

**ALY6040 DATA MINING APPLICATIONS**

**A red and white logo

Description automatically generated**

**MODULE 6 FINAL PROJECT**

**A REPORT ON WOMEN’S CLOTHING E-COMMERCE REVIEWS**

|  |  |
| --- | --- |
| **GROUP MEMBERS** | **NUID** |
| **WAI PHYO MAUNG** | **002737584** |
| **NANA FORDWOUH** | **002626026** |
| **SHREYANSH BHALODIYA** | **002664707** |
| **SURPREET MANN** | **002660066** |
| **MANJYOT KAUR GILL** | **002662913** |
| **PAPA EKOW ARMAH** | **002688073** |

**DR. CHINTHAKA PATHUM DINESH HERATH GEDARA (INSTRUCTOR)**

**OCTOBER 29, 2023**

1. **Introduction**

In the digital age, e-commerce platforms have become the go-to solution for many shoppers. The ability to browse, select, and purchase products from the comfort of one's home has revolutionized the shopping experience. One of the critical aspects of online shopping is customer reviews. These reviews provide insight into product quality, fit, and overall satisfaction. This analysis delves into a dataset containing reviews from a women's clothing e-commerce platform. Our primary objective is to understand the underlying patterns in the reviews and employ machine learning models to predict certain outcomes, possibly related to product recommendations.

1. **Initial Data Overview**

Upon our preliminary examination, the dataset revealed itself as a comprehensive collection of 22,641 reviews. Each review offers a unique perspective, captured across 10 distinct variables, ranging from the age of the reviewer to the nuances of their feedback.

1. **Data Preprocessing** 
   1. Understanding the Dataset's Structure

To lay the foundation for our analysis, we initiated our exploration by discerning the dataset's dimensions. By examining the shape of our data frame, df, we gained insights into its size, revealing the number of rows and columns it encompasses. This preliminary step is pivotal in setting the stage for subsequent data preprocessing tasks.

* 1. Tackling Missing Data

Addressing data gaps is pivotal in preprocessing. Using df.isnull().sum(), we identified columns with missing values, setting the stage for imputation strategies. We determined the percentage of missing data by comparing null counts with the dataset's size (23,468). The "Age" column stood out, missing 16% of its data, underscoring the need for thoughtful imputation. The "Title" column's voids can be filled using a title generator, deriving titles from review content. For other columns with a minor 3% missing data, we recommend deletion, ensuring the dataset's integrity.

* 1. Ensuring Data Uniqueness

Redundancy can be a silent adversary in data analysis. To ensure our dataset was devoid of such duplicities, we created a new data frame, df2, by purging duplicates from df using the drop\_duplicates() method. A juxtaposition of the shapes of df and df2 affirmed the absence of duplicate entries in our dataset. This validation is paramount, ensuring that our analysis remains untainted by repetitive data.

**A screenshot of a computer code

Description automatically generatedA screenshot of a computer

Description automatically generated**

Figure 1. Table of Null Values

1. **Outliers and Anomalies in Data**

In our analysis, we focused on potential outliers within the data frame, df. We extracted key columns: "Age," "Rating," and "Positive Feedback Count" into a new data frame, cr, and subsequently created a copy, df2. To pinpoint outliers, we employed the Z-score scaling method on each column, standardizing the data to have a mean of 0 and a standard deviation of 1. This transformation aids in identifying data points that deviate significantly from the norm. Visualizing this standardized data using a boxplot, it became evident that the "Positive Feedback Count" and "Age" columns contained pronounced outliers. Such anomalies warrant closer scrutiny as they can influence subsequent analyses or modeling endeavors.

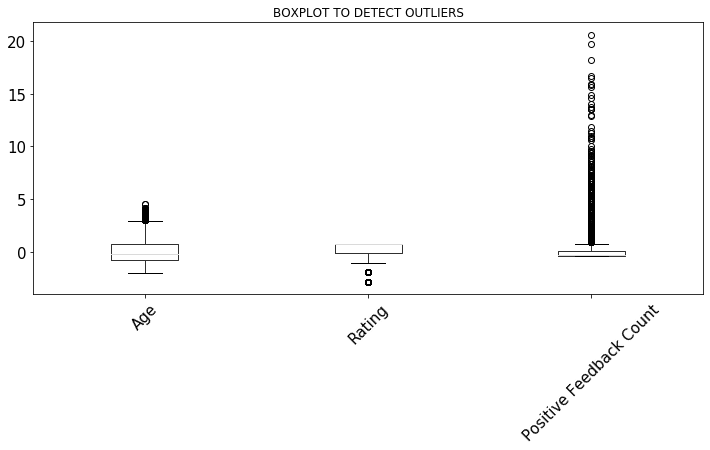


Figure 2. Box Plot to Detect Outliers

1. **Data Cleaning Approach**

To ensure data integrity, we adopted a rigorous cleaning process. We employed the Interquartile Range (IQR) method to identify and manage outliers. Specifically, we created a function, remove\_outliers\_iqr(df), which calculates the quartiles and IQR for each column. Instead of the conventional 1.5 times the IQR, we opted for an 8 times threshold, targeting only extreme outliers while retaining valuable data points. These outliers were replaced with NaN values. Post outlier removal, we standardized the data using Z-score scaling, ensuring uniformity in scale, and facilitating easier analysis. The final step involved visualizing the cleaned data with boxplots, offering a comparative view of the data before and after the cleaning process.

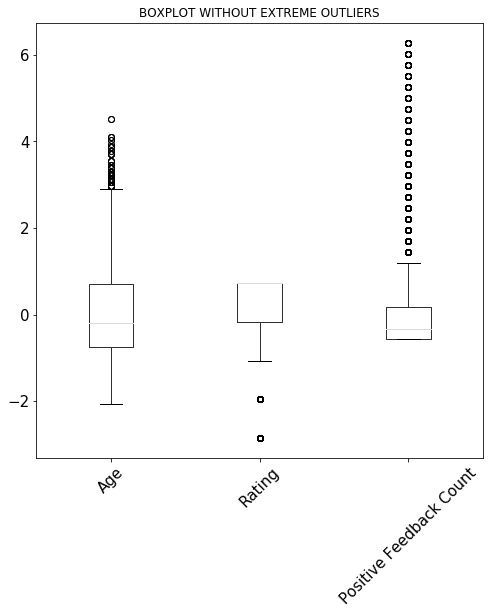


Figure 3. Box Plot Without Extreme Outliers

Our data cleaning approach was tailored to the analysis's specific needs, adjusting the outlier removal threshold. Recognizing that not all outliers are errors, we retained those that added analytical value. A closer look at the "Positive Feedback Count" column illustrates our method's efficacy. Originally, this column peaked at 122, hinting at extreme outliers. Post-cleaning, the maximum value was curtailed to 27, indicating successful outlier elimination. This refined data, with a more consistent feedback count distribution, ensures that subsequent analyses and models are grounded in reliable data, free from undue influences of extreme values.

A screenshot of a computer

Description automatically generated

Figure 4. Table of Summary Statistics

1. **Exploring the Dataset**
   1. Age Distribution of Reviewers

Reviewers' age distribution forms a bell curve, skewed towards younger age groups, with a peak around 35 years. The majority are aged 30-50. Reviews decline sharply for those 60 and above, with minimal contributions from those in their late 60s to early 70s. Essentially, the platform's core reviewing demographic lies between the mid-30s to late 40s, with other age groups less active in feedback.

**A graph of a number of customers

Description automatically generated with medium confidence**

Figure 5: Age Distribution of Reviewers

* 1. Rating Distribution

The retail store garnered predominantly five-star reviews, indicating high customer satisfaction. However, without industry comparisons, it's essential to consider competitor feedback for a comprehensive view. Reviews are subjective; occasional negative feedback might reflect isolated events rather than overall performance.

**A bar graph with different colored bars

Description automatically generated**

Figure 6: Rating Distribution

1. **Modelling**
   1. Product Recommendation

As the model selection process progressed, the dataset underwent a series of refinements for use in the machine learning model. Our target variable, the 'Recommended IND' column, was intentionally omitted because it contains binary data implying product recommendations.

This carefully curated dataset was constructed to include certain attributes that play a pivotal role in our modeling work. These attributes, namely "Rating," "Positive Feedback Count," "Polarity Score," and "Logistic\_Probability," were selected because of their importance in building predictive models that can evaluate product recommendations based on a variety of product characteristics. The training set accounts for 70% of the data and serves as the underlying component to direct and train the machine learning model. Conversely, the test set, which accounts for the remaining 30%, is used to evaluate the predictive performance of the model. This partitioning strategy was thoughtfully designed to reliably assess the model's capabilities.

Furthermore, it is worth noting that approximately 81.62% of the products in the training set were labeled as recommended, while approximately 82.47% of the products in the test set received recommendations. This distributional insight plays a fundamental role in our analysis because it provides an important context for understanding the balance between recommended and non-recommended products in our dataset. This understanding is pivotal when it comes to assessing the model's performance and its ability to make accurate recommendations. In the rapidly evolving world of e-commerce, understanding customer sentiments is paramount. This report delves into the application and findings of three machine learning models on the dataset of women's clothing e-commerce reviews. The models in focus are Naive Bayes, Support Vector Machine, and Logistic Regression.

* 1. Product Rating

The model proposed here focuses on sentiment analysis as a means of gaining insight from customer reviews. The goal is to understand how sentiment expressed in text relates to a given rating score and the likelihood that a product will be recommended. This approach allows us to distinguish obvious positive and negative comments from those that are more constructive in tone.

In this process, the model employs a supervised learning approach. Instead of analyzing consecutive words in the text, it focuses on the presence of specific words from a predefined dictionary in the customer comment corpus. To make the computation more efficient, the corpus contains about 9000 unique words, but only the top 5000 most frequently occurring words are considered. An important component is the "find\_features" function, which evaluates the presence of the top 5000 words in each text. This function uses a variable called "word\_features". This variable has been previously created to contain the most common words used by customers in the dataset.

In the "Apply Function to Data" section, a loop is used to implement this "find\_features" function for individual customer reviews. During this process, the model also maintains a label associated with each review, which is essential for subsequent analysis and classification tasks.

1. **Key Findings**
   1. Product Recommendation

The provided data presents the performance metrics for three different machine learning models: Logistic Regression, Support Vector Machines (SVM), and Naïve Bayes. These models are evaluated using key metrics such as accuracy, precision, recall, and F1-score.

* + 1. Logistic Regression

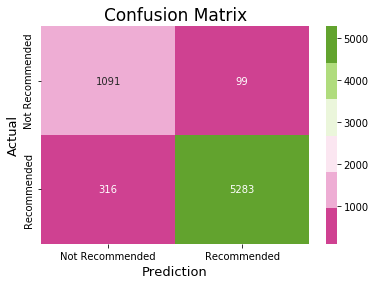


Figure 7: Confusion Matrix for Logistic Regression Model

In the evaluation of the Logistic Regression model, a comprehensive analysis was conducted, starting with the examination of its confusion matrix. This matrix revealed that the model yielded 1091 true negatives, 99 false positives, 316 false negatives, and 5283 true positives. Following this, critical performance metrics were calculated. The accuracy of the model was found to be approximately 93.89%, indicating that it correctly predicts the outcome in about 93.89% of cases. Furthermore, the precision of approximately 98.16% revealed that when the model predicts a positive outcome, it is correct about 98.16% of the time, underscoring its reliability. The recall, standing at around 94.36%, signifies that the model successfully identifies approximately 94.36% of actual positive cases. Lastly, the F1-score, which equates to about 0.9622, serves as a robust balance between precision and recall, solidifying the model's efficacy in striking that crucial equilibrium.

* + 1. Support Vector Machines

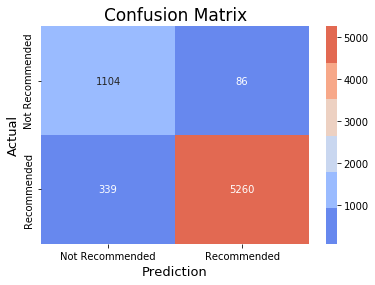


Figure 8: Confusion Matrix for Support Vector Machine Model

The matrix reveals that SVM yielded 1104 true negatives, 86 false positives, 339 false negatives, and 5260 true positives. This insightful analysis is complemented by key performance metrics. The accuracy of SVM is found to be approximately 93.74%, indicating that it achieves an impressive accuracy rate of about 93.74%. Moreover, the precision, standing at approximately 98.39%, illustrates that when SVM predicts a positive outcome, it is correct about 98.39% of the time, showcasing its robust predictive capabilities. The recall, reaching around 93.95%, signifies that SVM effectively identifies approximately 93.95% of actual positive cases, emphasizing its sensitivity. Finally, the F1-score, approximately 0.9612, mirrors Logistic Regression by serving as a balanced metric between precision and recall, underlining SVM's capacity to harmonize these vital aspects of model performance.

* + 1. Naïve Bayes

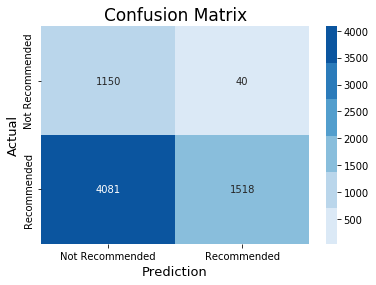


Figure 9: Confusion Matrix for Naïve Bayes Model

The assessment of the Naïve Bayes model reveals a distinct performance profile. Examining its confusion matrix shows that Naïve Bayes resulted in 1150 true negatives, 40 false positives, 4081 false negatives, and 1518 true positives. Notably, the model's accuracy is substantially lower at about 0.3930, indicating that its predictions are correct in only 39.30% of cases. On the positive side, the precision is relatively high, standing at approximately 0.9743, signifying that when Naïve Bayes predicts a positive outcome, it is correct about 97.43% of the time, demonstrating its precision in positive predictions. However, its recall is notably low at around 0.2711, indicating that Naïve Bayes identifies only about 27.11% of actual positive cases, suggesting a lack of sensitivity. The F1-score, although representing a balance between precision and recall, is also relatively low, approximately 0.4242, underlining the trade-off between these crucial performance aspects but at lower levels compared to the other models.

* 1. Product Rating

The provided analysis focuses on the sentiment of customer reviews by examining the most commonly occurring positive and negative words associated with different rating scores (specifically, scores of 4 and 5 for positive words and scores of 1 and 2 for negative words).

For positive words, the analysis identifies frequently used positive terms like "Love," "Glad," "Size," "Great," "Fit," "Perfect" and "Color" among high ratings (4 and 5). These words suggest that customers express high levels of satisfaction when they find products that are true to size, fit well, and have attractive colors. In essence, these terms are indicative of positive product attributes, contributing to favorable reviews.

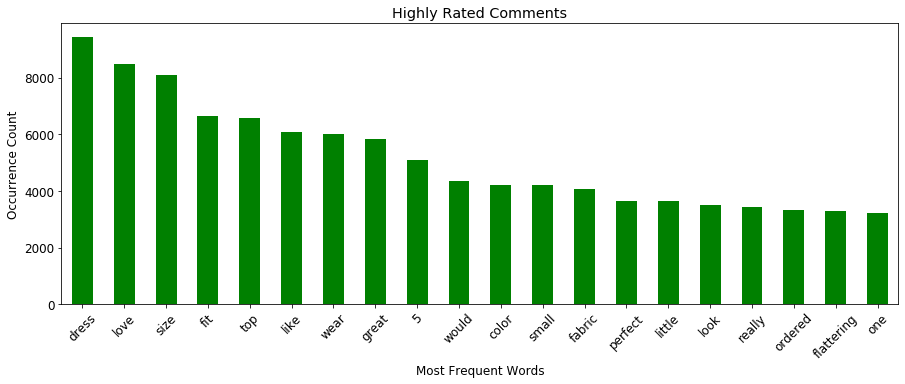


Figure 10: Bar plot for Positive Reviews

Conversely, when examining negative words associated with low ratings (1 and 2), the analysis highlights terms like 'disappointment,' 'tight,' 'small,' 'return,' and 'cheap.' These negative words indicate areas of concern for customers, such as issues related to product fit, size, quality, or overall disappointment. These terms are often associated with lower-rated products, highlighting aspects that lead to less favorable reviews.

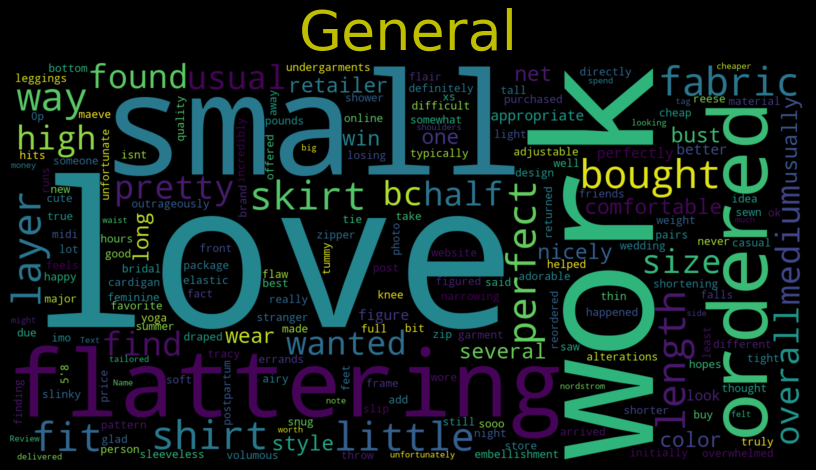


Figure 11: Word Cloud for Negative Reviews

1. **NLP-modelling**

Natural Language Processing (NLP) individual model for product recommendations. The primary goal of this model is to analyze customer reviews and provide recommendations based on sentiment analysis and other text processing techniques.

9.1 Input Column:

The model utilizes a review column as the input data source.

9.2 NLP Text Processing:

9.2.1 Sentiment Analysis:The model employs sentiment analysis to understand the emotional tone of the reviews, helping distinguish between positive and negative sentiments.

9.2.2 Word Cloud: A word cloud visualizes the most frequently occurring words in the reviews, offering insights into commonly mentioned terms.

N-grams: The model uses N-grams to analyze the frequency of word combinations in the text.

9.2.3 TF IDF Vectors: Term Frequency-Inverse Document Frequency vectors are used to represent the significance of words in the reviews.

9.3 Prediction - Product Recommendation (Yes/No):

The model predicts whether a product should be recommended to customers based on the NLP analysis.

9.4 Results

Validation Set Accuracy: 0.88

Validation Set ROC (Receiver Operating Characteristic): 0.84

These results indicate that the NLP model exhibits strong performance in identifying sentiments and making product recommendations. It is a valuable tool for businesses seeking to enhance their product recommendation systems and improve customer satisfaction.

**10. Final Modelling**

This report presents the final model developed for product recommendation, which leverages Natural Language Processing (NLP) text processing, sentiment analysis, and logistic regression modeling. The model's input data is based on customer reviews, and it produces two new variables, Polarity Score (Sentiment Score) and Logistic Regression Probabilities, ultimately leading to a recommendation of "Yes" or "No."

10.1 Final Model Data Preparation Steps:

10.1.1 Input: Reviews

10.1.2 NLP Text Processing: The reviews undergo NLP text processing, which includes sentiment analysis to determine polarity scores.

10.1.3 Logistic Regression Modeling: Logistic regression is applied to calculate probabilities.

10.1.4 Final Model Variables:

Rating

Product Positive Count

Polarity Score (Sentiment Score)

Logistic Regression Probability

Prediction:

The model predicts whether a product should be recommended (Yes) or not (No) based on the combined information from the variables.

10.2 Results:

The final model's performance was assessed using various evaluation metrics across different machine learning algorithms:

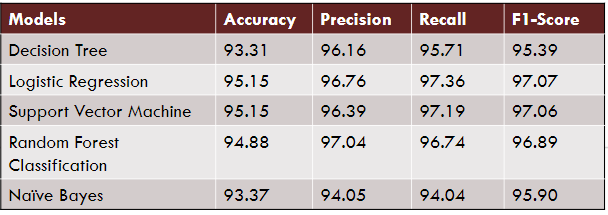


Figure 12. Performance of Final Models

The Logistic Regression model stands out with the highest accuracy and overall performance, making it the preferred choice for product recommendation. It achieved the best balance of precision and recall, ensuring accurate recommendations and customer satisfaction.

1. **Interpretation**

Both Logistic Regression and SVM demonstrate exceptional predictive capabilities, with significantly higher accuracy, precision, recall, and F1-scores when compared to Naïve Bayes. These models stand out for their proficiency in correctly identifying products that customers are likely to recommend, thereby ensuring a high level of accuracy in prediction. Importantly, the performance metrics of Logistic Regression and SVM exhibit striking similarities, underlining their effectiveness in delivering accurate product recommendations that align closely with customer preferences. This high level of consistency between the two models reinforces their position as the top choices for predicting customer recommendations effectively.

However, a noticeable performance gap becomes evident when contrasting these models with Naïve Bayes. Naïve Bayes falls behind, manifesting the lowest accuracy, recall, and F1-score among the trio of models. This performance shortfall implies that Naïve Bayes has limitations in accurately predicting which products customers are likely to recommend, resulting in an overall suboptimal performance in the context of product recommendation. It is noteworthy that while Naïve Bayes excels in achieving high precision, indicating that its positive predictions are likely to be correct, this accomplishment comes at the expense of a significantly low recall. This means that Naïve Bayes might miss a substantial number of products that customers would recommend, emphasizing a distinct trade-off between precision and recall, a trade-off that is far less pronounced in the other models. Thus, when seeking the optimal model for predicting customer recommendations, the consistent and superior performance of Logistic Regression and SVM positions them as the preferred choices, outperforming Naïve Bayes across various essential metrics.

A product recommendation usually serves as a clear and direct signal of positive sentiment in a review. When reviewers recommend a product, it means they are satisfied with it and feel it is valuable. Because this type of sentiment is relatively clear, it is a valuable indicator for companies to measure customer satisfaction and identify products that are highly favored by consumers.

Product ratings, on the other hand, tend to be more nuanced and multifaceted. Often, ratings on a scale of 1 to 5 can represent a broader range of feelings and opinions. In many cases, a product with a rating of around 3 may be accompanied by a review that combines constructive criticism with hopeful sentiment. Such reviews indicate that the product is neither particularly good nor particularly bad, but somewhere in the middle. Reviewers in this category may highlight what they appreciate about the product while also pointing out potential improvements.

1. **Recommendation**

For product recommendation, we recommend utilizing either the Logistic Regression or Support Vector Machines (SVM) models due to their consistent and superior performance in terms of accuracy, precision, recall, and F1-scores. These models are well-suited for predicting which products customers are likely to recommend, ensuring that the recommendations closely align with customer preferences. Their high accuracy and balanced precision and recall make them ideal for making accurate and reliable product recommendations.

For product rating analysis, we suggest employing sentiment analysis to understand how specific words in customer reviews relate to different rating scores. This approach allows for the identification of both positive and negative sentiments associated with different ratings, offering valuable insights into customer feedback. This analysis can help identify areas of improvement for products and understand the aspects that lead to positive or negative ratings. Additionally, it can be beneficial to explore the sentiments associated with middle-range ratings as these often contain constructive criticism and offer opportunities for product enhancement.

1. **Conclusion**

In conclusion, our analysis has provided valuable insights into customer reviews and product recommendations in the context of e-commerce. We began by thoroughly examining a dataset containing thousands of reviews, understanding its structure, addressing missing data, and ensuring data uniqueness. We then identified potential outliers and applied a rigorous data cleaning approach to maintain data integrity.

Exploring the dataset revealed key findings, particularly in the domain of product recommendation and product rating analysis. Logistic Regression and Support Vector Machines (SVM) emerged as top-performing models for product recommendation, consistently outperforming Naïve Bayes across essential metrics. These models exhibited high accuracy, precision, recall, and F1-scores, making them the preferred choices for predicting customer recommendations accurately. In contrast, Naïve Bayes demonstrated limitations in correctly predicting product recommendations, primarily due to a trade-off between precision and recall.

In the realm of product rating analysis, sentiment analysis proved effective in understanding how customer sentiment relates to different rating scores. Positive words like "Love," "Glad," "Size," and "Great" were indicative of high ratings, while negative words like 'disappointment' and 'return' were associated with lower ratings. This analysis offers valuable insights into customer feedback and areas for product improvement.

Based on our findings, we recommend using Logistic Regression or SVM for product recommendation due to their consistent and superior performance. For product rating analysis, employing sentiment analysis can provide valuable insights into customer feedback, helping businesses understand what aspects of a product are appreciated and where improvements are needed. This comprehensive approach ensures that companies can make data-driven decisions to enhance customer satisfaction and product quality in the competitive world of e-commerce.

**References**

Fakhitah Ridzuan, W. M. (2019). A Review on Data Cleansing Methods for Big Data. Retrieved from ScienceDirect: https://www.sciencedirect.com/science/article/pii/S1877050919318885

Sikibi, M. (2021, May 21). Use Data Mining Cleansing to Prepare Data for Strategic Decisions. Retrieved from IntechOpen: https://www.intechopen.com/chapters/77946

Women's E-Commerce Clothing Reviews. (n.d.). Retrieved from Kaggle: https://www.kaggle.com/datasets/nicapotato/womens-ecommerce-clothing-reviews