Data Appendix

Analysis Data File: GPT App Reviews Analysis

Unit of Observation:

Each row in this dataset represents a single review of the GPT application. The dataset contains information such as the app version, rating, comments, thumbs up count, and the results of various text analysis processes.

Overview of Analysis Data File:

The analysis data file `reviews_data_cleaned` is derived from the original file by dropping rows with missing values and resetting the index. This cleaned dataset is used for the subsequent analysis steps.

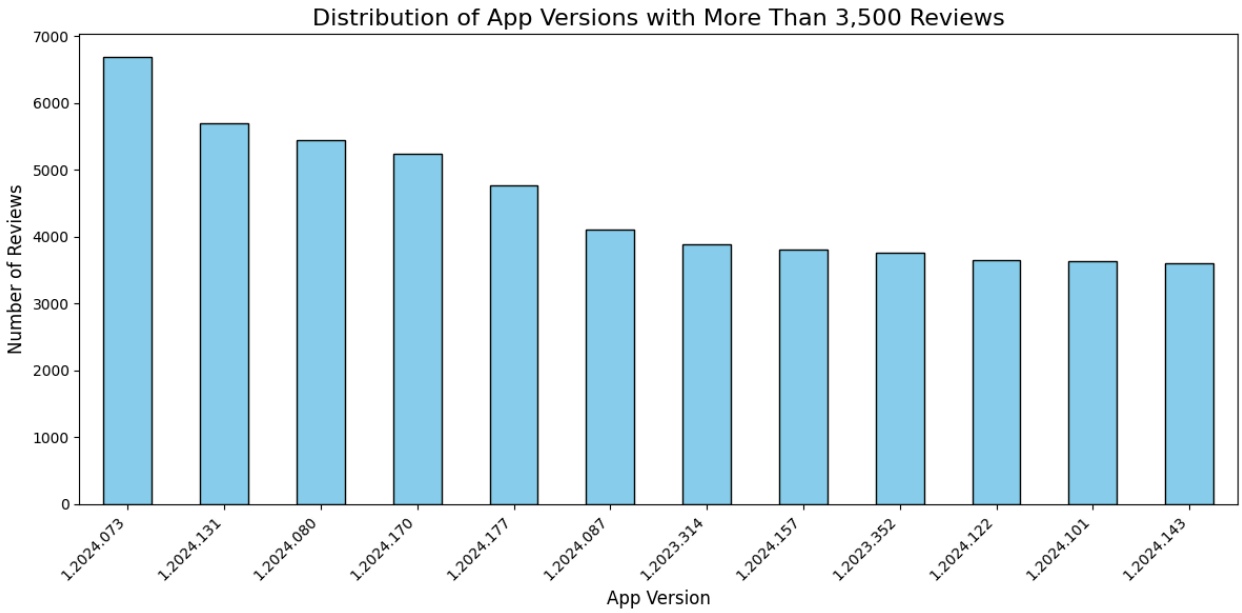
Variables in the Analysis Data File:

1. App Version

- **Definition**: The version of the GPT app that the review corresponds to.
- **Type**: Categorical
- **Coding Scheme**: String values representing different versions (e.g., "1.0.0", "2.0.1").
- **Summary Statistics**:
    - Frequency counts for each app version.

App Version	Frequency
1.2024.073	6692
1.2024.131	5699
1.2024.080	5442
1.2024.170	5232
1.2024.177	4764
1.2024.087	4110
1.2023.314	3890
1.2024.157	3803
1.2023.352	3761
1.2024.122	3645
1.2024.101	3640
1.2024.143	3597

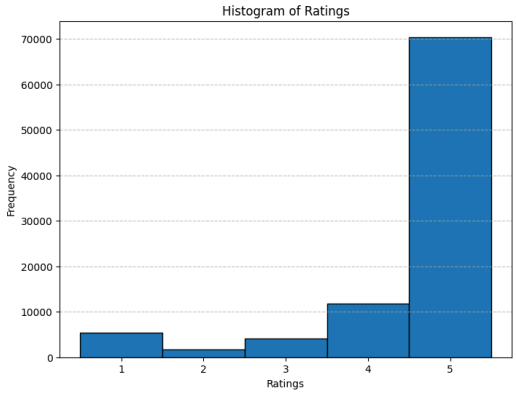
- **Visualization**: Bar chart showing the distribution of reviews across app versions.



2. Rating

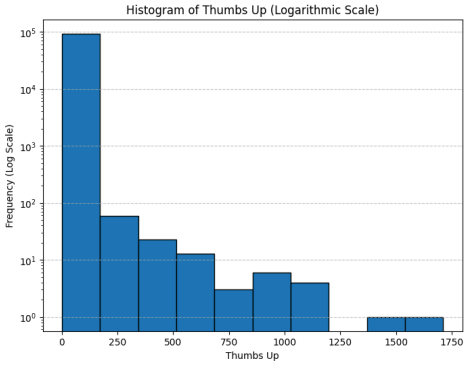
- **Definition**: The rating given by the user, on a scale of 1 to 5.
- **Type**: Quantitative
- **Summary Statistics**:
    - Count: 93268
    - Mean: 4.500847
    - Standard deviation: 1.068453
    - Min: 1
    - 25th / Median / 75th percentiles / Max : 5

- **Visualization**: Histogram showing the distribution of ratings.



3. Thumbs Up

- **Definition**: The number of thumbs up (likes) a review received.
- **Type**: Quantitative
- **Summary Statistics**:
    - Count: 93268
    - Mean: 18.244106
    - Standard deviation: 1.1
    - Min: 0
    - 25th / Median / 75th percentiles : 0
    - Max: 1712
- **Visualization**: Histogram showing the distribution of thumbs up (in a log scale)



4. Comment

- **Definition**: The review text left by the user.
- **Type**: Categorical
- **Additional Information**:
    - Preprocessed using techniques such as removal of stopwords and special characters.
    - Transformed using CountVectorizer for text analysis.

5. Preprocessed Comment

- **Definition**: A cleaned and normalized version of the original comments.
- **Type**: Categorical
- **Coding Scheme**: Lowercase text with stopwords and special characters removed.
- **Summary Statistics**: Not applicable (used for text analysis purposes).
    -

6. Compound Score

- **Definition**: A single sentiment score derived from the VADER sentiment analysis, ranging from -1 (most negative) to +1 (most positive).
- **Type**: Quantitative
- **Summary Statistics**:

  count 54275.000000

  mean 0.448871

  std 0.352049

  min -0.980800

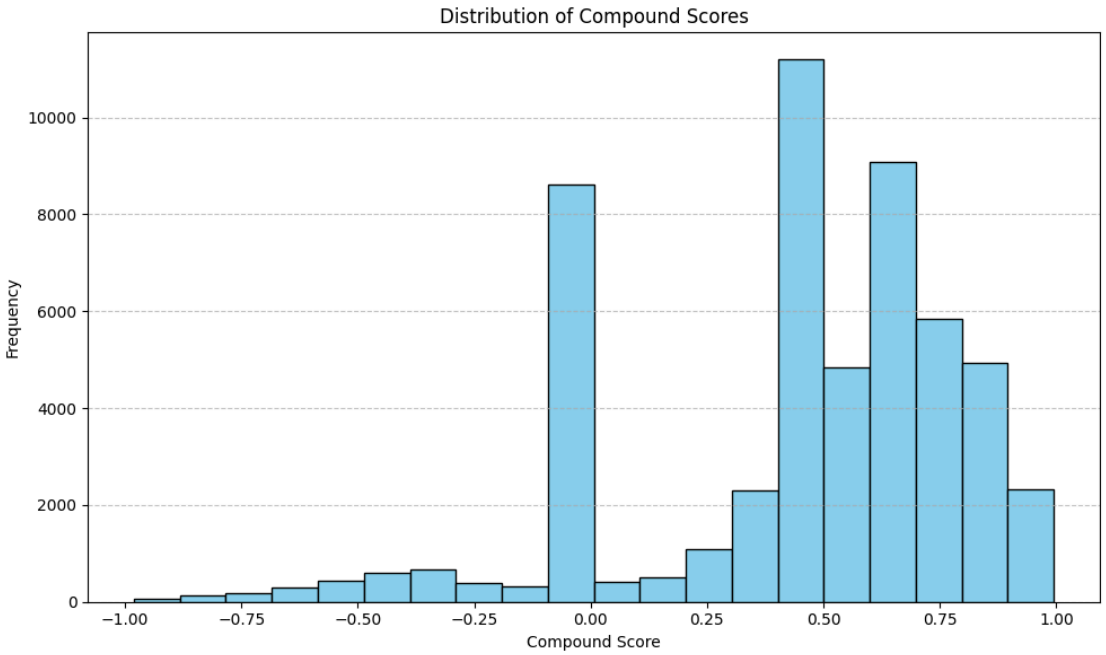
  25% 0.286350

  50% 0.493900

  75% 0.689200

  max 0.994600

- **Visualization**: Histogram showing the distribution of compound sentiment scores.

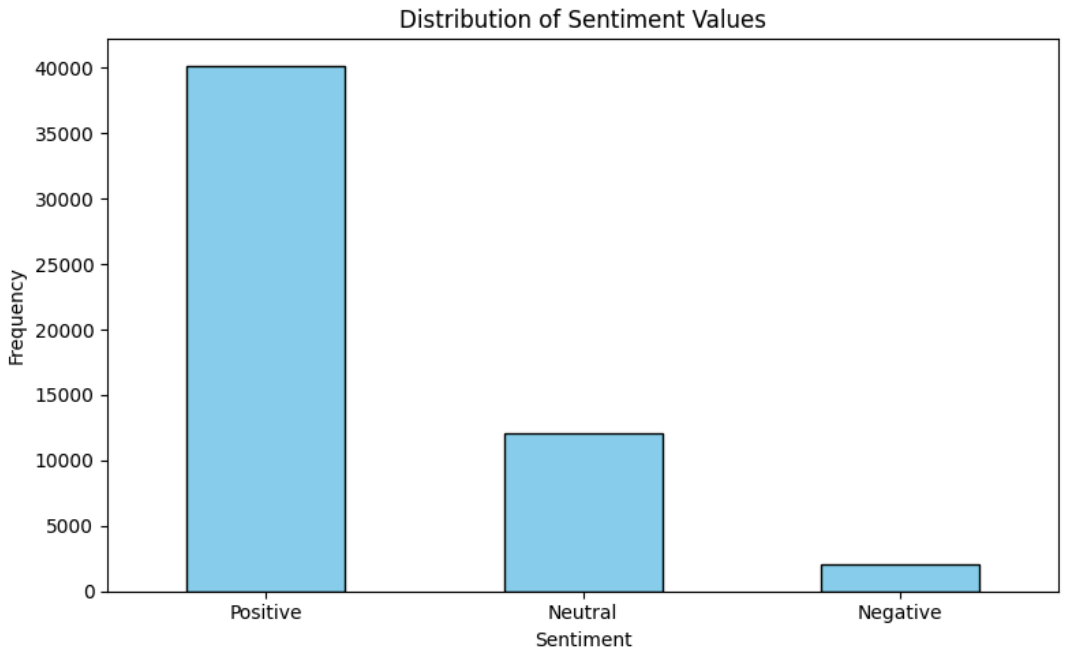


7. Sentiment

- **Definition**: Categorization of the compound score into 'Positive', 'Negative', or 'Neutral'.
- **Type**: Categorical
- **Coding Scheme**:
    - Positive: Compound score ≥ 0.33
    - Neutral: Compound score between -0.33 and 0.33
    - Negative: Compound score ≤ -0.33
- **Summary Statistics**:
    - Frequency count of each sentiment category.

Sentiment	Count
Positive	40177
Neutral	12046
Negative	2052

- **Visualization**: Bar chart showing the distribution of sentiments.

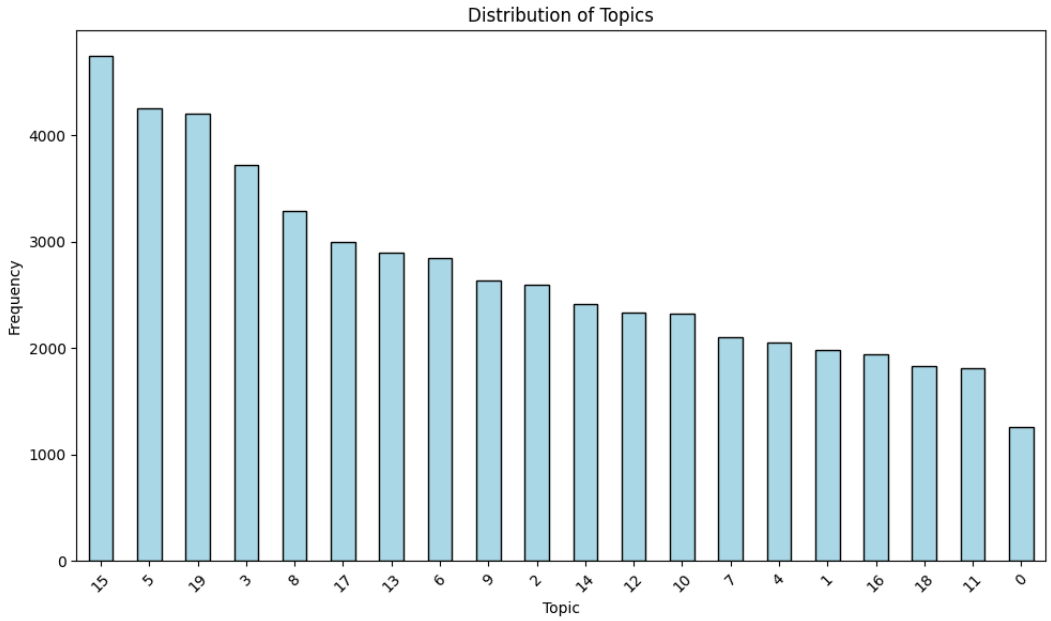


8. Topic

- **Definition**: Topic category assigned to each comment based on Latent Dirichlet Allocation (LDA) analysis.
- **Type**: Categorical
- **Coding Scheme**: Integer values representing different topics (e.g., 0, 1, 2, …, 19).
- **Summary Statistics**:
    - Frequency count of comments per topic.

Topic	Count
15	4749
5	4252
19	4202
3	3726
8	3293
17	3003
13	2895
6	2851
9	2642
2	2598
14	2414
12	2337
10	2324
7	2105
4	2051
1	1988
16	1946
18	1831
11	1809
0	1259

- **Visualization**: Bar chart showing the distribution of topics.



Data Cleaning and Preprocessing:

- Rows with missing values were removed.
- Text data was preprocessed by removing special characters and stopwords, and transforming text using `CountVectorizer`.
- Sentiment analysis was applied using the VADER lexicon.
- Topic modeling was performed using Latent Dirichlet Allocation (LDA).

Visualization and Summary:

The analysis data file was visualized using various plots to understand the distribution of key variables such as ratings, thumbs up counts, sentiment scores, and topic distributions. These visualizations are included in the Data Appendix for reference.