Team DataCart

Everyday Influencers:
Uncovering What Makes Amazon Reviews Helpful

Aikokul Abdivalieva
Brandon Escobar
Gabriela Gismondi
Nisso Hamralizoda

Reviews That Change The Internet

SW3K

No more winning for you, Mr. Banana!

Reviewed in the United States on March 3, 2011

For decades I have been trying to come up with an ideal way to slice a banana. "Use a knife!" they say. Well...my parole officer won't allow me to be around knives. "Shoot it with a gun!" Background check...HELLO! I had to resort to carefully attempt to slice those bananas with my bare hands. 99.9% of the time, I would get so frustrated that I just ended up squishing the fruit in my hands and throwing it against the wall in anger. Then, after a fit of banana-induced rage, my parole officer introduced me to this kitchen marvel and my life was changed. No longer consumed by seething anger and animosity towards thick-skinned yellow fruit, I was able to concentrate on my love of theatre and am writing a musical play about two lovers from rival gangs that just try to make it in the world. I think I'll call it South Side Story.

Banana slicer...thanks to you, I see greatness on the horizon.

59,056 people found this helpful

Helpful

"Those hysterical reviews certainly are creating more sales... We're all crying with laughter!"

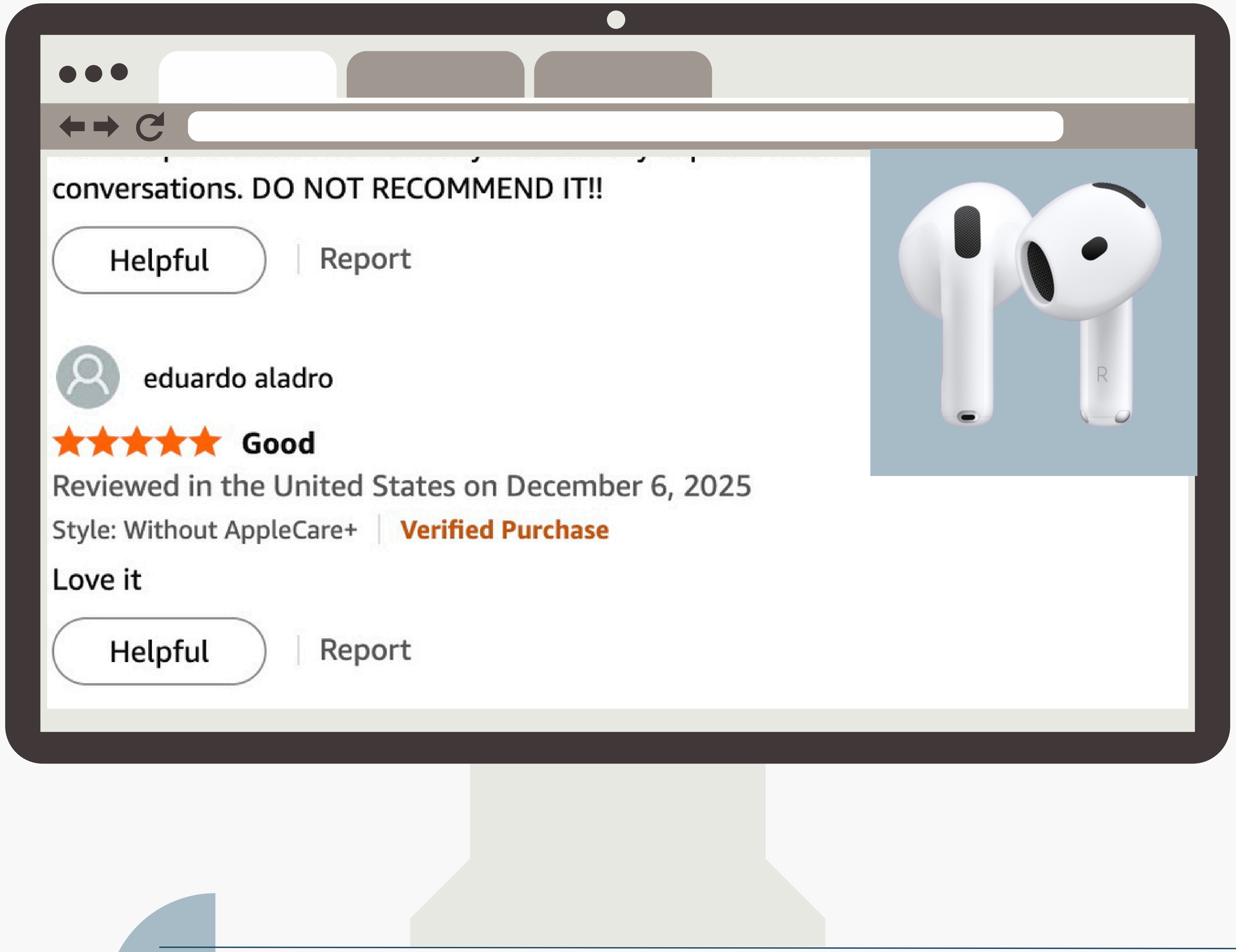
-Hutzler's Response

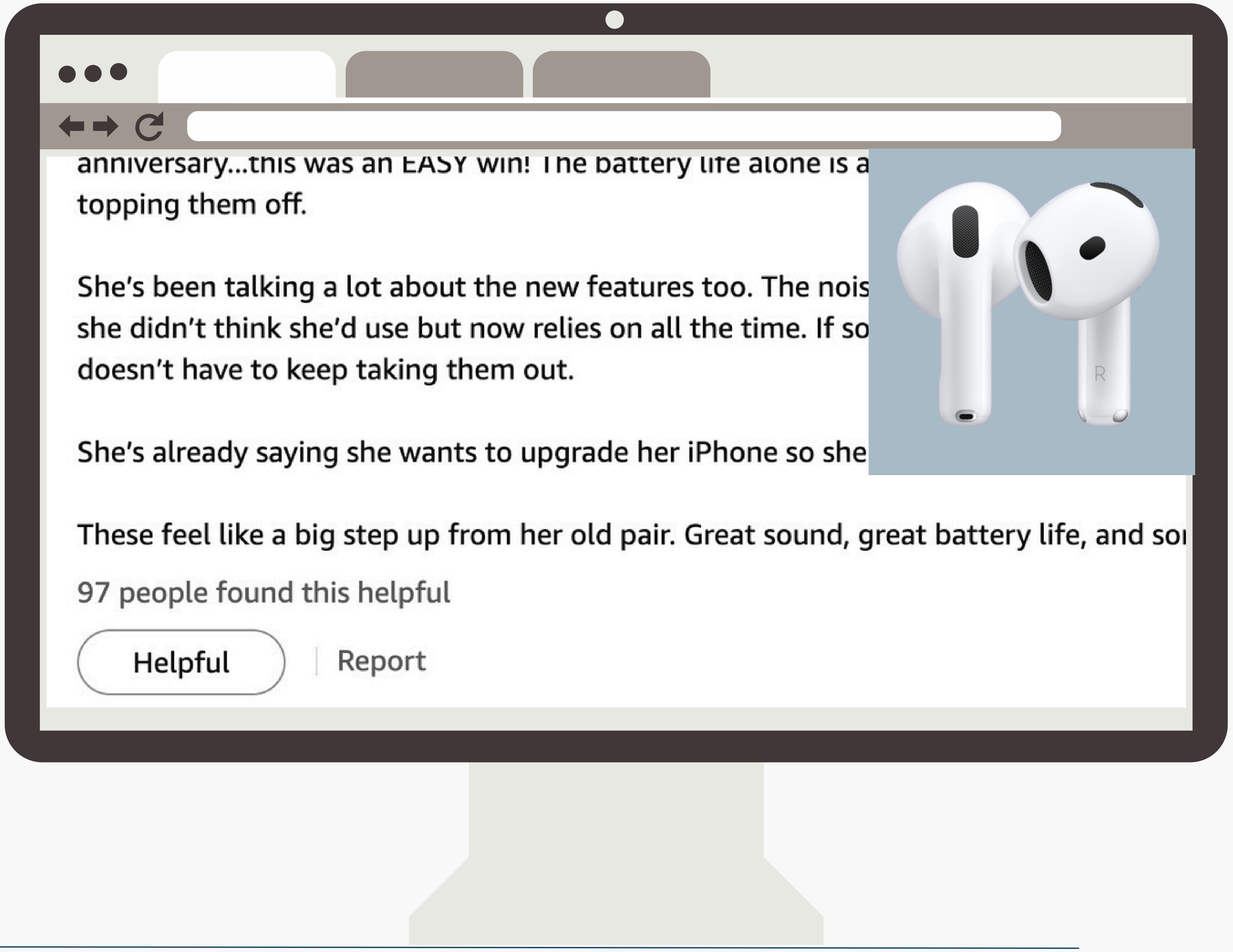
Cutting Through The Noise

- Ads
- Social media
- Promotions

- Average human validation

Why Some Reviews Matter More





Business Question

What differentiates highly helpful reviews from less helpful ones?

Business Value:

- Encourages detailed, authentic feedback → builds trust and conversions
- Reveals what makes a review useful and credible to drive sales
- Informs product and marketing teams on features that drive engagement

02

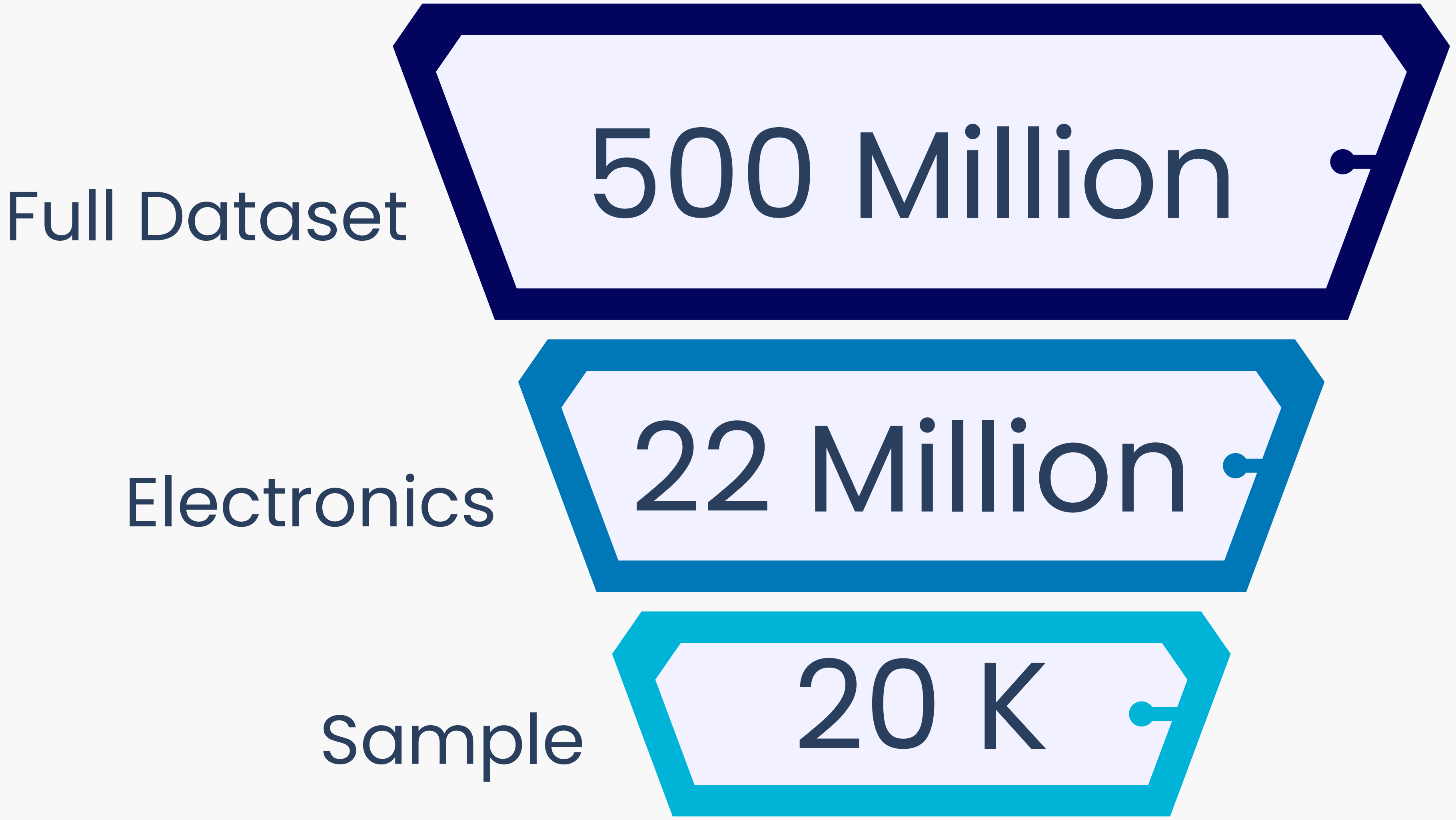
Data Preprocessing

2023 Amazon Reviews Dataset

https://amazon-reviews-2023.github.io/index.html

- McAuley Lab
- Over 500 million rows of data
- Focus on Electronics
- Amazon Review Dataset
- Amazon Metadata

Sampling & Validation



Stratified sampling by rating + helpfulness

Feature Engineering

Created:

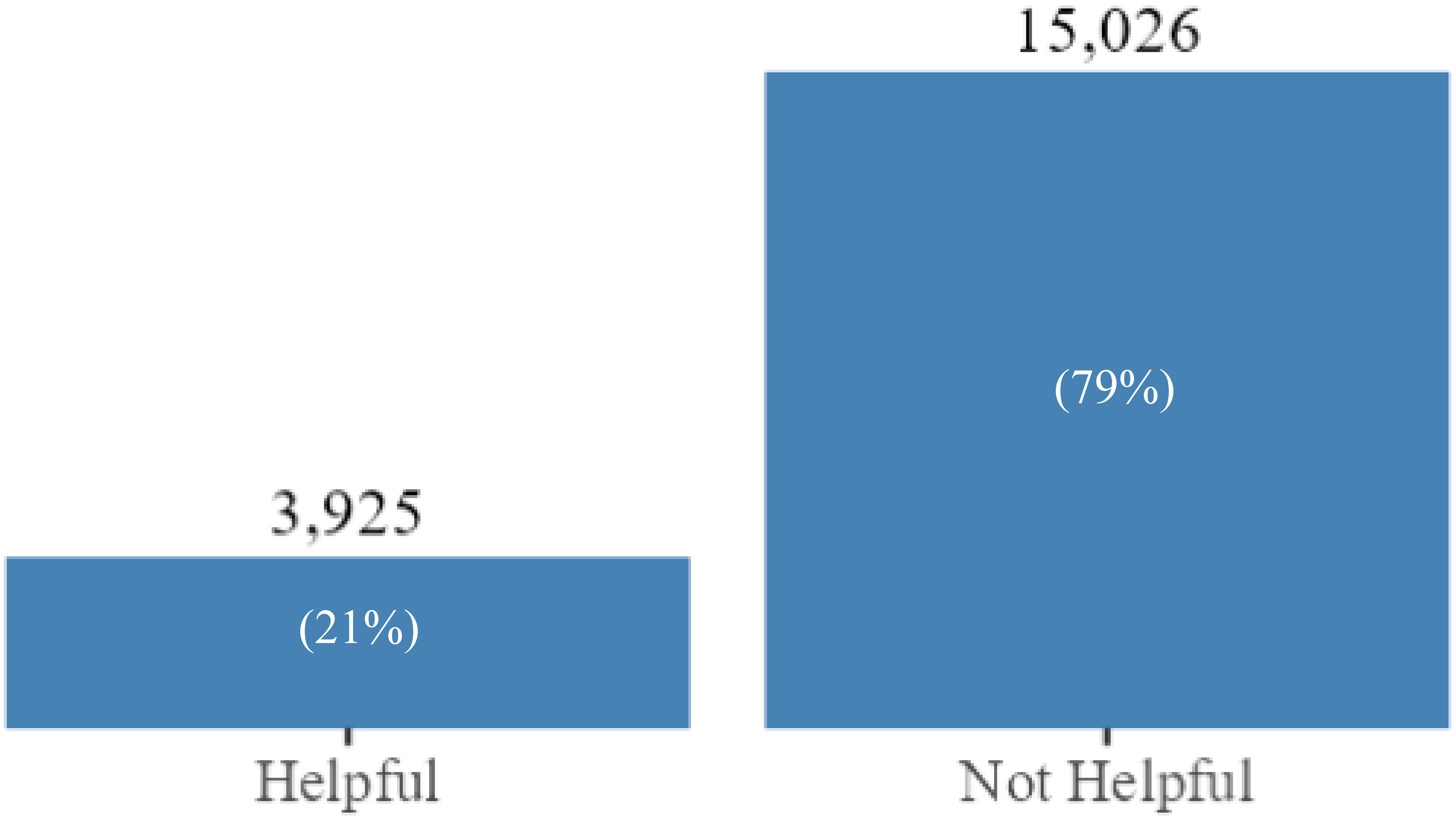
- Helpful vs Not Helpful (binary)
- Review & title word counts + log versions
- Logged helpful votes and price
- Includes image (yes/no)
- Brand groups (Big vs Small)
- Simplified product categories

04

Exploratory Data Analysis

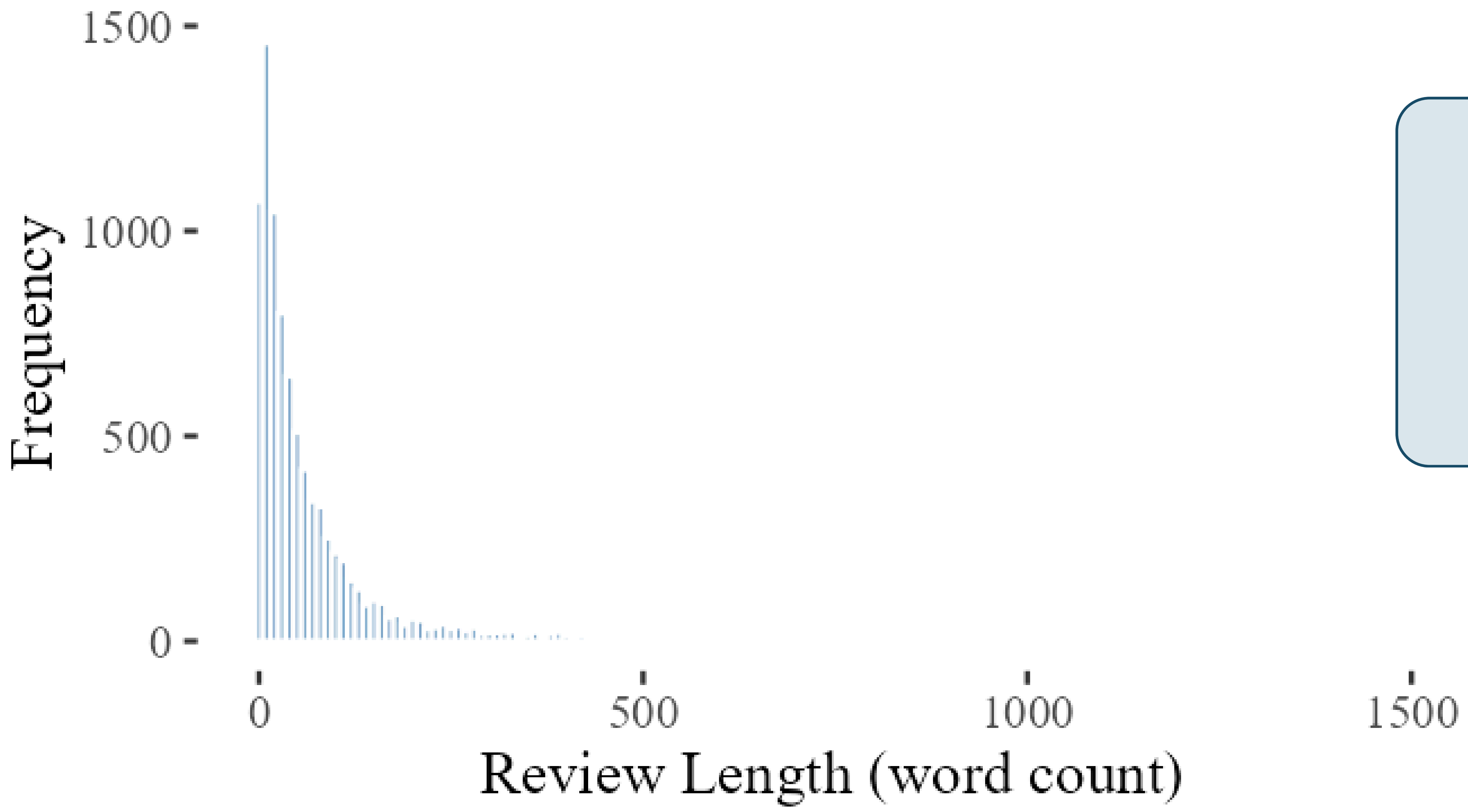
Behavior of Helpful Reviews

Distribution of Helpful Category



Word Counts

Histogram of Review Length by word count



Important Statistics:

Average = ***81*** *words*

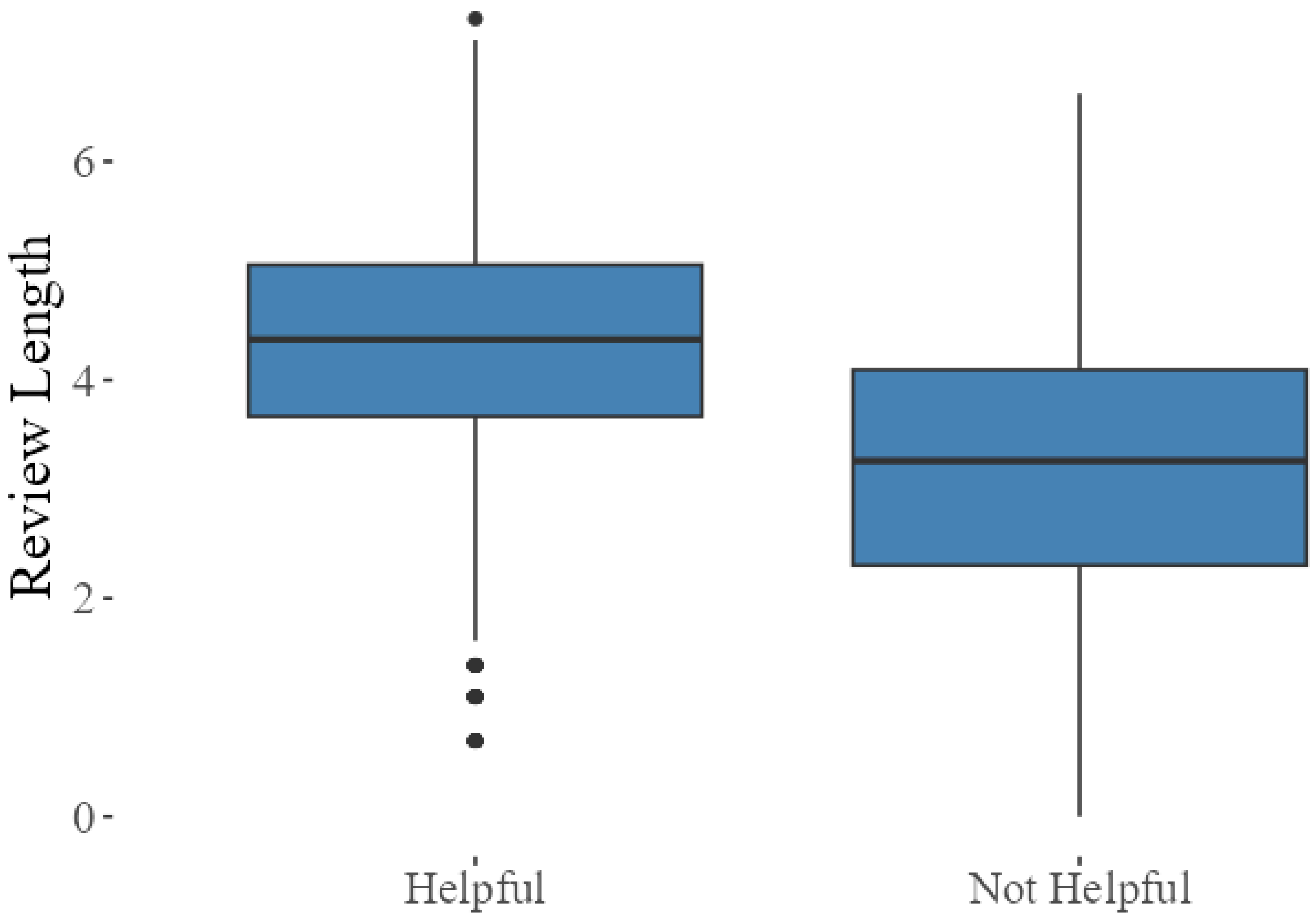
Min = ***1*** *word*

Max = ***1790*** *words*

Median = ***39*** *words*

Perceived Helpfulness

Average Review Length by Helpfulness Status



Welch Two Sample T-Test

- t-test statistic: 39.3
- Degrees of freedom: 4,670
- p-value < 0.01

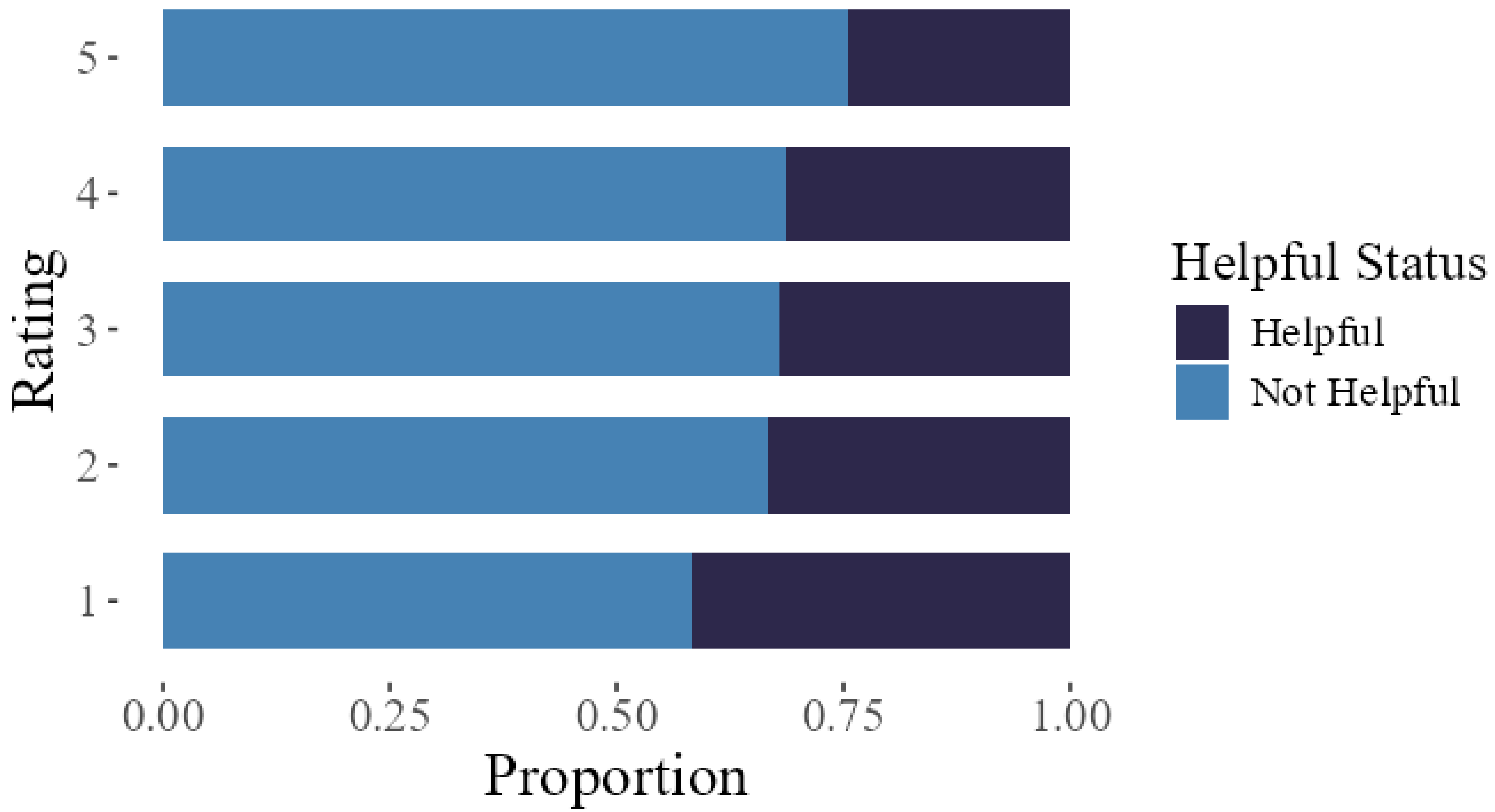
Reviews found helpful are **statistically significantly** longer than those found non-helpful.

Reviews With Images

Inclusion of Images vs Perceived Helpfulness of Reviews

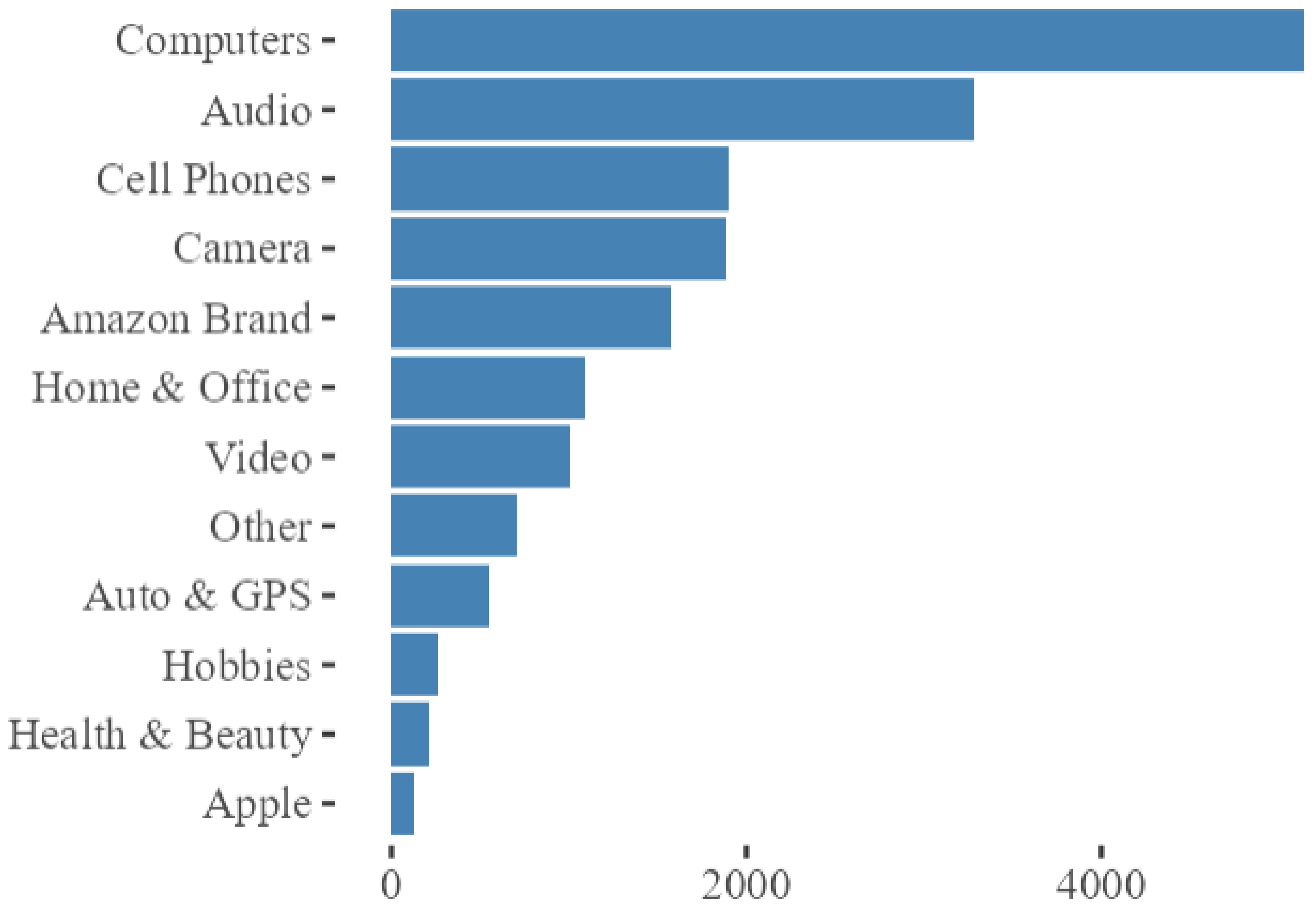
Mid-Range Ratings

Ratings vs Perceived Helpfulness of Reviews

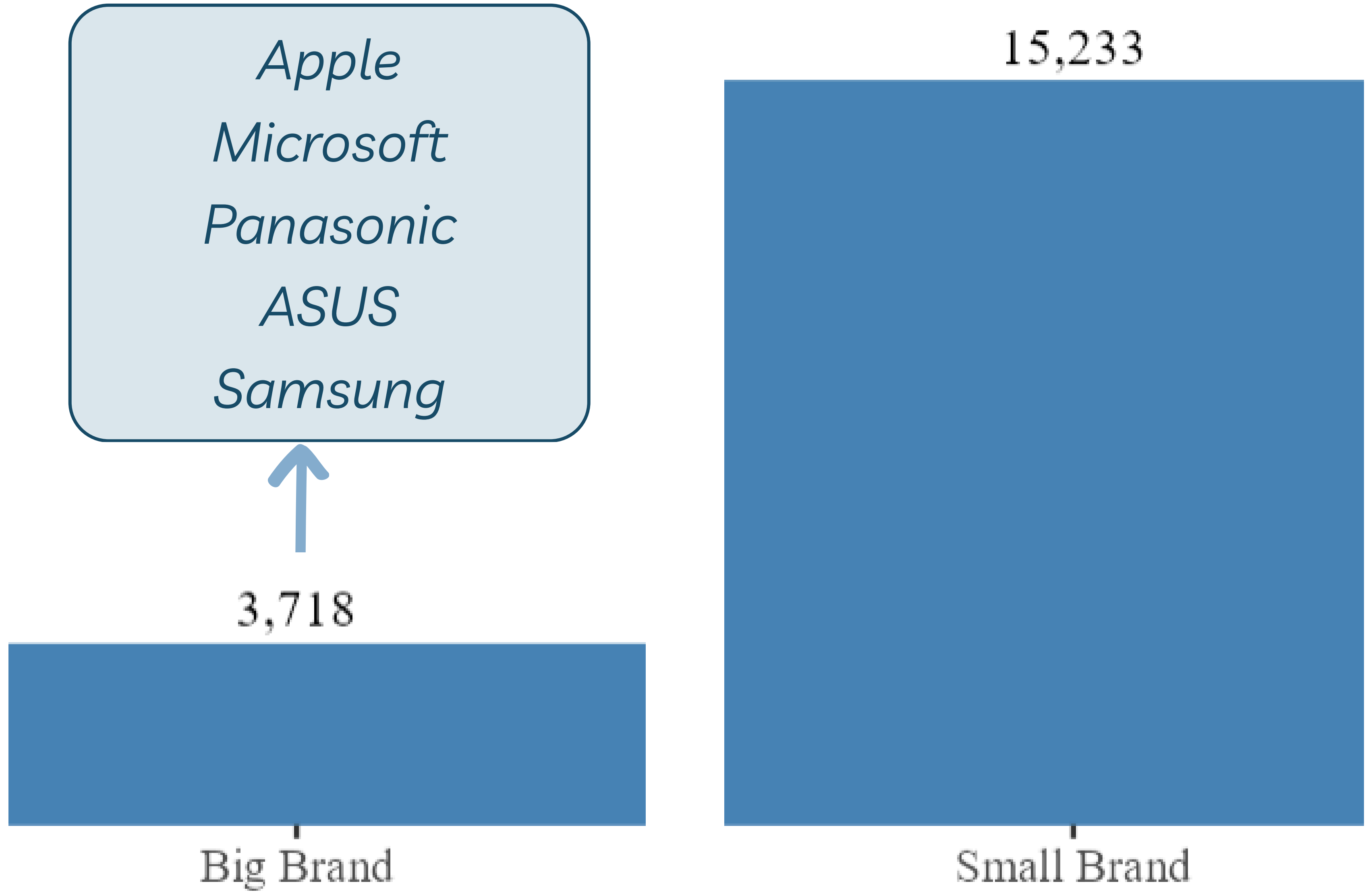


Product and Brand Types

Electronic Product Categories on Amazon



Type of Electronic Brands on Amazon



05

Modeling

Modeling Variables

Dependent variable:

Helpful Category

Independent variables:

Review Length (word count)
Title Length (word count)

Product Category

Includes Image

Type of Brand

Rating

Price

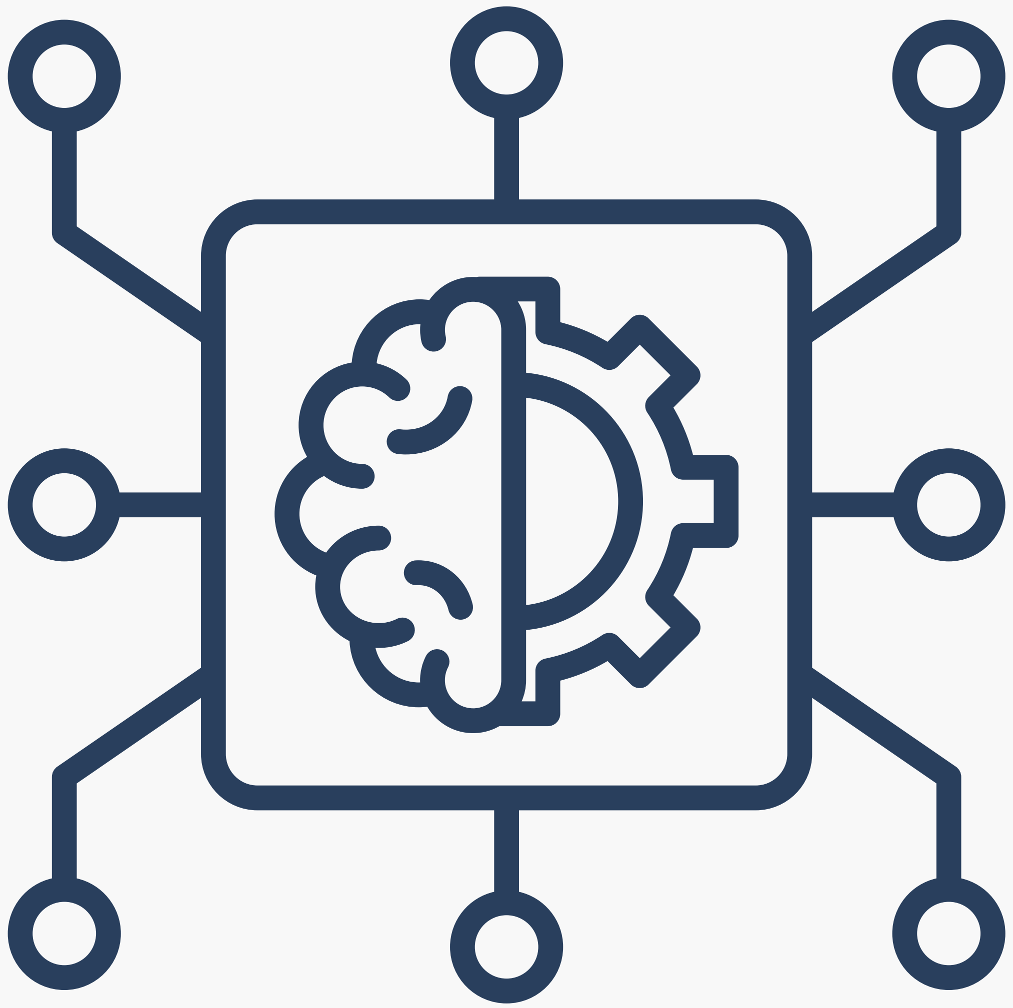
Modeling Overview

Why We Modeled

- Understand what drives "Helpful" reviews
- Compare different prediction methods
- Evaluate accuracy + interpretability

Models Explored

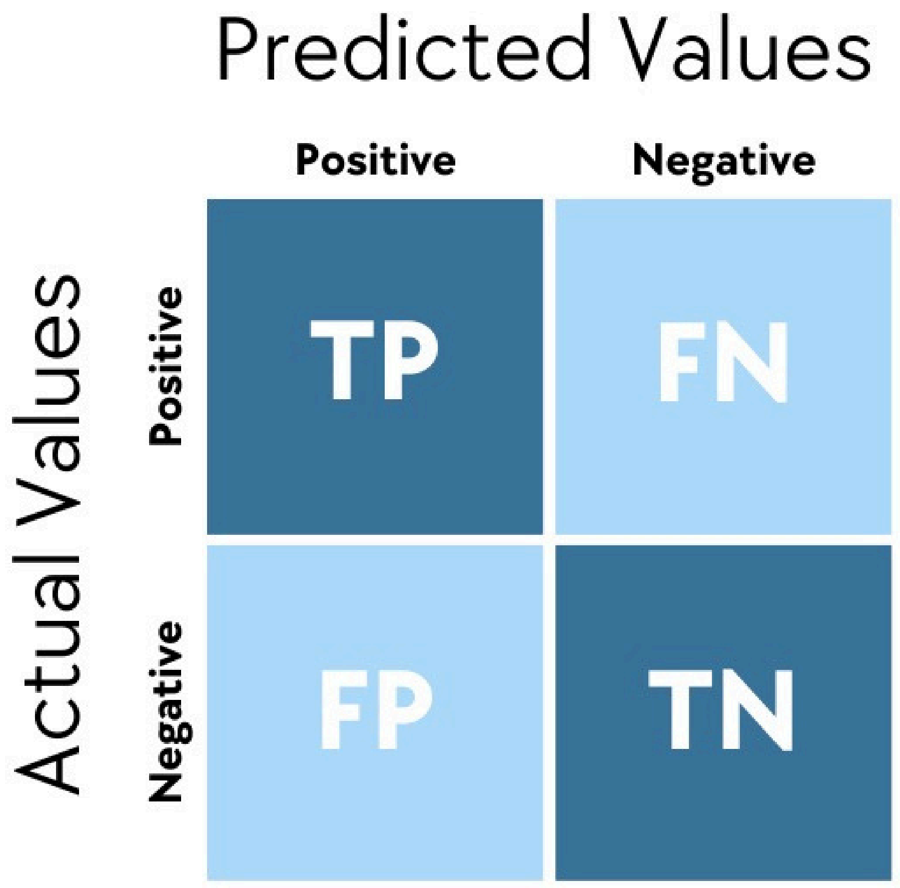
- Logistic Regression
- Weighted Logistic Regression
- Random Forests
- Ridge & Lasso

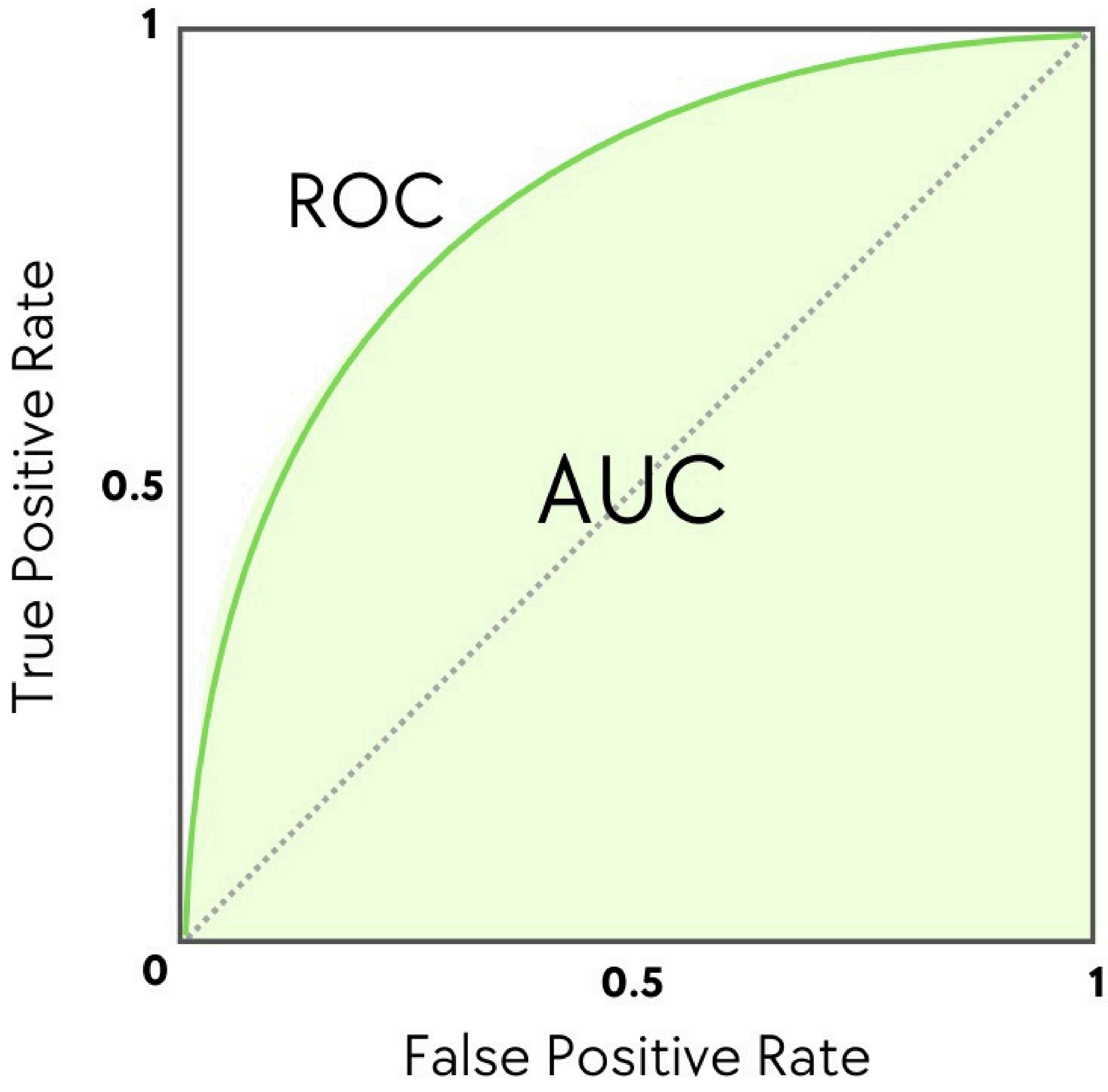


Results Methodology

Metrics Used

- F1 Score
- Precision & Recall
- ROC/AUC





06

Results & Findings

Our Best Models

Fit Statistics	Variable Selection Weighted Logistic Regression	Weighted Ridge Logistic Regression	Weighted Lasso Logistic Regression
ROC - AUC	0.76	0.66	0.66
Sensitivity (Recall)	0.71	0.68	0.68
Precision	0.36	0.37	0.37
F1 Score	0.47	0.39	0.39

Relevant Variables

Variables Included:

- Review Length (word count)
- Product Category
- Includes Image
- Rating
- Price

Best Model's Confusion Matrix:

Predicted \ Actual	Helpful	Not Helpful
Helpful	True Positive: **795**	False Positive: **1,434**
Not Helpful	False Negative: **335**	True Negative: **3,122**

70%

Best Model: Weighted Logistic Regression	
	Dependent variable:
Review Word Count (log)	0.697*** (0.015)
Category - Apple products	0.682*** (0.187)
Category - Audio	0.419*** (0.068)
Category - Auto and GPS	0.815*** (0.106)
Category - Camera and Photo	0.551*** (0.074)
Category - Cell phones	0.391*** (0.075)
Category - Computers	0.259*** (0.063)
Category - Health and Beauty	0.403** (0.160)
Category - Hobbies	0.769*** (0.138)
Category - Home and Office	0.748*** (0.085)
Category - Other	0.550*** (0.097)
Category - Video and TV	0.748*** (0.087)
Rating	-0.161*** (0.010)
Includes Images	0.759*** (0.067)
Price (log)	0.236*** (0.019)
Constant	-3.099*** (0.115)
Note:	*p<0.1; **p<0.05; ***p<0.01

Key Findings

Longer reviews are much more likely to be marked **helpful**

Reviews with **user-uploaded photos** are substantially more likely to be **helpful**

Higher star ratings are **less likely** to be voted **helpful.**

Reviews on **higher-priced products** are more often judged **helpful**, consistent with customers scrutinizing expensive purchases more closely.

Category context matters, but brand size does not

Client Recommendations

- Prompt reviews
- Incentivize image uploads

Limitations

- Analysis based only on Electronics category
- Unbalanced dataset
- Computational power

Future Opportunities

- Expand analysis to other categories
- Explore unorthodox
- Detecting fake reviews
- User demographics

Team DataCart

THANK YOU

Questions?

Bonus Slides

Missing Data & Outliers

Data primarily clean - except price.

Price missing? ~8,000 cases
→ Filled using Random Forest imputation
(using rating, avg rating, rating count, category)

Price outliers?
→ Kept only $5–$200
→ Removed extreme highs/lows

Text Cleaning & Tokens

Cleaned text (lowercase, decontraction, punctuation & HTML removal).

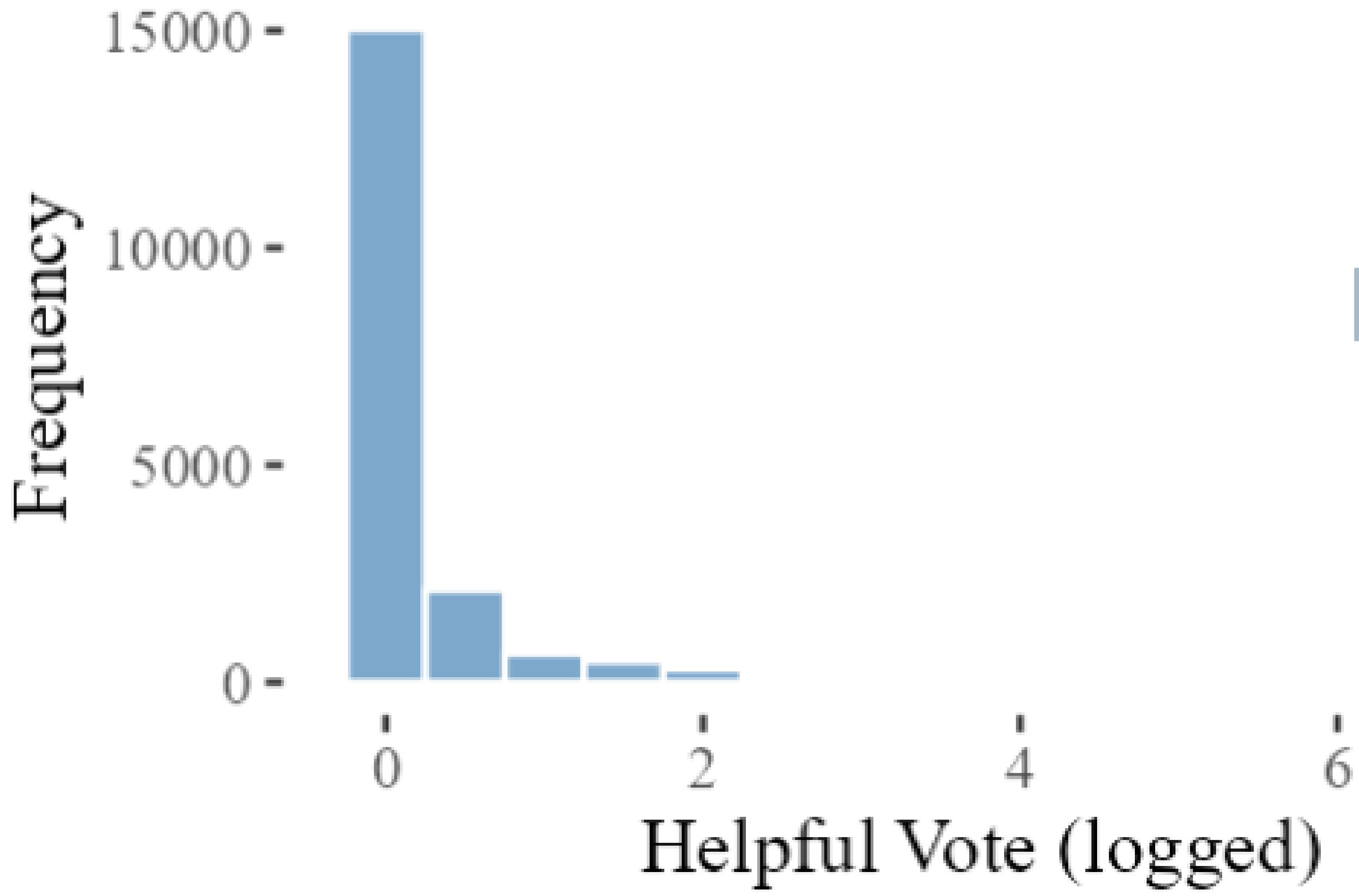
Created:

- Unigram tokens (lemmatized, stopwords removed)
- Bigram tokens (e.g., “not good”)

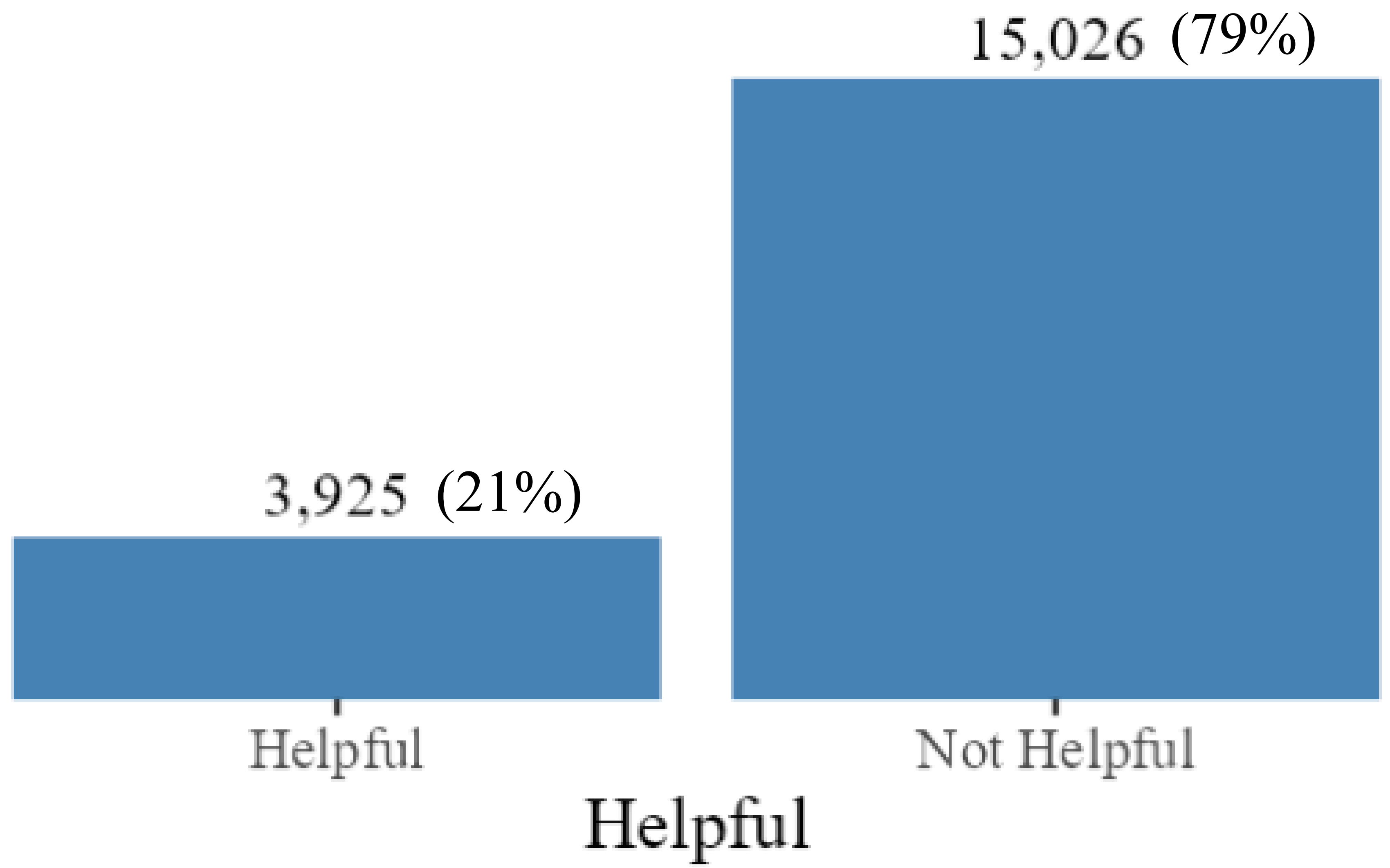
Final result: a clean, consistent, feature-rich dataset ready for modeling.

Helpful reviews behave like rare 'wins'

Histogram of Helpful Votes (Logged)

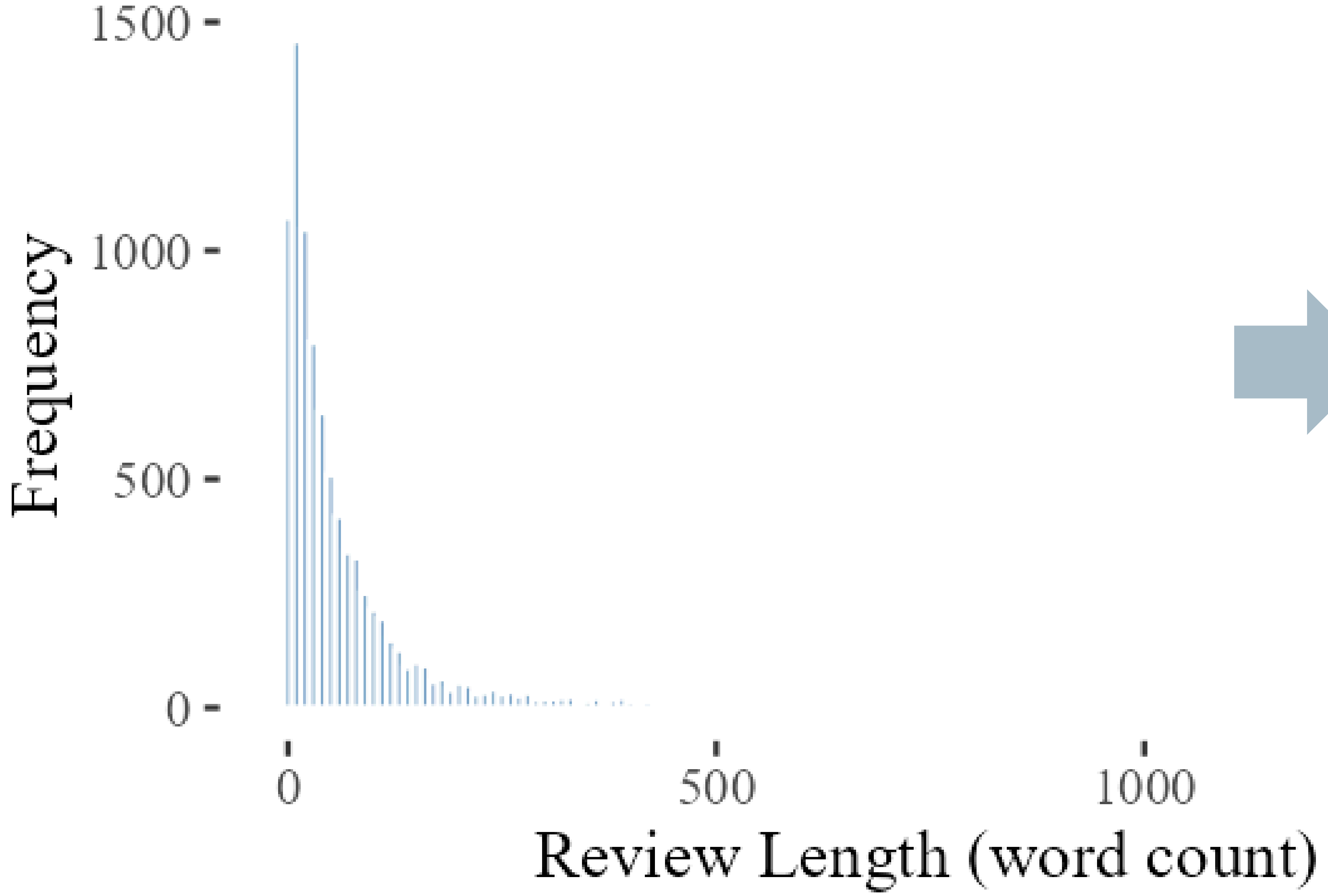


Distribution of Helpful Category

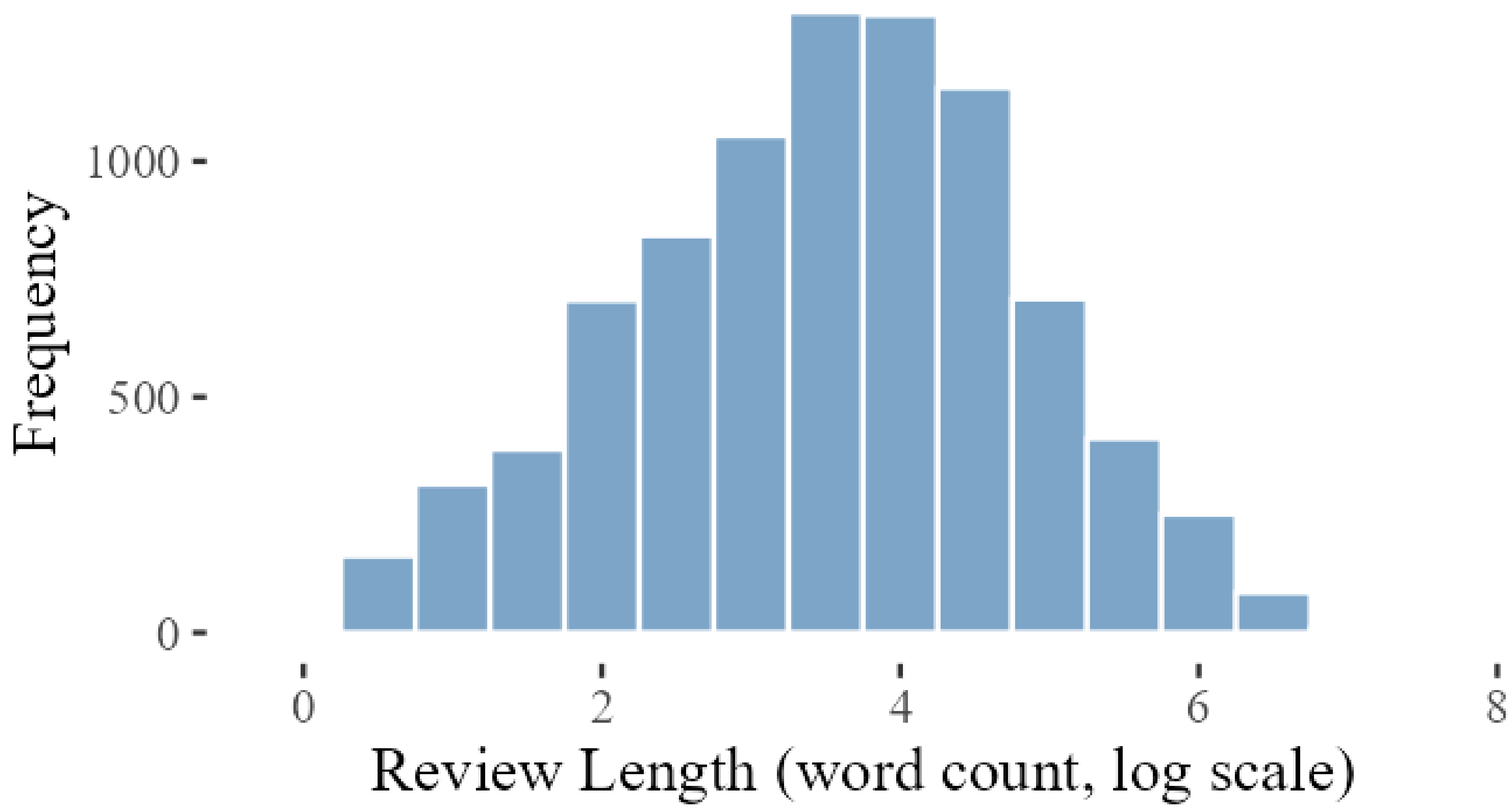


We used the log transformation of review length

Histogram of Review Length

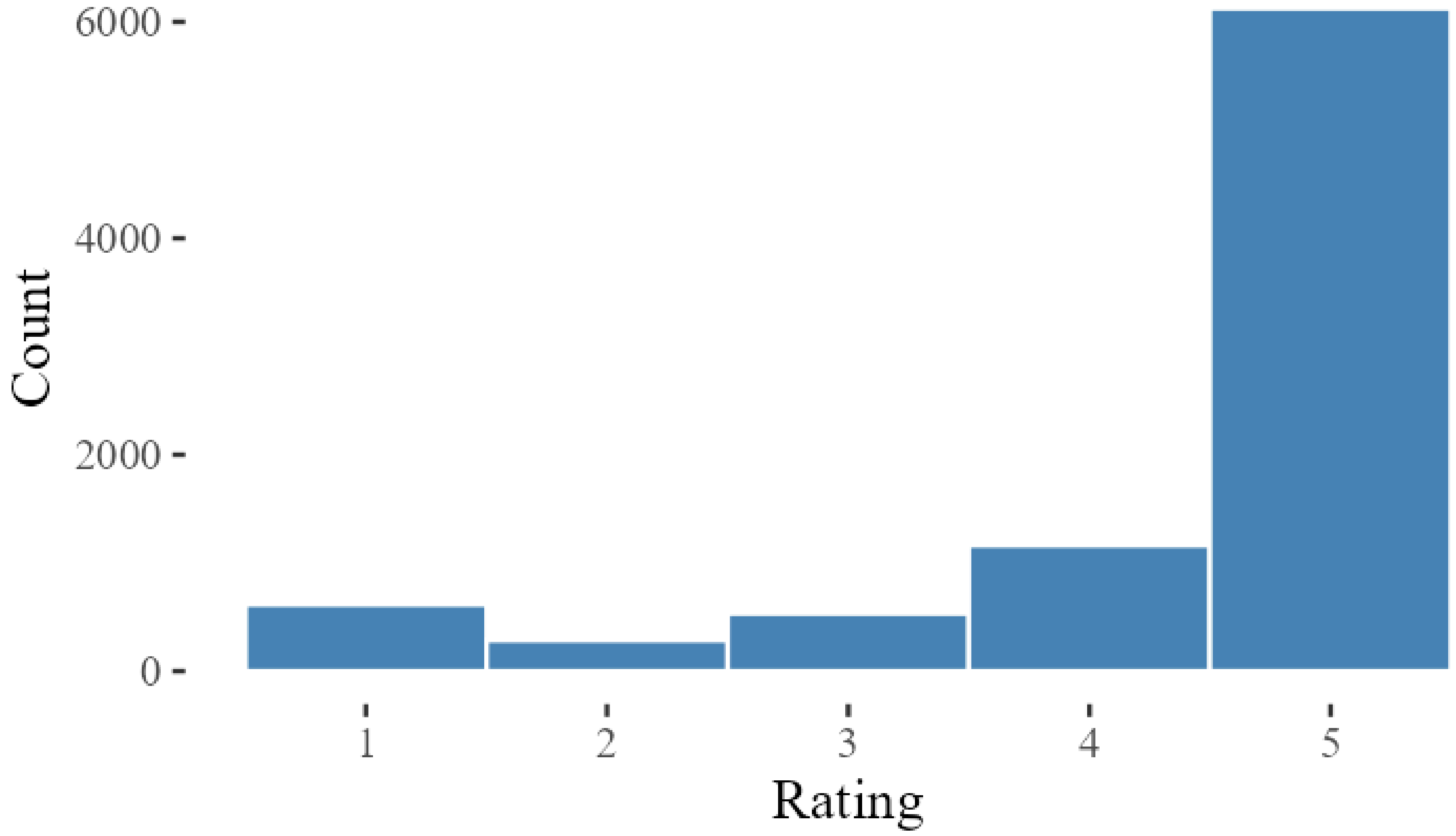


Histogram of Review Length (logged)

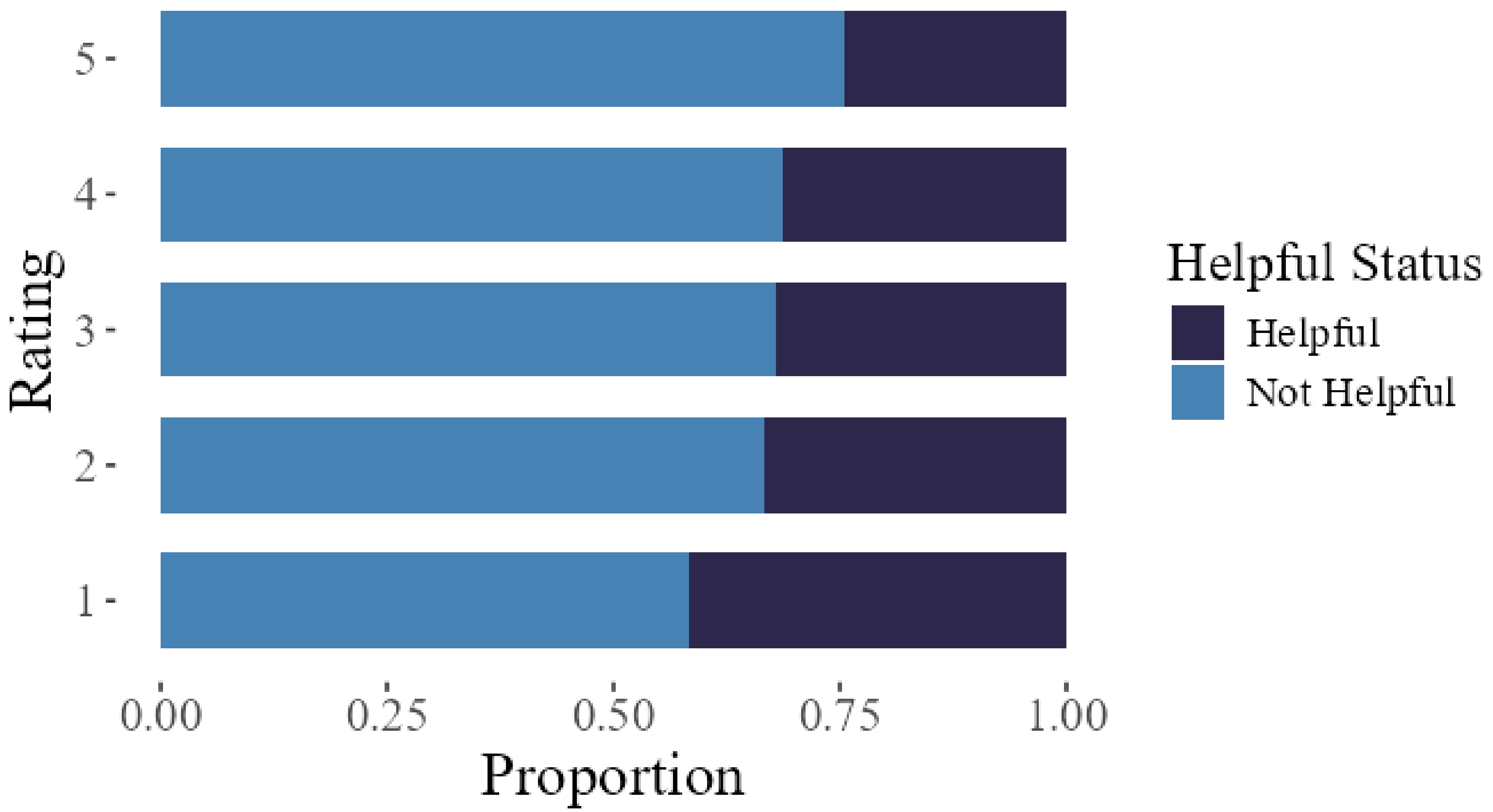


Mid-range ratings outperform 5-star reviews in helpfulness

Ratings Distribution in Electronic Products



Ratings vs Perceived Helpfulness of Reviews



What Each Model Contributes

Logistic Models

- Simple, interpretable
- Show how each factor affects helpfulness

Weighted Models

- Handle imbalance
- Better at detecting Helpful reviews

Random Forests

- Capture complex patterns
- Reveal strongest predictors

Penalized Models (Ridge and Lasso)

- Reduce overfitting
- Streamline variables