Master ICT - USTH
BC2.04b - Data Visualization – Week2
2024-2025

Clustering of numerical data with Tableau, and of categorical data with Formal Concept Analysis

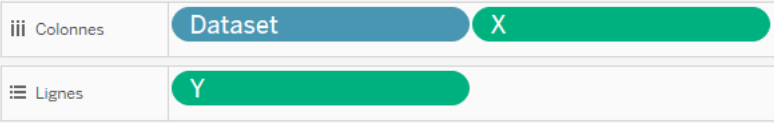
Objective : The objective of this practical work is to learn the basis of clustering for numerical data with Tableau, and the basis of formal concept analysis for categorical data

1. Clustering of numerical data with Tableau

Exercice 1. First we will use the *Datasauraus* dataset which contains several subsets. The file is composed of three columns:

- A dataset attribute that indicates the subdataset
- Two attributes x and y which indicate coordinates

Open the dataset with Tableau, view X in columns and Y in rows, add *Dataset* on the columns to get an overview of the 13 subdatasets.



Colonnes: Dataset, X

Lignes: Y

Exercice 2. Remove Dataset from the columns and use it as a filter to view only one dataset (the one of your choice. Show the summary (“Worksheet” / “Show summary”), look at the average and the 'standard deviation (find out how to display it if it's not there) then do the same thing again choosing another sub-dataset in the filters.

Exercice 3. Before clustering we often want to normalize the data, i.e. bring all the values between 0 and 1 or between -1 and 1, or set the mean to 0 and the variance to 1 (we then speak rather of standardization) while keeping all other data properties, including relative distances. To obtain an average of 0, simply subtract the average from each of the values. Intuitively, we want to create a new measure of the type:

```
 [X]-avg([X])
```

Create this measure. What appends ?

Level Of Detail

With Tableau, a calculated field corresponds to a calculation that is done independently for each row by combining elements of this row. For example, we will multiply a unit price by a quantity to obtain a total price on each line. We can also do a global calculation, for example SUM(X). Here we want to calculate a new value for each line but which also depends on other lines (the « average » over the whole dataset). Tableau states that it is not possible to mix aggregates and non-aggregates when creating a calculated field.

A simple solution is to precalculate these values with Excel or equivalent software and only use Tableau for the analysis/visualization part. However, it can be a bit tedious to constantly switch from one software to another. Tableau, since version 9, allows you to manage levels of detail (Level Of Detail or LOD):

https://help.tableau.com/current/pro/desktop/fr-fr/calculations_calculatedfields_lod_overview.htm

The concept is not so easy to understand and Tableau has explain everything, we will limit ourselves to the most basic use here, the use of a table level LOD

Exercice 4. On a new sheet, put *Dataset* in rows, put X on the text mark and replace SUM(X) by MIN(X). You must obtain a table with 13 rows and the minimum value for each data set (for example for away the minimum is 15.56).
Now create a calculated field "minx" containing the code below :

```
{FIXED [Dataset]: MIN([X])}
```

This formula allows us to say that we will calculate a fixed value, here the minimum of X, for each distinct value of Dataset. Add this new calculated field to the text marker and check that you have the same result as for the previous exercise. Remove [Dataset] in the formula, what changes and why?

Exercice 5. Using the previous principle, create a field *xnorm* containing X minus the average of X for each dataset, in the spirit of "[X]-avg([X])" but which works… Do the same for Y. Create a display with *x* and *xnorm* in columns, *y* and *ynorm* in rows filtering on circle. Check that everything is good and that the standardization went well.

Clustering

The general principle of clustering is to group the data by making the close data be in the same group and the distant data be in different groups. While there are a wide variety of methods for clustering data, Tableau only uses one method, which is simple and not always very effective, the k-means method:

- https://www.naftaliharris.com/blog/visualizing-k-means-clustering/

Tableau is not a tool for clustering, and if you want to use other methods you can use external software, calculate the clusters and then import the data into Tableau to visualize it. Here we will therefore limit ourselves to simple experiments.

To calculate groups, once a visualization has been set up, choose "cluster" in the analysis area and drag it onto the visualization. K-means needs two informations to calculate the groups:

- The attributes (coordinates) of the points to be clustered. On the wikipedia page there is an illustration of the algorithm in 2 dimensions but the algorithm works exactly the same way in 3, 4... dimensions, as soon as we know how to calculate the distances between the points .
- The number of groups. If the number of groups is not specified, Tableau is able to "guess" the optimal number of groups. If we have an idea of the number of groups we can indicate it.
- Tableau uses Euclidean distance (it draws a straight line and measures it) to know which points are near or far.

In each of the exercises below we will use a dataset. Use the "Dataset" field in the filter area to see only this one.

Normalisation – Dataset away

Tableau also normalizes values between 0 and 1 (min-max normalization). So if the values are in the interval [x,y], Tableau will normalize them in the interval [0,1]. A value x will become a value 0, a y will become a value 1 and the intermediate points will come between 0 and 1 proportionally.

Exercice 6. Create calculated fields xnorm and ynorm using the formulas in the previous question (min and max are aggregates). To verify that you have correctly created the fields, place X and Xnorm in columns, Y and Ynorm in rows and you should obtain the following figure on which we see that the clouds are identical but that those at the bottom left are well normalized between 0 and 1 with minimum and maximum values realigned to 0 and 1 on each axis:

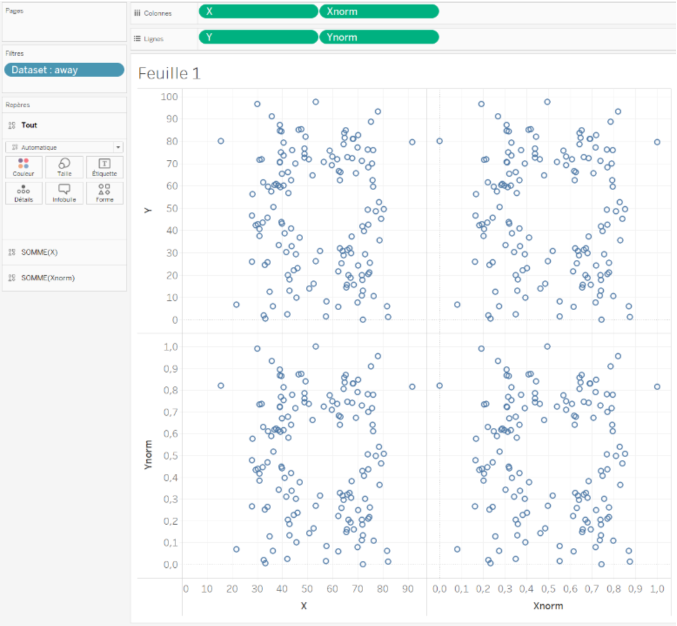


Tableau normalizes by default the data between 0 and 1 before doing the clustering so it is not useful to do it but congratulations for exercises 6 and 7 ;).

Dataset high_lines

This dataset contains two fairly well-separated groups that are easy to cluster.

Exercice 7. Run clustering on x and y with all default values. Tableau should normally find two groups. Drag and drop clusters onto dimensions for later and rename the new dimension so you can find it later. You can watch the video showing the creation of groups.

Exercice 8. Create a calculated field “yplus10” containing the formula below. What does this formula do? What is the meaning of 60? of 10 ?

```
IIF ([Y]<60, [Y]+10, [Y])
```

Restart clustering using x and yplus10. Does this change anything? You can use your original groups in the "shape" marker to see more clearly. Try with 20, 30, 40.

Dataset wide-lines of binary data

We are now going to study binary data (which can only take 2 values) and show that the clustering of this type of data with Tableau can be problematic. We will build a dataset to show this starting from wide_lines.

Exercice 9. Start by clustering wide_lines. Visually this dataset looks like high_lines but the clustering does not happen in the same way… To confirm this, compare the p-values of a clustering in two groups for wide_lines and for high_lines.

Exercice 10. Create a calculated field “xbinary” containing the formula below. This formula will create a new calculated field which will be 0 if X is less than 50 and 1 otherwise. Once the field is created, visualize to verify that everything is good.

```
IIF ([X]<50, 0, 1)
```

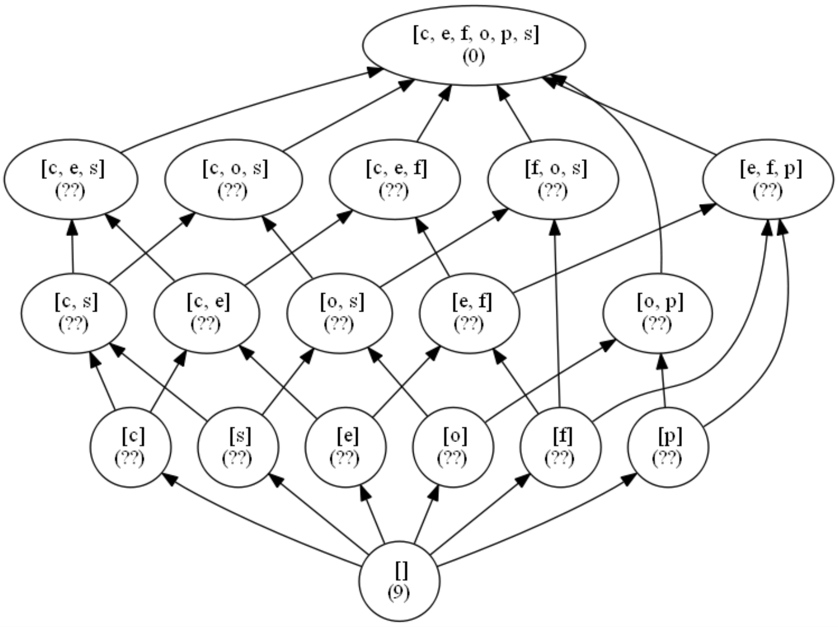
Exercice 11. Cluster the data into two groups using “xbinary” and “y”. Does the result seem satisfactory to you (go to describe the clusters)?
We must see here that with two groups, only the xbinary variable impacts the result. More generally, binary variables can have a strong impact on the results because the points of values 0 or 1 are very far from each other (due to min-max normalization) especially if the number of groups is low...

2. Hierarchy clustering of categorical data with Formal Concept Analysis

Exercice 12. Consider the data set *digits* which is a small binary dataset where numbers are described by the attributes c (compound number), e (even number), o (odd number), p (prime number), s (square number) and f (factorial number) and its concept lattice

- Complete the objects part of each concept
- Compute α(β([f,p]) to obtain the concept containing [f,p]
- Compute α(β([e,f]) to obtain the concept containing [e,f]
- Compute the immediate successors of the concept for [e,f] using the theorem of Bordat:
  The immediate successors of a closed set B are the inclusion minimal set of the family {α(β(x+F)) : x∉ F}

	c	e	o	p	f	s
1			X		X	X
2		X		X	X	
3			X	X		
4	X	X				X
5			X	X		
6	X	X			X	
7			X	X		
8	X	X				
9	X		X			X


[c, e, f, o, p, s]
(0)
[c, e, s]
(??)
[c, o, s]
(??)
[c, e, f]
(??)
[f, o, s]
(??)
[e, f, p]
(??)
[c, s]
(??)
[c, e]
(??)
[o, s]
(??)
[e, f]
(??)
[o, p]
(??)
[c]
(??)
[s]
(??)
[e]
(??)
[o]
(??)
[f]
(??)
[p]
(??)
[]
(9)