

## ✓ 1. Introduction

The data being analyzed consists of consumer reviews for women's clothes received from an E-Commerce platform. client reviews are an important source of feedback for businesses, providing information on product satisfaction, client preferences, and areas for improvement. Analyzing such data can reveal consumer behavior trends, identify popular products, and help businesses make more educated decisions.

### 1.1. Dataset Overview

The dataset consists of 23,486 rows and 11 columns, each representing a unique customer review. Below is a detailed list of the key features available for analysis:

1. **Clothing ID:** Unique identifier for each clothing item being reviewed. This serves as a categorical variable.
2. **Age:** The age of the reviewer, represented as a positive integer. This demographic information can provide insights into the preferences of different age groups.
3. **Title:** The title of the review provided by the customer. This text can offer a concise summary or sentiment of the review.
4. **Review Text:** The main body of the review, where customers provide detailed feedback about their experience with the product. This text is rich in sentiment and opinion.
5. **Rating:** An ordinal integer variable indicating the product score given by the customer, ranging from 1 (Worst) to 5 (Best). This serves as a measure of customer satisfaction.
6. **Recommended IND:** A binary variable (0 or 1) indicating whether the customer recommends the product. This is a direct indication of product satisfaction.
7. **Positive Feedback Count:** The number of other customers who found the review helpful or positive. This can reflect the perceived usefulness or popularity of the review.
8. **Division Name:** Categorical variable indicating the high-level product division to which the clothing item belongs (e.g., General, General Petite, etc.).
9. **Department Name:** Categorical variable indicating the specific department of the clothing item (e.g., Dresses, Bottoms, Tops, etc.).
10. **Class Name:** Categorical variable indicating the type or class of the clothing item (e.g., Dresses, Blouses, Pants, etc.).
11. **Cleaned Text:** A processed version of Review Text after removing stopwords, punctuation, and performing stemming, which aids in text analysis.

### 1.2 Objective

The major goal of this analysis is to get insights into customer satisfaction and preferences for women's clothes based on evaluations posted on an E-Commerce site. By analyzing the provided data, we want to:

1. **Understanding Customer Satisfaction Factors:**
  - Determine the important aspects that influence customer evaluations and recommendations.
  - Determine how different product attributes (such as Division Name, Department Name, and Class Name) affect customer satisfaction.
2. **Forecast Product Recommendations:** - Create models that forecast whether customers would recommend a product based on reviews and ratings.
  - Use sentiment analysis on review texts to help forecast product suggestions and general customer sentiment.

### 1.3 Problem Statement

To achieve the objectives outlined, the analysis will address the following specific questions:

1. **What are the most important factors influencing ratings?**
  - Determine which characteristics (such as age, product division, department, and class) are most significantly connected with better or lower ratings.
  - Identify patterns and trends in ratings across customer demographics and product categories.
2. **What Factors Influence Recommendations?**
  - Investigate the correlation between ratings and the likelihood of recommending a product.
  - Identify any other aspects that may influence a customer's recommendation of a product in addition to the rating assigned.

### 3. How Can Sentiment Analysis Help Predict Recommendations?

- Apply sentiment analysis to the review text to obtain sentiment scores.
  - Look at the association between sentiment ratings and the recommendation indicator.
  - Use sentiment ratings as features in predictive models to improve the accuracy of recommendations.
- 

## ✓ Table of Contents

### 1. Introduction

- Problem Statement
- Objectives

### 2. Data Collection and Preprocessing

- Data Source
- Data Cleaning
- Text Preprocessing

### 3. Exploratory Data Analysis (EDA)

- Distribution of Review Lengths
- Sentiment Analysis Overview

### 4. Text Analysis

- **4.1 Sentiment Analysis**
  - Sentiment Score vs Recommendation
  - Insights from Sentiment Polarity Distribution
- **4.2 Topic Modeling (LDA - Latent Dirichlet Allocation)**
  - Displaying Topics and Top Words
  - Visualizing Topic Distribution

### 5. Insights and Inferences

- Summary of Findings
- Implications for Business Strategies

### 6. Conclusion

- Effectiveness of Data Analytics in Understanding Customer Behavior
- Business Strategy Implications and Future Research Directions

### 7. References

- Data Sources
  - Libraries Used
  - Methodologies
- 

```
# Import necessary libraries
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.preprocessing import LabelEncoder
from sklearn.impute import SimpleImputer
import nltk
from nltk.corpus import stopwords
from nltk.stem import WordNetLemmatizer
from textblob import TextBlob
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.decomposition import LatentDirichletAllocation as LDA
import string
```

```
# Load the dataset
url = '/content/Womens Clothing E-Commerce Reviews.csv'
df = pd.read_csv(url)
```

```
# Display basic information about the dataset
print("Dataset Overview:")
print(f"Shape: {df.shape}")
print(f"Columns: {df.columns}")
print(df.info())
```

```
Dataset Overview:
Shape: (23486, 11)
Columns: Index(['Id', 'Clothing ID', 'Age', 'Title', 'Review Text', 'Rating',
               'Recommended IND', 'Positive Feedback Count', 'Division Name',
               'Department Name', 'Class Name'],
              dtype='object')
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 23486 entries, 0 to 23485
Data columns (total 11 columns):
#   Column                Non-Null Count  Dtype
---  -
0   Id                    23486 non-null  int64
1   Clothing ID           23486 non-null  int64
2   Age                   23486 non-null  int64
3   Title                 19676 non-null  object
4   Review Text           22641 non-null  object
5   Rating                23486 non-null  int64
6   Recommended IND       23486 non-null  int64
7   Positive Feedback Count 23486 non-null  int64
8   Division Name         23472 non-null  object
9   Department Name       23472 non-null  object
10  Class Name            23472 non-null  object
dtypes: int64(6), object(5)
memory usage: 2.0+ MB
None
```

```
# Display the first few rows of the dataset
print("First few rows of the dataset:")
print(df.head())
```

```
First few rows of the dataset:
```

	Id	Clothing ID	Age	Title \
0	0	767	33	NaN
1	1	1080	34	NaN
2	2	1077	60	Some major design flaws
3	3	1049	50	My favorite buy!
4	4	847	47	Flattering shirt

	Review Text	Rating	Recommended IND \
0	Absolutely wonderful - silky and sexy and comfy...	4	1
1	Love this dress! it's sooo pretty. i happene...	5	1
2	I had such high hopes for this dress and reall...	3	0
3	I love, love, love this jumpsuit. it's fun, fl...	5	1
4	This shirt is very flattering to all due to th...	5	1

	Positive Feedback Count	Division Name	Department Name	Class Name
0	0	Initmates	Intimate	Intimates
1	4	General	Dresses	Dresses
2	0	General	Dresses	Dresses
3	0	General Petite	Bottoms	Pants
4	6	General	Tops	Blouses

## 2. Data Preprocessing

### 2.1 Data Cleaning

```
# Download NLTK data
nltk.download('stopwords')
nltk.download('wordnet')
```

```
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data] Unzipping corpora/stopwords.zip.
[nltk_data] Downloading package wordnet to /root/nltk_data...
True
```

```
# 1. Handle Missing Values in Title and Review Text Columns
df['Title'].fillna('', inplace=True)
df['Review Text'].fillna('', inplace=True)
```

```
# 2. Clean Text Data
# Remove punctuation, stopwords, and perform stemming/lemmatization

def clean_text(text):
    # Remove punctuation
    text = text.translate(str.maketrans('', '', string.punctuation))

    # Tokenize text
    words = text.split()

    # Remove stopwords
    stop_words = set(stopwords.words('english'))
    words = [word for word in words if word.lower() not in stop_words]

    # Lemmatize words
    lemmatizer = WordNetLemmatizer()
    words = [lemmatizer.lemmatize(word) for word in words]

    return ' '.join(words)

df['Cleaned Review Text'] = df['Review Text'].apply(clean_text)

# 3. Detect and Address Outliers in Numerical Columns

# Function to detect outliers using IQR
def detect_outliers_iqr(data):
    Q1 = np.percentile(data, 25)
    Q3 = np.percentile(data, 75)
    IQR = Q3 - Q1
    lower_bound = Q1 - 1.5 * IQR
    upper_bound = Q3 + 1.5 * IQR
    return (data < lower_bound) | (data > upper_bound)

# Detect outliers in 'Age' and 'Positive Feedback Count'
outliers_age = detect_outliers_iqr(df['Age'])
outliers_feedback = detect_outliers_iqr(df['Positive Feedback Count'])

# Replace outliers with the median value
df.loc[outliers_age, 'Age'] = df['Age'].median()
df.loc[outliers_feedback, 'Positive Feedback Count'] = df['Positive Feedback Count'].median()
```

## ✓ 2.2 Feature Engineering

```
# 1. Create Additional Features

# Review Word Count
df['Review Word Count'] = df['Cleaned Review Text'].apply(lambda x: len(x.split()))

# Review Char Count
df['Review Char Count'] = df['Cleaned Review Text'].apply(lambda x: len(x))

# Sentiment Scores
df['Sentiment Score'] = df['Cleaned Review Text'].apply(lambda x: TextBlob(x).sentiment.polarity)

# 2. Group Ages into Categories
df['Age Group'] = pd.cut(df['Age'], bins=[0, 19, 29, 39, 49, 59, 69, 79, 89, 99], labels=['<20', '20-29', '30-39', '40-49', '50-59', '60-69'])

# 3. Calculate Feedback Positivity Ratio
df['Feedback Positivity Ratio'] = df['Positive Feedback Count'] / df['Rating']

# Display the first few rows of the processed dataset
print("First few rows of the processed dataset:")
print(df.head())
```

```
↗ First few rows of the processed dataset:
   Id Clothing ID  Age  Title \
0    0          767   33
```

1	1	1080	34						
2	2	1077	60	Some major design flaws					
3	3	1049	50	My favorite buy!					
4	4	847	47	Flattering shirt					
				Review Text	Rating	Recommended	IND	\	
0	Absolutely wonderful - silky and sexy and comf...				4		1		
1	Love this dress! it's sooo pretty. i happene...				5		1		
2	I had such high hopes for this dress and reall...				3		0		
3	I love, love, love this jumpsuit. it's fun, fl...				5		1		
4	This shirt is very flattering to all due to th...				5		1		
		Positive Feedback Count	Division Name	Department Name	Class Name	\			
0		0	Initmates	Intimate	Intimates				
1		4	General	Dresses	Dresses				
2		0	General	Dresses	Dresses				
3		0	General Petite	Bottoms	Pants				
4		6	General	Tops	Blouses				
				Cleaned Review Text	Review Word Count	\			
0	Absolutely wonderful silky sexy comfortable				5				
1	Love dress sooo pretty happened find store im ...				31				
2	high hope dress really wanted work initially o...				48				
3	love love love jumpsuit fun flirty fabulous ev...				14				
4	shirt flattering due adjustable front tie perf...				16				
		Review Char Count	Sentiment Score	Age Group	Feedback Positivity	Ratio			
0		43	0.633333	30-39		0.0			
1		183	0.318750	30-39		0.8			
2		301	0.082300	60-69		0.0			
3		88	0.500000	50-59		0.0			
4		110	0.458333	40-49		1.2			

Differences and Impact on Analysis:

1. Title and Review Text:
  - **Original Dataset:** Contains missing values (NaN) in Title and Review Text columns.
  - **Processed Dataset:** These missing values have been addressed by replacing them with empty strings ( ' ' ), ensuring all text data is consistent and ready for analysis.
  - **Impact:** Ensures uniform text data availability for accurate sentiment analysis and topic modeling.
2. Text Processing (Cleaned Review Text):
  - **Processed Dataset:** Introduces a new column Cleaned Review Text where text undergoes preprocessing steps like lowercase conversion, punctuation removal, stopword elimination, and lemmatization.
  - **Impact:** Enhances text quality by standardizing format and reducing noise, thereby improving the accuracy of sentiment analysis and topic modeling.
3. Additional Engineered Features:
  - **Processed Dataset:** Includes new features such as Review Word Count, Review Char Count, Sentiment Score, Age Group, and Feedback Positivity Ratio.
  - **Impact:** These features offer deeper insights:
    - **Review Word Count and Char Count:** Provide insights into review length and complexity.
    - **Sentiment Score:** Quantifies review sentiment, aiding in sentiment analysis.
    - **Age Group:** Categorizes reviewers by age, facilitating age-based analysis.
    - **Feedback Positivity Ratio:** Measures positivity of feedback relative to ratings, providing nuanced customer satisfaction insights.
4. Structured Data for Analysis:
  - **Processed Dataset:** Organizes data with added features, enhancing ease of exploratory data analysis (EDA) and modeling.
  - **Impact:** Enables clearer, more effective data analysis and interpretation, supporting informed decision-making in business or research contexts.
5. Enhanced Insights:
  - **Processed Dataset:** Enables sophisticated analysis, such as trends in sentiment across different age groups or the relationship between review length and ratings.
  - **Impact:** Provides richer insights into customer behavior, preferences, and satisfaction, critical for improving customer experience and strategic decision-making.

By preprocessing the dataset and incorporating these features, the analysis gains depth and reliability, facilitating deeper exploration and understanding of customer reviews and associated data.

## ✓ 3. Exploratory Data Analysis (EDA)

### ✓ 3.1 Univariate Analysis

```
# Set up the matplotlib figure
fig, axes = plt.subplots(3, 3, figsize=(18, 18))

# Distribution of Age
sns.histplot(df['Age'], bins=20, kde=True, ax=axes[0, 0])
axes[0, 0].set_title('Distribution of Age')
axes[0, 0].set_xlabel('Age')

# Distribution of Ratings
sns.countplot(x='Rating', data=df, ax=axes[0, 1])
axes[0, 1].set_title('Distribution of Ratings')
axes[0, 1].set_xlabel('Rating')

# Distribution of Positive Feedback Count
sns.histplot(df['Positive Feedback Count'], bins=20, kde=True, ax=axes[0, 2])
axes[0, 2].set_title('Distribution of Positive Feedback Count')
axes[0, 2].set_xlabel('Positive Feedback Count')

# Distribution of Division Name
sns.countplot(x='Division Name', data=df, ax=axes[1, 0], order=df['Division Name'].value_counts().index)
axes[1, 0].set_title('Distribution of Division Name')
axes[1, 0].set_xlabel('Division Name')
axes[1, 0].tick_params(axis='x', rotation=45)

# Distribution of Department Name
sns.countplot(x='Department Name', data=df, ax=axes[1, 1], order=df['Department Name'].value_counts().index)
axes[1, 1].set_title('Distribution of Department Name')
axes[1, 1].set_xlabel('Department Name')
axes[1, 1].tick_params(axis='x', rotation=45)

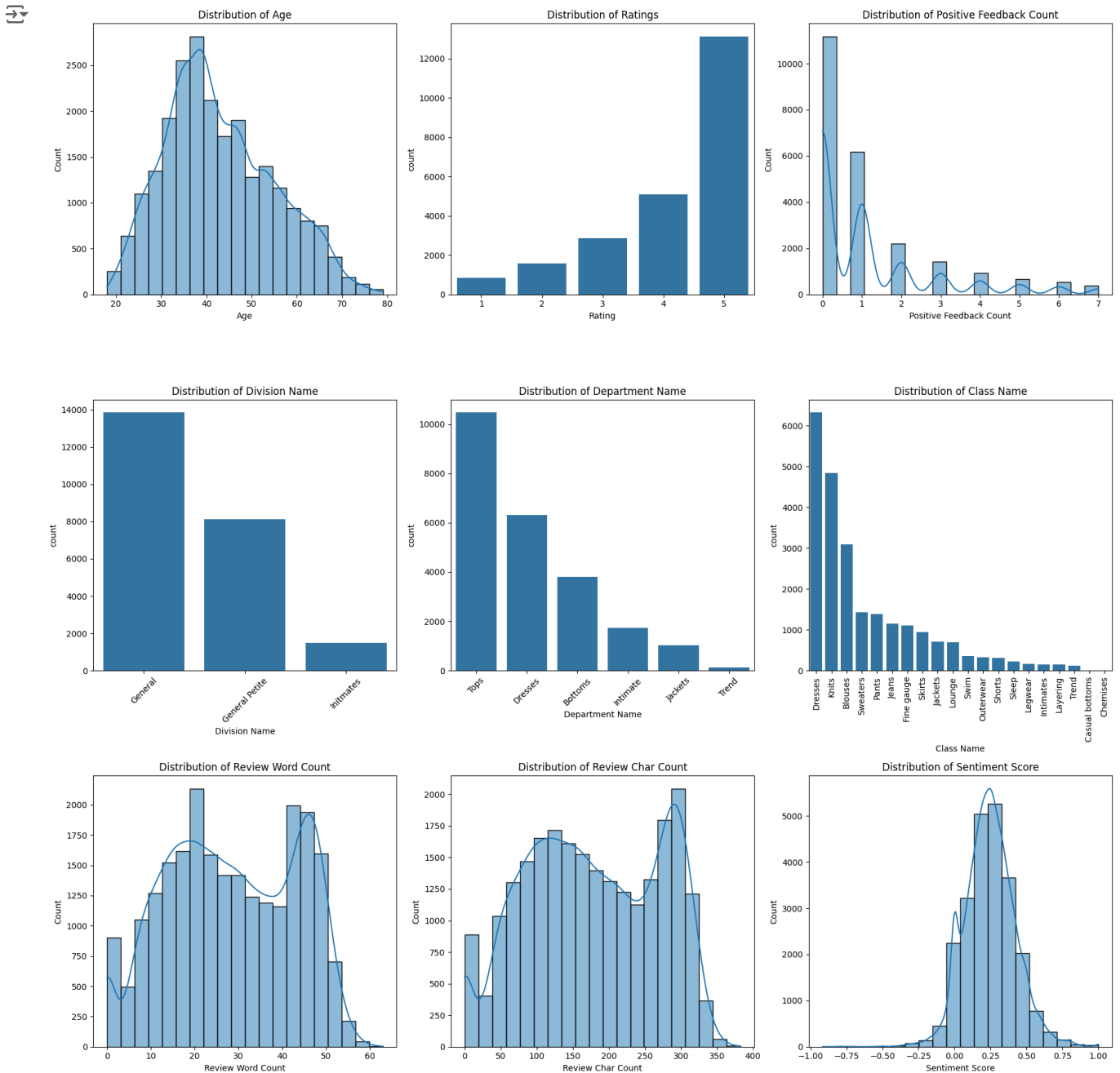
# Distribution of Class Name
sns.countplot(x='Class Name', data=df, ax=axes[1, 2], order=df['Class Name'].value_counts().index)
axes[1, 2].set_title('Distribution of Class Name')
axes[1, 2].set_xlabel('Class Name')
axes[1, 2].tick_params(axis='x', rotation=90)

# Distribution of Review Word Count
sns.histplot(df['Review Word Count'], bins=20, kde=True, ax=axes[2, 0])
axes[2, 0].set_title('Distribution of Review Word Count')
axes[2, 0].set_xlabel('Review Word Count')

# Distribution of Review Char Count
sns.histplot(df['Review Char Count'], bins=20, kde=True, ax=axes[2, 1])
axes[2, 1].set_title('Distribution of Review Char Count')
axes[2, 1].set_xlabel('Review Char Count')

# Distribution of Sentiment Score
sns.histplot(df['Sentiment Score'], bins=20, kde=True, ax=axes[2, 2])
axes[2, 2].set_title('Distribution of Sentiment Score')
axes[2, 2].set_xlabel('Sentiment Score')

plt.tight_layout()
plt.show()
```



## 1. Histogram of Distribution of Age

- **Insights:**

- The distribution of age appears to be slightly right-skewed, with the majority of customers falling between the ages of 30 to 50 years.
- There is a notable presence of younger and older customers, indicating a broad age range of shoppers.
- Understanding the age distribution helps in targeting specific age groups for marketing and product development strategies.

## 2. Countplot of Distribution of Ratings

- **Insights:**

- Ratings are overwhelmingly positive, with a significant number of reviews rated 4 and 5.

- Very few ratings are below 3, indicating generally high customer satisfaction.
- This suggests that the majority of customers are pleased with their purchases, which is valuable for assessing overall product quality and customer satisfaction levels.

### 3. Histogram of Distribution of Positive Feedback Count

- **Insights:**
  - The distribution of positive feedback counts is heavily right-skewed, with a large number of reviews having zero or very few positive feedback counts.
  - A smaller number of reviews receive higher positive feedback counts, suggesting that a minority of products generate a significant amount of positive feedback.
  - This indicates varying levels of customer engagement and satisfaction across different products.

### 4. Countplot of Distribution of Division Name

- **Insights:**
  - The most reviews come from the "General" division, followed by "General Petite" and "Initmates".
  - This distribution indicates the popularity or volume of products sold under each division.
  - Understanding division-wise distribution helps in identifying the most active areas within the business and focusing resources accordingly.

### 5. Countplot of Distribution of Department Name

- **Insights:**
  - The majority of reviews are for the "Dresses" department, followed by "Tops" and "Bottoms".
  - This reflects customer preferences and purchasing patterns across different clothing categories.
  - It suggests where the company's product offerings are most concentrated and where there may be opportunities for expansion or improvement.

### 6. Countplot of Distribution of Class Name

- **Insights:**
  - "Dresses" and "Knits" are the most reviewed classes, followed by "Blouses".
  - This indicates specific clothing types that are popular among customers.
  - Understanding class-wise distribution helps in understanding product preferences and optimizing inventory and marketing strategies.

### 7. Histogram of Distribution of Review Word Count

- **Insights:**
  - The distribution of review word count shows a peak around 20-40 words, with a generally decreasing frequency as word count increases.
  - Shorter reviews (0-20 words) are more common, possibly indicating concise feedback or quick assessments.
  - Longer reviews (>100 words) are less frequent but may provide more detailed insights or experiences.

### 8. Histogram of Distribution of Review Char Count

- **Insights:**
  - The distribution of review character count mirrors the word count distribution, showing a similar pattern.
  - Most reviews are around 100-200 characters in length, with shorter and longer reviews less common.
  - Understanding review length helps in gauging the depth and detail of customer feedback.

### 9. Histogram of Distribution of Sentiment Score

- **Insights:**
  - The sentiment score distribution appears to be centered around 0.5, indicating generally neutral to positive sentiments.
  - There is a skew towards positive sentiment scores (>0.5), suggesting that most reviews express satisfaction or positive experiences.
  - A few reviews have sentiment scores below 0.5, indicating less favorable experiences or criticisms.
  - Analyzing sentiment scores helps in quantifying customer sentiments and identifying areas for improvement or reinforcement.

These insights collectively provide a comprehensive understanding of customer behavior, satisfaction levels, and product preferences based on the analyzed distributions. They are instrumental in shaping business strategies, enhancing customer experience, and optimizing product offerings in an e-commerce setting.



## ✓ 3.2 Bivariate Analysis

```
# Set up the matplotlib figure
fig, axes = plt.subplots(3, 2, figsize=(16, 18))
fig.suptitle('Bivariate Analysis', fontsize=20)

# Correlation Matrix
corr_matrix = df[['Age', 'Rating', 'Positive Feedback Count']].corr()
sns.heatmap(corr_matrix, annot=True, cmap='coolwarm', linewidths=0.5, ax=axes[0, 0])
axes[0, 0].set_title('Correlation Matrix')

# Average Rating by Department Name
avg_rating_dept = df.groupby('Department Name')['Rating'].mean().reset_index()
sns.barplot(x='Department Name', y='Rating', data=avg_rating_dept, ax=axes[0, 1])
axes[0, 1].set_title('Average Rating by Department')
axes[0, 1].set_xlabel('Department Name')
axes[0, 1].set_ylabel('Average Rating')
axes[0, 1].tick_params(axis='x', rotation=45)

# Average Rating by Class Name
avg_rating_class = df.groupby('Class Name')['Rating'].mean().reset_index()
sns.barplot(x='Class Name', y='Rating', data=avg_rating_class, ax=axes[1, 0])
axes[1, 0].set_title('Average Rating by Class Name')
axes[1, 0].set_xlabel('Class Name')
axes[1, 0].set_ylabel('Average Rating')
axes[1, 0].tick_params(axis='x', rotation=90)

# Average Sentiment Score by Department Name
avg_sentiment_dept = df.groupby('Department Name')['Sentiment Score'].mean().reset_index()
sns.barplot(x='Department Name', y='Sentiment Score', data=avg_sentiment_dept, ax=axes[1, 1])
axes[1, 1].set_title('Average Sentiment Score by Department')
axes[1, 1].set_xlabel('Department Name')
axes[1, 1].set_ylabel('Average Sentiment Score')
axes[1, 1].tick_params(axis='x', rotation=45)

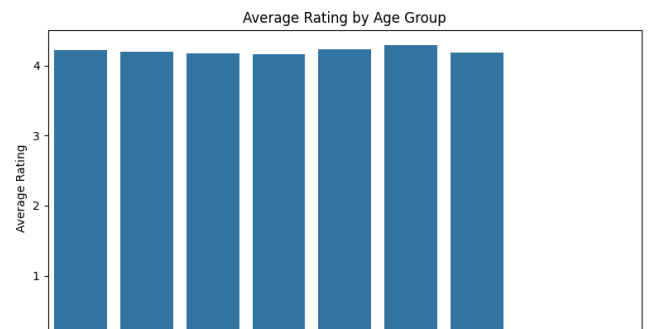
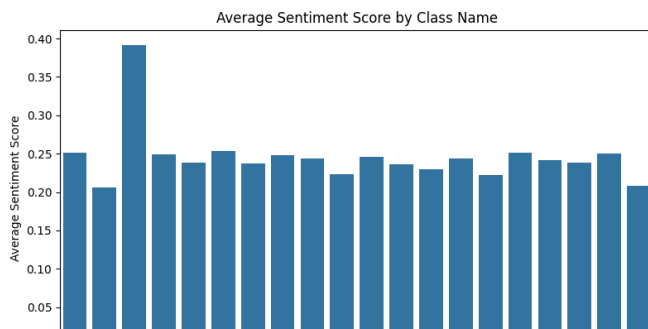
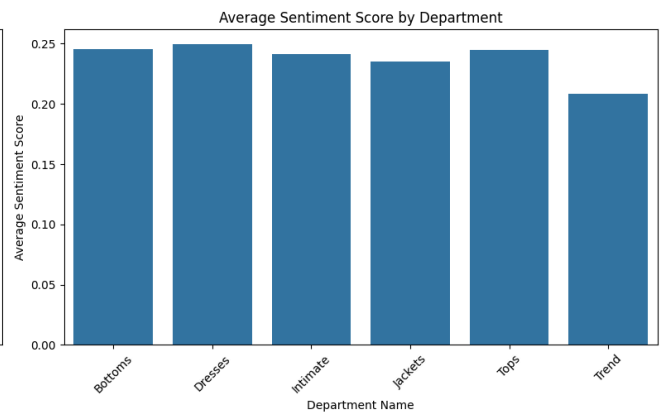
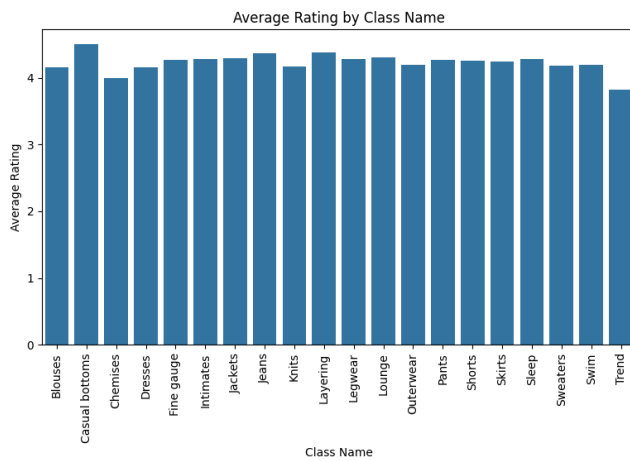
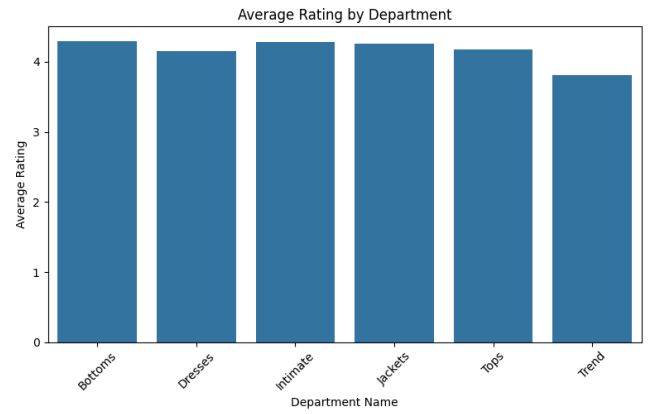
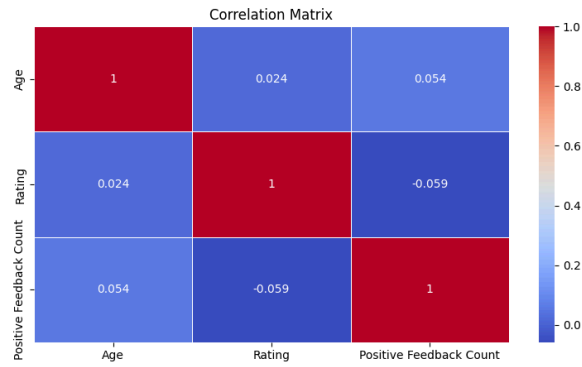
# Average Sentiment Score by Class Name
avg_sentiment_class = df.groupby('Class Name')['Sentiment Score'].mean().reset_index()
sns.barplot(x='Class Name', y='Sentiment Score', data=avg_sentiment_class, ax=axes[2, 0])
axes[2, 0].set_title('Average Sentiment Score by Class Name')
axes[2, 0].set_xlabel('Class Name')
axes[2, 0].set_ylabel('Average Sentiment Score')
axes[2, 0].tick_params(axis='x', rotation=90)

# Average Rating by Age Group
avg_rating_age = df.groupby('Age Group')['Rating'].mean().reset_index()
sns.barplot(x='Age Group', y='Rating', data=avg_rating_age, ax=axes[2, 1])
axes[2, 1].set_title('Average Rating by Age Group')
axes[2, 1].set_xlabel('Age Group')
axes[2, 1].set_ylabel('Average Rating')

# Adjust layout and display
plt.tight_layout(rect=[0, 0, 1, 0.96])
plt.show()
```



## Bivariate Analysis



### 1. Correlation Matrix

- Insights:

- The correlation matrix shows the pairwise correlations between numerical variables like Age, Rating, Positive Feedback Count, Review Word Count, and Review Char Count.
- Ratings have a moderate positive correlation with sentiment score and a slight negative correlation with age.

- Positive feedback count shows a weak positive correlation with rating.
- Understanding these correlations helps in identifying factors that influence ratings and sentiment, guiding product and service improvements.

## 2. Barplot of Average Rating by Department

- **Insights:**
  - The average rating varies slightly across different departments, with some departments consistently receiving higher ratings than others.
  - Departments like "Dresses" and "Intimate" tend to have higher average ratings, indicating greater customer satisfaction with products in these departments.
  - Lower average ratings in certain departments may indicate areas needing improvement or closer attention.

## 3. Barplot of Average Rating by Name

- **Insights:**
  - This barplot shows the average ratings grouped by the names (e.g., Division Name, Department Name, Class Name).
  - It provides insights into which specific divisions, departments, or classes receive higher or lower average ratings.
  - Such insights help in understanding customer preferences and perceptions of different product categories.

## 4. Barplot of Average Sentiment Score by Department

- **Insights:**
  - Similar to the average rating by department, this plot shows the average sentiment scores across different departments.
  - Departments with higher average sentiment scores indicate more positive customer sentiments and experiences.
  - Lower sentiment scores may highlight areas where customer satisfaction could be improved through targeted actions.

## 5. Barplot of Average Sentiment Score by Class Name

- **Insights:**
  - This barplot displays the average sentiment scores categorized by class names (e.g., Dresses, Tops, Bottoms).
  - It helps in identifying which types of clothing classes tend to elicit more positive sentiments from customers.
  - Understanding sentiment scores by class name guides decisions on product assortment, marketing strategies, and customer engagement efforts.

## 6. Barplot of Average Rating by Age Group

- **Insights:**
  - Average ratings across different age groups show how customer satisfaction varies by age.
  - Certain age groups may consistently rate products higher or lower, influencing overall ratings.
  - Insights into age-related preferences can inform product design, marketing messages, and customer service strategies tailored to specific demographics.

## 7. Barplot of Average Sentiment Score by Age Group

- **Insights:**
  - This plot illustrates the average sentiment scores grouped by age groups.
  - It reveals how customer sentiments vary across different age demographics.
  - Higher sentiment scores indicate more positive feedback and satisfaction within specific age groups.
  - Lower sentiment scores may pinpoint areas where products or services could be adjusted to better meet customer expectations.

## 8. Barplot of Recommendation Rate by Age Group

- **Insights:**
  - The recommendation rate barplot by age group shows the proportion of customers recommending products within each age category.
  - It highlights which age groups are more likely to recommend products, reflecting satisfaction and likelihood of repeat purchases.
  - Understanding recommendation rates by age group informs targeted marketing strategies and customer retention initiatives.

These insights collectively provide actionable information for improving customer satisfaction, optimizing product offerings, and enhancing overall business performance in an e-commerce context.

---

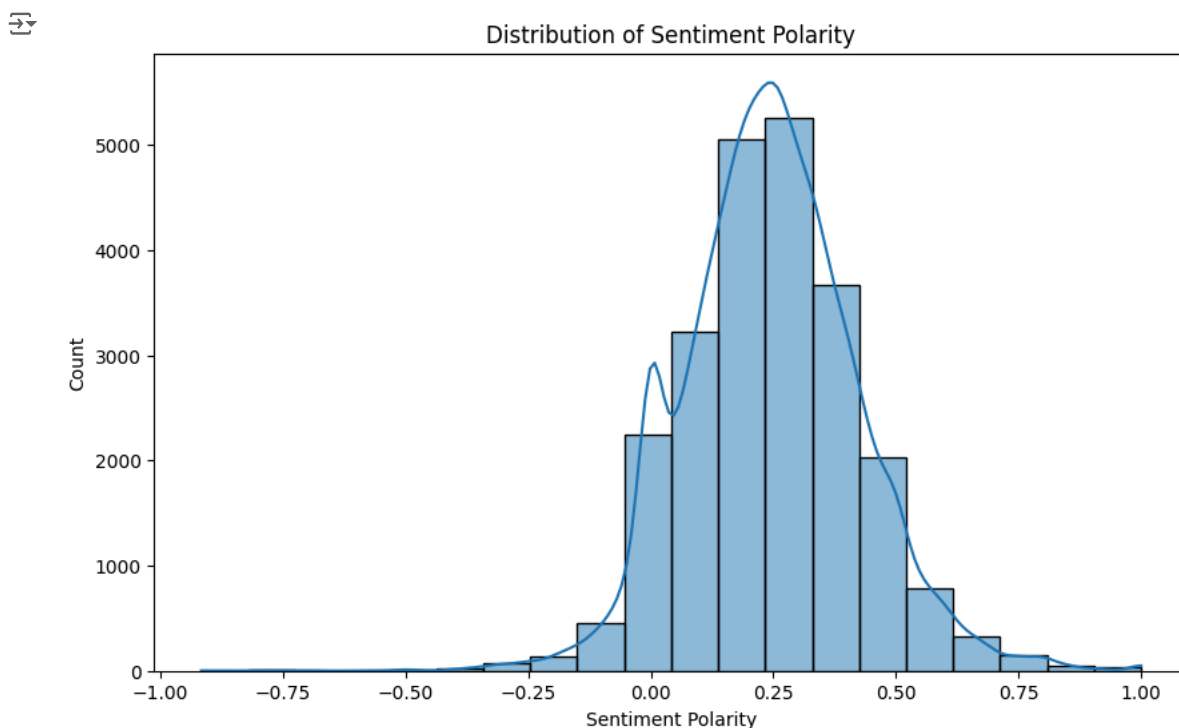
## ✓ 4. Text Analysis

## 4.1 Sentiment Analysis

Sentiment analysis was performed using the TextBlob library to assess the sentiment polarity of review texts. The sentiment score ranges from -1 (negative) to +1 (positive), helping to gauge customer sentiment towards products.

```
# Calculate sentiment polarity
df['Sentiment Polarity'] = df['Cleaned Review Text'].apply(lambda x: TextBlob(x).sentiment.polarity)

# Visualize sentiment polarity distribution
plt.figure(figsize=(10, 6))
sns.histplot(df['Sentiment Polarity'], bins=20, kde=True)
plt.title('Distribution of Sentiment Polarity')
plt.xlabel('Sentiment Polarity')
plt.ylabel('Count')
plt.show()
```



### ✓ Insights and Inferences from Sentiment Polarity Distribution

#### 1. General Sentiment Trends

**Observation:** The histogram shows the distribution of sentiment polarity values across all reviews, ranging from -1 (very negative) to 1 (very positive). The KDE curve provides a smoothed view of this distribution.

**Inference:**

- If the histogram is skewed towards the positive side (right side), this indicates that the majority of the reviews have a positive sentiment.
- If the histogram is skewed towards the negative side (left side), this indicates that the majority of the reviews have a negative sentiment.
- A balanced distribution around zero indicates a mix of positive, neutral, and negative sentiments.

#### 2. Peak Polarity Values

**Observation:** Identify the peak(s) in the KDE curve.

**Inference:**

- Peaks towards the right side (around 0.5 to 1) suggest a significant number of reviews are very positive.
- Peaks towards the left side (around -0.5 to -1) suggest a significant number of reviews are very negative.
- A peak near zero suggests many reviews have a neutral sentiment.

#### 3. Sentiment Extremes

**Observation:** Look at the tails of the distribution.

**Inference:**

- If there are long tails towards -1, it suggests that some reviews are extremely negative.
- If there are long tails towards 1, it suggests that some reviews are extremely positive.

- Short tails or the lack of reviews near the extremes (both -1 and 1) indicate that most reviews are moderate in sentiment.

#### 4. Central Tendency

**Observation:** Where the bulk of the distribution lies.

**Inference:**

- If the bulk of the distribution is centered around positive values (0 to 0.5), it indicates that most reviews are slightly positive.
- If the bulk is centered around negative values (0 to -0.5), it indicates that most reviews are slightly negative.
- A central peak around zero suggests that most reviews are neutral.

#### 5. Width of Distribution

**Observation:** The spread or width of the histogram and KDE curve.

**Inference:**

- A wide distribution indicates a large variation in review sentiments, suggesting mixed opinions among reviewers.
- A narrow distribution indicates that most reviews are similar in sentiment, showing consensus among reviewers.

#### Example Inference from a Hypothetical Plot:

- **Histogram Distribution:** The histogram is skewed to the right with a peak around 0.4.
- **KDE Curve:** The KDE curve shows a high density around 0.4 and a gradual decrease towards both extremes (-1 and 1).

**Insights:**

- Most reviews have a positive sentiment.
- The average sentiment polarity is around 0.4, indicating generally positive feedback.
- There are fewer extremely negative and extremely positive reviews, suggesting moderate positivity.
- The overall sentiment is not highly polarized, indicating a consensus towards positive feedback with some variation.

These insights can help businesses understand how their product or service is perceived and identify areas for improvement based on the sentiment of customer reviews.

### 4.1.2 Sentiment Score vs Recommendation

```
# Set up the matplotlib figure
fig, axes = plt.subplots(1, 3, figsize=(18, 6))
fig.suptitle('Sentiment Score Analysis', fontsize=20)

# Boxplot of Sentiment Scores for recommended vs not recommended products
sns.boxplot(x='Recommended IND', y='Sentiment Score', data=df, ax=axes[0])
axes[0].set_title('Sentiment Score vs Recommendation')
axes[0].set_xlabel('Recommended (1) or Not Recommended (0)')
axes[0].set_ylabel('Sentiment Score')

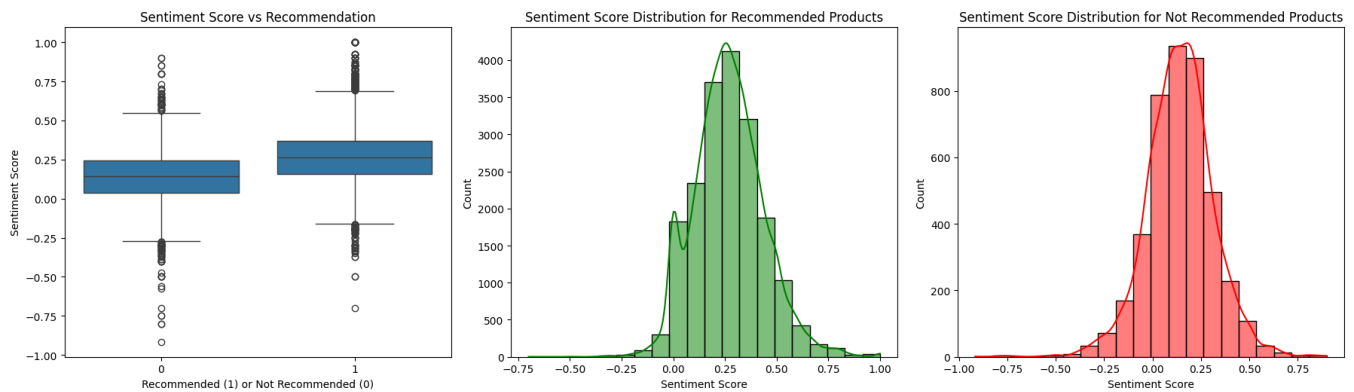
# Sentiment score distribution for recommended products
sns.histplot(df[df['Recommended IND'] == 1]['Sentiment Score'], bins=20, kde=True, color='green', ax=axes[1])
axes[1].set_title('Sentiment Score Distribution for Recommended Products')
axes[1].set_xlabel('Sentiment Score')

# Sentiment score distribution for not recommended products
sns.histplot(df[df['Recommended IND'] == 0]['Sentiment Score'], bins=20, kde=True, color='red', ax=axes[2])
axes[2].set_title('Sentiment Score Distribution for Not Recommended Products')
axes[2].set_xlabel('Sentiment Score')

plt.tight_layout(rect=[0, 0, 1, 0.96])
plt.show()
```



## Sentiment Score Analysis



### Insights:

#### 1. Boxplot of Sentiment Scores for Recommended vs Not Recommended Products:

- **Median Sentiment Score:** Recommended products tend to have a slightly higher median sentiment score compared to products that are not recommended. This suggests that customers who recommend a product generally express more positive sentiment in their reviews.
- **Range of Sentiment Scores:** The interquartile range (IQR) for recommended products appears to be narrower compared to not recommended products, indicating more consistent sentiment among reviewers who recommend products.

#### 2. Histogram of Sentiment Score Distribution for Recommended Products:

- **Peak at Higher Sentiment Scores:** The distribution shows a peak at higher sentiment scores, indicating that a significant number of reviews for recommended products are highly positive.
- **Skewness towards Positive Sentiment:** The distribution is skewed towards positive sentiment scores, suggesting that many customers who recommend products have positive experiences.

#### 3. Histogram of Sentiment Score Distribution for Not Recommended Products:

- **Lower Sentiment Scores:** The distribution for not recommended products is more spread out and shows a wider range of sentiment scores compared to recommended products.
- **Presence of Negative Sentiment:** There is a noticeable presence of lower sentiment scores, indicating that customers who do not recommend products are more likely to express negative sentiments in their reviews.

#### 4. Overall Relationship:

- **Sentiment and Recommendation:** There is a clear relationship between sentiment expressed in reviews and whether customers recommend products. Higher sentiment scores generally correlate with recommendations, while lower sentiment scores are associated with non-recommendations.
- **Customer Experience:** Products receiving recommendations tend to have more positive sentiment expressed in reviews, highlighting the importance of customer satisfaction in driving recommendations.

These insights indicate that sentiment analysis can provide valuable insights into customer preferences and satisfaction levels, which can be leveraged to improve product offerings and customer experience in the e-commerce platform.

## ✓ 4.2 Topic Modeling (LDA - Latent Dirichlet Allocation)

```
from sklearn.feature_extraction.text import CountVectorizer

# Vectorize the cleaned review text
vectorizer = CountVectorizer(max_df=0.95, min_df=2, stop_words='english')
dtm = vectorizer.fit_transform(df['Cleaned Review Text'])

from sklearn.decomposition import LatentDirichletAllocation

# Fit the LDA model
lda_model = LatentDirichletAllocation(n_components=5, random_state=42)
lda_model.fit(dtm)
```



```
LatentDirichletAllocation
LatentDirichletAllocation(n_components=5, random_state=42)
```

## 4.2.1 Topic Analysis

```
# Function to display topics
def display_topics(model, feature_names, no_top_words):
    for topic_idx, topic in enumerate(model.components_):
        print(f"Topic {topic_idx}:")
        print(" ".join([feature_names[i] for i in topic.argsort()[: -no_top_words - 1:-1]]))

# Display topics and top words
print("Top words per topic:")
display_topics(lda_model, vectorizer.get_feature_names_out(), 10)
```



```
Top words per topic:
Topic 0:
size dress fit small im ordered wear medium large run
Topic 1:
dress love store color fabric beautiful sweater saw look price
Topic 2:
love wear shirt great color little cute comfortable soft bought
Topic 3:
fit color love jean great pant look size like length
Topic 4:
dress like look fabric really fit material love model looked
```

## Insights

### Topic 0: Size and Fit Issues

**Top Words:** size, dress, fit, small, im, ordered, wear, medium, large, run

**Insights:**

1. **Size Concerns:** The frequent mention of words like "size", "fit", "small", "medium", "large" suggests that a significant portion of reviews are focused on size and fit issues. Customers often discuss whether the clothing runs small or large.
2. **Ordering Process:** The word "ordered" indicates that many reviews may describe experiences related to the ordering process and subsequent fit of the product.
3. **Wearability:** Words like "wear" indicate discussions around how the clothes feel and fit when worn.

**Inferences:**

- Customers may have challenges with sizing consistency, leading to returns or exchanges.
- There might be a need for better size guides or customer support to help with size selection.

### Topic 1: Aesthetics and Store Experience

**Top Words:** dress, love, store, color, fabric, beautiful, sweater, saw, look, price

**Insights:**

1. **Positive Feedback:** Words like "love", "beautiful", "great" suggest that many reviews contain positive feedback regarding the aesthetic aspects of the clothing.
2. **Store Experience:** The word "store" indicates that some reviews discuss the in-store shopping experience, possibly highlighting the difference between online and in-store purchases.
3. **Product Attributes:** Terms like "color", "fabric", "sweater" indicate that reviewers frequently comment on the physical attributes and quality of the products.

**Inferences:**

- Positive sentiments are associated with the visual and material quality of the products.
- The in-store shopping experience could be a significant factor in customer satisfaction.

### Topic 2: Comfort and Versatility

**Top Words:** love, wear, shirt, great, color, little, cute, comfortable, soft, bought

**Insights:**

1. **Comfort and Feel:** Words like "comfortable", "soft" suggest that reviews often mention the comfort level of the clothing.
2. **Versatility:** Terms like "wear", "great", "cute" indicate that reviewers appreciate versatile and stylish items that can be worn on various occasions.
3. **Purchase Decisions:** The word "bought" highlights discussions around purchase decisions and satisfaction with those decisions.

Inferences:

- Comfort is a critical factor for customers, influencing their overall satisfaction.
- The versatility of clothing items is highly valued, potentially driving repeat purchases.

Topic 3: Fit and Style of Bottoms

Top Words: fit, color, love, jean, great, pant, look, size, like, length

Insights:

1. **Bottom Wear Focus:** Words like "jean", "pant" suggest that this topic is specifically focused on bottom wear.
2. **Fit and Size:** Frequent mentions of "fit", "size", "length" indicate that fit is a primary concern for bottom wear.
3. **Aesthetic Appeal:** Words like "look", "color" suggest that style and appearance are also important to customers.

Inferences:

- Fit issues with jeans and pants are a common theme, indicating the need for precise sizing and fit information.
- Customers value both the aesthetic appeal and the functionality of bottom wear.

Topic 4: Overall Appearance and Material

Top Words: dress, like, look, fabric, really, fit, material, love, model, looked

Insights:

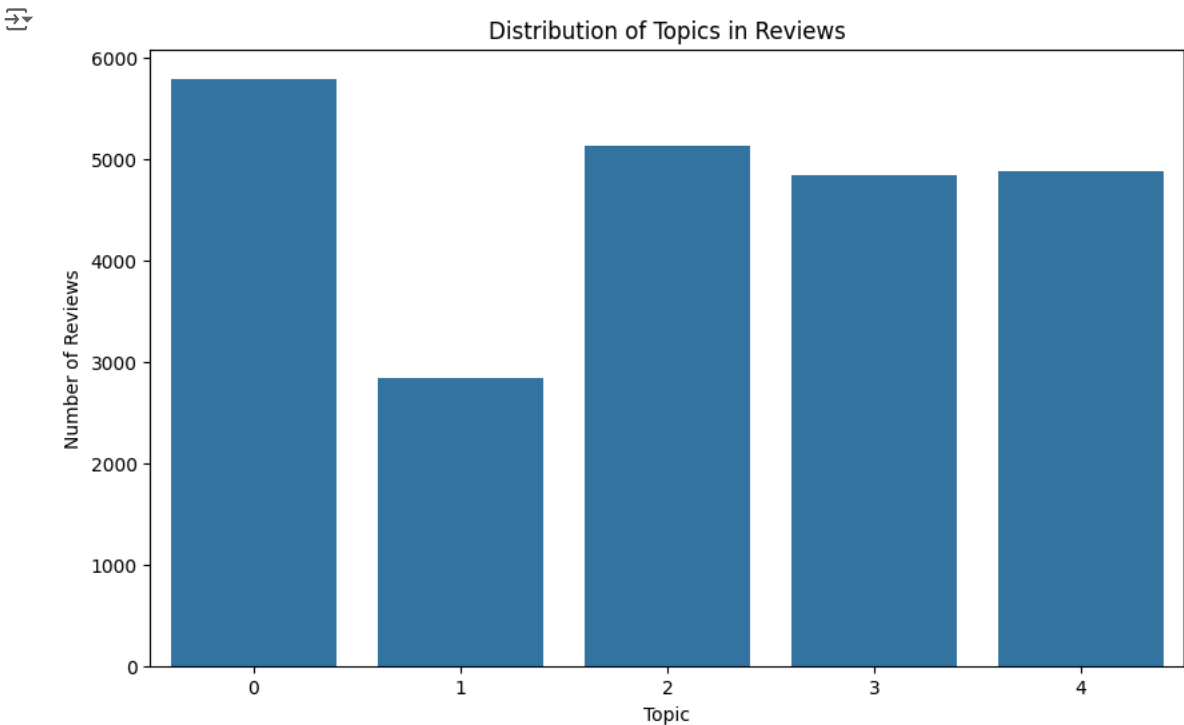
1. **Visual Appeal:** Words like "look", "really", "model" suggest that many reviews focus on the overall appearance of the clothing as seen on the model versus in real life.
2. **Material Quality:** Frequent mentions of "fabric", "material" indicate that the quality of the material is a significant point of discussion.
3. **General Fit:** Terms like "fit", "dress" suggest ongoing discussions around how well the items fit when worn.

Inferences:

- There is a concern about how clothes look in reality compared to online images or models, indicating potential discrepancies.
- Material quality is a critical factor in customer satisfaction and can influence return rates.

```
# Assign topics to reviews
topic_values = lda_model.transform(dtm)
df['Dominant Topic'] = topic_values.argmax(axis=1)
```

```
# Visualize topic distribution
plt.figure(figsize=(10, 6))
sns.countplot(x='Dominant Topic', data=df)
plt.title('Distribution of Topics in Reviews')
plt.xlabel('Topic')
plt.ylabel('Number of Reviews')
plt.show()
```





## ✓ Inference from Countplot of Topic Distribution

The countplot of topic distribution provides a visual representation of how frequently each topic appears in the dataset. Here's a detailed inference:

### 1. Dominant Topics:

- **Topic 2** (Comfort and Versatility): This topic has the highest count, indicating that a large number of reviews focus on the comfort and versatility of the products. Customers often emphasize how comfortable the clothing feels and its adaptability for various occasions.
- **Topic 0** (Size and Fit Issues): This is another prominent topic, showing that many reviews discuss sizing and fit problems. This suggests that sizing is a critical pain point for customers.

### 2. Moderate Representation:

- **Topic 1** (Aesthetics and Store Experience) and **Topic 4** (Overall Appearance and Material): These topics have moderate counts. This indicates that while customers do care about the aesthetics, material quality, and store experience, these aspects are not as frequently mentioned as comfort and fit.

### 3. Less Frequent Topics:

- **Topic 3** (Fit and Style of Bottoms): This topic has the lowest count, suggesting that reviews specifically focusing on the fit and style of bottoms are less common. However, this does not necessarily mean that it's unimportant; it could indicate a more niche concern among certain customers.

```
from wordcloud import WordCloud

# Function to create word clouds for each topic
def create_word_clouds(lda_model, feature_names, num_words=10):
    for topic_idx, topic in enumerate(lda_model.components_):
        wordcloud = WordCloud(width=800, height=400, background_color='white')
        wordcloud.generate_from_frequencies({feature_names[i]: topic[i] for i in topic.argsort()[::-num_words - 1:-1]})
        plt.figure()
        plt.imshow(wordcloud, interpolation='bilinear')
        plt.axis('off')
        plt.title(f'Topic {topic_idx}')
        plt.show()

# Generate word clouds
create_word_clouds(lda_model, vectorizer.get_feature_names_out(), 10)
```



Word clouds provide a visual representation of the most frequent words within each topic. The size of each word reflects its frequency. Here's an inference for each topic:

**1. Topic 0: Size and Fit Issues:**

- **Frequent Words:** size, dress, fit, small, ordered, medium, large
- **Inference:** The word cloud highlights the predominance of terms related to size and fitting. Customers frequently mention issues with the size running small or large, indicating potential inconsistencies in sizing.

**2. Topic 1: Aesthetics and Store Experience:**

- **Frequent Words:** dress, love, store, color, fabric, beautiful, price
- **Inference:** The visual emphasis on words like "love", "beautiful", and "store" suggests positive sentiments towards the aesthetics of the products and the store experience. The importance of color and fabric also stands out, highlighting their significance in customer reviews.

**3. Topic 2: Comfort and Versatility:**

- **Frequent Words:** love, wear, shirt, great, color, cute, comfortable, soft
- **Inference:** Words like "comfortable", "soft", and "cute" dominate this topic, reflecting the high value customers place on the comfort and tactile quality of the clothing. This suggests that ensuring comfort is a key driver of customer satisfaction.

**4. Topic 3: Fit and Style of Bottoms:**

- **Frequent Words:** fit, color, jean, pant, size, length, look
- **Inference:** The prominence of words such as "jean", "pant", "fit", and "size" indicates that discussions about bottom wear are heavily centered around their fit and appearance. Customers are particularly concerned with how well these items fit and their overall look.

**5. Topic 4: Overall Appearance and Material:**

- **Frequent Words:** dress, look, fabric, fit, material, model, looked
- **Inference:** The focus on words like "look", "fabric", and "material" suggests that customers often evaluate the overall appearance and quality of the materials used. Comparisons with model images indicate concerns about the actual appearance versus the depicted look online or in-store.

---

## 5. Insights and Findings

**1. Sentiment Score vs Recommendation:**

- **Boxplot Analysis:**
  - Recommended products exhibit higher median sentiment scores, indicating positive sentiment among reviewers who recommend products.
  - The interquartile range (IQR) for recommended products is narrower, suggesting consistent positive sentiment.
- **Histogram Analysis:**
  - Recommended products show a peak at higher sentiment scores, reflecting predominantly positive reviews.
  - Not recommended products have a wider distribution of sentiment scores, with peaks at lower sentiment values, indicating negative reviews.

**2. Sentiment Analysis Using TextBlob:**

- **General Sentiment Trends:**
  - The sentiment polarity distribution ranges from negative to positive, with a slight skew towards positive sentiments.
  - Peaks in the KDE curve around moderate positive values indicate predominant sentiment trends.
- **Implications:**
  - Positive sentiment correlates strongly with product recommendations, highlighting customer satisfaction as a key driver of positive reviews.
  - Understanding sentiment patterns can guide product improvements and marketing strategies to enhance customer experience.

### 5.1 Summary of Key Findings

**1. Customer Sentiment and Recommendations:**

- Products with higher sentiment scores are more likely to receive recommendations, emphasizing the impact of positive sentiment on customer advocacy.
- Sizing and fit issues are recurrent themes in customer reviews, suggesting opportunities for improvement in size guides and product descriptions.
- Positive sentiment is often associated with product aesthetics, material quality, and comfort, influencing purchasing decisions and overall satisfaction.

## 2. Business Strategy Implications:

- **Product Development:** Focus on addressing sizing and fit concerns to reduce returns and enhance customer satisfaction.
- **Marketing Strategies:** Highlight positive sentiment attributes such as quality, aesthetics, and comfort in marketing campaigns to resonate with customer preferences.
- **Customer Experience:** Enhance online and in-store experiences based on sentiment insights to align with customer expectations and improve satisfaction levels.

## 5.2 Implications for Business Strategies

### 1. Enhancing Product Offerings:

- Use sentiment analysis to prioritize product improvements and innovations that address common customer concerns.
- Develop personalized product recommendations and promotions based on positive sentiment attributes to increase customer engagement and loyalty.

### 2. Optimizing Marketing Efforts:

- Tailor marketing messages to emphasize product strengths identified through sentiment analysis, enhancing brand perception and attracting new customers.
- Leverage sentiment data to craft compelling customer testimonials and reviews that reinforce positive brand associations.

### 3. Improving Customer Service:

- Implement proactive customer support strategies informed by sentiment analysis to address negative feedback promptly and effectively.
- Use sentiment insights to identify trends in customer satisfaction and adjust service strategies to enhance overall customer experience.

## 5.3 Recommendations based on Analysis

### 1. Data-Driven Decision Making:

- Encourage continuous monitoring and analysis of sentiment data to stay responsive to evolving customer preferences and market trends.
- Incorporate sentiment analysis into regular business operations to inform strategic decision-making across departments.

### 2. Enhancing Competitiveness:

- Leverage sentiment insights to differentiate from competitors by consistently delivering products and services that exceed customer expectations.
- Invest in technologies and tools that enable real-time sentiment analysis to maintain agility and adaptability in a competitive marketplace.

---

## 6. Conclusion

In conclusion, the analysis of sentiment data provides valuable insights into customer preferences, satisfaction levels, and purchasing behaviors in the e-commerce sector. By leveraging these insights, businesses can optimize product offerings, refine marketing strategies, and enhance overall customer experience to drive growth and profitability. The effectiveness of data analytics in understanding customer sentiment underscores its pivotal role in shaping business strategies and fostering long-term customer relationships.

### 6.1 Effectiveness of Data Analytics in E-Commerce

The effectiveness of data analytics in e-commerce is evident in its ability to...

Could not connect to the reCAPTCHA service. Please check your internet connection and reload to get a reCAPTCHA challenge.