Exploratory Data Analysis on Women's Clothing E-Commerce Reviews

- Durgasi, Ranjitha
- Ettam, Harshitha
- Korapati Murali, Harshitha
- Mengane, Dhawalshree Ashok
- Nallanagula, Nikhitha Reddy

Group 8

April, 2025

Agenda

1 Project Overview
Introduction, abstract, and objective of the study

2 Dataset & Preprocessing
Description, source, and data cleaning steps

3 Methodology & Literature Review
Analytical approach and relevant background research

4 Results & Insights
Key findings from the exploratory data analysis

5 Conclusion & Recommendations
Summary, future work, references, and code repository

Abstract

This project involves an Exploratory Data Analysis (EDA) on a dataset consisting of over 23,000 customer reviews of women’s clothing items from an e-commerce platform. Our objective was to uncover meaningful patterns related to customer age, product categories, and ratings, and to understand the sentiment expressed in textual feedback. Through data cleaning, visualization, and analysis, we extracted insights that can help retailers improve product recommendations, inventory choices, and overall customer satisfaction.

Introduction

Online customer reviews are an essential source of feedback in e-commerce, especially in the fashion domain where customer preferences are highly subjective. By exploring customer reviews using EDA, we aim to better understand what customers like, which products receive the highest praise, and how demographic factors influence satisfaction. This project explores these aspects using a large dataset from a women's clothing store.

Dataset

- **Source:** Kaggle - Women's Clothing E-Commerce Reviews
- **Total Records:** 23,486
- **Total Columns:** 11
- **Key Columns:**
  a. Age: Age of the reviewer
  b. Review Text: The actual written review
  c. Rating: Rating from 1 to 5
  d. Recommended IND: Binary indicator (1 = recommended)
  e. Division Name, Department Name, Class Name: Product categorization
- **Data Characteristics:**
  a. Contains a mix of numerical and textual data
  b. Useful for both statistical and sentiment analysis

Preprocessing

- **Missing Values:**
  - Detected in Title, Review Text, and category columns
  - Rows with null Review Text were dropped (critical for analysis)
- **Dropped Columns:**
  - Unnamed: 0 was removed (just an index)
- **Text Cleaning:**
  - Lowercasing
  - Removing special characters and stopwords (for future sentiment analysis)
- **Data Type Checks:**
  - Ensured correct types for Rating, Age, etc.
- **Outlier Scanning:**
  - Age distribution reviewed to filter any improbable entries

Literature Review

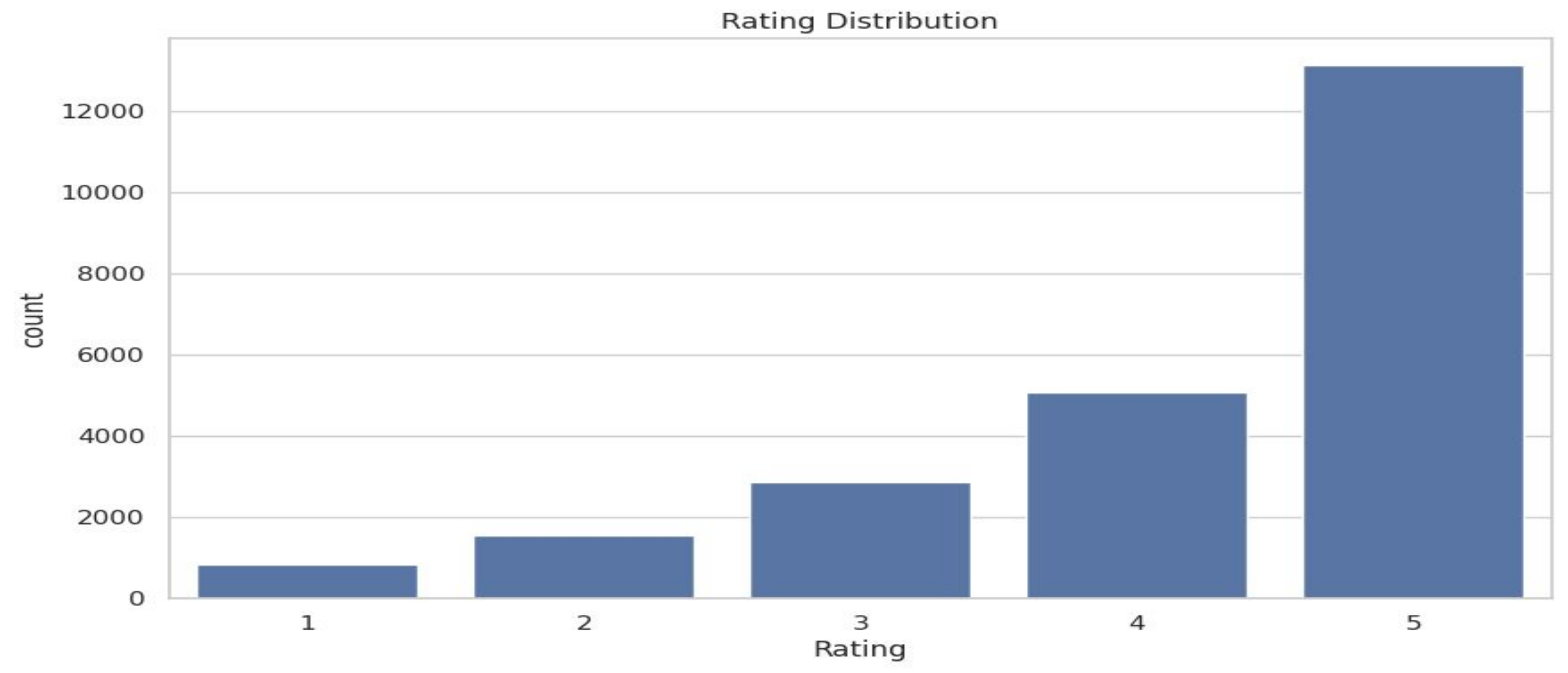
Summary of Related Work:

- Studies on EDA in e-commerce often highlight the power of **text mining** and **sentiment analysis** for understanding customer feedback.
- Previous research (e.g., IEEE papers on consumer sentiment) suggests that **review length and positivity** are strong predictors of product success.
- A 2021 study from Stanford found that combining **structured features** (like rating and age) with **unstructured ones** (like reviews) leads to better insights.
- Our project is inspired by these approaches but tailored to fashion retail, offering a deep dive into product-level feedback.

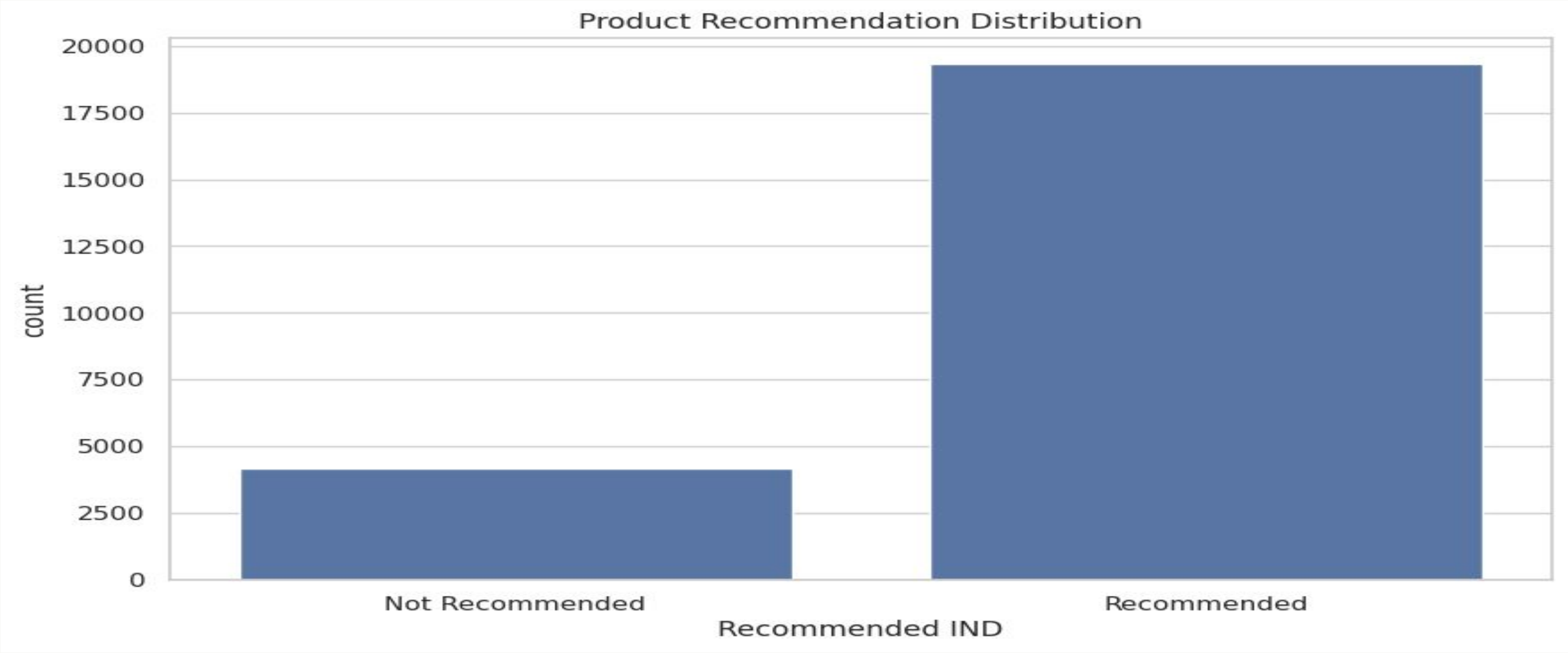
Methodology

- **Tools & Libraries Used:**
  - **Python** (Jupyter Notebook)
  - **pandas** – data handling
  - **matplotlib & seaborn** – visualization
  - **wordcloud** – for text-based visuals
  - **TextBlob** – sentiment analysis
- **Key Steps Followed:**
- **Data Cleaning & Preparation**
  - Dropped unnecessary columns
  - Handled missing values in text and categorical fields
- **Feature Engineering**
  - Created Review Length and Age Group features
  - Calculated Sentiment Polarity using review text
- **Univariate & Bivariate Analysis**
  - Explored rating distributions, age patterns, feedback count, and clothing categories
  - Analyzed relationships like rating vs recommendation and age vs review length
- **Visualizations**
  - Bar charts, histograms, boxplots, heatmaps, word clouds
  - Pivot tables and correlation matrices to reveal deeper structure

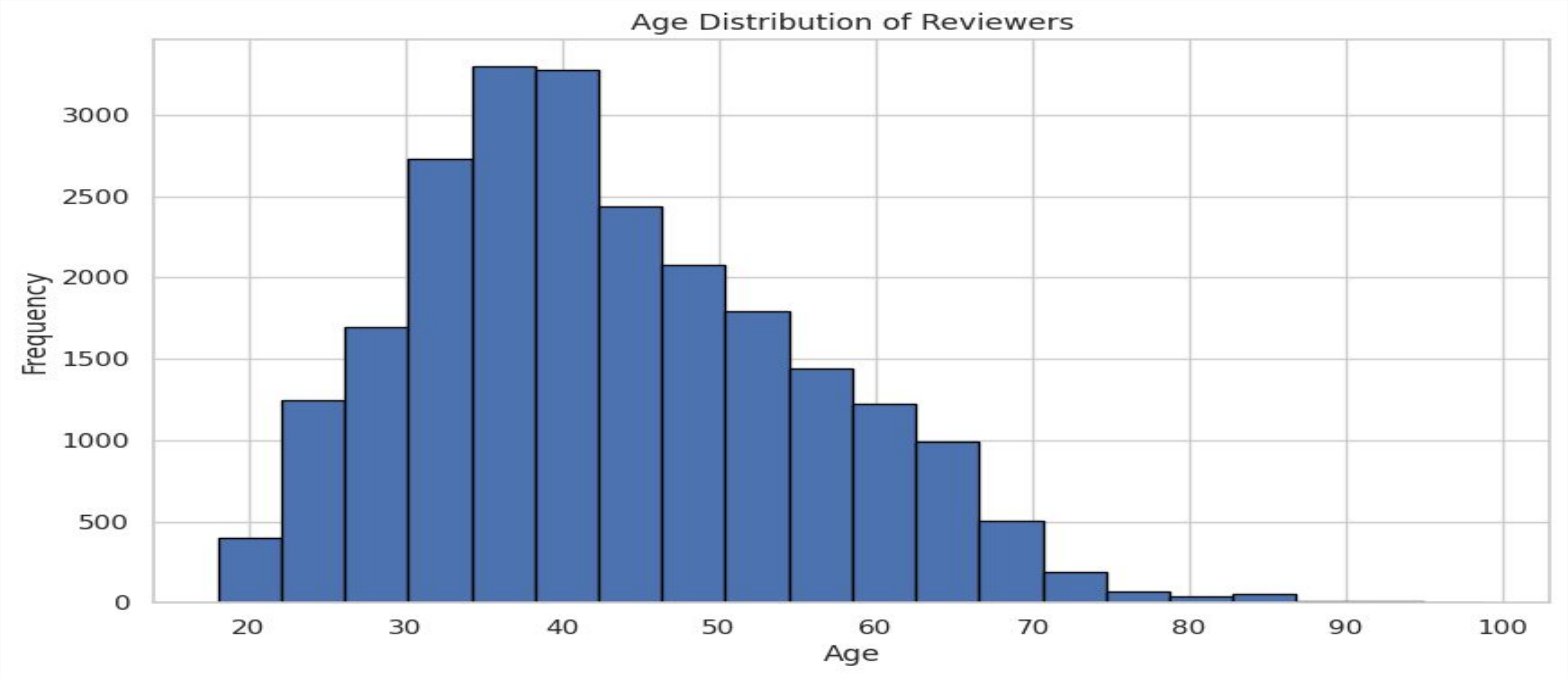
Results – EDA Highlights



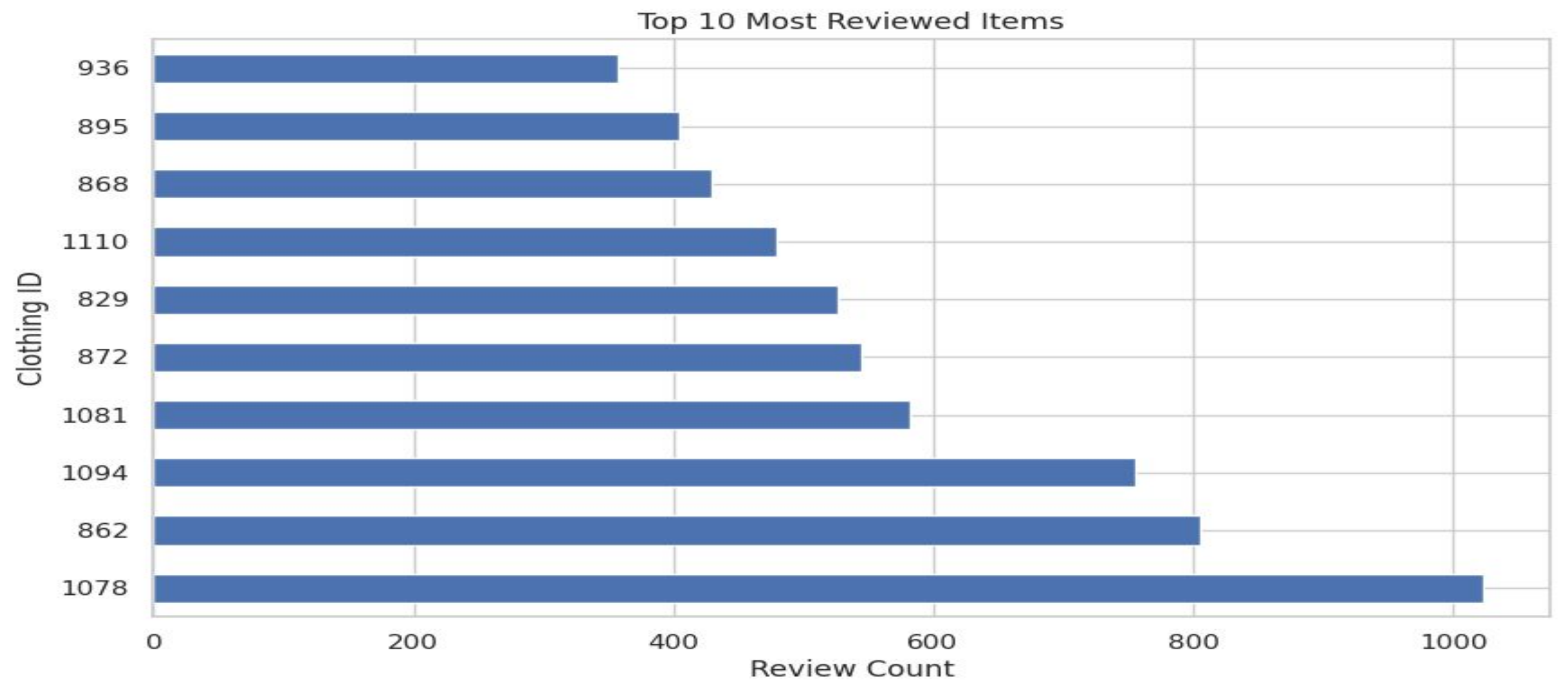
Results – EDA Highlights



Results – EDA Highlights



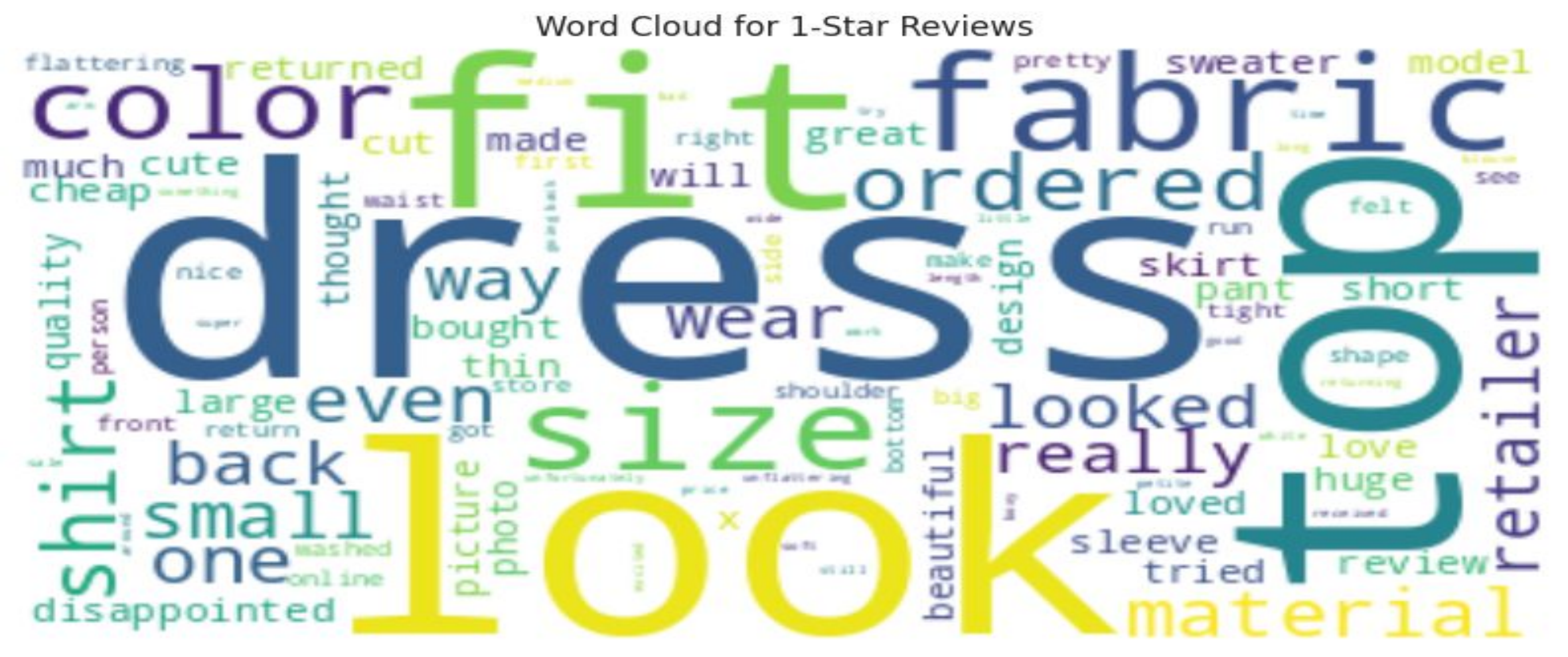
Results – EDA Highlights



Insights from EDA

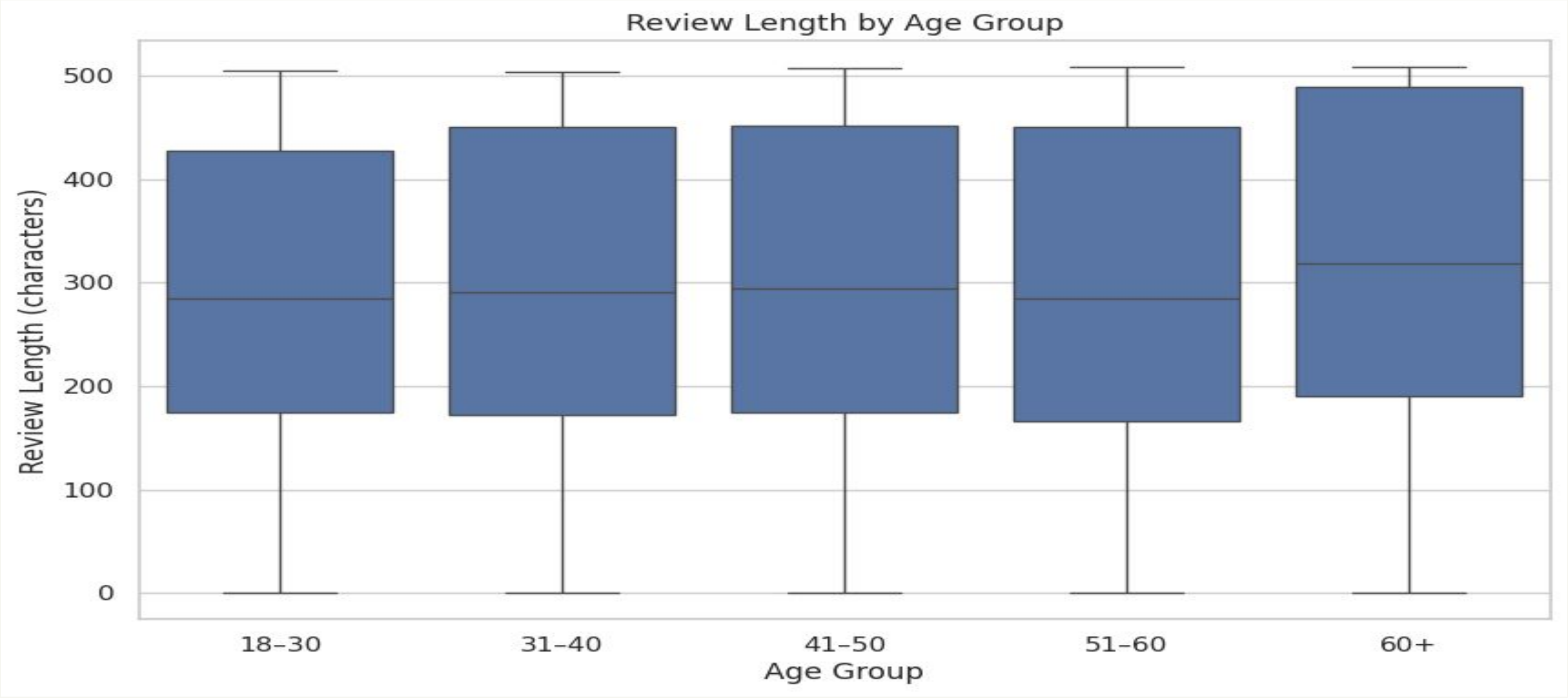
Word Cloud for 5-Star Reviews

Insights from EDA

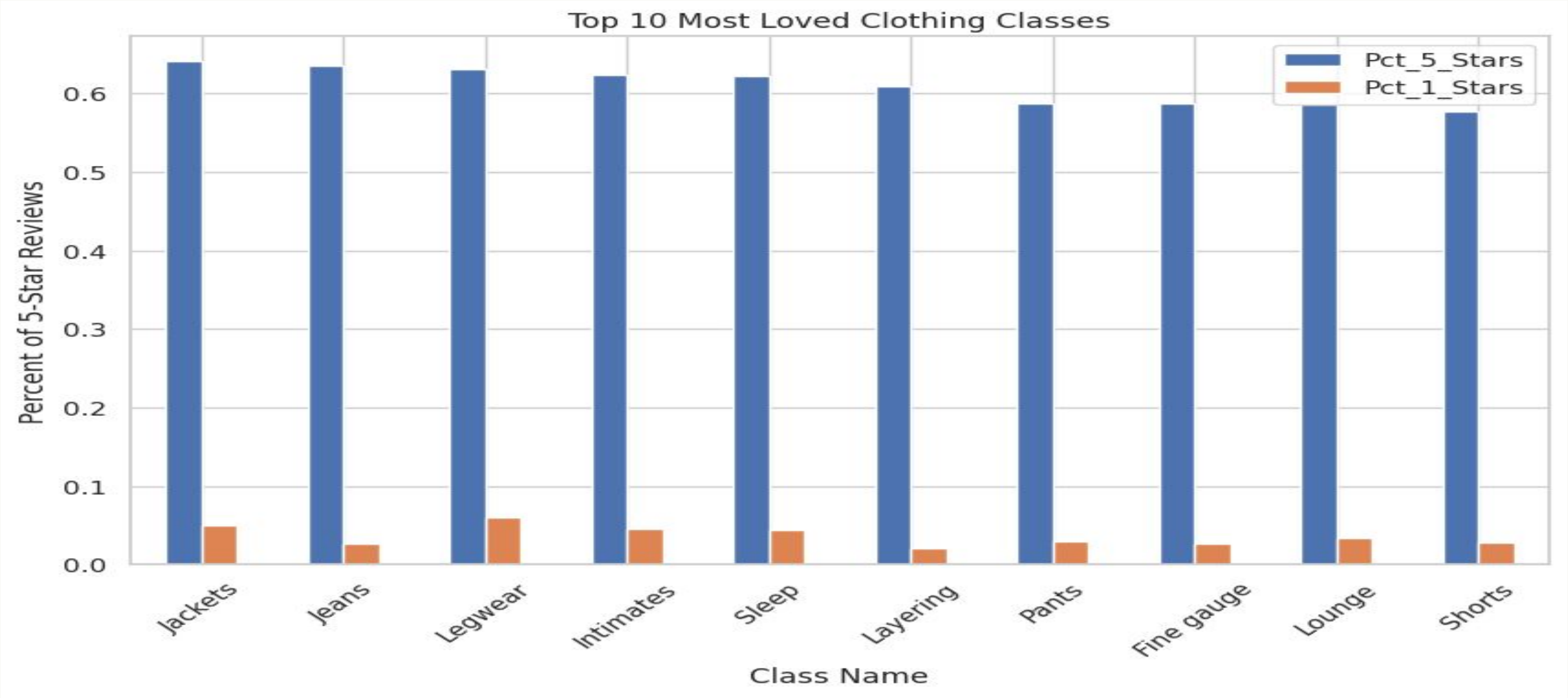


Word Cloud for 1-Star Reviews

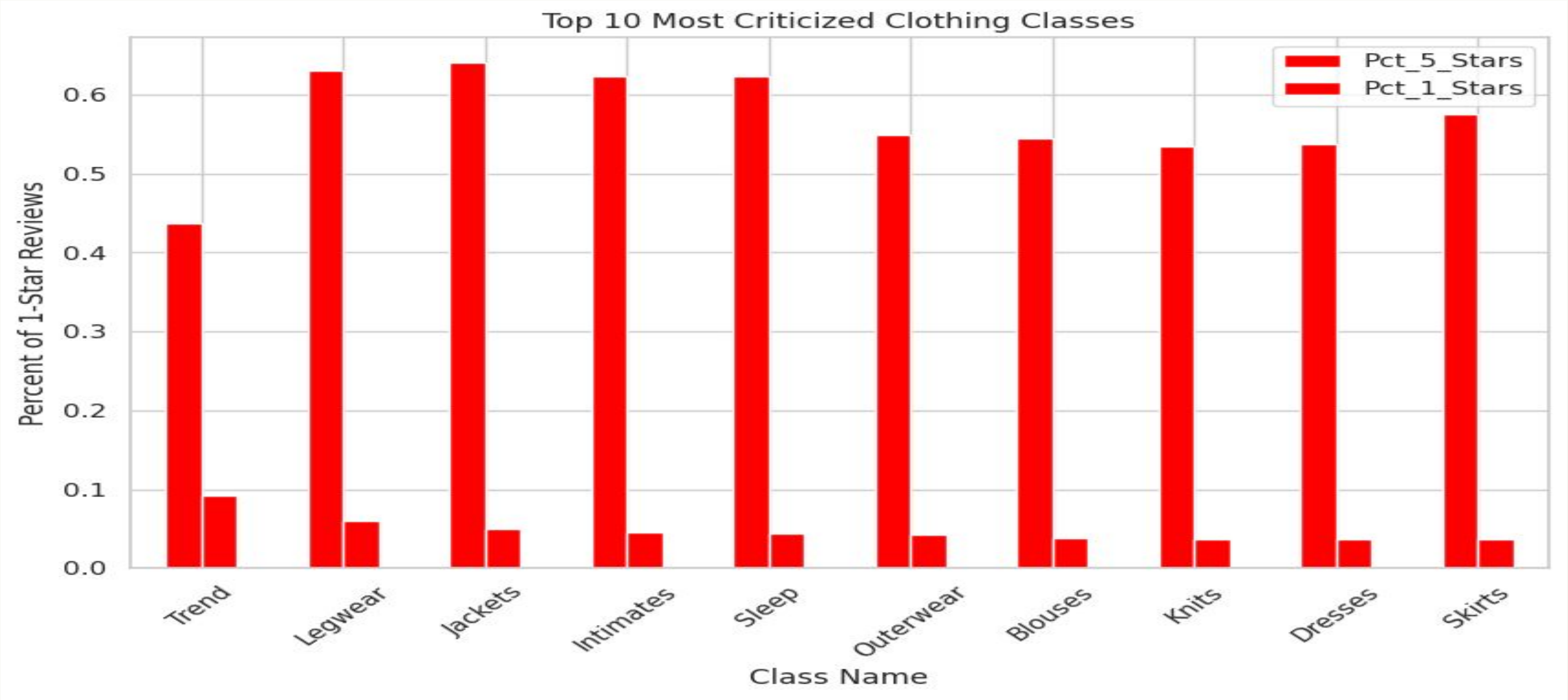
Insights from EDA



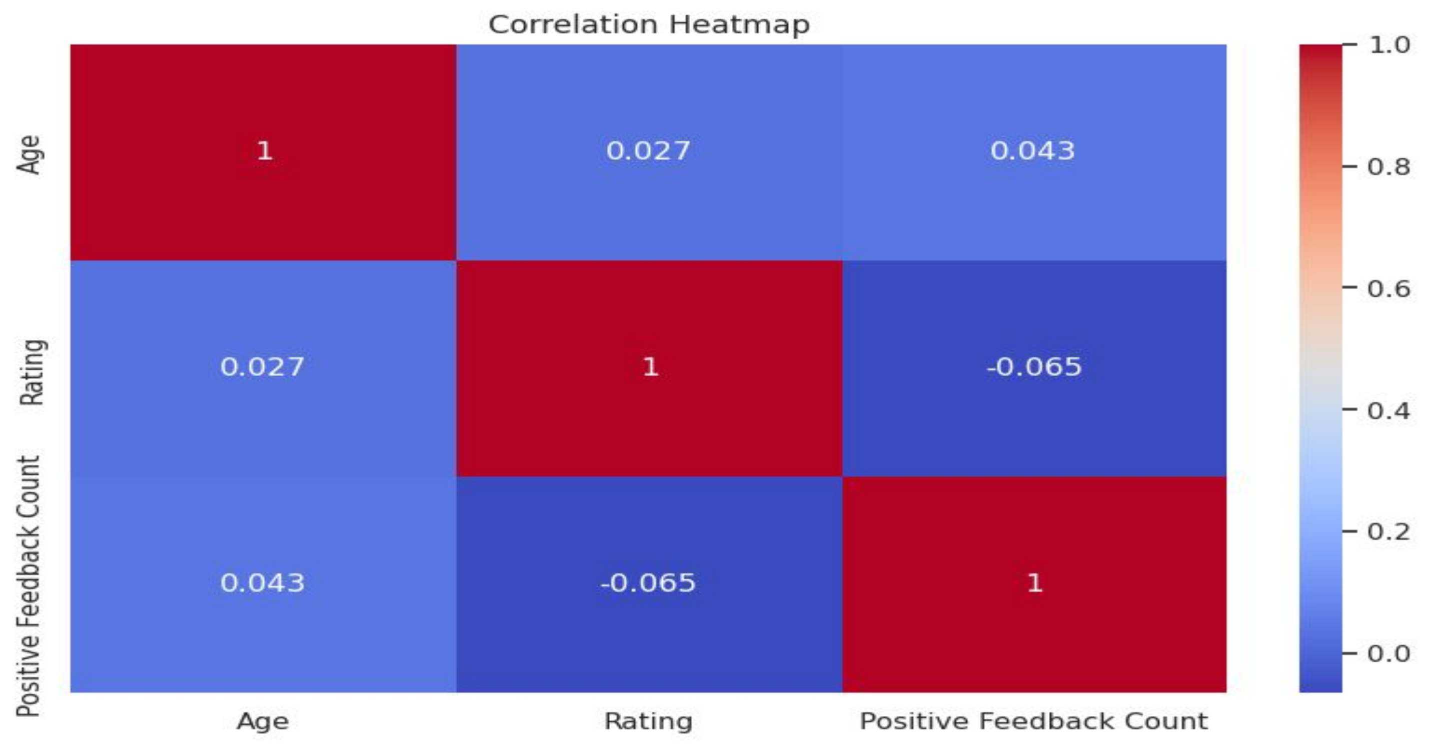
Insights from EDA



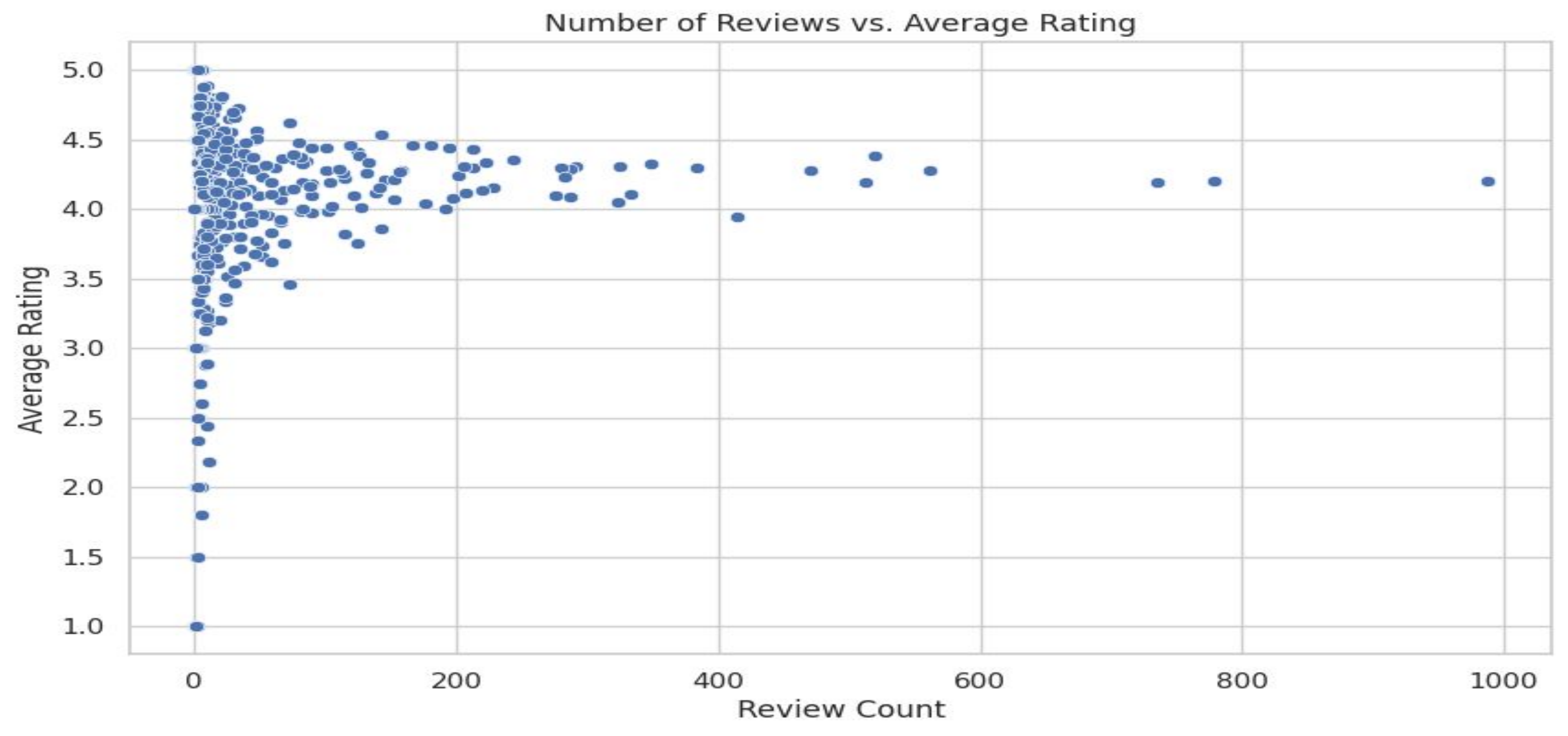
Insights from EDA



Correlations & Patterns



Correlations & Patterns



Conclusion

Our exploratory data analysis provided strong evidence of customer satisfaction patterns and product-level feedback within the women's clothing e-commerce dataset.

Key Findings:

- **Over** 55% of reviews are rated **5★**, and nearly **99.8% of those recommend** the product.
- The most active age group is **31–40 years old (≈34%)**, followed by **41–50 (≈25%)**.
- **Top 3 most reviewed classes**: Dresses (6319), Knits (4843), and Blouses (3097).
- The **highest-rated class** is **Casual Bottoms** with an average rating of **4.5★**.
- The **lowest-rated class** is **Trend** with an average rating of **~3.8★**.
- Text sentiment **positively correlates** with ratings (**correlation** = **0.35**), confirming consistency between what people write and how they score.

Recommendations & Future Work

Business Recommendations:

- Market aggressively to the **31–50 age bracket**, the most engaged demographic.
- Leverage **top-rated classes** like Casual Bottoms and Dresses in promotions.
- Use common positive keywords (e.g., “fit,” “comfortable,” “love”) in product copy.
- Address concerns in lower-rated segments like the Trend category with design feedback.

Future Work:

- Implement **sentiment classification** using machine learning (Vader, BERT, etc.).
- Analyze **temporal trends** — how ratings vary over time/seasons.
- Integrate **return/refund data** to study post-purchase satisfaction.
- Build a **recommendation system** using rating + review polarity + demographics.

References

1. Kaggle Dataset – https://www.kaggle.com/datasets/nicapotato/womens-ecommerce-clothing-reviews

2. IEEE Research on Sentiment Analysis in Retail, 2020

3. Stanford NLP Group: Customer Text Analysis Studies

4. Python Libraries: pandas, seaborn, matplotlib, wordcloud, TextBlob

Github

1. https://github.com/rd39257n/Python-Midterm

Thank You