Anitha R

Data Quality Assessment and Preparation Report

Project: Financial Document Anomaly Detection using SAP Data
Date: April 28, 2025
Phase: 1 - Data Quality Assessment and Preparation
Dataset: https://www.kaggle.com/datasets/sunithasiva/sap-dataset

1. Executive Summary

This report details the data quality assessment and preparation performed on core SAP financial table extracts (BKPF, BSEG, FAGLFLEXA), essential for understanding transactional integrity and enabling reliable financial analysis. Initial assessment revealed critical data quality issues, including duplicate document headers and line items which undermine financial accuracy, and significant financial imbalances in the classic BSEG ledger view. The remediation strategy involved leveraging the financially balanced New G/L table (FAGLFLEXA) alongside the document headers (BKPF), and programmatically removing exact duplicate records to ensure each business transaction is represented uniquely. Post-cleaning verification confirmed the successful resolution of these issues, resulting in a clean, consistent, and balanced dataset reflecting accurate double-entry postings, suitable for subsequent analysis and the development of anomaly detection models focused on identifying potential process deviations or errors.

2. Objective

The primary objective of this phase was to assess the quality, completeness, and suitability of the provided SAP financial data extracts (representing tables BKPF, BSEG, FAGLFLEXA) for building a robust anomaly detection model capable of identifying unusual financial postings or potential business process irregularities. This involved:

- Loading the transactional data into a Python environment.
- Performing comprehensive data quality checks focused on ensuring the completeness, consistency, validity, and uniqueness necessary for reliable financial reporting and analysis within an SAP context.
- Identifying and documenting data quality issues that could impact downstream analysis or model performance.
- Implementing a cleaning strategy to remediate identified issues and align the data with expected SAP data structures and accounting principles.
- Verifying the quality of the cleaned dataset to ensure its integrity and fitness for purpose.

Anitha R

- Producing a final, trustworthy dataset ready for feature engineering aimed at detecting financial anomalies.

3. Data Sources
The analysis utilized the following CSV data extracts, representing standard SAP financial tables central to Financial Accounting (**FI**) and Controlling (**CO**):

- bkpf.csv: Accounting Document Header (Table BKPF) – Contains header information for every financial posting, including metadata like document type, dates, user, and originating transaction code.
- bseg.csv: Accounting Document Segment / Line Items (Table BSEG - Classic G/L View) – Initially considered, provides detailed line items for FI documents but lacks some New G/L enhancements.
- faglflexa.csv: General Ledger Accounting: Actual Line Items (Table FAGLFLEXA - New G/L Actual Line-Item Table) – The primary source for New G/L reporting, containing actual postings with enhanced dimensions (e.g., Segment) essential for modern financial analysis.

4. Methodology

- **Environment:** Python 3.x with Pandas and NumPy libraries. Visualizations generated using Matplotlib/Seaborn.
- **Process:**
    1. **Loading:** Initial CSV files were loaded into Pandas DataFrames.
    2. **Initial Filtering:** DataFrames were filtered to retain essential fields relevant for linking documents (keys), financial analysis (amounts, accounts, dimensions), and understanding the business context (dates, users, document types, T-codes).
    3. **Type Conversion:** Data types for key fields were explicitly converted.
    4. **Initial Data Quality Analysis:** Comprehensive DQ checks were performed targeting key aspects of SAP data integrity.
    5. **Cleaning Strategy Definition:** Based on initial DQ findings, a strategy was formulated.
    6. **Cleaning Implementation:** The strategy was executed, resulting in cleaned DataFrames (df_bkpf_clean, df_faglflexa_clean) which were exported.
    7. **Post-Cleaning Data Loading & Verification:** The cleaned CSV files were loaded.

Anitha R

8. **Post-Cleaning Data Quality Re-Analysis:** Type conversions were re-applied, and all relevant DQ checks were re-run to verify cleaning effectiveness. Visualizations were generated.
9. **Final Assessment:** Results were evaluated to confirm suitability for the next phase.

5. Initial Findings (Pre-Cleaning)

The initial data quality analysis revealed several issues impacting the reliability of the data for representing actual business transactions:

- **Loading & Filtering:** Successful.
- **Uniqueness:** Significant duplicates found in BKPF (13,166 headers) and FAGLFLEXA (40,286 line items). Duplicate financial documents distort balances, impede reconciliation, and can lead to serious audit failures. (CRITICAL ISSUE)
- **Consistency:** Checks pending deduplication. Orphaned items (items without headers) would represent incomplete financial transactions. Redundancy checks were initially blocked.
- **Population & Genuineness:** While FAGLFLEXA dimensions (Cost Center, Profit Center, Segment) appeared populated, confirmation was needed to ensure these crucial fields for management reporting and profitability analysis contained genuine values, not placeholders.
- **Numeric Stats & Balance:**
    - FAGLFLEXA (hsl - Local Currency Amount): Sum close to zero, aligning with the fundamental double-entry accounting principle expected in the general ledger. Large range/standard deviation typical of diverse financial postings.
    - BSEG (dmbtr - Local Currency Amount): Showed a significant non-zero sum. This financial imbalance indicates the BSEG extract does not represent a complete or accurate set of balanced financial transactions, rendering it unsuitable for reliable analysis. (CRITICAL ISSUE for BSEG)
- **Value Distributions:** Distributions for fields like blart (Document Type, e.g., WE=Goods Receipt, RE=Invoice Receipt, AB=Accounting Document) and tcode (Transaction Code, e.g., MIRO, VF01, MB01) provided expected context on the underlying business processes generating the financial postings.

6. Cleaning Strategy & Implementation

Based on the critical issues impacting data trustworthiness, the following strategy was adopted:

Anitha R

1. **Discard BSEG:** Due to its financial imbalance failing to reflect double-entry principles, the BSEG extract was removed to avoid propagating inaccurate financial data.

2. **Focus on BKPF and FAGLFLEXA:** Utilize BKPF for header context and FAGLFLEXA as the reliable, balanced source for New G/L actual line items and enhanced dimensions.

3. **Handle Duplicates:** Programmatically remove exact duplicate rows using .drop_duplicates(), keeping the first instance of each unique record based on:

   - BKPF primary key: (bukrs, belnr, gjahr) to ensure each document header is unique.

   - FAGLFLEXA primary key: (rbukrs, docnr, ryear, docln) to ensure each line item is unique.

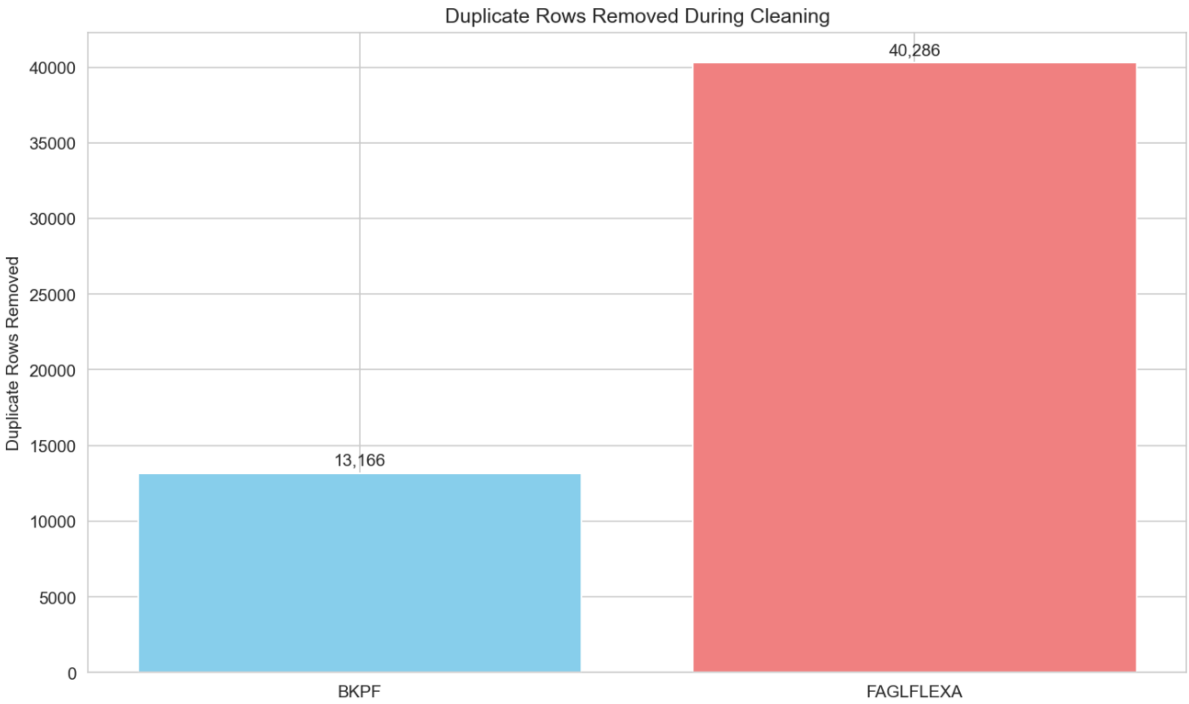
   - *(This ensures one consistent representation per document/line item in the dataset.)*

   - **Quantitative Impact:**

     - **BKPF:** 13,166 duplicate header rows removed.

     - **FAGLFLEXA:** 40,286 duplicate line-item rows removed.

     - *(See Figure: Duplicate Rows Removed During Cleaning)*

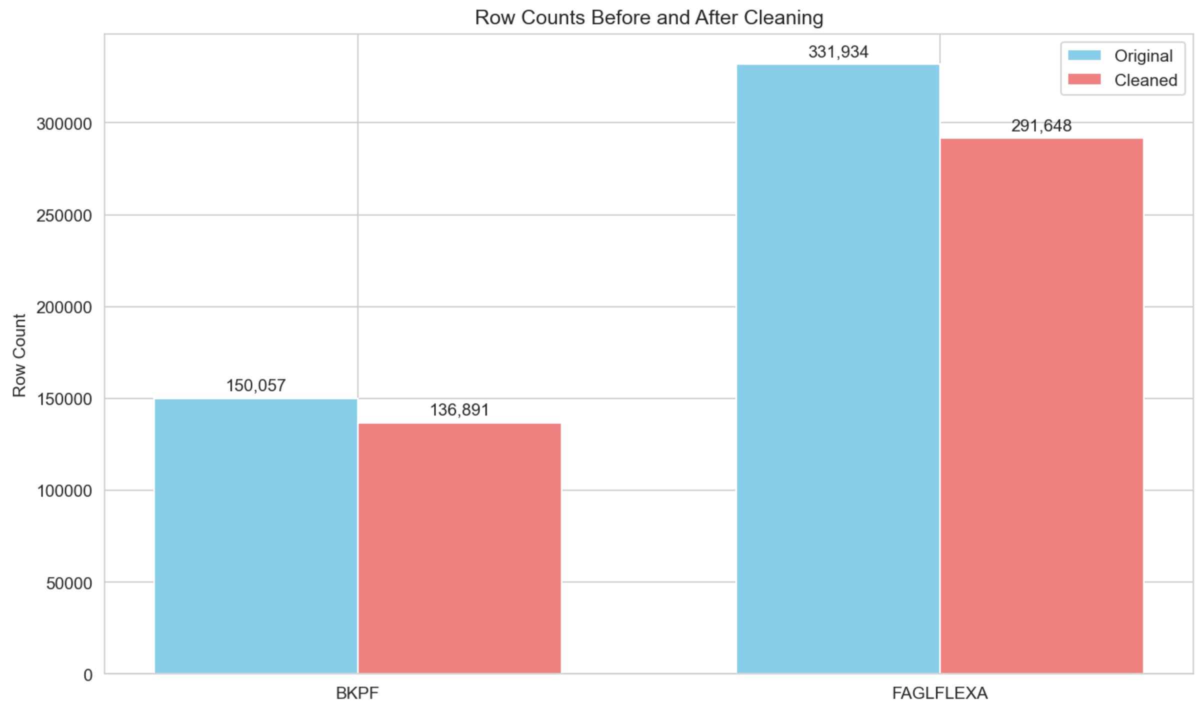


   - **Resulting Data Size:**

     - **Cleaned BKPF:** 136,891 unique headers.

Anitha R

    - **Cleaned FAGLFLEXA:** 291,648 unique line items.
    - *(See Figure: Row Counts Before and After Cleaning)*

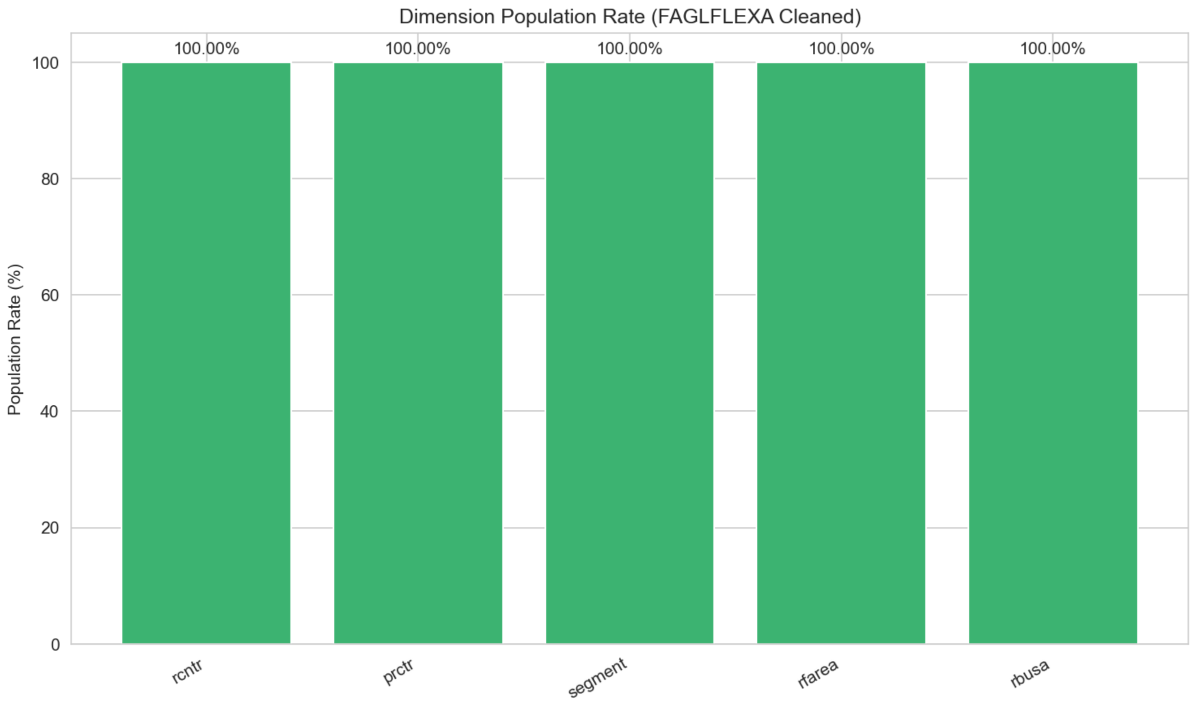


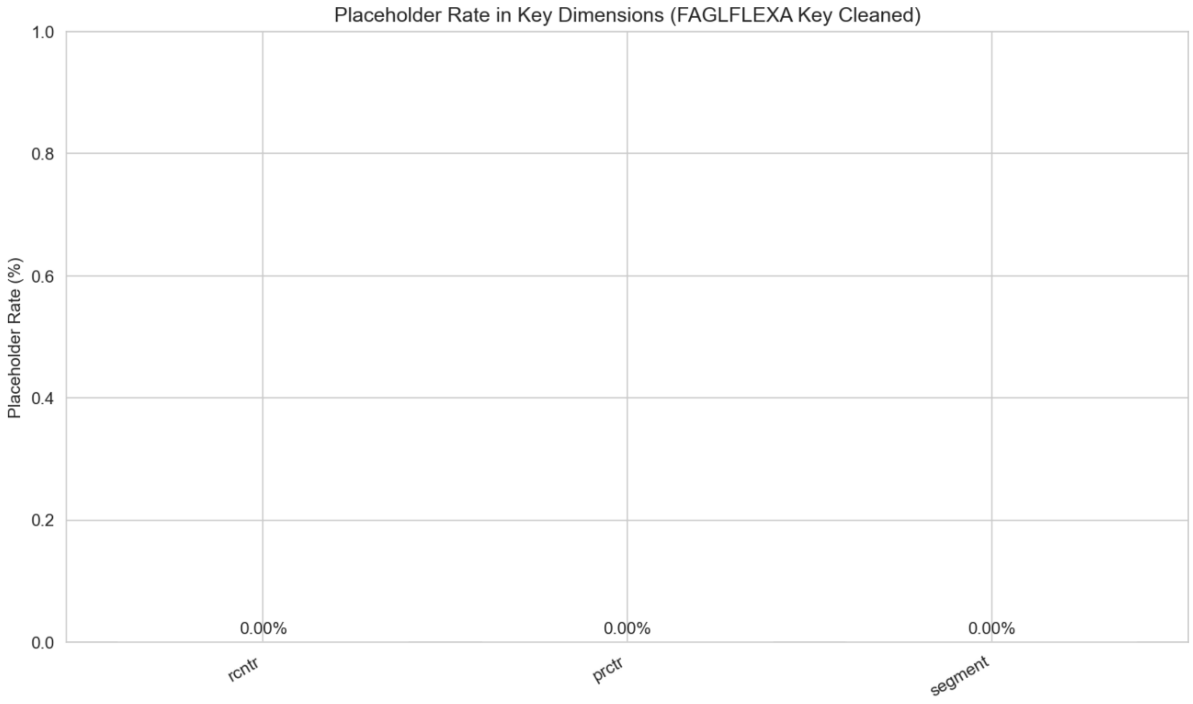
7. Post-Cleaning Analysis Results & Verification

Re-analysis of the cleaned BKPF and FAGLFLEXA datasets confirmed the effectiveness of the cleaning strategy:

- **Data Loading:** Cleaned files loaded successfully.
- **Uniqueness Verification:** Both BKPF and FAGLFLEXA confirmed 0 duplicates based on their primary keys (OK). Each financial document and line item is now uniquely represented.
- **Consistency Verification:** 0 orphaned FAGLFLEXA items found (OK). All line items correctly link to a unique document header, representing complete transactions.
- **Population & Genuineness Verification:** FAGLFLEXA dimensions (rcntr, prctr, segment, etc.) maintained 100.00% population with 0.00% placeholders found (OK). This confirms the availability of genuine dimension data crucial for accurate management reporting and analysis (e.g., profitability by segment, cost center accounting). *(See Figures: Dimension Population Rate, Placeholder Rate)*

Anitha R

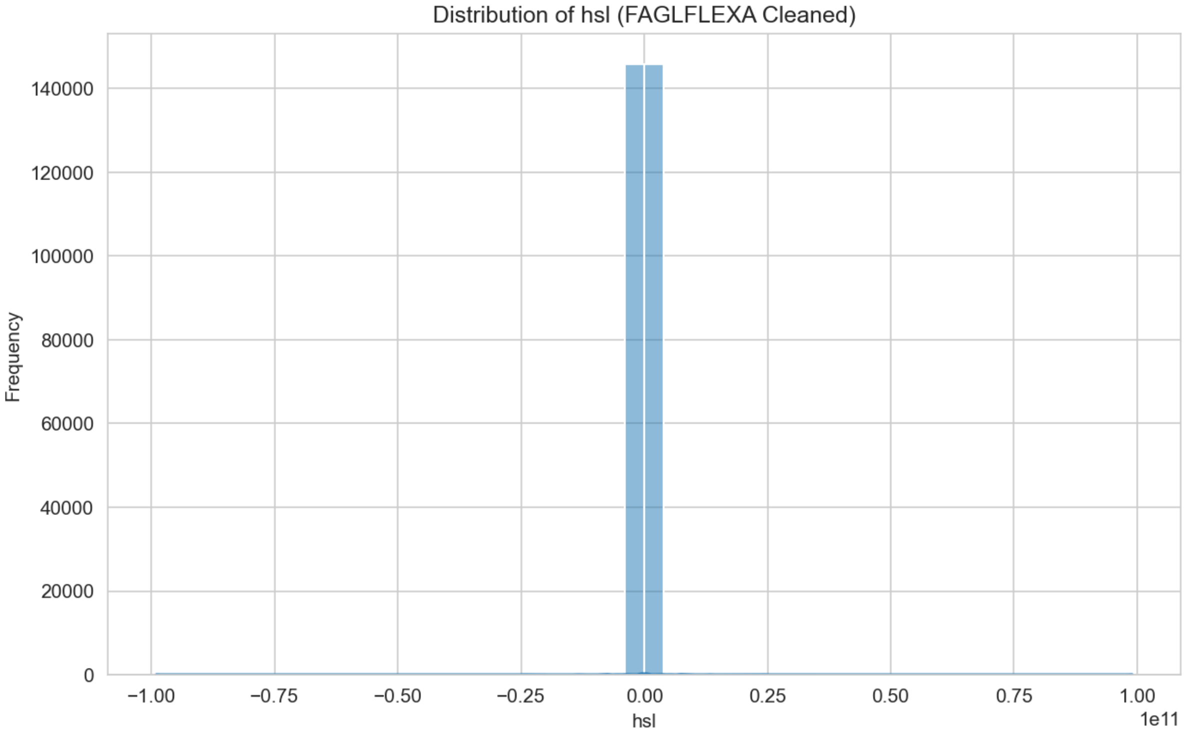




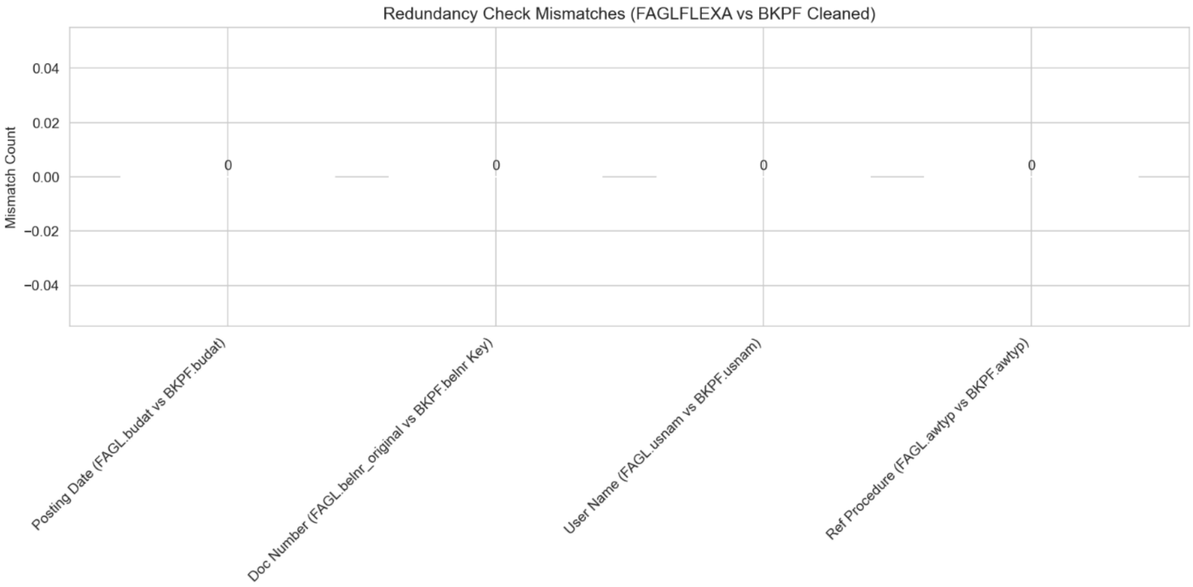
- **Value Distributions (Cleaned):** Distributions remained reasonable, reflecting typical business transaction patterns.

Anitha R

- **Numeric Stats & Balance Verification:** FAGLFLEXA (hsl) sum remained zero (OK: Balanced), confirming adherence to double-entry accounting principles. *(See Figure: Distribution of hsl)*



- **Redundancy Verification:** 0 mismatches found between cleaned BKPF and FAGLFLEXA for key overlapping fields (Posting Date, User Name, Reference Procedure, original Document Number) (OK). High consistency reinforces trust in the relationship between header and line item data. *(See Figure: Redundancy Check Mismatches)*



Anitha R

8. Final Assessment

The implemented cleaning strategy successfully addressed the critical data quality issues:

- Duplicate financial document headers and line items, which could lead to misstated reporting, were eliminated.
- Referential integrity, ensuring complete transactional data, is confirmed.
- The selected ledger data (FAGLFLEXA) accurately reflects balanced double-entry accounting.
- Crucial dimensions required for meaningful financial and management analysis are confirmed to be present and contain genuine values.
- Consistency between header and item data attributes is validated.

9. Conclusion

The data quality assessment identified critical issues impacting the reliability of the initial SAP extracts. By discarding the imbalanced BSEG data and systematically removing duplicate records from BKPF and FAGLFLEXA, a clean, consistent, and financially balanced dataset was achieved.

The resulting cleaned dataset, comprising df_bkpf_clean and df_faglflexa_clean, provides a trustworthy foundation representing unique business transactions and is deemed suitable for proceeding to the next phase: identifying potential anomalies through Exploratory Data Analysis and Feature Engineering.

Anitha R

10. Next Steps

- **Exploratory Data Analysis (EDA):** Analyze the cleaned, combined dataset to understand typical patterns in financial postings across different dimensions, time periods, users, and transaction types.
- **Feature Engineering:** Develop features based on the EDA findings and SAP business process knowledge. These features should aim to quantify potentially anomalous characteristics of financial documents (e.g., unusual amounts for account/user combinations, postings outside typical hours, missing expected dimensions for certain transactions, deviations from historical patterns).