Understanding Customer Behavior and Product Recommendations

Akhil Arekatika

Prof. Vaishali Goyal
AIT - 622 - DL1
George Mason University

Abstract:
This project focuses on analyzing a dataset sourced from Kaggle containing clothing reviews to extract valuable insights for business decision-making within the organization. The dataset encompasses various attributes such as Clothing ID, Age, Title, Review Text, Rating, and more, providing a rich source of information for analysis. Through this project, we aim to understand customer preferences, identify popular products, detect trends, and improve product offerings. The analysis will involve exploring the dataset to uncover patterns and correlations, performing statistical analyses to derive meaningful insights, and creating visualizations to effectively communicate findings. By leveraging this dataset and conducting thorough analyses, the organization seeks to enhance its understanding of customer behavior, optimize product strategies, and ultimately drive business growth in the competitive retail market.

Introduction:
In today's dynamic retail environment, the importance of data-driven decision-making cannot be overstated. With the exponential growth of online commerce and the abundance of customer feedback available through platforms like Kaggle, organizations have unprecedented access to valuable insights into consumer behavior and preferences. This project focuses on harnessing the power of a clothing review dataset sourced from Kaggle to glean actionable insights for strategic decision-making within the organization. By analyzing attributes such as Clothing ID, Age, Title, Review Text, and Rating, the goal is to uncover hidden patterns, identify trends, and refine product strategies to meet evolving customer needs. Through systematic analysis and advanced analytics techniques, the organization aims to enhance its competitive edge in the retail market and deliver superior value to customers. This research paper outlines a structured approach to extract, analyze, and interpret data, providing a roadmap for leveraging data analytics to drive business growth and innovation in the retail industry.

Description of the Problem:
In today's retail landscape, the abundance of data, particularly in the form of customer reviews, presents both opportunities and challenges for businesses. The problem at hand lies in effectively harnessing this wealth of data to extract meaningful insights that can drive strategic decision-making. Specifically, within the context of this research, the challenge revolves around analyzing a dataset of clothing reviews to uncover actionable insights for the retail sector.

The dataset contains a diverse array of variables, ranging from demographic information such as age to qualitative feedback in the form of review text and numerical ratings. However, the unstructured nature of review text, coupled with potential noise and missing values, presents significant hurdles in deriving accurate and reliable insights. Addressing this challenge requires the application of advanced data preprocessing techniques, sophisticated natural language processing (NLP) algorithms, and robust statistical analyses to extract valuable information from the dataset. Moreover, the dynamic nature of the retail industry necessitates not only understanding current consumer preferences but also predicting future trends to stay ahead of the competition. Thus, the problem extends beyond mere data analysis to encompass predictive modeling and strategic foresight to drive business success.

Project Objectives and Goals:
The goal of this Big data project is to analyze the 23,000 user reviews using Big Data analytic tools to gather user sentiment regarding different products and answer different questions. The Big Data project has several objectives:

- **Age Groups and Preferences:** Investigating how different age groups tend to prefer certain class names and department names, shedding light on consumer behavior patterns across demographic segments.
- **Sentiment Analysis and Product Recommendations:** Assessing the predictive power of sentiment analysis and review text length in determining the likelihood of product recommendations, and identifying key phrases associated with higher recommendation rates.
- **Age and Likelihood to Recommend:** Exploring the relationship between age and the likelihood of recommending products, providing insights into age-specific consumer tendencies.
- **Rating and Likelihood to Recommend:** Analyzing the correlation between product ratings and the likelihood of recommending products, offering insights into the influence of ratings on consumer recommendations.

Current Data Environment within the Organization:
The organization presently relies on external datasets, including those from platforms like Kaggle, to supplement its internal data. While this reflects an openness to integrating external data sources, the absence of centralized infrastructure and standardized processes for data management and analytics is notable. Data collection involves accessing and downloading datasets on an as-needed basis, potentially leading to inefficiencies and the creation of data silos.

Regarding data storage, individual analysts likely utilize their machines or local servers, resulting in decentralized data storage and potential challenges in data accessibility and sharing. Although basic data analysis capabilities exist using tools like Python or R, there is a clear need for investment in advanced analytics techniques and expertise, such as natural language processing (NLP) and predictive modeling, to elevate the organization's data maturity and readiness for comprehensive analytics.

Overall, the current data environment underscores the necessity for investment in data infrastructure, tools, and expertise to facilitate more holistic data management and advanced analytics capabilities. Initiatives like establishing a centralized data repository, implementing standardized data collection processes, and offering training and resources for advanced analytics techniques are crucial steps to enhance the organization's data maturity and readiness for advanced analytics.

Required Resources:
To ensure the successful execution of this project, a range of resources is indispensable to streamline tasks encompassing data cleaning, coding, statistical analysis, and visualization. Spyder IDE emerges as the pivotal tool for these endeavors, distinguished by its tailored integrated development environment optimized for scientific computing and Python-based data analysis.

1. **Spyder IDE:** Spyder provides a comprehensive environment for data analysis and scientific computing in Python. Its features include an interactive development environment, code editor, variable explorer,

and integrated console, making it well-suited for tasks such as data cleaning, coding, statistical analysis, and visualization.

2. **Python Language:** Python serves as the primary programming language for this project due to its versatility, ease of use, and extensive libraries for data analysis and visualization. Python libraries such as pandas, NumPy, matplotlib, seaborn, and scikit-learn are essential for data manipulation, statistical analysis, and visualization tasks.

3. **Data Cleaning Tools (Spyder IDE):** Spyder provides functionalities for data cleaning, including tools for handling missing values, removing duplicates, and transforming data. These tools are essential for preparing the dataset for analysis and ensuring data integrity throughout the project.

4. **Coding Environment (Spyder IDE):** Spyder's code editor offers features such as syntax highlighting, code completion, and debugging capabilities, facilitating efficient coding and script development for data analysis tasks.

5. **Statistical Analysis (Python Libraries):** Python libraries such as pandas and scikit-learn are used for statistical analysis tasks, including descriptive statistics, hypothesis testing, and predictive modeling. These libraries provide a wide range of statistical functions and algorithms for analyzing and interpreting data.

6. **Data Visualization (Python Libraries):** Matplotlib and seaborn, along with other visualization libraries in Python, are utilized for creating informative and visually appealing visualizations to communicate insights derived from the data analysis. These libraries offer a variety of plot types, customization options, and interactive features for effective data visualization.

Overall, the combination of Spyder IDE and Python language, along with relevant libraries and tools, provides the necessary resources to effectively clean, analyze, and visualize the dataset, enabling the successful completion of the project objectives.

Timeline for Completing the Project:

SL.No	Methodology	Timeline
1	Data Collection and Preprocessing	March 19 - March 21
2	Questions Selection and Engineering	March 21 - March 26
3	Model Development and Evaluation	March 27 - April 14
4	Deployment and Documentation.	April 14 - April 16

Description of Statistical Analyses

Various statistical methods were employed in this project to analyze the data and derive meaningful insights. Logistic regression was utilized to assess the accuracy of predicting product recommendations

based on the length of the review text. The analysis revealed a high accuracy level of approximately 0.88. This suggests a strong relationship between review text length and the likelihood of product recommendations. Specifically, the results indicated that longer review texts tend to correspond with more positive responses, as evidenced by the most frequently occurring text values.

```
Accuracy: 0.8860675645837933
Top 10 important features:
         Feature  Coefficient
599      perfect     4.555310
495         love     4.443361
177  comfortable     4.060183
184  compliments     4.054834
479       little     3.908177
359        great     3.715823
313         fits     3.594701
348         glad     3.167830
774         soft     2.995775
601    perfectly     2.946056
```

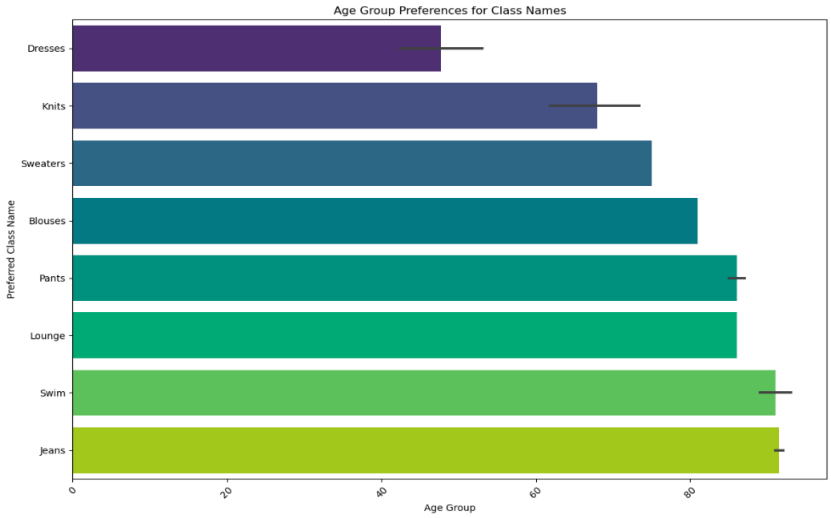
Furthermore, an logistic regression model was employed to investigate the relationship between product ratings and the likelihood of recommendations. The analysis yielded an accuracy level of approximately 0.93, indicating a robust association between higher ratings and an increased likelihood of product recommendations. This finding suggests that individuals who provide higher ratings are more inclined to recommend the product to others.

```
Accuracy: 0.9346434091410908
Classification Report:
              precision    recall  f1-score   support

           0       0.76      0.93      0.84       812
           1       0.98      0.94      0.96      3717

    accuracy                           0.93      4529
   macro avg       0.87      0.93      0.90      4529
weighted avg       0.94      0.93      0.94      4529
```

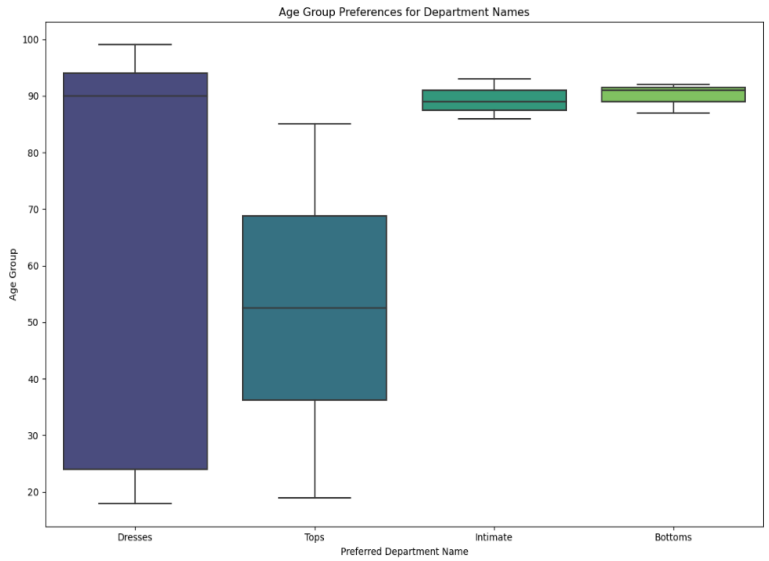
These statistical analyses provide valuable insights into consumer behavior and preferences, enabling a deeper understanding of the factors influencing product recommendations. By leveraging these findings, organizations can refine their marketing strategies, enhance product offerings, and ultimately improve customer satisfaction and loyalty.

Description of Visualizations

1.Age Groups and Preferences: Investigating how different age groups tend to prefer certain class names and department names, shedding light on consumer behavior patterns across demographic segments.



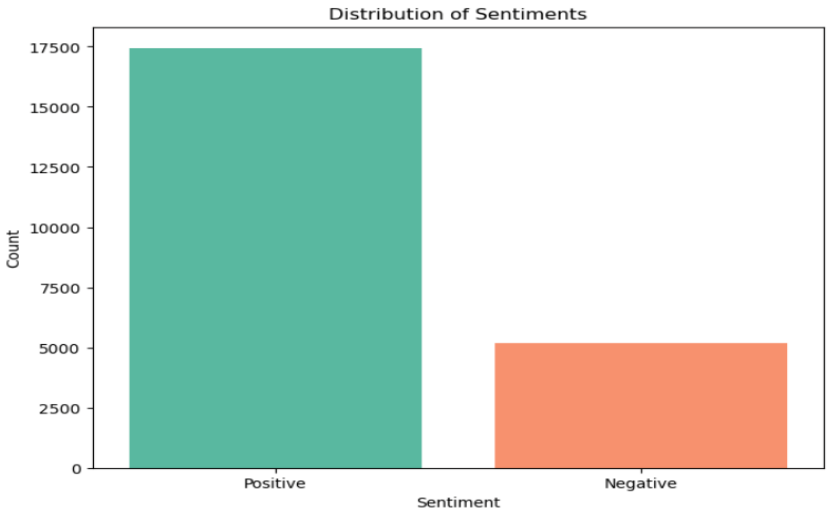
This research question focused on identifying the preferred clothing items across different age groups, a bar plot was utilized for clear visualization. Upon examination of the bar plot, it becomes evident that jeans exhibit widespread popularity across all age demographics, spanning from teenagers to individuals aged 80 and above. Following jeans, dresses maintain their popularity until the mid-50s age group. Subsequently, swim pants, blouses, sweaters, and nightwear follow suit, each in their respective order of popularity.



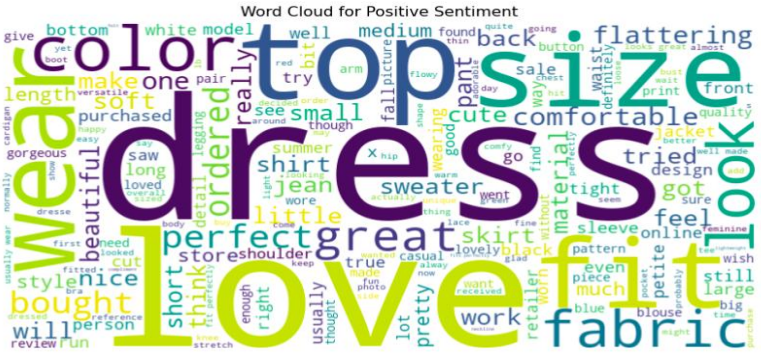
I used a box plot visualization approach to represent the age group references for the different department names. A box plot helps to illustrate the age range for each department, including the lowest, maximum, and average. Across all departments, the recommended age range is from the mid-20s to those aged 90 and

over, with an average age of 90. Following that, the next most desired age group is between 40 and 70, with an average of 50. Finally, the least desired age group, represented by the lower quartile, ranges from intimate dress options to those worn at the bottom of the hierarchy.

2.Sentiment Analysis and Product Recommendations: Assessing the predictive power of sentiment analysis and reviewing text length in determining the likelihood of product recommendations, and identifying key phrases associated with higher recommendation rates.



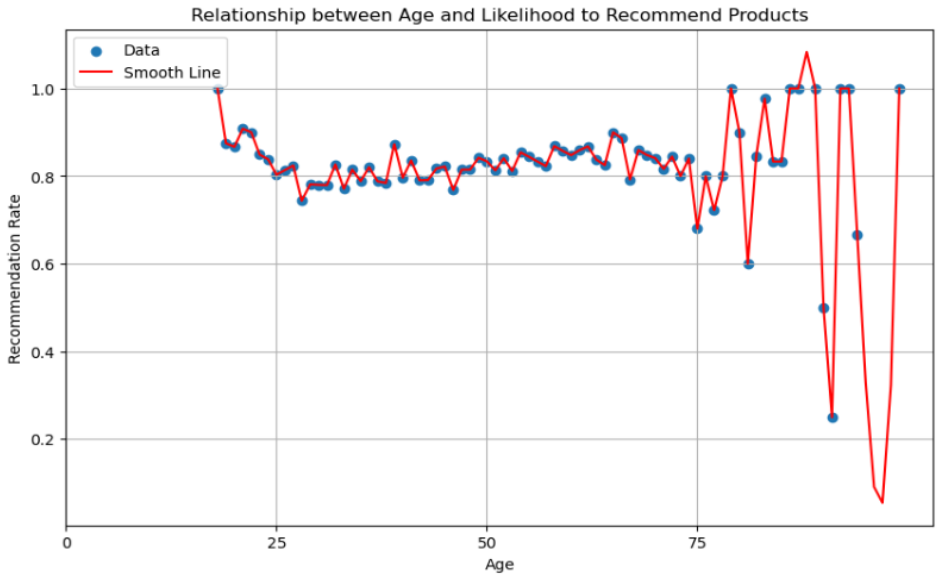
The task entails employing sentiment analysis and analyzing the length of review texts to gauge the probability of product recommendations. By utilizing a bar plot, the distribution of sentiments is visualized, indicating that longer texts tend to correlate with higher recommendation rates. Conversely, shorter texts have less influence.



Furthermore, the word cloud plot is used to identify common words in customer feedback, with concepts such as "dress," "love," "top," and others standing out. These key words shed light on the components of

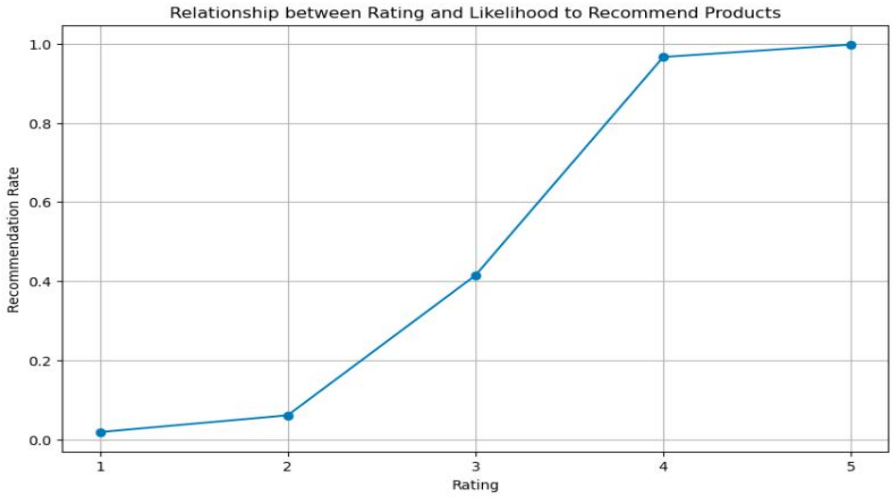
the product experience that are most meaningful to customers, allowing for a more in-depth understanding of consumer mood and preferences.

3.Age and Likelihood to Recommend: Exploring the relationship between age and the likelihood of recommending products, providing insights into age-specific consumer tendencies.



To establish the association between age and likely to recommend things, I used a scatter plot with a smooth line going through it. This technique reveals interesting tendencies in the link between age and product recommendation probability. The graph shows that the chance of recommending items falls after the age of 75. Thus, we may conclude that those under the age of 75 are more likely to promote things to others. Beyond the age of 75, the chance varies but eventually diminishes over time.

4.Rating and Likelihood to Recommend: Analyzing the correlation between product ratings and the likelihood of recommending products, offering insights into the influence of ratings on consumer recommendations.



To determine the link between the rating and the chance of recommending items, I used an interactive line plot within the graph. It is quite simple to derive information from this visualization. According to the research, as the rating grows, so does the possibility of a person making high recommendations, demonstrating a positive relationship between satisfaction and recommendation likelihood. When someone is entirely happy with a product or gives it a good rating, they are more likely to suggest it to others. As a result, consumer satisfaction emerges as a critical component within this group. The greater consumer satisfaction, the more likely they are to suggest the product.

Conclusion:

The project plan provides a structured approach for analyzing clothing review data to derive actionable insights. By following this plan, the organization can harness the power of data to make informed decisions and drive business growth in the competitive retail market.

References:

[1] Nicapotato. (2018). *Women's E-Commerce Clothing Reviews.* CC0: Public Domain [Dataset] Kaggle. https://www.kaggle.com/datasets/nicapotato/womens-ecommerce-clothing-reviews

[2] Barney, N. (2023, December 21). *What is sentiment analysis (opinion mining)?*. TechTarget. https://www.techtarget.com/searchbusinessanalytics/definition/opinion-mining-sentiment-mining

[3]*On-premises vs. Cloud Storage: Which is better?*. Helixstorm. (2021, March 9). https://www.helixstorm.com/blog/on-premises-vs-cloud-storage

[4] *What is natural language processing?*. IBM. (n.d.). https://www.ibm.com/topics/natural-language-processing