ID430B: Data Analytics for Designers 디자인 특강V <디자이너를 위한 데이터 분석>

Lecture 6
Data Visualization 2/3

Tak Yeon Lee <takyeonlee@kaist.ac.kr> (takyeonlee.com)
AI-Experience-Lab (reflect9.github.io/ael)

Things to Learn

1. **Dimensionality**
    1. How to visualize 1D (univariate), 2D, 3D, and higher-dimension dataset
2. **DO and DON'T**
    1. Heuristics of bad, good, and better visualizations
3. **Advanced Charts**
    1. Hierarchy, Flow, Network, Text
4. **Interactivity**
    1. Shneiderman's visual information seeking Mantra

Dimensionality

:= How many columns are used in one chart

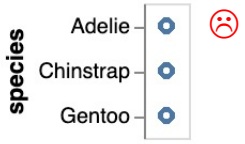
What are good / bad ways to visualize data of a specific dimension?

1D Nominal

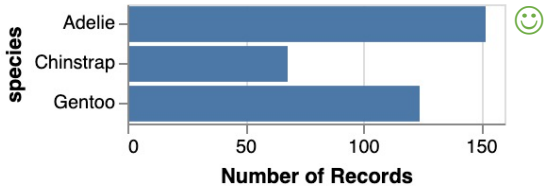
When you are interested in a single column containing nominal values (i.e. only frequency counting is allowed)
E.g. **species** column of the penguin dataset

Little can be done with the raw data

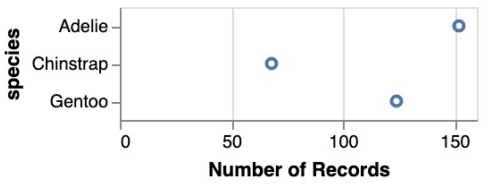
After aggregation (i.e. frequency counting), you can do basic visual analysis



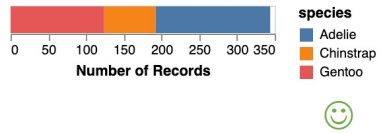
☹



☺

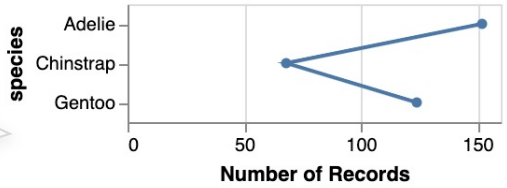


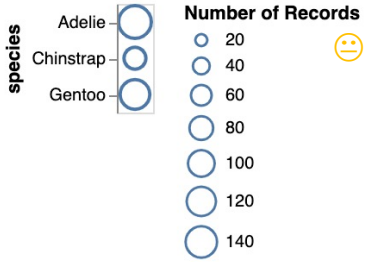
☺



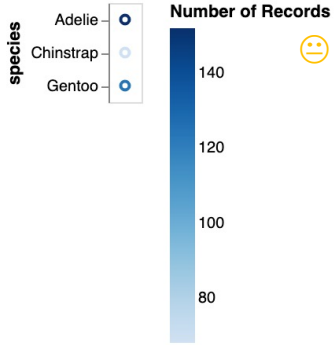
☺

☹ Why connect nominal values?





😐



😐

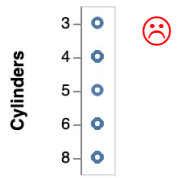
Examples created with https://vega.github.io/voyager/

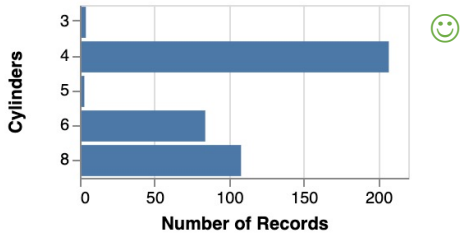
1D Ordinal

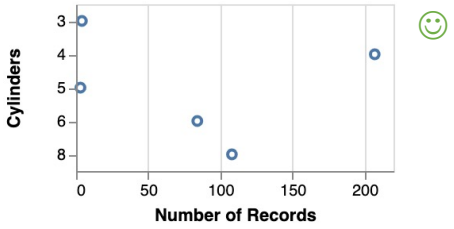
When you are interested in a single column containing ordinal values (i.e. counting and ranking are allowed)
E.g. **# of cylinders** column of the car dataset

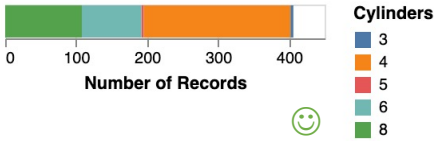
Little can be done with the raw data

After aggregation (i.e. frequency counting), you can do basic visual analysis

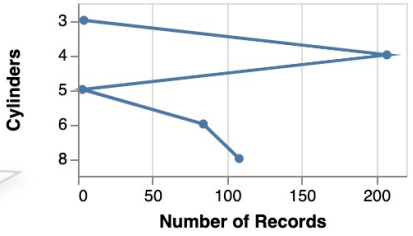


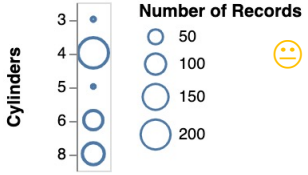


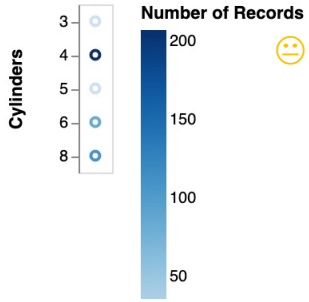




☺ Line chart makes sense as # cylinders have ordinal relationship



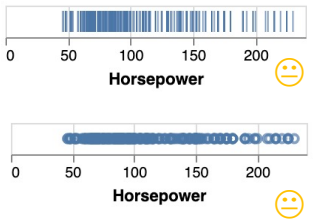




1D Quantitative

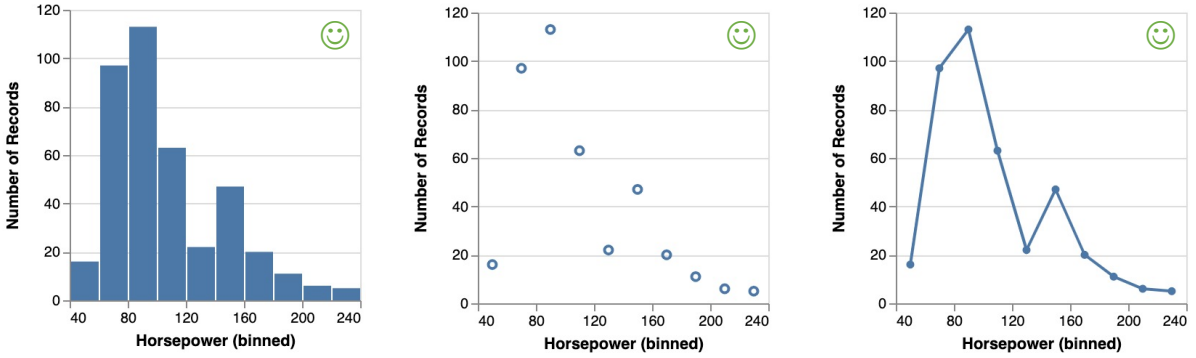
When you are interested in a single column containing quantitative (interval or ratio) values (i.e. numerical operations are allowed).
E.g. **horsepower** column of the car dataset

Little can be done with the raw data

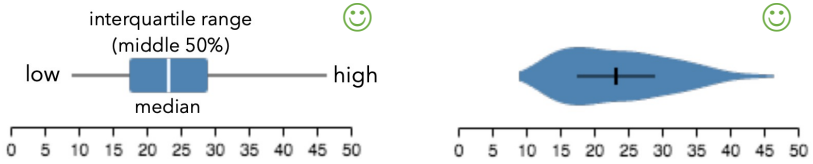


After binning into histogram, it makes much better sense

Bar chart, scatterplot, line chart all make sense



You can draw the distribution via descriptive statistics



(copied from) https://courses.cs.washington.edu/courses/cse442/20au/lectures/CSE442-VisualEncoding.pdf

Summary of 1D charts

Aggregation is the key to draw meaningful charts from 1D

- **Frequency counting** for nominals and ordinals
- **Binning** (to get histogram) or **Descriptive Statistics** (to get distribution) for quantitative values

EDA (Exploratory Data Analytics) begins with 1D charts

- Suitable for finding **outliers** or **incomplete** values
- Suitable for knowing **distribution** (mean, median, min, max)

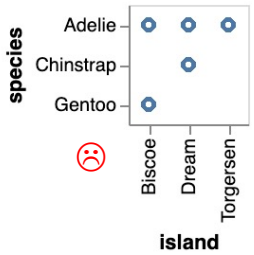
Once you found an interesting column(s), quickly move on to 2D

- If 1D is not interesting, adding another column in 2D is unlikely to be interesting
- Trial-and-errors of finding an interesting pair of columns is the core activity of EDA

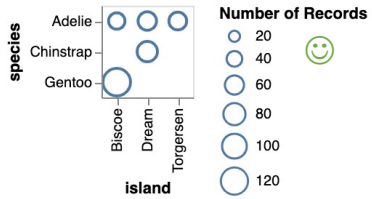
2D Nominal x Nominal

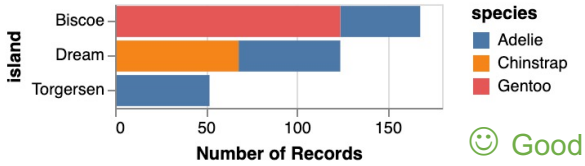
If you are interested in how two nominal columns are correlated
E.g. **species** and **island** columns of the penguin dataset

Little can be done with the raw data

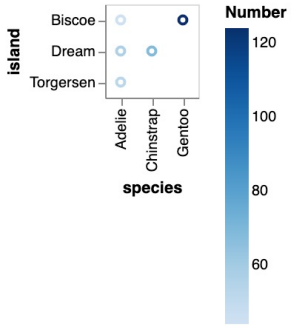


After aggregation (i.e. frequency counting), you can do basic visual analysis





☺ Good for accurate comparison of frequencies

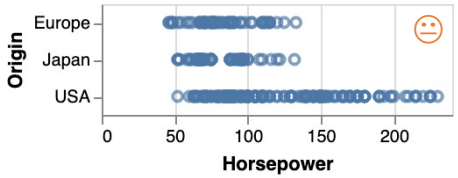


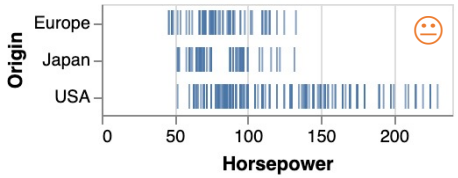
☺ Good for checking existence and rough comparison of frequency with small space

2D Nominal x Quantitative

If you are interested in how one nominal and one quantitative columns
E.g. **origin** and **horsepower** columns of the car dataset

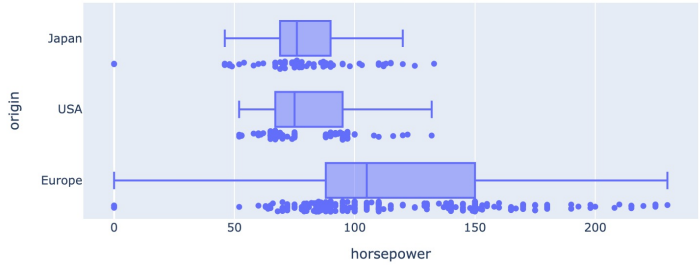
Raw data is already useful





A typical use case of scatterplot.
However, we can do better by applying descriptive stats on horsepower

With statistical summary on quantitative values, we can draw box, violin plots

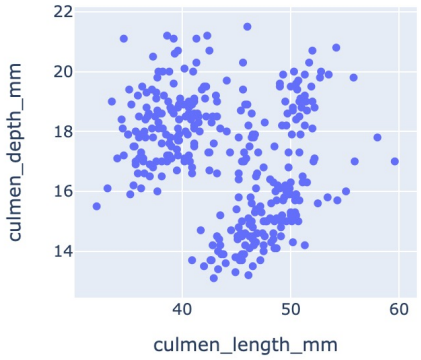


☺ We can compare distributions!

2D Quantitative x Quantitative

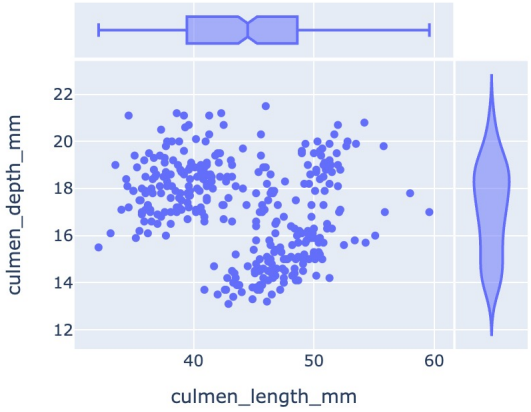
If you are interested in how two quantitative columns
E.g. **body mass** and **culmen length** columns of the penguin dataset

Raw data is already useful



☺ Correlation between two quantitative columns are clearly visible

Statistical summary can be drawn as marginal distribution option of Plotly

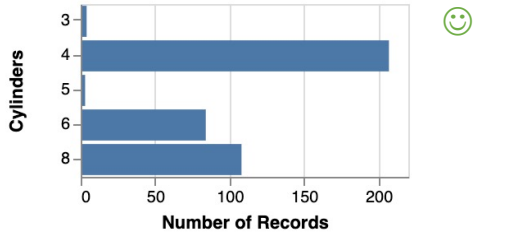


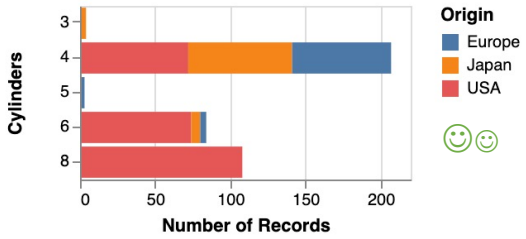
☺☺

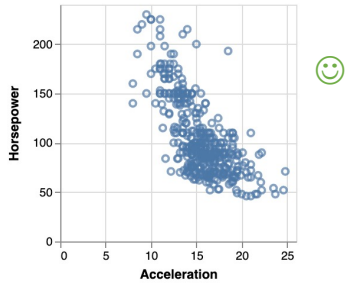
```
import plotly.express as px
fig = px.scatter(colDict, x="culmen_length_mm", y="culmen_depth_mm",
                 marginal_x="box", marginal_y="violin")
fig.show()
```

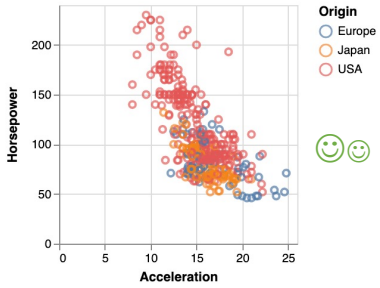
3D ANY

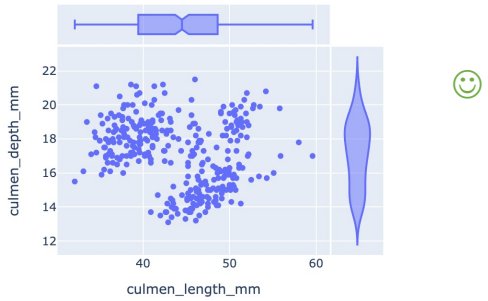
Each visualization can accommodate 1-2 extra columns with color or size encodings. Why not explore higher-dimensions?

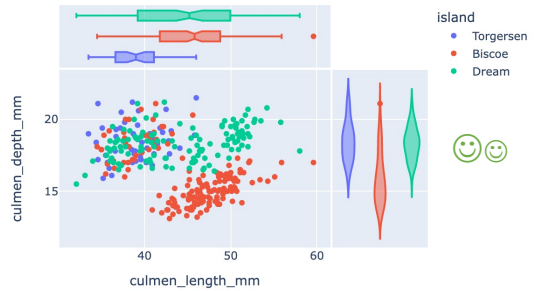






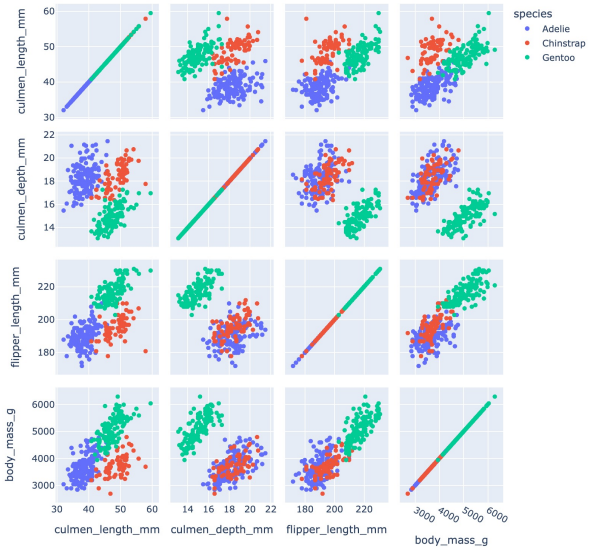




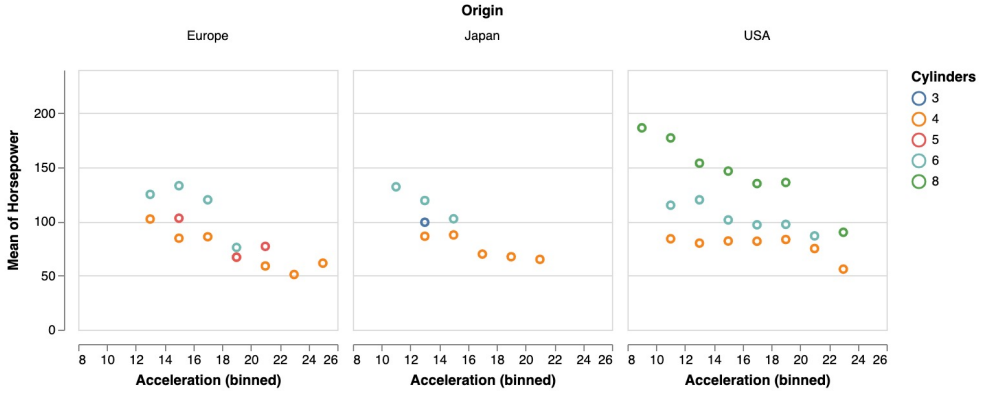


Higher Dimension

Single charts usually cannot accommodate larger than 5 dimensions. However, we can use composite charts. For example, we have used scatterplot matrix in the previous tutorial.



☺☺☺ Scatterplot matrix provides a good overview of five columns



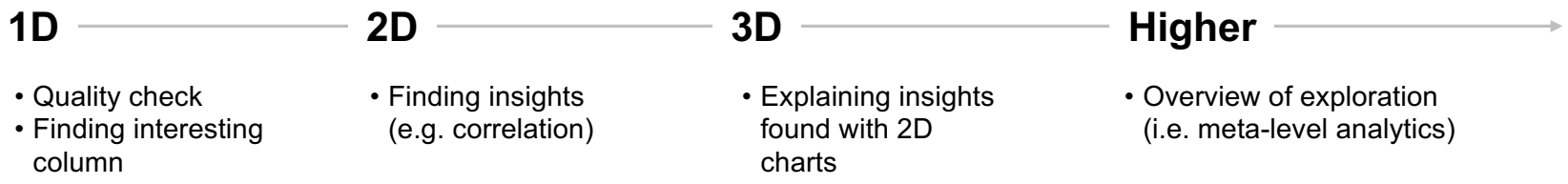
☺☺☺ Using subplots we can add another field

EDA Progression in general

Why did we learn 1D, 2D, 3D, and higher? It seems that higher dimensions are better.

1. **Data exploration usually starts with 1D** for…
   - Checking data quality of each column
   - Finding interesting column for further exploration

2. **# combinations grows very quickly** for higher dimensions
   E.g. If a dataset has 10 columns, there are 1000 combinations for 3D charts. Thus we need to narrow down columns to explore through 1D and 2D

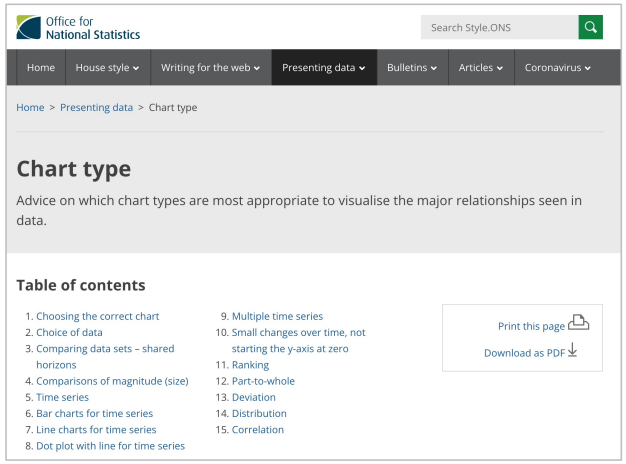


DO and DON’T

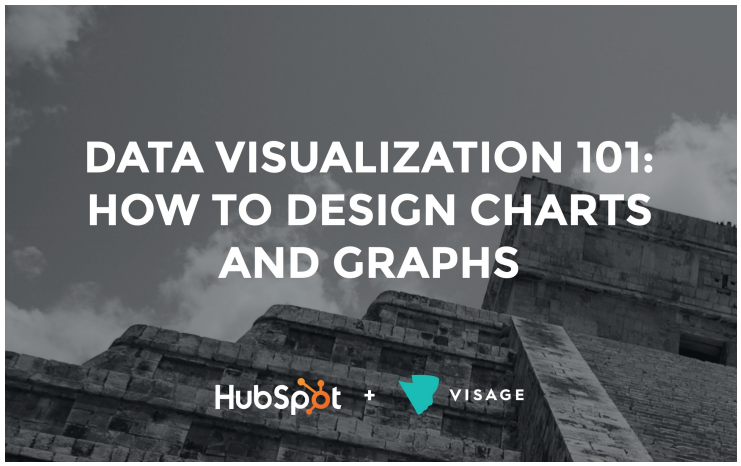
This is a good thing about data visualization. You can logically evaluate a specific visualization for a given dataset and an analytic goal. Let’s learn from good and bad practices.

Resources

I have collected the cases of good and bad visualizations from the following resources.



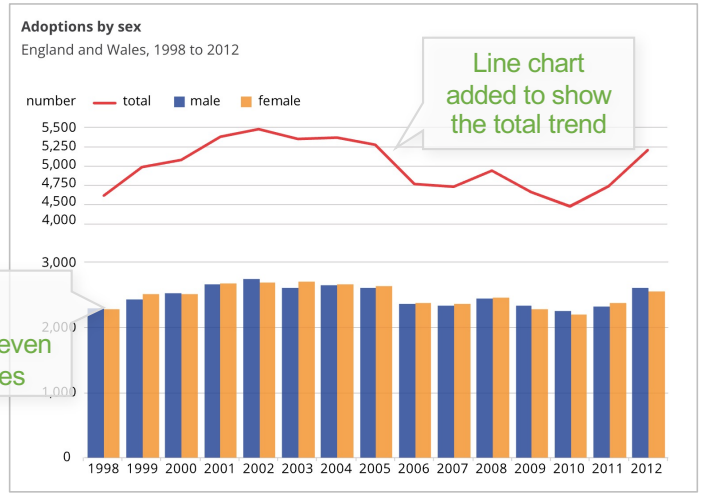
https://style.ons.gov.uk/category/data-visualisation/chart-type/#:~:text=Use%20bar%20charts%20to%20show,example%2C%20age%20or%20time).



https://cdn2.hubspot.net/hub/53/file-863940581-pdf/Data_Visualization_101_How_to_Design_Charts_and_Graphs.pdf

Place things to compare side-by-side

GOOD



☺ I can compare even tiny differences

Grouped Bar Chart

BAD

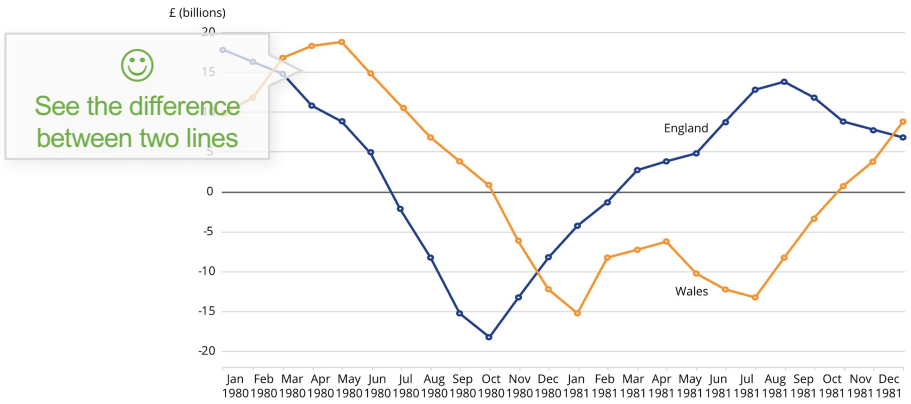
☹ Is the line at the exact middle?

Stacked Bar Chart

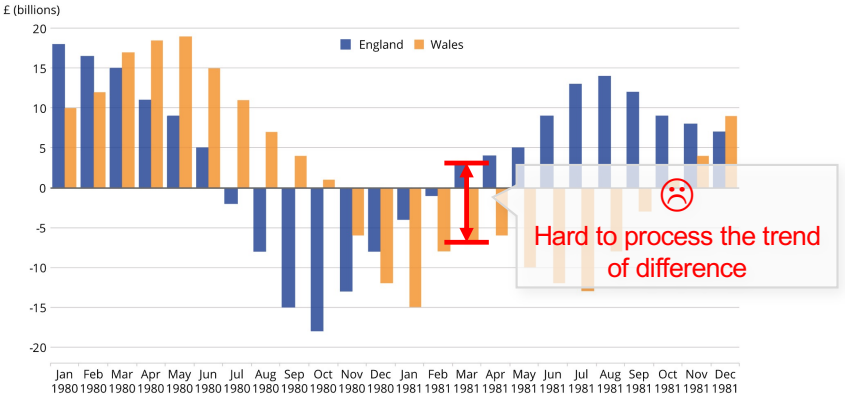
Minimize viewer's effort to catch the main signal

If the task is to see the **trend of difference** between two quantitative values over time

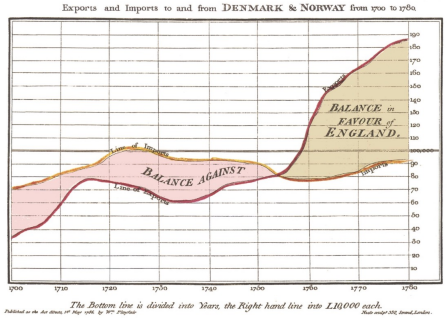
GOOD



BAD



BETTER



William Playfair's famous line chart

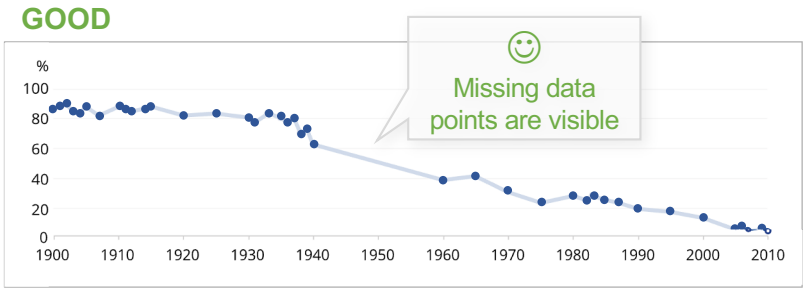
TAKEAWAY.

- Different analytic tasks might require different chart types for the same data
- Minimize the viewer's effort to catch the main signal

Beware of the common assumption of the chart type

Line chart assumes that data points are evenly distributed

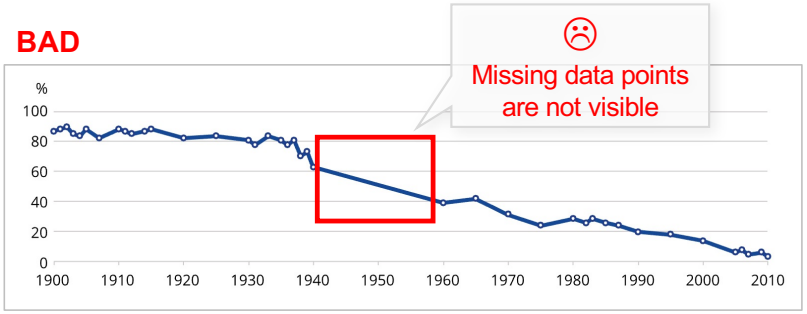
GOOD



Scatter plot + dimmed line chart

: The line implies the chart focuses on timeline trends. However, missing dots warn “This part of the timeline has no data point.”

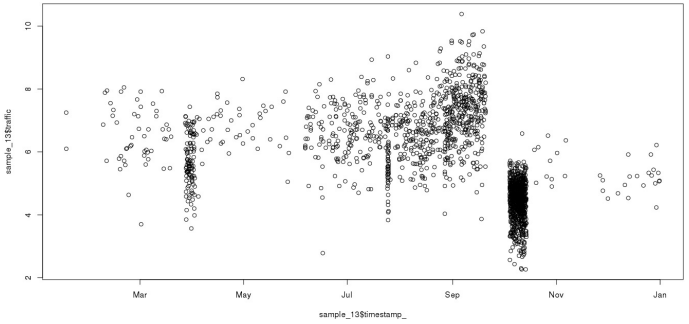
BAD



Simple line chart

: Line charts imply that the axis is regular (i.e. data points are evenly-spaced). Viewers may have a false assumption for unevenly-spaced time series.

An Extreme Case of Unevenly Spaced Time Series (link)

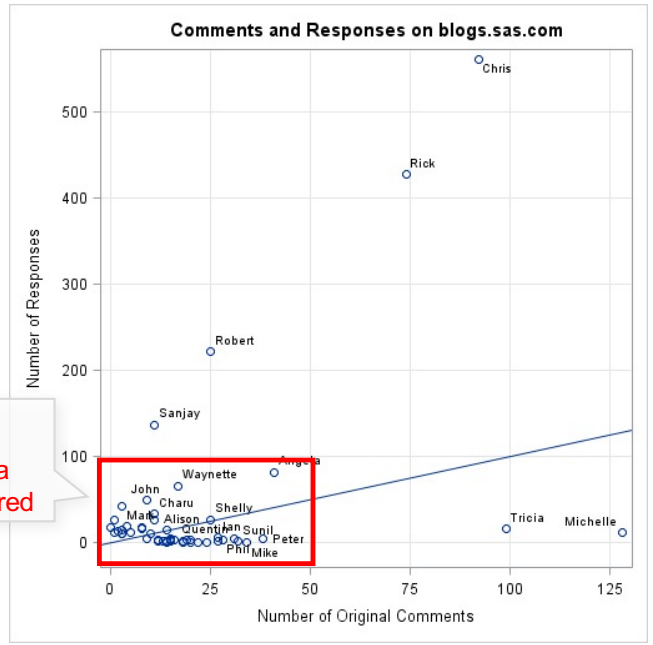


TAKEAWAY.

- Make missing data points visible.
- Beware of the common assumption of the chart type (e.g. line charts look like data points are evenly distributed)
- You can infer missing data points but that’s not part of visualization

Avoid overcrowded markers

BAD



☹ Too many data points are cluttered

GOOD

☺ Most data points are evenly spread

☹ Log axis might be hard to interpret for some audiences

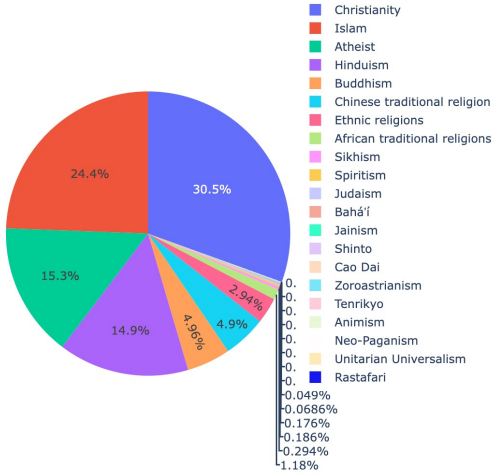
How to visualize Part-to-Whole

ratio
percentage
proportion
share

breakdown
make up
hierarchy

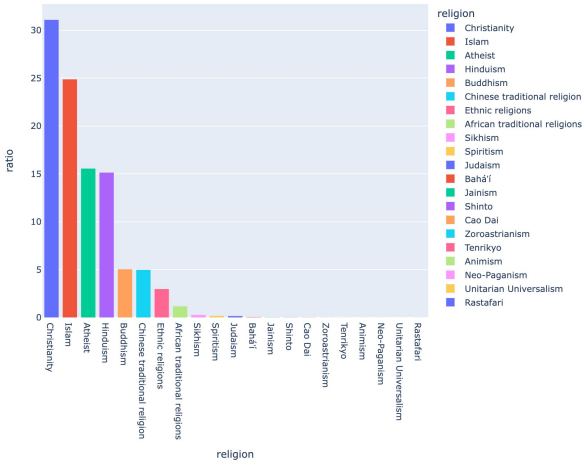
For large # of distinct items

BAD



Pie charts are sensitive to the cardinality (i.e. # of distinct items should not be much larger than six); Also it's not easy to visually compare ratios of items

GOOD



Bar charts are better with large # of distinct items.

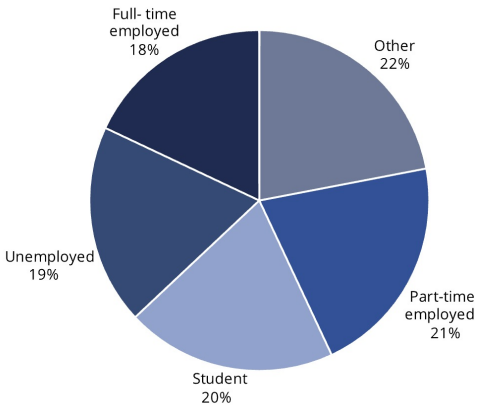
How to visualize Part-to-Whole

ratio
percentage
proportion
share

breakdown
make up
hierarchy

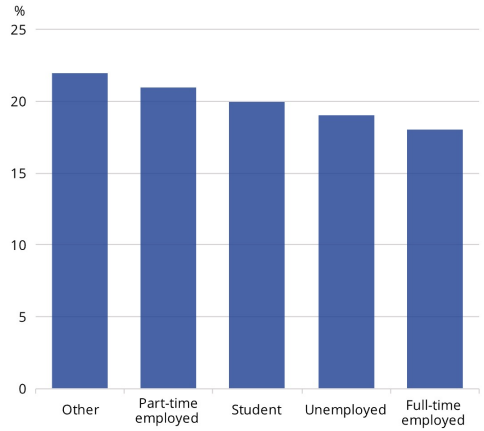
When there's no dominant item,

BAD



Pie charts are bad at comparing similar ratios. Labels are required then.

GOOD



Bar charts are better at comparing similar ratios (i.e. side-by-side comparison is the best)

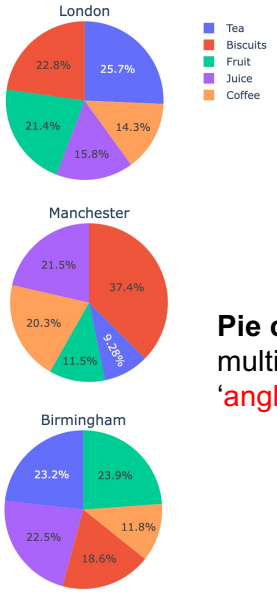
How to visualize Part-to-Whole

ratio
percentage
proportion
share

breakdown
make up
hierarchy

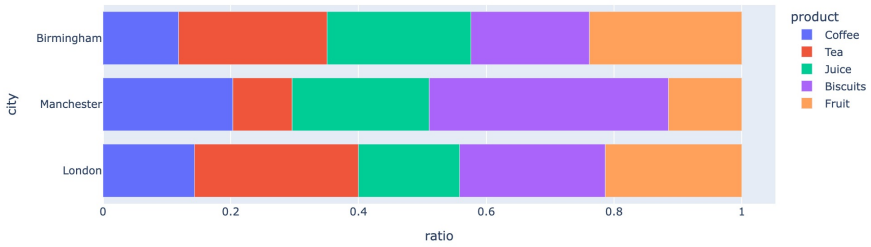
Comparing multiple sets of ratios

BAD



Pie charts are bad at comparing multiple sets of ratios because 'angle' is hard to compare.

GOOD



Stacked bar charts are good at comparing multiple sets of ratios. 'Length' is a good encoding to compare values.

If Pie charts have only bad things, why do people use it a lot?

There could be historical, aesthetic, and other reasons. My favorite explanation is that "*Pie charts are applicable to part-to-whole relationship only*". That means, even before reading text, readers would correctly guess the meaning of pie charts. In contrast, bar charts are effective for a wide range of data and questions – which means viewers have to read title, legends, and other textual information to interpret.

(optional reading) Why human love pie charts

How to visualize survey results on Likert scale

What is Likert scale?

- Commonly used in survey or user study; Captures the feeling of intensity for a given item
- Is Likert scale an **ordinal or interval** measure? Still lots of debates! [1,2] ☹ ☺

Likert is Ordinal ☹

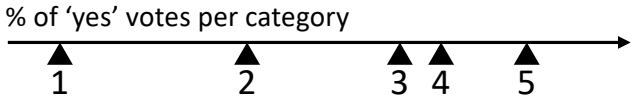
"Participants tend to avoid extreme responses. Thus the categories are not equally distributed."

"In many situations, participants have motivations to lean toward positive / negative responses."

To what extent do you agree with the following statement?

"I will vote for the policy A"

Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
1	2	3	4	5



Likert is Interval ☺

*"A well-designed question (i.e. having a clear neutral position, symmetrical and equidistant categories) can **approximate an interval measurement**, with large N"*

Take-away messages.

- Due to many potential biases, it's safe to treat them as ordinals.
- If you need interval values, carefully frame the question and categories so that they are symmetrical and equidistant.

1. amieson, Susan (2004). "Likert Scales: How to (Ab)use Them". Medical Education. **38** (12): 1217–1218.
2.^ Jump up to:[a] [b] Norman, Geoff (2010). "Likert scales, levels of measurement and the "laws" of statistics". Advances in Health Science Education. **15** (5): 625–632.

How to visualize survey results on Likert scale

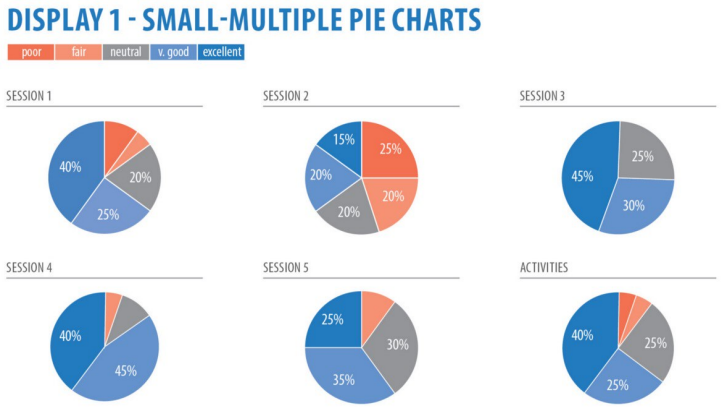
How to visualize Likert values as **Ordinal**

(Recommended Article)
https://medium.com/nightingale/seven-different-ways-to-display-likert-scale-data-d0c1c9a9ad59

Q. How would you rate the quality of the sessions and activities?

Responses

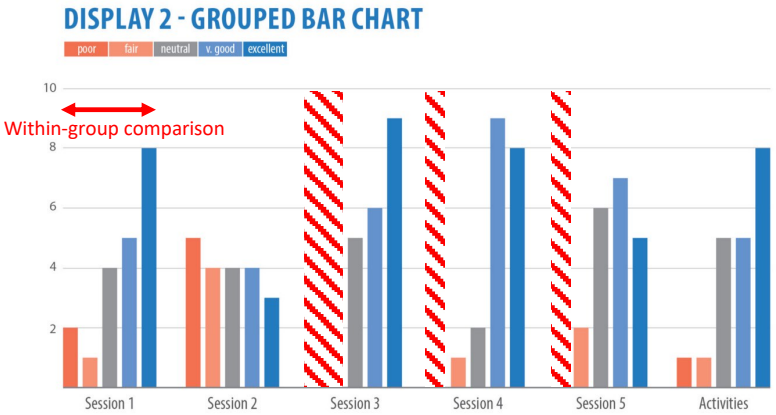
	Poor	Fair	Neutral	Very Good	Excellent
Session 1	2	1	4	5	8
Session 2	5	4	4	4	3
Session 3	0	0	5	6	9
Session 4	0	1	2	9	8
Session 5	0	2	6	7	5
Activities	1	1	5	5	8



☹ **Pie charts** are bad at comparing multiple sets of ratios because 'angle' is hard to compare. Making it even worse, pies are cyclic ('Poor' and 'Excellent' are placed right next to each other)
☺ No significant benefits of using pie charts

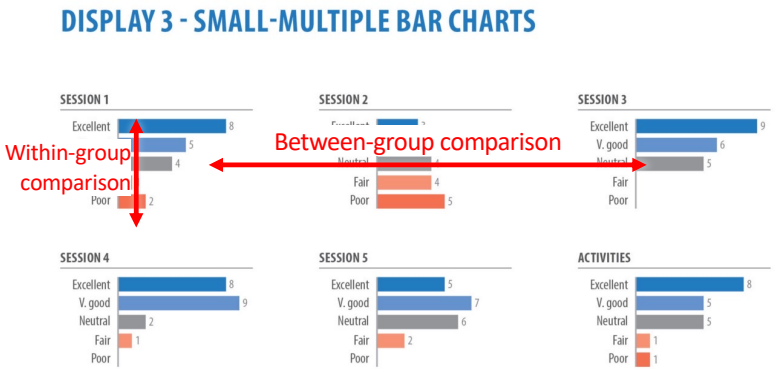
How to visualize survey results on Likert scale

How to visualize Likert values as **Ordinal**



☹ **Grouped bar charts** cannot represent categories with zero responses. Hard to compare between groups

☺ Easy to interpret distribution within a group

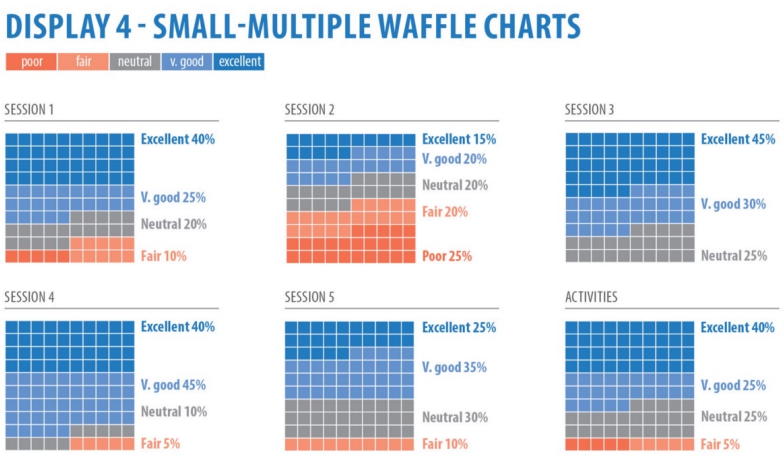


☹ Bar charts are very generic. Viewers have to interpret the meaning (“this is Part-to-Whole information of a Likert scale”) from text.

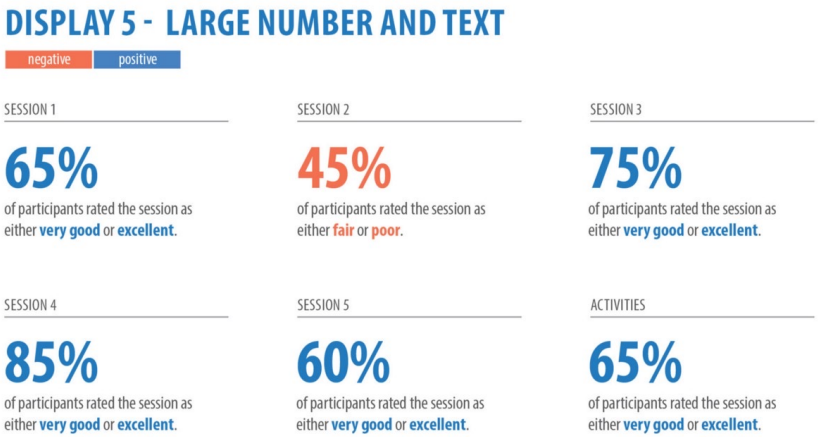
☺ Both between-group and within-group comparisons are well supported

How to visualize survey results on Likert scale

How to visualize Likert values as **Ordinal**



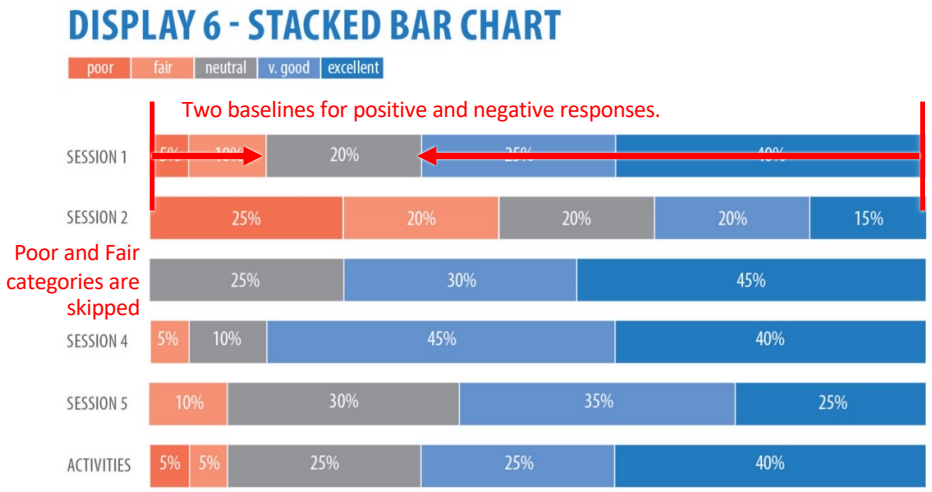
☺ Very Fancy. Each waffle is clearly part-to-whole information. (a perfect alternative of pie charts?!)
☹ Within-group or Between-group comparisons are not as easy as with bar charts; Empty categories are skipped (same with the group bar charts)



☺ Communicate specific insights with minimal cost; Useful as a summary of other charts
☹ Subjective; Huge information loss; Not allowing viewers to find their own insights

How to visualize survey results on Likert scale

How to visualize Likert values as **Ordinal**

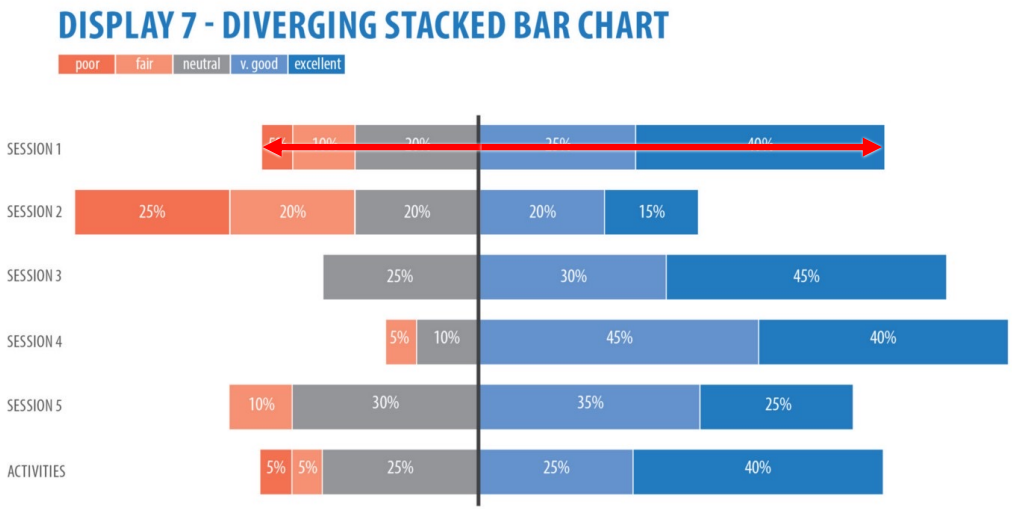


☺ Efficient use of screen real-estate (i.e. can show many sessions in small space); Both between and within-group comparison are well-supported; Since bars have the same total height, it's clear that they show part-to-whole information; Category ordering is maintained.

☹ Empty categories are skipped; There are two baselines for positive and negative categories (i.e. hard to compare).

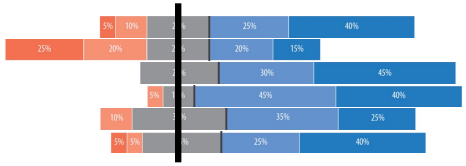
How to visualize survey results on Likert scale

How to visualize Likert values as **Ordinal**



☺ Aligning bars at the common baseline between neutral and v.good; Optimized for comparing positive / negative categories both within and between-groups

☹ Almost perfect. But why not align at the center of neutral?



Summary

- There is no silver-bullet chart. Each chart design has pros and cons.
  - Match with what analytic tasks should be supported / what message should be communicated
- Bar chart variations are very versatile
  - Grouped, Stacked, and Aligned bar charts can cover most analytic tasks
  - “Being widely used” makes bar charts hard to guess its meaning (as opposite to pie charts).
- Use multiple charts in combination for products or storytelling
  - E.g. coordinated charts; dashboard; data story; infographics, etc
  - We will learn them soonish…

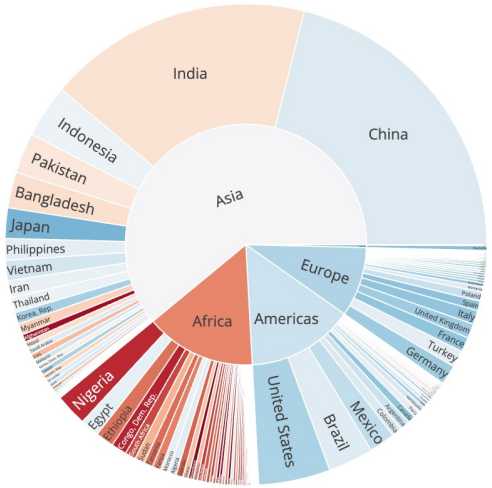
Advanced Charts

Take a quick look of advanced charts. While looking fancier, advanced charts are harder to interpret and suitable for very specific use cases.

Charts for **Hierarchical Data**

Useful for showing part-to-whole information with multiple depths

Sunburst



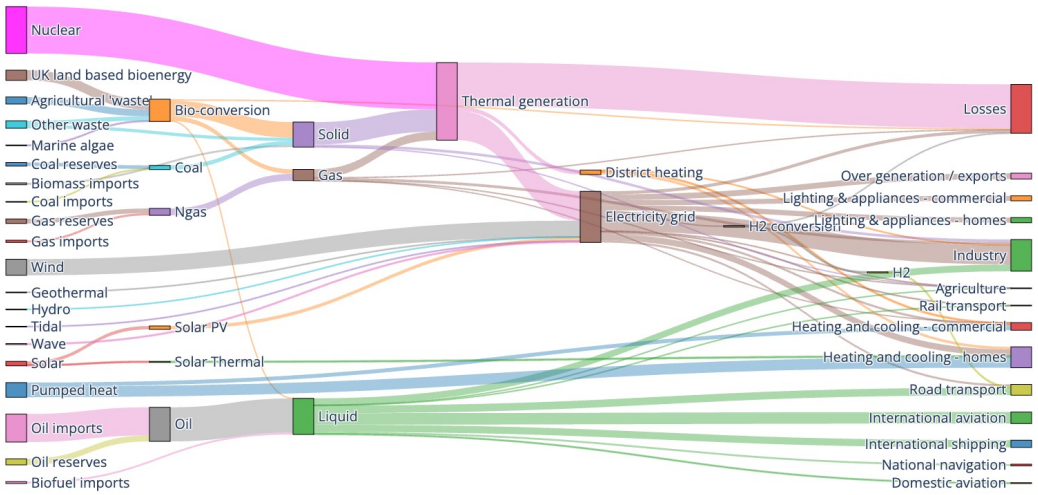
Treemap

Charts for **Flow**

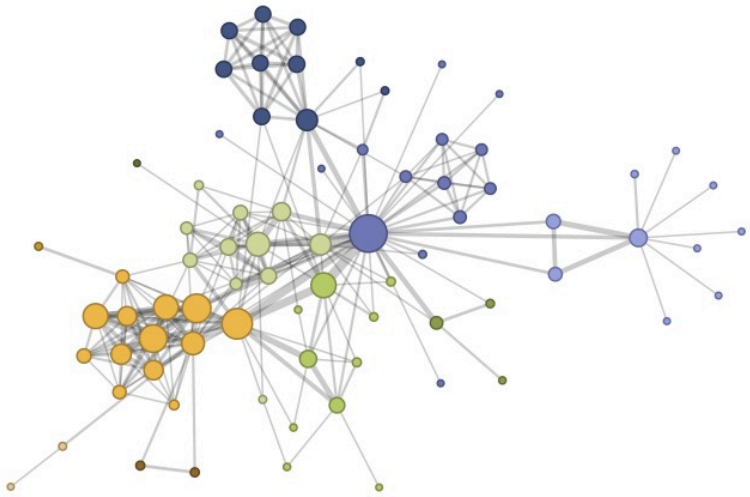
Useful for showing relationship between many states. For instance, customer journey (or funnel) can be drawn with Sankey or other charts for flow

Sankey diagram

Energy forecast for 2050
Source: Department of Energy & Climate Change, Tom Counsell via Mike Bostock

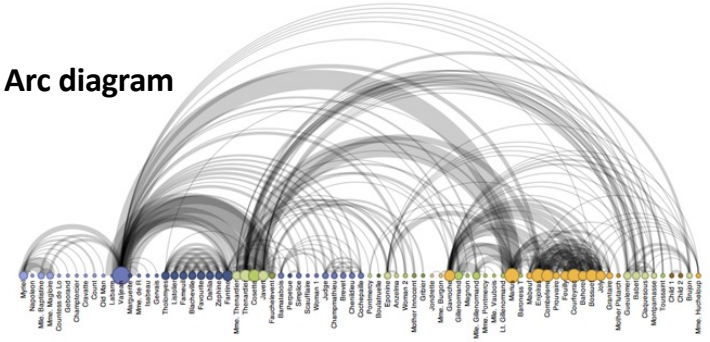


Charts for **Network**

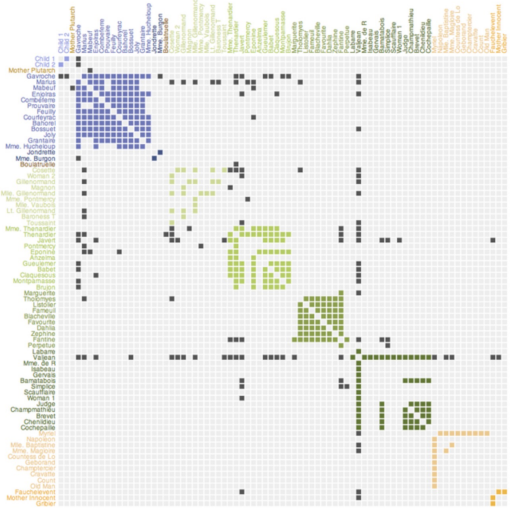


Node-link diagram with force-directed layout

Arc diagram



Adjacency matrix



Charts for **Text**

☺ Nice looking overview of tags and keywords; We will learn along with other NLP techniques

☺ Hard to control what viewers will perceive from a word cloud; Preprocessing is more important (and harder) than actual drawing.

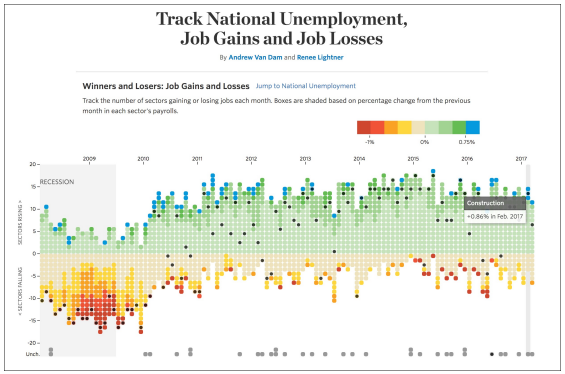
Interactivity

To overcome various limitations of single static visualizations, data visualization provides a variety of interactivity such as zoom, filter, brush, search, scale, get details, and more.

Ben Shneiderman's Mantra for Visual Information Seeking

1. Overview first 2. Zoom and Filter 3. Detail-on-demand

The initial view of a **dashboard** provides the data overview



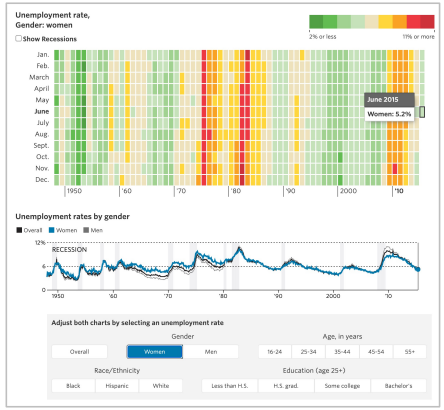
Data journalism article begins with an overview that introduces the topic and the dataset

The initial view of **an AI-supported visualization tool** (voyager) begins 1D charts per columns (i.e. **univariate** analysis)

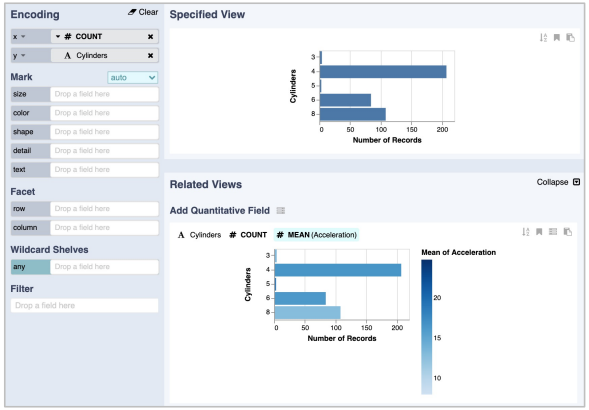
Ben Shneiderman's Mantra for Visual Information Seeking

1. Overview first
2. Zoom and Filter
3. Detail-on-demand

Brushing & Linking
Multiple charts in a dashboard are connected with each other. Brushing (i.e. selecting items) in one chart filter the same items in other charts.



Filtering UI
Data journalism articles often provide filtering UI for readers to find an interesting part of the dataset, and go deeper.

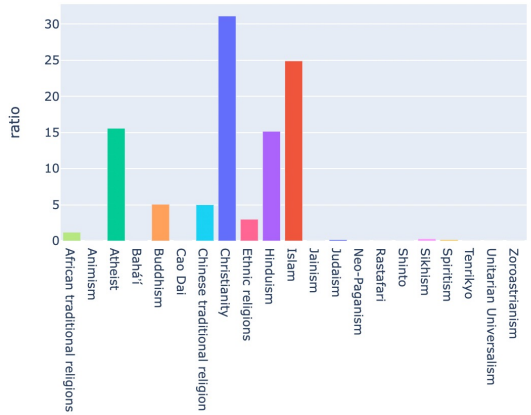


2D, 3D, and higher-dimension charts
After finding interesting columns, analysts would combine them to narrow down the scope and to find more specific insights (i.e. **multi-variate** analysis)

Ben Shneiderman's Mantra for Visual Information Seeking

1. Overview first
2. Zoom and Filter
3. Detail-on-demand

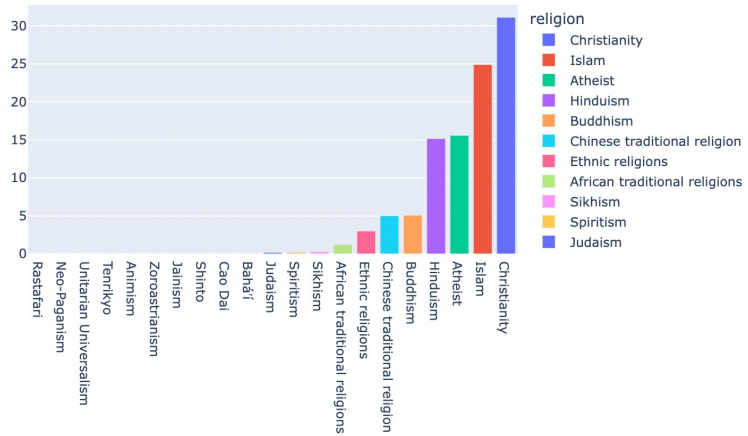
Sorted by Category (A->Z)



Switch Sorting Mechanism

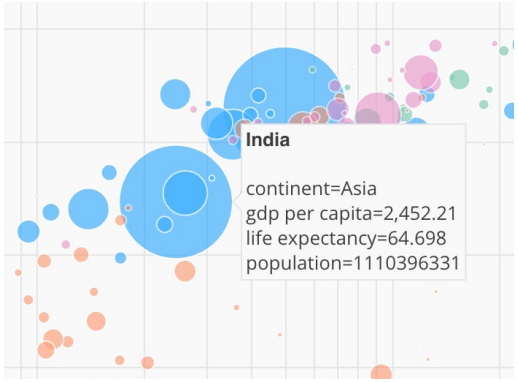
Different sorting methods helps viewers focus on different patterns and items. Thus sorting is a variation of zoom and filtering process

Sorted by Ratio (small->large)

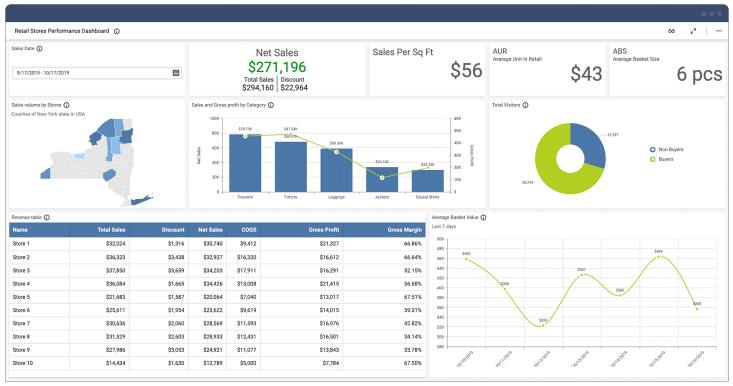


Ben Shneiderman's Mantra for Visual Information Seeking

1. Overview first
2. Zoom and Filter
3. **Detail-on-demand**



Mouse-over Tooltip
Most EDA tools allow viewers to see raw data of individual items as tool-tip



Data table filtered by user selection
Most EDA tools offer data tables that shows raw data of currently selected items

Summary of Interactivity [from Heer's Lecture Note]

Most visualizations are interactive

Even passive media elicit interactions

Good visualizations are task dependent

Pick the right interaction technique

Consider the semantics of the data domain

Fundamental interaction techniques

Selection / Annotation, Sorting, Navigation, Brushing & Linking, Dynamic Queries

End of Lecture 6