AI-Assisted Exploratory Data Mining — Report

Course: WIE 3007 — Data Mining (Financial Analytics Focus) Author: Generated from project workspace Date: 2025-11-28

Checklist - Dataset simulation approach documented - EDA visualizations & key statistics summarized (files referenced) - Preprocessing steps and LLM/SLM usage described - Business insights (AI-generated + heuristic) included - Example AI prompts and model responses provided

Executive summary This report documents the synthetic banking dataset generation, exploratory data analysis (EDA), AI-assisted summaries, and preprocessing performed for the WIE3007 assignment. A 500-customer monthly snapshot was simulated, then analyzed with conventional tools (pandas, seaborn) and supplemented with LLM/GenAI-generated insights. The outputs (CSV, EDA summaries, correlations, anomalies, and an AI summary) were saved under the `data/` folder and are ready to be presented in a short written report.

1. Dataset simulation (brief)

- Records: 500 customers, single month (2025-01-31) snapshot.
- File: data/sim_customers_seed_42.csv
- Key fields: customer_id, age, occupation, monthly_income, monthly_credit, monthly_debit, starting_balance, ending_balance, monthly_invest_gain, monthly_invest_return_pct, initial_credit_score, final_credit_score, transaction_category_major, customer_feedback.

Approach: - Random seeds set for reproducibility (seed = 42). Numerical fields are generated with clipped normal/exponential draws so values remain realistic (income, credits, debits, investment gains). Transaction categories and counts are sampled to produce plausible monthly spending patterns. - The generator also simulates salary credits, random transfers, investment gains (based on an "investment percentage"), daily debits, and occasional large spending events. - Textual customer feedback is produced via Generative AI (Gemini) in batched requests so that each customer receives a short (<=10 words) feedback note. The notebook includes strict prompts that request JSON-only outputs and retries on transient failures. - Controlled missingness and noise: a small fraction of entries were set to NaN for income/debits/credits (2–3%), a few credit scores were corrupted to simulate errors (1%), some feedback text removed (5%), and a handful of extreme outliers were injected for debits to evaluate anomaly detection.

2. Exploratory Data Analysis (EDA) Key numeric summaries:

- Rows: 500
- Monthly income — mean: 3866.87 (count 490, 10 missing)
- Monthly credit — mean: 4757.02 (count 485)
- Monthly debit — mean: 6235.75 (count 485)
- Starting balance — mean: 7899.51
- Ending balance — mean: 7640.06
- Monthly investment gain — mean: 127.12

- Monthly investment return (%) — mean: 0.8576
- Initial credit score — mean: 679.62
- Final credit score — mean: 678.12

Top categorical results: - Common occupations: Self-employed (62), Retail (61), Manager (58), Teacher (54), Finance (49) - Major transaction categories (top counts): shopping (100), education (86), healthcare (83) - Customer feedback: 25 entries are NaN (missing); several sample feedbacks include short actionable lines such as: "Dining out was excessive; need to cut back.", "Excellent savings this month, very happy with progress!", "Shopping spree impacted balance; will be more mindful." (these are examples present in the generated file).

Visualizations (produced in the notebook) - Missingness heatmap, histograms for income/credit/debit/ending balance, boxplots of monthly_debit by occupation, bar chart of major transaction categories, correlation heatmap for numeric features, and scatter/box/strip plots for relationships and outlier highlighting. These plots are generated inline in the notebook; to include them in the final document, export the PNGs from the notebook cells or re-run the notebook and save figures.

Anomalies - Anomalies were detected using a combined z-score (>3) and IQR rule. The anomalies include both extreme high-debit cases and a few extreme credit-score corruptions intentionally injected.

3. AI-assisted summaries (LLM outputs) The notebook attempts to use Google GenAI (Gemini) for two purposes: (a) generating short per-customer feedback strings during dataset simulation and (b) producing an overall 4–6 item insight summary that interprets EDA statistics and correlations.

Representative AI-generated summary: - Income & Credit Activity: There's a very strong positive correlation (0.94) between monthly income and monthly credit. This indicates that higher income is closely tied to increased credit activity, suggesting income significantly influences credit usage or availability. - Financial Balance Stability: The very strong positive correlation (0.91) between starting and ending balances highlights that initial financial health largely determines the final financial position. A strong starting balance is a key indicator of a healthy ending balance. - Credit Score Persistence: An individual's initial credit score has a strong positive correlation (0.70) with their final credit score. This suggests that an established creditworthiness tends to be consistent over time. - Income & Initial Financial Standing: Monthly income is strongly correlated (0.68) with the starting balance. This implies that higher earnings often contribute to a more robust financial position at the beginning of a period. - Investment Performance: Higher monthly investment return percentages are strongly linked to greater monthly investment gains (0.68 correlation). This underscores the importance of achieving a good return rate for maximizing investment growth. - Spending vs. Earning: On average, monthly debits (6235.75) significantly exceed both monthly income (3866.87) and monthly credit (4757.02). This suggests that individuals in this dataset may be spending more than they are earning or receiving through credit each month.

Interpretation (brief) - The dataset design and observed statistics produce clear relationships: income drives credit and starting balance, which then strongly predicts ending balance. Investment return % relates positively to investment gains. The average monthly debit exceeding income highlights either heavy use of credit and transfers or that the synthetic population includes households running down balances (or with high variability due to injected outliers).

4. Data preprocessing (what was done) All preprocessing steps were performed in the notebook and are summarized below; code snippets live in `sample.ipynb`.

Steps applied: - Missing value handling - Numeric columns: median imputation using sklearn.impute.SimpleImputer(strategy='median'). This addresses a small fraction of missing incomes/credits/debits. - Categorical columns: mode imputation (most frequent) applied to most categorical variables (excluding free text feedback fields). - Text fields: missing `customer_feedback` values were filled with the string "No feedback provided" before further text processing.

- Scaling and encoding
  - Numerical features were standardized (zero mean, unit variance) using sklearn.preprocessing.StandardScaler.
  - Categorical (non-text) columns were label-encoded via sklearn.preprocessing.LabelEncoder to create numeric codes for downstream modeling.
- Text processing with LLM
  - The notebook uses a batched Gemini call to classify feedback sentiment (positive/neutral/negative) and detect a topic from a closed list (savings, loan, investment, credit, dining, healthcare, utilities, shopping, education, transport, groceries, entertainment).
  - Batch calls are designed to request JSON-only outputs and perform retries on failures. Results are added back to the preprocessed frame as `feedback_sentiment` and `feedback_topic`.
- Outcome: The preprocessed DataFrame (`df_preproc`) contains standardized numerical fields, encoded categoricals, imputed text, and added LLM-derived text labels ready for modeling. The notebook displays a sample head of `df_preproc` after processing.

5. Example AI prompts and generated responses Below are the representative prompts (sanitized and paraphrased) that the notebook uses when calling a GenAI model. These are included so instructors can review the exact behavior used to generate synthetic text and labels.

A. Feedback generation prompt (used in dataset simulation, batched): - Purpose: For each customer summary line (ID and numeric summary), request a JSON array where each object contains {"customer_id": , "feedback": } and nothing else. The model was instructed to keep each feedback to approx 10 words.

B. Sentiment & topic classification prompt (used in preprocessing, batched): - Purpose: For a batch of short feedbacks, ask the model to classify sentiment as 'positive', 'neutral', or 'negative' and assign a topic from a closed taxonomy (savings, loan, investment, credit, dining, healthcare, utilities, shopping, education, transport, groceries, entertainment). The model is asked to return a JSON array with fields {"customer_id": , "sentiment": , "topic": } and nothing else.

Representative AI response is reproduced in section 3 above. Per-customer feedback examples (present in the dataset) include short actionable notes such as: - "Dining out was excessive; need to cut back." - "Utilities were costly; exploring ways to save energy." - "Excellent savings this month, very happy with progress!" - "Shopping spree impacted balance; will be more mindful."

6. Limitations, ethics, and next steps Limitations

- This is synthetic data: relationships and distributions reflect the generation design and seeded randomness rather than real-world banking populations. Use caution when generalizing results beyond the simulated cohort.
- The dataset intentionally includes injected noise and corrupt credit scores to test anomaly workflows; downstream models should either remove or explicitly model such corruption.
- LLM outputs are used for text generation and classification. These are helpful for pedagogy, but production systems should validate outputs for bias, hallucination, and accuracy.

Ethical considerations - No real user data is used. However, when applying LLMs to real financial text, ensure privacy protections, avoid exposing sensitive identifiers, and validate fairness across demographic slices.

Next steps / recommendations - Export visualizations (PNG) from the notebook and embed them into the final Word/PDF report. - Inspect data anomalies and decide which rows to drop/flag for downstream modeling. - Consider more robust text pipelines (tokenization, stopword handling) and small-model local alternatives if API access to GenAI is limited. - If modeling for debt or default risk, add temporal sequences (multiple months) and augment with demographic attributes.

End of report