Summer Fashion Trend Identification: Harnessing Clustering and Sentiment Analysis for Sustainable Trend Production

1. Executive Summary

In today's digital age, e-commerce and fashion retail are experiencing a significant transformation due to rapidly evolving consumer behaviors and expectations. Modern buyers are increasingly conscious of their purchasing decisions, considering not only the style, color, and price of a garment but also its quality, sustainability, and longevity. Therefore, the need for buyer trend identification is crucial to ensure that companies are optimizing revenue growth while also considering sustainability efforts through producing their most popular fashion trends.

In this project, we will be leveraging clustering, sentiment analysis, text mining, and visualizations in R to identify and capitalize on popular summer fashion trends for retailers to focus on. This proactive approach will allow us to optimize production schedules, reduce excess inventory, and lower production costs. This will enable retailers to enhance customer satisfaction by sustainably producing a curated selection of trendy clothing items, helping them stay at the forefront of the industry.

2. Statement of the Problem

In the rapidly evolving landscape of e-commerce and fashion retail, companies face significant challenges in aligning production with fluctuating consumer preferences and market demands. Modern consumers are not only seeking fashionable items but are also increasingly conscious of factors like sustainability and quality. This shift necessitates a robust method of identifying buyer fashion trends accurately and efficiently.

There is a critical need for a data-driven strategy that utilizes advanced analytical techniques such as clustering, text mining and sentiment analysis to identify the relationship between characteristics of clothing that make them trendy. This approach would enable fashion retailers to optimize production processes, minimize waste, and enhance customer satisfaction by offering products that align with the latest trends.

3. Literature Review

Techniques like clustering, sentiment analysis, text mining, and visualizations help retailers stay competitive, optimize production, and support sustainability. These methods enable the industry to meet modern consumer demands while reducing waste. Incorporating these analytical techniques can lead to a 15-20% improvement in operational efficiency and a potential 10-12% increase in market share,McKinsey & Company (2021)[1].

Clustering is widely used in fashion retail to categorize products based on the various attributes such as color, style, material, and price. Clustering, as noted by Xu and Wunsch (2005)[2], is essential for segmenting data into groups based on similarities. Algorithms like k-means and hierarchical clustering help categorize consumer preferences, aiding retailers in identifying and prioritizing popular styles, resulting in a more personalized shopping experience for customers and a better understanding of purchasing behavior. Sentiment analysis, discussed by Liu (2012)[3], determines the emotional tone in textual data, useful for

[1] *McKinsey & Company (2021). "What's now and next in analytics, AI, and automation.*

[2] *Survey of Clustering Algorithms, axon.cs.byu.edu/Dan/678/papers/Cluster/Xu.pdf. Accessed 4 Aug. 2024.*

[3] *Sentiment Analysis and Opinion Mining, www.cs.uic.edu/~liub/FBS/SentimentAnalysis-and-OpinionMining.pdf. Accessed 4 Aug. 2024.*

interpreting consumer feedback from social media and reviews. According to a survey by Deloitte, 45% of consumers are influenced by online reviews. This method helps retailers adjust products based on public perception and preferences, potentially increasing customer satisfaction by 10-15%. Text mining extracts valuable information from tags, reviews and social media. Feldman and Sanger (2007)[4] highlight Natural Language Processing (NLP) and topic modeling techniques that reveal emerging trends. A case study showed that text mining improved trend identification by 20%, allowing retailers to adapt their inventory to current consumer interests, leading to a 12% increase in sales.

Sustainability is increasingly vital in fashion. Consumer preferences now prioritize sustainability, quality, and ethical production. According to McKinsey & Company (2019)[5], 66% of global consumers and 73% of Millennials are willing to pay more for sustainable products. A report by Nielsen(2019)[6],indicated that sustainable practices could boost customer loyalty by 20%, with 50% of surveyed consumers prioritizing eco-friendly brands.

4. Research Questions, Hypotheses, and Effects

Research Question #1: Relative to summer fashion products in lower-priced clusters, do products in median and high priced clusters experience lower sales volumes?

- **Null Hypothesis (H0):** Products in median and high priced clusters experience the same or higher sales volumes as products in lower-priced clusters
- **Alternative Hypothesis (H1):** Products in median and high priced clusters experience lower sales volumes than products in lower-priced clusters.
- **Metrics**: A decrease in sales by 10% or more for median and high priced products compared to lower-priced products.

Research Question #2: Relative to summer fashion products with neutral or negative review sentiments, do products with predominantly positive review sentiments have higher sales volumes?

- **Null Hypothesis (H0):** Products with predominantly positive review sentiments do not have higher sales volumes compared to products with neutral or negative review sentiments.
- **Alternative Hypothesis (H1):** Products with predominantly positive review sentiments have higher sales volumes compared to products with neutral or negative review sentiments.
- **Metrics**: A sales increase of 12% or more for products with predominantly positive reviews.

5. Data Attributes

Two datasets are being utilized for this project; Dataset A consists of fashion product data while Dataset B consists of ratings and reviews of aforementioned products. Dataset A comprises 43 columns and encompasses a wide range of fashion products including clothing, accessories, footwear, and beauty products. It covers various categories such as casual wear, formal wear, athleisure, swimwear, lingerie, and more. Each category includes detailed information about product characteristics such as color, pattern,

---

[4] *(PDF) The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data, www.researchgate.net/publication/200504395_The_text_mining_handbook_Advanced_approaches_in_analyzing_unstructured_data. Accessed 4 Aug. 2024.*

[5] *Amed, I., Berg, A., Brantberg, L., & Hedrich, S. (2019). The State of Fashion 2019. McKinsey & Company. Retrieved from The State of Fashion 2019*

[6] *Nielsen. (2019). The Evolution of the Sustainability Mindset. Nielsen. Retrieved from The Evolution of the Sustainability Mindset*

material, price, product display page (PDP) links as well as numerical ratings (scale 1-5 stars). Dataset B consists of 11 columns, revolving around the reviews written by customers for fashion products in Dataset A. It also contains numeric rating information as well as product department, division, and class information.

6. Data Cleaning and Processing

Both datasets have been cleaned to account for missing values and inconsistencies within variables. Within Dataset A, less than 5% of the data consisted of missing or NULL values so these were omitted. Within the numerical rating columns, missing values were imputed with the mean. Duplicate columns and columns which will not be considered for further analysis are dropped bringing down the total variables in the final dataset to 28 (originally 43).

The product_color variable was cleaned by removing spaces, special characters and converting to lowercase. An analysis was done to see all the unique values for color in our dataset. These colors were then mapped to their "umbrella" color e.g. royal blue is mapped to blue, fuschia is mapped to pink. Product_variation_size_id was also manipulated where different variations of a size were mapped to the correct over arching size, for example SIZE 4XL is mapped to XXXXL to create consistency within the variable. Categorical variables such as badge_fast_shipping and badge_product_quality were converted to factor variables.

In Dataset B, numeric and string variables were handled differently in terms of missing/NULL values. Any row that had missing data in both the Review.Text and Title column was omitted. Any row that has data present in the Review.Text column but data missing in the Title column was replaced with an N/A value. Similarly the few missing values under Division.Name, Department.Name and Class.Name were given a default value of "Unknown". The Review.Text column is cleaned by first creating a corpus and then converting all data to lowercase, removing punctuations, stopwords and whitespace for consistency. A dictionary is created from all words present in Review.Text and similar words are grouped together. Terms appearing in less than 5% of the reviews are filtered out to aid in our analysis.

Dataset A and B are further transformed by creating four new columns; product_type (bell-bottom, v neck, cardigan etc), style (goth, grunge, boho, beach), material and pattern to aid in trend analysis. Four separate vectors of keywords were created, each consisting of words used to describe the aforementioned 4 columns. The grep function is then used to parse these keywords from the title, tags, review, division name, department name columns. If the keyword exists in these columns then it is added to the newly created product_type/style/pattern/material column; if not "unknown" is added as a placeholder.

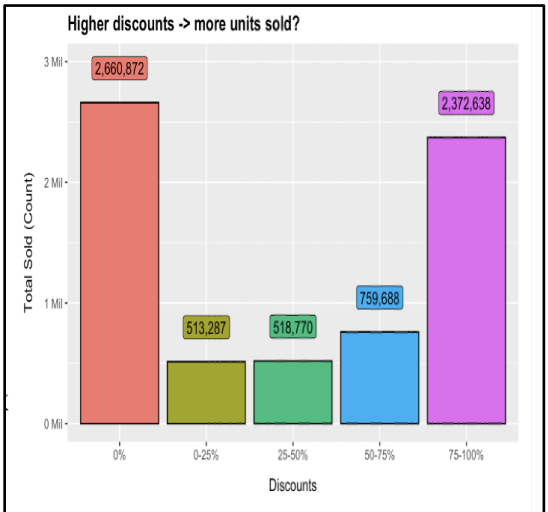

Figure #1: Discount versus Sales Volume

7. Exploratory Data Analysis

- **Clustering**

An in depth analysis is performed to understand the relationship between various attributes of an item and its sales volume and the impact of these attributes. It includes a univariate examination of both numeric and categorical variables in the dataset. Key questions addressed and analyzed in the EDA are:

1. Does a higher discount value (calculated as the difference between price column (sale price) and retail price column (original price)) lead to a higher number of units sold?

From Figure #1, it's observed that the higher the discounts,the more units are sold. Items that sold at a 0% discount were predominantly black in color and priced under €10. There are around 2M items sold at 0% discount rate. These could be items which are lower in price ,have good ratings or are sought after due to quality or style.

2. What are the top selling categories in each of the attributes - Color, Size, Patterns, Materials, Style, Product type?

Black, white, blue, green, gray and red have been identified as the most preferred colors. With respect to clothing sizes **XS, S, M and L** are the most popular sizes. With respect to pattern; **floral, printed, stripe, polka dot and plaid** emerge to be winners for the Summer. The preferred material is **cotton** followed by **chiffon, elastic, stretch, polyester and denim** which are primary materials used for outdoor activities such as running, sports or yoga.

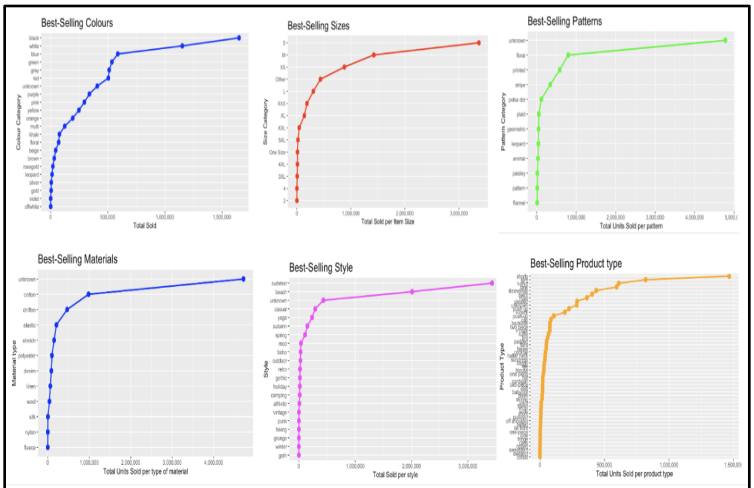

Figure #2: Analysis of Product Characteristics versus Sales Volume

As expected, **shorts, slim bottoms, t-shirts, tanks, sleeveless, swim and maxi** are the top products sold in summer. The styles associated with these products are **beach, casual,** and **yoga**. Autumn and spring styles were also identified as best sellers, which could be indicative of styles bought for traveling during the summer or pre-buying for the upcoming seasons.

3. What is the overall perception of products and sellers on the website? Do higher ratings result in a higher number of units sold?

We observed that most of the product and merchant ratings on the site for items priced < € 20 are between 3.5 to 5, which is a good indicator of perception of the products and their sellers on the site. Mid-tier and high-tier priced items also have ratings between 3.5 to 5, demonstrating the customer perception of high prices representing better quality

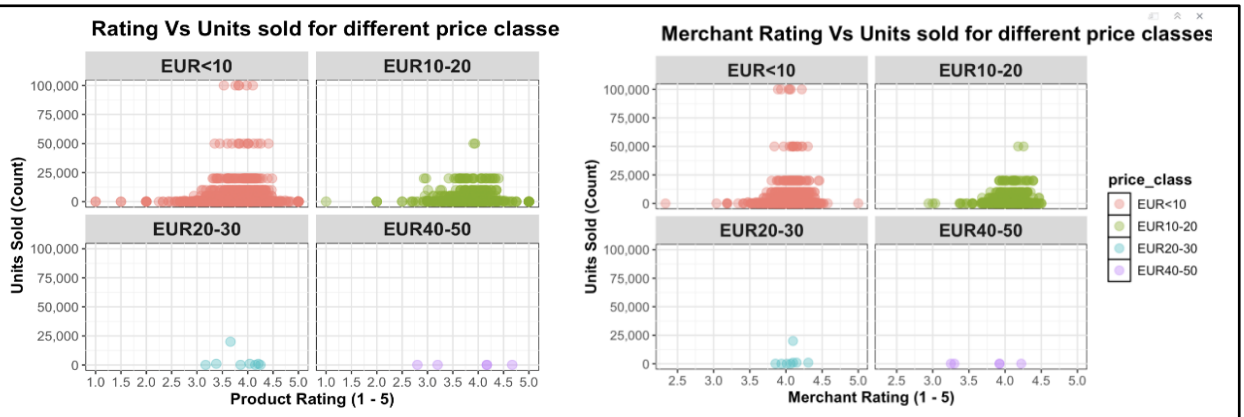

Figure #3: Product/Merchant Ratings versus Sales Volume

- **Text Mining of Tags**

The analysis reveals that for low-priced items, having fewer, well-chosen tags can significantly enhance sales. Across different price ranges, the total number of tags does not substantially impact sales, emphasizing the importance of tag relevance over quantity. Effective tags like "fashion," "women," and "casual" are strongly associated with higher sales, highlighting common consumer search terms. Retailers can optimize their tagging strategy by focusing on fewer, high-performing tags that align with consumer preferences to improve product visibility and sales performance.

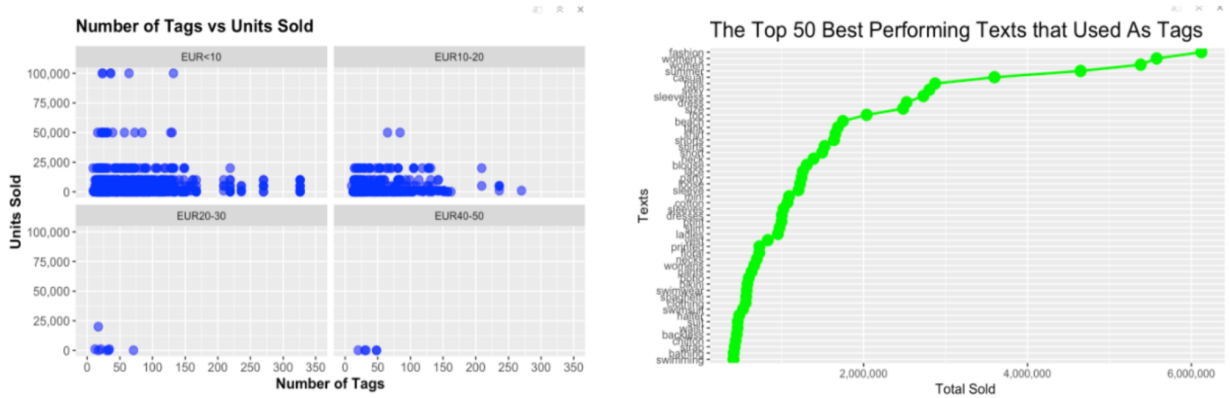

Figure #4: Relationship between Product Tags and Units Sold

- **Sentiment Analysis**

An initial analysis of the relationship between text reviews and their corresponding numeric ratings is performed. As seen in Figure #5, a majority of the text ratings left are for items that received a high numeric rating (4-5 stars), with the average rating score being a 4.183. This suggests that customers generally have positive experiences with the products, as reflected by both the high volume of positive text reviews and the overall high average rating.

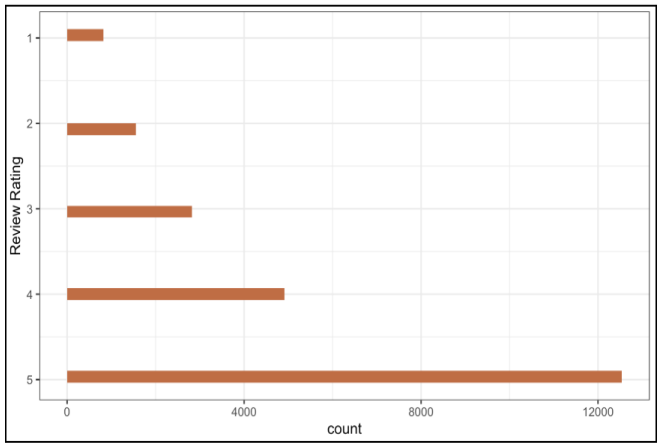

Figure #5: Distribution of Text Ratings Per Numeric Rating

A tibble: 1 × 3

characters <dbl>	words <dbl>	sentences <dbl>
308.7554	60.19668	4.718873

1 row

Figure #6 : Summary of Average Characters, Words, and Sentences Count

This relationship underscores the importance of identifying product characteristics that influence positive reviews, which in turn can enhance the product's reputation and attract potential buyers.

8. Clustering Analysis

- **Hierarchical Clustering**

Clustering analysis serves as a robust tool that segments the products into groups based on similarities in various attributes, such as customer ratings, style, material, and color. In the context of our data, clustering was chosen for its ability to handle complex, multi-dimensional data inherent to fashion

products. This approach facilitates developing targeted marketing strategies and production planning by identifying patterns that might not be immediately apparent through traditional analytics.

Following the in-depth data exploration, attributes were selectively chosen from Dataset A based on their significance to consumer preferences and purchasing behaviors. Numeric features, such as retail price, were utilized directly. Categorical data, including attributes like style, product_type, material, pattern, color, and tags, were processed using suitable encoding techniques to transform them into a numerical format conducive to clustering. The selected features, excluding rating, are scaled to ensure each variable contributes equally to the distance calculations. This ensures a holistic view of what characteristics impact a product's trendiness and popularity.

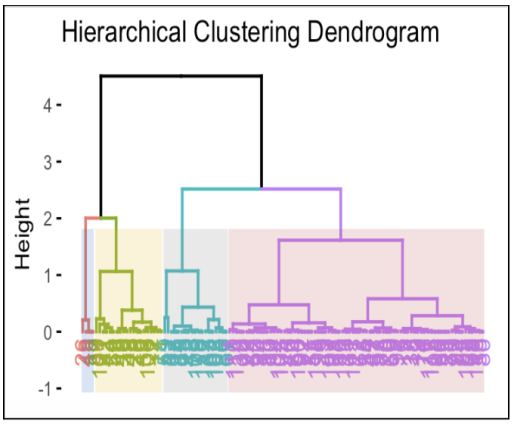

Figure #7 : Dendrogram

The Euclidean distance matrix is employed to calculate the distance between observations. Subsequently, hierarchical clustering is conducted using the Ward.D2 method, which aims to minimize the variance within each cluster. Upon analyzing the resulting dendrogram with k = 4 clusters (Figure #7), the identified clusters are reintegrated into the original dataset. Summary statistics are computed for each cluster, including means for numeric columns and modes for categorical columns. This approach illustrates the distribution of product features and facilitates deeper insights into each cluster's characteristics.

	cluster	mean_retail_price	mode_product_type	mode_style	mode_pattern	mode_material	mode_product_color
	<fct>	<dbl>	<chr>	<chr>	<chr>	<chr>	<chr>
1	1	19.0	shorts	summer	printed	cotton	black
2	2	37.2	shorts	summer	printed	cotton	pink
3	3	7.5	tee	summer	printed	cotton	pink
4	4	29.5	t-shirt	summer	printed	cotton	blue

Figure #8 : Hierarchical Cluster Summary Statistics

From Figure #8, it can be deduced that Cluster 1 represents a customer segment seeking affordable, casual summer shorts. The combination of a printed pattern and black color suggests that these items are versatile for casual wear. The low price point indicates that these products are likely appealing to budget-conscious consumers in need of practical, everyday clothing. In contrast, Cluster 2 also primarily features shorts but stands out due to a higher price point and the color pink. This cluster likely appeals to a customer segment that prioritizes style and femininity in their casual attire, willing to invest more in these qualities. The higher price also suggests a brand strategy appealing to a middle-income demographic.

Cluster 3 centers on highly affordable, casual tees marked by a summer style and the color pink, similar to Cluster 2 but at a substantially lower price point. This cluster likely attracts cost-sensitive consumers who desire stylish, seasonal garments without a high cost. The extremely low price implies that high volume sales may be a strategic focus for these products, as opposed to quality. Meanwhile, Cluster 4 features casual t-shirts at a higher price, indicating a premium customer

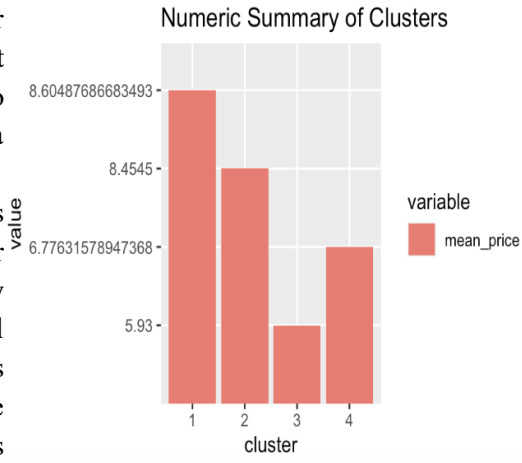

Figure #9: Retail Price Distribution Across Clusters

segment. The combination of a summer style, blue color, and printed design suggests that these t-shirts are marketed towards consumers willing to pay a premium for enhanced quality or brand value. This segment may also emphasize durability and superior cotton materials, justifying the higher pricing.

A manual segmentation analysis is conducted, where products are grouped based only on numerical rating (scale of 1-5). Insights from this rule-based clustering helps analyze characteristics of products directly defined by their ratings.

	cluster	mean_retail_price	mode_product_type	mode_style	mode_pattern	mode_material
	<fct>	<dbl>	<chr>	<chr>	<chr>	<chr>
1	4.22	6.5	"t-shirt"	beach	printed	cotton
2	4.23	65.5	"t-shirt"	casual	printed	cotton
3	4.24	34.7	"shorts"	summer	printed	cotton
4	4.28	32.7	"shorts"	athletic	printed	elastic
5	4.33	21.5	"slim"	summer	stripe	cotton
6	4.34	48	"tank"	summer	floral	chiffon
7	4.39	13	"tank"	summer	printed	cotton
8	4.46	18	"slim"	summer	floral	chiffon
9	4.47	9	" pleated"	yoga	flannel	elastic
10	4.75	10	"slim"	beach	printed	cotton
11	5	55	"v-neck"	beach	printed	silk

Figure #10 : Top 10 Highest Rating Clusters

From Figure #10, we gain a comprehensive understanding of the prevalent styles, materials, and product types among higher-rated versus lower-rated products. It is evident that items such as v-neck tops, slim jeans, pleated bottoms, and tank tops crafted from premium fabrics like silk, chiffon, and cotton are highly favored by the customer base. This insight can strategically inform marketing campaigns and product development efforts by focusing on features that customers most appreciate and value. Segmenting data by ratings facilitates the identification of areas for enhancement in lower-rated products while reinforcing the successful attributes of higher-rated ones.

Additionally, insights derived from the mean prices across each cluster can aid in refining pricing strategies to better align with consumer expectations and satisfaction levels. This type of segmentation analysis, combined with hierarchical clustering, provides valuable insights into the effectiveness of the current product suite and can guide immediate strategic business decisions.

- **K-Means**

In this analysis, product clustering, based on numerical features, is employed to guide pricing strategies and refine production and inventory management. To ensure each feature contributes equally to the clustering algorithm, variables such as price, retail price, units sold, rating, and rating count were scaled. The Elbow method to identify the optimal number of clusters, and as indicated in Figure #11, the total within-cluster sum of squares (WSS) shows a marked decrease up to three clusters, suggesting that three is the ideal number. Following this, K-means clustering was applied to the normalized data. The resultant clusters were visualized using Principal Component Analysis (PCA) and cluster plots with ellipses to delineate data distribution and define cluster boundaries. The PCA plot (Figure #12) demonstrates a distinct separation among clusters, each represented by a different color. The cluster plot with ellipses (Figure #12)

further clarifies the distribution and intersection of clusters, offering a more precise visualization of cluster demarcations.

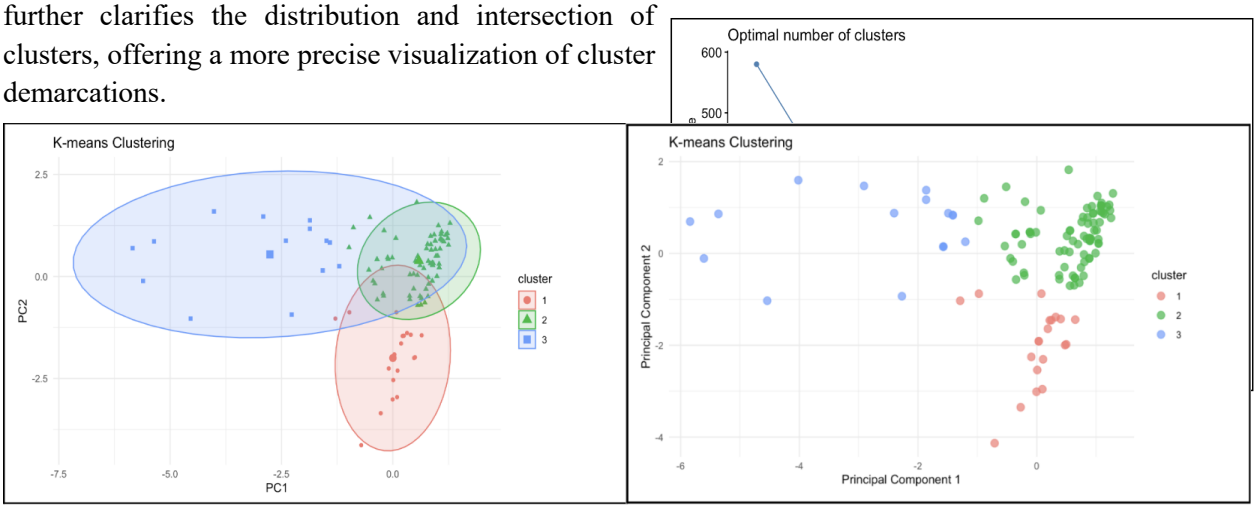

Figure #12: Principal Components Analysis Plot

Each cluster was characterized based on its numerical and categorical attributes. Cluster 1, characterized by higher average prices and retail prices with moderate sales volumes, likely represents luxury items perceived as higher-value products. Given their moderate sales volume but higher pricing, production can be finely tuned to prevent overproduction and reduce waste. Cluster 2 features lower prices but higher sales volumes, typical of mass-market products appealing to a broad customer base. Production processes for these items should be capable of scaling up efficiently to meet robust demand. Retail strategies might include dynamic pricing or bulk purchase discounts to boost revenue. Cluster 3, with moderately priced items but significantly higher sales volumes, seems to offer a balance between affordability and attractiveness, potentially representing a strong value proposition. This cluster would benefit from a flexible inventory management strategy that quickly adapts to fast-selling styles and colors identified earlier in the hierarchical clustering. Periodic promotional campaigns or discounts on less popular variants could sustain high demand levels.

9. Sentiment Analysis

Sentiment analysis is the computational task of automatically determining what feelings are associated with texts. A sentiment is often framed as a binary distinction (positive vs. negative), but it can also be more fine-grained and help identify the specific emotion an author is expressing. In an effort to pinpoint popular fashion items, text mining and sentiment analysis will be used on reviews (Dataset B) to better understand the nuances of how customers feel about different clothing pieces as well as their

attributes (i.e color, type, materials, etc.). Before analyzing the distribution of Dataset B, the overall sentiment valence (positive versus negative) and the emotions conveyed in our review texts is examined.

Overall, the sentiment expressed by customers toward the products listed was predominantly positive, as evidenced by a greater number of positive reviews and high ratings compared to negative feedback and low ratings. The ratio of negative to positive sentiment is approximately 20:80. This is further

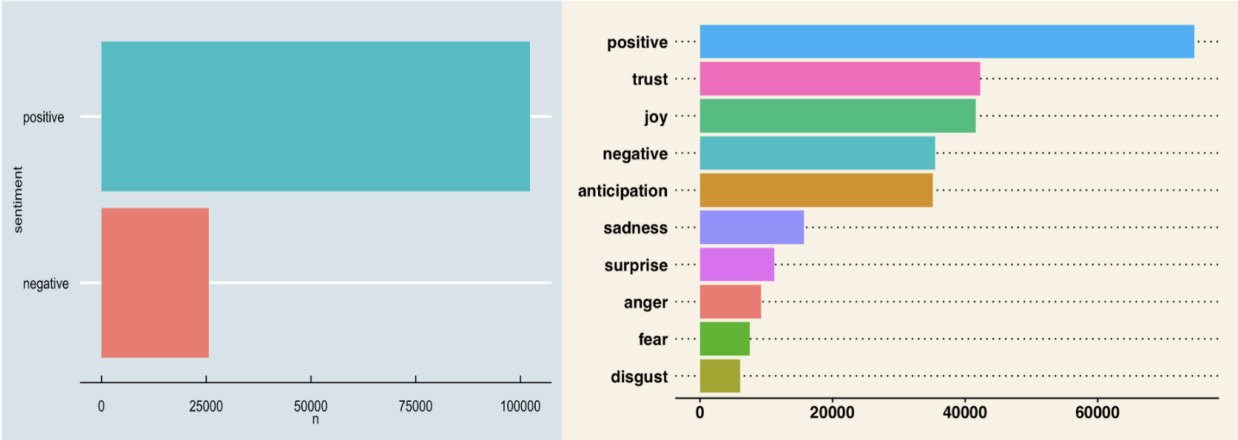

Figure #13: Overall Valence of Review Texts; Overall Sentiment Based on NRC Lexicon

validated by the emotions associated with our products, with more review texts reflecting positive emotions such as trust and joy. The emotion "anticipation" can be interpreted as either positive or negative, but an analysis of the correlation between ratings and text suggests that "anticipation" is generally associated with positive sentiments (linked to high ratings).

It's insightful to examine how review texts and ratings contribute to different valences and emotions, as this provides a deep understanding of customer perceptions towards our products. Identifying which product features and characteristics correlate with certain emotions can inform product improvements. To this end, "nouns" and "pronouns" are extracted from the review texts and analyzed using the udpipe_model, followed by integration with the NRC lexicon.

The analysis reveals that terms like "lace," "weight," and "bottom" tend to be associated with negative emotions (e.g., anger, fear), whereas "top," "compliment," and "store" are linked to positive emotions (e.g., anticipation, joy, trust).

From Figure #15, the top five popular colors—**red, floral (color/pattern), black, blue,** and **white**—are observed. Red, in particular, is often associated with positive emotions like joy, trust, and anticipation. The most popular materials are **denim, stretch, flannel, cotton**, and **wool**, with stretch, cotton, and wool showing the least association with negative emotions (e.g., sadness, negativity, disgust). The top

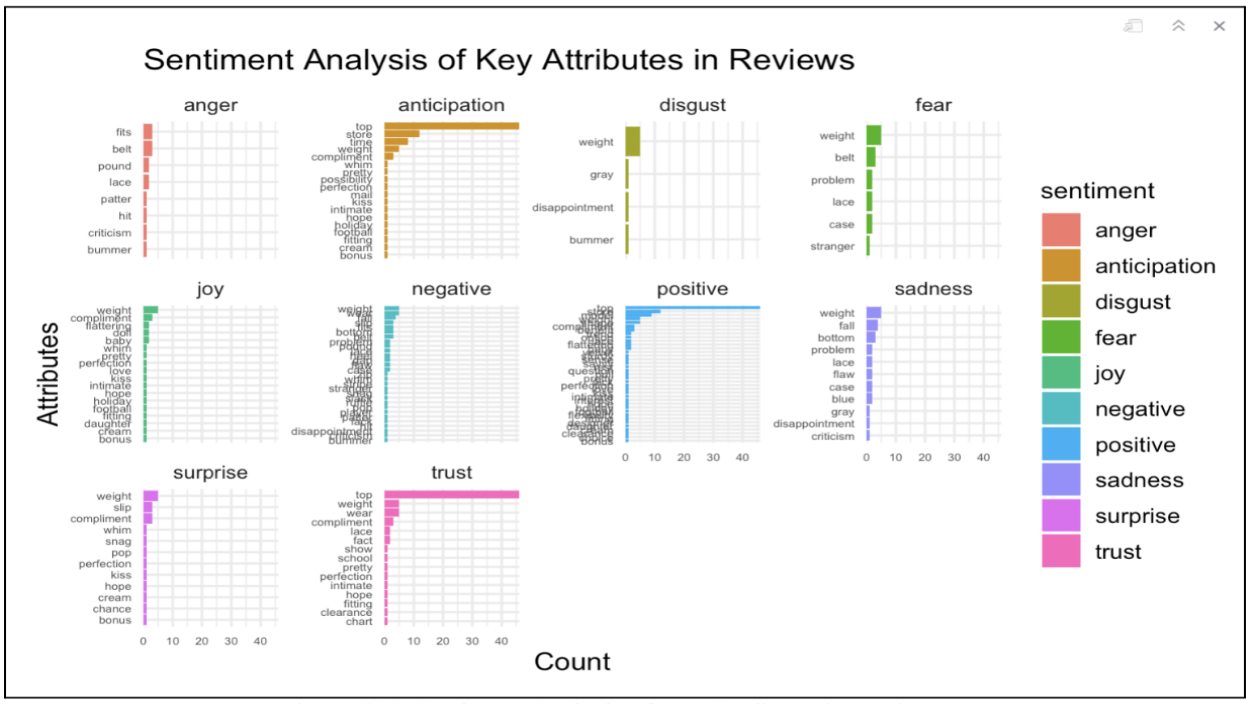

Figure #14: Sentiment Analysis of Key Attributes in Reviews

patterns are **floral, stripe, flannel, plaid**, and **printed**, and the most popular styles are **fall, mod, spring, casual,** and **summer**.

While sentiment analysis, including lemmatization and tokenization processes, is not flawless—often lacking granularity or context for specific sentiments—it remains a useful tool for summarizing text qualities, especially when the volume of text precludes detailed human analysis. Additionally, considering negation is crucial as it can reverse the polarity of expressions (e.g., "not bad" could be positive but might be tagged negatively when tokenized separately). In conjunction with the cluster analysis, the sentiment analysis offers valuable insights into what aspects of certain products customers particularly enjoy. This provides retailers with clear guidance on which features to emphasize in their design and production efforts.

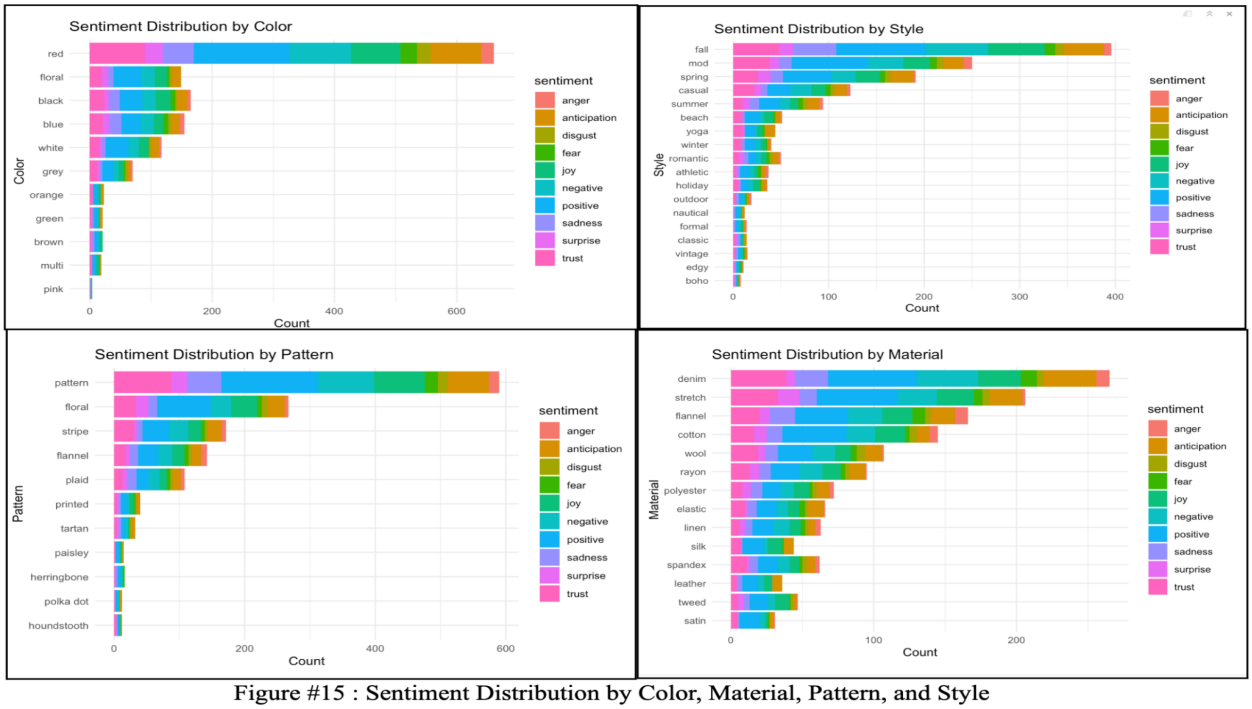

Figure #15 : Sentiment Distribution by Color, Material, Pattern, and Style

10. Conclusion & Recommendations

Through clustering and sentiment analysis, the findings of this study reveal significant influence of product characteristics such as color, style, and price, as well as the sentiment expressed in product reviews on consumer preferences and purchasing patterns. Within clustering, product segments provided actionable insights on specific characteristics within the current summer collection that made products desirable, as well as the optimal price points for different customer demographics. Higher-priced clusters demonstrated lower sales volumes. However, median-priced clusters demonstrated higher sales volumes than lower-priced clusters, suggesting a price sensitivity among summer fashion consumers. Implementing dynamic pricing models that adjust prices based on demand, competitor pricing, and inventory levels as well as strategic discounting will aid in maximizing profit margins while remaining competitive.

Products with predominantly positive review sentiments consistently showed higher sales volumes. This underscores the importance of customer perception and its direct impact on purchasing decisions. Prioritizing stocking such products will likely attract more customers and generate higher revenue. Insights from clustering and sentiment analysis about the top reviewed products can also inform product design decisions, materials used and other features for future collections. By adopting these recommendations retailers can not only enhance their operational efficiency and customer satisfaction but also strategically position themselves to capitalize on the dynamic nature of fashion trends, ultimately driving sustainable growth and profitability.