Data Valorization:
Data Visualization

Lionel Fillatre

fillatre@unice.fr

Topics

- Introduction
- Basic plots
- Distribution of a single attribute
- Relation between attributes
- Visualization of high-dimensional objects
- Conclusion

1 Introduction

Data Visualization

- Mapping of data to a visual format
  - Data objects, attributes, and relations
- Humans are good at understanding visual information
  - See patterns and trends
  - Detect outliers
- The adage "A picture is worth a thousand words" refers to the idea that a complex idea can be conveyed with just a single still image.
- This adage also characterizes one of the main goals of visualization: making it possible to absorb large amounts of data quickly.

Example: French invasion of Russia

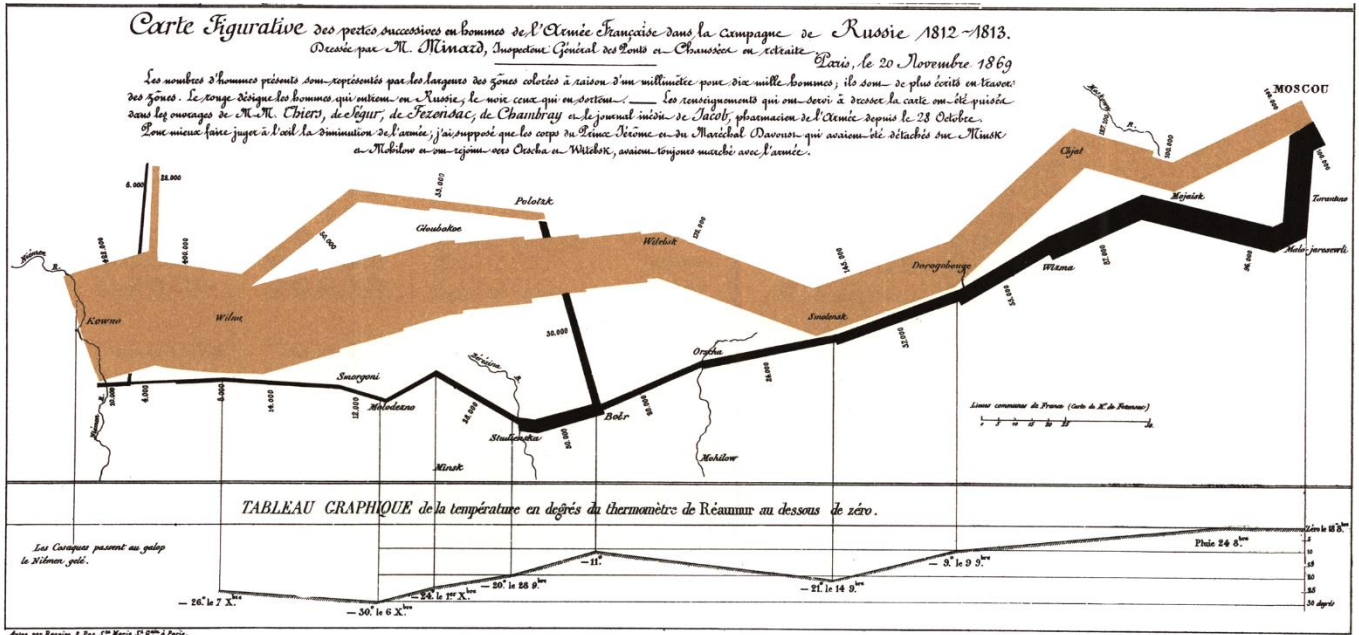
- The French invasion of Russia began on 24 June 1812 when Napoleon's Grande Armée crossed the Neman River in an attempt to engage and defeat the Russian army.

- The Grande Armée was a very large force, numbering 680,000 soldiers (including 300,000 of French departments). It was the largest army ever assembled in the history of warfare up to that point.

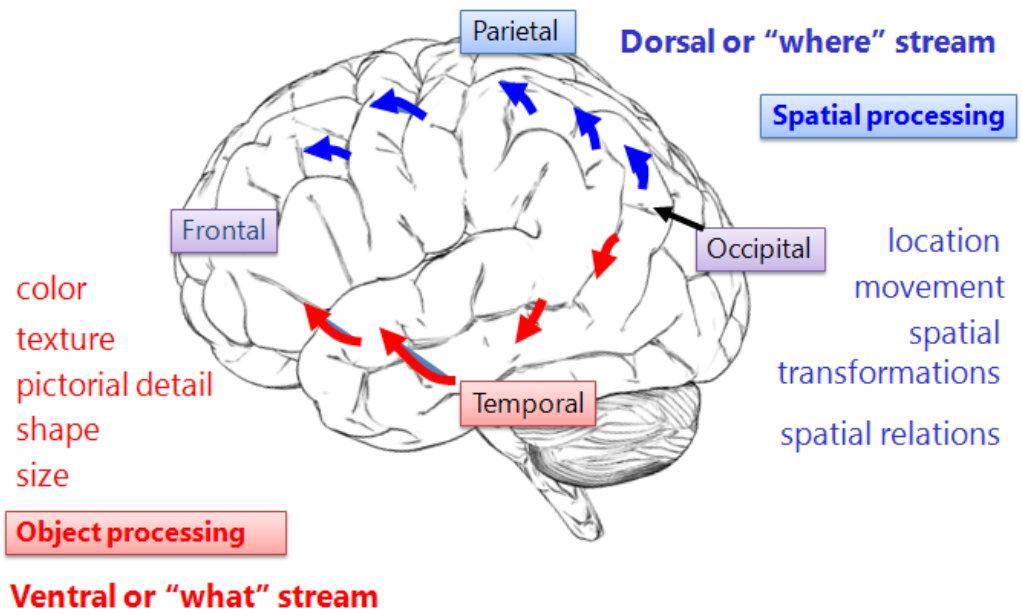
- Napoleon's invasion of Russia is listed among the most lethal military operations in world history.

- The Russian campaign was decisive for the Napoleonic Wars and led to Napoleon's defeat and exile on the island of Elba.

Famous example: Minard's graphic (1781-1870)



Human Visual System

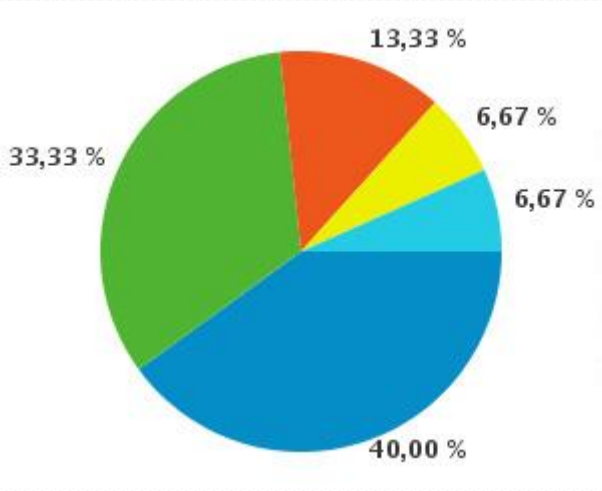


Main Elements

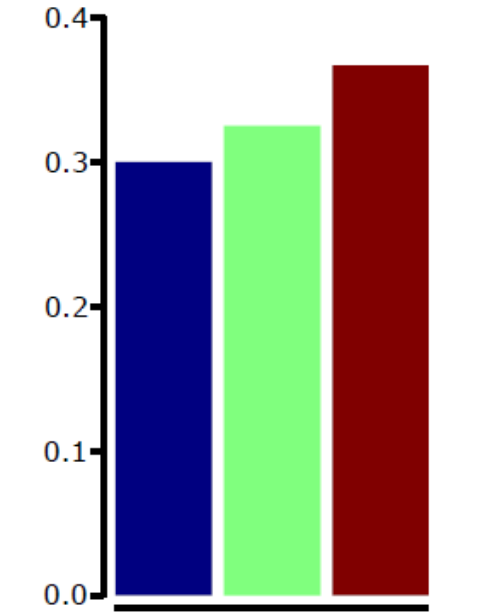
- Representation:
  - How will you map objects, attributes, and relations to visual elements?
  - Positions, lengths, colors, areas, orientation
- Arrangement:
  - How will you display the visual elements?
  - Viewpoint, transparency, separation, grouping
- Selection:
  - How will you handle a large number of attributes and data objects?
  - Display a subset, focus on a region of interest, show summaries

Representation

- **Area represents proportion (pie chart)**



- **Height represents proportion**



Arrangement

- Placement of visual elements
  - Can make a great difference in how easy it is to interpret data

	1	2	3	4	5	6
1	0	1	0	1	1	0
2	1	0	1	0	0	1
3	0	1	0	1	1	0
4	1	0	1	0	0	1
5	0	1	0	1	1	0
6	1	0	1	0	0	1
7	0	1	0	1	1	0
8	1	0	1	0	0	1
9	0	1	0	1	1	0

	6	1	3	2	5	4
4	1	1	1	0	0	0
2	1	1	1	0	0	0
6	1	1	1	0	0	0
8	1	1	1	0	0	0
5	0	0	0	1	1	1
3	0	0	0	1	1	1
9	0	0	0	1	1	1
1	0	0	0	1	1	1
7	0	0	0	1	1	1

Selection

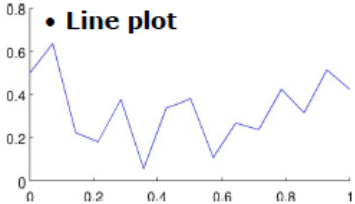
- Elimination or de-emphasis of certain objects or attributes
- A subset of attributes
  - Why? A graph can only show so many attributes – focus on the relevant
  - How?
    - Dimensionality reduction
    - Plot pairs of attributes
- A subset of objects
  - Why? A graph can only show so many objects – focus on the relevant
  - How?
    - Random sampling
    - Display of region of interest

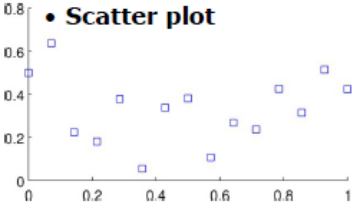
Main Types of plots

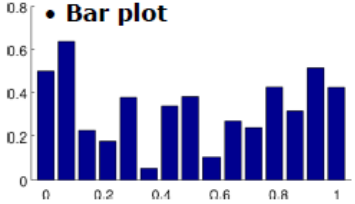
- Basic plots
- Distribution of a single attribute
  - Histogram
  - Empirical cumulative distribution
  - Percentile plots
  - Box plot
- Relation between attributes
  - 2-d histogram
  - Scatter plots
- Visualization of high-dimensional objects
  - Matrix plots
  - Parallel coordinates
  - Radar Charts (also known as star plots)
  - Force-directed network
  - Arc diagram

2 Basic plots

Basic plots


• Line plot
0.8
0.6
0.4
0.2
0
0
0.2
0.4
0.6
0.8
1

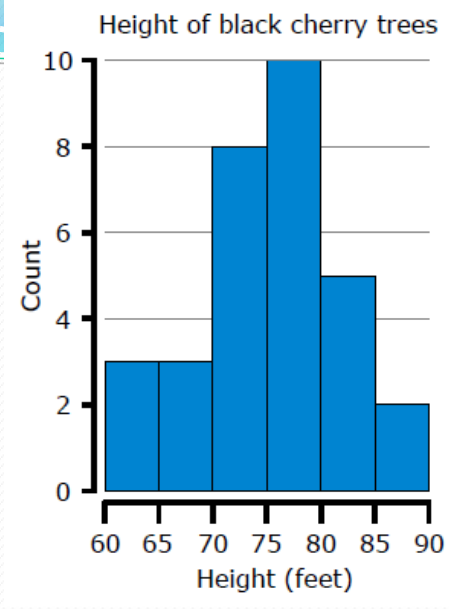

• Scatter plot
0.8
0.6
0.4
0.2
0
0
0.2
0.4
0.6
0.8
1


• Bar plot
0.8
0.6
0.4
0.2
0
0
0.2
0.4
0.6
0.8
1

3 Distribution of a single attribute

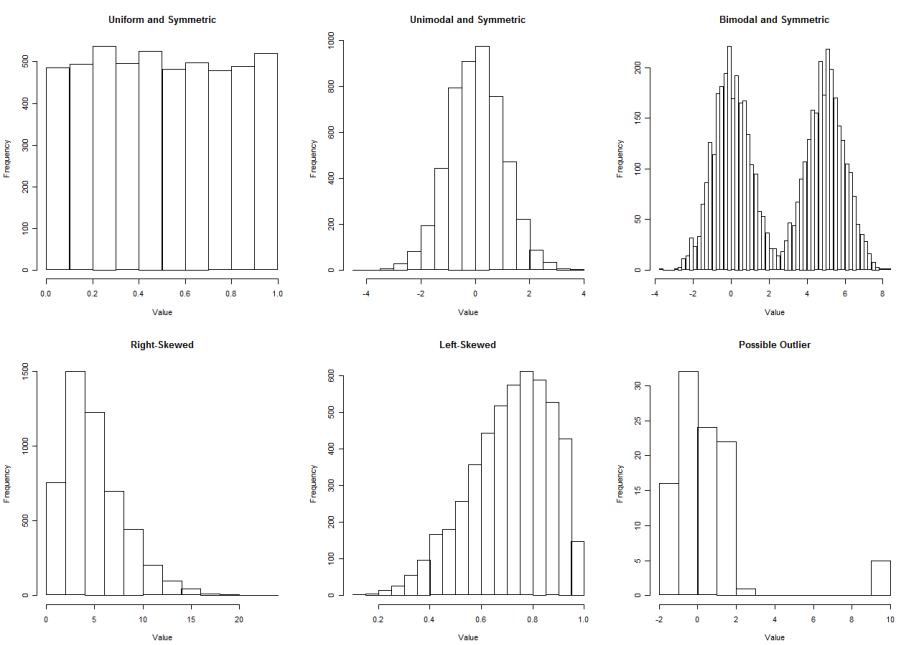
Histograms

- Shows distribution of a single variable
  - Divide the values into bins
  - Bar plot of the number of values in bin
  - Height indicates count of values
  - Shape determined by
    - Distribution of data
    - Number of bins / bin width



$$H = \{60, 64, 64, 66, 67, 69, 71, 72, 72, 72, 72, 73, 74, 74, 75, 75, 76, 76, 76, 77, 77, 78, 78, 79, 80, 80, 81, 82, 84, 85, 85, 89\}$$

Some examples



The Iris data set

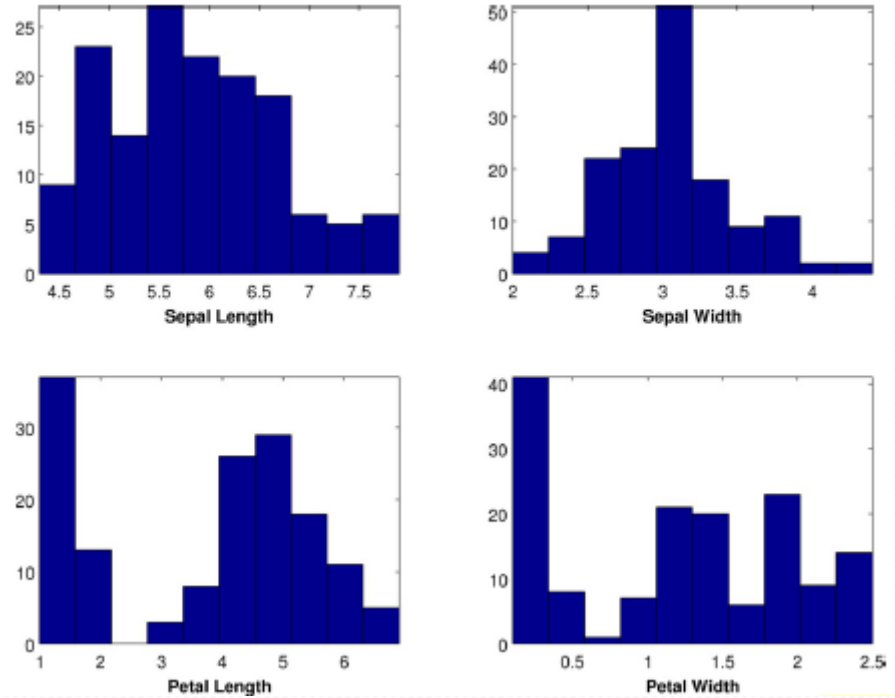
- Three flowers

50 instances of each class, 150 in total

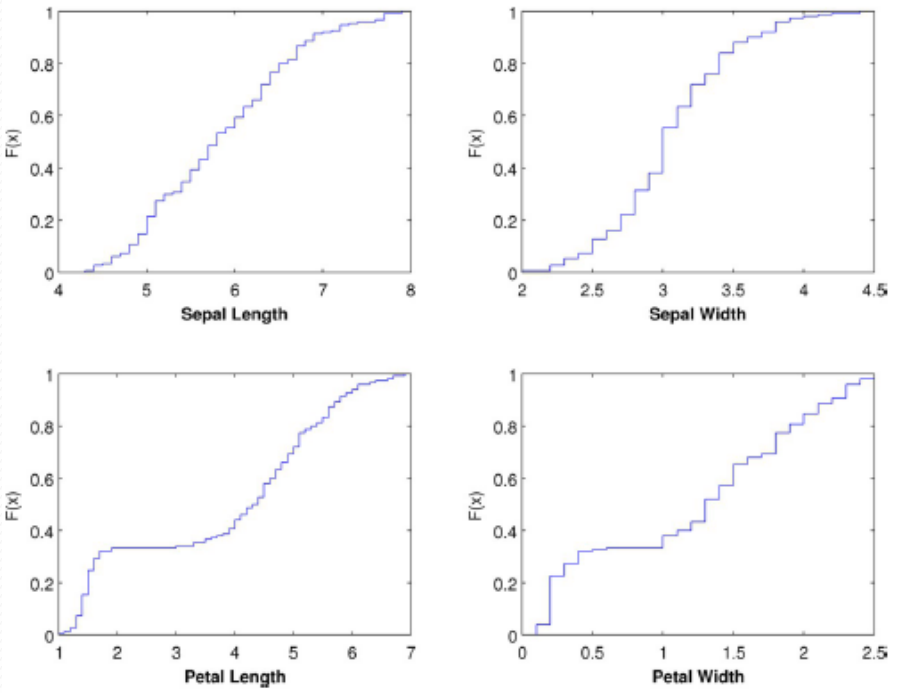
- Attributes
  - Sepal (outermost leaves)
    - length in cm
    - width in cm
  - Petal (innermost leaves)
    - length in cm
    - width in cm
  - Class of flower
    - Iris Setosa
    - Iris Versicolour
    - Iris Virginica

Fower ID	Attribute			
	Sepal Length	Sepal Width	Petal Length	Petal Width
1	5.1	3.5	1.4	0.2
2	4.9	3.0	1.4	0.2
3	4.7	3.2	1.3	0.2
4	4.6	3.1	1.5	0.2
.	.	.	.	.
.	.	.	.	.
150	5.9	3.0	5.1	1.8

Histograms of the Iris data attributes

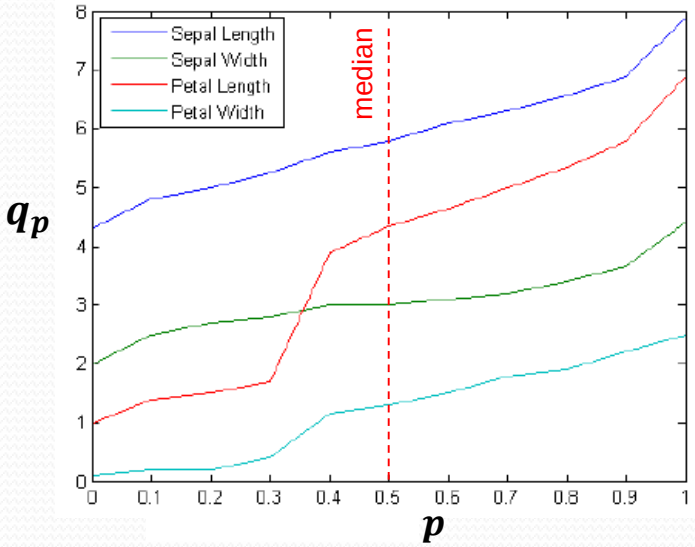


Empirical cumulative distributions

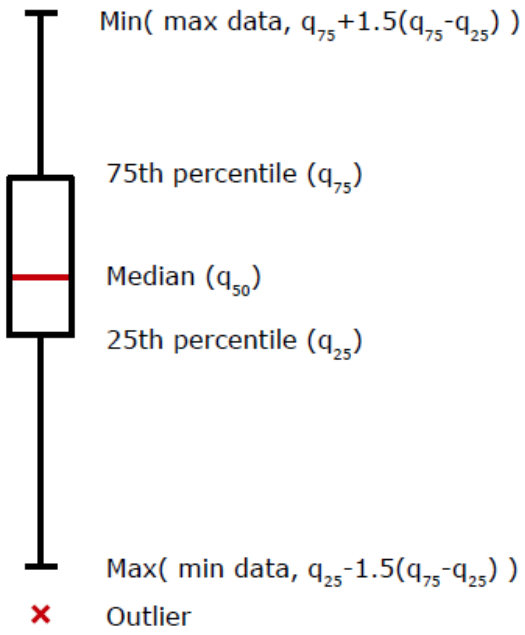


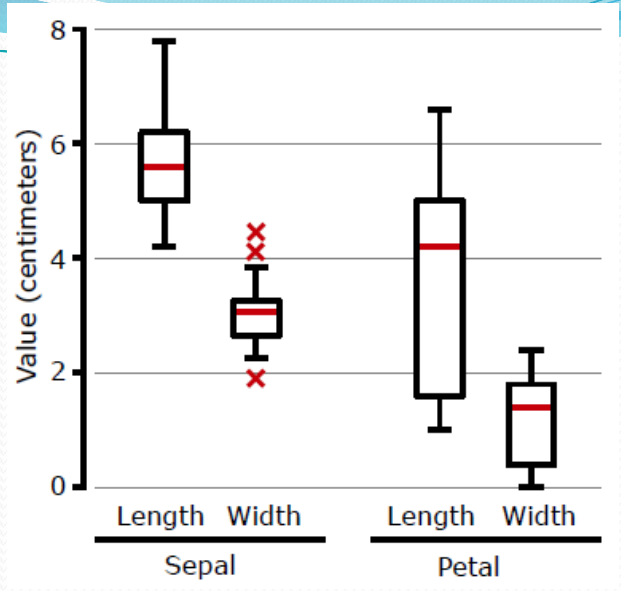
Percentile plots

- **Percentiles:** Given a continuous attribute $\boldsymbol{x}$ and a number $\boldsymbol{p}$ between 0 and 100 (or equivalently 0 and 1), the $\boldsymbol{p}$-th percentile is a value $\boldsymbol{q_p}$ of $\boldsymbol{x}$ such that $\boldsymbol{p}$ percent of the observed values of $\boldsymbol{x}$ are less than $\boldsymbol{q_p}$.



Box plots





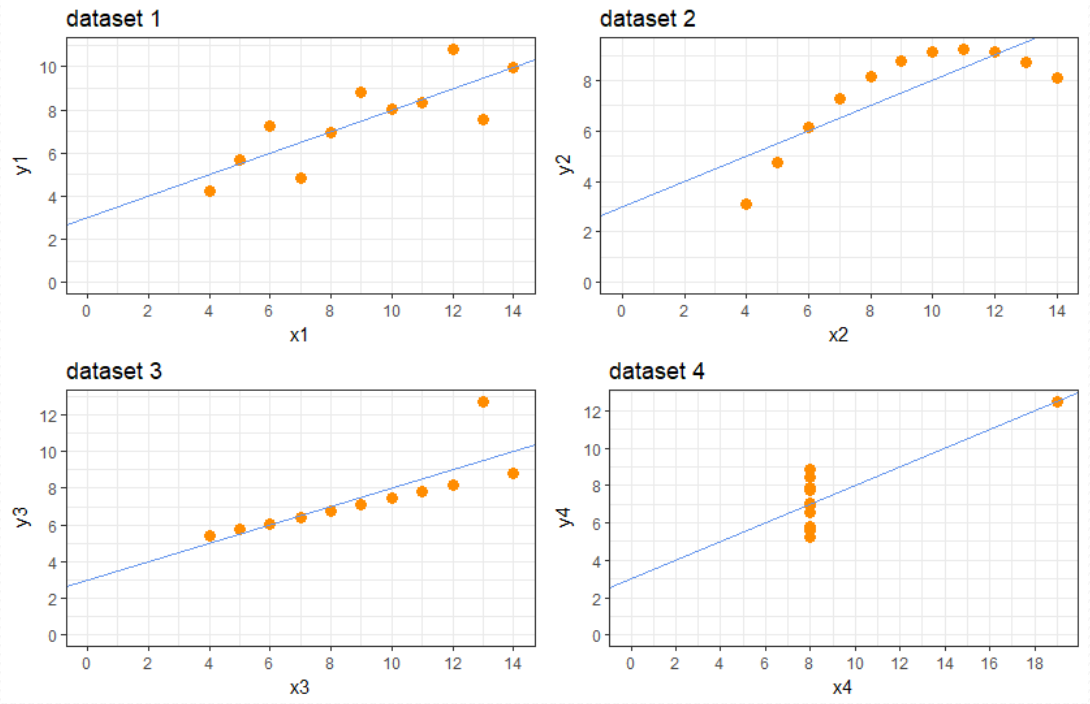
- $q_{75} - q_{25}$=inter-quartile range (IQR)
- Upper fence if outlier =$q_{75} + 1.5 * IQR$
- Lower fence if outlier =$q_{25} - 1.5 * IQR$

Illustration : Anscombe's Quartet

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

- Do you observe something?
- Let's see how they appear when we visualize them.

Anscombe's Quartet



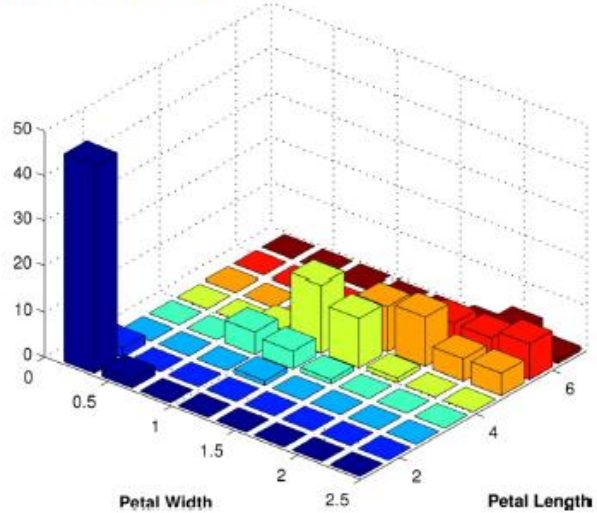
R Code to create plots

```
library(ggplot2)
library(gridExtra)
data(anscombe)

p1 <- ggplot(anscombe) + geom_point(aes(x1, y1), color = "darkorange", size = 3)
+ theme_bw() + scale_x_continuous(breaks = seq(0, 20, 2))
+ scale_y_continuous(breaks = seq(0, 12, 2))
+ geom_abline(intercept = 3, slope = 0.5, color = "cornflowerblue") + expand_limits(x = 0, y = 0) + labs(title = "dataset 1")

p2 <- ggplot(anscombe) + geom_point(aes(x2, y2), color = "darkorange", size = 3)
+ theme_bw() + scale_x_continuous(breaks = seq(0, 20, 2))
+ scale_y_continuous(breaks = seq(0, 12, 2))
+ geom_abline(intercept = 3, slope = 0.5, color = "cornflowerblue") + expand_limits(x = 0, y = 0) + labs(title = "dataset 2")

p3 <- ggplot(anscombe) + geom_point(aes(x3, y3), color = "darkorange", size = 3)
+ theme_bw() + scale_x_continuous(breaks = seq(0, 20, 2))
+ scale_y_continuous(breaks = seq(0, 12, 2))
+ geom_abline(intercept = 3, slope = 0.5, color = "cornflowerblue") + expand_limits(x = 0, y = 0) + labs(title = "dataset 3")

p4 <- ggplot(anscombe) + geom_point(aes(x4, y4), color = "darkorange", size = 3)
+ theme_bw() + scale_x_continuous(breaks = seq(0, 20, 2))
+ scale_y_continuous(breaks = seq(0, 12, 2))
+ geom_abline(intercept = 3, slope = 0.5, color = "cornflowerblue") + expand_limits(x = 0, y = 0) + labs(title = "dataset 4")

grid.arrange(p1, p2, p3, p4, ncol=2, nrow = 2)
```

4 Relation between attributes

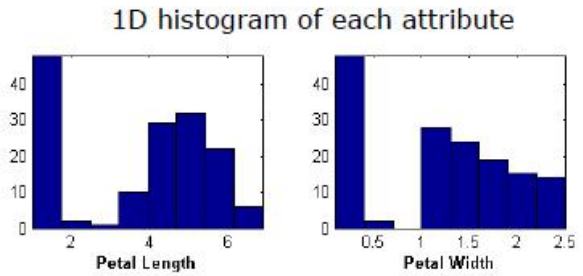
Two-dimensional histograms

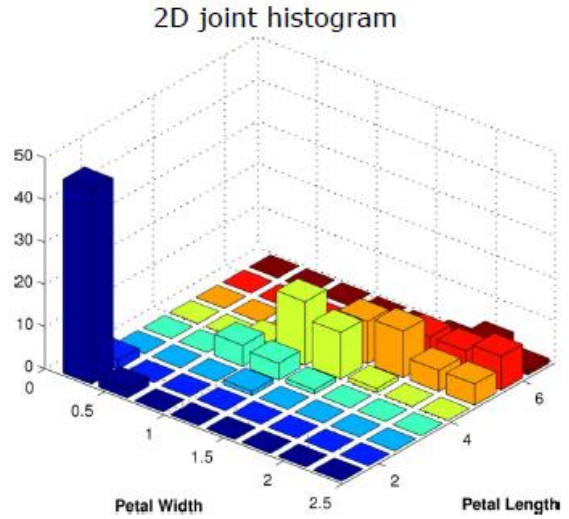
- Shows joint distribution of two variables



Two-dimensional histograms

Let B be the number of bins for each mode in the 1D and 2D histogram (B=8 in the figures below). How many values are there to be estimated in the 1D and 2D histogram as a function of B?

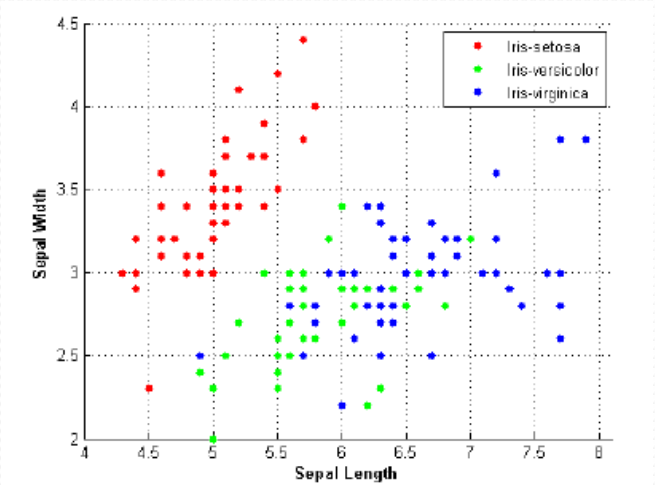


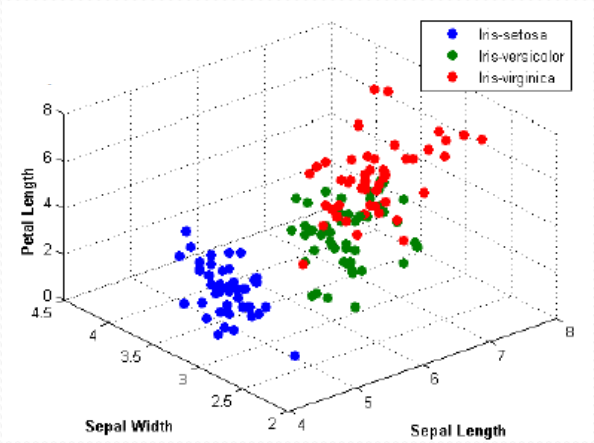


Scatter plots

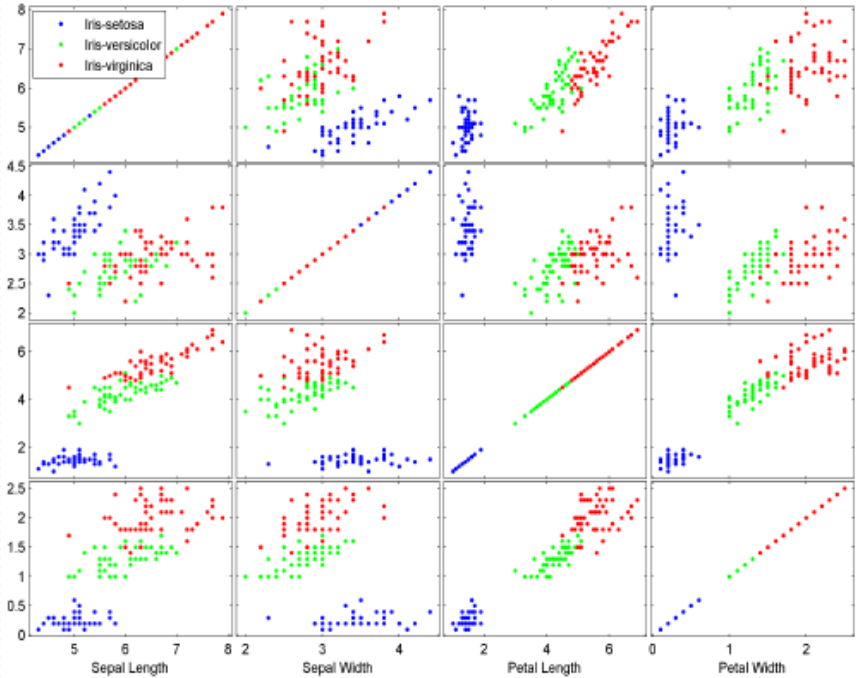
Shows relation between attributes :

- Assess dependence between attributes
- Used with classes to assess separability





Scatter plot matrix



WordCloud

- One of the best suited visualizations for textual data is the WordCloud.
- The wordcloud brings the more frequent ones to the center and enlarges them, giving us a clear picture of what the general idea of the text depicts.
- The following wordcloud shows that love is the most frequent positive term used in the analyzed tweet dataset.

5 Visualization of high-dimensional objects

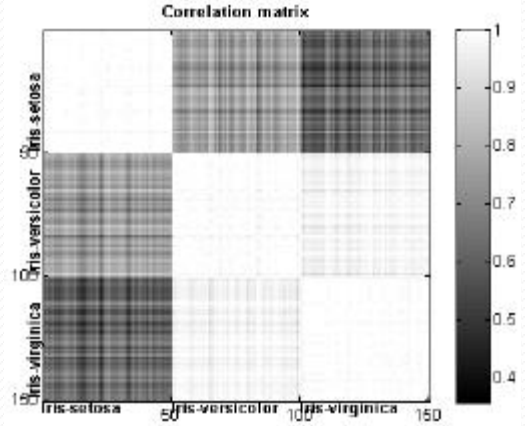
Matrix plots

Plot of raw data matrix

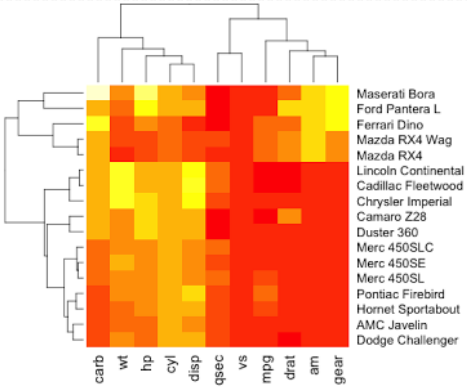
- Useful when objects are sorted according to class
- Typically, attributes are normalized

Plots of similarity matrices

- Useful for visualizing the relation between objects

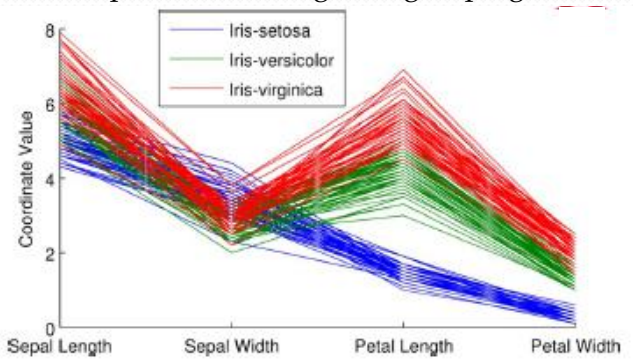


Heatmap (« mtcars » dataset)



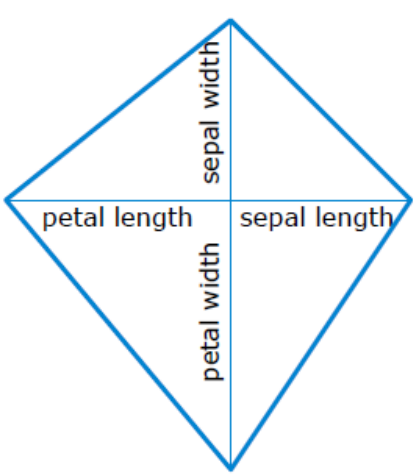
Parallel coordinates

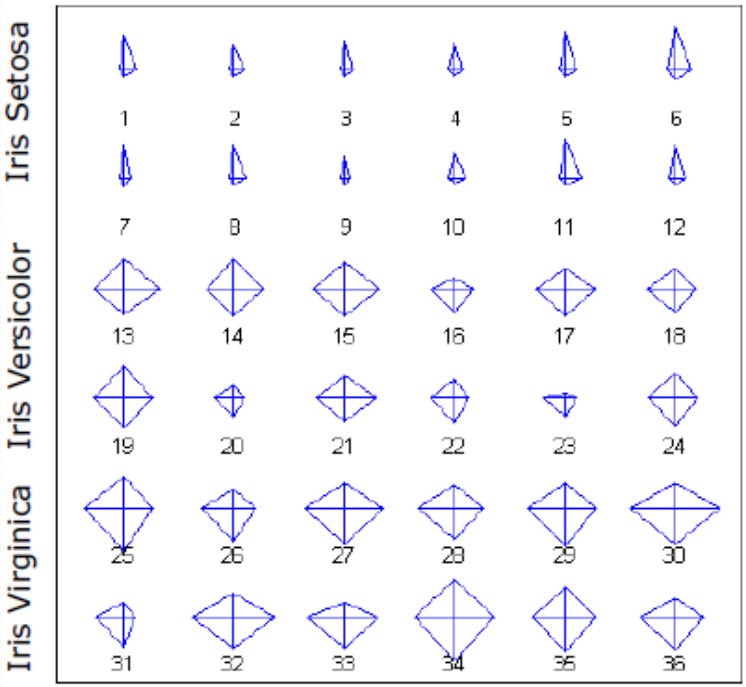
- Plot high-dimensional data
- Instead of perpendicular axes
  - Use parallel axes
- Attribute values are plotted as a point and the points are connected by a line
- Each object is represented as a line
- Lines representing a group of objects
  - Are similar in some sense
  - Ordering of attributes is important in seeing such groupings



Radar Charts

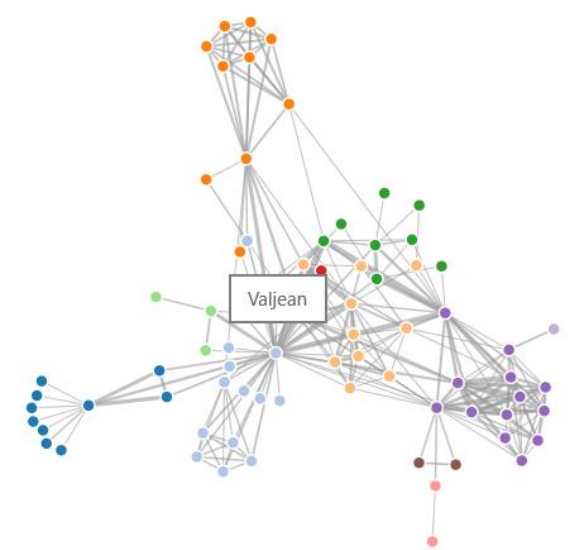
- Plot high-dimensional data
- Similar to parallel coordinates
  - Axes radiate from center
  - Connecting line is a polygon





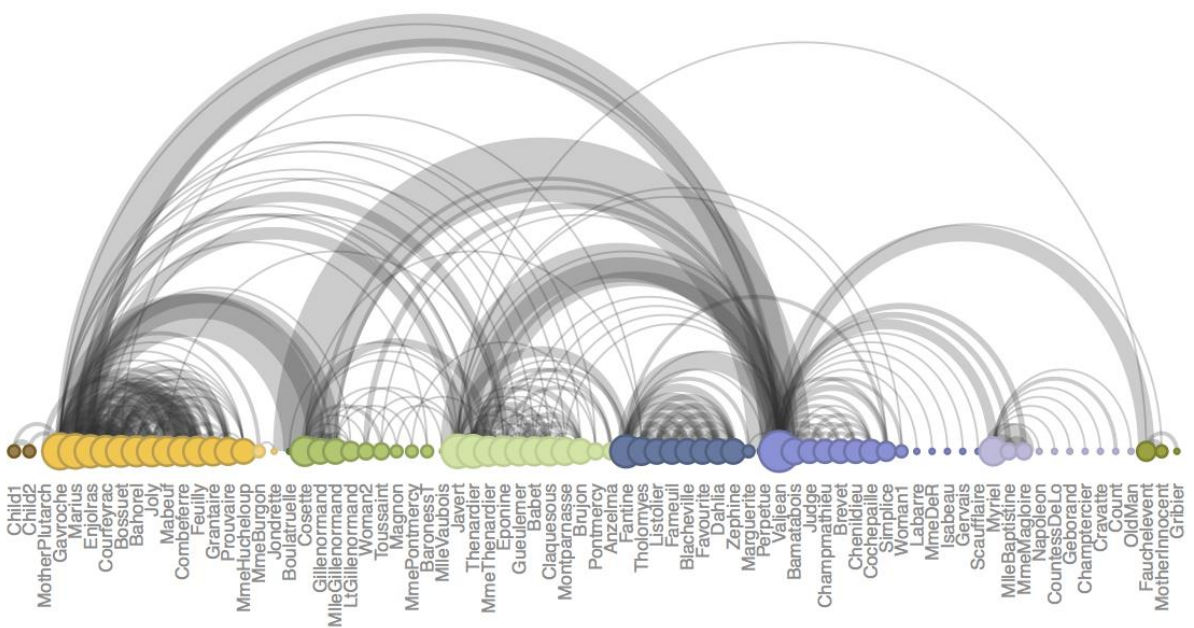
Force-Directed Graph

- Their purpose is to position the nodes of a graph in two-dimensional or three-dimensional space so that
  - all the edges are of more or less equal length and there are as few crossing edges as possible,
  - by assigning forces among the set of edges and the set of nodes, based on their relative positions,
  - and then using these forces either to simulate the motion of the edges and nodes or to minimize their energy



Data based on character coappearence in Victor Hugo's Les Misérables, compiled by Donald Knuth (places related characters in closer proximity, while unrelated characters are farther apart).

Arc Diagram


Child1
Child2
MotherPlutarch
Gavroche
Marius
Enjolras
Courfeyrac
Bossuet
Bahorel
Joly
Mabeuf
Combeferre
Feuilly
Grantaire
Prouvaire
MmeHucheloup
MmeBurgon
Jondrette
Boulatruelle
Cosette
Gillenormand
MlleGillenormand
LtGillenormand
Woman2
Toussaint
Magnon
MmePontmercy
BaronessT
MlleVaubois
Javert
Thenardier
MmeThenardier
Eponine
Gueulemer
Babet
Claquesous
Montparnasse
Brujon
Pontmercy
Anzelma
Fantine
Tholomyes
Listolier
Fameuil
Blacheville
Favourite
Dahlia
Zephine
Marguerite
Perpetue
Valjean
Bamatabois
Judge
Champmathieu
Brevet
Chenildieu
Cochepaille
Simplice
Woman1
Labarre
MmeDeR
Isabeau
Gervais
Scaufflaire
Myriel
MlleBaptistine
MmeMagloire
Napoleon
CountessDeLo
Geborand
Champtercier
Cravatte
Count
OldMan
Fauchelevent
MotherInnocent
Gribier

6 Conclusion

Key Points: ACCENT

- **A**pprehension
  - Is it easy to see what is important in the graph?
- **C**larity
  - Are the most important elements visually most prominent?
- **C**onsistency
  - Have you used the same colors, shapes, etc. as in other graphs?
- **E**fficiency
  - Does it convey its information in the most simple and efficient way?
- **N**ecessity
  - Are all elements of the graph necessary to represent data?
- **T**ruthfulness
  - Does the graph represent the data correctly?

Conclusion

- Visual summaries of the data
- A picture is worth a thousand words
- Complex ideas communicated with clarity, precision, and efficiency
- Compare the graphs using ACCENT
- Making good data visualizations is an art
- This is the first step to
  - Make efficient and pretty dashboard
  - Prepare relevant and pleasant storytelling.