|  |  |  |  |
| --- | --- | --- | --- |
| Logo, company name  Description automatically generated | | **DS 2022** | |
| Data Science Project | | | |
| **Team nr:** 21 | **Student 1 :** \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ | | **IST nr:** \_\_\_\_\_\_\_ |
| **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

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

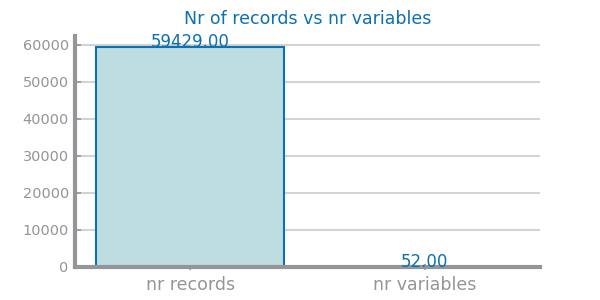
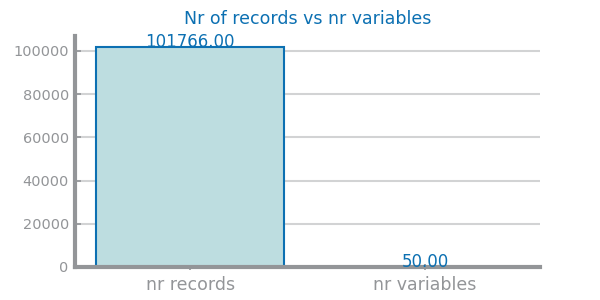


Figure 2: records vs variables dataset 1

Figure 3: records vs variables dataset 2

Figure 4 Nr variables per type for dataset 1 (left) and dataset 2 (right)

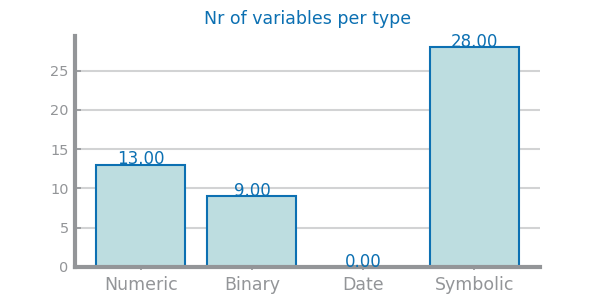
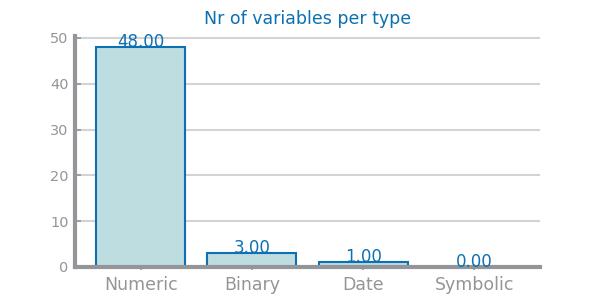


Figure 5: nr variables per type – dataset 1 Figure 6: Nr variables per type - dataset 2

Figure 7 Nr missing values for dataset 1 (left) and dataset 2 (right)

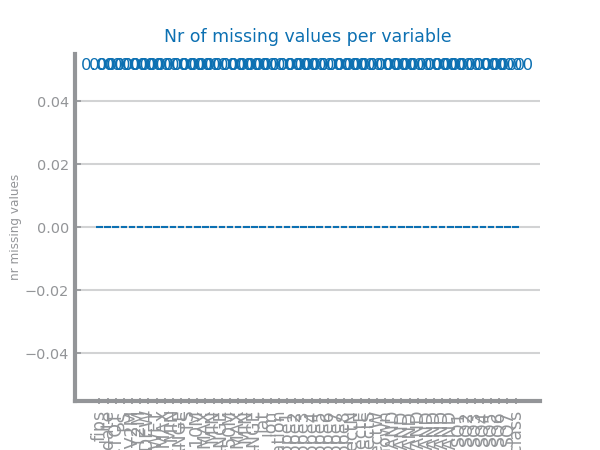
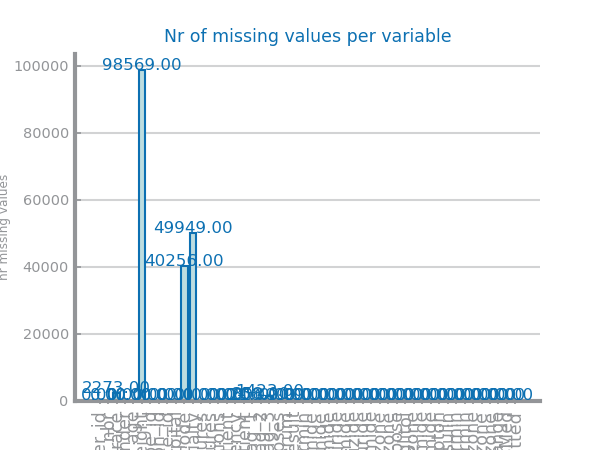


Figure 8: missing values dataset 1

Figure 9: Missing values dataset 2

## Data Distribution

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

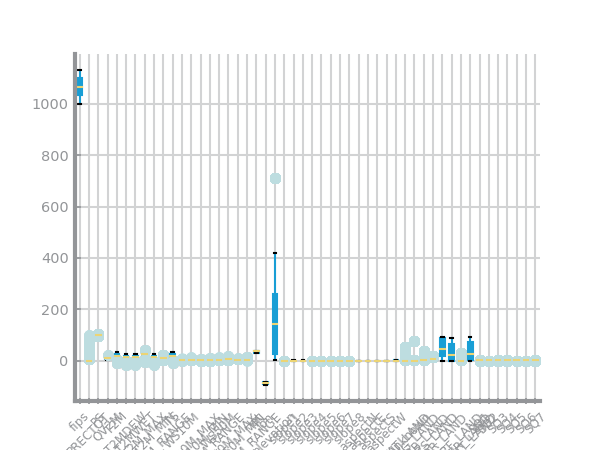
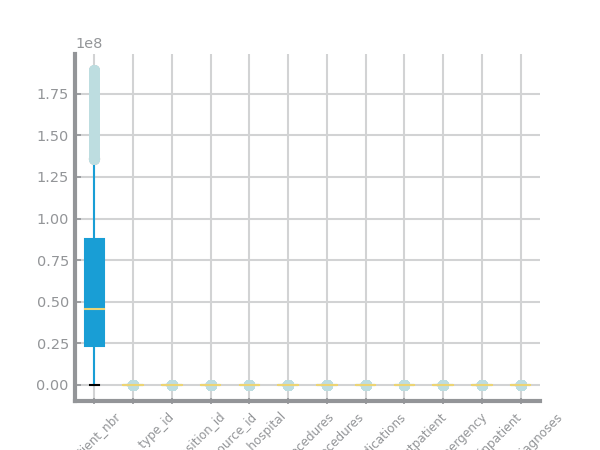


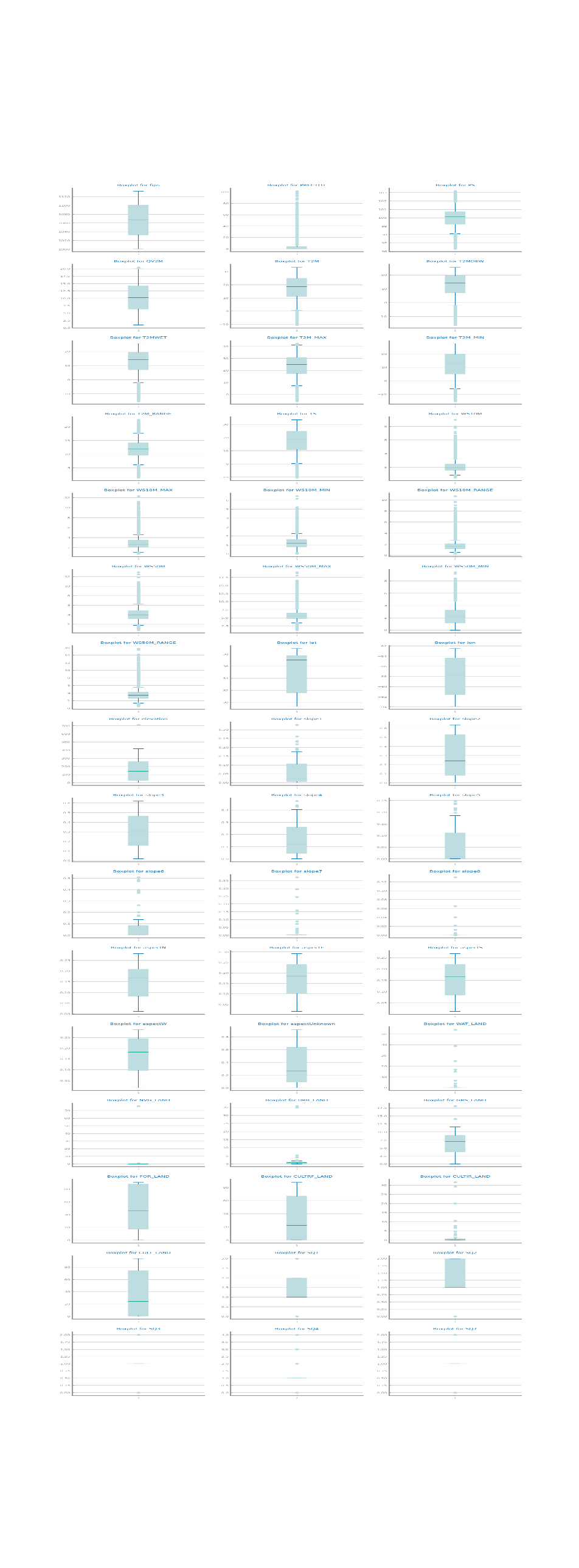
Figure 11: global boxplot dataset 2

Figure 12: global boxplot dataset 1

Figure 13 Single variable boxplots for dataset 1



Figure 14 Single variable boxplots s for dataset 2



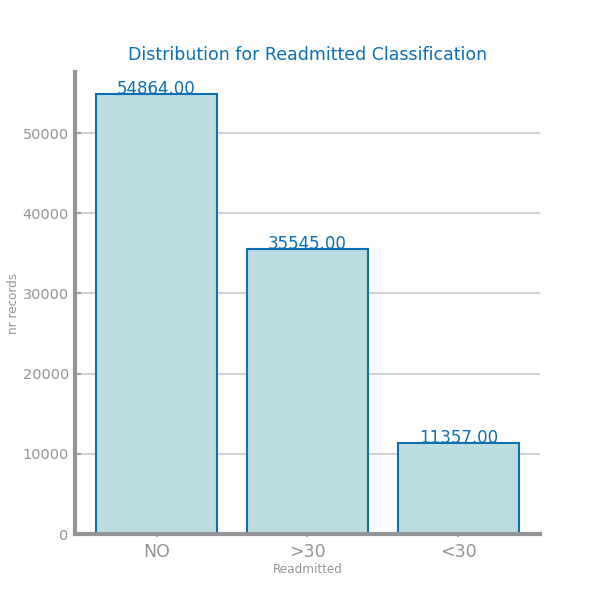
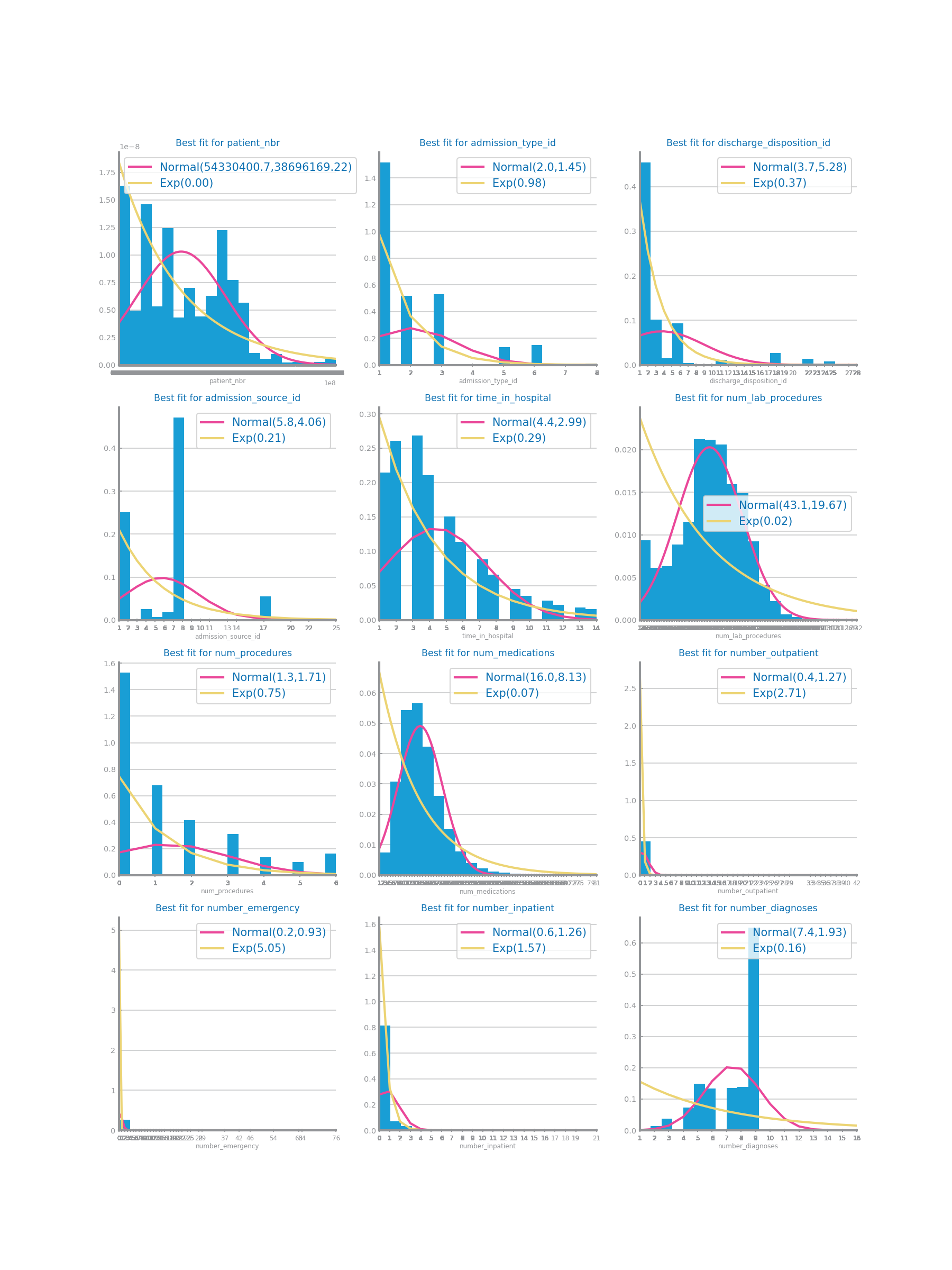
Figure 16 Histograms for dataset 1

Figure 15:class distribution dataset 1

Figure 17: numeric distribution - Dataset 1

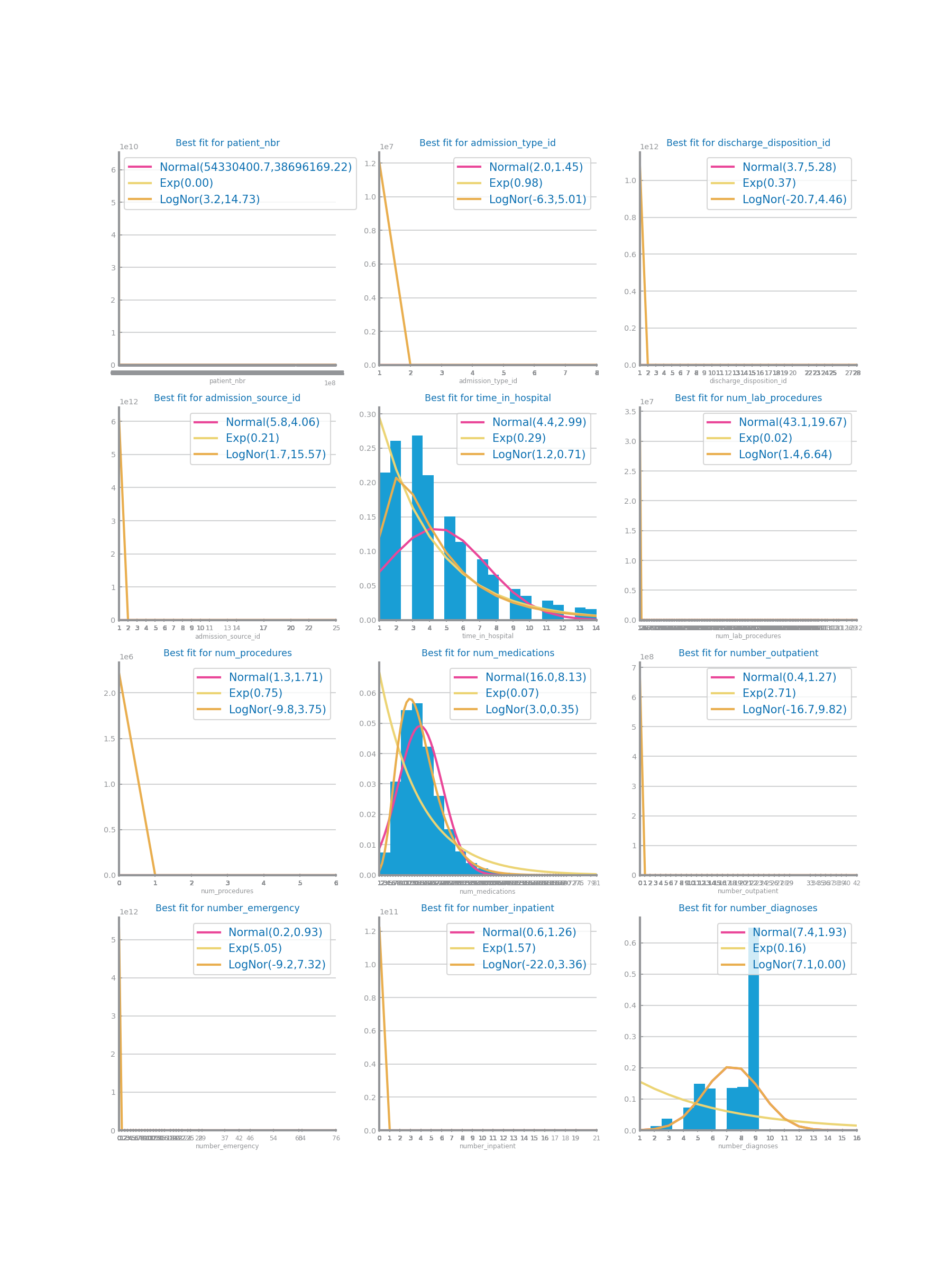


Figure 18: Numeric distribution with log - dataset 1

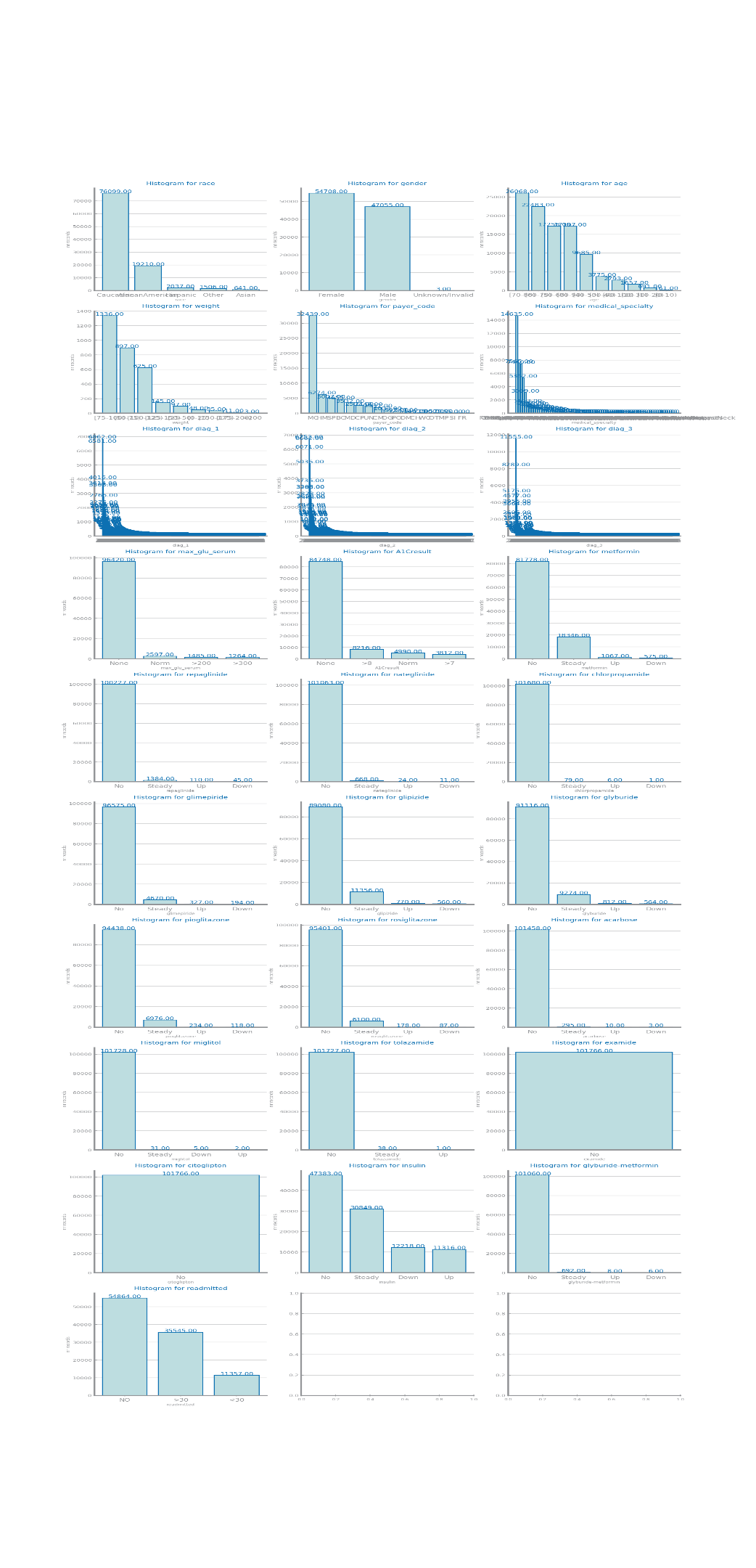


Figure 19: symbolic variables histograms dataset 1

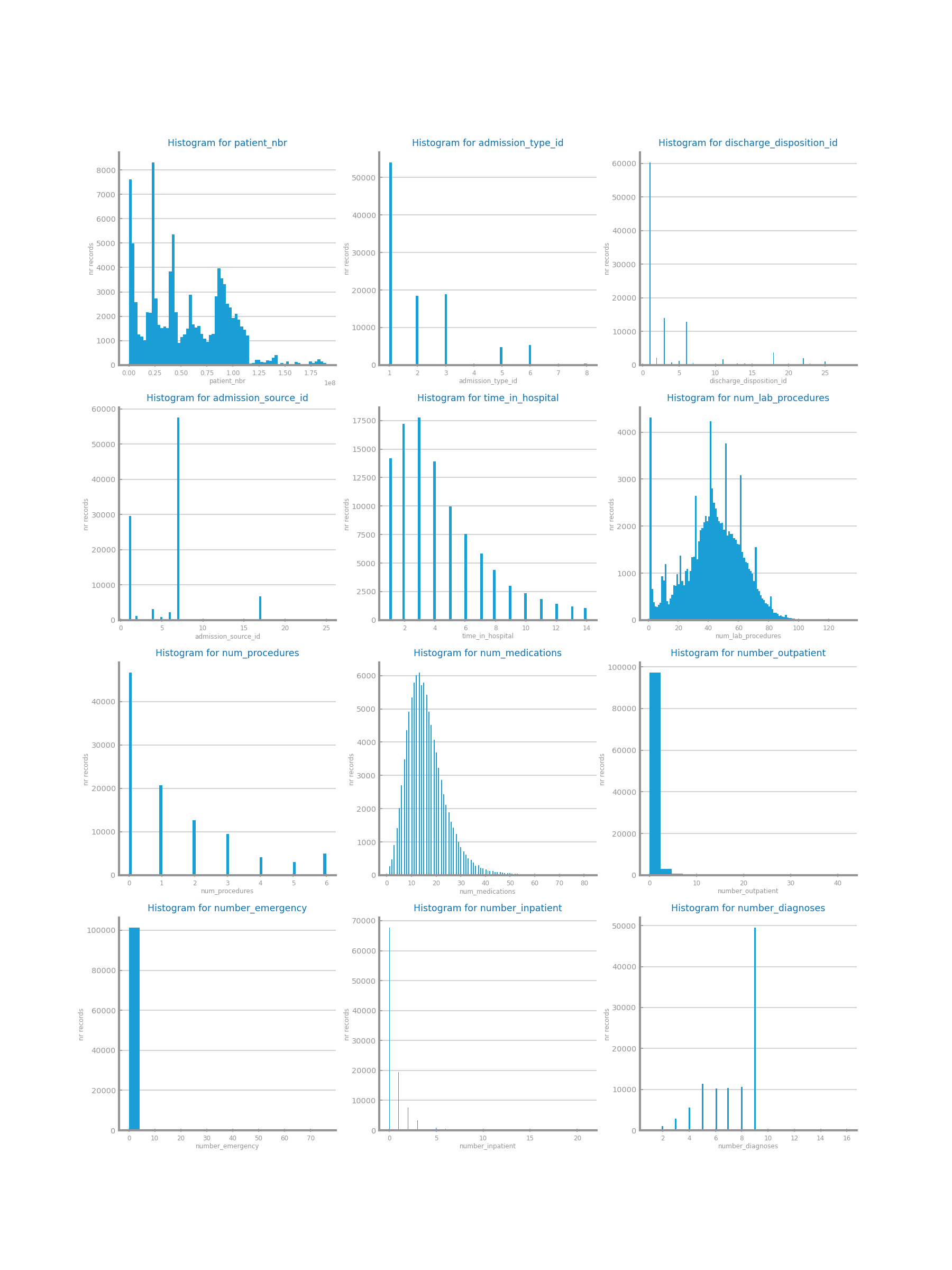


Figure 20: single histograms numeric - dataset 1

Figure 21 Histograms for dataset 2

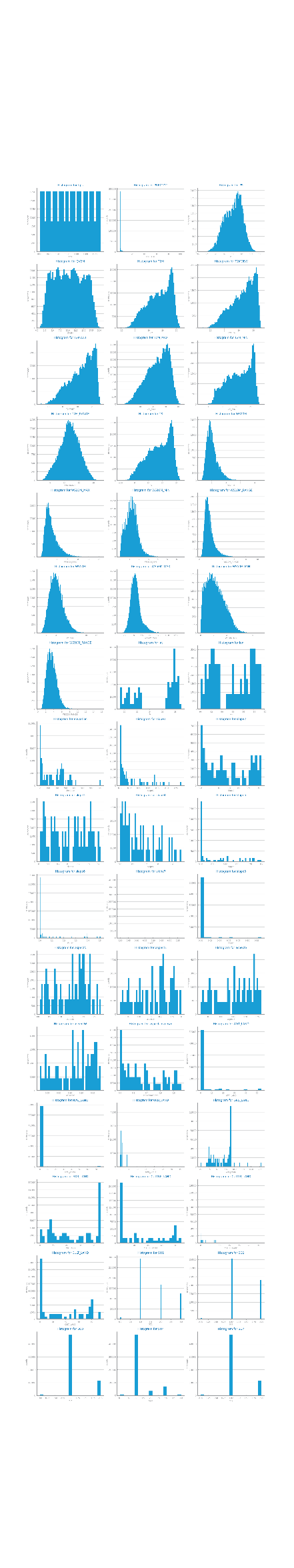


Figure 22: single histograms numeric dataset 2

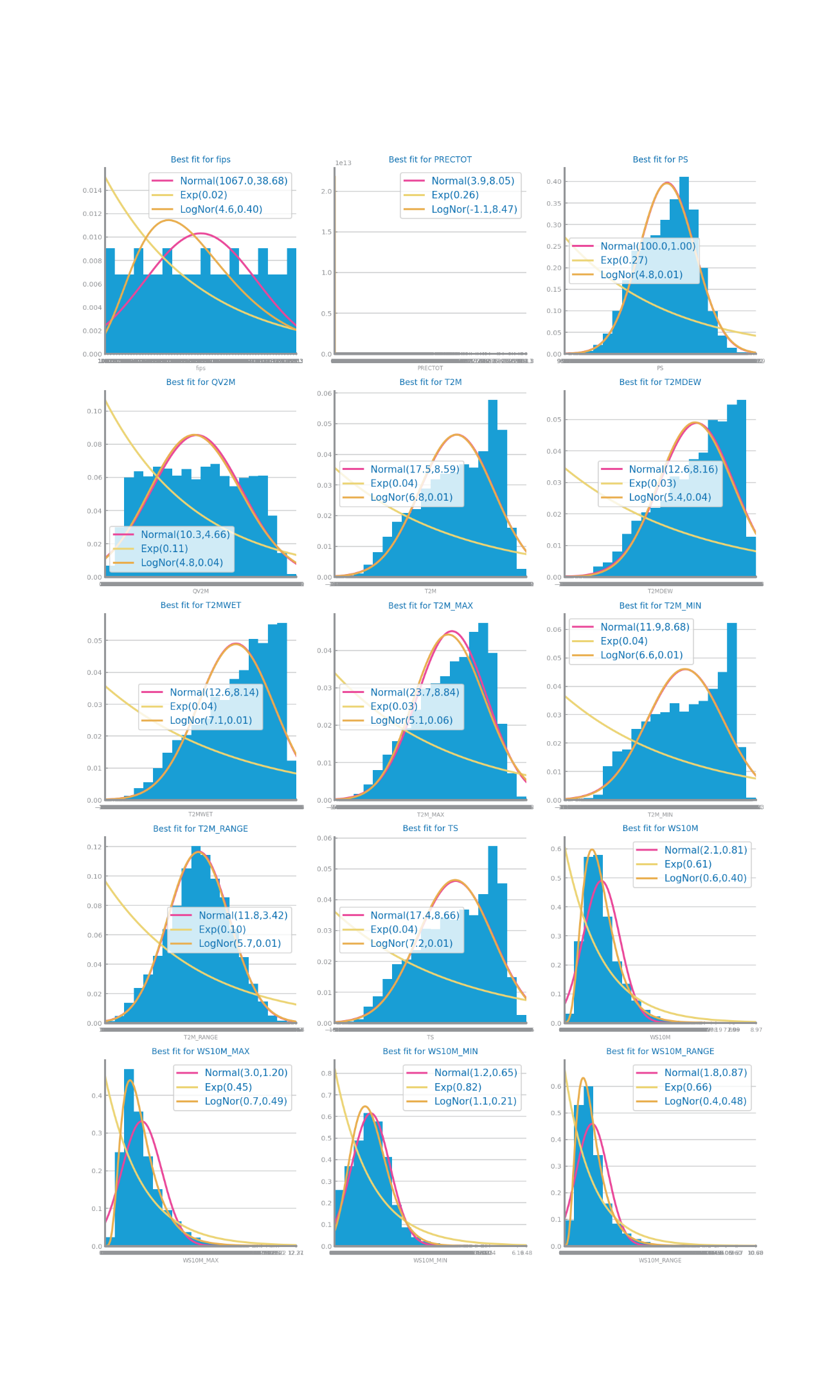


Figure 23: histograms with log - dataset 2

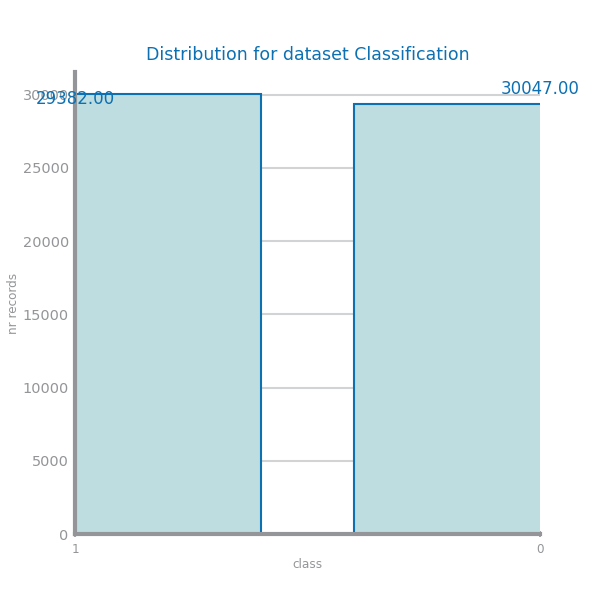


Figure 24: class distribution - dataset 2

## Data Granularity

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 25 Granularity analysis for dataset 1

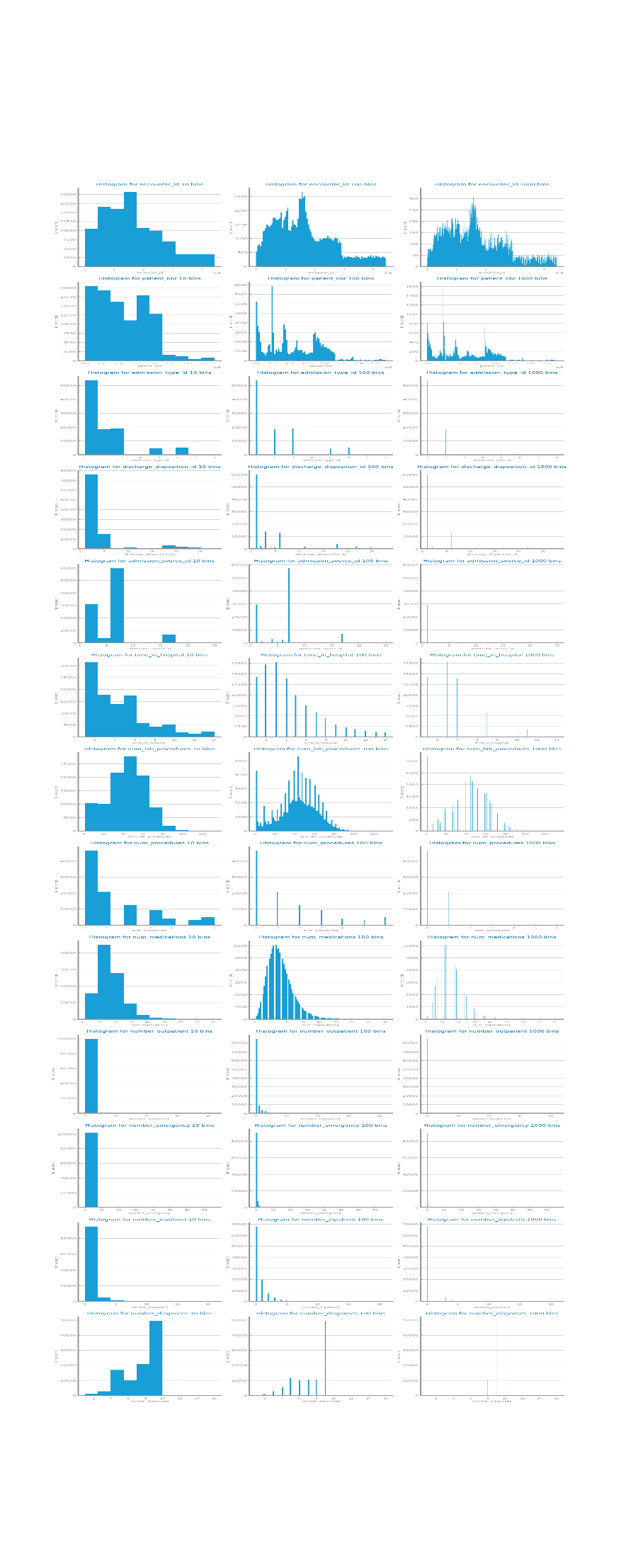


Figure 26: granularity study numeric variables dataset 1

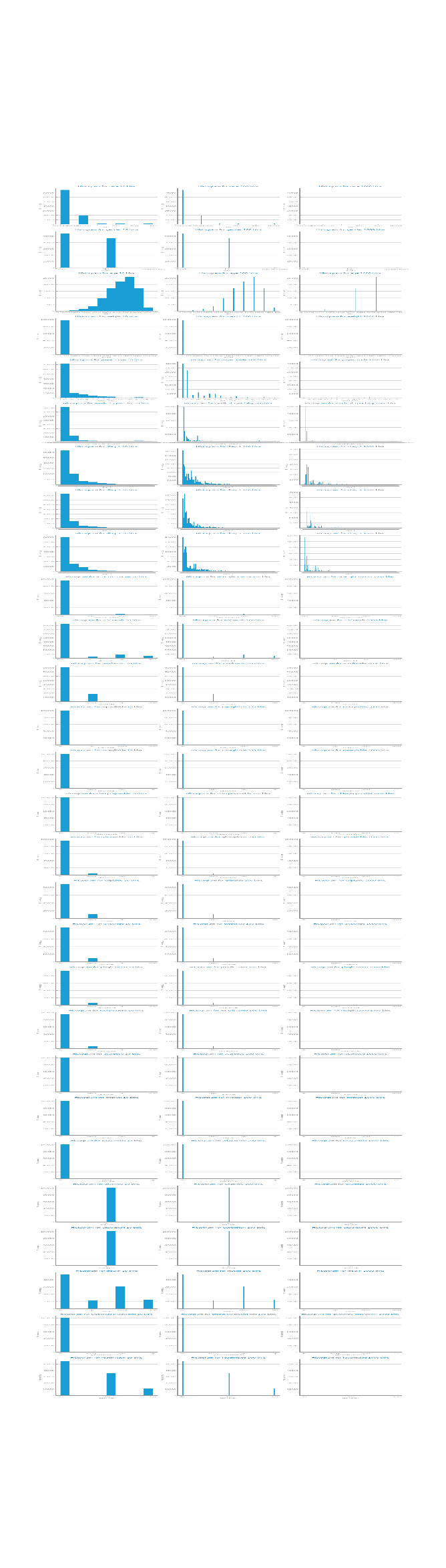


Figure 27: granularity study symbolic variables dataset 1

Figure 28 Granularity analysis for dataset 2

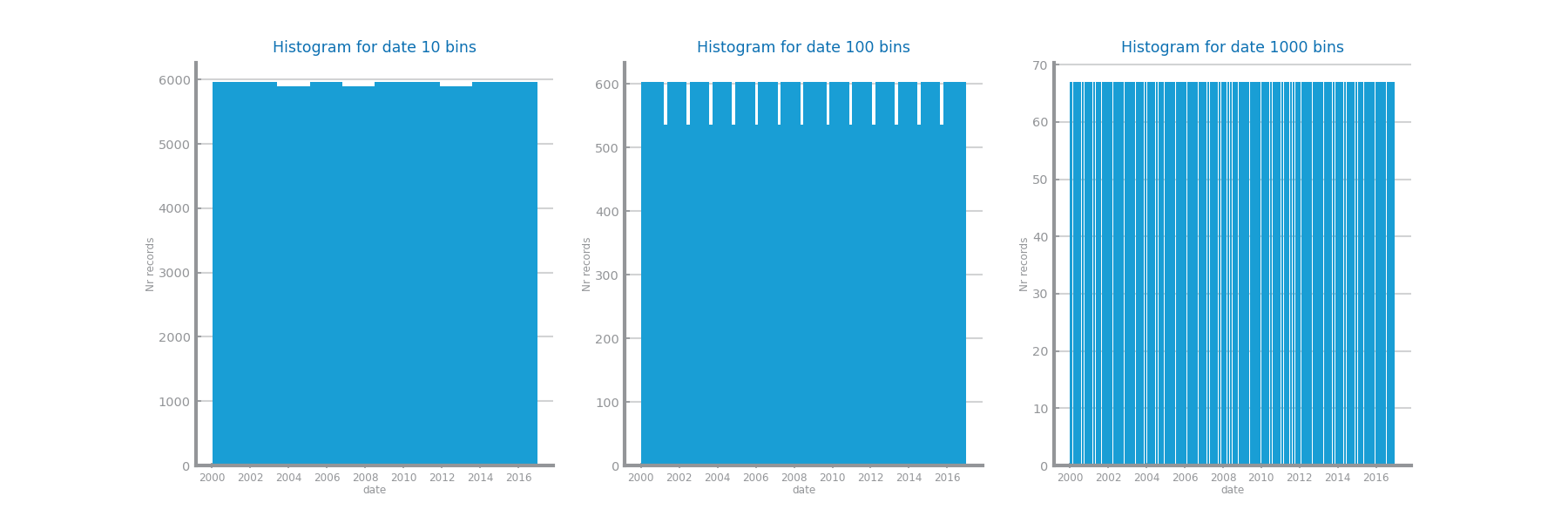


Figure 29: granularity study date - dataset 2

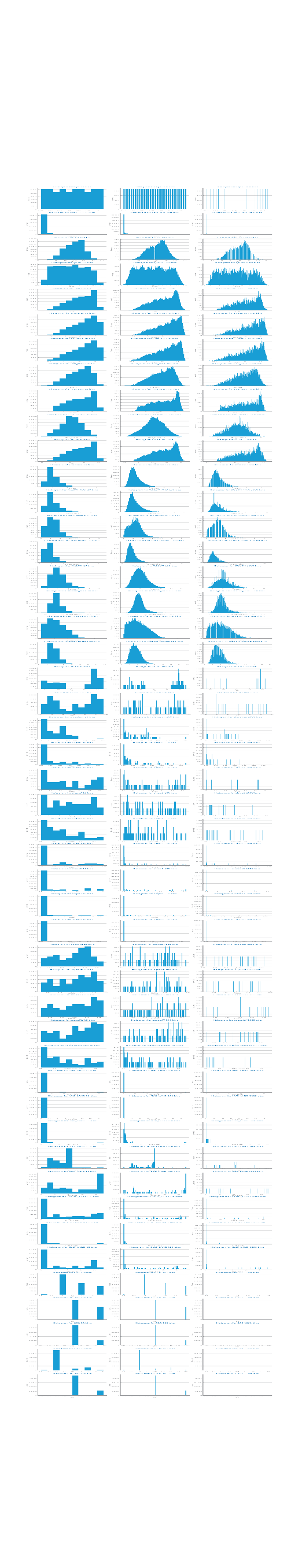


Figure 30: granularity analysis numeric variables dataset 2

## 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 31 Sparsity analysis for dataset 1



Figure 32: Sparsity study numeric variables dataset 1

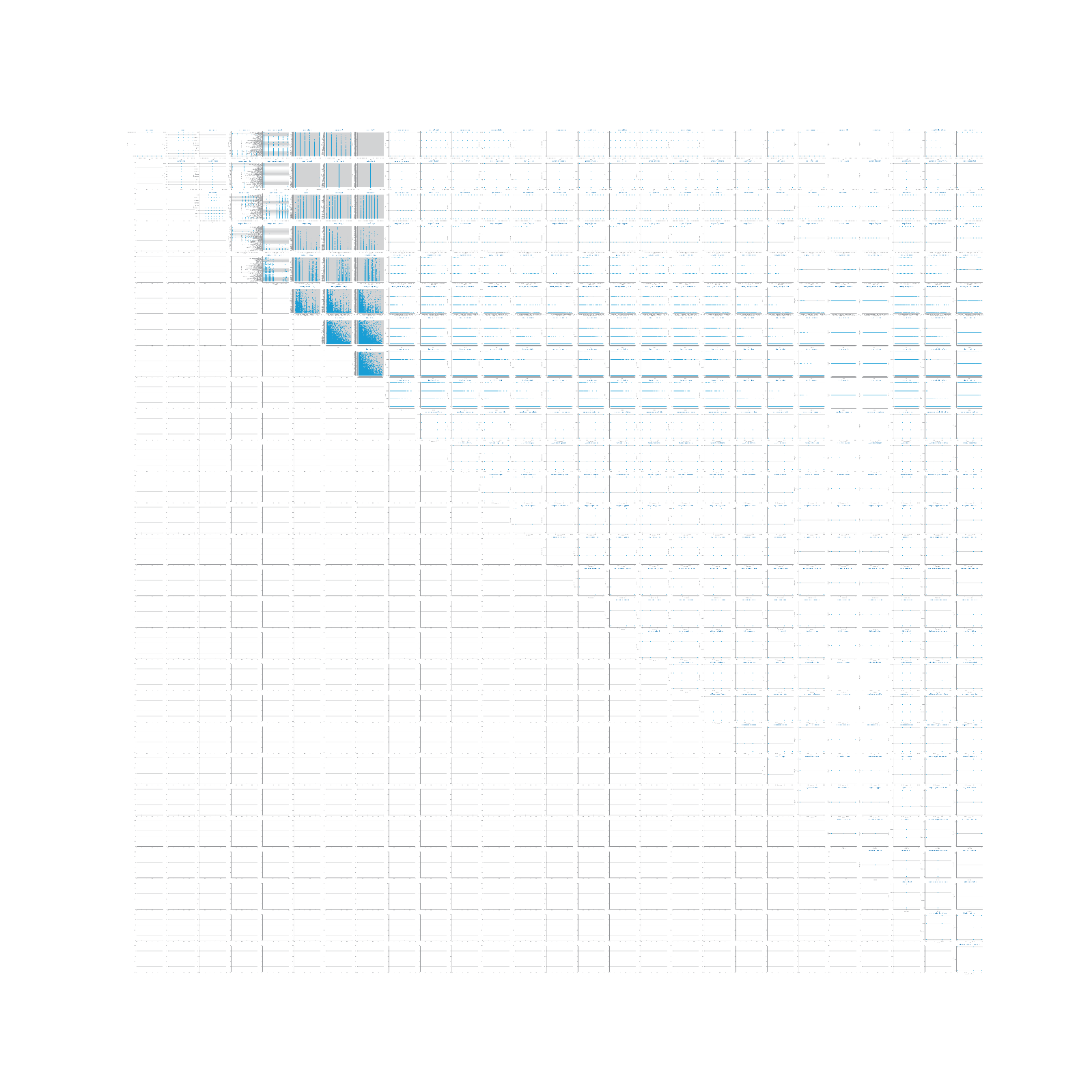


Figure 33: Sparsity study symbolic variables dataset 1

Figure 34 Sparsity analysis for dataset 2

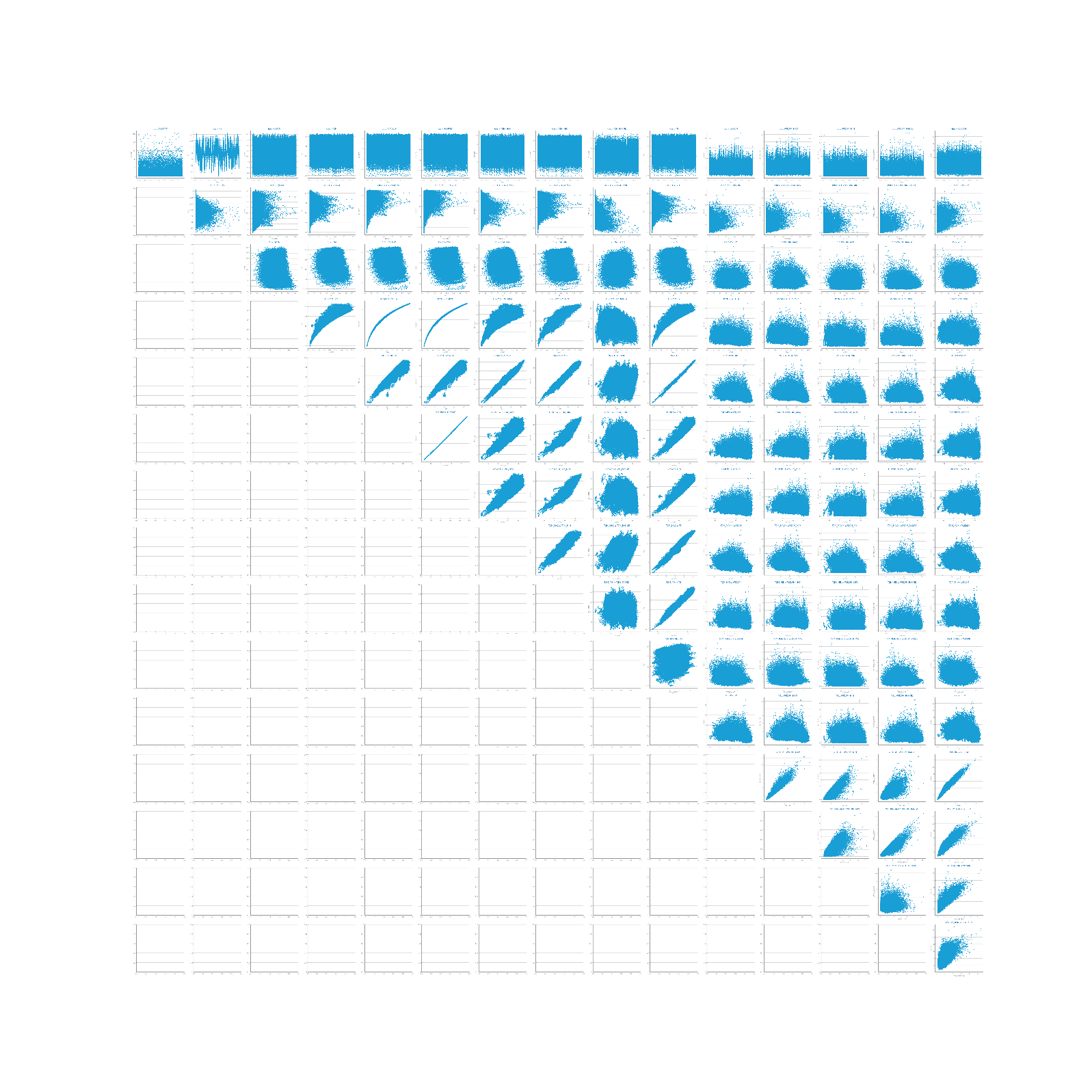


Figure 35: sparsity study dataset 2

Figure 36 Correlation analysis for dataset 1

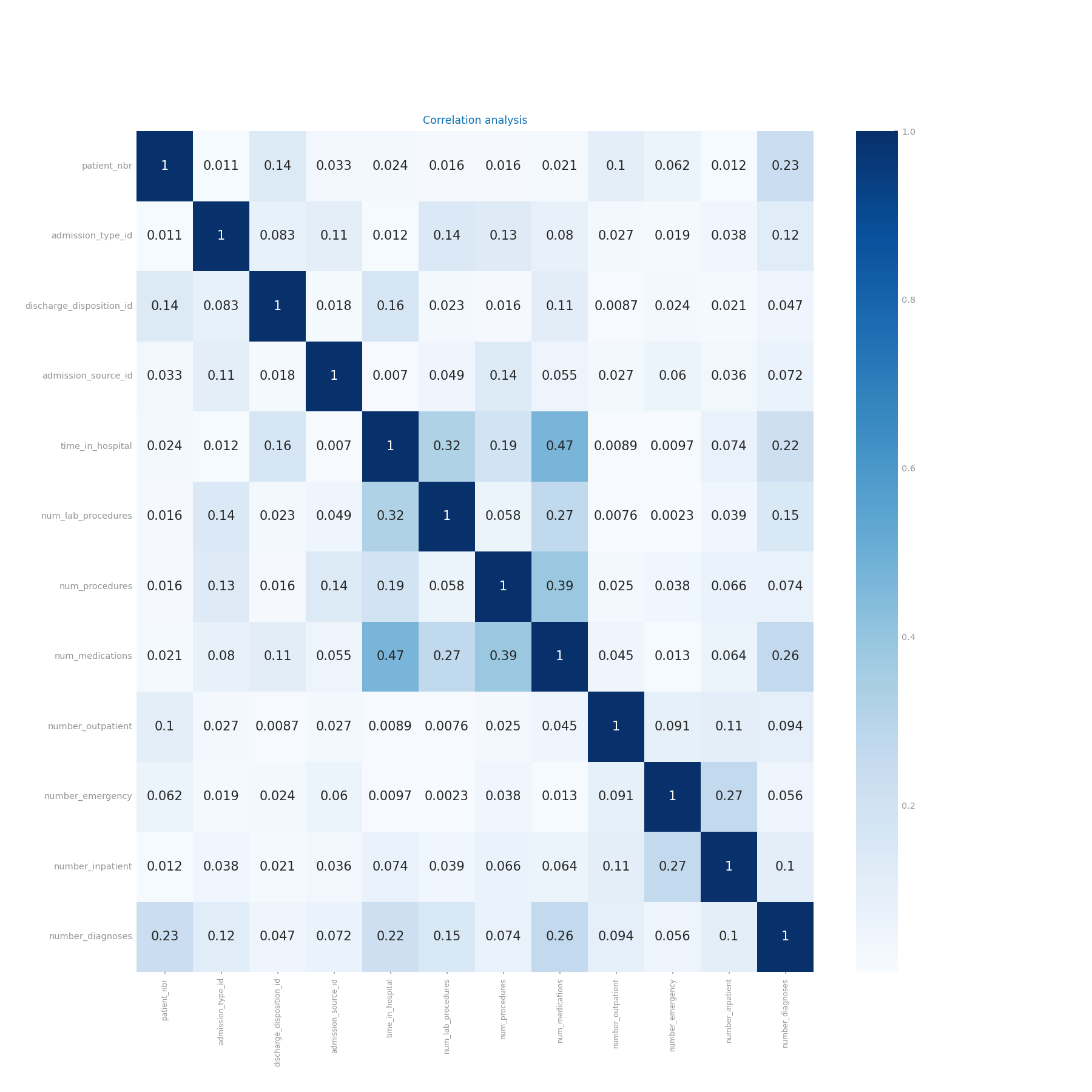


Figure 37: correlation analysis dataset 1

Figure 38 Correlation analysis for dataset 2

Add the improved one

# Data Preparation

## Variables Encoding

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

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 39 Missing values imputation results with different approaches for dataset 1

Figure 40 Missing values imputation results with different approaches for dataset 2

## Outliers Treatment

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 41 Outliers imputation results with different approaches for dataset 1

Figure 42 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 43 Scaling results with different approaches for dataset 1

Figure 44 Scaling results with different approaches for dataset 2

## 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 45 Balancing results with different approaches for dataset 1

Figure 46 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 47 Feature selection of redundant variables results with different parameters for dataset 1

Figure 48 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 49 Principal components analysis and feature extraction results for dataset 1

Figure 50 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 51 Feature generation results for dataset 1

Figure 52 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 53 Naïve Bayes alternatives comparison for dataset 1

Figure 54 Naïve Bayes alternative comparison for dataset 2

Figure 55 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 56 KNN different parameterizations comparison for dataset 1

Figure 57 KNN different parameterizations comparison for dataset 2

Figure 58 KNN overfitting analysis for dataset 1 (left) and dataset 2 (right)

Figure 59 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 60 Decision Trees different parameterizations comparison for dataset 1

Figure 61 Decision Trees different parameterizations comparison for dataset 2

Figure 62 Decision Trees overfitting analysis for dataset 1 (left) and dataset 2 (right)

Figure 63 Decision trees best model results for dataset 1 (left) and dataset 2 (right)

Figure 64 Best tree for dataset 1

Figure 65 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 66 Random Forests different parameterizations comparison for dataset 1

Figure 67 Random Forests different parameterizations comparison for dataset 2

Figure 68 Random Forests overfitting analysis for dataset 1 (left) and dataset 2 (right)

Figure 69 Random Forests best model results for dataset 1 (left) and dataset 2 (right)

Figure 70 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 71 Gradient boosting different parameterizations comparison for dataset 1

Figure 72 Gradient boosting different parameterizations comparison for dataset 2

Figure 73 Gradient boosting overfitting analysis for dataset 1 (left) and dataset 2 (right)

Figure 74 Gradient boosting best model results for dataset 1 (left) and dataset 2 (right)

Figure 75 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 76 MLP different parameterizations comparison for dataset 1

Figure 77 MLP different parameterizations comparison for dataset 2

Figure 78 MLP overfitting analysis for dataset 1 (left) and dataset 2 (right)

Figure 79 Loss curves analysis for dataset 1 (left) and dataset 2 (right)

Figure 80 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 81 Time series 1 at the most granular detail

Figure 82 Time series 1 at the second chosen granularity

Figure 83 Time series 1 at the third chosen granularity

Figure 84 Time series 2 at the most granular detail

Figure 85 Time series 2 at the second chosen granularity

Figure 86 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 87 Boxplot(s) for time series 1

Figure 88 Boxplot(s) for time series 2

Figure 89 Histogram(s) for time series 1

Figure 90 Histogram(s) for time series 2

Figure 91 Stationarity study for time series 1

Figure 92 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.**

Figure 93 Forecasting plots after different aggregations on time series 1

Figure 94 Forecasting results after different aggregations on time series 1

Figure 95 Forecasting plots after different aggregations on time series 2

Figure 96 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 97 Forecasting plots after different smoothing parameterizations on time series 1

Figure 98 Forecasting results after different smoothing parameterizations on time series 1

Figure 99 Forecasting plots after different smoothing parameterizations on time series 2

Figure 100 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 101 Forecasting plots after first and second differentiation of time series 1

Figure 102 Forecasting results after first and second differentiation of time series 1

Figure 103 Forecasting plots after first and second differentiation of time series 2

Figure 104 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 105 Forecasting plots obtained with Simple Average model over time series 1

Figure 106 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 107 Forecasting plots obtained with Persistence model over time series 1

Figure 108 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 109 Forecasting study over different parameterizations of the rolling mean algorithm over time series 1

Figure 110 Forecasting plots obtained with the best parameterization of rolling mean algorithm, over time series 1

Figure 111 Forecasting results obtained with the best parameterization of rolling mean algorithm, over time series 1

Figure 112 Forecasting study over different parameterizations of the rolling mean algorithm over time series 2

Figure 113 Forecasting plots obtained with the best parameterization of rolling mean algorithm, over time series 2

Figure 114 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 115 Forecasting study over different parameterizations of the ARIMA algorithm over time series 1

Figure 116 Forecasting plots obtained with the best parameterization of ARIMA algorithm, over time series 1

Figure 117 Forecasting results obtained with the best parameterization of ARIMA algorithm, over time series 1

Figure 118 Forecasting study over different parameterizations of the ARIMA algorithm over time series 2

Figure 119 Forecasting plots obtained with the best parameterization of ARIMA algorithm, over time series 2

Figure 120 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 121 Forecasting study over different parameterizations of LSTMs over time series 1

Figure 122 Forecasting plots obtained with the best parameterization of LSTMs, over time series 1

Figure 123 Forecasting results obtained with the best parameterization of LSTMs, over time series 1

Figure 124 Forecasting study over different parameterizations of the LSTMs over time series 2

Figure 125 Forecasting plots obtained with the best parameterization of LSTMs, over time series 2

Figure 126 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.**