

Combining Purchase Value and Product Preferences for Fashion Retail Customer Segmentation

Darren Waluya Ardianto¹, I Kadek Dwi Nuryana²

^{1,2}Universitas Negeri Surabaya, Surabaya, Indonesia

darren.22085@mhs.unesa.ac.id, dwinuryana@unesa.ac.id

ABSTRACT

The rapid growth of transactions in the fashion retail industry produces large volumes of customer data whose behavioral patterns remain difficult to exploit when segmentation is performed manually and descriptively. Uniform mass marketing is particularly ineffective for premium fashion, where preserving brand exclusivity demands a more personal approach. This study develops a customer segmentation model that does not stop at grouping customers by transaction value, but also reveals the shopping-style preferences embedded within each segment, and presents the results through an analytical dashboard for business stakeholders. Recency, Frequency, and Monetary (RFM) parameters were derived from transaction data and transformed using logarithmic scaling and a Robust Scaler to handle skewed distributions and outliers. The parameters were then grouped with K-Means clustering, whose optimal number of clusters was determined through the Elbow Method, Silhouette Score, and Davies-Bouldin Index. Within each resulting cluster, Latent Dirichlet Allocation (LDA) was applied independently to uncover hidden purchase topics, which were interpreted as fashion-style personas. The findings show that 1,152 customers were divided into four tiers, with K=4 as the optimal configuration, supported by a Silhouette Score of 0.5952 and a Davies-Bouldin Index of 0.5644. LDA identified distinct style personas across tiers, with topic coherence shaped by each segment's vocabulary diversity. The resulting Streamlit dashboard passed all 45 functional test scenarios, confirming that integrating value-based and preference-based analysis yields richer customer profiles for targeted marketing.

Keyword: Customer Segmentation, RFM Analysis, K-Means Clustering, Latent Dirichlet Allocation, Fashion Retail, Personalized Marketing, Analytical Dashboard.

Article Info:

Article history:

Received 04 08, 2026

Revised 05 12, 2026

Accepted 06 10, 2026

Corresponding Author

Darren Waluya Ardianto

Universitas Negeri Surabaya, Surabaya, Indonesia

darren.22085@mhs.unesa.ac.id

1. INTRODUCTION

Digital transformation has enabled retailers to accumulate transaction data at massive volume, yet much of it remains dark data: raw records that are stored but never analyzed, leaving valuable behavioral patterns untapped [1]. This gap is most costly in fashion retail, where uniform mass marketing is increasingly obsolete. Blanket price promotions are especially damaging for affordable luxury products, whose high-value customers equate exclusivity with social status, so aggressive discounting can erode brand equity [2]. Retaining these customers without sacrificing exclusivity requires precise, personalized marketing built

on fine-grained segmentation [3], a demand intensified by the multi-dimensional and often hidden nature of fashion purchasing behavior.

Effective personalization must move beyond demographic attributes such as age or gender, which offer only coarse context; behavioral segmentation grounded in transaction history gives a far more objective view of customer interests [4]. The RFM (Recency, Frequency, Monetary) model is among the most established behavioral frameworks, quantifying how recently, how often, and how much a customer purchases [5], yet many firms still derive it manually with spreadsheets, an approach prone to human error and unable to scale.

To cluster large-scale RFM parameters, K-Means is widely used for its efficiency [6], but as a distance-based method it is highly sensitive to outliers [7], which are common in the right-skewed spending of retail data. This study therefore applies a logarithmic transformation and a Robust Scaler to neutralize outlier influence before clustering. The number of clusters is determined not by the subjective Elbow Method alone [8] but by combining it with the Silhouette Score and the Davies-Bouldin Index, providing mathematically objective validation of compact, well-separated segments [9].

Although RFM-based K-Means reveals who the valuable customers are, it cannot reveal what products they consistently buy, and an extensive diversity of taste remains within each financially homogeneous cluster [10]. To uncover these latent preferences, Latent Dirichlet Allocation (LDA) is applied to each segment, treating every customer as a document and every purchased product as a word so that co-occurring purchase patterns emerge as latent topics [11]. Applying LDA independently within each cluster yields richer insight than applying it to the whole population at once [10]. Because raw model output remains an abstract exercise until it is made interpretable, the results are delivered through an interactive Business Intelligence dashboard that turns complex segmentation into intuitive, customer-centric visualizations [12].

Despite the maturity of these techniques, they are seldom integrated into one coherent pipeline for fashion retail: prior work computes RFM manually with little outlier handling, validates K-Means with a single metric, confines LDA to review text rather than transaction records, and stops at static results without an interactive interface. This study addresses those gaps with an integrated pipeline of RFM feature engineering, outlier-robust preprocessing, three-metric K-Means validation, per-segment LDA interpreted as fashion-style personas, and a web-based Streamlit dashboard, applied to real data from YBG (Yours Beyond Glamour), an affordable-luxury fashion retailer whose records from 1,152 customers had remained unexploited as dark data. The novelty lies in combining behavioral value segmentation with product-preference discovery inside a single validated, dashboard-ready framework, producing customer profiles that are both statistically robust and behaviorally rich for targeted marketing.

2. METHODS

The research was conducted quantitatively and organized according to the Cross Industry Standard Process for Data Mining (CRISP-DM), which structures an analytical project into six iterative phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. This framework was adopted because its phased and cyclical nature fits a study that moves from raw retail transactions toward a deployed decision support tool, allowing earlier phases to be revisited as insights emerge [13]. The entire pipeline, from data handling to modeling and visualization, was implemented in Python, with the

resulting analytical product presented through a web-based dashboard developed in Streamlit. Figure 1 illustrates the sequence of stages followed throughout the study.

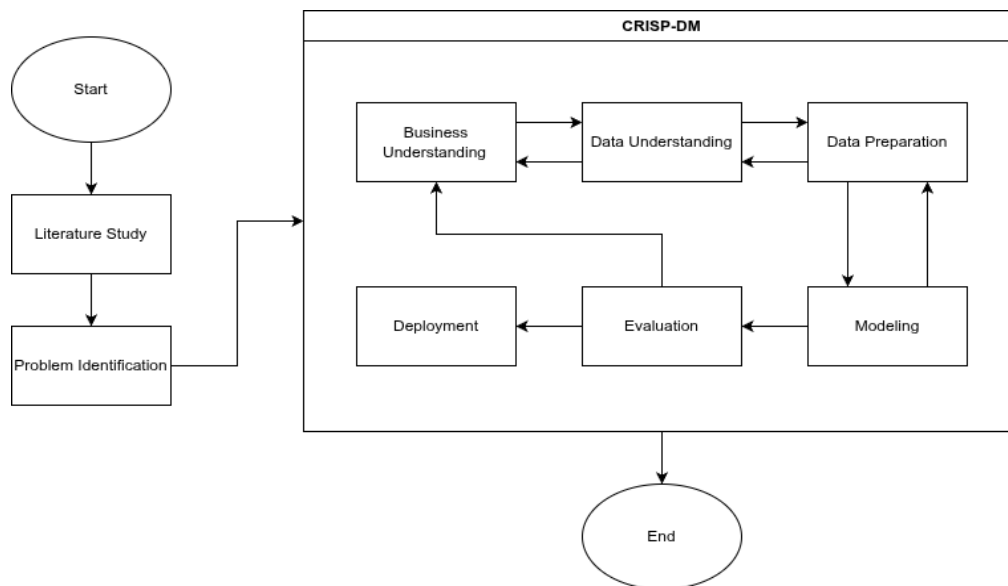


Figure 1. Research workflow based on the CRISP-DM framework.

The workflow begins with a literature study and problem identification, which establish the need to combine purchase value with product preference in fashion retail segmentation, followed by business understanding that frames the analytical goal in terms of actionable customer groups. In the data understanding stage, the transaction records of YBG, a fashion retail business covering the period from August 2023 to August 2025, are examined to assess their quality and completeness before any transformation. Data preparation then proceeds along two tracks. The first track builds the Recency, Frequency, and Monetary (RFM) features from the cleaned transactions and stabilizes their skewed distributions through logarithmic transformation, after which a Robust Scaler based on the median and interquartile range is applied so that extreme values do not dominate the distance computation [14], [15]. The second track prepares the transactional text for topic modeling by standardizing product and brand vocabulary and constructing a bag of words representation, in which each customer is treated as a document and each purchased item as a word, while terms that are too rare or too frequent across documents are filtered out [16], [17], [18].

The modeling stage applies two complementary techniques. K-Means clustering groups customers on the scaled RFM features, with the number of clusters guided by the Elbow Method, the Silhouette Score, and the Davies-Bouldin Index [8], [9], [19], and the resulting structure projected into two principal components for visual inspection. Latent Dirichlet Allocation is then applied independently within each customer segment to uncover latent purchase topics, which are interpreted as fashion personas according to the most prominent product and brand keywords in each topic [10], [11]. In the evaluation stage, the segmentation quality is assessed through the Silhouette Score and the Davies-Bouldin Index, while the topic models are assessed through the coherence score to determine the most interpretable number of topics per segment. Finally, in the deployment stage, the validated results are integrated into a single interactive dashboard whose functionality is verified through black box testing before release.

3. RESULTS AND DISCUSSION

The findings of this study are reported phase by phase in line with the CRISP-DM workflow, beginning with the framing of the business problem and the examination of the raw transaction data, continuing through feature construction, the formation and validation of customer segments, the extraction of purchase preferences within each segment, and ending with the deployed dashboard. Each phase is first described through its concrete output and then interpreted, so that the path from unstructured retail transactions toward actionable customer groups and their dominant style preferences can be followed clearly.

3.1 Business Understanding

YBG operates in the affordable luxury fashion retail segment, where customers vary widely both in how much they spend and in the kinds of products and brands they prefer. Under manual and intuition based handling, the business could not translate its accumulating transaction records into a clear picture of who its most valuable customers are or what each group actually tends to buy. This condition motivated the need for a segmentation approach that does not rely on a single dimension of value alone, but instead links the monetary contribution of each customer with the product and brand preferences expressed through their purchases, so that marketing effort can be directed more precisely rather than applied uniformly across the entire customer base.

Based on this need, the analytical objective of the study was defined as producing customer segments that are both measurable and interpretable. Measurable means that each segment is formed from quantitative purchasing behavior, namely how recently, how often, and how much a customer buys, while interpretable means that each segment can be further characterized by the dominant style of products it favors. This dual objective became the reference point for the subsequent stages, guiding the construction of behavioral features, the grouping of customers, the extraction of purchase preferences within each group, and finally the presentation of the results in a form that business users can explore directly.

3.2 Data Understanding

The study draws on two related sources exported from the operational system of YBG, namely a customer file containing identity attributes and a transaction file recording every purchase. Because the modeling in this study is driven by purchasing behavior and product preference, the transaction file serves as the primary source, while the customer file is used only to link purchases to unique individuals. The transaction records span the period from August 2023 to August 2025 and originally comprised 15,241 rows.

An initial inspection revealed that a large portion of these rows did not represent valid completed purchases, containing incomplete entries, non purchase activities, and placeholder or error values in several operational fields. After these invalid rows were removed, 3,775 valid transactions belonging to 1,152 unique customers remained, meaning that roughly three quarters of the raw rows were discarded before any feature was constructed. This substantial reduction reflects the presence of dark data, the unused and low quality records that accumulate in operational systems and can distort analysis when they are left unfiltered [1]. Each record also carries more than twenty operational fields, many of which relate to loyalty points, payment handling, and administrative notes that are not relevant to segmentation, so only the attributes that support the construction of behavioral features and the modeling of product preference were retained. Table 1 summarizes these selected attributes together with their role in the analysis.

Table 1. Description of the transaction dataset attributes

Attribute	Description	Role in Analysis
order_date	Date of each transaction	Recency and Frequency
customer_id	Unique customer identifier	Aggregation unit
sku_name	Purchased product name	Topic modeling
Brand	Product brand	Topic modeling
Category	Product category	Topic modeling
qty_ordered	Quantity purchased	Frequency support
after_discount	Final transaction value after discount	Monetary

As shown in Table 1, the retained attributes fall into two complementary groups that anticipate the modeling stage. The transaction date, quantity, and final value provide the behavioral basis for the RFM features that quantify how recently, how often, and how much each customer buys, while the product name, brand, and category provide the textual basis for uncovering the style preferences expressed in each customer's purchases. Understanding the data at this level of detail was essential before any transformation, since it determined which fields were meaningful to carry forward and confirmed that the cleaned dataset was sufficiently rich to support both the segmentation and the topic modeling performed in the following stages [1]. These attributes therefore become the raw material prepared in the next stage.

3.3 Data Preparation

From the cleaned transactions, three behavioral features were constructed for every customer. Recency was measured as the number of days between a customer's most recent purchase and the end of the observation period, Frequency as the total number of transactions made by the customer, and Monetary as the total value of those transactions after discount. Table 2 reports the descriptive statistics of these features across the 1,152 customers. The values reveal a strongly right skewed profile in the Frequency and Monetary dimensions, where the mean sits far above the median and a small number of customers reach extreme values, such as a maximum spending of more than 345 million rupiah against a median of about 6.5 million, both yielding a skewness of 5.37. Recency, in contrast, is far more evenly spread, with a skewness of only 0.58.

Table 2. Descriptive statistics of the RFM variables before transformation

Variable	Min	Max	Mean	Median	Std. Dev	Skewness
Recency	1	752	220,2	150.0	222.5	0.58
Frequency	1	47	2.47	1.0	3.88	5.37
Monetary	50,000	345,180,000	14,610,851	6,565,000	28,622,876	5.37

Because distance based clustering is sensitive to such imbalance, the features were transformed before modeling. A logarithmic transformation was first applied to compress the long tails of Frequency and Monetary [14], followed by a Robust Scaler that centers each feature on its median and scales it by the interquartile range, an approach more resilient to outliers than conventional min-max scaling [15]. Figure 2 contrasts the distributions before and

after this process. The aim of the transformation was not to force perfect normality but to curb the dominance of extreme values and place all three features on a comparable scale, and the figure shows that this was achieved most clearly for Monetary, whose sharply concentrated raw values become far more spread out. Some structure understandably remains, as Frequency still shows a strong mass at its lowest value, reflecting the large share of customers who purchased only once, while Recency separates into two groups of recent and long inactive customers. Rather than being a shortcoming, these residual patterns hint at natural groupings in the customer base that the clustering stage is expected to capture.

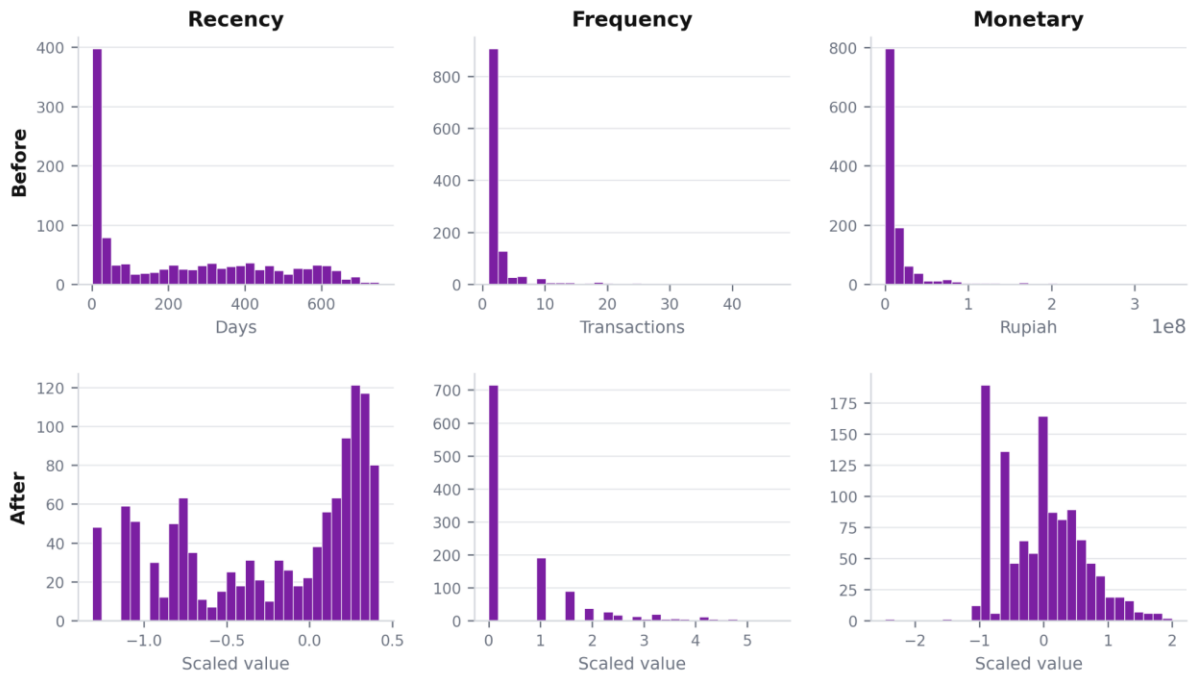


Figure 2. RFM variable distributions before and after transformation.

In parallel, a second track prepared the transactional text for topic modeling. The product, brand, and category fields of each customer were gathered so that every customer could be represented as a single document and every purchased item as a word within it. Product and brand names were standardized to merge inconsistent spellings into unified vocabulary items [18], after which the collection was converted into a bag of words representation, with terms that appeared in fewer than two customer documents or in more than eighty five percent of them filtered out to remove both noise and overly common items [16], [17]. This produced two aligned inputs from the same cleaned data, a scaled numerical matrix describing purchasing behavior and a textual matrix describing product preference, which together feed the modeling stage.

3.4 Modeling

The prepared RFM matrix was grouped using K-Means clustering, with the number of clusters set to four based on the evaluation described in Section 3.5. The algorithm was run with a fixed random seed and multiple initializations to ensure stable and reproducible groupings. Because the segmentation is based on three features, namely the scaled Recency, Frequency, and Monetary values, the resulting groups were projected from this three dimensional feature space onto two principal components to make them visually interpretable, with the two components together accounting for 97 percent of the total variance. Figure 3 shows this projection, in which the four segments occupy distinct regions along the first component, confirming that customers are separated primarily by the combined

strength of their purchasing behavior. Table 3 summarizes the average RFM profile of each segment.

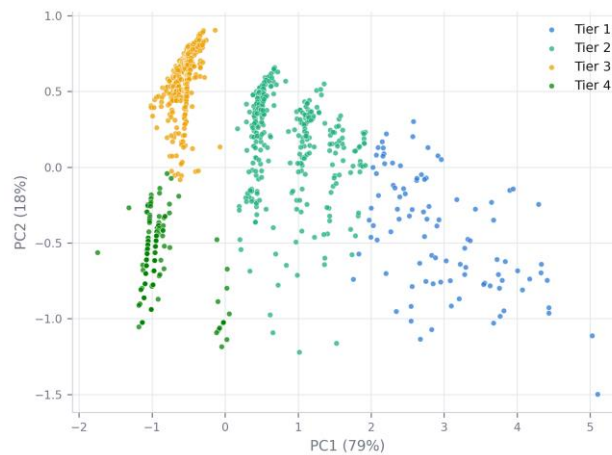


Figure 3. PCA visualization of RFM customer clusters.

Beyond this visual separation, the segments also differ substantially in their underlying behavior. Table 3 reports the average Recency, Frequency, and Monetary profile of each segment together with its size.

Table 3. Average RFM profile and size of each customer segment.

Segment	Customer	Recency (Days)	Frequency	Monetary (Rupiah)	Characterization
Tier 1	97	140	12	83,978,771	High value loyal customers
Tier 2	325	277	3	17,473,965	Mid value, cooling customers
Tier 3	378	390	1	6,402,164	Low value, long dormant buyers
Tier 4	352	7	1	1,666,770	Recently acquired one-time buyers

The four segments describe clearly different relationships with the business. Tier 1 is the smallest but most valuable group, purchasing about twelve times and spending close to 84 million rupiah on average while remaining relatively active. Tier 2 represents a mid value group whose members buy occasionally but have not returned for several months, signaling a risk of gradual disengagement. Tier 3 is the largest group and consists mainly of one time buyers who have been inactive for more than a year, marking them as dormant. Tier 4, in contrast, is defined by very recent activity of only about a week, yet with a single low value purchase, which identifies it as newly acquired customers whose long term potential is not yet established. This profile confirms that value and engagement, rather than demographic attributes, form the primary axis along which the customer base divides.

While the RFM segments explain how much and how recently customers buy, they do not reveal what those customers actually prefer. To address this, Latent Dirichlet Allocation was applied independently within each segment, treating every customer as a document and every purchased product or brand as a word, so that the latent purchase topics uncovered reflect the taste of that specific value group [10], [11]. Each resulting topic was

then interpreted as a fashion persona according to its most prominent product and brand keywords. Table 4 summarizes these personas across the four segments.

Table 4. Extracted fashion personas for each customer segment using LDA.

Segment	Topics	Dominant fashion personas	Signature Brands
Tier 1	3	Prestige heritage leather goods; exclusive heritage accessories; collector vintage timepieces	Chanel, Hermès, Dior, Loro Piana, watches
Tier 2	8	Contemporary quiet luxury, heritage flap and saddle bags, iconic monogram wallets, stealth wealth menswear	Céline, Loewe, Saint Laurent, Balenciaga, Chanel, Louis Vuitton
Tier 3	2	Accessible heritage leather goods; accessible classic flap bags	Louis Vuitton, Dior, Chanel, Gucci, Celine
Tier 4	6	Practical daily totes, casual leather hobos, everyday wallet and coin essentials	Longchamp (Le Pliage), functional totes, wallets

Reading the two tables together reveals a pattern that neither method could show alone. Each value tier carries a distinct brand DNA. The most valuable customers in Tier 1 gravitate toward heritage houses such as Chanel, Hermès, and Dior and even collectible timepieces, whereas the newly acquired customers in Tier 4 concentrate on practical and accessible pieces led by Longchamp. The mid value Tier 2 is the most heterogeneous, spreading across eight personas that mix quiet luxury with iconic monogram styles, while the dormant Tier 3 collapses into only two broad personas centered on accessible classics. This shows that customers of similar monetary standing do not necessarily share the same taste, and that a high spending segment is not stylistically uniform. Combining purchase value with product preference therefore produces a segmentation that is both financially and stylistically meaningful, which is the central contribution of this study and the basis for the persona driven views presented in the dashboard.

3.5 Evaluation

The number of customer segments was determined by testing K-Means with between two and eight clusters and examining three complementary indicators. The Elbow Method observes the within cluster sum of squares (WCSS), looking for the point where its decline begins to level off, while the Silhouette Score rewards well defined clusters with higher values and the Davies-Bouldin Index rewards compact and well separated clusters with lower values. Table 5 reports all three across the tested range. As shown, the WCSS falls sharply up to four clusters and then flattens, producing a visible elbow at that point. Because the location of an elbow is read visually and can be subjective [8], this indication was confirmed with the other two measures, both of which reached their best value at exactly four clusters. Four segments were therefore adopted for the analysis.

Table 5. Performance evaluation metrics of K-Means clustering across various values of K.

Number of Clusters	WCSS	Silhouette Score	Davies-Bouldin Index
2	948.3	0.5489	0.7042
3	552.1	0.5298	0.6623
4	290.7	0.5952	0.5644
5	217.8	0.5861	0.6460
6	187.0	0.5706	0.7458
7	162.5	0.5801	0.7120
8	144.0	0.5037	0.7847
9	130.5	0.4232	0.8540

Reading the three indicators together explains why four clusters were chosen. The WCSS drops steeply from 948.3 at two clusters to 290.7 at four, after which each additional cluster brings only a small further reduction, so the marginal benefit of splitting the data more finely fades beyond this point. At the same configuration, the Silhouette Score of 0.5952 indicates that the four segments possess a reasonably strong structure in which customers are closer to their own group than to neighboring ones, consistent with the interpretation applied in earlier RFM segmentation studies [9]. The Davies-Bouldin Index of 0.5644, being well below one, further confirms that the segments are compact internally and clearly separated from one another [19]. The agreement of all three indicators on the same number of clusters shows that the segmentation is statistically sound rather than an arbitrary division of the customer base.

Beyond the customer segments, the topic models built within each segment were also evaluated. The number of topics in each segment was not fixed in advance but selected empirically using the coherence score, which measures how semantically consistent the words within a topic are. Each segment was modeled repeatedly with between two and eight topics, and the value giving the highest coherence was retained. Figure 4 presents these coherence curves for all four segments, with the selected number of topics marked on each. The resulting optimal structures were three topics for Tier 1, eight for Tier 2, two for Tier 3, and six for Tier 4.

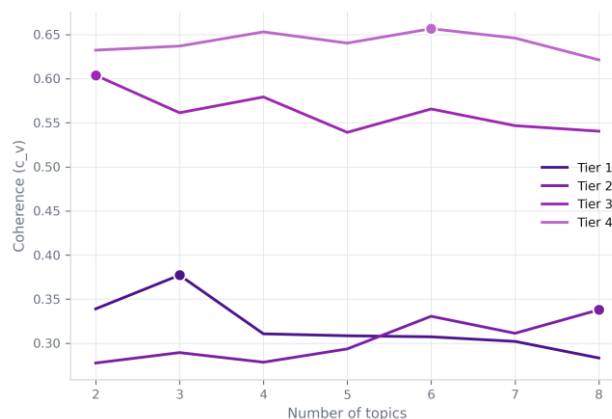


Figure 4. Coherence score line chart for topic number selection in each segment.

The coherence results reveal a meaningful contrast between the segments. Tier 3 and Tier 4 reached noticeably higher coherence, of about 0.60 and 0.66, because their purchases revolve around a narrower and more consistent vocabulary of accessible and practical items, which produces cleaner topics. Tier 1 and Tier 2 settled at lower coherence, of about 0.38 and 0.34, since their far richer mix of luxury brands and product types naturally yields topics that overlap more, and Tier 2 in particular required as many as eight topics to accommodate its diverse taste. Importantly, coherence was treated as a guide for selecting the number of topics rather than as the sole measure of validity. Following the practice of interpreting topics through domain knowledge [17], every topic was additionally examined qualitatively, and even in the lower scoring segments the highest probability product and brand keywords could still be read and labeled as a sensible fashion persona. A lower coherence value therefore reflects broader stylistic variety within a high value segment rather than a failure of the model, and the personas remained meaningful and usable throughout. With both the segmentation and the topic models validated numerically and qualitatively, the results were consolidated into the dashboard described in the following stage.

3.6 Deployment

The validated models were consolidated into a single web based analytical dashboard built with Streamlit, so that the results could be explored by business users without any programming. The dashboard is organized into four pages. The Overview page presents a high level summary of the customer base and overall sales activity. The Segmentation page displays the four RFM segments together with their profiles and the principal component view. The Purchase Pattern page translates the per-segment LDA output into readable fashion personas, showing the distribution of customers across personas and the signature keywords of each. The Customer 360 page brings these two perspectives together for an individual customer, combining the value tier with the dominant style persona. Figure 5 shows the Overview page of the dashboard.

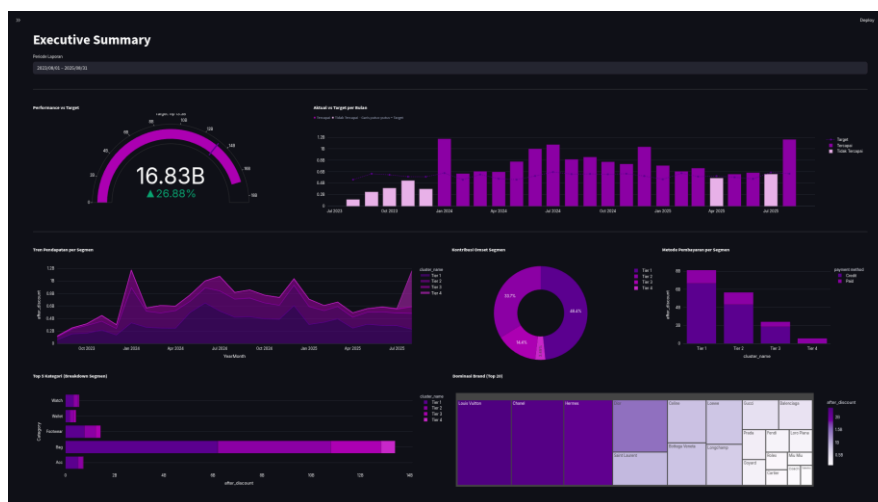


Figure 5. Overview page of the analytical dashboard.

To ensure that the dashboard behaves as intended, its functionality was verified through black box testing, which checks each feature against its expected output without inspecting the underlying code. The testing covered the four pages across 45 scenarios, including navigation, filtering, segment selection, persona display, and individual customer lookup, and all of them produced the expected results, giving a 100 percent success rate. This confirms that the analytical pipeline is not only statistically valid but also operationally reliable when delivered to end users.

Taken as a whole, the value of this study lies less in any single technique than in the way the three reinforce one another. RFM reduces each customer to three behavioral numbers and K-Means arranges those numbers into value tiers, which is where many previous efforts stop, including studies that pair RFM and K-Means with an interactive interface [20], [21]. Such tiers answer who the important customers are but remain silent on what they actually want, and the per-segment LDA layer is what closes this gap. Its payoff is visible in the contrast between the tiers, since the most valuable customers in Tier 1 are not a uniform audience but split into heritage leather goods buyers, accessory buyers, and even collectors of vintage timepieces, whereas the newly acquired customers in Tier 4 converge on practical and accessible pieces led by Longchamp. A value only segmentation would have treated the whole of Tier 1 identically and missed these distinctions entirely.

By modeling preference within each tier rather than across the whole base, the approach keeps every persona anchored to a comparable level of spending power, which makes it directly actionable, since an offer aimed at Tier 1 heritage buyers carries a very different price expectation from one aimed at Tier 4 everyday buyers. This speaks directly to the personalization problem that motivated the study, as the business can now align its offers with both the spending capacity and the taste of each group, protecting the exclusivity that premium fashion depends on while still engaging lower tiers with relevant everyday products. The dashboard makes this operational by letting a non technical user move from an aggregate view down to a single customer and see both the value tier and the style persona at once, advancing beyond earlier RFM and K-Means dashboards that report value tiers alone [20].

CONCLUSION

This study set out to move customer segmentation in fashion retail beyond a single dimension of transaction value, and its findings meet that aim. Using real transaction data from YBG, RFM features combined with K-Means clustering produced four distinct customer segments whose validity was confirmed by three agreeing indicators, reaching a Silhouette Score of 0.5952 and a Davies-Bouldin Index of 0.5644, while Latent Dirichlet Allocation applied independently within each segment uncovered the shopping style preferences hidden inside every value tier and translated them into interpretable fashion personas that differ markedly from one tier to another. Both outcomes were consolidated into a four page interactive dashboard that passed forty five black box testing scenarios with a perfect success rate. The main theoretical contribution lies in coupling value based and preference based segmentation within a single pipeline, extending the analysis from who the valuable customers are to what they actually prefer, and doing so by applying topic modeling to structured transaction data rather than the review text on which such models are usually trained. Together with a cluster validation that relies on three complementary metrics instead of a single subjective criterion, this produces a segmentation that is at once statistically sound and behaviorally rich, offering a replicable way to turn ordinary sales records into segments that are financial and stylistic at the same time.

Several directions remain open for further development. Because the analysis draws on data from a single business, applying the same pipeline to other retailers would help test how far the approach generalizes across different product ranges and customer bases. The segmentation is also static in time, so incorporating temporal and predictive elements, such as customer lifetime value estimation or churn prediction, would add a forward looking dimension to the current profile. Finally, the per-segment personas provide a natural foundation for a recommendation layer, in which products that consistently appear together within a tier could

drive targeted cross-selling, product bundling, and personalized campaigns that bring the segmentation still closer to day to day marketing practice in fashion retail.

REFERENCES

- [1] O. F. Olaitan *et al.*, "Dark data in business intelligence: A systematic review of challenges, opportunities, and value creation potential," *J. Econ. Bus. Commerce*, vol. 2, no. 2, pp. 135–142, 2025.
- [2] K. Jing, L. Chen, and Y. Mei, "The effectiveness of price promotions in purchasing affordable luxury products: An event-related potential study," *Front. Neurosci.*, vol. 16, art. 818503, 2022.
- [3] E. Yıldız, C. Güngör Şen, and E. E. Işık, "A hyper-personalized product recommendation system focused on customer segmentation: An application in the fashion retail industry," *J. Theor. Appl. Electron. Commer. Res.*, vol. 18, no. 1, pp. 571–596, 2023.
- [4] A. U. Makhtibekov, "Segmentation of the audience based on behavioral and demographic data: Prospects for personalized marketing," 2025, doi: 10.24412/2411-0450-2025-7-125-130.
- [5] T. Ho, S. Nguyen, H. Nguyen, N. Nguyen, D.-S. Man, and T.-G. Le, "An extended RFM model for customer behaviour and demographic analysis in retail industry," *Bus. Syst. Res. J.*, vol. 14, no. 1, pp. 26–53, 2023.
- [6] O. P. Barus, C. Nathasya, and J. J. Pangaribuan, "The implementation of RFM analysis to customer profiling using K-Means clustering," *Math. Model. Eng. Probl.*, vol. 10, no. 1, pp. 298–303, 2023.
- [7] J. Heidari, N. Daneshpour, and A. Zangeneh, "A novel K-means and K-medoids algorithms for clustering non-spherical-shape clusters non-sensitive to outlier," *Pattern Recognit.*, vol. 155, art. 110639, 2024.
- [8] E. Schubert, "Stop using the elbow criterion for k-means and how to choose the number of clusters instead," *ACM SIGKDD Explor. Newsl.*, vol. 25, no. 1, pp. 36–42, 2023.
- [9] M. K. Delimayanti, A. Hidayati, D. Kurniawati, N. Z. Rahmadani, and M. B. Faydhurrahman, "Customer segmentation in steel sales: Optimizing profitability with RFM model and K-Means algorithm," in *Proc. Int. Conf. Artif. Intell., Blockchain, Cloud Comput., Data Anal. (ICoABCD)*, 2024, pp. 267–272.
- [10] M. Nilashi, S. Samad, B. Minaei-Bidgoli, F. Ghabban, and E. Supriyanto, "Online reviews analysis for customer segmentation through dimensionality reduction and deep learning techniques," *Arab. J. Sci. Eng.*, vol. 46, no. 9, pp. 8697–8709, 2021.
- [11] R. Tachibana and H. Yamashita, "An analysis model for the relationship between consumers' purchase and sense of value data," in *Proc. IEEE Int. Conf. Big Data, Cloud Comput., Data Sci. Eng. (BCD)*, 2019, pp. 136–141.
- [12] M. Siek and V. V. Lim, "Implementation of customer relationship management analytical dashboard with customer segmentation," in *Proc. 15th Int. Conf. Inf. Commun. Technol. Syst. (ICTS)*, 2025, pp. 1–6.
- [13] C. Schröer, F. Kruse, and J. M. Gómez, "A systematic literature review on applying CRISP-DM process model," *Procedia Comput. Sci.*, vol. 181, pp. 526–534, 2021.
- [14] T. P. Reddy and G. K. N., "Predictive vs descriptive segmentation: Markov and K-Means clustering for RFM analysis," in *Proc. Int. Conf. Electron., Comput., Commun. Control Technol. (ICECCC)*, 2025, pp. 1–9.
- [15] M. D. V. Prasad and S. T., "Enhancing K-Means clustering accuracy through modified robust scaling technique," 2024, doi: 10.20944/preprints202411.1245.v1.
- [16] S. Shashank and R. K. Behera, "Factors influencing recommendations for women's

- clothing satisfaction: A latent Dirichlet allocation approach using online reviews," *J. Retail. Consum. Serv.*, vol. 81, art. 104011, 2024.
- [17] Z. Rosadi and A. Solichin, "Topic modeling tugas akhir mahasiswa menggunakan metode latent Dirichlet allocation dengan Gibbs sampling," *J. Ticom: Technol. Inf. Commun.*, vol. 13, no. 1, pp. 38–44, 2024.
- [18] Y. Zou, C. Zhao, P. Childs, D. Luh, and X. Tang, "User experience design for online sports shoe retail platforms: An empirical analysis based on consumer needs," *Behav. Sci.*, vol. 15, no. 3, art. 311, 2025.
- [19] K. Mehrez, "RFM–K-OPT based machine learning framework for customer segmentation and behavioral profiling in direct marketing," *Int. J. Adv. Comput. Sci. Appl.*, vol. 16, no. 12, 2025.
- [20] F. Alzami et al., "Implementation of RFM method and K-Means algorithm for customer segmentation in e-commerce with Streamlit," *ILKOM J. Ilm.*, vol. 15, no. 1, pp. 32–44, 2023.
- [21] A. Nugraha, Y. A. Effendi, N. Nicholas, Z. Tao, M. Afifuddin, and N. Nuzulita, "K-Means clustering interpretation using recency, frequency, and monetary factor for retail customers segmentation," *TELKOMNIKA*, vol. 23, no. 2, p. 435, 2025.