Grant Agreement Number: 825225

Safe-DEED

www.safe-deed.eu

D4.2 Baseline prototypes for data valuation

Deliverable number	*D4.2*
Dissemination level	*Public*
Delivery date	*30 November 2019*
Status	*Final*
Author(s)	*Mihnea Tufiş, Ludovico Boratto*

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825225.

Changes Summary

Date	Author	Summary	Version
12.11.2019	Mihnea Tufiş	First draft	0.1
14.11.2019	Patrick Ofner	Review	0.2
14.11.2019	Mark de Reuver	Review	0.3
15.11.2019	Mihnea Tufiş	Comments integration	0.4

Executive summary

This document supports deliverable D4.2 - a demonstrator of baseline prototypes for data valuation. This is a first implementation of the architecture described in deliverable D4.1 and provides a skeleton of the Data Valuation Component's layers. The functionality of the demonstrator has been tested as part of the data trials documented in deliverable D6.1. The current version of the demonstrators located on the Safe-DEED Git repository and can be made publicly available upon request; future versions will be published on the Safe-DEED website.

The rest of the document is structured as follows: Section 2 describes the implementation details of the Data Valuation Component (DVC), including a description of the sub-components, the class diagram of the solution, the data flow between sub-components and an explanation of the output of the DVC. Section 3 describes the structure, the dependencies and how to run the demonstrator package. Section 4 concludes the document and discusses the next steps in the development of the DVC.

Table of Contents

1 Introduction 6

2 Implementation of the DVC 6

2.1 Sub-components............ 6

2.2 Class diagram 8

2.3 Data flow between sub-components............ 8

2.4 Output explanation............ 9

3 Deliverable details 11

3.1 Package structure 11

3.2 Implementation notes and library requirements............ 11

3.3 Running the DVC 11

4 Conclusion and next steps............ 12

5 References 12

6 Annex DVC - output 13

List of Figures

Figure 1 : DVC architecture [2]. 7

Figure 2 : UML class diagram of the DVC. 8

Figure 3 : Data flow diagram (DFD) of the Data Valuation Component (DVC). 9

Figure 4 : Tree structure of the DVC package. 11

Figure 5 : Output example – Descriptive statistics of the FORTHNET CRM dataset. 13

Figure 6 : Example of output dataset profile. Left: Variables section, with toggled details (*ACCNT_ID*, *CONTRACT_ID*) and warnings (*DATE_ISSUED*). Right: Correlations section between pairs of 2 variables (Pearson and Spearman matrices). 14

Abbreviations

DVC Data Valuation Component

ADAS Automatic Data Analysis and Scoring

S2VM Score-to-Value Mapping

CPL Communication and Presentation Layer

QDSC Qualitative Data Scoring Component

WP Work Package

ML Machine Learning

1 Introduction

This document provides a description of the development tasks within deliverable D4.2 - a demonstrator of the baseline prototypes for data valuation. The code is currently stored on the Safe-DEED git repository and is available on request; more refined versions will later be available on the Safe-DEED website.

This deliverable represents a first implementation of the high-level architecture described in deliverable D4.1 [2]. It provides a skeleton of the Data Valuation Component (DVC) and includes basic functionalities of each of the component's layers. The functionality has been validated in D6.1 [1] using three data sets: two internal data sets provided by FORTHNET (CRM and viewership) and one openly available data set (Titanic passengers).

2 Implementation of the DVC

In this section we provide a description of each of the sub-components of the DVC, as they have been described in the DVC architecture (Fig. 1) defined in deliverable D4.1 [2].

We are then presenting the class diagram of the DVC implementation (Fig. 2) and the data flow diagram (Fig. 3) describing the typical flow of data between components. We conclude with an illustration and explanation of a sample output of the data valuation process.

2.1 Sub-components

The DVC comprises of the following sub-components, as they are illustrated in the platform architecture laid out in deliverable D4.1 [2] (Fig. 1):

1. Qualitative information extraction and Data Scoring sub-component (QDSC);
2. Data Ingestion Layer (DIL);
3. Automatic Data Analysis and Scoring (ADAS);
4. Score-to-Value Mapping (S2VM);
5. Communication and Presentation Layer (CPL).

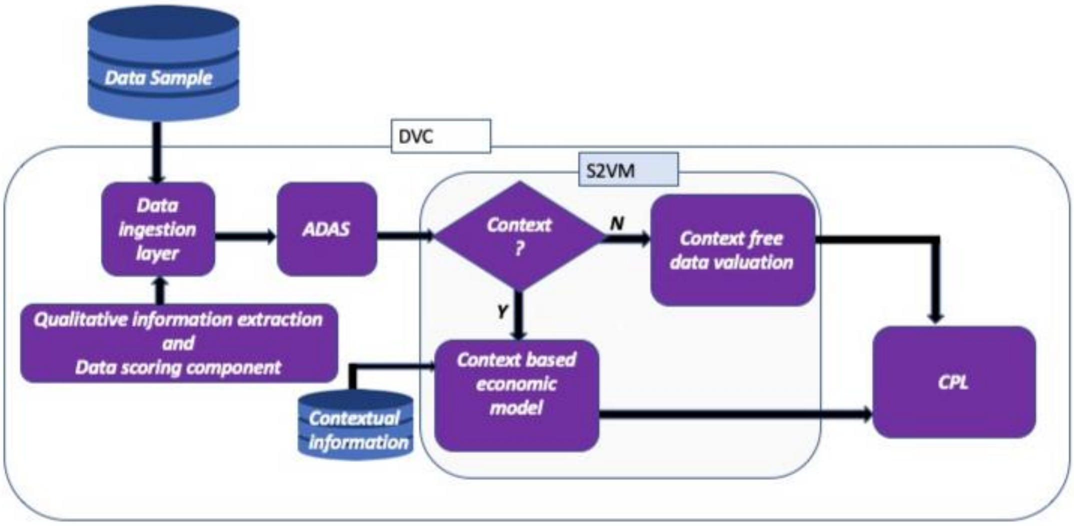

Data Sample
DVC
S2VM
Data ingestion layer
ADAS
Context ?
N
Y
Context free data valuation
Qualitative information extraction and Data scoring component
Context based economic model
Contextual information
CPL

Figure 1 : DVC architecture [2].

Next, we give a description of the roles of each of these modules in the data valuation process.

Qualitative information extraction and Data Scoring sub-component (QDSC)

The first stage of the data valuation process directly involves the user, who is required to provide information about the context in which they wish to valuate an input dataset. It comprises of a series of questionnaires focused on data provenance, acquisition cost, business impact. The context is exchanged in a JSON format which is fed to this component. Finally, the component computes and returns a context-based score.

Data Ingestion Layer (DIL)

DIL represents the entry-point of the data into the platform. The layer detects the data format and performs the suitable operations for ingesting it. Currently the only supported formats are **CSV** and **XLS(X)**, with the capacity of choosing a specific datasheet for the latter. If available, metadata extraction is also performed at this step. On a long term, the intention is for this layer to extract available metadata directly from marketplaces (e.g., Data Market Austria).

Automatic Data Analysis and Scoring (ADAS)

Prepares the loaded data and performs a set of analytic operations to extract the intrinsic properties of the dataset:

- data shape and size;
- data type inference;
- the profile of each field:
  - missing values;
  - distribution of the data from each field;
- exploitability by means of different machine learning algorithms:
  - regression;
  - classification;
  - clustering;

Based on these results, a new score will be generated, considering data quality assessment criteria such as data completeness, uniqueness, timeliness, validity, accuracy, consistency, and data relevance.

Score-to-Value Mapping (S2VM)

If the market segment to be addressed is provided and a pricing model can be applied , the S2VM will attempt to generate an economic value for the dataset. In addition to the two previously computed scores, this layer also considers market dynamics and a history of previous, similar datasets, among other external factors. This module will implement a rule-based economic model which considers all previously enumerated factors to output the economic value of the input dataset.

Communication and Presentation Layer (CPL)

This layer acts as an interface with the user, to report the final results of the data valuation process. It displays in a clear manner the following:

- the QDSC and ADAS scores;
- the economic value of the dataset;
- a report of the dataset profile;
- a report of the applicable machine learning models.

2.2 Class diagram

Fig. 2 illustrates the UML class diagram of the current DVC implementation.

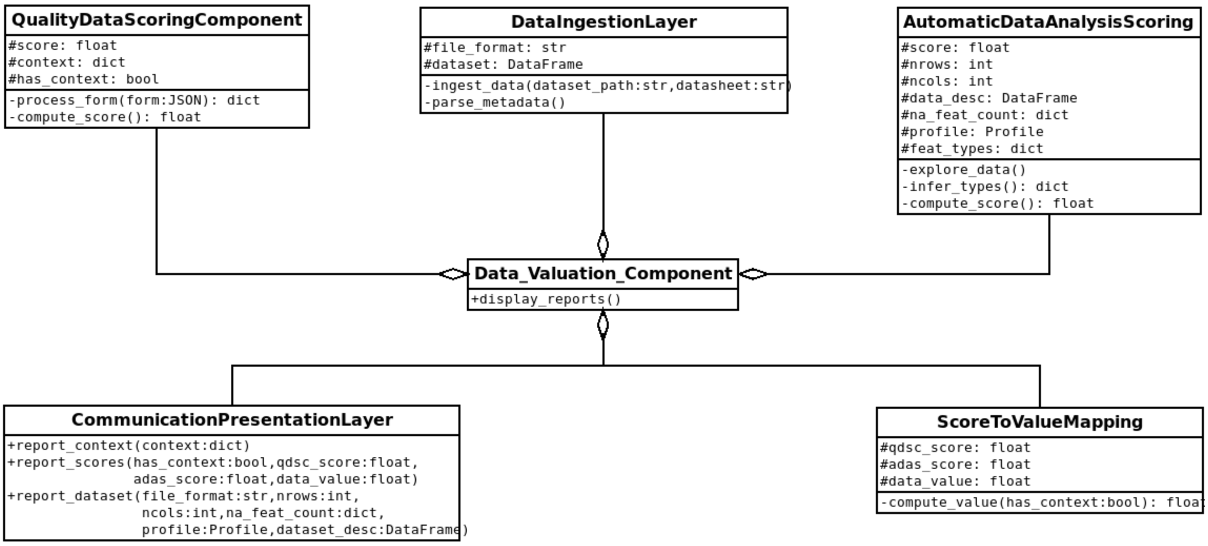


Figure 2 : UML class diagram of the DVC.

2.3 Data flow between sub-components

In this section we present an overview of the input and output of the DVC together with a data flow diagram for the entire component.

Input:

- A dataset (or snapshot thereof). Only CSV and XLS(X) are supported for this version of the component;
- User provided context, in the form of a JSON file.

Output:

- A set of scores, which evaluate the input dataset from three perspectives: quality, exploitability through Machine Learning (ML), economic value;
- A set of reports based on the analysis of the intrinsic properties of the dataset (format, shape, data types, missing values, duplicates).

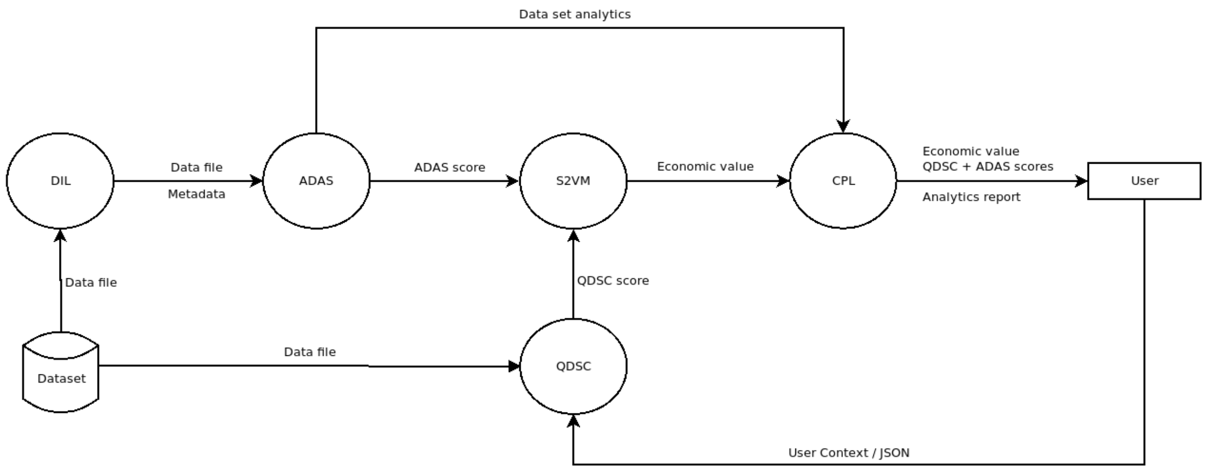


Figure 3 : Data flow diagram (DFD) of the Data Valuation Component (DVC).

2.4 Output explanation

In the following, we give a sample output of the DVC, together with an explanation of each of the output sections.

For this version of the DVC, the focus has been on illustrating basic functionality of the CPL output. Once the development of the DVC will advance, we are considering the presentation of the output in a more user-friendly manner. Thus, there are two types of outputs being generated:

1) console output (which in a future version will be exported in a more standard fashion);
2) the data set profiles in HTML format.

Console output

1. A textual confirmation of each component being properly initialized:

```
QDSC

DIL

ADAS

S2VM

CPL
```

2. The content of the user-provided context file (JSON). If no such file is supplied, we are in the case of non-contextual valuation. This output is destined to be simply a reminder of the context in which the valuation is performed.

```
{'enterprise': {'previous_use': 'true', 'added_value': 'true',
'collection_cost': '10', 'storage_cost': '2', 'processing_cost': '4',
'improvement_potential': 'true', 'customer_reach': 'true',
'dept_contribution': 'marketing', 'hierarchy_use': 'executive'},
'metadata': {'provenance': 'survey', 'description': 'medical',
'features_desc': 'false', 'aggregated': 'false', 'unique': 'false',
'transformed': 'false', 'anonymized': 'true', 'encrypted': 'false'},
'legal': {'contract_bound': 'false', 'user_consent': 'true',
'license': 'open', 'access_difficultty': 'easy'}, 'applications':
{'exploration': 'true', 'regression': 'true', 'classification':
'false', 'clustering': 'false', 'supervised': 'true', 'unsupervised':
'true'}}
```

3. The scores computed by each of the sub-components (QDSC, ADAS) and the final score (Dataset Value):

```
QDSC score:  1.0          ADAS score:  1.0       Dataset Value:  2.0
```

4. Dataset summary. First, the inferred file format and its shape (number of records, number of fields):

```
File format:  .xlsx

Data shape: Rows = 1048575, Cols = 22
```

5. Followed by descriptive statistics (Fig. 5 in Annex) for each column of the dataset: number of records, number of unique values, top value, frequency of the top value, mean, standard deviation, minimum and maximum values, value at the main quantiles.

6. Number of missing values for each field of the dataset:

```
Missing values (fields):

MONTH                      0         Unnamed_11             0
CUSTOMER_ID                0         Unnamed_12             0
ASSET_ID                   0         Unnamed_13             0
PROVISION_ID               0         Unnamed_14             0
CONTRACT_ID                0         Unnamed_15             0
DATE_ISSUED                0         Unnamed_16             0
START_DATE                 0         Unnamed_17             0
END_DATE                   0         Unnamed_18             0
REVENUE_TYPE_ID            0         Unnamed_19          2385
REVENUE_TYPE_DESC          0         Unnamed_20        126264
REVENUE                    0         Unnamed_21       1046739
```

HTML profile

In order to have an overview of the properties of each of the columns in a dataset, we generate a profile of the dataset (Fig. 6 in Annex), comprising of the following sections:

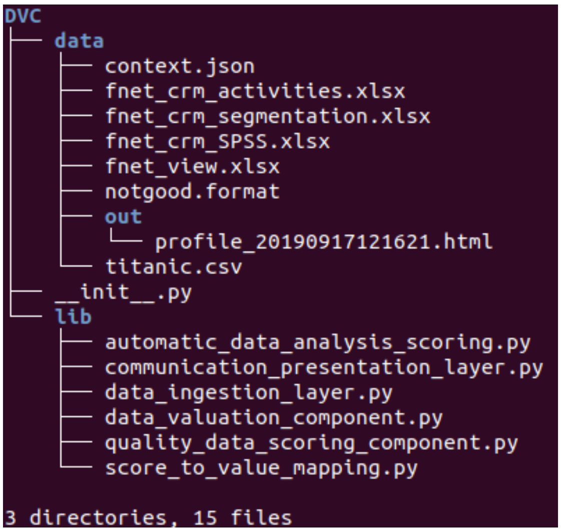
1. Dataset info: size, shape, duplicate percentage;
2. Variables types: how many columns of each type;
3. Warnings: if columns contain a large proportion of a particular value, or large proportion of 0s, etc.
4. Variables: a summary of the descriptive stats for each column, including a histogram of the distribution, number of unique values, number of missing values, mean, standard deviation, max, min. Then, if you toggle the details of each column, you can see detailed quantile stats, descriptive stats, the histogram of the distribution, common and extreme values (top-5 and bottom-5).
5. Correlations: a set of correlation matrices based on different correlation coefficients between all pairs of columns. This informs the suggestion to discard from model design those columns that have a high correlation coefficient.
6. Missing values: a histogram of how many values are present in each of the columns.
7. Sample: a sample of the head and tail records of the dataset.

3 Deliverable details

3.1 Package structure

The deliverable is now available on the internal Git repository and can be publicly accessed upon request. Future versions will be made publicly available on the Safe-DEED website.

The structure of the package that forms deliverable D4.2 is the following:



```
DVC
├── data
│   ├── context.json
│   ├── fnet_crm_activities.xlsx
│   ├── fnet_crm_segmentation.xlsx
│   ├── fnet_crm_SPSS.xlsx
│   ├── fnet_view.xlsx
│   ├── notgood.format
│   ├── out
│   │   └── profile_20190917121621.html
│   └── titanic.csv
├── __init__.py
└── lib
    ├── automatic_data_analysis_scoring.py
    ├── communication_presentation_layer.py
    ├── data_ingestion_layer.py
    ├── data_valuation_component.py
    ├── quality_data_scoring_component.py
    └── score_to_value_mapping.py

3 directories, 15 files
```

Figure 4 : Tree structure of the DVC package.

3.2 Implementation notes and library requirements

The current version of the prototype was developed using Python 3.7 and in order to run requires the following external libraries. At this stage, there are no particular requirements with respect to the operating system:

json 0.8.4

matplotlib 3.0.3

numpy 1.16.2

pandas 0.24.2

pandas-profiling 1.4.1

3.3 Running the DVC

This section describes how to run the current DVC prototype.

1. Download the DVC package from the Safe-DEED git repository.
2. Place a *context.json* (context file) and the desired input file in the *./data/* folder;
3. Open a terminal window, navigate to the project folder and run the following command. The –datasheet is available only for XLS(X) files and allows the import and valuation of the specified datasheet only. If none is specified, the first datasheet will be used by default.

```
> python3 __init__.py FILE_NAME [--datasheet DATASHEET_NAME]
```

4. In the current prototype, the output is displayed in the terminal window. Additionally, an HTML report of the data profile is generated in the *./data/out/* folder.

4 Conclusion and next steps

Deliverable D4.2 demonstrates baseline functionality of the sub-components of the DVC. This is an initial implementation of the architecture described in deliverable D4.1 [2]. The output of the CPL provides us with a sample of profiling, qualitative and structural information about an input data set.

We have now proceeded to a literature review of metrics and algorithms for valuating data from different perspectives: quality, exploitability, economical. These will result in an ensemble of techniques for valuating a data set within a context, as well as in a context-free situation. Subsequently, the implementation of these techniques will be carried out by extending the baseline functionality delivered in D4.2.

Some of the more complex functionalities that we envision for the subsequent versions of the DVC are:

- increase the range of input files and expand the file format interpretation functionality (DIL);
- development of the data sampling functionality, including a de-anonymization attempt of sensitive fields, to be performed prior to the data analysis stage;
- basic ML algorithms (regression, classification, clustering) to allow for a proper valuation of the usability of the input data set (ADAS);
- development of a rule based economic model for the valuation of the economic aspects of the input data set (S2VM).

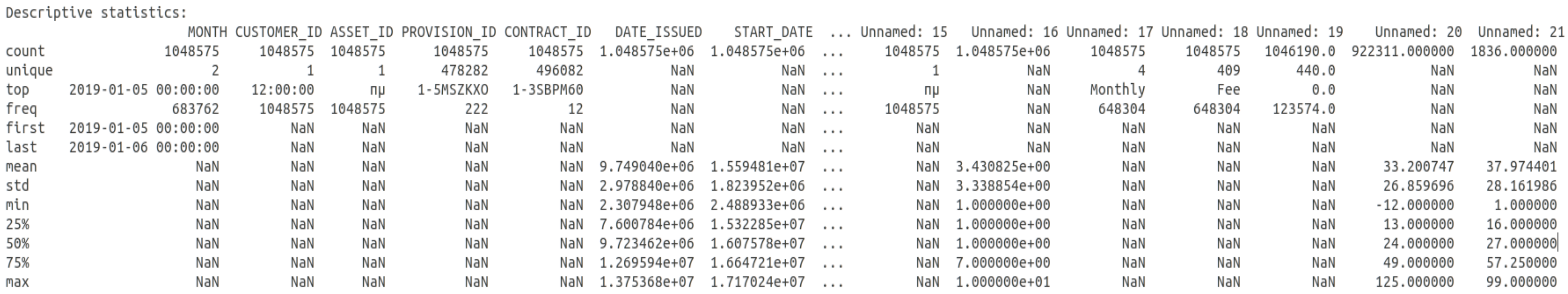
5 References

[1] Markopoulos I 2019, Initial phase personal data trials. Safe-DEED Deliverable D6.1 (draft). Source: TBC when final.

[2] Paraschiv M., Boratto L., Kassa Y., Tufiş M. (2019): Report on Requirements and Design. Safe-DEED Deliverable D4.1. Source: https://tinyurl.com/r4xlmeu (accessed November 12, 2019).

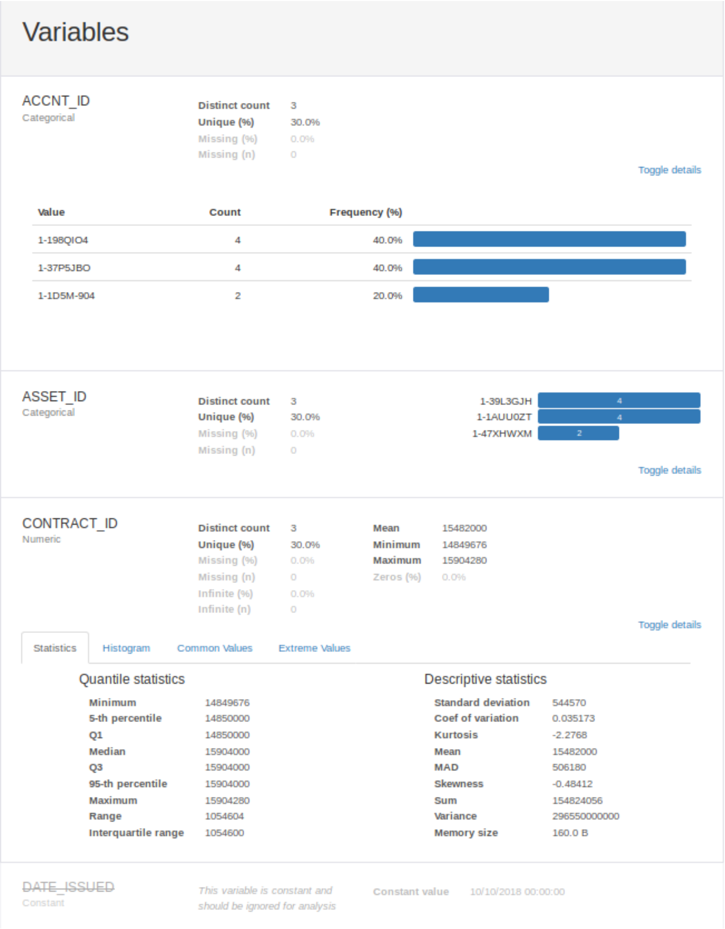
6 Annex DVC - output

In this section we illustrate screen captures of 2 particular output sections of the DVC applied to a test data set: the descriptive statistics section (Fig. 5) and the data set profile section (Fig. 6).



```
Descriptive statistics:
                     MONTH CUSTOMER_ID ASSET_ID PROVISION_ID CONTRACT_ID   DATE_ISSUED    START_DATE ... Unnamed: 15   Unnamed: 16 Unnamed: 17 Unnamed: 18 Unnamed: 19   Unnamed: 20 Unnamed: 21
count              1048575     1048575  1048575      1048575     1048575  1.048575e+06  1.048575e+06 ...     1048575  1.048575e+06     1048575     1048575   1046190.0 922311.000000 1836.000000
unique                   2           1        1       478282      496082           NaN           NaN ...           1           NaN           4         409       440.0           NaN         NaN
top    2019-01-05 00:00:00    12:00:00       пμ    1-5MSZKXO   1-3SBPM60           NaN           NaN ...          пμ           NaN     Monthly         Fee         0.0           NaN         NaN
freq                683762     1048575  1048575          222          12           NaN           NaN ...     1048575           NaN      648304      648304    123574.0           NaN         NaN
first  2019-01-05 00:00:00         NaN      NaN          NaN         NaN           NaN           NaN ...         NaN           NaN         NaN         NaN         NaN           NaN         NaN
last   2019-01-06 00:00:00         NaN      NaN          NaN         NaN           NaN           NaN ...         NaN           NaN         NaN         NaN         NaN           NaN         NaN
mean                   NaN         NaN      NaN          NaN         NaN  9.749040e+06  1.559481e+07 ...         NaN  3.430825e+00         NaN         NaN         NaN     33.200747   37.974401
std                    NaN         NaN      NaN          NaN         NaN  2.978840e+06  1.823952e+06 ...         NaN  3.338854e+00         NaN         NaN         NaN     26.859696   28.161986
min                    NaN         NaN      NaN          NaN         NaN  2.307948e+06  2.488933e+06 ...         NaN  1.000000e+00         NaN         NaN         NaN    -12.000000    1.000000
25%                    NaN         NaN      NaN          NaN         NaN  7.600784e+06  1.532285e+07 ...         NaN  1.000000e+00         NaN         NaN         NaN     13.000000   16.000000
50%                    NaN         NaN      NaN          NaN         NaN  9.723462e+06  1.607578e+07 ...         NaN  1.000000e+00         NaN         NaN         NaN     24.000000   27.000000
75%                    NaN         NaN      NaN          NaN         NaN  1.269594e+07  1.664721e+07 ...         NaN  7.000000e+00         NaN         NaN         NaN     49.000000   57.250000
max                    NaN         NaN      NaN          NaN         NaN  1.375368e+07  1.717024e+07 ...         NaN  1.000000e+01         NaN         NaN         NaN    125.000000   99.000000
```

Figure 5 : Output example – Descriptive statistics of the FORTHNET CRM dataset.



Variables

ACCNT_ID
Categorical

Distinct count	3
Unique (%)	30.0%
Missing (%)	0.0%
Missing (n)	0

Toggle details

Value	Count	Frequency (%)
1-198QIO4	4	40.0%
1-37P5JBO	4	40.0%
1-1D5M-904	2	20.0%

ASSET_ID
Categorical

Distinct count	3
Unique (%)	30.0%
Missing (%)	0.0%
Missing (n)	0

1-39L3GJH	4
1-1AUU0ZT	4
1-47XHWXM	2

Toggle details

CONTRACT_ID
Numeric

Distinct count	3	Mean	15482000
Unique (%)	30.0%	Minimum	14849676
Missing (%)	0.0%	Maximum	15904280
Missing (n)	0	Zeros (%)	0.0%
Infinite (%)	0.0%		
Infinite (n)	0		

Toggle details

Statistics Histogram Common Values Extreme Values

Quantile statistics

Minimum	14849676
5-th percentile	14850000
Q1	14850000
Median	15904000
Q3	15904000
95-th percentile	15904000
Maximum	15904280
Range	1054604
Interquartile range	1054600

Descriptive statistics

Standard deviation	544570
Coef of variation	0.035173
Kurtosis	-2.2768
Mean	15482000
MAD	506180
Skewness	-0.48412
Sum	154824056
Variance	296550000000
Memory size	160.0 B

~~DATE_ISSUED~~
Constant

This variable is constant and should be ignored for analysis

Constant value 10/10/2018 00:00:00

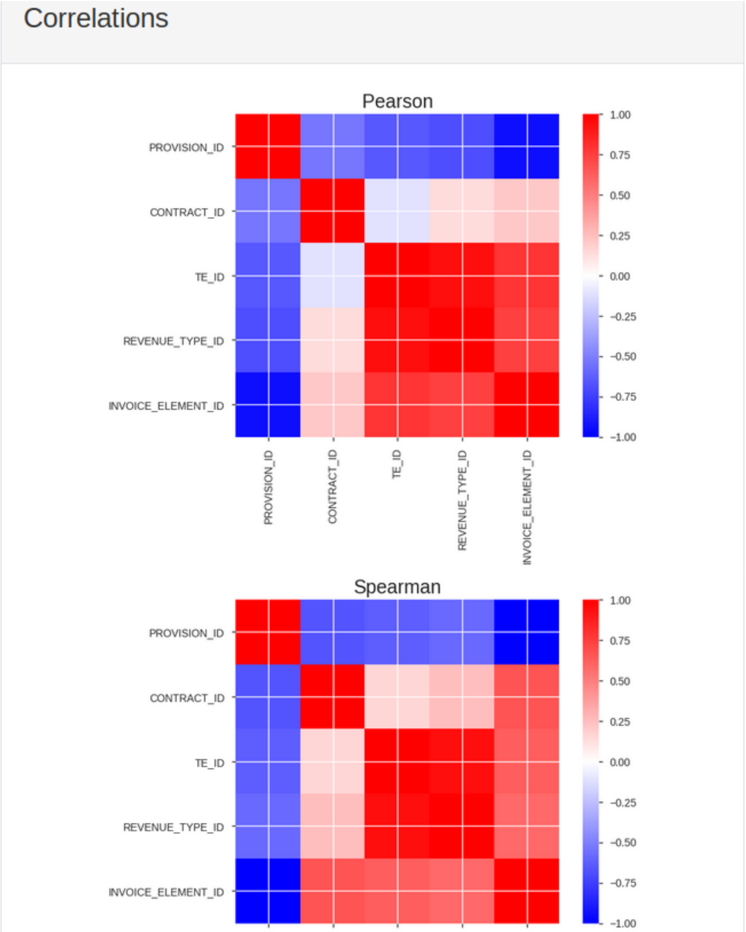


Figure 6 : Example of output dataset profile. Left: Variables section, with toggled details (*ACCNT_ID*, *CONTRACT_ID*) and warnings (*DATE_ISSUED*). Right: Correlations section between pairs of 2 variables (Pearson and Spearman matrices).