Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

A Comprehensive Statistical and Machine Learning Study on Myntra Fashion Products

Rohan J. Kende
Department of Statistics
Dr. D. Y. Patil Arts, Commerce and Science College,
Pune, India

Pranit. B. Bolli
Department of Statistics
Dr. D. Y. Patil Arts, Commerce and Science College,
Pune, India

Abstract - The rapid growth of e-commerce platforms has generated large volumes of product and customer interaction data, creating opportunities for data-driven decision-making in online fashion retail. This study presents a comprehensive analytical framework for examining Myntra fashion product data using statistical methods, exploratory data analysis (EDA), natural language processing (NLP), and machine learning techniques. A dataset containing product attributes such as brand, category, price, discount percentage, ratings, reviews, and textual descriptions was cleaned and preprocessed to ensure analytical reliability. Exploratory analysis revealed strong patterns in pricing distribution, discount strategies, brand performance, color and size preferences, and rating behavior. Statistical hypothesis testing, including ANOVA, t-tests, chi-square tests, correlation, and regression analysis, confirmed significant differences in discount percentages across brands and a strong positive relationship between list

price and sale price. NLP techniques such as TF-IDF keyword extraction, text
clustering, and sentiment analysis identified dominant product themes and predominantly positive description sentiment. Multiple machine learning models, including Decision Tree, Support
Vector Machine, Random Forest, and KNN, were applied for price segment and discount-level prediction, showing high classification performance. A content-based recommendation system was also developed using category, price, ratings, and review features to generate personalized product suggestions. The results demonstrate that integrated statistical and machine learning approaches can effectively extract actionable insights, support pricing and promotion strategies, and enhance recommendation capabilities in online fashion retail environments.

Keywords
E-commerce Analytics, Myntra, Fashion Products, Exploratory Data Analysis, NLP, Sentiment Analysis, Machine Learning, Recommendation System, Statistical Testing, Pricing and Discount Analysis.

I. INTRODUCTION

In recent years, online fashion platforms have transformed the retail industry by providing large-scale digital catalogs and personalized shopping experiences. Platforms such as Myntra host thousands of products across brands, categories, and price ranges. This generates rich datasets that can be analyzed to understand consumer preferences, pricing structures, and promotional strategies.

This study focuses on applying statistical analysis, exploratory data analysis, natural language processing, and machine learning methods to Myntra fashion product data. The purpose is to extract meaningful insights related to pricing patterns, brand discount strategies, rating behavior, and textual product descriptions. The research also aims to build predictive and recommendation models that replicate real-world e-commerce intelligence systems.

II. OBJECTIVES

- Identify product and brand trends
- Analyze price and discount behavior
- Apply hypothesis testing and regression
- Perform NLP on product descriptions
- Build clustering and sentiment models
- Develop a recommendation system

III. DATA COLLECTION

The dataset contains about 4000 Myntra product records with attributes such as brand, category, color, size, list price, sale price, discount percentage, ratings, and descriptions, supporting both statistical and NLP analysis.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

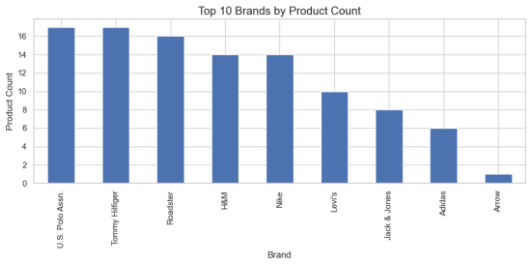
IV. DATA CLEANING AND PREPROCESSING

Data cleaning included duplicate removal, missing value handling, numeric conversion, text normalization, and feature engineering such as weighted ratings, TF-IDF vectors, and price segments.

V. EXPLORATORY DATA ANALYSIS

EDA shows major brand dominance, mid-range price concentration, black color preference, and mostly low-to-moderate discount distribution. Insert PPT graphs: brand column chart, word map, pie chart, histogram, and Pareto chart here as figures.

Fig. 1 — Brand Product Count Column Chart

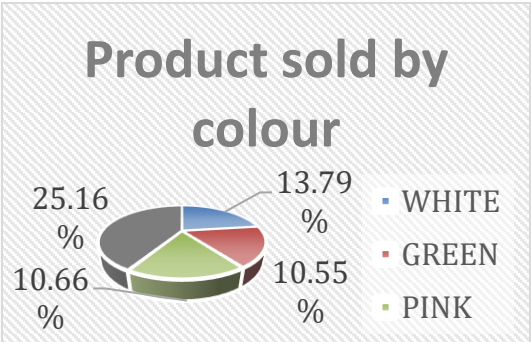


Conclusion: From above graph we can say that most product are sold by U.S. Polo , Tommy Hilfiger, Roaster.

Fig. 2 — Brand Word Map

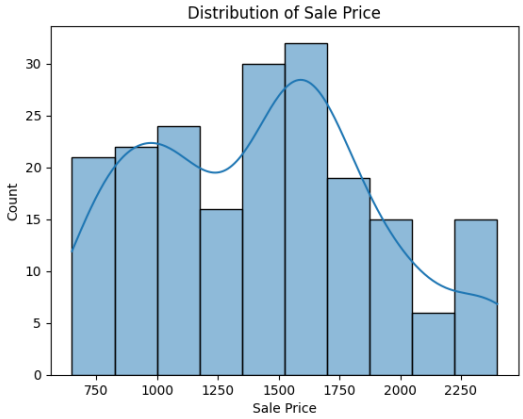
Conclusion: This word map shows that the project is related to all these brands

Fig. 3 — Product Color Pie Chart



Conclusion : From above graph we can say that Black coloured products are sold mostly also we can say that peoples likes black colour as compared to others.

Fig. 4 — Sale Price Histogram



Conclusion: From above graph, we can say that most products are priced betn 1000 to 1750 and few have higher price than 2000.

Fig. 5 — Discount Pareto Chart

Conclusion : This graph shows that most products are have low discounts mainly between 0 to 20%. A smaller number of products fall in the higher discount categories.

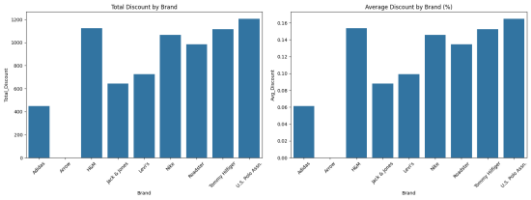
The trend line decreases sharply, which indicates that as discount increases the number of items offering that discount decreases.

VI. PRICE AND DISCOUNT ANALYSIS

Analysis shows sale price is consistently below list price, strong positive price correlation, and stable ratings even at higher discounts. Insert PPT scatter plots, line charts, heatmaps, and boxplots here.

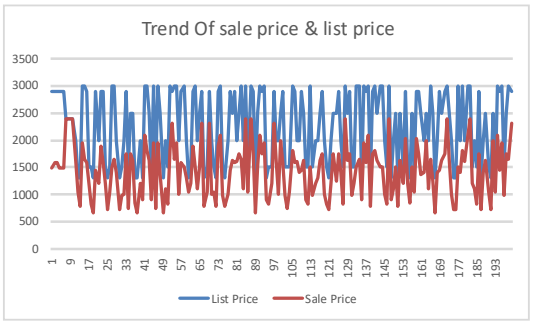
International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

Fig. 6 — Brand Discount Column Chart



Conclusion : This graph shows U.S. Polo, H&M , Tommy Hillfiger gives more discount as compared to other brands.

Fig. 7 — List vs Sale Price Line Chart

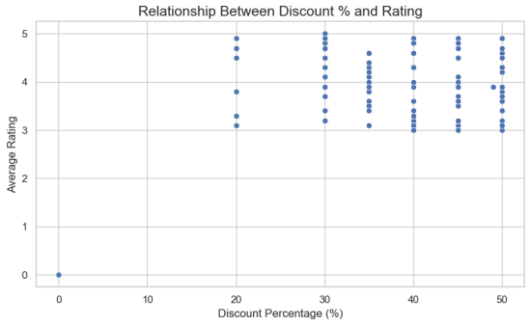


Conclusion : The graph compares the List Price and Sale Price trends across multiple products. It clearly shows that the Sale Price is consistently lower than the List Price indicating that most products are sold at a discount.

Fig. 8 — List vs Sale Price Scatter Plot

Conclusion : The scatter plot shows a strong positive relationship between list price and sale price. As the list price increases, the sale price also increases, but sale prices always remain lower than list prices.

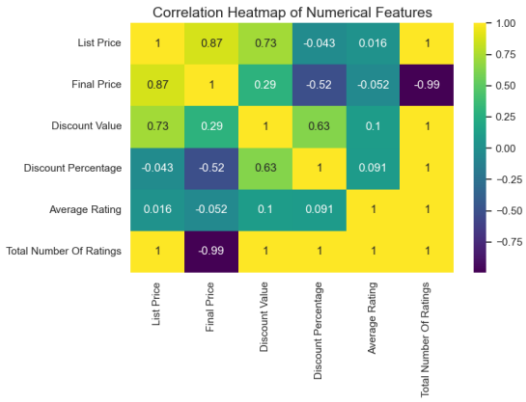
Fig. 9 — Discount vs Rating Scatter Plot



Conclusion :

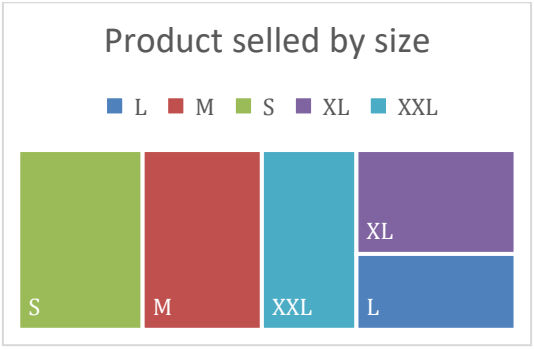
The scatter plot shows that products with moderate to high discounts (20%–50%) still maintain strong average ratings, mostly between 3.0 and 5.0. This indicates that higher discounts positively affect customer ratings.

Fig. 10 — Correlation Heatmap



Conclusion : Correlation heatmap confirms strong positive correlation between list and final price and negative relation with discount percent.

Fig. 11 — Size Tree Map



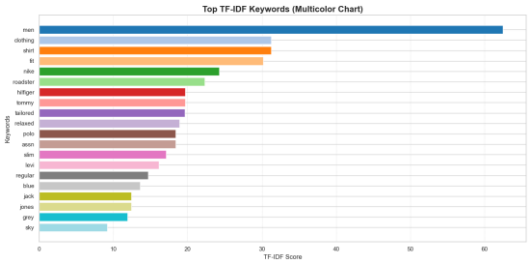
Conclusion : This shows that most products of size S, M are sold as compared to other sizes.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

VII. NLP Analysis

TF-IDF reveals dominant keywords related to men's apparel and brands. Sentiment analysis shows mostly positive product descriptions. Insert TF-IDF and sentiment graphs from PPT here.

Fig. 12 — TF-IDF Keyword Chart



Conclusion : The TF-IDF keyword chart highlights the most important and frequently appearing terms in the product descriptions.

Fig. 13 — Sentiment Score Output

	description	sentiment
0	Black printed Kurta with Palazzos with dupatta...	0.083333
1	Orange solid Kurta with Palazzos with dupatta<...	0.233333
2	Navy blue embroidered Kurta with Trousers with...	0.060000
3	Red printed kurta with trouser and dupatta ...	0.185714
4	Black and green printed straight kurta, has a ...	-0.033333
5	Stately and versatile, this kurta set will be ...	0.334694
6	When in doubt, rock this simple yet stylish ku...	0.111722
7	Yellow and white printed kurta with palazzos<b...	0.032857
8	Green & pink printed kurta with palazzos &...	0.016667
9	This set includes: Kurta and Trousers A...	0.105952

	sentiment_label
0	Positive
1	Positive
2	Positive
3	Positive
4	Negative
5	Positive
6	Positive
7	Positive
8	Positive
9	Positive

Conclusion : The sentiment analysis of product descriptions shows that the majority of the kurta-related descriptions carry a positive sentiment.
Most sentiment polarity values are above 0, indicating favourable wording such as "stylish," "versatile," "embroidered," and "printed," which contribute to a positive perception of the products.

VIII. STATISTICAL TESTING

t-test shows significant difference between list and sale price ($p < 0.05$).
ANOVA shows significant difference in discount percentages across brands.
Chi-square shows no association between brand and sub-category.
Correlation between list and sale price ≈ 0.84.

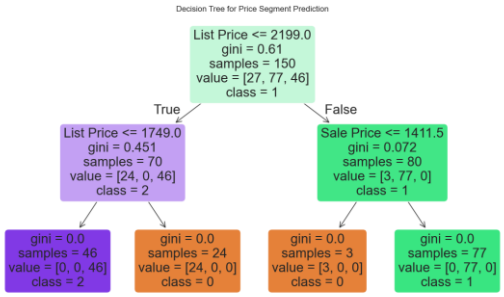
Regression model:

Y = 603.87 + 1.15X
$R^2 \approx 0.70$.

IX. MACHINE LEARNING MODELS

Decision Tree, SVM, Random Forest, KNN, and Regression models achieved high classification performance for price and discount segments.

Fig. 15 — Decision Tree Diagram



Conclusion : The model clearly learns that product price (both list and sale price) strongly determines the price segment. Most classifications are highly pure, showing that the decision tree effectively separates budget, mid-range, and premium products using only price-related features.

X. RECOMMENDATION SYSTEM

A content-based recommendation system using category, price, ratings, and reviews provides similar product suggestions. Insert recommendation output figure here.

Fig. 16 — Recommendation Output Table

	Product_id	BrandName	Category	Individual_category
5	2490950	Mast & Harbour	Western	tops
414	11373682	Roadster	Western	tops
1585	8330195	Roadster	Western	tops
730	2490953	Mast & Harbour	Western	tops
1187	11373642	Roadster	Western	tops
380	11735666	her by invictus	Western	tops
1336	2490952	Mast & Harbour	Western	tops
713	807931	La Zoire	Western	tops
276	12009564	Tokyo Talkies	Western	tops
1628	10964506	Roadster	Western	tops

	category_by_Gender	OriginalPrice (in Rs)	Discount (in Rs)
5	Women	599.0	359.40
414	Women	599.0	269.55
1585	Women	499.0	299.40
730	Women	599.0	329.45
1187	Women	599.0	239.60
380	Women	699.0	349.50
1336	Women	599.0	359.40
713	Women	699.0	559.20
276	Women	799.0	423.47
1628	Women	599.0	329.45

	Discount (in %)	Ratings	Reviews	SizeOption
5	40	4.4	999	XS, S, M, L, XL
414	55	4.2	918	XS, S, M, L, XL
1585	40	4.2	749	XS, S, M, L, XL
730	45	4.2	866	XS, S, M, L, XL
1187	60	4.5	798	XS, S, M, L, XL
380	50	4.3	925	S, M, L, XL, XXL
1336	40	4.3	782	XS, S, M, L, XL
713	20	4.3	868	S, M, L, XL
276	47	4.4	946	S, M, L, XL
1628	45	4.2	742	XS, S, M, L, XL

	Description	weighted_score
5	mast harbour women yellow solid tank top	7.338230
414	roadster women white black solid a line pure ...	6.436728
1585	roadster women mustard yellow styled back pure...	6.304208
730	mast harbour women black solid top	6.072120
1187	roadster women black white solid a line pure ...	5.994992
380	her by invictus navy blue white floral pleate...	5.690272
1336	mast harbour women pink solid tank top	5.613689
713	la zoire green sheer knotted top	5.339628
276	tokyo talkies navy blue white vertical stripe...	5.209512
1628	the roadster lifestyle co women mustard yellow...	5.202671

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

Conclusion : A content-based recommendation system was developed using product category, price, ratings, and reviews. The system computes similarity and suggests related products. This demonstrates how recommendation engines can be built using product metadata without user history.

XI. CONCLUSION

Statistical analysis, NLP, and machine learning methods effectively analyze fashion e-commerce data and reveal useful business insights. The study identifies strong relationships between list price, sale price, and discount strategies across major brands. Results show that higher discounts do not reduce customer ratings, indicating maintained perceived quality. NLP techniques highlight positive sentiment and strong brand emphasis in product descriptions. Machine learning models successfully predict price and discount segments, while the recommendation system provides relevant product suggestions. Overall, these approaches support better pricing decisions, targeted promotions, personalization, and improved customer experience in online fashion retail platforms

XII. REFERENCE

[1] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.
[2] Benesty, J., Chen, J., Huang, Y., & Cohen, I. (2009). Pearson Correlation Coefficient. Springer.
[3] Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning.
[4] Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). Classification and Regression Trees. CRC Press.
[5] Ricci, F., Rokach, L., & Shapira, B. (2011). Recommender Systems Handbook. Springer.
[6] Salton, G., & McGill, M. J. (1983). Introduction to Modern Information Retrieval. McGraw-Hill
[7] MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. University of California Press.
[8] Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. Springer.
[9] Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.
[10] Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.
[11] Jurafsky, D., & Martin, J. H. (2023). *Speech and Language Processing*. Pearson.
[12] Aggarwal, C. C. (2015). *Data Mining: The Textbook*. Springer.
[13] Aggarwal, C. C. (2016). *Recommender Systems: The Textbook*. Springer.
[14] Liu, B. (2012). *Sentiment Analysis and Opinion Mining*. Morgan & Claypool.
[15] Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques*. Morgan Kaufmann.
[16] Pedregosa, F. et al. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*.
[17] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space