HelpfulLens EDA Appendix

Skew, Exposure, and Context for Helpful Votes

Team HelpfulLens

December 14, 2025

1 Key Takeaways

- Helpful votes are zero-inflated and heavy-tailed; a hurdle/log setup is appropriate.
- Popularity effects matter: cool/funny votes and author fans correlate with helpfulness.
- Content length, category, and city explain meaningful variation; temporal effects are mild.
- No single feature dominates; multi-signal models are necessary.

2 Figures (from `reports/eda/figures`)

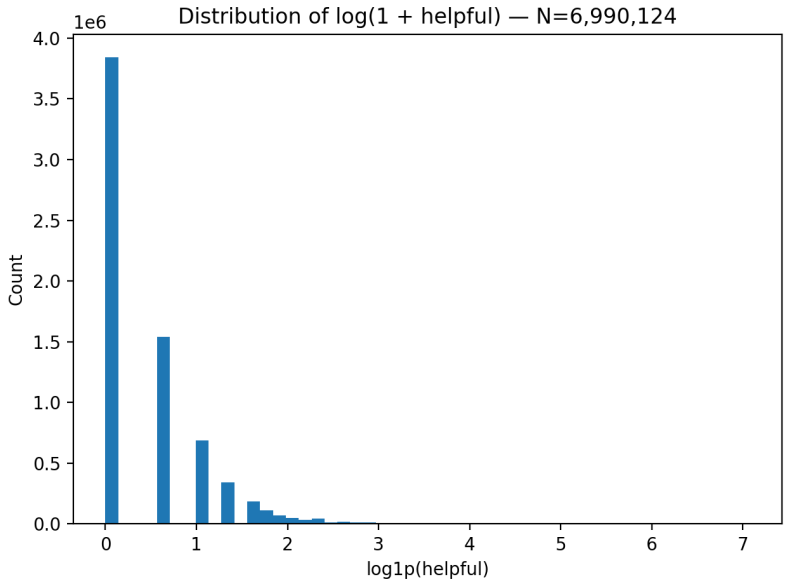

Figure 1: Distribution of $\log(1 + helpful)$.

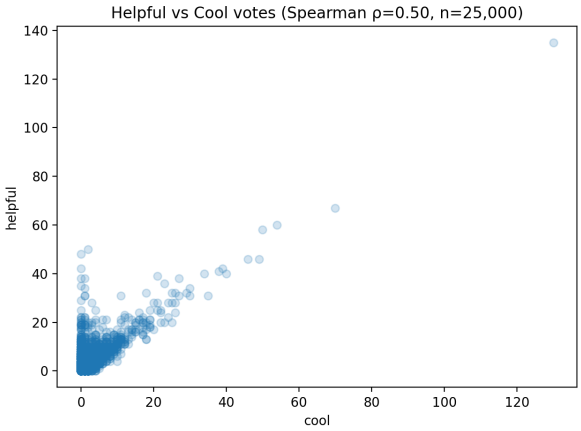

Figure 2: Helpful vs. cool votes (sample).

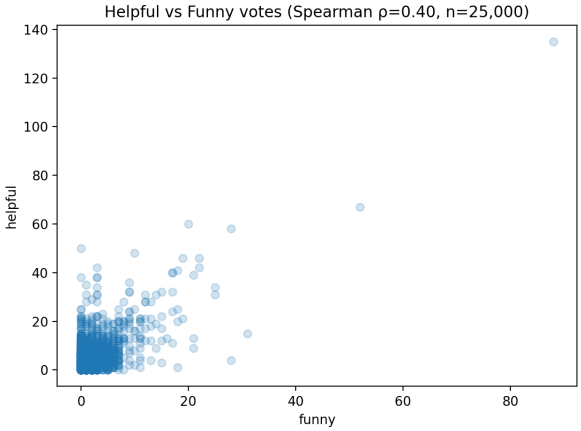

Figure 3: Helpful vs. funny votes (sample).

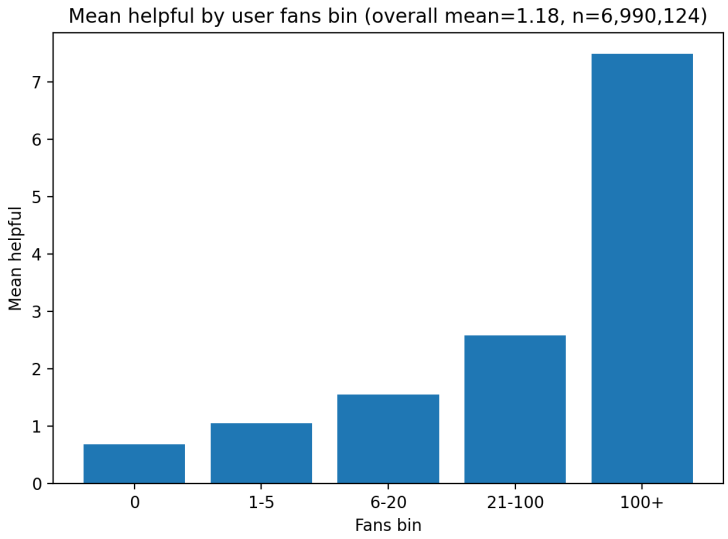

Figure 4: Mean helpful by user fan bins.

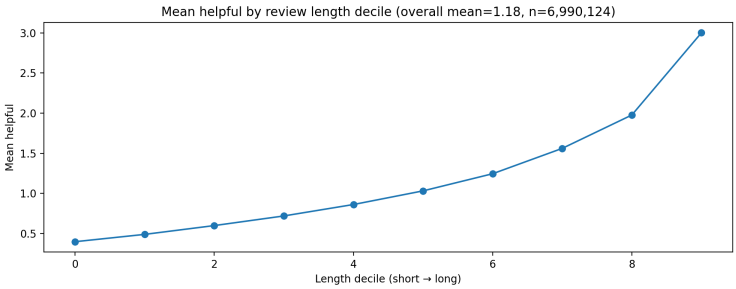

Figure 5: Mean helpful by review length decile.

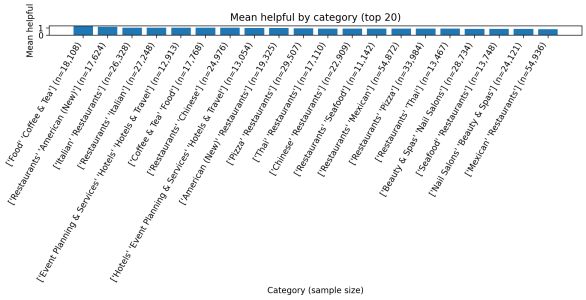

Figure 6: Mean helpful by category (top 20).

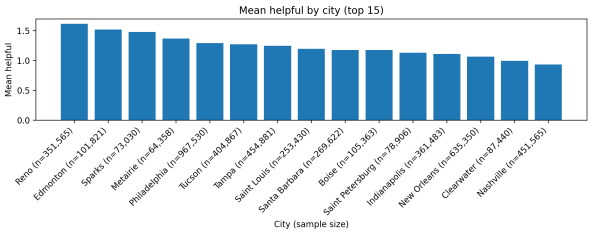

Figure 7: Mean helpful by city (top 15).

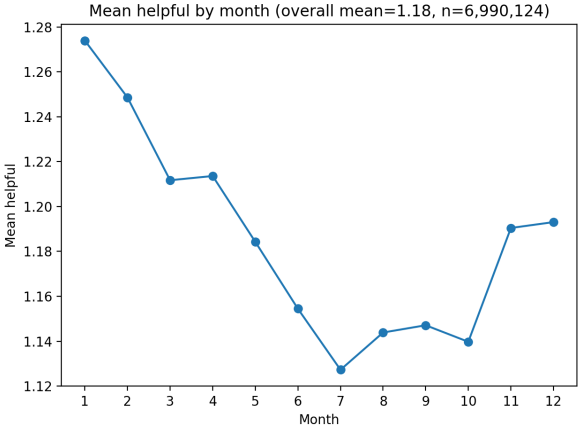

Figure 8: Mean helpful by month.

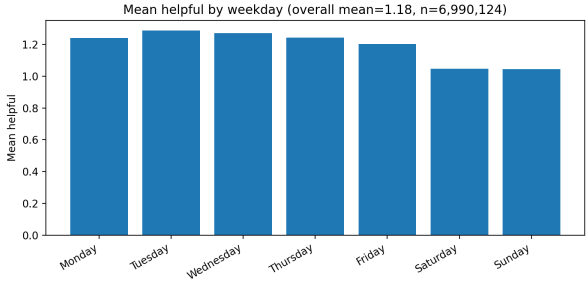

Figure 9: Mean helpful by weekday.

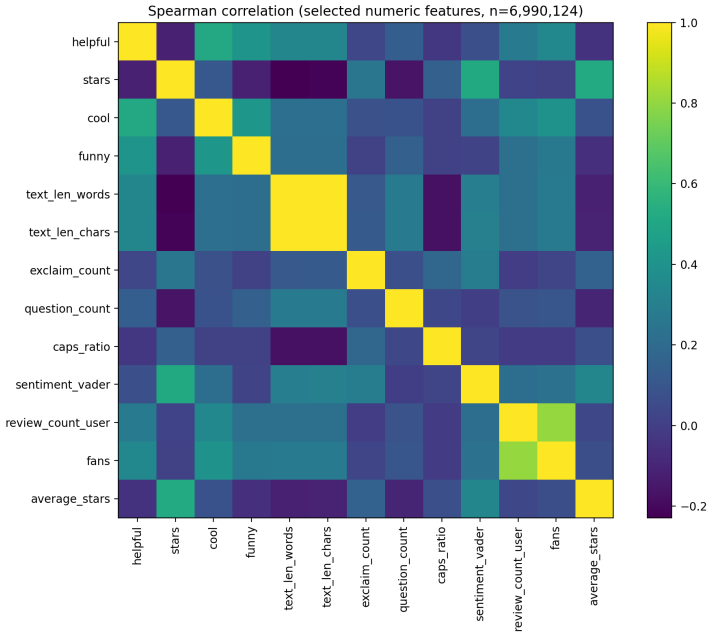

Figure 10: Spearman correlation among numeric features.

3 How to Regenerate

Run the refined EDA helper (uses cleaned parquet + normalized categories):

```
python scripts/eda_refined.py \
  --review-path data/cleaned/reviews_clean.parquet \
  --business-path data/cleaned/business_clean.parquet \
  --user-path data/cleaned/users_clean.parquet \
  --out-dir reports/eda_refined
```

This writes refreshed figures and `reports/eda_refined/eda_refined_summary.md`.

4 Refined EDA: Exposure vs. Helpfulness

This section summarizes the refined exploratory analysis built from cleaned parquets (7M reviews). Figures come from `reports/eda_refined/figures` (generated via `scripts/eda_refined.py`).

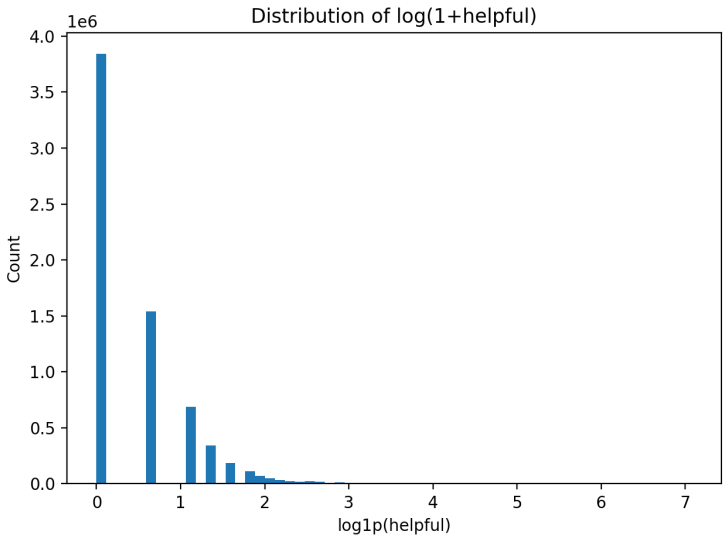

Figure 11: Heavy skew in helpful votes; most reviews have zero, a small tail exceeds 1,000.

Skew and zero inflation.

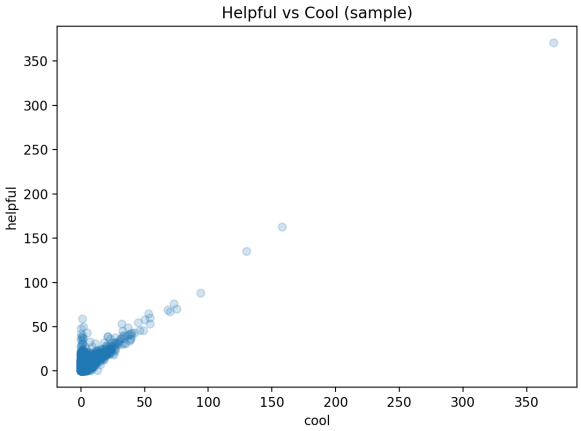

Figure 12: Helpful vs. cool votes (sample).

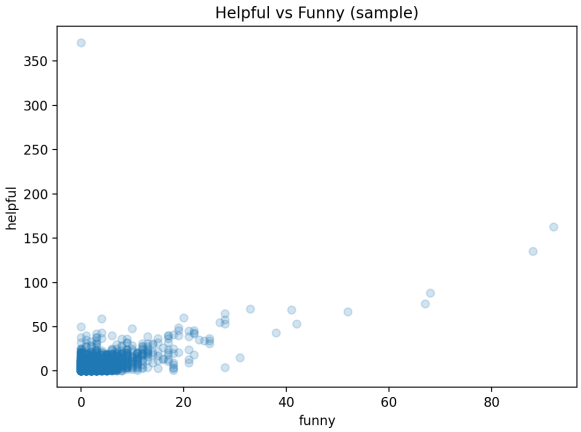

Figure 13: Helpful vs. funny votes (sample).

Popularity vs. helpfulness. Helpful correlates with cool/funny (shared exposure), so these are exposure controls, not targets.

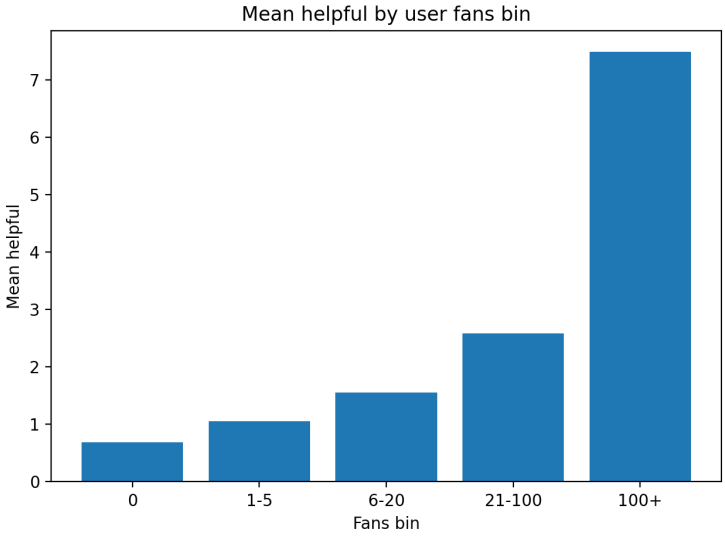

Figure 14: Mean helpful by user fan bins.

Author reputation. High-fan authors attract more helpful votes; fan count is a key confounder for visibility.

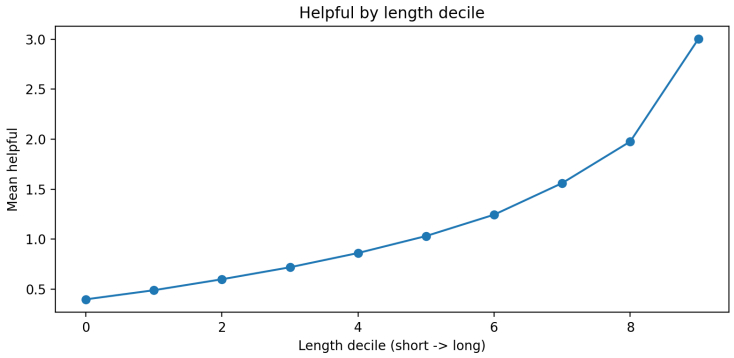

Figure 15: Mean helpful by review length decile.

Content length. Longer reviews earn more helpful votes (monotonic trend), but length alone is not sufficient.

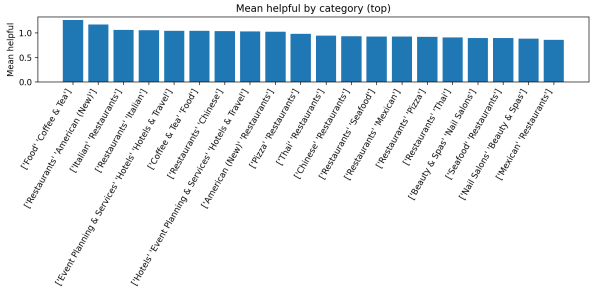

Figure 16: Mean helpful by business category (top 20).

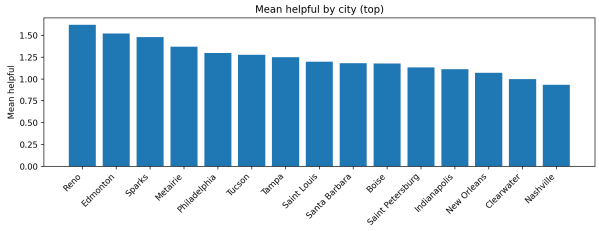

Figure 17: Mean helpful by city (top 15).

Place and category. Certain cuisines and locales drive higher helpfulness, reflecting audience/density effects.

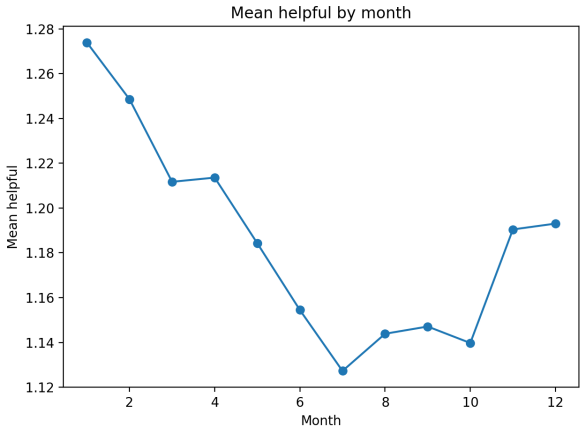

Figure 18: Mean helpful by month.

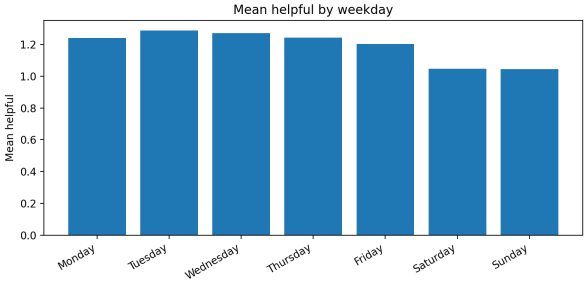

Figure 19: Mean helpful by weekday.

Temporal patterns. Seasonality/weekday effects are mild; included as controls.

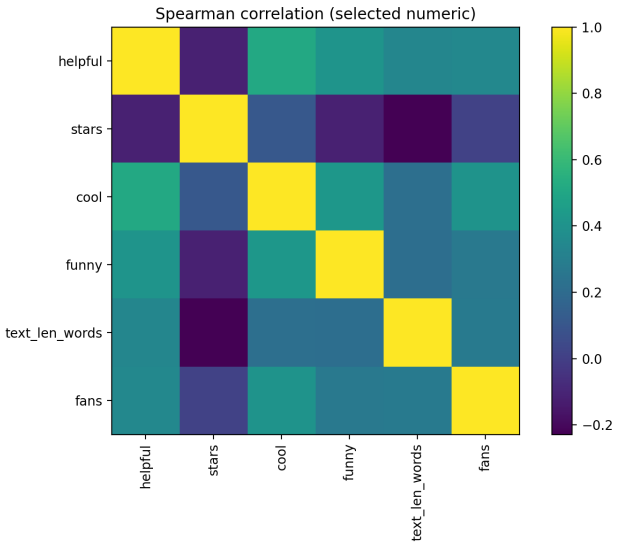

Figure 20: Spearman correlation among numeric features.

Correlations. Length correlates moderately with helpful; star ratings are weak; no single feature dominates, supporting multi-signal models (length + reputation + context).

Reproducibility. Regenerate figures with:

```
python scripts/eda_refined.py \
  --review-path data/cleaned/reviews_clean.parquet \
  --business-path data/cleaned/business_clean.parquet \
  --user-path data/cleaned/users_clean.parquet \
  --out-dir reports/eda_refined
```