**Customer Sentiment Analysis**

Murtuja Afshar

Debug-Tech Task 2

February 23, 2025

**Data Collection:**

This dataset provides a comprehensive analysis of customer reviews for various products on an e-commerce platform. It is an invaluable resource for market analysts, product managers, and researchers interested in understanding consumer behavior and product performance. Each record in the dataset includes detailed feedback from customers, aiding in product evaluation and market research.

Key Features:

Reviewer ID: Unique identifier for the reviewer.

ASIN: Amazon Standard Identification Number for the product.

Reviewer Name: Name of the reviewer.

Helpful: Number of helpful votes the review received.

Review Text: The content of the review written by the customer.

Overall Rating: The overall rating given to the product (ranging from 1 to 5 stars).

Summary: A brief summary of the review.

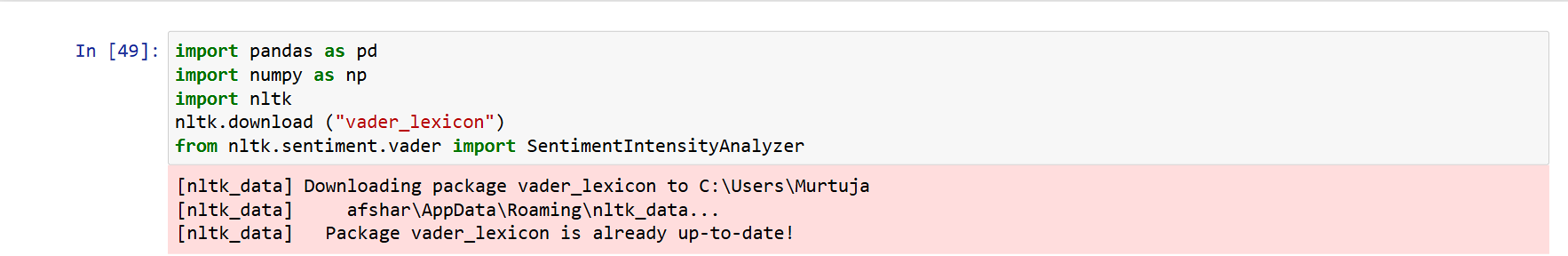
Unix Review Time: The time the review was posted in Unix timestamp format.

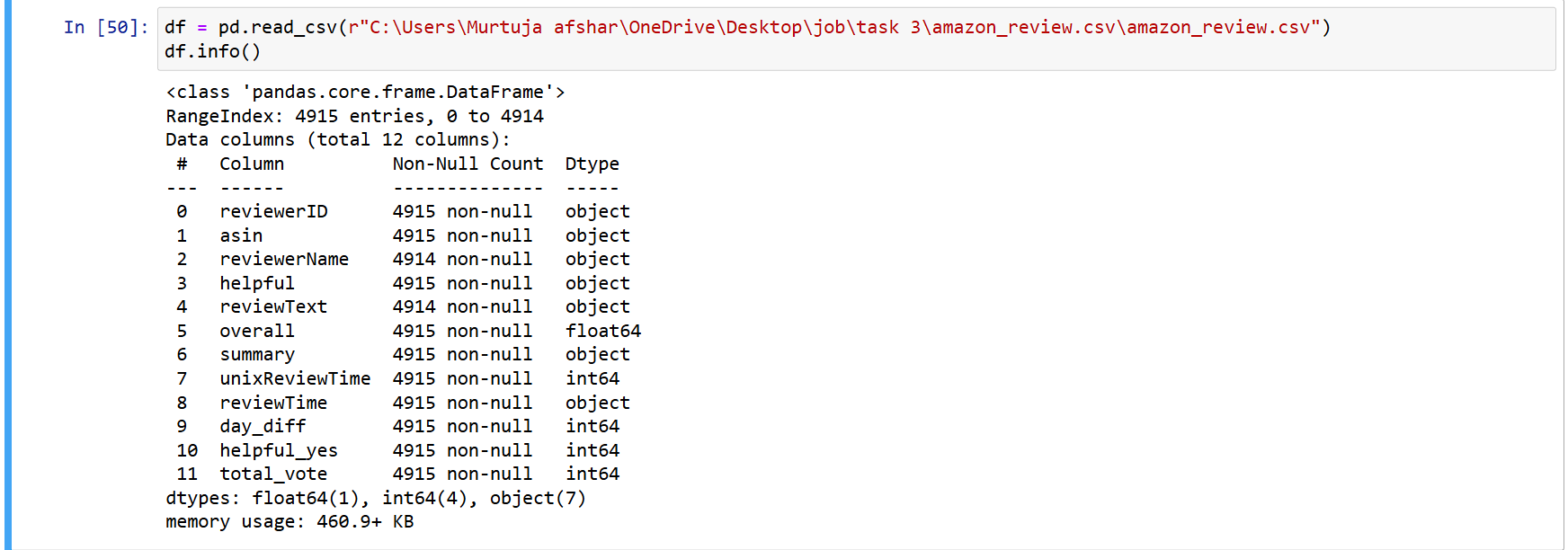
Review Time: The time the review was posted in a readable date format.

Day Difference: The number of days between the review date and the current date.

Helpful Yes: Number of positive helpful votes.

Total Votes: Total number of votes the review received.





**Data Cleaning:**

Data quality is often compromised by missing values, which can negatively impact model performance. The first step is to handle **NA (Not Available) values**. In text data, these often appear as blank entries in review columns. To clean the data, we typically “**Remove rows with missing values:”** This can be done using methods like dropna() in pandas, which drops rows containing any missing values in specified columns.

Converting all text to lowercase is a standard practice in text cleaning. This ensures that words like "Great" and "great" are treated as the same, preventing discrepancies due to case differences.

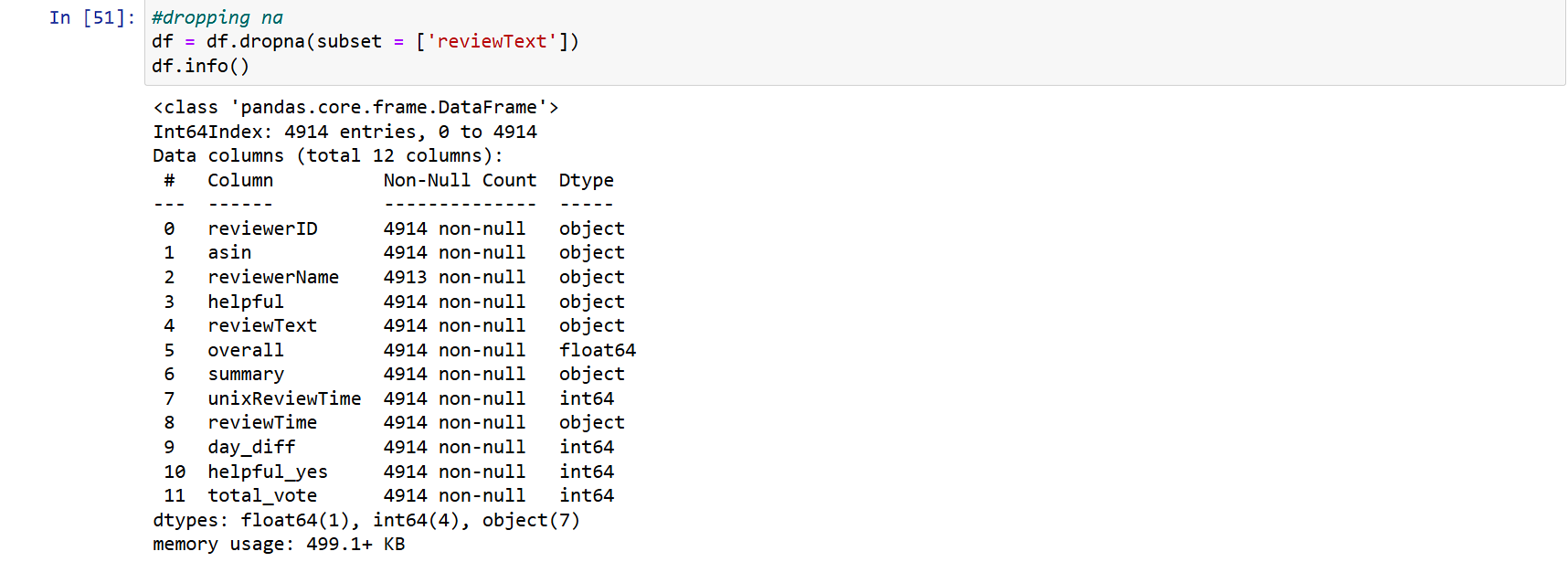
Special characters (like punctuation, symbols, and numbers) often don’t contribute to sentiment analysis or other NLP tasks and may interfere with model accuracy. Hence, we remove:

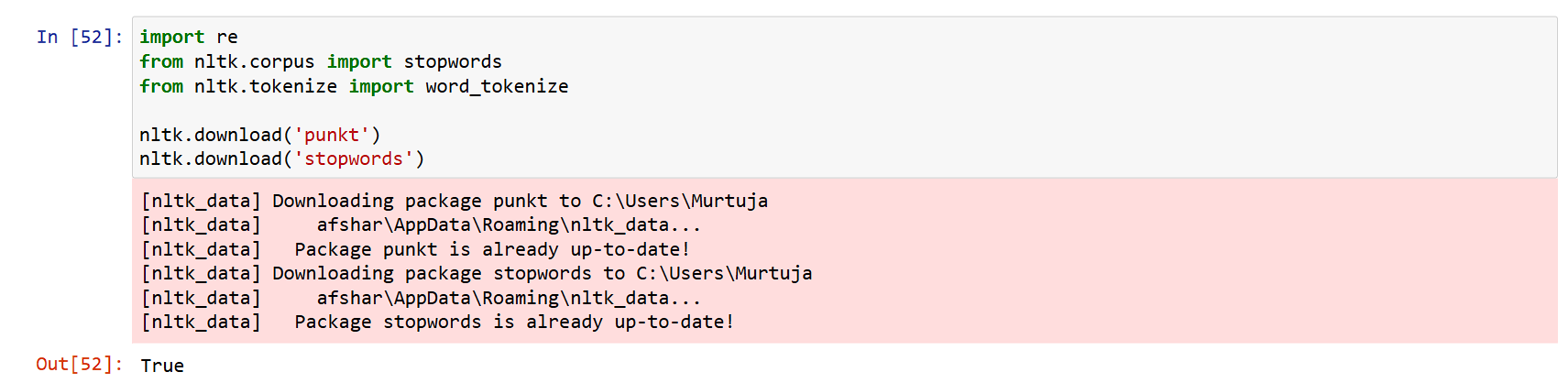
* **Punctuation:** Symbols like periods, commas, and exclamation marks.
* **Extra Spaces:** Redundant spaces in text can lead to inconsistent tokenization.

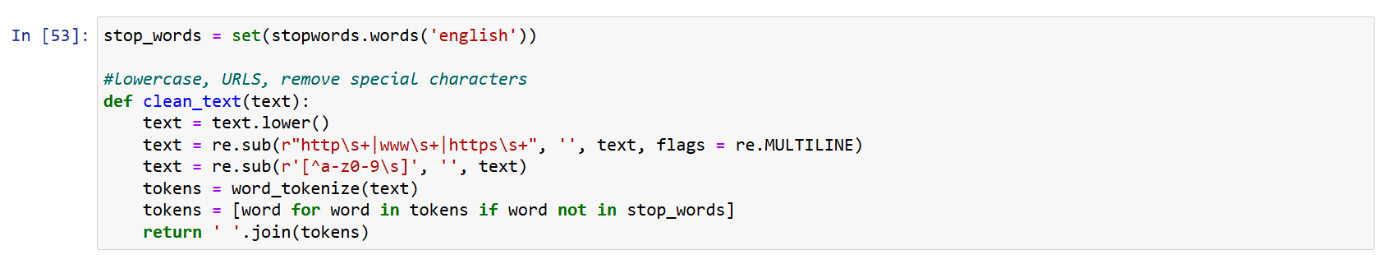
Tokenization is the process of splitting text into individual words or tokens. It’s crucial for many NLP applications as it breaks down the text into manageable units. We can use libraries like NLTK.

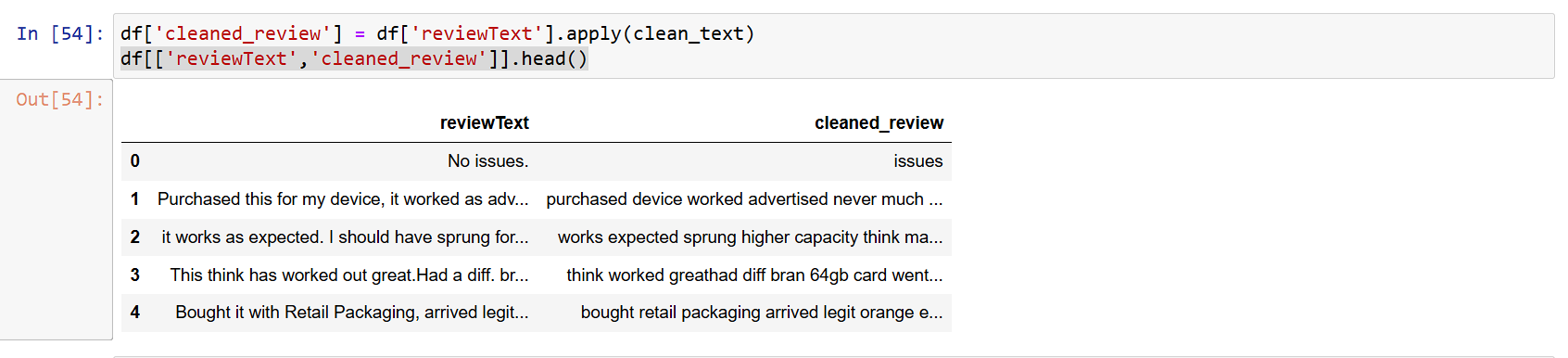
Stopwords are common words like “the”, “and”, “is”, etc., that don’t carry significant meaning in text analysis. Removing stopwords reduces the dimensionality of the dataset and focuses on meaningful words.

After these steps, the data is ready for analysis. The cleaned text will be uniform, tokenized, and free from irrelevant information, allowing for more accurate analysis in machine learning models or sentiment analysis.









**Sentiment Analysis:**

After cleaning the text, the VADER (Valence Aware Dictionary and sEntiment Reasoner) tool from the NLTK library was used to perform sentiment analysis. VADER is well-suited for short text like product reviews as it can handle informal language, emoticons, and punctuation marks effectively. It works by scoring the sentiment in a review using a compound score, which indicates whether the review is positive, neutral, or negative.

VADER outputs a compound score that ranges from -1 (most negative) to +1 (most positive). The compound score is a normalized value that aggregates the scores for individual words in the text and adjusts for context, such as negations (e.g., "not good" is interpreted differently than "good").

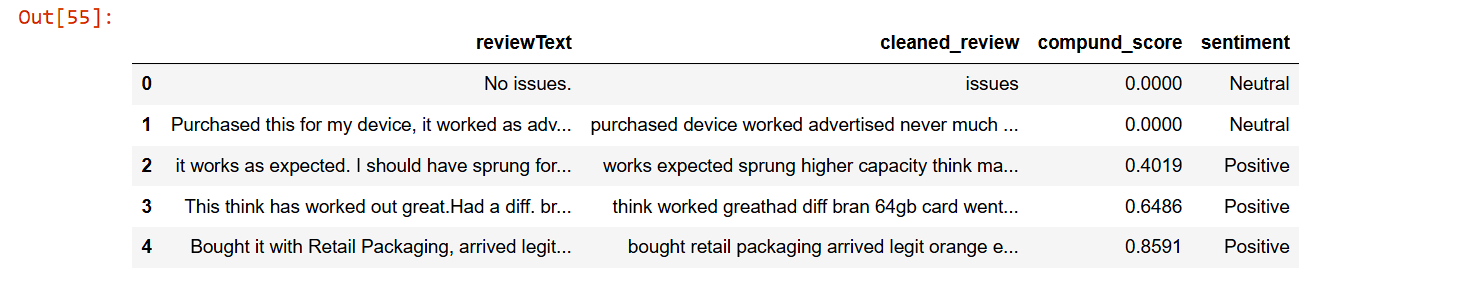
Once the compound score was calculated, each review was classified into one of three sentiment categories: Positive, Neutral, or Negative. The classification was based on the value of the compound score:

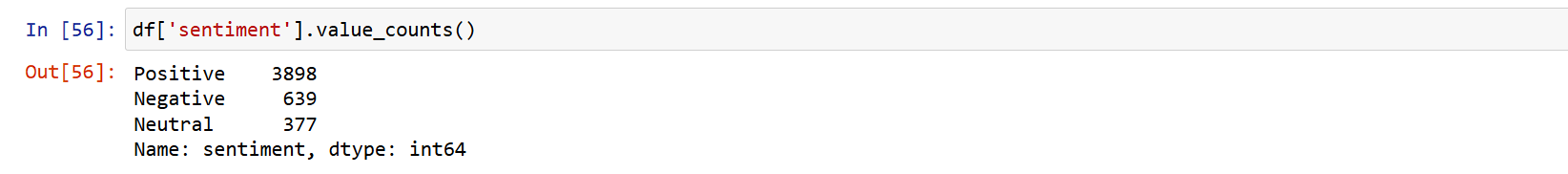
* Positive sentiment: If the compound score was greater than 0.05.
* Neutral sentiment: If the compound score was between -0.05 and 0.05.
* Negative sentiment: If the compound score was less than -0.05.

The final output of the sentiment analysis included the original review text, the cleaned version of the review, the compound score, and the corresponding sentiment classification. This process effectively enabled you to categorize each review into its sentiment type based on the overall emotional tone of the text.

In summary, sentiment analysis in this case involved cleaning the text, calculating the compound score using VADER, and classifying the reviews into sentiment categories based on that score. This process provides a way to quantify and categorize the emotions expressed in customer reviews.







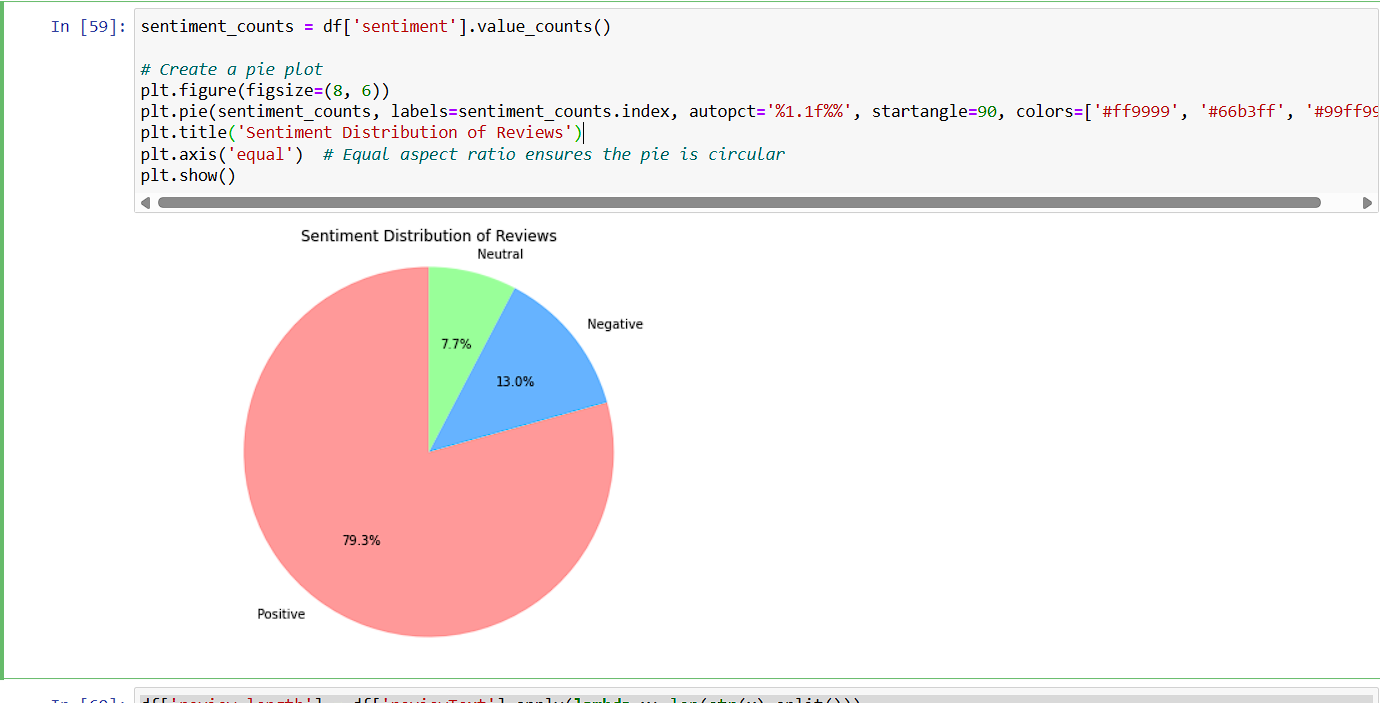
**Visualization:**

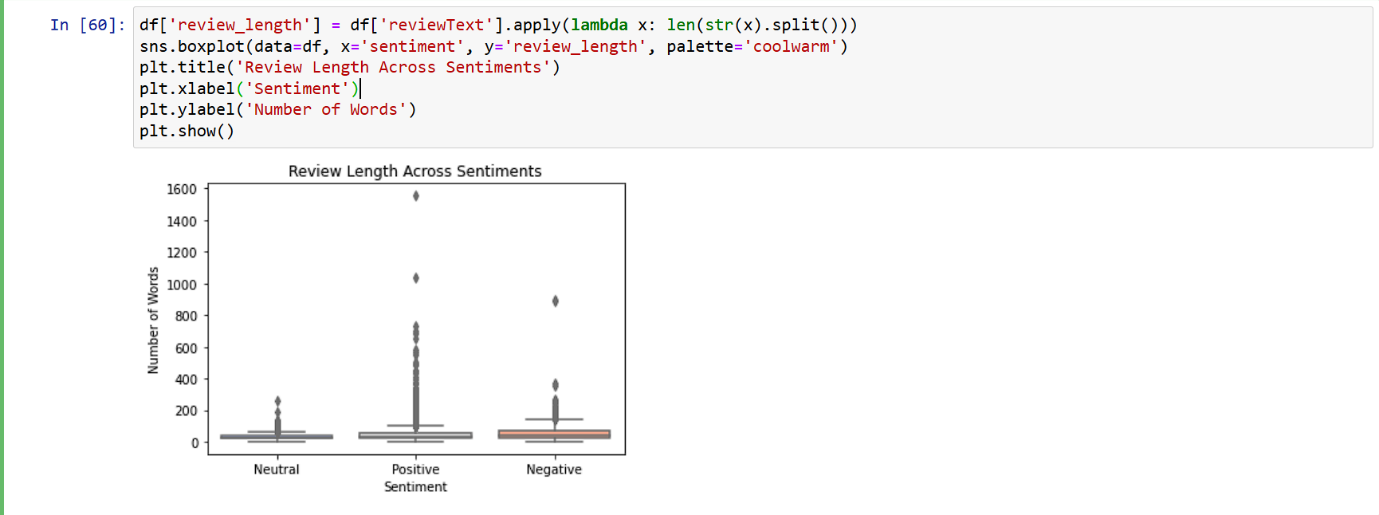
Three visualizations were carried out:  
“Sentiment Distribution (Count Plot): The first plot is a count plot that displays the distribution of sentiment categories (Positive, Neutral, Negative) in the dataset. It uses seaborn to create a simple bar chart, where each bar represents the count of reviews for each sentiment category. This visualization helps to quickly understand the balance or imbalance in sentiment, showing how many reviews fall under each category. The use of the pastel color palette gives a clear and aesthetically pleasant visualization of the sentiment data.

Sentiment Distribution (Pie Chart): The second plot is a pie chart that provides a more visual breakdown of sentiment distribution. It displays the proportion of Positive, Neutral, and Negative reviews as a percentage of the total. The autopct function is used to display percentages on the chart. The chart's colors differentiate the sentiment categories, offering a clearer representation of their relative proportions. The equal aspect ratio ensures the pie chart remains circular, making it visually balanced.

Review Length Across Sentiments (Box Plot): The third plot is a box plot that compares the review length (in terms of the number of words) across different sentiment categories. By calculating the review length and using seaborn's boxplot, the plot highlights the central tendency (median) and spread of review lengths. The coolwarm palette helps distinguish sentiment categories, and the box plot makes it easy to see if certain sentiments correlate with longer or shorter reviews.







**Conclusion:**

The sentiment analysis of reviews reveals a stark imbalance, with a dominant 79.3% classified as positive. This overwhelming positivity suggests strong customer satisfaction or a potential skew in the data towards positive reviews. Negative sentiment accounts for 13%, indicating a significant minority expressing dissatisfaction or concerns. Neutral sentiment comprises a small 7.7%, suggesting most reviews hold a clear opinion.

This distribution implies a generally favorable reception, but the substantial negative portion warrants attention. Businesses should investigate the root causes of negative reviews to address issues and maintain overall positive sentiment. Further analysis could explore the content of these reviews to identify specific areas of praise or criticism.