TÉCNICO LISBOA

DS 2025

Data Science Project

Team nr: 26	**Student 1:**	Alexandre Cachera	**IST nr:** 116285
	Student 2:	Jaime Gosai	**IST nr:** 99239
	Student 3:	Fredrik Preus Dovland	**IST nr:** 116071
	Student 4:	Lukas Bruns	**IST nr:** 116926

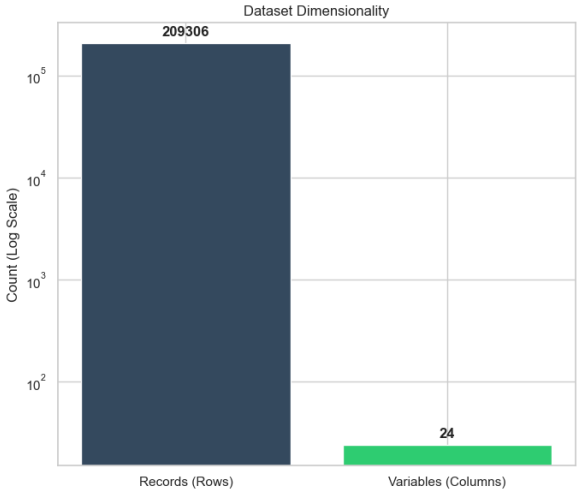
CLASSIFICATION

1 DATA PROFILING

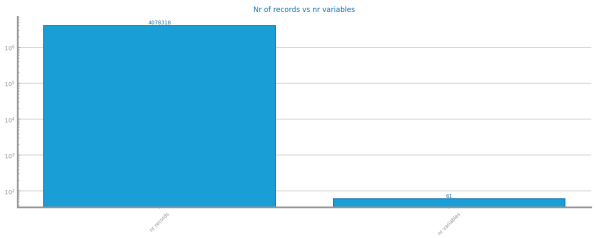
Identifying data leakage in the Traffic dataset led us to remove post-accident variables. This ensures the model only uses pre-crash data, preventing artificial performance inflation.

Data Dimensionality

Both datasets show high record-to-variable ratios, supporting model stability. Dataset 1 (Traffic security) contains many categorical variables and implicit missing values (e.g., "UNKNOWN"), requiring encoding and specialized cleaning. Dataset 2 (Flights) is mostly numerical but includes categorical features and 3% missing values in several variables.

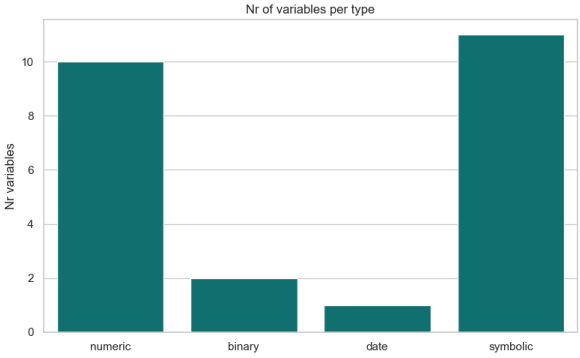


(a) Dataset 1

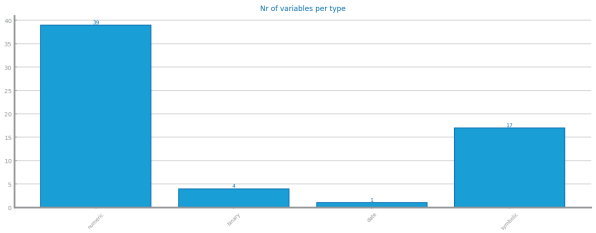


(b) Dataset 2

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

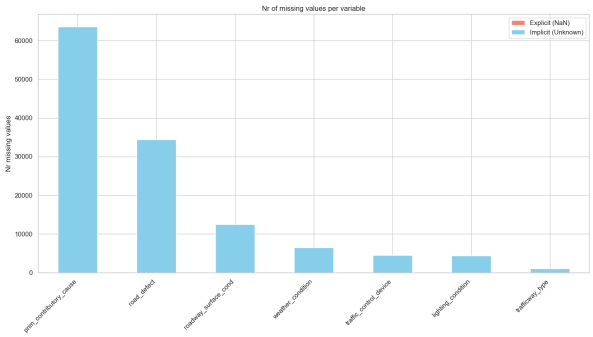


(a) Dataset 1

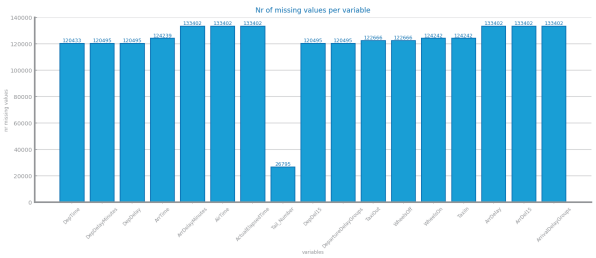


(b) Dataset 2

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



(a) Dataset 1

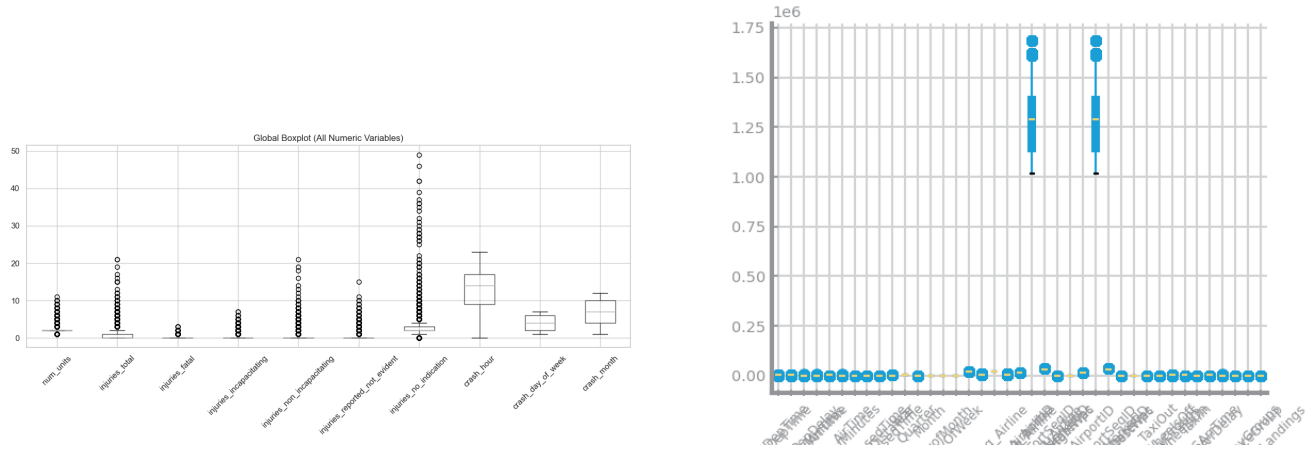


(b) Dataset 2

Figure 3: Nr missing values for dataset 1 (left) and dataset 2 (right)

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.**



(a) Dataset 1

(b) Dataset 2

Figure 4: Global boxplots dataset 1 (left) and dataset 2 (right)

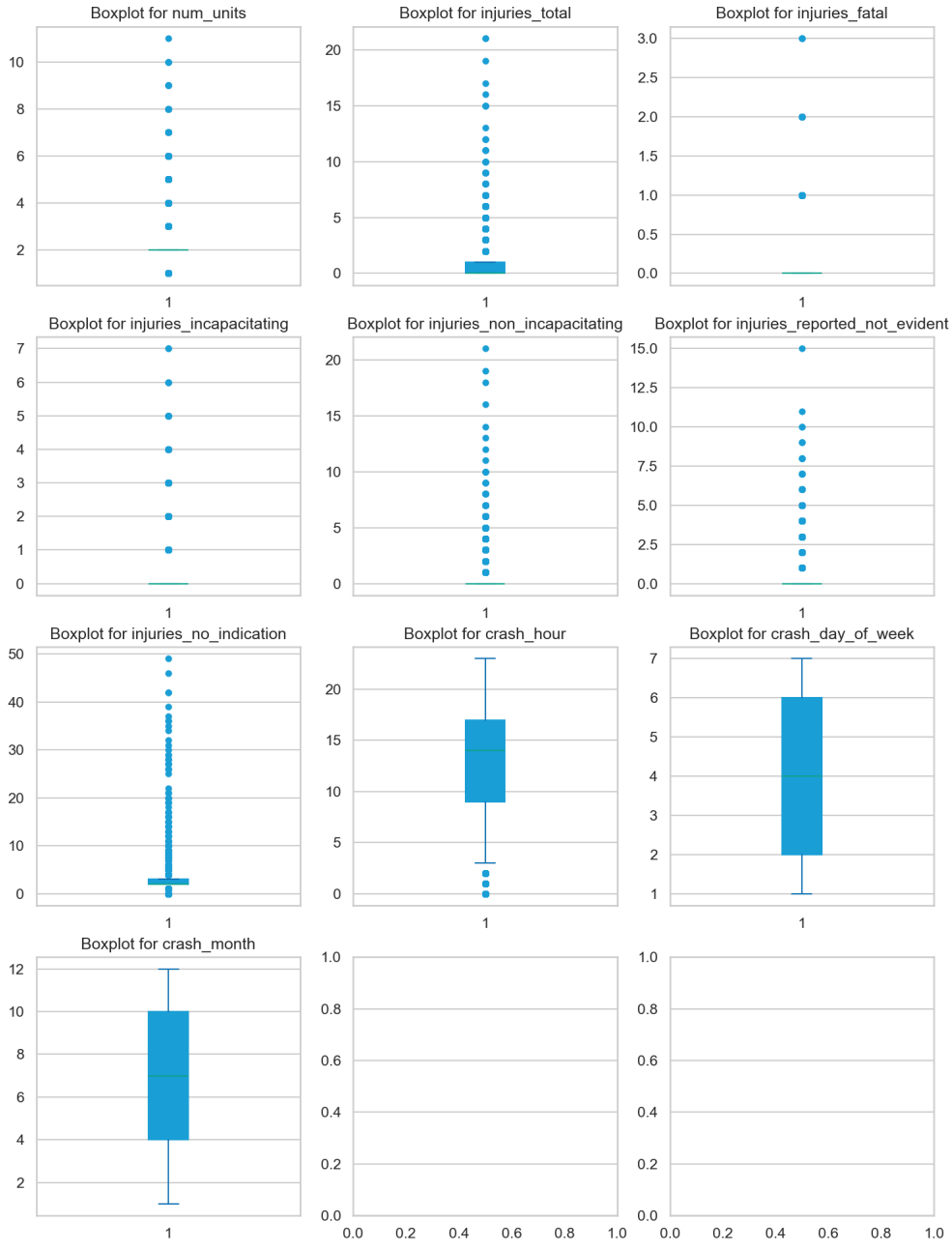


Figure 5: Single variables boxplots for dataset 2

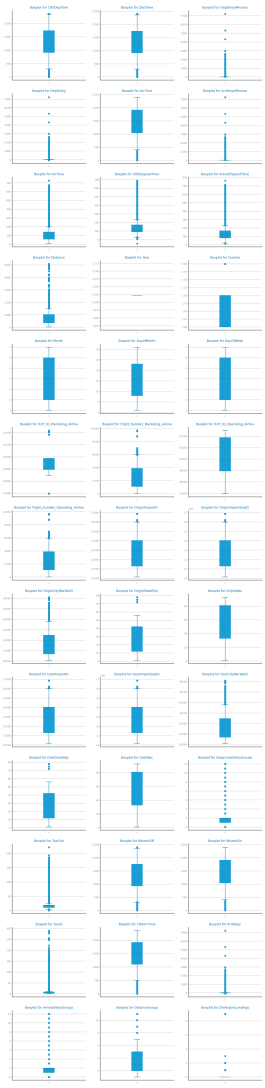


Figure 6: Single variables boxplots for dataset 2

Figure 7: Histograms for dataset 1

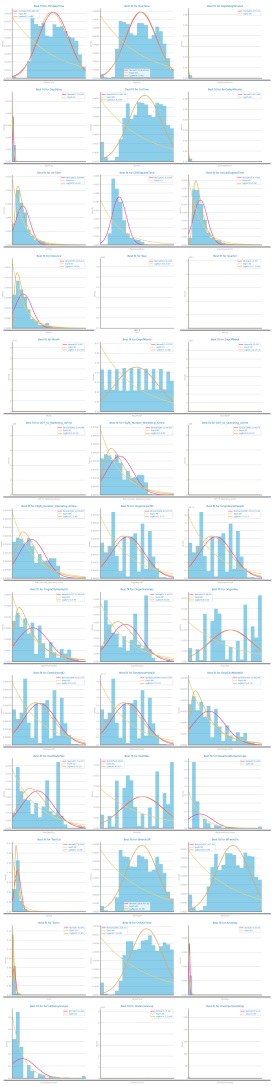


Figure 8: Histograms for dataset 2

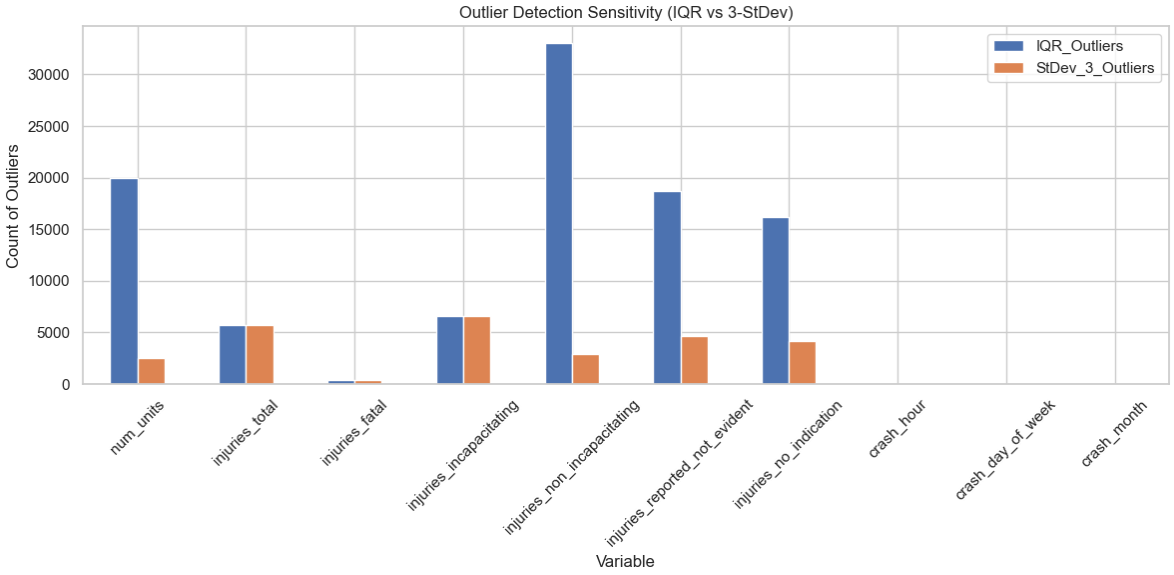


Figure 9: Outliers study dataset 1

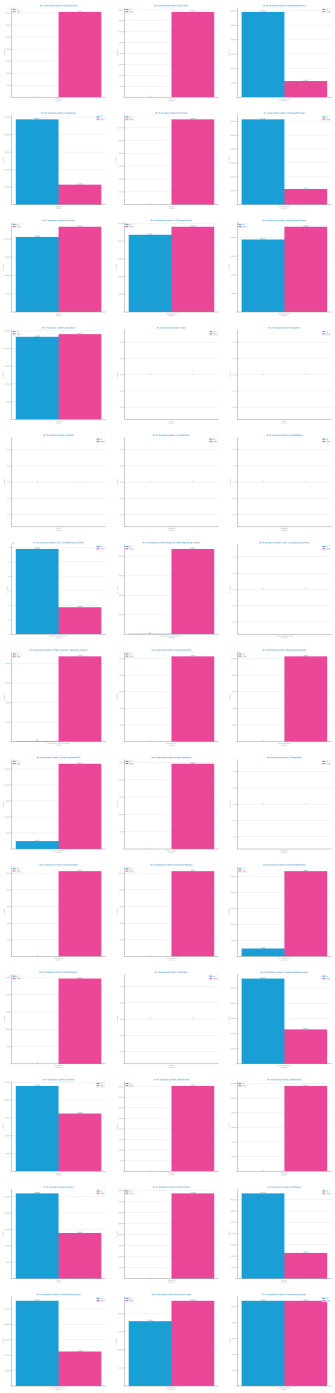


Figure 10: Outliers study dataset 2

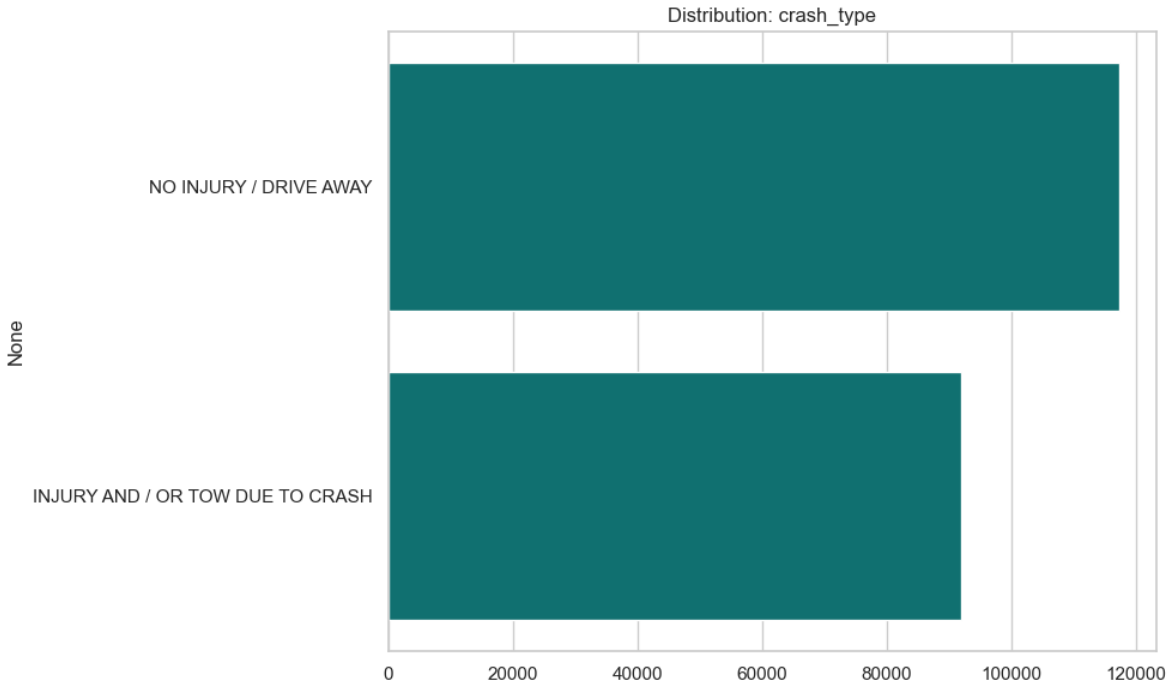


Figure 11: Class distribution for dataset 1

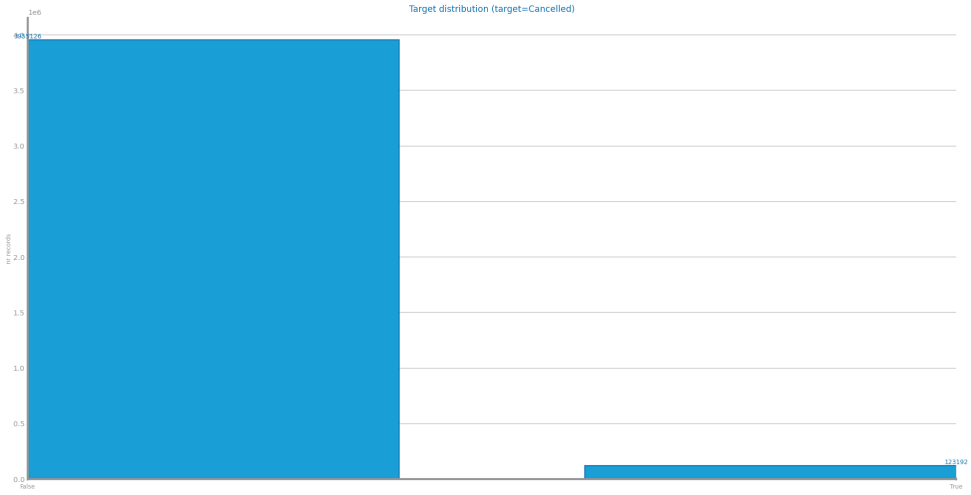


Figure 12: Class distribution for dataset 2

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 500 characters.**

Figure 13: Granularity analysis for dataset 1

Figure 14: Granularity analysis for 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 500 characters.**

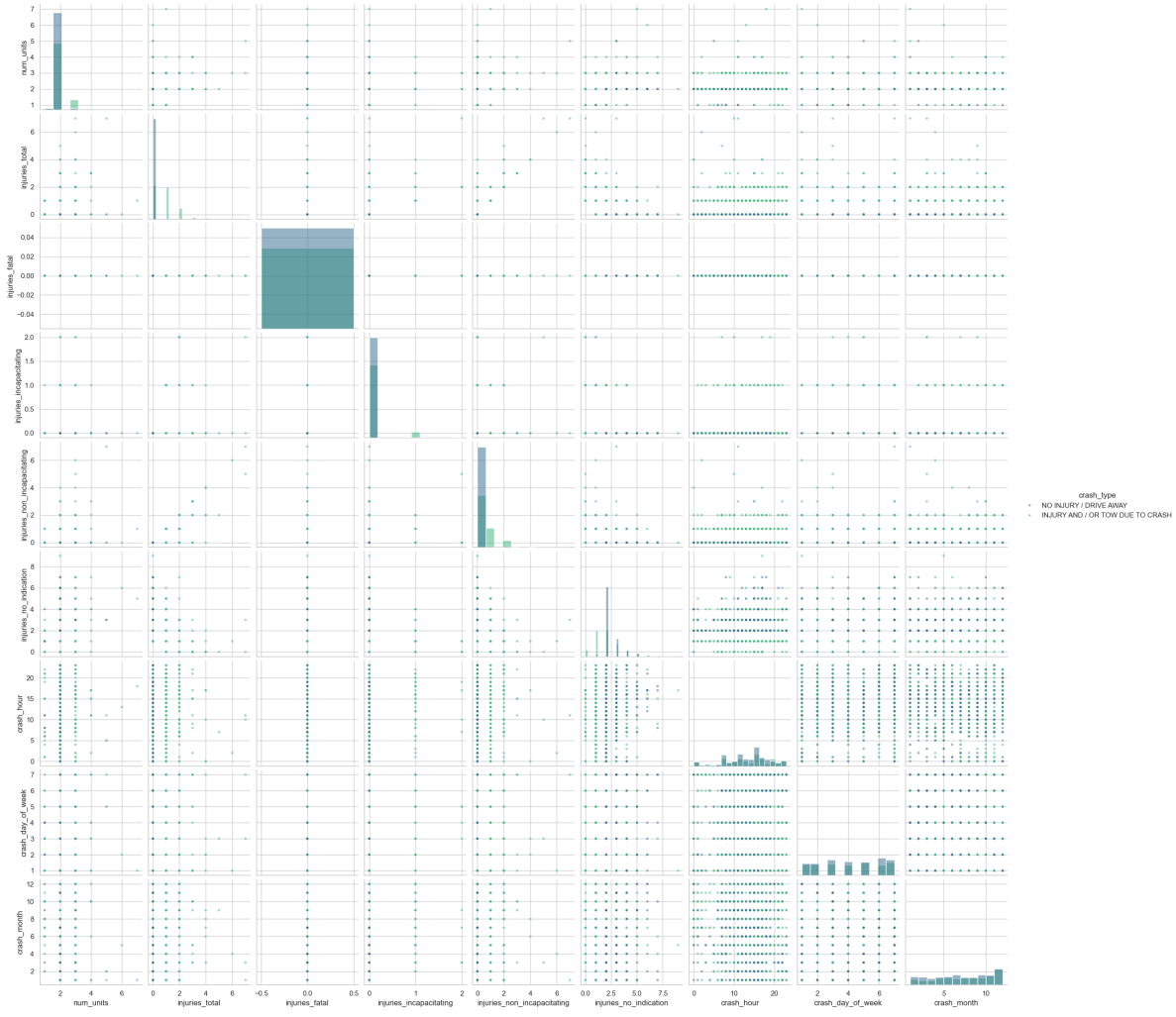


Figure 15: Sparsity analysis for dataset 1

Figure 16: Sparsity analysis for dataset 2

Figure 17: Correlation analysis for dataset 1

Figure 18: Correlation analysis for dataset 2

2 DATA PREPARATION

Variables Encoding

Dataset 1 transformations include Binary Mapping for 'intersection related i' and Cyclical Encoding (sine/cosine) for 'crash hour', 'crash month', and 'crash day of week'. Ordinal Encoding was applied to 'damage' and 'most severe injury' to preserve their inherent hierarchy. Infrequent labels in categorical variables were grouped to mitigate sparsity found during distribution profiling. All remaining categorical variables underwent One-Hot Encoding. Dataset 2 used Target Encoding for 'Hub Airline', 'Route', 'Airline', 'Origin', and 'Dest' to handle high cardinality effectively. Cyclical Encoding (sine/cosine) was applied to 'Month', 'DayOfWeek', 'DayofMonth', 'CRSDepTime', and 'CRSArrTime' to preserve periodic continuity. One-Hot Encoding was not applied to symbolic variables because the excessive unique values in routes and airports would have created an unmanageable, high-dimensional, and sparse feature space.

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 500 characters.**

	Constant		**Statistical**	
	KNN	NB	KNN	NB
Precision				
Recall				
F1				

Table 1: Missing values imputation results with different approaches for dataset 1

	Constant		**Statistical**	
	KNN	NB	KNN	NB
Precision				
Recall				
F1				

Table 2: 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 500 characters.**

	Replace		**Truncate**	
	KNN	NB	KNN	NB
Precision				
Recall				
F1				

Table 3: Outliers imputation results with different approaches for dataset 1

	Replace		Truncate	
	KNN	NB	KNN	NB
Precision				
Recall				
F1				

Table 4: 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.**

	Normalization		Standardization	
	KNN	NB	KNN	NB
Precision				
Recall				
F1				

Table 5: Scaling results with different approaches for dataset 1

	Normalization		Standardization	
	KNN	NB	KNN	NB
Precision				
Recall				
F1				

Table 6: 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 500 characters.**

	Oversampling		Smote	
	KNN	NB	KNN	NB
Precision				
Recall				
F1				

Table 7: Balancing results with different approaches for dataset 1

	Oversampling		Smote	
	KNN	NB	KNN	NB
Precision				
Recall				
F1				

Table 8: 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 (based on correlation) and relevant (based on variation) 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 500 characters.**

Figure 19: Feature selection of redundant variables results with different parameters for dataset 1

Figure 20: Feature selection of redundant variables results with different parameters for dataset 2

Figure 21: Feature selection of relevant variables results with different parameters for dataset 1 (variance study)

Figure 22: Feature selection of relevant variables results with different parameters for dataset 2 (variance study)

Additional Feature Generation

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 summarise all variables generated and the formula used to derive them (in a table). **Shall not exceed 200 characters.**

Figure 23: Feature generation results for dataset 1

Figure 24: Feature generation results for dataset 2

3 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 500 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 25: Naïve Bayes alternatives comparison for dataset 1

Figure 26: Naïve Bayes alternative comparison for dataset 2

Figure 27: 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 parameterisations. 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 500 characters.**

Figure 28: KNN different parameterisations comparison for dataset 1

Figure 29: KNN different parameterisations comparison for dataset 2

Figure 30: KNN overfitting analysis for dataset 1 (left) and dataset 2 (right)

Figure 31: KNN best model results for dataset 1 (left) and dataset 2 (right)

Decision Trees

Shall be used to present the results achieved through different parameterisations 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 32: Decision Trees different parameterisations comparison for dataset 1

Figure 33: Decision Trees different parameterisations comparison for dataset 2

Figure 34: Decision Trees overfitting analysis for dataset 1 (left) and dataset 2 (right)

Figure 35: Decision trees best model results for dataset 1 (left) and dataset 2 (right)

Figure 36: Best tree for dataset 1

Figure 37: Best tree for dataset 2

Random Forests

Shall be used to present the results achieved through different parameterisations 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 38: Random Forests different parameterisations comparison for dataset 1

Figure 39: Random Forests different parameterisations comparison for dataset 2

Figure 40: Random Forests overfitting analysis for dataset 1 (left) and dataset 2 (right)

Figure 41: Random Forests best model results for dataset 1 (left) and dataset 2 (right)

Figure 42: Random Forests variables importance for dataset 1 (left) and dataset 2 (right)

Gradient Boosting

Shall be used to present the results achieved through different parameterisations 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 43: Gradient boosting different parameterisations comparison for dataset 1

Figure 44: Gradient boosting different parameterisations comparison for dataset 2

Figure 45: Gradient boosting overfitting analysis for dataset 1 (left) and dataset 2 (right)

Figure 46: Gradient boosting best model results for dataset 1 (left) and dataset 2 (right)

Figure 47: 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 parameterisations 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. Shall be used to present the evaluation of the best model achieved. **Shall not exceed 500 characters.**

Figure 48: MLP different parameterisations comparison for dataset 1

Figure 49: MLP different parameterisations comparison for dataset 2

Figure 50: MLP overfitting analysis for dataset 1 (left) and dataset 2 (right)

Figure 51: Loss curve analysis for dataset 1 (left) and dataset 2 (right)

Figure 52: MLP best model results for dataset 1 (left) and dataset 2 (right)

4 CRITICAL ANALYSIS

Shall be used to present a summary of the results achieved with the different modelling 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 ANALYSIS

5 DATA PROFILING

Data Dimensionality and Granularity

May be used to identify the most atomic granularity and two other different granularities to consider. **Shall not exceed 500 characters.**

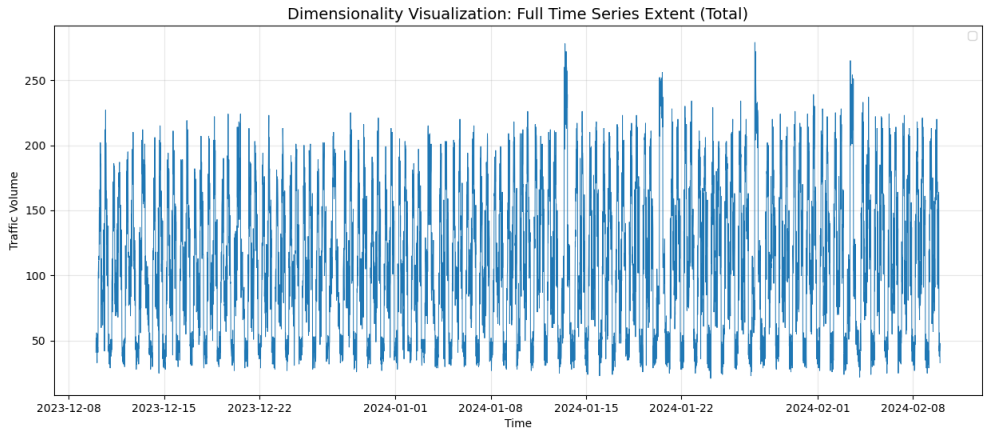


Figure 53: Time series 1 at the most granular detail

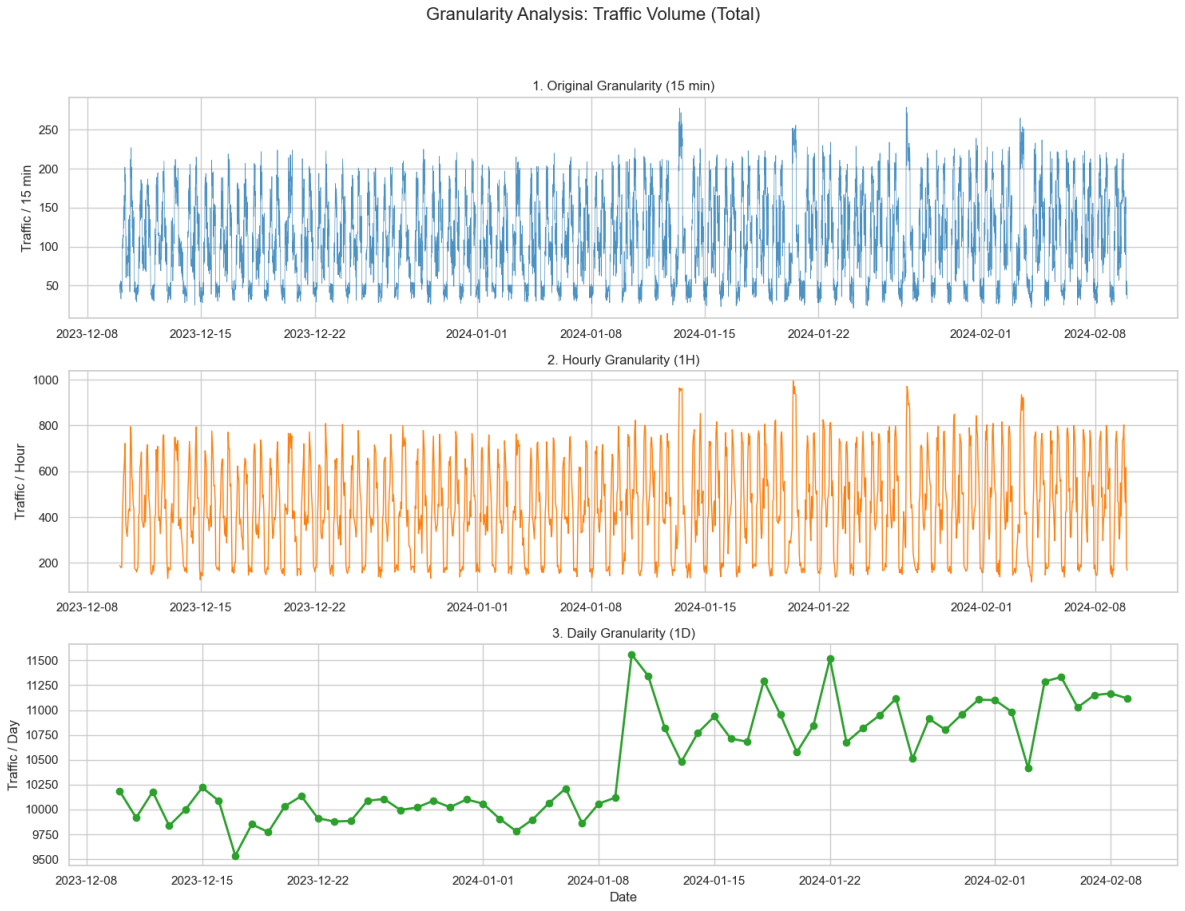


Figure 54: Time series 1 granularities

Data Distribution

Shall be used to perform the data analysis at those three different granularities, concerning the series distribution. **Shall not exceed 500 characters.**

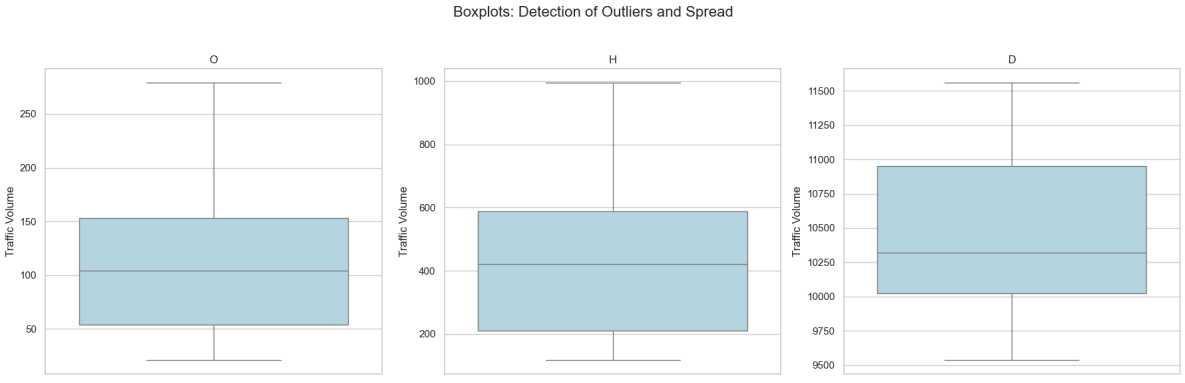


Figure 55: Boxplot(s) for time series 1

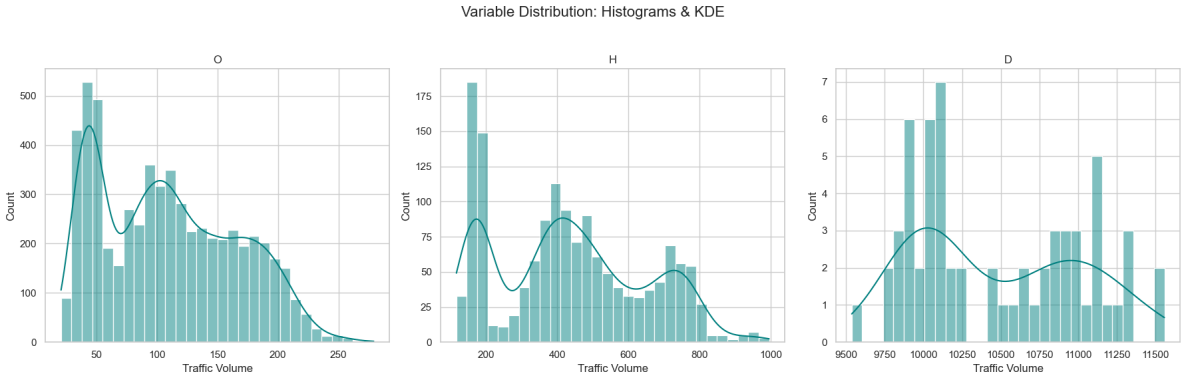


Figure 56: Histogram(s) for time series 1

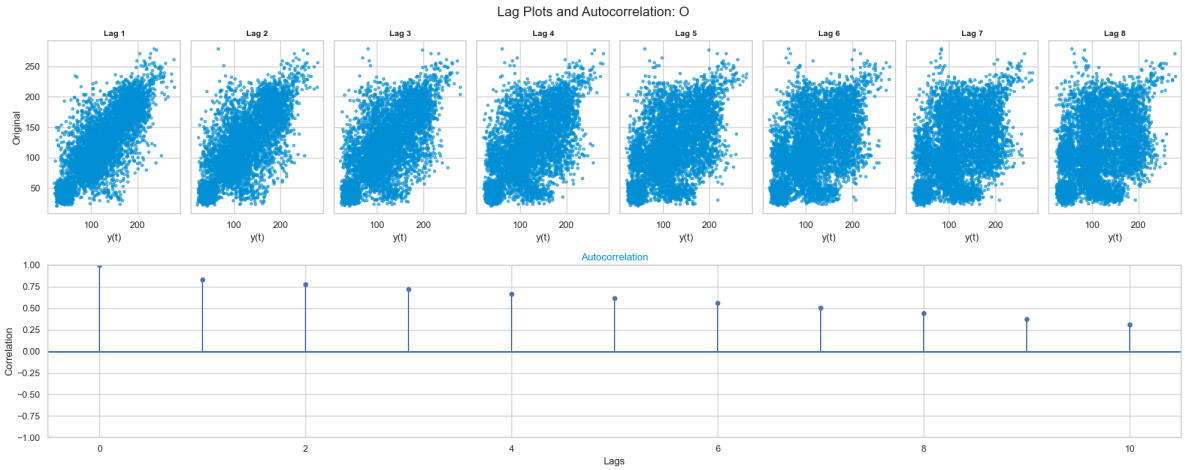
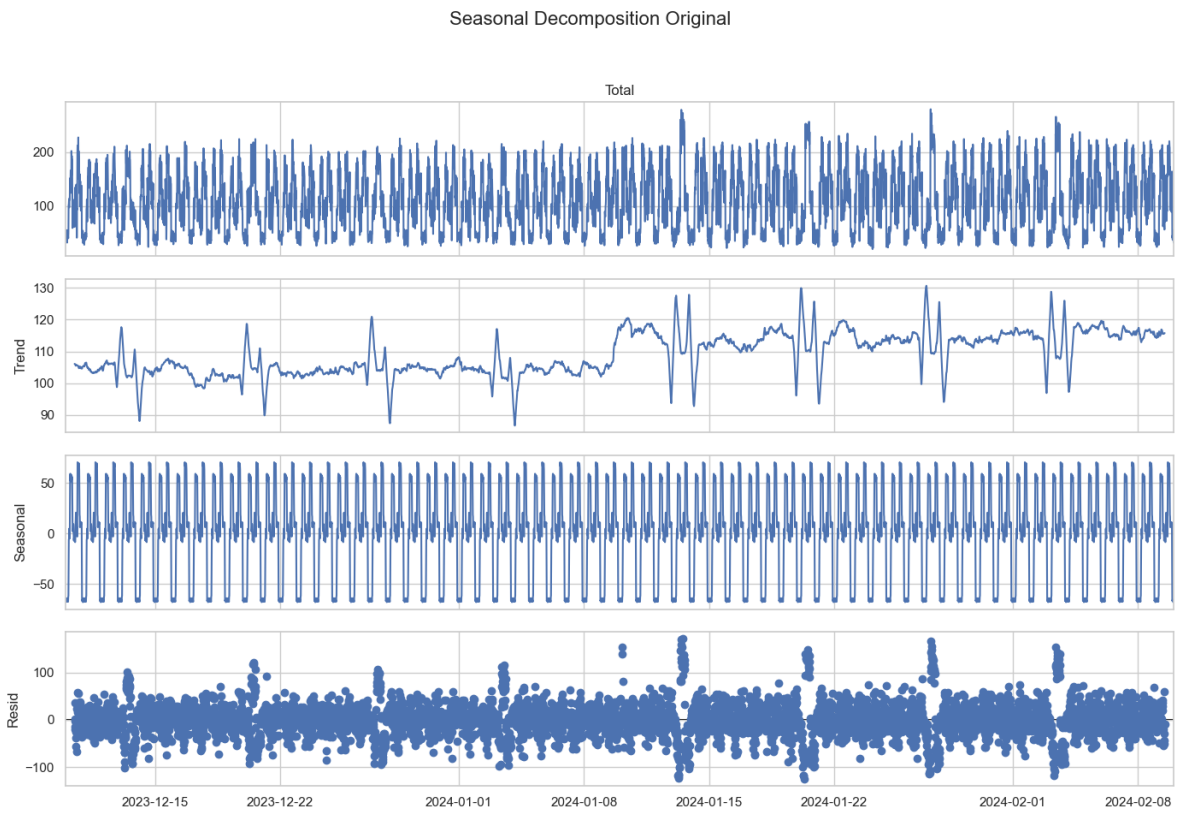


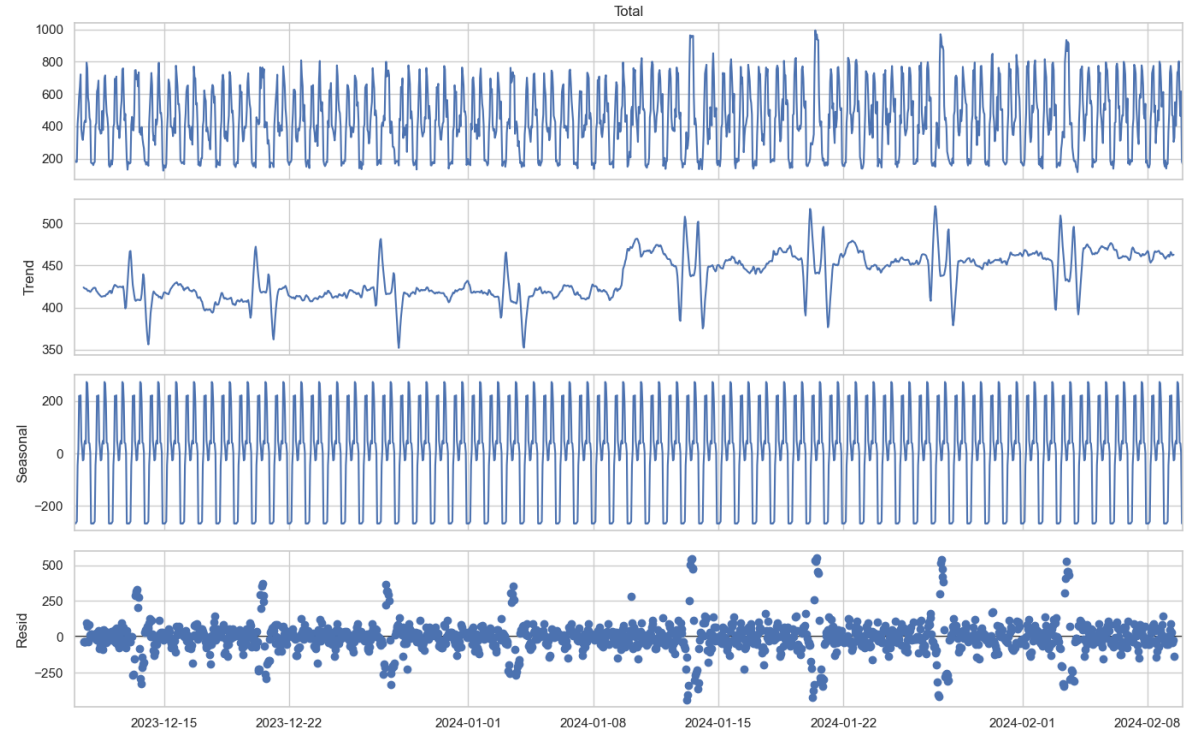
Figure 57: Autocorrelation lag-plots and correlogram for original time series 1

Data Stationarity

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

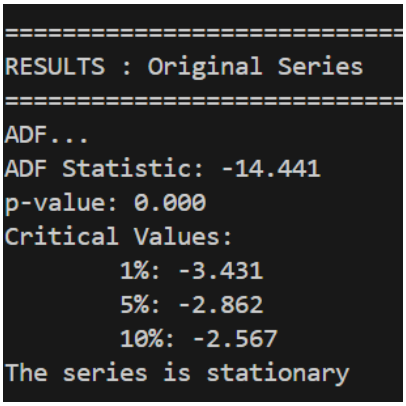


(a) Original

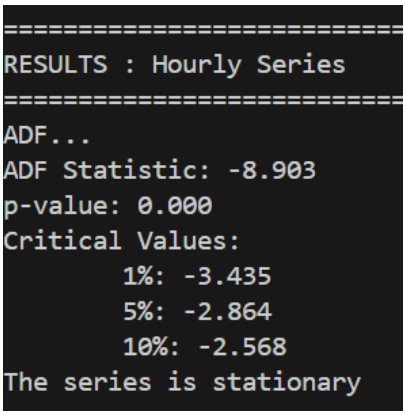


(b) Hourly

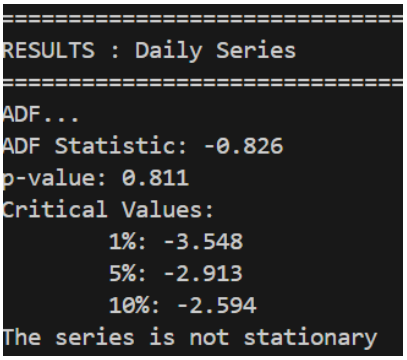
Seasonal Decomposition Daily



(a) Original



(b) Hourly



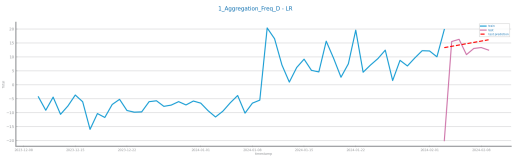
(c) Daily

Figure 59: Stationarity study for time series 1

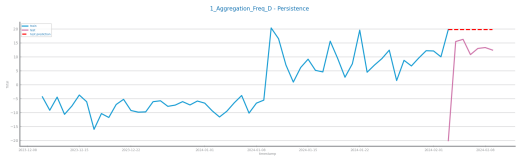
6 DATA TRANSFORMATION

Aggregation

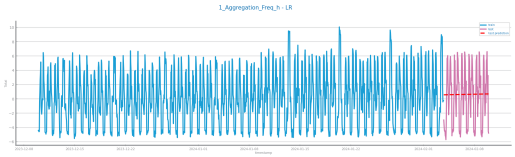
Shall describe the results of applying three different aggregations over both datasets, and identifying the granularity chosen to proceed. **Shall not exceed 300 characters.**



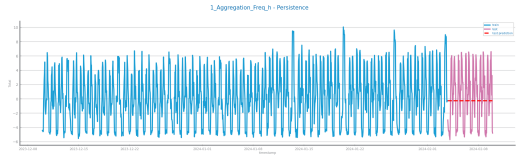
(a) Daily - LR



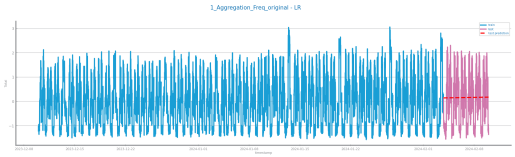
(b) Daily - Persistence



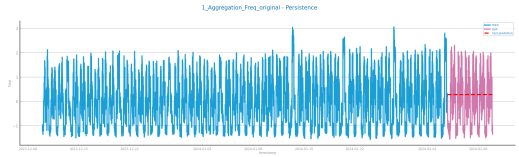
(c) Hourly - LR



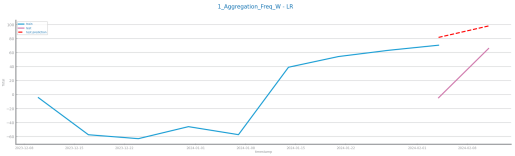
(d) Hourly - Persistence



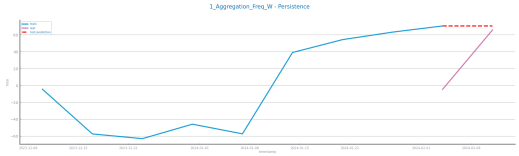
(e) Original - LR



(f) Original - Persistence

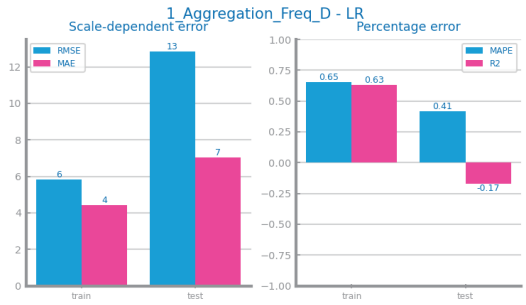


(g) Weekly - LR

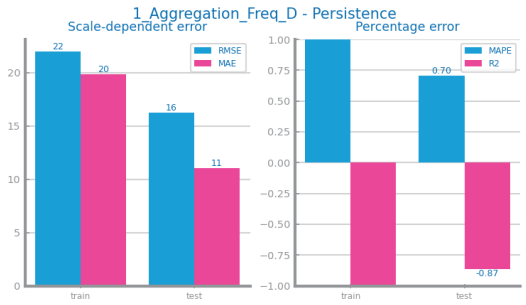


(h) Weekly - Persistence

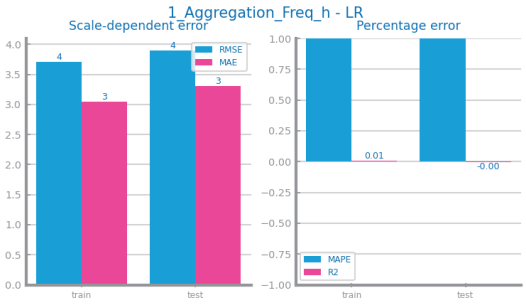
Figure 60: Forecasting plots after different aggregations on time series 1



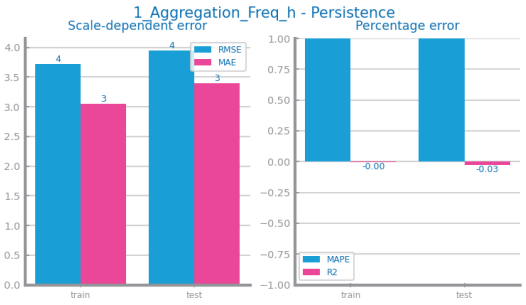
(a) Daily - LR



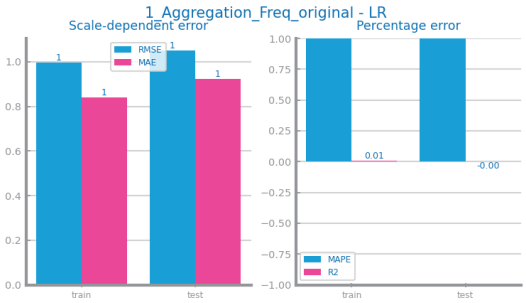
(b) Daily - Persistence



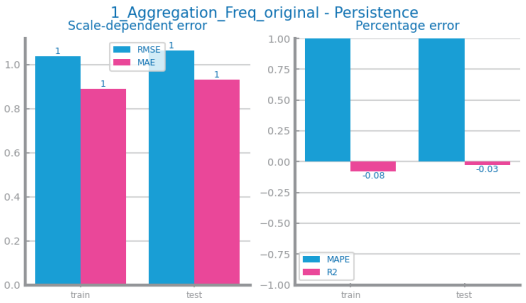
(c) Hourly - LR



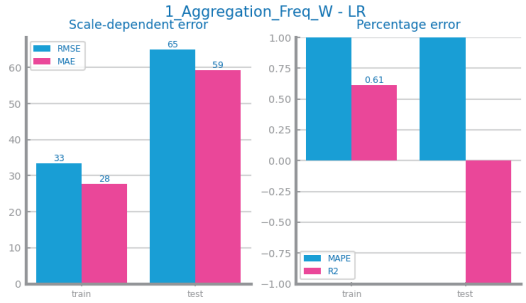
(d) Hourly - Persistence



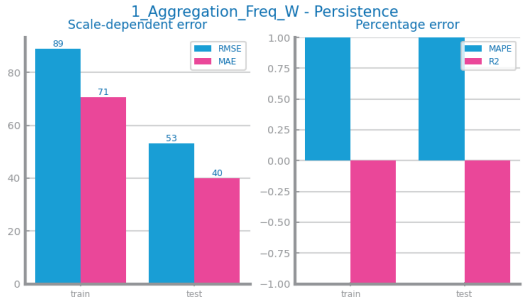
(e) Original - LR



(f) Original - Persistence



(g) Weekly - LR



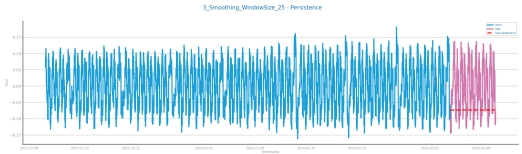
(h) Weekly - Persistence

Figure 61: Forecasting results after different aggregations on time series 1

Smoothing

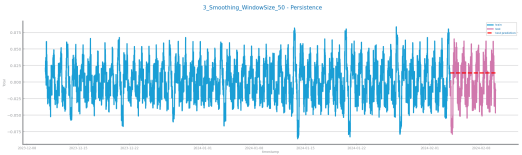
Shall describe the results of applying smoothing transformations over both datasets, and identifying the best result to proceed. **Shall not exceed 300 characters.**

(a) WindowSize 25 - LR



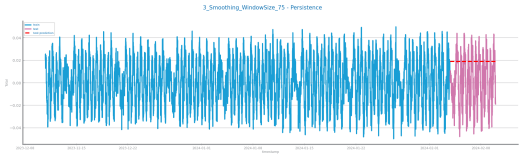
(b) WindowSize 25 - Persistence

(c) WindowSize 50 - LR

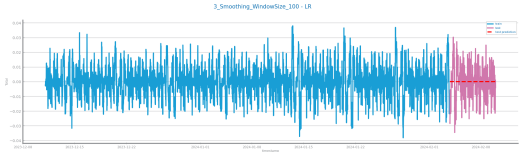


(d) WindowSize 50 - Persistence

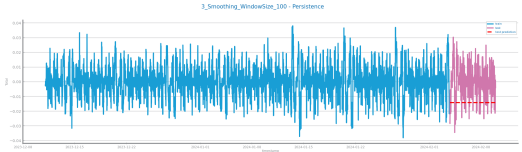
(e) WindowSize 75 - LR



(f) WindowSize 75 - Persistence

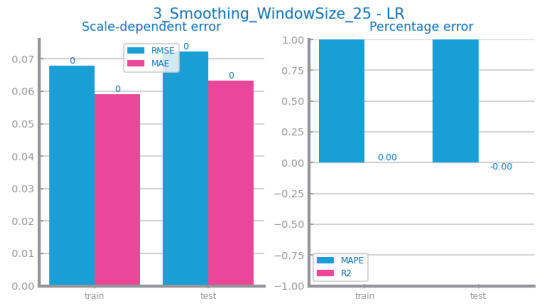


(g) WindowSize 100- LR

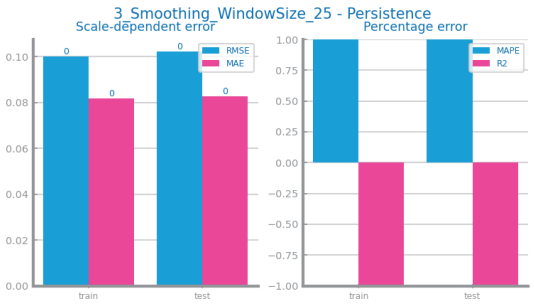


(h) WindowSize 100 - Persistence

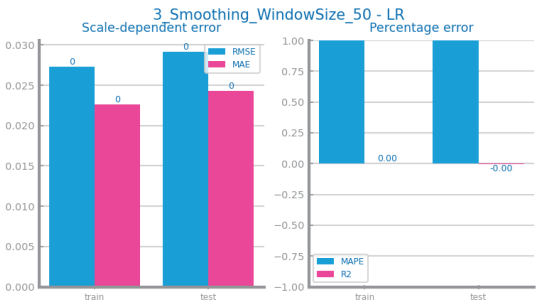
Figure 62: Forecasting plots after different smoothing parameterisations on time series 1



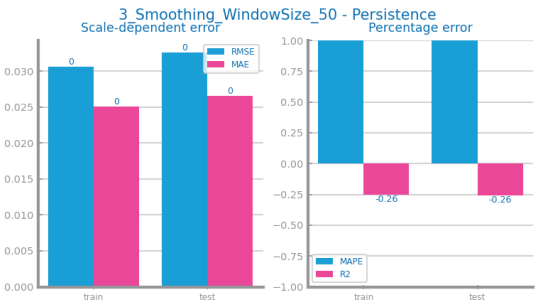
(a) WindowSize 25 - LR



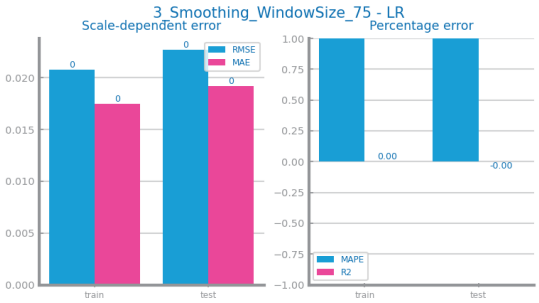
(b) WindowSize 25 - Persistence



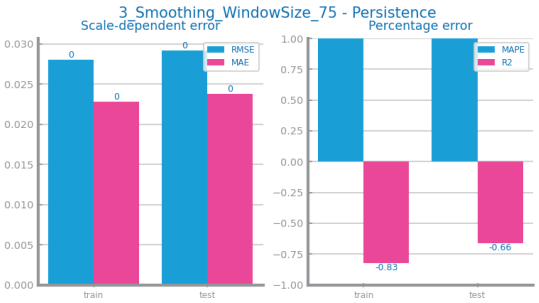
(c) WindowSize 50 - LR



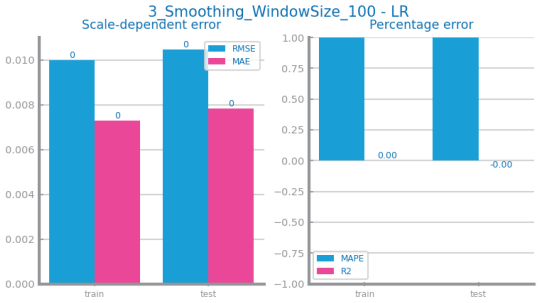
(d) WindowSize 50 - Persistence



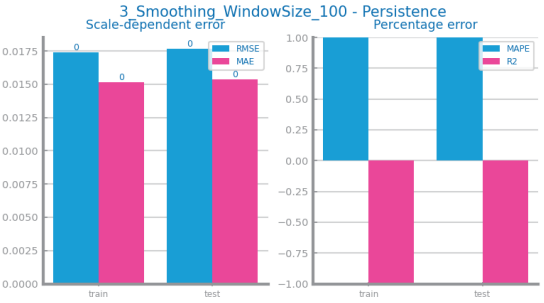
(e) WindowSize 75 - LR



(f) WindowSize 75 - Persistence



(g) WindowSize 100 - LR

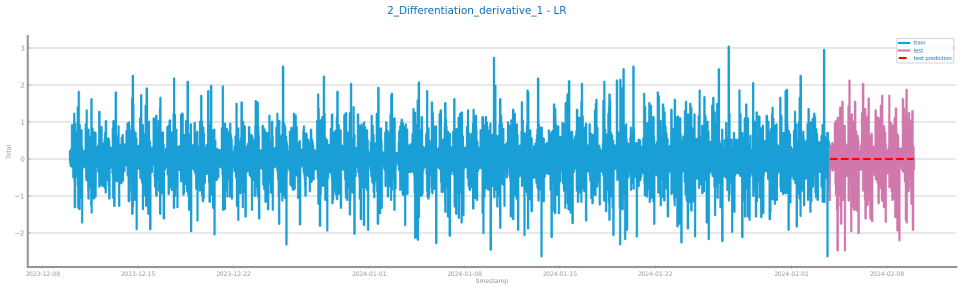


(h) WindowSize 100 - Persistence

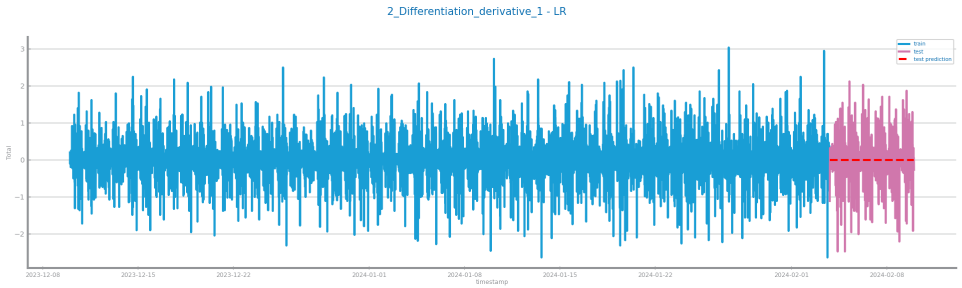
Figure 63: Forecasting results after different smoothing parameterisations on time series 1

Differentiation

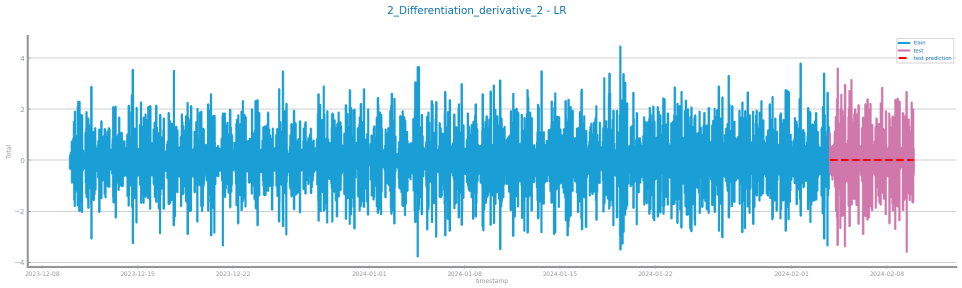
Shall describe the results of applying two consecutive differentiation of both datasets, and identifying the best result to proceed. **Shall not exceed 300 characters.**



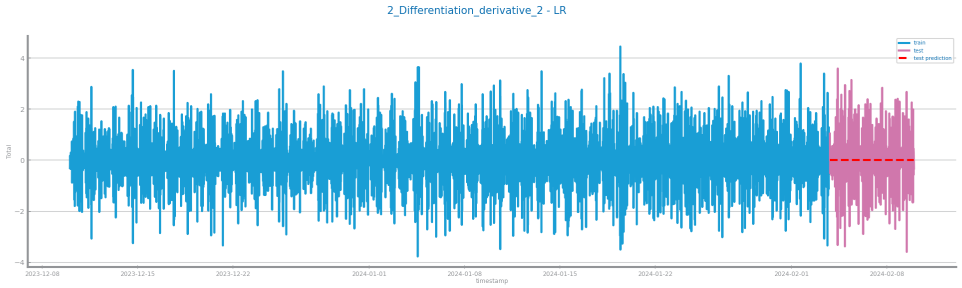
(a) Rist derivative - LR



(b) First derivative - Persistence

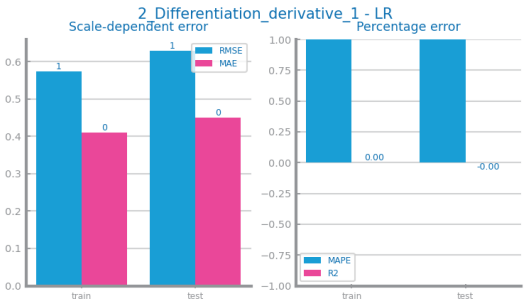


(c) Second derivative - LR

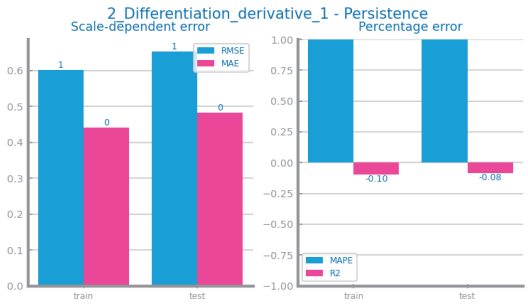


(d) Second derivative - Persistence

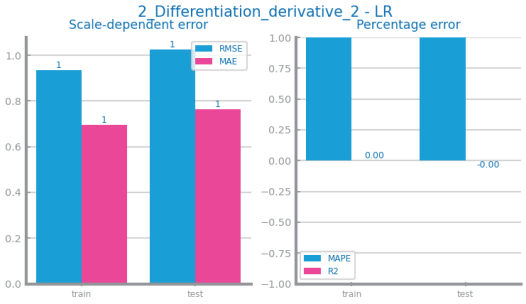
Figure 64: Forecasting plots after first and second differentiation of time series 1



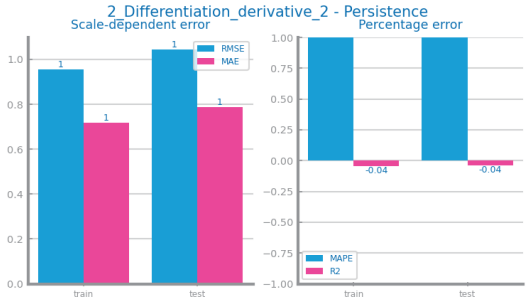
(a) First derivative - LR



(b) First derivative - Persistence



(c) Second derivative - LR



(d) Second derivative - Persistence

Figure 65: Forecasting results after first and second differentiation of time series 1

7 MODELS' EVALUATION

Shall be used to summarise the transformations done over the original time series. **Shall not exceed 500 characters.**

Simple Average Model

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

Figure 66: Forecasting plots obtained with Simple Average model over time series 1

Figure 67: Forecasting results obtained with Simple Average model over time series 1

Persistence Model

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

Figure 68: Forecasting plots obtained with Persistence model (long term) over time series 1

Figure 69: Forecasting plots obtained with Persistence model (one-set-behind) over time series 1

Figure 70: Forecasting results obtained with Persistence model in both situations over time series 1

Rolling Mean Model

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

Figure 71: Forecasting study over different parameterisations of the Rolling Mean algorithm over time series 1

Figure 72: Forecasting plots obtained with the best parameterisation of Rolling Mean algorithm, over time series 1

Figure 73: Forecasting results obtained with the best parameterisation of Rolling Mean algorithm, over time series 1

Exponential Smoothing Model

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

Figure 74: Forecasting study over different parameterisations of the Exponential Smoothing algorithm over time series 1

Figure 75: Forecasting plots obtained with the best parameterisation of Exponential Smoothing algorithm, over time series 1

Figure 76: Forecasting results obtained with the best parameterisation of Exponential Smoothing algorithm, over time series 1

Linear Regression Model

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

Figure 77: Forecasting plots obtained with Linear Regression model over time series 1

Figure 78: Forecasting results obtained with Linear Regression model over time series 1

ARIMA Model

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

Figure 79: Forecasting study over different parameterisations of the ARIMA algorithm over time series 1, only with the target variable

Figure 80: Forecasting plots obtained with the best parameterisation of ARIMA algorithm, over time series 1, only with the target variable

Figure 81: Forecasting results obtained with the best parameterisation of ARIMA algorithm, over time series 1, only with the target variable

Figure 82: Forecasting study over different parameterisations of the ARIMA algorithm with multiple variables over time series 1

Figure 83: Forecasting plots obtained with the best parameterisation of ARIMA algorithm with multiple variables over time series 1

Figure 84: Forecasting results obtained with the best parameterisation of ARIMA algorithm with multiple variables over time series 1

LSTMs Model

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

Figure 85: Forecasting study over different parameterisations of LSTMs over time series 1, only with the target variable

Figure 86: Forecasting plots obtained with the best parameterisation of LSTMs, over time series 1, only with the target variable

Figure 87: Forecasting results obtained with the best parameterisation of LSTMs, over time series 1, only with the target variable

Figure 88: Forecasting study over different parameterisations of LSTMs with multiple variables over time series 1

Figure 89: Forecasting plots obtained with the best parameterisation of LSTMs with multiple variables over time series 1

Figure 90: Forecasting results obtained with the best parameterisation of LSTMs with multiple variables over time series 1

8 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.**