# Graph Construction

## Visualizzazione dell'Informazione Quantitativa

https://softeng.polito.it/courses/VIQ

SoftEng http://softeng.polito.it

Version 3.2.0
© Marco Torchiano, 2021

# Grammar of Graphics

- Theory behind graphics construction
  - Separation of data from aesthetic
  - Definition of common chart elements
  - Composition of such 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*)
- AESTHETIC: attributes (*e.g.*, *color, position*)
- GUIDE: guides (e.g., *axes*, *legends*)

# Specification for a scatter plot

- DATA: x, y, group
- TRANS: identity
- SCALE: *linear*(*dim*(1)), *linear*(*dim*(2))
- COORD:*rect*(*dim*(1, 2))
- ELEMENT: *point*()
- AESTHETIC: *position*(x*y)
- GUIDE: *axis*(*dim*(1)), *axis*(*dim*(2))

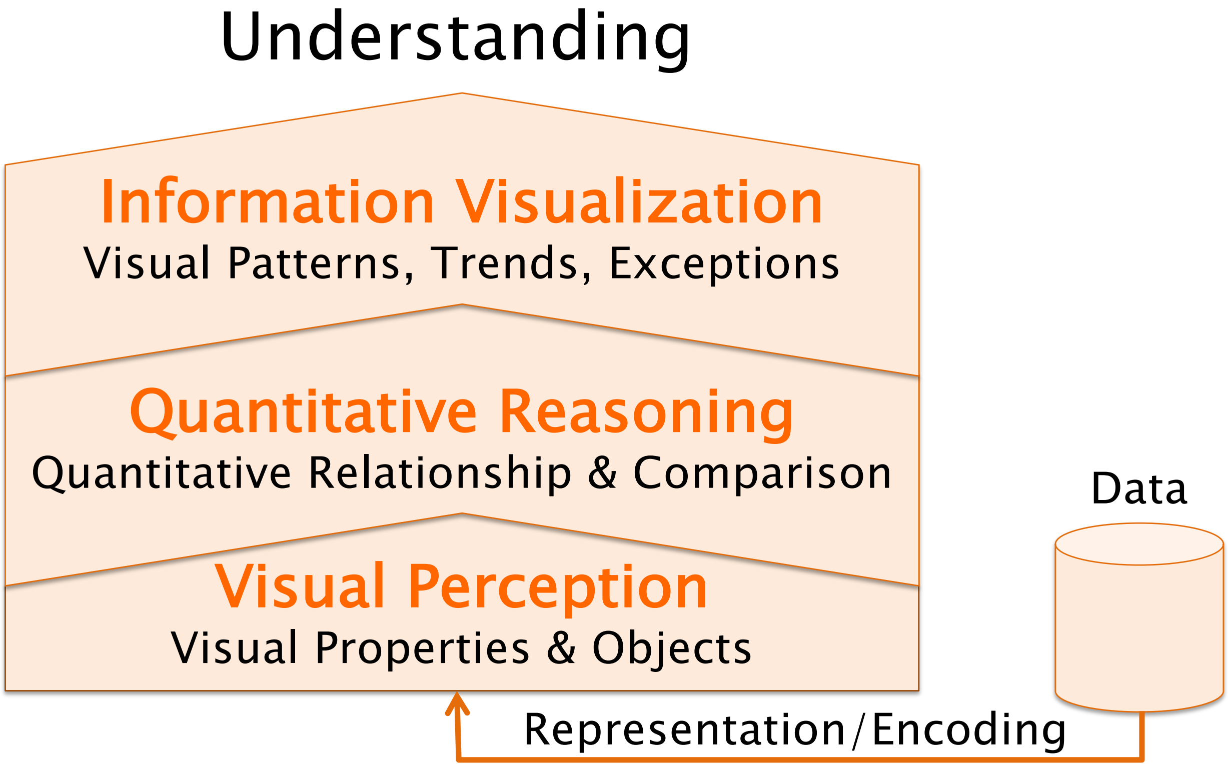
# Graph visual components

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

Visualizzazione dell'Informazione Quantitativa

# VISUAL RELATIONSHIPS

# Data Visualization


Understanding
Information Visualization
Visual Patterns, Trends, Exceptions
Quantitative Reasoning
Quantitative Relationship & Comparison
Visual Perception
Visual Properties & Objects
Data
Representation/Encoding

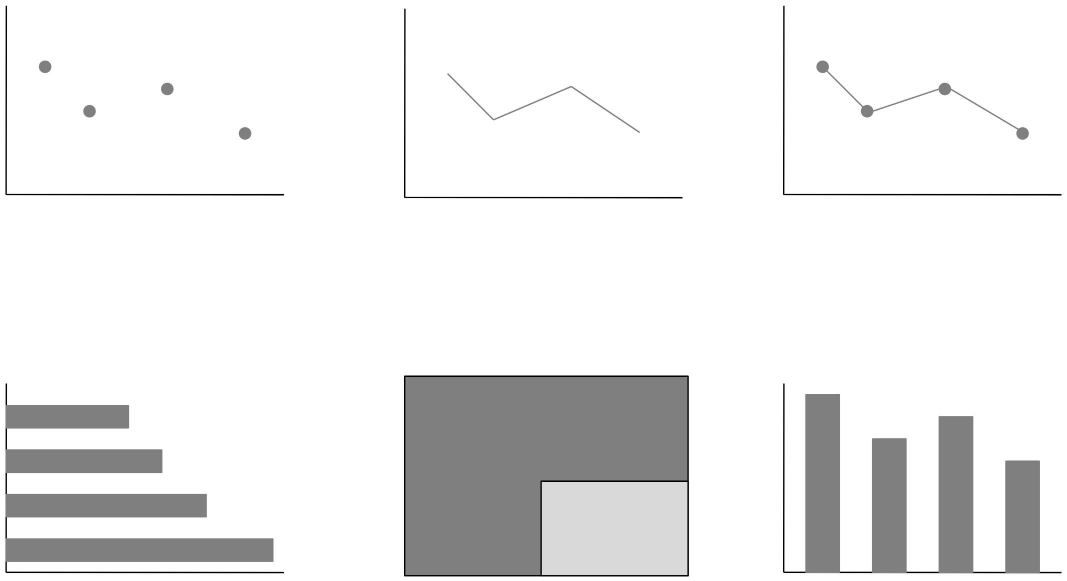
# Visual Encoding

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

# Relationships

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

# Quantitative encoding

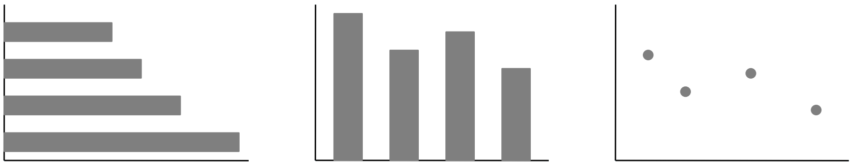


# Sample data

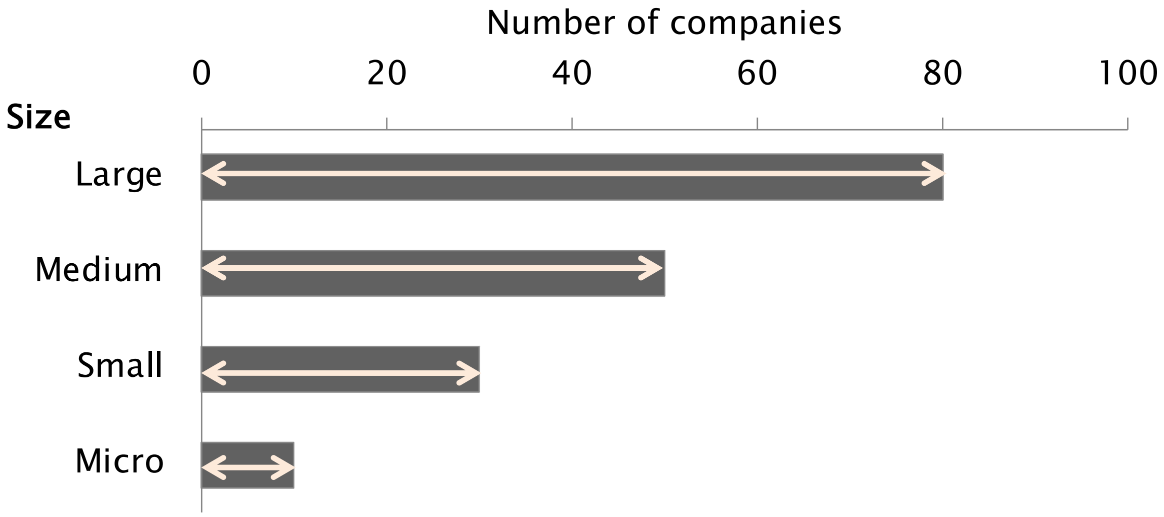
| Region | turnout (2018) | turnout (2013) |
|---|---|---|
| ABRUZZO | 75.3% | 75.9% |
| BASILICATA | 71.1% | 69.5% |
| CALABRIA | 63.6% | 63.1% |
| CAMPANIA | 68.2% | 67.9% |
| EMILIA-ROMAGNA | 78.3% | 82.1% |
| FRIULI-VENEZIA GIULIA | 75.1% | 77.2% |
| LAZIO | 72.6% | 77.5% |
| LIGURIA | 72.0% | 75.1% |
| LOMBARDIA | 76.8% | 79.6% |
| MARCHE | 77.3% | 79.8% |
| MOLISE | 71.6% | 78.1% |
| PIEMONTE | 75.2% | 77.3% |
| PUGLIA | 69.1% | 69.9% |
| SARDEGNA | 65.5% | 68.5% |
| SICILIA | 62.8% | 64.6% |
| TOSCANA | 77.5% | 79.2% |
| TRENTINO-ALTO ADIGE | 74.3% | 81.0% |
| UMBRIA | 78.2% | 79.5% |
| VALLE D'AOSTA | 72.3% | 77.0% |
| VENETO | 78.7% | 81.8% |

# Nominal comparison

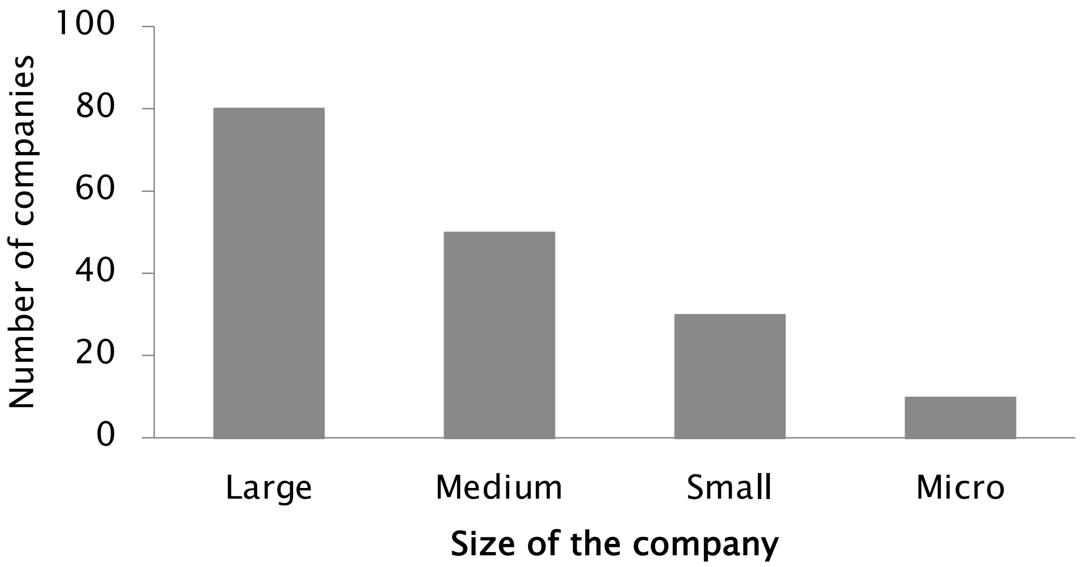
- Compare quantitative values corresponding to categorical levels
  - Small differences are difficult to see
    - Non zero-based scale can emphasize
  - Dot plots can be used for small differences
    - They do not require zero based scale



# Line length – Bars chart



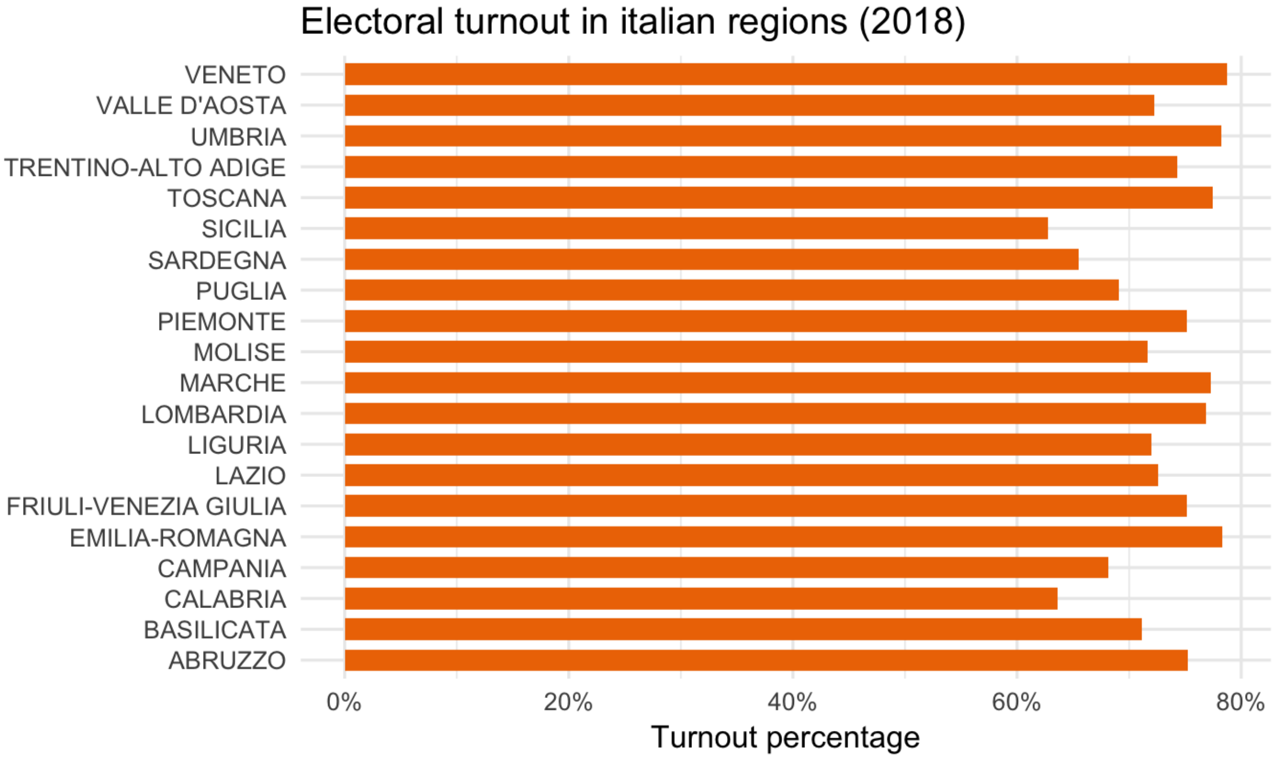
# Vertical Bars (aka Columns)



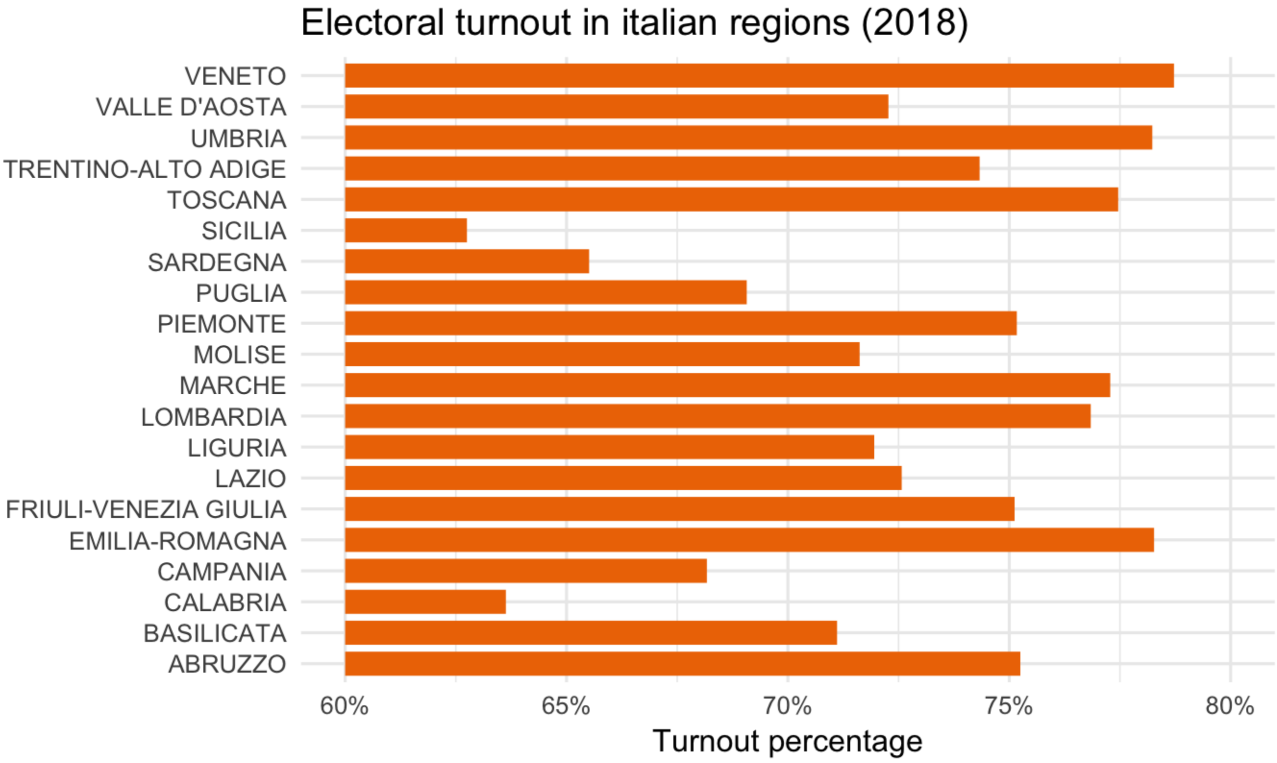
# Bar charts

- Categorical values are encoded as position along an axis
- Quantitative values are encoded only as length of the bars
  - The axis is a supporting element
- Width of bars plays no role
  - Bars are just very thick lines
- Bars require a zero-based scale
  - See: Lie factor!

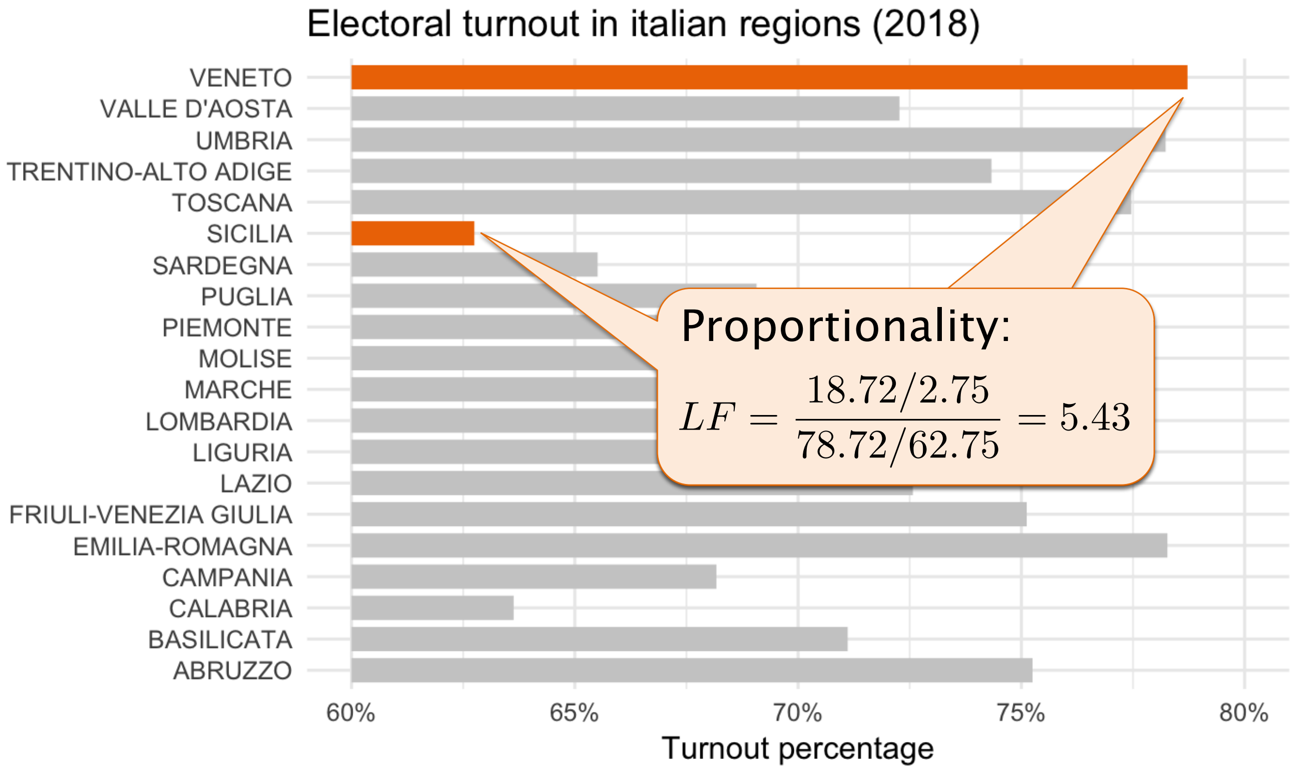
# Comparison – Barplot



# Barplot (non zero based scale)



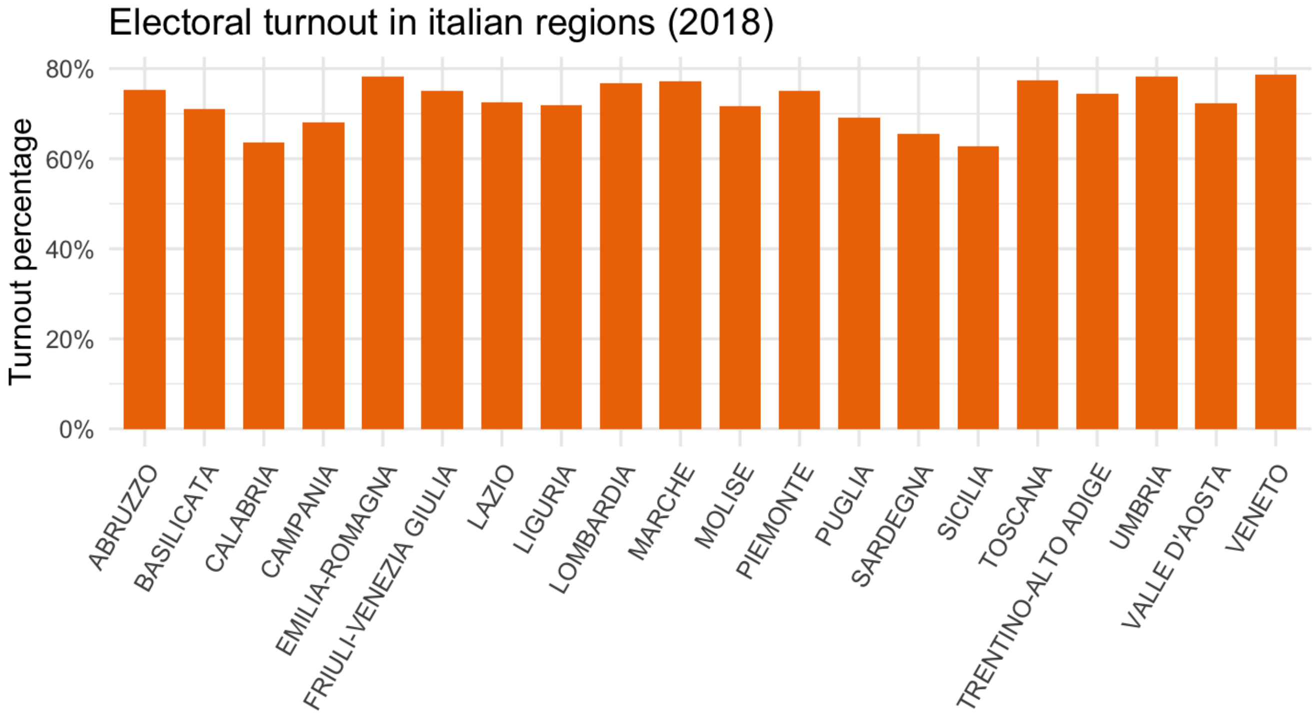
# Barplot (non zero based scale)



Proportionality:

$$LF = \frac{18.72/2.75}{78.72/62.75} = 5.43$$

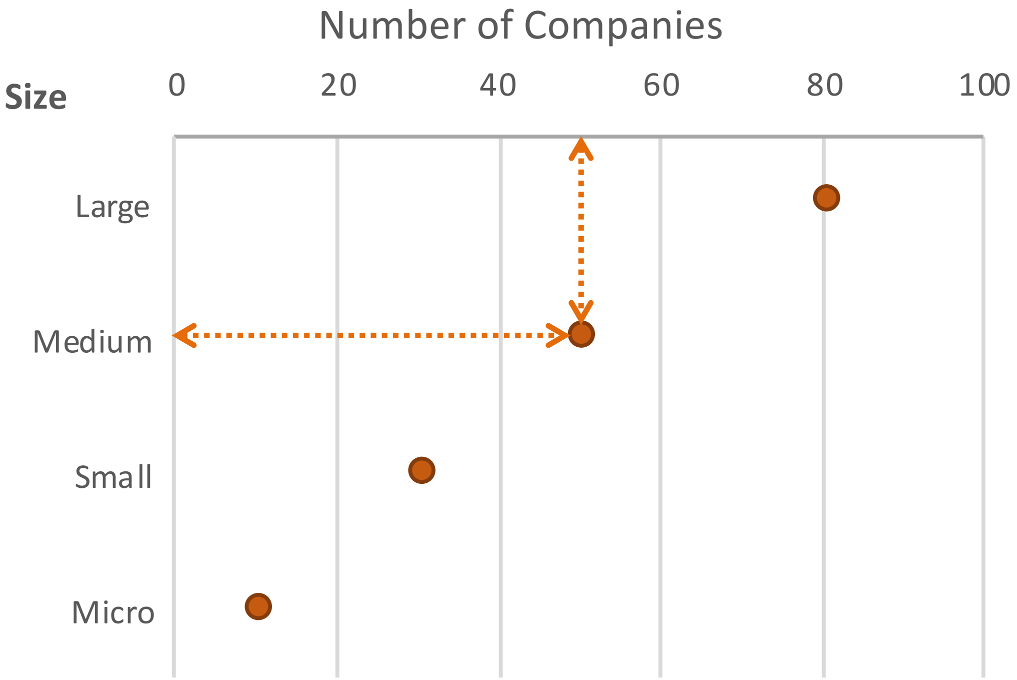
# Barplot vertical labels



# Bars Guidelines

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

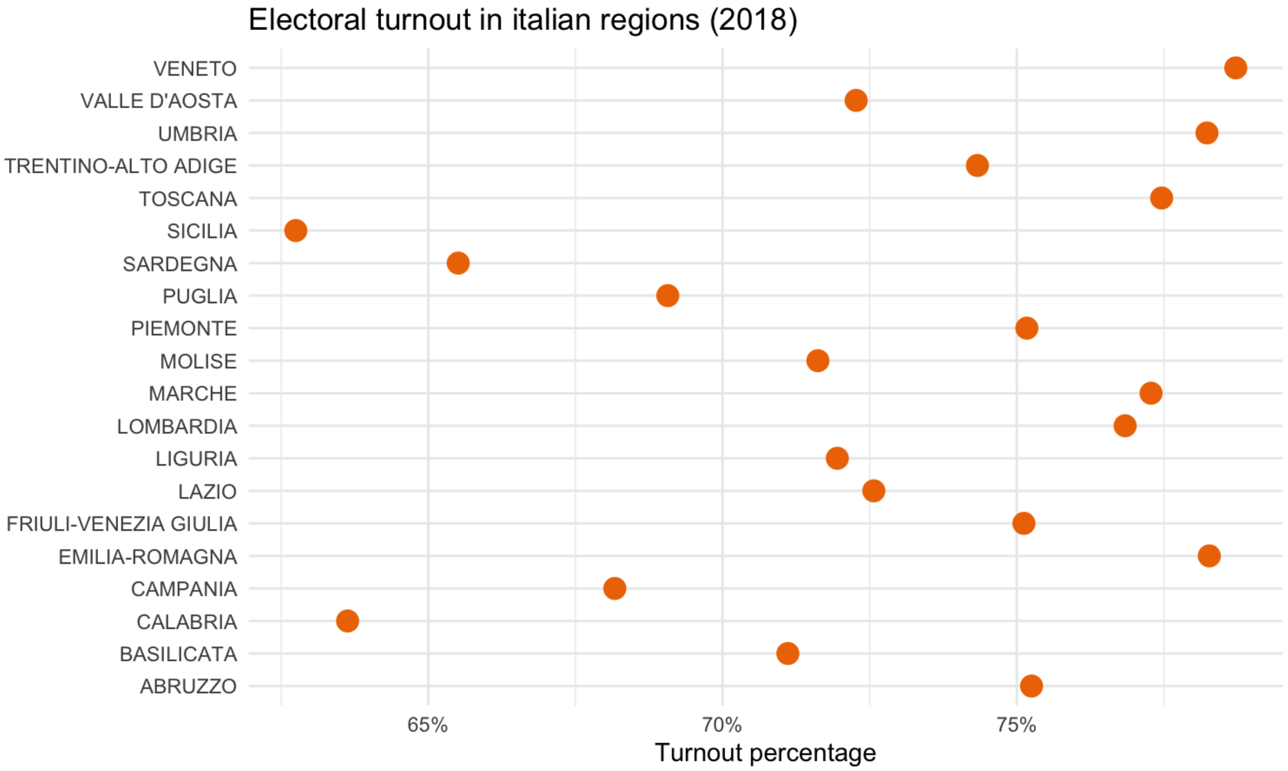
# Position – Dots plot



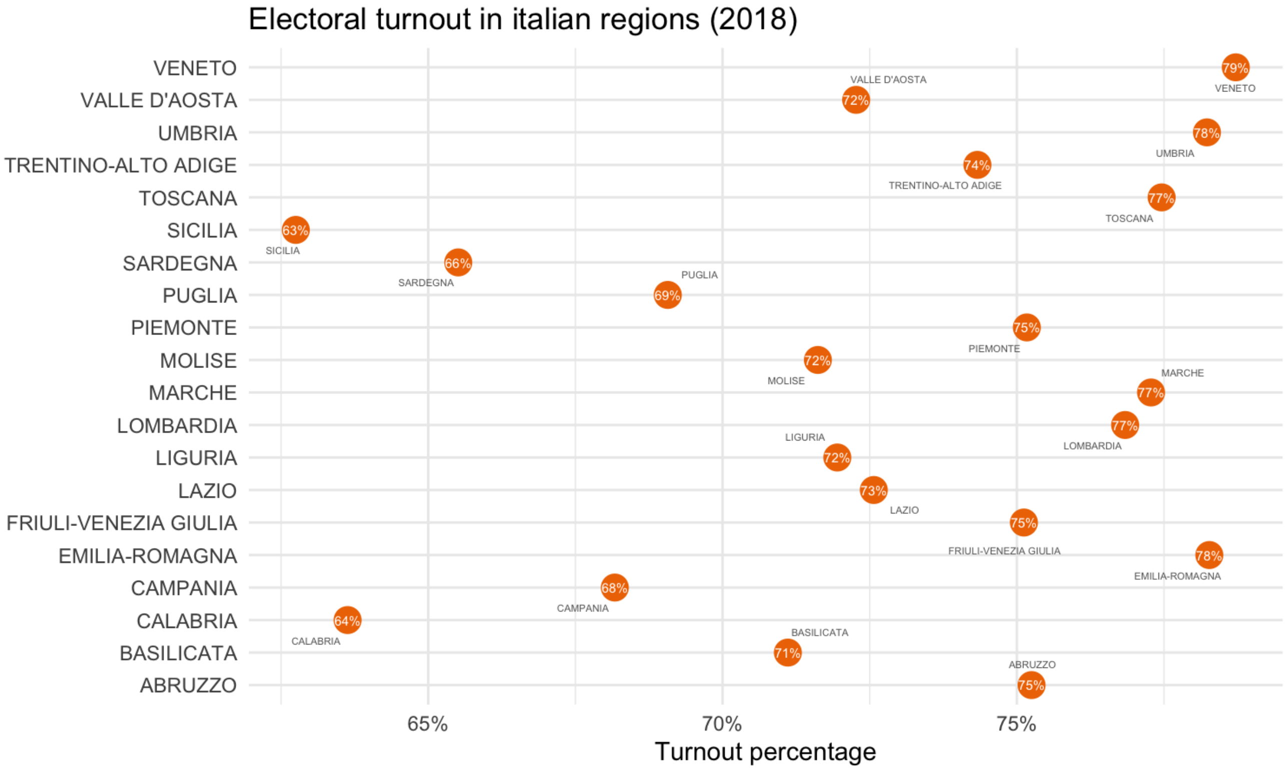
# Dot plots

- Categorical values are encoded as position along an axis
- Quantitative values are encoded as position along an axis
  - There is no need to have a zero based axis range

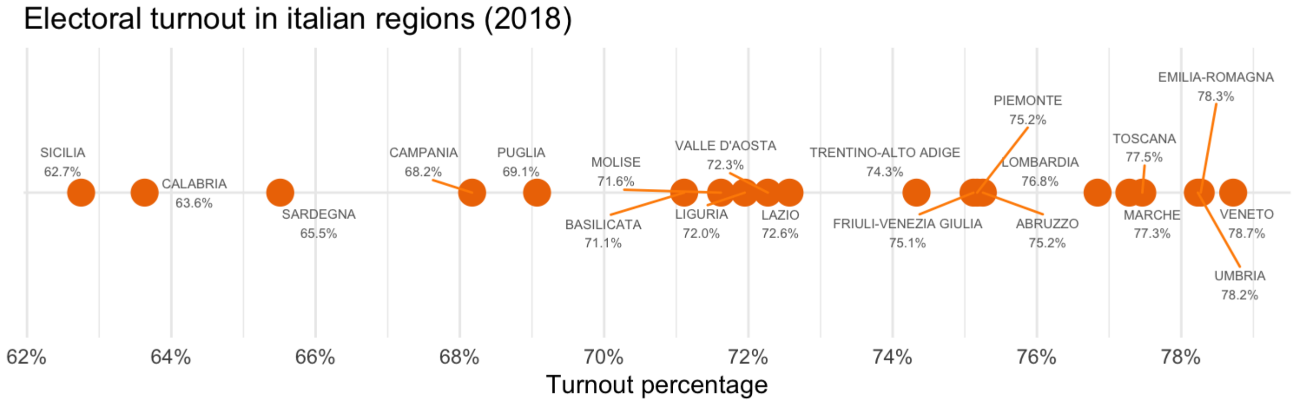
# Comparison – Dot plot



# Comparison – Dot plot



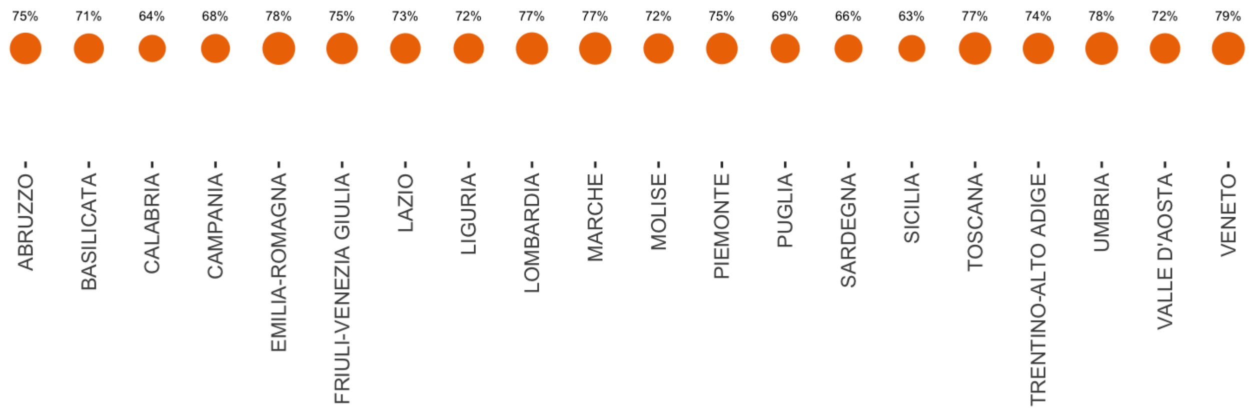
# Comparison – Strip plot



# Comparison – Area – Bubbles

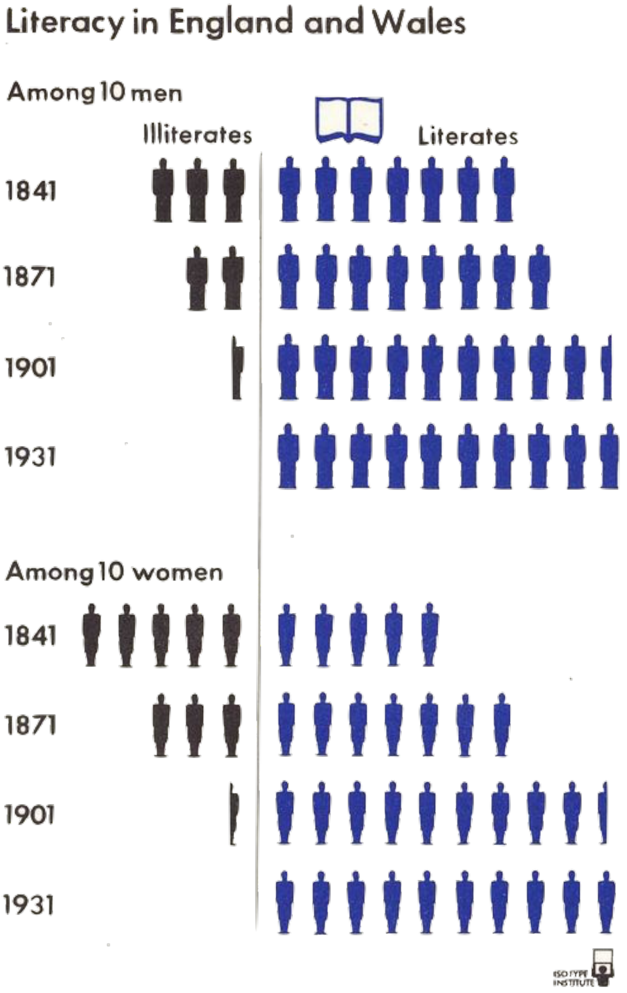
Extremely difficult to compare size

Electoral turnout in italian regions (2018)



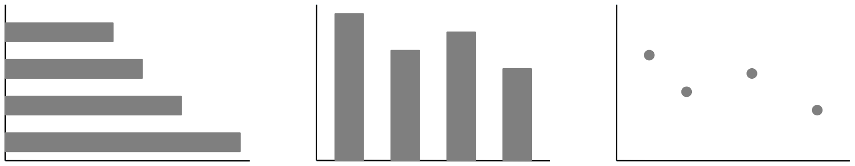
# Count – Isotype

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



# Ranking

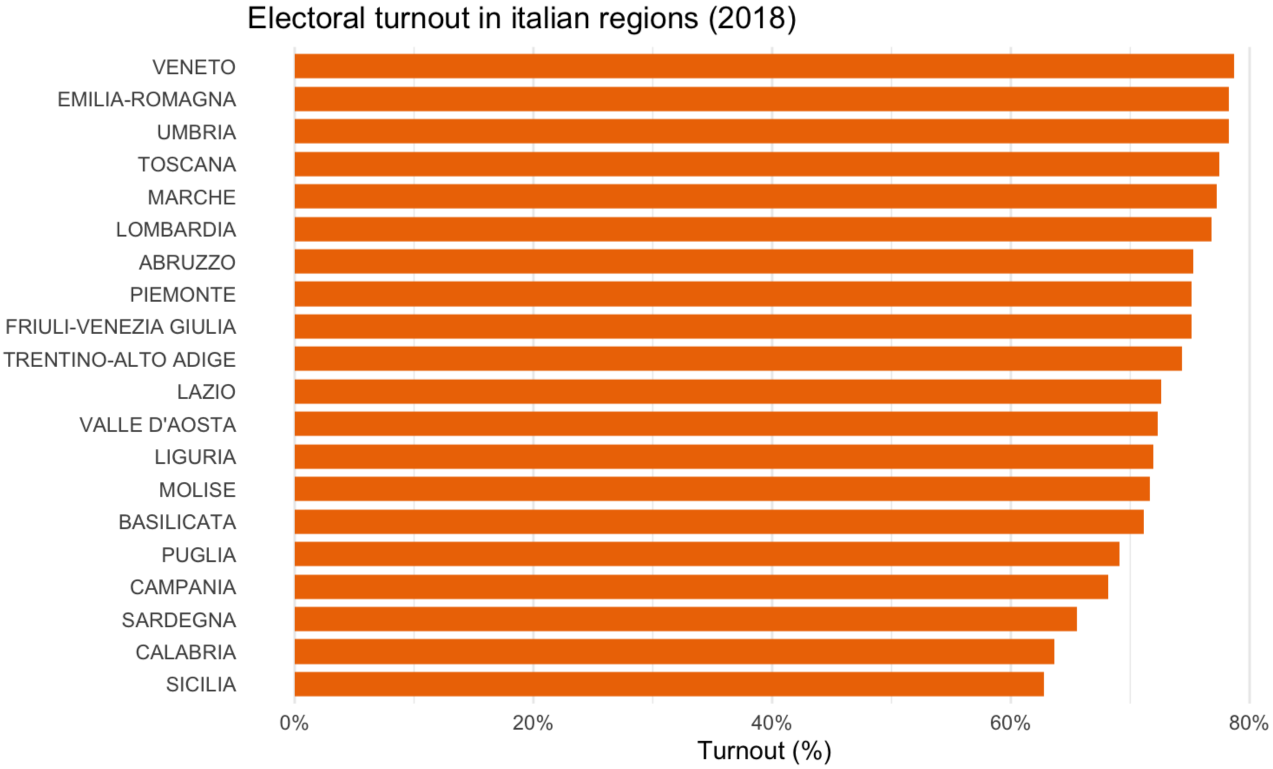
- Same type as nominal comparison
- Pay attention to order
  - Bar graphs
  - Dot plot
    - Allow non zero-based axes



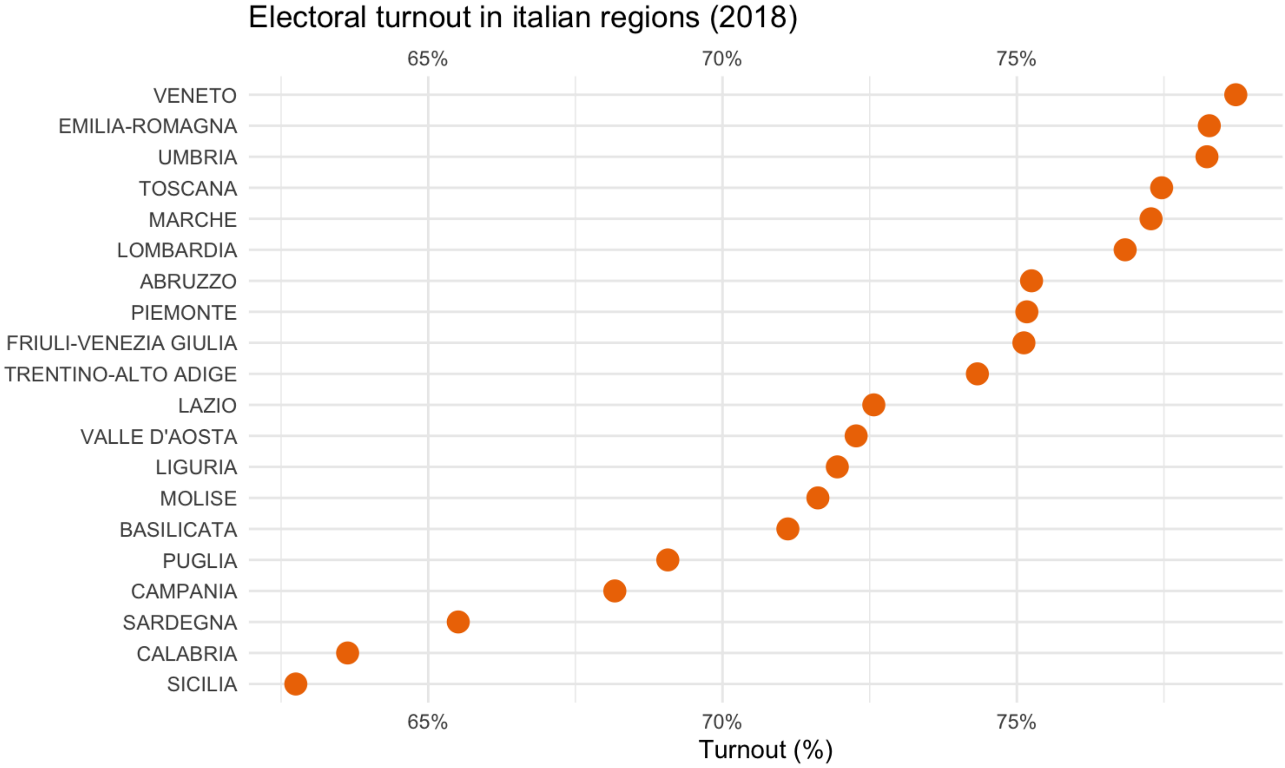
# Ranking

| Purpose | Sort order | Chart orientation |
|---|---|---|
| Highlight the highest value | Descending | H: highest on top<br>V: highest on left |
| Highlight the lowest value | Ascending | H: lowest on top<br>V: lowest on left |

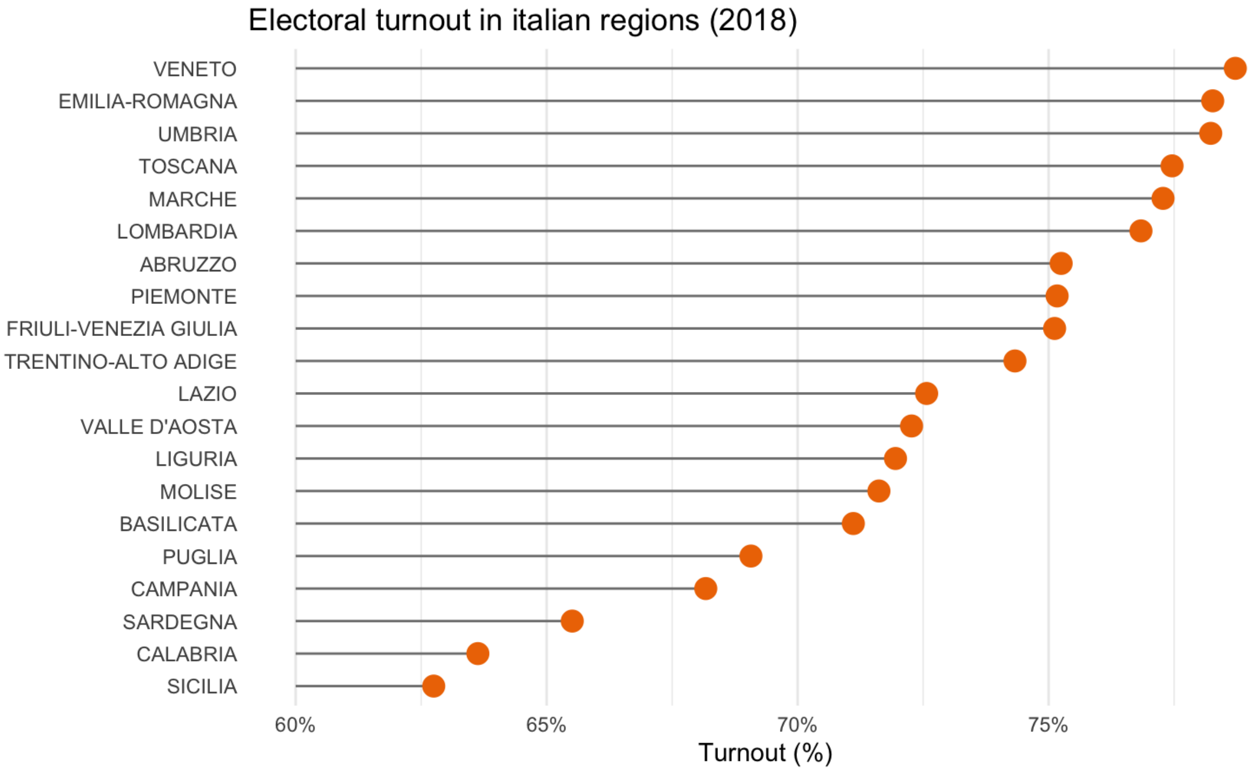
# Ranking – Barplot



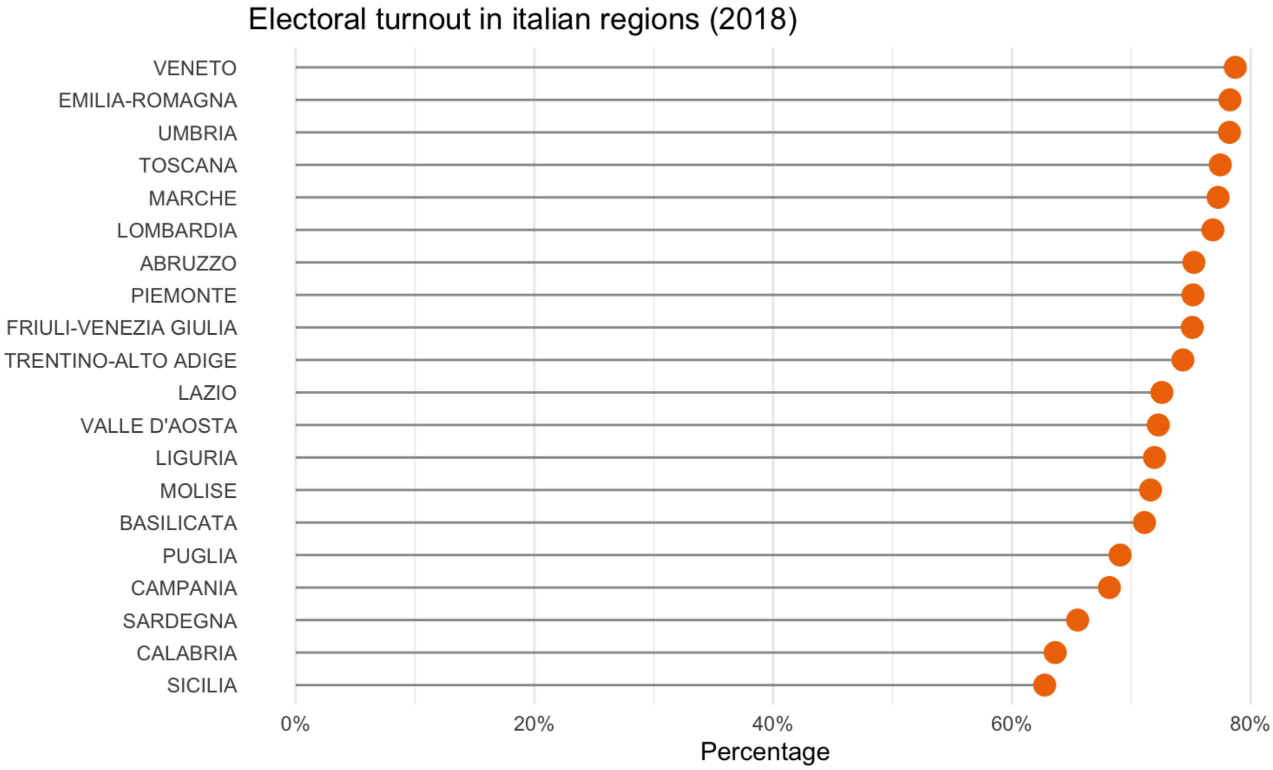
# Ranking – Dot plot



# Lollypop (nonzero based scale)

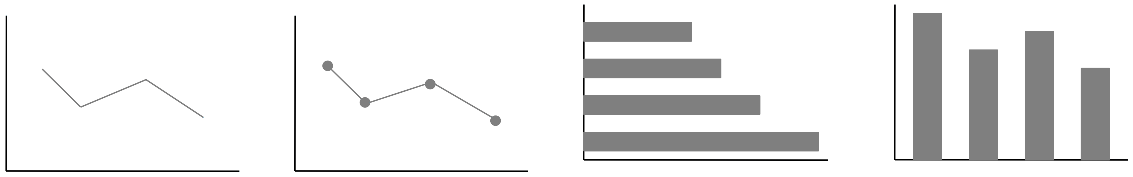


# Lollypop (zero based scale)

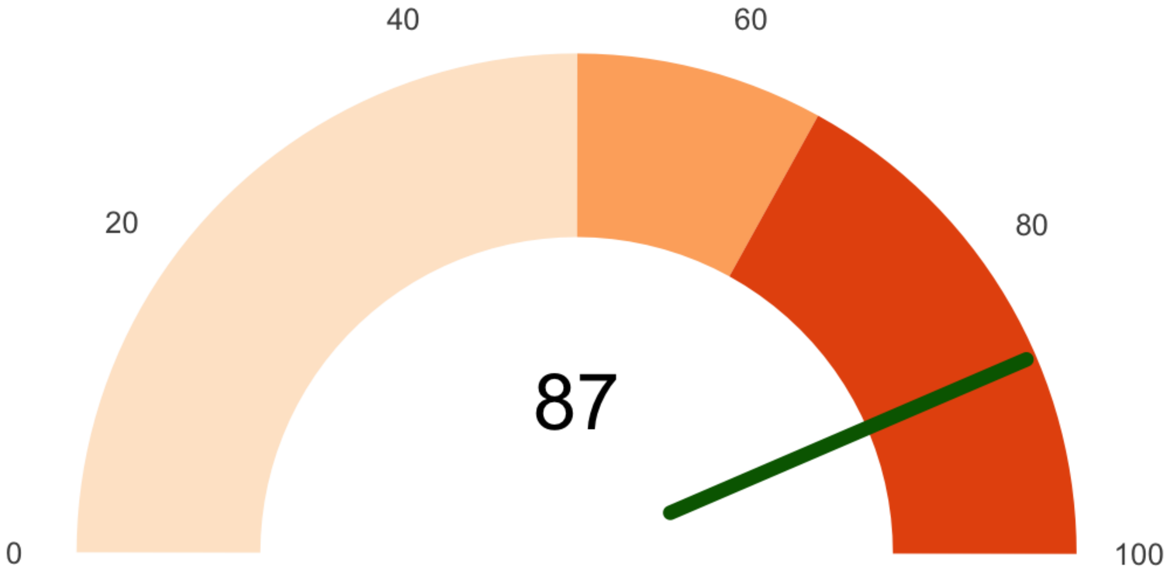


# Deviation

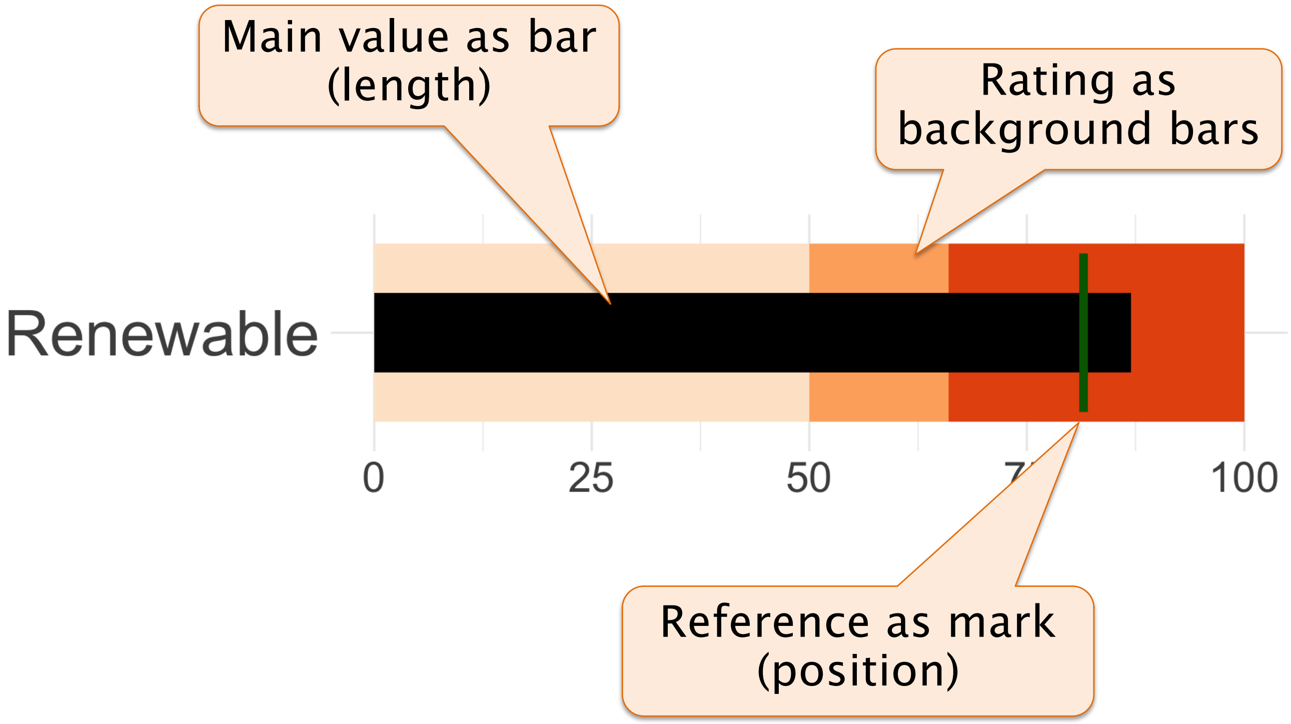
- To what degree one or more sets of values differ in relation to primary values.
  - Points (dots)
  - Gauge
  - Bars
  - Bullet



# Angle + Position – Gauge



# Length+Position– Bullet Graph



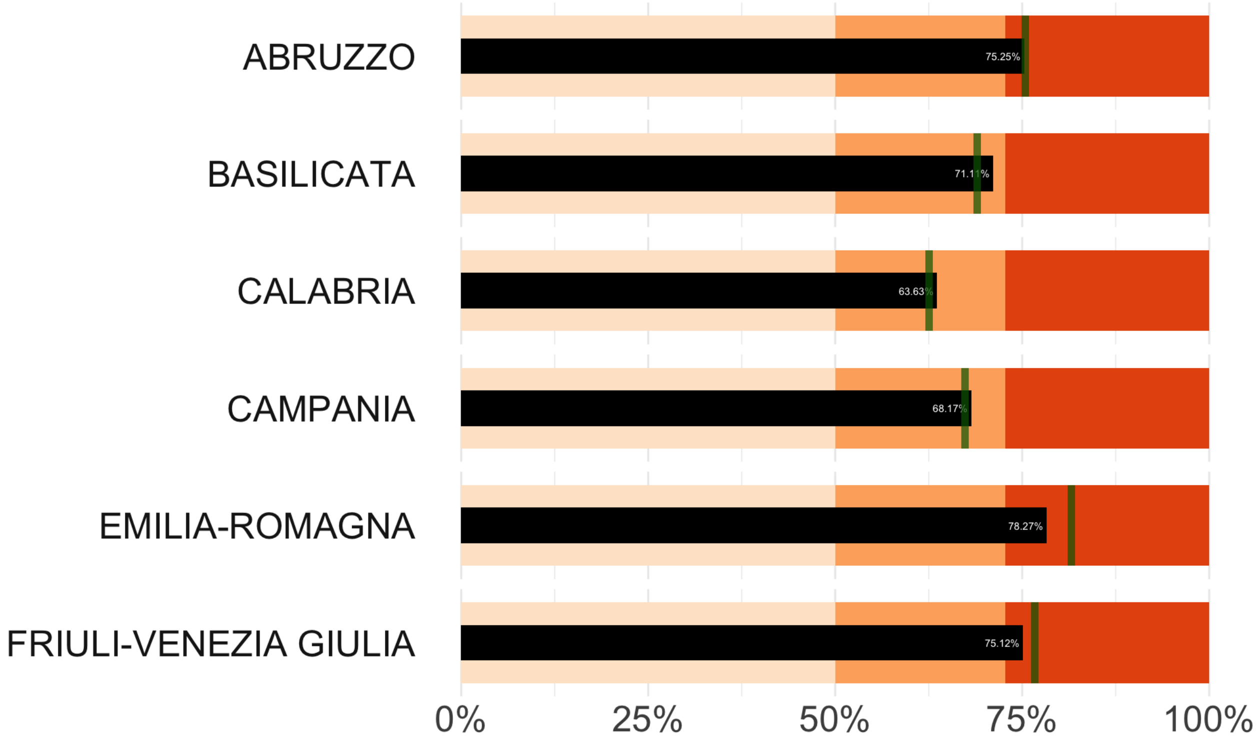
Main value as bar (length)

Rating as background bars

Reference as mark (position)

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

# Length+Position– Bullet Graph

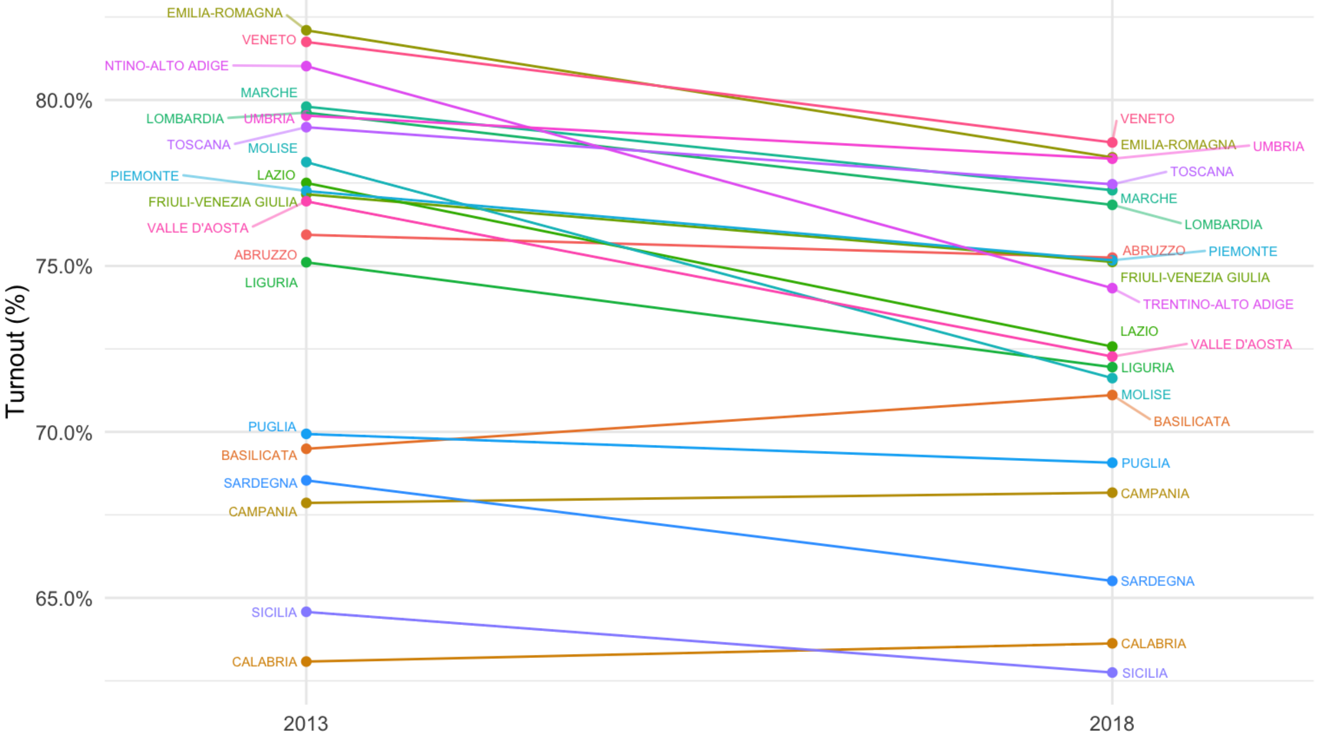


# Pre-post variation

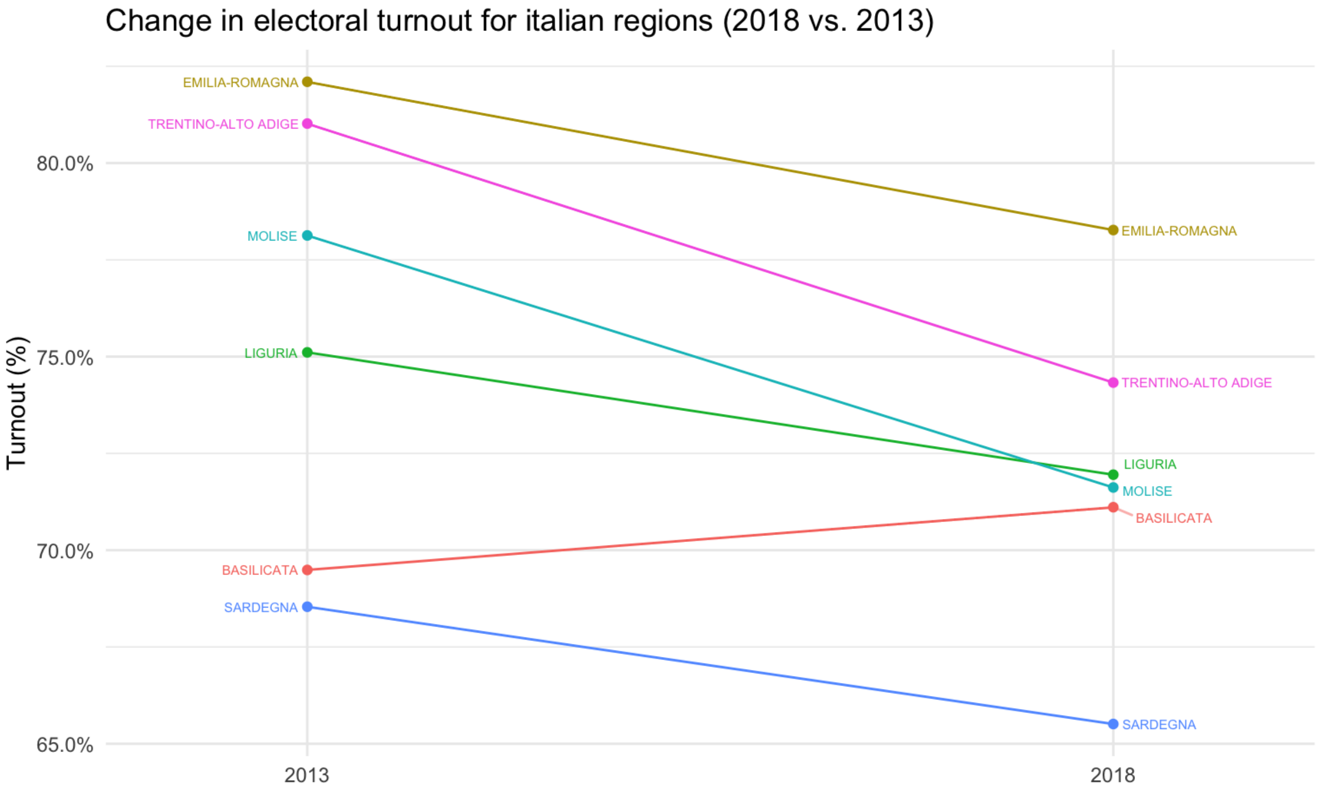
- Comparing several categorical values typically two conditions
  - Pre vs. post
  - With vs. without
  - …

# Slope chart

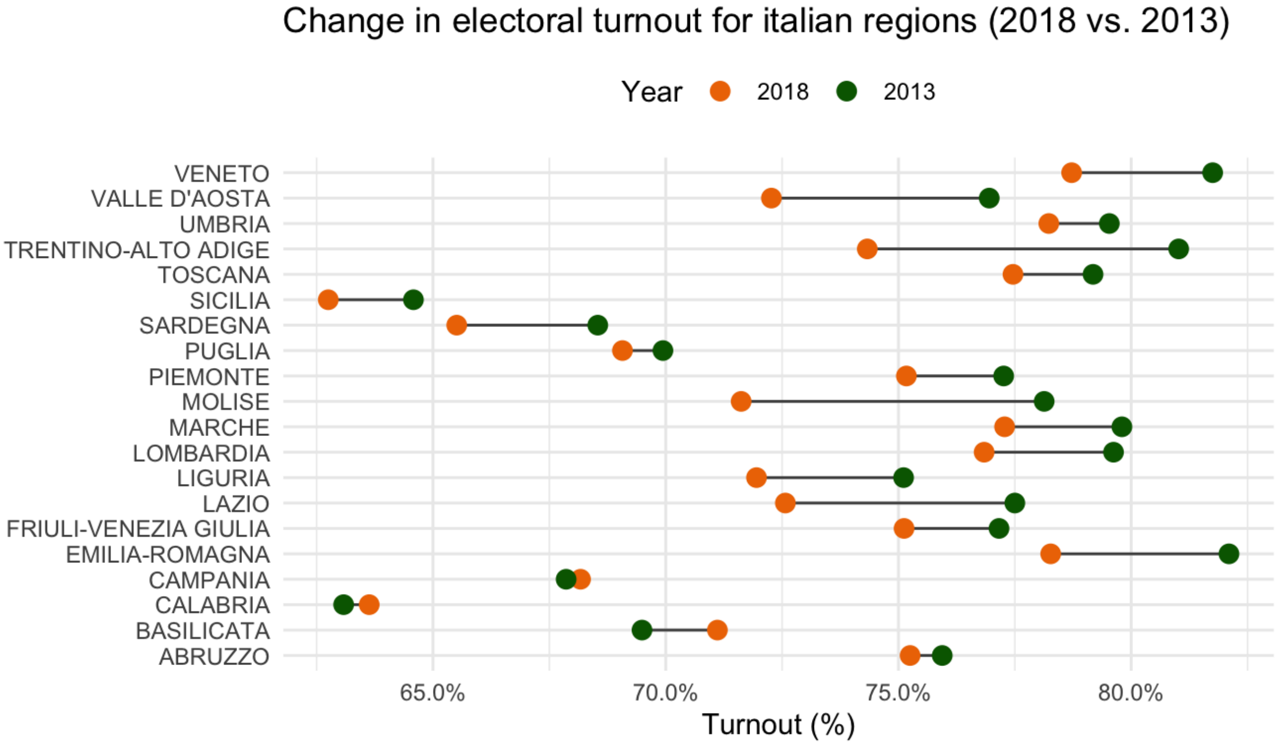
Change in electoral turnout for italian regions (2018 vs. 2013)



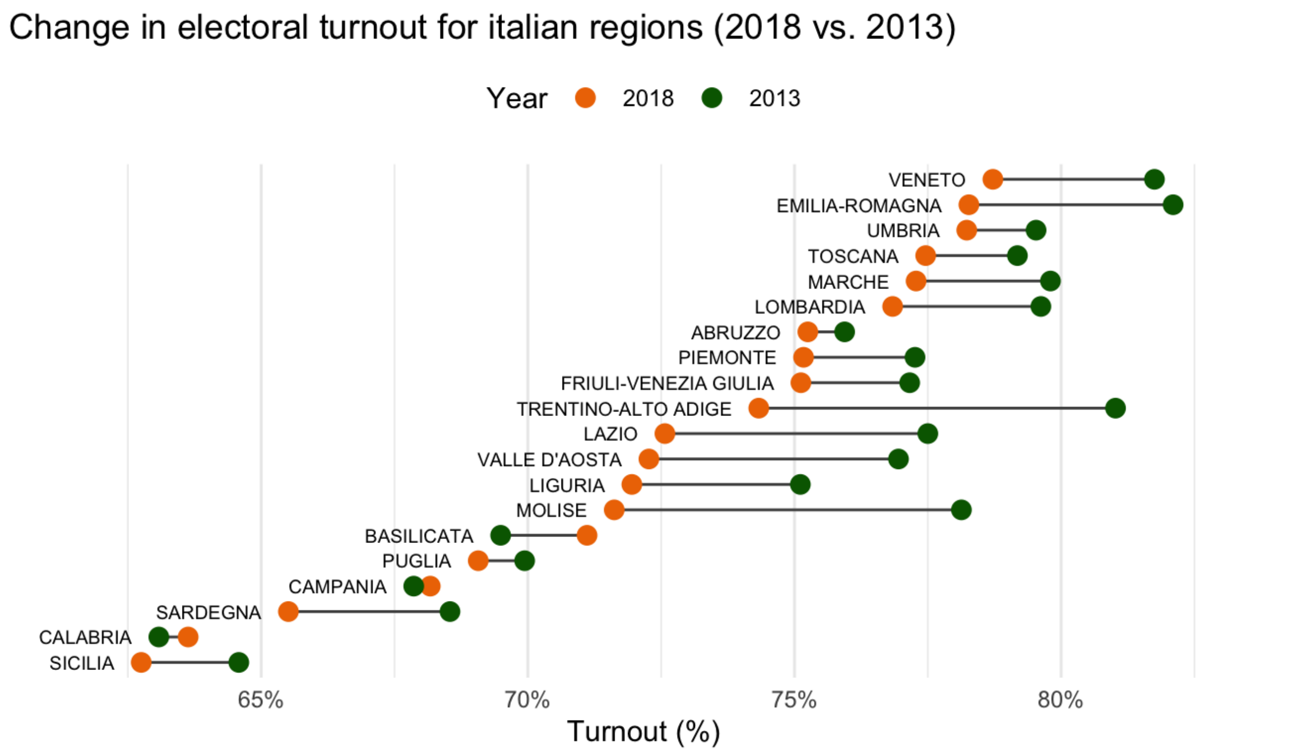
# Slope chart



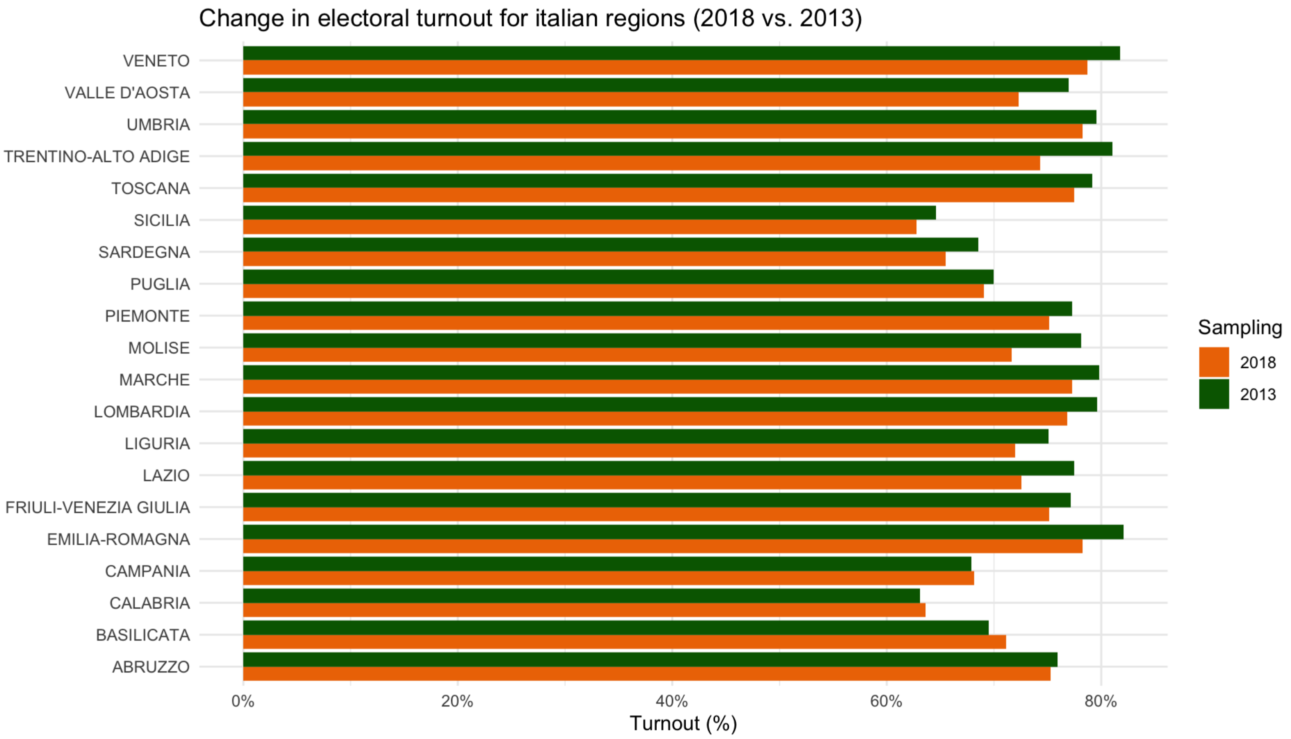
# Dumbbell plot



# Dumbbell plot (sorted)



# Clustered bars

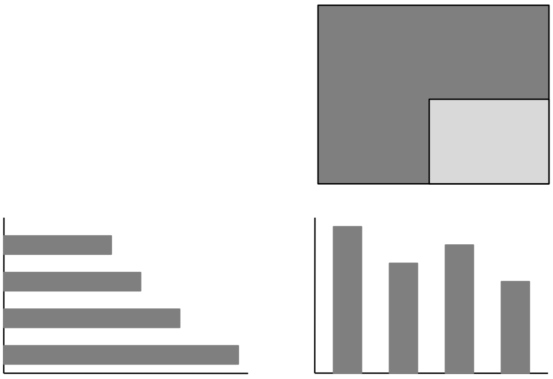


# Proportion (Part-to-whole)

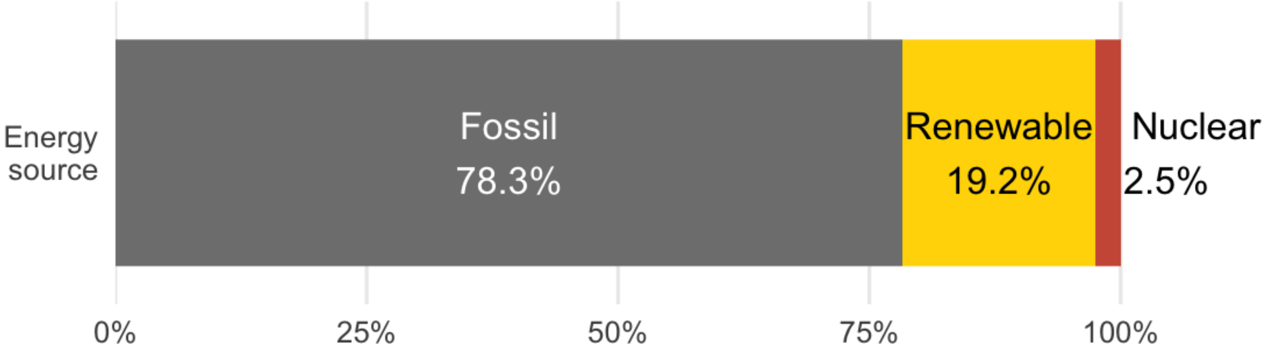
- Represent the frequency of different values within a given category
  - Be careful to use all values within the same category
- Can be used to compare frequency distribution across different categories sharing the same levels

# Proportion (Part-to-whole)

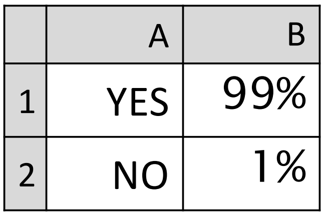
- Best unit: percentage
- Stacked bar graph
  - Difficult to read individual values
- Stacked area
- Treemap
- Gridplot
- Pie / Donut
- Marimekko



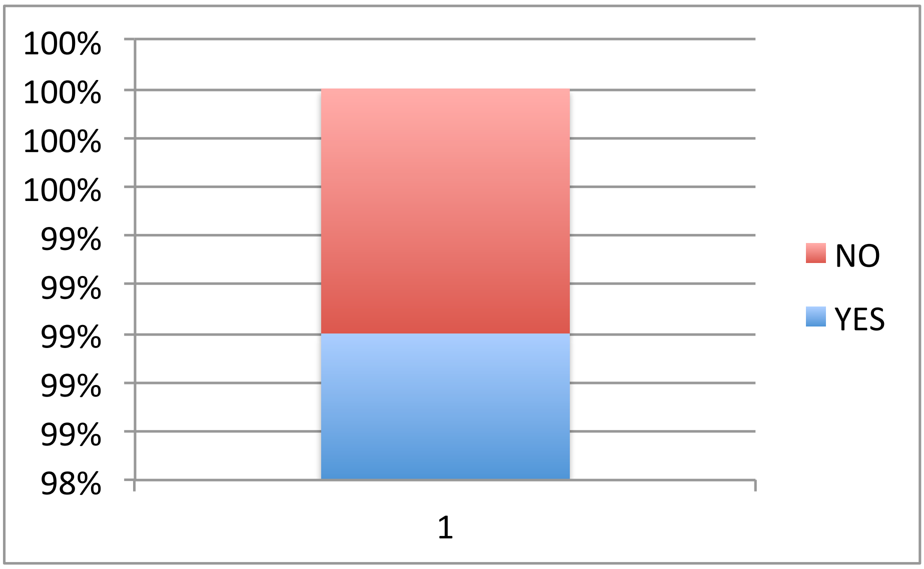
# Length – Stacked Bar



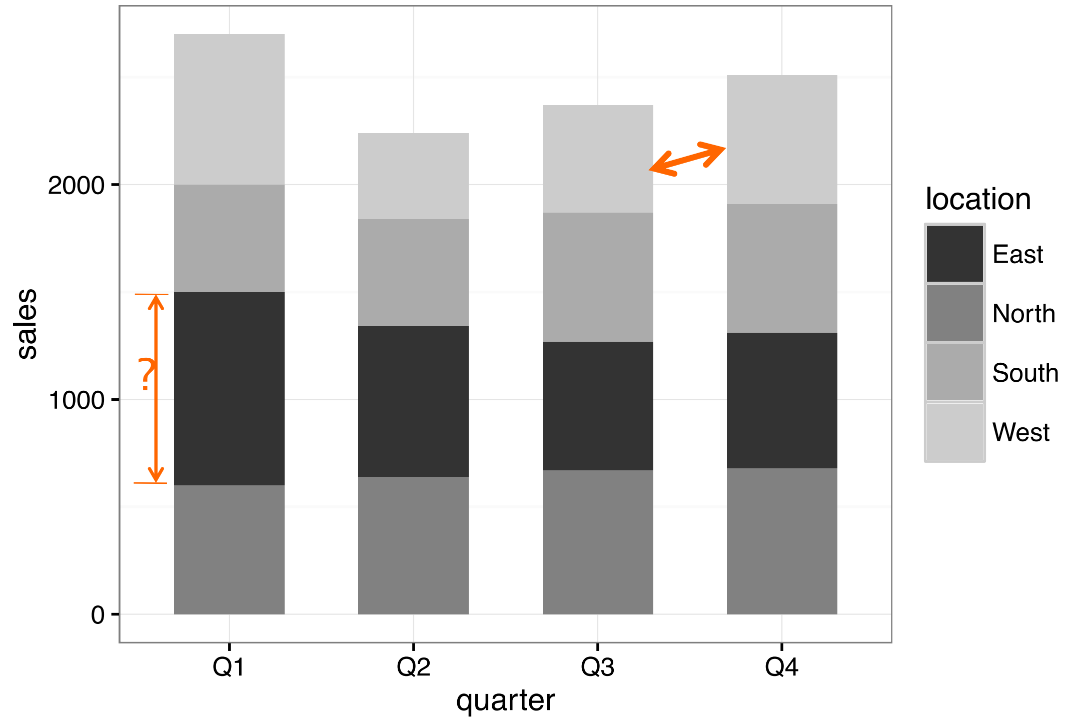
# Beware of MS-Excel Defaults



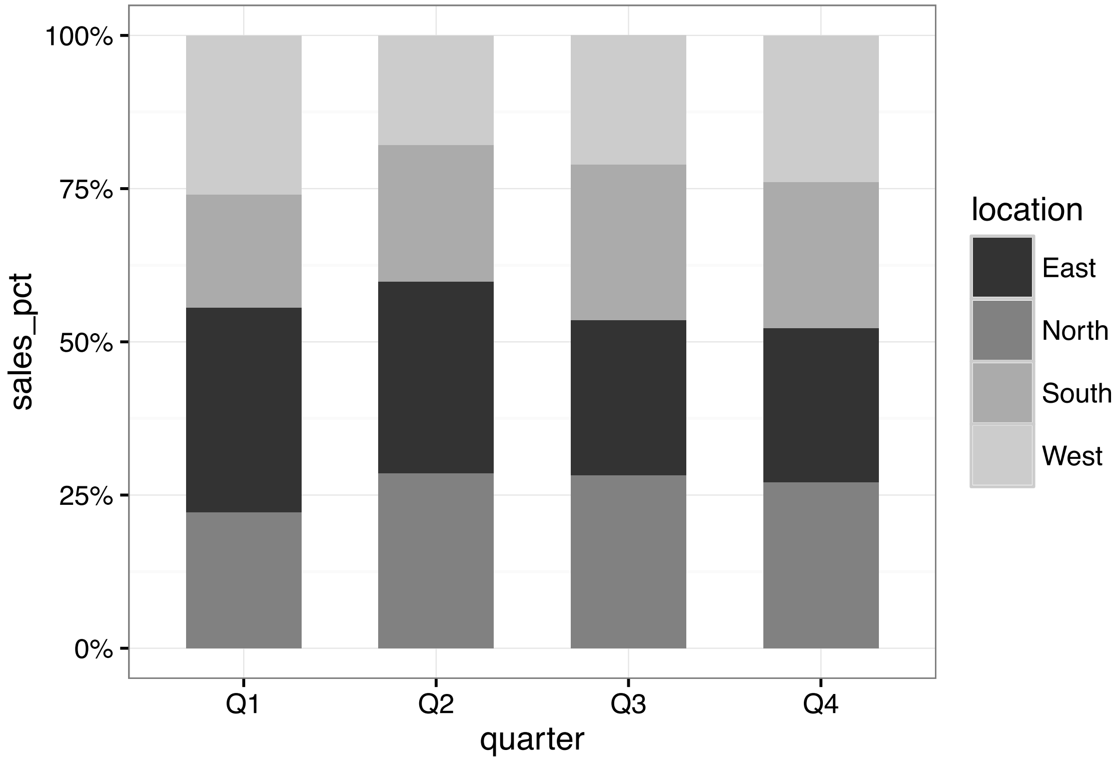
| | A | B |
|---|---|---|
| 1 | YES | 99% |
| 2 | NO | 1% |



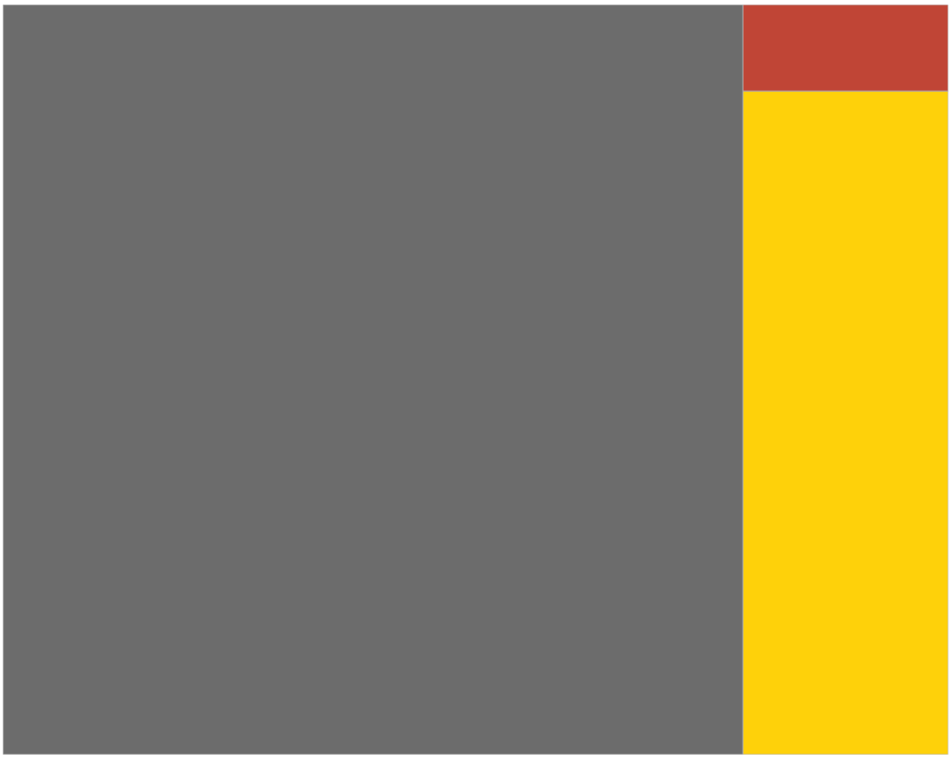
# Stacked bar graph



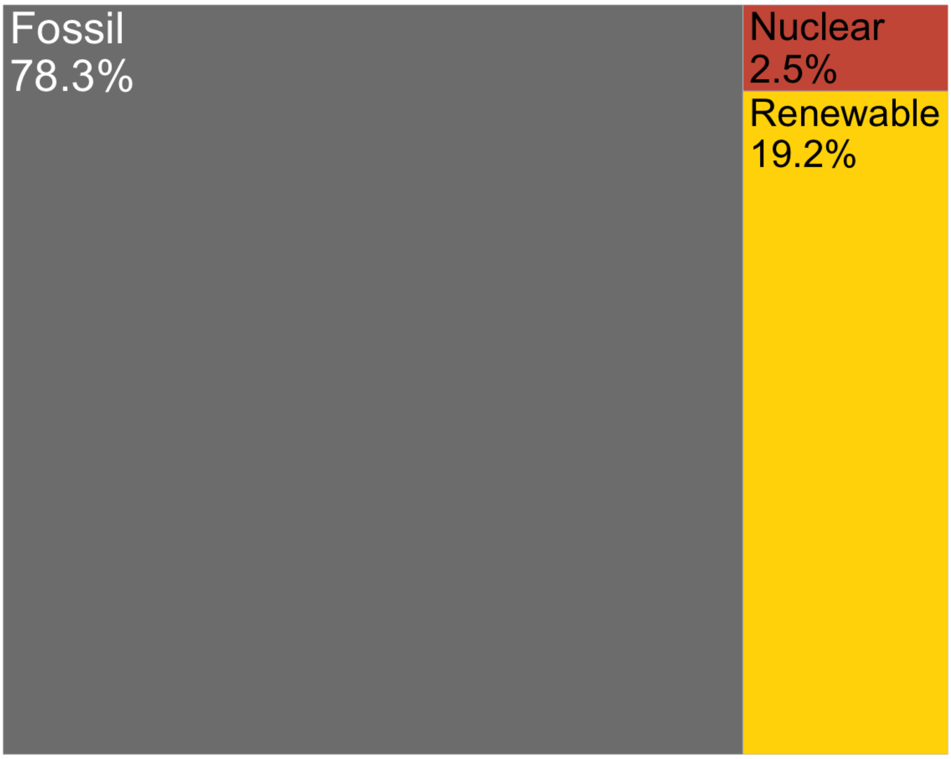
# Stacked bars w/percentage



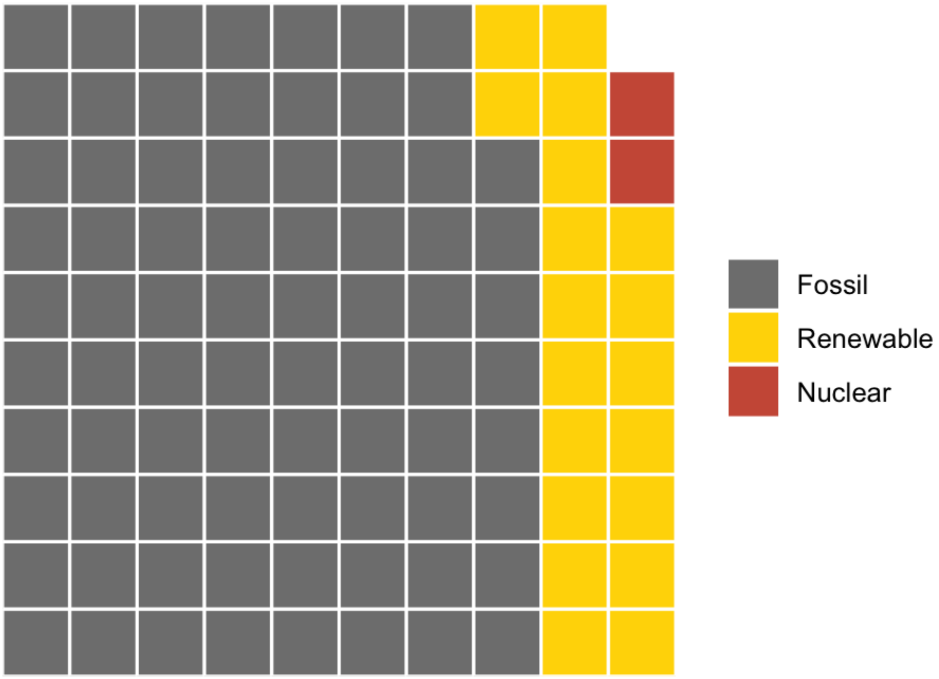
# Area – Treemap



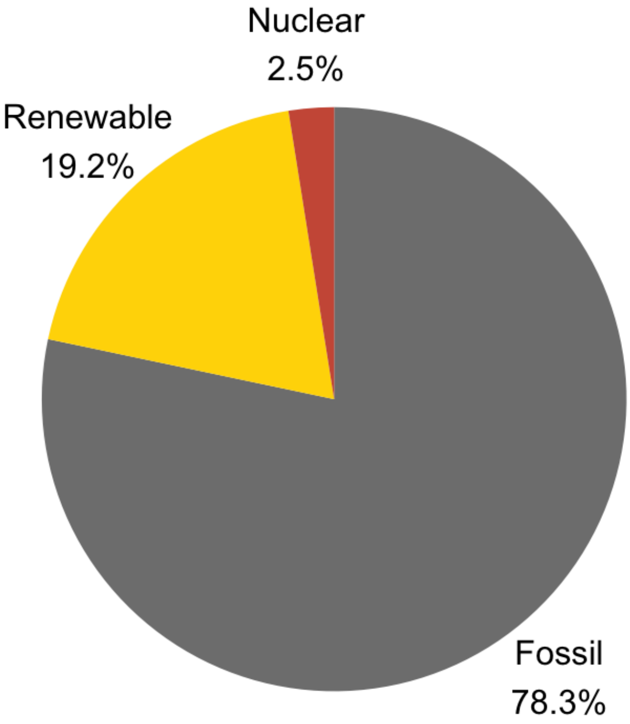
# Area – Treemap



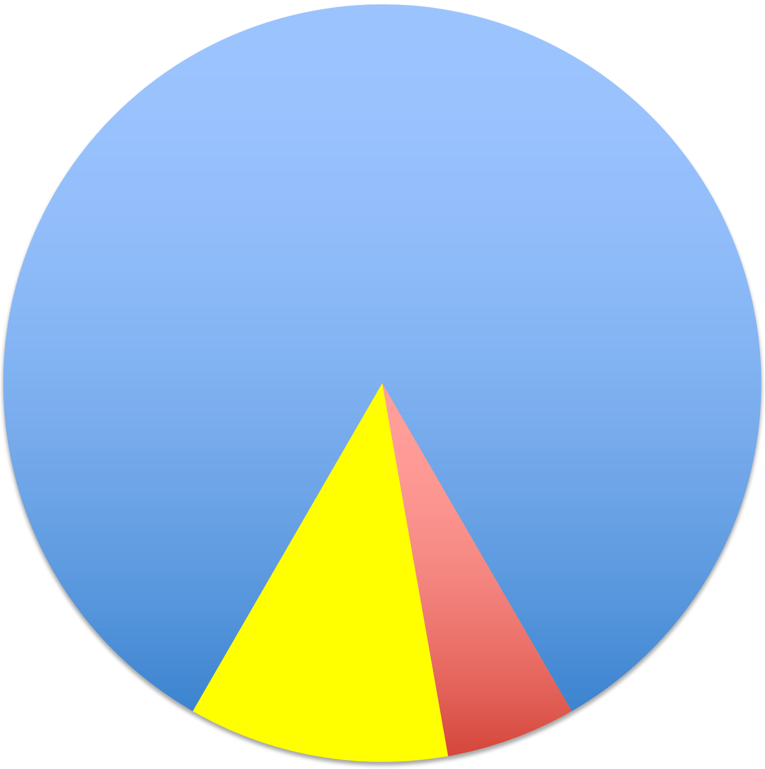
# Area + Count – Waffle / Grid



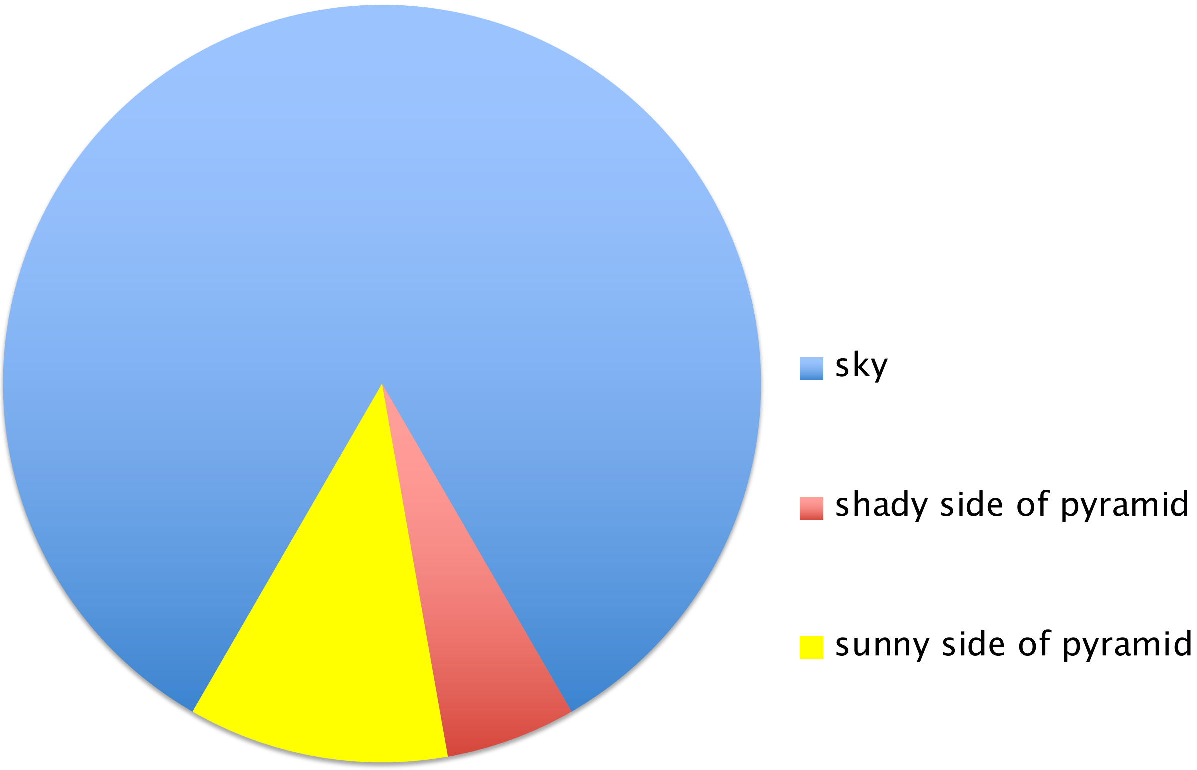
# Area + Angle – Pie Chart



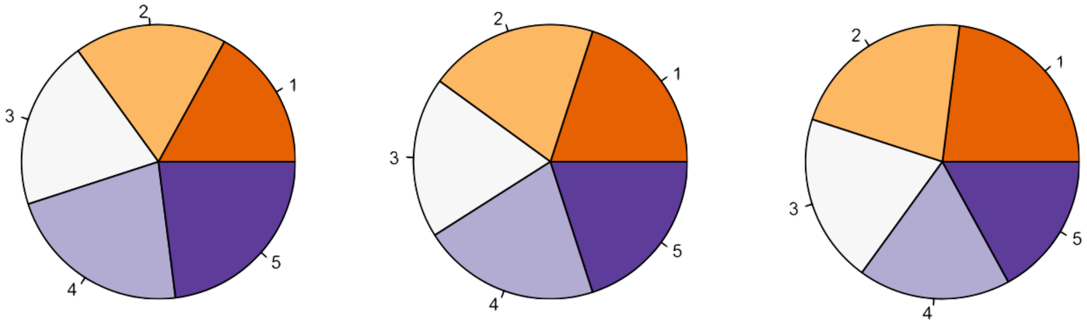
# Pie charts



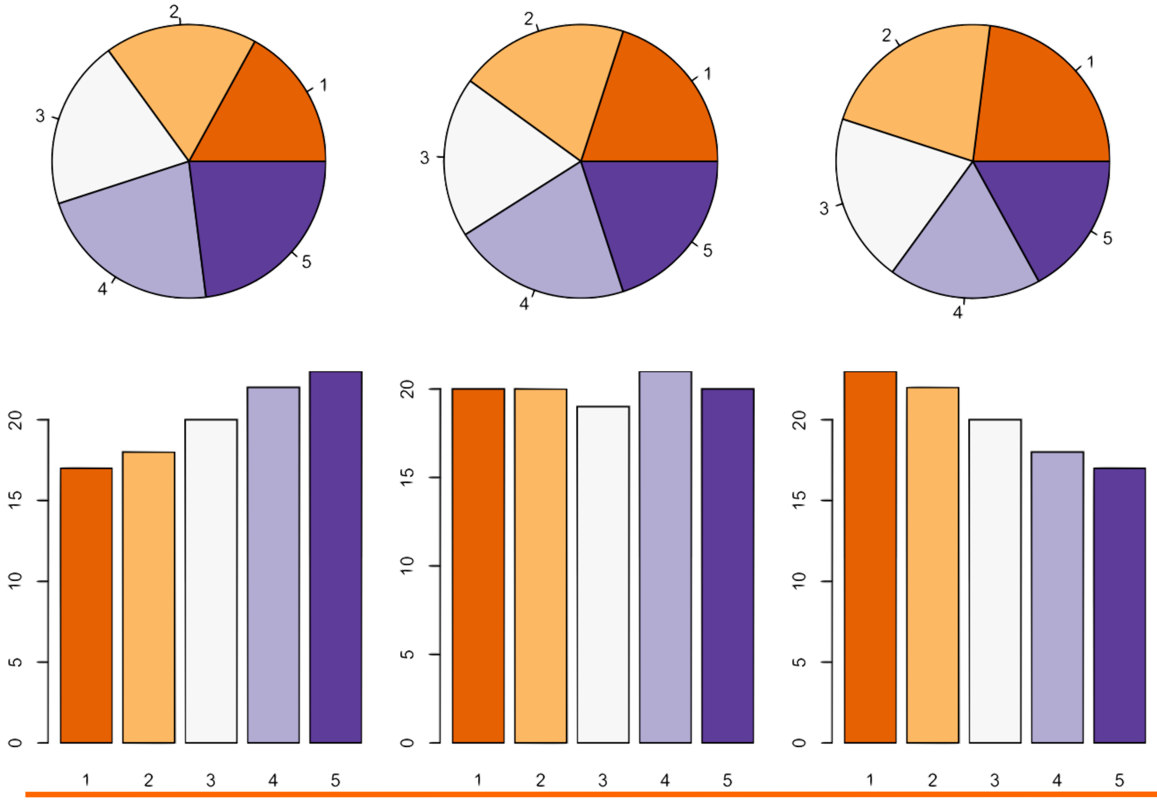
# Pie charts



# Pies



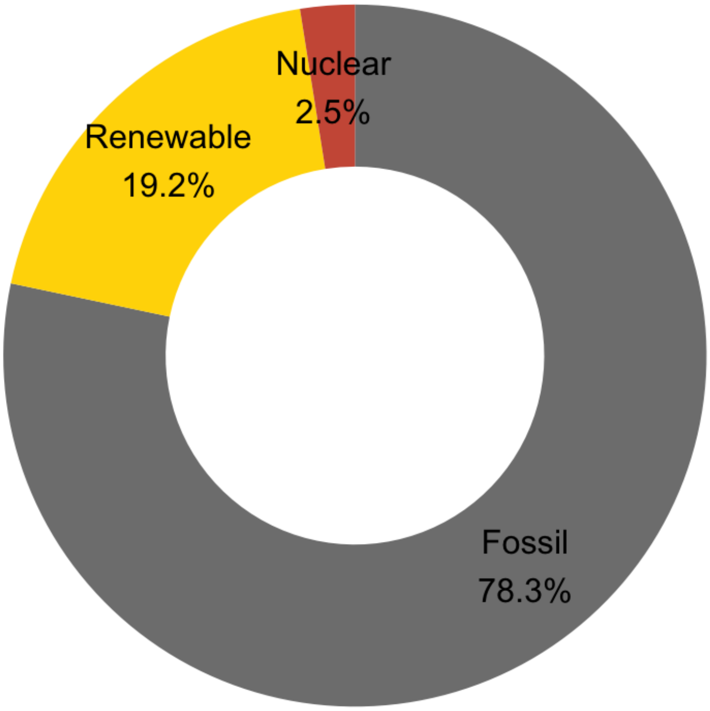
# Pies vs. Bars



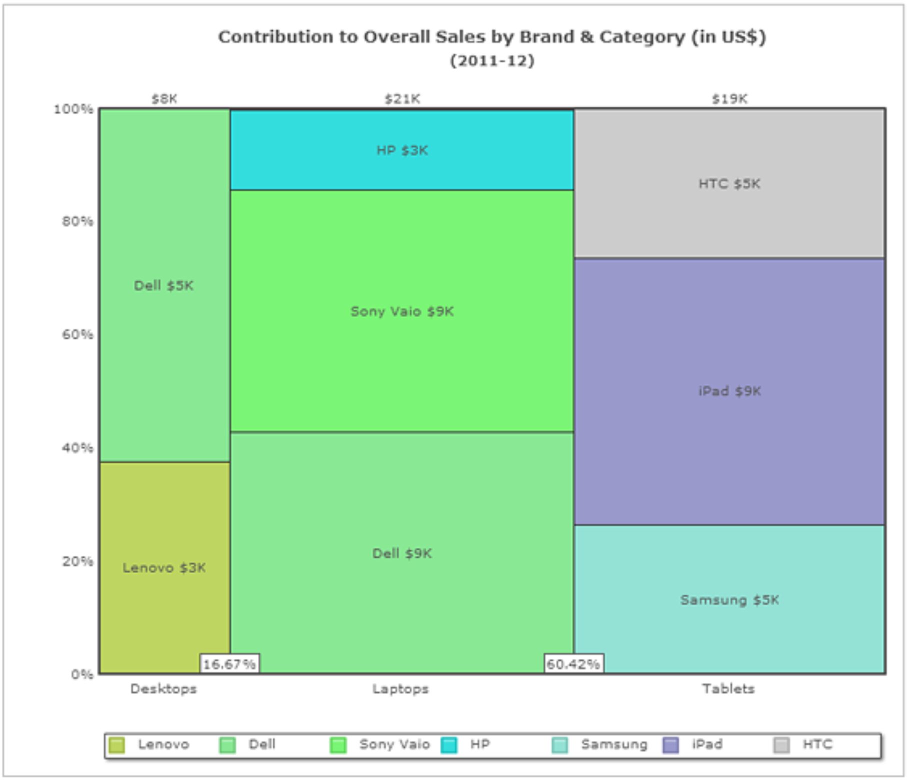
# Pie Charts: guidelines

- 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

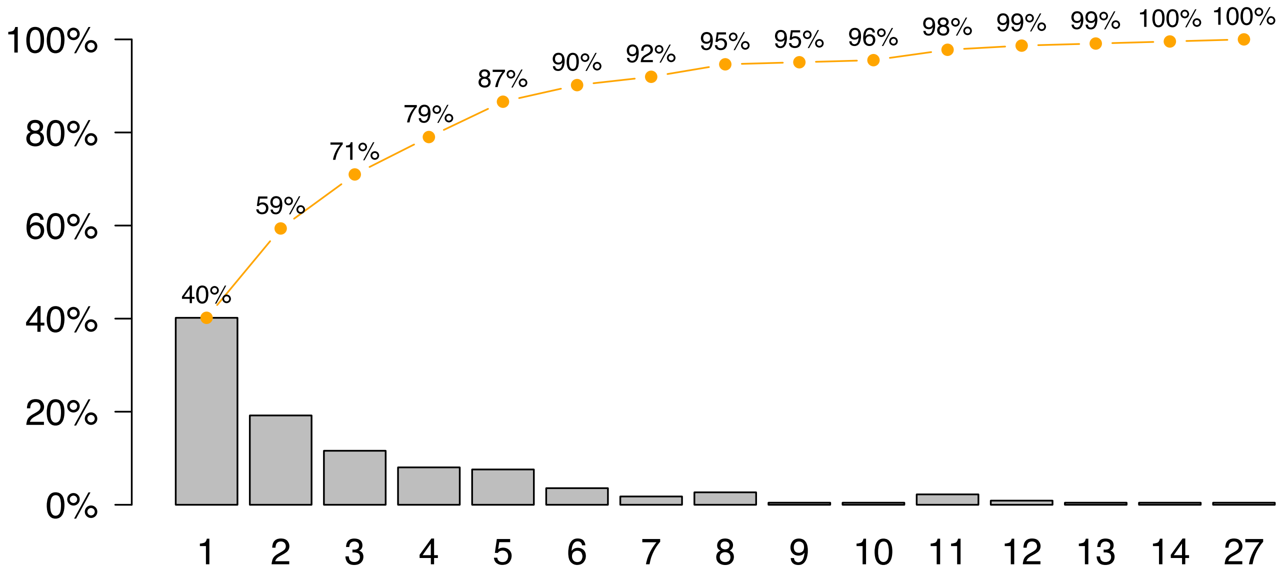


# Marimekko Chart



https://www.fusioncharts.com/chart-primers/marimekko-chart/

# Pareto chart



# Distribution

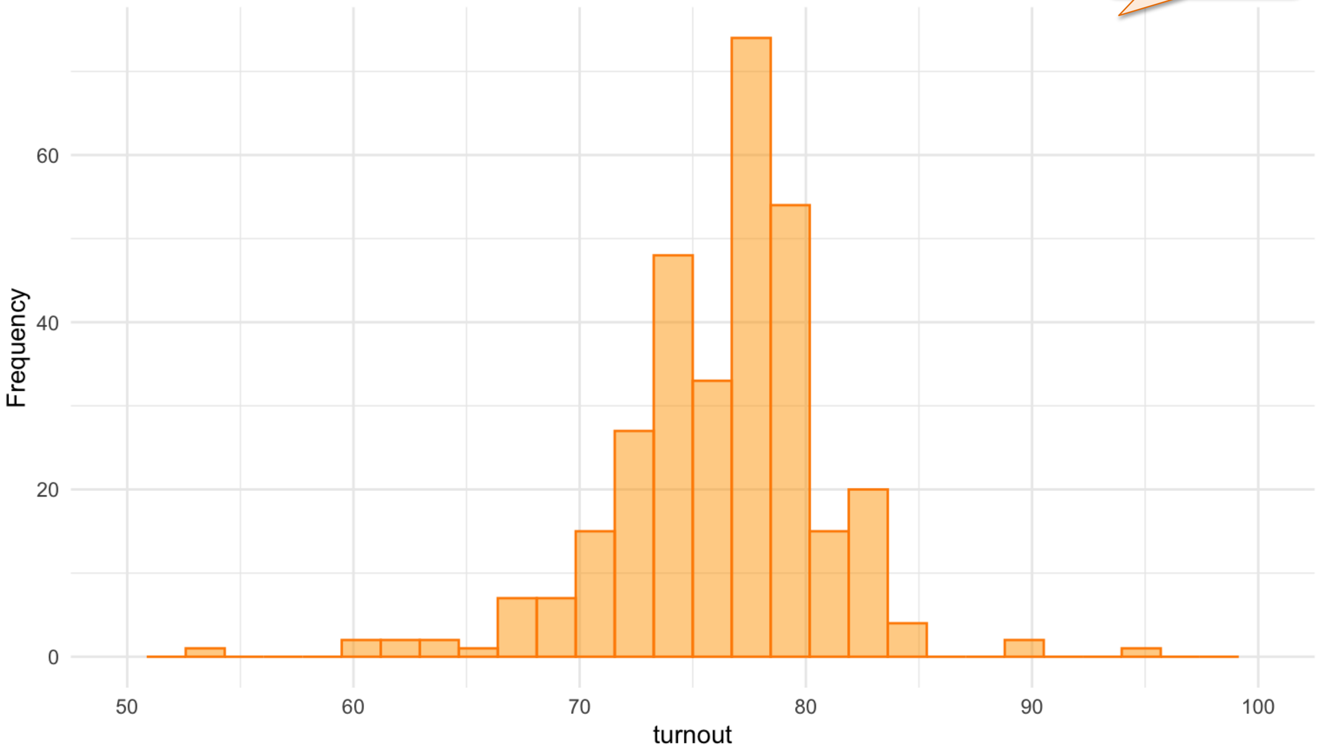
- Continuous values
  - Show distribution of single set of values
  - Show and compare two or more distributions

# 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

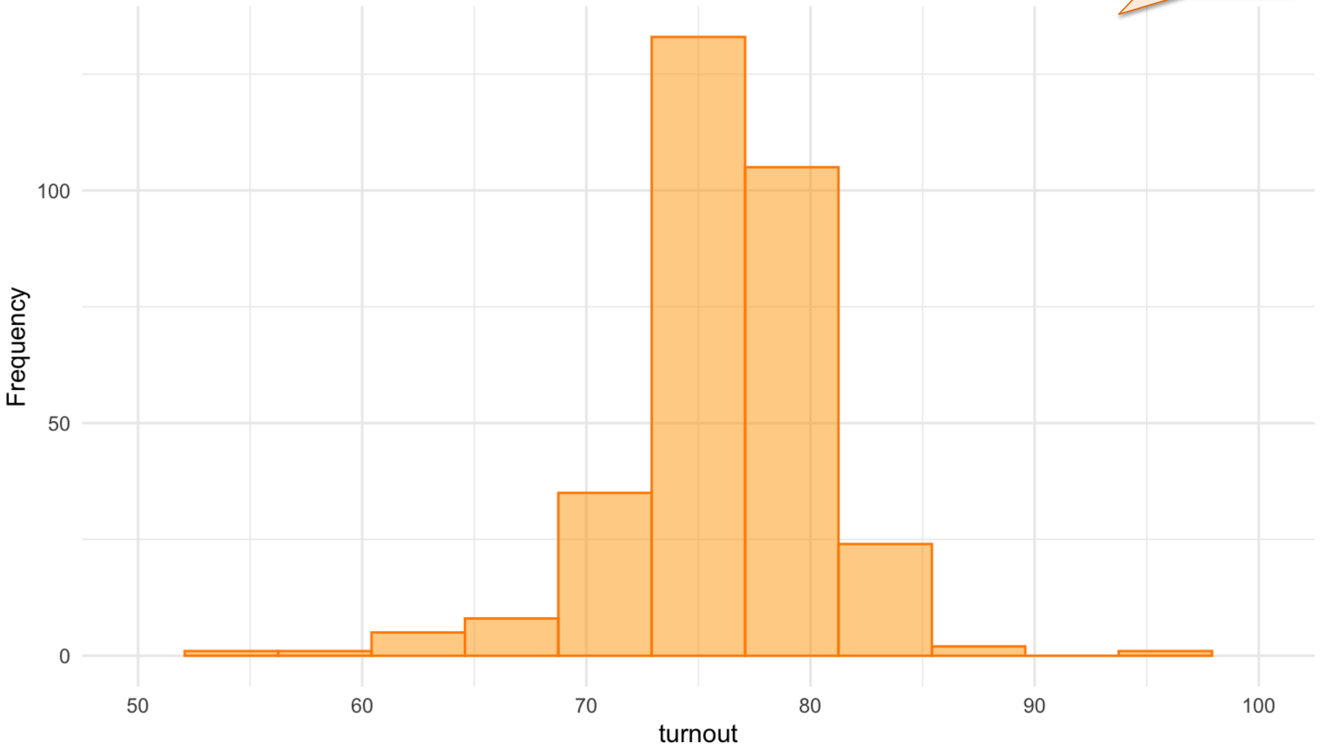
# Histogram

30 bins



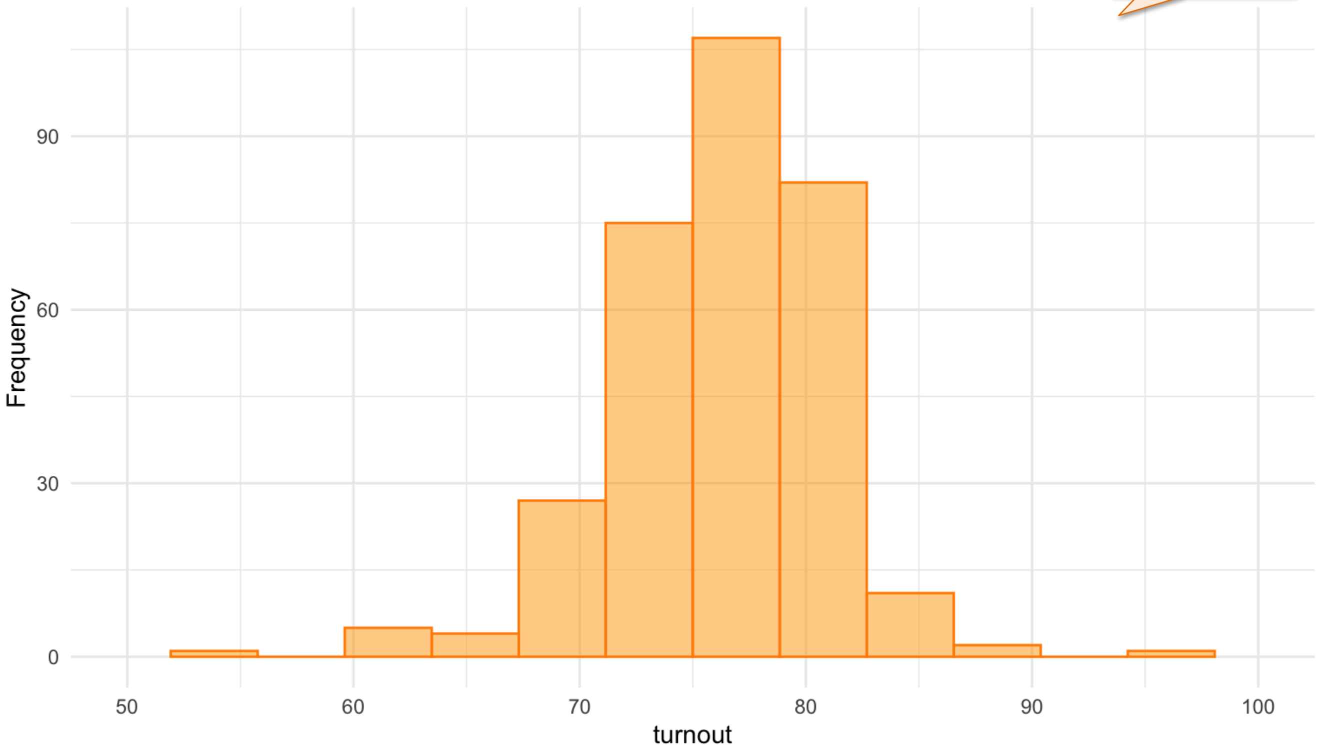
# Histogram

13 bins

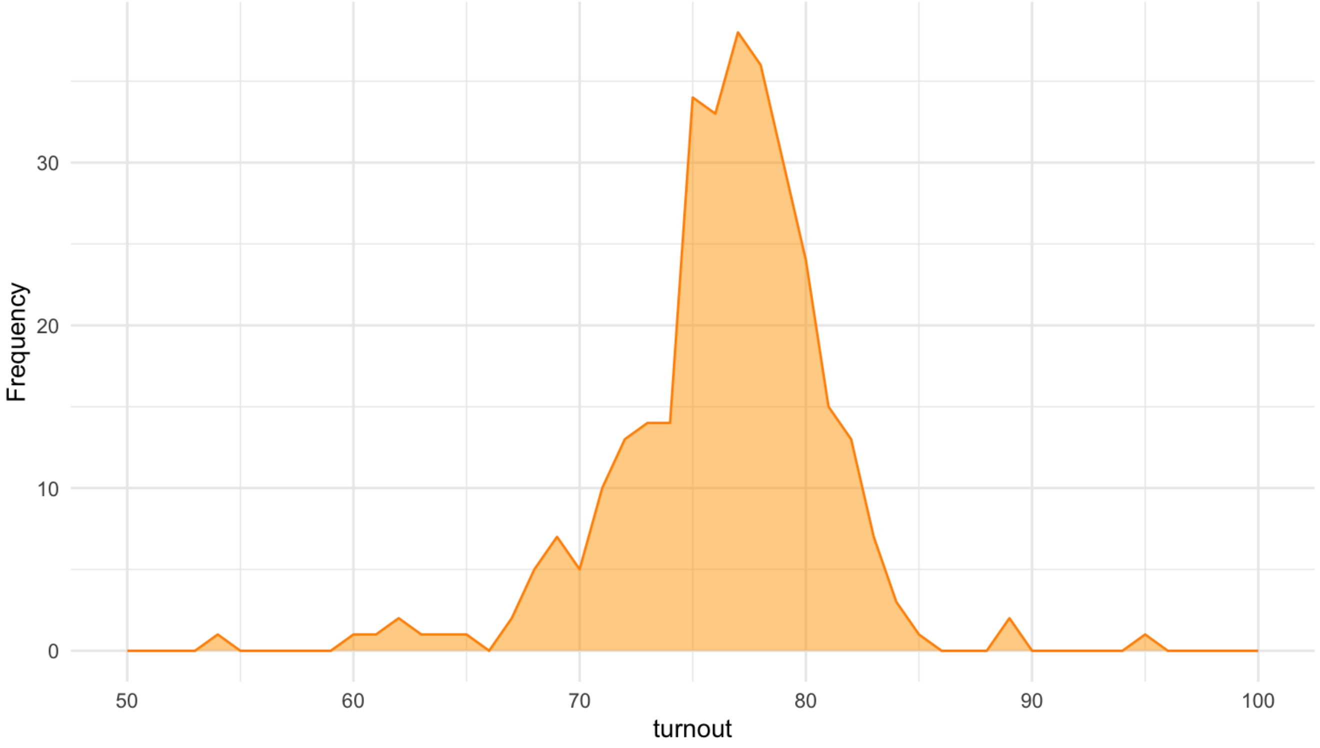


# Histogram

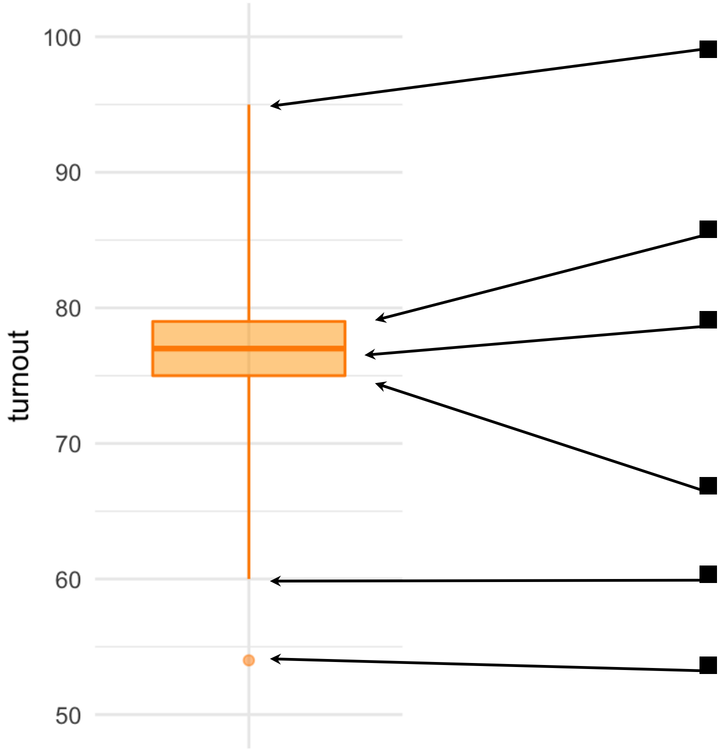
14 bins



# Frequency polygon

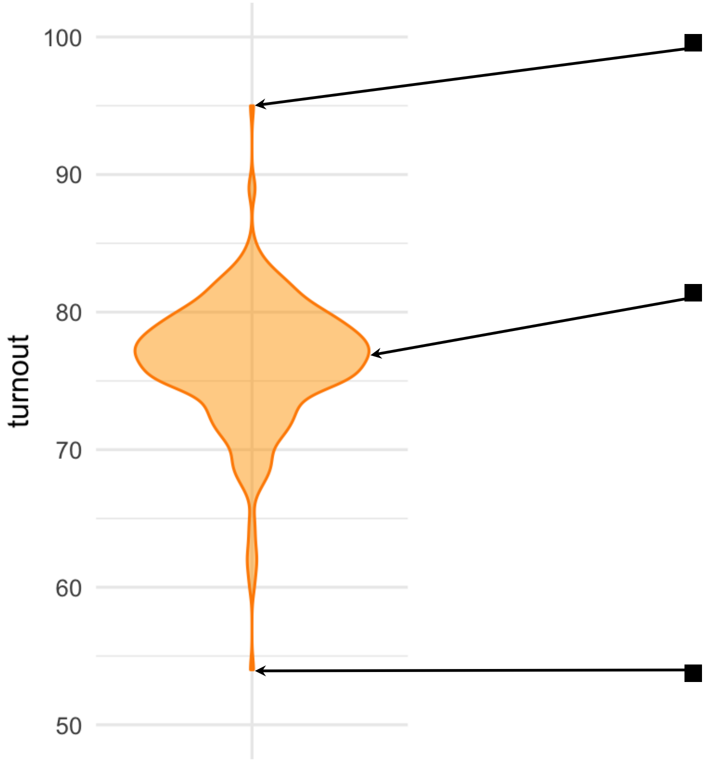


# Boxplot



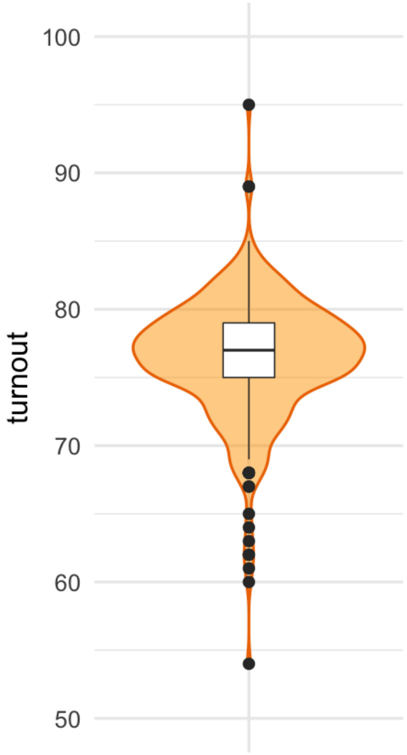
- Max value
- 3rd quartile
- Median
  - 2nd quartile
- 1st quartile
- Min value
- Outlier

# Violin plot



- Max value
- Frequency polygon
  - mirrored
- Min value

# Violin + Boxplot

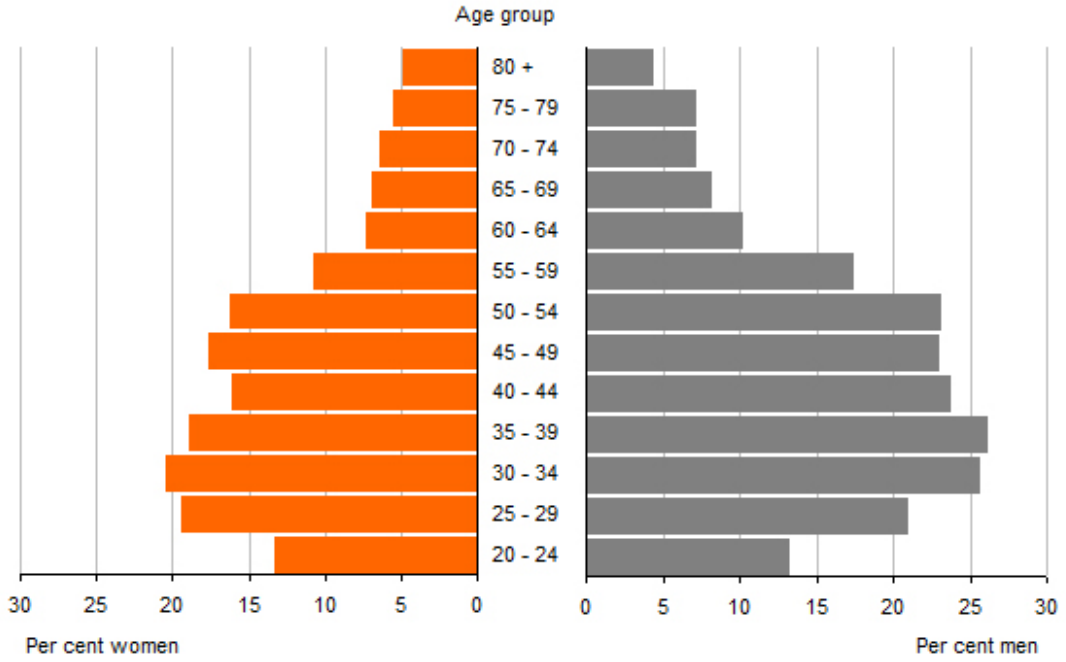


- Overlaying a box plot over the violin provides additional details

# Multiple distribution

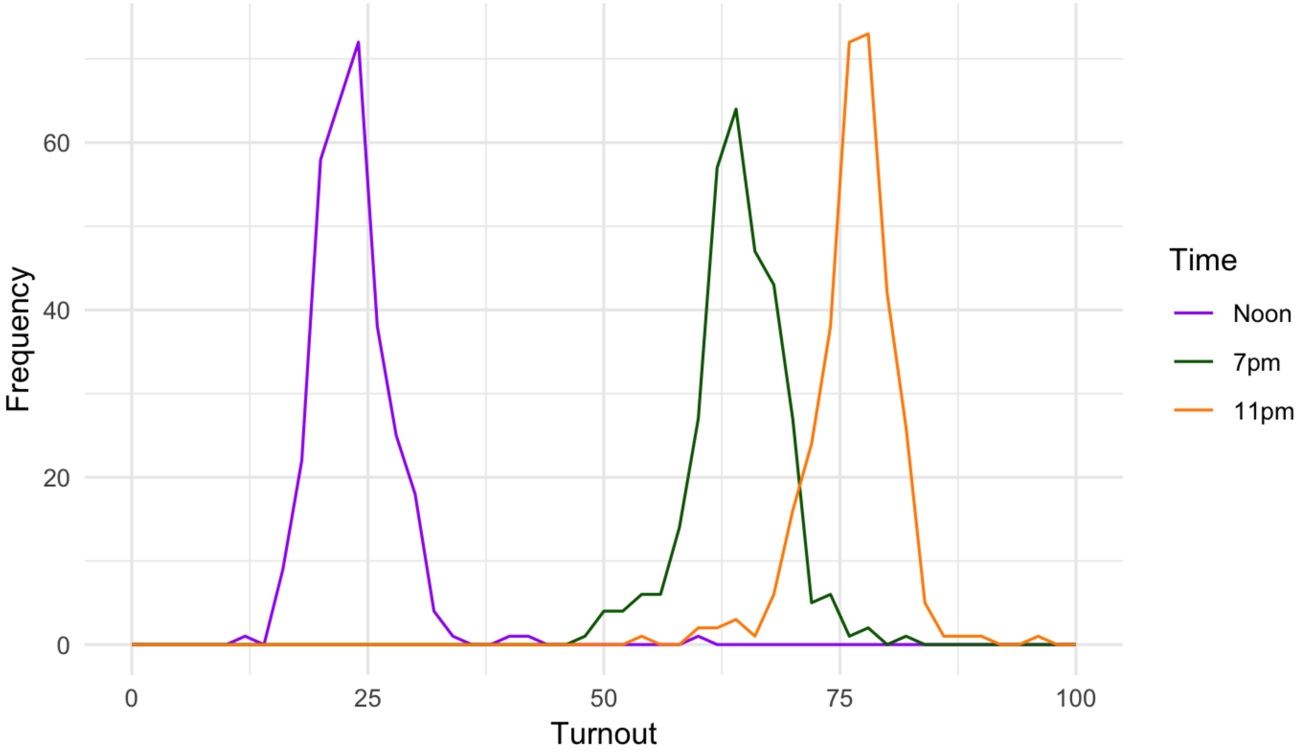
- Histogram is not suitable
- Frequency polygon
  - Line graphs
  - Frequency density function
- Boxplot
  - Summary
  - Less distracting with high number of categories

# Paired diverging bargraph

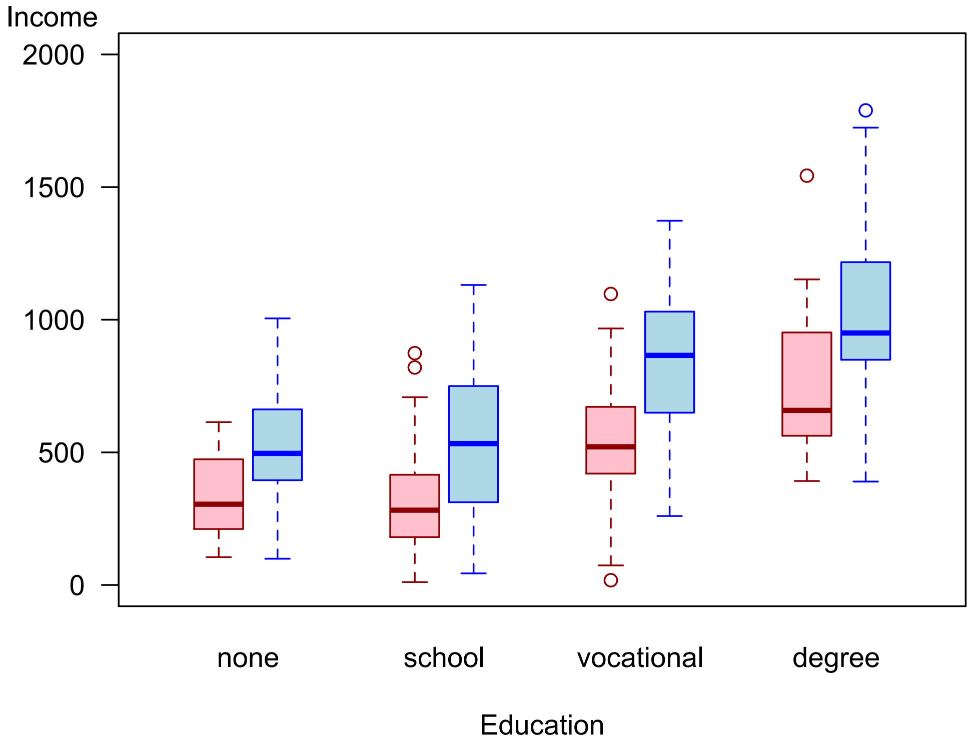


https://unstats.un.org/unsd/genderstatmanual/Print.aspx?Page=Presentation-of-gender-statistics-in-graphs

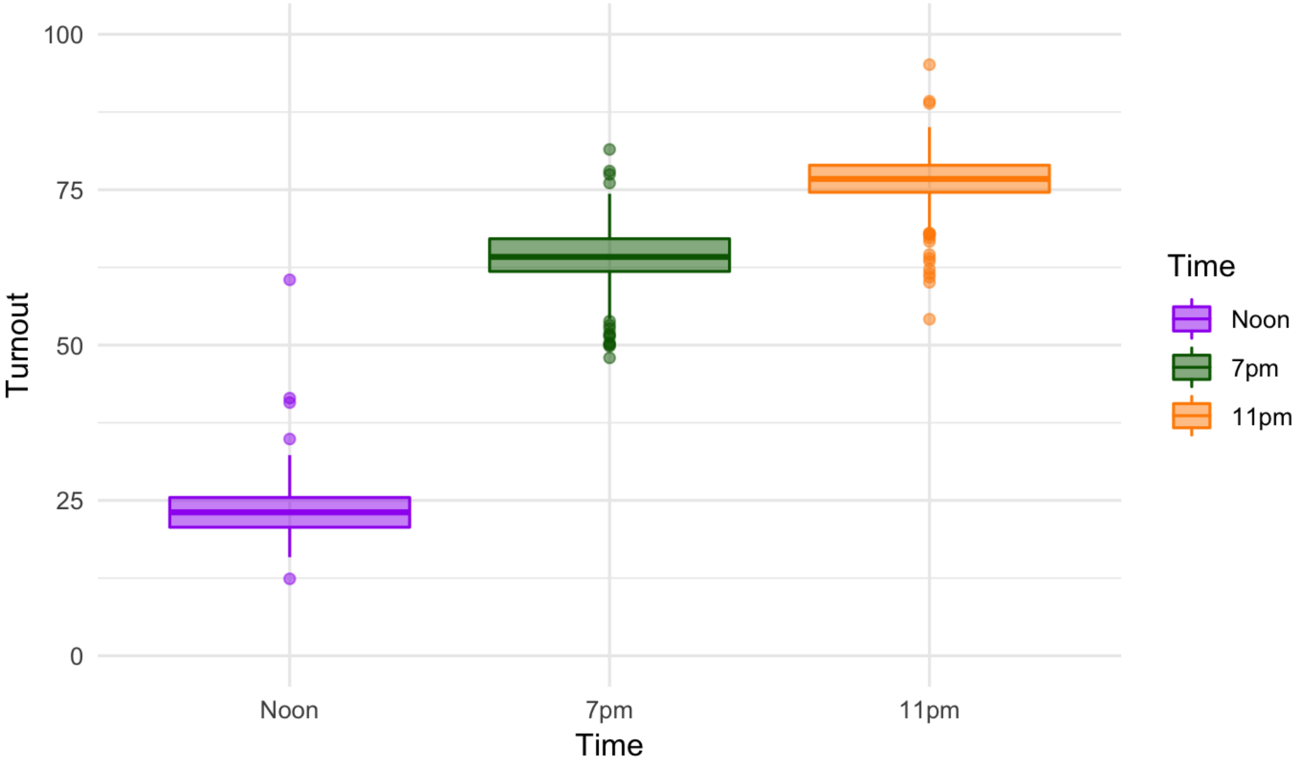
# Multiple Frequency polygons



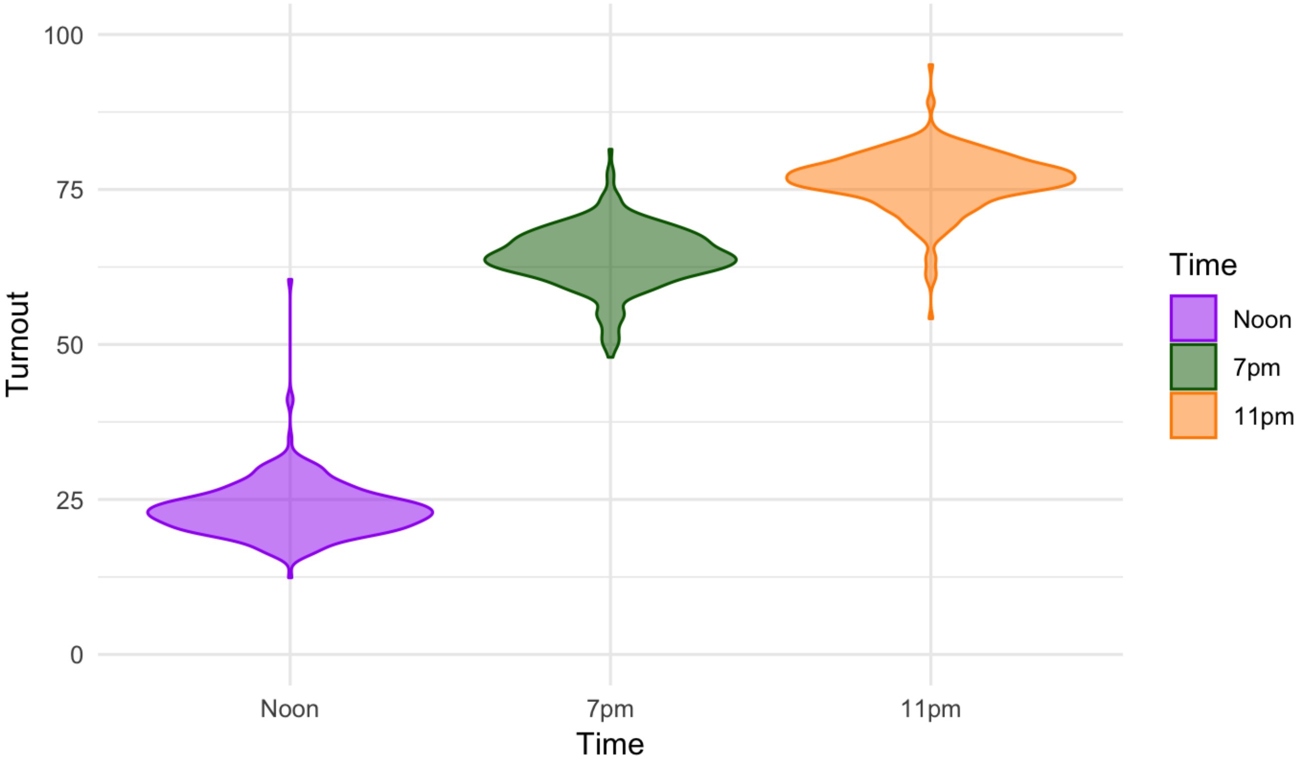
# Box plot



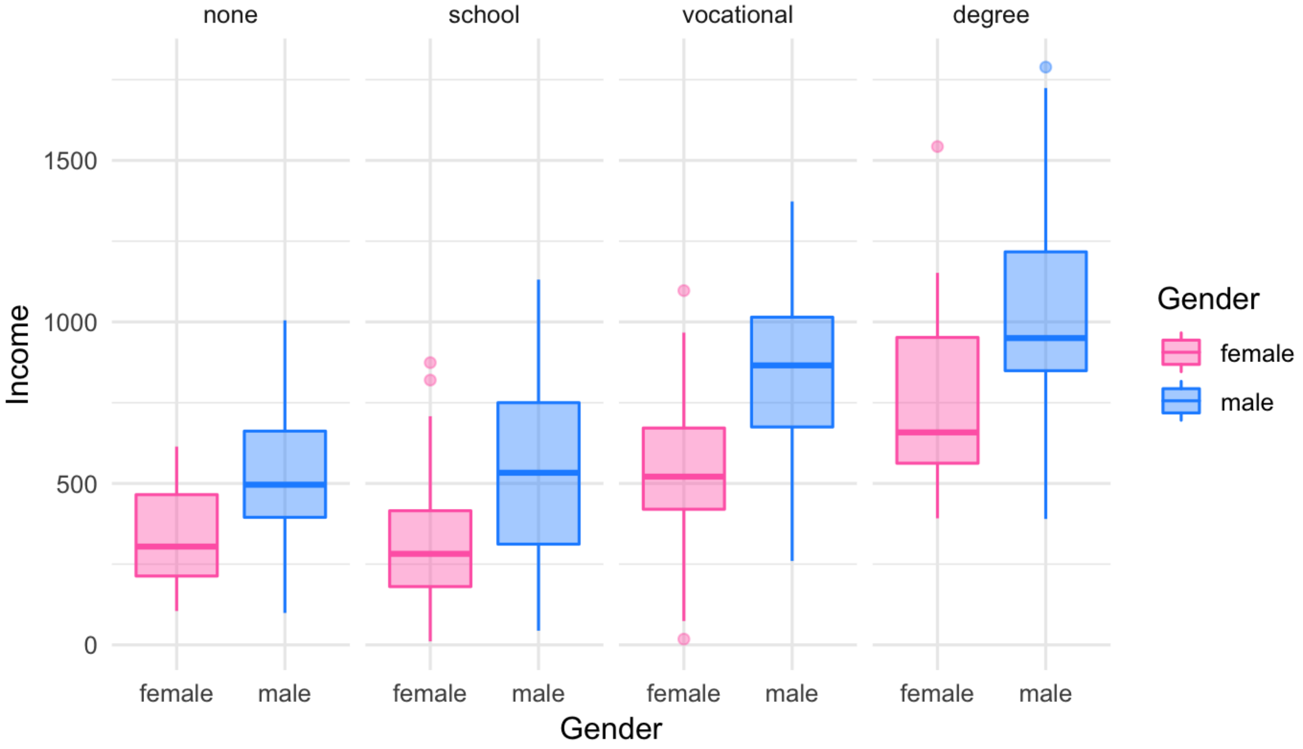
# Multiple Box plot



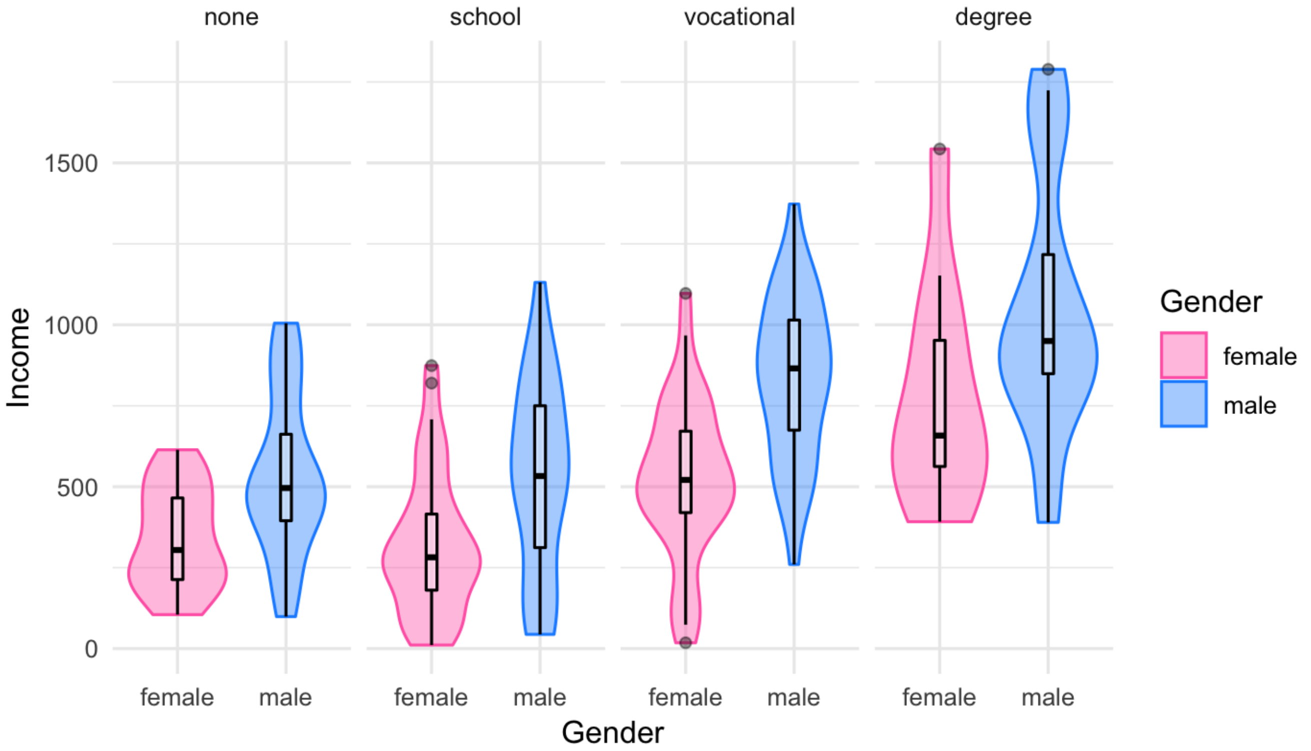
# Violin plot



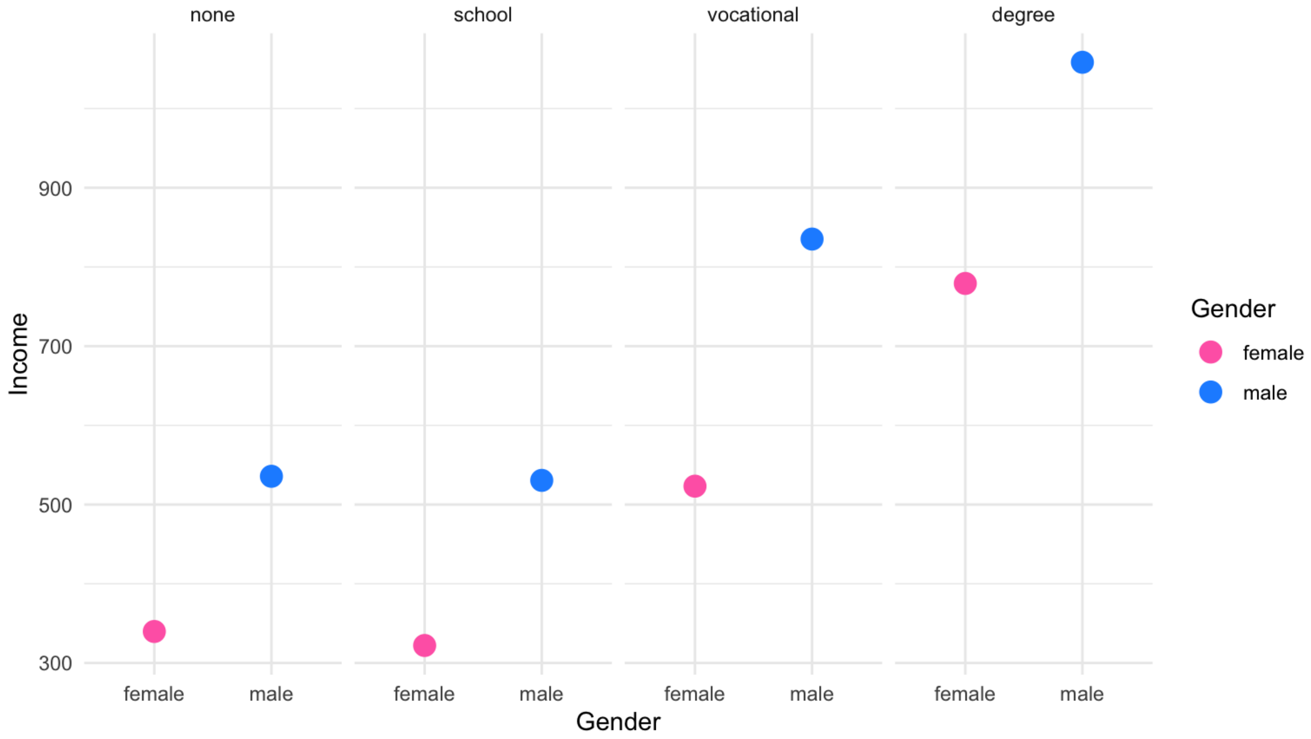
# Multiple box plots



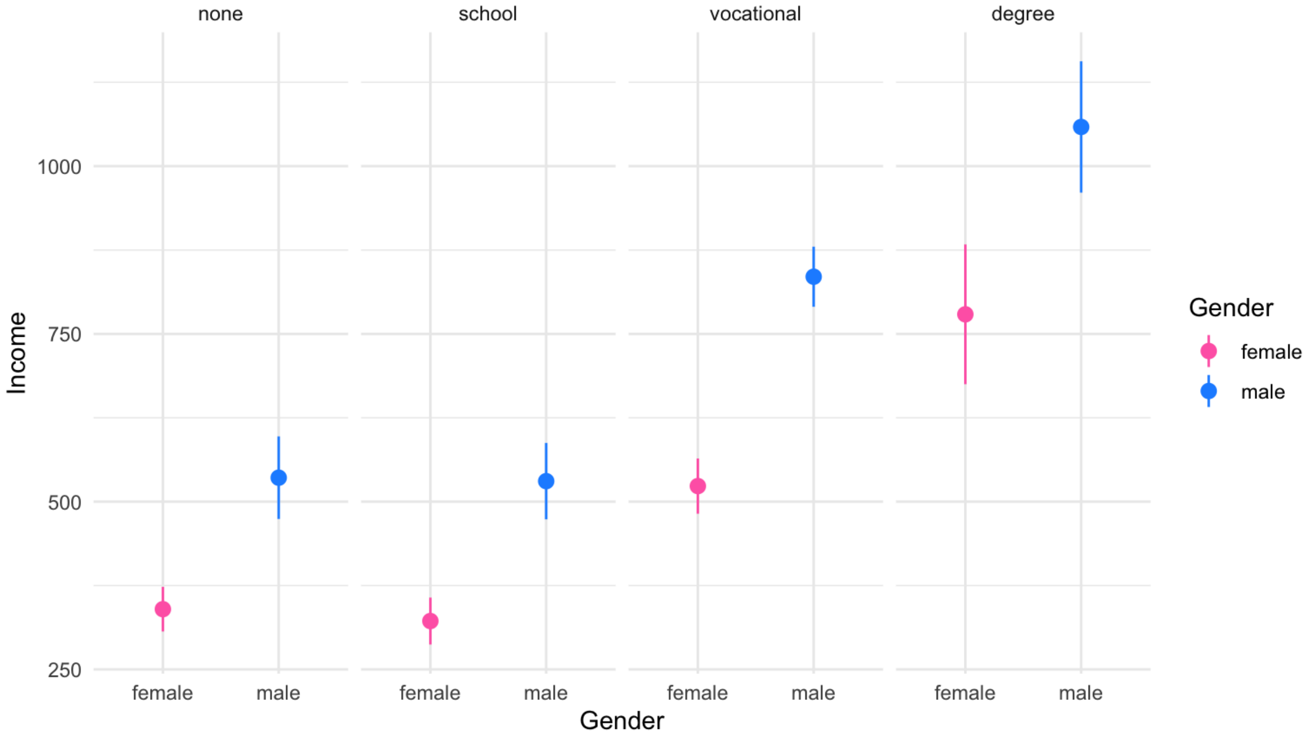
# Multiple violin plots



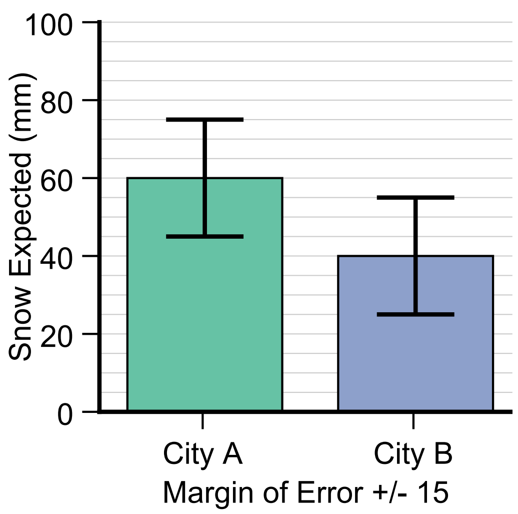
# Just dots for mean values

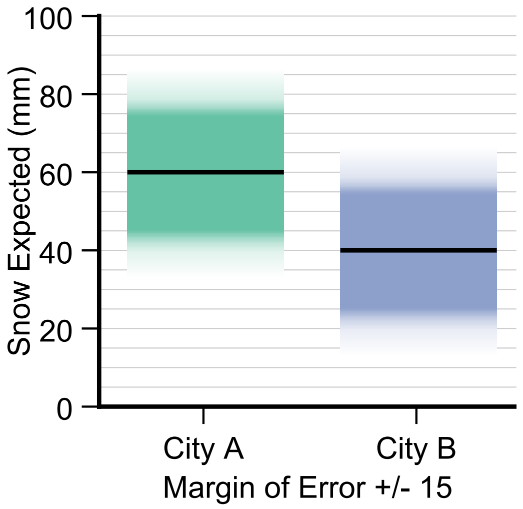


# Confidence intervals



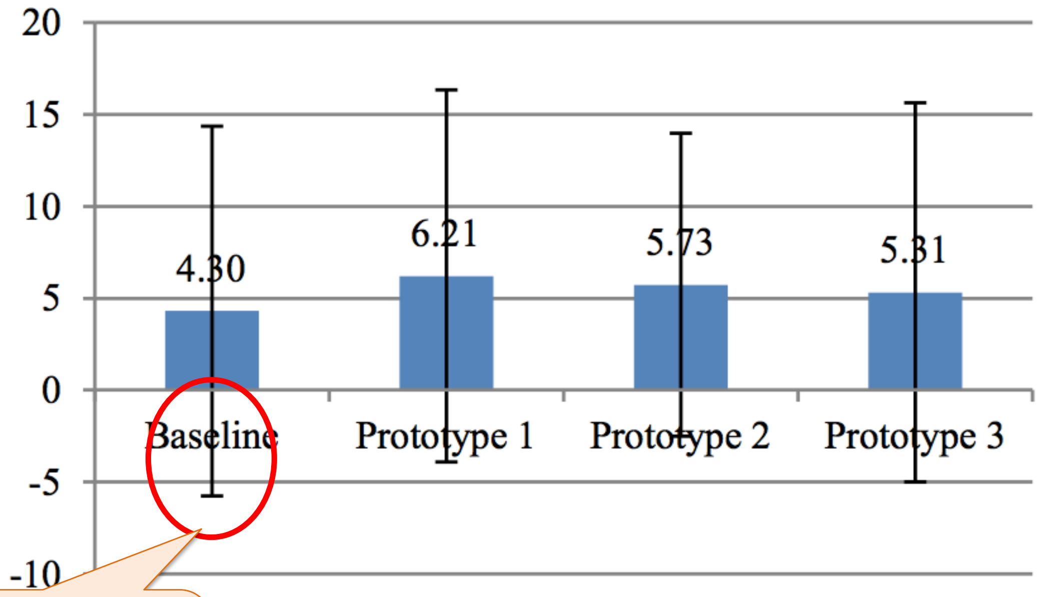
# Confidence Intervals





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

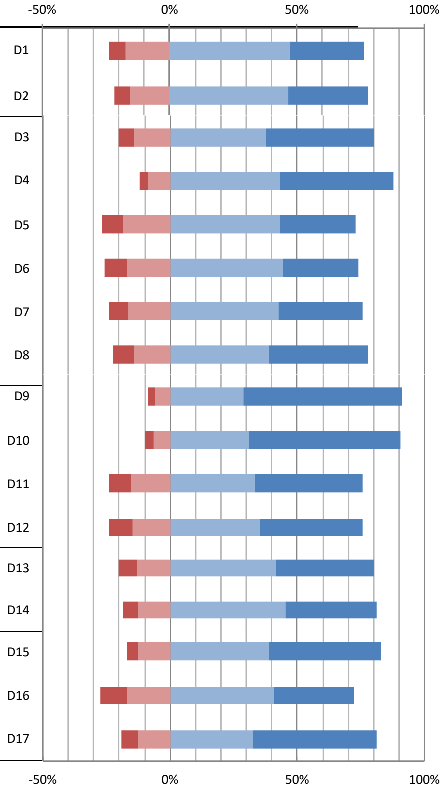
Figure 5. Mean files per changeset.

# 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

| Macroarea | N° | Domanda |
|---|---|---|
| Organizzazione del periodo didattico | 1 | Il carico di studio complessivo degli insegnamenti previsti nel periodo didattico è accettabile? |
| | 2 | L'orario degli insegnamenti del periodo didattico è ben organizzato? |
| Organizzazione di questo insegnamento | 3 | Le regole d'esame, gli obiettivi e il programma dell'insegnamento sono stati resi noti in modo chiaro? |
| | 4 | L'insegnamento è stato svolto in maniera coerente con quanto dichiarato sul portale della didattica? |
| | 5 | Le conoscenze preliminari da me possedute sono risultate sufficienti per la comprensione della materia ? |
| | 6 | Il carico di studio richiesto da questo insegnamento è proporzionato ai crediti assegnati? |
| | 7 | Il materiale didattico, indicato o fornito, è adeguato per lo studio della materia? |
| | 8 | Le attività didattiche integrative (esercitazioni, lab, seminari, visite, ecc.) sono utili per l'apprendimento della materia? |
| Efficacia del docente | 9 | Il docente rispetta gli orari di svolgimento dell'attività didattica? |
| | 10 | Il docente è disponibile a fornire chiarimenti e spiegazioni? |
| | 11 | Il docente interagisce efficacemente con gli studenti, stimolando l'interesse verso la materia? |
| | 12 | Il docente espone gli argomenti in modo chiaro? |
| Infrastrutture | 13 | Le aule in cui si svolgono le lezioni sono adeguate? |
| | 14 | I locali e le attrezzature per le attività didattiche integrative sono adeguati? |
| Interesse e soddisfazione | 15 | Sono interessato agli argomenti di questo insegnamento? (indipendentemente da come è stato svolto) |
| | 16 | Sono soddisfatto di come è stato svolto questo insegnamento? |
| | 17 | Al fine dell�apprendimento, la frequenza alle attività didattiche è utile? |

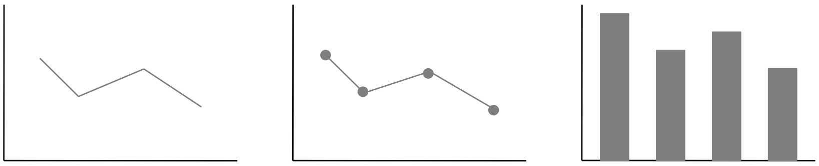


# Time series

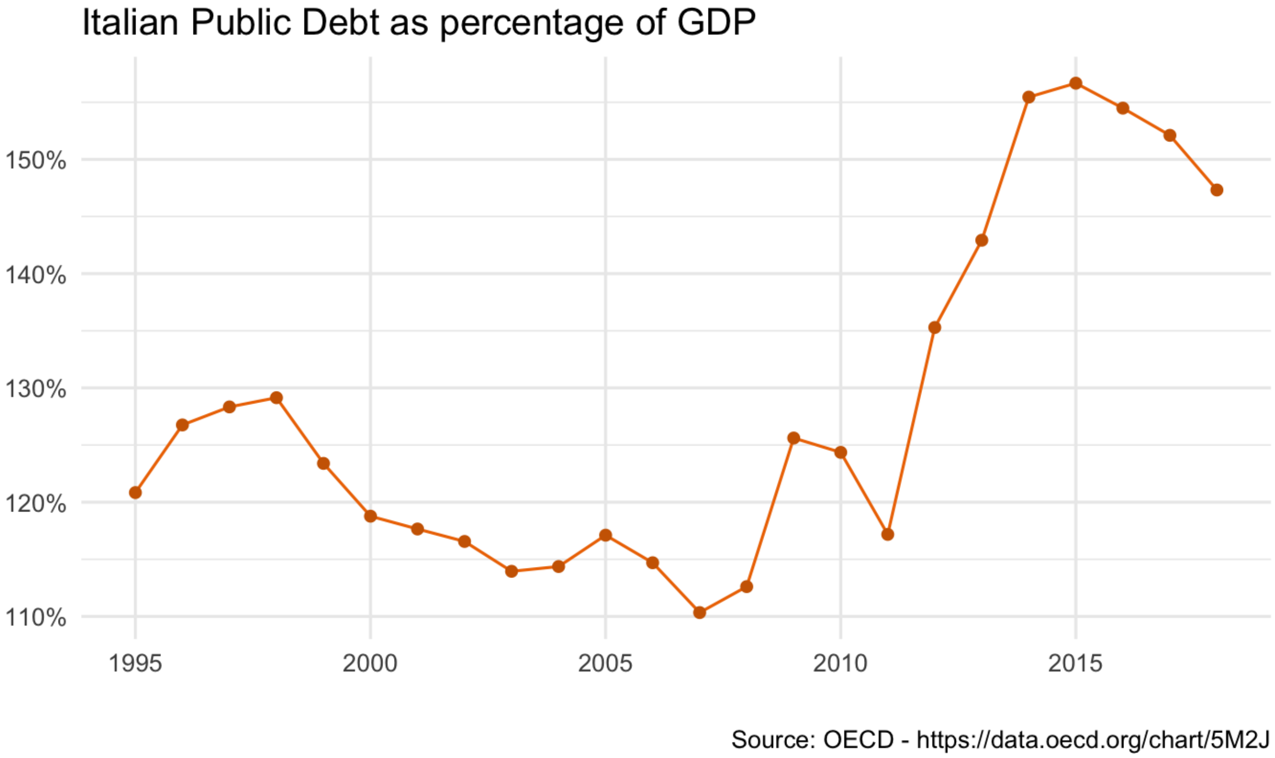
- Series of relationships between quantitative values that are associated with categorical subdivisions of time
- Communicate
  - Change
  - Rise
  - Increase
  - Fluctuate
  - Grow
  - Decline
  - Decrease
  - Trend

# Time series

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

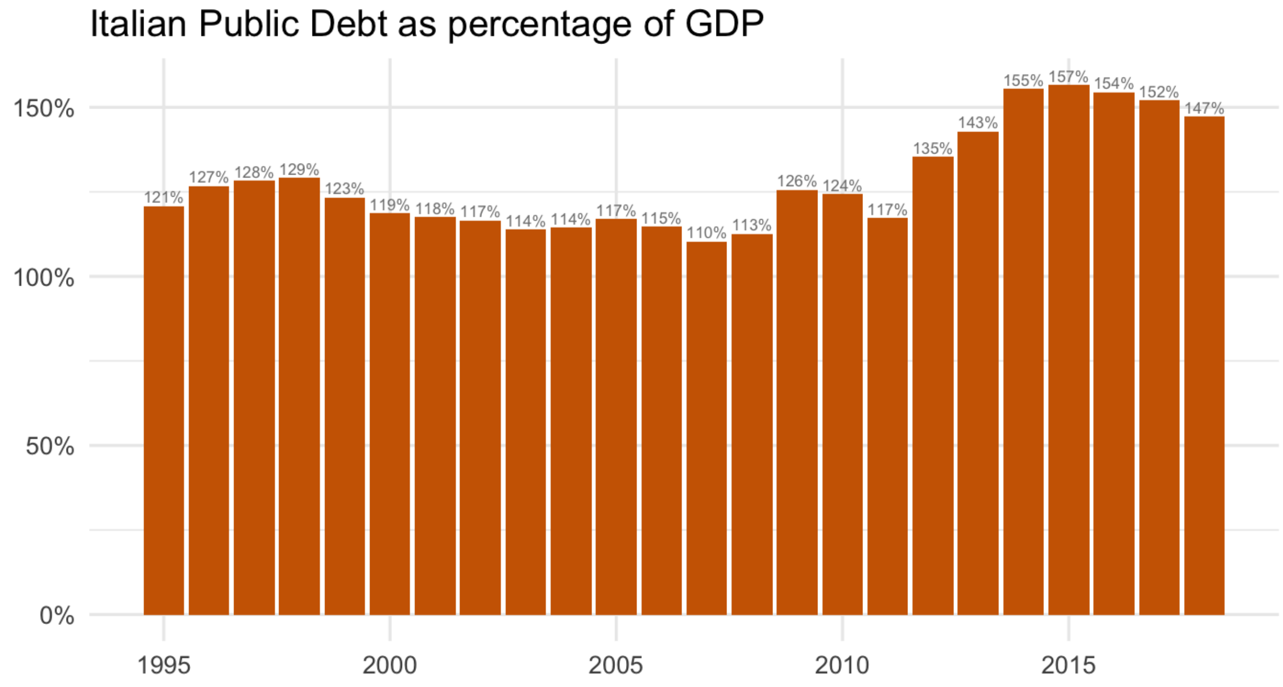


# Line plot



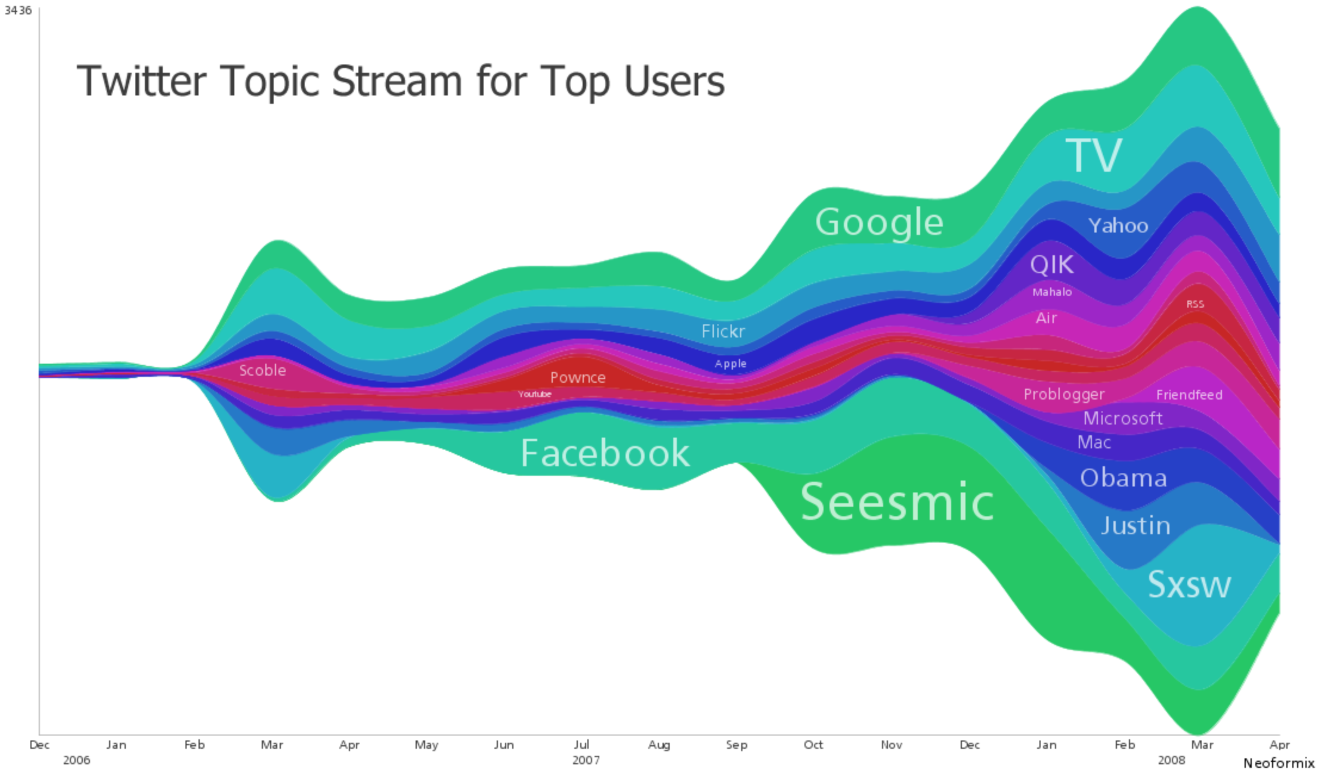
Source: OECD - https://data.oecd.org/chart/5M2J

# Bars



Source: OECD - https://data.oecd.org/chart/5M2J

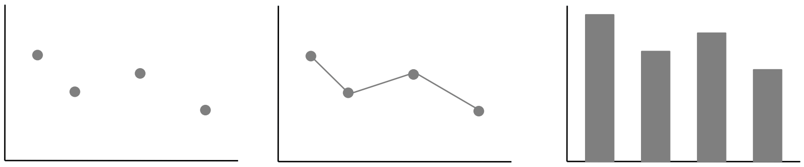
# Streamgraph



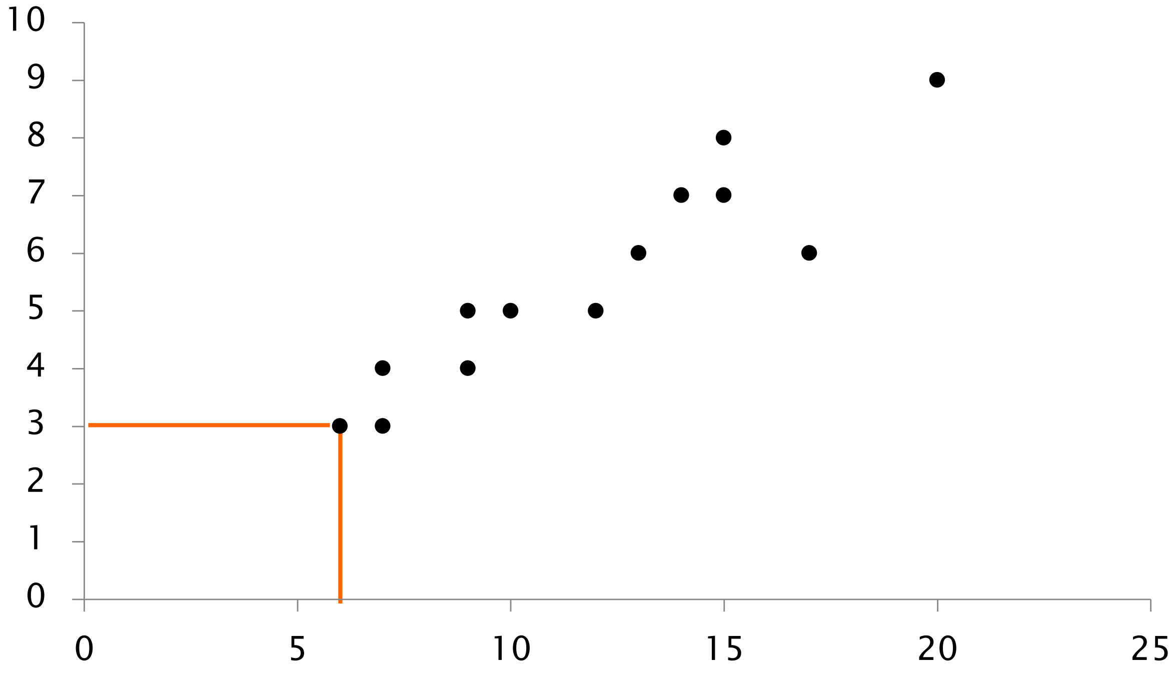
http://www.neoformix.com/2008/TwitterTopicStream.html

# Correlation

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



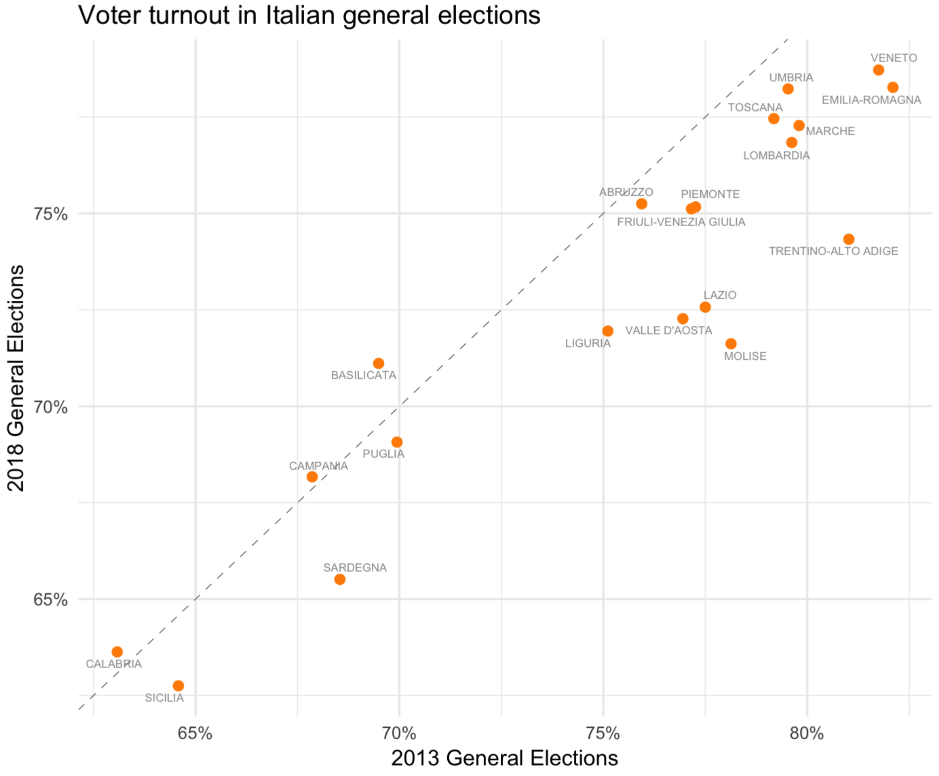
# Points



# 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

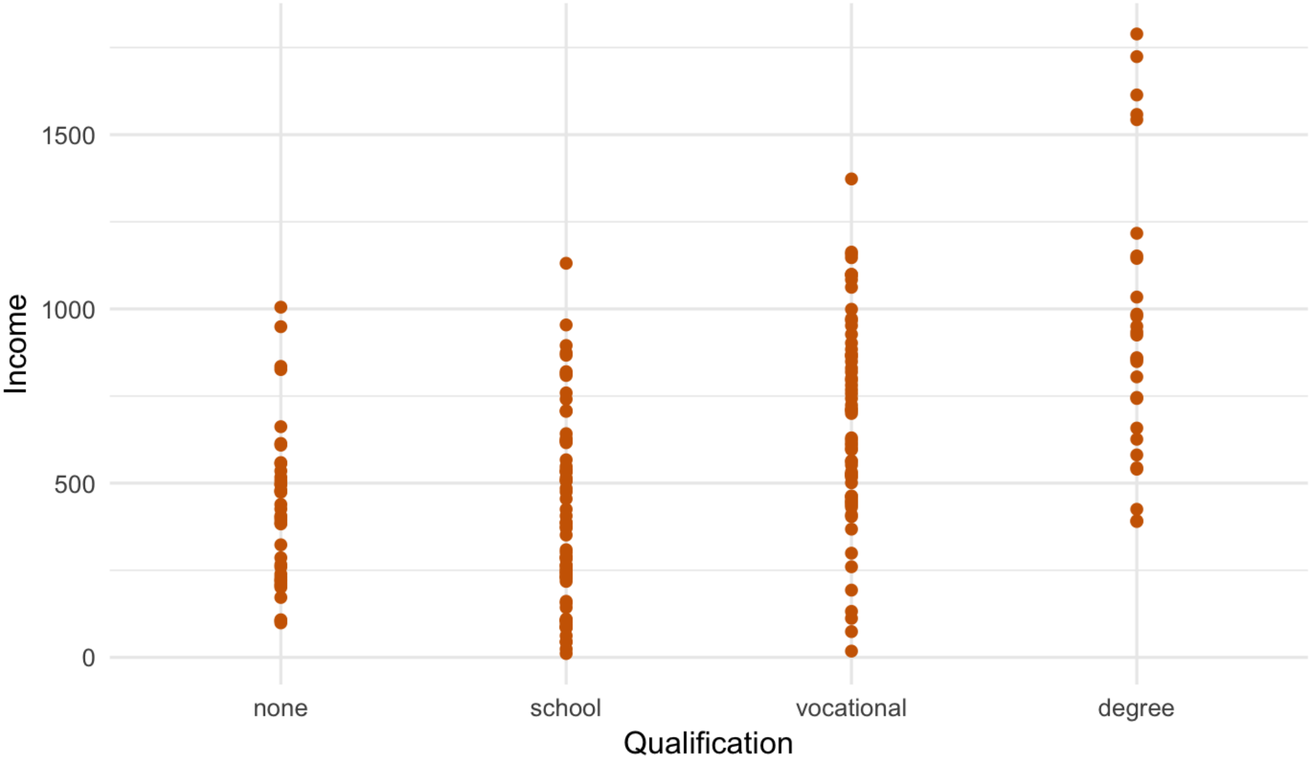
# Scatter plot



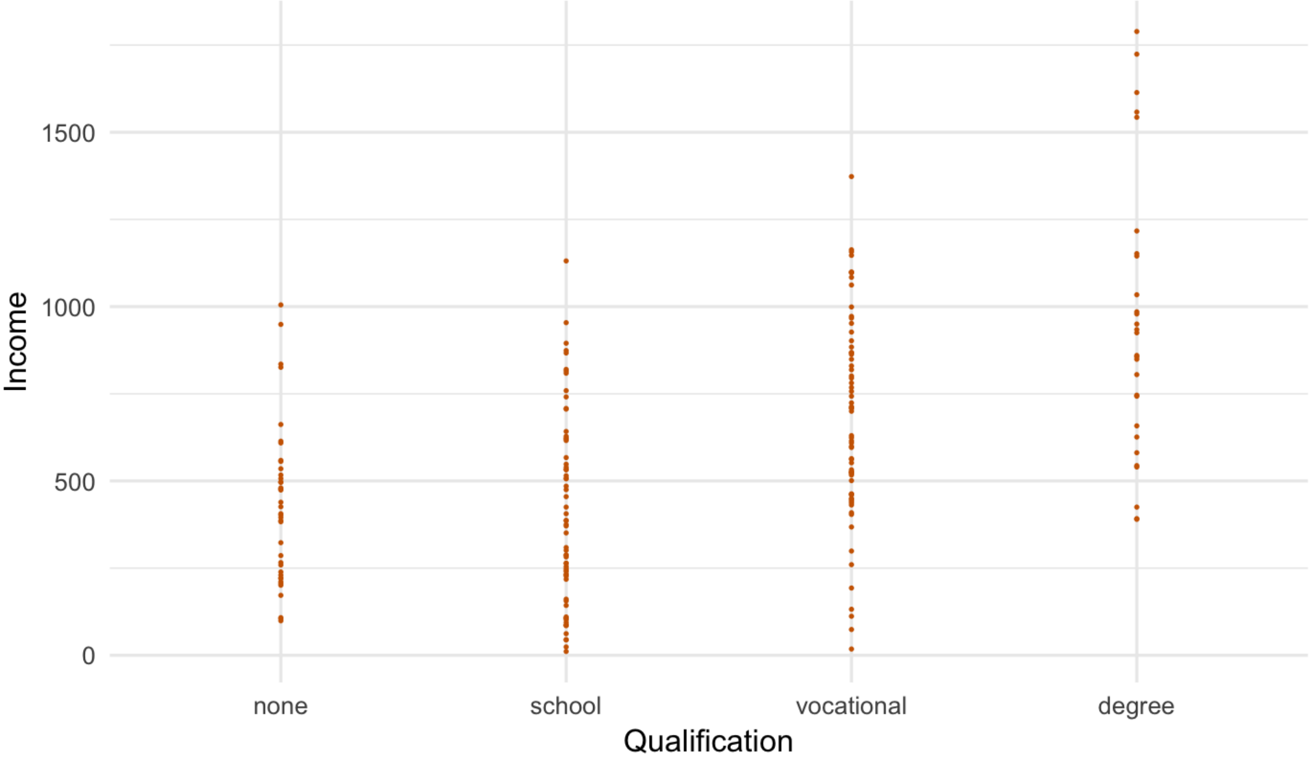
# Overplotting

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

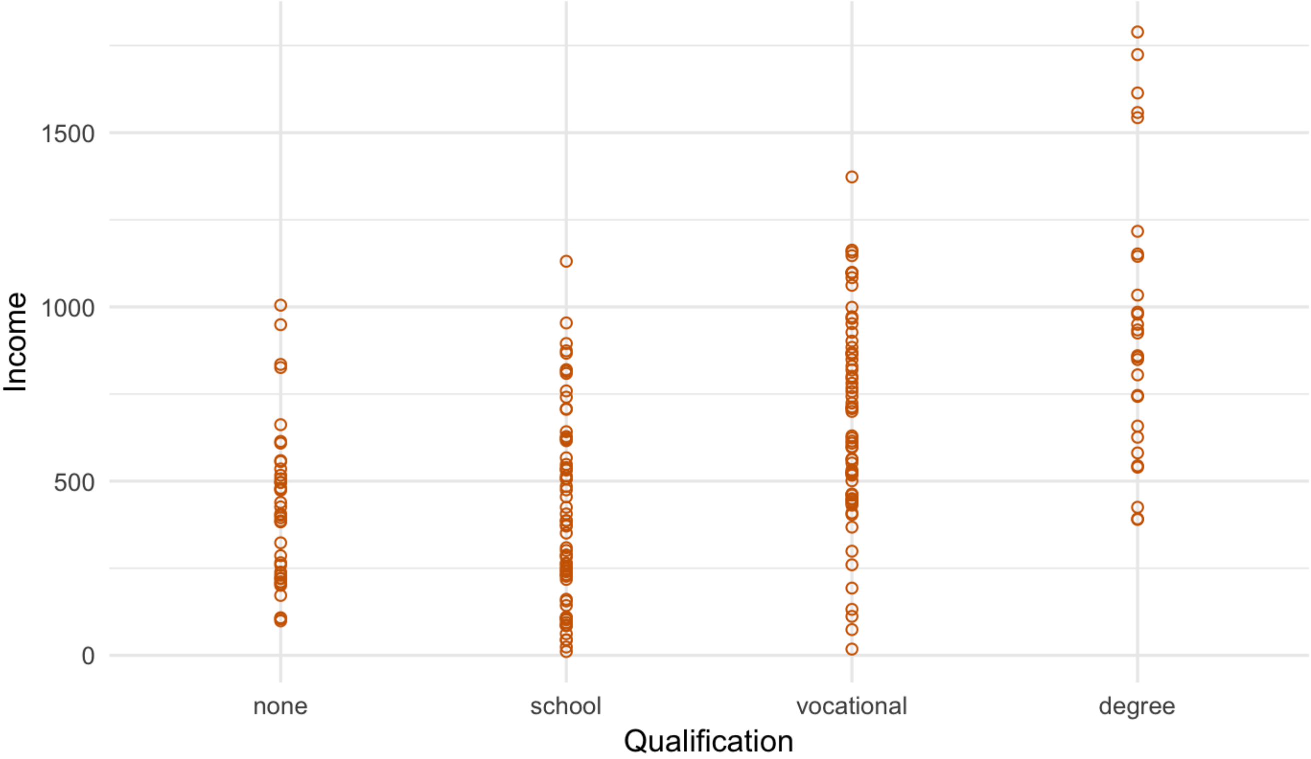
# Overplotting example



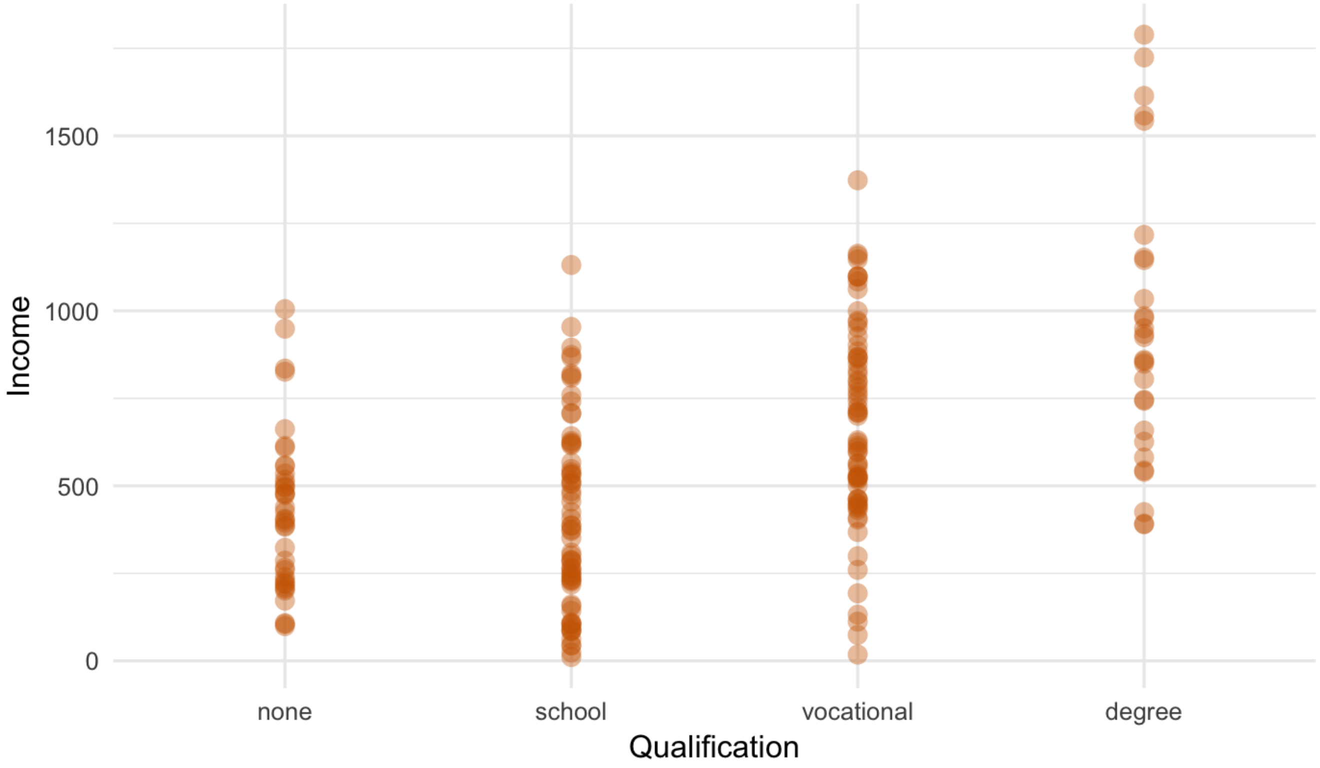
# Overplotting – Small



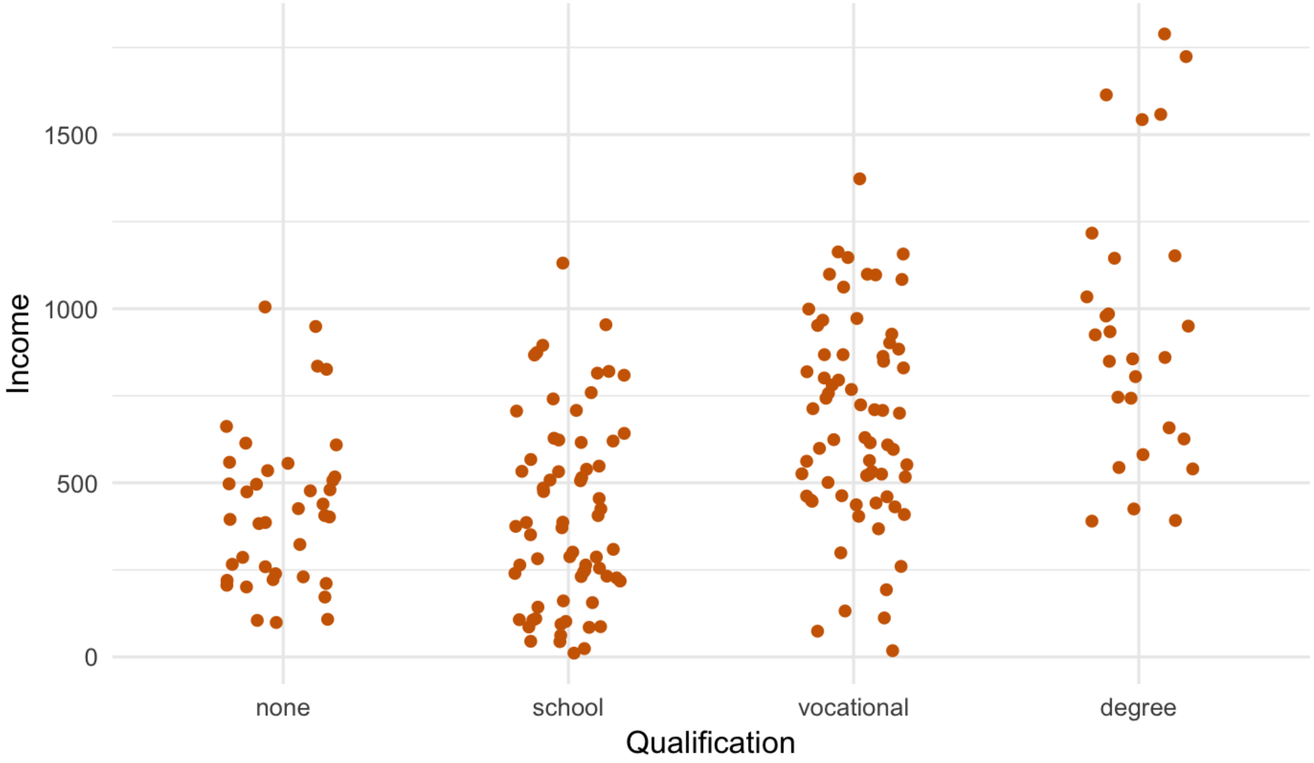
# Overplotting – Outlined



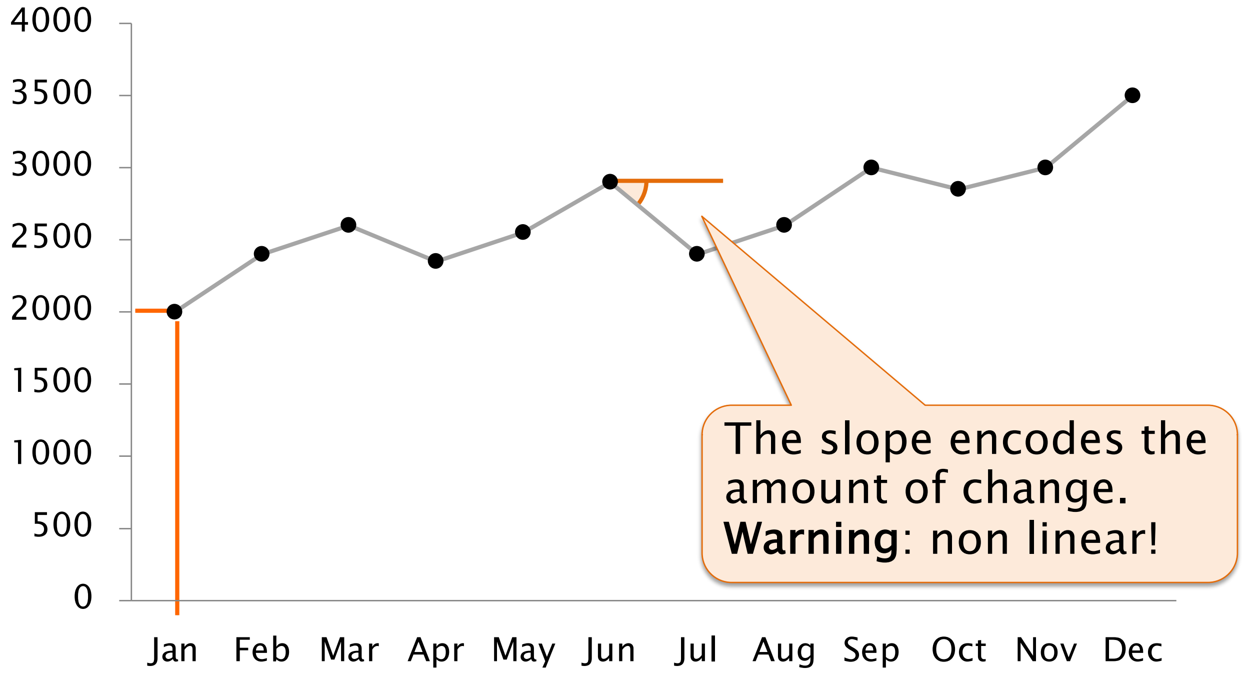
# Overplotting – Transparent



# Overplotting – Jittering

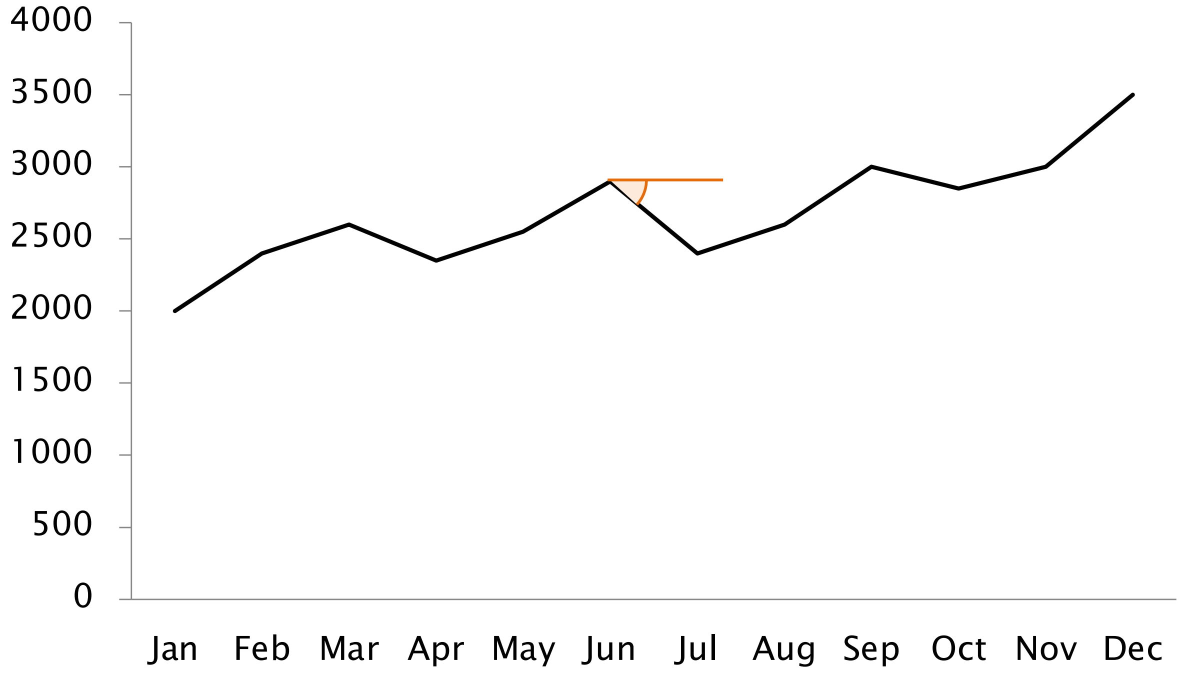


# Points and Lines

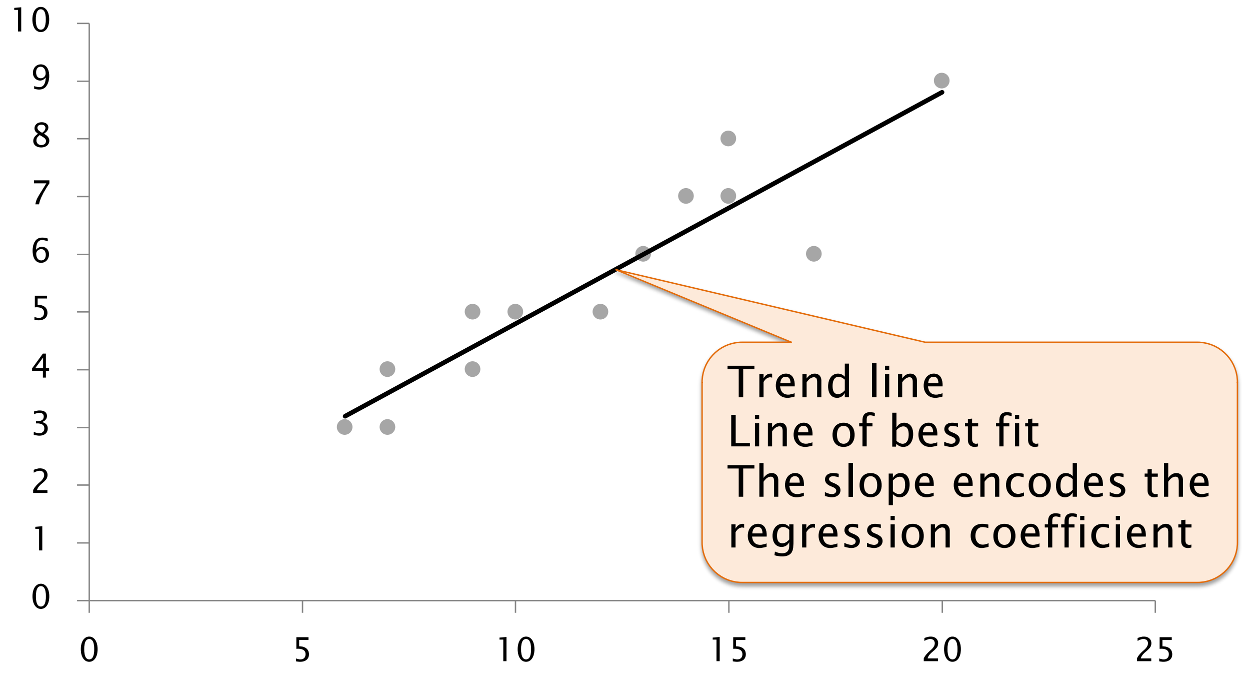


The slope encodes the amount of change. **Warning**: non linear!

# Slope of lines



# Slope of lines

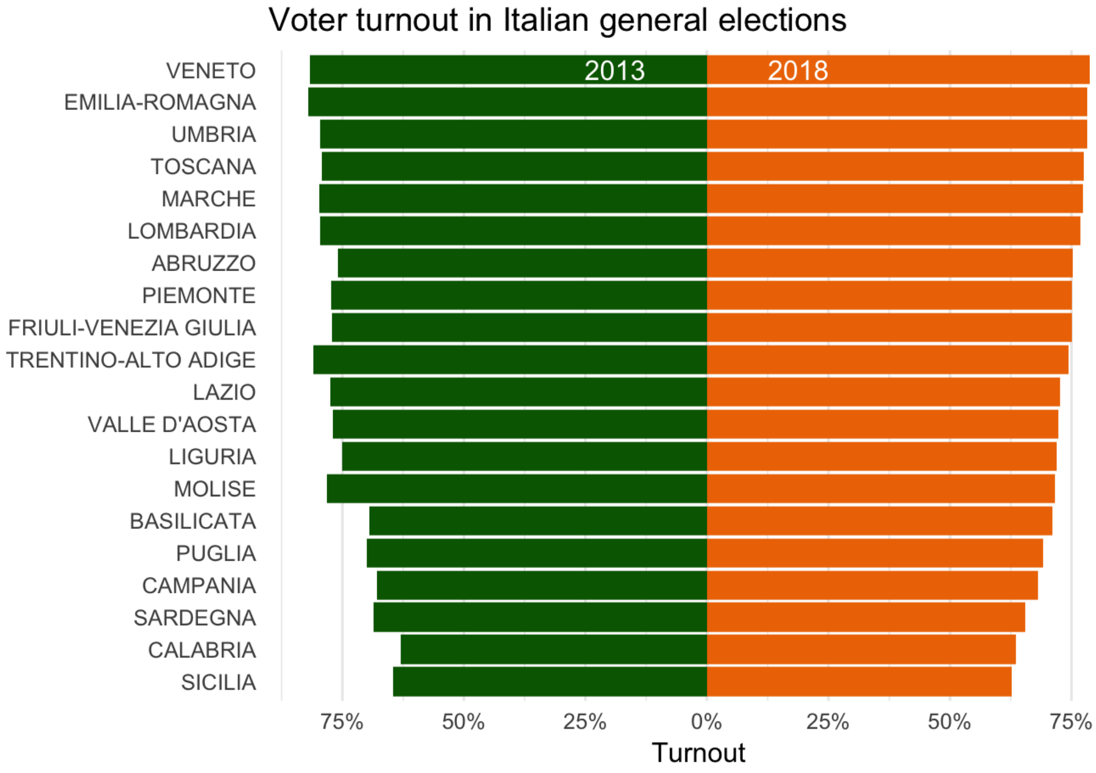


Trend line
Line of best fit
The slope encodes the regression coefficient

# Lines

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

# Paired diverging bars

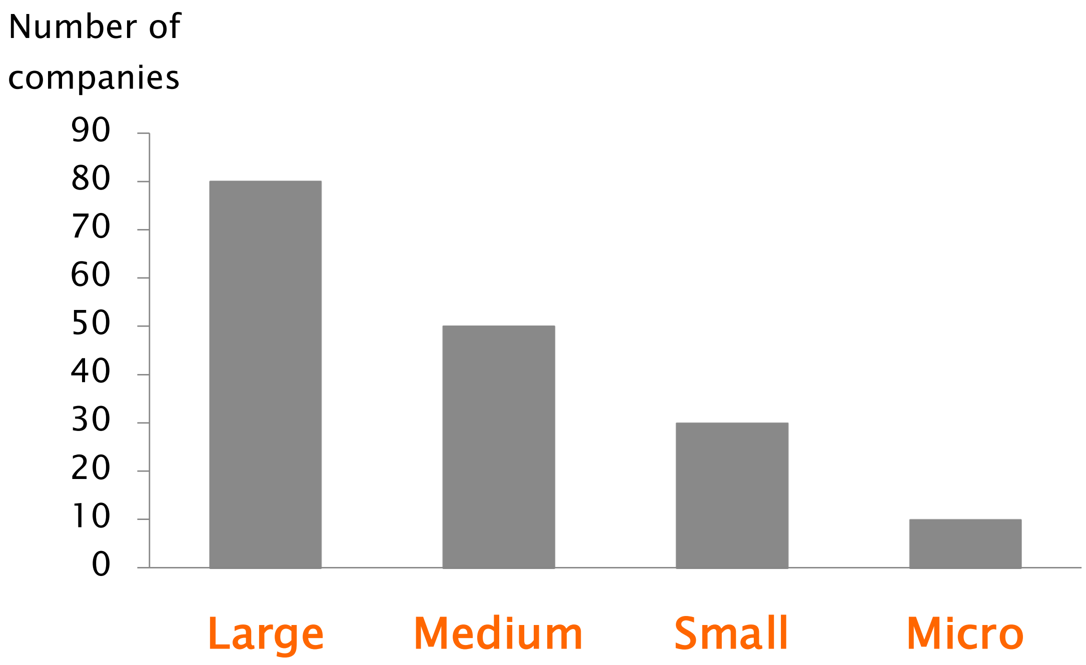


# Categorical encoding attributes

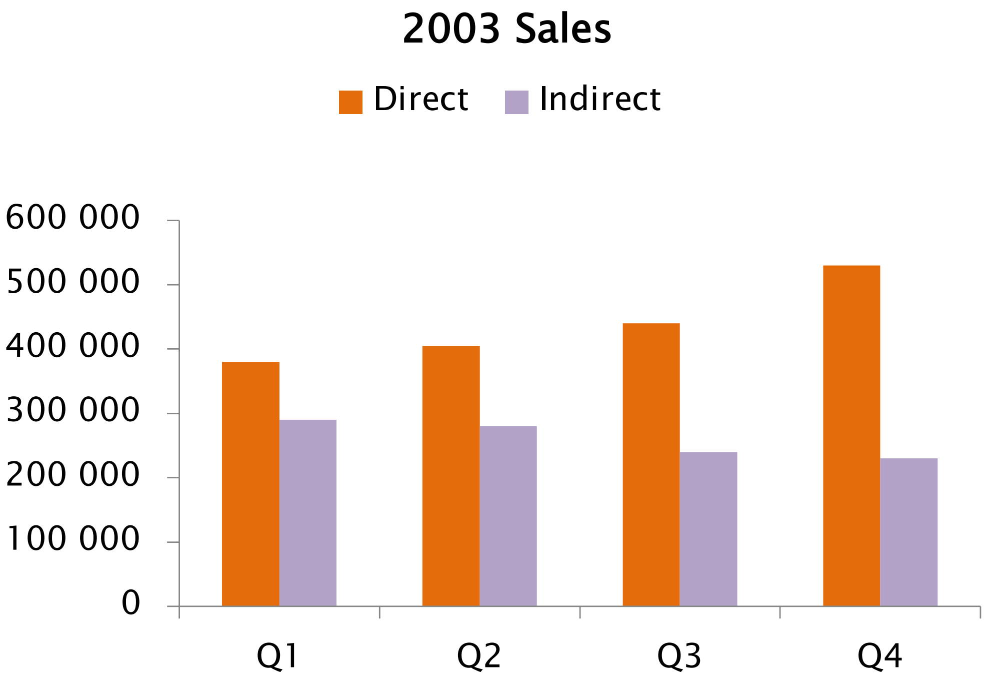
- Encoding of categorical levels
  - Position (along an axis)
  - Size
  - Color
    - Intensity
    - Saturation
    - Hue
  - Shape
  - Fill pattern
  - Line style

Ordinal

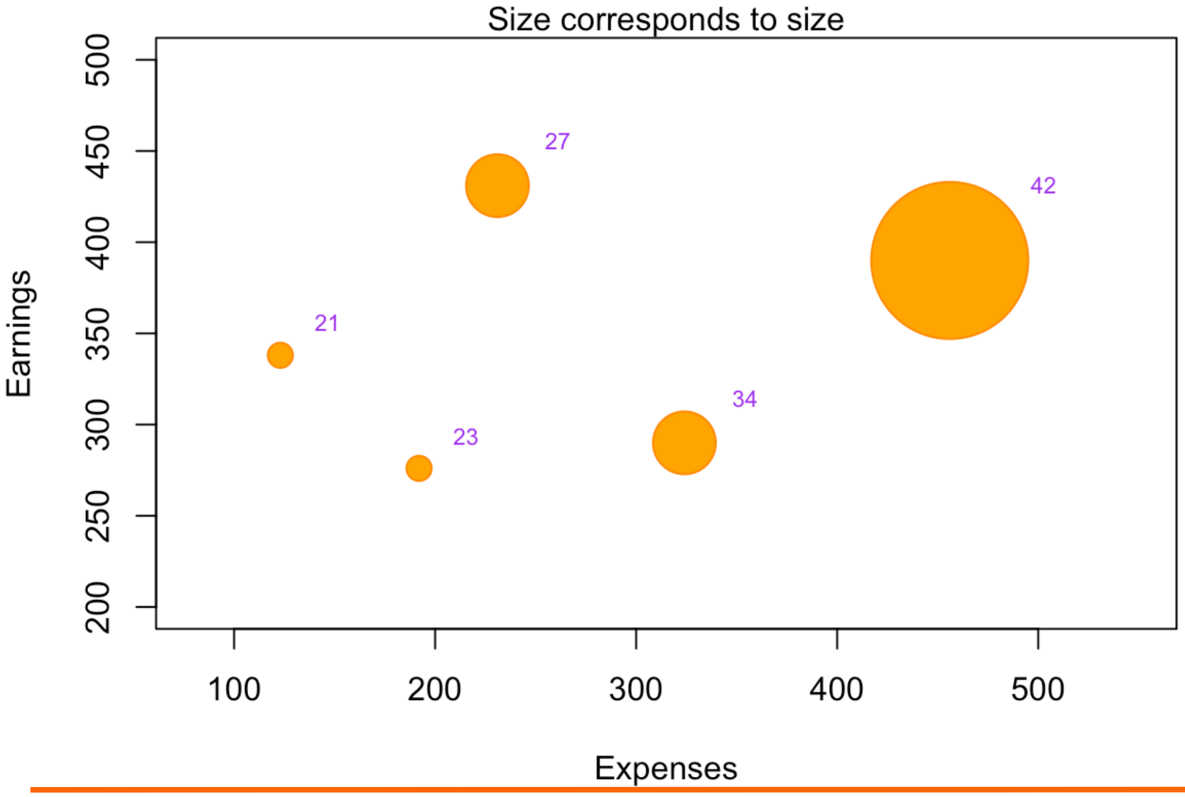
# Position



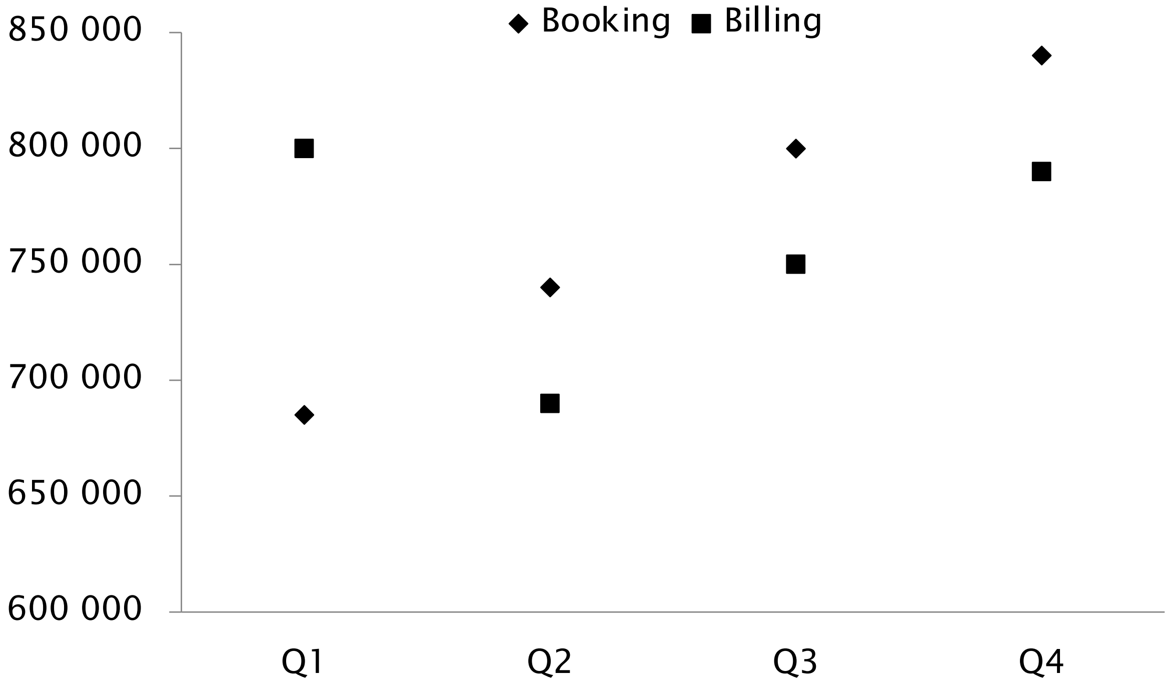
# Position × Color (hue)



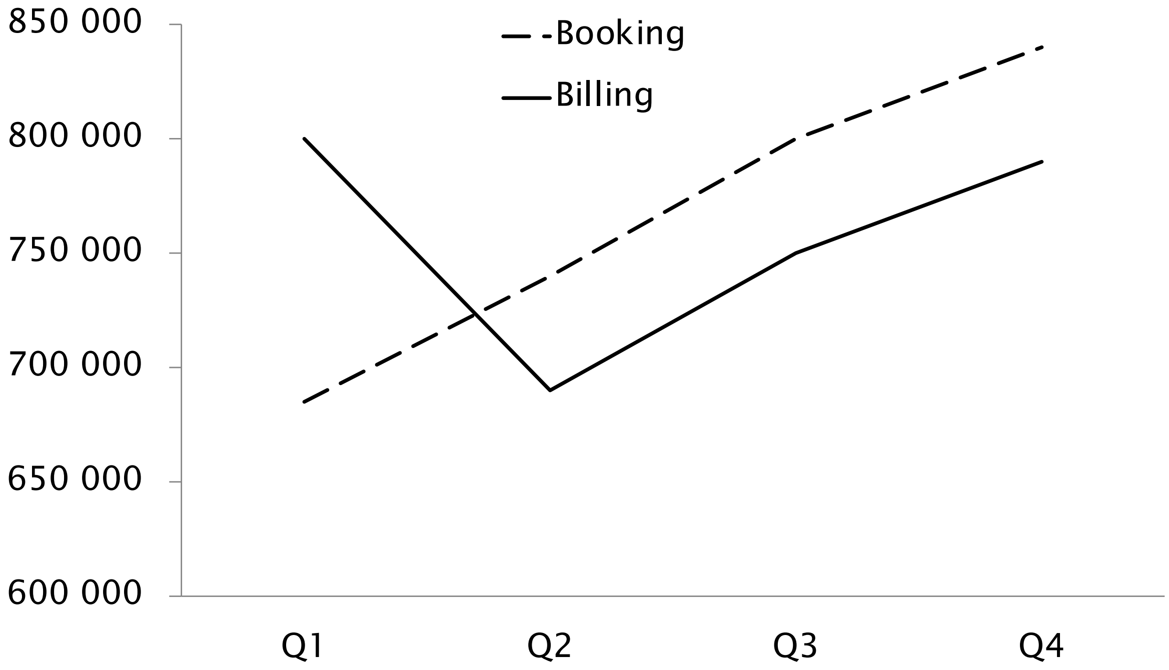
# Size



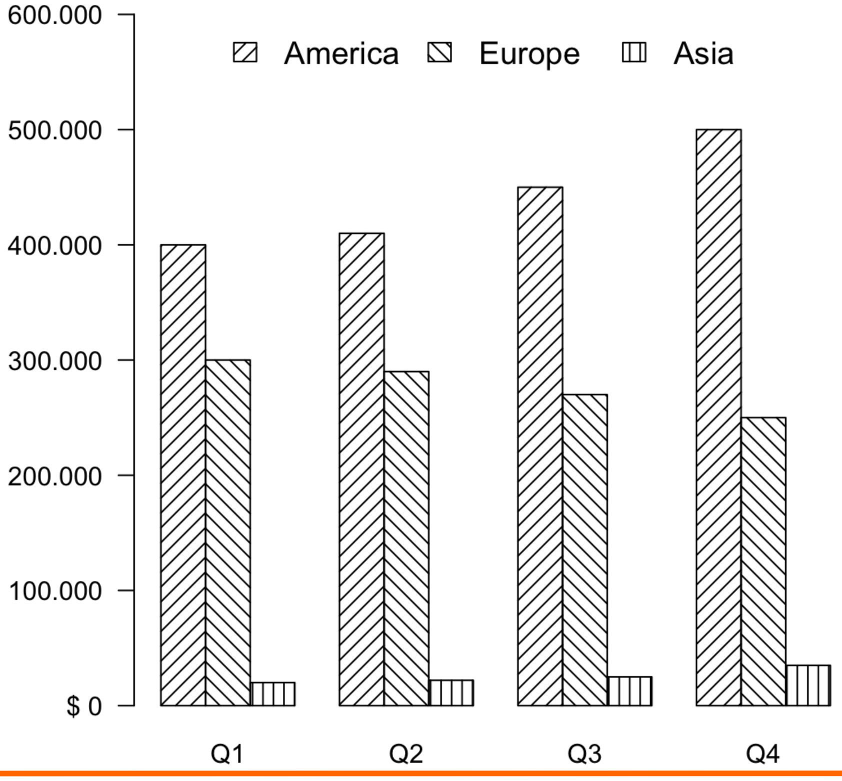
# Point shape



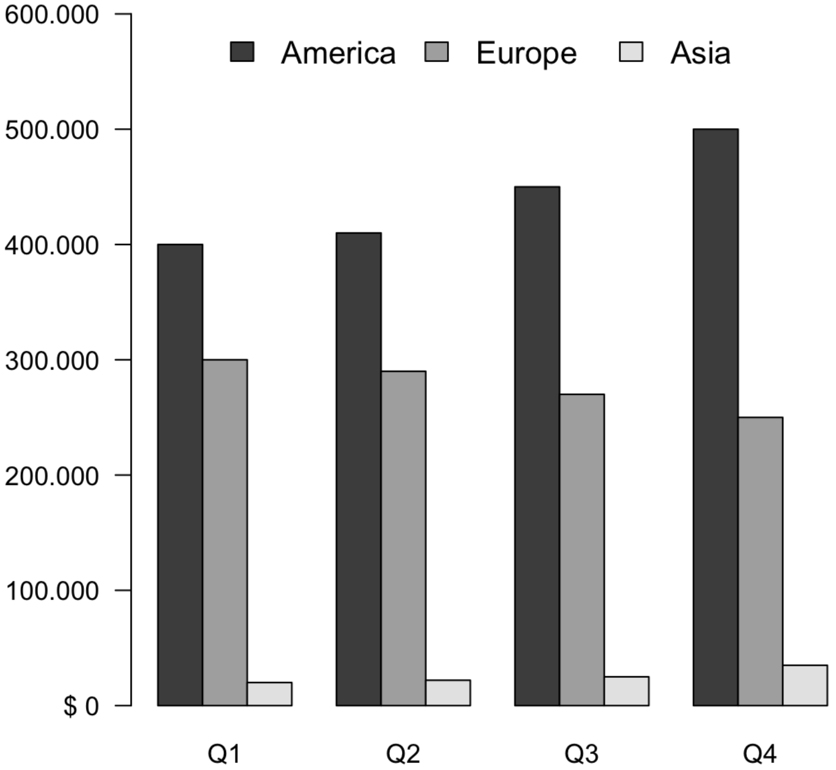
# Line style



# Fill Texture



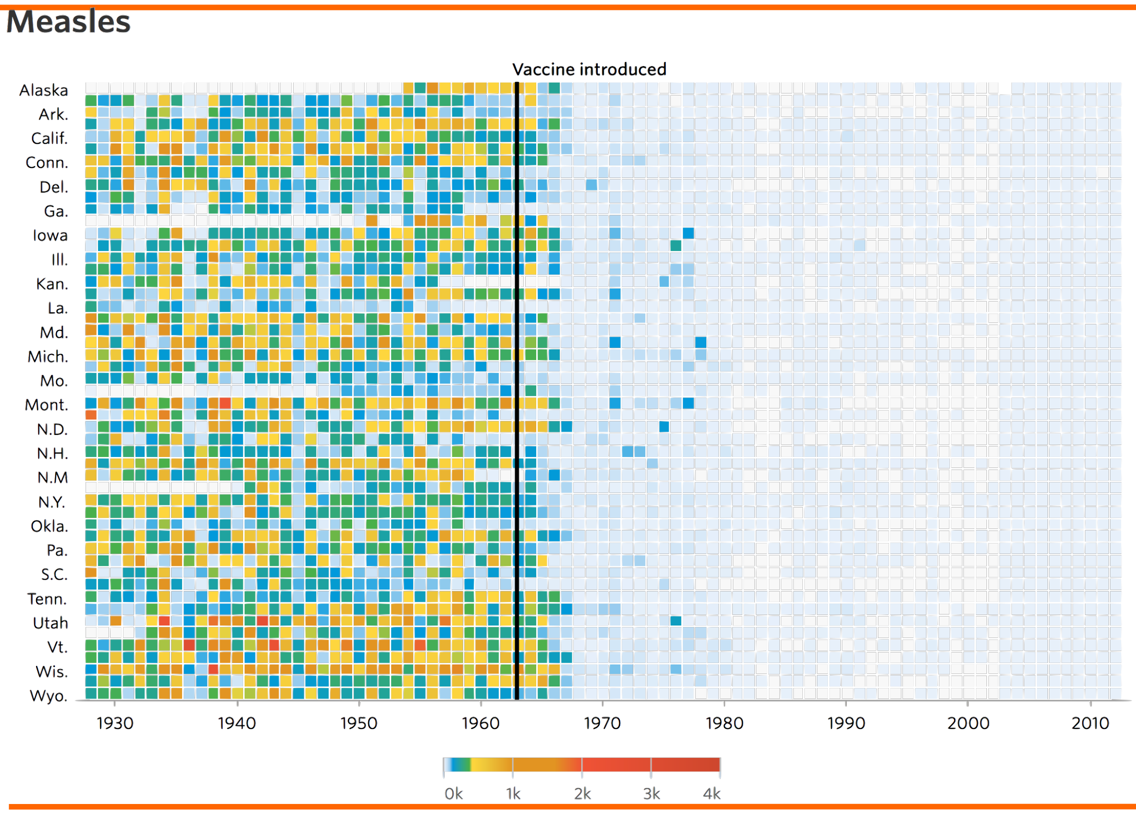
# Fill Gray Levels



# 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

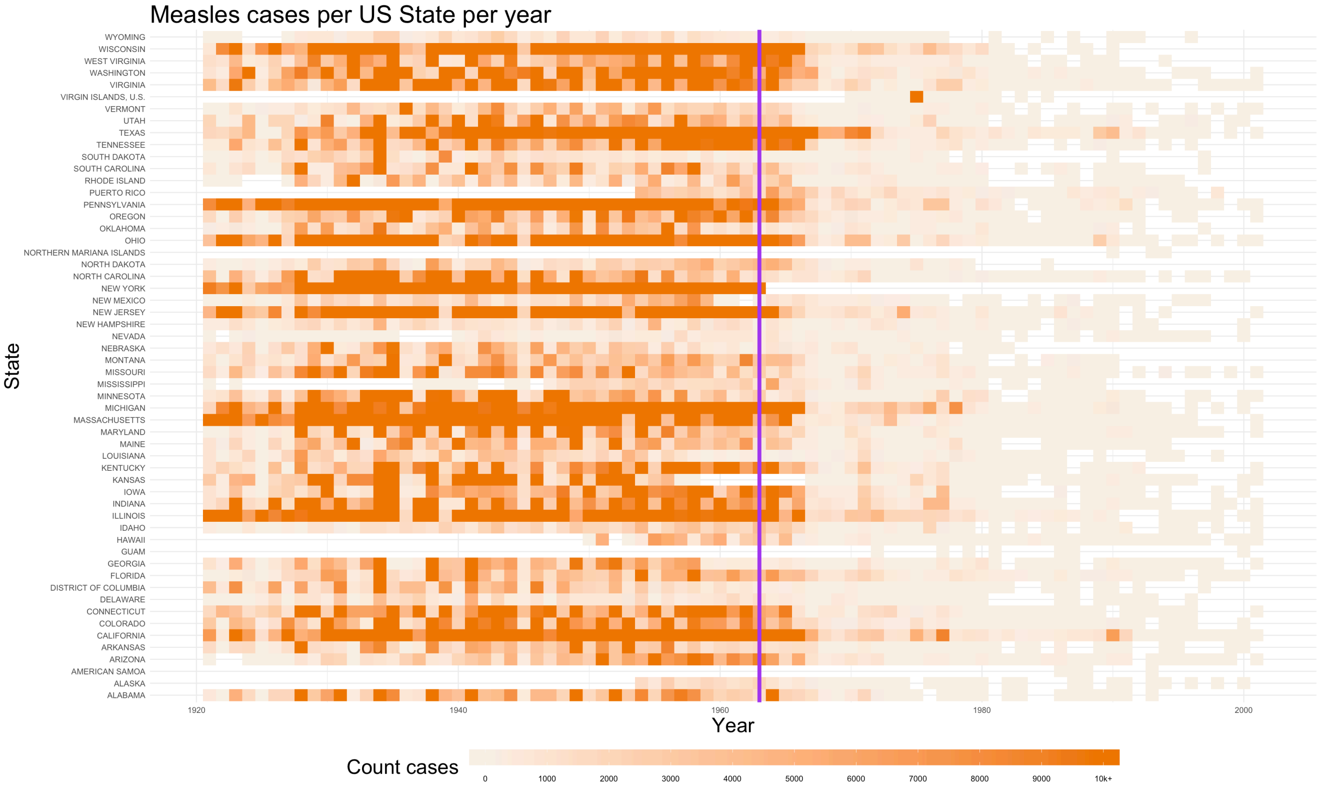


http://graphics.wsj.com/infectious-diseases-and-vaccines/

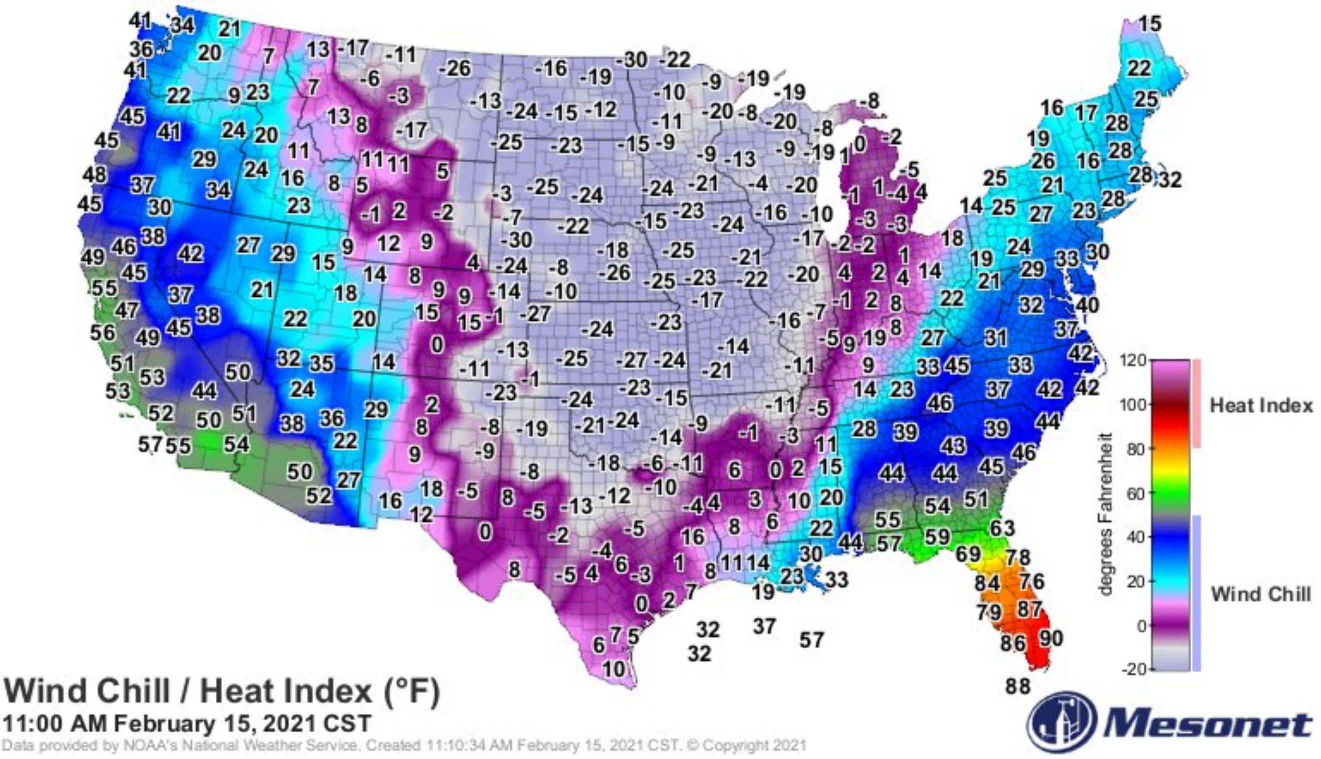
# Heatmaps

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

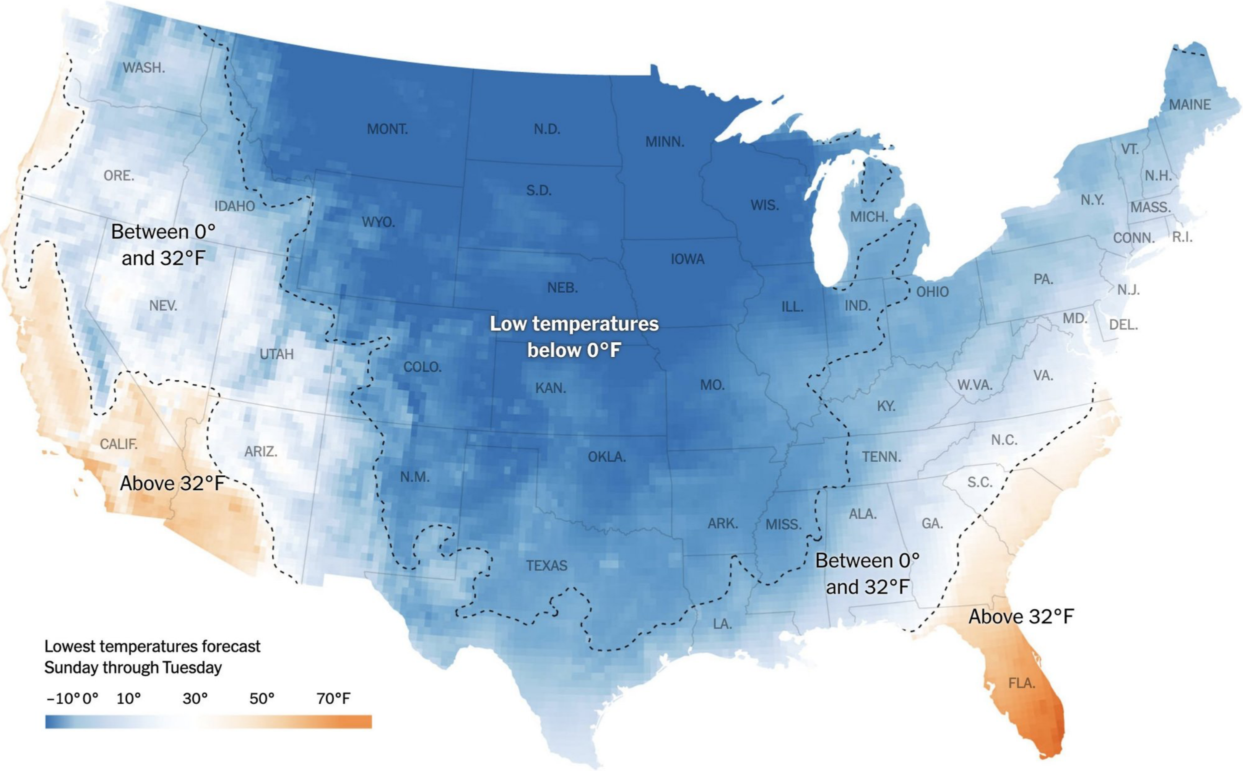
# Heatmaps



# Rainbow palette


120
100
80
60
40
20
0
-20
degrees Fahrenheit
Heat Index
Wind Chill
Wind Chill / Heat Index (°F)
11:00 AM February 15, 2021 CST
Data provided by NOAA's National Weather Service. Created 11:10:34 AM February 15, 2021 CST. © Copyright 2021
Mesonet

# Gradient palette



## 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

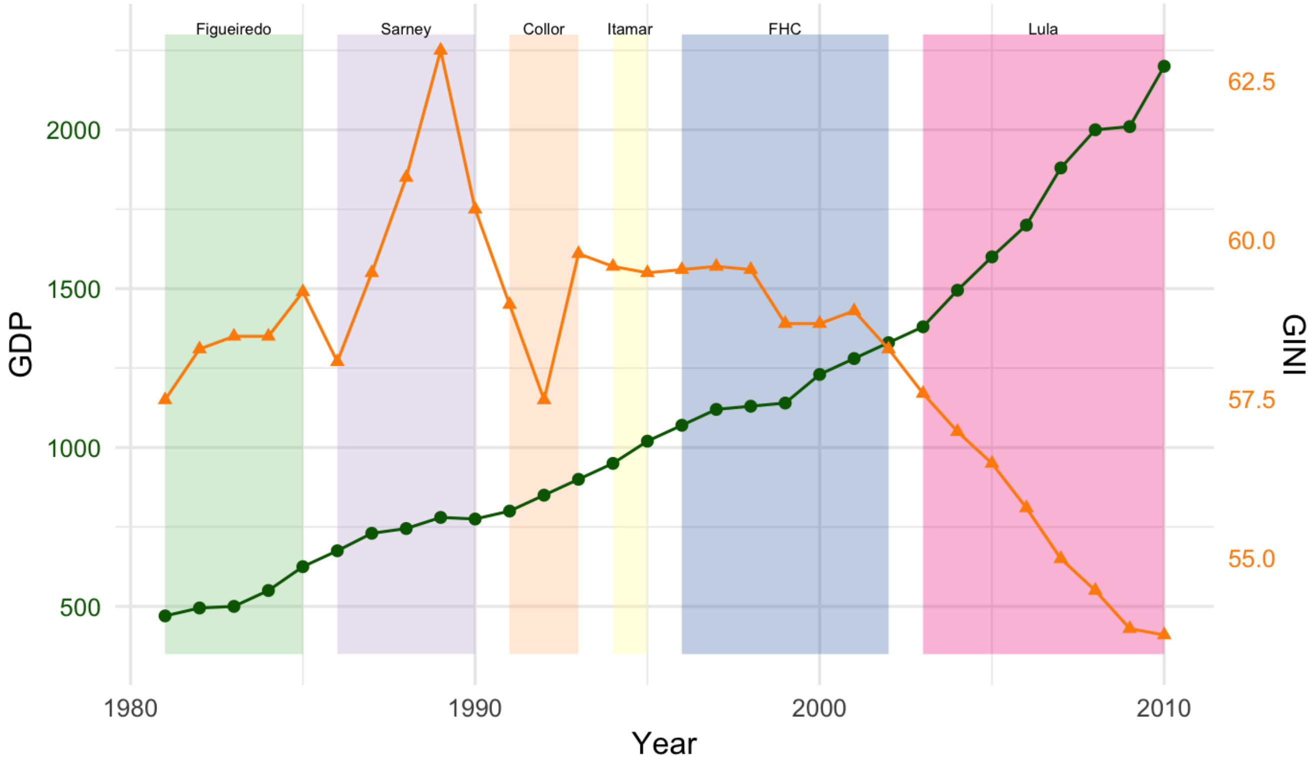
# Tick marks

- 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

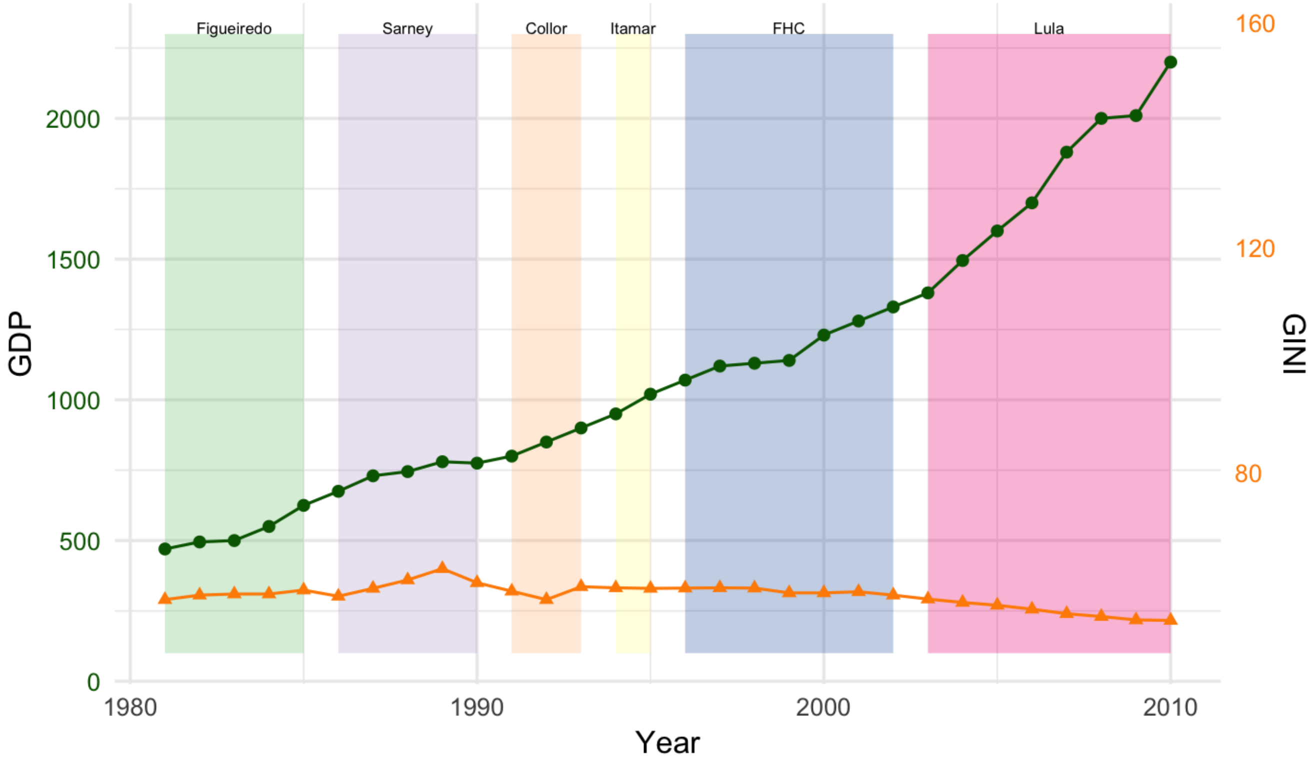
# Multiple variables

- 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

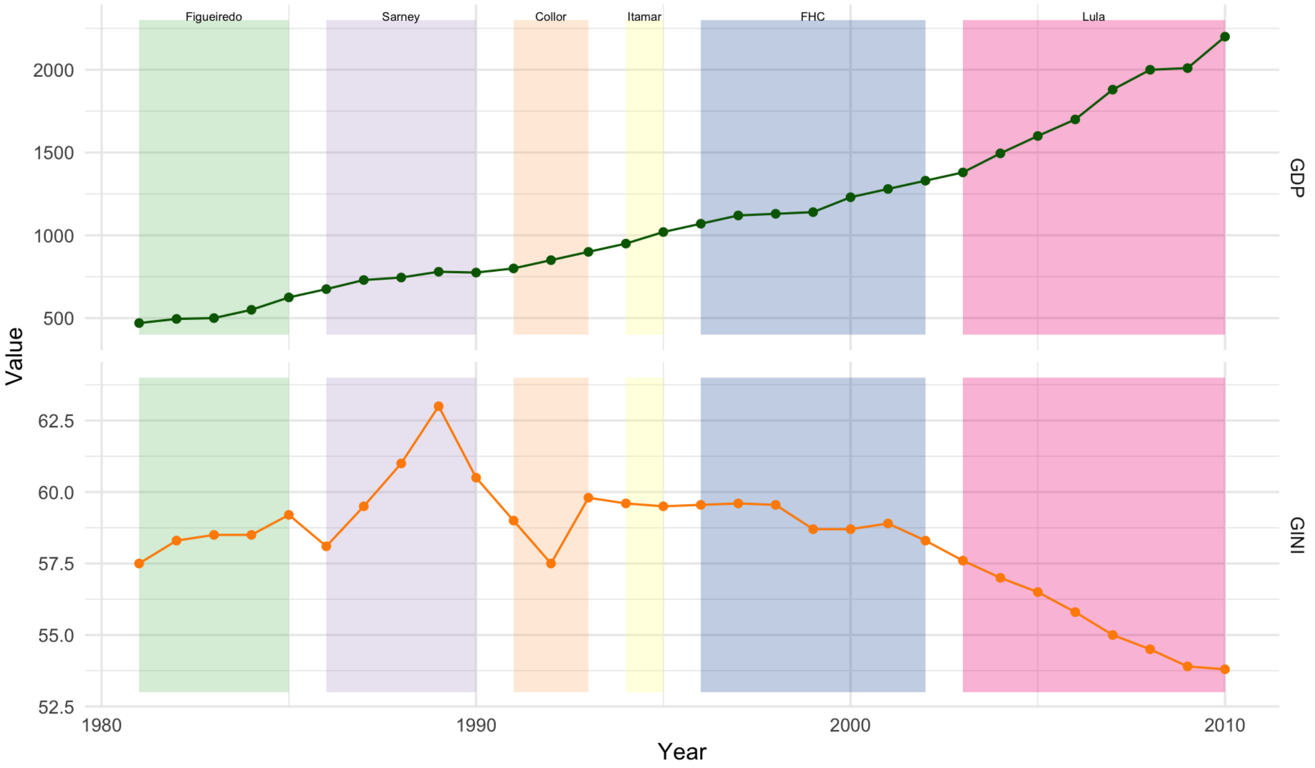
# Double scale



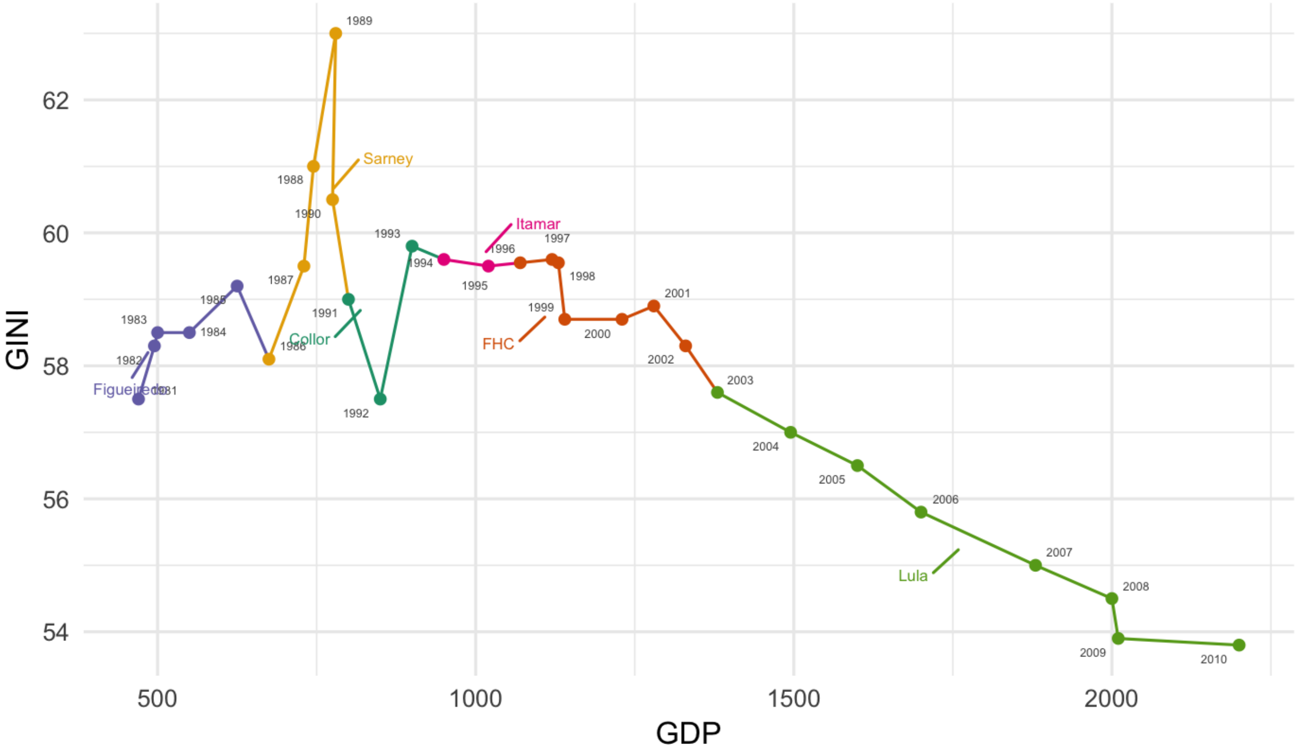
# Double scale (alternative)



# Multiple graphs



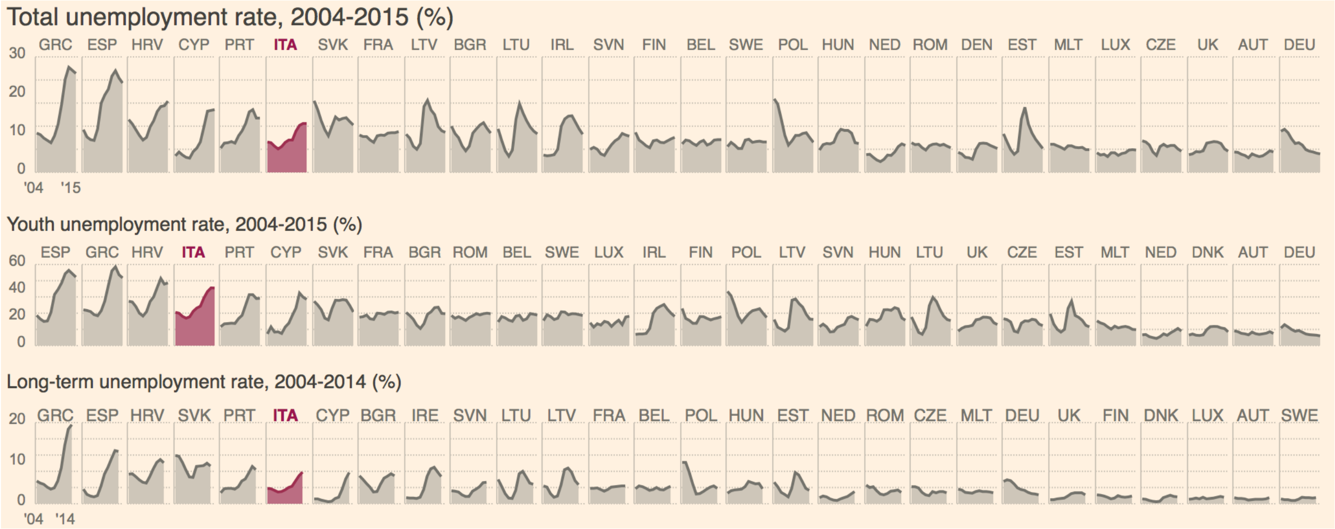
# Path



# 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



FT EU unemployment tracker
http://blogs.ft.com/ftdata/2015/04/17/eu-unemployment-tracker/

# Small multiples

- Consistency
  - Same scale
  - Same categorical levels
  - Same ordering of categorical levels
- Arrangement
  - Align axis that involve comparison
    - Possibly along a matrix

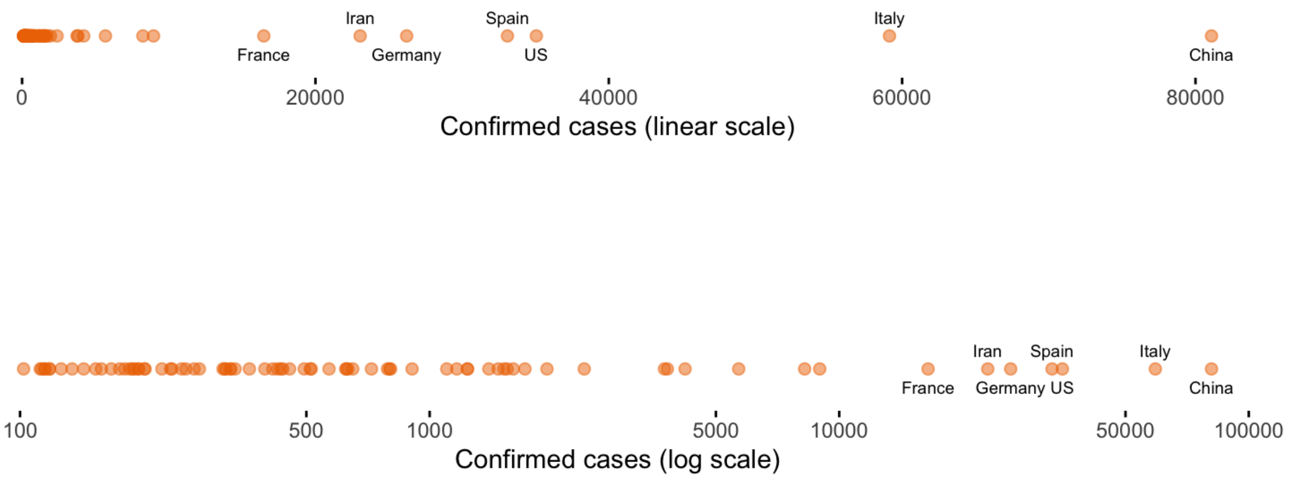
# Trellis

- 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

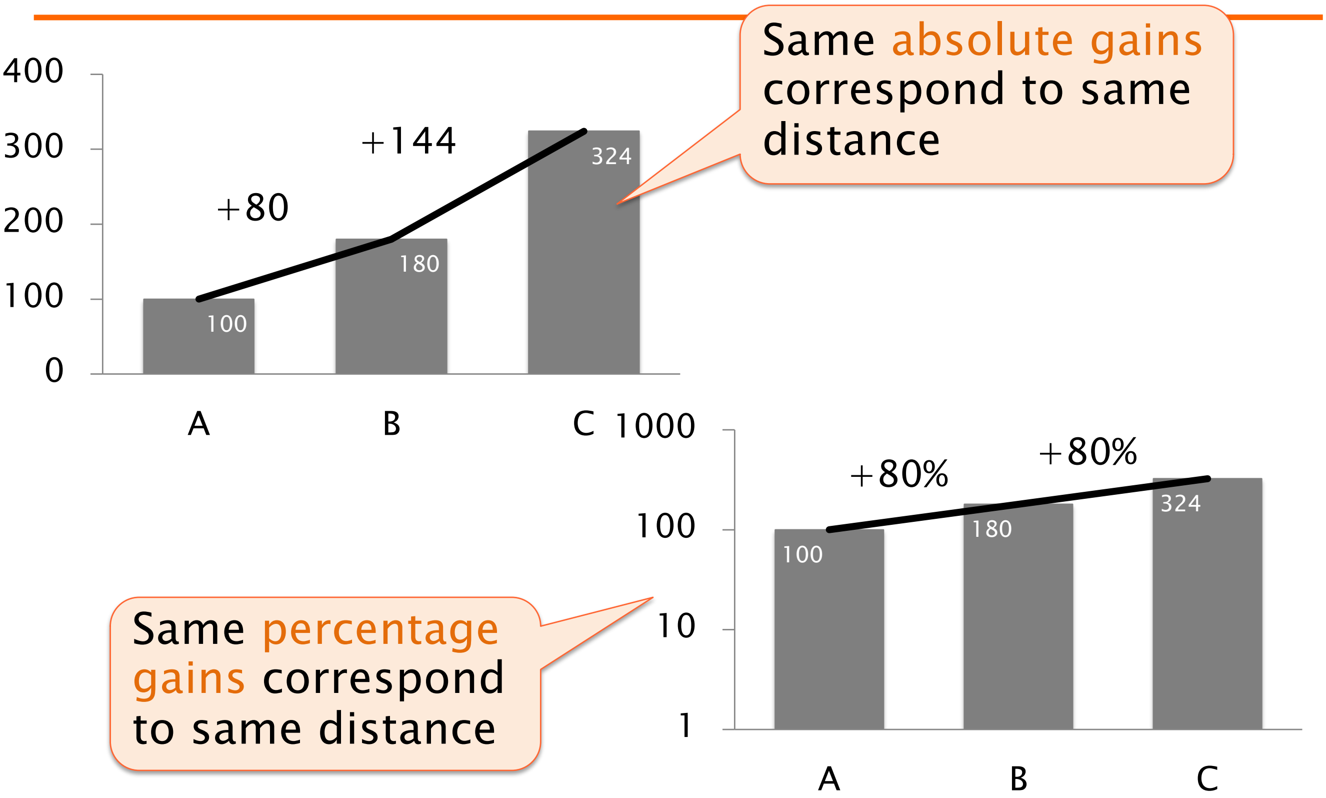
# Log scale

- Reduce visual difference between quantitative data sets with significantly wide ranges
- Differences are proportional to percentages
- Constant percentage increase correspond to a line

# Log scale – wide range



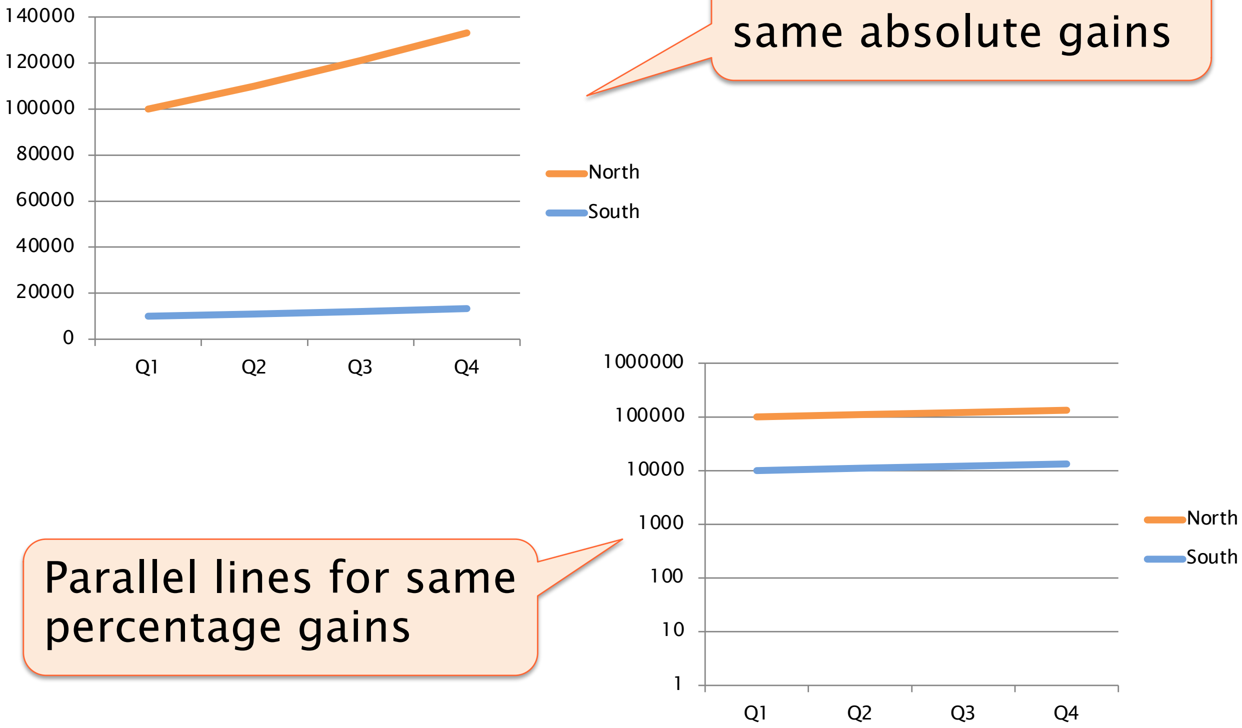
# Log scale – differences



Same absolute gains correspond to same distance

Same percentage gains correspond to same distance

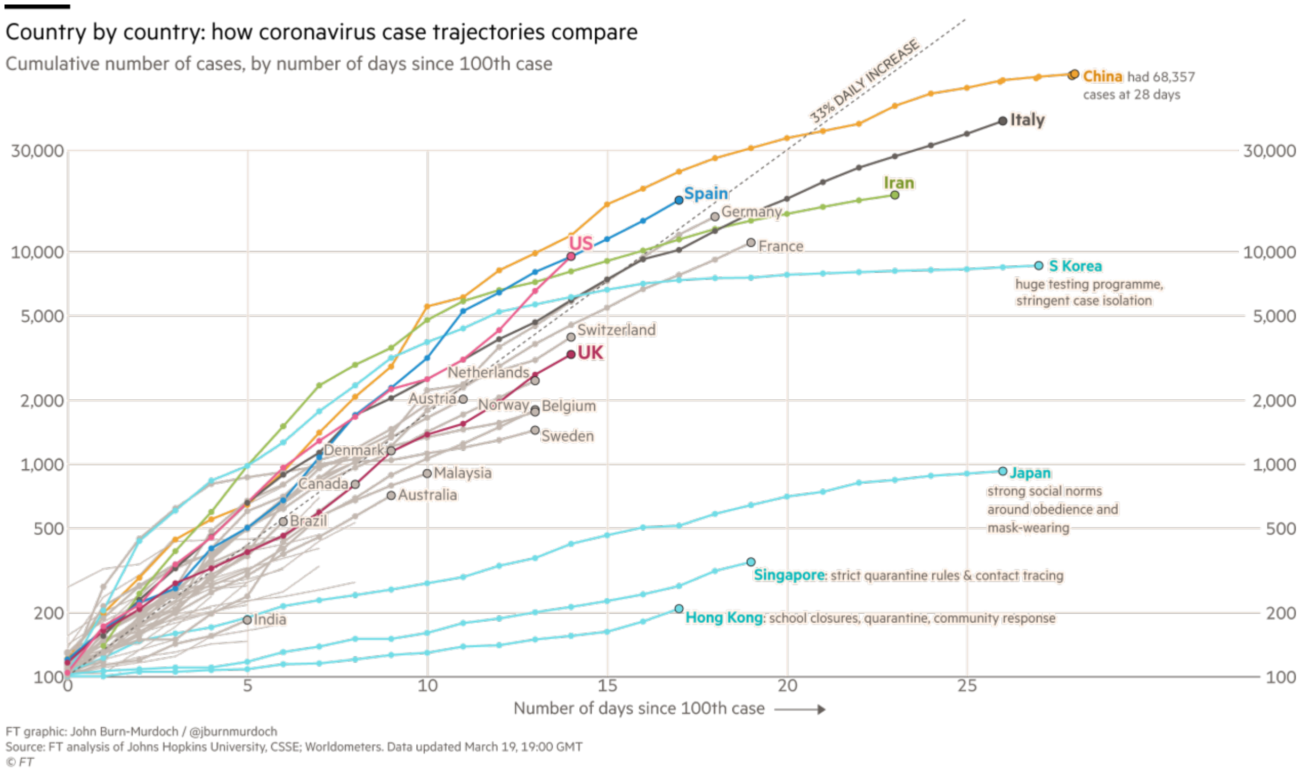
# Log scale – variation



Parallel lines for same absolute gains

Parallel lines for same percentage gains

# Log scale



Country by country: how coronavirus case trajectories compare

Cumulative number of cases, by number of days since 100th case

FT graphic: John Burn-Murdoch / @jburnmurdoch

Source: FT analysis of Johns Hopkins University, CSSE; Worldometers. Data updated March 19, 19:00 GMT

© FT

https://www.ft.com/content/a26fbf7e-48f8-11ea-aeb3-955839e06441

# 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

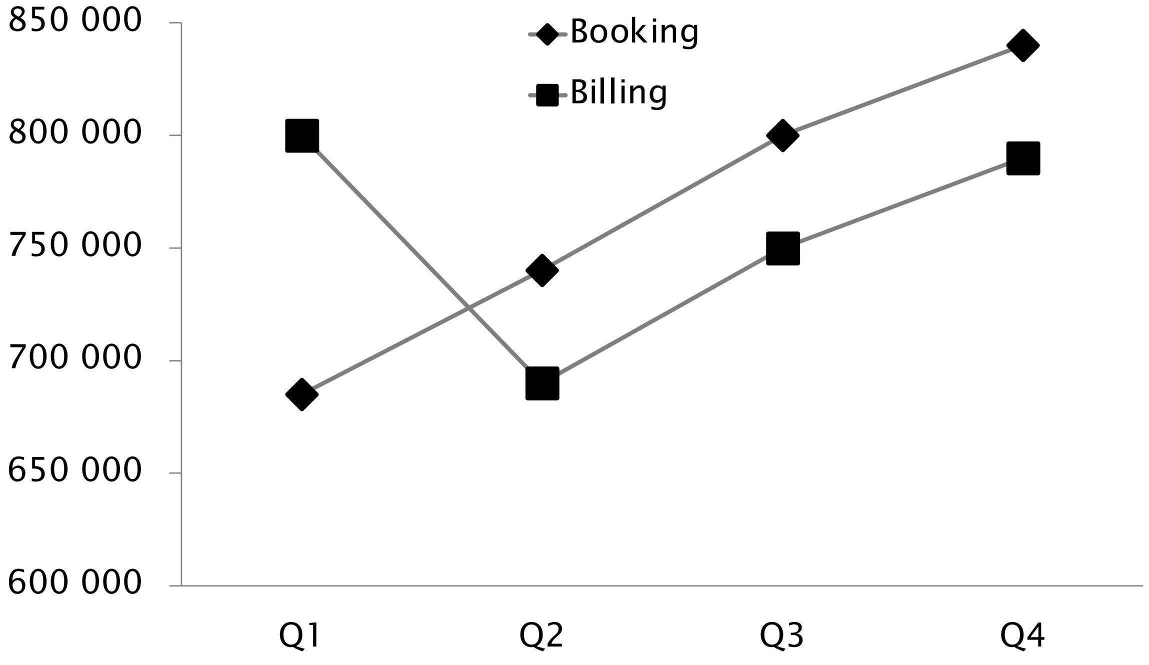
# Legends

- 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

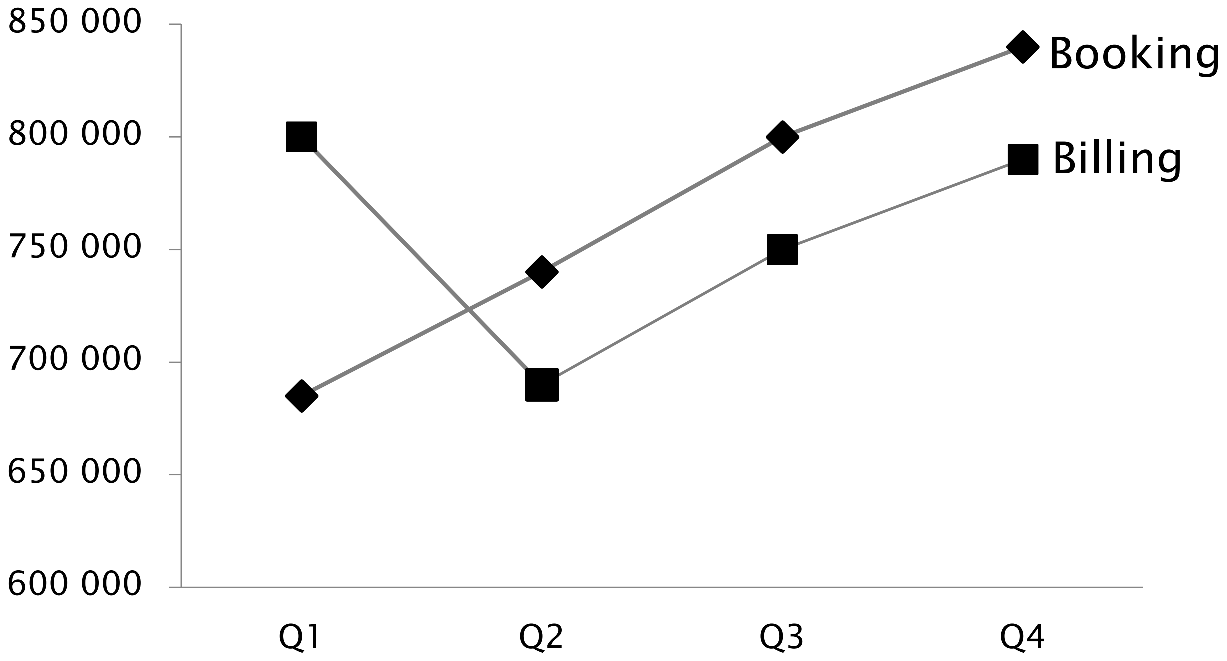
# Legends

- 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

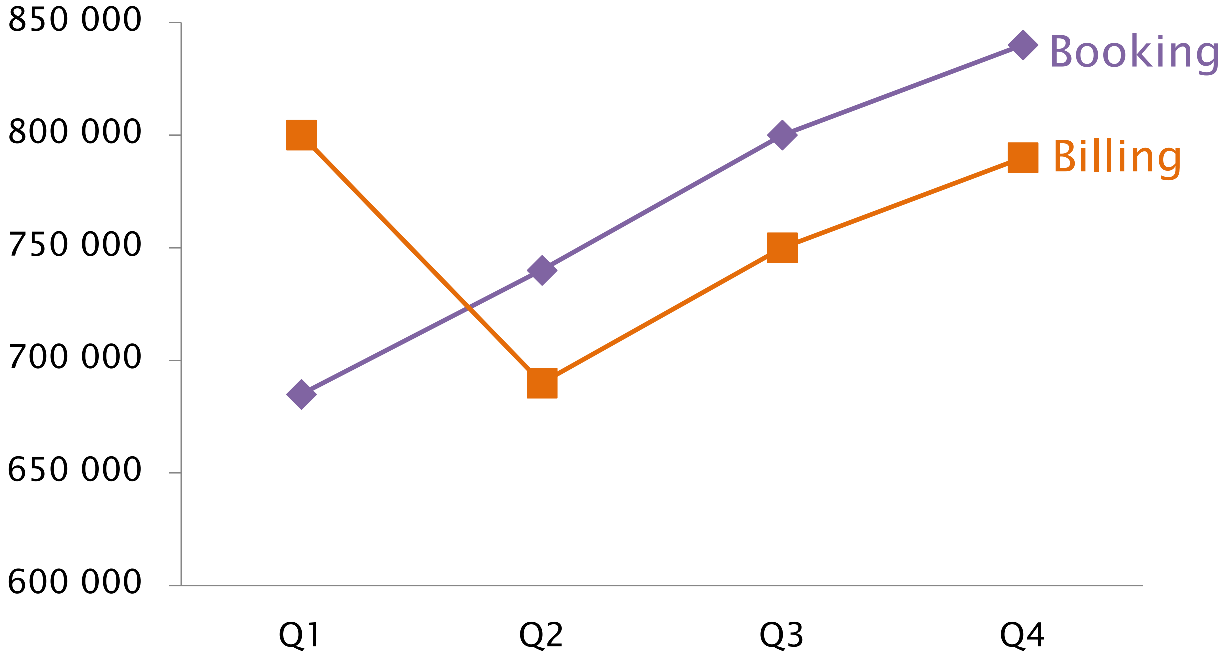
# Legend



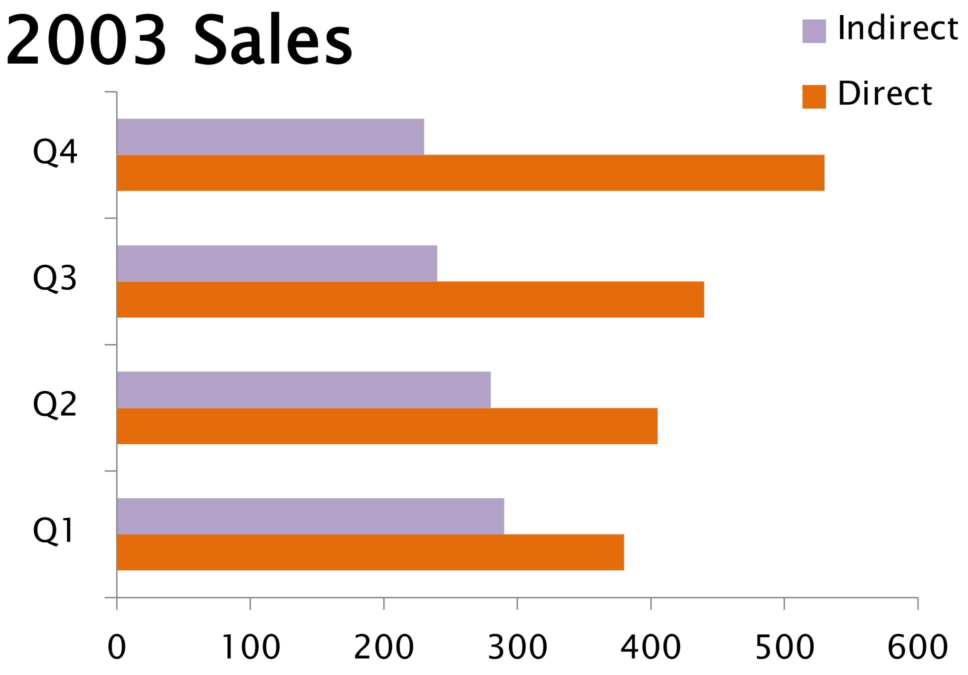
# Direct labeling



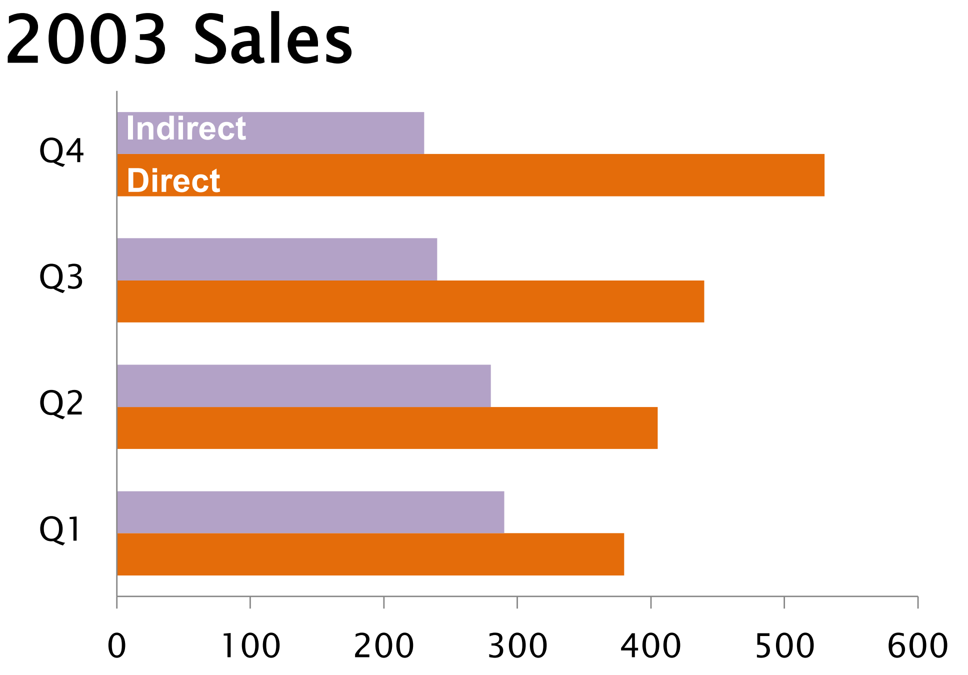
# Direct labeling and color



# Legend



# Direct labeling



# 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

# 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

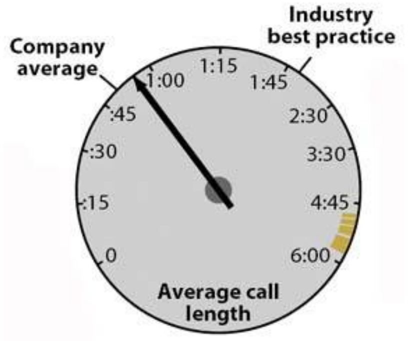
# Dashboard

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

# Provide context for data

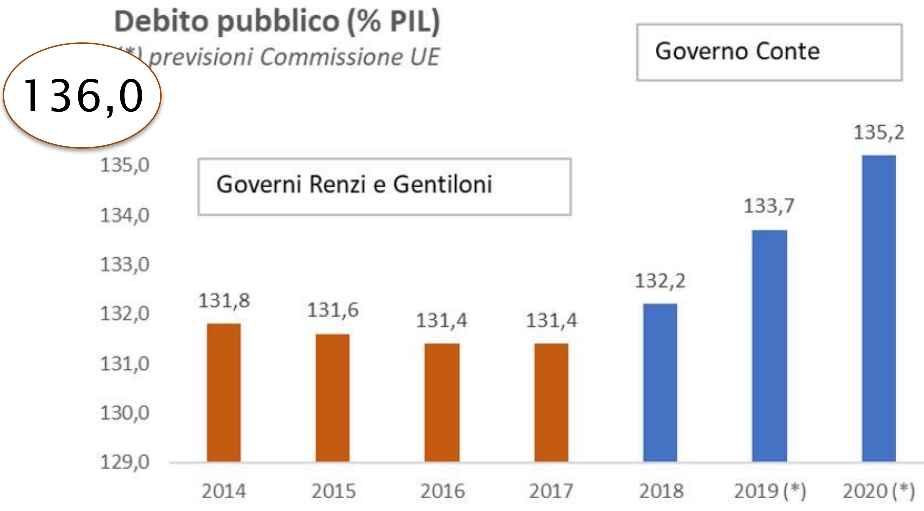
- References allow judging the data

7,822
YTD Units



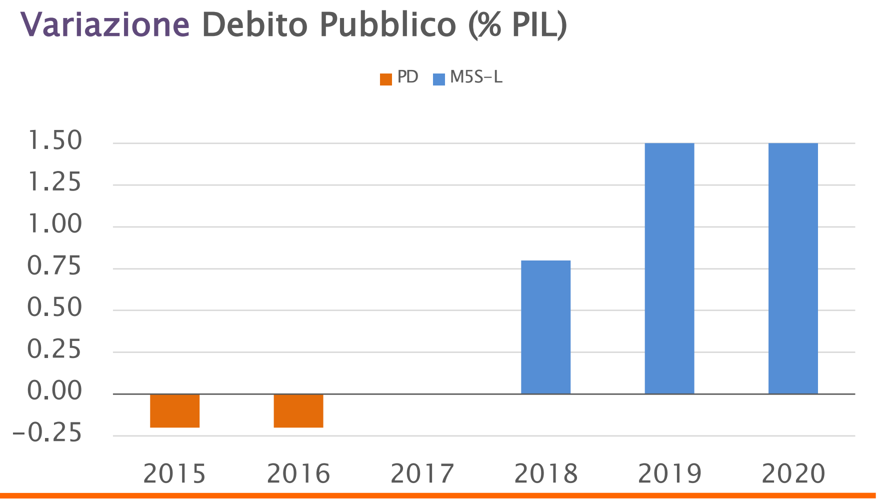
# Use appropriate detail

- Typical counterexamples
  - Dates with seconds detail
  - Decimals



# Use the right measures

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



# 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

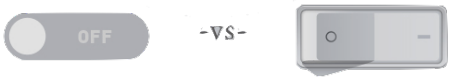
- Graphs
  - Visual information

# Avoid decorations

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

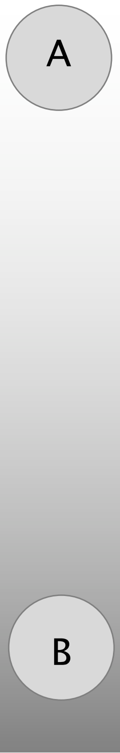
# Avoid decorations

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



# Avoid decorations

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



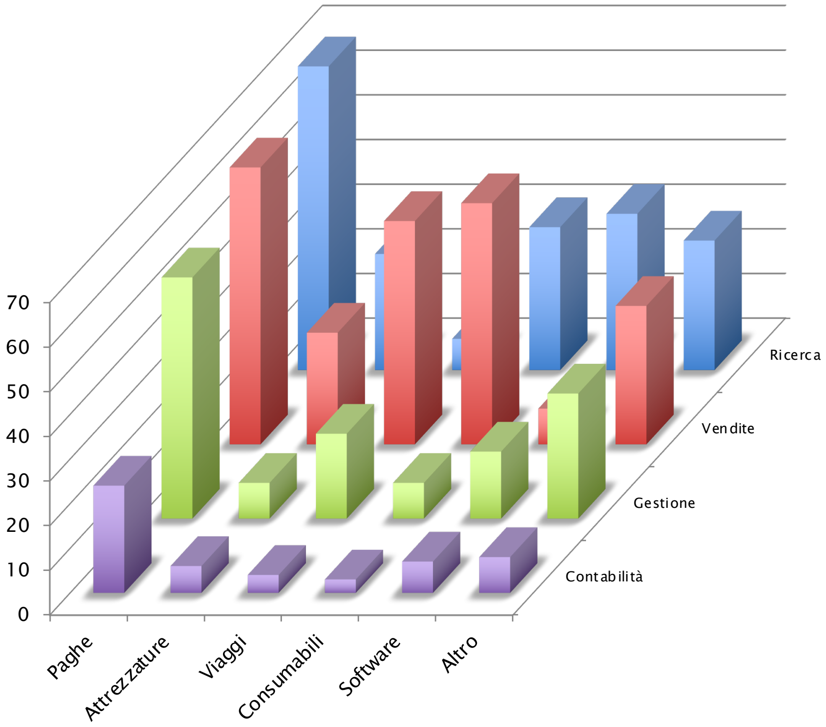
# Avoid decorations

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

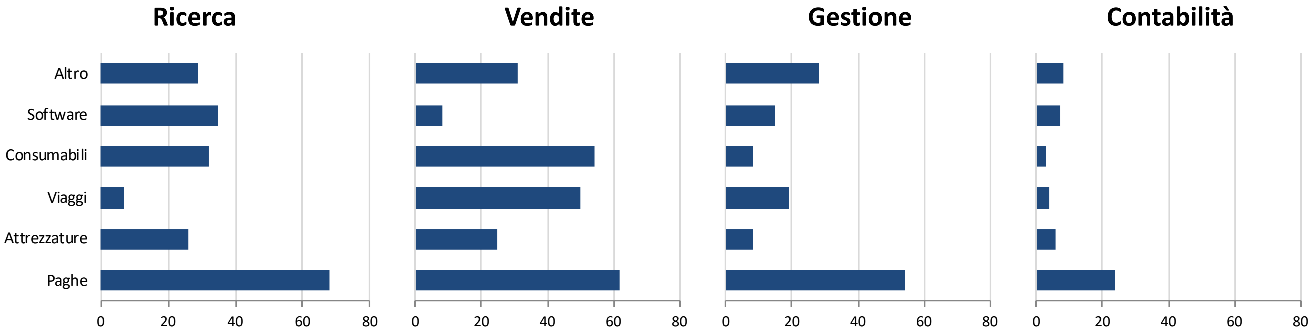
# 3D diagrams

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

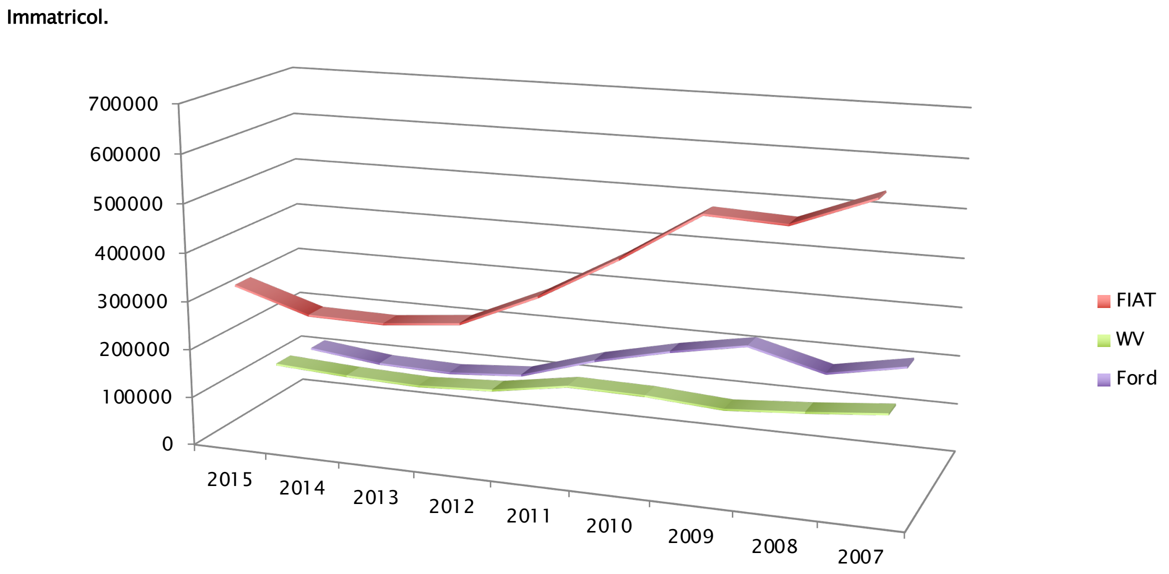
# Encoding 3D



# Encoding 3D → 2D

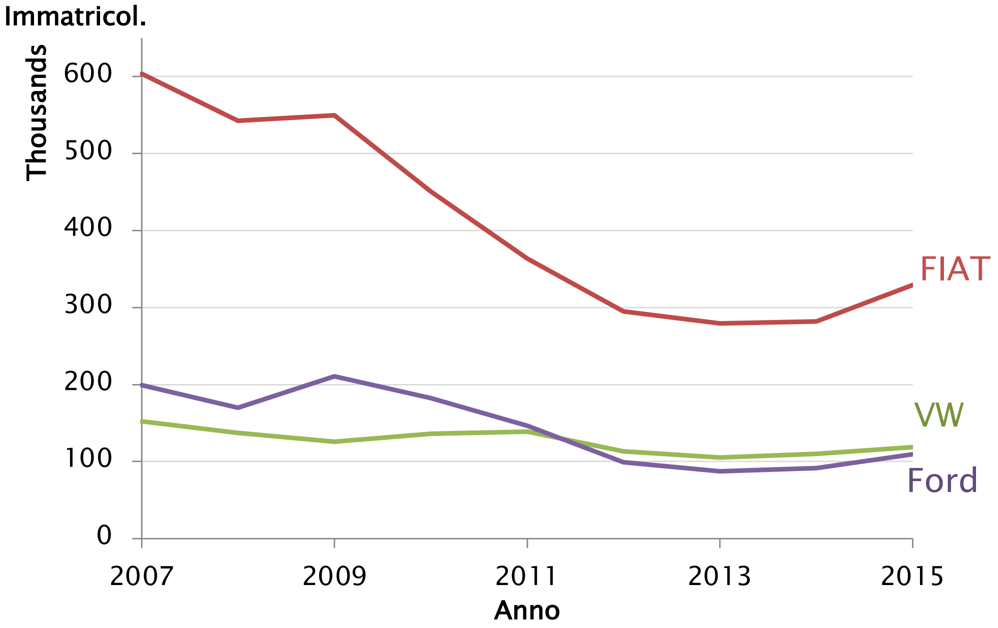


# Decorative 3D



# Decorative 3D → 2D

## Immatricolazioni auto per marchio sul mercato italiano



# 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