HELPFUL REVIEWS

According to the Machine

By Corey Wade, September 1, 2018

THE PROBLEM

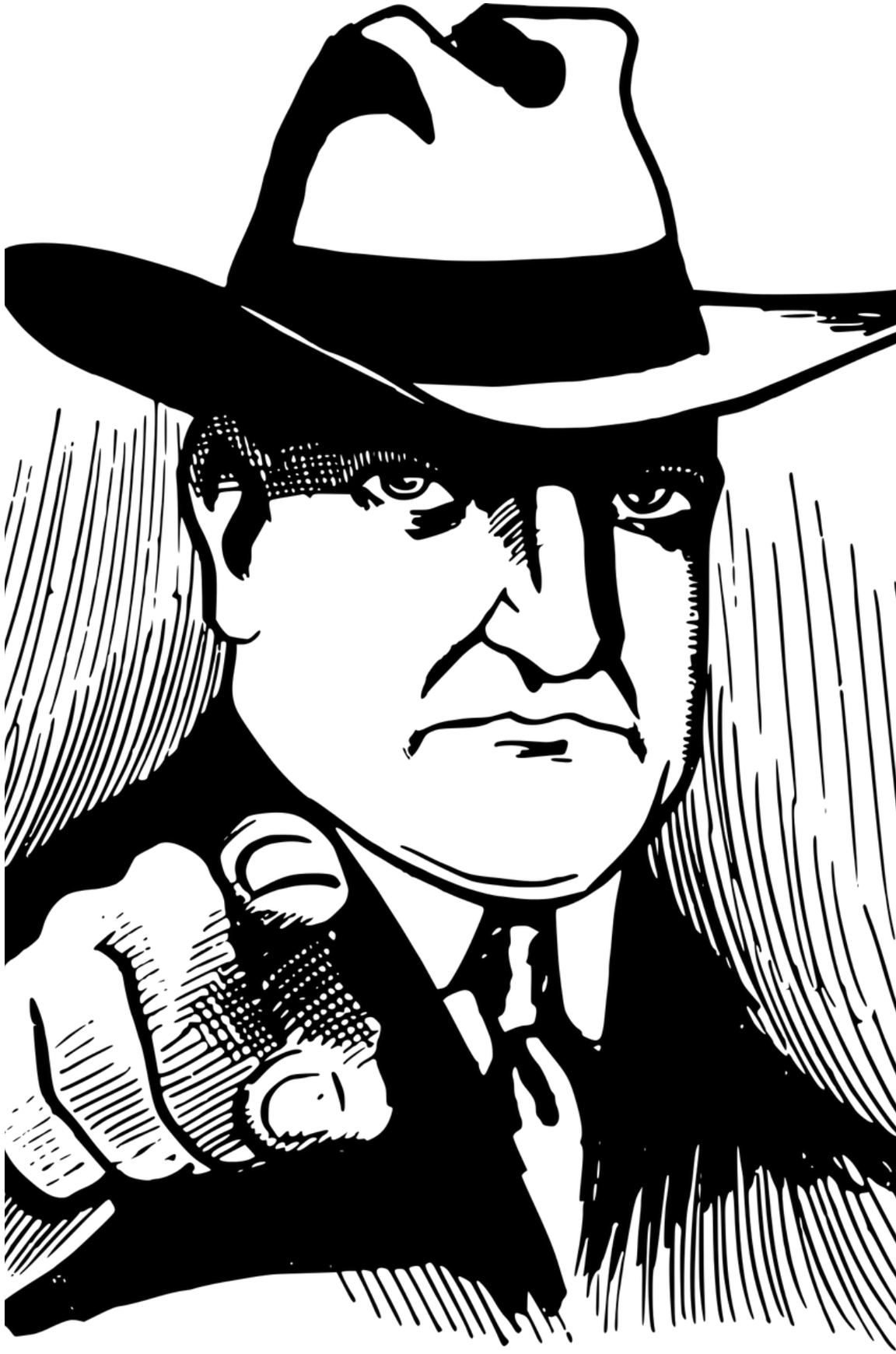
- Amazon's Fake Review Problem Is Now Worse Than Ever, Study Suggests
  - Forbes
- How merchants use Facebook to flood Amazon with fake reviews
  - Washington Post
- Scammers elude Amazon crackdown on fake reviews with new tricks
  - New York Post

THE MISSION

- Step one: determine whether given reviews are helpful.
- The starting point will be Amazon book reviews.
- A new metric must be created to classify helpful reviews.
- The metric will be used later to determine helpful reviewers.
- Identifying helpful reviewers help protect consumers from fraud.

THE DATA

- Amazon Book Reviews Dataset.
- 8.9 million rows.
- May 1996 - July 2014.
- Limited to users and books with at least 5 reviews.
- Relevant columns include star rating, text of review, and helpful/unhelpful votes.
- See http://jmcauley.ucsd.edu/data/amazon/ for more info and datasets.



SAMPLE DATA

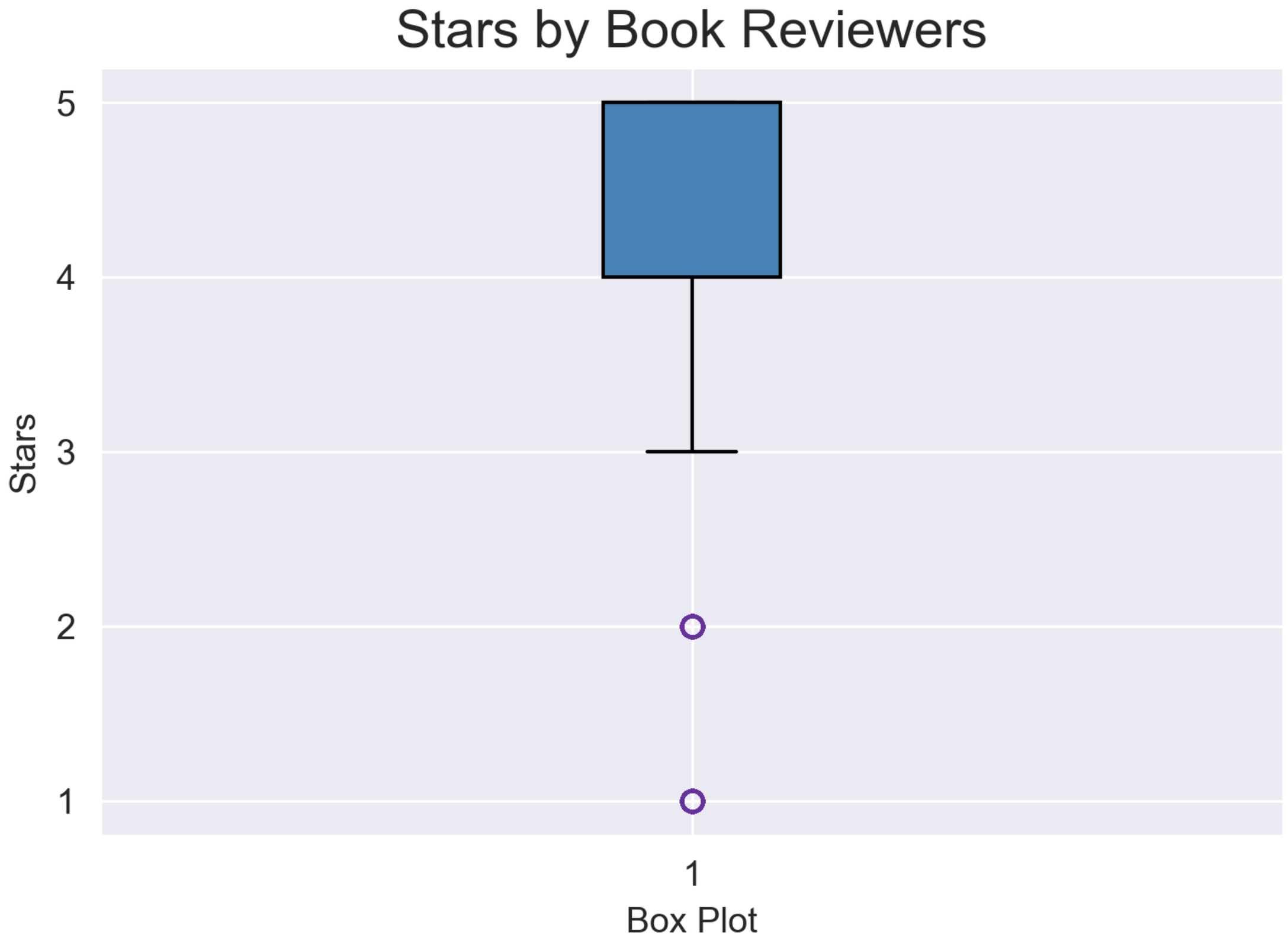
	asin	helpful	overall	reviewText
0	000100039X	[0, 0]	5	Spiritually and mentally inspiring! A book tha...
1	000100039X	[0, 2]	5	This is one my must have books. It is a master...
2	000100039X	[0, 0]	5	This book provides a reflection that you can a...
3	000100039X	[0, 0]	5	I first read THE PROPHET in college back in th.
4	000100039X	[7, 9]	5	A timeless classic. It is a very demanding an...

SAMPLE DATA CONT'D

reviewTime	reviewerID	reviewerName	summary	unixReviewTime
12 16, 2012	A10000012B7CGYK OMPQ4L	Adam	Wonderful!	1355616000
12 11, 2003	A2S166WSCFIFP5	adead_poet@ hotmail.com	close to god	1071100800
01 18, 2014	A1BM81XB4QHOA3	Ahoro Blethends	Must Read for Life	1390003200
09 27, 2011	A1MOSTXNIO5MPJ	Alan Krug	Timeless for every good	1317081600
10 7, 2002	A2XQ5LZHTD4AFT	Alaturka	A Modern Rumi	1033948800

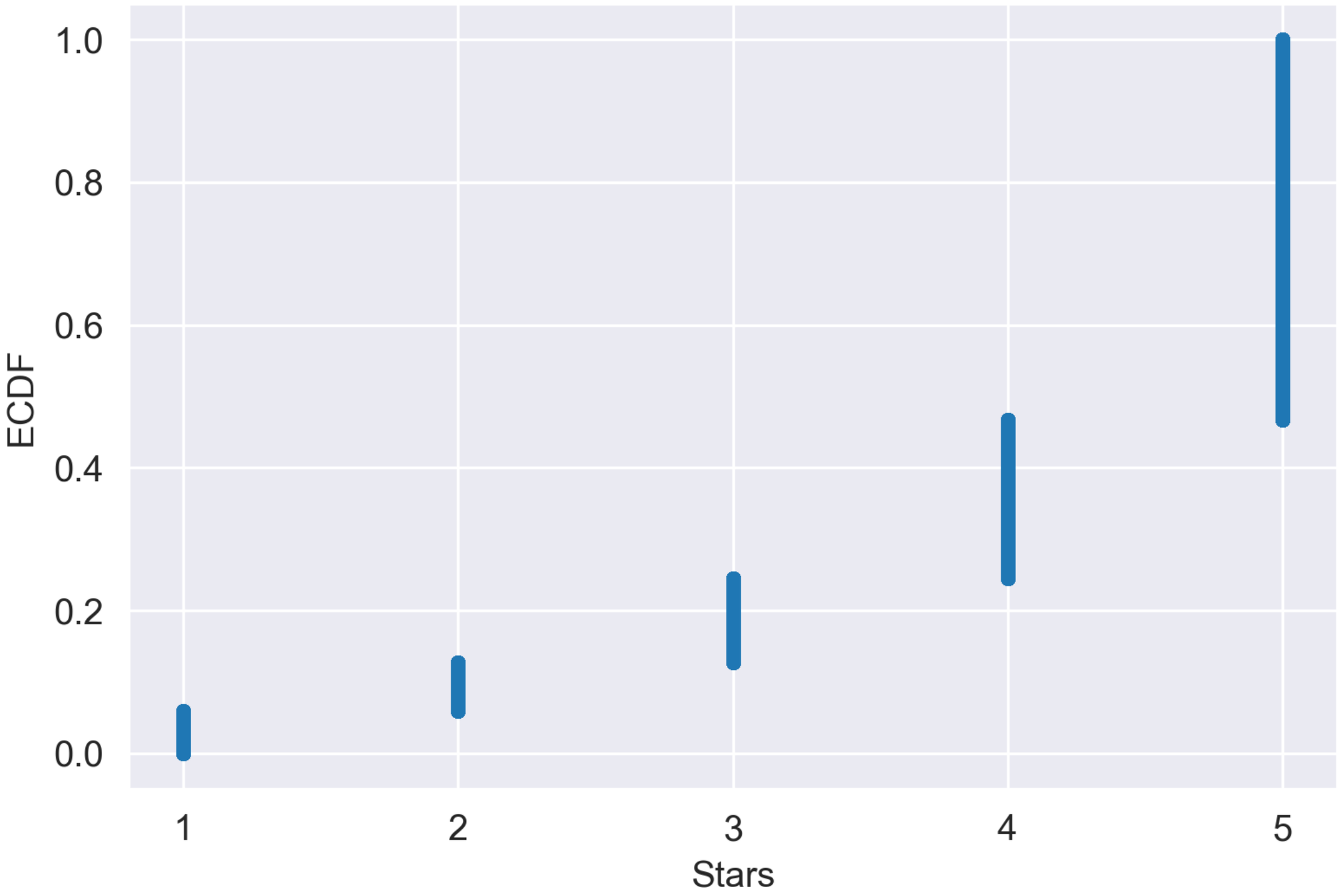
THE FACTS

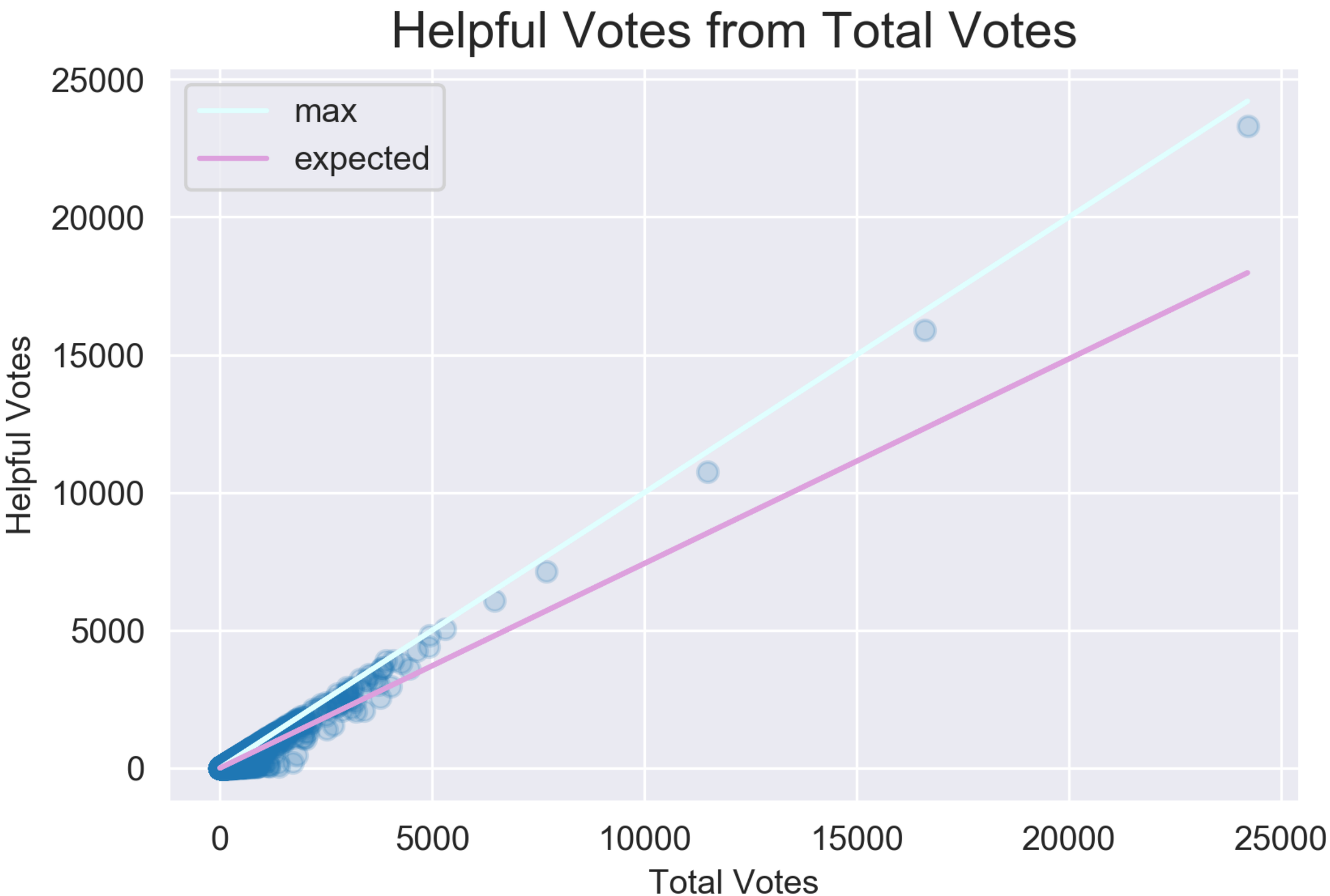
- Median review is 5 stars.
- Mean review is 4.25 stars.
- 75% are 4 stars or higher.
- Reviews under 3 stars are outliers.
- Most reviews receive 0 or 1 helpful/unhelpful votes.
- 74.3% of votes are helpful.
- Top review has 23,311 helpful votes.
- Fun reviews with most helpful votes published on Medium: 5 Reviews to Make You Blush 5 Shades of Crimson.

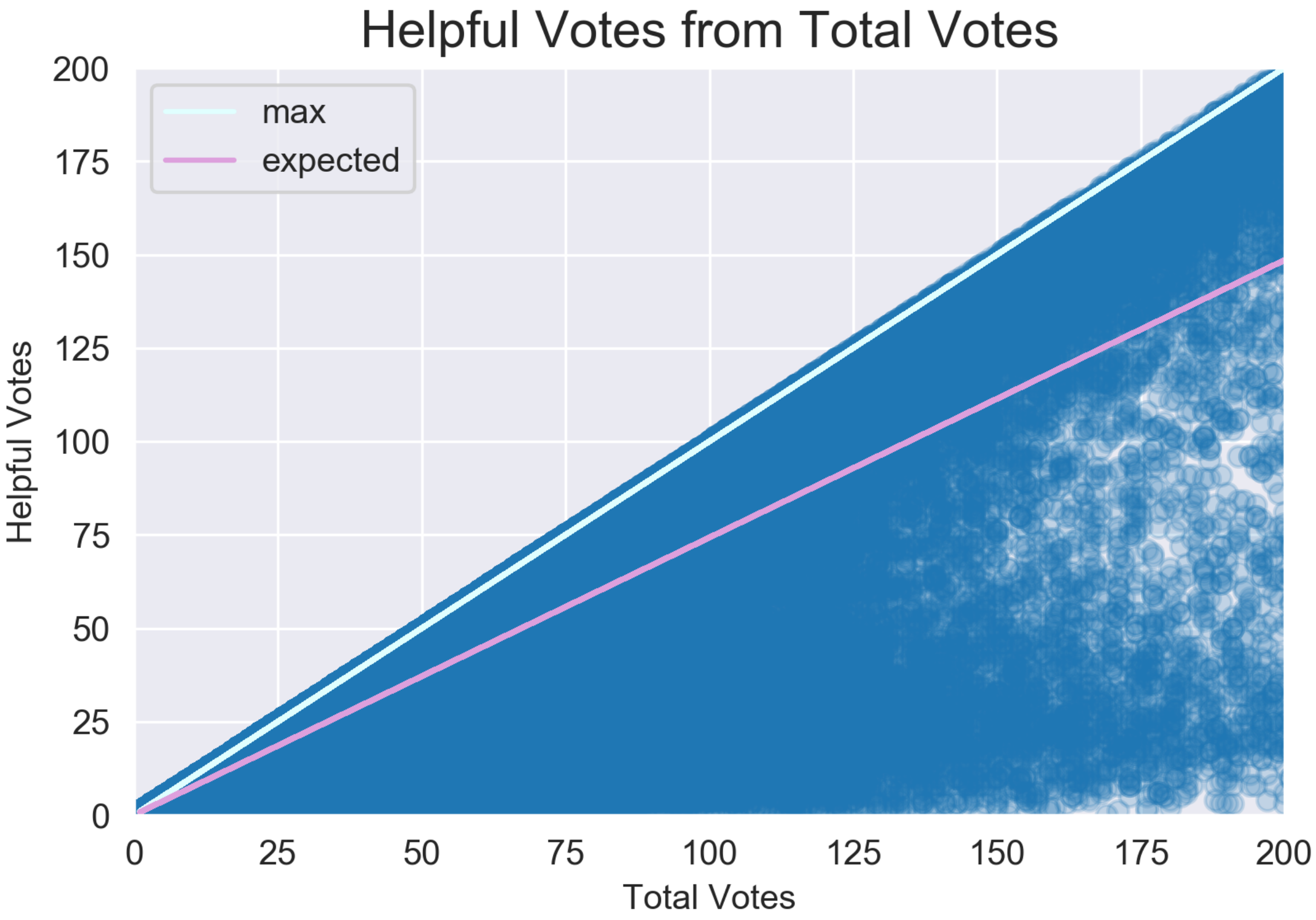

Stars by Book Reviewers
5
4
3
2
1
Stars
1
Box Plot

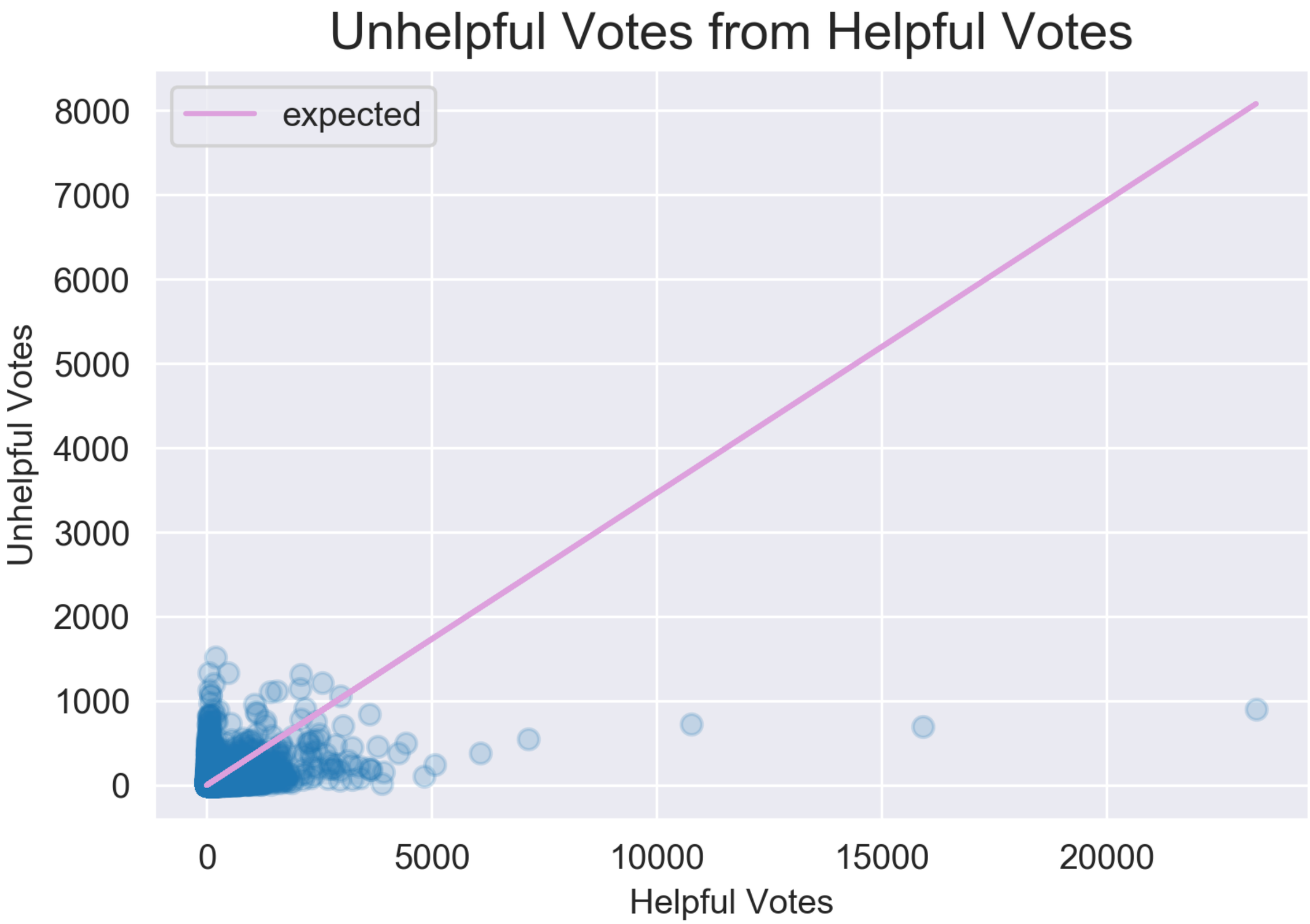

Stars by Book Reviewers
2500000
2000000
1500000
1000000
500000
0
1
2
3
4
5
Stars

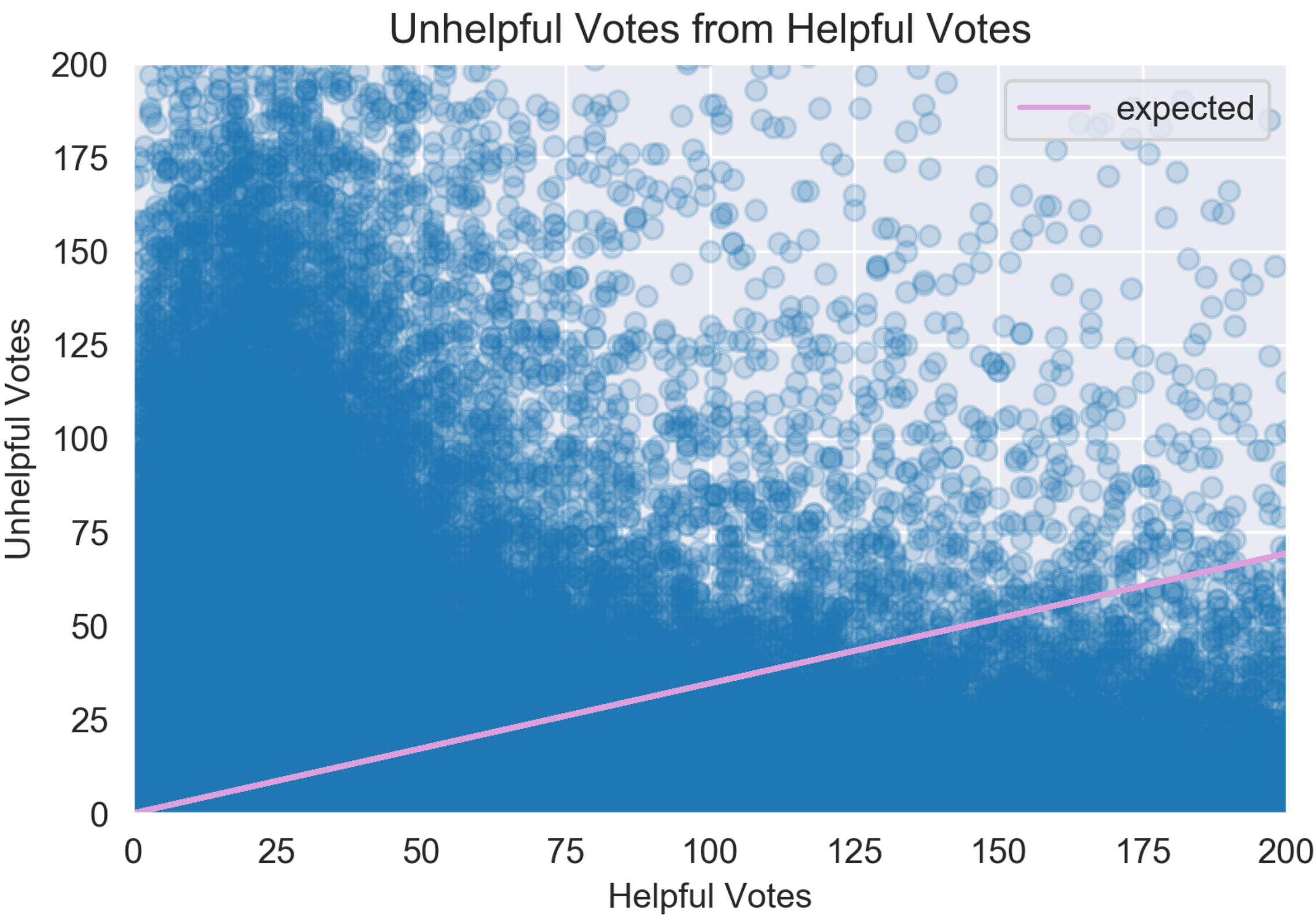
Stars by Book Reviewers

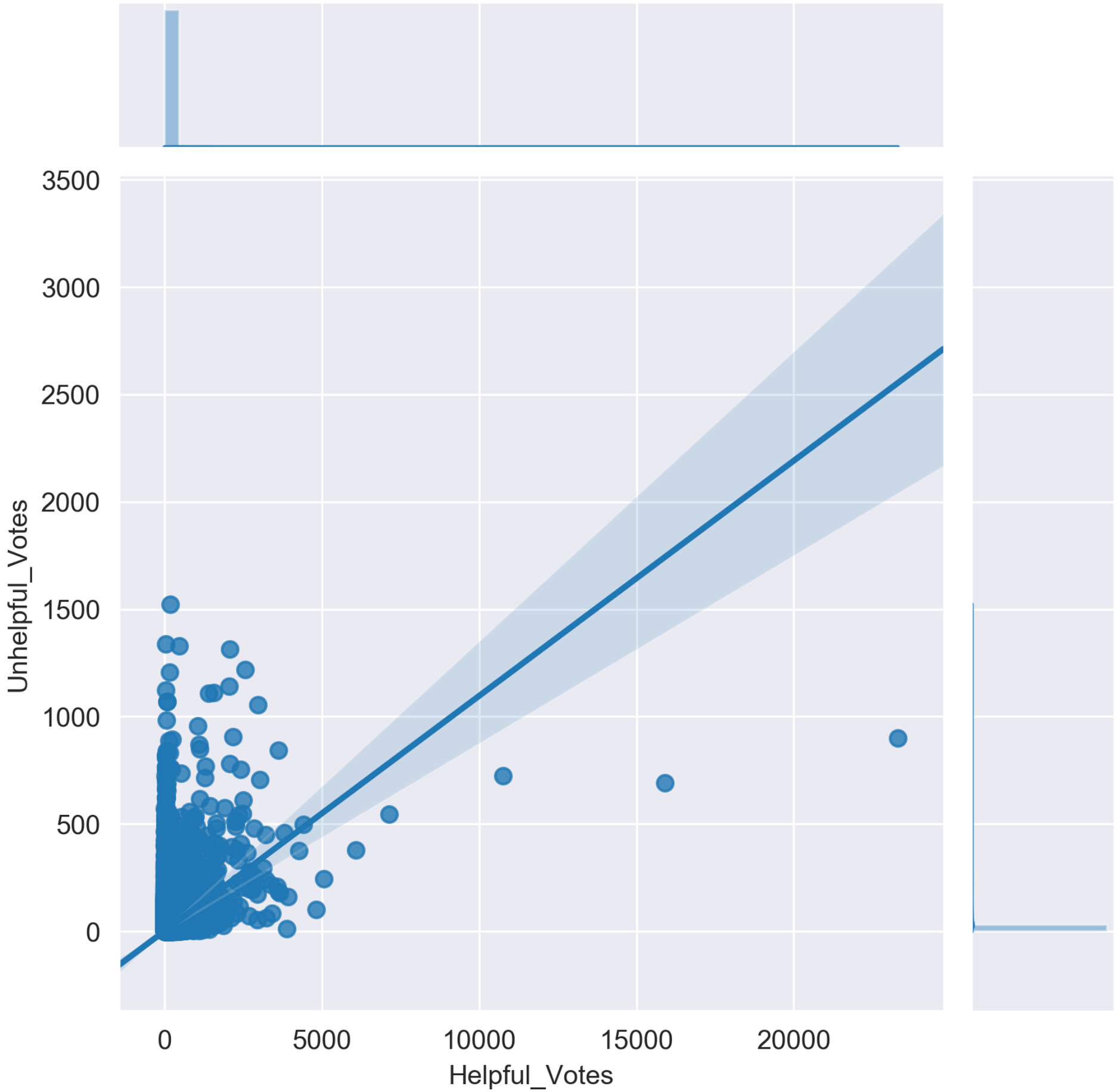



Helpful Votes from Total Votes
max
expected
25000
20000
15000
10000
5000
0
Helpful Votes
0
5000
10000
15000
20000
25000
Total Votes


Helpful Votes from Total Votes
max
expected
Helpful Votes
Total Votes
0
25
50
75
100
125
150
175
200


Unhelpful Votes from Helpful Votes
expected
8000
7000
6000
5000
4000
3000
2000
1000
0
Unhelpful Votes
0
5000
10000
15000
20000
Helpful Votes


Unhelpful Votes from Helpful Votes
expected
Unhelpful Votes
Helpful Votes
0
25
50
75
100
125
150
175
200


3500
3000
2500
2000
1500
1000
500
0
Unhelpful_Votes
0
5000
10000
15000
20000
Helpful_Votes

NEW METRIC

- There is currently no helpful rating metric.
- Strategy: combine helpful votes and helpful percentage.

CAVEATS

- Helpful Votes must be converted into numbers between 0 and 1.
- The Helpful Votes column is very right skewed.
- The final numbers should reflect meaningful percentages: 50% about average, 90% great.

SKEWED_TO_LINEAR FUNCTION

- Requirements: straight line function that avoids division by 0, Helpful Votes column, number of pivots.
- Pivots: percentiles spread across the data using logspace.
- Y-values: percentiles from the first pivot to the last.
- X-values: number of helpful votes associated with each percentile.
- Piecewise function: Adjacent (x,y) points are connected via a straight line.
- Results: All x-values receive percentile ranking from corresponding y-values on graph.
- Transformation: Helpful votes scaled between 0 and 1 with skewness intact.

See full function and all relevant jupyter notebooks at https://github.com/coreyjwade/Helpful_Reviews.

PIECEWISE LINEAR TRANSFORMATION: 50 PIVOTS

y-values

```
percents= [  1. 9.79891606  17.82362289  25.1422425   31.81690291
  37.90426555  43.45600627  48.51925398  53.13699077  57.34841678
  61.18928294  64.69219452  67.88688785  70.8004828   73.45771297
  75.88113568  78.09132347  80.10703869  81.94539282  83.62199171
  85.15106808  86.54560229  87.81743261  88.97735565  90.03521804
  91.          91.87989161  92.68236229  93.41422425  94.08169029
  94.69042656  95.24560063  95.7519254   96.21369908  96.63484168
  97.01892829  97.36921945  97.68868879  97.98004828  98.2457713
  98.48811357  98.70913235  98.91070387  99.09453928  99.26219917
  99.41510681  99.55456023  99.68174326  99.79773557  99.9035218
 100.  ]
```

x-values

```
helpful_votes = [0.0, 0.0, 1.0, 1.0, 1.0, 1.0, 1.0, 2.0, 2.0, 2.0,
2.0, 3.0, 3.0, 4.0, 4.0, 4.0, 5.0, 6.0, 6.0, 7.0, 8.0, 9.0, 9.0, 10.0,
11.0, 12.0, 14.0, 15.0, 16.0, 18.0, 19.0, 21.0, 23.0, 25.0, 28.0,
30.0, 33.0, 36.0, 40.0, 44.0, 48.0, 54.0, 60.0, 67.0, 77.0, 89.0,
104.0, 126.0, 163.0, 243.0, 23311.0]
```

HELPFUL RATING METRIC

- Helpful Votes scaled between 0 and 1 and Helpful Vote Percentage can now be combined.
- Percentage of Helpful Reviews per Book added to balanced popularity.
- Multiplying the standard deviation of each column by 100 and rounding down gives 0.42 for Helpful Votes Scaled and 0.56 for Helpful Percentage.
- The remaining percentage is left for Percentage of Helpful Reviews per Book.

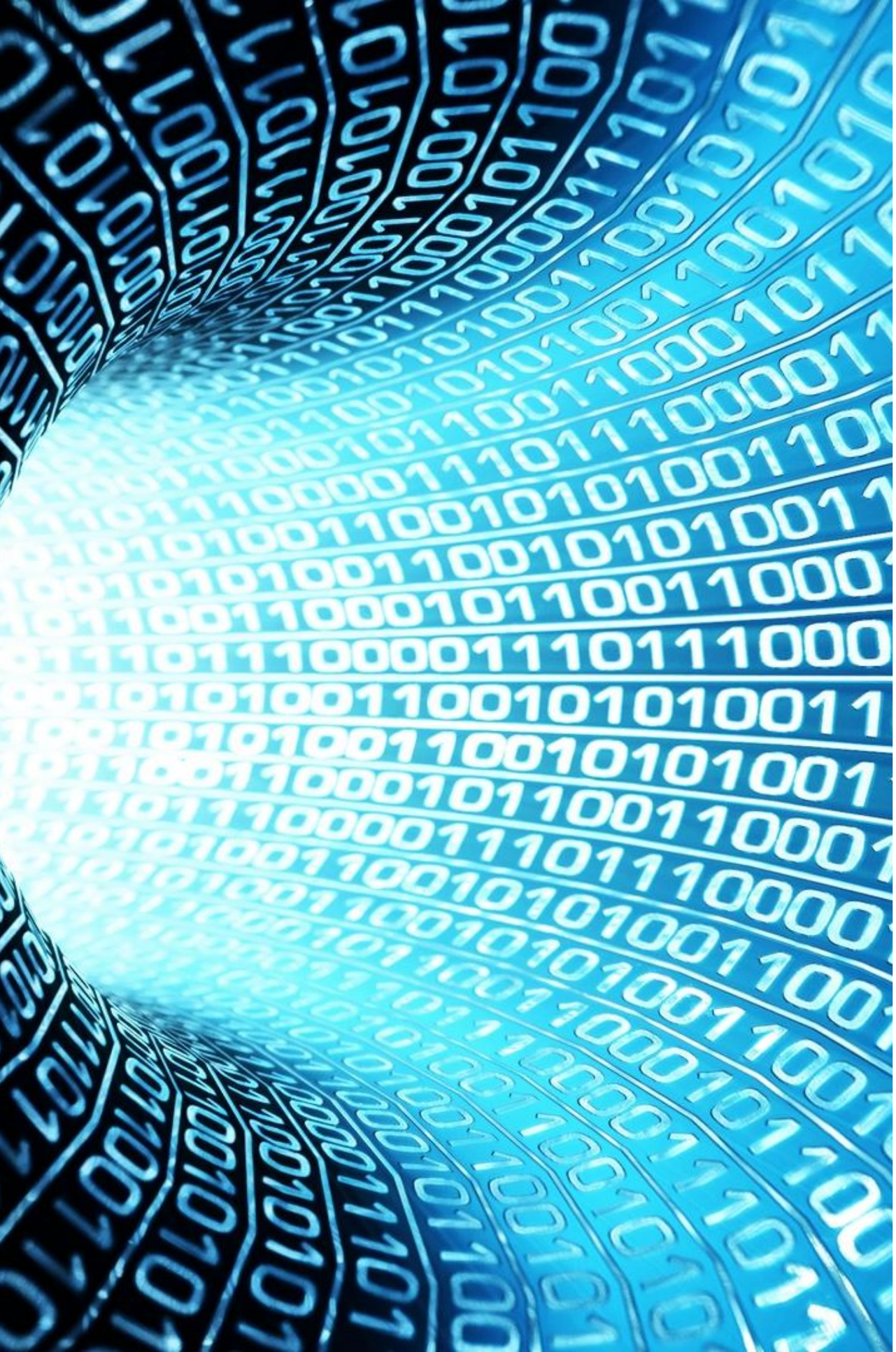
$$Helpful\_Rating = 0.42 * Helpful\_Votes\_Scaled + 0.56 * Helpful\_Percentage + 0.02 * Percentage\_Helpful\_Reviews\_Book$$

NATURAL LANGUAGE PROCESSING

- Before making predictions, book reviews must be converted into a corpus.
- Normalize corpus with lowercase letters; eliminate stop words and special characters.
- Use CountVectorizer and TfidfVectorizer to convert individual reviews into a sparse matrix.
- Iterate over n-grams. One word, is the default. Also try 2 word combinations and 3 word combinations.
- Best results consistently came from CountVectorizer(ngram_range=(1,2)) using, 1 and 2 word combinations.

PREDICTIONS

- Question: Is a particular review helpful?
- Y is the helpful rating.
- Instead of predicting an exact rating, the data is split into helpful and unhelpful scores.
- Reviews with a helpful rating of over 85% are helpful.
- Reviews with a helpful rating of under 50% are not helpful.
- Leaving out the middle is justified because these reviews could go either way. User results may not be as accurate due to bias.



MACHINE LEARNING

- Naive Bayes, Random Forests, Decisions Trees and Logistic Regression were all attempted to make predictions.
- Logistic Regression consistently delivered the best results, followed by Naive Bayes.
- Logistic Regression Cross-Validation had AUC means of over 90%.
- Confusion Matrix precision of unhelpful ratings were over 80%, and precision of helpful ratings were over 90%.
- Hyperparameter C tuned as 0.007742636826811269.

RESULTS

- Validation sets returned 91% accuracy.
- Test sets returned 88% accuracy.
- Changes in min_df and max_df led to minimal gains.
- Star ratings were predicted with greater accuracy. This can be used to flag users who gave the wrong amount of stars.
- Deep learning did not initially outperform Logistic Regression, but more tests and reviews deliver better results.
- The helpful rating metric can be applied to any product that counts votes (traditionally thumbs up / thumbs down).
- The next step is to use a similar pipeline to determine helpful reviewers.

REFERENCES

All reports, data wrangling, data analysis, and machine learning jupyter notebooks are on github.

https://github.com/coreyjwade/Helpful_Reviews

Publicly available Amazon datasets.

http://jmcauley.ucsd.edu/data/amazon/

My personal website.

coreyjwade.com