Data Appendix

Section 1: Movies_reviews_reduced.jsonl

Unit of Observation: Each row in the data frame represents a single review provided by an Amazon customer for a Movie or Television show that is found on their website.

- In total there are **154,639** Reviews in the reduced dataset

Subsection 1.1 *Rating*

- This column documents the customer's rating of a movie / television show on the scale of 1 to 5. These ratings are finite with a 1 representing a review where the customer is most dissatisfied and a 5 representing a review where a customer is most satisfied.
    - *Type:* Quantitative (Discrete Value)
    - Range: 1 to 5
    - Example: 2

Rating Statistics	**Value**
Mean	4.29
Median	5
Standard Deviation	1.24
Max	5
Min	1

Subsection 1.2 *Title*

- This column documents the title associated with a customer's review. Each review has only one title.
    - *Type*: Text Data
    - Example: "The Best Movie!"

Subsection 1.3 *Text*

- This column includes the body text associated with a review which is generally longer and more comprehensive than the review title text found above. This text for each review was passed through the VADER package to determine the sentiment score of each review as highlighted in Subsection 1.5 next.
    - *Type:* Text Data
    - Example: “I really enjoyed the movie. It had a great beginning, middle, and end which kept my family entertained for hours on a rainy day.”

Review Body Text Statistics	**Value**
Mean Character Length	240 Characters
Median Character Length	91 Characters
Standard Deviation of Character Length	572 Characters
Max Character Length	30,078 Characters
Min Character Length	1 Characters

Subsection 1.5 Sentiment_score

- This column was added to the dataframe during analysis and stores the sentiment score associated with each review determined by the VADER package. All sentiment scores exist on a spectrum between -1 and 1. A sentiment score of -1 represents the most negative sentiment, and a sentiment score of 1 represents the most positive sentiment. A score around 0 represents more neutral emotions within text.
    - *Type:* Quantitative (Continuous)
    - Range: -1 to 1
    - Example: 0.95

Sentiment Score Statistics	**Value**
Mean	0.48
Median	0.63

Standard Deviation	0.48
Max	1
Min	-1