

## Data Appendix

### Original Dataset

The unit of observation in this dataset is Amazon reviewers. Each row represents information about the reviewer as well as information about their review of a product.

#### *Reviewer Name*

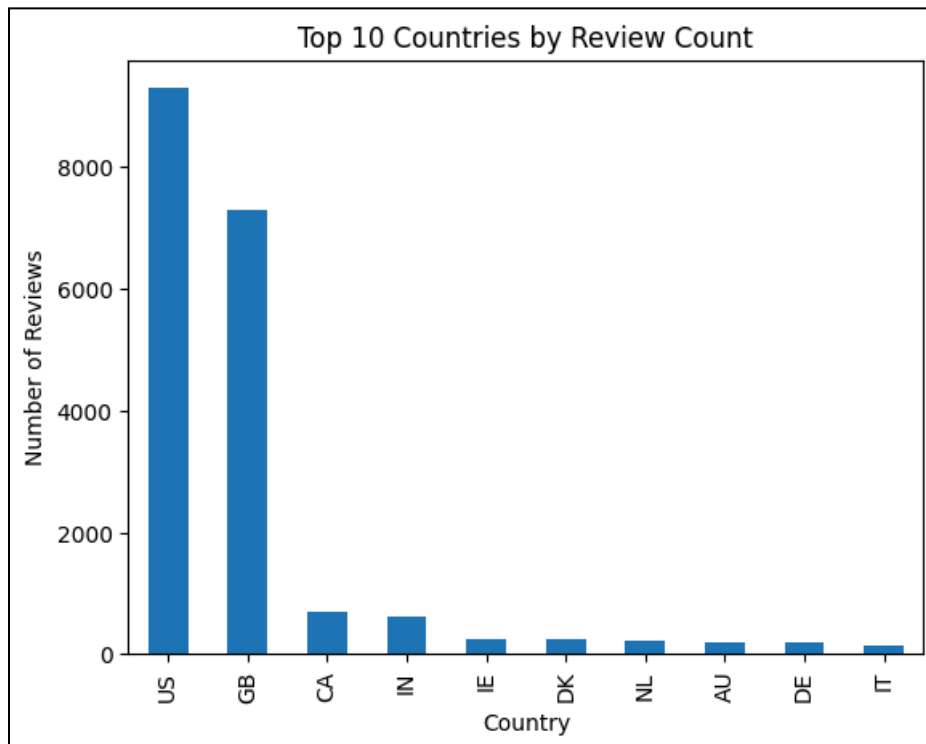
- The name or pseudonym of the reviewer
- No other variables were used to construct this variable
- 13914(0)

#### *Profile Link*

- A text link to the reviewer's Amazon profile for additional insights
- No other variables were used to construct this variable
- 13914(51)

#### *Country*

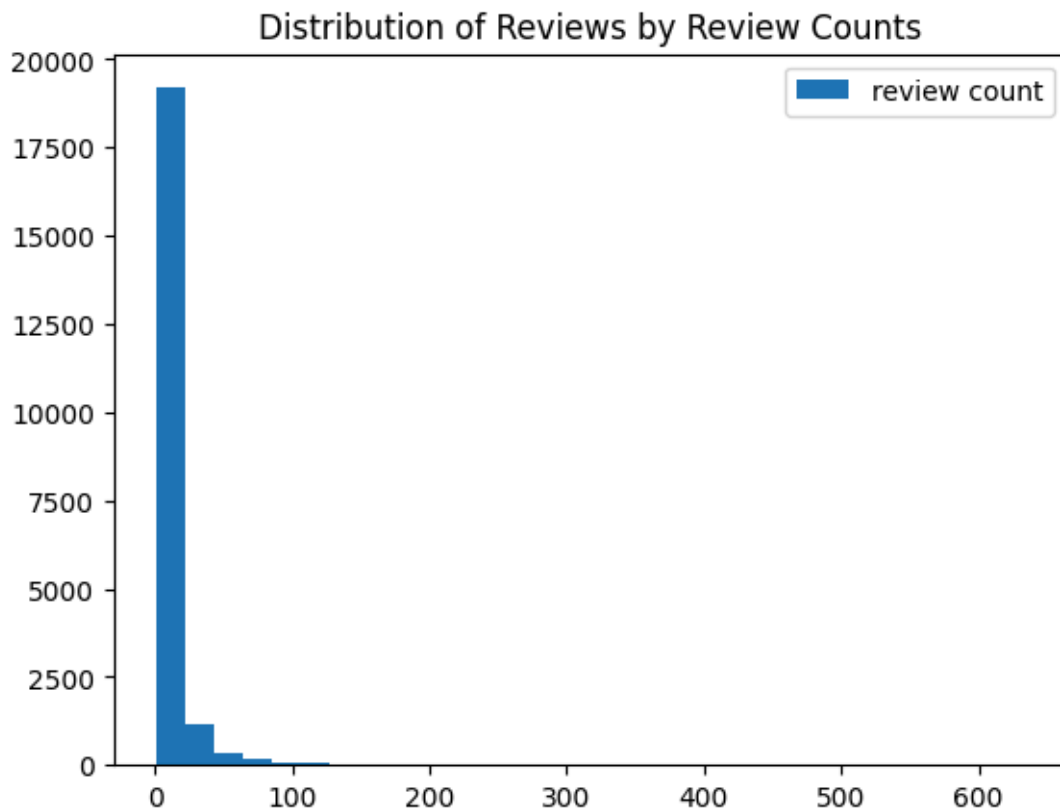
- The country of the reviewer's location, in text (e.g., "US")
- No other variables were used to construct this variable
- 13914(160)





### *Review Count*

- Number of reviews by the same user, written as solely a number
- This variable was derived from the original “review count” variable in the dataset, with the text removed
- 13914(159)
- Summary statistics:
  - Mean: 9.33
  - Standard deviation: 19.44
  - Minimum: 1
  - 25 percentile: 1
  - 50 percentile: 3
  - 75 percentile: 9
  - Max: 633



### *Review Date*

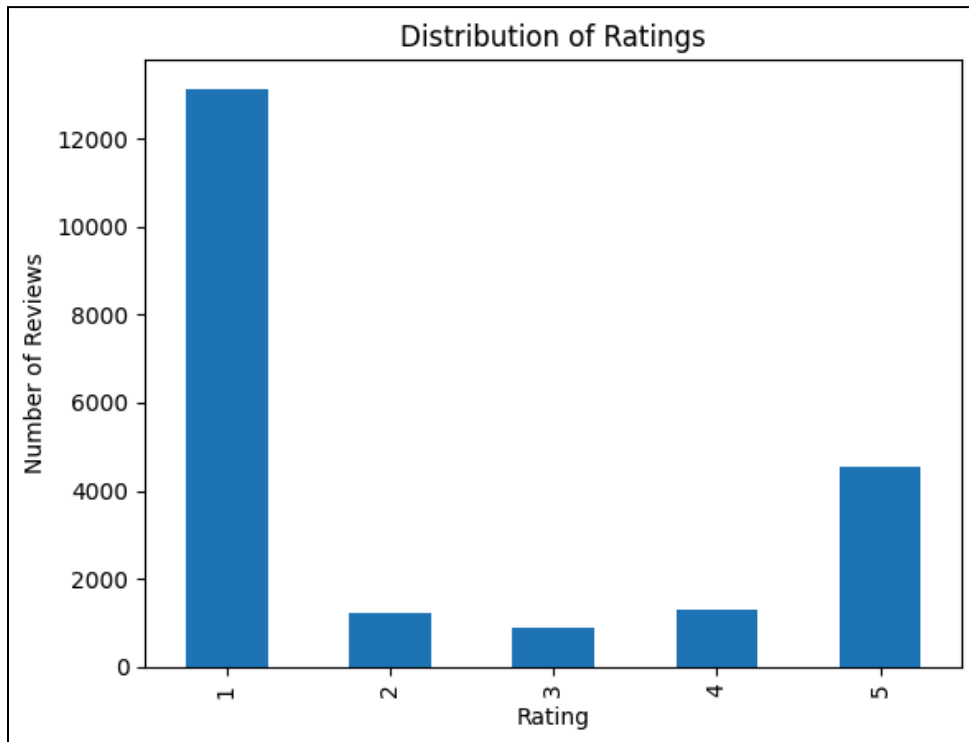
- Date (year/month/day) and time for when the review was posted
- No other variables were used to construct this variable
- 13914(159)

### *Rating*

- A numerical measure of satisfaction with the product or service, on a scale of 1 to 5
- No other variables were used to construct this variable



- 13914(159)
- Summary statistics:
  - Mean: 1.6
  - Standard deviation: 1.29
  - Minimum: 1
  - 25 percentile: 1
  - 50 percentile: 1
  - 75 percentile: 1
  - Max: 5



#### *Review Title*

- Text that aims to summarize the review sentiment
- No other variables were used to construct this variable
- 13914(159)

#### *Review Text*

- Detailed customer feedback of the product or service
- No other variables were used to construct this variable
- 13914(159)

#### *Date of Experience*

- Date of when the product or service was experienced, as text (e.g., "September 14th, 2024")
- No other variables were used to construct this variable
- 13914(267)

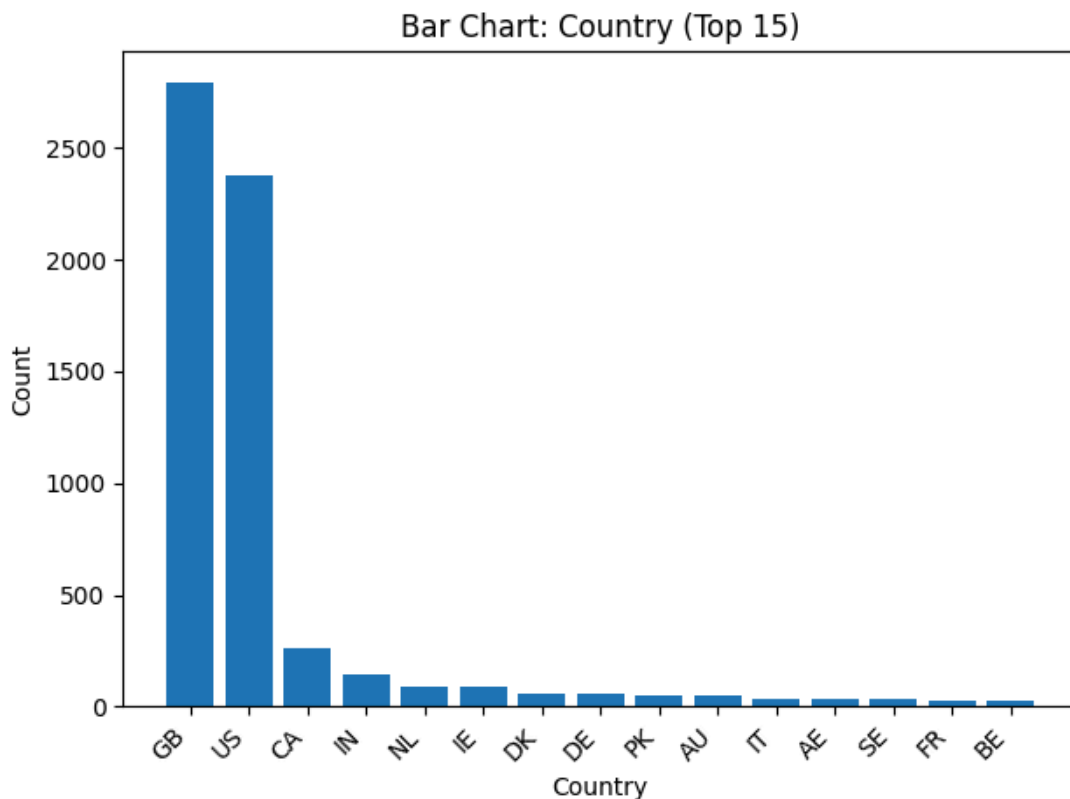


## Cleaned Dataset

The unit of observation in this dataset is Amazon reviewers. Each row represents information about the reviewer as well as information about their review of a product.

### Country

- The country of the reviewer's location
- No other variables were used in the creation of this variable, but this variable was filtered to only include countries in reviews from 2023 and 2024
- 6574(0)



### Review Title

- Text that aims to summarize the review sentiment
- No other variables were used to construct this variable, but this variable was filtered to only include reviews from 2023 and 2024
- 6574(0)

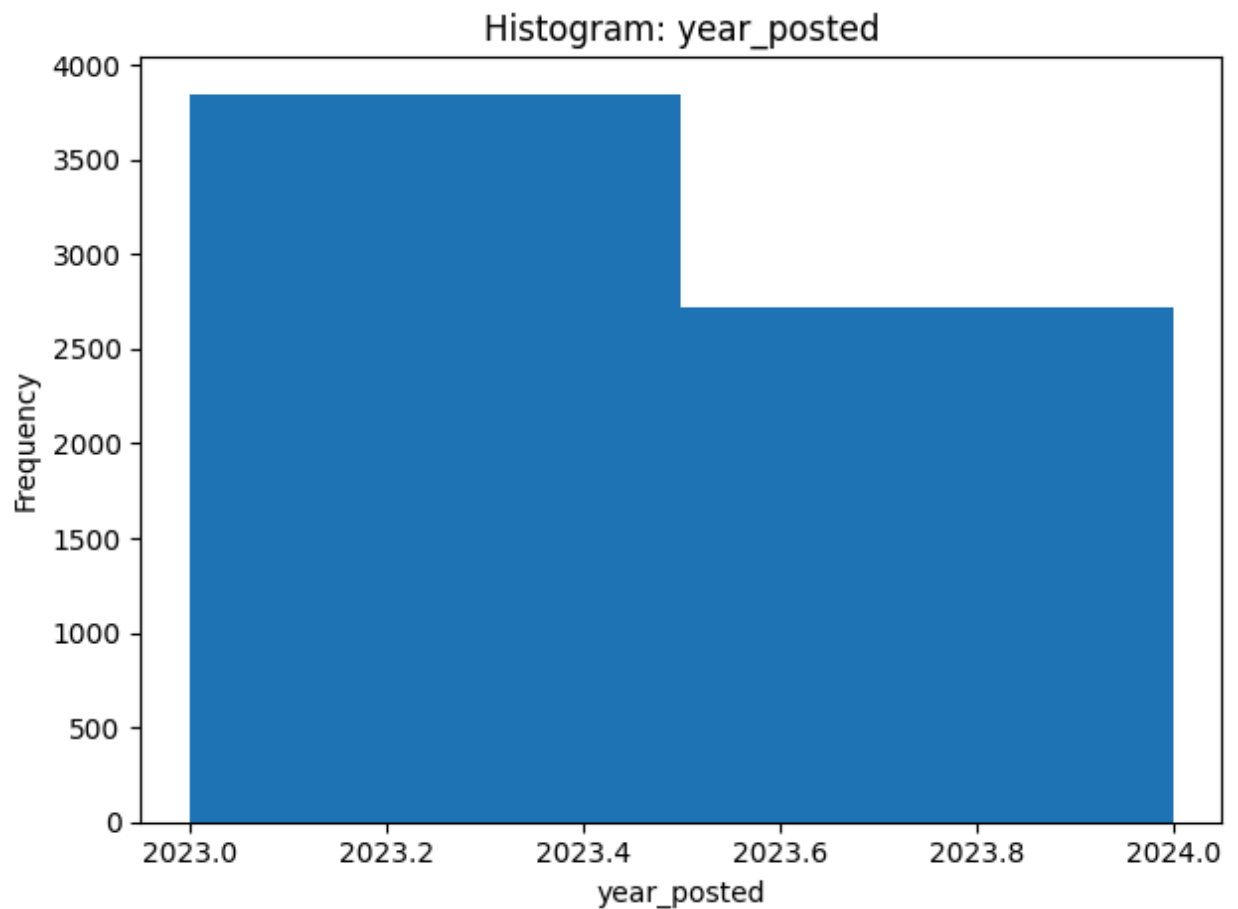
### Review Text

- Detailed customer feedback of the product or service
- No other variables were used to construct this variable, but this variable was filtered to only include reviews from 2023 and 2024.
- 6574(0)



### *Year\_posted*

- The year that the review was posted on Amazon
- This variable was extracted from the “Review Date” variable, and filtered to only include 2023 and 2024
- 6574(0)
- Summary statistics:
  - Mean: 2023.41
  - Std: 0.49
  - Min: 2023
  - 25 percentile: 2023
  - 50 percentile: 2023
  - 75 percentile: 2024
  - Max: 2024



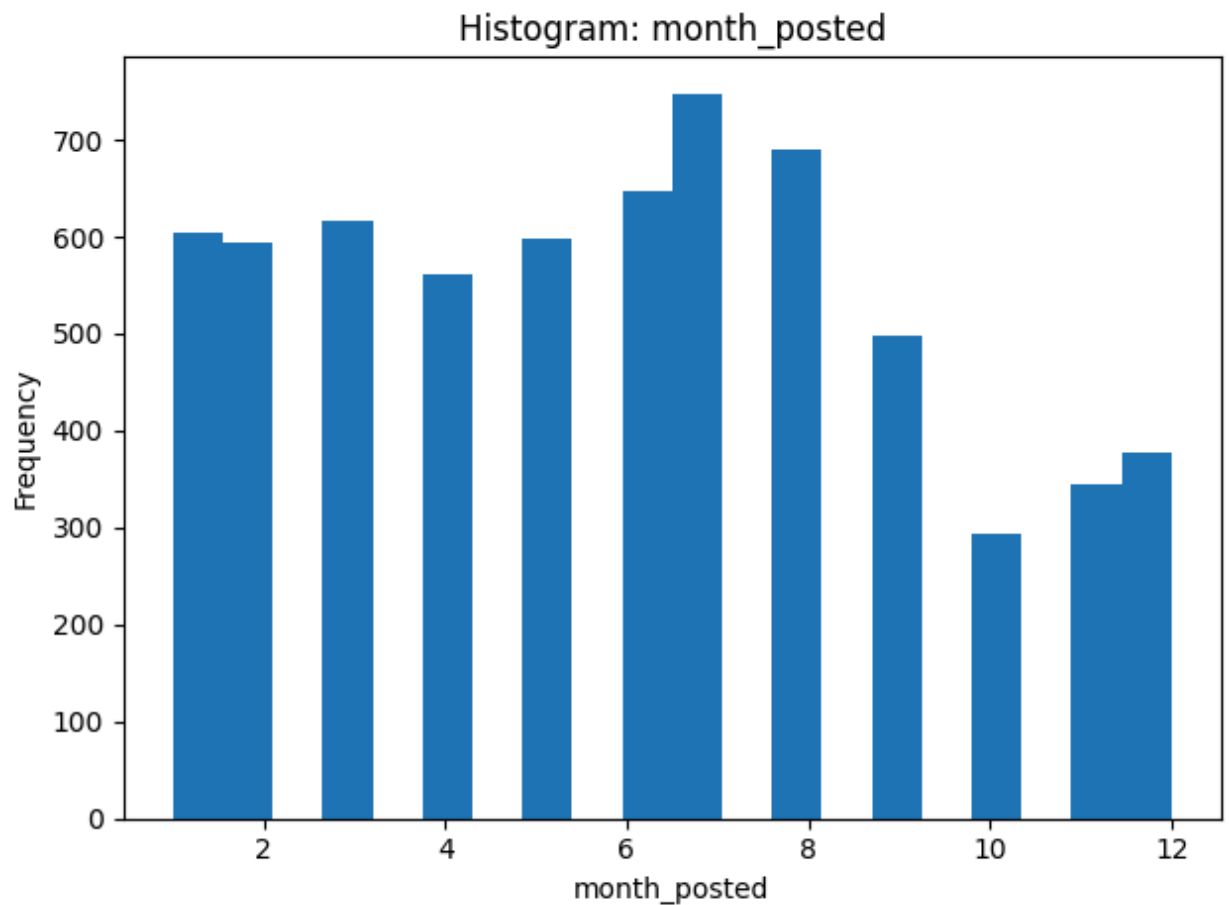
•

### *Month\_posted*

- The month that the review was posted on Amazon
- This variable was extracted from the “Review Date” variable, and filtered to only include 2023 and 2024
- 6574(0)
- Summary statistics:

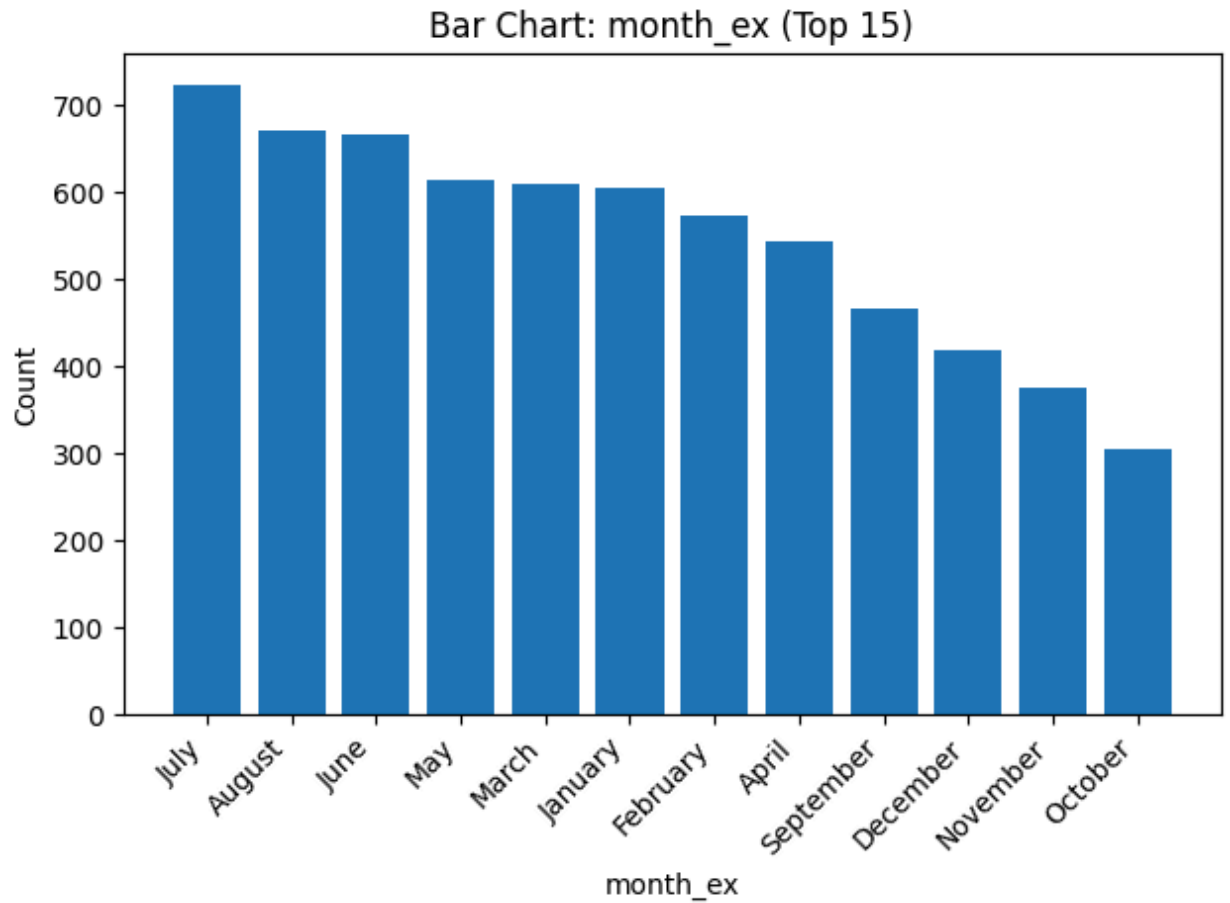


- Mean: 5.97
- Std: 3.21
- Min: 1
- 25 percentile: 3
- 50 percentile: 6
- 75 percentile: 8
- Max: 12



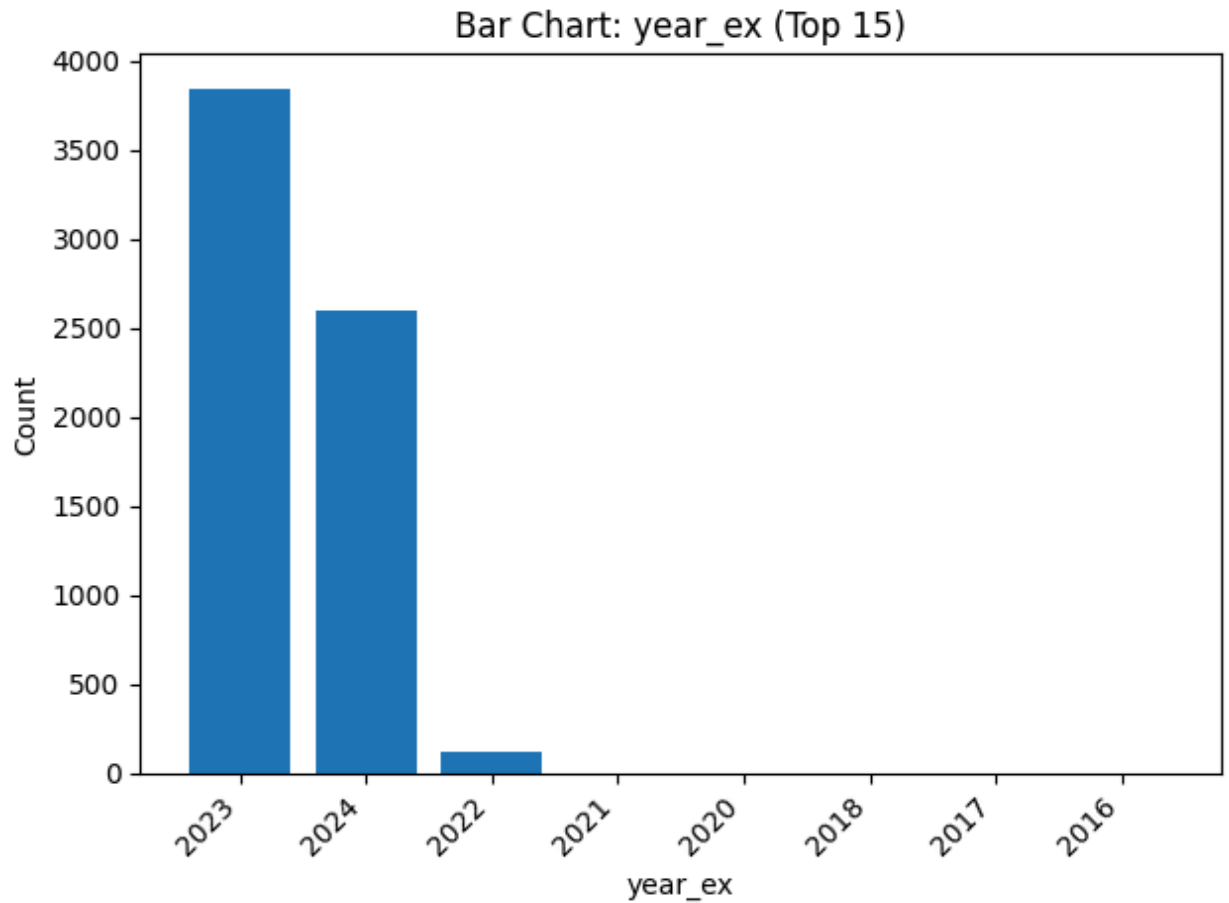
- 
- *Month\_ex*
  - Month that the product or service was experienced, as text
  - Derived from the “Date of Experience” variable and filtered to only include months from reviews posted in 2023 and 2024
  - 6574(0)
-





- 
- Year\_ex
  - Year that the product or service was experienced
  - Derived from the “Date of Experience” variable and filtered to only include years from reviews posted in 2023 and 2024
  - 6574(0)



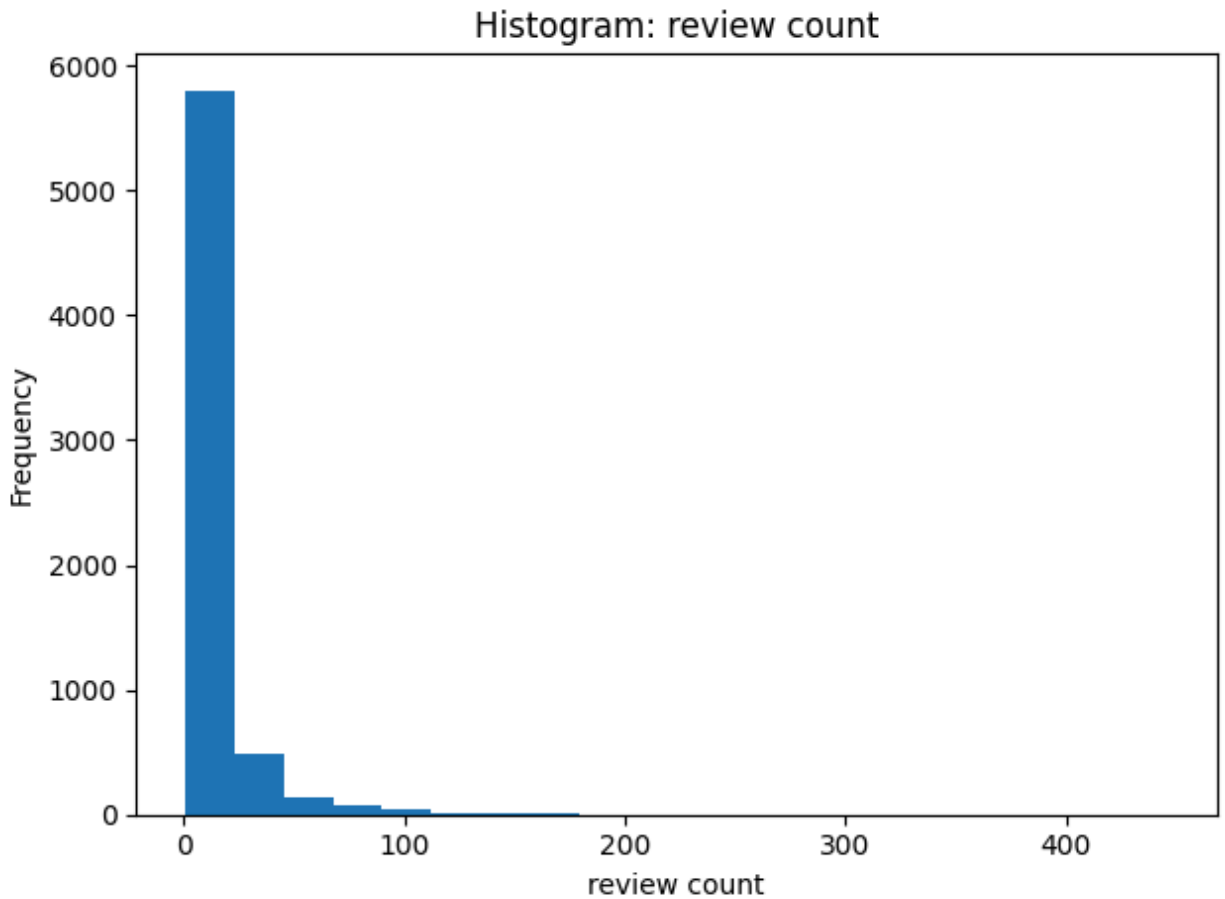


- 

*Review count*

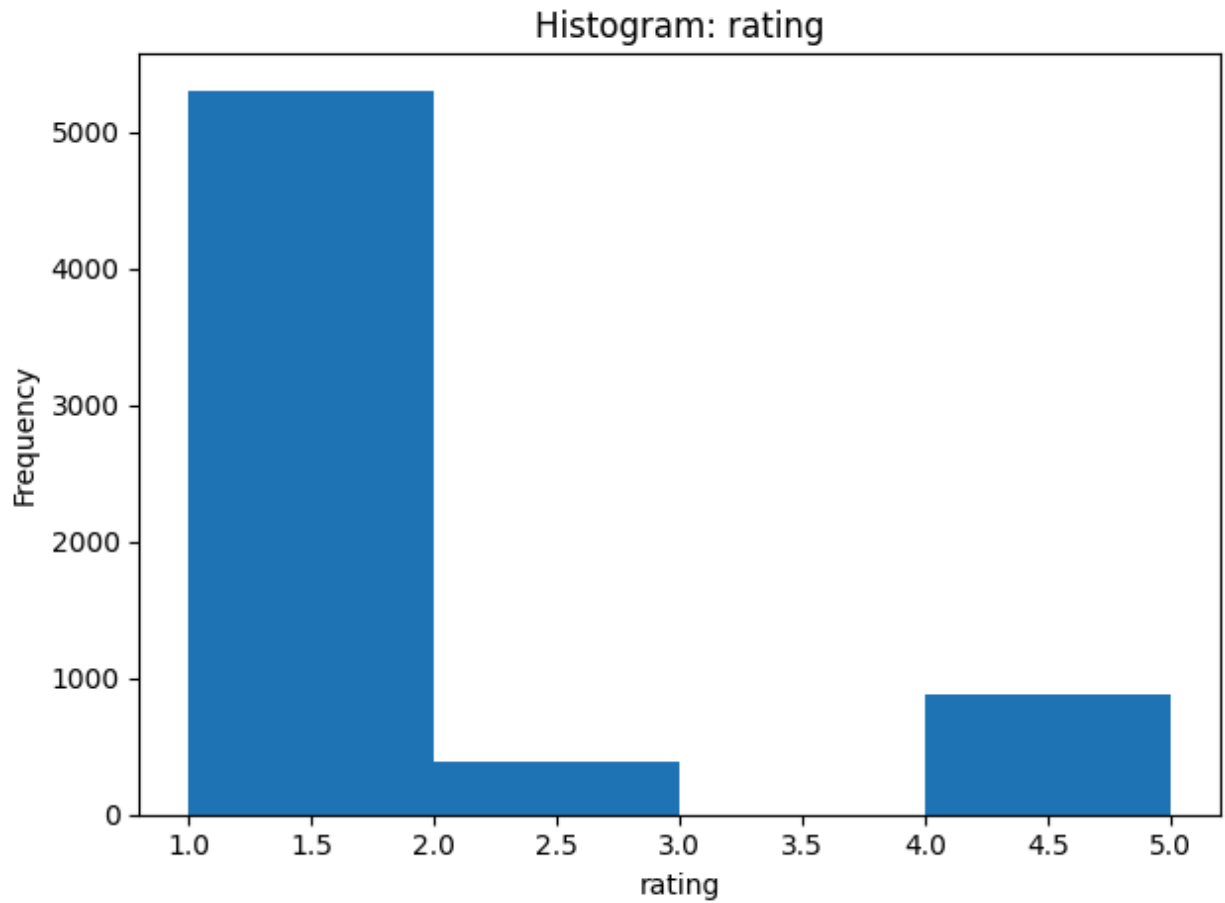
- Number of reviews by the same user, written as solely a number
- This variable was derived from the original “review count” variable in the dataset, with the text removed. Filtered to only include reviews posted in 2023 and 2024
- 6574(0)
- Summary statistics:
  - Mean: 10.6
  - Standard deviation: 20.9
  - Min: 1
  - 25 percentile: 2
  - 50 percentile: 4
  - 75 percentile: 11
  - Max: 446





- 
- *Rating*
  - A numerical measure of satisfaction with the product or service, on a scale of 1 to 5
  - No other variables were used to construct this variable. However, ratings of “3” were removed (for the binary classifier)
  - 6574(0)
  - Summary statistics:
    - Mean: 1.56
    - Standard deviation: 1.28
    - Min: 1
    - 25 percentile: 1
    - 50 percentile: 1
    - 75 percentile: 1
    - Max: 5

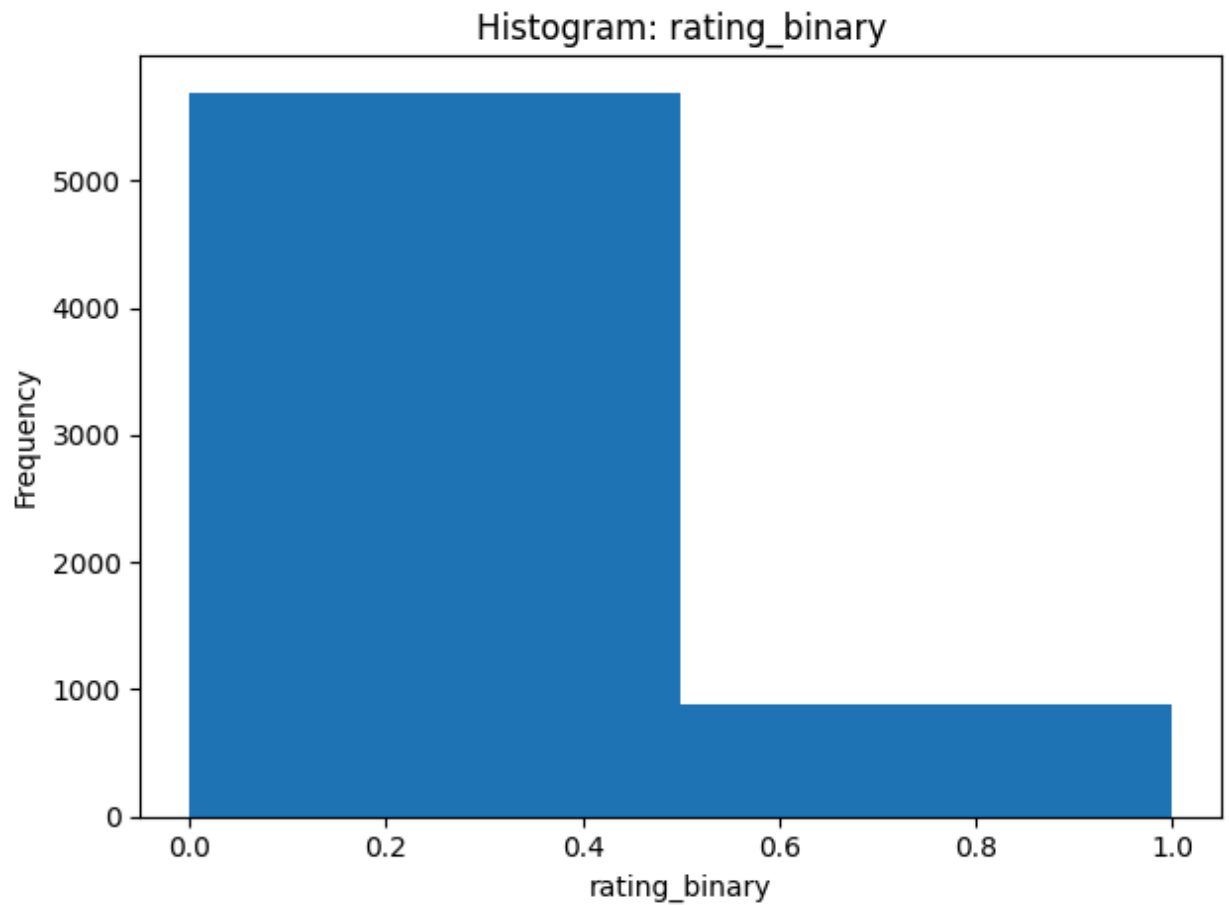




#### *Rating\_binary*

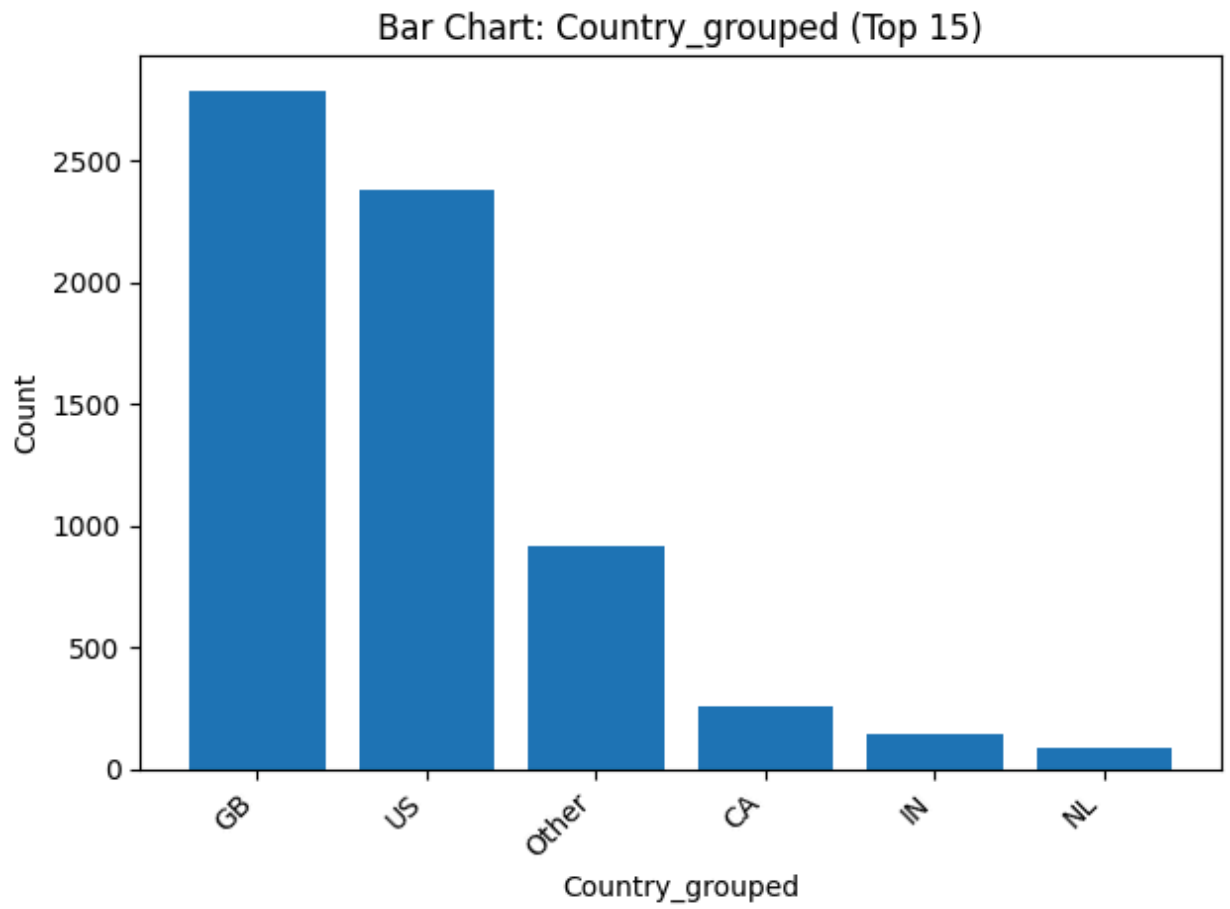
- A binary classification of rating as "0", meaning low (1-2) or "1", meaning high (4-5).
- Derived from the "rating" variable. Ratings of "3" were removed
- 6574(0)





- *Country\_grouped*
  - A classification of country as either the name of a country that has a top 5 most ratings in the dataset, or "other"
  - This is derived from the country variable
  - 6574(0)





•