Group 6 Clustering Analysis

Chaland Pauline, Gangloff Romain, Maige Jason, Pandya Vivek

Data Science and Organizational Behaviour

Burgundy School of Business, Dijon

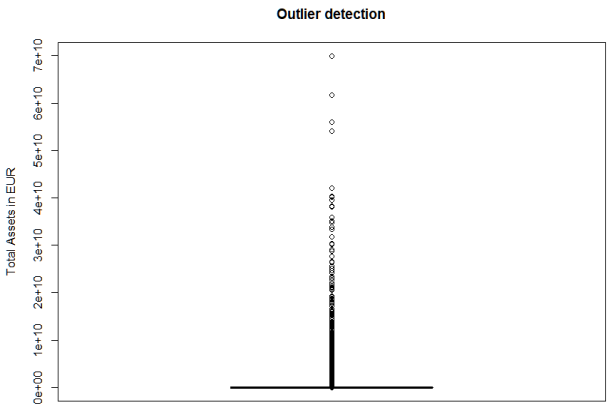
18 April 2023

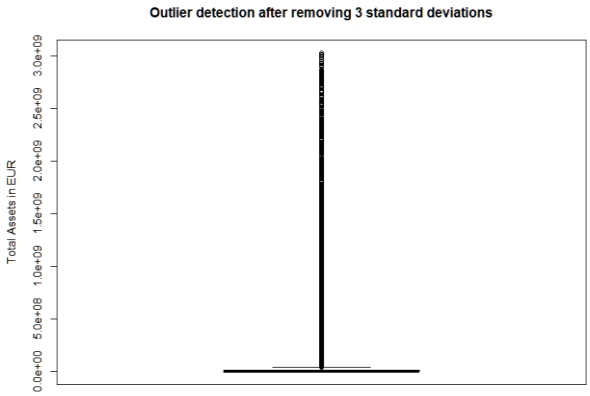
Swayam Sarkar

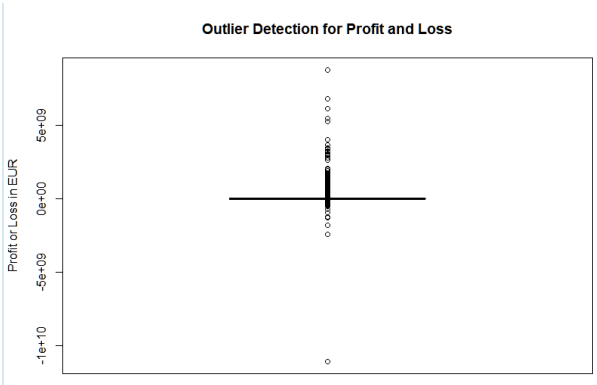
Contents

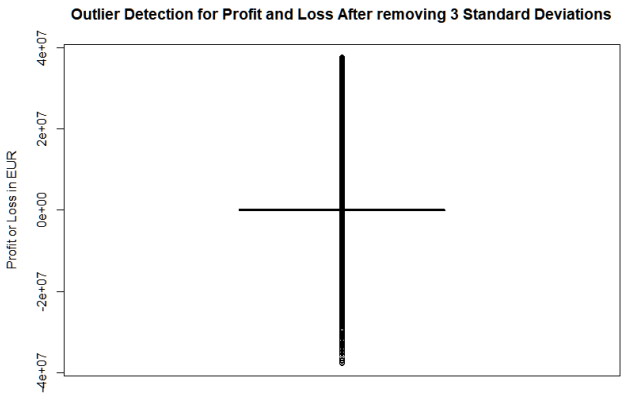
1.1 Data Cleaning, Standardization and Normalization

At the initial look at the data, it was identified that certain currencies were not converted to Euros. It was communicated with Swayam Sarkar on an email exchange, and we were provided with an updated data set. First, data was cleansed to remove the null values/NAs from the Asset and ProfitLoss columns. The next step involved conducting basic statistical tests to identify the extreme values. We plotted basic box plots to identify the outliers to have a cleaner data set. We tried to remove them initially by eliminating everything beyond the interquartile range. However, it still contained outliers. Hence, we used the method of eliminating everything beyond three standard deviations of the data. This reduced the number of outliers and provided us with cleaner data. The exact process was then applied to the Profit Loss column. After this operation, the reduced data set had 96% values from the original. Furthermore, we used the standard scalar function on these columns to scale and normalize them.





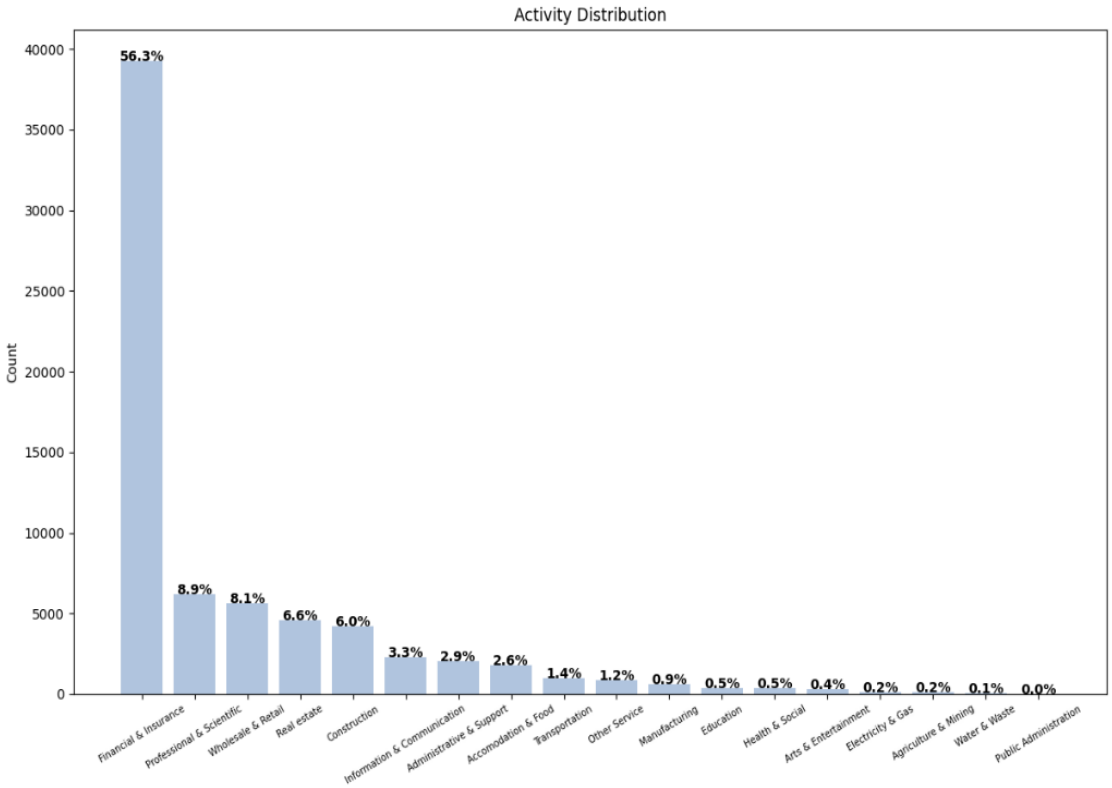


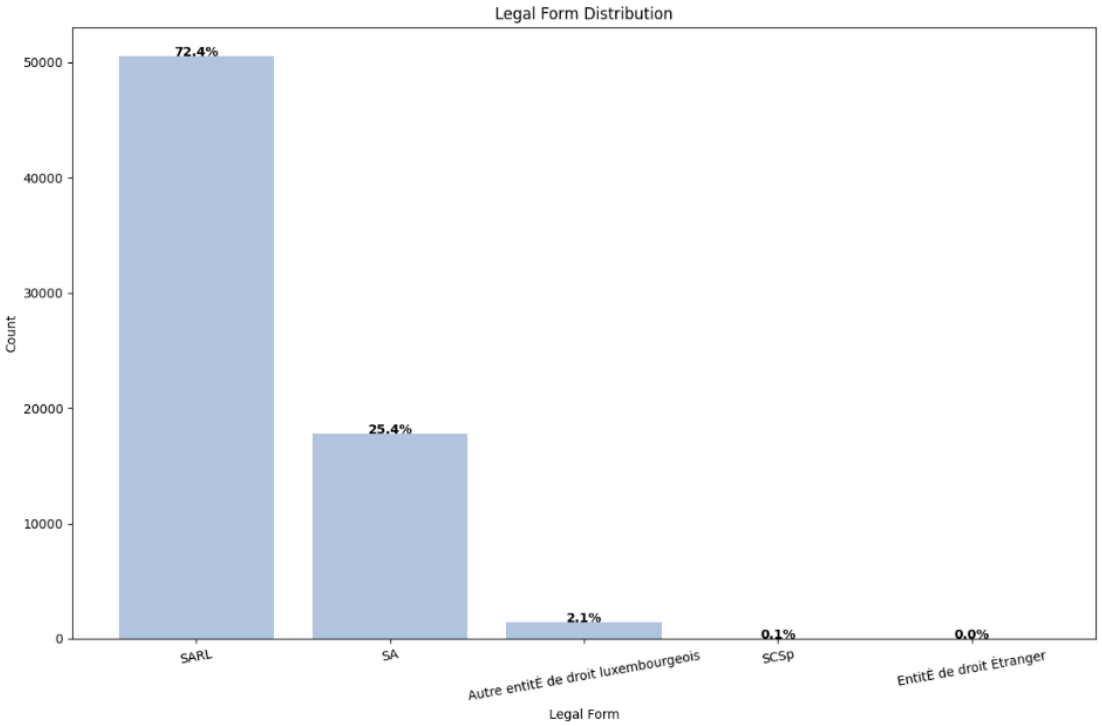


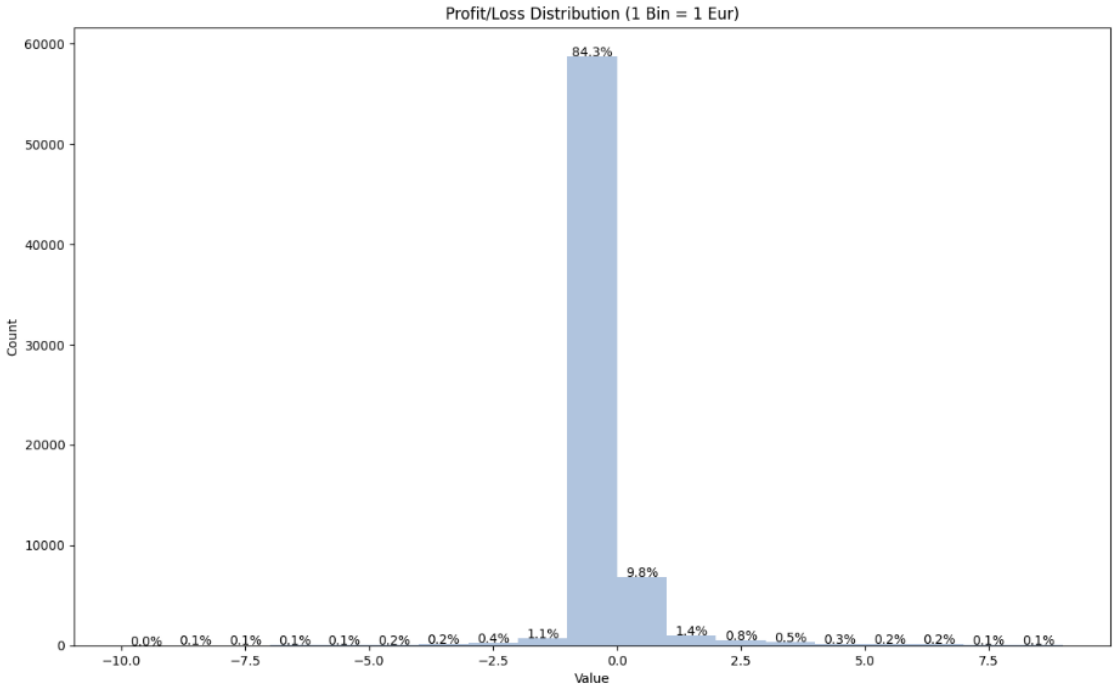
We created multiple output files with fundamental differences to use different clustering techniques. In the first case, we chose not to separate the profit and the loss values from the ProfitLoss column. In the second case, we separated them and converted the negative values (loss column) to positive ones. Finally, in the third case, we treated the financial data as categorical data by hot encoding the categorical fields and generating a new area that displayed the company's profitability on account of total assets and profit and loss.

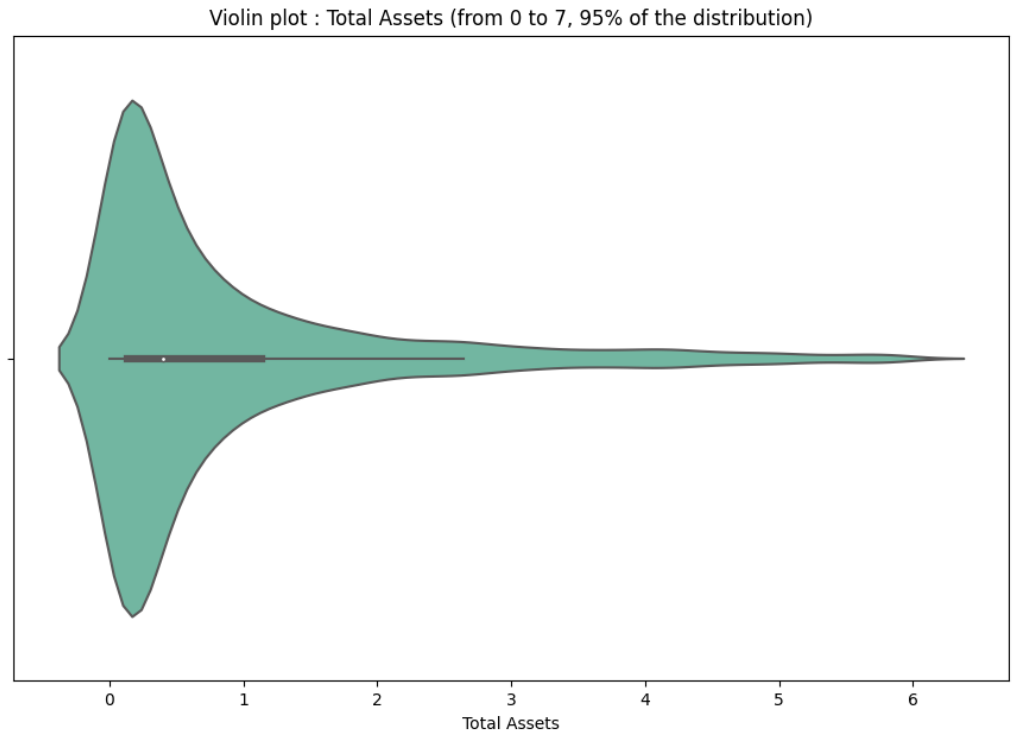
2.1 Summary Statistics

Concerning the statistics part of the clustering, the method used was quite simple, i.e., going through the distributions of the numerical columns to understand the data more. Coupled with splits with the categorical data, the overview allows us to learn which paths we will use for clustering.

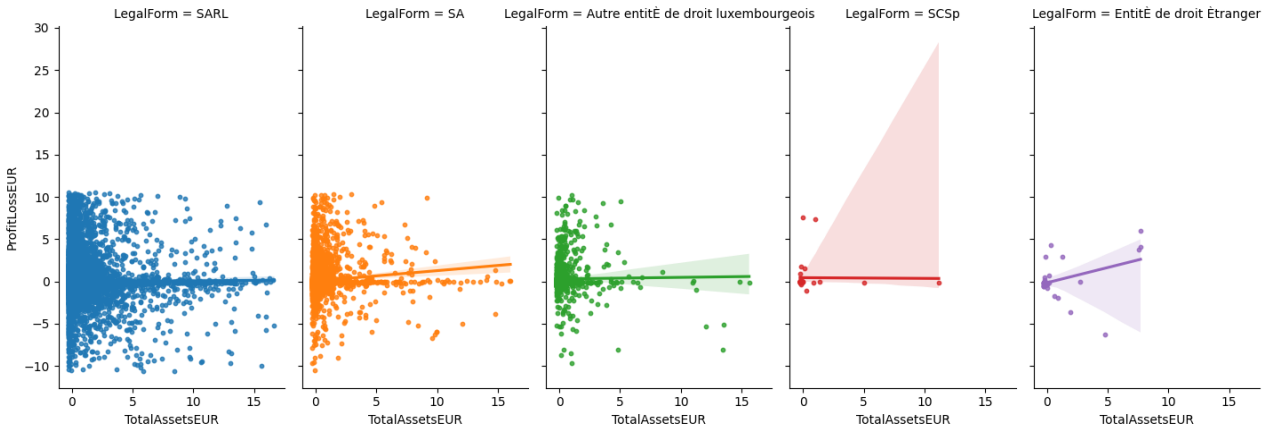



Legal Form Distribution
72.4%
25.4%
2.1%
0.1%
0.0%
Count
50000
40000
30000
20000
10000
0
SARL
SA
Autre entitÉ de droit luxembourgeois
SCSp
EntitÉ de droit Étranger
Legal Form


Profit/Loss Distribution (1 Bin = 1 Eur)
Count
60000
50000
40000
30000
20000
10000
0
84.3%
9.8%
1.4%
1.1%
0.8%
0.5%
0.4%
0.3%
0.2%
0.2%
0.2%
0.1%
0.1%
0.1%
0.1%
0.1%
0.1%
0.0%
−10.0
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
Value

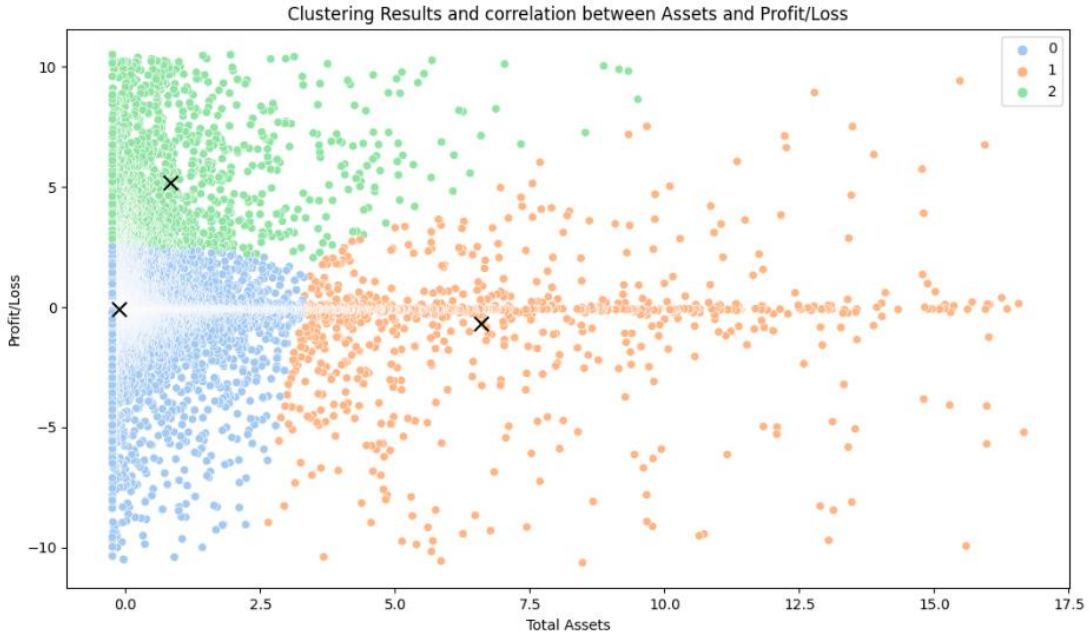


The understanding of the dataset lies in the significant weight taken by the distribution of the activity (56% of Finance and insurance) and its massive representation of SARL (75%). The profit distribution shows us a vast, gigantic loss between 0 and -1 euros (85%) and a gain between 0 and 1 euros (10%), spreading for the rest between -10eur and + 8eur. The distribution of assets by the view of the violin chart shows the dominance of low-value assets (between 0 and 1 euros).



The point cloud of assets and profits by legal form shows the preponderance of SARLs in the dataset. All the regression lines do not show significant results of correlations between assets and earnings except for slightly positive SAs (25% of the total distribution). In addition, we observe that SARLs and SCSp follow the same lines of regression, namely neutral or even slightly positive.

The clustering created based solely on profit and assets shows three distinct groups:



- **Group 0** - very homogeneous with a profit/loss between -10 and +2.5 for assets purchased between 0 and 2.5.
- **Group 1** - very heterogeneous with substantial disparities with profits between -10 and + 10 for assets obtained between 2.5 and 17.
- **Group 2 -** homogeneous with a profit between +2.5 and +10 for an asset purchase between 0 and 7.5 concentrated around 1.

Based on this alone, group 2 does the best 'good deal,' unlike group 0, which is pretty similar in asset purchase but different in income obtained. However, group 1 is very heterogeneous and may need to be more credible when comparing the other groups.

3.1 Clustering

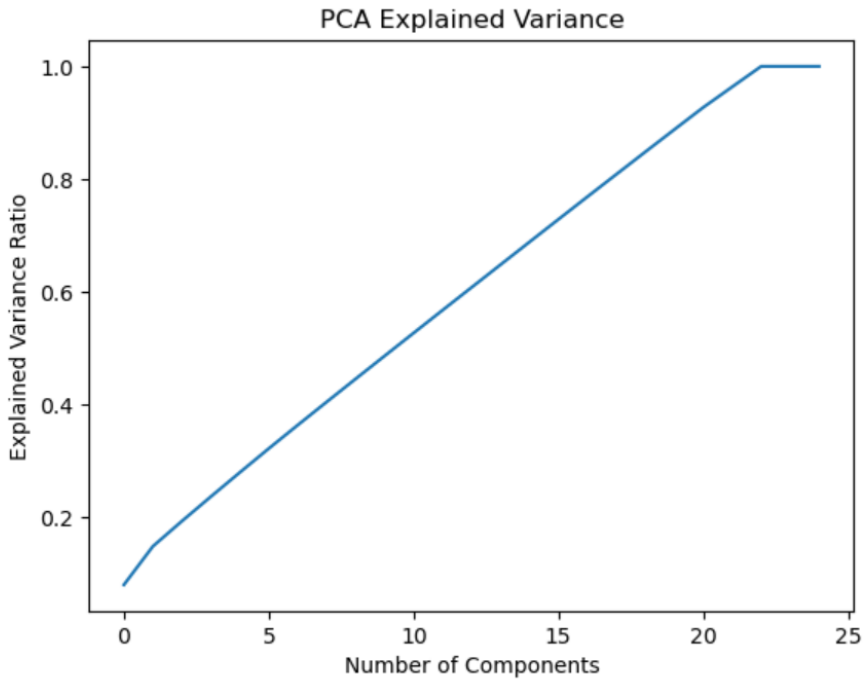
Firstly, to cluster this dataset, we have decided only to keep four variables which are the following: Activity, Company type, Profit and Asset. As a result, the main issue was to realize a clustering with a combination of numerical and categorical variables. After some research, we realized that the K-prototype is the best way of clustering for a mix of numerical and categorical variables, so we tried it.

K-prototypes

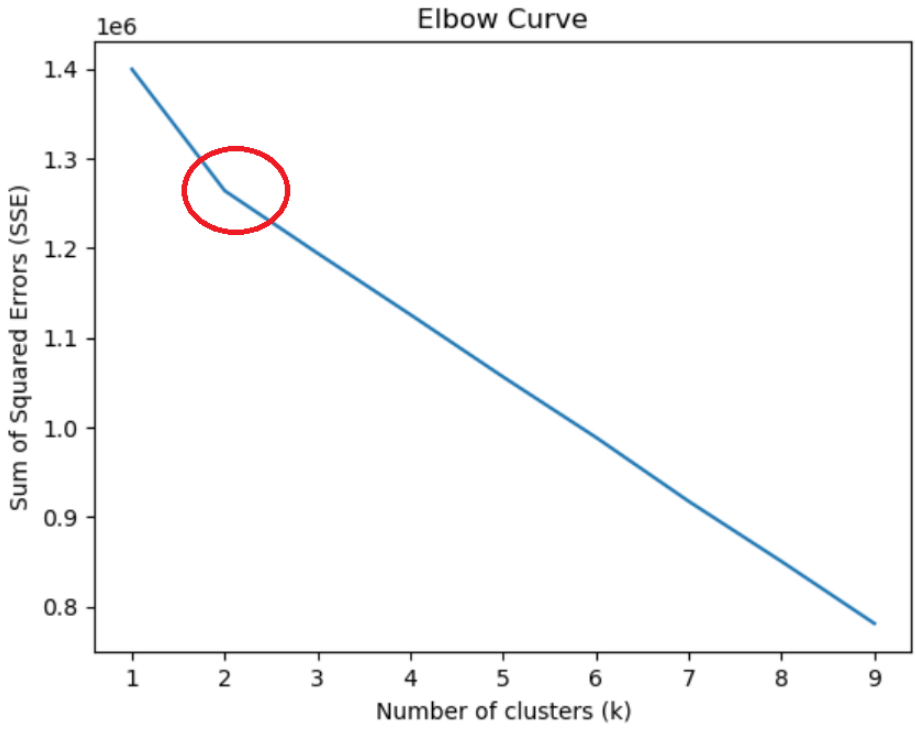
Unfortunately, it has been a failure: 3 clusters have been created, but 98 % of the companies were in the same cluster (67 377 out of 68611). Therefore, even if it was not recommended, we tried implementing K-mean even with categorical variables.

3.1.1 K-means

The idea was to realize a hot encoding and to make the hot encoding variables continuous by standardizing them. However, after doing a PCA, the number of dimensions was still too great to explain the variance. Indeed 18 dimensions were necessary to get 80 % of the conflict, as shown in the graph below.



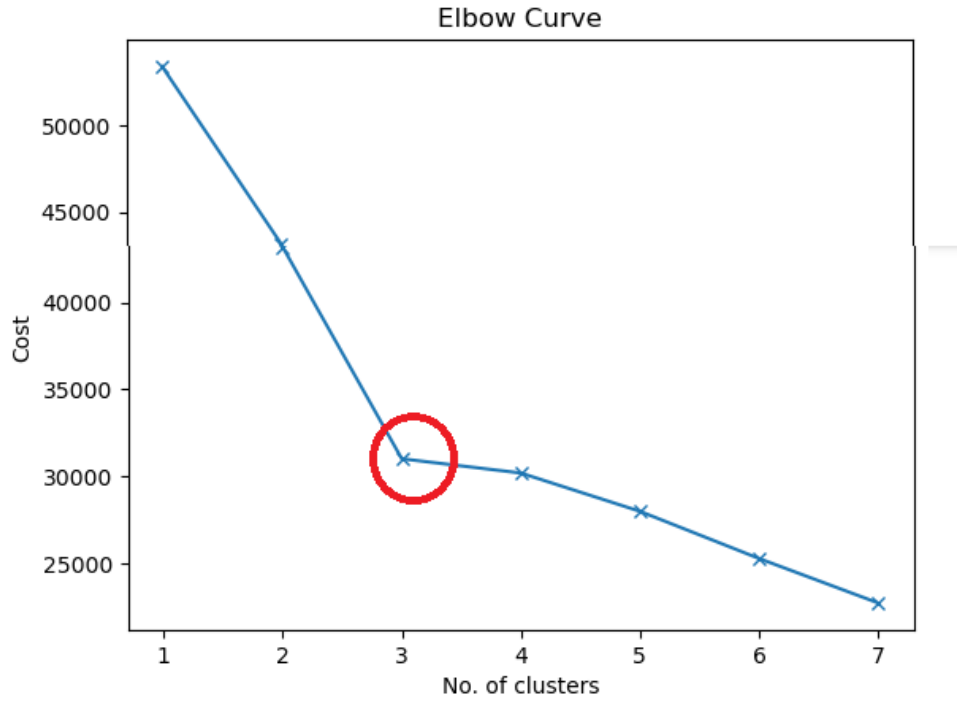
Moreover, after doing the elbow curve, the optimal number of clusters was two, which could have been more interesting.



Therefore, we decided to switch our way of solving the problem by transforming the numerical variables into categorical variables to realize a K mode.

3.1.2 K-Mode

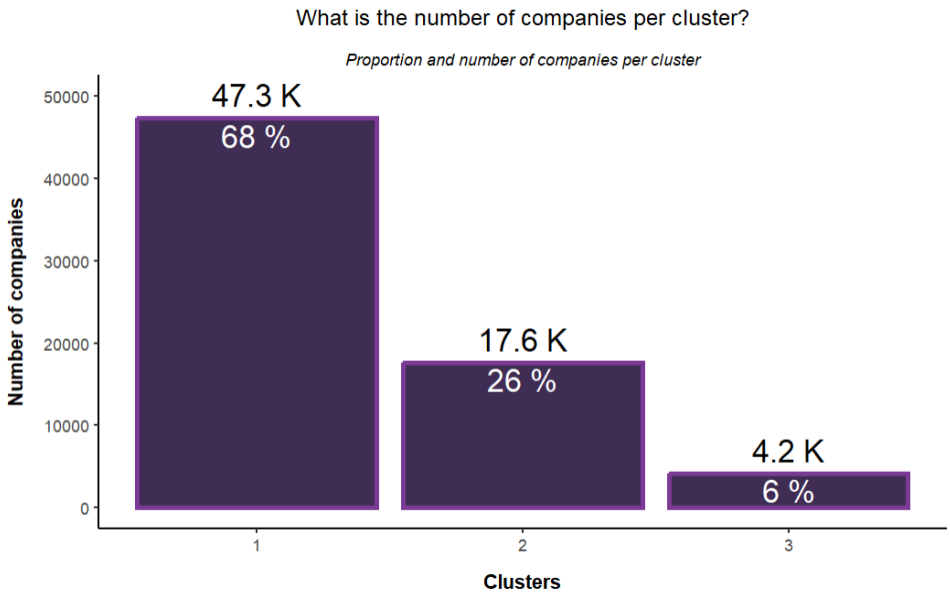
To transform the numerical variables into categorical variables, we divided the interval of the variables into five groups. So small profits are called Profit1, then we go until Profit5, and the same for the asset. However, it's essential to understand that each group does not have the same number of companies. Indeed, if we had chosen quartiles, it would not have represented the reality that the clustering would have been biased. By choosing intervals with the same lengths, we keep the shape of the original dataset, which is not spread out equally.



Then thanks to the info from the elbow curve, we realize a clustering of 3 different clusters thanks to the K-mode algorithm.

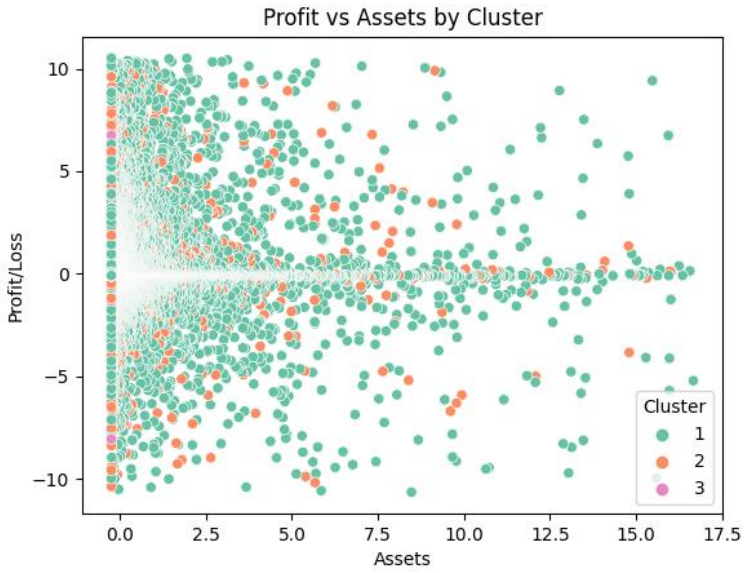
3.2 Clustering visualizations

We notice immediately that cluster 1 is composed of a significant part of the companies:



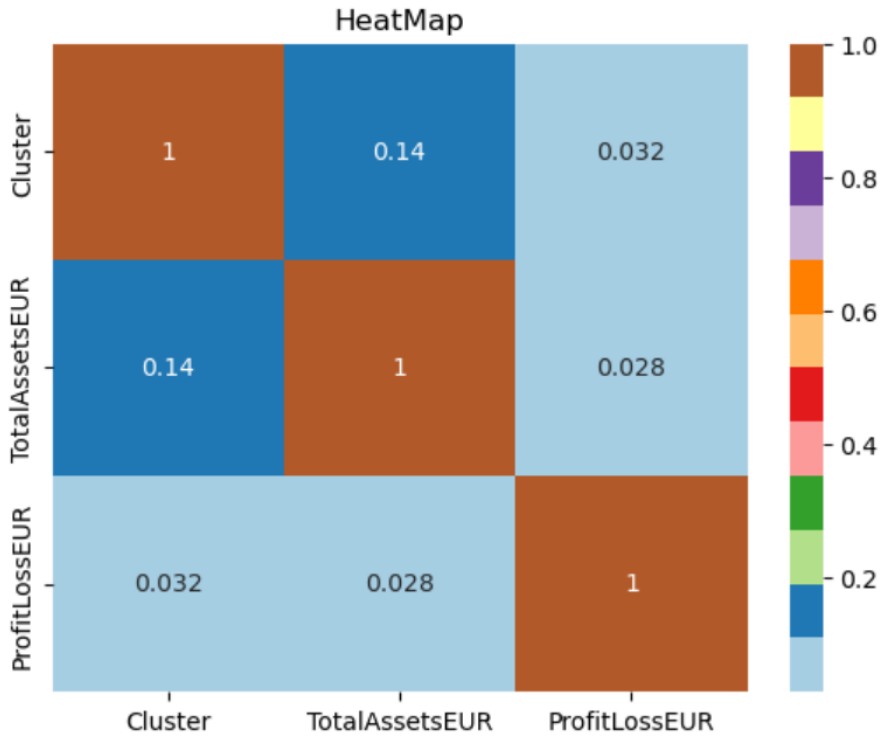
Let's try to understand what our clusters are composed of and why K mode has built these clusters like this. Let's begin with the profit and asset influences:

3.2.1 Profit Asset Analysis



We note that by considering the legal form and the type of activity, we observe different clusters than we proposed previously when we only thought of profit and assets. At this stage, it isn't easy to discern a pattern as we did when we took a few variables into account.

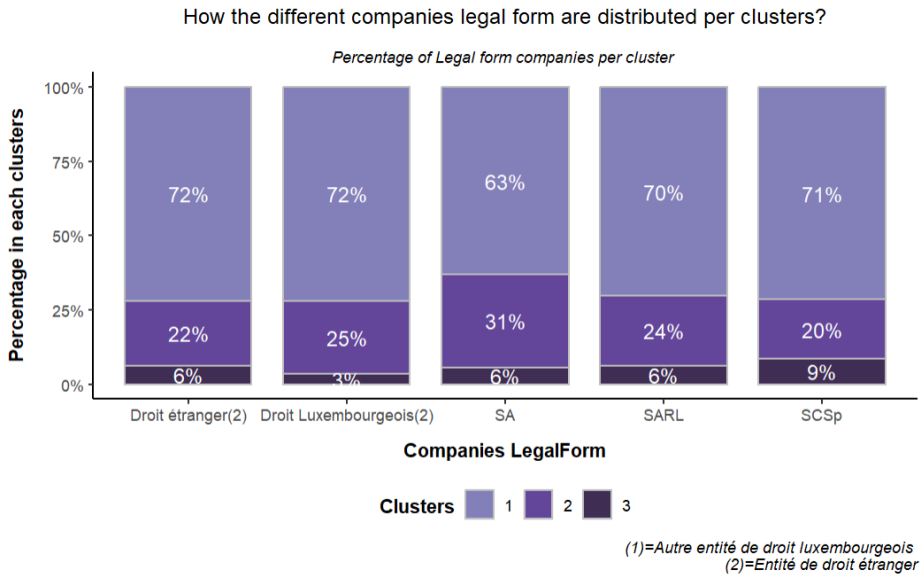
By splitting by Cluster. The 3 clusters are similar at this stage, and their sizes can only differentiate themes. However, we notice some tendencies when comparing them: the more assets bought at a higher price, the more neutral they are in profit or loss. And conversely, the more assets that have been purchased at a low cost, the more they tend to have a profit or a significant loss. This gives the impression that the left of the 3 clusters is prolonged on the abscissa axis.

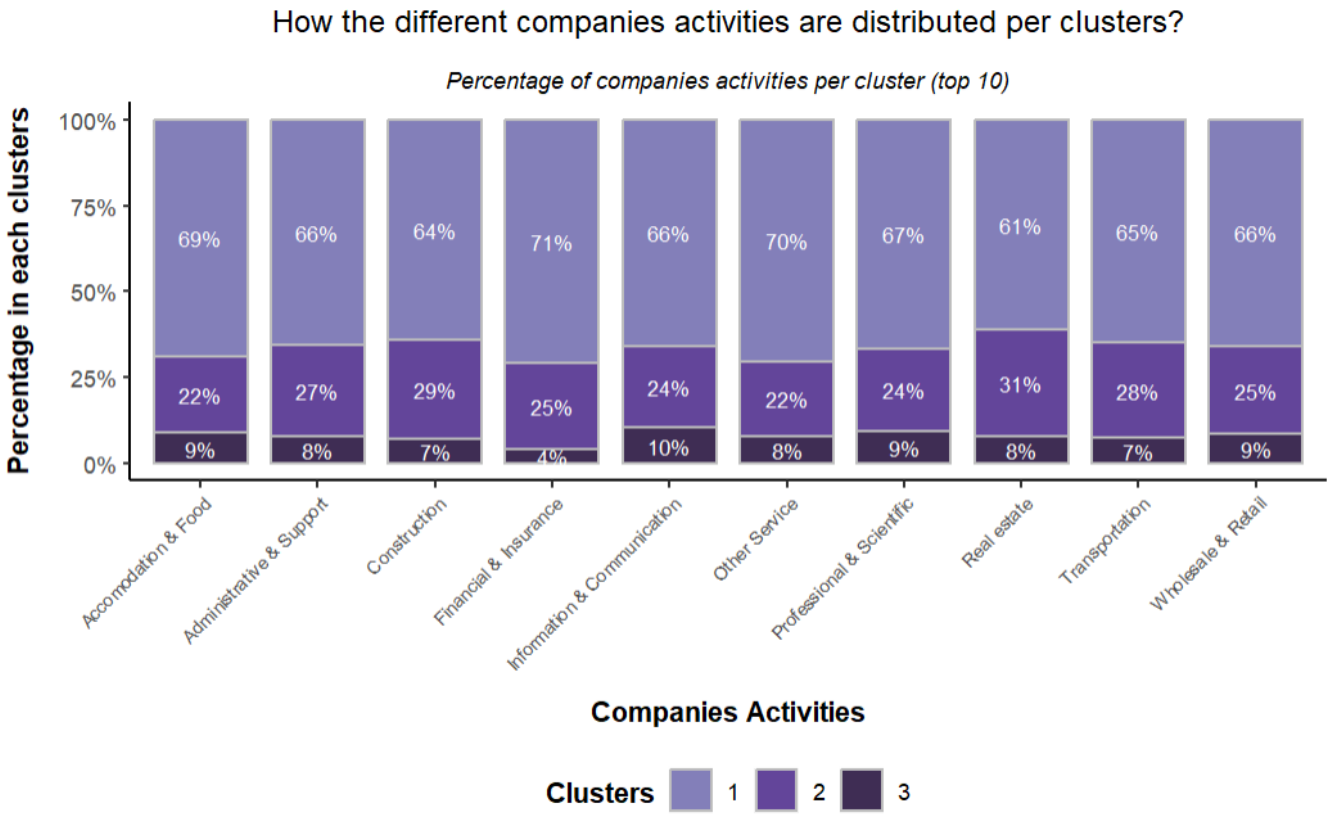


In addition, this weak correlation between assets and clusters is visible on the heatmap (dark blue).

3.2.2 Activities Legal Form Analysis

Now let's check if one of the activities or Legal forms is overrepresented in one cluster:





As you can see in these graphs, activities and legal forms are proportionately represented in the charts, so there is little influence on actions and legal documents.

4.1 Practical implementation of our findings

A large majority of the companies have similar characteristics. There is no singularity. This reflects a solid business base.

Moreover, the clusters are mostly very dense. Therefore, we would need more information to make further recommendations.

In summary, analyzing and clustering the financial information available for multiple entities for a given year can provide valuable insights into the health and performance of companies. Furthermore, by using clustering techniques and in-depth analysis, similar entities can be grouped, and vital financial characteristics of each group can be highlighted, which can help inform investment and risk management decisions.