INT 303 BIG DATA ANALYTICS

Lecture5: Data Visualization

Jia WANG

Jia.wang02@xjtlu.edu.cn

Xi'an Jiaotong-Liverpool University

西交利物浦大學

OUTLINE

- Visualization motivation
- Principle of Visualization
- Types of Visualization
- Example

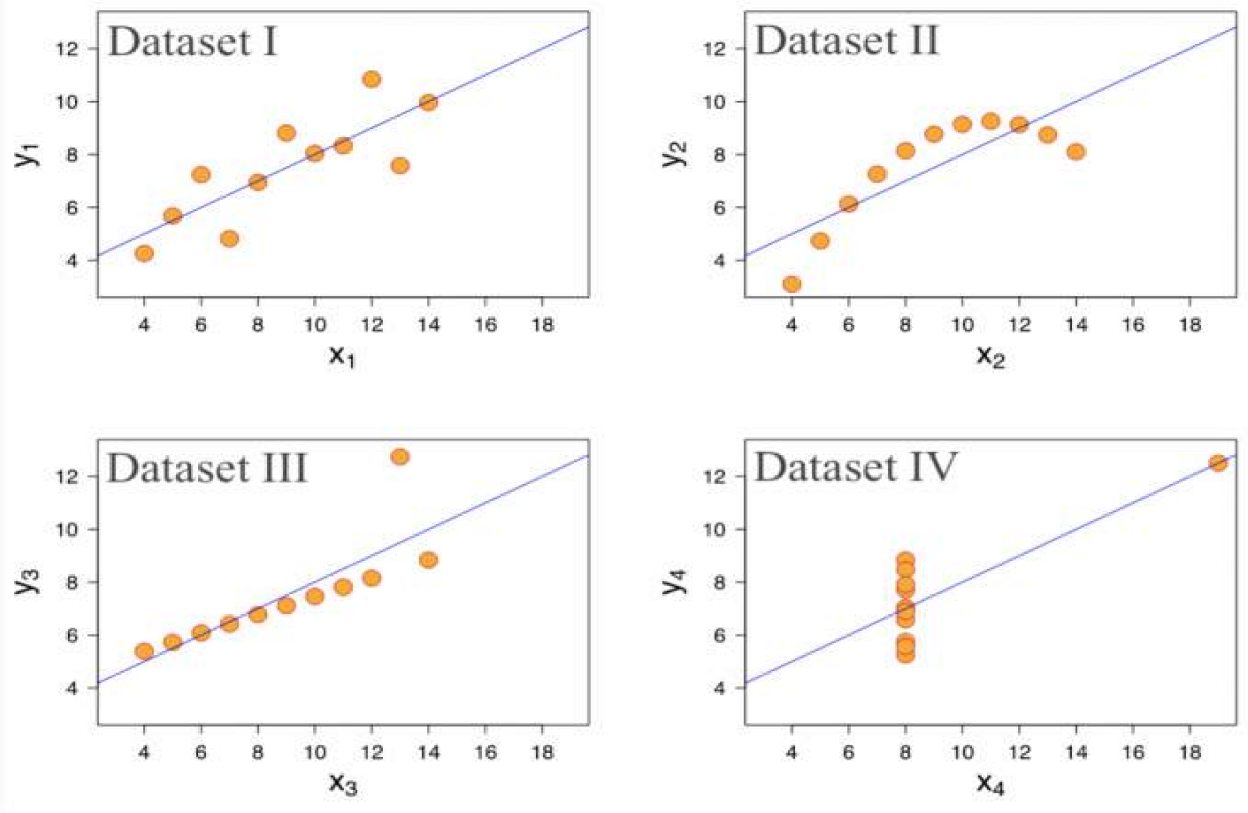
Anscombe's Data

The following four data sets comprise the Anscombe's Quartet; all four sets of data have identical simple summary statistics.

	Dataset I		Dataset II		Dataset III		Dataset IV	
	x	y	x	y	x	y	x	y
	10	8.04	10	9.14	10	7.46	8	6.58
	8	6.95	8	8.14	8	6.77	8	5.76
	13	7.58	13	8.74	13	12.74	8	7.71
	9	8.81	9	8.77	9	7.11	8	8.84
	11	8.33	11	9.26	11	7.81	8	8.47
	14	9.96	14	8.1	14	8.84	8	7.04
	6	7.24	6	6.13	6	6.08	8	5.25
	4	4.26	4	3.1	4	5.39	19	12.5
	12	10.84	12	9.13	12	8.15	8	5.56
	7	4.82	7	7.26	7	6.42	8	7.91
	5	5.68	5	4.74	5	5.73	8	6.89
Sum:	99.00	82.51	99.00	82.51	99.00	82.51	99.00	82.51
Avg:	9.00	7.50	9.00	7.50	9.00	7.50	9.00	7.50
Std:	3.32	2.03	3.32	2.03	3.32	2.03	3.32	2.03

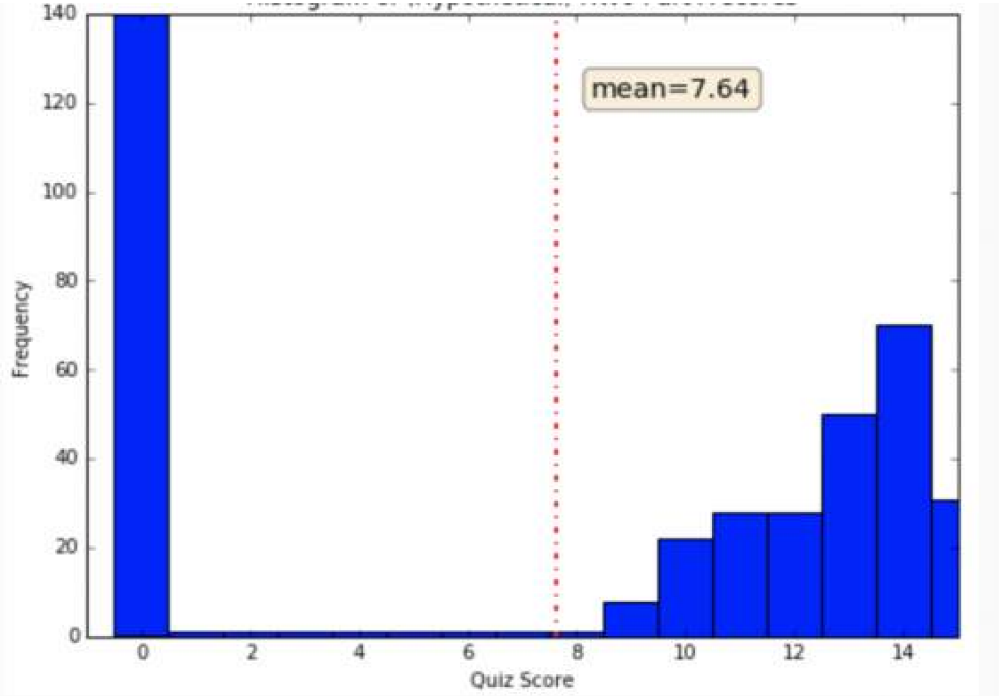
Anscombe's Data (cont.)

Summary statistics clearly don't tell the story of how they differ. But a picture can be worth a thousand words:



VISUALIZATION MOTIVATION

If I tell you that the average score for Homework 0 was: 7.64 last year, what does that suggest?



And what does the graph suggest?

VISUALIZATION MOTIVATION

Visualizations help us to analyze and explore the data. They help to:

- Identify hidden patterns and trends
- Formulate/test hypotheses
- Communicate any modeling results
  - Present information and ideas succinctly
  - Provide evidence and support
  - Influence and persuade
- Determine the next step in analysis/modeling

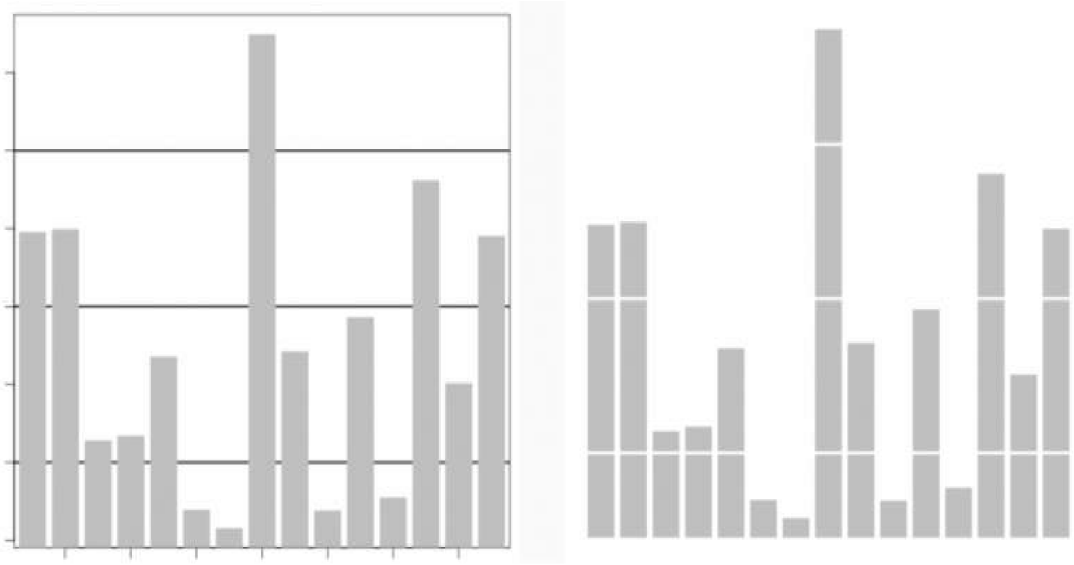
PRINCIPLES OF VISUALIZATIONS

PRINCIPLES OF VISUALIZATIONS (1)

Some basic data visualization guidelines from Edward Tufte:

1. Maximize data to ink ratio: show the data.

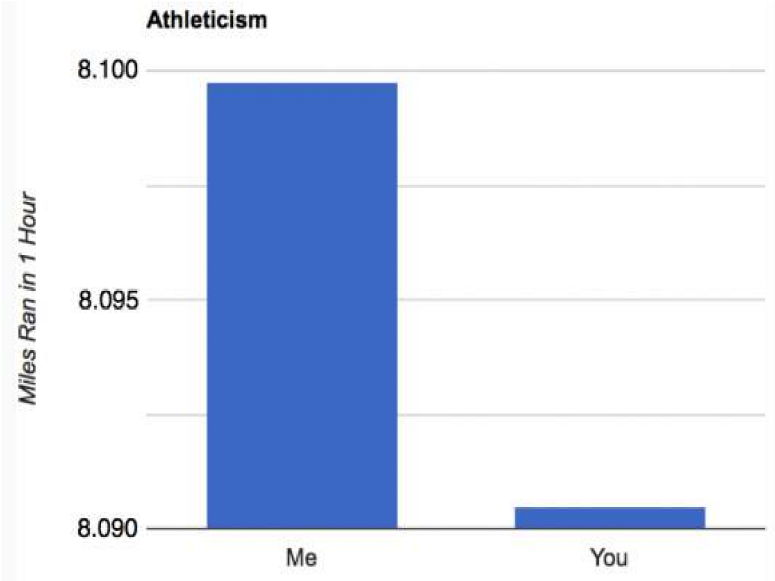
Which is better?

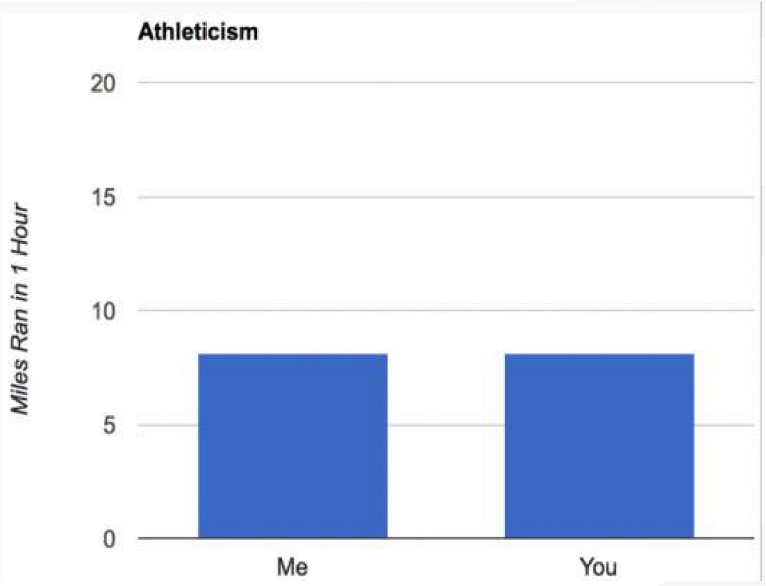


PRINCIPLES OF VISUALIZATIONS (2)

Some basic data visualization guidelines from Edward Tufte:

1. Maximize data to ink ratio: show the data
2. Don’t lie with scale (Lie Factor)

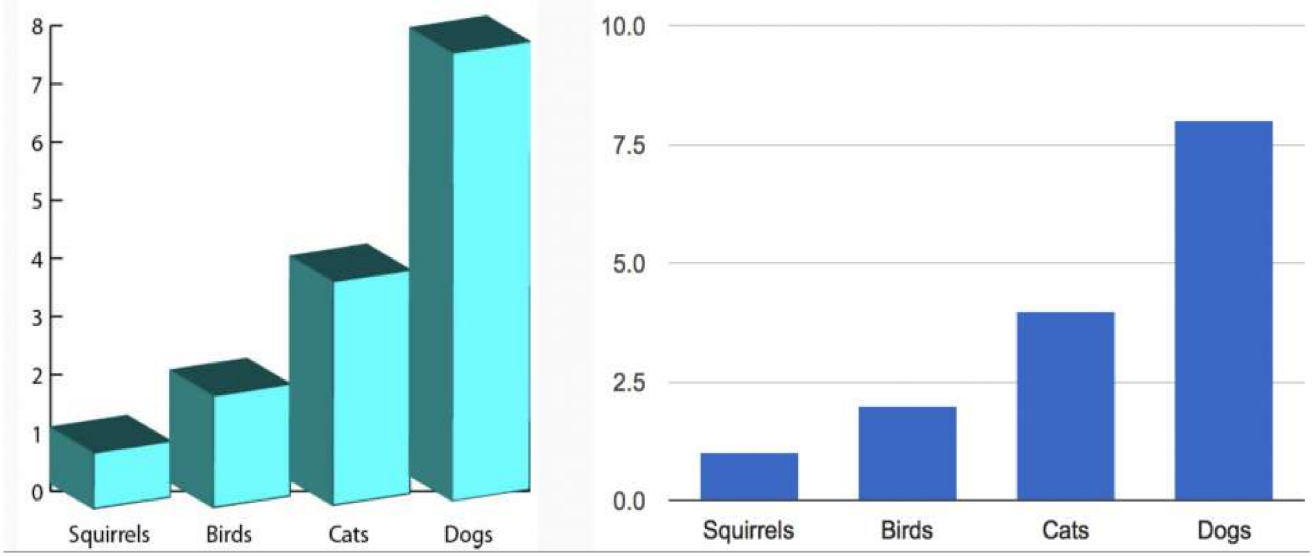




PRINCIPLES OF VISUALIZATIONS (3)

Some basic data visualization guidelines from Edward Tufte:

1. Maximize data to ink ratio: show the data
2. Don’t lie with scale: minimize
3. Minimize chart-junk: show data variation, not design variation



PRINCIPLES OF VISUALIZATIONS (4)

Some basic data visualization guidelines from Edward Tufte:

1. Maximize data to ink ratio: show the data
2. Don’t lie with scale: minimize
3. Minimize chart-junk: show data variation, not design variation
4. Clear, detailed and thorough labeling

TYPES OF VISUALIZATIONS

TYPES OF VISUALIZATIONS

What do you want your visualization to show about your data?

- **Distribution:** how a variable or variables in the dataset distribute over a range of possible values.
- **Relationship:** how the values of multiple variables in the dataset relate
- **Composition:** how a part of your data compares to the whole.
- **Comparison:** how trends in multiple variable or datasets compare

What would you like to show?

COMPARISON

- Among Items
  - Two Variables per Item → VARIABLE WIDTH COLUMN CHART
  - Many Categories → TABLE WITH EMBEDDED CHARTS
  - Few Categories → BAR CHART HORIZONTAL
- Over Time
  - Cyclical Data → CIRCULAR AREA CHART
  - Non-Cyclical Data → LINE CHART
- One Variable per Item
  - Single or Few Categories → BAR CHART VERTICAL
  - Many Categories → LINE CHART

RELATIONSHIP

- Two Variables → SCATTER PLOT
- Three or more Variables → SCATTER PLOT BUBBLE SAZE

DISTRIBUTION

- Single Variable
  - Few Data Points → BAR HISTOGRAM
  - Many Data Points → LINE HISTOGRAM
- Two Variables → SCATTER PLOT

COMPOSITION

- Changing Over Time
  - Few Periods
    - Only Relative Differences Matter → STACKED 100% BAR CHART
    - Relative and Absolute Differences Matter → STACKED BAR CHART
  - Many Periods
    - Only Relative Differences Matter → STACKED AREA 100% CHART
    - Relative and Absolute Differences Matter → STACKED AREA CHART
- Static
  - Simple Share of Total → PIE CHART
  - Accumulation or Subtraction to Total → WATERFALL CHART
  - Components of Components → STACKED 100% BAR CHART WITH SUBCOMPONENTS
  - Accumulation to total and absolute difference matters → TREE MAP

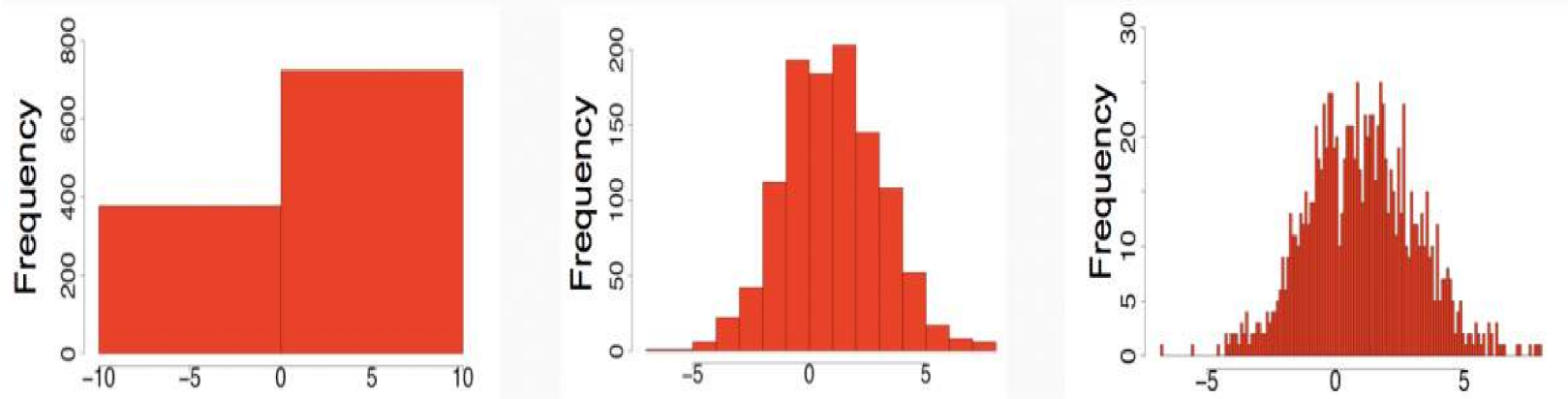
DISTRIBUTION

- When studying how quantitative values are located along an axis, distribution charts are the way to go.

- By looking at the shape of the data, the user can identify features such as value range, central tendency and outliers.

HISTOGRAMS TO VISUALIZE DISTRIBUTION

A **histogram** is a way to visualize how 1-dimensional data is distributed across certain values.

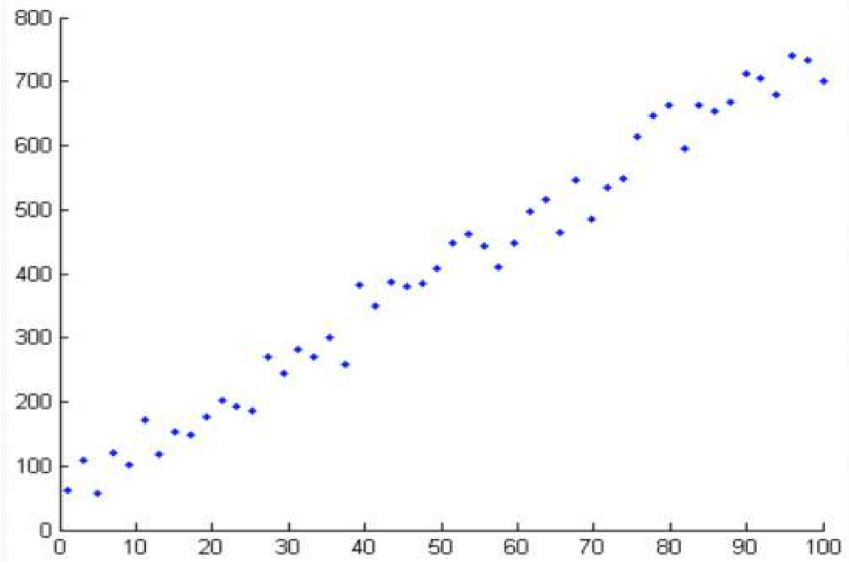


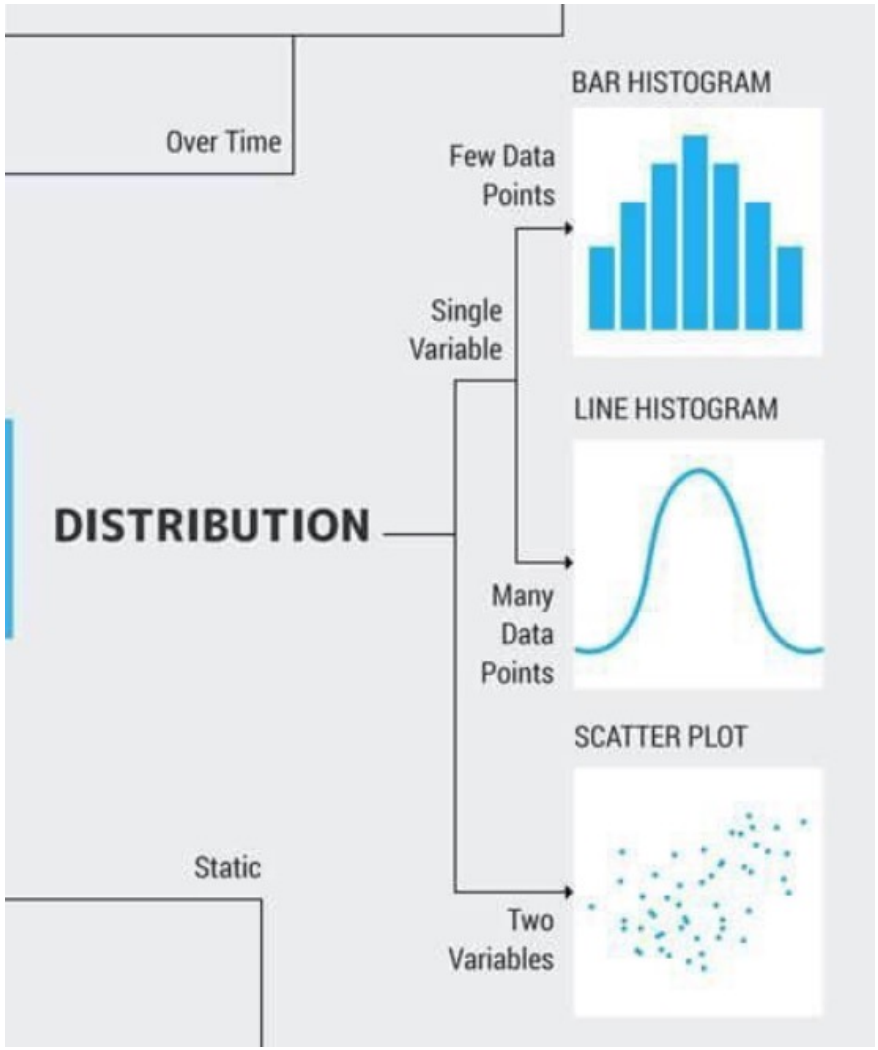
Note: Trends in histograms are sensitive to number of bins.

SCATTER PLOTS TO VISUALIZE RELATIONSHIPS

A **scatter plot** is a way to visualize how multi-dimensional data are distributed across certain values.

A scatter plot is also a way to visualize the relationship between two different attributes of multi-dimensional data.



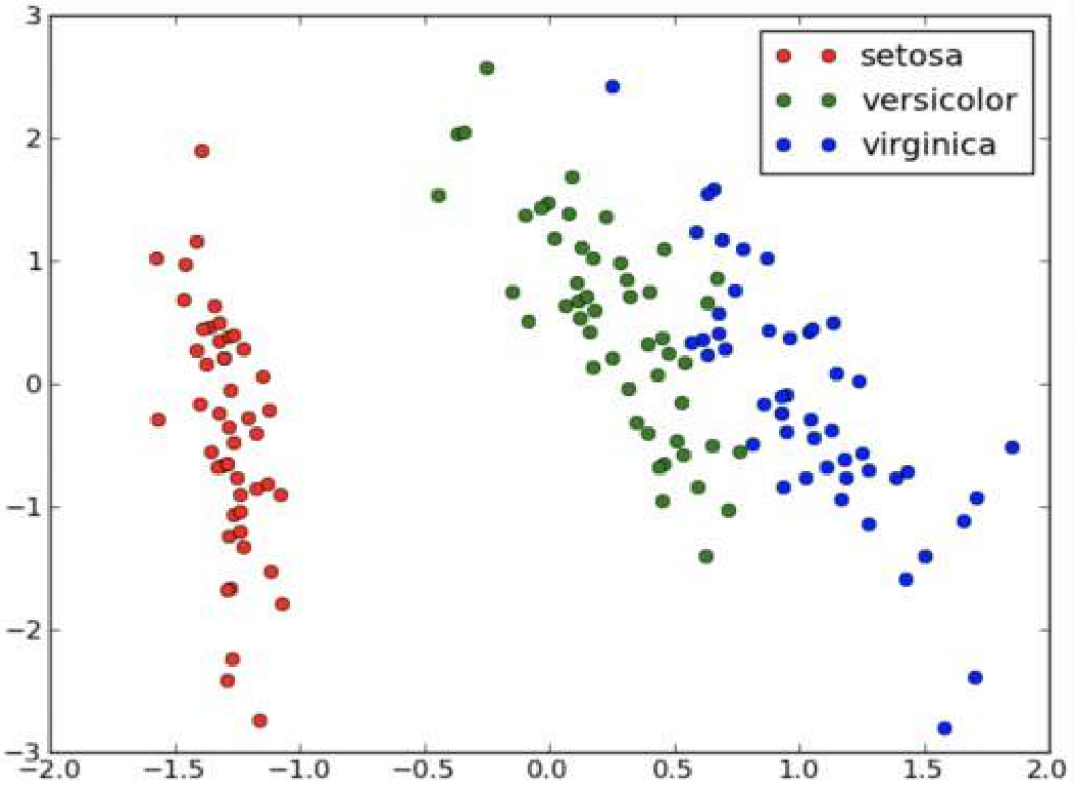

Over Time
BAR HISTOGRAM
Few Data Points
Single Variable
LINE HISTOGRAM
DISTRIBUTION
Many Data Points
SCATTER PLOT
Static
Two Variables

RELATIONSHIP

- They are used to find correlations, outliers, and clusters in your data.

- While the human eye can only appreciate three dimensions together, you can visualize additional variables by mapping them to the size, color or shape of your data points.

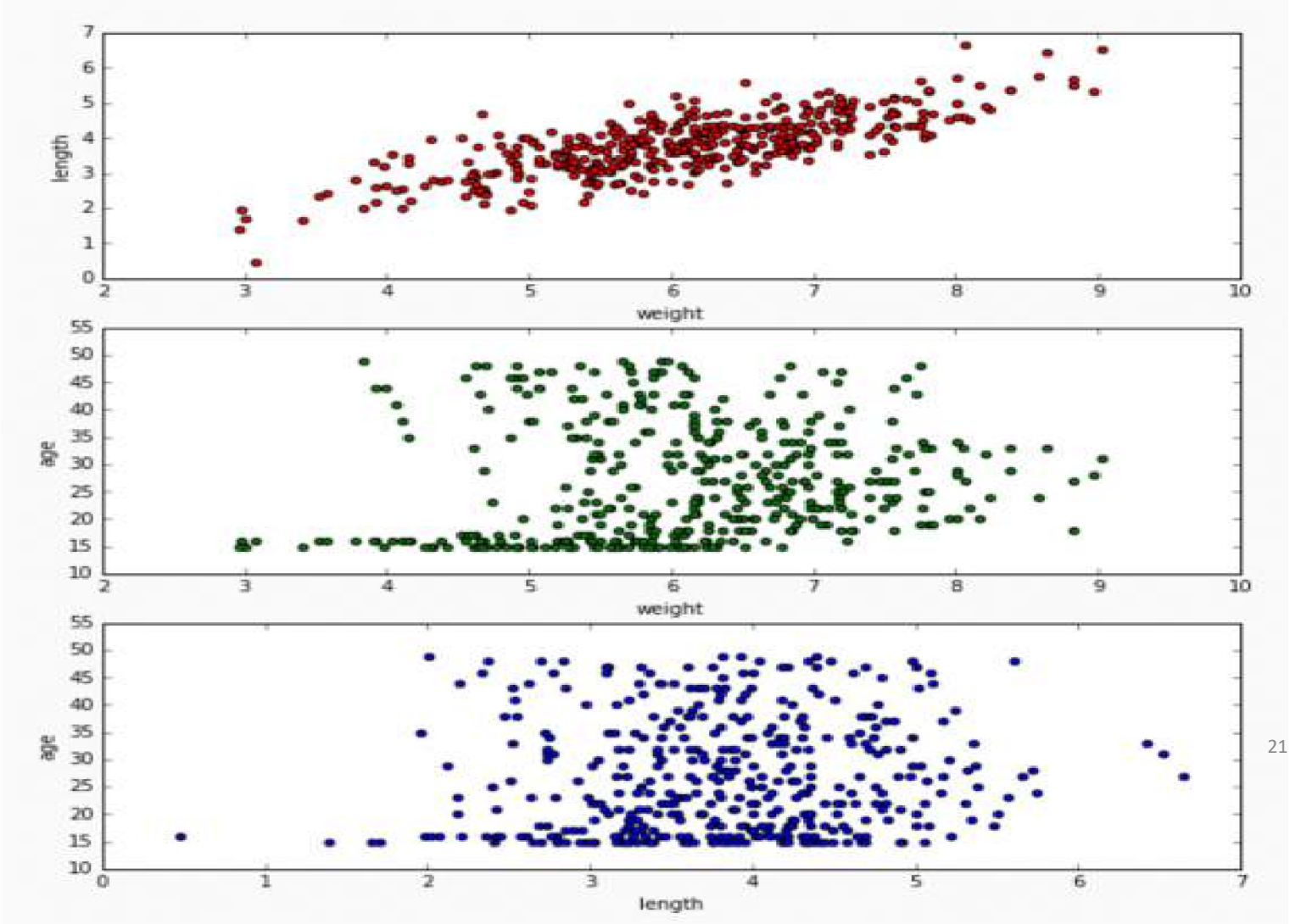
For 3D data, color coding a categorical attribute can be "effective"



This visualizes a set of Iris measurements. The variables are: petal length, sepal length, Iris type (setosa, versicolor, virginica).

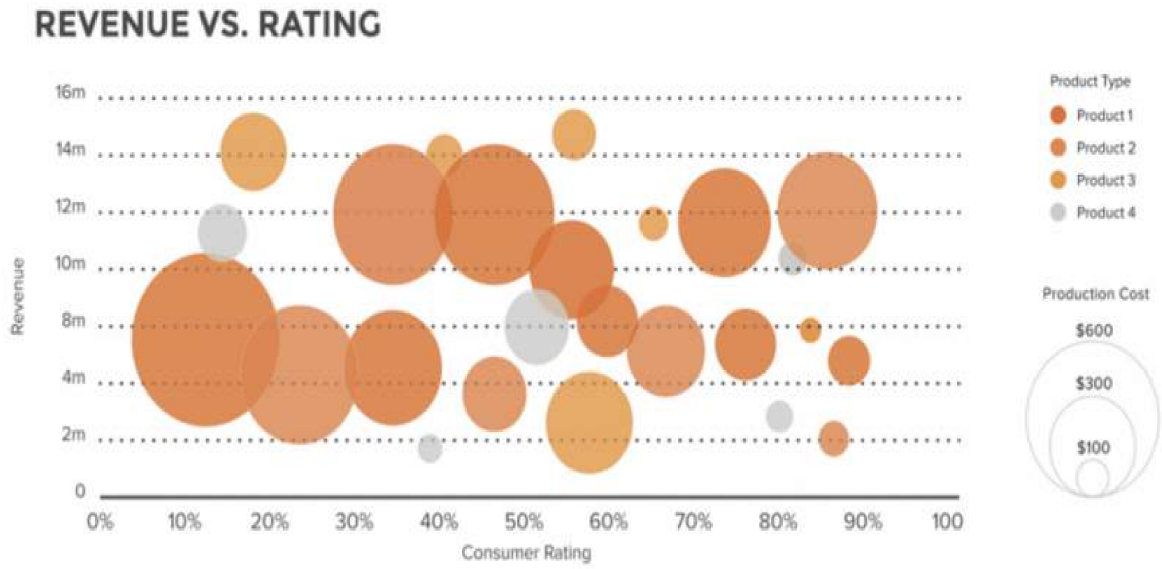
Except when it's not effective.
What could be a better choice?

Relationships may be easier to spot by producing multiple plots of lower dimensionality.

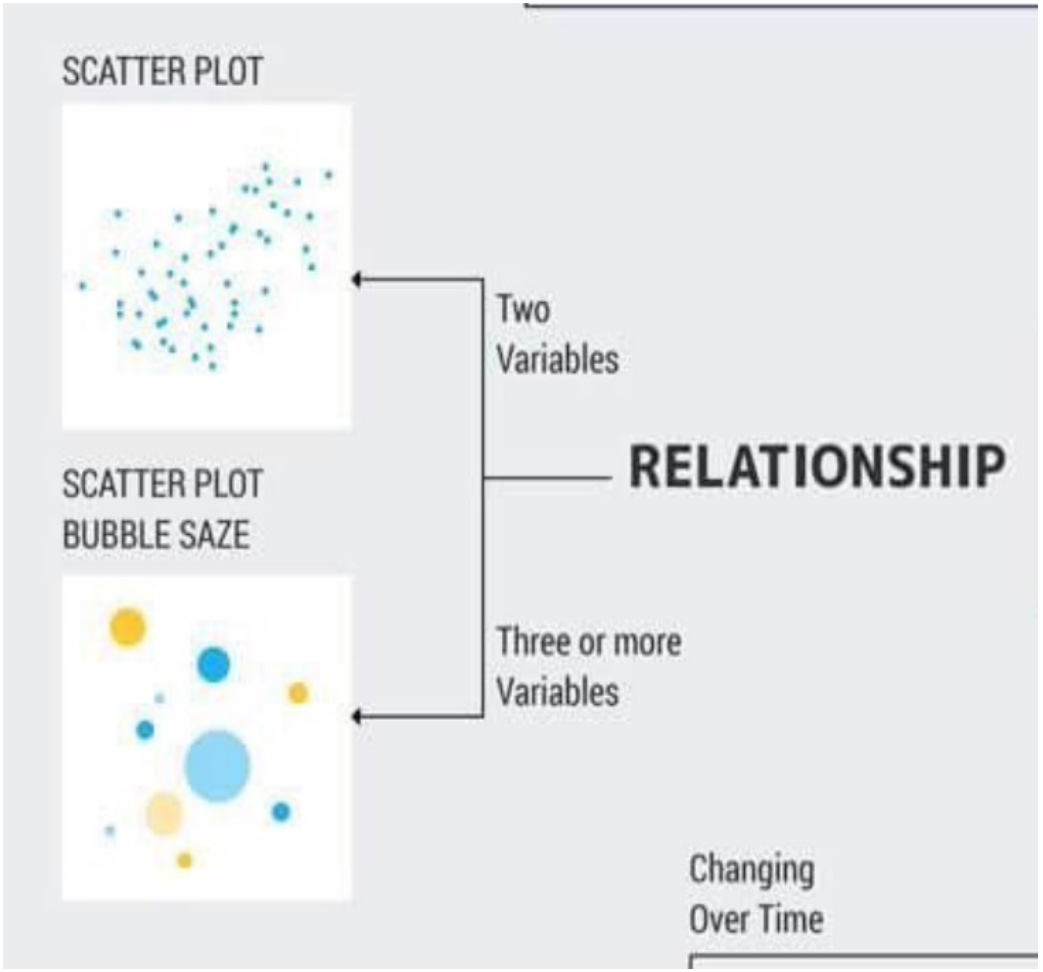


3D CAN WORK

For 3D data, a quantitative attribute can be encoded by size in a bubble chart.



The above visualizes a set of consumer products. The variables are: revenue, consumer rating, product type and product cost.


SCATTER PLOT
Two
Variables
RELATIONSHIP
SCATTER PLOT
BUBBLE SAZE
Three or more
Variables
Changing
Over Time

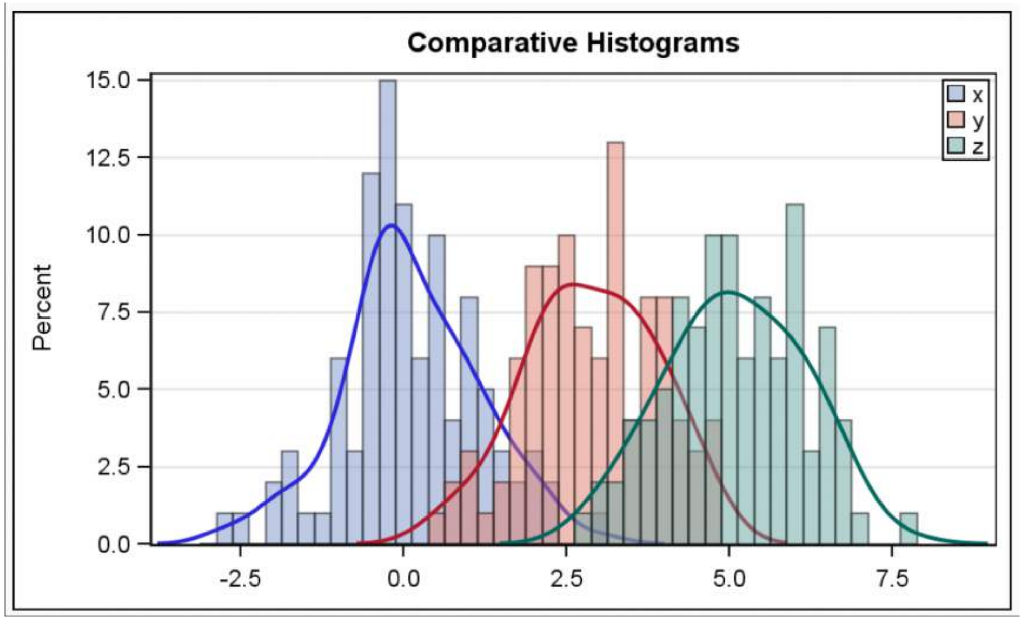
COMPARISON

- These are used to compare the magnitude of values to each other and to easily identify the lowest or highest values in the data.

- If you want to compare values over time, line or bar charts are often the best option.
  - Bar or column charts → Comparisons among items,.
  - Line charts →A sense of continuity.
  - Pie charts for comparison as well

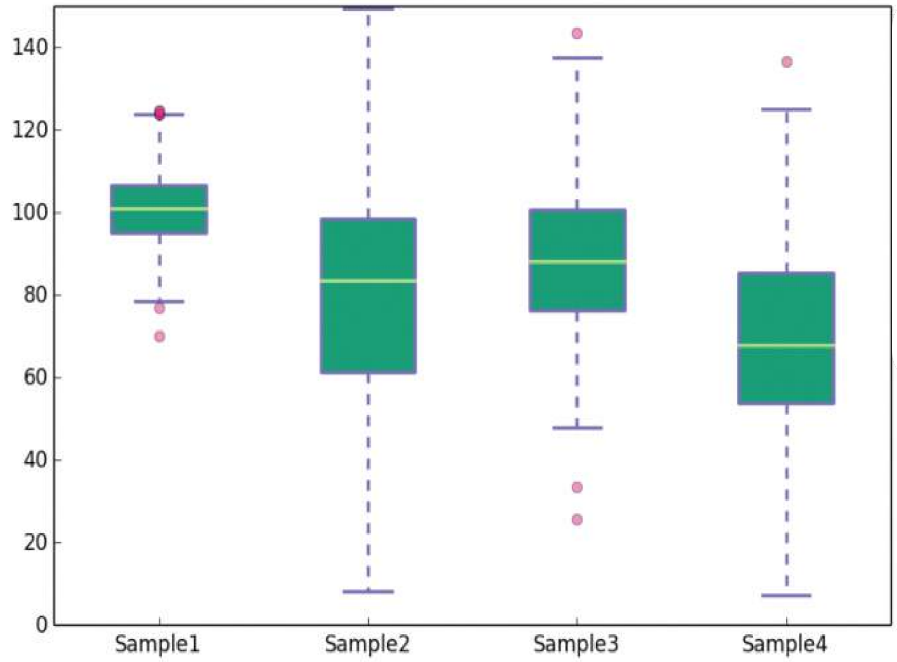
MULTIPLE HISTOGRAMS

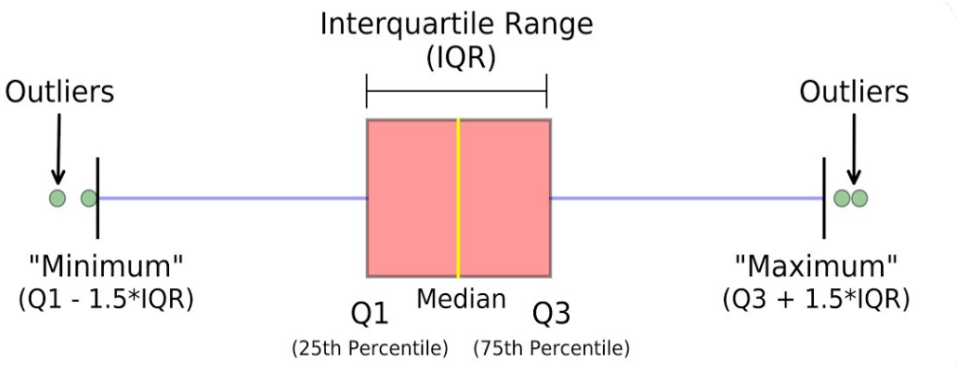
Plotting **multiple histograms** (and **kernel density estimates** of the distribution, here) on the same axes is a way to visualize how different variables compare (or how a variable differs over specific groups).



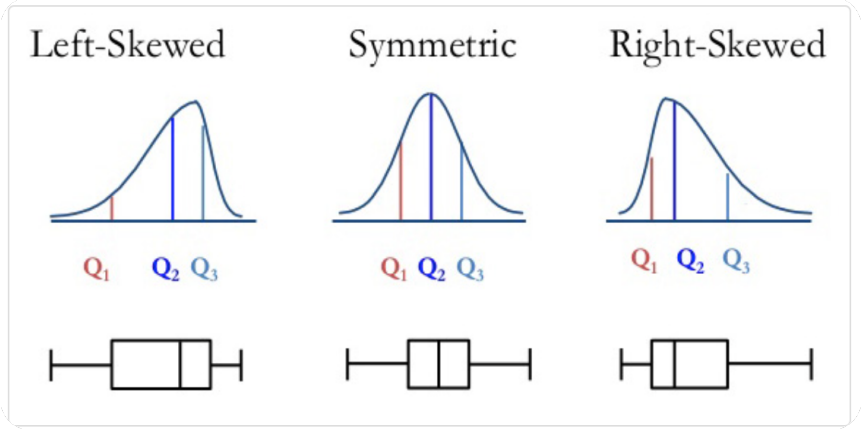
BOXPLOTS

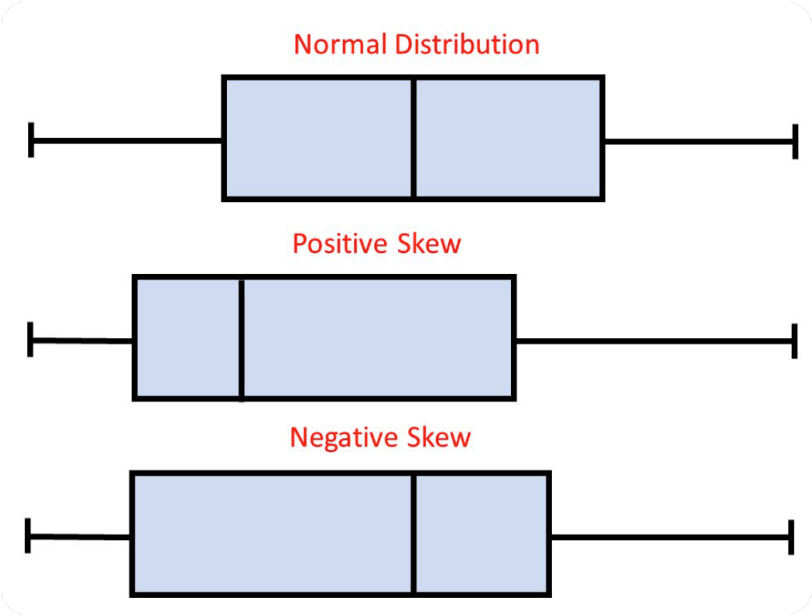
A **boxplot** is a simplified visualization to compare a quantitative variable across groups. It highlights the range, quartiles, median and any outliers present in a data set.

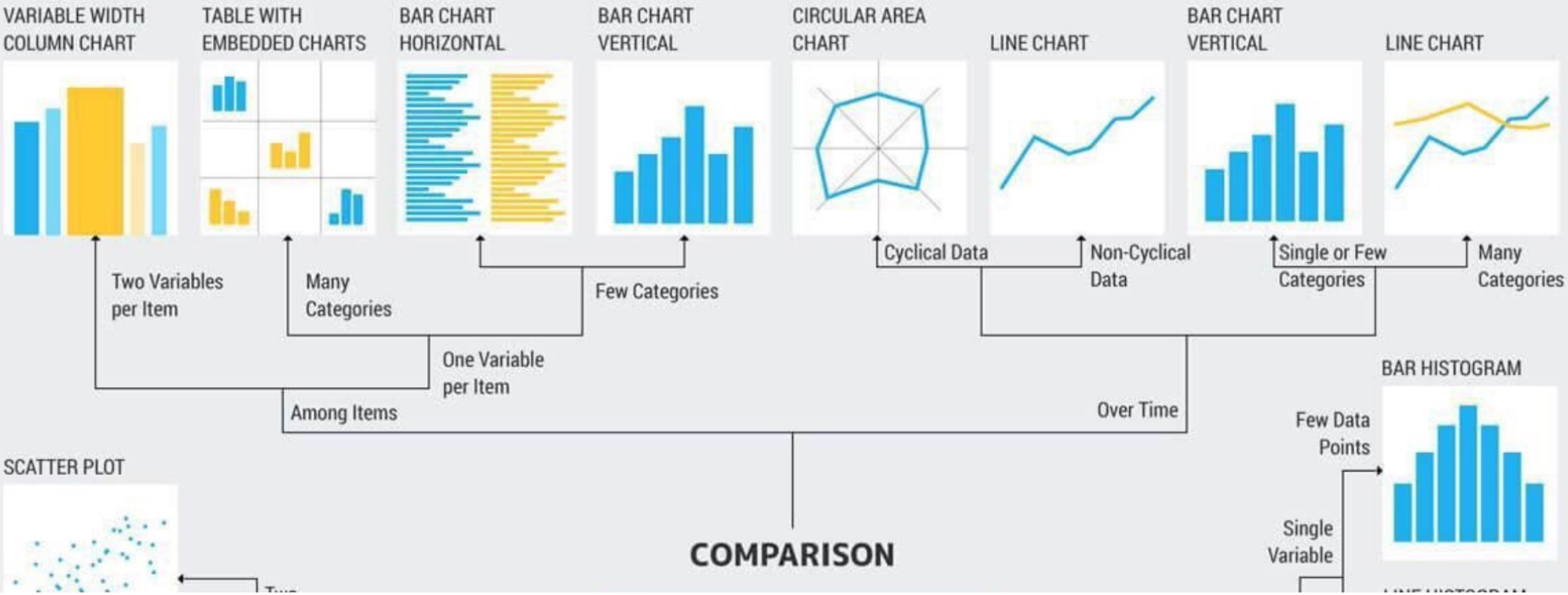




If the data do not appear to be symmetric, does each sample show the same kind of asymmetry?





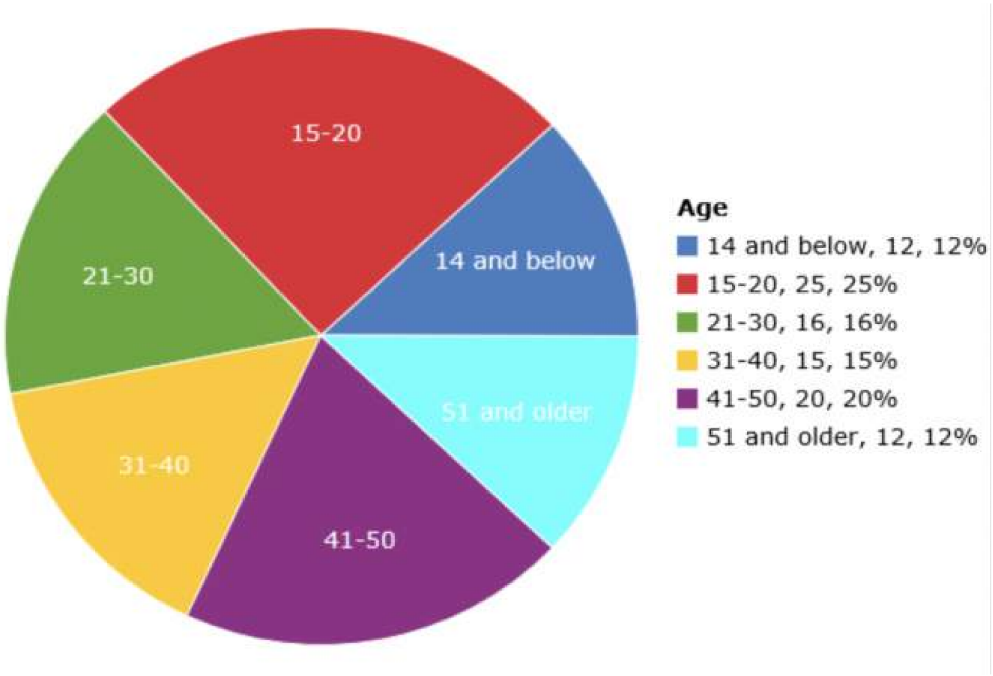

VARIABLE WIDTH COLUMN CHART
TABLE WITH EMBEDDED CHARTS
BAR CHART HORIZONTAL
BAR CHART VERTICAL
CIRCULAR AREA CHART
LINE CHART
BAR CHART VERTICAL
LINE CHART
Two Variables per Item
Many Categories
Few Categories
Cyclical Data
Non-Cyclical Data
Single or Few Categories
Many Categories
One Variable per Item
Among Items
Over Time
BAR HISTOGRAM
Few Data Points
Single Variable
SCATTER PLOT
COMPARISON

COMPOSITION

- Composition charts are used to see how a part of your data compares to the whole.
- Show relative and absolute values.
- They can be used to accurately represent both static and time-series data.

PIE CHART FOR A CATEGORICAL VARIABLE?

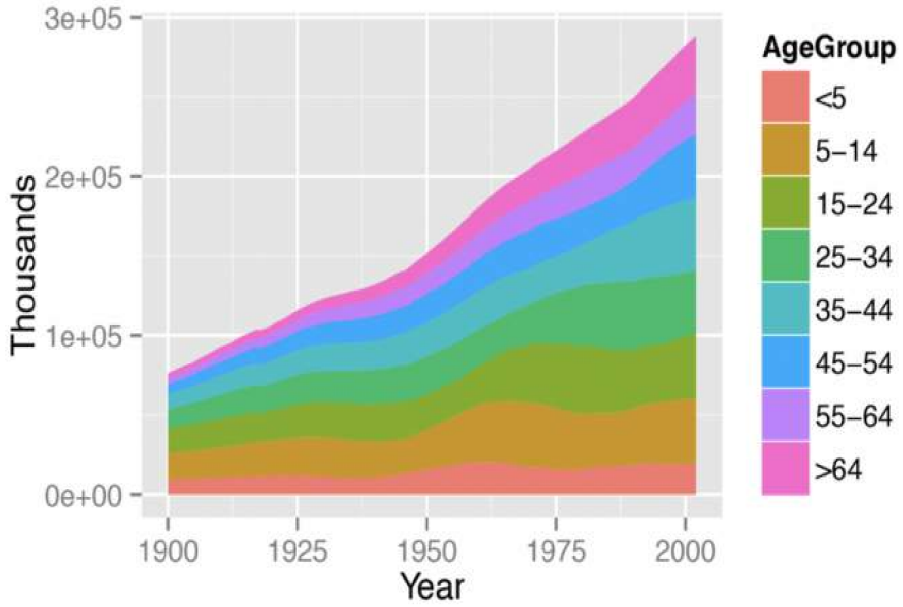
A **pie chart** is a way to visualize the static composition (aka, distribution) of a variable (or single group).

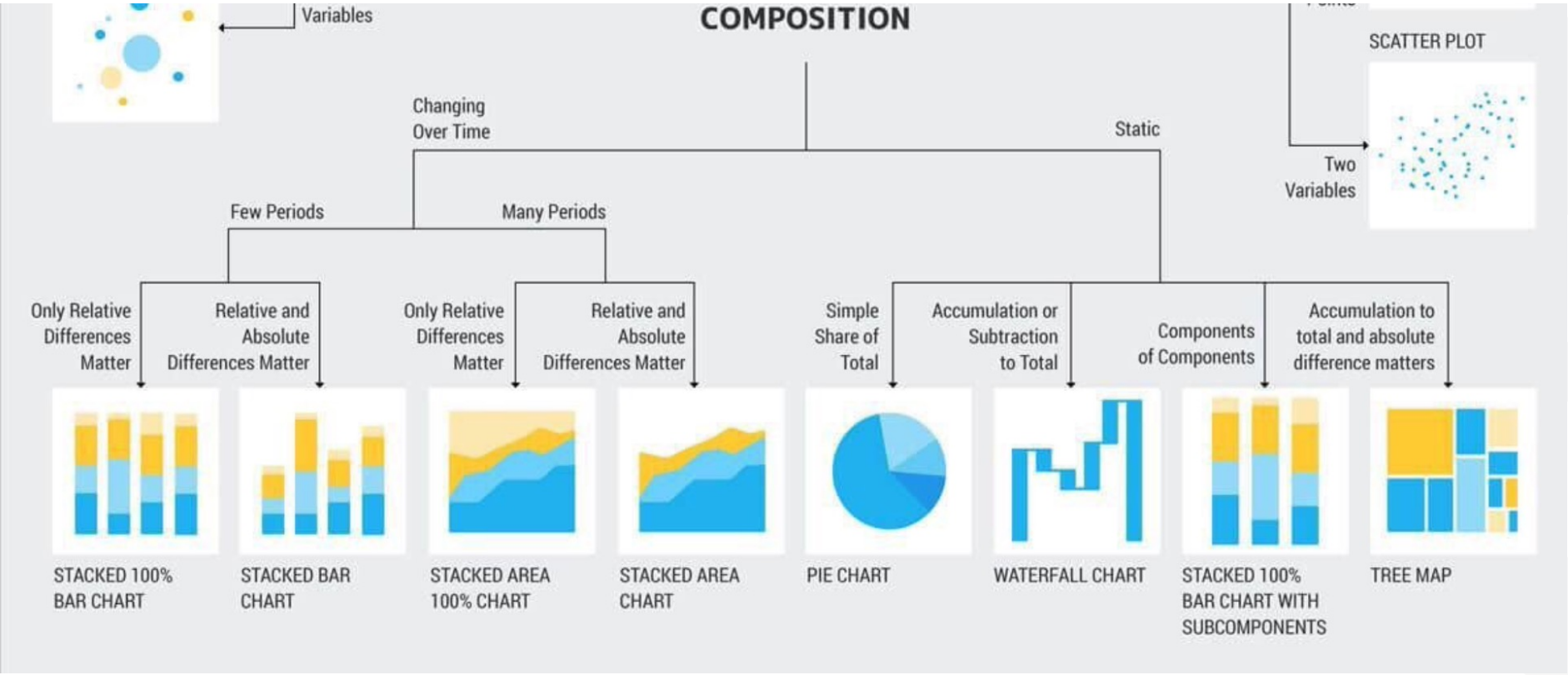


Pie charts are often frowned upon (and bar charts are used instead). Why?

STACKED AREA GRAPH TO SHOW TREND OVER TIME

A **stacked area graph** is a way to visualize the composition of a group as it changes over time (or some other quantitative variable). This shows the relationship of a categorical variable (AgeGroup) to a quantitative variable (year).




Variables
COMPOSITION
SCATTER PLOT
Changing Over Time
Static
Two Variables
Few Periods
Many Periods
Only Relative Differences Matter
Relative and Absolute Differences Matter
Only Relative Differences Matter
Relative and Absolute Differences Matter
Simple Share of Total
Accumulation or Subtraction to Total
Components of Components
Accumulation to total and absolute difference matters
STACKED 100% BAR CHART
STACKED BAR CHART
STACKED AREA 100% CHART
STACKED AREA CHART
PIE CHART
WATERFALL CHART
STACKED 100% BAR CHART WITH SUBCOMPONENTS
TREE MAP

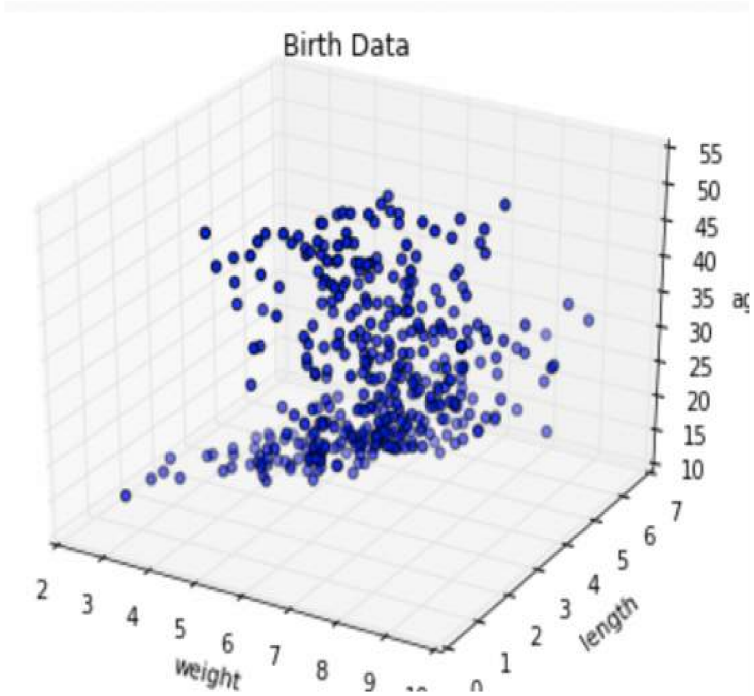
[NOT] ANYTHING IS POSSIBLE!

Often your dataset seem too complex to visualize:

- Data is too high dimensional (how do you plot 100 variables on the same set of axes?)
- Some variables are categorical (how do you plot values like Cat or No?)

More dimensions not always better

When the data is high dimensional, a scatter plot of all data attributes can be impossible or unhelpful



An Example

AN EXAMPLE

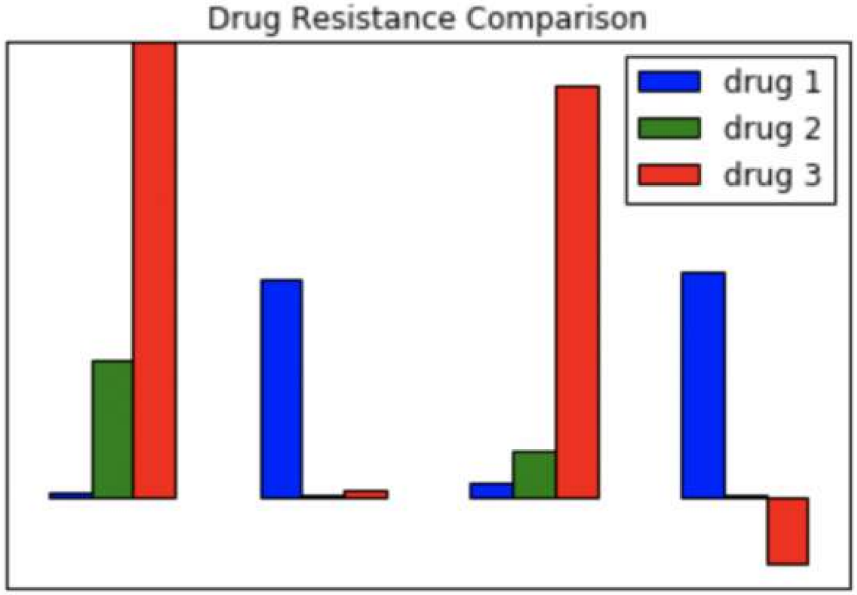
Use some simple visualizations to explore the following dataset:

Bacteria Name	Group No.	Res. to Drug 1	Res. to Drug 2	Res. to Drug 3
Brucella abortus	1	0.1	3	49
Diplococcus pneumoniae	2	4.75	0.007	0.125
Aerobacter aerogenes	1	0.3	1	47.2
Streptococcus viridans	2	4.9	0.03	-1.45

How should we begin?

AN EXAMPLE (1)

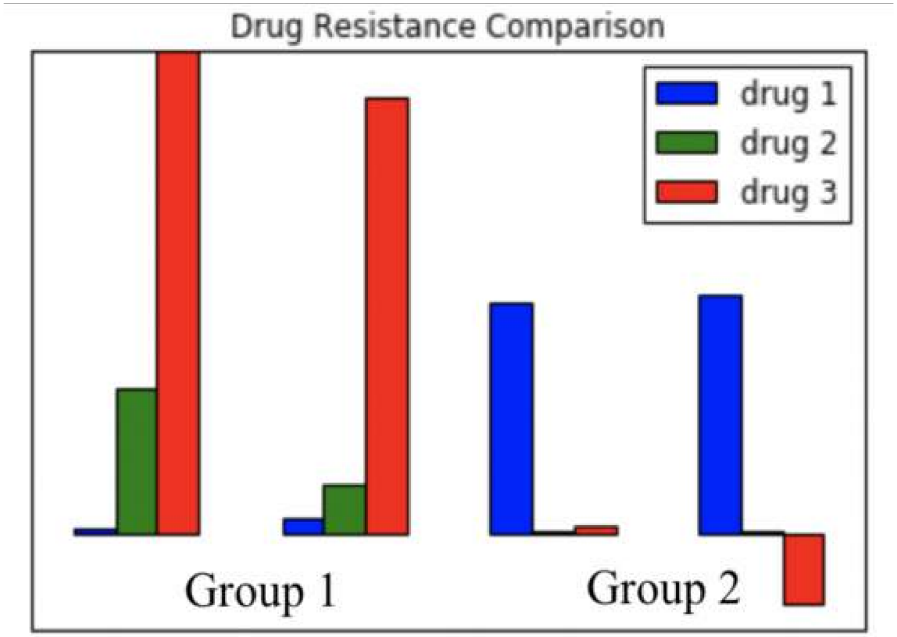
A bar graph showing resistance of each bacteria to each drug:



What do you notice?

AN EXAMPLE (2)

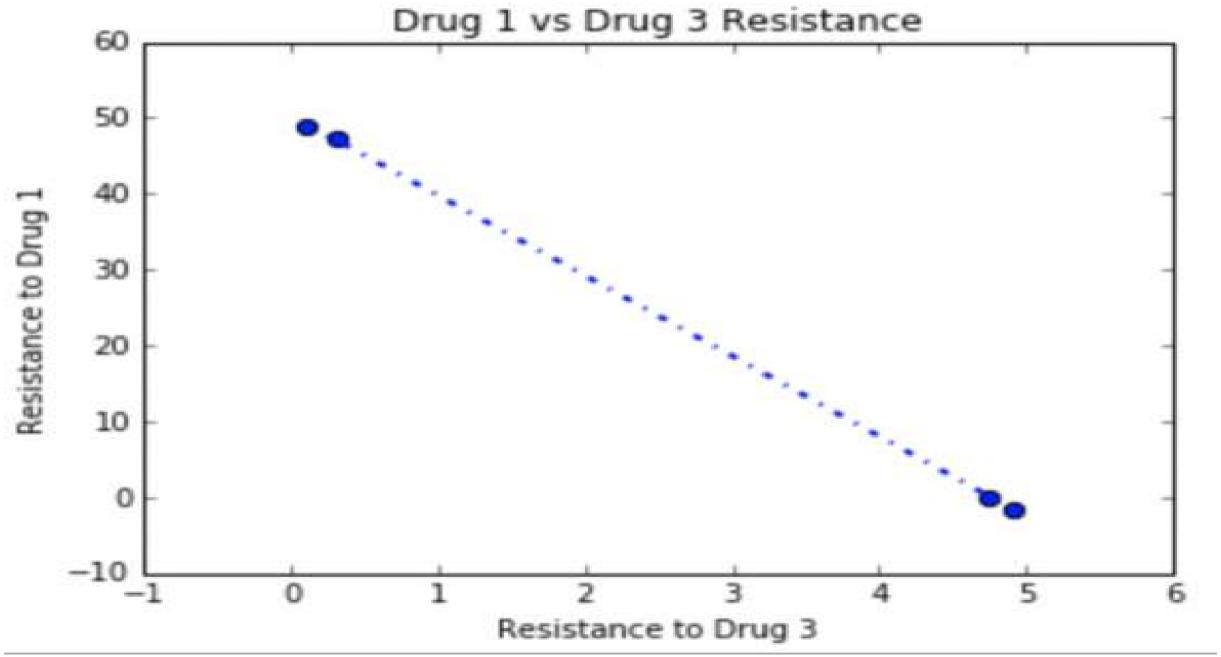
Bar graph showing resistance of each bacteria to each drug (grouped by Group Number):



Now what do you notice?

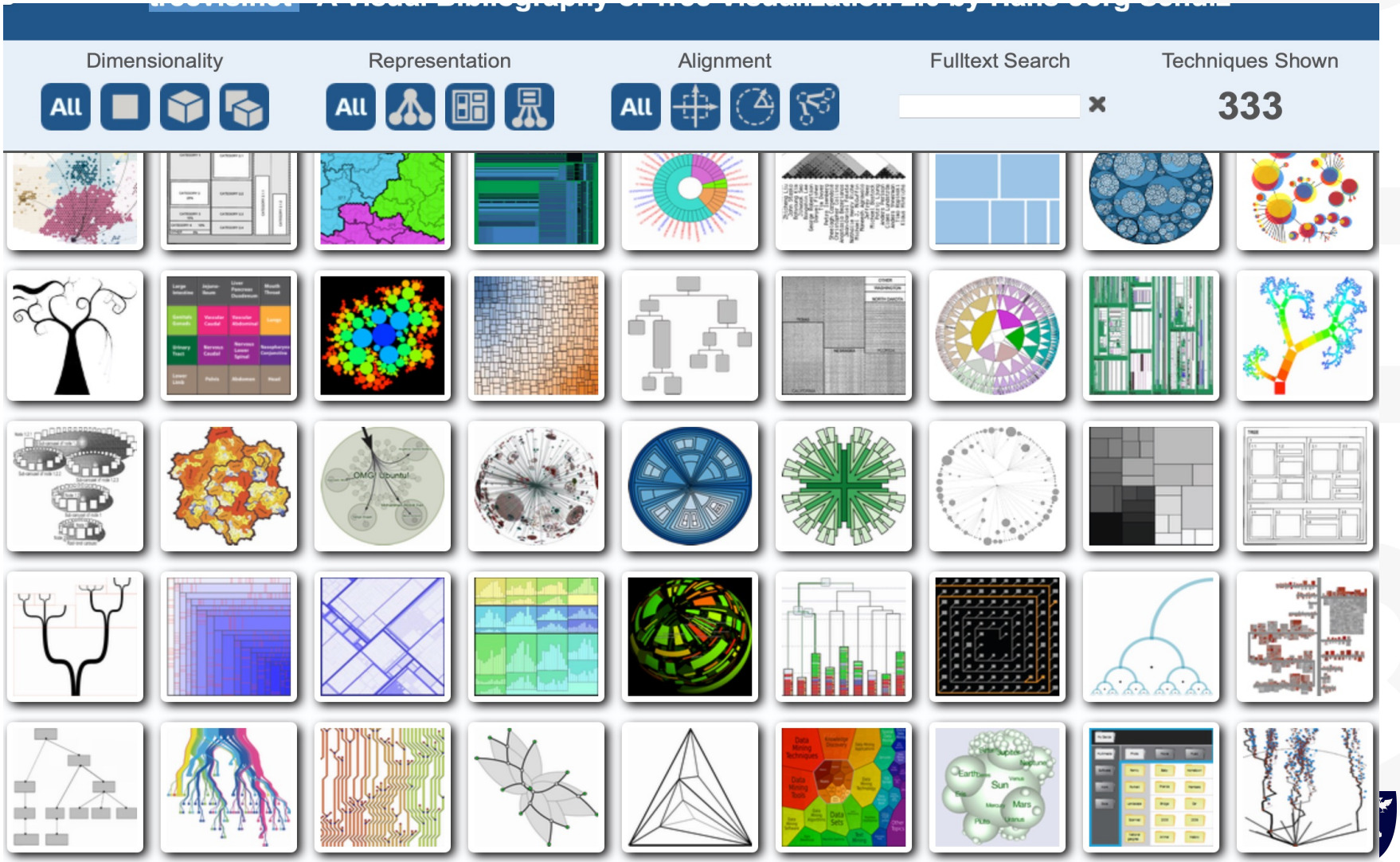
AN EXAMPLE (3)

Scatter plot of Drug #1 vs Drug #3 resistance:



Key: the process of data exploration is iterative (visualize for trends, re- visualize to confirm)!

TREEVIS.NET


Dimensionality
All
Representation
All
Alignment
All
Fulltext Search
Techniques Shown
333

GEPHI.ORG

gephi makes graphs handy

Home Features Learn Develop Plugins Services Consortium

The Open Graph Viz Platform

Gephi is the leading visualization and exploration software for all kinds of graphs and networks. Gephi is open-source and free.

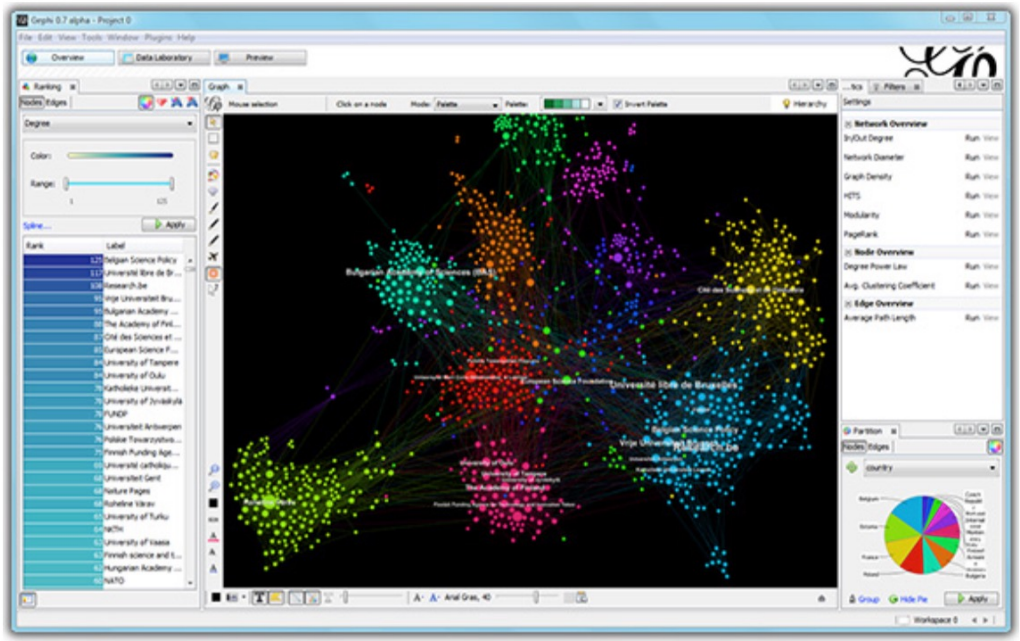
Runs on Windows, Mac OS X and Linux.

Learn More on Gephi Platform »

Download FREE
Gephi 0.9.2

Release Notes | System Requirements

- Features
- Quick start
- Screenshots
- Videos



Support us! We are non-profit. Help us to innovate and empower the community by donating only 8€:

Donate

TOOLS FOR INTERACTIVE GRAPHICS

- R/shiny
- plotly/dash
- Tableau
- d3.js
- vega-lite/vega

Q1: WHAT ARE SOME IMPORTANT FEATURES OF A GOOD DATA VISUALIZATION?

The data visualization should be light and must highlight essential aspects of the data; looking at important variables, what is relatively important, what are the trends and changes. Besides, data visualization must be visually appealing but should not have unnecessary information in it.

One can answer this question in multiple ways: from technical points to mentioning key aspects, but be sure to remember saying these points:

1. Maximize data to ink ratio: show the data
2. Don't lie with scale: minimize
3. Minimize chart-junk: show data variation, not design variation
4. Clear, detailed and thorough labeling

QUESTION2: WHAT IS A SCATTER PLOT? FOR WHAT TYPE OF DATA IS SCATTER PLOT USUALLY USED FOR?

- A scatter plot is a chart used to plot a correlation between two or more variables at the same time. It’s usually used for numeric data.

QUESTION3: WHAT TYPE OF PLOT WOULD YOU USE IF YOU NEED TO DEMONSTRATE “RELATIONSHIP” BETWEEN VARIABLES/PARAMETERS?

- When we are trying to show the relationship between 2 variables, scatter plots or charts are used. When we are trying to show “relationship” between three variables, bubble charts are used.

THANK YOU

VISIT US

WWW.XJTLU.EDU.CN

FOLLOW US

@XJTLU

Xi'an Jiaotong-Liverpool University
西交利物浦大學