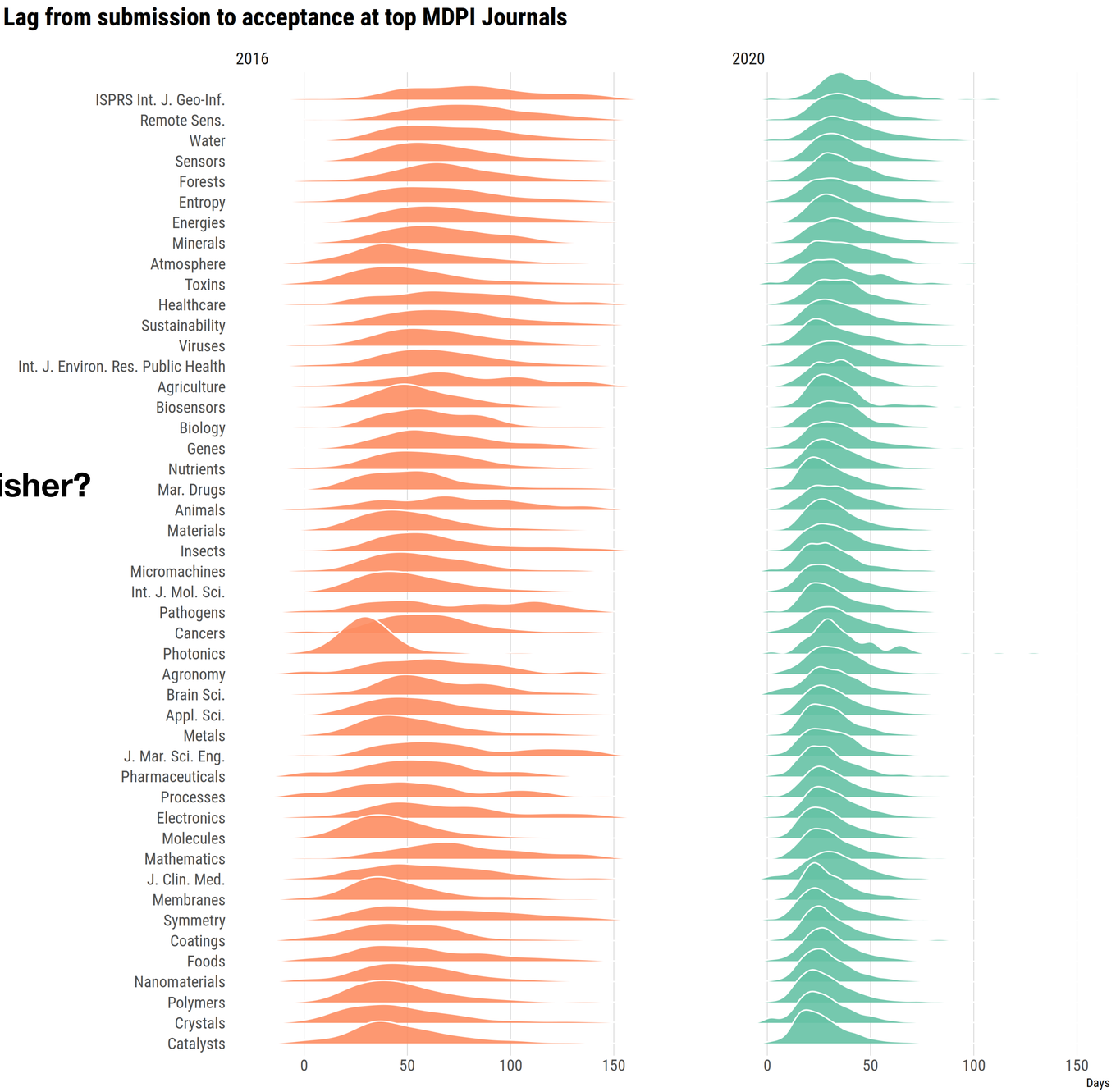
RIDGELINE PLOT

Data to ink

Ridgeline plot

Paolo Crosetto - Is MDPI a predatory publisher?

https://paolocrosetto.wordpress.com/2021/04/12/is-mdpi-a-predatory-publisher/



Data to ink

Which of these have you used, or would you use?
Which are more intuitive to read, and which are too complex?

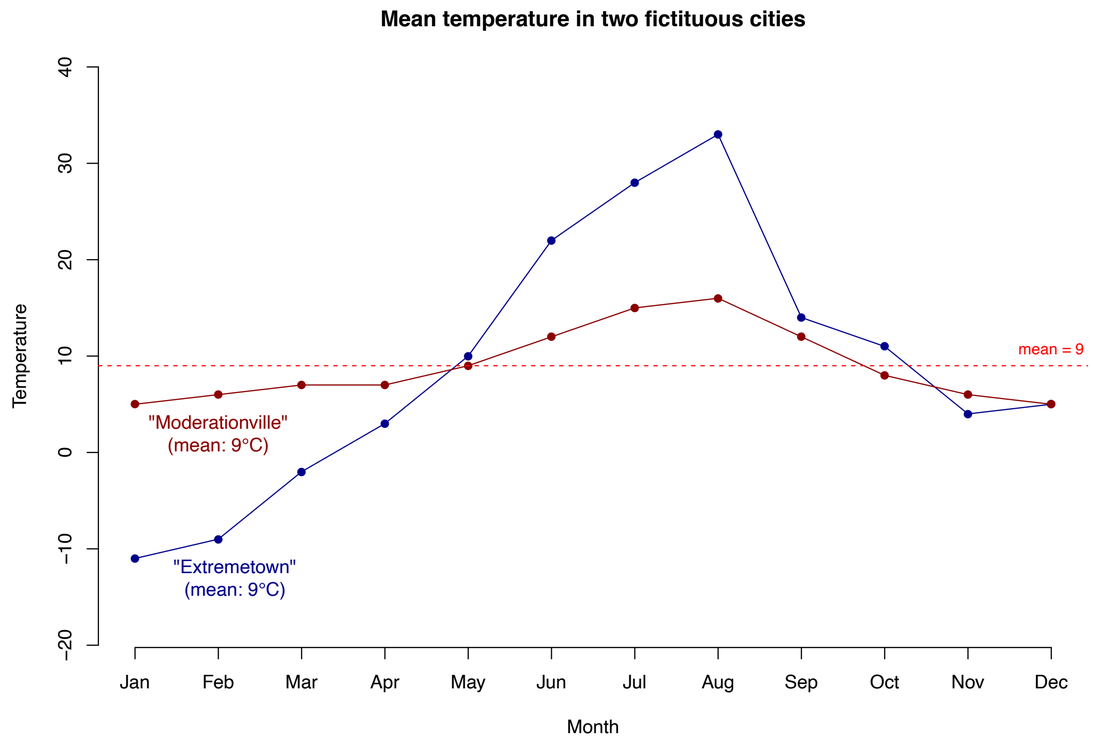
- Barplot
- Dot plot
- Lineplot
- Boxplot
- Density plot
- Scatterplot
- Count plot
- Violin plot
- Violin-scatter plot
- Bee swarm plot Raincloud plot
- Ridgeline plot
- Heat map
- Choropleths
- Association plots
- Mosaic plots
- Sankey diagrams or alluvial plots

Data to ink

1. Whiskers as extra information
2. Trend lines as extra information

Data to ink

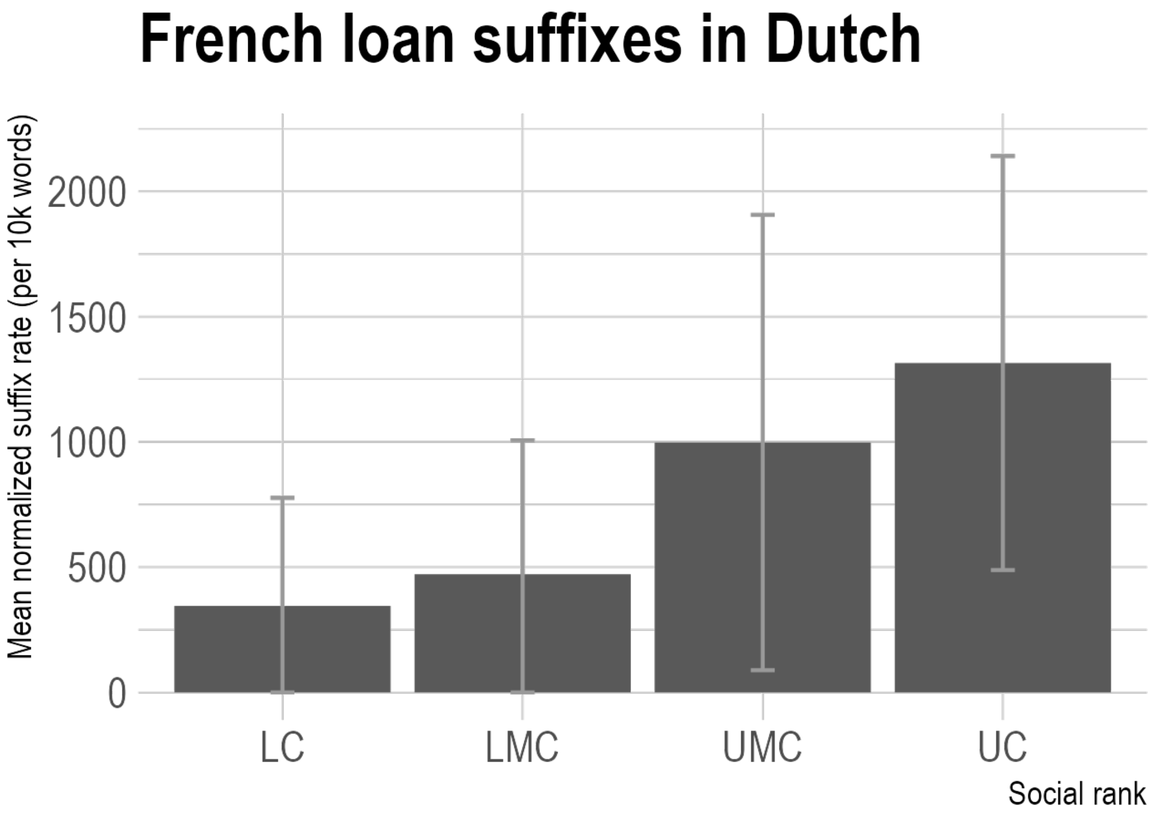
Adding a measure of dispersion or uncertainty
– Standard deviation → scatter in the sample
– Confidence interval → Confidence in the result
(always mention which is shown)



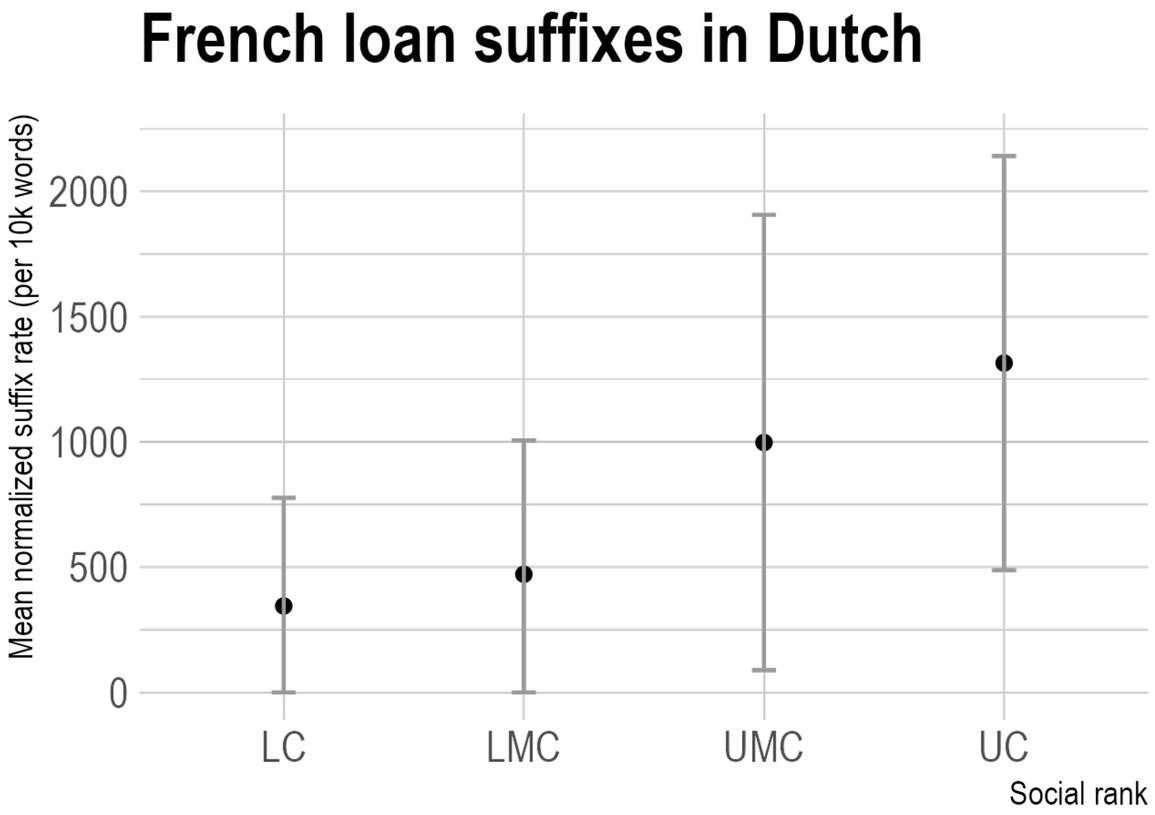
$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

Data to ink

Adding a measure of dispersion or uncertainty — standard deviation



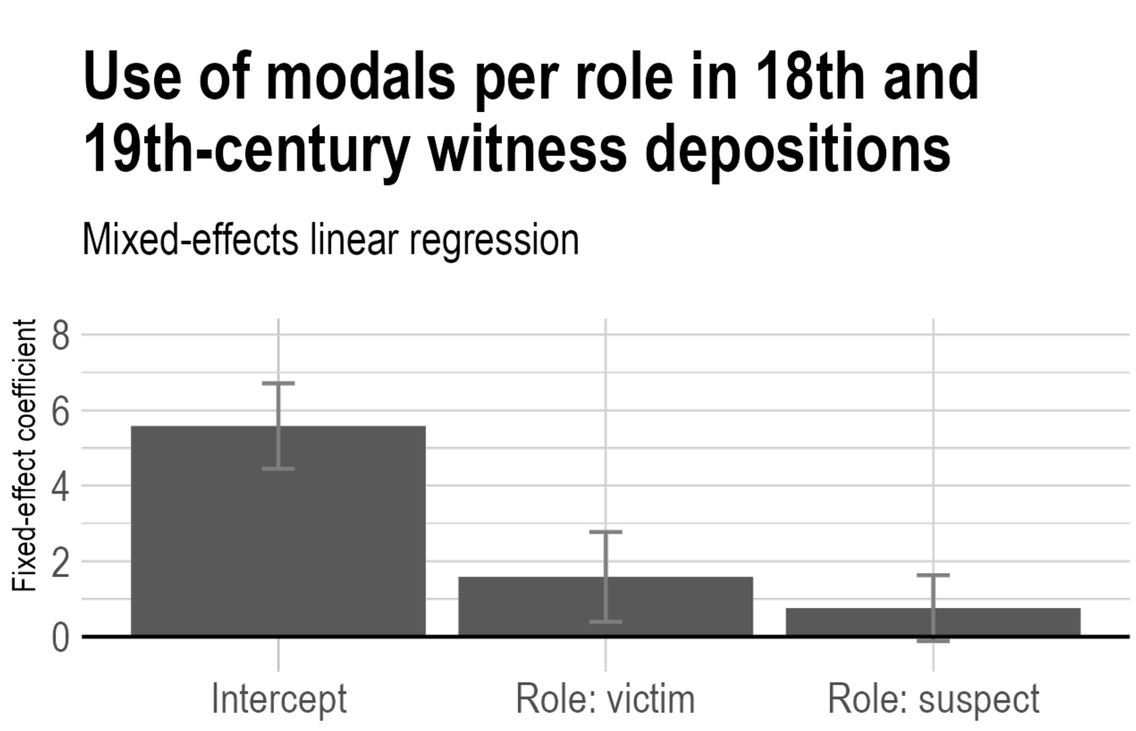
Data: Letters as Loot Corpus



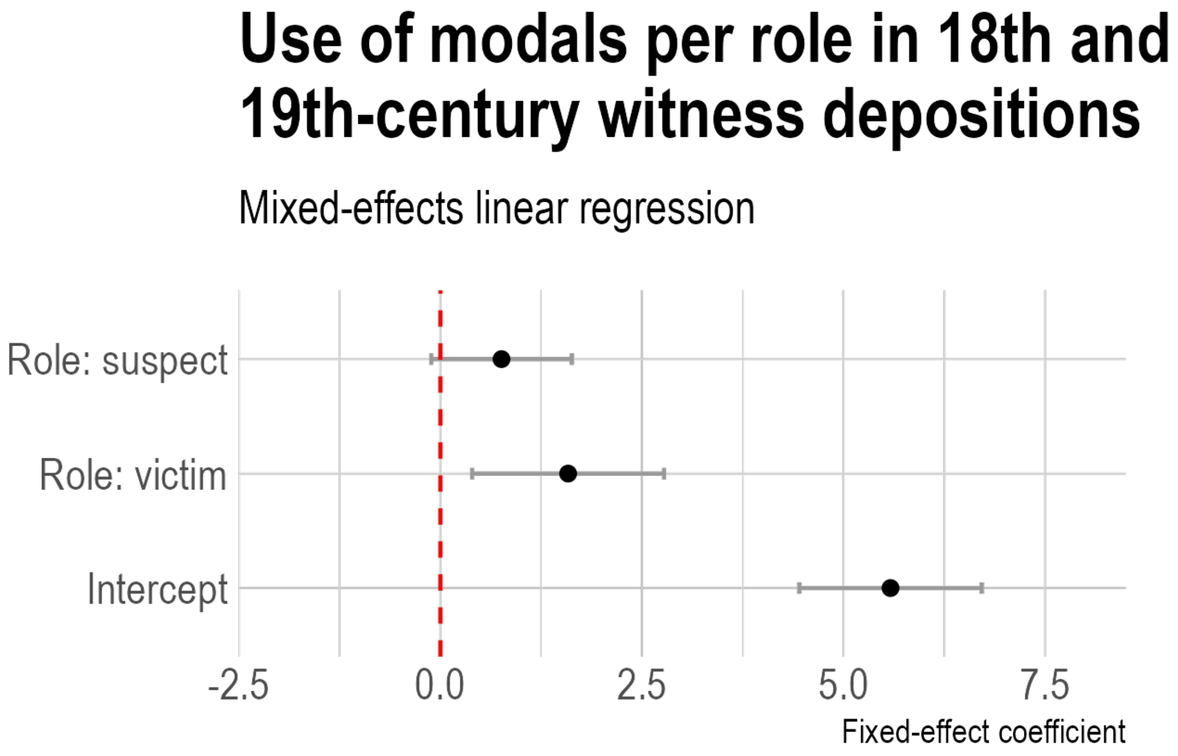
Data: Letters as Loot Corpus

Data to ink

Adding a measure of dispersion or uncertainty — CI 95%



Reference level - Role: witness
Whiskers represent 95% CI
Data: Serwadczak (in prep.)



Reference level - Role: witness
Whiskers represent 95% CI
Data: Serwadczak (in prep.)

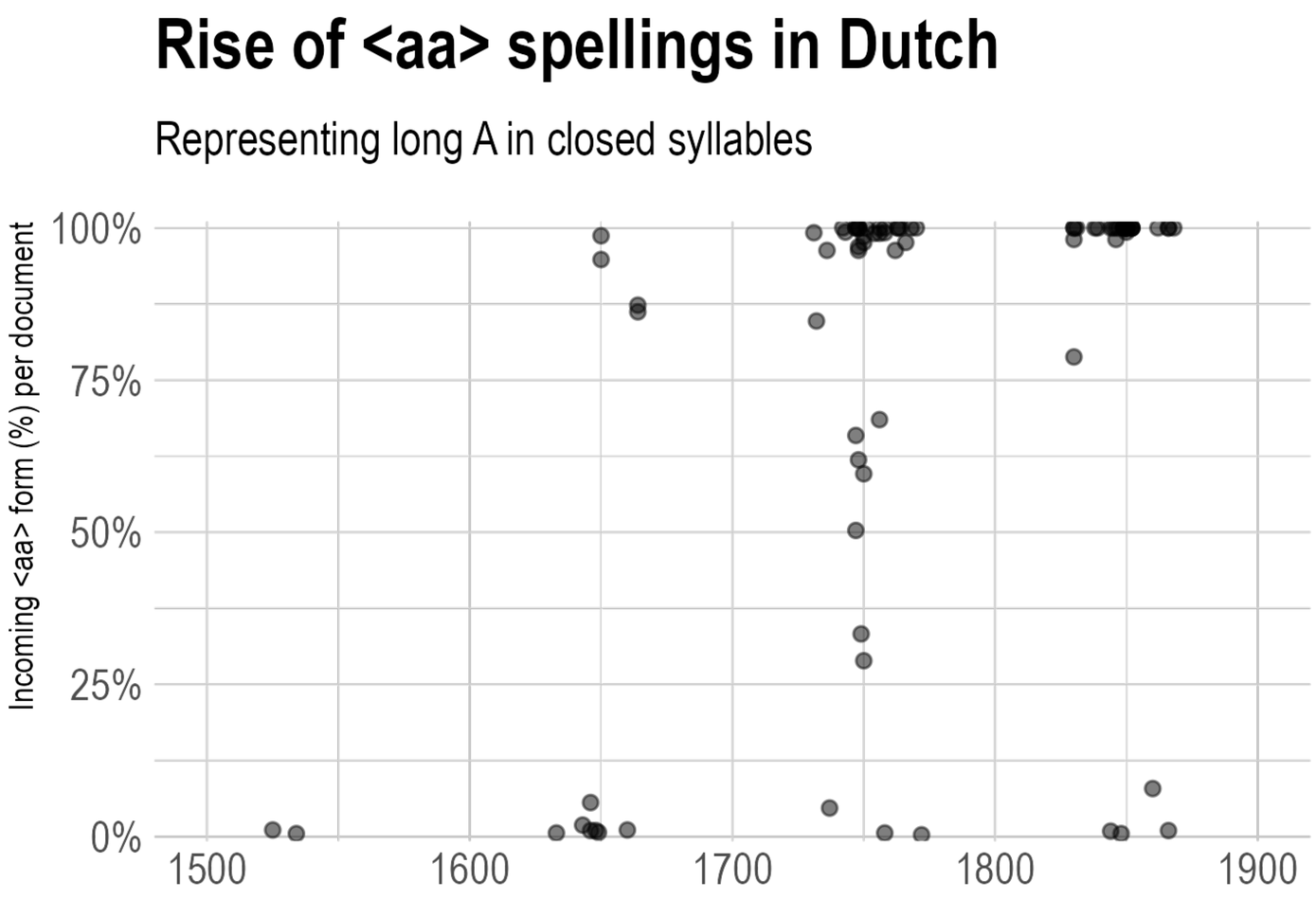
Data to ink

1. Whiskers as extra information
2. Trend lines as extra information

    ~ regression model

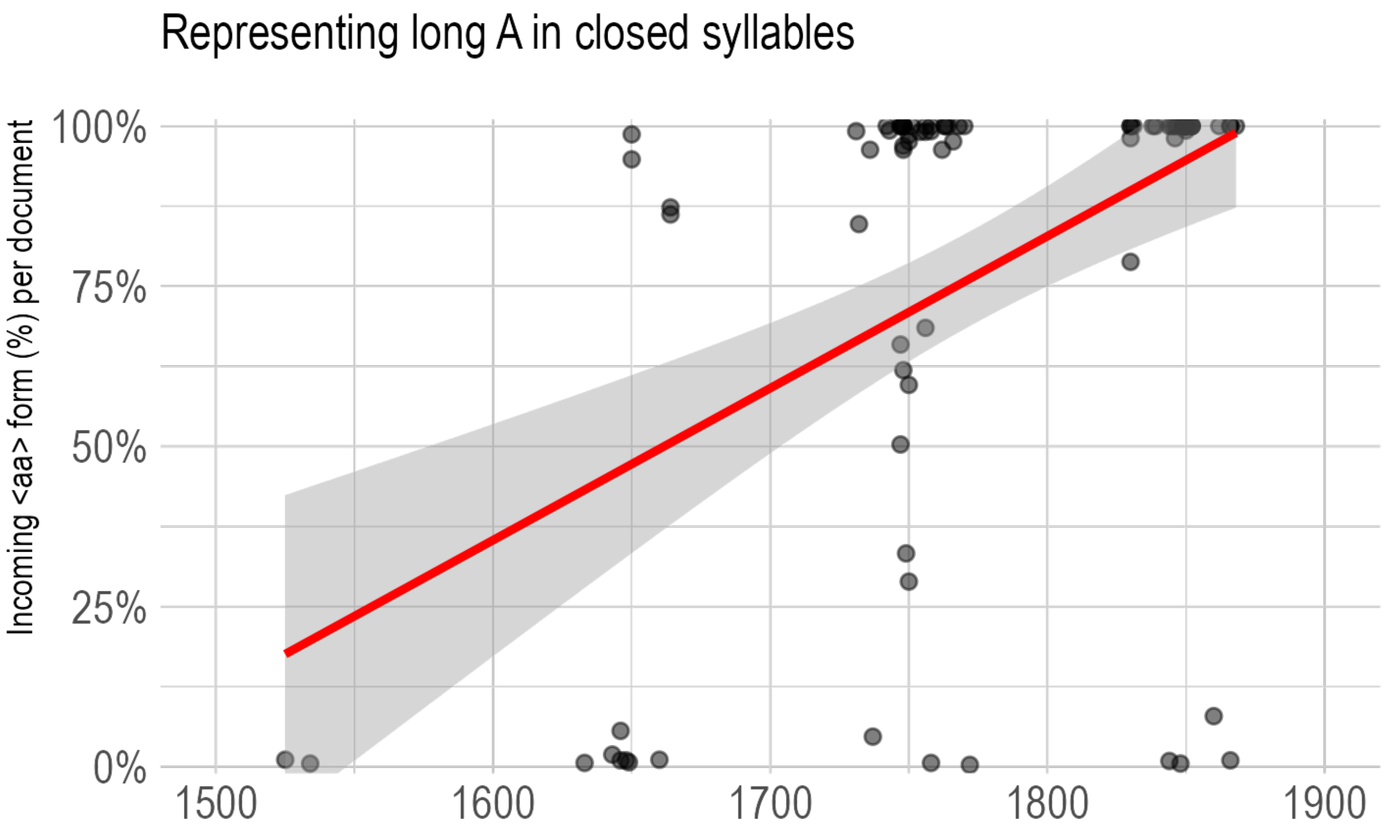
    think about most appropriate graphical summary

Data to ink

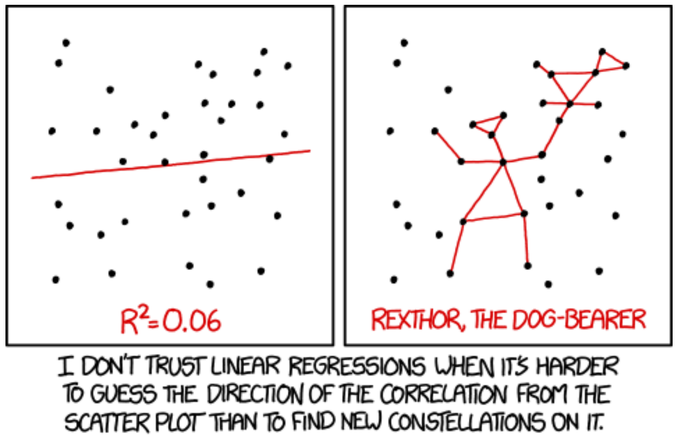


Data: Van de Voorde (2021)

Data to ink

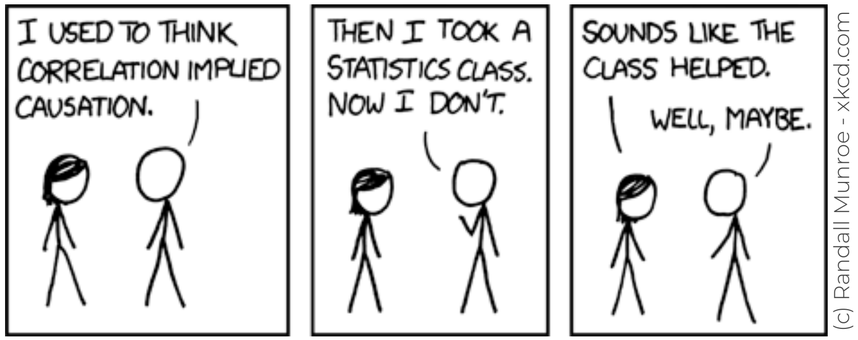


Data: Van de Voorde (2021)
Smoother: linear



(c) Randall Munroe - xkcd.com

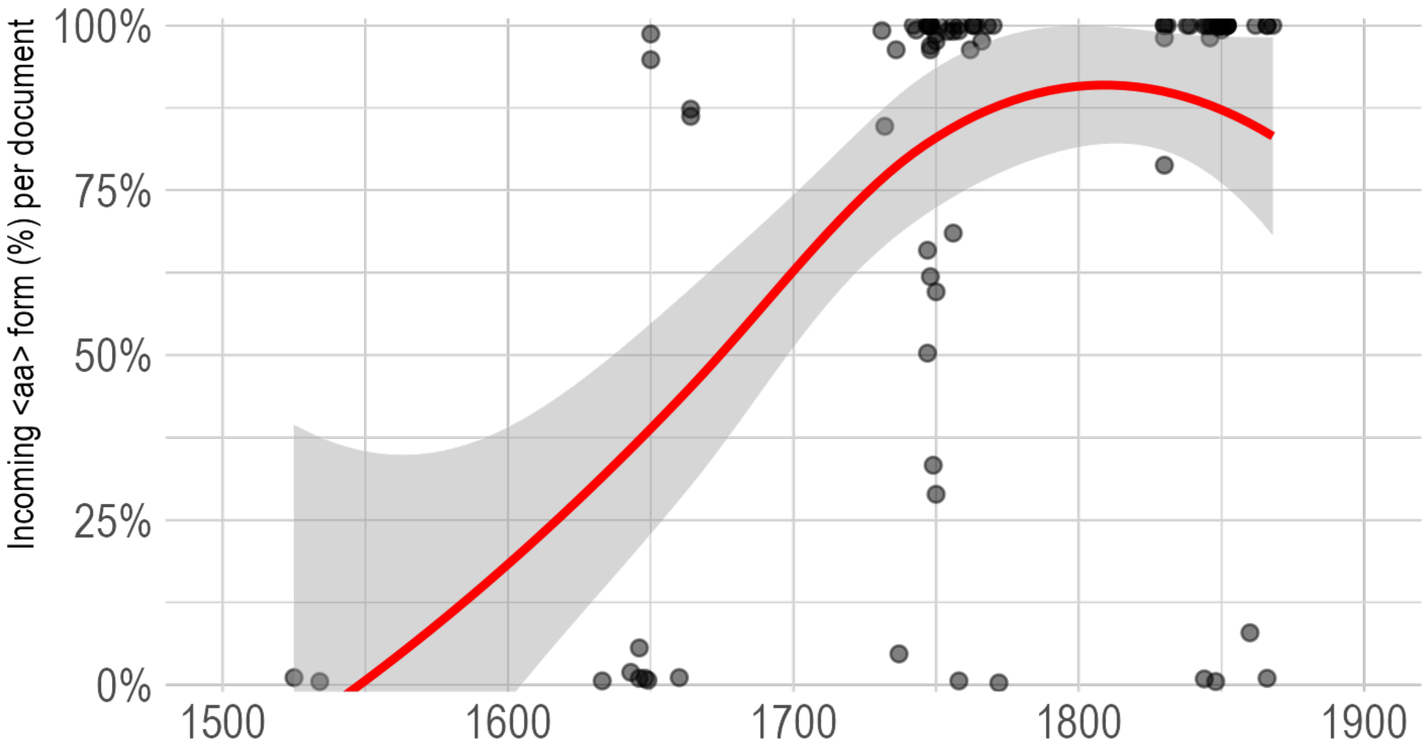
Data to ink



(c) Randall Munroe - xkcd.com

Rise of <aa> spellings in Dutch

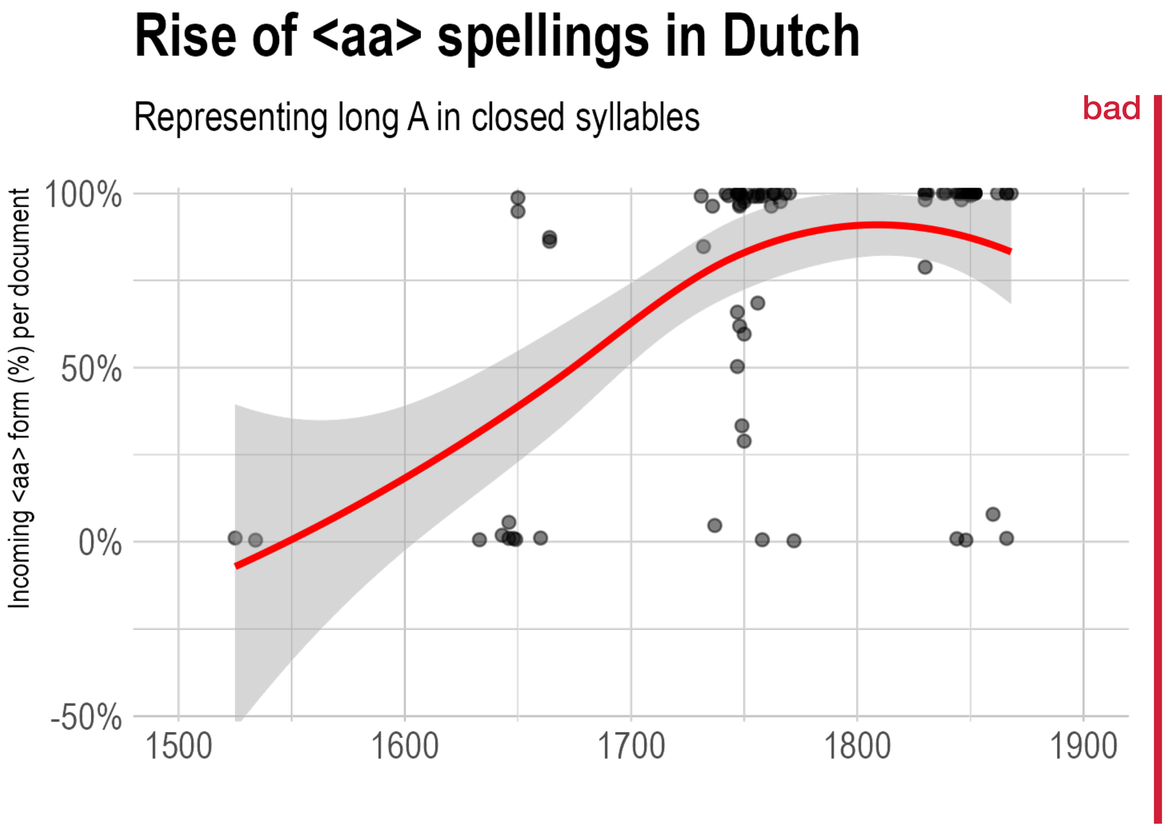
Representing long A in closed syllables



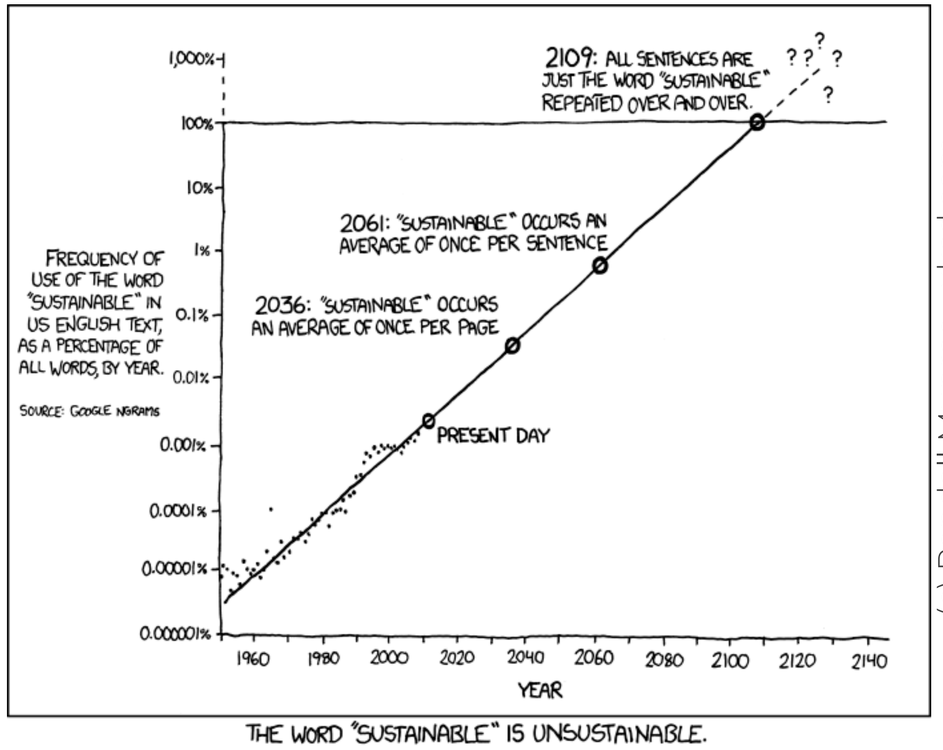
Data: Van de Voorde (2021)
LOESS smoother (span: 1)

Data to ink

Careful: don't let trend lines predict impossible values

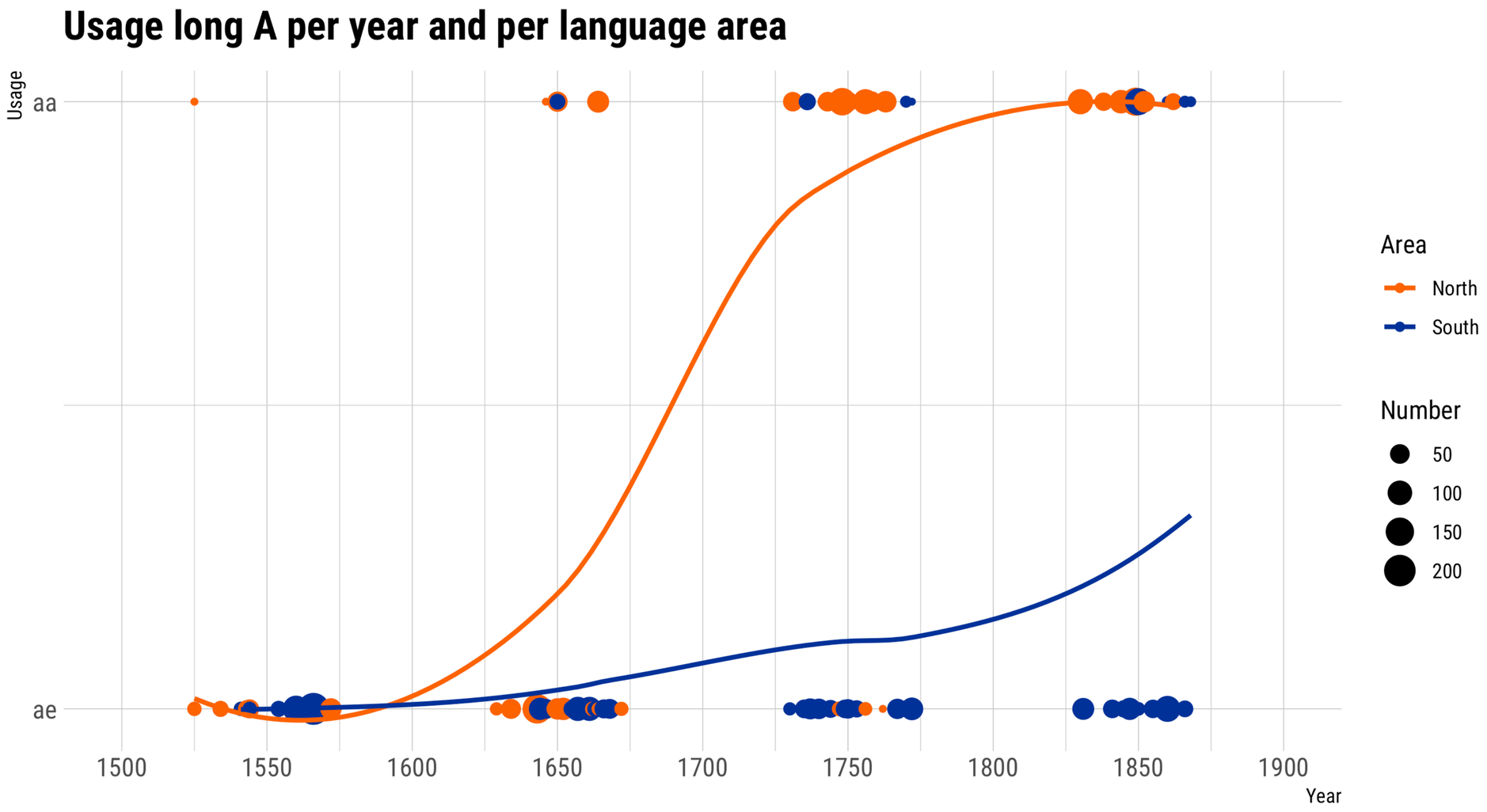


Data: Van de Voorde (2021)
LOESS smoother (span: 1)



(c) Randall Munroe - xkcd.com

Data to ink



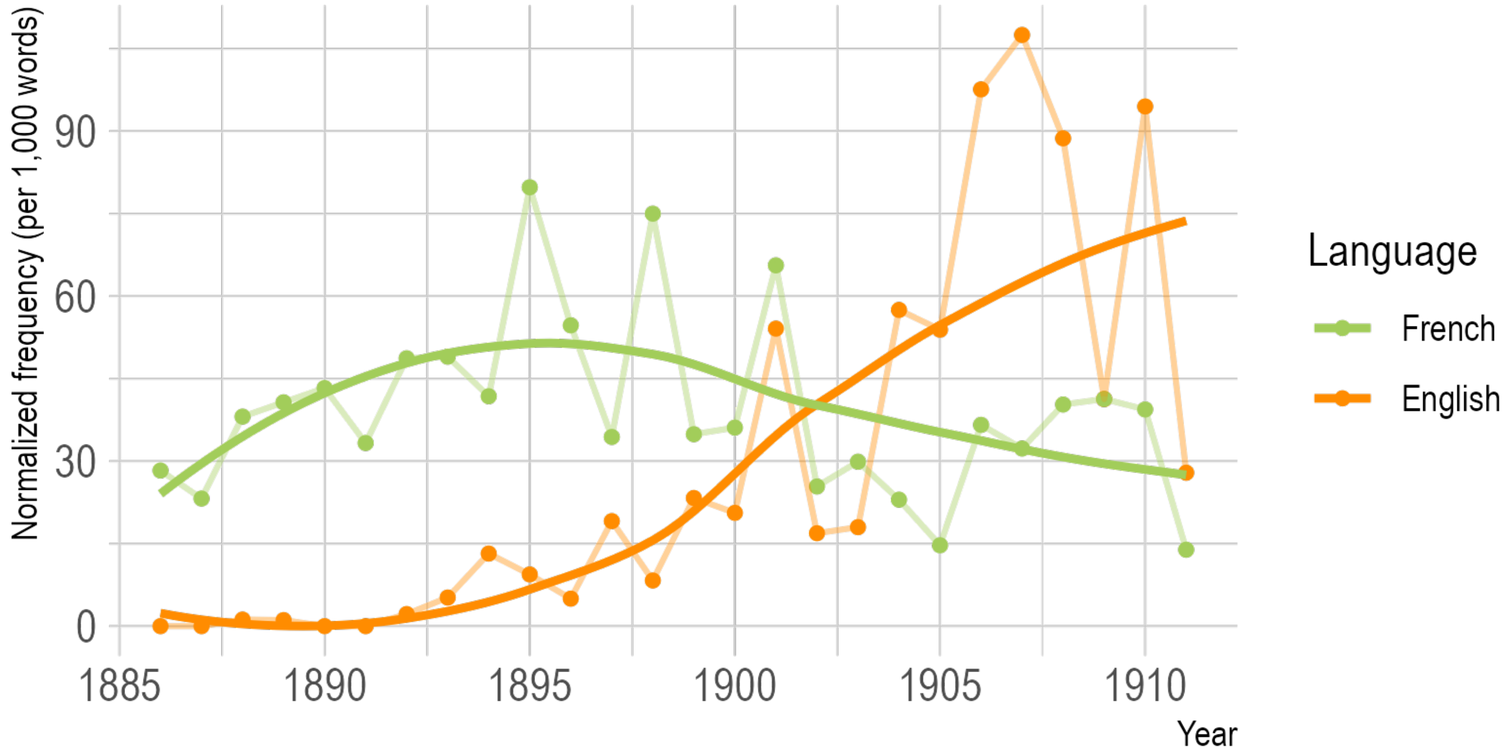
Lismont, Rutten & Vosters (forthcoming)

N=4751; LOESS smoother span: 0.75

Data to ink

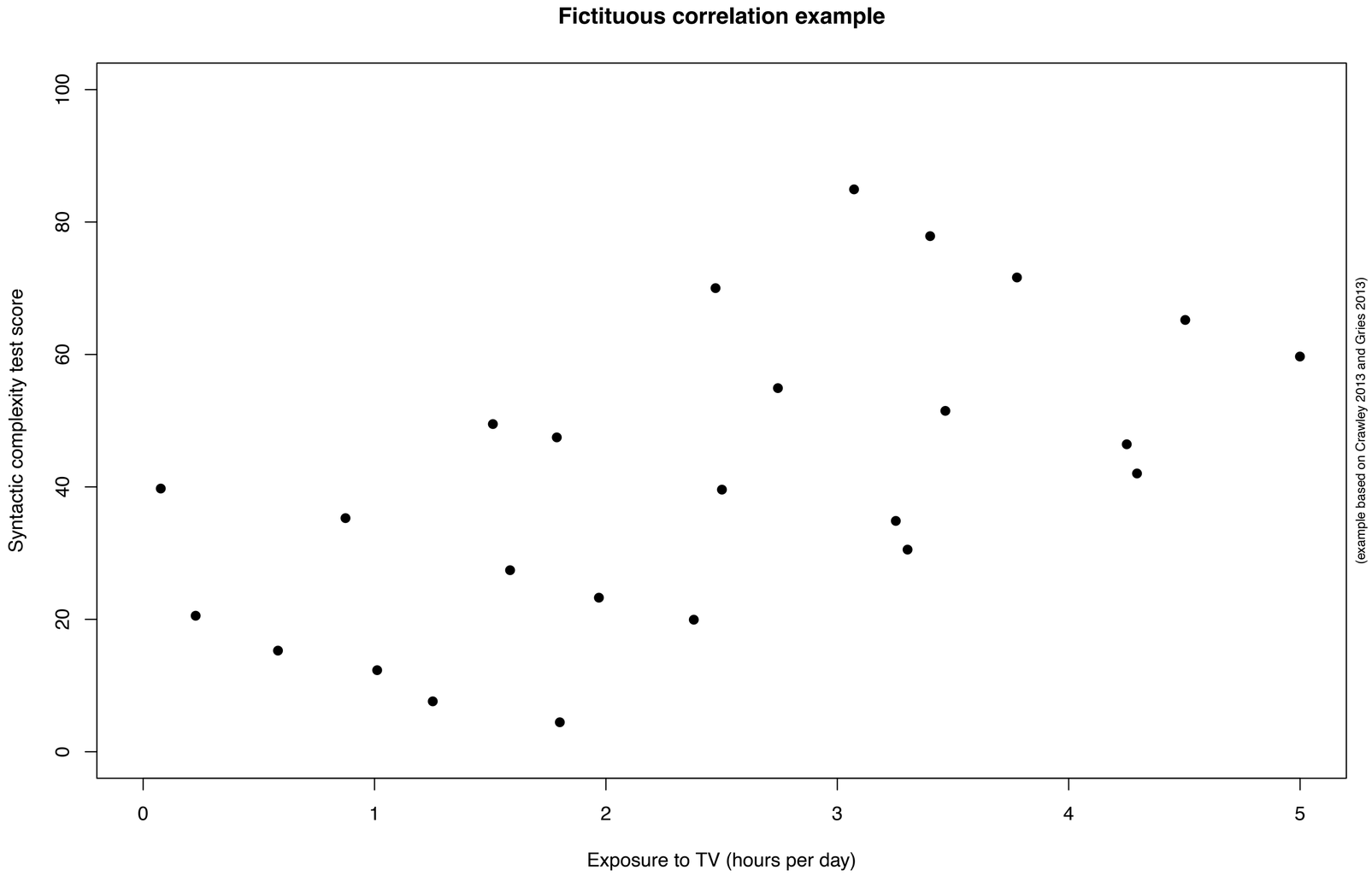
Lexical borrowings

Diary of Brother John, Flemish priest in North America

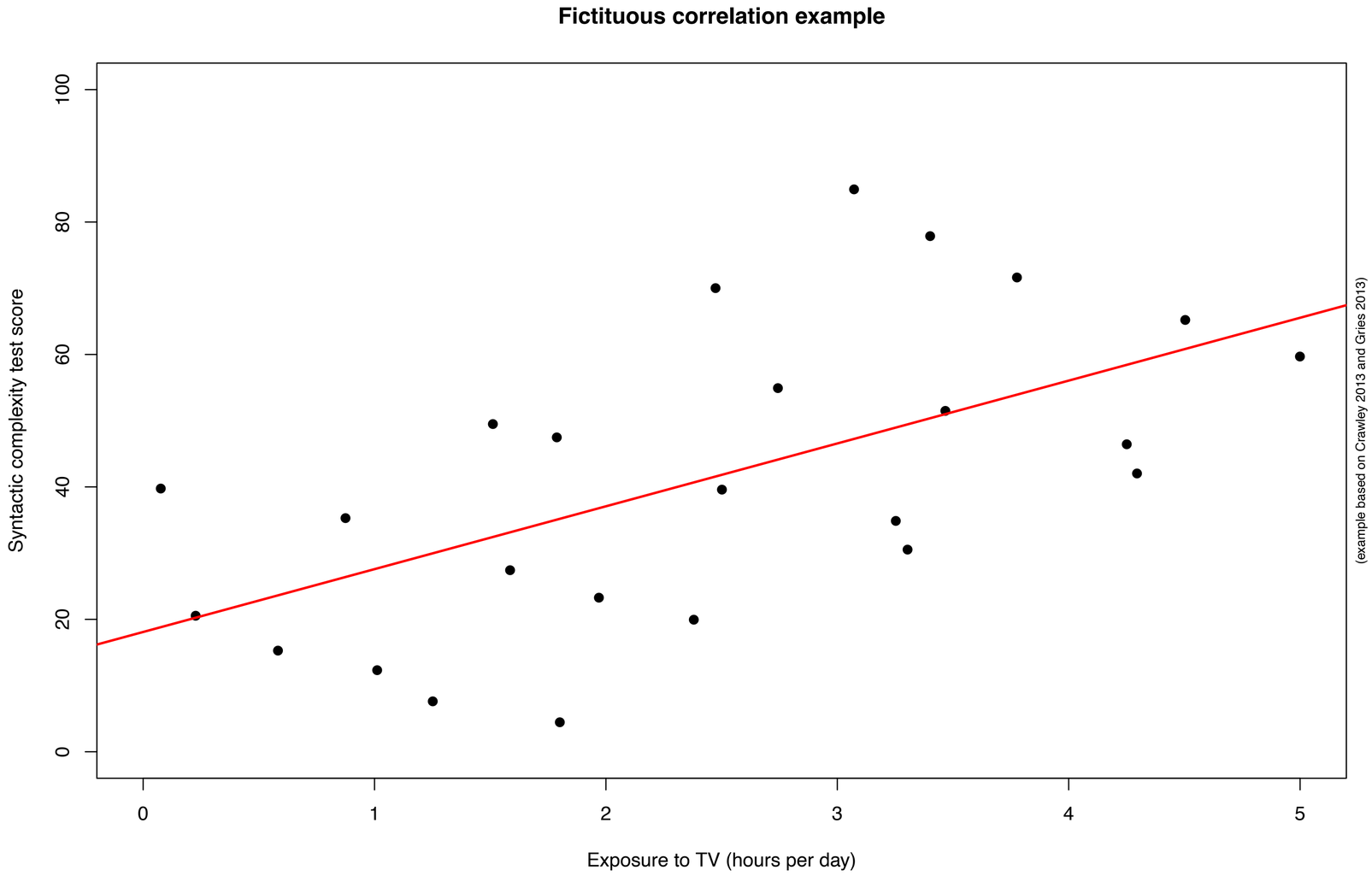


Source: Crombez, Vandenbussche & Vosters (submitted)
LOESS smoother (span: 1)

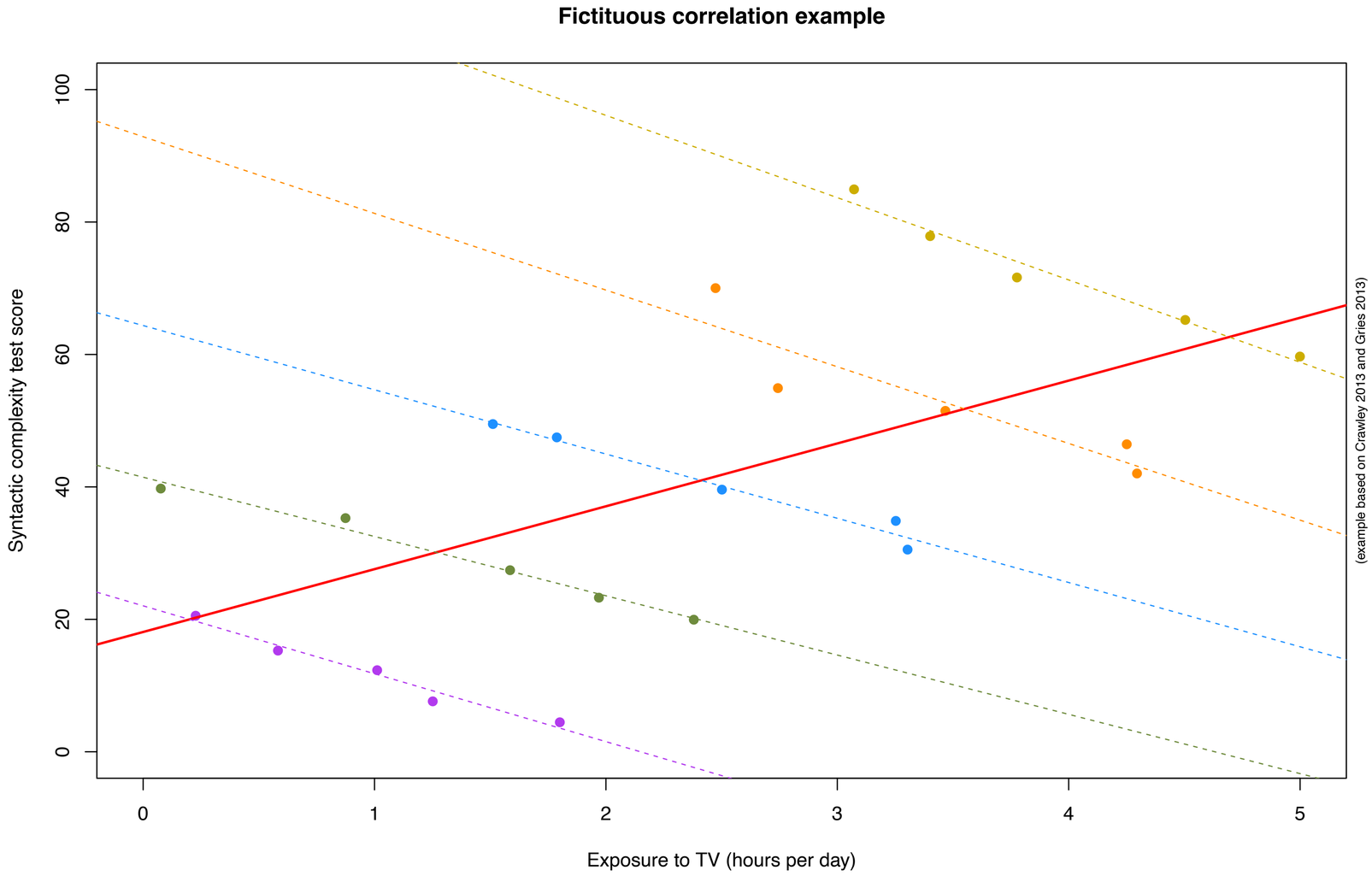
Data to ink


Fictituous correlation example
Syntactic complexity test score
0
20
40
60
80
100
0
1
2
3
4
5
Exposure to TV (hours per day)
(example based on Crawley 2013 and Gries 2013)

Data to ink


Fictituous correlation example
Syntactic complexity test score
0
20
40
60
80
100
0
1
2
3
4
5
Exposure to TV (hours per day)
(example based on Crawley 2013 and Gries 2013)

Data to ink


Fictituous correlation example
Syntactic complexity test score
0
20
40
60
80
100
0
1
2
3
4
5
Exposure to TV (hours per day)
(example based on Crawley 2013 and Gries 2013)

Proportional ink

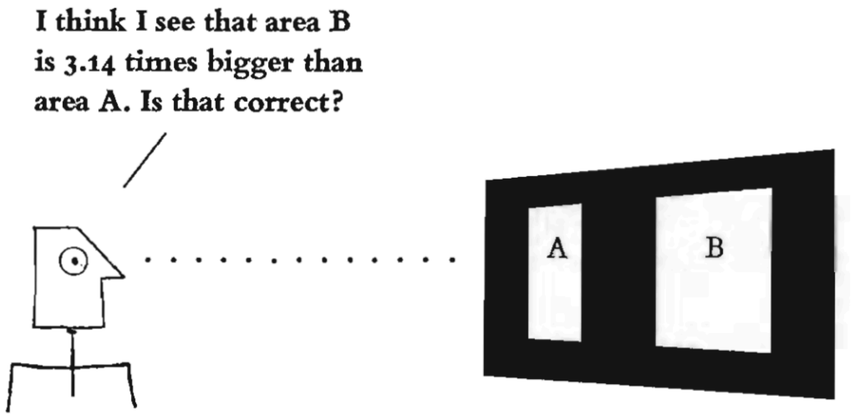
2. Principle of proportional ink

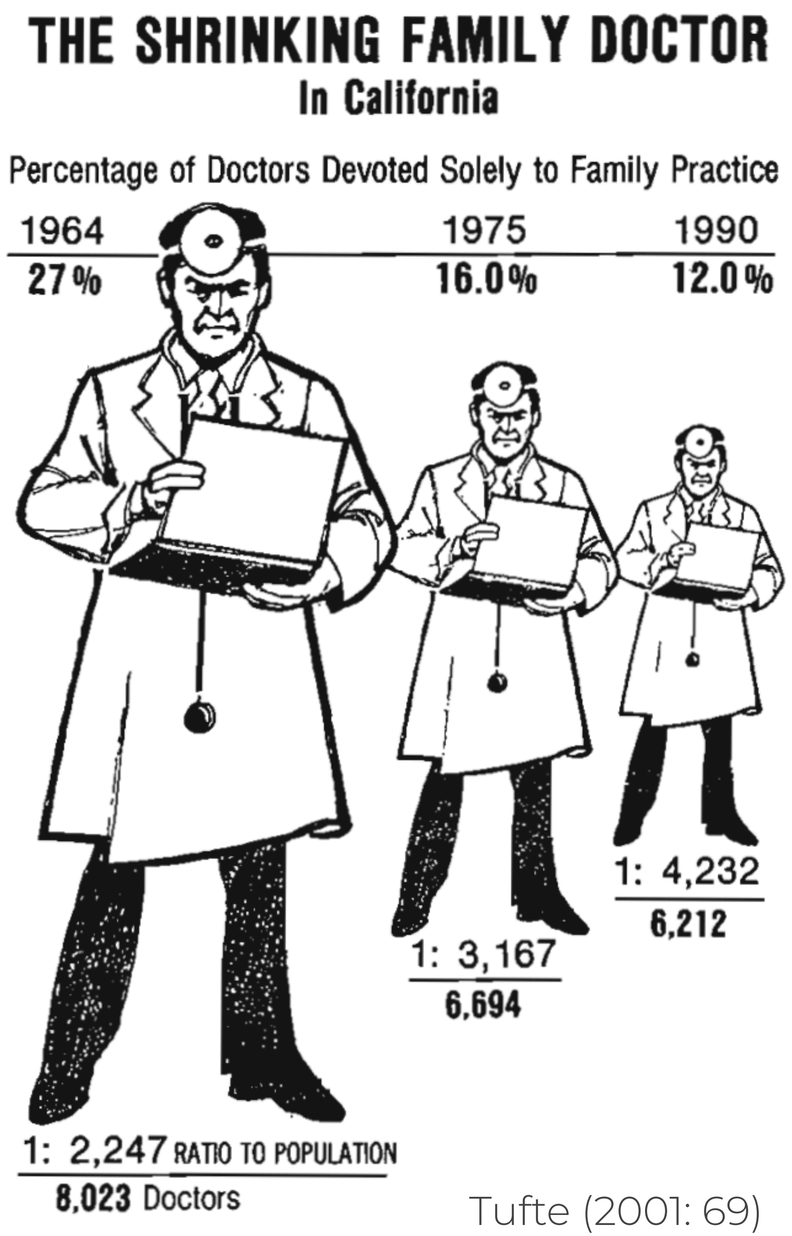
   Tufte (2001: 77):

   "The **representation** of numbers, as physically measured on the surface of the graphic itself, should be directly **proportional** to the numerical **quantities represented**"

DESIGN PRINCIPLES

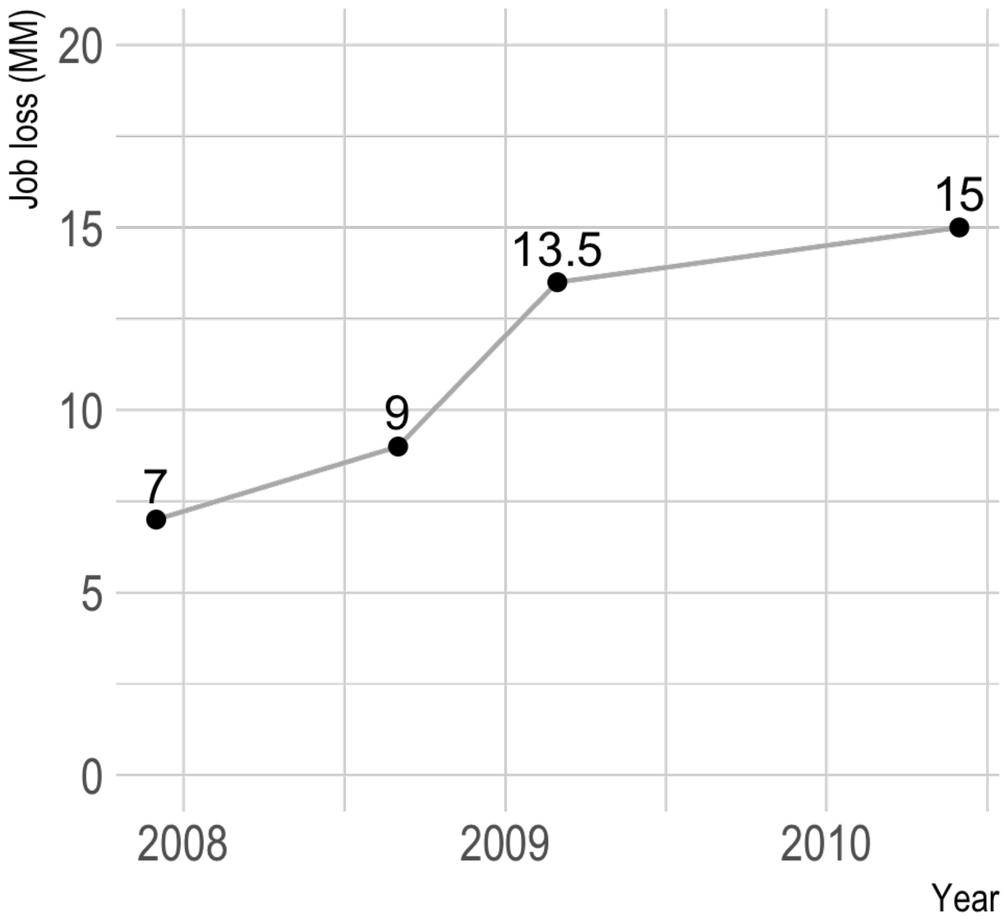
Proportional ink





Tufte (2001: 69)

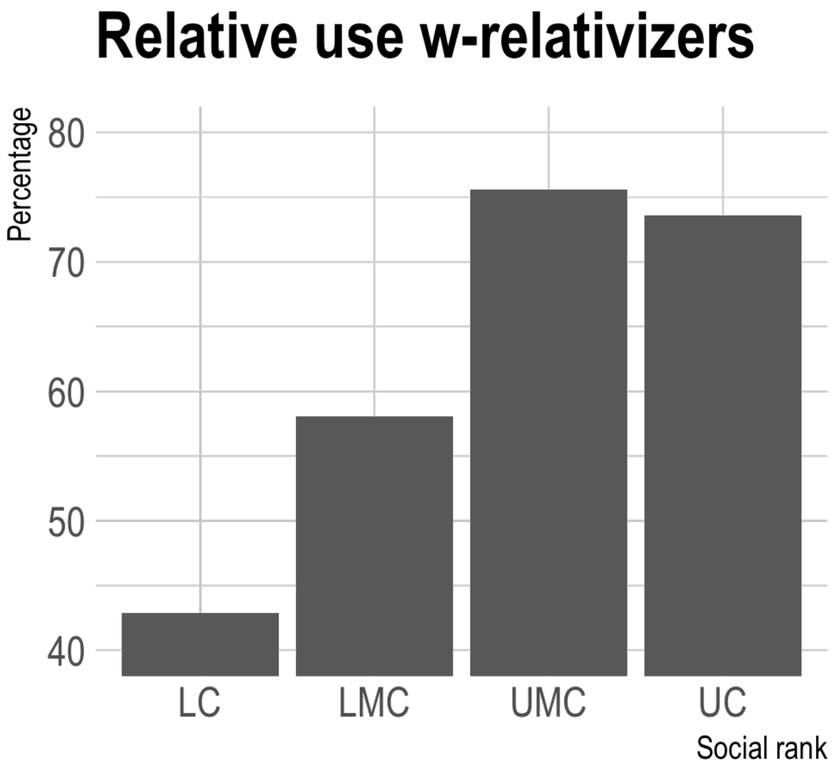
Proportional ink



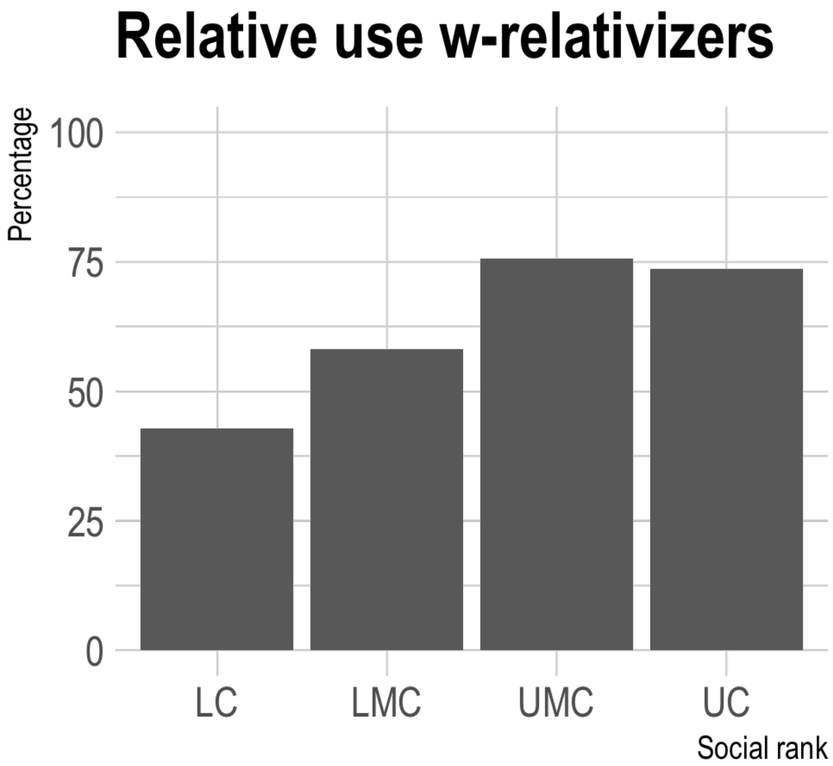
Proportional ink

2. Principle of proportional ink

But also:



Data: Rutten & Van der Wal (2014: 298)

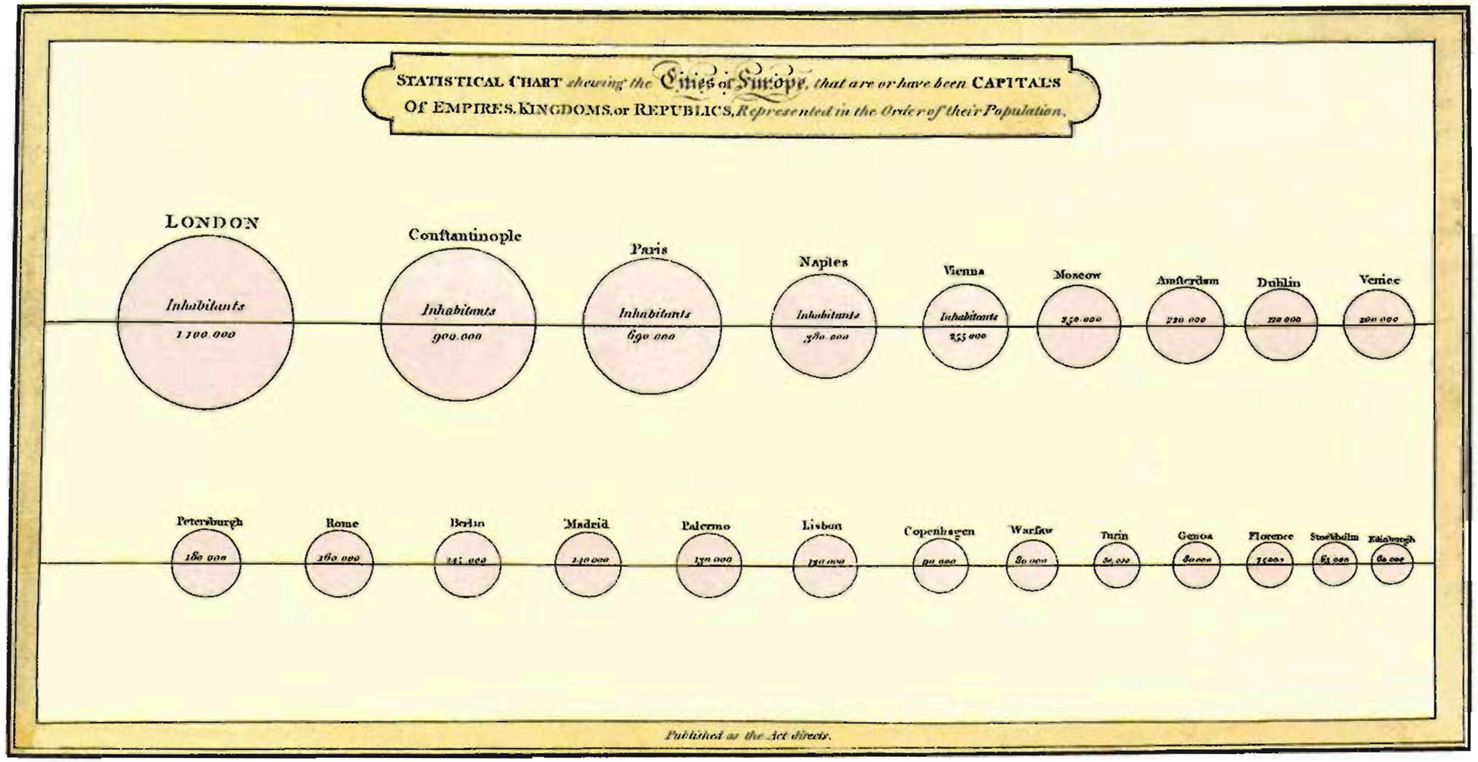


Data: Rutten & Van der Wal (2014: 298)

Proportional ink

2. Principle of proportional ink

   ~ Do not use **two-dimensional** visuals to represent **one-dimensional** data



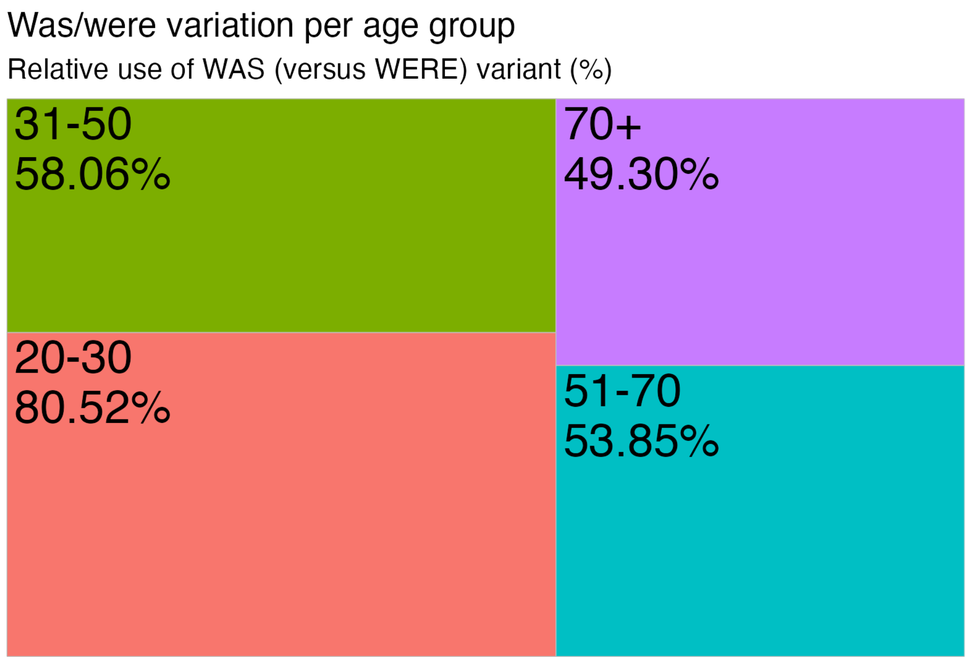
Tufte (2001: 73)

Proportional ink

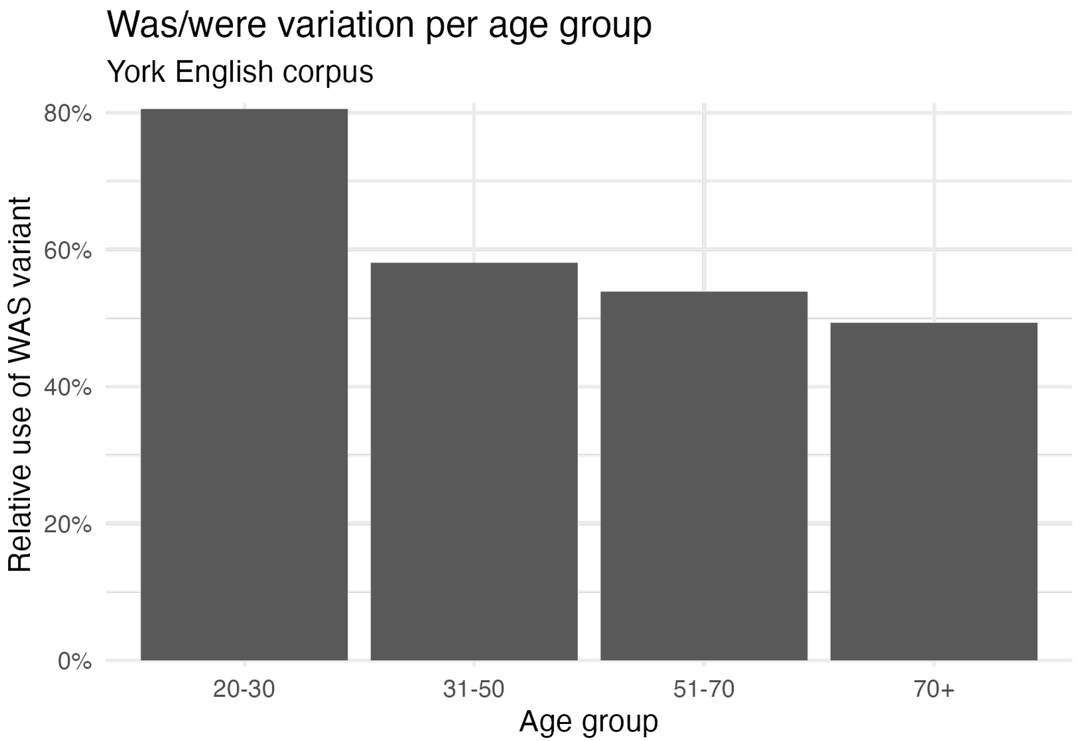
2. Principle of proportional ink

~ Do not use **two-dimensional** visuals to represent **one-dimensional** data

In any case, human perception: judging distances > judging areas



(Data: Tagliamonte & Baayen 2012)

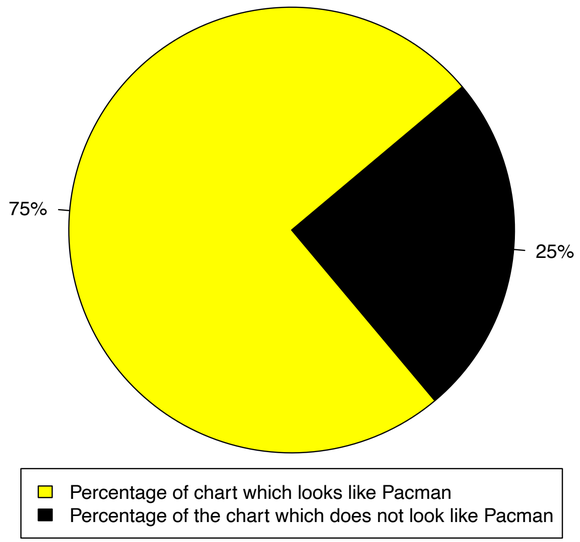


(Data: Tagliamonte & Baayen 2012)

Proportional ink

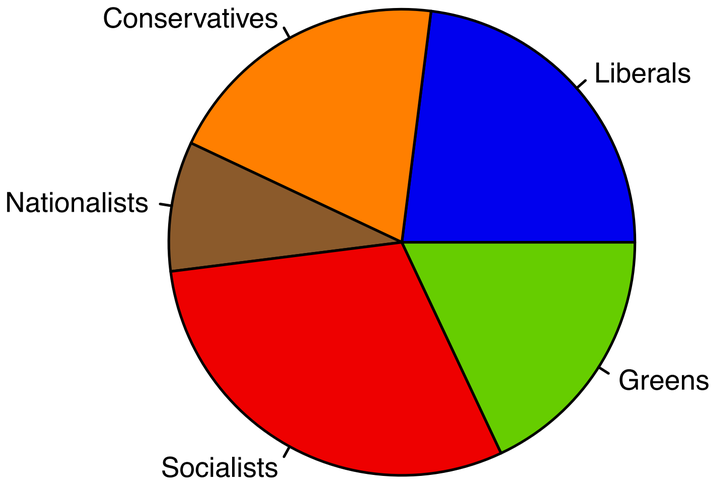
2. Principle of proportional ink

Pie charts often not useful

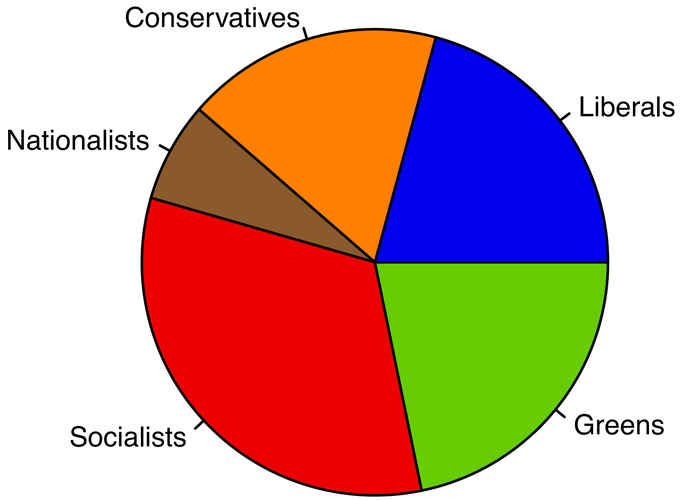


Fictitious election results

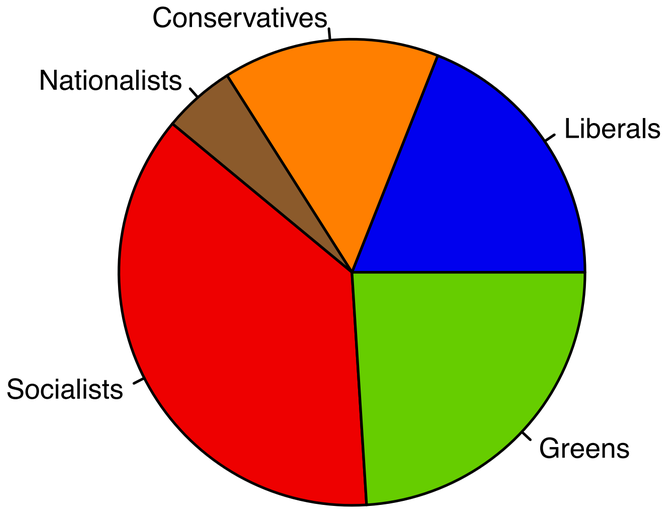
Pie charts



2005 election results



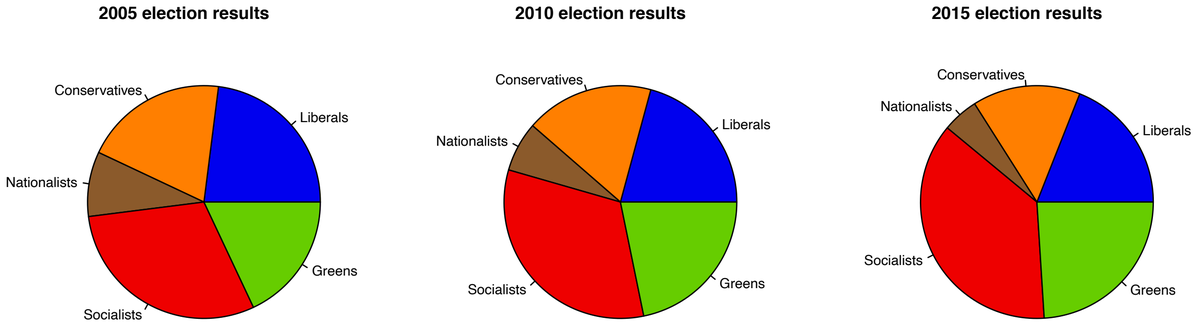
2010 election results

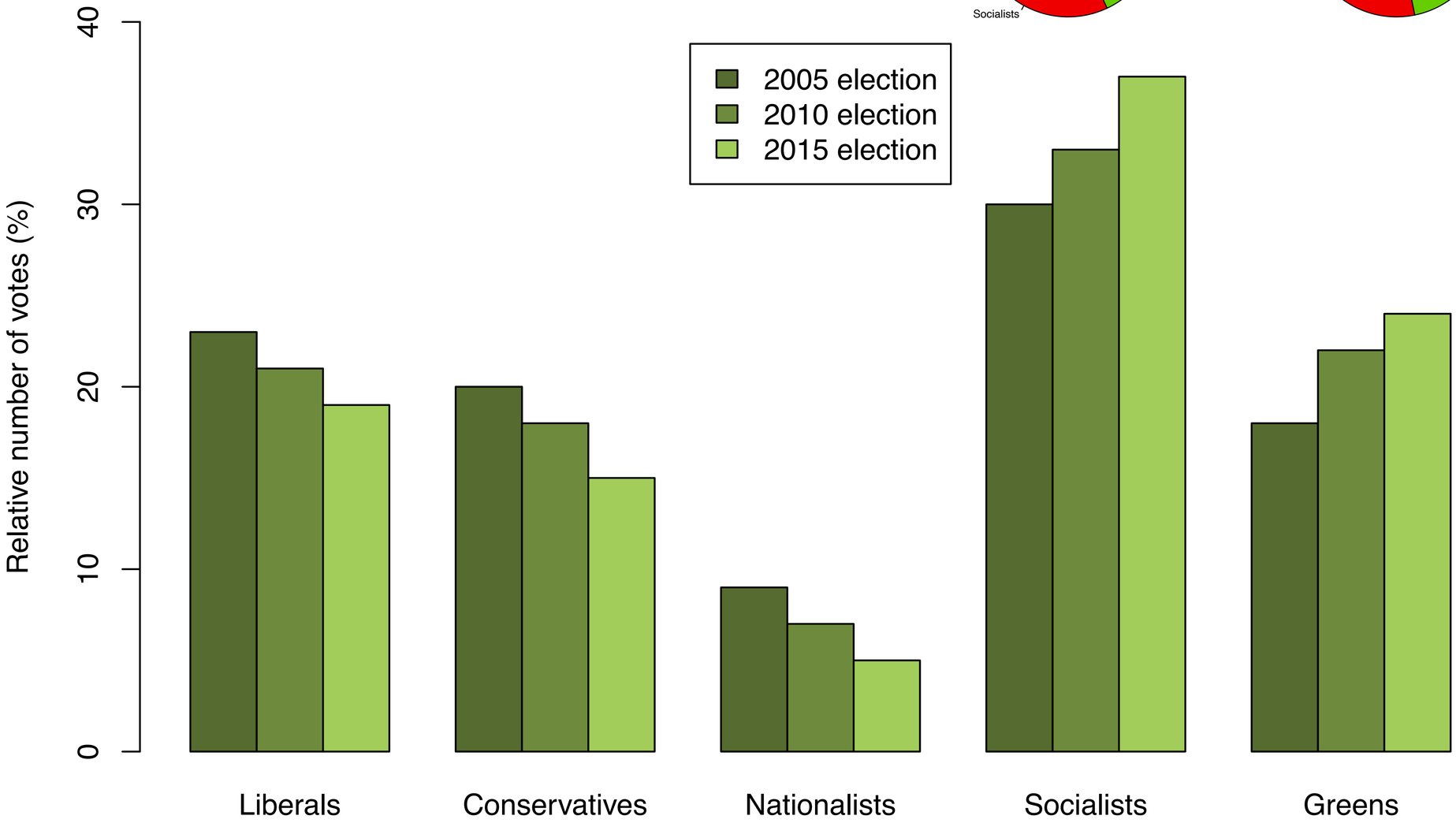


2015 election results

Ficticious election results

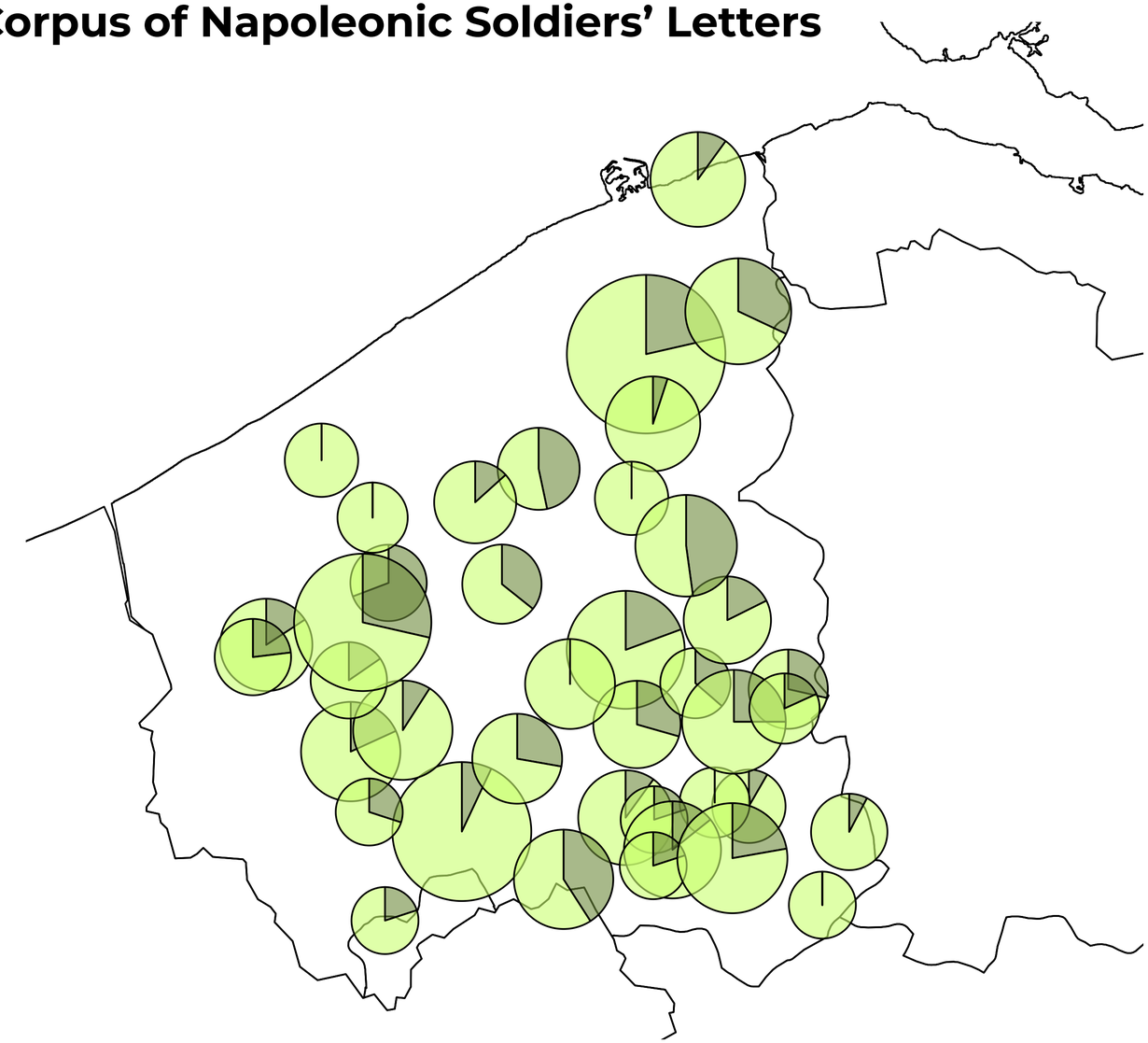
Barplots





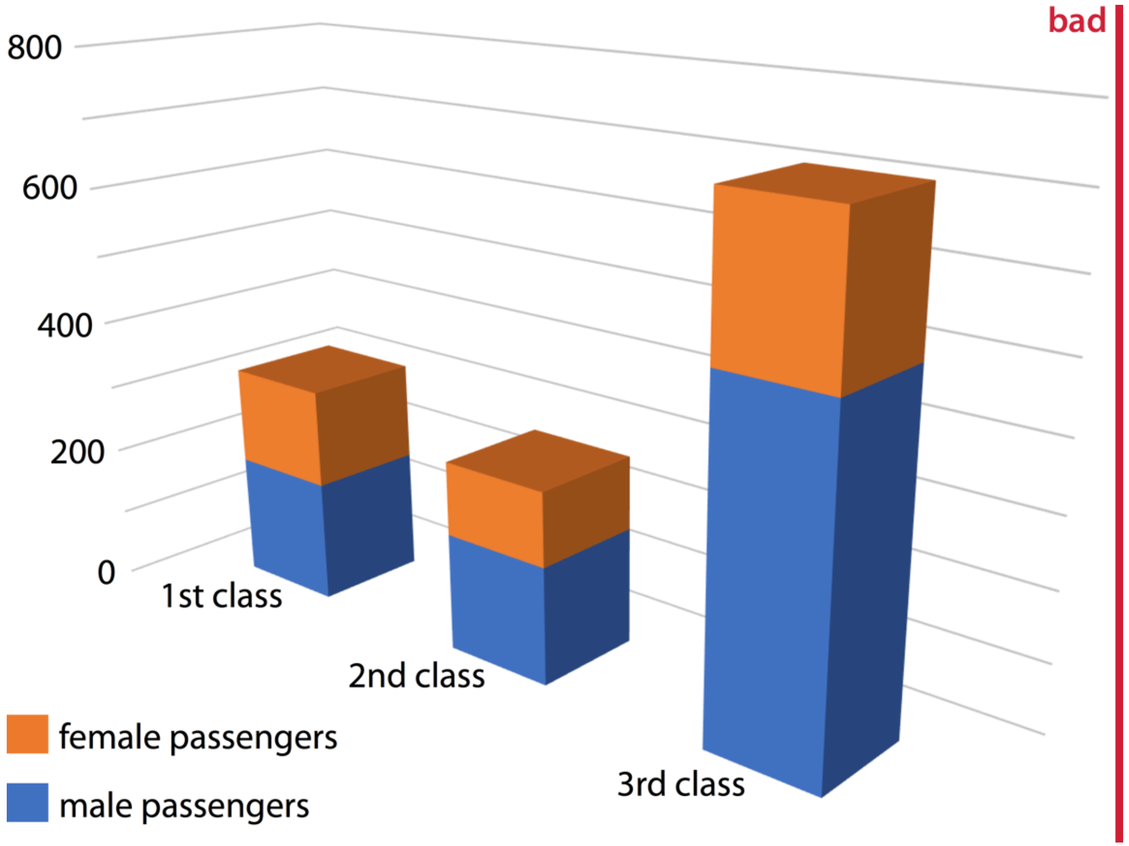
Bipartite versus single negation per locality

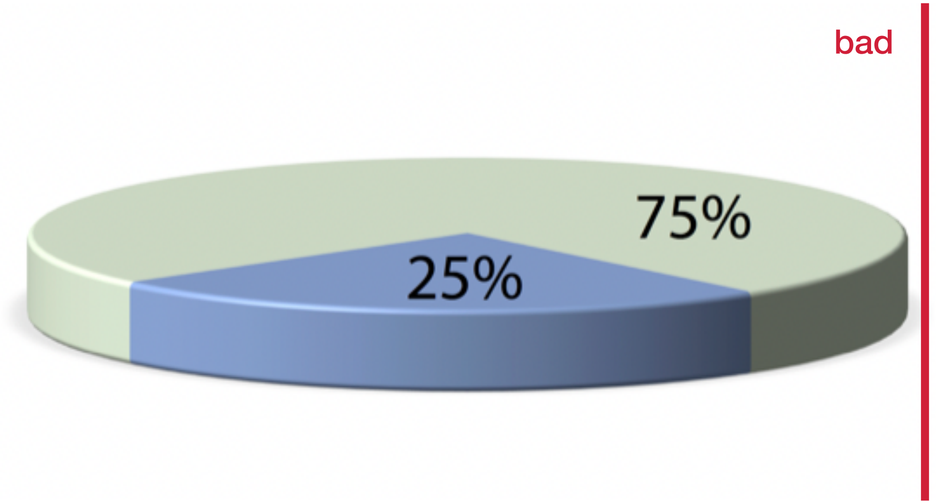
Corpus of Napoleonic Soldiers' Letters



Only including localities with 10 or more tokens

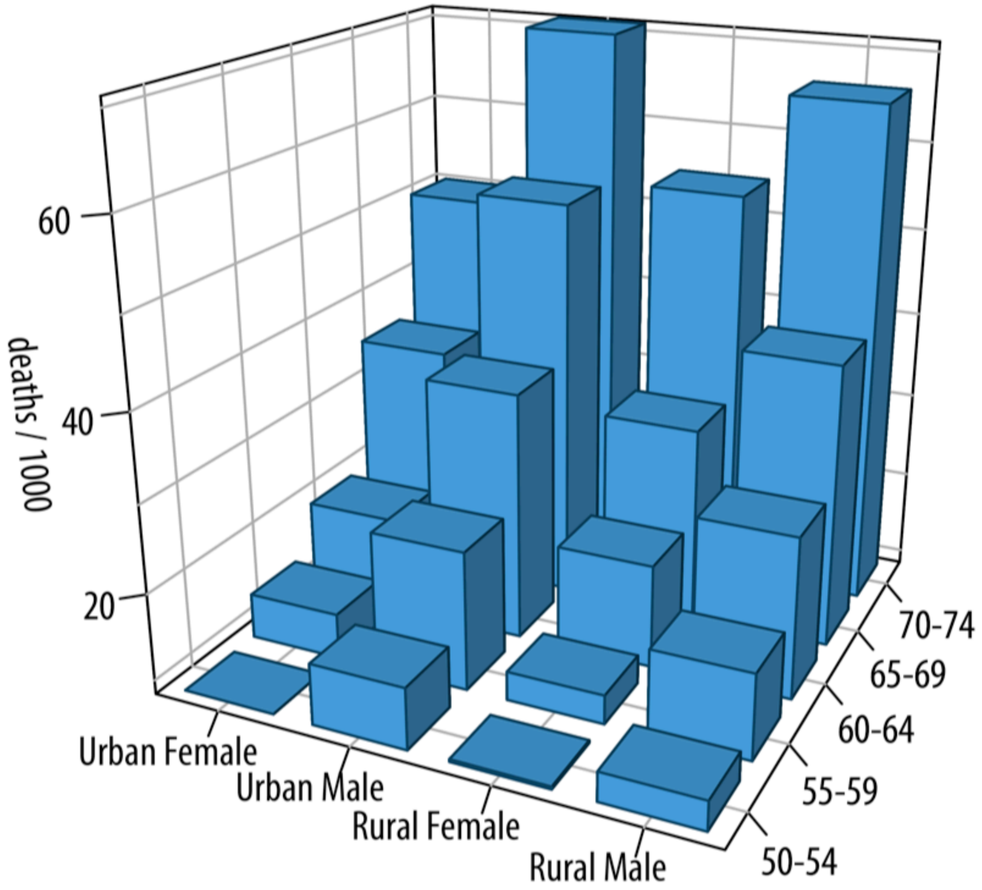
Proportional ink





(Wilke 2019)

Proportional ink

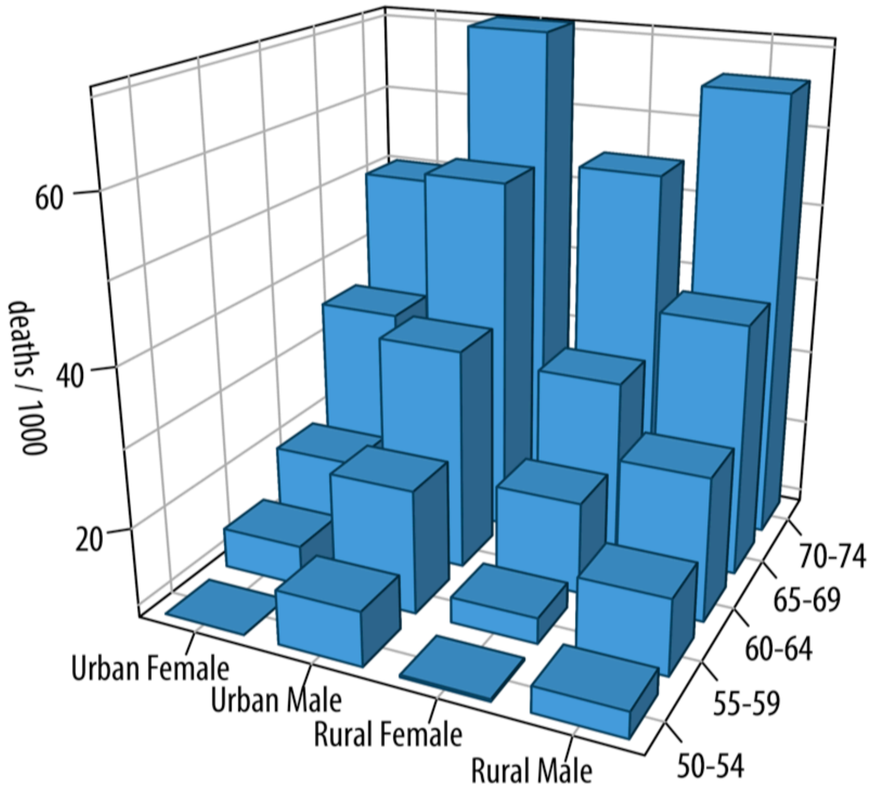


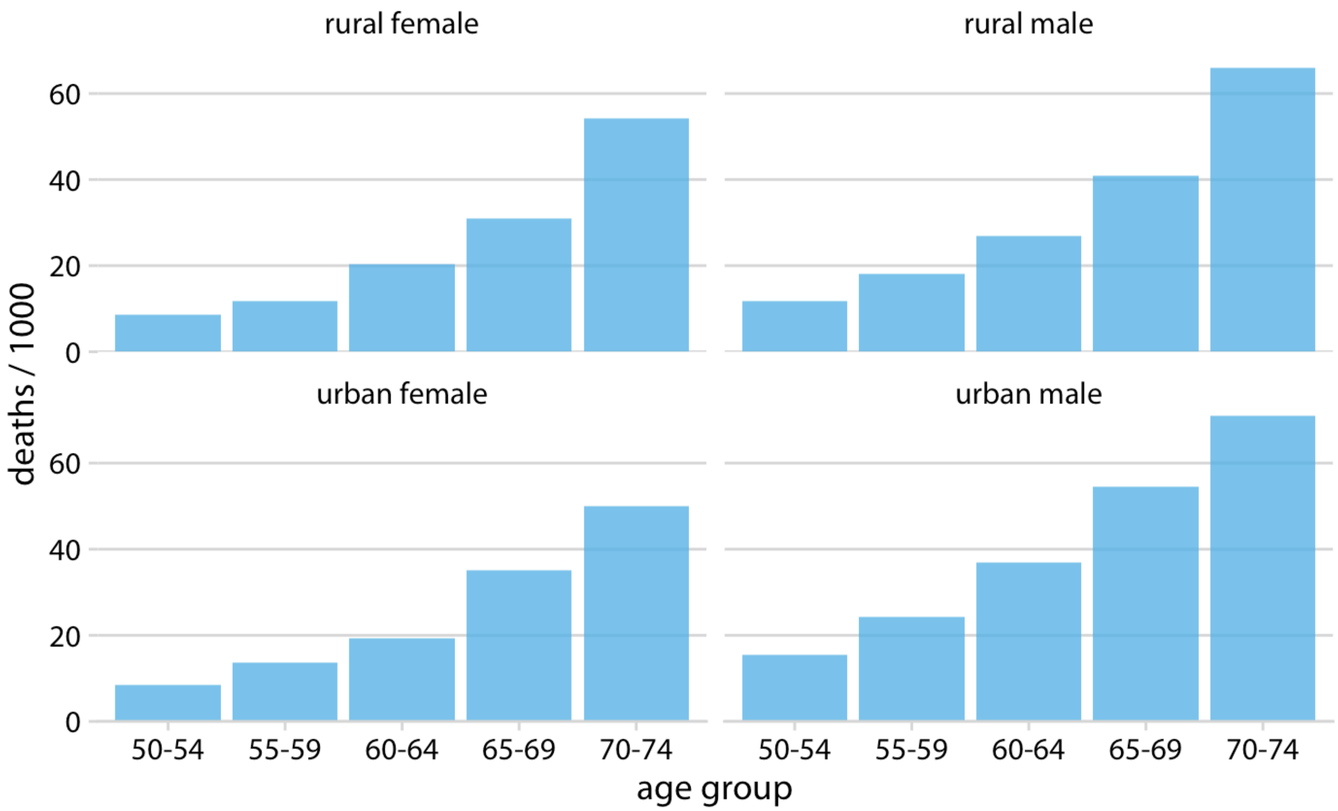
bad

(Wilke 2019)

Proportional ink

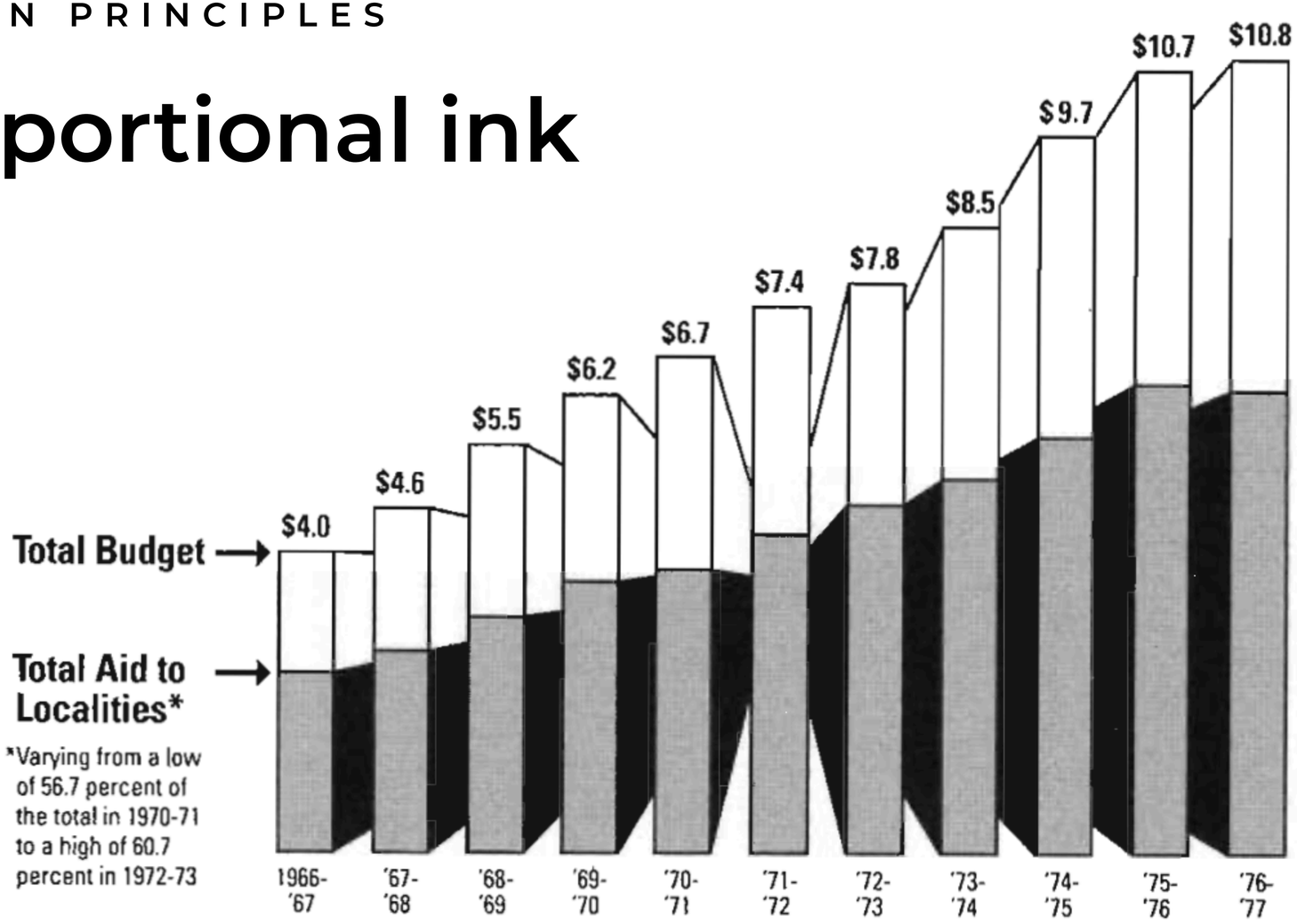
bad





(Wilke 2019)

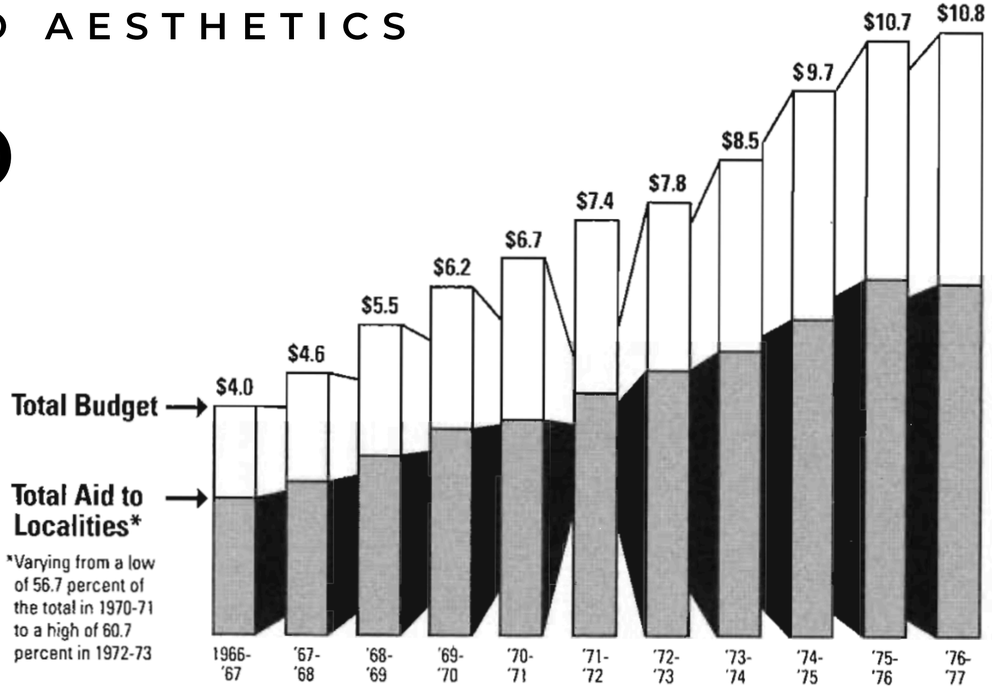
Proportional ink



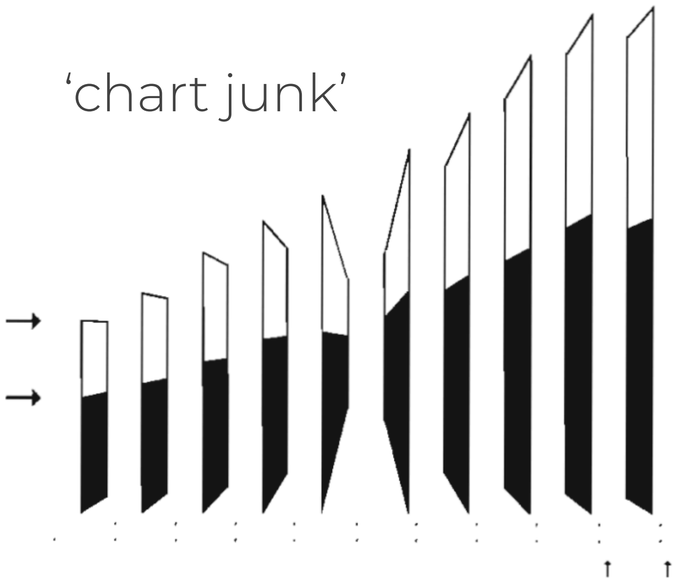
(Tufte 2001: 67)

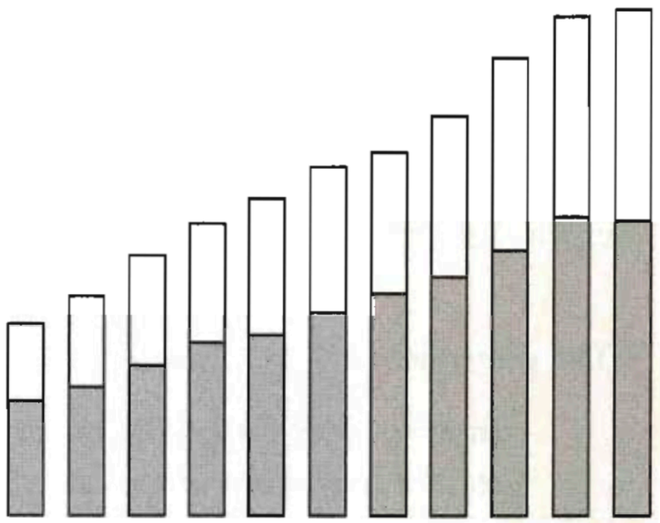
Tufte (2001: 71):
Do not use **two or three dimensions** to represent **one-dimensional** data

Avoid 3D



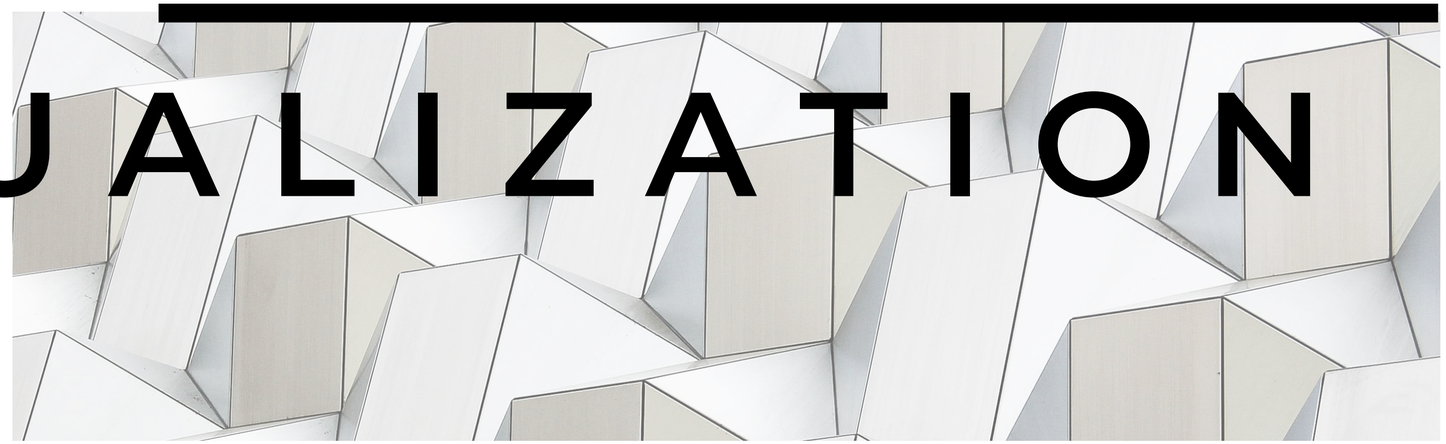
'chart junk'





(Tufte 2001: 67)

DATA VISUALIZATION

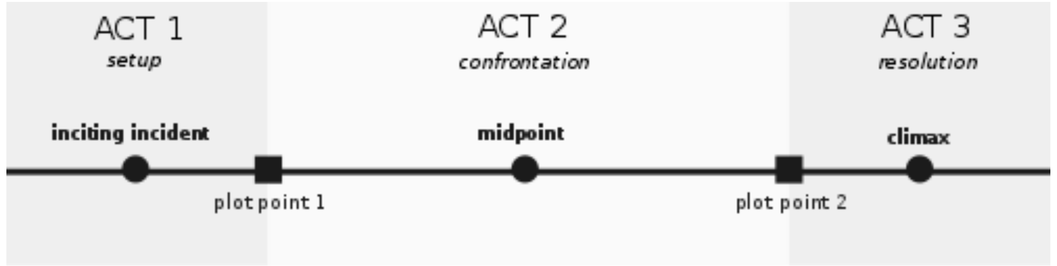


STORYTELLING

Introduction

Purpose of data viz: communication

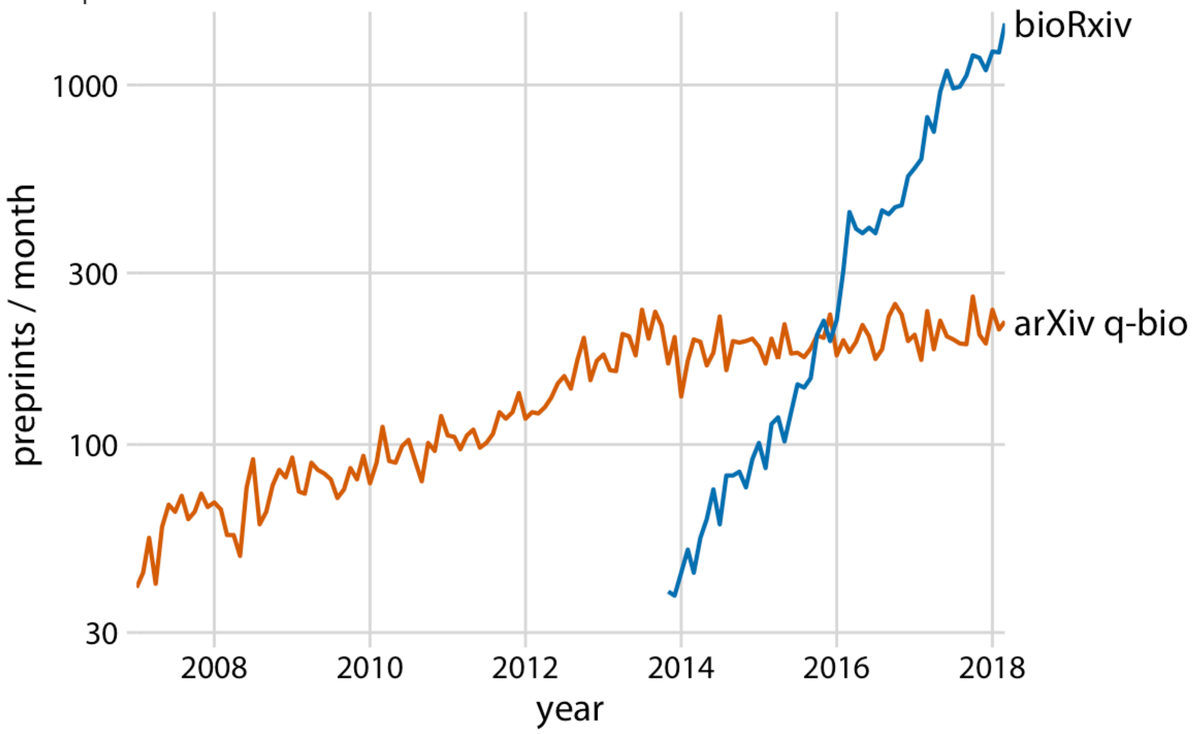
- Capture an audience with an interesting and meaningful story
- Story: set of observations presented in a specific way or order, triggering an emotional response (Wilke 2019)
  - just one visualization is often not sufficient
  - 3-Act Structure: Setup/Lead > Confrontation/Development > Resolution
    Background Challenge Resolution
- Aristotle: δέσις → λύσις



Example

Wilke (2019): preprint submission to arXiv.org in quantitative biology

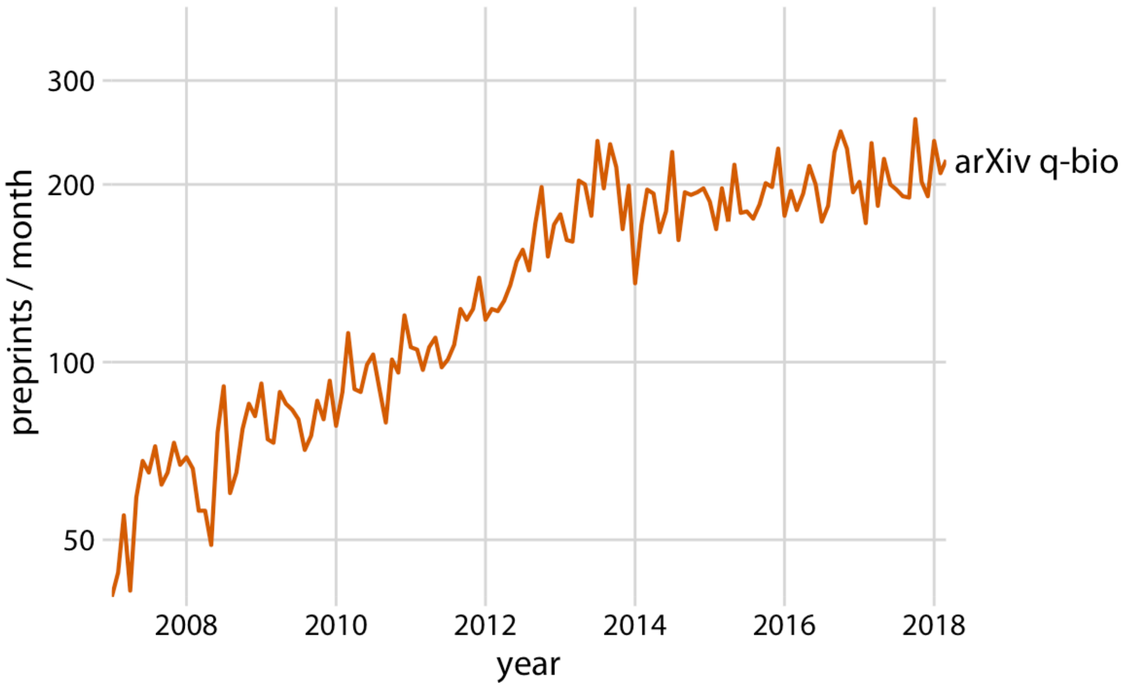
- grew exponentially until 2013
- then, a biology-specific preprint server bioRxiv was launched by a respected publisher

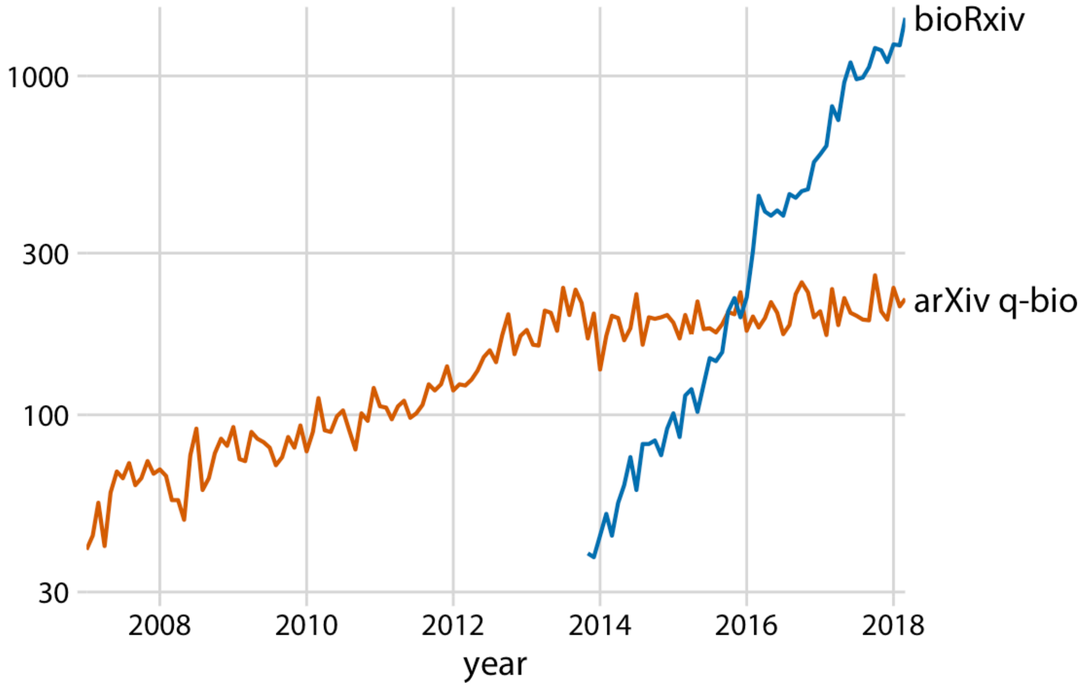


Example

Wilke (2019): preprint submission to arXiv.org in quantitative biology

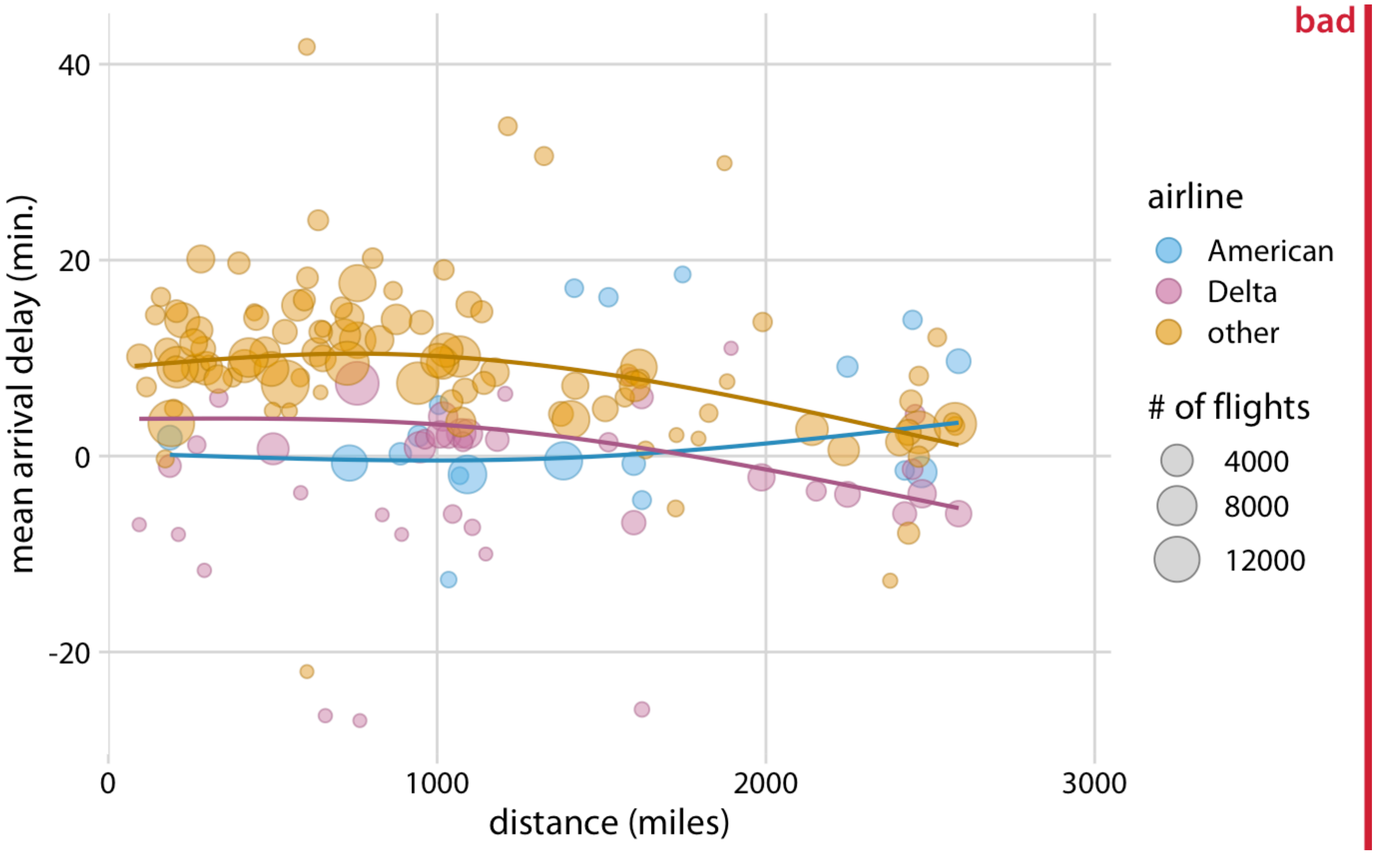
- grew exponentially until 2013
- then, a biology-specific preprint server bioRxiv was launched by a respected publisher





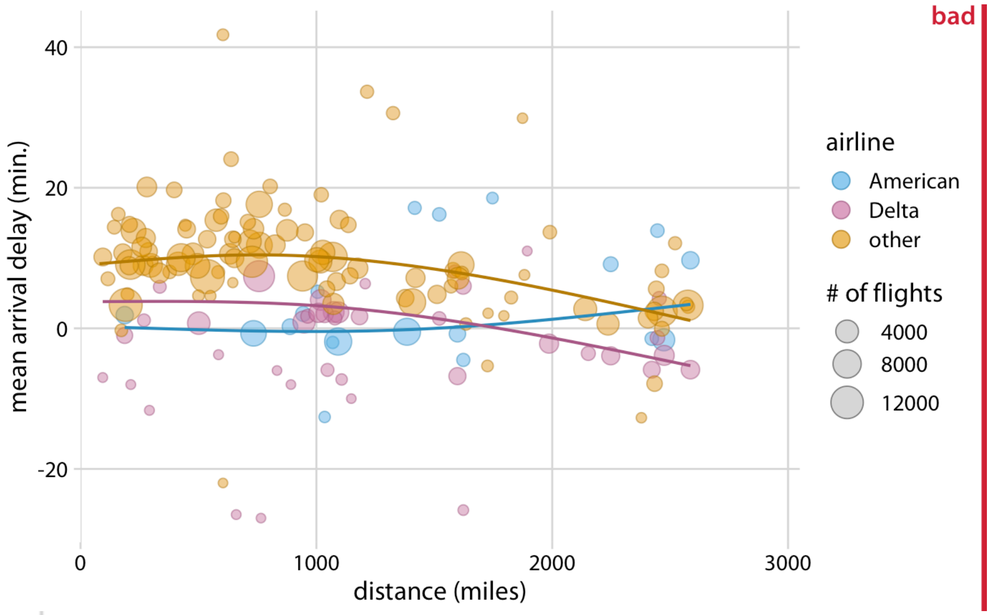
For the generals

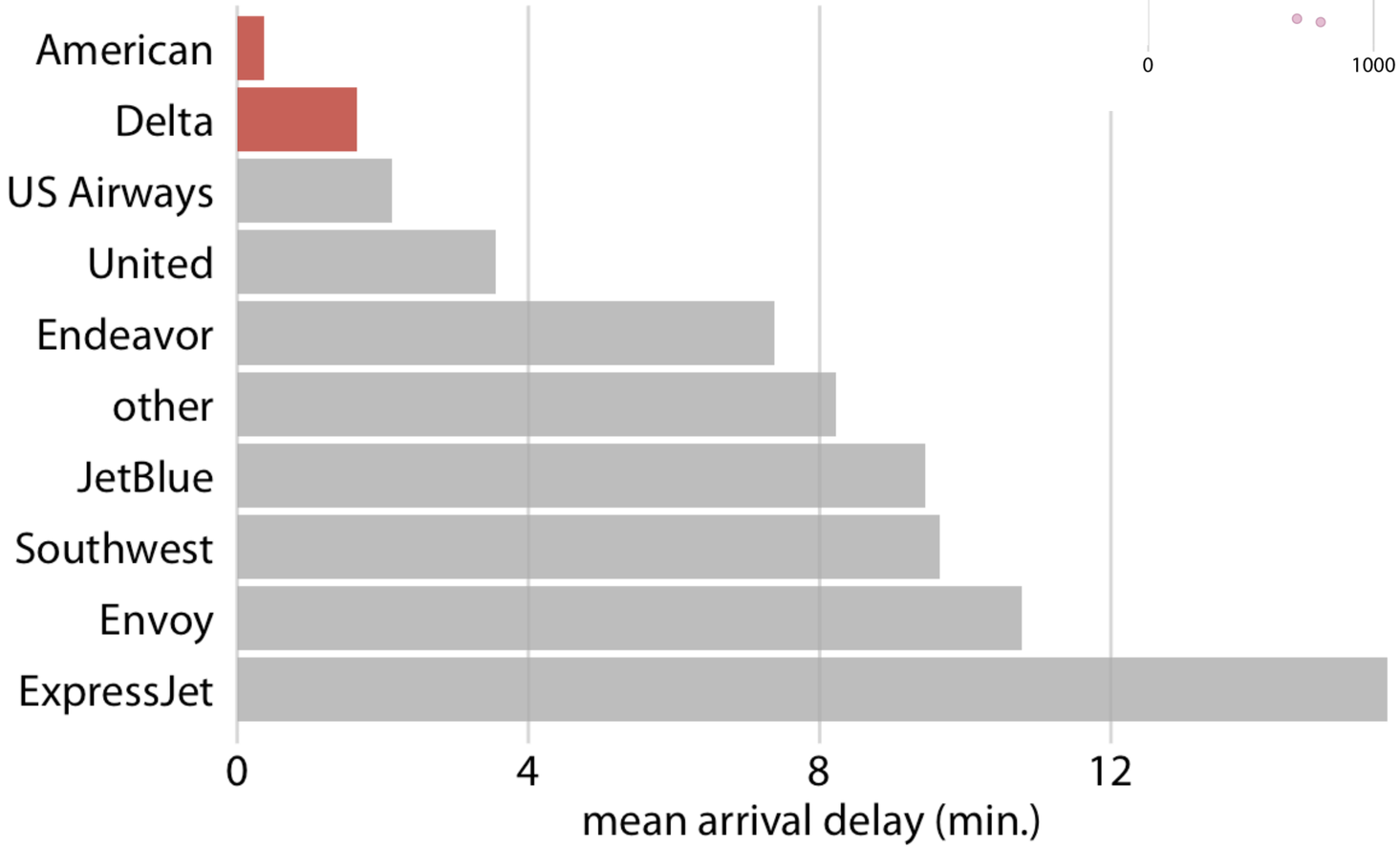
Making a figure for the generals (Wilke 2019)



For the generals

Making a figure for the generals (Wilke 2019)

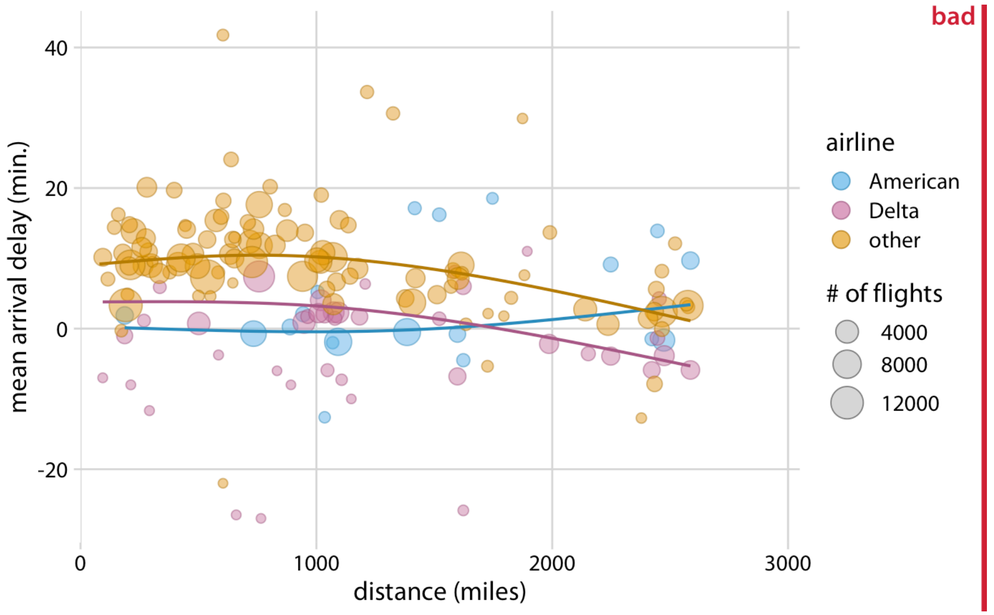


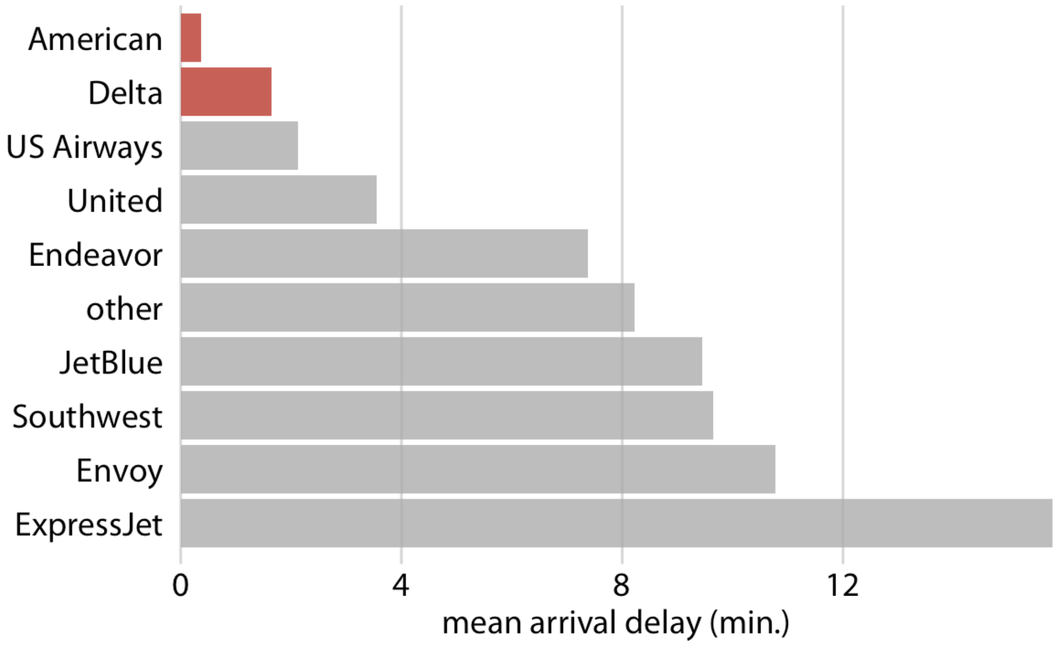


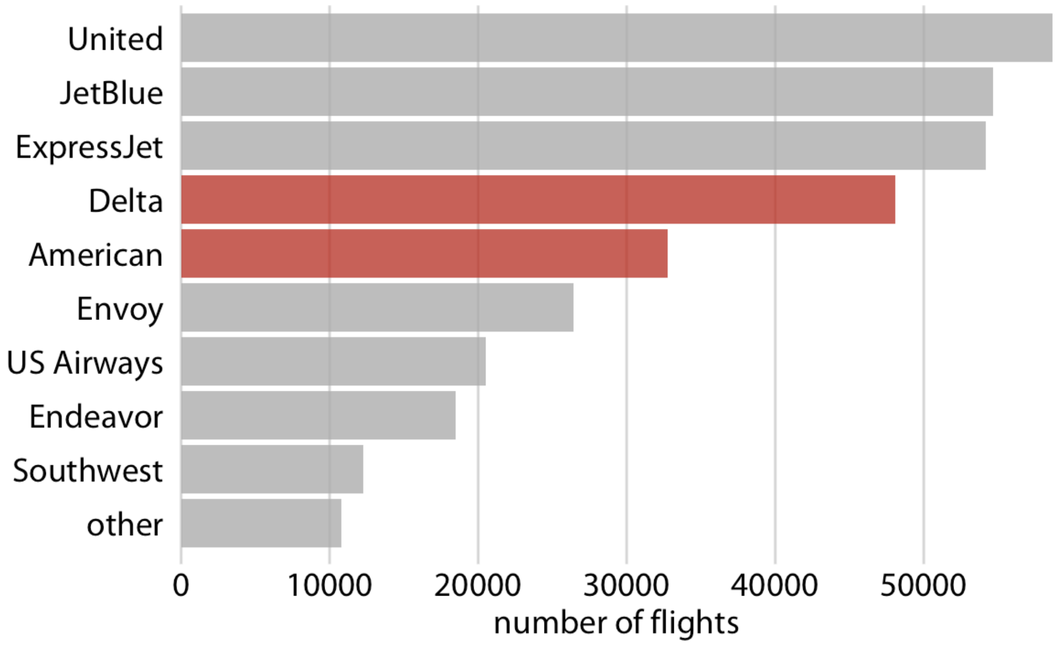
STORYTELLING

For the generals

Making a figure for the generals (Wilke 2019)

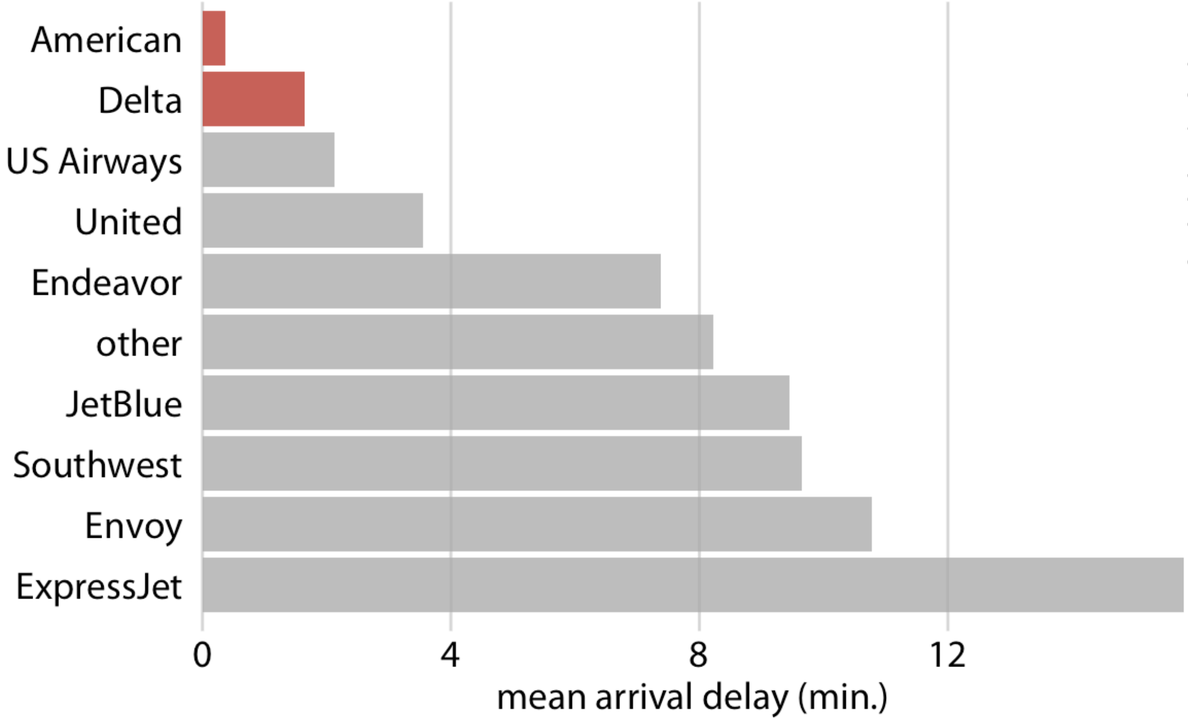


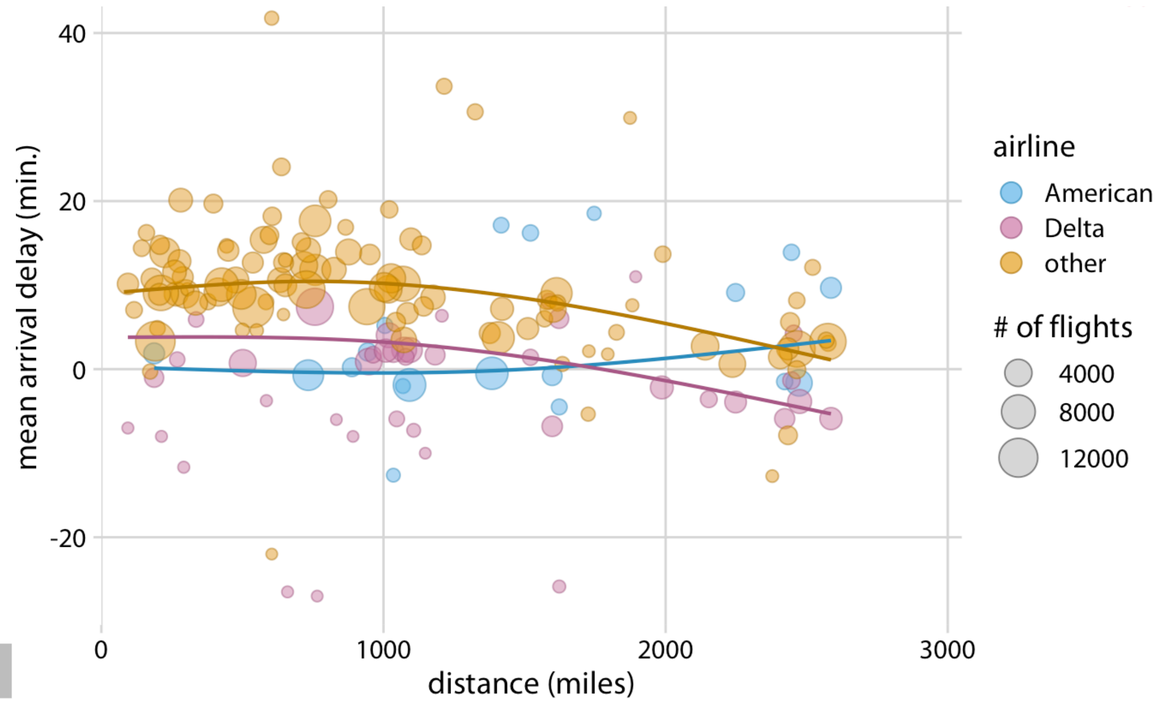




For the generals

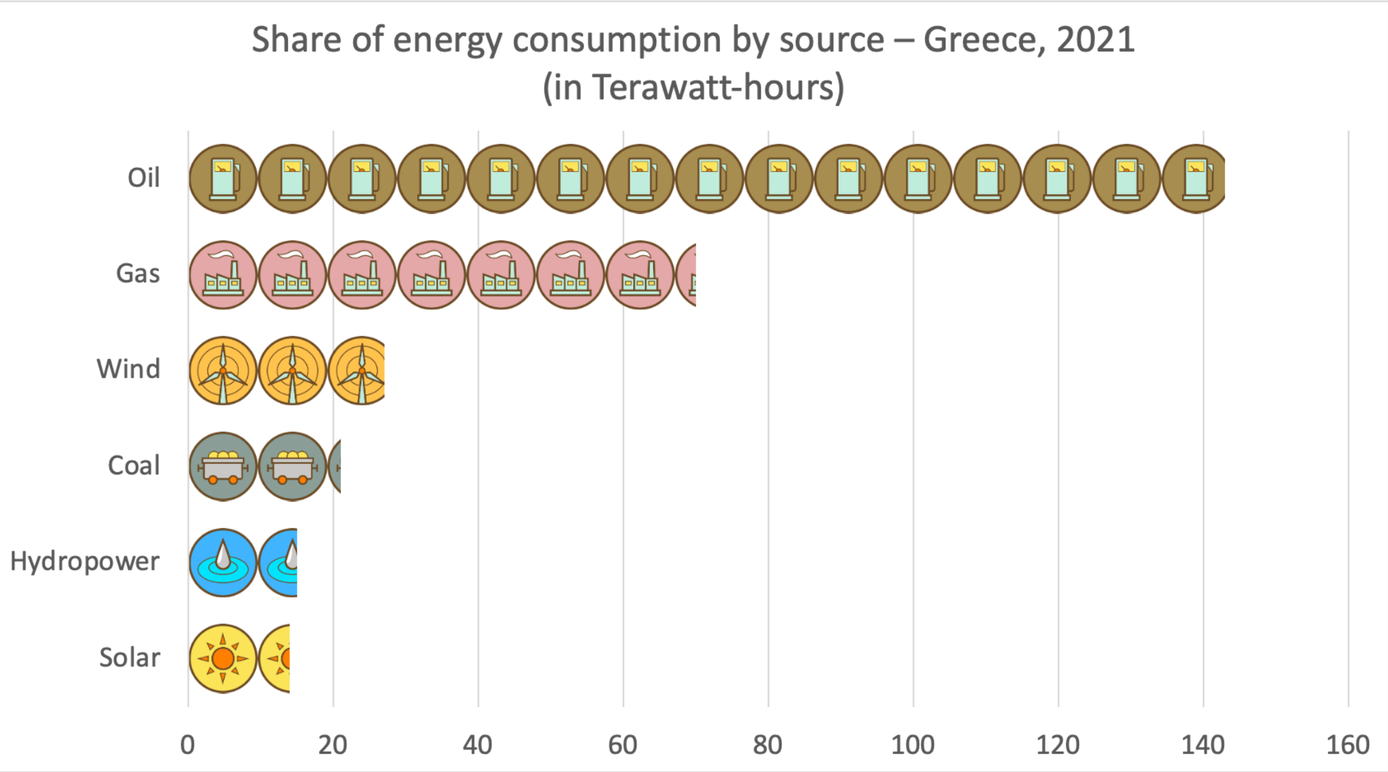
Gradually build up the level of complexity



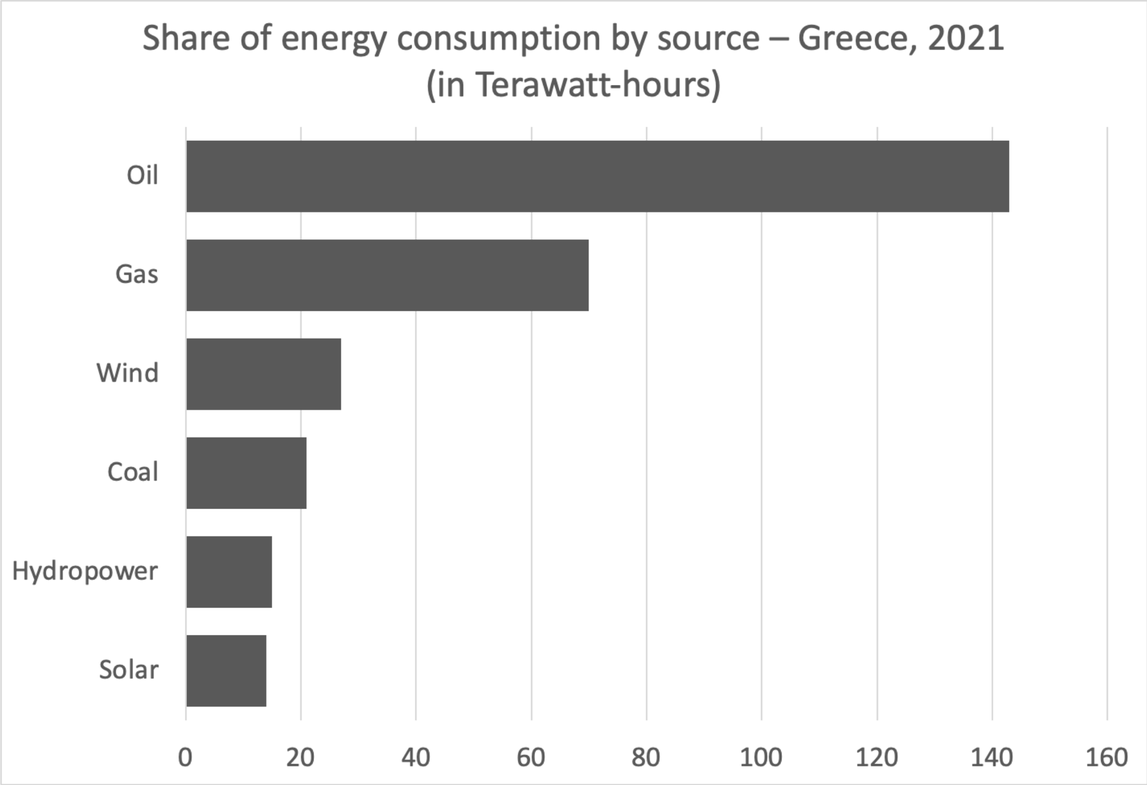


STORYTELLING

Make it memorable

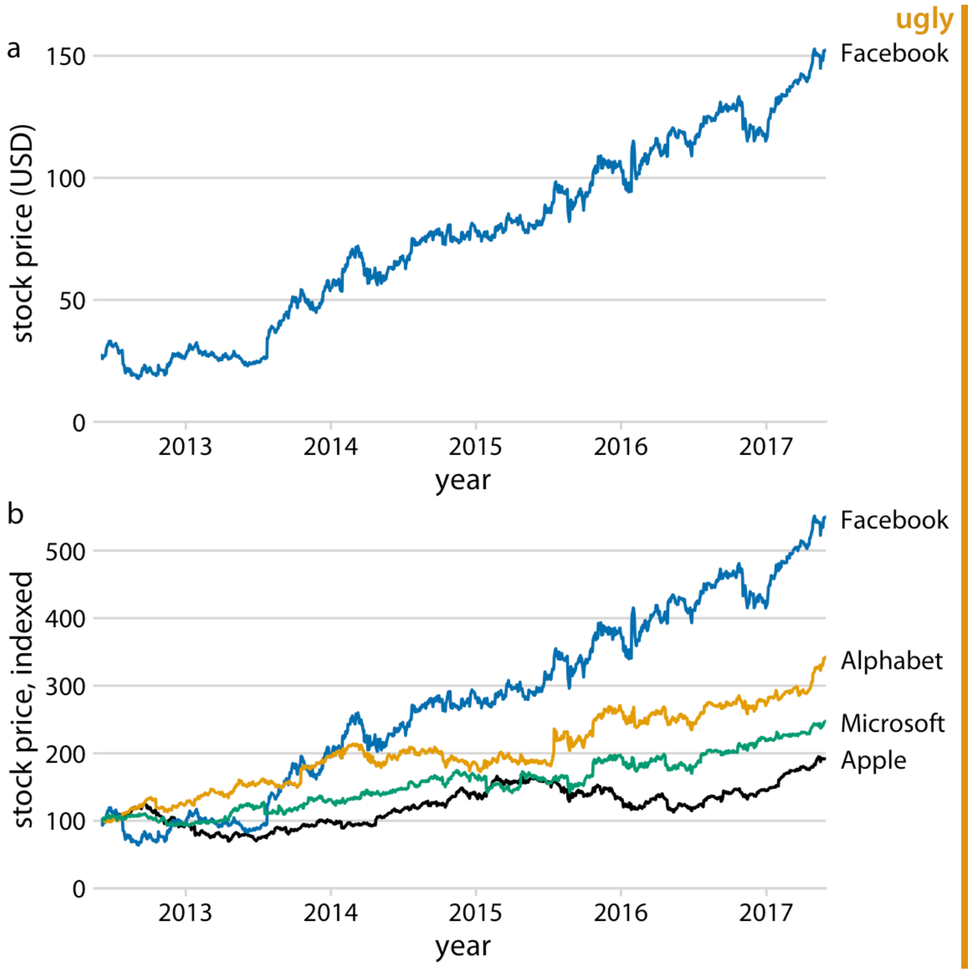


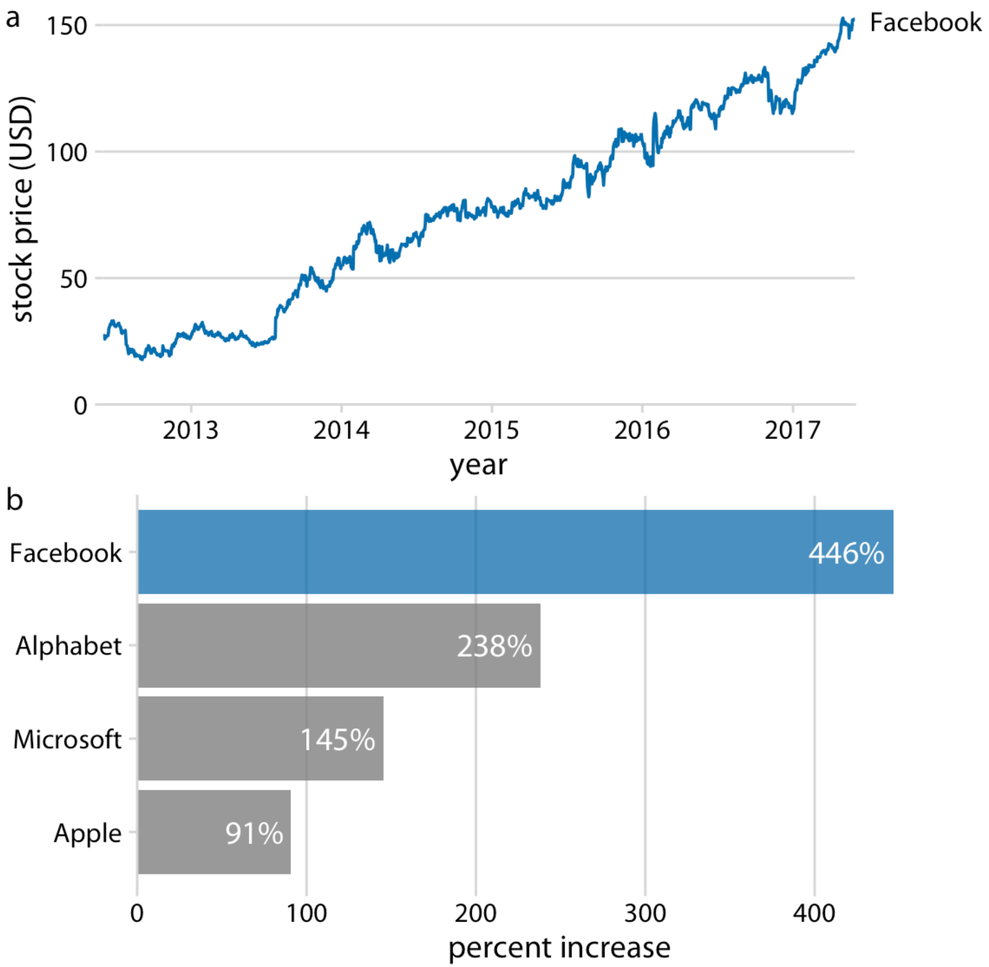
Data: BP Statistical Review of World Energy & Ember



Data: BP Statistical Review of World Energy & Ember

Consistent but not repetitive





Examples

What's the story, how was it built up, and how did the visualizations fit in?

What types of visualizations did they use?

What did you like? What could be improved?

Examining inequality — https://www.gatesfoundation.org/goalkeepers/report/2019-report/
(Only 1st part: e.g. 'Health and education…' + 'Gender inequality stacks…')

30 Years of American anxieties — https://bit.ly/viz_anxieties

Measuring gender imbalances — https://bit.ly/viz_creative

10 years on — https://graphics.reuters.com/FINANCIAL-CRISIS2008/0100805F0BK/
(with audio – select: 'Politics' - by Noah Barkin)

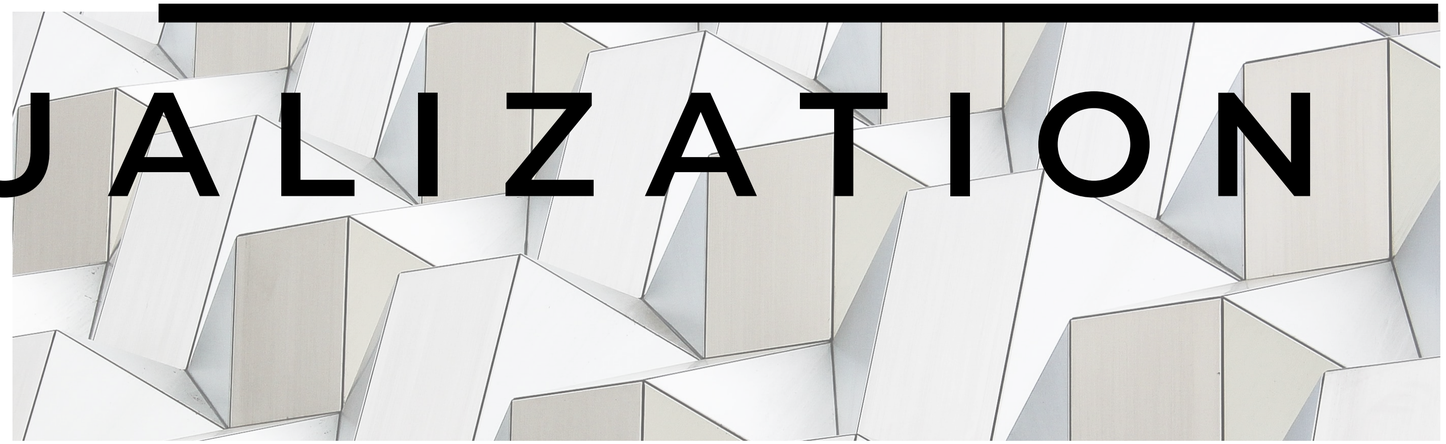
Wine and math — https://bit.ly/viz_wine

Rhythm of food — https://bit.ly/viz_food

Why do cats and dogs… — https://bit.ly/viz_cats

See also https://www.juiceanalytics.com/writing/20-best-data-storytelling-examples

DATA VISUALIZATION



A BRIEF NOTE ON SOFTWARE

A BRIEF NOTE ON

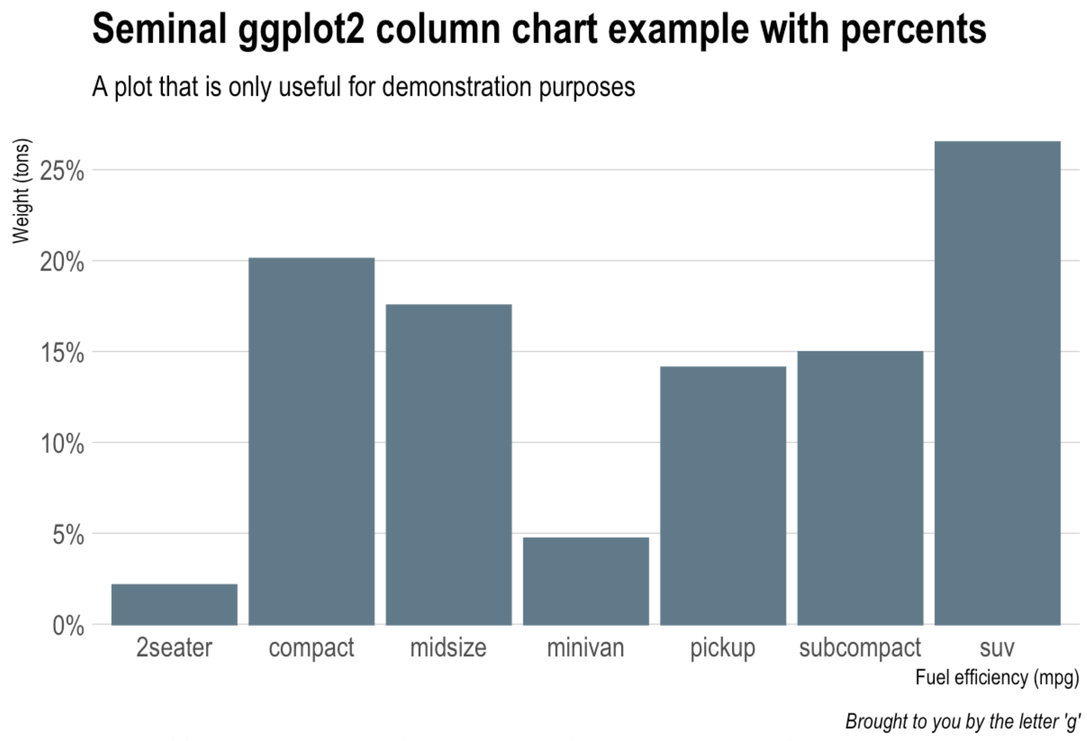
Software

"The best visualization software is the one that allows you to make the figures you need" (Wilke 2019)

- R and RStudio as GUI
- packages `library(tidyverse)` and `library(ggplot2)` in particular (Wickham et al. 2019)
- reproducibility and repeatability
- code available afterwards
- Excel, Python, SPSS, ...
- websites
  (e.g. flourish.studio)
- data exploration versus presentation

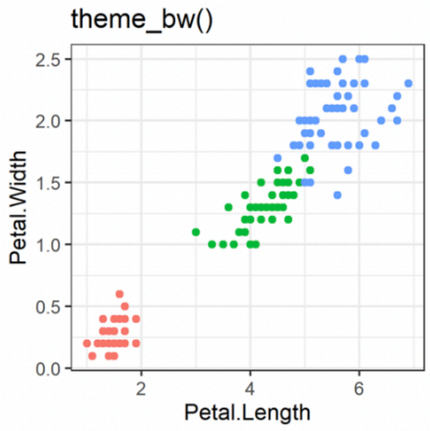
Software for plotting

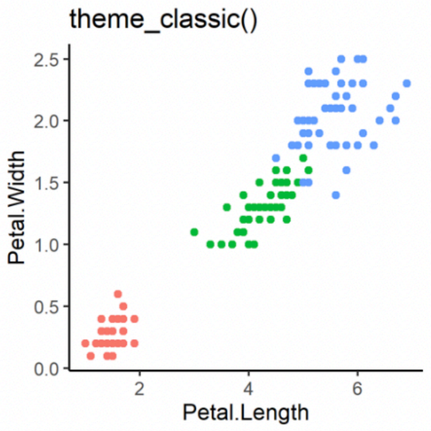
R themes – library(hrbrthemes) (Rudis 2020) → theme_ipsum()

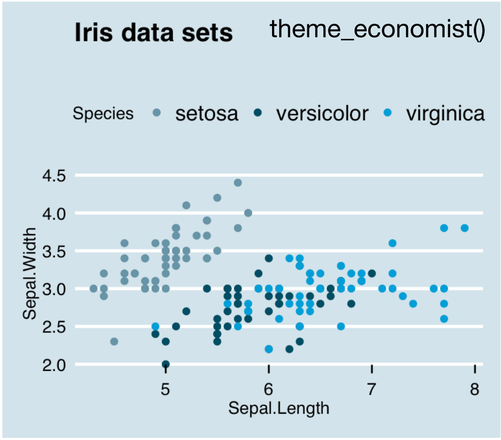


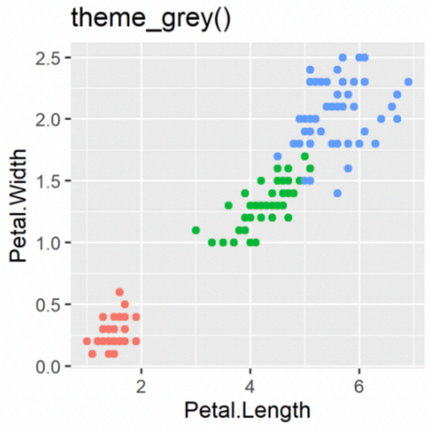
https://github.com/hrbrmstr/hrbrthemes/

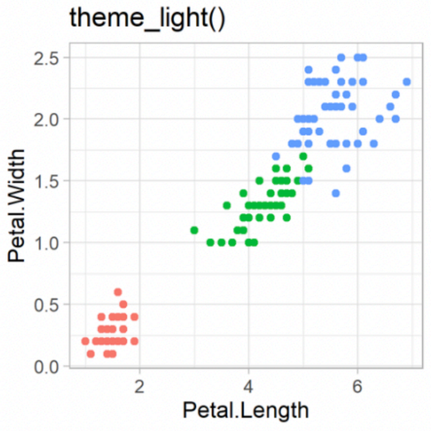
IF YOU REALLY HATE SOMEONE, TEACH THEM TO RECOGNIZE BAD KERNING.

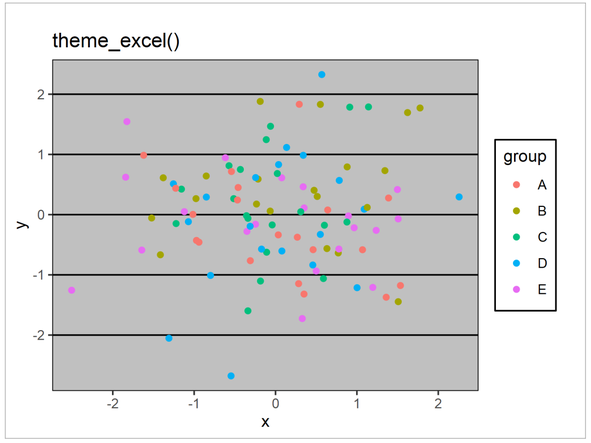

theme_bw()
Petal.Width
Petal.Length

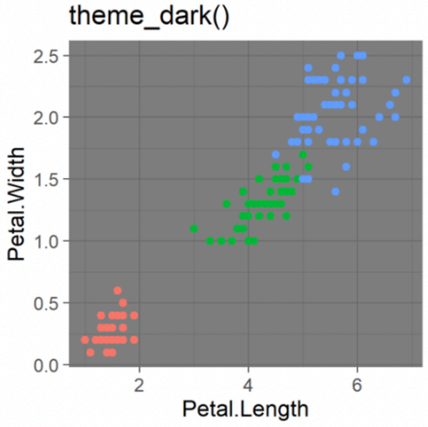

theme_classic()
Petal.Width
Petal.Length

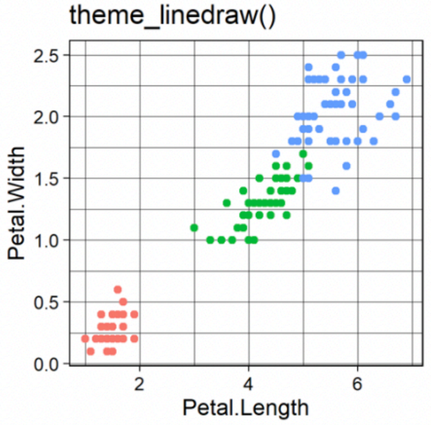

Iris data sets
theme_economist()
Species
setosa
versicolor
virginica
Sepal.Width
Sepal.Length

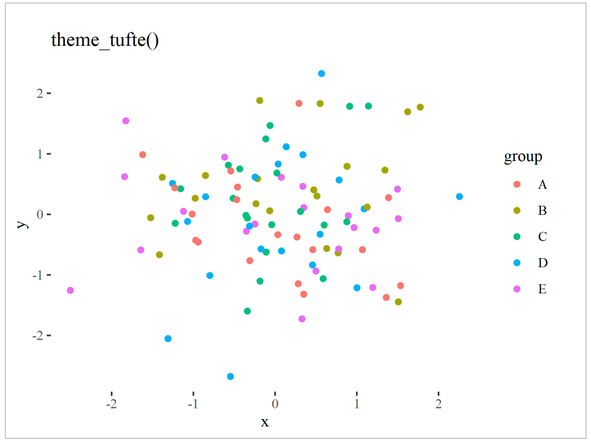

theme_grey()
Petal.Width
Petal.Length


theme_light()
Petal.Width
Petal.Length

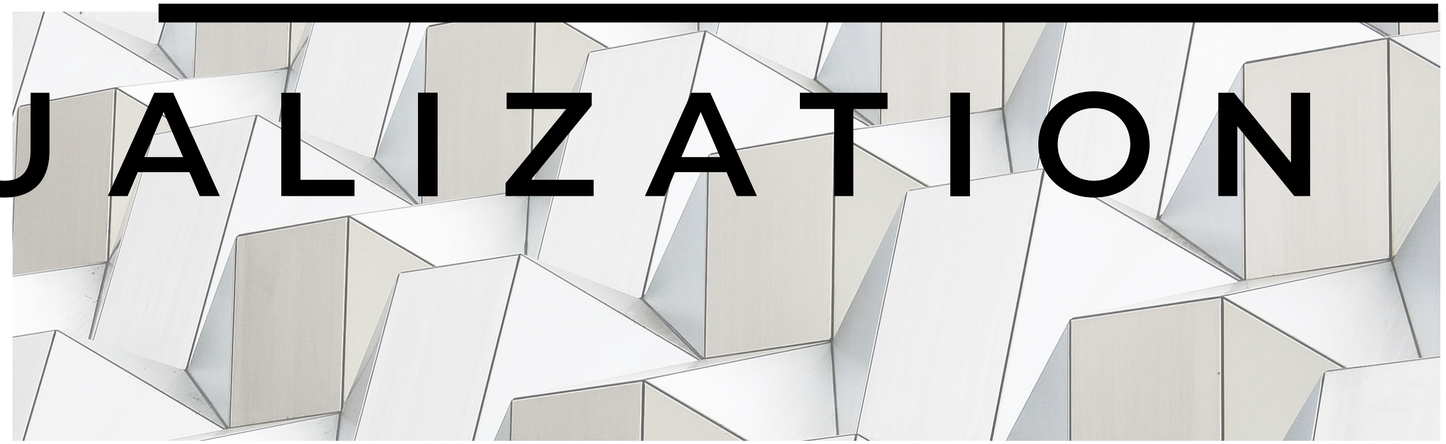

theme_excel()
group
A
B
C
D
E
y
x


theme_dark()
Petal.Width
Petal.Length


theme_linedraw()
Petal.Width
Petal.Length


theme_tufte()
group
A
B
C
D
E
y
x

DATA VISUALIZATION



USING COLOR

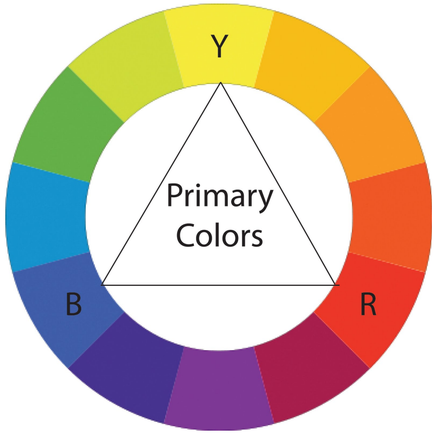
Color scales

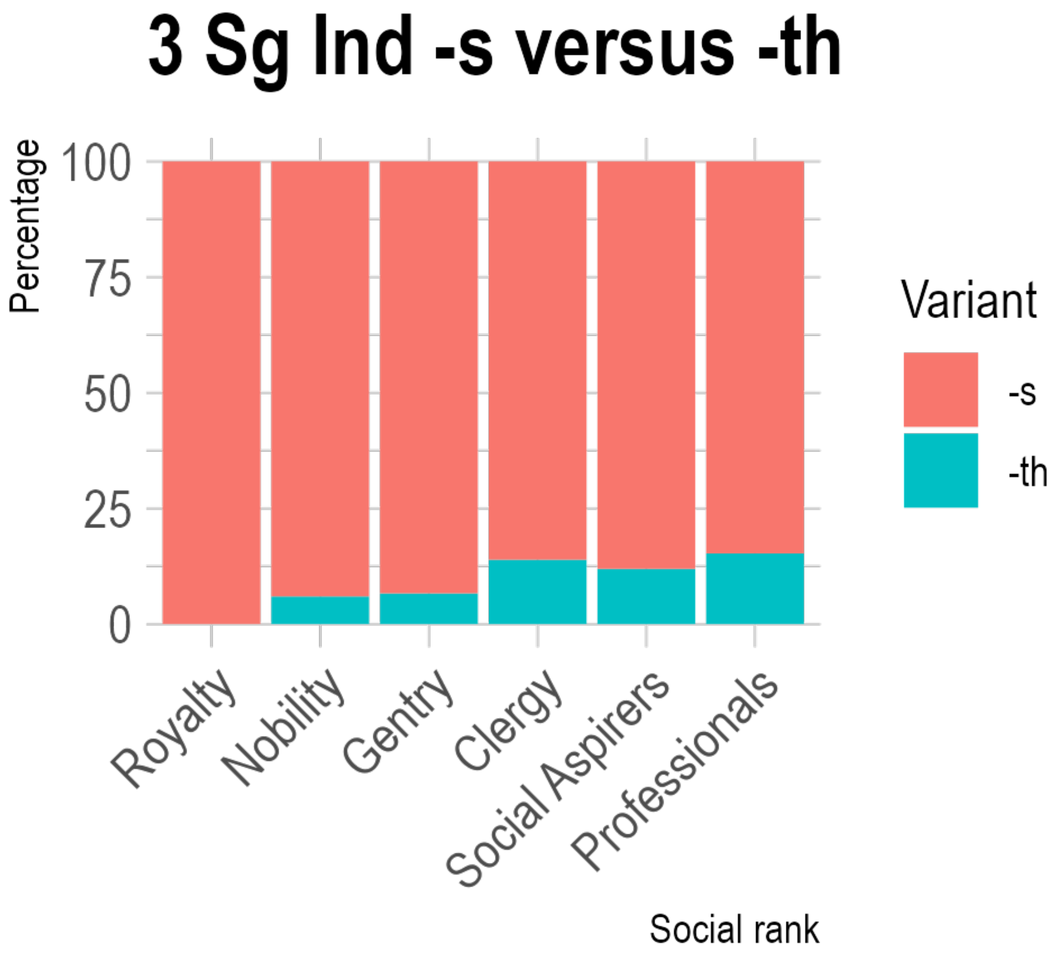
For nominal data:

discrete color scales

— similar colors:
no one standing out

— no pattern or sequence





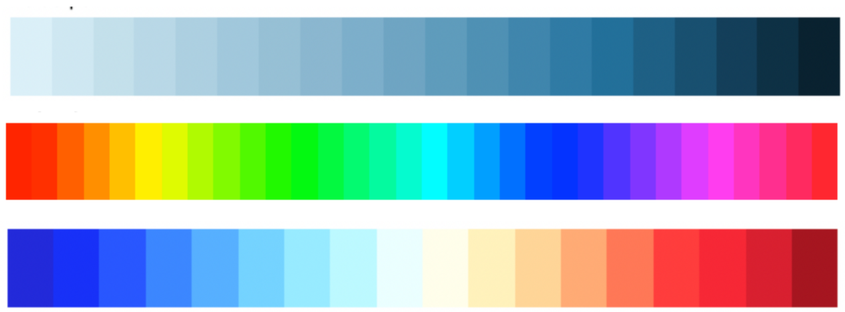
Data: Nevalainen & Raumolin-Brunberg (2016: 140); male informants, 1640–1681

Color scales

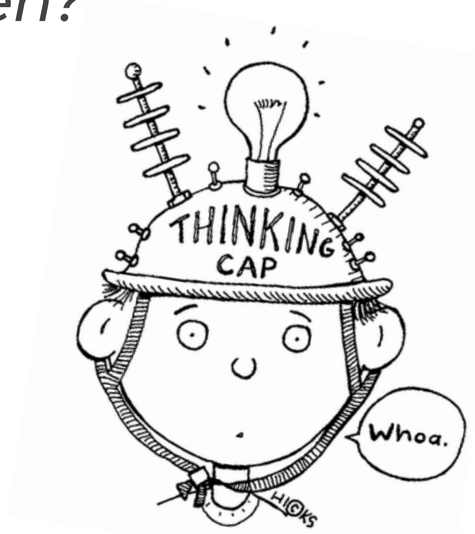
For ordinal or numerical data:

sequential color scales

- monochromatic
- multi-hue
- diverging



when?



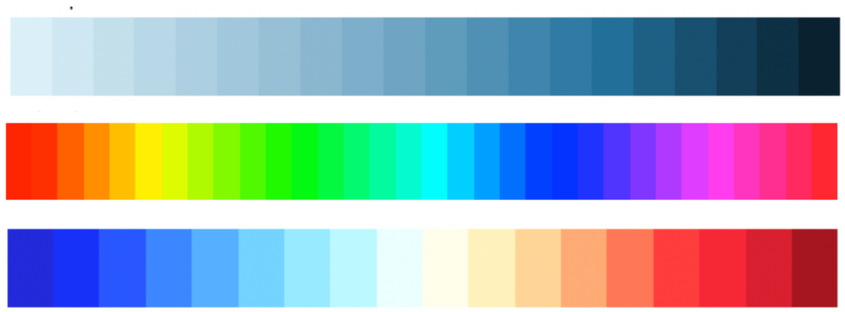
Overview for R?
https://awesomeopensource.com/project/EmilHvitfeldt/r-color-palettes

Color scales

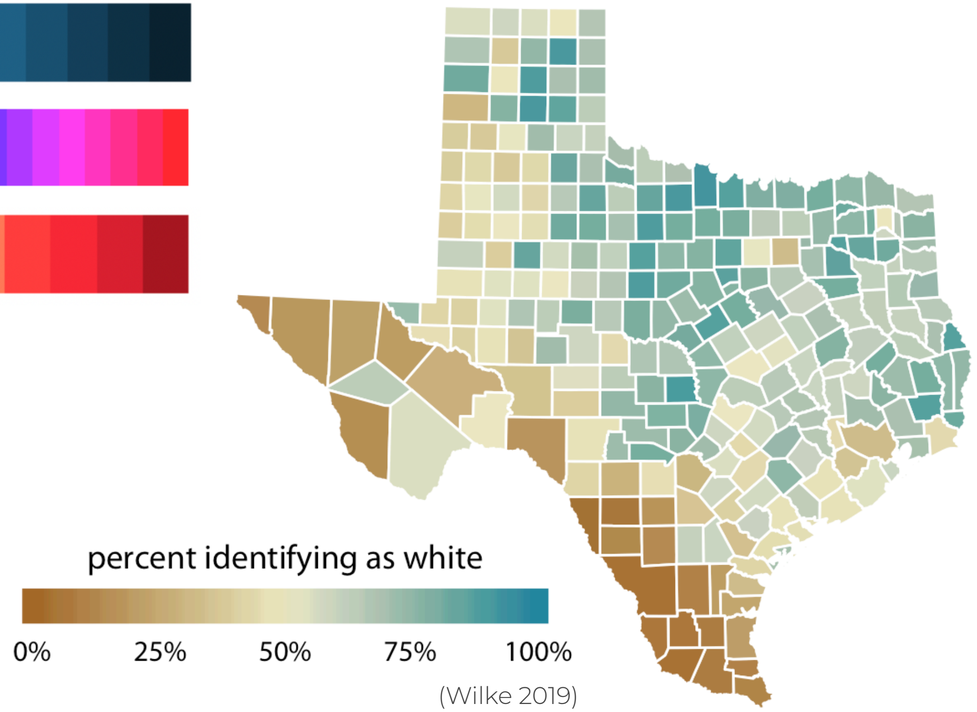
For ordinal or numerical data:

sequential color scales

- monochromatic
- multi-hue
- diverging



CHOROPLETH



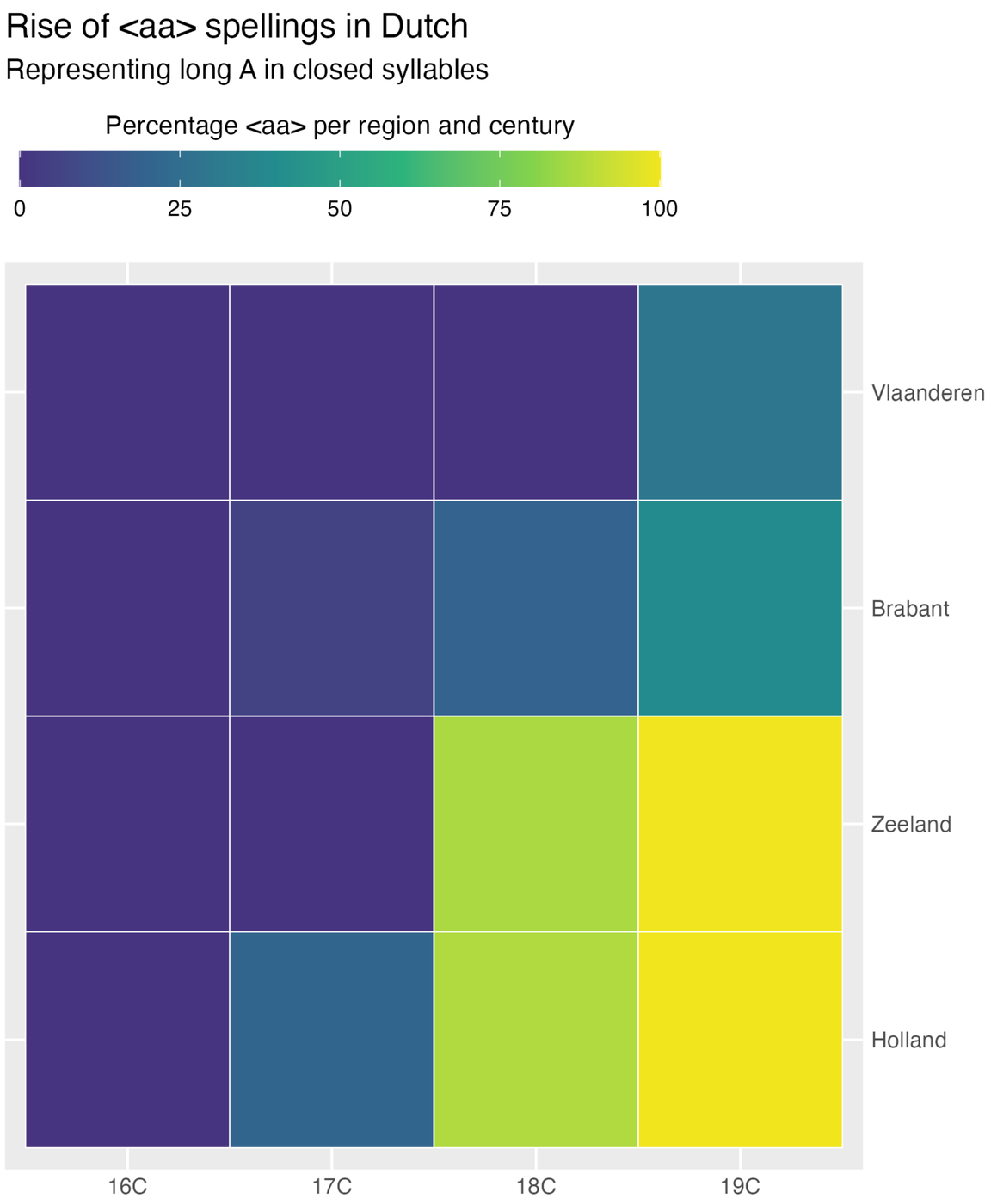
(Wilke 2019)

USING COLOR

Color scales

For ordinal or numerical data:

sequential color scales



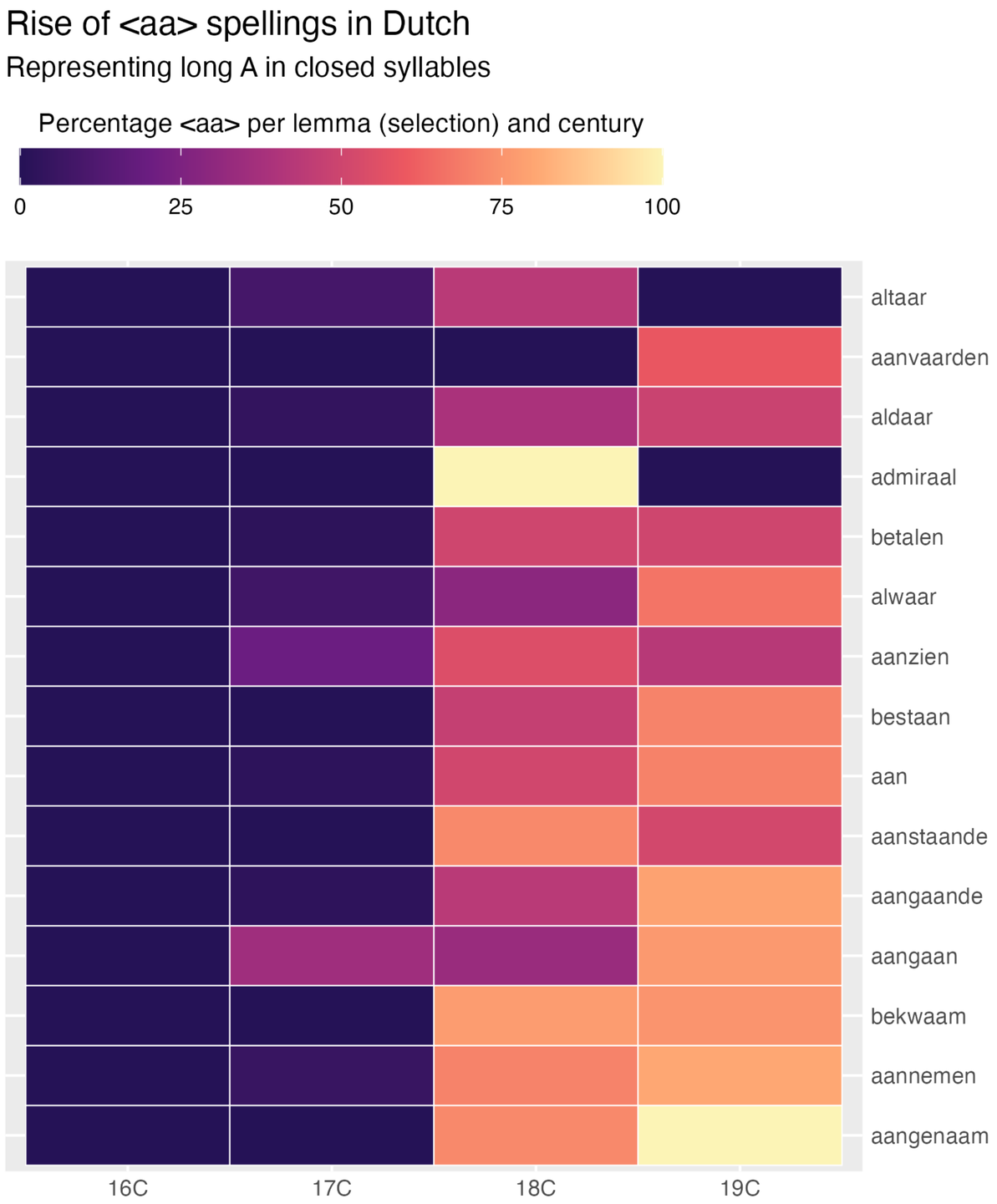
HEAT MAP

USING COLOR

Color scales

For ordinal or numerical data:

sequential color scales

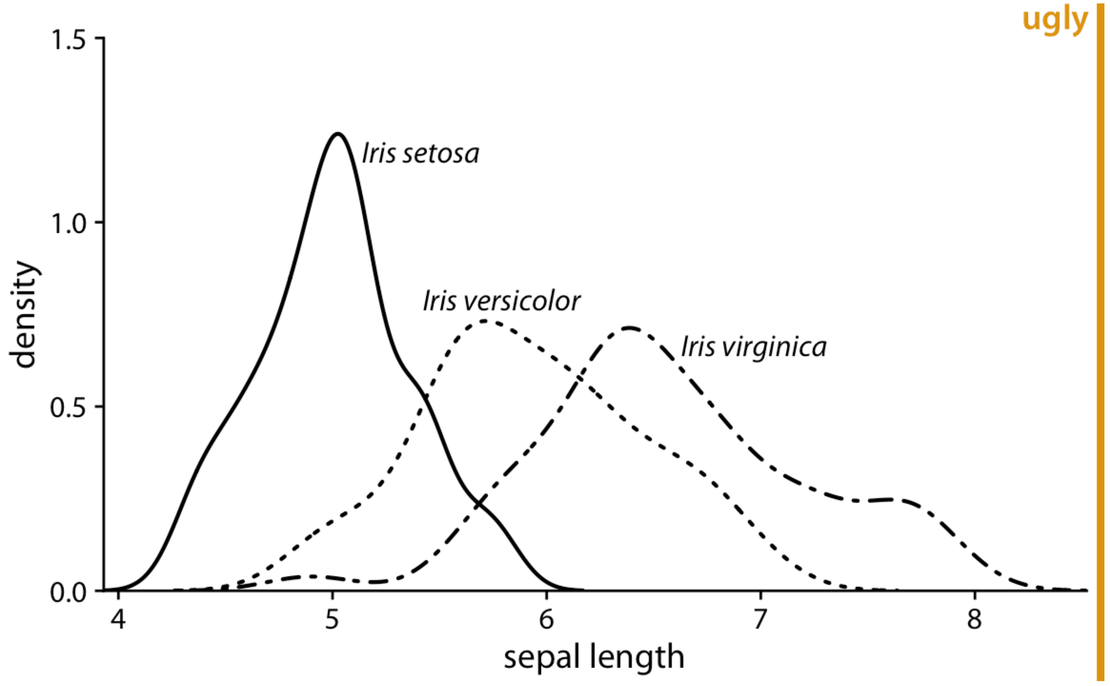


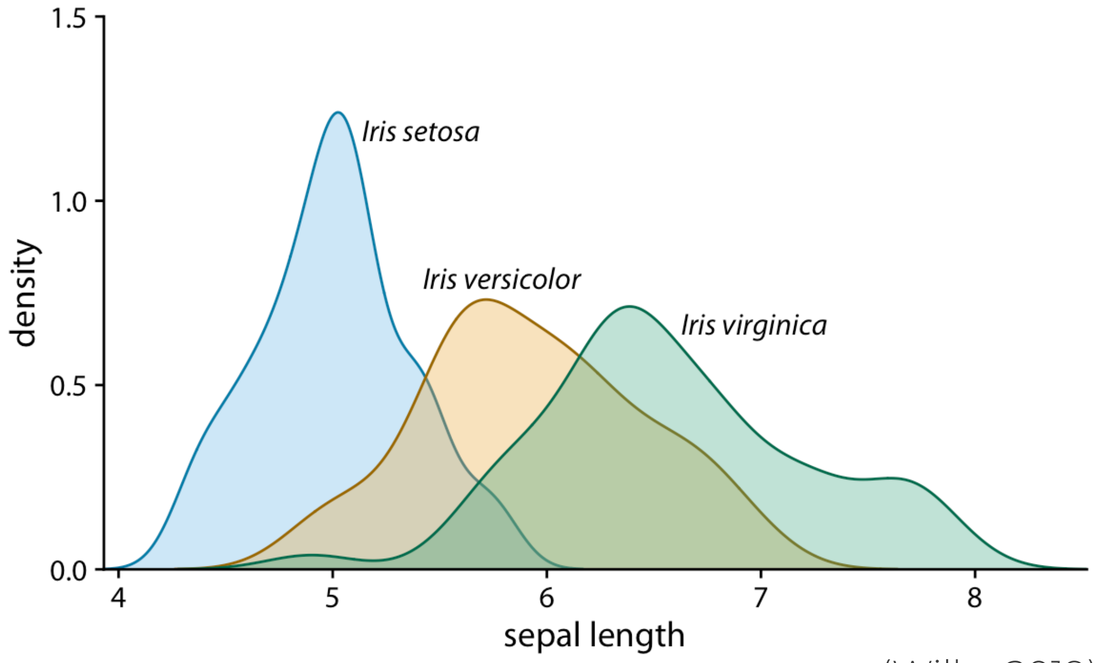
HEAT MAP

Lines versus solid colors

Perceptual preference for **solid colors** and shapes over **line drawings**

(= artifacts from **hand-drawn** visualizations)





(Wilke 2019)

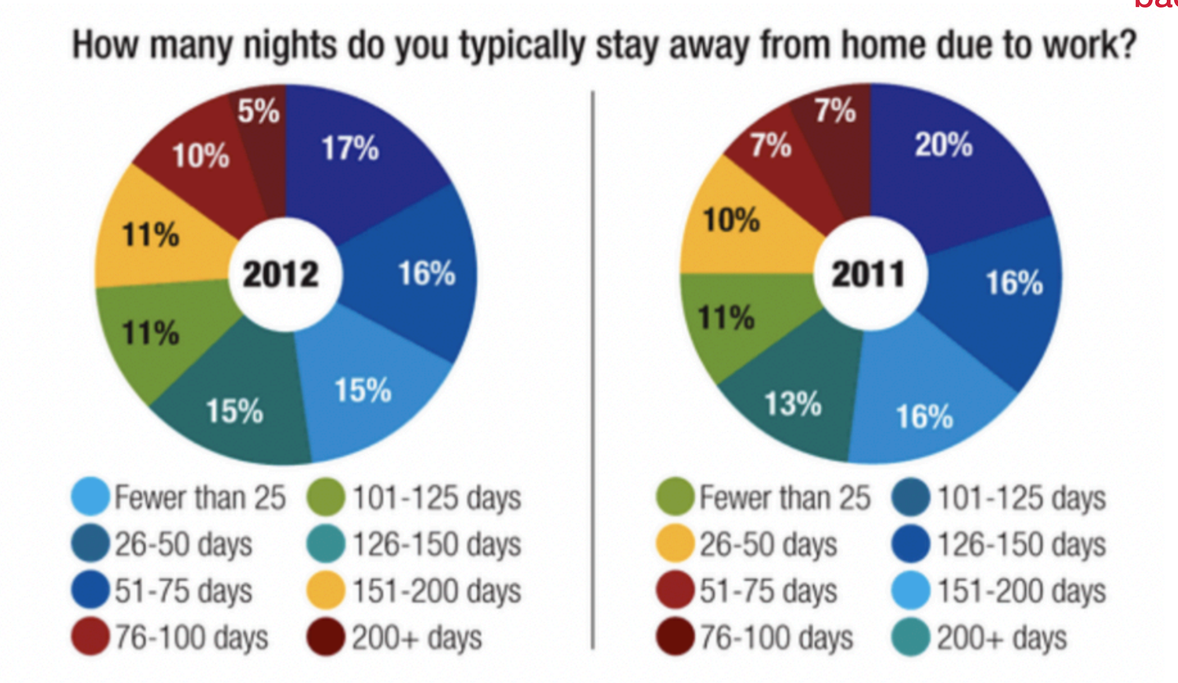
DENSITY PLOT

USING COLOR

Pitfalls

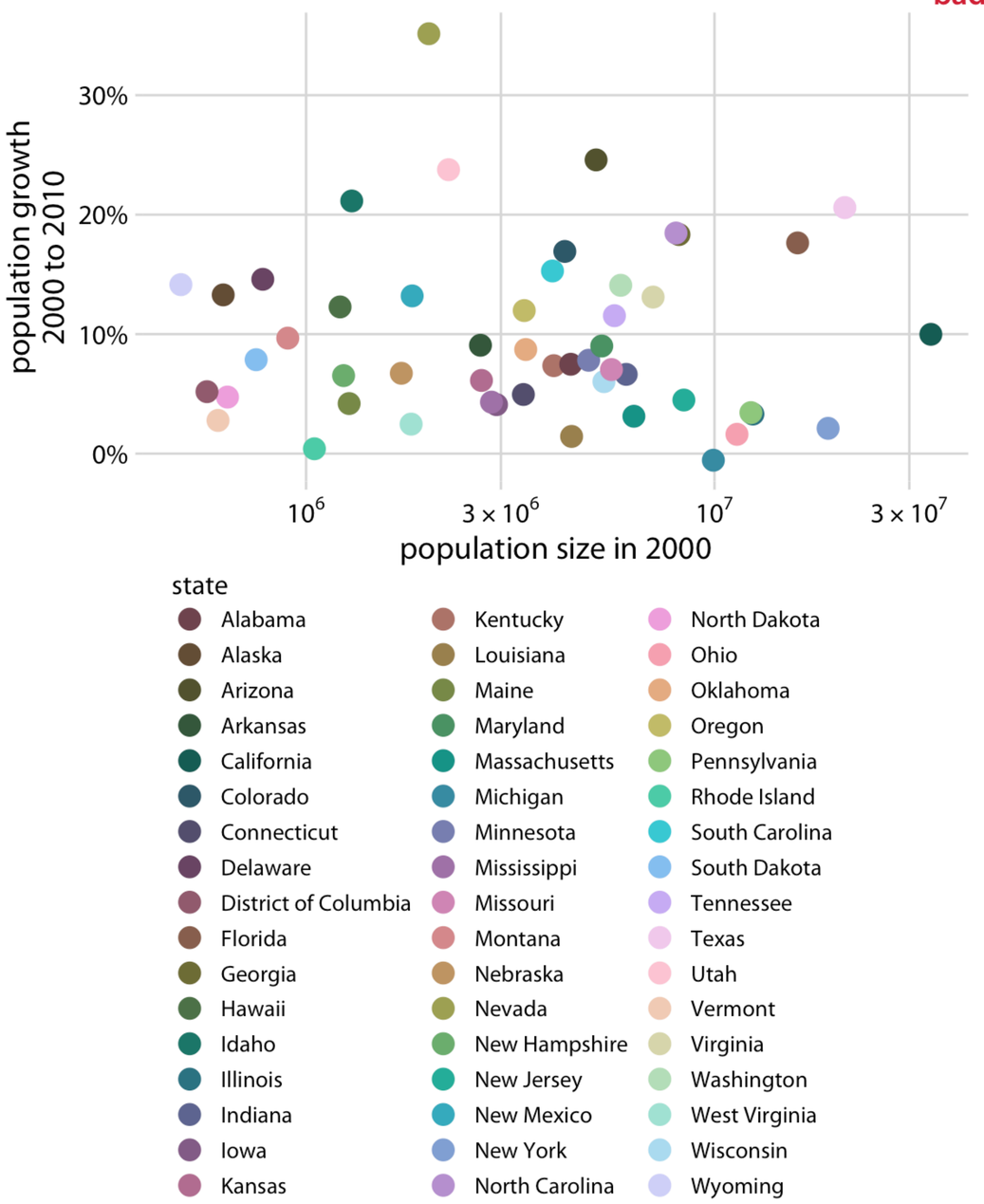
Pitfall 1: **TMI**

bad



https://towardsdatascience.com/color-in-data-visualization-less-how-more-why-348514a3c4d8

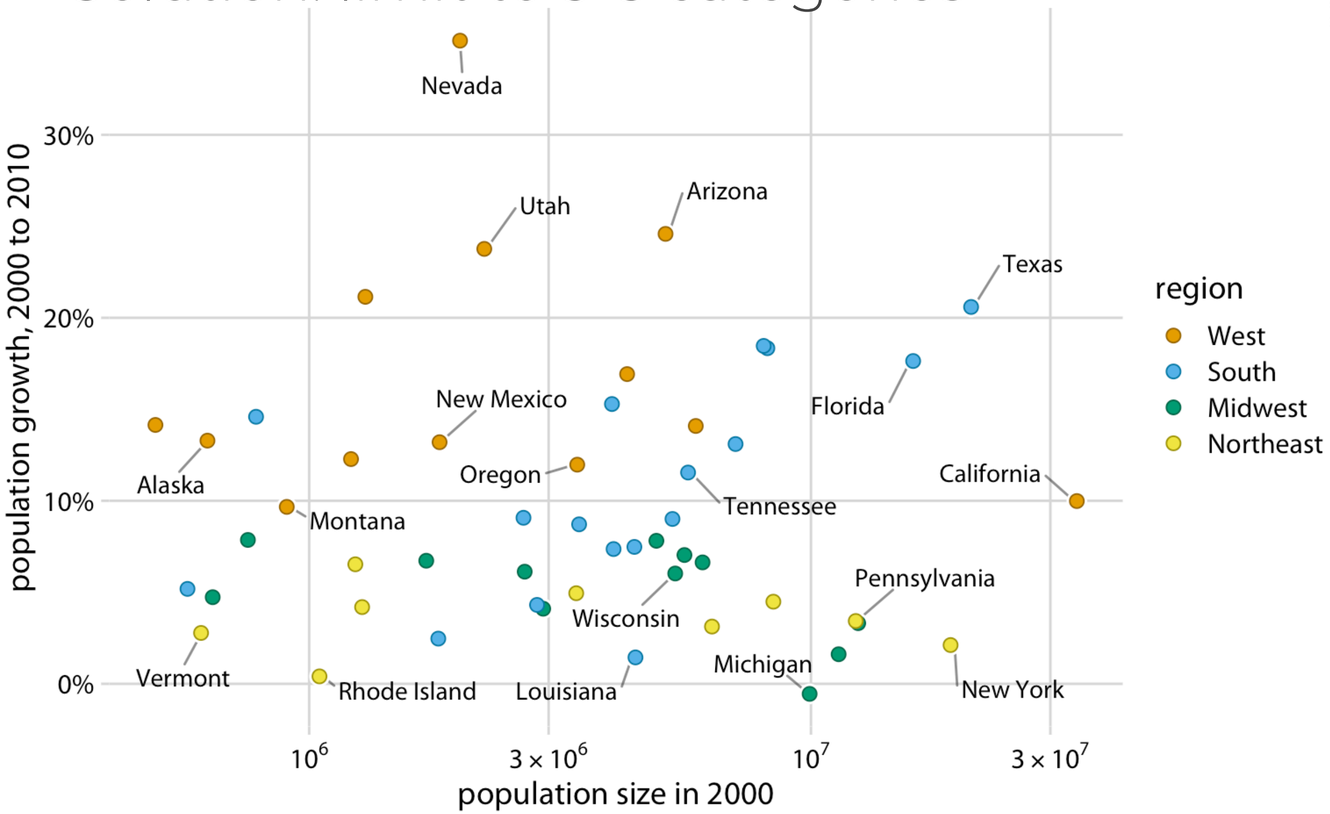
bad

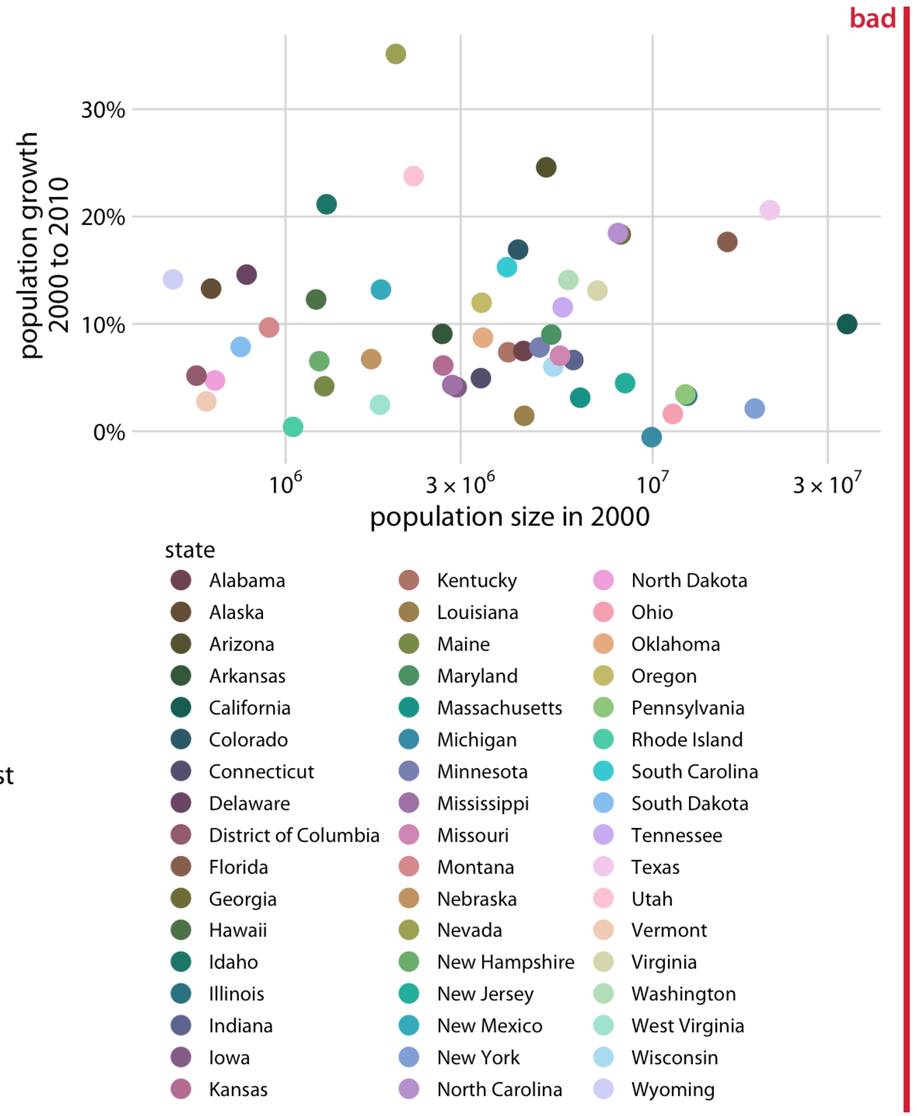


(Wilke 2019)

Pitfalls

Solution: limit to 3-5 categories



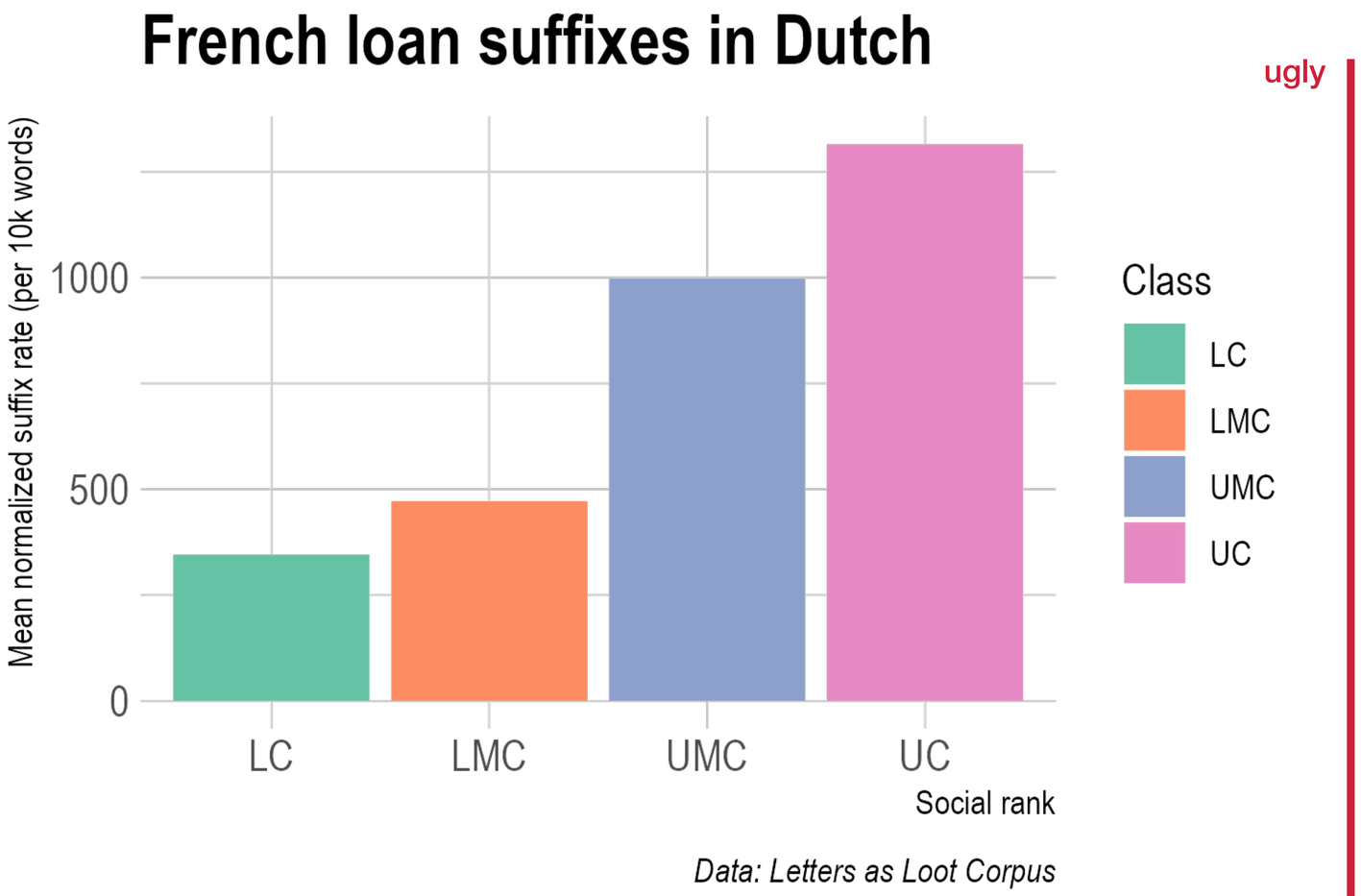


⇒ *Use color to explore!*

(Wilke 2019)

Pitfalls

Pitfall 2: **No information**

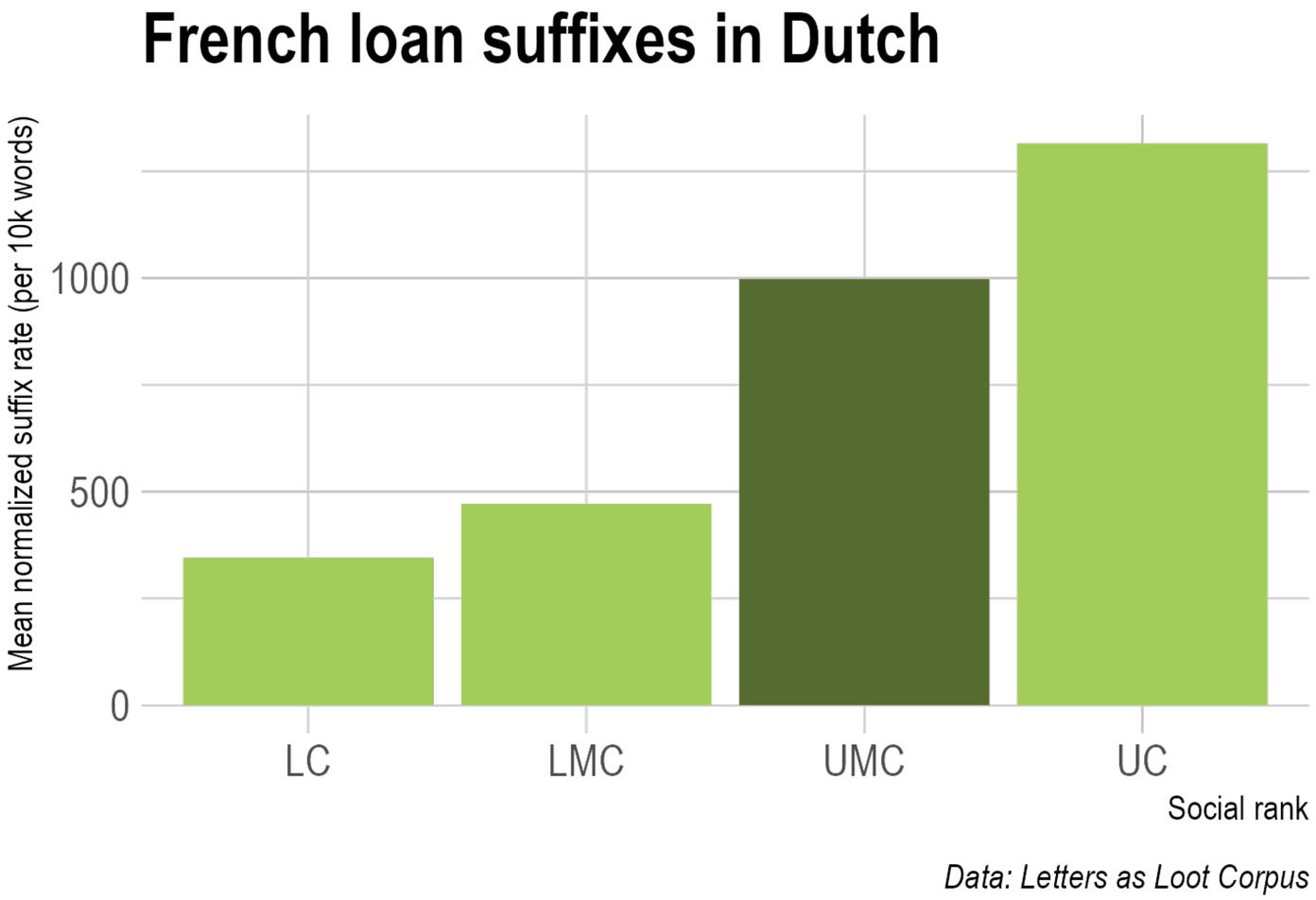


ugly

Pitfalls

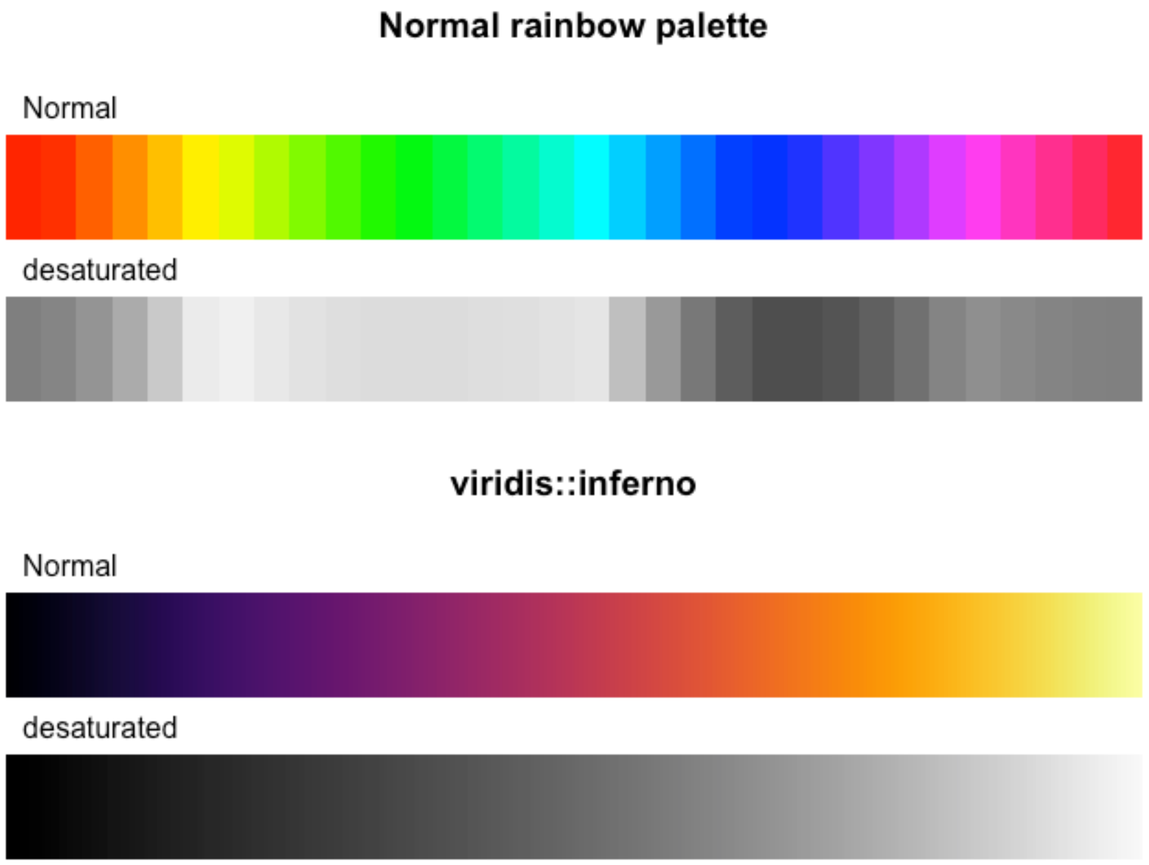
Pitfall 2: **No information**

⇒ *Use color to highlight!*



Pitfalls

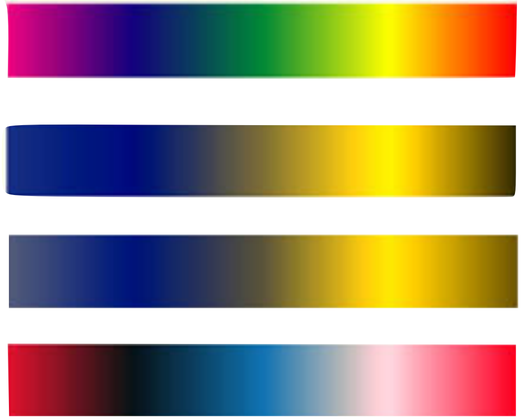
Pitfall 3: **Printing in black and white**



(Wilke 2019)

Pitfalls

Pitfall 4: **Accessibility for people with color-blindness**



trichromatic vision

protanomaly (*red)

deuteranomaly (*green)

tritanomaly (*blue/yellow)

https://jfly.uni-koeln.de/color/

#E69F00	#56B4E9	#009E73	#F0E442	#0072B2	#D55E00	#CC79A7	#000000

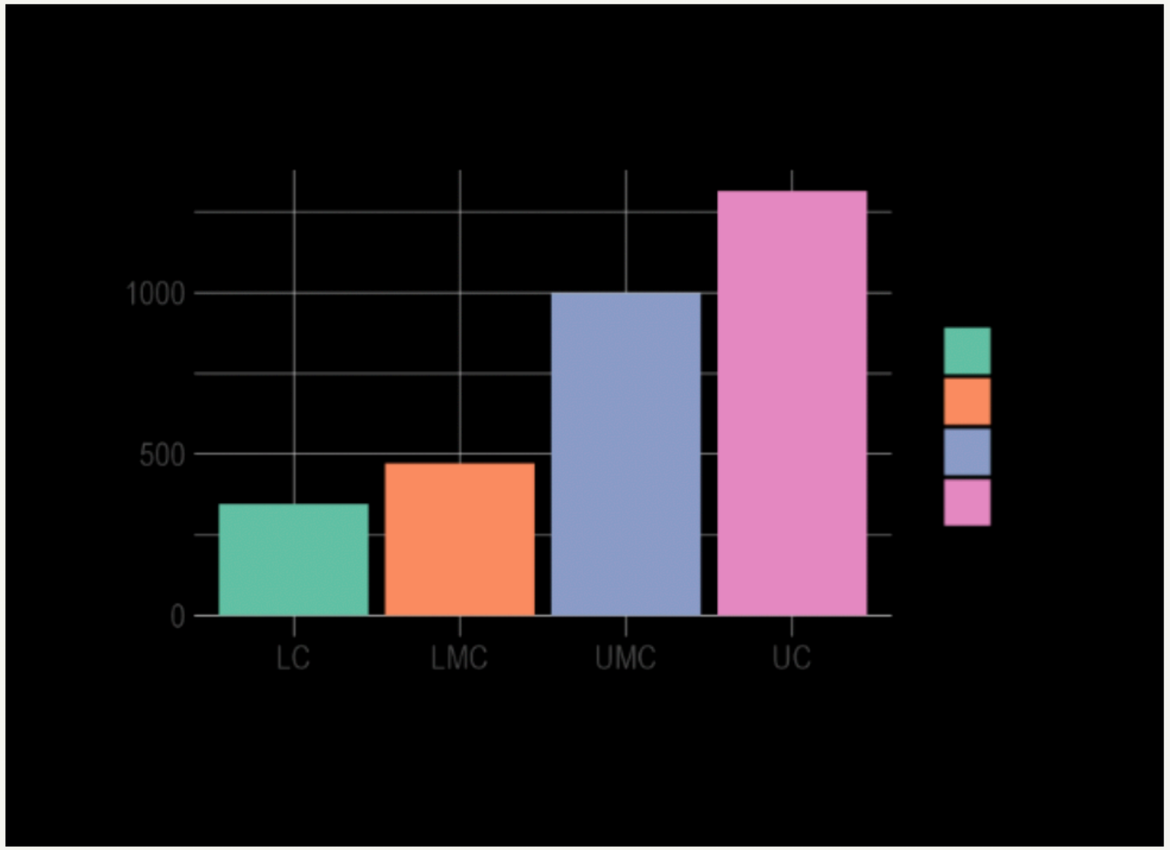
Okabe and Ito 2008

Process Your Own! Load Image *Current:* *Overview_bar9.png*

Privacy Note: processing your images on this page **does not upload them to any server.** All processing happens locally on your machine in this webapp — we'll never see it.

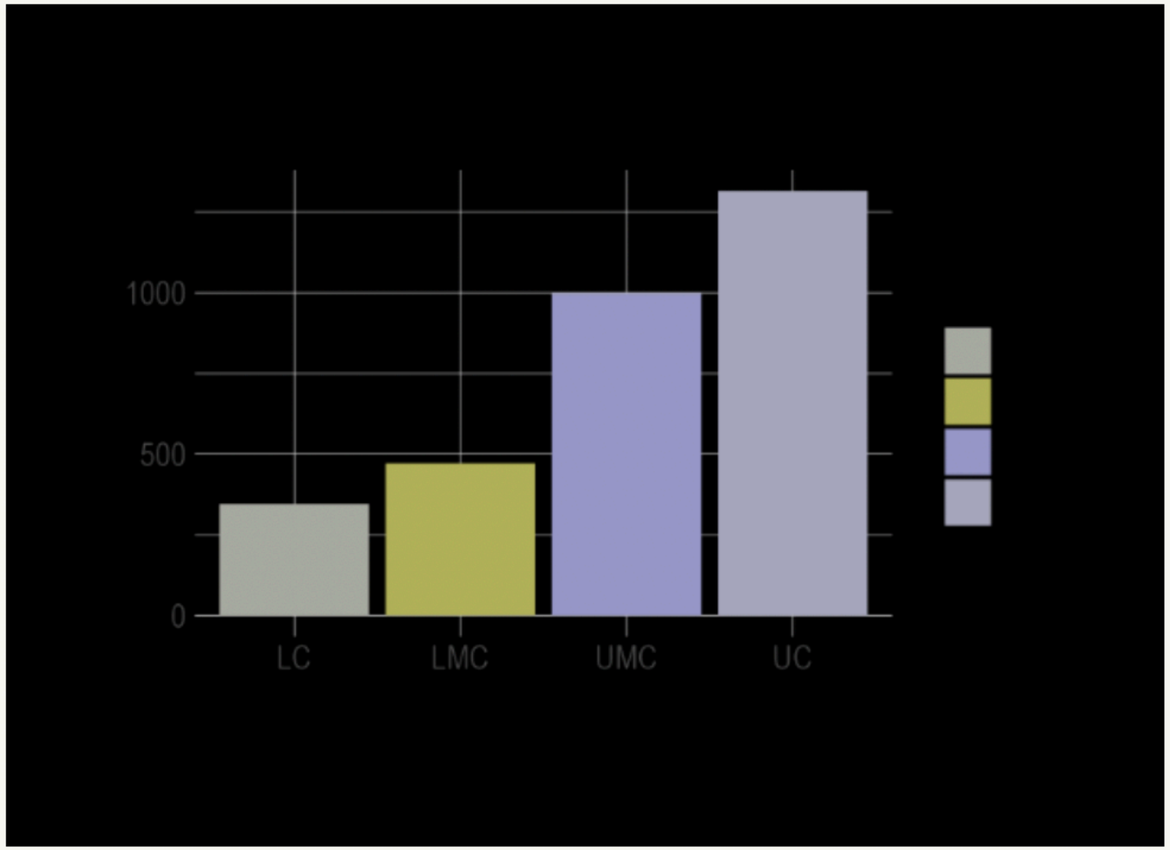
Trichromatic Vision

Normal LMS Cone Response



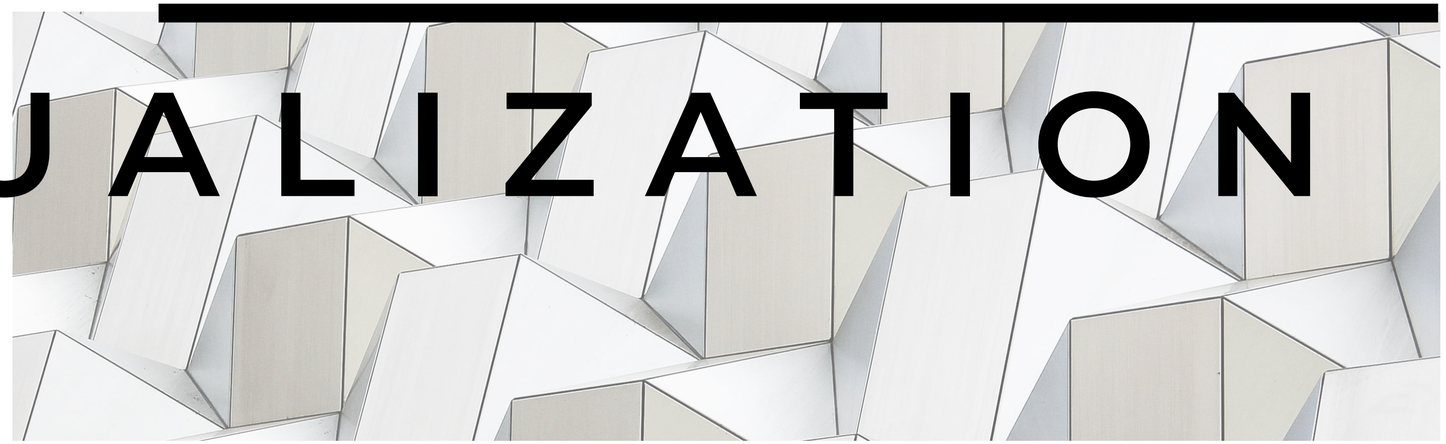
Deuteranopia

Missing M Cones (Green Cones)



https://www.myndex.com/CVD/four

DATA VISUALIZATION



GROUP WORK

Assignment

In groups of two or three,
go back to some of the data you have (or one of you has),
and find different appropriate visualizations
to tell us a story with your data.

Tool: http://www.flourish.studio or: Excel or: hand-draw (approx.)!

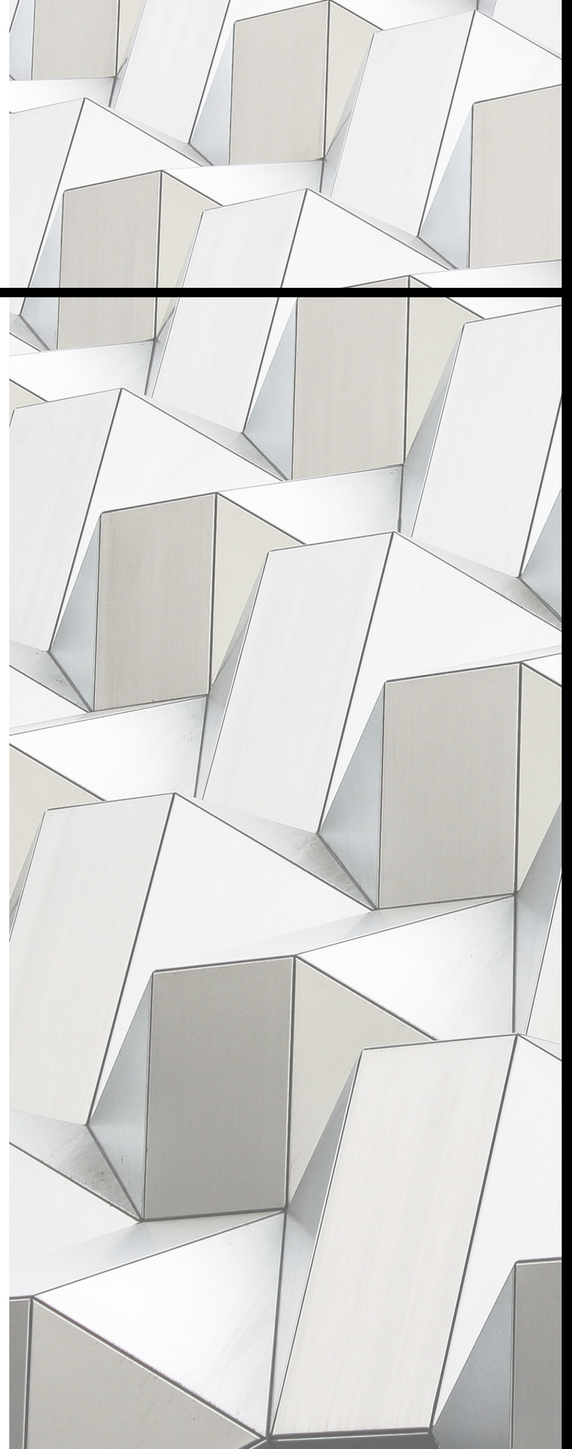
Work with other people. If none of you have any data,
feel free to use Tagliamonte & Baayen (2012):

https://bit.ly/TagliamonteBaayen2012

See also https://www.juiceanalytics.com/writing/20-best-data-storytelling-examples

THANK YOU

Rik Vosters
Vrije Universiteit Brussel



FEEL FREE TO GET IN TOUCH

Email

Rik.Vosters@vub.be

Website

www.rikvosters.be

www.historicalsociolinguistics.be