Lab 2
Data Profiling

GOAL

To explore the data in order to identify the data characteristics and the required preparation tasks to apply next.

WORK TO BE DONE

To explore the data following the four perspectives:

Data dimensionality;
Data distribution;
Data granularity for non-numeric data (temporal or symbolic);
Data sparsity.

REPORT TO DELIVER

PDF file only with charts and tables. No analysis or justification is needed at this point.
Suggested charts per dataset:

Dimensionality:
Number of records x Number of variables;
Types of variables
Missing values

Distribution:
Numerical variables: boxplots, histograms, and probability density functions.
Binary variables: histograms.
Other symbolic variables: histograms at the most atomic level.
Outliers study
Class distribution.

Granularity:
Temporal or *Other symbolic variables:* histograms at different levels of detail.

Sparsity:
Scatter-plots for pairs of variables, preferentially

Correlation:
Correlation matrix, including the class variable.

Good work!!!