An Analysis of Amazon Reviews

Joao Carreira

Outline

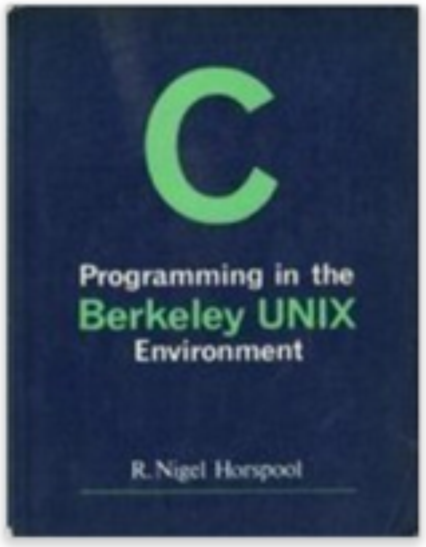
- Dataset and Methodology
- Sanity checks
- Dataset Analysis
    1. Characterization
    2. Products
    3. Users/Reviews
- Conclusion

Dataset - Overview

- Amazon founded in 1994
- Amazon reviews 1995-2013 (18 year span)
- 34M reviews, 7M users, 2M products
- 35Gb of uncompressed data
- Dataset is available for research purposes [1]
- An analysis of review text is available [2]

[1] https://snap.stanford.edu/data/web-Amazon.html
[2] J. McAuley and J. Leskovec. Hidden factors and hidden topics: understanding rating dimensions with review text. RecSys, 2013.

Dataset - User Reviews



product/productId: 0131097601
product/title: C Programming in the Berkeley Unix Environment
product/price: unknown
review/userId: A1KLBWKUQHSQVW
review/profileName: Eugene Mah "physics geek"
review/helpfulness: 0/0
review/score: 4.0
review/time: 994291200
review/summary: indispensible title on my computer bookshelf
review/text: This has been one of those books that I constantly refer to. Not only is it good for learning some of the unique C things that apply to Unix, but you can also learn how to get around in Unix. This is the book I learned C from, and it's still one of the first ones I go to when I need to refresh my brain about something.

Dataset - Other Records

1. Product Brand

B0000C2LFS Gifted Horse

2. Product Categories

0131097601
Books, Computers & Technology, Microsoft, Development, C & C++ Windows Programming
Books, Computers & Technology, Programming, APIs & Operating Environments, Unix
Books, Computers & Technology, Programming, Languages & Tools
Books, Computers & Technology, Software
Books, Education & Reference
Books, Science & Math, Mathematics

3. Product description

product/productId: 1878972405
product/description: Portuguese author Fernando Pessoa (1888-1935) published little in his lifetime, but his rediscovery
in the 1990s has been as central to postmodernism as the rediscovery of Kafka in the 1950s was to modernism.

4. Related products

B000K85RMI also purchased 0684803305 0805062904

Methodology

- Exploratory analysis of the dataset
- This analysis focus on products and users
- No textual analysis - NLP - of reviews
- Perl + R
- Code, graphs and slides available @

  github.com/jcarreira/amazon-study

Sanity Checks

Sanity Check	Description	Check ?
Correct timestamps	Time between 95 and ‘13	✗
Helpfulness <= 1	Helpfulness factor at most 1	✗
Price	Price is positive (and reasonable)	✗
Score 1-5	Score is a 1-5 value	✓
Review entries complete	All reviews have all entries	✗
Product price fluctuation	Different reviews for the same product may have different prices	✗
Review product title consistency	Review product title matches product title	✓
Daily activity cycle	Less reviews during night and more during day	✗
Products categories	All products have categories	✗

Sanity Checks

- Timestamps: Some are missing (e.g., “-1” entries)
- Timestamp hour at 4pm or 5pm
- Helpfulness: Some factors are > 1

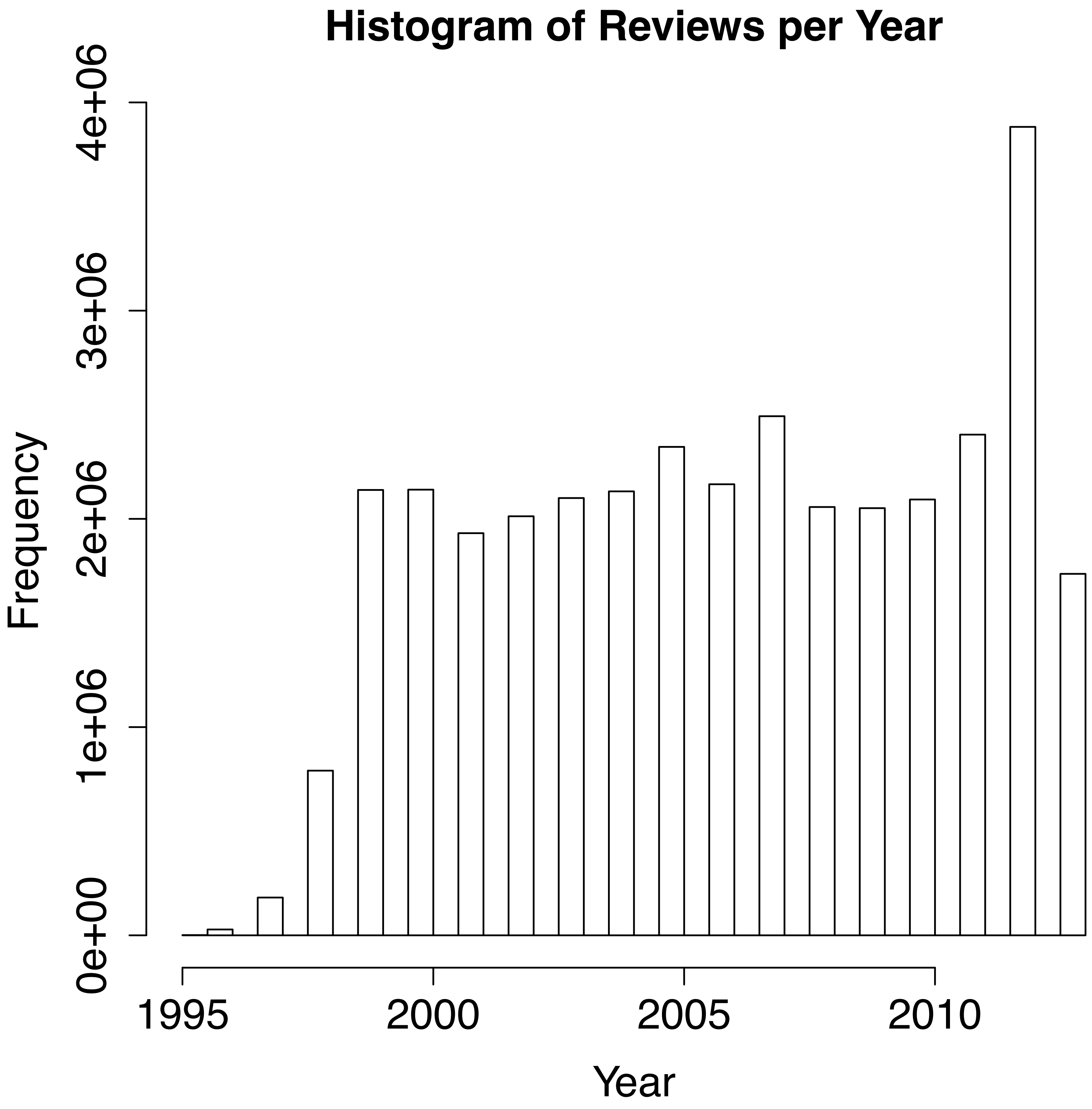
```
product/productId: 1930771142
product/title: You Can Have Your Cheese and Eat It Too!
product/price: unknown
review/userId: A1VYC3XNQU72RF
review/profileName: William Cottringer
review/helpfulness: 2/1
```

- Price: Some products have price 0$. Others “unknown”
- Product price: prices are constant through time — not what happens in reality
- Some reviews do not have text (just summary)
- Some products have no category

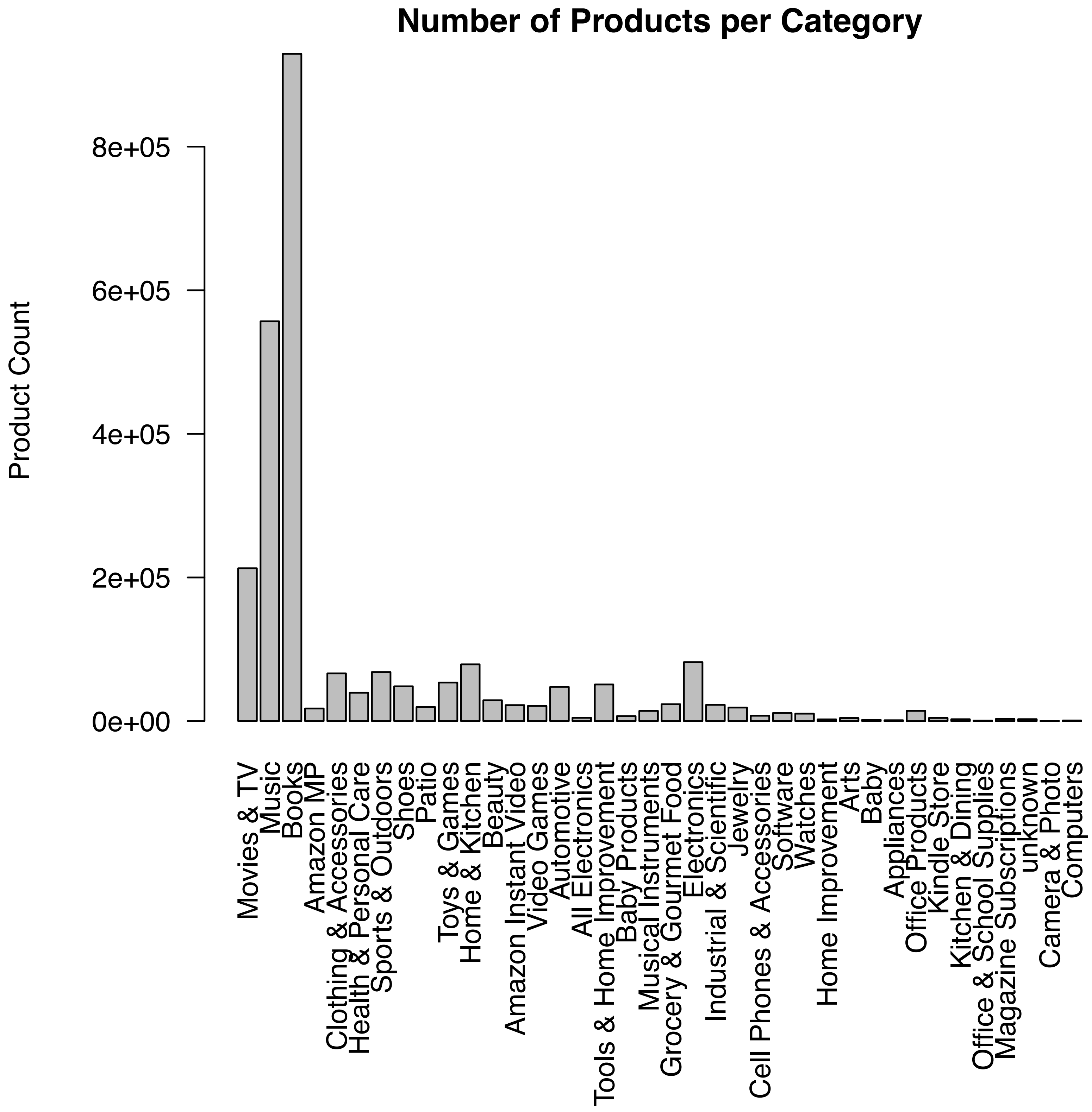
Dataset Characterization

- How many reviews are made per year?
- What are the “biggest” products in amazon?
- How much do products cost?
- What are the most expensive categories?
- How often do users review products?

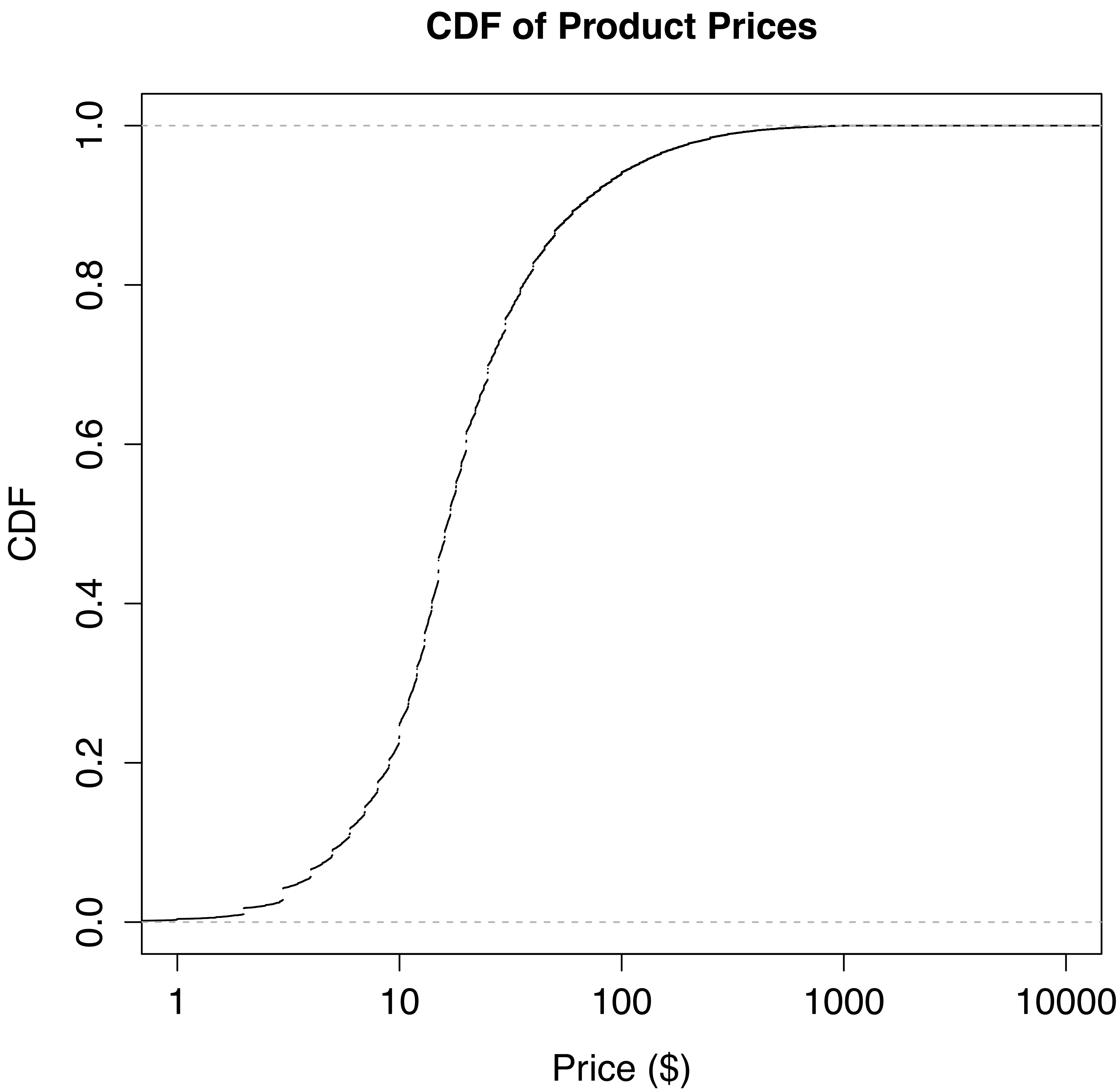
Reviews per Year



Product Categories

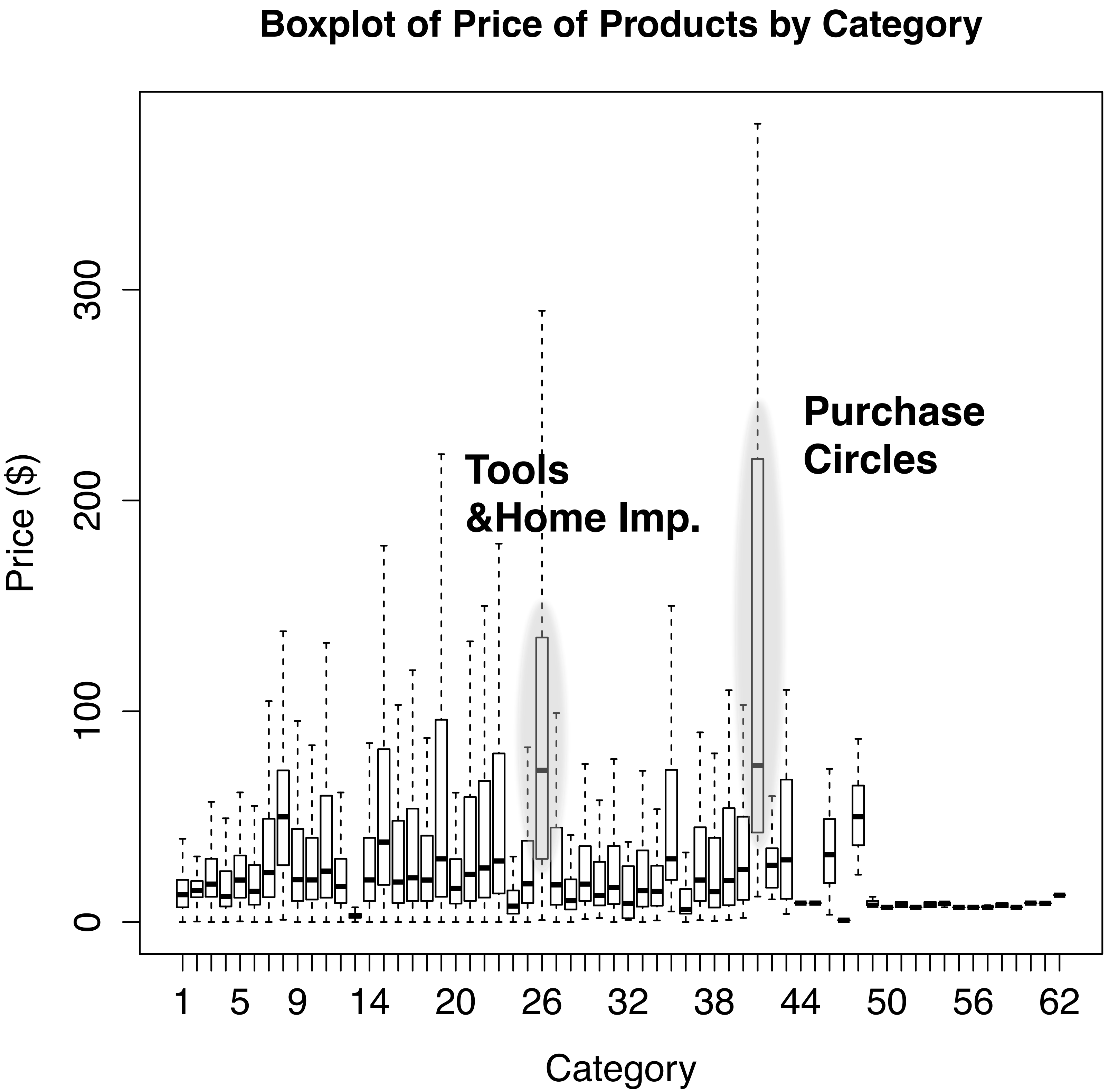


Product Prices



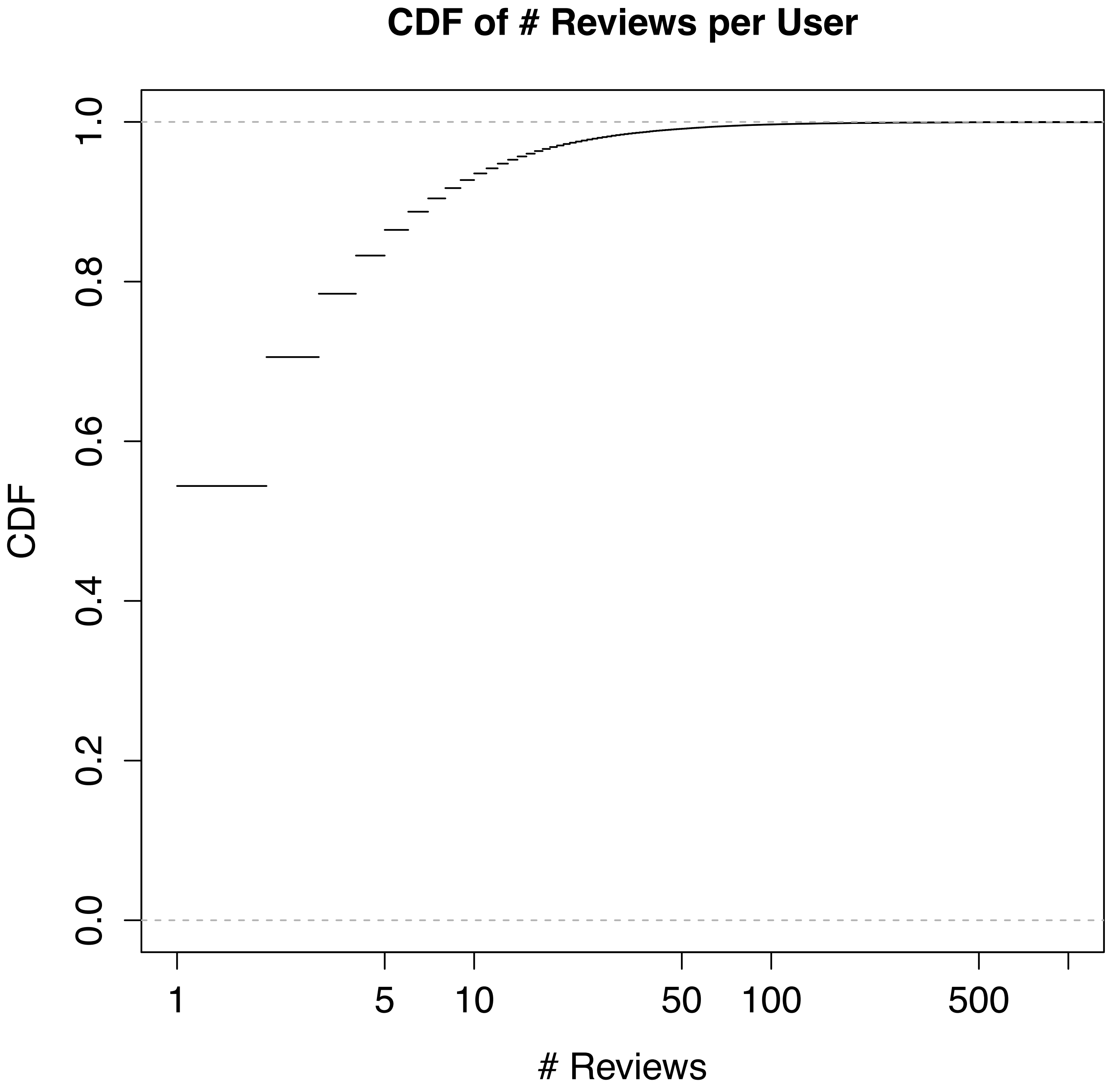
- Most products cost < 50$
- Prices capped at 999.99$

Product Prices



- Outliers ignored
- Purchase circles - bestsellers lists for specific groups

Users Reviews



> 80% of users do not review more than 5 times

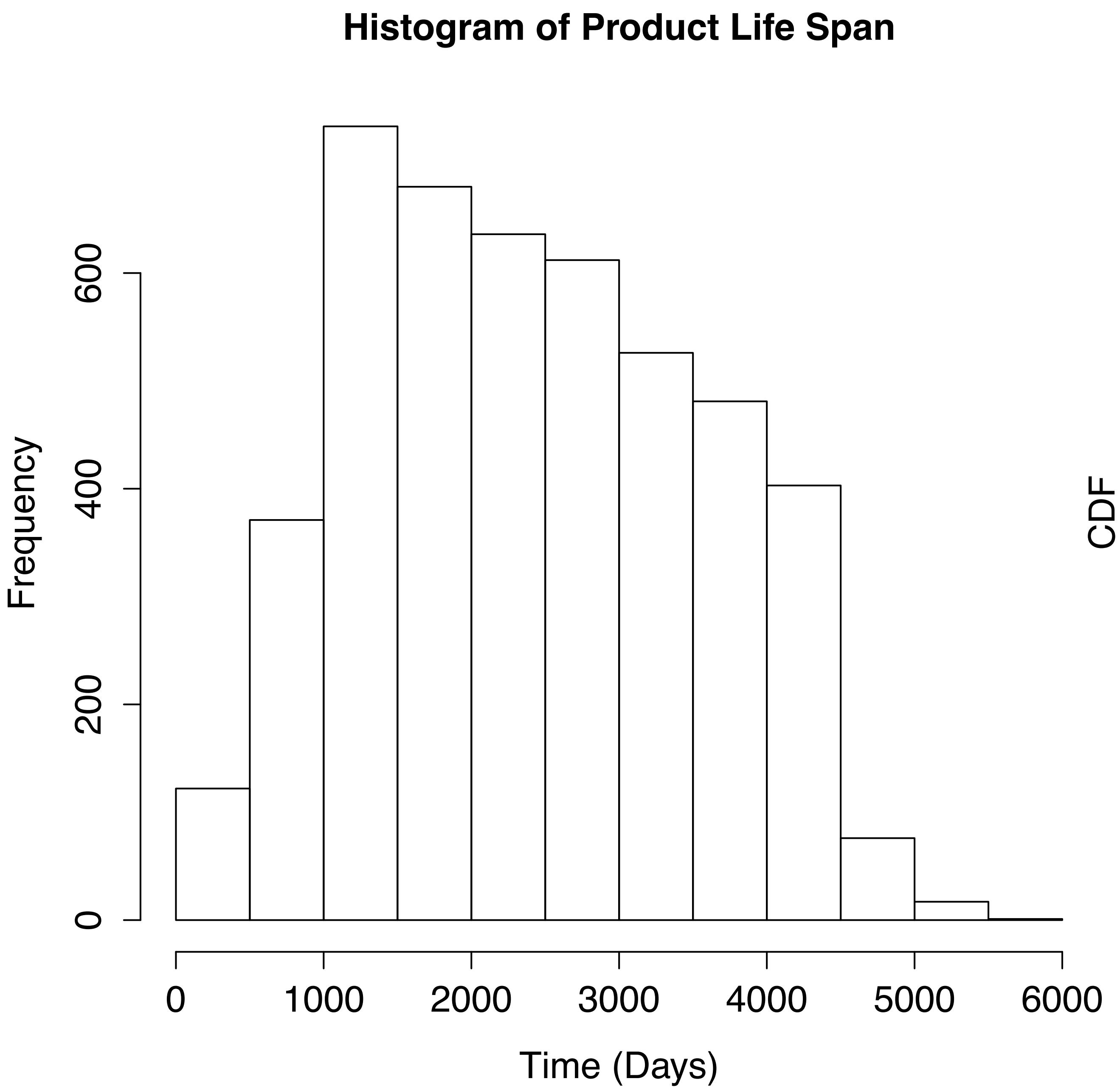
Products - Questions

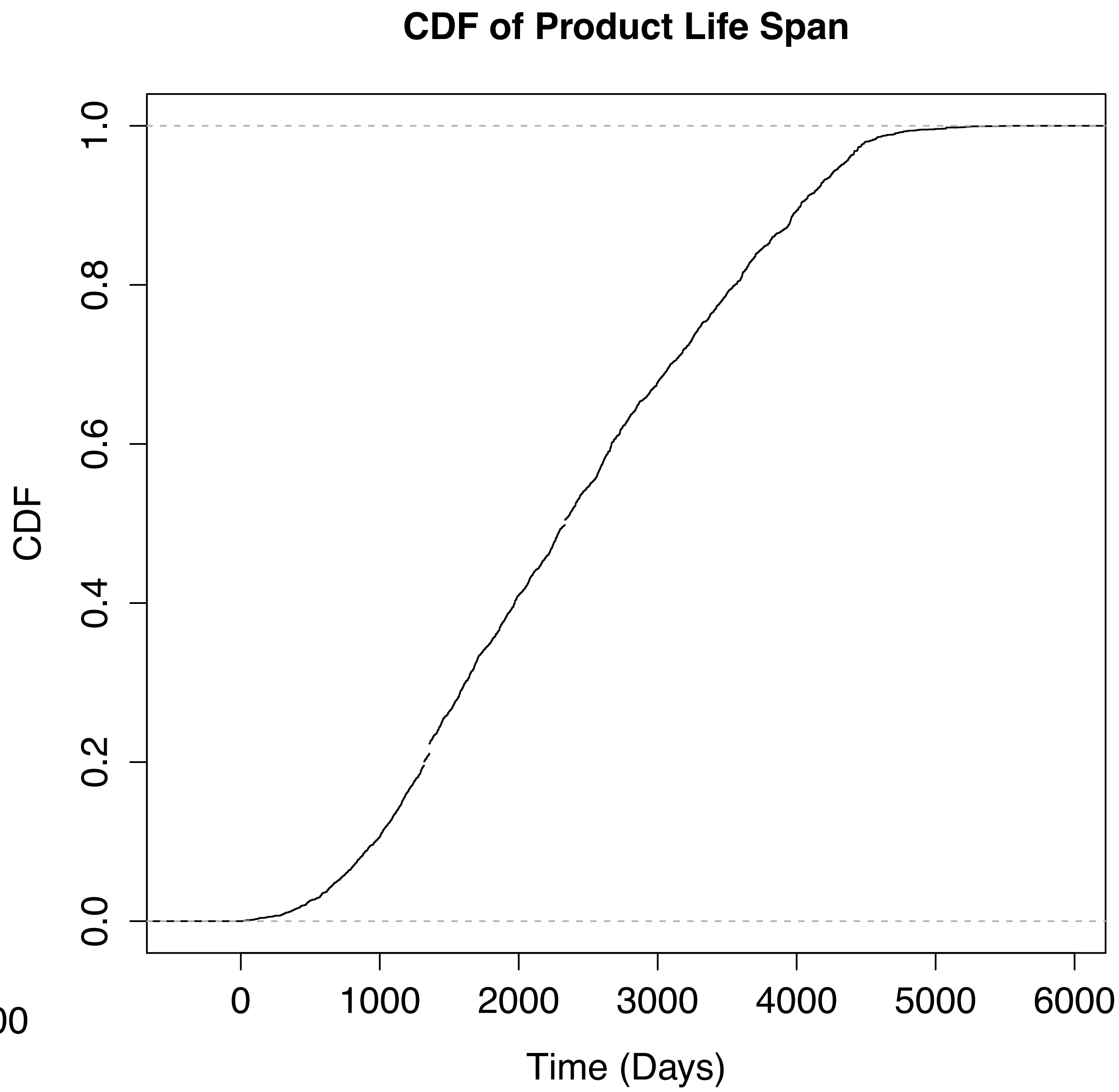
Subject	Question	Expectations
Life Expectancy	What is the life expectancy of a product?	Strong variations
	Do reviews affect the life expectancy of products?	Probably
	Do product life expectancy varies per product category?	Yes (e.g., books vs technology)
Reviews	Do review scores decay over time?	Depends on product category
	Do reviews cluster at specific times (e.g., product launch)?	Should follow curve of adoption

Products - Life Expectancy

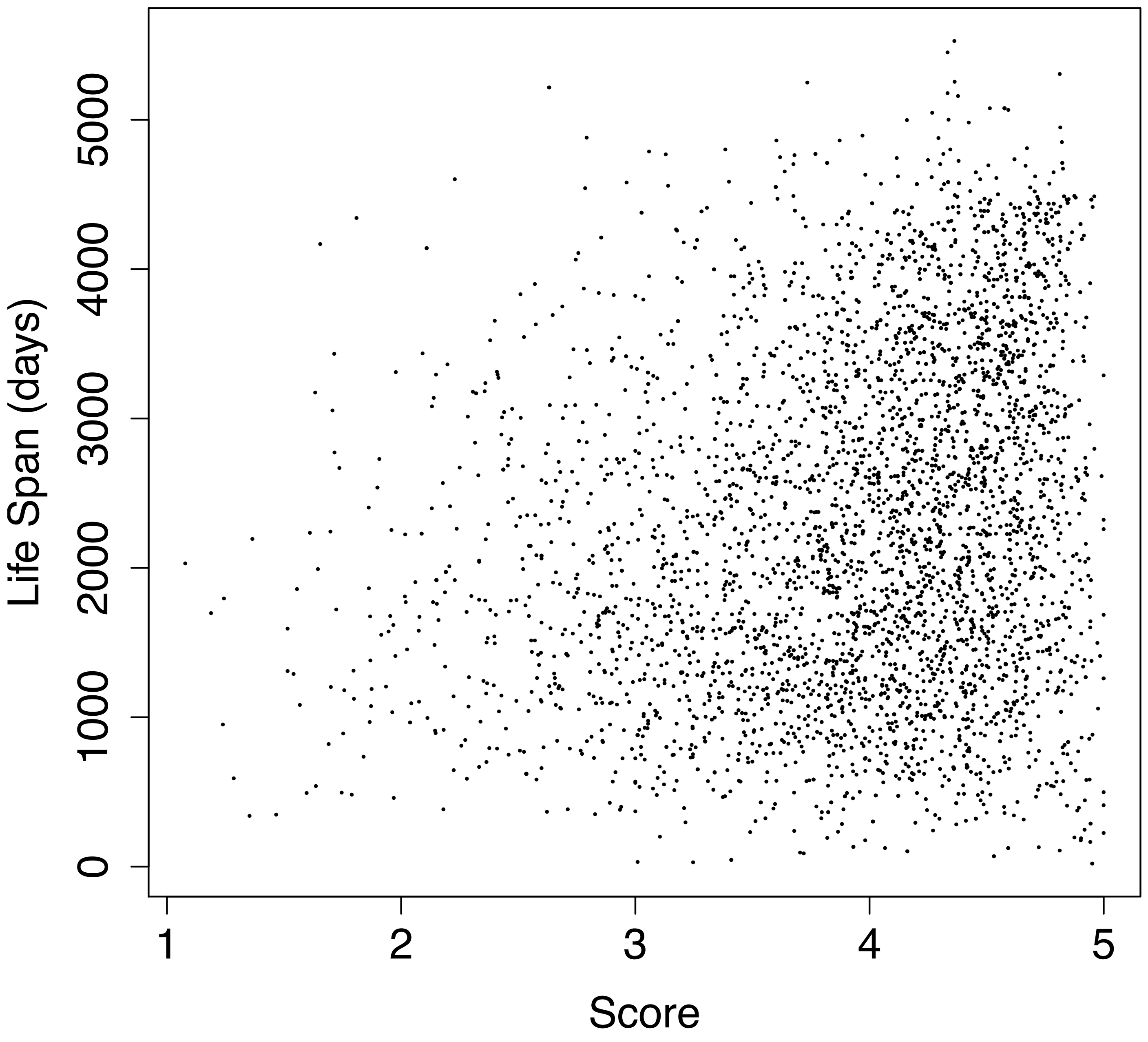
- Life expectancy: average number of years of life
- Considered only products with
  - > 50 reviews (frequently reviewed products)
  - last review before 2010 (no review likely means the product ‘died’)
- This filters reviews down to 4K products

Products - Life Span



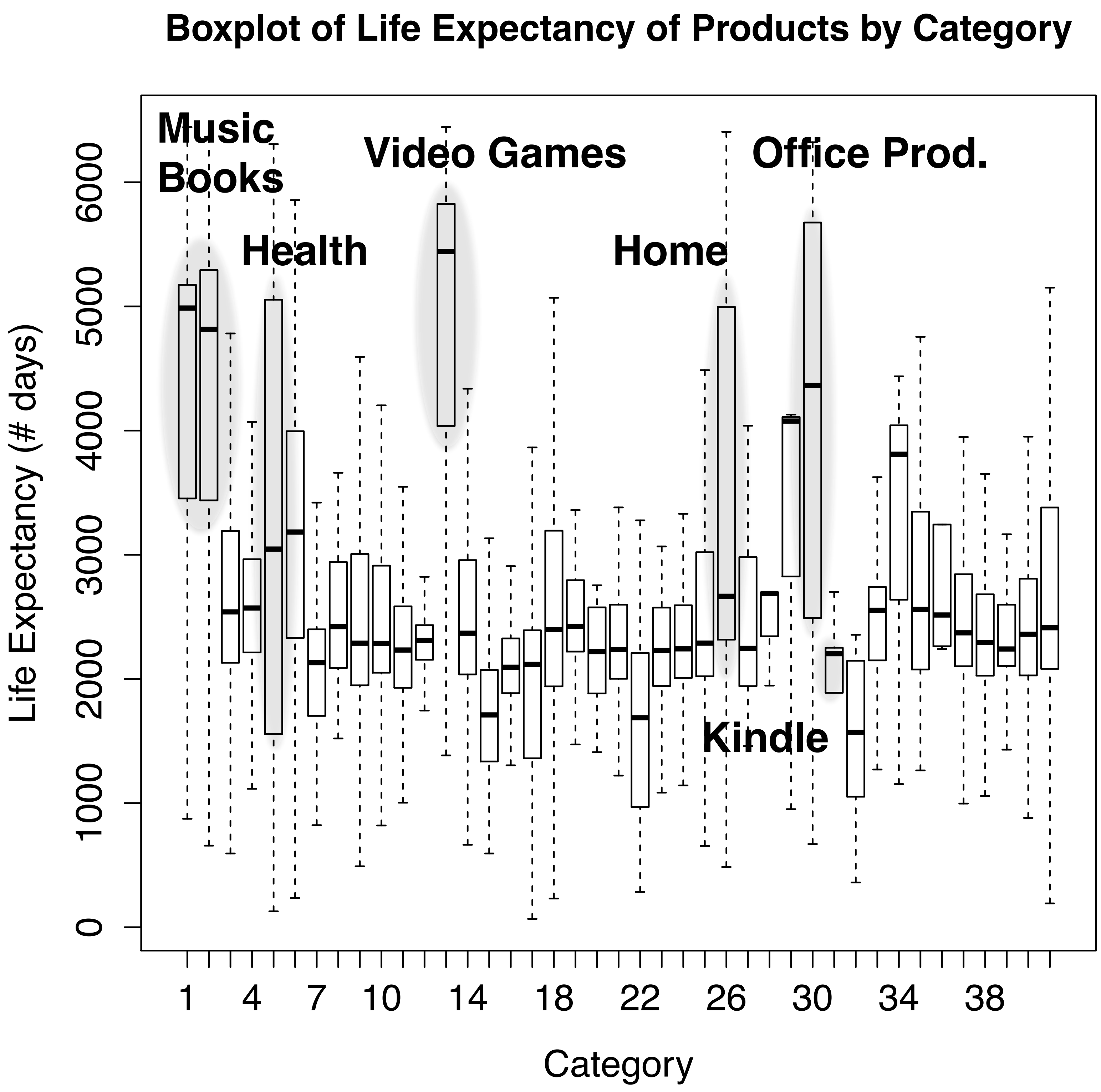


Products - Scores vs Life Expectancy



Correlation coefficient = 0.22 -> Scores do not affect life expectancy

Product Life Expectancy by Category

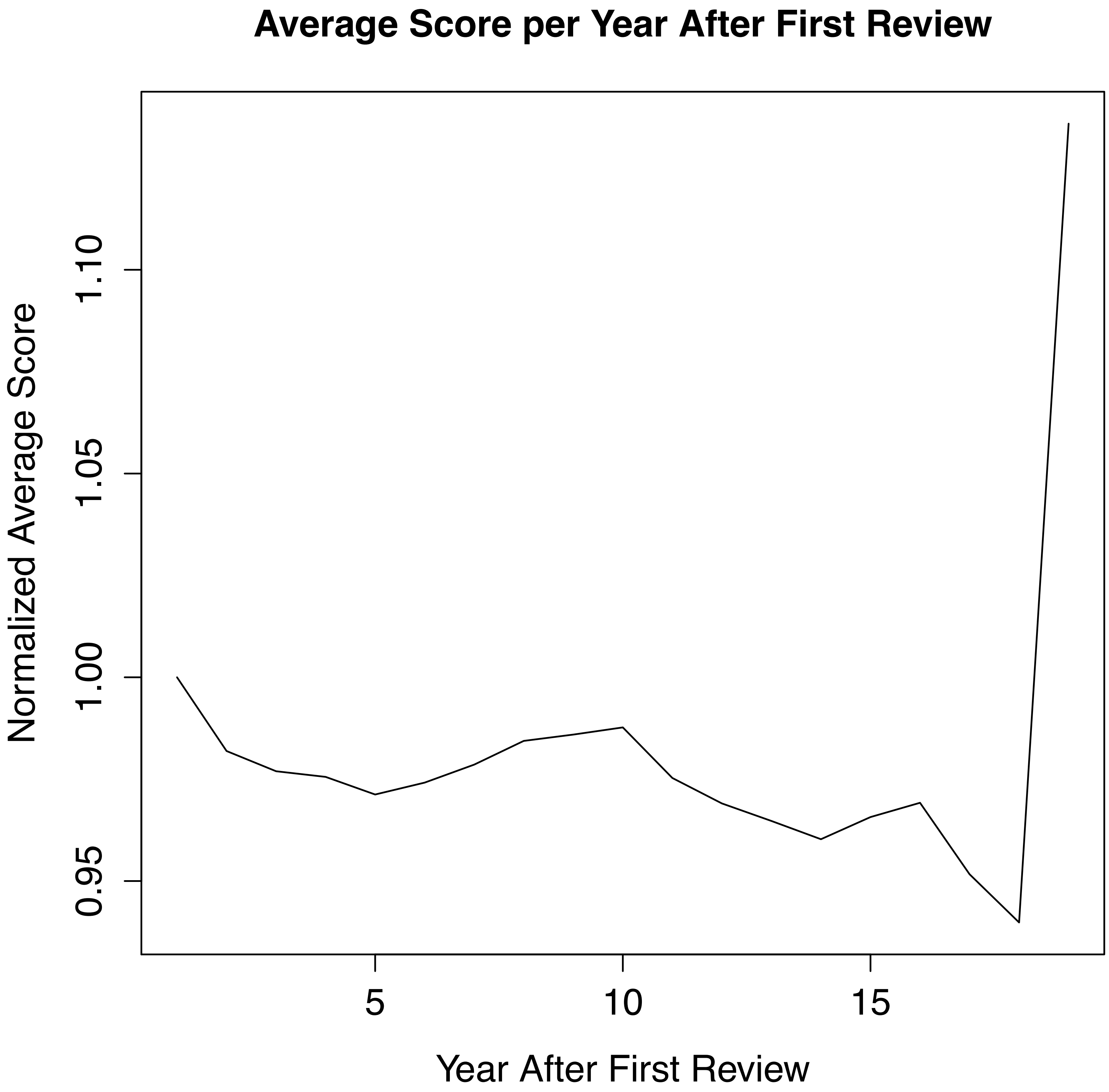


- Cross-classification of books and kindle

Review Scores Decay

- Compute the average decay of review scores over the years
- For each product scores are normalized to the first year average score
- Normalized scores are averaged per year after a product’s first review
- Products with less than 5 years of reviews and 3 reviews per year are ignored
- -> 28976 products

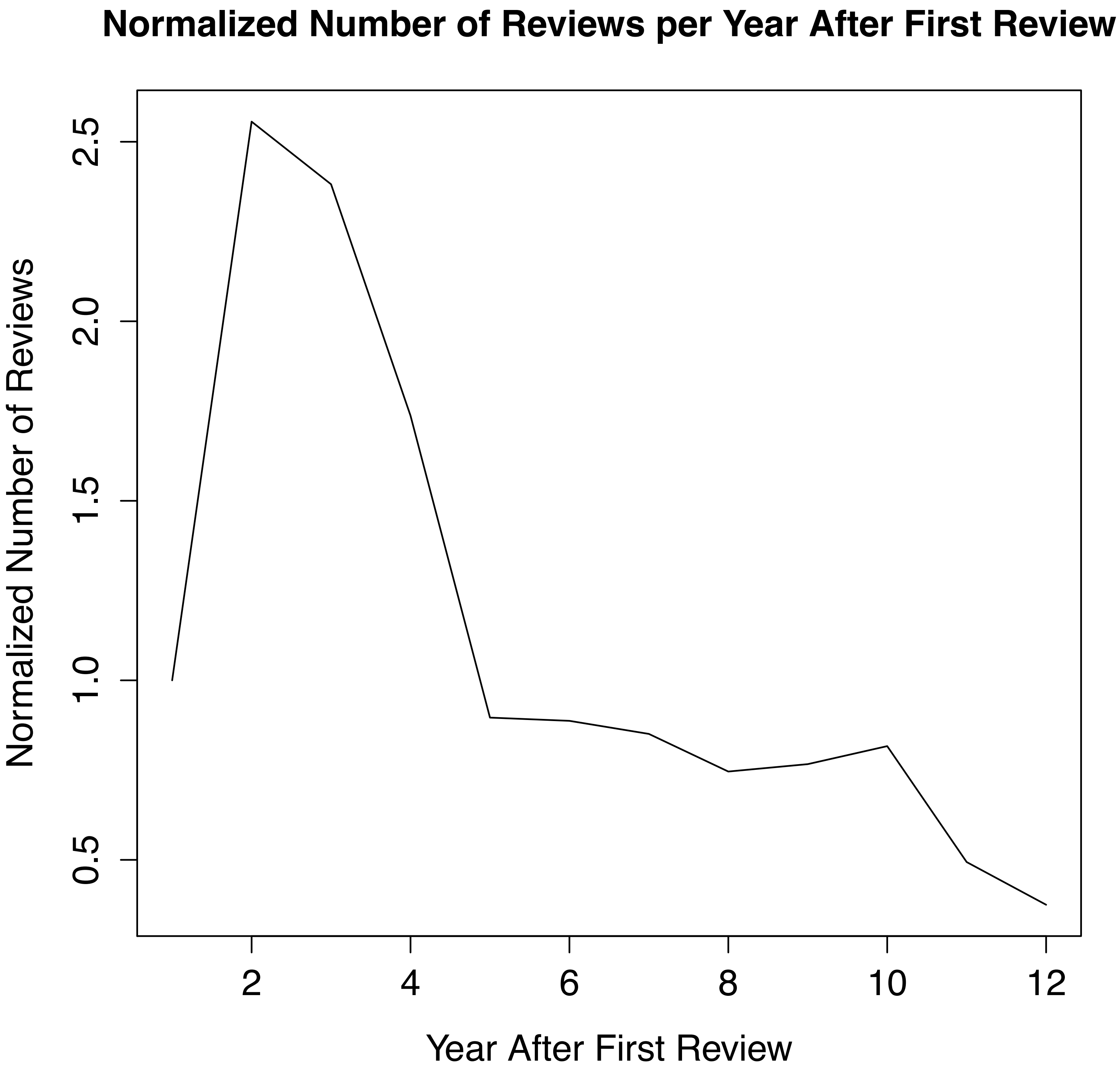
Review Scores Decay



Reviews Curve

- Compute reviews clustering throughout a product’s life — should follow curve of adoption
- For each product # of reviews is normalized
- # of reviews is averaged per year after a product’s first review
- Only “dead” products with no “holes” and at least 3 reviews per year considered
- -> 136 products

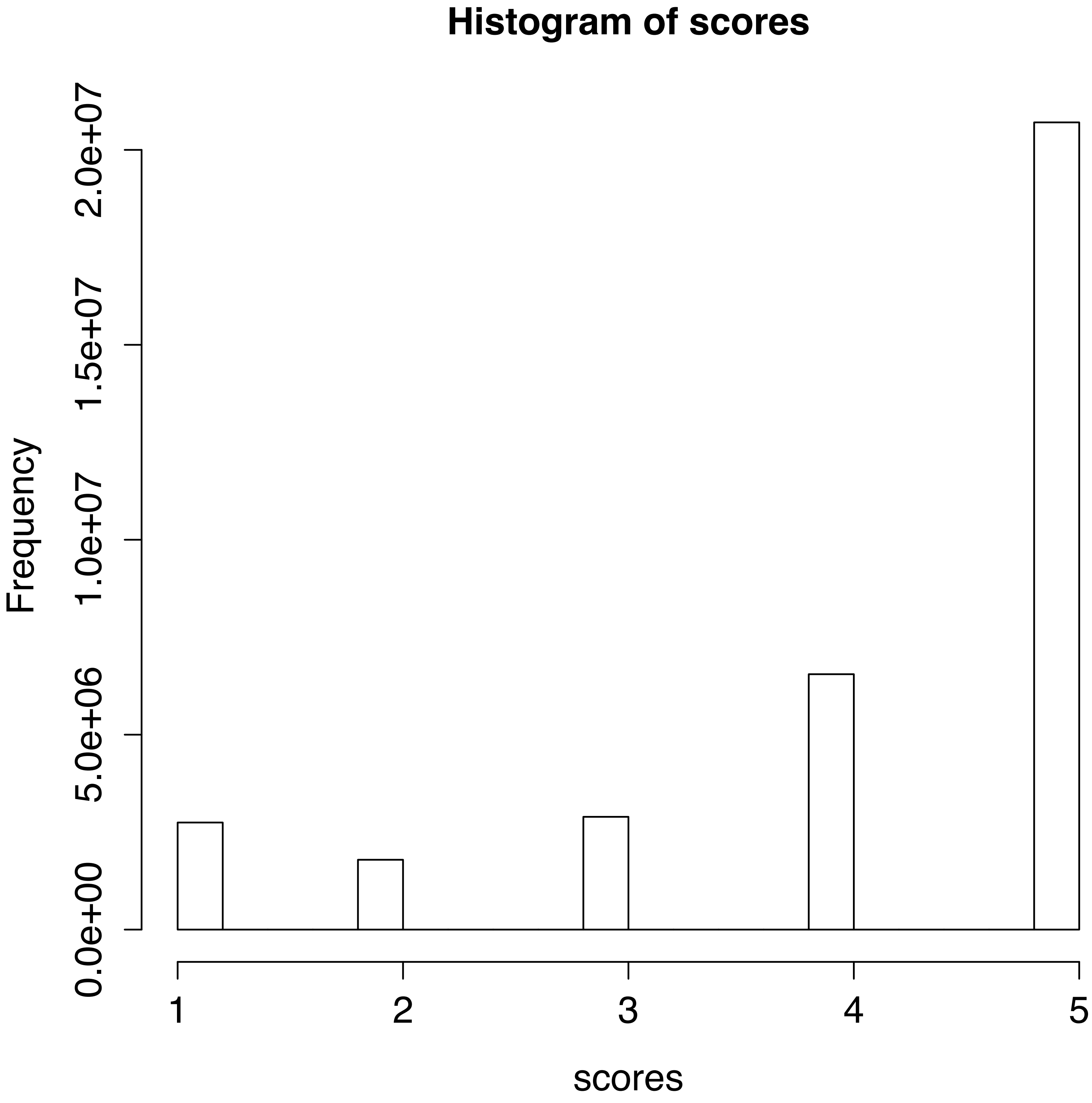
Reviews Curve



User Reviews - Questions

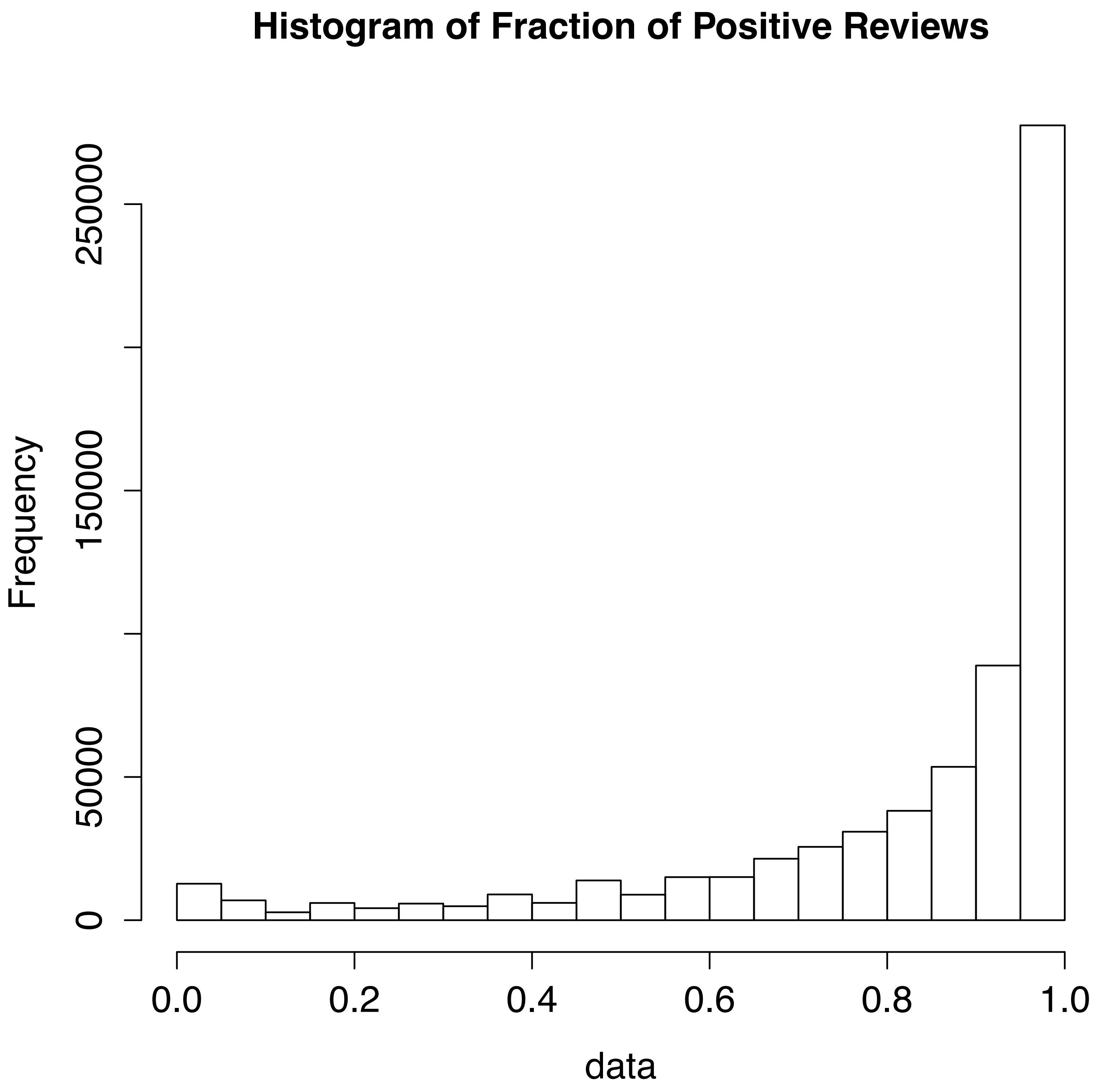
Question	Expectations
Do users tend to review a product when they are either very satisfied or unsatisfied?	Yes
Do positive / negative reviews tend to cluster in individual users, i.e., are there 'negative' users and 'positive' users?	Probably yes
Do users review products in a specific area of expertise or across different product categories?	Don’t know
Do users tend to be active reviewers over long periods of time?	No
What features of a review make it helpful?	Probably user experience and reviewer depth

Users - Scores



- Most reviews are positive

Users - Positive vs Negative Reviews

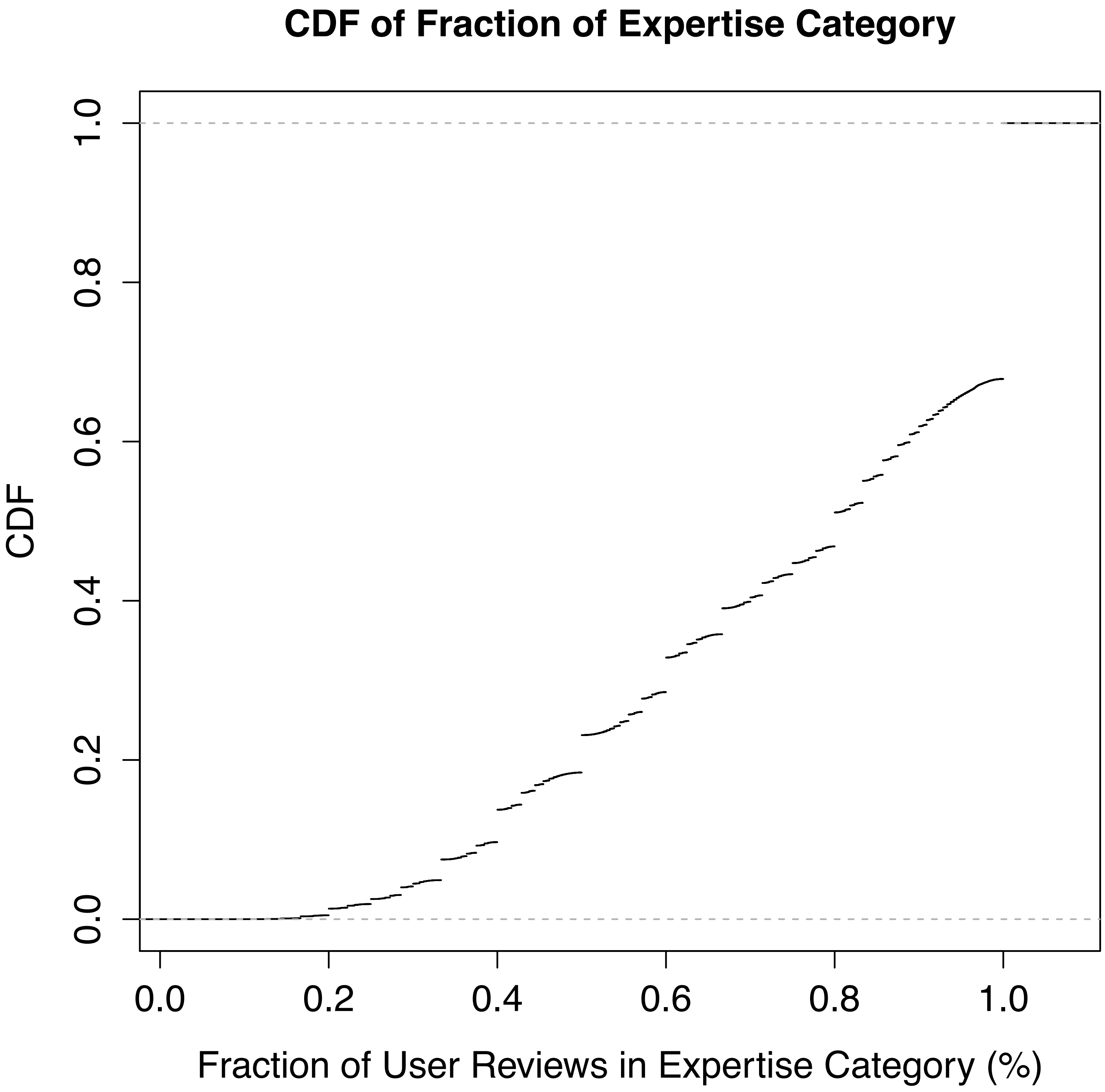


- Users with less than 10 reviews not considered
- Many “positive” users

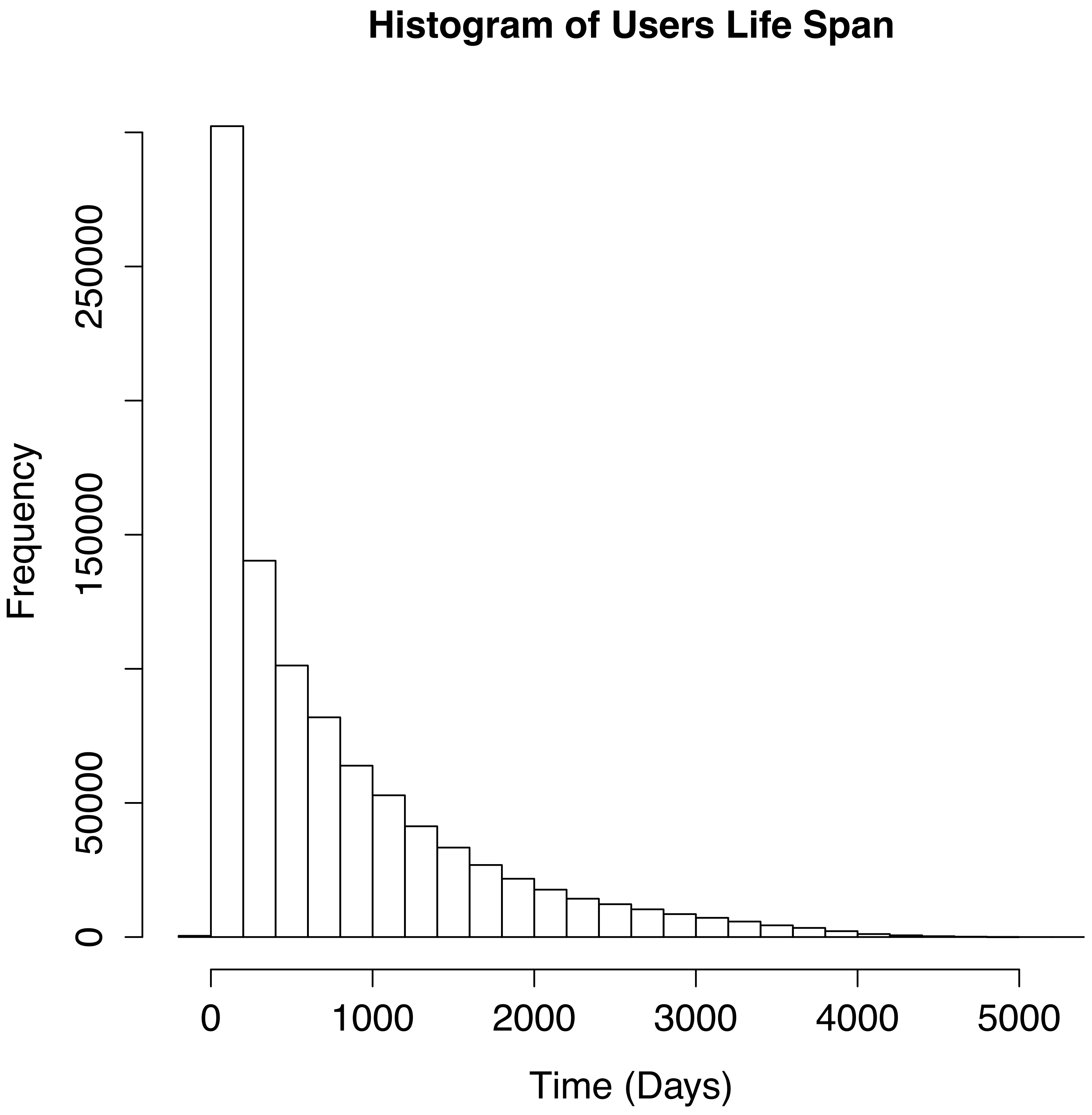
Are Reviewers (1 Cat.) Experts?

- Check how many reviews are focused on a single category for each reviewer
- Ignore reviewers with less than 5 reviews

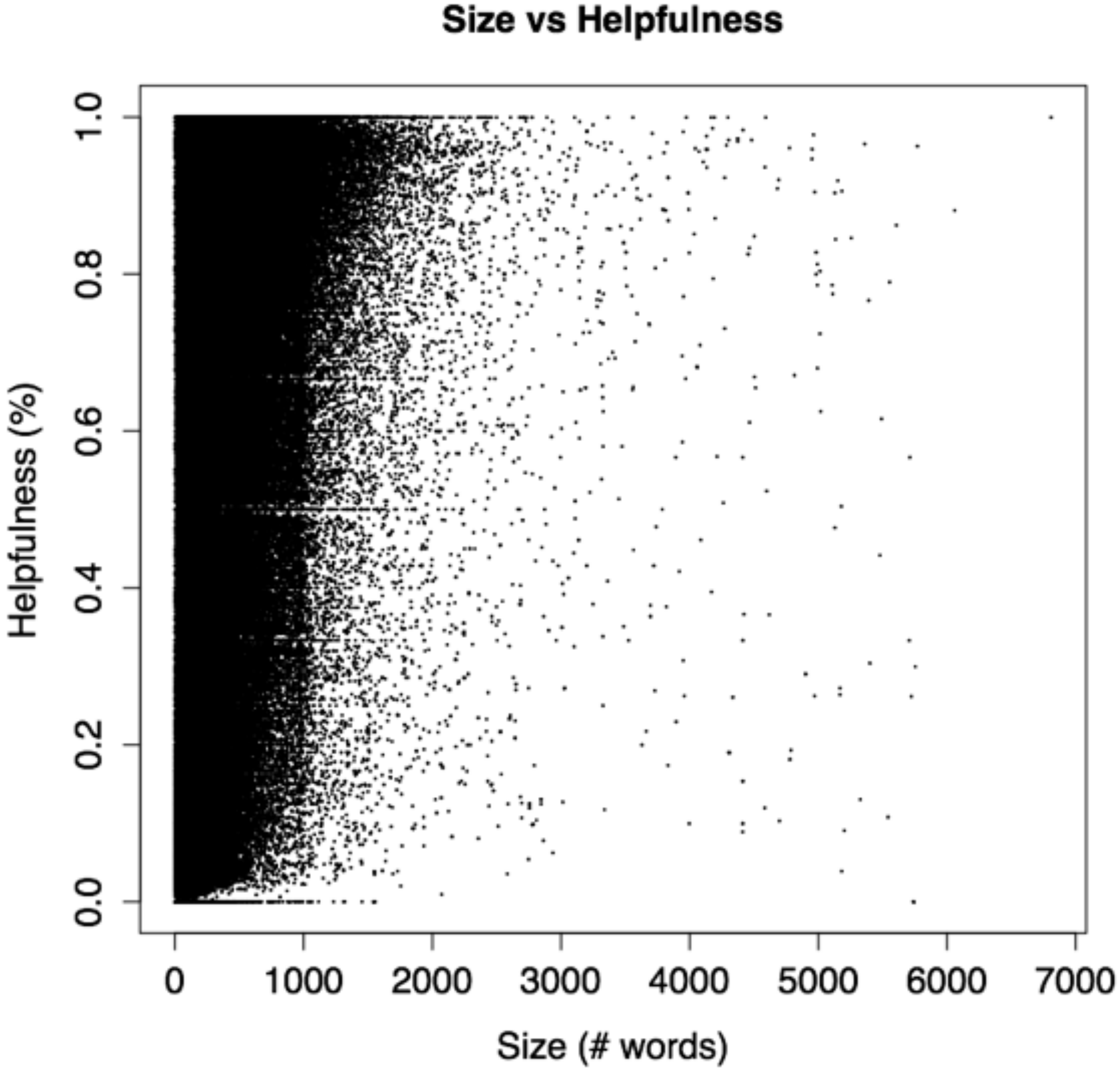
Are Reviewers (1 Cat.) Experts?



Users Life Expectancy

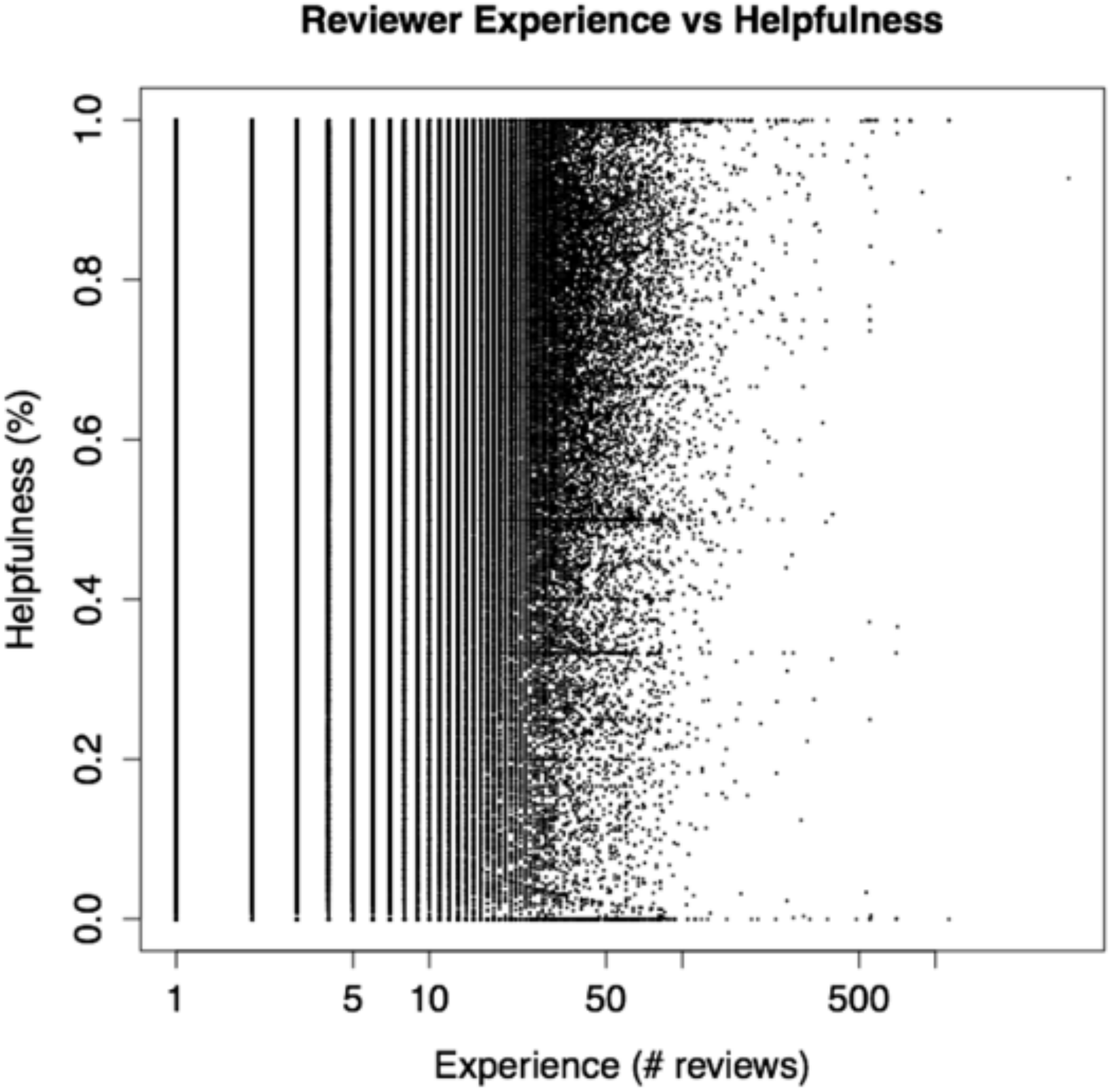


Reviews Size vs Helpfulness



- Correlation coefficient = 0.24

Reviewer Experience vs Helpfulness

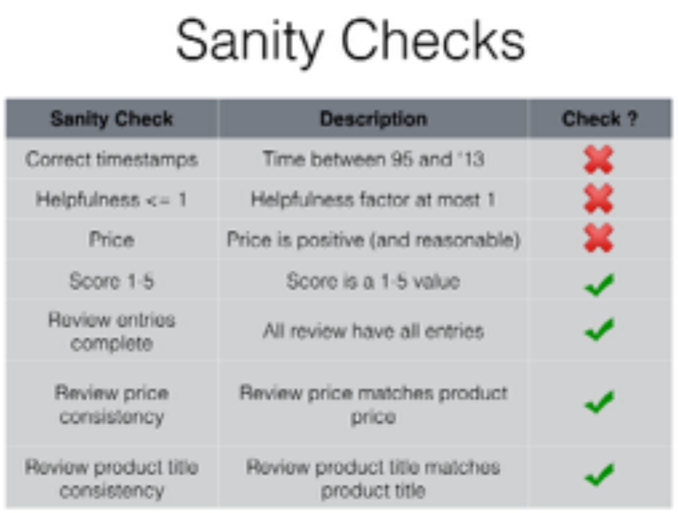


- Correlation coefficient = -0.041

Questions?

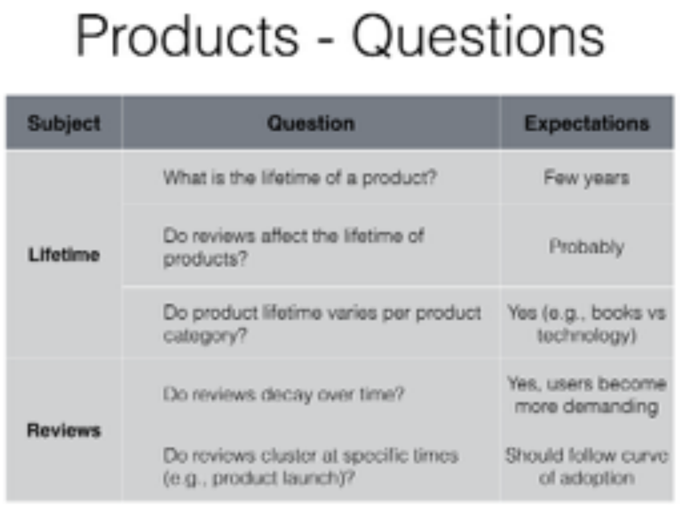
Dataset - Review Records

product/productId: 0131097601
product/title: C Programming in the Berkeley Unix Environment
product/price: unknown
review/userId: A1KLBWKUQHSQVW
review/profileName: Eugene Mah "physics geek"
review/helpfulness: 0/0
review/score: 4.0
review/time: 994291200
review/summary: indispensible title on my computer bookshelf
review/text: This has been one of those books that I constantly refer to. Not only is it good for learning some of the unique C things that apply to Unix, but you can also learn how to get around in Unix. This is the book I learned C from, and it's still one of the first ones I go to when I need to refresh my brain about something.



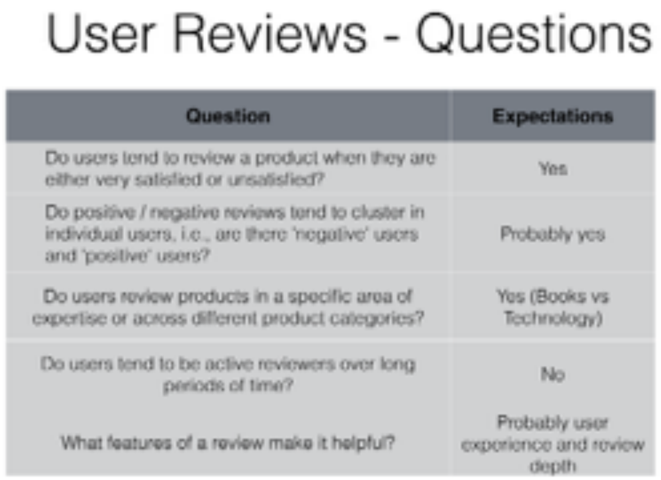
Sanity Checks

Sanity Check	Description	Check ?
Correct timestamps	Time between 95 and '13	✗
Helpfulness <= 1	Helpfulness factor at most 1	✗
Price	Price is positive (and reasonable)	✗
Score 1-5	Score is a 1-5 value	✓
Review entries complete	All review have all entries	✓
Review price consistency	Review price matches product price	✓
Review product title consistency	Review product title matches product title	✓



Products - Questions

Subject	Question	Expectations
Lifetime	What is the lifetime of a product?	Few years
	Do reviews affect the lifetime of products?	Probably
	Do product lifetime varies per product category?	Yes (e.g., books vs technology)
Reviews	Do reviews decay over time?	Yes, users become more demanding
	Do reviews cluster at specific times (e.g., product launch)?	Should follow curve of adoption



User Reviews - Questions

Question	Expectations
Do users tend to review a product when they are either very satisfied or unsatisfied?	Yes
Do positive / negative reviews tend to cluster in individual users, i.e., are there 'negative' users and 'positive' users?	Probably yes
Do users review products in a specific area of expertise or across different product categories?	Yes (Books vs Technology)
Do users tend to be active reviewers over long periods of time?	No
What features of a review make it helpful?	Probably user experience and review depth

- github.com/jcarreira/amazon-study