The Hidden Palette

Mining Co-Occurring Fashion Attributes

2025

RLV
Lena Choi, Robert Collis, Vaishnavi Venkataraghavan
University of Michigan, School of Information
Fall 2025

Overview

- *Briefly describe your research question or objective.*
    - **Research Question:** What attribute combinations frequently occur together across fashion products?
        - Can we uncover fashion item clusters using categorical features?
    - We want to explore fashion and find a clear story about what goes together and how trends change over time. Using the data, we'll look for items that might be commonly found together. Maybe summer items often show up as tees in bright colors, or winter items are black coats. It turns everyday fashion labels (color, season, type) into patterns that we can see and explain.
    - Why it Matters: Fashion is a multi billion dollar industry, and understanding how colors and categories rise or fall across seasons and years can help plan next season's palette, decide what to stock, and design better shopping filters and websites.
- *Explain why this topic or dataset is interesting or important.*
    - The fashion industry is constantly changing and evolving with patterns across categories, colors, and styles. By mining concurrent attributes and segmenting products, we hope to reveal how the organization of fashion items, which would inevitably be useful for recommender systems.
- *Mention any expected challenges or limitations (e.g., data sparsity, preprocessing needs, computational complexity).*
    - Our main challenge will be finding useful patterns and making sure the groups we find actually make sense. Another challenge with this dataset is the uneven year distribution. Most products cluster around a few key years, which can skew trends and make some year-to-year comparisons unreliable?
    - May need to do some additional data cleaning, especially for the “productDisplayName” category. There may be some inconsistent or noisy text, with brand names or style words.

Background / Related Work

- Retail uses "market-basket" analysis to find things that appear together and clustering to reveal style groups. We'll combine both: mine clear "if-then" rules and map products into a small space to find bigger segments, then check that the rules help explain each cluster over time. This goes a step beyond basic exploratory data analysis by tying patterns to clear, visual groups over time & seasons.

- This builds on a project Robert Collis did by building a brand similarity network by price/style/fit, but shifts from brand networks to product-attribute rules plus clustering items, with a focus on colors, seasons, and item types over time.
- Link: ThreadLines on GitHub

Data Representation & Dataset

- *Identify which of the seven data representations you are using.*
  - Itemsets: Each product in styles.csv will be treated as a set of categorical items, where every attribute-value pair becomes an "item"
  - Each product will treated as a transaction/"basket" of categorical attributes
- *Provide details about your dataset:*
  - Our data comes from the Kaggle Fashion Product Images (Small) collection. The file is in a CSV format complete with images (but we will not use those in this project). There are 10 columns, **[id, gender, masterCategory, subCategory, articleType, baseColour, season, year, usage, and productDisplayName]**. It has 44.4k rows with no missing data. The dataset is made up of 50% men, 42% women, and 8% other for gender. The primary categories are made up of 48% apparel, 25% accessories, and 26% other.
- *Source (with URL or reference)*
  - Fashion Product Images (Small)
- *Structure (number of records, attributes, time period, etc.)*
  - 10 Columns: id, gender, masterCategory, subCategory, articleType, baseColour, season, year, usage, and productDisplayName
  - 44.4K items
  - Time Period: Year(2007 - 2019), season (Fall, Winter, Spring, Summer)
- *Data format (CSV, JSON, text, etc.)*
  - CSV file format
  - Dataset does contain images (.jpg), but we will not be utilizing
- *Describe necessary preprocessing steps such as cleaning, transformation, or integration.*
  - Data Cleaning: Doesn't appear to contain any missing values. The dataset is well-formatted.
  - Text Tokenization: Brand names, style words from "productDisplayName" category
  - Binary Encoding: Convert itemsets into a transaction-item matrix for Apriori, MCA, and clustering.

Analytical Techniques

- *List the techniques you will use (at least two).*
- *Reference the corresponding week or lecture topic where each method was discussed.*

- *Explain how the techniques complement each other in addressing your data question.*

1. **Association Rule Mining (Apriori):**
    a. Apriori ties to Week 3: Mining Itemsets and Lab 2: Mining Itemsets. We'll use it to find items that go together with support, confidence, and lift.
2. **Dimensionality Reduction:**
    a. Ties to Week 4: Mining Matrix Data & Lab 3: Mining Matrix Data . We'll compress many categories into a simple low-dimensional map so similar products land near each other for easy plotting.
3. **Clustering:**
    a. Ties to Week 4: Mining Matrix Data & Lab 1: ML Modeling Refresher & Lab 3: Mining Matrix Data. We'll group products into clear segments and name each cluster by its common attributes.

****These techniques complement each other by providing a multi-layered view of fashion data: Apriori finds the rules, dimensionality reduction unveils the structure, and clustering groups similar transactions.**

Evaluation & Visualization

- *Describe how you plan to assess the quality or significance of your findings.*
- *Specify any metrics, interpretive frameworks, or visualizations you will use (e.g., heatmaps, network diagrams, time plots, or pattern summaries).*
    - Evaluation Metrics:
        - **Apriori:** Support, confidence, lift, pruning of redundant/insignificant rules
        - **Dimensionality Reduction:** Variance explained by first components
        - **Clustering:** Silhouette score
    - Visualization:
        - **MCA Map**: Scatter plot of products, colored by cluster
        - **Cluster Profiles**: Small bar charts or heatmap showing each cluster's top colors, item types, and seasons
        - **Top rules Table:** 10-15 Apriori rules with support, confidence, and lift.
        - **Color Trend Line:** Share of top colors by year, with an optional season view (Spring/Summer/Fall/Winter)

Team Plan & Timeline

I. Team Members & Responsibilities

- **Lena Choi: Project Manager**
    1. Scope, timeline, task board, standups
    2. Data governance (naming, folders), version control

3. Final quality assurance
4. Submit all deliverables (report, slides, code link)

- **Robert Collis: Visualization & Reporting Lead**
  1. Trend charts, information visualizations, captions
  2. Writes methods/results
  3. Creates slide deck/ design visual styles
  4. Presents findings
- **Vaishnavi Venkataraghavan: Data & Code Owner**
  1. Data cleaning/feature prep
  2. Implements apriori & analysis functions
  3. Implements clustering
  4. Maintains notebook, README.md file

II. Milestones: Subject to Change

- **Proposal Submission:** October 27, 2025 (Due: November 07, 2025)
- **Data Cleaning:** November 10 - 12, 2025
- **Analysis:** November 13 - 25, 2025
- **Report Writing:** December 1, 2025
- **Final Report & Video:** December 5, 2025 (Due: December 09, 2025)