Report of data analysis:

Dataset description:

Dataset : amason_consumer_preview.csv

This dataset comprises **5,000 consumer reviews** sourced from Datafiniti's Product Database. Downloaded from Kaggle.The reviews encompass a variety of Amazon products, including popular choices like the Kindle and Fire TV Stick.
Each review entry provides valuable insights into customer sentiment and purchasing behavior, including the following attributes:
•**Basic product information:** Product identifiers, categories, or other relevant details.
•**Rating:** Numerical score assigned by the reviewer to the product (e.g., 1-5 star rating).
•**Review text:** Unstructured text containing the reviewer's detailed feedback and opinions.
•My analysis primarily focused on the review.text column, which contains the actual customer reviews. To enrich the analysis, I've leveraged two additional columns as descriptive features: Name (Product name) and review.rating.

Details of the pre-processing steps:

Initial Data Inspection (Optional):
While not required for the program's functionality, I utilized Jupyter Notebook to perform an initial inspection of the data. This exploration provided a preliminary understanding of the data structure and content, aiding in the following pre-processing steps. I have included the note book as description of my though process.

The programmatic pre-processing included following steps:

Data Subset Creation:
To focus on the relevant information for analysis, I created a sub-DataFrame containing only the following three columns:
•Name (Product name)

•review.text (customer review text)
•review.rating

Text Preprocessing:

1.**Text Extraction:** A pandas Series was created containing only the review.text column. This Series served as the basis for the text preprocessing steps.
2.**NLP load:** The review.text text data was loaded line by line into spaCy NLP model en_core_web_sm
3.**Lowercasing:** All characters were converted to lowercase for consistency in text processing.
4.**Whitespace Removal:** Any leading, trailing, or excessive whitespace characters were eliminated to improve text cleaning.
5.**Lemmatization:** Words were converted to their base forms to capture the core meaning and reduce variations.
6.**Stop Word Removal:** Common words with a minimal semantic meaning were removed from the text.
7.**Punctuation Removal:** Punctuation marks were removed.

Evaluation of results:

Sentiment Analysis Case Study: Amazon Kindle E-Reader (8th Generation)

While my program is capable of analyzing more than the required two reviews, I've chosen to focus on the initial entries for a detailed exploration. These reviews are for the same product: the Amazon Kindle E-Reader 6" Wifi (8th Generation, 2016).

Review 1: Negative Sentiment:

The sentiment analysis program classified the first review as negative, with a polarity score of -0.016666666666666663. Examining the review text, it appears the customer might have selected the wrong product and is expressing dissatisfaction.

Review 1 text:

“I thought it would be as big as small paper but turn out to be just like my palm. I think it is too small to read on it... not very comfortable as regular Kindle. Would definitely recommend a paperwhite instead.”

Potential Explanations for Low Score:
•TextBlob might have focused on the negative aspects (size, comfort) and disregarded the potentially neutral comparison with "small paperwhite"
•The overall structure of the review, despite mentioning a positive alternative (paperwhite), might have communicate a more negative tone.
•3 star rating do indicate that customer was not totally dissatisfied, however I would accept result from program .

Review 2: Positive sentiment:
The sentiment analysis program classified the second review as positive, with a polarity score of 0.2777777777777778 Examining the review text, it appears the customer is happy to use the product on the beach.
Possible Reasons for Positive Score:
•TextBlob might have focused on the positive keywords ("light," "easy to use," "beach") and the enthusiastic tone of the customer.
•The briefness of the review could have limited the ability to detect any negativity.
•However the 5 star rating confirms the prediction.

Insights into models strengths and limitations:

SpaCy:
•**Advantages:**
•Offers better NLP pipeline with features like tokenization, named entity recognition, and dependency parsing. This allows for deeper text analysis.
•**Disadvantages:**
•Requires more complex setup and code compared to TextBlob.
TextBlob:
•**Advantages:**
•Simpler and easier to use, requiring minimal setup beyond installation.
•Well-suited for basic sentiment analysis tasks.
•**Disadvantages:**
•Provides a less advanced approach to sentiment analysis compared to spaCy.
en_core_web_sm Language Model:
•Both spaCy and TextBlob can be used with the en_core_web_sm language model for sentiment analysis.

•This pre-trained model is a good starting point for English text analysis, but keep in mind its limited
•For more specialized tasks, exploring other spaCy language models might be better option.