DATA APPENDIX

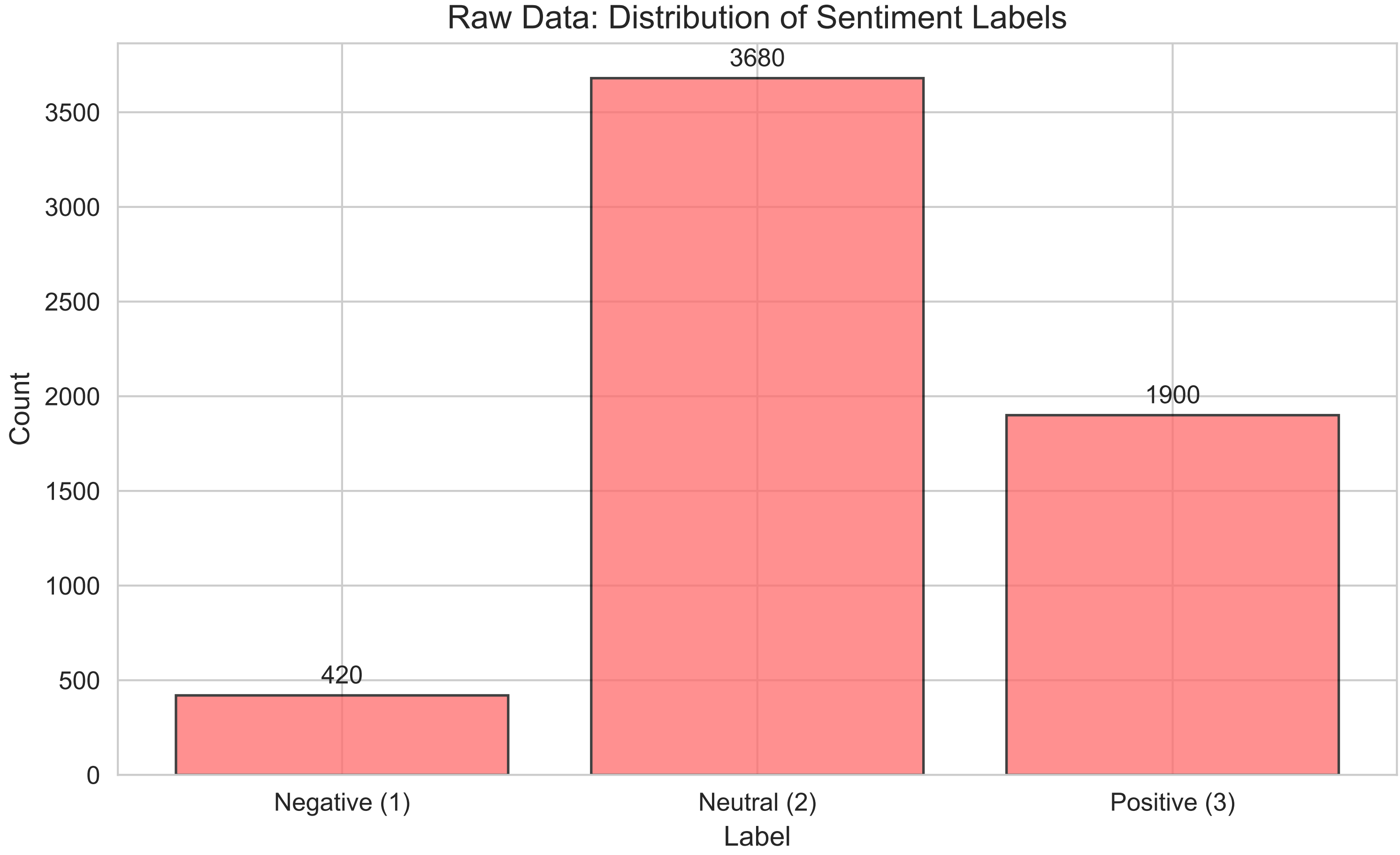
Project: Sentiment Shifts in Vaccine-Related Tweets by Theme
Group: Model Citizens
Members: Shaina Banduri, Neil Parikh, Nishana Dahal
Course: DS 4002
Date: Feb. 2026

This appendix documents every variable in the datasets used in this project, following the TIER Protocol 4.0.

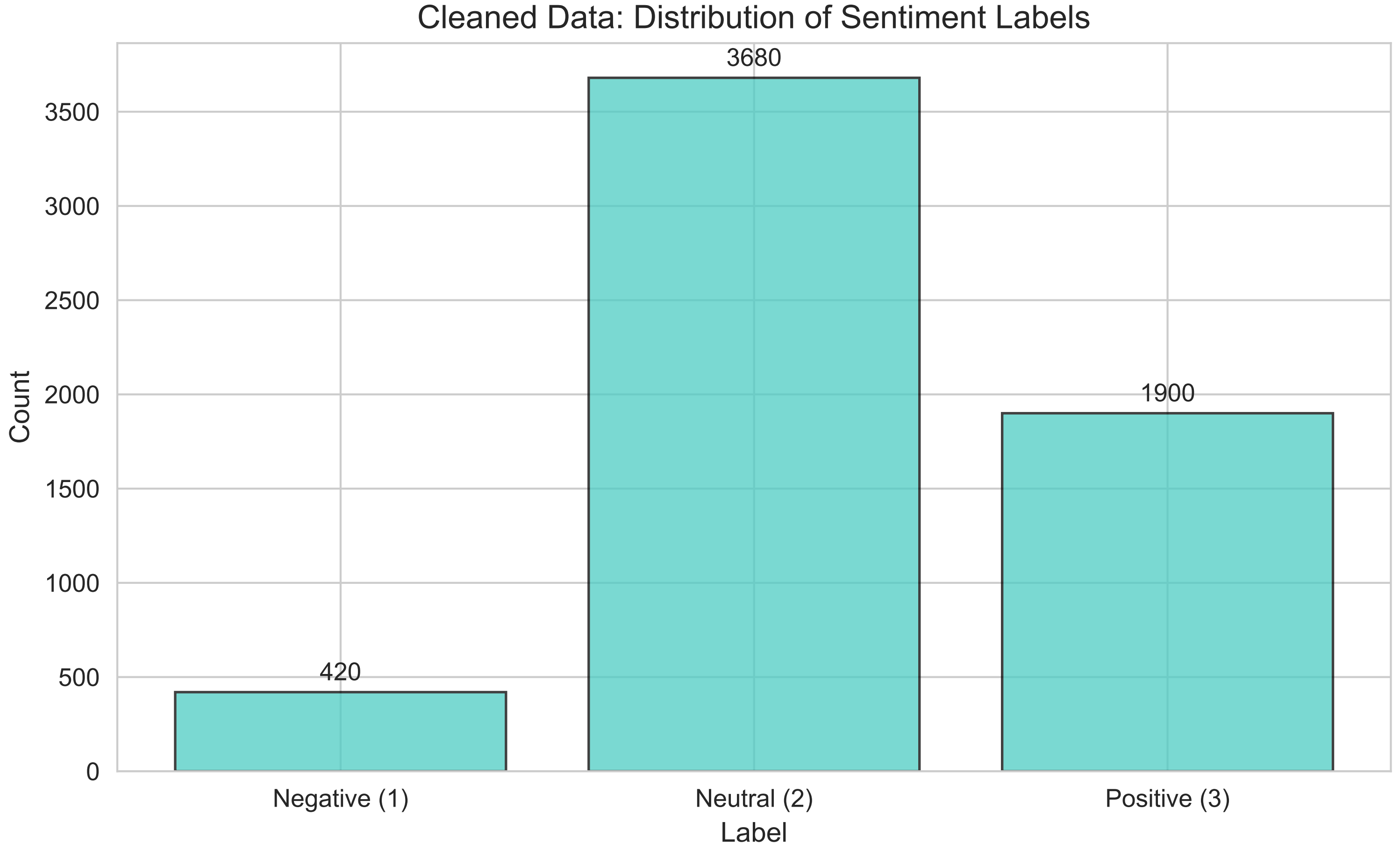
Datasets documented:

1. covid-19_vaccine_tweets_with_sentiment.csv (raw)
2. covid19_vaccine_tweets_cleaned.csv   (preprocessed)
3. covid19_vaccine_tweets_analyzed.csv   (final analysis)

```
============================================================
DATASET 1: Raw Data
File: covid-19_vaccine_tweets_with_sentiment.csv
Rows: 6,000   Columns: 3

Unit of observation: One tweet from Twitter related to
COVID-19 vaccines, with a human-annotated sentiment label.

--- Variable: tweet_id (float64) ---
  Unique numerical ID for each tweet.
  Source: Original Kaggle dataset.
  Observations: 6000   Missing: 0

--- Variable: label (int64) ---
  Human-annotated sentiment label.
  Values: 1 = Negative, 2 = Neutral, 3 = Positive.
  Source: Original Kaggle dataset (human annotators).
  Observations: 6000   Missing: 0
  Distribution:
    1 (Negative): 420
    2 (Neutral):  3680
    3 (Positive): 1900

--- Variable: tweet_text (string) ---
  Full text content of the tweet including hashtags,
  URLs, and @mentions.
  Source: Original Kaggle dataset.
  Observations: 6000   Missing: 0
```

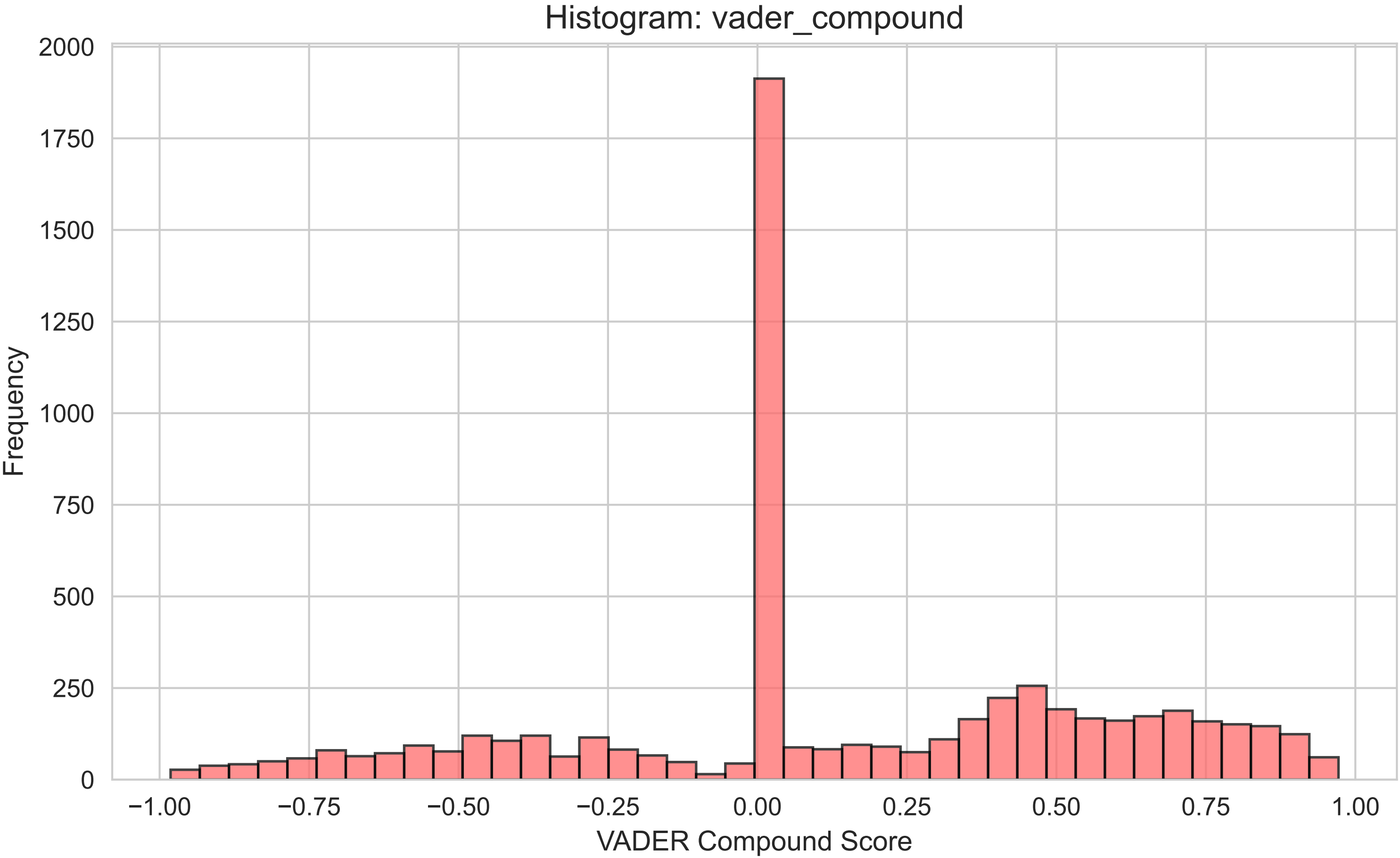

Raw Data: Distribution of Sentiment Labels
Count
3500
3000
2500
2000
1500
1000
500
0
420
3680
1900
Negative (1)
Neutral (2)
Positive (3)
Label

```
============================================================
DATASET 2: Cleaned Data
File: covid19_vaccine_tweets_cleaned.csv
Rows: 6,000   Columns: 3

Unit of observation: One cleaned tweet. Rows with missing
tweet_text were removed during preprocessing. Text was
lowercased, URLs removed, @mentions removed, # symbols
stripped (hashtag text preserved), whitespace normalized.

--- Variable: tweet_id (float64) ---
  Same as raw data. Unique tweet identifier.
  Observations: 6000   Missing: 0

--- Variable: label (int64) ---
  Same as raw data.
  Observations: 6000   Missing: 0
  Distribution:
    1 (Negative): 420 (7.0%)
    2 (Neutral):  3680 (61.3%)
    3 (Positive): 1900 (31.7%)

--- Variable: tweet_text (string) ---
  Cleaned tweet text. All lowercase, no URLs, no @mentions,
  no # symbols (hashtag text preserved), single-spaced.
  Observations: 6000   Missing: 0
```

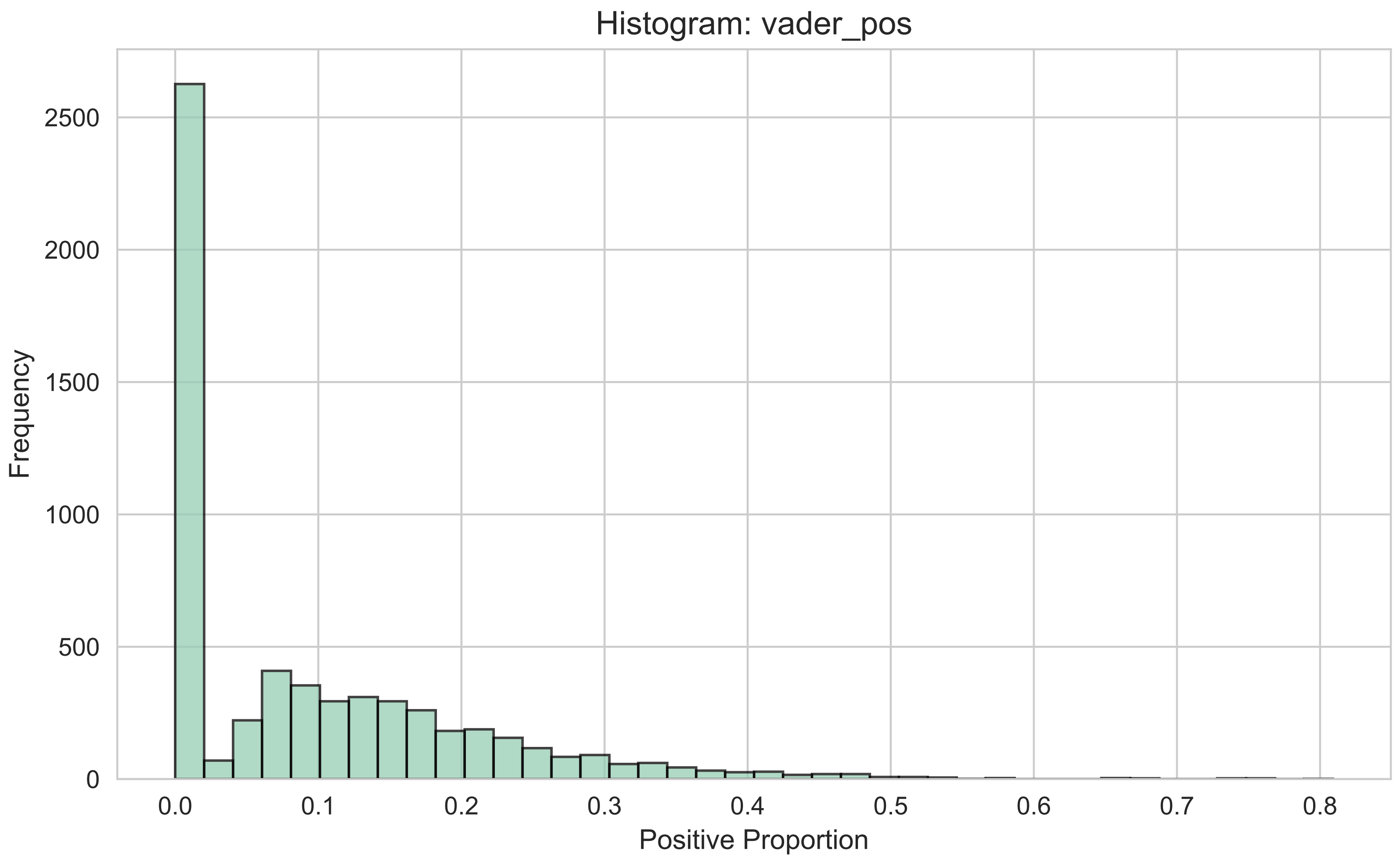

Cleaned Data: Distribution of Sentiment Labels
Count
0
500
1000
1500
2000
2500
3000
3500
420
3680
1900
Negative (1)
Neutral (2)
Positive (3)
Label

```
============================================================
DATASET 3: Analyzed Data
File: covid19_vaccine_tweets_analyzed.csv
Rows: 6,000   Columns: 13

Unit of observation: One cleaned tweet enriched with VADER
sentiment scores and theme/brand indicator flags computed
by analysis.py.

Original variables (same as cleaned data):
  tweet_id, label, tweet_text -- see Dataset 2 above.

The following pages document each added variable with
summary statistics and visualizations.
```

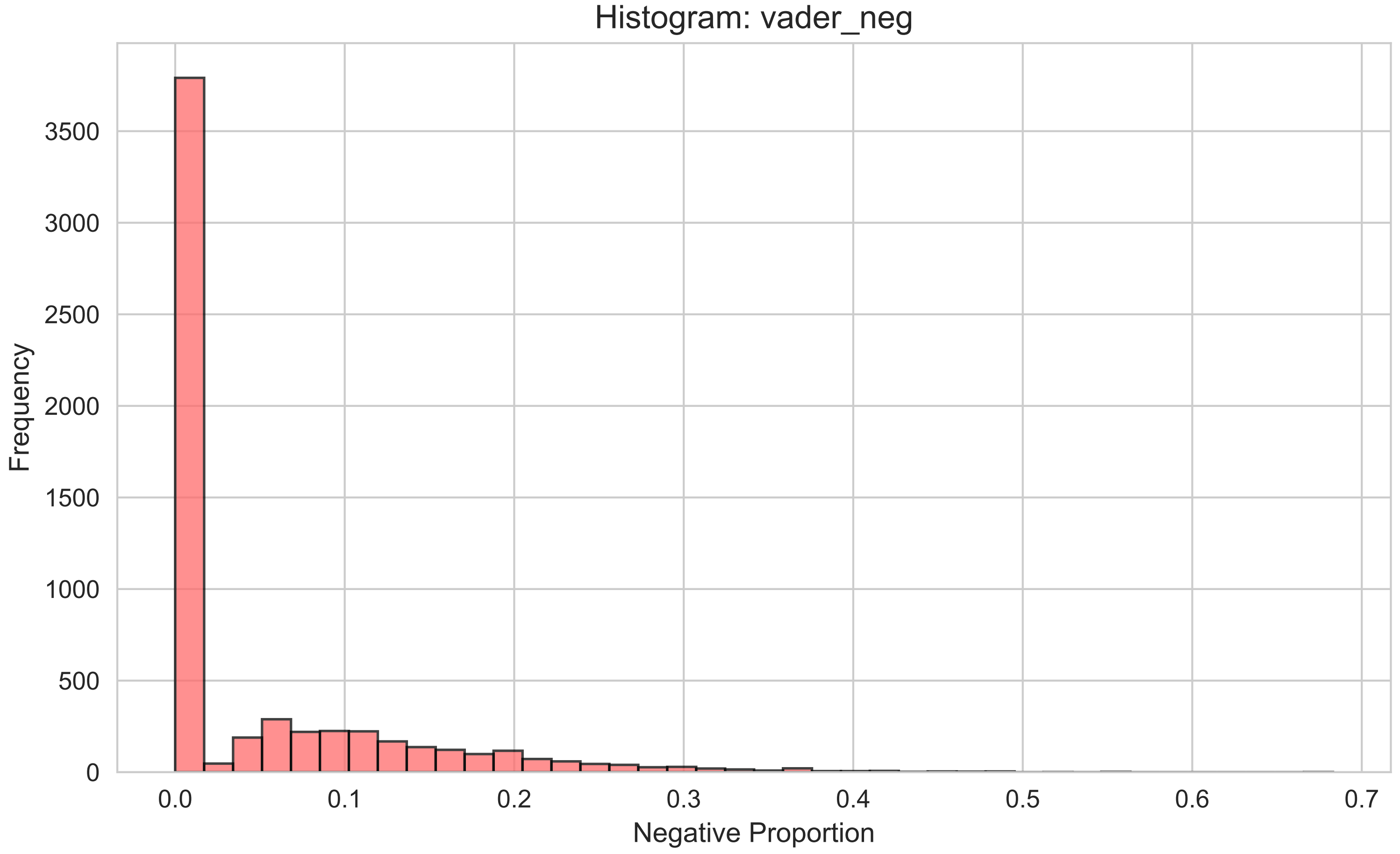
```
--- Variable: vader_compound (float64) ---
  VADER normalised compound sentiment score.
  Range: [-1, 1]. -1 = most negative, +1 = most positive.
  Source: Computed from tweet_text using vaderSentiment.

    Observations: 6000   Missing: 0
    Mean:   0.1313
    Std:    0.4457
    Min:    -0.9816
    25th:   0.0000
    Median: 0.0000
    75th:   0.4926
    Max:    0.9718
```


Histogram: vader_compound
Frequency
0
250
500
750
1000
1250
1500
1750
2000
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
VADER Compound Score

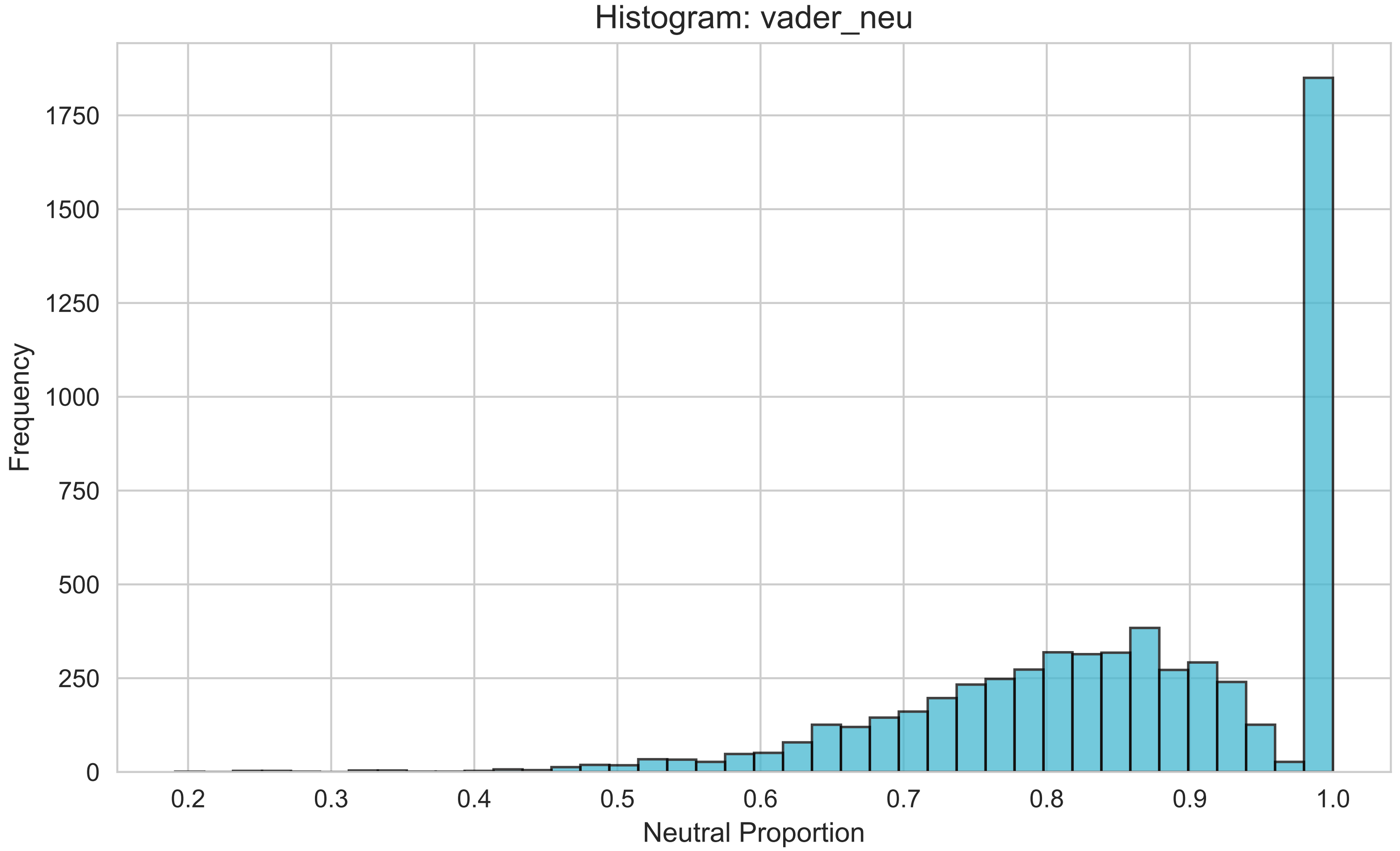
```
--- Variable: vader_pos (float64) ---
  Proportion of text tokens with positive sentiment [0, 1].
  Source: Computed from tweet_text using vaderSentiment.

    Observations: 6000   Missing: 0
    Mean:   0.0941
    Std:    0.1142
    Min:    0.0000
    25th:   0.0000
    Median: 0.0660
    75th:   0.1560
    Max:    0.8090
```


Histogram: vader_pos
Frequency
0
500
1000
1500
2000
2500
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Positive Proportion

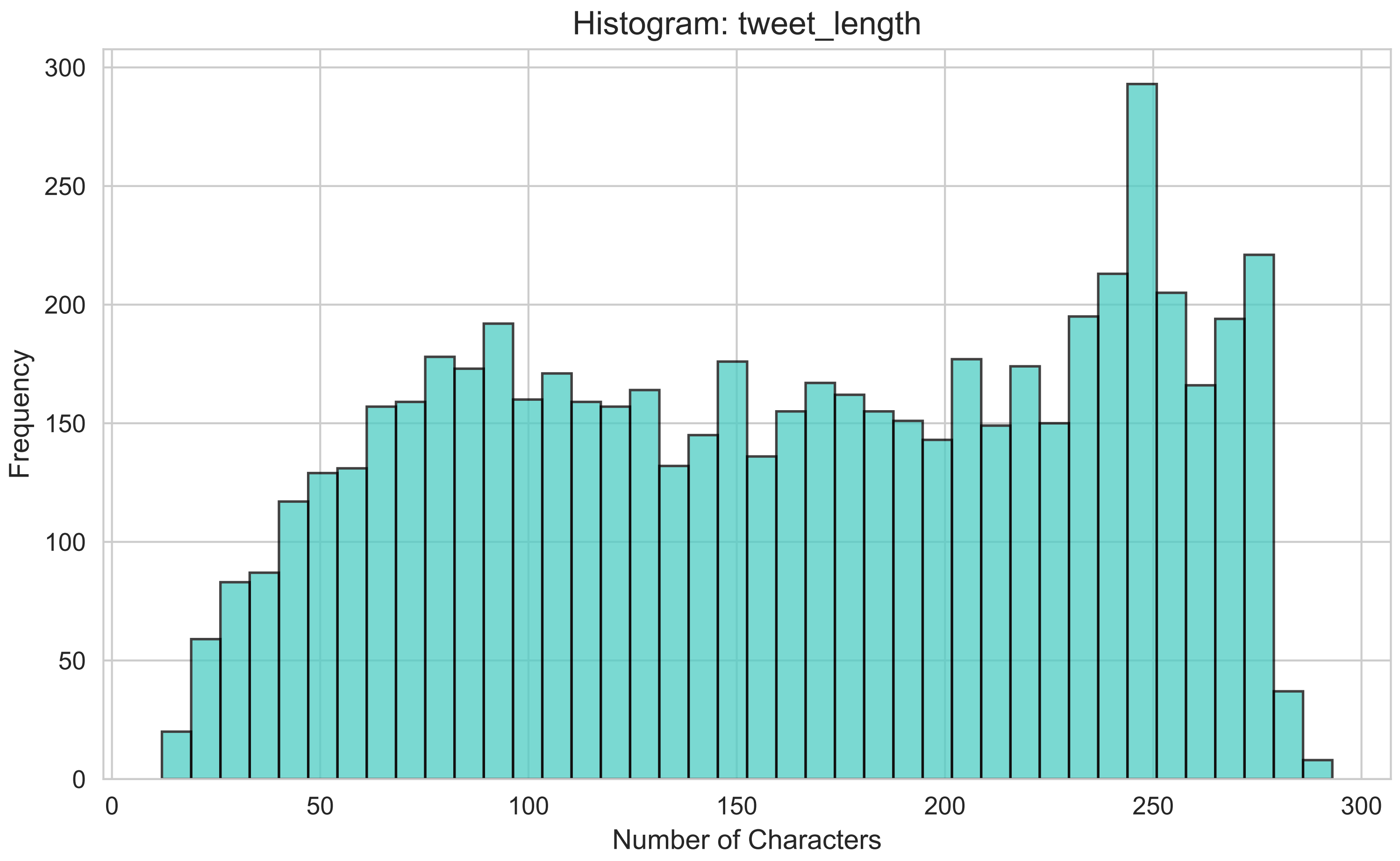
```
--- Variable: vader_neg (float64) ---
  Proportion of text tokens with negative sentiment [0, 1].
  Source: Computed from tweet_text using vaderSentiment.

    Observations: 6000   Missing: 0
    Mean:   0.0494
    Std:    0.0825
    Min:    0.0000
    25th:   0.0000
    Median: 0.0000
    75th:   0.0820
    Max:    0.6830
```


Histogram: vader_neg
Frequency
Negative Proportion
0
500
1000
1500
2000
2500
3000
3500
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7

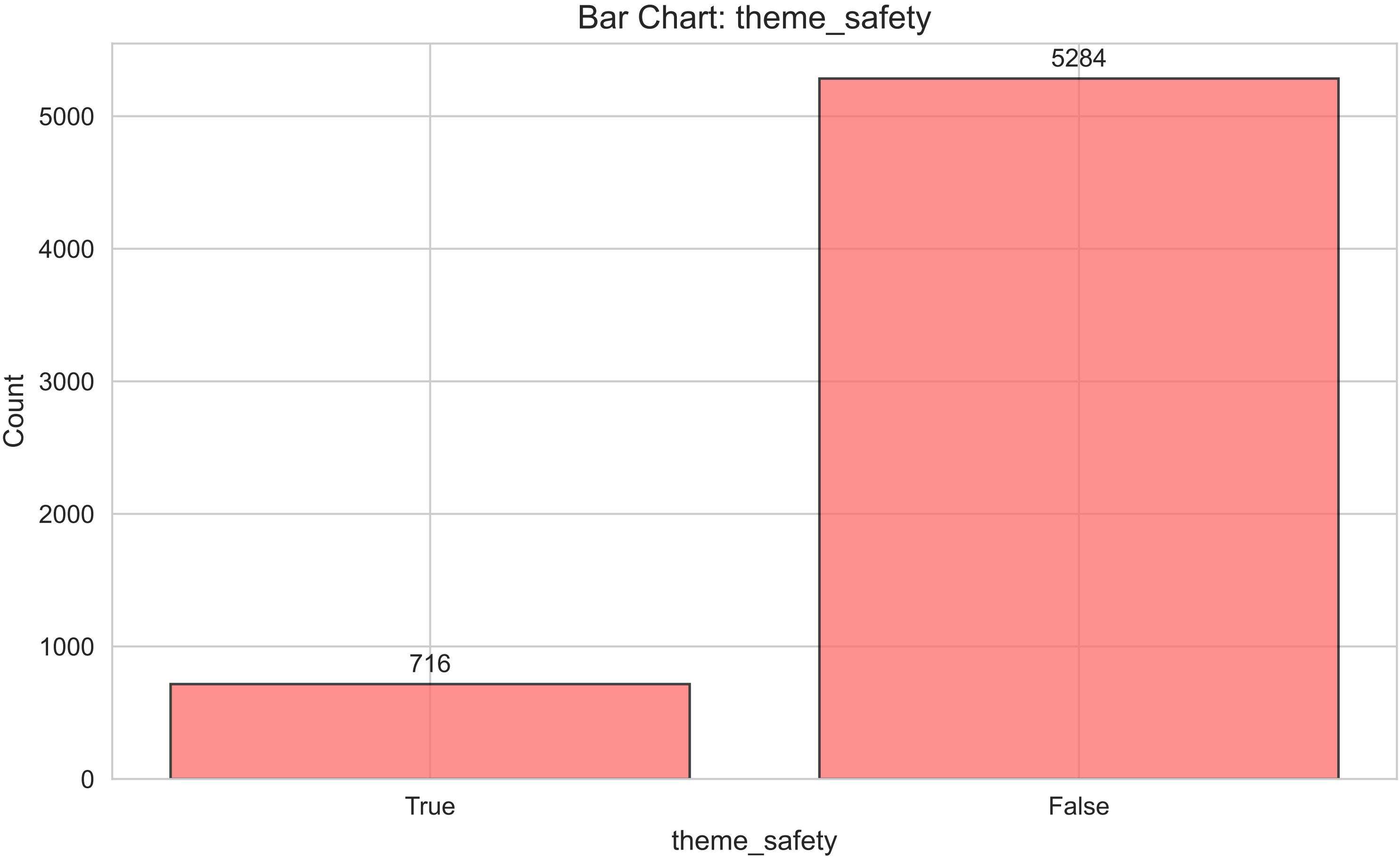
```
--- Variable: vader_neu (float64) ---
  Proportion of text tokens with neutral sentiment [0, 1].
  Source: Computed from tweet_text using vaderSentiment.

    Observations: 6000   Missing: 0
    Mean:   0.8565
    Std:    0.1326
    Min:    0.1910
    25th:   0.7690
    Median: 0.8670
    75th:   1.0000
    Max:    1.0000
```


Histogram: vader_neu
Frequency
1750
1500
1250
1000
750
500
250
0
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Neutral Proportion

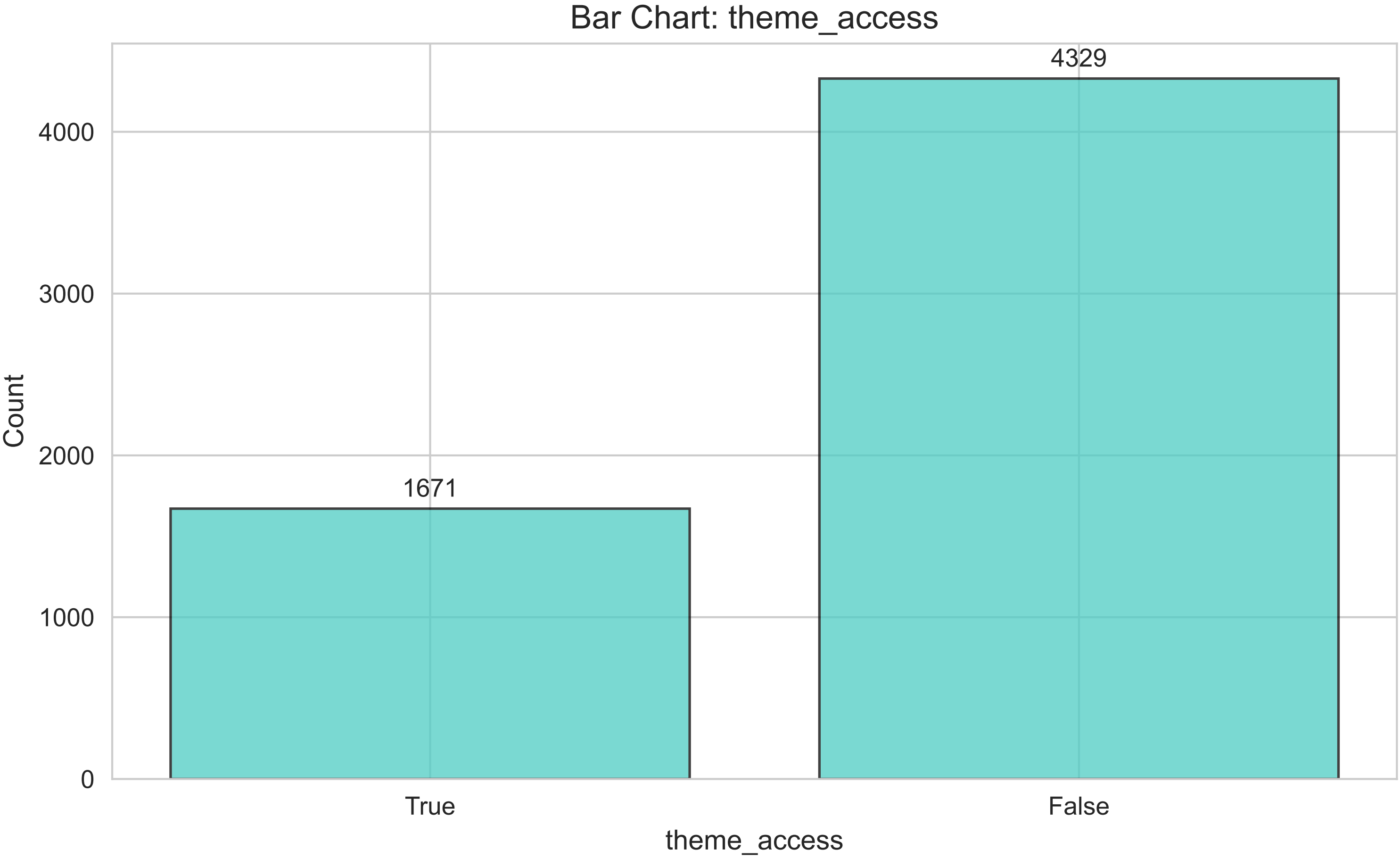
```
--- Variable: tweet_length (int64) ---
  Character count of the cleaned tweet text.
  Source: Computed as len(tweet_text) in analysis.py.

    Observations: 6000   Missing: 0
    Mean:   161.8418
    Std:    73.8684
    Min:    12.0000
    25th:   97.0000
    Median: 164.0000
    75th:   230.0000
    Max:    293.0000
```


Histogram: tweet_length
Frequency
0
50
100
150
200
250
300
Number of Characters
0
50
100
150
200
250
300

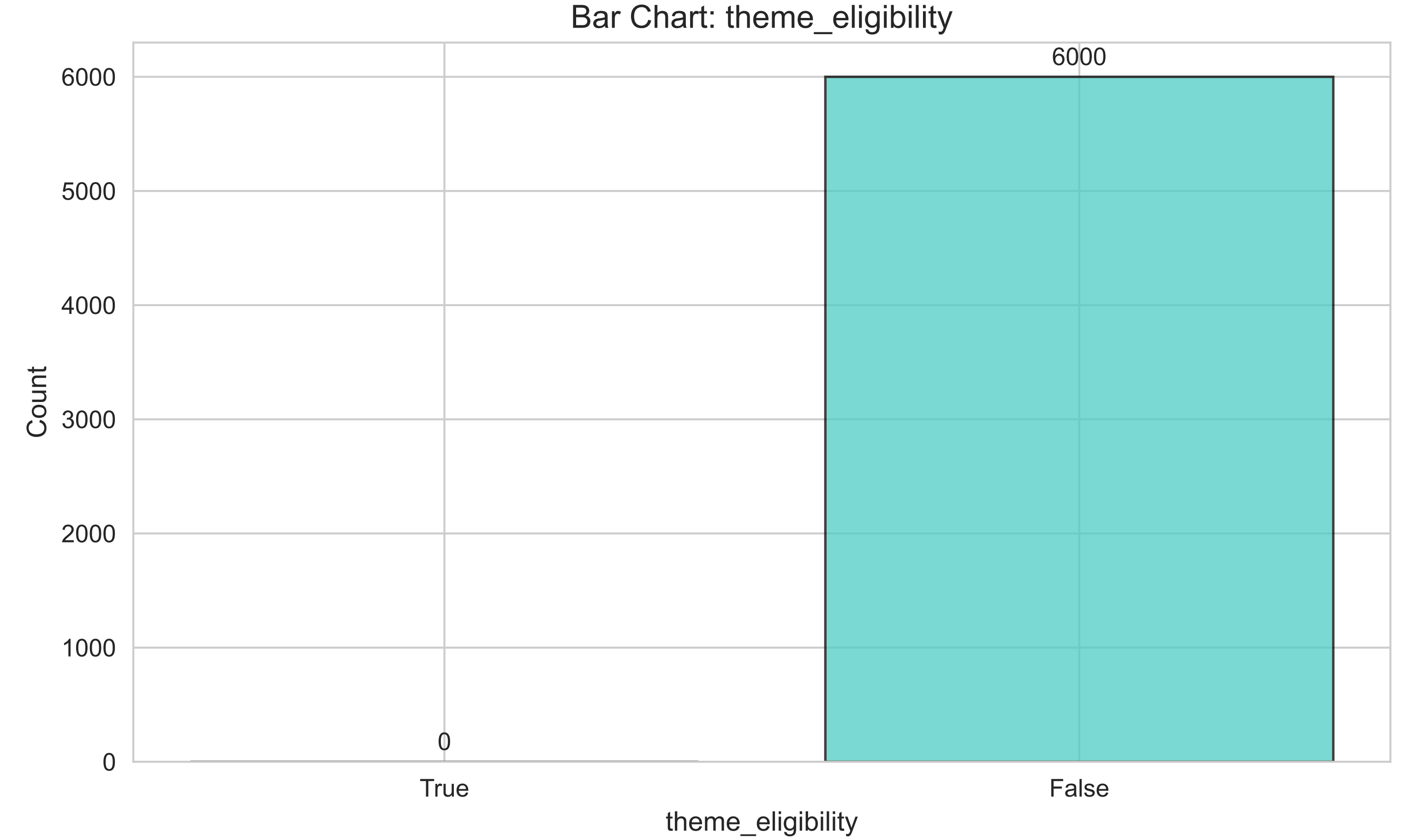
```
--- Variable: theme_safety (bool) ---
  True if tweet contains a safety/side-effects keyword.
    Source: Keyword matching in analysis.py.
  Observations: 6000   Missing: 0
  Frequency table:
    True:  716 (11.9%)
    False: 5284 (88.1%)
```


Bar Chart: theme_safety
Count
0
1000
2000
3000
4000
5000
716
5284
True
False
theme_safety

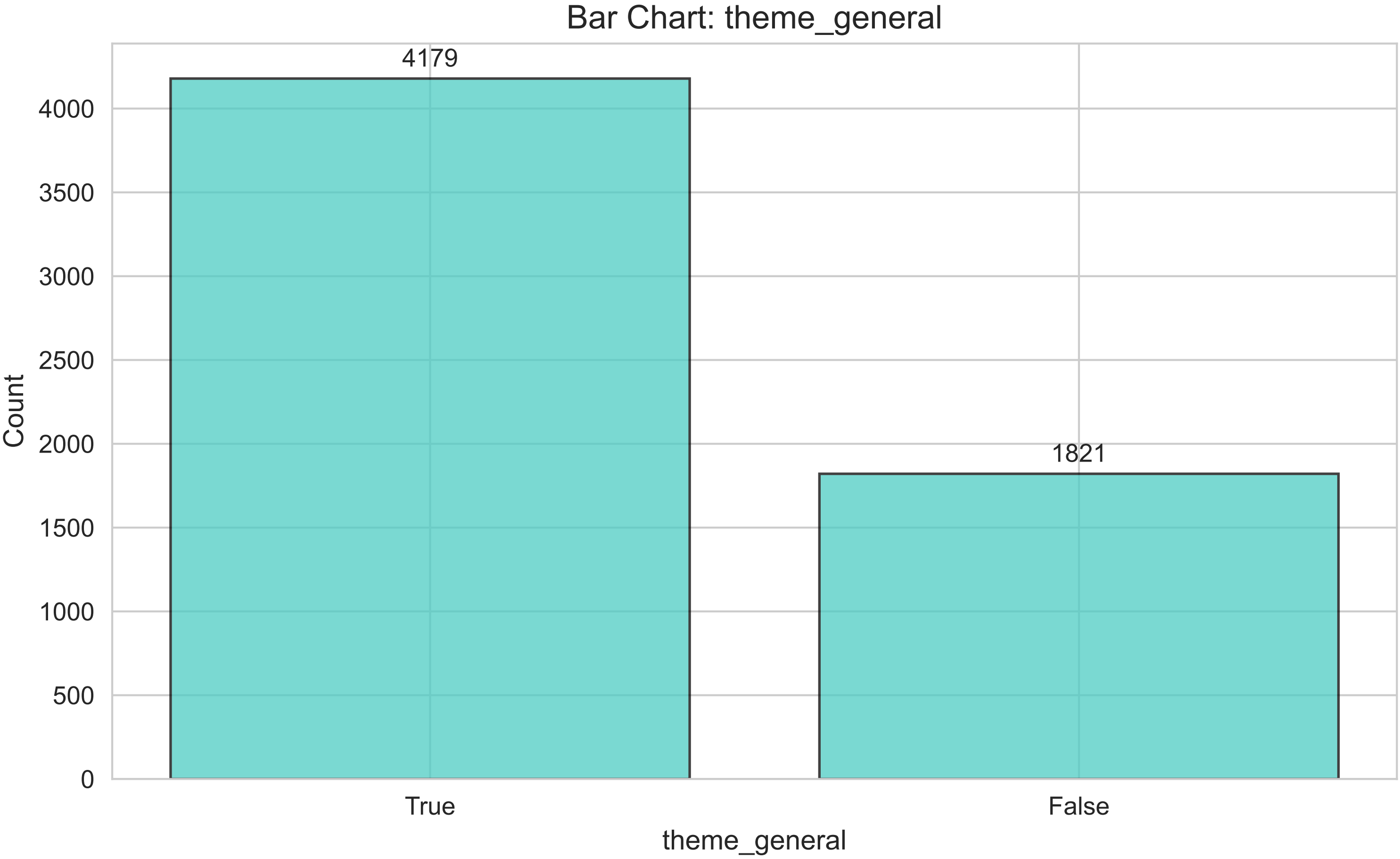
```
--- Variable: theme_access (bool) ---
  True if tweet contains an access/appointments keyword.
    Source: Keyword matching in analysis.py.
  Observations: 6000   Missing: 0
  Frequency table:
    True:  1671 (27.9%)
    False: 4329 (72.2%)
```


Bar Chart: theme_access
Count
0
1000
2000
3000
4000
1671
4329
True
False
theme_access

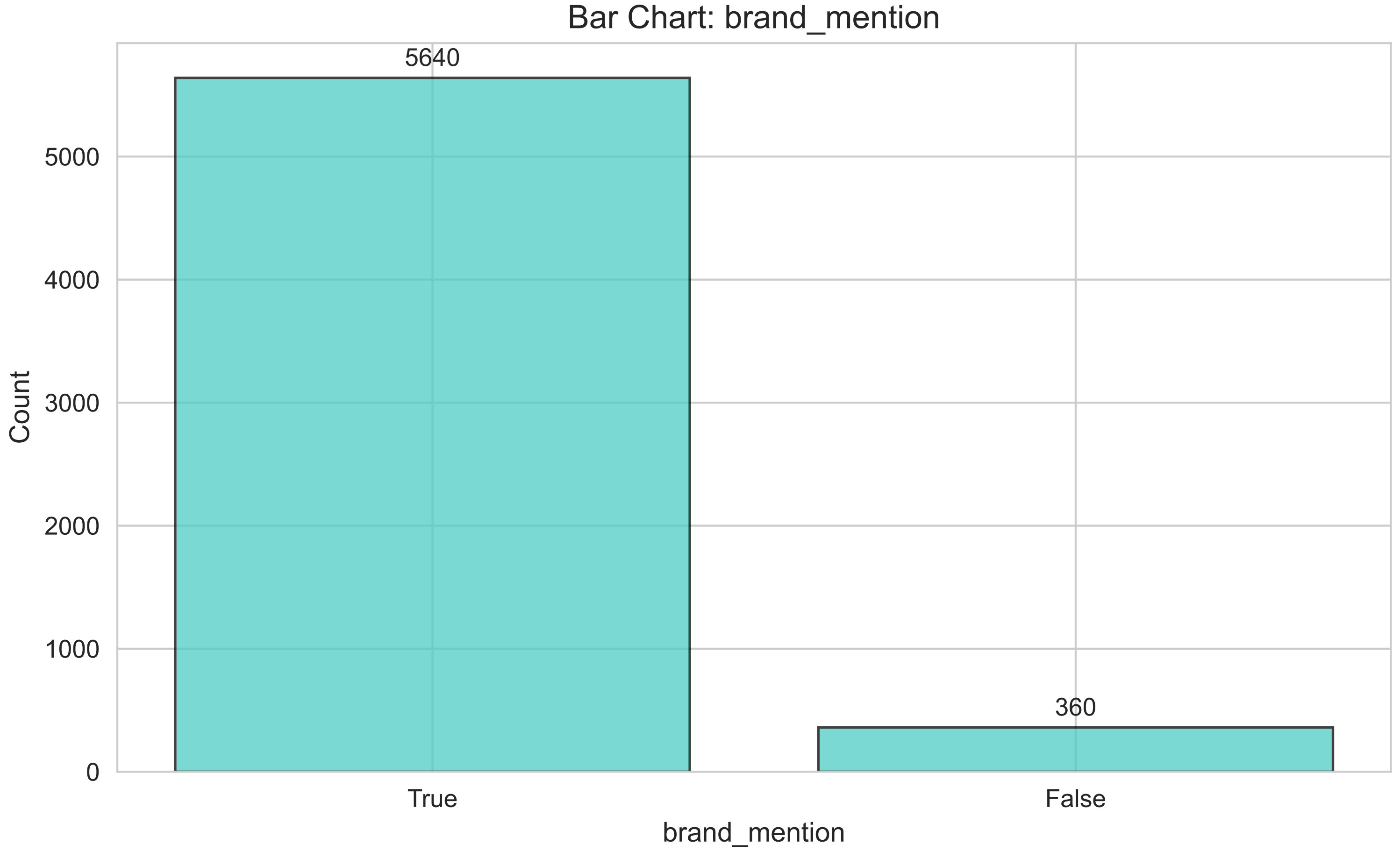
```
--- Variable: theme_eligibility (bool) ---
  True if tweet contains an eligibility keyword.
    Source: Keyword matching in analysis.py.
  Observations: 6000   Missing: 0
  Frequency table:
    True:  0 (0.0%)
    False: 6000 (100.0%)
```


Bar Chart: theme_eligibility
Count
0
1000
2000
3000
4000
5000
6000
True
False
theme_eligibility
0
6000

```
--- Variable: theme_general (bool) ---
  True if tweet contains a general-information keyword.
    Source: Keyword matching in analysis.py.
  Observations: 6000   Missing: 0
  Frequency table:
    True:  4179 (69.7%)
    False: 1821 (30.3%)
```


Bar Chart: theme_general
Count
theme_general
4179
1821
True
False
0
500
1000
1500
2000
2500
3000
3500
4000

```
--- Variable: brand_mention (bool) ---
  True if tweet mentions a vaccine brand name.
    Source: Keyword matching in analysis.py.
  Observations: 6000   Missing: 0
  Frequency table:
    True:  5640 (94.0%)
    False: 360 (6.0%)
```


Bar Chart: brand_mention
5640
360
0
1000
2000
3000
4000
5000
Count
True
False
brand_mention

```
============================================================
END OF DATA APPENDIX
============================================================
```