Python and SQL: intro / SQL platforms

Ewa Weychert

Class 7: Visualization, "A picture is worth a thousand words"

Goals of today's class

1. Why is data visualization important?
2. Data types vs. chart types
3. Bad examples of visualization
4. How to fix bad visualization examples
5. Good examples of visualization

What are one-dimensional charts?

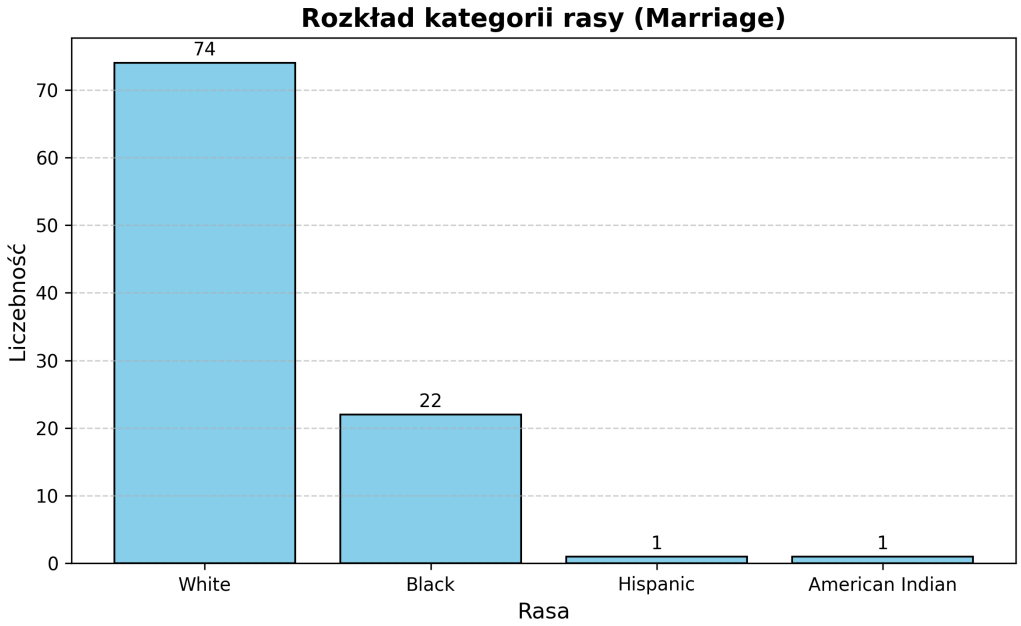
- One-dimensional charts present the distribution of data for a single variable.
- The variable may be:
  - **categorical** (e.g., race, gender),
  - **quantitative** (e.g., age, weight).
- The goal is to show how often particular values or intervals occur.

Categorical variable

- Typical ways of visualization:
  - bar chart,
  - pie chart,
  - less commonly – tree map.
- Example: distribution of marriage participants by race in Mobile County, Alabama.

Bar chart

- Shows the count or percentage for each category.
- Colors, axis labels, and titles can be modified.
- Charts can be sorted in ascending or descending order to make interpretation easier.
- Numeric values are often added directly on the bars.

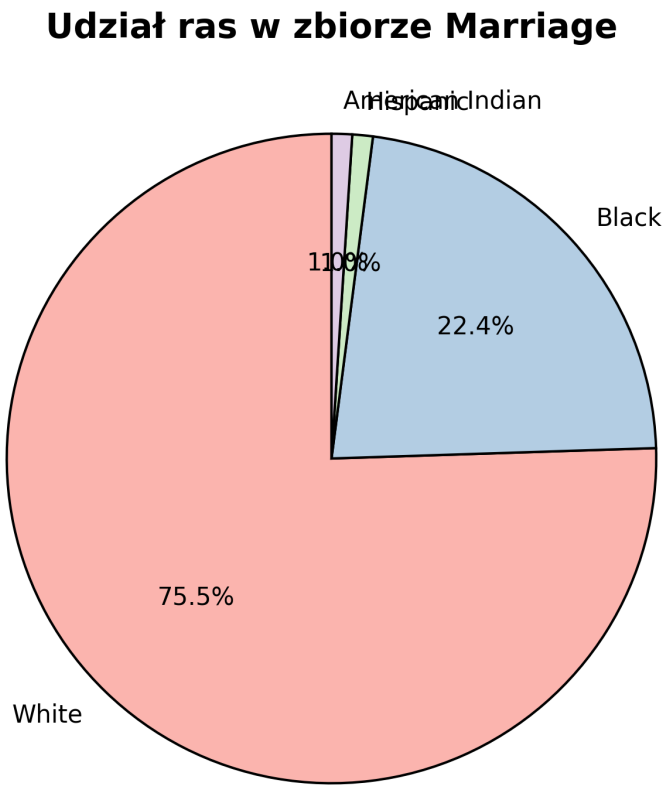


Percentages and ordering categories

- Bars can represent **percentages** instead of counts.
- Categories can be ordered by frequency.
- Labels may overlap — in such cases:
  - rotate the axes (*horizontal bar chart*),
  - rotate the labels,
  - or shorten/wrap the label text.

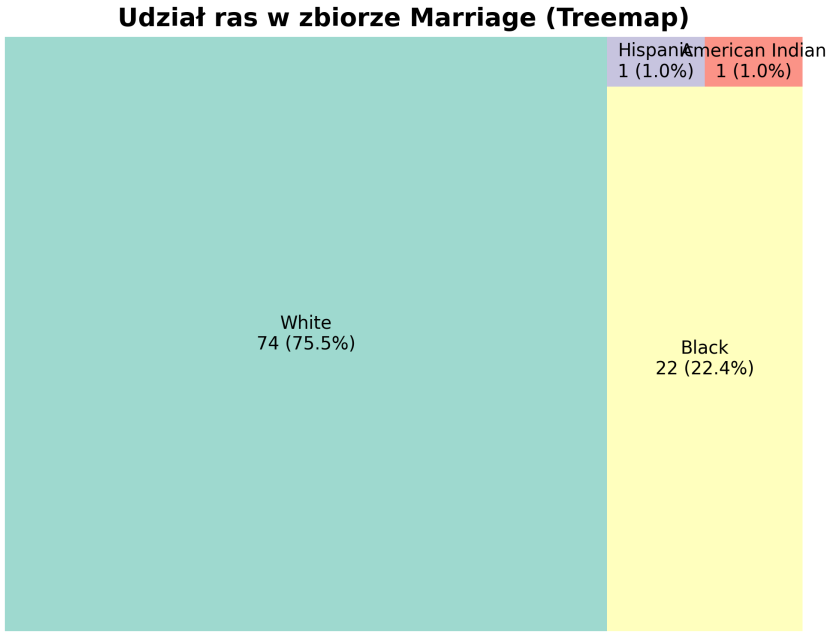
Pie chart

- Used to show parts of a whole (percentage share of categories).
- Better to use only when the number of categories is small.
- People compare bar lengths better than angles — therefore a bar chart is often preferable.
- Percentage labels can be added for clarity.



Treemap

- A treemap shows the share of categories using rectangles with proportional areas.
- It works well when there are many categories.
- Each rectangle represents a category, and its area corresponds to its numerical value.
- It is easy to spot dominant groups.

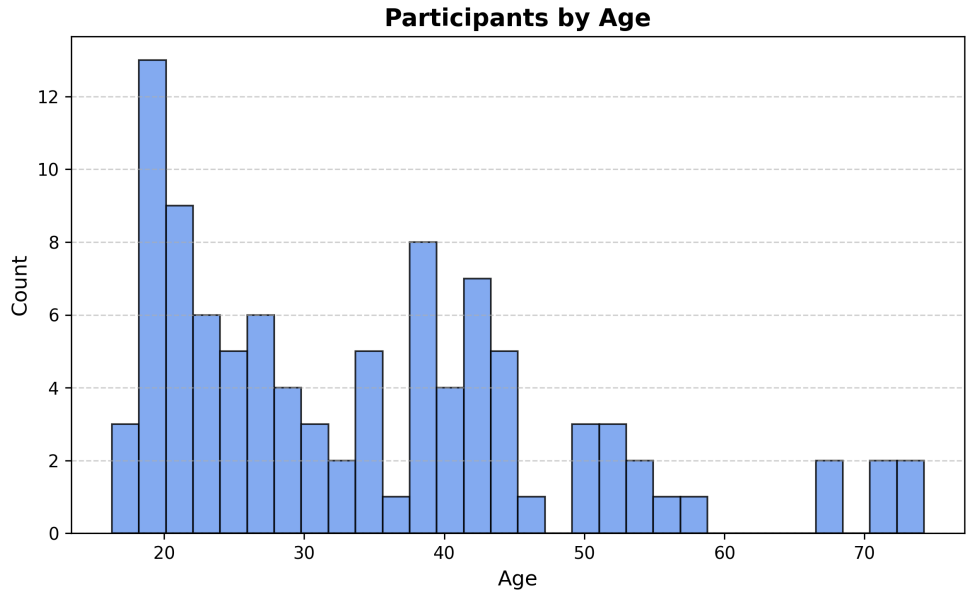


Quantitative variable

- To visualize the distribution of a single quantitative variable, one uses:
  - a histogram,
  - a density plot (kernel density plot),
  - a dot plot.

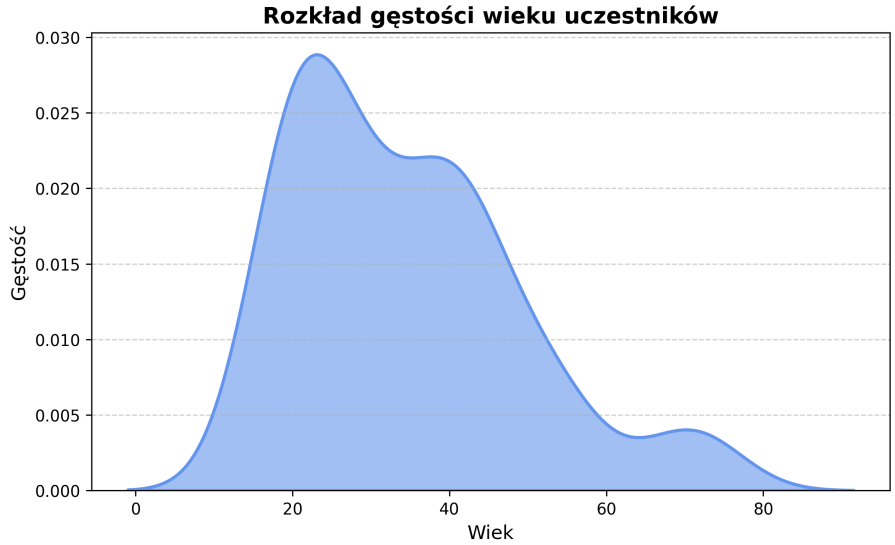
Histogram

- Shows how often values occur within specified intervals.
- The number of intervals (*bins*) affects the appearance of the plot.
- The fill color and border color can be customized.
- A histogram can display counts or percentages.
- It helps identify the shape of the distribution (e.g., normal, skewed, multimodal).



Density plot (Kernel density plot)

- Shows a smoothed shape of the distribution of a continuous variable.
- It is an alternative to a histogram – instead of bars, it shows a continuous curve.
- The area under the curve sums to 1.
- The **bandwidth** parameter controls the degree of smoothing:
  - smaller — more detailed, less smooth plot,
  - larger — more smoothed, less detailed.



Source: *Marriage.csv* data, own elaboration

Summary

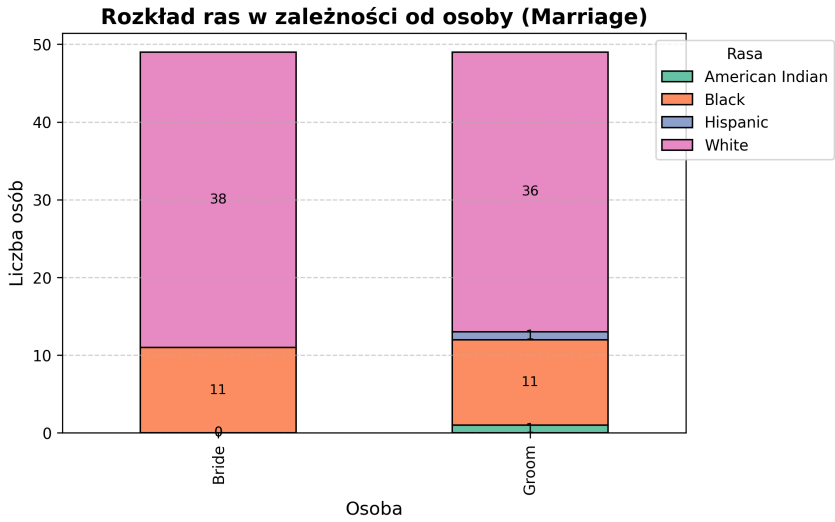
- The choice of chart depends on the data type and the goal of the analysis.
- For categorical data: bar charts, pie charts, treemaps.
- For quantitative data: histograms, density plots.
- Good practices:
  - maintain readability,
  - add labels and units,
  - avoid unnecessary embellishments.

4.1 Categorical vs. categorical

- Relationships between two **categorical** variables are most often shown using:
  - **bar charts**: *stacked*, *grouped*, *segmented* (each bar = 100%).
  - Less commonly: **mosaic plot**.
- The choice depends on whether we are comparing *counts* or *proportions*.

Stacked bar chart

- Shows how the **distribution of races** changes depending on the **person** (*Bride / Groom*).
- Each bar corresponds to one person category, and the colors show the share of each race.
- We see that **White individuals dominate** in both groups, while other races appear only sporadically.
- A *stacked bar* chart effectively illustrates both total counts and the internal structure of each category.
- With a larger number of categories, one may consider the percentage version (*100% stacked bar chart*) to compare proportions.

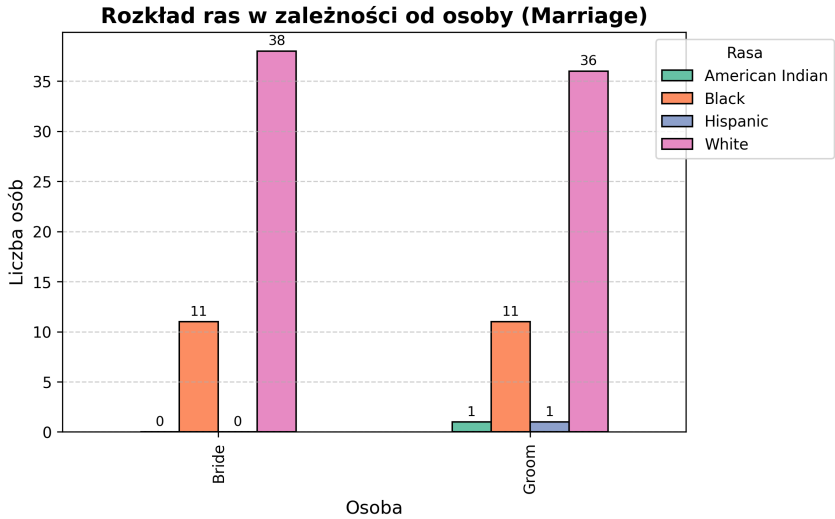


Data source: *Marriage.csv*

Numbers inside the bars = number of individuals in each group.

Grouped bar chart

- For each category A, we draw side-by-side bars for the levels of B.
- This makes it easier to **compare categories of B** along the same axis height.
- Make sure to:
  - keep *zero-count* categories (so they don't disappear),
  - use a clear legend layout and consistent colors.
- The example shows the distribution of races (*race*) depending on the person (*Bride* / *Groom*) from the file `Marriage.csv`.

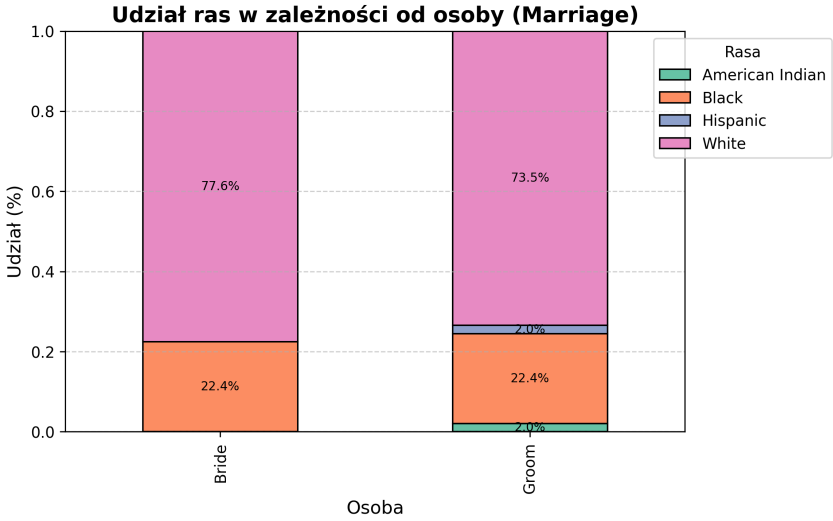


Data source: *Marriage.csv*

Numbers above bars = number of individuals in the group.

Segmented bar chart (each bar = 100%)

- The height of each bar is scaled to 100% — we show **proportions**, not counts.
- Ideal for answering the question: "What
- Makes it easier to compare proportions between levels of variable A, regardless of differences in sample size.
- In this example, we compare the **share of races** in the *Bride* and *Groom* groups.



Data source: *Marriage.csv*

Each bar = 100%, values = percentage shares.

Colors, labels, and category order

- **Category order** (reorder): set a logical order (e.g., 2seater → suv → pickup).
- **Category names**: change them to understandable ones (e.g., *f*, *r*, *4* → "front", "rear", "4×4").
- **Y-axis**: for 100% charts, use percentage labels (0, 20, . . . , 100).
- **Color scheme**: use qualitative (*qual*) palettes that are consistent and colorblind-friendly.
- **Background and grid**: a minimalist style improves readability.

Value labels on segments

- Adding
- Good practices:
  - round to 0–1 decimal places,
  - use text contrast with the segment background,
  - avoid cluttering labels for very small shares (use a hiding threshold).

When to use which chart?

- **I want to compare counts** between groups → *stacked* or *grouped*.
- **I want to compare proportions within groups** → *segmented* (100%).
- **Very many categories** → consider a *mosaic* plot or a contingency table with annotations.
- **Readability is crucial** → reduce the number of levels, merge rare ones, sort logically.

Interpretation tips

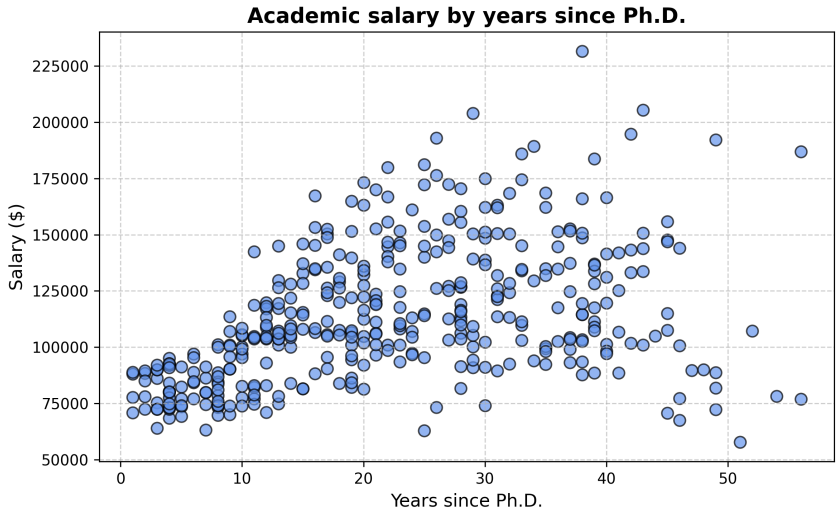
- Check for **dominant categories** and **exceptions** (e.g., the only type with 0 cases).
- Pay attention to **trend patterns** (e.g., increasing share of front-wheel drive).
- Remember **sample sizes** — a 100% chart does not show whether groups are large or small.
- Add **descriptions/legend/source** — it helps clarify the context.

4.2 Quantitative vs. quantitative

- The relationship between two quantitative variables is most often shown using:
  - a **scatterplot**,
  - a **line plot** (when one axis represents time).
- The goal is to assess the direction, strength, and shape of the relationship, and to detect outliers.

Scatterplot — basics

- Each point = one observation: x on the horizontal axis, y on the vertical axis.
- We read:
  - **direction** (increasing/decreasing),
  - **shape** (linear, nonlinear, thresholds),
  - **spread** (strength of the relationship, clusters),
  - **outliers**.
- Example: the relationship between years since PhD and salary level.



Data source: *Salaries.csv*

Each point = 1 faculty member.

Scatterplot — readability

- Point aesthetics: *color*, *size*, *shape*, *transparency* (alpha).
- **Transparency** helps when points overlap.
- Axis scales:
  - choosing ranges and tick marks,
  - label formats (e.g., currency, percentages),
  - fixed step on a time axis or years of experience.

Fit lines (best fit)

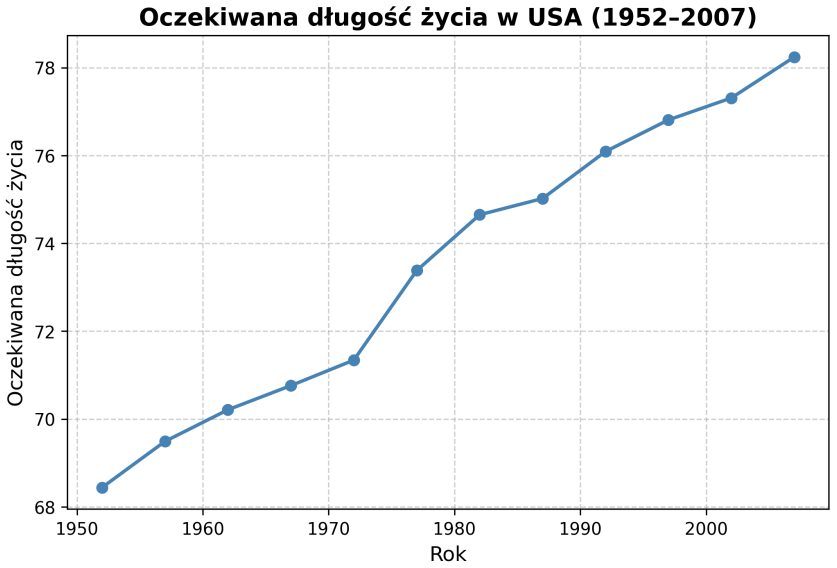
- **Linear** — a quick summary of the trend; usually shown with a 95% confidence band.
- **Polynomial** (e.g., quadratic, cubic) — when the relationship has bends/curvature.
- **Nonparametric** — local smoothing, good for nonlinearity without assuming a specific shape.
- Note: a fit line does not replace the need to inspect the raw points.

What to watch for in scatterplots

- **Nonlinearity**: a straight line may be misleading — check LOESS/polynomial fits.
- **Heteroscedasticity**: the spread increases/decreases with x (funnel shape).
- **Clusters** and **confounding** — different groups within one cloud of points.
- **Influential points** — single observations that change the fit.
- **Scales and transformations** (e.g., log) can reveal patterns.

Line plot (when one variable is time)

- The line connects points ordered in time — it shows changes and trends.
- Good practices:
  - clear axis labels and units,
  - highlight points (dots) for longer series,
  - include a title and data source; optionally annotate important events.
- Remember: time series data are discussed in more detail in time-series analysis.



Data source: *gapminder dataset*

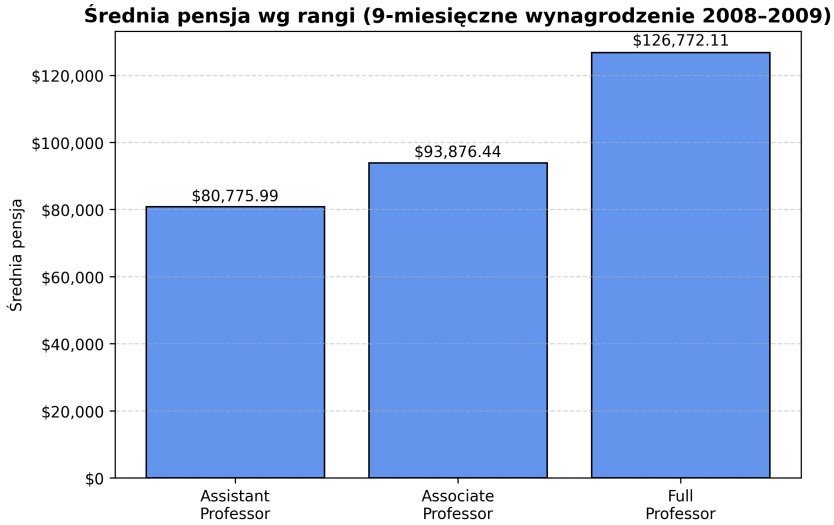
Life expectancy in the USA (1952–2007).

Summary

- Scatterplot: the primary tool for two quantitative variables.
- Fit lines (linear, polynomial, LOESS) help *summarize* the relationship.
- Line plot: when the x-axis represents time.
- Always ensure readability, proper scaling, labeling, and show uncertainty (confidence bands).

Bar chart of means

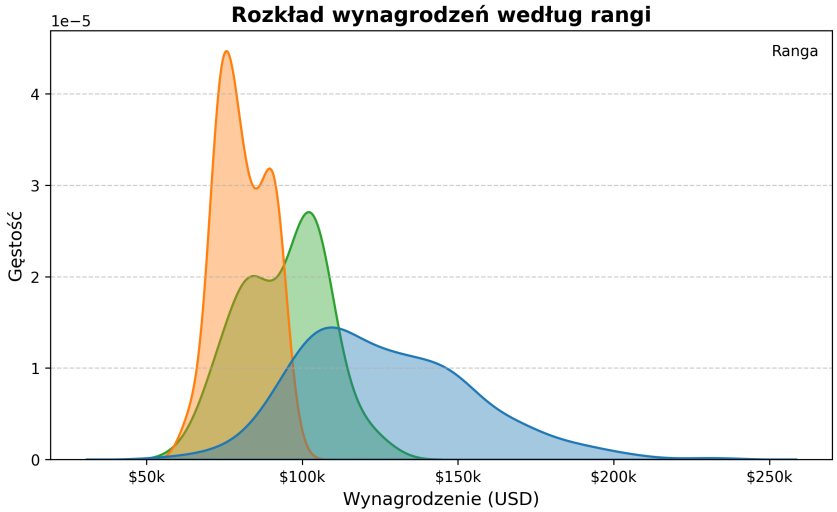
- Used to show the **means** (or medians) of a quantitative variable for each category.
- Example: average salary of professors by rank.
- Bars show summary values, but **not the distribution of the data**.
- Value labels and units (e.g., $ or PLN) are often added.



Source: *Salaries.csv*

Density plot for groups

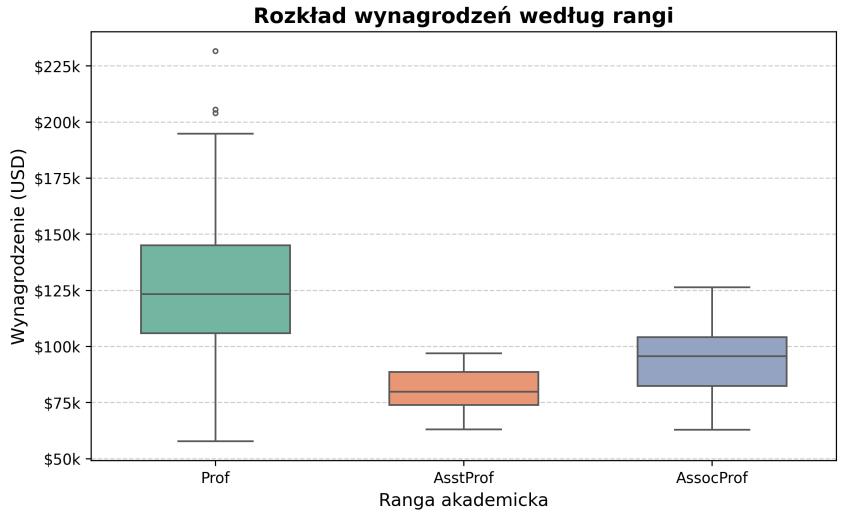
- Shows the distribution of a quantitative variable separately for each group (e.g., academic rank).
- Lines are semi-transparent (`alpha`) — this allows overlapping areas to remain visible.
- Enables assessment of differences in the shapes of distributions between groups.



Each color = a different rank.

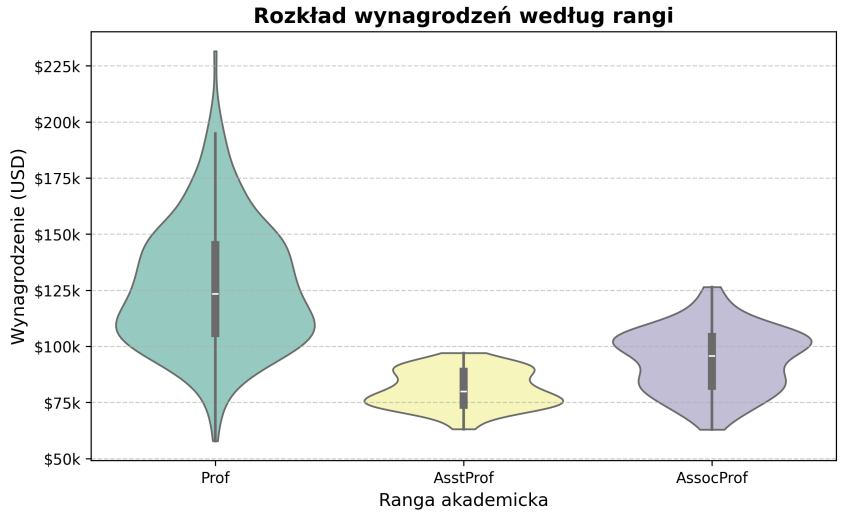
Boxplot

- Shows the median, quartiles, and outliers.
- Allows quick comparison of distributions across groups.
- The “notched” version allows assessing whether medians differ significantly.
- Box height = interquartile range (IQR).



Violin plot

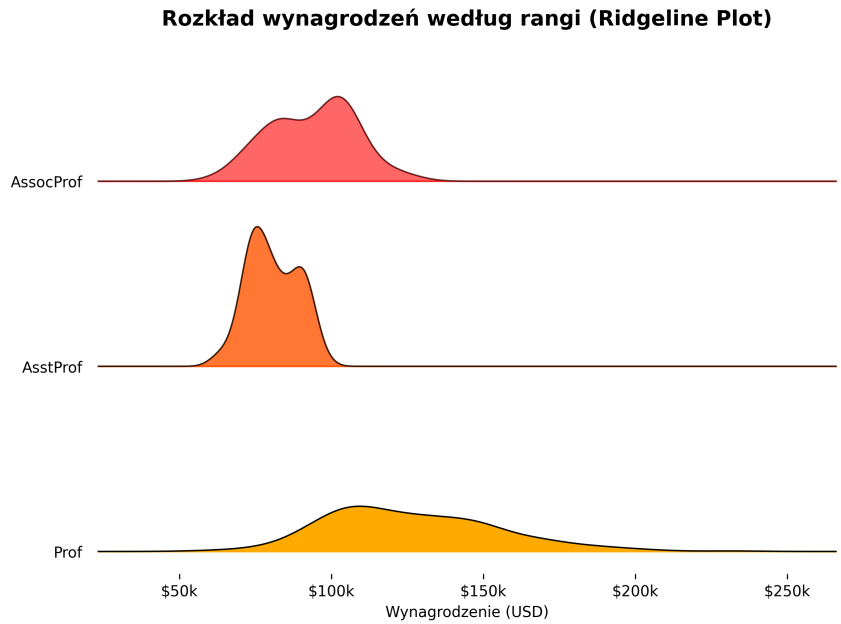
- Combines the advantages of a boxplot and a density plot.
- Shows the shape of the distribution (symmetry, skewness) within each category.
- A boxplot can be overlaid for additional information.



Blue — distribution, orange — boxplot.

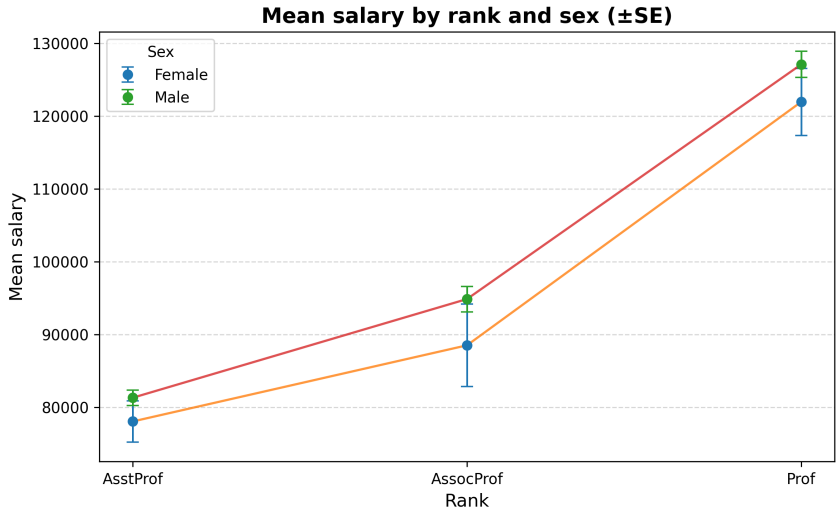
Ridgeline plot

- Shows distributions of many groups stacked one above another — saves space.
- Useful when there are many categories (e.g., car classes).
- Each ridge = distribution of one category of the variable.



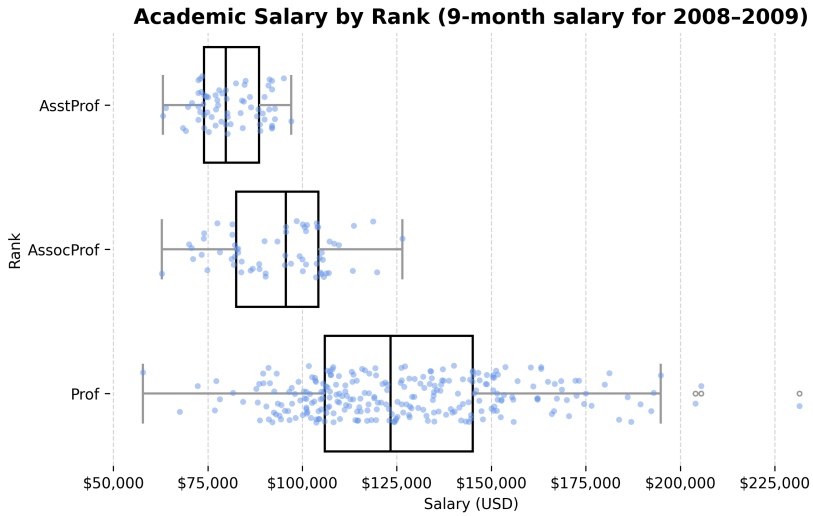
Mean plot with errors (Mean ± SE)

- Shows the mean and an error interval (e.g., *standard error* or 95% CI).
- Helps visually compare groups.
- The line connects the means, and the error bars show uncertainty.
- Useful when comparing groups by sex, rank, etc.



Categorical scatterplot (Strip / Jitter)

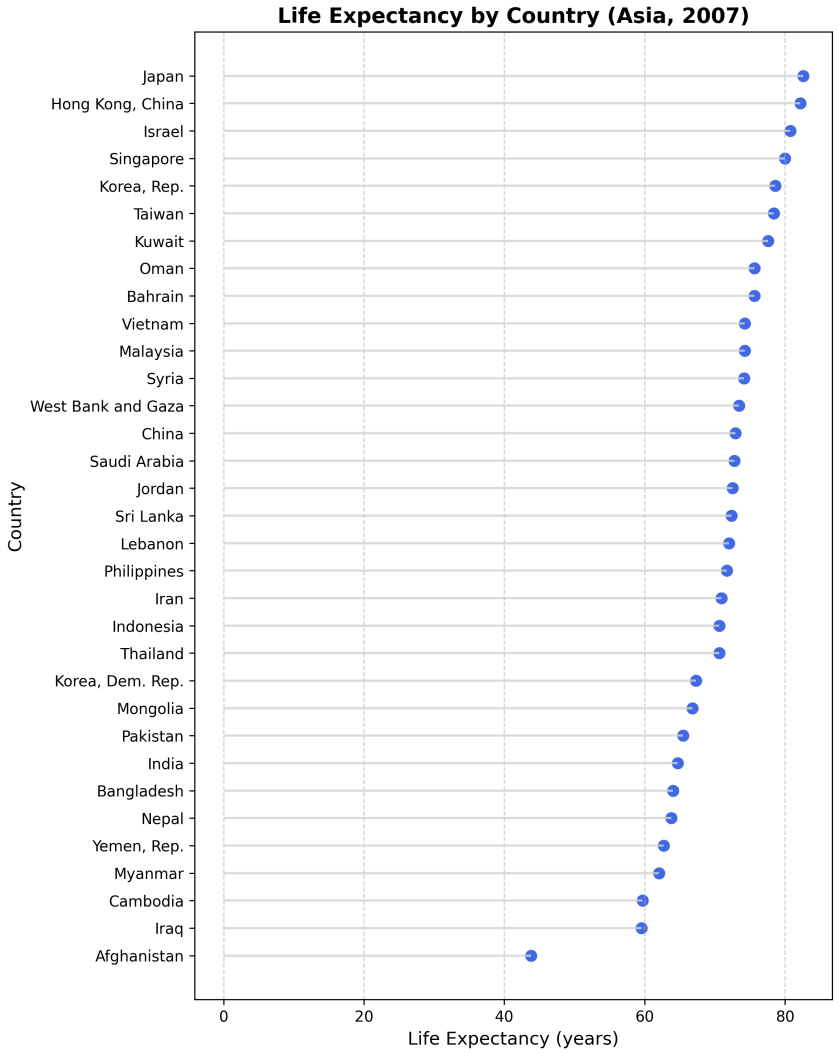
- Each point = a single observation.
- **Strip plot** — points overlap, **jitter** — points are slightly spread out.
- Helps visualize data density and variation.



Each point = one professor.

Cleveland plot (Dot / Lollipop)

- Shows quantitative values for many categories.
- Makes precise comparisons between groups easier.
- A connecting segment can be added between points (a *lollipop plot*).
- Example: life expectancy in Asian countries (2007).

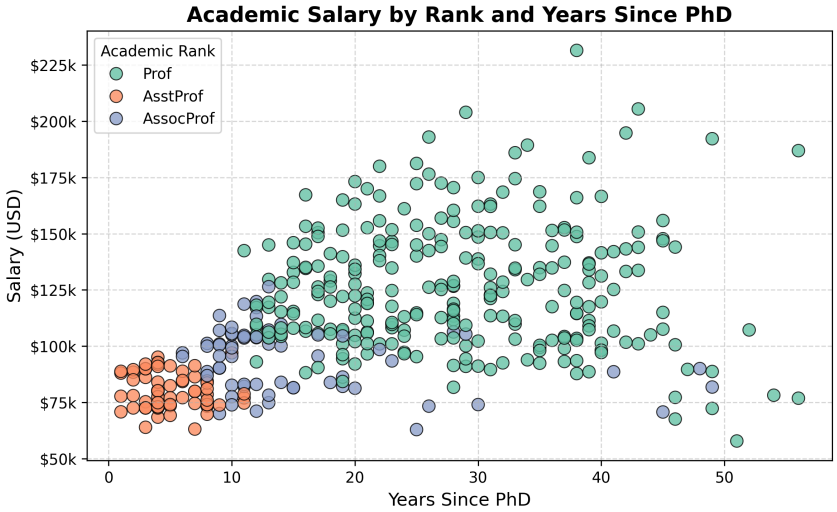


Multivariate plots — the idea

- Multivariate plots show the relationships among **3+ variables**.
- Two popular techniques:
  - **Grouping** — mapping additional variables to visual features.
  - **Faceting** — *small multiples*: many small plots.
- The choice depends on the goal: a single-plot synthesis vs. comparison across panels.

Grouping — aesthetic mapping

- Axes: x and y take 2 variables.
- Additional variables can be mapped to:
  - **color** (e.g., group categories),
  - **point shape** (different types of observations),
  - **size** (a quantitative variable — e.g., a *bubble plot*),
  - **line type** or **transparency** (*alpha*).
- Allows showing several dimensions of data in a single plot.
- Note: too many variables may cause *"overloading"* of the plot.



Example: salary distribution with color = sex, size = experience.

Grouping — good practices

- Use **transparency** when points overlap.
- Use **qualitative palettes** for categories; label the legend clearly and concisely.
- For size mapped to a quantitative variable: provide a **readable size legend**.
- Limit the number of groups — too many colors/shapes reduce readability.

Example: Scatter + color / shape / size

- x = years since PhD, y = salary.
- Color = **rank** (Asst/Assoc/Prof) — separates groups.
- Shape = **sex** — distinguishes M/F (watch for readability).
- Size = **years of service** — "bubble plot"; requires a size legend.
- Conclusions: strong relationship between *years since PhD* and *years of service*; distributions differ across ranks and sexes.

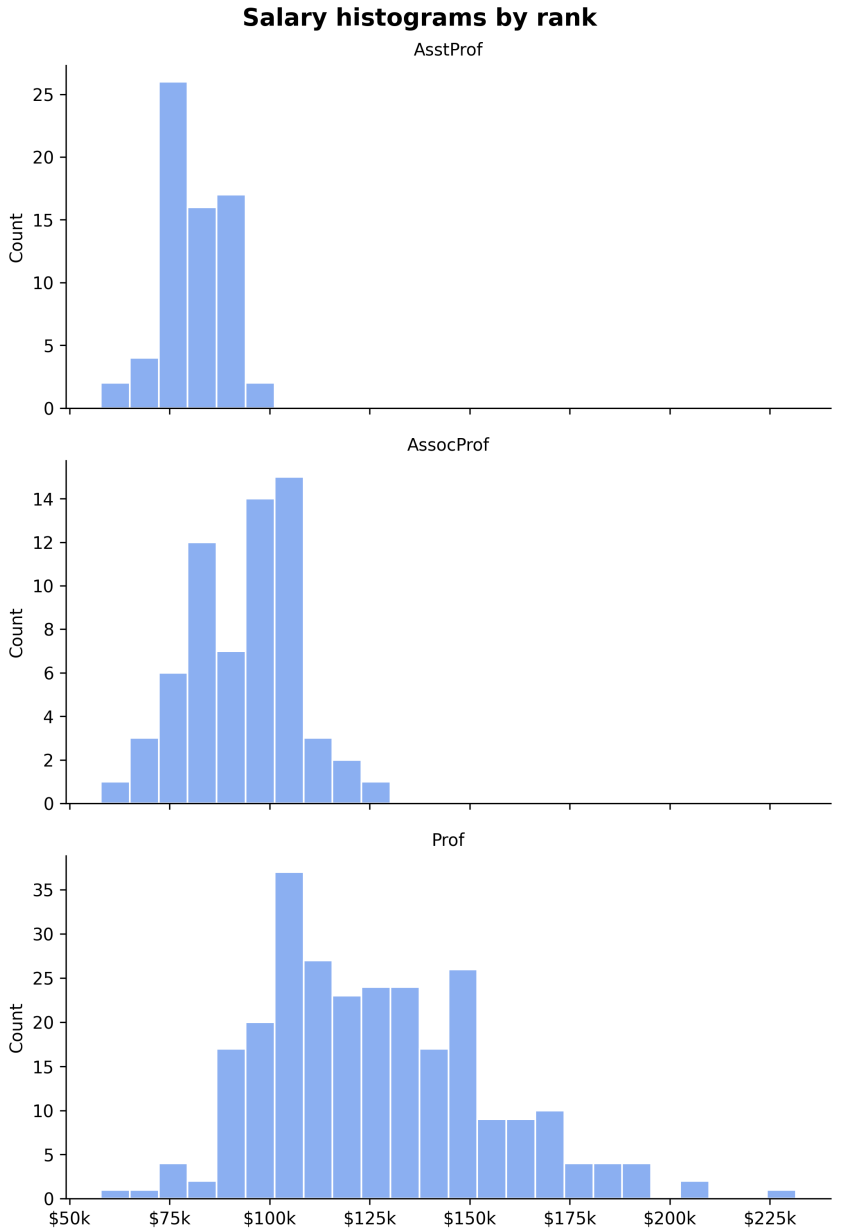
Example: salary distribution with color = rank, shape = sex, size = experience.

Faceting — definition

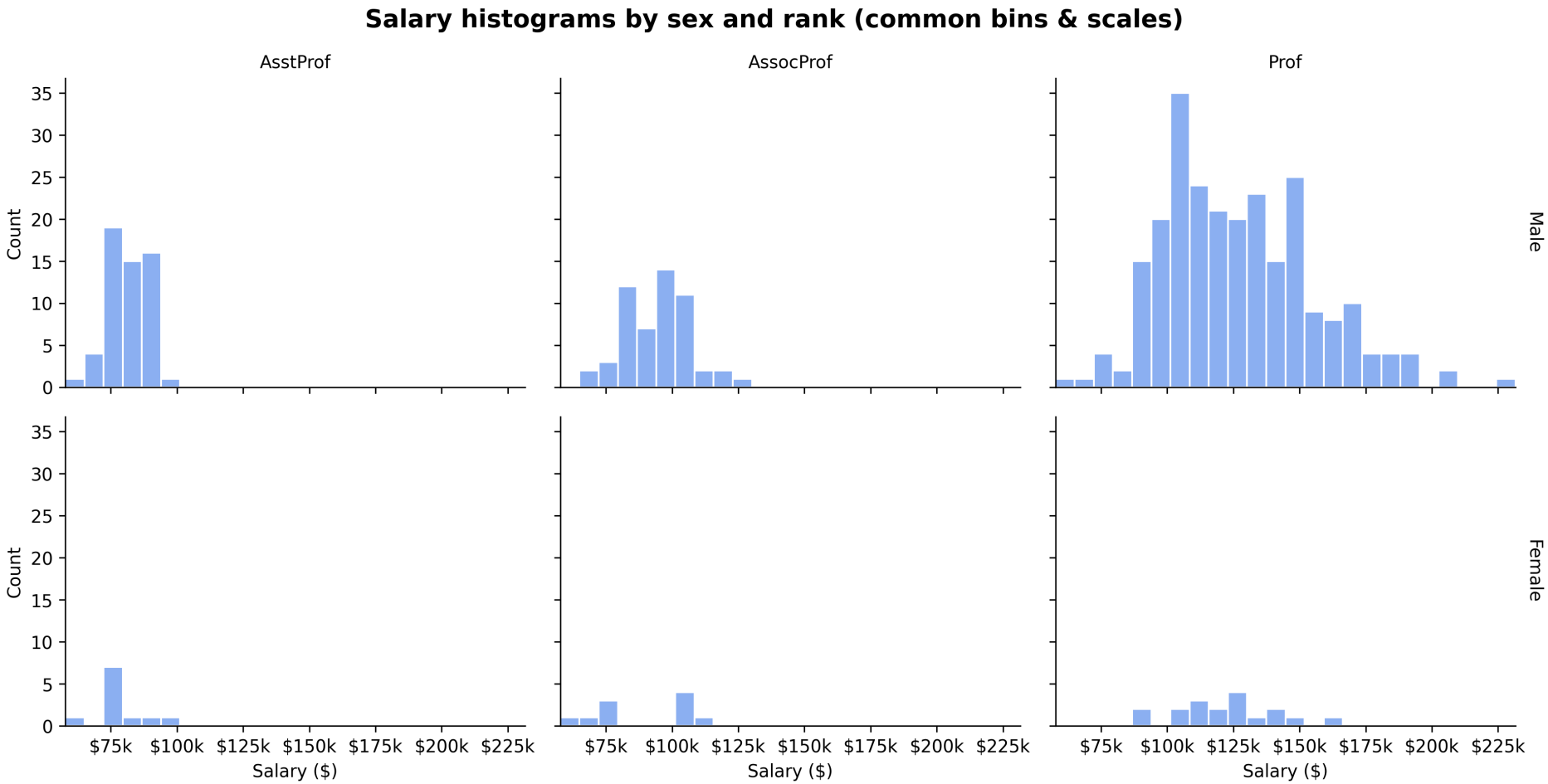
- We create **separate panels** for the levels of a third variable (or combinations of variables).
- Advantages:
  - order and clarity (without overloading with colors),
  - easy comparison of **distribution shapes** and trends across panels.
- Variants:
  - **facet_wrap** — a grid based on one factor,
  - **facet_grid** — rows and columns based on two factors.

Faceting — tips

- Set **common axis scales** when you want to compare levels.
- Limit the number of panels — too many "mini-plots" strain the eyes.
- Add **clear panel headers** and, if needed, **source information**.
- Combine with *grouping* sparingly (e.g., color + faceting by 2 dimensions).
- Use faceting when you want to **compare distributions across groups**.

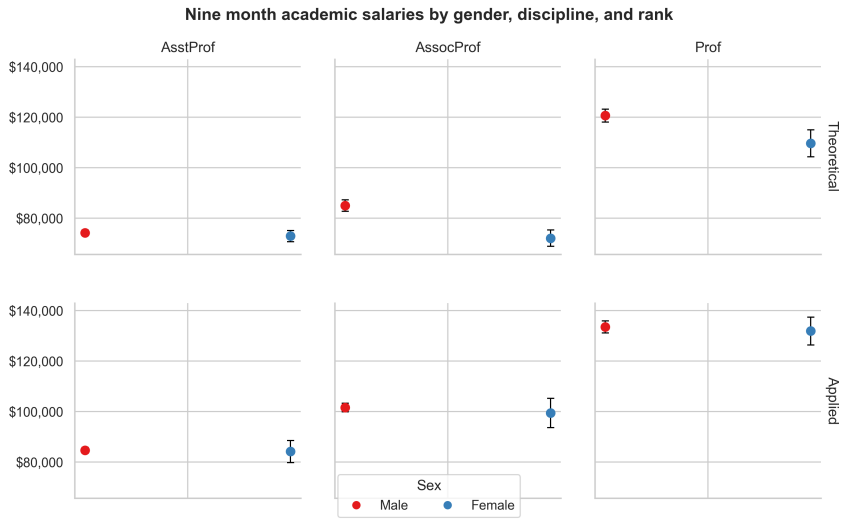


Faceting — another example



Combining grouping + faceting

- Example: mean salaries with standard error (**Mean/SE**) by **sex** (color), and faceting by **rank** and **discipline** (theory/applied).
- Advantage: simultaneous comparison **between groups** (sex, discipline) and **within each panel** (rank).
- We see that mean salaries increase with rank; differences between sexes are moderate.
- Risk: symbols/fonts may become too small — useful to **scale** and **simplify the background**.



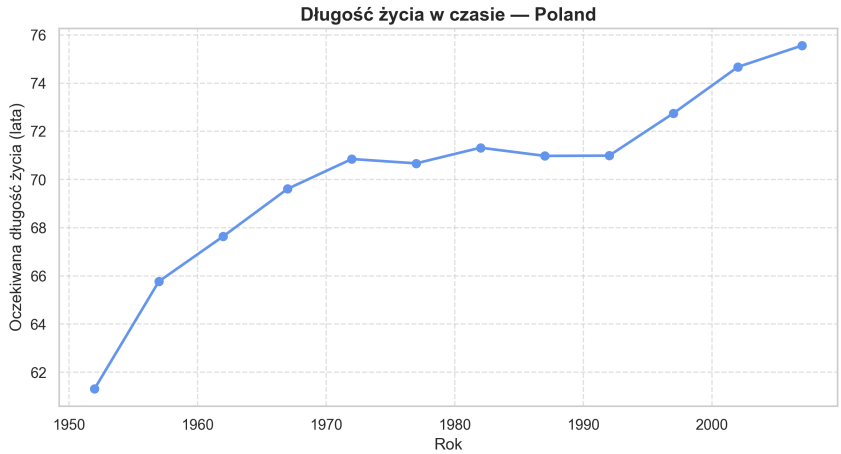
Source: *Salaries.csv* (average 9-month salary by sex, rank, and discipline).

When to use grouping vs. faceting?

- **Grouping** — when you want to see relationships among many groups **at a glance**.
- **Faceting** — when the priority is **comparing shapes** across groups without the clutter of many colors.
- Often a **hybrid** (e.g., color within panels) gives the best compromise.

Time series — basics

- A **line plot** shows change over time (x-axis = time).
- Good practices:
  - clear **date ticks** and **label formatting** (e.g., every 5–10 years),
  - axis labels and **data source**.
- Example: change in life expectancy in Poland from 1952–2007 (Gapminder).

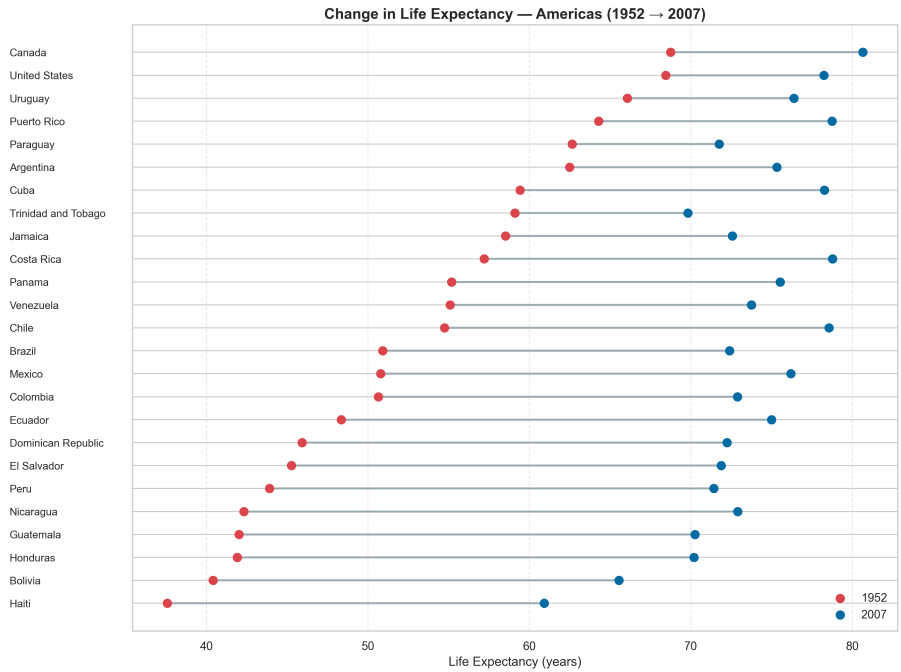


Source: *Gapminder dataset (1952–2007)*

Values: life expectancy (years).

Dumbbell chart — change between two points in time

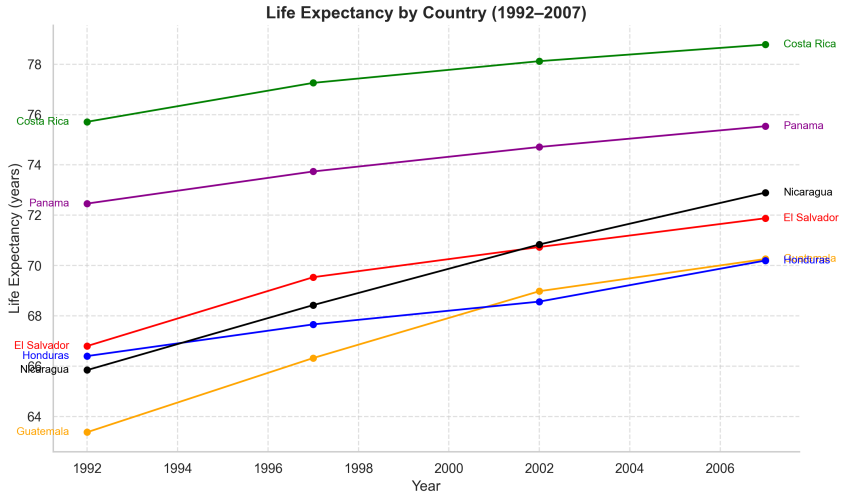
- For many units (e.g., countries) we compare **values in year A and B**.
- A line connects the two points; sorting by the initial value improves readability.
- Useful for **showing increases/decreases** and the **ranking** of units.
- Point colors: 1952 vs. 2007.
- The chart shows an increase in life expectancy in all countries of the Americas.



Source: *Gapminder (1952–2007)*

Slope graph — several points in time

- A few selected years on the x-axis; each line = one group.
- Labels placed **at the beginning and end** of the lines instead of a legend.
- Excellent for "*before and after*" stories or short time sequences.
- Watch out for **line crossing** — adjust colors and label placement accordingly.

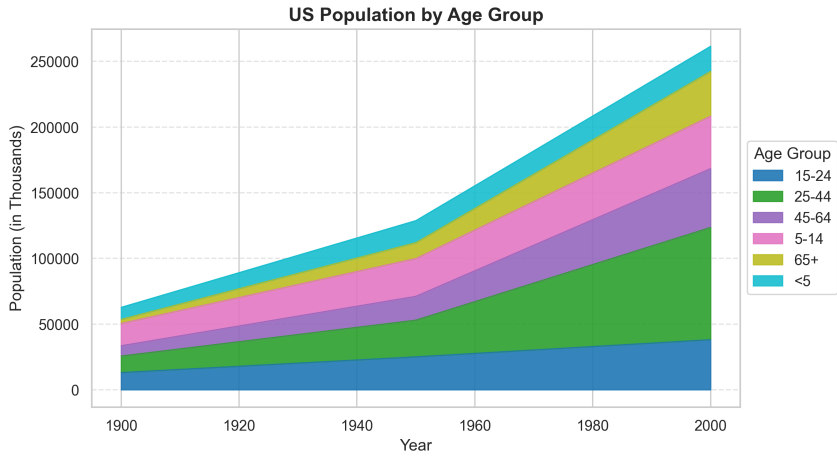


Source: *Gapminder (1992–2007)*

Central American countries — life expectancy.

Area chart — filled line plot

- A **single** area chart emphasizes the **level and trend** of a value over time.
- A **stacked** area chart shows **group shares** and the **total sum**.
- Good practices:
  - use meaningful **units** (avoid scientific notation; convert thousands → millions),
  - **order layers** from the most important at the bottom,
  - clear **boundaries** and consistent color palettes.



Source: *gcookbook::uspopage*

Stacked chart: U.S. population by age group (1900–2000).

Summary

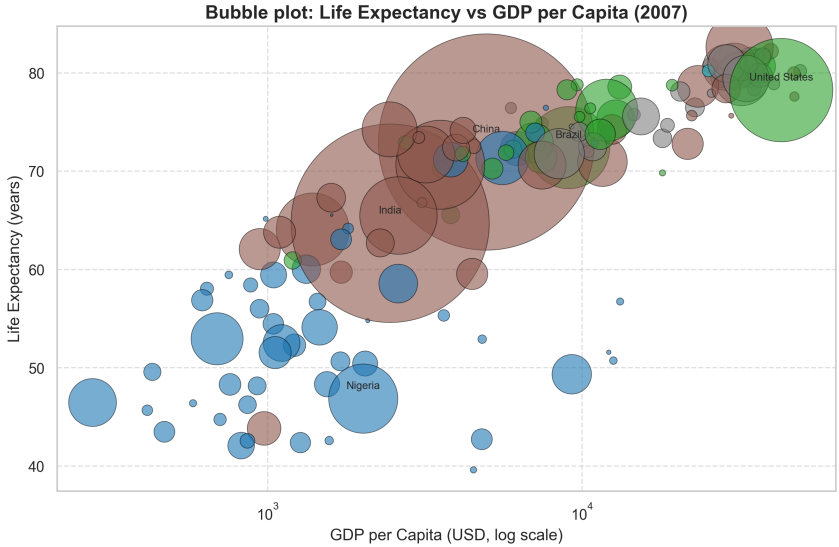
- Map variables to visual features **intentionally** (color/shape/size/line/alpha).
- Use **facets** to organize comparisons across groups.
- In time series, pay attention to **dates**, **scales**, and **readability** (optional LOESS).
- Choose the chart type according to the **analytical question**: change over time, differences between years, structure of shares, etc.

Other Chart Types

- Useful in specific situations, harder to fit into classic categories.
- Often combine multiple dimensions / require additional design decisions.
- Key factors: **readability** and **appropriateness** for the analytical question.

Bubble chart

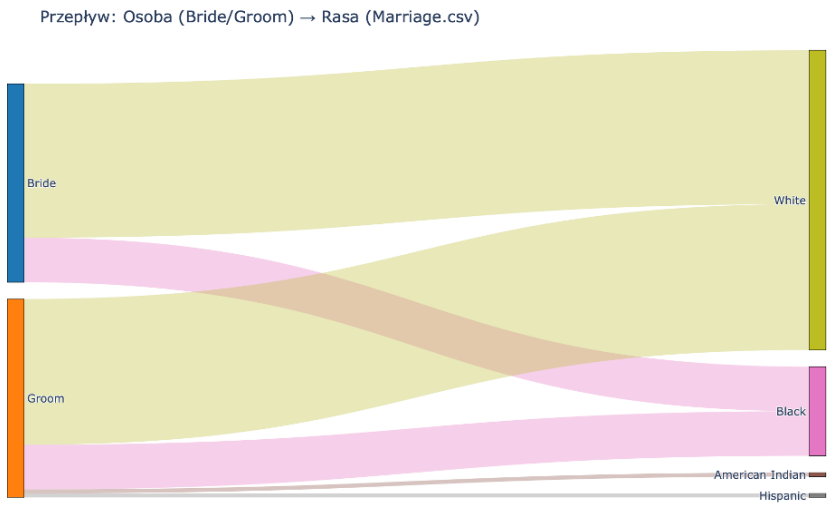
- A scatterplot where the **size** of the point encodes a third quantitative variable.
- Advantages: dense information in a single plot; easy detection of outliers.
- Disadvantages: people judge **area** worse than **length** (possible misleading impressions).
- Good practices:
  - size-scale legend,
  - moderate transparency,
  - limit the number of points.



Source: *Gapminder data (year 2007).*

Flow diagrams: Sankey & Alluvial

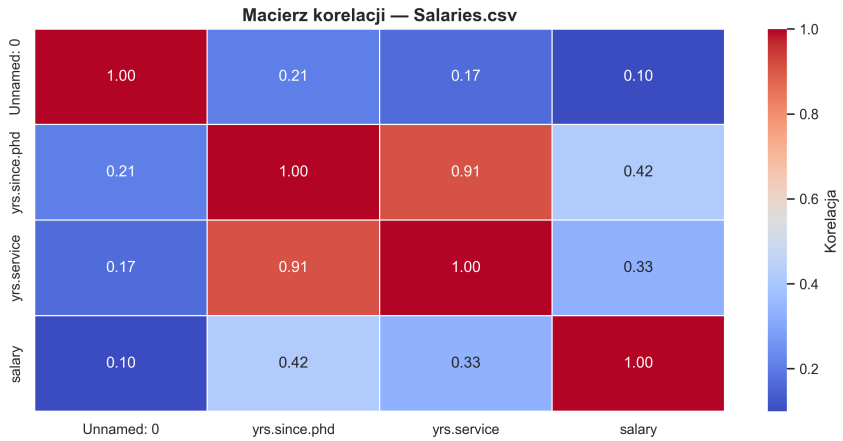
- **Sankey**: shows the flow of quantities from sources to targets — line width = magnitude.
- **Alluvial**: a variant of Sankey with ordered "layers" (e.g., stages, years, levels).
- Applications:
  - population migration, energy flows, user paths,
  - conversion analysis or category transitions over time.
- Tips:
  - group rare categories (*Other*),
  - add labels for units and values,
  - use a consistent color palette for categories.



Example: flow from Bride/Groom → race (data: *Marriage.csv*).

Heatmap

- Tiles encode **values** using color for many variables and observations.
- Often preceded by **standardization** (e.g., z-scores) of columns or rows.
- Add-ons: **clustermap** (clustering of rows/columns with a dendrogram).
- Note: the color scale and ordering have a major impact on interpretation.

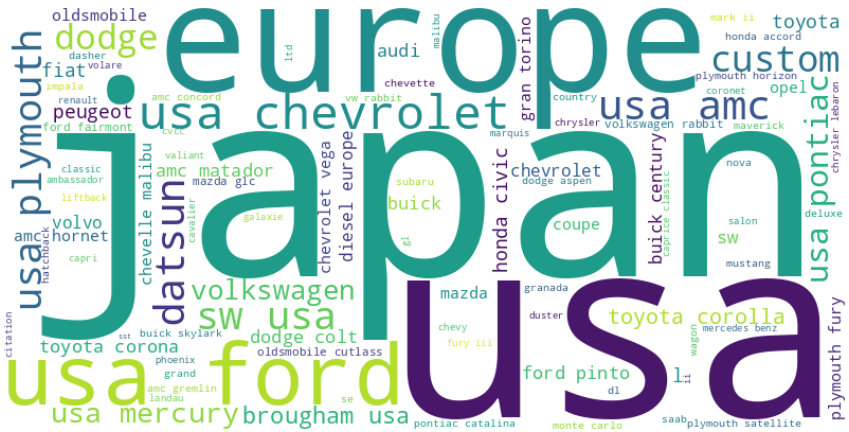


Source: *Salaries.csv*

Color = strength of correlation between variables.

Word cloud

- Word frequency → **size** of the word; a quick overview of themes.
- Necessary steps: **stopwords** (removal of function words), text cleaning and normalization.
- Limitations: visually appealing but **not very precise** quantitatively; best combined with bar charts.



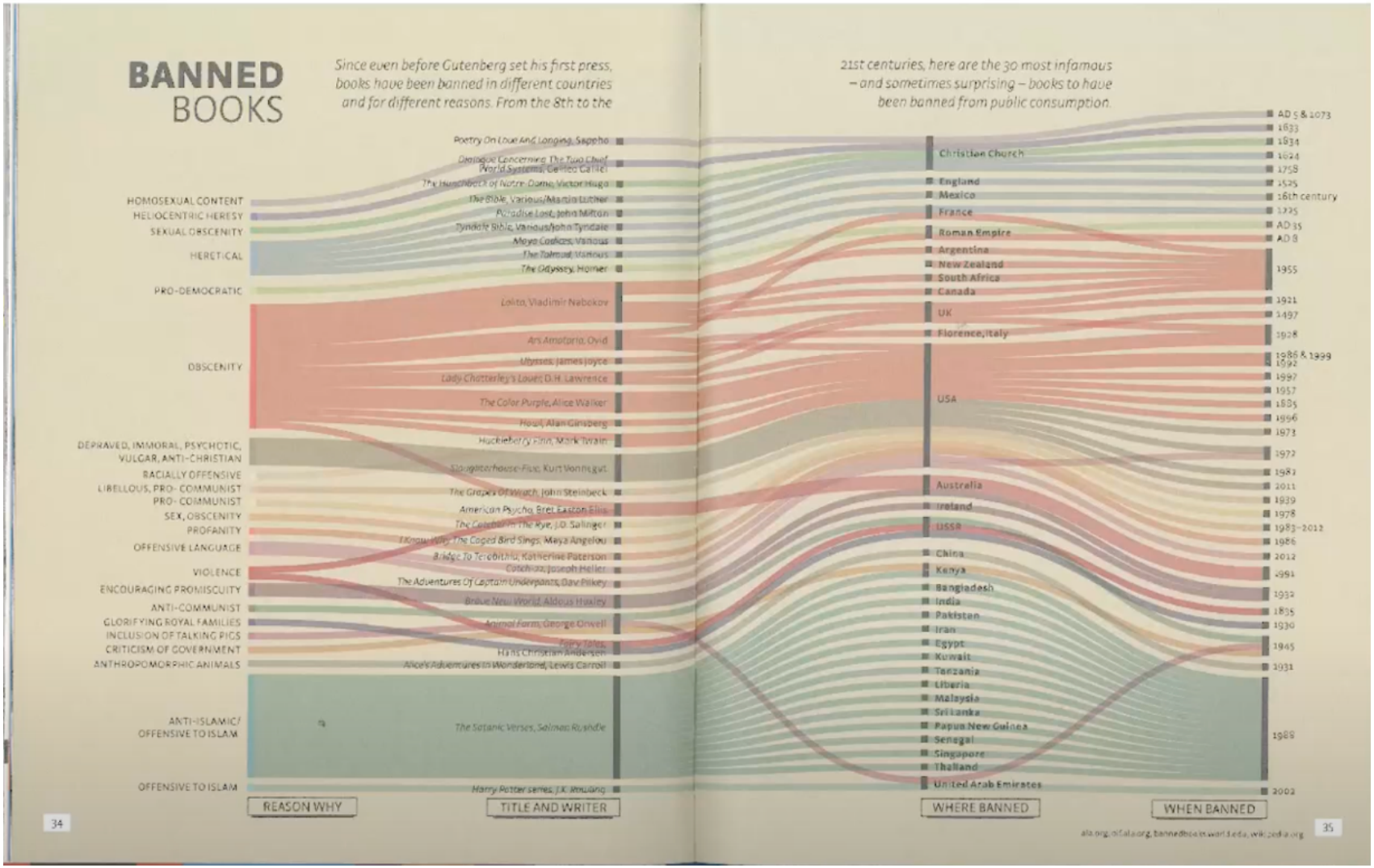
General good practices

- Always match the chart type to the **question**: order/shares/trend/flow/profile.
- Take care of **scales**, **legends**, **axis labels**, **sources**, and **units**.
- Limit the number of colors/series; use **transparency** and **sorting**.
- Avoid optical illusions (unnecessary 3D, excessive decorations).

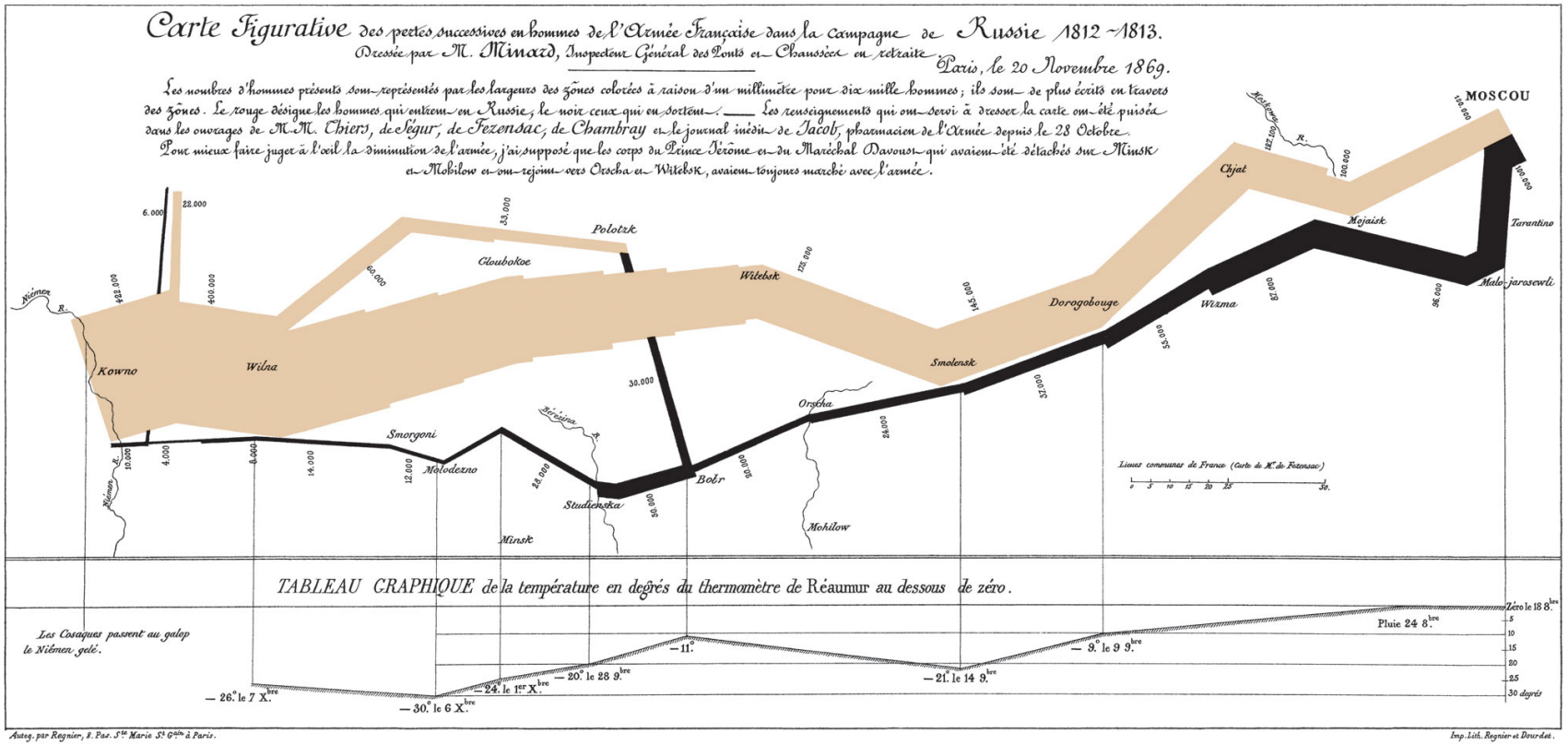
What is visualization?

- *Based on (non-visual) data.* The goal of visualization is to communicate data. This excludes photography and image processing. Visualization transforms what is invisible into something visible.
- *Creation of an image.* It may seem obvious that visualization must create an image, but this is not always straightforward. The image must be the main medium of communication — other modalities can only provide supplementary information. If the image is just a small part of the process, it is not visualization.
- *The result must be readable and interpretable.* The most important criterion is that visualization must allow us to learn something about the data. Every transformation of complex data into an image omits some information, but it must preserve at least the essential aspects that can be interpreted.

Bad Visualizations 1/6



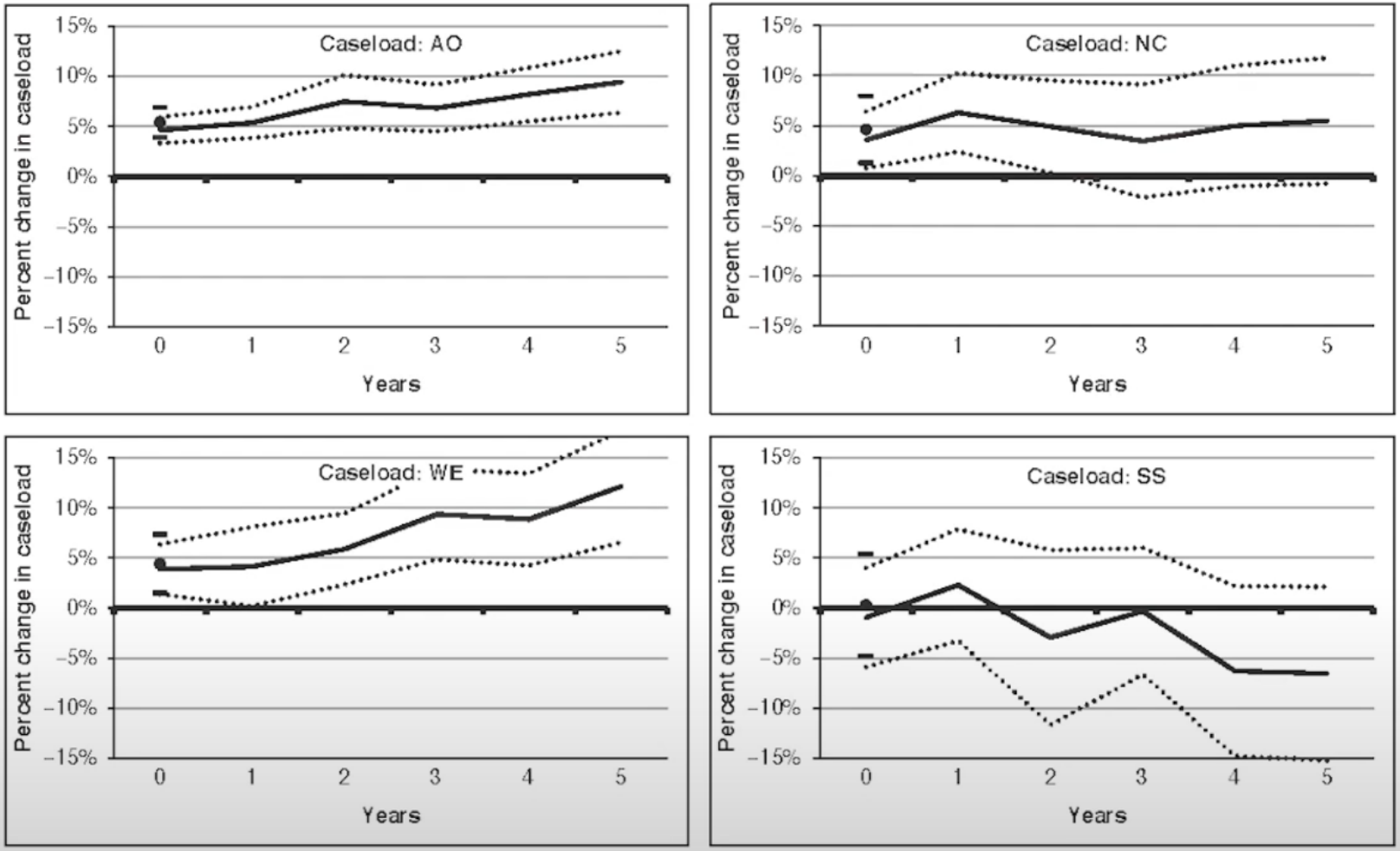
Bad Visualizations 2/6



Bad Visualizations 3/6

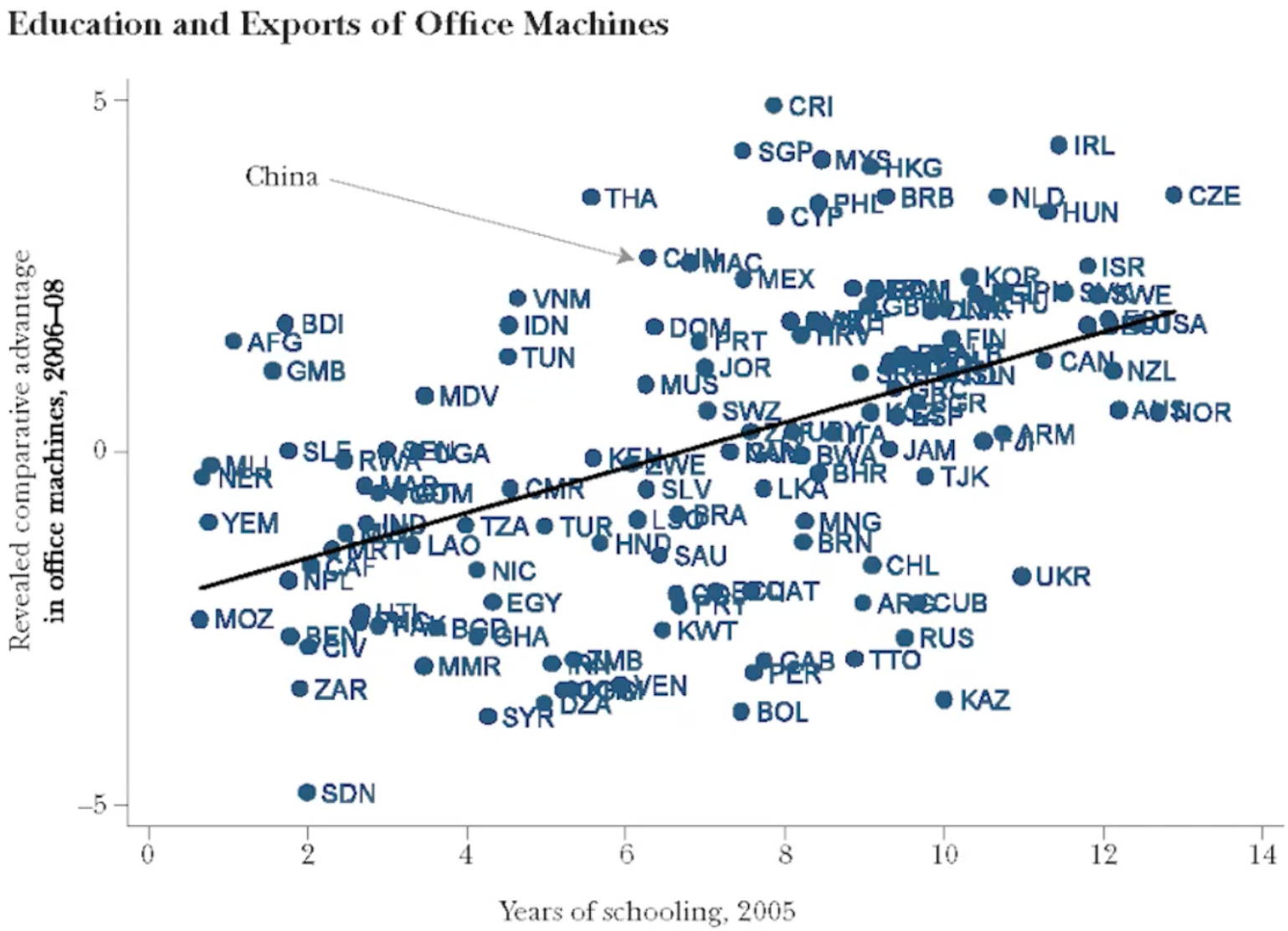
Figure 1A

An Original Line Chart

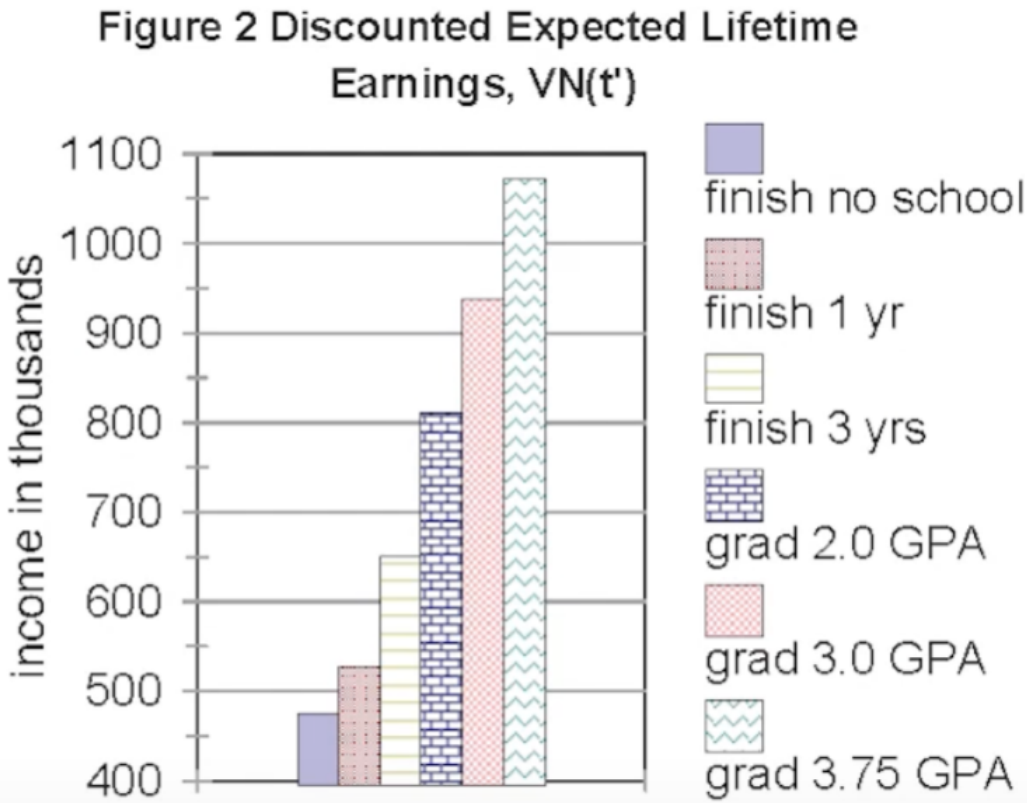


Source: Klerman and Danielson (2011).

Bad Visualizations 4/6

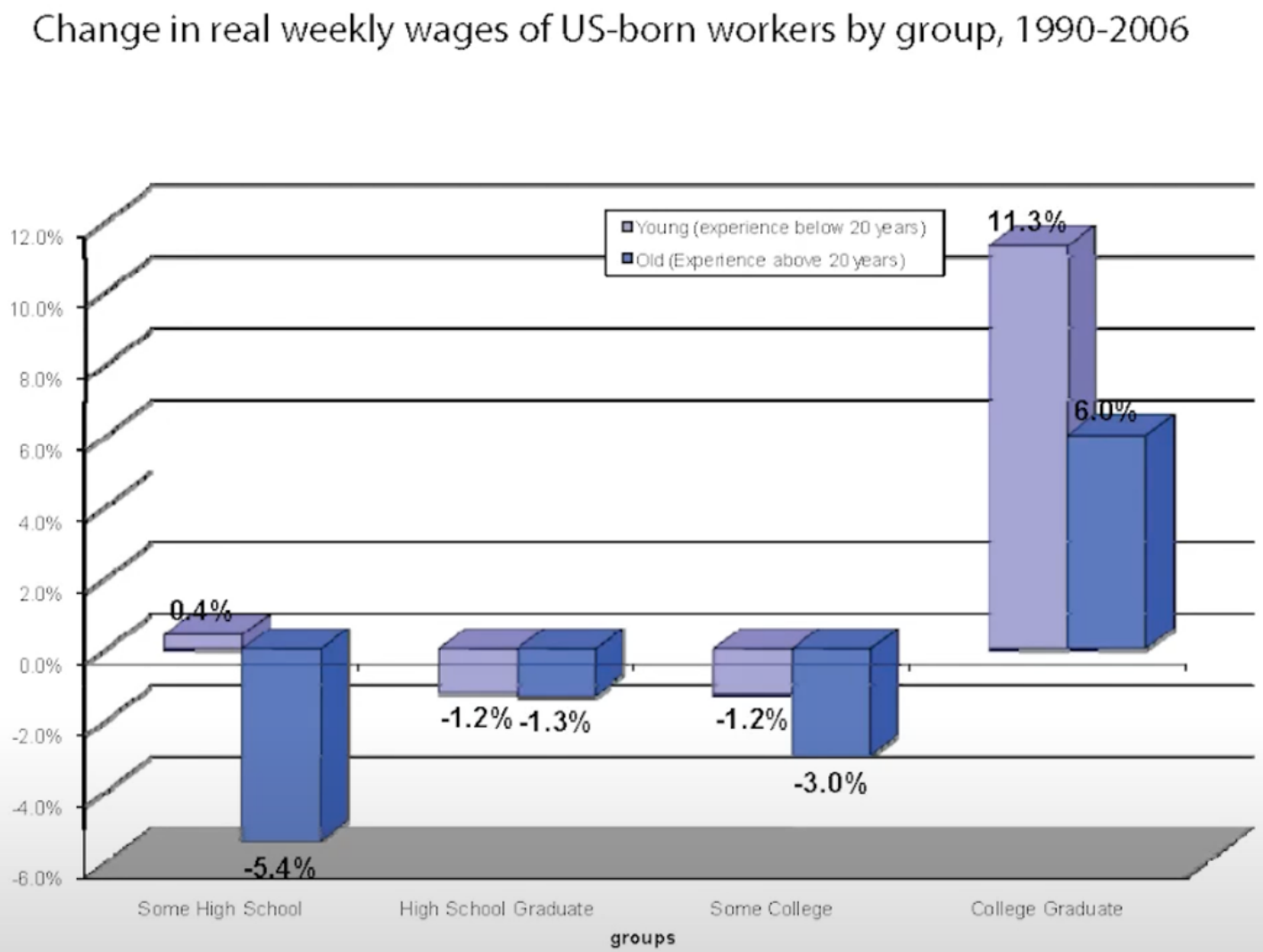


Bad Visualizations 5/6



Source: Stinebrickner and Stinebrickner (2013).

Bad Visualizations 6/6

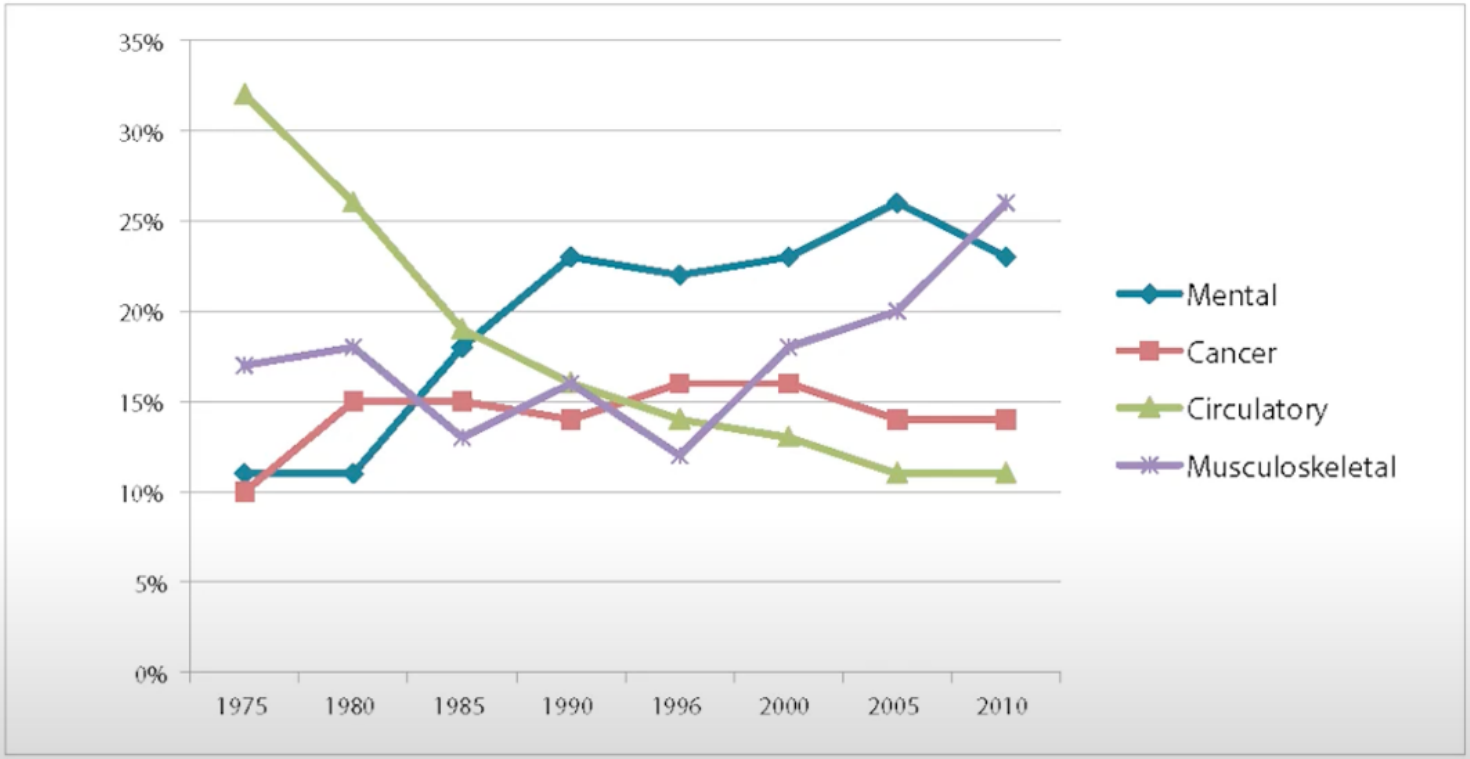


Source: Ottaviano and Peri (2008).

Bad Visualizations 7/6

Figure 6A
A Spaghetti Chart

27. Initial DI Worker Awards by Major Cause of Disability—Calendar Years 1975-2010



Source: Social Security Advisory Board (2012).

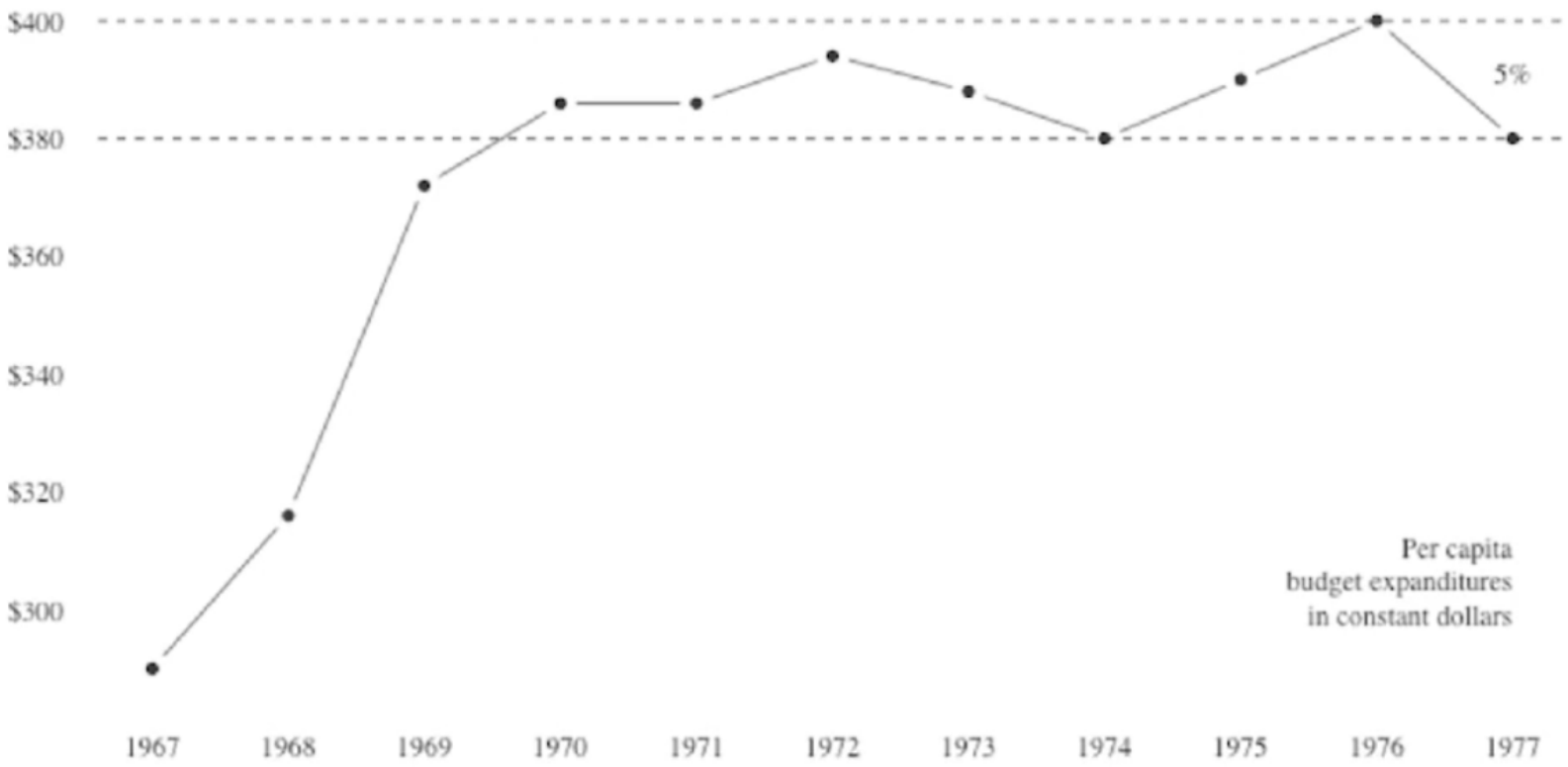
Edward Tufte

- The foundational source, widely discussed and developed by many authors — not necessarily the most up-to-date...
- *The Visual Display of Quantitative Information*
- *Envisioning Information*
- Common sense, but take it with a grain of salt ;)

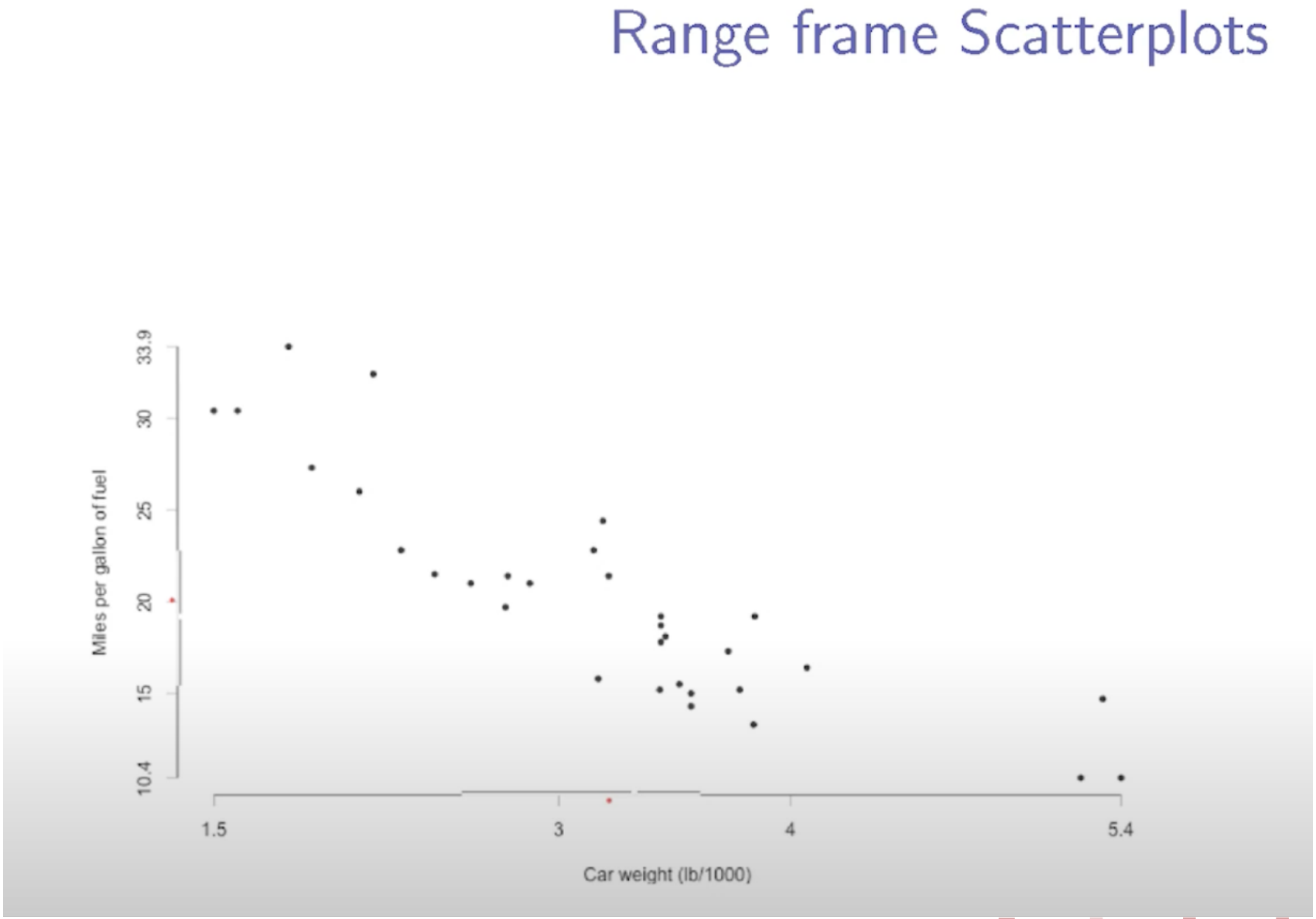
Edward Tufte

- Show the data
- Maximize the ratio of data-ink to total ink
- Remove non-data ink (as much as possible)
- Remove redundant data ink
- Avoid chartjunk (moire effects, decorative "ducks")
- Aim to increase data density
- Graphics should preferably have a horizontal orientation
- Where this leads (Chapter 6 of *The Visual Display of Quantitative Information*)

Edward Tufte example 1



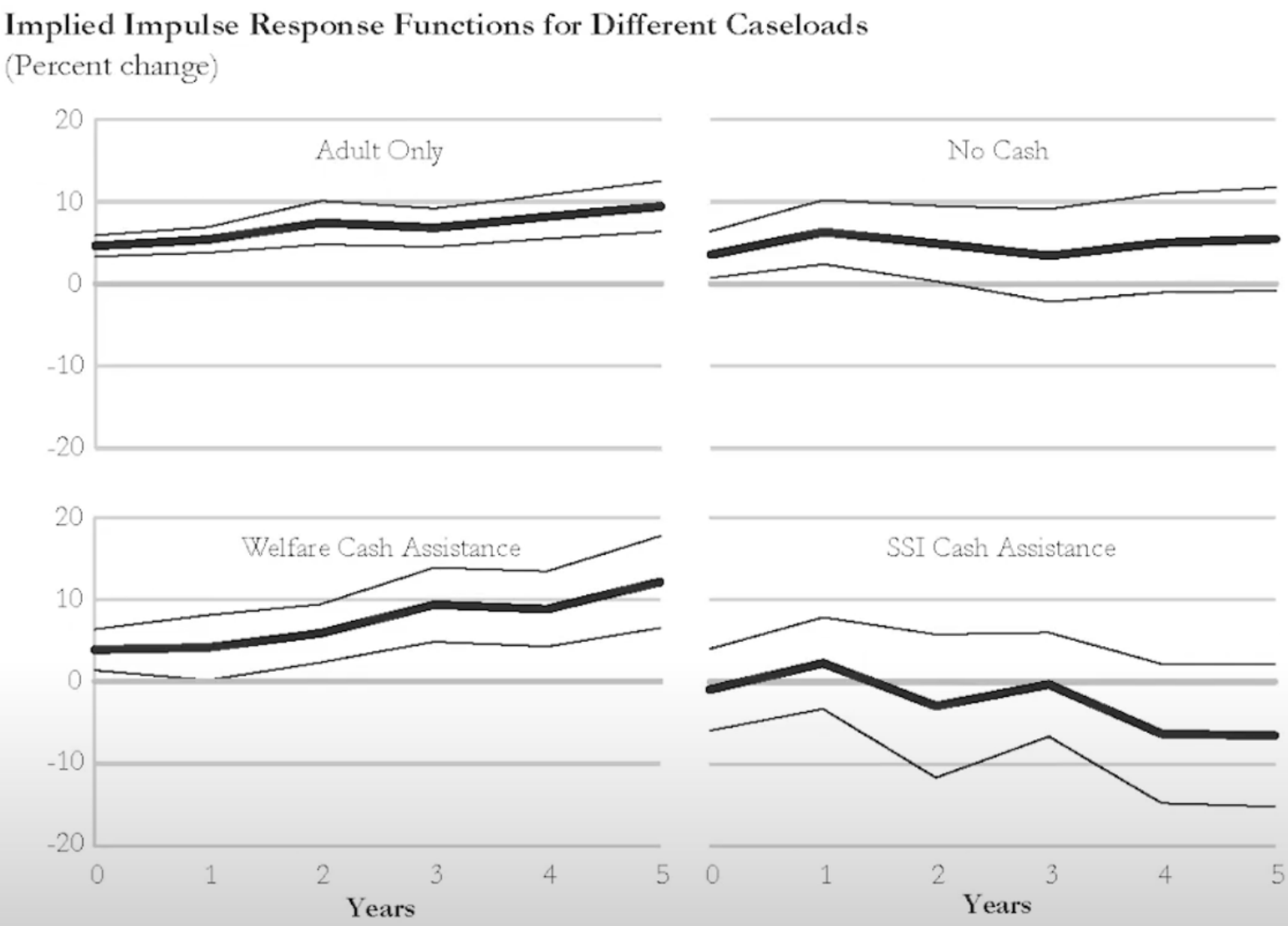
Edward Tufte example 2



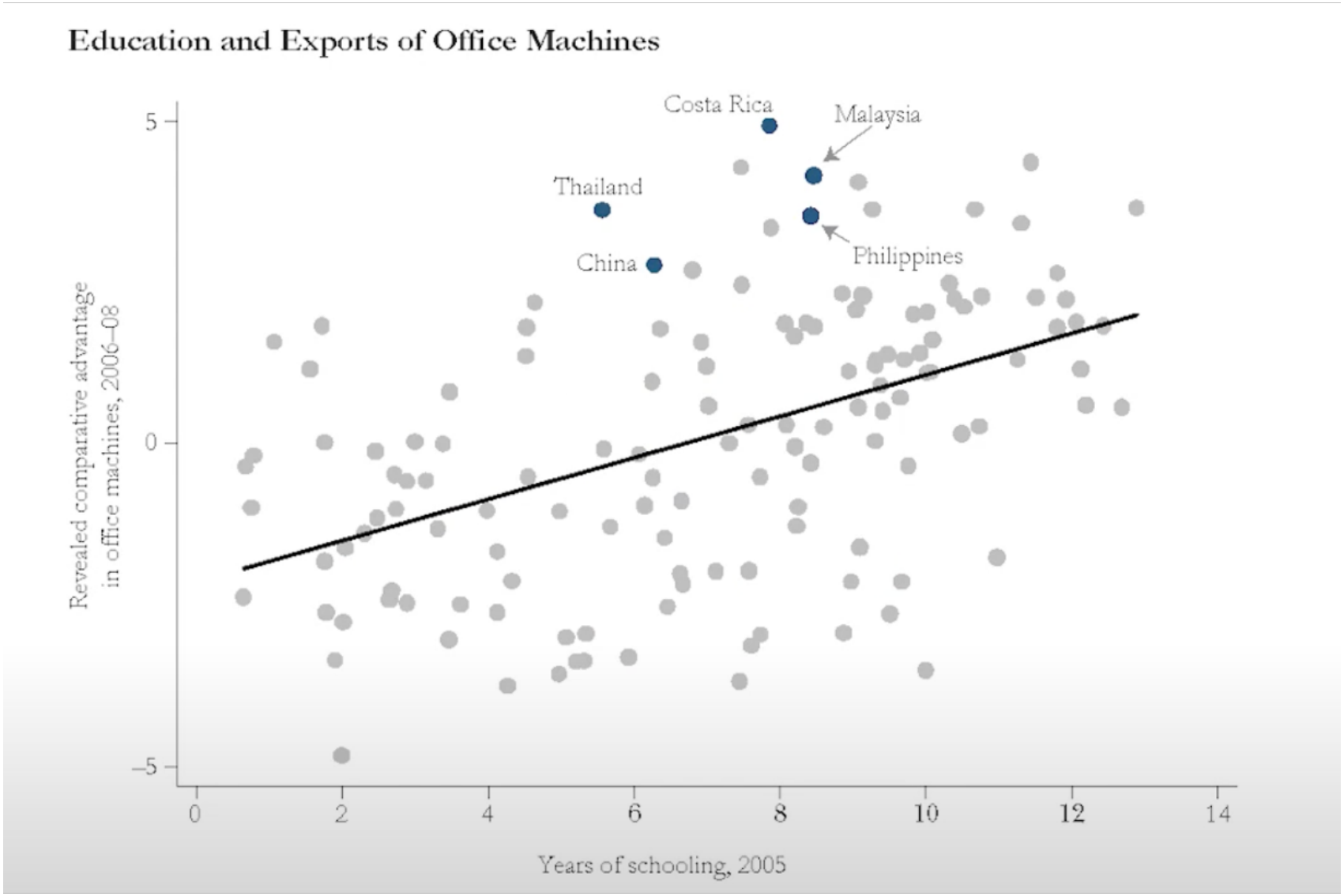
Some rules about charts (based on "Flowingdata")

- Bar charts must start at zero [most axes should start at zero]
- Avoid too many slices in pie charts
- Respect the part-to-whole principle (in pie charts)
- Show the data: use small multiples, aggregation, transparency
- A chart should be self-explanatory:
  - Explain the encoding (what represents what)
  - Label the axes
  - Include units
  - Size circles by area, not diameter

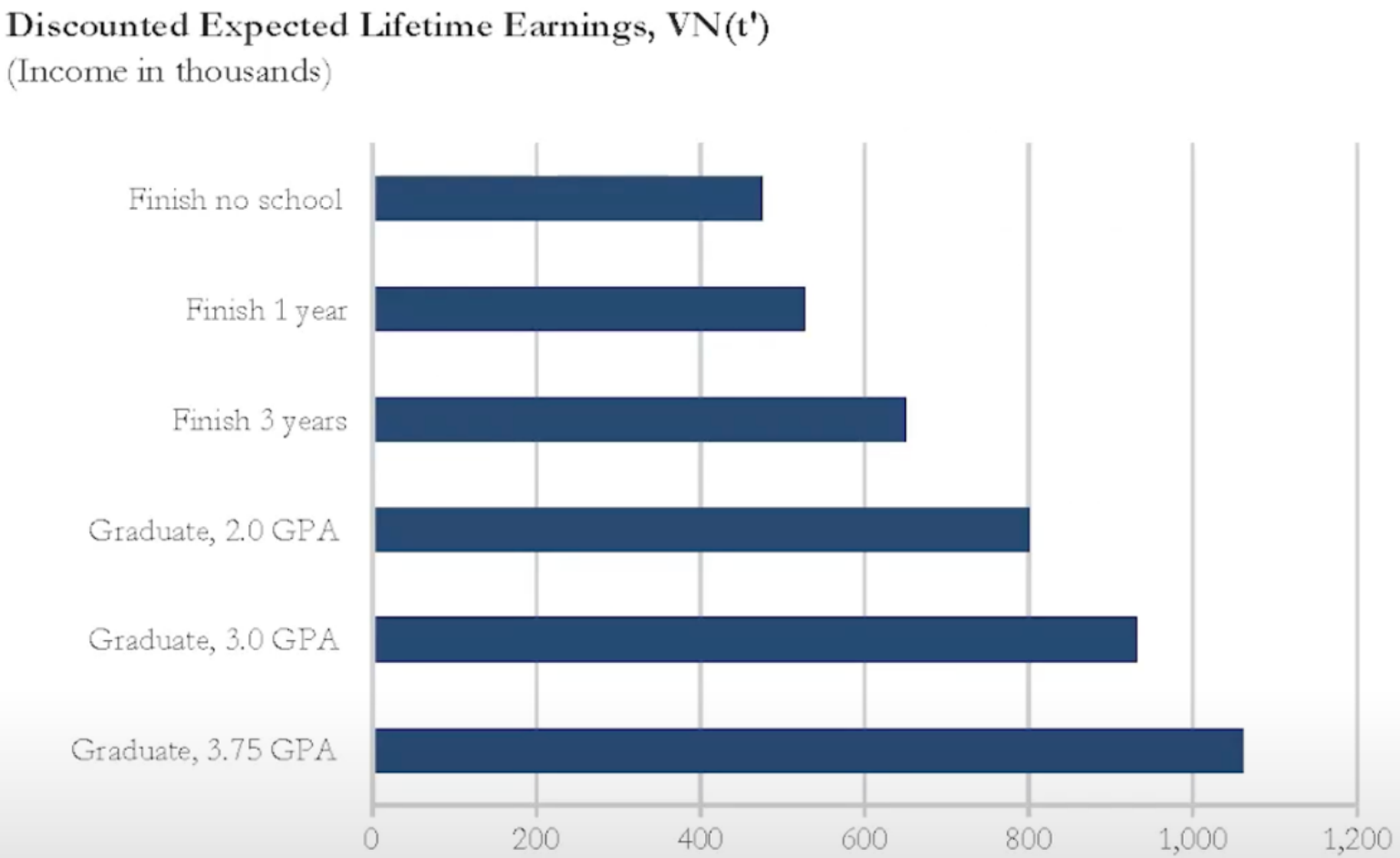
Bad Visualization Improved 1



Bad Visualization Improved 2



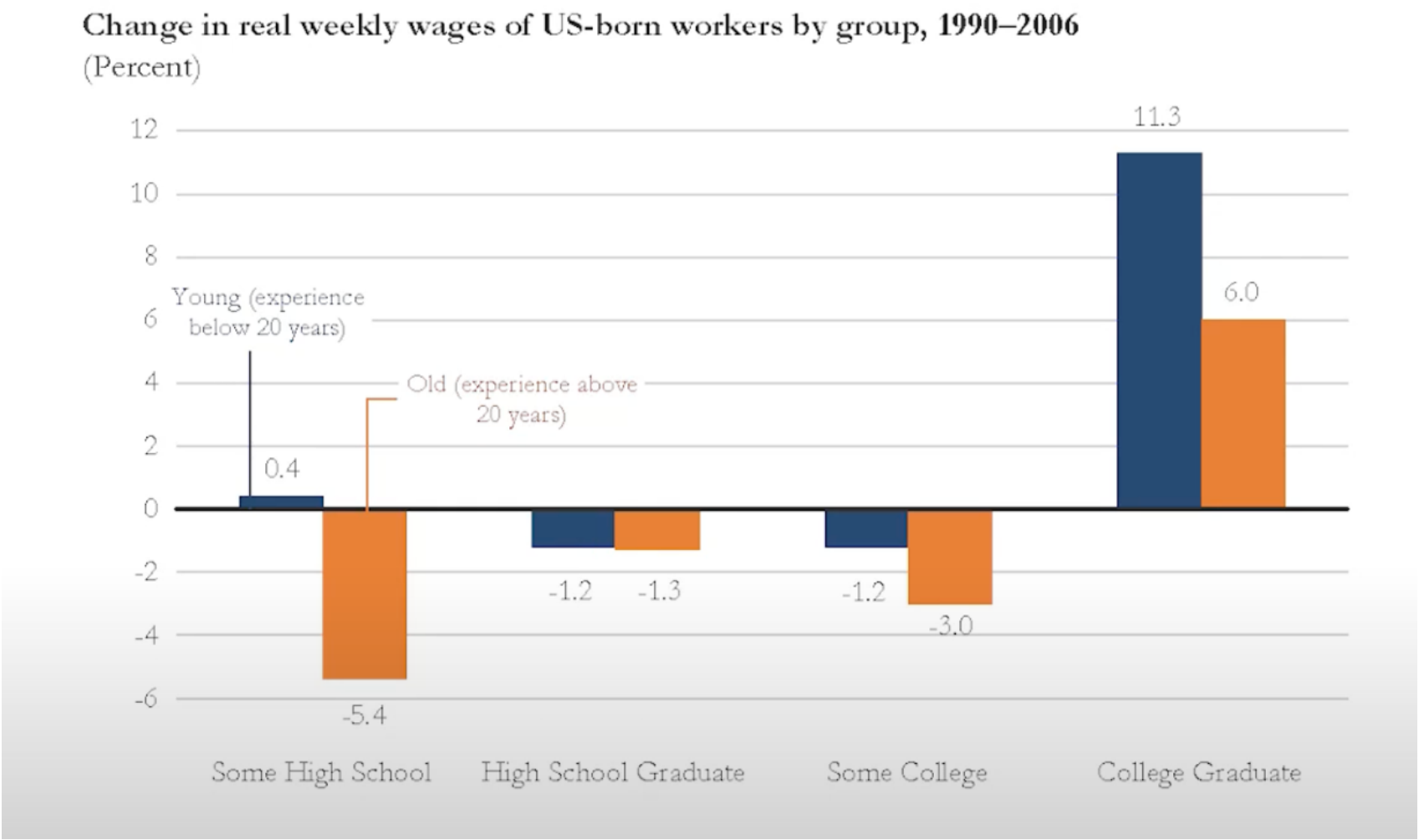
Bad Visualization Improved 3



Source: Author's calculations using numbers inferred from text in Stinebrickner and Stinebrickner (2013).

Bad Visualization Improved 4

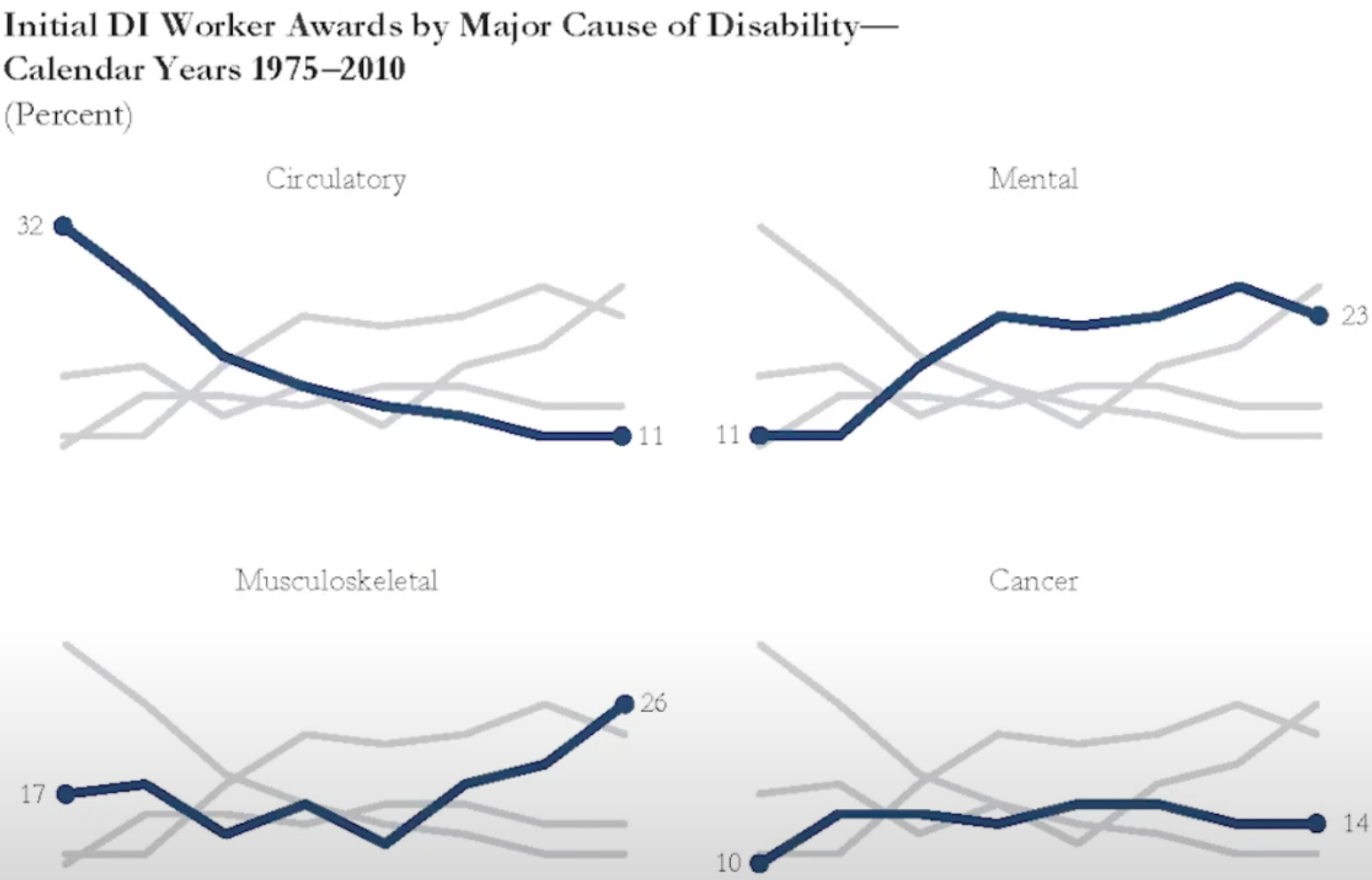
Flattening a 3D Chart



Bad Visualization Improved 5

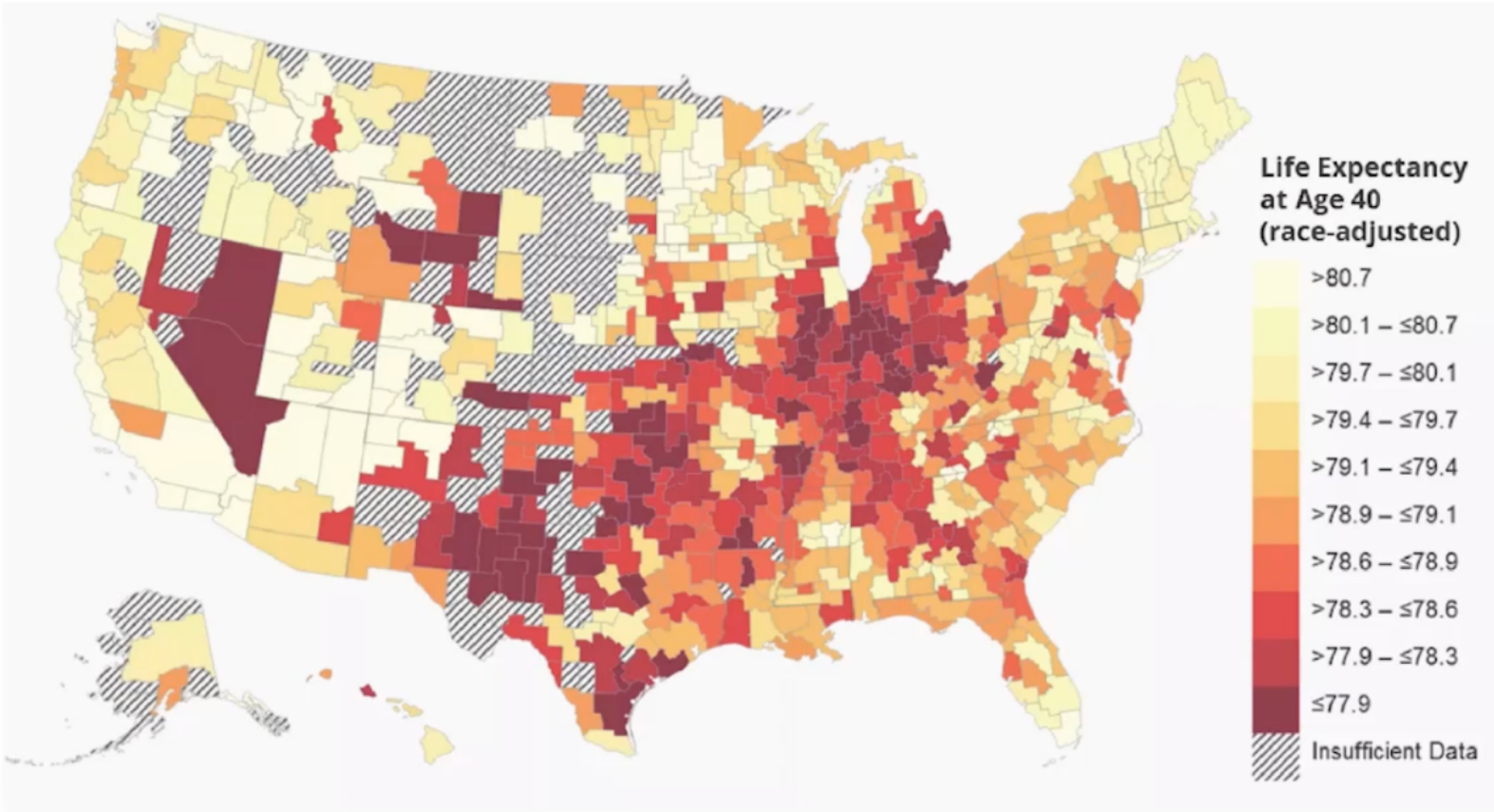
Figure 6B

Revising the Spaghetti Chart



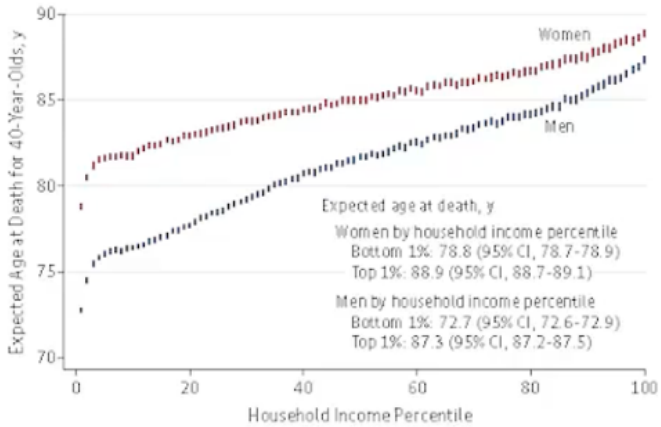
Life Expectancy

A data set in pictures: Chetty et al, JAMA, 2016



A data set in pictures: Chetty et al, JAMA, 2016

Figure 2. Race- and Ethnicity-Adjusted Life Expectancy for 40-Year-Olds by Household Income Percentile, 2001-2014



Mean household income in thousands, $[a]					
Women	24	45	71	112	1.9 million
Men	26	50	77	119	2.0 million

Life expectancies were calculated using survival curves analogous to those in Figure 1. The vertical height of each bar depicts the 95% confidence interval. The difference between expected age at death in the top and bottom income percentiles is 10.1 years (95% CI, 9.9-10.3 years) for women and 14.6 years (95% CI, 14.4-14.8 years) for men. To control for differences in life expectancies across racial and ethnic groups, race and ethnicity adjustments were calculated using data from the National Longitudinal Mortality Survey and estimates were reweighted so that each income percentile bin has the same fraction of black, Hispanic, and Asian adults.

[a] Averaged across years and ages. The data are in thousands unless otherwise indicated.

What are we trying to achieve?

- Show the data
- Do not lie about them
- Illustrate a story (causality? patterns? turning point?)
- Convince *and* inform
- Below are examples of not very good charts (from *Schwabish: An economist's guide to visualizing data*)
- What do you think the problems are?

Why visualize data?

- Two main goals:
  - **For yourself:** explore distributions, patterns, outliers (histograms, densities, nonparametric fits).
  - **For others:** tell a clear, truthful story about the results.
- Academic and non-academic success depends heavily on the ability to communicate using effective visualizations.
- Good charts often transfer well between articles and presentations (unlike long tables).

Know your audience

- Academic peers vs. general readers (e.g., the *NYT*) may require different aesthetic choices and explanations.
- Journals differ: some prefer a more technical style, others more artistic; aim for clarity acceptable in various contexts.

Main principles

- **Make charts self-contained:** axes, units, labels, sources; understandable even without the text.
- **Show the data honestly:** avoid distortions (e.g., bar charts should begin at zero).
- **Reduce clutter:** remove unnecessary elements (grid, thick axes, decorative textures) if they do not help readability.
- **Choose a purpose:** every chart should add more than a table and support your narrative.

Inspirations and sources

- Edward Tufte: maximize *data-ink*, minimize *chartjunk*; dense yet readable charts; prefer horizontal layouts when possible.
- Stephen Few (FlowingData/other sources): practical rules and tutorials in R; emphasis on clarity and self-explanatory design.
- Robert Kosara (EagerEyes), Jonathan Schwabish: blogs, examples, and before/after corrections.

Common mistakes (found in practice)

- **Lack of self-explanation:** unclear labels/abbreviations (e.g., "AO", "NC" with no legend).
- **Overly strong grid and axis lines:** thick zero lines, excessive ticks dominating the data.
- **Cluttered scatterplots:** hundreds of overlapping labels; consider highlighting only key cases.
- **Bar charts not starting at zero:** dramatically exaggerate differences.
- **3D for 2D data:** adds distortion without information.
- **"Spaghetti" line charts:** too many series to follow.
- **Incorrect use of color:** red/green combinations difficult for colorblind users; ensure grayscale readability.

Literature

- Slatkin, B. (2019). *Effective Python: 90 specific ways to write better Python.* Addison-Wesley Professional.
- Beazley, D., & Jones, B. K. (2013). *Python Cookbook: Recipes for mastering Python 3.* O'Reilly Media.
- Lecture 23: *Visualizing Data*, MIT OpenCourseWare — Esther Duflo, Nobel laureate in economics: YouTube
- Wikipedia: Data and Information Visualization

Additional materials — data visualization

- *Yau: Flowing Data* contains many R tutorials (flowingdata.com/category/tutorials) and a four-week self-guided minicourse.
- R can also handle and create maps.
- At motioninsocial.com/tufte you can find Tufte-style charts created in R.

e.weychert@uw.edu.pl