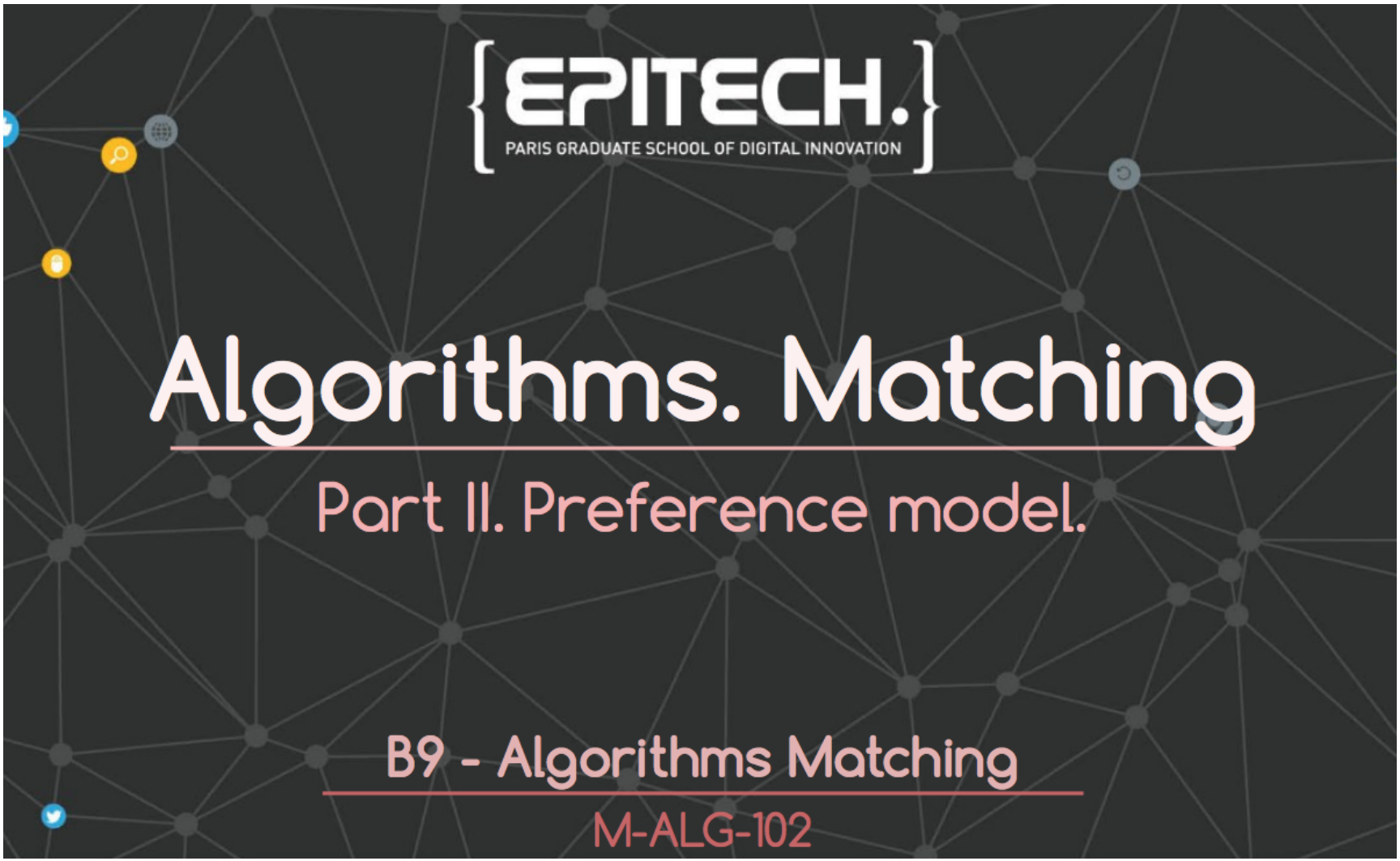

EPITECH.
PARIS GRADUATE SCHOOL OF DIGITAL INNOVATION
Algorithms. Matching
Part II. Preference model.
B9 - Algorithms Matching
M-ALG-102

Overview of day 2

- Classic visualization methods
  - Classification of data types
  - Classic methods
- Intermediate methods
  - Parallel coordinate plot
  - Miscellaneous
  - Hierarchies
- Visualization platforms
- Advanced methods
  - Principal component analysis
- Challenges

Classification of data

Let us summarize the different types of data that could be represented:

Classification of data

Let us summarize the different types of data that could be represented:

- univariate data (time series)

Classification of data

Let us summarize the different types of data that could be represented:

- univariate data (time series)
- two-dimensional data (geographical coordinates)

Classification of data

Let us summarize the different types of data that could be represented:

- univariate data (time series)
- two-dimensional data (geographical coordinates)
- multidimensional data (experiments, climatic data)

Classification of data

Let us summarize the different types of data that could be represented:

- univariate data (time series)
- two-dimensional data (geographical coordinates)
- multidimensional data (experiments, climatic data)
- text (newspaper, web pages)

Classification of data

Let us summarize the different types of data that could be represented:

- univariate data (time series)
- two-dimensional data (geographical coordinates)
- multidimensional data (experiments, climatic data)
- text (newspaper, web pages)
- hierarchical and relational data (web links, hierarchical relationships in an organization)

Classic visualization methods

We will now go trough a number of visualization methods for data and discuss their assets and drawbacks, and when they are relevant.

Pie chart

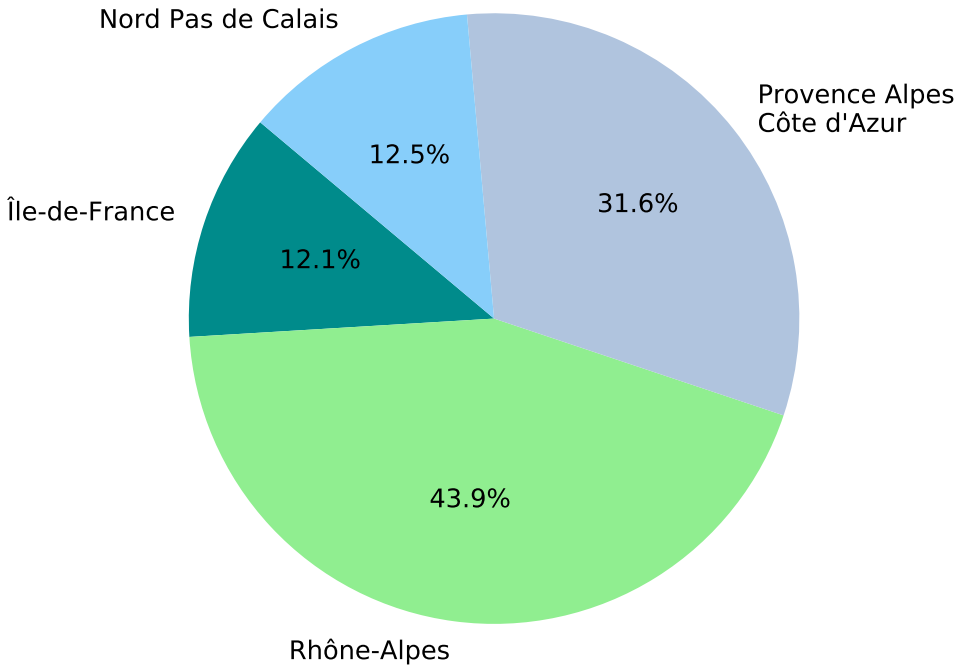


Figure: Areas of four region in France

Pie chart

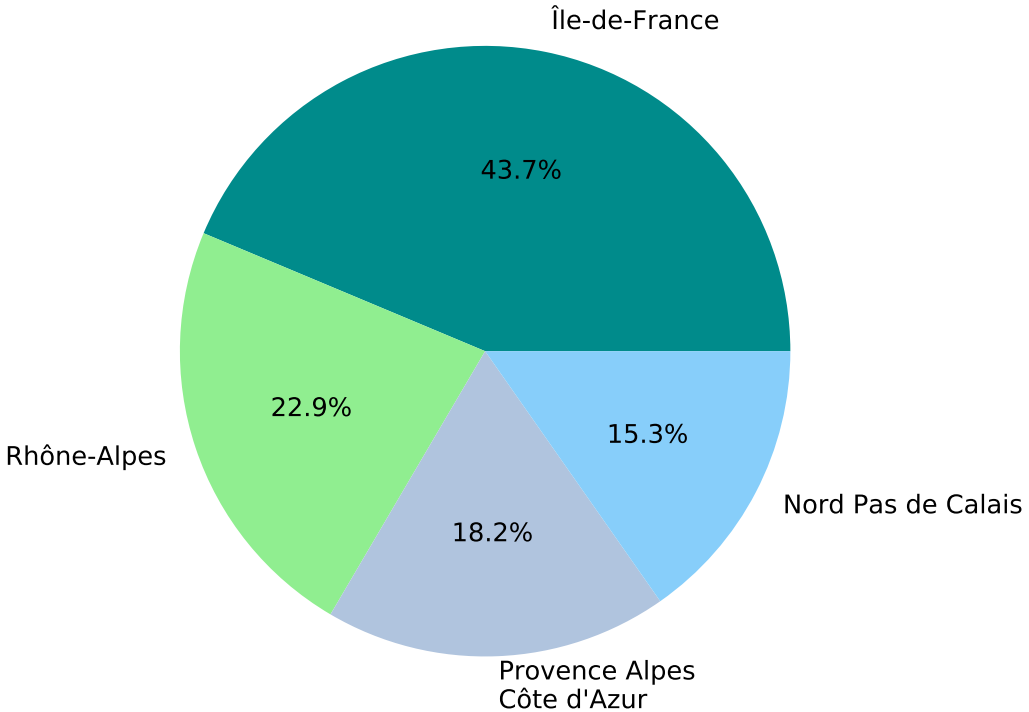


Figure: Population of four region in France

Pie chart

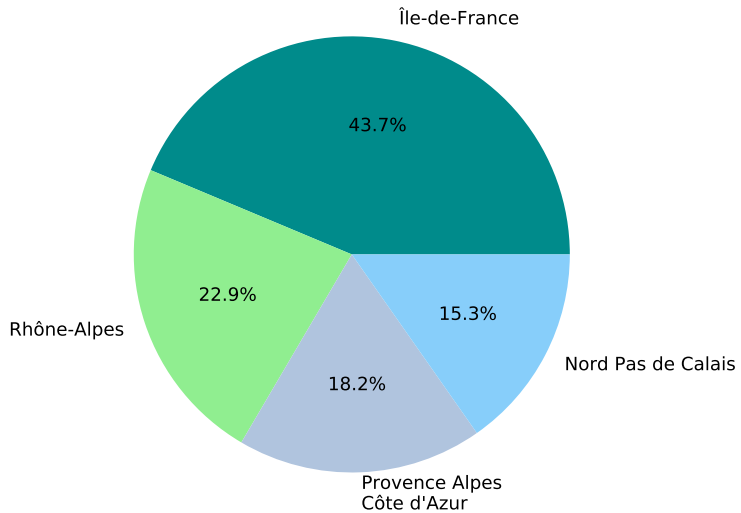


Figure: Population of four region in France

You have a pie chart template in **simple_graphs/pie_graph.py**

Pie chart

The pie chart can only represent relative sizes of sets.

Line graph (graphique linaire)

We studied the line graph yesterday.

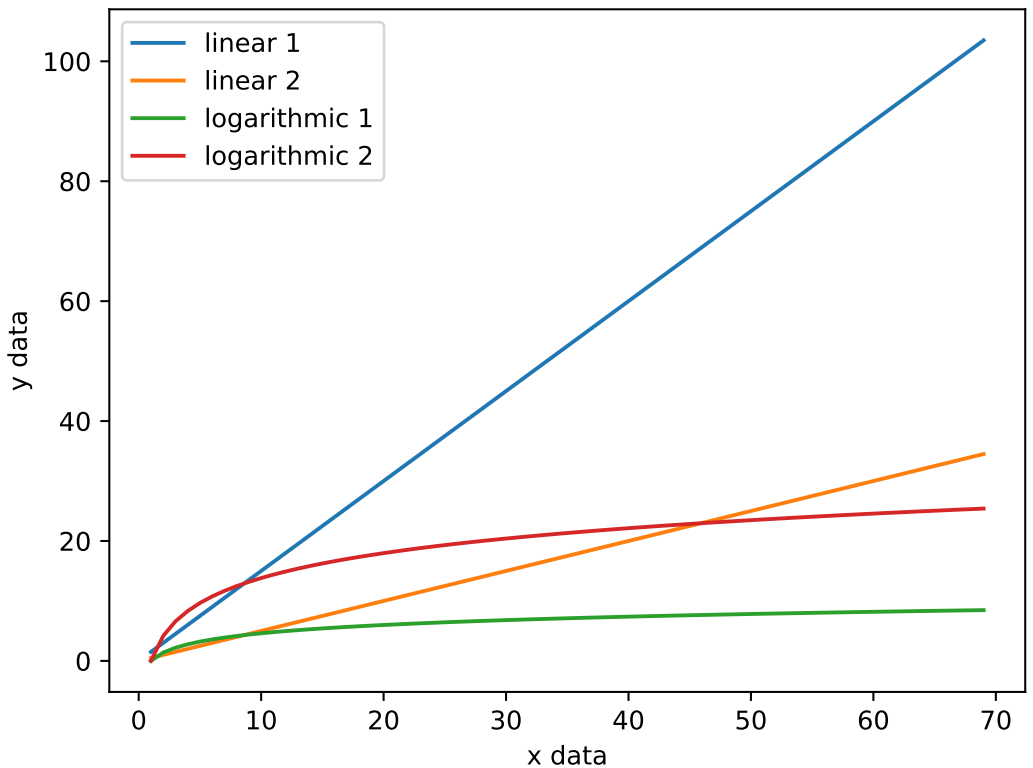
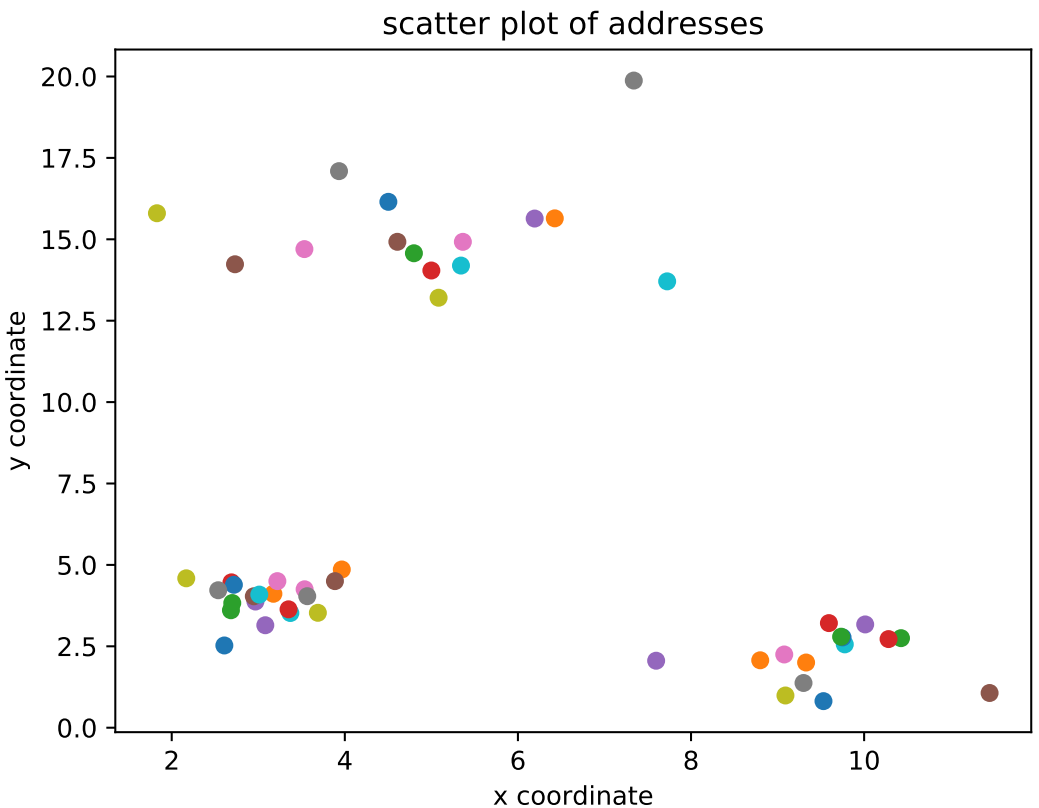


Figure: Logarithmic function

Scatter plot (nuage de points)

We also studied the scatter plot yesterday.



Scatter plot (nuage de points)

However, we need to discuss the case of a multivariate random variable.
For instance the classic iris dataset.

Scatter matrix

- Let us generalize the idea of a scatter plot.

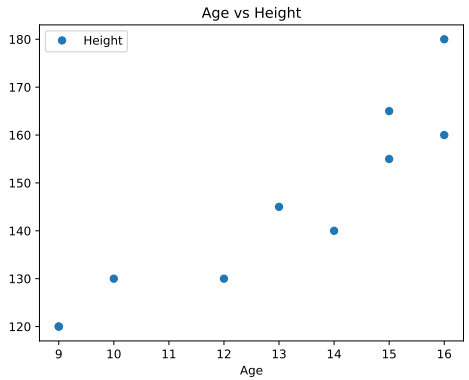


Figure: Exemple scatter plot

- Given a dataset, we can build all the possible scatter plots and store them in a matrix, called the **scatter matrix**.

Scatter matrix

- Let us generalize the idea of a scatter plot.

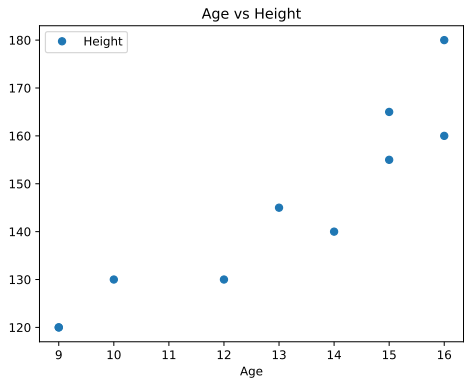


Figure: Exemple scatter plot

- Given a dataset, we can build all the possible scatter plots and store them in a matrix, called the **scatter matrix**.
- What will happen on the diagonal of the matrix ?

Scatter matrix

- Let us generalize the idea of a scatter plot.
- Given a dataset, we can build all the possible scatter plots and store them in a matrix, called the **scatter matrix**.
- What will happen on the diagonal of the matrix ?

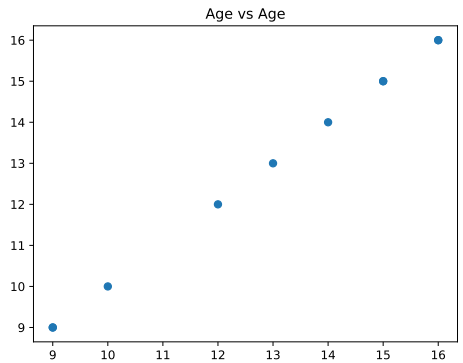


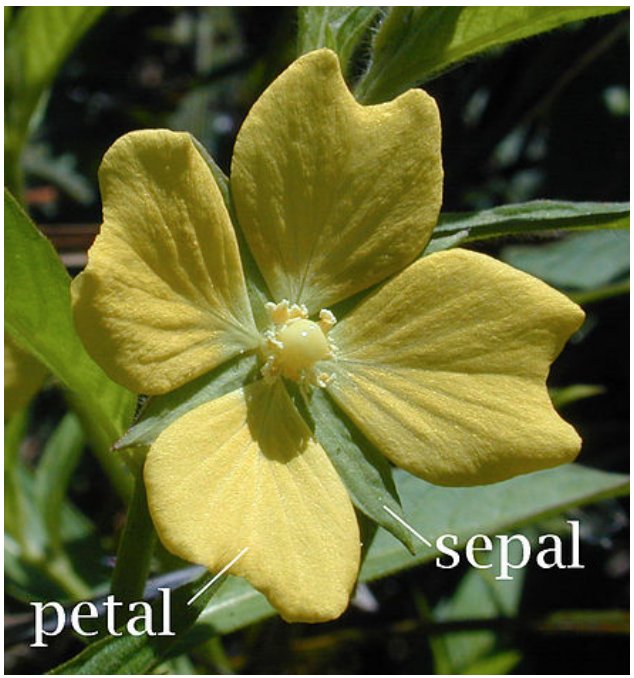
Figure: Variable plotted against itsself : All the points are on the $y = x$ line

Scatter matrix

- Let us generalize the idea of a scatter plot.
- Given a dataset, we can build all the possible scatter plots and store them in a matrix, called the **scatter matrix**.
- On the diagonal, one can plot histograms or the density probability
- The scatter plot can be a good way to start analyzing a dataset when we don't know which variables could be correlated

Iris dataset

- 150 samples of iris flower
- 3 species
- 4 attributes : petal width and length, sepal width and length



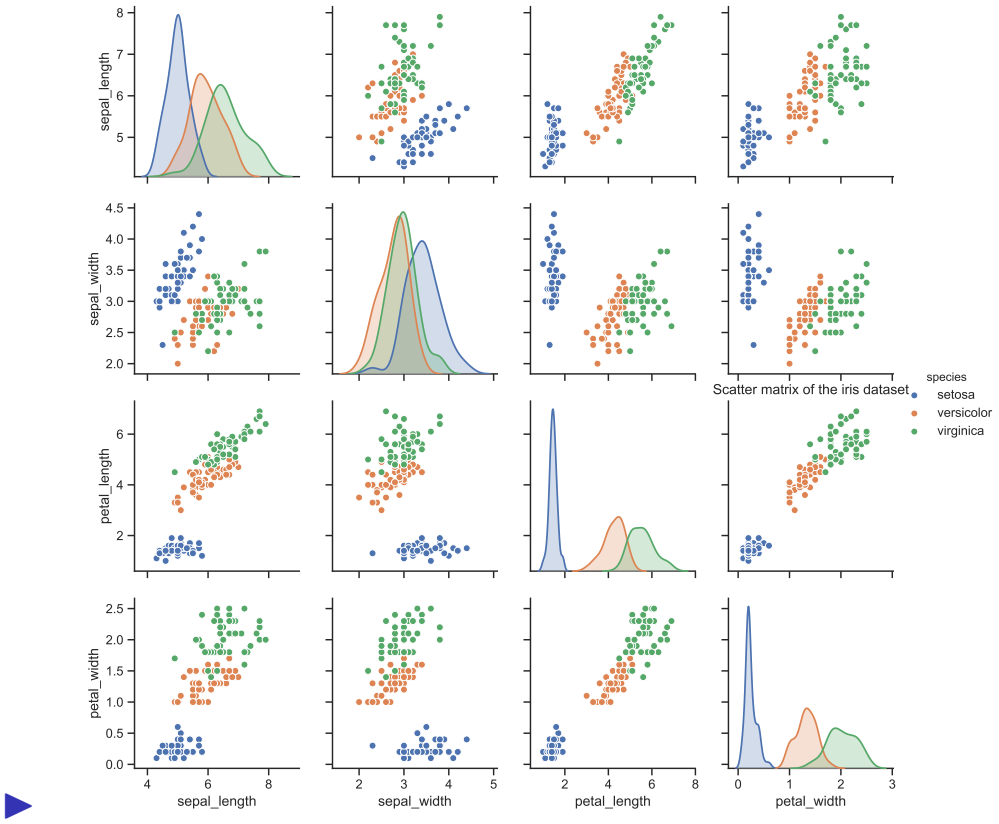


Figure: Plotted in **iris_scatter_matrix.py**. Is there a variable that can discriminate between the species ?

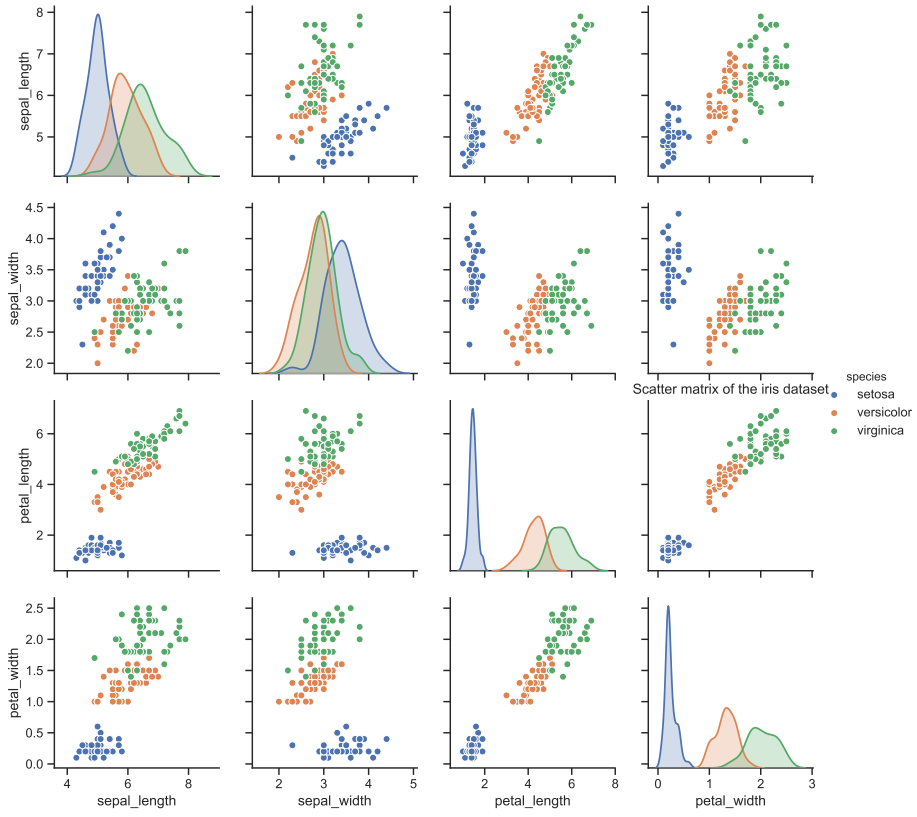


Figure: It seems that **petal width** is a parameter that separates the three species. On the contrary, **sepal width** is not able to do so.

Scatter plot

The scatter plot is considered to be relevant to represent around 10000 points at most.

Scatter plot

The scatter plot is considered to be relevant to represent around 10000 points at most.
If we have a large dataset, we must select a fraction of the points to represent.

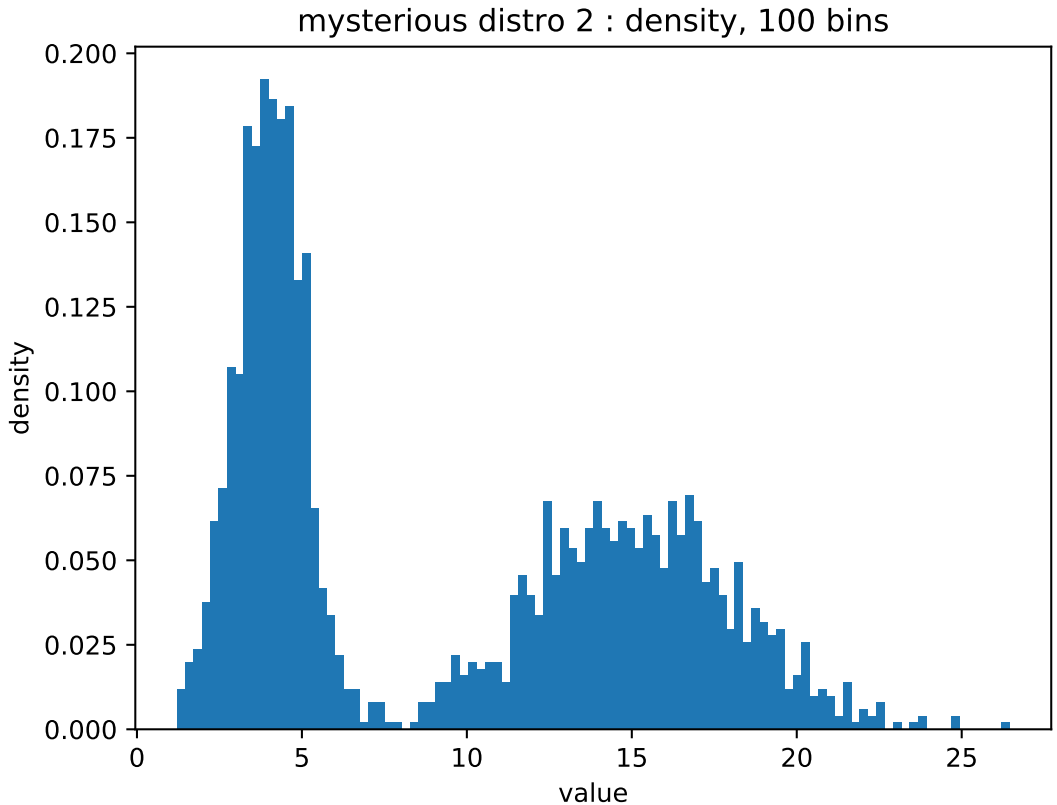
Scatter plot

The scatter plot is considered to be relevant to represent around 10000 points at most.
If we have a large dataset, we must select a fraction of the points to represent.

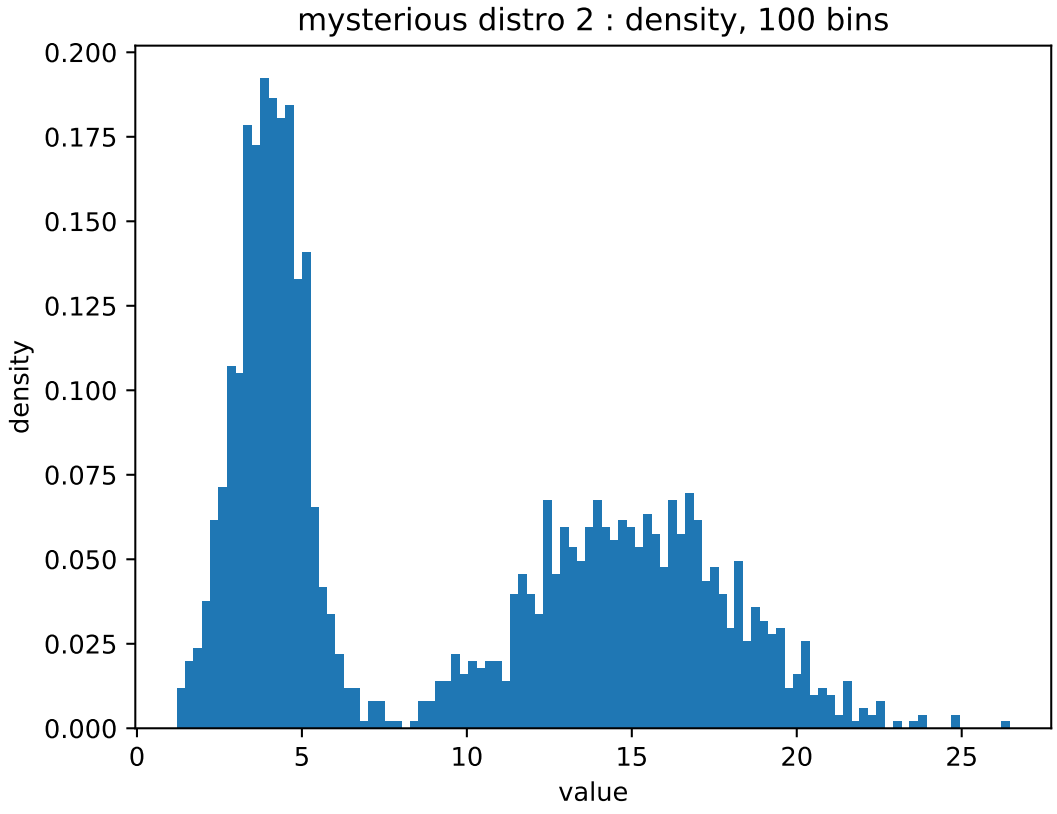
Exercice 1 : **Number of possible plots**

How does the number of possible scatter plot depend on the number of datapoints ? and on the number of dimensions ?

Histogram


mysterious distro 2 : density, 100 bins
density
value
0.200
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
0
5
10
15
20
25

Histogram



As in the case of the pie chart, this histogram can only represent **one dimension**.

Parallel coordinate plot

The parallel coordinate plot is a tool used to show multidimensional data.

Parallel coordinate plot

The parallel coordinate plot is a tool used to show multidimensional data.
Each dimension will correspond to one column.
Each datapoint is represented by a line between columns.

Parallel coordinate plot

Exercice 2 : **Showing the plot**

cd parallel_coordinate/ and use **parallel_coordinate.py** in order to to a parallel coordinate plot of the data from the file **data.npy**.

Parallel coordinate plot

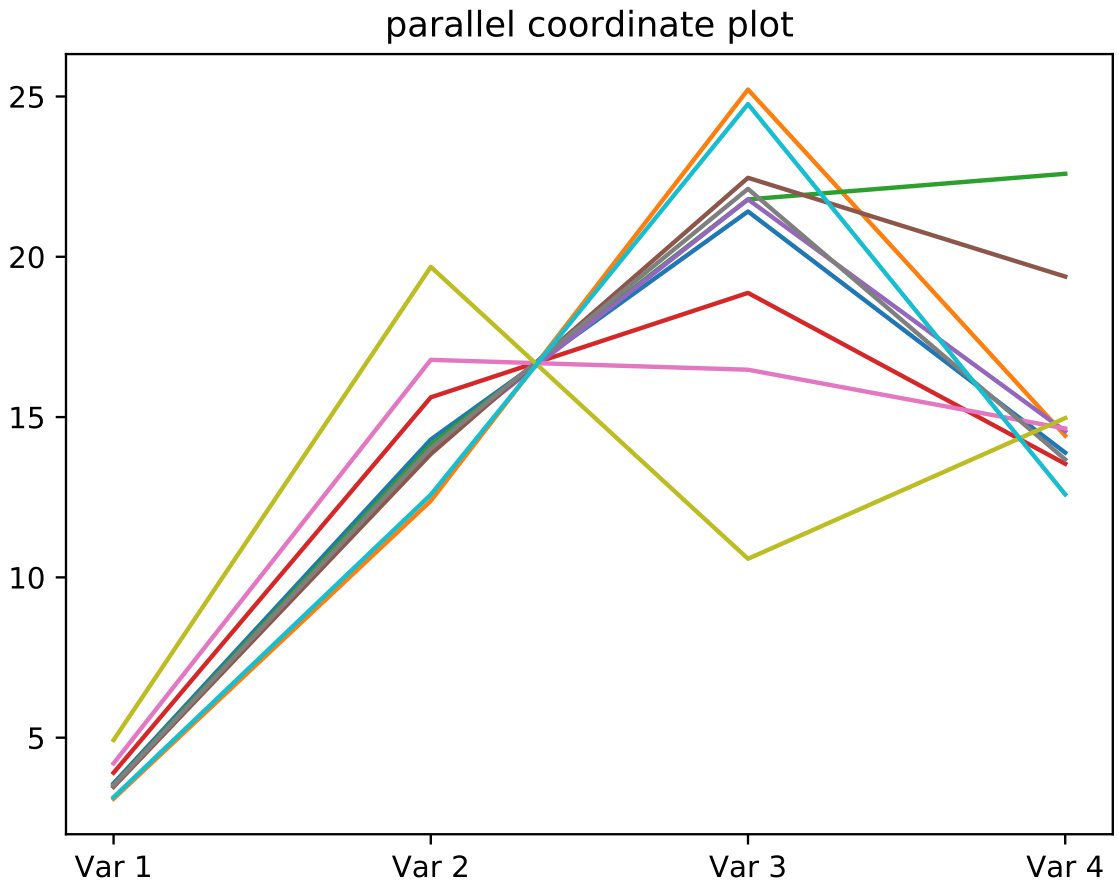


Figure: What are the relationships between the variables ?

Parallel coordinate plot

Assets:

- Can help identify patterns or correlations in the data.

Drawbacks:

- Can be overcluttered

Parallel coordinate plot

We will use plotly to do "brushing". (pip install plotly)
`https://plot.ly/`

Parallel coordinate plot with plotly

Use the file **parallel_coordinate_plotly.py** in order to show an interactive plot of the iris dataset.

Practical application

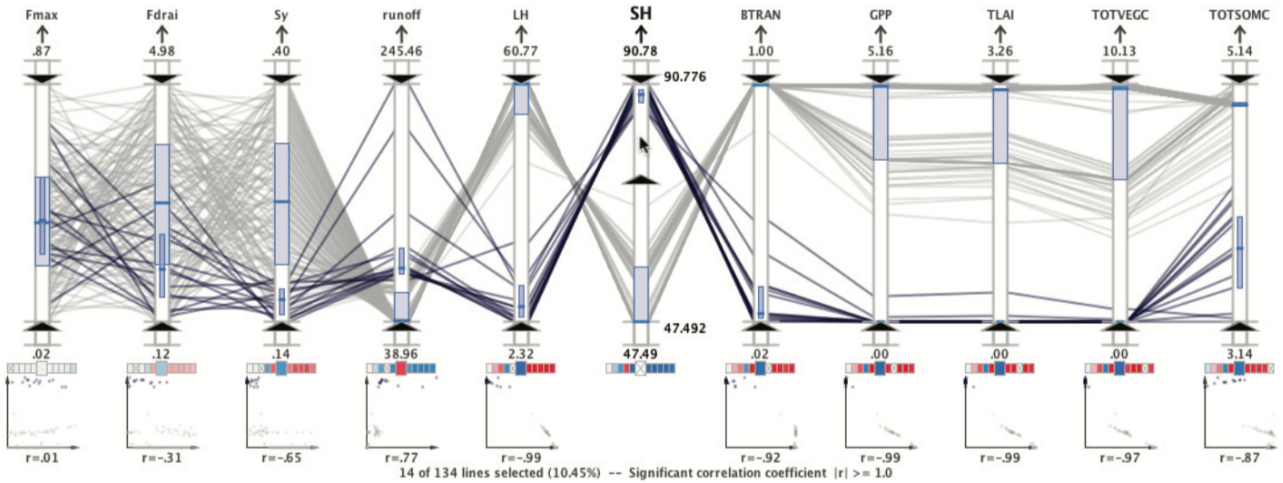


Figure: [Steed et al., 2014]

Radar plot

Figure: [Analytics,]

Heat map

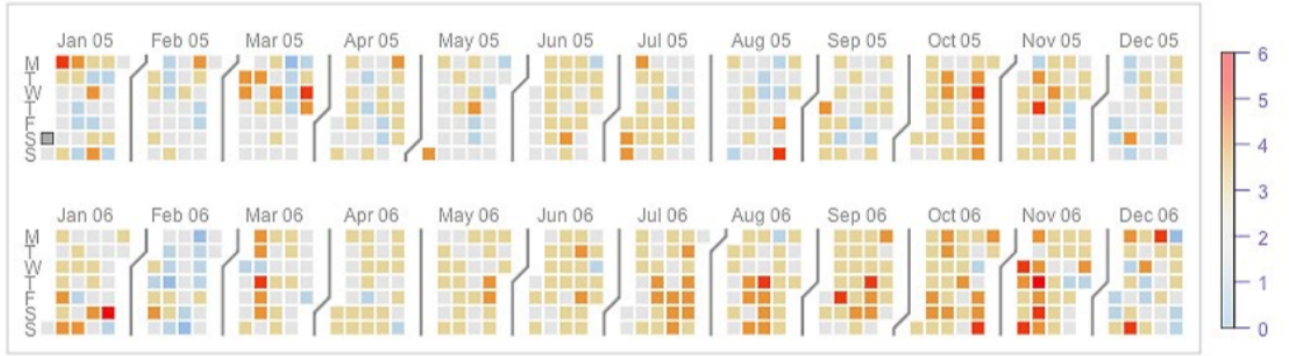


Figure 3: Calendar heat map example that shows two years of changes (in percentages) in customer web orders by year (row), month (column), day of week (sub row), week (sub column) and day.

Figure: [Analytics,]

Networks visualization

Graphs are useful in order to represent relationships between datapoints.

Networks

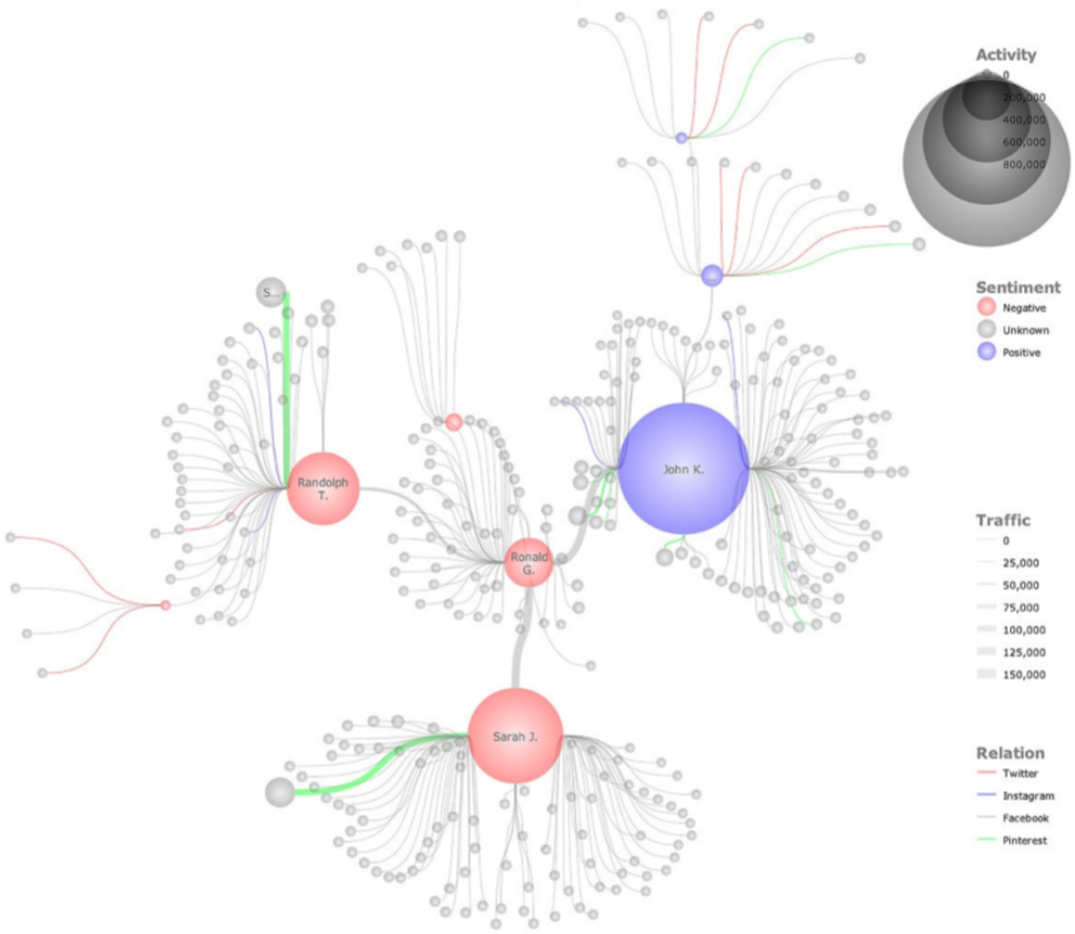


Figure: [Analytics,]

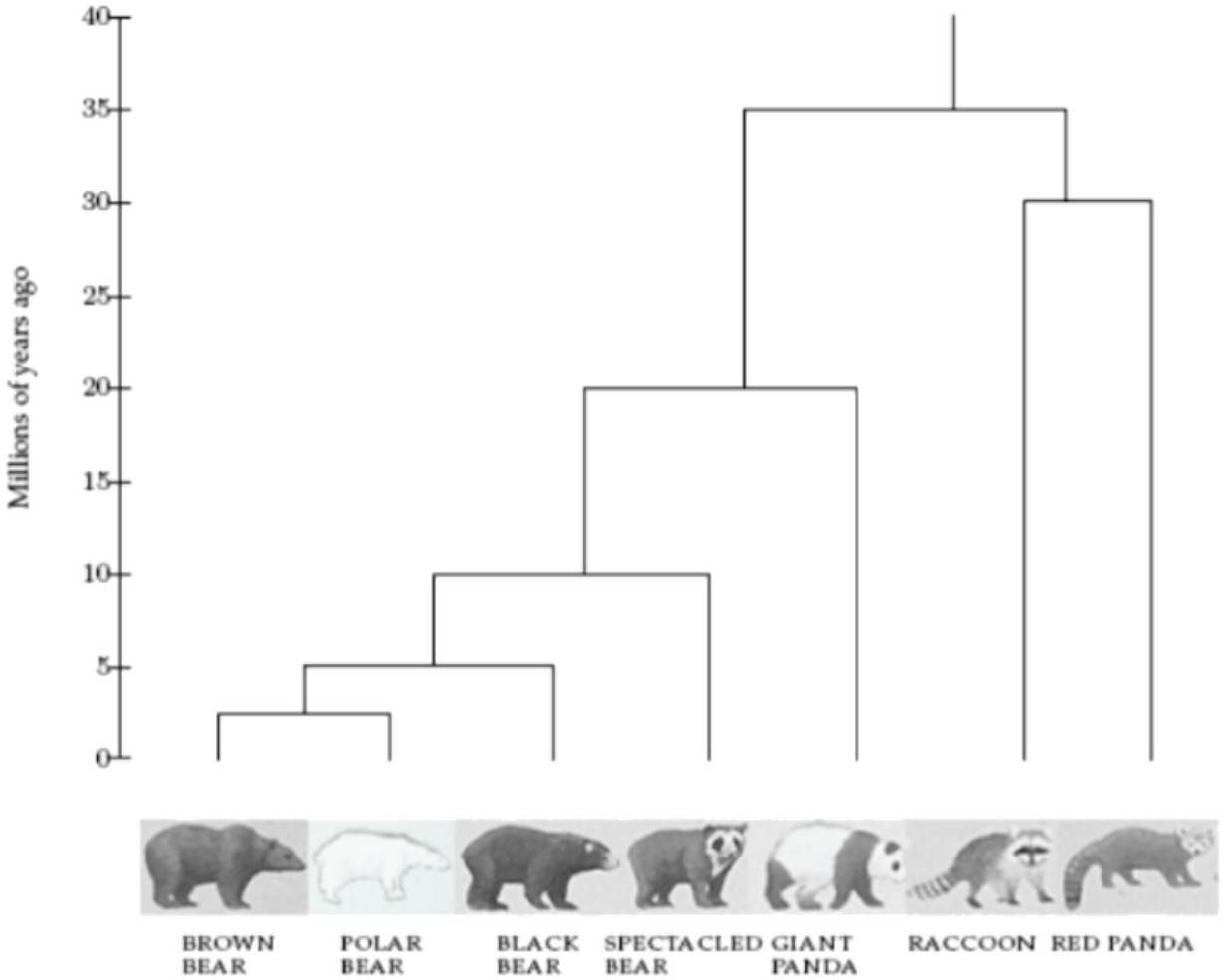
Hierarchies

The hierarchical clustering we performed yesterday can also be performed with libraries such as **sklearn** and **scipy**.
`https://scikit-learn.org/stable/auto_examples/cluster/plot_agglomerative_dendrogram.html`

Hierarchies

https://docs.scipy.org/doc/scipy/reference/cluster.hierarchy.html

Example application of hierarchical clustering



Treemaps

A **Treemap** is a another representation of hierarchical data in the two-dimensional space.

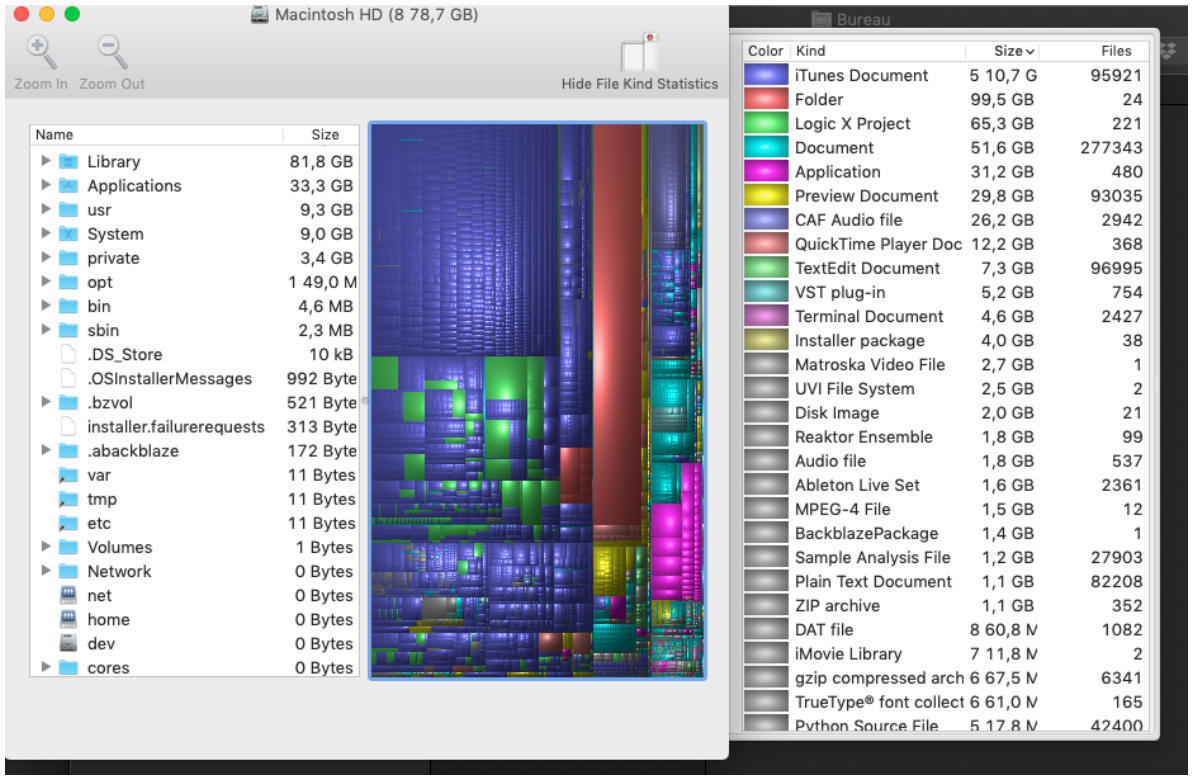


Figure: Disk Inventory X `http://www.derlien.com/`

Treemaps

A **Treemap** is a another representation of hierarchical data in the two-dimensional space.

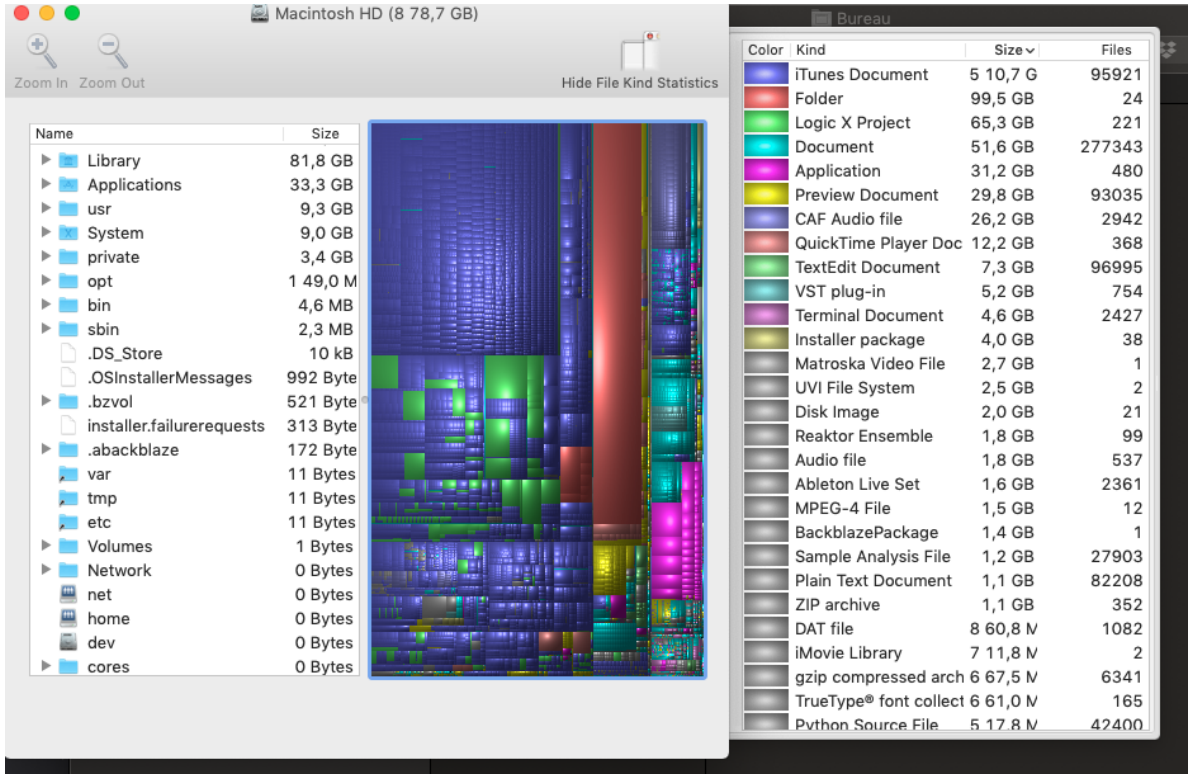


Figure: Disk Inventory X `http://www.derlien.com/`

Building a tree map

cd treemap/ and use **build_treemap.py** in order to build a treemap of your Desktop.

Building a tree map

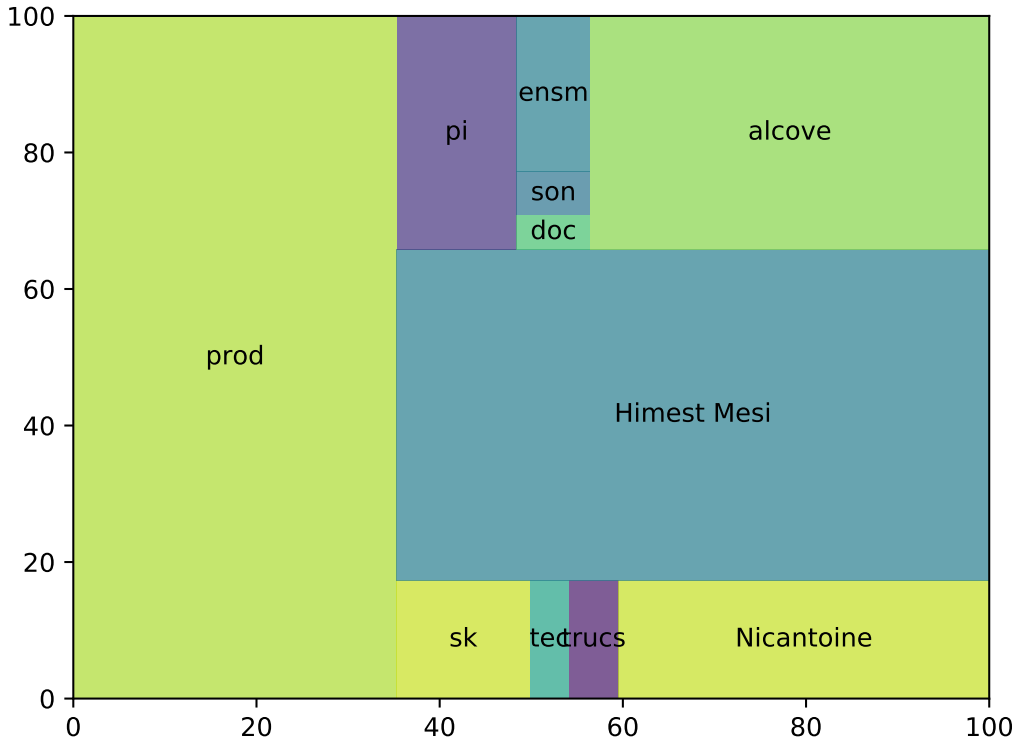


Figure: Treemap of desktop computer (desktop folder)

Building a tree map

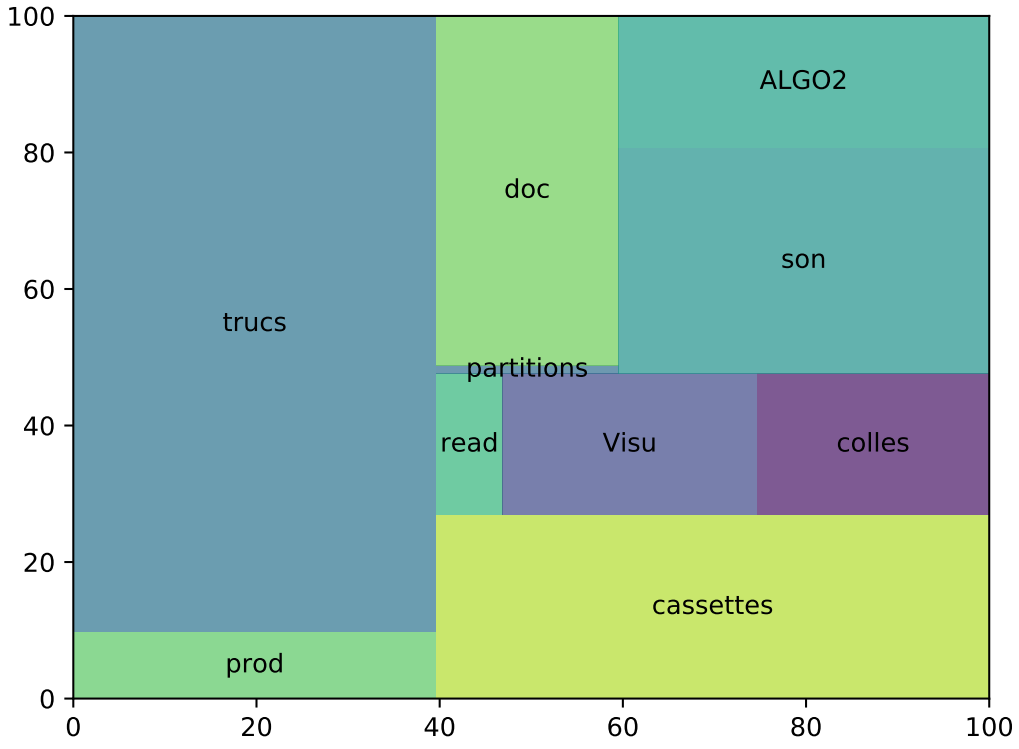
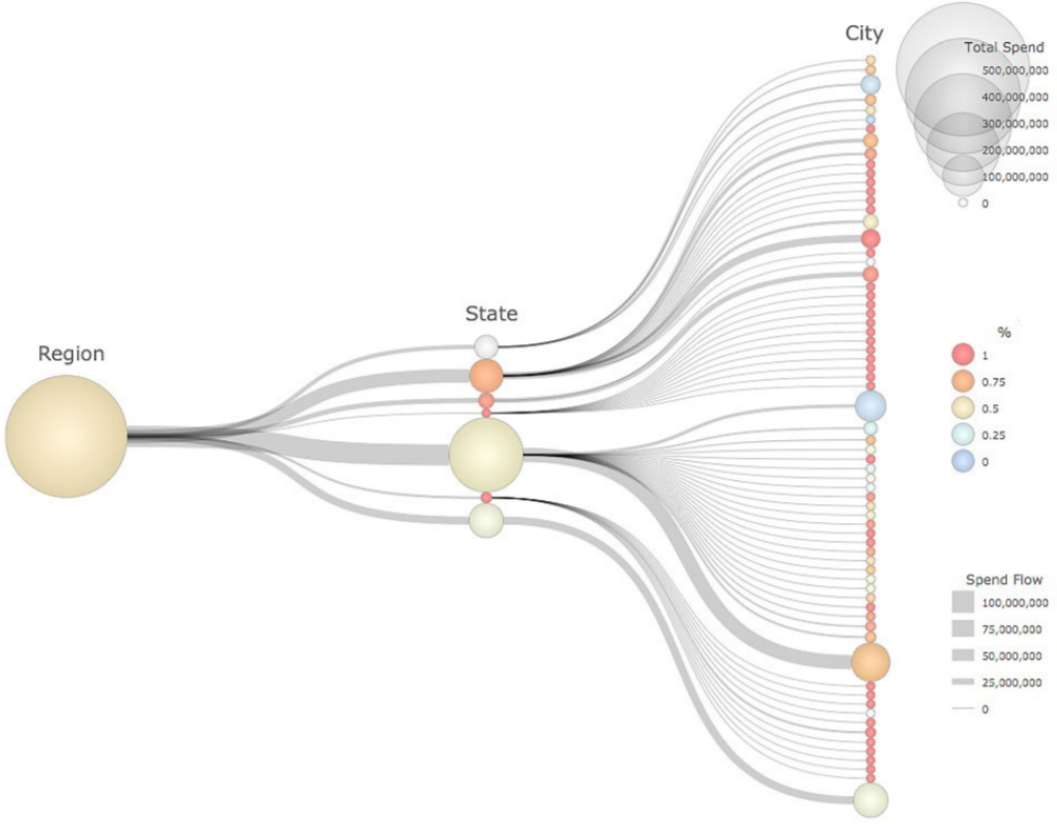


Figure: Treemap of laptop (desktop folder)

Hierarchy


City
Total Spend
500,000,000
400,000,000
300,000,000
200,000,000
100,000,000
0
State
Region
%
1
0.75
0.5
0.25
0
Spend Flow
100,000,000
75,000,000
50,000,000
25,000,000
0

Remak on visualization

The Gestalt Principles give guidelines for a good visualization.

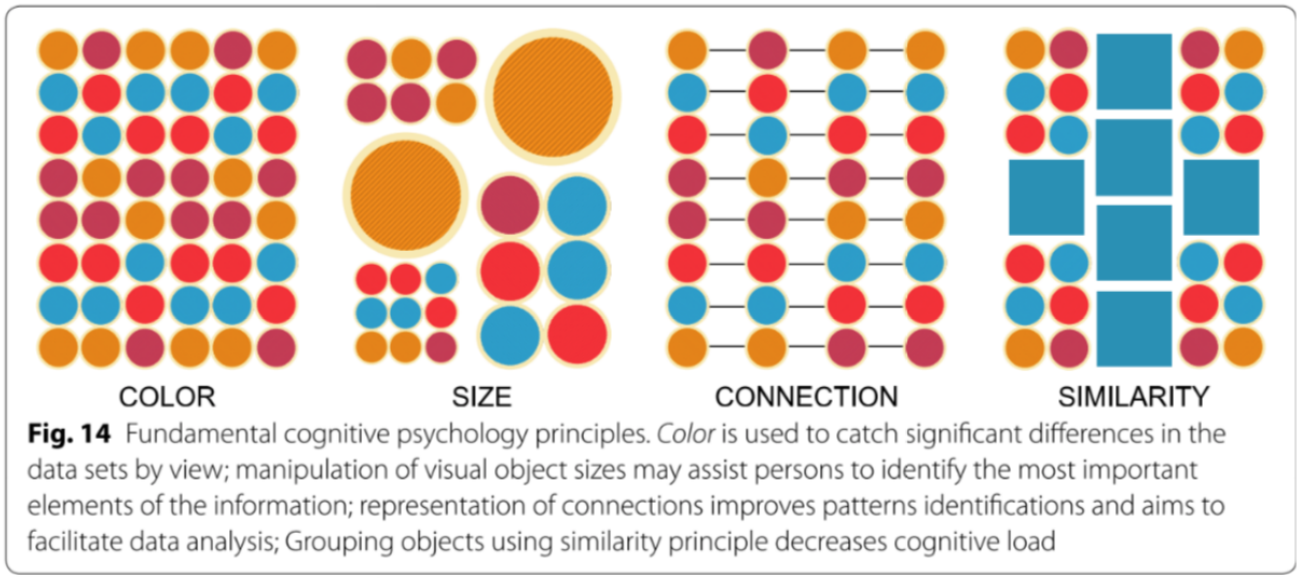


Fig. 14 Fundamental cognitive psychology principles. *Color* is used to catch significant differences in the data sets by view; manipulation of visual object sizes may assist persons to identify the most important elements of the information; representation of connections improves patterns identifications and aims to facilitate data analysis; Grouping objects using similarity principle decreases cognitive load

Figure: [Olshannikova et al., 2016]

Visalization platforms

Here are some useful links for data visualization and processing.

Links

https://fcpython.com/

Links

https://www.dataiku.com/

Links

https://www.tylertech.com/products/socrata

Links

https://dev.socrata.com/

Links

https://cytoscape.org/

Links

https://www.tableau.com/

Links

`http://vis.stanford.edu/wrangler/`

Links

https://noduslabs.com/infranodus/

Advanced methods

- Kohonen maps (carte auto adaptive, self organizing map)
- t SNE [Van Der Maaten and Hinton, 2008]
- clustergram
- Manifold learning and nonlinear dimensionality reduction

Challenges

Some people argue that augmented reality or virtual reality may help us to gain more insight from data visualization [Olshannikova et al., 2016].

Project description

Description of the project.

References


Analytics, B.
Analytics : The real-world use of big data.


Olshannikova, E., Ometov, A., Koucheryavy, Y., and Olsson, T. (2016).
Chapter 4 VISUALIZING BIG DATA.
Number October.


Steed, C. A., Shipman, G., Thornton, P., Ricciuto, D., Erickson, D., and Branstetter, M. (2014).
Practical Application of Parallel Coordinates for Climate Model Analysis.
Procedia - Procedia Computer Science, 9(December 2012):877–886.


Van Der Maaten, L. J. P. and Hinton, G. E. (2008).
Visualizing high-dimensional data using t-sne.