|  |  |  |  |
| --- | --- | --- | --- |
| Logo, company name  Description automatically generated | | **DS 2022** | |
| Data Science Project | | | |
| **Team nr:** 21 | **Student1 :** Basanta Poudel | | **IST nr:** 80894 |
| **Student 2 :** \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ | | **IST nr:** \_\_\_\_\_\_\_ |
| **Student 3 :** \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ | | **IST nr:** \_\_\_\_\_\_\_ |

This document presents a template for the Data Science Project report. It specifies the mandatory format and suggests the structure to follow. All text with grey background shall be replaced with the analysis made over the datasets.

Classification

# Data Profiling

May be used to describe any useful observation about the data, and that was used in the current project. An example is the use of any domain knowledge to process the data or evaluate the results. **Shall not exceed 200 characters.**

## Data Dimensionality

For both datasets, there are plenty records for analysis. However, there are also many variables (50 & 52). This may cause problems regarding the curses of dimensionality and overfitting. It is key to select the right features later on, to reduce the number of dimensions. Also, there are many missing values for three variables in dataset 1. Instances with ‘’?’’ are counted as a missing value. Strategies have to be created how to deal with these.

For dataset 2, there are no missing values (also no NaN or ‘’?’’), and none of them are symbolic.

Shall contain all relevant information and charts respecting to the data dimensionality perspective, such as the number of records and number of dimensions, and their impact on the following analysis. **Shall not exceed 200 characters.**

Figure Nr Records x Nr variables for dataset 1 (left) and dataset 2 (right)

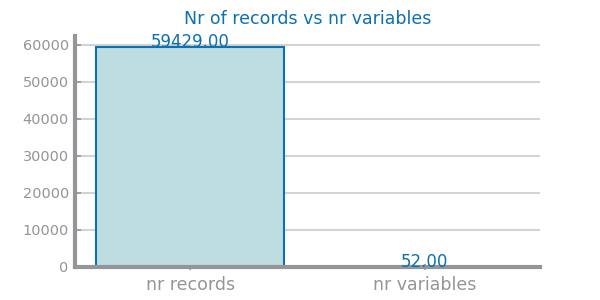
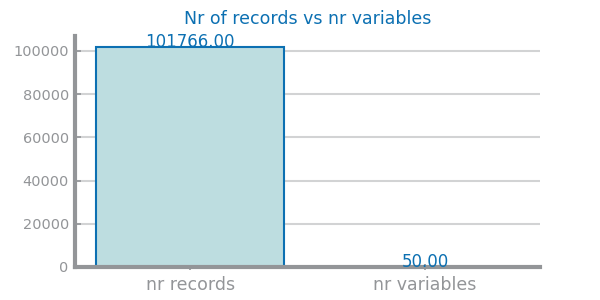


Figure : records vs variables dataset 1 Figure : records vs variables dataset 2

Figure Nr variables per type for dataset 1 (left) and dataset 2 (right)

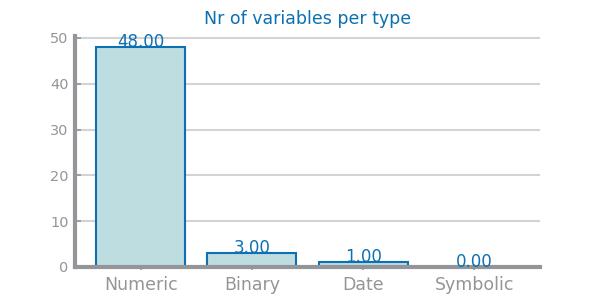
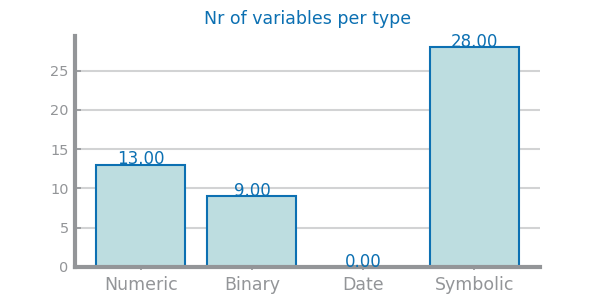


Figure : nr variables per type – dataset 1 Figure : Nr variables per type - dataset 2

Figure Nr missing values for dataset 1 (left) and dataset 2 (right)

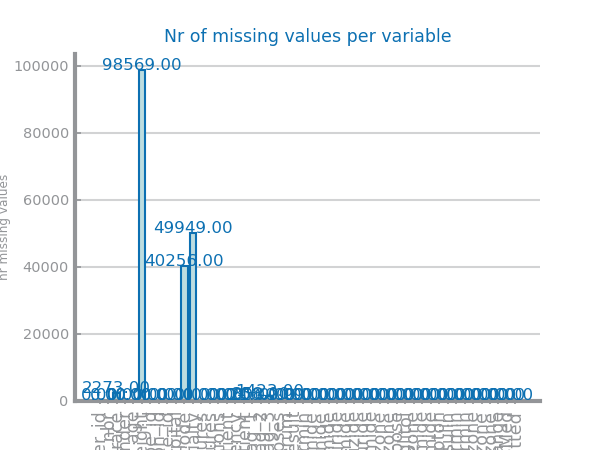


Figure : missing values dataset 1

Figure : Missing values dataset 2

## Data Distribution

* **Class distribution**
* **Outliers**
* **Magnitude of variables**

**Dataset 1**

The most striking observation for dataset 1 is the number of variables which contain outliers. This can be seen in the graph ‘’nr outliers per variable’’ and can be confirmed in the boxplots and histograms for individual variables. This is an issue for further analysis, since it distorts the distribution of the variables. Strategies need to be taken carefully to deal with this. Furthermore, the classes are significantly uneven distributed. This will cause problems with model training; for example the records from class ‘’<30’’ to be ignored by the model.

The boxplots also clearly show the different ranges of the variables. For model training, this should be taken care of, since it is not desired that variables of higher magnitude are considered more important.

The variables ‘’time in hospital’’, ‘’num procedures’’, ‘’outpatient’’, ‘’inpatient’’ and ‘’emergency’’ seem to follow an exponential distribution, as can be expected. The rest of the numeric variables seem to follow a normal distribution, though some distorted by the many outliers.

Most variables about the medical features (symbolic) for medications are very unevenly distributed. However, this does not have a negative impact on model training.

**Dataset 2**

The number of outliers is also an issue for dataset 2. Same as for dataset 1, these have to be treated.

For class distribution almost seems to be perfectly distributed. Balancing the classes seems not to be needed for this dataset.

The boxplots do again show a large difference in magnitude for all variables. This has to be scaled, in order to be used for further processing.

All the numeric variables of this dataset seem to follow the normal distribution best. There is no variable which is distributed exponentially.

Shall contain all relevant information and charts respecting to the data distribution perspective, such as each variable distribution, type, domain and range. May be used to describe any useful observation about the data, and that was used in the current project. **Shall not exceed 500 characters**.

Figure Global boxplots dataset 1 (left) and dataset 2 (right)

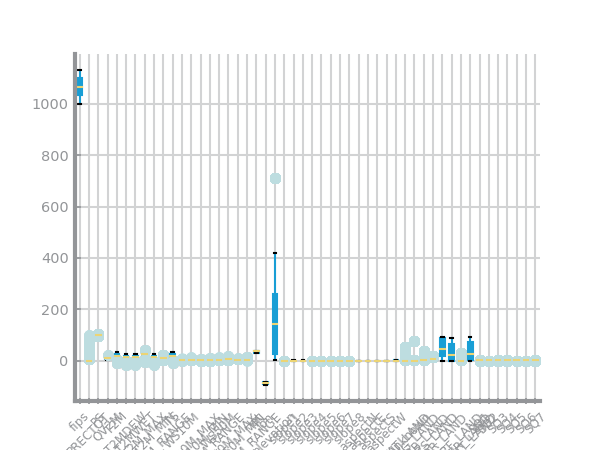
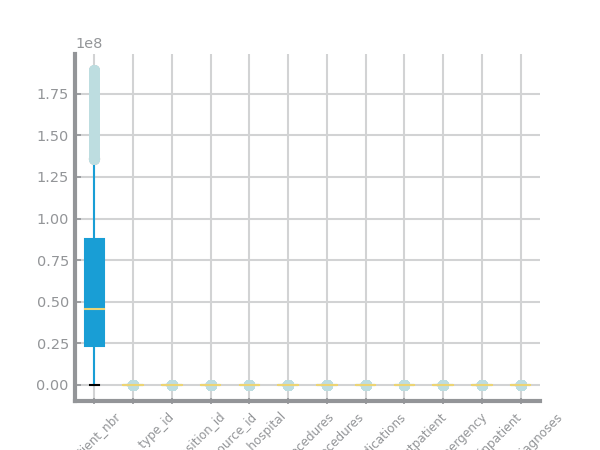


Figure : global boxplot dataset 2

Figure : global boxplot dataset 1

Figure Single variable boxplots for dataset 1



Figure Single variable boxplots s for dataset 2



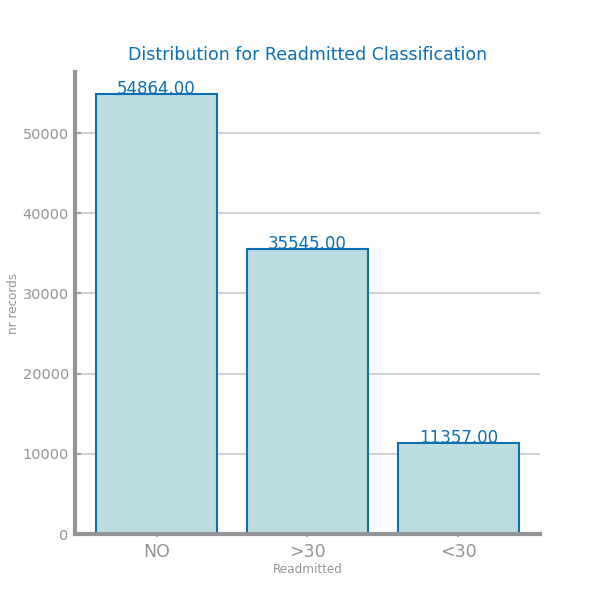
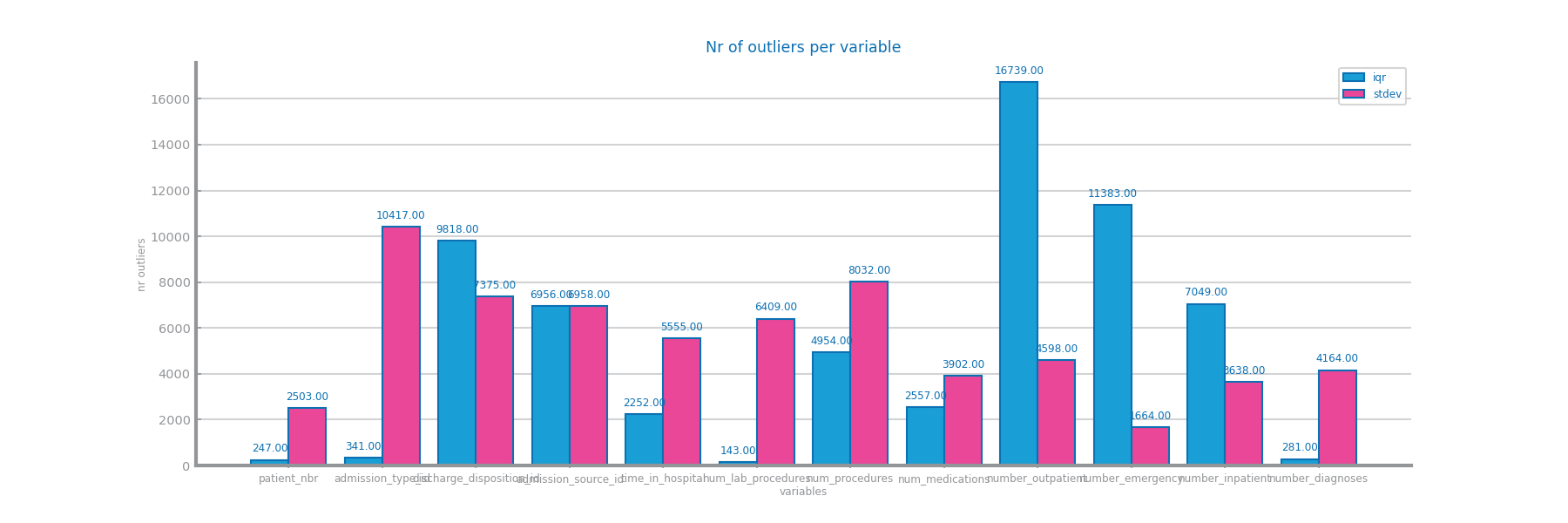
Figure Histograms for dataset 1

Figure : Number of outliers per variable - dataset 1

Figure :class distribution dataset 1

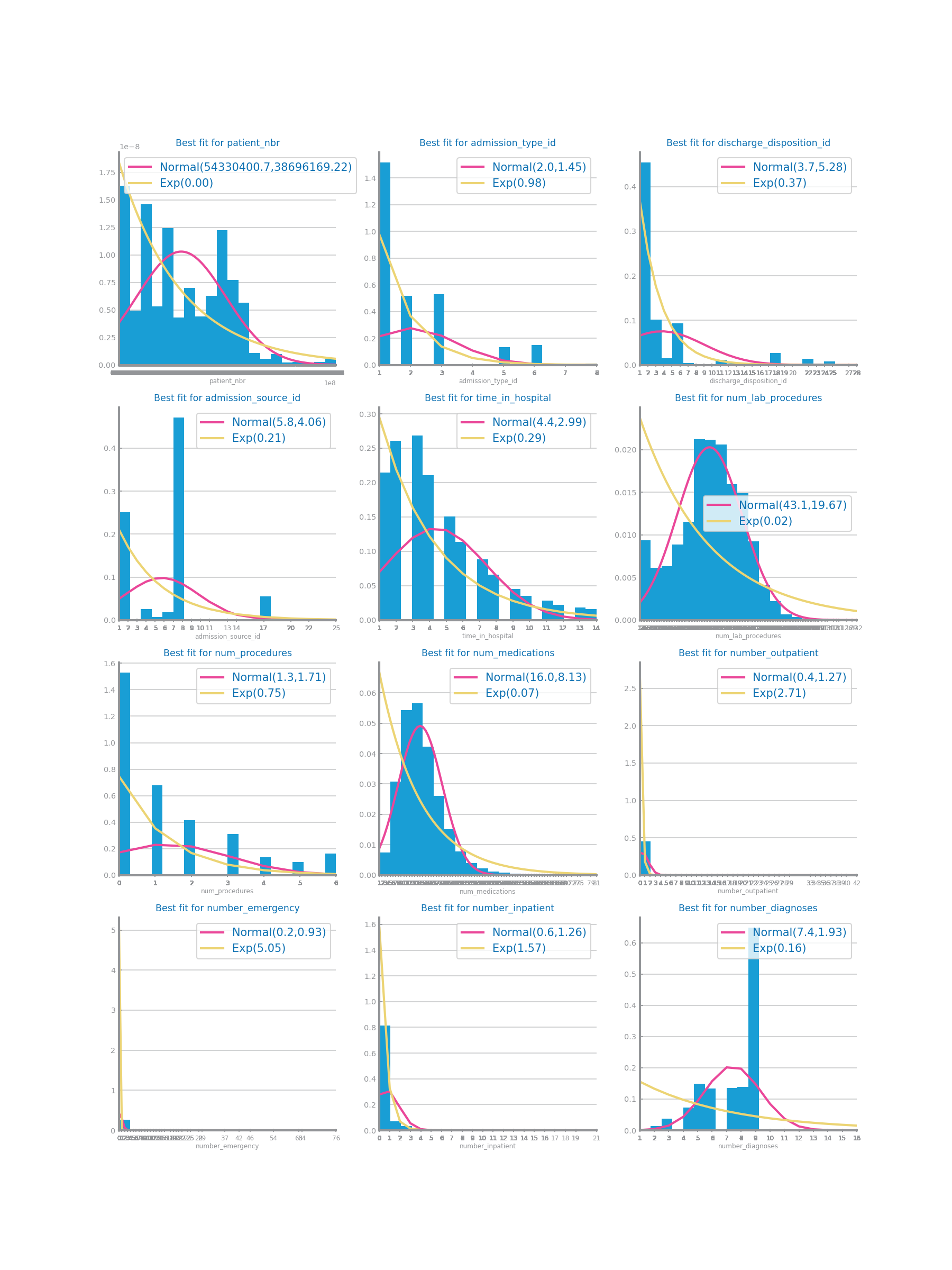


Figure : numeric distribution - Dataset 1

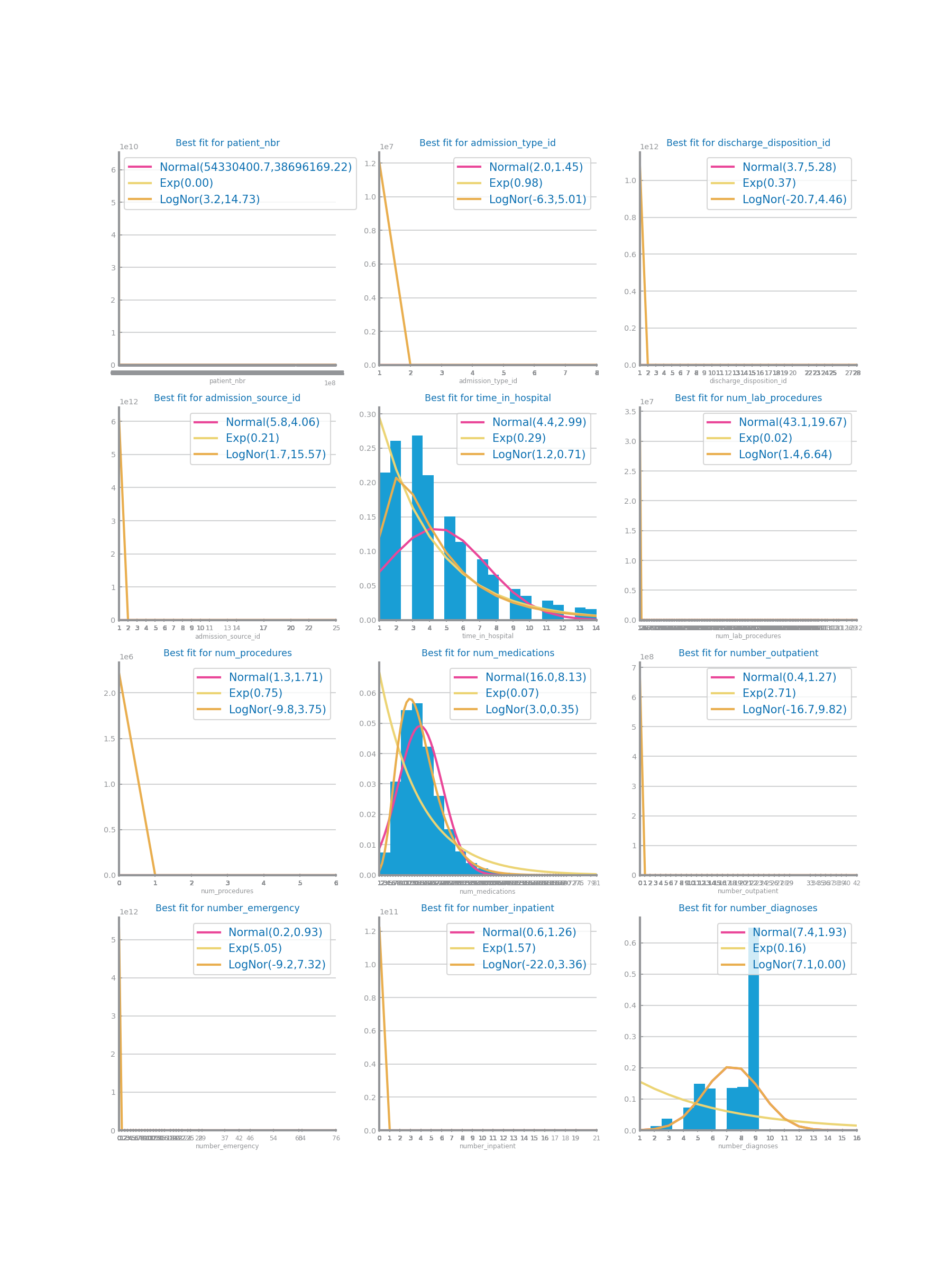


Figure : Numeric distribution with log - dataset 1

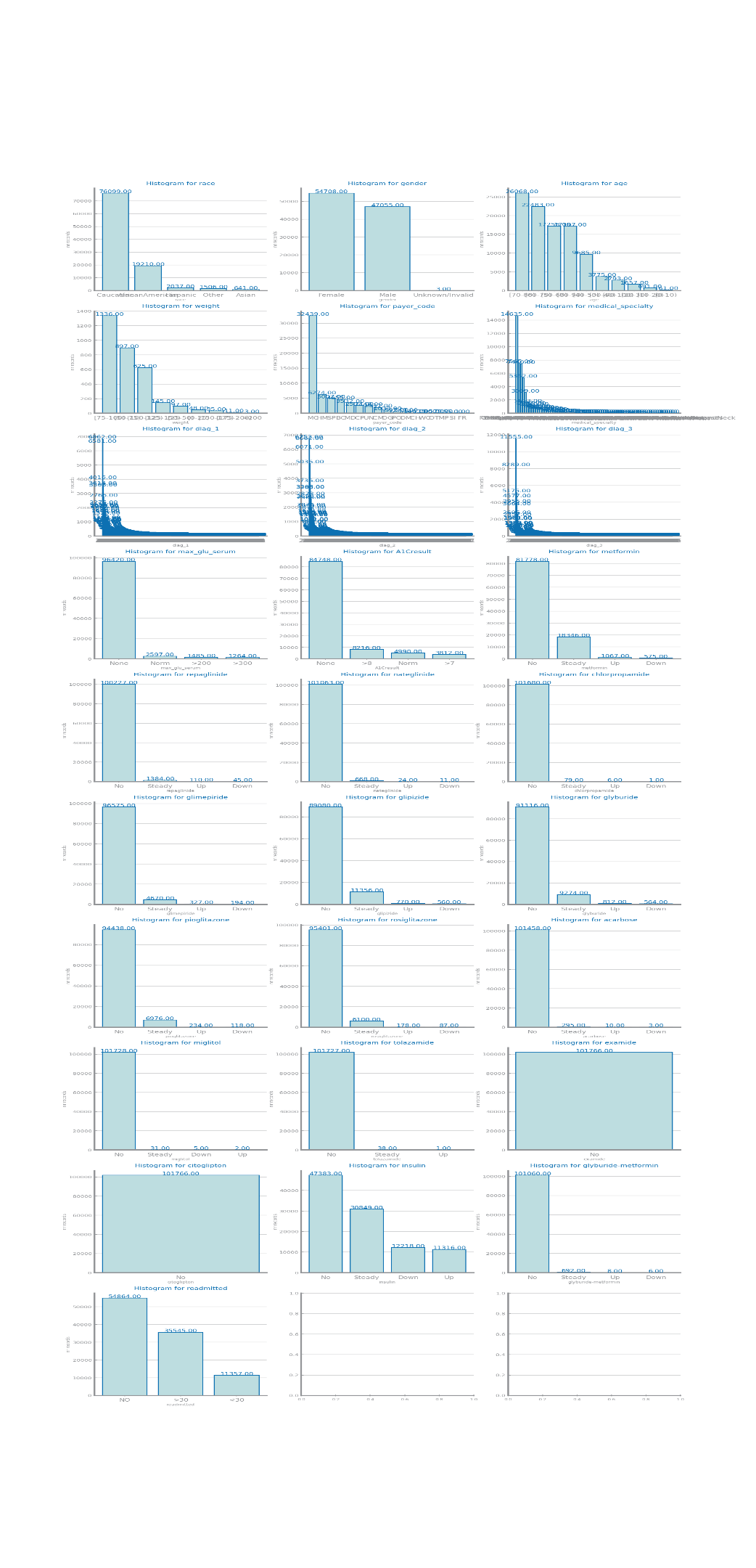


Figure : symbolic variables histograms dataset 1

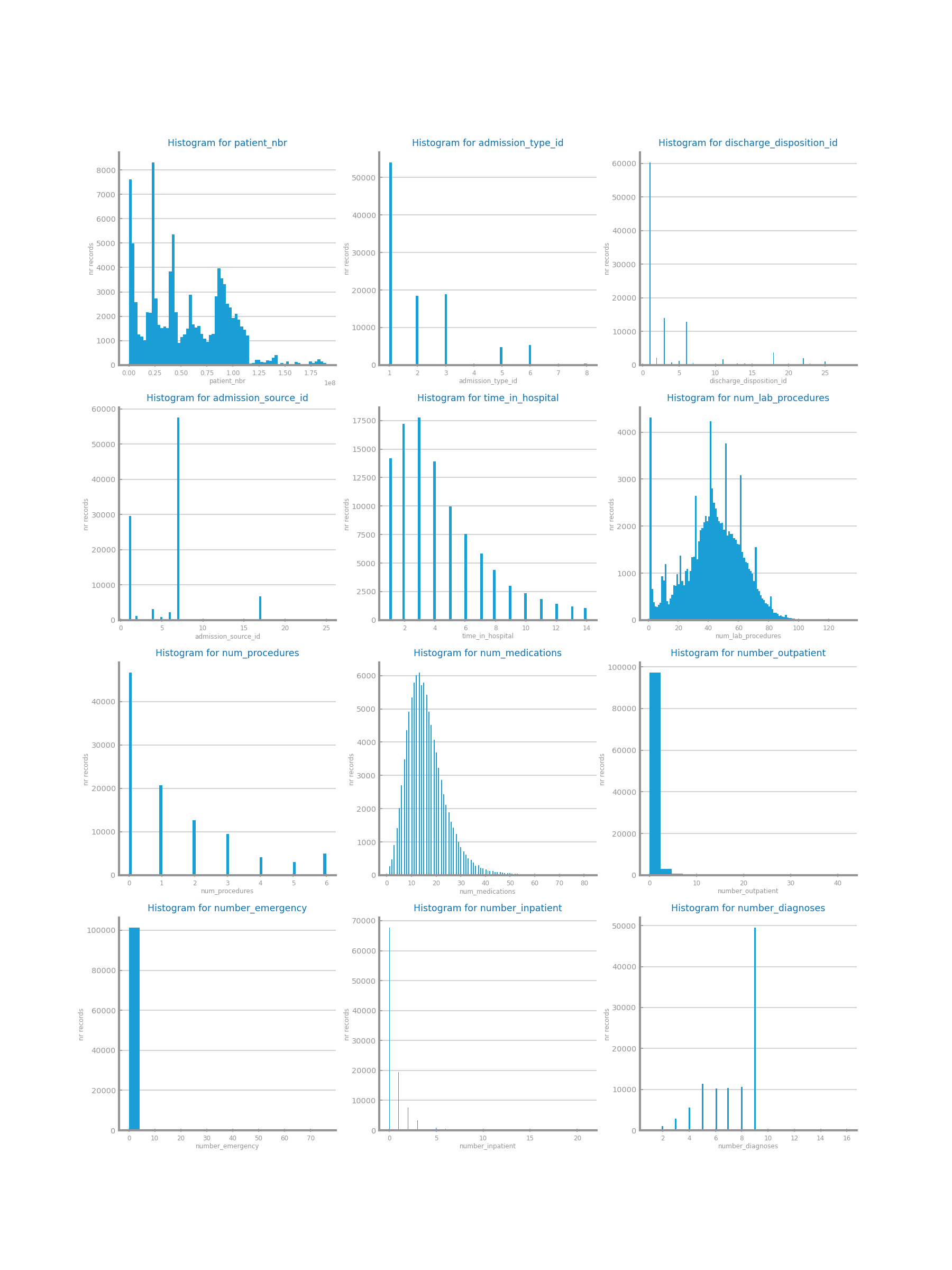


Figure : single histograms numeric - dataset 1

Figure Histograms for dataset 2

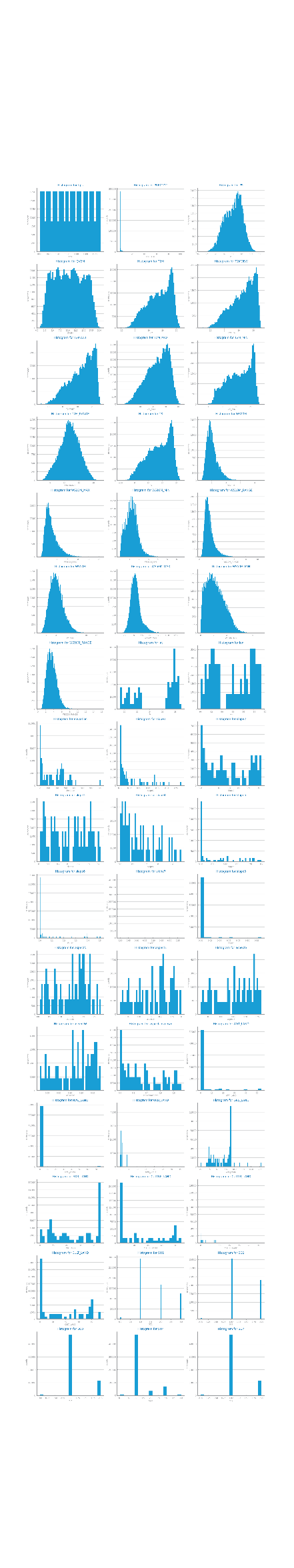


Figure : single histograms numeric dataset 2

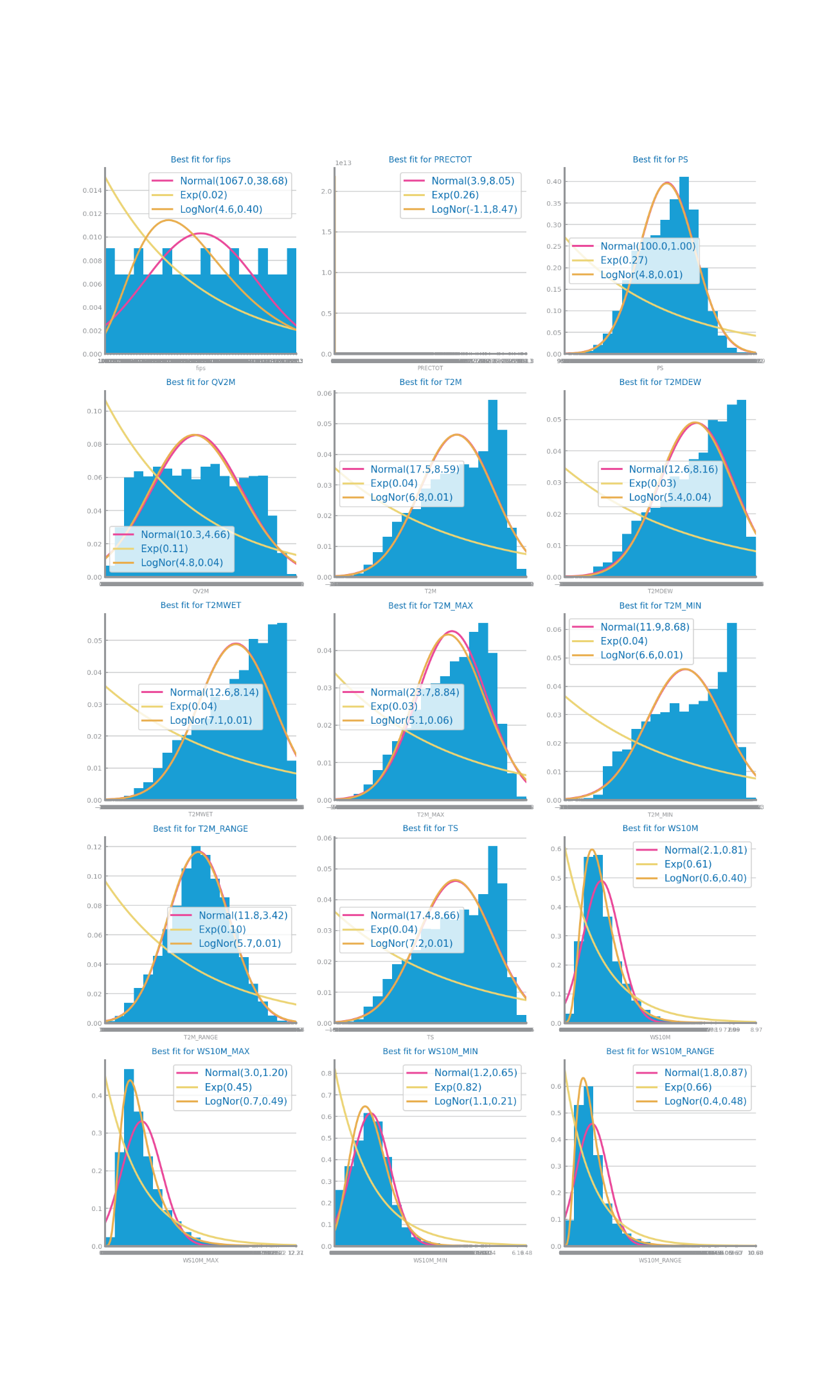


Figure : histograms with log - dataset 2

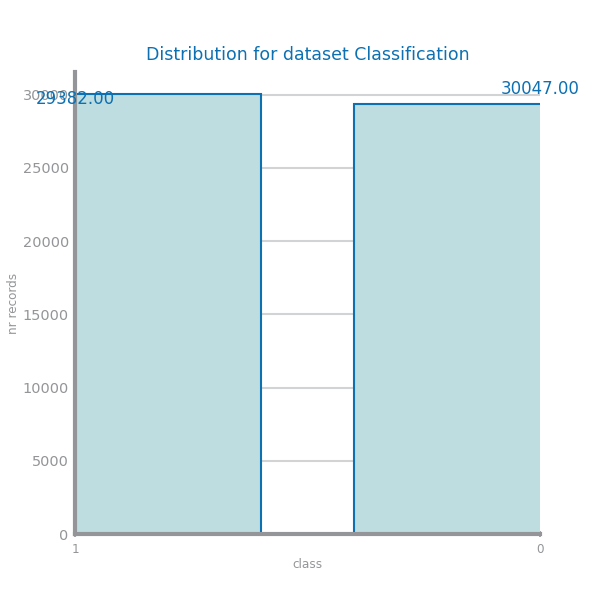


Figure : class distribution - dataset 2

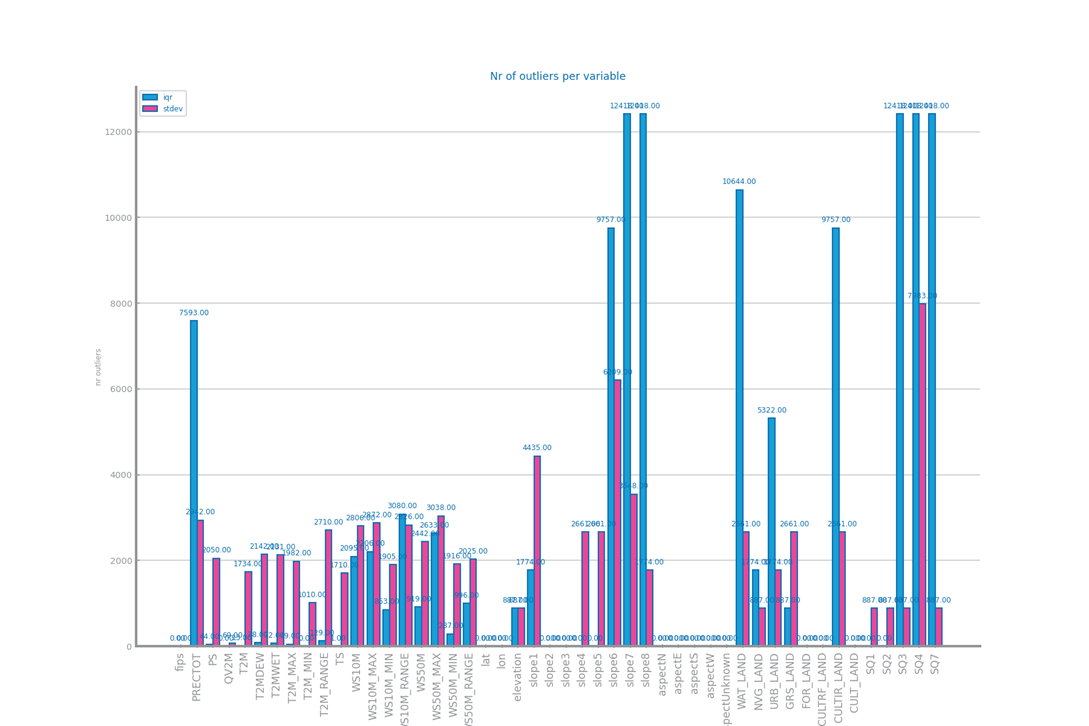


Figure : outliers for variables dataset 2

## Data Granularity

Most of the numeric variables of dataset 1 are best represented with 100 bins. This means that there is not a very fine level of granularity (except for the id types, but they are not considered of numeric value). So, there is probably no discretization needed.

The symbolic variables are best represented with 10 bins, which means that there are generally not many categories. This is however not true for the ‘’diag’’ variables, due to their many number of categories. Climbing up the concept hierarchy can be difficult for these variables, since they are distinct diagnoses.

Dataset 2 shows another image. Here, the variables differ more in granularity. The variables ‘QV2M’, ‘T2-’, ‘TS’ and ‘WS’ show high levels of granularity. Problems occurring can be a high storage space, and worse results. Storage space is not an issue for now, since the programming software handles the data fine. If the model building becomes very slow later on, aggregating these variables may be an option. It is however possible that results improve with a lower level of granularity. This will be tested during the model building.

Shall contain all relevant information and charts respecting to the data granularity perspective, such as the impact of different granularities considered for each variable. May present additional taxonomies if needed. **Shall not exceed 200 characters.**

Figure Granularity analysis for dataset 1

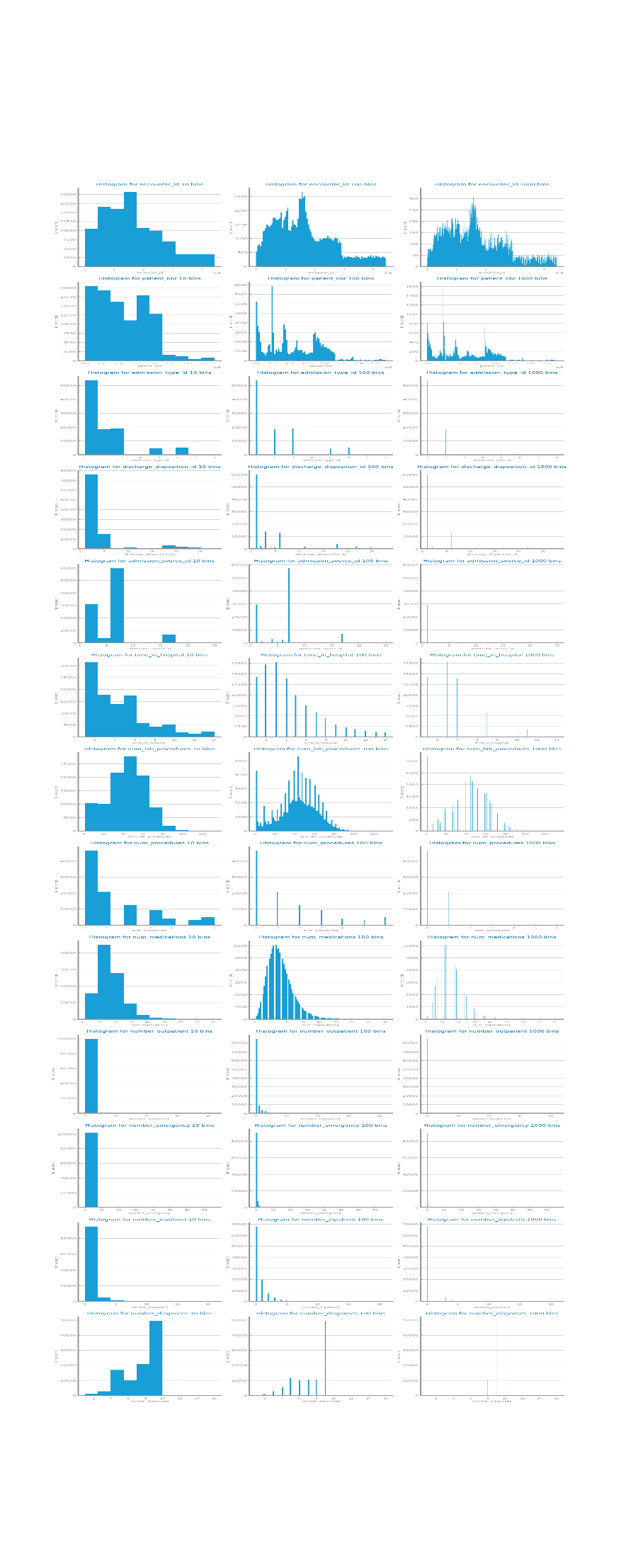


Figure : granularity study numeric variables dataset 1

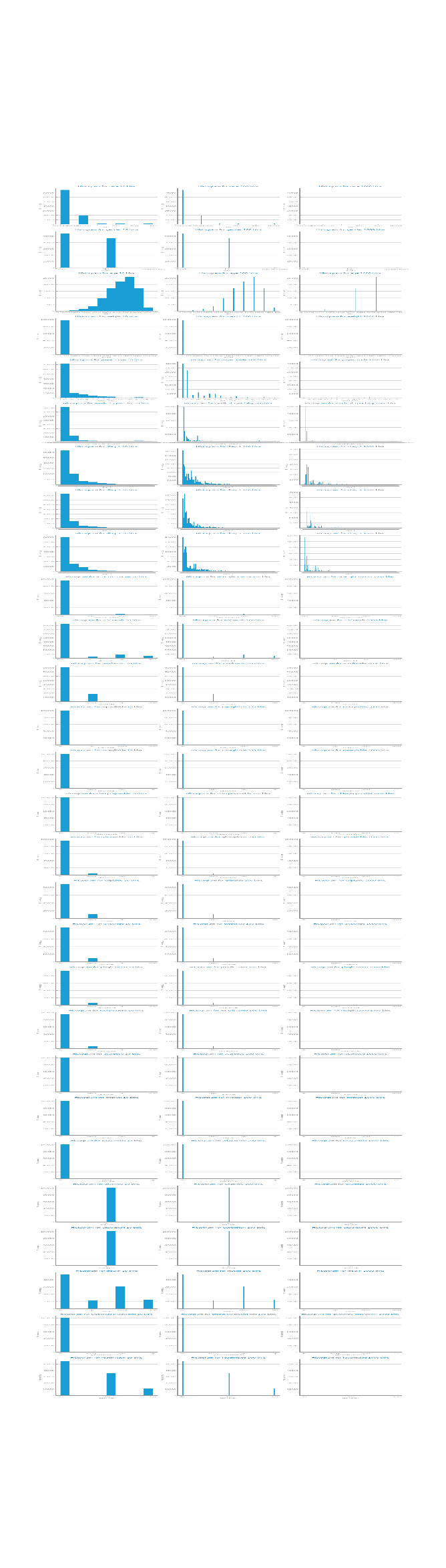


Figure : granularity study symbolic variables dataset 1

Figure Granularity analysis for dataset 2

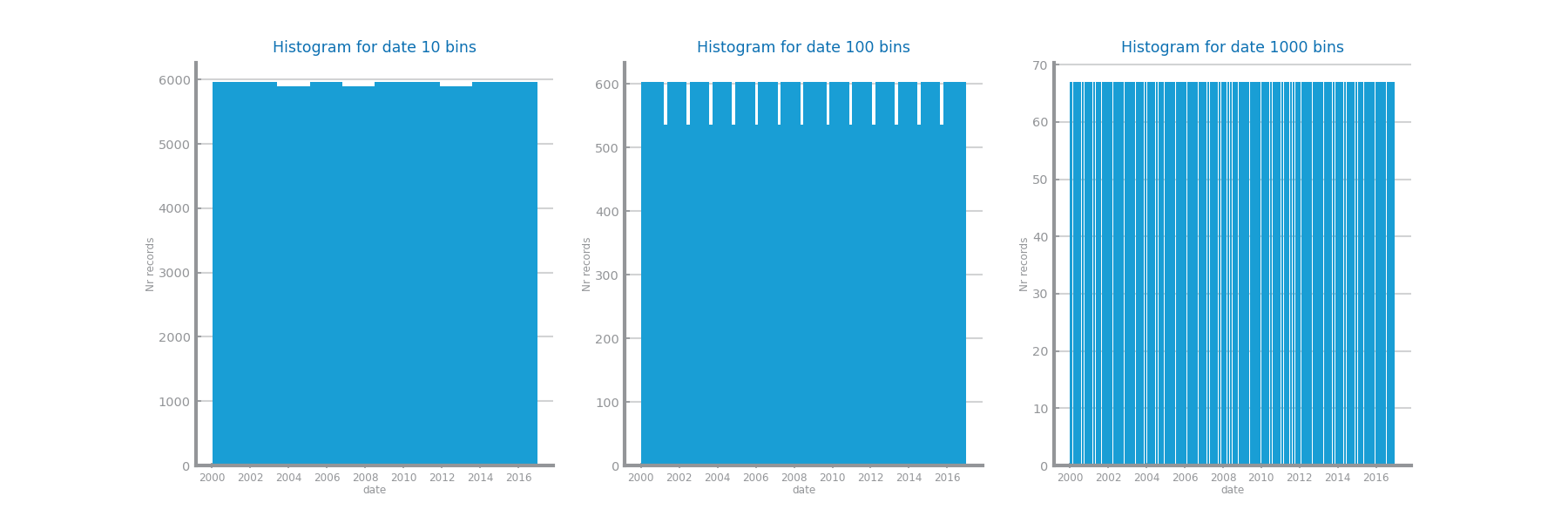


Figure : granularity study date - dataset 2

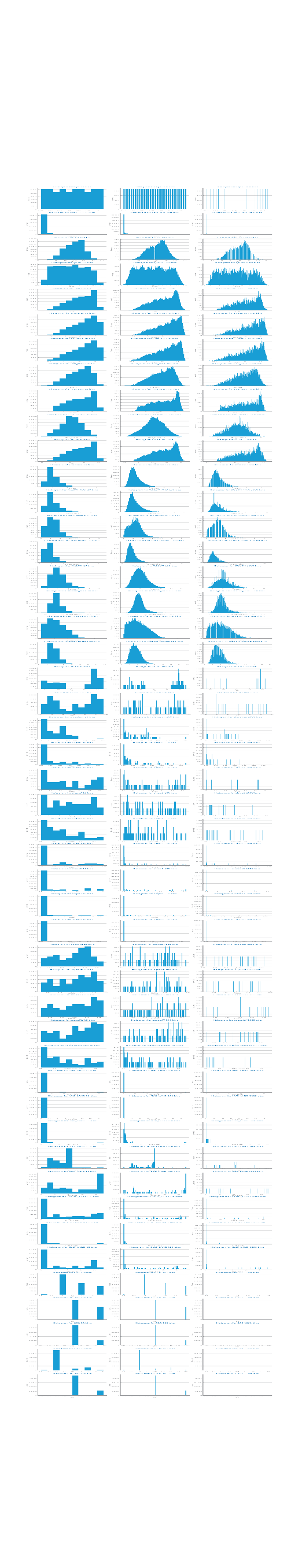


Figure : granularity analysis numeric variables dataset 2

## Data Sparsity

Shall contain all relevant information and charts respecting to the data sparsity perspective, such as domain coverage and correlation among variables. **Shall not exceed 300 characters.**

Figure Sparsity analysis for dataset 1



Figure : Sparsity study numeric variables dataset 1

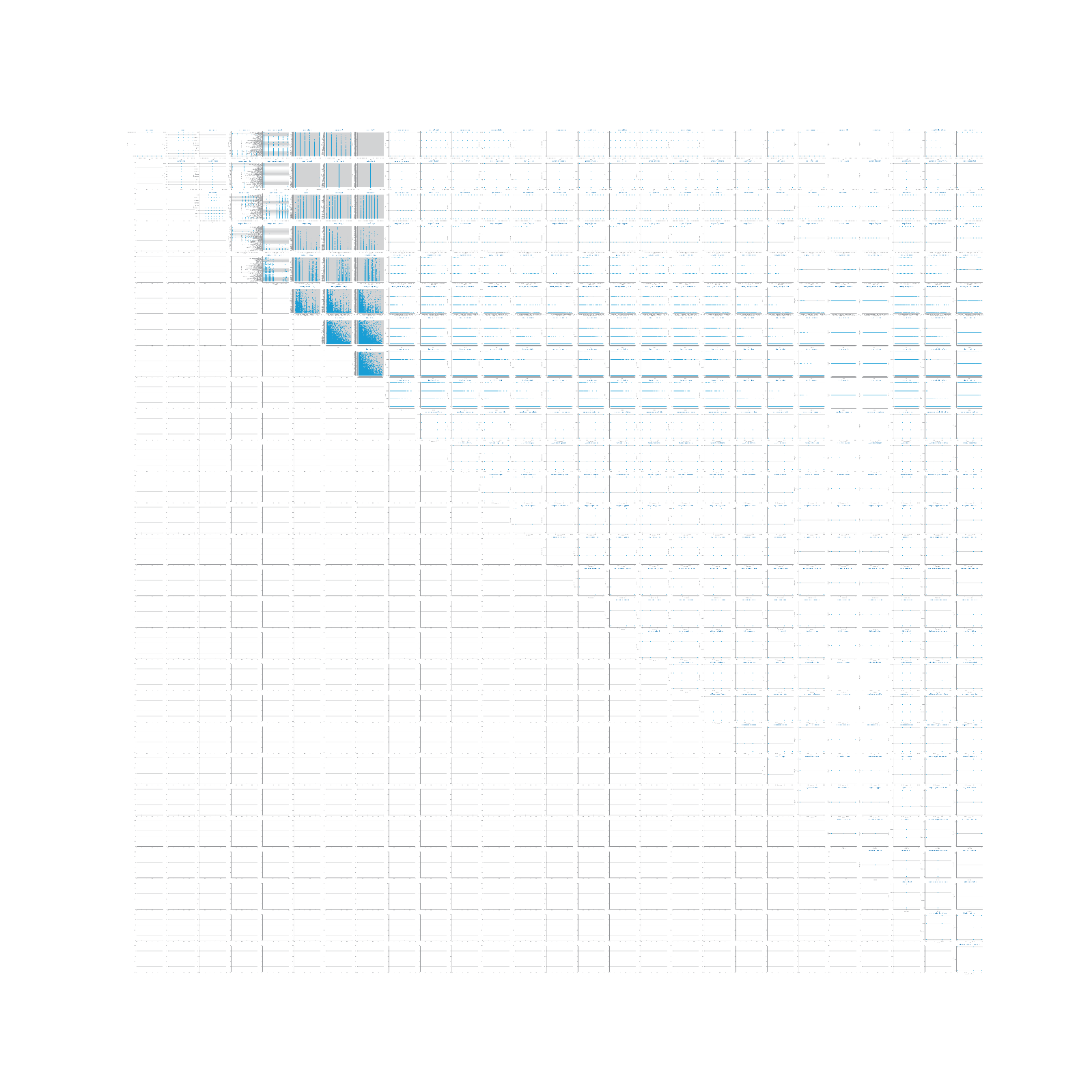


Figure : Sparsity study symbolic variables dataset 1

Figure Sparsity analysis for dataset 2

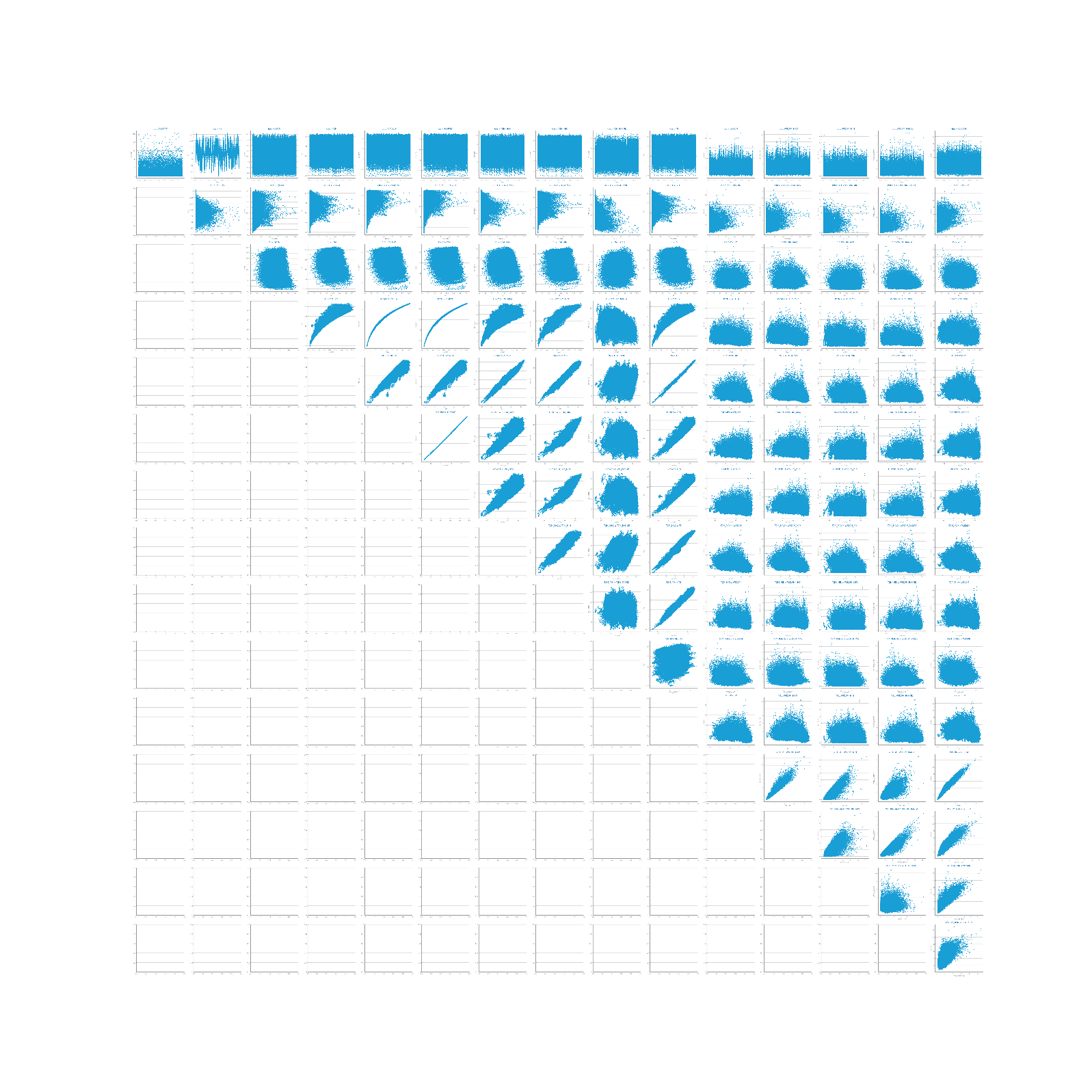


Figure : sparsity study dataset 2

Figure Correlation analysis for dataset 1

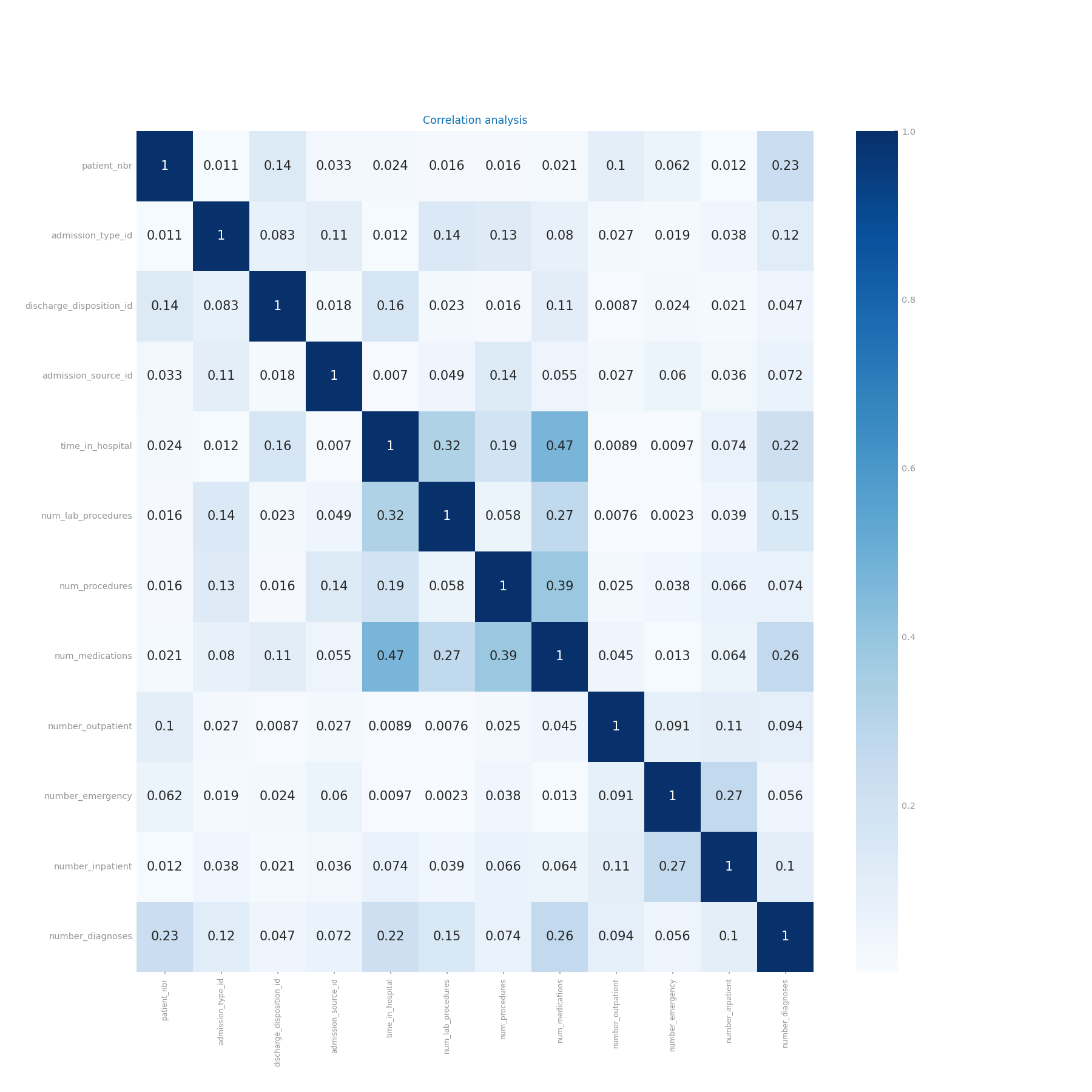


Figure : correlation analysis dataset 1

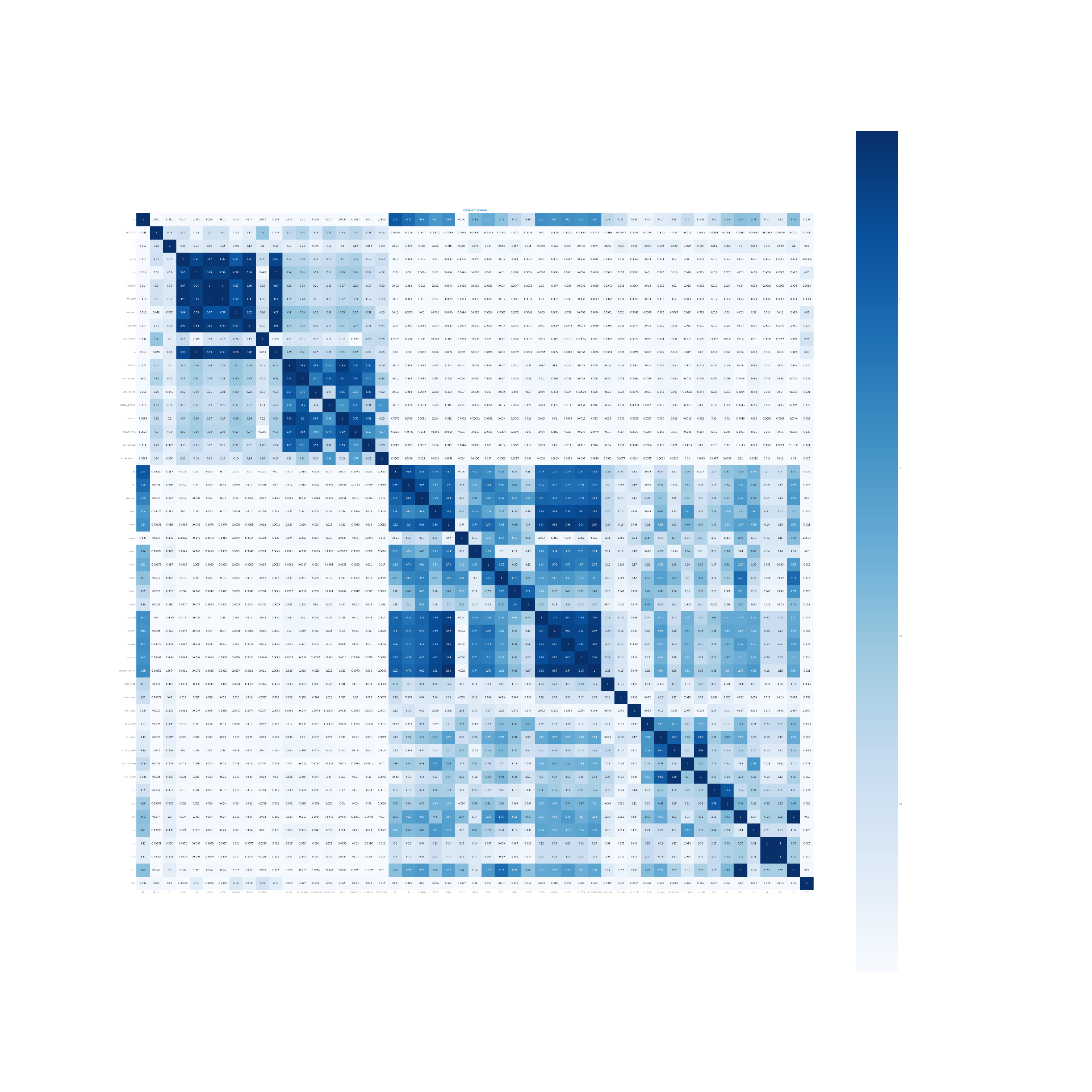


Figure Correlation analysis for dataset 2

# Data Preparation

## Variables Encoding

For any model to be trained by the data, it is important that the data is numeric. Also, the data should be encoded in a logical way, such that the highest value for a variable, receives the highest code number.

**Dataset 1**

Dataset 1 contains a large number of symbolic variables, which need to be encoded. Different methods have been used for different variables. The table below displays the overview:

|  |  |
| --- | --- |
| **Transformation method** | **Variables** |
| Dummification | None |
| Random code assignment | ['race', 'payer\_code', 'medical\_specialty'] |
| Binary code [0,1] | ['gender',] [diabetes\_med and change]?? |
| Encoding according to ordinal value | ['age', 'max\_glu\_serum', 'A1Cresult', 'metformin', 'repaglinide', 'nateglinide', 'chlorpropamide', 'glimepiride', 'glipizide', 'glyburide', 'pioglitazone', 'rosiglitazone', 'acarbose', 'miglitol', 'tolazamide', 'insulin', ‘readmitted’] |
| Encoding according to dataset description | ['diag\_1', 'diag\_2', 'diag\_3'] |

**Dataset 2**

For dataset 2, this section is somewhat easier. All variables are numeric except for one, which is date. Each date value is transformed as follows: date = 10.000\*year + 100\*month + day. This way, the variable is encoded to numeric and a larger code means a later date.

Shall contain all relevant information respecting to the transformation of variables, including *dummification*. The list of variables under each one of the transformations, shall be presented. If not applied explain the reason for that, based on data characteristics. **Shall not exceed 500 characters.**

## Missing Value Imputation

Missing values can be imputated with different strategies and combinations of them. For dataset 1, two strategies were tested. Both strategies start with dropping the variable ‘’weight’’. This variable contains 97% missing values. The 3% values can in no way be used to fill the missings. The decision was made to leave the variables payer\_code and medical\_specialty in the data, because 50% missing values may still be fillable.

The fill strategies ‘’mean’’ and ‘’most\_frequent’’ were tested against each other. The decision was made to not test constant fill, because this is not adding any valuable information to the data, and may even distort the data too much.

Which one was best?

**Dataset 2**

As there are no missing values for dataset 2, no imputation is needed.

Shall contain all relevant information and charts respecting to missing values imputation, such as the choices made and the impact of the different approaches on modelling results. Shall also clearly reveal the approach selected to proceed with the processing. If not applied explain the reason for that, based on data characteristics. **Shall not exceed 200 characters.**

Figure Missing values imputation results with different approaches for dataset 1

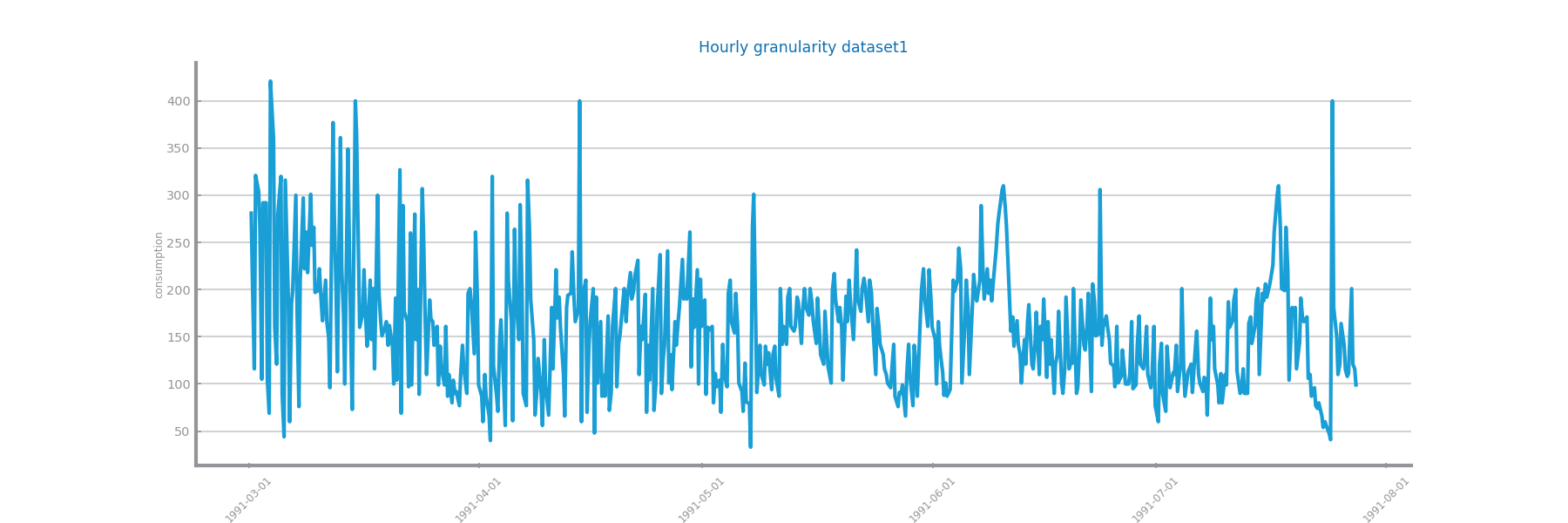


Figure Missing values imputation results with different approaches for dataset 2

For dataset 2, there are no missing values. Therefore, we don’t present any graphs for missing value imputation dataset 2.

## Outliers Treatment

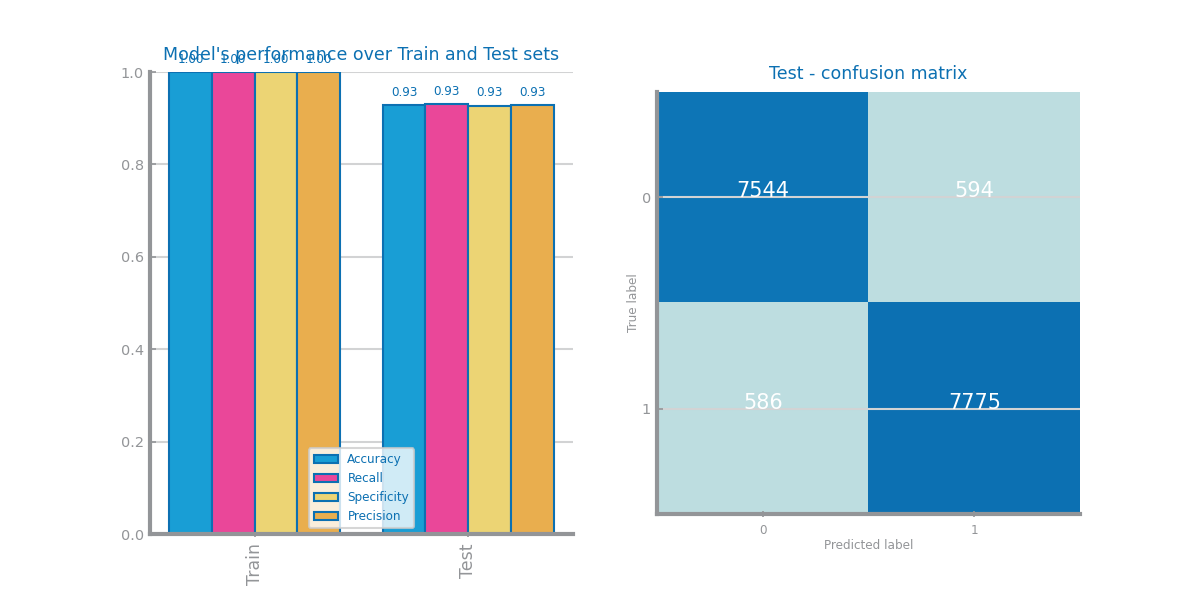
As could be seen in the section ‘’data distribution’’, both datasets contain a lot of outliers. Two treatment options are considered: using the interquartile range (iqr) and the standard deviation. The outlier parameter has been set to 1.5. For the first treatment, all values above the 75% quantile + iqr are considered as outlier. The same goes for all values below 25% quantile – iqr. For the standard deviation approach, they are all values mean +- 1,5\* stdev.

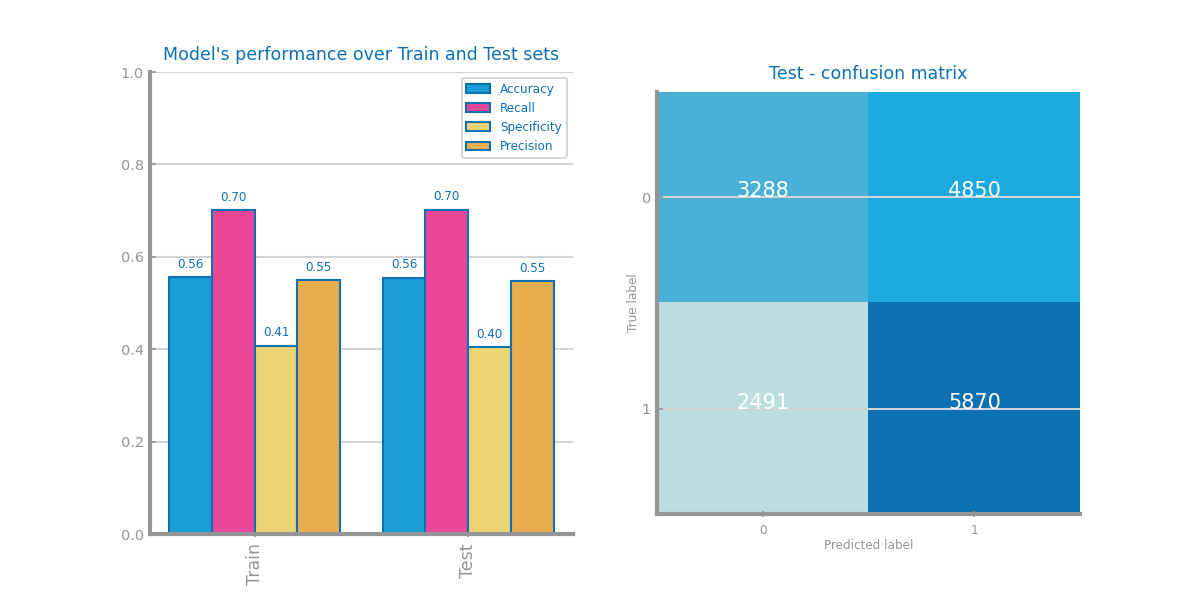
Best approach?

Shall contain all relevant information and charts respecting to outliers imputation, such as the choices made and the impact of the different approaches on modelling results. Shall also clearly reveal the approach selected to proceed with the processing. If not applied explain the reason for that, based on data characteristics. **Shall not exceed 200 characters**.

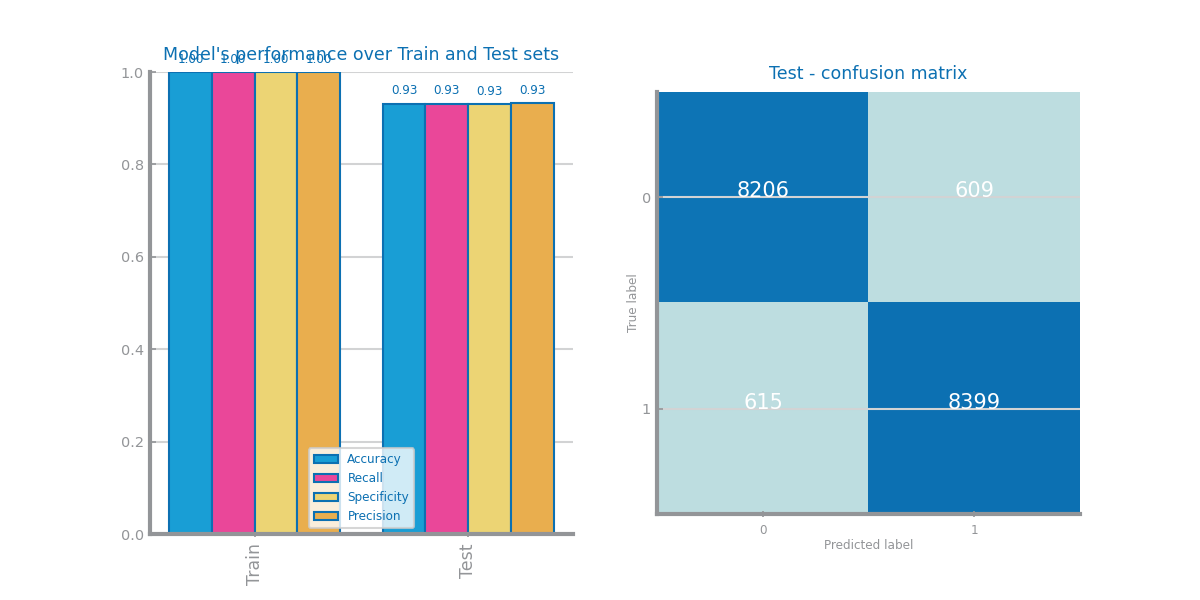
Figure Outliers imputation results with different approaches for dataset 1

Results for drop outliers dataset 2: first knn, then nb





Results for truncate outliers dataset 2: first knn, then nb



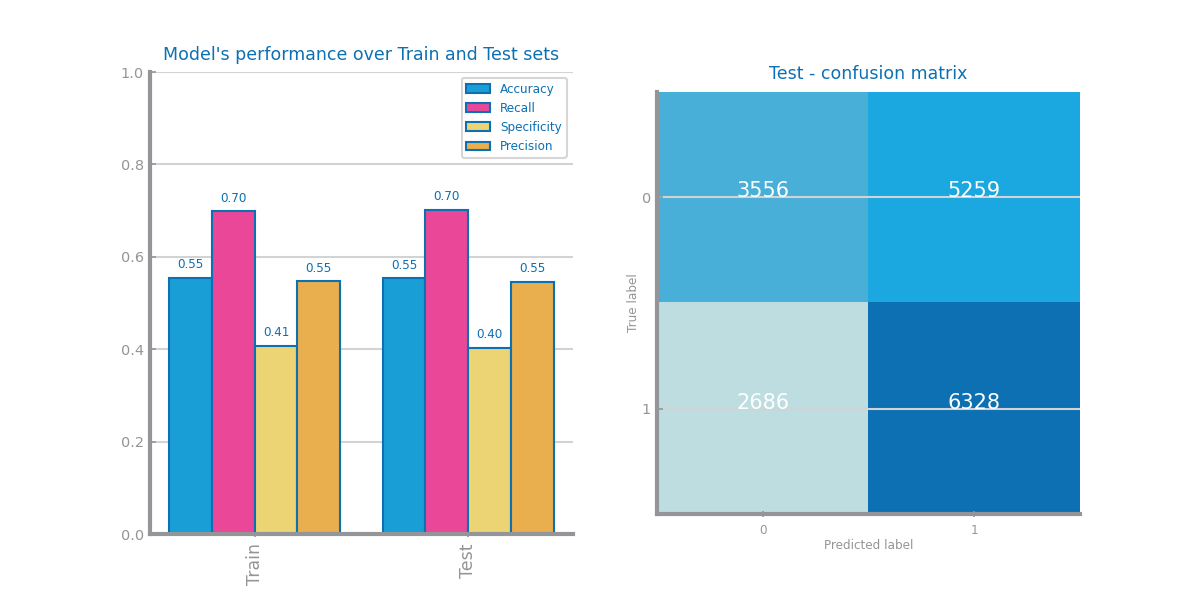


Figure Outliers imputation results with different approaches for dataset 2

## Scaling

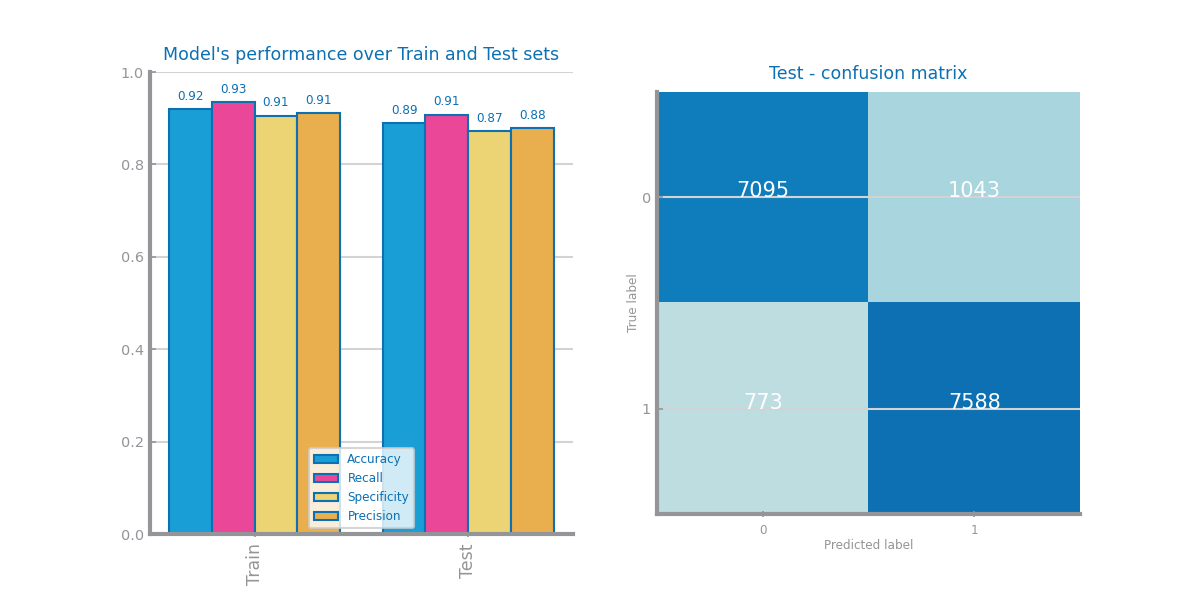
Shall contain all relevant information and charts respecting to scaling transformation, such as the choices made and the impact of the different approaches on modelling results. Shall also clearly reveal the approach selected to proceed with the processing. If not applied explain the reason for that, based on data characteristics. **Shall not exceed 200 characters.**

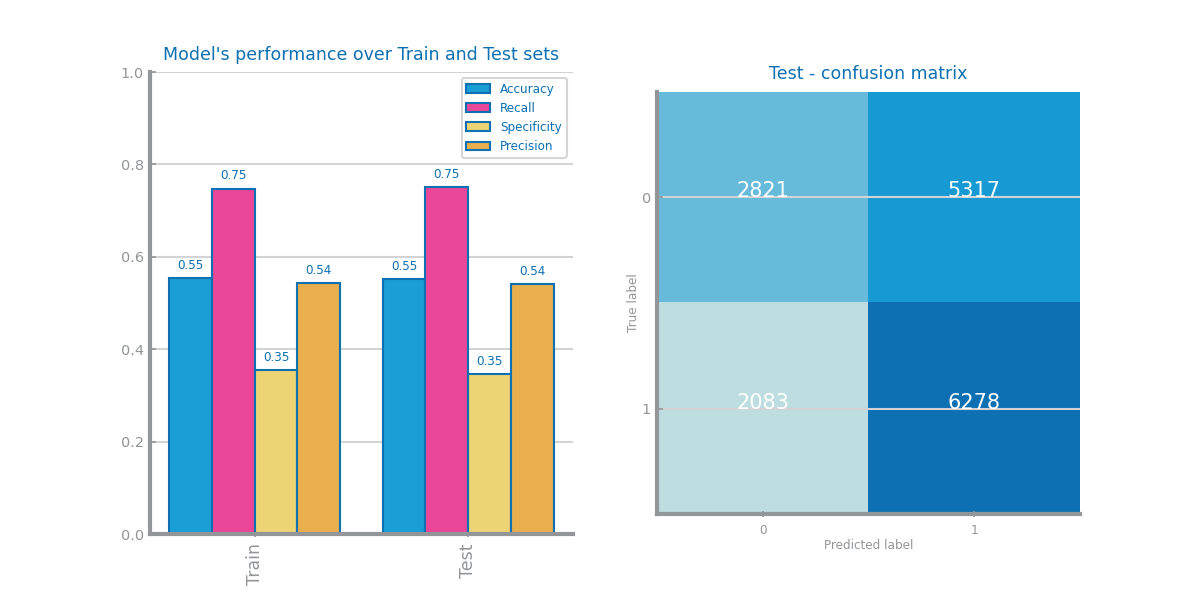
Figure Scaling results with different approaches for dataset 1

Figure Scaling results with different approaches for dataset 2

Results for minmax scaling dataset 2: first knn, then nb

Results for zscore scaling dataset 2, first knn, then nb





## Balancing

Shall contain all relevant information and charts respecting to balancing transformation, such as the choices made and the impact of the different approaches on modelling results. Shall also clearly reveal the approach selected to proceed with the processing. If not applied explain the reason for that, based on data characteristics. **Shall not exceed 200 characters**.

Figure Balancing results with different approaches for dataset 1

Figure Balancing results with different approaches for dataset 2

## Feature Selection

Shall contain all relevant information and charts respecting to feature selection based on filtering out **redundant** variables. The different choices and their impact on the modelling results shall be presented and explained. Should also clearly reveal the approach selected to proceed with the processing. All explanations shall be based on data characteristics. **Shall not exceed 200 characters.**

Figure Feature selection of redundant variables results with different parameters for dataset 1

Figure Feature selection of redundant variables results with different parameters for dataset 2

## Feature Extraction (optional)

Shall contain all relevant information and charts respecting to feature extraction, in particular PCA. The different choices and their impact on the modelling results shall be presented and explained. **Shall not exceed 200 characters.**

Figure Principal components analysis and feature extraction results for dataset 1

Figure Principal components analysis and feature extraction results for dataset 2

## Feature Generation (optional)

Shall contain all relevant information and charts respecting to feature generation. The different choices and their impact on the modelling results shall be presented and explained. Shall summarize all variables generated and the formula used to derive them (in a table). **Shall not exceed 300 characters.**

Figure Feature generation results for dataset 1

Figure Feature generation results for dataset 2

# Models’ Evaluation

Shall be used to point out any important decision taken during the training, including training strategy and evaluation measures used. **Shall not exceed 300 characters**

## Naïve Bayes

Shall be used to present the results achieved with each one of Naïve Bayes implementations, comparing and proposing explanations for them. If any of the implementations is not used, a justification for it shall be presented.

Shall be used to present the evaluation of the best model achieved.

**Shall not exceed 300 characters.**

Figure Naïve Bayes alternatives comparison for dataset 1

Figure Naïve Bayes alternative comparison for dataset 2

Figure Naïve Bayes best model results for dataset 1 (left) and dataset 2 (right)

## KNN

Shall be used to present the results achieved through different similarity measures and KNN parameterizations. The results shall be compared and explanations for them shall be presented. The justification for the chosen similarity measures shall be presented. Shall be used to address the *overfitting* phenomenon, studying the conditions under which models face it. Shall be used to present the evaluation of the best model achieved. **Shall not exceed 400 characters**

Figure KNN different parameterizations comparison for dataset 1

Figure KNN different parameterizations comparison for dataset 2

Figure KNN overfitting analysis for dataset 1 (left) and dataset 2 (right)

Figure KNN best model results for dataset 1 (left) and dataset 2 (right)

## Decision Trees

Shall be used to present the results achieved through different parameterizations for the train of decision trees. The results shall be compared and explanations for them shall be presented. Shall be used to address the *overfitting* phenomenon, studying the conditions under which models face it. Shall be used to present the evaluation of the best model achieved. Shall be used to present the best tree achieved and its succinct description. **Shall not exceed 500 characters**

Figure Decision Trees different parameterizations comparison for dataset 1

Figure Decision Trees different parameterizations comparison for dataset 2

Figure Decision Trees overfitting analysis for dataset 1 (left) and dataset 2 (right)

Figure Decision trees best model results for dataset 1 (left) and dataset 2 (right)

Figure Best tree for dataset 1

Figure Best trees for dataset 2

## Random Forests

Shall be used to present the results achieved through different parameterizations for the train of random forests. The results shall be compared and explanations for them shall be presented. Shall be used to address the *overfitting* phenomenon, studying the conditions under which models face it. Shall be used to present the evaluation of the best model achieved. May be used to present the most important variables in the model. **Shall not exceed 500 characters**

Figure Random Forests different parameterizations comparison for dataset 1

Figure Random Forests different parameterizations comparison for dataset 2

Figure Random Forests overfitting analysis for dataset 1 (left) and dataset 2 (right)

Figure Random Forests best model results for dataset 1 (left) and dataset 2 (right)

Figure Random Forests variables importance for dataset 1 (left) and dataset 2 (right)

## Gradient Boosting

Shall be used to present the results achieved through different parameterizations for the train of gradient boosting. The results shall be compared and explanations for them shall be presented. Shall be used to address the *overfitting* phenomenon, studying the conditions under which models face it. Shall be used to present the evaluation of the best model achieved. May be used to present the most important variables in the model. **Shall not exceed 500 characters**

Figure Gradient boosting different parameterizations comparison for dataset 1

Figure Gradient boosting different parameterizations comparison for dataset 2

Figure Gradient boosting overfitting analysis for dataset 1 (left) and dataset 2 (right)

Figure Gradient boosting best model results for dataset 1 (left) and dataset 2 (right)

Figure Gradient boosting variables importance for dataset 1 (left) and dataset 2 (right)

## Multi-Layer Perceptrons

Shall be used to present the results achieved through different parameterizations for the train of MLPs. The results shall be compared and explanations for them shall be presented. Shall be used to address the *overfitting* phenomenon, studying the conditions under which models face it. In particular by analysing the loss\_curve\_ available at the end of each train. Shall be used to present the evaluation of the best model achieved. **Shall not exceed 500 characters**

Figure MLP different parameterizations comparison for dataset 1

Figure MLP different parameterizations comparison for dataset 2

Figure MLP overfitting analysis for dataset 1 (left) and dataset 2 (right)

Figure Loss curves analysis for dataset 1 (left) and dataset 2 (right)

Figure MLP best model results for dataset 1 (left) and dataset 2 (right)

# Critical Analysis

Shall be used to present a summary of the results achieved with the different modeling techniques, and the impact of the different preparation tasks on their performance.

A cross-analysis of the different models may also be presented, identifying the most relevant variables common to all of them (when possible) and the relation among the patterns identified within the different classifiers.

A critical assessment of the best models shall be presented, clearly stating if the models seem to be good enough for the problem at hand.

**Additional charts may be presented here. Shall not exceed 2000 characters.**

Time Series Forecasting

# Data Profiling

## Data Granularity

May be used to identify the most atomic granularity and two other different granularities to consider. **Shall not exceed 300 characters.**

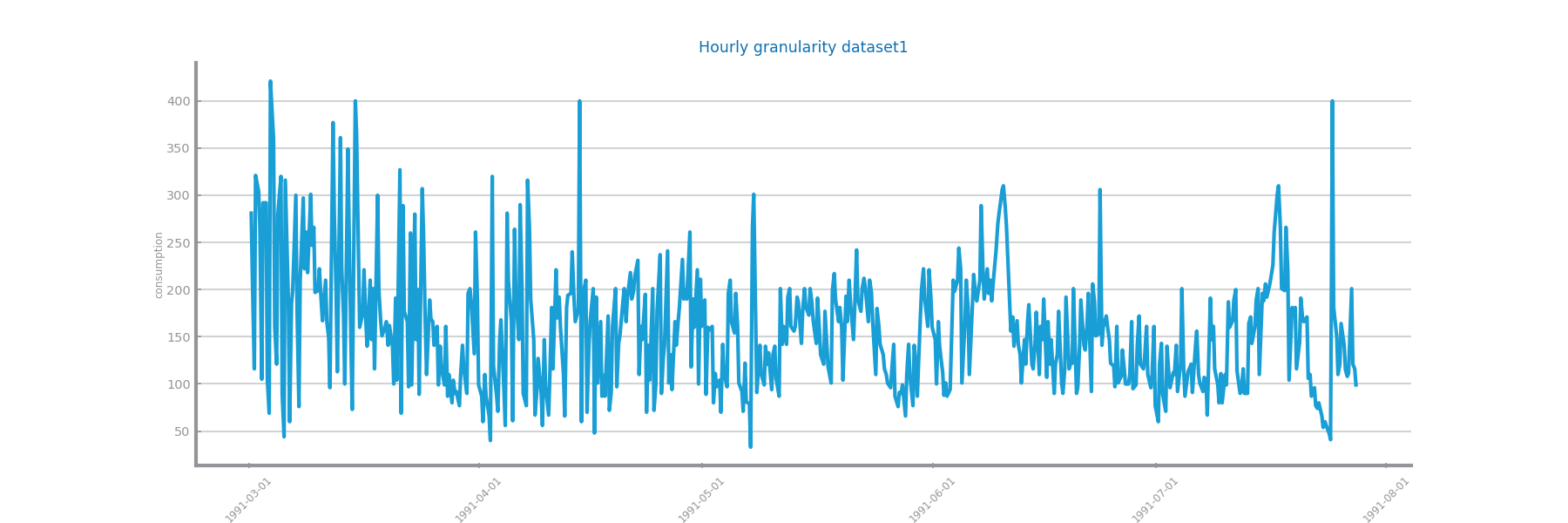


Figure Time series 1 at the most granular detail

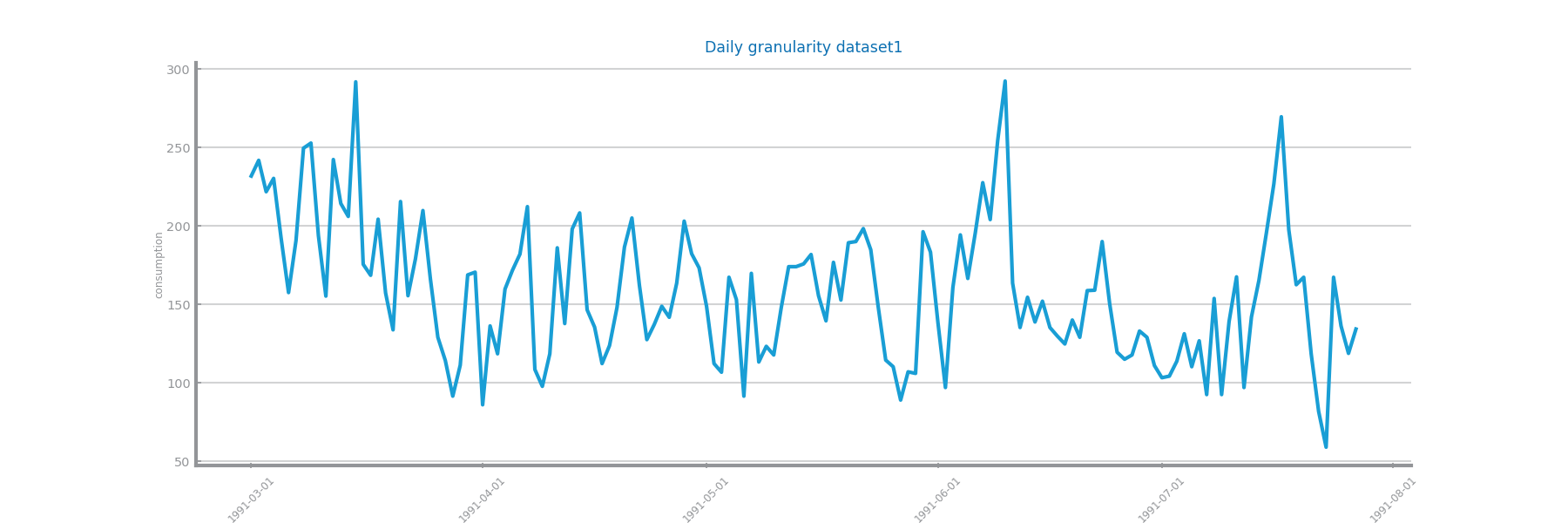


Figure Time series 1 at the second chosen granularity

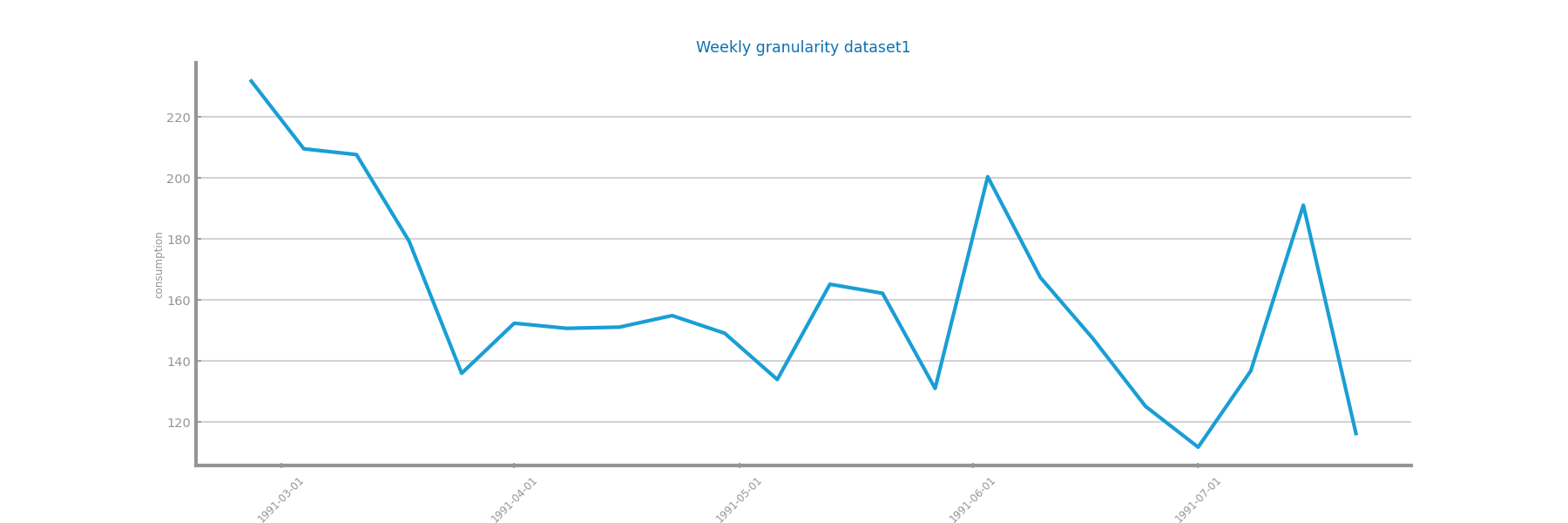


Figure Time series 1 at the third chosen granularity

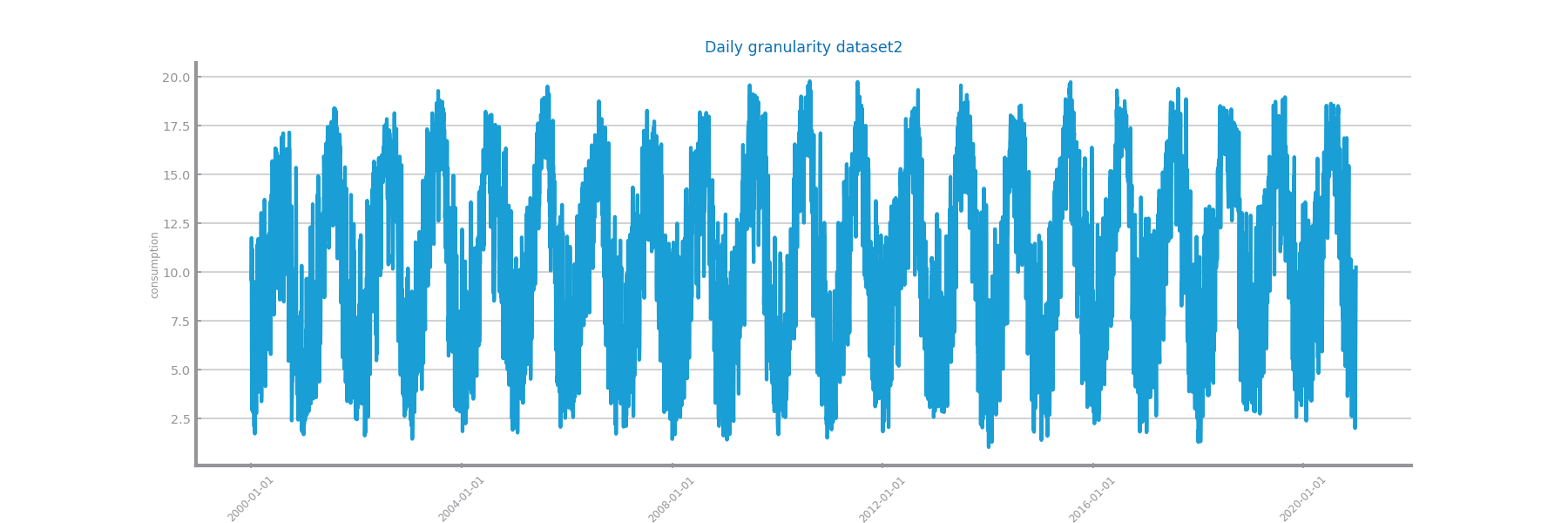


Figure Time series 2 at the most granular detail

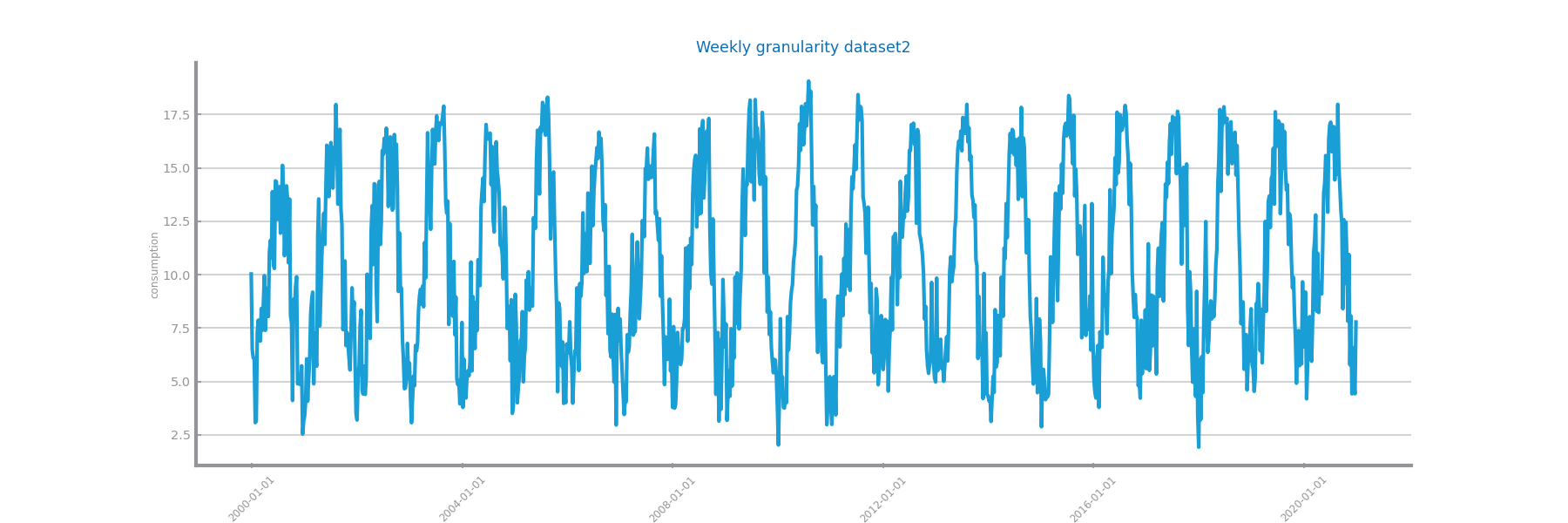


Figure Time series 2 at the second chosen granularity

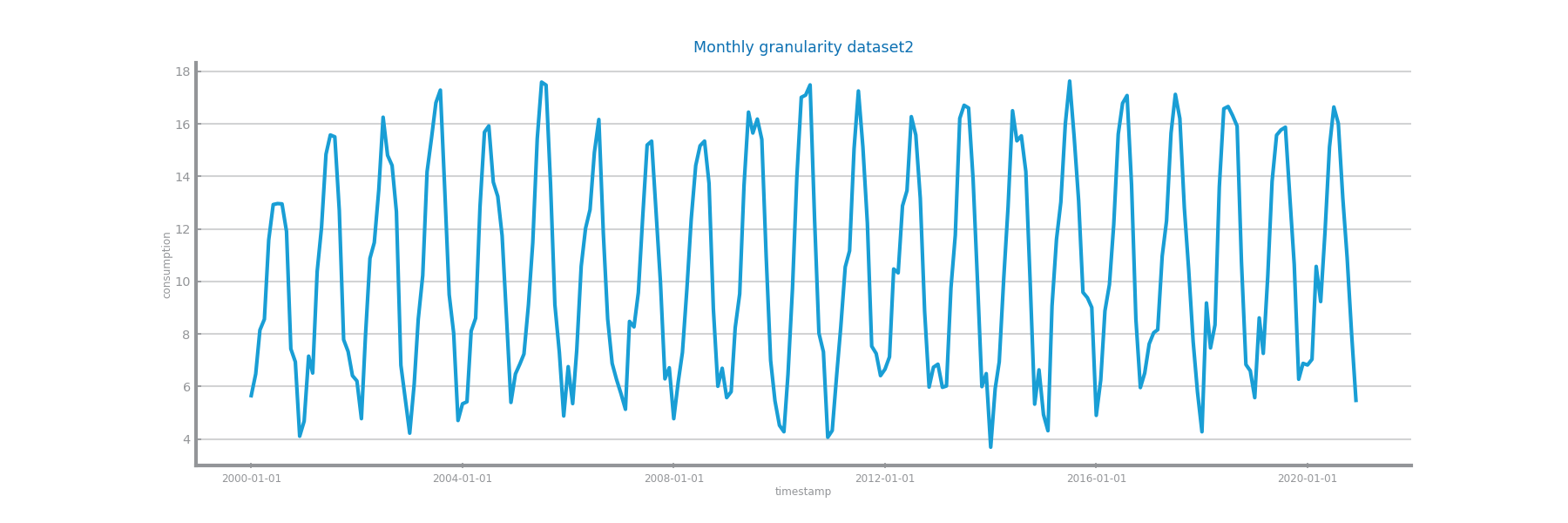


Figure Time series 2 at the third chosen granularity

## Data Distribution and Stationarity

Shall be used to perform the data analysis at those three different granularities, namely the series distribution and stationarity. **Shall not exceed 300 characters.**

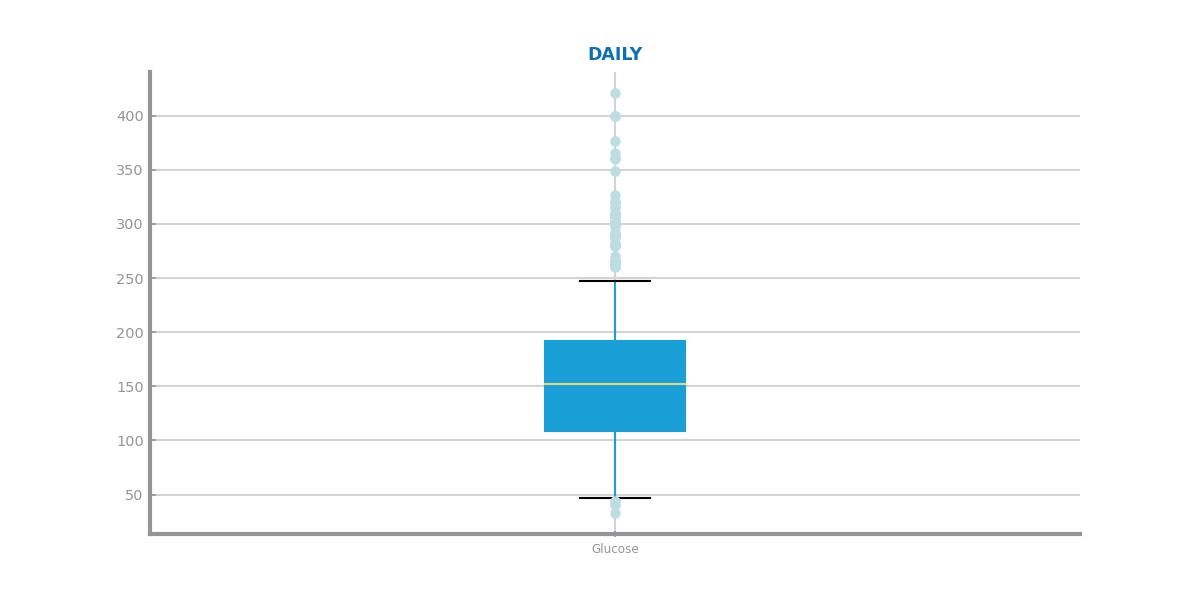


Figure Boxplot(s) for time series 1

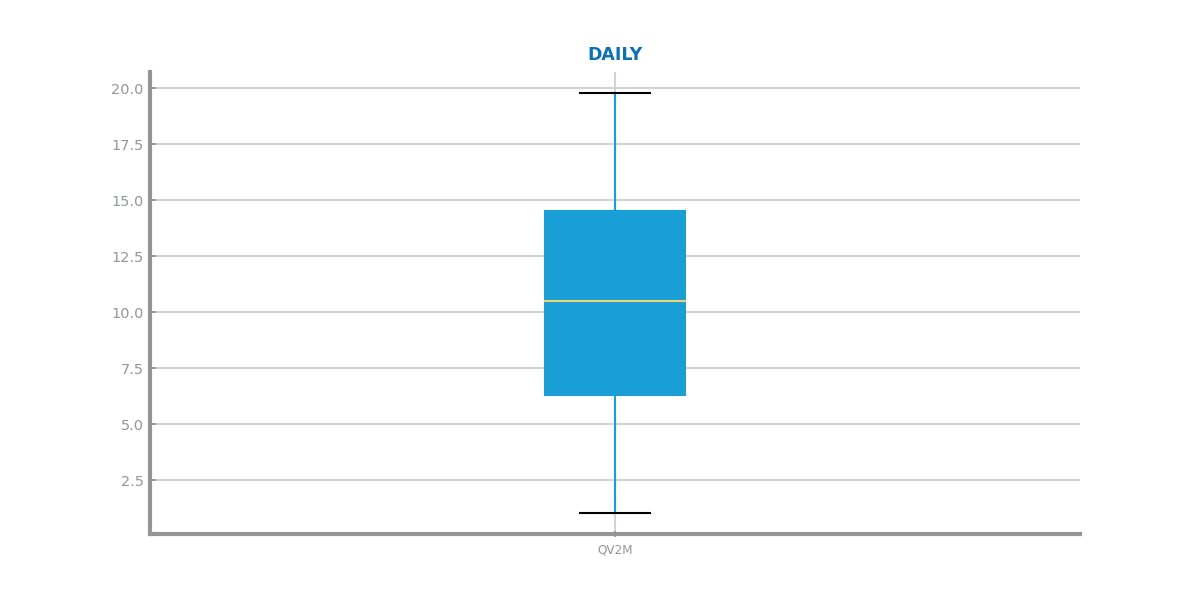
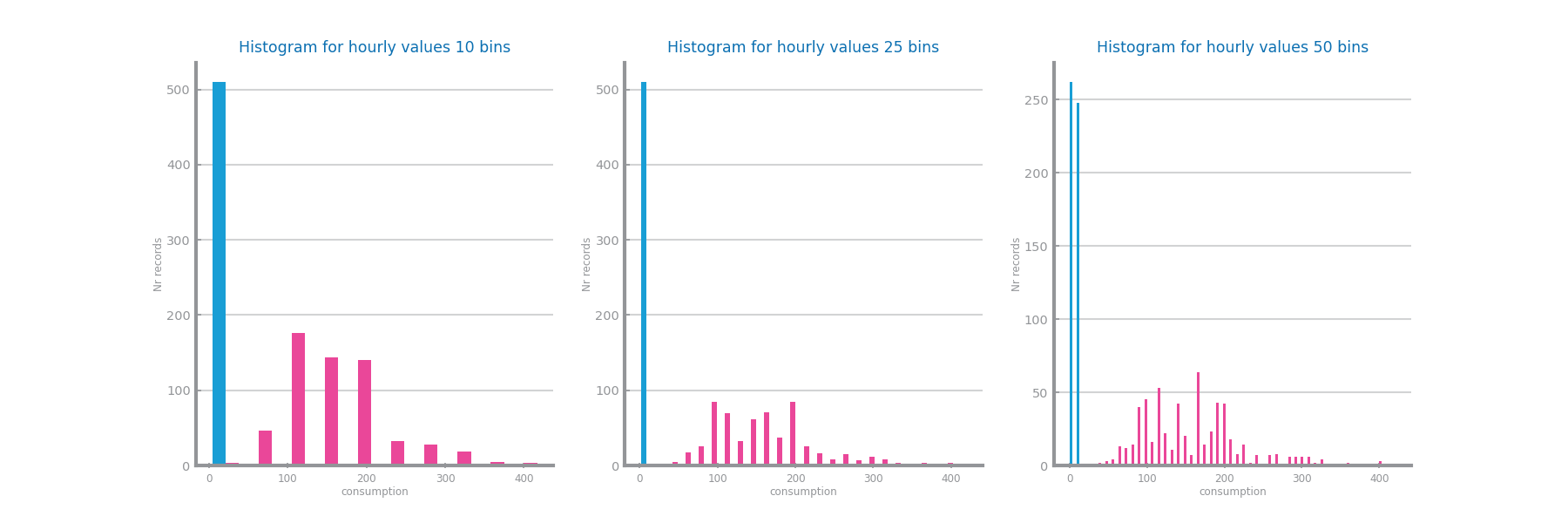


Figure Boxplot(s) for time series 2



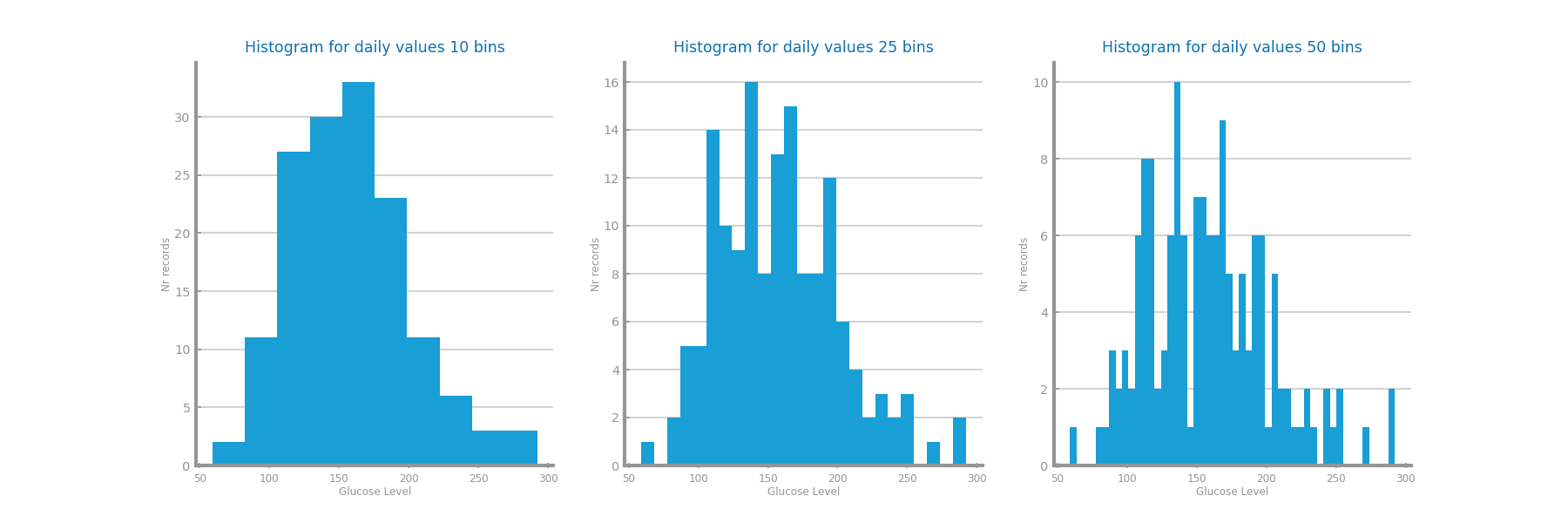
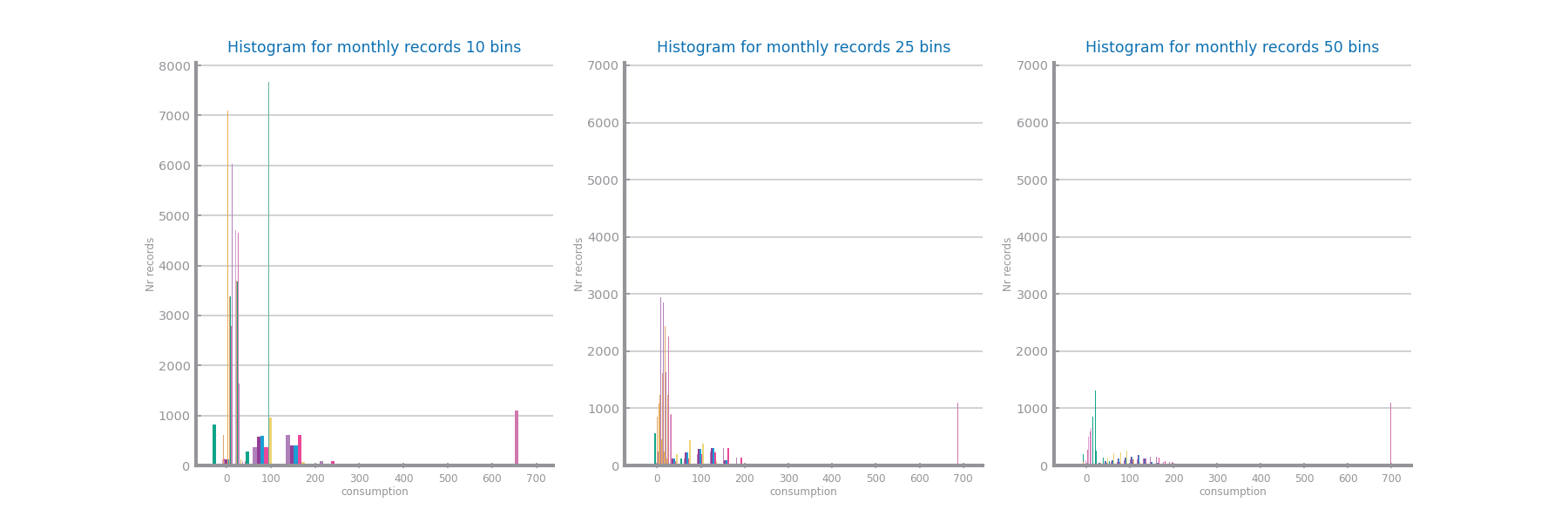




Figure Histogram(s) for time series 1



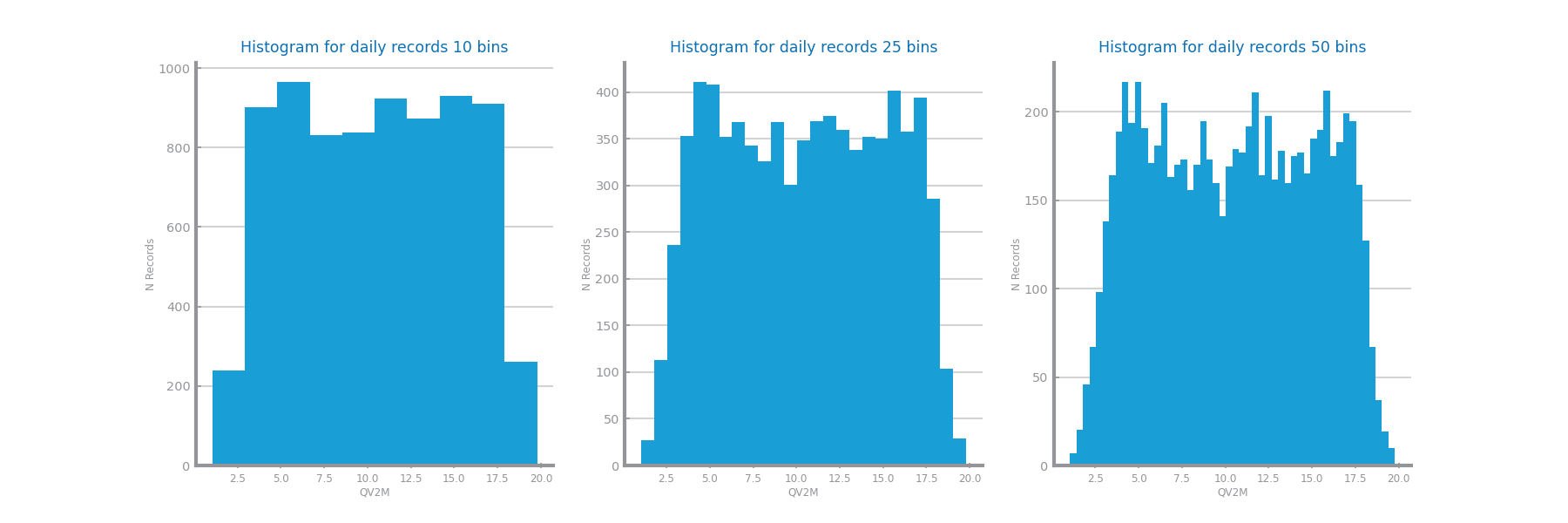
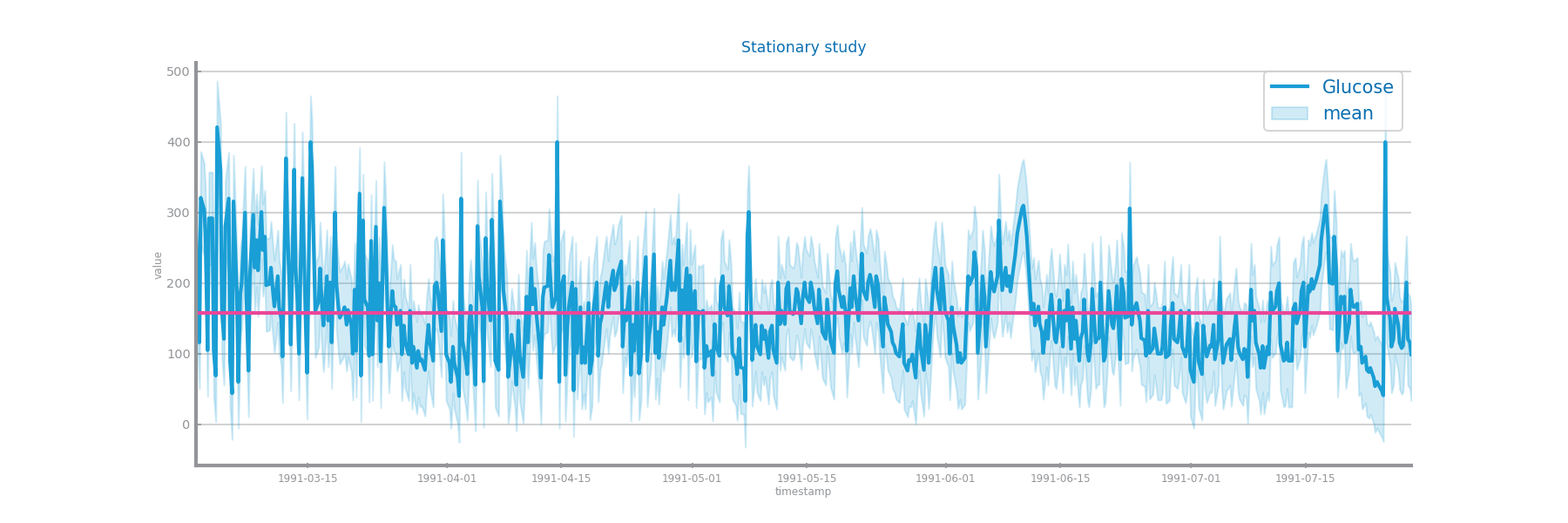


Figure Histogram(s) for time series 2



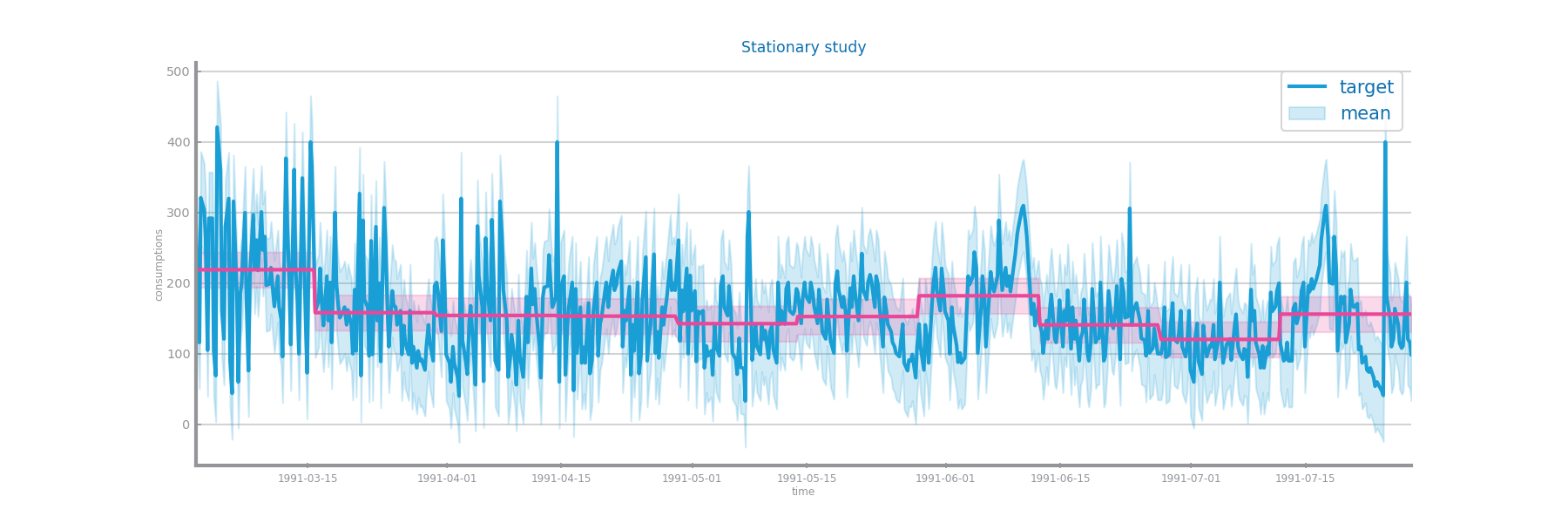
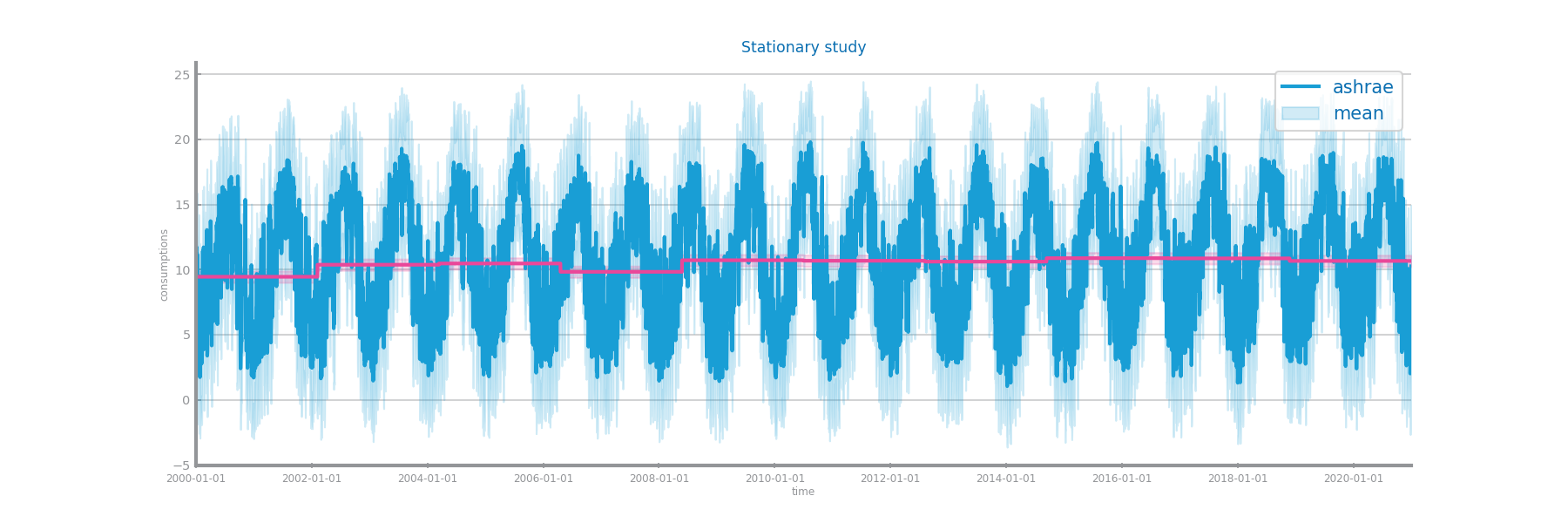


Figure Stationarity study for time series 1



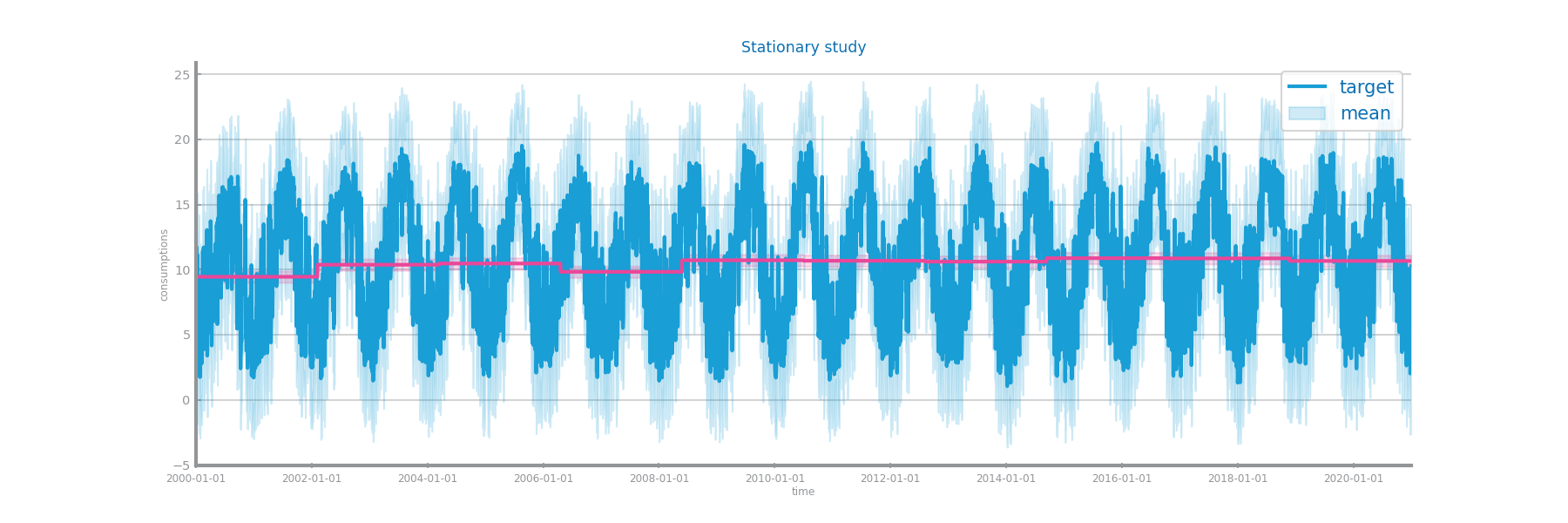


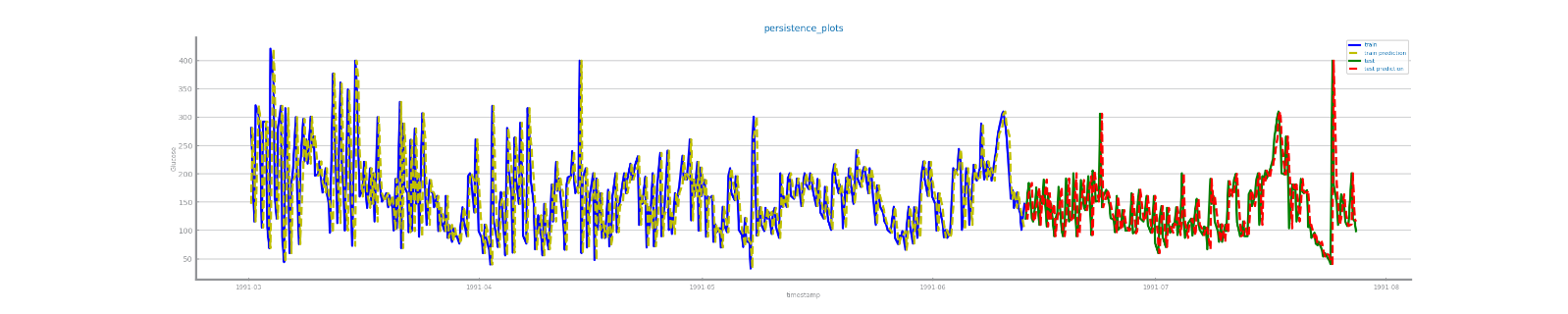
Figure Stationarity study for time series 2

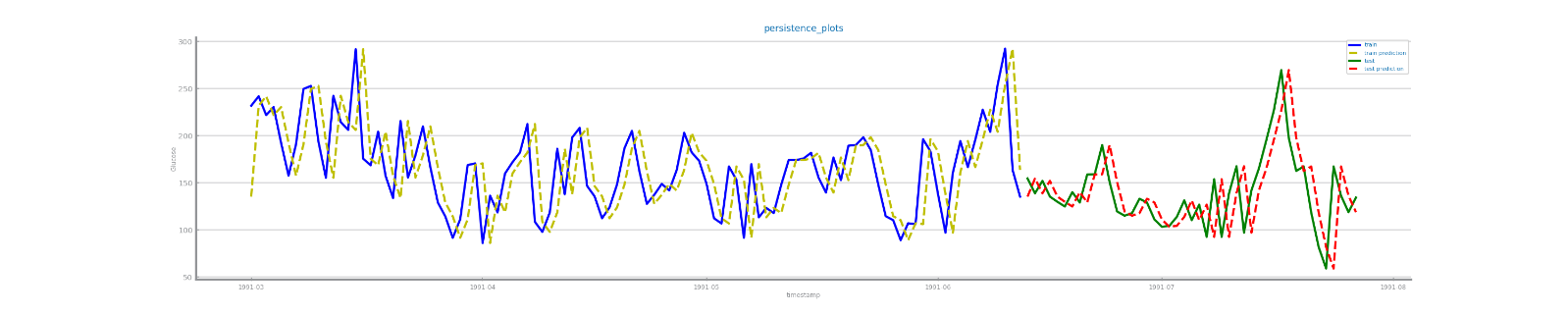
# Data Transformation

## Aggregation

Shall describe the results of applying the persistence model over the three different aggregations over both datasets, and identifying the granularity chosen to proceed. **Shall not exceed 200 characters.**

Persistence plots hourly, daily, weekly:





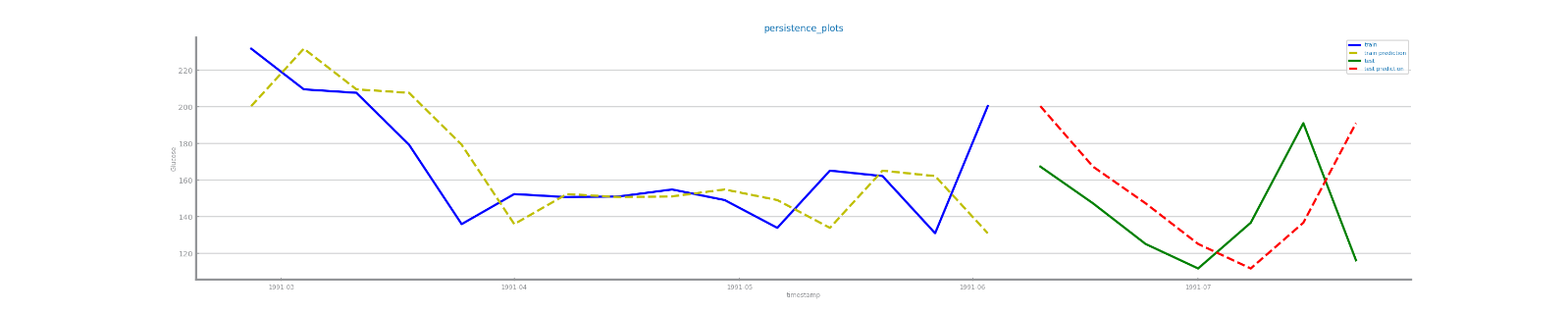
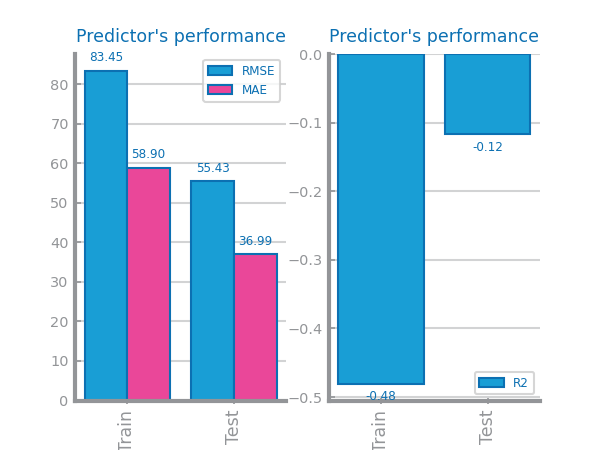
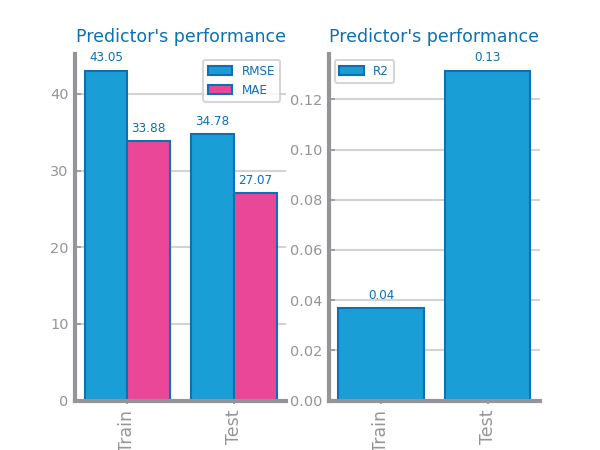


Figure Forecasting plots after different aggregations on time series 1

Persistance evaluations hourly, daily, weekly:





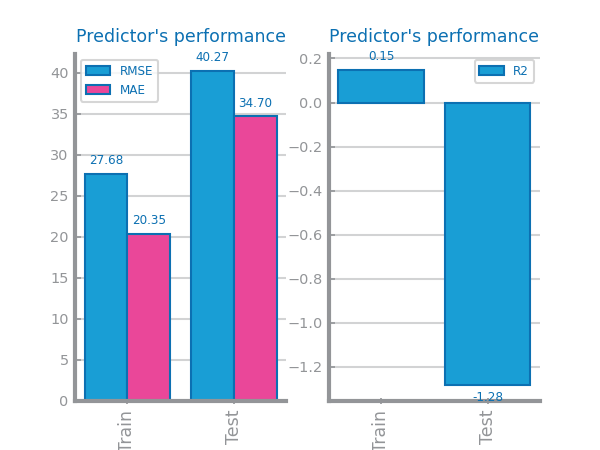


Figure Forecasting results after different aggregations on time series 1

Persistence plots daily, weekly, monthly:

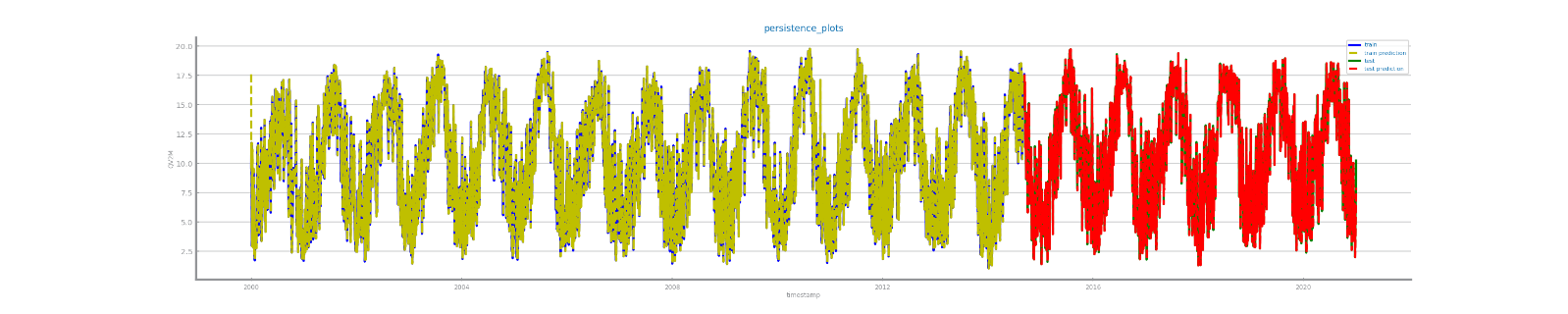


Figure Forecasting plots after different aggregations on time series 2

Figure Forecasting results after different aggregations on time series 2

## Smoothing

Shall describe the results of applying the persistence model over different smoothing transformations over both datasets, and identifying the best result to proceed. **Shall not exceed 200 characters.**

Figure Forecasting plots after different smoothing parameterizations on time series 1

Figure Forecasting results after different smoothing parameterizations on time series 1

Figure Forecasting plots after different smoothing parameterizations on time series 2

Figure Forecasting results after different smoothing parameterizations on time series 2

## Differentiation

Shall describe the results of applying the persistence model over two consecutive differentiation of both datasets, and identifying the best result to proceed. **Shall not exceed 200 characters.**

Figure Forecasting plots after first and second differentiation of time series 1

Figure Forecasting results after first and second differentiation of time series 1

Figure Forecasting plots after first and second differentiation of time series 2

Figure Forecasting results after first and second differentiation of time series 2

# Models’ Evaluation

Shall be used to summarize the transformations done over the original time series. **Shall not exceed 200 characters.**

## Simple Average Model

Shall be used to present the results achieved through the simple average model. **Shall not exceed 200 characters.**

Figure Forecasting plots obtained with Simple Average model over time series 1

Figure Forecasting plots obtained with Simple Average model over time series 2

## Persistence Model

Shall be used to present the results achieved through the persistence model. **Shall not exceed 200 characters.**

Figure Forecasting plots obtained with Persistence model over time series 1

Figure Forecasting plots obtained with Persistence model over time series 2

## Rolling Mean Model

Shall be used to present the results achieved through the rolling mean forecasting algorithms. **Shall not exceed 500 characters.**

Figure Forecasting study over different parameterizations of the rolling mean algorithm over time series 1

Figure Forecasting plots obtained with the best parameterization of rolling mean algorithm, over time series 1

Figure Forecasting results obtained with the best parameterization of rolling mean algorithm, over time series 1

Figure Forecasting study over different parameterizations of the rolling mean algorithm over time series 2

Figure Forecasting plots obtained with the best parameterization of rolling mean algorithm, over time series 2

Figure Forecasting results obtained with the best parameterization of rolling mean algorithm, over time series 2

## ARIMA Model

Shall be used to present the results achieved through the ARIMA forecasting algorithms. **Shall not exceed 500 characters.**

Figure Forecasting study over different parameterizations of the ARIMA algorithm over time series 1

Figure Forecasting plots obtained with the best parameterization of ARIMA algorithm, over time series 1

Figure Forecasting results obtained with the best parameterization of ARIMA algorithm, over time series 1

Figure Forecasting study over different parameterizations of the ARIMA algorithm over time series 2

Figure Forecasting plots obtained with the best parameterization of ARIMA algorithm, over time series 2

Figure Forecasting results obtained with the best parameterization of ARIMA algorithm, over time series 2

## LSTMs Model

Shall be used to present the results achieved through LSTMs. **Shall not exceed 500 characters.**

Figure Forecasting study over different parameterizations of LSTMs over time series 1

Figure Forecasting plots obtained with the best parameterization of LSTMs, over time series 1

Figure Forecasting results obtained with the best parameterization of LSTMs, over time series 1

Figure Forecasting study over different parameterizations of the LSTMs over time series 2

Figure Forecasting plots obtained with the best parameterization of LSTMs, over time series 2

Figure Forecasting results obtained with the best parameterization of LSTMs, over time series 2

# Critical Analysis

Shall be used to present a summary of the results achieved with the different forecasting techniques, and the impact of the different preparation tasks on their performance.

A critical assessment of the best models shall be presented, clearly stating if the models seem to be good enough for the problem at hand.

**Additional charts may be presented here. Shall not exceed 2000 characters.**