Analyzing Product Ratings on Amazon

Danielle Ferstl, Jethro Chan, Xuehan Wang, Udaya Gadparthi, Brady Tilkens

SUMMARY OF DATA

Kaggle Dataset: Amazon customer reviews from 1995 to 2015

- 37 TSV files, each file for a category of product:
- Ex. apparel, tech, groceries, toys

Questions to Explore

- What is the highest rated category of product?
- What types of products received the most helpful reviews?
- Is there a relationship between star rating and helpfulness?

Relevance

- Help businesses improve product quality, manage customer expectations
- Help customers better interpret star ratings and helpfulness when purchasing product

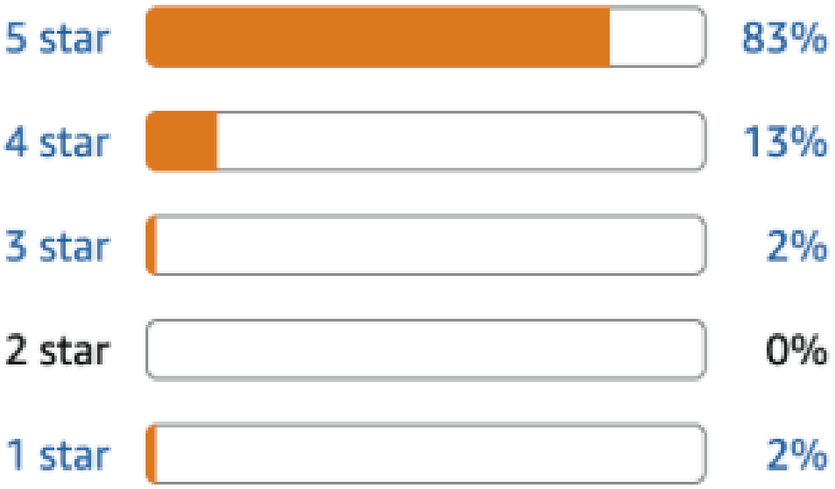
DATA CLEANUP

Import dataset from Kaggle to CHTC group directory	Cleanup Relevant columns:	shell script
Each file consisted of millions of reviews	• “review_id”: unique ID of review • “product_title”: product title • “product_category”: broad product category that can be used to group reviews • “star_rating”: 1-5 star rating of review • “helpful_votes”: number of helpful votes • “total_votes”: number of total votes review received	to clean each TSV file and isolate relevant columns

Customer reviews

4.7 out of 5

138 global ratings



How customer reviews and ratings work

Review this product

Share your thoughts with other customers

Write a customer review

Top reviews from the United States

Bryce Williamson

It is exactly what was articulated

Reviewed in the United States on April 14, 2022

Verified Purchase

And that's all I wanted :)

Helpful | Report

Stormy Jones

Perfect.

Reviewed in the United States on September 29, 2021

Verified Purchase

Exactly what we wanted.

Helpful | Report

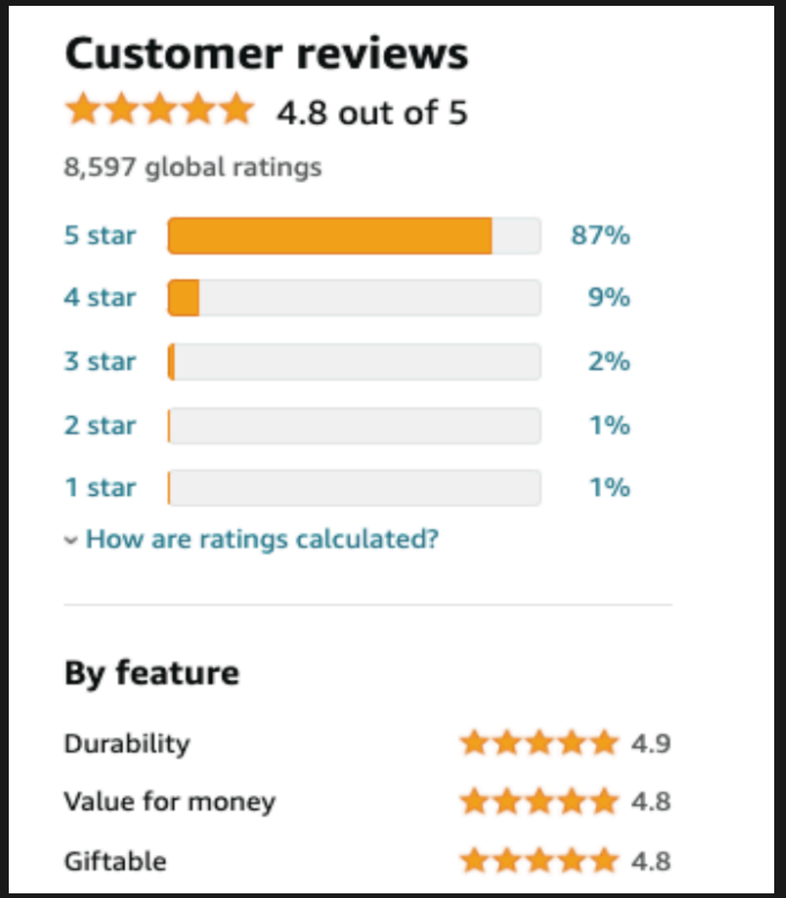
Question 1:

Highest and Lowest Ratings

Script: analyze_ratings.R

- Calculates the average star rating per product category

Condor Job: submit_ratings.sub



Question 2:

Most Helpful Reviews

Script: analyze_helpfulness.R

- Calculates average helpful votes per product category
- Sorts categories from most to least helpful

Condor Job: submit_helpfulness.sub

Question 3:

Correlation Between Rating and Helpfulness

Script: analyze_correlation.R

- Calculates correlation between star rating and helpful votes.

Condor Job: submit_correlation.sub

Combining Results

Merged outputs using cat command:

cat *_avg_ratings.tsv > all_avg_ratings.tsv

cat *_avg_helpful.tsv > all_avg_helpful.tsv

cat *_correlation.txt > all_correlations.txt

Created summary files for further analysis and visualization

.log Analysis (Per Job)

Merged outputs using cat command:

Memory Used: 2 GB (2048 MB)

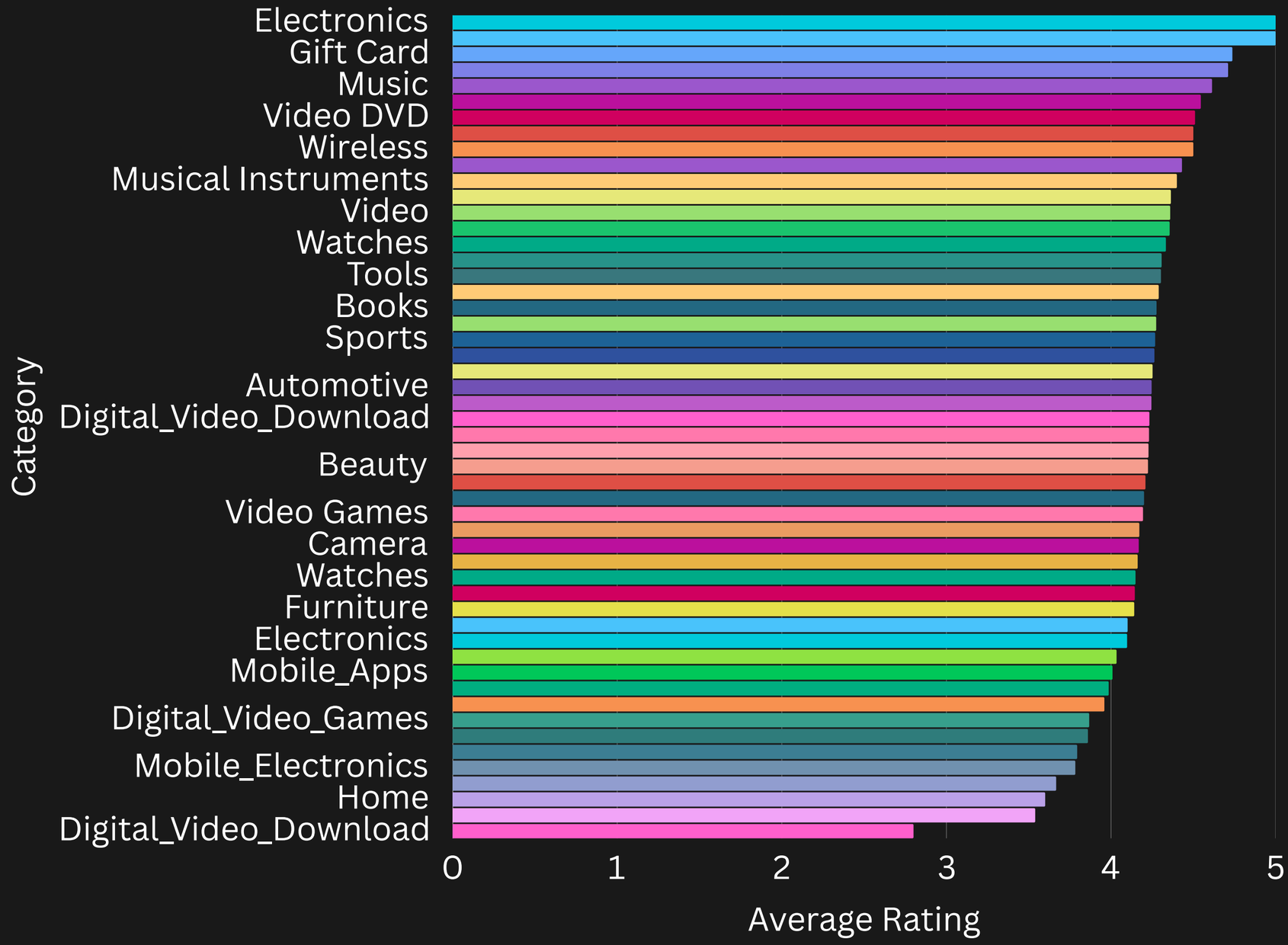
Disk Used: 2 GB (~2,098,305 KB)

CPUs: 1

Time to Execute (running): 13 seconds

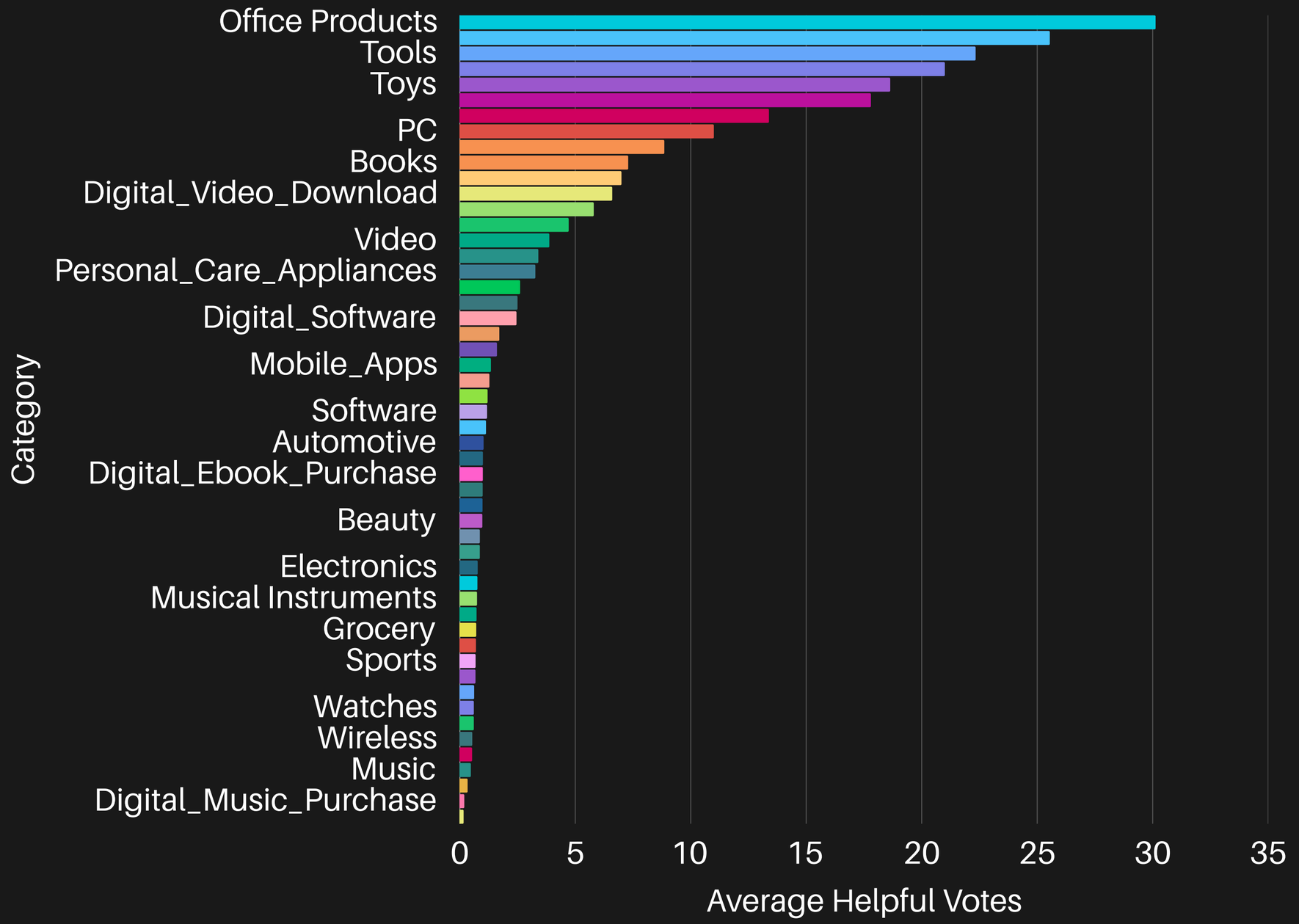
Total Slot Time (Time Slot Busy): 78 seconds

Average Star Rating by Category

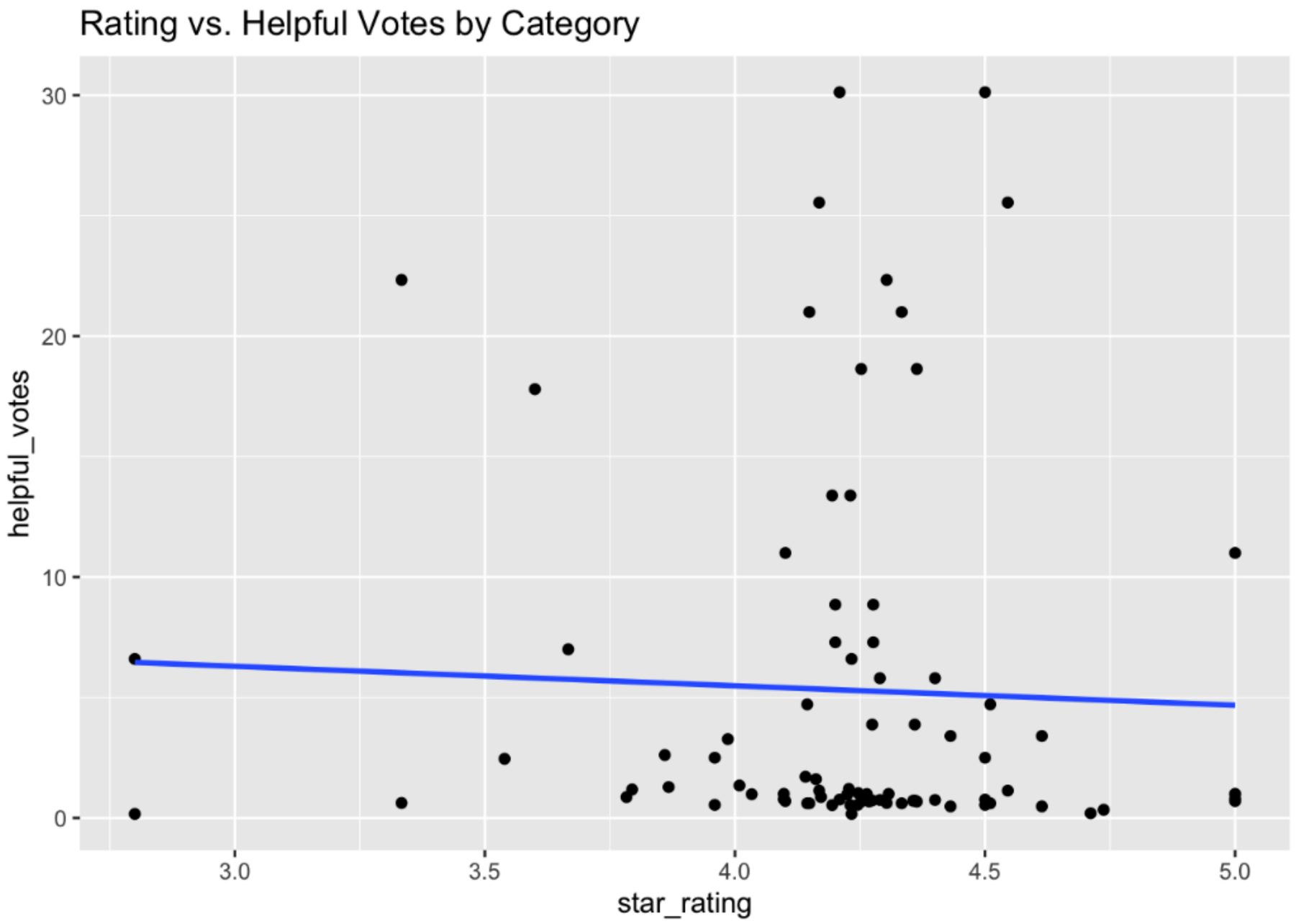


- Digital_Video_Download products have the lowest average star rating (≈ 2.8)
- Electronics has the highest average star rating (5.0)

Average Helpful Votes by Category



- Digital purchases (e.g. Digital_Music_Purchase, Digital_Video_Download) consistently rank lowest in helpfulness, often below 1 vote on average.
- Office Products ranked the highest with an average of 30.12 helpful votes



Very weak negative correlation (r = -0.04, R^2 = 0.002)

Star ratings do NOT predict review helpfulness

Helpful reviews likely focus on depth and relevance, not positivity

CONCLUSION

Highest Star Rating: Gift Card
Lowest Star Rating: Digital/Video/Download

Most Helpful Votes: Office Products
Least Helpful Votes: Digital Music Purchase

Correlation between star rating and helpful votes was found to be insignificant.
Positive correlation between product complexity and buyer engagement.

BROADER PERSPECTIVES

- Illustrates the significance of informational infrastructure in retail.
- Complex product consumer feedback loop.
- Evolving consumer expectations for quality of products received.
- Value in future research that encompass price points, demographics, and geographical regions.
- Value in future research related to review linguistics.