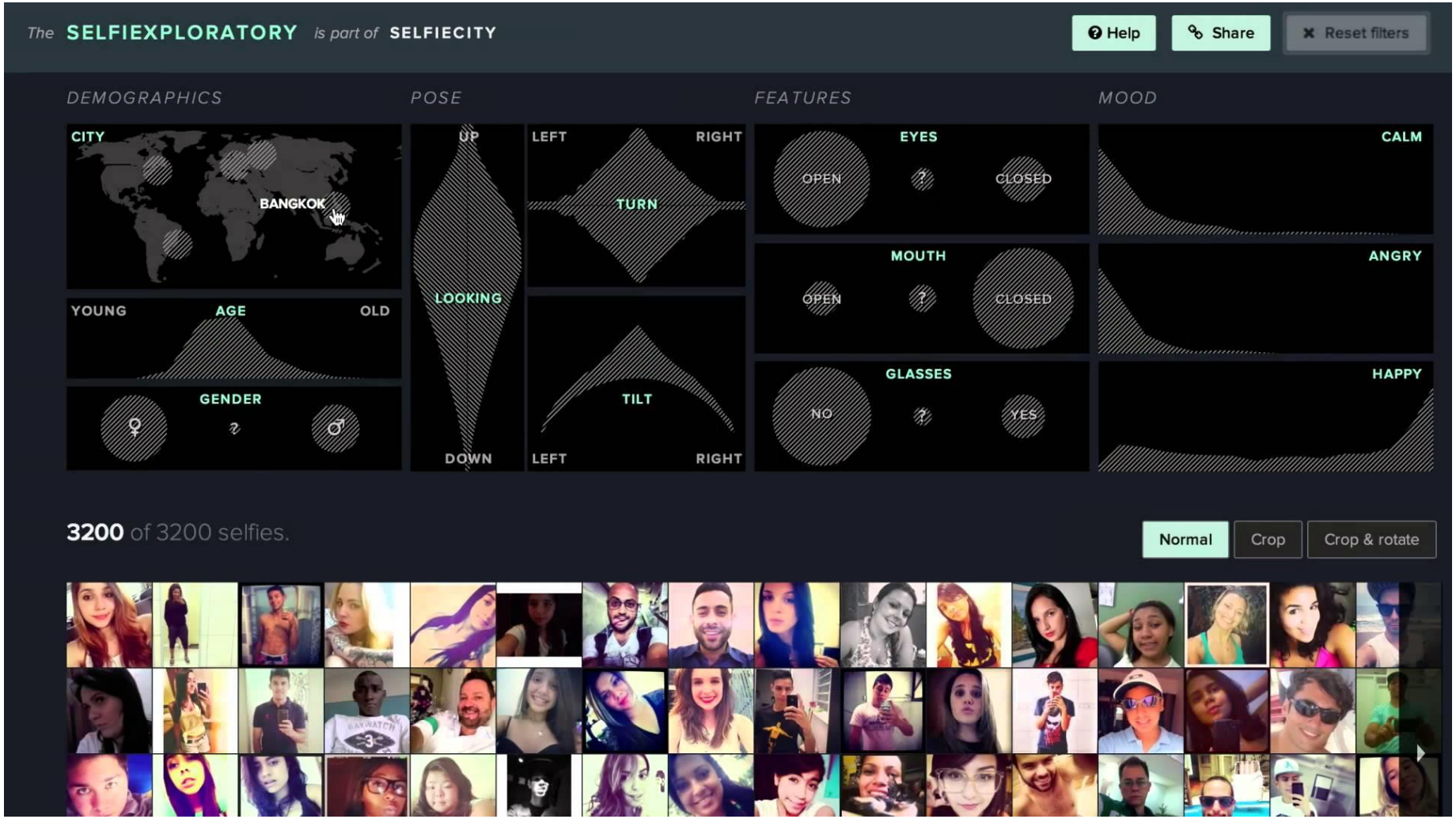
Data Visualization

Introduction

- Data visualization is a graphic representation that expresses the significance of data.
- It reveals insights and patterns that are not immediately visible in the raw data.
- Data visualization produces visual representation of abstract data to reinforce human cognition
- Ideal visualization shows the right amount of data, in the right order, in the right visual form, to convey the high priority information.

SELFIECITY



PC: https://cdnl.tblsft.com/sites/default/files/pages/9_selfiecity.jpg

http://selfiecity.net/

Goal of data visualization

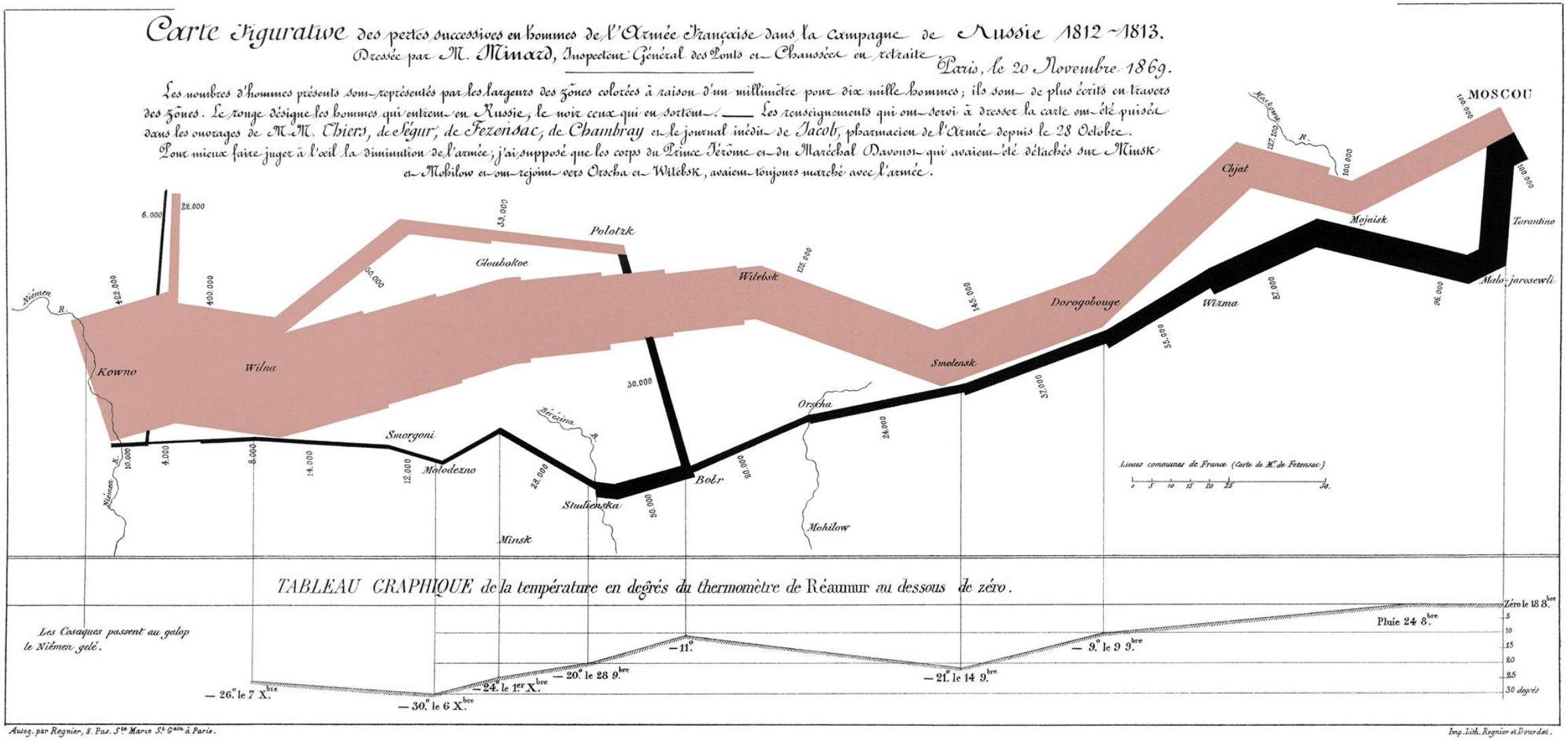
- Communicate information
- Make Decision
- To see data in context
- To convey ideas effectively
- Tell a story
- Inspire

Visual Data Representation

- Infographics
- Data Visualization

Charles Joseph Minard: Napoleon's Retreat From Moscow (The Russian Campaign 1812-1813)

An Interactive Chart



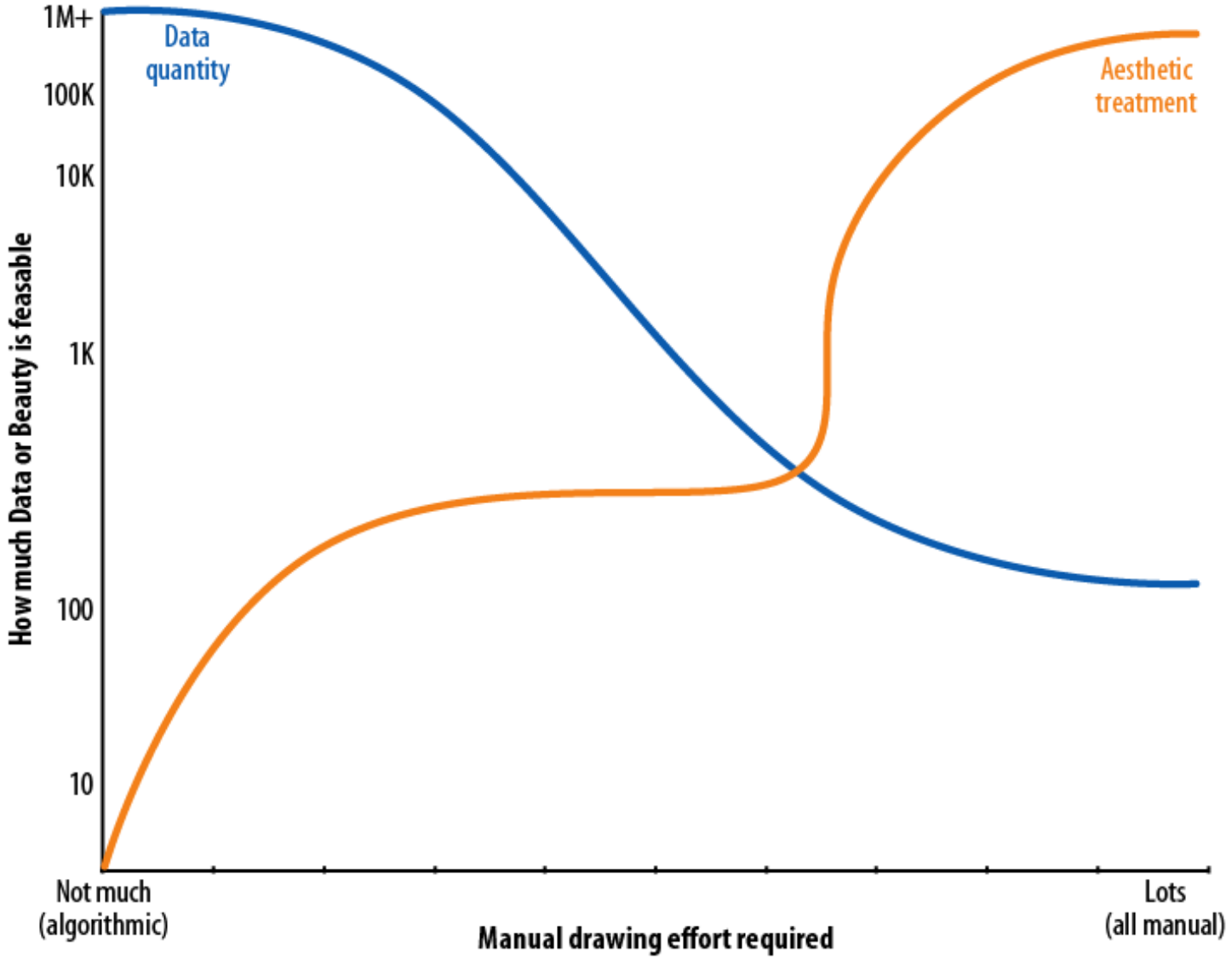
PC: https://masswerk.at/minard/images/minard-campagne-de-russie-1812_hr.png

Visualizing Global Health Data

https://youtu.be/hVimVzgtD6w

Infographics vs Data Visualization

- Infographics
  - manually drawn
  - specific to the data at hand
  - aesthetically rich
  - relatively data-poor
- Data Visualization
  - Algorithmically drawn
  - Easy to regenerate with different data
  - Often aesthetically barren
  - Relatively data-rich



PC: https://www.oreilly.com/library/view/designing-data-visualizations/9781449314774/httpatomoreillycomsourceoreillyimages898012.png

Why data visualization is powerful tool

- Intuitive: Presenting a graph instantly makes sense, even to people who have never worked with graphs before.
- Fast: It is fast because our brains are great at identifying patterns, but only when data is presented in a tangible format
- Flexible: The world is densely connected, so as long as there is an interesting relationship in your data somewhere, you will find value in graph visualization.
- Insightful: Exploring graph data interactively allows users to gain more in-depth knowledge, understand the context and ask more questions, compared to static visualization or raw data.

When

- Data can be presented in the form of rectangular tables, or it can be presented in colorful graphs of various types.
- Edverd Tufte says that “Small, non-comparative, highly-labeled data sets usually belong in tables”.
- As the amount of data grows, graphs are preferable.
- Graphics help give shape to data.

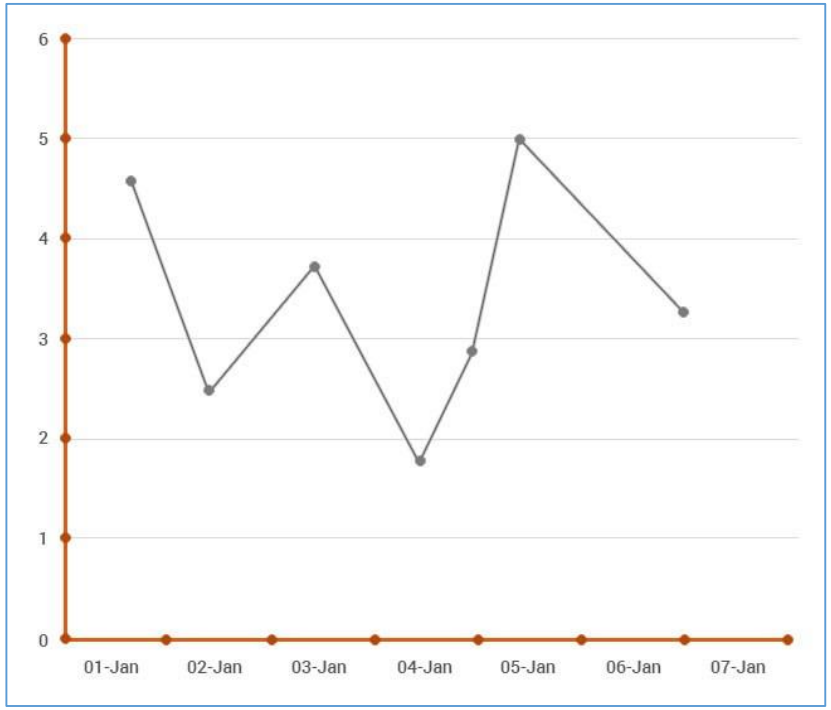
Types of Charts

Charts

- Line graph
- Scatter plot
- Bar graph
- Stacked Bar graphs
- Histograms
- Pie charts
- Box charts
- Bubble Graph
- Dials
- Geographical Data maps
- Pictographs

Line graph

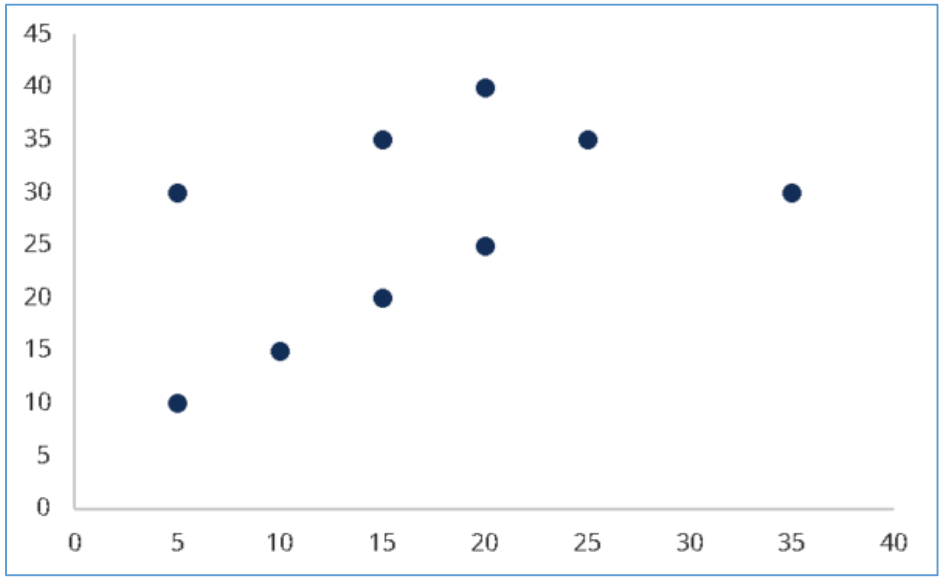
- Basic and most popular type of displaying information.
- Shows data as a series of points connected by straight line segments.
- If mining with time-series data, time is usually shown on the x-axis. Multiple variables can be represented on the same scale on y-axis to compare of the line graphs of all the variables.



PC: https://www.edelweiss.in/ewwebimages/WebImages/Learner/Line_Chart_Stocks~b2869c5e-d36c-4bdb-80e1-07d4805e70e0.jpg

Scatter plot

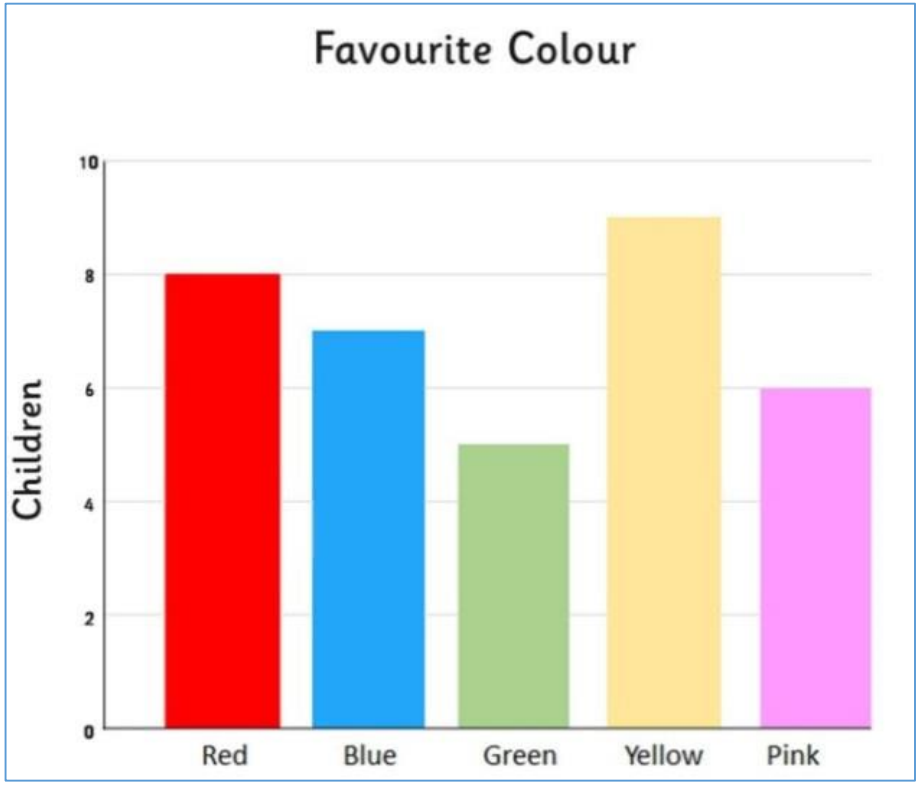
- A scatter plot is a chart type that is normally used to observe and visually display the relationship between variables.
- The values of the variables are represented by dots.
- The positioning of the dots on the vertical and horizontal axis will inform the value of the respective data point
- Hence, scatter plots make use of Cartesian coordinates to display the values of the variables in a data set.



PC: https://cdn.corporatefinanceinstitute.com/assets/scatter-plot2-600x367.png

Bar graph

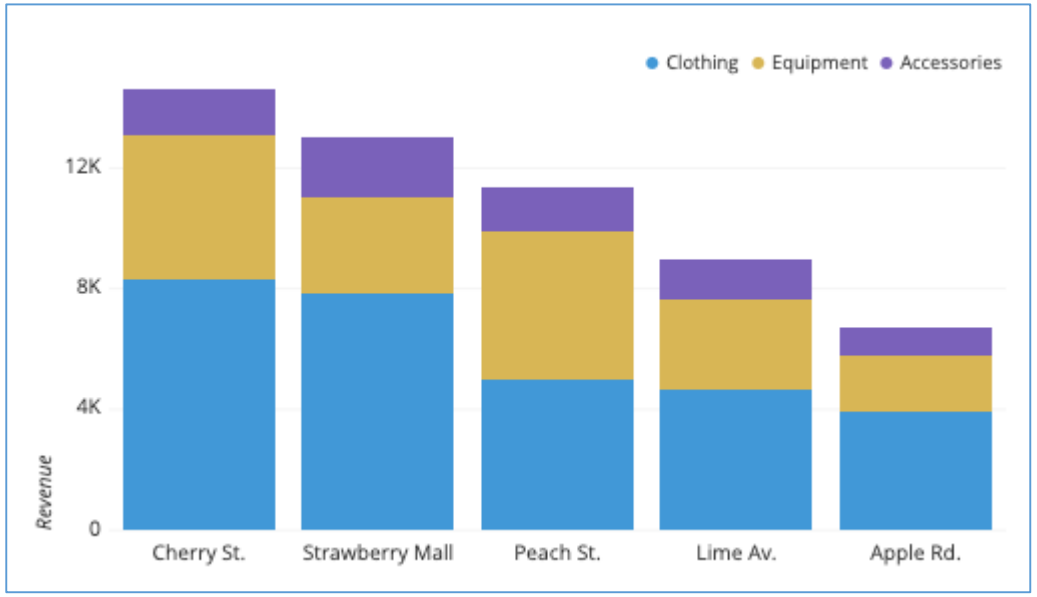
- A bar graph shows thin colorful rectangular bars with their lengths being proportional to the values represented.
- The bars can be plotted vertically or horizontally.
- The bar graphs use a lot of more ink than the line graph and should be used when line graphs are inadequate



PC: https://images.twinkl.co.uk/tw1n/image/private/t_630/u/ux/barchart_ver_1.jpg

Stacked Bar graphs

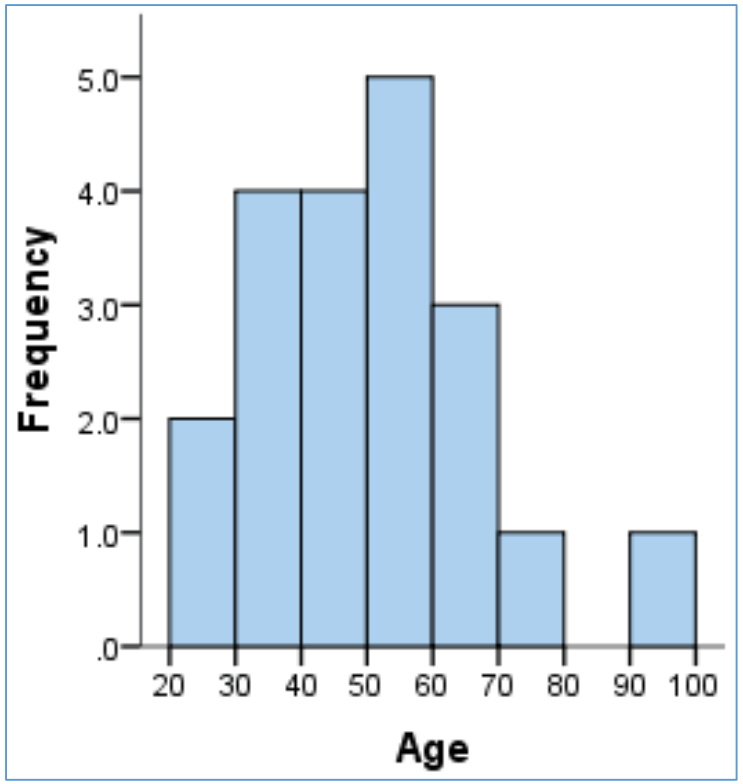
- These are a particular method of doing bar graphs. Values of multiple variables are stacked one on top of the other to tell an interesting story.
- Bars can also be normalized such as the total height of every bar is equal, so it can show the relative composition of each bar.



https://chartio.com/learn/charts/stacked-bar-chart-complete-guide/

Histograms

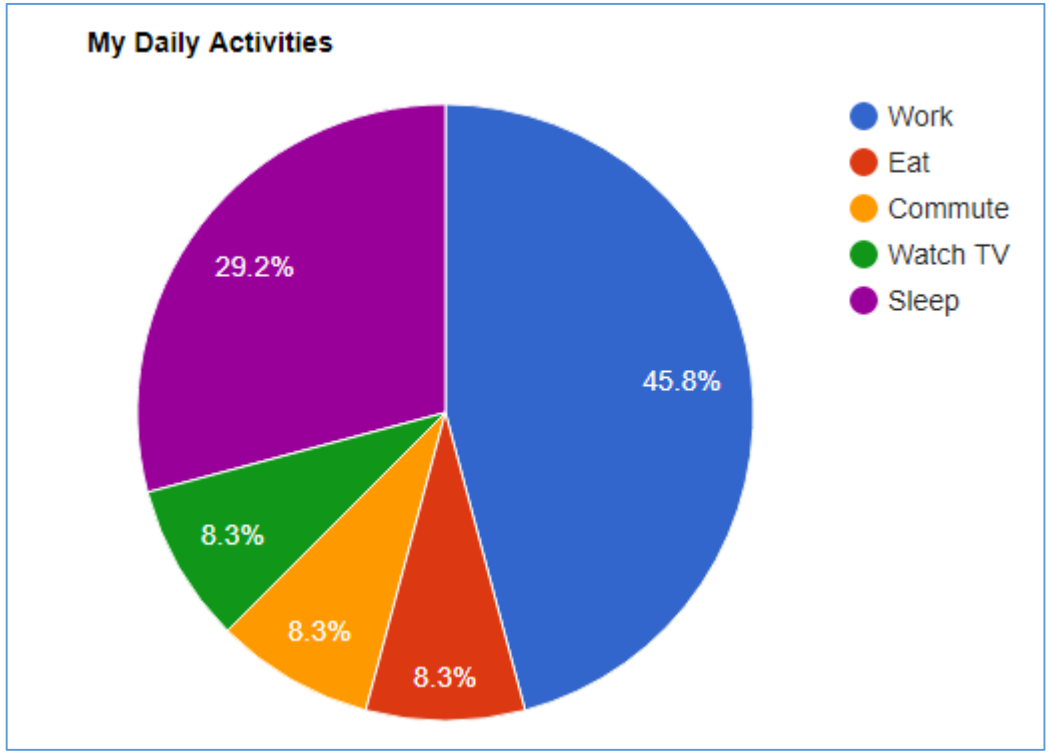
- These are like bar graphs, except that they are useful in showing data frequencies or data values on classes (or ranges) of a numerical variable.



PC: https://statistics.laerd.com/statistical-guides/img/uh/laerd-statistics-example-histogram-frequencies-for-age.png

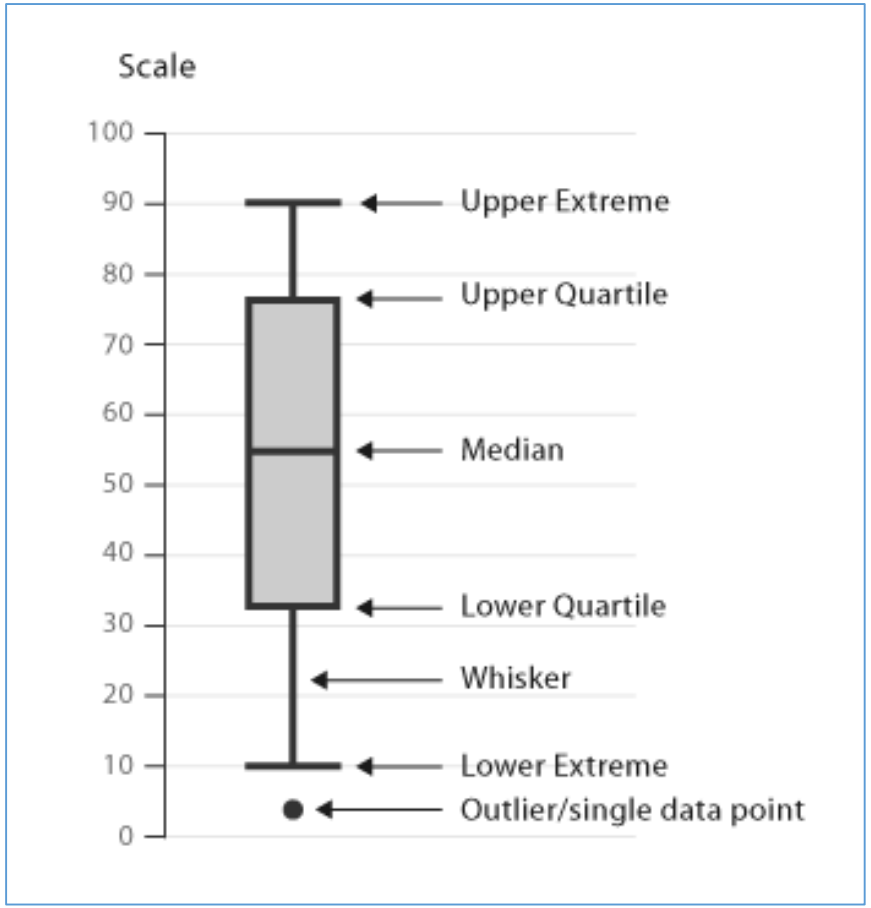
Pie charts

- These are very popular to show the distribution of a variable, such as sales by region. The size of a slice is representative of the relative strengths of each value.



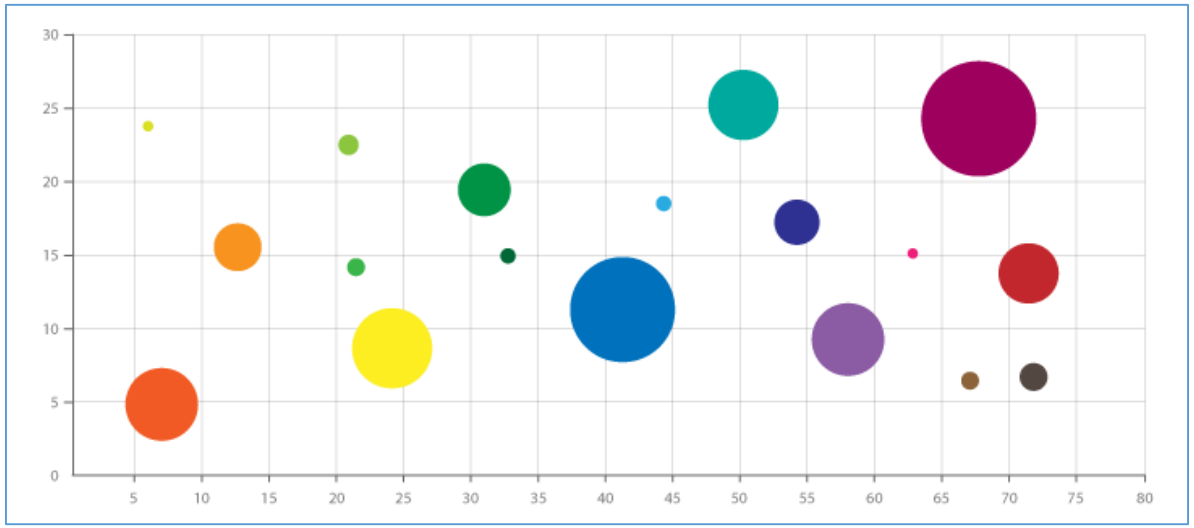
Box charts

- These are special form of charts to show the distribution of variables.
- The box shows the middle half of the values, while whiskers on both sides extend to the extreme values in either direction.



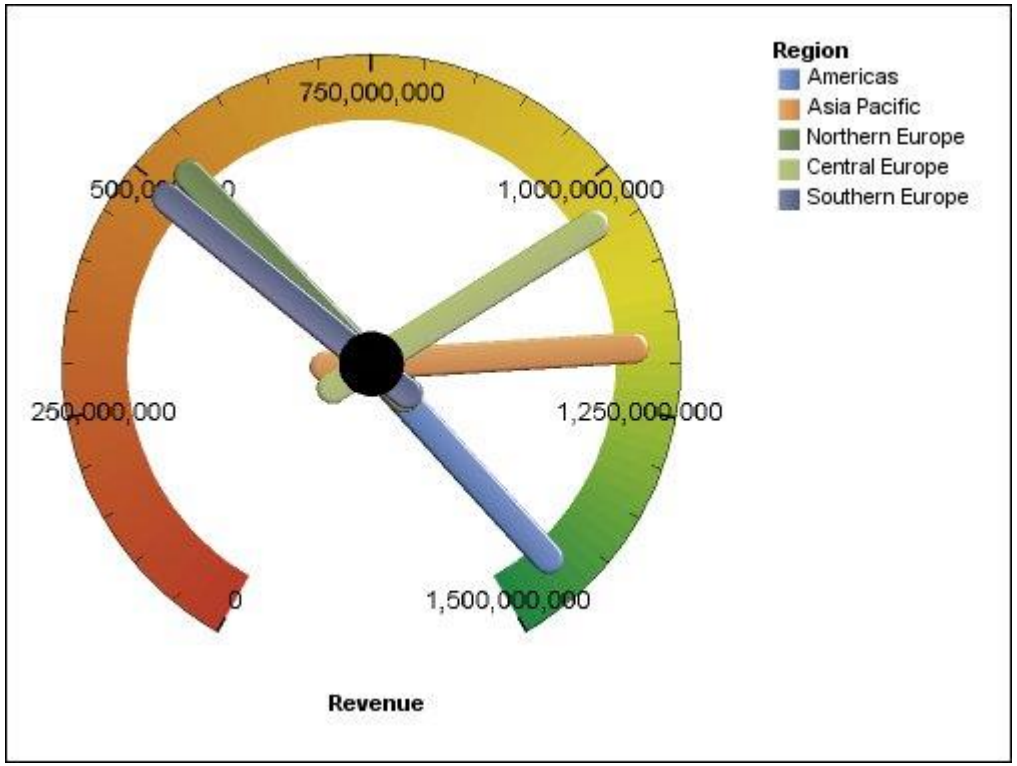
Bubble Graph

- This is an interesting way of displaying multiple dimensions in one chart.
- It is a variant of a scatter plot with many data points marked on two dimensions.
- Imagine that each data point on the graph is a bubble (or a circle) the size of the circle and the color fill in the circle could represent two additional dimensions.

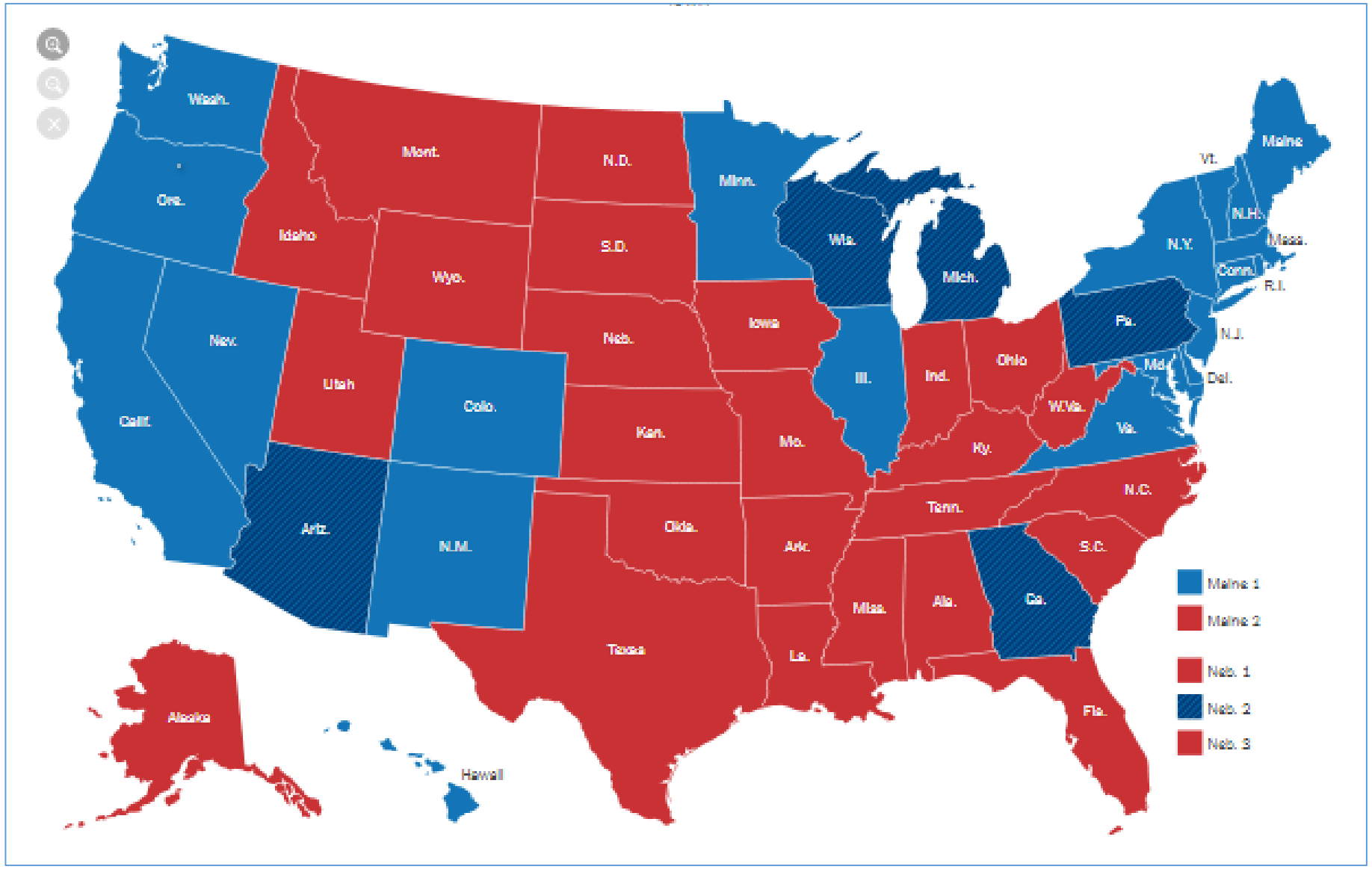


Dials

- These are charts like the speed dial in the car, that shows whether the variable value (such as sales number) is in the low range, medium range, or high range.
- These ranges could be colored red, yellow and gree to give an instant view of the data.

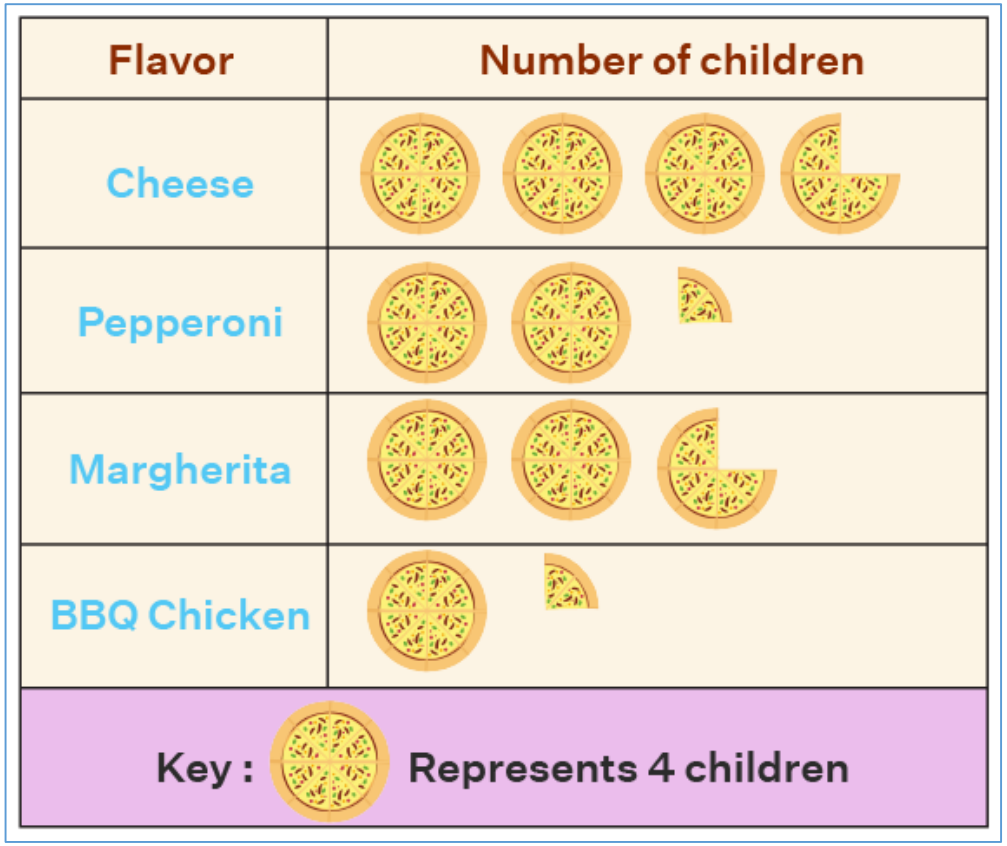


Geographical Data maps



Pictographs

- One can use pictures to represent data, where images are used to show the product for easy reference.
- A survey was conducted for 40 children by a fast food junction to understand the demand for different flavors of pizza available in their outlet.



Design Principles

Introduction

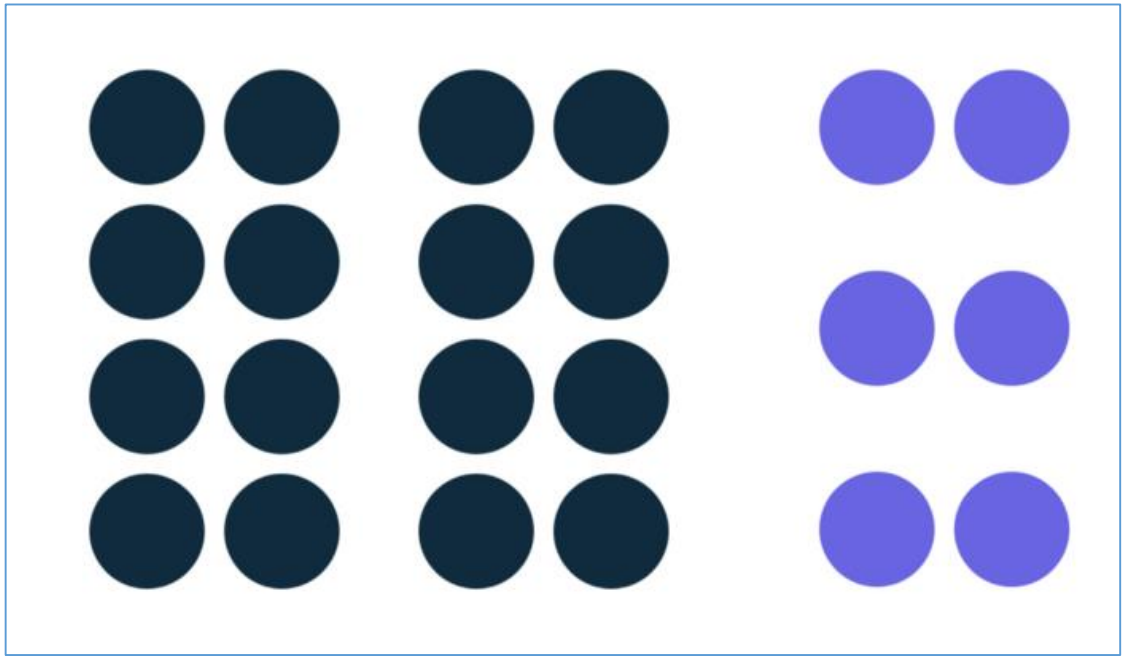
- Communicating the data effectively is an art.
- The role of data visualization in communicating the complex insights hidden inside data is vital.
- This is becoming more and more important since the audience for data visualizations is also expanding along with the size of data.
- While data scientists and analysts have an eye for digging out the key insights from even complex visualizations, a top business stakeholder or an average person might not be able to do the same.
- Here are some of the key design principles for creating beautiful and effective data visualizations.

Gestalt Principles of Design

- Proximity
- Similarity
- Enclosure
- Symmetry
- Closure
- Continuity
- Connection
- Figure and ground

Proximity

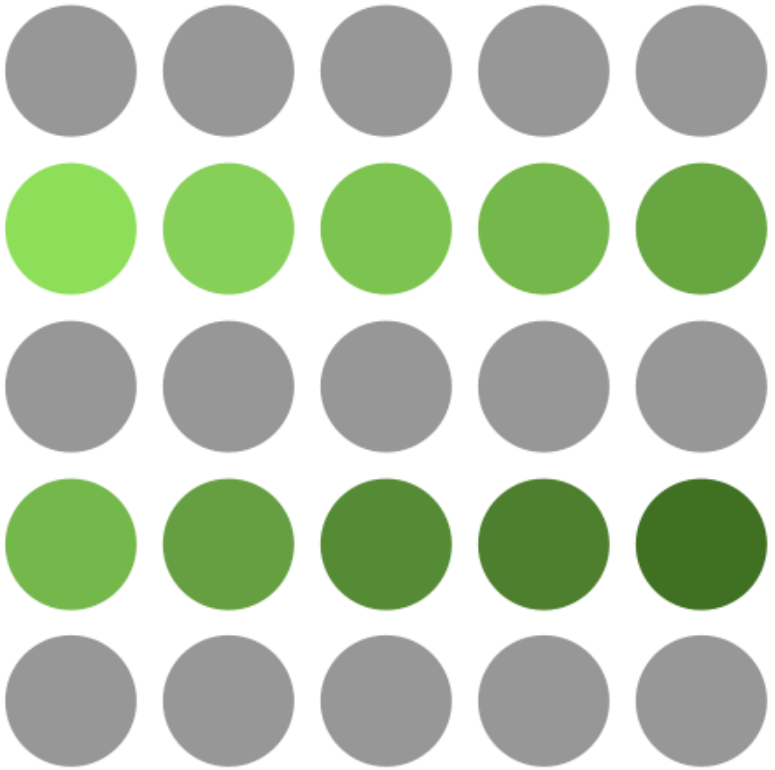
- We are accustomed to placing charts and graphs near each other or in some organized arrangement.
- The Proximity Principle states that objects arranged close together are perceived as more related that those placed further apart.



PC: https://miro.medium.com/max/700/1*rb6h4RwbjHPjQl2qRLntHA.png

Similarity

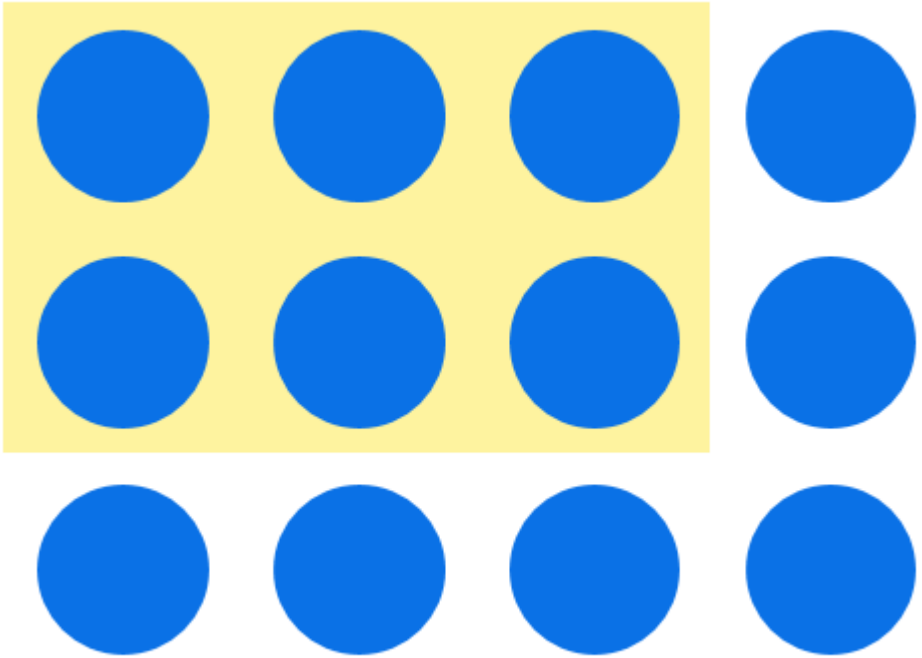
- When objects share similar attributes, such as color, direction, or shape, they are perceived as being part of a group.
- The Similarity Principle can help someone more readily identify which groups the displayed data belong with.



PC: https://evolvingweb.ca/sites/default/files/inline-images/3-gestalt-principle-2_0.png

Enclosure

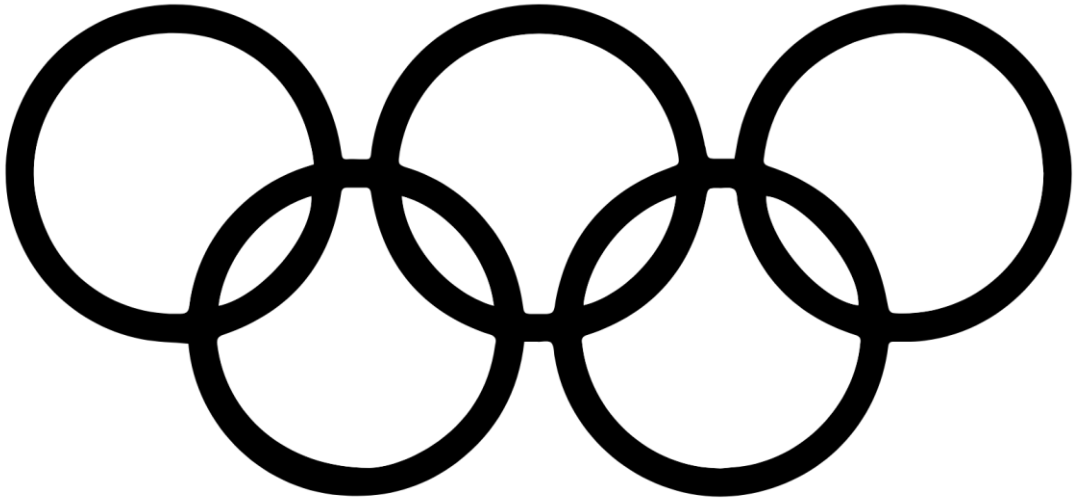
- Objects collected within a boundary-like structure are perceived as a group.
- By placing a line or shading around visual elements on a dashboard, signals that objects within the boundary form one, or belong together.



PC: https://www.topcoder.com/blog/gestalt-principles-for-data-visualization/

Symmetry

- Principle which states that symmetrical elements tend to be perceived as belonging together regardless if they are far apart.



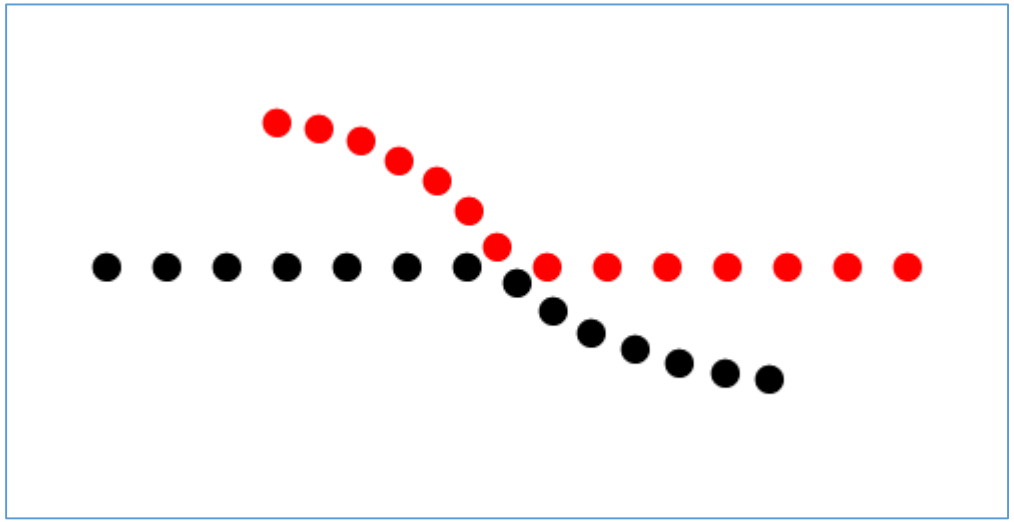
Closure

- We tend to complete shapes and paths even if part of them is missing
- It’s the idea that brain will fill in the missing parts of a design or image to create a whole.

PC: https://wwf.panda.org/

Continuity

- We perceive objects as belonging together, as part of a single whole, if they are aligned with one another or appear to form a continuation of one another.
- It's like the closure principle, but besides the visual connection to form shape, we also attach visual direction as part of the continuation.



Connection

- Helps group elements together
- Objects are perceived as lines that move along the smoothest path

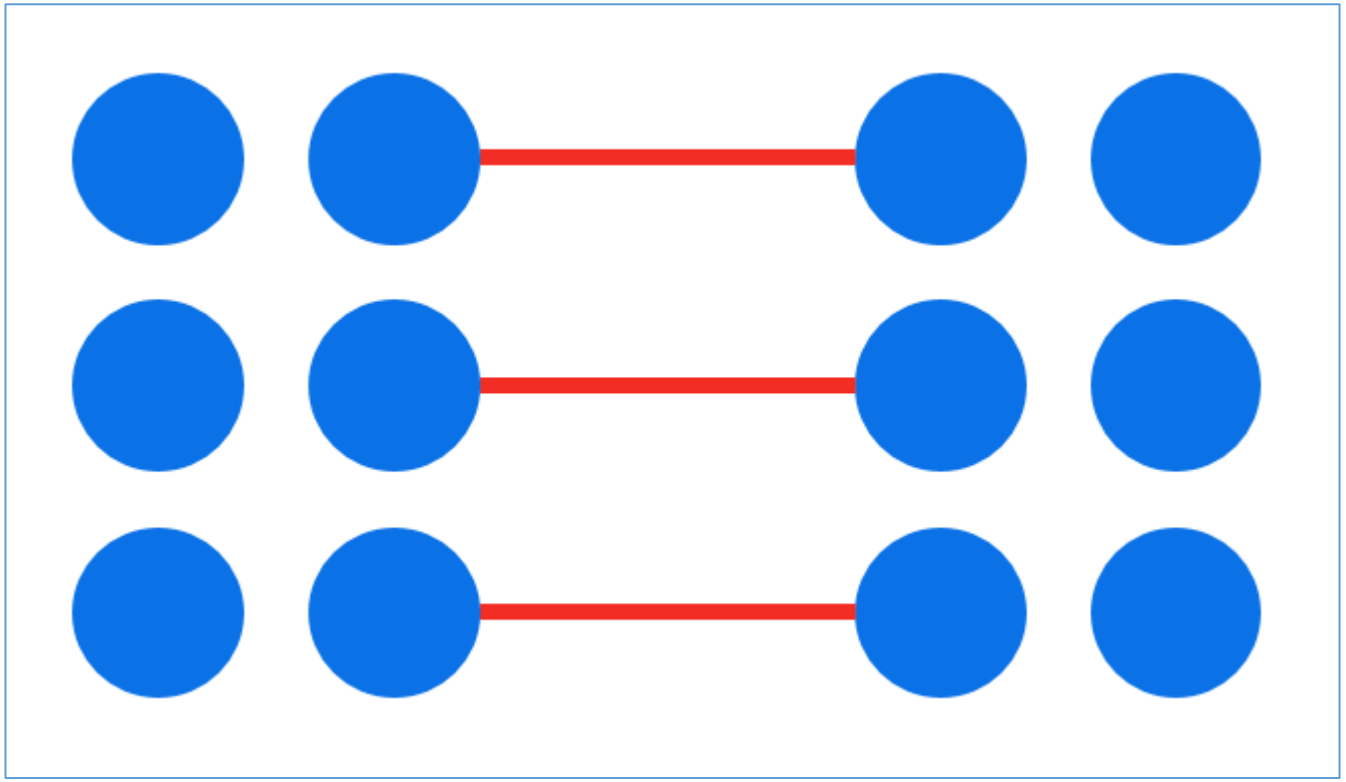


Figure and ground

- We typically notice only one of several main visual aspects of a graph
- what we do notice becomes the figure, and everything else becomes the “background”.

Can you Identify..?

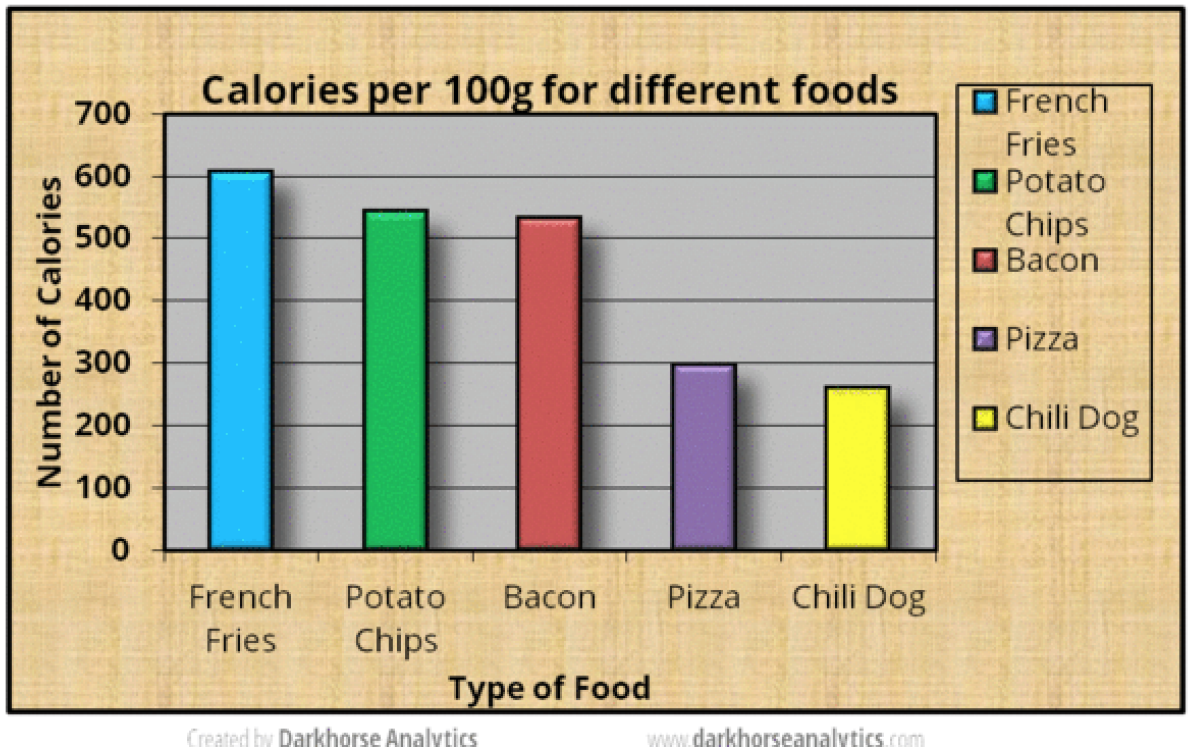
Tufte's Principles of Design

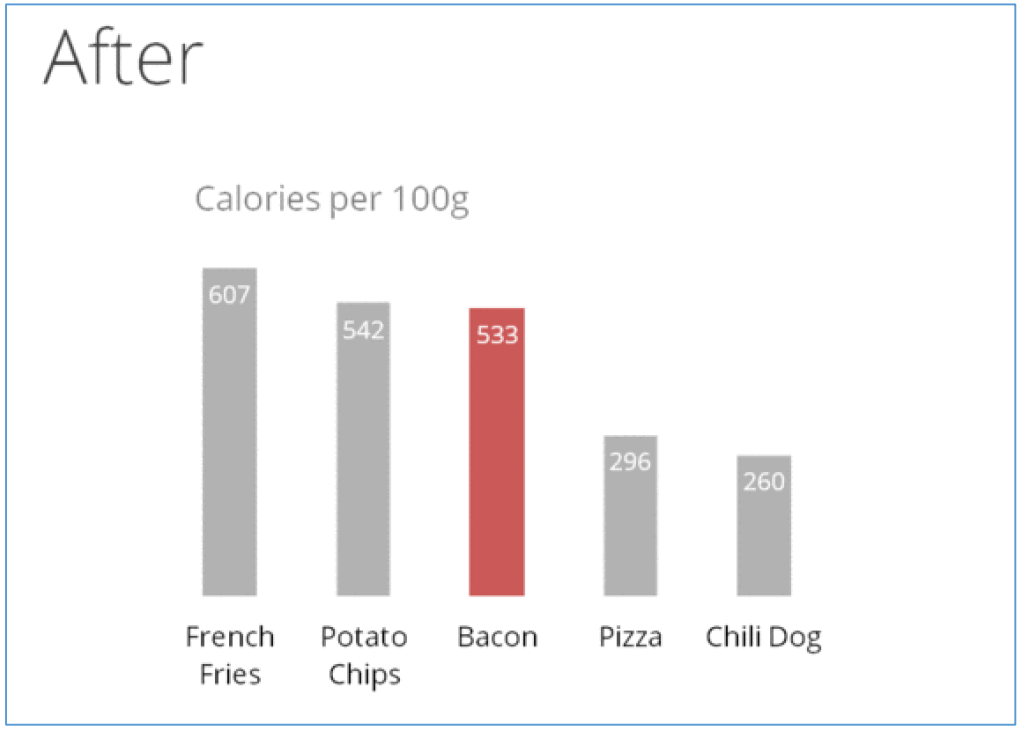
- Principle 1: Maximizing the data-ink ratio, within reason
- Principle 2: Mobilize every graphical element, perhaps several times over, to show the data.
- Principle 3: Maximize data density and the size of the data matrix, within reason.
- Principle 4: Escape flatland - small multiples, parallel sequencing.
- Principle 5: Provide the user with an overview and details on demand.
- Principle 6: Utilize Layering & Separation.

Principle 1: Maximizing the data-ink ratio, within reason

- Data-ink is the non-erasable core of a graphic, the non-redundant ink arranged in response to variation in the numbers represented.
- It is also the proportion of graphic's ink devoted to the non-redundant display of data-information.
- Data Ink Ratio = Data Ink / Total Ink

Example

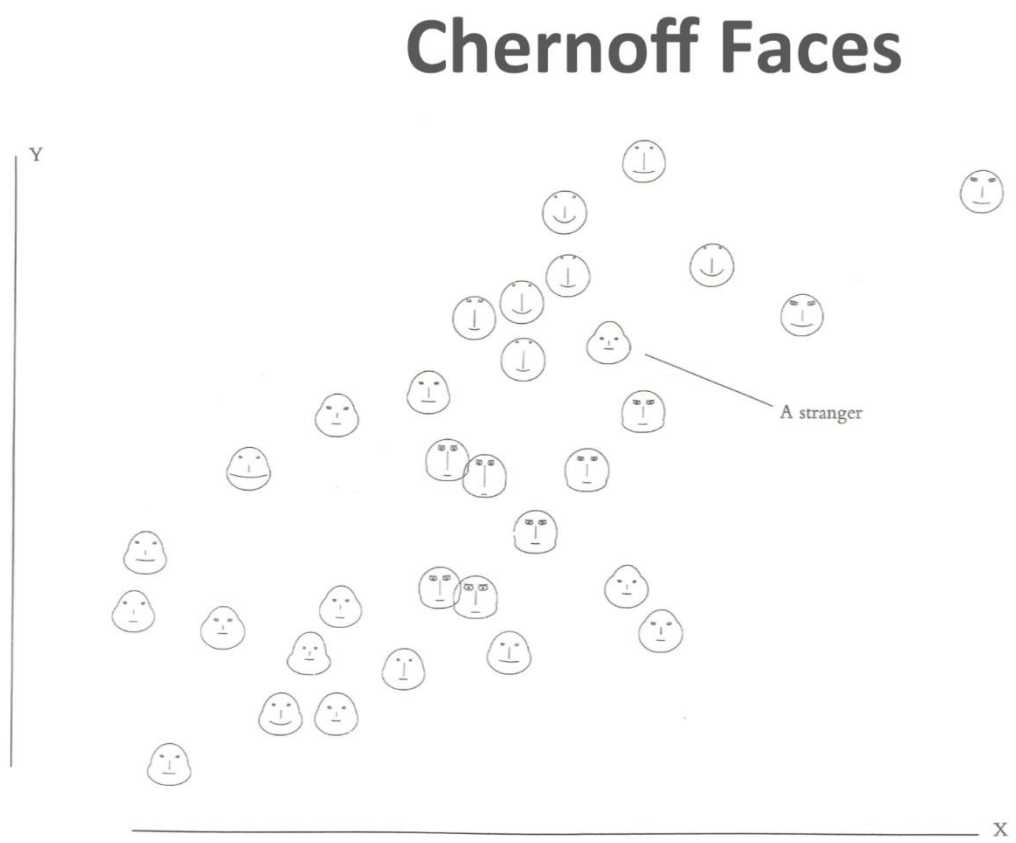

Calories per 100g for different foods
Number of Calories
700
600
500
400
300
200
100
0
French Fries
Potato Chips
Bacon
Pizza
Chili Dog
Type of Food
Created by Darkhorse Analytics
www.darkhorseanalytics.com


After
Calories per 100g
607
542
533
296
260
French Fries
Potato Chips
Bacon
Pizza
Chili Dog

Principle 2: Mobilize every graphical element, perhaps several times over, to show the data.

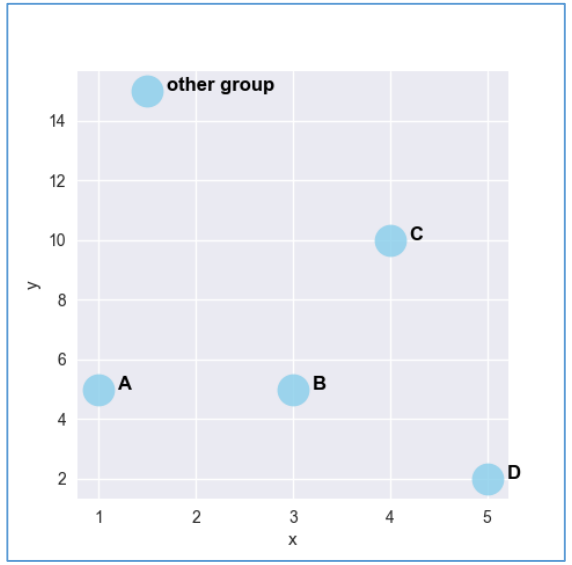
- The graphical element that actually locates or plots the data is the data measure, the ink of the data measure can can itself carry data
- Multiple layers of information are created by multiple viewing depths and multiple viewing angles.
- Thus design techniques for enhancing graphical clarity in the face of complexity must be developed along with multifunctioning elements.
- In other words, we should try to make all present graphical elements data encoding elements. We must make every graphical element effective.

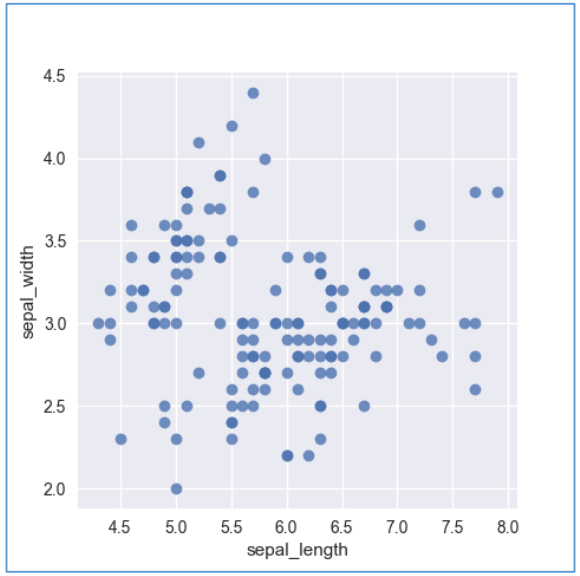
Example



Principle 3: Maximize data density and the size of the data matrix, within reason.

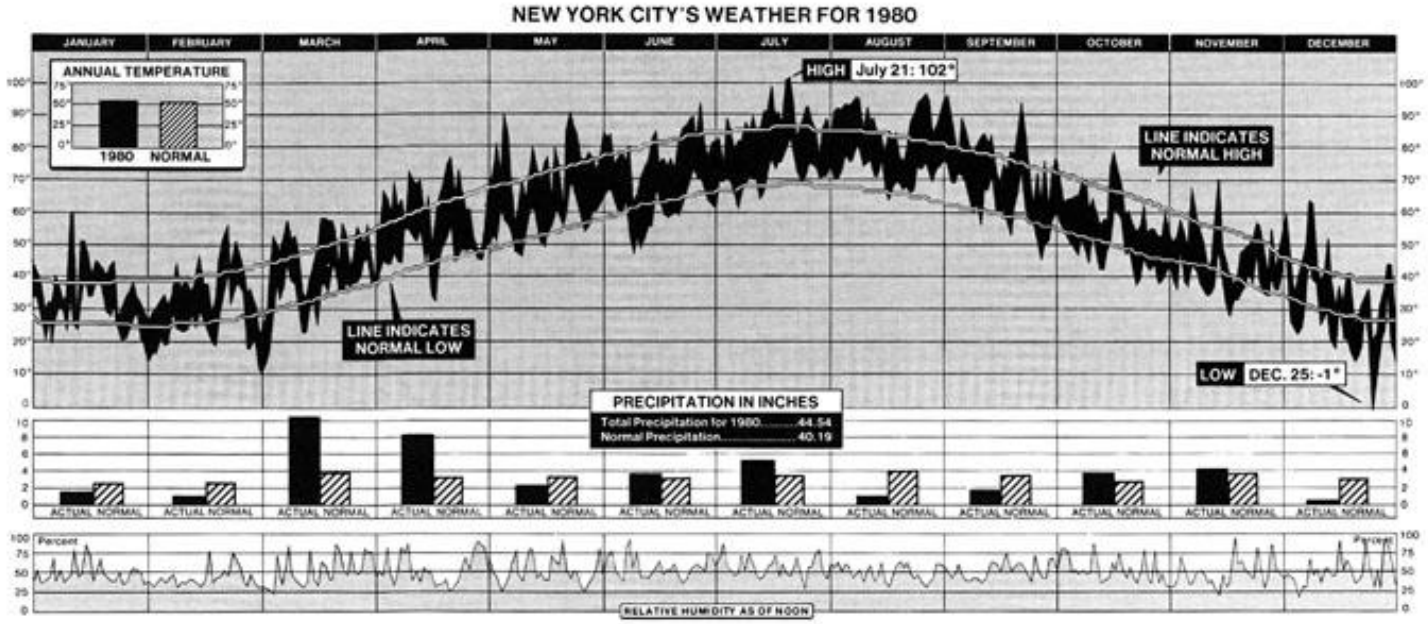
- One way of achieving this is through the Shrink Principle.
- As the volume of data increases, data measures must shrink
- Graphs must be shrunk way down without losing legibility or information.





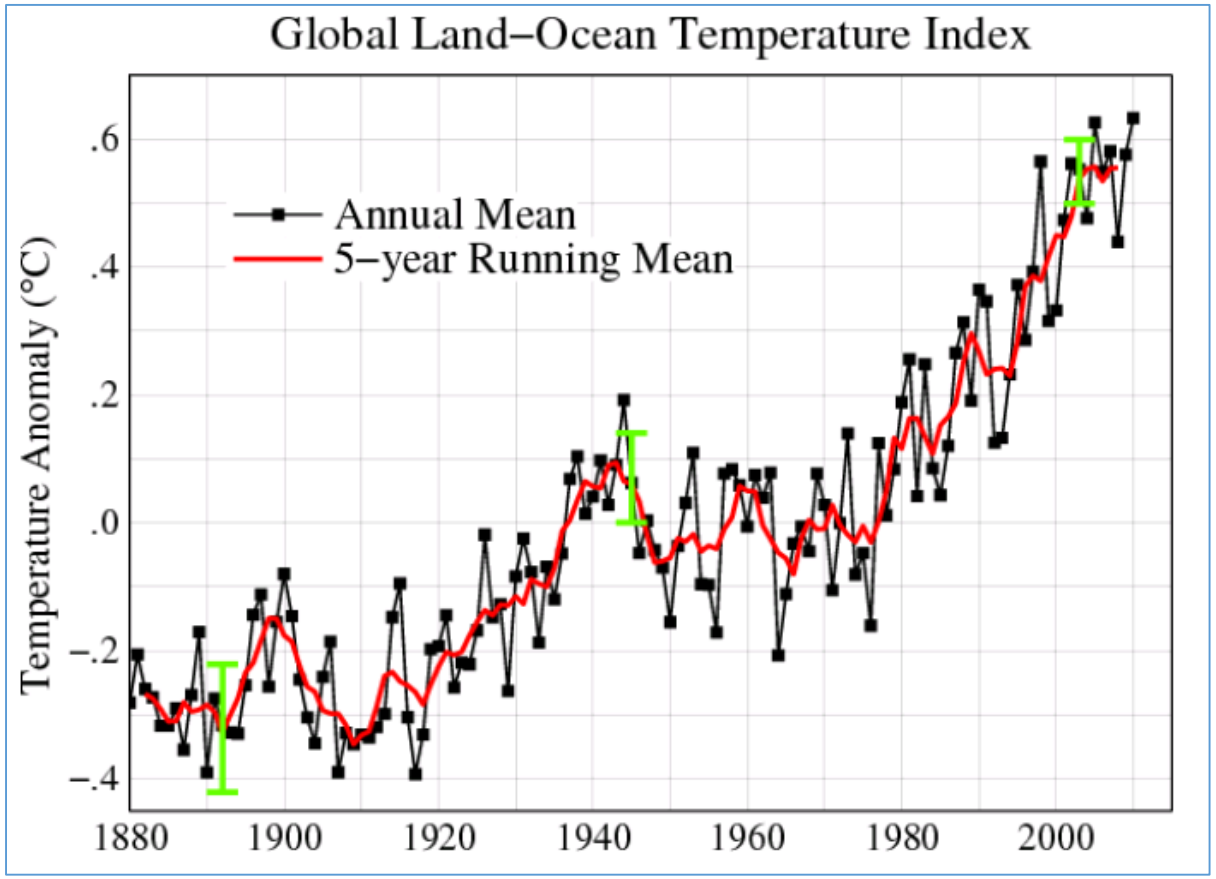
Principle 4: Escape flatland - small multiples, parallel sequencing.

- Data is multivariate doesn't necessarily mean 3D projection.
- How can we enhance multivariate data on inherently 2D surfaces?
- We can use small multiple graphs or parallel sequencing skill.



Principle 5: Provide the user with an overview and details on demand.

- A carefully designed view can show a macrostructure (overview) as well as microstructure (detail) in one space.



Principle 6: Utilize Layering & Separation.

- Supported by Gestalt law of grouping
  - Grouping with colors
  - Using Color to separate

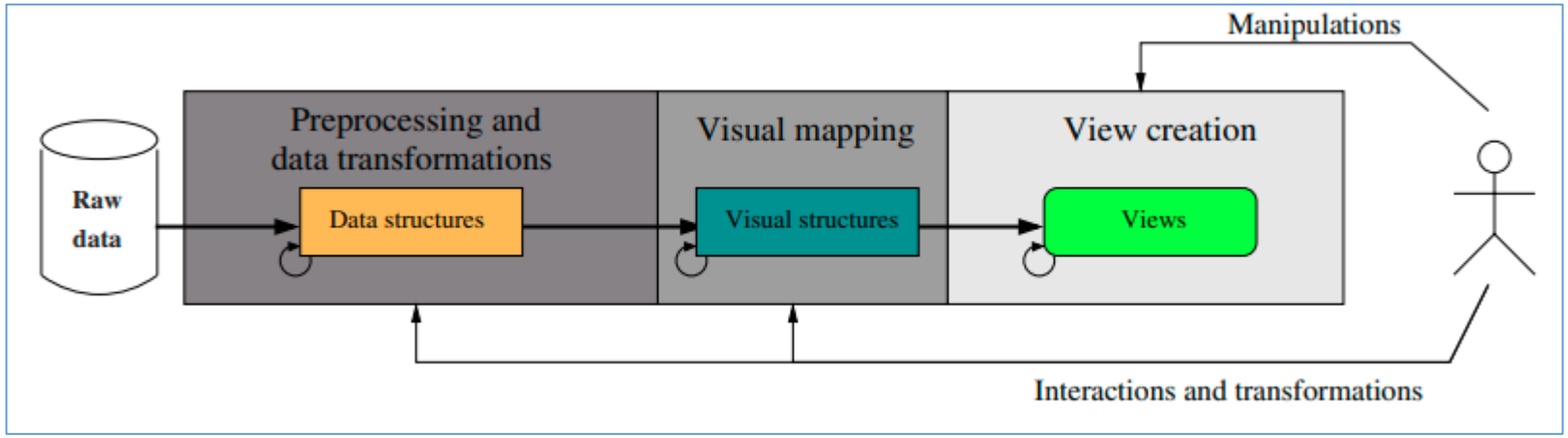
Creating Visual Representations

- Let’s imagine that we have at our disposal a collection of data on which we’d like to carry out explorative analysis to identify any possible unknown tendencies or relationships.
- How can we go about a creating visual representation from this data?

How to do..?

- We don’t want to attempt designing everything from scratch
- Computers can help us greatly and many visualization software that can provide us with a complete series of visual templates.
- But how do these programs work?

The process of generating a graphical representation.



- Preprocessing and data transformations
- Visual mapping
- View creation

Preprocessing and Data Transformations

Example

- Find the age distribution of employees in the organization
- Find the income of employees age considering their age
- Whether the age and Gender are effecting on sales?
- How total salary distributed among the employees.

	EMPID	Gender	Age	Sales	BMI	Income
0	E001	M	34	123	Normal	350
1	E002	F	40	114	Overweight	450
2	E003	F	37	135	Obesity	169
3	E004	M	30	139	Underweight	189
4	E005	F	44	117	Underweight	183
5	E006	M	36	121	Normal	80
6	E007	M	32	133	Obesity	166
7	E008	F	26	140	Normal	120
8	E009	M	32	133	Normal	75
9	E010	M	36	133	Underweight	40

Visual Mapping

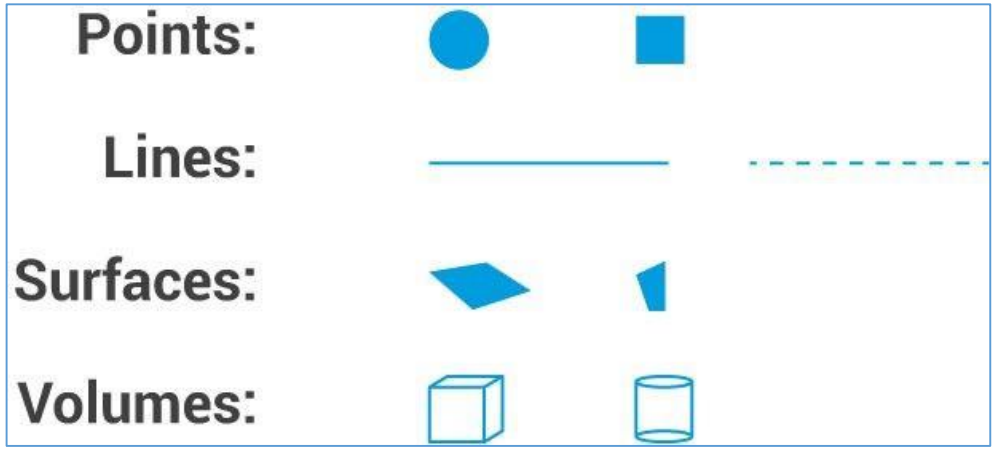
- There are three elements to define for an information visualization
- The Spatial Substrate
- The Graphical Elements
- The Graphical Properties

The Spatial Substrate

- Space in which we're going to create our visualization (2D/3D).
- It's important to consider which axes you will use and what kind of data will map to each individual axis.
  - Quantitative.
  - Ordinal
  - Nominal

The Graphical Elements

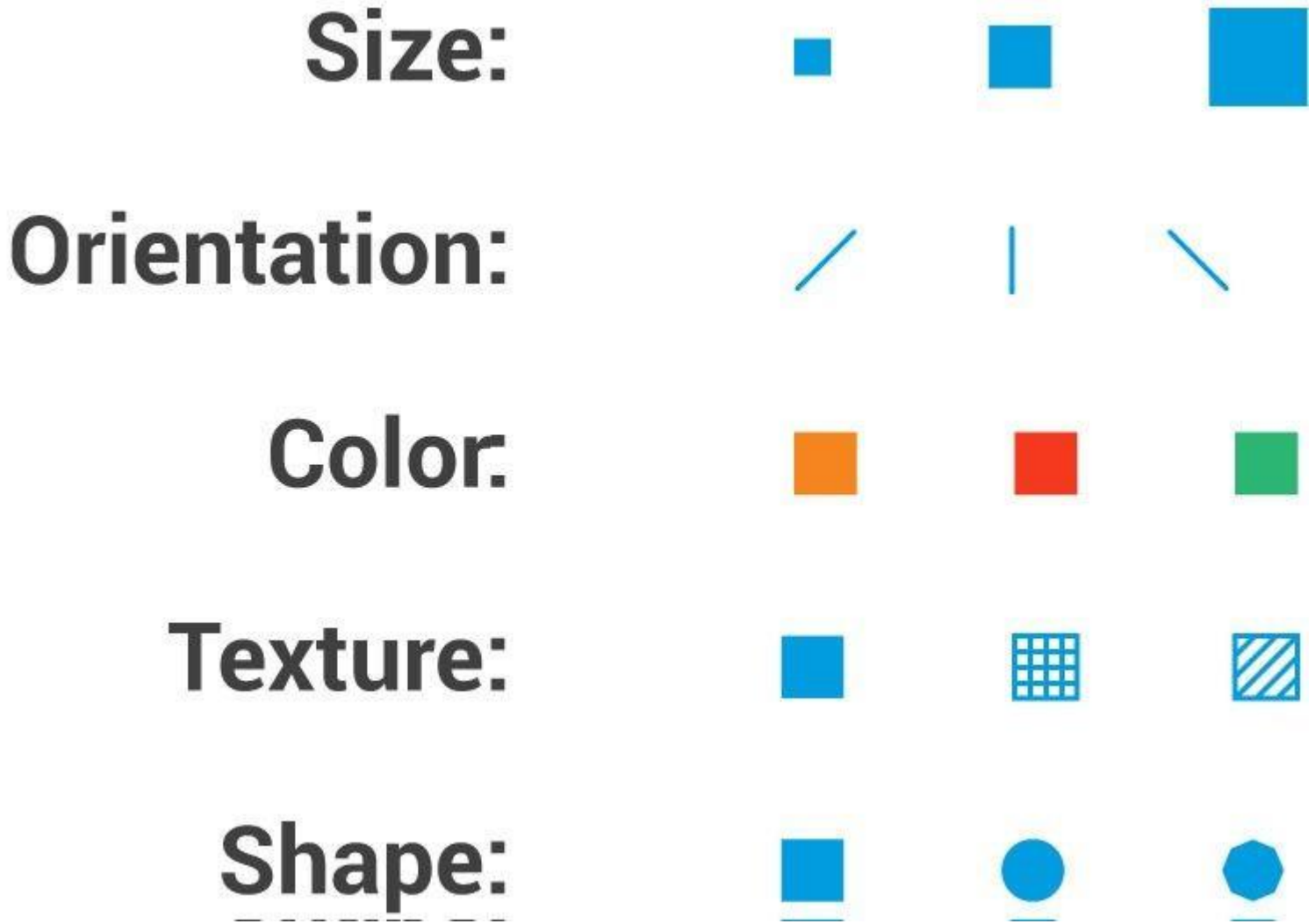
- Graphical elements are the visual elements that will appear in the spatial substrate. There are four kinds of visual elements:
- Points
- Lines
- Surfaces
- Volumes



PC: https://www.interaction-design.org/literature/article/visual-mapping-the-elements-of-information-visualization

The Graphical Properties

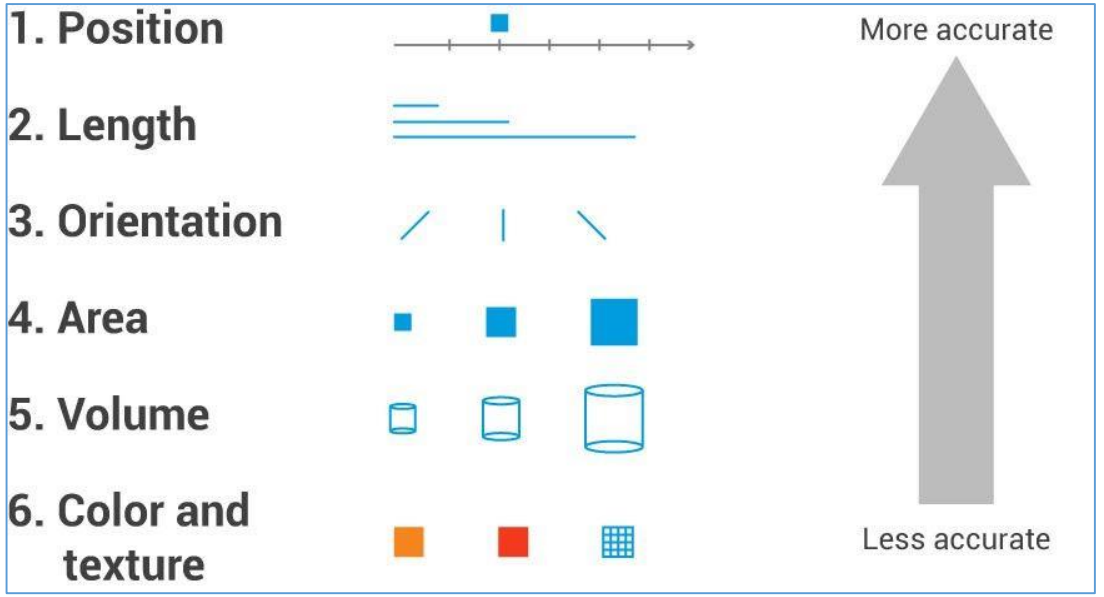
- **Size** – the size of a point, surface or volume can convey information about significance or weight
- **Orientation** – the orientation of a line or volume in space can help clarify the purpose of the line or volume
- **Color** – color coding can be used to provide a visual key to data or to provide a “call to attention” to the eye
- **Textures** – textures can be used to provide visual keys to data
- **Shapes** – hexagons, circles, triangles, etc. can be used to provide a visual key to data



PC: https://www.interaction-design.org/literature/article/visual-mapping-the-elements-of-information-visualization

Cleveland and McGill graphical properties

- Cleveland and McGill said some graphical properties are more effective than others when it comes to conveying information.
  - Position on the axis
  - Length of line
  - Orientation of line or object
  - Area of shape
  - Volume of shape
  - Color and textural indicators



PC: https://www.interaction-design.org/literature/article/visual-mapping-the-elements-of-information-visualization

Views

- The views are the final result of the generation process.
- They are the result of the mapping of data structures to the visual structures, generating a visual representation in the physical space represented by the computer.
- They are what we see displayed on the computer screen

Designing a Visual Application

- The procedure to follow, when creating the visual representations of abstract data, can be outlined in the following steps:
  - Define the problem
  - Examine the nature of the data to represent.
  - Number of dimensions
  - Data structures
  - Type of interaction