Final Project

CS:4980:15 – Big Data Technologies

Tony Andrys

Amazon Review Data Analysis

CS:4980:15 – Big Data Technologies // Final Project

Tony Andrys

I. Overview

The goal of this project was to answer the following questions about Amazon reviews, both in the specific category of music and over all reviews:

1. How are reviews distributed about individual products? How are products rated?

2. How are scores distributed? Do reviews tend to be positive, negative, or average?

3. Is there a correlation between the length of a review and the score it assigns a product?

4. How do other users perceive another review (judged by "helpfulness" scores assessed by other Amazon users)? Are reviews considered "more helpful" when they are positive, negative, or neutral?

5. Do early reviews of a product affect the scores of later reviews? In other words, is there evidence of confirmation bias among review authors?

Work on this project occurred in two phases. The first phase consisted of reformulating the data and answering a subset of the research questions from the Basic Plan. During this period, only reviews from the Music category were analyzed. All work was able to be performed on the class cluster and data was stored on the shared s3 bucket.

The second phase involved answering the remaining questions from the Basic plan, as well as expanding analysis to the rest of the dataset. Due to the size of the full dataset, work on the class cluster was almost impossible. As a result, the remainder of the project was completed on an individual cluster. All data is stored as gzips and are accessible at <s3://andrys-cs4980>.

II. Hardware Configuration

A. Amazon Web Services

To provide the computational resources necessary for this project, Amazon's suite of virtual web services (AWS) were utilized.

Elastic Compute Cloud (EC2) and Elastic Map Reduce (EMR) provided the virtual machines, networking infrastructure, and clustering frameworks necessary for distributed computing. Amazon's Simple Storage Service (S3) was used for remote storage of the dataset, logs, and the final results.

B. Cluster Configuration

The Spark cluster used for data processing and analysis contained one master node and four worker nodes.

TABLE I
CLUSTER RESOURCES

Instance Type	vCPUs	RAM	SSD Storage
m3.xlarge	4	16 GB	80 GB
r3.xlarge	4	32 GB	160 GB
r3.xlarge	4	32 GB	160 GB
r3.xlarge	4	32 GB	160 GB
r3.xlarge	4	32 GB	160 GB

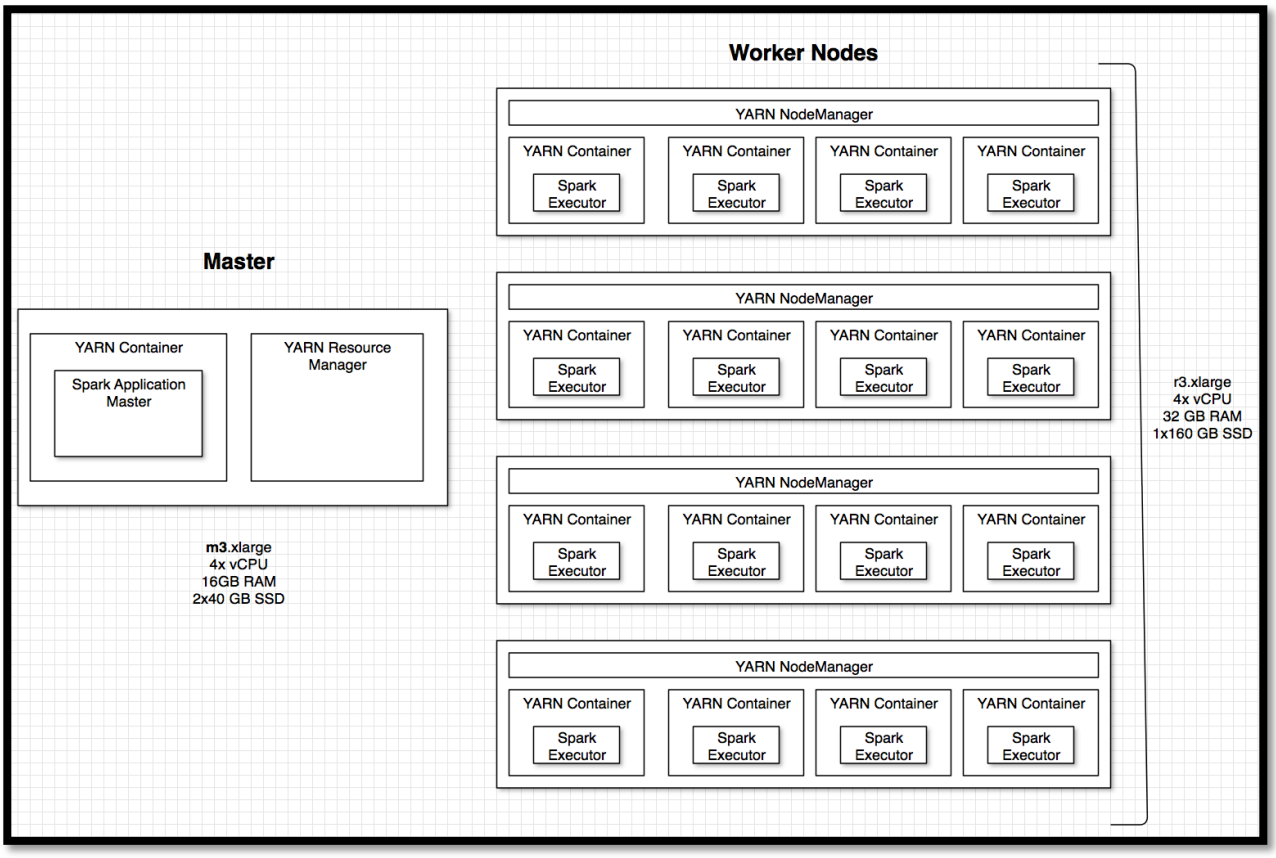


Fig 1. Cluster configuration and software stack

III. Review Data

A. Overview

The data used in this project was compiled and used internally by members of the Stanford Network Analysis Project[1], a research group interested in social network analysis. The Amazon review data was first made available to third parties after McAuley and Leskovec's RecSys paper[2] was published in 2013. This collection contains approximately 34,686,770 reviews over 2,441,053 products available on Amazon. The reviews range from June 1995 to March 2013.

B. Data Size

The data is distributed as 27 files, one file corresponding to each top-level product category on Amazon. Each file contains a subset of the total collection of reviews.

TABLE II
ENUMERATION OF REVIEW DATA

Product Category	**Review Count**	**Uncompressed Filesize**
Amazon Instant Video	717,651	839.8 MB
Arts	27,980	19.6 MB
Automotive	188,728	129.8 MB
Baby	184,887	42 MB
Beauty	252,056	171.9 MB
Books	12,886,488	14390 MB
Cell Phones & Accessories	78,930	69.4 MB
Clothing & Accessories	681,933	372.9 MB
Electronics	1,241,778	1110 MB
Gourmet Foods	154,635	106.3 MB
Health	428,781	313.1 MB
Home & Kitchen	991,794	757.6 MB
Industrial & Scientific	137,042	85.7 MB
Jewelry	58,621	34.9 MB
Kindle Store	160,793	190.6 MB
Movies & TV	7,850,072	8770 MB
Musical Instruments	85,405	68.1 MB
Music	6,396,350	6510 MB
Office Products	138,084	107.7 MB
Patio	206,250	158.7 MB
Pet Supplies	217,170	164.6 MB
Shoes	389,877	240.4 MB
Sports & Outdoors	510,991	368.3 MB
Tools & Home Improvement	409,449	321.2 MB
Toys & Games	435,996	314.7 MB
Video Games	463,669	485.2 MB
Watches	68,356	51.5 MB
Total	**35,358,850**	**36.31 GB**

C. Data Format

Each review in the dataset contained 10 properties (or attributes). The first three attributes relate to the *product* being reviewed, while the remaining seven are characteristics about the review itself.

TABLE III
REVIEW PROPERTIES

#	Property (key)	Description	Notes
1	product/productId	Alphanumeric product identifier	Amazon Standard Identification Number (ASIN). Has a well defined form: alphanumeric string of length 10. Maps to URL: http://amazon.com/dp/<ASIN>
2	product/title	Title of product	As printed on the product's webpage.
3	product/price	Price in dollars	Simple floating point value – no leading $ included.
4	review/userId	Alphanumeric user identifier	Maps to URL: http://www.amazon.com/gp/cdp/member-reviews/<review/userId>
5	review/helpfulness	Ratio reflecting the number of users who voted this review as "helpful" or "unhelpful".	$h = \frac{yesVotes - noVotes}{totalNumberOfVotes}$
6	review/profileName	Display name of author	Appears above each review authored by the user.
7	review/score	Score of review	Valid scores: 1.0, 2.0, 3.0, 4.0, 5.0
8	review/time	Date and time this review was published	Represented as UNIX/POSIX time (ex: 1431568566)
9	review/summary	Heading of this review	On Amazon.com, this is displayed in bold font directly above the body of the review.
10	review/text	Full text of the review	Stored exactly as submitted by the author in terms of capitalization and punctuation. However, any existing control characters ('\n', '\t', etc) were not stored.

Attributes of a review were stored as ten key–value pairs, and review boundaries are denoted by a single blank line.

product/productId: B000H9LE4U
product/title: Copper 122-H04 Hard Drawn Round Tubing, ASTM B75, 1/2" OD, 0.436" ID, 0.032" Wall, 96" Length
product/price: 22.14
review/userId: ABWHUEYK6JTPP
review/profileName: Robert Campbell
review/helpfulness: 0/0
review/score: 1.0
review/time: 1339113600
review/summary: Either 1 or 5 Stars. Depends on how you look at it.
review/text: Either 1 or 5 Stars. Depends on how you look at it.1 Star because they sent 6 feet of 2" OD copper pipe.0 Star because they won't accept returns on it.5 stars because I figure it's actually worth $12-15/foot and since they won't take a return I figure I can sell it and make $40-50 on this deal

product/productId: B000LDNH8I
product/title: Bacharach 0012-7012 Sling Psychrometer, 25?F to 120?F, red spirit filled

Fig 2. Example of review data representation from Industrial & Scientific category.

D. Reformulation

The format of the data raised many concerns.

While the attributes of each review were human readable, data processing frameworks such as Hadoop and Spark are written to expect all associated attributes of an object (a review) on one line. The use of a colon as the key–value separator would lead to parsing problems if a review's body text contained a colon.

Finally, from a resource standpoint, this choice of data structure was extremely wasteful. In our input file, the keys exist only to separate review properties and are discarded as the data is read. Because Spark will never use the key names, there is no need for anything but a single character to separate review attributes.

To illustrate the waste inherent in this format, consider the amount of storage necessary required to store the key names of one review. Each review requires an extra 157B (on top of the space required to store its attributes) to store in a file. Overhead of this size is nontrivial when we recall that we need to store over 35,000,000 reviews. When 157B of overhead is scaled to this level, data structure wastes 5.44GB on information that Spark will immediately throw out!

To avoid these problems, it was necessary to write a series of Python scripts to transform the data from a series of key—value pairs to a list of tab separated values.

IV. Results

A. *Product Review Distribution - Music*

The calculated product review distribution was in line with my predictions. More than half (60%) of products have 1-3 reviews, 83% of products have been reviewed 1-10 times, and the vast majority (92%) have anywhere from 1-25 reviews.

TABLE IV
PRODUCT REVIEW FREQUENCIES - MUSIC

Number of Reviews/Product	**Frequency**	**% Of Sample**
1	187443	33.66%
2	92476	16.61%
3	55130	9.90%
4	36689	6.59%
5	26100	4.69%
6	19728	3.54%
7	14947	2.68%
8	12074	2.17%
9	9625	1.73%
10	8024	1.44%
Total	**462236**	**83.01%**

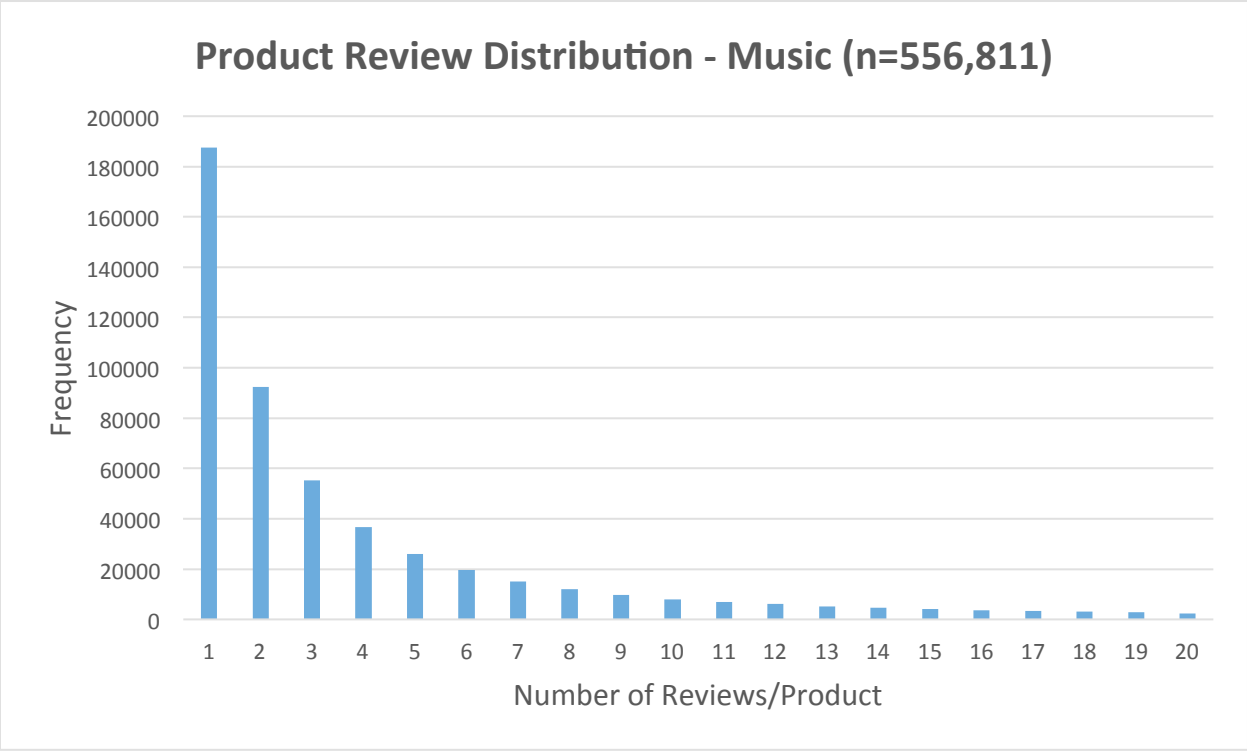


Fig 3. Music product review frequencies from [1,25]

B. Product Review Distribution – All Categories

The product/review distribution observed in the Music category is similar to the distribution received when all categories are considered. 57.7% of products have 1-3 reviews, 80.6% of products have 1-10 reviews, and 90.7% have 1-25 reviews.

TABLE V
PRODUCT REVIEW FREQUENCIES – MUSIC

Number of Reviews/Product	**Frequency**	**% Of Sample**
1	801013	33.14%
2	375910	15.55%
3	217238	8.99%
4	153922	6.37%
5	108740	4.50%
6	87477	3.62%
7	64165	2.65%
8	56728	2.35%
9	43333	1.79%
10	38797	1.61%
Total	**1947323**	**80.56%**

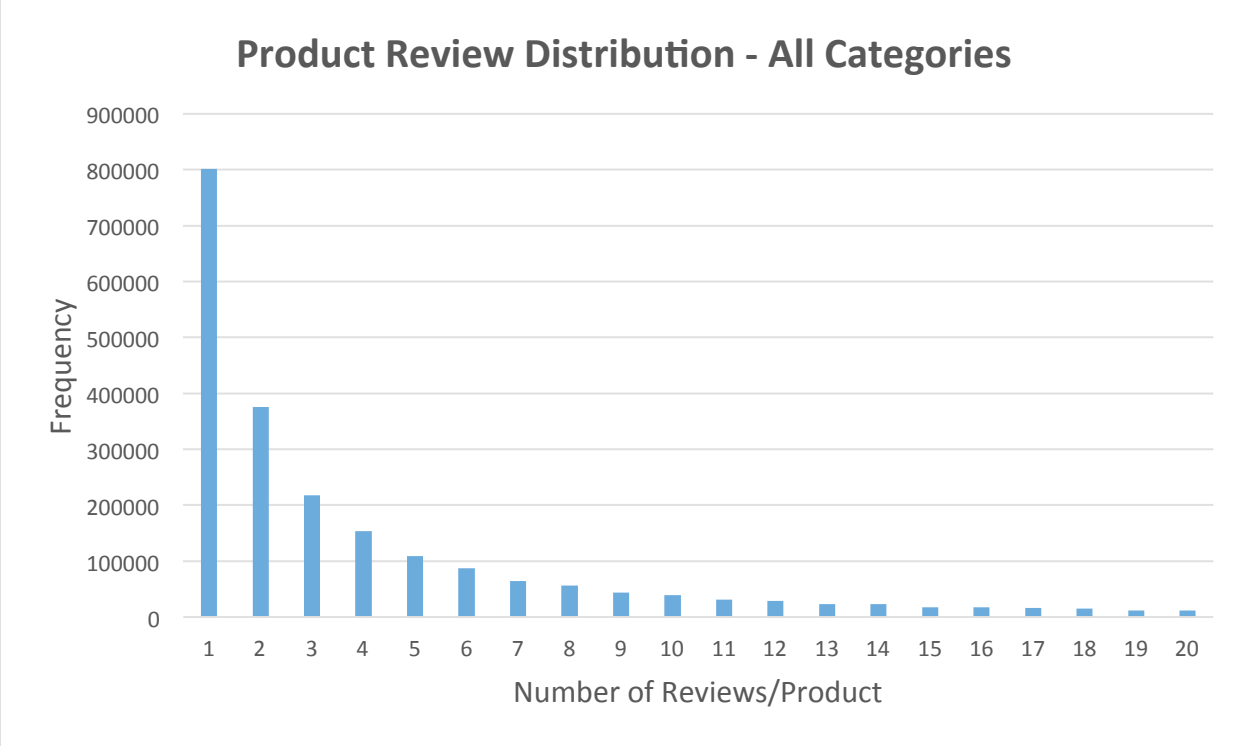


Fig 4. All product review frequencies from [0,20]

C. Score Distribution

One of the more surprising results of this project came in the form of the score distribution. Contrary to popular belief, the vast majority of Amazon reviews are overwhelmingly positive. More than three quarters (78.5%) of all reviews in the dataset give their products above average (4.0) or excellent (5.0) scores. This trend is even more apparent in Music, where more than four-fifths of reviews (83.5%) assign scores of either 4 or 5.

TABLES VI, VII
REVIEW SCORE DISTRIBUTIONS – MUSIC & ALL CATEGORIES

Music			**All Categories**		
Review Score	Frequency	% Of Sample	Review Score	Frequency	% Of Sample
1.0	355037	5.55%	1.0	2799514	7.92%
2.0	246367	3.85%	2.0	1827169	5.17%
3.0	448143	7.00%	3.0	2949238	8.34%
4.0	1148156	17.95%	4.0	6671675	18.88%
5.0	4198641	65.64%	5.0	21083236	59.67%

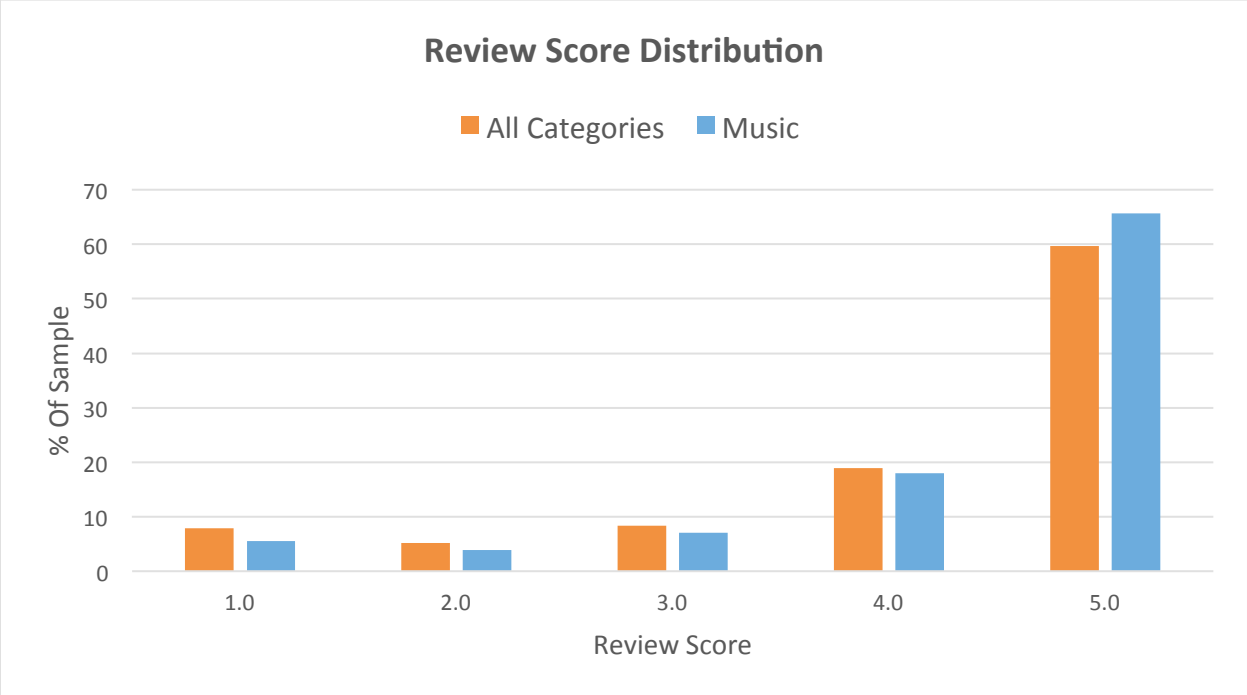


Fig 5. Score distributions for Music and All Categories

D. Score vs. Word Count

There is an intuitive argument for the existence of a relationship between score and review length. If we agree that all reviews are a reflection of a *real* consumer's opinion of a product (and not authored by a paid employee of its manufacturer or an author who picked a score at random), it is safe to assume the author holds a strong opinion. A strong opinion seems to imply that the author would have more to say (good or bad), and would write a longer review.

However, the data says that this is not the case. Surprisingly enough, users who rate products as average (3.0) tend to write longer reviews. In the case of Music, average reviews are about 42 words longer than 1.0 or 5.0 reviews. In general, average reviews are about 40 words longer.

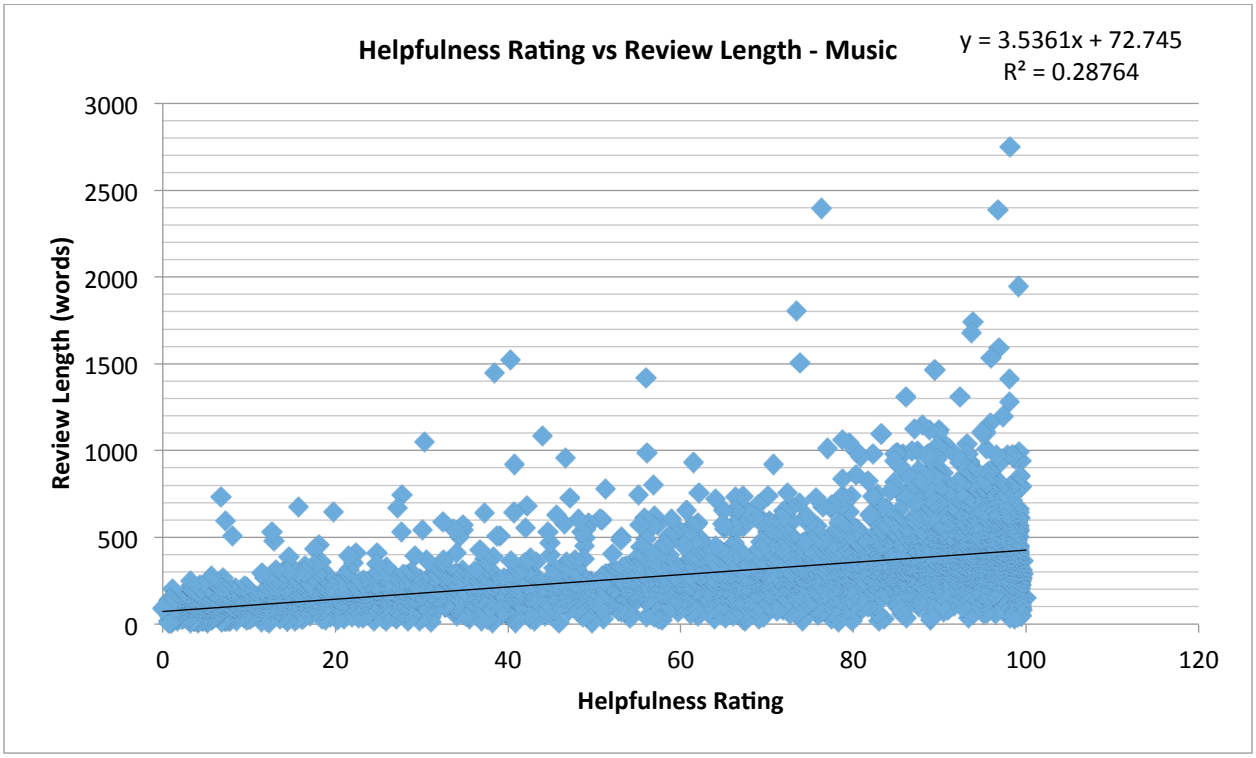
TABLES VIII, IX
AVERAGE LENGTH OF REVIEW VS SCORE – MUSIC & ALL CATEGORIES

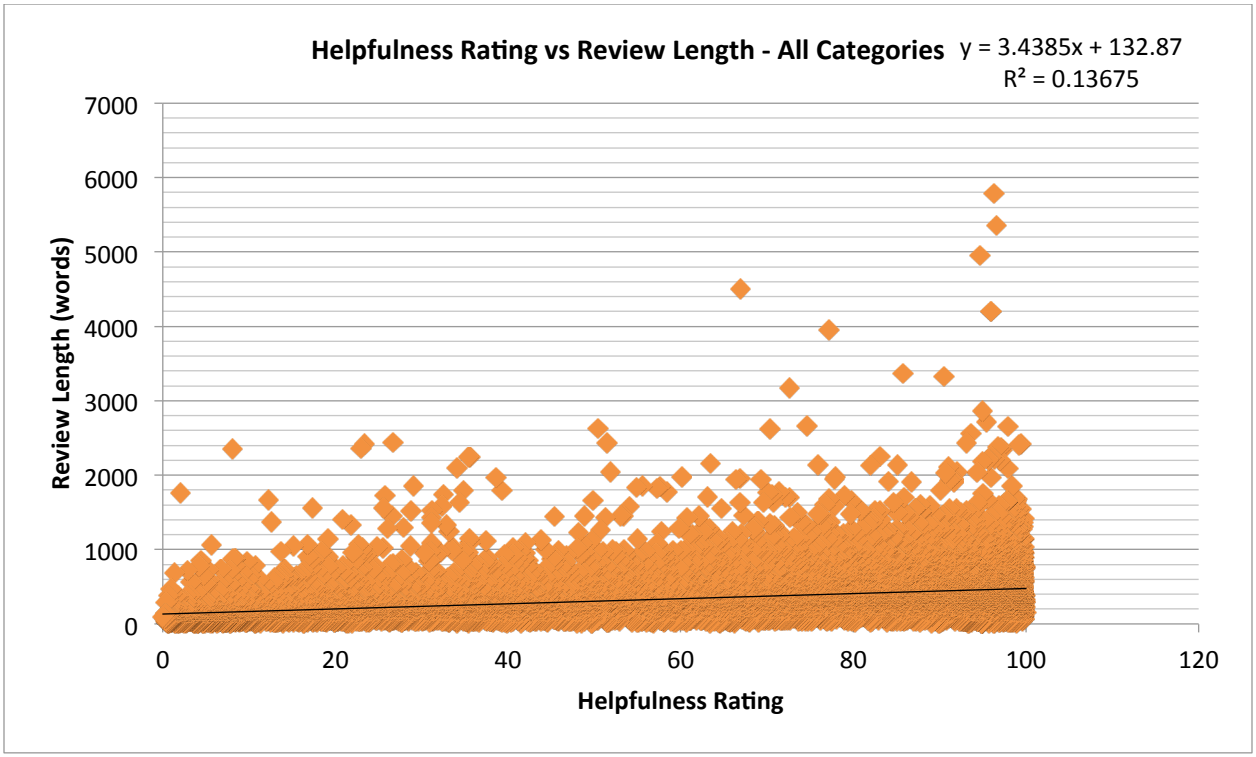
Music		**All Categories**	
Score	Average Word Count/Review	Score	Average Word Count/Review
1.0	105.7017607	1.0	117.9418956
2.0	139.0781558	2.0	149.3101098
3.0	157.2016432	3.0	159.3762552
4.0	154.0655181	4.0	152.403799
5.0	124.02789	5.0	119.1872612

Fig 6. Average length of review wrt score.

E. Helpfulness Rating vs Word Count

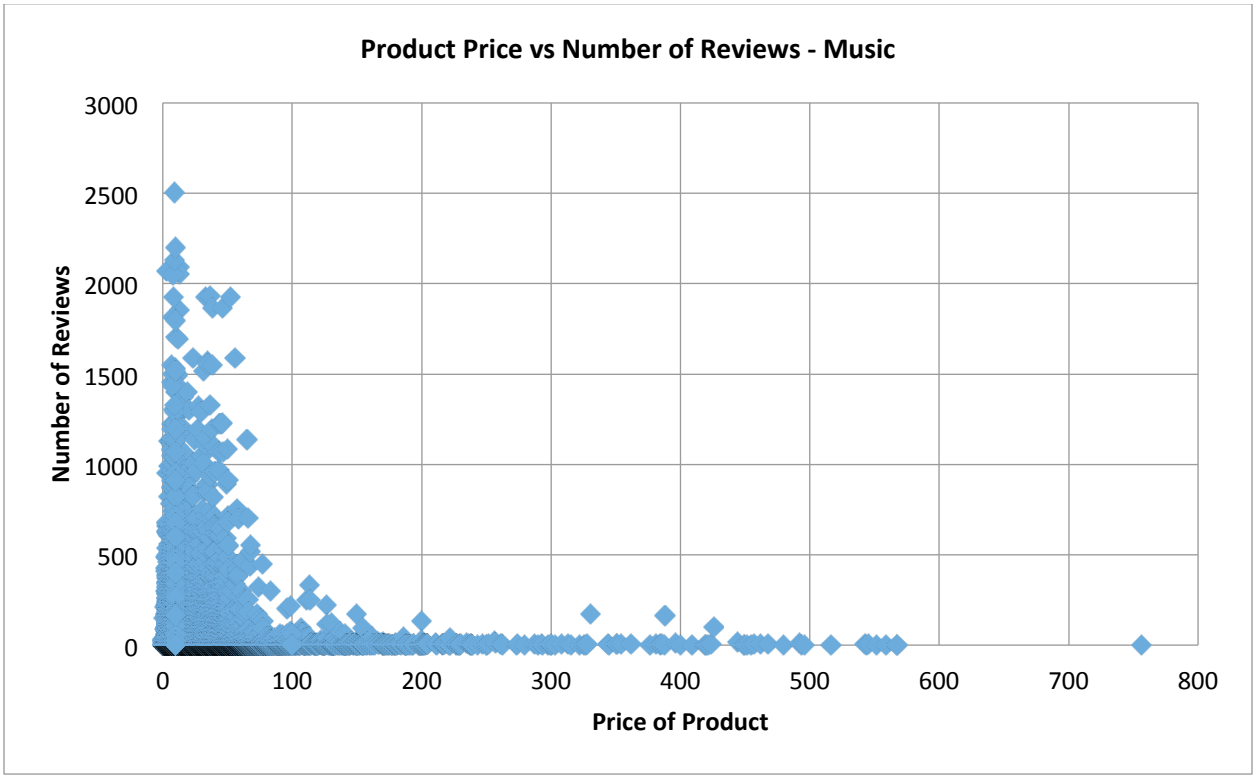
As reviews grow in length, their helpfulness scores tend to increase. This makes sense: A longer, more in-depth review is more likely to help make a purchase decision than a few disjoint comments or a short rant about a product.





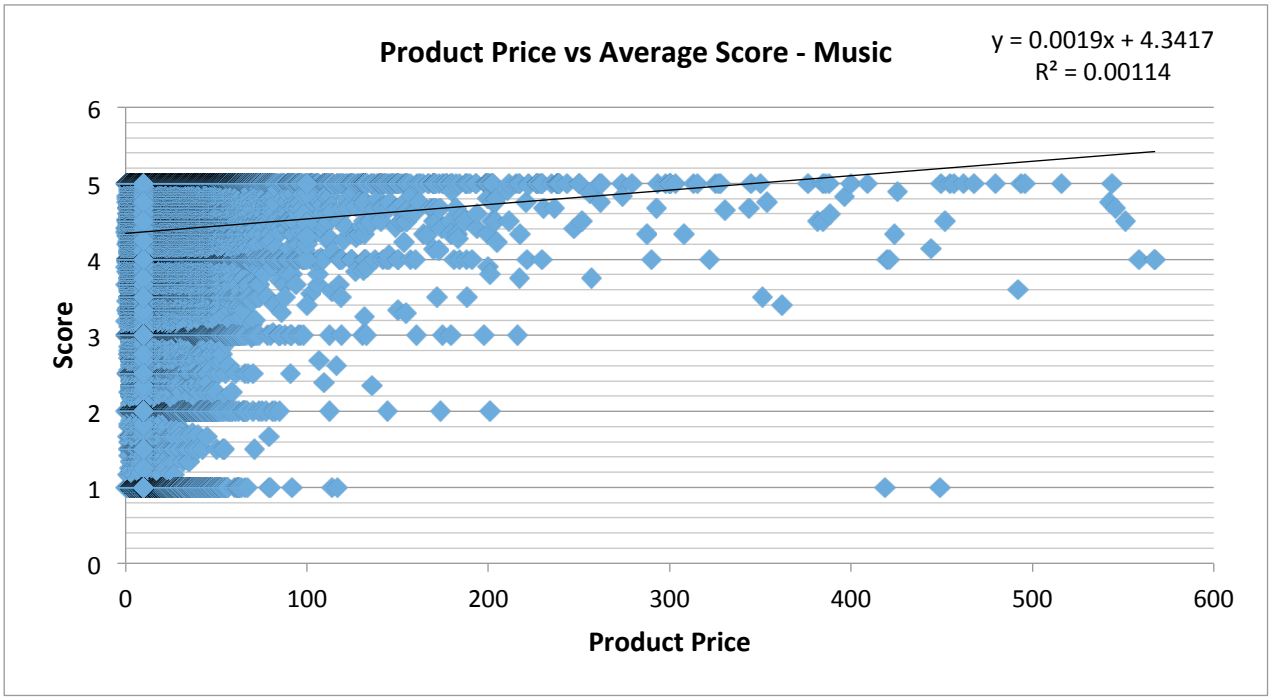
F. Number of Reviews vs. Price

The data presents a case for an inverse relationship between the price of a product and the number of times it has been reviewed. At the risk of engaging in *post hoc* argument, it seems reasonable to think that high prices leads to fewer sales, and fewer sales would lead to a smaller number of reviews on Amazon. The answer is likely to be much more complex than this.



G. Product Price vs Average Score

In the Music category, there is a positive correlation between the price of a product and its mean rating, but not a particularly strong one. It would be interesting to control for the differences in review counts (as demonstrated in the previous section) to see if this relationship would hold. No correlation was observed when considering the entire data set*.



**This graph served as an excellent reminder of the horrible experience one can encounter when trying to process large amounts of data using software that is not designed to handle it. Excel took approximately 9.5 minutes to render this scatter plot on a relatively modern machine with a 2.6GHz i7 and 16GB of RAM.*

H. Can a product's first reviews affect its later reviews?

This was measured by recording each product's initial review score (excluding any product with less than ten reviews). Then, the whole set of each product's review scores was sorted by their publish dates in ascending order. After this, we attempt to fit a curve (via least-squares polynomial fit) to the data. The slope of the trend line is calculated. Finally, all slopes associated with products with the *same* initial score were grouped together and averaged. The resulting slopes are extremely small, leading to the belief that the answer to this question is no. However, the result of this computation could be considered an approximation at best, and it is very likely that this is not the best way to quantify this.

TABLE X
PRODUCT MEAN LINEAR REGRESSIONS BY SCORE – MUSIC

Score	Slope of Trend Line (average)
1.0	$3.11 * 10^{-10}$
2.0	$2.96 * 10^{-9}$
3.0	$9.85 * 10^{-11}$
4.0	$-9.02 * 10^{-10}$
5.0	$1.03 * 10^{-9}$

TABLE XI
PRODUCT MEAN LINEAR REGRESSIONS BY SCORE – ALL CATEGORIES

Score	Slope of Trend Line (average)
1.0	$3.85 * 10^{-10}$
2.0	$2.18 * 10^{-9}$
3.0	$7.27 * 10^{-10}$
4.0	$-6.57 * 10^{-10}$
5.0	$-2.20 * 10^{-9}$

References

[1] Jure Leskovec, Andrej Krevl, SNAP Datasets: Stanford Large Network Dataset Collection, 2014, http://snap.stanford.edu/data

[2] J. McAuley and J. Leskovec. Hidden factors and hidden topics: understanding rating dimensions with review text. RecSys, 2013.