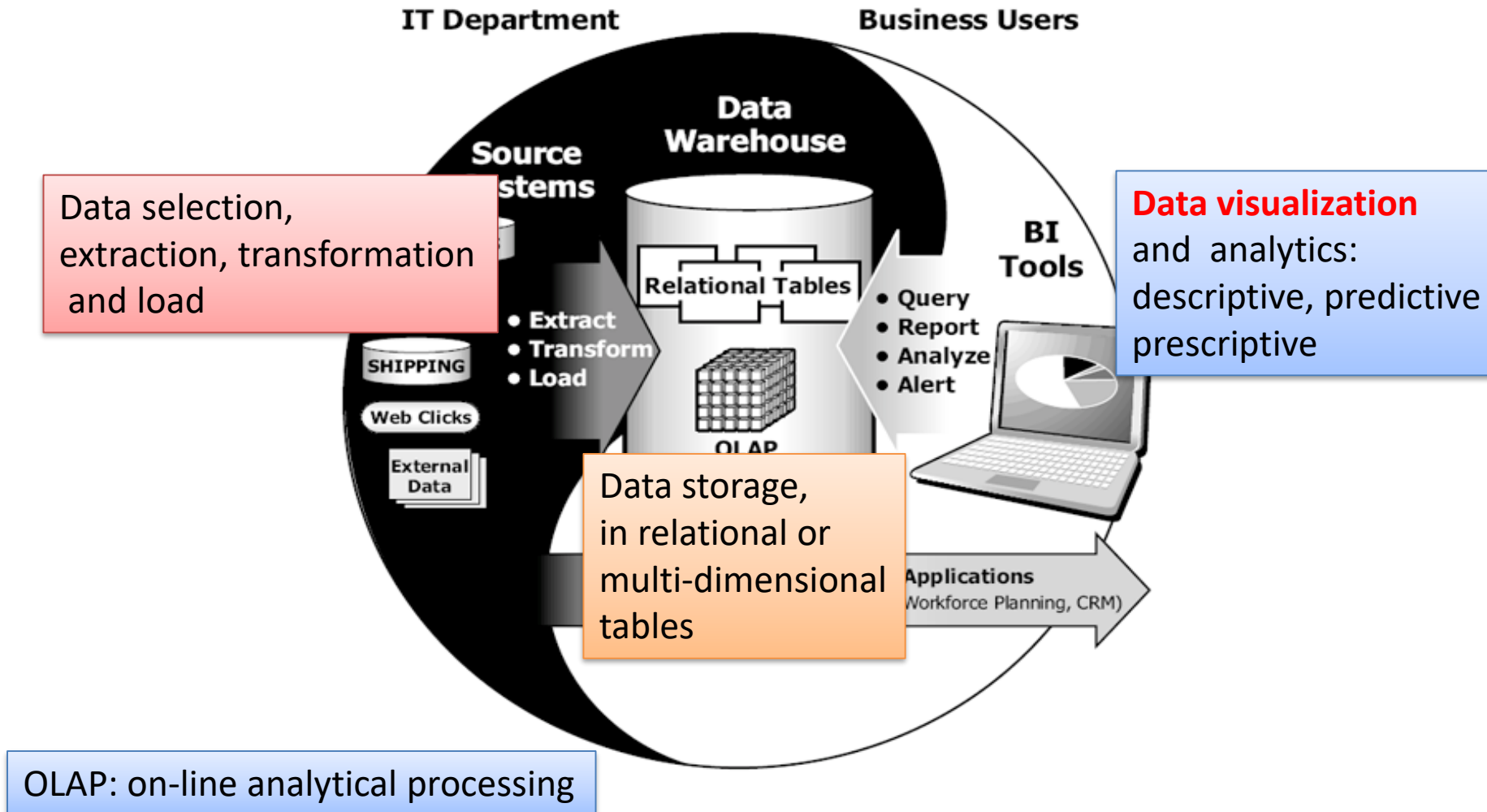


DATA VISUALIZATION



Architecture of a BI system



Outline


What is Visualization?

Why Visualize?

Data → Visualization

How to tell if a visualization is
“appropriate”

Guidelines



Visualization: what is?

- Visual **representations** of data that reinforce human **cognition**

Perhaps a
more helpful
question:

What are some ways
a “visualization” can
be **useful**?

But first off: Value of Visualization

- Provably **much better than written reports**, since:
- Reduce Memory Load
 - Working memory is limited
 - Offload storage/organization to the diagram
- Reduce Search Time
 - Pre-attentive (constant-time) search
 - Spatially-indexed patterns store the “facts”
- Enable Perceptual Inference
 - Map inference to pattern finding

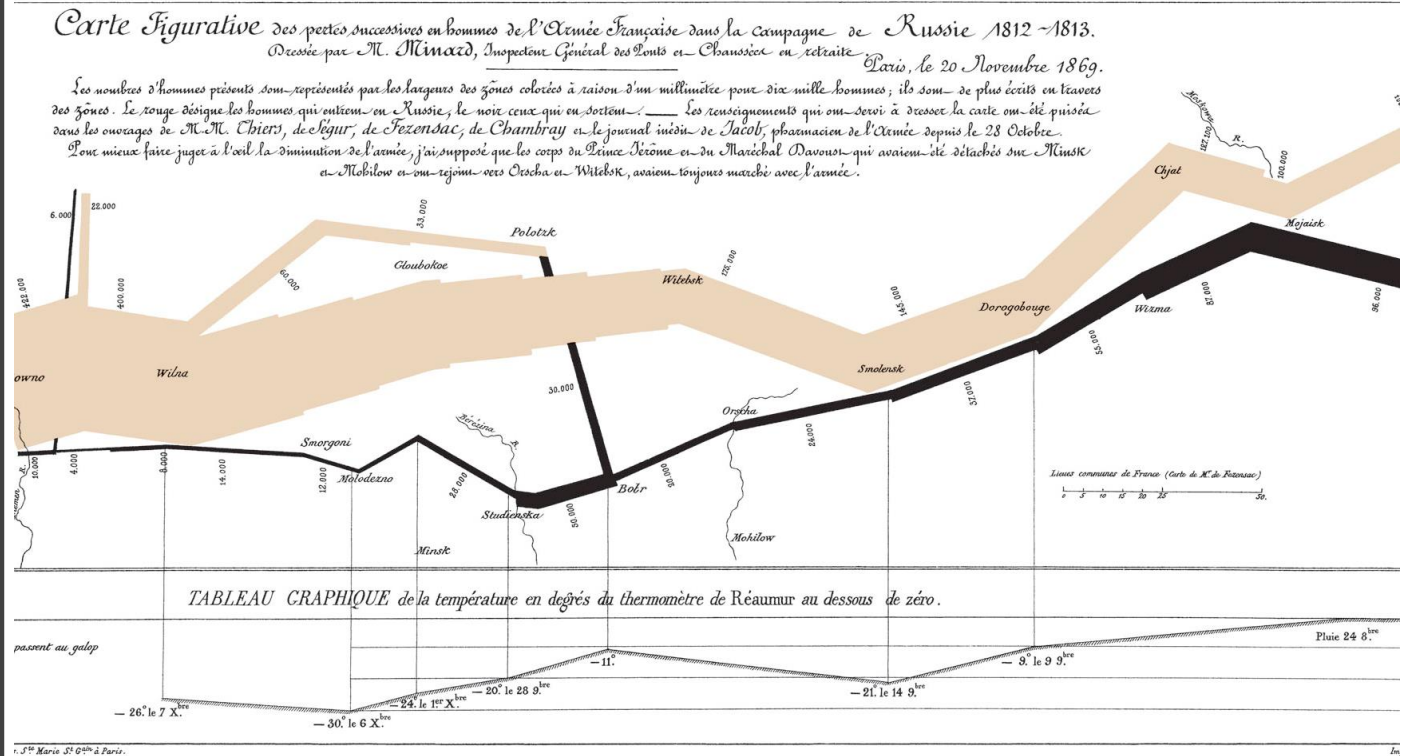
Some very “old” examples: cholera outbreak in 1854

- In 1854, cholera broke out in London
 - 127 people died
 - 616 people became ill
- Initial explanation: “Miasma”
- Dr. John Snow identified the outbreak
- How did he do it?
 - Talked to people
 - Identified cases
 - Used maps
 - Convinced others



Another old example: 1869

- “Figurative Map of the successive losses in men of the French Army in the Russian campaign 1812-1813.”
- Drawn by Mr. Charles Minard, Inspector General of Bridges and Roads in retirement.
- Paris, 20 November 1869.



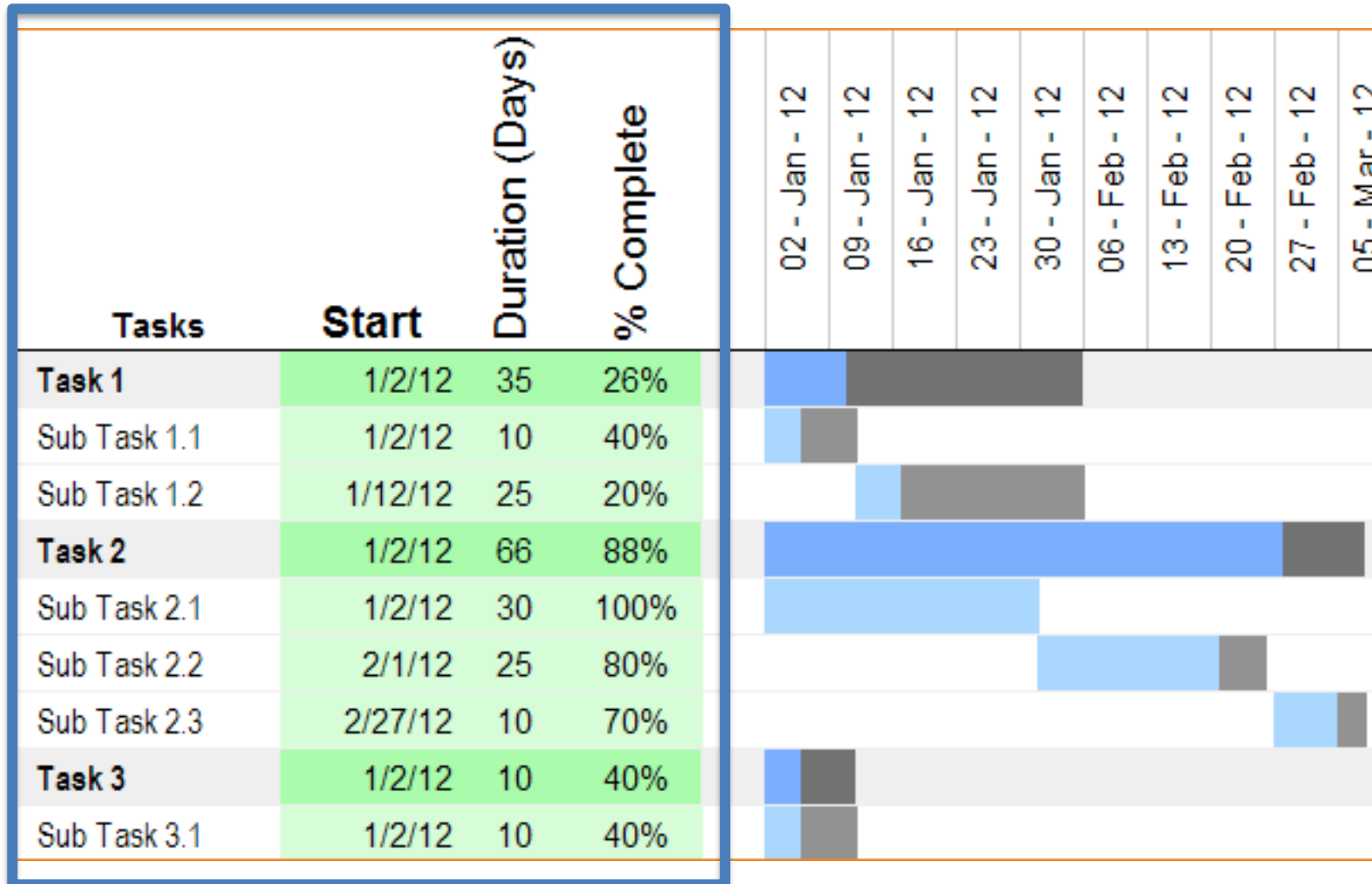
Dimensions: temperature, route of the troup, human losses,
 Directions (black ←, red →)

Minard's multi- dimensional map

- Minard was a pioneer of the use of graphics in engineering and statistics.
- He is most well known for his cartographic depiction of numerical data on a map of Napoleon's disastrous losses suffered during the Russian campaign of 1812
- The illustration depicts Napoleon's army departing the Polish-Russian border. A thick band illustrates the size of his army at specific geographic points during their advance and retreat.
- It displays **six types of data in two dimensions**:
 - the number of Napoleon's troops;
 - the distance traveled;
 - temperature;
 - latitude and longitude;
 - direction of travel;
 - and location relative to specific dates without making mention of Napoleon;

So when vizualizations are useful?

Example: Traditional excel table vrs “fancy” visualization





Useful=help
us
understanding
the data

- What types of data?
- What kind of “understanding” we want to convey?
- How to connect data with (good) visualizations?



Types of data

- Data can be classified in three groups
- Qualitative (Attributes)
 1. Nominal
 2. Ordinal
- Quantitative (Metrics)
 3. Numeric

Qualitative:Nominal data

- Data that be counted, but **not** ordered or aggregated.
- Examples:
 - Products – Books, Movies, Music
 - Gender – Male, Female
 - State – Virginia, Nevada, California



Qualitative: Ordinal data

- Data that can be counted and **ordered**, but not aggregated.
- Examples:
 - Date – 1/1/2014, 1/2/2014...
 - Grades – A, B, C...
 - Ranks – Like, Neutral, Dislike

Metrics

- Quantitative data that can be **counted, ordered, and aggregated**.
- Examples:
 - Revenue, Cost, Profit
 - Number of Customers
 - Temperature
 - Time



Ordinal Attributes and Metrics

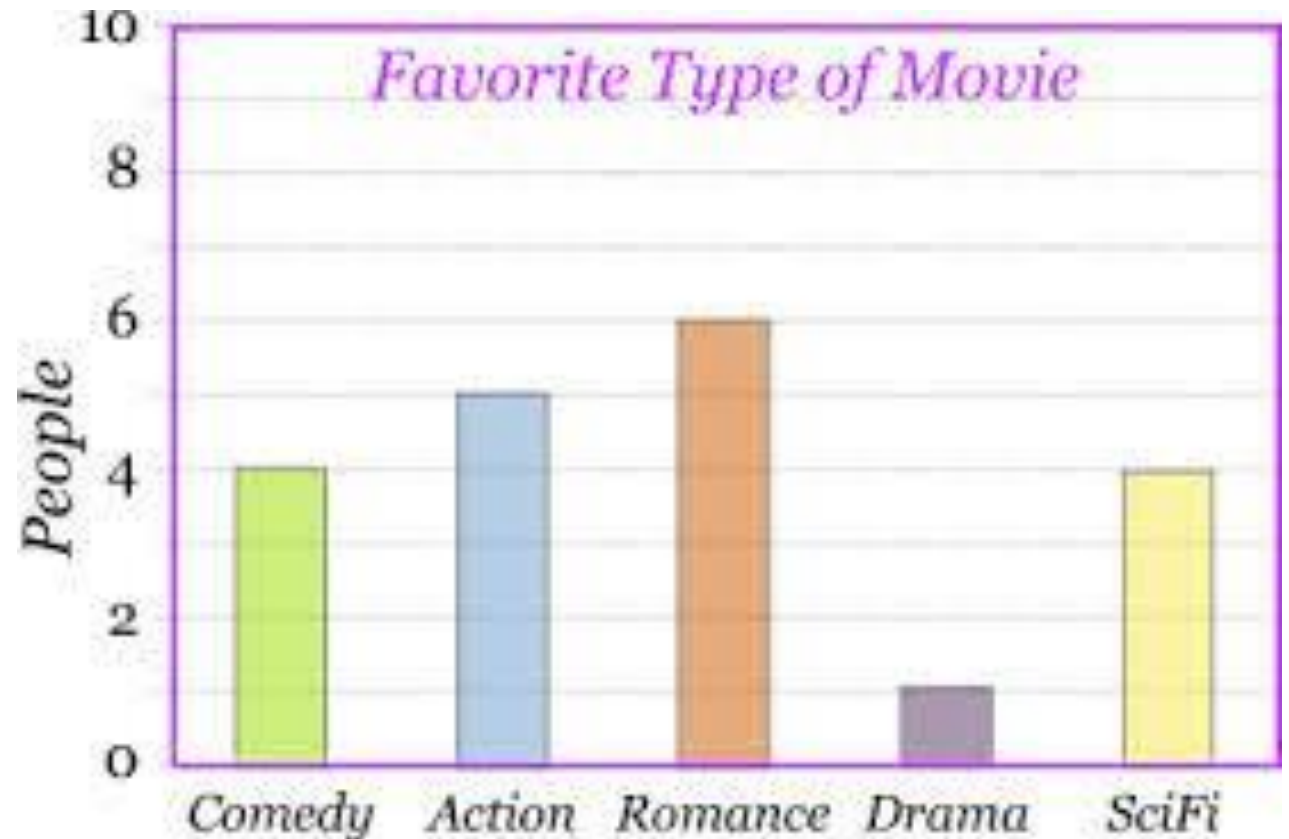
- Some data can be used as **either ordinal or metrics**. Their classification is dependent on usage.
- Examples:
 - Age
 - Scores

Types of visualizations

- Bar charts (histograms)
- Line charts
- Scatterplots
- Maps
- Pie Charts
- Network (graphs)
- ...many others (will see creative examples)

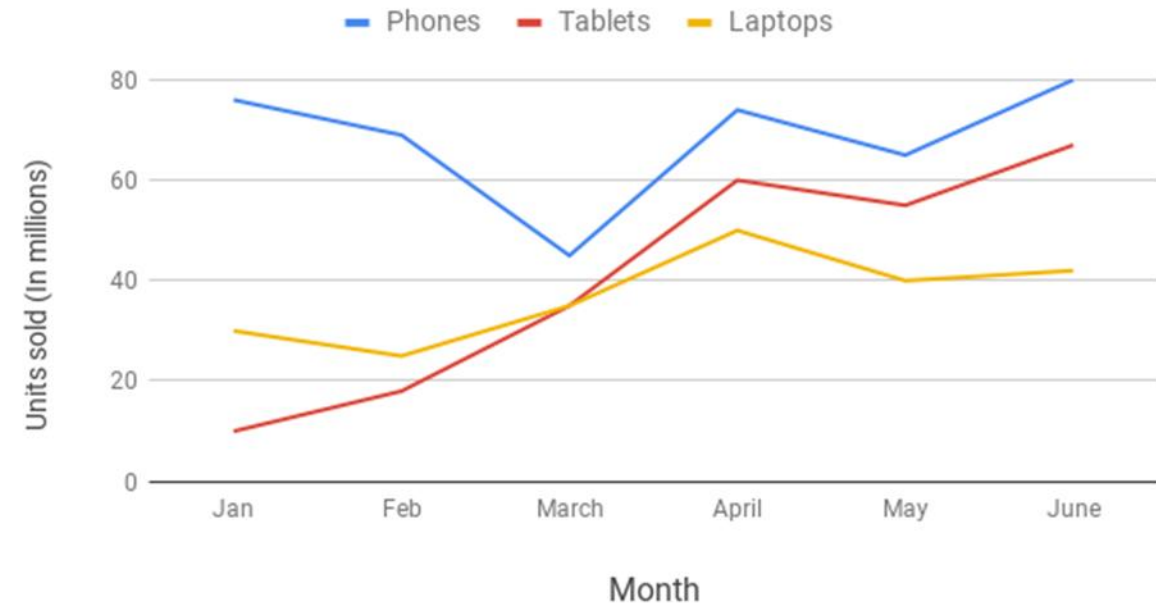
Bar Charts (histograms)

- Vertical bar charts are useful to **compare** different attributes of type **nominal** (categorical or discrete), such as age groups, classes, schools, etc., as long as there are not too many categories to compare.
- X represents the nominal variable, y is a metrics (e.g., number of people that likes Comedy)



Line charts

Device sales

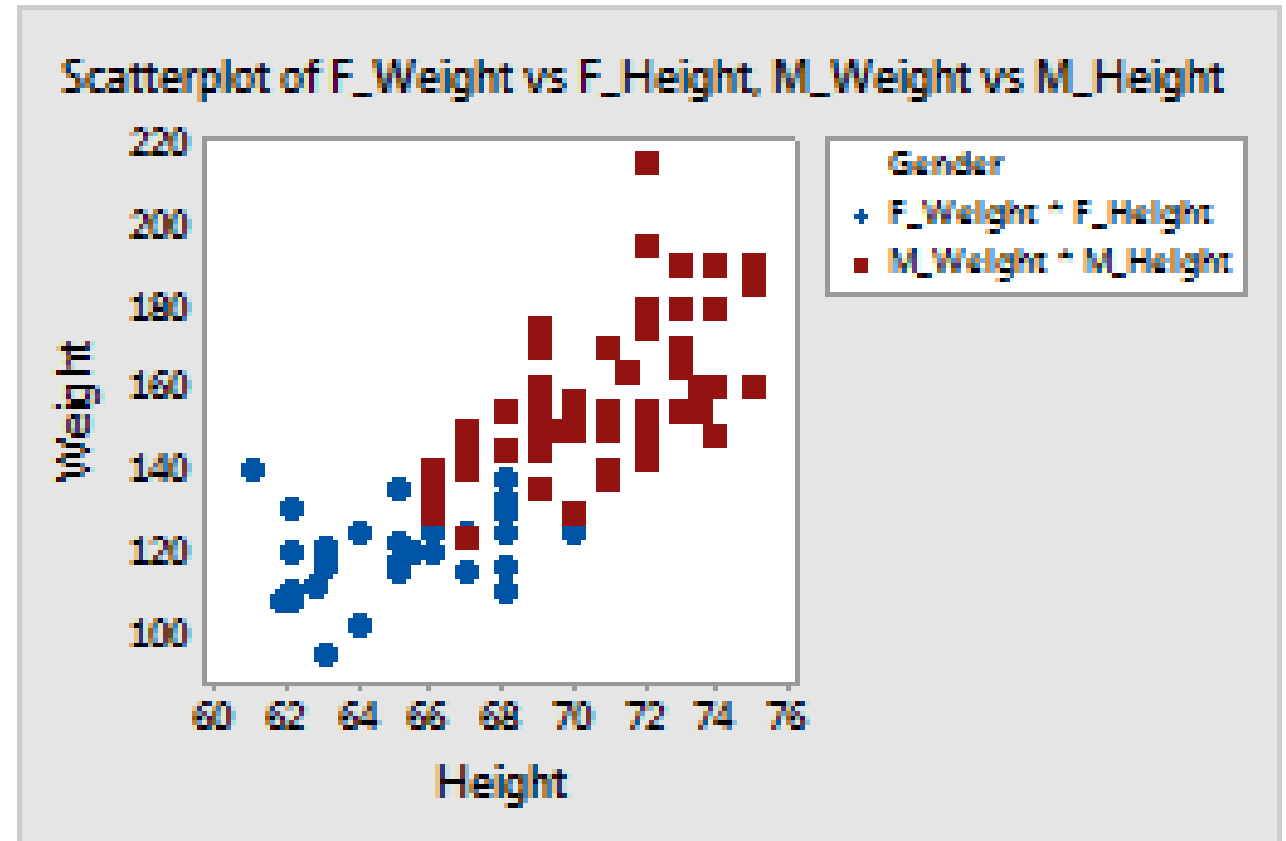


Line charts show *trends* of numerical data (metrics)

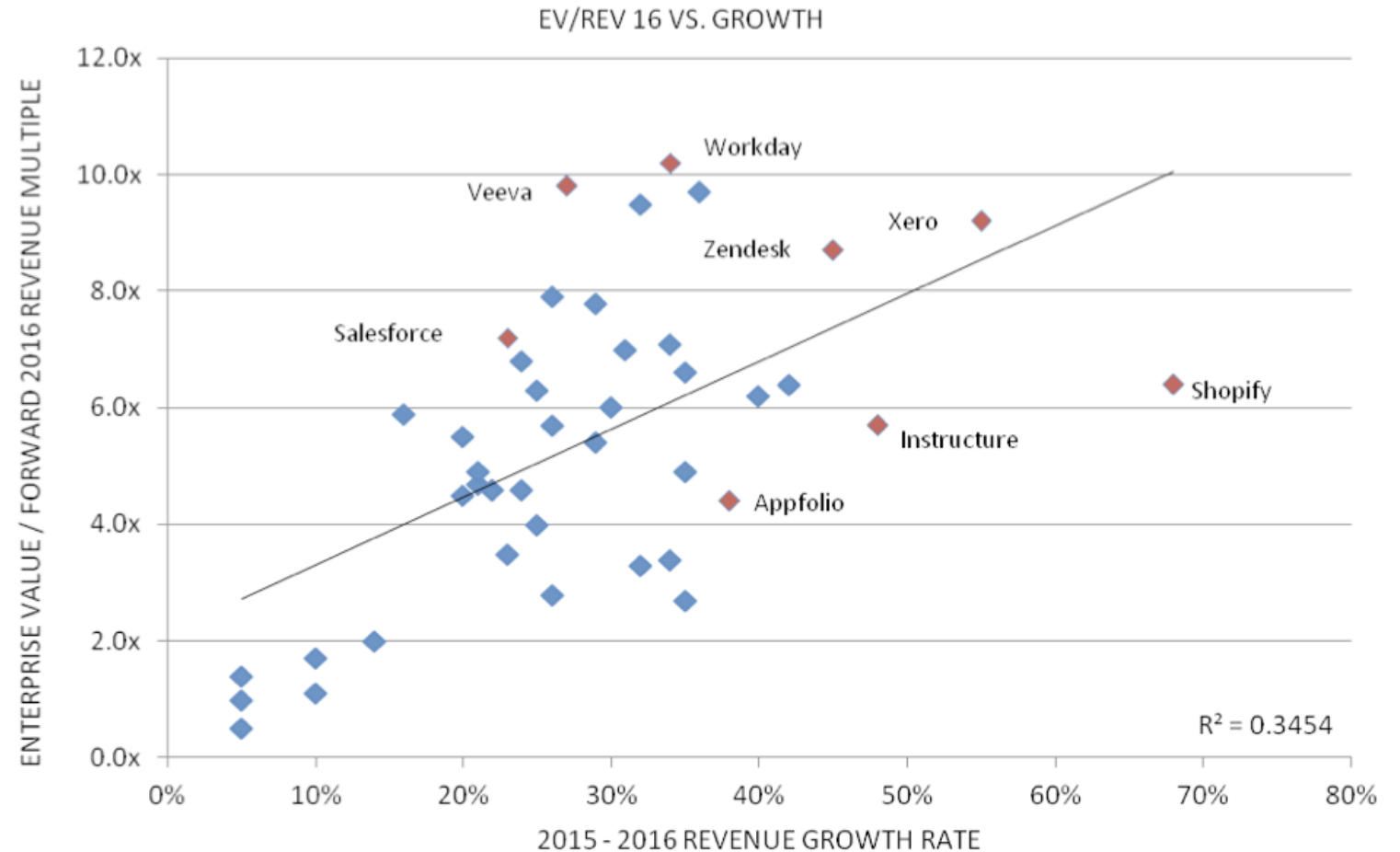
X is a metrics or ordinal, y is a metrics

Scatterplots

- Shows the relationship between two **continuous variables (x,y metrics)**
- Each point in the plot represents an observation
- You can change color of points to **highlight nominal attributes** (e.g., gender)
- So here we have 3 types of attributes shown: two metrics, one nominal

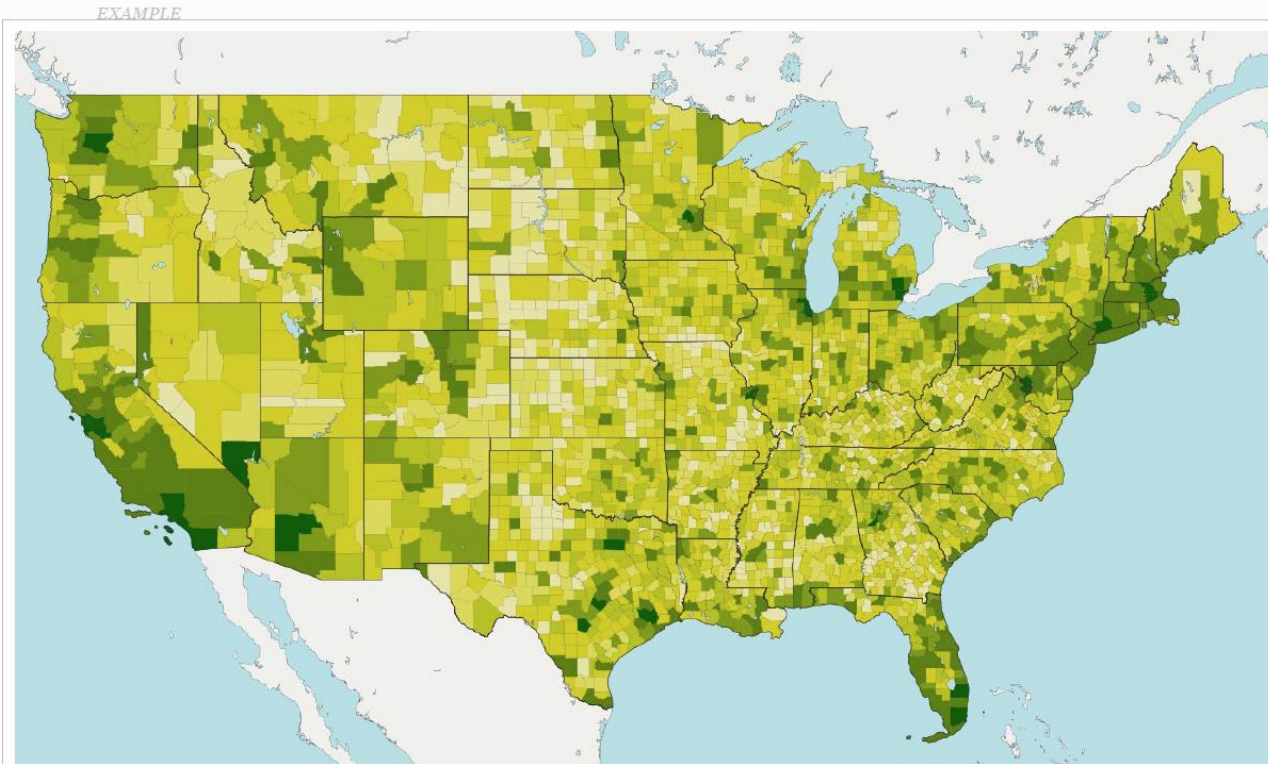


Another
example
scatterplot
(enterprise
value/revenue
vs growth)



Sometimes it is nice to show a trend line in the scatterplot

Maps

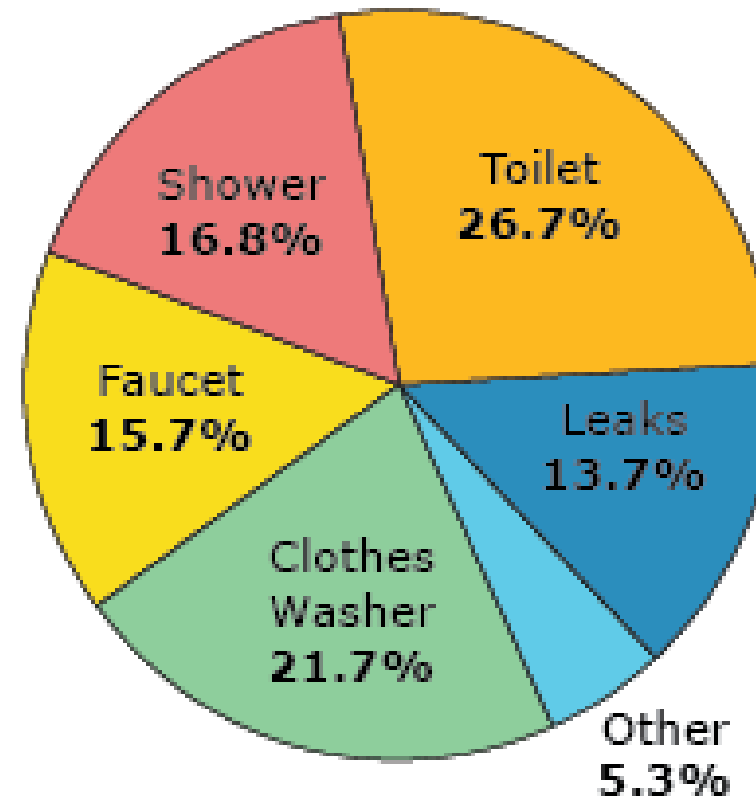


- Useful for analysis with a strong geographic component
- Remember: color scale comparisons are harder for humans than size comparisons. Keep this in mind as you choose between a map or another layout.

Pie Charts

- Almost never the right choice (angular comparison is hard)
- Use only if the following 2 conditions are met:
 - You want to show **the relative relationship** between 2-3 attributes
 - At least 1 is a metrics the other is a nominal
 - The count of metrics must add up to 100% (it should be a percentage)

How Much Water Do We Use?



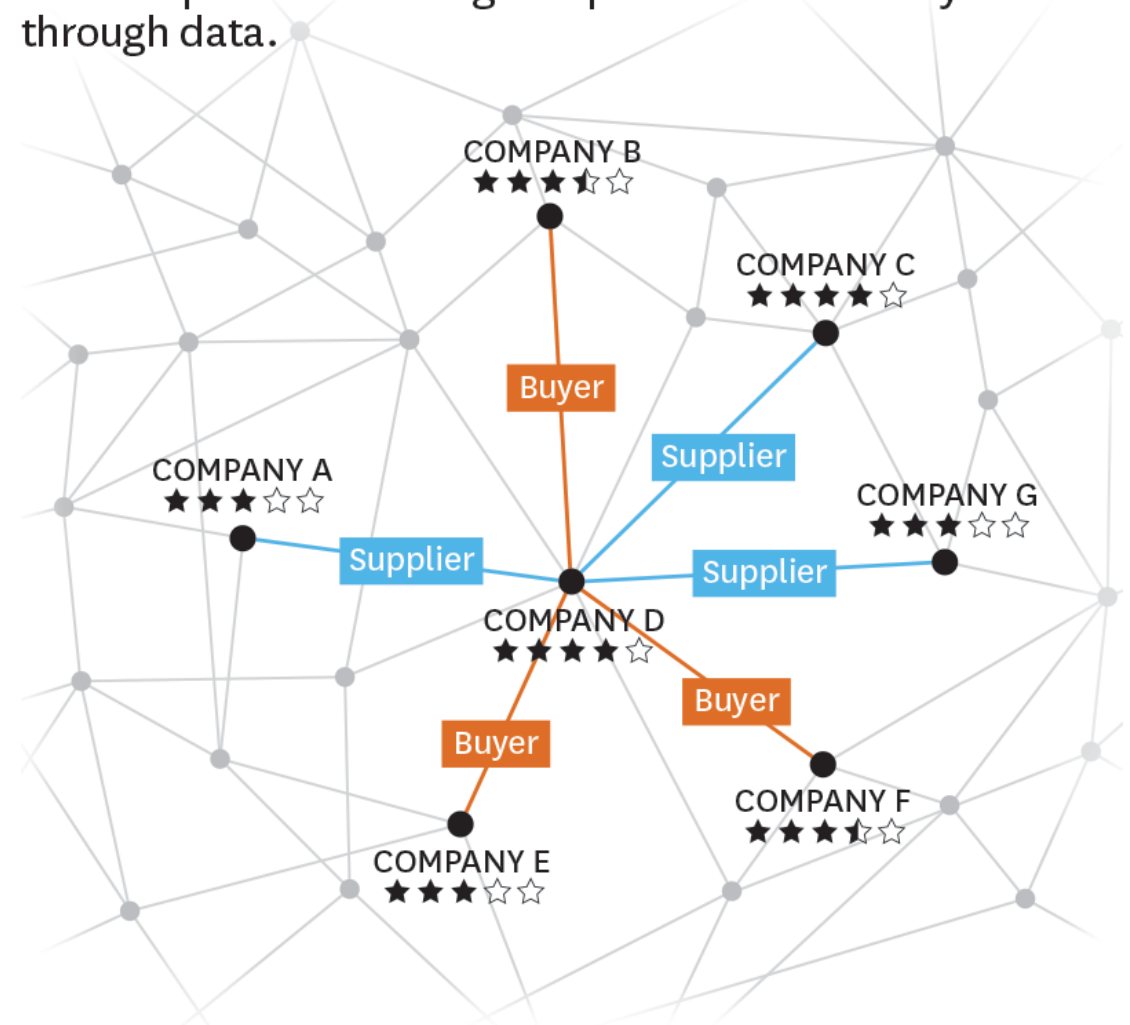
Source: American Water Works Association Research Foundation, "Residential End Uses of Water." 1999

Networks (graphs)

- Useful for showing the **relationships** between entities (both nodes and relations are labelled with nominal attributes)
- Can use color, size, etc. to encode additional – nominal - attributes about nodes/edges (e.g. here suppliers vrs buyers)
- Caveat: network diagrams quickly become hairballs for large, dense data.

THE COMMERCIAL GRAPH

An example of visualizing complex business ecosystems through data.



SOURCE PLATFORM THINKING LABS

HBR.ORG

Connecting Data To Visualization

- Data have types
- Visualizations have types
- Reports have communication objectives
- **How do we map one onto the other?**
 1. **Depending on the information we want to convey**
 2. **Depending on the type of data we want to visualize**

SELECTING VISUALIZATIONS

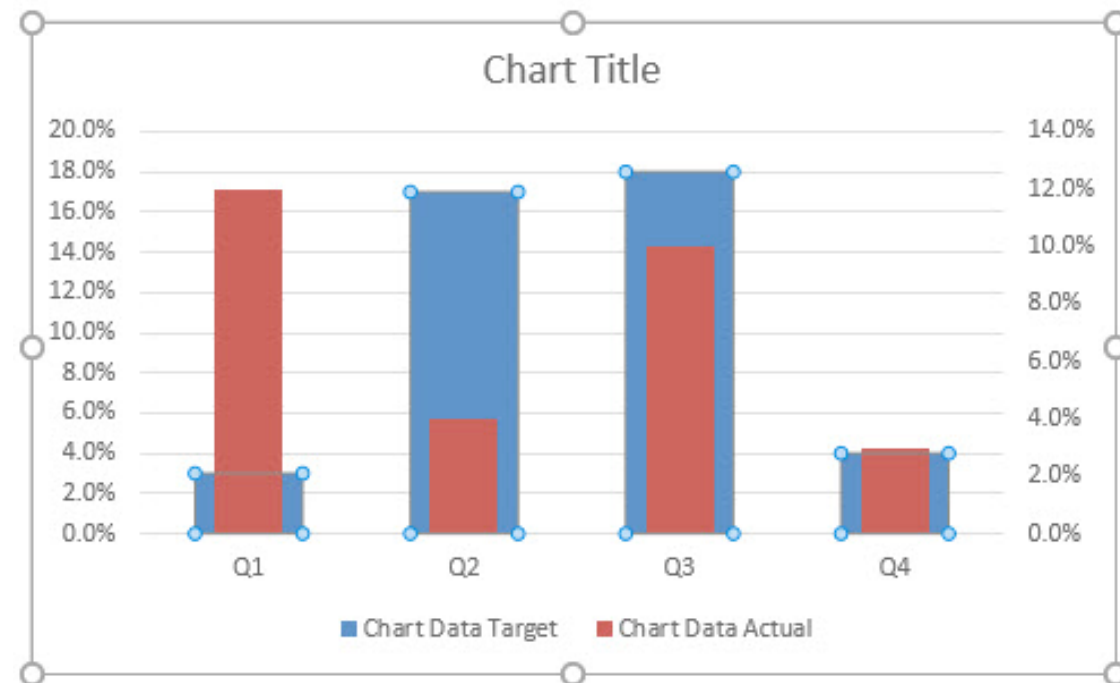
**Depending on the
information we want
to convey**

Do you want to **compare** values?

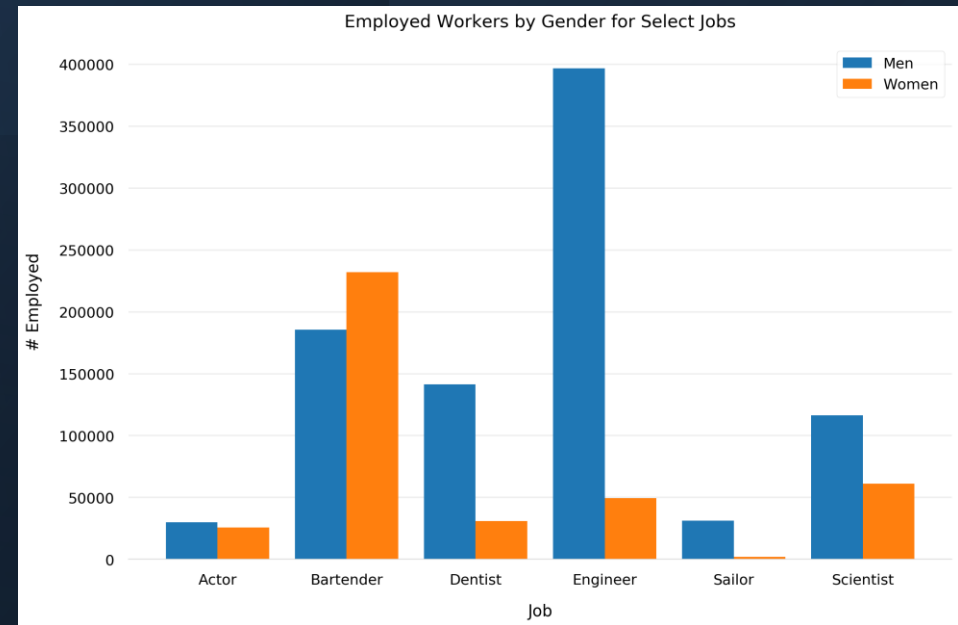
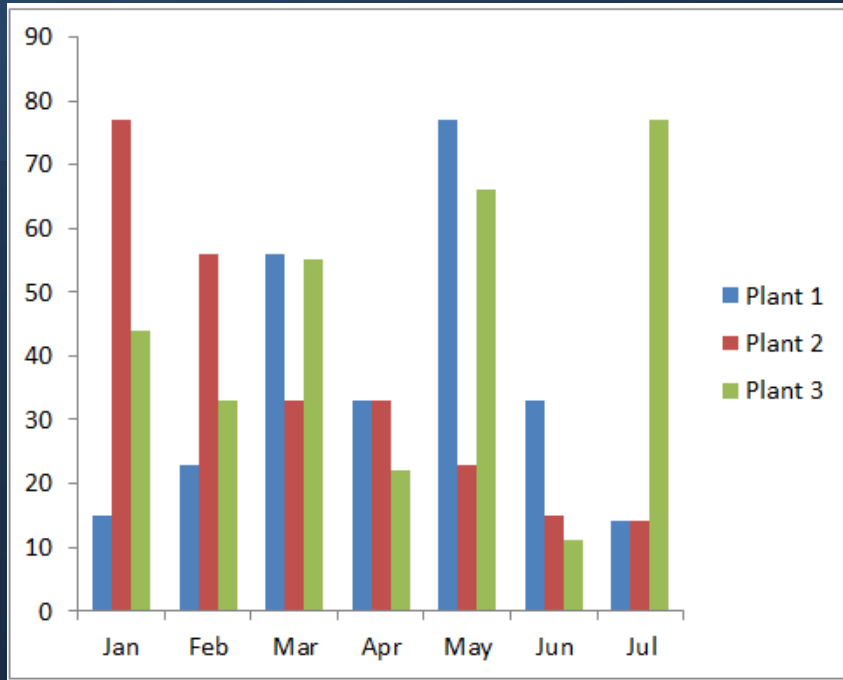
- Charts are perfect for comparing one or many value sets, and they can easily show the low and high values in the data sets.
- To create a **comparison chart**, use these types of graphs:
 - Column
 - Bar
 - Circular Area
 - Line
 - Scatter Plot
 - Bullet

A **column chart** is used to show a **comparison** among different items, or it can show a comparison of items over time.

- Design Best Practices for Column Charts:
- Use consistent colors throughout the chart, selecting accent colors to highlight meaningful data points or changes over time.
- Use horizontal labels to improve readability.
- Start the y-axis at 0 to appropriately reflect the values in your graph.



Other ways of representing comparisons with histograms (bar charts)



A horizontal column bar chart should be used to avoid clutter when one data label is long (e.g., “Individual contributors” or if you have more than 10 items to compare.

Visitors in museums in 2020

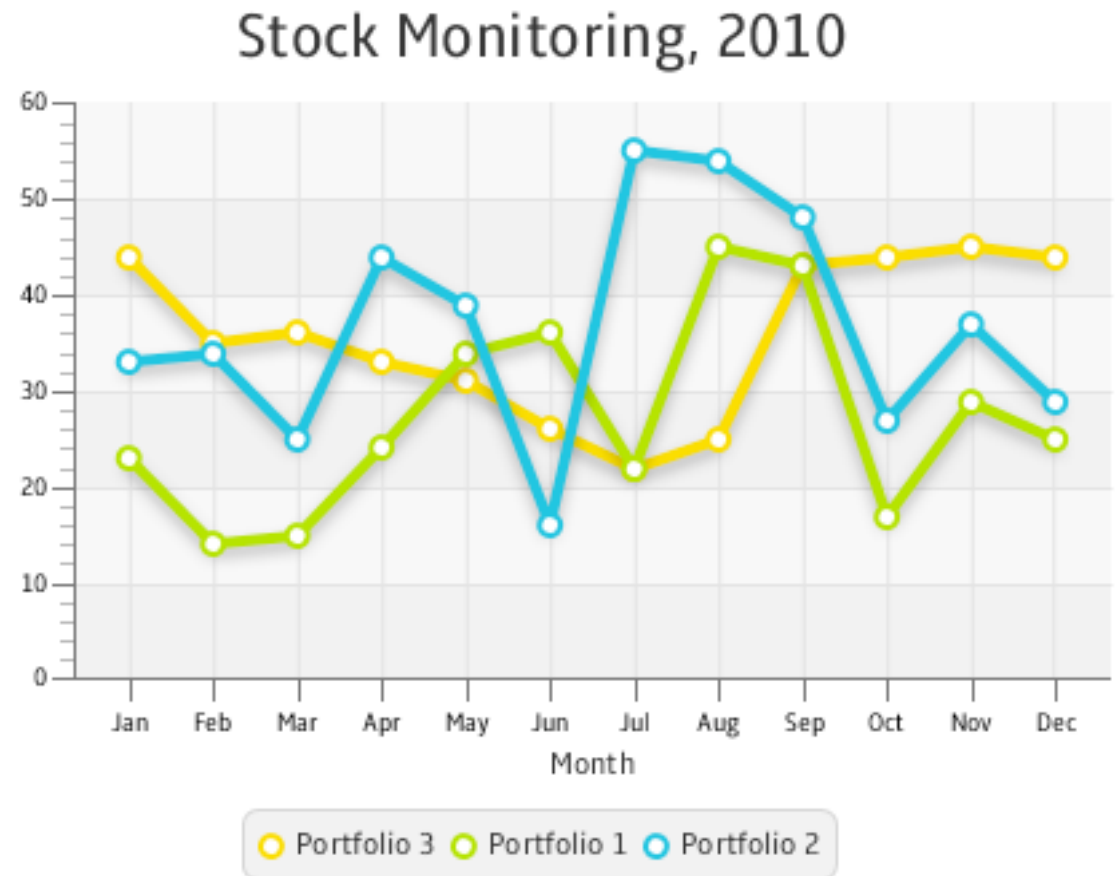
Musée du Louvre	10,200,000
National Museum of China	8,610,092
Metropolitan Museum of Art	6,953,927
Vatican Museums	6,756,186
Tate Modern	5,868,562
British Museum	5,820,000
National Gallery	5,735,831
National Gallery of Art	4,404,212
State Hermitage Museum	4,220,000
Victoria and Albert Museum	3,967,566

Source: [Wikipedia](#) • [Get the data](#) • Created with [Datawrapper](#)

Excuses for being late to class

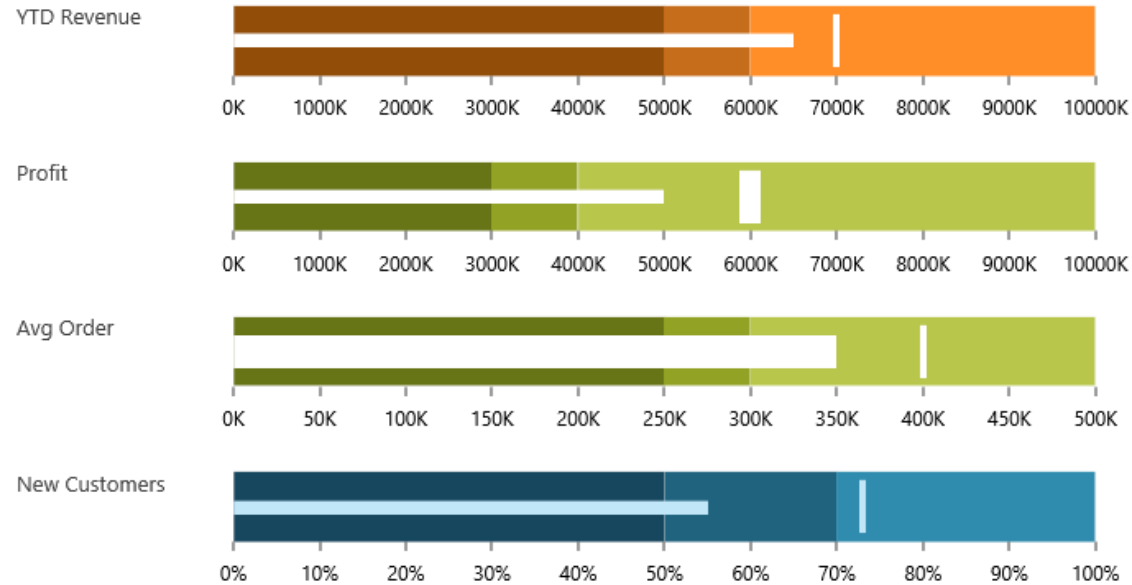


A **line chart** reveals trends or progress over time and can be used to show many different categories of data. You should use it when you chart a **continuous variable (metrics)**.



A bullet graph reveals progress toward a goal, compares this to another measure, and provides context in the form of a rating or performance.

Sunkost: Sales Target of 2010



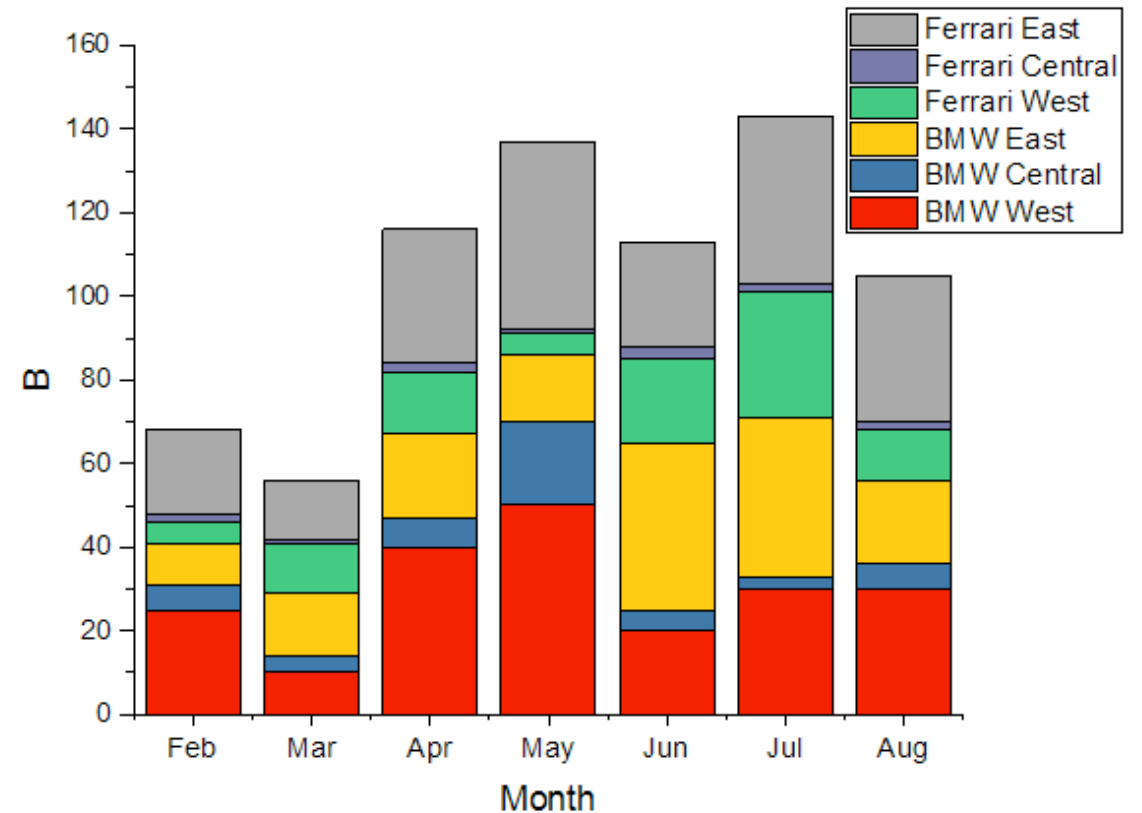
The sales totals to date for 2013 (white horizontal line), clearly exceed the total sales for all of 2012 (the beginning of the middle range). The 2013 sales numbers suggest that our new marketing campaign is successful, resulting in increased product penetration and a significant sales boost, working our way up to the targeted goals for the whole year (vertical white lines).

- In this example, the vertical white bar is the target, the horizontal white line is what has been achieved so far

Do you want to show the «composition» of something?

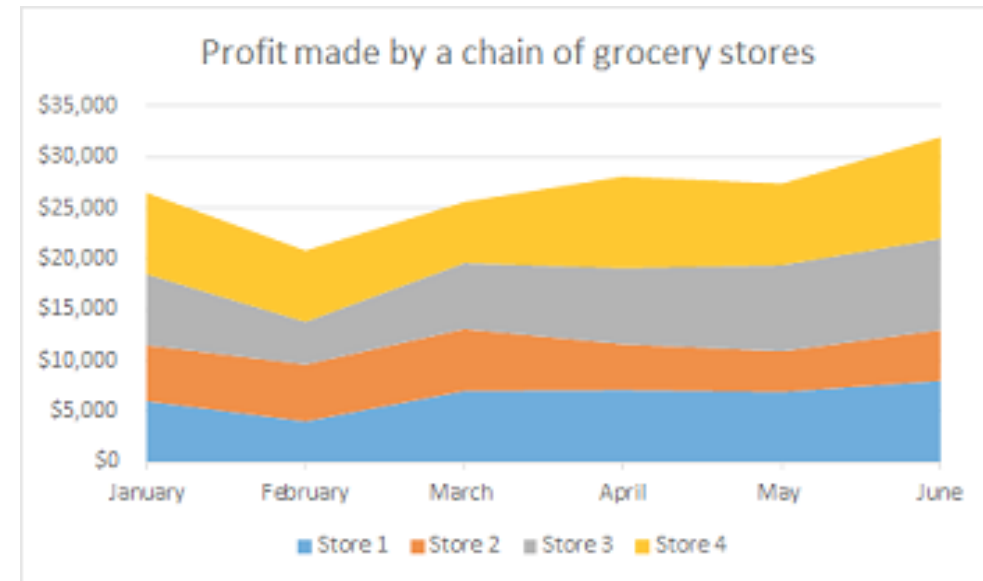
- Use this type of chart to show **how individual parts make up the whole** of something, such as the device type used for mobile visitors to your website, or total sales broken down by sales departments. Better suited to compare categories.
- To show composition, use these charts:
 - Pie
 - Stacked Bar
 - Stacked Column
 - Area
 - Waterfall

A **stacked chart** is used to break down and **compare parts of a whole**. Each bar in the **chart** represents a whole, and segments in the bar represent different parts or categories of that whole.



The «whole» here are the total parts sold (in a given month).

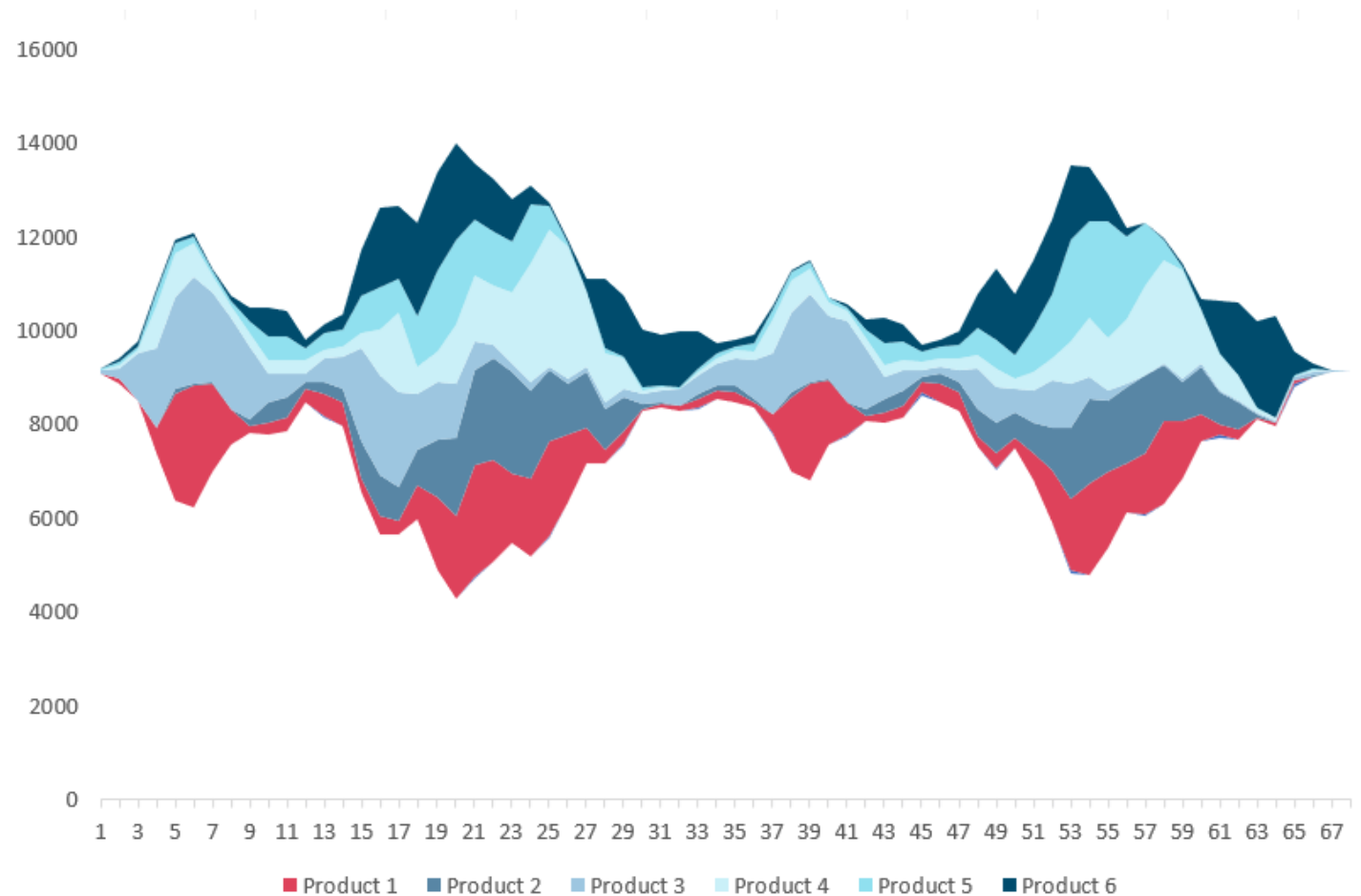
An **area chart** is basically a line chart, but the space between the x-axis and the line is filled with a color or pattern. It is useful for showing **part-to-whole relations**, such as showing individual sales reps' contribution to total sales for a year. It helps you analyze both overall and individual trend information.



The thickness of a line for a given month shows how a specific Store has contributed to total sales

Stream charts (a.k.o. area charts)

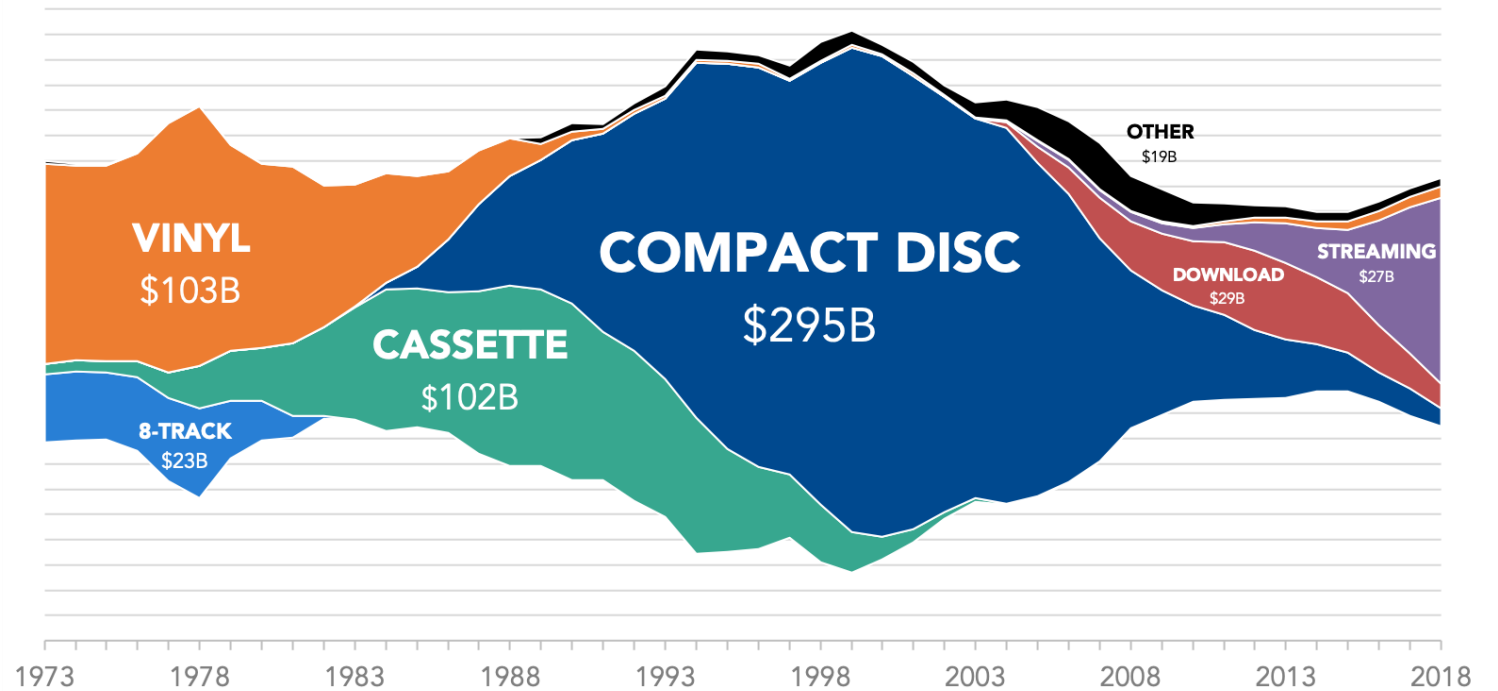
- It shows the percentage of sales for different products



Another example: US music sales by format

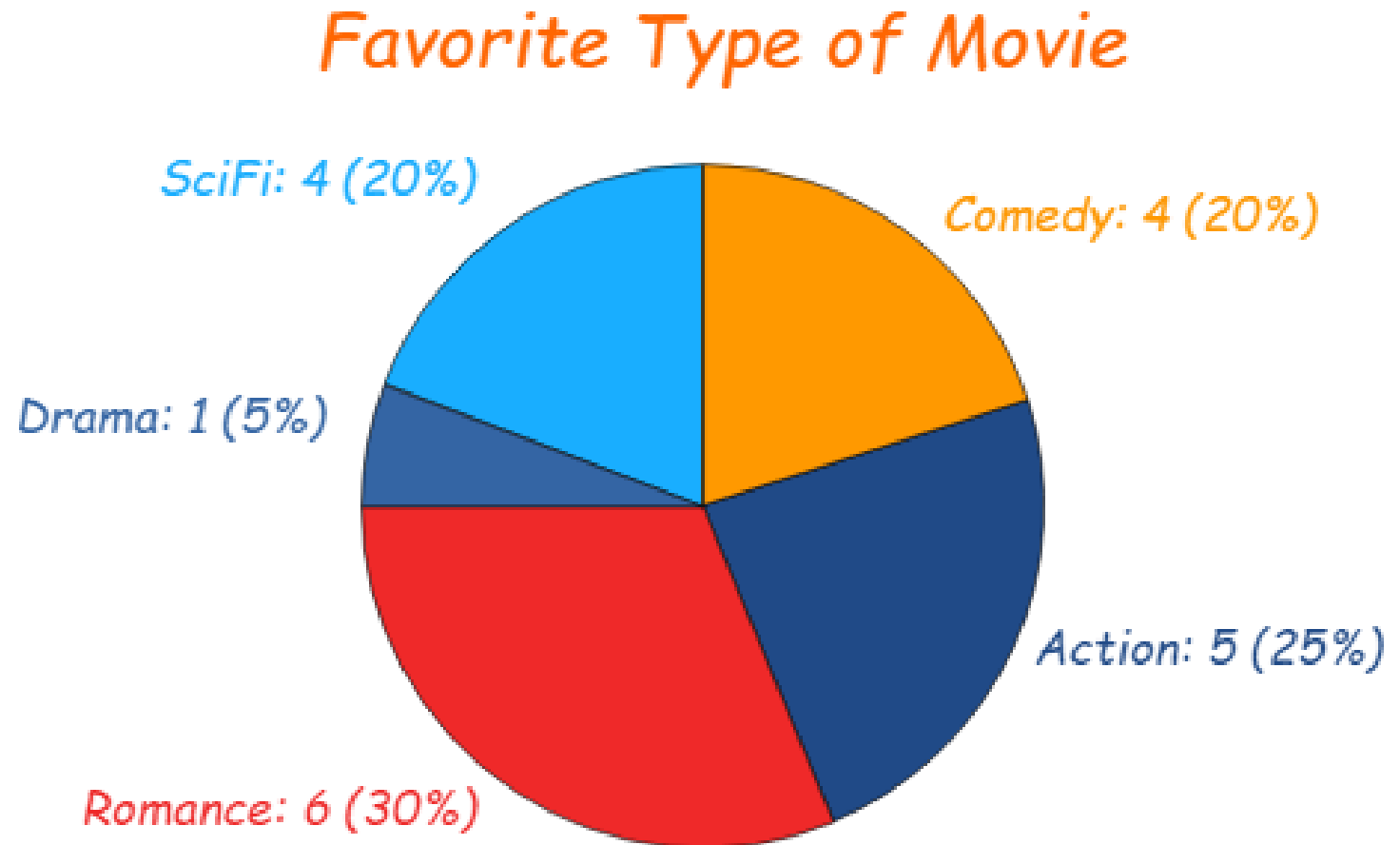
US music sales by format (inflation-adjusted)

EACH INTERVAL = \$1 BILLION (USD)

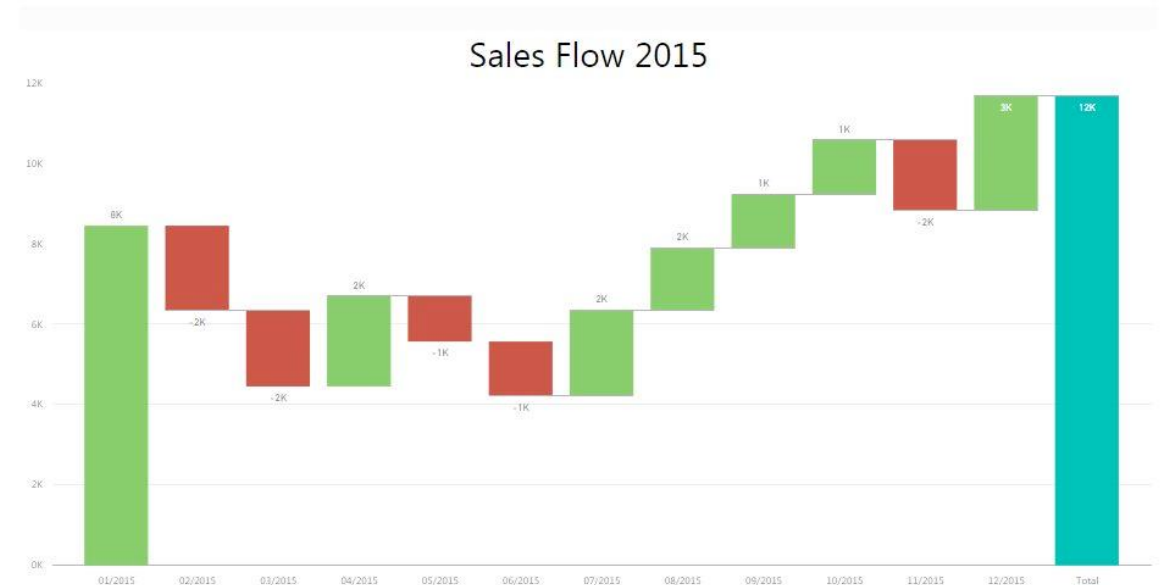


SOURCE: Recording Industry Association of America

A **pie chart** shows a static number and how **categories** represent part of a whole -- the composition of something. A pie chart represents numbers in percentages, and the **total sum of all segments needs to equal 100%.**



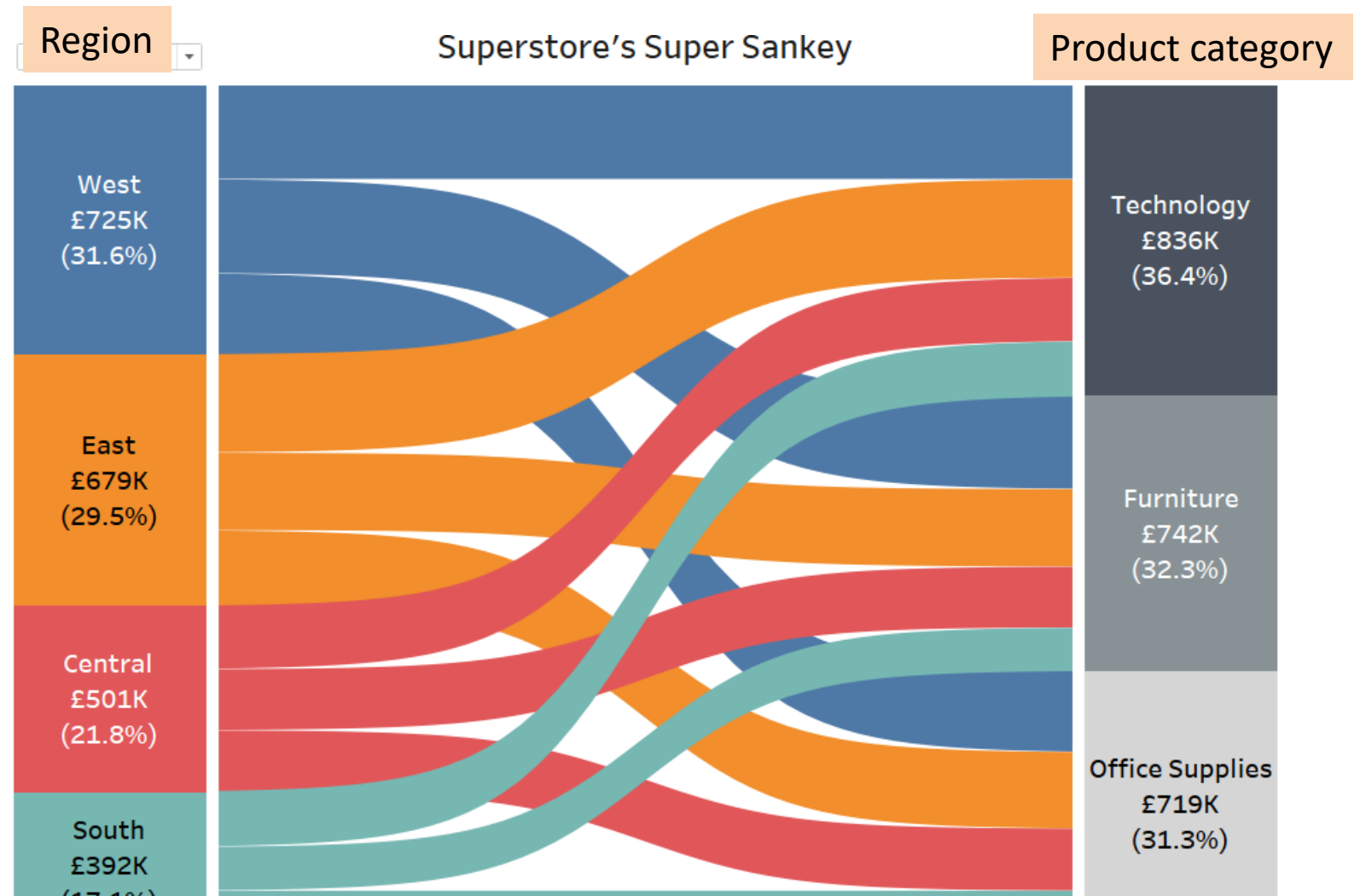
A **waterfall chart** should be used to show **how an initial value is affected by intermediate values** -- either positive or negative -- and **resulted in a final value**. This should be used to reveal the composition of a number.



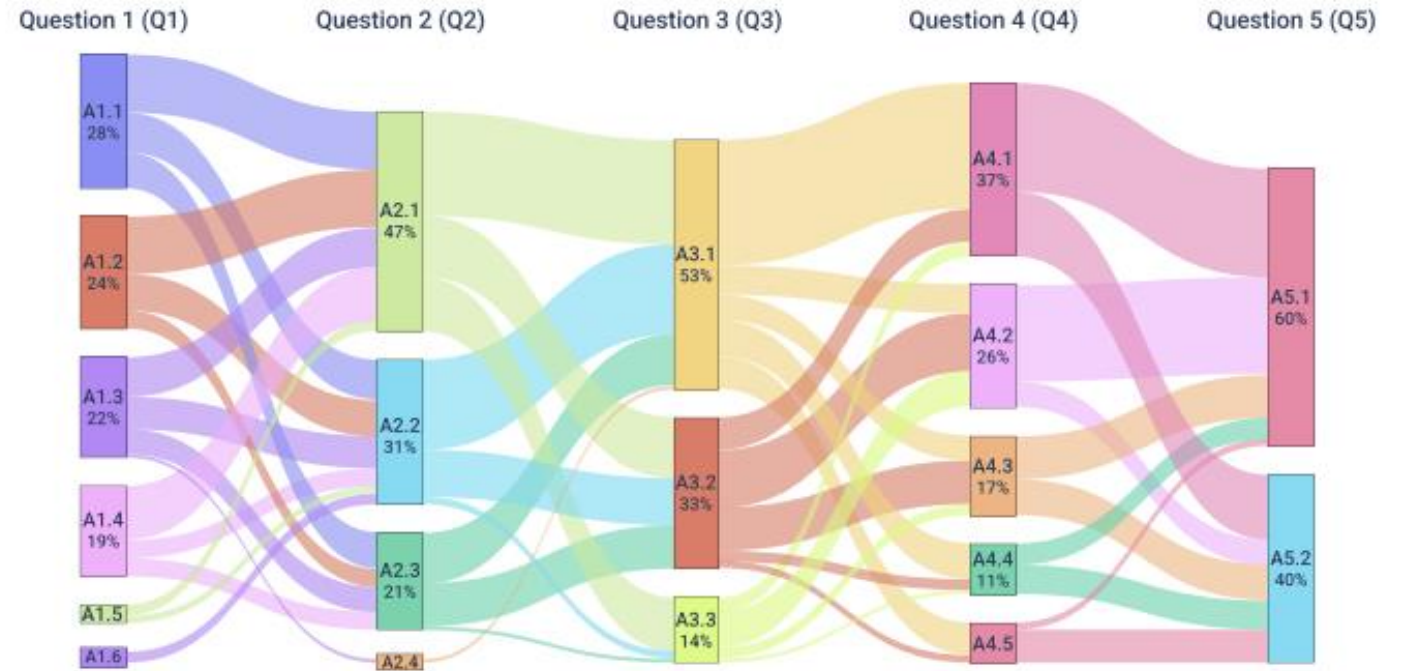
The graph showcases how subsequent movements affected the total balance (increases in value are coloured green and decreases are coloured red).

Sankey diagrams depict a flow from one set of values to another

Here, we have two attributes: Region and Category (two “wholes”), and the flow from one set of % to the other.



Sankey diagrams are also useful to represent closed-answer questionnaires



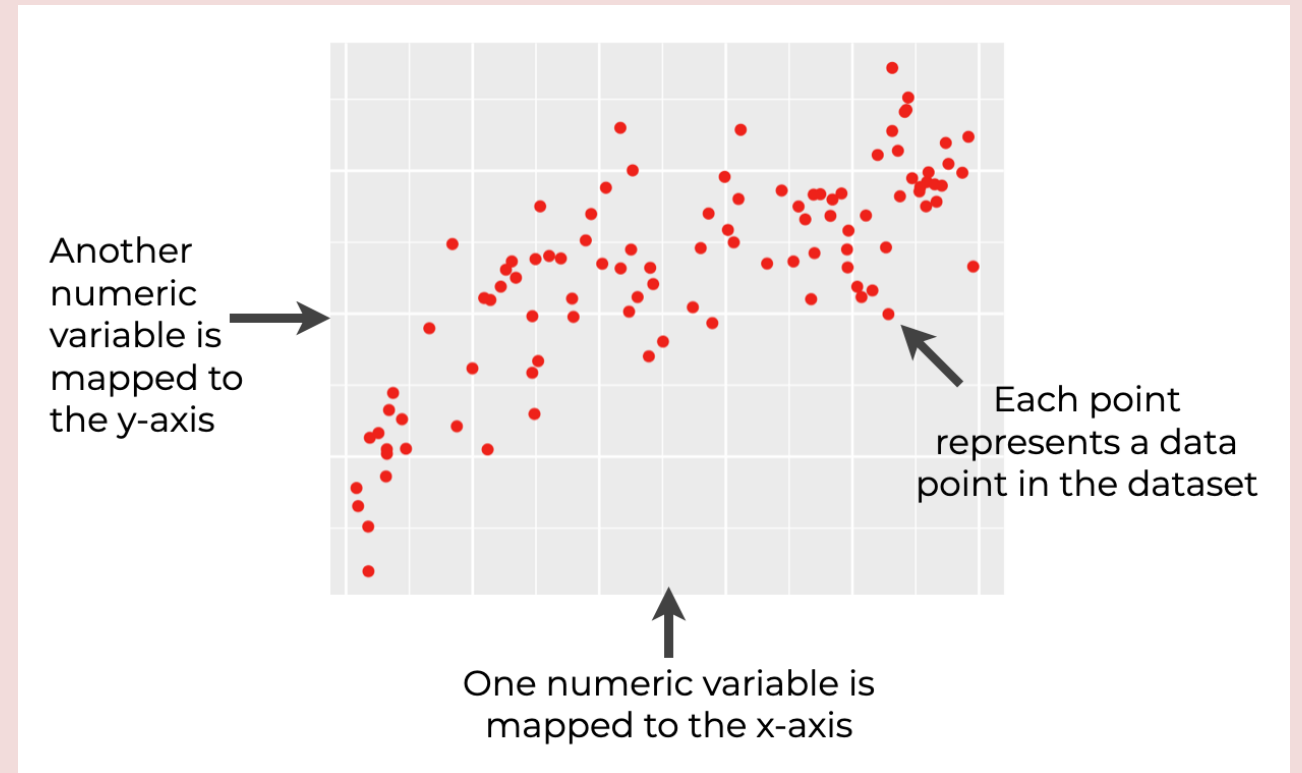
Do you want to understand the distribution of your data?

- Distribution charts help you to understand outliers, the **normal tendency**, and the range of information in your values.
- Use these charts to show distribution:
 - Scatter Plot
 - Line
 - Column
 - Bar

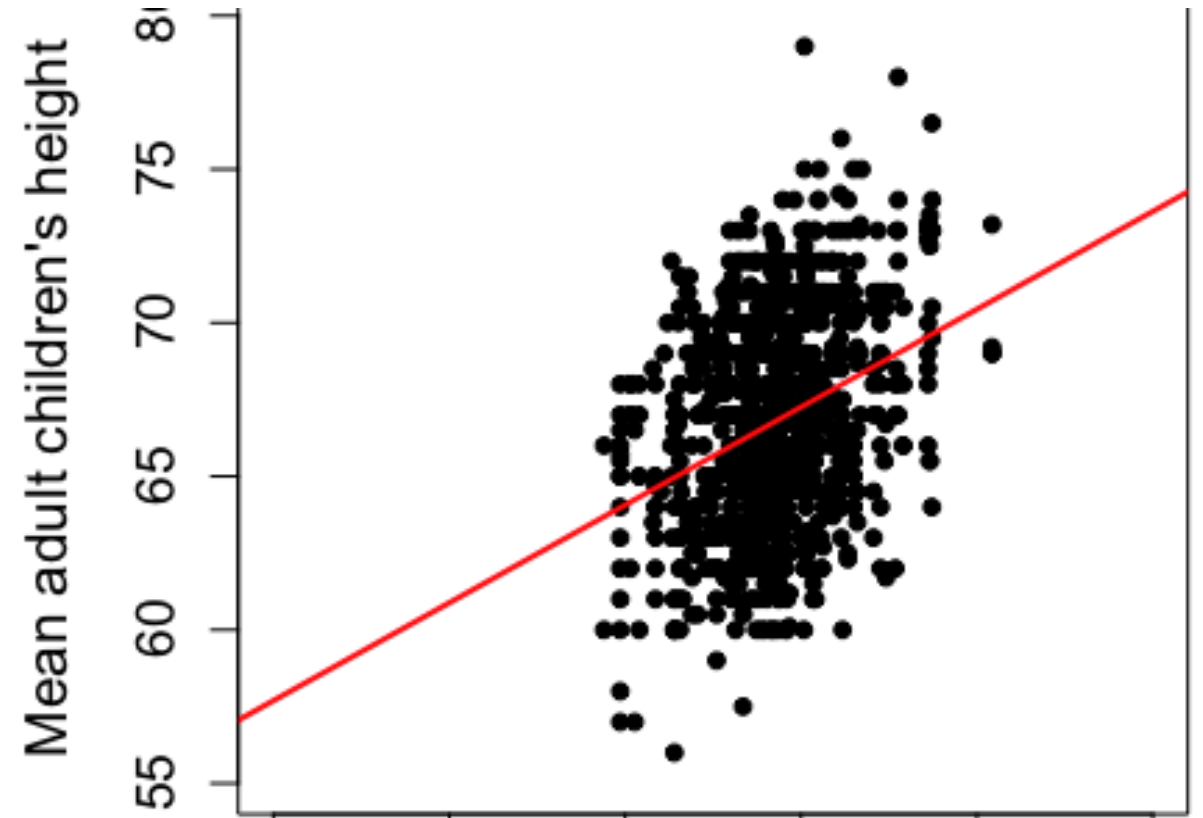
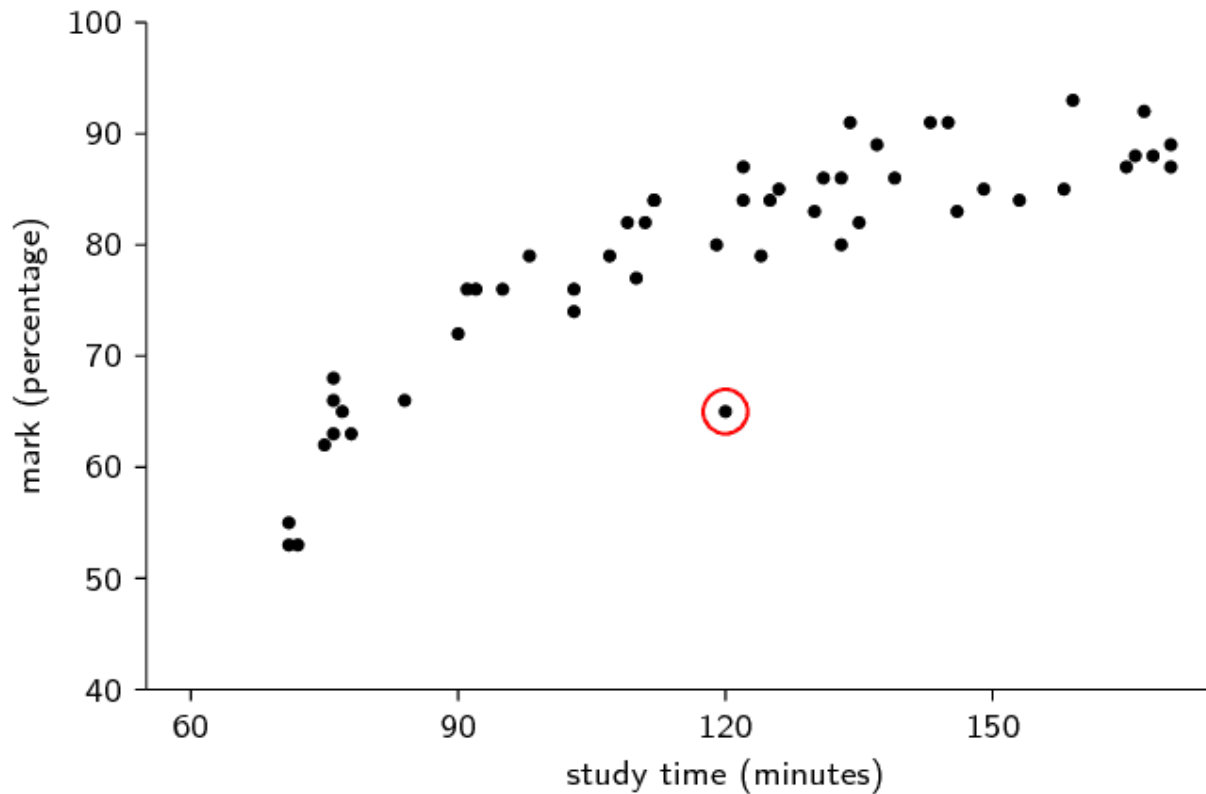
A **scatter plot** will show the relationship between two different variables (metrics) or it can reveal the distribution trends.

It should be used when there are many different data points, and you want to highlight similarities in the data set.

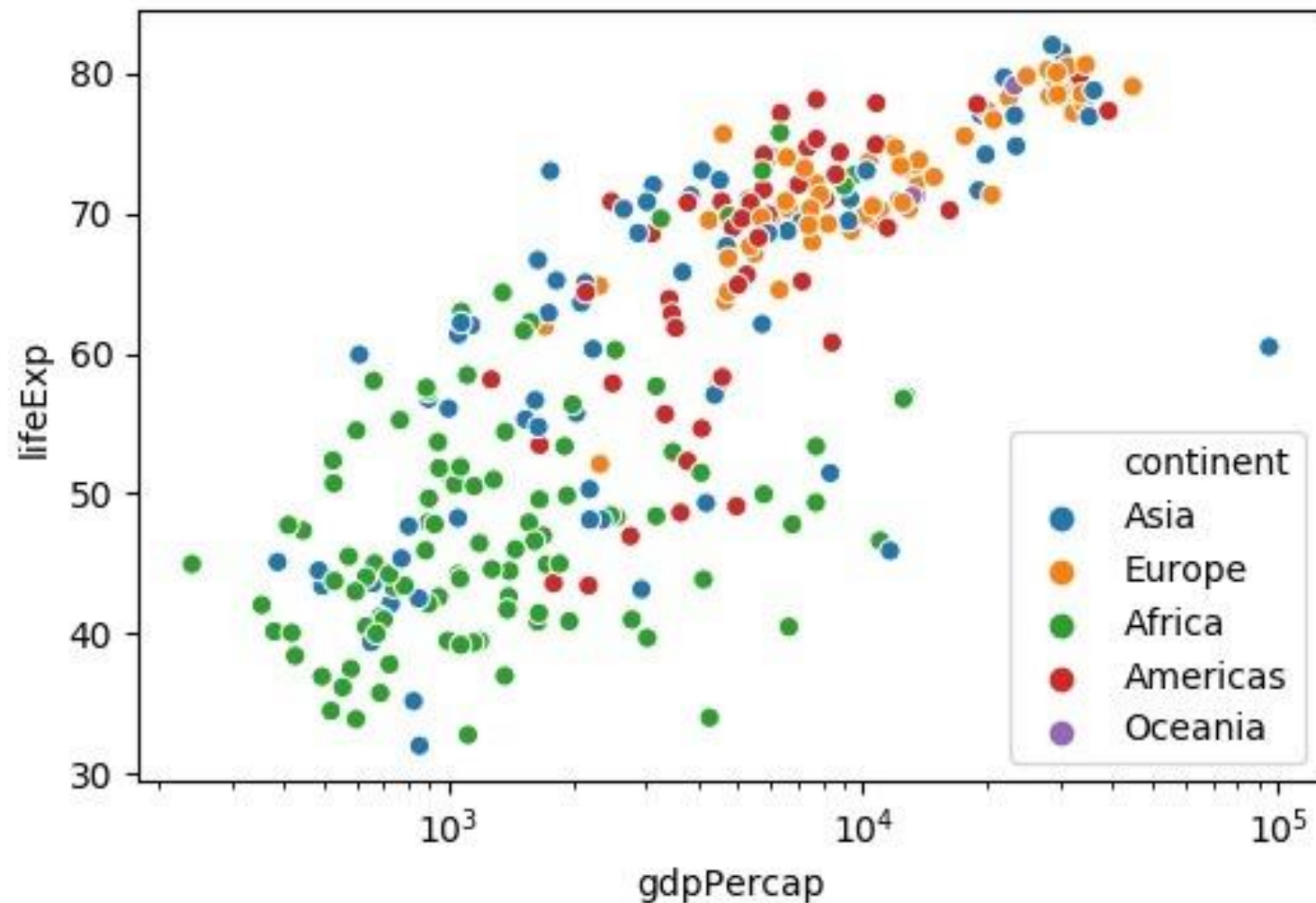
This is also useful when looking for outliers or for understanding the distribution of your data.



Examples of scatterplots

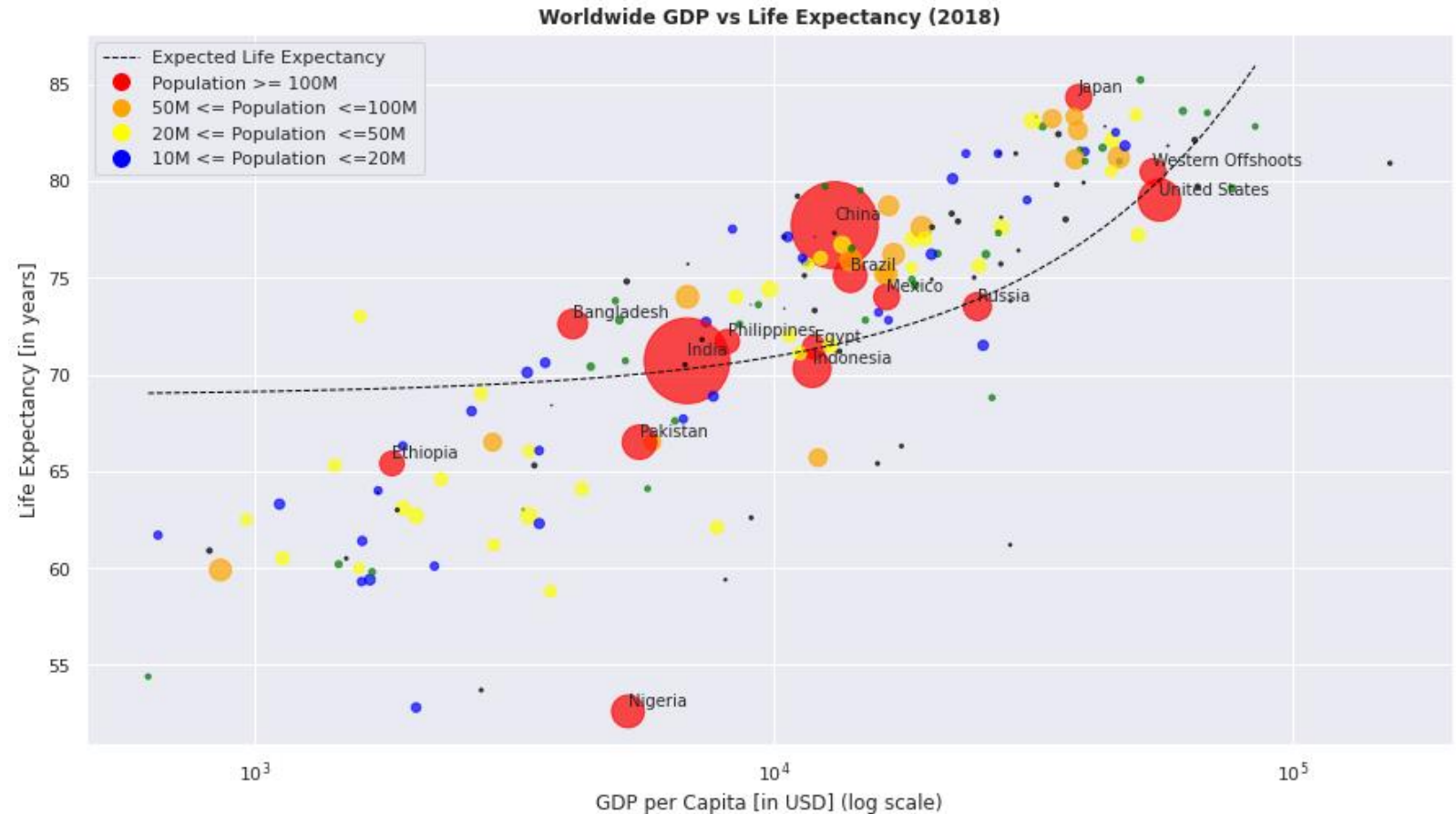


We can add a
third
categorical
variable
(attribute)



And a fourth variable (a metrics) represented by the dimension of the bubble.

These are also called bubble charts



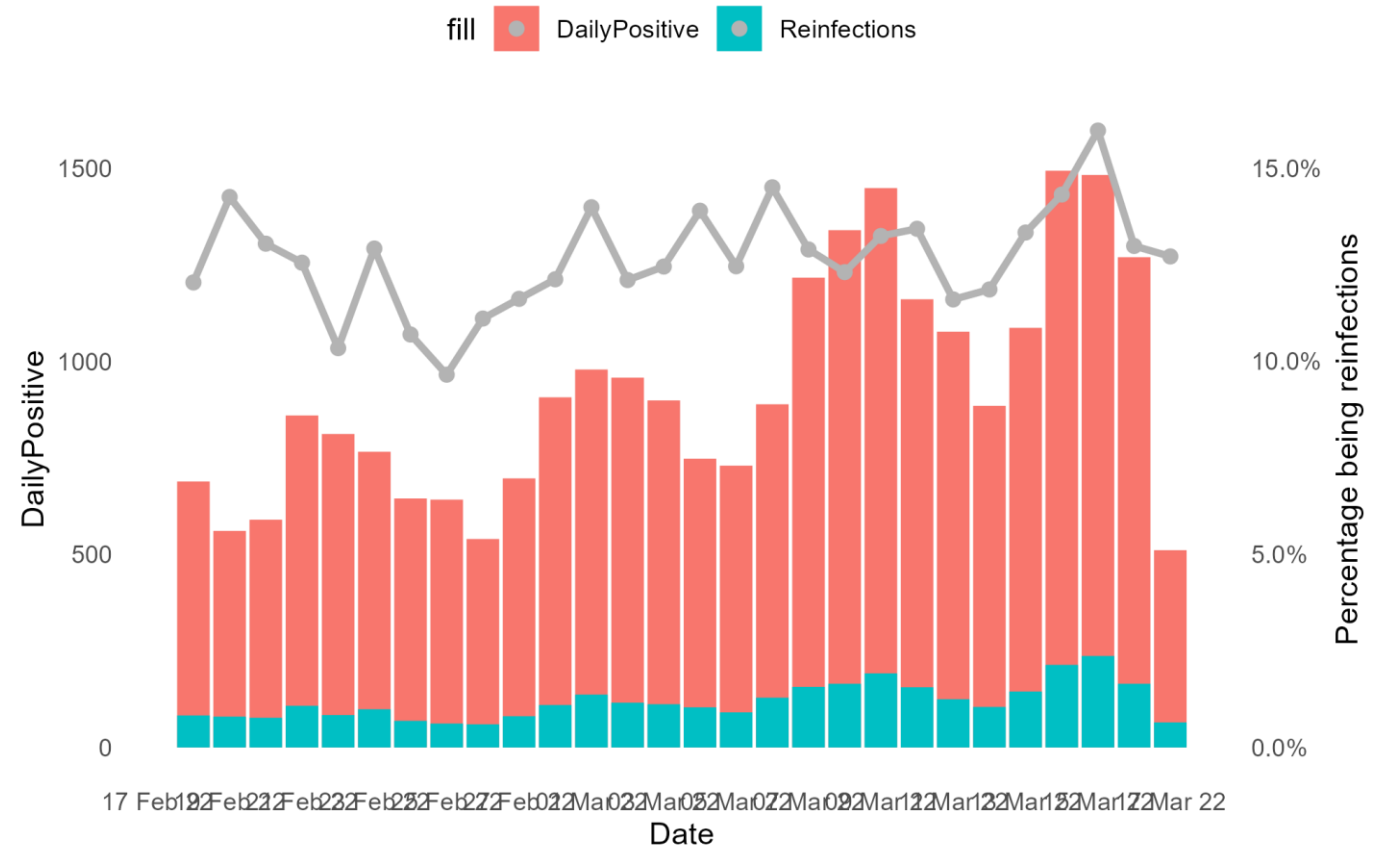
Are you interested in analyzing **trends** in your data set?

- If you want to know more information about how a data set performed during a specific time period, there are specific chart types that do extremely well.
- You should choose a:
 - Line
 - Dual-Axis Line
 - Funnels
 - Column

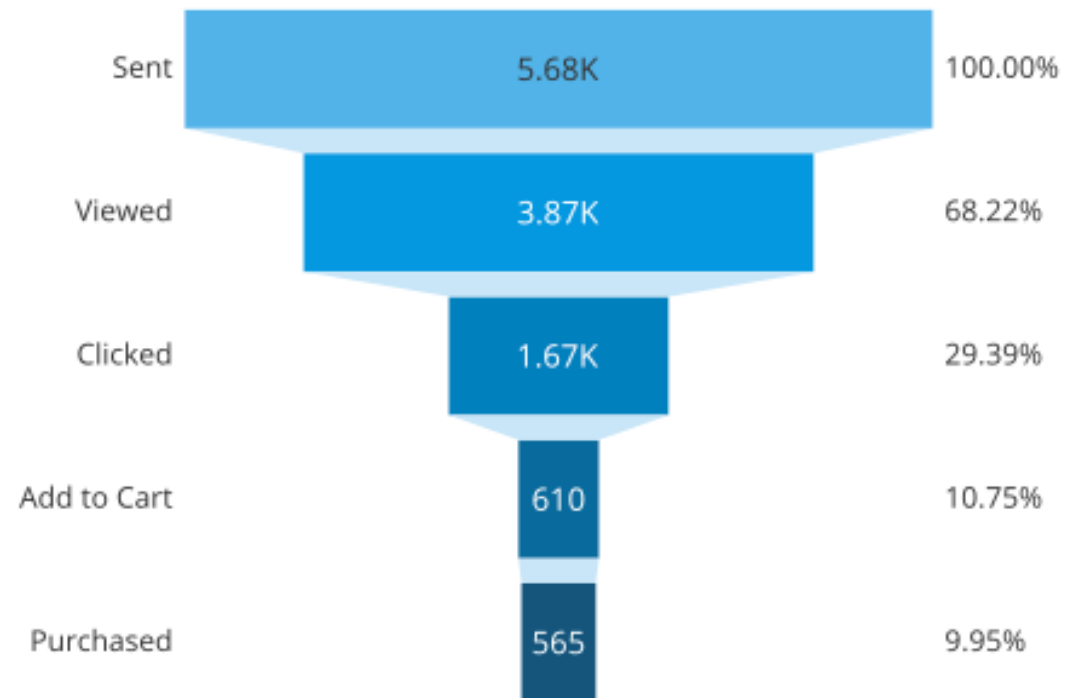
A **dual axis** chart allows you to plot data using **two y-axes** and a **shared x-axis**. It's used **with three types of variables**, two of which are **continuous** metrics and another which is a categorical or ordinal attribute. This should be used **to visualize a correlation or the lack thereof** between these three variables



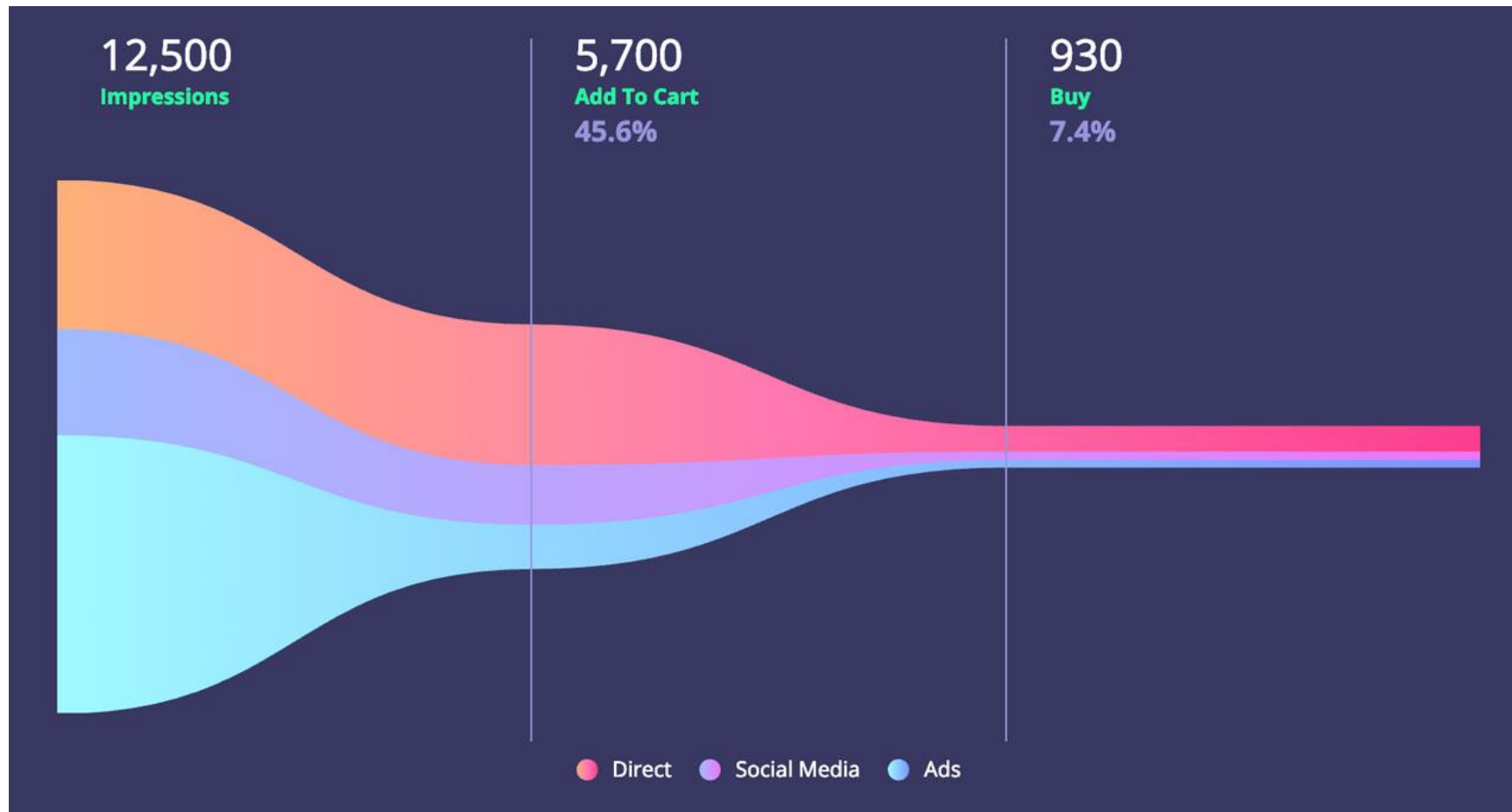
Another example:
x= ordinal (time)
y1=daily positive to
COVID (gray line)
y2=% reinfected



A **funnel chart** shows a series of steps and the **completion rate** for each step. This can be used to track the sales process or the conversion rate across a series of pages or steps (e.g. from contacts to contracts).



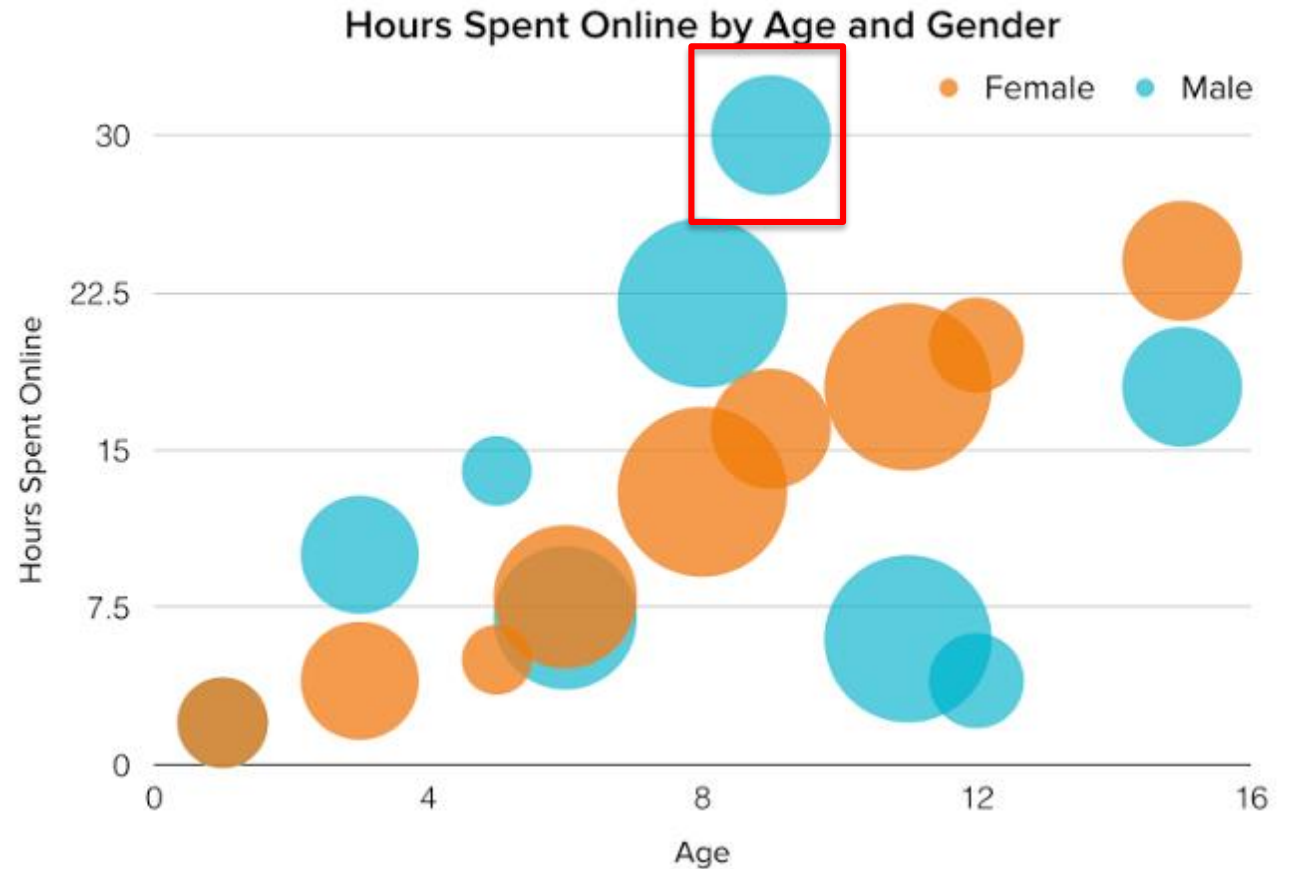
Another example of funnel



Do you want to better understand the relationship between value sets?

- Relationship charts are suited to showing **how one variable relates to one or numerous different variables.**
- You could use this to show **how something positively effects, has no effect, or negatively effects another variable.**
- When trying to establish the relationship between things, use these charts:
 - Scatter Plot
 - Heat maps
 - Bubble
 - Line
 - Networks
 - Spirals

A **bubble chart** is similar to a scatter plot in that it can show distribution or relationship. There are 4 dimensions here: x and y are numeric variables, the colour allows to incorporate discrete (symbolic) variables, e.g., gender, and the size of the bubble is a fourth numeric variable.



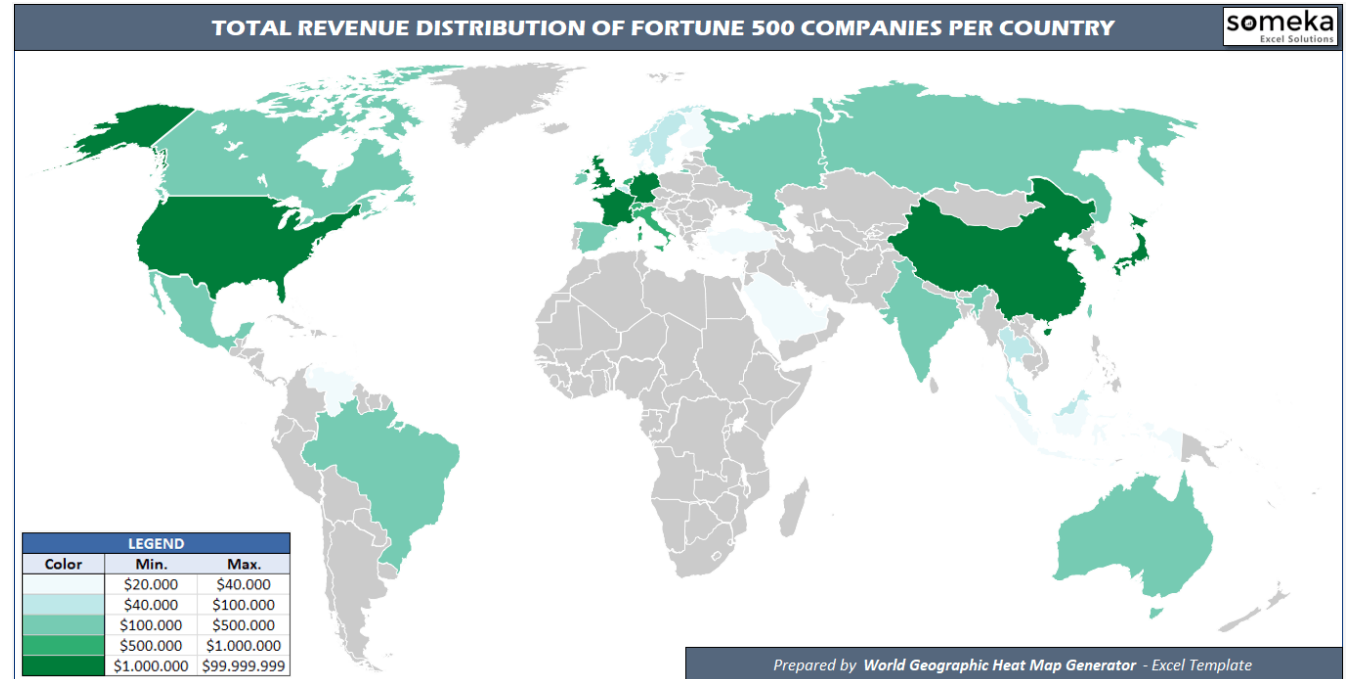
The dimension of the bubble indicates the dimension of the sample (e.g. how many females aged 8 spend 30 hours online in a week)

The rating information is displayed using varying colors or saturation. In the example, one variable is a metric (temperature at central park) the other is an ordinal (month and year)

Red indicates hottest periods, green the coldest

[illegible]

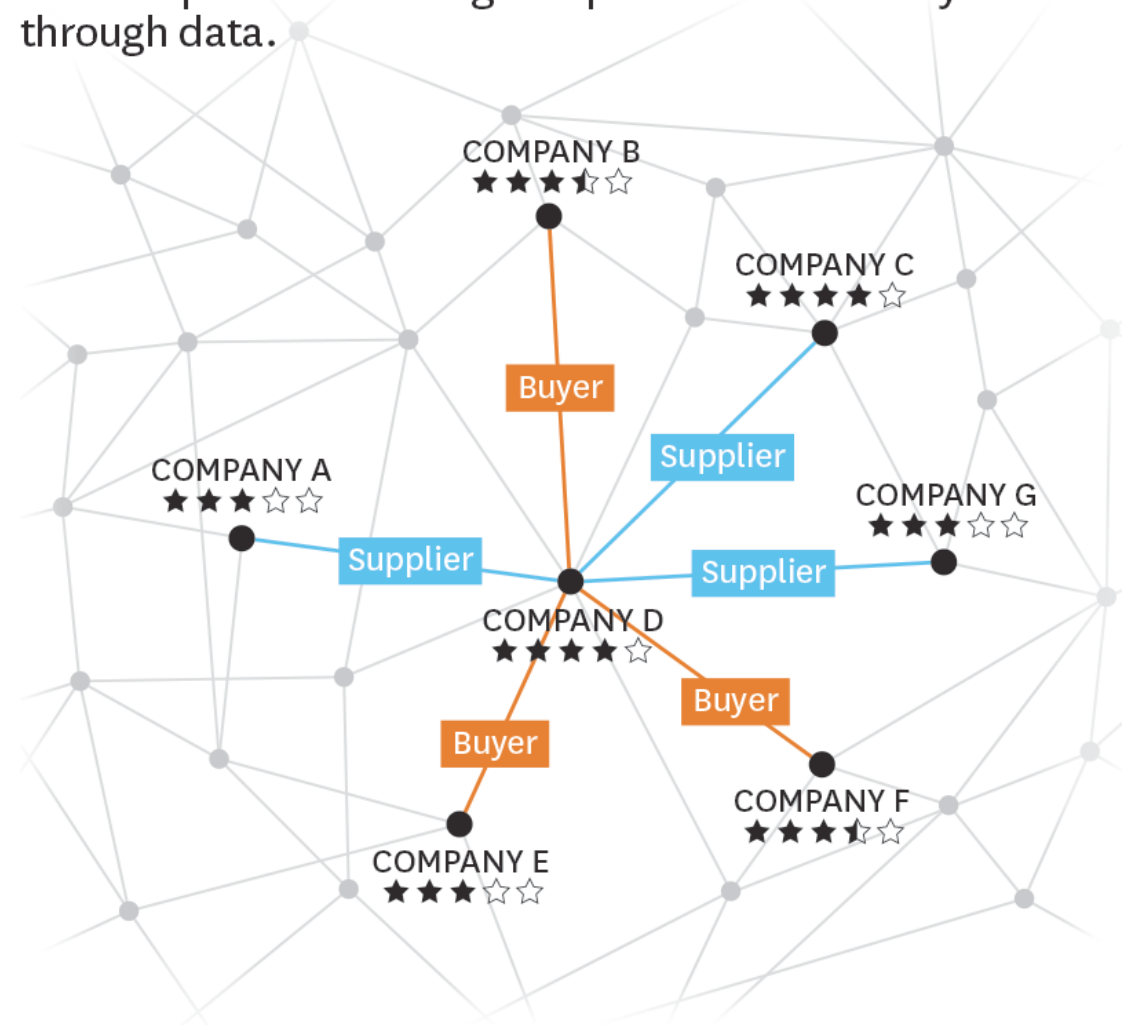
Heat maps can also be geographical maps (we add the attribute “region”)



Network graphs
are useful to
show complex,
non-numerical
relations
between entities

THE COMMERCIAL GRAPH

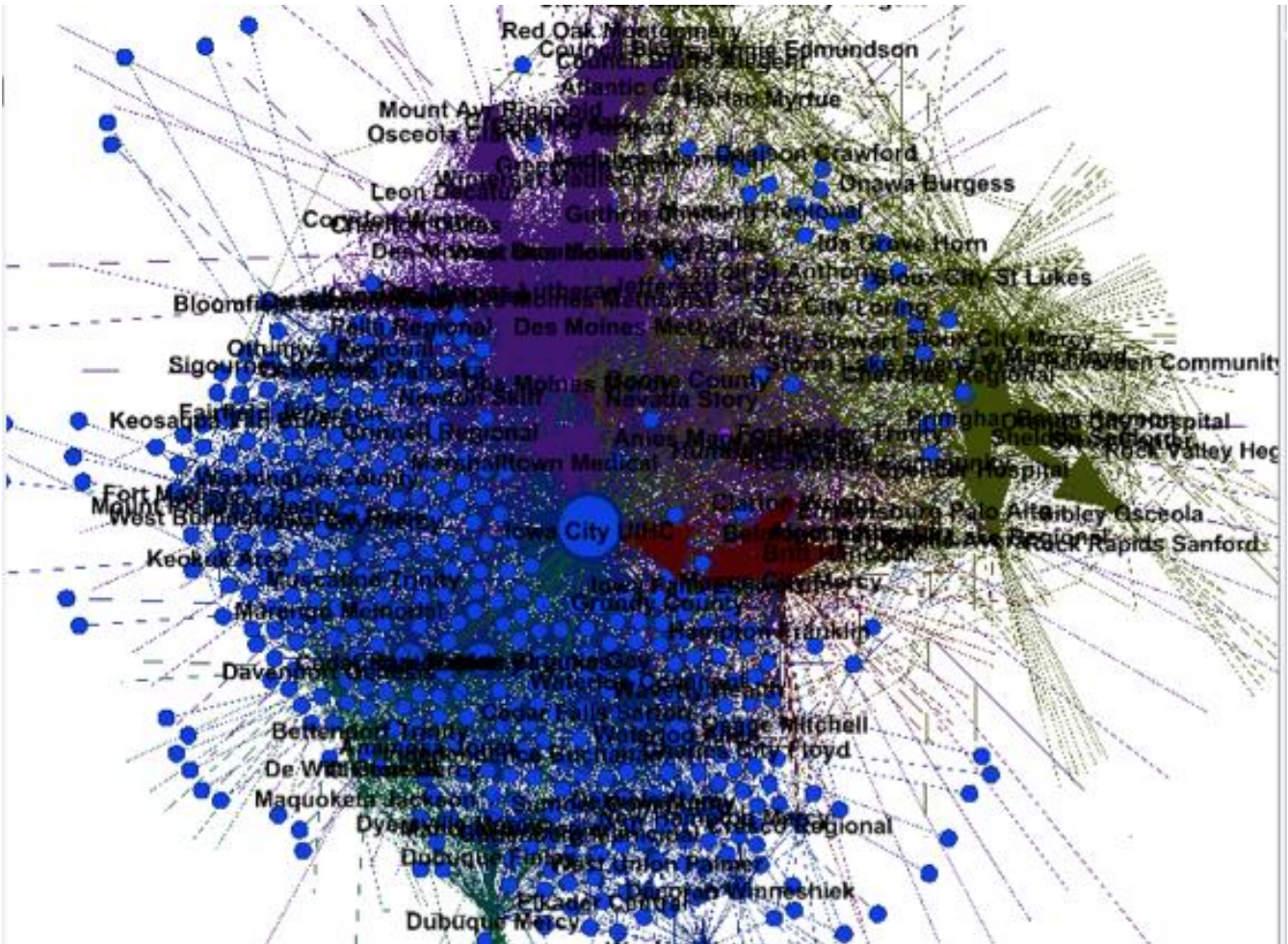
An example of visualizing complex business ecosystems through data.



SOURCE PLATFORM THINKING LABS

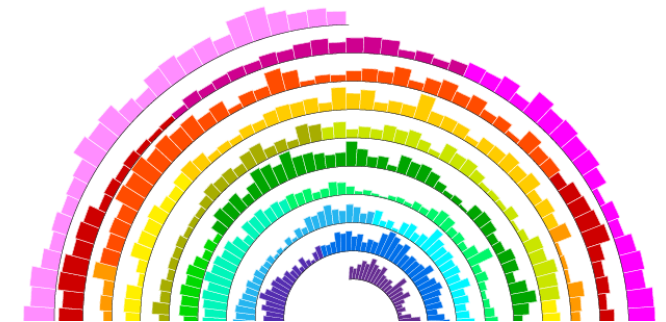
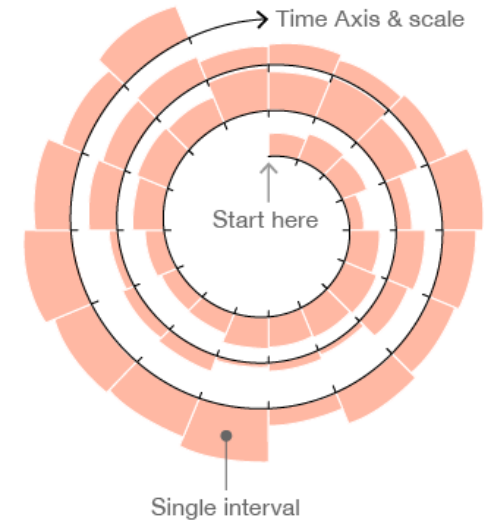
HBR.ORG

Although they quickly become unreadable

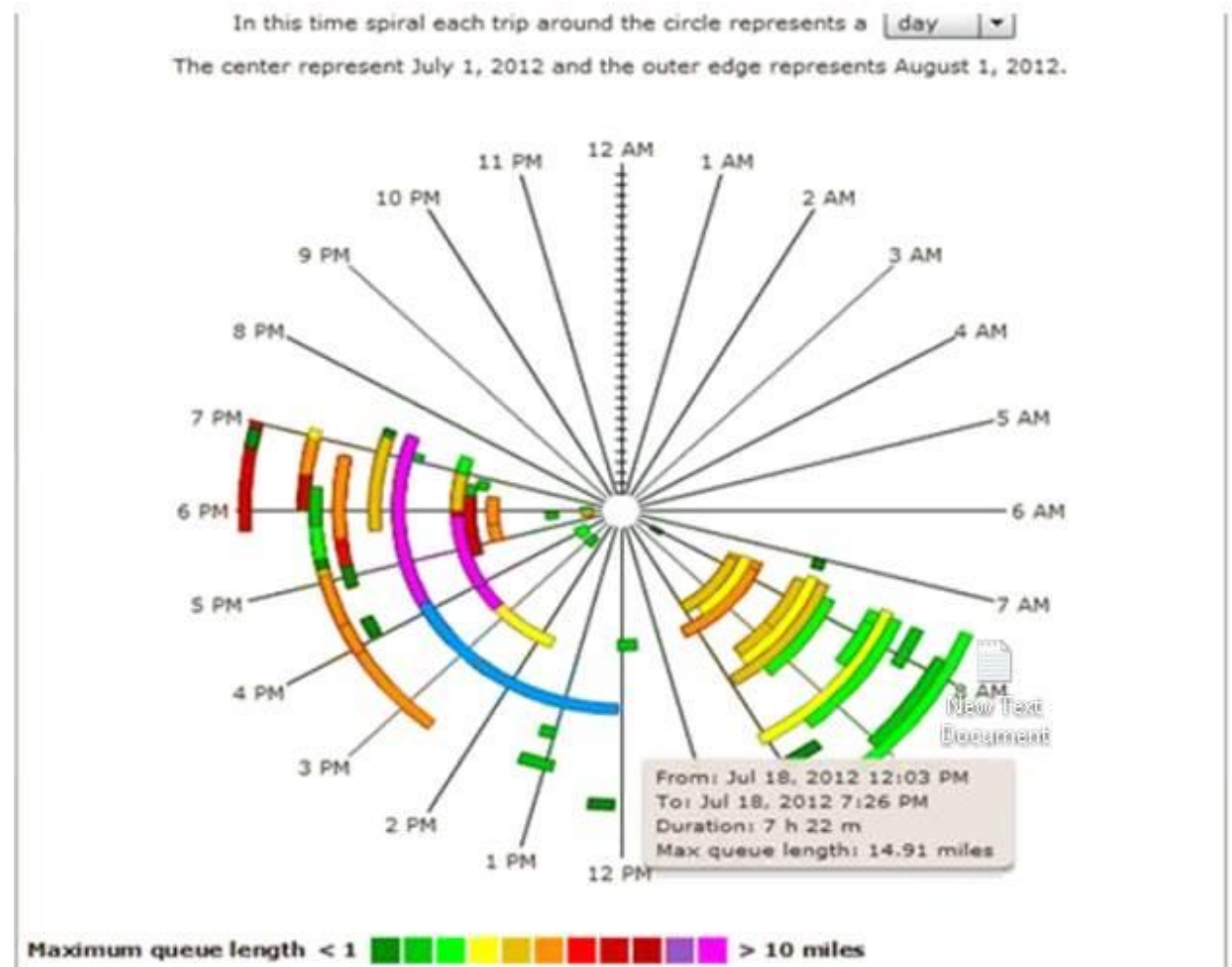


Spirals

- Spiral Plots are ideal for showing large data sets, usually to show **trends over a large time period**. This makes Spiral Plots great for displaying **periodic patterns**. Colour can be assigned to each period to break them up and to allow some comparison between each period.
- So for example, if we were to show data over a year, we could assign a colour for each month on the graph.



Spirals uses



Here each trip around the circle represent a day, colors represent queue lengths in streets

SELECTING VISUALIZATIONS

Depending on the type of data we want to visualize

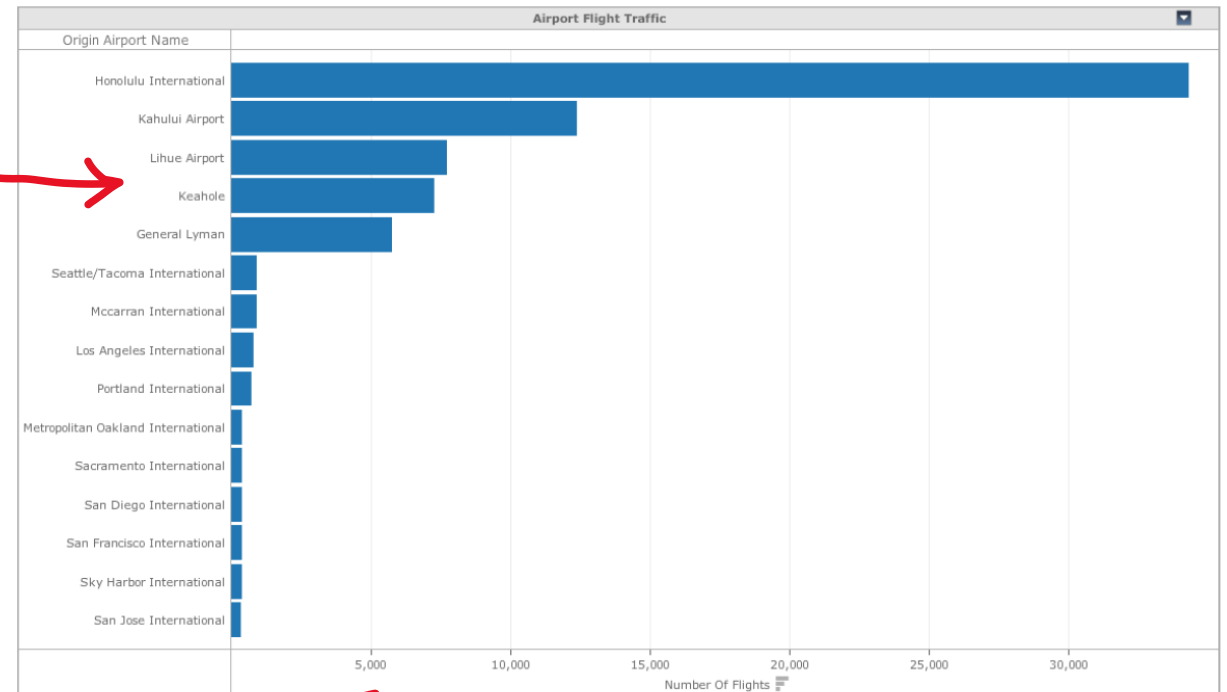
Types of visualizations also depends on the type of data

Remember the 3 types of variables in a dataset

- Qualitative (Attributes)
 - Nominal
 - Ordinal
- Quantitative (Metrics)
 - Numeric

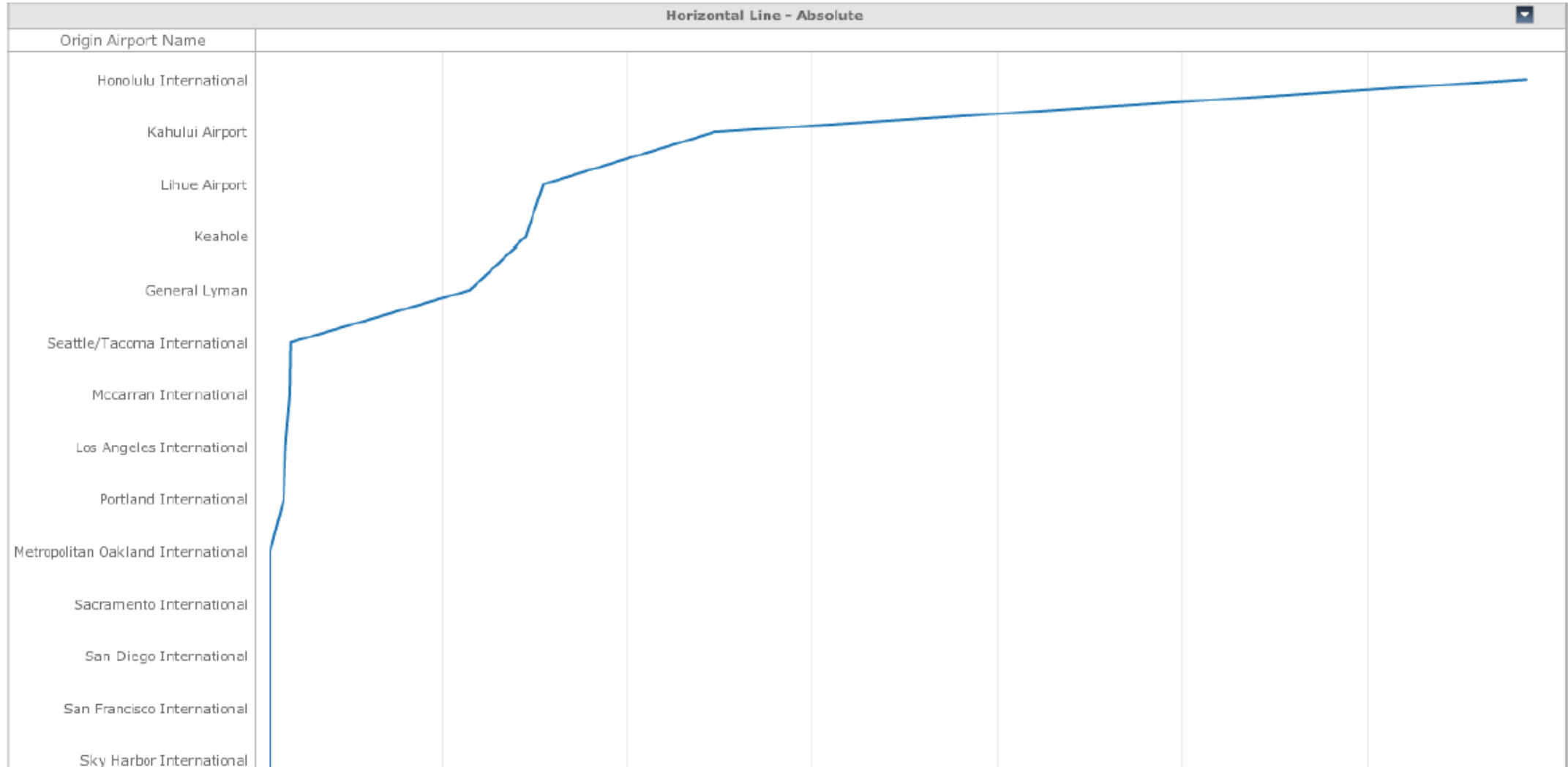
Attributes
(nominal) and
metrics

Comparative Analysis Bar Chart - Sorted



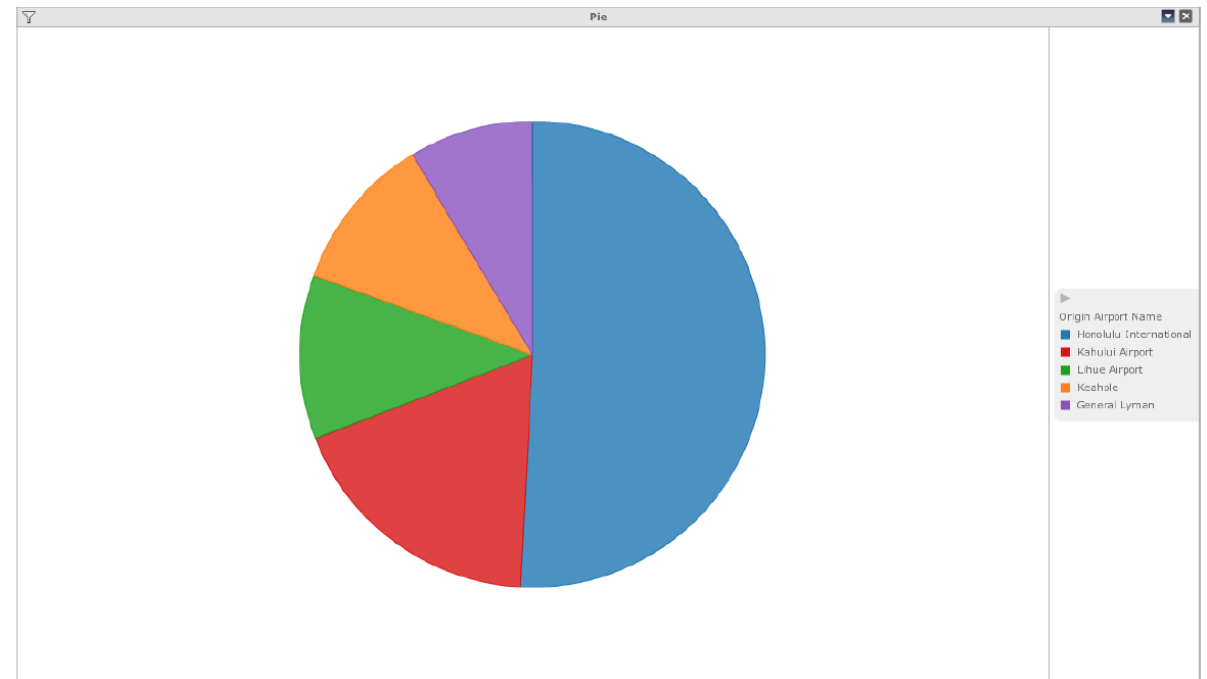
Comparative Analysis

Avoid: Line Chart – Implies continuity between points



Attribute (Nominal) and Metric

Contribution Analysis – Few Elements Pie Chart

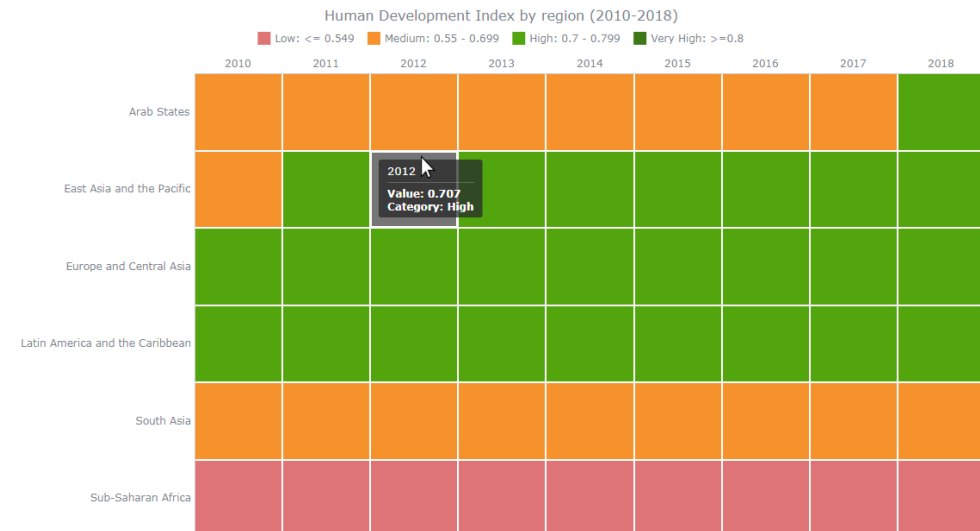
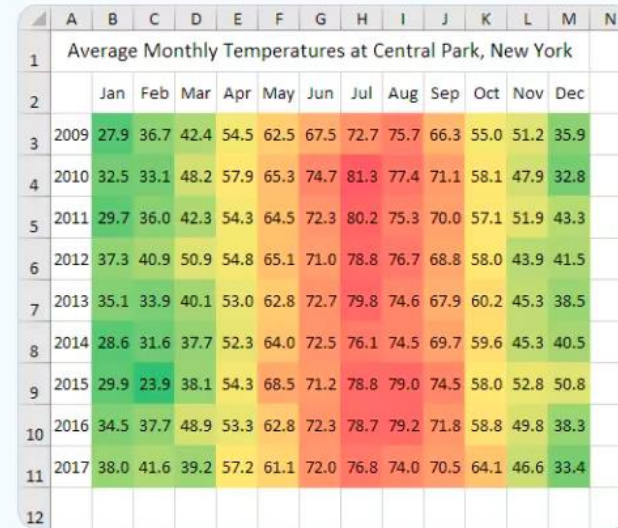


In a pie, colors indicate the values of the nominal attribute, the width of each slice the percentage (metric)

Nominal, ordinal) and Metric

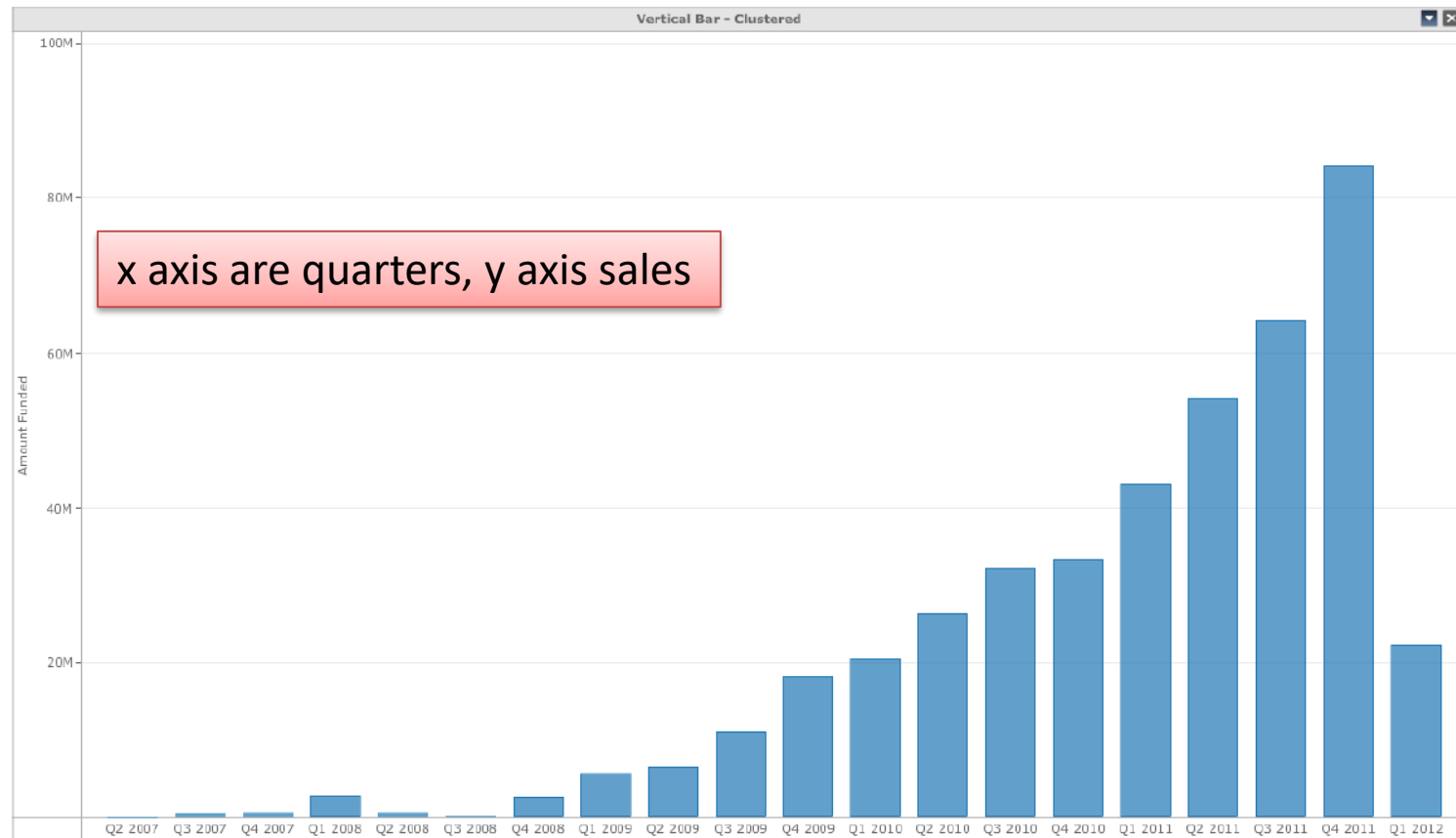
Many elements: heat map

X and Y are **attributes**,
either nominal or ordinal;
the "heat" (intensity of a
colour) symbolizes a **metric**,
which might or might not
be explicitly represented



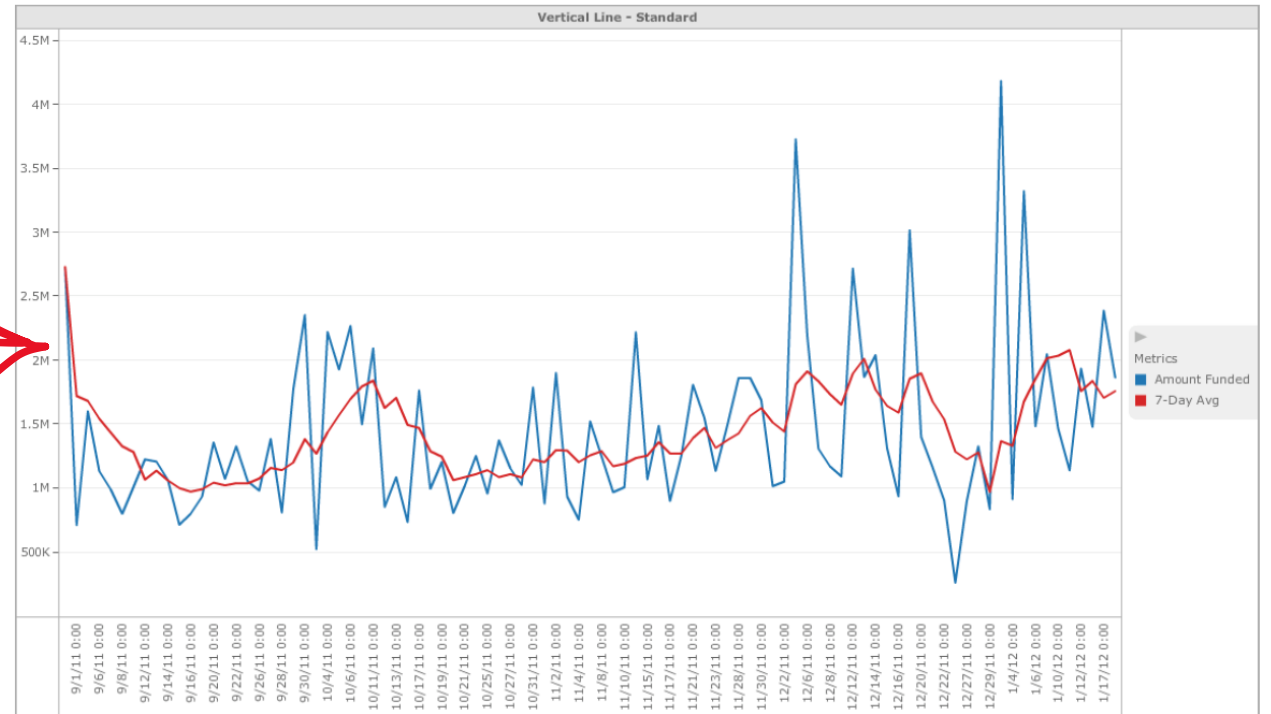
Attribute (Ordinal) and Metric

Time-Series Analysis – Few Elements Column Chart



Attribute
(Ordinal) and
Metric

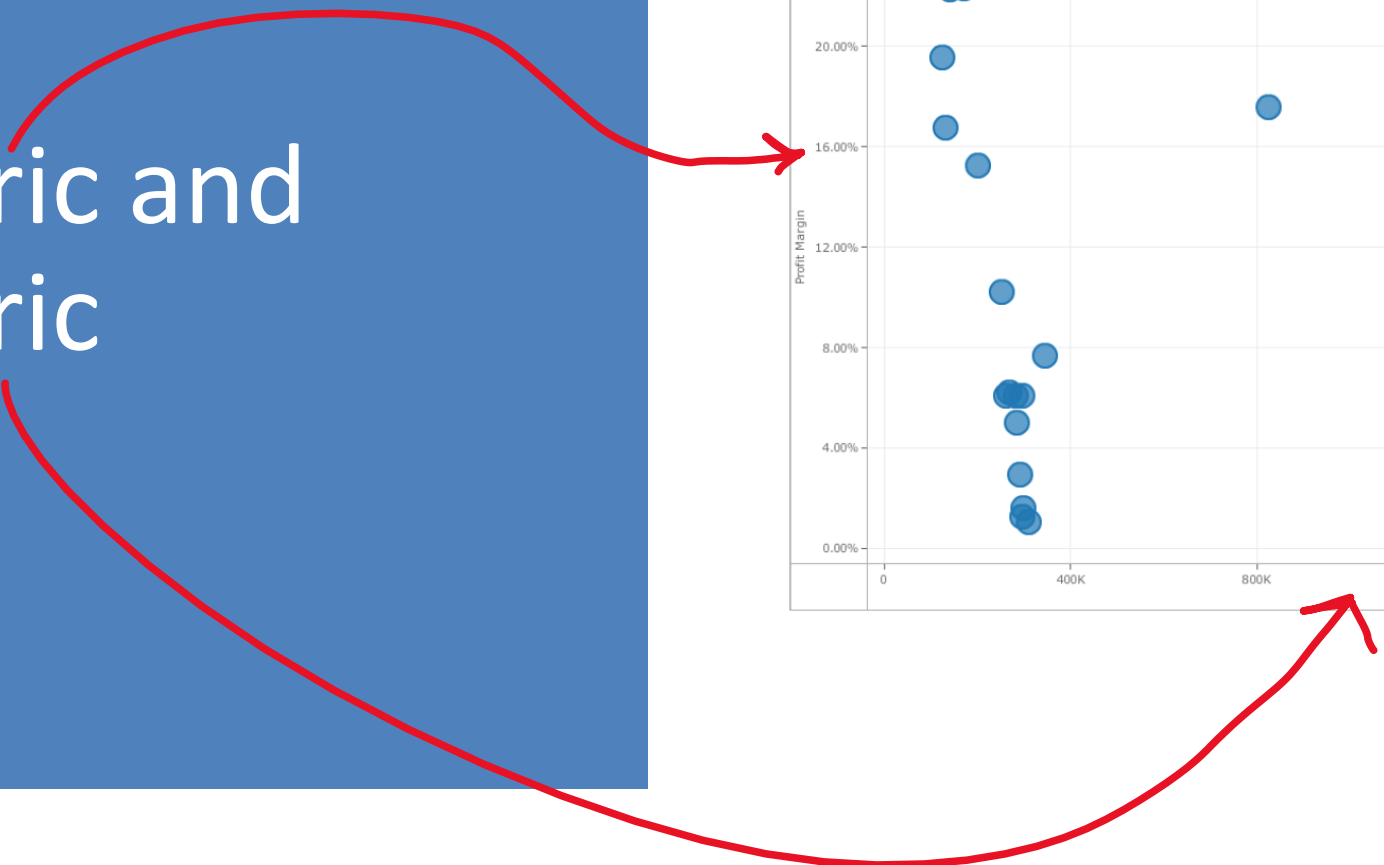
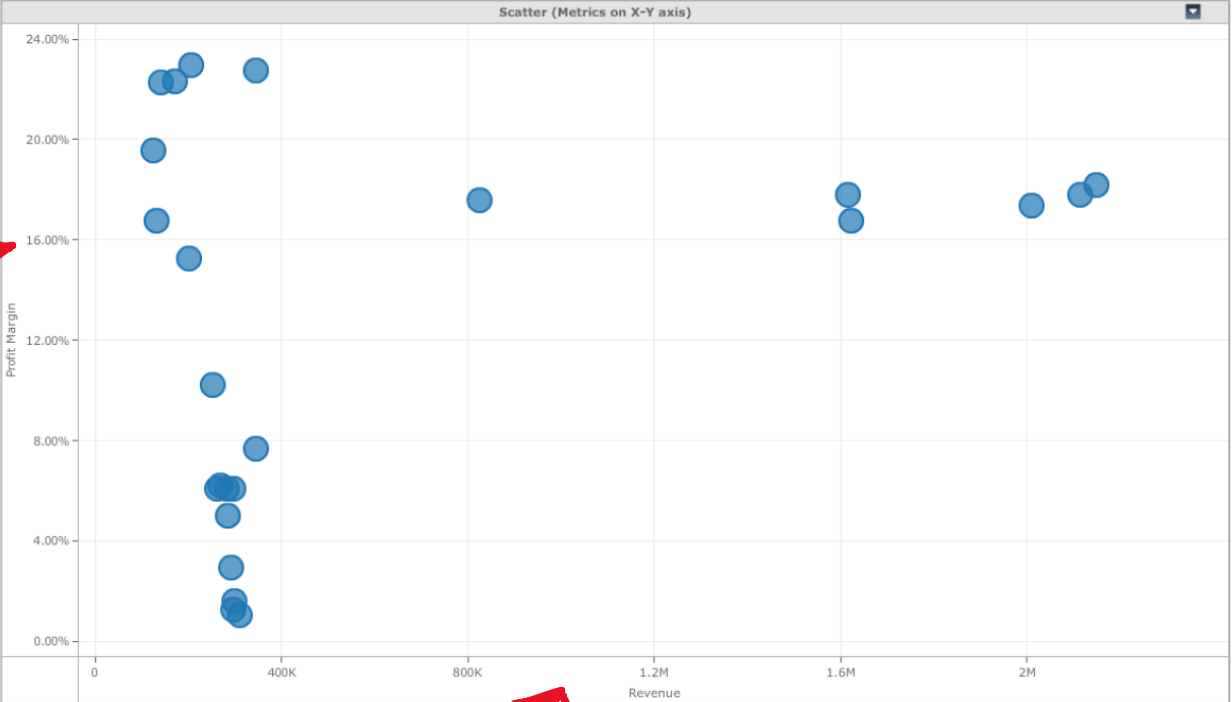
Time-Series Analysis – Many Elements Line Chart



Metric and
Metric

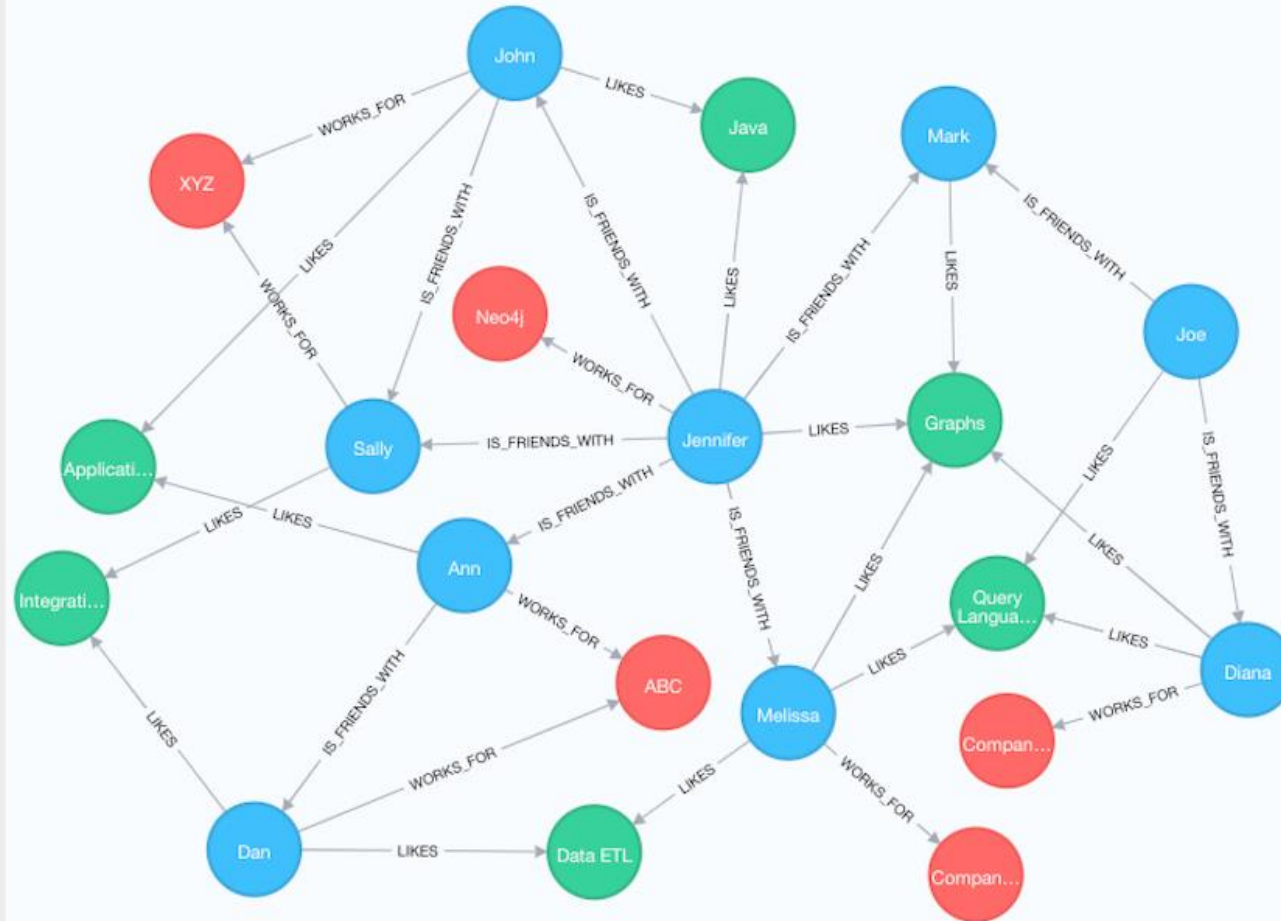
Correlation Analysis

Scatter Plot



Attribute (Nominal) and Attribute (Nominal)

The variables in a graph, or network, are circles (x) and edges (y). Here, values of circles are either persons or companies, and edges are types of relations





How to **improve** a visualization

- Colors
- Saturations
- Size
- Interpretability
- Performance
- Layout
- Interactivity



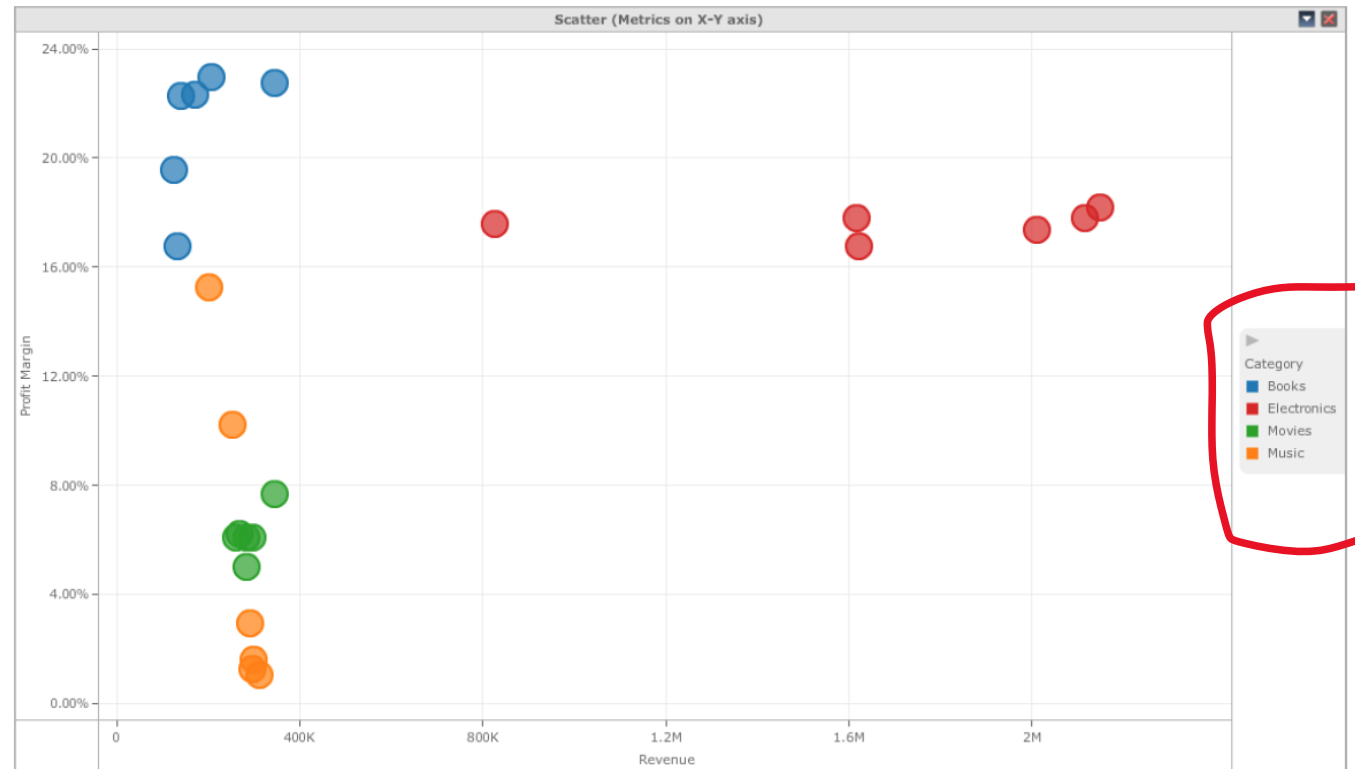
Enhancing Visualizations for Additional Insights

Appropriate Visual Enhancements

	Attribute (Nominal)	Attribute (Ordinal)	Metric
Color Hue	x	x	x
Color Saturation		x	x
Size		x	x

what colours, what intensity, what size?

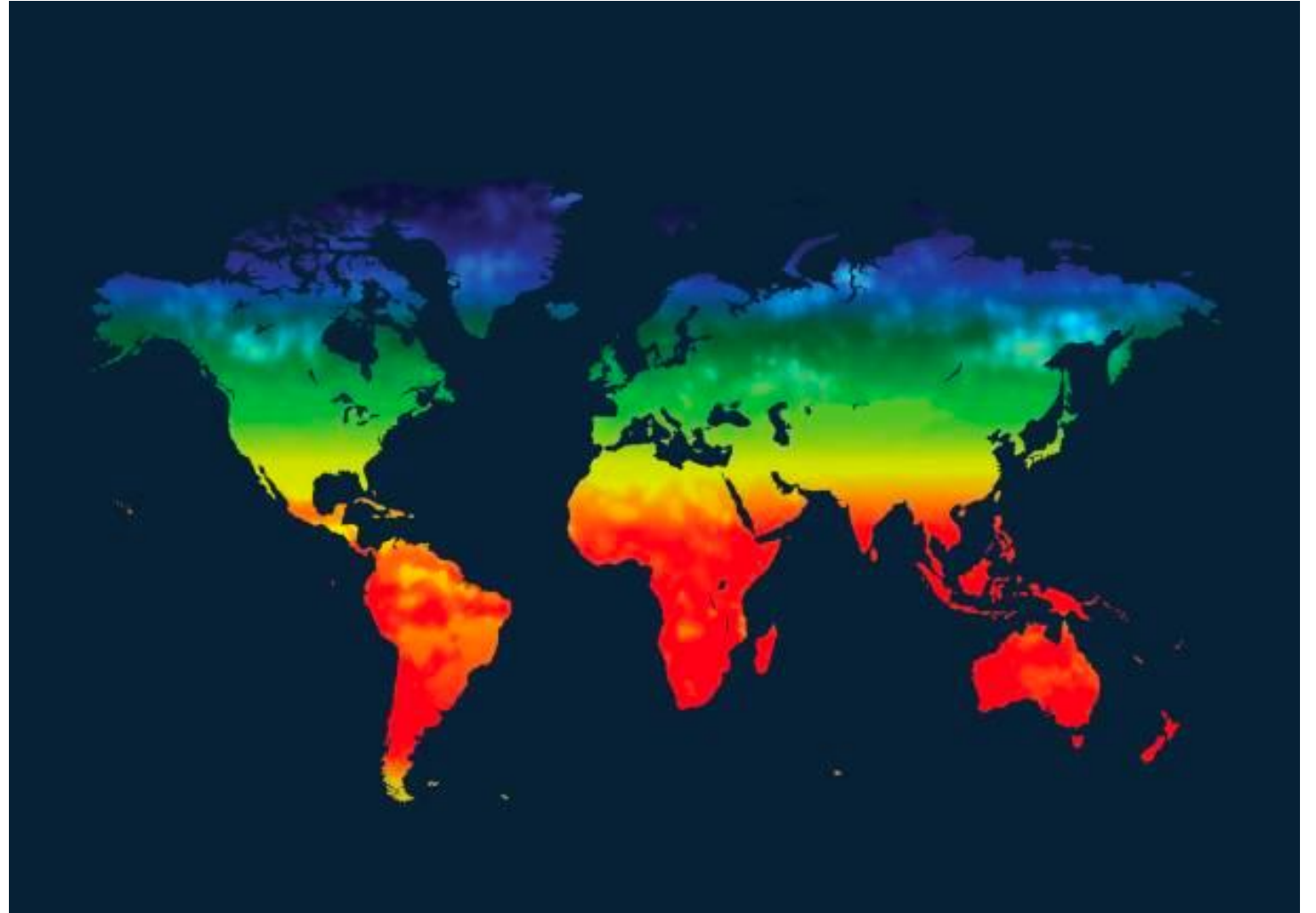
Color (Hue) to
Identify different
values of nominal
attributes



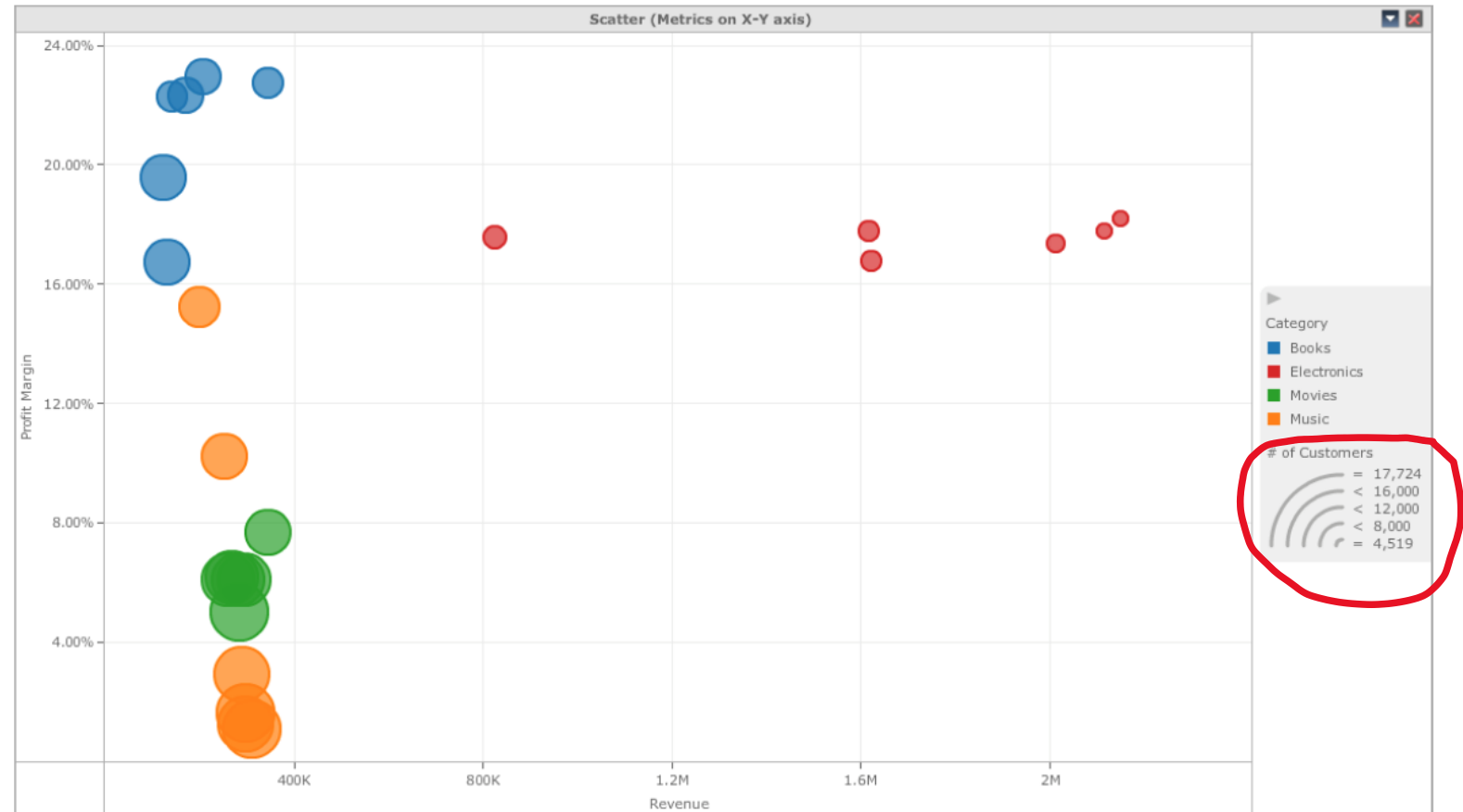
At first glance, we can tell the difference between groups identified by a nominal variable

Color (Saturation) to Highlight Metric Patterns

- Here, the metric is "temperature"
- At first glance, we can tell where the hottest regions are located



Adding Size to Emphasize Metric Trends

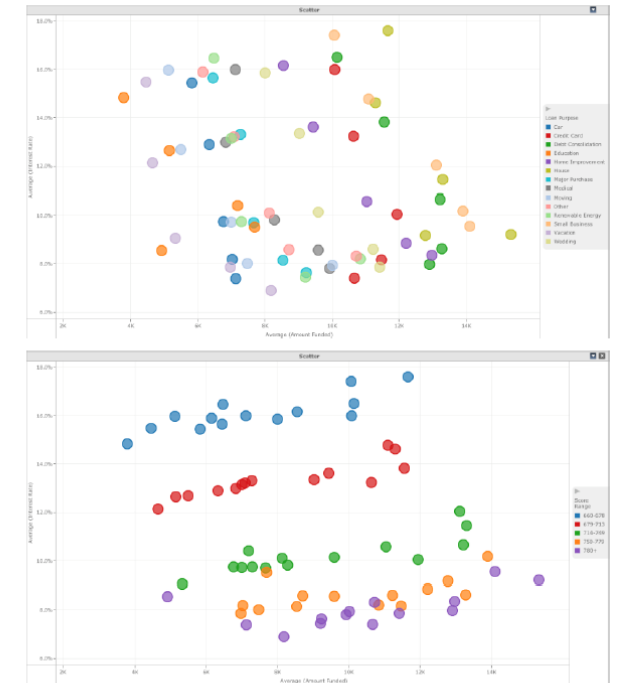


Size of balls indicates the number of customers. At first glance, we understand what are the biggest groups. Later, we might look for precise numbers

More hints on colors

Colors Should Enhance Data Comprehension, Not Distract

Use Fewer Than 6 Colors

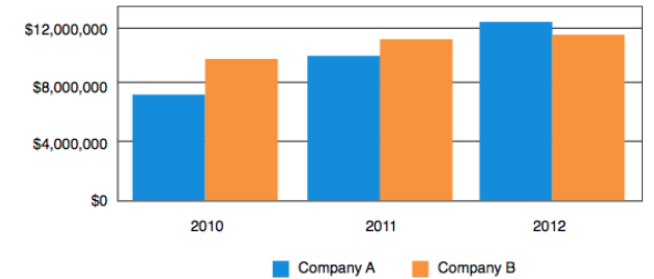
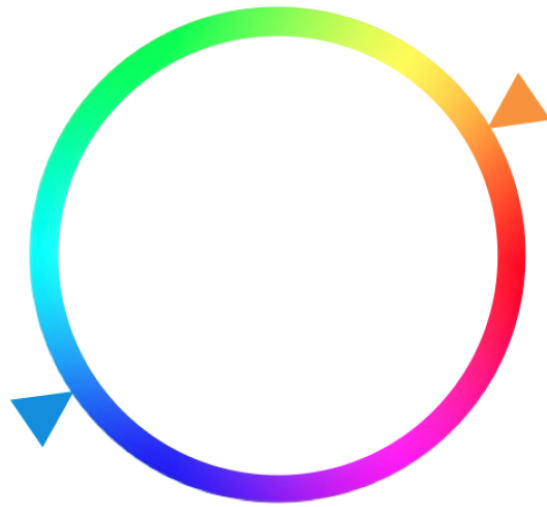


Otherwise the difference is difficult to perceive

More hints on colors

Use Colors to Emphasize Comparisons

Use Opposing Colors for Comparisons



More hints on colors

Avoid color confusion!

Why should we both change from our home kits?

Home Jerseys

Wales



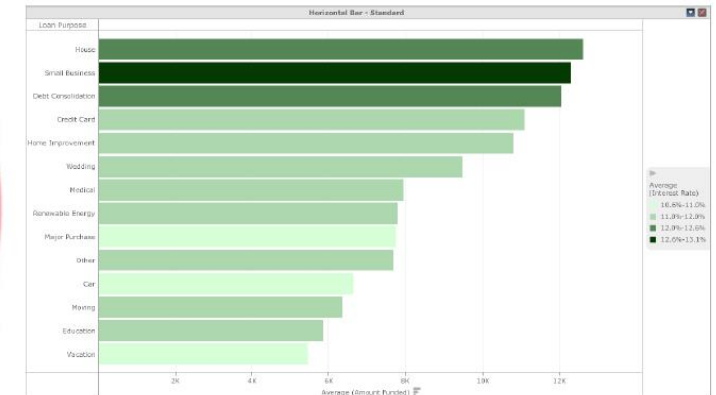
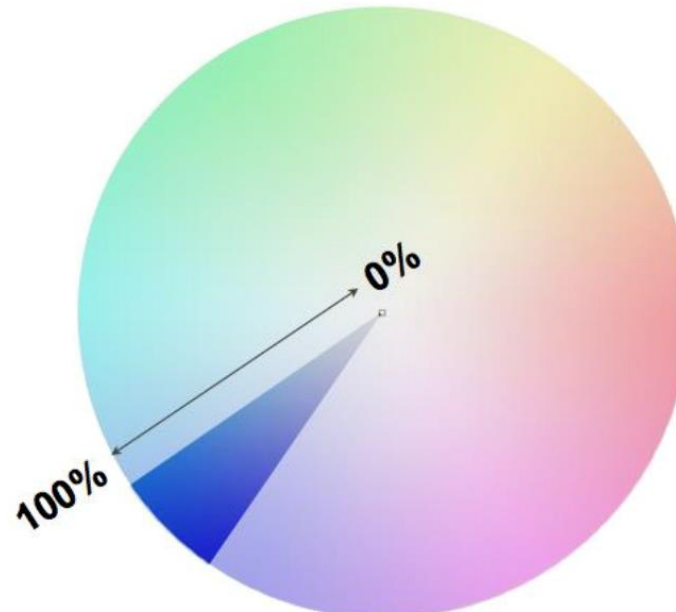
Portugal



More hints
on colors

Use Color Saturation Correctly

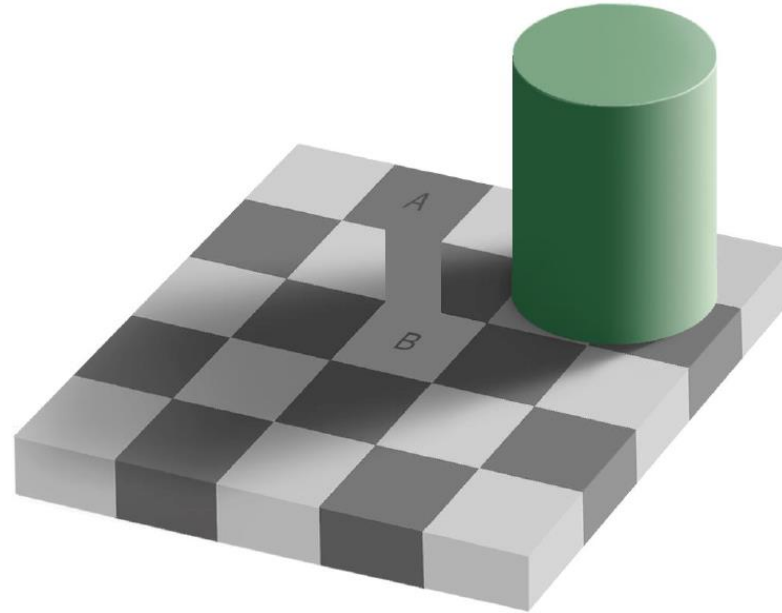
Less Saturation: Smaller Values
More Saturation: Greater Values



More hints on
colors

Color Constancy Can Confound Data Comprehension

Avoid Color Gradients for Backgrounds



Interpretability: avoid users do the math

Business questions

- Which months were below target?
- Which months were above target?
- And by how much?

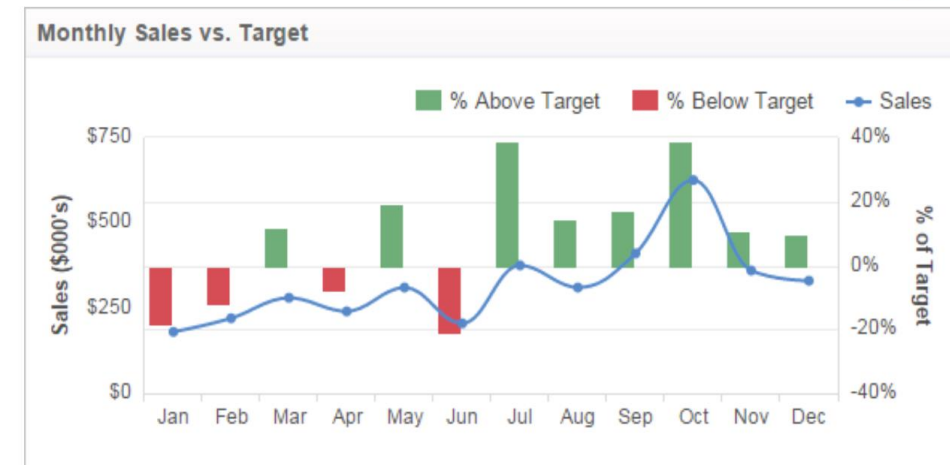


The graph answers these questions, but requires humans to "visually compute" the difference between the two curves

Interpretability: avoid users do the math

- Shows Actuals
- Shows Good and Bad months
- Quantifies good and bad
- Uses and Overlay and dual Y Axis.

This graph instead shows the same information in a much more intuitive way



Interpretability: sloppy labelling

Make it idiot proof, I mean “self explanatory”!

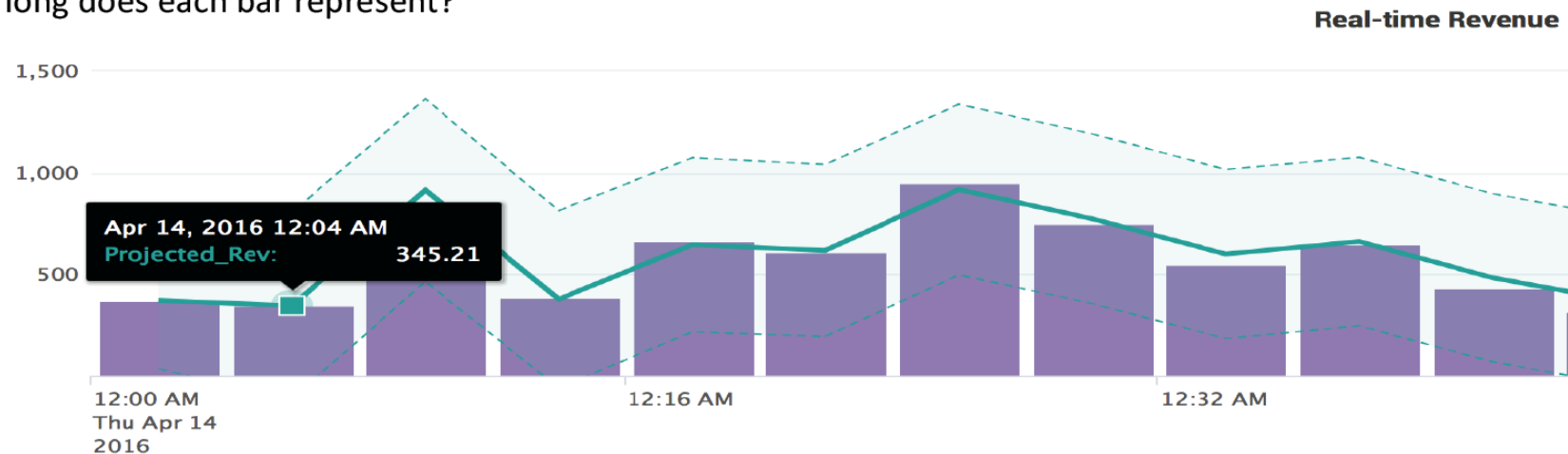
- What is the unit of measure on the Y axis?
- Are you using a log scale? (If so, mention it)
- Are the numbers shown in K's, M's or B's?
- If Currency, which currency?
- Net Revenue or Gross Revenue?
- How long is the rolling window being shown?
- How long does each bar represent?



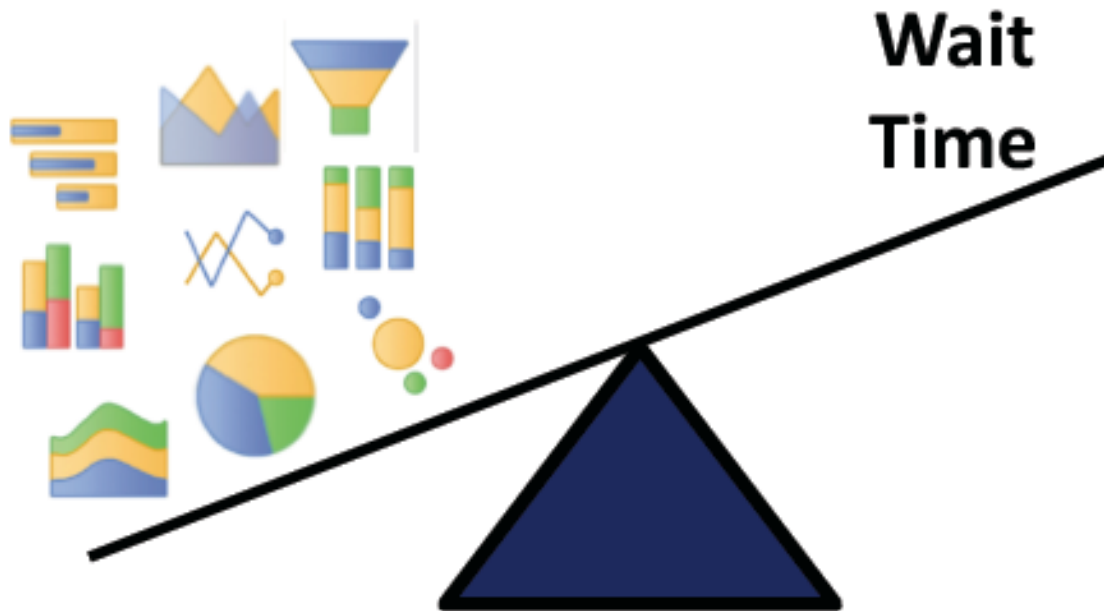
Real Time Revenue



Real Time Gross Revenue (\$) last 1 hour in 4 minute buckets



Performance: reduce load times

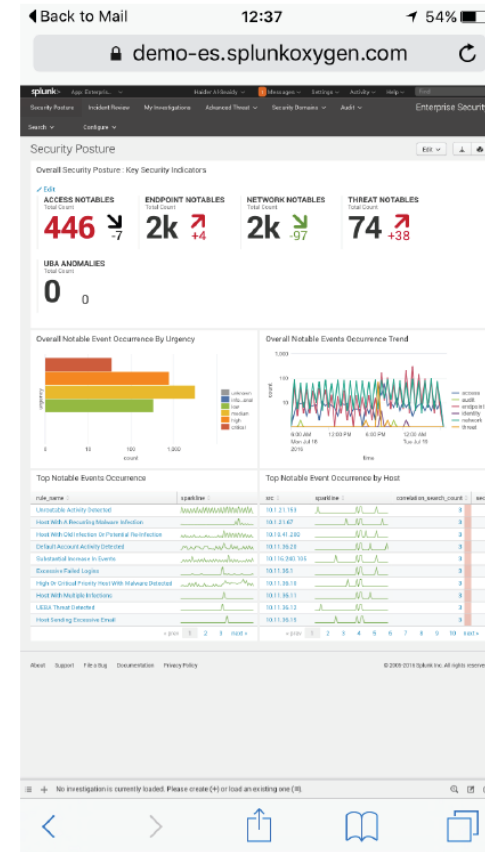
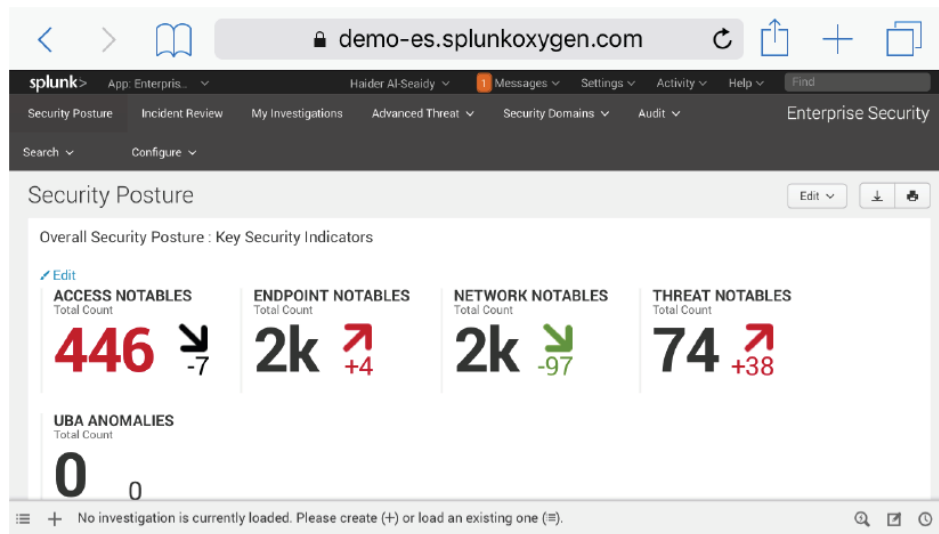


- **Limit the number of objects on a single screen**
- Limit real-time searches
- Specify filters to reduce the data
- Use **summary indexes** to reduce the search load of the dashboard
- **Arrange in a single screen related data** (e.g. all plots that are meant to identify gender differences, or differences among point of sales)

Layout: size to the right width

Orientation

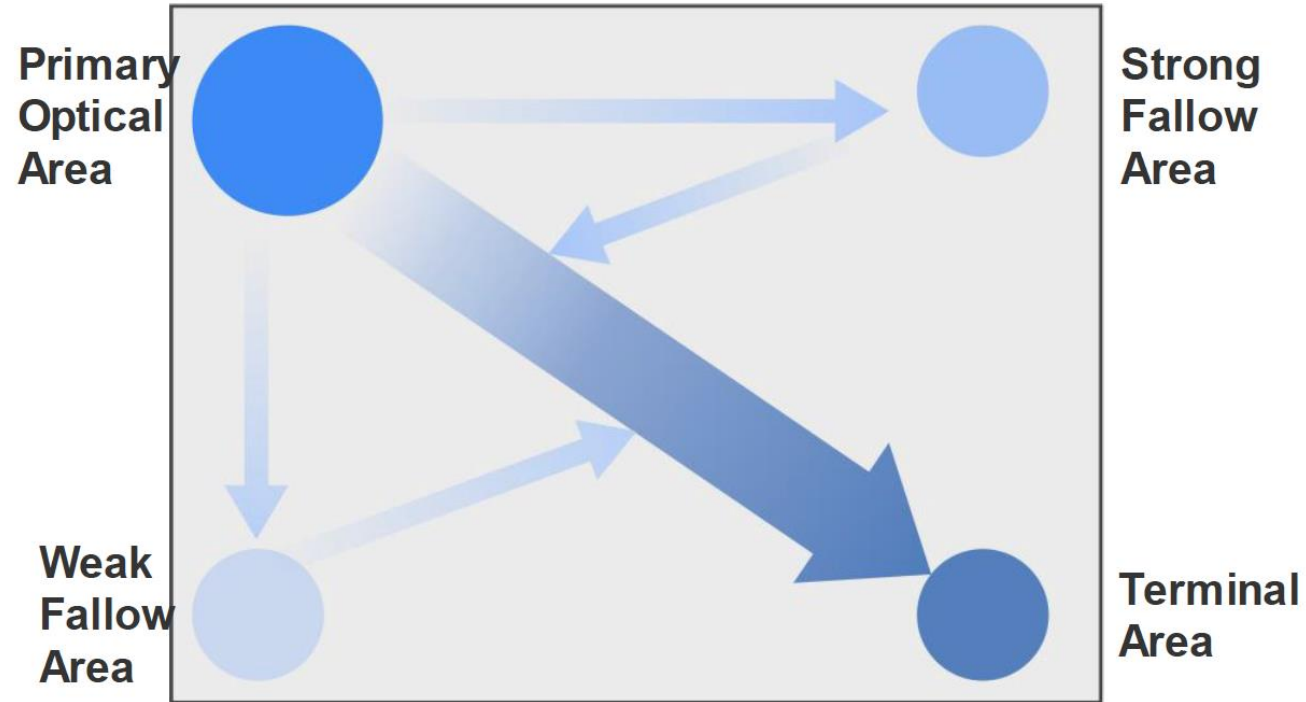
- Landscape mode on a mobile will require scrolling to see content
- Are your key metrics at the top?
- Sizes to the width of the app



Layout:

People Have a Bias in How They Read and Scan Content (this depends on cultures, of course)

Reading Gravity



Layout: place most relevant content in primary optical area

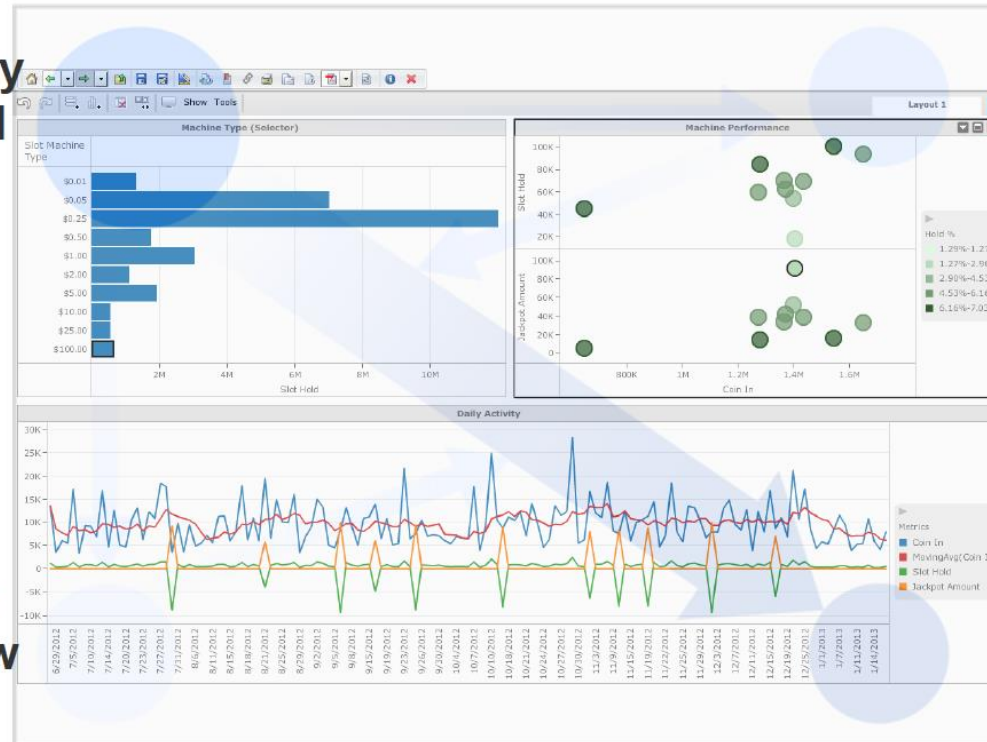
Reading Gravity

Primary
Optical
Area

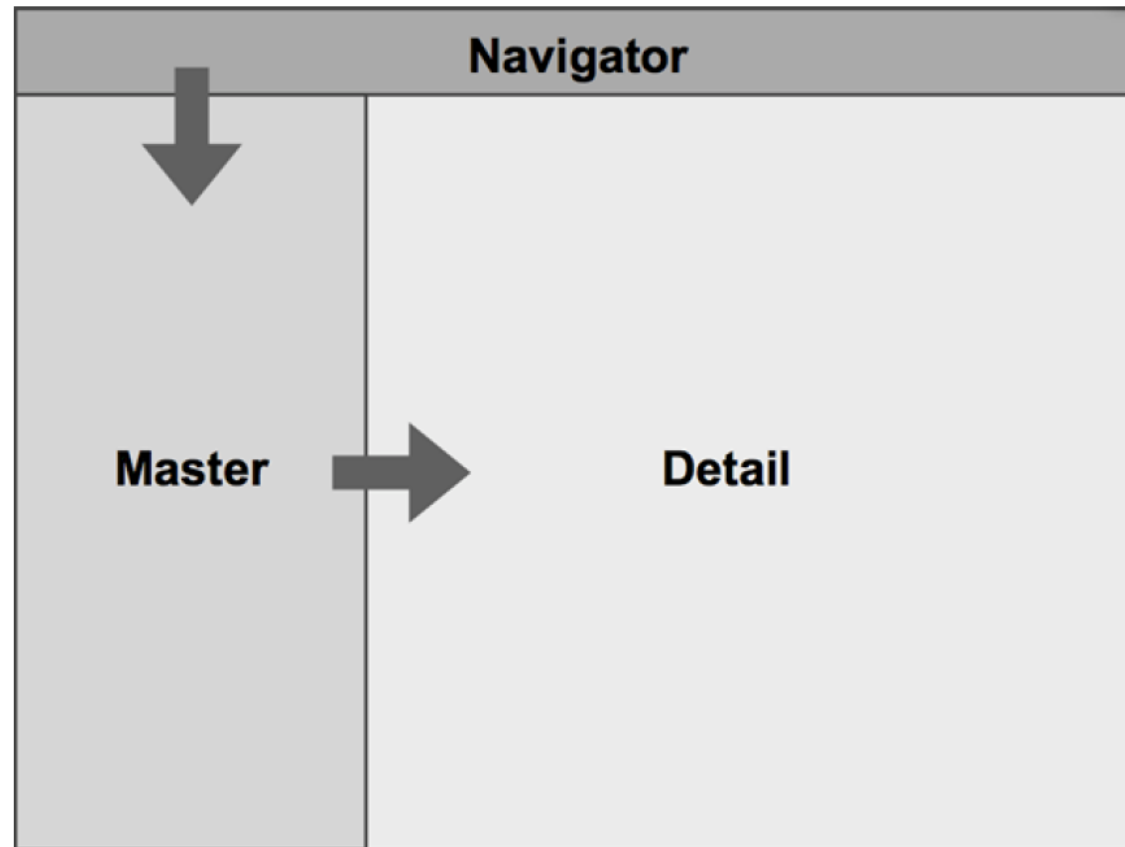
Strong
Fallow
Area

Weak
Fallow
Area

Terminal
Area

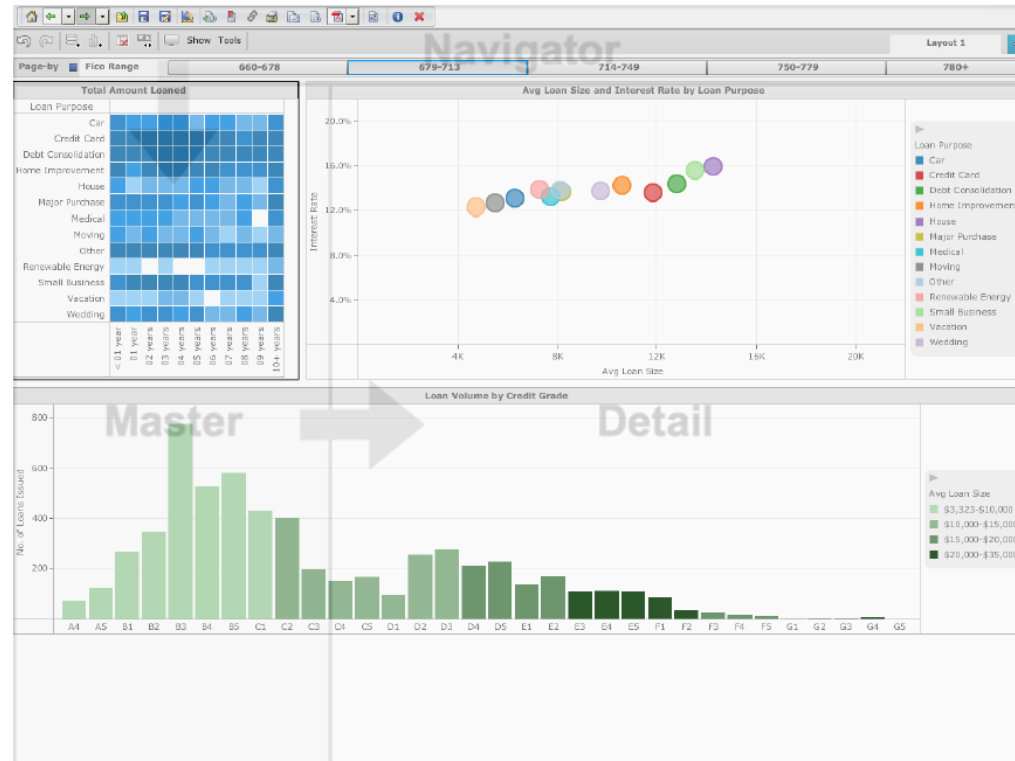


Layout: present content **hierarchically**



Layout: present content hierarchically

Present Data Hierarchically

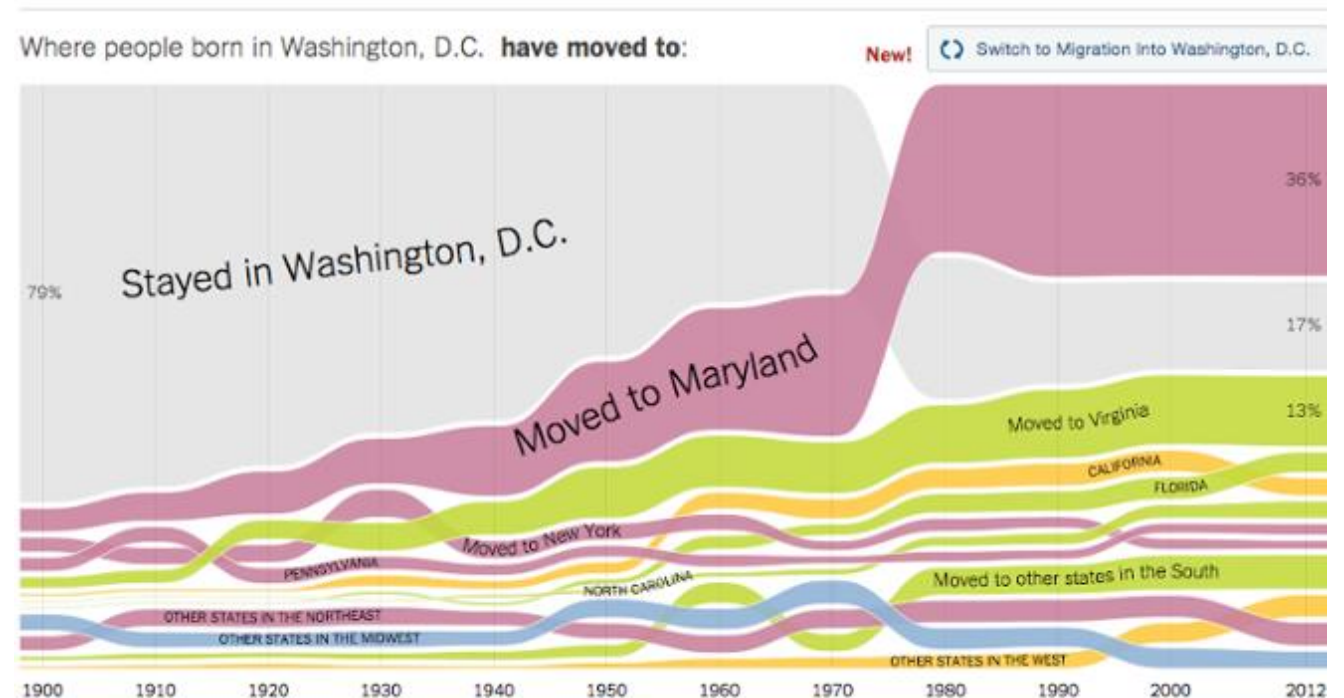


Interactivity & animation

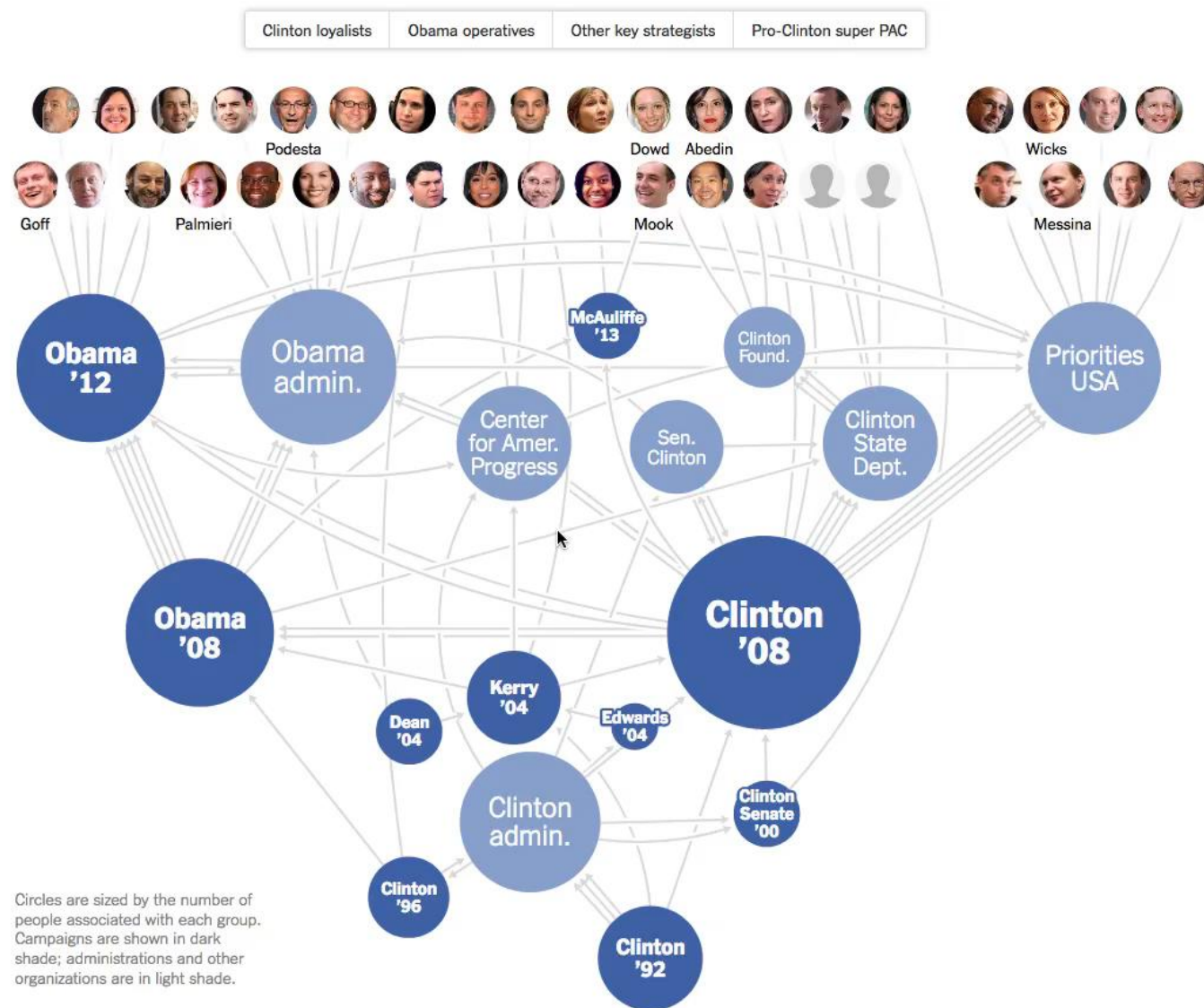
- Interactivity and animation are the latest and coolest features for presenting information
- User can adapt a visualization to his/her own needs and curiosity, interacting with the map
- Interactive maps –when well designed – greatly improve the efficacy of an interaction
- See here some inspiring example <https://infogram.com/blog/map-examples-from-the-web/>

The New York Times' project on where people born in a state move to

- It visualizes a large amount of data accumulated during more than
- 100 years. Yet, it is easy to understand, and it clearly highlights
- interesting trends.

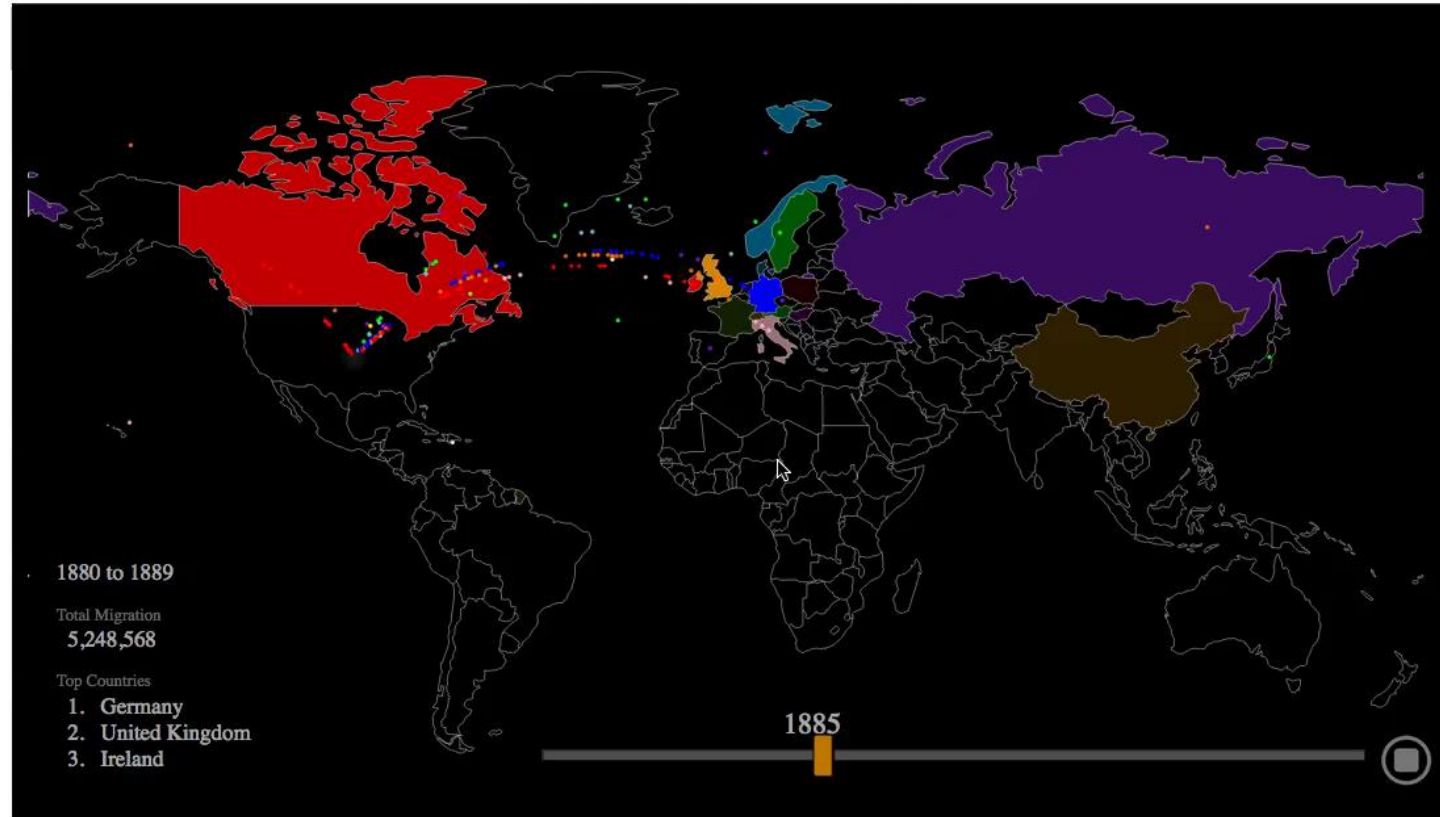


Politicians & Political campaigns



Immigration

Two Centuries of U.S. Immigration (1 dot = 10,000 people)



[Full screen interactive map / HD video](#)

International Trades

Click on a country to see its share of trade alone, or spin/navigate the globe by using your mouse.



Interactive: Mapping the Flow of International Trade

SPONSORED INFOGRAPHICS



Demos and examples

<https://www.highcharts.com/demo>

<https://www.datapine.com/blog/best-data-visualizations/>

Tools to create nice visualizations

- In addition to watson, for your project's data analytics you can use this free tool
- <https://flourish.studio/>

Flourish

- Go to <https://flourish.studio/> and register for free
- Choose a template among the very many
- For every template, an explanation of the use is provided

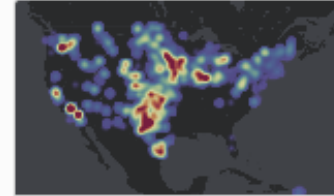
3D map

Map for displaying regions, lines and points with optional time slider

STARTING POINTS ?



Animated points



Heatmap



Lines



Point map v

Hierarchy

A template for grouping data and visualising it hierarchically

STARTING POINTS ?



Hierarchical bars



Packed circles



Radial tree



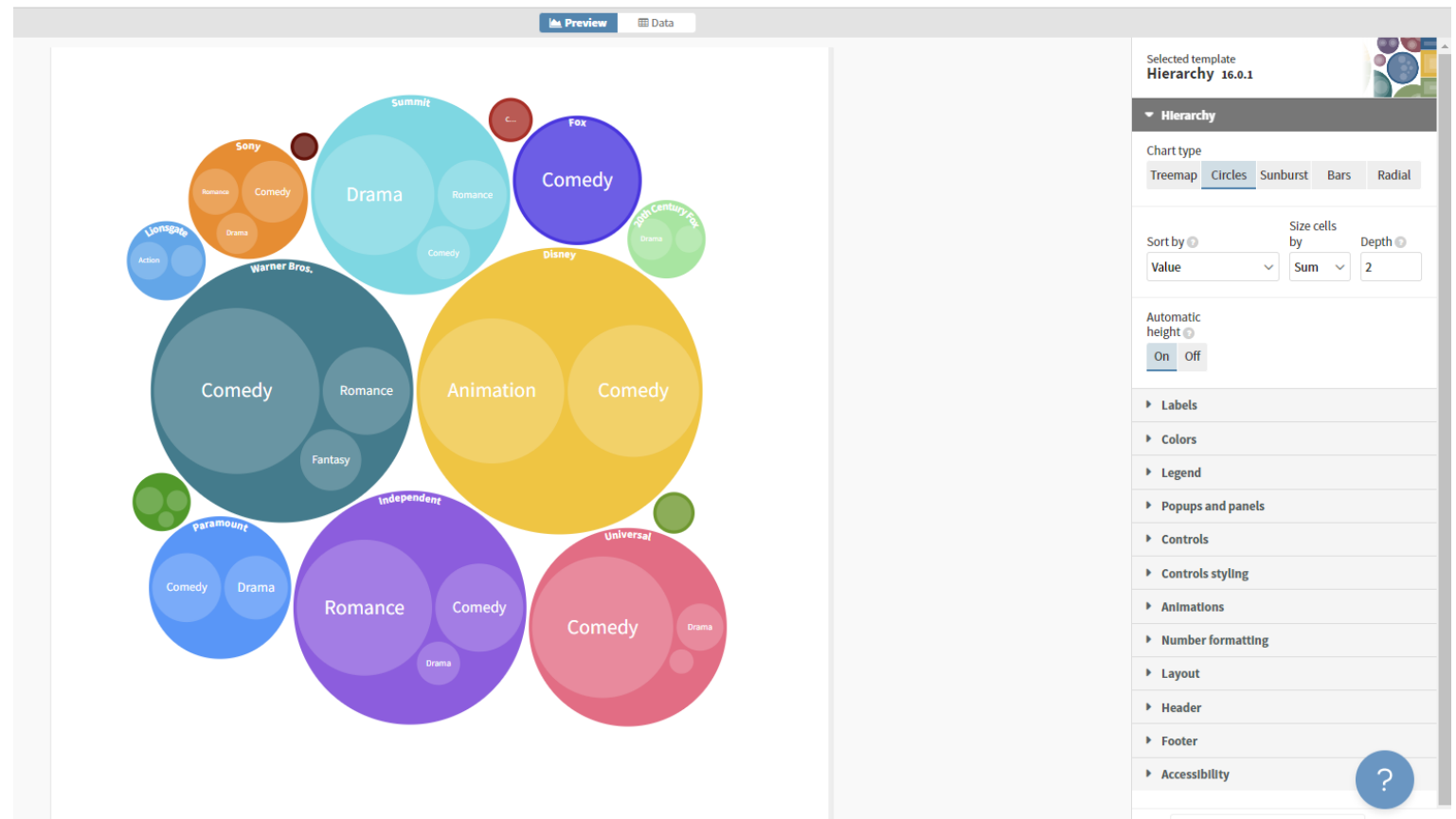
St

Create some visualization with Flourish (use sample data first)

Creating a visualization in Flourish is a matter of seconds!

1. To create a visualization, go to your projects page (create a project page first), and click: **NEW VISUALIZATION**
2. This will bring up **the Flourish template chooser**, which will show you some visualization templates (example: area charts).
3. Select, e.g., the Area chart (stacked) as a **starting point**.
 1. TIP: Starting points are just **examples** made with a template. Once you open a starting point, you can adjust the template-specific settings, so your project resembles a different starting point.
 2. TIP: Not sure which template is the right choice for your data? Check out our help guide for more information.
4. Choosing a template opens the visualization editor. Each Flourish template comes with some sample data, so you can see the template in action.
5. The visualization editor has four main parts:
 - **Preview tab** – this is where you can see what your project will look like when it's published.
 - **Data tab** – this is where **you should upload your own data**. If you edit or replace any information within these cells, your visualization will instantly reflect these changes.
 - Some Data tabs have more than one data sheet. If you are not sure how your data should be structured, or what each sheet affects in the visualization, we recommend reading the template-specific help docs. You can also access them through the chat symbol in the bottom right-hand corner of every page!

Example: select a template and explore options to the right



See the data and explore options

Untitled visualisation
by Paola Velardi Private

Create a storyExport & publish

PreviewData

Data

Data

	A	B	C	D	E	F	G	H
1	Film	Genre	Lead Studio	Audience score %	Profitability	Rotten Tomatoes %	Worldwide Gross	Year
2	27 Dresses	Comedy	Fox	71	5.3436218	40	160.308654	2008
3	(500) Days of Summer	Comedy	Fox	81	8.096	87	60.72	2009
4	A Dangerous Method	Drama	Independent	89	0.44864475	79	8.972895	2011
5	A Serious Man	Drama	Universal	64	4.382857143	89	30.68	2009
6	Across the Universe	Romance	Independent	84	0.652603178	54	29.367143	2007
7	Beginners	Comedy	Independent	80	4.471875	84	14.31	2011
8	Dear John	Drama	Sony	66	4.5988	29	114.97	2010
9	Enchanted	Comedy	Disney	80	4.005737082	93	340.487652	2007
10	Fireproof	Drama	Independent	51	66.934	40	33.467	2008
11	Four Christmases	Comedy	Warner Bros.	52	2.022925	26	161.834	2008
12	Ghosts of Girlfriends Past	Comedy	Warner Bros.	47	2.0444	27	102.22	2009
13	Gnomeo and Juliet	Animation	Disney	52	5.387972222	56	193.967	2011
14	Going the Distance	Comedy	Warner Bros.	56	1.3140625	53	42.05	2010
15	Good Luck Chuck	Comedy	Lionsgate	61	2.36768512	3	59.192128	2007
16	He's Just Not That Into You	Comedy	Warner Bros.	60	7.1536	42	178.84	2009
17	High School Musical 3: Senior Year	Comedy	Disney	76	22.91313646	65	252.044501	2008
18	I Love You Phillip Morris	Comedy	Independent	57	1.34	71	20.1	2010
19	It's Complicated	Comedy	Universal	63	2.642352941	56	224.6	2009
20	Jane Eyre	Romance	Universal	77		85	30.147	2011
21	Just Wright	Comedy	Fox	58	1.797416667	45	21.569	2010

1

more rows

Upload data

Data

SELECT COLUMNS TO VISUALISE

Categories/nesting C-A

Size by G

Filter

Info for popups

Uploading your data

Today just play with Flourish
data, but in your final project,
you can upload your own data

[illegible]



Homework

- Select a dataset of your choice
 - Use [Flourish](#) to generate 5 visualizations, of which at least
 - 1 comparison
 - 1 whole-parts analysis
 - 1 trend analysis
 - 1 visualization to understand relations between at least 3 variables (2 metrics, one attribute)
-