

Graph Construction

Data Management and Visualization



SoftEng
<http://softeng.polito.it>

Version 2.7.0
© Marco Torchiano, 2022





This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

To view a copy of this license, visit

<http://creativecommons.org/licenses/by-nc-nd/4.0/>.

You are free: to copy, distribute, display, and perform the work

Under the following conditions:



Attribution. You must attribute the work in the manner specified by the author or licensor.



Non-commercial. You may not use this work for commercial purposes.



No Derivative Works. You may not alter, transform, or build upon this work.

- For any reuse or distribution, you must make clear to others the license terms of this work.
- Any of these conditions can be waived if you get permission from the copyright holder.

Your fair use and other rights are in no way affected by the above.

Data Visualization

Understanding

Information Visualization

Visual Patterns, Trends, Exceptions

Quantitative Reasoning

Quantitative Relationship & Comparison

Visual Perception

Visual Properties & Objects

Representation/Encoding



Grammar of Graphics

- Theory behind graphics construction
 - ◆ Separation of data from aesthetic
 - ◆ Definition of common plot/chart elements
 - ◆ Composition of such common elements
- Building a graphic involves
 1. Specification
 2. Assembly
 3. Display

Leland Wilkinson, *The grammar of graphics*

Specification

- **DATA**: a set of data operations that create variables from datasets
 - ◆ Link variables (e.g., *by index* or *id*)
- **TRANS**: variable transformations (e.g., *rank*)
- **SCALE**: scale transformations (e.g., *log*)
- **COORD**: a coordinate system (e.g., *polar*)
- **ELEMENT**: visual objects (e.g., *points*) and their aesthetic attributes (e.g., *color, position*)
- **GUIDE**: guides (e.g., *axes, legends*)

Specification for a scatter plot

- DATA: $x = x$
- DATA: $y = y$
- TRANS: $x = x$
- TRANS: $y = y$
- SCALE: $\text{linear}(\text{dim}(1))$
- SCALE: $\text{linear}(\text{dim}(2))$
- COORD: $\text{rect}(\text{dim}(1, 2))$
- GUIDE: $\text{axis}(\text{dim}(1))$
- GUIDE: $\text{axis}(\text{dim}(2))$
- ELEMENT: $\text{point}(\text{position}(x^*y))$

Graph visual components

- Data components
 - ◆ Visual objects associated to measures
 - ◆ Visual attributes
- Layout
 - ◆ Positioning rules (e.g. cartesian coord)
- Support components
 - ◆ Axes
 - ◆ Labels
 - ◆ Legends

VISUAL ENCODING

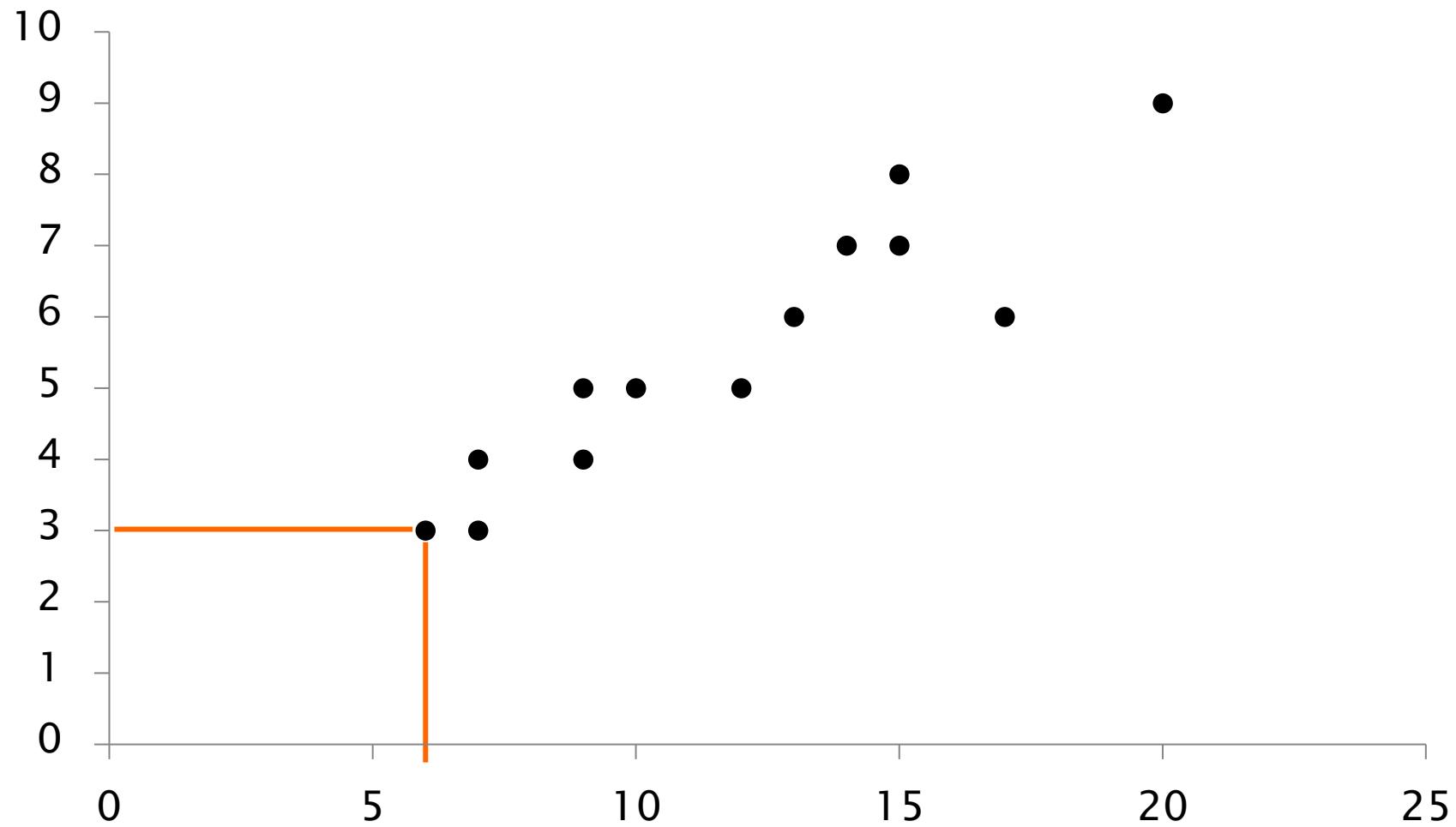
Visual Encoding

- Given a variable (measure), identify:
 - ◆ Visual object
 - ◆ Visual attribute
- Main distinction
 - ◆ Quantitative (interval, ratio, absolute)
 - ◆ Categorical (nominal, ordinal)

Visual Encoding

Object	Attribute
Point	Position (w.r.t. axis/axes)
Line	Length Position (w.r.t. axis/axes) Slope
Bar	Length
Shape	Size (area) Count

Points



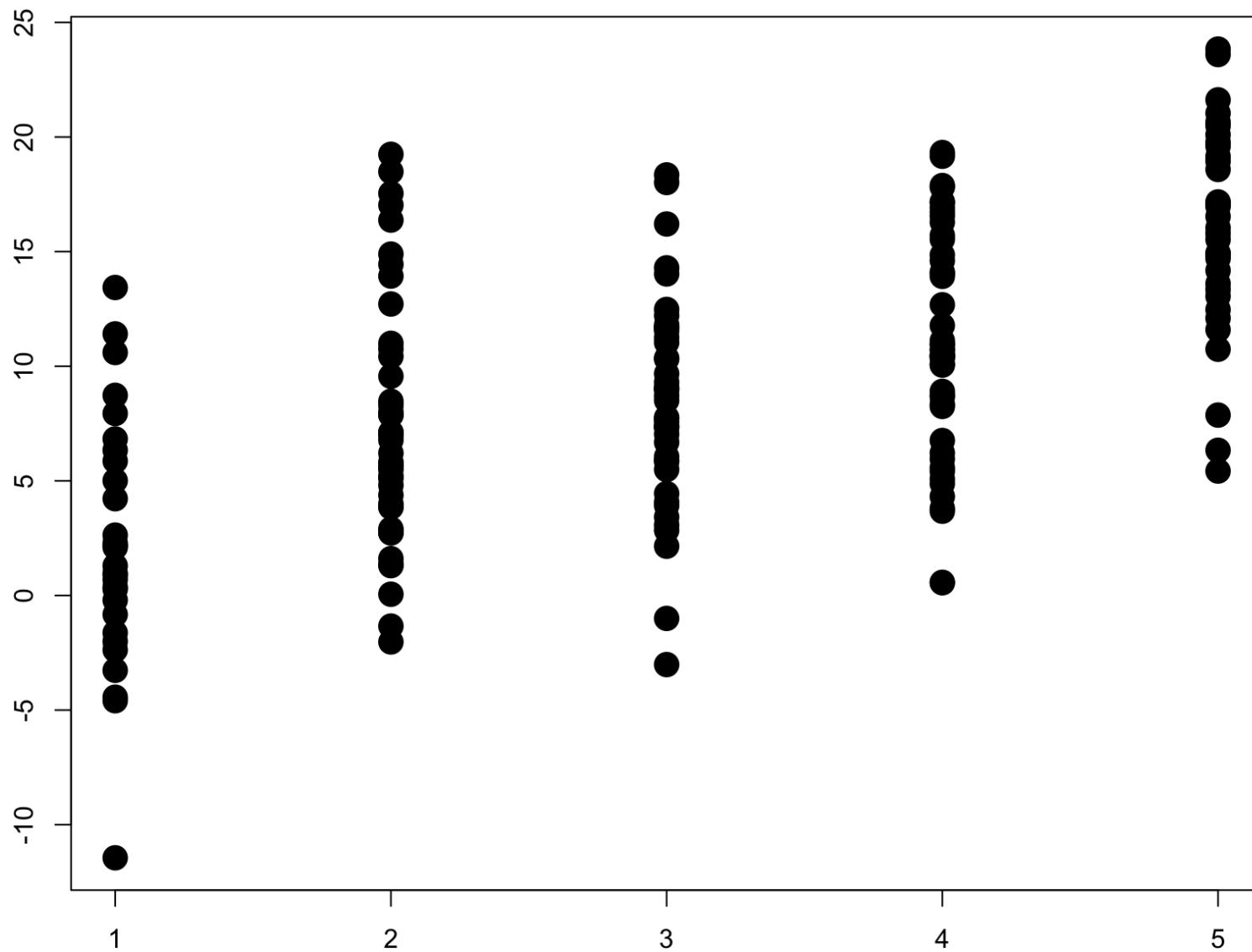
Points Guidelines

- Points must be clearly distinguished
 - ◆ Enlarge points
 - ◆ Select radically distinct shapes (+ ○)
 - ◆ Balance size of points and graph
 - ◆ Use outlined shapes
- Lines must not obscure points

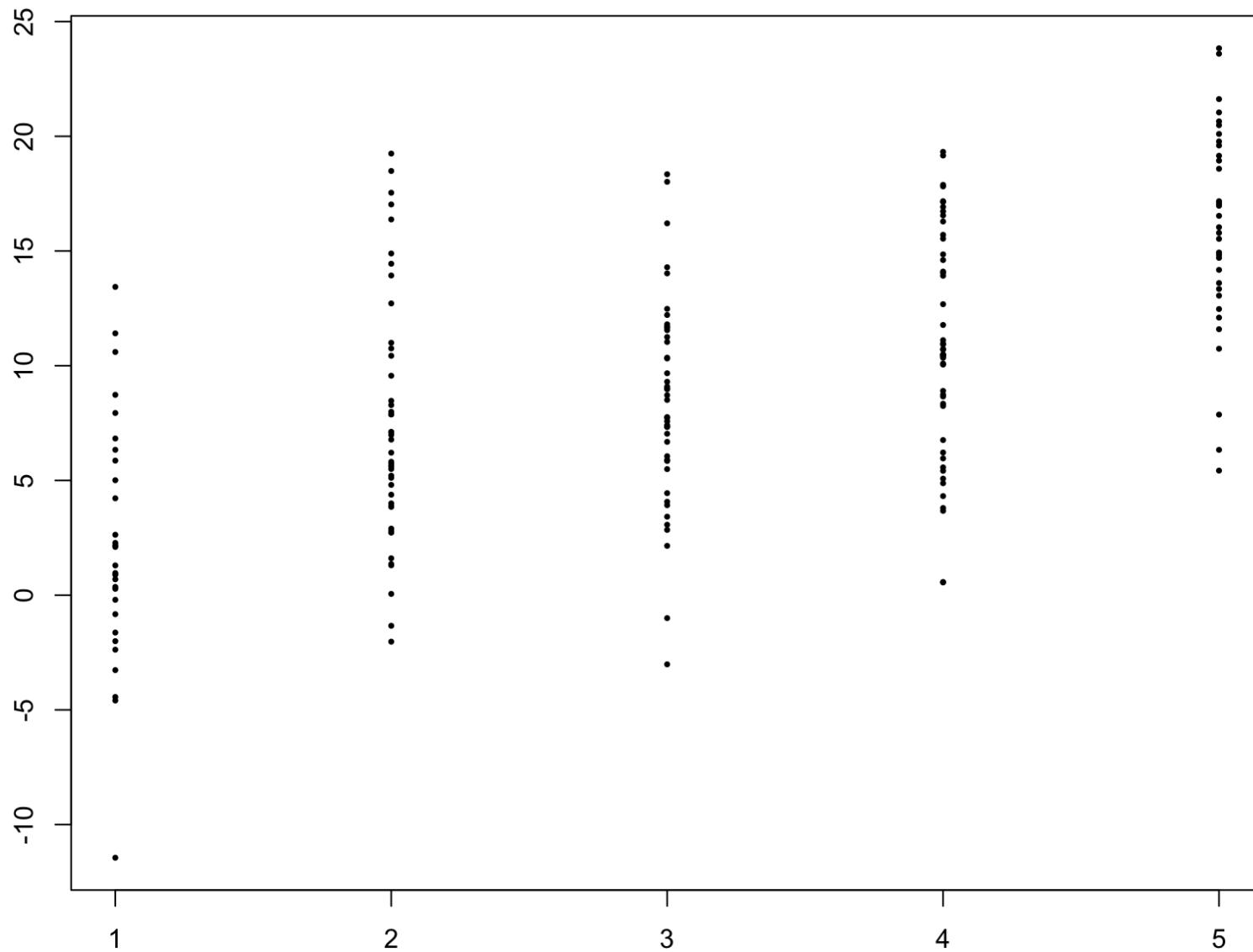
Overplotting

- Phenomenon related to multiple points (or shapes) overlapping
 - ◆ Discrete (integer) measure
 - ◆ Very large dataset
- Solutions
 - ◆ Small shapes
 - ◆ Outlined shapes
 - ◆ Transparent shapes (alpha)
 - ◆ Jittering

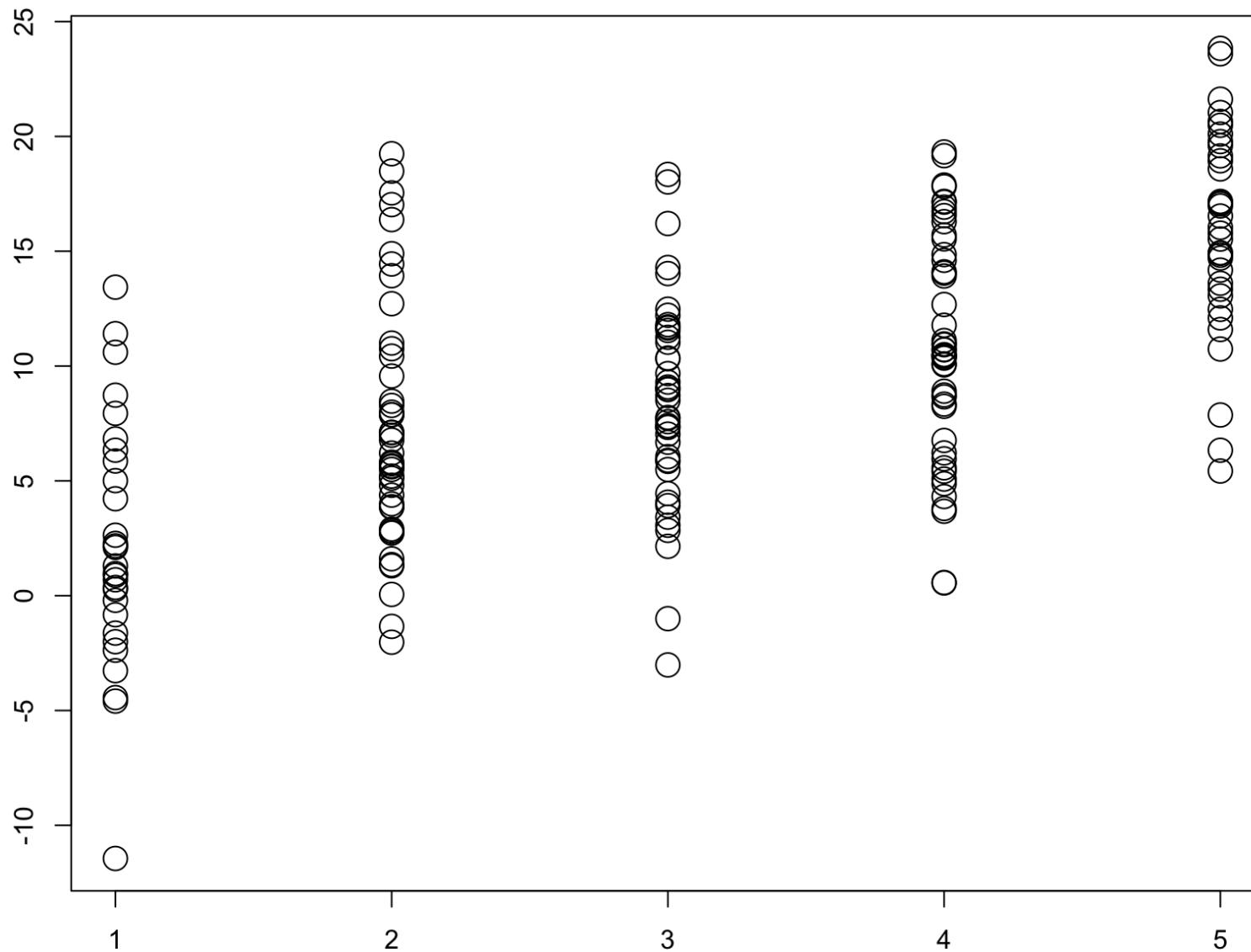
Overplotting example



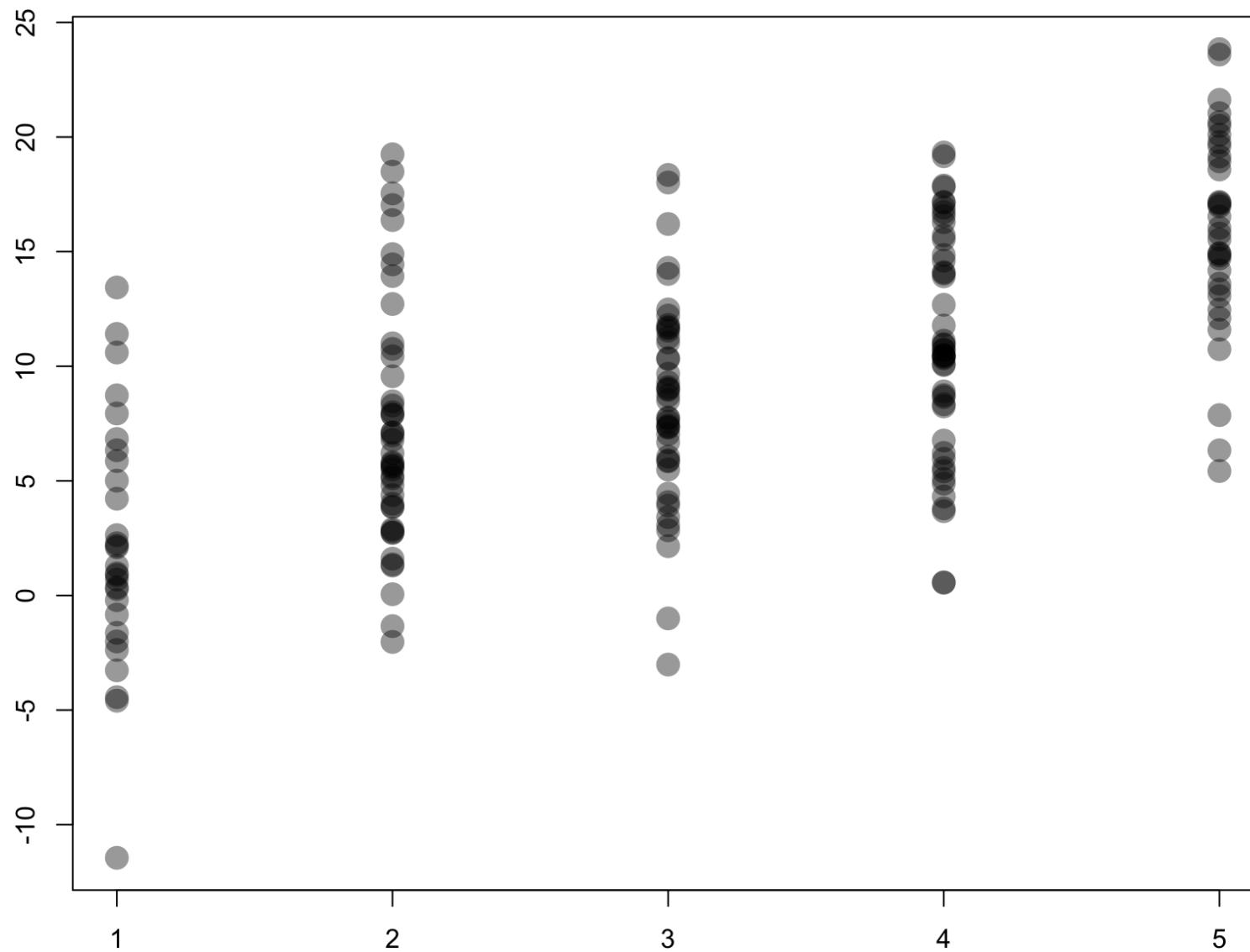
Overplotting – Small



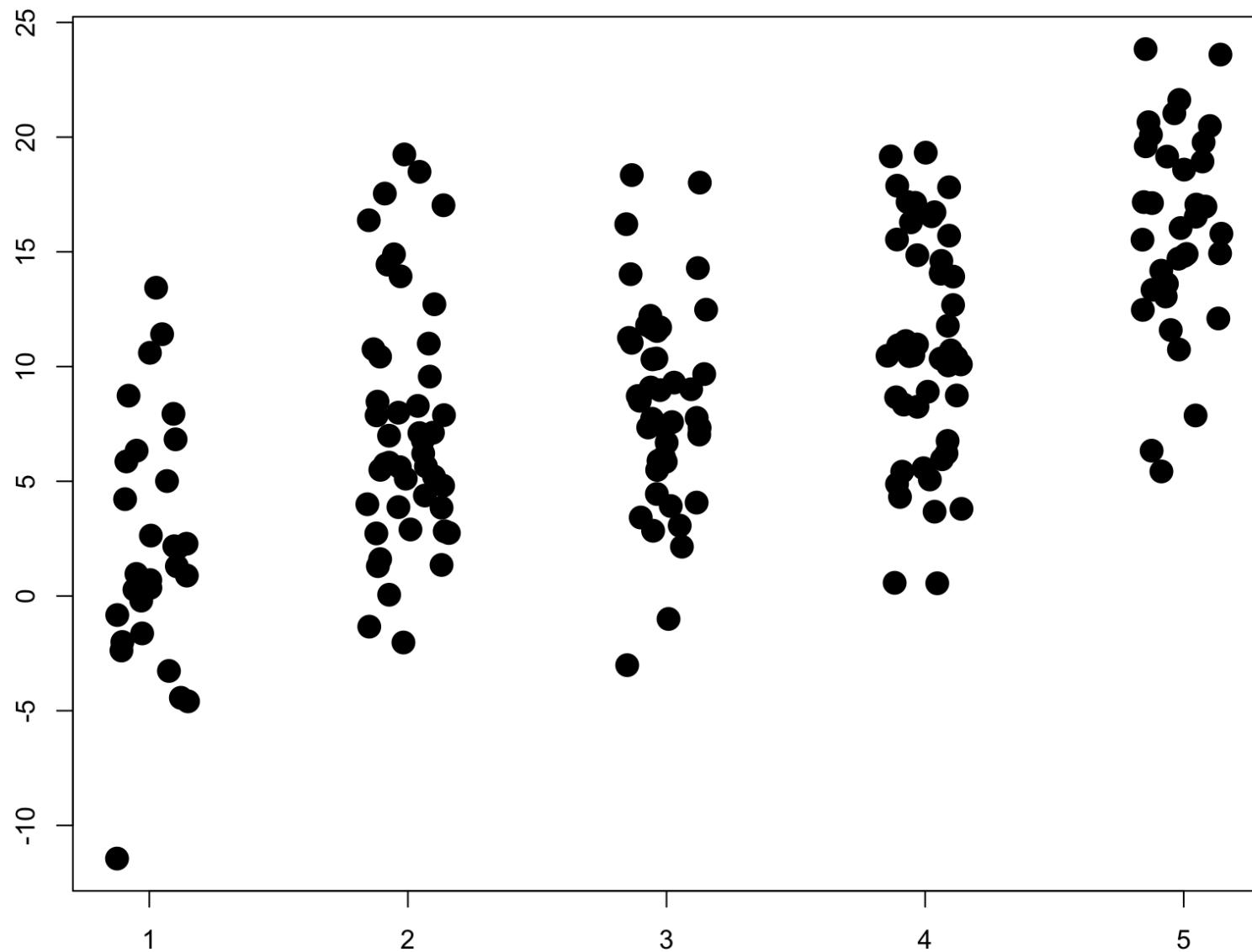
Overplotting - Outlined



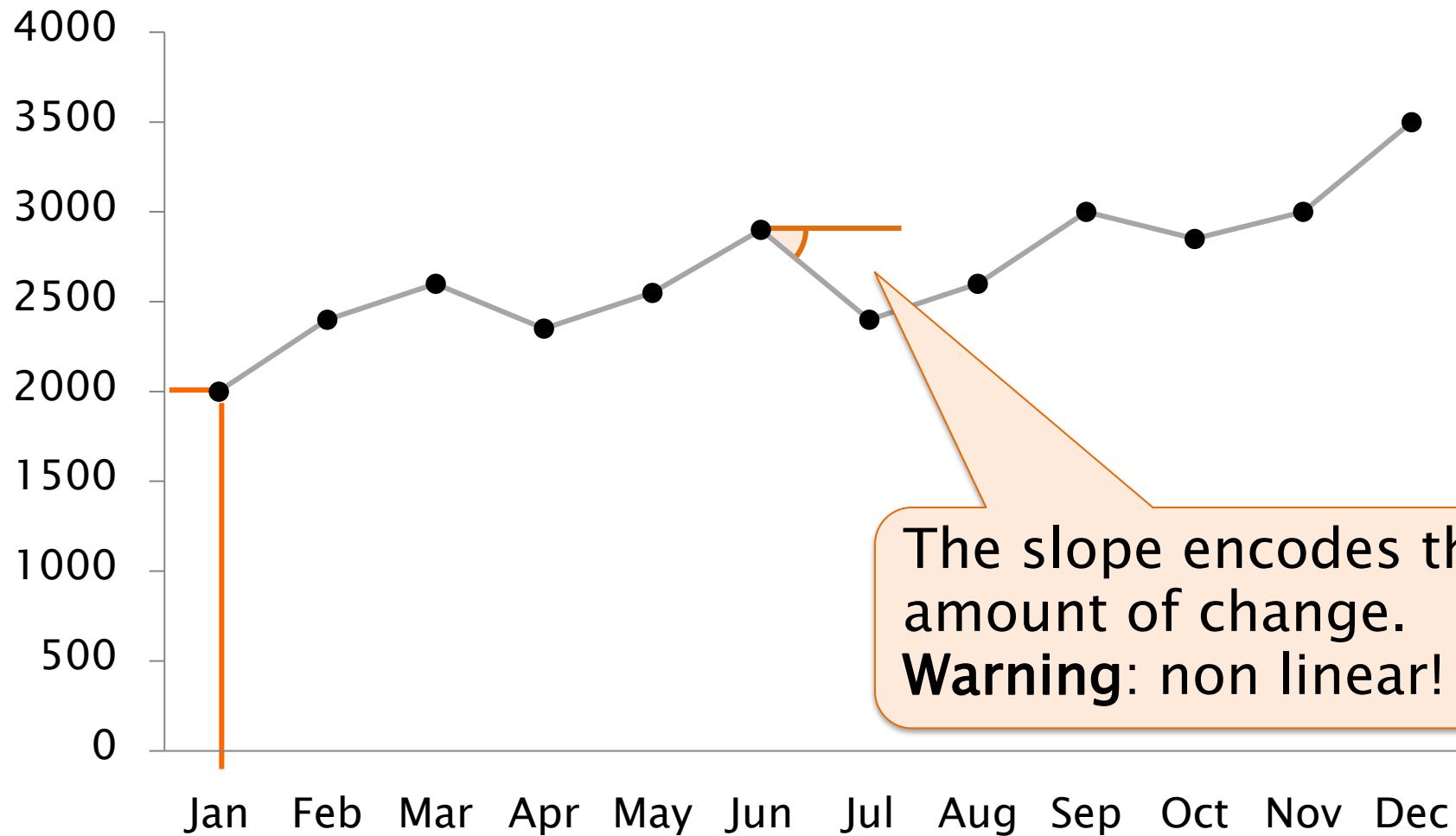
Overplotting – Transparent



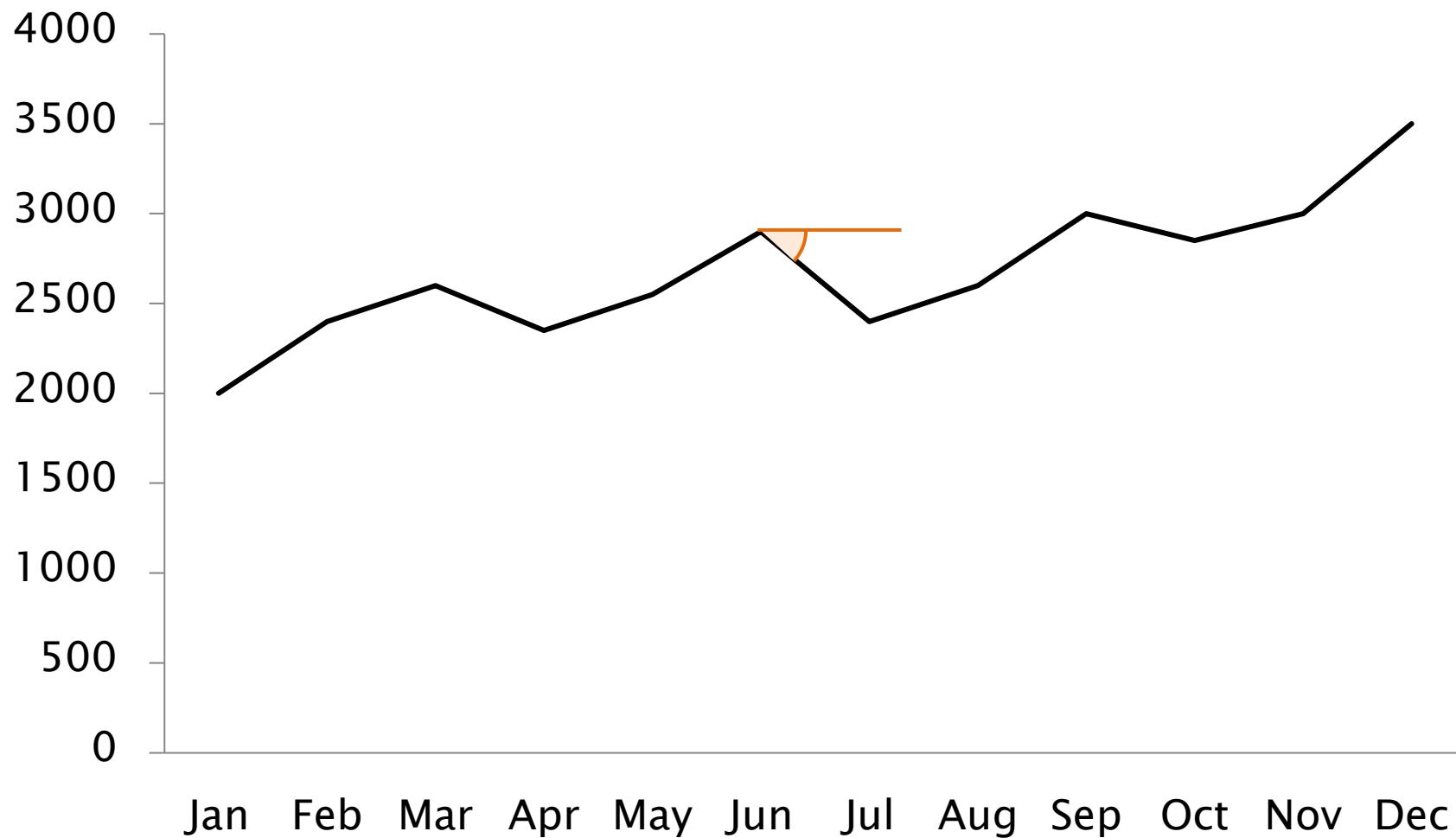
Overplotting – Jittering



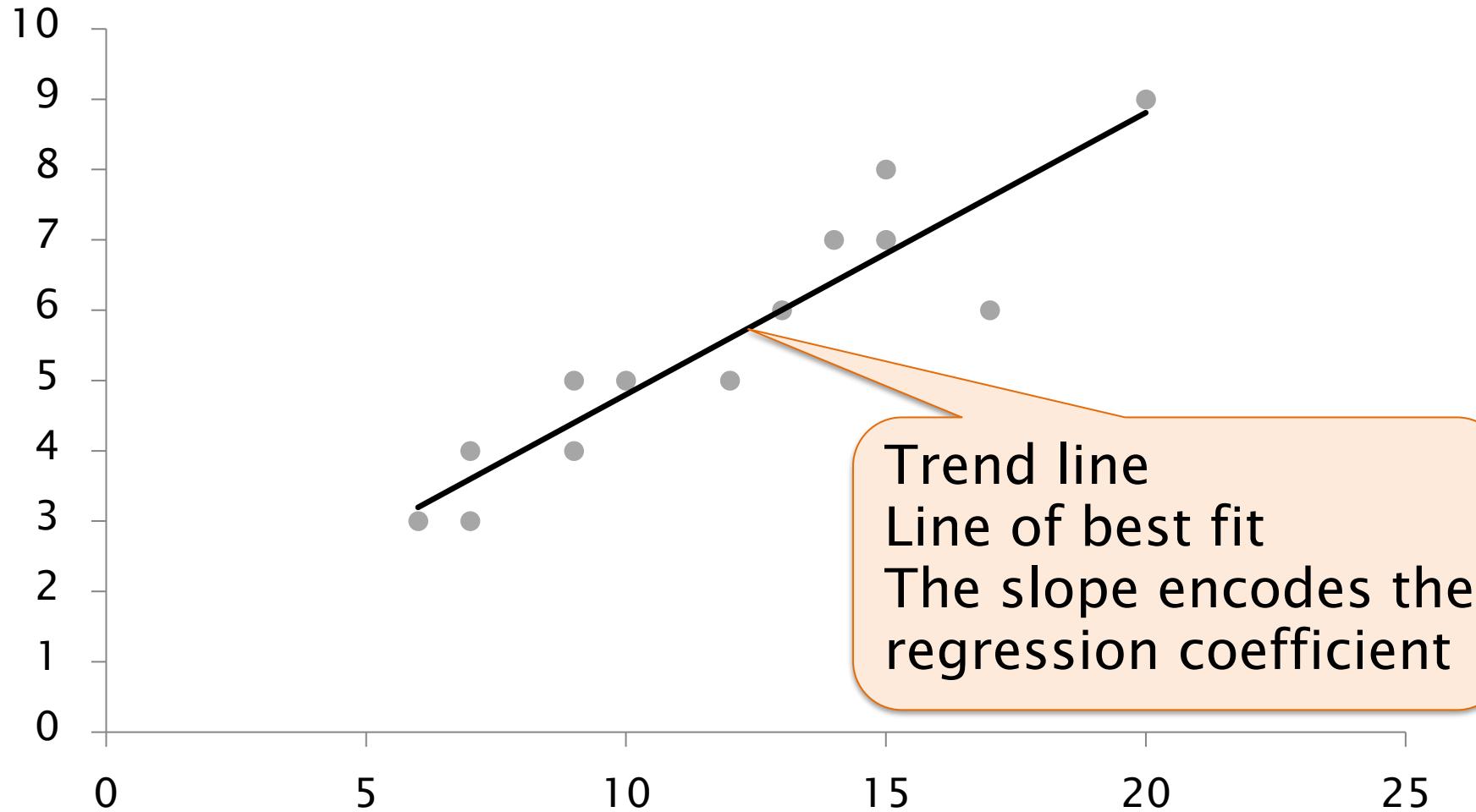
Points and Lines



Slope of lines



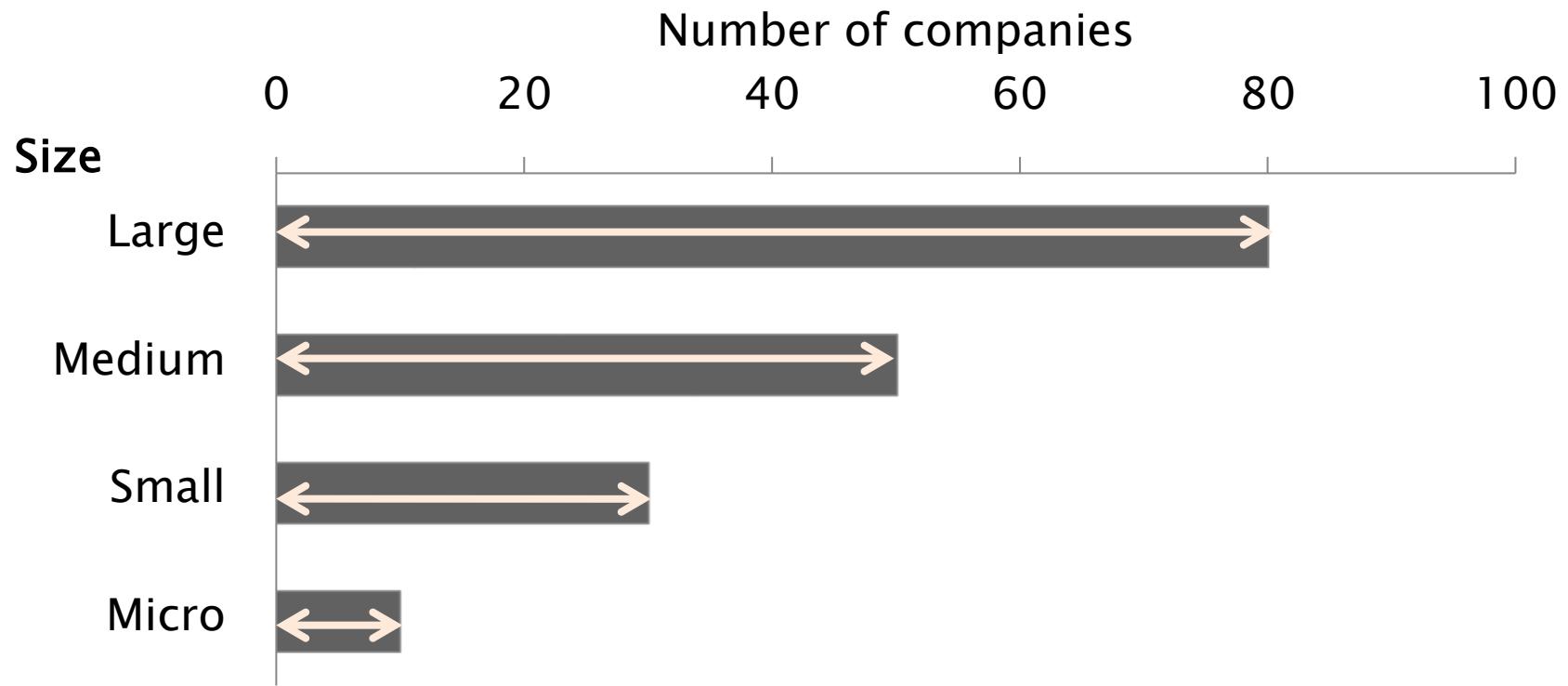
Slope of lines



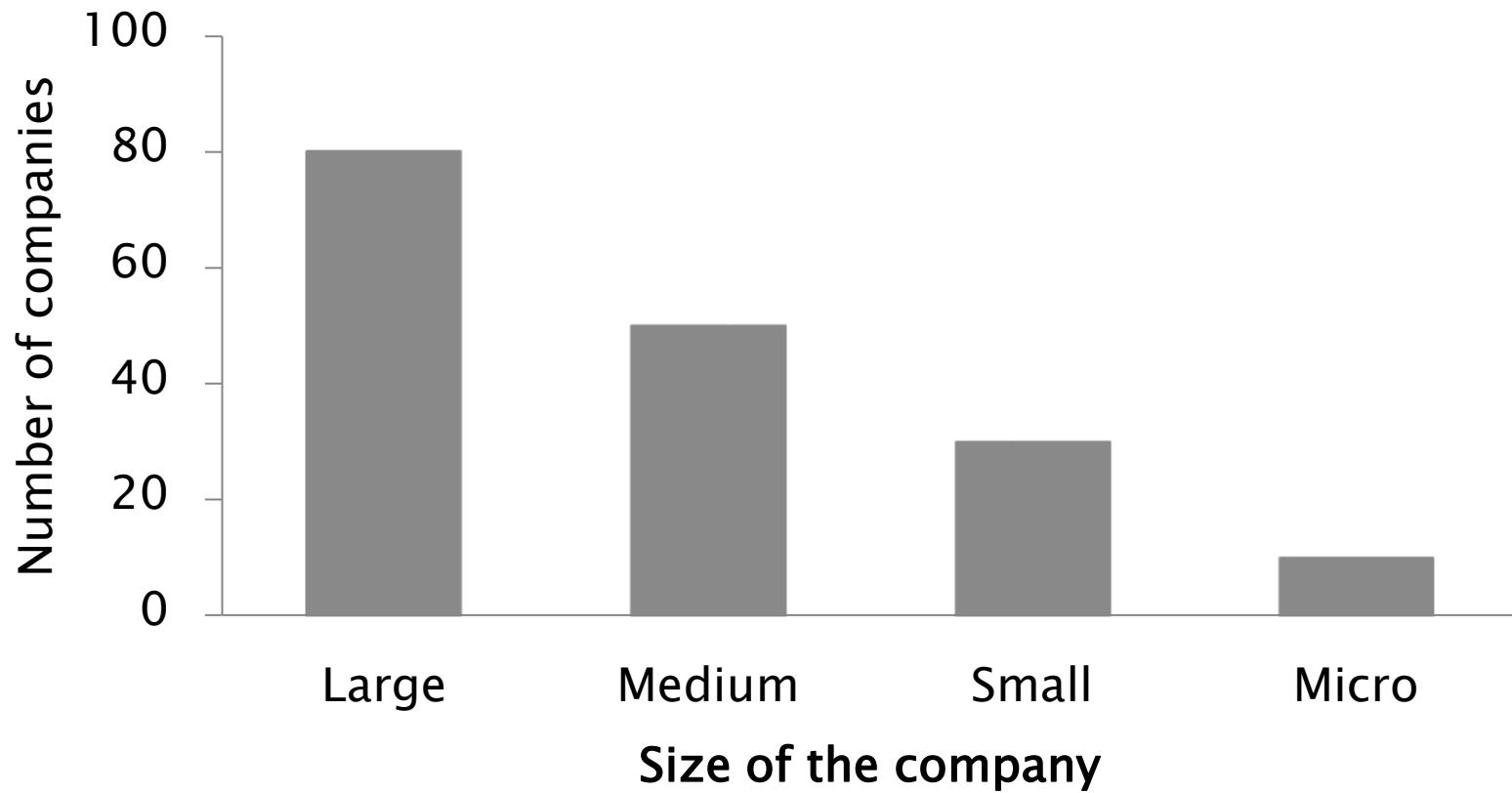
Lines

- Easy perception of trends and overall shape of data
- Best suited for time series
- Variation encoded as slope
 - ◆ Clear direction
 - ◆ Approximate magnitude

Bars (line length)



Vertical Bars (aka Columns)



Bars

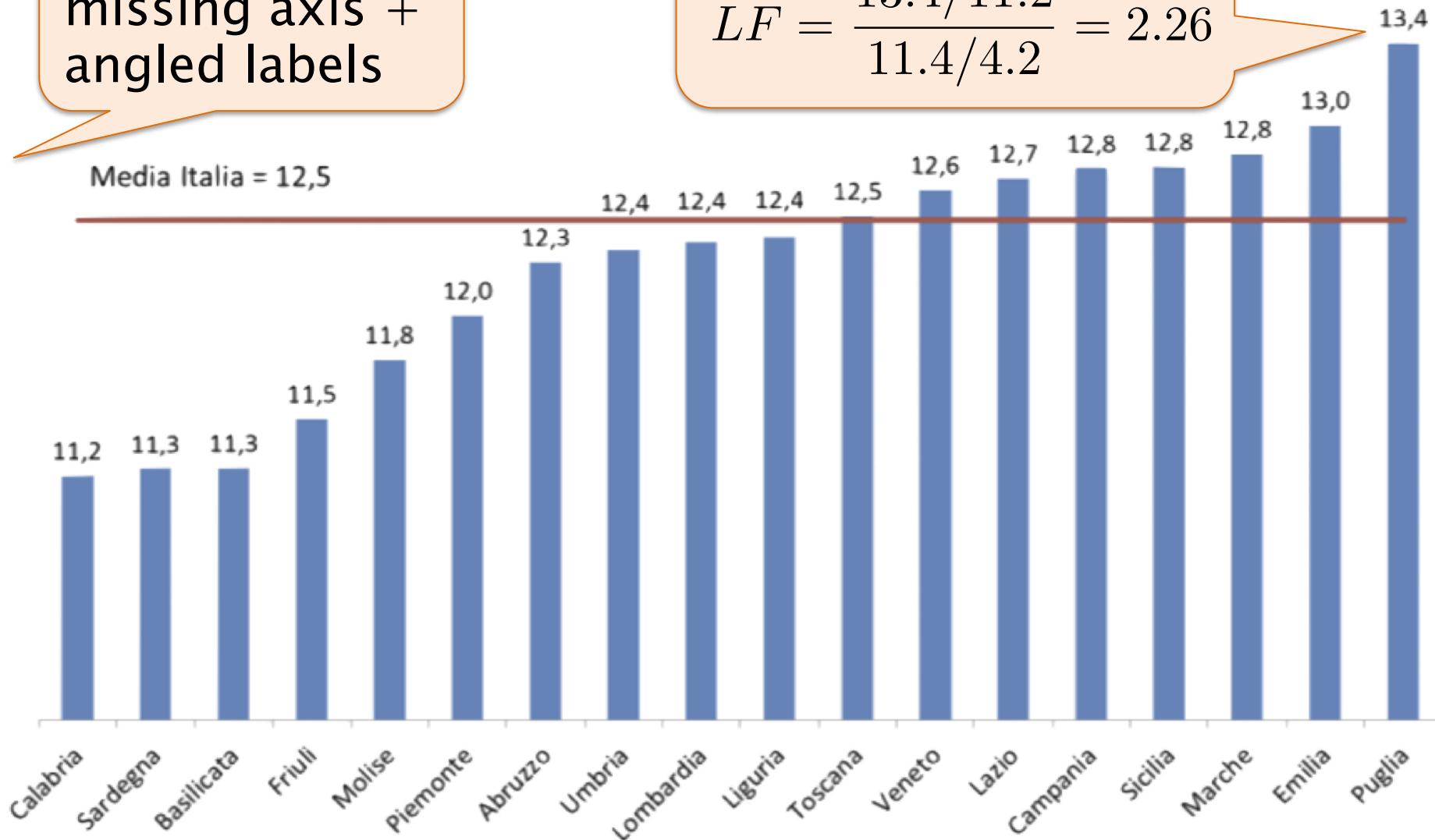
- Quantitative values are encoded only as length of the bars
- Width of bars plays no role
 - ◆ Bars are just very thick lines
- Bars require a zero-based scale
 - ◆ See: Lie factor!

Bar must be zero based

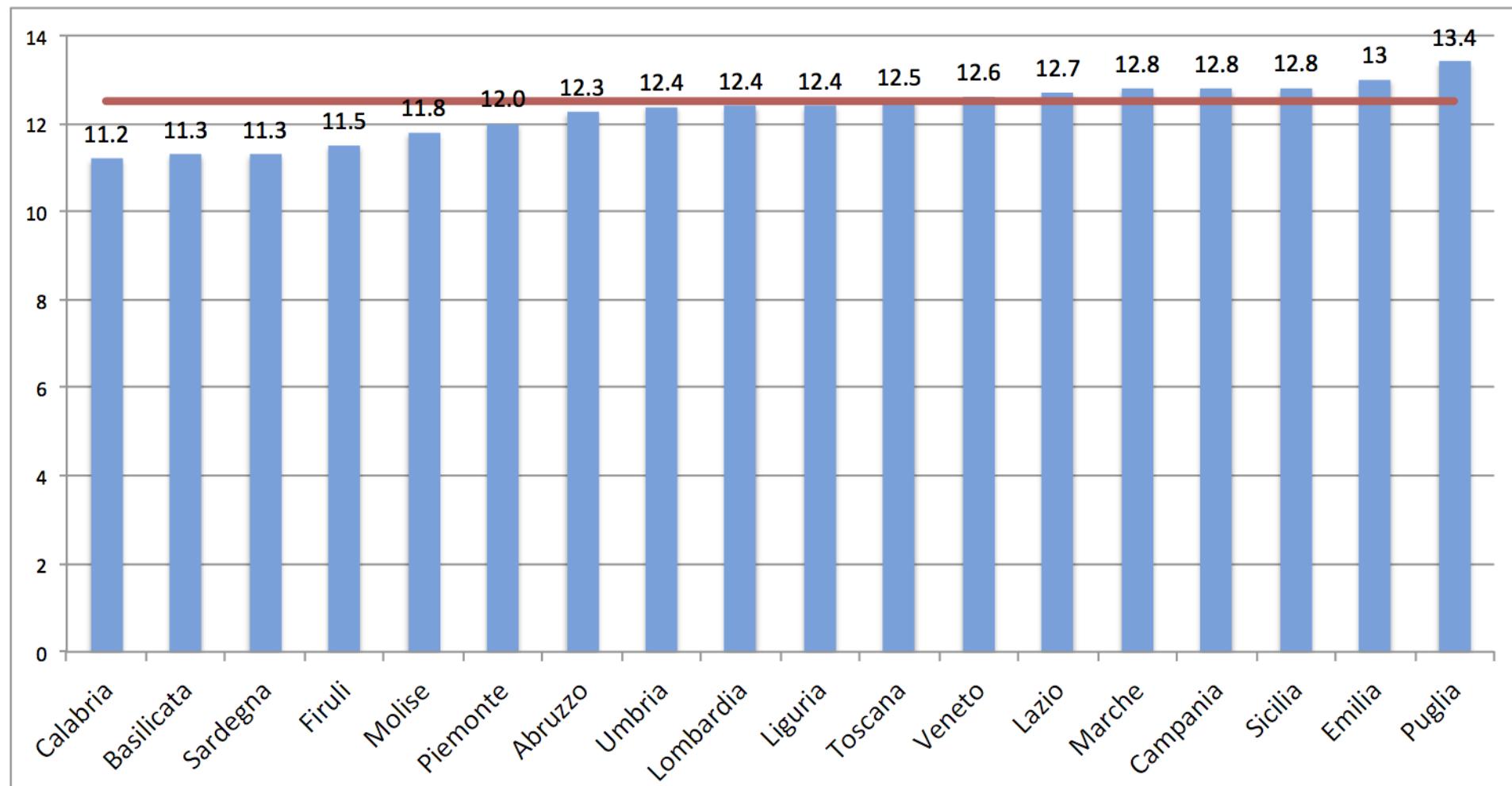
Clarity:
missing axis +
angled labels

Proportionality:

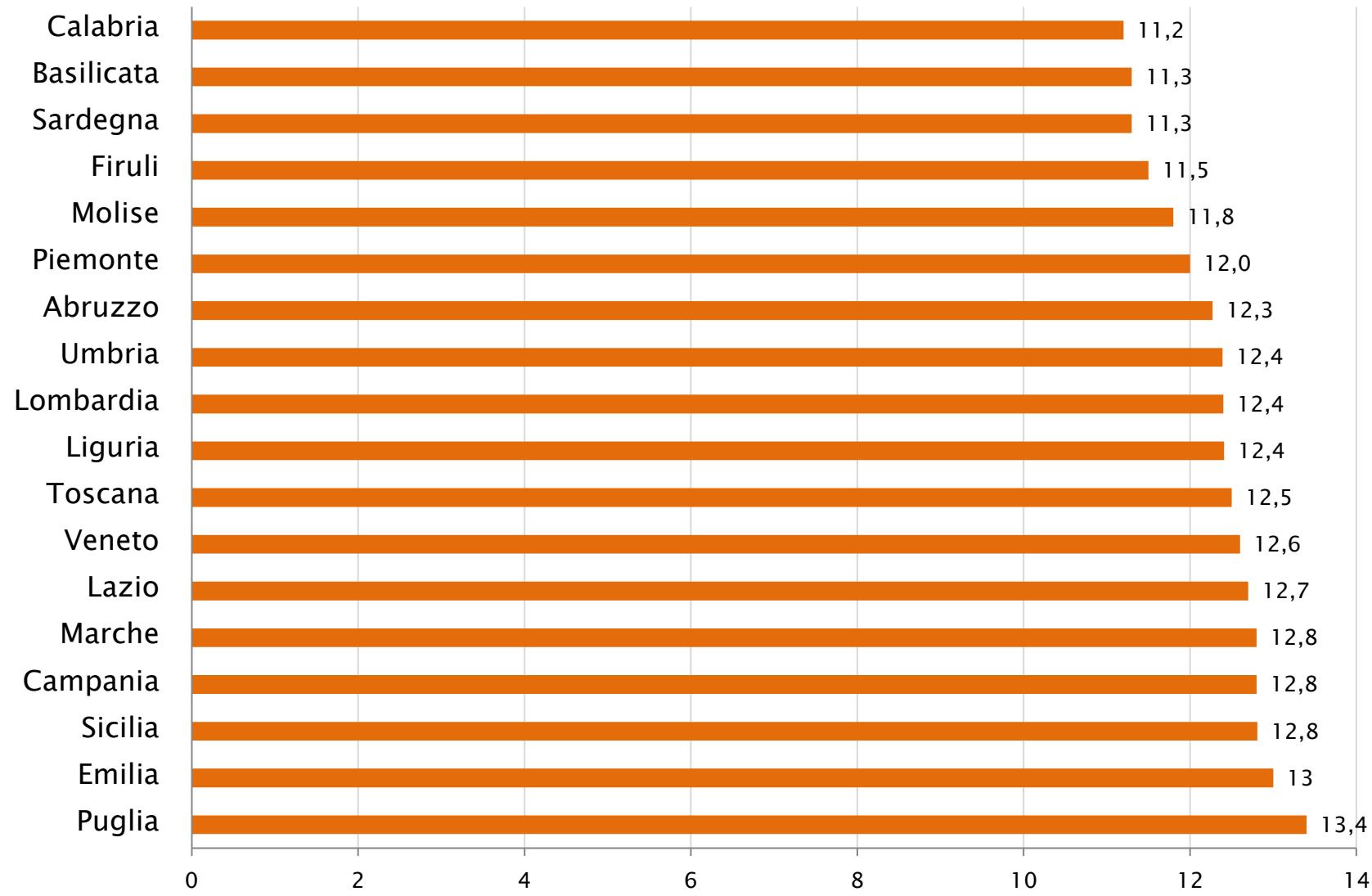
$$LF = \frac{13.4/11.2}{11.4/4.2} = 2.26$$



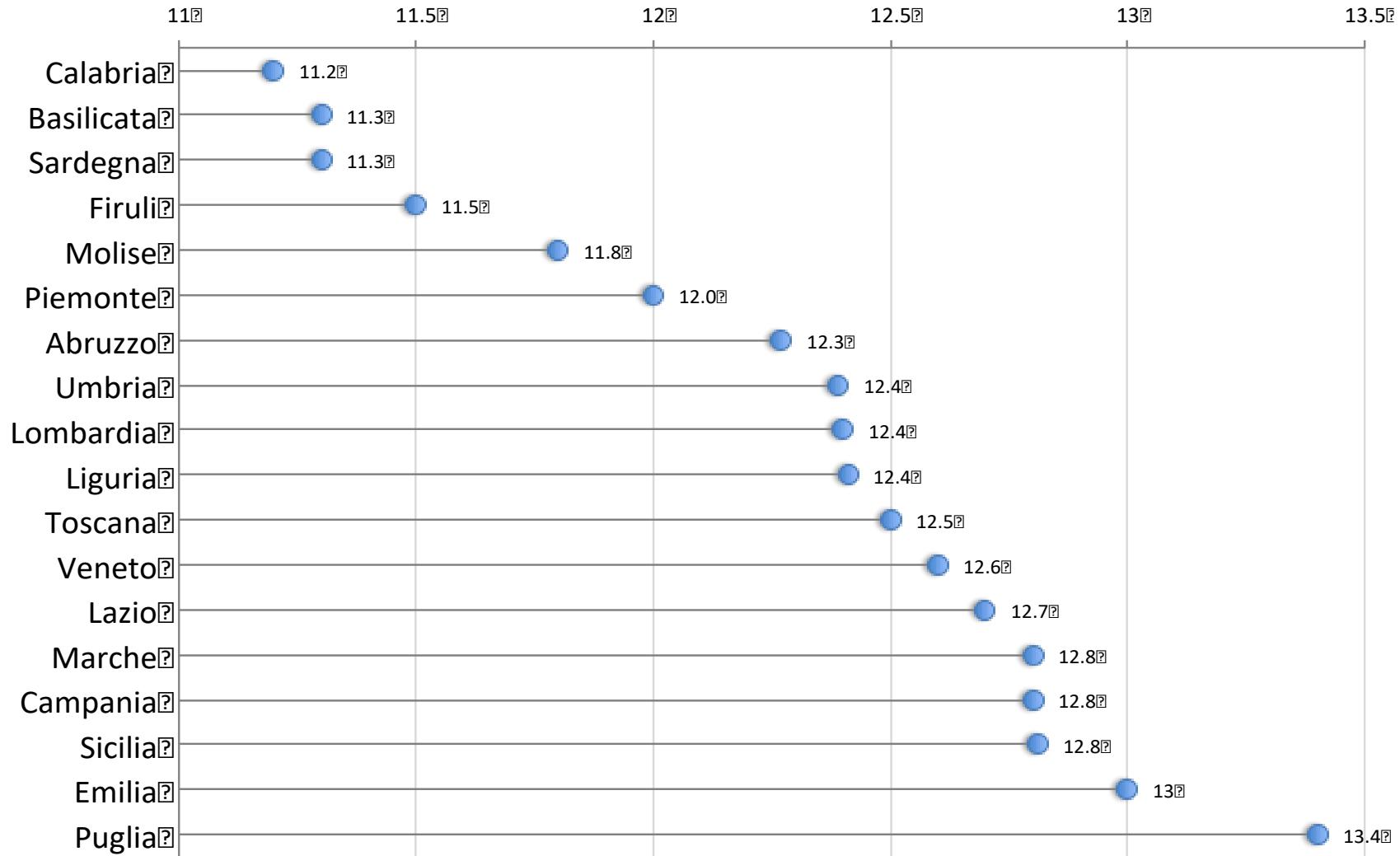
Bar are zero based



Horizontal let longer labels

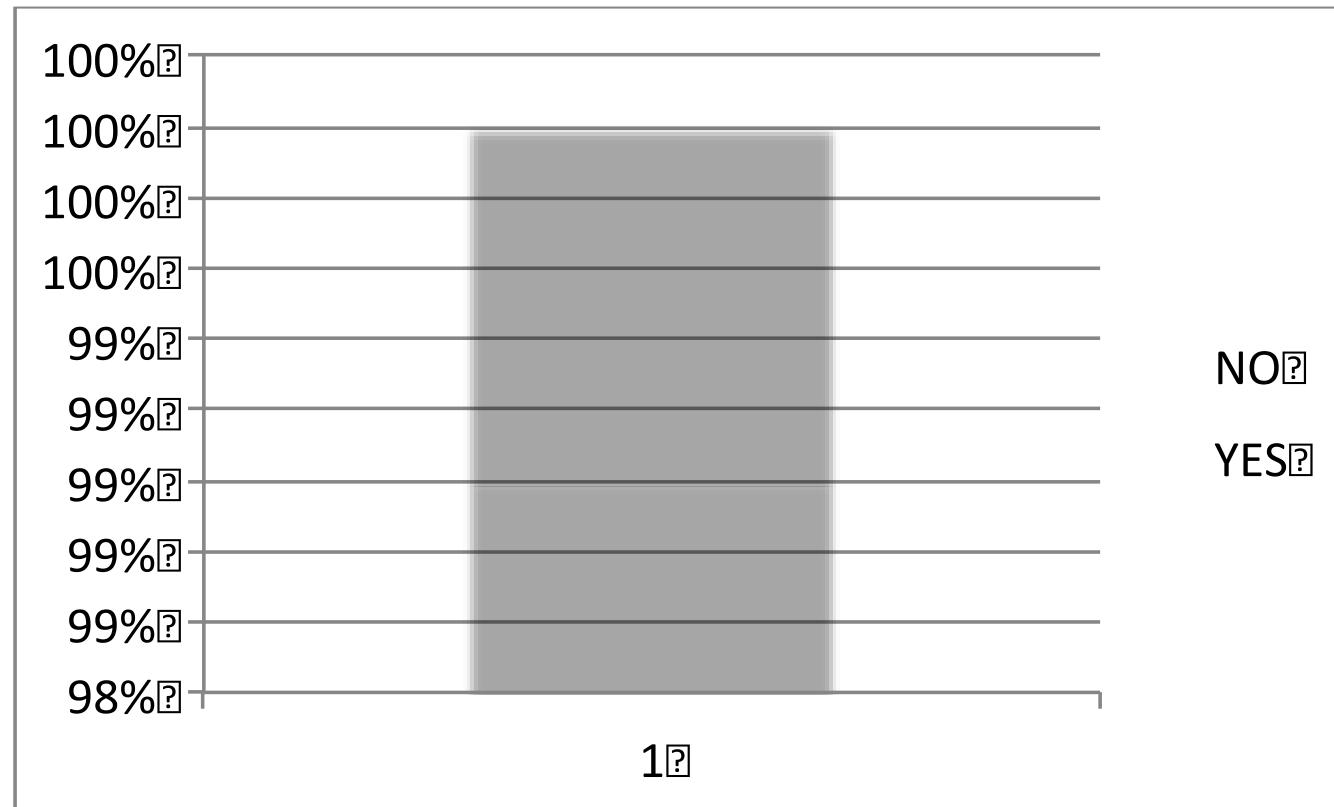
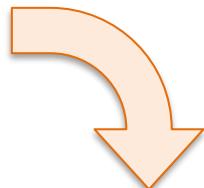


Dot plot (Lollipop)



Beware MS-Excel Default

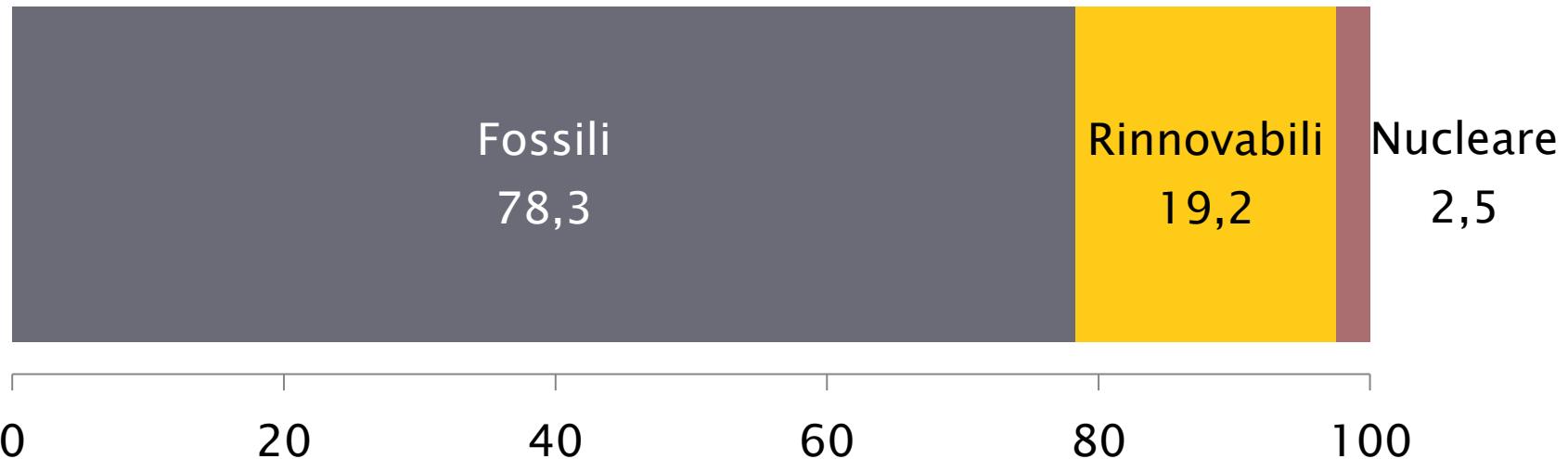
	A	B
1	YES	99%
2	NO	1%



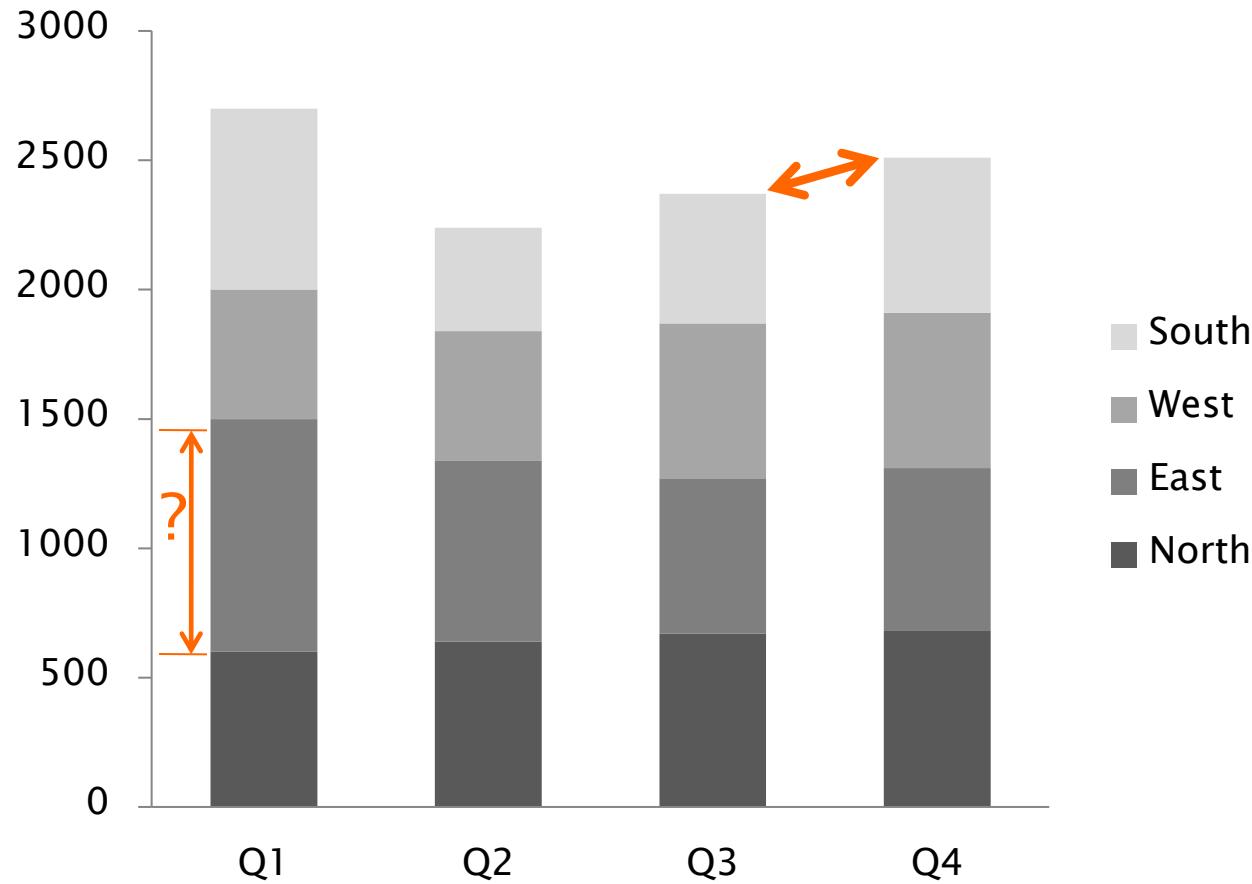
Bars Guidelines

- Use horizontal bars when
 - ◆ A descending order ranking
 - ◆ Categorical label don't fit
- Proximity
 - ◆ Use a 1:1 bar:spacing ratio $\pm 50\%$
 - ◆ No spacing between bars that are not labeled on the axis (legend categories)
 - ◆ No overlapping bars

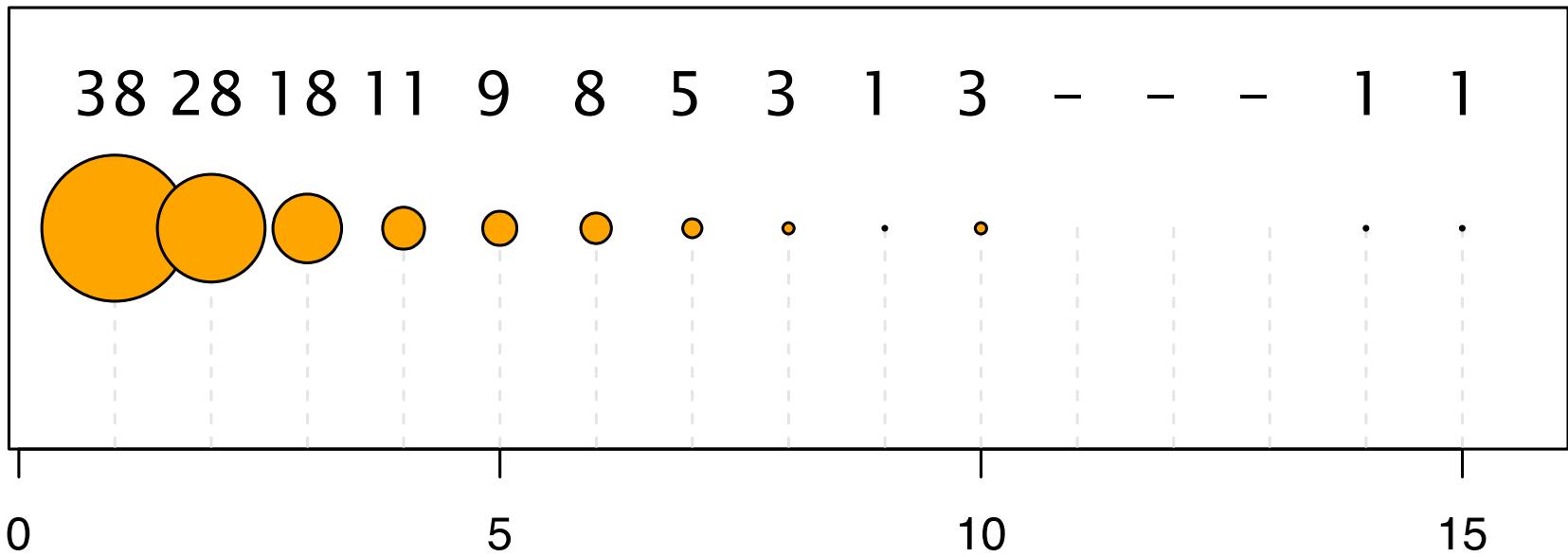
Length - Stacked Bars



Nonaligned bars – Stacked



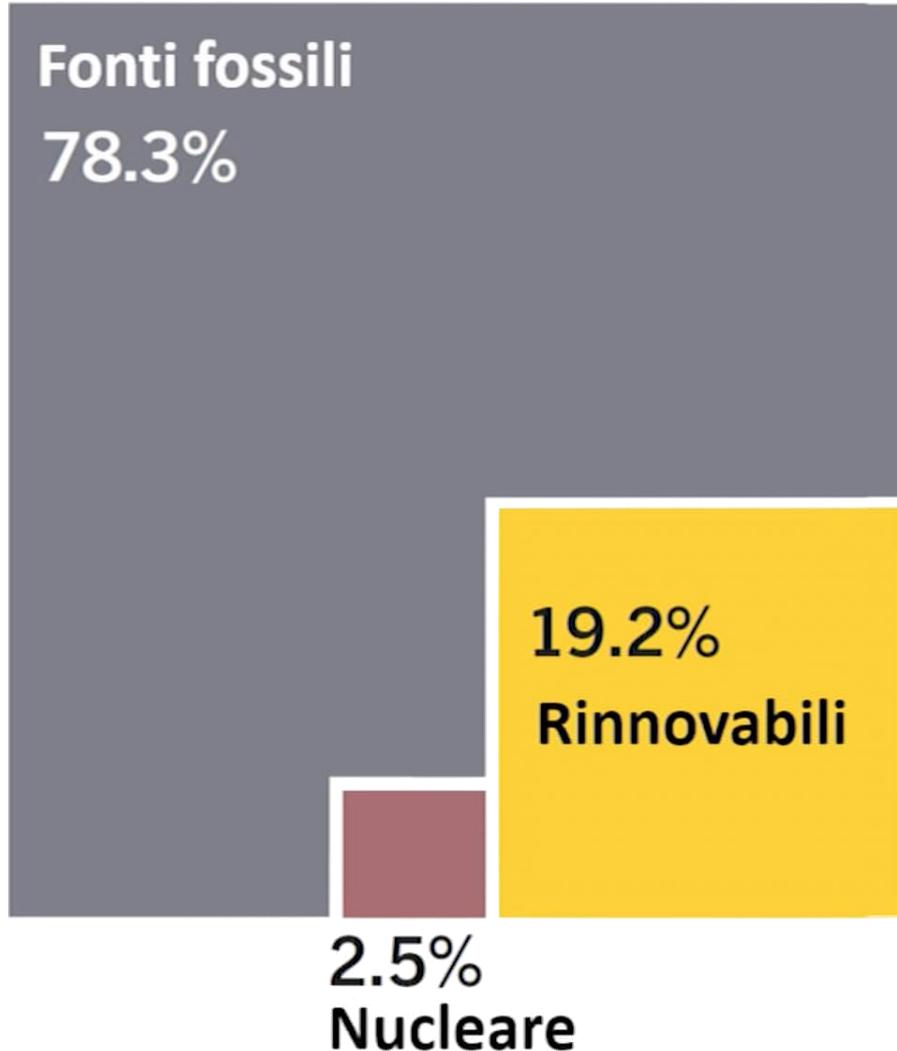
Shape area – Bubbles



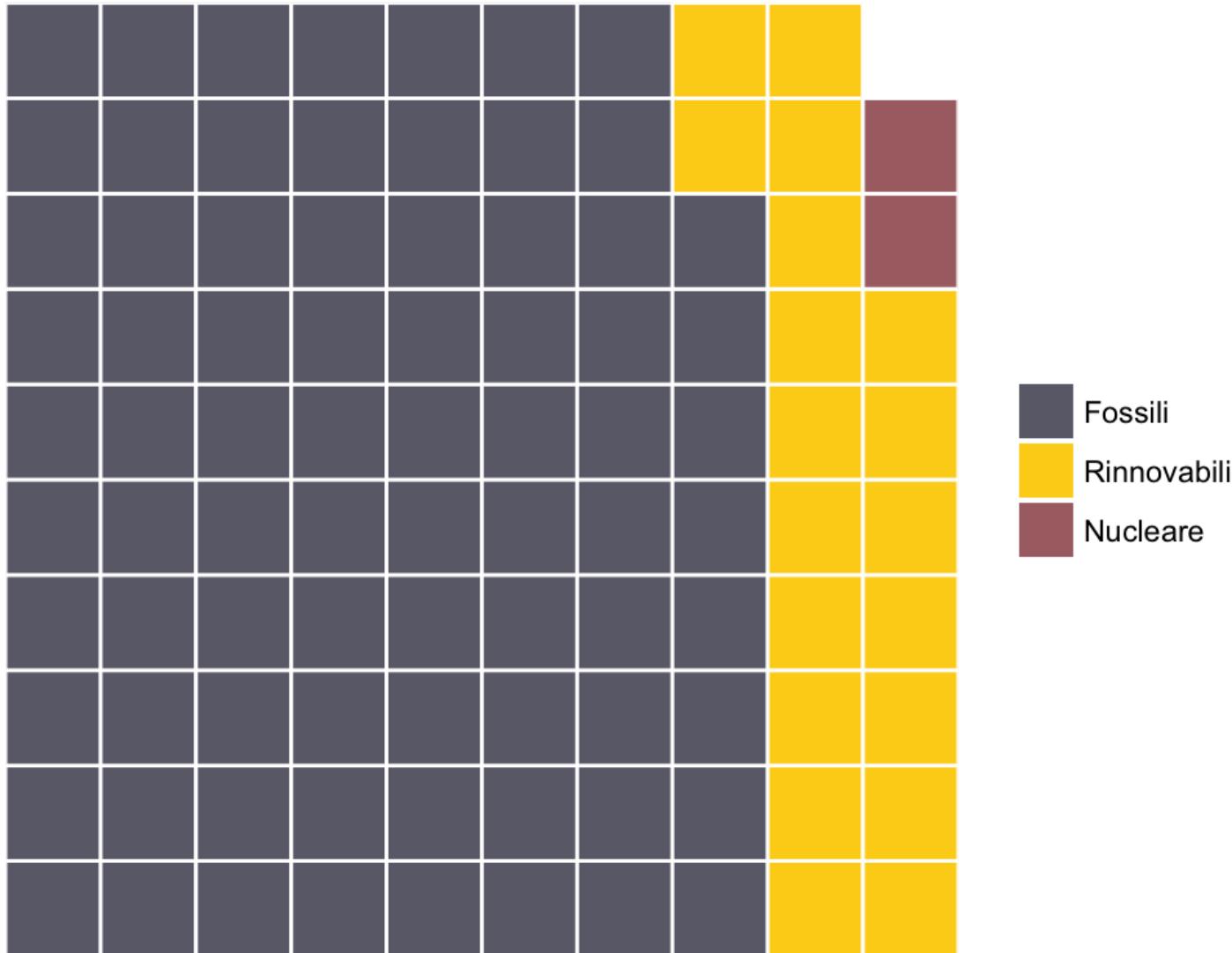
Area – Treemap



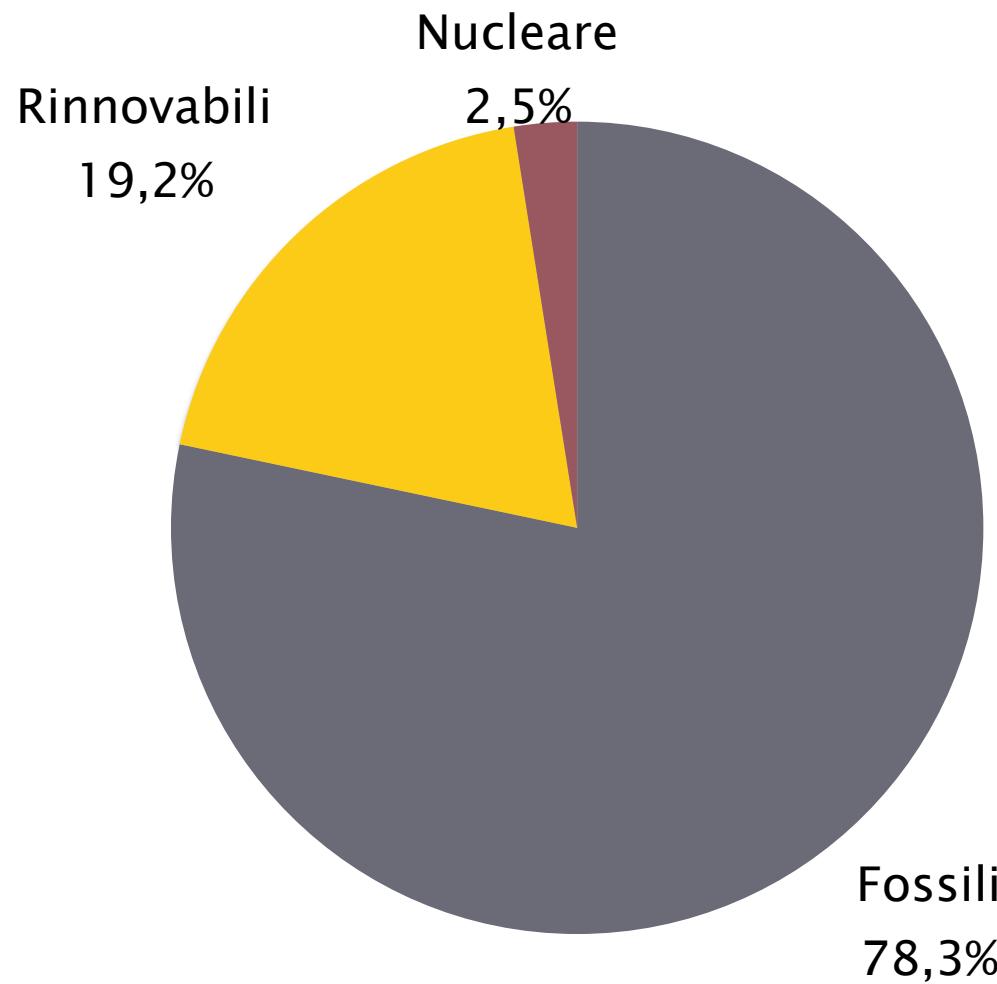
Area – Treemap



Area + Count – Waffle / Grid



Area + Angle – Pie Chart



Count – Isotype

- Isotype
 - ◆ International System Of Typographic Picture Education
- Marie and Otto Neurath
 - ◆ Vienna, 1936

Literacy in England and Wales

Among 10 men

Illiterates

1841



1871



1901



1931



Among 10 women

1841



1871



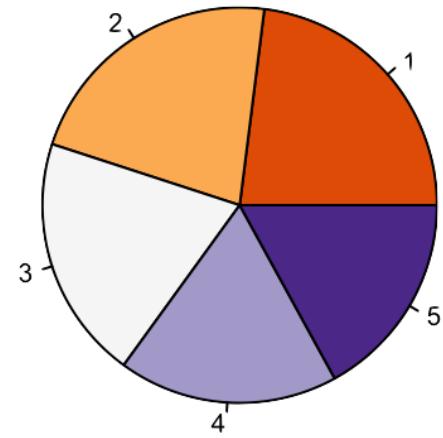
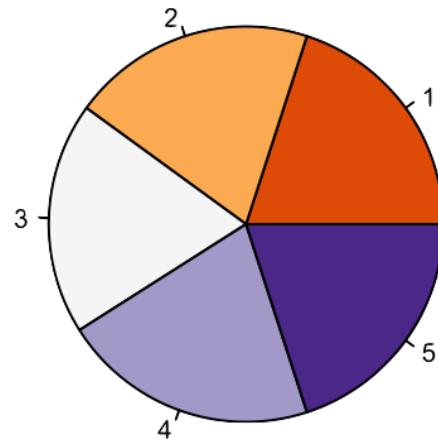
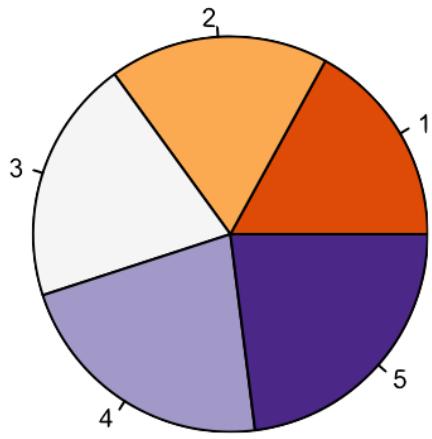
1901



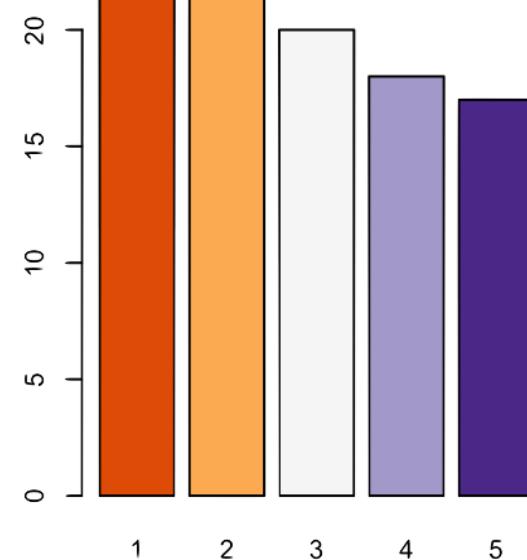
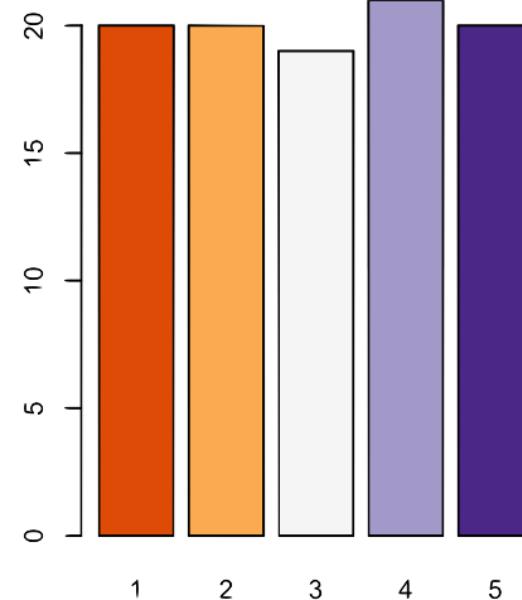
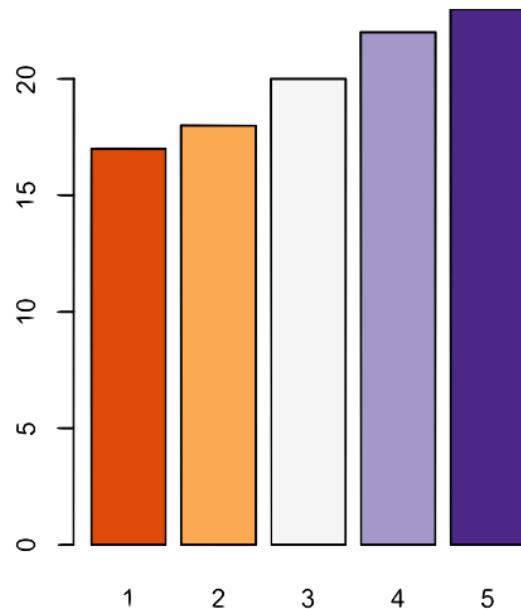
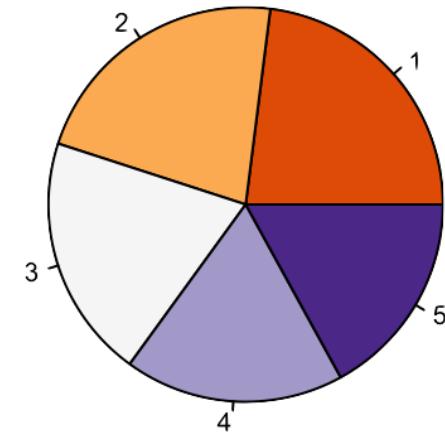
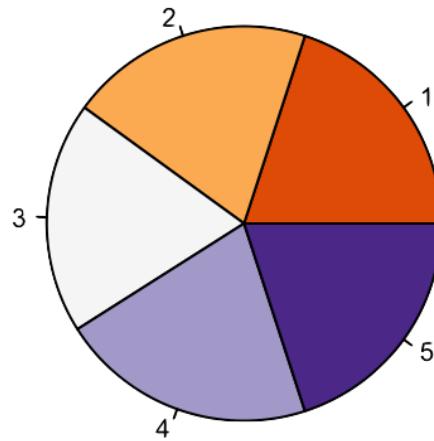
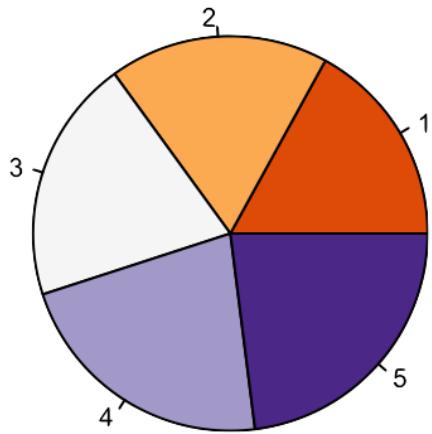
1931



Pies



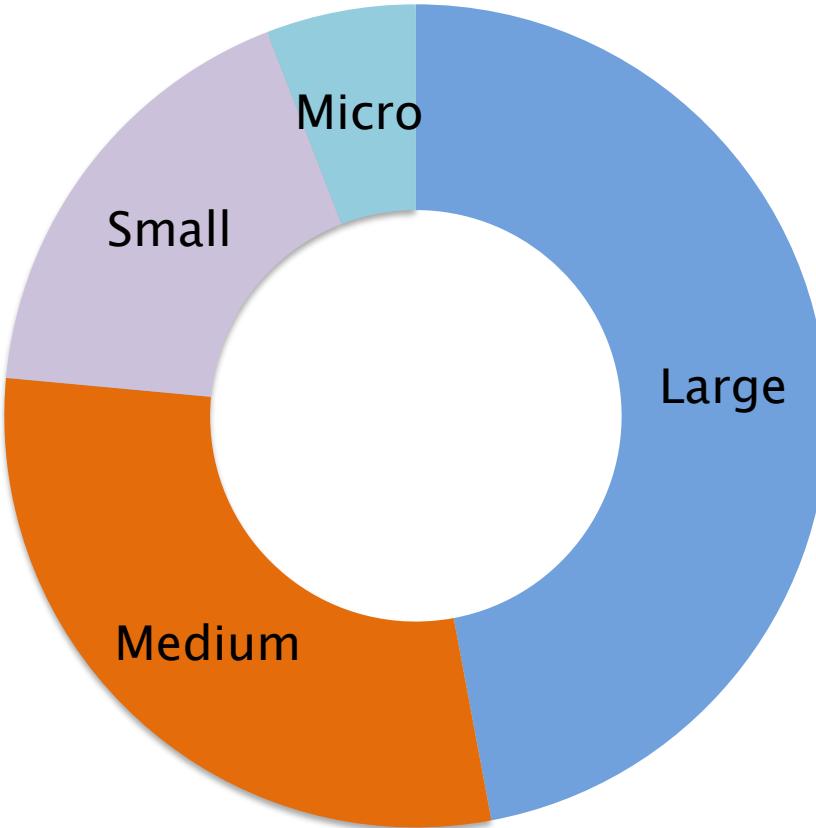
Pies vs. Bars



Pie Charts: guidelines

- Have serious limitations
 - ◆ To represent part–whole relationship
 - ◆ Only with a small number of categories
 - Up to four
 - Avoid rainbow pie
 - ◆ When proportions are distinct enough
- Remember to ease reading
 - ◆ Labels placed close to slices
 - ◆ Labels include values (percentages)

Area/Angle/Length – Donut



Categorical encoding attributes

- Encoding of categorical levels

- ◆ Position (along an axis)

- ◆ Size

- ◆ Color

- Intensity

- Saturation

- Hue

- ◆ Shape

- ◆ Fill pattern

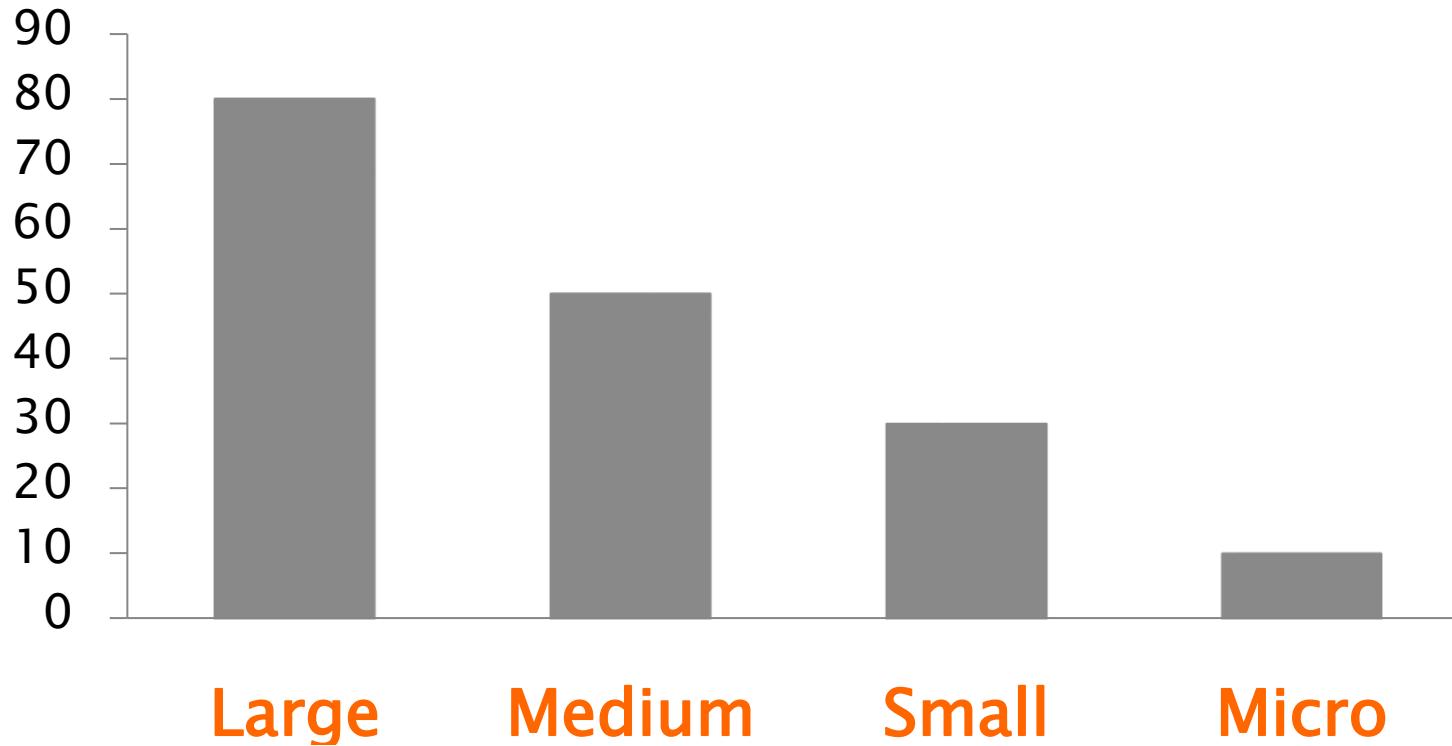
- ◆ Line style



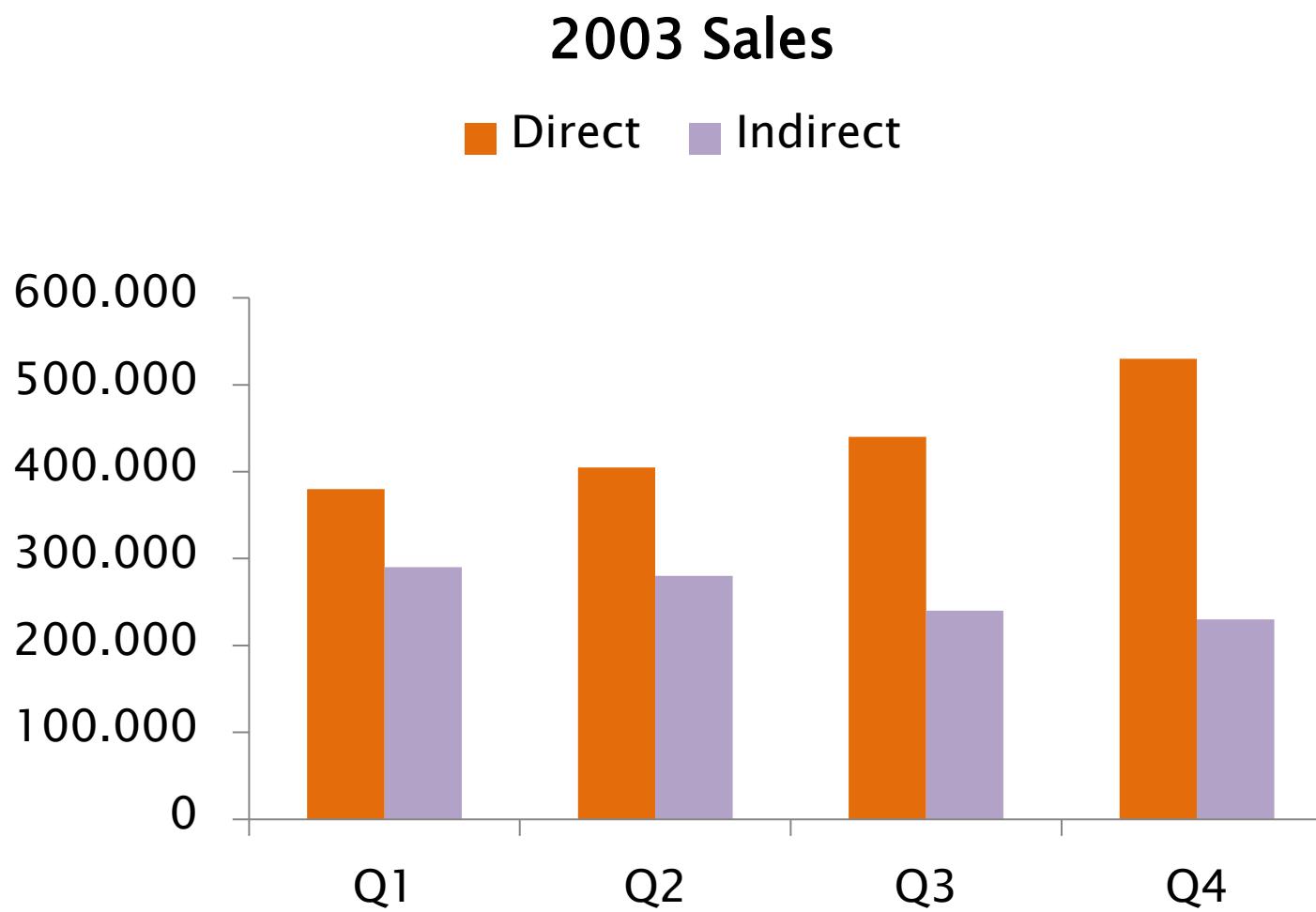
Ordinal

Position

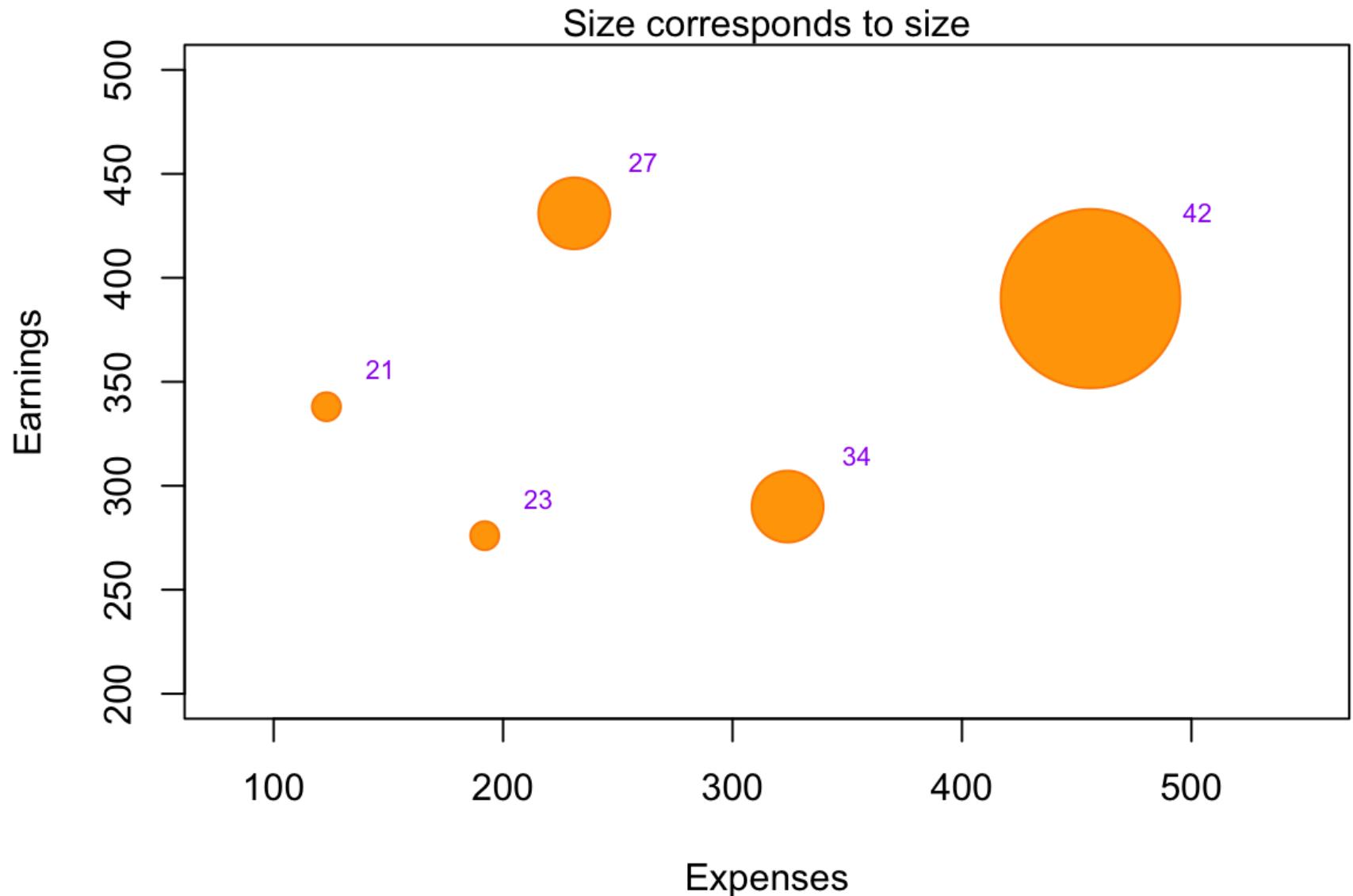
Number of
companies



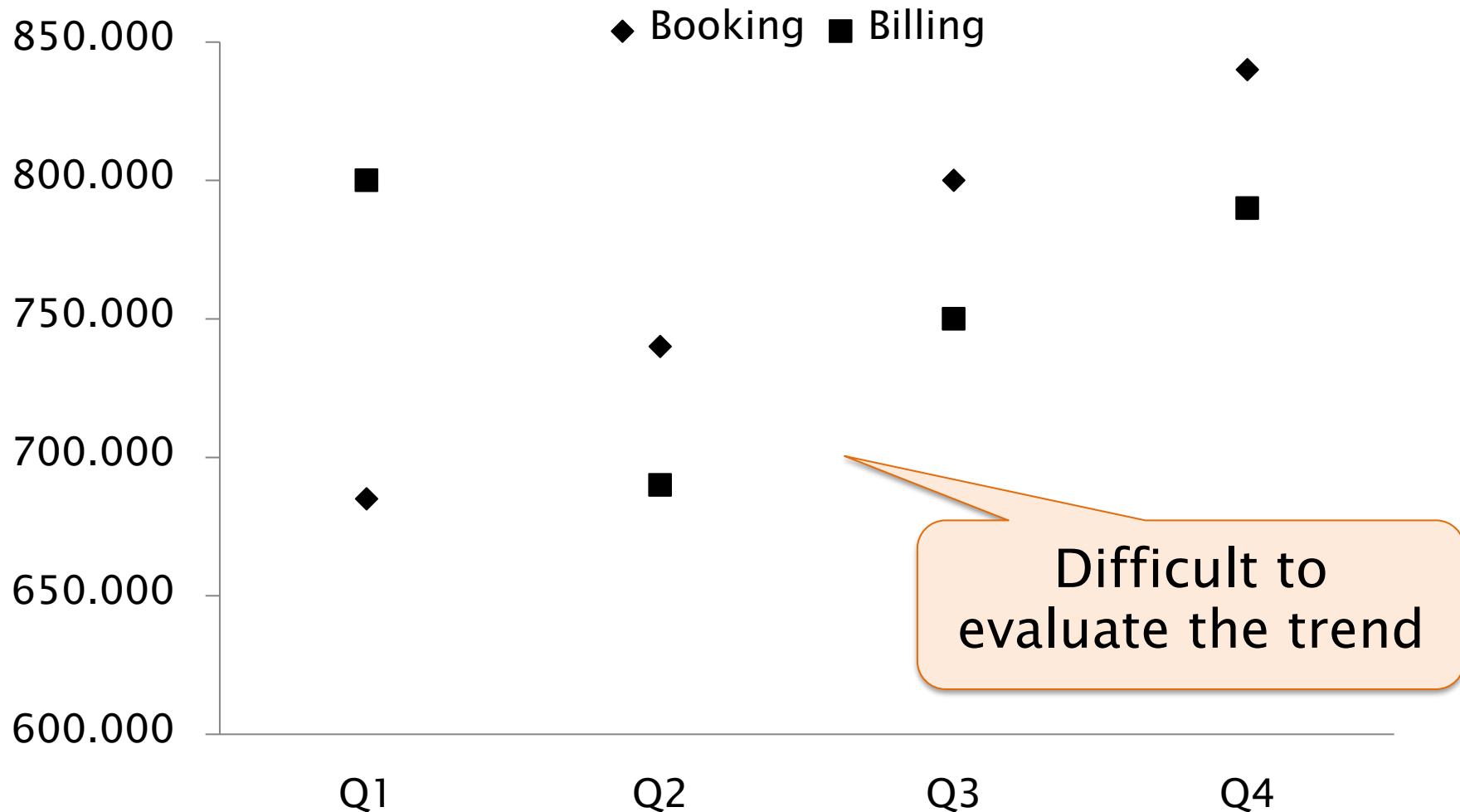
Position + Color (hue)



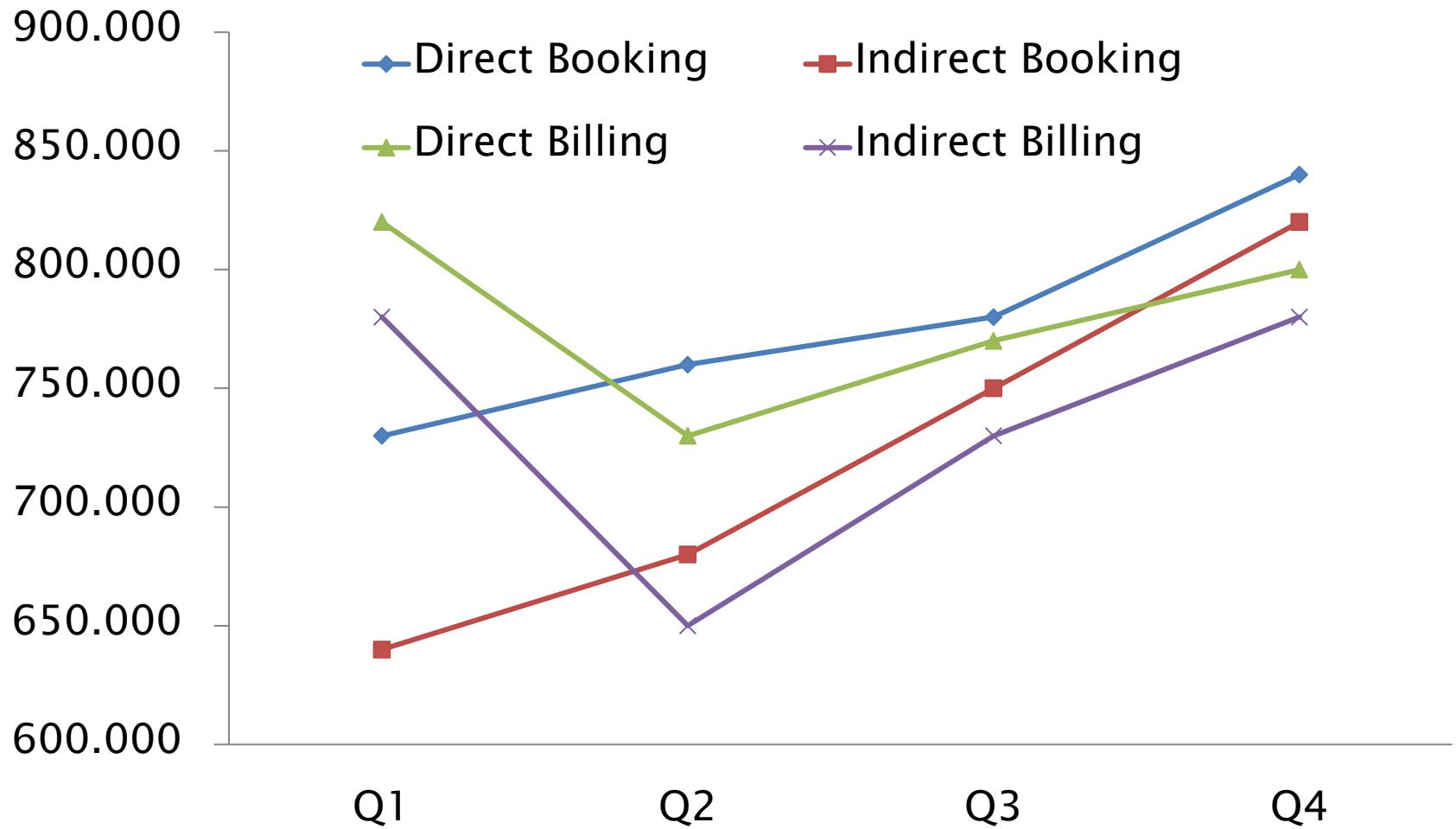
Size



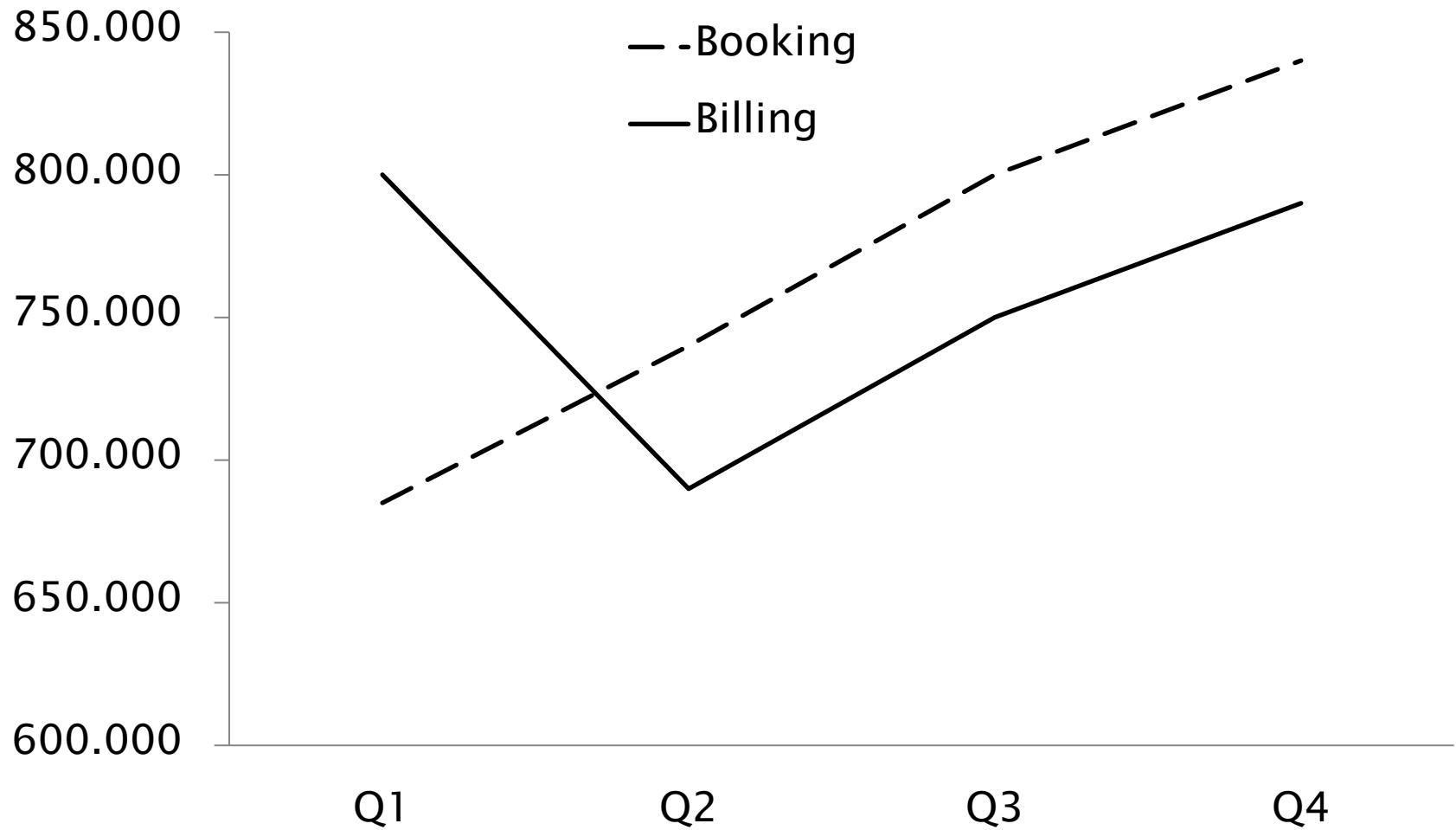
Point shape



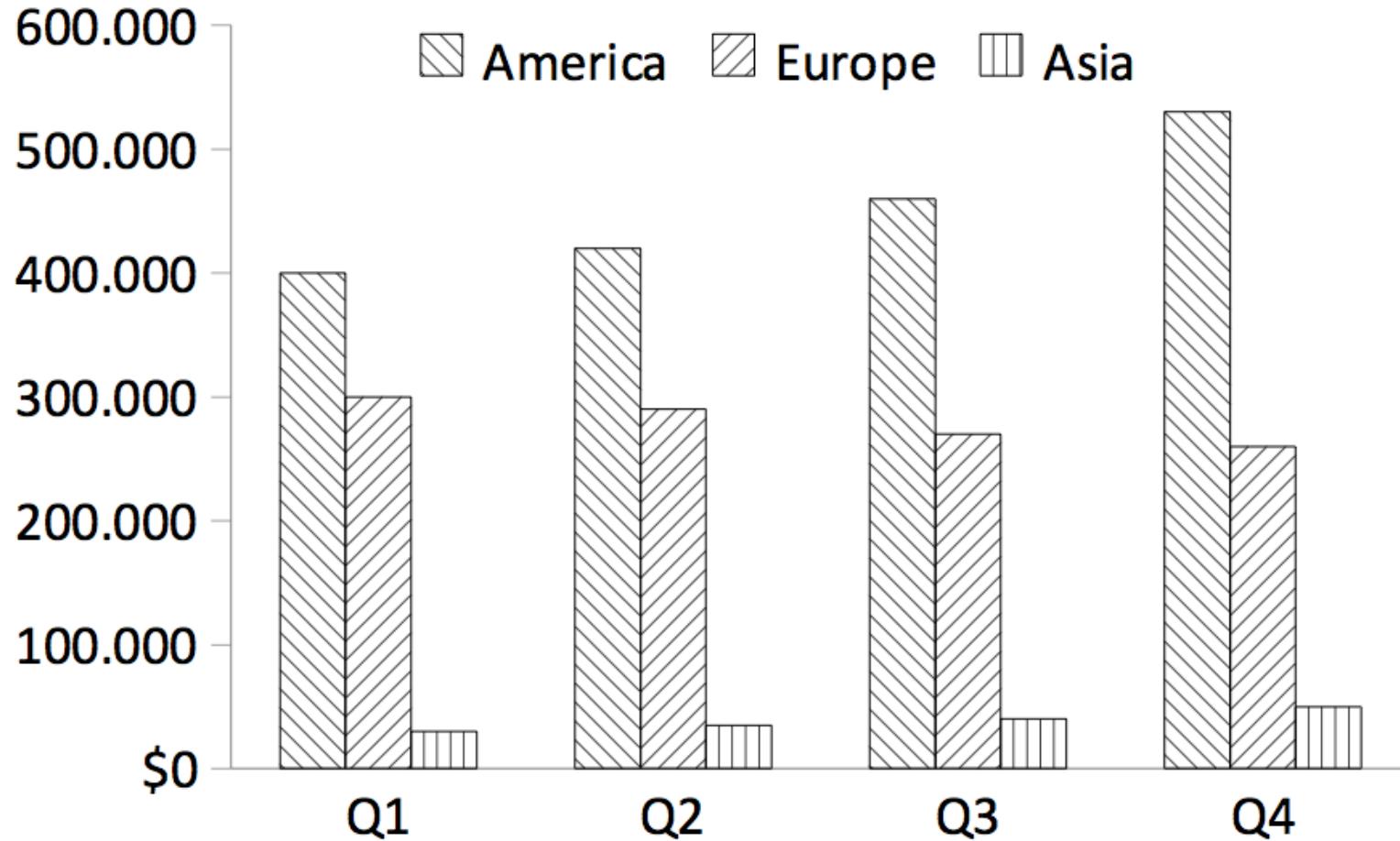
Point shape + Color



Line style



Fill Texture

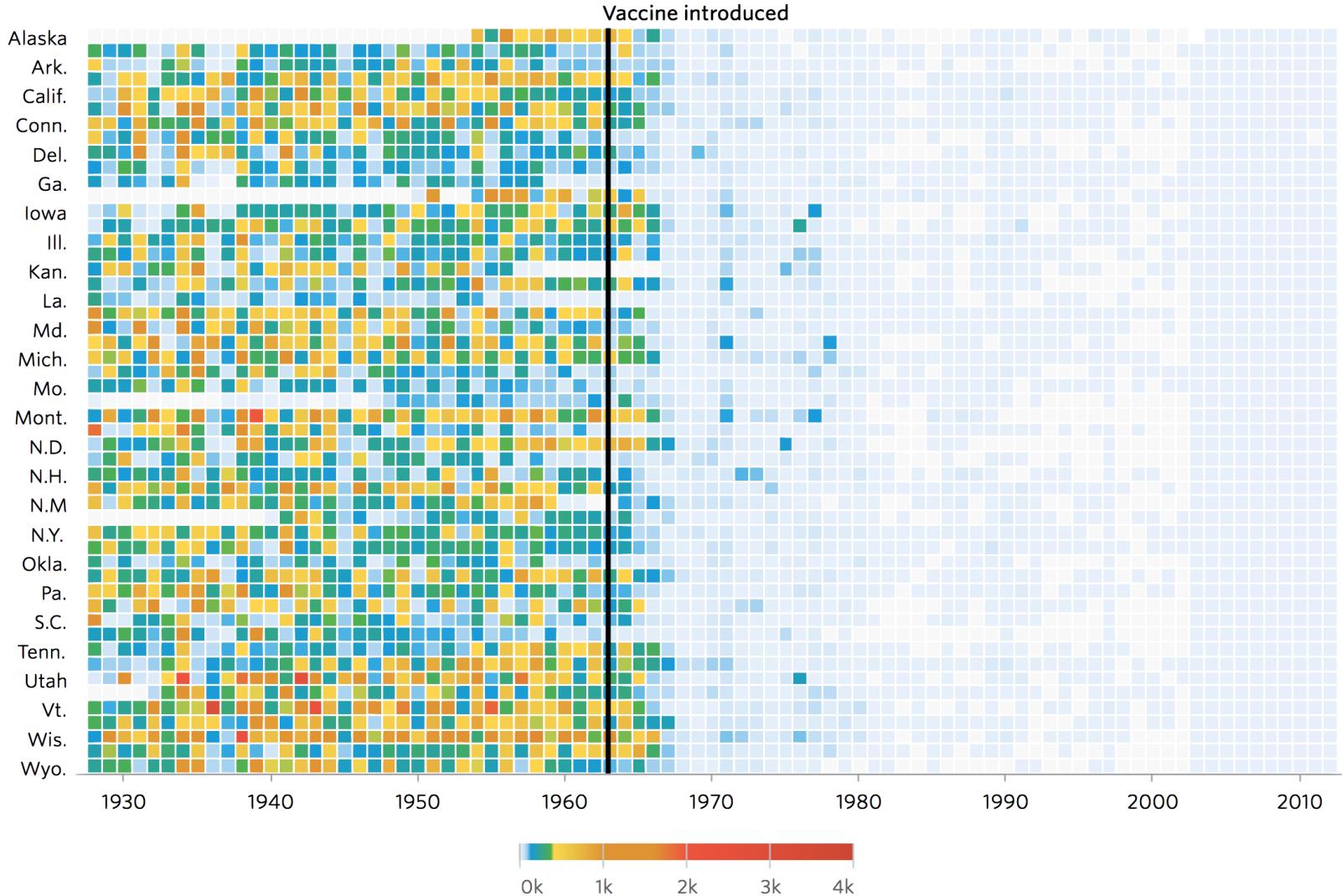


Discretization / Quantization

- A data transformation that maps a quantitative measure into an ordinal one
 - ◆ Based on the definition of intervals
- Discretized measures can be encoded using an ordinal-friendly visual attribute
 - ◆ Size
 - ◆ Color
- Warning: details are lost in the process

Heatmaps

Measles

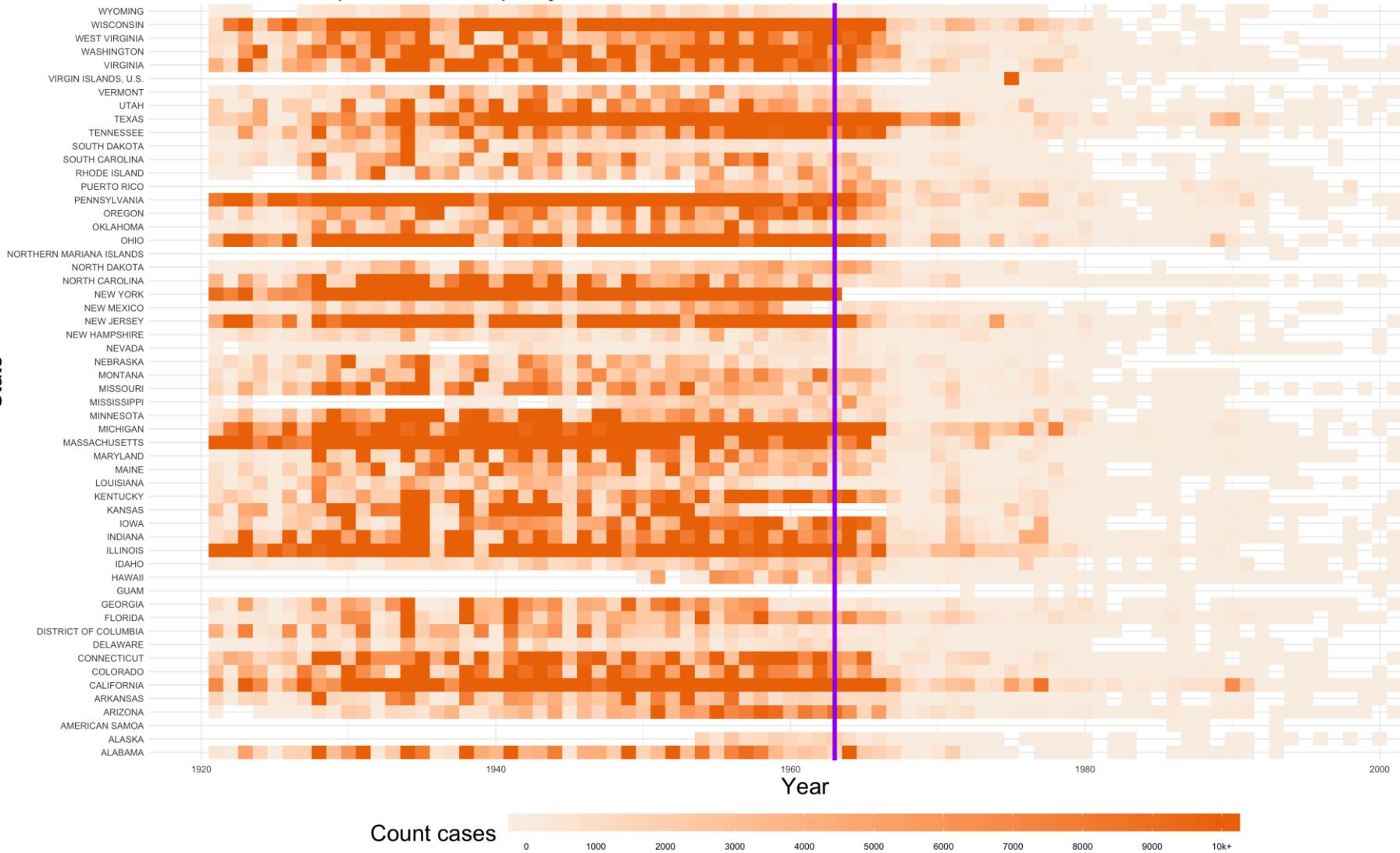


Heatmaps

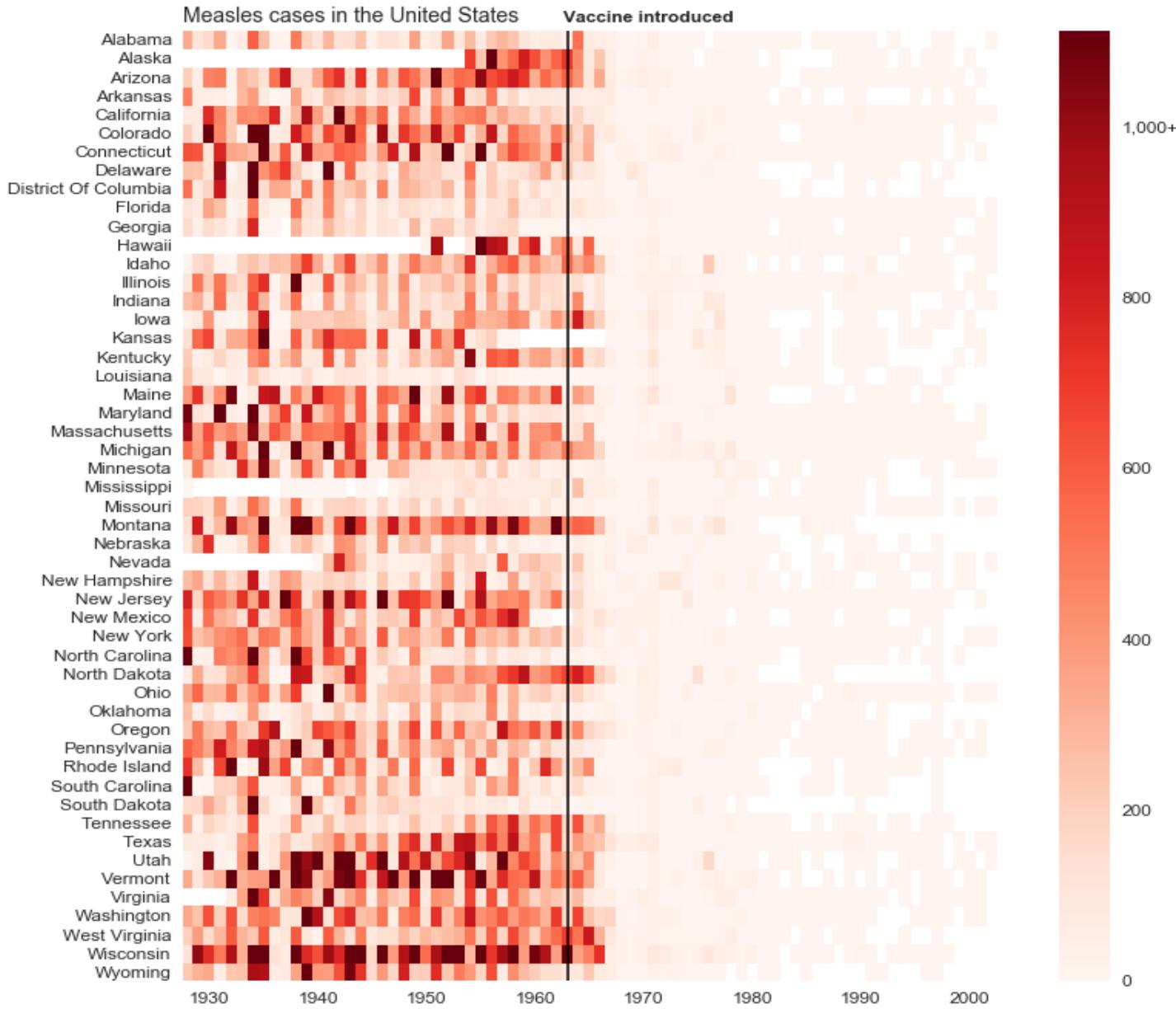
- Hues have no unique order semantics
 - ◆ Only intensity has one
- Rainbow palette have serious problems for color blinds
 - ◆ Roughly 5% of the population

Heatmaps

Measles cases per US State per year



Heatmap



Data source: Project TYCHO (tycho.pitt.edu) | Author: Randy Olson (randalolson.com / @randal_olson)

<http://www.randalolson.com/2016/03/04/revisiting-the-vaccine-visualizations/>

SUPPORT ELEMENTS

Support elements

- Axes
 - ◆ Ticks
- Graph area
 - ◆ Grids
- Labels
- Legends
- References
- Trellies

Axes

- Allow positioning of elements
 - ◆ Points
 - ◆ Extremes of bars and lines
- Labeled
 - ◆ What is the measure?
- Number of axis should be 2
 - ◆ 1 is fine for bars
 - continuity gestalt principle

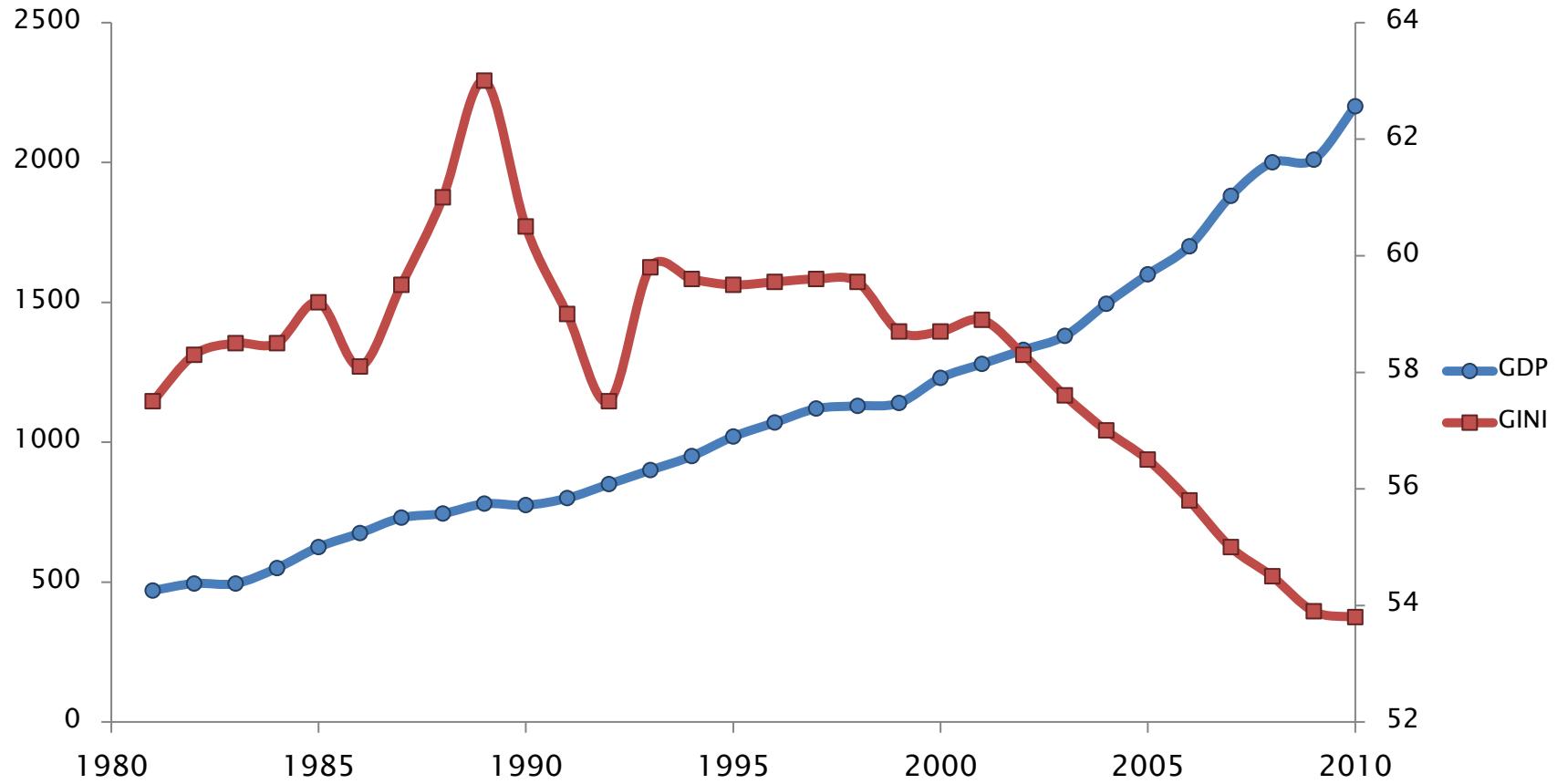
Tick marks

- Must not obscure data objects
- Outside the data region
- Avoid for categorical scales
- Balanced number
 - ◆ Too many clutter the graph
 - ◆ Too few make difficult to discern reference for data objects
 - ◆ Intervals must be equally spaced

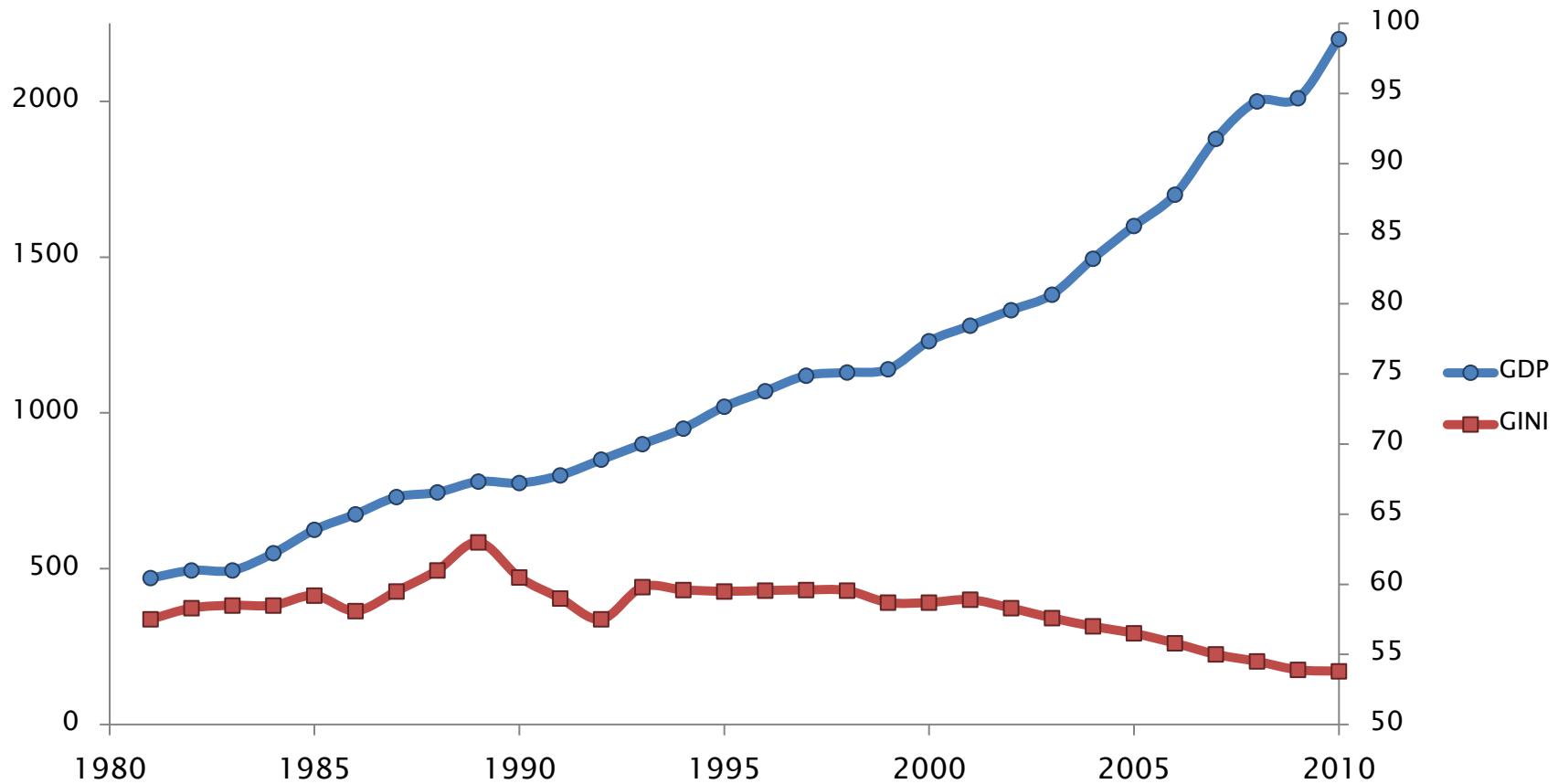
Multiple variables

- Correlation between 3+ variables
 - ◆ E.g. two measures in time series
- Multiple units of measure
 - ◆ Double quantitative (y) axis
 - ◆ Multiple graphs
 - ◆ One variable not encoded explicitly

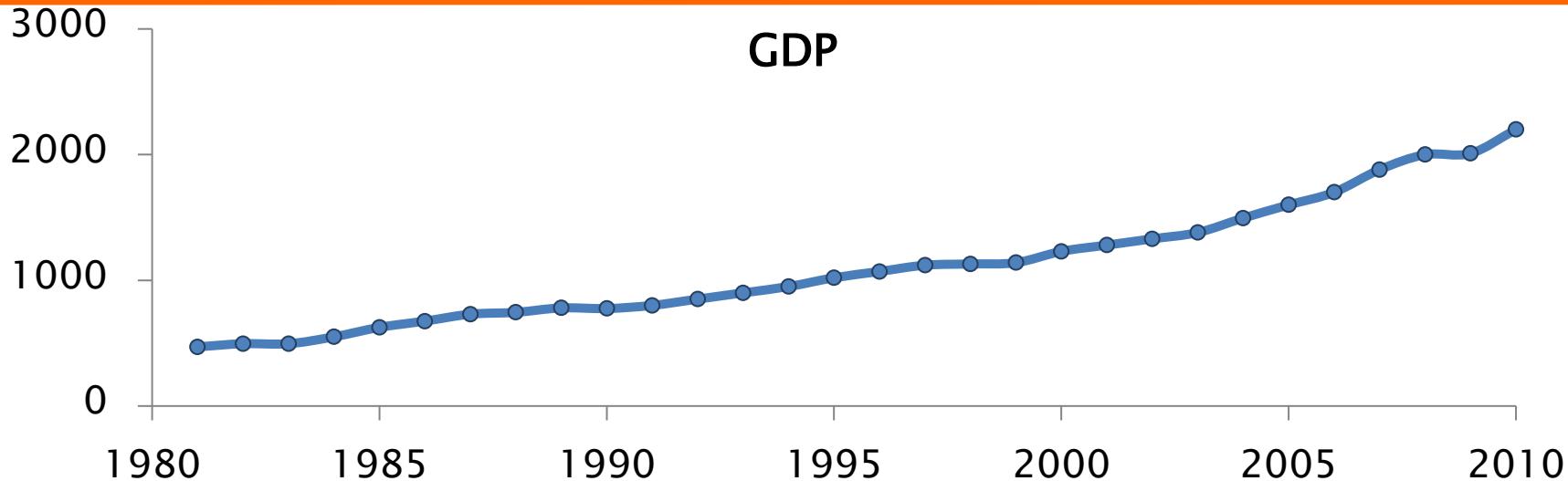
Double scale



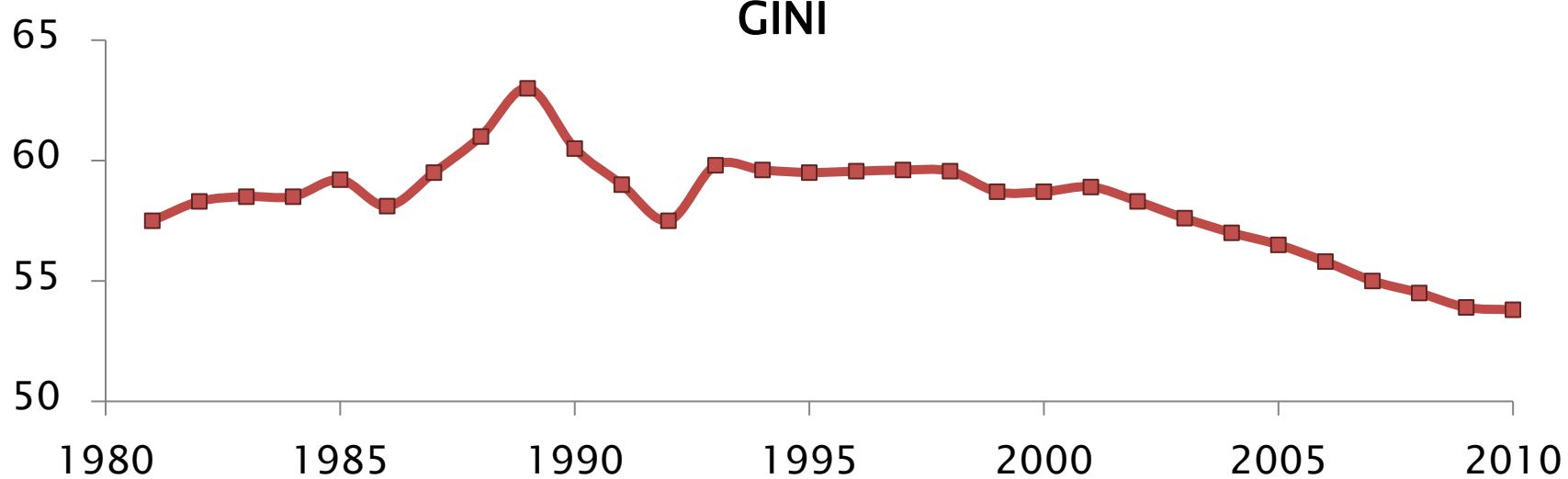
Double scale



Multiple graphs



GINI



Small multiples

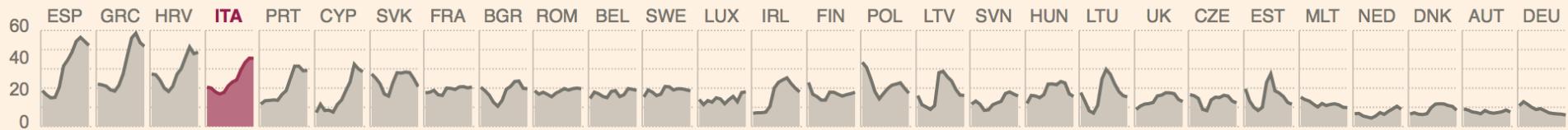
- A.k.a.
 - ◆ Trellis
 - ◆ Lattice
 - ◆ Grid
- Set of aligned graphs sharing (at least one) scale and axis
 - ◆ Enable ease of comparison among different measures

Small multiples

Total unemployment rate, 2004-2015 (%)



Youth unemployment rate, 2004-2015 (%)



Long-term unemployment rate, 2004-2014 (%)



FT EU unemployment tracker

<http://blogs.ft.com/ftdata/2015/04/17/eu-unemployment-tracker/>

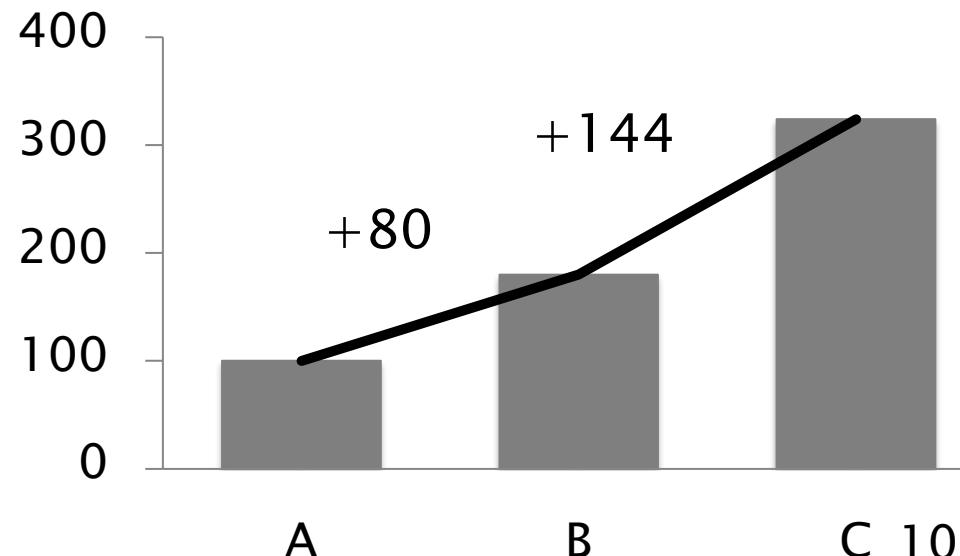
Trellis

- Sequence
 - ◆ Intrinsic order
 - ◆ Order of relevance
 - ◆ Order by some quantitative attribute
- Rules and grids
 - ◆ Use when spacing is not enough
 - ◆ Can direct the reader to scan graphs horizontally or vertically

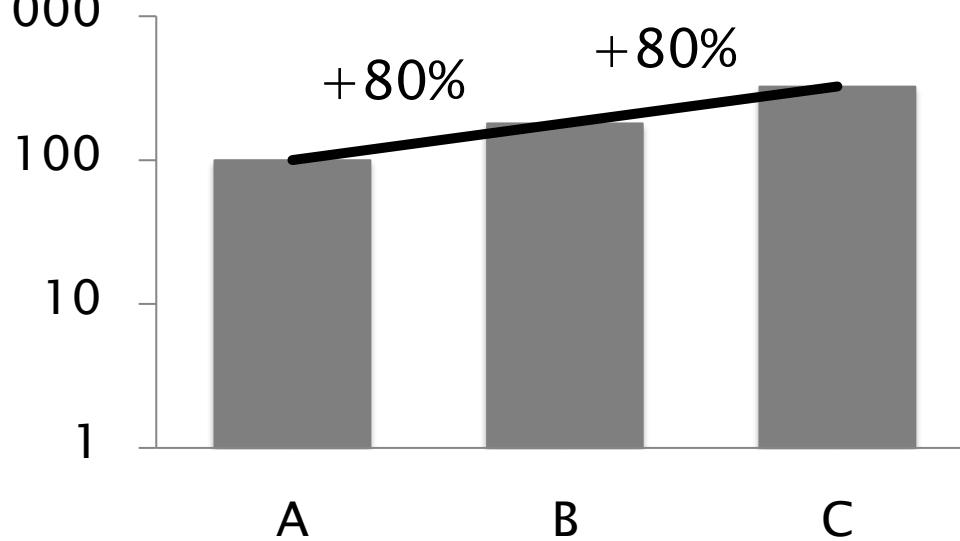
Log scale

- Reduce visual difference between quantitative data sets with significantly wide ranges
- Differences are proportional to percentages

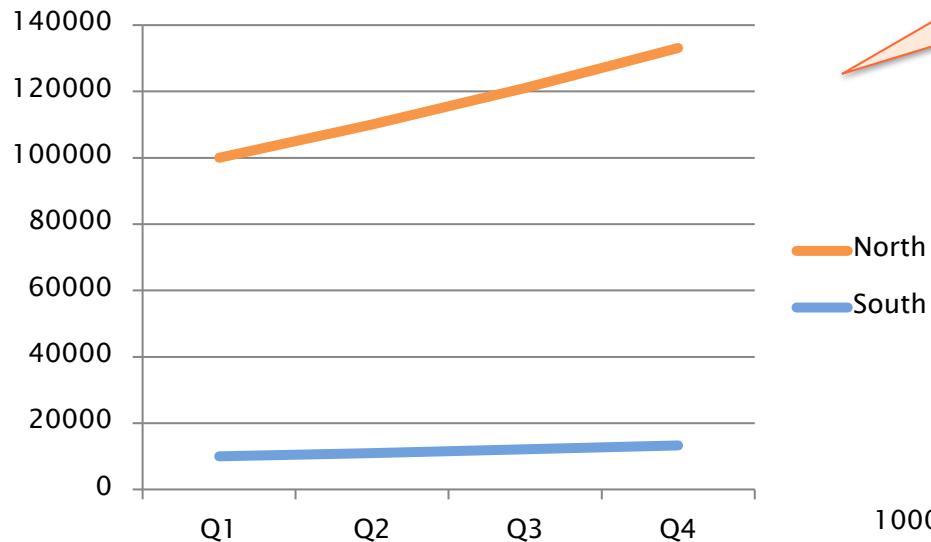
Log scale



C 1000

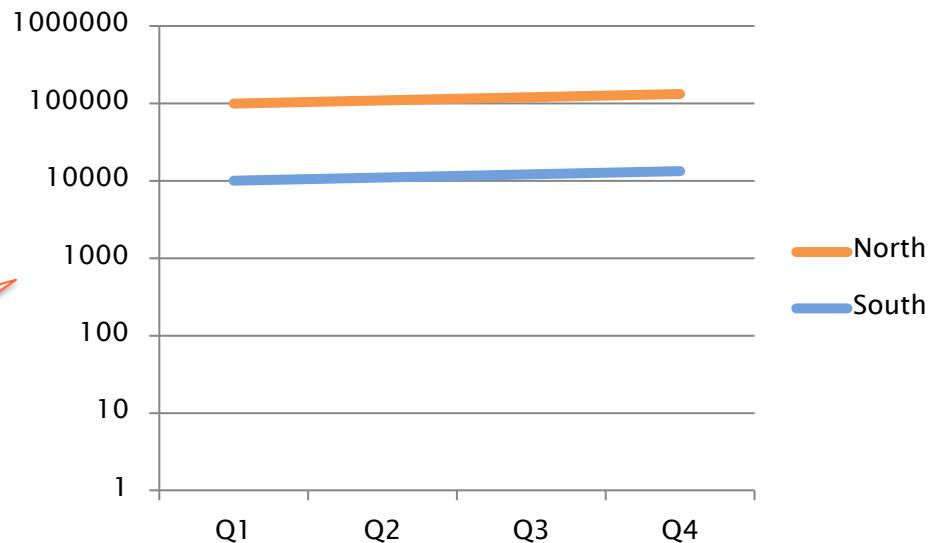


Log scale



Absolute Gains

Percentage Gains



Graph area

- Aspect ratio should not distort perception
 - ◆ Typically wider than taller
 - ◆ Scatter plots may be squared
- Grid lines must be thin and light
 - ◆ Useful to look-up values
 - ◆ Enhance comparison of values
 - ◆ Enhance perception of localized patterns

Labels

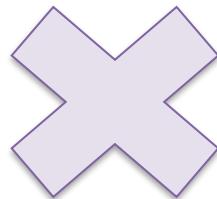
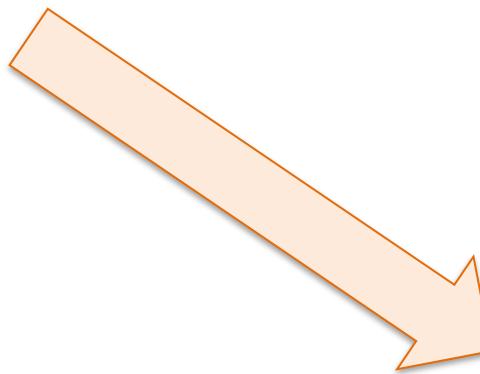
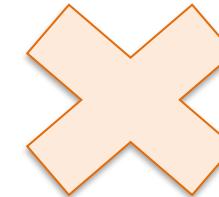
- Important elements (e.g. titles) should be prominent
 - ◆ Top
 - ◆ Larger

Gutenberg Diagram

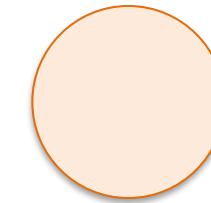
Primary area



Strong fallow area

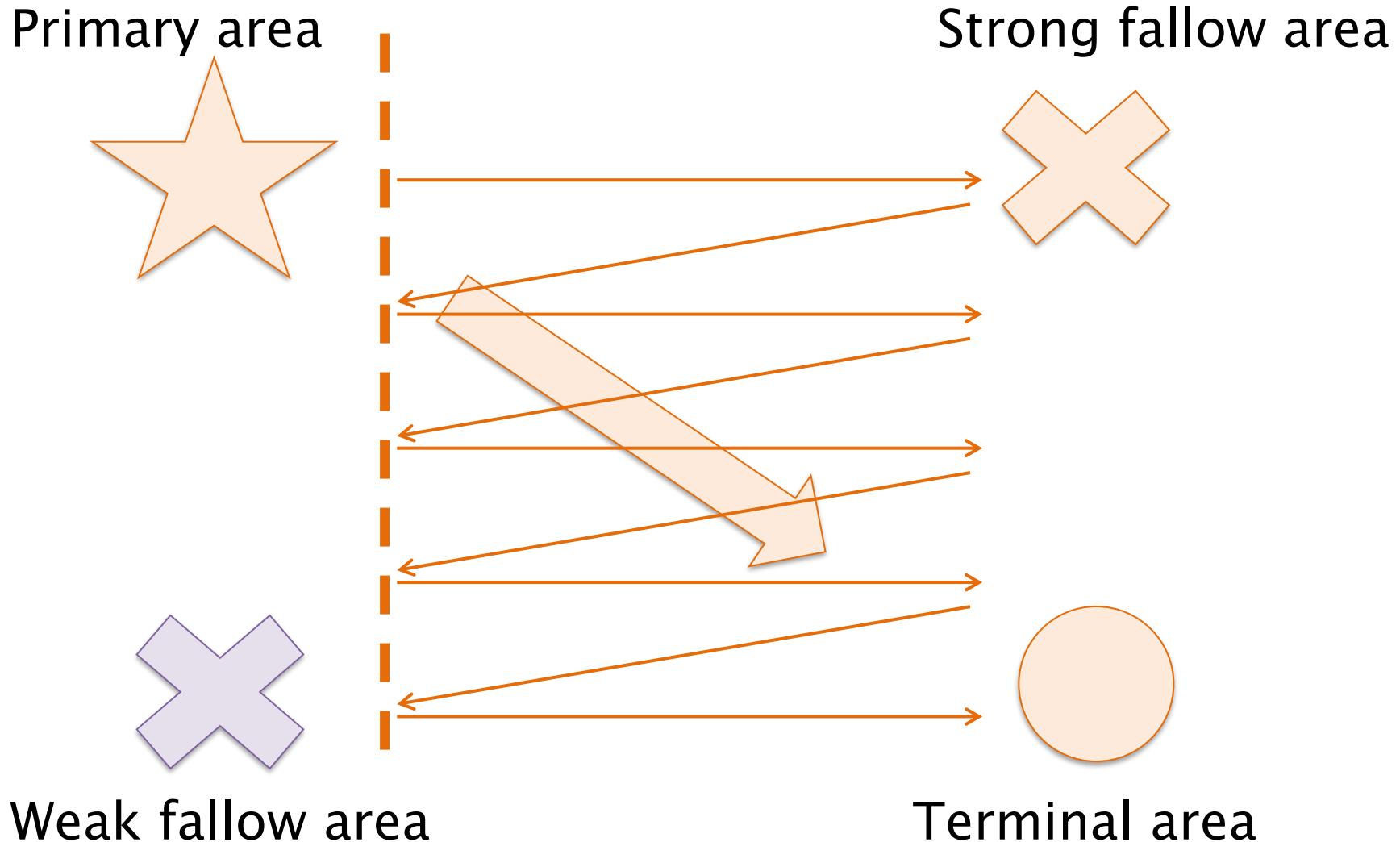


Weak fallow area



Terminal area

Gutenberg Diagram



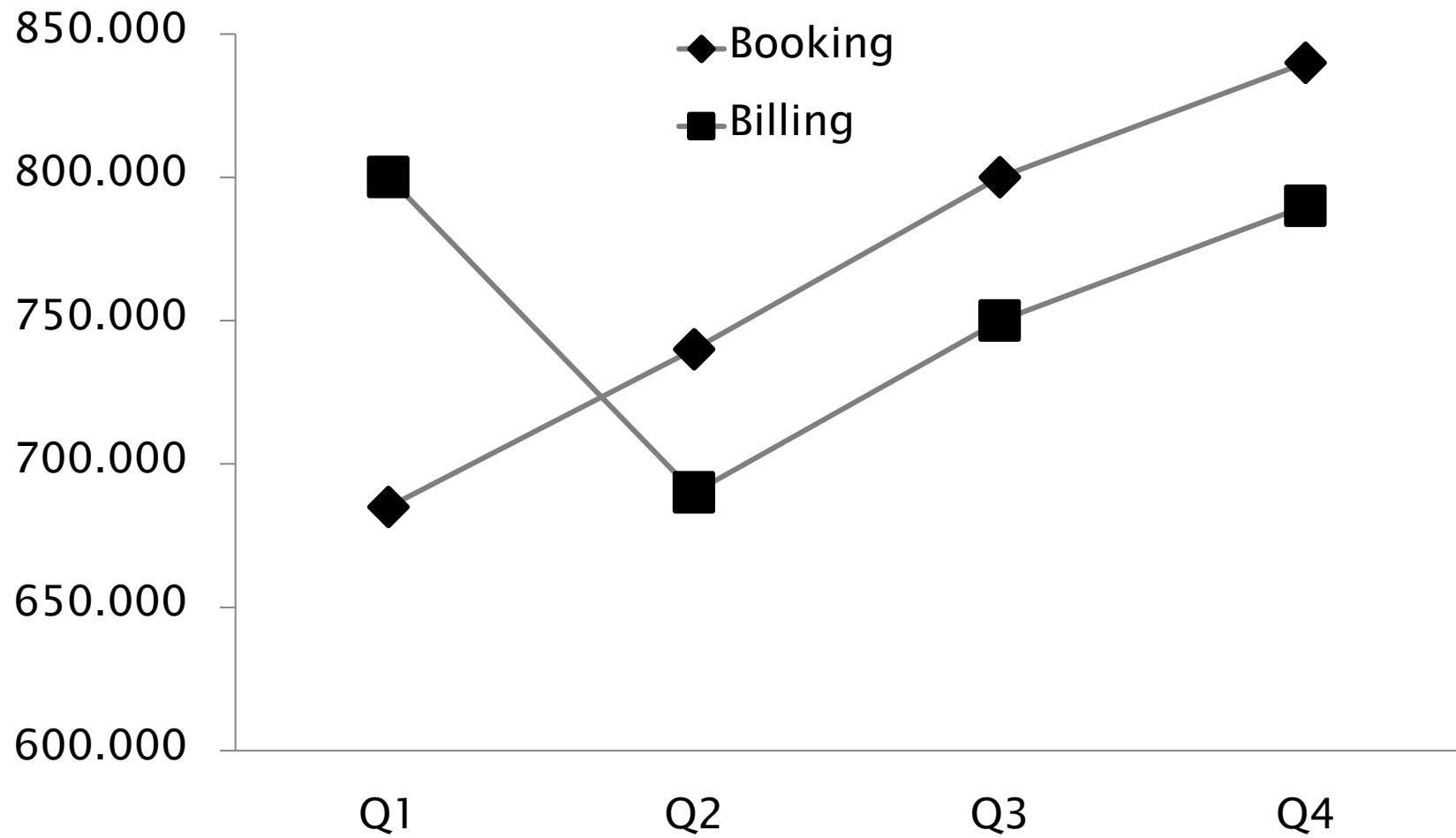
Legends

- Used for categorical attributes not associated to any axis
- As close as possible to the objects
- Less prominent than data objects
- Borders are used only when necessary to separate from other elements

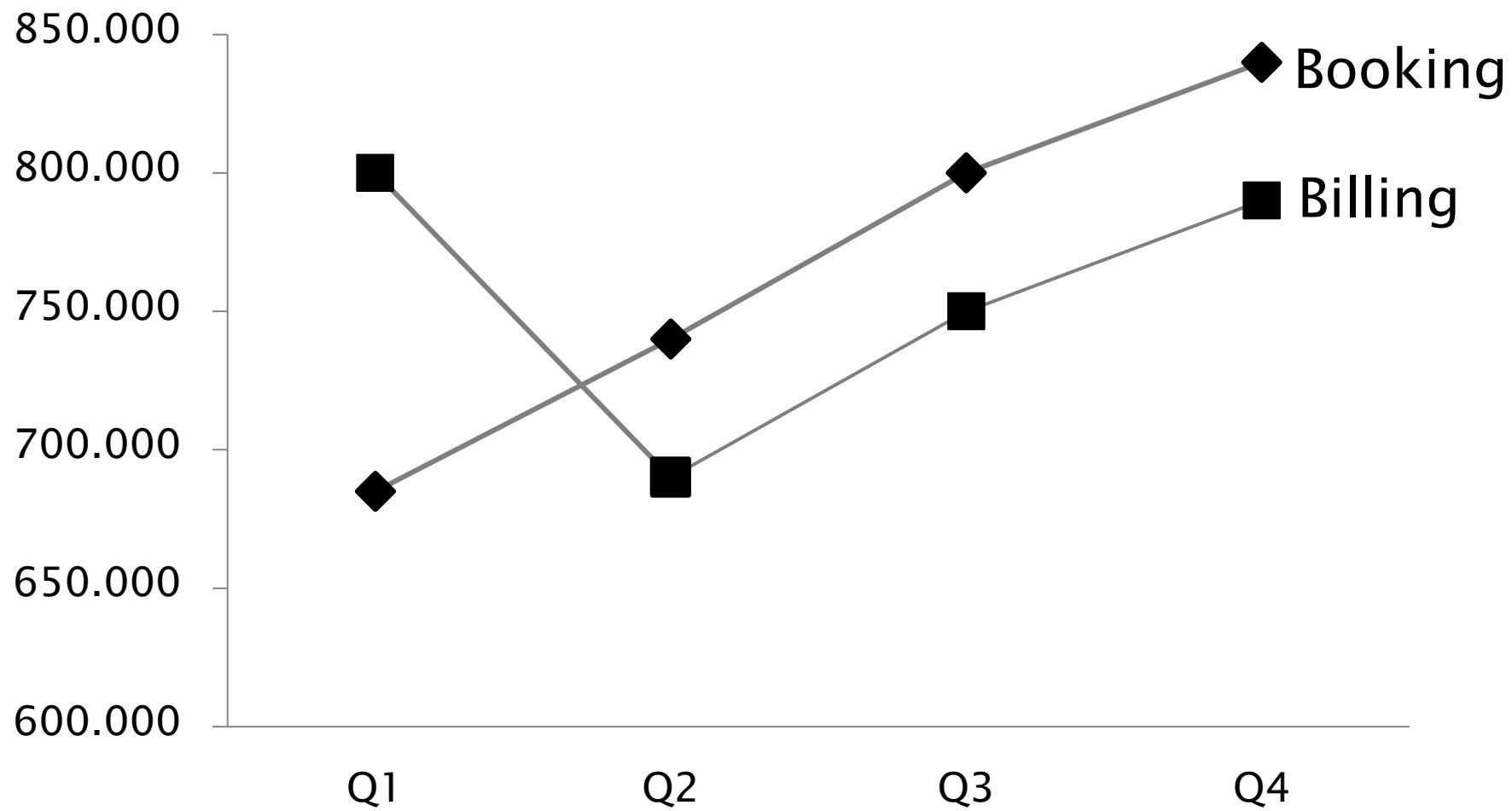
Legends

- Text should be as close as possible to the object it complements
 - ◆ Prefer direct labeling to separate legends
- Number of categorical subdivisions
 - ◆ Perceptual limit is between 5 and 8
 - ◆ Limit is independent of the visual attribute used to encode it
 - ◆ Joint use of attributes ease discrimination

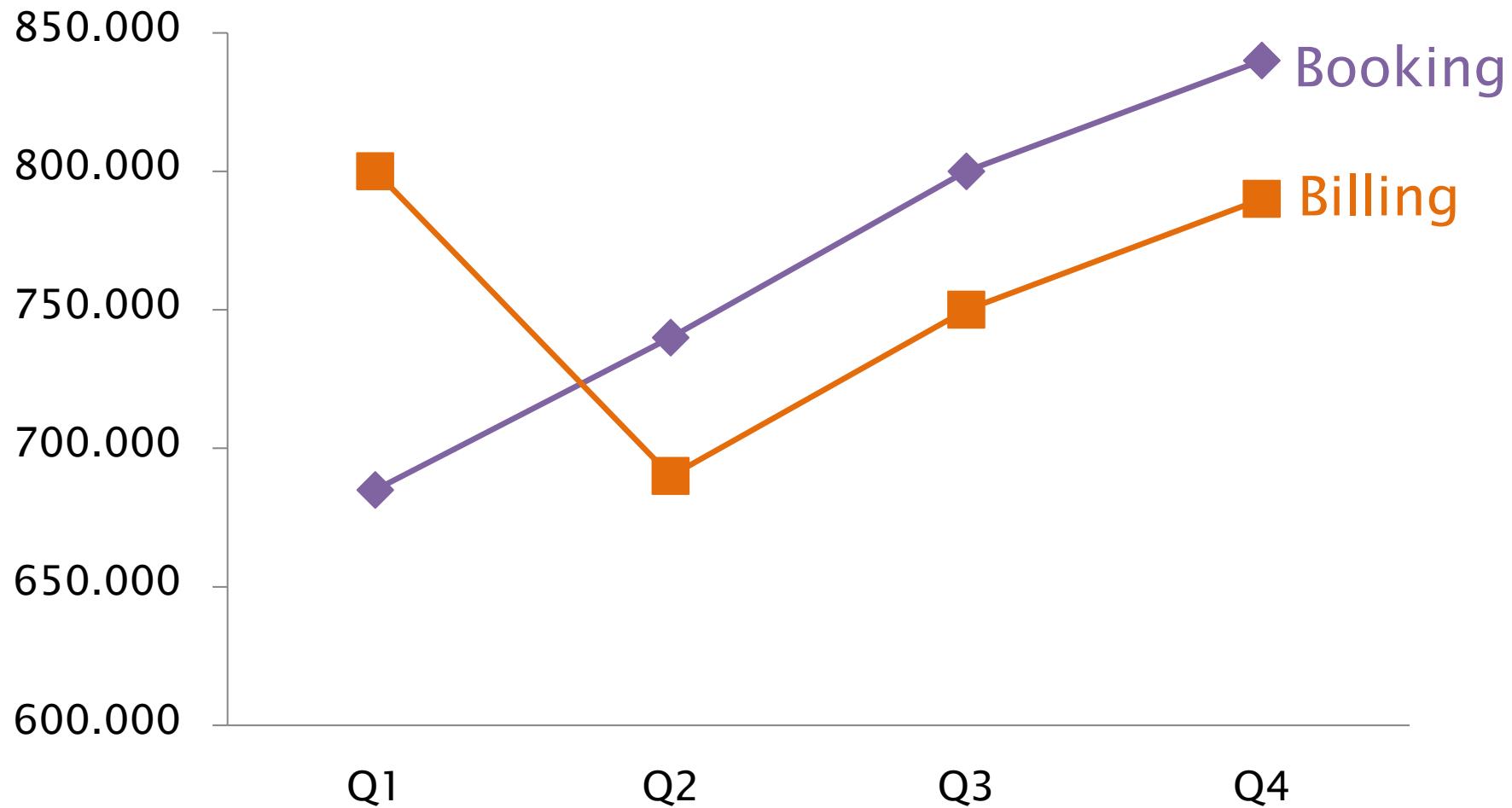
Legend



Direct labeling

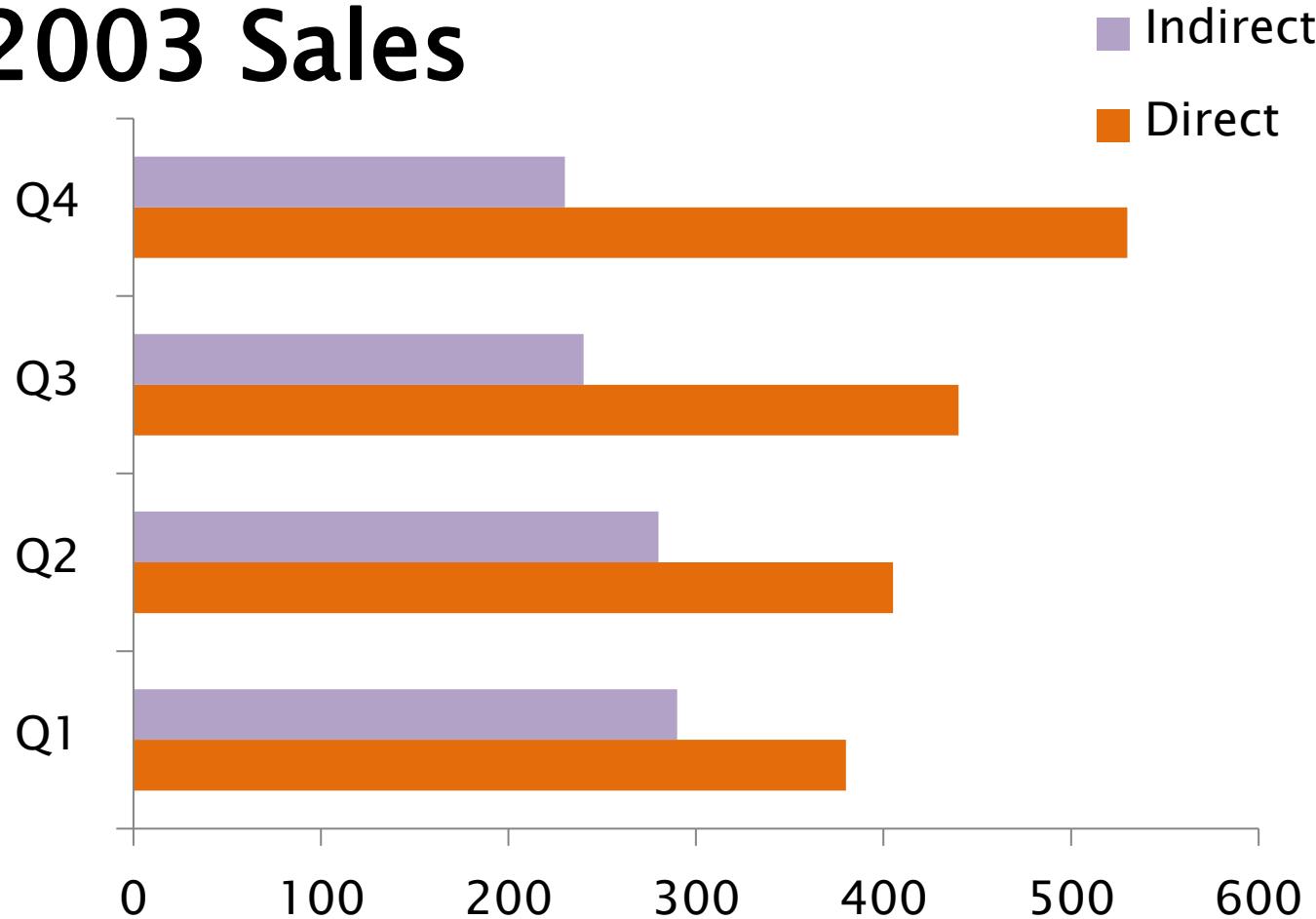


Direct labeling and color



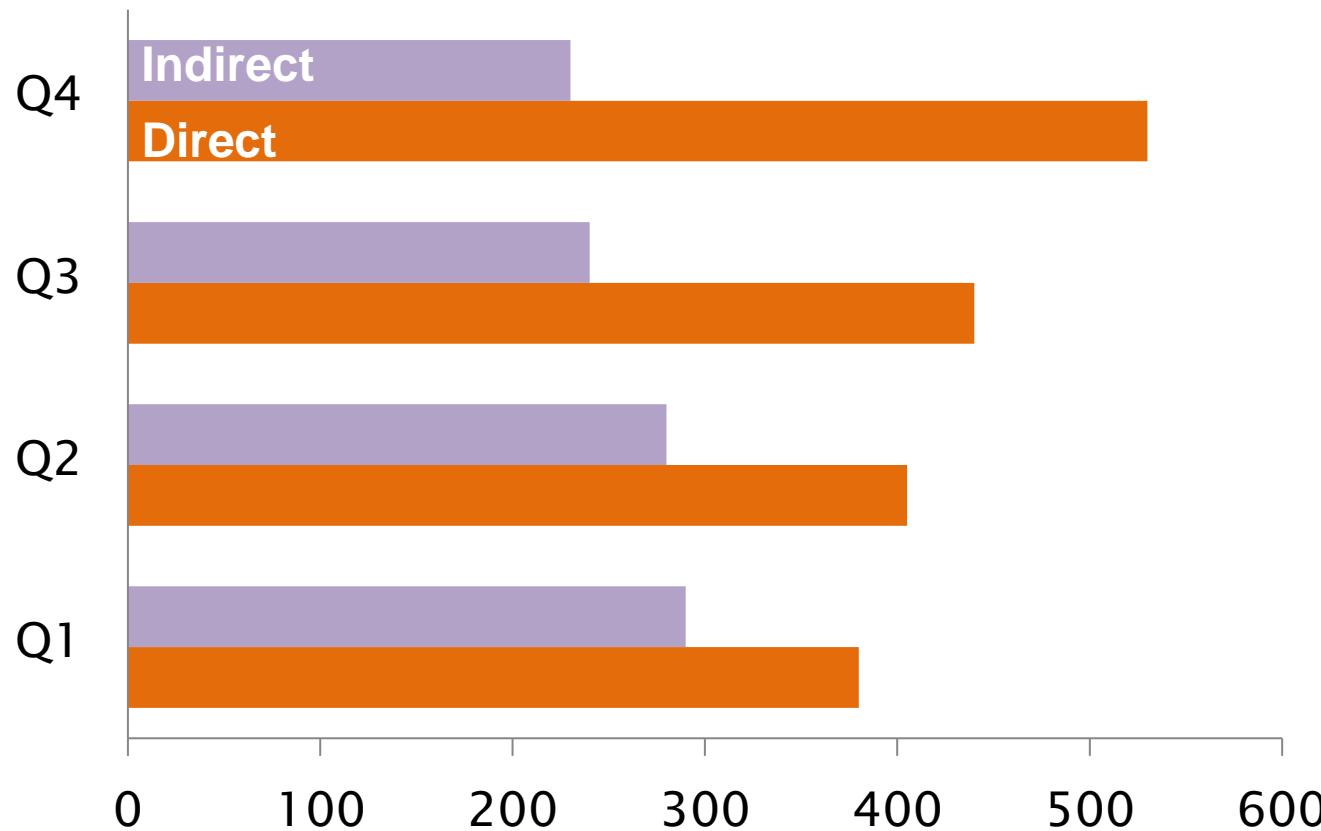
Legend

2003 Sales



Direct labeling

2003 Sales



Reference lines and regions

- Reference lines support an easy comparison to a given value
 - ◆ Mean
 - ◆ Threshold
- Reference regions allow comparison with several values
 - ◆ Use background color

VISUAL RELATIONSHIPS

Data Visualization

Understanding

Information Visualization

Visual Patterns, Trends, Exceptions

Quantitative Reasoning

Quantitative Relationship & Comparison

Visual Perception

Visual Properties & Objects

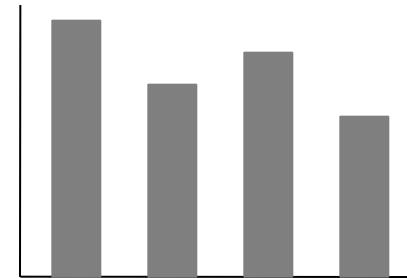
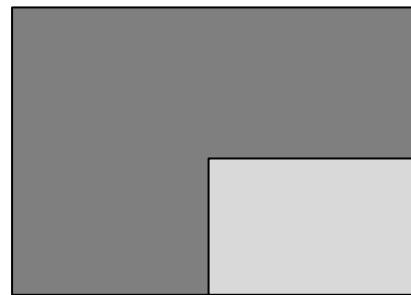
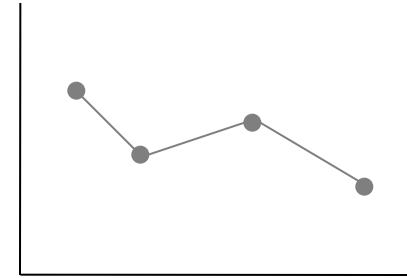
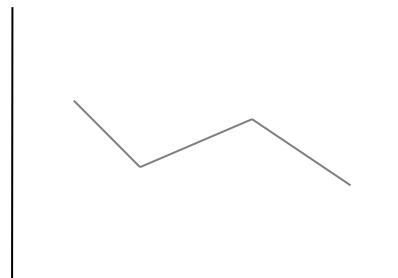
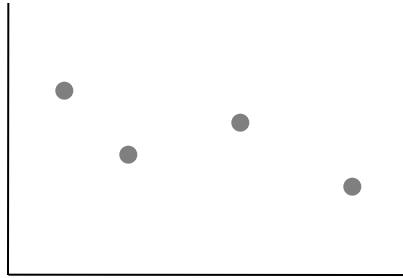
Data

Representation/Encoding

Relationships

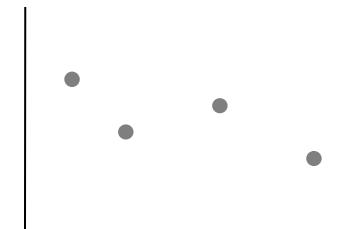
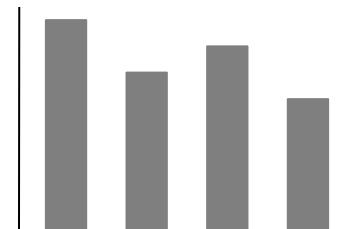
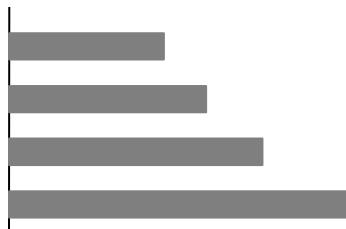
- Within a category
 - ◆ Nominal comparison
 - ◆ Ranking
 - ◆ Part-to-whole
 - ◆ Distribution
- Between measures
 - ◆ Time series
 - ◆ Deviation
 - ◆ Correlation

Quantitative encoding

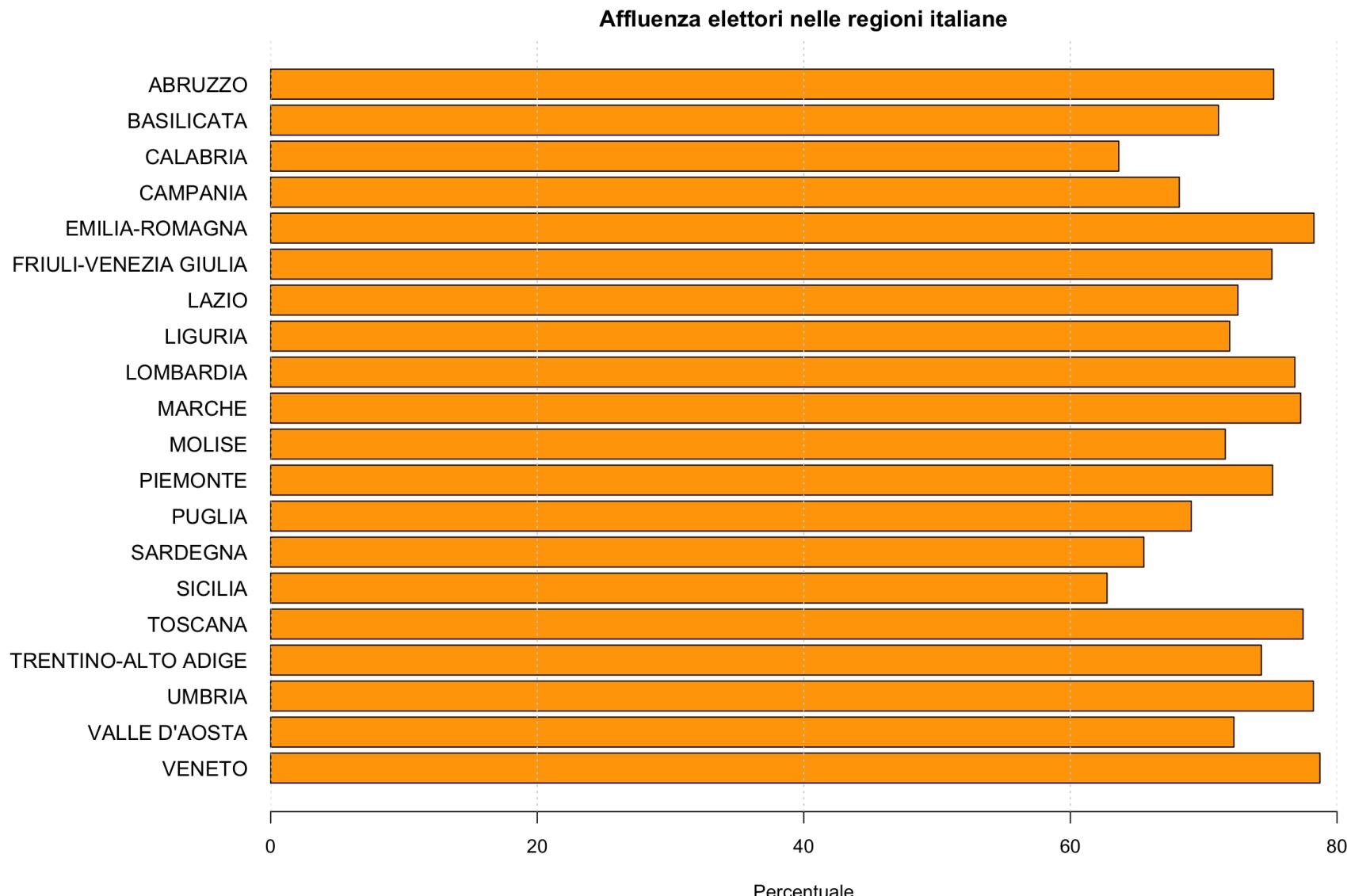


Nominal comparison

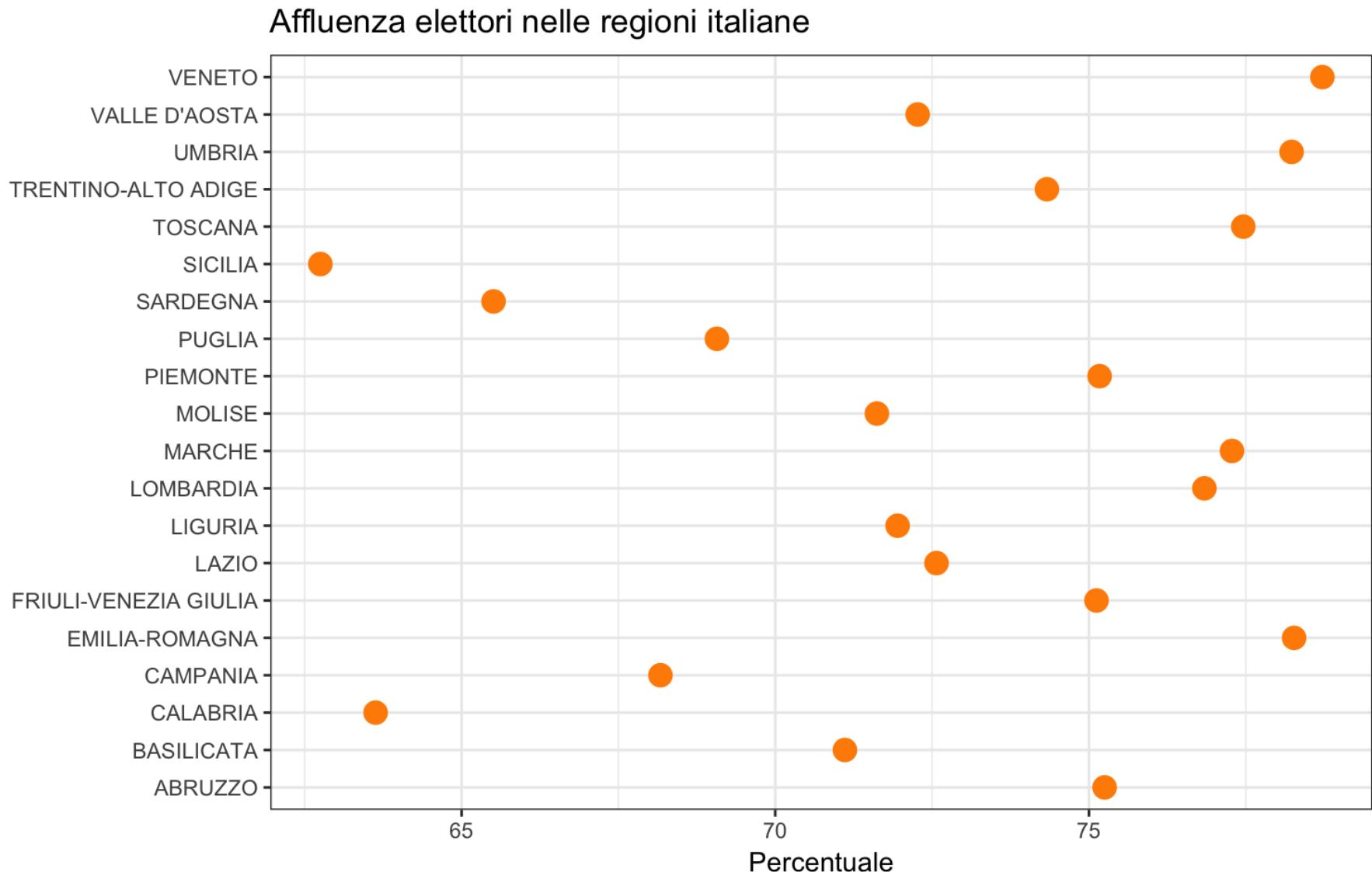
- Compare quantitative values corresponding to categorical levels
 - ◆ Small differences are difficult to see
 - Non zero-based scale can emphasize
 - ◆ Dot plots can be used for small differences
 - They do not require zero based scale



Comparison – Barplot



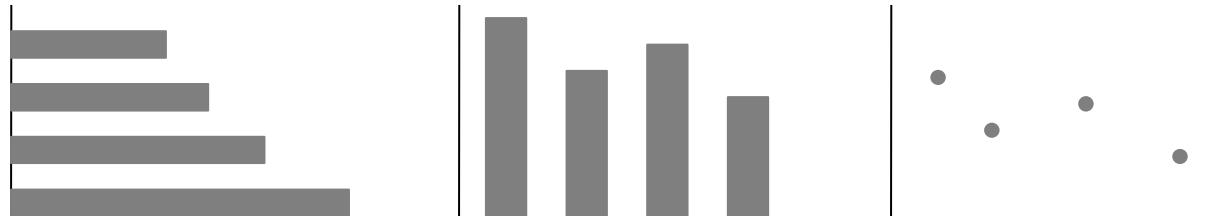
Comparison – Dot plot



Ranking

- Same type as nominal comparison
- Pay attention to order

Purpose	Sort order	Bars orientation
Highlight the highest value	Descending	H: highest on top V: highest on left
Highlight the lowest value	Ascending	H: lowest on top V: lowest on left

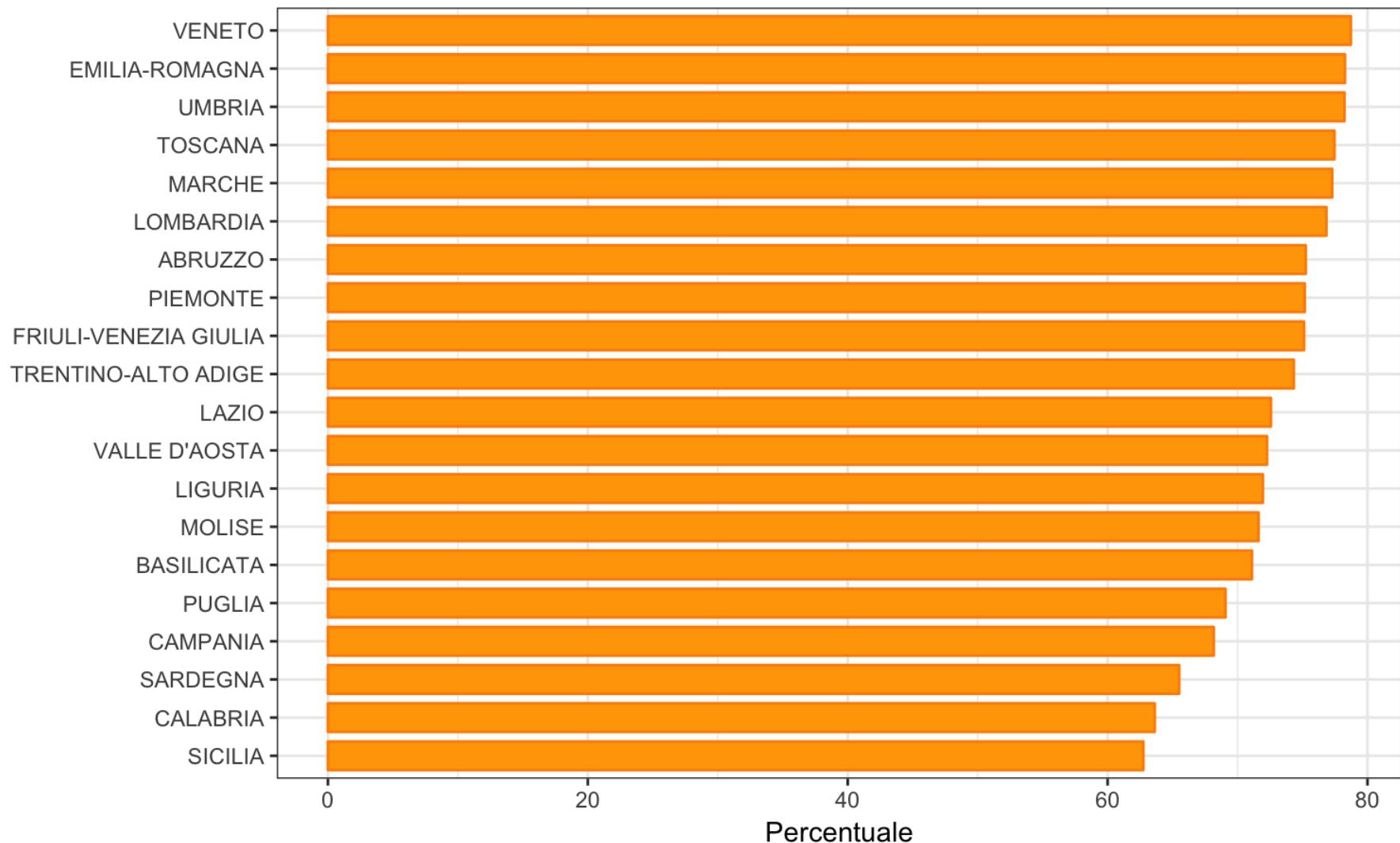


Ranking

- Bar graphs
- Dot plot
 - ◆ Allow non zero-based axes
- Line charts
 - ◆ Show evolution in time

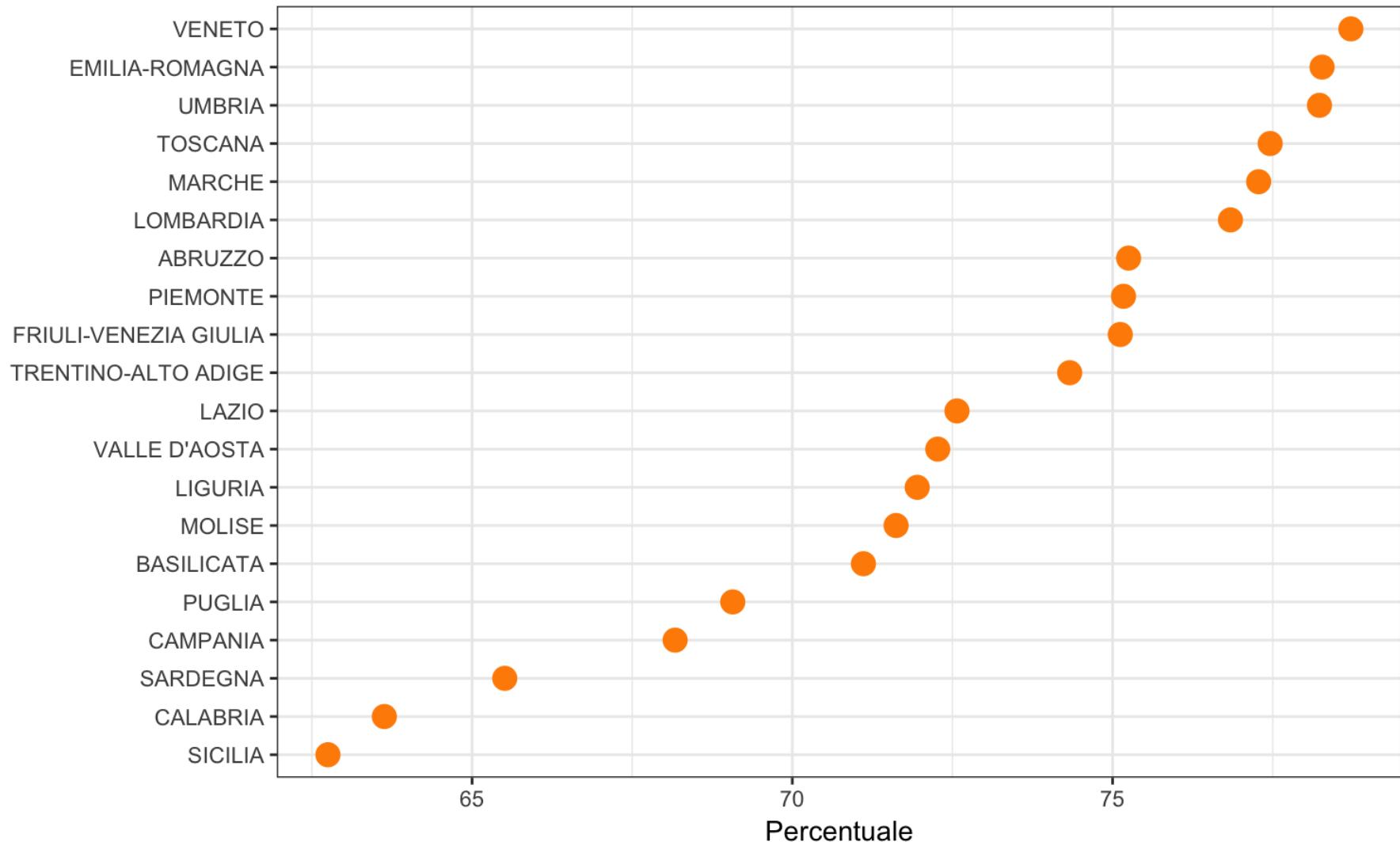
Ranking – Barplot

Affluenza elettori nelle regioni italiane



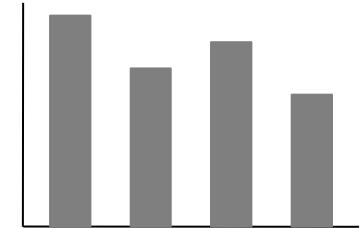
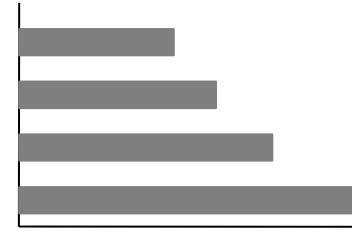
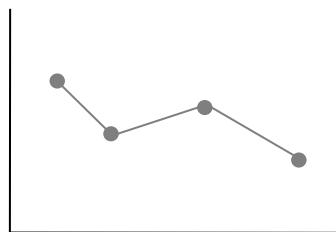
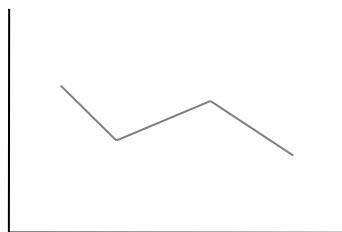
Ranking – Dot plot

Affluenza elettori nelle regioni italiane



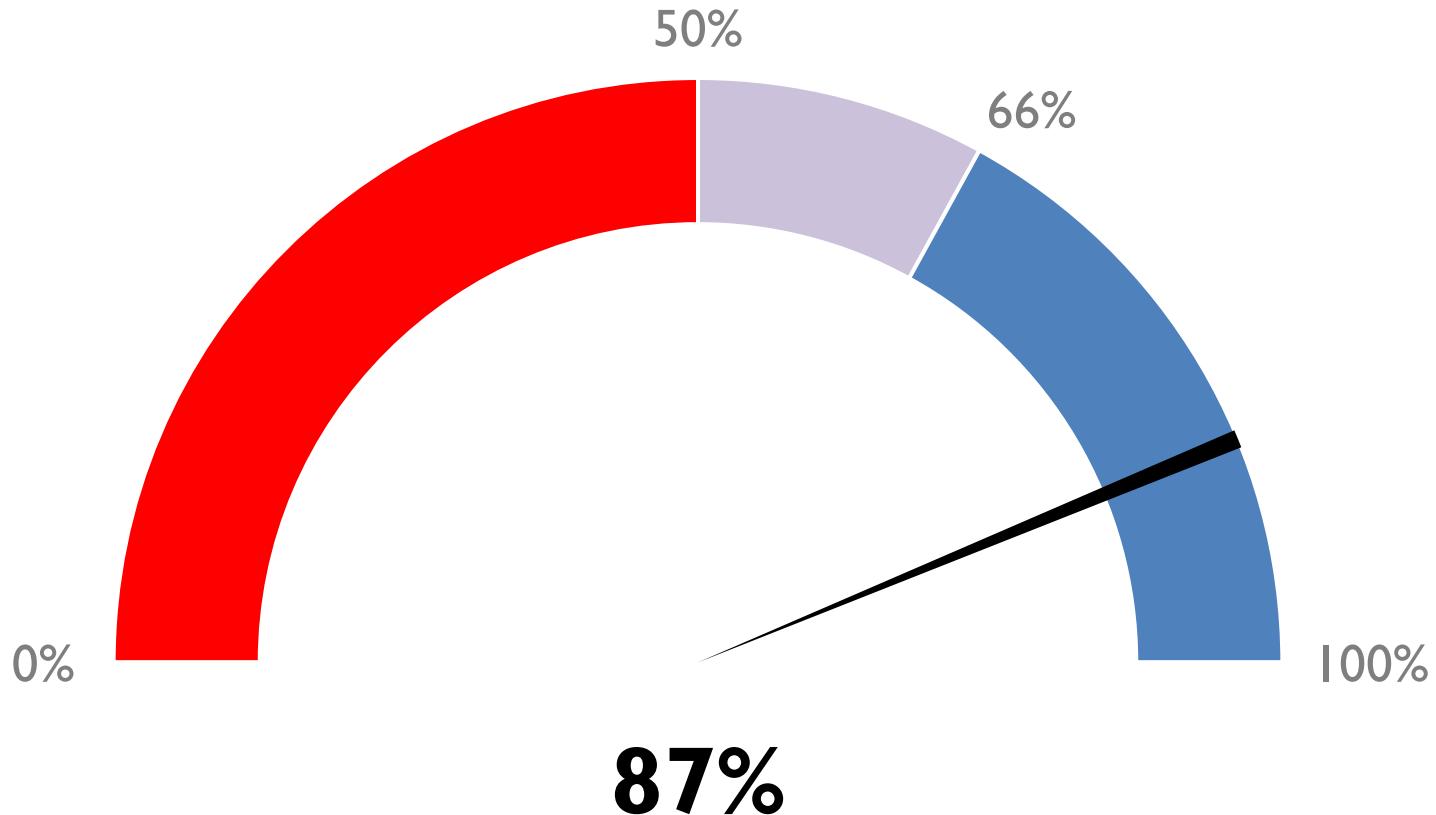
Deviation

- To what degree one or more sets of values differ in relation to primary values.
 - ◆ Points
 - ◆ Gauge
 - ◆ Bars
 - ◆ Bullet

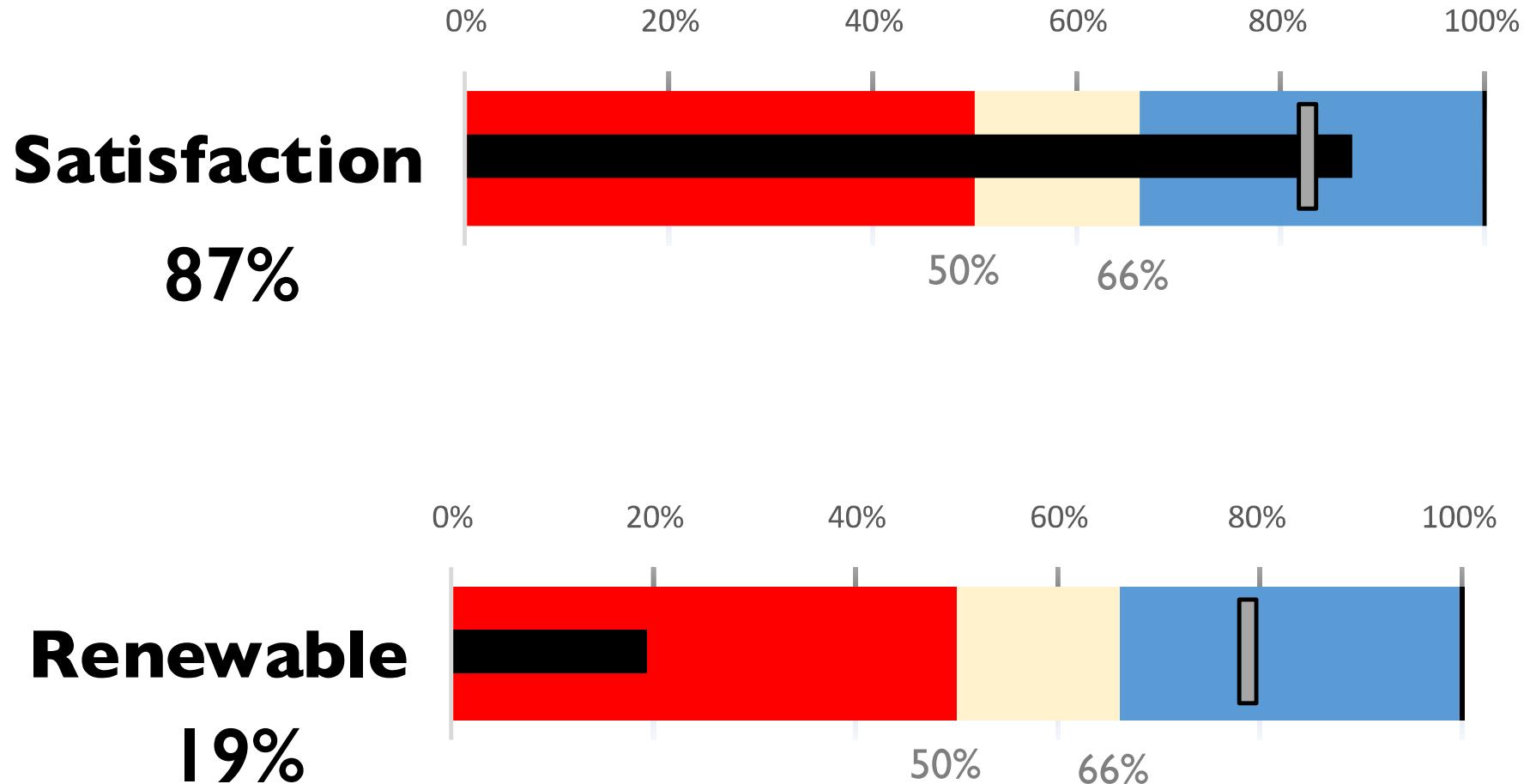


Angle + Position – Gauge

Satisfaction



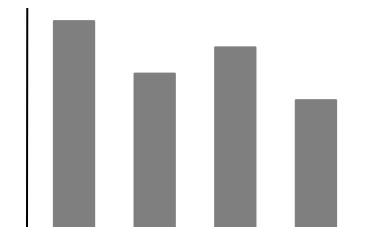
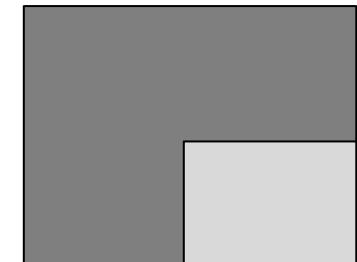
Length+Position- Bullet Graph



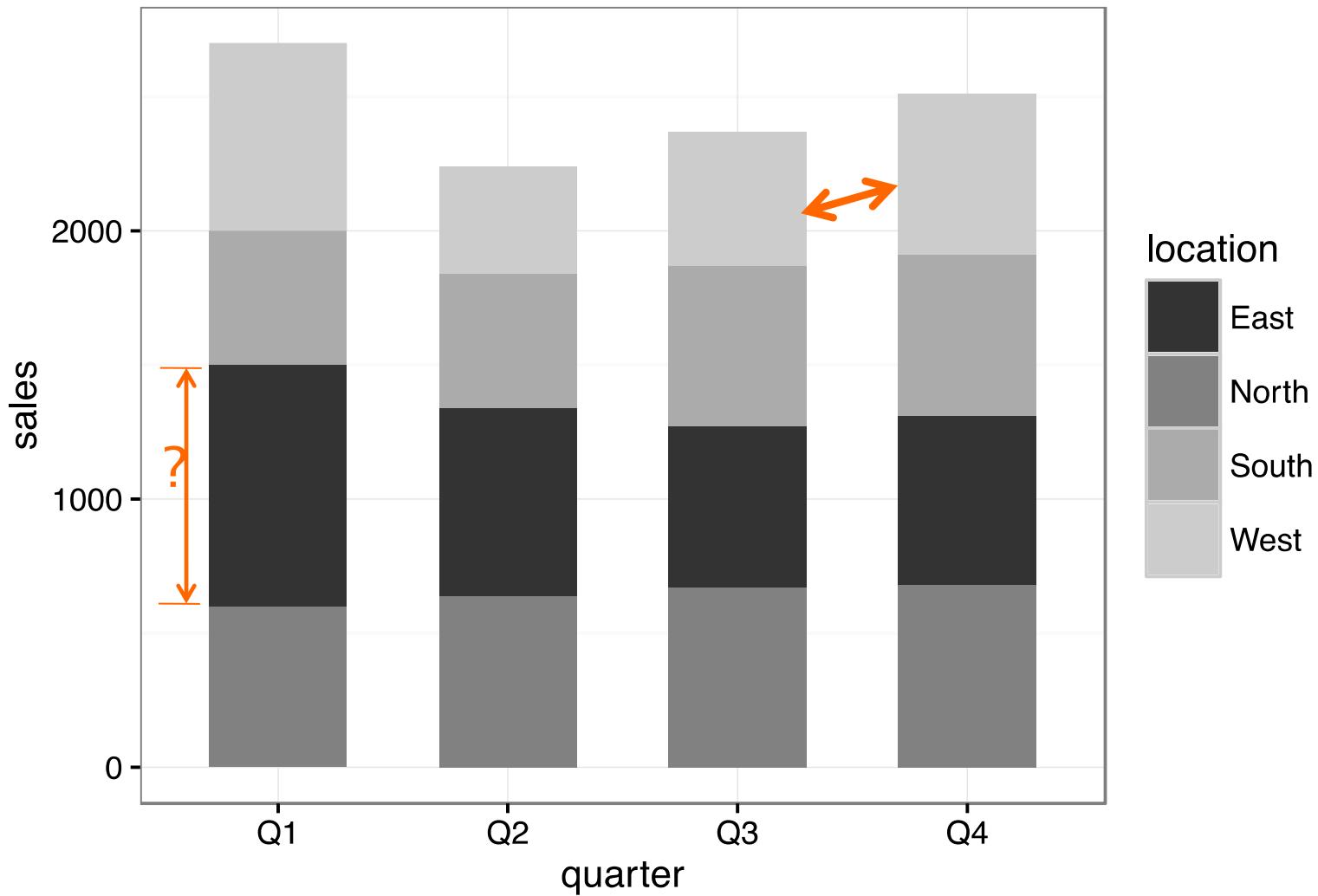
https://www.perceptualedge.com/articles/misc/Bullet_Graph_Design_Spec.pdf

Proportion (Part-to-whole)

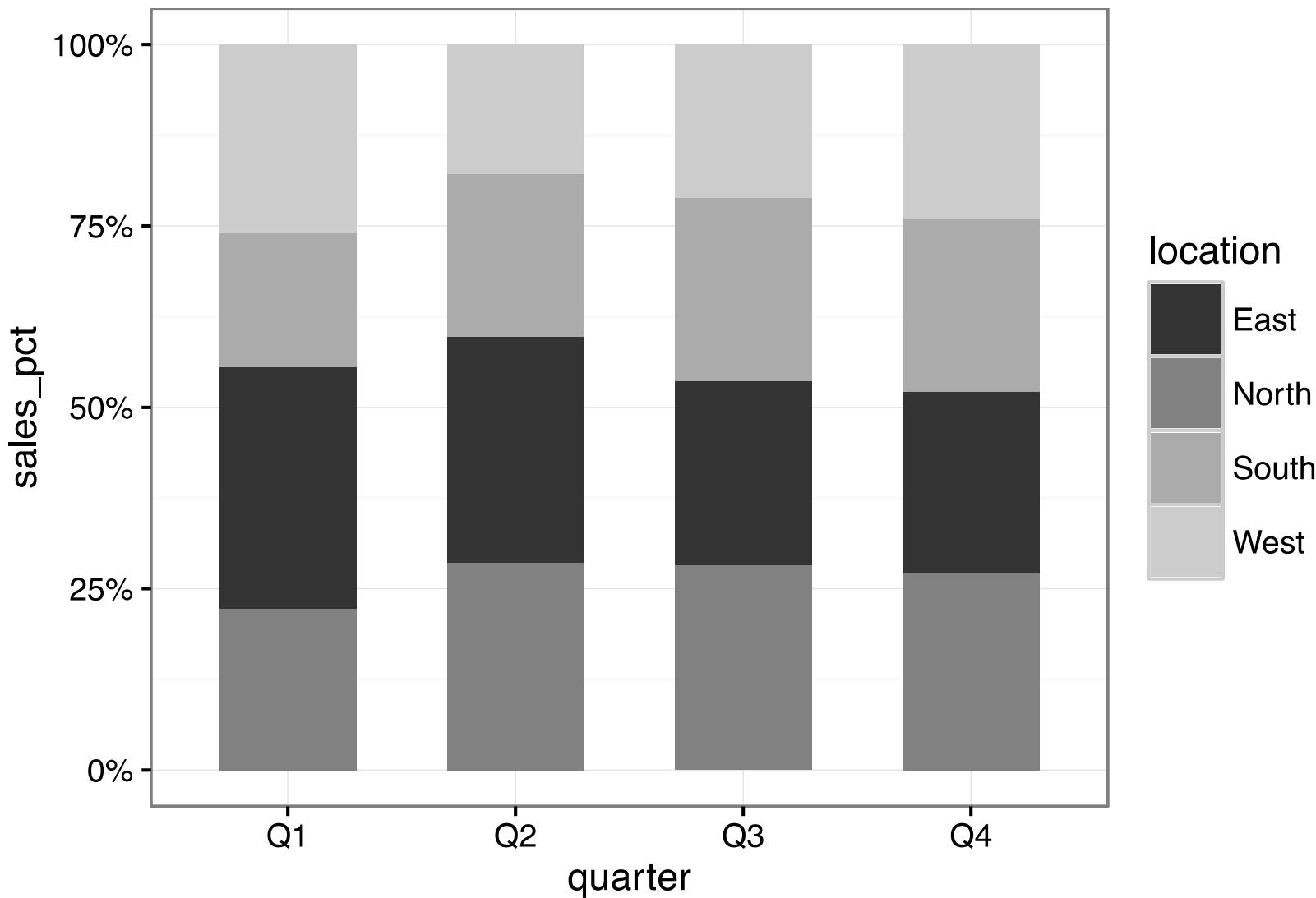
- Best unit: percentage
- Stacked bar graph
 - ◆ Difficult to read individual values
- Stacked area
- Treemap
- Gridplot
- Pie / Donut
- Marimekko



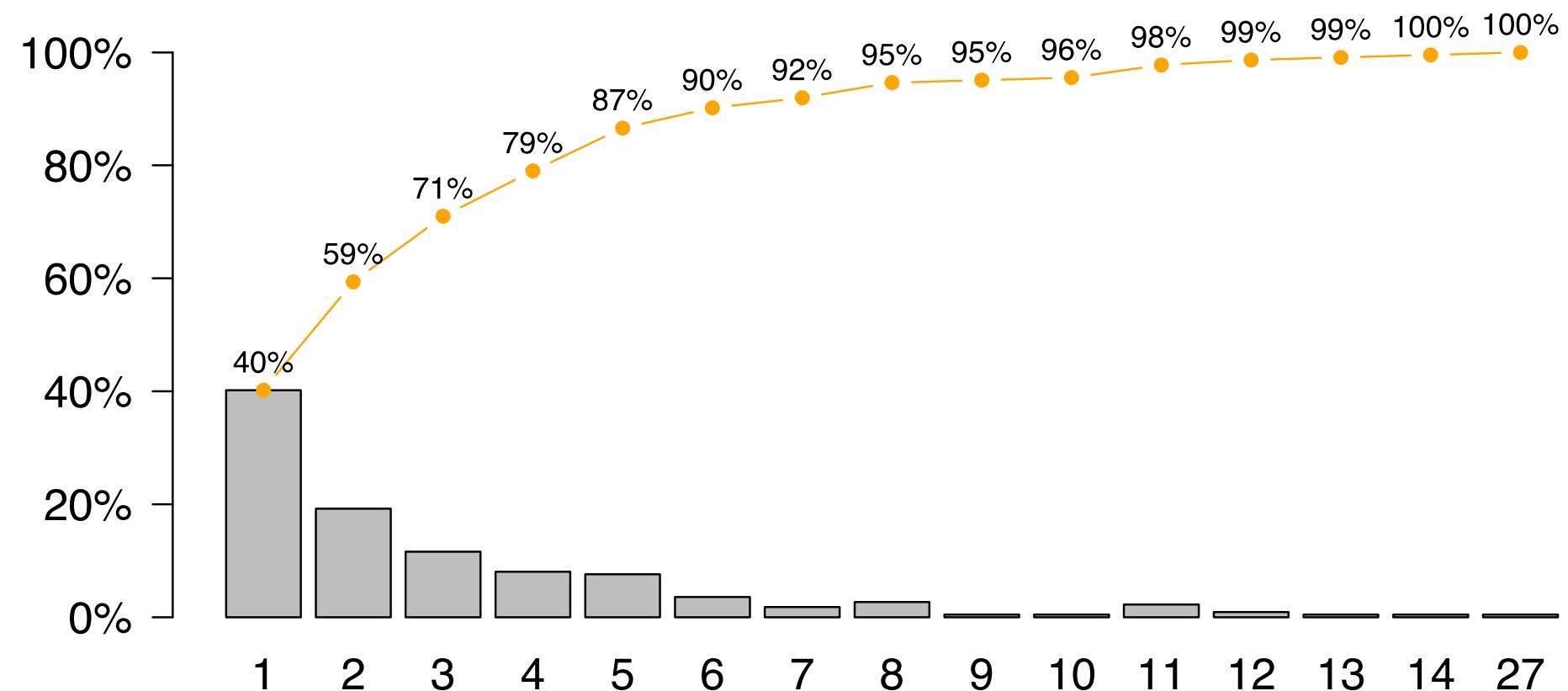
Stacked bar graph



Stacked bar graph percentage



Pareto chart

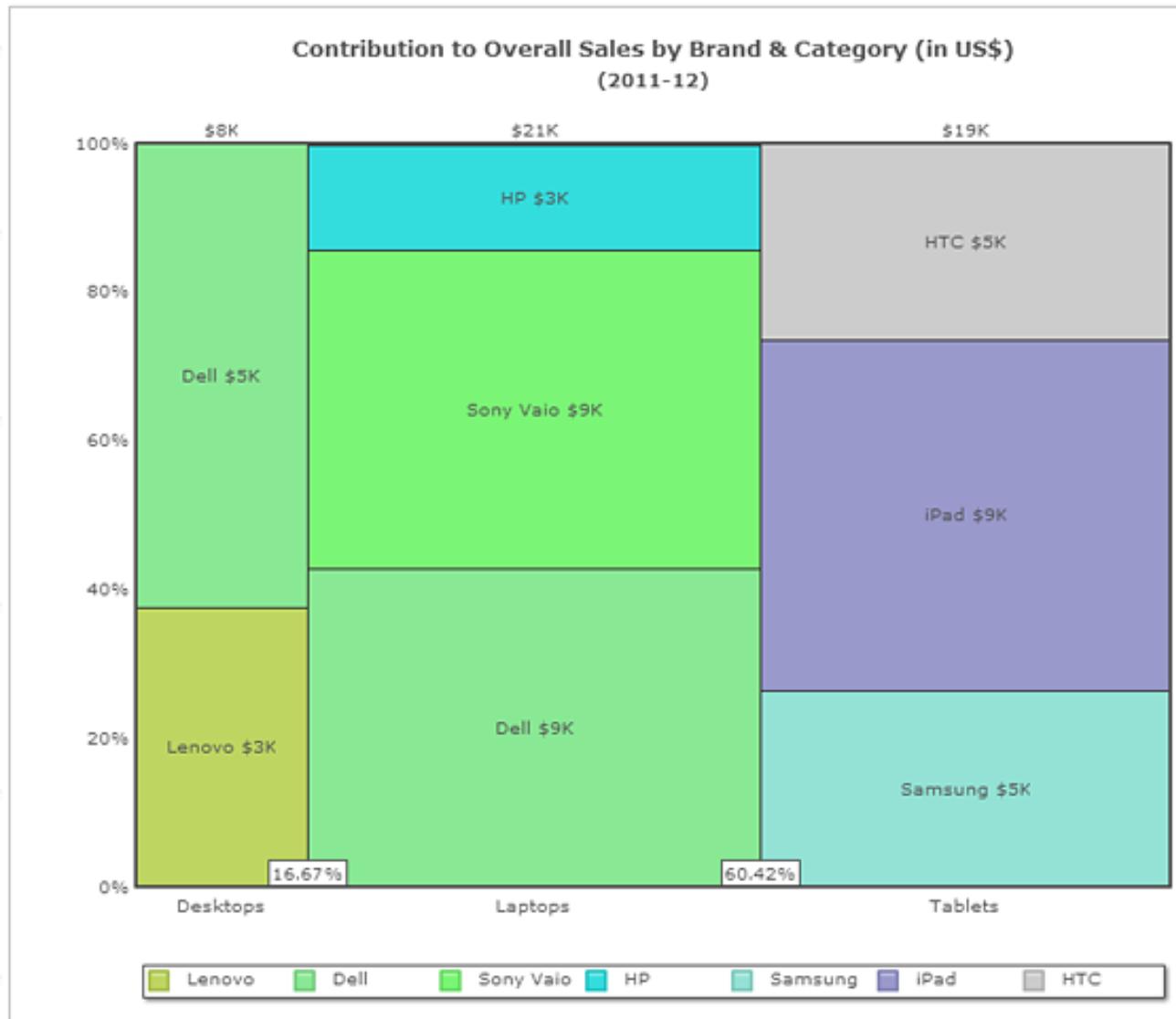


Treemap

proporzione



Marimekko Chart



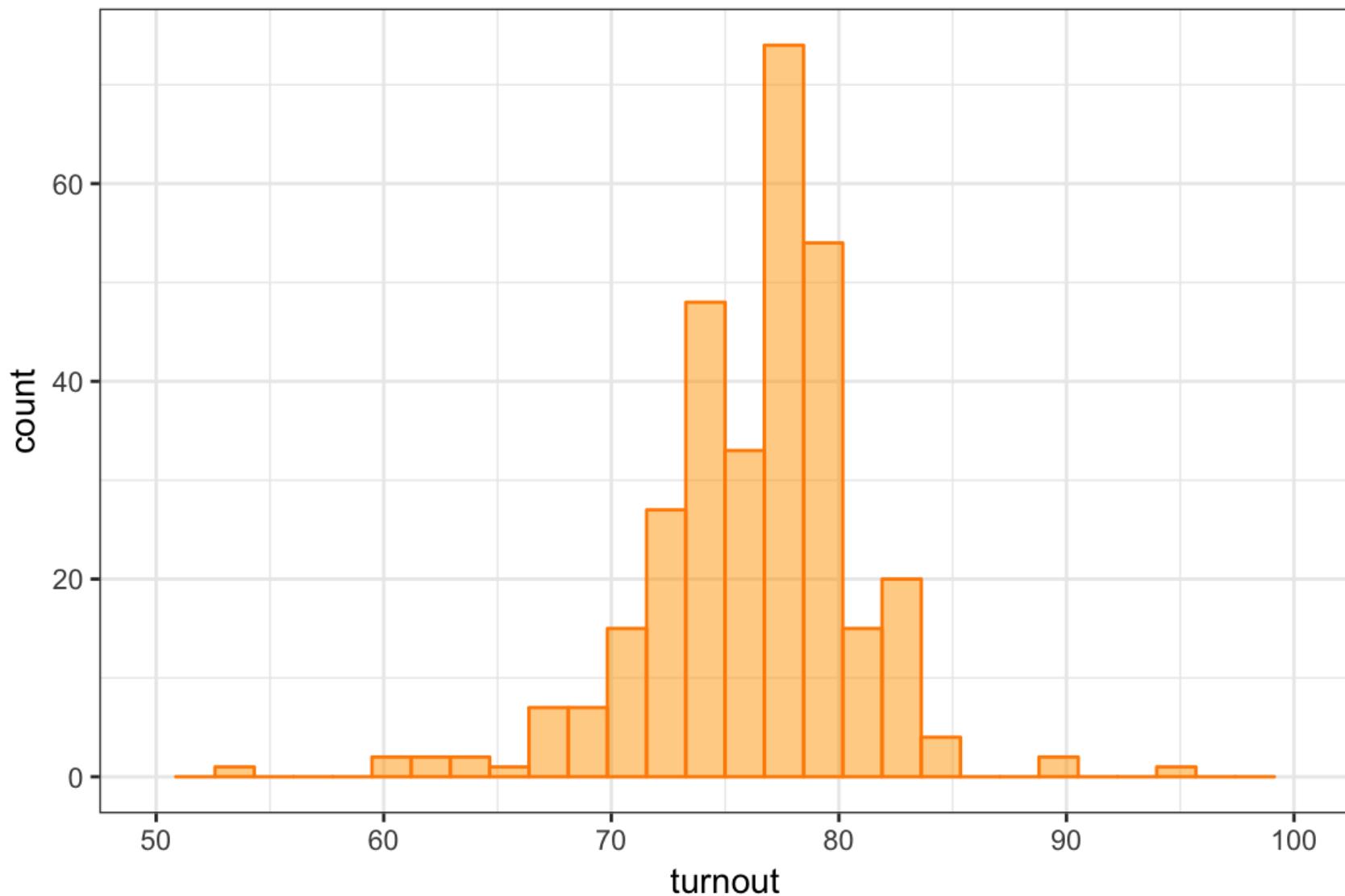
Distribution

- Two main types
 - ◆ Show distribution of single set of values
 - ◆ Show and compare two or more distributions

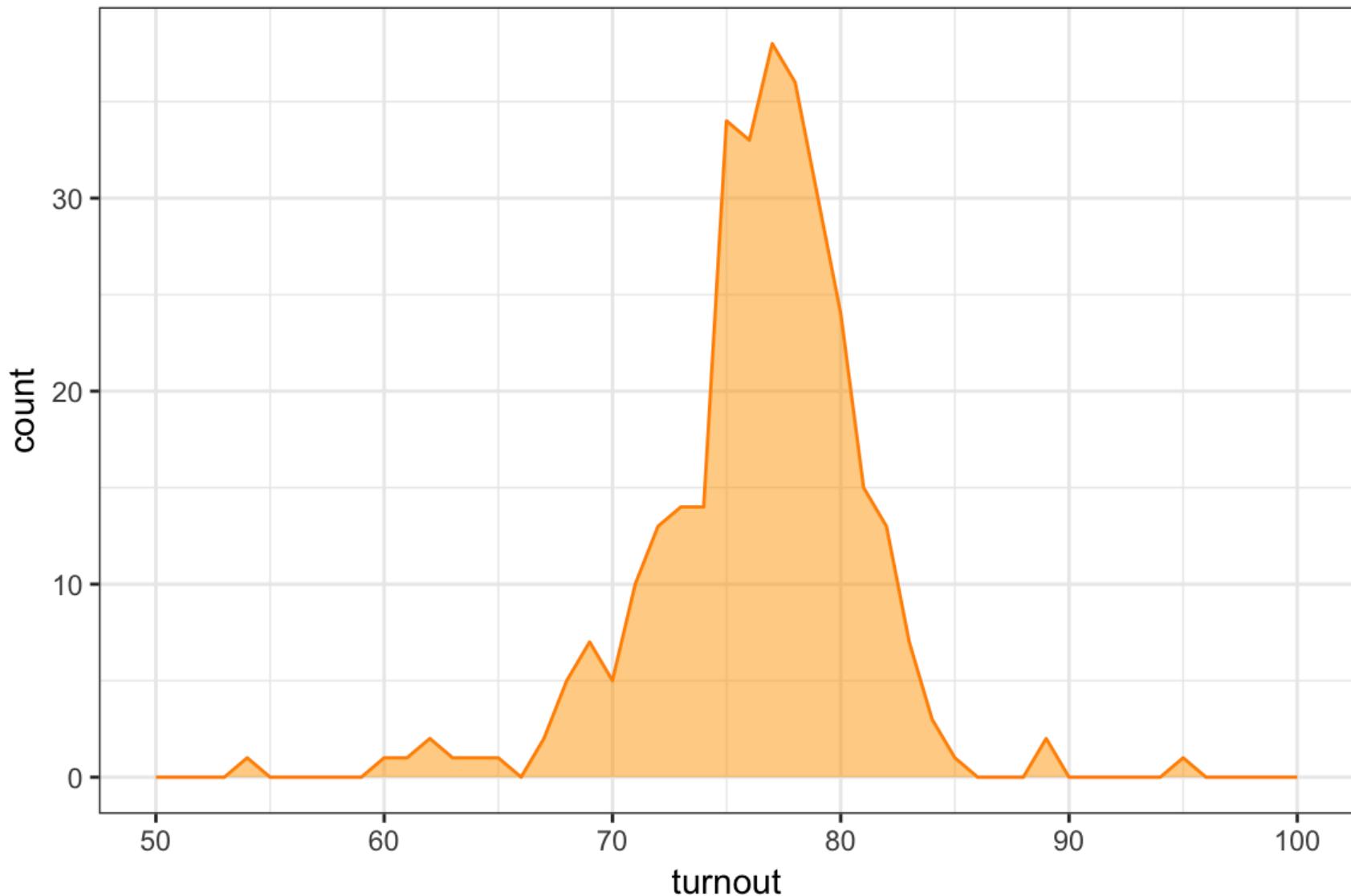
Single distribution

- Histogram
 - ◆ Vertical bar graph
 - ◆ Frequency for subdivision
 - Quantitative ranges
 - Categories
 - ◆ Emphasis on number of occurrences
- Frequency polygon
 - ◆ Line graphs
 - ◆ Frequency density function
 - ◆ Emphasis on the shape of the distribution
- Boxplot
 - ◆ Summary

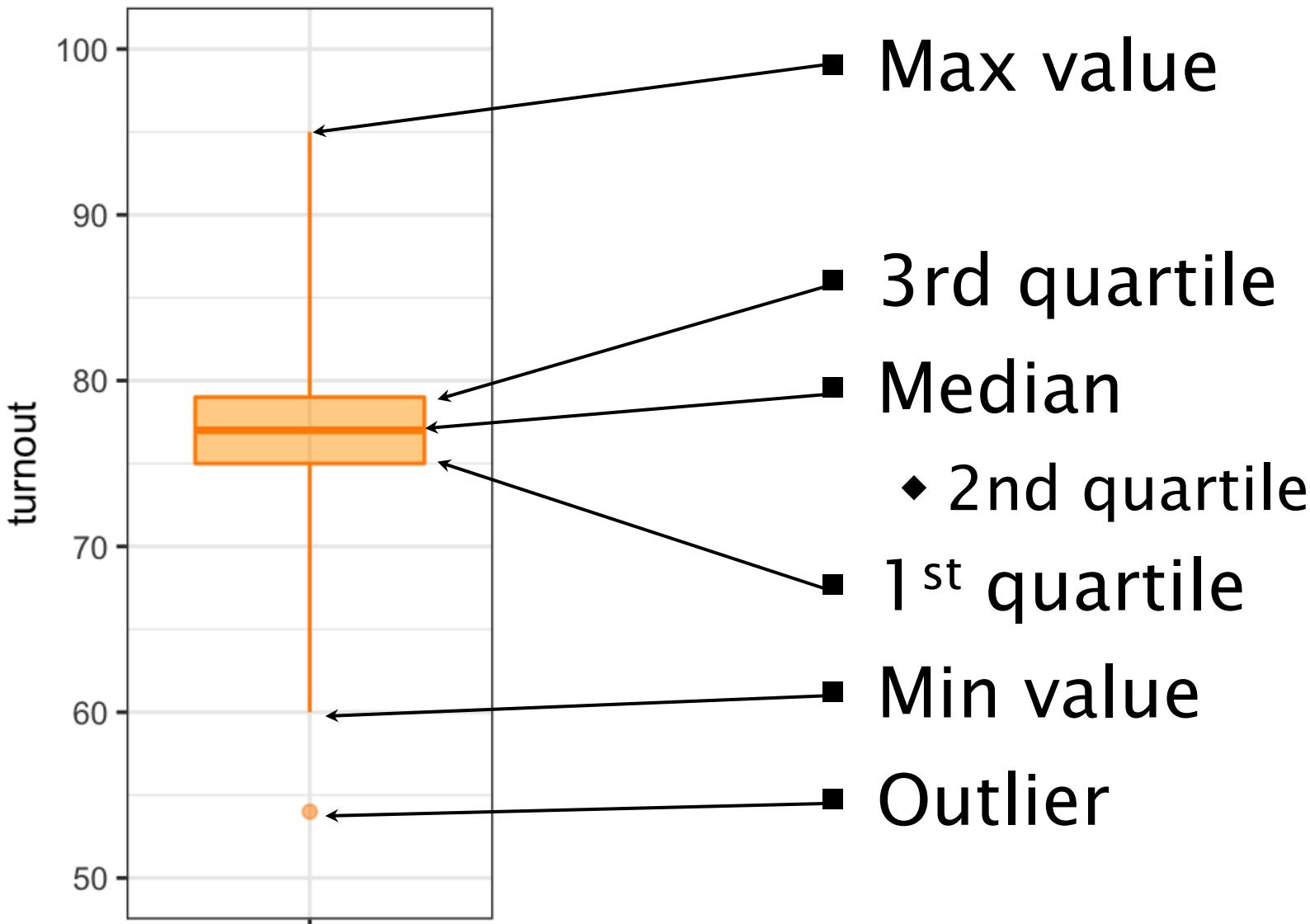
Histogram



Frequency polygon



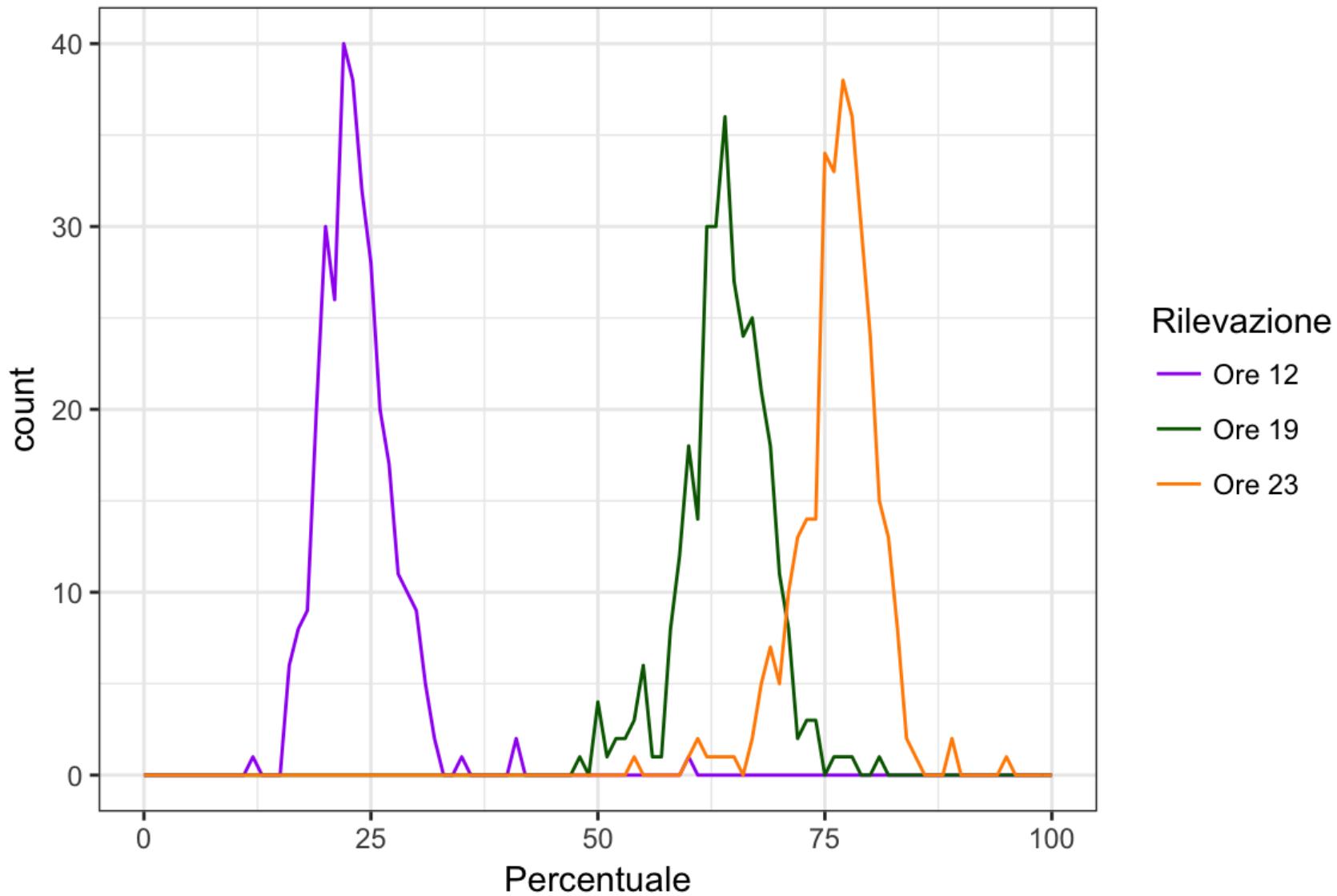
Boxplot



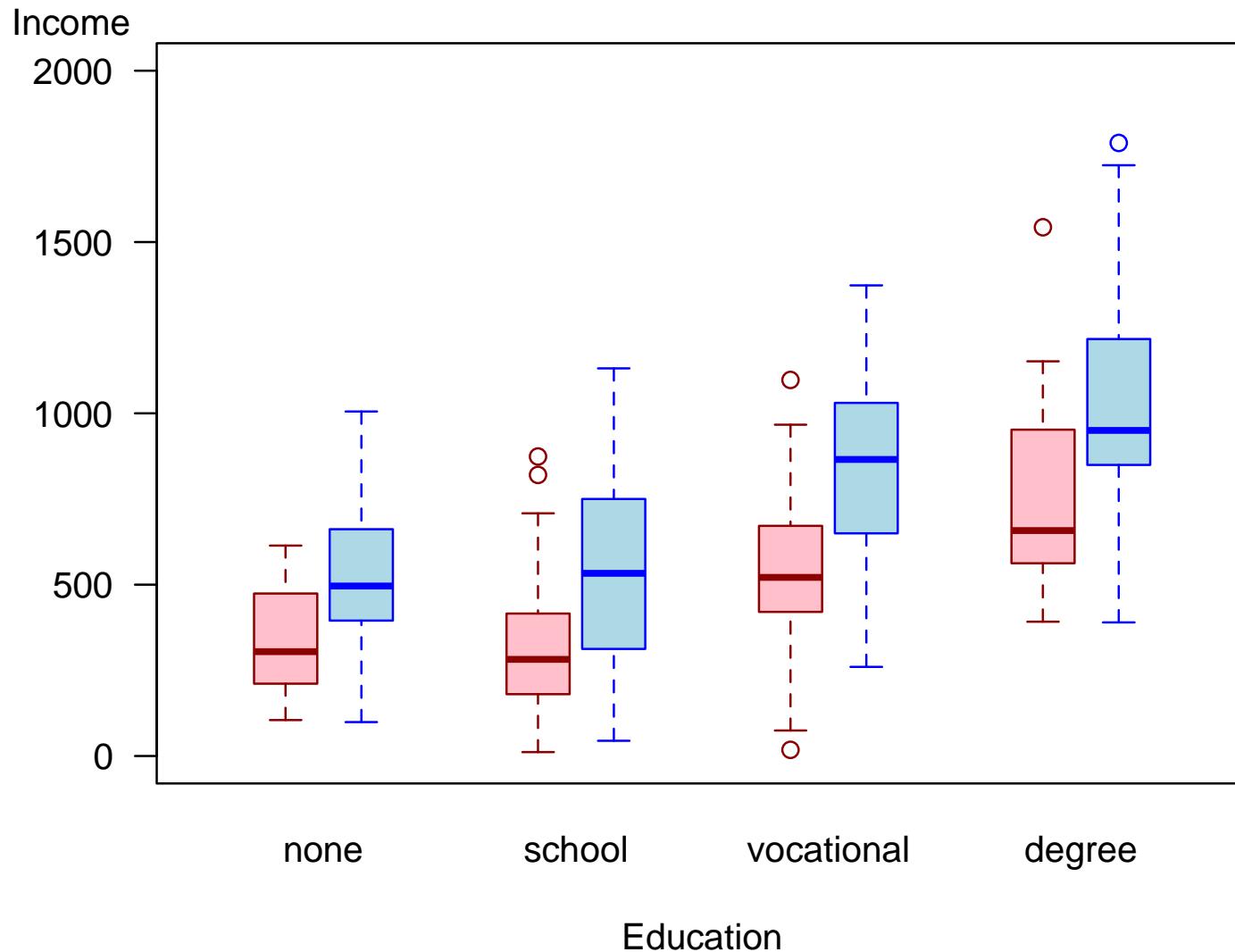
Multiple distribution

- Histogram is not suitable
- Frequency polygon
 - ◆ Line graphs
 - ◆ Frequency density function
- Boxplot
 - ◆ Summary
 - ◆ Less distracting with high number of categories

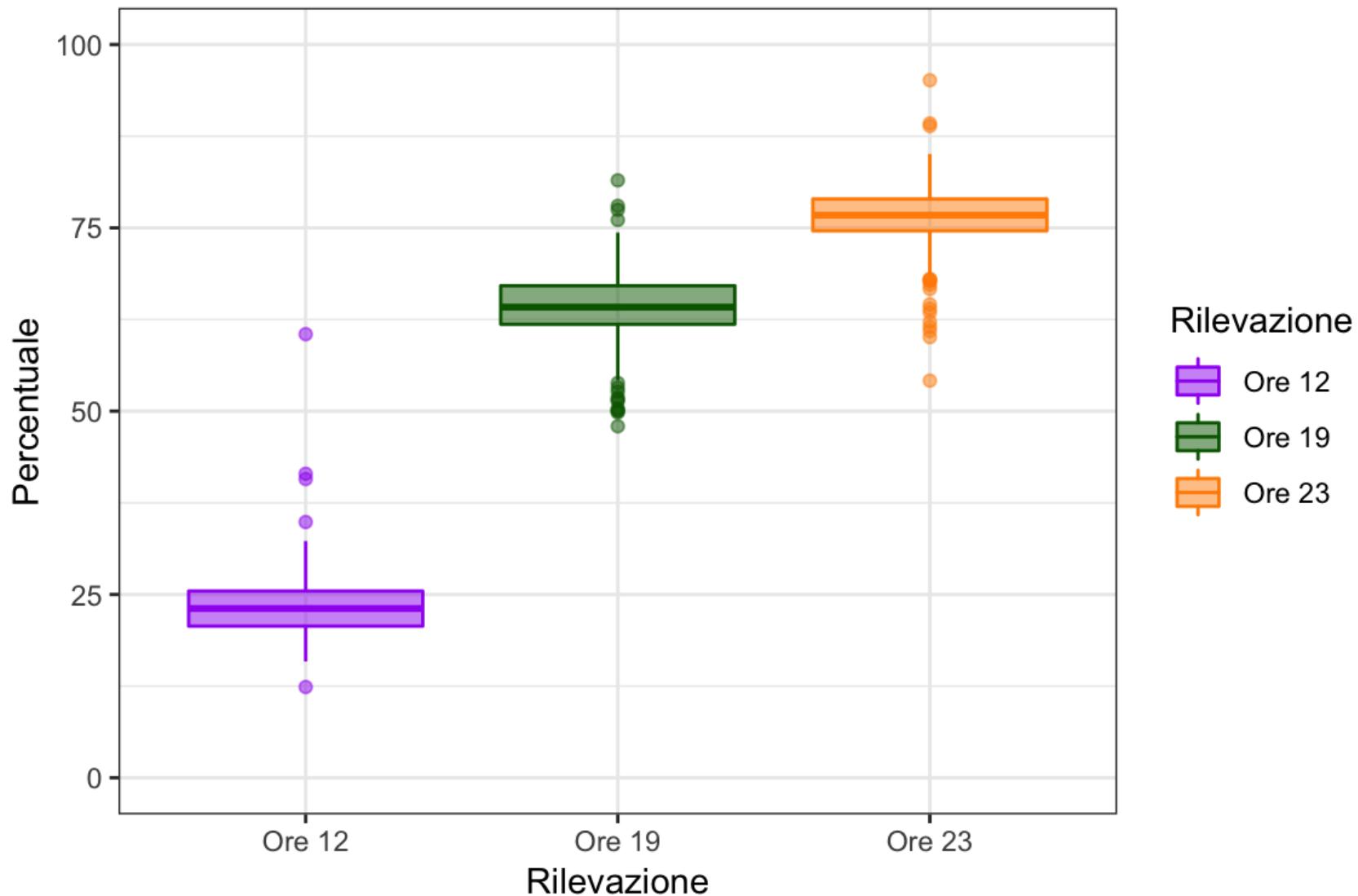
Multiple polygons



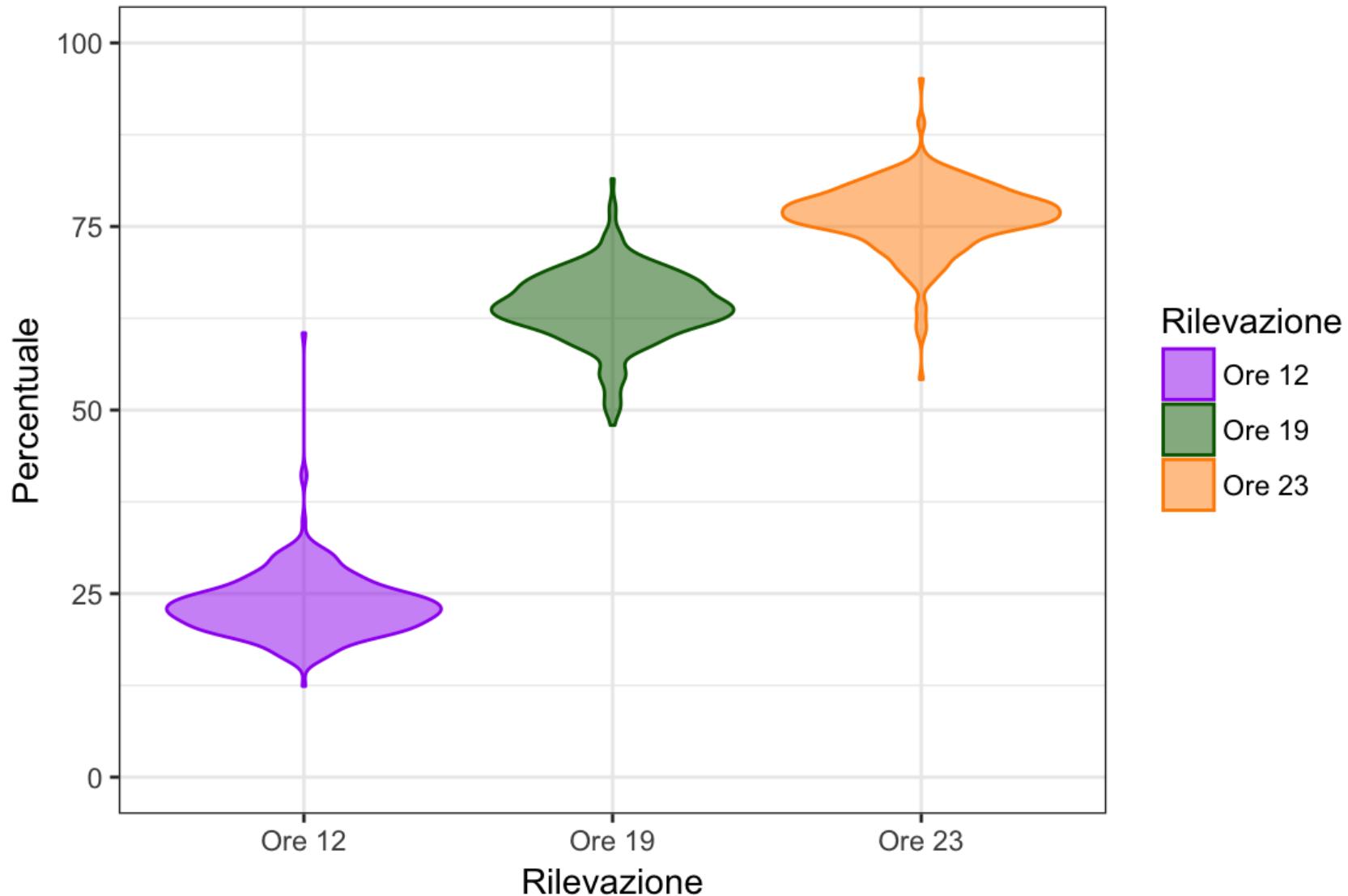
Box plot



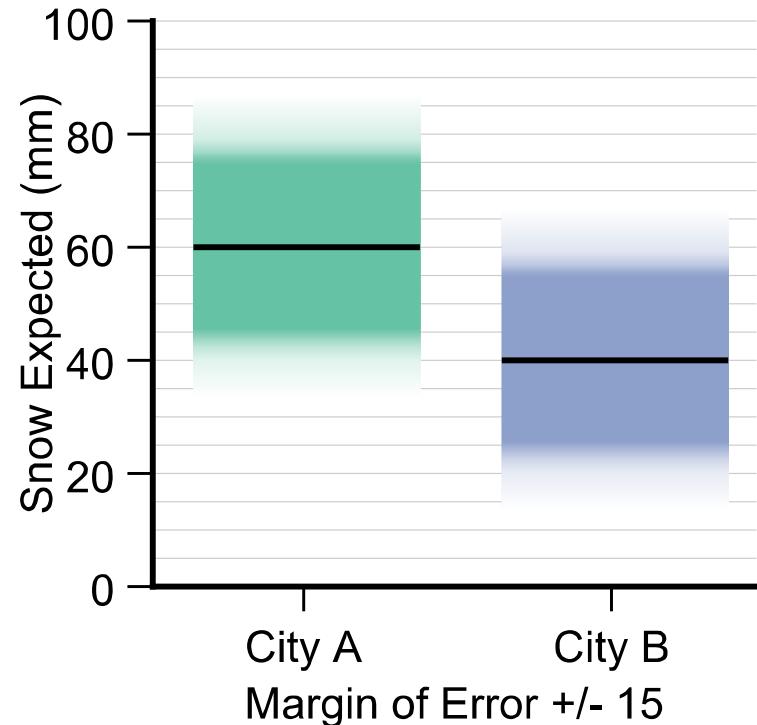
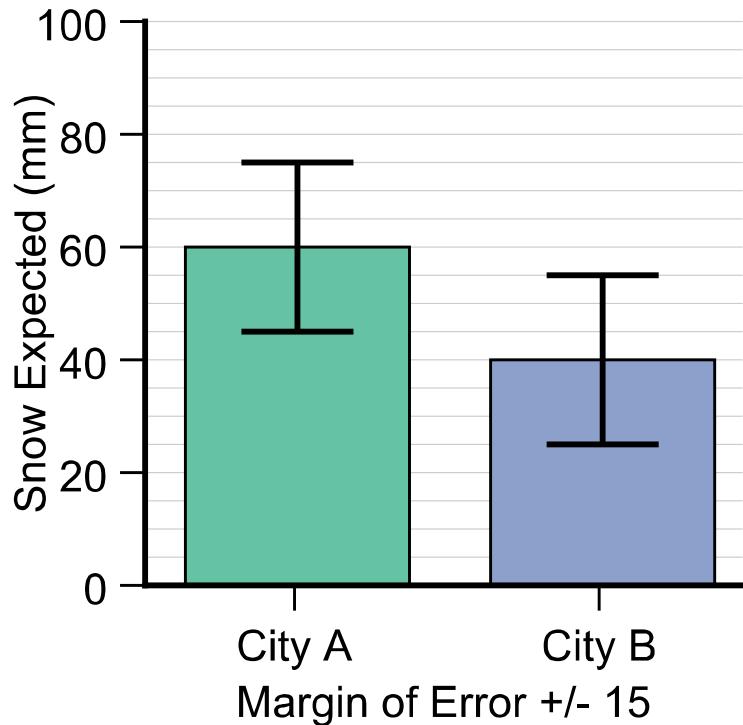
Box plot



Violin plot

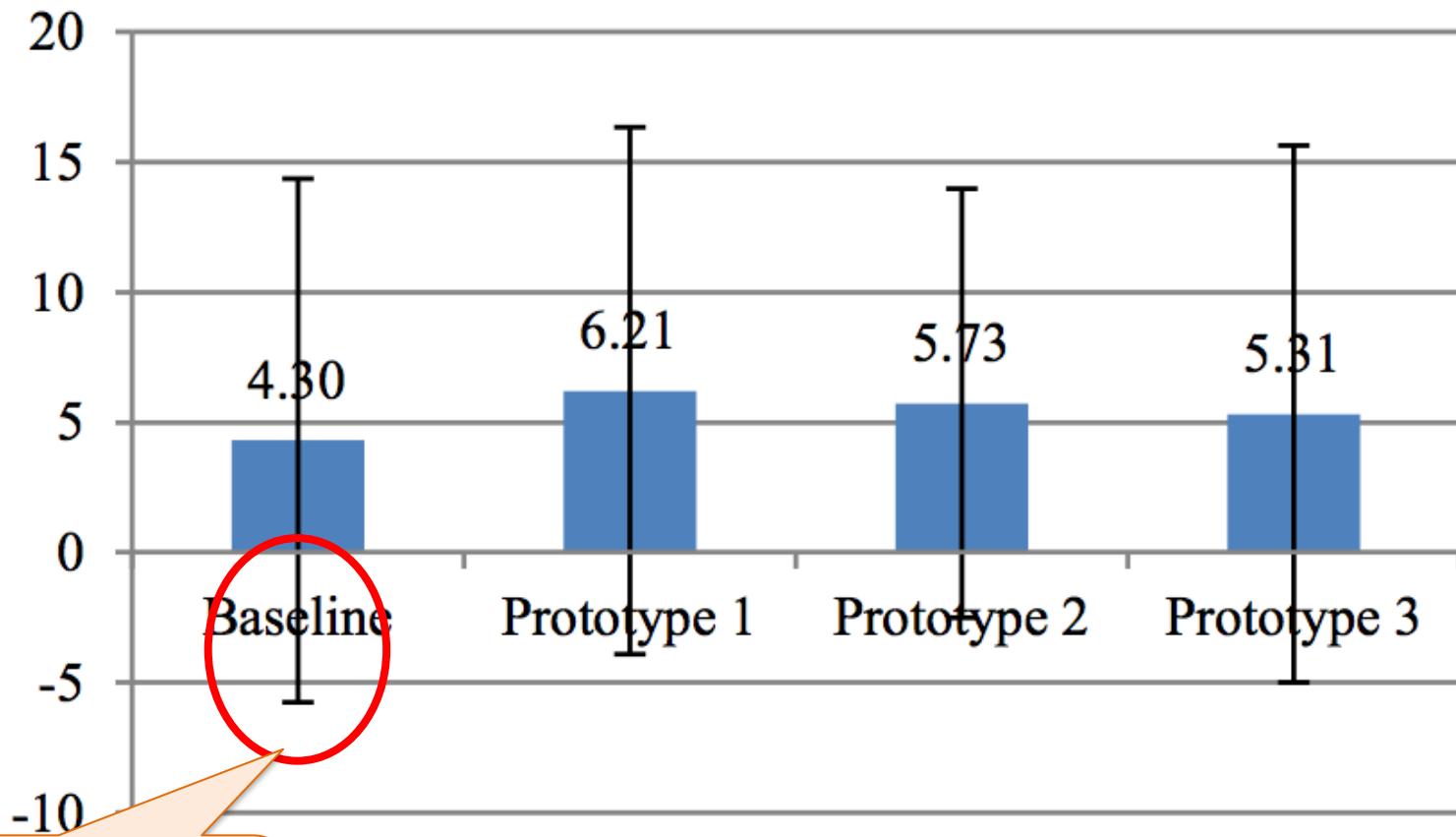


Confidence Intervals



Error Bars Considered Harmful: Exploring Alternate Encodings for Mean and Error
Michael Correll, and Michael Gleicher
IEEE Transactions on Visualization and Computer Graphics, Dec. 2014

Interval may be Asymmetric



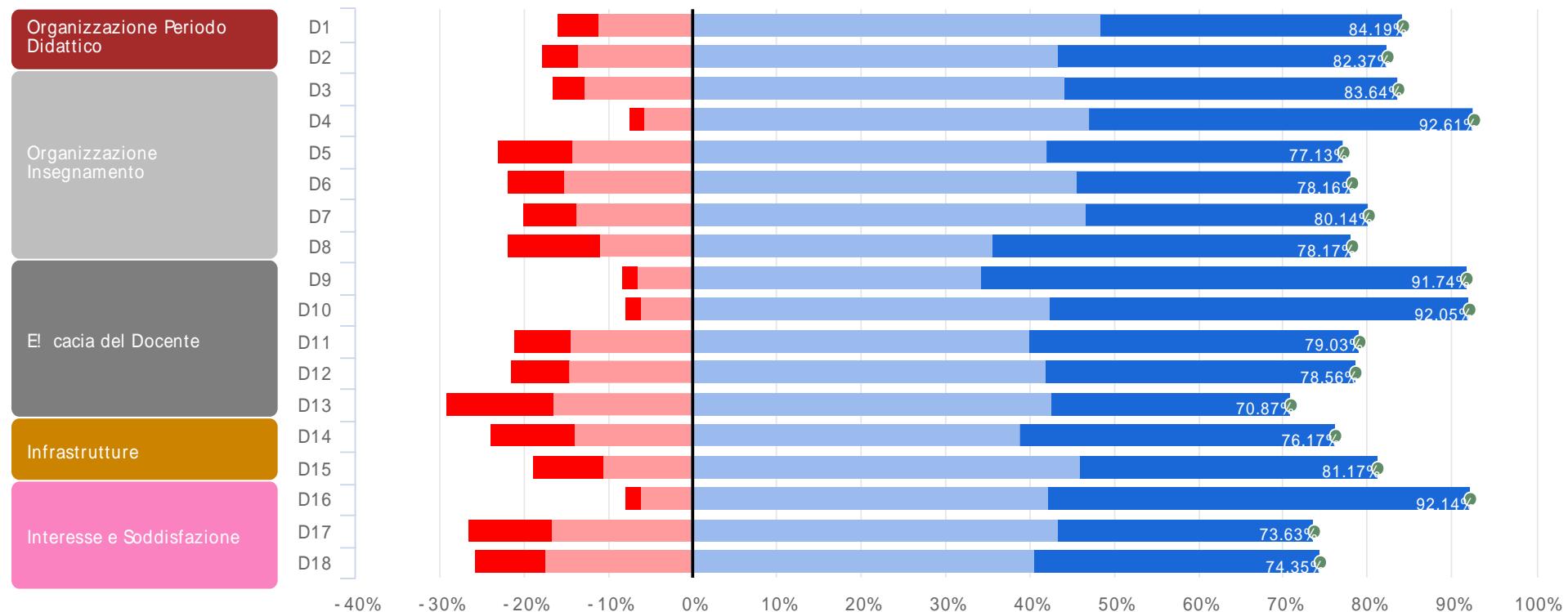
It is physically impossible to modify -6 files

Figure 5. Mean files per changeset.

Likert / Agreement

- Likert scale:
 - ◆ Measures agreement / disagreement with a given statement
 - ◆ Response on an ordinal scale, e.g.
 - Definitely No
 - Mostly No
 - Undecided
 - Mostly Yes
 - Definitely Yes
- Often used to measure positive vs. negative perception

Diverging Stacked Bars

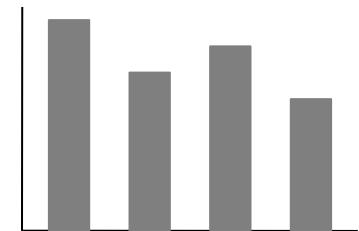
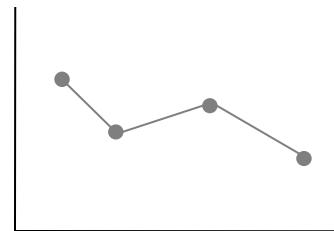
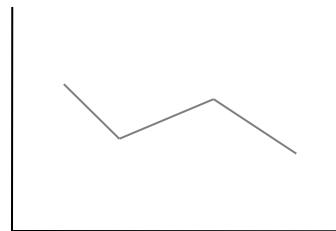


Time series

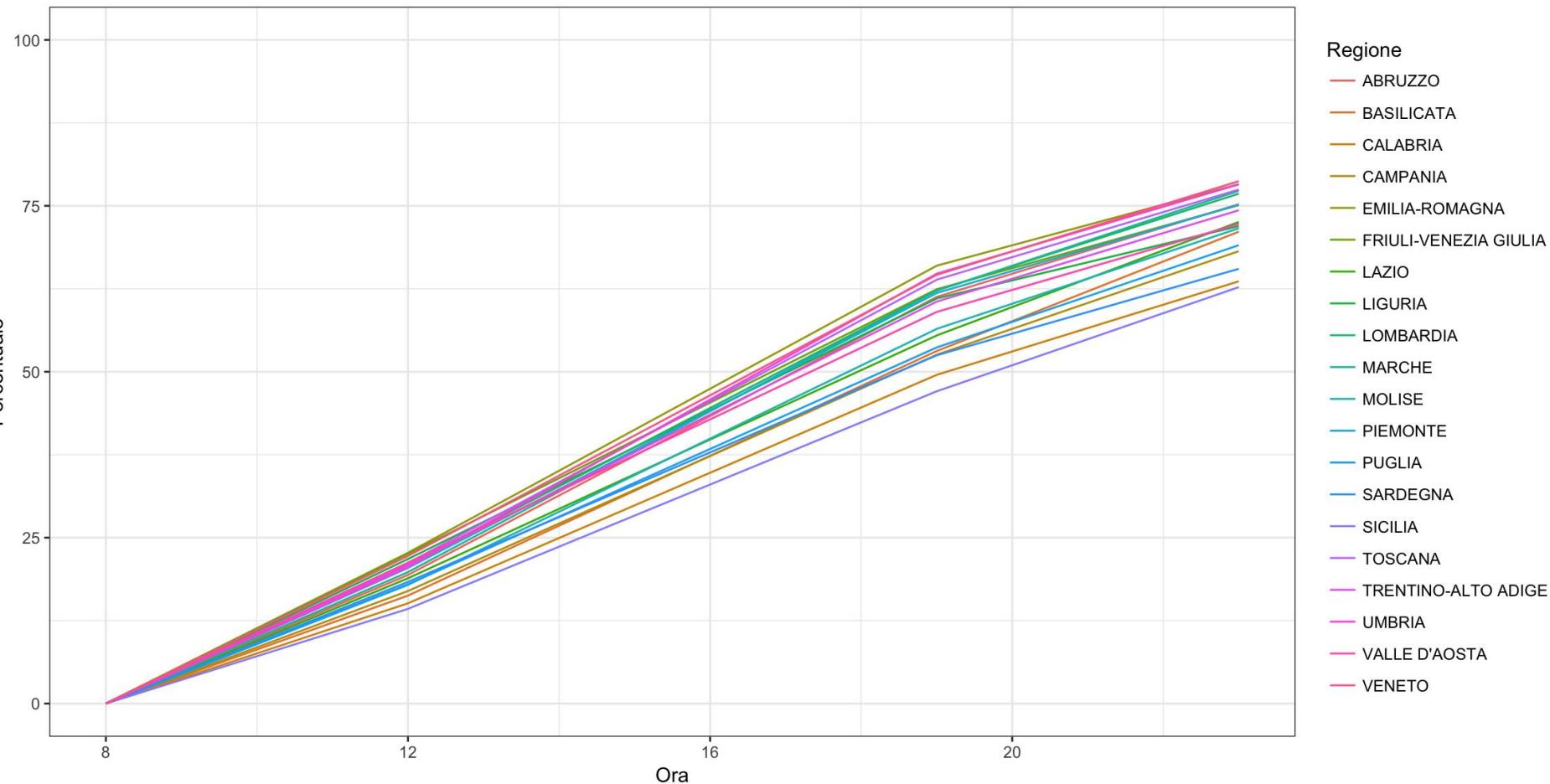
- Series of relationships between quantitative values that are associated with categorical subdivisions of time
- Communicate
 - ◆ Change
 - ◆ Rise
 - ◆ Increase
 - ◆ Fluctuate
 - ◆ Grow
 - ◆ Decline
 - ◆ Decrease
 - ◆ Trend

Time series

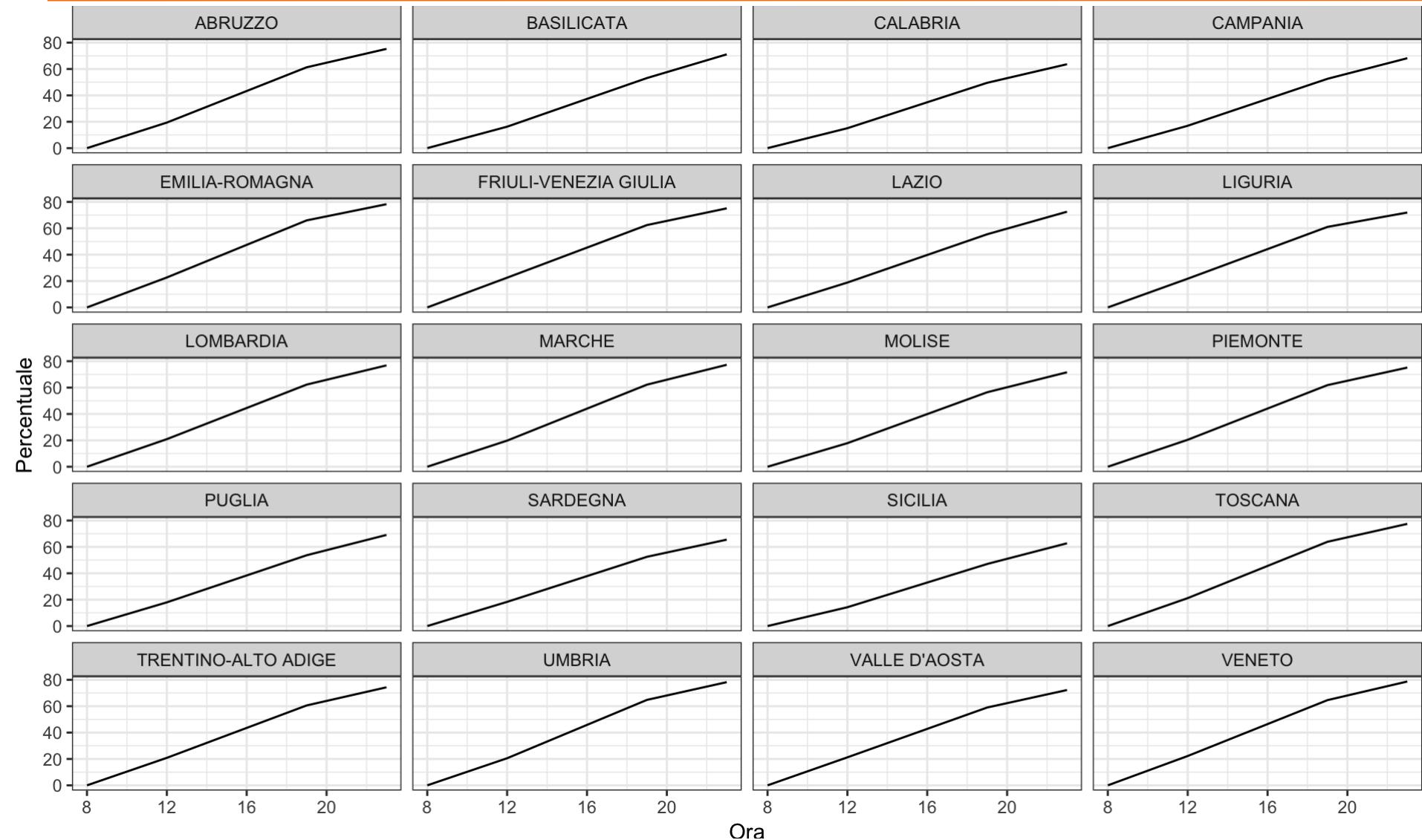
- Time grows from left to right
 - ◆ Cultural convention
- Vertical bars
 - ◆ highlight individual points in time
 - ◆ hide overall trend



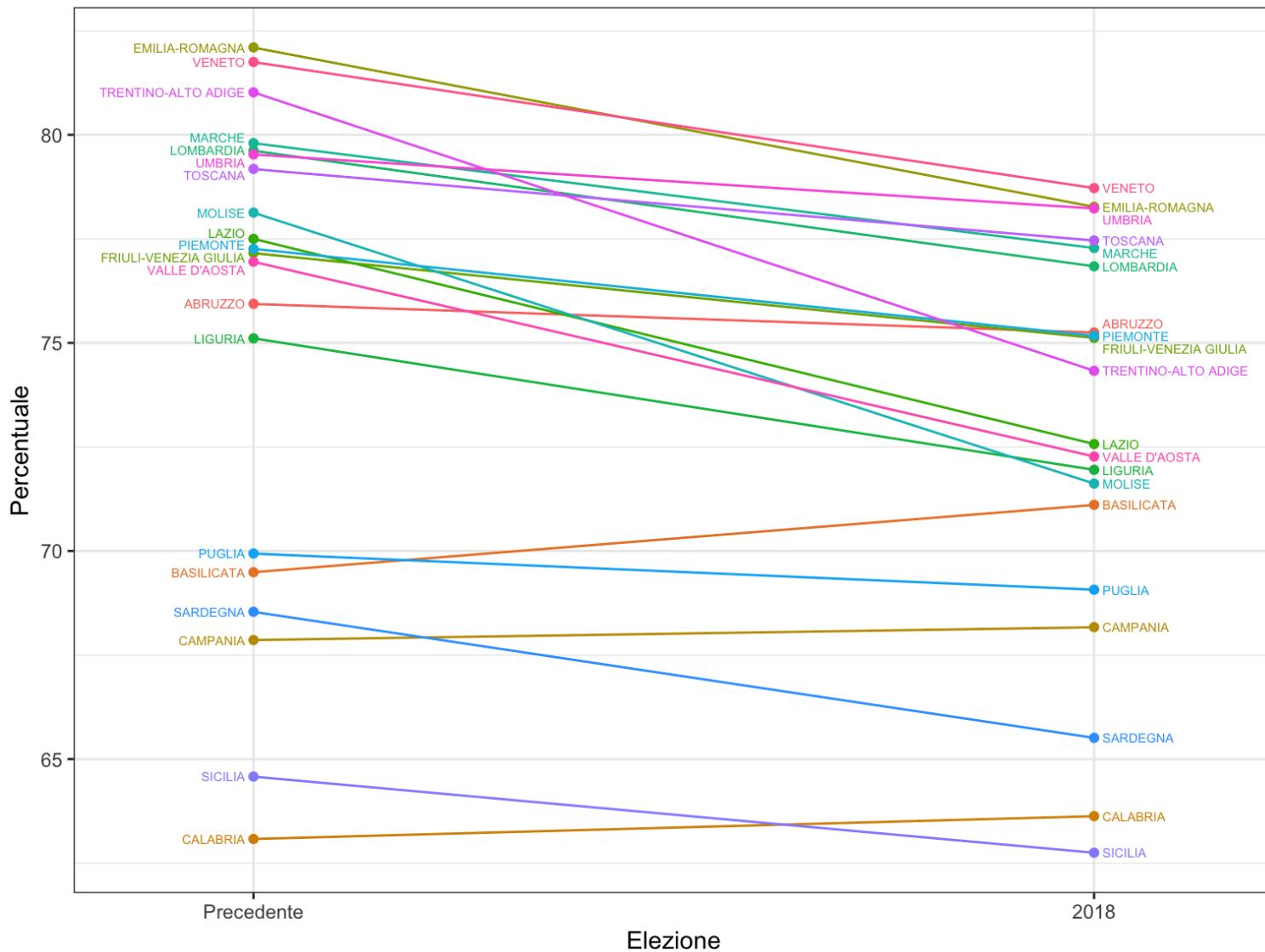
Trends



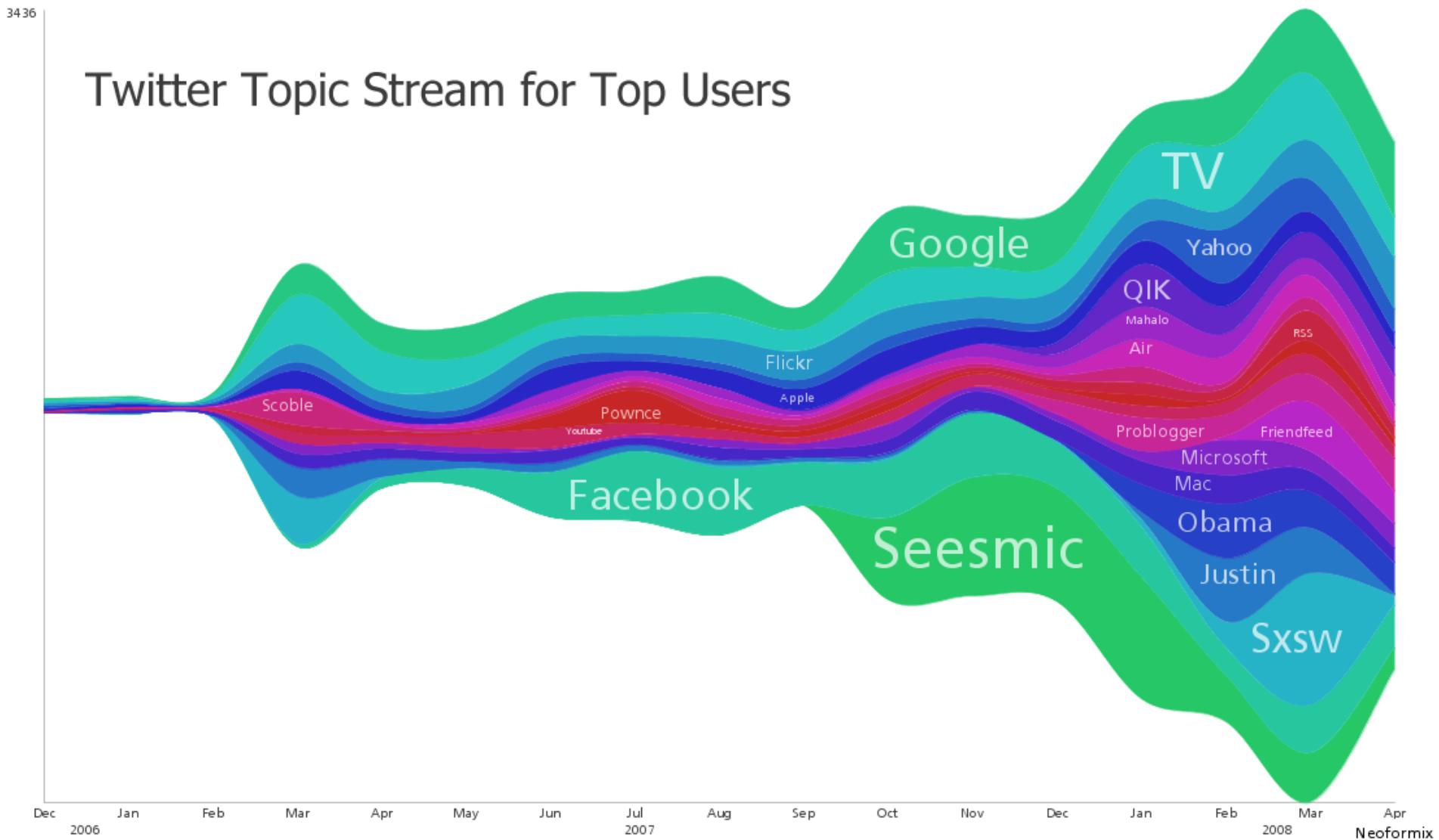
Lines small multiples



Slope chart



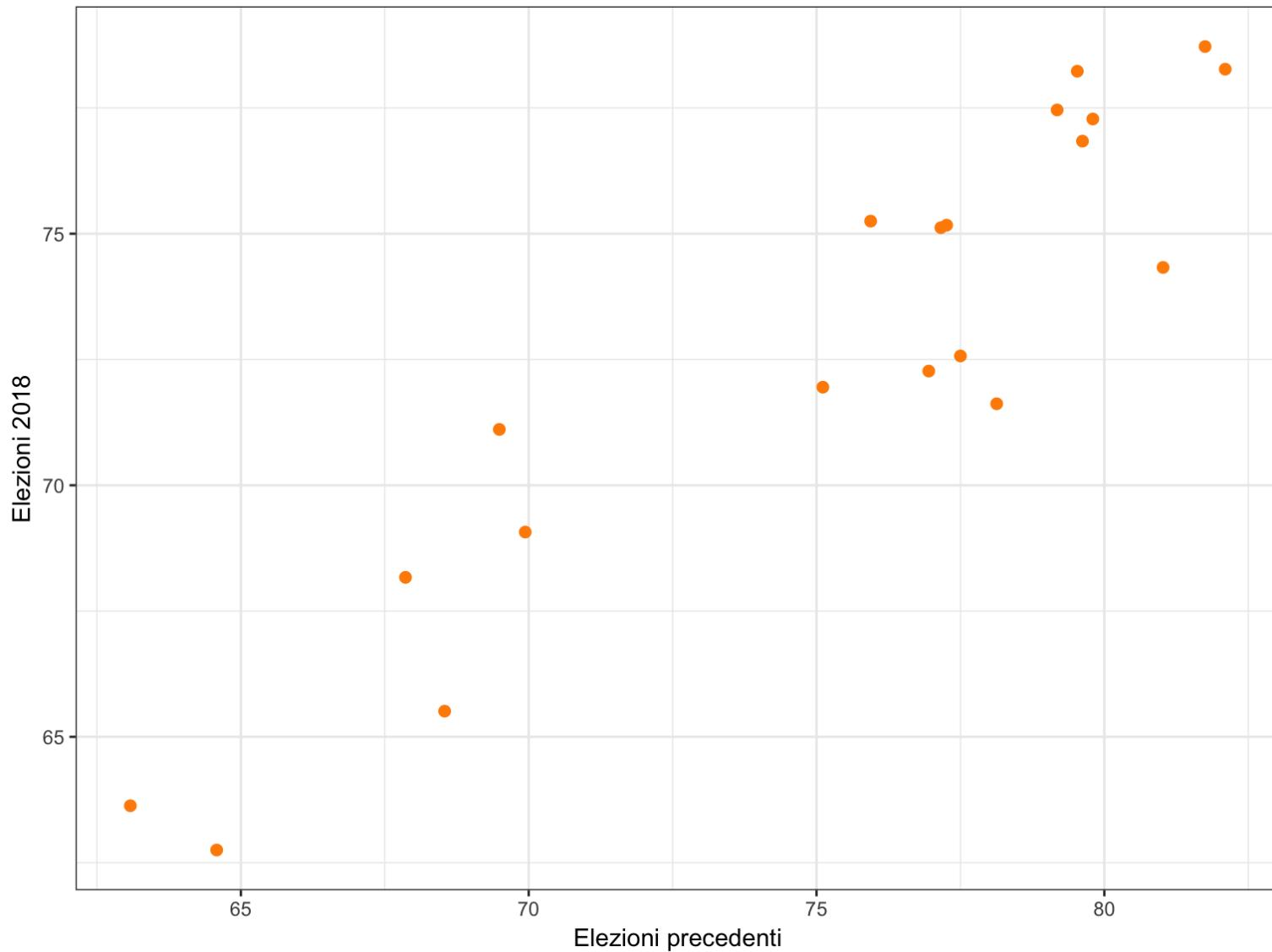
Streamgraph



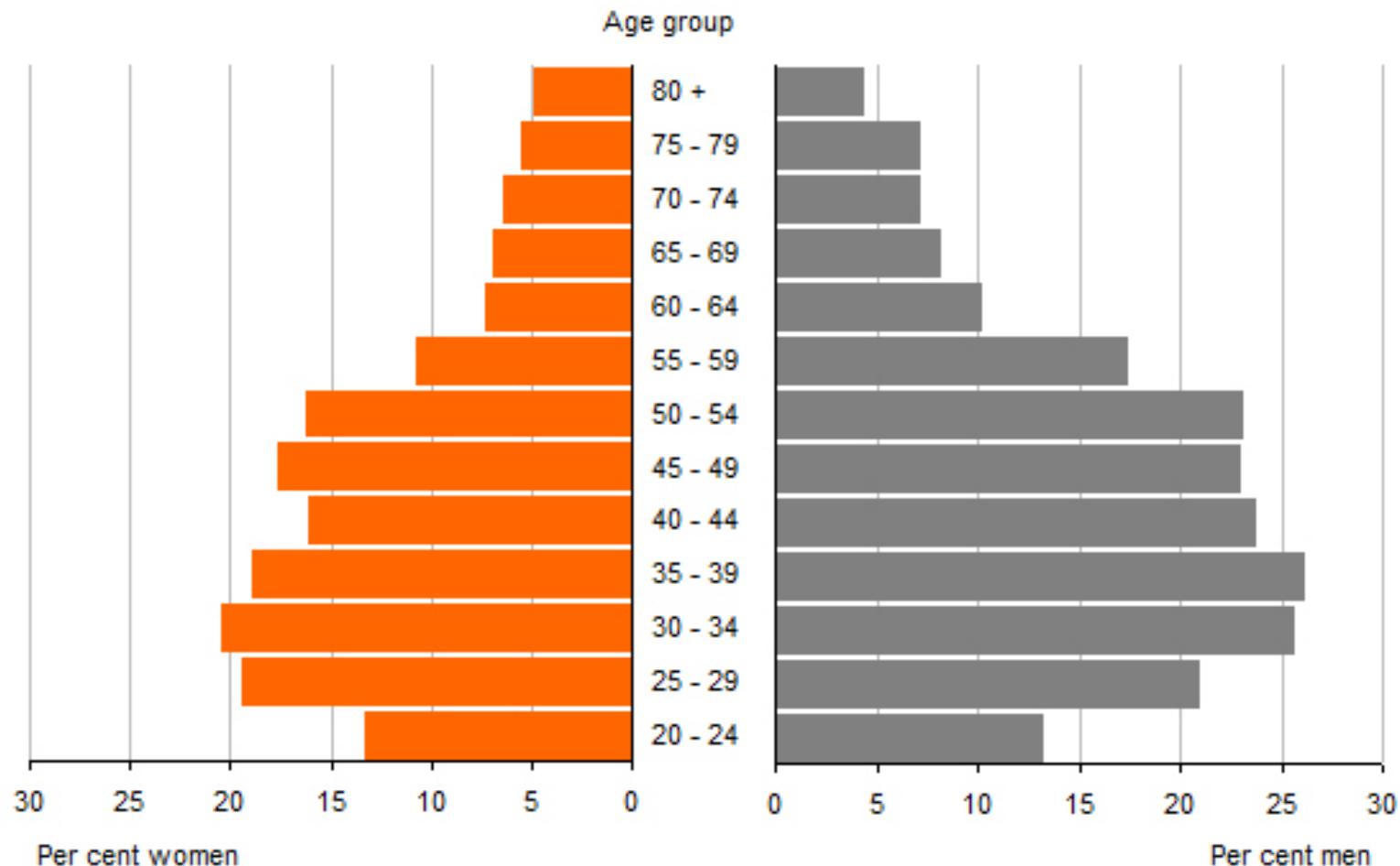
Correlation

- Relationships between two paired sets of quantitative values
 - ◆ Scatter plot w/possible trend line
 - Ok for educated audience
 - ◆ Paired bar graph

Scatter plot



Paired bargraph



[https://unstats.un.org/unsd/genderstatmanual/
Print.aspx?Page=Presentation-of-gender-statistics-in-graphs](https://unstats.un.org/unsd/genderstatmanual/Print.aspx?Page=Presentation-of-gender-statistics-in-graphs)

DASHBOARD

Dashboard

Visualization of the most relevant information

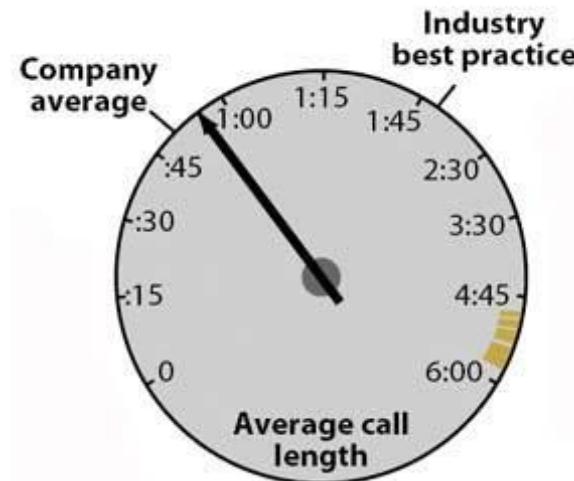
needed to achieve one or more goals
which fits entirely on a single screen
so it can be monitored at a glance

Dashboard

- Dashboards display mechanisms are
 - ◆ small
 - ◆ concise
 - ◆ clear
 - ◆ intuitive
- Dashboards are customized
 - ◆ To suit the goals of person, group, function

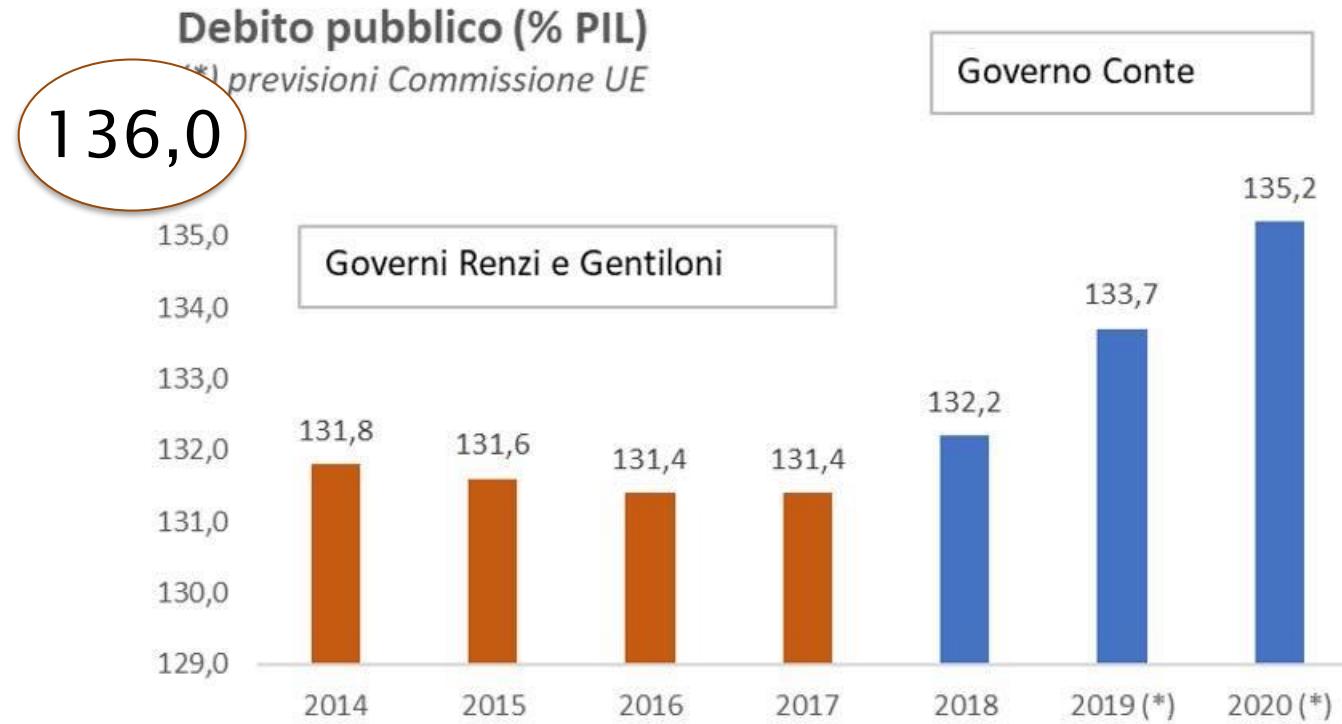
Provide context for data

- References allow judging the data



Use appropriate detail

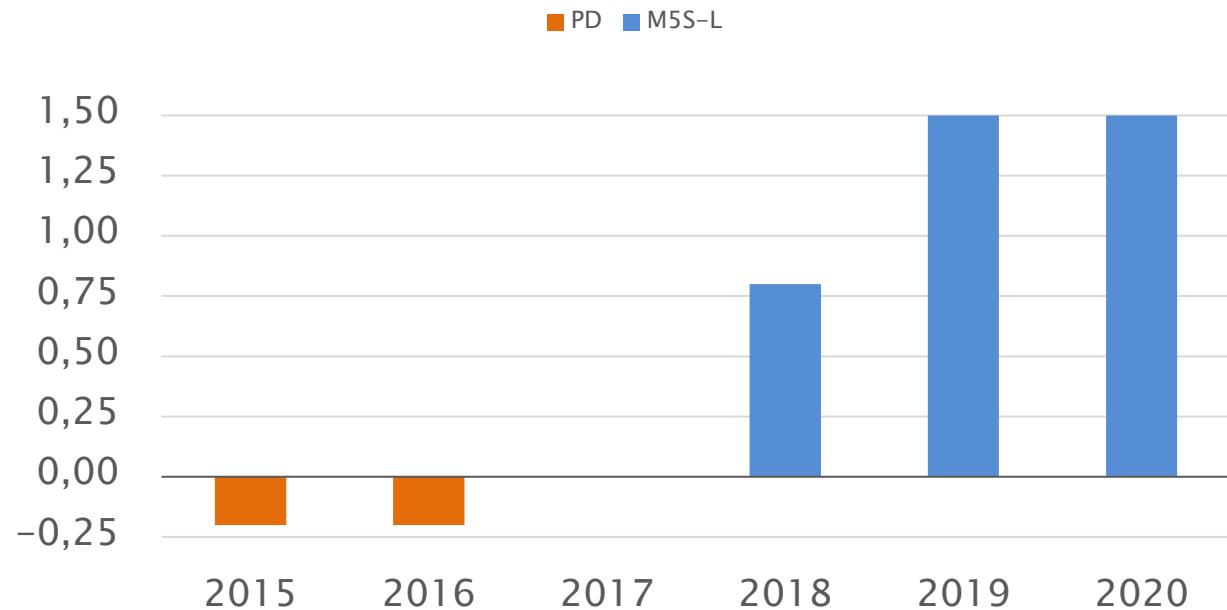
- Typical counter-examples
 - ◆ Dates with seconds detail
 - ◆ Decimals



Use the right measures

- If you are interested in e.g. the difference, ratio, variation show such derived measure

Variazione Debito Pubblico (% PIL)



Use appropriate visualization

- Typical errors:
 - ◆ Any chart when a table would be better
 - ◆ Pie-charts not representing part-whole
 - ◆ Bubble charts

Visualization instruments

- Tables
 - ◆ Textual information
- Graphs
 - ◆ Visual information

Avoid decorations

- Skeuomorphic design
- Backgrounds motives
- Color gradients
- Variations not encoding any measure
 - ◆ Typically color

Avoid decorations

- Skeuomorphic design
- Backgrounds motives
- Color gradients
- Variations not encoding any measure
 - ◆ Typically color



-VS-



Avoid decorations

- Skeuomorphic design
- Backgrounds motives
- Color gradients
- Variations not encoding any measure
 - ◆ Typically color

A

B

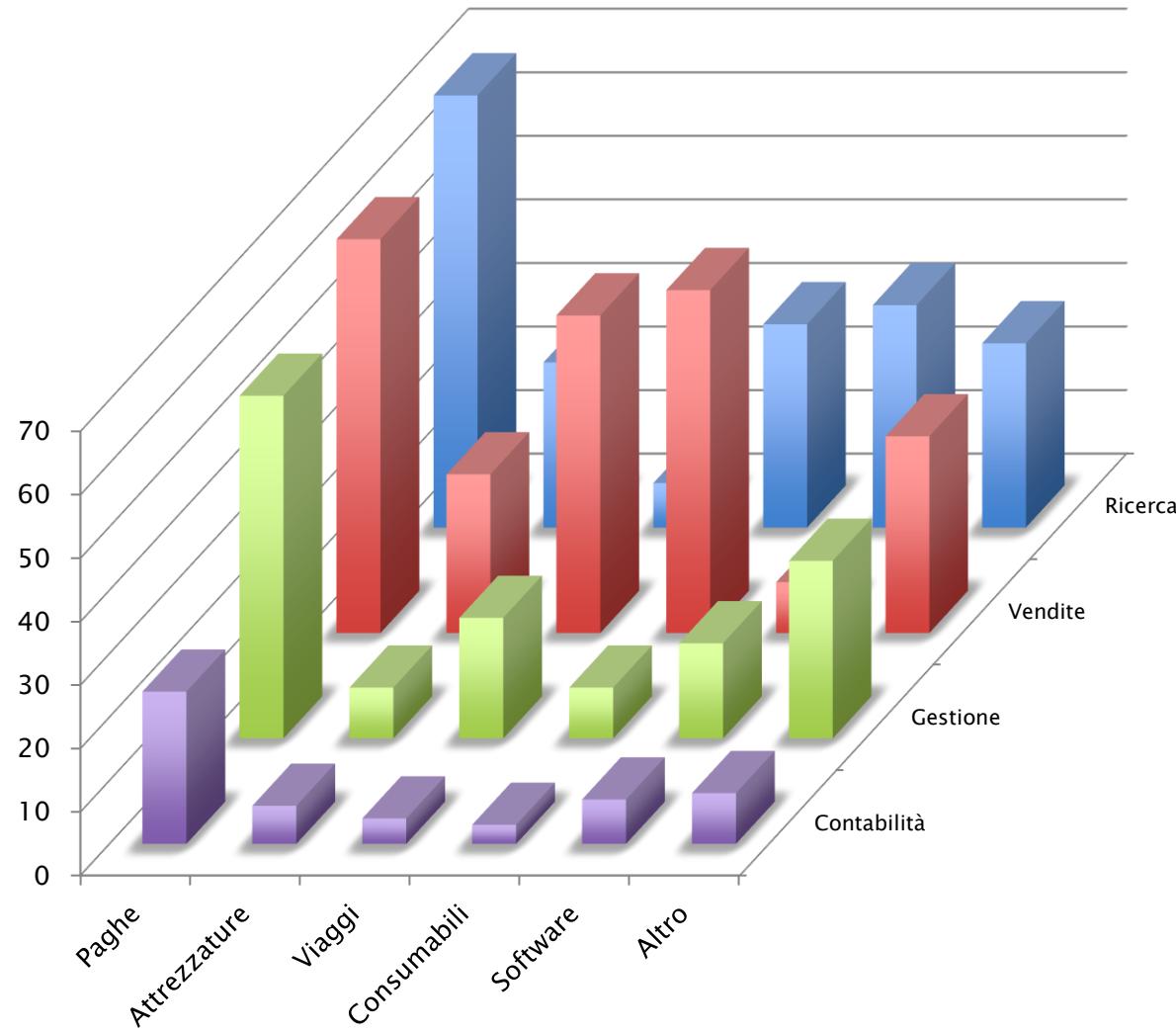
Avoid decorations

- Skeuomorphic design
- Backgrounds motives
- Color gradients
- Variations not encoding any measure
 - ◆ Typically color

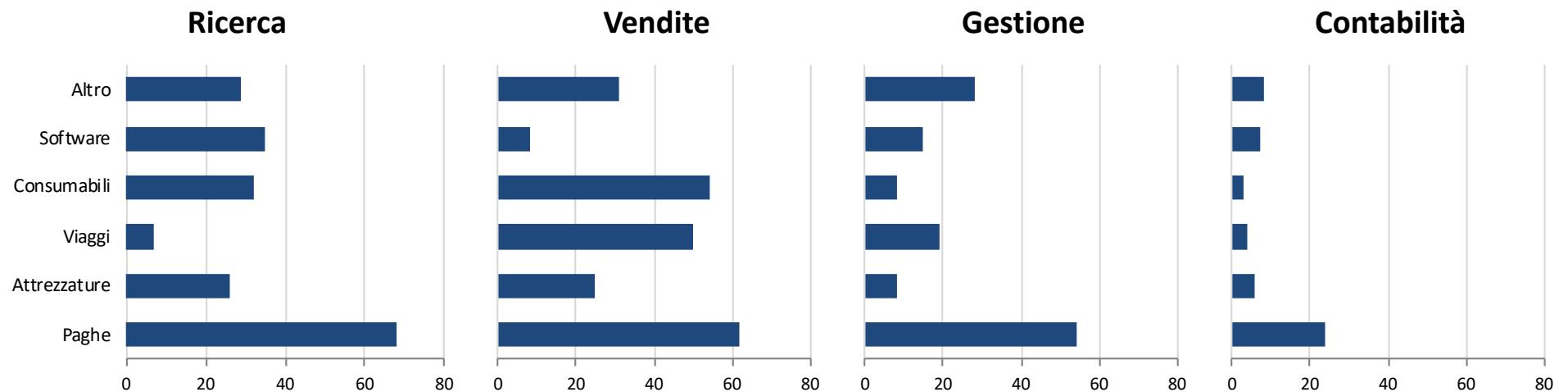
3D diagrams

- Encoding
 - ◆ Axonometry typically hides some data and makes comparison hard
- Not encoding
 - ◆ Perspective deform dimensions
 - ◆ Depth or height distract and make comparison more difficult

Encoding 3D

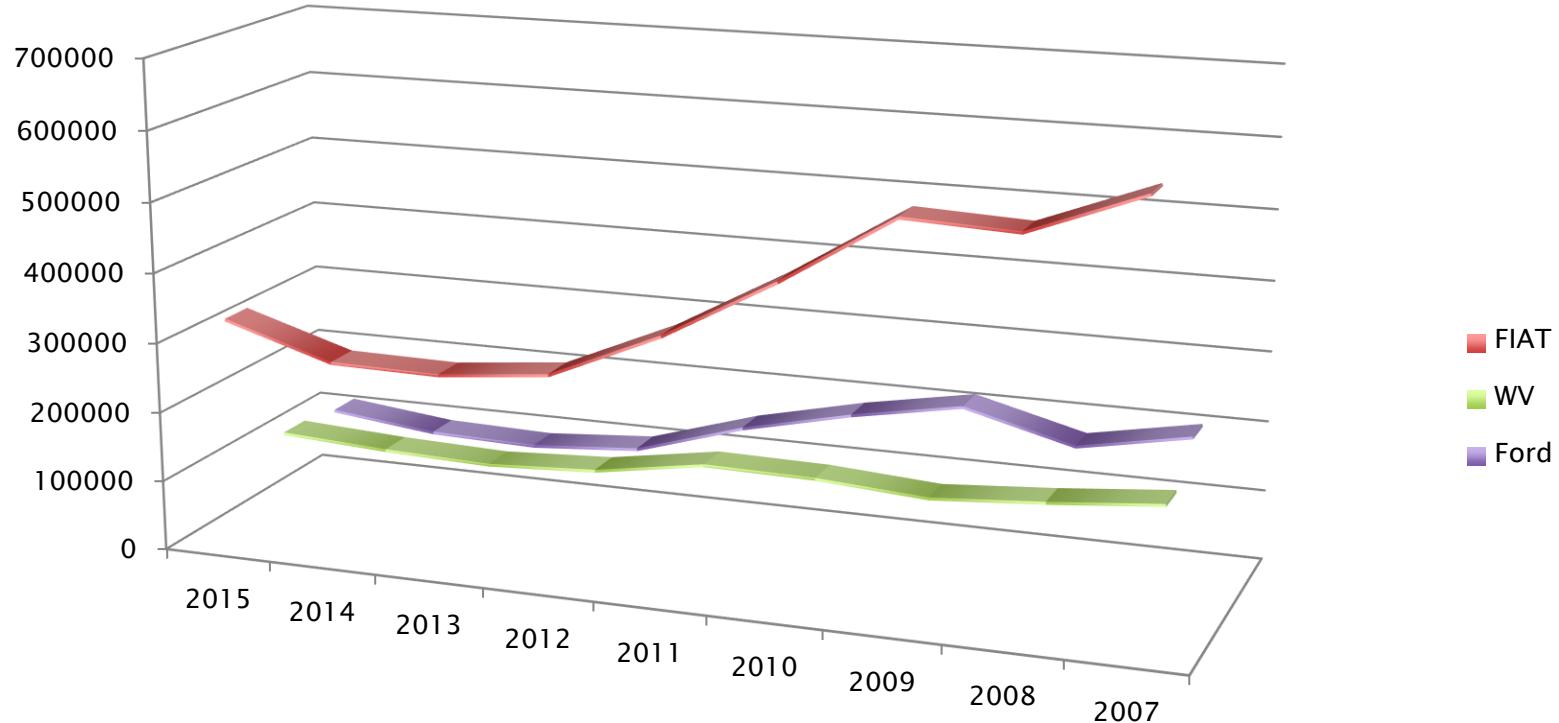


Encoding 3D → 2D



Decorative 3D

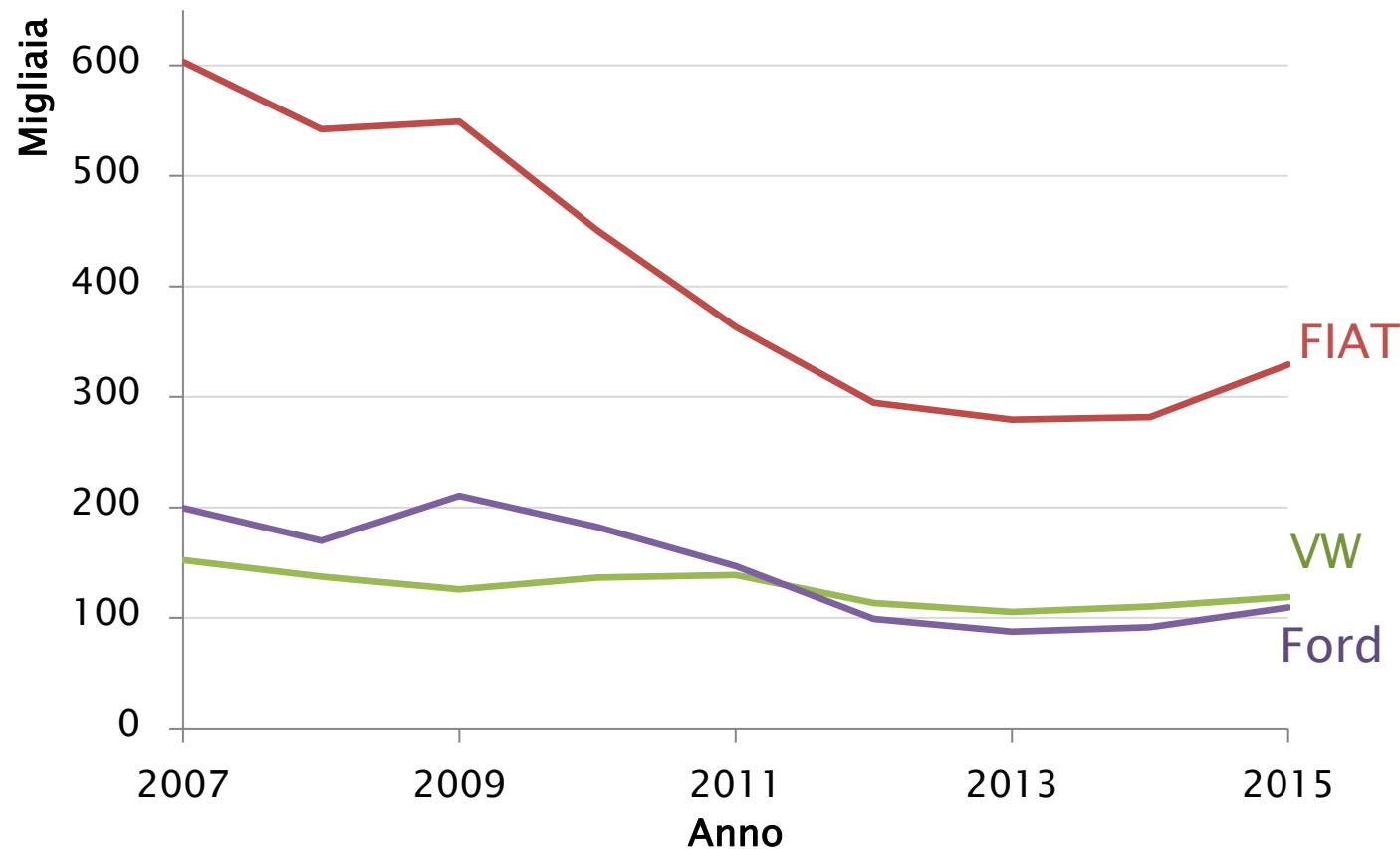
Immatricol.



Decorative 3D → 2D

Immatricolazioni auto per marchio sul mercato italiano

Immatricol.



References

- Stephen Few, 2004. Show me the numbers. Analytics Press.
 - ◆ <http://www.perceptualedge.com/blog/>
- Edward R. Tufte, 1983. The Visual Display of Quantitative Information. Graphics Press.

References

- Wilkinson, L. (2006). *The grammar of graphics*. Springer Science & Business Media.
- Wickham, H. (2010). A layered grammar of graphics. *Journal of Computational and Graphical Statistics*, 19(1), 3–28.
- Visual Vocabulary
<http://ft.com/vocabulary>

References

- R.Olson. Revisiting the vaccine visualization
 - ◆ <http://www.randalolson.com/2016/03/04/revisiting-the-vaccine-visualizations/>
- Nathan Yau. 9 Ways to Visualize Proportions – A Guide
 - ◆ <http://flowingdata.com/2009/11/25/9-ways-to-visualize-proportions-a-guide/>
- M.Correll, and M.Gleicher. Error Bars Considered Harmful: Exploring Alternate Encodings for Mean and Error *IEEE Transactions on Visualization and Computer Graphics*, Dec. 2014
 - ◆ <http://graphics.cs.wisc.edu/Papers/2014/CG14/Preprint.pdf>