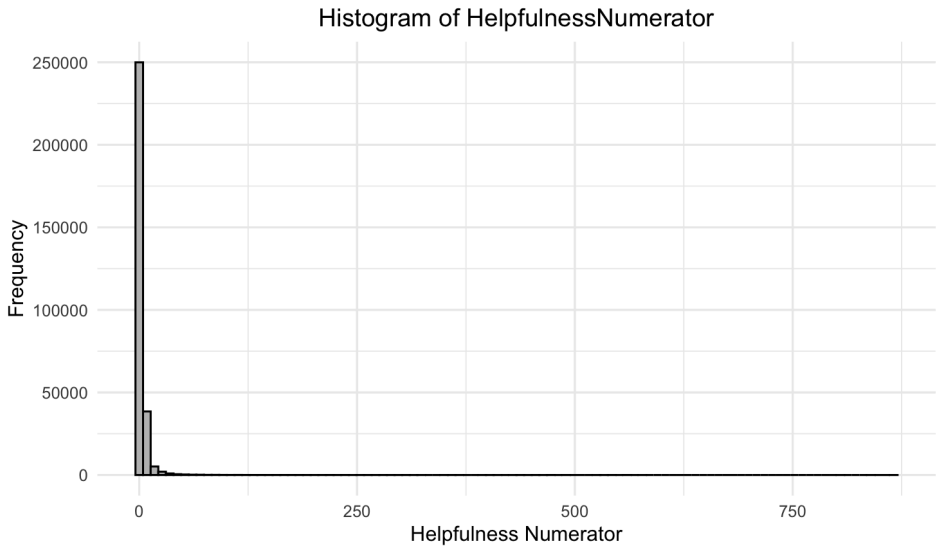
Data Appendix
Dataset: Amazon Fine Food Reviews

The data file contains eleven variables, with each row containing information about a different review. The unit of observation is the Amazon reviews.

Variables:

HelpfulnessNumerator

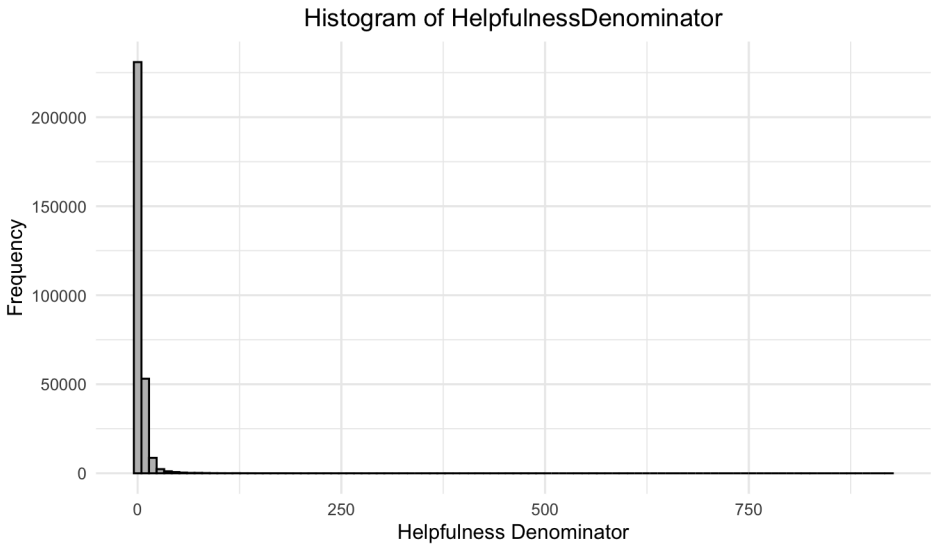
- Definition: Integer that indicates the amount of users who marked a review was helpful
- Units of Measurement: Review count
- Exact Name of Variable in original data file: HelpfulnessNumerator
- 568,452(0)
- Steps for processing the variable for cleaned dataset: No changes were made
- Minimum: 0.000
- 1st Quartile: 1.000
- Median: 1.000
- Mean: 3.322
- 3rd Quartile: 3.000
- Maximum: 866.000
- Standard Deviation: 10.28834



HelpfulnessDenominator

- Definition: Integer that indicates the amount of users who marked that a review was helpful
- Units of Measurement: Review count
- Exact Name of Variable in original data file: HelpfulnessDenominator
- 568,452(0)

- Steps for processing the variable for cleaned dataset:
    1. Deleted rows where the denominator was zero, meaning no one rated the review as helpful/not helpful
- Minimum: 1.000
- 1st Quartile: 1.000
- Median: 2.000
- Mean: 4.246
- 3rd Quartile: 4.000
- Maximum: 923.000
- Standard Deviation: 11.06105



Text

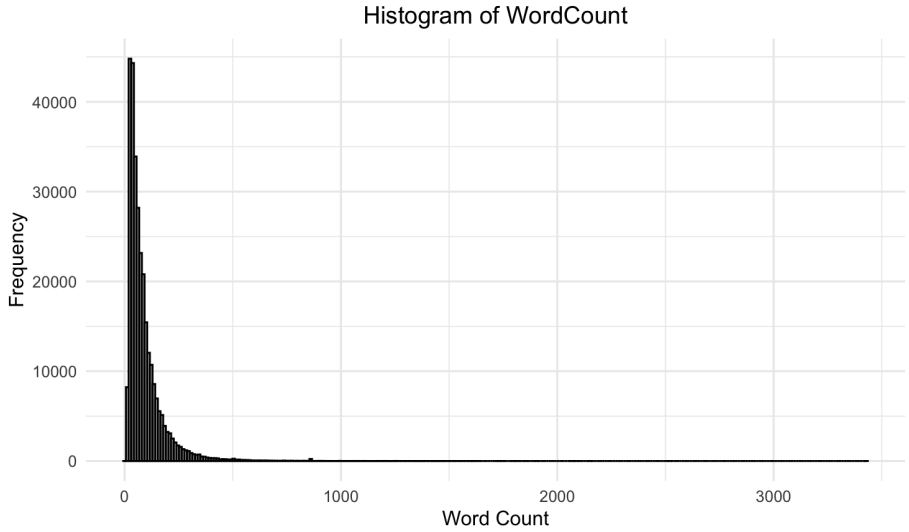
- Definition: The entire text of the review
- Units of Measurement: N/A
- Exact Name of Variable in original data file: Text
- 568,452(0)
- Steps for processing the variable for cleaned dataset:
    1. Remove text reviews which are duplicates of one another.
    2. Remove text reviews containing hyperlinks or irregular characters. Inclusion of these text variables could introduce negative helpfulness scores which skew the data

WordCount

- Definition: The word count of the review
- Units of Measurement: Number of words
- Exact Name of Variable in original data file: N/A
- 568,452(0)
- Steps for processing the variable for cleaned dataset:

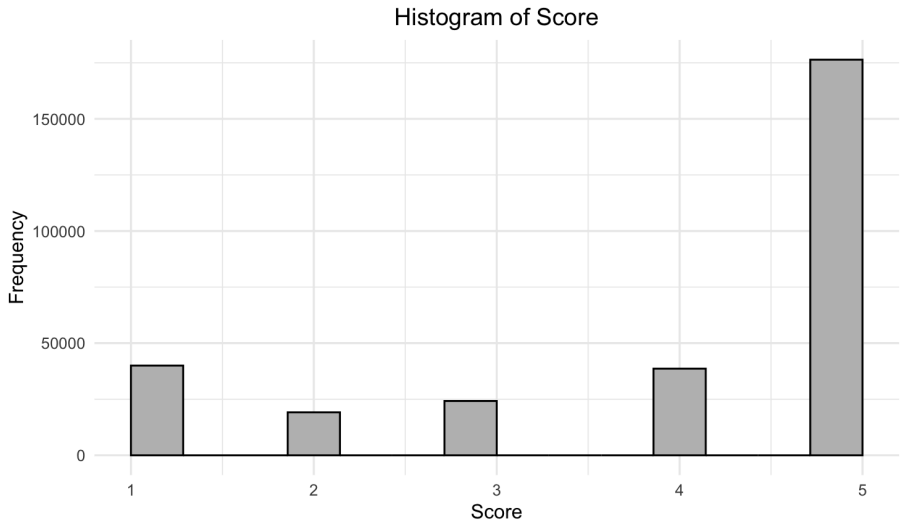
1. Using the review variable, use the “len” function in python to calculate the word count for each review

- Mean: 89.66
- Standard deviation:
- Minimum: 3.00
- 25th percentile: 36.00
- Median: 63.00
- 75th percentile: 109.0
- Maximum: 3432.0
- Standard Deviation: 91.3374



Score

- Definition: A rating one through five, one being the worst and five being the best, of the Amazon review’s rating of the product purchased
- Units of Measurement: Numerical value (1-5)
- Exact Name of Variable in original data file: Score
- 568,452(0)
- Steps for processing the variable for cleaned dataset: No changes were made.
- Mean: 3.979
- Minimum: 1.0
- 25th percentile: 3.0
- Median: 5.0
- 75th percentile: 5.0
- Maximum: 5.0
- Standard Deviation: 1.461233



Id

- Definition: Row Id
- Units of Measurement: N/A
- Exact Name of Variable in original data file: Id
- 568,452(0)
- Steps for processing the variable for cleaned dataset: No changes were made

ProductId

- Definition: 10-character string with a random combination of letters and numbers used to identify a product
- Units of Measurement: N/A
- Exact Name of Variable in original data file: ProductId
- 568,452(0)
- Steps for processing the variable for cleaned dataset: No changes were made

UserId

- Definition: 14-character string with a random combination of letters and numbers used to identify a user
- Units of Measurement: N/A
- Exact Name of Variable in original data file: UserId
- 568,452(0)
- Steps for processing the variable for cleaned dataset: No changes were made

ProfileName

- Definition: Username of the individual's profile in which they uploaded the review on; can be any length and may contain letters, numbers, spaces, or other special characters
- Units of Measurement: N/A
- Exact Name of Variable in original data file: ProfileName
- 568,452(0)
- Steps for processing the variable for cleaned dataset: No changes were made

Time

- Definition: Timestamp of when the review is posted
- Units of Measurement: N/A
- Exact Name of Variable in original data file: Time
- 568,452(0)
- Steps for processing the variable for cleaned dataset: No changes were made

Summary

- Definition: A short synopsis of the review contents; headline of each review
- Units of Measurement: N/A
- Exact Name of Variable in original data file: Summary
- 568,452(0)
- Steps for processing the variable for cleaned dataset: No changes were made.