Phase 2 Report: Exploratory Data Analysis & Feature Engineering

Project: Financial Document Anomaly Detection using SAP Data
Date: April 28, 2025
Phase: 2 - Exploratory Data Analysis (EDA) & Feature Engineering
Input Data:

- bkpf_cleaned_after_step1_20250428_165811.csv (Cleaned Headers from Phase 1)
- faglflexa_cleaned_after_step1_20250428_165811.csv (Cleaned Line Items from Phase 1)
  Dataset Origin: https://www.kaggle.com/datasets/sunithasiva/sap-dataset

1. Executive Summary

This report details the Exploratory Data Analysis (EDA) and Feature Engineering phase, building upon the cleaned SAP financial data (BKPF headers, FAGLFLEXA line items) prepared in Phase 1. The EDA focused on understanding baseline patterns within the reliable dataset, examining transaction timing, user activity, account usage, process context (Document Types, T-Codes), and monetary values, aided by visualizations. Key insights revealed typical business hour posting distributions (concentrated 8 AM - 1 PM), clear weekly (minimal weekend) and monthly (mid-month and near month-end peaks) posting cycles, dominance by specific users (BHUSHAN, MOUNIKAPALI) and integrated MM/SD transaction codes/document types (e.g., MIRO/RE, MB01/WE), and expectedly wide, balanced value ranges for financial postings (hsl) where normalcy strongly depends on context (user, doc type). Based on these insights, a set of 16 features was engineered to quantify potential deviations from normalcy, such as unusual posting times, outlier amounts relative to user or account history (using log-deviation), rarity of user-transaction code combinations, and contextual flags (e.g., missing cost objects for presumed expenses - 0 instances found). This process resulted in a feature-rich dataset (sap_engineered_features.csv) containing 291,648 records and 30 columns, forming a robust input for the subsequent anomaly detection modeling phase (Phase 3).

2. Objectives

- Perform Exploratory Data Analysis (EDA) on the cleaned and merged BKPF/FAGLFLEXA dataset to gain insights into data characteristics and typical patterns.
- Engineer a set of relevant features based on EDA findings and SAP domain expertise to capture potential anomaly indicators.
- Produce a final dataset enriched with these features for Phase 3.

3. Methodology

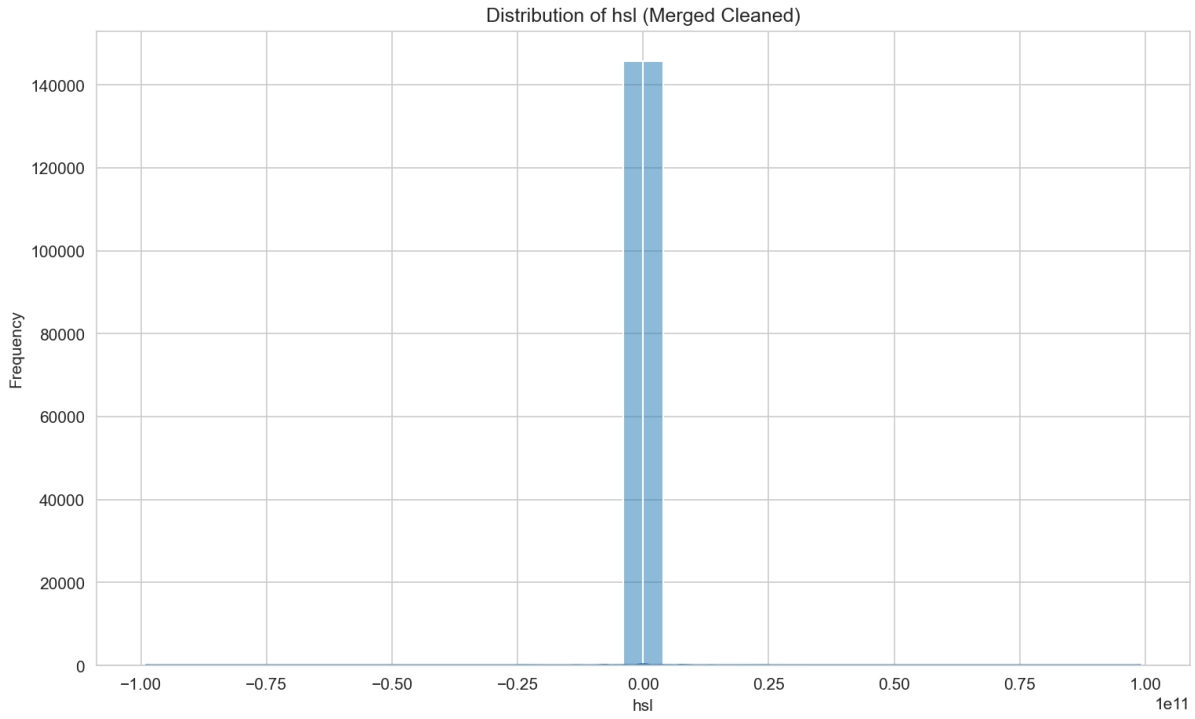
- **Environment:** Python 3.12 with Pandas, NumPy, Matplotlib, and Seaborn.

- **Process:** Loaded cleaned data; Re-applied type conversions; Renamed keys and handled potential belnr conflict for merging FAGLFLEXA and BKPF; Performed EDA with visualization generation; Defined and implemented feature engineering strategy; Prepared and exported the final dataset.

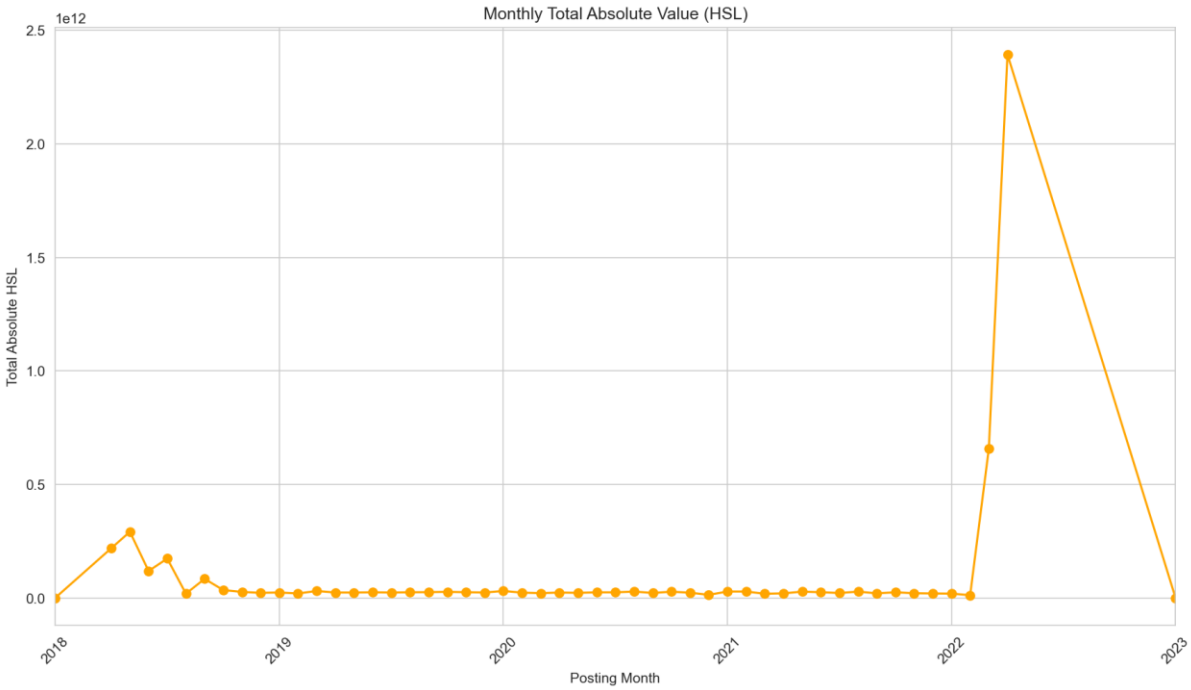
4. Exploratory Data Analysis (EDA) Findings

Visualizations were generated to aid in understanding distributions and patterns within the cleaned, merged dataset. Plots were saved to the EDA_Plots_20250428_205234 directory.

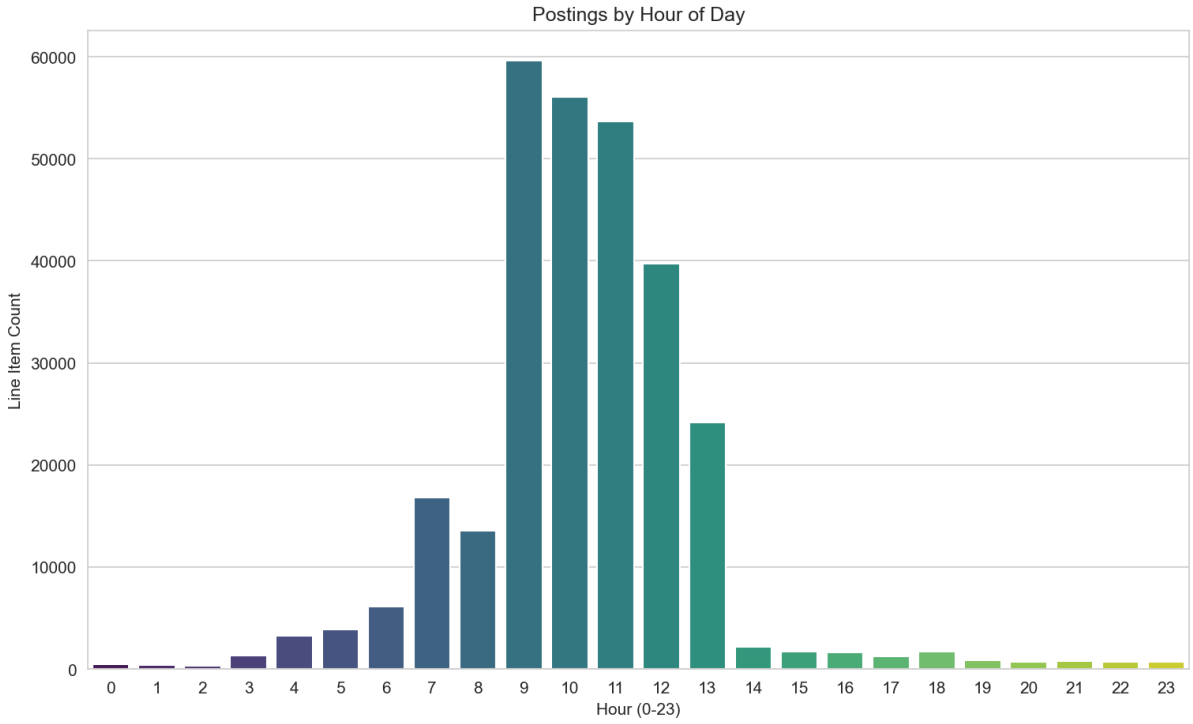
- Loading & Merging:
    - Successfully loaded bkpf_cleaned_after_step1_20250428_165811.csv (136,891 rows) and faglflexa_cleaned_after_step1_20250428_165811.csv (291,648 rows).
    - Merge successful, resulting in merged_df (291,648 rows, 37 columns). Verification confirmed 0 orphaned FAGLFLEXA rows (OK).
- Transaction Volume & Value:
    - HSL Distribution: The histogram confirmed an extremely wide range (Min/Max ≈ +/- 9.9e+10), typical for financial data, with values highly concentrated near zero, necessitating log transformations or relative deviation features for effective analysis. *(See Figure: merged_hsl_distribution.png)*



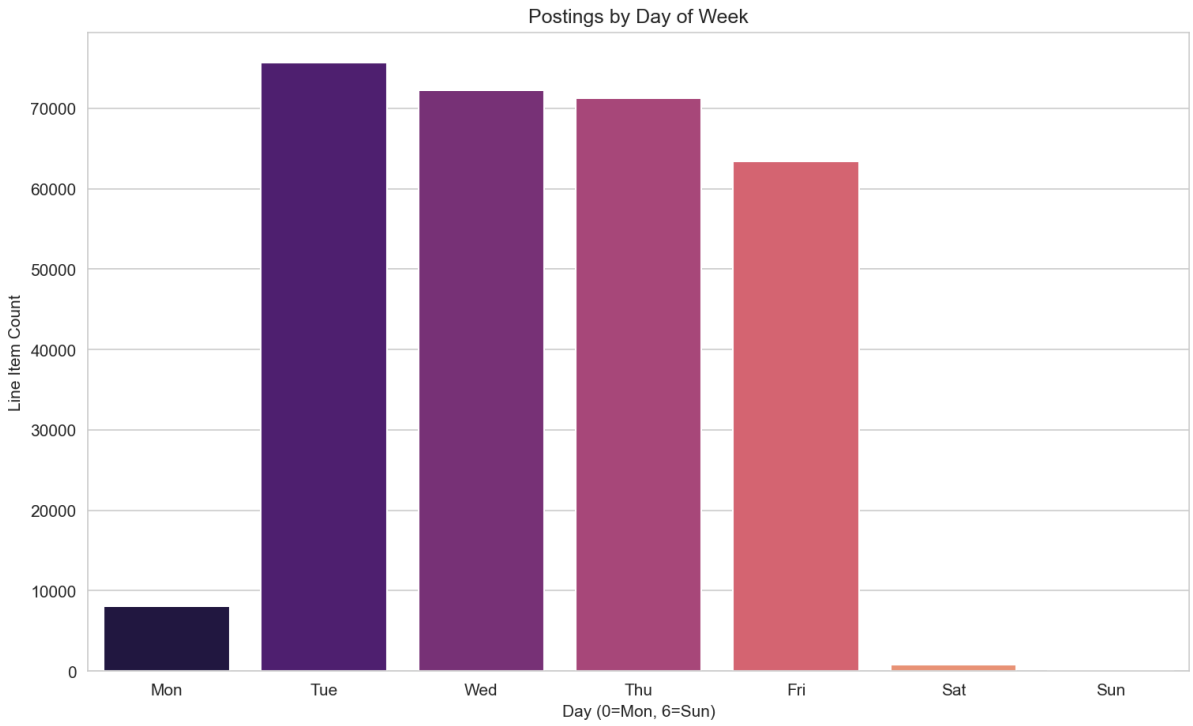
- Monthly Activity: Line plots reveal distinct temporal patterns. Line item volume shows recurring peaks, suggesting seasonality or closing cycles *(See Figure: monthly_volume.png)*. A significant value spike occurred around mid-late 2022, indicating potential regime shifts or significant business events *(See Figure: monthly_value.png)*.



- Timing:
  - Posting Hour (based on Entry Time): Activity peaks strongly during standard business hours (8 AM - 1 PM), with notable secondary activity from 5 AM to 7 AM. Activity is minimal overnight, validating time-based features as postings outside typical hours are suspicious. *(See Figure: posting_hour_distribution.png)*

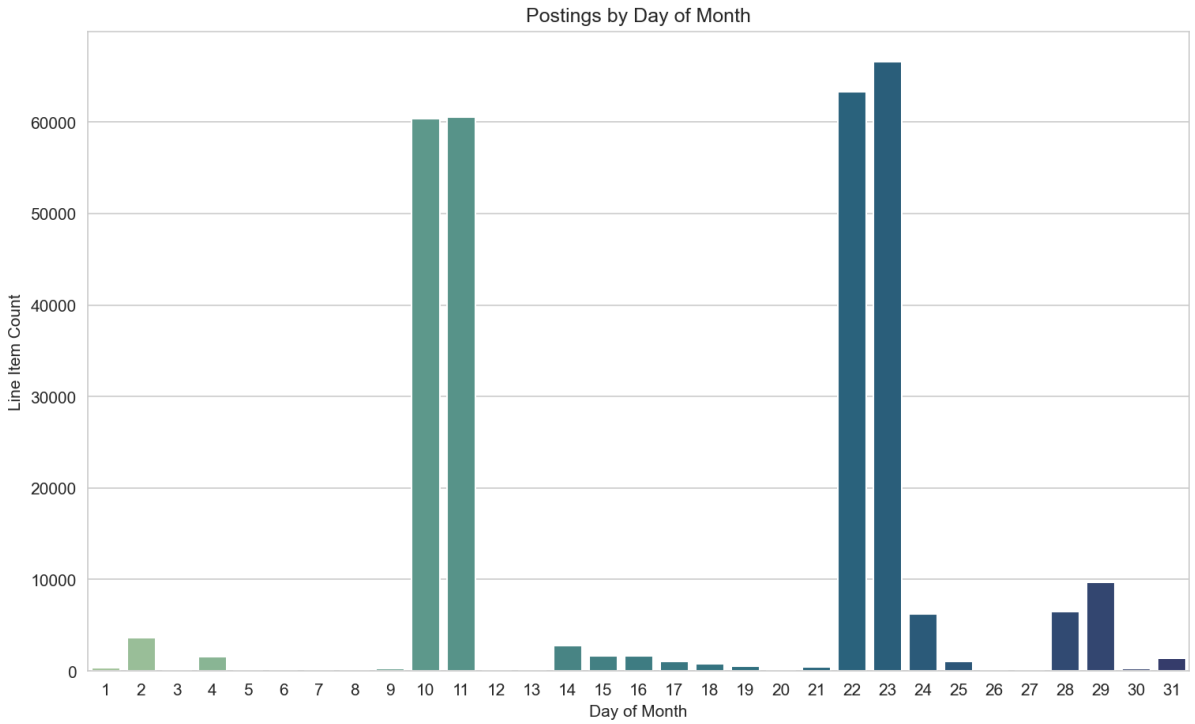


- Posting Day of Week: A clear business week pattern emerges, with activity concentrated Tuesday-Thursday and minimal activity on Saturday and Sunday. Weekend postings are strong candidates for anomalous events. *(See Figure: posting_dayofweek_distribution.png)*

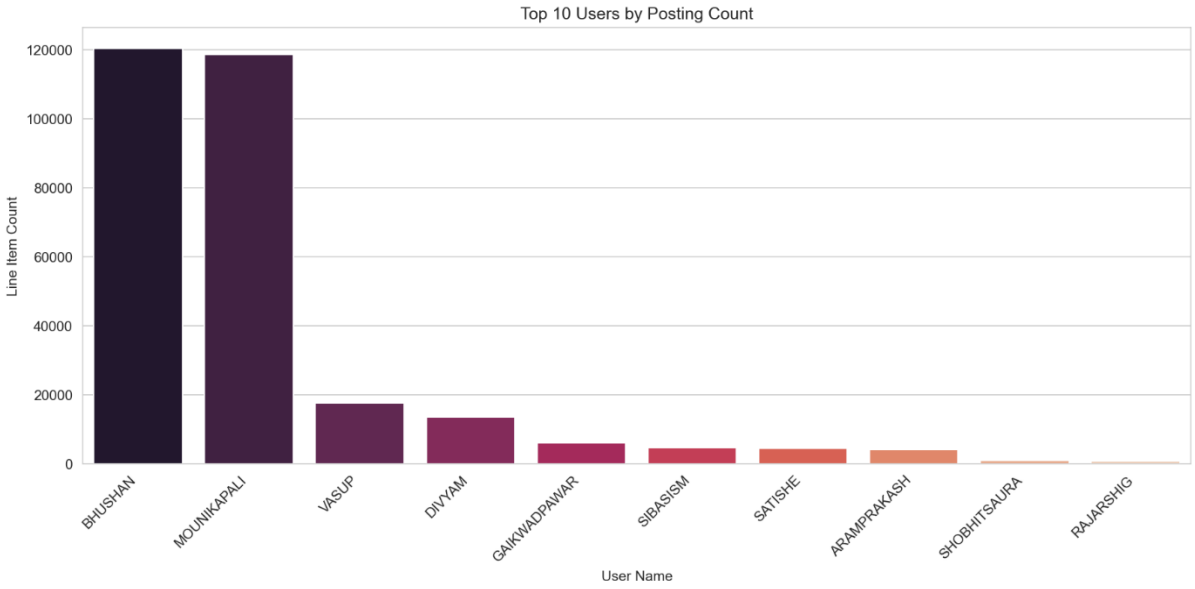


- Posting Day of Month: Activity is not uniform, showing distinct peaks around the 10th-11th and 22nd-23rd, alongside increased activity towards month-

end. Significant deviations from this rhythm could indicate unusual posting behavior. *(See Figure: posting_dayofmonth_distribution.png)*

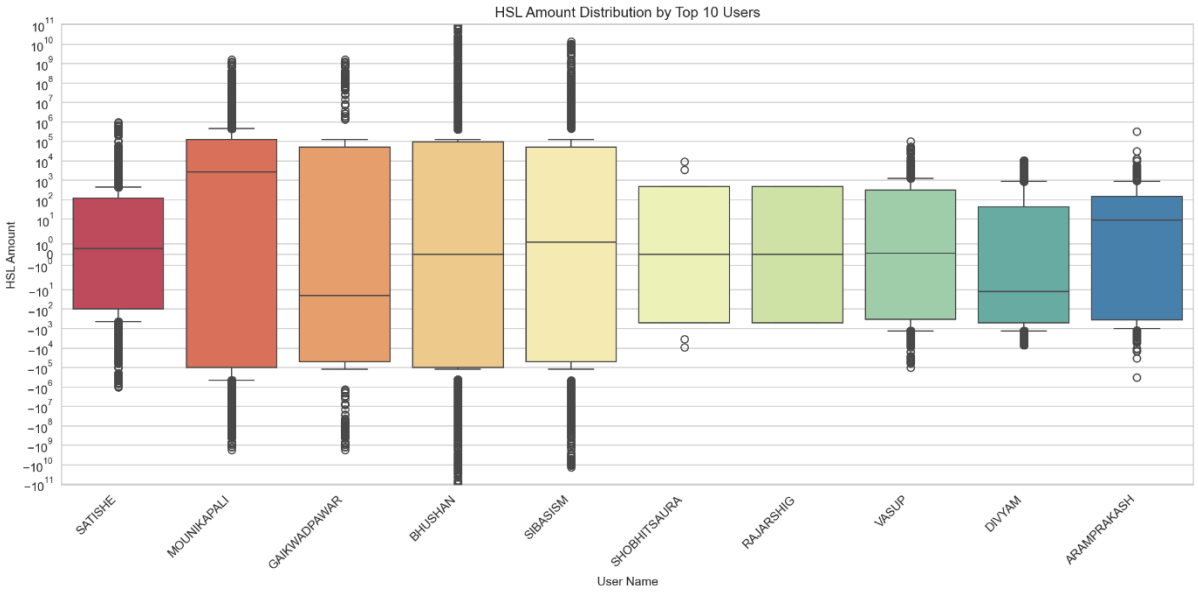


- User Activity:
    - Posting Volume: Activity is highly concentrated, dominated by users BHUSHAN and MOUNIKAPALI, with a steep drop-off for others. This supports features quantifying user frequency. *(See Figure: top_users_by_count.png)*

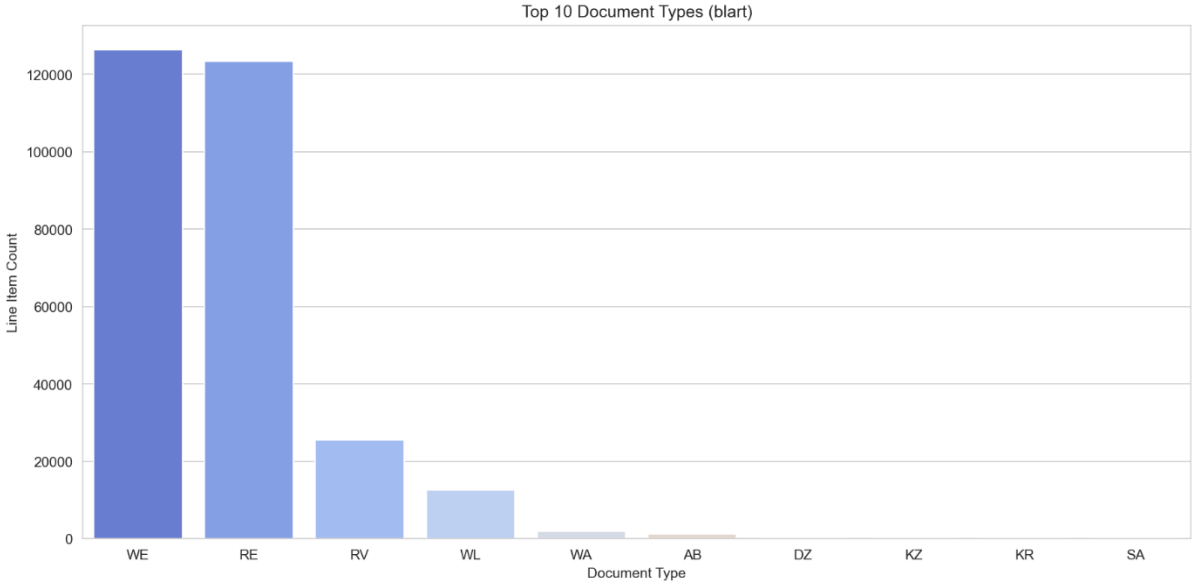


    - Amount by User: The box plot clearly demonstrates that different users have distinct posting amount profiles (median, spread, outliers), strongly justifying

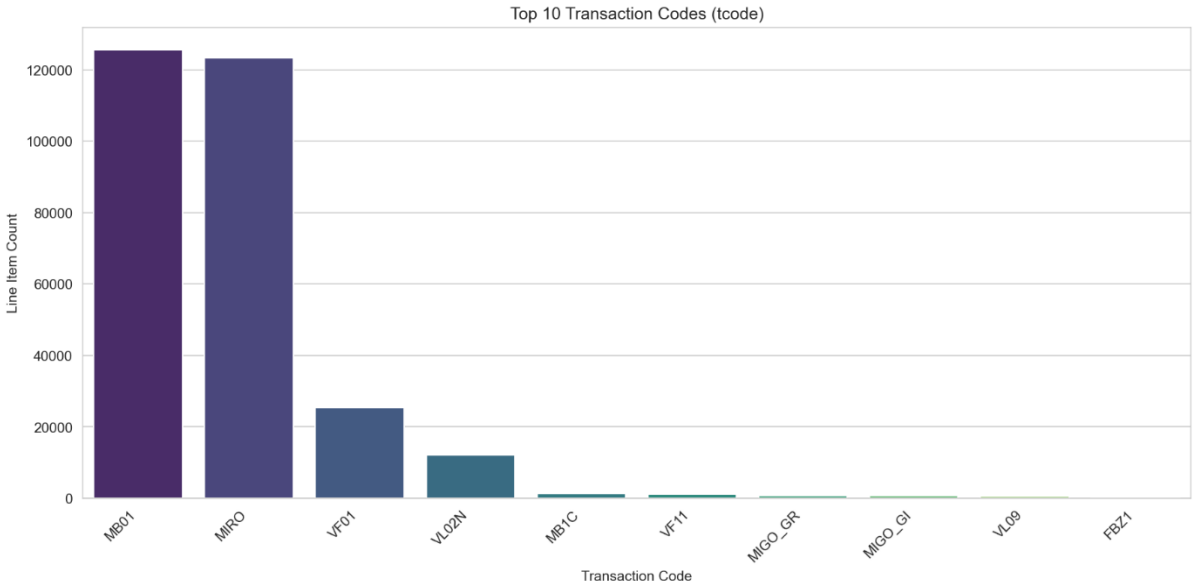
features measuring amount deviation relative to a specific user's history. *(See Figure: hsl_boxplot_by_user.png)*



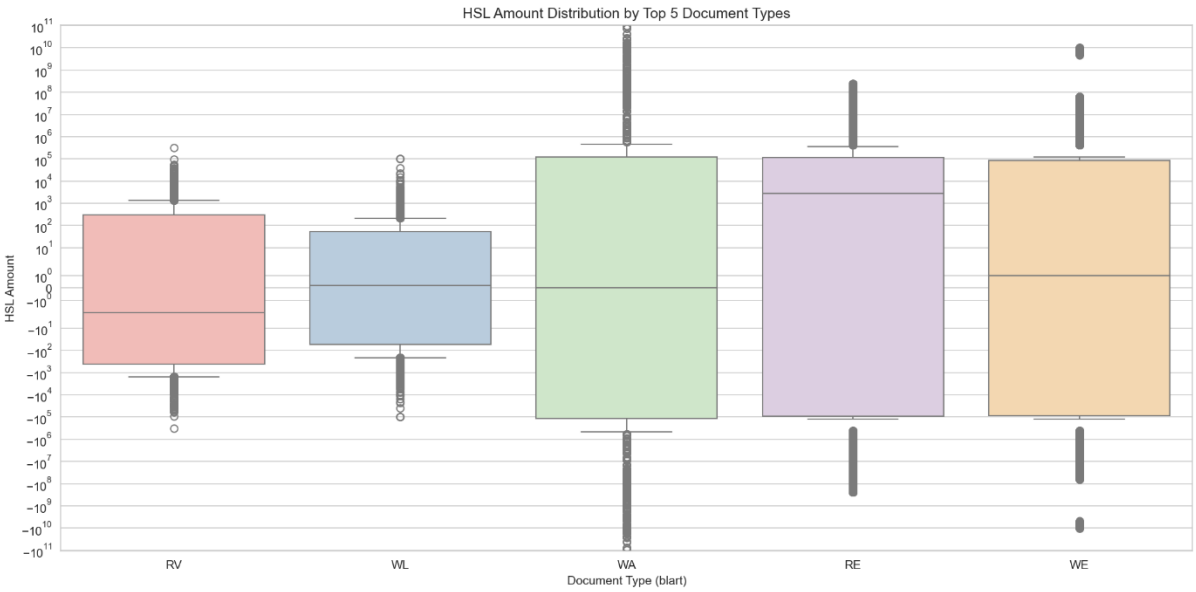
- Process Context:
    - Document Types (blart): Dominated by WE (Goods Receipt) and RE (Invoice Receipt), followed by RV (Billing Doc Transfer) and WL (Goods Issue), confirming MM and SD integration drive most postings. Rare document types are potential flags. *(See Figure: top_document_types.png)*



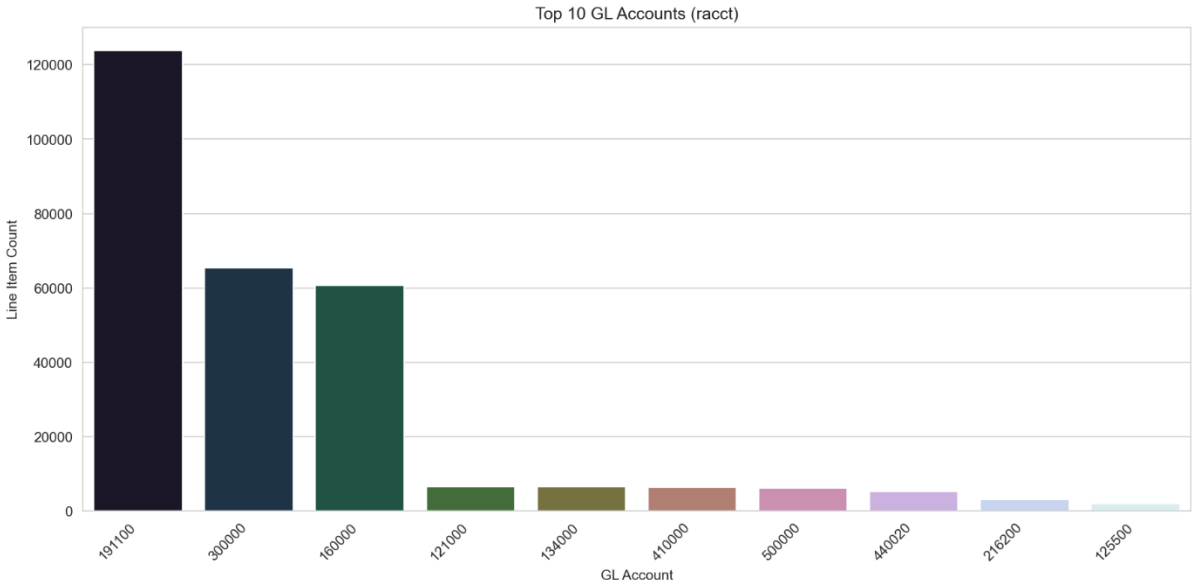
- Transaction Codes (tcode): Mirroring blart, MB01 (GR) and MIRO (Invoice) dominate, followed by VF01 (Billing) and VL02N (Delivery). Rare TCodes or unexpected User-TCode combinations are suspicious. *(See Figure: top_tcodes.png)*



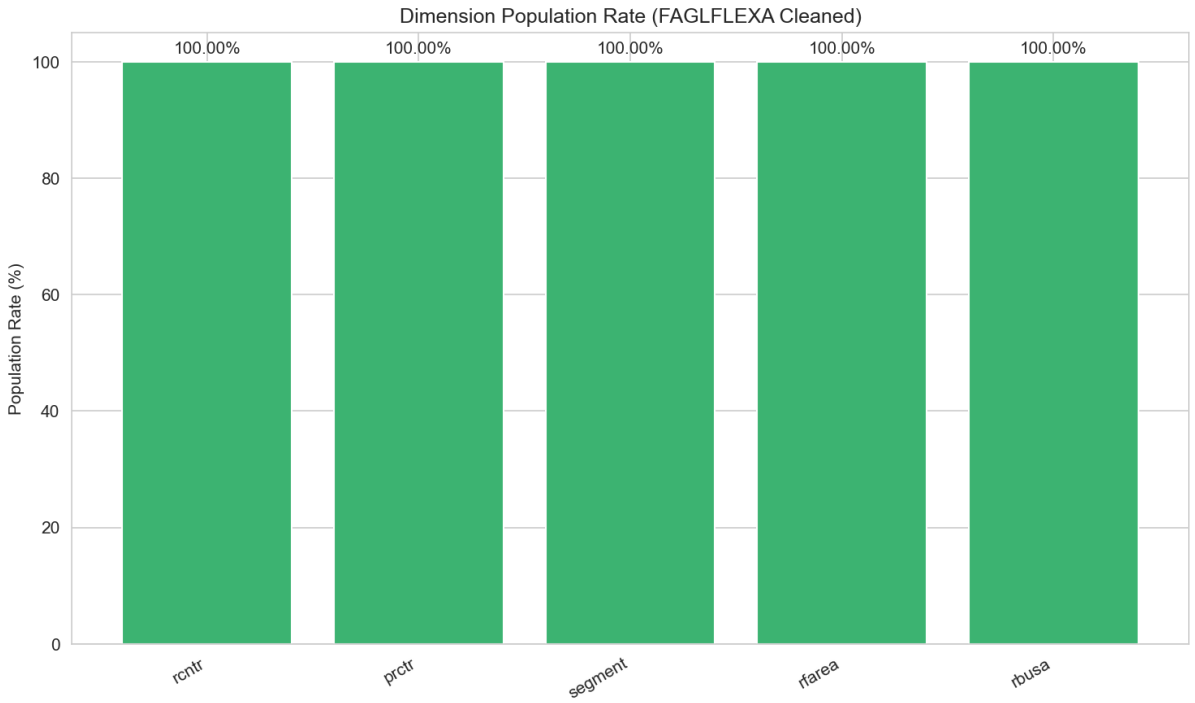
- Amount by Doc Type: The box plot showed different value profiles for top 5 document types, reinforcing that judging an amount requires knowing the process context (blart). *(See Figure: hsl_boxplot_by_blart.png)*

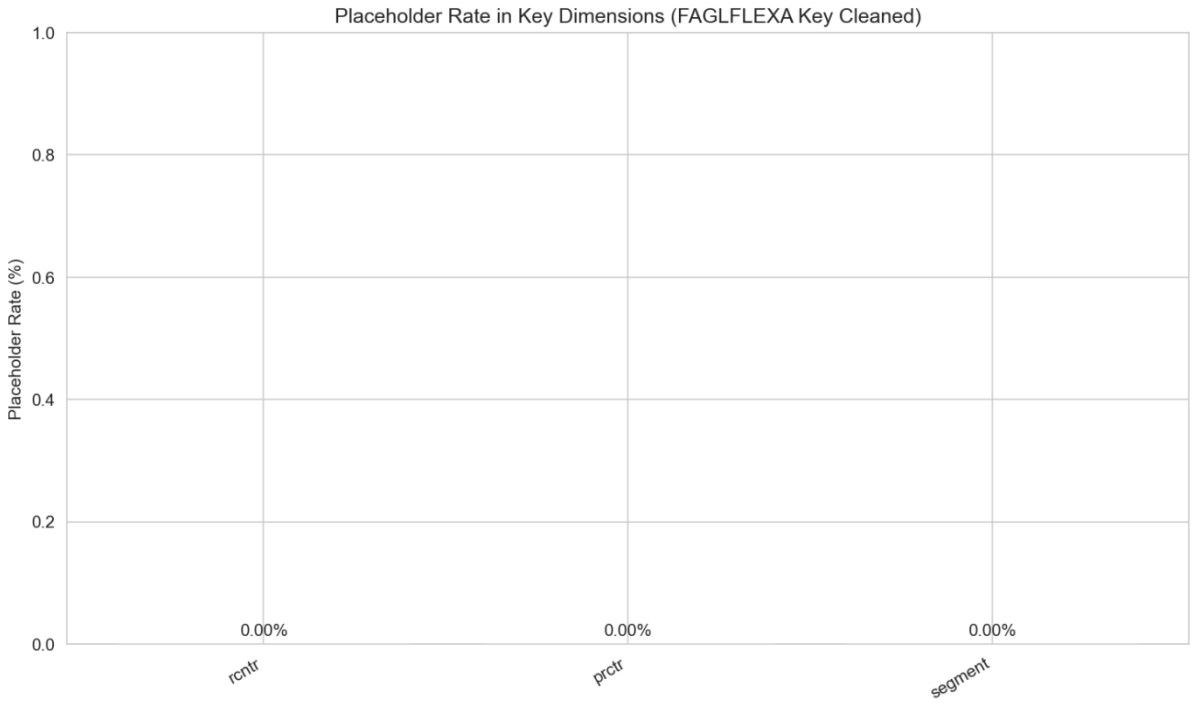


- Account & Dimension Analysis:
    - GL Accounts (racct): High activity concentrated in likely clearing/recon accounts (191100, 300000, 160000). Anomalies might occur in postings to less frequent accounts or amounts unusual for any given account. *(See Figure: top_gl_accounts.png)*

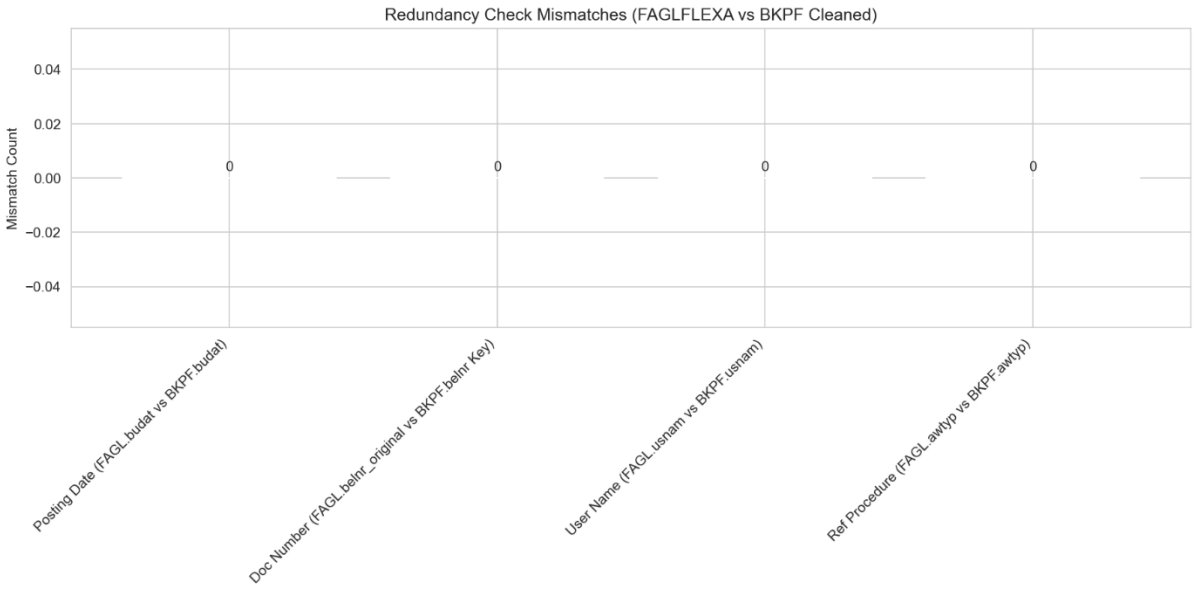


    - Dimensions: Key FAGLFLEXA dimensions (rcntr, prctr, segment) confirmed at 100% population and 0.00% placeholders, providing a rich basis for analysis. *(See Figures: faglflexa_dimension_population.png, faglflexa_key_placeholder_rates.png)*





- Redundancy Verification (Post-Cleaning):
    - Comparison of overlapping fields (Posting Date, User Name, etc.) between BKPF and FAGLFLEXA showed 0 mismatches across all compared fields. *(See Figure: redundancy_mismatches.png)*



5. Feature Engineering Strategy & Implementation

Based on EDA insights and SAP knowledge, the following 16 features were engineered:

- **Timing Features (from BKPF cpudt/cputm or budat):**
    - FE_PostingHour, FE_PostingDayOfWeek, FE_IsOutsideBusinessHours, FE_IsWeekend (Captures deviations from typical work times).
- **Magnitude Features (from FAGLFLEXA hsl):**
    - FE_AbsoluteAmount, FE_LogAmount (Provides absolute impact and a distribution more suitable for modeling).
- **User-Based Features (from BKPF usnam_bkpf, tcode):**
    - FE_UserPostingFrequency (Identifies activity level).
    - FE_UserAvgLogAmount, FE_AmountDeviationFromUserMean (Contextualizes amount based on user history).
    - FE_IsRareTCodeForUser (Flags unusual process execution by a user, threshold < 5 postings).
- **Account/Dimension-Based Features (from FAGLFLEXA racct, rcntr):**
    - FE_AccountPostingFrequency (Identifies high/low volume accounts).
    - FE_AccountAvgLogAmount, FE_AmountDeviationFromAccountMean (Contextualizes amount based on account history).
    - FE_IsMissingCostCenterForExpense: Flag (1/0) if rcntr is missing/placeholder AND racct starts with defined expense prefixes ('4','6','7').
        - *Observation:* **0 instances** found. This likely reflects strong system configuration preventing such postings in this environment, rather than an absence of the control concept itself. The feature remains structurally valid.
- **Contextual Features (from BKPF blart, tcode):**
    - FE_DocTypeFrequency, FE_TCodeFrequency (Quantify rarity of the overall process).

Features Considered but Not Implemented: IsManualPosting (requires system rules), HighValuePostingFlag (relative deviation preferred over arbitrary threshold), Text Analysis Features (scope/data availability).

- **IsManualPosting:** Requires predefined business rules or naming conventions specific to the source SAP system to map blart or usnam to manual vs. automated postings. This information was not available. *Potential Enhancement:* Could be added if system-specific rules are obtained.

- **HighValuePostingFlag:** While FE_AbsoluteAmount and FE_LogAmount capture magnitude, creating a specific flag requires defining somewhat arbitrary thresholds (e.g., top 1% of values, fixed amount like >$1M), possibly per document type (blart). This was deferred, as relative deviation features (FE_AmountDeviation...) often provide more contextualized anomaly signals. *Potential Enhancement:* Thresholds could be defined based on further analysis or business input.
- **Text Analysis Features (SGTXT):** Analyzing item text (SGTXT or similar fields) for unusual keywords (e.g., "Test", "Manual Adj.", "Plug") requires Natural Language Processing (NLP) techniques (like TF-IDF, keyword extraction). This was considered out of scope for this phase and also dependent on the availability and quality of text data (BSEG was discarded, FAGLFLEXA text fields were not included in the initial column selection). *Potential Enhancement:* If relevant text fields exist and NLP is feasible, this could add value.

6. Final Engineered Dataset Summary

- **Shape:** 291,648 rows × 30 columns.
- **Content:** Includes identifiers, selected original fields, and 16 engineered features (FE_...).
- **NaNs:** 1 NaN identified in FE_AbsoluteAmount/FE_LogAmount (minimal impact).
- **Export:** Saved as sap_engineered_features.csv.

Sample Engineered Data (First 5 Rows):
(Same sample table as provided in the previous output)


sample data output for the report.csv

7. Conclusion
Phase 2 successfully transformed the cleaned SAP financial data into a feature-rich dataset suitable for anomaly detection. The EDA, supported by visualizations, confirmed expected business patterns and highlighted areas of variability where anomalies might occur (e.g., timing, user/account context for amounts). The engineered features quantify deviations from these patterns across multiple dimensions. This engineered dataset (sap_engineered_features.csv) provides a strong foundation for building and evaluating anomaly detection models in the next phase.

The visualizations strongly support the conclusions drawn in the report. They provide clear, visual evidence of the data characteristics and patterns that underpin feature engineering strategy. We have successfully identified:

- **Baseline Norms:** Typical posting times, user activity levels, process flows (TCode/Blart), and account usage.
- **Context Dependency:** The critical importance of considering user, document type, and account when evaluating transaction amounts.
- **Potential Anomalies:** Deviations from these norms (odd times, rare processes, amounts unusual for the context) are clearly identifiable areas to target.

This visual analysis confirms the dataset is well understood and the engineered features are well-grounded in the data's behaviour.

8. Next Steps

- **Anomaly Detection Model Building & Evaluation (Phase 3):** Utilize the engineered dataset (sap_engineered_features.csv) as input for various anomaly detection algorithms. Train, tune, and evaluate models, analyzing the characteristics of detected anomalies to understand their business significance.