Association Analysis (Apriori) GUI — User Guide

Version: 1.0 | Generated: 2025-10-08

1. Purpose

This user guide explains how to install, configure, and operate the "Association Analysis (Apriori) GUI". The application loads transactional spare-parts data from Excel, mines frequent itemsets and association rules using the Apriori algorithm (mlxtend), and provides an interactive interface to filter, sort, visualize, and export results.

2. System Requirements

• **Operating System:** Windows, macOS, or Linux with Python 3.9+ installed.
• **Python packages:** pandas, openpyxl, matplotlib, networkx, mlxtend, tkinter (standard library), and optional: openpyxl for Excel I/O.
• **Hardware:** Recommended ≥ 8 GB RAM for medium datasets; more memory improves performance on large transaction sets.
• **Display:** At least 1440×900 recommended for comfortable layout.

3. Installation

1) Install Python 3.9+.
2) Install dependencies in your environment:
   pip install pandas openpyxl matplotlib networkx mlxtend
3) Save the provided Python script as, for example, app.py.
4) Launch the application with: python app.py

4. Input Data Format (Excel)

The application expects an Excel file (.xlsx/.xls) with at least three columns in the first worksheet:
• Column 1 — Transaction ID (e.g., Service Order ID)
• Column 2 — Item (Spare part name)
• Column 3 — Order No. (e.g., internal material or order number)

Optional columns:
• Column 4 — Consumption (numeric). If present and 'Include ordered spare parts (consumption = 0)' is disabled, rows with consumption ≤ 0 are excluded.
• Column 5 — Price (numeric). If present, unit prices are derived per item (price/consumption when valid; otherwise total price). These power the cost metrics.

5. Typical Workflow

1) Select an Input file and Output file.
2) Set minimum Support and Confidence thresholds.
3) (Optional) Enable 'Include ordered spare parts (consumption = 0)'. Use 'Filters & Columns' to define include/exclude text filters and choose visible metrics.
4) Click 'Run analysis' (or press F5). Progress and messages appear in the Log panel.
5) Review results in the Results tab. Use Quick search, column sorting, and metric visibility.
6) (Optional) Open Charts: Top 20 bar chart or Rule Graph.
7) Export filtered results using 'Save filtered…'.

6. GUI Layout & Controls

6.1 Inputs Tab

• Input file (Excel): Path to the Excel workbook to analyze. [Browse…] opens a file chooser (Ctrl+O).
• Output file (Excel): Path for the rules export. [Save as…] opens a save dialog.
• Min support: Fraction (0–1] defining the minimum itemset frequency across transactions.
• Min confidence: Fraction (0–1] defining the minimum rule confidence.
• Include ordered spare parts (consumption = 0): If checked and a consumption column exists, rows with zero/invalid consumption are kept; otherwise they are removed.
• Run analysis: Executes Apriori mining with the current settings (F5). Buttons are temporarily disabled while running.
• Stop: Present but not active in this version (the analysis runs synchronously).
• Progress bar: Indicates ongoing processing.

6.2 Filters & Columns Tab

• Text filters:
  – Include only: Keep rules where any term occurs in antecedents or consequents.
  – Exclude terms: Remove rules where any term occurs in antecedents or consequents.
  – Apply filter / Reset: Apply or clear text filters.
• Visible metrics: Checkboxes to toggle additional columns in the Results table. 'Antecedents', 'Consequents', and 'Confidence' are always shown.

6.3 Results Area

• Results tab: Interactive table of rules.
  – Quick search: Live filter on antecedents/consequents (Ctrl+F). 'Clear' resets the search.
  – Save filtered…: Exports the currently visible subset (to .xlsx or .csv) (Ctrl+S).
  – Table: Click headers to sort; right-click for context menu (Copy row, Copy cell, Export visible columns…).
• Charts tab: Buttons to open matplotlib charts in separate windows: 'Show Top 20' and 'Show graph'.

• Log panel: Expand/Collapse via 'Hide log v' / 'Show log >'; 'Clear log' empties the log.
• Status bar: Shows the total rule count, visible count, and last export path.

7. Keyboard Shortcuts

Shortcut	Action
Ctrl+O	Open input file dialog (Browse...)
Ctrl+S	Save filtered results
F5	Run analysis
Ctrl+F	Focus Quick search

8. Metrics & Columns

The following columns may appear in the Results table. Visibility is controlled via the 'Visible metrics' panel (except those always visible).

• **Antecedents:** Left-hand side (LHS) item(s) of the rule; always visible.

• **Consequents:** Right-hand side (RHS) item(s) of the rule; always visible.

• **Support:** P(LHS ∪ RHS). Fraction of transactions containing all items in the rule.

• **Confidence:** P(RHS | LHS) = support(LHS ∪ RHS) / support(LHS). Always visible.

• **Lift:** confidence / support(RHS). >1 suggests positive association; ≈1 independence; <1 negative.

• **Leverage:** support(LHS ∪ RHS) − support(LHS) × support(RHS). Absolute deviation from independence.

• **Conviction:** (1 − support(RHS)) / (1 − confidence). Higher means stronger implication.

• **Zhangs Metric:** Zhang's Z measure of rule strength (bounded, symmetric properties).

• **Combination Count:** Absolute count of transactions where all LHS and RHS items co-occur.

• **Cost Antecedents (EUR):** Sum of derived unit prices for LHS items (missing prices treated as 0).

• **Cost Consequents (EUR):** Sum of derived unit prices for RHS items (missing prices treated as 0).

• **Cost Consequents x Combination Count (EUR):** Cost Consequents multiplied by Combination Count (weighted impact).

• **Cost Antecedent+Consequents (EUR):** Sum of LHS + RHS unit prices.

• **Mat_combination:** Dash-joined list of distinct order numbers (from column 3) across LHS∪RHS.

• **Mat_combination_items:** Comma-separated list of distinct item names across LHS∪RHS.

• **Mat_combination_id:** Deterministic 8-digit ID derived from Mat_combination using SHA-256 modulo 10^8 (leading zeros replaced by 9s).

• **Different Items:** Cardinality of distinct items across LHS∪RHS.

9. Charts

• Show Top 20: Aggregates cost_consequents_weighted per Mat_combination_id and displays the top 20 as a horizontal bar chart. Labels prefer Mat_combination_items (fallback: Mat_combination), truncated if long.
• Show graph: Builds a directed graph where nodes are rule sides (antecedents/consequents as strings) and edges are rules weighted by confidence. Node sizes scale with cumulative Combination Count. Edge labels show confidence.

10. Data Processing Details

• Pre-filters:
 – Include terms: Keep only rows where any term occurs in the Item column before mining (case-insensitive).
 – Exclude terms: Remove rows where any term occurs in the Item column before mining (case-insensitive).
• Consumption handling: If a consumption column exists and the 'Include ordered spare parts (consumption = 0)' checkbox is OFF, rows with consumption ≤ 0 are removed.
• Price derivation: If a price column exists, unit prices are approximated per item from price/consumption when valid; otherwise the total price is used. Missing/invalid prices are treated as zero in cost sums.
• Transactions: Transactions are grouped by the first column (Transaction ID). Items are one-hot encoded (mlxtend.TransactionEncoder) prior to Apriori.
• Rule mining: Uses mlxtend.apriori(min_support) to compute frequent itemsets and mlxtend.association_rules(metric='confidence', min_threshold=min_confidence).

11. Results Table Behavior

• Sorting: Click any header to sort ascending/descending.
• Formatting: Numeric columns are formatted—6 decimals for probabilities (support, confidence, lift, leverage, conviction, zhangs_metric); 2 decimals for cost metrics; integers for counts.

• Copy: Right-click → Copy row / Copy cell to copy tab-separated values.
• Export: 'Save filtered…' exports only the currently visible subset and columns.

12. Error Handling & Messages

• Input validation: Min support/confidence must be in (0,1]. Non-numeric inputs are rejected.
• Structural checks: Requires ≥3 columns. If include/exclude filters remove all rows, a clear error is raised.
• Empty rules: If thresholds are too strict, no rules may be found (log will inform you).
• Exceptions: Unexpected errors are printed to the Log and also shown via a pop-up message.

13. Tips & Best Practices

• Start with a low min_support (e.g., 0.001) to discover rare but relevant combinations; then tune upward for precision.
• Combine Confidence and Lift to prioritize impactful rules (e.g., confidence ≥ 0.2 and lift ≥ 1.2).
• Use text filters to focus on specific subsystems or part families.
• If cost metrics are important, ensure Price and (ideally) Consumption columns are present and numeric.

14. Known Limitations

• The analysis runs synchronously; the Stop button is a placeholder in this version.
• Graph labels use the full antecedents/consequents strings—large rulesets may produce dense plots.
• Price derivation is heuristic when consumption is missing/invalid.

15. Data Privacy

The application processes local Excel files and writes results locally. Ensure your inputs contain no personal data, or process them in accordance with your organization's data policies.

Appendix A — Default Column Mapping

• Column 1 → Transaction ID
• Column 2 → Item (name/description)
• Column 3 → Order number (material/bestellnummer)
• Column 4 → Consumption (optional)
• Column 5 → Price (optional)

Appendix B — Keyboard & Mouse Reference

- Right-click in table: copy/export options
- Ctrl+Click on table (macOS: Cmd+Click): also opens context menu