Imbalanced Data Learning:

a Python implementation of Random Over-Sampling Example algorithm.

Table of contents

BvD.ID.number
Company.name
Country.ISO.Code
Postcode
City
NACE codes
NACE.Rev..2.main.section
NACE.Rev..2.Core.code..4.digits.
NACE.Rev..2.Primary.code.s.
Cons..code
BvD.Independence.Indicator
BvD.major.sector
Standardised.legal.form
Category.of.the.company
No.of.companies.in.corporate.group
No.of.recorded.shareholders
No.of.recorded.subsidiaries
No.of.recorded.branch.locations
Fixed.assets.th.EUR.2010
Intangible.fixed.assets.th.EUR.2010
Tangible.fixed.assets.th.EUR.2010
Other.fixed.assets.th.EUR.2010
Current.assets.th.EUR.2010
Stock.th.EUR.2010
Debtors.th.EUR.2010
Other.current.assets.th.EUR.2010
Total.assets.th.EUR.2010
Shareholders.funds.th.EUR.2010
Capital.th.EUR.2010
Other.shareholders.funds.th.EUR.2010
Non.current.liabilities.th.EUR.2010
Long.term.debt.th.EUR.2010
Other.non.current.liabilities.th.EUR.2010
Current.liabilities.th.EUR.2010
Creditors.th.EUR.2010
Other.current.liabilities.th.EUR.2010
Cash...cash.equivalent.th.EUR.2010
Operating.revenue..Turnover..th.EUR.2010
Sales.th.EUR.2010
Operating.P.L...EBIT..th.EUR.2010
Financial.revenue.th.EUR.2010
Financial.expenses.th.EUR.2010
Financial.P.L.th.EUR.2010
P.L.before.tax.th.EUR.2010
Taxation.th.EUR.2010
P.L.after.tax.th.EUR.2010
P.L.for.period...Net.income..th.EUR.2010

1. Introduction

> *"It is the time you have wasted for your rose that makes your rose so important."* - Antoine de Saint-Exupery

Imbalanced learning refers to a classification or regression problems where the dataset classes are not represented equally. Common examples of such problems are churn, fraud and anomaly detection and clinical data, when one of the classes is rare because problematic, costly, unethical, dangerous to produce, or unexpected. Class unbalancing, specified as the proportion in the number of samples in different classes, can reach values in the orders of $10^2 \div 10^4 : 1$ and up to $10^5 : 1$ [1]

Fig. _ Examples of balanced and imbalanced datasets.

Most real datasets does not have exactly equal number of instances in each class, and while a small difference seldom matters, a heavy imbalance can quickly become a bottleneck in model training. Most learning methods have been conceived to identify the classification rules that better fits the data by some global accuracy criteria. Their target is to minimize the global error, that may be influenced enough from the minority class. Some methods, like the broadly used logistic regression, is more vulnerable to this effect, but even non-parametric methods, like trees, and association rules, are immune from this effect. For example tree generated from imbalanced datasets will have an high accuracy on the prevalent class and a very low precision in identifying the rare event. It appears evident how things become worse when the minority class is the event of interest, like a positive diagnosis of cancer in a patient.

A brief description of the sections of this work:

More data, less data. The most heard sentence in machine learning community is "You need more data!". Still, a large dataset might indeed expose a different, and perhaps more balanced perspective on the classes: more minority examples can indeed be useful. Other strategies may include considering more than once one or more minority samples. Chapter 1 will review the bibliography about solution for this problem, giving a view over cost-sensitive learning and different oversampling and undersampling methods, their advantages and disadvantage.

Random Over Sampling Examples In chapter 2 we will focus on one of these techniques, henceforth named just by its acronym ROSE, that propose a smart, albeit simple, way to generate new data from existing ones.

The Accuracy Paradox. To assess the performance of a solution a metric is needed. When a class represent almost the totality of a dataset, a learning algorithm can achieve a good accuracy by classifying every test sample as belonging to the majority class. To avoid this problem, different metrics has been developed to assess the real model utility and assessing capabilities. Chapter 3 will review available metrics that can be used to effectively evaluate performance of resampling methods.

A method is as good as it is available. To make ROSE easy to use, the main activity of this work involved incorporating it in the most used Python machine learning package, `scikit-learn` , and in particular in its sub-project `imbalanced-learn`. We will overview the development methods, CI/CD, software testing and documentation, in chapter 4.

But is it good? In chapter 5 will set up a wide testing framework for evaluating performance of resampling methods over 27 different famous standard datasets commonly used for classification problems. Different supervised models has been trained and tested on imbalanced and balanced data, and their performance reported. But toy datasets are usually easier to balance. In chapter 6 we used ROSE to dramatically improve classification capabilities of different models in a real-world dataset, with the aim of forecasting the economic outcome of small firms.

How can I use it? In appendix 1 we will show code snippets, use cases and links to repositories that will facilitate user's experience with ROSE, and guarantee repeatability of all experiments included in the present world.

2. Imbalanced learning

2.1. Imbalanced dataset problem

Despite the fact that in literature most imbalanced learning problems are traditionally referred to binary datasets, real world datasets can often be multiclass, as in microarray research[2], protein classification [3], medical diagnostics[4], activity recognition[5], target detection[6], and video mining[7]. Extending imbalanced classification to multi-class scenarios is a natural path, then. As the number of classes increases, so does the challenge of representing the whole problem space accurately, and the need for taking into account the presence of multi-minority and multi-majority classes [8].

In many problem, imbalancing is intrinsically tied to the nature of the data, and not due to lack of sampling, bias, or other sampling errors. In other cases no enough samples of a specific class exists at all.

Most learning methods' loss functions are supposed to be minimized globally, under the assumptions that all class have the same weight. When data are imbalanced, the learning process often achieves this objective by focusing on majority class, leading to bad performance [9], with higher errors on minority classes.

The lack of model effectiveness in prediction of rare classes has been deeply discussed in literature. Both parametric and non parametric methods appear to be sensitive to imbalancing. As an example in logistic regression, one of the most used for binary classification, this effect depends from an underestimation of conditional probability of the rare class [10],[11].

Even the more flexible non-parametric methods, like classification trees and association rules are immune from the effect of asymmetric class distribution. Trees, for example, are being grown finding the recursive divisions of the parameter space that maximize the impurity reduction. The imbalance found is the dataset will be often mirrored in the imbalance of the accuracy over different classes [12], [13]. Even association rules, being selected by their supports, tend to underperform[14],[15].

2.2. Treating imbalanced datasets

Many solution has been advanced to treat imbalanced data problems. Most fall in one of the following two approaches: using cost-sensitive learning models, and resampling the data.

2.2.1. Cost-sensitive learning

Cost sensitive learning is an umbrella term for algorithms in whose objective function it is possible to assign a different cost to misclassification of different classes. An intuitive example of this approach can be imagined when talking about a binary clinical cancer test: a false positive will lead to some extra exam, while a false negative will probably cost a life. The most logical decision is to estimate the relationship between these costs, and assign a larger (*hopefully, much larger*) cost to a false negative.

For multiclass data, a cost matrix $\mathbf{C}$ is computed, where $\mathbf{C}_{i,j}$ will be the cost of misclassifying a sample belonging to the class j as it were belonging to the class i [16], [17].

2.2.2. Resampling

A different, alternative approach against imbalancing can be tried by preprocessing the data, instead of modifying the learning rules, using sampling methods. This approach has consistently proven effective, with different degrees [18],[19]. Different resampling methods has been proposed, falling in two categories:

- undersampling methods, where majority class samples are being randomly discarded to remove imbalancing, at the price of sample size, in a non-heuristic way;
- oversampling methods, where different techniques can be used to generate new minority samples from the existing ones. The following sub-chapters (1.1.3 and 1.1.4) gives an overview of these methods.

Oversampling and undersampling presents different pro and cons, leading to the need of an empirical comparison between different methodologies.

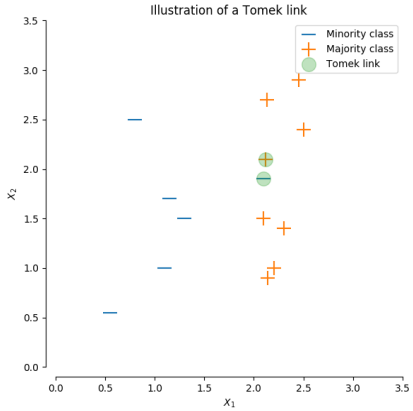
Methods	**Pros**	**Cons**
undersampling	faster learning	loss of sample size
oversampling	slower learning higher computational costs	introduction of artifacts possible overfitting

Despite those problems, resampling is more commonly used than cost-sensitive learning, that is not supported for all learning methods.

2.2.3. Undersampling strategies

Undersampling reduces the size of majority class to avoid imbalancing. In this paragraph we will provide an overview over commonly available undersampling strategies.

- **Random UnderSampler (RUS)**: it works by simply choosing random samples from the majority class.
- **Condensed NN**: [20] uses a 1-nearest neighbor rule to iteratively decide if a sample should be removed or not. It is sensitive to noise and will generate noisy samples.
- **One Sided Selection** [21] and **Tomek Links** instead tends to remove noisy samples.



- **Edited NN** and **Repeated Edited NN** [22] apply (respectively one of more times) a nearest-neighbors algorithm and "edit" the dataset by removing samples which do not agree "enough" with their neighborhood. For each sample in the class to be under-sampled, the nearest-neighbors are computed and if the selection criterion is not fulfilled, the sample is removed. The criterium can be based on majority, or totality of nearest neighbors belonging to the same class of the inspected sample to be kept in the dataset.
- **All KNN** is another iterative process that does the same of the latter, but incrementing at each iteration the number of considered neighbors.
- **Near Miss** [23] is a collection of three different algorithms that respectively:
    - selects the majority samples for which the average distance to the k *nearest* neighbors of the minority class is the *smallest*, or
    - selects the majority samples for which the average distance to the k *farthest* neighbors of the minority class is the *smallest*, or
    - first keel the M-nearest neighbors are kept, then, the majority samples selected are the one for which the average distance from the k nearest neighbors is the *largest*.
- **Neighborhood Cleaning Rule**[^Laurikkala, 2001] focuses on cleaning the data without condensing them.
- **Instance Hardness Threshold** [24] trains any classifier on the data, and the samples with lower probabilities are removed from the dataset. It is not guaranteed to output a balanced dataset, though.

2.2.4. Synthetic data generations

In this section we present the most commonly used oversampling techniques and their variants:

- Synthetic Minority Oversampling TEchnique (SMOTE) based
- ADAptative SYNthetic sampling (ADASYN).

2.2.4.1. SMOTE based methods

SMOTE[25] is a class of resampling algorithms that use the following approach:

- a random sample from the minority class is chosen
- his k-neighbors are found (default k=5)
- lines are drawn from the original sample to the neighbors
- new examples are drawn randomly along these lines, with $x_{new} = x_i + \lambda * (x_{nn} - x_i)$, where λ is drawn from $Uniform(0, 1)$, or other distributions.

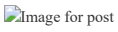


Fig. _ : SMOTE resampling concept.

There are many variants of SMOTE that has been developed to improve its performance.

Borderline SMOTE [26] will classify each sample x_i to be:

1. *noise*, when all k-neighbors are of a different class from x_i
2. *in danger*, when at least half of the neighbors belongs to the same class
3. *safe*, when all neighbors belongs to the same class.

The algorithm will then use "*in danger*" samples to generate new samples, with the same procedure of SMOTE.

K-Means SMOTE [27] uses a K-Means clustering method before to apply SMOTE. The clustering will group samples together and generate new samples depending of the cluster density.

SMOTENC [25] slightly change the way a new sample is generated by performing something specific for the categorical features. In fact, the categories of a new generated sample are decided by picking the most frequent category of the nearest neighbors present during the generation.

SVMSMOTE [28] fits a Support Vector Classifier to find support vectors and generate samples considering them. Tuning the C parameter of the SVM classifier allows to select more or less support vectors.

2.2.4.2. ADASYN

ADASYN [29] works similarly to the regular SMOTE. However, the number of samples generated for each x_i is proportional to the number of samples which are not from the same class than x_i in a given neighborhood. Therefore, more samples will be generated in the area where the nearest neighbor rule is not respected.

2.2.4.3. Combination and Ensemble methods

Combinations of different methods can be used efficiently. SMOTE based methods can generate noise when generating point between marginal outliers and inliers. After the resampling this issue can be solved by cleaning the space resulting from oversampling.

Two methods used for this purpose are:

- **Tomek's links**: [30] an undersampling technique used to remove unwanted overlaps between classes, where majority class links are removed until minimally-distanced neighbors pairs belong to the same class. Two instances form a Tomek's link if:
    - one of them is noise (*see Borderline SMOTE definition of noise)*, or
    - both are near a border

  In other words, if they are each other's closest neighbor, and of different classes.
- **Edited nearest-neighbors** [31] uses asymptotic convergence properties of nearest neighbor rules that use an editing procedure to reduce the number of preclassified samples and to improve performance [32]

Ensemble methods can be used generate undersampled subsets of many different oversampled datasets, or by bagging different undersamplers. Additionally, pipelines can be assembled, to chain different methods.

3. Random over-sampling examples (ROSE)

ROSE provides a different methodology to deal with imbalanced samples. As its alternatives do, it alters the distribution of the classes, using the following solution, based on the generation of new artificial data from the classes, according to a smoothed bootstrap approach [33]. It focuses on $\mathcal{X}$ domains included in $\mathbb{R}^d$, that is $P(\boldsymbol{x}) = f(\boldsymbol{x})$, a probability density function on $\mathcal{X}$. We consider that $n_j < n$ is the size of $\mathcal{Y}_j, j = 0, 1$. The ROSE procedure to generate a single new artificial sample consists in drawing a sample from $K_{\mathbf{H}_j}(\bullet, \mathbf{x}_i)$, with $K_{\mathbf{H}_j}$ a probability distribution centered at $\mathbf{x}_i$, and $\mathbf{H}_j$ a matrix of scale parameters, determining the width of the extracted sample neighborhood.

Usually $\mathbf{H}_j$ is chosen in the set of unimodal symmetric distributions. Once a class has been selected,

$$\begin{aligned}\hat{f}(\mathbf{x}|y = \mathcal{Y}_j) &= \sum_{i=1}^{n_j} p_i Pr(\mathbf{x}|\mathbf{x}_i) \\ &= \sum_{i=1}^{n_j} \frac{1}{n_j} Pr(\mathbf{x}|\mathbf{x}_i) \\ &= \sum_{i=1}^{n_j} \frac{1}{n_j} K_{\mathbf{H}_j}(\mathbf{x} - \mathbf{x}_i).\end{aligned}$$

such as, in this framework, the generation of new examples from the class $\mathcal{Y}_j$ will correspond to the generation of data from the kernel density estimate of $f(\mathbf{X}|\mathcal{Y}_j)$, to generate a new synthetic balanced training set $\mathbf{T}_m^*$. Usually m is set to the size of majority class, but can be set lower to perform under-sampling. The choice of K and $\mathbf{H}_j$ was addressed by a large specialized literature on kernel density estimation [34]. By choosing $\mathbf{H}_j \to 0$, ROSE collapses to a standard combination of over- and under-sampling.

Apart from enhancing learning, the generation of synthetic examples from an estimate of conditional densities of the classes may aid the estimation of learner accuracy and overcome the limits of both resubstitution and holdout methods. Resampled datasets can be efficiently employed in leave-K-out or bootstrap estimation.

3.1. Assumptions

ROSE requires that the resampled variables are of numeric type, being impossible to fit a multivariate kernel on unordered categorical variables. This can include variables with limited numeric support (e.g. $\{0, 1\}$, or percentage values.).

Variables belonging to $\mathbb{N}$ could generate non-integer samples. This problem can be contained by rounding.

Variables belonging to $\mathbb{N}^+$ or $\mathbb{R}^+$ domains poses another problem, since samples drawn from the kernel function are not guaranteed to be positive. This particular problem can be contained by a log-transform of the original dataset parameters.

Relatively to our work described in Chapter 4, future development of ROSE will consider the option to extend the class by including type inference or by collecting `numpy.array` and `pandas.DataFrame` dtypes data to dynamically change the random sampling function.

3.2. Kernel methods

Since 90s estimation and learning methods using positive definite kernels have become popular, particularly in machine learning [35]. Real world analysis problems ofter require nonlinear methods to detect the kind of dependencies that allows successful prediction of properties of interests.

The operational use of ROSE requires a prior specification of the $\mathbf{H}_j$ matrices. In principle this leads to a criticality, since different choices of the smoothing matrices leads to larger or smaller $K_{\mathbf{H}_j}$, namely larger or smaller neighborhoogs of the observations from which the synthetic samples are generated. There is a large body of literature on methods of choice of the smoothing parameters [36] , [34]. The idea beyond these methods is to minimize an optimality criterion, as the asymptotic mean integrated squared error (AMISE).

$$AMISE(h;r) = \frac{R(K^{(r)})}{nh^{2r+1}} + \frac{1}{4}h^4\mu_2^2(K)r(f^{(r+2)})$$

Among all possible alternatives, Menardi and Torelli's proposal is to use Gaussian Kernels with diagonal smoothing matrices $\mathbf{H}_j = diag(h_1^{(j)}, \ldots, h_d^{(d)})$, and minimize AMISE.

This leads to:

$$h_q^{(j)} = \left(\frac{4}{(d+2)n}\right)^{1/(d+4)} \hat{\sigma}_q^{(j)} (q = 1, \ldots, d; j = class)$$

where $\hat{\sigma}_q^{(j)}$ is the sample estimate of the standard deviation of the qth dimension of the observation belonging to the class $\mathcal{Y}_j$. Despite the naivety of this approach, authors reports good results, since the only interest is producing a reasonable neighborhood where to sample the new data from, and happens to perform well even if $f(\mathbf{x}|y = \mathcal{Y}_j)$ is not $Normal$, just unimodal.

Choice of $\mathbf{H}_j$ smoothing matrix gives control on data generation:

In the following image we generated three blobs of examples from multivariate normal distributions. For the three classes, n will be respectively 33, 50 and 170.

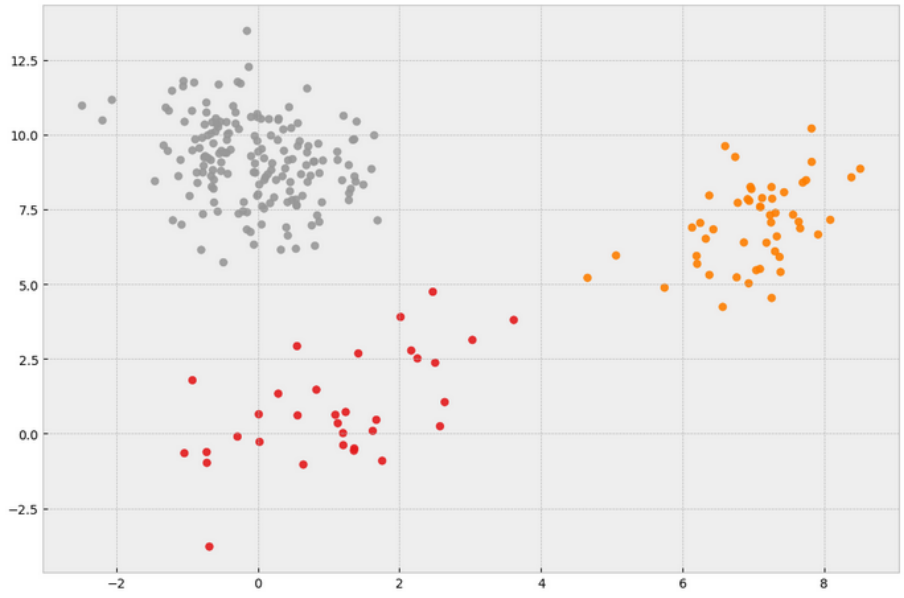


Fig: _ unbalanced classes.

In the next figure, we used ROSE to rebalance the datasets, and bring n to a total of 300 examples per class.

Fig _ : rebalanced datasets. Original data points are marked with grey"X".

Rose can use a **shrink factor** vector, to shrink kernels independently for each class. The following figure show how, decreasing the shrink factors, new data will be more and more closely clustered around original data points.

Fig _ Using shrink factors: $grey = 0.2, orange = 0.5, red = 1$.

4. Metrics

Evaluating performance is a critical part of building a machine learning model. In this chapter we will describe some of these tools, and how to choose the best one for our purposes in imbalanced data problems.

4.1. Confusion matrix

Confusion Matrices (henceforth CM) are tables that can be used to describe the performance of a classifier on a test set of data for which true values are known. They are detailed and simple to understand, but does not summarize well the performance.

n = 165	Predicted: NO	Predicted: YES
Actual: No	50	10
Actual: Yes	5	100

Table _ : An example of a confusion matrix for a binary classifier.

On the diagonal we find correctly predicted samples (true negatives, or TN, and true positives, or TP), leaving misclassified data on other cells (false positives, or FP, and false negatives, or FN). Confusion matrices can be extended to multiclass classifiers, their size becoming $j \times j$, for classes in $\mathcal{Y}_j$. Sums over rows and column will describe the total of actual vs predicted predictions. We have seen how secondary indexes can be computed form these values and their ratios.

When describing a model's performance, the most common classification metric is its $Accuracy$, defined as

$$Accuracy = (TP + TN)/(TP + TN + FP + FN)$$

This can be misleading, when the problem uses imbalanced data. Consider a sample with a 100:1 imbalance ratio. Classifying all values as the majority class will gives a $\sim 99\% Accuracy$ score. [37]. Different solutions has been proposed to solve this issue. For example, $BalancedAccuracy$ score, defined as

$$Balanced\ Accuracy = \frac{\frac{TP}{P} + \frac{TN}{TN+FP}}{2}$$

can help. Another metric is $Predicted\ positive\ condition\ rate$, defined as

$$Predicted\ positive\ condition\ rate = \frac{TP + FP}{TP + FP + TN + FN}$$

which identifies the proportion of the total population correctly identified. Two other commonly used index is $F1$ score and Matthews correlation coefficient.

More informative visualizations of model performances can be given not by indices, but by plots, like Receiver Operating Characteristics and $Precision$ vs $Recall$ plots, that deserves a dedicated description in the following sub-chapters.

Additional metrics that can be extracted from CM are

- Cohen's Kappa, that is a measure of how well the classifier performed as compared to how well it would have performed simply by chance. We left it out

after bibliography reported unreliable results due to high sensitivity to the distribution of the marginal totals [38]

- Null Error Rate, that is how often you would have been wrong if you always predicted the majority class. This can be used as a useful baseline metric to compare a classifier against. Still, the Accuracy Paradox tells us that sometimes the best classifier will still have an high error rate than the null error rate.
- F_1 score. Since we will use it in our test suite later, we will dedicate next sub-chapter to its description.
- K measure, a theoretically grounded measure that relies on a strong axiomatic base.[39]
- confusion entropy, a statistical score comparable with Matthews correlation coefficient, treated below.
- Power's informedness and markedness [40], a couple of interesting alternative metrics that respectively describe how a binary predictor is informed in relation to the opposite condition, and the probability that the predictor correctly marks a specific condition.
- Mattew's correlation Coefficient (MCC), exhaustively treated in a following sub-chapter.

Despite their effectiveness, most of the aforementioned measures does not appear to have achieved such a diffusion in the literature to be considered a solid alternative to MCC and F_1 score. They are good single-valued indicators of performance, supported by a strong bibliography, and useful to compare large numbers of tests.

To have a deeper comprehension of a model's performance we used other two plotted tools: Receiving Operator Characteristic and Precision/Recall plots. The following sub-chapters will describe our four tools in depth.

4.2. F_1 Score

Called also F-score or F-measure, is an accuracy metric, calculated from the precision and recall of the test.

$$
\begin{aligned}
Precision &= \frac{TP}{TP + FP} \\
Recall &= \frac{TP}{TP + FN} \\
F_1 &= 2 * \frac{precision * recall}{precision + recall} \\
&= \frac{2 * TP}{2 * TP + FP + FN}
\end{aligned}
$$

It is a particular case of the more general F_β score, defined as

$$
\begin{aligned}
F_\beta &= (1 + \beta^2) \frac{precision * recall}{(\beta^2 * precision) + recall} \\
&= \frac{(1 + \beta^2) * TP}{(1 + \beta^2) * TP + \beta^2 * FN + FP}
\end{aligned}
$$

where recall is considered β times as important as precision. a $\beta > 1$ will increase recall importance, while $0 < \beta < 1$ will weight recall lower than precision [41]. It has recently been criticized as less informative and truthful than Mattees Correlation Coefficient (see below), especially for imbalanced classes.[42], and the adoption of new metrics is being suggested, like Informedness (Youden's J statistic)[43] and Markedness[44], in fields like biology and linguistics. When using geometric mean instead of harmonic mean of recall and precision it is known as Fowlkes-Mallows index [45]. In multiclass cases, researchers can employ the F_1 micro-macro averaging procedure. [46]. Micro-averaging puts more emphasis on common labels in the dataset, since it gives each sample the same importance, measuring F_1 score of the aggregated contribution of all classes. In macro-averaging the same importance is instead given at every class, regardless of their frequency: a separate F_1 score is computed for each class, and then they are averaged. It may overestimate the score for imbalanced problems.

4.3. Matthews correlation coefficient (MCC)

Accuracy and F_1 score computed on confusion matrices have been (and still are) among the most popular adopted metrics in binary classification task [42]. However these measures can show overoptimistic inflated results, especially on imbalanced datasets. The Mattews correlation coefficient (henceforth, MCC) is instead a more reliable statistical rate which encompass all for confusion matrix categories (TP, FP, TN, FN), proportionally both to the size of positive and negative elements in the dataset.

$$\begin{aligned} MCC &= \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \\ &= \sqrt{\frac{\chi^2}{n}} \end{aligned}$$

It derives from Guilford's ϕ coefficient[47]. Originally developed by Mattews in 1975 for comparing chemical structures, it has been re-proposed by Baldi et al [48] as a standard performance metric in the multiclass case, and American Food and Drug Administration (FDA) employed it as main evaluation measure in Microarray II / Sequencing Quality Contro (MAQC/SEQC)[49]. Nonetheless, it has been reported to suffers from instability in case of imbalanced outcomes. [50]. Despite the existence of Bayesian based improvements and mathematical workarounds, they have not been adopted yet.

4.4. Receiver Operating Characteristic (ROC) and AUC

A Receiver Operating Characteristic (ROC) curve is a plot that summarizes the performance of a binary classification model on the positive class. The x-axis indicates the False Positive Rate and the y-axis indicates the True Positive Rate.

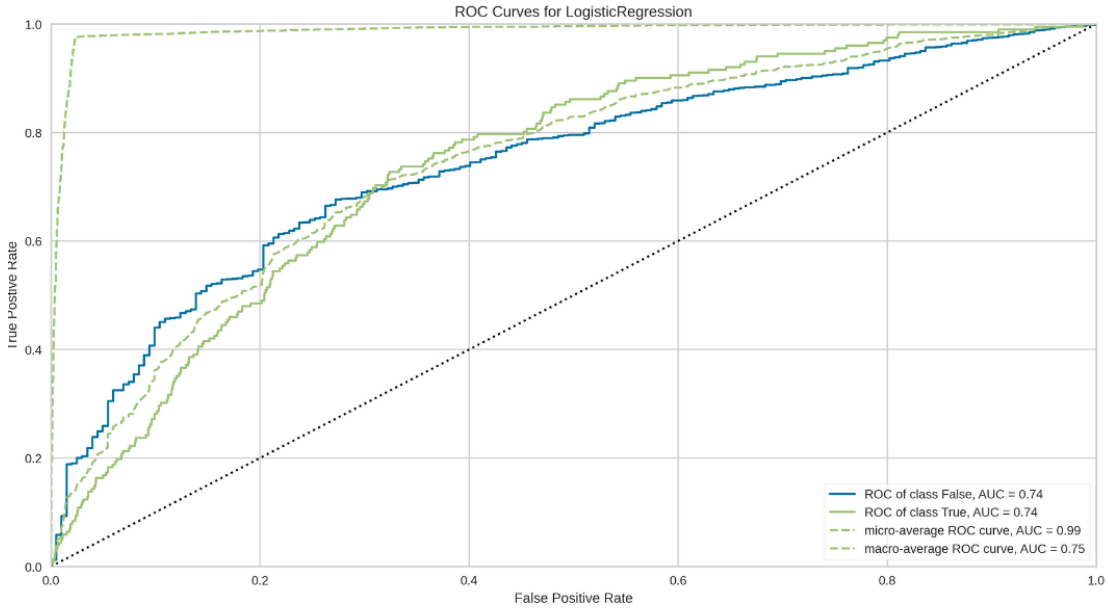


Fig _ : Example of ROC curve. AUC (Area under the curve) are shown in the bottom-right legend.

A ROC gives an intuitive visualization of a classifier performance: the dotted diagonal represent a classifier with no discriminative power, and the more the curve tends to the upper-left corner, the better the classifier is. The area under the curve (AUC) gives a commonly used single-valued index of performance. The threshold is applied to the cut-off point in probability between the positive and negative classes, which by default for any classifier would be set at 0.5, halfway between each outcome (0 and 1). A trade-off exists between the TP rate an FP rate, such that changing the threshold of classification will change the balance of predictions towards improving the TP rate at the expense of FP rate, or the reverse case.

By evaluating the true positive and false positives for different threshold values, the ROC curve is drawn. An interesting property is that the ROC is unbiased towards model that performs well on the minority class at the expense of the majority class, or vice versa, making it an interesting choice when dealing with imbalanced data.

4.5. Precision-recall plots

Precision-recall plots are a powerful visualization tool to evaluate binary classifiers, closely related to the Receiver Operating Characteristic described in the precedent sub-chapter. It shows the relation between these indexes, at the variation of a threshold

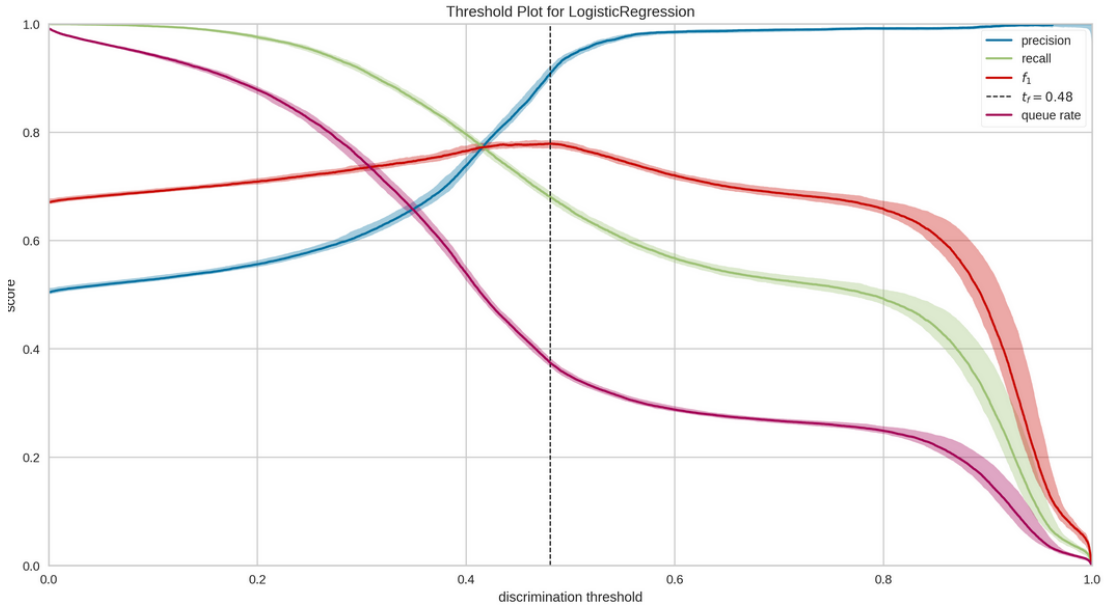


Fig _ Precision-recall plot of a logistic regression model. Bands are confidence interval around values. Queue rate can be seen as the "spam folder" or the inbox of the fraud investigation desk. This metric describes the percentage of instances that must be reviewed. If review has a high cost (e.g. fraud prevention) then this must be minimized with respect to business requirements; if it doesn't (e.g. spam filter), this could be optimized to ensure the inbox stays clean.

5. Implementation of ROSE in the `imbalanced-learn` Python package

As we said, a tool is useful only if it available. ROSE has an already available R implementation [51]. Despite R being the favored programming language among statistician, Python is quickly rising in popularity, and over the years tens of thousands of packages were offered to help researches in mathematic and statistic fields. We decided to avoid contributing on closed source, expensive or ineffective softwares like MatLab, Excel, Stata, SPSS, and contributing to the community by choosing Python.

As of the date of this writing, the best way to start an argument in a group of data scientists is posing the question "So, Python or R?". This work will stay as far as possible from taking a side in this dilemma, both languages offering many pro and cons, opportunities and flaws.

Instead of the simpler choice of publishing a stand-alone library, we decided to maximize the availability of the code extending the already-available `imbalanced-learn` library[52], that is a contributor of the well known `scikit-learn` project[53].

This package offers a lot of functionalities, models and mathematical tools, and its main characteristic is the standard API of its classes, that makes them versatile.

Computationally speaking, ROSE resampling is obtained with the following algorithm (pseudocode):

```
define make_samples (X,y,n,h_shrink):
     n = number of samples to be created
     p = number of features
     S = subset of samples randomly selected from X
     minAMISE = (4/((p+2)*n))**(1/(p+4))
     vars = variance/convariance matrix of all classes
     hOPT=h_shrink*minAMISE*vars
     randoms = multivariate_normal(size=(n,p))
     rose = randoms*hOPT + S
     return rose
```

It uses the well known `numpy` library for matrix calculations and sampling.

5.1. `scikit-learn` context

`scikit-learn` (also known as `sklearn`) is a free software machine learning library for Python. It features algorithms for classification, regression, and clustering, including Support vector machines, tree-based models, boosted models, k-means, and DBSCAN. It is built around the famous `numpy` and `scipy` packages, with some routines written in Cython, to improve performance. Some functions are just wrapping other libraries, like LIBSVM or LIBLINEAR.

It was born in 2007 for the Google Summer of Code competition as "SciKit" (SciPy Toolkit), a third party extension of `SciPy`. The original codebase has been rewritten in 2010 by Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort and Vincent Michel.

It offers a curated integration with different other Python libraries, like `matplotlib` or `plotly` for plotting, `Pandas` dataframes, `sparse` arrays, `numpy` objects, `scipy`, `tensorflow`, `keras`, and more. Among these API-compatible packages we can find `imbalanced-learn`.

At the moment of this writing, the last version number is 0.23.0 (released in May 2020).

5.2. Test driven development

Development in `imbalanced-learn` packages follows strict guidelines, as explained in the project documentation. Pull requests are to be submitted at https://github.com/scikit-learn-contrib/imbalanced-learn/pulls .

If accepted, they can be marked for review by the sender. With a fast and effective peer review process, they enter the project Continuous Integration / Continuous Deployment process (henceforth, CI/CD).

At the moment there are 1588 test units for `imbalanced-learn`, embracing library compatibility, mathematical correctness, error tolerance and numeric problems.

Most test units already encompass mathematical correctness, but we still added a unit to check if the variance/covariance matrix of resampled data is similar to the one of the original dataset, and some check about correct handling of sparse arrays and Pandas dataframes.

Extra test units verify PEP-8[54] compliance about linting and code style. Commit history and review process is available. [55] An example of a successful pipeline build can be read at https://lgtm.com/projects/g/scikit-learn-contrib/imbalanced-learn/logs/languages/lang:javascript

5.3. GitHub and Kubernetes CI/CD

CI/CD is a modern DevOps tool. Code is automatically and continuously pushed to the master branch of the project's repository (we used Github, but Gitlab and other repositories offer the same service). `imbalanced-learn` employs an Azure pipeline.

When a code change is detected, the CI/CD pipeline starts:

- the CI/CD cluster reads a YAML file, with a matrix of configurations: different operative systems, different versions of Python, different version of any used library.
- for every combination, a virtual machine (deployed as Kubernetes containers, in our case) are instantiated.
- at the launch, the pod loads the configuration, and runs all the code test units
- the results of the test units are fed back to the repository
- if all tests are passed, the code can be merged.

Our implementation has been correctly merged, and will be published with the next release of the library. Meanwhile, it can be imported from the `ROSE` branch of the official `imbalanced-learn` repository.

5.4. Documentation

Documentation correctness is integral part of the review process. Functions API are automatically harvested from the code by the `sphinx` documentation library, while theoretical descriptions, application and user guide has been written by the author, and can be found at the official website of the project's documentation, at https://imbalanced-learn.readthedocs.io .

6. Empirical analysis

With the aim of obtaining benchmark the effectiveness of ROSE, a simple test suite has been written, in a Jupyter Notebook.

6.1. Materials & methods

The pipeline evaluates the performance every combination on a grid of models, resampling methods, and parameters.

6.1.1. Datasets

A total of 27 datasets has been used. Data comes from the following repositories, and are available for repeatability. All datasets are loaded from Zenodo repository through `imblearn.datasets.fetch_datasets()` API. Additional informations can be found on `imbalanced-learn` repository documentation.

Short name	Source	Website
UCI	UCI Machine Learning Repository, University of Californa, School of Information and Computer Science	http://archive.ics.uci.edu/ml
LIBSVM	National Taiwan University	https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/
KDD	SIGKDD International Conference on Knowledge Discovery and Data Mining	https://www.biostat.wisc.edu/~craven/kddcup/index.html

Table _ : data sources for empirical testing.

ID	Name	Source repository	Target	Shape(n,p)	imbalance ratio
1	ecoli	UCI	imU	(336,7)	8.6:1
2	optical_digits	UCI	8	(5620,64)	9.1:1
3	satimage	UCI	4	(6435,36)	9.3:1
4	pen_digits	UCI	5	(10992,16)	9.4:1
5	abalone	UCI	7	(4177,10)	9.7:1
6	sick_euthyroid	UCI	sick euthyroid	(3163,42)	9.8:1
7	spectrometer	UCI	≥ 44	(531,93)	11:1
8	car_eval_34	UCI	good, v.good	(1728,21)	12:1
9	isolet	UCI	A,B	(7797,617)	12:1
10	us_crime	UCI	> 0.65	(1994,100)	12:1
11	yeast_ml8	LIBSVM	8	(2417,103)	13:1
12	scene	LIBSVM	>1 label	(2407,294)	13:1
13	libras_move	UCI	1	(360,90)	14:1
14	thyroid_sick	UCI	sick	(3772,52)	15:1
15	coil_2000	KDD	minority	(9822,85)	16:1
16	arrhytmia	UCI	06	(452,278)	17:1
17	solar_flare_m0	UCI	M->0	(1389,32)	19:1

ID	Name	Source repository	Target	Shape(n,p)	imbalance ratio
18	oil	UCI	minority	(937,49)	22:1
19	car_eval_4	UCI	vgood	(1728,21)	26:1
20	wine_quality	UCI	≤ 4	(4898,11)	26:1
21	letter_img	UCI	Z	(20000,16)	26:1
22	yeast_me2	UCI	ME2	(1484,8)	28:1
23	webpage	LIBSVM	minority	(34780,300)	33:1
24	ozone_level	UCI	ozone	(2536,72)	34:1
25	mammography	UCI	minority	(11183,6)	42:1
26	protein_homo	KDD	minority	(145751,74)	11:1
27	abalone_19	UCI	19	(4177,10)	130:1

Table _ : Details on dataset used for empirical test. Columns are internal ID, short name, source repository (see above for complete reference), target column, or value, of the binary classifier, dataset shape, and imbalanced ratio, as numbers of samples in the majority class divided by numbers of samples in the minority class.

6.1.2. Models

The following list of models has been trained for every dataset/resampler combination. All models used `scikit-learn` implementation.

- k-neighbors classifier
    - k=3
- Support Vector Classifier (linear kernel)
    - C=0.025
    - max_iterations = 4000
- Support Vector Classifier (RBF kernel)
    - $\gamma = 2$
    - C=1
    - max_iterations = 4000
- Decision Tree classifier
    - max_depth = 5
- Gaussian Naive Bayes Classifier
- Random Forest Classifier
    - max_depth = 5
    - n. of estimators = 10
    - max_features = 1
- Multi layer perceptron
    - 1 hidden layer, 30 neurons
    - learning rate = adaptive
    - alpha = 1
    - max_iterations = 1000

- ADABoost classifier
- Quadratic Discriminant Analysis

All unspecified parameters were left at default values. For further details on models and default parameters, check `scikit-learn` API reference.

An additional model was tested, a Gaussian Process Classifier, but it was found to be too computational heavy for the large number of tests required.

6.1.3. Resamplers

We have tested the following already described resamplers:

- no resampling (original dataset)
- Random Over Sampler (ROS)
- Random Under Sampler (RUS)
- SMOTE
- ADASYN
- and, of course, ROSE.

6.1.4. Chosen metrics

The following metrics has been collected, for every model/dataset/resampler combination:

- precision
- recall
- F_1
- support
- AUC
- Matthews correlation coefficient

6.2. Results

We report the tables of Matthews Correlation Coefficient for every model.

KNeighborsClassifier

	SMOTE	ROS	ROSE	ADASYN	RUS
ecoli	0.59	0.59	0.59	0.59	0.59
optical_digits	0.81	0.81	0.81	0.81	0.81
satimage	0.54	0.54	0.54	0.54	0.54
pen_digits	0.96	0.96	0.96	0.96	0.96
abalone	0.2	0.2	0.2	0.2	0.2
sick_euthyroid	0.81	0.81	0.81	0.81	0.81
spectrometer	0.69	0.69	0.69	0.69	0.69
car_eval_34	0.88	0.88	0.88	0.88	0.88
isolet	0.61	0.61	0.61	0.61	0.61
us_crime	0.35	0.35	0.35	0.35	0.35
yeast_ml8	-0.0	-0.0	-0.0	-0.0	-0.0
scene	0.06	0.06	0.06	0.06	0.06
libras_move	0.2	0.2	0.2	0.2	0.2
thyroid_sick	0.87	0.87	0.87	0.87	0.87
coil_2000	0.1	0.1	0.1	0.1	0.1
arrhythmia	0.86	0.86	0.86	0.86	0.86
solar_flare_m0	0.11	0.11	0.11	0.11	0.11
oil	0.3	0.3	0.3	0.3	0.3
car_eval_4	0.94	0.94	0.94	0.94	0.94
wine_quality	0.32	0.32	0.32	0.32	0.32
letter_img	0.92	0.92	0.92	0.92	0.92
yeast_me2	0.22	0.22	0.22	0.22	0.22
webpage	0.65	0.65	0.65	0.65	0.65
ozone_level	0.03	0.03	0.03	0.03	0.03
mammography	0.6	0.6	0.6	0.6	0.6
protein_homo	0.74	0.74	0.74	0.74	0.74
abalone_19	-0.01	-0.01	-0.01	-0.01	-0.01

SVC(linear kernel)

	SMOTE	ROS	ROSE	ADASYN	RUS
ecoli	0.8	0.77	0.8	0.74	0.57
optical_digits	0.93	0.83	0.73	0.88	0.84
satimage	0.4	-0.15	0.21	0.22	0.29
pen_digits	0.03	0.06	0.34	0.16	0.35
abalone	0.52	0.5	0.48	0.49	0.41
sick_euthyroid	0.43	0.37	0.47	0.48	0.74
spectrometer	0.98	0.98	-0.15	0.98	0.91
car_eval_34	0.91	0.91	0.91	0.9	0.81
isolet	0.96	0.96	0.92	0.96	0.93
us_crime	0.78	0.73	0.7	0.75	0.74
yeast_ml8	0.28	0.21	0.27	0.32	0.0
scene	0.59	0.59	0.56	0.6	0.54
libras_move	0.94	0.88	0.8	0.77	0.0
thyroid_sick	0.63	0.66	0.52	0.64	0.7
coil_2000	0.27	0.27	0.17	0.33	0.35
arrhythmia	0.92	0.92	0.78	0.92	0.1
solar_flare_m0	0.6	0.47	0.52	0.54	0.33
oil	-0.12	0.15	0.31	-0.23	-0.23
car_eval_4	0.95	0.95	0.95	0.96	0.78
wine_quality	0.37	0.06	0.05	0.12	0.44
letter_img	0.92	0.91	0.83	0.86	0.89
yeast_me2	0.69	0.69	0.66	0.67	0.0
webpage	0.74	0.75	0.27	0.63	0.85
ozone_level	0.75	0.53	0.34	0.78	0.64
mammography	0.76	0.76	0.73	0.64	0.72
protein_homo	0.37	-0.06	-0.03	-0.02	0.51
abalone_19	0.39	0.39	0.36	0.38	0.48

SVC (RBF kernel)

	SMOTE	ROS	ROSE	ADASYN	RUS
ecoli	0.82	0.82	0.87	0.82	0.8
optical_digits	0.4	1.0	0.0	0.66	0.0
satimage	0.43	1.0	0.47	0.43	0.0
pen_digits	0.56	1.0	0.56	0.84	0.33
abalone	0.63	0.64	0.85	0.61	0.53
sick_euthyroid	0.62	1.0	0.12	0.56	0.16
spectrometer	0.0	1.0	0.0	0.0	0.0
car_eval_34	1.0	1.0	0.97	1.0	0.81
isolet	0.75	1.0	0.93	0.79	0.0
us_crime	0.99	1.0	0.69	0.98	0.58
yeast_ml8	1.0	1.0	0.91	1.0	0.2
scene	0.89	1.0	0.21	0.88	0.15
libras_move	1.0	1.0	0.91	1.0	0.84
thyroid_sick	0.63	1.0	0.05	0.63	0.0
coil_2000	0.77	0.98	0.64	0.75	0.1
arrhythmia	0.22	1.0	0.0	0.17	0.0
solar_flare_m0	0.91	0.85	0.84	0.86	0.07
oil	0.0	1.0	0.0	0.05	0.0
car_eval_4	1.0	1.0	0.86	1.0	0.14
wine_quality	0.81	1.0	0.44	0.83	0.25
letter_img	0.99	1.0	0.91	0.98	0.46
yeast_me2	0.75	0.78	0.7	0.75	0.64
webpage	0.99	1.0	0.73	0.99	0.34
ozone_level	0.44	1.0	0.0	0.41	0.0
mammography	0.91	0.92	0.93	0.87	0.75
protein_homo	0.01	1.0	0.01	0.01	0.17
abalone_19	0.8	0.76	0.96	0.8	0.67

DecisionTreeClassifier

	SMOTE	ROS	ROSE	ADASYN	RUS
ecoli	0.85	0.84	0.82	0.81	0.33
optical_digits	0.89	0.88	0.92	0.88	0.77
satimage	0.8	0.78	0.73	0.78	0.72
pen_digits	0.93	0.93	0.85	0.82	0.88
abalone	0.64	0.68	0.9	0.65	0.59
sick_euthyroid	0.95	0.95	0.92	0.93	0.88
spectrometer	0.88	0.94	0.83	0.88	0.63
car_eval_34	0.91	0.91	0.93	0.9	0.89
isolet	0.87	0.85	0.91	0.89	0.78
us_crime	0.81	0.82	0.9	0.82	0.6
yeast_ml8	0.45	0.43	0.53	0.44	0.15
scene	0.68	0.7	0.85	0.65	0.21
libras_move	0.99	0.96	0.82	0.95	0.71
thyroid_sick	0.97	0.98	0.97	0.98	0.89
coil_2000	0.57	0.41	0.94	0.66	0.24
arrhythmia	0.97	0.99	0.93	0.98	1.0
solar_flare_m0	0.78	0.6	0.94	0.74	0.09
oil	0.93	0.94	0.9	0.92	0.47
car_eval_4	0.95	0.95	0.95	0.96	0.88
wine_quality	0.59	0.67	0.65	0.58	0.27
letter_img	0.93	0.92	0.89	0.93	0.89
yeast_me2	0.83	0.87	0.95	0.87	0.55
webpage	0.65	0.6	0.98	0.63	0.58
ozone_level	0.83	0.89	0.84	0.73	0.46
mammography	0.84	0.85	0.89	0.79	0.82
protein_homo	0.88	0.88	0.98	0.84	0.81
abalone_19	0.82	0.88	0.99	0.8	-0.13

RandomForestClassifier

	SMOTE	ROS	ROSE	ADASYN	RUS
ecoli	0.59	0.59	0.59	0.59	0.59
optical_digits	0.8	0.8	0.8	0.8	0.8
satimage	0.54	0.54	0.54	0.54	0.54
pen_digits	0.95	0.95	0.95	0.95	0.95
abalone	0.21	0.21	0.21	0.21	0.21
sick_euthyroid	0.81	0.81	0.81	0.81	0.81
spectrometer	0.6	0.6	0.6	0.6	0.6
car_eval_34	0.9	0.9	0.9	0.9	0.9
isolet	0.6	0.6	0.6	0.6	0.6
us_crime	0.35	0.35	0.35	0.35	0.35
yeast_ml8	0.05	0.05	0.05	0.05	0.05
scene	0.08	0.08	0.08	0.08	0.08
libras_move	0.59	0.59	0.59	0.59	0.59
thyroid_sick	0.88	0.88	0.88	0.88	0.88
coil_2000	0.1	0.1	0.1	0.1	0.1
arrhythmia	0.66	0.66	0.66	0.66	0.66
solar_flare_m0	0.1	0.1	0.1	0.1	0.1
oil	0.37	0.37	0.37	0.37	0.37
car_eval_4	0.94	0.94	0.94	0.94	0.94
wine_quality	0.28	0.28	0.28	0.28	0.28
letter_img	0.92	0.92	0.92	0.92	0.92
yeast_me2	0.21	0.21	0.21	0.21	0.21
webpage	0.65	0.65	0.65	0.65	0.65
ozone_level	0.02	0.02	0.02	0.02	0.02
mammography	0.64	0.64	0.64	0.64	0.64
protein_homo	0.73	0.73	0.73	0.73	0.73
abalone_19	-0.01	-0.01	-0.01	-0.01	-0.01

MLPClassifier

	SMOTE	ROS	ROSE	ADASYN	RUS
ecoli	0.84	0.76	0.74	0.81	0.8
optical_digits	1.0	1.0	0.98	0.98	0.92
satimage	0.59	0.41	0.5	0.28	0.44
pen_digits	1.0	1.0	0.98	1.0	0.98
abalone	0.54	0.53	0.66	0.53	0.53
sick_euthyroid	0.88	0.85	0.83	0.81	0.71
spectrometer	0.94	0.93	0.9	0.91	0.45
car_eval_34	0.98	0.98	0.94	0.98	0.58
isolet	0.99	0.97	0.97	0.97	0.93
us_crime	0.81	0.79	0.83	0.79	0.76
yeast_ml8	0.66	0.68	0.59	0.62	0.16
scene	0.84	0.82	0.77	0.86	0.62
libras_move	0.99	0.99	0.91	0.99	0.84
thyroid_sick	0.89	0.88	0.86	0.83	0.79
coil_2000	0.71	0.62	0.93	0.72	0.28
arrhythmia	0.92	0.93	0.77	0.97	-0.02
solar_flare_m0	0.8	0.8	0.92	0.76	0.32
oil	0.53	0.49	0.74	0.51	0.74
car_eval_4	0.97	0.97	0.97	0.98	0.88
wine_quality	0.54	0.51	0.66	0.55	0.35
letter_img	0.98	0.98	0.93	0.98	0.89
yeast_me2	0.78	0.73	0.65	0.7	0.85
webpage	0.91	0.9	0.94	0.9	0.89
ozone_level	0.2	0.62	0.56	0.55	-0.5
mammography	0.83	0.81	0.86	0.74	0.8
protein_homo	0.96	0.97	0.99	0.97	0.82
abalone_19	0.65	0.62	0.85	0.68	0.52

AdaBoostClassifier

	SMOTE	ROS	ROSE	ADASYN	RUS
ecoli	0.8	0.91	0.84	0.85	0.57
optical_digits	0.96	0.92	0.98	0.96	0.85
satimage	0.84	0.8	0.76	0.81	0.72
pen_digits	0.96	0.95	0.96	0.96	0.92
abalone	0.63	0.65	0.9	0.64	0.57
sick_euthyroid	0.94	0.92	0.96	0.92	0.85
spectrometer	0.97	0.98	0.92	0.95	0.91
car_eval_34	0.98	0.99	0.97	0.99	0.97
isolet	0.95	0.95	0.95	0.95	0.87
us_crime	0.87	0.87	0.9	0.83	0.71
yeast_ml8	0.58	0.64	0.76	0.59	0.17
scene	0.74	0.76	0.88	0.71	0.37
libras_move	0.99	0.99	0.91	0.98	0.6
thyroid_sick	0.96	0.94	0.97	0.96	0.84
coil_2000	0.81	0.42	0.94	0.8	0.21
arrhythmia	0.97	0.99	0.98	0.96	1.0
solar_flare_m0	0.81	0.51	0.95	0.8	0.27
oil	0.95	0.97	0.92	0.96	0.63
car_eval_4	0.98	0.98	0.98	0.99	0.88
wine_quality	0.68	0.68	0.75	0.67	0.33
letter_img	0.98	0.96	0.97	0.99	0.95
yeast_me2	0.84	0.92	0.96	0.88	0.7
webpage	0.91	0.84	0.98	0.89	0.86
ozone_level	0.91	0.96	0.93	0.9	0.65
mammography	0.81	0.84	0.94	0.74	0.78
protein_homo	0.94	0.93	0.99	0.93	0.9
abalone_19	0.78	0.9	0.99	0.78	0.52

GaussianNB

	SMOTE	ROS	ROSE	ADASYN	RUS
ecoli	0.64	0.64	0.64	0.64	0.64
optical_digits	0.81	0.81	0.81	0.81	0.81
satimage	0.51	0.51	0.51	0.51	0.51
pen_digits	0.95	0.95	0.95	0.95	0.95
abalone	0.14	0.14	0.14	0.14	0.14
sick_euthyroid	0.79	0.79	0.79	0.79	0.79
spectrometer	0.64	0.64	0.64	0.64	0.64
car_eval_34	0.88	0.88	0.88	0.88	0.88
isolet	0.6	0.6	0.6	0.6	0.6
us_crime	0.34	0.34	0.34	0.34	0.34
yeast_ml8	0.01	0.01	0.01	0.01	0.01
scene	0.08	0.08	0.08	0.08	0.08
libras_move	0.46	0.46	0.46	0.46	0.46
thyroid_sick	0.88	0.88	0.88	0.88	0.88
coil_2000	0.09	0.09	0.09	0.09	0.09
arrhythmia	0.76	0.76	0.76	0.76	0.76
solar_flare_m0	0.16	0.16	0.16	0.16	0.16
oil	0.31	0.31	0.31	0.31	0.31
car_eval_4	0.94	0.94	0.94	0.94	0.94
wine_quality	0.24	0.24	0.24	0.24	0.24
letter_img	0.92	0.92	0.92	0.92	0.92
yeast_me2	0.21	0.21	0.21	0.21	0.21
webpage	0.64	0.64	0.64	0.64	0.64
ozone_level	0.03	0.03	0.03	0.03	0.03
mammography	0.63	0.63	0.63	0.63	0.63
protein_homo	0.72	0.72	0.72	0.72	0.72
abalone_19	-0.01	-0.01	-0.01	-0.01	-0.01

QuadraticDiscriminantAnalysis

	SMOTE	ROS	ROSE	ADASYN	RUS
ecoli	0.65	0.2	-0.1	0.26	0.47
optical_digits	0.41	0.36	0.26	0.45	0.43
satimage	0.67	0.63	0.89	0.67	0.6
pen_digits	0.96	0.97	0.95	0.92	0.97
abalone	0.61	0.51	0.9	0.57	0.52
sick_euthyroid	0.42	0.41	0.91	0.42	0.26
spectrometer	1.0	1.0	0.91	0.99	0.23
car_eval_34	0.0	0.0	0.0	0.0	0.0
isolet	0.99	1.0	0.91	0.96	0.17
us_crime	0.94	0.94	0.91	0.97	0.52
yeast_ml8	0.92	0.92	0.9	0.91	0.24
scene	1.0	1.0	0.84	1.0	0.11
libras_move	1.0	1.0	0.94	1.0	0.84
thyroid_sick	0.31	0.26	0.94	0.28	0.08
coil_2000	0.17	0.15	0.07	0.17	0.22
arrhythmia	1.0	1.0	0.42	1.0	0.3
solar_flare_m0	0.59	0.43	0.95	0.49	0.08
oil	1.0	1.0	0.93	1.0	0.14
car_eval_4	0.0	0.0	0.0	0.0	0.0
wine_quality	0.54	0.51	0.75	0.54	0.34
letter_img	0.93	0.91	0.89	0.94	0.87
yeast_me2	0.08	0.06	0.01	0.15	-0.31
webpage	0.68	0.57	0.98	0.61	0.76
ozone_level	1.0	1.0	0.85	1.0	0.23
mammography	0.7	0.68	0.78	0.57	0.64
protein_homo	0.91	0.88	0.99	0.88	0.9
abalone_19	0.72	0.67	0.99	0.69	0.63

Other metrics are reported in Appendix _.

7. ORBIS Dataset: a real-world application

Benchmark test datasets usually are convenient. The data are clean, there is an actual relationship, all the variables are used.

We decided to test ROSE in a real world problem belonging to a field considered difficult to handle: econometrics.

7.1. Problem description

In this particular project the Client asked the following question:

> Is it possible to foresee which firm have potential for becoming an High Growth Firm, given their economic status at the first year of activity?

7.2. Dataset description

The Client provided a dataset that is a subset of ORBIS database, a collection of information on listed company across the globe, curated by Bureau Van Djik (henceforth BvD), a Moody's Analytics controlled private society. BvD collects information from about 375 millions of public and private firms in a standardized way, allowing for comparison and analytics. ORBIS data comes from more than 160 providers and hundreds of internal sources. The firm activity itself revolves about the reconstruction of proprietary assets and recognition of effective owners, providing firm structure hierarchy diagrams to rebuild dependencies among groups and controlled societies. Those data can be used to find informations about a firm, can be filtered to find firms that satisfy certain criteria, analyze peer groups, retrieve market informations about competitors and potential collaborations, and analyze stakeholders interdependence and financial strength.

ORBIS is used by enterprise, governments and public administrations, academic entities, financial institutes and professional studies, and is focused on efficiency aimed at decisional processes. Different targets can be optimized by ORBIS data:

- Credit risk
- Compliance and financial frauds
- Supply chain risk
- Transfer pricing
- Commercial development
- M&A and corporate finance
- Master Data Management projects

We had no direct source to the original data, that were provided as a CSV archived version with the data of 115840 firms. Being expensive data, we are not allowed to publish them for repeatability, but we included a MD5 checksum of the provided file.

```
HGFfinal.merge.csv
Size: 44,7 MB (44745032 bytes)
MD5 checksum: 420d345c68dc3998b8403ab07d0fecf8
```

Our datasets encompassed 3 different categories of information:

- Company information, like name, location, contacts, sector, NACE code, etc.
- Economic information:
    - Balance sheet
    - Profit and Loss (P&L) statement
- BvD evaluations, like trust level, default chances, and independence score.

For most of them, where numeric values were not available, qualitative informations where provided. Still, a lot of data were missing.

7.2.1. Exploratory Data Analysis

We report here the procedure of data import and cleaning that has been done before performing any other test.

7.2.1.1. Data import

Data has been imported in a Pandas DataFrame, and analyzed in a dedicated Python 3.6 `conda` environment in a Jupyter Notebook on a local Linux machine.

Numeric data has been parsed to `int` and `float` data types accordingly, while ordered categories, like `BvD.Independence.Indicator` has been cast in `pd.api.types.CategoricalDtype()` format.

7.2.1.2. Variables Description

7.2.1.2.1. BvD.ID.number

This is our dataset primary key. Unique (cardinality = n). It is composed by 2 letters and 8÷12 digits. The two letters appear to be the `Country.ISO.Code`.

```
CategoricalIndex(['IS4203100990', 'GR997722505', 'BG201066368', 'BG201251947',
                  'BG201331418', 'IS4102101180', 'BG201222746', 'BG201124711',
                  'BG201005899', 'BA4281217330002',
                  ...
                  'SK45369747', 'SK45284300', 'SK45349304', 'SK45457824',
                  'SK45432112', 'SK45480371', 'SK45452245', 'SK45407851',
                  'SK45430268', 'SK45418527'],
                 categories=['AT9010104250', 'AT9030242392', 'AT9070278738',
'AT9070279036', 'AT9090150166', 'AT9110712698', 'AT9110713446', 'AT9110713447',
...], ordered=False, name='BvD.ID.number', dtype='category', length=115840)
```

7.2.1.2.2. Company.name

`dtype: string`

Contains the firm name. All caps, sometimes includes firm's juridic form.

```
BvD.ID.number
BG201056516               AW TRONICS OOD
IT02637960606       HOME DESIGN - S.R.L.
FR519561369                FINANCIERE HL
SK45371296          EU MANAGEMENT, S.R.O.
FR529210692      TOTAL E&P WELL RESPONSE
Name: Company.name, dtype: string
```

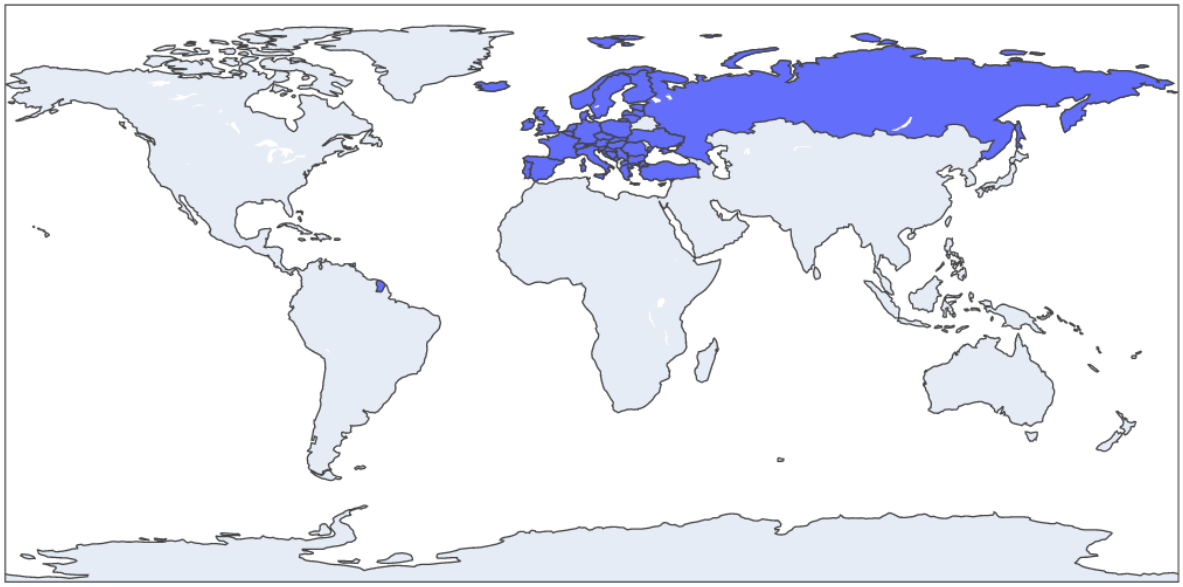
7.2.1.2.3. Country.ISO.Code

`dtype: string`

Contains a two-letter ISO 3166 alpha-2 code from 38 countries.

```
data['Country.ISO.Code'].unique()
```

```
['IS', 'GR', 'BG', 'BA', 'BE', 'IE', 'CY', 'DE', 'AT', 'DK', 'GB', 'CH', 'CZ',
 'EE', 'ES', 'FI', 'FR', 'HR', 'HU', 'IT', 'RS', 'PL', 'UA', 'ME', 'NL', 'LU',
 'MT', 'MK', 'TR', 'LT', 'LV', 'NO', 'PT', 'RO', 'RU', 'SE', 'SI', 'SK']
 Length: 38, dtype: string
```



Involved countries

```
Country.ISO.Code  AT  BA   BE   BG  CH  CY    CZ  DE  DK    EE  ...   PL  \
HGF                                                              ...
0                 19  97  190  626   2   3  1985  28   1  2489  ...  446
1                  0  14   11   31   0   0    60   0   0   114  ...   46

Country.ISO.Code    PT     RO   RS     RU    SE    SI    SK  TR   UA
HGF
0                 6535  15500  361  11731  2512  1693  3497   5  374
1                  265    758   20    192    76    92   106   3    6

[2 rows x 38 columns]

Chi² =  4.79e+02
p    =   3.4e-78
degrees of freedom = 37
```

7.2.1.2.4. Postcode

Postcode of the firm. Refers to a different encoding for each country.

```
BvD.ID.number
FR522454743     81640
RU66322917     129085
EE11951827      10151
FR519806830     18700
RO27831630        nan
Name: Postcode, dtype: category
Categories (29914, object): [00-024, 00-042, 00-066, 00-102, ..., YO19 6ED, YO26
4GB, YO41 5NS, nan]
```

Present only in 107506 rows. Missing in 8334 rows.

7.2.1.2.5. City

All caps name of the firm's city. Missing in 88 entries.

7.2.1.2.6. NACE codes

Statistical Classification of Economic Activities in the European Community code, known as NACE, is the industry standard classification of European Union. Established by Regulation (EC) No 1893/2006, it uses four hierarchical levels:

- Level 1: 21 sections identified by alphabetical letters A to U;
- Level 2: 88 divisions identified by two-digit numerical codes (01 to 99);
- Level 3: 272 groups identified by three-digit numerical codes (01.1 to 99.0);
- Level 4: 615 classes identified by four-digit numerical codes (01.11 to 99.00).

The first four digits of the code, which is the first four levels of the classification system, are the same in all European countries. National implementations may introduce additional levels. The fifth digit might vary from country to country and further digits are sometimes placed by suppliers of databases.

links: Reference to all NACE codes , Wikipedia: NACE codes.

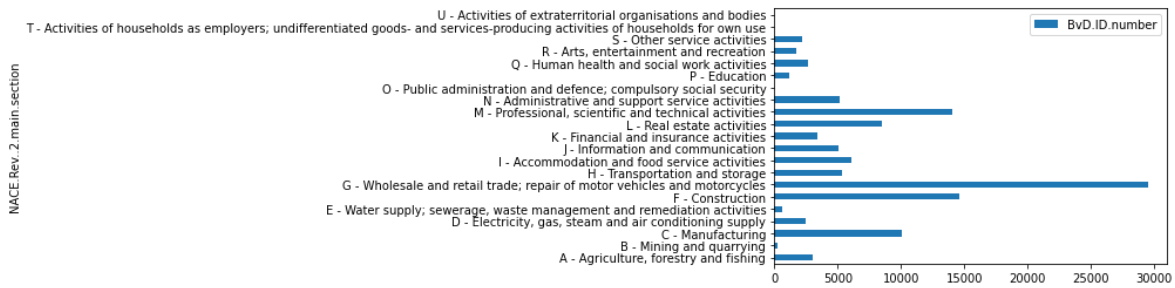
7.2.1.2.7. NACE.Rev..2.main.section

Level 1 NACE code. A letter, and the sector description.

```
BvD.ID.number
FR521201111                                        F - Construction
FR523714624      G - Wholesale and retail trade; repair of moto...
ESB85891794                            S - Other service activities
IT03972880235        I - Accommodation and food service activities
PT509412866      M - Professional, scientific and technical act...
Name: NACE.Rev..2.main.section, dtype: category
```

All 21 sections are being represented.

```
pd.DataFrame(data['NACE.Rev..2.Core.code..4.digits.']).reset_index().groupby('NACE.Rev..2.main.section').count().plot.barh()
```



number of entries per section

```
NACE      A    B     C     D    E      F      G     H     I     J  ...     L  \
HGF                                                              ...
0      2938  213  9621  2429  593  14212  28334  5038  5942  4779  ...  8340
1        99    8   444    33   43    385   1192   294   154   266  ...   146

NACE      M     N   O     P     Q     R     S  T  U
HGF
```

```
0     13611  4899  16  1116  2504  1646  2151  2  1
1       460   244   0    66   129    75    64  0  0

[2 rows x 21 columns]

Chi² =  3.96e+02
p    =   1.6e-71
degrees of freedom = 20
```

7.2.1.2.8. NACE.Rev..2.Core.code..4.digits.

4 digit NACE code. 729 different categories.

```
BvD.ID.number
NO995138697      4110
PT509493599      4339
RO27726219       4711
FR520957143       161
IT03102890831    4120
Name: NACE.Rev..2.Core.code..4.digits., dtype: category
Categories (729, int64): [100, 110, 111, 112, ..., 9609, 9700, 9810, 9900]
```

```
NACE  100   110   111   112   113   115   119   120   121   122   ...  9529  \
HGF                                                               ...
0      18    60   571     8   227     1    64    17   117     1  ...    34
1       2     1    20     0     5     0     1     1     7     0  ...     1

NACE  9600  9601  9602  9603  9604  9609  9700  9810  9900
HGF
0        4   131   957   129   186   351     1     1     1
1        0     3    20     4     6    16     0     0     0

[2 rows x 729 columns]

Chi² =  1.69e+03
p    =   9.4e-78
degrees of freedom = 728
```

7.2.1.2.9. NACE.Rev..2.Primary.code.s.

4 digit NACE code. Similar to the former, it contains duplicates. 729 different categories.

```
BvD.ID.number
RU65230449      4633
NO995818558     4759
SE5568024276    7311
NO995182165     3312
RO27703886      4613
Name: NACE.Rev..2.Primary.code.s., dtype: category
Categories (729, int64): [100, 110, 111, 112, ..., 9609, 9700, 9810, 9900]
```

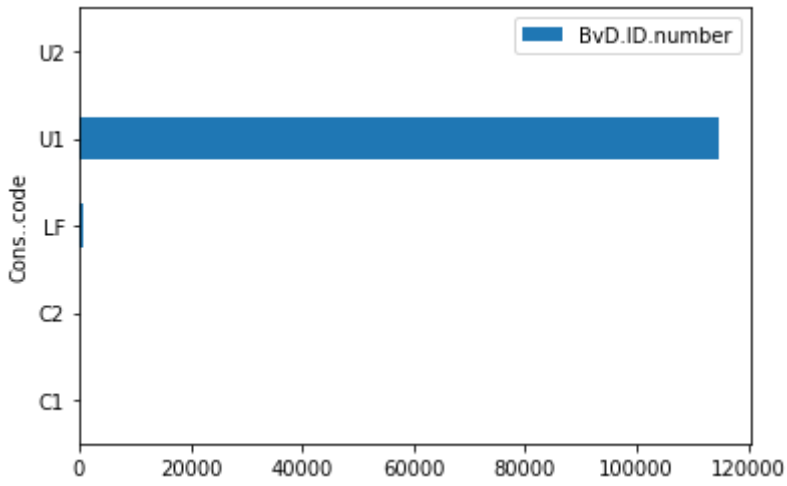
7.2.1.2.10. Cons..code

Bankscape Consolidation Code. It indicates the level of consolidation for the different financial statements

- **C1**: statement of a mother company integrating the statements of its controlled subsidiaries or branches with no unconsolidated companion,

- **C2**: statement of a mother company integrating the statements of its controlled subsidiaries or branches with an unconsolidated companion,
- **U1**: statement not integrating the statements of the possible controlled subsidiaries or branches of the concerned company with no consolidated companion.
- **U2**: statement not integrating the statements of the possible controlled subsidiaries or branches of the concerned company with an consolidated companion.
- **LF**: limited financials: information based on rounded figures officially available, sometimes collected from other directories or websites.

```
BvD.ID.number
RO26440005       U1
IT03232890982    U1
IT11001531000    U1
RU67267304       LF
PT509588980      U1
Name: Cons..code, dtype: category
Categories (5, object): [C1, C2, LF, U1, U2]

```



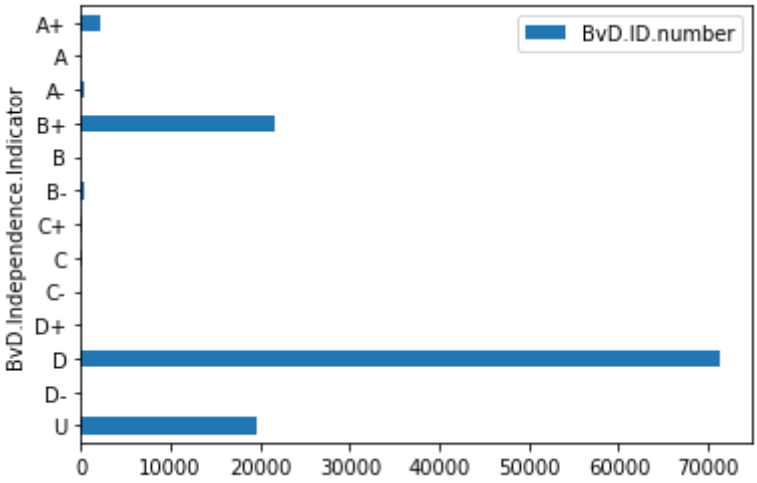

```
CONS    C1   C2   LF      U1   U2
HGF
0      116  103  784  110504  140
1        7    3    7    4170    6

Chi² =      18.9
p    =   0.00084
degrees of freedom = 4
```

7.2.1.2.11. BvD.Independence.Indicator

It characterizes the degree of independence of a company with regard to its shareholders. It has been mapped to an ordered category, with null value (U) being set at the lowest value.

links: Variable description

```
BvD.ID.number
RO26798381       D
RU67068071       D
FR527515381      U
RU64795818       D
IT06649101216    D
Name: BvD.Independence.Indicator, dtype: category
Categories (13, object): [U < D- < D < D+ ... B+ < A- < A < A+]
```



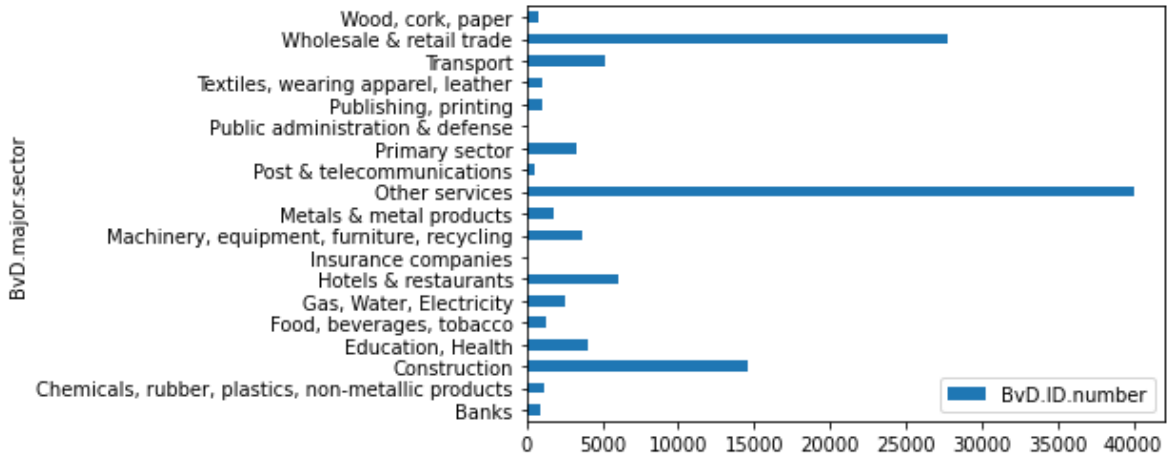

```
INDEP  -   A    A+   A-  B     B+   B-    C   C+      D      U
HGF
0      3  20  2170  257  1  20777  346  104  197  68757  19015
1      0   1    63   21  0    837   16   13    8   2588    646

Chi² =      46.4
p    =   1.2e-06
degrees of freedom = 10
```

7.2.1.2.12. BvD.major.sector

A different sector encoding from BvD. It encompass 19 categories.

```
BvD.ID.number
RO26474245                                          Other services
HR76526891156    Chemicals, rubber, plastics, non-metallic prod...
RU68874348                     Textiles, wearing apparel, leather
CZ28117956                                          Primary sector
RO15587044                                          Other services
Name: BvD.major.sector, dtype: category
Categories (19, object): [Banks, Chemicals, rubber, plastics, non-metallic
prod..., Construction, Education, Health, ..., Textiles, wearing apparel,
leather, Transport, Wholesale & retail trade, Wood, cork, paper]
```

There are no missing values.

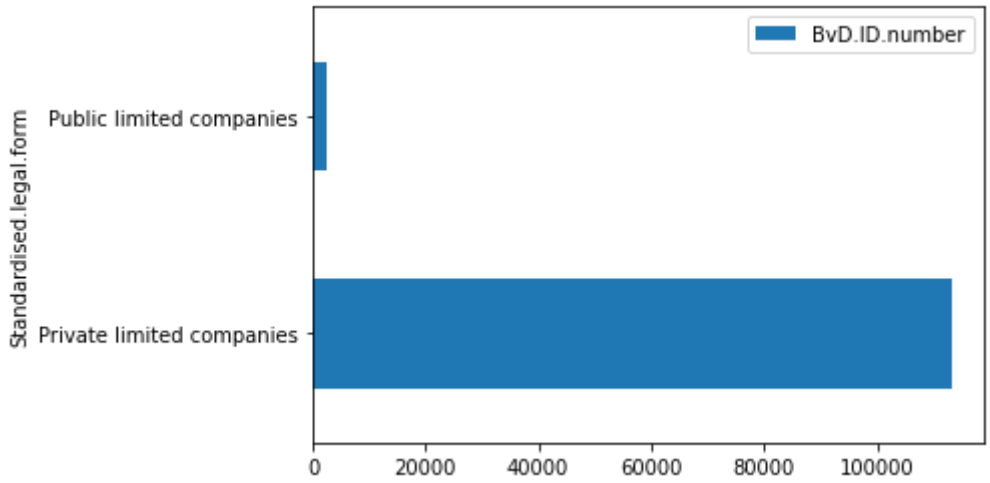



```
SECT  Banks  Chemicals,  Constructi  Education,  Food, beve  Gas, Water  \
HGF
0       947        1156       14212        3809        1252        2494
1        22          58         385         211          74          36

SECT  Hotels & r  Insurance   Machinery,  Metals & m  Other serv  Post & tel  \
HGF
0           5942          19        3466        1712       38654         447
1            154           2         149          64        1340          39

SECT  Primary se  Public adm  Publishing  Textiles,   Transport  Wholesale   \
HGF
0           3148          17        1042        1000       4919       26666
1            107           0          43          52        286        1137

SECT  Wood, cork
HGF
0            745
1             34

Chi² =   2.7e+02
p    =   7.4e-47
degrees of freedom = 18
```

7.2.1.2.13. Standardised.legal.form

Two level factor, stating if the company is public or private.

```
BvD.ID.number
ESB85955946    Private limited companies
FR519336481    Private limited companies
PT509284035     Public limited companies
EE11920653     Private limited companies
NO995237563    Private limited companies
Name: Standardised.legal.form, dtype: category
Categories (2, object): [Private limited companies, Public limited companies]
```




```
FORM  Private  Public
HGF
0      109204     2443
1        4056      137

Chi² =      21.6
p    =   3.3e-06
degrees of freedom = 1
```

7.2.1.2.14. Category.of.the.company

4 level factor stating the dimension of the company. It was impossible to retrieve information about objective inclusion criteria anywhere. Different legislations use different criteria, and despite their similarity, this does not allow a unequivocal definition.

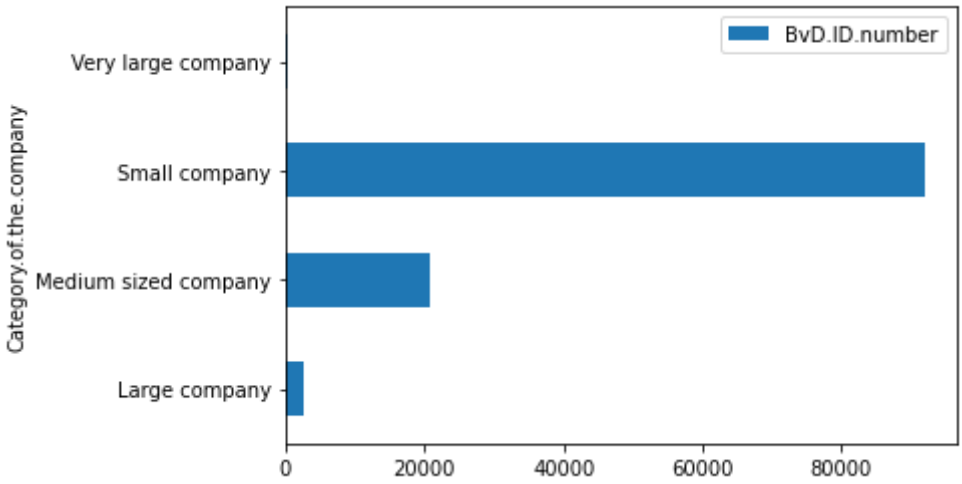
To give an approximation of this classification, we will report Australian's definition of large company. A company is considered large if it satisfies at least two of the following criteria:

- the consolidated revenue for the financial year of the company and the companies it controls is AU$50 millions or more,
- the value of the consolidated gross assets at the end of the financial year of the company and any entities it controls is AU$25 millions or more, and
- the company and any entity it controls have 100 or more employees at the end of the fiscal year.

European Union EUROSTAT website reports a different classification, based only on employees:

number of employees	**enterprise size**
< 10	micro enterprise
$10 \leq e < 50$	small enterprise
$50 \leq e < 250$	medium-sized enterprise
$e \geq 250$	large enterprise

```
BvD.ID.number
RO26541921             Small company
RO26543973             Small company
RO27249764      Medium sized company
PT509553923            Small company
SE5568008675           Small company
Name: Category.of.the.company, dtype: category
Categories (4, object): [Large company, Medium sized company, Small company,
Very large company]
```




```
CAT  Large compan  Medium sized  Small compan  Very large c
HGF
0             2485         19276         89542           344
1              267          1522          2373            31

Chi² =  1.41e+03
p    =  2.3e-306
degrees of freedom = 3
```

7.2.1.2.15. No.of.companies.in.corporate.group

Number of companies in the corporate group. The largest part of entries has 0 companies in the group (assuming: no group).

outliers:

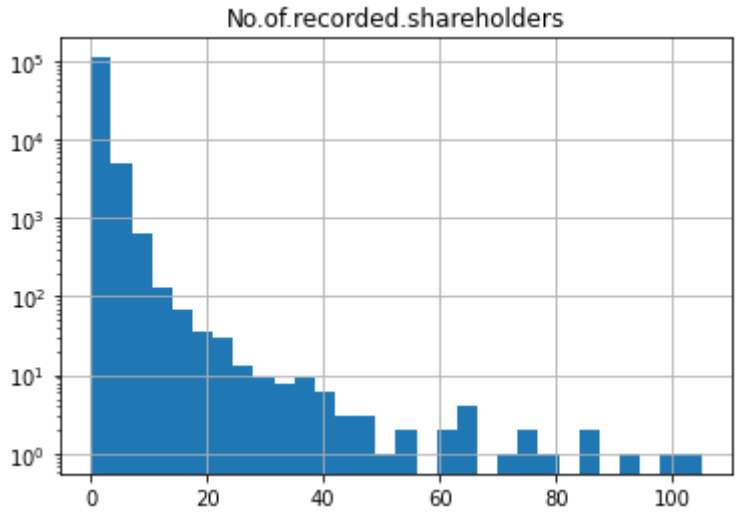
```
BvD.ID.number
IE488184         SKY HIGH III LEASING DESIGNATED ACTIVITY COMPANY
IT07063570969                                      HB SERVIZI S.R.L
IT07182390968                        POLIAMBULATORIO BICOCCA S.R.L.
NO995590271                                          ELKEM RANA AS
RO1590899                          ADAMA AGRICULTURAL SOLUTIONS SRL
RO25221180                                        EDPR ROMANIA SRL
Name: Company.name, dtype: string
```

```
HGF vs non-HGF for No.of.companies.in.corporate.group
Welch's t-test statistic = -0.7536
p-value = 0.4511
```

```
Optimization terminated successfully.
         Current function value: 0.155660
         Iterations 7
                           Logit Regression Results
==============================================================================
Dep. Variable:                    HGF   No. Observations:               115840
Model:                          Logit   Df Residuals:                   115838
Method:                           MLE   Df Model:                            1
Date:                Mon, 29 Jun 2020   Pseudo R-squ.:                1.205e-05
Time:                        15:23:49   Log-Likelihood:                -18032.
converged:                       True   LL-Null:                       -18032.
Covariance Type:            nonrobust   LLR p-value:                    0.5098
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -3.2825      0.016   -208.368      0.000      -3.313      -3.252
CORPGRP     2.408e-05   3.39e-05      0.710      0.478   -4.24e-05    9.06e-05
==============================================================================
```

7.2.1.2.16. No.of.recorded.shareholders

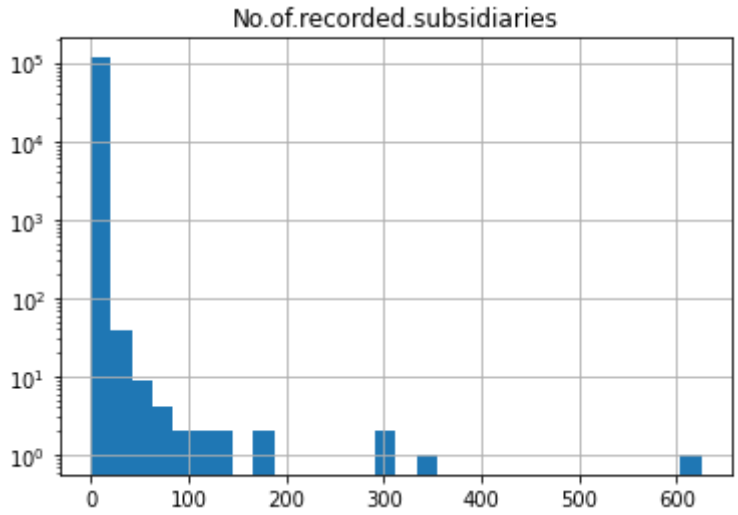
Numbers of enterprise recorded stakeholders.




```
HGF vs non-HGF for No.of.recorded.shareholders
Welch's t-test statistic = -4.809
p-value = 1.571e-06

Optimization terminated successfully.
         Current function value: 0.155556
         Iterations 7
                           Logit Regression Results
==============================================================================
Dep. Variable:                    HGF   No. Observations:               115840
```

```
Model:                          Logit   Df Residuals:                   115838
Method:                           MLE   Df Model:                            1
Date:                Mon, 29 Jun 2020   Pseudo R-squ.:               0.0006832
Time:                        15:25:05   Log-Likelihood:                -18020.
converged:                       True   LL-Null:                       -18032.
Covariance Type:            nonrobust   LLR p-value:                 6.913e-07
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -3.3319      0.018   -183.049      0.000      -3.368      -3.296
SHA            0.0315      0.006      5.698      0.000       0.021       0.042
==============================================================================

```

7.2.1.2.17. No.of.recorded.subsidiaries



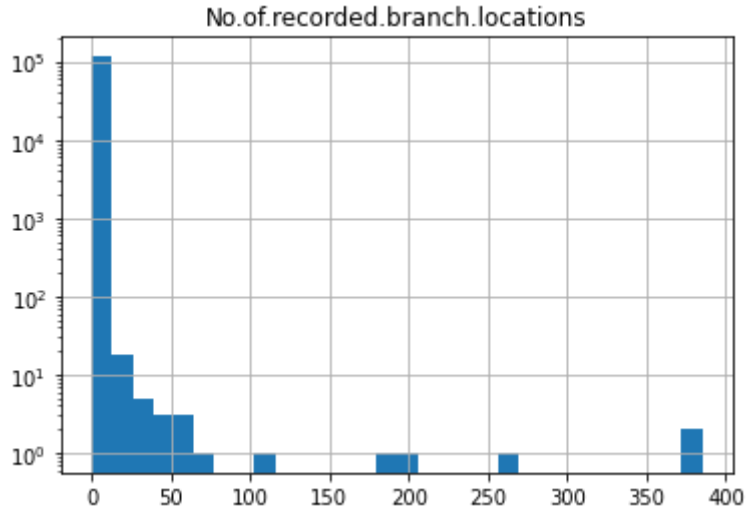
outlier:

```
BvD.ID.number
RS20661283    AKCIONARSKI FOND
Name: Company.name, dtype: string
```

```
HGF vs non-HGF for No.of.recorded.subsidiaries
Welch's t-test statistic = -3.294
p-value = 0.0009946
```

```
                           Logit Regression Results
==============================================================================
Dep. Variable:                    HGF   No. Observations:                115840
Model:                          Logit   Df Residuals:                    115838
Method:                           MLE   Df Model:                            1
Date:                Mon, 29 Jun 2020   Pseudo R-squ.:               8.499e-05
Time:                        15:25:58   Log-Likelihood:                -18030.
converged:                       True   LL-Null:                       -18032.
Covariance Type:            nonrobust   LLR p-value:                   0.08000
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -3.2831      0.016   -208.501      0.000      -3.314      -3.252
SUB            0.0052      0.003      2.045      0.041       0.000       0.010
==============================================================================
```

7.2.1.2.18. No.of.recorded.branch.locations

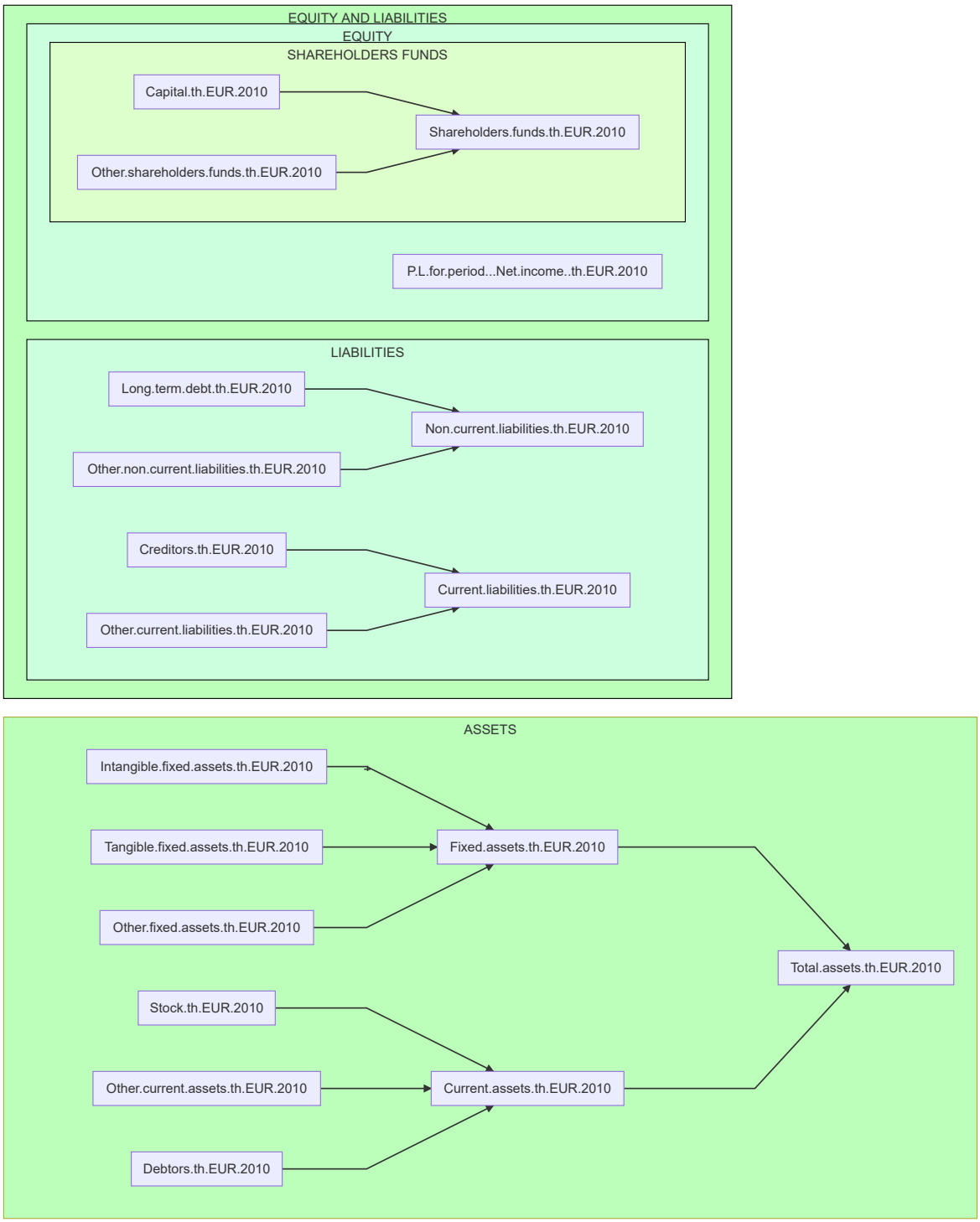


outliers:

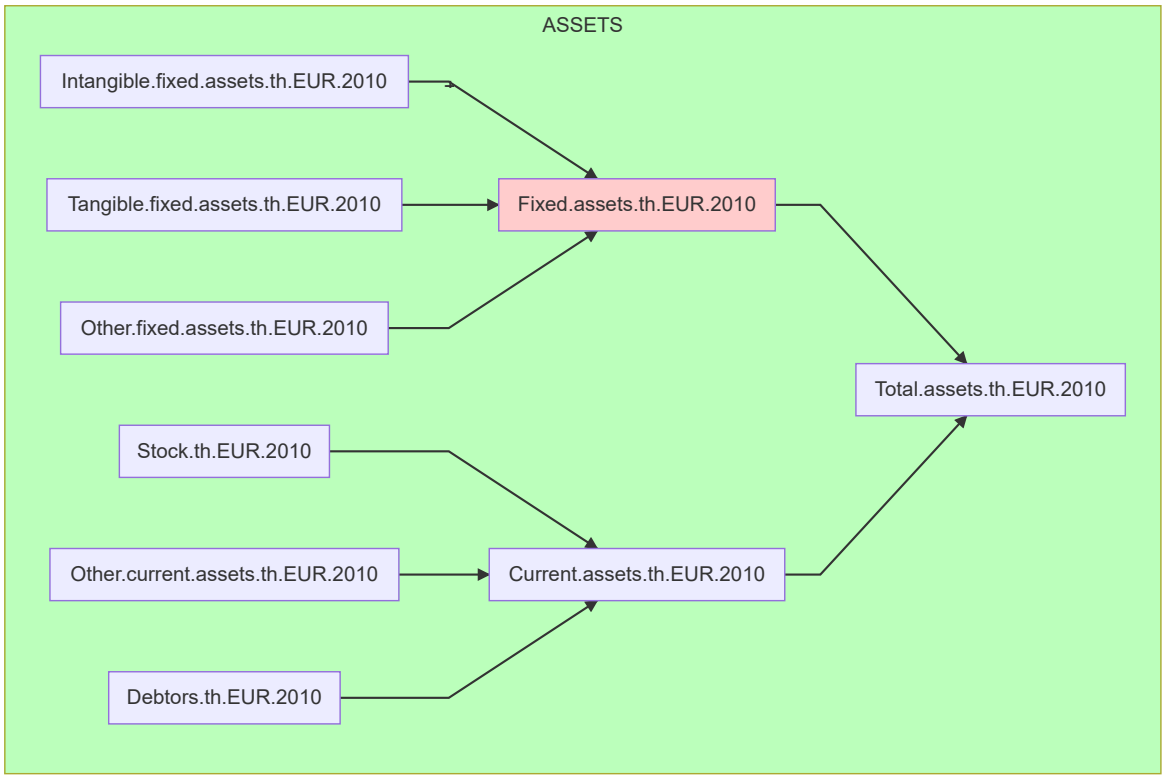
```
BvD.ID.number
FR528648892                          CHAUSSON MATERIAUX
FR524237351    S A S LOT AGRICULTURE ET ENERGIE SOLAIRE
Name: Company.name, dtype: string
```

```
HGF vs non-HGF for No.of.recorded.branch.locations
Welch's t-test statistic = -3.193
p-value = 0.001417


Optimization terminated successfully.
         Current function value: 0.155622
         Iterations 7
                           Logit Regression Results
==============================================================================
Dep. Variable:                    HGF   No. Observations:               115840
Model:                          Logit   Df Residuals:                   115838
Method:                           MLE   Df Model:                            1
Date:                Mon, 29 Jun 2020   Pseudo R-squ.:              0.0002588
Time:                        15:26:33   Log-Likelihood:                -18027.
converged:                       True   LL-Null:                       -18032.
Covariance Type:            nonrobust   LLR p-value:                  0.002250
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -3.2833      0.016   -208.582      0.000      -3.314      -3.252
BRA            0.0107      0.003      3.303      0.001       0.004       0.017
==============================================================================
```

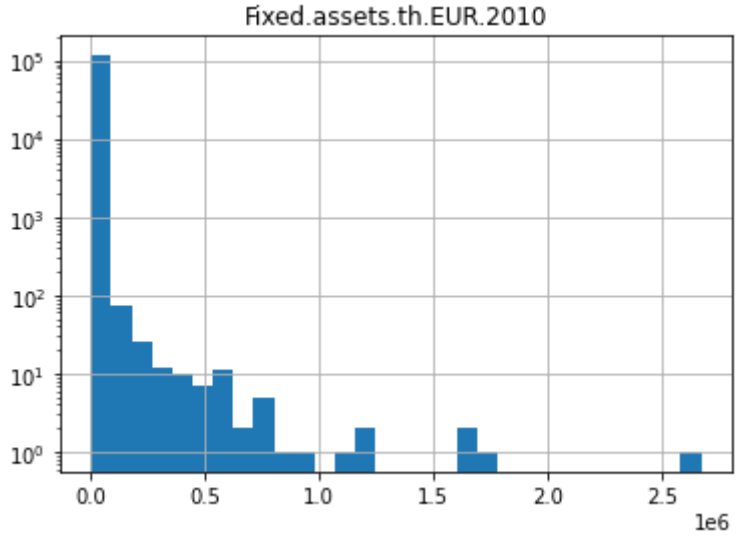
Balance sheet structure:



7.2.1.2.19. Fixed.assets.th.EUR.2010



A fixed asset is a long-term tangible piece of property or equipment that a firm owns and uses in its operations to generate income. Fixed assets are not expected to be consumed or converted into cash within a year. Fixed assets most commonly appear on the balance sheet as property, plant, and equipment (PP&E).

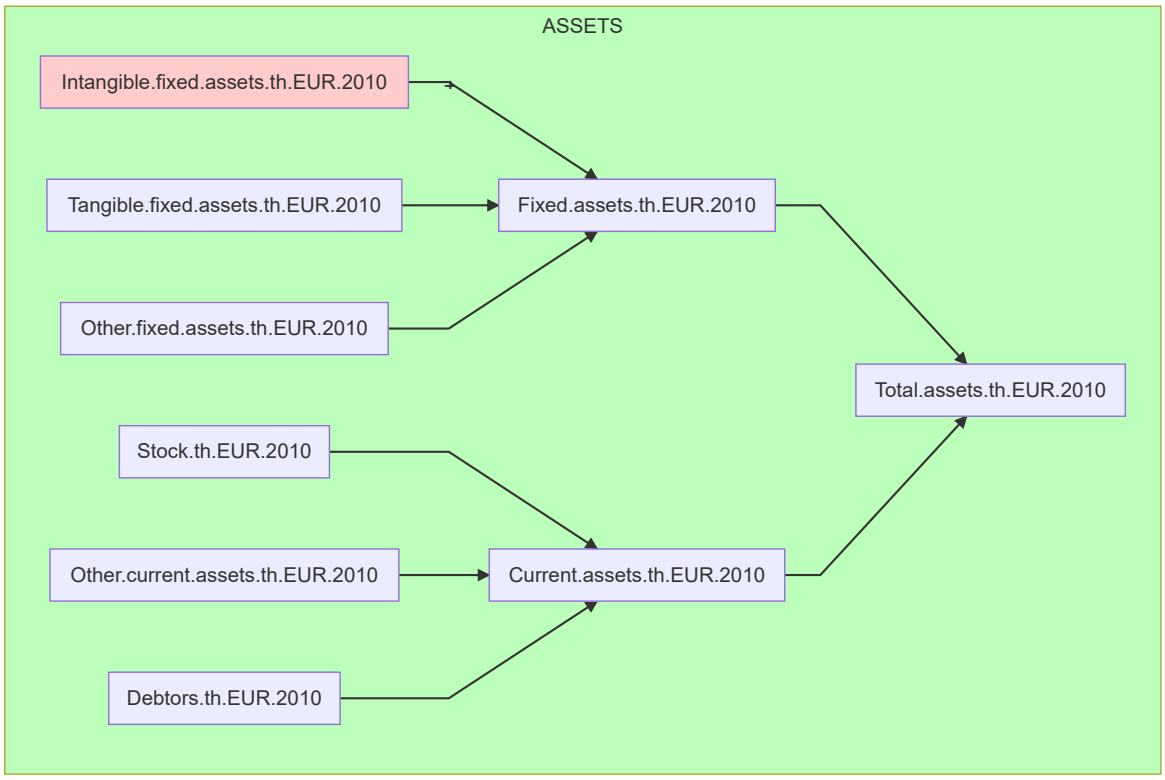



```
outlier:
BvD.ID.number
IE486605    SAP IRELAND US-FINANCIAL SERVICES DESIGNATED A...
Name: Company.name, dtype: string
```

```
HGF vs non-HGF for Fixed.assets.th.EUR.2010
Welch's t-test statistic = 3.679
p-value = 0.0002362

Optimization terminated successfully.
         Current function value: 0.155638
         Iterations 9
                           Logit Regression Results
```

```
==============================================================================
Dep. Variable:                    HGF   No. Observations:               115840
Model:                          Logit   Df Residuals:                   115838
Method:                           MLE   Df Model:                            1
Date:                Mon, 29 Jun 2020   Pseudo R-squ.:               0.0001559
Time:                        15:27:08   Log-Likelihood:                -18029.
converged:                       True   LL-Null:                       -18032.
Covariance Type:            nonrobust   LLR p-value:                   0.01773
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -3.2793      0.016   -207.847      0.000      -3.310      -3.248
FX         -6.137e-06   3.82e-06     -1.607      0.108   -1.36e-05    1.35e-06
==============================================================================
```

7.2.1.2.20. Intangible.fixed.assets.th.EUR.2010

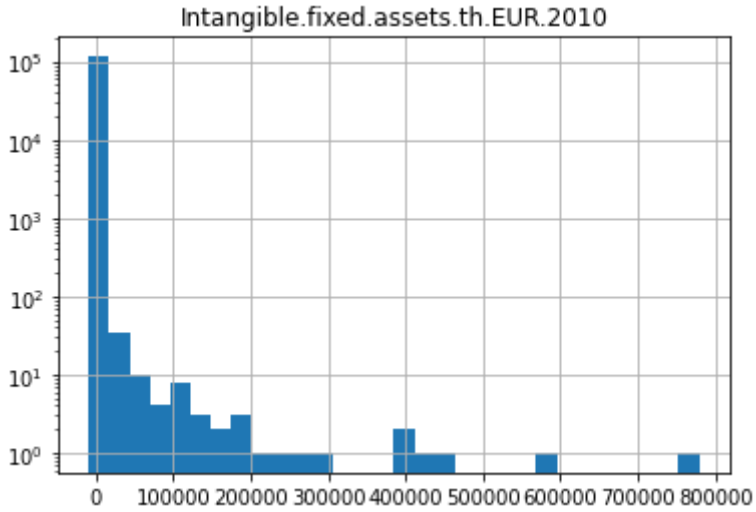


Intangible assets include operational assets that lack physical substance. For example, goodwill is a fixed asset, as are patents, copyrights, trademarks and franchises. A company's intangible assets are often not reported on a company's financial statements, or they may be reported at significantly less than their actual value. This is because assets are accounted for at their historical cost.

Unlike tangible fixed assets such as a building or machinery, intangibles are often developed internally without any direct, measurable cost that can be capitalized. When an intangible is purchased, however, or when costs can be directly traced to the development of the asset, the cost is recorded as an intangible asset on the balance sheet.

Intangible assets are valued at their cost of acquisition. A purchased intangible is valued based on the amount paid for the asset. Research and development costs associated with developing an intangible are expensed for the year in which they were incurred.

However, costs of registering patents or trademarks and legal fees incurred to defend a company's right of use are included in the cost of acquisition, which is reported as an intangible asset on the balance sheet.



x 10^3 €

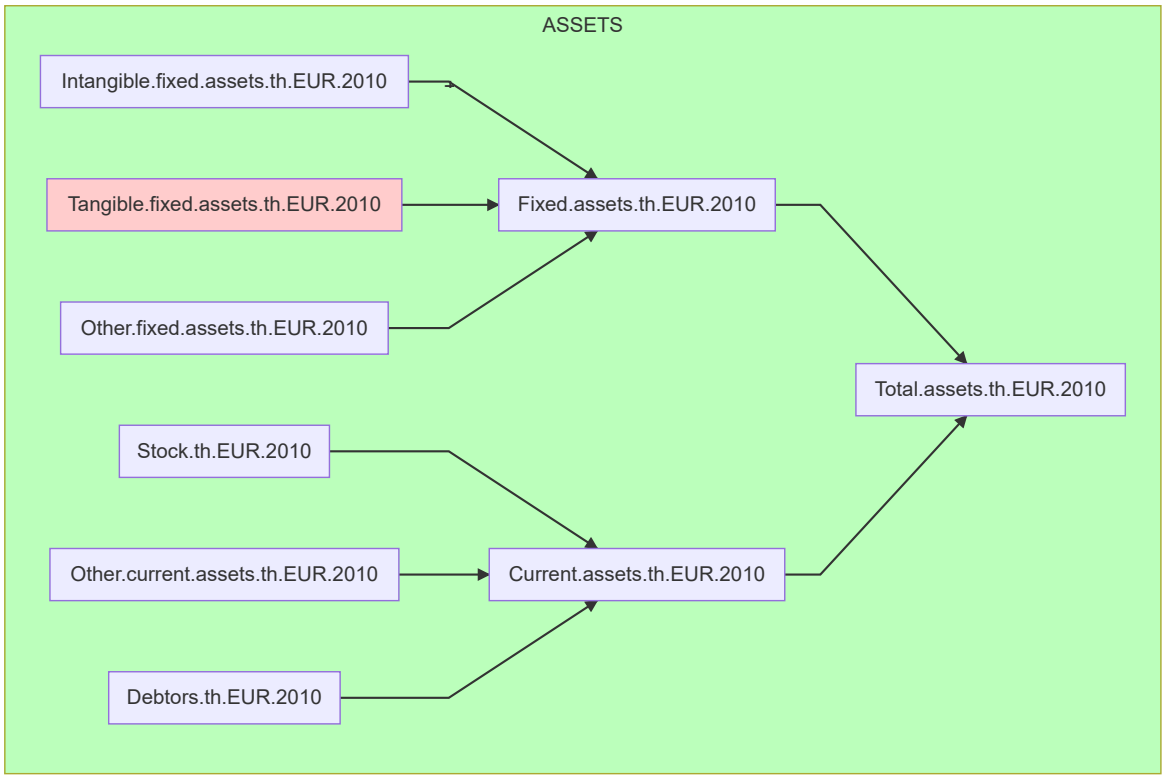
outlier:

```
BvD.ID.number
IT10969001006    LOTTERIE NAZIONALI S.R.L.
Name: Company.name, dtype: string
```

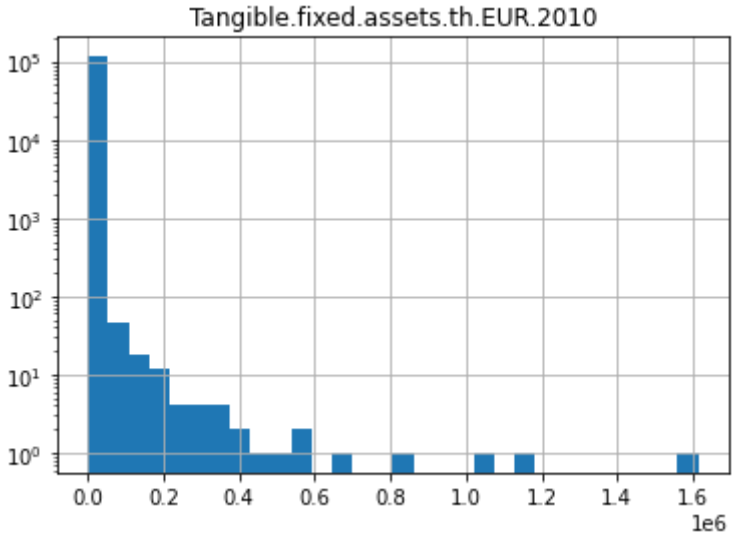
```
HGF vs non-HGF for Intangible.fixed.assets.th.EUR.2010
Welch's t-test statistic = 1.822
p-value = 0.06857
```

```
Optimization terminated successfully.
         Current function value: 0.155657
         Iterations 9
                           Logit Regression Results
==============================================================================
Dep. Variable:                    HGF   No. Observations:               115840
Model:                          Logit   Df Residuals:                   115838
Method:                           MLE   Df Model:                            1
Date:                Mon, 29 Jun 2020   Pseudo R-squ.:               3.442e-05
Time:                        15:28:39   Log-Likelihood:                -18031.
converged:                       True   LL-Null:                       -18032.
Covariance Type:            nonrobust   LLR p-value:                    0.2652
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -3.2814      0.016   -208.487      0.000      -3.312      -3.251
INT        -9.435e-06   1.24e-05     -0.760      0.447   -3.38e-05    1.49e-05
==============================================================================
```

7.2.1.2.21. Tangible.fixed.assets.th.EUR.2010



Tangible fixed assets generally refer to assets that have a physical value. Examples of this are your business premises, equipment, inventory and machinery. Tangible fixed assets have a market value that needs to be accounted for when you file your annual accounts. Some of these assets, for example computer equipment, will incur depreciation, which needs to be factored into your accounts.



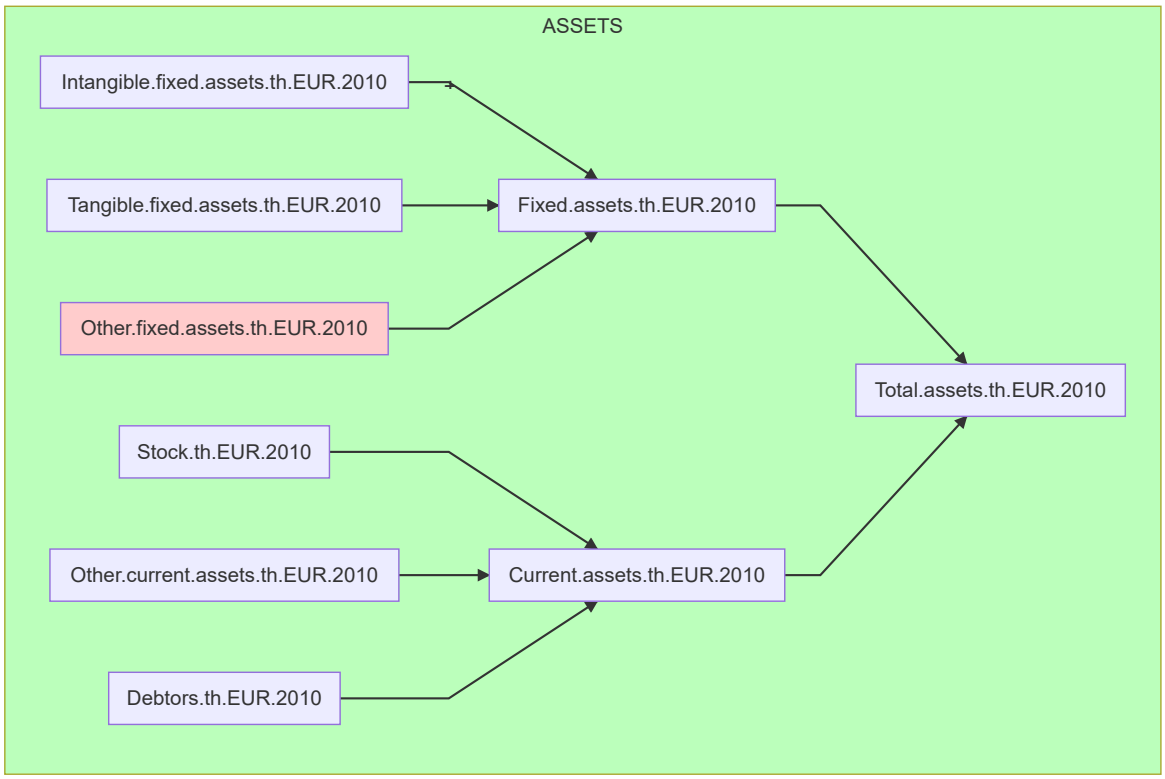
outlier:

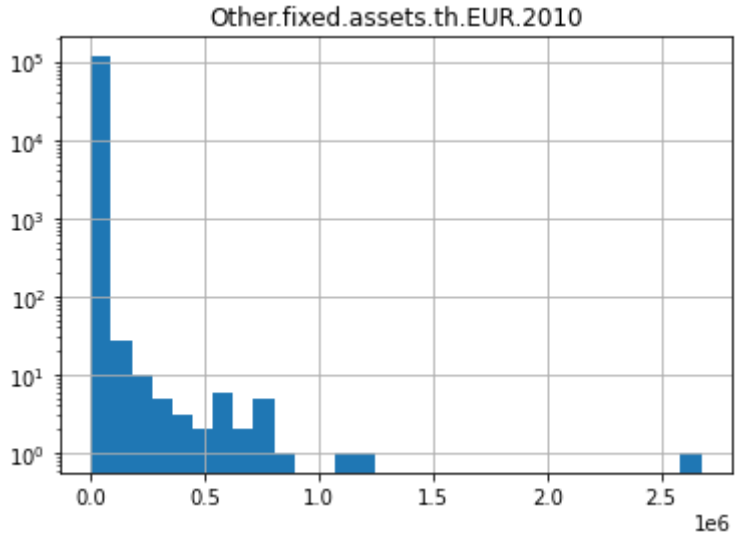
```
BvD.ID.number
GB07145051    CAPITAL & COUNTIES PROPERTIES PLC
Name: Company.name, dtype: string
```

```
HGF vs non-HGF for Tangible.fixed.assets.th.EUR.2010
Welch's t-test statistic = 1.394
p-value = 0.1635

Optimization terminated successfully.
```

```
         Current function value: 0.155655
         Iterations 8
                           Logit Regression Results
==============================================================================
Dep. Variable:                    HGF   No. Observations:               115840
Model:                          Logit   Df Residuals:                   115838
Method:                           MLE   Df Model:                            1
Date:                Mon, 29 Jun 2020   Pseudo R-squ.:               4.600e-05
Time:                        15:29:04   Log-Likelihood:                -18031.
converged:                       True   LL-Null:                       -18032.
Covariance Type:            nonrobust   LLR p-value:                    0.1977
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -3.2808      0.016   -208.167      0.000      -3.312      -3.250
T          -4.673e-06   4.82e-06     -0.969      0.332   -1.41e-05    4.78e-06
==============================================================================
```

7.2.1.2.22. Other.fixed.assets.th.EUR.2010



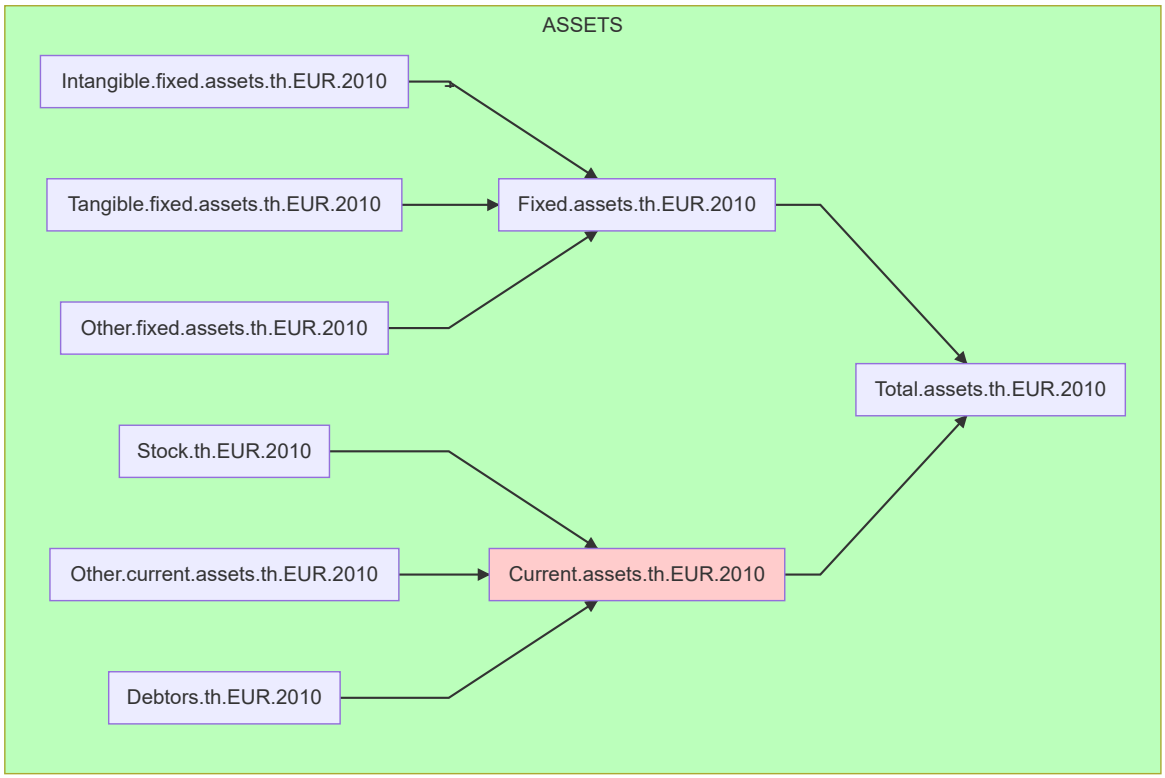


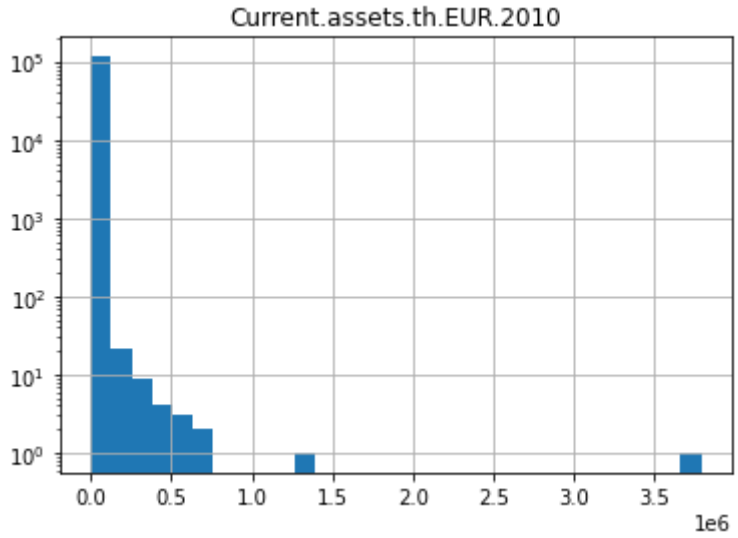
outlier:

```
BvD.ID.number
IE486605    SAP IRELAND US-FINANCIAL SERVICES DESIGNATED A...
Name: Company.name, dtype: string
```

```
HGF vs non-HGF for Other.fixed.assets.th.EUR.2010
Welch's t-test statistic = 4.793
p-value = 1.65e-06

Optimization terminated successfully.
         Current function value: 0.155642
         Iterations 10
                           Logit Regression Results
==============================================================================
Dep. Variable:                    HGF   No. Observations:               115840
Model:                          Logit   Df Residuals:                   115838
Method:                           MLE   Df Model:                            1
Date:                Mon, 29 Jun 2020   Pseudo R-squ.:               0.0001271
Time:                        15:29:34   Log-Likelihood:                -18030.
converged:                       True   LL-Null:                       -18032.
Covariance Type:            nonrobust   LLR p-value:                   0.03230
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -3.2804      0.016   -208.259      0.000      -3.311      -3.249
OF         -1.366e-05   1.07e-05     -1.272      0.203   -3.47e-05    7.38e-06
==============================================================================
```

7.2.1.2.23. Current.assets.th.EUR.2010





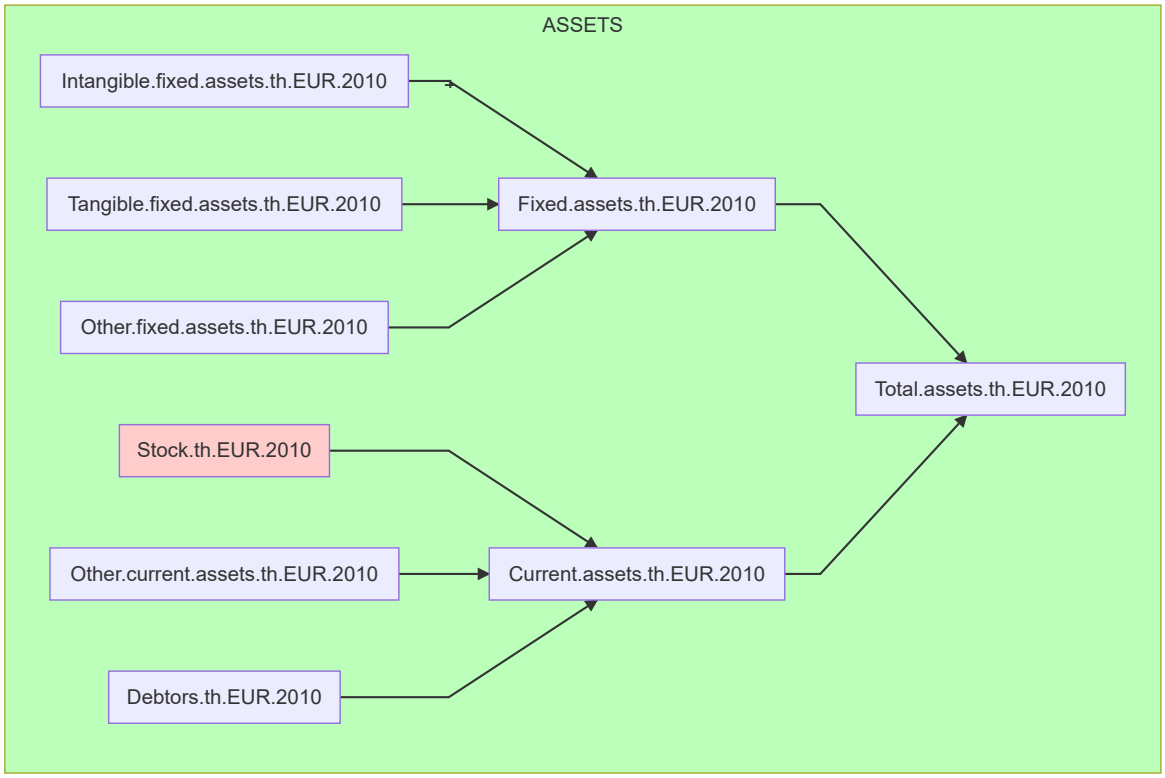

```
outlier:
BvD.ID.number
GB07450219    LONG ISLAND ASSETS LIMITED
Name: Company.name, dtype: string
```

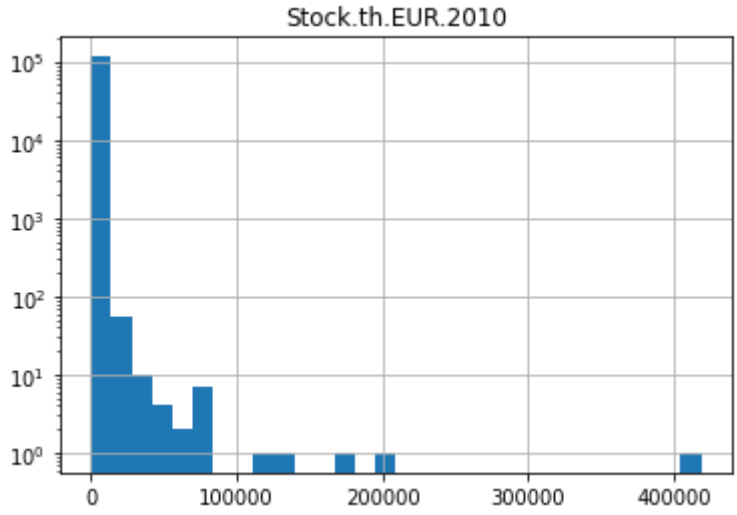
```
HGF vs non-HGF for Current.assets.th.EUR.2010
Welch's t-test statistic = 6.732
p-value = 1.687e-11
```

```
Optimization terminated successfully.
         Current function value: 0.155487
         Iterations 11
                           Logit Regression Results
==============================================================================
Dep. Variable:                    HGF   No. Observations:                115840
Model:                          Logit   Df Residuals:                    115838
Method:                           MLE   Df Model:                             1
```

```
Date:                Mon, 29 Jun 2020   Pseudo R-squ.:                0.001126
Time:                        15:30:01   Log-Likelihood:                -18012.
converged:                       True   LL-Null:                       -18032.
Covariance Type:            nonrobust   LLR p-value:                 1.856e-10
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -3.2613      0.016   -201.717      0.000      -3.293      -3.230
CA            -0.0001    2.7e-05     -4.235      0.000      -0.000   -6.15e-05
==============================================================================
```

7.2.1.2.24. Stock.th.EUR.2010



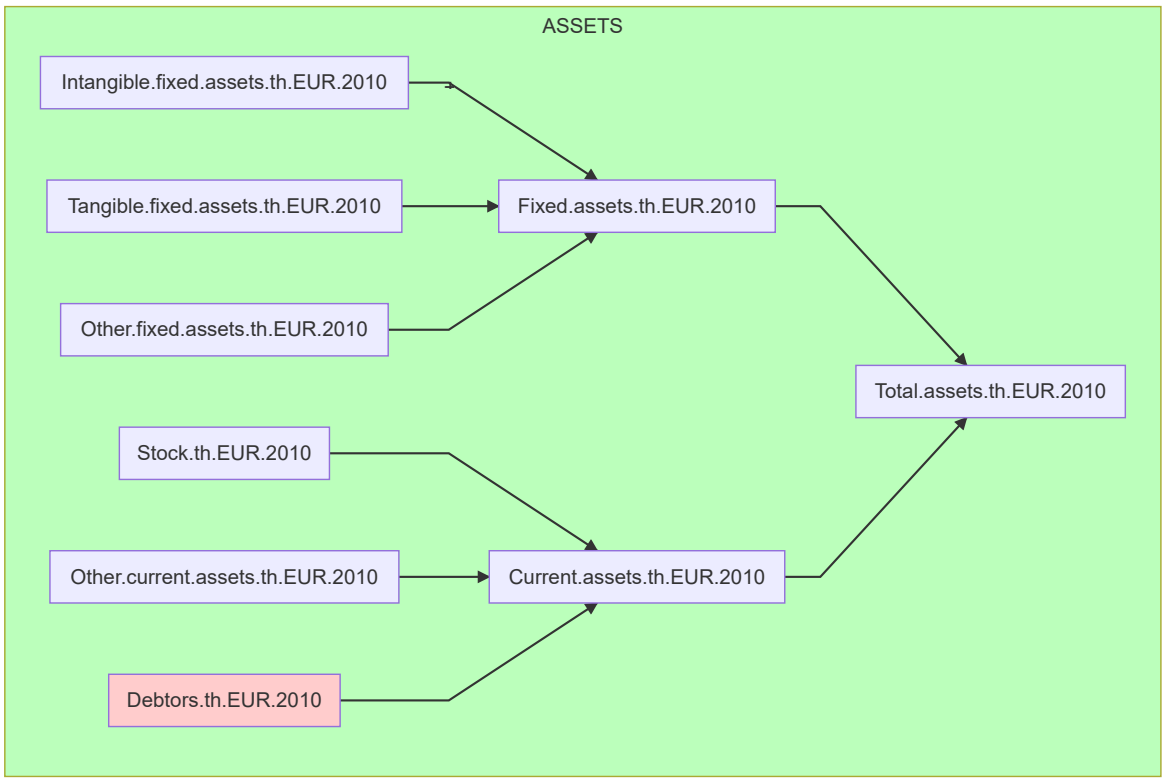


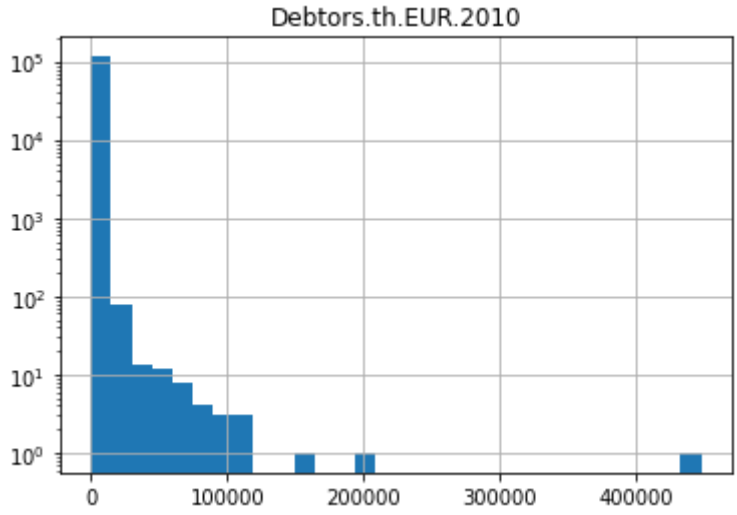
outlier:

```
BvD.ID.number
IT07099900966    MILANOSESTO S.P.A.
Name: Company.name, dtype: string
```

```
HGF vs non-HGF for Stock.th.EUR.2010
Welch's t-test statistic = 10.69
p-value = 1.155e-26

Optimization terminated successfully.
         Current function value: 0.155221
         Iterations 11
                           Logit Regression Results
==============================================================================
Dep. Variable:                    HGF   No. Observations:               115840
Model:                          Logit   Df Residuals:                   115838
Method:                           MLE   Df Model:                            1
Date:                Mon, 29 Jun 2020   Pseudo R-squ.:                0.002831
Time:                        15:30:21   Log-Likelihood:                -17981.
converged:                       True   LL-Null:                       -18032.
Covariance Type:            nonrobust   LLR p-value:                5.302e-24
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -3.2474      0.016   -201.604      0.000      -3.279      -3.216
S             -0.0015      0.000     -6.485      0.000      -0.002      -0.001
==============================================================================
```

7.2.1.2.25. Debtors.th.EUR.2010



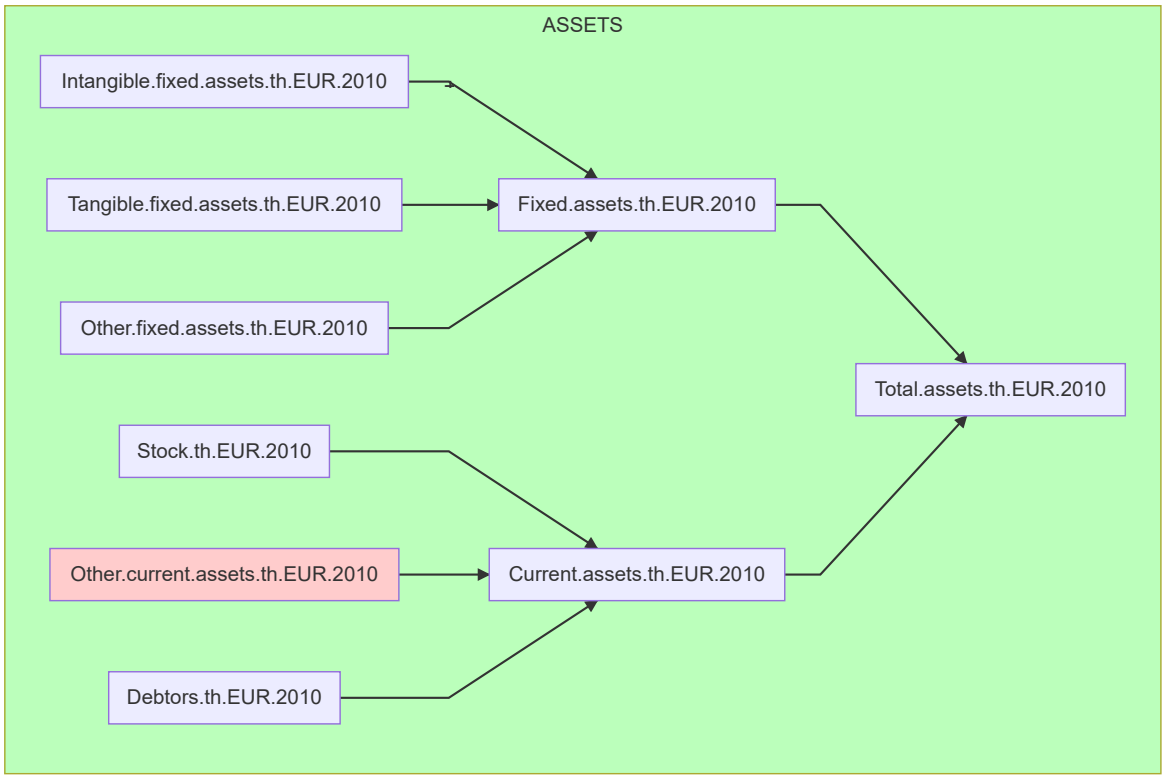


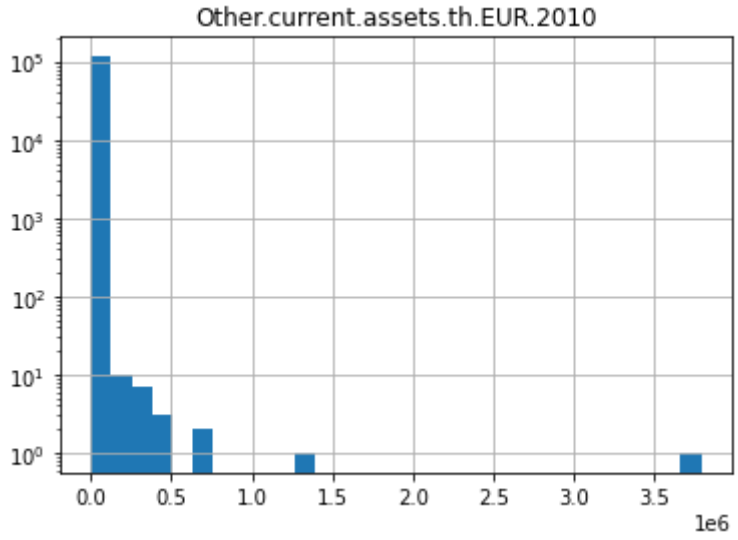
outlier:

```
BvD.ID.number
GB07246104    MACSCO 22 LIMITED
Name: Company.name, dtype: string
```

```
HGF vs non-HGF for Debtors.th.EUR.2010
Welch's t-test statistic = 7.135
p-value = 1.014e-12

Optimization terminated successfully.
         Current function value: 0.155553
         Iterations 10
                           Logit Regression Results
==============================================================================
Dep. Variable:                    HGF   No. Observations:               115840
Model:                          Logit   Df Residuals:                   115838
Method:                           MLE   Df Model:                            1
Date:                Mon, 29 Jun 2020   Pseudo R-squ.:               0.0007008
Time:                        15:30:49   Log-Likelihood:                -18019.
converged:                       True   LL-Null:                       -18032.
Covariance Type:            nonrobust   LLR p-value:                 4.972e-07
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -3.2697      0.016   -204.720      0.000      -3.301      -3.238
DEB           -0.0002    7.3e-05     -3.356      0.001      -0.000      -0.000
==============================================================================
```

7.2.1.2.26. Other.current.assets.th.EUR.2010





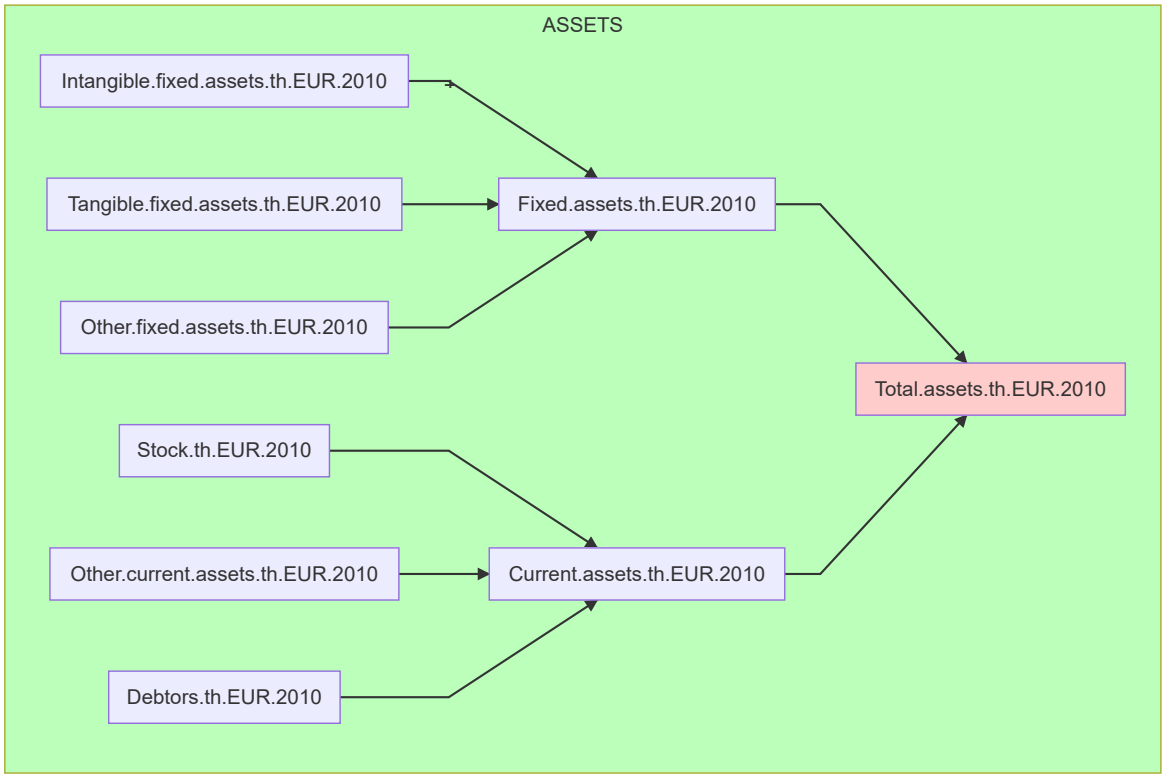
outlier:

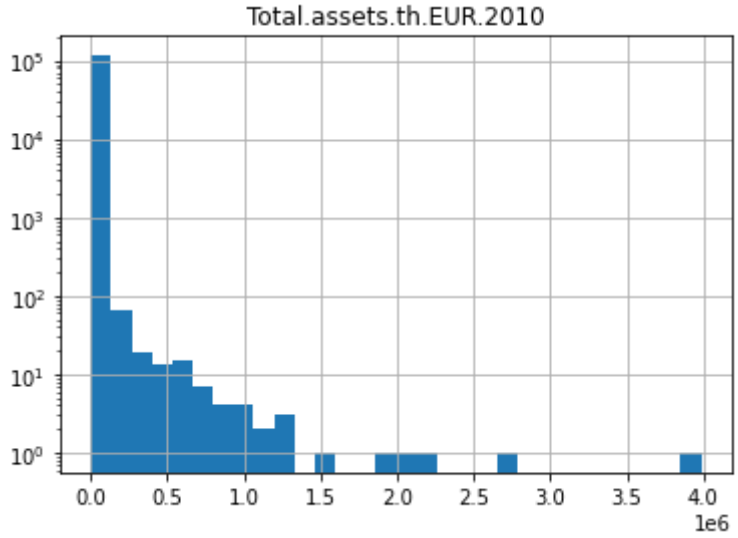
```
BvD.ID.number
GB07450219    LONG ISLAND ASSETS LIMITED
Name: Company.name, dtype: string
```

```
HGF vs non-HGF for Other.current.assets.th.EUR.2010
Welch's t-test statistic = 4.135
p-value = 3.56e-05

Optimization terminated successfully.
         Current function value: 0.155609
         Iterations 10
                           Logit Regression Results
==============================================================================
Dep. Variable:                    HGF   No. Observations:                115840
Model:                          Logit   Df Residuals:                    115838
```

```
Method:                           MLE   Df Model:                            1
Date:                Mon, 29 Jun 2020   Pseudo R-squ.:               0.0003421
Time:                        15:31:18   Log-Likelihood:                -18026.
converged:                       True   LL-Null:                       -18032.
Covariance Type:            nonrobust   LLR p-value:                 0.0004442
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -3.2750      0.016   -206.020      0.000      -3.306      -3.244
OCA        -6.435e-05   2.72e-05     -2.366      0.018      -0.000   -1.11e-05
==============================================================================
```

7.2.1.2.27. Total.assets.th.EUR.2010



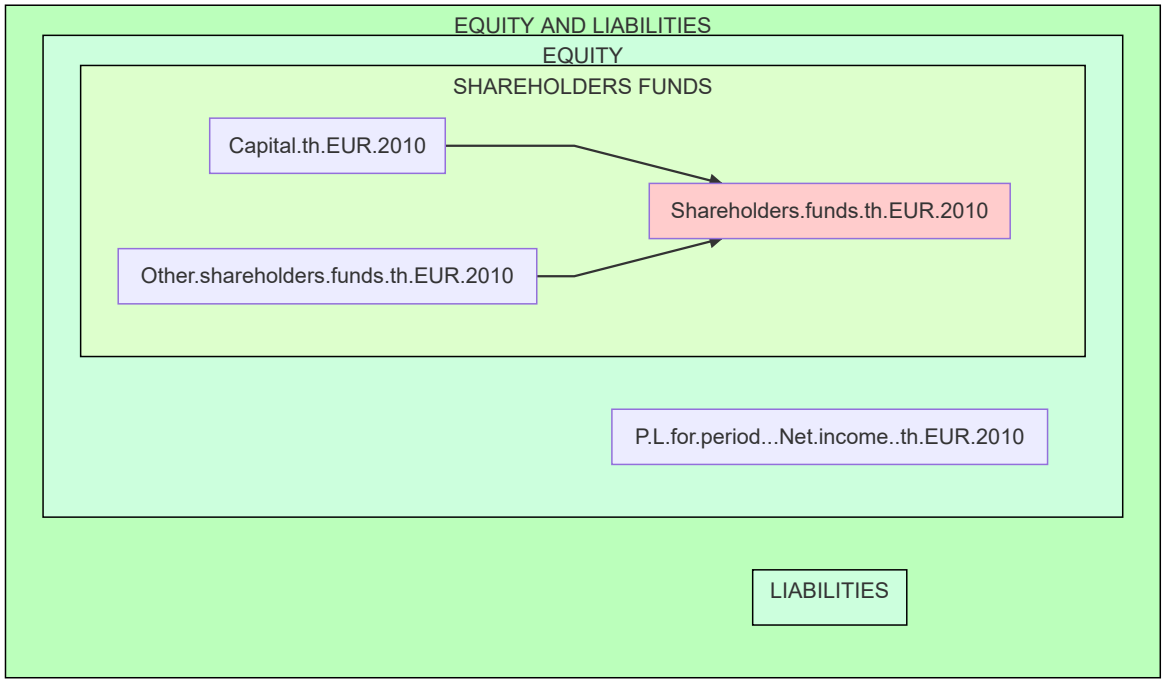


outlier:

```
BvD.ID.number
GB07450219    LONG ISLAND ASSETS LIMITED
Name: Company.name, dtype: string
```

```
HGF vs non-HGF for Total.assets.th.EUR.2010
Welch's t-test statistic = 5.512
p-value = 3.645e-08

Optimization terminated successfully.
         Current function value: 0.155604
         Iterations 10
                           Logit Regression Results
==============================================================================
Dep. Variable:                    HGF   No. Observations:               115840
Model:                          Logit   Df Residuals:                   115838
Method:                           MLE   Df Model:                            1
Date:                Mon, 29 Jun 2020   Pseudo R-squ.:               0.0003744
Time:                        15:31:50   Log-Likelihood:                -18025.
converged:                       True   LL-Null:                       -18032.
Covariance Type:            nonrobust   LLR p-value:                 0.0002385
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -3.2751      0.016   -206.277      0.000      -3.306      -3.244
TA         -1.137e-05   4.72e-06     -2.406      0.016   -2.06e-05   -2.11e-06
==============================================================================
```

7.2.1.2.28. Shareholders.funds.th.EUR.2010



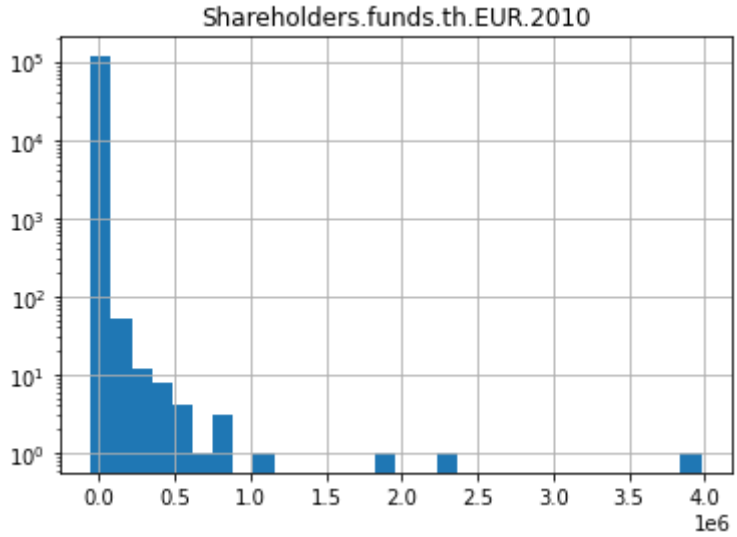
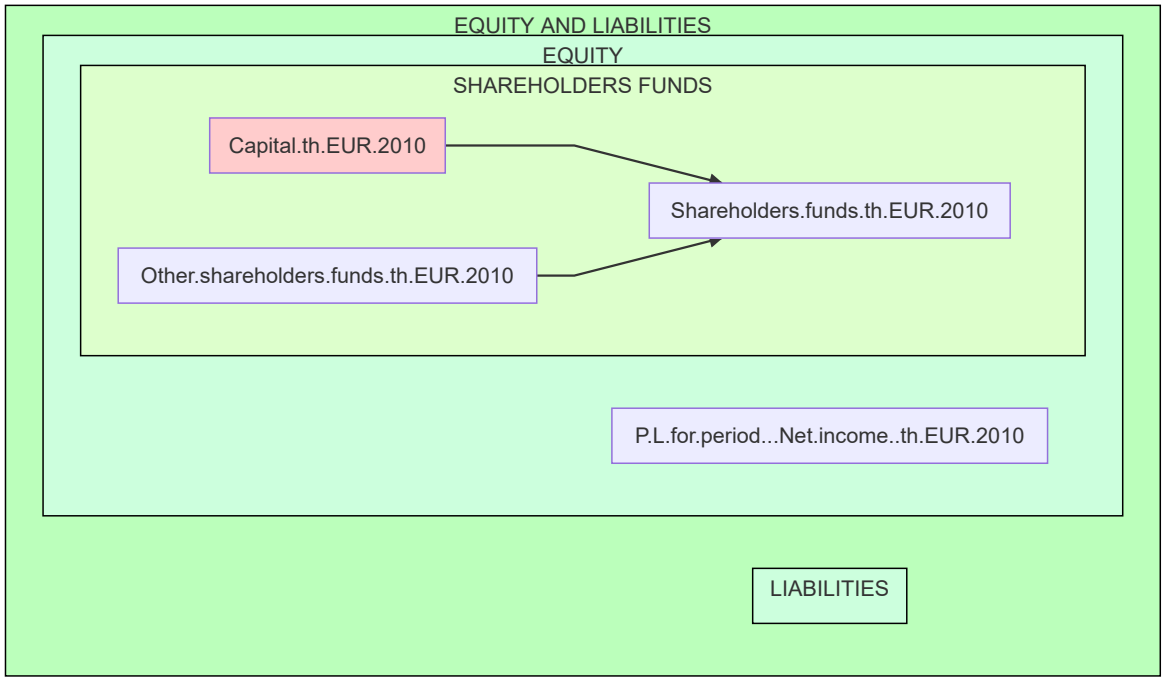


image-20200624142754096

```
outlier:

BvD.ID.number
GB07450219    LONG ISLAND ASSETS LIMITED
Name: Company.name, dtype: string

HGF vs non-HGF for Shareholders.funds.th.EUR.2010
Welch's t-test statistic = 1.909
p-value = 0.05634

Optimization terminated successfully.
         Current function value: 0.155657
         Iterations 8
                           Logit Regression Results
==============================================================================
Dep. Variable:                    HGF   No. Observations:               115840
Model:                          Logit   Df Residuals:                   115838
Method:                           MLE   Df Model:                            1
Date:                Mon, 29 Jun 2020   Pseudo R-squ.:               3.438e-05
Time:                        15:32:19   Log-Likelihood:                -18031.
converged:                       True   LL-Null:                       -18032.
Covariance Type:            nonrobust   LLR p-value:                    0.2655
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -3.2812      0.016   -208.351      0.000      -3.312      -3.250
SHA        -2.844e-06   3.41e-06     -0.834      0.404   -9.53e-06    3.84e-06
==============================================================================
```

7.2.1.2.29. Capital.th.EUR.2010



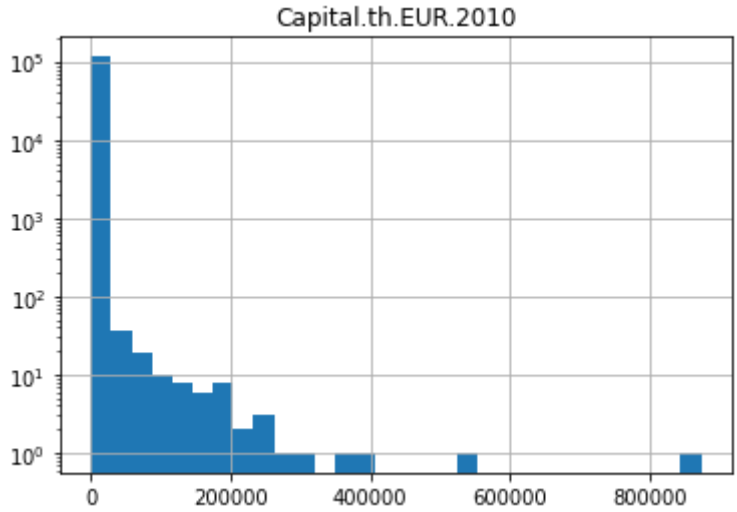


image-20200624142543360

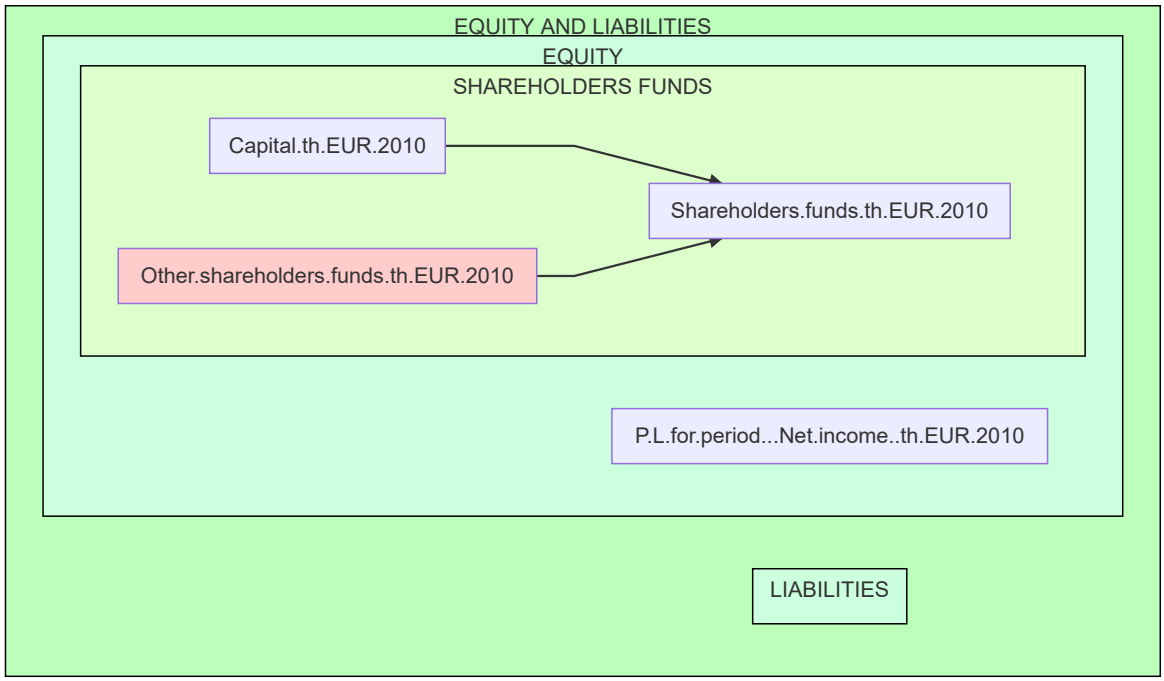
outlier:

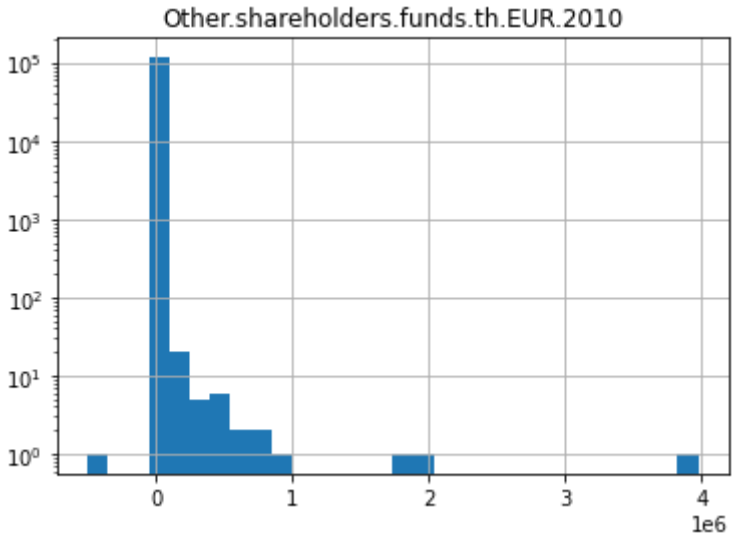
```
BvD.ID.number
FR519720643    IRIDIUM FRANCE
Name: Company.name, dtype: string
```

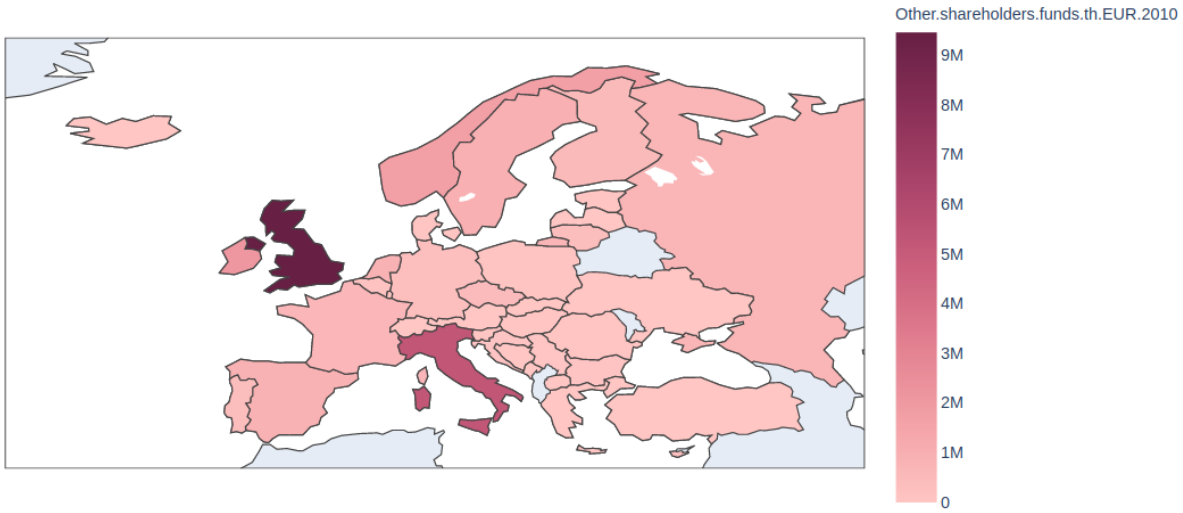
```
HGF vs non-HGF for Capital.th.EUR.2010
Welch's t-test statistic = 4.127
p-value = 3.699e-05

Optimization terminated successfully.
         Current function value: 0.155645
         Iterations 9
                           Logit Regression Results
==============================================================================
Dep. Variable:                    HGF   No. Observations:                115840
Model:                          Logit   Df Residuals:                    115838
Method:                           MLE   Df Model:                             1
```

```
Date:                Mon, 29 Jun 2020   Pseudo R-squ.:                0.0001079
Time:                        15:32:46   Log-Likelihood:                -18030.
converged:                       True   LL-Null:                       -18032.
Covariance Type:            nonrobust   LLR p-value:                    0.04852
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -3.2801      0.016   -208.097      0.000      -3.311      -3.249
CAP        -1.822e-05   1.36e-05     -1.337      0.181   -4.49e-05    8.49e-06
==============================================================================
```

7.2.1.2.30. Other.shareholders.funds.th.EUR.2010





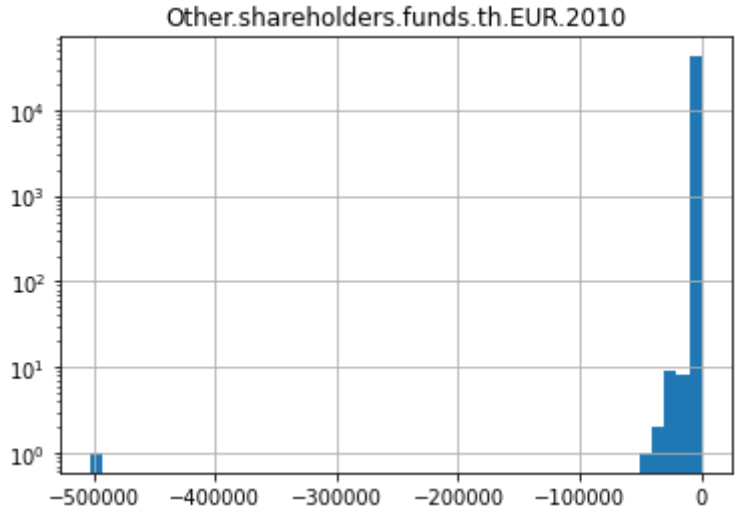


42359 companies has negative values in the range (-)

outlier:

```
BvD.ID.number
GB07450219    LONG ISLAND ASSETS LIMITED
Name: Company.name, dtype: string
```

Distribution of negative values:



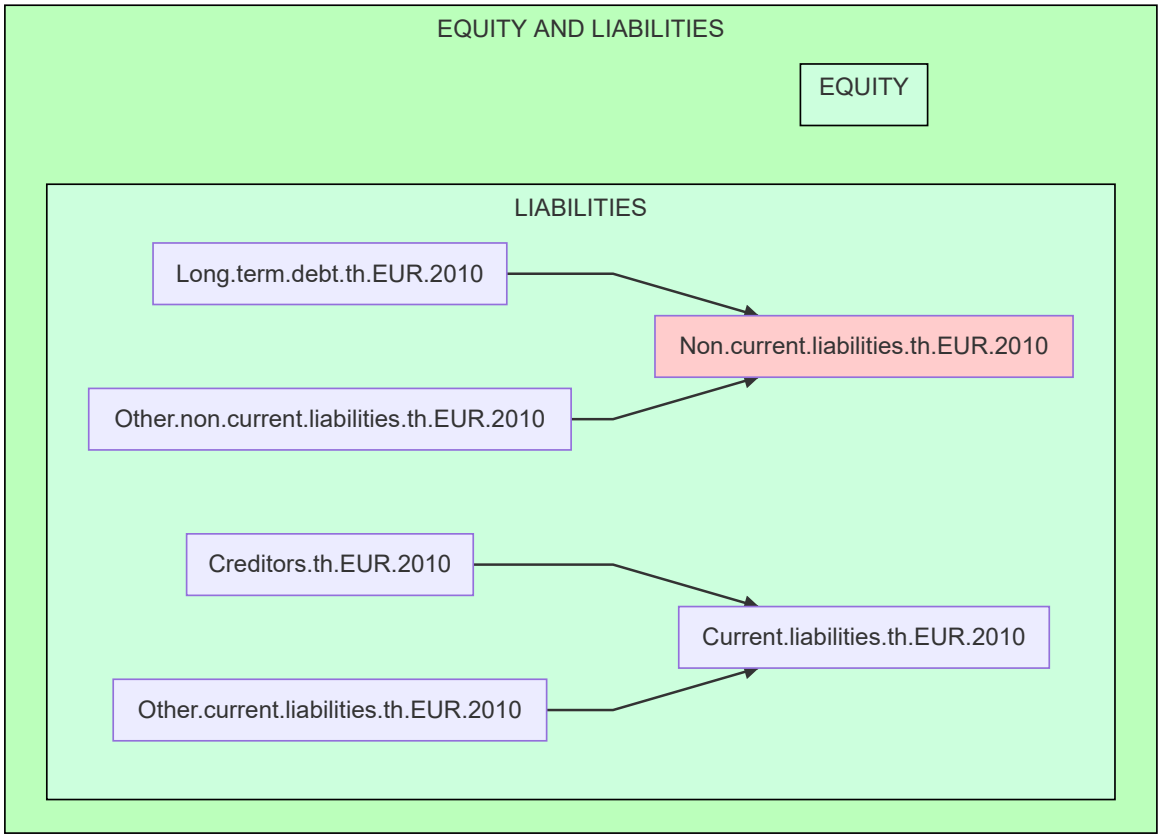
outlier:

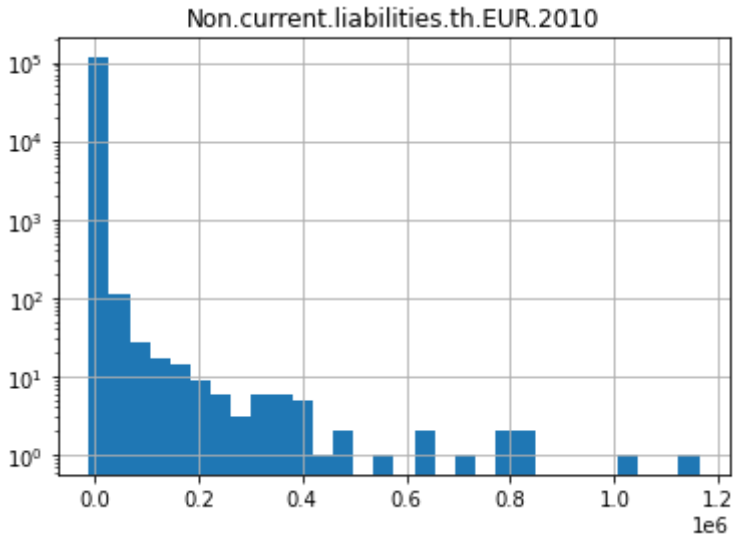
```
vD.ID.number
FR519720643    IRIDIUM FRANCE
Name: Company.name, dtype: string BvD.ID.number
FR519720643   -502140.0
Name: Other.shareholders.funds.th.EUR.2010, dtype: float64
```

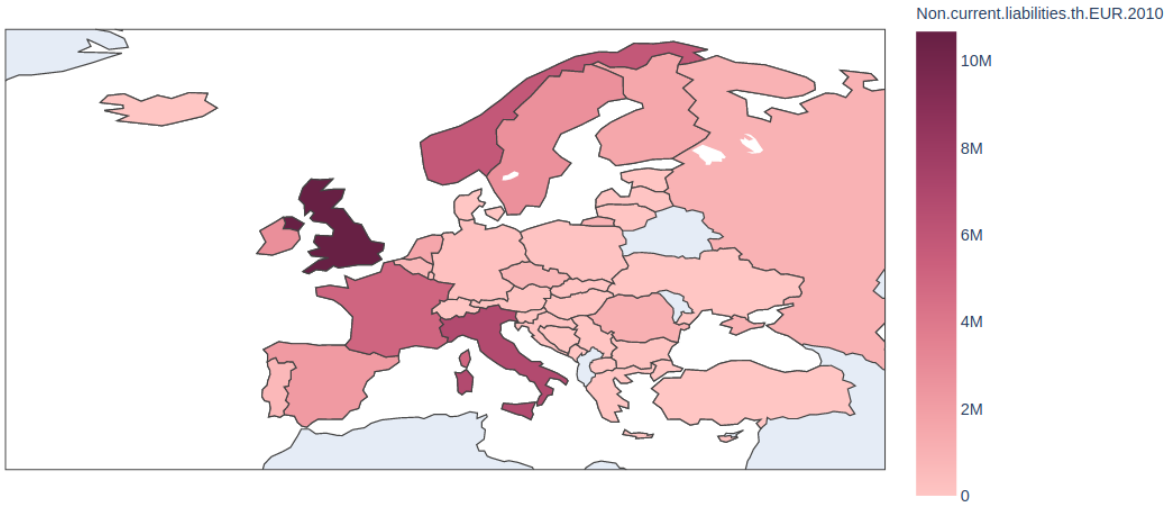
```
HGF vs non-HGF for Other.shareholders.funds.th.EUR.2010
Welch's t-test statistic = 0.8908
p-value = 0.3731

Optimization terminated successfully.
         Current function value: 0.155661
         Iterations 7
                           Logit Regression Results
==============================================================================
Dep. Variable:                    HGF   No. Observations:                115840
```

```
Model:                          Logit   Df Residuals:                   115838
Method:                           MLE   Df Model:                            1
Date:                Mon, 29 Jun 2020   Pseudo R-squ.:               4.778e-06
Time:                        15:33:13   Log-Likelihood:                -18032.
converged:                       True   LL-Null:                       -18032.
Covariance Type:            nonrobust   LLR p-value:                    0.6781
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -3.2818      0.016   -208.591      0.000      -3.313      -3.251
OSF        -6.672e-07      2e-06     -0.334      0.739   -4.59e-06    3.25e-06
==============================================================================
```

7.2.1.2.31. Non.current.liabilities.th.EUR.2010

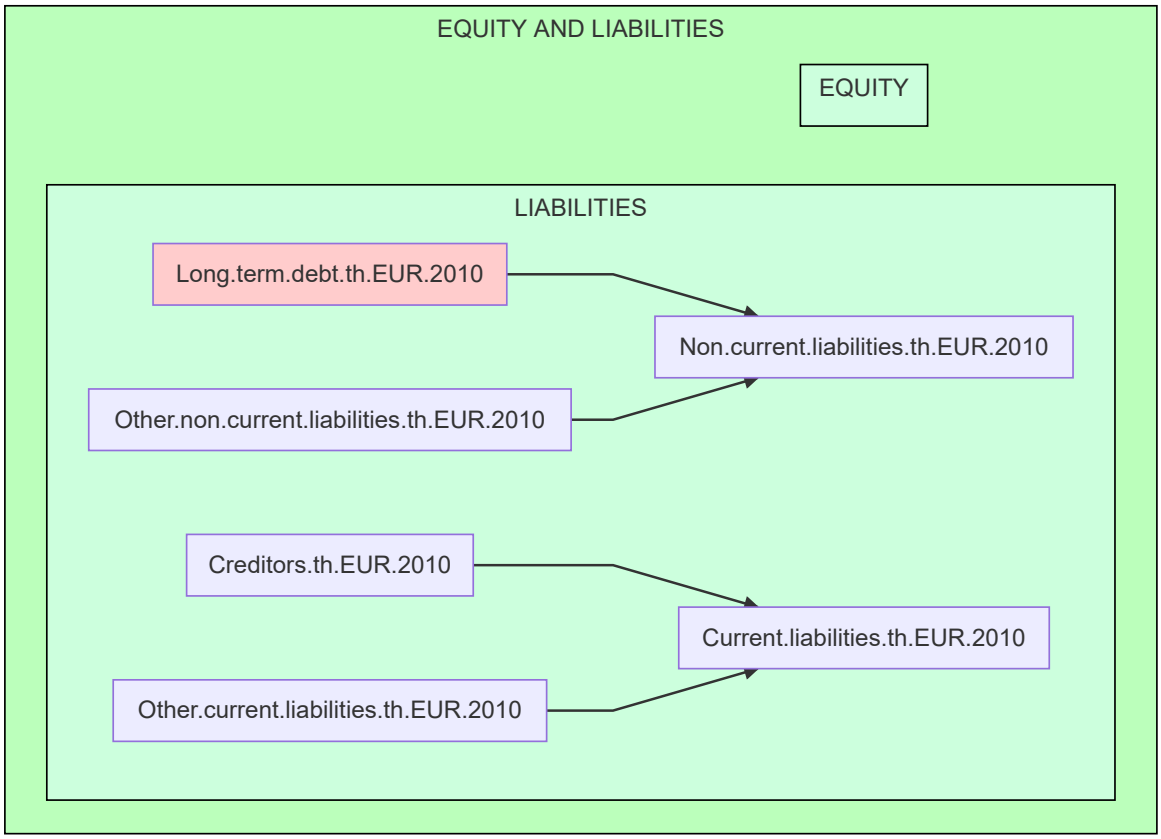


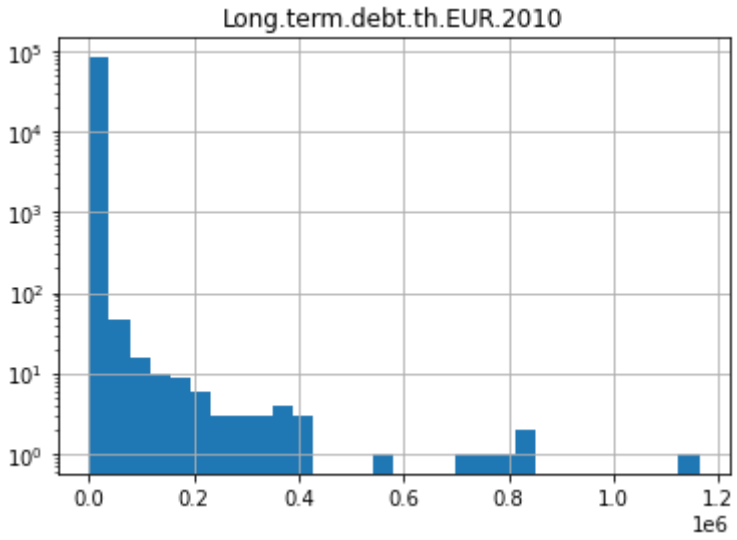


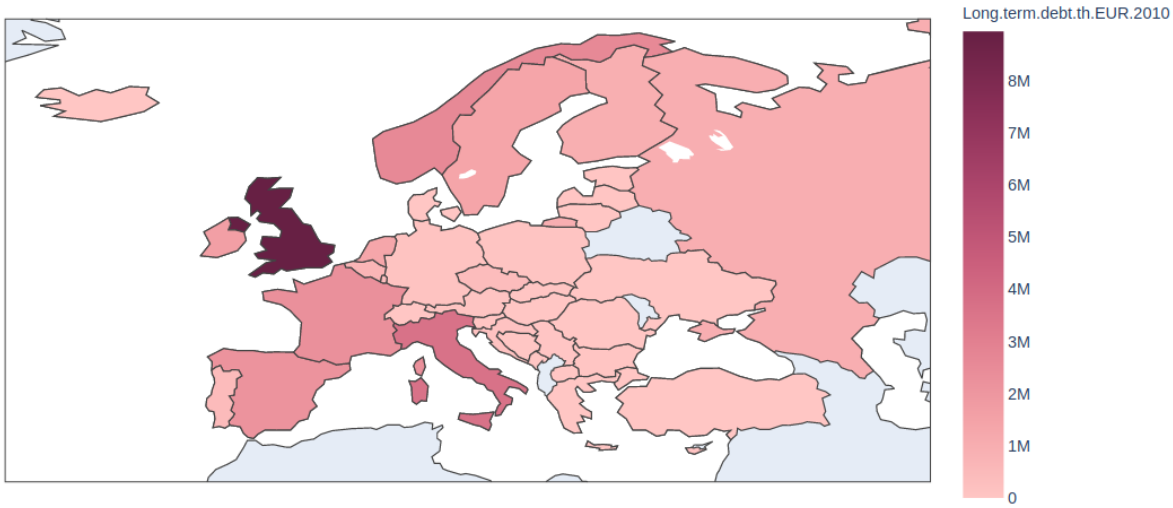



```
HGF vs non-HGF for Non.current.liabilities.th.EUR.2010
Welch's t-test statistic = 7.868
p-value = 3.731e-15

Optimization terminated successfully.
         Current function value: 0.155577
         Iterations 10
                           Logit Regression Results
==============================================================================
Dep. Variable:                    HGF   No. Observations:               115840
Model:                          Logit   Df Residuals:                   115838
Method:                           MLE   Df Model:                            1
Date:                Mon, 29 Jun 2020   Pseudo R-squ.:               0.0005448
Time:                        15:34:00   Log-Likelihood:                -18022.
converged:                       True   LL-Null:                       -18032.
Covariance Type:            nonrobust   LLR p-value:                 9.307e-06
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -3.2743      0.016   -206.612      0.000      -3.305      -3.243
NCL        -5.339e-05   1.92e-05     -2.774      0.006   -9.11e-05   -1.57e-05
==============================================================================

```

7.2.1.2.32. Long.term.debt.th.EUR.2010


EQUITY AND LIABILITIES
EQUITY
LIABILITIES
Long.term.debt.th.EUR.2010
Non.current.liabilities.th.EUR.2010
Other.non.current.liabilities.th.EUR.2010
Creditors.th.EUR.2010
Current.liabilities.th.EUR.2010
Other.current.liabilities.th.EUR.2010


Long.term.debt.th.EUR.2010
10^5
10^4
10^3
10^2
10^1
10^0
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1e6

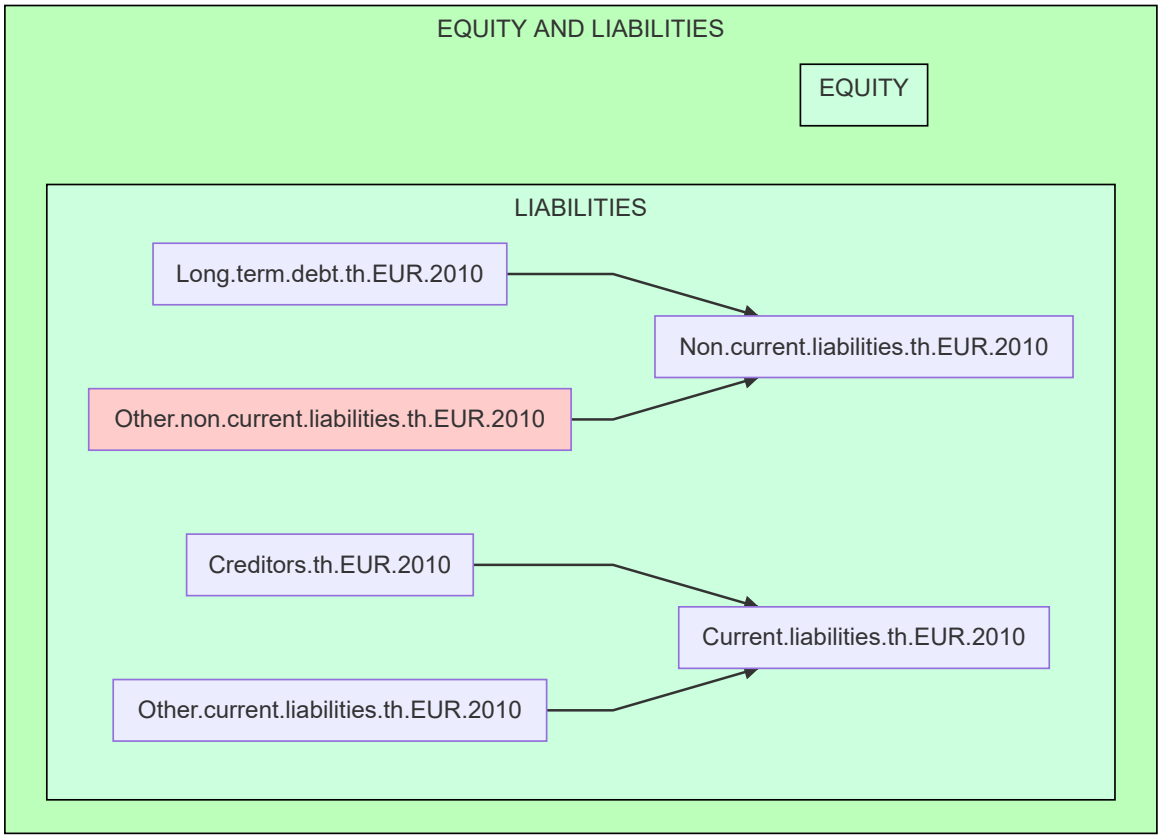

Long.term.debt.th.EUR.2010
8M
7M
6M
5M
4M
3M
2M
1M
0

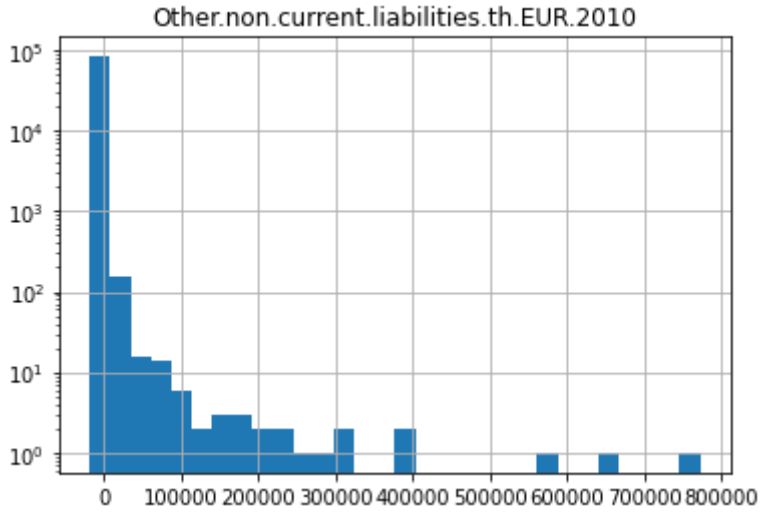
outlier:

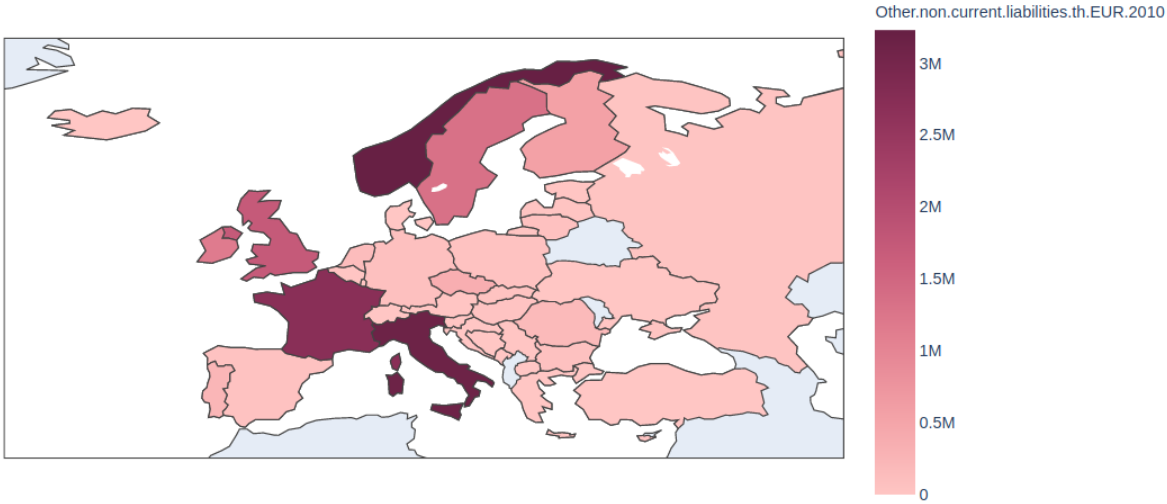
```
BvD.ID.number
GB07251526    TESCO PROPERTY FINANCE 3 PLC
Name: Company.name, dtype: string
```

```
Optimization terminated successfully.
         Current function value: 0.143145
         Iterations 10
                           Logit Regression Results
==============================================================================
Dep. Variable:                    HGF   No. Observations:                83074
Model:                          Logit   Df Residuals:                    83072
Method:                           MLE   Df Model:                            1
Date:                Mon, 29 Jun 2020   Pseudo R-squ.:               0.0004459
Time:                        15:34:25   Log-Likelihood:                -11892.
converged:                       True   LL-Null:                       -11897.
Covariance Type:            nonrobust   LLR p-value:                  0.001125
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -3.3885      0.020   -172.120      0.000      -3.427      -3.350
LTD        -4.824e-05   2.41e-05     -1.998      0.046   -9.56e-05   -9.12e-07
==============================================================================
```

7.2.1.2.33. Other.non.current.liabilities.th.EUR.2010



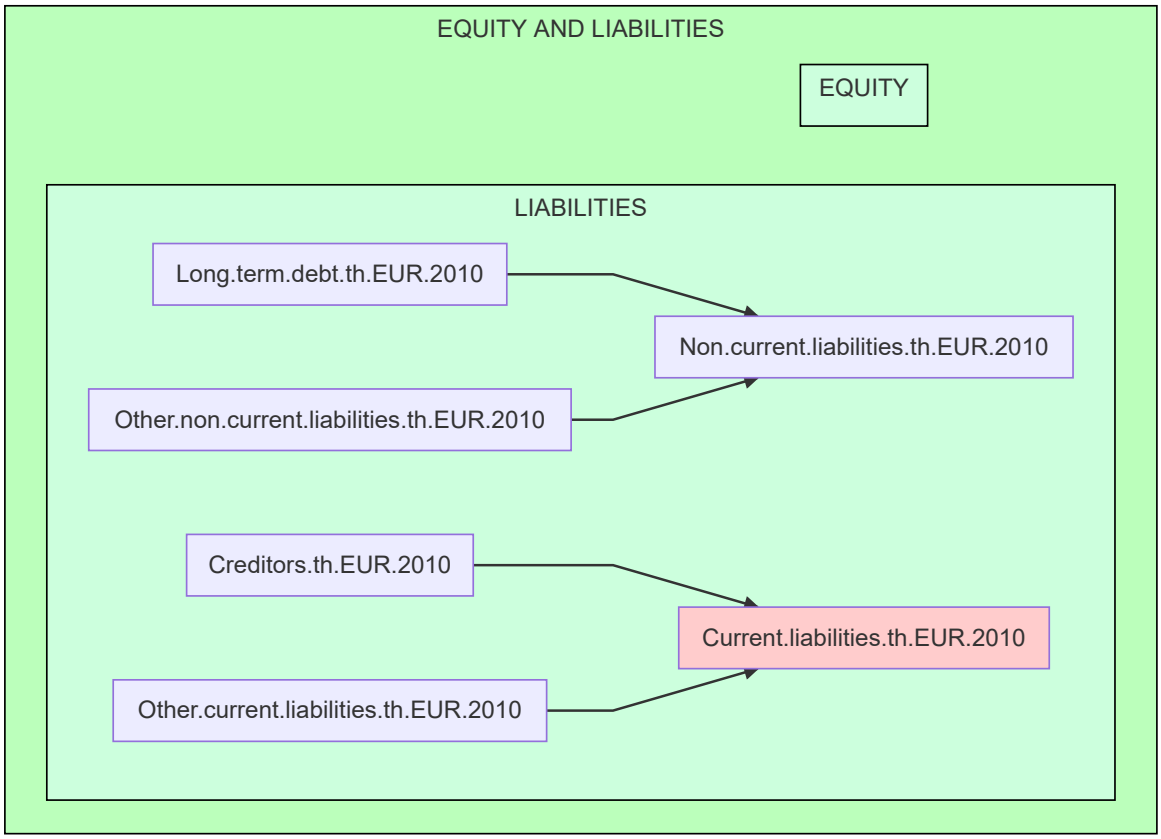


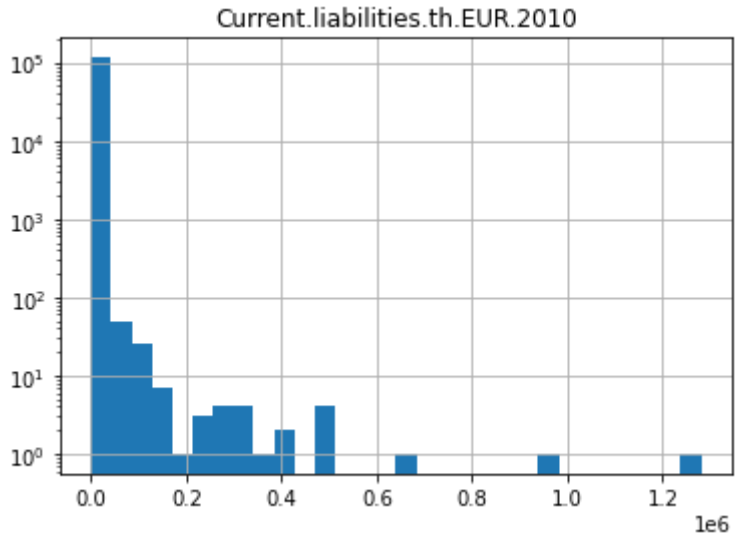


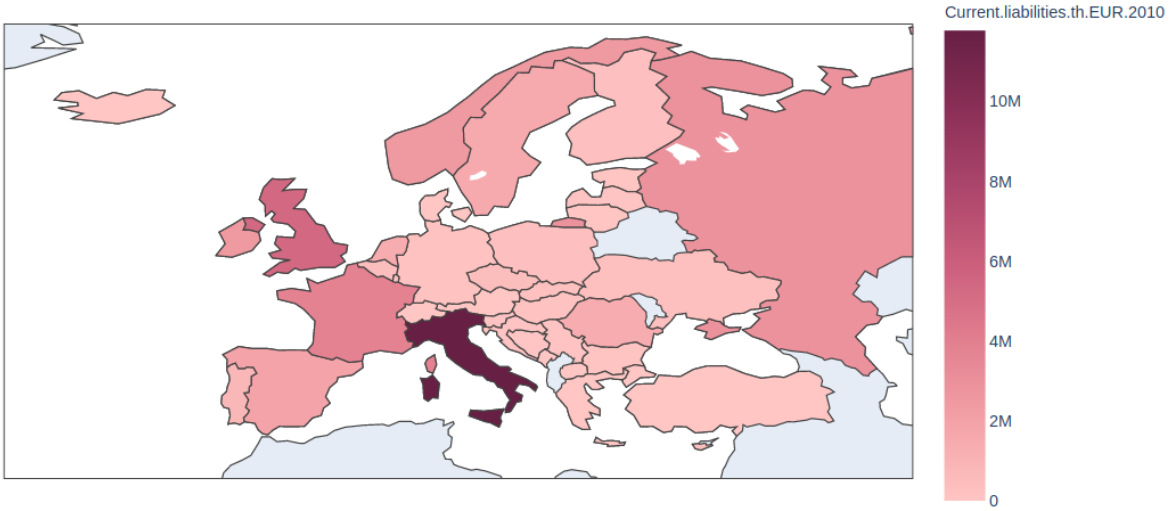
outliers:

```
vD.ID.number
IE487769       SCF CAPITAL DESIGNATED ACTIVITY COMPANY
FR521029926                              DIACINE FRANCE
NO995633604                  INDUSTRIINVESTERINGER AS

Optimization terminated successfully.
         Current function value: 0.143132
         Iterations 10
                           Logit Regression Results
==============================================================================
Dep. Variable:                    HGF   No. Observations:                83074
Model:                          Logit   Df Residuals:                    83072
Method:                           MLE   Df Model:                            1
Date:                Mon, 29 Jun 2020   Pseudo R-squ.:               0.0005412
Time:                        15:34:51   Log-Likelihood:                -11891.
converged:                       True   LL-Null:                       -11897.
Covariance Type:            nonrobust   LLR p-value:                 0.0003325
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -3.3869      0.020   -171.972      0.000      -3.426      -3.348
ONCL          -0.0001   4.63e-05     -2.467      0.014      -0.000   -2.35e-05
==============================================================================
```

7.2.1.2.34. Current.liabilities.th.EUR.2010


EQUITY AND LIABILITIES
EQUITY
LIABILITIES
Long.term.debt.th.EUR.2010
Non.current.liabilities.th.EUR.2010
Other.non.current.liabilities.th.EUR.2010
Creditors.th.EUR.2010
Current.liabilities.th.EUR.2010
Other.current.liabilities.th.EUR.2010


Current.liabilities.th.EUR.2010
10^5
10^4
10^3
10^2
10^1
10^0
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1e6

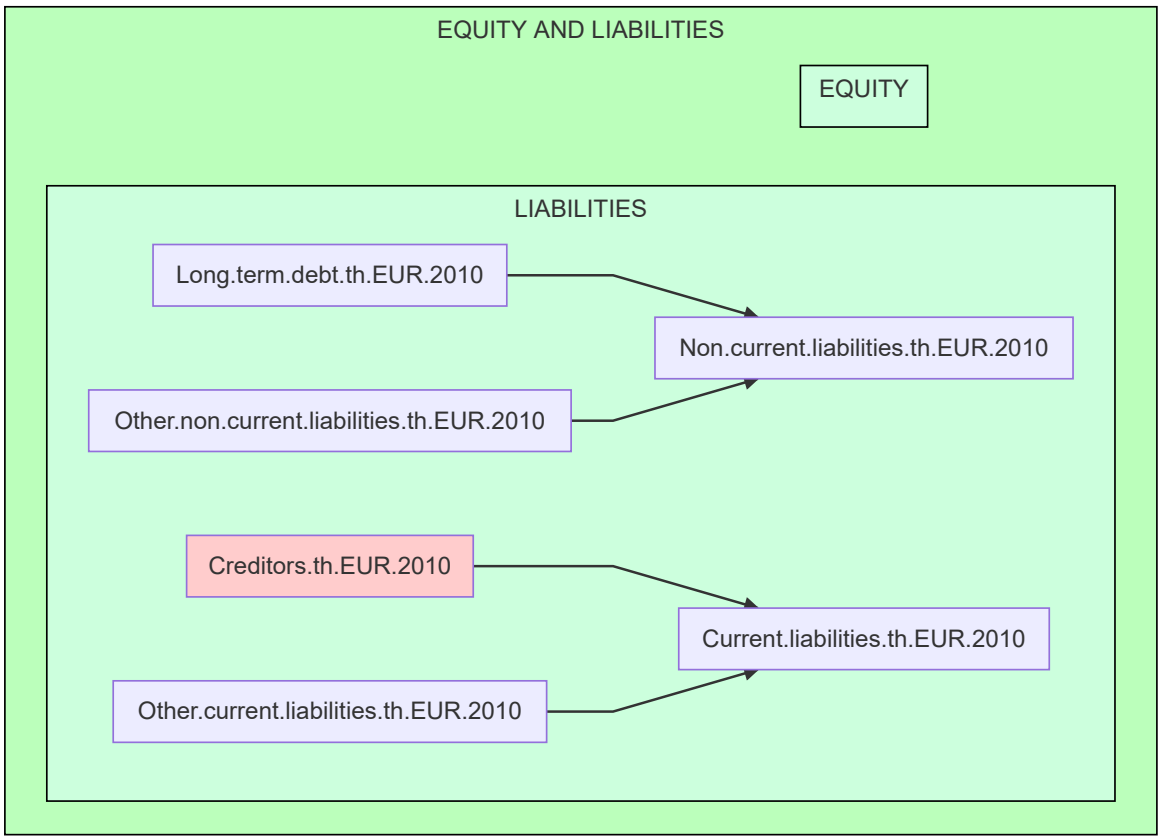

Current.liabilities.th.EUR.2010
10M
8M
6M
4M
2M
0

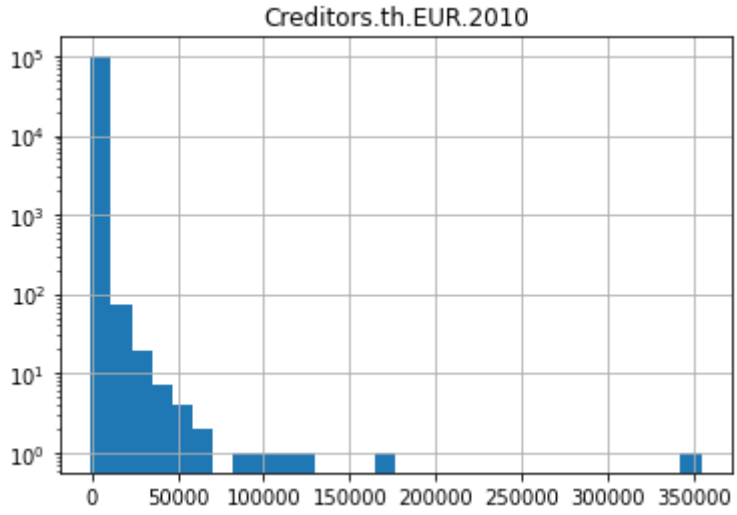
outliers:

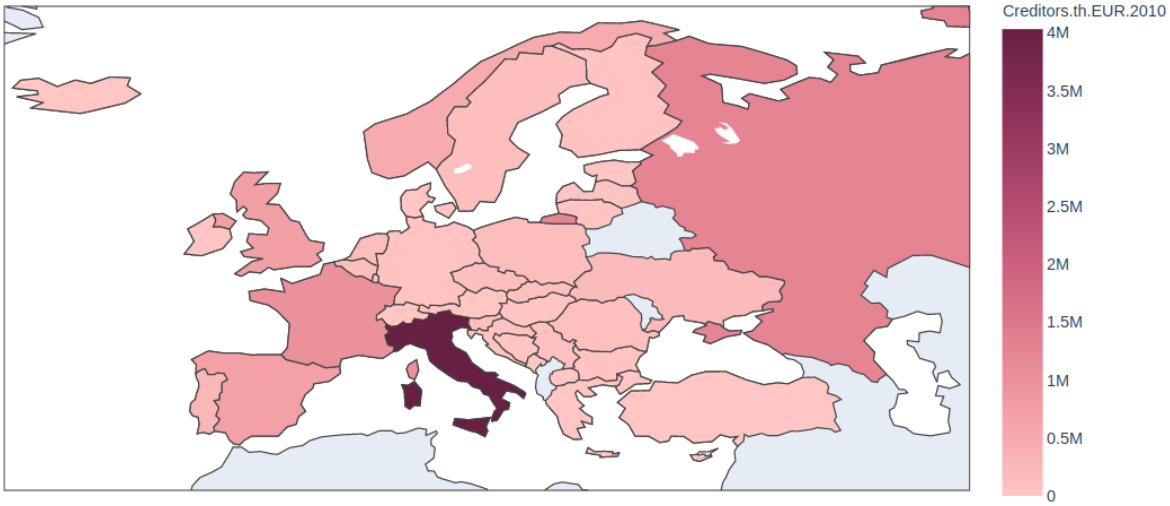
```
BvD.ID.number
IE480184          ESB FINANCE DESIGNATED ACTIVITY COMPANY
IE486122                      ICG EOS LOAN FUND I LIMITED
GB07193500           PREMIER LOTTERIES CAPITAL UK LIMITED
GB07202475       PREMIER LOTTERIES INVESTMENTS UK LIMITED
IT10319310016                    INFRATRASPORTI.TO S.R.L.
Name: Company.name, dtype: string
```

```
HGF vs non-HGF for Current.liabilities.th.EUR.2010
Welch's t-test statistic = 7.677
p-value = 1.691e-14

Optimization terminated successfully.
         Current function value: 0.155527
         Iterations 10
                           Logit Regression Results
==============================================================================
Dep. Variable:                    HGF   No. Observations:               115840
Model:                          Logit   Df Residuals:                   115838
Method:                           MLE   Df Model:                            1
Date:                Mon, 29 Jun 2020   Pseudo R-squ.:               0.0008644
Time:                        15:35:15   Log-Likelihood:                -18016.
converged:                       True   LL-Null:                       -18032.
Covariance Type:            nonrobust   LLR p-value:                 2.359e-08
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -3.2648      0.016   -202.670      0.000      -3.296      -3.233
CL          -9.74e-05   2.52e-05     -3.868      0.000      -0.000    -4.8e-05
==============================================================================
```

7.2.1.2.35. Creditors.th.EUR.2010



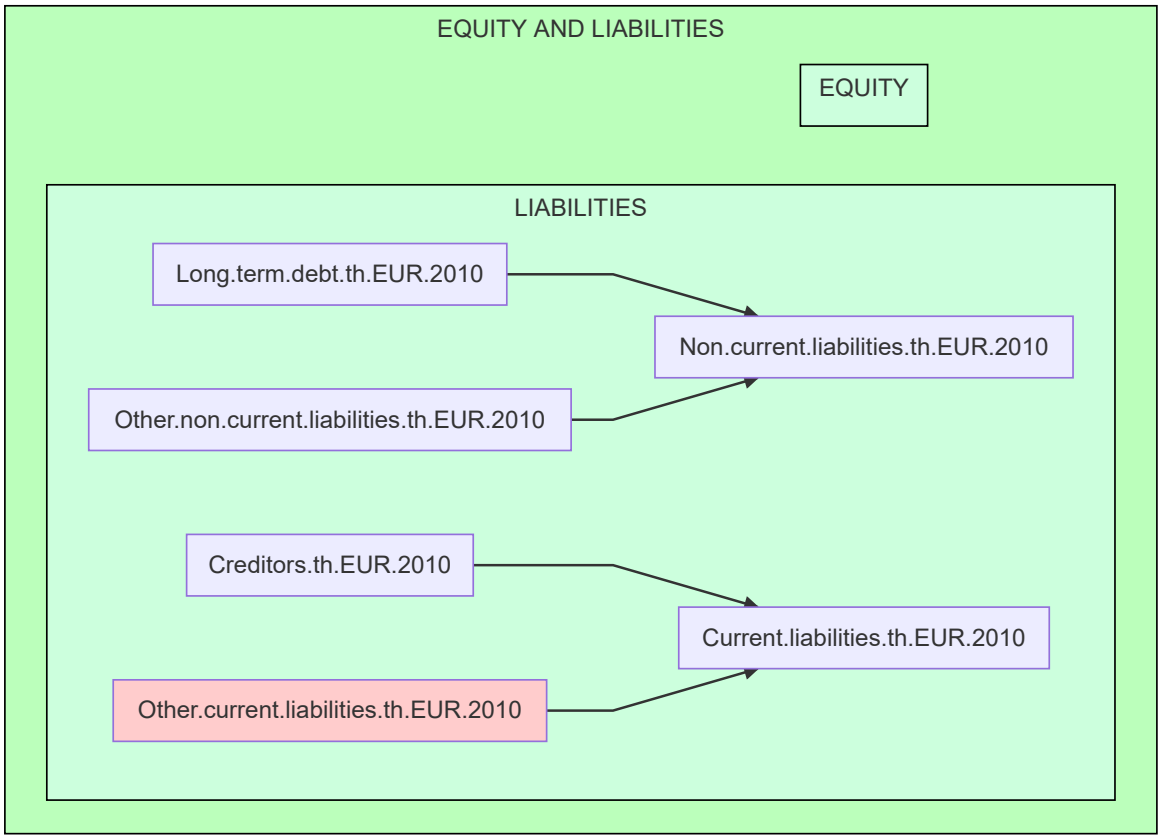


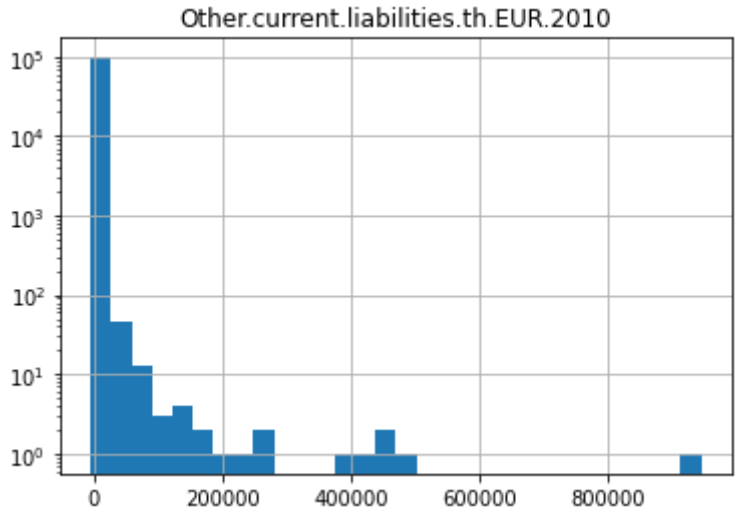


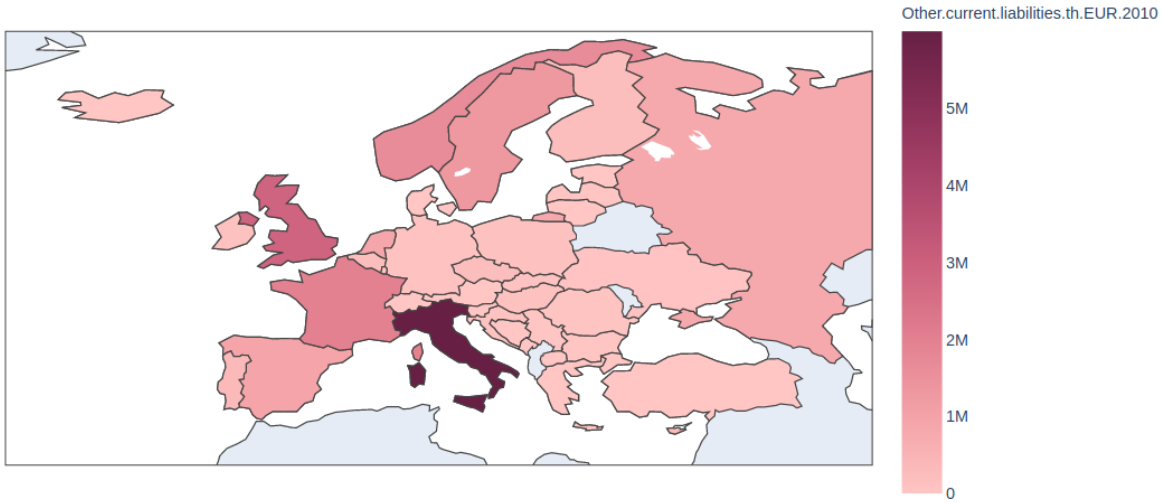
outliers:

```
BvD.ID.number
GB07254605        ED BROKING GROUP LIMITED
IT10969001006    LOTTERIE NAZIONALI S.R.L.
Name: Company.name, dtype: string

Optimization terminated successfully.
         Current function value: 0.149458
         Iterations 10
                           Logit Regression Results
==============================================================================
Dep. Variable:                    HGF   No. Observations:                98541
Model:                          Logit   Df Residuals:                    98539
Method:                           MLE   Df Model:                            1
Date:                Mon, 29 Jun 2020   Pseudo R-squ.:                0.001558
Time:                        15:35:36   Log-Likelihood:                -14728.
converged:                       True   LL-Null:                       -14751.
Covariance Type:            nonrobust   LLR p-value:                 1.211e-11
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -3.3077      0.018   -183.576      0.000      -3.343      -3.272
CR            -0.0005      0.000     -4.689      0.000      -0.001      -0.000
==============================================================================
```

7.2.1.2.36. Other.current.liabilities.th.EUR.2010


EQUITY AND LIABILITIES
EQUITY
LIABILITIES
Long.term.debt.th.EUR.2010
Non.current.liabilities.th.EUR.2010
Other.non.current.liabilities.th.EUR.2010
Creditors.th.EUR.2010
Current.liabilities.th.EUR.2010
Other.current.liabilities.th.EUR.2010

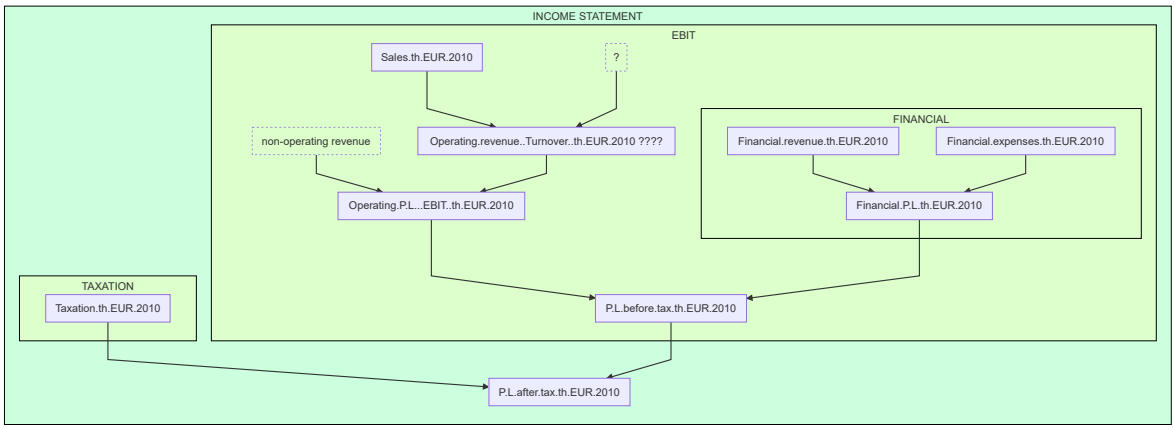

Other.current.liabilities.th.EUR.2010
10^5
10^4
10^3
10^2
10^1
10^0
0
200000
400000
600000
800000


Other.current.liabilities.th.EUR.2010
5M
4M
3M
2M
1M
0

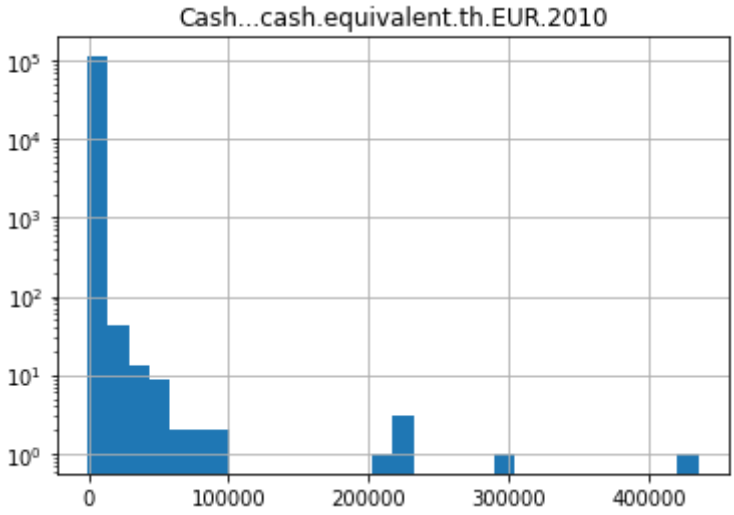
outlier:

```
BvD.ID.number
IT10319310016    INFRATRASPORTI.TO S.R.L.
Name: Company.name, dtype: string

Optimization terminated successfully.
         Current function value: 0.148949
         Iterations 10
                           Logit Regression Results
==============================================================================
Dep. Variable:                    HGF   No. Observations:                97463
Model:                          Logit   Df Residuals:                    97461
Method:                           MLE   Df Model:                            1
Date:                Mon, 29 Jun 2020   Pseudo R-squ.:               0.0006357
Time:                        15:36:15   Log-Likelihood:                -14517.
converged:                       True   LL-Null:                       -14526.
Covariance Type:            nonrobust   LLR p-value:                 1.727e-05
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -3.3277      0.018   -185.296      0.000      -3.363      -3.292
OCL           -0.0001   4.46e-05     -3.012      0.003      -0.000   -4.69e-05
==============================================================================
```

Income statement graph:



7.2.1.2.37. Cash...cash.equivalent.th.EUR.2010



outliers:

400+K bin:

```
BvD.ID.number
NO995216604    WALLENIUS WILHELMSEN ASA
Name: Company.name, dtype: string
```

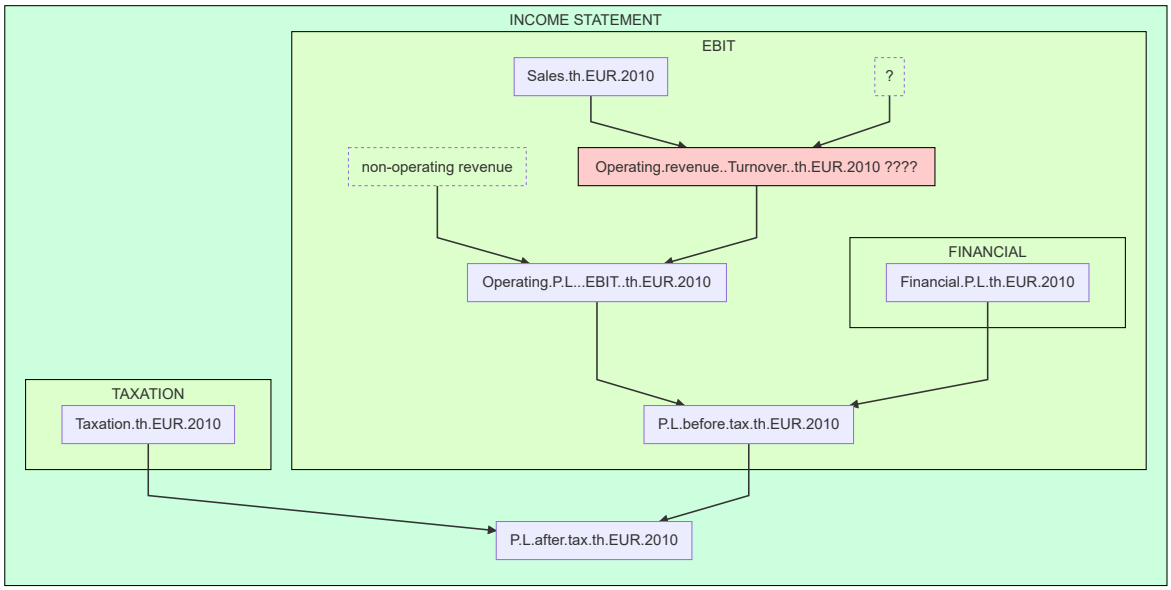
300K bin:

```
BvD.ID.number
GB07123187            ACACIA MINING PLC
Name: Company.name, dtype: string
```

200K+ bin:

```
BvD.ID.number
BE0831465984                            XIX-INVEST
GB07145051      CAPITAL & COUNTIES PROPERTIES PLC
GB07254605               ED BROKING GROUP LIMITED
GB07283266              HIGHBRIDGE COBALT LIMITED
Name: Company.name, dtype: string

Optimization terminated successfully.
         Current function value: 0.157093
         Iterations 9
                           Logit Regression Results
==============================================================================
Dep. Variable:                    HGF   No. Observations:               110446
Model:                          Logit   Df Residuals:                   110444
Method:                           MLE   Df Model:                            1
Date:                Mon, 29 Jun 2020   Pseudo R-squ.:              0.0001870
Time:                        15:36:36   Log-Likelihood:                -17350.
converged:                       True   LL-Null:                       -17354.
Covariance Type:            nonrobust   LLR p-value:                   0.01086
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -3.2645      0.016   -201.981      0.000      -3.296      -3.233
OCL        -9.186e-05   5.11e-05     -1.797      0.072      -0.000    8.33e-06
==============================================================================
```

7.2.1.2.38. Operating.revenue..Turnover..th.EUR.2010



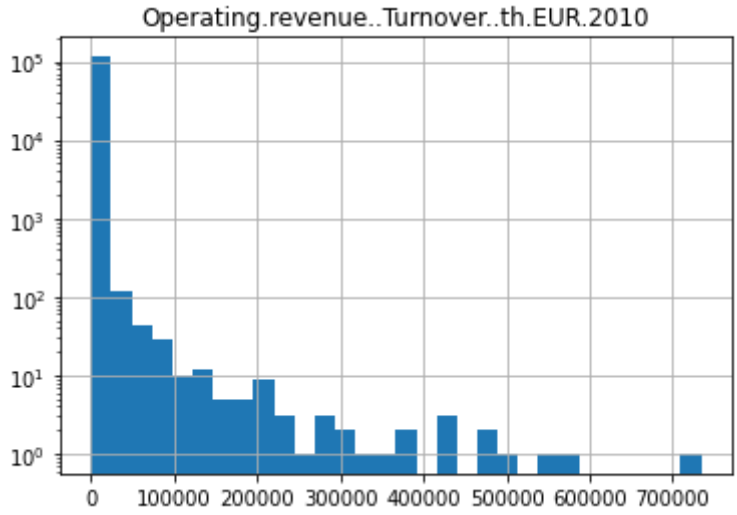
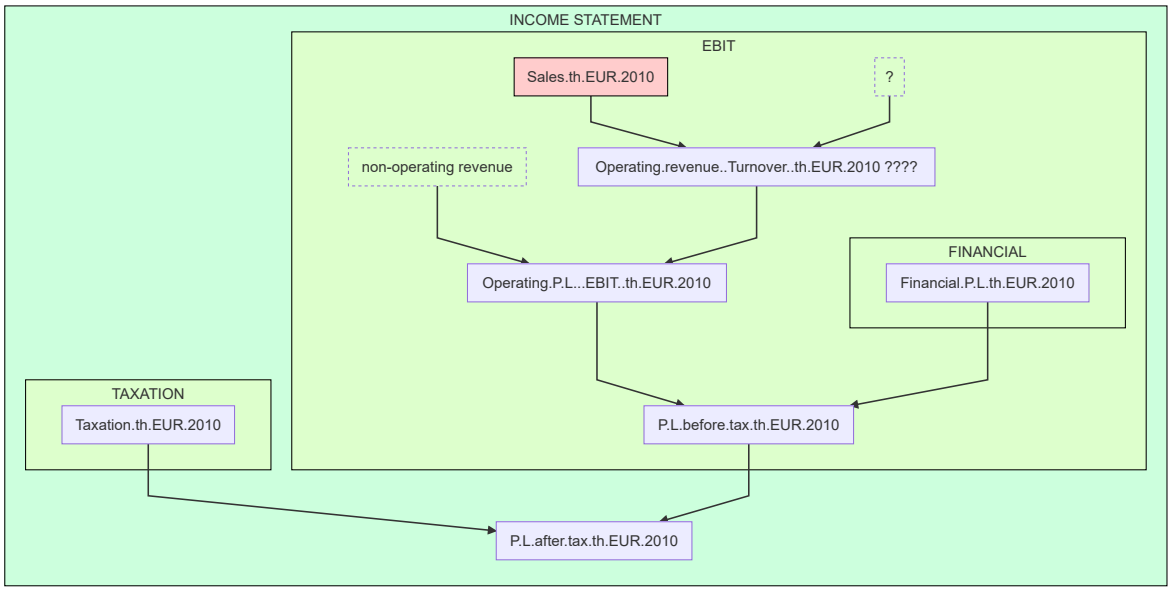


image-20200624143439652

```
HGF vs non-HGF for Operating.revenue..Turnover..th.EUR.2010
Welch's t-test statistic = 9.251
p-value = 2.661e-20

Optimization terminated successfully.
         Current function value: 0.155219
         Iterations 11
                           Logit Regression Results
==============================================================================
Dep. Variable:                    HGF   No. Observations:                115840
Model:                          Logit   Df Residuals:                    115838
Method:                           MLE   Df Model:                             1
Date:                Mon, 29 Jun 2020   Pseudo R-squ.:                0.002844
Time:                        15:37:05   Log-Likelihood:                -17981.
converged:                       True   LL-Null:                       -18032.
Covariance Type:            nonrobust   LLR p-value:                 4.183e-24
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
```

```
------------------------------------------------------------------------------
Intercept     -3.2318      0.017   -193.739      0.000      -3.265      -3.199
ORT           -0.0003   4.57e-05     -6.834      0.000      -0.000      -0.000
==============================================================================
```

7.2.1.2.39. Sales.th.EUR.2010



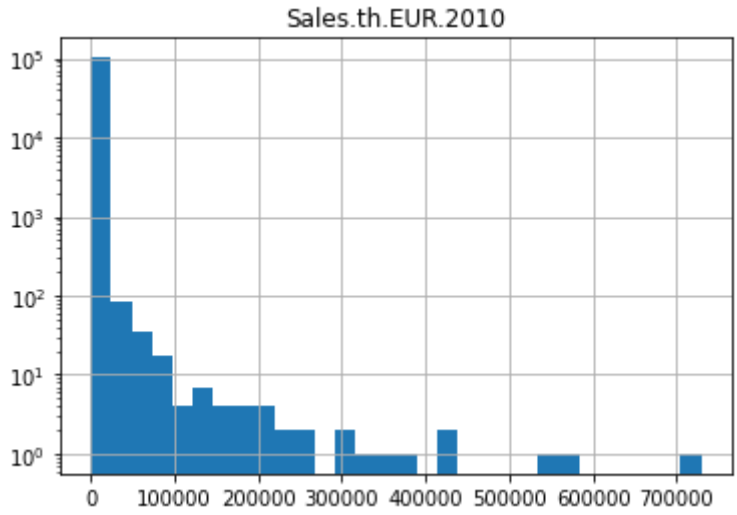


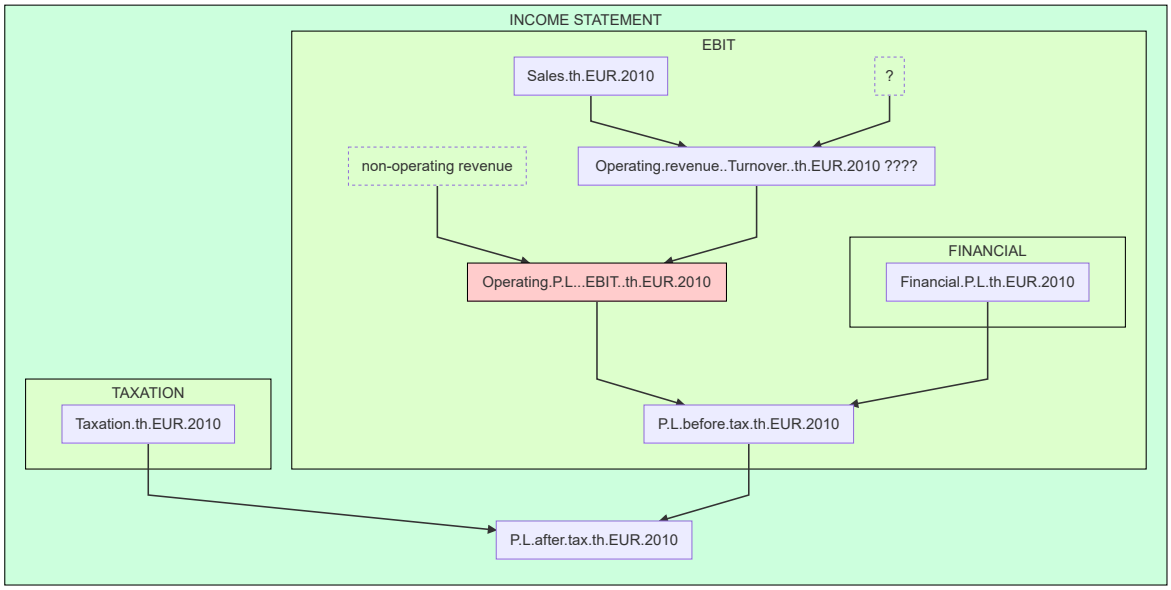
image-20200624143528781

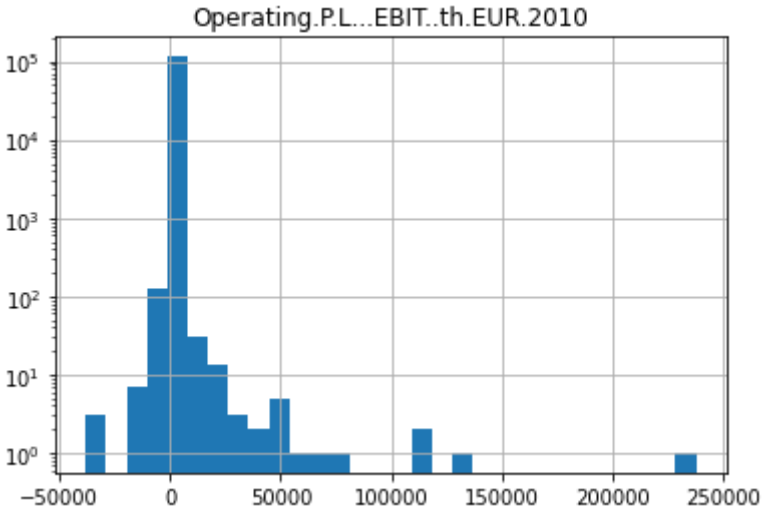
outliers:

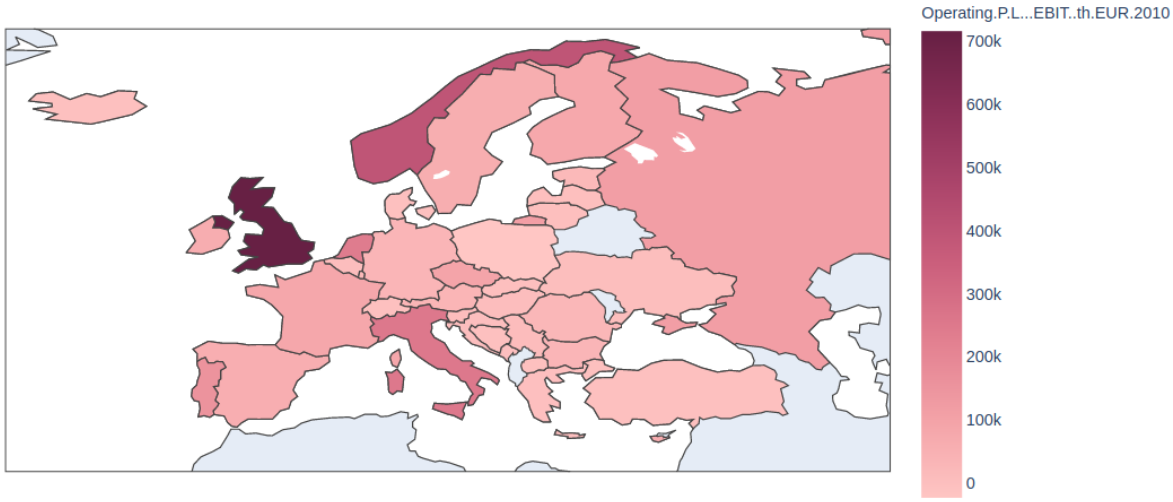
```
BvD.ID.number
CYC266578        HMS HYDRAULIC MACHINES & SYSTEMS GROUP PLC
GB07123187                                 ACACIA MINING PLC
IT10813301008                                     EOS S.R.L.
Name: Company.name, dtype: string

Optimization terminated successfully.
         Current function value: 0.160957
         Iterations 11
                           Logit Regression Results
```

```
==============================================================================
Dep. Variable:                    HGF   No. Observations:                102550
Model:                          Logit   Df Residuals:                    102548
Method:                           MLE   Df Model:                             1
Date:                Mon, 29 Jun 2020   Pseudo R-squ.:                0.005149
Time:                        15:37:31   Log-Likelihood:                -16506.
converged:                       True   LL-Null:                       -16592.
Covariance Type:            nonrobust   LLR p-value:                 4.804e-39
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -3.1452      0.018   -177.808      0.000      -3.180      -3.111
SAL           -0.0007   8.11e-05     -9.127      0.000      -0.001      -0.001
==============================================================================
```

7.2.1.2.40. Operating.P.L...EBIT..th.EUR.2010







Positive outliers:

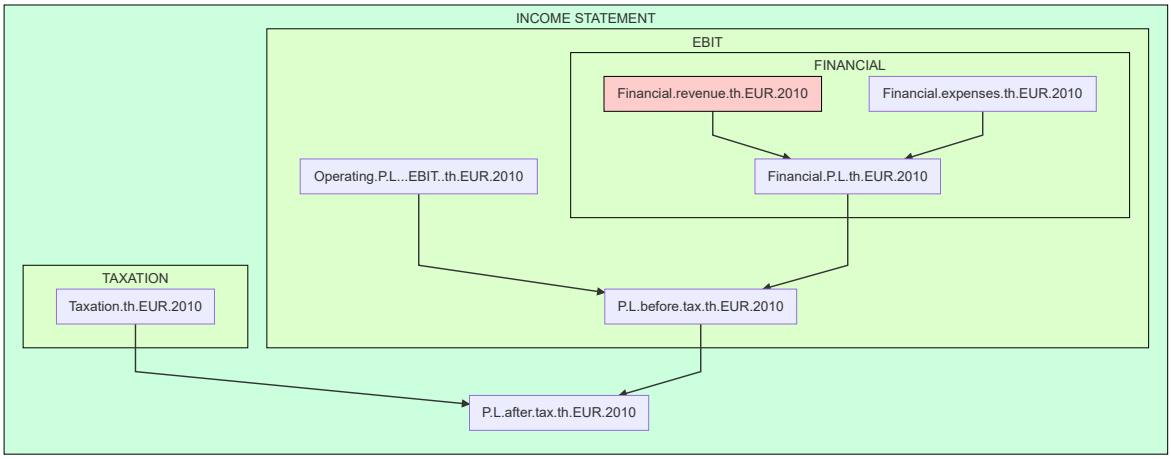
```
BvD.ID.number
GB07123187                    ACACIA MINING PLC
NL50397931     ATLANTIC AURUM INVESTMENTS B.V.
NO995216604            WALLENIUS WILHELMSEN ASA
PT509444229     MOTA-ENGIL AFRICA - SGPS, S.A.
Name: Company.name, dtype: string
```

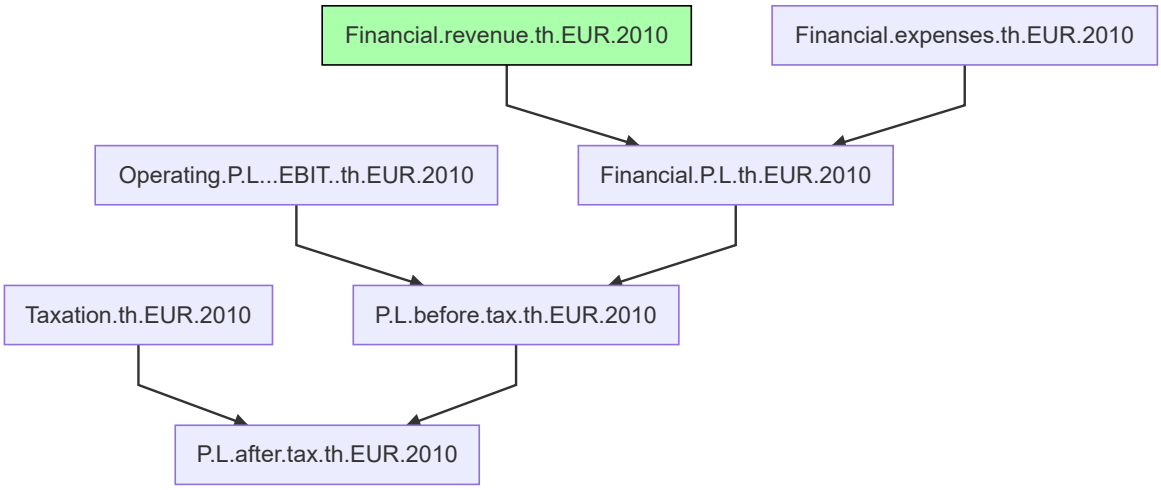
Negative outliers:

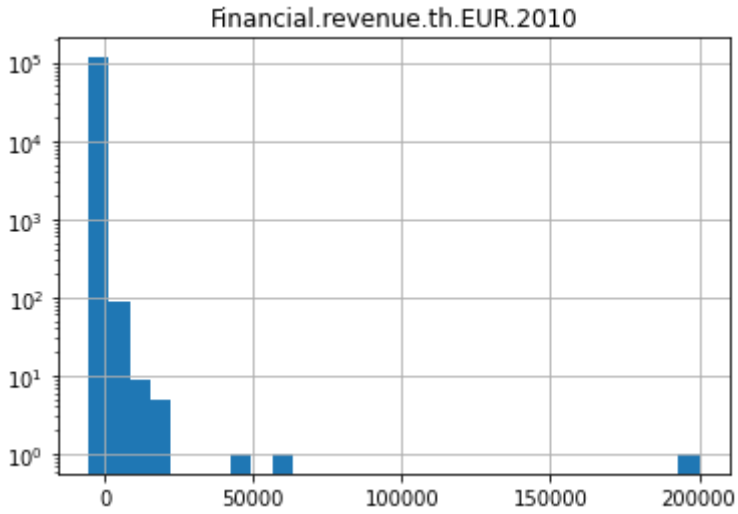
```
BvD.ID.number
IE507678             HORIZON THERAPEUTICS PUBLIC LIMITED COMPANY
PL301339040    SAMSUNG ELECTRONICS POLAND MANUFACTURING SP. Z...
RU65519055                  AKTSIONERNOE OBSHCHESTVO TATTEPLOSBYT
Name: Company.name, dtype: string
```

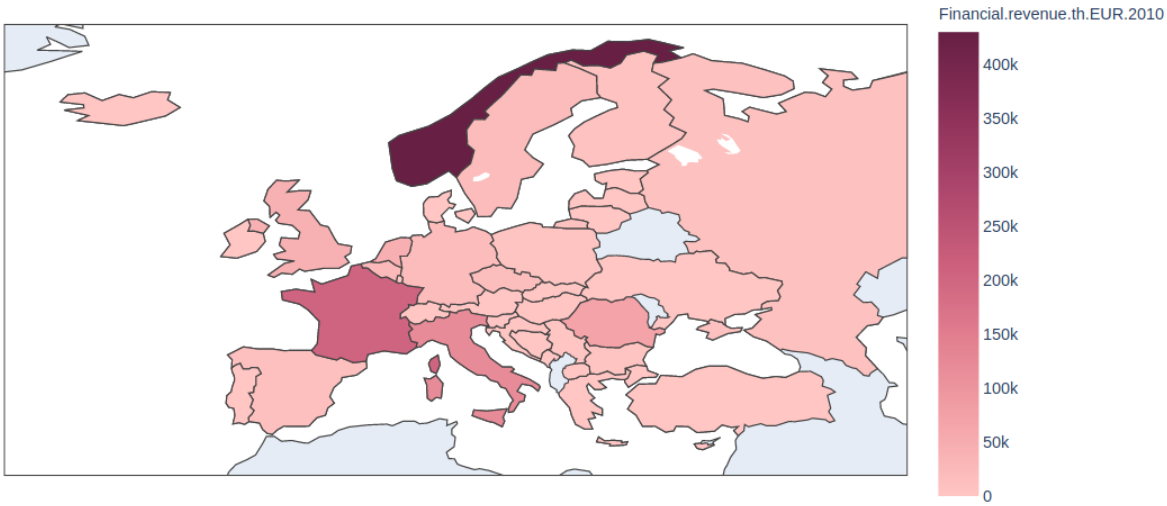
```
HGF vs non-HGF for Operating.P.L...EBIT..th.EUR.2010
Welch's t-test statistic = 4.539
p-value = 5.771e-06

Optimization terminated successfully.
         Current function value: 0.155585
         Iterations 7
                           Logit Regression Results
==============================================================================
Dep. Variable:                    HGF   No. Observations:               115840
Model:                          Logit   Df Residuals:                   115838
Method:                           MLE   Df Model:                            1
Date:                Mon, 29 Jun 2020   Pseudo R-squ.:               0.0004971
Time:                        15:37:58   Log-Likelihood:                -18023.
converged:                       True   LL-Null:                       -18032.
Covariance Type:            nonrobust   LLR p-value:                 2.296e-05
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -3.2814      0.016   -208.555      0.000      -3.312      -3.251
EBIT          -0.0001   3.08e-05     -4.079      0.000      -0.000   -6.54e-05
==============================================================================
```

7.2.1.2.41. Financial.revenue.th.EUR.2010


INCOME STATEMENT
EBIT
FINANCIAL
Financial.revenue.th.EUR.2010
Financial.expenses.th.EUR.2010
Financial.P.L.th.EUR.2010
Operating.P.L...EBIT..th.EUR.2010
TAXATION
Taxation.th.EUR.2010
P.L.before.tax.th.EUR.2010
P.L.after.tax.th.EUR.2010


Financial.revenue.th.EUR.2010
Financial.expenses.th.EUR.2010
Operating.P.L...EBIT..th.EUR.2010
Financial.P.L.th.EUR.2010
Taxation.th.EUR.2010
P.L.before.tax.th.EUR.2010
P.L.after.tax.th.EUR.2010


Financial.revenue.th.EUR.2010
10^5
10^4
10^3
10^2
10^1
10^0
0
50000
100000
150000
200000



outliers:

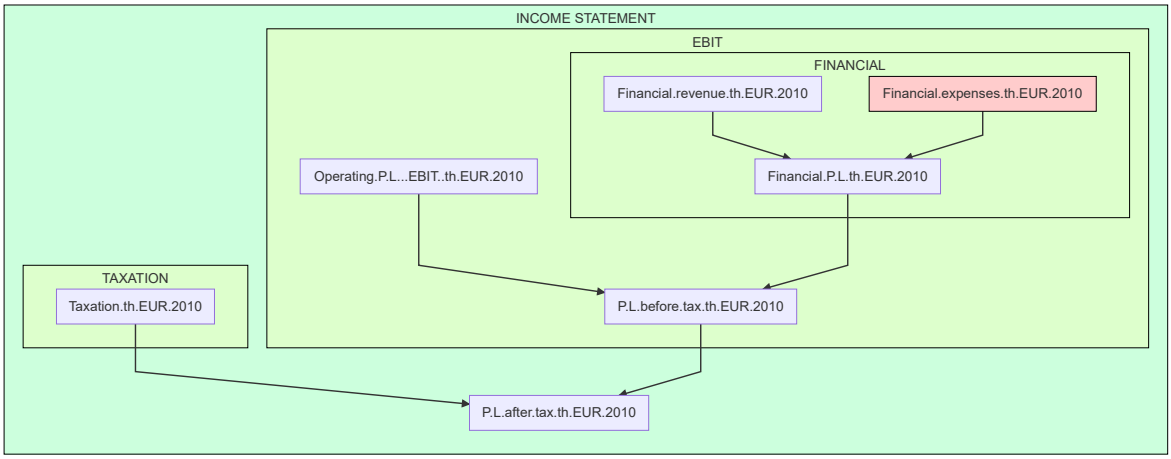
50K cluster:

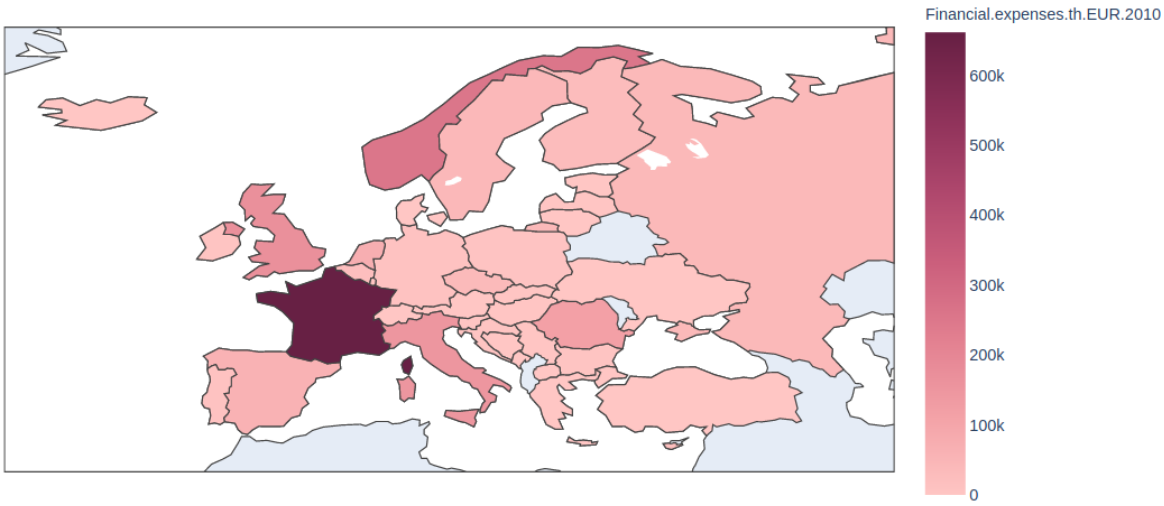
```
BvD.ID.number
FR527925143                      SOFAQUE
NO996031454           DOLPHIN INVEST AS
Name: Company.name, dtype: string
```

200K cluster:

```
BvD.ID.number
NO995633604    INDUSTRIINVESTERINGER AS
Name: Company.name, dtype: string

Optimization terminated successfully.
         Current function value: 0.155644
         Iterations 9
                           Logit Regression Results
==============================================================================
Dep. Variable:                    HGF   No. Observations:               115840
Model:                          Logit   Df Residuals:                   115838
Method:                           MLE   Df Model:                            1
Date:                Mon, 29 Jun 2020   Pseudo R-squ.:              0.0001129
Time:                        15:38:23   Log-Likelihood:                -18030.
converged:                       True   LL-Null:                       -18032.
Covariance Type:            nonrobust   LLR p-value:                   0.04361
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -3.2807      0.016   -208.461      0.000      -3.312      -3.250
FR            -0.0003      0.000     -1.646      0.100      -0.001    5.98e-05
==============================================================================
```

7.2.1.2.42. Financial.expenses.th.EUR.2010





positive outlier:

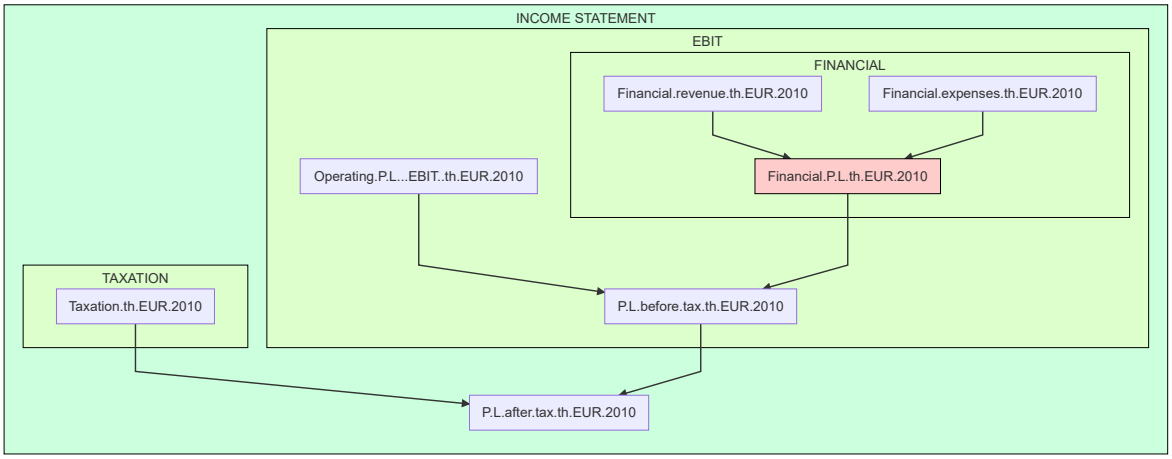
```
BvD.ID.number
FR519720643    IRIDIUM FRANCE
Name: Company.name, dtype: string
```

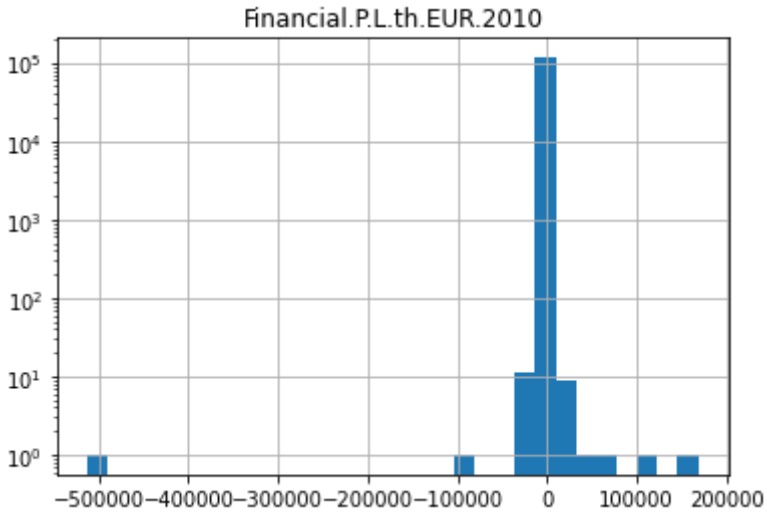
negative outlier:

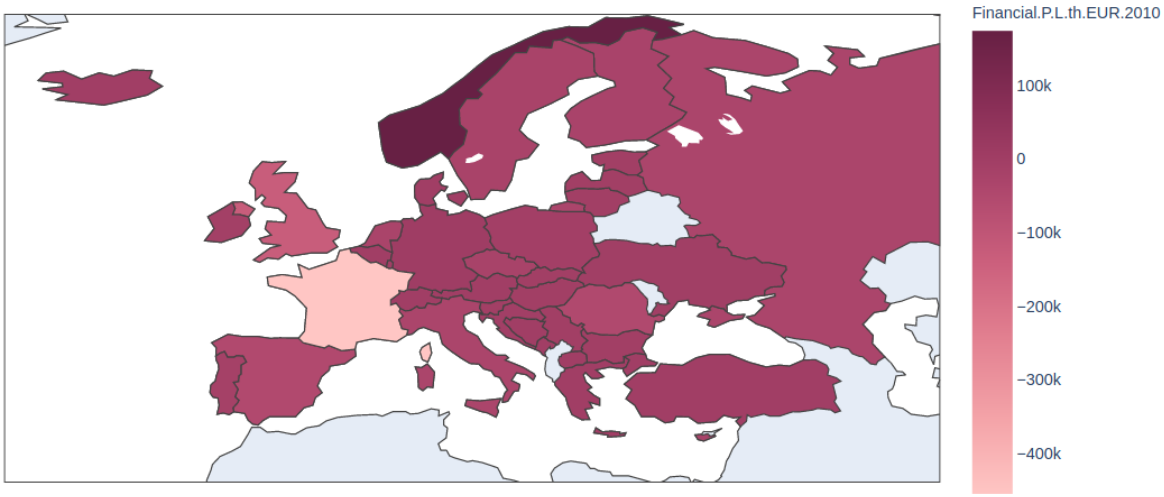
```
BvD.ID.number
GB07145051    CAPITAL & COUNTIES PROPERTIES PLC
Name: Company.name, dtype: string
```

```
HGF vs non-HGF for Financial.expenses.th.EUR.2010
Welch's t-test statistic = 2.789
p-value = 0.005292

Optimization terminated successfully.
         Current function value: 0.155655
         Iterations 7
                           Logit Regression Results
==============================================================================
Dep. Variable:                    HGF   No. Observations:               115840
Model:                          Logit   Df Residuals:                   115838
Method:                           MLE   Df Model:                            1
Date:                Mon, 29 Jun 2020   Pseudo R-squ.:               4.376e-05
Time:                        15:38:43   Log-Likelihood:                -18031.
converged:                       True   LL-Null:                       -18032.
Covariance Type:            nonrobust   LLR p-value:                    0.2090
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -3.2817      0.016   -208.616      0.000      -3.313      -3.251
FE         -2.787e-05   1.96e-05     -1.425      0.154   -6.62e-05    1.05e-05
==============================================================================
```

7.2.1.2.43. Financial.P.L.th.EUR.2010





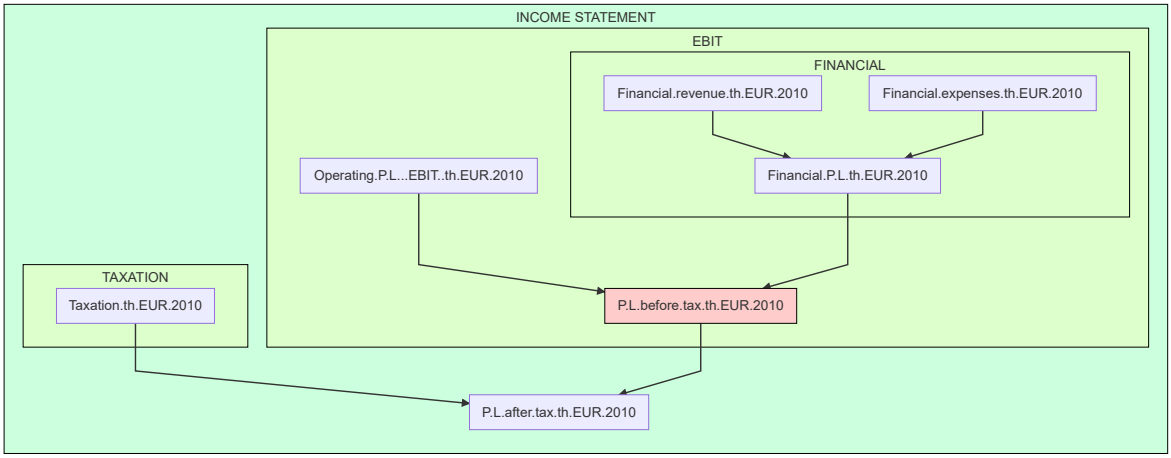


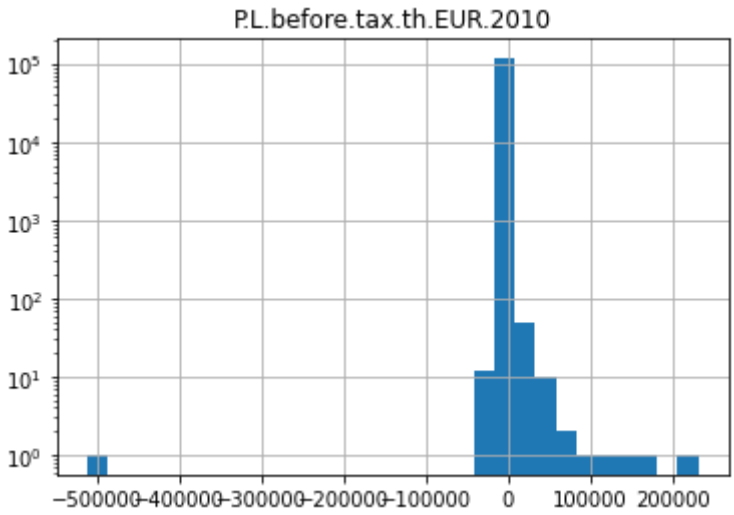
Outlier:

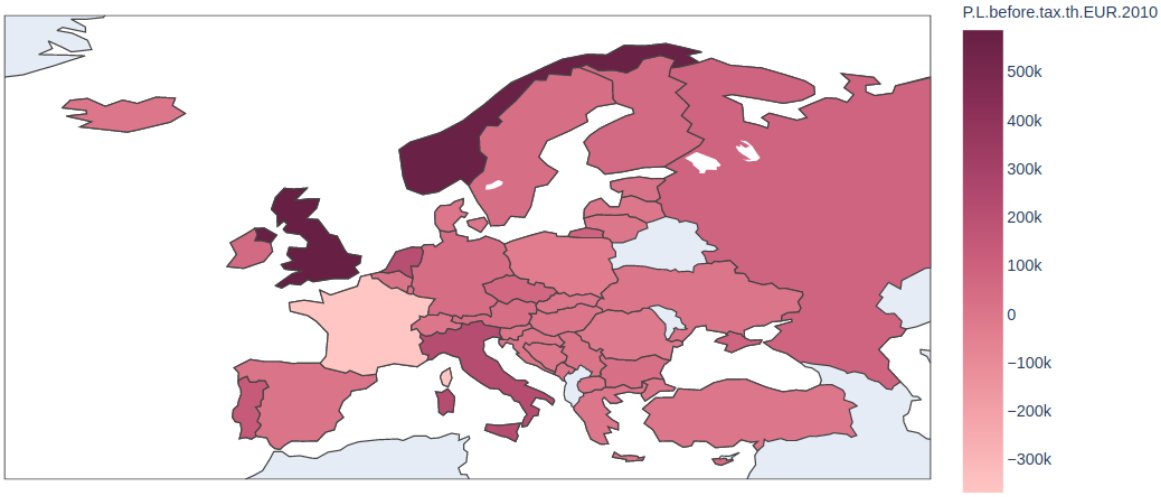
```
BvD.ID.number
FR519720643    IRIDIUM FRANCE
Name: Company.name, dtype: string
```

```
HGF vs non-HGF for Financial.P.L.th.EUR.2010
Welch's t-test statistic =  -1.5
p-value = 0.1337

Optimization terminated successfully.
         Current function value: 0.155661
         Iterations 7
                           Logit Regression Results
==============================================================================
Dep. Variable:                    HGF   No. Observations:               115840
Model:                          Logit   Df Residuals:                   115838
Method:                           MLE   Df Model:                            1
Date:                Mon, 29 Jun 2020   Pseudo R-squ.:               5.126e-06
Time:                        15:39:06   Log-Likelihood:                -18032.
converged:                       True   LL-Null:                       -18032.
Covariance Type:            nonrobust   LLR p-value:                    0.6672
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -3.2819      0.016   -208.633      0.000      -3.313      -3.251
FPL         6.226e-06   1.53e-05      0.408      0.684   -2.37e-05    3.62e-05
==============================================================================
```

7.2.1.2.44. P.L.before.tax.th.EUR.2010





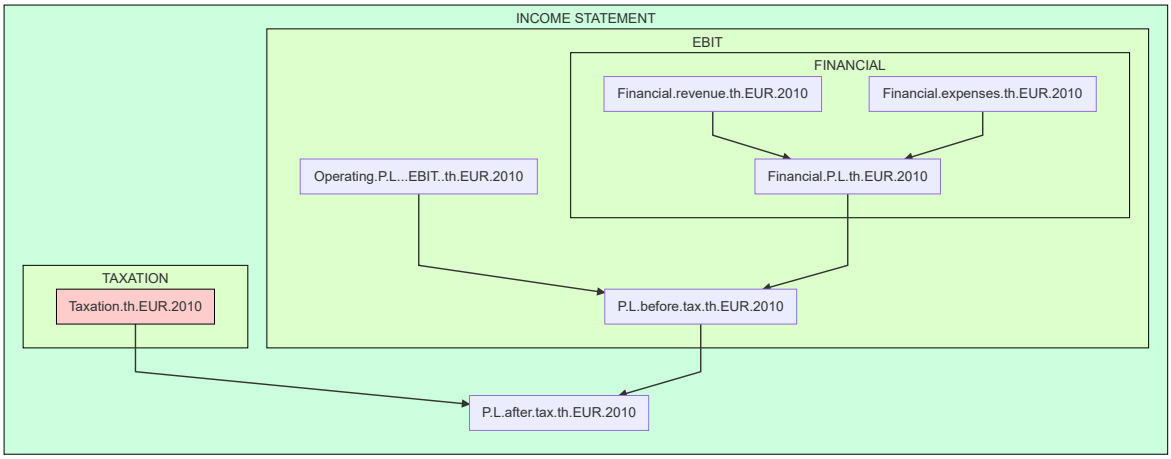


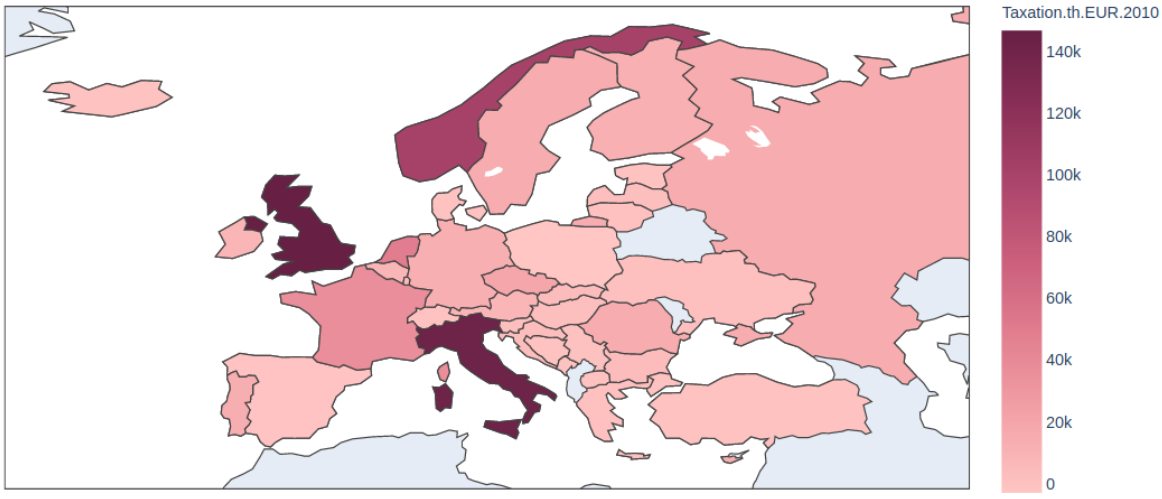
outlier:

```
BvD.ID.number
FR519720643    IRIDIUM FRANCE
Name: Company.name, dtype: string
```

```
HGF vs non-HGF for P.L.before.tax.th.EUR.2010
Welch's t-test statistic = 3.793
p-value = 0.000149

Optimization terminated successfully.
         Current function value: 0.155659
         Iterations 7
                           Logit Regression Results
==============================================================================
Dep. Variable:                    HGF   No. Observations:                115840
Model:                          Logit   Df Residuals:                    115838
Method:                           MLE   Df Model:                             1
Date:                Mon, 29 Jun 2020   Pseudo R-squ.:               1.779e-05
Time:                        15:39:22   Log-Likelihood:                -18032.
converged:                       True   LL-Null:                       -18032.
Covariance Type:            nonrobust   LLR p-value:                    0.4232
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -3.2819      0.016   -208.631      0.000      -3.313      -3.251
FPL        -4.184e-06   4.37e-06     -0.957      0.339   -1.28e-05    4.39e-06
==============================================================================
```

7.2.1.2.45. Taxation.th.EUR.2010





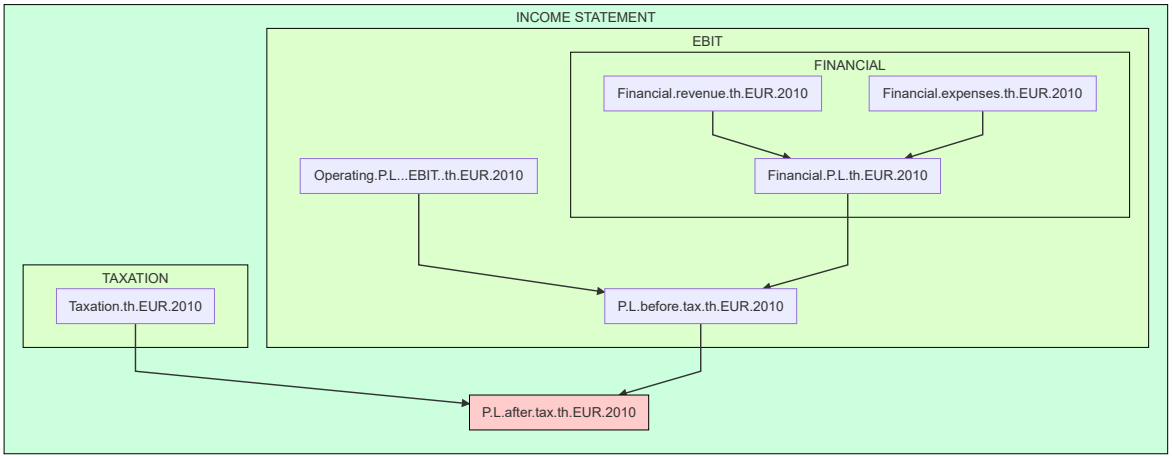
outlier:

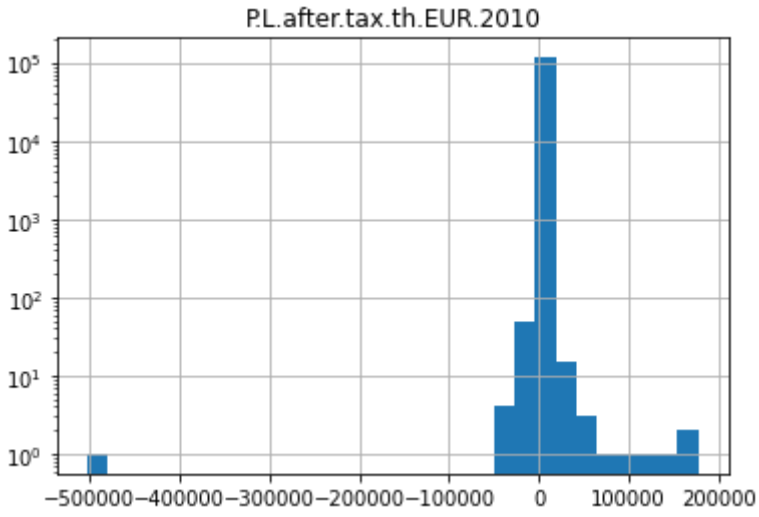
```
BvD.ID.number
GB07123187    ACACIA MINING PLC
Name: Company.name, dtype: string
```

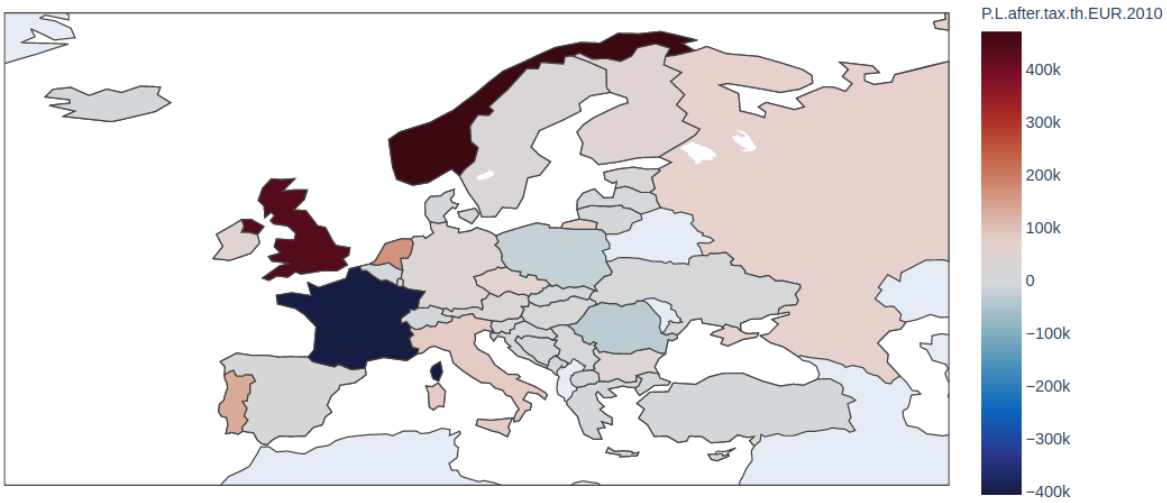
```
HGF vs non-HGF for Taxation.th.EUR.2010
Welch's t-test statistic = 5.486
```

```
p-value = 4.128e-08

Optimization terminated successfully.
         Current function value: 0.155649
         Iterations 7
                           Logit Regression Results
==============================================================================
Dep. Variable:                    HGF   No. Observations:               115840
Model:                          Logit   Df Residuals:                   115838
Method:                           MLE   Df Model:                            1
Date:                Mon, 29 Jun 2020   Pseudo R-squ.:               8.533e-05
Time:                        15:39:38   Log-Likelihood:                -18030.
converged:                       True   LL-Null:                       -18032.
Covariance Type:            nonrobust   LLR p-value:                    0.07940
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -3.2814      0.016   -208.591      0.000      -3.312      -3.251
FPL           -0.0002      0.000     -1.794      0.073      -0.000    1.68e-05
==============================================================================
```

7.2.1.2.46. P.L.after.tax.th.EUR.2010





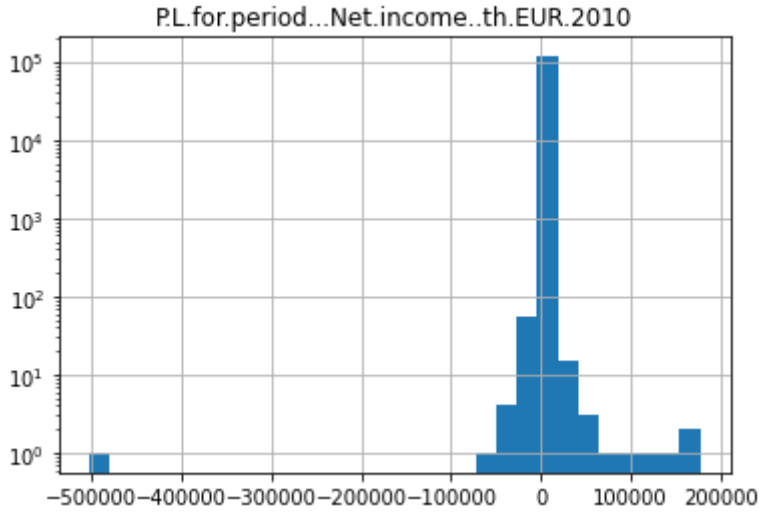


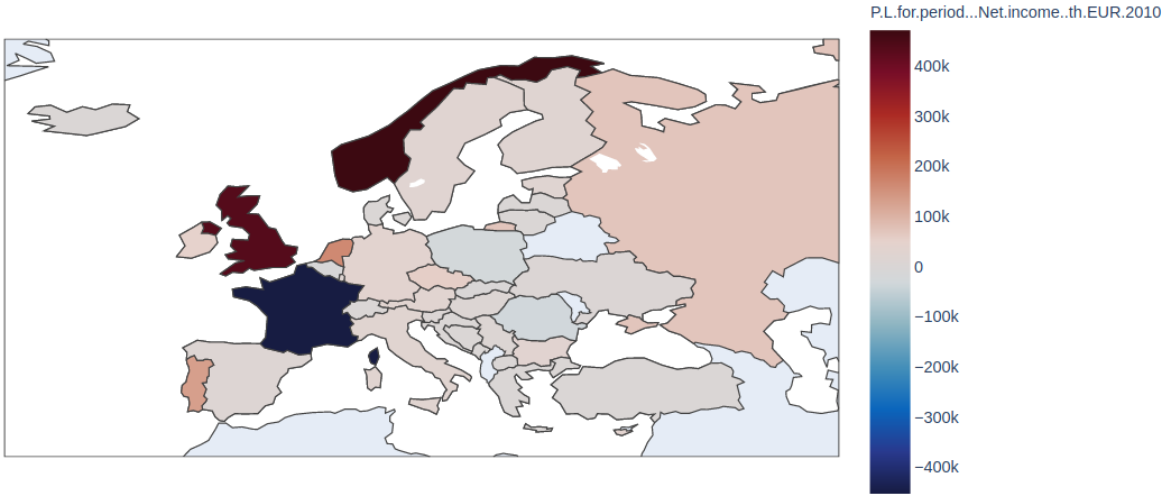
outlier:

```
BvD.ID.number
FR519720643    IRIDIUM FRANCE
Name: Company.name, dtype: string
```

```
HGF vs non-HGF for P.L.after.tax.th.EUR.2010
Welch's t-test statistic = 3.352
p-value = 0.0008042

Optimization terminated successfully.
         Current function value: 0.155660
         Iterations 7
                           Logit Regression Results
==============================================================================
Dep. Variable:                    HGF   No. Observations:               115840
Model:                          Logit   Df Residuals:                   115838
Method:                           MLE   Df Model:                            1
Date:                Mon, 29 Jun 2020   Pseudo R-squ.:               1.429e-05
Time:                        15:40:03   Log-Likelihood:                -18032.
converged:                       True   LL-Null:                       -18032.
Covariance Type:            nonrobust   LLR p-value:                    0.4728
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -3.2819      0.016   -208.632      0.000      -3.313      -3.251
PL         -3.991e-06   4.65e-06     -0.858      0.391   -1.31e-05    5.13e-06
==============================================================================
```

7.2.1.2.47. P.L.for.period...Net.income..th.EUR.2010



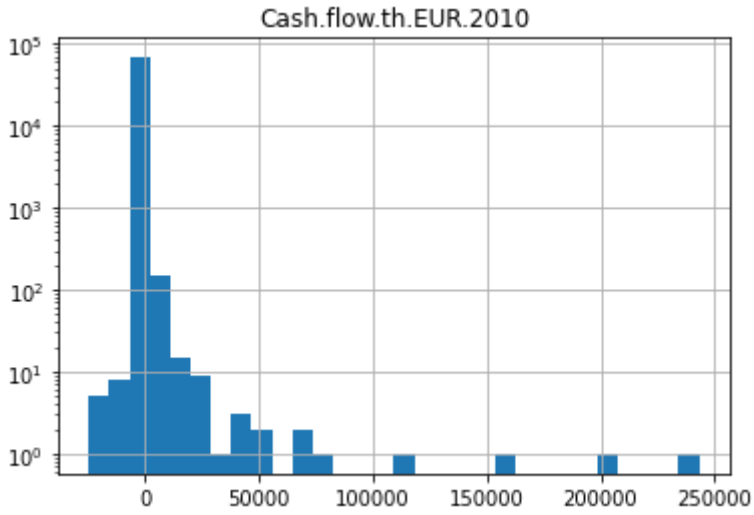


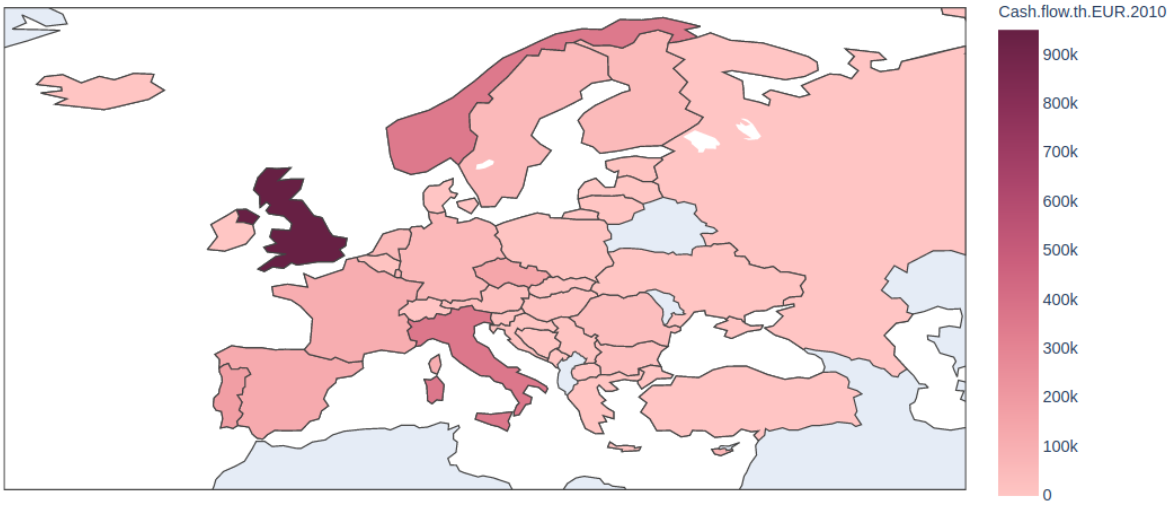
outlier:

```
BvD.ID.number
FR519720643    IRIDIUM FRANCE
Name: Company.name, dtype: string
```

```
HGF vs non-HGF for P.L.for.period...Net.income..th.EUR.2010
Welch's t-test statistic = 3.123
p-value = 0.001791

Optimization terminated successfully.
         Current function value: 0.155660
         Iterations 7
                           Logit Regression Results
==============================================================================
Dep. Variable:                    HGF   No. Observations:               115840
Model:                          Logit   Df Residuals:                   115838
Method:                           MLE   Df Model:                            1
Date:                Mon, 29 Jun 2020   Pseudo R-squ.:               1.267e-05
Time:                        15:40:20   Log-Likelihood:                -18032.
converged:                       True   LL-Null:                       -18032.
Covariance Type:            nonrobust   LLR p-value:                    0.4991
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -3.2819      0.016   -208.632      0.000      -3.313      -3.251
PL          -3.82e-06   4.75e-06     -0.804      0.421   -1.31e-05    5.49e-06
==============================================================================
```

7.2.1.2.48. Cash.flow.th.EUR.2010

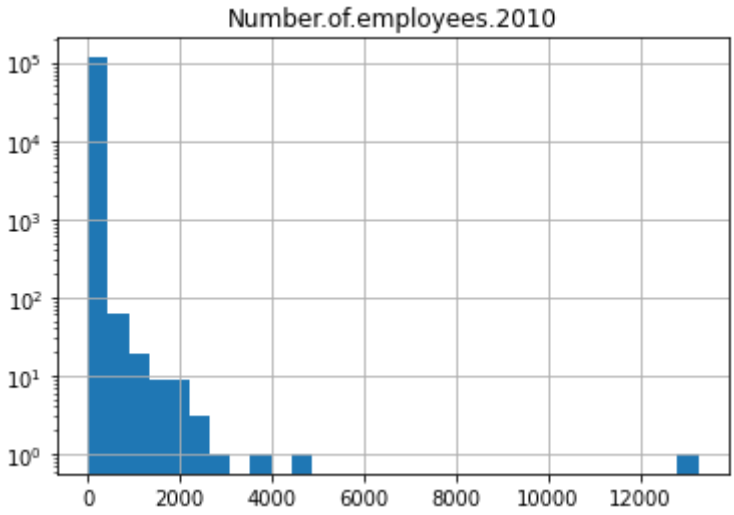


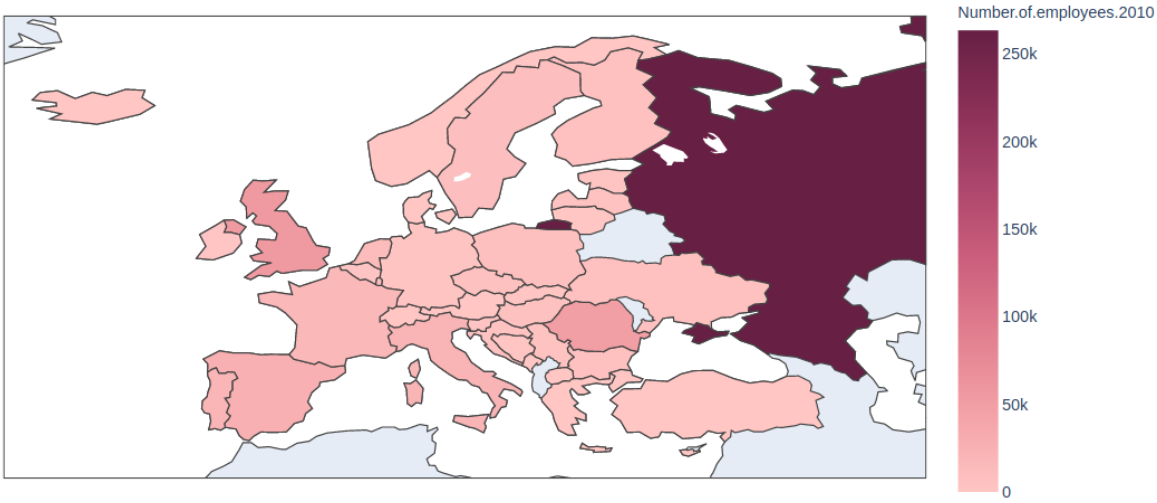


outliers:

```
BvD.ID.number
GB07123187                       ACACIA MINING PLC
GB07145051     CAPITAL & COUNTIES PROPERTIES PLC
GB07140891                           ENQUEST PLC
PT509444229       MOTA-ENGIL AFRICA - SGPS, S.A.
Name: Company.name, dtype: string

Optimization terminated successfully.
         Current function value: 0.162521
         Iterations 7
                           Logit Regression Results
==============================================================================
Dep. Variable:                    HGF   No. Observations:                68021
Model:                          Logit   Df Residuals:                    68019
Method:                           MLE   Df Model:                            1
Date:                Mon, 29 Jun 2020   Pseudo R-squ.:              0.0007085
Time:                        15:40:47   Log-Likelihood:                -11055.
converged:                       True   LL-Null:                       -11063.
Covariance Type:            nonrobust   LLR p-value:                 7.517e-05
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -3.2206      0.020   -161.248      0.000      -3.260      -3.181
CF            -0.0001   3.46e-05     -4.110      0.000      -0.000   -7.44e-05
==============================================================================

```

7.2.1.2.49. Number.of.employees.2010



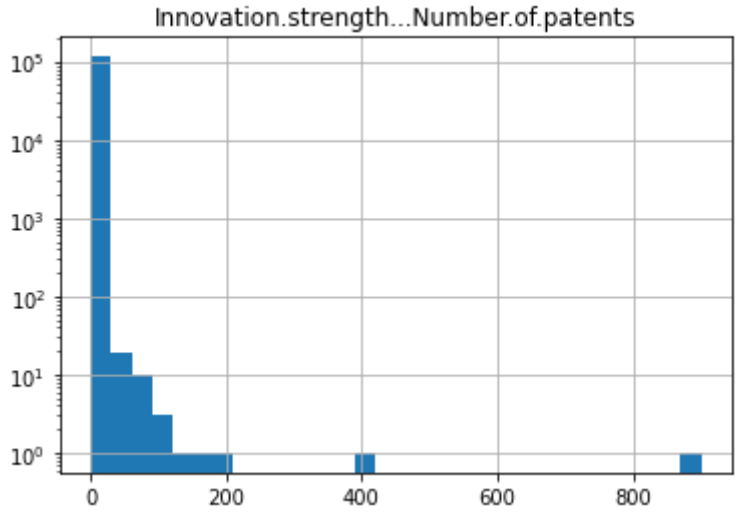


outlier:

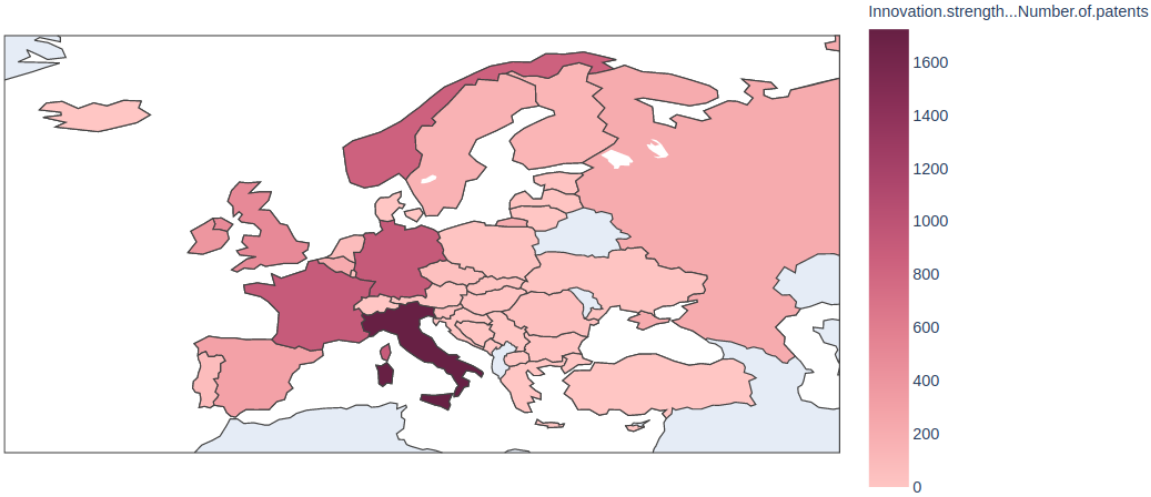
```
BvD.ID.number
GB07158140    CARE UK HEALTH & SOCIAL CARE INVESTMENTS LIMITED
Name: Company.name, dtype: string
```

```
HGF vs non-HGF for Number.of.employees.2010
Welch's t-test statistic =  8.77
p-value = 2.022e-18

Optimization terminated successfully.
         Current function value: 0.155502
         Iterations 9
                           Logit Regression Results
==============================================================================
Dep. Variable:                    HGF   No. Observations:               115840
Model:                          Logit   Df Residuals:                   115838
Method:                           MLE   Df Model:                            1
Date:                Mon, 29 Jun 2020   Pseudo R-squ.:                0.001025
Time:                        15:41:13   Log-Likelihood:                -18013.
converged:                       True   LL-Null:                       -18032.
Covariance Type:            nonrobust   LLR p-value:                 1.198e-09
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -3.2594      0.016   -201.417      0.000      -3.291      -3.228
E             -0.0070      0.001     -4.796      0.000      -0.010      -0.004
```

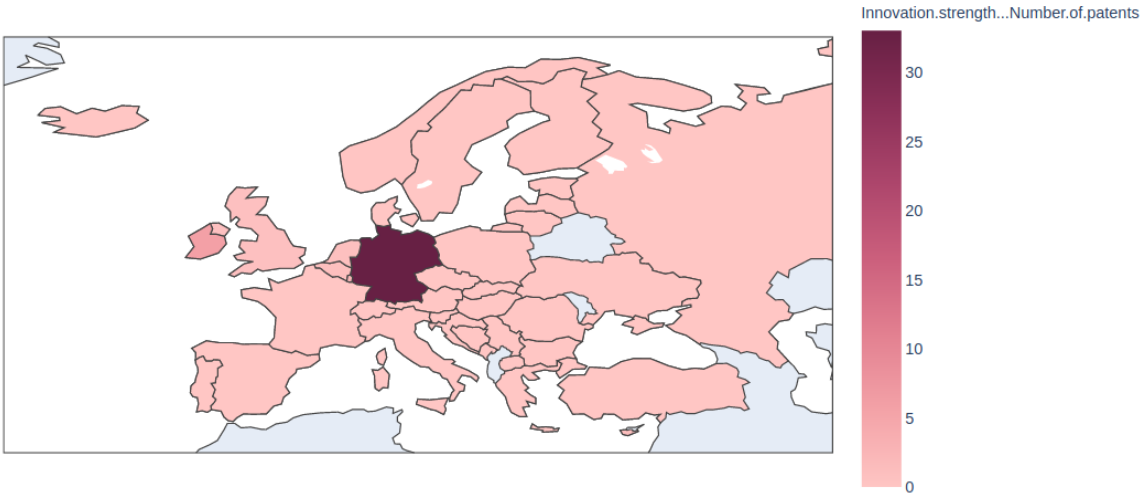
7.2.1.2.50. Innovation.strength...Number.of.patents



sum:



mean:

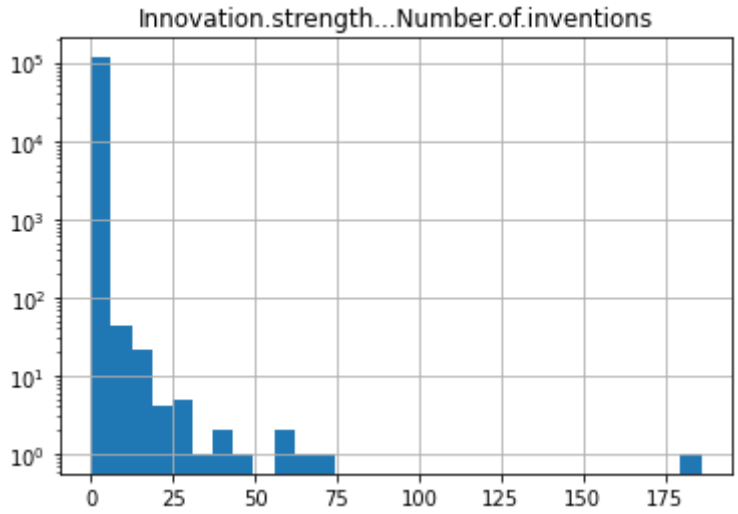


outliers:

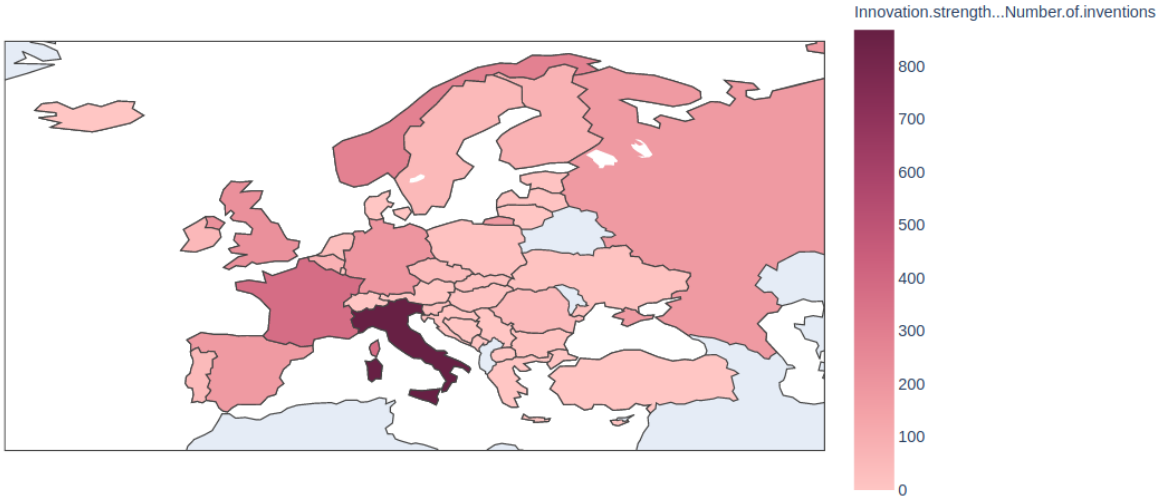
```
BvD.ID.number
IE507678        HORIZON THERAPEUTICS PUBLIC LIMITED COMPANY
DE8190460728                                    WAVELIGHT GMBH
Name: Company.name, dtype: string
```

```
HGF vs non-HGF for Number.of.patents
Welch's t-test statistic = -2.068
p-value = 0.03868

Optimization terminated successfully.
         Current function value: 0.155654
         Iterations 7
                           Logit Regression Results
==============================================================================
Dep. Variable:                    HGF   No. Observations:               115840
Model:                          Logit   Df Residuals:                   115838
Method:                           MLE   Df Model:                            1
Date:                Mon, 29 Jun 2020   Pseudo R-squ.:               5.332e-05
Time:                        15:41:34   Log-Likelihood:                -18031.
converged:                       True   LL-Null:                       -18032.
Covariance Type:            nonrobust   LLR p-value:                    0.1655
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -3.2822      0.016   -208.621      0.000      -3.313      -3.251
PAT            0.0034      0.002      1.641      0.101      -0.001       0.008
==============================================================================
```

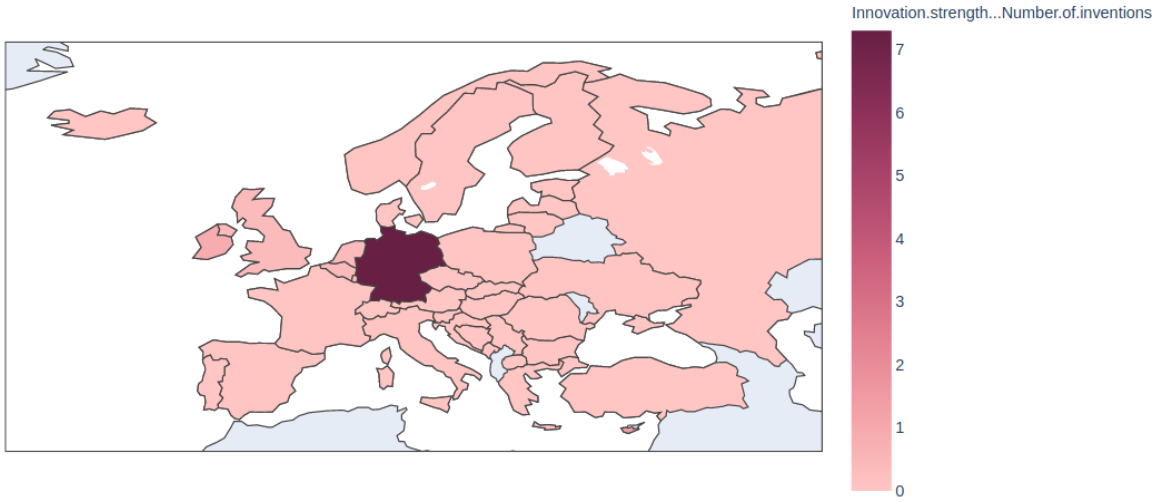
7.2.1.2.51. Innovation.strength...Number.of.inventions



sum:



mean:

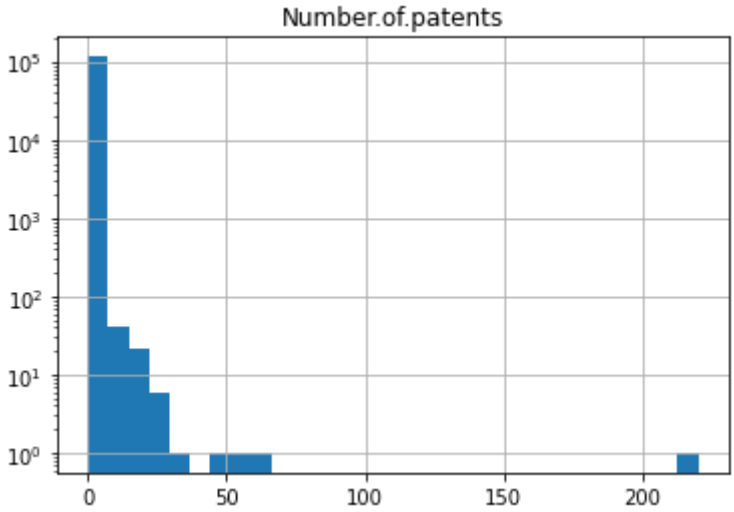


outlier:

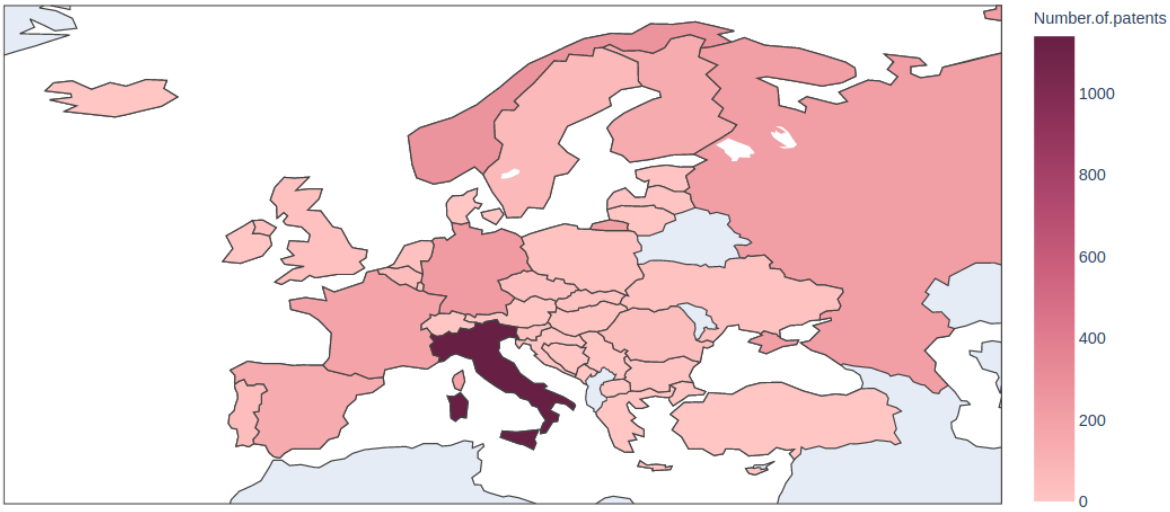
```
BvD.ID.number
DE8190460728    WAVELIGHT GMBH
Name: Company.name, dtype: string

```

```
HGF vs non-HGF for Innovation.strength...Number.of.inventions
Welch's t-test statistic = -1.361
p-value = 0.1737

Optimization terminated successfully.
         Current function value: 0.155656
         Iterations 7
                           Logit Regression Results
==============================================================================
Dep. Variable:                    HGF   No. Observations:               115840
Model:                          Logit   Df Residuals:                   115838
Method:                           MLE   Df Model:                            1
Date:                Mon, 29 Jun 2020   Pseudo R-squ.:               3.980e-05
Time:                        15:41:56   Log-Likelihood:                -18031.
converged:                       True   LL-Null:                       -18032.
Covariance Type:            nonrobust   LLR p-value:                    0.2309
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -3.2824      0.016   -208.594      0.000      -3.313      -3.252
INV            0.0141      0.010      1.425      0.154      -0.005       0.034
```

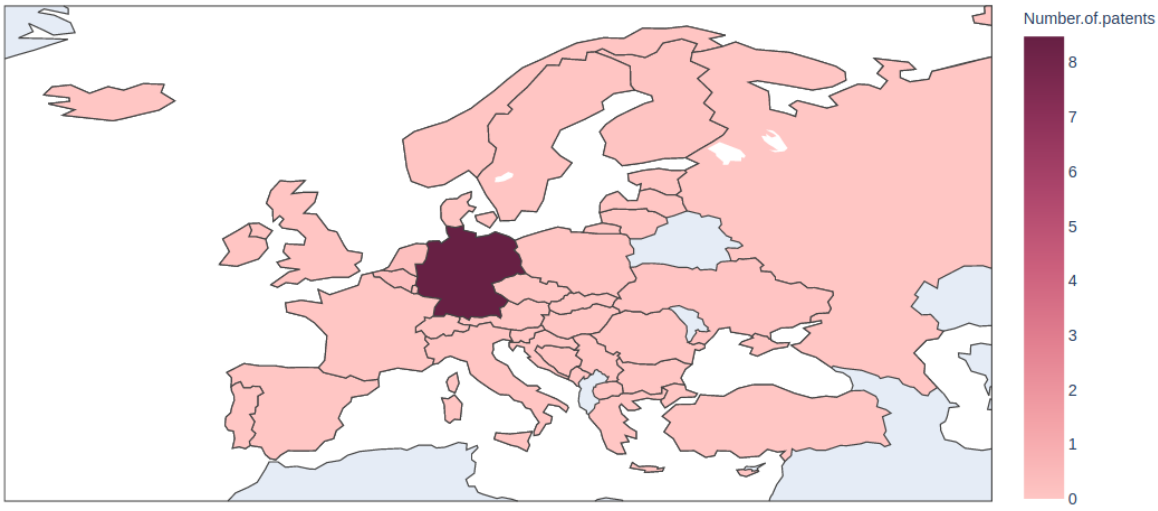
7.2.1.2.52. Number.of.patents



sum:



mean:

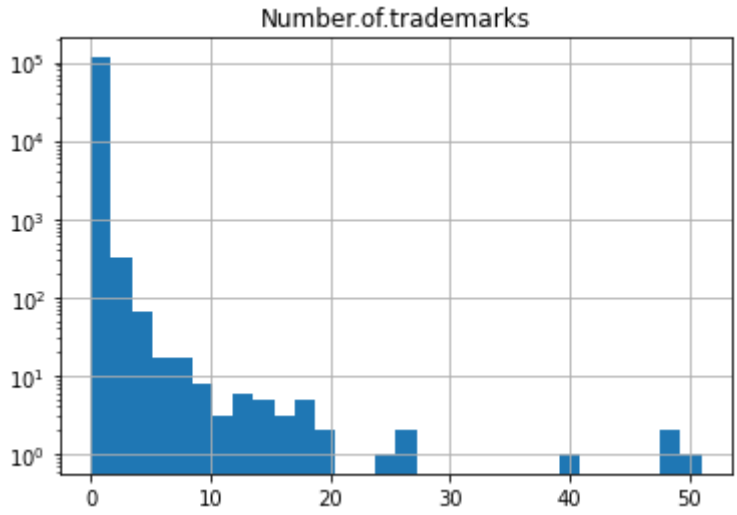


outlier:

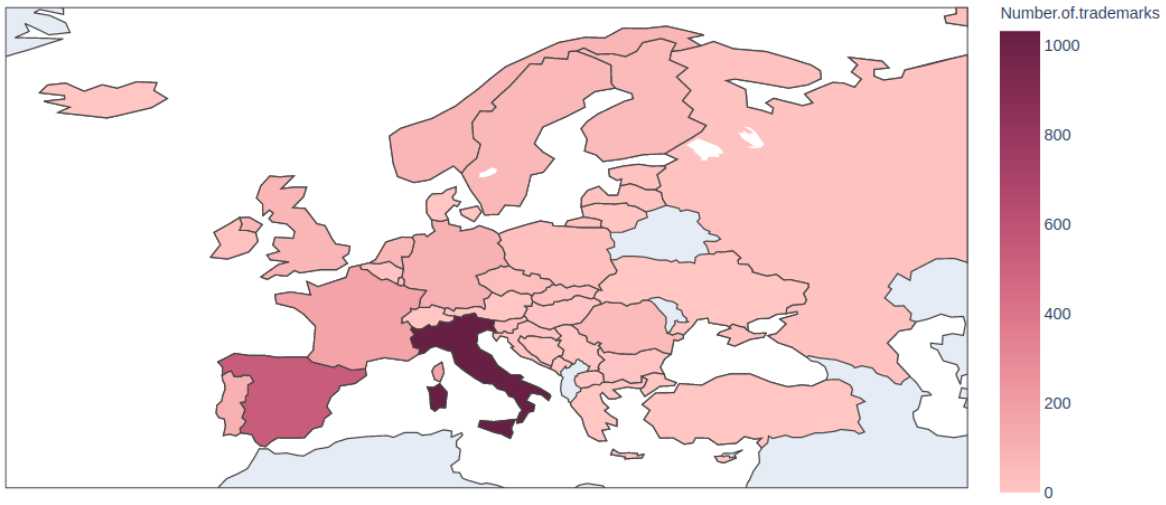
```
BvD.ID.number
DE8190460728    WAVELIGHT GMBH
Name: Company.name, dtype: string
```

```
HGF vs non-HGF for Number.of.patents
Welch's t-test statistic = -2.068
p-value = 0.03868

Optimization terminated successfully.
         Current function value: 0.155657
         Iterations 7
                           Logit Regression Results
==============================================================================
Dep. Variable:                    HGF   No. Observations:               115840
Model:                          Logit   Df Residuals:                   115838
Method:                           MLE   Df Model:                            1
Date:                Mon, 29 Jun 2020   Pseudo R-squ.:               3.073e-05
Time:                        15:42:17   Log-Likelihood:                -18031.
converged:                       True   LL-Null:                       -18032.
Covariance Type:            nonrobust   LLR p-value:                    0.2925
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -3.2823      0.016   -208.605      0.000      -3.313      -3.251
P              0.0116      0.009      1.267      0.205      -0.006       0.029
==============================================================================
```

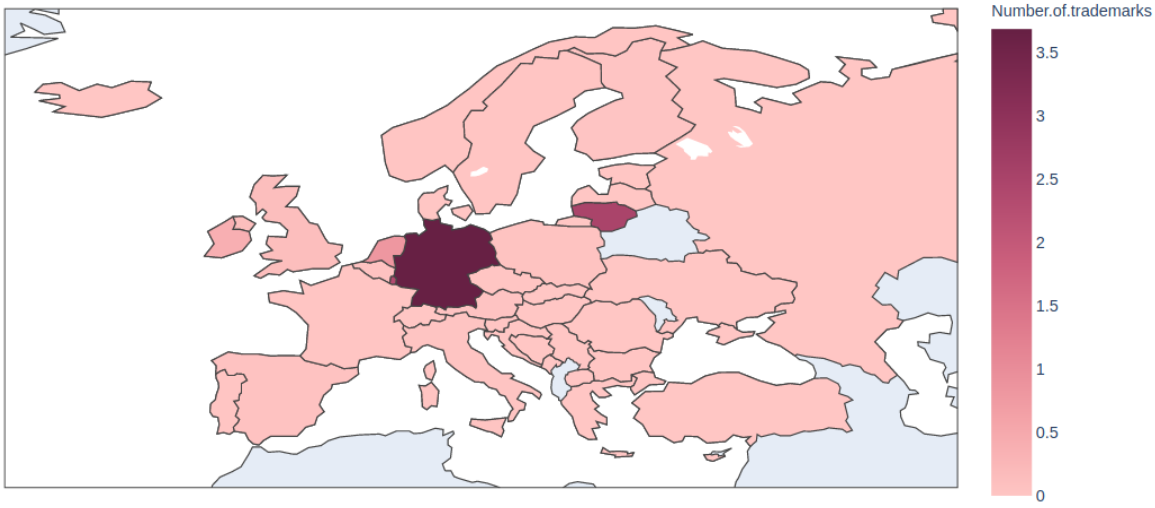
7.2.1.2.53. Number.of.trademarks



sum:



mean (Lituania!):



outliers:

```
BvD.ID.number
DE4250480683          VOLMARY GMBH
DE8190460728       WAVELIGHT GMBH
IT07237530964        SALROS S.R.L.
LULB157784        MEDA PHARMA SARL
Name: Company.name, dtype: string
```

```
HGF vs non-HGF for Number.of.trademarks
Welch's t-test statistic = -5.38
p-value = 7.824e-08

Optimization terminated successfully.
         Current function value: 0.155560
         Iterations 7
                           Logit Regression Results
==============================================================================
Dep. Variable:                    HGF   No. Observations:                115840
Model:                          Logit   Df Residuals:                    115838
Method:                           MLE   Df Model:                             1
Date:                Mon, 29 Jun 2020   Pseudo R-squ.:               0.0006523
Time:                        15:42:41   Log-Likelihood:                -18020.
converged:                       True   LL-Null:                       -18032.
Covariance Type:            nonrobust   LLR p-value:                 1.234e-06
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
```

```
Intercept     -3.2852       0.016   -208.489       0.000       -3.316       -3.254
T              0.0918       0.018      5.037       0.000        0.056        0.128
==============================================================================
```

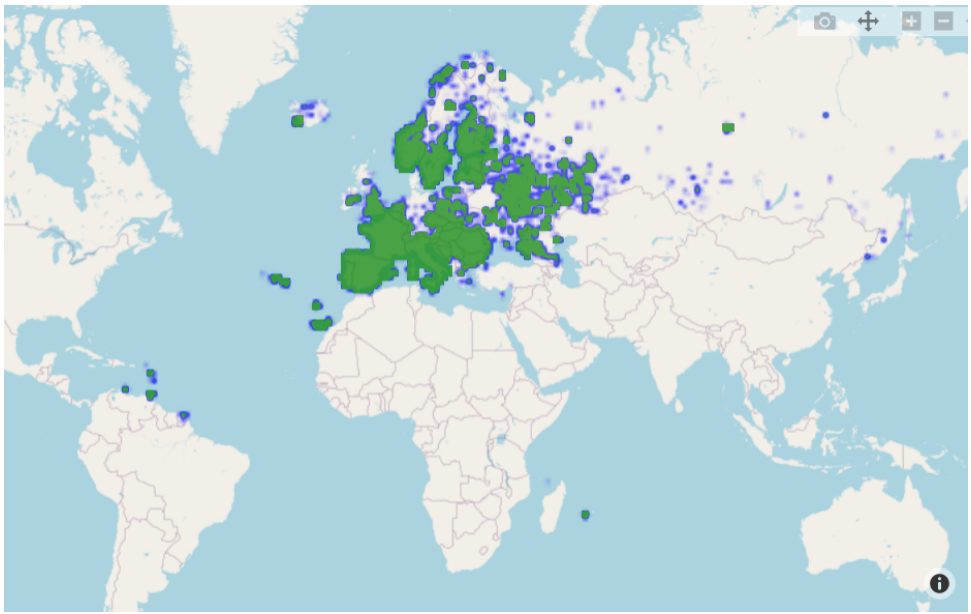
7.2.1.2.54. Trademarks...Type

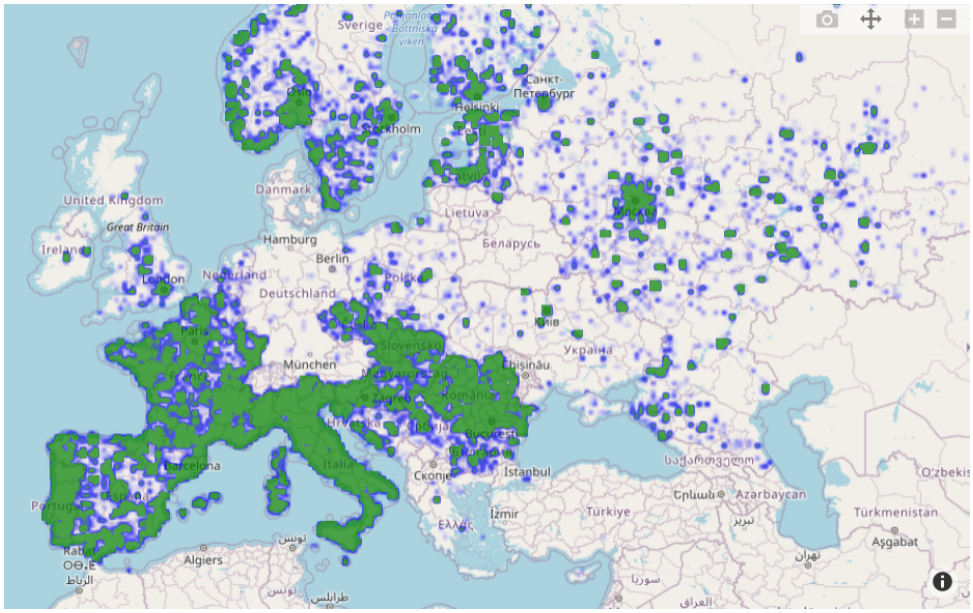
```
BvD.ID.number
RO26941545       No
PT509482171      No
IT06789671218    No
FR519321806      No
NO995113562      No
Name: Trademarks...Type, dtype: category
Categories (4, object): [Figurative, No, Other, Word]
```

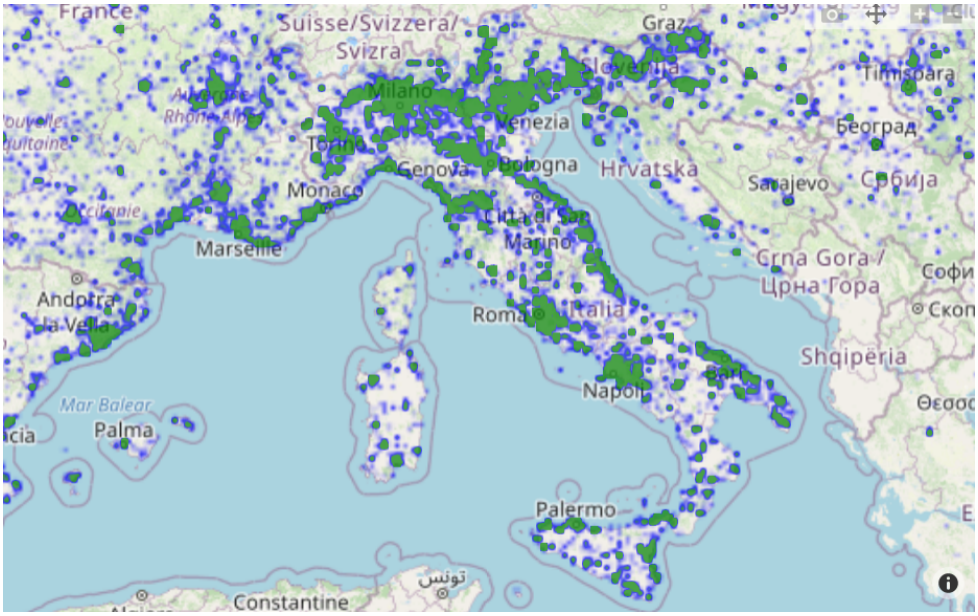
Trademarks...Type	**n**
Figurative	779
Word	400
Other	3
No	114658
Total	**115840**

7.2.1.2.55. lat / lon

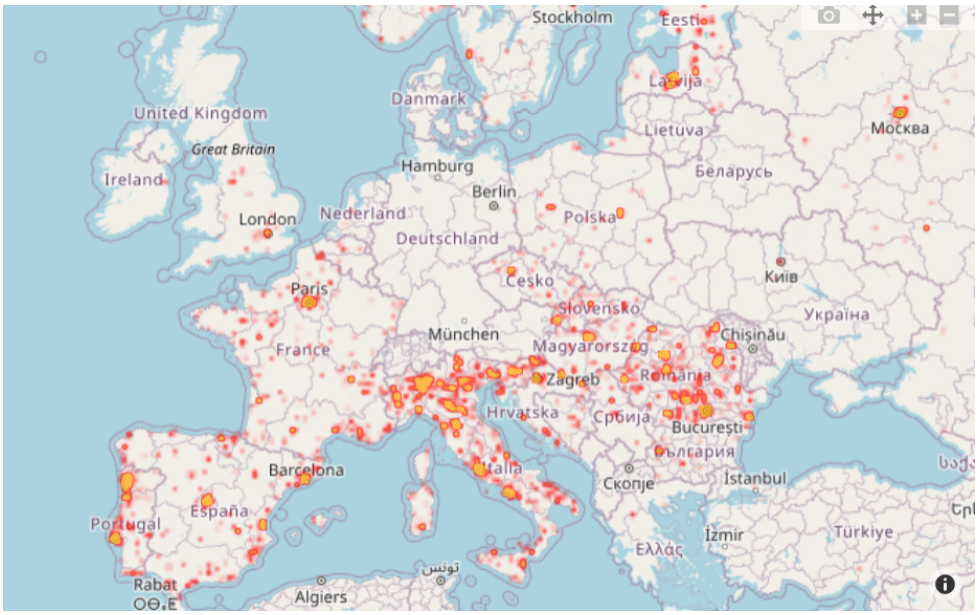
Geographic coordinates of the firm. Not available for ~8000 samples.







HGF=True:



(These maps are navigable like Google Maps on my pc, I can provide an interactive version.)

7.2.1.2.56. trust

Trust evaluation, as string made of two parts.

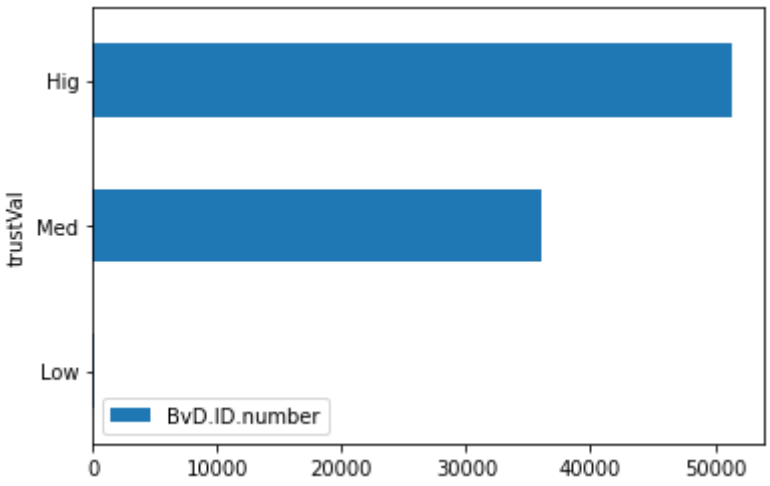
- A class {Low, Medium, High}
- A firm name, maybe the evaluator?

```
BvD.ID.number
RO26547053                                                    <NA>
HR05541138225    Medium: [DAKOVO, 31400, Croatia] instead of [C...
IT06738671210    High: [ESSECI ITALIA S.R.L., POGGIOMARINO, 800...
ESB72172646          High: [NUMENTI SL, PUERTO REAL, 11510, Spain]
IT02439920352      High: [SAN TOMMASO S.R.L., CANTU, 22063, Italy]
Name: trust, dtype: string
```

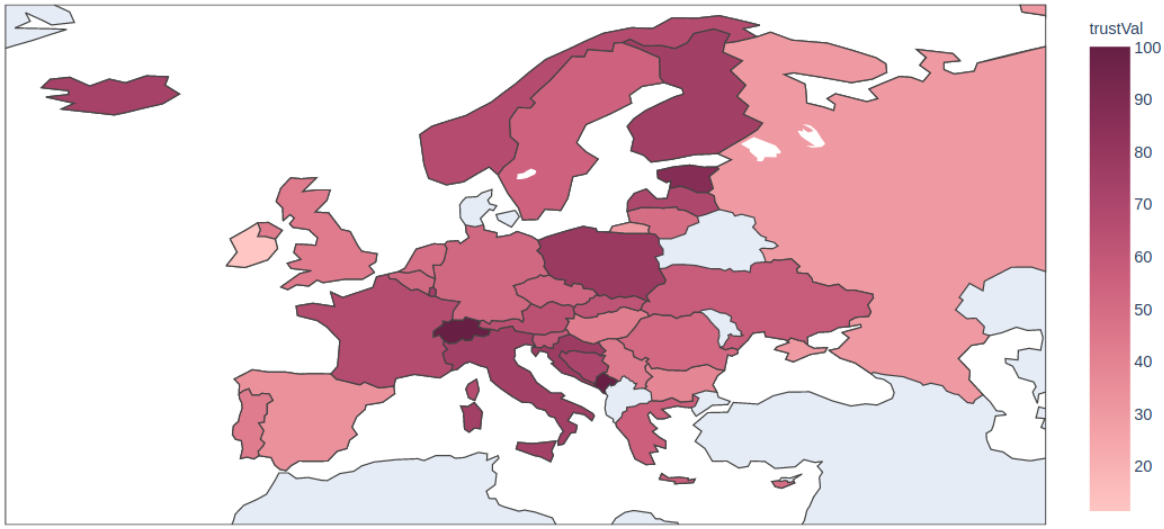
7.2.1.2.57. trustVal

To give meaning to `trust`, We have transformed the variable, creating a variable with the ranked level of trust. Class labels are in {'Low','Med','Hig'}.

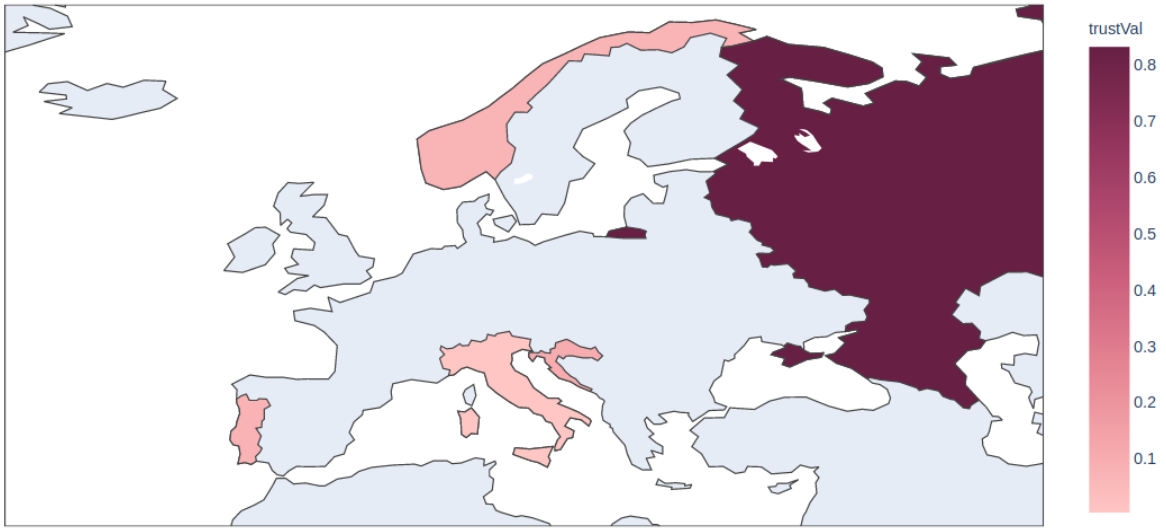
```
BvD.ID.number
IT04844800658    Hig
SE5568281116     Med
IT07016280963    Med
GB07234353       NaN
FR519497507      Hig
Name: trustVal, dtype: category
Categories (3, object): [Low < Med < Hig]
```



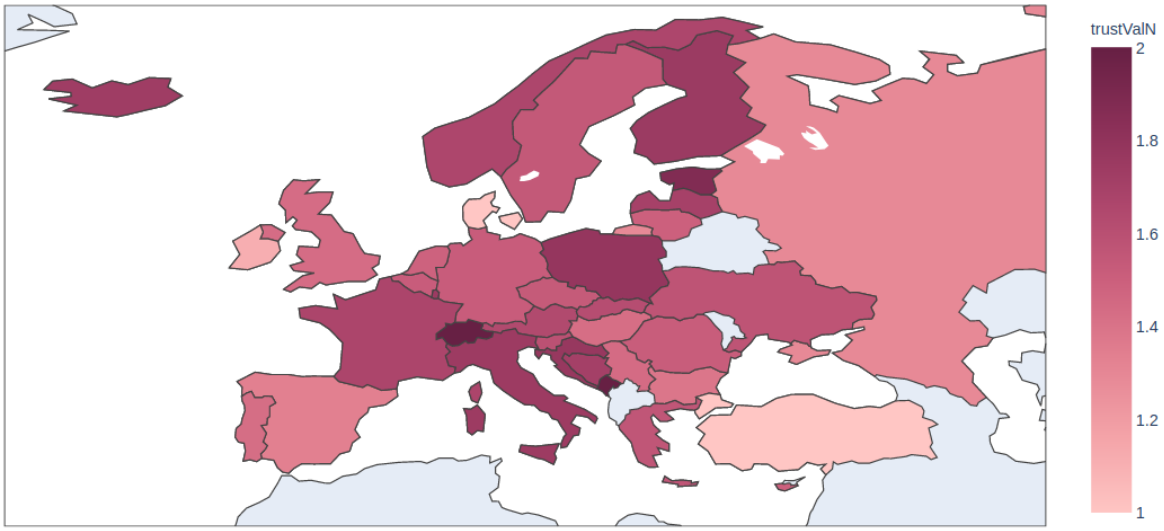
Percentage of companies with "High" rating:



Percentage of companies with "Low" rating (*note the different y axis*):



mean trust: (high = 2, low = 0)



7.3. High Growth Firms

The first issue to track was negotiating with the Client an objective definition of High Growth Firm (henceforth, HGF). HGF is a dichotomic variable, defining whether a firm is a good performer.

There are multiple definition of HGF in literature, that leads to the choice of our metric.

7.3.1. HGF metrics

There are three different accepted definition of HGF:

1. **Compound Annual Growth Rate (CAGR)**. Companies with an average growth rate ≥20% for the first 5 years:

$$CAGR = \left(\frac{turnover_{2014}}{turnover_{2010}} \right)^{1/4} - 1 \geq 20\% \, \$$$

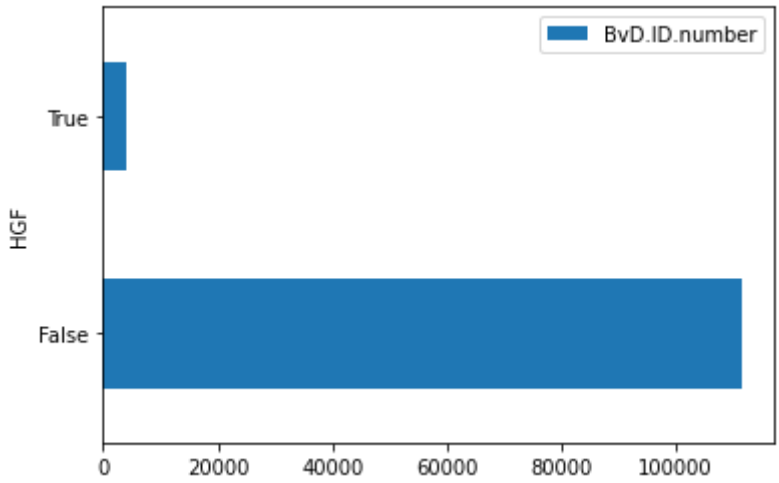
2. **Gazelle**[56] gazelles are firm with a growth rate that remains >20% for the first 5 years

$$Gazele = all \left(\frac{turnover_t}{turnover_{t-1}} \geq 20\% \right), for \, t = 2010, \ldots, 2014$$

3. **Eurostat**[57], employed by Eurostat, being HGF means having a growth rate ≥ 20% for 3 consecutive years.

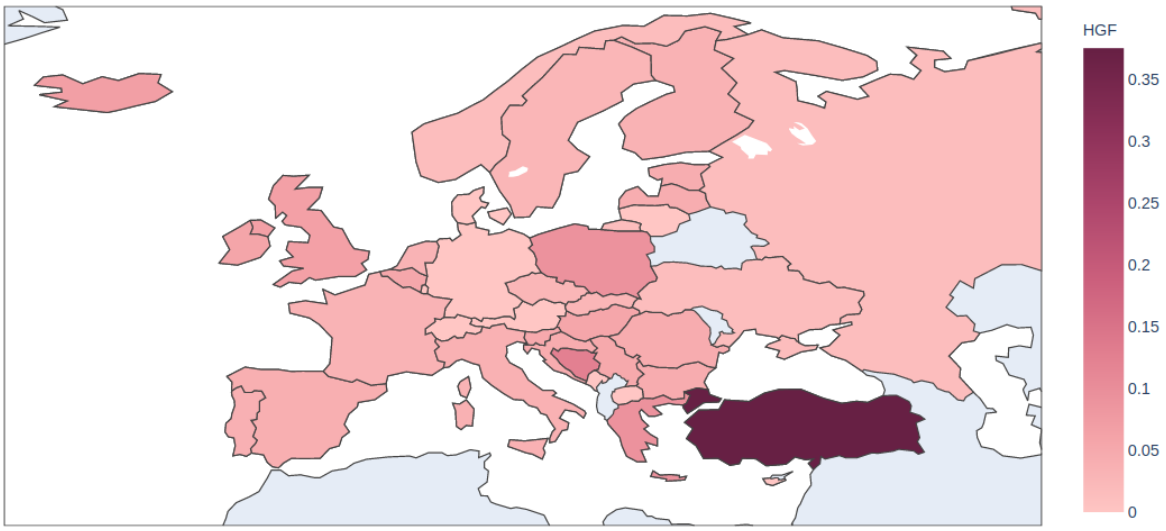
$$\begin{aligned} Eurostat = \exists \, t \in \{2010, 2011, 2012\} : \\ \left(\frac{turnover_t}{turnover_{t-1}} \geq 20\% \wedge \right. \\ \frac{turnover_t + 1}{turnover_t} \geq 20\% \wedge \\ \left. \frac{turnover_t + 2}{turnover_{t+1}} \geq 20\% \right) \end{aligned}$$

In this dataset the Client chose to compute HGF by the second option, **Gazelle**.



The dataset is heavily unbalanced, with a ratio of $\sim 33:1$ against non-HGF firm.

$p(HGF|country)$:



7.4. Using ROSE on ORBIS dataset

We applied ROSE on the dataset, and checked the performance of different models pre- and post- resampling.

7.4.1. data cleaning

From bibliography and understanding the dataset we recognized that a lot of columns where just sum of other columns. The graphs in variable descriptions helps understanding this collinearity. It was judged safe to just drop derived variables, keeping only original ones. The following variables were dropped:

```
'Fixed.assets.th.EUR.2010',
'Current.assets.th.EUR.2010',
'Total.assets.th.EUR.2010',
'Shareholders.funds.th.EUR.2010',
'Non.current.liabilities.th.EUR.2010',
'Other.current.liabilities.th.EUR.2010',
'Sales.th.EUR.2010',
'Financial.revenue.th.EUR.2010',
'Financial.expenses.th.EUR.2010',
'Taxation.th.EUR.2010',
'Cash.flow.th.EUR.2010',
```

For the same reason, we dropped variables derived from NACE code:

```
'NACE.Rev..2.main.section',
'NACE.Rev..2.Primary.code.s.'
```

The client wanted to focus only on private small companies, so the dataset was filtered in that sense, and the variable was dropped.

```
df = df[df['Standardised.legal.form']=='Private limited companies']
df = df.drop('Standardised.legal.form', axis=1)

df = df[df['Category.of.the.company']=='Small company']
df = df.drop('Category.of.the.company', axis=1)
```

Consolidation Code was deemed irrelevant by the Client, and hence dropped. Given ROSE inability to work on string values, and the excessive cardinality of postcodes, the following variables were dropped:

```
"Company.name",
"City",
"trust",
"Postcode",
"Postcode2"
```

Categorical variables were then one-hot-encoded:

```
for var in ["Country.ISO.Code",
            "BvD.Independence.Indicator",
            "NACE.Rev..2.Core.code..4.digits.",
            "BvD.major.sector",
             "trustVal",
             "Trademarks...Type",]:

    temp = pd.get_dummies(df[var])
    df = df.join(temp)
    df = df.drop(var, axis=1)
```

After the cleaning, the dataset's shape was 90711 examples, with 832 variables, most of them due to the one-hot-encoding. Of these, in 2343 samples $HGF = True$, while in 88368 $HGF = False$.

7.4.2. Data visualization

Given the high dimensionality, we used t-distributed stochastic neighbor embedding (t-SNE) to plot a representation of the original data. The parameters of the t-SNE were: $perplexity = 100, iterations = 250, n\_components = 2$. Extra iteration and different perplexities has been tested, withtout significant improvement.

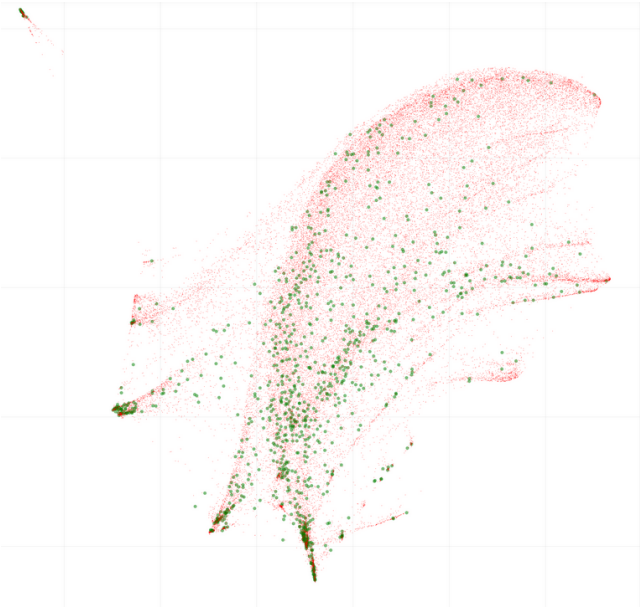


Fig _ t-SNE of original dataset. Green sample ($HGF = True$) size has been exaggerated on purpose.

7.4.3. ROSE Resampling

A default `imblearn.over_sampling.ROSE()` instance has been generated, with `random_state` parameter set on 42 for reasons pertaining the life, the universe, and everything else. We used the same t-SNE methodology as above to visualize the balanced dataset.

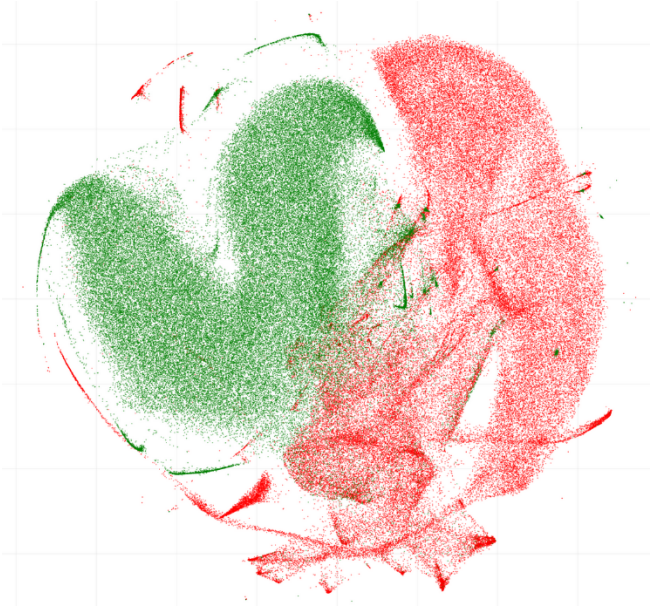


Fig _ *t*-SNE plot of resampled dataset

The resampler was used to even the classes, and different models has been tested, without optimization. To begin, we tested a Gaussian Naive Bayes model:

Performance on original dataset:

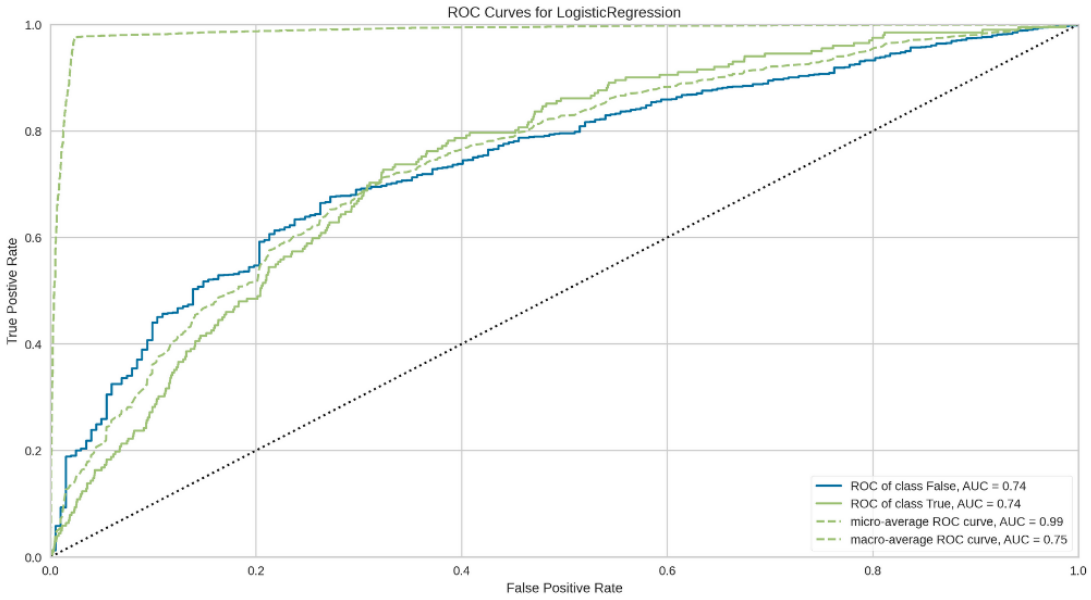
HGF	**Precision**	**Recall**	**F1**	**Support**
True	0.029	0.762	0.055	202
False	0.985	0.376	0.545	8343

Performance on balanced dataset:

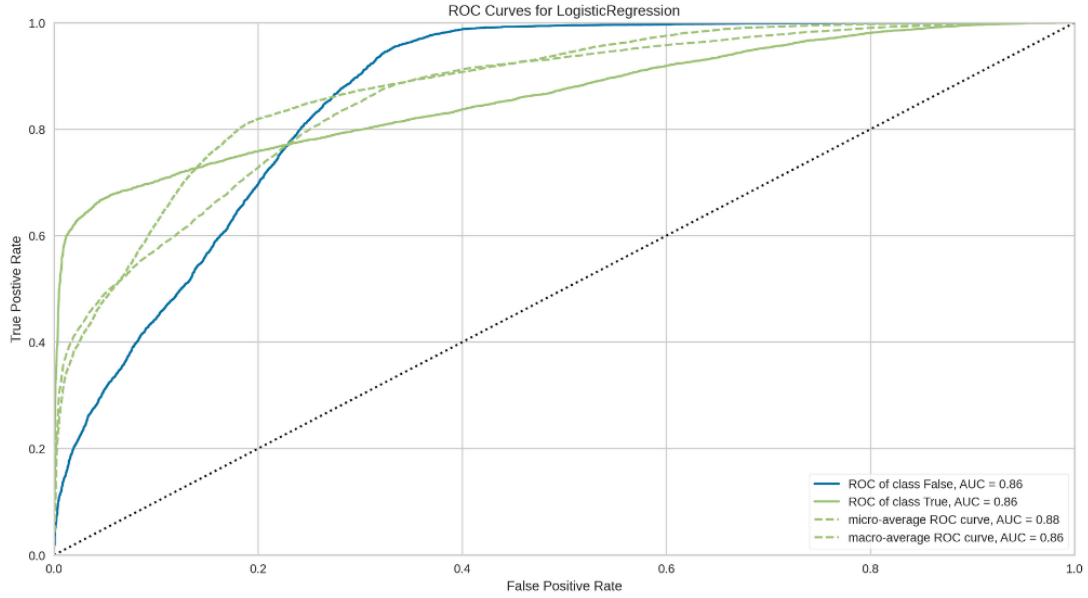
HGF	**Precision**	**Recall**	**F1**	**Support**
True	0.691	0.956	0.738	8332
False	0.894	0.367	0.521	8358

Then we checked ROC Curves for a non-optimized logistic regression model, encompassing all variables.

Performance on original dataset:

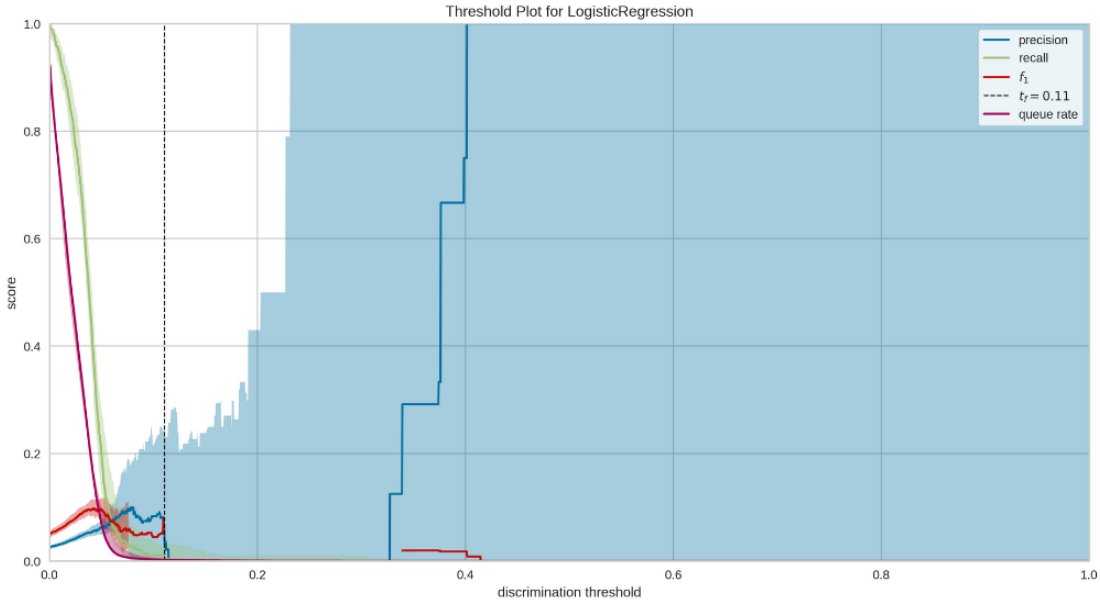


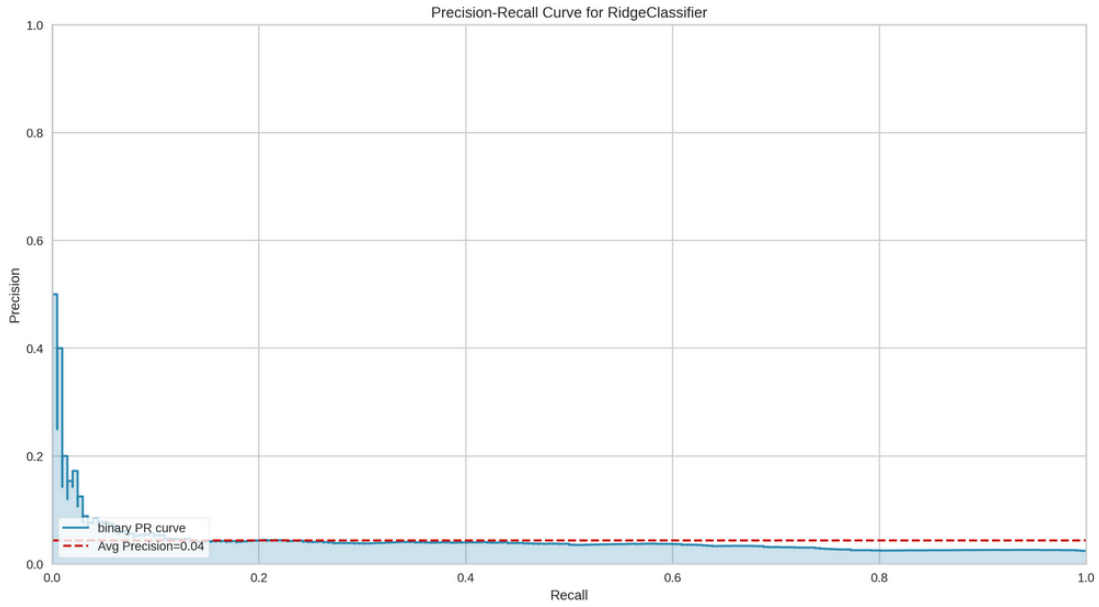
Performance on balanced dataset:



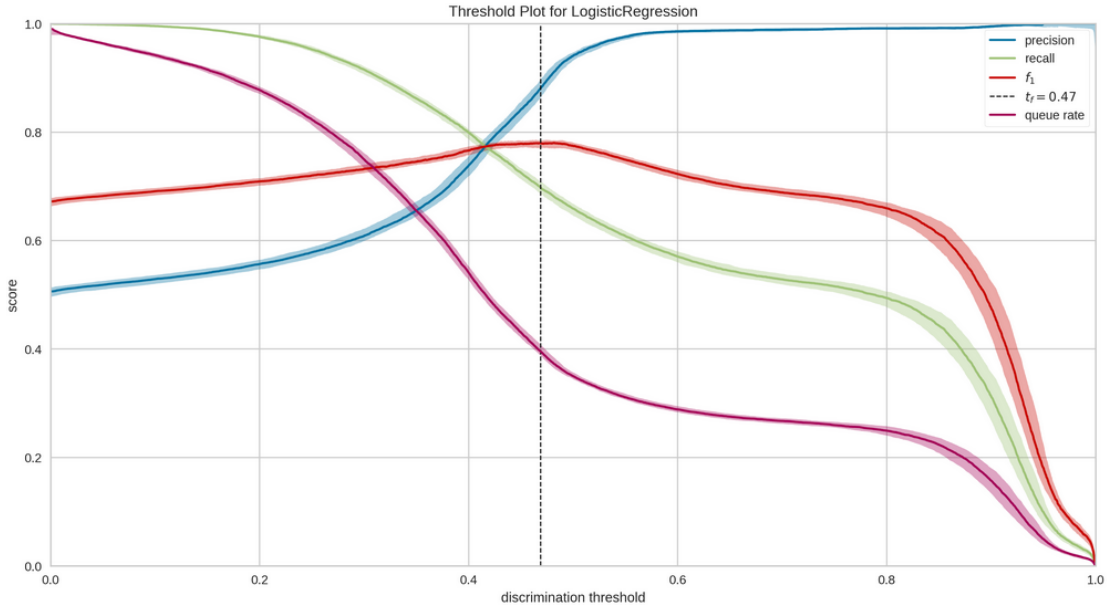
To better visualize the tradeoff between precision and recall in both models, we plotted threshold plots of the logistic classifier, and precision/recall curves of a ridge classifier:

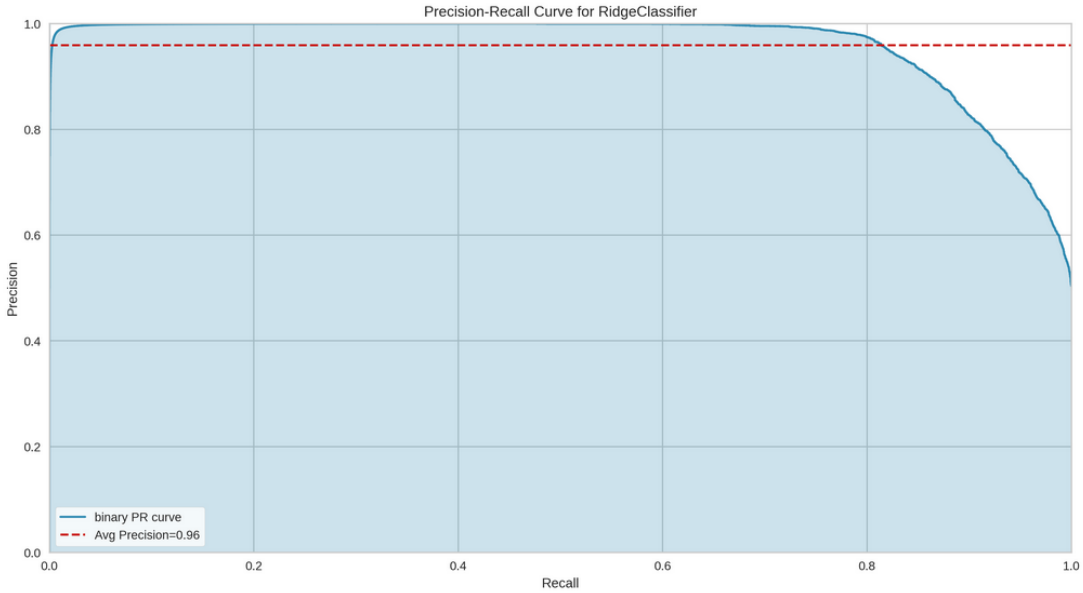
Performance on original dataset:





Performance on balanced dataset:




Precision-Recall Curve for RidgeClassifier
Precision
Recall
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
binary PR curve
Avg Precision=0.96

8. Discussion

This work's first objective, ROSE implementation in Python's package `imbalanced-learn` has been successfully achieved, and with the next release it will be available for all users.

Binary classifier metrics evaluation and choice has proven a big challenge. Matthews correlation coefficient has proven a severe judge, performing better than F_1 score in describing, in a single number, the model performance.

Additional models could have been tested, like bigger ANNs, different NN architectures, or Gaussian Process classifiers, but additional computational power is required to do that, given the number of models to train and compare. By expanding the set, given the high repeatability of the tests, we could be able to propose a standard suite for testing resamplers.

Testing ROSE under different datasets and algorithms shown that, in some cases, its performance can equal and even be better than other resamplers. The difference is exacerbated when the imbalance ratio of the dataset is higher.

This is only the first part of ROSE development for Python. The algorithm still have unsolved issues, like incapacity of treating categorical data, or variables with limited support. Ideas for solution has been discussed, and will be implemented in the future, but their implementation and validation was out of scope for this project.

9. Bibliography

10. Appendix 1: other compared metrics

In this section we will show tables of results, analogue to the ones shown in Chapter 6.2, for other computed metrics on the same experimental setup.

10.1. ROC-AUC

KNeighborsClassifier

	SMOTE	ROS	ROSE	ADASYN	RUS
ecoli	0.95	0.95	0.87	0.91	0.72
optical_digits	1.0	1.0	0.99	1.0	0.99
satimage	0.94	0.96	0.86	0.94	0.89
pen_digits	1.0	1.0	0.99	1.0	0.99
abalone	0.89	0.92	0.92	0.89	0.74
sick_euthyroid	0.87	0.92	0.75	0.87	0.69
spectrometer	0.98	0.99	0.81	0.98	0.94
car_eval_34	0.94	0.91	0.97	0.93	0.86
isolet	0.95	0.98	0.97	0.95	0.92
us_crime	0.93	0.96	0.74	0.92	0.85
yeast_ml8	0.74	0.92	0.92	0.73	0.58
scene	0.8	0.93	0.77	0.8	0.58
libras_move	0.99	0.99	0.85	0.99	0.94
thyroid_sick	0.89	0.94	0.81	0.89	0.71
coil_2000	0.9	0.92	0.7	0.9	0.61
arrhythmia	0.86	0.92	0.75	0.86	0.74
solar_flare_m0	0.86	0.88	0.73	0.87	0.62
oil	0.87	0.93	0.9	0.83	0.63
car_eval_4	0.95	0.94	0.97	0.95	0.82
wine_quality	0.91	0.97	0.77	0.91	0.65
letter_img	1.0	1.0	0.98	1.0	0.97
yeast_me2	0.95	0.98	0.85	0.96	0.92
webpage	0.96	0.97	0.88	0.96	0.86
ozone_level	0.93	0.96	0.86	0.92	0.73
mammography	0.96	0.98	0.96	0.94	0.87
protein_homo	0.98	1.0	0.89	0.98	0.84
abalone_19	0.96	0.99	0.97	0.96	0.69

SVC(linear kernel)

	SMOTE	ROS	ROSE	ADASYN	RUS
ecoli	0.89	0.88	0.89	0.86	0.78
optical_digits	0.96	0.91	0.86	0.94	0.92
satimage	0.68	0.43	0.59	0.61	0.64
pen_digits	0.51	0.53	0.67	0.58	0.68
abalone	0.75	0.75	0.74	0.74	0.71
sick_euthyroid	0.69	0.67	0.7	0.73	0.86
spectrometer	0.99	0.99	0.44	0.99	0.94
car_eval_34	0.95	0.95	0.95	0.95	0.89
isolet	0.98	0.98	0.96	0.98	0.96
us_crime	0.89	0.87	0.85	0.88	0.87
yeast_ml8	0.64	0.61	0.63	0.65	0.5
scene	0.79	0.79	0.78	0.8	0.77
libras_move	0.97	0.94	0.9	0.89	0.5
thyroid_sick	0.81	0.82	0.73	0.81	0.85
coil_2000	0.63	0.63	0.59	0.66	0.67
arrhythmia	0.96	0.96	0.89	0.96	0.55
solar_flare_m0	0.8	0.73	0.76	0.77	0.66
oil	0.44	0.53	0.6	0.43	0.45
car_eval_4	0.98	0.98	0.98	0.98	0.9
wine_quality	0.68	0.53	0.52	0.56	0.72
letter_img	0.96	0.95	0.91	0.93	0.94
yeast_me2	0.84	0.84	0.82	0.83	0.5
webpage	0.85	0.86	0.57	0.78	0.92
ozone_level	0.87	0.76	0.66	0.89	0.81
mammography	0.88	0.88	0.86	0.82	0.85
protein_homo	0.66	0.48	0.49	0.49	0.73
abalone_19	0.66	0.66	0.67	0.65	0.69

SVC (RBF kernel)

	SMOTE	ROS	ROSE	ADASYN	RUS
ecoli	0.91	0.91	0.92	0.9	0.89
optical_digits	0.64	1.0	0.5	0.8	0.5
satimage	0.66	1.0	0.69	0.66	0.5
pen_digits	0.74	1.0	0.74	0.91	0.6
abalone	0.81	0.81	0.92	0.8	0.76
sick_euthyroid	0.77	1.0	0.51	0.74	0.53
spectrometer	0.5	1.0	0.5	0.5	0.5
car_eval_34	1.0	1.0	0.99	1.0	0.9
isolet	0.86	1.0	0.96	0.88	0.5
us_crime	1.0	1.0	0.83	0.99	0.77
yeast_ml8	1.0	1.0	0.96	1.0	0.59
scene	0.94	1.0	0.54	0.94	0.55
libras_move	1.0	1.0	0.95	1.0	0.94
thyroid_sick	0.78	1.0	0.5	0.78	0.5
coil_2000	0.88	0.99	0.8	0.87	0.52
arrhythmia	0.55	1.0	0.5	0.53	0.5
solar_flare_m0	0.95	0.92	0.92	0.93	0.53
oil	0.5	1.0	0.5	0.5	0.5
car_eval_4	1.0	1.0	0.93	1.0	0.52
wine_quality	0.9	1.0	0.66	0.91	0.56
letter_img	0.99	1.0	0.95	0.99	0.67
yeast_me2	0.88	0.89	0.85	0.87	0.82
webpage	0.99	1.0	0.87	0.99	0.62
ozone_level	0.67	1.0	0.5	0.65	0.5
mammography	0.95	0.96	0.96	0.93	0.88
protein_homo	0.5	1.0	0.5	0.5	0.53
abalone_19	0.89	0.88	0.98	0.9	0.81

DecisionTreeClassifier

	SMOTE	ROS	ROSE	ADASYN	RUS
ecoli	0.92	0.92	0.91	0.89	0.67
optical_digits	0.95	0.94	0.96	0.94	0.88
satimage	0.9	0.88	0.86	0.88	0.85
pen_digits	0.96	0.96	0.92	0.9	0.94
abalone	0.82	0.82	0.95	0.82	0.79
sick_euthyroid	0.97	0.97	0.96	0.96	0.94
spectrometer	0.94	0.97	0.91	0.94	0.82
car_eval_34	0.95	0.95	0.96	0.95	0.94
isolet	0.93	0.92	0.95	0.94	0.89
us_crime	0.91	0.91	0.95	0.91	0.8
yeast_ml8	0.72	0.7	0.75	0.71	0.57
scene	0.83	0.84	0.92	0.81	0.6
libras_move	0.99	0.98	0.91	0.98	0.88
thyroid_sick	0.98	0.99	0.98	0.99	0.94
coil_2000	0.78	0.7	0.97	0.83	0.62
arrhythmia	0.99	1.0	0.96	0.99	1.0
solar_flare_m0	0.89	0.8	0.97	0.87	0.55
oil	0.97	0.97	0.95	0.96	0.72
car_eval_4	0.98	0.98	0.98	0.98	0.95
wine_quality	0.79	0.84	0.82	0.79	0.63
letter_img	0.97	0.96	0.95	0.96	0.94
yeast_me2	0.92	0.94	0.97	0.93	0.77
webpage	0.81	0.79	0.99	0.8	0.78
ozone_level	0.91	0.94	0.92	0.87	0.72
mammography	0.92	0.93	0.95	0.89	0.9
protein_homo	0.94	0.94	0.99	0.92	0.9
abalone_19	0.9	0.94	1.0	0.89	0.44

RandomForestClassifier

	SMOTE	ROS	ROSE	ADASYN	RUS
ecoli	0.93	0.92	0.92	0.9	0.83
optical_digits	0.93	0.92	0.95	0.92	0.9
satimage	0.9	0.87	0.88	0.87	0.9
pen_digits	0.98	0.98	0.95	0.96	0.97
abalone	0.82	0.82	0.84	0.81	0.82
sick_euthyroid	0.81	0.94	0.95	0.84	0.86
spectrometer	0.95	0.99	0.95	0.94	0.94
car_eval_34	0.95	0.95	0.96	0.95	0.92
isolet	0.88	0.87	0.95	0.88	0.84
us_crime	0.89	0.88	0.93	0.89	0.83
yeast_ml8	0.76	0.82	0.79	0.73	0.59
scene	0.82	0.83	0.95	0.8	0.63
libras_move	0.99	1.0	0.92	0.99	0.69
thyroid_sick	0.87	0.8	0.97	0.85	0.83
coil_2000	0.82	0.68	0.97	0.8	0.6
arrhythmia	0.9	0.89	0.93	0.88	0.61
solar_flare_m0	0.87	0.76	0.97	0.83	0.71
oil	0.96	0.94	0.96	0.96	0.81
car_eval_4	0.96	0.95	0.99	0.96	0.9
wine_quality	0.81	0.82	0.82	0.85	0.75
letter_img	0.96	0.94	0.96	0.96	0.88
yeast_me2	0.92	0.95	0.94	0.91	0.76
webpage	0.86	0.76	0.98	0.85	0.74
ozone_level	0.9	0.89	0.88	0.89	0.8
mammography	0.91	0.93	0.95	0.89	0.88
protein_homo	0.93	0.92	0.97	0.91	0.91
abalone_19	0.86	0.9	0.97	0.85	0.56

MLPClassifier

	SMOTE	ROS	ROSE	ADASYN	RUS
ecoli	0.91	0.88	0.87	0.89	0.89
optical_digits	1.0	1.0	0.99	0.99	0.96
satimage	0.78	0.7	0.72	0.61	0.7
pen_digits	1.0	1.0	0.99	1.0	0.99
abalone	0.77	0.76	0.83	0.76	0.76
sick_euthyroid	0.94	0.93	0.91	0.9	0.85
spectrometer	0.97	0.96	0.95	0.95	0.73
car_eval_34	0.99	0.99	0.97	0.99	0.74
isolet	0.99	0.98	0.99	0.99	0.96
us_crime	0.9	0.9	0.92	0.89	0.88
yeast_ml8	0.83	0.84	0.79	0.81	0.58
scene	0.92	0.91	0.89	0.93	0.81
libras_move	0.99	0.99	0.96	0.99	0.94
thyroid_sick	0.94	0.94	0.93	0.91	0.89
coil_2000	0.85	0.81	0.96	0.86	0.64
arrhythmia	0.96	0.96	0.88	0.99	0.49
solar_flare_m0	0.9	0.9	0.96	0.88	0.65
oil	0.73	0.7	0.87	0.74	0.85
car_eval_4	0.99	0.99	0.99	0.99	0.95
wine_quality	0.77	0.76	0.83	0.77	0.67
letter_img	0.99	0.99	0.96	0.99	0.95
yeast_me2	0.89	0.86	0.82	0.85	0.92
webpage	0.96	0.95	0.97	0.95	0.95
ozone_level	0.55	0.8	0.77	0.76	0.25
mammography	0.92	0.9	0.93	0.87	0.9
protein_homo	0.98	0.99	1.0	0.99	0.91
abalone_19	0.82	0.81	0.92	0.83	0.75

AdaBoostClassifier

	SMOTE	ROS	ROSE	ADASYN	RUS
ecoli	0.9	0.95	0.92	0.91	0.78
optical_digits	0.98	0.96	0.99	0.98	0.93
satimage	0.92	0.9	0.88	0.9	0.86
pen_digits	0.98	0.98	0.98	0.98	0.96
abalone	0.81	0.82	0.95	0.81	0.78
sick_euthyroid	0.97	0.96	0.98	0.96	0.93
spectrometer	0.98	0.99	0.96	0.98	0.94
car_eval_34	0.99	1.0	0.99	0.99	0.99
isolet	0.97	0.97	0.98	0.97	0.94
us_crime	0.93	0.94	0.95	0.92	0.85
yeast_ml8	0.79	0.82	0.88	0.8	0.59
scene	0.87	0.88	0.94	0.85	0.69
libras_move	0.99	0.99	0.95	0.99	0.81
thyroid_sick	0.98	0.97	0.98	0.98	0.92
coil_2000	0.91	0.71	0.97	0.9	0.61
arrhythmia	0.99	1.0	0.99	0.98	1.0
solar_flare_m0	0.9	0.75	0.97	0.9	0.64
oil	0.97	0.99	0.96	0.98	0.81
car_eval_4	0.99	0.99	0.99	0.99	0.95
wine_quality	0.84	0.84	0.88	0.83	0.66
letter_img	0.99	0.98	0.99	0.99	0.98
yeast_me2	0.92	0.96	0.98	0.94	0.85
webpage	0.95	0.92	0.99	0.94	0.93
ozone_level	0.96	0.98	0.96	0.95	0.83
mammography	0.91	0.92	0.97	0.87	0.89
protein_homo	0.97	0.96	1.0	0.96	0.95
abalone_19	0.89	0.95	1.0	0.89	0.75

GaussianNB

	SMOTE	ROS	ROSE	ADASYN	RUS
ecoli	0.88	0.88	0.84	0.87	0.89
optical_digits	0.67	0.6	0.57	0.64	0.67
satimage	0.85	0.84	0.8	0.82	0.87
pen_digits	0.76	0.74	0.76	0.85	0.76
abalone	0.76	0.77	0.75	0.76	0.71
sick_euthyroid	0.66	0.64	0.65	0.65	0.6
spectrometer	0.67	0.69	0.94	0.66	0.74
car_eval_34	0.95	0.95	0.95	0.95	0.94
isolet	0.93	0.89	0.91	0.89	0.89
us_crime	0.86	0.85	0.92	0.84	0.84
yeast_ml8	0.74	0.63	0.95	0.72	0.52
scene	0.71	0.71	0.86	0.71	0.76
libras_move	0.96	0.88	0.95	0.85	0.81
thyroid_sick	0.68	0.64	0.64	0.67	0.63
coil_2000	0.56	0.54	0.56	0.55	0.52
arrhythmia	0.68	0.66	0.66	0.67	0.61
solar_flare_m0	0.75	0.72	0.87	0.72	0.53
oil	0.57	0.82	0.8	0.63	0.72
car_eval_4	0.99	0.99	0.99	0.99	1.0
wine_quality	0.76	0.74	0.8	0.75	0.7
letter_img	0.88	0.87	0.88	0.87	0.87
yeast_me2	0.58	0.53	0.52	0.58	0.54
webpage	0.91	0.89	0.96	0.85	0.85
ozone_level	0.79	0.76	0.74	0.79	0.8
mammography	0.86	0.86	0.88	0.77	0.88
protein_homo	0.88	0.88	1.0	0.88	0.89
abalone_19	0.73	0.73	0.69	0.74	0.88

QuadraticDiscriminantAnalysis

	SMOTE	ROS	ROSE	ADASYN	RUS
ecoli	0.78	0.58	0.49	0.56	0.72
optical_digits	0.64	0.61	0.57	0.66	0.65
satimage	0.82	0.8	0.94	0.82	0.78
pen_digits	0.98	0.98	0.97	0.96	0.99
abalone	0.8	0.75	0.95	0.78	0.76
sick_euthyroid	0.65	0.65	0.95	0.66	0.58
spectrometer	1.0	1.0	0.95	1.0	0.62
car_eval_34	0.5	0.5	0.5	0.5	0.5
isolet	1.0	1.0	0.96	0.98	0.59
us_crime	0.97	0.97	0.95	0.98	0.76
yeast_ml8	0.96	0.96	0.95	0.95	0.58
scene	1.0	1.0	0.92	1.0	0.51
libras_move	1.0	1.0	0.97	1.0	0.94
thyroid_sick	0.59	0.56	0.97	0.57	0.52
coil_2000	0.53	0.52	0.5	0.53	0.6
arrhythmia	1.0	1.0	0.65	1.0	0.65
solar_flare_m0	0.79	0.68	0.97	0.72	0.53
oil	1.0	1.0	0.97	1.0	0.57
car_eval_4	0.5	0.5	0.5	0.5	0.5
wine_quality	0.77	0.75	0.87	0.77	0.66
letter_img	0.97	0.96	0.94	0.97	0.94
yeast_me2	0.52	0.51	0.5	0.55	0.42
webpage	0.82	0.75	0.99	0.77	0.87
ozone_level	1.0	1.0	0.92	1.0	0.61
mammography	0.85	0.84	0.89	0.79	0.82
protein_homo	0.95	0.94	1.0	0.94	0.95
abalone_19	0.86	0.82	1.0	0.83	0.81

10.2. F_1 score (mean of the two classes)

KNeighborsClassifier

	SMOTE	ROS	ROSE	ADASYN	RUS
ecoli	0.95	0.95	0.87	0.92	0.72
optical_digits	1.0	1.0	0.99	1.0	0.99
satimage	0.94	0.96	0.85	0.94	0.9
pen_digits	1.0	1.0	0.99	1.0	0.99
abalone	0.89	0.92	0.92	0.89	0.74
sick_euthyroid	0.87	0.92	0.75	0.86	0.69
spectrometer	0.98	0.99	0.81	0.98	0.95
car_eval_34	0.93	0.91	0.97	0.93	0.86
isolet	0.95	0.98	0.97	0.95	0.92
us_crime	0.93	0.96	0.73	0.92	0.85
yeast_ml8	0.73	0.92	0.92	0.7	0.57
scene	0.79	0.93	0.76	0.79	0.58
libras_move	0.99	0.99	0.83	0.99	0.91
thyroid_sick	0.89	0.94	0.8	0.89	0.71
coil_2000	0.9	0.92	0.68	0.9	0.61
arrhythmia	0.86	0.92	0.72	0.86	0.75
solar_flare_m0	0.86	0.88	0.71	0.87	0.63
oil	0.86	0.93	0.9	0.83	0.61
car_eval_4	0.95	0.93	0.97	0.95	0.79
wine_quality	0.91	0.97	0.76	0.91	0.65
letter_img	1.0	1.0	0.98	1.0	0.97
yeast_me2	0.95	0.98	0.86	0.96	0.92
webpage	0.96	0.97	0.88	0.96	0.85
ozone_level	0.93	0.96	0.86	0.92	0.73
mammography	0.96	0.98	0.96	0.94	0.87
protein_homo	0.98	1.0	0.89	0.98	0.84
abalone_19	0.96	0.99	0.97	0.96	0.68

SVC(linear kernel)

	SMOTE	ROS	ROSE	ADASYN	RUS
ecoli	0.9	0.88	0.9	0.86	0.77
optical_digits	0.96	0.91	0.86	0.94	0.92
satimage	0.68	0.42	0.56	0.6	0.64
pen_digits	0.51	0.53	0.67	0.57	0.68
abalone	0.75	0.75	0.74	0.74	0.7
sick_euthyroid	0.67	0.65	0.67	0.73	0.86
spectrometer	0.99	0.99	0.37	0.99	0.95
car_eval_34	0.95	0.95	0.95	0.95	0.89
isolet	0.98	0.98	0.96	0.98	0.96
us_crime	0.89	0.87	0.85	0.87	0.87
yeast_ml8	0.63	0.61	0.63	0.64	0.32
scene	0.79	0.79	0.78	0.8	0.76
libras_move	0.97	0.94	0.89	0.88	0.25
thyroid_sick	0.8	0.82	0.72	0.8	0.84
coil_2000	0.62	0.63	0.58	0.65	0.67
arrhythmia	0.96	0.96	0.89	0.96	0.54
solar_flare_m0	0.8	0.73	0.76	0.77	0.66
oil	0.44	0.41	0.55	0.33	0.3
car_eval_4	0.97	0.97	0.97	0.98	0.88
wine_quality	0.67	0.52	0.52	0.56	0.71
letter_img	0.96	0.95	0.91	0.93	0.94
yeast_me2	0.84	0.84	0.82	0.83	0.32
webpage	0.85	0.86	0.48	0.78	0.92
ozone_level	0.87	0.76	0.65	0.89	0.78
mammography	0.88	0.88	0.86	0.82	0.85
protein_homo	0.64	0.39	0.4	0.46	0.72
abalone_19	0.62	0.63	0.66	0.61	0.65

SVC (RBF kernel)

	SMOTE	ROS	ROSE	ADASYN	RUS
ecoli	0.91	0.91	0.93	0.9	0.89
optical_digits	0.58	1.0	0.33	0.79	0.31
satimage	0.61	1.0	0.66	0.61	0.32
pen_digits	0.72	1.0	0.72	0.91	0.52
abalone	0.8	0.81	0.92	0.79	0.76
sick_euthyroid	0.77	1.0	0.36	0.72	0.38
spectrometer	0.33	1.0	0.33	0.33	0.28
car_eval_34	1.0	1.0	0.99	1.0	0.9
isolet	0.85	1.0	0.96	0.88	0.32
us_crime	1.0	1.0	0.83	0.99	0.77
yeast_ml8	1.0	1.0	0.95	1.0	0.57
scene	0.94	1.0	0.42	0.94	0.48
libras_move	1.0	1.0	0.95	1.0	0.91
thyroid_sick	0.77	1.0	0.34	0.77	0.32
coil_2000	0.88	0.99	0.8	0.87	0.39
arrhythmia	0.42	1.0	0.33	0.39	0.28
solar_flare_m0	0.95	0.92	0.91	0.93	0.5
oil	0.32	1.0	0.32	0.34	0.32
car_eval_4	1.0	1.0	0.93	1.0	0.34
wine_quality	0.9	1.0	0.62	0.91	0.45
letter_img	0.99	1.0	0.95	0.99	0.64
yeast_me2	0.88	0.89	0.85	0.87	0.81
webpage	0.99	1.0	0.87	0.99	0.57
ozone_level	0.62	1.0	0.33	0.59	0.3
mammography	0.95	0.96	0.96	0.93	0.88
protein_homo	0.33	1.0	0.33	0.33	0.39
abalone_19	0.89	0.87	0.98	0.9	0.81

DecisionTreeClassifier

	SMOTE	ROS	ROSE	ADASYN	RUS
ecoli	0.92	0.92	0.91	0.9	0.67
optical_digits	0.95	0.94	0.96	0.94	0.88
satimage	0.9	0.88	0.86	0.88	0.86
pen_digits	0.96	0.96	0.92	0.9	0.94
abalone	0.82	0.82	0.95	0.82	0.79
sick_euthyroid	0.97	0.97	0.96	0.96	0.94
spectrometer	0.94	0.97	0.91	0.94	0.82
car_eval_34	0.95	0.95	0.96	0.95	0.94
isolet	0.93	0.92	0.95	0.94	0.89
us_crime	0.91	0.91	0.95	0.91	0.8
yeast_ml8	0.72	0.7	0.74	0.69	0.57
scene	0.83	0.83	0.92	0.81	0.6
libras_move	0.99	0.98	0.91	0.98	0.83
thyroid_sick	0.98	0.99	0.98	0.99	0.94
coil_2000	0.78	0.7	0.97	0.83	0.62
arrhythmia	0.99	1.0	0.96	0.99	1.0
solar_flare_m0	0.89	0.8	0.97	0.87	0.53
oil	0.96	0.97	0.95	0.96	0.71
car_eval_4	0.97	0.97	0.97	0.98	0.94
wine_quality	0.79	0.83	0.82	0.79	0.62
letter_img	0.97	0.96	0.95	0.96	0.94
yeast_me2	0.91	0.93	0.97	0.93	0.77
webpage	0.8	0.79	0.99	0.8	0.77
ozone_level	0.91	0.94	0.92	0.87	0.7
mammography	0.92	0.93	0.95	0.89	0.9
protein_homo	0.94	0.94	0.99	0.92	0.9
abalone_19	0.9	0.93	1.0	0.89	0.44

RandomForestClassifier

	SMOTE	ROS	ROSE	ADASYN	RUS
ecoli	0.93	0.93	0.92	0.9	0.83
optical_digits	0.93	0.92	0.95	0.92	0.9
satimage	0.9	0.88	0.88	0.87	0.9
pen_digits	0.98	0.98	0.95	0.96	0.97
abalone	0.82	0.82	0.84	0.81	0.82
sick_euthyroid	0.81	0.94	0.95	0.84	0.86
spectrometer	0.95	0.99	0.95	0.94	0.95
car_eval_34	0.94	0.95	0.96	0.95	0.92
isolet	0.88	0.87	0.95	0.88	0.84
us_crime	0.89	0.88	0.93	0.89	0.83
yeast_ml8	0.76	0.82	0.79	0.73	0.58
scene	0.82	0.83	0.95	0.8	0.63
libras_move	0.99	1.0	0.92	0.99	0.7
thyroid_sick	0.87	0.8	0.97	0.84	0.83
coil_2000	0.82	0.68	0.97	0.8	0.59
arrhythmia	0.9	0.89	0.93	0.88	0.61
solar_flare_m0	0.87	0.76	0.97	0.83	0.72
oil	0.96	0.94	0.96	0.96	0.81
car_eval_4	0.96	0.95	0.99	0.96	0.88
wine_quality	0.81	0.82	0.82	0.85	0.75
letter_img	0.96	0.94	0.96	0.96	0.87
yeast_me2	0.92	0.94	0.94	0.91	0.76
webpage	0.85	0.76	0.98	0.85	0.73
ozone_level	0.9	0.89	0.88	0.89	0.78
mammography	0.91	0.93	0.95	0.89	0.88
protein_homo	0.93	0.91	0.97	0.91	0.91
abalone_19	0.86	0.89	0.97	0.85	0.56

MLPClassifier

	SMOTE	ROS	ROSE	ADASYN	RUS
ecoli	0.92	0.88	0.87	0.9	0.89
optical_digits	1.0	1.0	0.99	0.99	0.96
satimage	0.78	0.7	0.69	0.57	0.69
pen_digits	1.0	1.0	0.99	1.0	0.99
abalone	0.76	0.76	0.83	0.76	0.76
sick_euthyroid	0.94	0.93	0.91	0.9	0.85
spectrometer	0.97	0.96	0.95	0.95	0.73
car_eval_34	0.99	0.99	0.97	0.99	0.74
isolet	0.99	0.98	0.99	0.99	0.96
us_crime	0.9	0.9	0.92	0.89	0.88
yeast_ml8	0.83	0.84	0.79	0.81	0.58
scene	0.92	0.91	0.89	0.93	0.81
libras_move	0.99	0.99	0.95	0.99	0.91
thyroid_sick	0.94	0.94	0.93	0.91	0.9
coil_2000	0.85	0.81	0.96	0.86	0.64
arrhythmia	0.96	0.96	0.88	0.99	0.46
solar_flare_m0	0.9	0.9	0.96	0.88	0.65
oil	0.7	0.66	0.87	0.73	0.85
car_eval_4	0.99	0.99	0.99	0.99	0.94
wine_quality	0.77	0.75	0.83	0.77	0.67
letter_img	0.99	0.99	0.96	0.99	0.95
yeast_me2	0.89	0.86	0.82	0.85	0.92
webpage	0.96	0.95	0.97	0.95	0.95
ozone_level	0.43	0.8	0.76	0.75	0.25
mammography	0.92	0.9	0.93	0.87	0.9
protein_homo	0.98	0.99	1.0	0.99	0.91
abalone_19	0.82	0.81	0.92	0.83	0.75

AdaBoostClassifier

	SMOTE	ROS	ROSE	ADASYN	RUS
ecoli	0.9	0.95	0.92	0.92	0.77
optical_digits	0.98	0.96	0.99	0.98	0.92
satimage	0.92	0.9	0.88	0.9	0.86
pen_digits	0.98	0.98	0.98	0.98	0.96
abalone	0.81	0.82	0.95	0.81	0.77
sick_euthyroid	0.97	0.96	0.98	0.96	0.93
spectrometer	0.98	0.99	0.96	0.98	0.95
car_eval_34	0.99	1.0	0.99	0.99	0.99
isolet	0.97	0.97	0.97	0.97	0.94
us_crime	0.93	0.94	0.95	0.92	0.85
yeast_ml8	0.79	0.82	0.88	0.8	0.58
scene	0.87	0.88	0.94	0.85	0.69
libras_move	0.99	0.99	0.95	0.99	0.75
thyroid_sick	0.98	0.97	0.98	0.98	0.92
coil_2000	0.91	0.71	0.97	0.9	0.6
arrhythmia	0.99	1.0	0.99	0.98	1.0
solar_flare_m0	0.9	0.75	0.97	0.9	0.64
oil	0.97	0.98	0.96	0.98	0.81
car_eval_4	0.99	0.99	0.99	0.99	0.94
wine_quality	0.84	0.84	0.88	0.83	0.66
letter_img	0.99	0.98	0.99	0.99	0.98
yeast_me2	0.92	0.96	0.98	0.94	0.85
webpage	0.95	0.92	0.99	0.94	0.93
ozone_level	0.96	0.98	0.96	0.95	0.81
mammography	0.91	0.92	0.97	0.87	0.89
protein_homo	0.97	0.96	1.0	0.96	0.95
abalone_19	0.89	0.95	1.0	0.89	0.75

GaussianNB

	SMOTE	ROS	ROSE	ADASYN	RUS
ecoli	0.89	0.89	0.85	0.88	0.89
optical_digits	0.64	0.53	0.48	0.59	0.64
satimage	0.85	0.85	0.8	0.82	0.87
pen_digits	0.75	0.73	0.76	0.85	0.75
abalone	0.76	0.76	0.75	0.76	0.71
sick_euthyroid	0.61	0.6	0.6	0.61	0.55
spectrometer	0.66	0.67	0.94	0.65	0.75
car_eval_34	0.95	0.95	0.95	0.95	0.94
isolet	0.93	0.89	0.91	0.89	0.89
us_crime	0.86	0.85	0.92	0.84	0.84
yeast_ml8	0.73	0.63	0.95	0.72	0.52
scene	0.7	0.7	0.86	0.71	0.75
libras_move	0.96	0.87	0.95	0.85	0.81
thyroid_sick	0.65	0.6	0.59	0.64	0.57
coil_2000	0.47	0.44	0.45	0.47	0.41
arrhythmia	0.65	0.62	0.62	0.64	0.61
solar_flare_m0	0.75	0.72	0.87	0.72	0.5
oil	0.48	0.81	0.79	0.58	0.71
car_eval_4	0.99	0.99	0.99	0.99	1.0
wine_quality	0.75	0.74	0.8	0.75	0.7
letter_img	0.88	0.87	0.88	0.87	0.87
yeast_me2	0.49	0.4	0.37	0.48	0.44
webpage	0.91	0.89	0.96	0.84	0.85
ozone_level	0.79	0.75	0.74	0.79	0.78
mammography	0.86	0.86	0.88	0.77	0.88
protein_homo	0.88	0.88	1.0	0.88	0.89
abalone_19	0.72	0.72	0.68	0.72	0.87

QuadraticDiscriminantAnalysis

	SMOTE	ROS	ROSE	ADASYN	RUS
ecoli	0.79	0.49	0.29	0.47	0.71
optical_digits	0.59	0.54	0.46	0.62	0.62
satimage	0.82	0.8	0.94	0.82	0.78
pen_digits	0.98	0.98	0.97	0.96	0.99
abalone	0.79	0.75	0.95	0.77	0.75
sick_euthyroid	0.6	0.59	0.95	0.61	0.51
spectrometer	1.0	1.0	0.95	1.0	0.61
car_eval_34	0.33	0.33	0.33	0.33	0.35
isolet	1.0	1.0	0.95	0.98	0.59
us_crime	0.97	0.97	0.95	0.98	0.76
yeast_ml8	0.96	0.96	0.95	0.96	0.51
scene	1.0	1.0	0.92	1.0	0.37
libras_move	1.0	1.0	0.97	1.0	0.91
thyroid_sick	0.51	0.46	0.97	0.48	0.4
coil_2000	0.41	0.39	0.35	0.4	0.57
arrhythmia	1.0	1.0	0.61	1.0	0.62
solar_flare_m0	0.78	0.66	0.97	0.7	0.5
oil	1.0	1.0	0.97	1.0	0.57
car_eval_4	0.33	0.33	0.33	0.33	0.28
wine_quality	0.77	0.74	0.87	0.77	0.65
letter_img	0.97	0.96	0.94	0.97	0.94
yeast_me2	0.39	0.36	0.36	0.49	0.28
webpage	0.81	0.74	0.99	0.76	0.87
ozone_level	1.0	1.0	0.92	1.0	0.61
mammography	0.85	0.84	0.89	0.79	0.81
protein_homo	0.95	0.94	1.0	0.94	0.95
abalone_19	0.85	0.82	1.0	0.83	0.81

1. Provost, Foster & Fawcett, Tom. (2001). Robust Classification for Imprecise Environments. Machine Learning. 42. 203-231. 10.1023/A:1007601015854. ↩

2. Yu, H., Hong, S., Yang, X., Ni, J., Dan, Y., Qin, B.: Recognition of Multiple imbalanced cancer types based on DNA microarray data using ensemble classifiers. BioMed Res. Int. 2013, 1–13 (2013) ↩

3. Zhao, X.M., Li, X., Chen, L., Aihara, K.: Protein classification with imbalanced data. Proteins Struct. Funct. Bioinf. 70(4), 1125-1132(2008) ↩

4. Cerf, L., Gay, D., Selmaoui-Folcher, N., Crémilleux, B., Boulicaut, J.F.: Parameter-free classification in multi-class imbalanced data sets. Data Knowl. Eng. 87, 109–129 (2013) ↩

5. Gao, X., Chen, Z., Tang, S., Zhang, Y., Li, J.: Adaptive weighted imbalance learning with application to abnormal activity recognition. Neurocomputing 173, 1927–1935 (2016) ↩

6. Razakarivony, S., Jurie, F.: Vehicle detection in aerial imagery: a small target detection benchmark. J. Vis. Commun. Image Represent. 34, 187–203 (2016) ↩

7. Gao, Z., Zhang, L., Chen, M.-yu., Hauptmann, A.G., Zhang, H., Cai, A.N.: Enhanced and hierarchical structure algorithm for data imbalance problem in semantic extraction under massive video dataset. Multimed. Tools Appl. 68(3), 641–657 (2014) ↩

8. Wang, S., Chen, H., Yao, X.: Negative correlation learning for classification ensembles. In: 2010 International Joint Conference on Neural Networks (IJCNN), pp. 1–8. IEEE (2010) ↩

9. Ganganwar, Vaishali. (2012). An overview of classification algorithms for imbalanced datasets. International Journal of Emerging Technology and Advanced Engineering. 2. 42-47. ↩

10. King, Gary, and Langche Zeng. "Logistic regression in rare events data." *Political analysis* 9.2 (2001): 137-163. ↩

11. Menardi, G., Torelli, N. (2009) Some issues in building and assessing classification rules with extremely skewed datasets. Proceedings of the 7th Meeting of the Classification and data Analysis Group of the Italian Statistical Society (Invited Papers), Catania, ISBN 978-88-6129-406-6, Cleup (Padova). ↩

12. Menardi, G., and N. Torelli. "Training and assessing classification rules with unbalanced data) Working Paper Series." *N* 2 (2010): 2010. ↩

13. Chawla, Nitesh V., et al. "SMOTEBoost: Improving prediction of the minority class in boosting." *European conference on principles of data mining and knowledge discovery*. Springer, Berlin, Heidelberg, 2003. ↩

14. Gue, Kevin R. "A dynamic distribution model for combat logistics." *Computers & Operations Research* 30.3 (2003): 367-381. ↩

15. Ndour, Cheikh, Aliou Diop, and Simplice Dossou-Gbété. "Classification approach based on association rules mining for unbalanced data." *arXiv preprint arXiv:1202.5514* (2012). ↩

16. Liu, Xu-Ying, and Zhi-Hua Zhou. "The influence of class imbalance on cost-sensitive learning: An empirical study." *Sixth International Conference on Data Mining (ICDM'06)*. IEEE, 2006. ↩

17. Zhou, Zhi-Hua, and Xu-Ying Liu. "On multi-class cost-sensitive learning." *Computational Intelligence* 26.3 (2010): 232-257. ↩

18. He, Haibo, and Edwardo A. Garcia. "Learning from imbalanced data." *IEEE Transactions on knowledge and data engineering* 21.9 (2009): 1263-1284. ↩

19. Weiss, Roger D., et al. "Long-term outcomes from the national drug abuse treatment clinical trials network prescription opioid addiction treatment study." *Drug and alcohol dependence* 150 (2015): 112-119. ↩

20. P. Hart, “The condensed nearest neighbor rule,” In Information Theory, IEEE Transactions on, vol. 14(3), pp. 515-516, 1968. ↩

21. M. Kubat, S. Matwin, “Addressing the curse of imbalanced training sets: one-sided selection,” In ICML, vol. 97, pp. 179-186, 1997. ↩

22. Wilson, Dennis L. "Asymptotic properties of nearest neighbor rules using edited data." *IEEE Transactions on Systems, Man, and Cybernetics* 3 (1972): 408-421. ↩

23. I. Mani, I. Zhang. “kNN approach to unbalanced data distributions: a case study involving information extraction,” In Proceedings of workshop on learning from imbalanced datasets, 2003. ↩

24. D. Smith, Michael R., Tony Martinez, and Christophe Giraud-Carrier. “An instance level analysis of data complexity.” Machine learning 95.2 (2014): 225-256. ↩

25. N. V. Chawla, K. W. Bowyer, L. O.Hall, W. P. Kegelmeyer, “SMOTE: synthetic minority over-sampling technique,” Journal of artificial intelligence research, 321-357, 2002. ↩ ↩

26. Han, Hui, Wen-Yuan Wang, and Bing-Huan Mao. "Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning." *International conference on intelligent computing*. Springer, Berlin, Heidelberg, 2005. ↩

27. Felix Last, Georgios Douzas, Fernando Bacao, “Oversampling for Imbalanced Learning Based on K-Means and SMOTE” https://arxiv.org/abs/1711.00837 ↩

28. H. M. Nguyen, E. W. Cooper, K. Kamei, “Borderline over-sampling for imbalanced data classification,” International Journal of Knowledge Engineering and Soft Data Paradigms, 3(1), pp.4-21, 2009. ↩

29. He, Haibo, Yang Bai, Edwardo A. Garcia, and Shutao Li. “ADASYN: Adaptive synthetic sampling approach for imbalanced learning,” In IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), pp. 1322-1328, 2008. ↩

30. G. Batista, R. C. Prati, M. C. Monard. “A study of the behavior of several methods for balancing machine learning training data,” ACM Sigkdd Explorations Newsletter 6 (1), 20-29, 2004. ↩

31. G. Batista, B. Bazzan, M. Monard, “Balancing Training Data for Automated Annotation of Keywords: a Case Study,” In WOB, 10-18, 2003. ↩

32. Wilson, Dennis L. "Asymptotic properties of nearest neighbor rules using edited data." *IEEE Transactions on Systems, Man, and Cybernetics* 3 (1972): 408-421. ↩

33. Tibshirani, Robert J.; Efron, Bradley. An introduction to the bootstrap. *Monographs on statistics and applied probability*, 1993, 57: 1-436. ↩

34. Bowman, Adrian W., and Adelchi Azzalini. *Applied smoothing techniques for data analysis: the kernel approach with S-Plus illustrations*. Vol. 18. OUP Oxford, 1997 ↩ ↩

35. Hofmann, Thomas, Bernhard Schölkopf, and Alexander J. Smola. "Kernel methods in machine learning." *The annals of statistics* (2008): 1171-1220. ↩

36. Silverman, Bernard W. *Density estimation for statistics and data analysis*. Vol. 26. CRC press, 1986. ↩

37. Mower, Jeffrey P. "PREP-Mt: predictive RNA editor for plant mitochondrial genes." *BMC bioinformatics* 6.1 (2005): 96. ↩

38. Flight L, Julious SA. The disagreeable behaviour of the kappa statistic. Pharm Stat. 2015; 14:74–8. ↩

39. Sebastiani F. An axiomatically derived measure for the evaluation of classification algorithms. In: Proceedings of ICTIR 2015 – the ACM SIGIR 2015 International Conference on the Theory of Information Retrieval. New York City: ACM: 2015. p. 11–20. ↩

40. Powers DMW. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness & correlation. J Mach Learn Technol. 2011; 2(1):37–63. ↩

41. Van Rijsbergen, Cornelis J. "A new theoretical framework for information retrieval." *Acm Sigir Forum*. Vol. 21. No. 1-2. New York, NY, USA: ACM, 1986. ↩

42. Chicco, Davide, and Giuseppe Jurman. "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation." *BMC genomics* 21.1 (2020): 6. ↩ ↩

43. Youden, William J. "Index for rating diagnostic tests." *Cancer* 3.1 (1950): 32-35. ↩

44. Henning, Andersen. "Markedness: The First 150 Years." *Markedness in Synchrony and Diachrony, Olga M. Tomic (ed.), Mouton de Gruyter, Berlin–Germany* (1989): 11-46. ↩

45. Fowlkes, Edward B., and Colin L. Mallows. "A method for comparing two hierarchical clusterings." *Journal of the American statistical association* 78.383 (1983): 553-569. ↩

46. Tague-Sutcliffe J. The pragmatics of information retrieval experimentation, revisited. Informa Process Manag. 1992; 28:467–90. ↩

47. Guilford, Joy Paul. "Psychometric methods." (1954). ↩

48. Baldi P, Brunak S, Chauvin Y, Andersen CA, Nielsen H. Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics. 2000; 16(5):412–24. ↩

49. The MicroArray Quality Control (MAQC) Consortium. The MAQC-II Project: a comprehensive study of common practices for the development and validation of microarray-based predictive models. Nat Biotechnol. 2010; 28(8):827–38. ↩

50. Brown JB. Classifiers and their metrics quantified. Mol Inform. 2018; 37:1700127 ↩

51. https://cran.r-project.org/web/packages/ROSE/ROSE.pdf ↩

52. https://imbalanced-learn.org/stable/ ↩

53. https://scikit-learn.org/stable/ ↩

54. https://www.python.org/dev/peps/pep-0008/ ↩

55. https://github.com/scikit-learn-contrib/imbalanced-learn/pull/754 ↩

56. Birch, David L., Anne Haggerty, and William Parsons. *Who's creating jobs?: 1995*. Cognetics, Inc., 1995. ↩

57. Chianca, Thomaz. "The OECD/DAC criteria for international development evaluations: An assessment and ideas for improvement." *Journal of Multidisciplinary Evaluation* 5.9 (2008): 41-51. ↩