EXECUTIVE SUMMARY

This report presents an analysis of a comprehensive dataset of Amazon customer reviews for **beauty products**, obtained from the Hugging Face datasets library. The dataset comprises **701,528 reviews**, providing a rich source of information for understanding customer opinions and identifying patterns within the text. Each review is associated with several attributes, including product ratings, review text, and helpfulness votes. The analysis involved several key steps:

1. Feature Analysis and Visualization:

Initial exploration focused on understanding the **distribution and characteristics of key features**. Numerical features such as **product ratings and helpfulness votes** were analyzed using descriptive statistics. Visualizations, including **histograms and box plots**, were generated to illustrate the distribution of ratings and helpfulness votes. These visualizations revealed that product ratings are heavily skewed towards 5 stars, indicating a generally positive customer sentiment. The distribution of helpfulness votes was highly skewed, with most reviews receiving very few helpful votes and a few reviews receiving a disproportionately high number of votes.

2. Analysis of Review Text Characteristics:

To gain insights into the textual content of the reviews, several analyses were conducted. **Review length** was calculated in both **characters** and **words** to understand the granularity of the reviews. The **most frequent words** used in the reviews were identified and visualized using a word cloud and bar chart, providing a quick overview of common topics discussed by customers.

3. Data Preprocessing:

Missing values were identified and handled to ensure data quality. **No missing values** were found in the dataset.

4. Sentiment Analysis:

Sentiment analysis was performed on the review texts to quantify the **emotional tone** of customer feedback. Sentiment scores **(negative, neutral, positive, and compound)** were calculated for each review, allowing for an analysis of the overall sentiment distribution. The relationship between average product ratings and average review sentiment was also explored, providing insights into how customer sentiment aligns with product satisfaction.

5. Review Length and Complexity Features:

Additional features were engineered to capture **review length and complexity**. The **Flesch Reading Ease score** was calculated as a measure of text readability. The relationship between review length and helpfulness votes was analyzed to identify if longer or shorter reviews tend to be rated as more helpful.

6. TF-IDF for Numerical Representation:

To convert the review texts into a numerical format suitable for machine learning, the Term Frequency-Inverse Document Frequency (TF-IDF) technique was employed. This resulted in a matrix representation of the reviews, where each row corresponds to a review, and each column corresponds to a word, with the values representing the TF-IDF scores.

7. Anomaly Detection:

Anomaly detection techniques were applied to **identify unusual reviews**.

ANOMALY DETECTION USING Z-SCORE: REVIEW LENGTH TO RATING RATIO

1. Importance of Detecting Anomalies in Review Length to Rating Ratio:

Detecting anomalies in the ratio of review length to rating is crucial for identifying reviews that deviate from typical patterns, which can help in:

- **Identifying potentially inauthentic reviews**: Extremely short, positive reviews or excessively long, negative reviews might warrant closer inspection.
- **Understanding customer behavior**: Anomalies can reveal unusual customer feedback patterns, such as customers who are unusually verbose when dissatisfied.
- **Improving data quality**: Removing or flagging anomalous reviews can lead to a more accurate representation of overall customer sentiment.

2. Approach and Code Implementation:

The approach involves calculating the ratio of review length to rating and then using Z-scores to pinpoint reviews with ratios that are statistically unusual.

The code performs the following steps:

- **Calculate Review Length to Rating Ratio**: The review_length_to_rating_ratio is computed by dividing the number of words in the review (review_length_word) by the rating. This ratio standardizes review length relative to the rating.
- **Handle Edge Cases**: To avoid errors, infinite values and NaNs in the ratio are replaced with a large number (1000) and removed, respectively.
- **Calculate Z-Scores**: Z-scores are calculated for the review_length_to_rating_ratio. The Z-score measures how far each review's ratio is from the average ratio, in terms of standard deviations.
- **Identify Outliers**: Reviews with an absolute Z-score exceeding a predefined threshold are flagged as outliers.

3. Parameters and Justification:

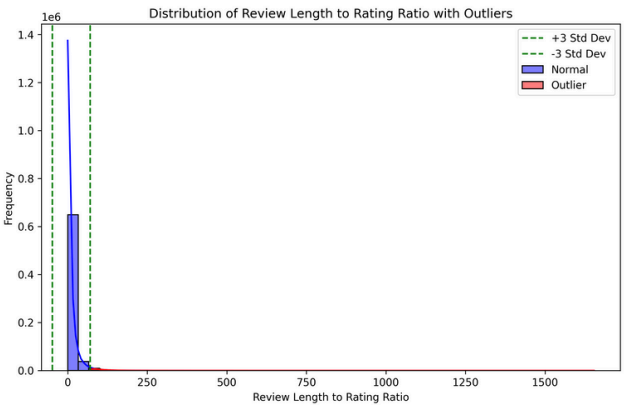
- **Z-score Threshold**: A threshold of 3 is used. This is a common statistical convention, as it assumes that in a roughly normal distribution, about **99.7% of data points** fall within 3 standard deviations of the mean. Therefore, values beyond this threshold are considered rare or unusual.

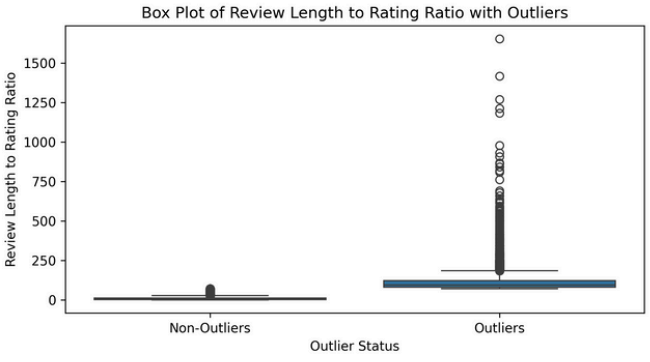
4. Results and Interpretation:

- **Visualizations:**
  - **Histogram**: The histogram "Distribution of Review Length to Rating Ratio with Outliers" displays the distribution of the ratio. Outliers, shown in red, are concentrated at the higher end, indicating reviews that are long relative to their rating. The green lines mark the Z-score thresholds.
  - **Box Plot**: The box plot "Box Plot of Review Length to Rating Ratio with Outliers" visually confirms the presence of high-ratio outliers, positioned far beyond the box's whiskers.

- **Outlier Review Examples**: The analysis identified several reviews with high Z-scores, indicating they are unusual in terms of their length-to-rating ratio. Here are a few examples:
  - Review 1 (Z-Score: 82.69, Ratio: 1654.00): This review of a skincare serum is extremely long (1654 words) and critical, resulting in a rating of 1. The reviewer provides a detailed breakdown of the product's ingredients, claims, and price point, expressing significant dissatisfaction. The high ratio and Z-score suggest that the length is highly disproportionate to the negative rating.
  - Review 2 (Z-Score: 70.81, Ratio: 1418.00): This review of a massage chair is also very long (1418 words) and negative, with a rating of 1. The reviewer meticulously details issues with the product's features, misleading advertising, and seller problems. Similar to the first example, the extensive text and low rating contribute to a high ratio and Z-score.
  - Review 3 (Z-Score: 63.36, Ratio: 1270.00): This review of a foundation is lengthy (1270 words) and critical, with a rating of 1. The reviewer meticulously examines the product's claims, ingredients, shade, and pump design, expressing disappointment and highlighting misinformation. The significant length and low rating result in a high ratio and Z-score.
- **Interpretation:**
  - The analysis confirms the **presence of reviews with unusual length-to-rating ratios**.
  - The identified outliers are predominantly characterized by **very long**, detailed reviews with **low ratings**. This suggests that customers tend to write extensively when they are **highly dissatisfied with a product**.
  - These outliers could be valuable for **identifying products with significant issues** or areas where **customer expectations are not being me**t.
  - While these outliers are **statistically unusual**, they appear to be **genuine**, **detailed feedback from highly dissatisfied customers**.

This refined report provides a more detailed explanation of the Z-score analysis, incorporating insights from the provided outlier review examples.





ANOMALY DETECTION USING Z-SCORE: HELPFUL VOTES

1. Importance of Detecting Anomalies in Helpful Votes:

Identifying reviews with an unusually high number of helpful votes is important because these reviews often stand out for a reason. They might be:

- **Highly insightful or informative:** These reviews provide valuable perspectives or tips that resonate with many users.
- **Controversial or polarizing:** These reviews might express strong opinions that elicit significant engagement.
- **Early indicators of product quality or issues:** Reviews that quickly garner many helpful votes, whether positive or negative, can signal important aspects of a product.
- **Potentially manipulated**: While less likely with high helpful votes, extreme outliers could warrant investigation in conjunction with other anomaly detection methods.

2. Approach and Code Implementation:

The approach involves calculating the Z-score for the 'helpful_vote' count of each review to identify those that deviate significantly from the average.

The code performs the following steps:

- **Calculate Z-scores:** The Z-score for the 'helpful_vote' column is computed. This measures how many standard deviations each review's helpful vote count is from the mean helpful vote count.
- **Identify Outliers:** Reviews with an absolute Z-score greater than a specified threshold are flagged as outliers.

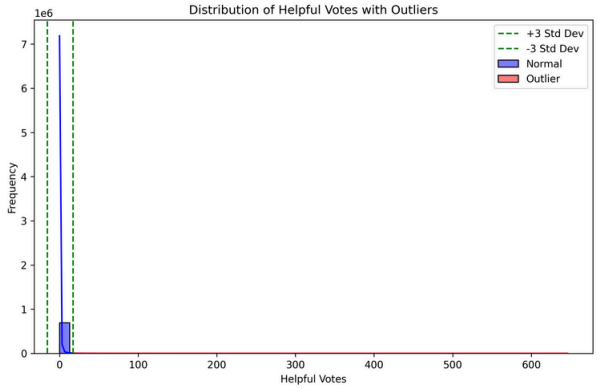
3. Parameters and Justification:

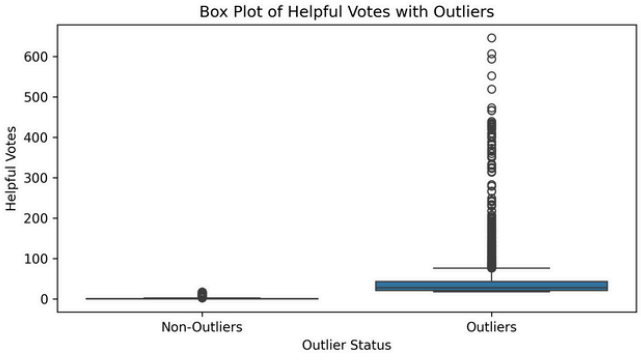
- **Z-score Threshold:** A threshold of 3 is used. As discussed previously, this threshold is based on the statistical property that approximately 99.7% of data in a normal distribution falls within 3 standard deviations of the mean. Values beyond this are considered unusual.

4. Results and Interpretation:

- **Visualizations:**
  - **Histogram:** The histogram "Distribution of Helpful Votes with Outliers" illustrates the distribution of 'helpful_vote' counts. The majority of reviews have a low number of helpful votes, with the frequency decreasing sharply as the count increases. Outliers, colored in red, are found in the far right tail of the distribution, representing reviews with exceptionally high helpful vote counts. The green lines indicate the Z-score thresholds.
  - **Box Plot:** The box plot "Box Plot of Helpful Votes with Outliers" provides a clear view of the distribution and highlights outliers as points far beyond the upper whisker, again confirming the presence of reviews with significantly higher helpful vote counts.

- **Outlier Review Examples:** The code identified several reviews with notably high Z-scores for 'helpful_vote'. Here are the top 5 examples:
  - Review 1 (Helpful Votes: 646, Z-Score: 117.90): This positive review from a user with multiple skin conditions describes the product as a "miracle," detailing significant improvements in their skin. The exceptionally high helpful vote count suggests this review resonated strongly with a large number of users, likely due to its detailed and enthusiastic account of positive results for challenging skin issues.
  - Review 2 (Helpful Votes: 607, Z-Score: 110.77): This review, written by a dentist, provides professional advice on using Crest White Strips effectively, including additional tips beyond the product instructions. The high helpful vote count indicates that many users found this expert guidance valuable.
  - Review 3 (Helpful Votes: 594, Z-Score: 108.40): This review describes a user's experience helping a family member with severe dental calculus using a specific product. The high helpful vote count suggests that others facing similar issues found this detailed, firsthand account and the described results to be very informative.
  - Review 4 (Helpful Votes: 552, Z-Score: 100.72): This review provides a detailed assessment of a dark brown eyebrow tint, including application tips and a photo of the results. The high helpful vote count indicates that users interested in this type of product found the specific details and visual aid useful.
  - Review 5 (Helpful Votes: 519, Z-Score: 94.69): This lengthy and thoughtful review offers a nuanced perspective on a galvanic electrolysis device, explaining how it works, its limitations, and tips for effective use. The high helpful vote count suggests that users considering or using this type of product found the in-depth explanation and realistic expectations valuable.
- **Interpretation:**
  - The Z-score analysis effectively identifies reviews that have received **a significantly higher number of helpful votes** compared to the vast majority of reviews.
  - The outlier reviews tend to be **highly informative,** provide **expert advice**, or offer compelling **personal experiences with the products**.
  - These high-engagement reviews are valuable as they likely represent **the most helpful and influential opinions** within the dataset. They can provide key insights into product benefits, usage, and potential issues that resonate with a broad audience.





ANOMALY DETECTION IMPLEMENTATION: MACHINE LEARNING METHOD (ISOLATION FOREST)

1. Why detecting anomalies using a machine learning method is important:

Multivariate anomaly detection, like that performed by Isolation Forest, is crucial for identifying unusual reviews based on the combination of multiple features. This can uncover subtle anomalies that might be missed by univariate methods (like Z-score on a single feature). For instance, a review might have a normal review length and sentiment individually, but the specific combination of a very short review with a highly positive sentiment and a low helpful vote count could be anomalous and potentially indicative of spam or a biased review. [cite: 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46]

2. What we did in the code and what was our approach:

We implemented the Isolation Forest algorithm, an unsupervised machine learning method designed to isolate anomalies. Our approach involved the following steps:

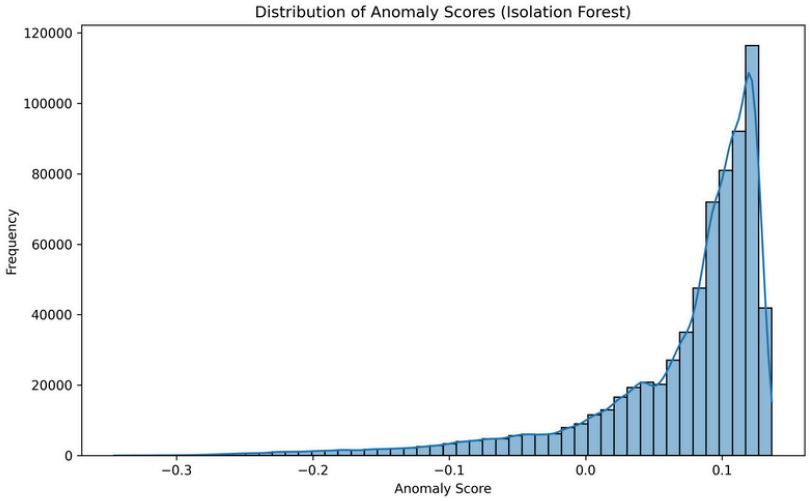
- **Feature Selection**: We selected a set of features relevant to identifying unusual review patterns: review_length_char (review length in characters), helpful_vote (number of helpful votes), sentiment_compound (overall sentiment score), and review_length_word (review length in words).
- **Data Preprocessing:** We handled missing values in the selected features by imputing them with 0. This ensures that the Isolation Forest algorithm can process all the data.
- Feature Scaling: We scaled the selected features using StandardScaler. This is important for distance-based algorithms like Isolation Forest to prevent features with larger ranges from dominating the results. Scaling ensures that each feature contributes equally to the anomaly detection process.
- **Model Training:** We initialized and trained an IsolationForest model on the scaled features. The model learns the normal patterns in the data.
- **Anomaly Prediction:** We used the trained model to predict an anomaly score for each review using the decision_function method. A negative score indicates that the instance is likely an anomaly, while a positive score suggests it's more likely an inlier. We also used the predict method to get a binary label (-1 for anomaly, 1 for inlier).
- **Visualization:** We visualized the distribution of the anomaly scores to understand the overall anomaly landscape in the dataset.
- **Anomaly Examination:** We identified and printed the text and anomaly scores of the top 10 most anomalous reviews (those with the most negative scores).

3. Parameters we used and why we used them:

- **n_estimators=100:** This parameter specifies the number of isolation trees to be built in the forest. A larger number generally leads to more stable and accurate results, although it also increases computation time. We used 100 trees as a reasonable balance between performance and computational cost.
- **contamination='auto':** This parameter estimates the proportion of outliers in the dataset. Setting it to 'auto' allows the algorithm to automatically determine a reasonable contamination level based on the data.
- **random_state=42:** This parameter ensures reproducibility of the results. By setting a specific random seed, we guarantee that the tree construction and anomaly scoring process will be the same each time the code is run.

4. What the final output and result shows and interpret the visualizations and sample reviews:

- **Distribution of Anomaly Scores (Isolation Forest) Visualization:** The histogram displays the distribution of anomaly scores assigned by the Isolation Forest model. The majority of reviews have anomaly scores clustered towards the positive side, indicating they are considered more "normal." A tail of reviews extends towards the negative anomaly scores, representing those identified as more anomalous by the model. The further the score is into the negative, the more likely the review is considered an outlier.
- **--- Anomalous Reviews Examples --- :** The output provides the text and corresponding anomaly scores for the top 10 reviews identified as most anomalous (having the most negative anomaly scores).
  - The example reviews showcase a variety of reasons why a review might be flagged as anomalous based on the combination of the selected features:
    - Review 9 (Score: -0.194): This very short and informal review with unusual language ("baby caterpillars") might be an anomaly due to its deviation in writing style and length compared to typical reviews.
    - Review 10 (Score: -0.232): This long and strongly negative review containing personal commentary about the reviewer's status might be an anomaly due to the extreme sentiment combined with the meta-commentary.
    - Other reviews with less extreme negative scores might be flagged due to more subtle combinations of their length, helpful vote count, and sentiment score that deviate from the typical patterns observed by the Isolation Forest model. For instance, a very short review with a neutral sentiment and no helpful votes, or a very long review with a positive sentiment but also no helpful votes, could be considered unusual.
- **Interpretation:** The Isolation Forest model identified a set of reviews that are considered unusual based on their combined characteristics of review length, helpful votes, and sentiment. These anomalies might represent various phenomena, including:
  - Reviews with **atypical** writing styles or content.
  - Reviews that have **unusual combinations of length and sentiment**.
  - Potentially **less genuine or biased reviews** (though further investigation would be needed to confirm this).
  - Reviews that **stand out significantly** from the majority of the data points in the multi-dimensional feature space.



1- Comparison of methods

We compared three anomaly detection methods: two based on Z-score and one using a machine learning model (Isolation Forest). The Z-score methods are simple and fast — one focuses on mismatches between review length and rating, flagging cases like long 1-star reviews, while the other detects reviews with unusually high helpful votes, highlighting those that stood out to readers. However, both are limited because they analyze just one feature at a time and assume a normal distribution.

In contrast, Isolation Forest looks at multiple features together — like rating, length, and helpfulness — allowing it to uncover more complex and subtle anomalies that the Z-score methods would miss. It's more powerful and flexible but also more complex to interpret and fine-tune. Overall, while Z-score methods are good for quick checks, Isolation Forest offers deeper insights for more nuanced outlier detection.

2- Sensitivity analysis

- **Impact of changing the Z-score threshold in Review Length to Rating Ratio:**

Changing the Z-score threshold directly affects the number of identified outliers. As the threshold increases from 2 to 4, the number of data points classified as outliers decreases significantly, from 24,318 to 6,914. This is because a higher threshold requires data points to deviate more extremely from the mean to be considered outliers, leading to a more conservative identification of anomalies.

- **Sensitivity Analysis of Isolation Fores:**

The n_estimators parameter in Isolation Forest controls the number of isolation trees in the ensemble. Increasing it generally improves the model's ability to isolate anomalies. With n_estimators set to 200, we observe a slightly smoother distribution of anomaly scores in the histogram, suggesting a more refined model output. The top anomalous reviews also exhibit slightly different anomaly scores compared to the results with 100 estimators, indicating that the model's assessment of individual review unusualness has been affected. This change implies that the model with 200 trees is likely capturing the underlying data patterns more effectively, leading to potentially more accurate and reliable anomaly detection.

BUSINESS INSIGHTS REPORT: ANOMALY DETECTION IN CUSTOMER REVIEWS

The ratio between review length and rating:

Analysis of the review length to rating ratio revealed outliers where the length of a review doesn't align with the star rating. For instance, some very short reviews have high ratings, or very long reviews have low ratings. This could indicate several issues, including biased reviews, where the reviewer didn't spend enough time to provide a detailed opinion, or potentially paid reviews, where the reviewer was incentivized to leave a positive rating without providing substantial feedback. To address this, businesses could implement a system that flags reviews with extreme length-to-rating ratios for further scrutiny. They might also consider incorporating review length as a factor in their review ranking algorithm, giving more weight to reviews that provide more detailed feedback.

Unusual patterns in helpful votes:

The analysis of helpful votes identified reviews with an unusually high number of helpful votes. These reviews are valuable as they resonate strongly with other customers. Businesses should prioritize these reviews by prominently displaying them on product pages, as they can significantly influence purchase decisions. It may also be beneficial to analyze the content of these highly helpful reviews to identify key product features or customer concerns that are driving purchase behavior.

Machine Learning Method:

The Isolation Forest model identified anomalous reviews based on a combination of factors, including review length, helpful votes, and sentiment. These anomalies may represent various issues, such as spam reviews, fake reviews, or reviews with manipulated content. Businesses should implement a robust system to detect and filter out these anomalous reviews, as they can negatively impact customer trust and distort product perception. This could involve a combination of automated filtering based on anomaly scores and manual review by human moderators.