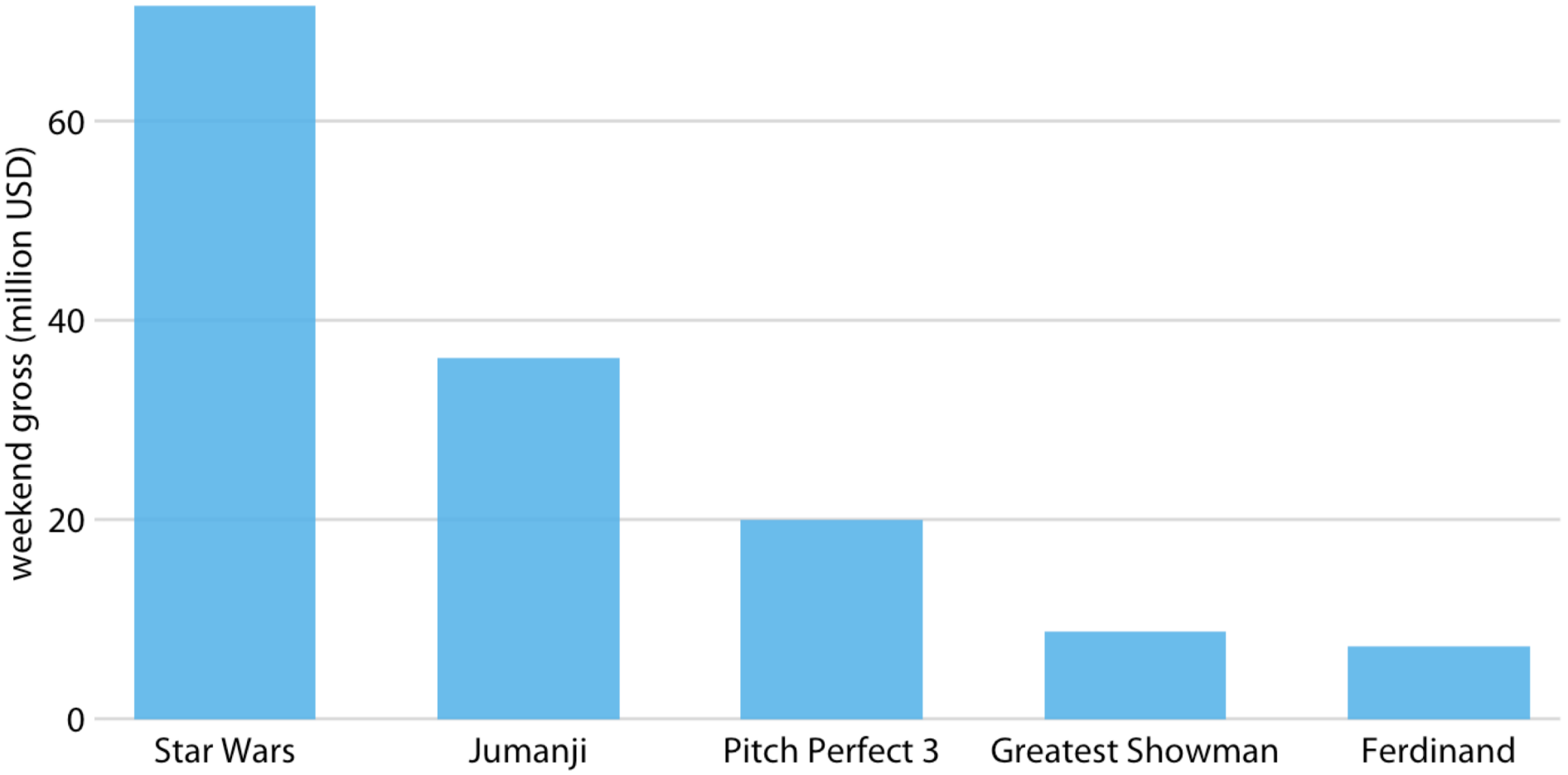
VISUALIZATION TECHNIQUES

PERE-PAU VÁZQUEZ – VIRVIG GROUP – UPC

OUTLINE

- **Displaying quantities**
- Displaying distributions
- Displaying proportions
- Displaying relationships
- Displaying time series
- Displaying geospatial data
- Other charts
- Uncertainty

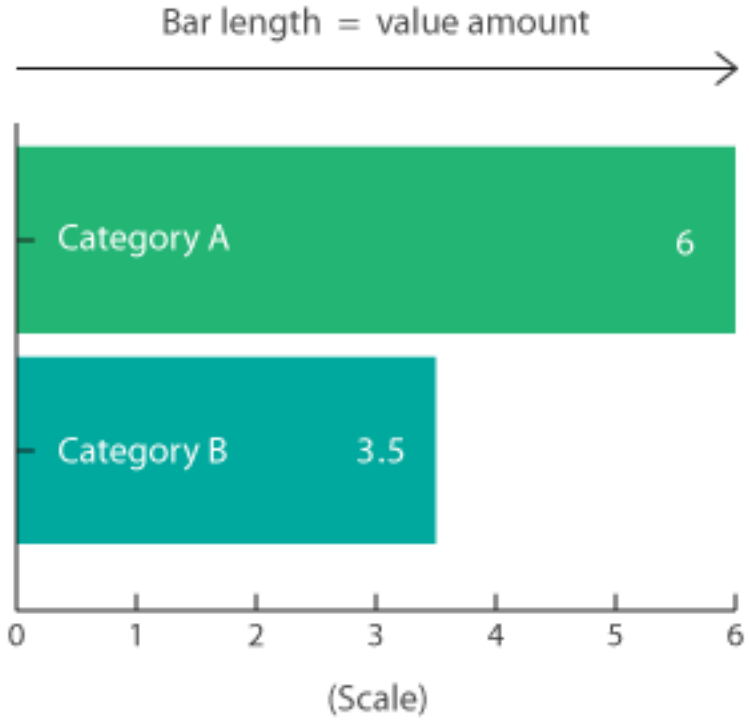
DISPLAYING QUANTITIES.
BAR CHARTS



https://clauswilke.com/dataviz

DISPLAYING QUANTITIES. BAR CHARTS

- Data: One key, one value
  - Data: 1 category attribute, 1 quantity attribute
  - Mark: lines
- Tasks:
  - Compare/lookup (really easy)
  - Scales to hundreds of elements



https://datavizcatalogue.com/

DISPLAYING QUANTITIES. BAR CHARTS

- Tips
  - Label orientation: For long labels, consider rotating bars, not labels
    - Vertical labels are difficult to read
  - Adequate order:
    - Alphabetical order if we want to make label search easier
    - By quantity if we want to facilitate value search

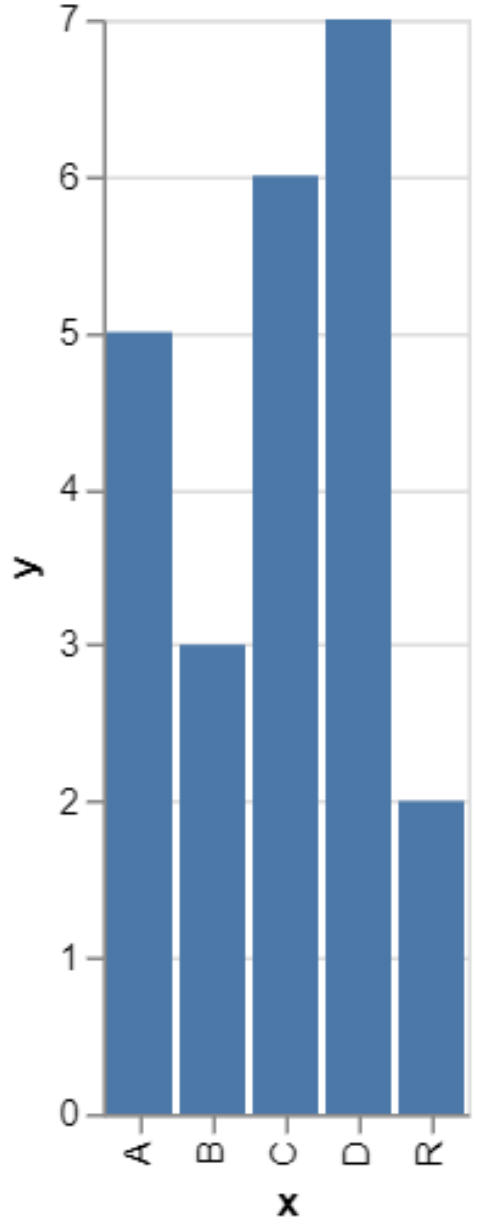
DISPLAYING QUANTITIES. BAR CHARTS

- Guidelines:
  - Always start at zero
  - Make labels easy to read (horizontal if possible)
  - Order based on data or labels
  - By default, prefer neutral colors
  - Gridlines if precision is required
  - If data is ordered (e.g., in time), line charts typically better
  - Don't use hundreds of bars

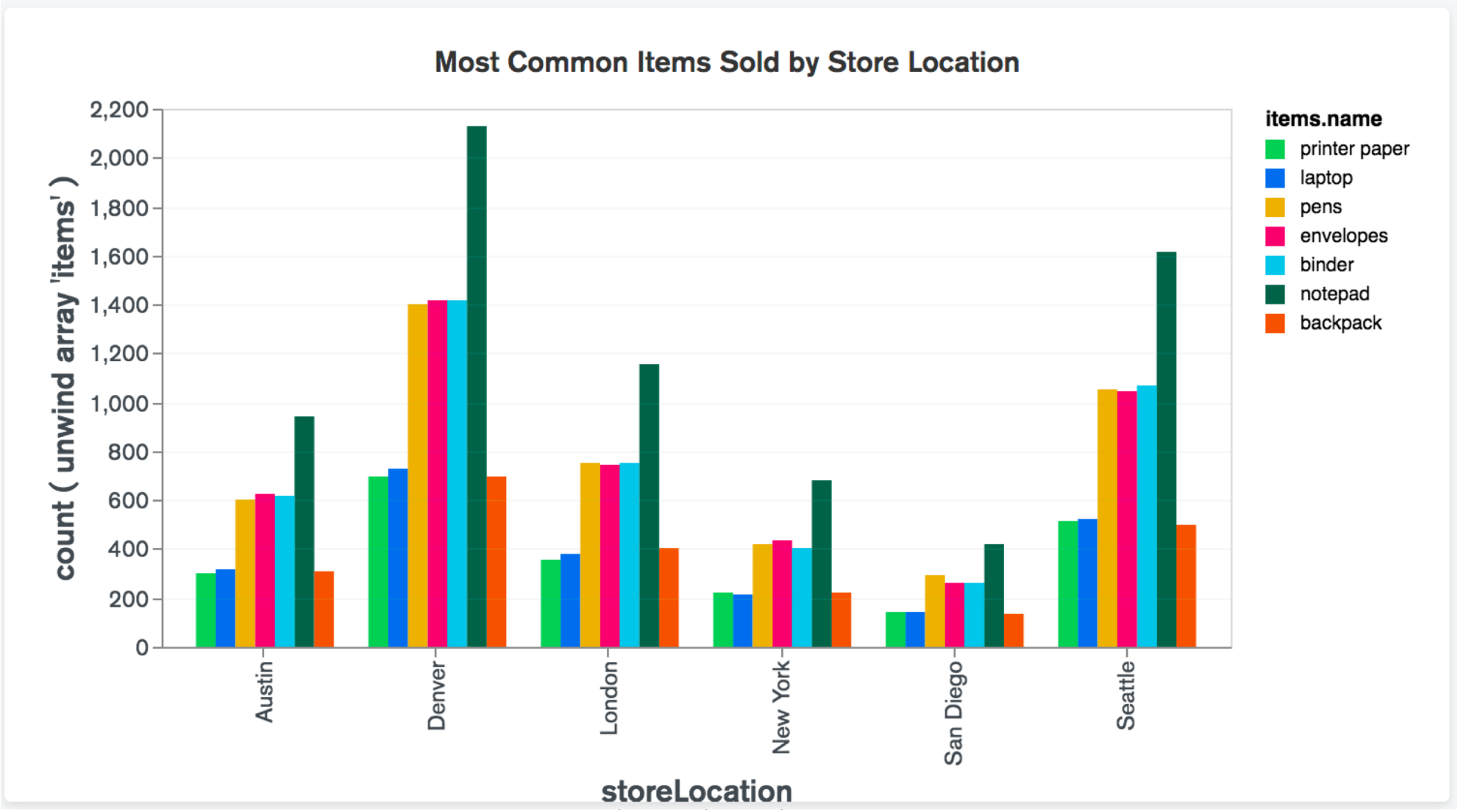
DISPLAYING QUANTITIES. BAR CHARTS

```
data = alt.Data(values=[{'x': 'A', 'y': 5},
                        {'x': 'B', 'y': 3},
                        {'x': 'C', 'y': 6},
                        {'x': 'D', 'y': 7},
                        {'x': 'R', 'y': 2}])

alt.Chart(data).mark_bar().encode(
    x='x:O',
    y='y:Q'
)
```

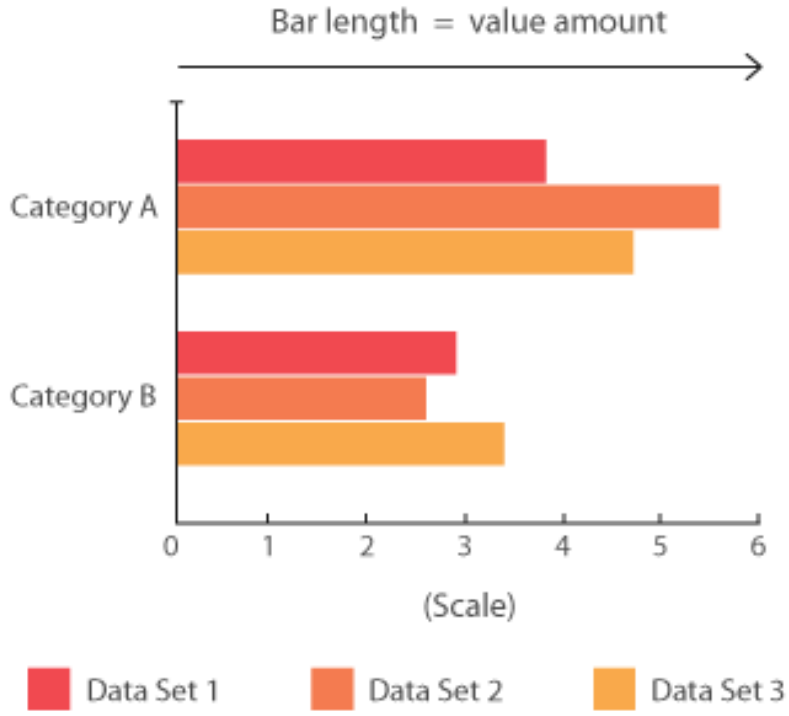


DISPLAYING QUANTITIES.
PAIRED BAR CHARTS (AKA GROUPED)



DISPLAYING QUANTITIES. PAIRED BAR CHARTS

- Data: One value (quant.), two keys (cat.)
- Marks: lines
- Channels:
  - Length to express quantity
  - Color hue for one key
  - Spatial regions for the other key
    - Separated horizontally, aligned vertically
    - Sometimes ordered by quant attribute, sometimes only on the keys
      - By label (alphabetical)



https://datavizcatalogue.com/

DISPLAYING QUANTITIES. PAIRED BAR CHARTS

- Tasks:
  - Compare
    - Within same key (easier)
    - Among keys (not so easy)
    - Lookup values
- Scalability:
  - ≈ one dozen secondary keys (less than bar charts)

DISPLAYING QUANTITIES. PAIRED BAR CHARTS

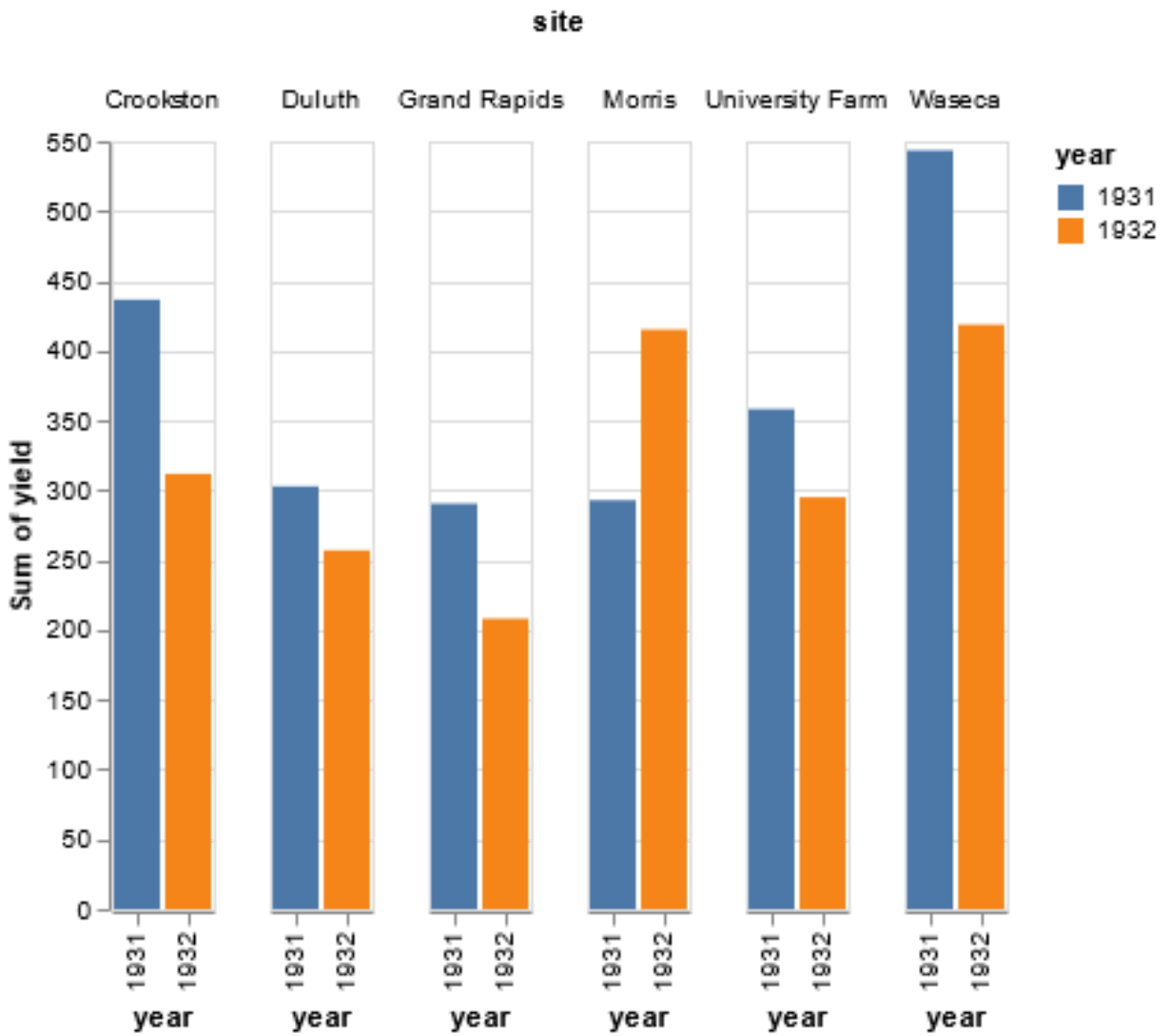
- Guidelines:
  - Always start at zero
  - Bar charts guidelines also apply
  - Don't use them if one category is time

DISPLAYING QUANTITIES. PAIRED BAR CHARTS

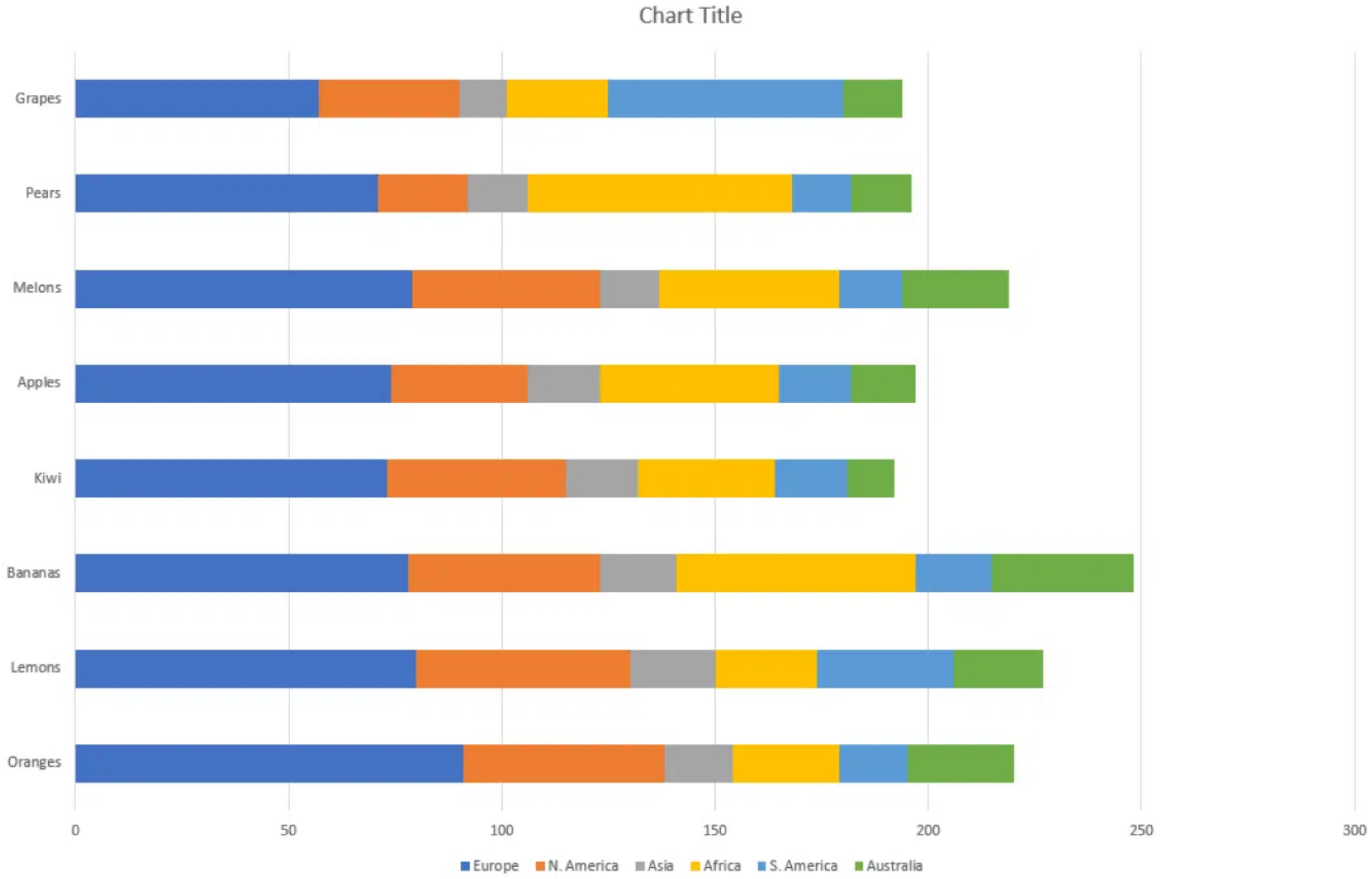
```
import altair as alt
from vega_datasets import data

source = data.barley()

alt.Chart(source).mark_bar().encode(
    x='year:O',
    y='sum(yield):Q',
    color='year:N',
    column='site:N'
)
```

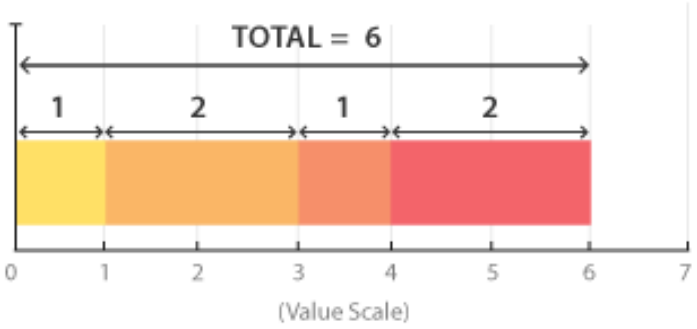


DISPLAYING QUANTITIES. STACKED BAR CHART



DISPLAYING QUANTITIES. STACKED BAR CHART

- Data: One value, two keys
- Channels:
  - Same as before, but bars stacked vertically
  - Color hue for one key, spatial regions for the other key
- Tasks:
  - Compare
    - Within same key (difficult), among keys (also difficult), lookup values
- Scalability:
  - ≈ one dozen secondary keys (less than bar charts)



https://datavizcatalogue.com/

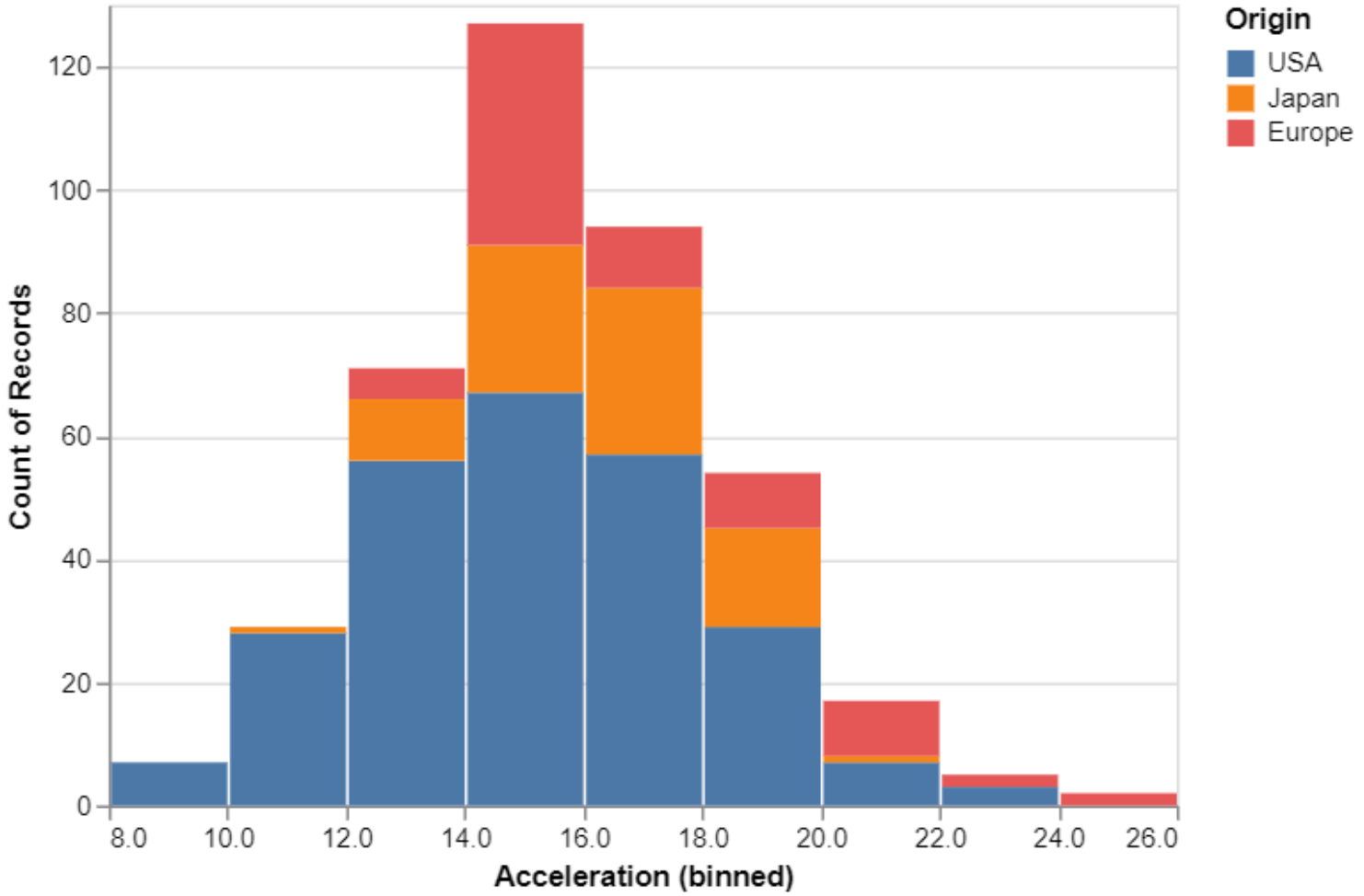
DISPLAYING QUANTITIES. STACKED BAR CHART

- Guidelines:
  - Always start at zero
  - Same guidelines as bar charts
  - Difficult to compare between groups
  - Difficult to compare within groups
  - Not use when total quantity does not make sense
  - Use with few categories

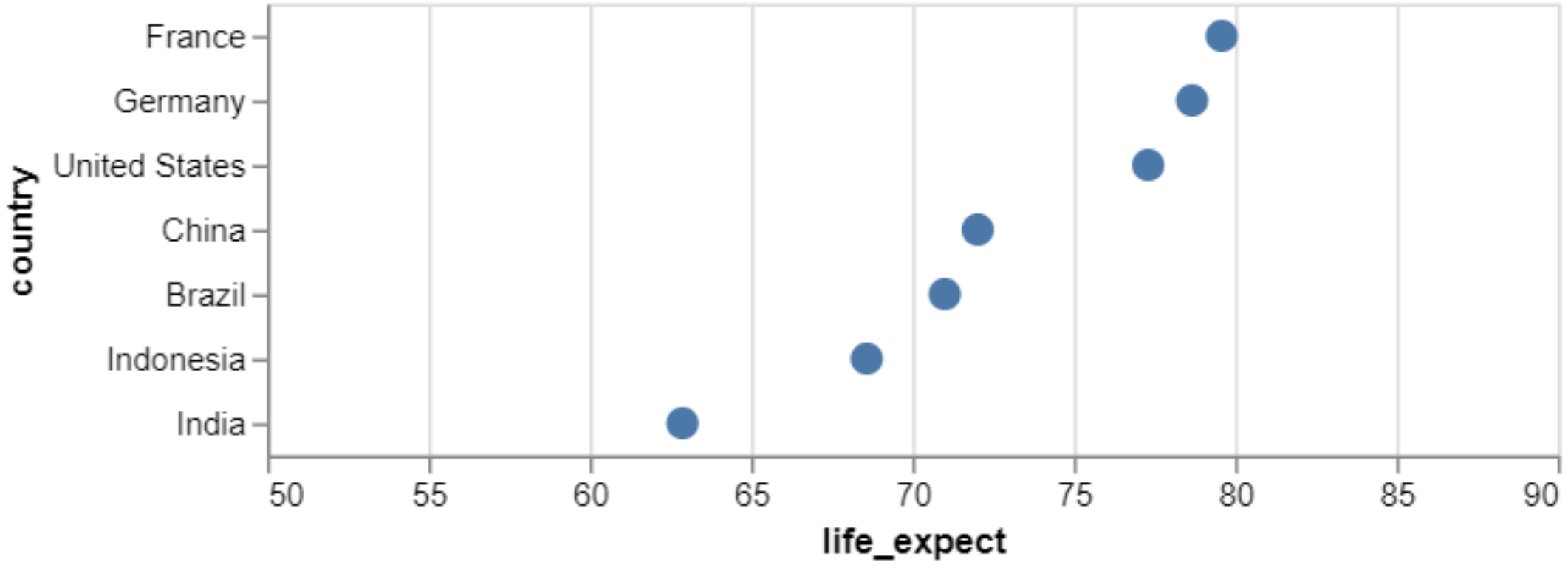
DISPLAYING QUANTITIES. STACKED BAR CHART

```
df = data.cars.url

alt.Chart(df).mark_bar(
).encode(
    x = alt.X('Acceleration:Q', bin=True),
    y = alt.Y('count():Q'),
    color = alt.Color('Origin:N', sort='-y'),
)
```

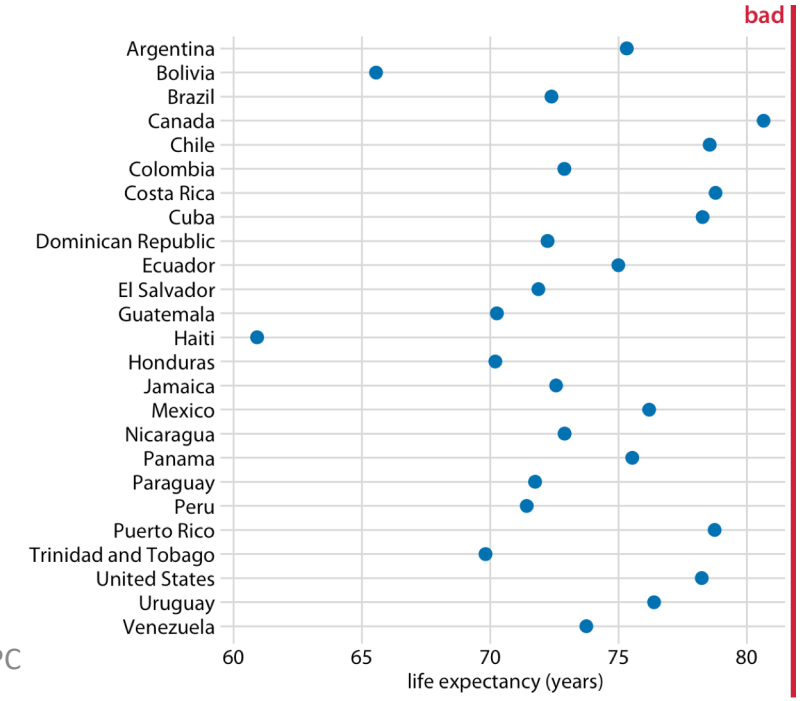


DISPLAYING QUANTITIES.
DOT PLOT



DISPLAYING QUANTITIES.
DOT PLOT

- Do not need to start at zero
  - Suitable when small differences must be shown
  - Bar charts might lead the attention away from those differences
- Tips:
  - Must be ordered by quantities
    - The opposite makes the chart difficult to read

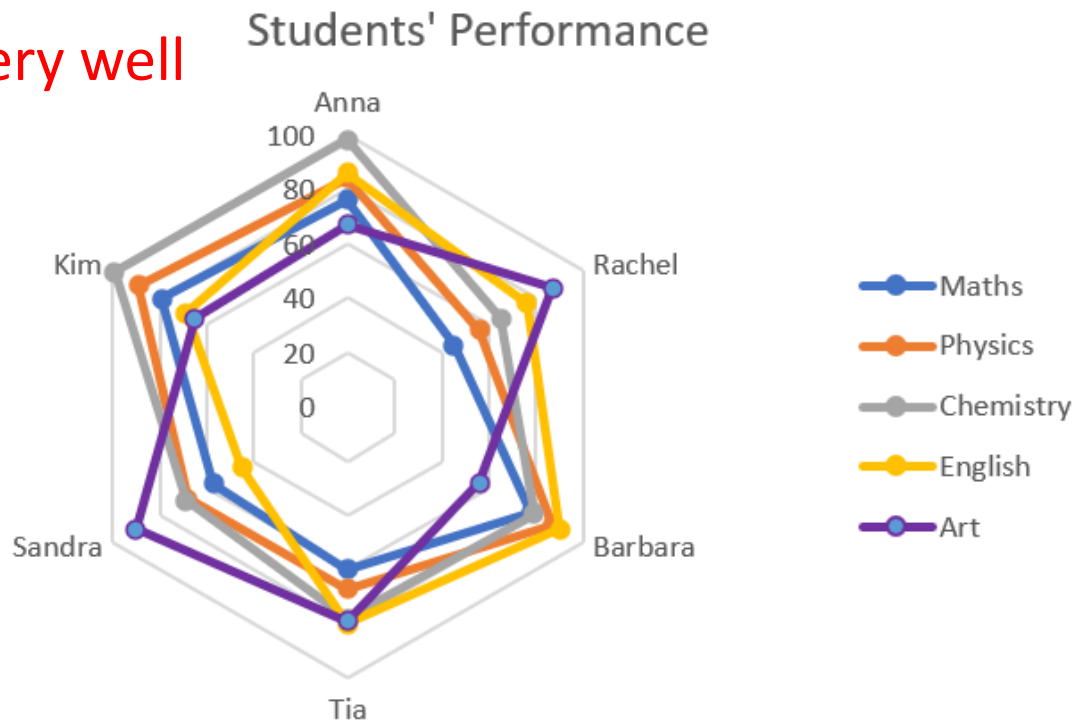


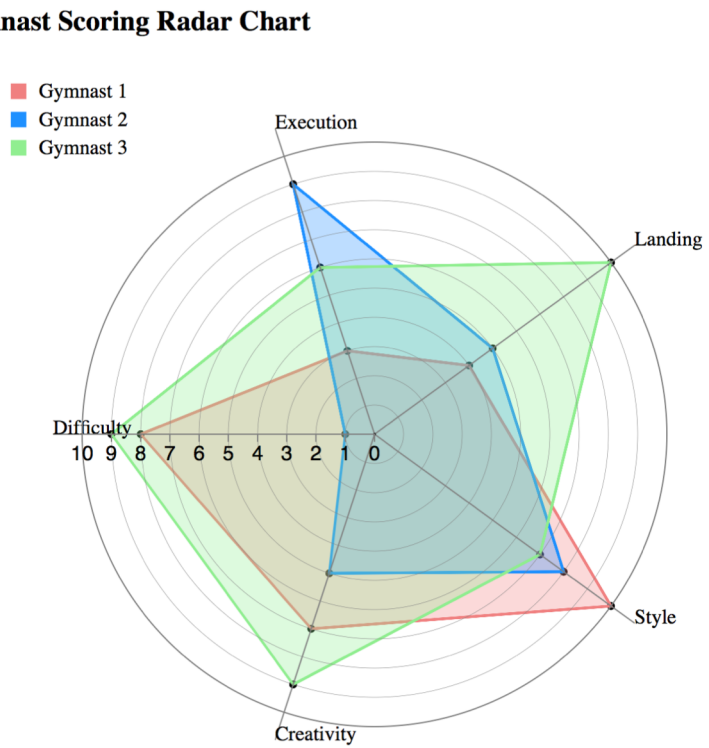
DISPLAYING QUANTITIES. DOT PLOT

- Guidelines:
  - Do not need to start at zero
  - Must be ordered by quantity
  - Suitable when small differences must be shown
  - If values are relevant, label axes suitably

DISPLAYING QUANTITIES. RADAR CHART

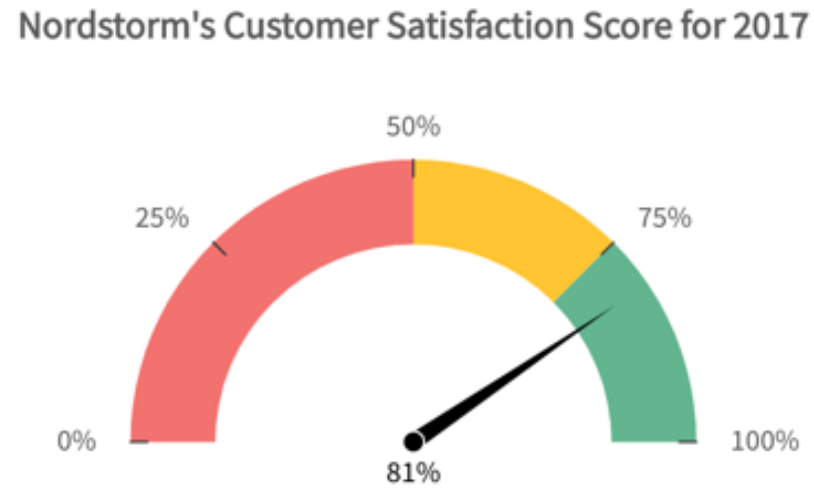
- Analogous to paired/grouped column charts
  - Multiple dimensions
  - Space efficient
  - Different designs
  - Do not scale very well
  - Can be small

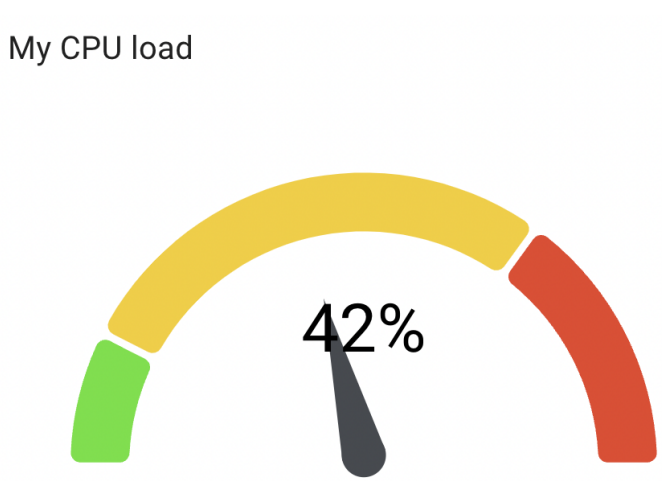




DISPLAYING QUANTITIES. GAUGE & BULLET CHARTS

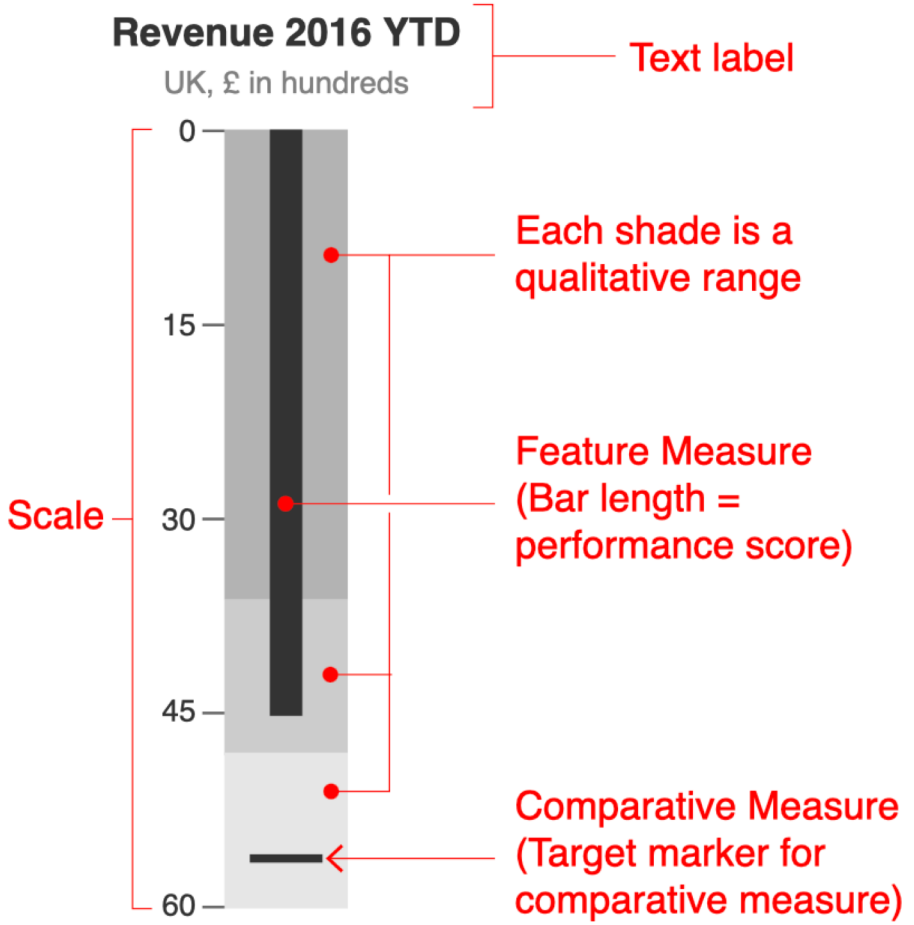
- Gauge charts:
  - Adaptation of real gauges
  - Very common in business analytics
    - Used to display KPIs
  - Current value (front) vs reference (background)
  - Using angle to encode values
    - Less optimal than other visual variables
  - Use too much space
  - Commonly include the data in text too





DISPLAYING QUANTITIES. GAUGE & BULLET CHARTS

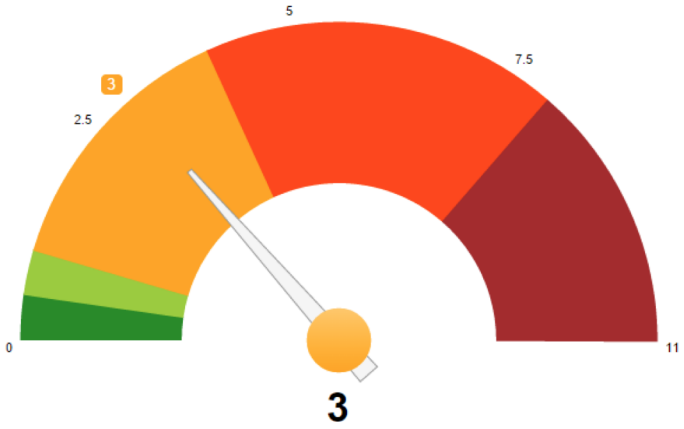
- Bullet charts:
  - Version of gauge charts using bars
  - Using the background of the bar chart to encode the reference value(s)
  - Space efficient
    - May encode multiple values in the same space
  - Better for perception (comparing lengths instead of angles)

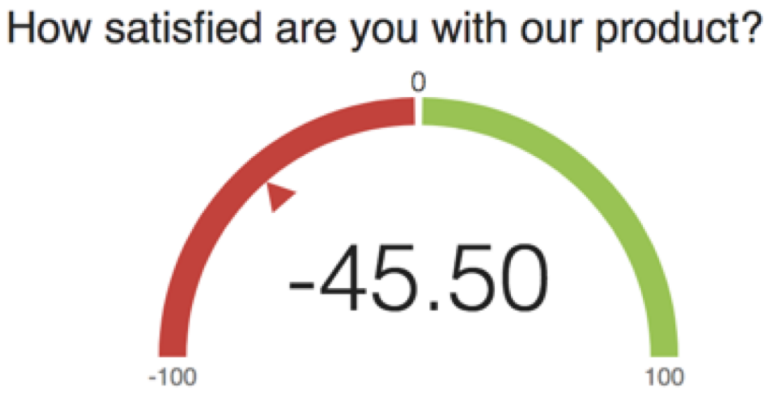


https://datavizcatalogue.com/

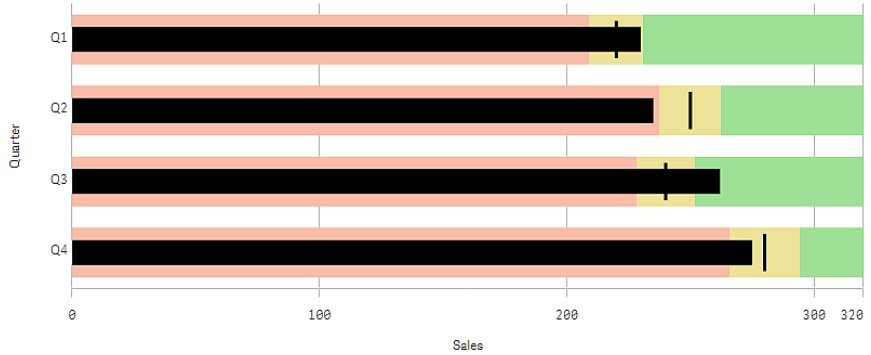
DISPLAYING QUANTITIES. GAUGE VS BULLET CHARTS

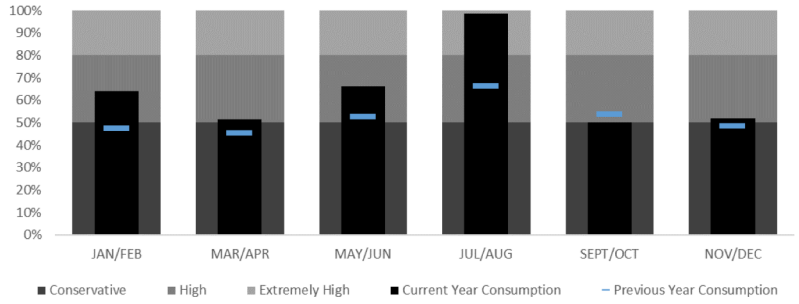
Gauge charts





Bullet charts





OUTLINE

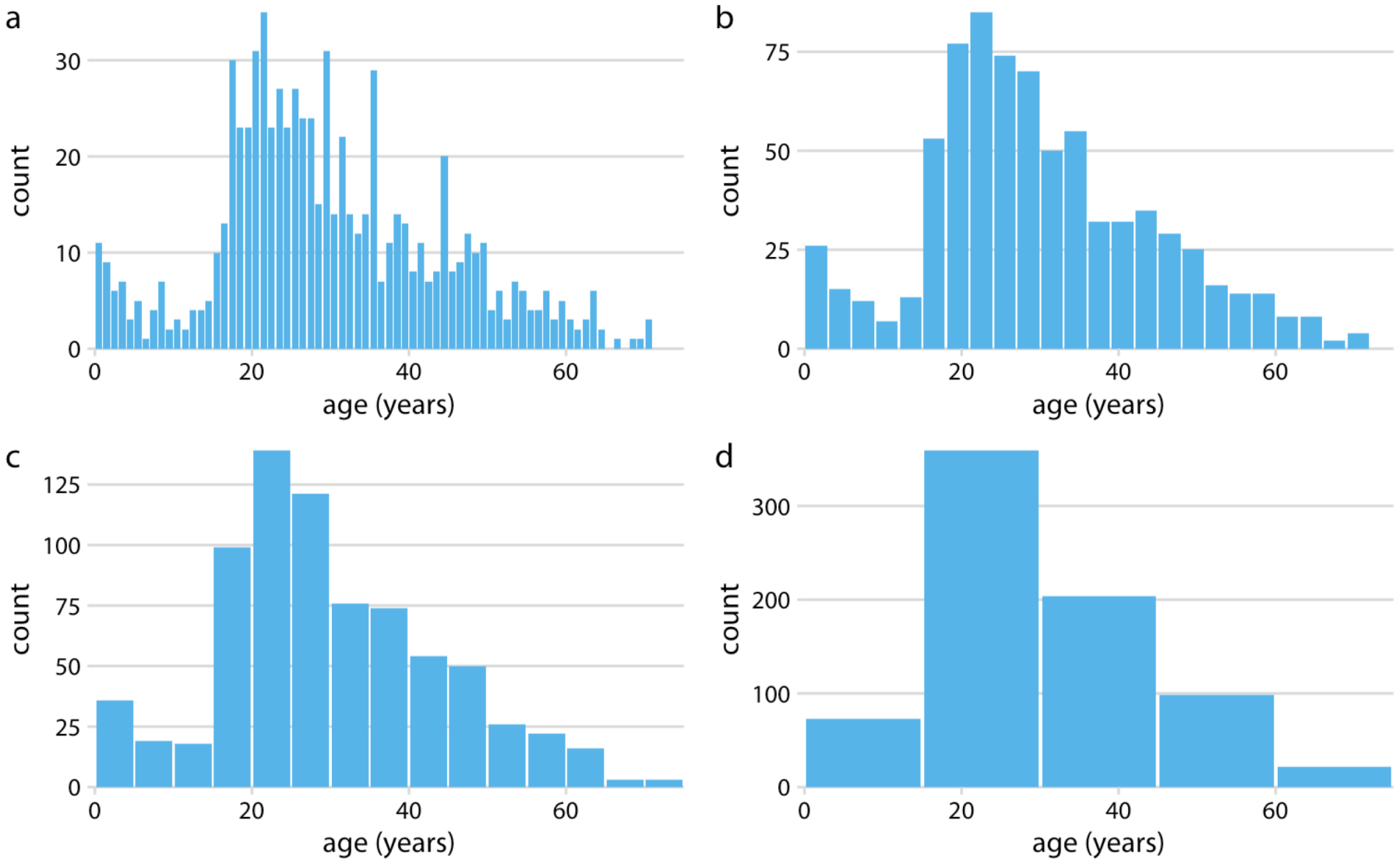
- *Displaying quantities*
- **Displaying distributions**
- Displaying proportions
- Displaying relationships
- Displaying time series
- Displaying geospatial data
- Other charts
- Uncertainty

DISTRIBUTIONS. HISTOGRAMS

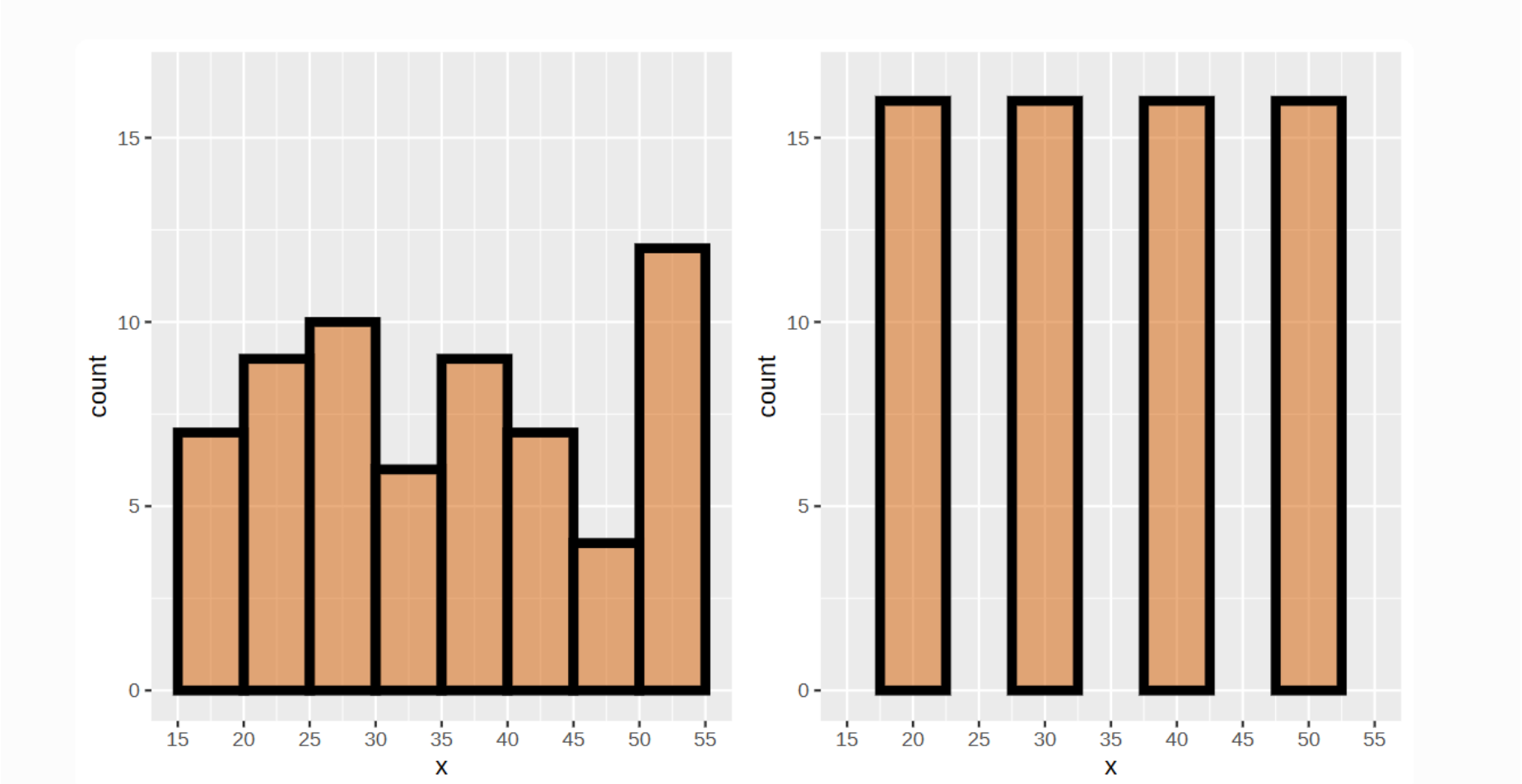
- Individual observations may be not adequate
  - Too many
  - Interested in trends/cumulative values…

- Binning information in value ranges: Histograms
  - Bar charts with binned information
  - Binning sizes may be important: Different distributions may appear
    - Explore different bin sizes

DISTRIBUTIONS. HISTOGRAMS



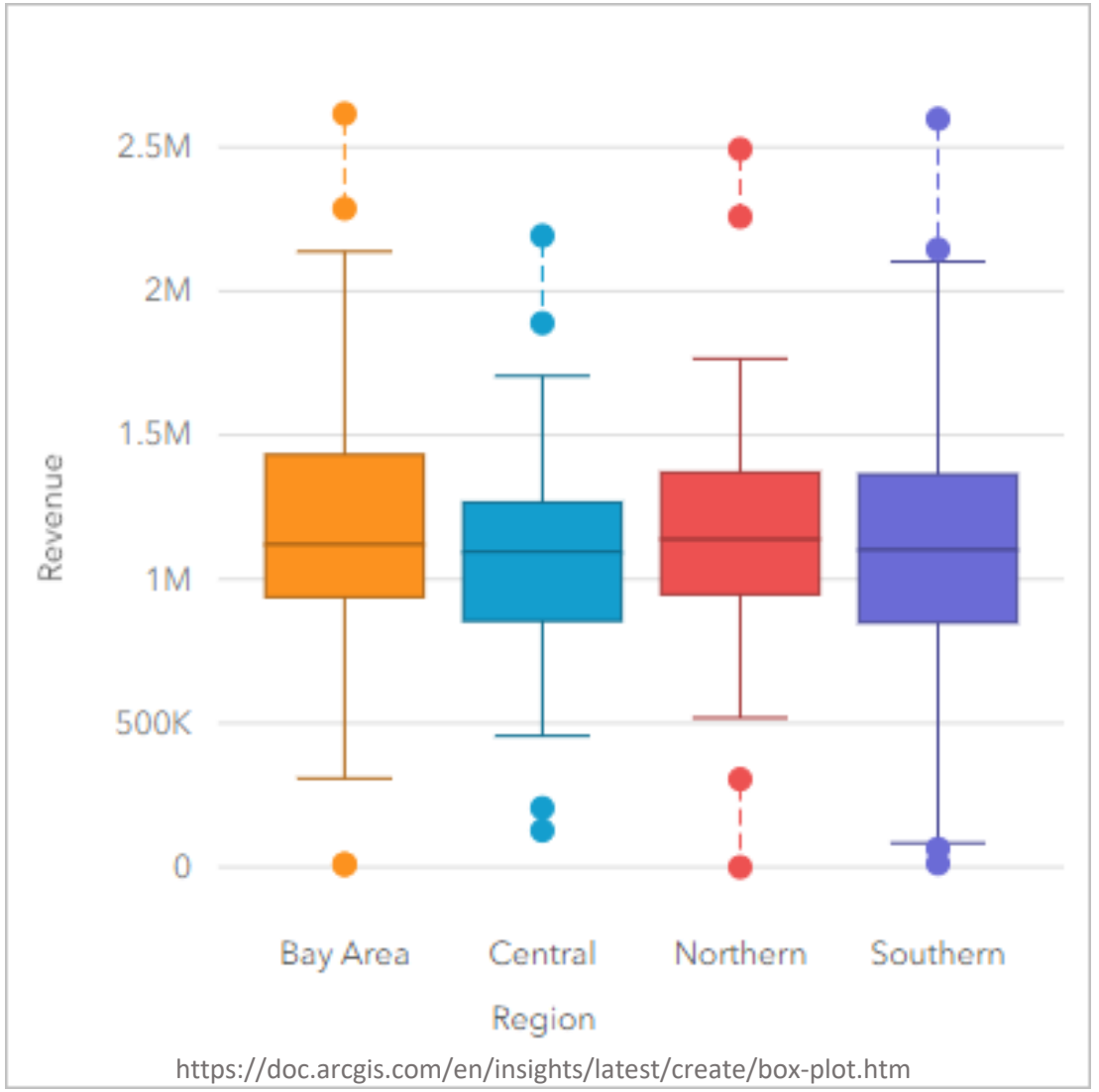
DISTRIBUTIONS. HISTOGRAMS



DISTRIBUTIONS.
BOX PLOTS

- Useful for several distributions at the same time
- Gives insights on data distribution inside
  - Statistical descriptors
    - Median, minimum, maximum, outliers…

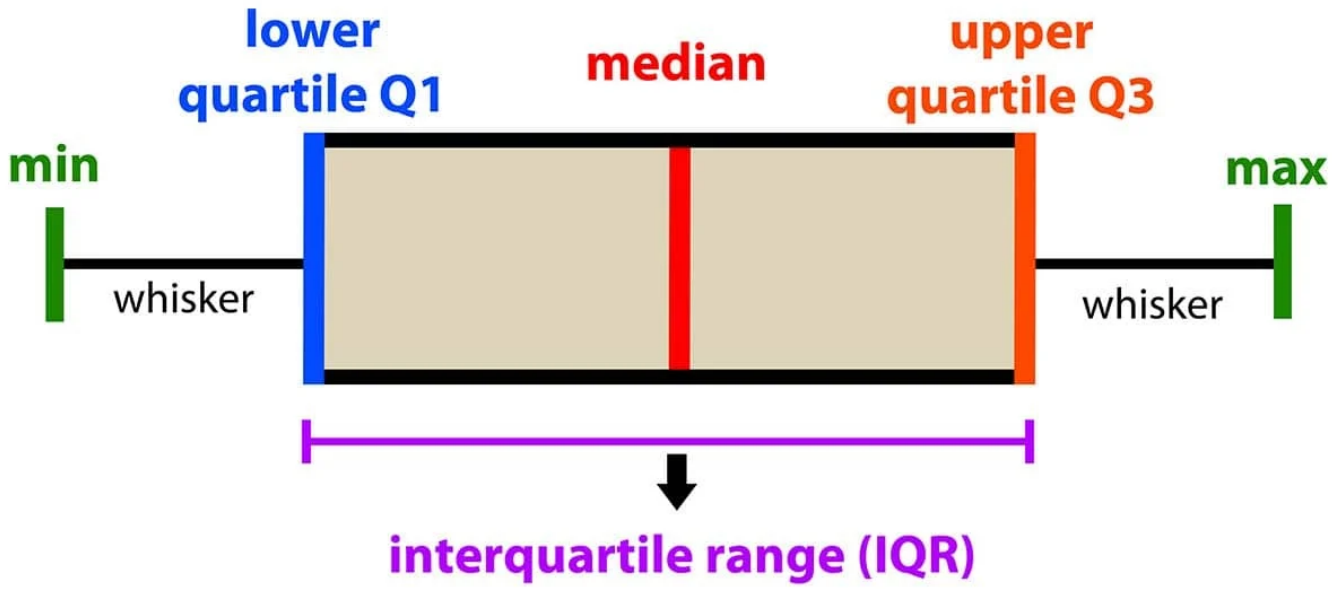
DISTRIBUTIONS.
BOX PLOTS



https://doc.arcgis.com/en/insights/latest/create/box-plot.htm

DISTRIBUTIONS. BOX PLOTS

- Construction



https://www.simplypsychology.org/boxplots.html

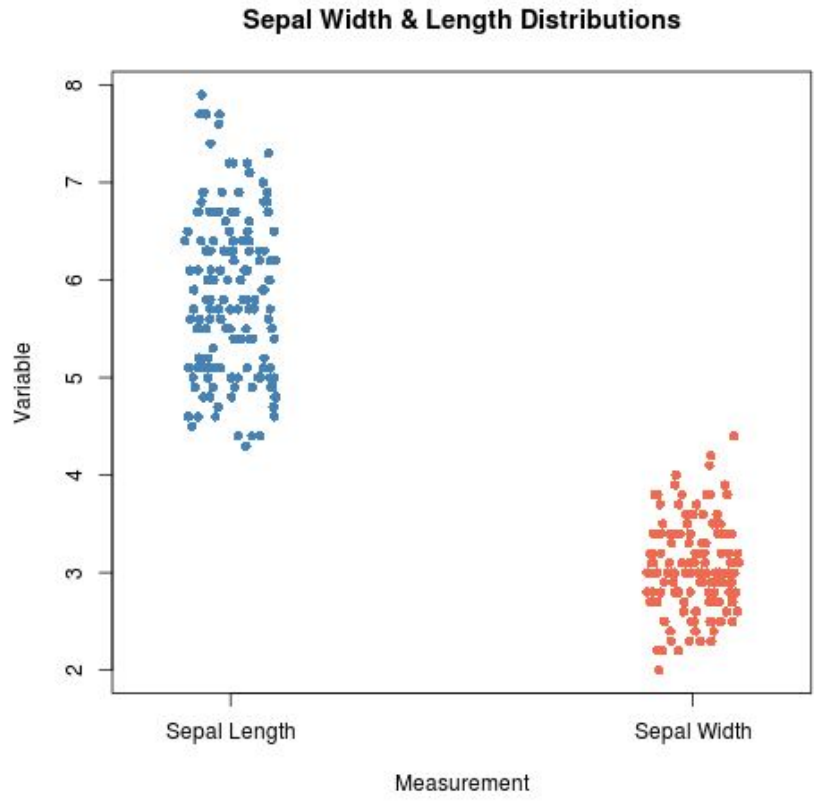
DISTRIBUTIONS.
BOX PLOTS

- Box plots hide/abstract too much data
- Hidden information may be relevant
- We can use alternative charts to show the internal distribution, e.g., violin plots, strip charts, beeswarm…

https://www.statology.org/strip-chart-r/

DISTRIBUTIONS.
VIOLIN PLOTS & STRIP CHARTS

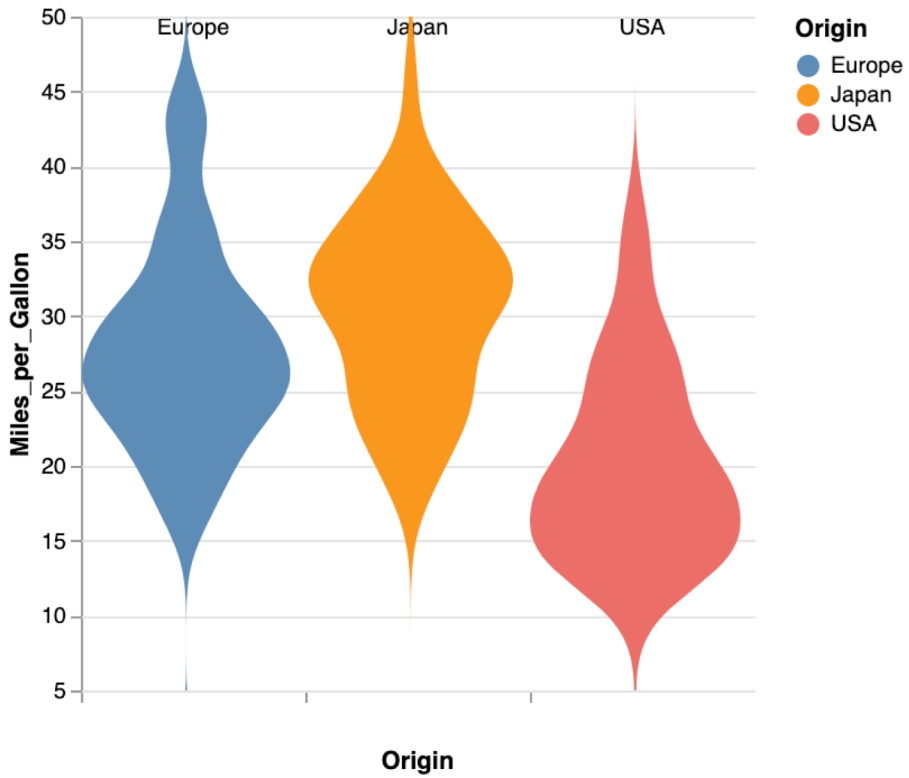
- Strip charts show all the data points
- Use random positioning in one axis (e.g., X) to avoid overlapping
- Positioning can be calculated in different ways
  - Random (jittering)
  - Using blue noise



https://www.statology.org/strip-chart-r/

DISTRIBUTIONS.
VIOLIN PLOTS & STRIP CHARTS

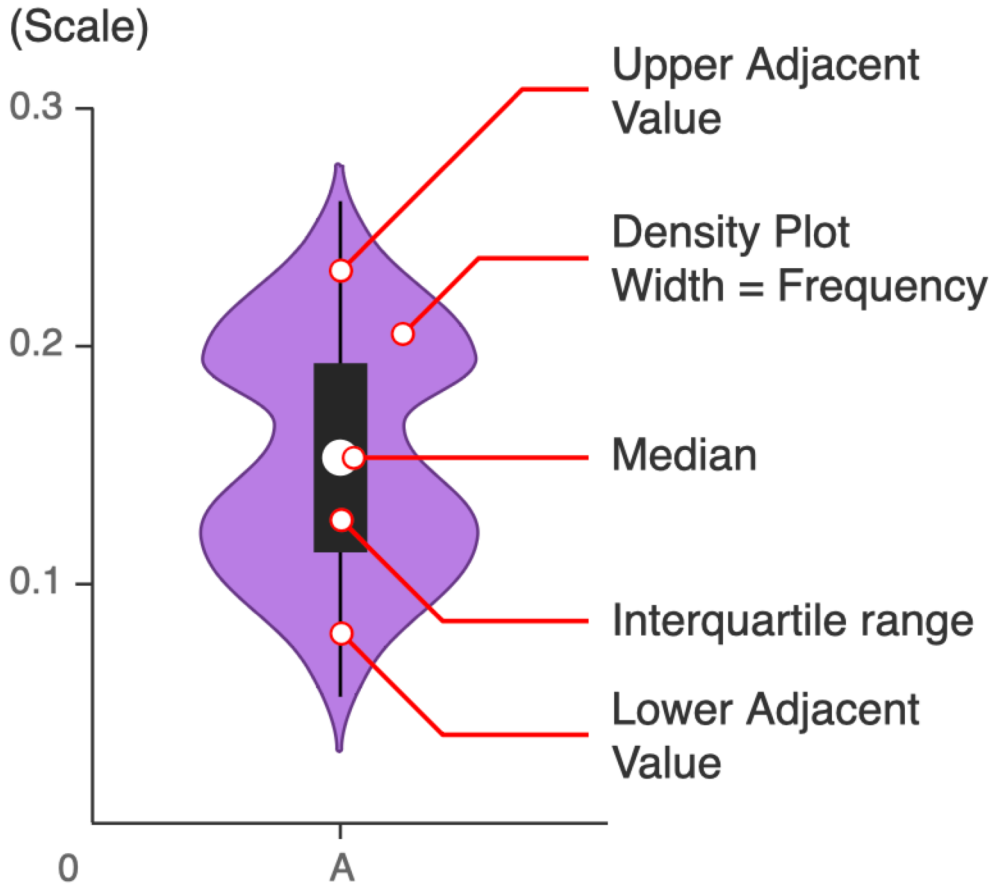
- Violin plots show a density chart to illustrate the distribution
  - Need to calculate the shape
  - May still hide some data



https://www.statology.org/strip-chart-r/

DISTRIBUTIONS.
VIOLIN PLOTS & STRIP CHARTS

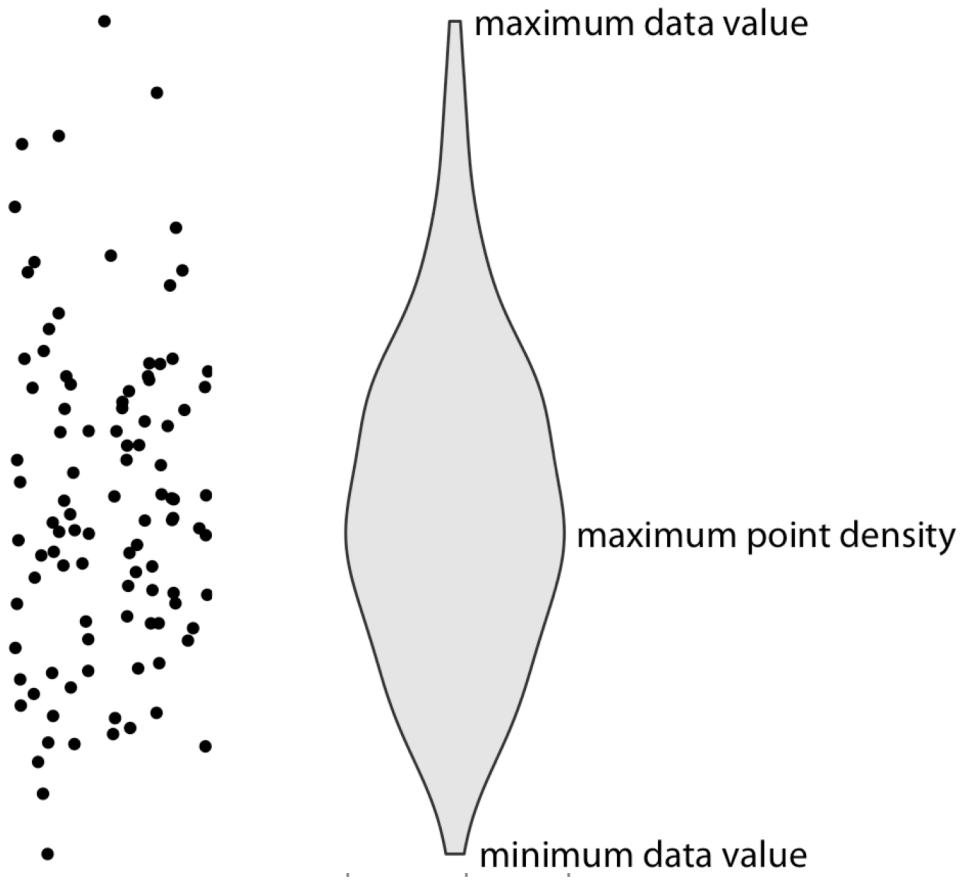
- Violin plots. Original design:
  - Includes the distribution **and a box plot**
  - Density is mirrored around the axis
    - For aesthetical reasons



https://datavizcatalogue.com/

DISTRIBUTIONS.
VIOLIN PLOTS & STRIP CHARTS

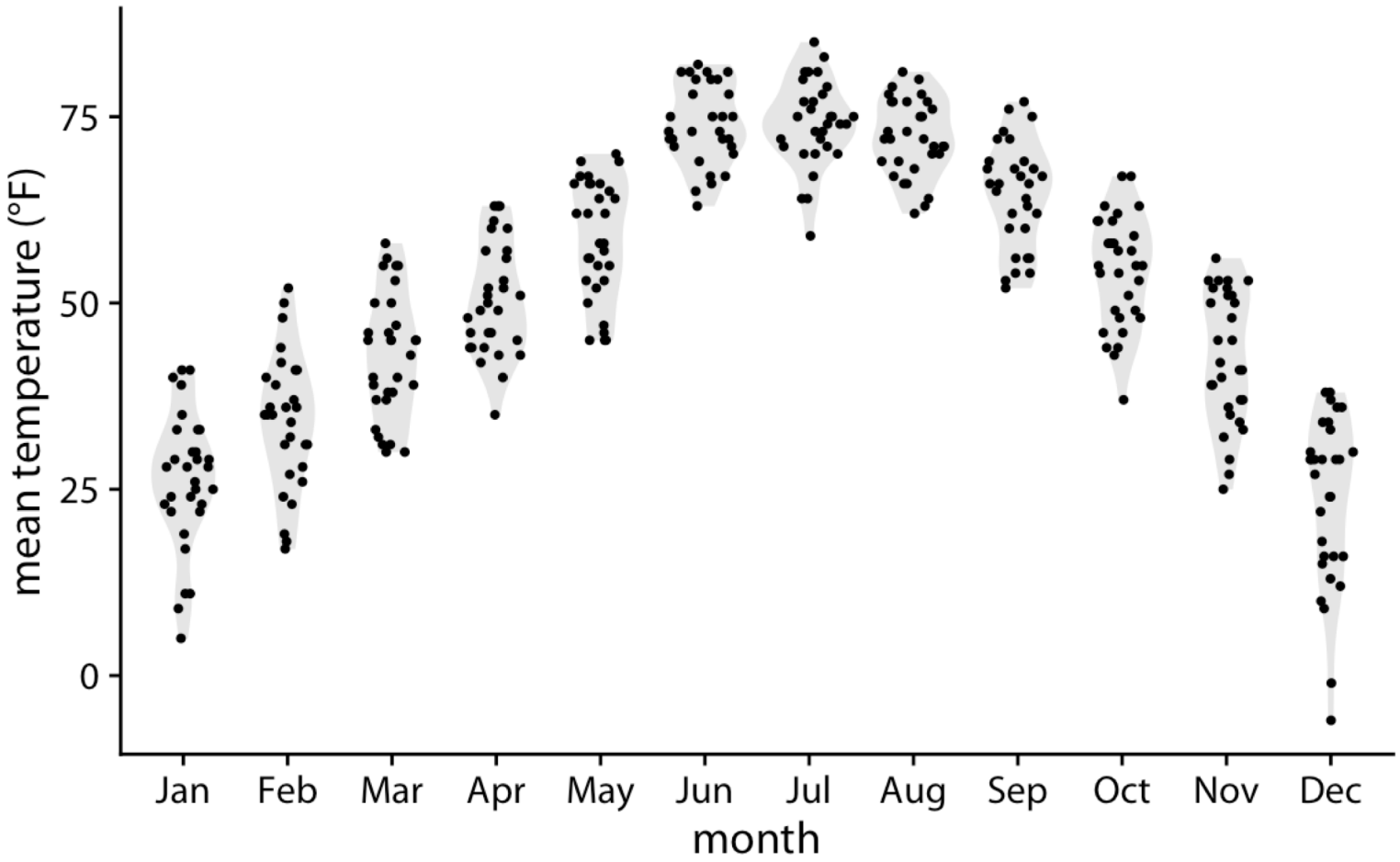
- Violin plots. Construction from a distribution:



https://clauswilke.com/dataviz

DISTRIBUTIONS.
VIOLIN PLOTS & STRIP CHARTS

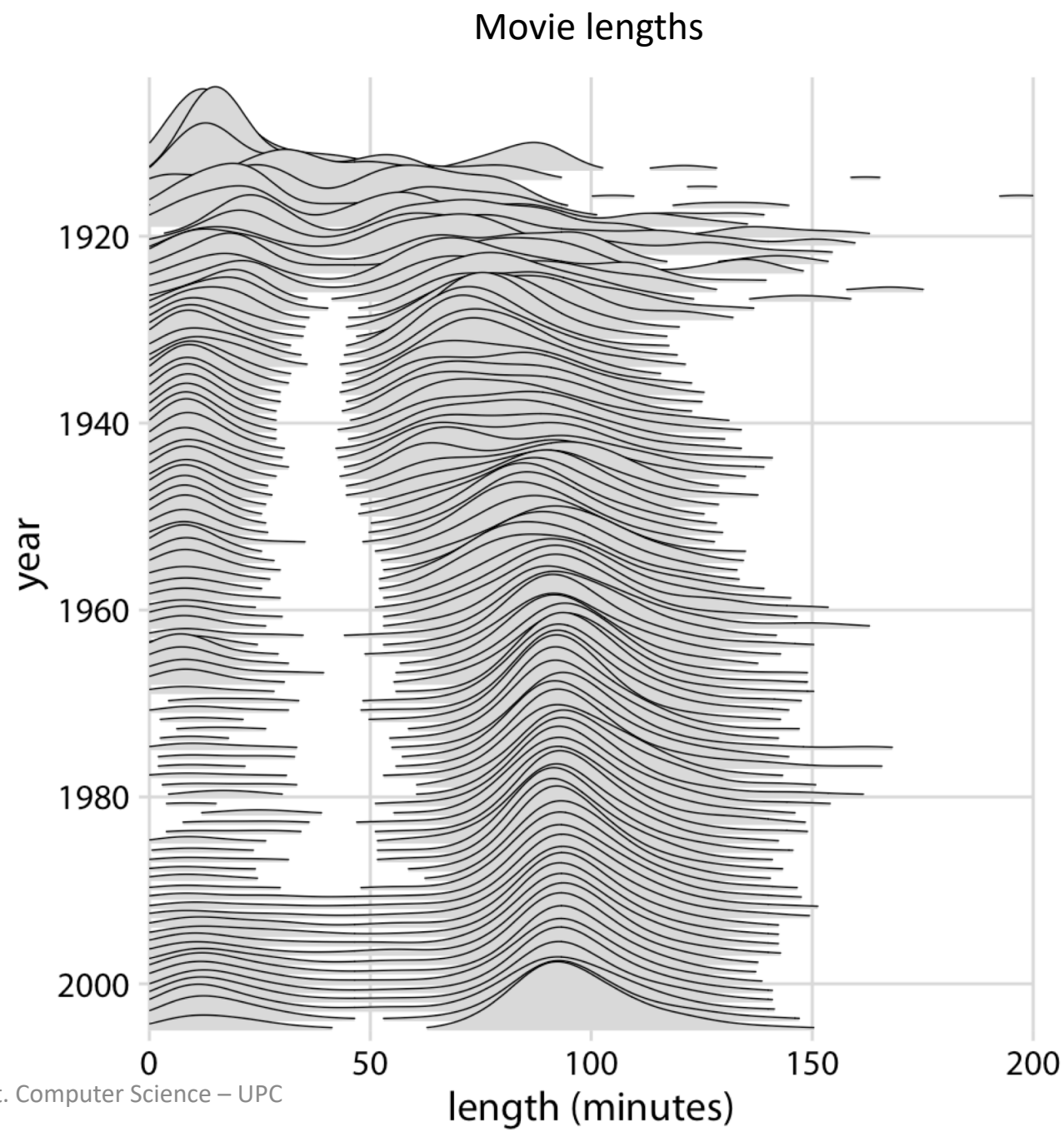
- Can combine both



:tps://clauswilke.com/dataviz

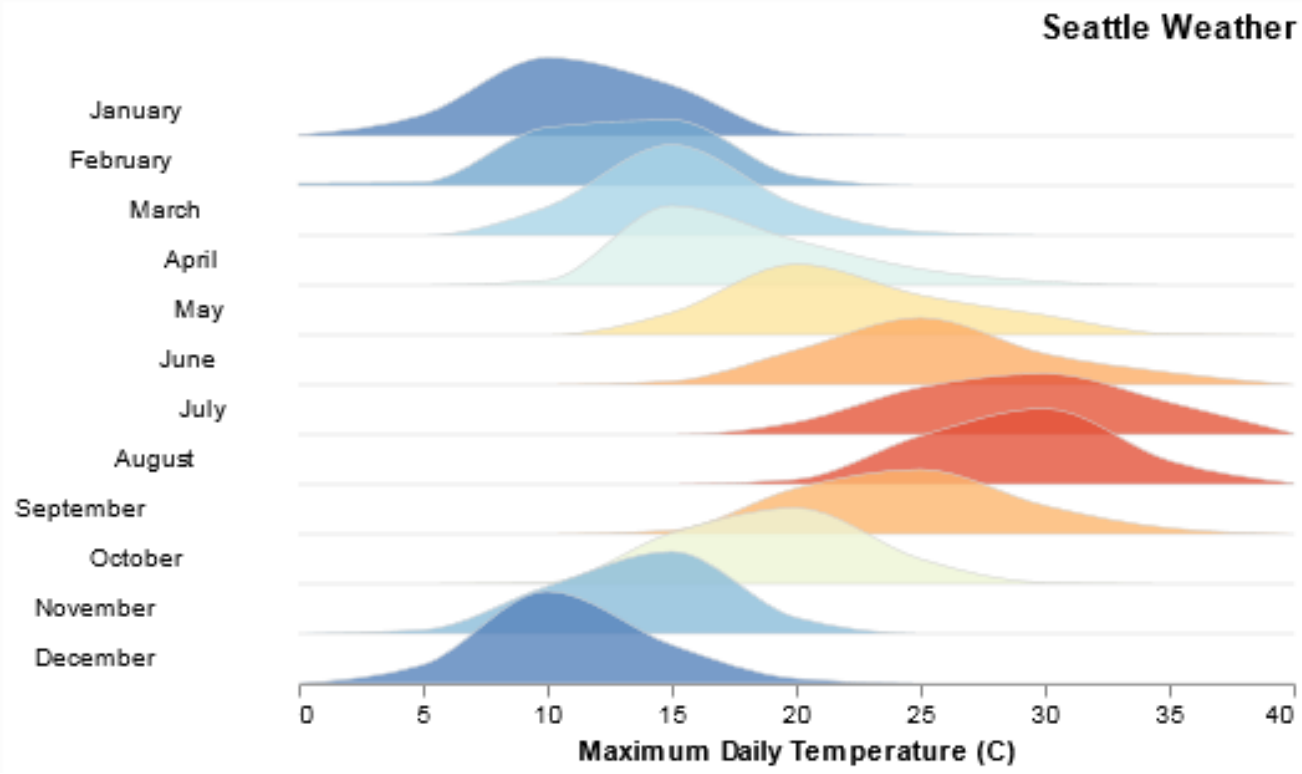
DISTRIBUTIONS. RIDGE PLOTS

- Like a half violin plot in horizontal (violin plots are symmetric)
  - Allows more data
  - Even overlapping if done carefully
  - No accurate value estimation possible



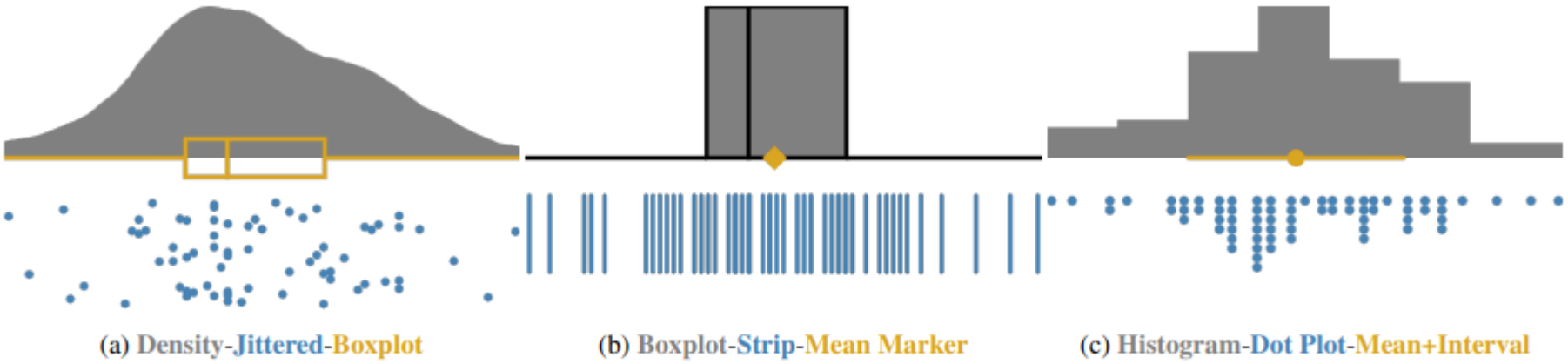
DISTRIBUTIONS.
RIDGE PLOTS

- Like a half violin plot in horizontal (violin plots are symmetric)
  - Allows more data
  - Even overlapping if done carefully



DISTRIBUTIONS. RAINCLOUD PLOTS

- Combination of three plots to show both the distribution and the key statistics



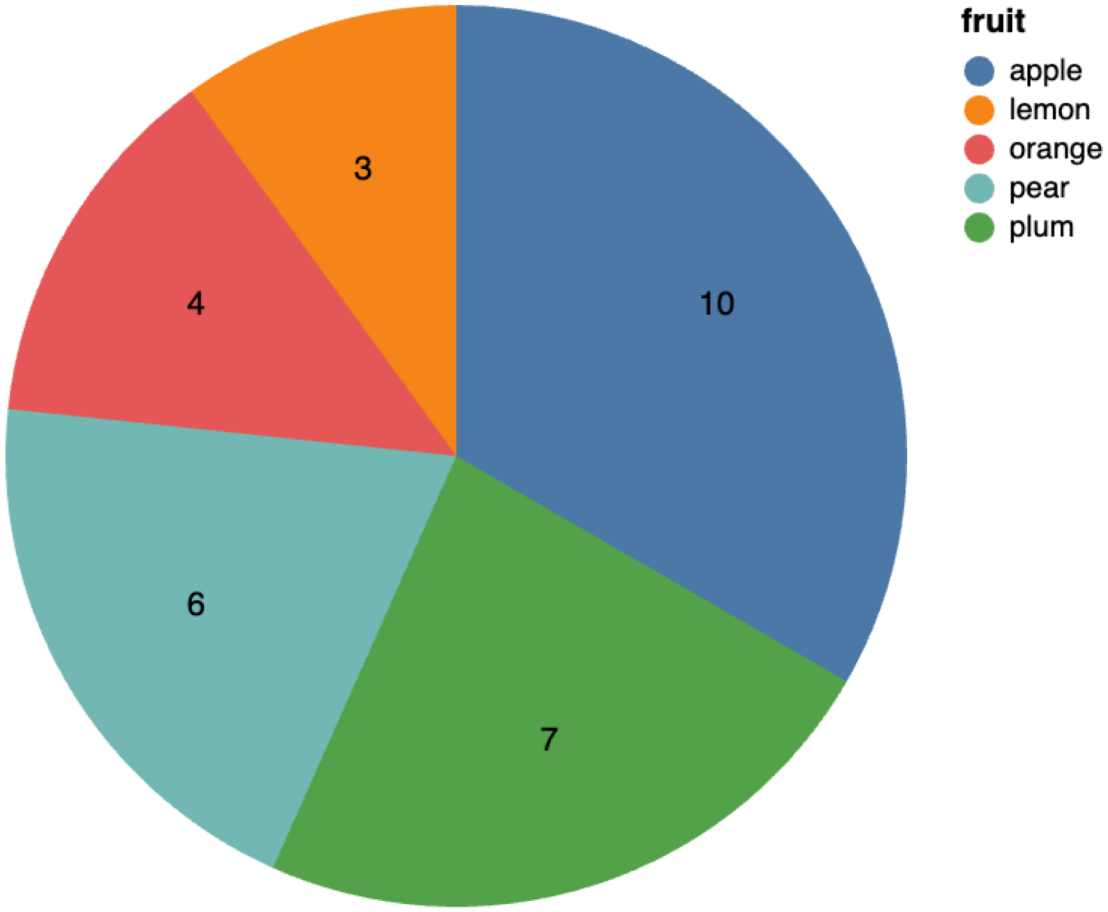
OUTLINE

- *Displaying quantities*
- *Displaying distributions*
- **Displaying proportions**
- Displaying relationships
- Displaying time series
- Displaying geospatial data
- Other charts
- Uncertainty

PROPORTIONS.
PIE CHARTS

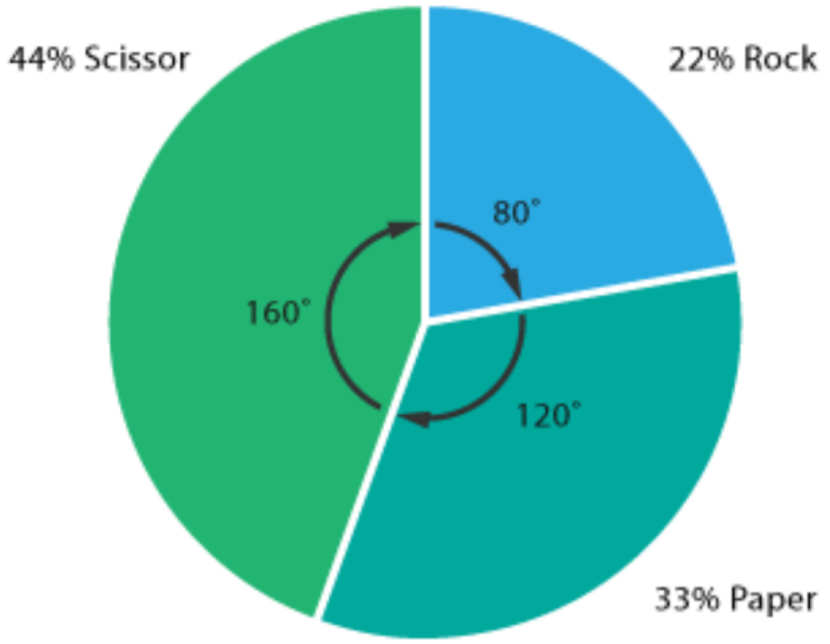
- Pie chart
  - Area marks with **angle** channel (*angle* encodes quantity)
  - Accuracy: angle/area much less accurate than line length
- Data
  - 1 key attribute, 1 quantitative value attribute
- Task
  - Part-to-whole judgements
- **Perceptual issues!!!!**

PROPORTIONS.
PIE CHARTS



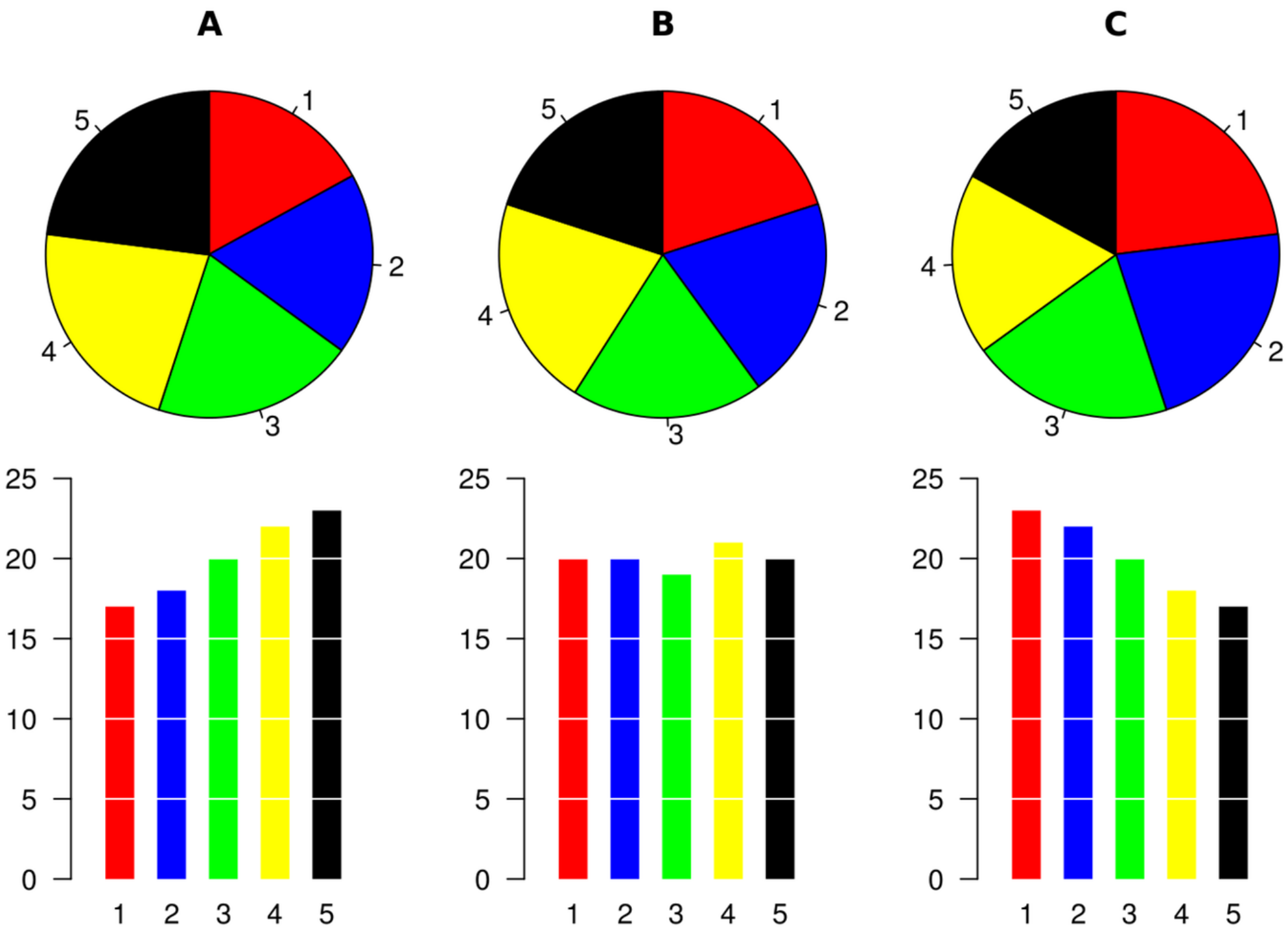
PROPORTIONS.
PIE CHARTS

- Data: One key, one value
- Channels:
  - Angle: quantity
  - Color hue for other key
- Tasks:
  - Part-to-whole comparison
    - Difficult for similar regions, difficult to estimate values
- Scalability:
  - ≈ few categories (3-5), different



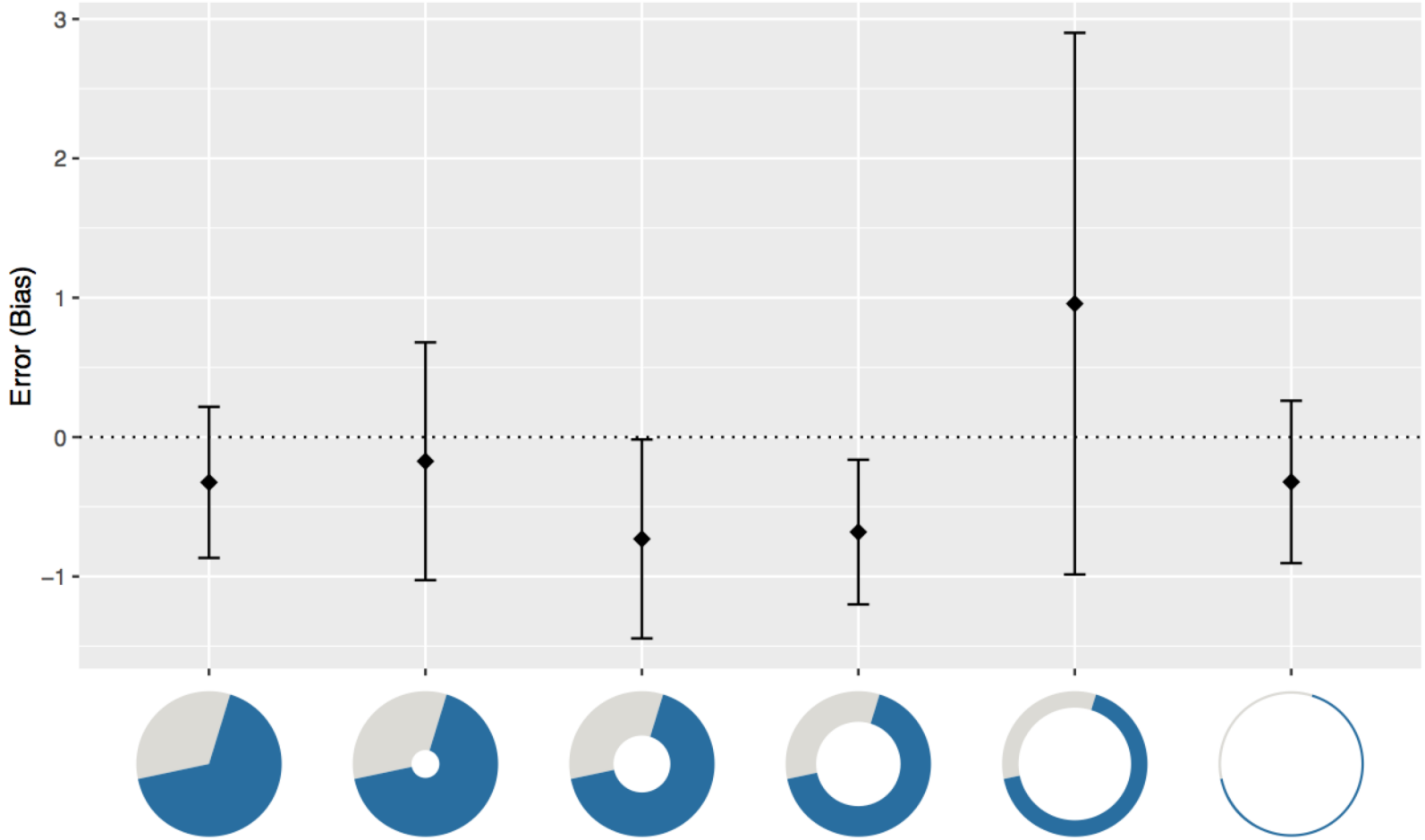
Data			
Rock	Paper	Scissor	TOTAL
2	3	4	9
To calculate percentages			
2/9=22%	3/9=33%	4/9=44%	100%
Degrees for each "pie slice"			
(2/9) x 360 = 80°	(3/9) x 360 = 120°	(4/9) x 360 = 160°	360°

https://datavizcatalogue.com/


A
B
C
1
2
3
4
5
25
20
15
10
5
0

PROPORTIONS.
PIE CHARTS

- Donut charts seem to be equally good (cf. [Kosara 2016])

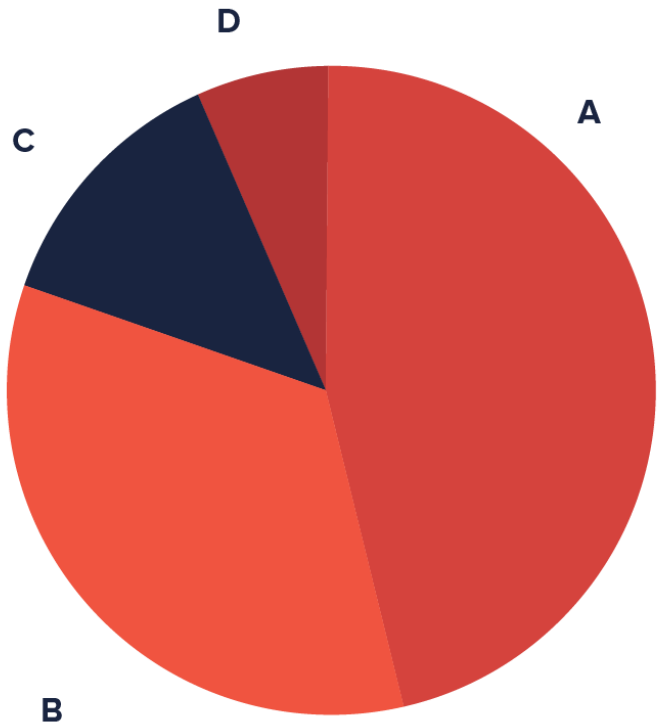


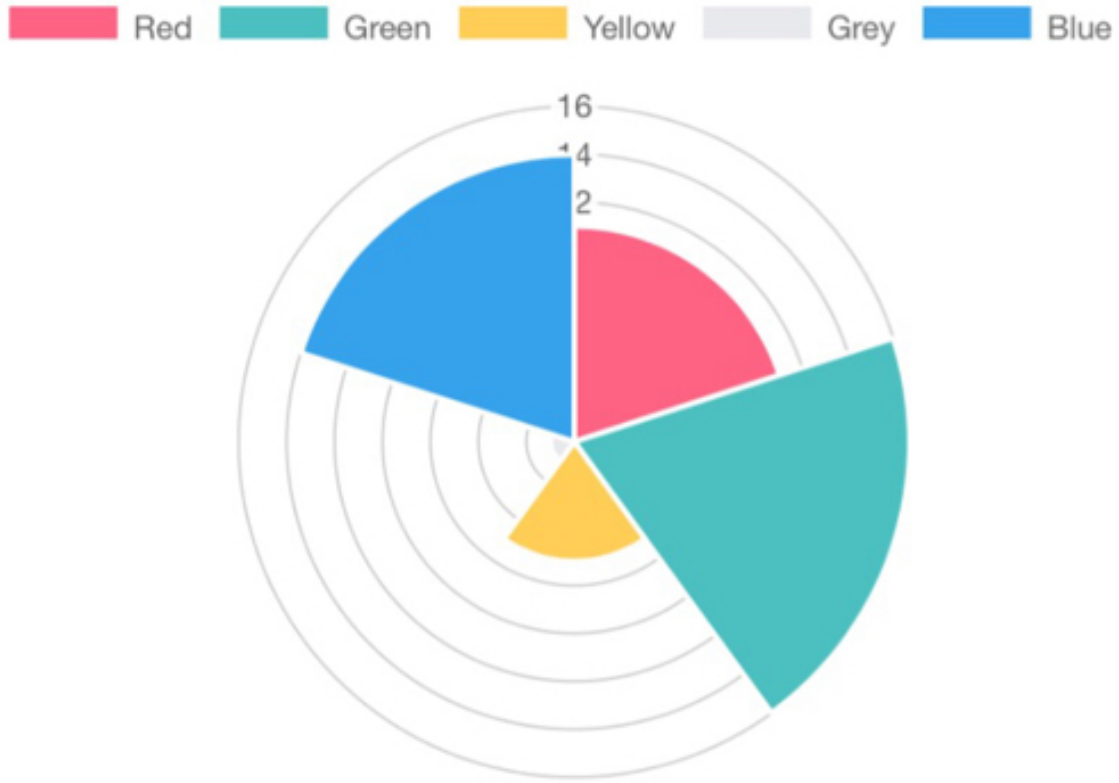
PROPORTIONS.
PIE CHARTS

- Guidelines:
  - They are proportions, should add to 100%
  - Use few categories (max 3-5?)
  - Start at 12
  - Sort from large to small (clockwise)
  - Very similar values will be difficult to appreciate visually
  - Certain key proportions (1/4th, half) may be easier to read
  - It is difficult to get them well
  - **99.85% of the visualization community hates them completely**

PROPORTIONS.
PIE CHARTS

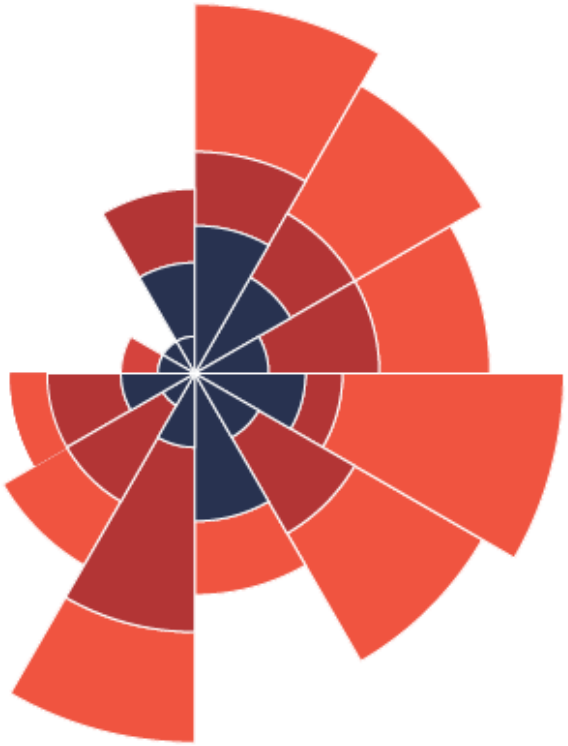
- Different from polar area charts
  - Sectors are equal size, quantities encoded using radius





PROPORTIONS.
PIE CHARTS

- Polar charts can also stack:
  - Same problems (or worse) than with stack bar charts

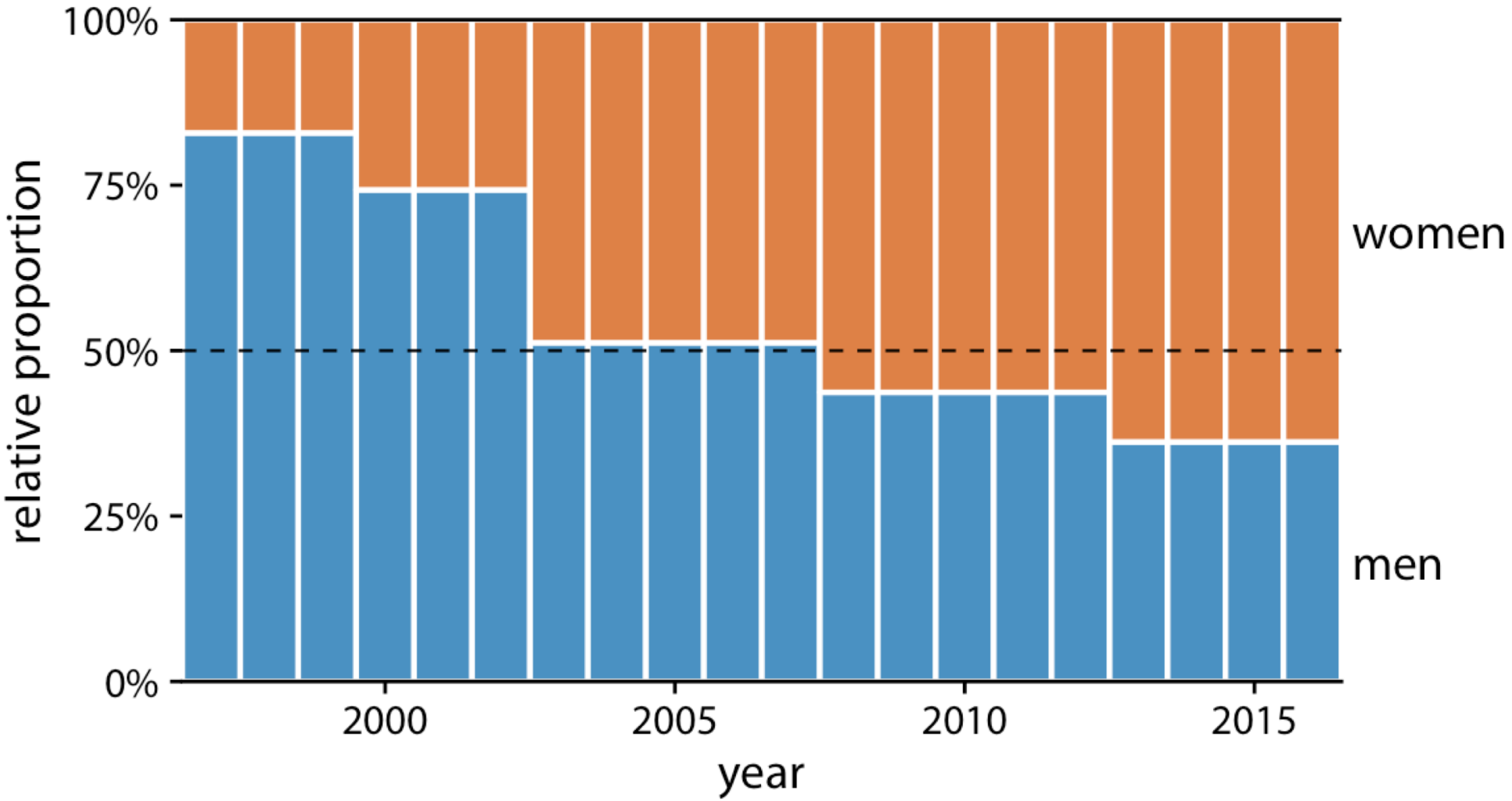


PROPORTIONS.
NORMALIZED STACK BARS

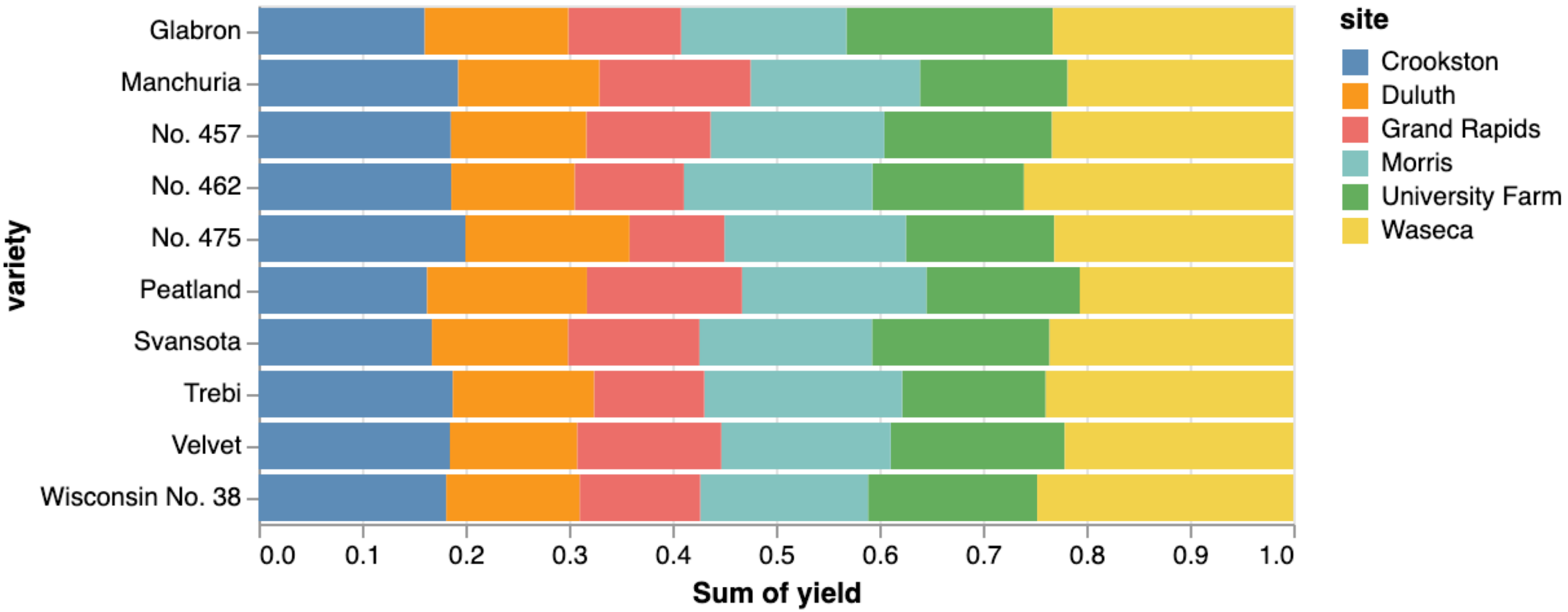
- May work for small amount of stacked categories
  - Need to facilitate comparison
    - Many stacked categories are difficult
  - Two categories align on top and bottom

- Stacked area charts also commonly used
  - Same problems than stacked bars
  - Might be useful if converted to small multiples

PROPORTIONS.
NORMALIZED STACK BARS



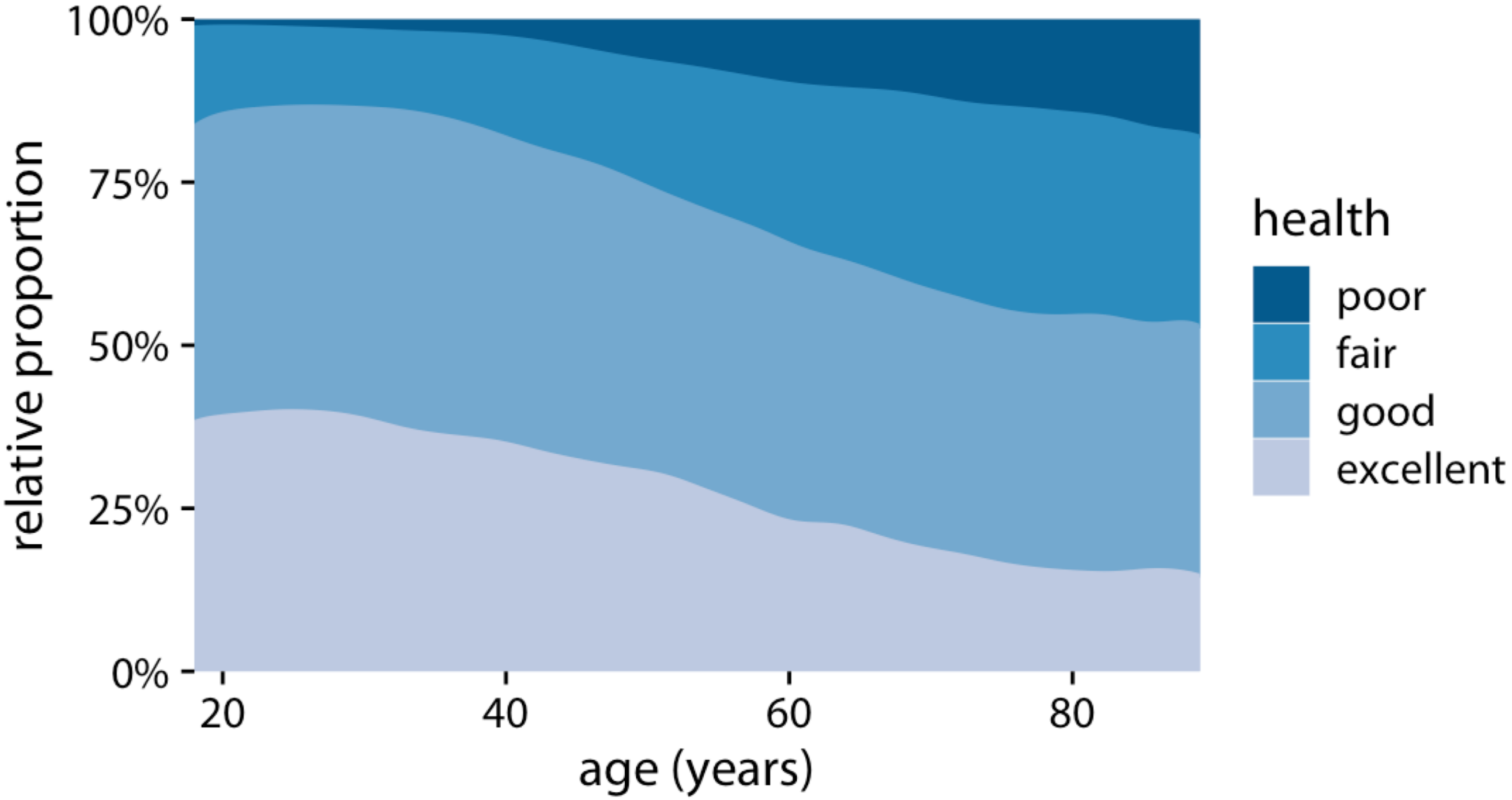
PROPORTIONS.
NORMALIZED STACK BARS



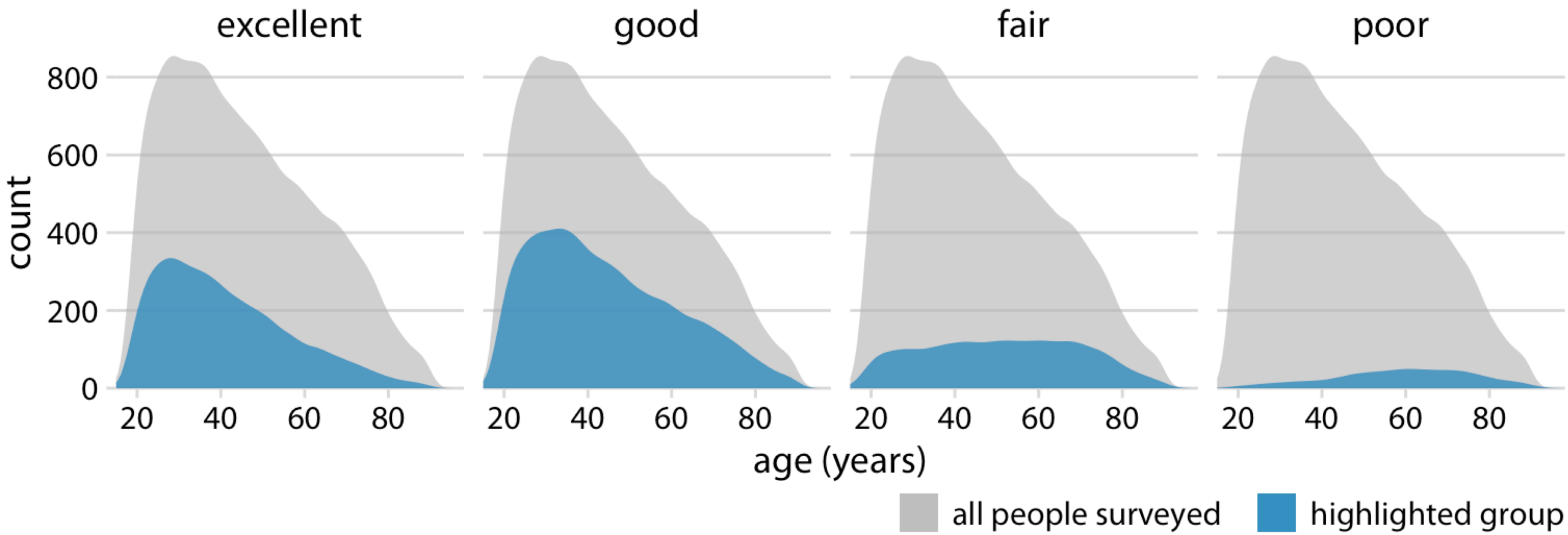
PROPORTIONS.
PIES VS STACKED BARS VS BARS

Purpose	Pie charts	Stacked bar charts	Bar charts
Show data as proportions	Yes	Yes	Difficult
Easy visual comparison of relative proportions	Only if simple	No	Yes
Easy identification of simple fractions	Yes	No	Maybe
Visually appealing even for very small datasets	Yes	No	Yes
Works for many elements	No	No	Yes
Works for many sets of proportions or time series of proportions	No	Only global idea	No

PROPORTIONS.
NORMALIZED STACKED DENSITY PLOTS



PROPORTIONS.
OVERLAID DENSITY PLOTS

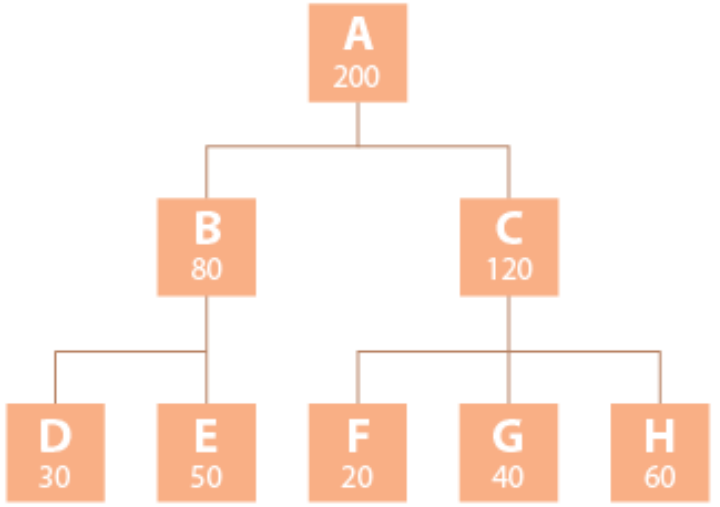


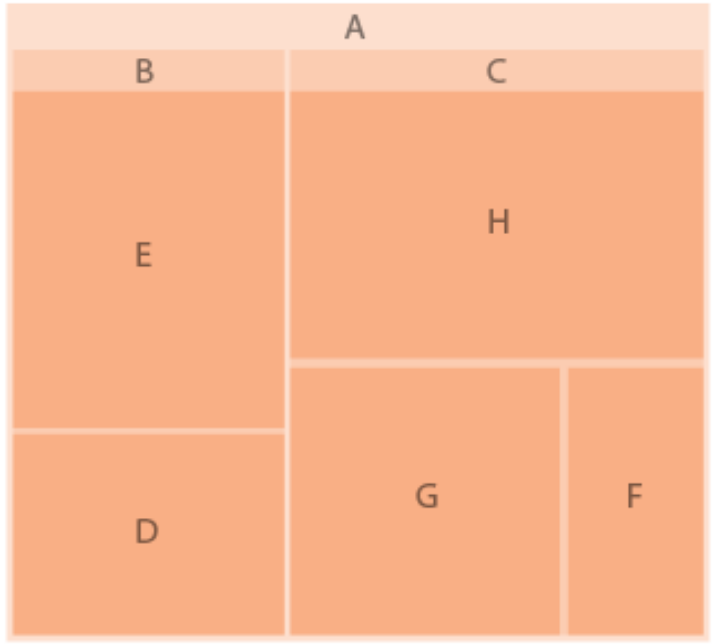
PROPORTIONS. TREEMAPS

- AKA enclosure diagram

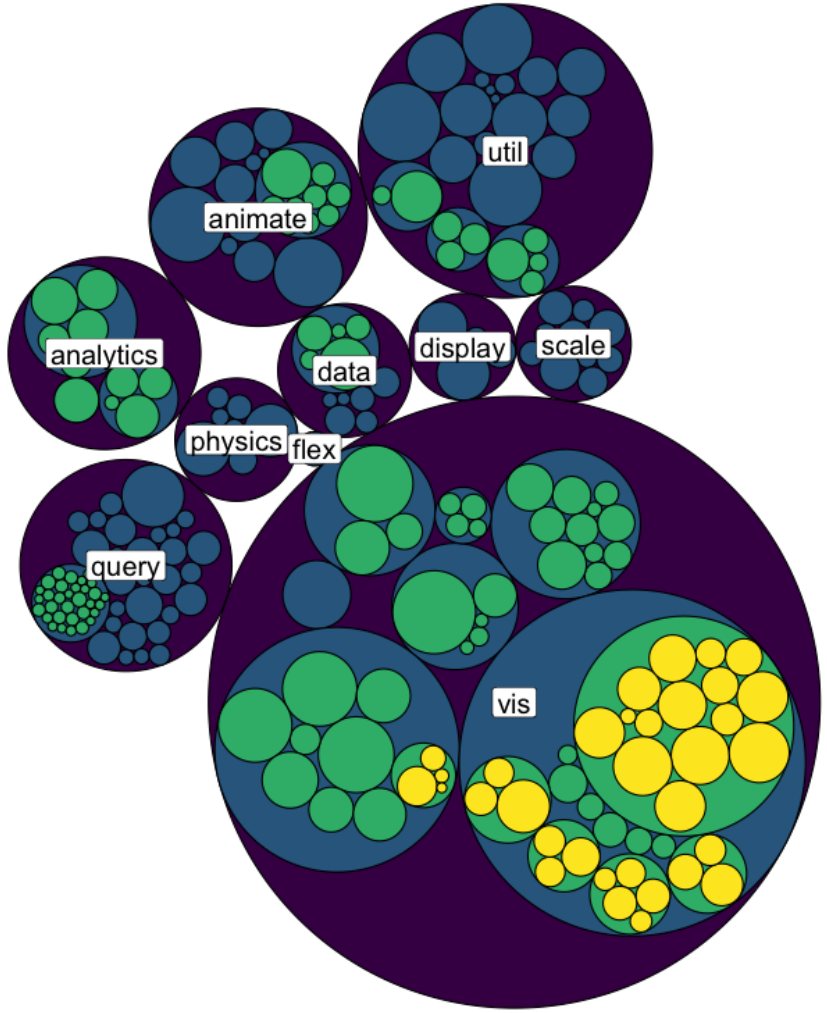
PROPORTIONS. TREEMAPS

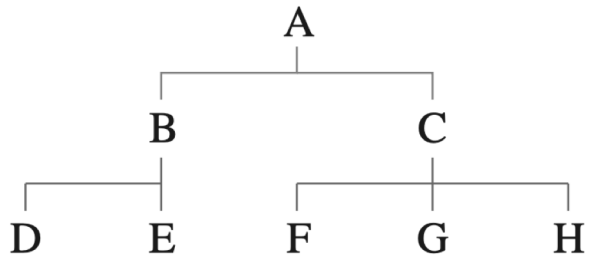
- Construction

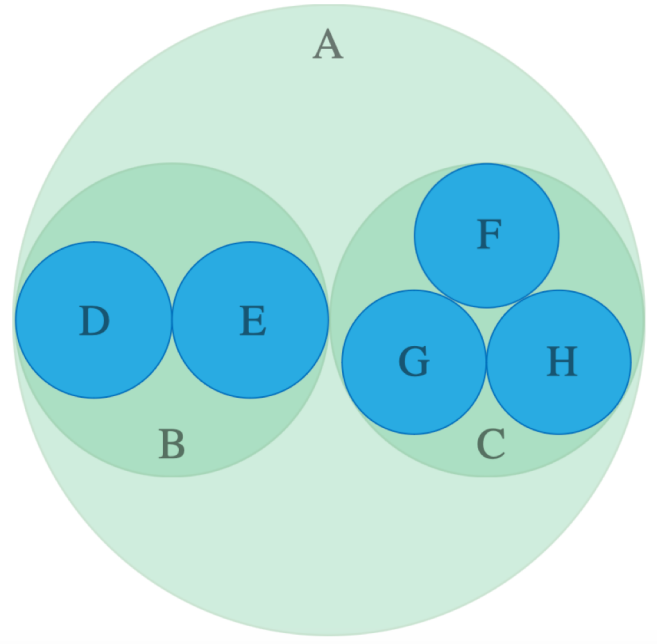




PROPORTIONS.
CIRCLE PACKING





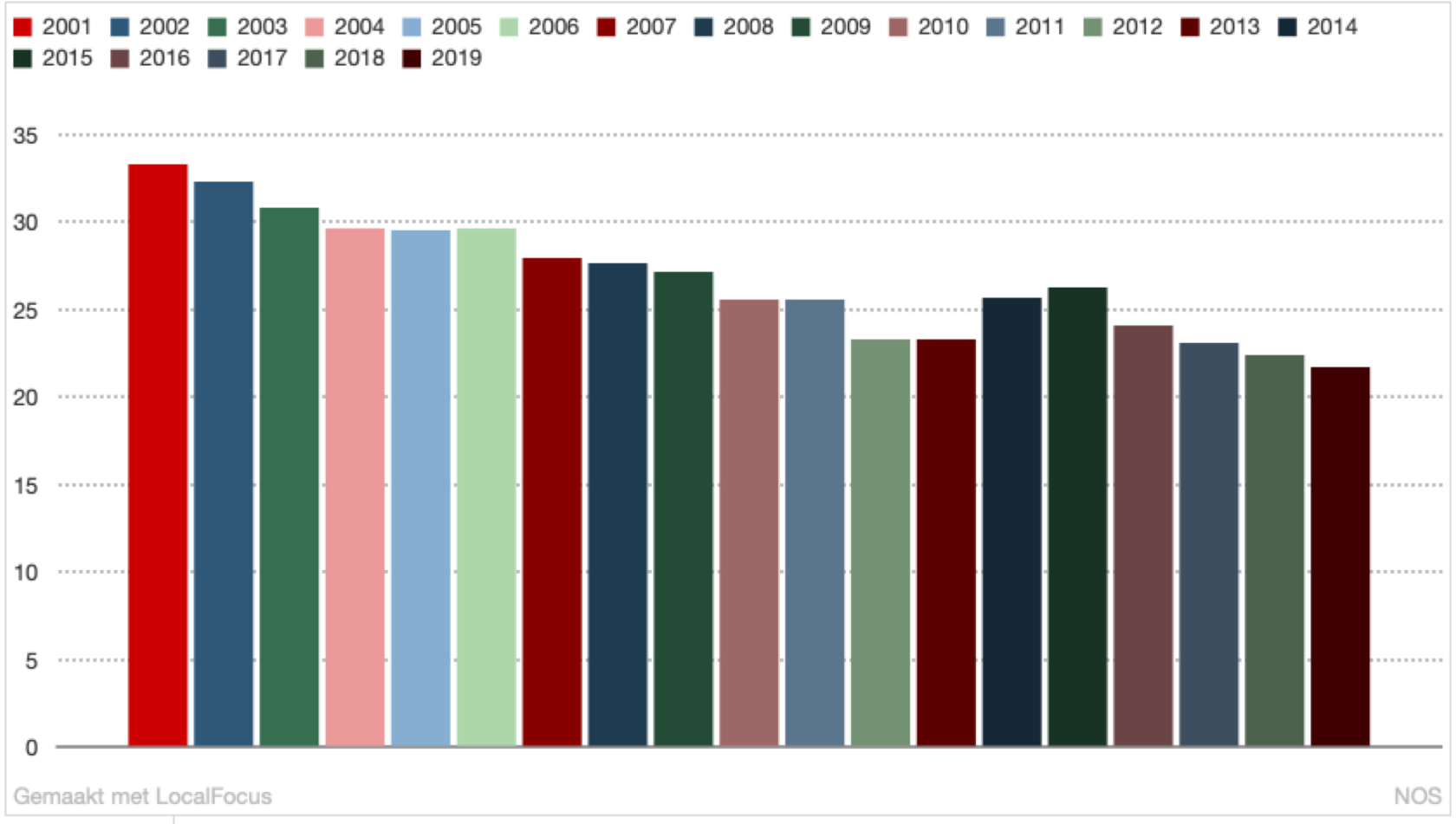


https://datavizcatalogue.com/

EXERCISES

PERE-PAU VÁZQUEZ – VIRVIG GROUP – UPC

VISUALIZATION

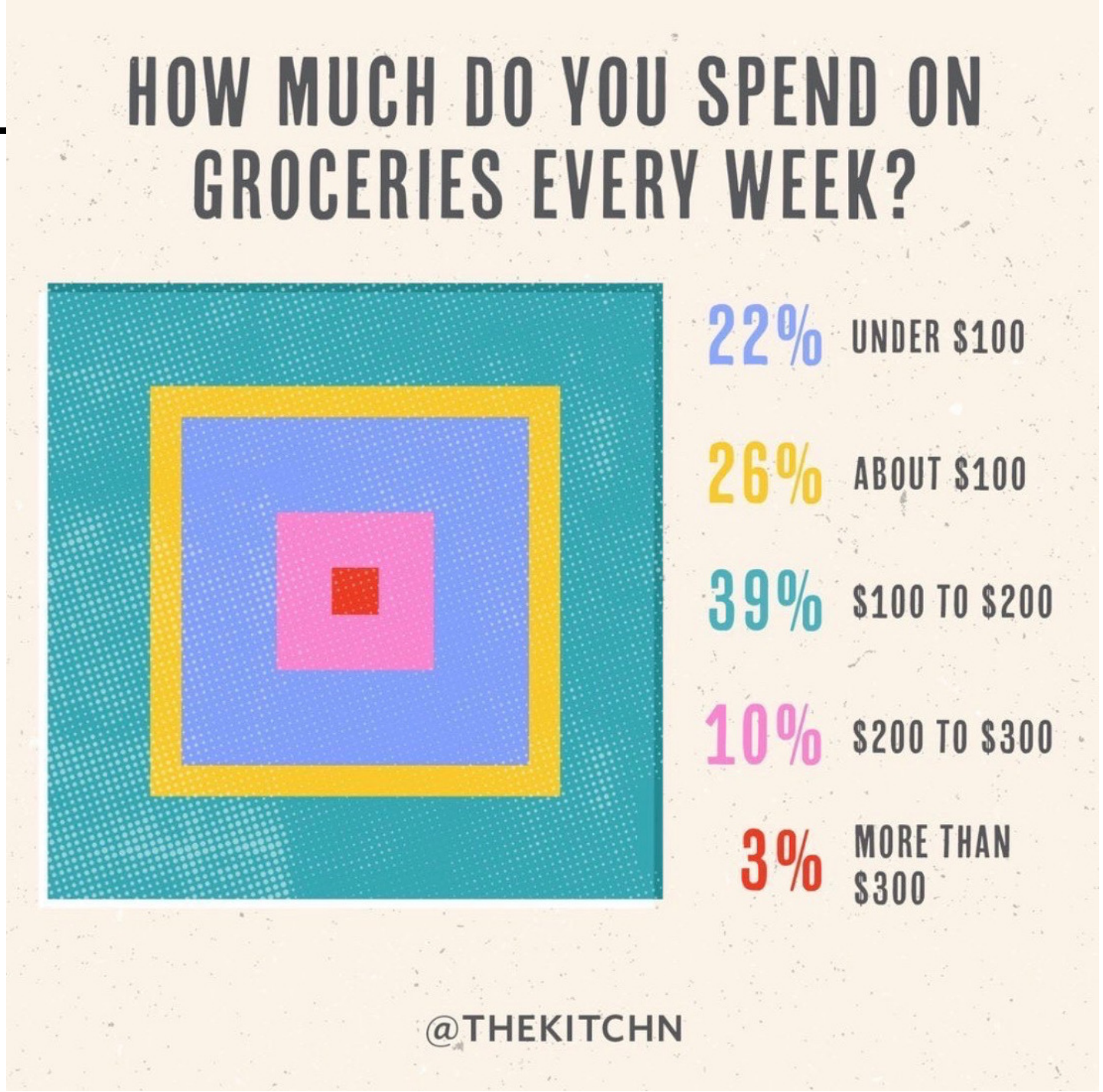


Percentage rokers in Nederland volgens het CBS

EXERCISE


No Facebook Arthur é responsávelpor 40% de toda interação envolvendo candidatos à prefeitura.
40%
34%
14%
No Instagram a vantagem é ainda maior, onde Arthur do Val acumula 44%.
44%
36%
11%
A FORÇA DE ARTHUR DO VAL NAS REDES
★★ PREFEITO ★★
ARTHUR 51
VICE ADELAIDE

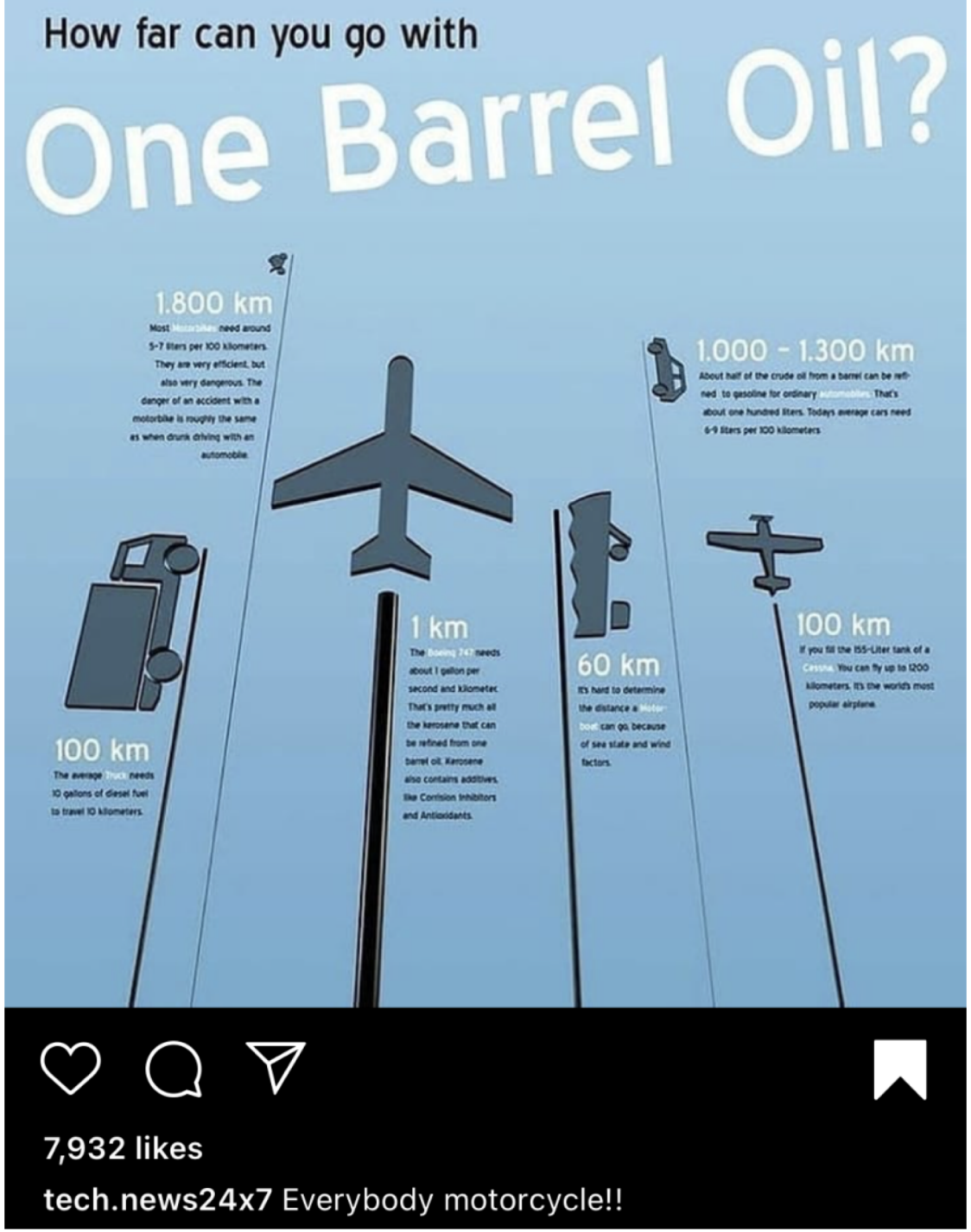
VISUALIZAT


HOW MUCH DO YOU SPEND ON GROCERIES EVERY WEEK?
22% UNDER $100
26% ABOUT $100
39% $100 TO $200
10% $200 TO $300
3% MORE THAN $300
@THEKITCHN

VISUALIZA


PESQUISA IBOPE CONFIRMA:
O POVO DO CABO É 40!
O HOMEM DISPAROU!
37%
33%
21%
CONSIDERANDO OS VOTOS VÁLIDOS
IBOPE inteligência
40
PREFEITO
LULA
CABRAL
VICE EDNA GOMES
FRENTE POPULAR DO CABO | PSB - PP - PDT - PTC - PSL - REPUBLICANOS PROS - AVANTE - PMN | CNPJ: 38.513.901/0001-50
@lulacabral40

VISUALIZATION



EXERCISE

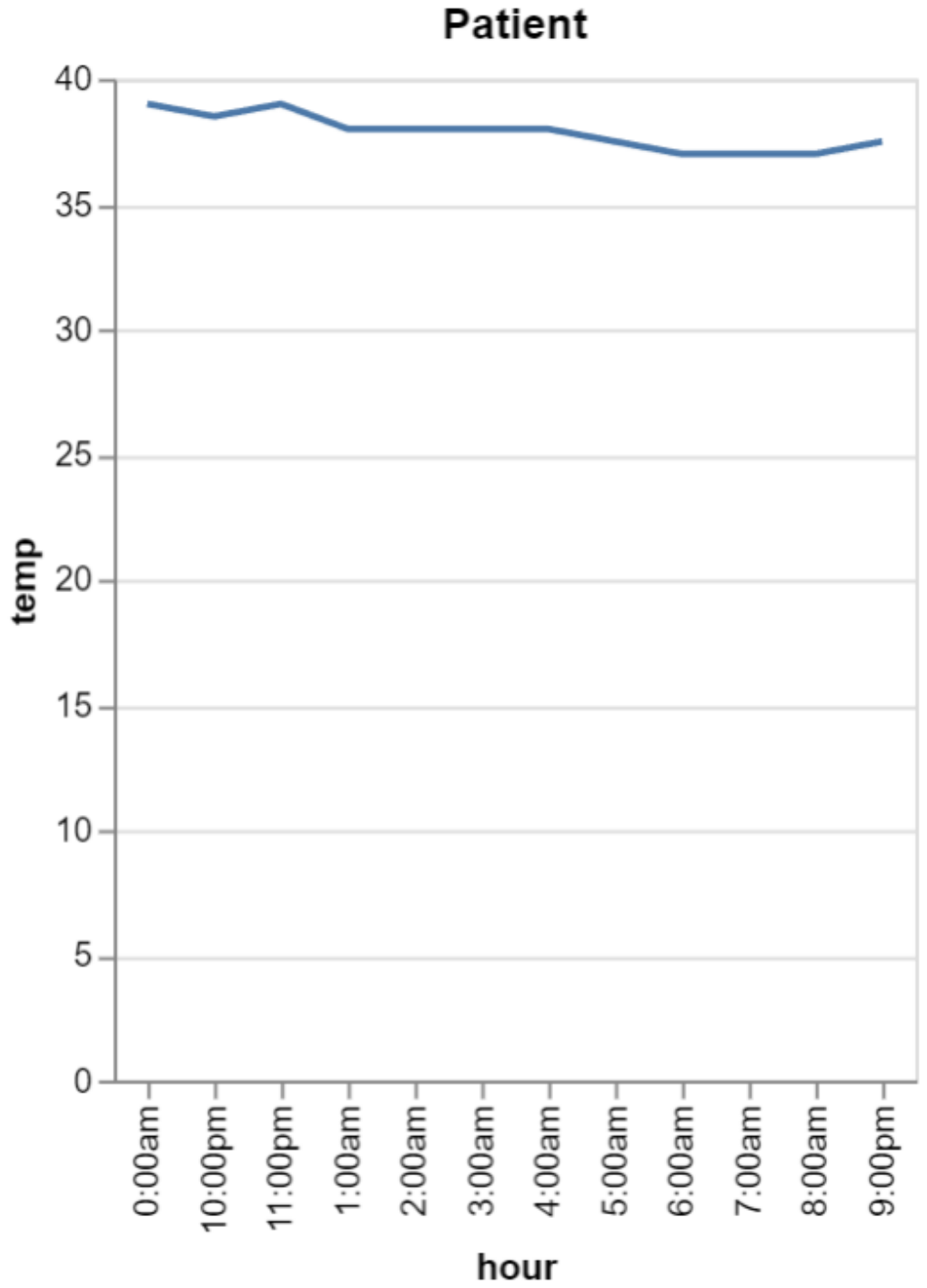
- **Give examples of data sets with the following characteristics:**
  - **with and without an ordering relationship**
  - **with and without a distance metric**
  - **with and without an absolute zero**

EXERCISE

- **We want to monitor the temperature of a patient along the night. One GECD student suggests to use a bar chart. Another suggests to use a line chart. Who is right? Why?**

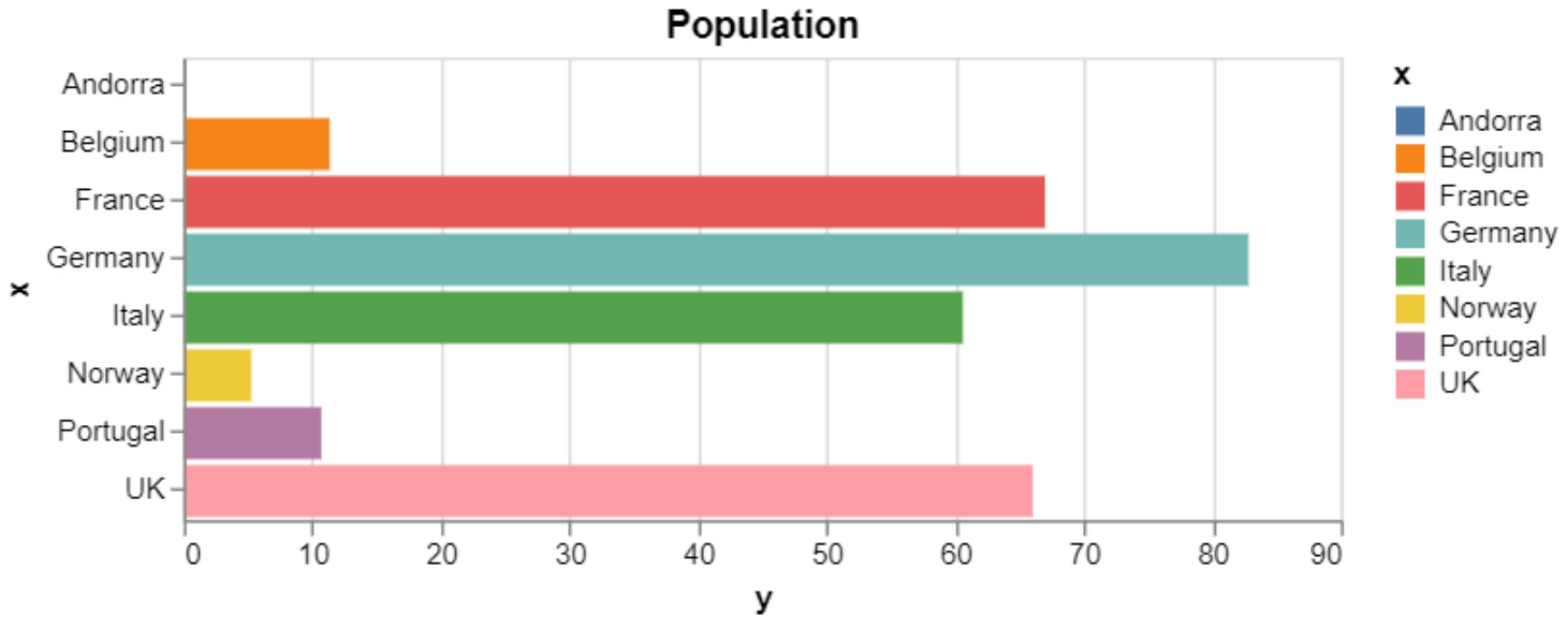
EXERCISE

- **The GECD student that was right created this chart, is it correct? Why?**



EXERCISE

- Do you think this chart is right? If not, what things should be corrected? If it is ok, why?



EXERCISE

- Can you encode multiple keys in a bar chart?

EXERCISES

PERE-PAU VÁZQUEZ – VIRVIG GROUP – UPC

OUTLINE

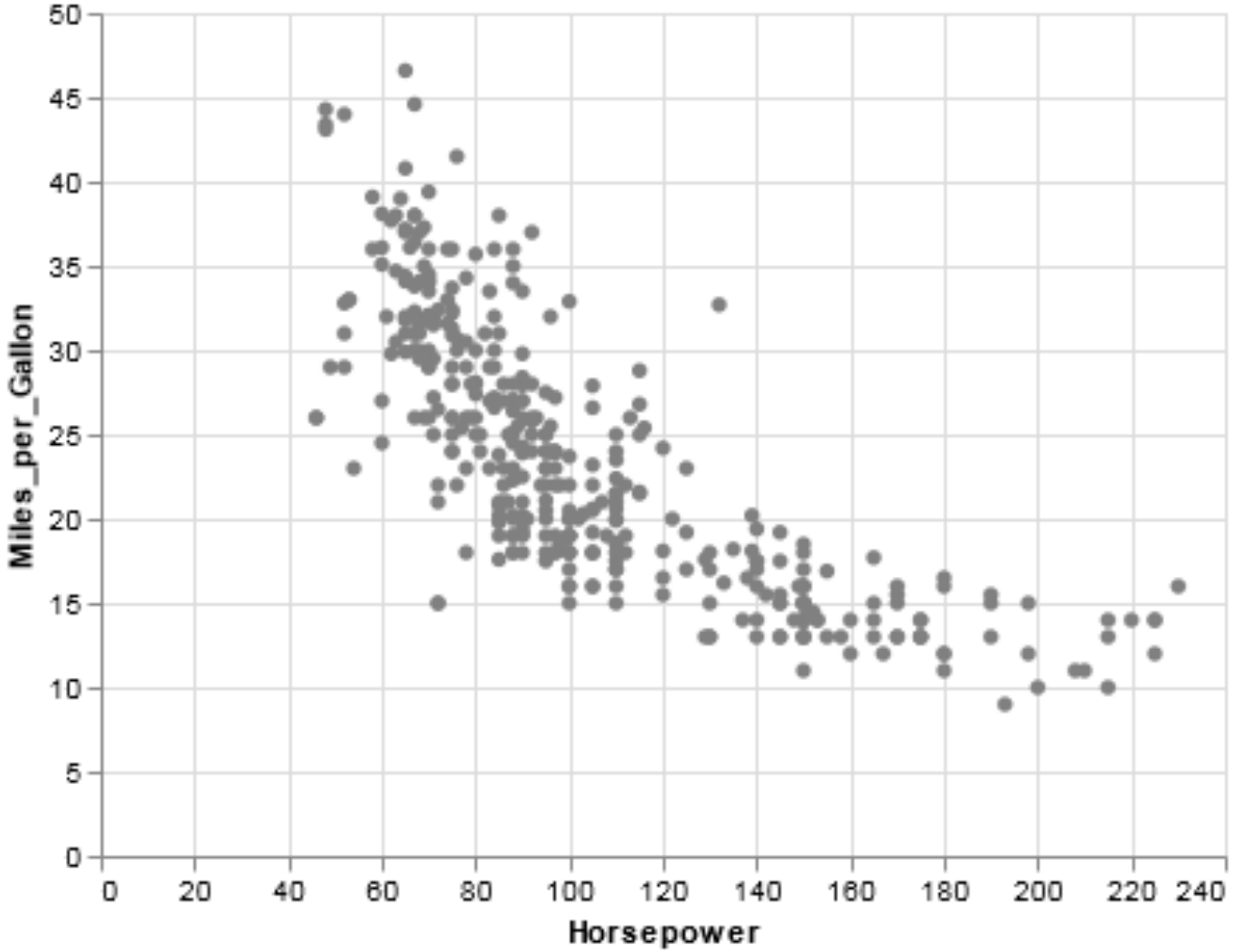
- *Displaying quantities*
- *Displaying distributions*
- *Displaying proportions*
- **Displaying relationships**
- Displaying time series
- Displaying geospatial data
- Other charts
- Uncertainty

RELATIONSHIPS. SCATTERPLOTS

- Two values, no keys
  - Data
    - 2 quantitative values (e.g., horsepower, miles per gallon)
  - Marks: point
    - Using position
  - Channels
    - Position
  - Task
    - Find correlations

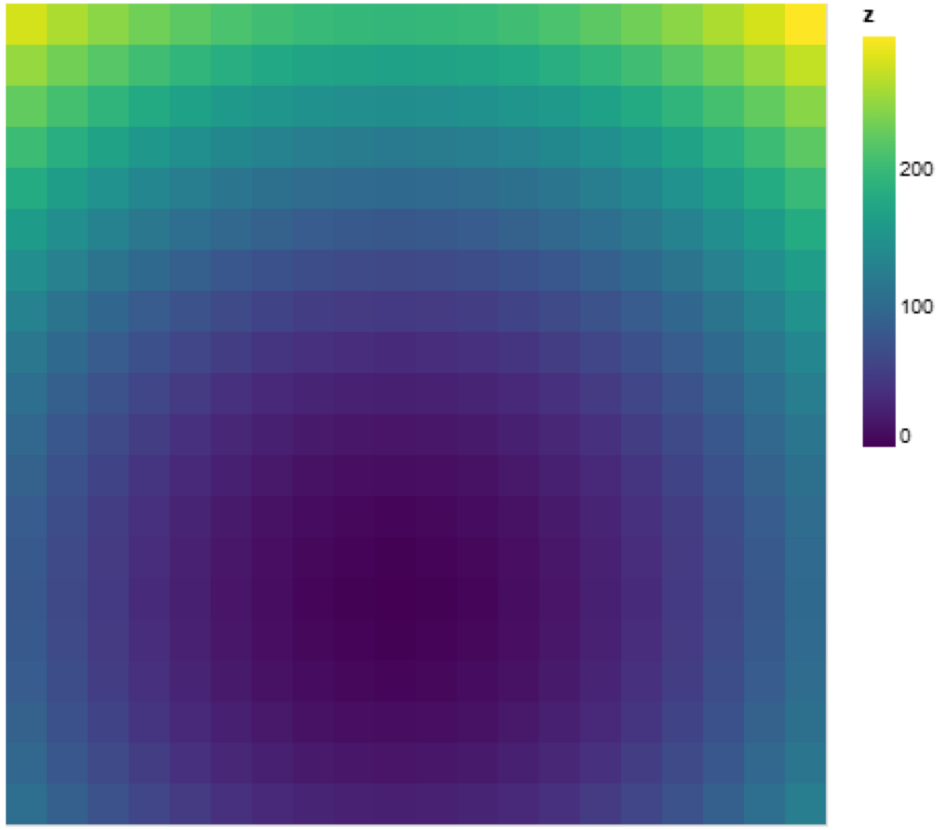
  - Be careful with clutter

RELATIONSHIPS. SCATTERPLOTS



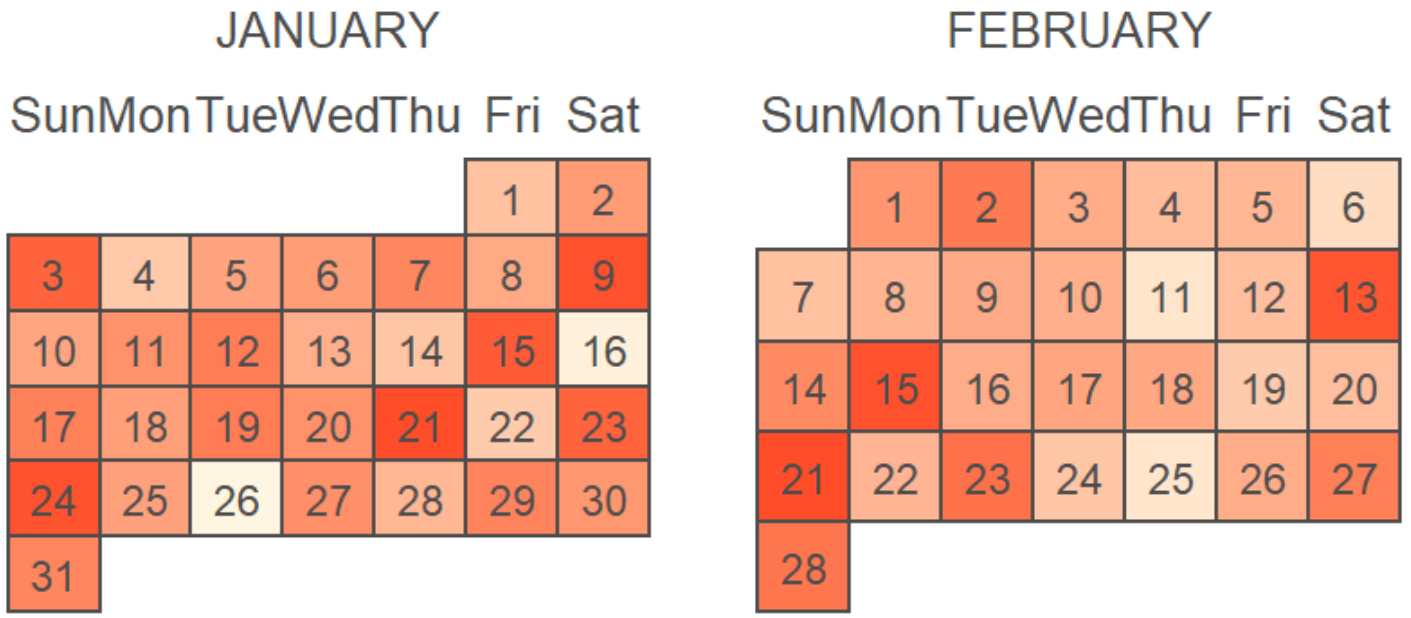
RELATIONSHIPS. HEAT MAPS

- Two keys, one value
  - data
    - 2 categorical attributes
    - 1 quant attribute
  - Marks: area
    - Separate and align in 2D matrix
      - Indexed by 2 categorical attributes
  - Channels
    - Color by quantitative attribute (ordered colormap)
  - Task
    - Find clusters, outliers
  - Scalability
    - 1M items, 100s of category levels, ~10 quant attribute levels

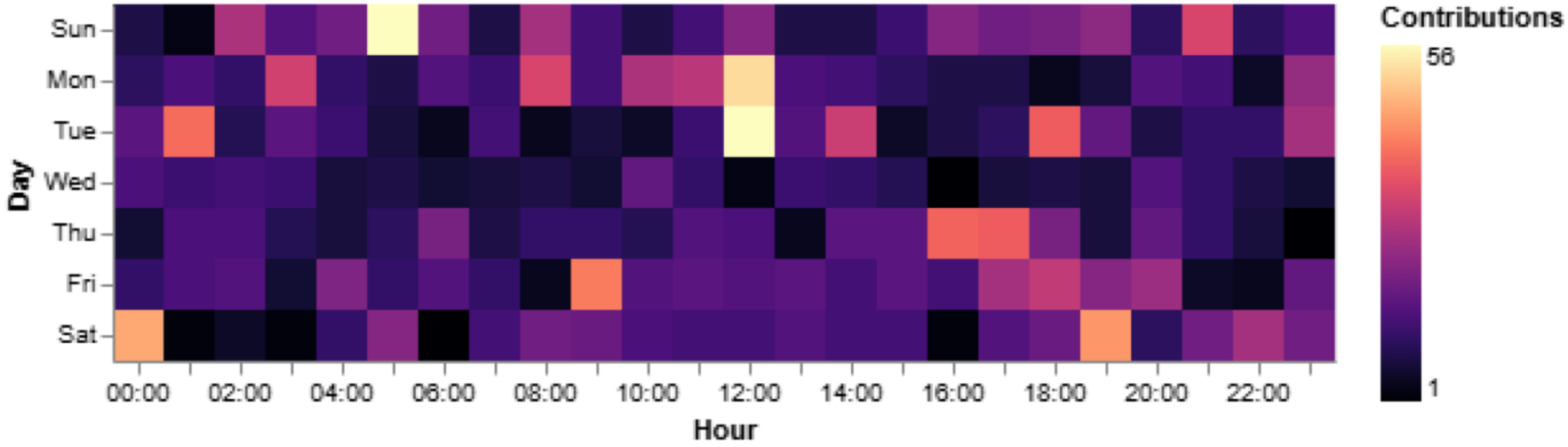


RELATIONSHIPS. HEAT MAPS

- Most general version:
  - AKA reorderable matrix
  - Can swap or reorder any axis
- Other versions:
  - Ordered data
  - Calendar heatmap



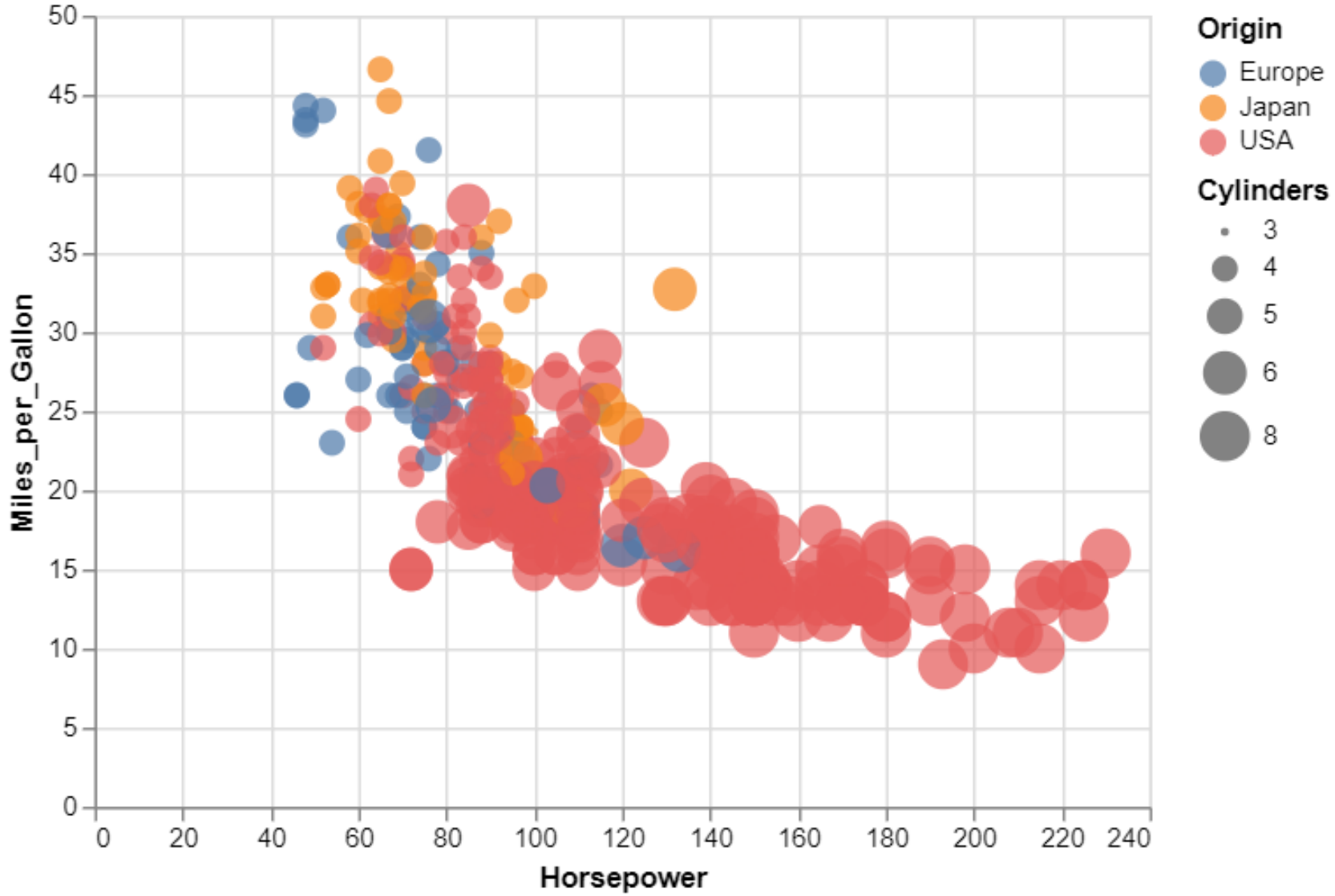
https://r-charts.com/evolution/calendar-heatmap/



RELATIONSHIPS. BUBBLE CHARTS

- Scatterplots with more attributes as other visual channels
  - Size
  - Color
  - …
- May easily generate overlapping and clutter
  - Depends on the distribution
  - May change opacity to alleviate
    - Only a bit

RELATIONSHIPS.
BUBBLE CHARTS



RELATIONSHIPS.
SCATTERPLOT MATRICES (SPLOM)

RELATIONSHIPS.
SCATTERPLOT MATRICES (SPLOM)

- Matrix showing all relations between variables
  - Selection / brushing, drilling down, linked brushing
- Tasks:
  - Find correlations between pairs of variables

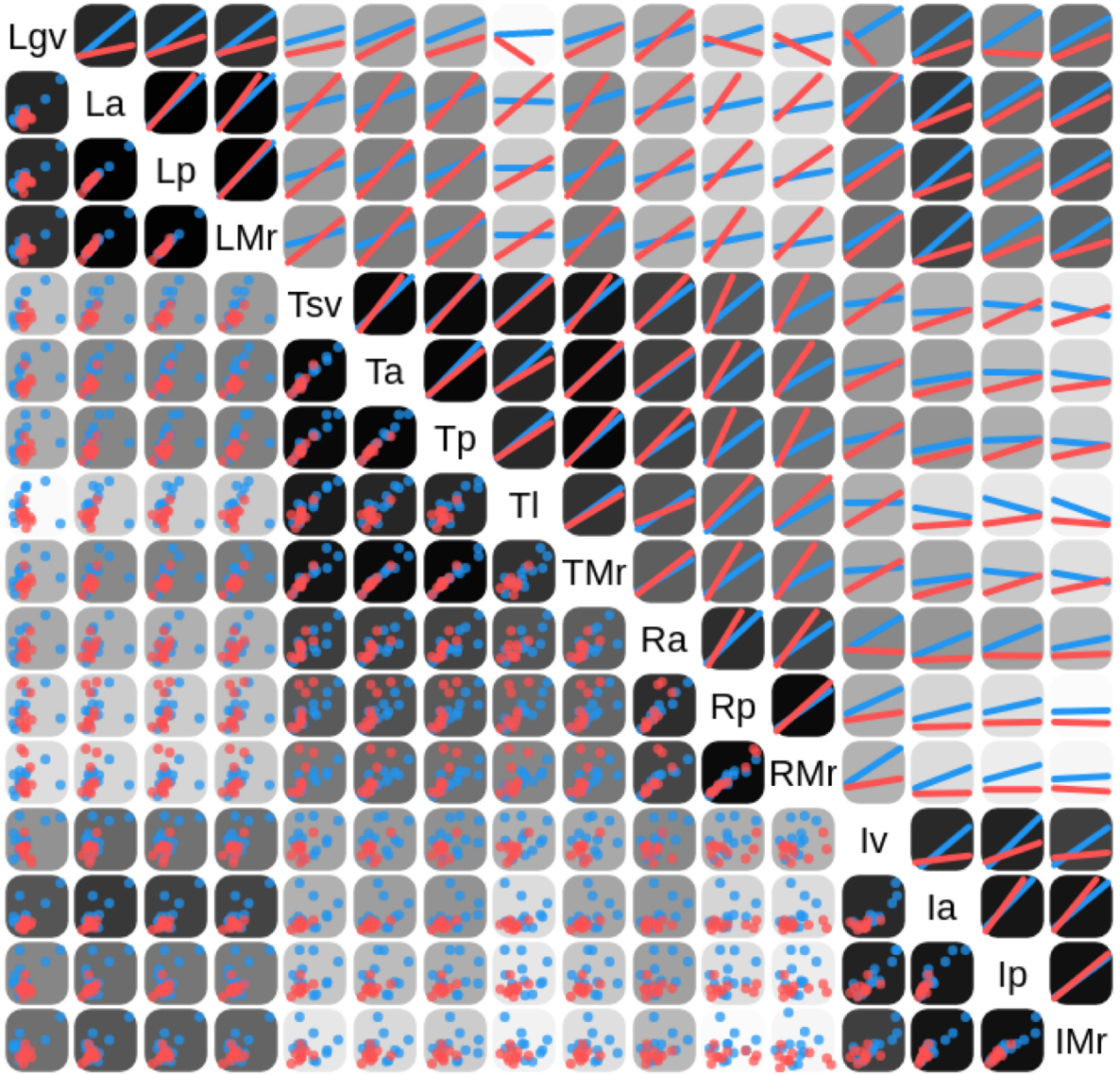
RELATIONSHIPS. SCATTERPLOT MATRICES (SPLOM)

- Scatterplot matrix (SPLOM)
  - Rectilinear axes, point mark
  - All possible pairs of axes
  - Scalability
    - One dozen attributes
    - Dozens to hundreds of items
  - Issues
    -

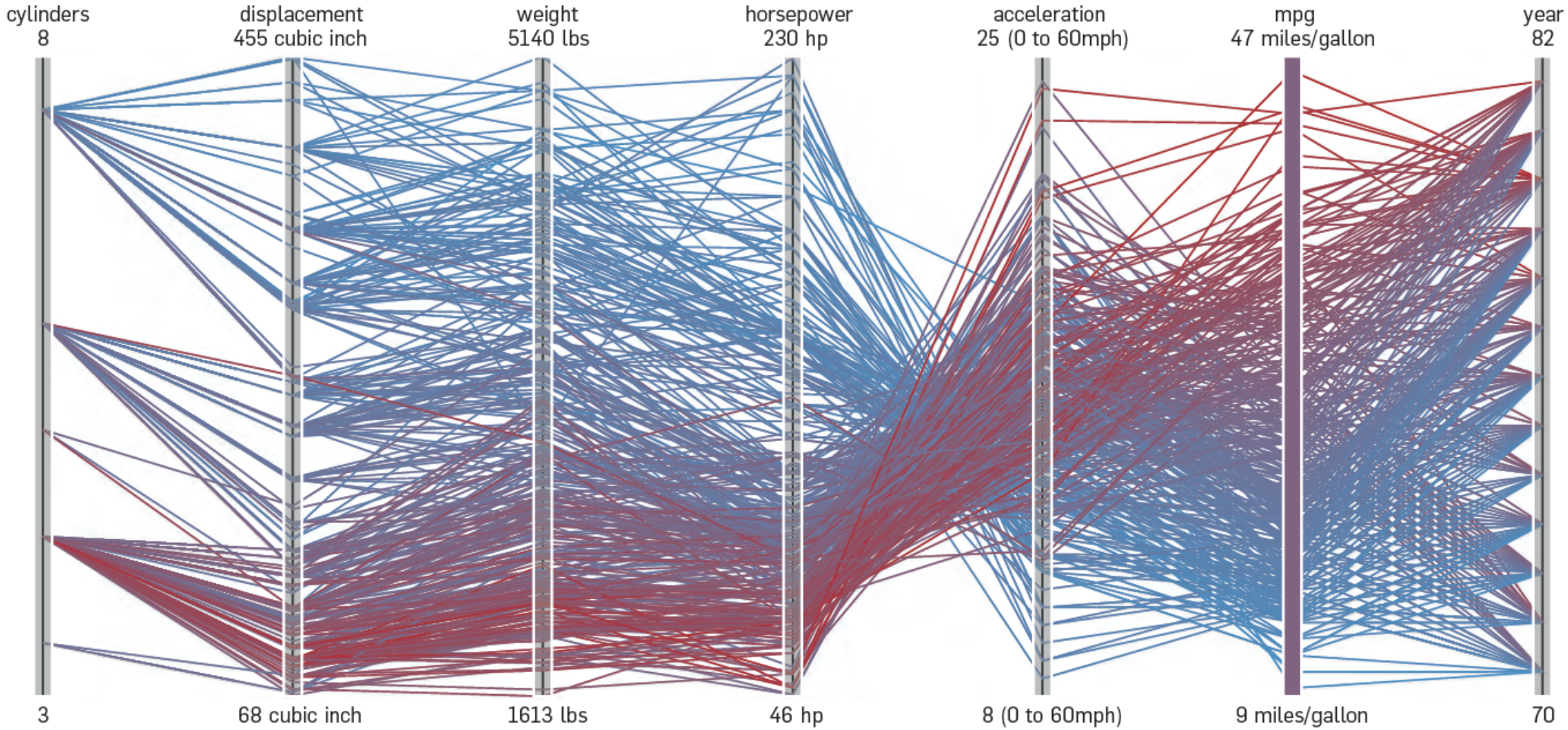
RELATIONSHIPS.
SCATTERPLOT MATRICES (SPLOM)

- Discussion:
  - Space
  - Repeated charts
  - Diagonal

RELATIONSHIPS.
SCATTERPLOT MATRICES (SPLOM)

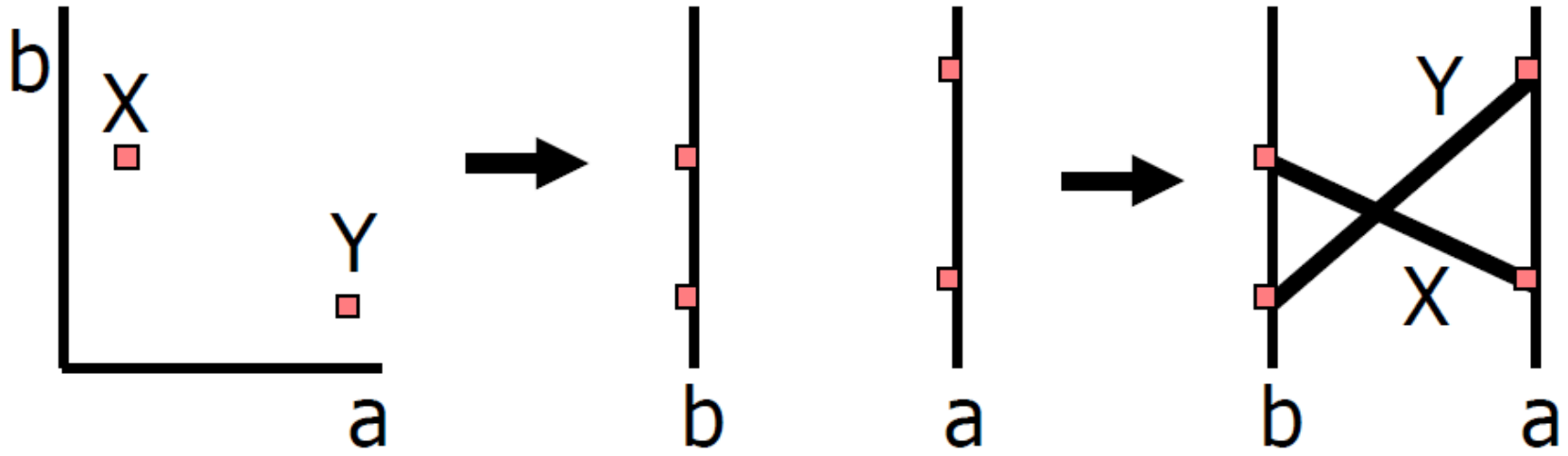


RELATIONSHIPS.
PARALLEL COORDINATE PLOTS



RELATIONSHIPS.
PARALLEL COORDINATE PLOTS

- All axes parallel
- Each sample is a line in this space
- Axes scaled to min/max range of data
- All dimensions can be shown at the same time
  - Positive / negative correlation == parallel / crossing lines



RELATIONSHIPS.
PARALLEL COORDINATE PLOTS

- Data:
  - Several keys, quantitative/categorical values
- Scalability:
  - Up to one dozen of keys
  - Hundreds of items (not thousands)
    - Larger sizes require special techniques, e.g. blending, or hierarchical approaches

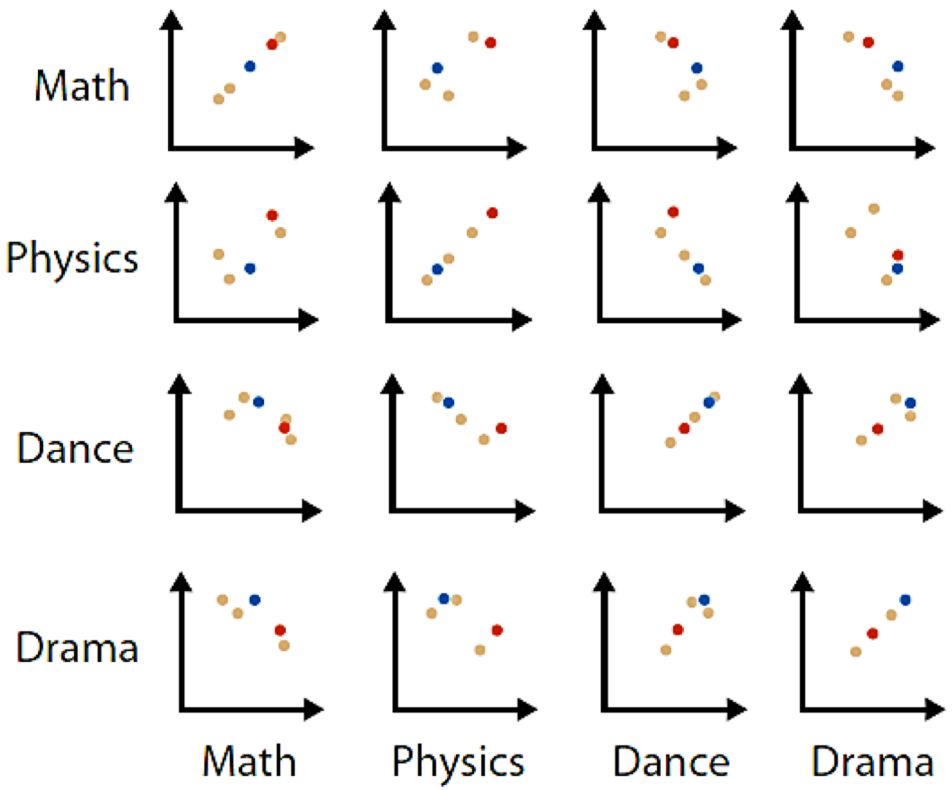
RELATIONSHIPS.
PARALLEL COORDINATE PLOTS

- Discussion. Relatively compact

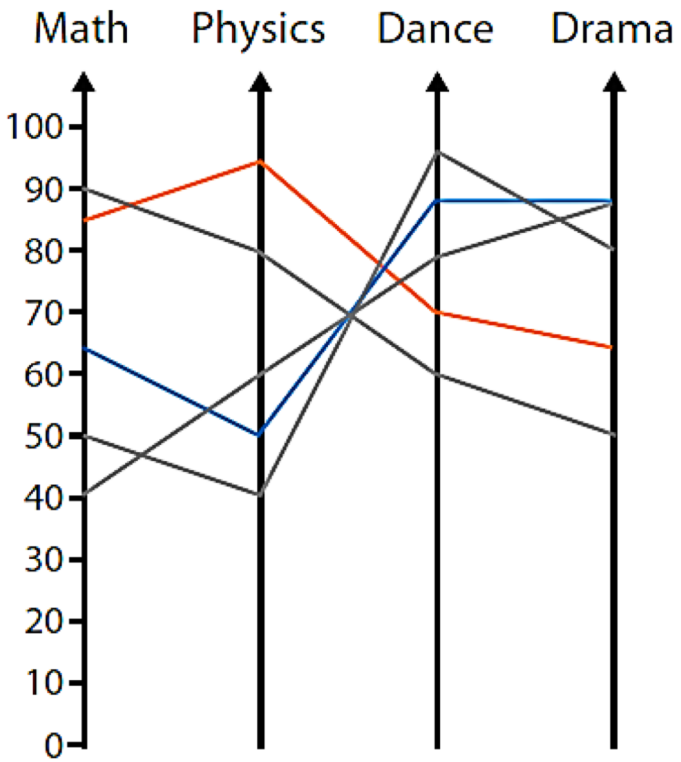
Table

Math	Physics	Dance	Drama
85	95	70	65
90	80	60	50
65	50	90	90
50	40	95	80
40	60	80	90

Scatterplot Matrix



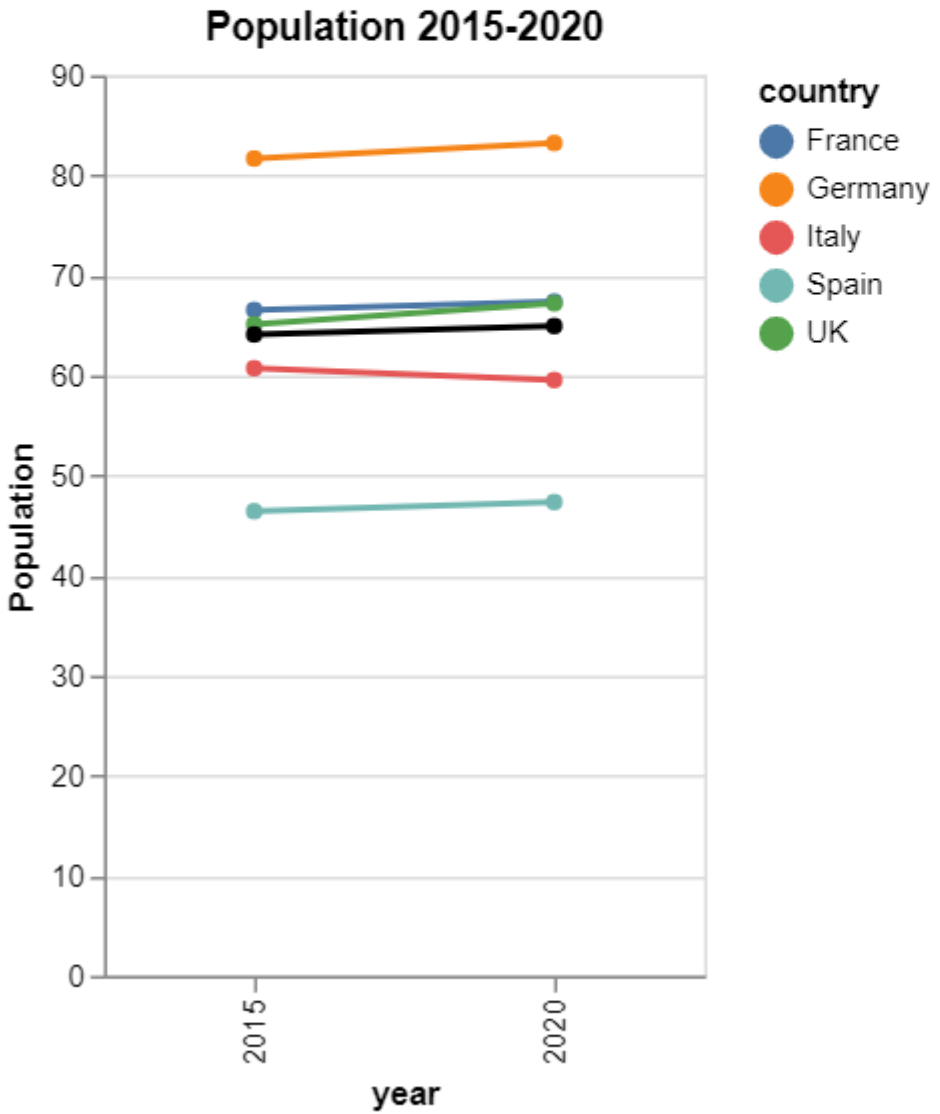
Parallel Coordinates



RELATIONSHIPS. SLOPE CHARTS

- Data:
  - Two values (typically) in two moments of time
- Marks:
  - Points (connected by a line)
- Scalability:
  - Dozens of items?

RELATIONSHIPS.
SLOPE CHARTS



OUTLINE

- *Displaying quantities*
- *Displaying distributions*
- *Displaying proportions*
- *Displaying relationships*
- **Displaying time series**
- Displaying geospatial data
- Other charts
- Uncertainty

REPRESENTATIONS. LINE CHARTS

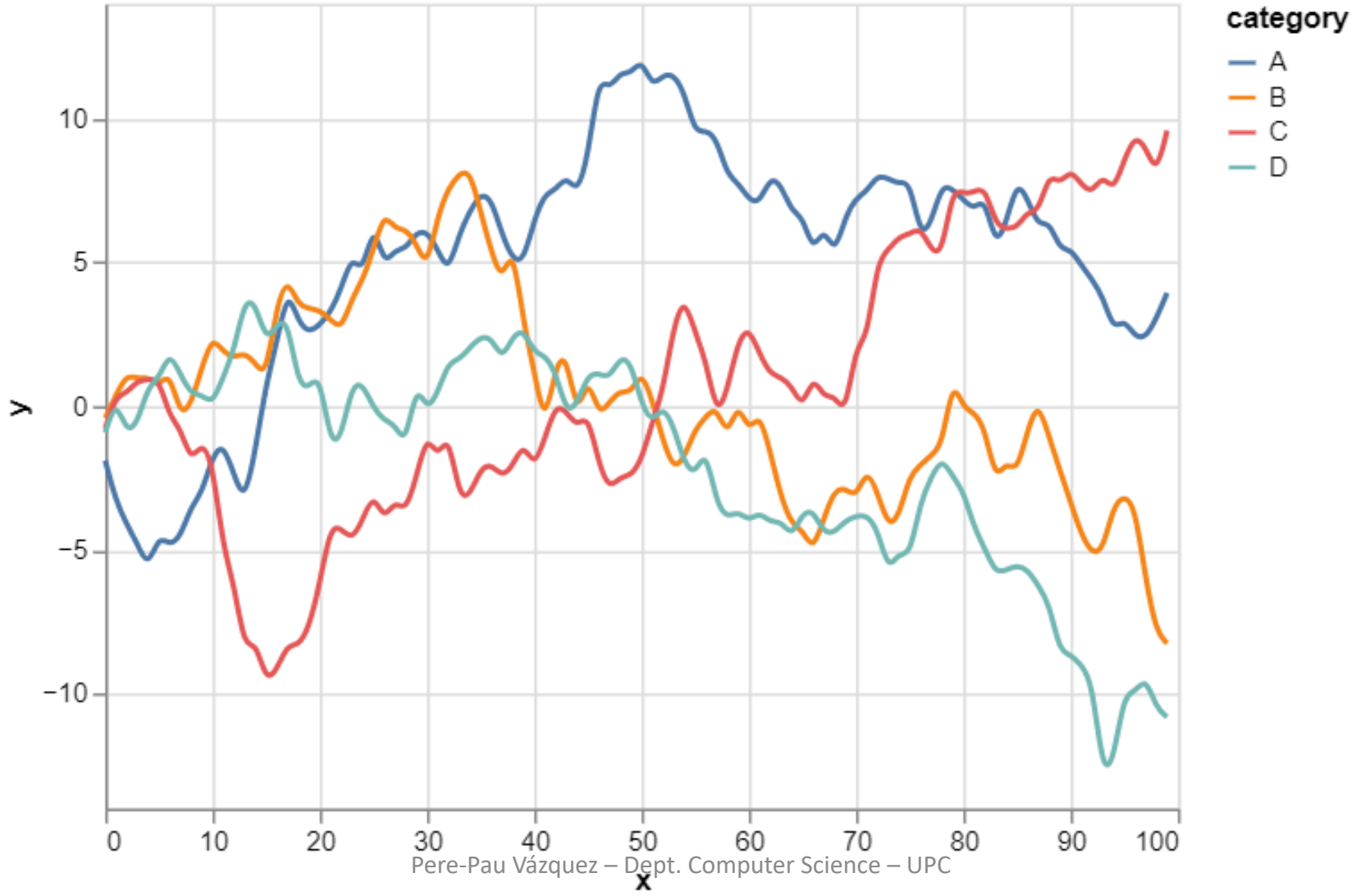
- One key, one value
- Data: 2 quant attributes
- Mark: points
  - Line connection marks between them
- Channels
  - Aligned lengths to express quant value
  - Separated and ordered by key attribute into horizontal regions
- Tasks: Trends
  - Connection marks emphasize ordering of items

TIME SERIES.
LINE CHARTS

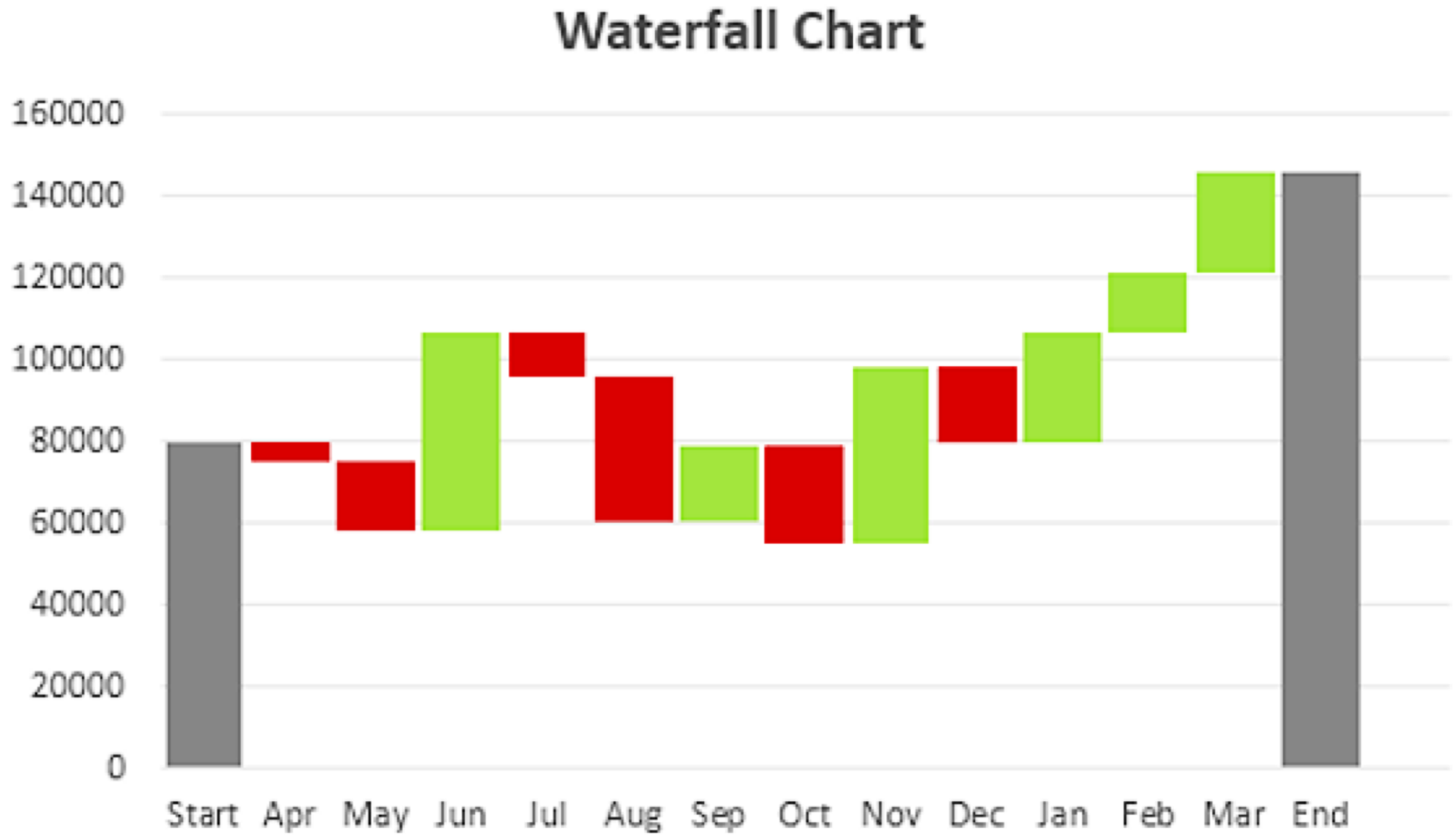
```
np.random.seed(35)
source = pd.DataFrame(np.cumsum(np.random.randn(100, 4), 0).round(2),
                      columns=['A', 'B', 'C', 'D'],
                      index=pd.RangeIndex(100, name='x'))
source = source.reset_index().melt('x', var_name='category', value_name='y')


alt.Chart(source).mark_line(interpolate='basis').encode(
    x='x:Q',
    y='y:Q',
    color='category:N'
)
```

TIME SERIES.
LINE CHARTS



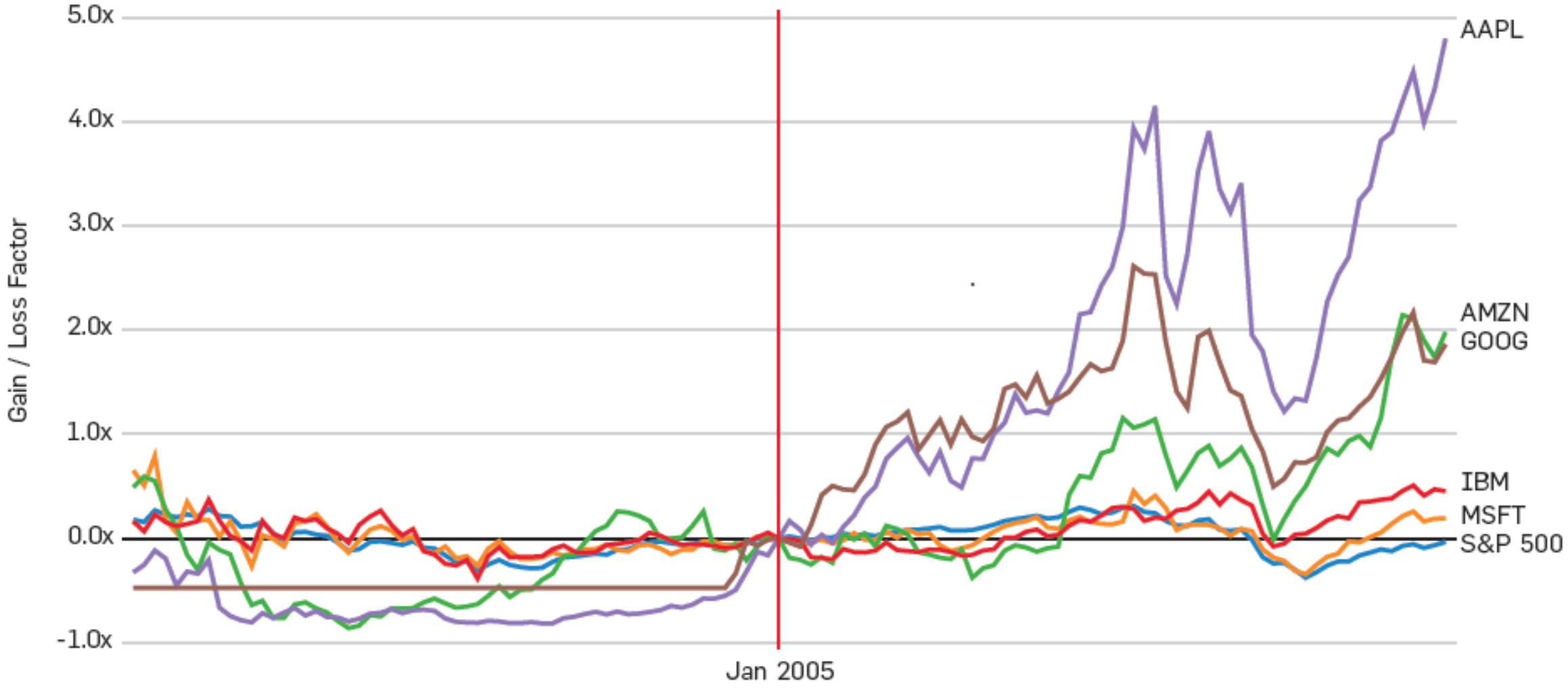
TIME SERIES.
WATERFALL CHART



TIME SERIES. INDEX CHARTS

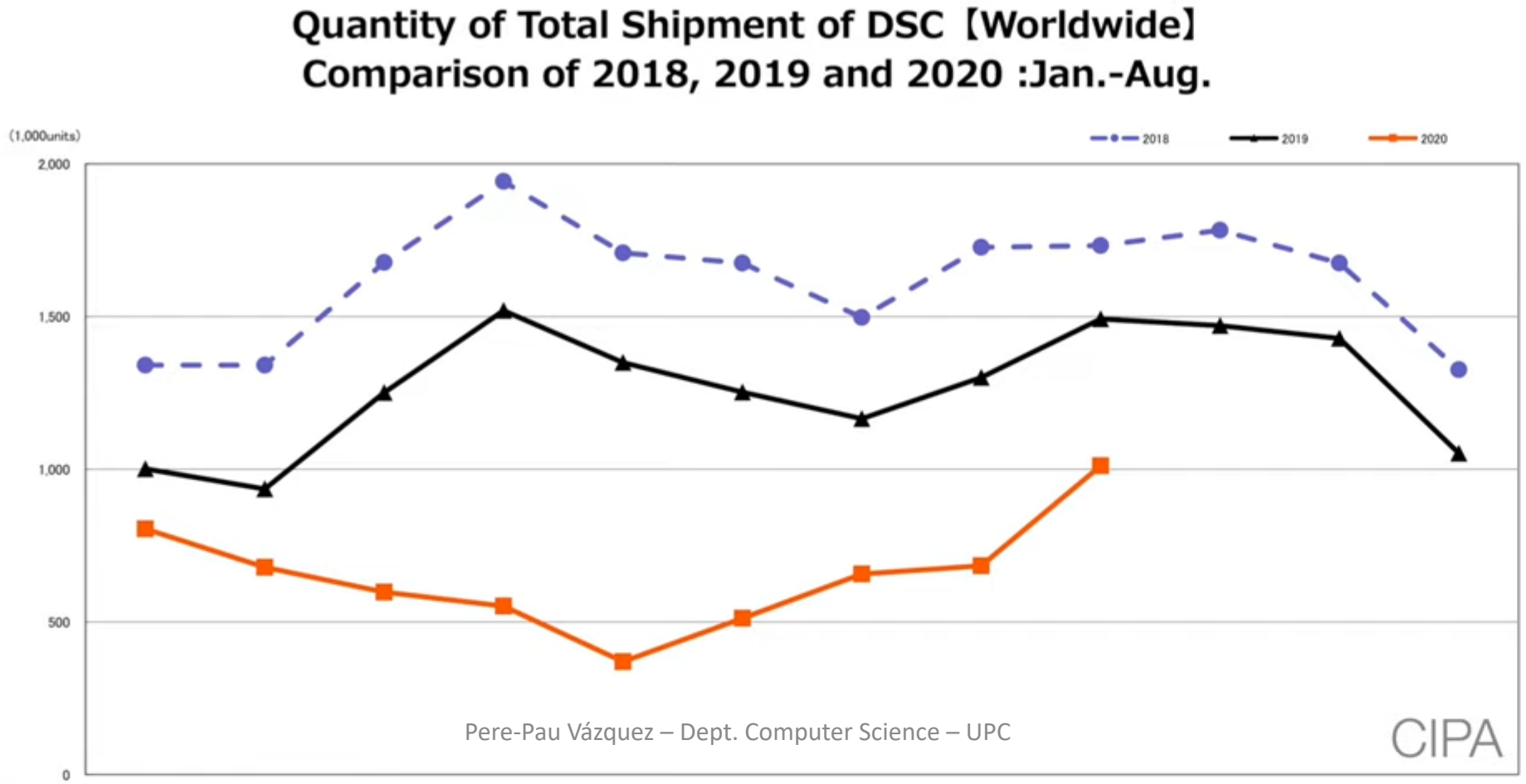
- Shows percentage changes for a collection of time-series data
  - Based on a selected index point
  - Suitable for displaying relative changes (e.g. stock market)
    - May lose context

TIME SERIES.
INDEX CHARTS



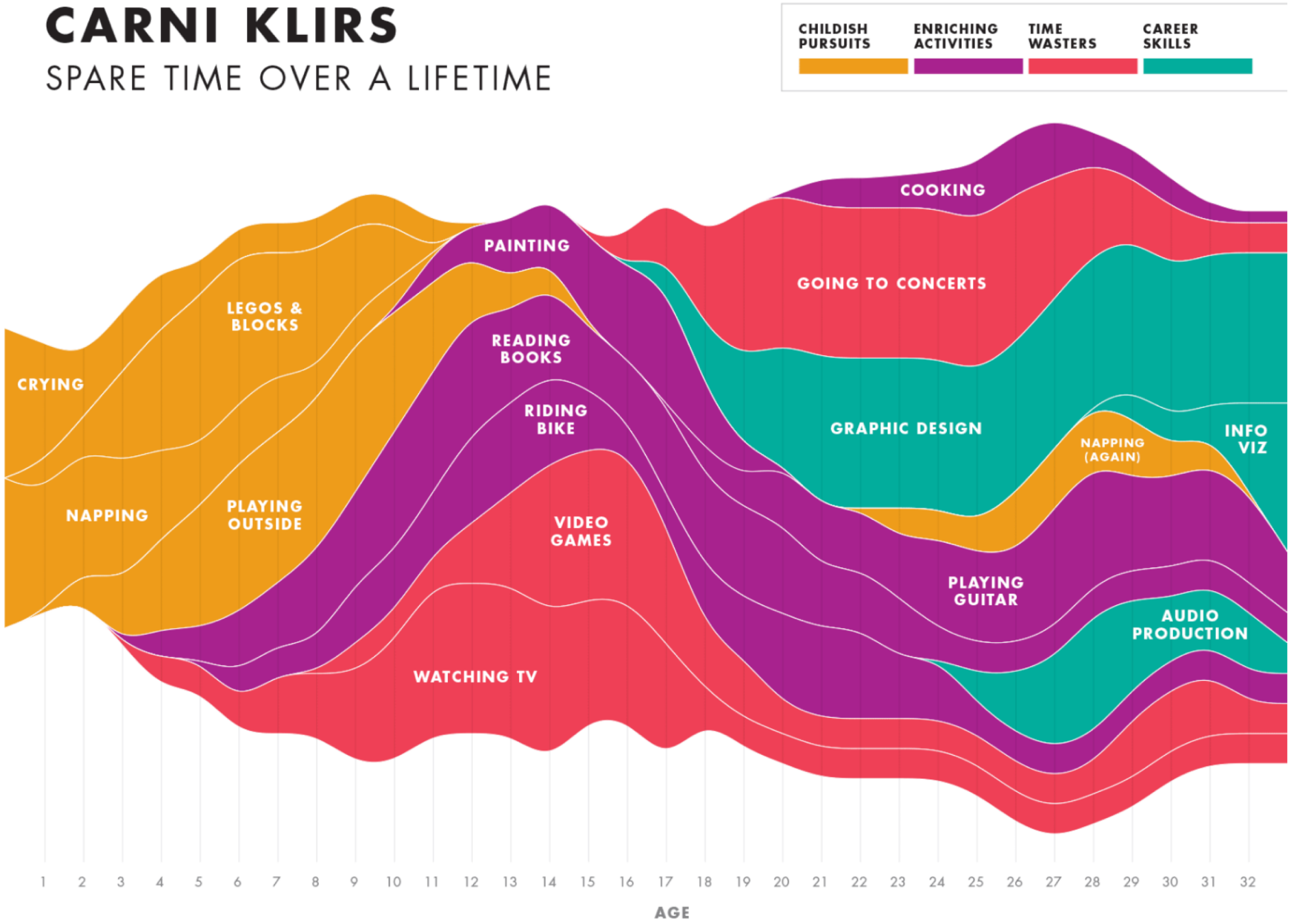
TIME SERIES.
INDEX CHARTS

- Indexing over the time axis



TIME SERIES. STREAMGRAPH

- Generalized stacked graph



https://www.carniklirs.com/project/spare-time-over-a-lifetime

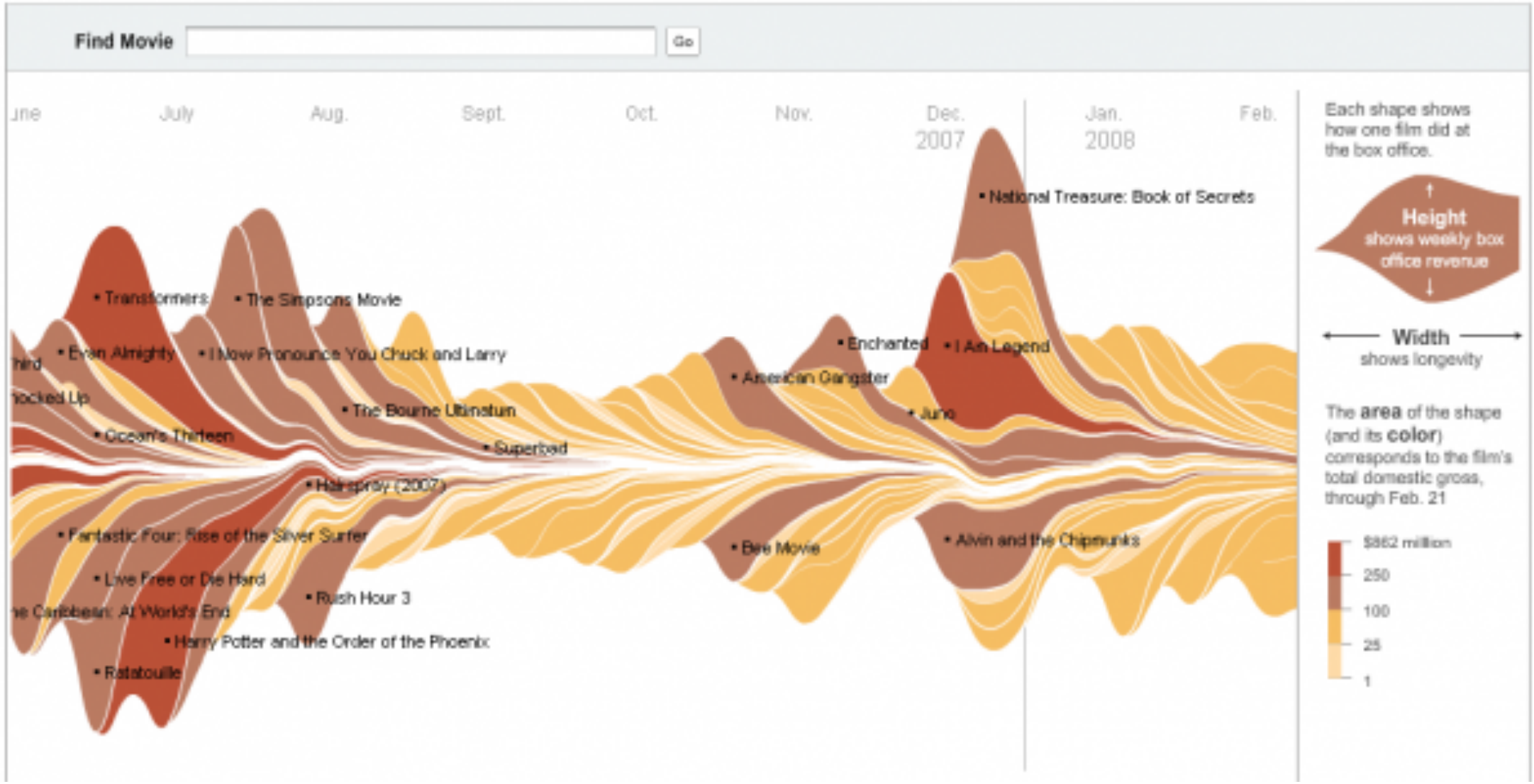
TIME SERIES. STREAMGRAPH

- Streamgraph
  - Emphasizing horizontal continuity vs vertical items
  - Data
    - 1 categ key attrib (e.g. artist)
    - 1 ordered key attrib (e.g. time)
    - 1 quant value attrib (e.g. counts)
  - Derived data
    - geometry: layers, where height encodes counts
    - 1 quant attrib (layer ordering)
  - scalability
    - hundreds of time keys
    - dozens to hundreds of artist keys
    - more than stacked bars, since most layers don’t extend across whole chart

TIME SERIES. STREAMGRAPH

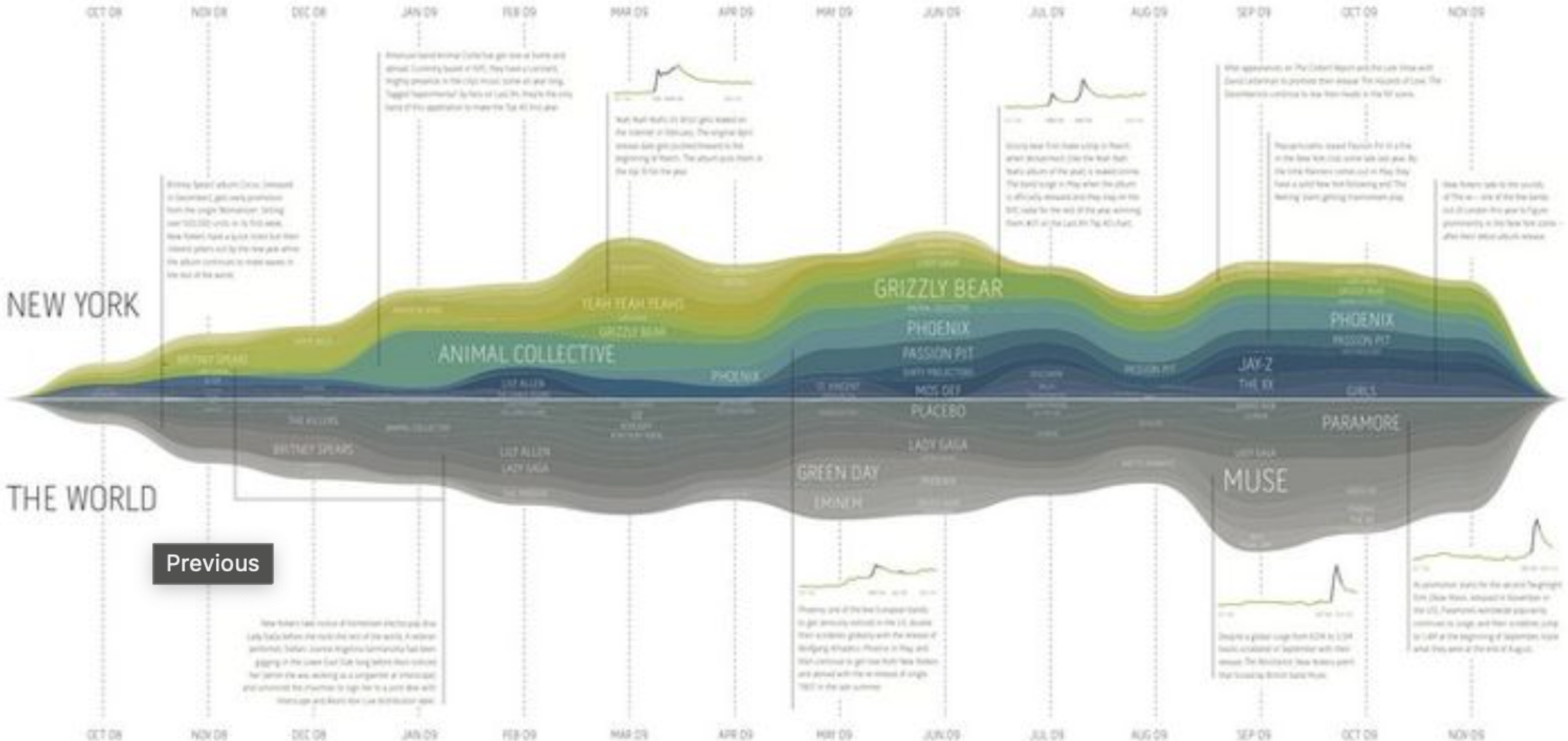
- Discussion:
  - Does not support negative values
  - Does not support data that cannot be added (e.g. temperatures)
  - Trends may be difficult to interpret
    - E.g. at the top of the curves

TIME SERIES.
STREAMGRAPH



NEW YORK VS. THE WORLD

MONTH-BY-MONTH LISTENING TRENDS FOR NEW YORK IN 2009*



Previous

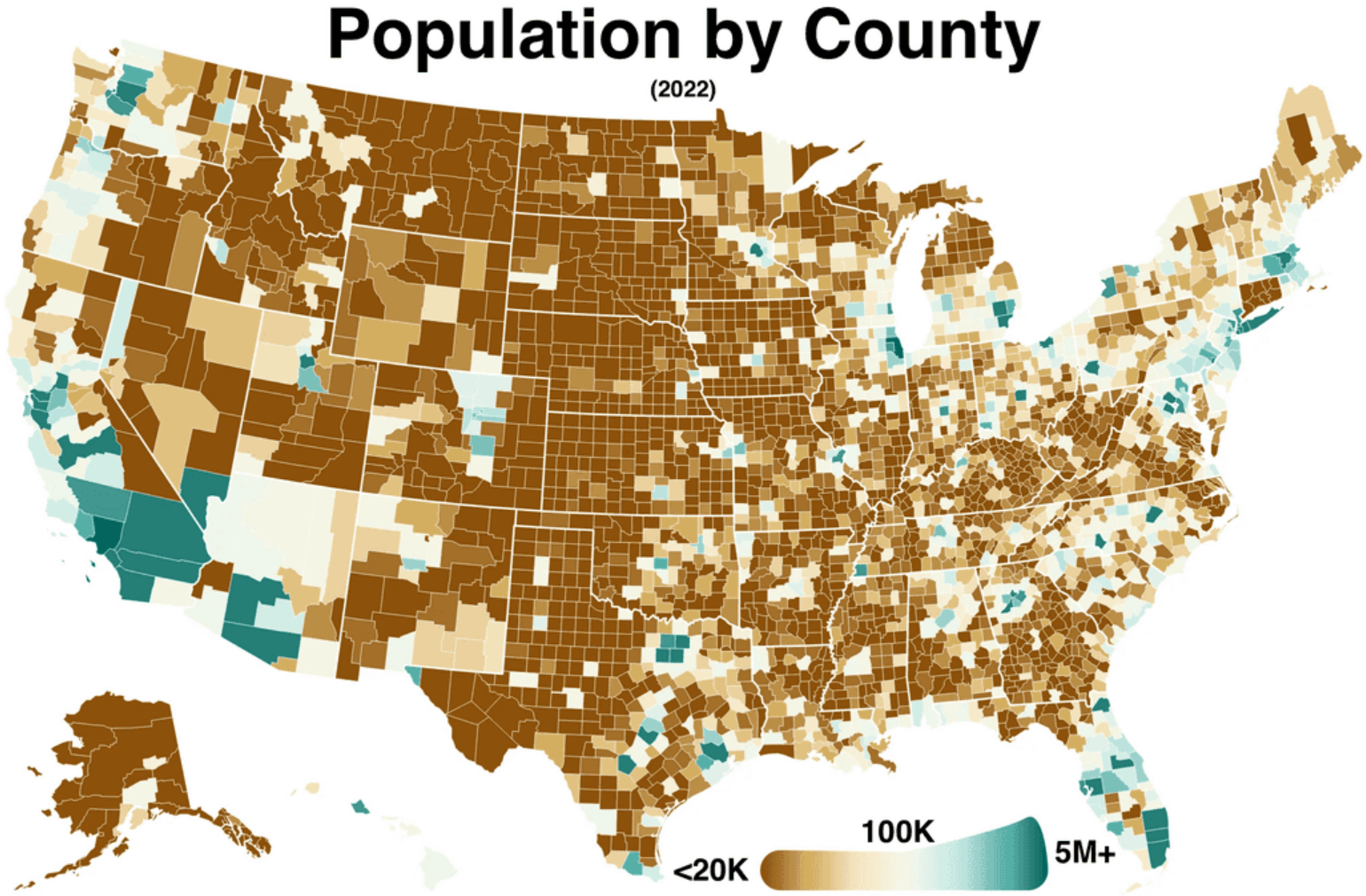
OUTLINE

- *Displaying quantities*
- *Displaying distributions*
- *Displaying proportions*
- Displaying relationships
- *Displaying time series*
- **Displaying geospatial data**
- Other charts
- Uncertainty

GEOSPATIAL DATA. CHOROPLETH MAPS

- Choropleth maps (Greek: choro = area, pleth = value)
  - Assume that the mapped attribute is uniformly distributed in the regions
  - Used to emphasize the spatial distribution of one or more geographic attributes.
  - Data normalization and color or grayscale mapping are important design decisions
  - Similar types: dasymetric, isarithmic, isometric, isopleth…

GEOSPATIAL DATA. CHOROPLETH MAPS



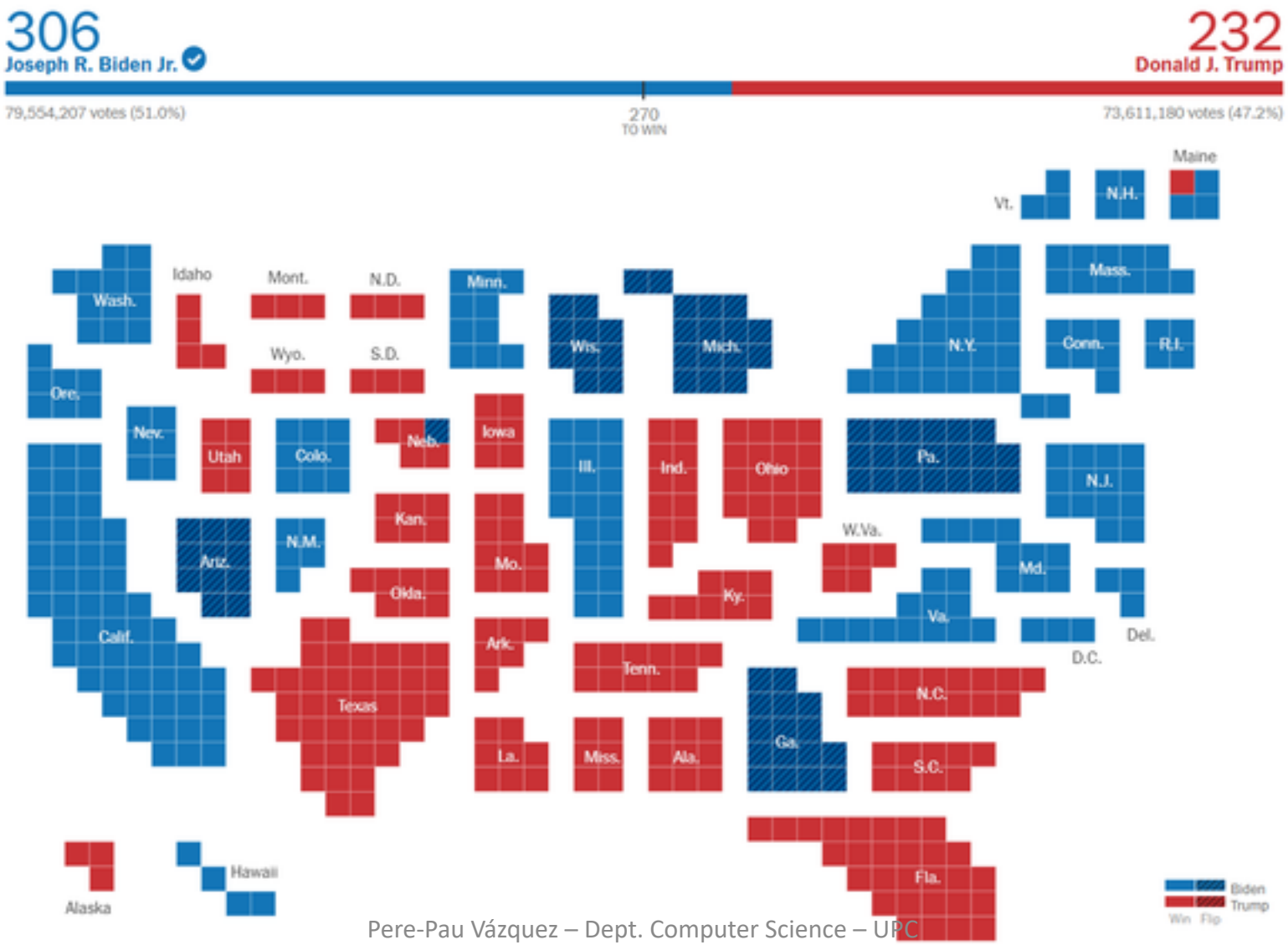
GEOSPATIAL DATA. CHOROPLETH MAPS

- Use colors to encode geographically aggregated data
- It is difficult to use properly:
  - Normalized data is often required (e.g. to communicate population density)
  - Shaded values can be perceived differently according to the area of the geographic region
- Encode: Color: Sequential segmented colormap

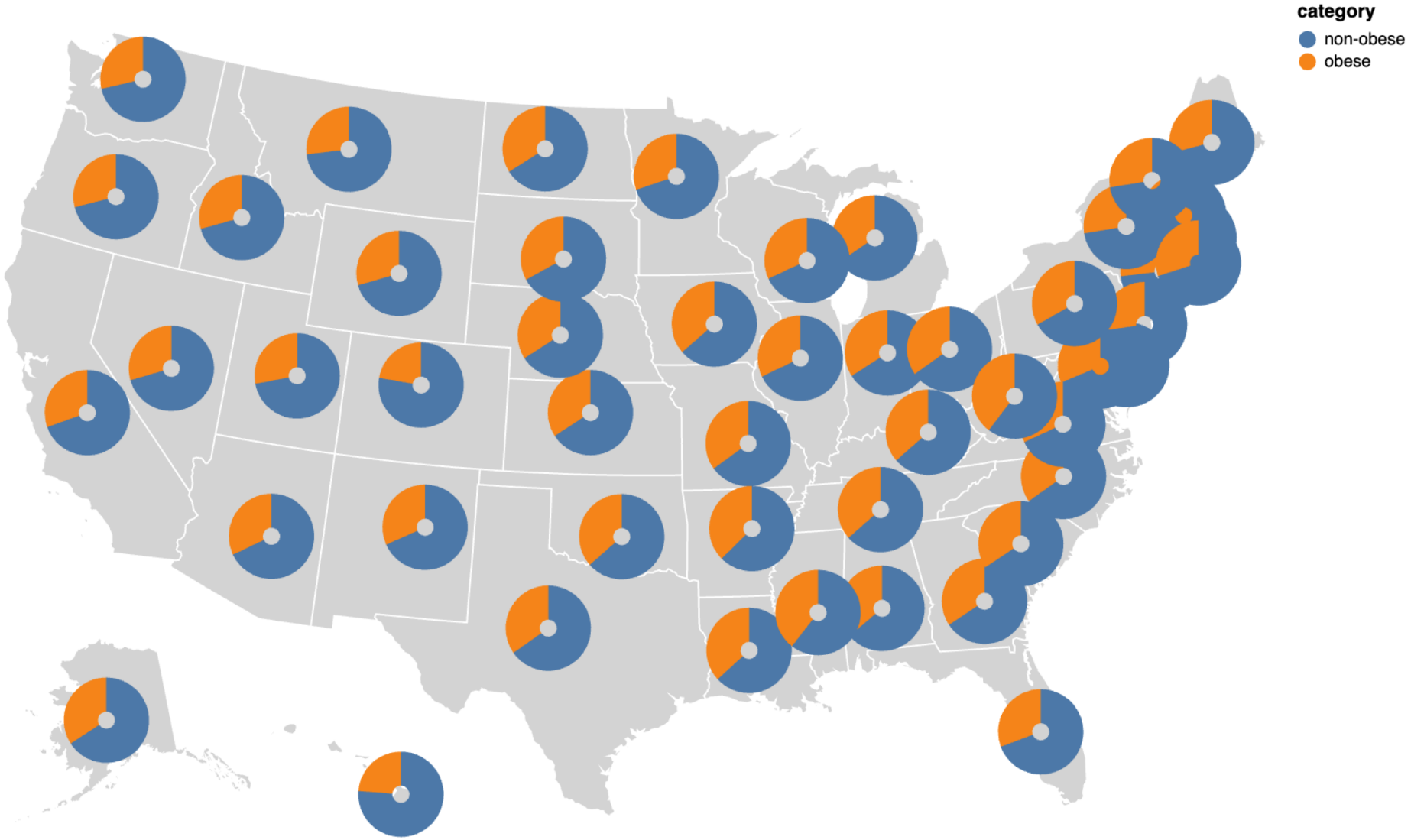
GEOSPATIAL DATA. CHOROPLETH MAPS

- Issues
  - Larger regions → larger importance
  - Must (commonly) take into account population density
  - Color issues

- https://www.nytimes.com/interactive/2020/10/30/opinion/election-results-maps.html


306
Joseph R. Biden Jr.
79,554,207 votes (51.0%)
270
TO WIN
232
Donald J. Trump
73,611,180 votes (47.2%)
Maine
Vt.
N.H.
Wash.
Idaho
Mont.
N.D.
Minn.
Mass.
Ore.
Wyo.
S.D.
Wis.
Mich.
N.Y.
Conn.
R.I.
Nev.
Utah
Colo.
Neb.
Iowa
Ill.
Ind.
Ohio
Pa.
N.J.
Kan.
N.M.
Ariz.
Mo.
W.Va.
Md.
Ky.
Okla.
Va.
Del.
Calif.
Ark.
D.C.
Tenn.
N.C.
Texas
La.
Miss.
Ala.
Ga.
S.C.
Fla.
Alaska
Hawaii
Biden
Trump
Win Flip

GEOSPATIAL DATA.
GRADUATED SYMBOL MAPS

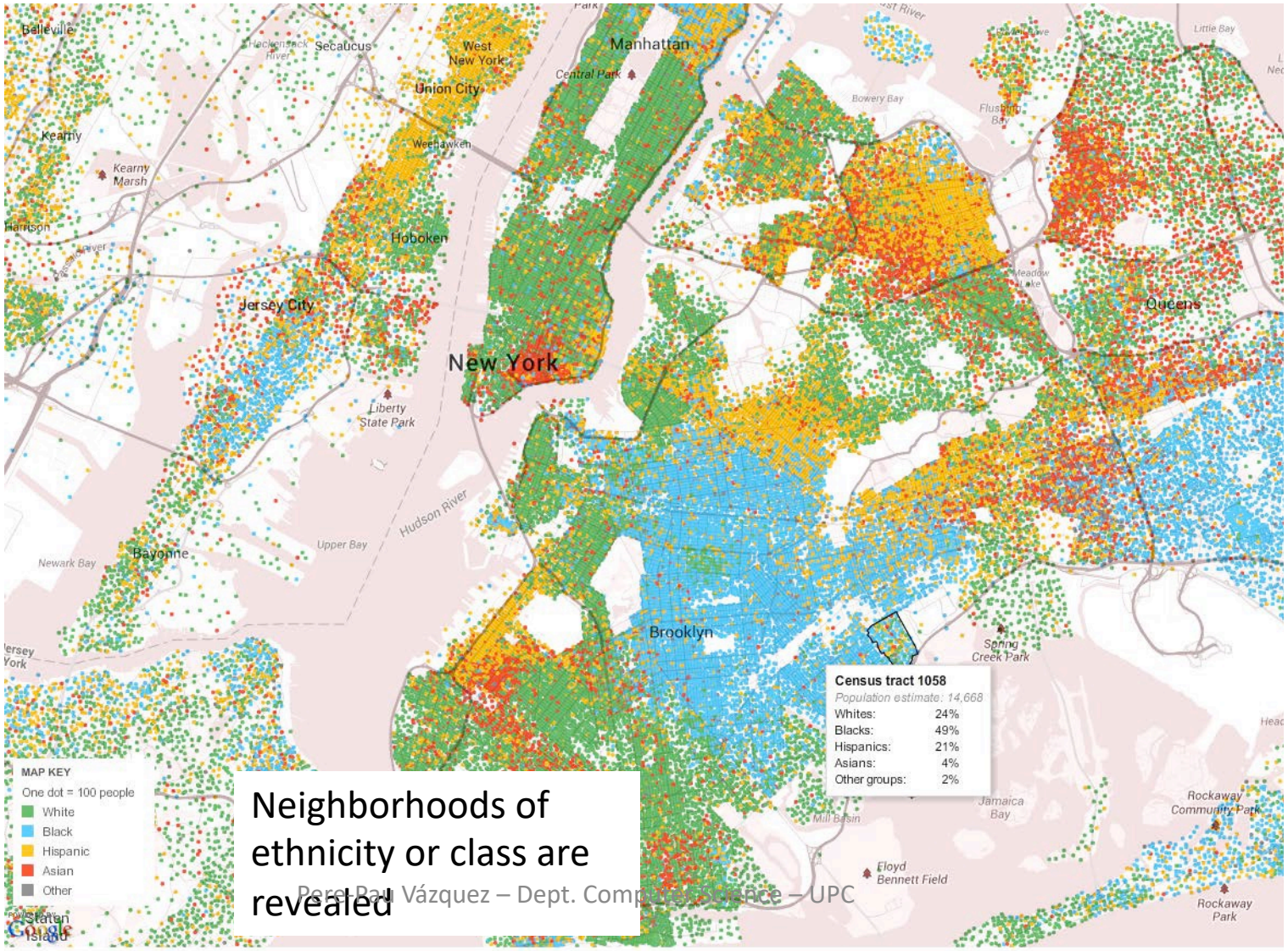


GEOSPATIAL DATA.
GRADUATED SYMBOL MAPS

- Places symbols on the underlying map
- Avoids confounding geographic area with data values
- Enables visualizing more dimensions

GEOSPATIAL DATA. DOT MAPS

- Census data



Neighborhoods of ethnicity or class are revealed

GEOSPATIAL DATA.
DOT MAPS

- Dot maps. Issues:
  - If the size of the symbol is used to represent a quantitative parameter, scaling may present perception issues
    - Perceived size also depends on their local neighborhood
  - If color is used to represent a quantitative parameter, problems of color perception also appear
  - Large data cause overlap or overplotting problems
    - Especially in highly populated areas, while low-population areas are virtually empty

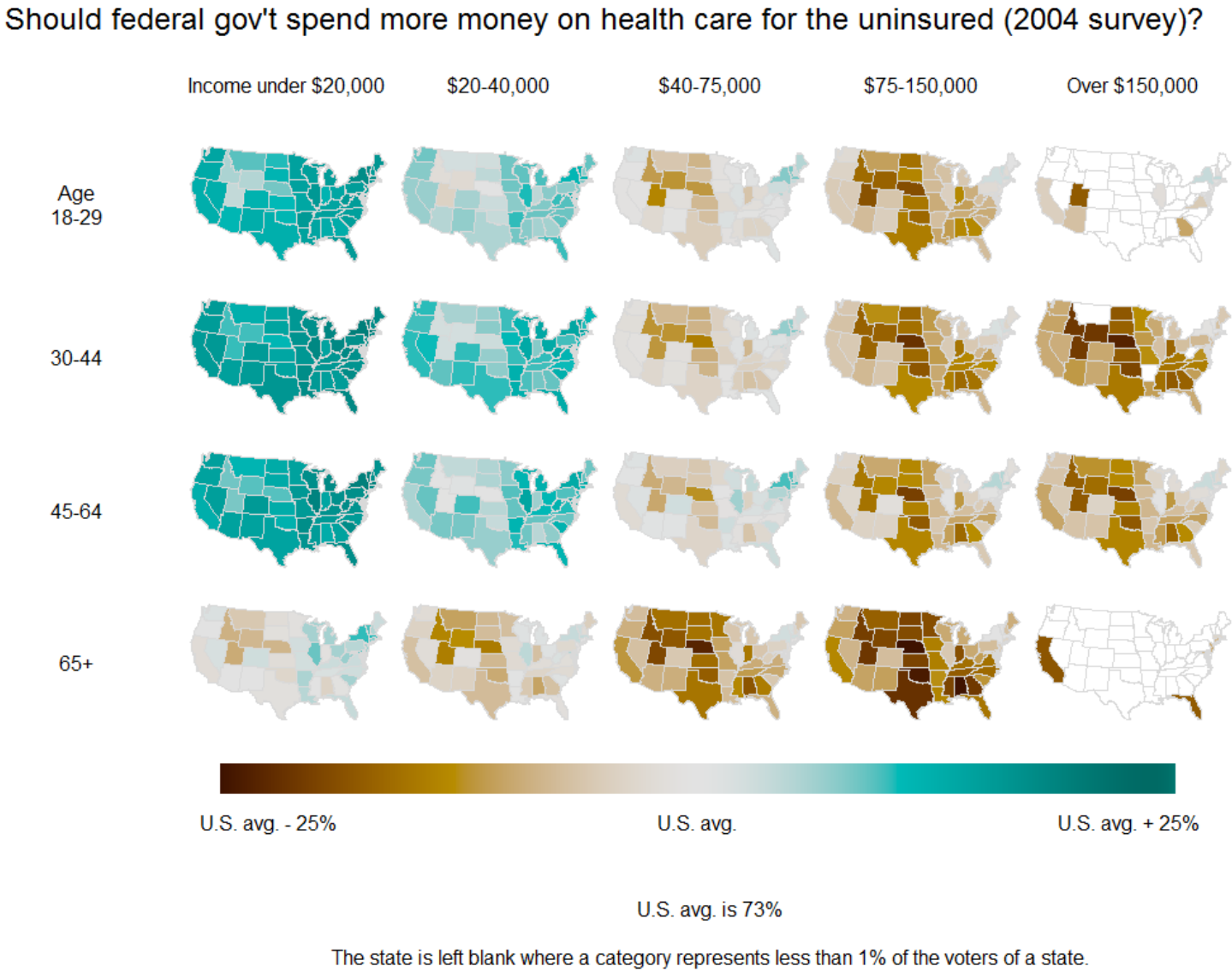
OUTLINE

- *Displaying quantities*
- *Displaying distributions*
- *Displaying proportions*
- Displaying relationships
- *Displaying time series*
- *Displaying geospatial data*
- **Other charts**
- Uncertainty

OTHER CHARTS.
MULTIPLE VARIABLES. SMALL MULTIPLES

- Grid (or column, or row) of small(er) charts
- Its goal is to ease comparison
- Need to preserve axes, alignments
  - Key to comparisons
- Sometimes can compare in both dimensions

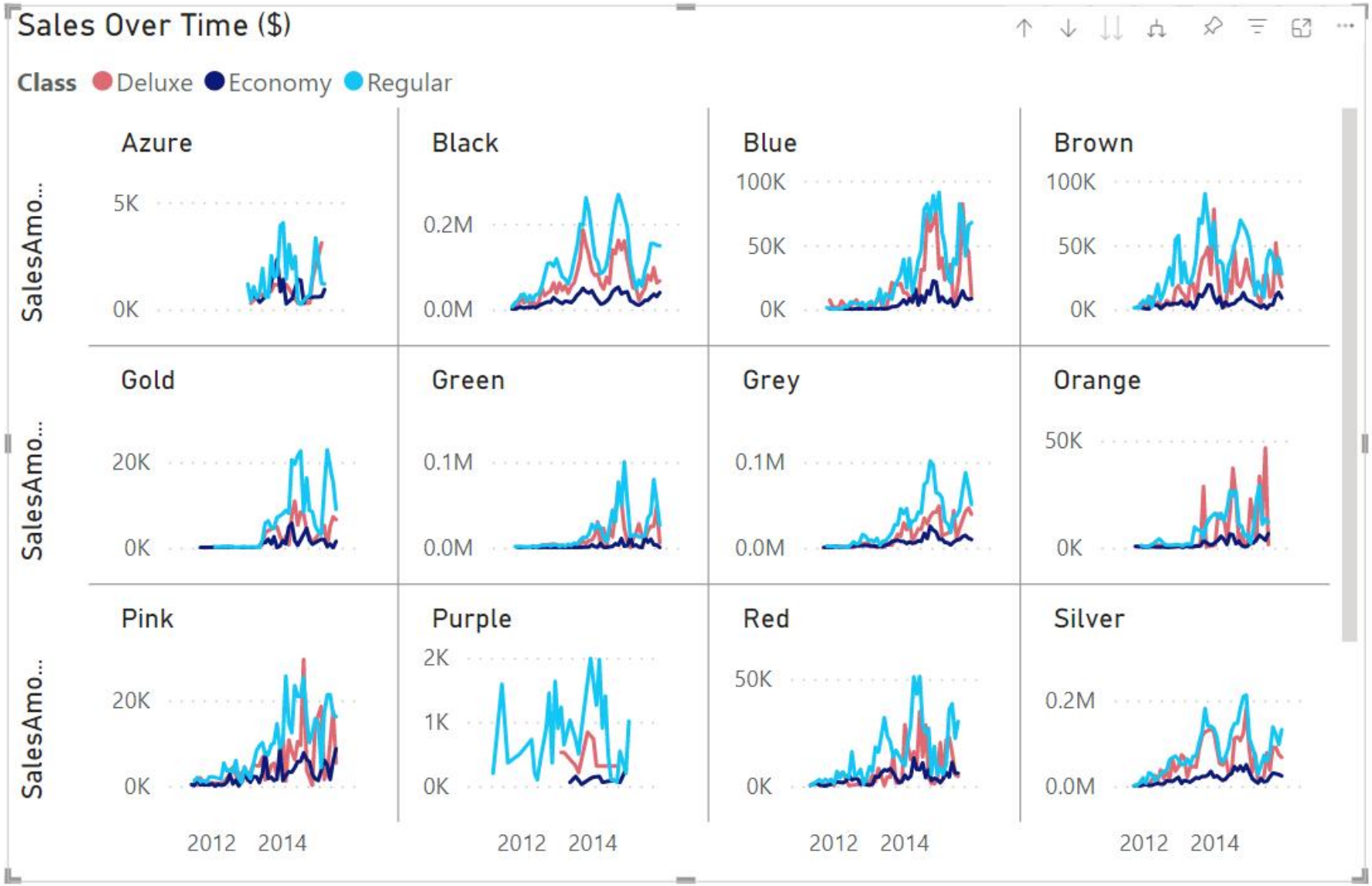
OTHER CHARTS.
MULTIPLE VARIABLES. SMALL MULTIPLES



OTHER CHARTS.
TIME SERIES. SMALL MULTIPLES

- Instead of stacking, placing series together
  - Same axes
- Task:
  - Find trends, seasonal patterns
- Can use any type of visualization

OTHER CHARTS.
TIME SERIES. SMALL MULTIPLES

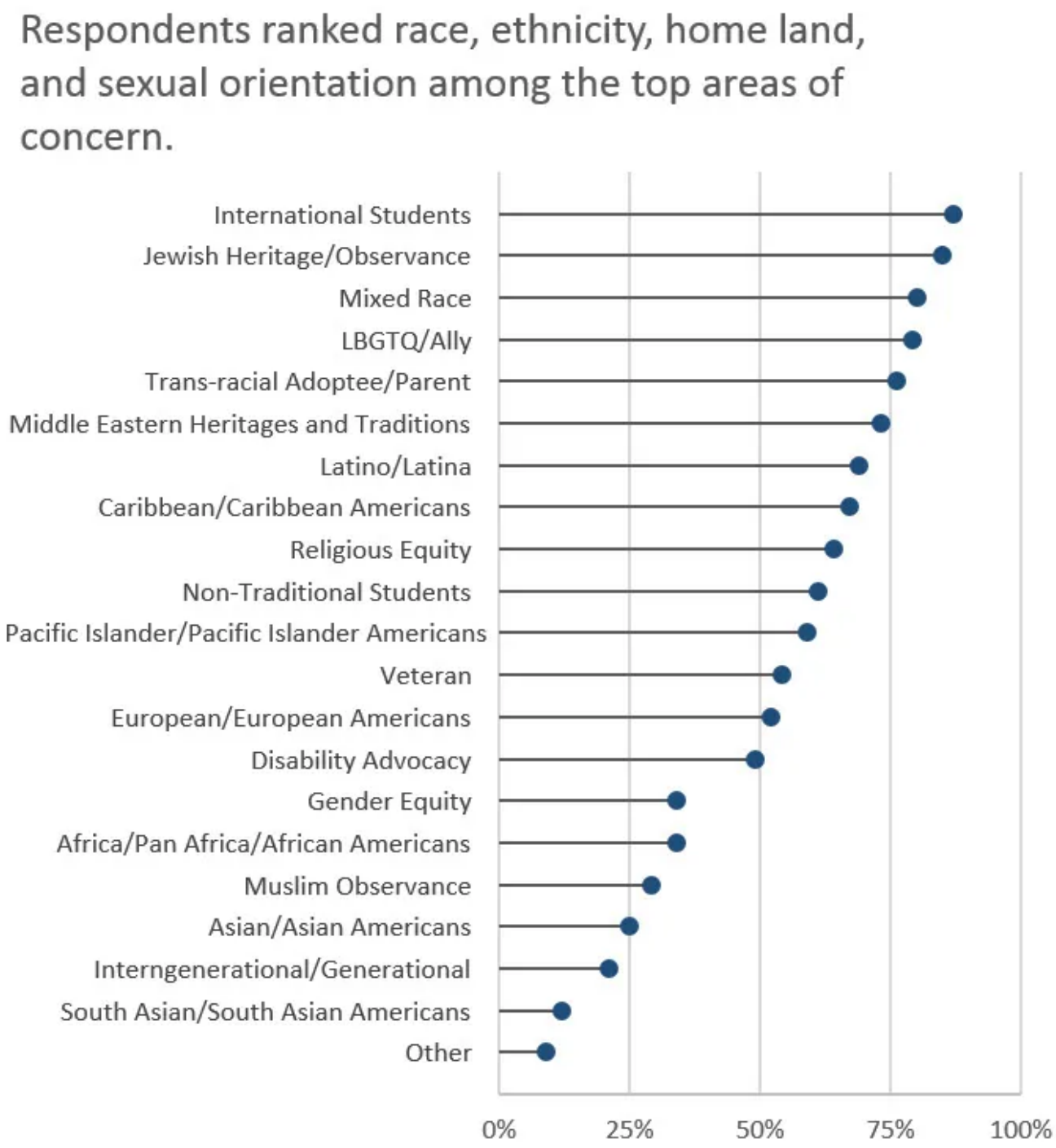


https://sumproduct.com/thought/unshared-and-unsynchronised-axes-for-small-multiples-charts/

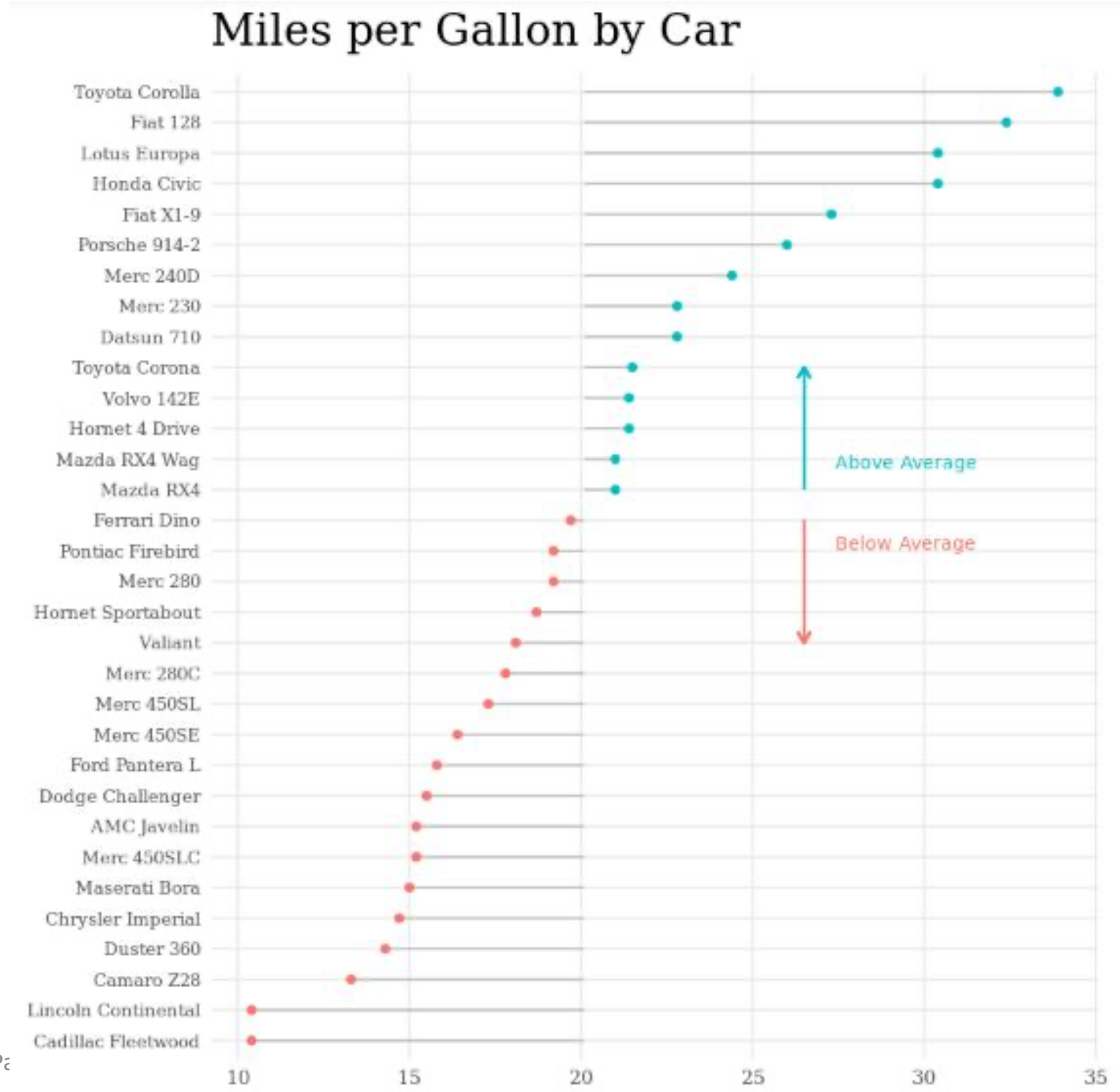
OTHER CHARTS. QUANTITIES. LOLLIPOP

Might not start at zero

https://stephanieevergreen.com/lollipop/



OTHER CHARTS. QUANTITIES. LOLLIPOP

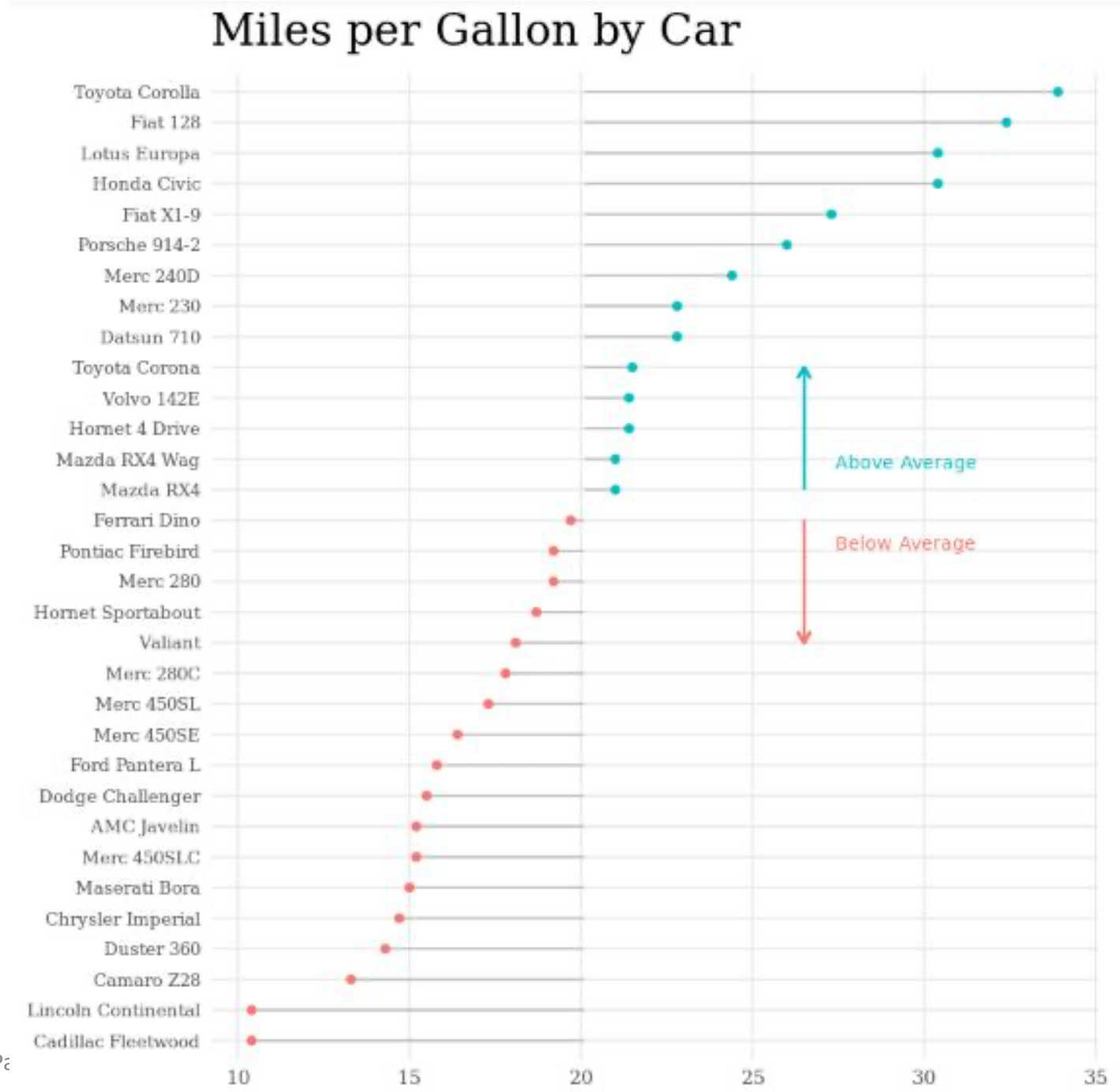


https://www.statology.org/lollipop-chart-r/

OTHER CHARTS. QUANTITIES. LOLLIPOP

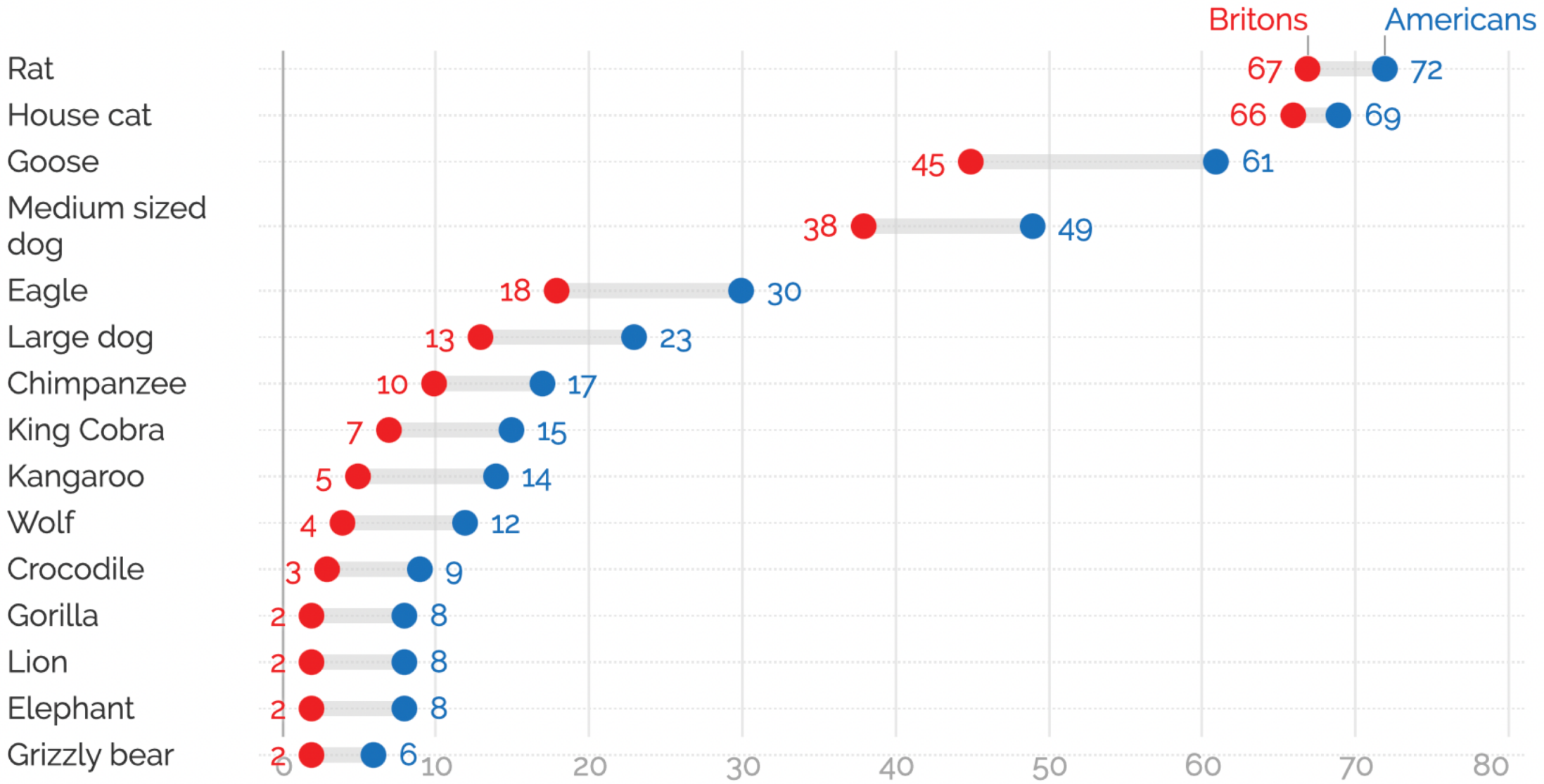
Are horizontal gridlines needed?

Increase contrast



https://www.statology.org/lollipop-chart-r/

OTHER CHARTS. QUANTITIES.
RANGE PLOT: DOT PLOT WITH TWO VALUES

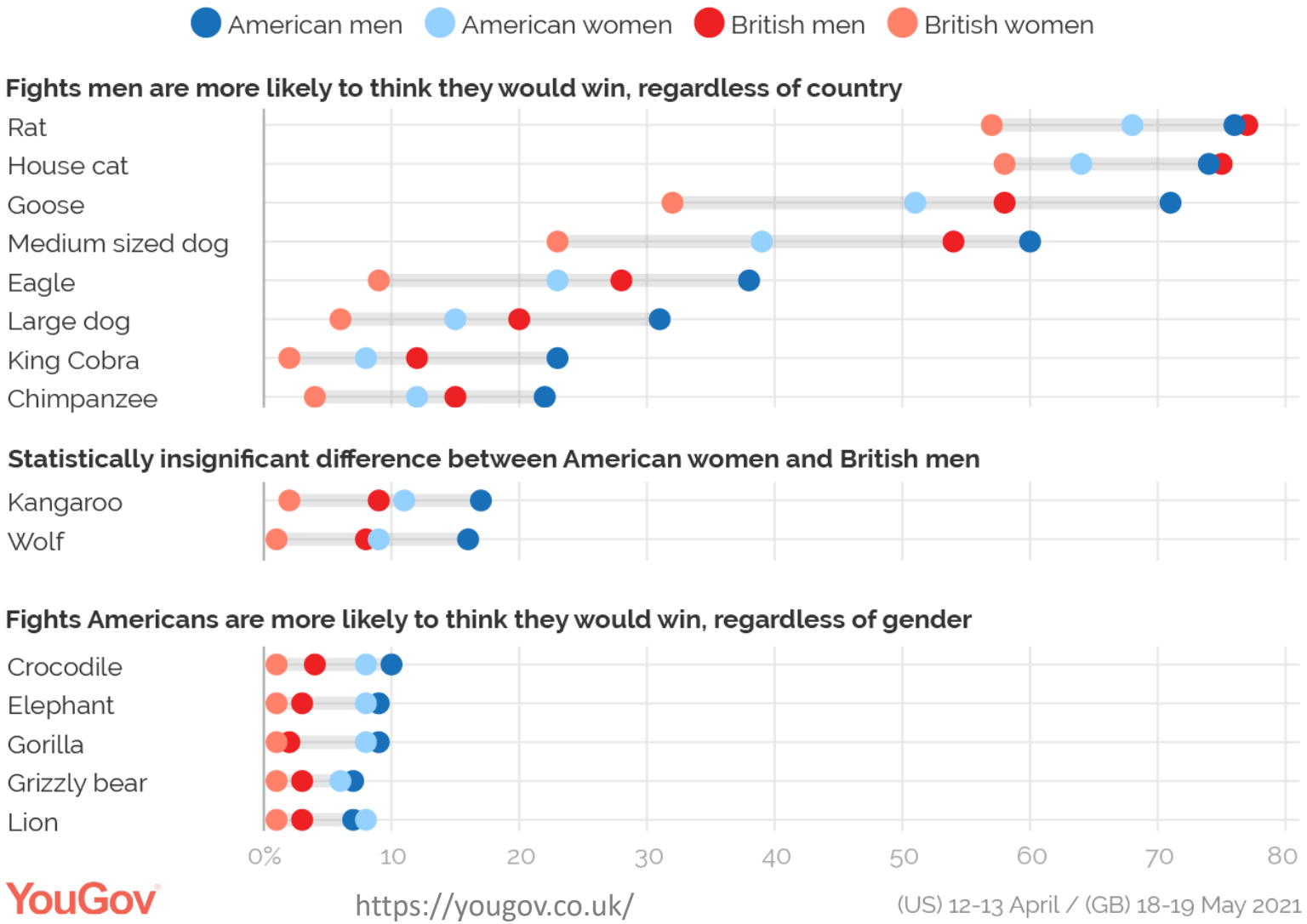


https://yougov.co.uk/

OTHER CHARTS
RANGE PLOT: D

British women are less likely than American women to think they could beat animals in a fight, and, when it comes to the scariest animals, so are British men

Which of the following animals, if any, do you think you could beat in a fight if you were unarmed?
% of 2,082 GB adults



YouGov https://yougov.co.uk/ (US) 12-13 April / (GB) 18-19 May 2021

OUTLINE

- *Displaying quantities*
- *Displaying distributions*
- *Displaying proportions*
- Displaying relationships
- *Displaying time series*
- *Displaying geospatial data*
- *Other charts*
- **Uncertainty**

Visual vocabulary

Designing with data

There are so many ways to visualise data - how do we know which one to pick? Use the categories across the top to decide which data relationship is most important in your story, then look at the different types of chart within the category to form some initial ideas about what might work best. This list is not meant to be exhaustive, nor a wizard, but is a useful starting point for making informative and meaningful data visualisations.

FT graphic: Alan Smith; Chris Campbell; Ian Bott; Liz Faunce; Graham Parish; Billy Ehrenberg-Shannon; Paul McCallum; Martin Stabe
Inspired by the Graphic Continuum by Jon Schwabish and Severino Ribecca

ft.com/vocabulary

Deviation

Emphasise variations (+/-) from a fixed reference point. Typically the reference point is zero but it can also be a target or a long-term average. Can also be used to show sentiment (positive/neutral/negative).

Example FT uses
Trade surplus/deficit, climate change

Diverging bar

A simple standard bar chart that can handle both negative and positive magnitude values.

Diverging stacked bar

Perfect for presenting survey results which involve sentiment (eg disagree/neutral/agree).

Spine

Splits a single value into two contrasting components (eg male/female).

Surplus/deficit filled line

The shaded area of these charts allows a balance to be shown – either against a baseline or between two series.

Correlation

Show the relationship between two or more variables. Be mindful that, unless you tell them otherwise, many readers will assume the relationships you show them to be causal (i.e. one causes the other).

Example FT uses
Inflation and unemployment, income and life expectancy

Scatterplot

The standard way to show the relationship between two continuous variables, each of which has its own axis.

Column + line timeline

A good way of showing the relationship between an amount (columns) and a rate (line).

Connected scatterplot

Usually used to show how the relationship has changed over time.

Bubble

Like a scatterplot, but adds additional detail by sizing the circles according to a third variable.

XY heatmap

A good way of showing the patterns between 2 categories of data, less effective at showing fine differences in amounts.

Ranking

Use where an item's position in an ordered list is more important than its absolute or relative value. Don't be afraid to highlight the points of interest.

Example FT uses
Wealth, deprivation, league tables, constituency election results

Ordered bar

Standard bar charts display the ranks of values much more easily when sorted into order.

Ordered column

See above.

Ordered proportional symbol

Use when there are big variations between values and/or seeing fine differences between data is not so important.

Dot strip plot

Dots placed in order on a strip are a space-efficient method of laying out ranks across multiple categories.

Slope

Perfect for showing how ranks have changed over time or vary between categories.

Lollipop

Lollipops draw more attention to the data value than standard bar/column and can also show rank and value effectively.

Bump

Effective for showing changing rankings across multiple dates. For large datasets, consider grouping lines using colour.

Distribution

Show values in a dataset and how often they occur. The shape (or 'skew') of a distribution can be a memorable way of highlighting the lack of uniformity or equality in the data.

Example FT uses
Income distribution, population (age/sex) distribution, revealing inequality

Histogram

The standard way to show a statistical distribution – keep the gaps between columns small to highlight the 'shape' of the data.

Dot plot

A simple way of showing the change or range (min/max) of data across multiple categories.

Dot strip plot

Good for showing individual values in a distribution, can be a problem when too many dots have the same value.

Barcode plot

Like dot strip plots, good for displaying all the data in a table, they work best when highlighting individual values.

Boxplot

Summarise multiple distributions by showing the median (centre) and range of the data

Violin plot

Similar to a box plot but more effective with complex distributions (data that cannot be summarised with simple average).

Population pyramid

A standard way for showing the age and sex breakdown of a population distribution; effectively, back to back histograms.

Cumulative curve

A good way of showing how unequal a distribution is: y axis is always cumulative frequency, x axis is always a measure.

Frequency polygons

For displaying multiple distributions of data. Like a regular line chart, best limited to a maximum of 3 or 4 datasets.

Beeswarm

Use to emphasise individual points in a distribution. Points can be sized to an additional variable. Best with medium-sized datasets

Change over Time

Give emphasis to changing trends. These can be short (intra-day) movements or extended series traversing decades or centuries. Choosing the correct time period is important to provide suitable context for the reader.

Example FT uses
Share price movements, economic time series, sectoral changes in a market

Line

The standard way to show a changing time series. If data are irregular, consider markers to represent data points.

Column

Columns work well for showing change over time – but usually best with only one series of data at a time.

Column + line timeline

A good way of showing the relationship over time between an amount (columns) and a rate (line).

Slope

Good for showing changing data as long as the data can be simplified into 2 or 3 points without missing a key part of story.

Area chart

Use with care – these are good at showing changes to total, but seeing change in components can be very difficult.

Candlestick

Usually focused on day-to-day activity, these charts show opening/closing and high/low points of each day.

Fan chart (projections)

Use to show the uncertainty in future projections - usually this grows the further forward to projection.

Connected scatterplot

A good way of showing changing data for two variables whenever there is a relatively clear pattern of progression.

Calendar heatmap

A great way of showing temporal patterns (daily, weekly, monthly) – at the expense of showing precision in quantity.

Priestley timeline

Great when date and duration are key elements of the story in the data.

Circle timeline

Good for showing discrete values of varying size across multiple categories (eg earthquakes by continent).

Vertical timeline

Presents time on the Y axis. Good for displaying detailed time series that work especially well when scrolling on mobile.

Seismogram

Another alternative to the circle timeline for showing series where there are big variations in the data.

Streamgraph

A type of area chart; use when seeing changes in proportions over time is more important than individual values

Magnitude

Show size comparisons. These can be relative (just being able to see larger/bigger) or absolute (need to see fine differences). Usually these show a 'counted' number (for example, barrels, dollars or people) rather than a calculated rate or per cent.

Example FT uses
Commodity production, market capitalisation, volumes in general

Column

The standard way to compare the size of things. Must always start at 0 on the axis.

Bar

See above. Good when the data are not time series and labels have long category names.

Paired column

As per standard column but allows for multiple series. Can become tricky to read with more than 2 series.

Paired bar

See above.

Marimekko

A good way of showing the size and proportion of data at the same time – as long as the data are not too complicated.

Proportional symbol

Use when there are big variations between values and/or seeing fine differences between data is not so important.

Isotype (pictogram)

Excellent solution in some instances – use only with whole numbers (do not slice off an arm to represent a decimal).

Lollipop

Lollipop charts draw more attention to the data value than standard bar/column – does not have to start at zero (but preferable).

Radar

A space-efficient way of showing value of multiple variables– but make sure they are organised in a way that makes sense to reader.

Parallel coordinates

An alternative to radar charts – again, the arrangement of the variables is important. Usually benefits from highlighting values.

Bullet

Good for showing a measurement against the context of a target or performance range.

Grouped symbol

An alternative to bar/column charts when being able to count data or highlight individual elements is useful.

Part-to-whole

Show how a single entity can be broken down into its component elements. If the reader's interest is solely in the size of the components, consider a magnitude-type chart instead.

Example FT uses
Fiscal budgets, company structures, national election results

Stacked column/bar

A simple way of showing part-to-whole relationships but can be difficult to read with more than a few components.

Marimekko

A good way of showing the size and proportion of data at the same time – as long as the data are not too complicated.

Pie

A common way of showing part-to-whole data – but be aware that it's difficult to accurately compare the size of the segments.

Donut

Similar to a pie chart – but the centre can be a good way of making space to include more information about the data (eg total).

Treemap

Use for hierarchical part-to-whole relationships can be difficult to read when there are many small segments.

Voronoi

A way of turning points into areas – any point within each area is closer to the central point than any other centroid.

Arc

A hemicycle, often used for visualising parliamentary composition by number of seats.

Gridplot

Good for showing % information, they work best when used on whole numbers and work well in small multiple layout form.

Venn

Generally only used for schematic representation.

Waterfall

Can be useful for showing part-to-whole relationships where some of the components are negative.

Spatial

Aside from locator maps only used when precise locations or geographical patterns in data are more important to the reader than anything else.

Example FT uses
Population density, natural resource locations, natural disaster risk/impact, catchment areas, variation in election results

Basic choropleth (rate/ratio)

The standard approach for putting data on a map – should always be rates rather than totals and use a sensible base geography.

Proportional symbol (count/magnitude)

Use for totals rather than rates – be wary that small differences in data will be hard to see.

Flow map

For showing unambiguous movement across a map.

Contour map

For showing areas of equal value on a map. Can use deviation colour schemes for showing +/- values

Equalised cartogram

Converting each unit on a map to a regular and equally-sized shape – good for representing voting regions with equal value.

Scaled cartogram (value)

Stretching and shrinking a map so that each area is sized according to a particular value.

Dot density

Used to show the location of individual events/locations – make sure to annotate any patterns the reader should see.

Heat map

Grid-based data values mapped with an intensity colour scale. As choropleth map – but not snapped to an admin/political unit.

Flow

Show the reader volumes or intensity of movement between two or more states or conditions. These might be logical sequences or geographical locations.

Example FT uses
Movement of funds, trade, migrants, lawsuits, information, relationship graphs.

Sankey

Shows changes in flows from one condition to at least one other; good for tracing the eventual outcome of a complex process.

Waterfall

Designed to show the sequencing of data through a flow process, typically budgets. Can include +/- components.

Chord

A complex but powerful diagram which can illustrate 2-way flows (and net winner) in a matrix.

Network

Used for showing the strength and inter-connectedness of relationships of varying types.

FURTHER READING & OTHER RESOURCES

- Jon Schwabish's "One chart at a time" Youtube series: https://www.youtube.com/watch?v=gFFj22kjlZk
- https://datavizproject.com/data-type/
- https://github.com/ft-interactive/chart-doctor/tree/master/visual-vocabulary
- https://public.tableau.com/es-es/s/gallery/visual-vocabulary
- https://clauswilke.com/dataviz
- https://blog.hubspot.com/marketing/types-of-graphs-fordata-visualization
- Timeline of datavis milestones http://www.datavis.ca/milestones/

VISUALIZATION TECHNIQUES

PERE-PAU VÁZQUEZ – VIRVIG GROUP – UPC

EXERCISES

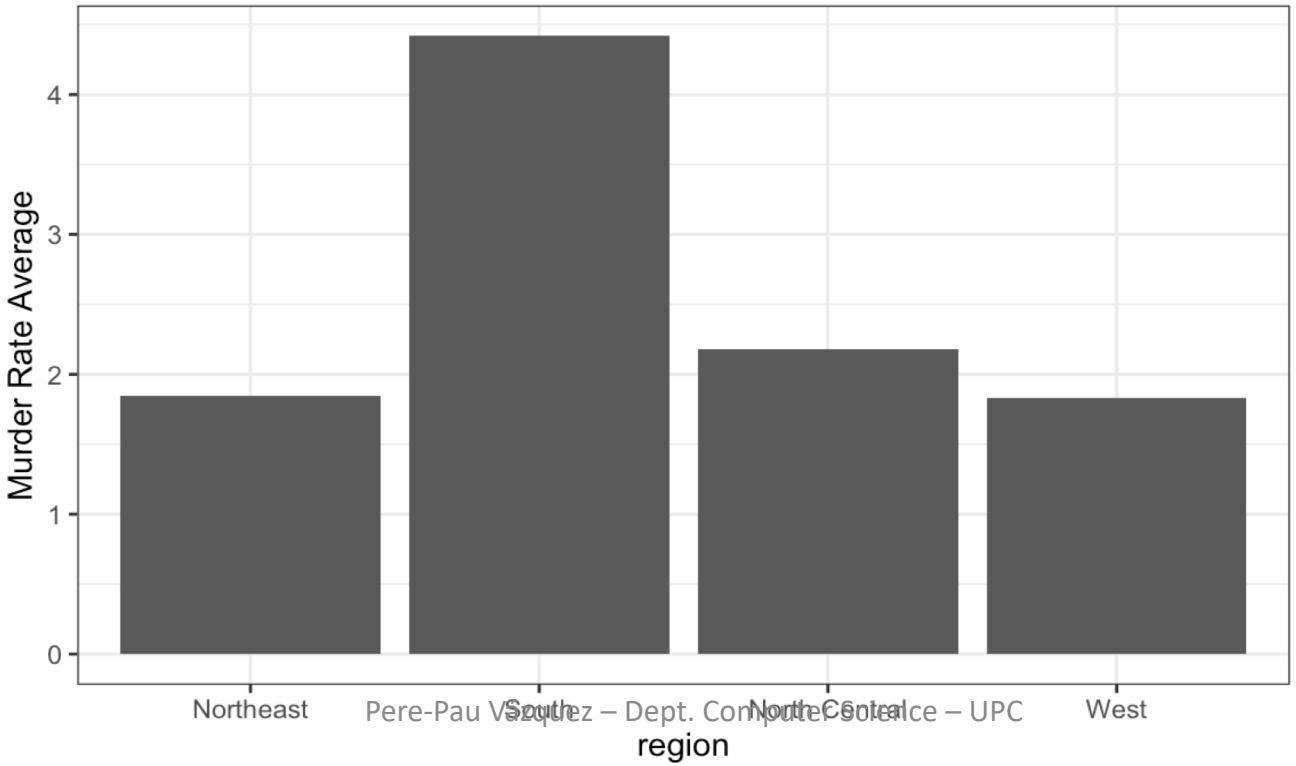
PERE-PAU VÁZQUEZ – VIRVIG GROUP – UPC

EXERCISE

- **We have a dataset with driving tickets. In some cases, we have missing values in the amount of the fine. Comment on the following two strategies to deal with missing values. Discuss strengths and weaknesses:**
  a. Erase the rows
  b. Encode the value with a particular number, such as a negative, e.g. -10000

REPRESENTATIONS. EXERCISE

- Say we are interested in comparing gun homicide rates across regions of the US. We see this plot:
  - What region would you consider safer? Why? Discuss its design.

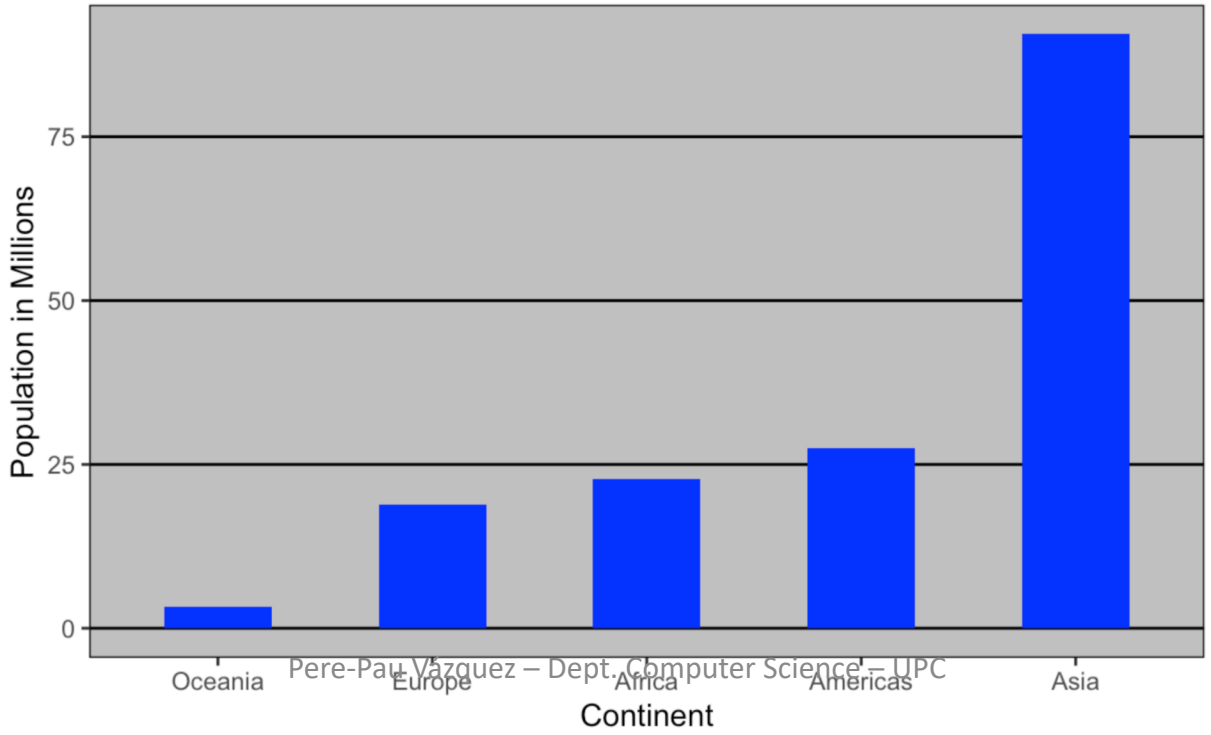


EXERCISE

- **We want to compare the evolution of the energy consumption at UPC with the price of the electricity along one year. Do you think a line plot with dual axis could work?**

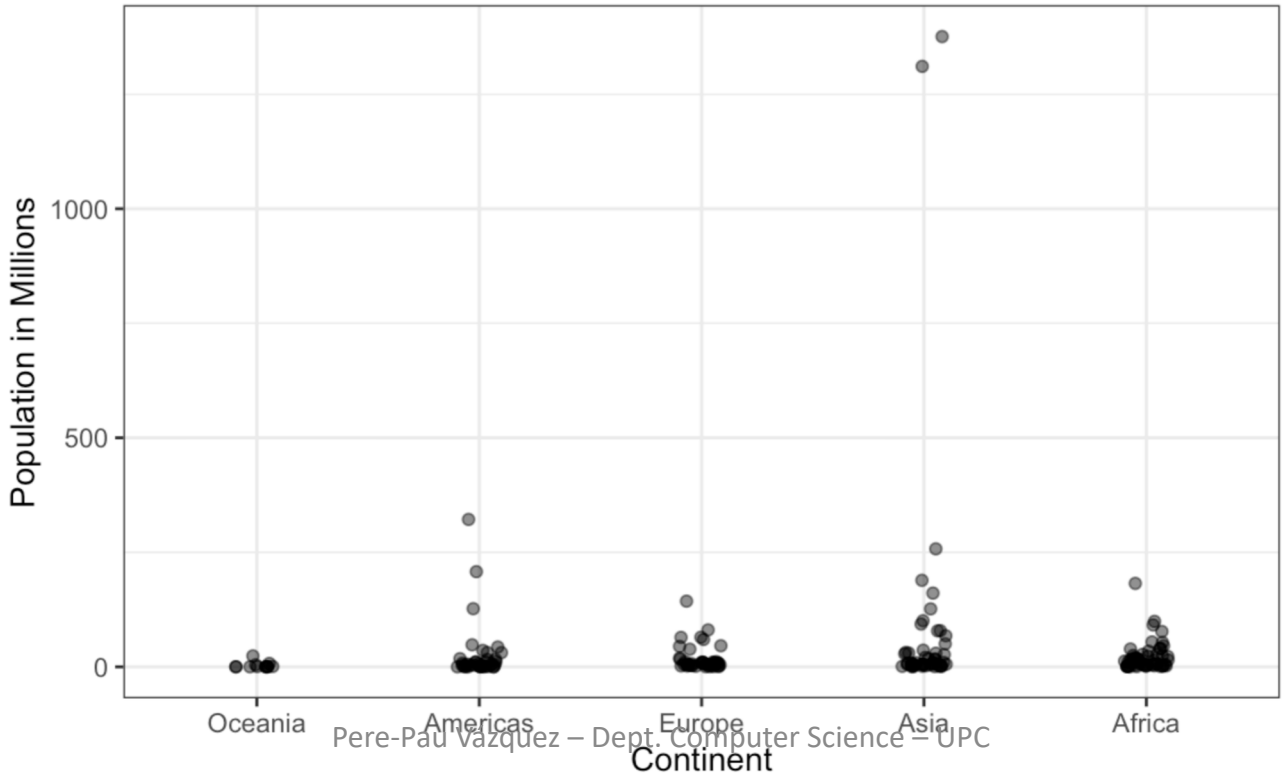
EXERCISE

- Consider the following chart that shows the population of each continent in 2015. Do you think it is faithfully conveying the data? Why?



EXERCISE

- What if we represent the data in the following way? Do you think it is effective at showing the data? Why?

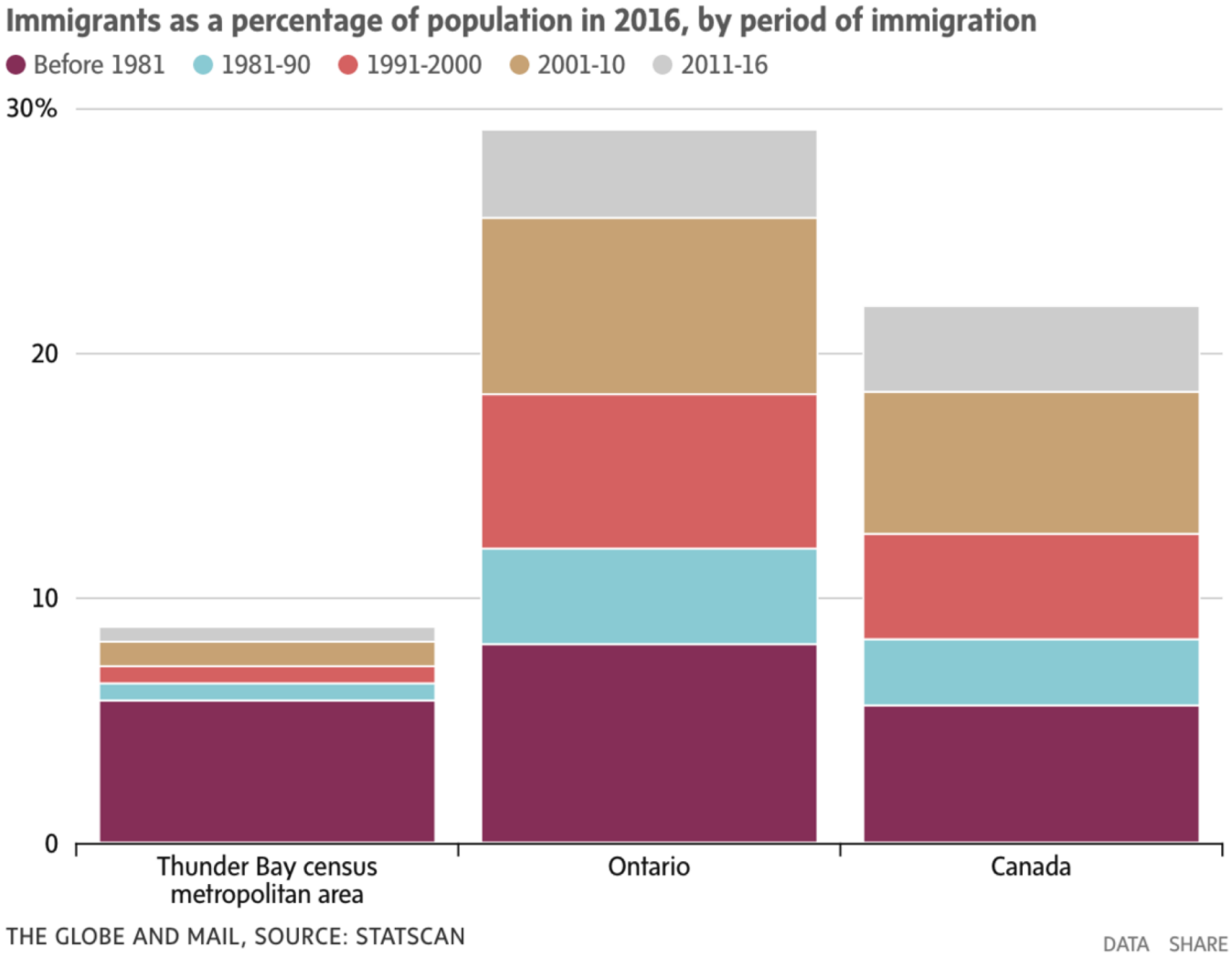


EXERCISE

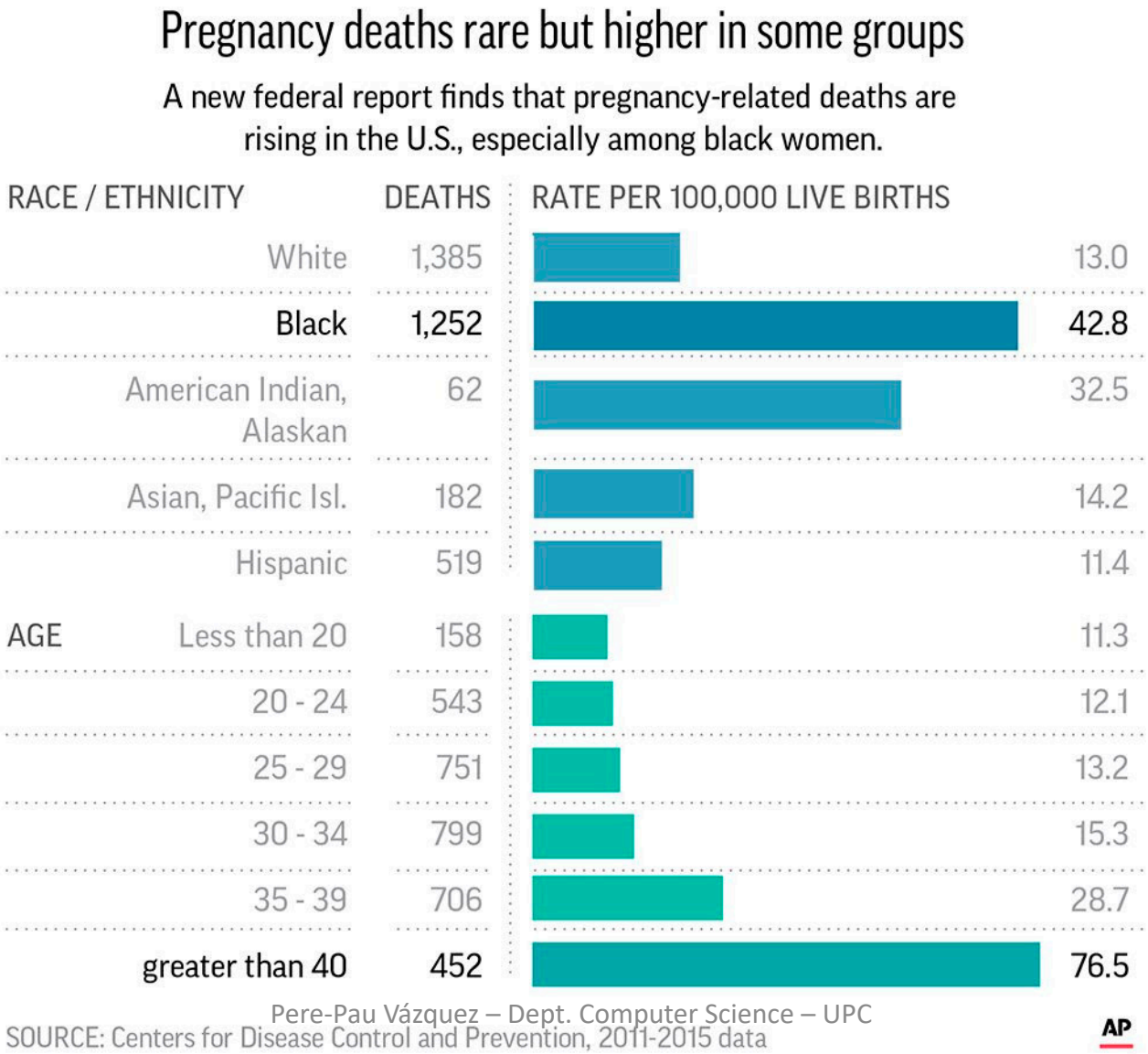
We have installed a noise sensor next to a train station. We have captured noise information for 24 hours and we have labelled the different sources of noise that seem to be the noise origin every second. Every second we have one label, and there are a total of 12 different labels. How would you represent the information visually?

a. Bar chart with colors for the different labels
b. Line chart with colors for the different labels
c. Heatmap with colors for the different labels
d. None of the previous ones

EXERCISE


Immigrants as a percentage of population in 2016, by period of immigration
Before 1981
1981-90
1991-2000
2001-10
2011-16
30%
20
10
0
Thunder Bay census metropolitan area
Ontario
Canada
THE GLOBE AND MAIL, SOURCE: STATSCAN
DATA
SHARE

EXERCISE


Pregnancy deaths rare but higher in some groups
A new federal report finds that pregnancy-related deaths are rising in the U.S., especially among black women.
RACE / ETHNICITY
DEATHS
RATE PER 100,000 LIVE BIRTHS
White 1,385 13.0
Black 1,252 42.8
American Indian, Alaskan 62 32.5
Asian, Pacific Isl. 182 14.2
Hispanic 519 11.4
AGE
Less than 20 158 11.3
20 - 24 543 12.1
25 - 29 751 13.2
30 - 34 799 15.3
35 - 39 706 28.7
greater than 40 452 76.5
SOURCE: Centers for Disease Control and Prevention, 2011-2015 data
AP

EXERCISE

- **We want to compare the revenue of the 50 most relevant content creators living in Andorra from the last 10 years. How would you do it? Describe and justify the design. Discuss advantages and shortcomings.**

EXERCISES

PERE-PAU VÁZQUEZ – VIRVIG GROUP – UPC