NOVA IMS Information Management School

MDSAA

Master's Degree Program in

Data Science and Advanced Analytics

DeepPAY: A Framework for Dimensionality Reduction and Interactive Exploration

Diana Ferreira da Silva

Project Work

presented as partial requirement for obtaining a Master's Degree in Data Science and Advanced Analytics

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

DeepPAY: A Framework for Dimensionality Reduction and Interactive Exploration

by

Diana Ferreira da Silva

Project Work presented as partial requirement for obtaining the Master's degree in Data Science and Advanced Analytics, with a with a specialization in Data Science

Supervised by

Fernando Bação, PhD, NOVA Information Management School

July, 2025

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism, any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledged the Rules of Conduct and Code of Honor from the NOVA Information Management School.

[Lisbon, July 2025]

Diana Silva

ACKNOWLEDGEMENTS

I would like to thank my advisor, Professor Fernando Bação, and also Professor Farina Pontejos, for their valuable feedback, guidance, and support throughout the development of this project. Their input and advice were crucial throughout the different stages of this work, from defining the approach to its implementation.

A special thanks to Vânia Gramaça Silva and Luís Oliveira, from Banco de Portugal, for their support during the project's important phases, as well as their motivation and constructive input.

In addition, I wish to acknowledge my family and friends. I appreciate their support and commitment to encouraging me throughout, checking in with me regularly. Their presence made everything much more feasible and meaningful.

Lastly, I offer my thanks for everyone who contributed to this project in one way or another, directly and indirectly.

ABSTRACT

As payments become more digitalized and interconnected, the complexity and volume of the data they generate continue to grow, resulting in new challenges for analysis and interpretation. These challenges are critical in contexts where timely insight and transparency are essential. This study introduces a visual analytics framework designed to explore high-dimensional payment data through a combination of dimensionality reduction (PCA and UMAP), clustering techniques, and an autoencoder–isolation forest model for anomaly detection. The solution is implemented as a modular Dash application that supports dynamic interaction, enabling users to uncover latent structures, identify behavioural clusters, and detect anomalous or inconsistent records. Two usage scenarios are explored, using payments data collected by Banco de Portugal: one that examines the clustering patterns and network topology of interbank transactions, and another that focuses on detecting abrupt changes and irregularities in payment series. Together, these use cases demonstrate the system's potential to support both structural and temporal analyses. Although the framework was developed with a focus on payment systems, the approach is sufficiently general to be applied to other domains involving high-dimensional transactional information, such as stock exchange operations or insurance records. By bringing together analytical depth and interpretability, this work contributes to the design of transparent and flexible tools for navigating complex data ecosystems.

KEYWORDS

anomaly detection; clustering; dash application; high-dimensional data; network analysis; payment data

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

BdP	*Banco de Portugal*
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DR	Dimensionality reduction
MSE	Mean Squared Error
PCA	Principal Component Analysis
PSP	Payment service provider
SICOI	*Sistema de Compensação Interbancária*
SQL	Structured Query Language
t-SNE	t-Distributed stochastic neighbor embedding
UMAP	Uniform Manifold Approximation and Projection

1. INTRODUCTION

The increasing digitalization and complexity of payments is leading to a sharp rise in both the volume and dimensionality of payment data. In particular, the evolution of payments, driven by the adoption of real-time settlement mechanisms and the integration of digital platforms, has led to the generation of increasingly large and complex datasets (BdP, 2022). In Portugal, *Sistema de Compensação Interbancária* (SICOI), the Portugal's national system that processes and clears retail payments, processed 4.7 billion operations worth €776.7 billion in 2024. With an average of 11.5 million card transactions per day, card payments, representing 89.5% of all retail transactions processed by SICOI, illustrate the central role of electronic instruments in the payment ecosystem. Instant transfers stand out due to the increase of 46.4% in volume and 47.2% in value compared to the previous year (BdP, 2024).

Payment data are an invaluable source of information for central banks due to their timeliness and granularity. They contribute not only to the estimation of key macroeconomic indicators but also to the modelling of agents' behaviour in response to shocks or policy actions and to promotion of smooth payment system functioning and financial stability. While the developments in payments offer many advantages, they also introduce new analytical challenges. As the volume and complexity of data increase, traditional monitoring tools often fall short in identifying subtle shifts or emerging risks. Analytical approaches must evolve to support timely insight and interpretability in increasingly dynamic environments (BdP, 2022).

While these developments offer many advantages, they also introduce new challenges. Analysing such data has become significantly more demanding, both from a technical and interpretative perspective. As payments grow more complex, traditional analytical approaches often fail to uncover underlying patterns or detect early signals of systemic change (BdP, 2022).

In this context, visual analytics has emerged as a promising paradigm, offering a bridge between computational methods and human reasoning. Yet despite its potential, its application to payment data remains relatively underexplored, particularly in the design of interactive and interpretable tools capable of supporting real-time analysis. Given the strategic role of Banco de Portugal in monitoring payment system stability, the development of tools that can turn granular transactional data into useful insight is especially relevant. This study builds on data collected by Banco de Portugal under PAY, highlighting the importance of such data for supervisory and analytical functions.

Extracting insights from payment data plays a key role in understanding the structure of system participants networks, identifying risk patterns, and uncovering shifting trends within the broader payments ecosystem. Yet the high dimensionality of these datasets often hides the most relevant patterns. This becomes particularly critical in the context of central bank oversight, since payment systems are critical pillars of financial stability. As these systems grow more complex, bringing together a broader set of participants, payment instruments,

and operational dynamics, the need for flexible, scalable tools becomes more evident. Still, many analyses continue to depend on static dashboards and aggregated indicators which, while useful for basic monitoring, are not always capable of revealing irregularities and early behavioural shifts that may signal reporting issues.

This trade-off between data richness and interpretability motivates the development of DeepPAY, an interactive visual analytics framework for the exploration of high-dimensional payment data that uses the state-of-the-art dimensionality reduction techniques.

The goal is not only to apply advanced computational methods, but to make their outputs meaningful and accessible within central banks contexts. In doing so, this work seeks to bridge the gap between technical complexity and practical insight and transforming complex data into interpretable visual representations without oversimplifying or losing analytical depth. Dimensionality reduction methods such as Principal Component Analysis (PCA), t-Distributed Stochastic Neighbour Embedding (t-SNE), and Uniform Manifold Approximation and Projection (UMAP) allowed projecting high-dimensional spaces into two or three dimensions, where visual inspection becomes feasible. However, these methods also introduce distortions, and their results can be sensitive to parameter choices and data preprocessing strategies. Furthermore, the interpretability of the reduced space is not always straightforward, particularly for end-users unfamiliar with the mathematical principles behind the algorithms. For these reasons, it is not sufficient to merely apply such techniques, they must be embedded within an interactive and intelligible environment that supports exploration, hypothesis generation, and critical scrutiny.

This project proposes the design and implementation of an interactive framework for visual analytics, specifically for high-dimensional payment data. The proposed solution integrates dimensionality reduction techniques, such as PCA and UMAP, clustering, anomaly detection, and data shift monitoring into a modular Dash-based interface. Unlike traditional static tools, DeepPAY supports dynamic exploration of high-dimensional payment data, allowing users to interact with latent structures, track behavioural changes, and identify potential inconsistencies. Instead of providing fixed outputs, the system is designed to be an exploratory environment, where analysts can test hypotheses, follow patterns, spot unusual cases, and identify irregularities or unexpected signals in the data. It supports key analytical capabilities, including anomaly scoring, temporal drift tracking, cluster profiling, and network-based flow visualization.

This research contributes to the existing literature in several ways. It helps clarify how dimensionality reduction techniques can be used to explore complex payment data in a way that remains interpretable and useful for analysis. It also delivers a flexible and modular framework, capable of handling large transactional datasets, simplifying them, and presenting their structure through interactive visualizations. On a practical level, it includes a working prototype that demonstrates how these tools can support tasks such as detecting anomalies or identifying behavioural patterns. Although the work focuses on payment data, the

approach is broad enough to be adapted to other context involving high-dimensional data. The architecture of DeepPAY is general enough to be applied in other sectors characterized by complex transactional patterns, such as stock exchange, financial, fraud, insurance, telecommunications, or cyber risk sectors.

This investigation begins with a review of the relevant literature and conceptual foundations that support the integration of visual analytics and payment data monitoring (Section 2). Section 3 describes the methodological framework of the system, including data collection, preprocessing, and the development of analytical modules for anomaly detection, clustering, and time series variation. Section 4 details the design of the interactive visualization layer, describing the interface elements and interaction models developed to explore high-dimensional payment data. Section 5 presents the application of the system to real-world use cases using payment data collected by Banco de Portugal, illustrating the framework's utility in cluster/network exploration and time series quality monitoring. The project concludes with a synthesis of the main findings, current limitations, and suggestions for future research directions (Section 6).

2. LITERATURE REVIEW

The big data revolution has changed the way information is generated, stored, and analysed, leading to more complex and high-dimensional datasets. Large or high-dimensional data is characterized by many attributes or features. As the number of features or attributes increases, the dimensionality of the dataset also increases (Hasugian et al., 2023).

Such data represents substantial challenges for analysis and visualization because the high dimensionality leads to sparsity, where data points become increasingly dispersed, making it difficult to identify meaningful patterns. Additionally, the distances between points lose their interpretability, which complicates tasks like clustering and classification (van der Maaten & Hinton, 2008).

This is also applicable to the payment ecosystem, where similar challenges arise. The ongoing digitization of the economy further amplifies both the volume and richness of payment data, enabling more granular insights into consumption behaviour, financial intermediation, and systemic resilience. Payment transaction data is inherently complex, characterized by high dimensionality, temporal patterns, and considerable variability across users and systems (León, 2020). Common features include transaction amounts, timestamps, merchant categories, geolocations, and user demographics (Ardizzi et al., 2024).

Despite these complexities, payment data has become an asset for economic analysis, forecasting, and policy evaluation. As highlighted by Ardizzi et al. (2024)understanding consumer adoption and usage patterns of payment technologies are essential for monitoring economic activity and ensuring safe, efficient, and reliable payment systems, particularly in the context of innovations such as instant payments. In addition, payment data is frequently used for other applications such as fraud detection, customer segmentation, and credit risk assessment, while also offering insights into evolving trends within the payment ecosystem, ensuring regulatory compliance, mitigating systemic risks, and protecting consumers (Putrevu & Mertzanis, 2023).

2.1 HIGH DIMENSIONAL DATA ISSUES

The high-dimensional nature of data, such as payment data, presents multiple challenges to traditional analytical methods, including sparsity, noise, overplotting, computational complexity, and reduced interpretability (Hasugian et al., 2023). These challenges are part of a broader phenomenon known as the “curse of dimensionality”, first introduced by Bellman (2015).

The curse of dimensionality describes how, as the number of dimensions increases, data points become equidistant, reducing the effectiveness of traditional distance-based methods used for clustering, classification, and regression. As a result, it becomes more difficult to

identify meaningful patterns or correlations, which ultimately complicates both data analysis and visualization (Hasugian et al., 2023).

High-dimensional datasets are often sparse, meaning that most data points carry limited significance or contain missing values. This sparsity complicates feature selection and reduces the effectiveness of traditional analytical techniques (Donoho, 2000). Additionally, overplotting occurs when multiple data points overlap in visualizations, obscuring patterns and reducing interpretability (Hasugian et al., 2023). This is problematic in exploratory data analysis and visualization tasks.

Algorithms designed for data analysis and visualization often scale poorly with the number of dimensions. As the dimensionality increases, the computational resources required grow exponentially, limiting the practicality of traditional methods (Aggarwal, 2002). Furthermore, in high-dimensional spaces, understanding the relationships between variables becomes increasingly difficult. This lack of interpretability poses challenges for decision-making processes and the design of intuitive visualizations.

Another critical issue is overfitting, where models capture noise instead of meaningful patterns. Overfitting reduces generalization capabilities and the reliability of predictive models (Kelleher et al., 2015). This problem becomes particularly pronounced as the number of features relative to observations grows.

These challenges highlight the necessity of dimensionality reduction techniques that can effectively preserve the underlying structure of the data while mitigating the negative effects of high dimensionality.

2.2 DIMENSIONALITY REDUCTION TECHNIQUES

Dimensionality reduction (DR) techniques are crucial for reducing the complexity of high-dimensional datasets while maintaining as much of the intrinsic structure of the data as possible. These methods are often employed to visualize, analyse, and process data that would otherwise be too large or difficult to interpret. By reducing the number of dimensions while preserving the underlying structure, DR techniques facilitate more efficient analysis and visualization, enabling the extraction of meaningful patterns and relationships from complex datasets (Hasugian et al., 2023).

Dimensionality reduction (DR) methods can be broadly categorized into linear and non-linear approaches. Among the methods employed are Principal Component Analysis (Jolliffe, 2002), t-Distributed Stochastic Neighbour Embedding (van der Maaten & Hinton, 2008), and Uniform Manifold Approximation and Projection (McInnes & Healy, 2018). Each of these techniques offers multiple advantages and presents specific challenges.

2.2.1 LINEAR METHODS

Principal Component Analysis (PCA) is a linear dimensionality reduction technique that transforms the original variables into a new set of uncorrelated variables called principal components. These components are ordered by the amount of variance they capture from the data, ensuring that the first principal component retains the maximum variance, the second component captures the maximum variance orthogonal to the first, and so on. This transformation reorients the data into a new coordinate system, effectively reducing dimensionality while preserving variance. (Hasugian et al., 2023; Jolliffe, 2002).

PCA identifies orthogonal directions in the data that maximize variance, making it computationally efficient and widely used in various applications. However, PCA assumes linear relationships among variables, which may limit its ability to preserve complex structures in high-dimensional or nonlinear datasets (Hasugian et al., 2023), like payment data.

PCA is widely used due to its computational efficiency, especially when compared with non-linear techniques such as t-SNE and UMAP. It is particularly useful in applications requiring noise reduction or pattern detection, as it focuses on components with the highest variance and discards those that contribute less. Because it is a linear method, the resulting components remain interpretable in terms of the original feature space, which adds to its transparency and ease of use.

However, PCA comes with limitations. Its assumption of linearity may restrict its ability to capture more complex, non-linear interactions, something commonly observed in high-dimensional and structured datasets (Salmanian et al., 2024), like payment data. Moreover, PCA is sensitive to the scaling of features, meaning that variables with larger ranges can dominate the principal components unless proper standardization is applied beforehand. Another limitation lies in its handling of categorical variables. Since PCA relies on Euclidean distances and variance, it does not naturally support categorical data. While encoding techniques, such as one-hot encoding, can be applied, they often lead to increased sparsity and dimensionality, which may dilute meaningful structures or introduce noise into the analysis (Hasugian et al., 2023).

Other linear methods for dimensionality reduction (DR) exist, which will not be discussed in this literature review.

2.2.2 NON-LINEAR METHODS

t-Distributed Stochastic Neighbor Embedding (t-SNE) is a nonlinear dimensionality reduction technique designed primarily for visualizing high-dimensional data. Unlike linear methods such as PCA, t-SNE focuses on preserving the local structure of the data, making it particularly effective in revealing clusters and hidden patterns that may remain undetected with linear methods (van der Maaten & Hinton, 2008).

The algorithm works by converting similarities between data points into joint probabilities and then minimizing the divergence between these probabilities in the original and reduced space. As a result, similar data points are mapped close together in a lower-dimensional space, typically two or three dimensions, while dissimilar points are mapped further apart. This makes t-SNE particularly effective for exploratory data analysis, as it can capture non-linear relationships between data points and reveal intricate structures such as outliers and clusters or manifolds in complex datasets (Hasugian et al., 2023).

Nonetheless, it is computationally expensive, making it less suitable for very large datasets, and is highly sensitive to hyperparameters such as perplexity and learning rate, which can significantly influence the output structure.

Despite its strengths, t-SNE also poses several challenges. The method is computationally intensive and scales poorly with larger datasets, as it relies on pairwise distance computations and iterative optimization. Furthermore, t-SNE often distorts global relationships between clusters, meaning that the distance between cluster may not be meaningful or difficult to interpret. It is also highly sensitive to parameter choices, particularly perplexity and learning rate, which can significantly affect the resulting embedding. Additionally, its stochastic nature means that different runs may yield different visualizations unless a fixed random seed is used (Hasugian et al., 2023).

When it comes to categorical data, t-SNE requires prior encoding, such as one-hot encoding, which often results in sparse matrices. This sparsity can reduce the algorithm's effectiveness, particularly in datasets with high-cardinality categorical features. Additionally, t-SNE's tendency to preserve local relationships can lead to misleading visualizations when categorical variables have significant global structure.

Alternatively, UMAP is a nonlinear dimensionality reduction technique that constructs a weighted graph of the high-dimensional data and then optimizes a low-dimensional projection that preserves both local and global data structure, enabling visualization and clustering (Hasugian et al., 2023; McInnes & Healy, 2018).

Compared to t-SNE, UMAP is generally faster, more scalable, and more consistent across runs, making it well-suited for large datasets and real-time interactive applications while allowing reproducibility. It also achieves a better balance between preserving local neighbourhoods and maintaining the global structure of the data. This makes UMAP valuable in contexts like payment systems, where both local anomalies and broader behavioural clusters can provide valuable insights for analysis (Hasugian et al., 2023).

However, UMAP is not without limitations. Its mathematical formulation is more complex than that of PCA or t-SNE, which may limit its accessibility to non-specialist users. Moreover, like t-SNE, its performance depends on the tuning of hyperparameters, such as the number of neighbours and minimum distance, which control how local or global the resulting embedding will be (Hasugian et al., 2023).

In terms of handling categorical data, UMAP performs relatively well when paired with encoding techniques. It is considered more robust than PCA and t-SNE in dealing with sparse or mixed-type datasets and is often the preferred choice when interpretability and scalability are both required.

2.2.3 Comparative Analysis

The choice of a dimensionality reduction technique depends largely on the characteristics of the dataset and the specific objectives of the analysis. Comparative studies highlight that PCA, t-SNE, and UMAP each serve distinct purposes. For exploratory visualization, t-SNE and UMAP are generally preferred due to their ability to capture non-linear relationships and reveal latent structures within high-dimensional data. In contrast, PCA is more effective in reducing noise and feature redundancy, given its computational efficiency and ability to retain global variance structures (Hasugian et al., 2023).

Among non-linear techniques, UMAP often outperforms t-SNE in terms of scalability, computational efficiency, and preservation of meaningful structures, making it particularly suited for large datasets (Hasugian et al., 2023). UMAP also balances local and global structure preservation, whereas t-SNE focuses primarily on local structures. This difference makes UMAP more effective for clustering tasks.

However, despite their strengths, t-SNE and UMAP introduce interpretability challenges. These methods transform the data into lower-dimensional representations without maintaining a direct, mathematically explainable relationship with the original feature space (Salmanian et al., 2024).

For datasets containing categorical variables, UMAP is generally the most suitable method, as it efficiently handles sparse matrices and mixed data types. In contrast, PCA requires encoding and normalization of categorical variables, as its reliance on variance-based projections is not necessarily compatible with categorical data. T-SNE, on the other hand, often struggles with categorical variables due to the sparsity introduced by one-hot encoding, which can distort neighbourhood relationships.

2.3 Data Quality and Preprocessing for Dimensionality Reduction

The quality of high-dimensional data plays an important role in the effectiveness of dimensionality reduction techniques. Raw datasets often contain errors, missing values, and inconsistencies that can distort patterns and relationships and affect the accuracy of visualization and analysis. Preprocessing is thus essential to ensure data integrity, efficiency, and interpretability before modelling.

One of the primary preprocessing tasks is data cleaning, which involves identifying and correcting missing values, duplicate records, and inconsistencies. Common approaches include imputing missing values using statistical methods, normalizing numerical attributes to maintain consistency, and detecting outliers through robust anomaly detection techniques (Milani et al., 2021).

Beyond cleaning, feature selection and engineering are critical in optimizing dimensionality reduction performance. High-dimensional datasets often contain redundant or irrelevant features that can increase computational complexity and obscure important patterns. Correlation analysis, mutual information, and variance thresholding are commonly employed to identify and retain the most informative attributes (Liu et al., 2010).

Scalability is another major challenge in preprocessing. Traditional preprocessing pipelines struggle with high-throughput transactional data, necessitating the adoption of parallelized and distributed processing frameworks, such as Apache Spark. These systems enable batch and real-time data processing, ensuring that dimensionality reduction remains feasible for datasets containing millions of transactions (Hasugian et al., 2023; Karimov et al., 2018).

The ability to detect outliers also plays a key role in improving data quality, as outliers often signal rare events or hidden structures in high-dimensional datasets. Identifying these anomalies is important not only for cleaning purposes but also for uncovering patterns that would otherwise remain obscured by noise or scale. Traditional anomaly detection techniques such as Density-Based Spatial Clustering of Applications with Noise (DBSCAN) and Isolation Forest presents some strengths and limitations. DBSCAN, for instance, excels at identifying clusters of varying shapes and isolating noise points but can be sensitive to parameter settings and struggles with varying densities across dimensions (Ester et al., 1996). Additionally, Chandola et al. (2009) highlight that DBSCAN is not specifically optimized for anomaly detection, as its primary objective is clustering rather than identifying rare or isolated instances. Isolation Forest, by contrast, is better suited for high-dimensional data, requiring less tuning and offering linear time complexity. It isolates anomalies by recursively partitioning the feature space using randomly selected attributes, resulting in shorter path lengths for outliers (Yang et al., 2022).

However, as dimensionality increases, the effectiveness of these algorithms tends to reduce. This is particularly problematic in complex domains where meaningful patterns are often difficult to identify due to redundancy and data sparsity. In this context, hybrid models have emerged as promising alternatives. Among them, AE-Iforest, a method that combines deep autoencoders with Isolation Forest, has shown superior performance across multiple benchmarks. The autoencoder first learns a compact, low-dimensional representation of the data by minimizing reconstruction error. This encoding not only compresses but also denoises the input. Then, Isolation Forest Is applied in this reduced space to compute anomaly scores. By integrating reconstruction loss with path length, AE-Iforest improves sensitivity to subtle anomalies while mitigating the curse of dimensionality (Yang et al., 2022).

The interpretability of such hybrid models can also be improved by examining the latent features learned by the autoencoder and correlating them with original variables. While techniques like t-SNE and UMAP offer compelling visualizations of these latent spaces, models like AE-Iforest go one step further by scoring each data point not just based on geometry, but also on how well the system can reconstruct it (Yang et al., 2022).

2.4 VISUALIZATION OF HIGH-DIMENSIONAL DATA

Data visualization is essential for exploring and interpreting high-dimensional datasets. In its early stages, visualization primarily relied on basic graphical representations, such as bar charts, pie charts, and line graphs (Hasugian et al., 2023). However, as data complexity increased, these traditional methods struggled with overplotting, information loss, and limited scalability, making it difficult to represent meaningful patterns in large-scale datasets (Krause et al., 2016).

One of the main challenges of visualizing high-dimensional data is representing multiple attributes simultaneously. Since human cognition is limited in its ability to interpret more than three dimensions at a time, selecting, filtering, and transforming the data before visualization becomes essential. Common approaches for high-dimensional data visualization include scatter plots, parallel coordinates, and heatmaps, which have been widely used to represent complex structures (Hasugian et al., 2023). However, these techniques do not scale efficiently beyond 10 to 20 dimensions (Krause et al., 2016), necessitating more sophisticated dimensionality reduction techniques such as PCA, t-SNE, and UMAP.

Recent research in interactive visualization frameworks has been addressing scalability and interpretability challenges by integrating dimensionality reduction with real-time interaction mechanisms. These approaches allow analysts to dynamically adjust parameters, filter data subsets, and uncover latent patterns (Salmanian et al., 2024).

Several frameworks, such as PrAVA, SeekAView and DimVis, have emerged to support these goals, each addressing a different set of challenges.

PrAVA, proposed by Milani et al. (2021), focuses on integrating data quality assessment directly into the visualization pipeline. Instead of applying PCA, t-SNE, or UMAP to raw data, PrAVA first performs data profiling to identify issues such as missing values, noise, and redundant features. These assessments are then used to guide preprocessing decisions within the same analytical environment. This connection between cleaning and visualization guarantees that results remain interpretable and scalable, even in domains with millions of records.

SeekAView, introduced by Krause et al. (2016), tackles the problem of transparency in non-linear dimensionality reduction. Because algorithms like t-SNE and UMAP transform data into abstract representations, analysts often struggle to understand how those projections relate

to the original features. SeekAView allows users to interactively modify reduction parameters and view changes in real time, enabling a trial-and-error process that helps identify stable, meaningful visualizations. This is particularly useful when selecting parameter configurations that preserve relevant data structures.

To deepen the interpretability of DR outputs, DimVis integrates dimensionality reduction with explainable machine learning. It uses Explainable Boosting Machines (EBMs) to model the composition of clusters in UMAP embeddings. Analysts can select a cluster and immediately view which features most strongly contribute to its formation. This design bridges the gap between visual patterns and feature-level understanding, something increasingly important in fields where explainability and accountability are critical, including finance, cybersecurity, and healthcare (Salmanian et al., 2024).

While these frameworks improve post-hoc analysis, a growing demand exists for real-time exploration and alerting mechanisms, particularly in the context of financial systems where the timeliness of insight is paramount. High-frequency, high-dimensional transaction data poses some challenges for streaming pipelines and interactive dashboards. It is not enough to visualize static snapshots, systems must ingest, process, and present data in real time to enable immediate detection of abnormal behaviours.

In this regard, tools like Apache Kafka and Spark Streaming provide the infrastructure needed for continuous data ingestion and low-latency computation. When integrated with interactive visualization systems, these tools make it possible to generate live dashboards that update as new data arrives, and issue alerts when anomaly scores exceed defined thresholds (Karimov et al., 2018).

Despite these advances, real-time visualization systems still face significant challenges. At the same time, continuous data accumulation introduces further demands on data quality assurance. Ensuring the reliability of insights extracted in real time requires a robust data management strategy, which includes scalable infrastructure, efficient preprocessing pipelines, and rigorous validation practices (Hasugian et al., 2023). Moreover, it is still difficult to fully integrate real-time anomaly detection models into these pipelines (Yang et al., 2022).

2.5 Anomaly Detection in Payment Systems

The increasing complexity of payments has promoted interest in advanced analytics capable of detecting atypical patterns in payments behaviour. Given their high frequency and dimensionality, payments produce complex temporal and topological structures, making anomaly detection both essential and technically challenging. In this context, detecting deviations in such environments is essential for maintaining stability, identifying operational failures, and flagging suspicious activities (León, 2020).

Dimensionality reduction methods such as PCA and neural autoencoders have been employed to denoise and compress payment data into lower-dimensional representations. For instance, Triepels et al. (2017) show how such embeddings can distinguish between normal and anomalous subspaces based on reconstruction error. Similarly, Arévalo et al. (2022) combined PCA with clustering (e.g., k-means and DBSCAN) to detect outlier clusters in El Salvador's real-time gross settlement system.

Clustering plays a central role in these applications. DBSCAN, for example, excels at detecting noise points and irregular cluster shapes but may struggle with varying density and high-dimensional noise. K-means, while more scalable, often assumes spherical clusters and requires a predefined number of clusters. Both approaches can benefit from preprocessing with PCA or neural encoding to reduce sparsity and highlight meaningful signals (Arévalo et al., 2022).

Another approach is the integration of anomaly detection with interpretability tools. For example, Random Forests have been used to rank the importance of features within detected clusters, offering insights into the variables driving anomalous behaviour (Arévalo et al., 2022). More recently, hybrid models like AE-Iforest combine autoencoders with Isolation Forest to detect subtle outliers in compressed latent space. This architecture is more robust to the curse of dimensionality and supports visual explanations using tools like UMAP. (León, 2020; Triepels et al., 2017).

In addition to these methods, recent studies have approached payment data as dynamic network structures, where nodes represent institutions and directed edges encode transaction flows, forming time-evolving networks. This abstraction allows the application of network-based clustering and dimensionality reduction to reveal latent structures (e.g., a compressed representation of the original feature space) and detect anomalies (Arévalo et al., 2022; León, 2020; Soramäki & Cook, 2013).

Degree centrality, for example, provides a simple and informative indicator of node connectivity. In a directed network, in-degree reflects the number of institutions from which a participant receives payments, while out-degree captures the number of institutions to which payments are sent. These measures are particularly useful in understanding an entity's position in the system: high out-degree may indicate a liquidity provider or payment originator, whereas high in-degree may signal aggregation or settlement roles (Berndsen & Heijmans, 2020).

The use of network-based indicators has gained prominence, particularly in supervisory context. Glowka et al. (2025) propose a methodology for identifying critical participants in large-value payment systems using a combination of traffic-based indicators and network centrality metrics, such as eigenvector centrality and degree centrality. Through hierarchical clustering, their approach generates participant groupings that reflect both transactional volume and systemic interconnectedness, supporting threshold calibration without relying on

prior labels or simulations. This methodology provides a valuable contribution by integrating graph-based metrics to identify structurally important or at-risk institutions within payment systems. The use of dendrograms increases the interpretability of clustering results, offering a greater degree of visual transparency, particularly relevant when analytical findings need to be communicated to supervisory authorities or operational stakeholders. Additionally, the hierarchical clustering approach enables flexible thresholding and adaptable granularity, making it especially effective in capturing the structural heterogeneity characteristic of complex payment networks.

While these studies demonstrate the value of combining clustering and dimensionality reduction, their operational deployment requires robust preprocessing pipelines, adaptable anomaly thresholds, and interactive visualization tools that allow users to interact with latent data structures.

This research aims to bridge these gaps by developing a scalable and interactive visual analytics framework (DeepPAY) tailored to high-dimensional payment data. By incorporating dimensionality reduction, clustering, anomaly detection and quality checks, the framework seeks to improve interpretability of analyses within complex payment ecosystems.

2.6 TIME SERIES QUALITY PROFILING

A time series is defined as an ordered sequence of data points measured over time, where each point describes the evolution of a given quantity or process based on a continuous or periodic measure (Schmidl et al., 2022).

The task of profiling time series for evaluating quality and structural changes is increasingly central. Schmidl et al. (2022) provide a comprehensive evaluation of 71 state-of-the-art anomaly detection algorithms across 976 datasets. The authors highlight the diversity of time series anomaly types, point anomalies, subsequence anomalies, and correlation anomalies in multivariate settings, and conclude that no single detection method is universally superior. Additionally, the authors observe that simple statistical methods often achieve similar performance levels to more complex deep learning approaches, despite the latter's significantly higher computational cost.

Parallel to the anomaly detection literature, the phenomenon of concept drift, where the statistical properties of a time series evolve over time, is widely studied in the field of data stream mining. As reviewed by Agrahari and Singh (2022), drift can manifest in various forms: abrupt or gradual, recurring or non-repeating, real or virtual. Detection mechanisms typically rely on sliding windows and distributional tests to capture these transitions. Their review underscores the value of drift-aware models in domains where dynamic shifts can signal either changes in behaviour or breakdowns in the reporting process.

These studies suggest that a multi-indicator approach is better suited for practical applications, particularly when the objective is not automated classification, but rather interactive exploration and human-centred quality monitoring. By integrating drift-aware metrics, flatline detection, and z-score diagnostics alongside simpler completeness indicators, the proposed system improves anomaly detection while preserving interpretability. Presenting these indicators in a user-friendly format allows users to quickly identify structurally irregular or behaviourally suspicious series.

3 METHODOLOGY

This section presents the methodological approach adopted for the development of a modular and interactive application for the visual exploration of high-dimensional payment data. The system is implemented using the Dash framework and follows a multi-page architecture, structured to promote scalability, maintainability, and clear separation of analytical concerns. The architecture integrates both backend computations and frontend interactivity within a single unified environment.

The application combines three analytical components, cluster analysis, anomaly detection, and series variation analysis, within an integrated visualization framework. As shown in Figure 3.1, the pipeline begins with data collection and initial exploration, followed by specific modules responsible for quality checks, clustering structures, and statistical drift detection. Each component is supported by preprocessing routines and interactive visualizations that allow users to explore complex datasets dynamically. The integration of dimensionality reduction, unsupervised learning, and behaviour profiling provides the foundation for understanding payment system behaviour and detecting irregularities.

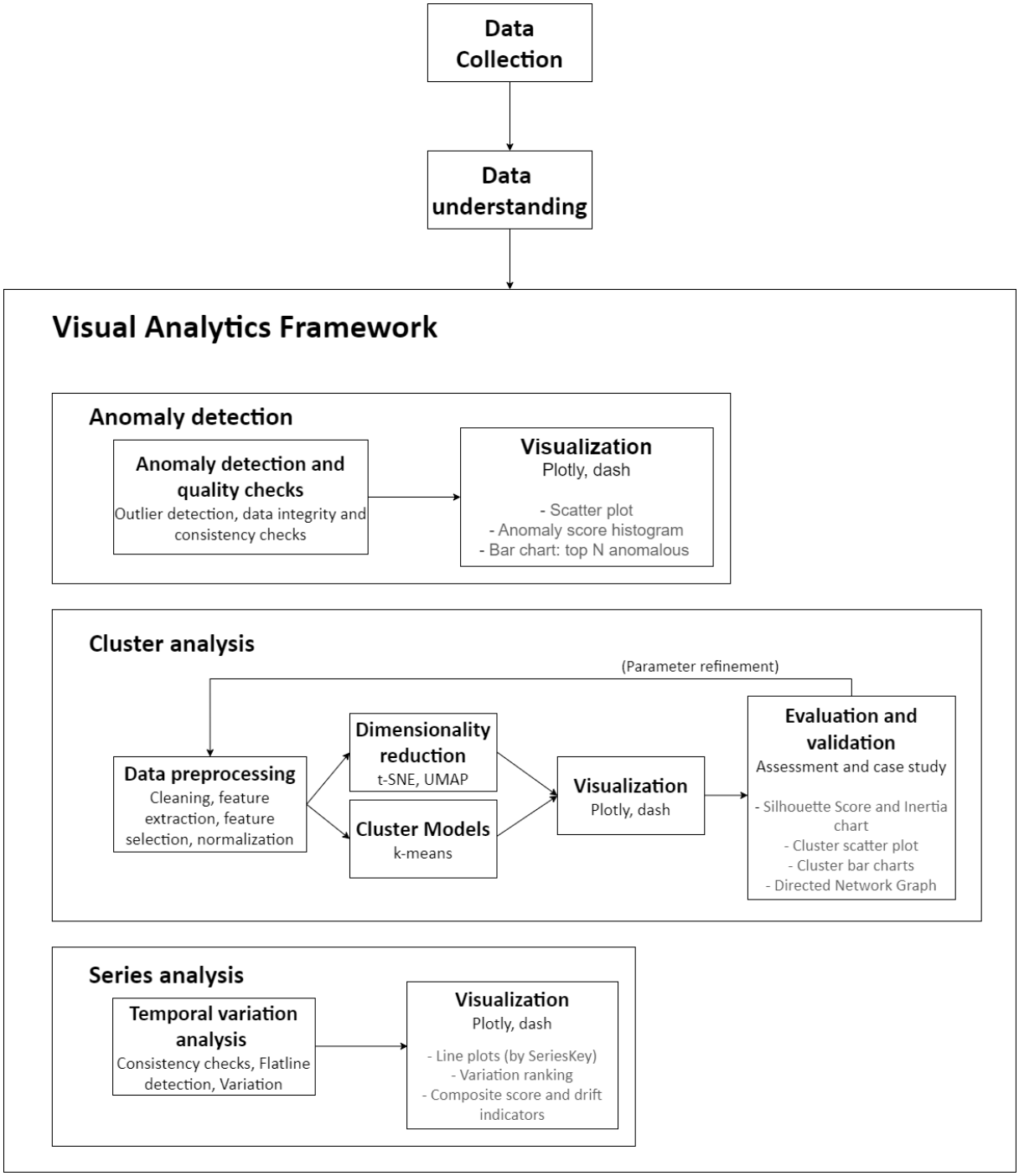


Figure 3.1 – Overview of the Visual Analytics Framework

3.1 DATA COLLECTION AND UNDERSTANDING

The present investigation is based on payment data collected by Banco de Portugal under Instruction No. 19/2012 (as updated by Instruction No. 9/2023), within the scope of the PAY. Under this scope, data are collected daily and at the individual transaction level for cheques,

credit transfers, direct debits, and payments card, with a reporting delay of one to three working days. In general terms, each transaction can be reported by up to three entities: the payment service provider (PSP) of the fund originator, the PSP of the beneficiary, and the entity that processes the payment. These multiple perspectives allow for cross-validation of the same transaction across different sources, ensuring data quality and consistency.

To demonstrate the applicability and flexibility of the proposed visual analytics framework (DeepPAY), two usage scenarios were constructed, each drawing from distinct data sources and analytical objectives.

The first scenario focuses on the clustering structure and network topology of the interbank payment system. It is based on high-dimensional transactional data involving credit transfer payments exchanged between payment institutions, specifically operations processed through SICOI, the Portugal's national system that processes and clears retail payments. The data covers the period from January 1st, 2025, to May 31st, 2025, and was retrieved directly from PAY. It comprises approximately 168 million individual credit transfer records, each capturing detailed transactional information, including transaction amount, volume, and characteristics of both the transaction and its participants. The transactional dataset includes categorical features such as channel (e.g., in-person, online banking, …), transfer type, and geographic region, which together provide contextual information about each payment. This level of granularity enables the exploration of behavioural clusters and flow dynamics within the payment system. It is important to note that the dataset complies with the General Data Protection Regulation (RGPD) and does not contain any personal or sensitive information. As such, it is not possible to identify individual clients based on the data used. A detailed description of the variables used is provided in Table A.1 in Appendix A.

The second scenario investigates temporal dynamics and anomalies in aggregated payment behaviour, with a particular focus on direct debits. For this purpose, a total of four monthly time series covering a seven-month period were retrieved from BPStat, Banco de Portugal's official statistical platform. These series cover multiple categories of direct debit instructions and rejections, and are used to assess month-over-month variation, identify flatline behaviour, and detect structural breaks in reported values. The selection of the four series was guided by their coverage of distinct operational categories, their behavioural diversity (patterns of flatness and variation) and their representativeness in the context of direct debit usage in Portugal.

These datasets reflect two complementary perspectives, micro-level transaction data and macro-level reporting series, that support both structural and temporal analyses within the DeepPAY framework.

3.2 SYSTEM ARCHITECTURE AND DESIGN

The DeepPAY framework is architected as a modular and scalable system for visual analytics over high-dimensional payment data. It integrates anomaly detection, cluster analysis, and time series variation into an interactive application developed with Dash. The system is designed to allow components to be reused across different tasks, keep the logic of each part clearly separated to improve maintainability, and provide the user with control over how to explore and interact with the data.

Some components of the application are common to all analytical modules but are not connected. As a result, changes in one module do not automatically affect the others, requiring each module to independently import data and instantiate shared elements as needed. This design allows for modular independence while still promoting reuse and standardization.

The shared services include a flexible data ingestion layer, supporting both file uploads (CSV or Excel) and direct SQL queries. This ensures that the application can operate in a wide range of institutional or operational contexts. In addition, in each module, checks for specific fields are performed that are required for its respective analytical tasks, such as transaction amounts, quantities, time references, or identifiers. These checks help ensure internal consistency and support the harmonization of critical variables across the system.

A responsive layout with uniform navigation is implemented to maintain a coherent interface across devices. The landing page (index module) serves as the homepage of the application, offering a visually clean and intuitive design that provides general instructions and facilitates access to the different analytical modules, namely, anomaly detection, clustering, and time series. Navigation buttons and concise module descriptions are presented to guide users effectively, boost usability, and reinforce code clarity and functional independence across components. An illustration of this navigation interface is shown in Figure 3.2.

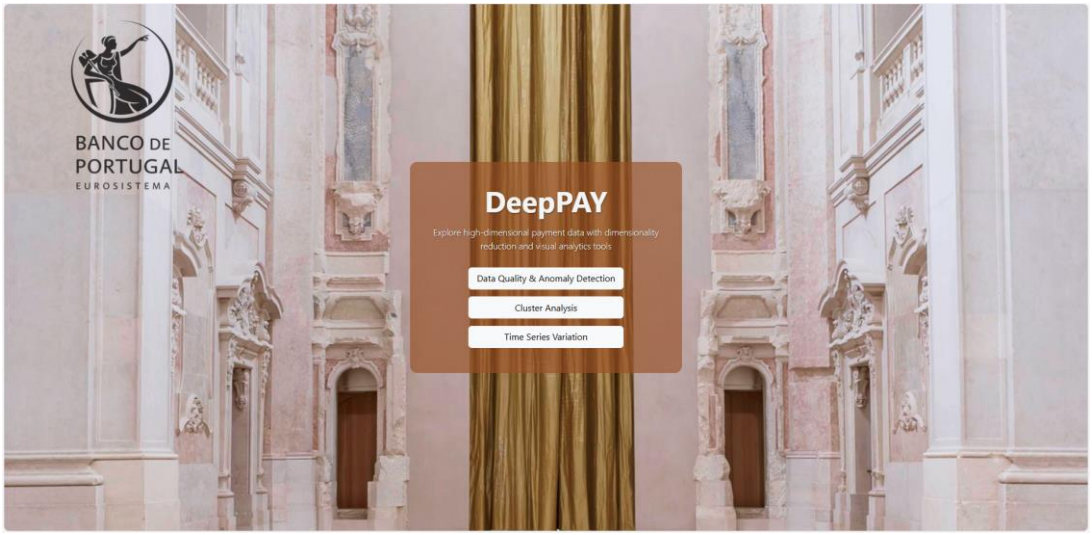


Figure 3.2 - Navigation interface of the DeepPAY system

3.2.1 DATA QUALITY AND ANOMALY DETECTION MODULE

This subchapter presents the methodology used to detect anomalous observations in payment data, integrating a hybrid model that combines Autoencoder-based feature compression with Isolation Forest for outlier detection. The methodological selection is informed by findings in the literature, particularly Yang et al. (2022).

The implementation was structured within a modular pipeline in Python (Figure 3.3).

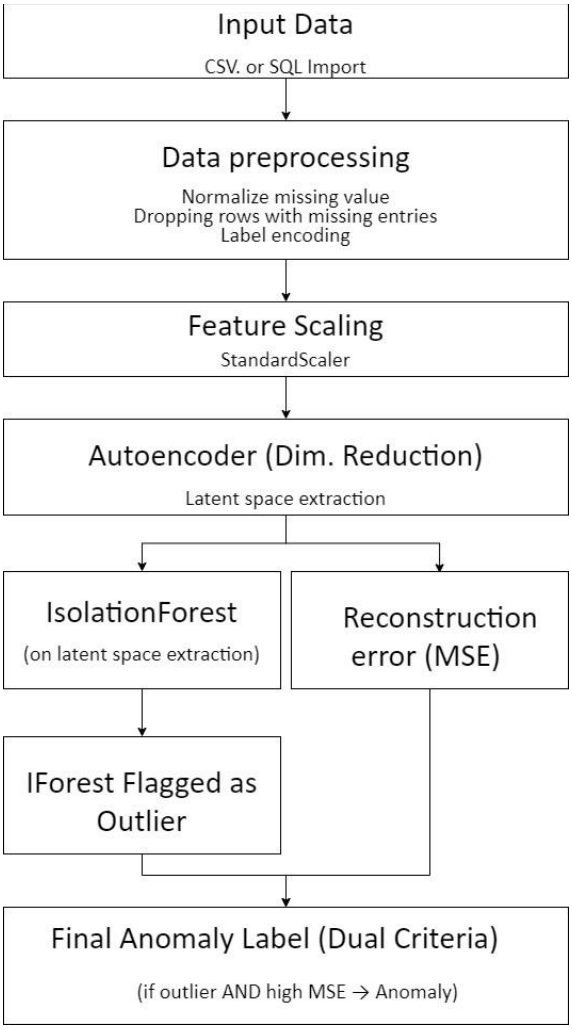


Figure 3.3 - Pipeline for Anomaly Detection in payment data

The preprocessing phase ensures the integrity and standardization of the data, including normalization of missing value representations (e.g., "NULL", "NA", "?", among others) into a standardized NaN format. Rather than imputing values, rows containing missing entries in critical fields were removed to preserve analytical reliability. Additionally, categorical variables were encoded numerically using label encoding, and type coercion was applied where

necessary to ensure consistency across the dataset. Subsequently, all numerical variables were standardized using z-score scaling (StandardScaler). These steps ensured compatibility with the modeling tools and reduced the risk of model bias due to formatting inconsistencies or variables with larger magnitudes.

The modelling process was structured in two main steps. First, a dense Autoencoder neural network was trained in an unsupervised approach to compress and reconstruct the input data. By learning to compress and then rebuild the data, the Autoencoder was able to measure reconstruction error using mean squared error as an initial anomaly indicator. Observations with high reconstruction error typically deviate from the learned data structure and are thus candidates for being anomalous. In parallel, the latent representation extracted from the encoder was used as input to an Isolation Forest model. This ensemble method isolates data points through recursive partitioning, identifying those more easily separated from the rest. The final classification of a transaction as anomalous was only made when both methods agreed that is, when the reconstruction error exceeded the 95th percentile and the Isolation Forest simultaneously flagged the observation as an outlier. This dual condition increased the reliability of the detection process by ensuring that anomalies are not only structurally inconsistent but also statistically rare.

3.2.2 CLUSTER ANALYSIS MODULE

The Cluster Analysis Module implements an unsupervised learning pipeline designed to uncover hidden structure in high-dimensional payment data. It achieves this by combining dimensionality reduction techniques with clustering algorithms, enabling the segmentation of observations based on patterns in user-selected variables.

Regardless of the input method, the data must contain at minimum the following fields: “DtLiq” (the transaction date), “Mont” (the transaction amount), and “Quant” (the transaction quantity). The system validates the presence of these mandatory fields and applies type coercion to convert “Mont” and “Quant” to numeric format and “DtLiq” to datetime. Records with missing or invalid values in any of these columns are excluded.

Following validation, all numeric columns in the dataset are identified. High-cardinality numeric features are considered quantitative variables for clustering, whereas low-cardinality numerical attributes are classified as semantic features and treated as categorical proxies in downstream visualizations such as stratified bar charts or network-based graphs.

Before clustering, a correlation matrix is computed across the numeric features. This is used to detect multicollinearity and guide the user in selecting a minimal and non-redundant feature set. This matrix is rendered as a heatmap with coefficients, enabling users to identify multicollinearity and potential variables for dimensionality reduction.

Following this diagnostic step, users are prompted to select the features to be included in the clustering process. Selected features are standardized via z-score normalization using StandardScaler. This transformation ensures that each feature contributes equally to the Euclidean distance calculations during clustering. The standardized variance of each feature is also computed and visualized, assisting users in evaluating the relative informativeness of the selected variables.

To project the selected feature space into a two-dimensional visualization, users may choose between two dimensionality reduction methods. PCA is a linear transformation that converts the original variables into a new set of orthogonal components that sequentially capture the maximum variance in the data. Alternatively, UMAP is a non-linear technique that aims to preserve both the global and local structure of the data manifold. UMAP's behaviour can be adjusted through the *n_neighbors* and *min_dist* hyperparameters, which control the balance between local detail and global shape preservation. When UMAP is selected, PCA is automatically applied beforehand as a preprocessing step. This design choice helps to reduce noise and improve computational efficiency by first projecting the data into a lower-dimensional space while retaining most of its variance.

After dimensionality reduction, the user-defined number of clusters *k* is passed to the K-Means algorithm. The algorithm partitions the data by minimizing intra-cluster variance and assigns each observation to the nearest cluster centroid. The resulting labels are appended to the original dataset and serve as the basis for further quantitative and graphical analysis.

To support the evaluation of clustering quality and the selection of an appropriate number of clusters, the module computes two key diagnostic metrics across a range of *k* values. The first, inertia, measures the within-cluster sum of squared distances to the cluster centroids, providing an indication of how compact the clusters are. The second, the silhouette score, quantifies how similar each data point is to its own cluster relative to other clusters, with values ranging from -1 to 1; higher scores suggest better-defined and more separable clusters.

To complement the feature-based clustering analysis, the module incorporates a graph-based modelling approach that captures the topological structure of payment flows. A directed network is constructed from the transaction dataset by identifying each unique entity involved in a transaction, specifically, the sending institution and receiving institution, as nodes. Directed edges are established from source to target to reflect the directionality of each payment, and edges are weighted by the transaction volume ("Quant").

The network is illustrated using a directed graph object. Node attributes are augmented with cluster labels inherited from the feature-based K-Means clustering, enabling comparison between spatial and topological segmentations. Furthermore, node centrality metrics, such as in-degree and out-degree, are computed to assess the structural prominence of each node in the network. These attributes are later used to scale node size and support systemic role identification.

The graph layout is determined using a force-directed algorithm (spring layout), which assigns spatial coordinates to nodes based on their connectivity. This methodology allows the system to capture both transactional intensity (via edge weights) and structural centrality (via node degree), facilitating an additional layer of clustering analysis grounded in network topology.

Feature-space clusters are derived from selected features such as “Mont” and “Quant”, visualized via PCA or UMAP. Network clusters emerge from the transaction graph and highlight participants with similar connectivity patterns. This dual approach enriches the analytical value of the module, allowing both behavioural segmentation and systemic role identification.

3.2.3 TIME SERIES VARIATION

The time series variation module was designed to support the characterization of each individual series in the dataset through a collection of robust and interpretable indicators. This functionality is particularly relevant in the context of payment data, where inconsistencies in temporal reporting, sudden pattern changes, long periods of inactivity, or abnormal stability and drift may signal either underlying transactional anomalies or data quality issues related to reporting. By combining metrics that capture both structural and behavioural deviations over time, this module contributes to a more informed and diagnostic exploration time series.

The first step in this process involves ensuring a consistent data structure. For this module to function correctly, the data imported must contain, at a minimum, the following fields: year (“Ano”), month (“Mes”), a unique series identifier (“SeriesKey”), and two numeric measures: quantity (“Quant”) and amount (“Mont”). These are then used to construct a proper time axis by combining the year and month into a complete date, allowing the system to align each series along a monthly frequency.

Once passed this initial control, the time series variation module incorporates a short-term analysis focused on detecting recent deviations. Specifically, the system computes the relative percentage change in the selected measure between the two most recent months for each series. These changes are then compiled into a ranked table of the Top Varying Series, filtered by a user-defined variation threshold. This component acts as a lightweight anomaly detection mechanism, drawing immediate attention to series experiencing abrupt short-term shifts, even if their broader temporal profile remains stable.

In parallel with the short-term analysis, the system computes a structural profiling analysis, where each time series is evaluated across several dimensions, as described in the Table 1. These dimensions were chosen for their ability to capture both structural characteristics and behavioural irregularities of the series over time.

Table 3.1 - Time Series Profiling Indicators for Variation Analysis

Indicator	**Description**	**Purpose**
Missing values (%)	Percentage of missing values in the quantity/amount dimension over the selected period.	Evaluate reporting completeness.
Flatline (%)	Proportion of months where quantity/amount value did not change from the previous month.	Detect cases where values remain unchanged from one month to the next.
Mean Drift	Absolute difference in mean quantity/amount between first and second halves of the series.	Detect global structural change.
Std Drift	Absolute difference in quantity/amount standard deviation between first and second halves of the series.	Identify shifts in variance.
Drift Ratio	Quantity/amount mean drift normalized by the overall standard deviation of the series.	Provide scale-independent drift metric.
Local Drift	Absolute difference in mean quantity/amount between the last two months and the preceding two.	Highlight recent changes in behaviour.
Last Z-Score	Z-score of the latest quantity/amount observation compared to the historical mean and standard deviation (excluding the last value).	Detect recent outliers.
Composite Score	Mean of all normalized indicators (scaled between 0 and 1).	Prioritize series with multiple quality concerns or structural shifts.

The completeness analysis was operationalized by measuring the proportion of missing values for both quantity (QT) and amount (MT) dimensions over the selected period. To complement this, flatline detection was conducted to flag sequences where no change occurred over time, typically suggestive of non-reporting or erroneous data reporting. These values were then complemented by various drift metrics, such as normalized mean drift, local variation between recent months, and the Z-score of the last observed value.

Although these indicators are designed to be robust, the quality and interpretability of results depend on the time coverage of each series. A minimum of five monthly observations is required to compute the variation metrics. This threshold ensures that the series can be divided into segments for drift analysis and still retain sufficient data points to compute changes in mean, variance, and recent fluctuations. Fewer than five data points would compromise the reliability of the estimates, while setting a higher threshold (e.g., six or more) could exclude shorter but still informative series.

4 INTERATIVE VISUALIZATION LAYER

The visual layer of the framework plays a fundamental role in translating complex analytical outputs into accessible and actionable insights. Built using Dash, this component provides an interactive interface that complements the underlying analytical modules by enabling users to explore patterns, identify anomalies, and interpret results with clarity. The following subsections describe how each analytical module is visually represented, focusing on the layout, features, and user features of the interface.

4.1 DATA QUALITY AND ANOMALY DETECTION MODULE

The layout was designed to offer clarity and immediate feedback, enabling users to navigate, explore, and export results with ease. An overview of the dashboard can be seen in Figure 4.1, which illustrates the integration of anomaly visualizations, summary statistics, and download functionality.

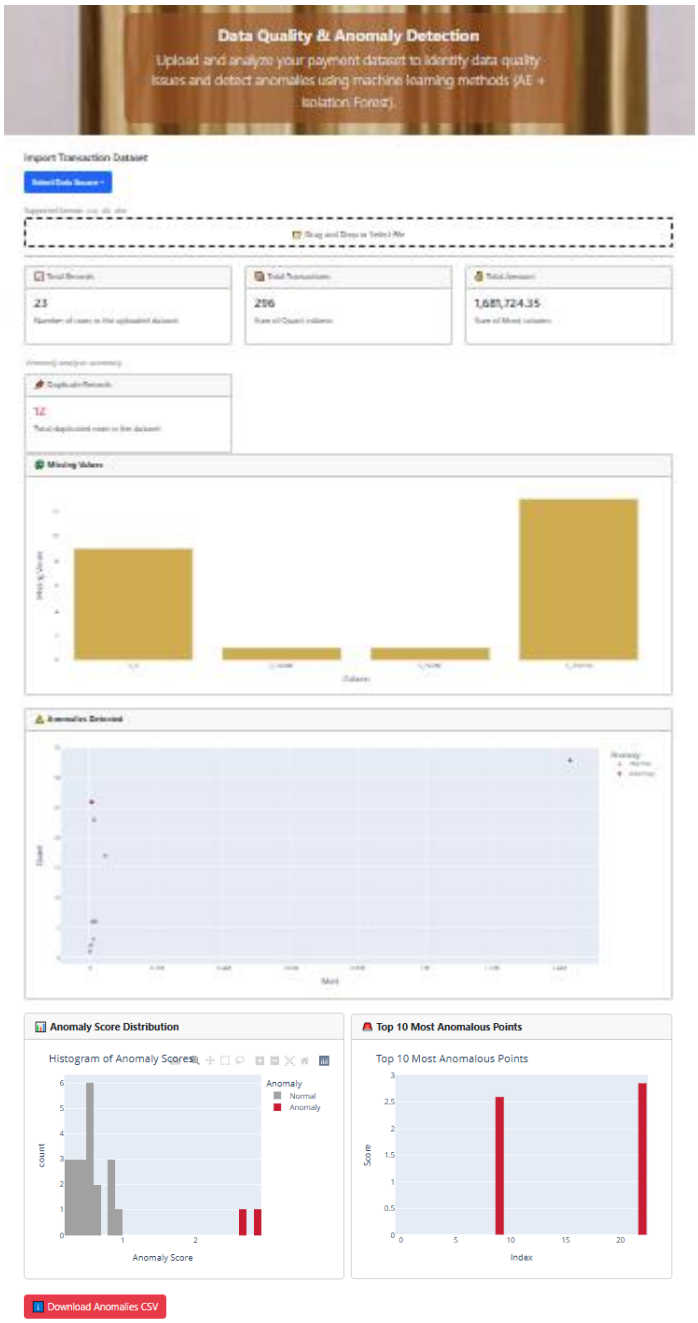


Figure 4.1 - Dashboard interface showing anomaly detection outputs

The interface presents a scatter plot of transaction amount (“Mont”) versus transaction quantity (“Quant”), with anomalies highlighted using a distinct colour scheme: red for anomalies and grey for normal transactions. This simple visualization immediately reveals the location and density of outliers relative to the main distribution.

A second component displays a histogram of anomaly scores, allowing the user to evaluate the distribution of reconstruction errors across the dataset. This helps contextualize the threshold used in the detection phase and assess the intensity of deviations.

To support focused analysis, a third component offers a bar chart showing the top ten most anomalous transactions. These are selected based on their anomaly scores, providing a concise list of observations that merit closer inspection.

In addition to these visual components, the module includes the possibility to download all detected anomalies as a .csv file. The exported data preserves the original values, prior to any encoding or standardization, ensuring that further analysis or external validation is conducted on the native representation of the information.

Each visual element was further enriched with contextual information through a tooltip. By hovering over any point in the plots, the user gains access to the key fields that identify the transaction, as well as the anomaly score associated with that observation. This level of granularity increases interpretability and provides a transparent basis for informed decision-making.

4.2 CLUSTER ANALYSIS MODULE

The visualization layer of the Cluster Analysis Module supports interactive exploration of unsupervised learning results by combining summary statistics with topological representations of transaction data. Its graphical outputs are designed to help users interpret both the internal structure of high-dimensional feature clusters and the systemic relationships between entities in the network.

The clustered data, obtained through dimensionality reduction and K-means, is displayed in a two-dimensional scatterplot, where each point represents an observation coloured by its assigned cluster label. This projection allows users to visually assess the separation between clusters and identify structural patterns in the data distribution.

Each cluster is further characterized through summary metrics including the total number of transactions, the average transaction amount (“Mont”), and the average quantity (“Quant”) These metrics are presented both in tabular format and through interactive horizontal bar charts, enabling intuitive comparisons across clusters.

To support the selection of an appropriate number of clusters, the interface includes a dual-axis line chart that visualizes two diagnostic metrics computed over a range of *k* values: inertia, which quantifies within-cluster compactness, and the silhouette score, which measures cohesion and separation.

To complement the cluster analysis and provide insights into payments flows, the module includes a Directed Network Graph visualization, where nodes represent payment participants (i.e., PSP). Node colour is got from the cluster label assigned in the scatterplot analysis, while node size is scaled using centrality such as in-degree and out-degree, highlighting participants

with high transactional activity within the network. Edges represent transactions, where the direction reflects the flow (e.g., from payer to payee), weighted by “Quant”.

The network is laid out using a force-directed algorithm, which positions nodes based on their transactional connectivity to reduce overlap and reveal communities. This visualization enables users to identify whether observed clusters are structurally coherent and to detect hubs, bridges, or isolated nodes within the transaction topology. By integrating feature-based and topology-based segmentation, the network graph offers a dual perspective that improves interpretability and systemic insight.

This graph bridges the clustering results with structural payment flow information, offering a second modality of clustering analysis based on topological clusters as opposed to feature-space clusters.

4.3 TIME SERIES VARIATION

The visualization interface designed for this module seeks to support an intuitive and scalable overview of time series variation. The layout is structured around two main components: a grid of line charts and a summary table of quality indicators.

Each chart in the grid represents a unique series and displays either the quantity or amount over time, depending on user selection. The user can specify a date range and switch between “Quant” and “Mont” views dynamically. The choice of visual encoding is consistent across all charts (same layout, colour scheme, and axis configuration). This design supports comparative pattern recognition while avoiding visual noise.

In addition, the interface also includes a compact module labelled *Top Varying Series*, a component that highlights the most abrupt relative changes between the most recent and penultimate months within the selected period. Users can sort the variation list in ascending or descending order and apply a threshold filter to exclude minor fluctuations. The presence of this view helps users detect short-term anomalies even if the overall structural metrics remain stable.

Additionally, below the chart gallery, the user is presented with a comprehensive table summarizing the diagnostics computed for each series. This includes the percentage of missing values, flatline rates, multiple forms of drift, the latest z-score, and the composite anomaly score. To make these indicators more interpretable, each column header includes a contextual tooltip that succinctly explains the corresponding metric. For example, the tooltip for the “QT Drift Ratio” clarifies that it represents the normalized difference in average values between the first and second halves of the series, while “QT Flatline (%)” denotes the proportion of months with no observable change.

The user may sort and filter this table by any metric or focus exclusively on series whose quality metrics exceed a predefined threshold. In doing so, the dashboard shifts from being a passive reporting tool to a dynamic exploration interface. This dual-layered approach, combining visual timelines with multidimensional profiling, offers a rich environment for anomaly detection and ongoing data quality assessment.

5 RESULTS AND DISCUSSION

To evaluate the effectiveness of the proposed framework, real payment data collected under the PAY was used. The three main analytical modules of the DeepPAY application, anomaly detection, cluster analysis, and time series exploration, were used to examine different aspects of the data. The results are presented in terms of the patterns discovered, interpretability of the visualizations, and responsiveness of the interface.

5.1 SYSTEM OVERVIEW

Before proceeding with any analysis, the DeepPAY application validates the structure of the uploaded dataset to ensure compatibility with each module. In this scenario, a file was submitted in the anomaly detection and quality assessment module without the required columns “Mont” (amount) and “Quant” (quantity). As shown in Figure 5.1, the application immediately flags this issue, returning a clear error message to the user.

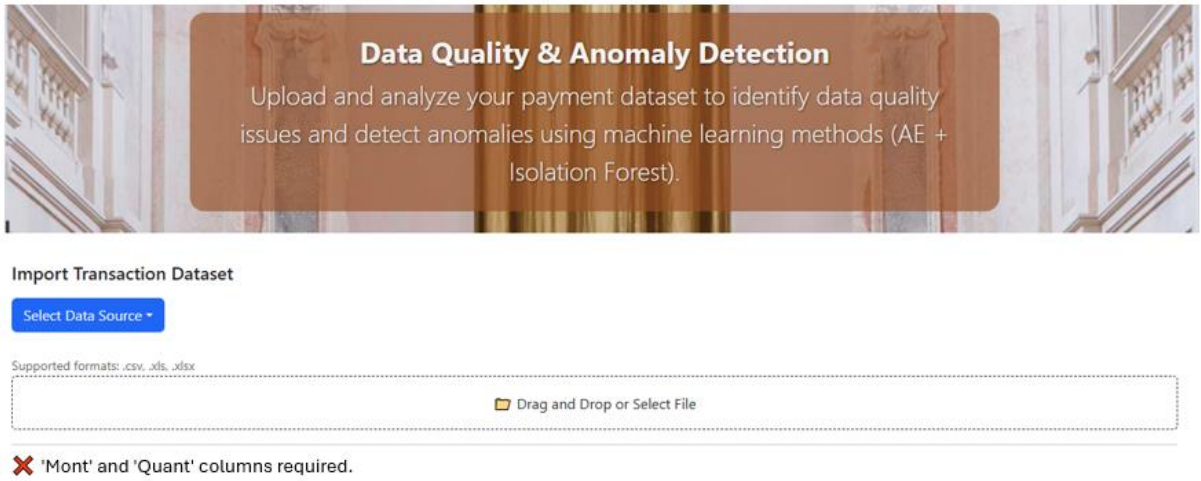


Figure 5.1 - Upload Error: Missing Mandatory Fields

This initial validation step is critical to ensuring downstream components (such as the Autoencoder + Isolation Forest pipeline) function as intended. It prevents runtime errors and guides the user toward providing a valid dataset.

5.2 USE CASE 1: CLUSTER AND NETWORK ANALYSIS OF THE PAYMENT SYSTEM

Understanding the structure of interactions between participants in a payment system is essential to uncovering operational patterns, network dependencies, and systemic vulnerabilities. In this use case, a combination of clustering and network analysis techniques

was applied to high-dimensional payment data to identify distinct behavioral profiles among entities and to visualize the underlying payment topology.

The dataset comprises credit transfer transactions exchanged between institutions during the period from January 1st to May 31st 2025. As described in Section 3.1, the input includes operational and contextual variables such as transaction amount, volume, type, channel, and geographic indicators. A complete description of the features is provided in Table A.1 (Appendix A).

To assist the variable selection step for clustering, the system calculates the standardized variance of each feature after preprocessing. The selected features as shown in Figure 5.2 have roughly the same level of standardized variance which shows that these features balance and contribute towards the variability in the dataset. This step ensures that no single feature outweighs the others when it comes to determining the clusters formed. From the original dataset, only the features with non-zero variance were kept, as variables exhibiting no variation provide no discriminatory power in unsupervised learning and do not contribute to meaningful cluster formation.

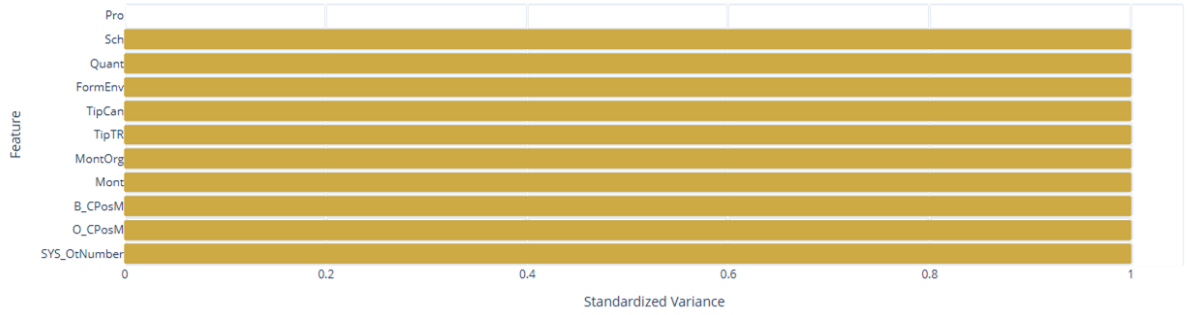


Figure 5.2 - Feature Variance (Standardized)

The selected features were then used to build a two-step clustering pipeline. First, data dimensionality reduction was performed using UMAP, which projects high-dimensional data to a latent space that is more suitable for clustering. Subsequently, behavioural clustering was accomplished using K-means. The optimum number of clusters was derived from both the Silhouette Score and the Elbow Method. The results are shown in Figure 5.3, illustrating that the silhouette curve (red line), reaches its maximum at k = 3, indicating strong cohesion within clusters and good separation between them. At the same time, the inertia curve (grey line) exhibits a sharp decline up to k = 3, after which the incremental gains in explained variance become marginal. The convergence of these two criteria supports the selection ok k = 3 as the most appropriate clustering configuration for the dataset under analysis.

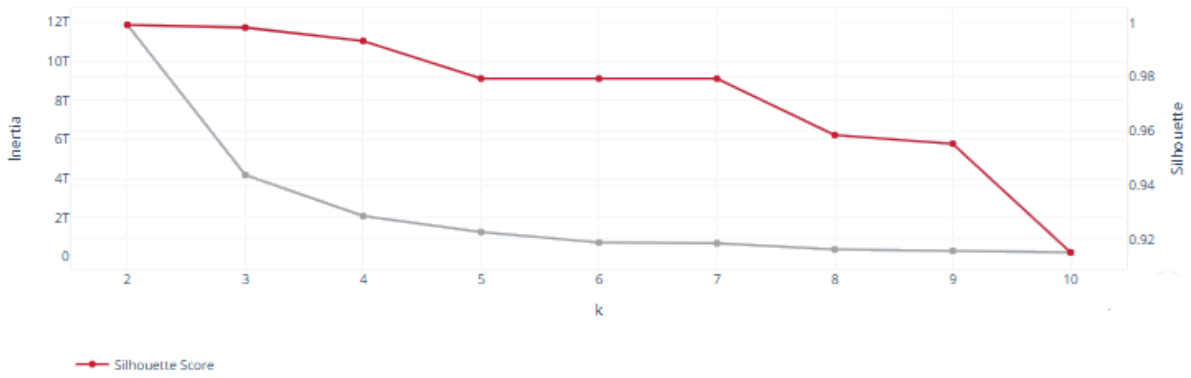


Figure 5.3 - Cluster Evaluation Using Silhouette Score and Elbow Method

To complement the feature-based segmentation, a directed network graph was constructed (Figure 5.4). In this representation, each node corresponds to a payment system participant, and each edge represents a transaction from the originator's PSP to the beneficiary's PSP. Edge weights reflect transaction volumes ("Quant"), and node size indicates total degree (i.e., sum of in- and out-degrees). Node colours correspond to cluster labels from the feature-based segmentation, enabling a layered interpretation of both behavioural and topological roles.

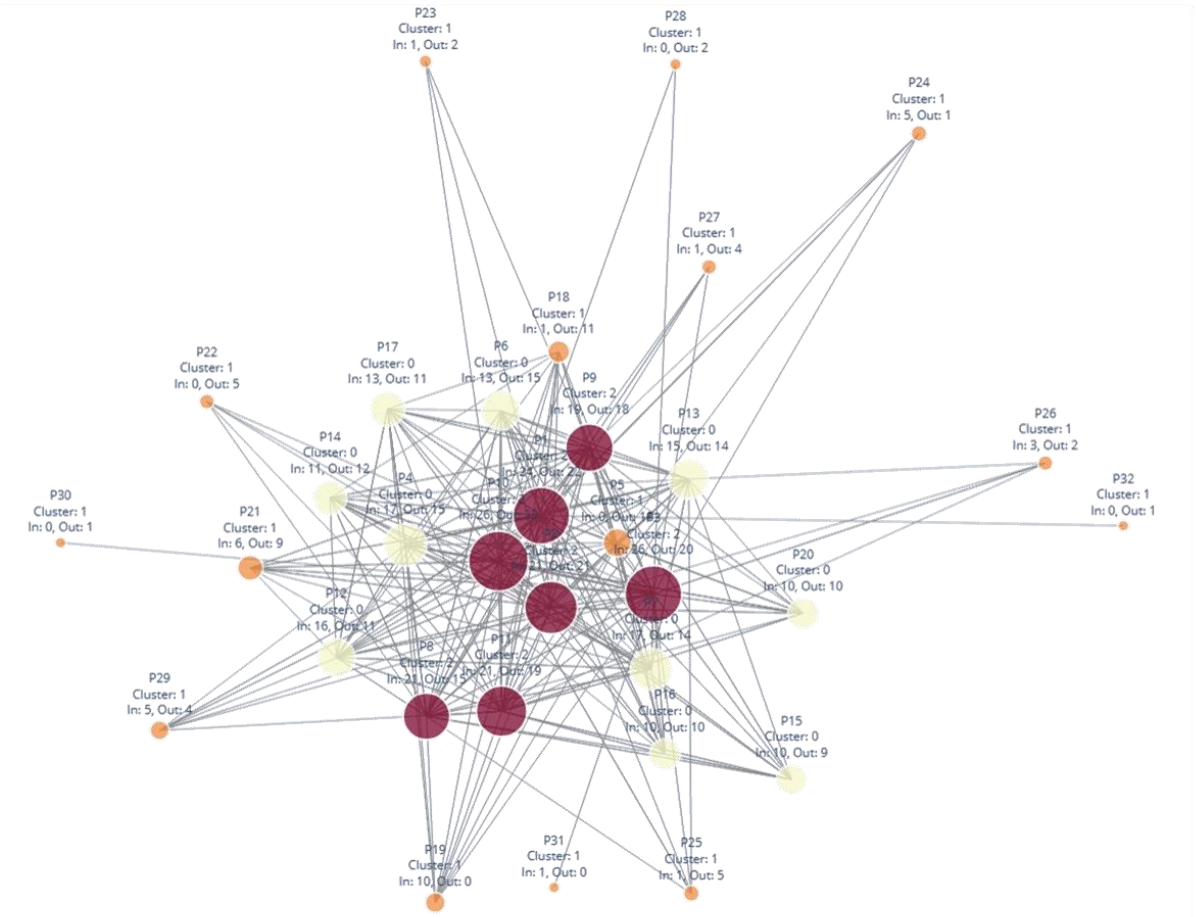


Figure 5.4 - Directed Network Graph of Payment System Participants

The resulting network reveals a dense and highly interconnected structure, particularly around Cluster 2 (dark red), which aggregates the most active entities in terms of connectivity and volume exchange. These nodes likely correspond to institutions with central operational roles, serving as major liquidity providers within the system. The presence of numerous overlapping edges highlights the high volume of bilateral interactions concentrated in this cluster, reinforcing their systemic importance.

Surrounding this core, Cluster 0 (light yellow) contains several mid-sized nodes exhibiting moderate degrees of connectivity. These entities appear to function as intermediaries, linking the highly active core to the network's periphery. Their positioning and flow patterns suggest a functionally diverse set of roles, either facilitating payment routing between major and minor participants, or as participants with specialized bilateral flows.

At the periphery of the graph, Cluster 1 (orange) includes participants with comparatively low connectivity and smaller node sizes. These entities are characterized by asymmetric flow

patterns, often exhibiting higher out-degree than in-degree. This asymmetry suggests that entities within this group act primarily as originators of transactions and receive comparatively few incoming transfers. Such a pattern may be associated with institutions executing specific payment functions, acting as payment initiators rather than intermediaries or endpoints in the network. This peripheral position in the network contrasts, both topologically and behaviourally, with the more balanced and densely connected profiles observed in Clusters 0 and 2.

5.3 USE CASE 2: TEMPORAL SERIES VARIATION

This use case focuses on monitoring temporal variation across individual series to identify patterns such as abrupt changes, flatline behaviour, or statistical drift, phenomena that may reflect operational shifts or signal data quality concerns. The goal is to support the early detection of inconsistencies or trends within payment records.

Figure 5.5 illustrates the Series Variation Analysis module, where users can visually inspect monthly fluctuations for selected series and identify the most varying indicators over a specified period. In this example, April 2025 is compared to March 2025. The right-hand panel ranks series by percentage change, while interactive charts on the left display historical evolution.

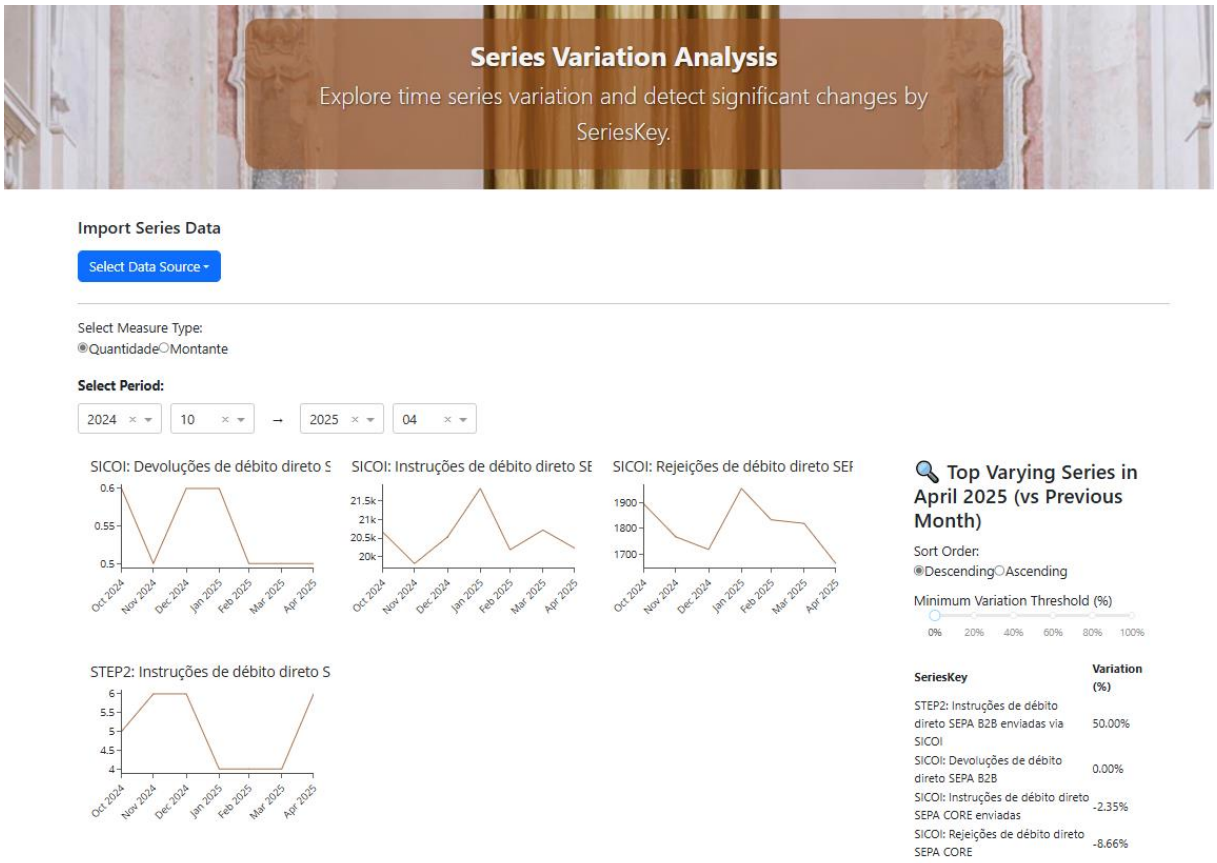
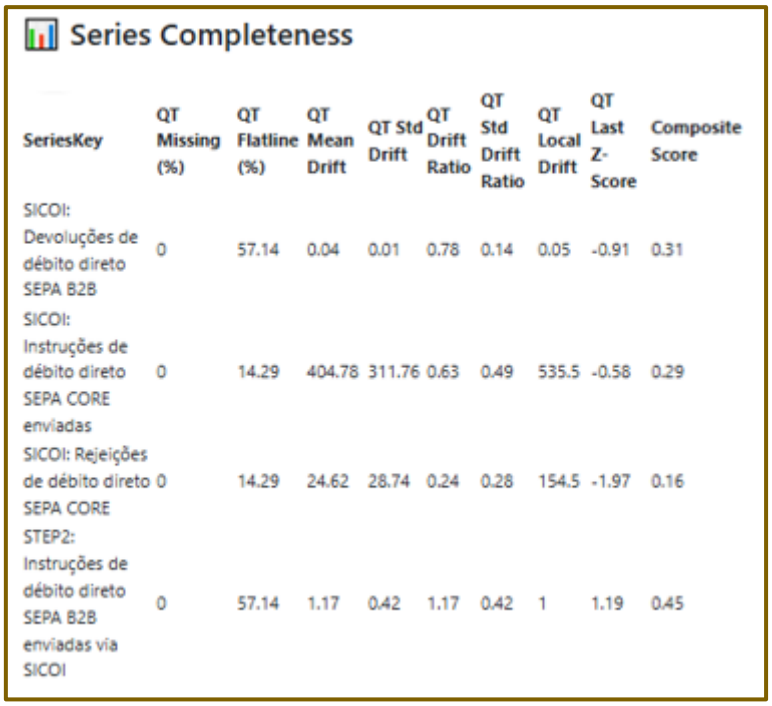


Figure 5.5 - Series Variation Analysis Module

The ranking shows that the series "STEP2: Instruções de débito direto SEPA B2B enviadas via SICOI" increased by 50.00%, while others such as "*SICOI: Rejeições de débito direto SEPA CORE*" and "*SICOI: Instruções de débito direto SEPA CORE enviadas*" recorded moderate negative variations of −8.66% and −2.35%, respectively. In contrast, some series remained flat during the selected period, including "*SICOI: Devoluções de débito direto SEPA B2B*", which showed 0% change.

As a complement to the variation monitoring, the Series Completeness panel (Figure 5.6) at the bottom of the interactive charts presents combined completeness metrics like drift scores and Z-score variations. For instance, "*STEP2: Instruções de débito direto SEPA B2B enviadas via SICOI*" series, despite its 50.00% increase, also shows a QT flatline of 57.14% and a local Z-score of 1.19, indicating that the increase in April may represent a sudden change following multiple months of inactivity. This is reinforced by low values in both QT mean drift (1.17) and QT standard deviation (0.42), suggesting a sharp one-time jump rather than a sustained trend.



Series Completeness

SeriesKey	QT Missing (%)	QT Flatline (%)	QT Mean Drift	QT Std Drift	QT Drift Ratio	QT Std Drift Ratio	QT Local Drift	QT Last Z-Score	Composite Score
SICOI: Devoluções de débito direto SEPA B2B	0	57.14	0.04	0.01	0.78	0.14	0.05	-0.91	0.31
SICOI: Instruções de débito direto SEPA CORE enviadas	0	14.29	404.78	311.76	0.63	0.49	535.5	-0.58	0.29
SICOI: Rejeições de débito direto SEPA CORE	0	14.29	24.62	28.74	0.24	0.28	154.5	-1.97	0.16
STEP2: Instruções de débito direto SEPA B2B enviadas via SICOI	0	57.14	1.17	0.42	1.17	0.42	1	1.19	0.45

Figure 5.6 - Series Completeness and Drift Metrics

The "*SICOI: Rejeições de débito direto SEPA CORE*" series, which recorded a decrease of −8.66%, exhibits a relatively high level of regular fluctuation. This is reflected in its QT mean drift of 24.62, combined with a stable drift ratio and a local Z-score of −1.97, suggesting that the observed change falls within the expected range of variation and does not necessarily indicate an anomaly.

By combining visual trend inspection with statistical drift analysis, this module enables users to distinguish between genuine shifts in behaviour and data-related irregularities. The ability to isolate flatline effects, measure standardized variation, and evaluate Z-scores in parallel supports more reliable interpretation.

While traditional tools such as spreadsheets or ad hoc scripts can perform basic calculations, they often lack the scalability and interactivity required in central bank analysis. These limitations are exacerbated when working with thousands of time series, where manual or static approaches are no longer feasible. By embedding statistical drift metrics within an interactive visual interface, the Series Variation Analysis module supports scalable monitoring, increases interpretability, and enables anomaly triage at scale. It provides a reproducible and efficient framework to track reporting consistency and behavioural change over time, capabilities that are essential to unsure data quality in complex payment infrastructures.

6 CONCLUSIONS AND FUTURE RESEARCH

This research introduces DeepPAY, a visual analytics framework designed to detect anomalies and assess data quality in large-scale transactional datasets. By integrating dimensionality reduction, clustering, and anomaly detection within an interactive environment, the framework facilitates the exploration of high-dimensional payment data and supports the identification of hidden structures, behavioural patterns, and reporting inconsistencies.

A key contribution of this work lies in its modular and scalable architecture, which supports data exploration. The framework's design enables users to intuitively navigate complex data landscapes, identify outliers, and assess data quality through multiple perspectives, including statistical variation, clustering behaviour, and topological flow.

One of the primary advantages of DeepPAY is the combination of unsupervised learning with network representations which increases interpretability. Specifically, the clustering results, combined with transaction flow topology, help reveal the roles that different institutions play, such as identifying those that are more central or critical within the payment system. In addition, the system is scalable, which allows the monitorization of thousands of time series simultaneously. Unlike static tools, which include spreadsheets or ad hoc scripts, DeepPAY places embedded statistical drift metrics within a visual and reproducible framework, enabling efficient anomaly prioritization and monitoring consistency in reports over time. Furthermore, the identification of outliers enables the detection of structurally atypical events, such as electricity outages, mass gatherings like concerts, or large-scale disruptions like pandemics, which often manifest as abrupt changes in payment behaviour. In addition, it can uncover more subtle or previously unnoticed shifts in transaction patterns that may not be easily identifiable through traditional monitoring methods.

Although the current implementation provides a solid foundation for supervisory analysis, it also has some limitations. Future research could further improve this foundation by embedding the current framework within a streaming infrastructure, enabling real-time anomaly detection at scale. Furthermore, while the use of dimensionality reduction and clustering improves pattern recognition, the interpretability of latent spaces remains a challenge, especially for non-technical users. Incorporating explainable models into the pipeline would help bridge the gap between automation and interpretability, providing users not only with alerts about anomalous behaviour but also with insights into the underlying causes.

In addition, future work could include controlled experiments to assess the practical effectiveness of the system. For example, evaluating whether analysts are better able to identify data inconsistencies, anomalous variations, or structural irregularities when using DeepPAY compared to conventional tools would provide valuable insight into its real-world utility. These user-centric evaluations could guide further design improvements and boost the tool’s impact in supervisory or operational settings.

In summary, by aligning analytical capabilities with user-centric design, this work contributes with a practical, more transparent and scalable, solution to the challenge of analysing high-dimensional payment data, balancing computational complexity with user-centric design. Its flexibility makes it not only suitable for central banking oversight but also adaptable to other domains facing similar data complexity and interpretability requirements.

BIBLIOGRAPHICAL REFERENCES

Aggarwal, C. C. (2002). Towards effective and interpretable data mining by visual interaction. *SIGKDD Explor. Newsl.*, *3*(2), 11–22. https://doi.org/10.1145/507515.507518

Agrahari, S., & Singh, A. K. (2022). Concept Drift Detection in Data Stream Mining : A literature review. *Journal of King Saud University - Computer and Information Sciences*, *34*(10, Part B), 9523-9540. https://doi.org/10.1016/j.jksuci.2021.11.006

Ardizzi, G., Bruno, G., Marcucci, J., Iannaccone, R., Moauro, F., Righi, A., & Zurlo, D. (2024). The use of payment transaction data for economic forecasts. *IFC Bulletins chapters*, *61*.

Arévalo, F., Barucca, P., Téllez-León, I.-E., Rodríguez, W., Gage, G., & Morales, R. (2022). Identifying clusters of anomalous payments in the salvadorian payment system. *Latin American Journal of Central Banking*, *3*(1), 100050. https://doi.org/10.1016/j.latcb.2022.100050

BdP. (2022). *Relatório dos Sistemas de Pagamentos*. B. d. Portugal. https://www.bportugal.pt/sites/default/files/anexos/pdf-boletim/rsp2022.pdf

BdP. (2024). *Relatório dos Sistemas de Pagamentos*. B. d. Portugal. https://www.bportugal.pt/sites/default/files/documents/2025-05/rsp2024.pdf

Bellman, R. E. (2015). *Adaptive Control Processes: A Guided Tour*. Princeton University Press. https://books.google.pt/books?id=iwbWCgAAQBAJ

Berndsen, R., & Heijmans, R. (2020). Near-real-time monitoring in real-time gross settlement systems: a traffic light approach. *Journal of Risk*.

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly Detection: A Survey. *ACM Comput. Surv.*, *41*. https://doi.org/10.1145/1541880.1541882

Donoho, D. L. (2000). High-dimensional data analysis: The curses and blessings of dimensionality. *AMS math challenges lecture*, *1*(2000), 32.

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). *A density-based algorithm for discovering clusters in large spatial databases with noise* Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, Portland, Oregon.

Glowka, M., Müller, A., & Weber, A. (2025). The hierarchy of critical participants: A clustering approach utilising network-based indicators for payment systems. *Latin American Journal of Central Banking*, 100169. https://doi.org/10.1016/j.latcb.2025.100169

Hasugian, P., Mawengkang, H., Sihombing, P., & Efendi, S. (2023). *Review of High-Dimensional and Complex Data Visualization*. https://doi.org/10.1109/ICoSNIKOM60230.2023.10364377

Jolliffe, I. T. (2002). *Principal Component Analysis*. Springer. https://books.google.pt/books?id=TtVF-ao4fI8C

Karimov, J., Rabl, T., Katsifodimos, A., Samarev, R., Heiskanen, H., & Markl, V. (2018). *Benchmarking Distributed Stream Data Processing Systems*. https://doi.org/10.1109/ICDE.2018.00169

Kelleher, J., Mac Namee, B., & D'Arcy, A. (2015). *Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies*.

Krause, J., Dasgupta, A., Fekete, J.-D., & Bertini, E. (2016). SeekAView: An intelligent dimensionality reduction strategy for navigating high-dimensional data spaces. *2016 IEEE 6th Symposium on Large Data Analysis and Visualization (LDAV)*, 11-19.

León, C. (2020). Detecting anomalous payments networks: A dimensionality-reduction approach. *Latin American Journal of Central Banking*, *1*(1), 100001. https://doi.org/10.1016/j.latcb.2020.100001

Liu, H., Motoda, H., Setiono, R., & Zhao, Z. (2010). Feature Selection : An Ever Evolving Frontier in Data Mining.

McInnes, L., & Healy, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. https://doi.org/10.48550/arXiv.1802.03426

Milani, A. M. P., Loges, L. A., Paulovich, F. V., & Manssour, I. H. (2021). PrAVA: Preprocessing profiling approach for visual analytics. *Information Visualization*, *20*(2-3), 101-122. https://doi.org/10.1177/14738716211021591

Putrevu, J., & Mertzanis, C. (2023). The adoption of digital payments in emerging economies: challenges and policy responses. *Digital Policy, Regulation and Governance*, *26*. https://doi.org/10.1108/DPRG-06-2023-0077

Salmanian, P., Chatzimparmpas, A., Karaca, A. C., & Martins, R. (2024). *DimVis: Interpreting Visual Clusters in Dimensionality Reduction With Explainable Boosting Machine*. https://doi.org/10.2312/mlvis.20241125

Schmidl, S., Wenig, P., & Papenbrock, T. (2022). Anomaly detection in time series: a comprehensive evaluation. *Proc. VLDB Endow.*, *15*(9), 1779–1797. https://doi.org/10.14778/3538598.3538602

Soramäki, K., & Cook, S. (2013). SinkRank: An Algorithm for Identifying Systemically Important Banks in Payment Systems. *Economics*, *7*(1). https://doi.org/10.5018/economics-ejournal.ja.2013-28

Triepels, R., Daniels, H., & Heijmans, R. (2017). *Anomaly Detection in Real-Time Gross Settlement Systems*. https://doi.org/10.5220/0006333004330441

van der Maaten, L., & Hinton, G. (2008). Viualizing data using t-SNE. *Journal of Machine Learning Research*, *9*, 2579-2605.

Yang, J., Yang, X., & Zhang, Z. (2022). A High-dimensional Anomaly Detection Algorithm Based on IForest with Autoencoder. 2022 4th International Conference on Data-driven Optimization of Complex Systems (DOCS)

APPENDIX A

Table A.1 - Data Description

Feature	**Description**	**Type of feature**
Ref	Transfer Reference	varchar(50)
Ord	Sender's PSP	varchar(4)
Ben	Receiver's PSP	varchar(4)
PayID	PISP Payment ID	varchar(36)
BICOrd	BIC of the Sender's PSP	varchar(11)
BICBen	BIC of the Receiver's PSP	varchar(11)
PasOrd	Country of the Sender's PSP	varchar(2)
PasBen	Country of the Receiver's PSP	varchar(2)
LEIOrd	LEI of the Sender's PSP	varchar(20)
LEIBen	LEI of the Receiver's PSP	varchar(20)
Sch	Scheme	int
Pro	Processor	int
DtLiq	Settlement Date	date
DtPISP	PISP Operation Acceptance Date	date
TsOrd	Sender's Timestamp	datetime
TsBen	Receiver's Timestamp	datetime
TipTR	Type of Transfer	varchar(1)
Div	Transaction Currency	varchar(3)
Mont	Transaction Amount (in euros)	decimal(17,2)
MontOrg	Amount in Original Currency	decimal(17,2)
TipCan	Channel Type	int

FormEnv	Transmission Format	int
OperElet	Electronic Operation	varchar(1)
OperRem	Remote Operation	varchar(1)
OperECom	E-commerce Operation	varchar(1)
IniPISP	PISP Initiated	varchar(1)
ModAc	Access Model	int
SCA	SCA - Strong Customer Authentication	varchar(1)
MNonSCA	Non-SCA Reason	int
O_SI	Sender’s Client's Institutional Sector	varchar(30)
O_PasMC	Sender’s Client's Address Country	varchar(2)
O_PasNC	Sender’s Client’s Nationality Country	varchar(2)
O_CPosM	Postal Code of the Sender’s Client [if PT]	varchar(4)
O_TipDoc	Sender’s Type of Legal Entity ID Document [if company]	varchar(15)
O_NumDoc	Sender’s Legal Entity Identification Number [if company]	varchar(255)
O_LEIC	Sender’s Client’s LEI [if company]	varchar(20)
O_ENI	Sender’s Sole Proprietor	varchar(1)
B_SI	Receiver’s Client's Institutional Sector	varchar(30)
B_PasMC	Receiver’s Client's Address Country	varchar(2)
B_PasNC	Receiver’s Client's Nationality Country	varchar(2)
B_CPosM	Postal Code of the Receiver’s Client [if PT]	varchar(4)
B_TipDoc	Receiver’s Type of Legal Entity ID Document [if company]	varchar(15)
B_NumDoc	Receiver’s Legal Entity Identification Number [if company]	varchar(255)
B_LEIC	Receiver’s Client’s LEI [if company]	varchar(20)

B_ENI	Receiver's Sole Proprietor	varchar(1)
InfUltBen	Information on Final Beneficiary	varchar(11)
InfUltOrd	Information on Final Originator	varchar(11)
CatMotTR	Transfer Reason Category (ISO Code)	varchar(4)
MotISO	Reason (ISO Code)	varchar(4)
CodOper	Operation Code (TNS or SCT)	int
SYS_Count_RT	Return Indicator	int
SYS_Count_FR	Fraud Indicator	int
SYS_Count_PR	Loss Indicator	int
SYS_O_Ot	Reported from Sender's PSP Perspective	int
SYS_B_Ot	Reported from Receiver's PSP Perspective	int
SYS_I_Ot	Reported from PISP's Perspective	int
SYS_P_Ot	Reported from Processor's Perspective	int
SYS_OtNumber	Number of Reporting Perspectives Indicator	int
SYS_Status	Record Status	int
CreatedDate	Creation Date	datatime
UpdatedDate	Update Date	datatime

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa