Exploring the Mechanisms that Drive Product Ratings on Amazon

Phoebe Wong, Nick Stern, Zizheng Xu, Yiming Xu
Stat 139 Fall 2018 Final Project

Introduction

Amazon showcases more than 560 million products as of January 2018.[1] This immense repository gives rise to transaction rates that brought in nearly 178 billion dollars of net revenue for the company in 2017 alone.[2] Amazon's influence has reshaped how products are discovered, sold, and shipped, catalyzing a major shift in balance from traditional retail to e-commerce. One interesting byproduct of the rising e-commerce trend is an increase in transparency surrounding the quality of products. Amazon facilitates this through their product reviews, a feature where customers who have verifiably purchased the product can comment on their experience, ascribing it an integer numerical value from 1 to 5. Presumably better ratings are correlated with better products, but are ratings also correlated with price? Are ratings correlated with the number of reviews? Are they dependent upon the type of product?

The underlying psychology of how people rate products is ambiguous. When price is involved, one line of logic may be that we should expect expensive products to be higher quality, and thus earn higher ratings. On the other hand, if the product is exorbitant, the reviewer may be more critical in their judgement of its worth. In the context of categorical comparisons, one would expect customers to be more critical of products that have a greater impact on their wellbeing. For example, someone who buys a computer with a spotty internet connection would likely leave a lower rating than someone who buys a fancy chair that wobbles. This is because the inability to browse the internet has more of a negative impact than being uncomfortable in a chair, even though the two objects may cost the same. In short, our project aims to clarify these psychological quandaries using data on Amazon product reviews, prices, and ratings.

Hypotheses of Interest

There are two main hypotheses we investigate in our report:

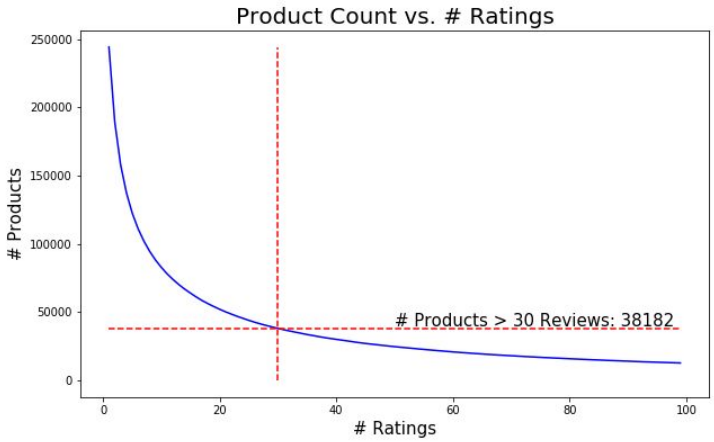
1. Is there a correlation between product ratings on Amazon and various product-related features such as price, number of comments, and the superlatives used in comments?
2. Is there a significant difference in mean rating between product categories or between company brands within a single category?

Data

The dataset we used was sourced with the permission of Professor Julian McAuly from the University of California San Diego. A publicly available sample of the datasets can be found here. This dataset contains product reviews and metadata from Amazon. In total the data accounts for 142.8 million reviews spanning May 1996 to July 2014. The structure is as follows:

- Reviews:
    - Reviewer id - Unique identifier for each review
    - Product id - Unique identifier for each product
    - Reviewer name - Account name of the reviewer
    - Review text - Text of the review
    - Review summary - Summary of the review
    - Product rating - Rating given to the product for each review
    - Helpfulness rating - Number of people that found this review helpful
    - Unix review time - Unix timestamp of review
    - Review time - Datetime timestamp of review
- Product metadata:
    - Product id - Unique identifier for each product
    - Title - Name of product
    - Price - Price of product
    - Image url - URL of image
    - Related products - a list of suggested related products
    - salesRank - salesRank information
    - Brand - Brand to which the product belongs
    - Categories - Categories to which the product belongs

We merged the reviews and metadata into a single dataframe and isolated products that had more than 30 reviews. This was to ensure products with just a few reviews wouldn't become outliers that disproportionately affect the methods that weight the mean rating of each product equally. The visual on the right shows the evolution of the volume of products vs. the number of reviews per product, along with the cut we made for the electronics category. Note, some of the titles had null values. We decided to purge the products with no title to allow for brand comparisons.



Categorical Comparisons

1. Inter-Category:

In our second hypothesis we seek to address the question of whether there is a significant difference in mean rating between categories of products. This is to potentially help elucidate the question over whether people are more critical of products that have a greater capacity for failure.

1.1. Method:

In order to carry out multiple comparisons between different categories, we considered doing an ANOVA test or a series of non-parametric tests with a Bonferroni correction. We made the decision to weight each product in each category equally. To do so we grouped the data in each category by product and averaged the ratings. We chose to examine comparisons between the following four categories: Baby Products, Musical Instruments, Sports/Outdoor Gear, and Electronics. A visualization of the distributions is provided below:

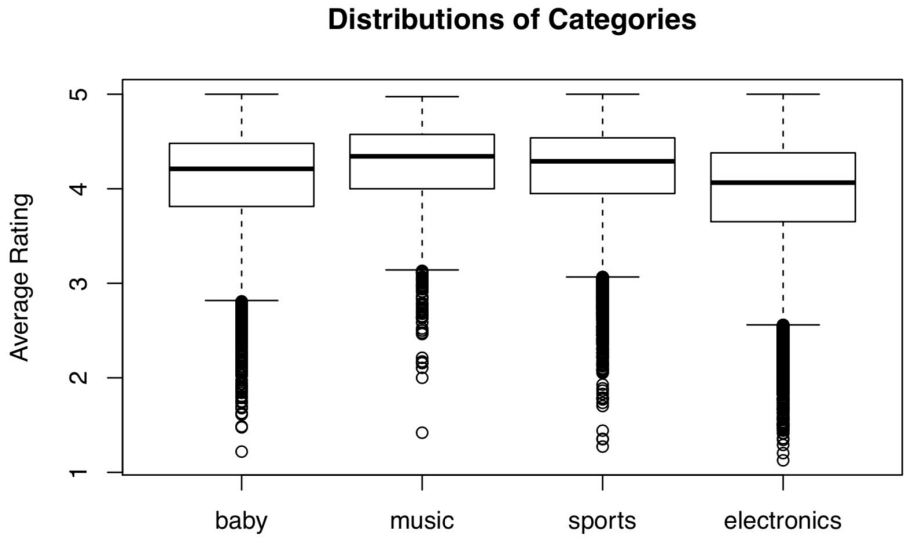


Figure 1: Distributions of 4 product categories.

It seems that the spreads are more or less the same, but each distribution is clearly left skewed. To be conservative and to introduce variety into our report, we decided to run a series of wilcoxon signed-rank tests with a Bonferroni correction to test whether the group means were different or not.

1.2. Results:

The results of the signed-rank tests are summarized in Table 1 below:

Multiple Comparisons Results for α = .0083				
Group 1	Group 2	p-value	Significant?	Direction
Baby	Music	8.90E-33	Yes	Baby < Music
Baby	Sports	1.06E-30	Yes	Baby < Sports
Baby	Electronics	4.79E-70	Yes	Baby > Electronics
Music	Sports	6.26E-07	Yes	Music > Sports
Music	Electronics	9.86E-149	Yes	Music > Electronics
Sports	Electronics	0	Yes	Sports > Electronics

Table 1: Multiple comparisons results across categories.

It turns out that the sample sizes were so large for each category, the p-values were miniscule and the medians were all determined to be significantly different. The pecking order for medians from highest to lowest were music, sports, baby and electronics in that order. Furthermore, the median rating for electronics was substantially lower than the other categories.

1.3. Discussion:

While the differences in the median rating for each category were all significant, it is interesting to see how much lower the median rating for electronics is than the rest. This could be an indication that people, in general, are more critical of electronic products. Harkening back to the example from the introduction, it is possible that this is because the cost of failure is higher for electronics than for similarly priced sports gear, musical instruments, or baby products. As another example, compare a $90 cell phone to a $90 pair of soccer cleats. If your cleats turn out to be uncomfortable and/or fall apart easily, the worst case scenario is you cannot play soccer for the time being. If your cell phone freezes or breaks easily, you cannot make any calls, use it to browse the internet, or a number of key features that people expect cell phones to have these days. However, it is important to remember that this is speculation, and we haven't controlled for several factors that may also influence the median rating in a category. One example of a possible set of confounding variables are the differing

demographics of the people shopping in each category. It is possible that the average age of people shopping for electronics is different than those who are buying baby toys, as young parents who are taking care of children may not have as much time or money to invest in a new speaker system or gaming console. Overall, it is interesting to get directional reads from the categorical comparisons, but in the next section we will take a deeper dive to gain new insights from making intra-category comparisons.

2. Intra-Category:

Average ratings of a company's products on Amazon gives us a measurement of shopping experience of this brand on Amazon. For this part, we want to investigate whether there is significant difference between shopping experiences of brands on Amazon.

To control for variances between product categories, we focused on testing the difference between brands in the same category, for example musical instruments. The brands who receive the most reviews are what we are interested in, not only because they are popular and has more widespread influence, but also it gives us a reasonable way to assume normality using the Central Limit Theorem. We selected the top ten most reviewed brands to perform two kinds of hypothesis tests.

2.1. Methods:

We investigate the problem from two perspectives using:

A. ANOVA test

By performing an ANOVA test, we will have a general idea of whether the ten companies receive the same ratings.

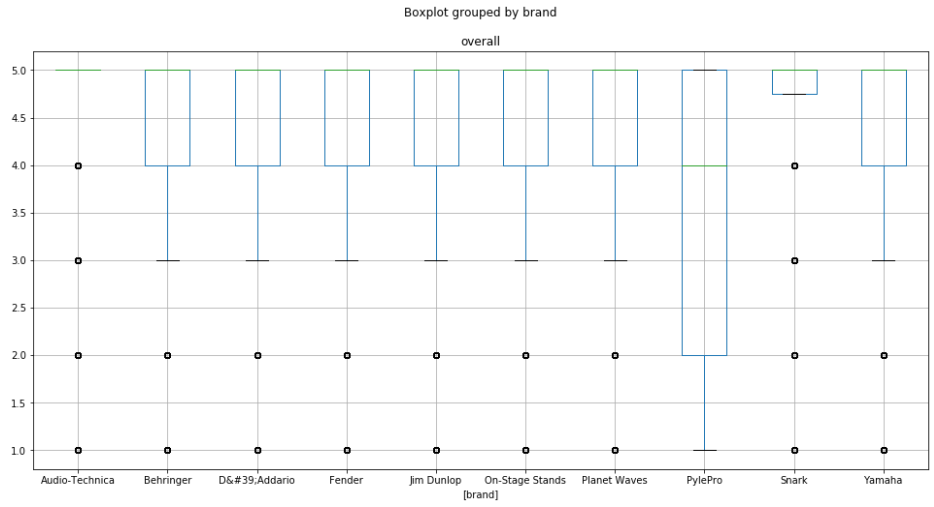


Figure 2: Distributions of top 10 reviewed musical instrument brands.

From this box plot we can see the box plots have similar quantiles except for PylePro, Snark and Audio-Technica. Although the distribution of ratings are right skewed, the number of reviews are more than 3000, and variances are very similar. Therefore, our Anova test results are not likely to be severely affected. A detailed check of assumptions is shown in the Appendix. We carried out ANOVA tests both on the ten groups, and the seven groups that exhibit the same boxplot shape.

B. Tukey HSD test

This test tells us which companies are rated different, and gives us pairwise comparison results, shown in the next section.

2.2. Results:

a. ANOVA tests

ANOVA test for the ten companies Behringer, Planet Waves, PylePro, On-Stage Stands, Jim Dunlop, Yamaha, Audio-Technica, Fender, "D'Addario", Snark:
F-statistic=339.89468397137284, p-value=0.0
ANOVA test for the seven companies Behringer, Planet Waves, On-Stage Stands, Jim Dunlop, Yamaha, Fender, "D'Addario":
F-statistic=62.109158071416374, p-value=5.6075009653532724e-77

It turns out that the companies do not receive the same ratings by customers. We turn to Tukey HSD test the pairwise test to look into details.

b. TukeyHSD results

The results are summarized in the R output table on the following page.

2.3. Discussion:

We can see from the ANOVA test that these companies have different mean ratings over their received reviews. Furthermore, the Tukey HSD test tells us that each company received a rating that is significantly different from the mean rating across all companies. Ranking their ratings from high to low, we get to know that Audio-Technica, Snark and D'Addario receive the best three average ratings among the ten, while PylePro gets a significantly low average rating. After performing the two tests, we know not only which companies are giving out the best online experiences, but also how significantly they differ from other brands.

Multiple Comparison of Means - Tukey HSD,FWER=0.05

group1	group2	meandiff	lower	upper	reject
Audio-Technica	Behringer	-0.4606	-0.5284	-0.3928	True
Audio-Technica	D'Addario	-0.107	-0.1884	-0.0255	True
Audio-Technica	Fender	-0.2841	-0.3595	-0.2086	True
Audio-Technica	Jim Dunlop	-0.1296	-0.2008	-0.0584	True
Audio-Technica	On-Stage Stands	-0.2685	-0.3395	-0.1974	True
Audio-Technica	Planet Waves	-0.2774	-0.3458	-0.209	True
Audio-Technica	PylePro	-1.0119	-1.0819	-0.9418	True
Audio-Technica	Snark	-0.0722	-0.1577	0.0133	False
Audio-Technica	Yamaha	-0.331	-0.4022	-0.2598	True
Behringer	D'Addario	0.3536	0.277	0.4302	True
Behringer	Fender	0.1765	0.1063	0.2468	True
Behringer	Jim Dunlop	0.331	0.2654	0.3966	True
Behringer	On-Stage Stands	0.1921	0.1267	0.2576	True
Behringer	Planet Waves	0.1832	0.1207	0.2457	True
Behringer	PylePro	-0.5513	-0.6157	-0.4869	True
Behringer	Snark	0.3884	0.3075	0.4693	True
Behringer	Yamaha	0.1296	0.064	0.1952	True
D'Addario	Fender	-0.1771	-0.2605	-0.0936	True
D'Addario	Jim Dunlop	-0.0226	-0.1022	0.057	False
D'Addario	On-Stage Stands	-0.1615	-0.241	-0.082	True
D'Addario	Planet Waves	-0.1704	-0.2475	-0.0933	True
D'Addario	PylePro	-0.9049	-0.9835	-0.8263	True
D'Addario	Snark	0.0348	-0.0578	0.1274	False
D'Addario	Yamaha	-0.224	-0.3036	-0.1444	True
Fender	Jim Dunlop	0.1545	0.081	0.228	True
Fender	On-Stage Stands	0.0156	-0.0577	0.0889	False
Fender	Planet Waves	0.0066	-0.0641	0.0774	False
Fender	PylePro	-0.7278	-0.8002	-0.6554	True
Fender	Snark	0.2119	0.1245	0.2992	True
Fender	Yamaha	-0.0469	-0.1204	0.0266	False
Jim Dunlop	On-Stage Stands	-0.1389	-0.2078	-0.07	True
Jim Dunlop	Planet Waves	-0.1478	-0.214	-0.0816	True
Jim Dunlop	PylePro	-0.8823	-0.9502	-0.8144	True
Jim Dunlop	Snark	0.0574	-0.0263	0.1411	False
Jim Dunlop	Yamaha	-0.2014	-0.2705	-0.1323	True
On-Stage Stands	Planet Waves	-0.0089	-0.0749	0.0571	False
On-Stage Stands	PylePro	-0.7434	-0.8112	-0.6756	True
On-Stage Stands	Snark	0.1963	0.1127	0.2799	True
On-Stage Stands	Yamaha	-0.0625	-0.1314	0.0064	False
Planet Waves	PylePro	-0.7345	-0.7994	-0.6695	True
Planet Waves	Snark	0.2052	0.1239	0.2866	True
Planet Waves	Yamaha	-0.0536	-0.1198	0.0126	False
PylePro	Snark	0.9397	0.8569	1.0225	True
PylePro	Yamaha	0.6809	0.613	0.7489	True
Snark	Yamaha	-0.2588	-0.3425	-0.175	True

['Audio-Technica' 'Behringer' 'D'Addario' 'Fender' 'Jim Dunlop'
 'On-Stage Stands' 'Planet Waves' 'PylePro' 'Snark' 'Yamaha']

Figure 3: Unique groups: [Behringer, Planet Waves, PylePro, On-Stage Stands, Jim Dunlop, Yamaha, Audio-Technica, Fender, D'Addario, Snark]

Modeling of Ratings

1. Method:

1.1 Features in Linear Regression

Dependent Variable of Interest: Amazon Product Rating

Our response variable is the mean rating for each product (0-5). However, it's quite left-skewed in its raw unit, which challenges the assumption for constant variance in linear regression. In order to transform the response and at the same time maintain some interpretation power, we flip the response variable around its mean to make it right-skewed and then take log on it (1). The log transformation allows us to interpret the mean rating on the median scale.

$$X_{rating\ transformed} = log(2 * \overline{x_{rating}} - x_{rating}) \quad (1)$$

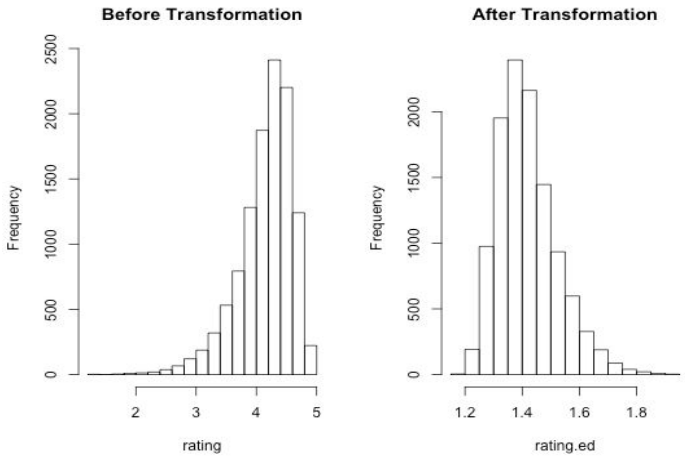


Figure 4: The transformation of the response variable

1.2 Predictors and their transformation:

To examine what determines product rating, we consider the following product-related features:

1. Mean product price
2. Number of reviews per product

Linguistic features

From our raw data, we have data on individual user reviews of each product. We aggregate reviews per product and calculate the follow linguistic features per product:

1. Mean sentiment score per product

2. Mean character count
3. Mean word count
4. Mean number of exclamations and question marks
5. Mean number of words in all caps

We used the exclamation marks and all-caps word counts as a rough measure of extreme emotion in reviews which might have an interaction with the sentiment of the review (e.g., positive reviews are more positive if it has strong emotions, same for negative reviews).

We derived the sentiment score per review using the AFINN dataset. The AFINN sentiment lexicon provides numeric positivity scores for each word. The scores range between -5 and 5, with negative scores indicating negative sentiment and positive scores indicating positive sentiment. We then aggregated the sentiment score per product and used that as a feature on the basis of product ratings might be higher if the average sentiment score of the reviews is positive.

Including the text related features, the predictors have non-linear relationship with the response. We can also see that the constant variance assumption has been clearly violated from the residuals-fitted value plot. Thus, we need to transform some of our features as follow:

- Log-transformed: Mean product price, Number of reviews per product, Mean character count, Mean word count, Mean number of exclamations and question marks.
- Add one and then log-transformed: Mean number of words in all caps.

When we re-plot the residuals-fitted value plot, the non-constant variance problem has been alleviated greatly (figure 6).

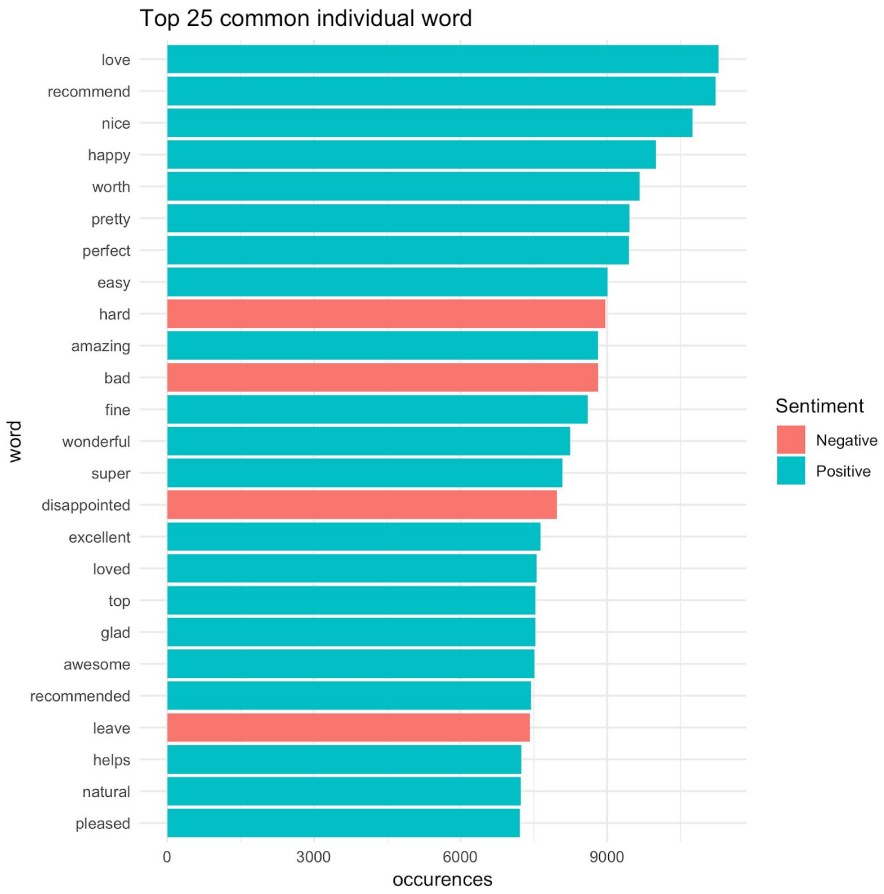


Figure 5: Top 25 common words in reviews of Beauty products, colored by the sentiments of the individual word

Table 2 shows the summary statistics of each predictor in our model after feature engineering. The data is from the beauty category in Amazon with 11336 products.

Table 2: Summary Statistics of All Variables (N=11336)

Variabel Name	Description	Mean	SD
rating.ed	$log(2 * \overline{rating} - rating)$	1.41	0.10
price.ed	log(mean price)	2.67	0.87
num_char.ed	log(mean character count)	5.77	0.32
num_words.ed	log(mean word count)	4.12	0.32
num_allcaps.ed	log(100*Mean number of words in all caps)	0.45	0.31

num_allmark.ed	log(Mean number of exclamations and question marks+1)	4.27	0.45
num_review.ed	log(number of reviews)	4.21	0.68
sentiment	mean sentiment score	1.14	0.48

2. Results:

2.1 Model building

We built our first simple linear regression model (model 1), using the transformed price as the predictor and the transformed rating as the response.

$$Rating_{transformed} = \beta_0 + \beta_1 * Price_{transformed}$$

After inspecting the result, the assumptions of the model fitted well. There is a weak positive relationship (0.0151, $p < .001$) between the product rating and price, which is thought-provoking. Note that we've flipped the rating around its mean, we have to take the opposite direction when interpreting the coefficients in the original unit.

We then incorporated the natural language processing (NLP) features of reviews into our model. Firstly, we included all the main effects of our model (model 2). Model 2 result shows a weak (but significant) positive relationship between price and rating. Meanwhile, most of the NLP features also has significant relationship with rating.

ESS F-test result suggested including the NLP features significantly improves the model with a p-value of <2.2e-16.

2.2 Model selection: Stepwise model selection

To select the best combination between the NLP features and the price predictor, we added all main effects and the 2nd order interactions terms from model 2 (model 3). We then conducted a stepwise model selection with model 1, model 3 as the lower and upper bound of the scope and let the result of the model selection as our final model.

After checking the assumptions for the final model, the constant variance assumption has been violated. From the residuals vs fitted plot and the scale location plot, we observe that variance increases with fitted value, which indicates that a weighted least square approach might help to adjust our model.

Since our independent variables are aggregated,, it's natural to use $\frac{1}{(number\ of\ reviews)}$ as the model weight. However, the non-constant variance problem still exists (see figure 7). To get a more reliable result on our inference, we ran 500 simulations to construct the 95% bootstrapped confidence intervals for each coefficients.

	Final model WLS	Bootstrap CI Lower Bound	Bootstrap CI Upper Bound
price.ed	-0.182 (.076)	-0.363	0.049
num_char.ed	**-1.746*** (.345)	-2.443	-0.876
num_words.ed	**1.152*** (.355)	0.224	1.872
num_allcaps.ed	**0.162*** (.055)	0.044	0.310
num_review.ed	**0.188*** (.118)	0.037	0.457
sentiment	**-1.580*** (.141)	-2.029	-1.309
num_char.ed:num_words.ed	**0.063*** (.007)	0.052	0.083
sentiment:num_allcaps.ed	**-0.036*** (.006)	-0.055	-0.022
price.ed:num_char.ed	0.095* (.048)	-0.051	0.211
num_allcaps.ed:num_char.ed	**-0.038*** (.009)	-0.060	-0.021
num_allcaps.ed:num_review.ed	**0.024*** (. 005)	0.017	0.035
num_char.ed:num_review.ed	-0.096 (.073)	-0.252	0.005
price.ed:num_allcaps.ed	**0.009*** (.003)	0.003	0.015
price.ed:num_words.ed	-0.090. (.049)	-0.211	0.059
num_words.ed:num_review.ed	0.085 (.075)	-0.019	0.240
sentiment:num_char.ed	**0.907*** (.087)	0.730	1.178
sentiment:num_words.ed	**-0.908*** (.099)	-1.190	-0.727
price.ed:sentiment	**-0.005*** (.002)	-0.010	-0.0004
Intercept	**5.431*** (.568)	4.001	6.551
R-squared	0.350		

*Notes: Standard errors in parentheses. 0.05 significance indicated by *

Table 3: Linear Regression of final model with 18 features (N=11336)

From the table above we can see that, the estimate from final model indicates that price still has a positive relationship with the rating, however the confidence interval indicates that this relationship is not significant controlling for all important confounding variables.

From the model, we can see that log-transformed price has a positive (insignificant) relationship with mean log-transformed product rating. Interestingly, most of the linguistic features have significant relationship with rating. Particularly, sentiment score of reviews has a significant positive relationship (with a coefficient of -1.58) with product rating, meaning that the more positive the reviews are (indicated by sentiment scores), the higher the product rating.

3. Discussion:

Natural Language Processing Features

We have applied standard natural language processing on the product reviews. However, we did not have explore different processing features. For example, we used the number of all cap words as a measure of extreme emotion, however, some common emotion-neutral acronyms are all caps, e.g., PDF, DVD which we did not account for.

We have only considered unigram (one word unit) in our features, based on how the ASFINN lexicon is trained and computation simplicity. We can also consider different ngrams (e.g., batch of words) to account for sentiments of batch of words (e.g., consider “not good” as negative) in the future.

We used ASFINN lexicon to weigh the sentiments of individual words in the review. ASFINN lexicon was trained on Twitter which might not contain domain-specific words that are used in reviews in product. Since we can only estimate the sentiment score of reviews that contains words that exist within ASFINN lexicon (2477 words), we might lose some information of reviews that use domain-specific that is not included in the ASFINN lexicon. If possible, we can look for lexicon that is trained on review domains and particular product-related domain (i.e., in our case, lexicon trained on beauty product review) and calculate the sentiment of the Amazon beauty product reviews based on the domain-specific lexicon.

Weighted function

We have applied the weighted least squares to adjust for the non-constant variance problem. However, we’ve tried different functions of the weight, such as 1/(number of reviews), 1/(predicted values), 1/(number of reviews*predicted values), none of them has a good performance on alleviating the problems and even worsen the distribution of the residuals. Therefore, we turned into another way of doing this, that is conducting WLS with unknown

weights and did the weighted process by Iteratively Reweighted Least Squares. However, that still performs poorly on our problem.
Therefore, we came to use bootstrap to get a reliable results on the inference. We only used the estimates from the WLS for our final model and made judgement on whether the coefficients are significant or not on the confidence intervals drawn by bootstrap.

Bibliography

1. https://www.scrapehero.com/many-products-amazon-sell-january-2018/
2. https://www.statista.com/statistics/266282/annual-net-revenue-of-amazoncom/
3. http://cs229.stanford.edu/proj2014/Jordan%20Rodak,%20Minna%20Xiao,%20Steven%20Longoria,%20Predicting%20Helpfulness%20Ratings%20of%20Amazon%20Product%20Reviews.pdf
4. Lars Kai Hansen, Adam Arvidsson, Finn Årup Nielsen, Elanor Colleoni, Michael Etter, "Good Friends, Bad News - Affect and Virality in Twitter", The 2011 International Workshop on Social Computing, Network, and Services (SocialComNet 2011).

Appendix

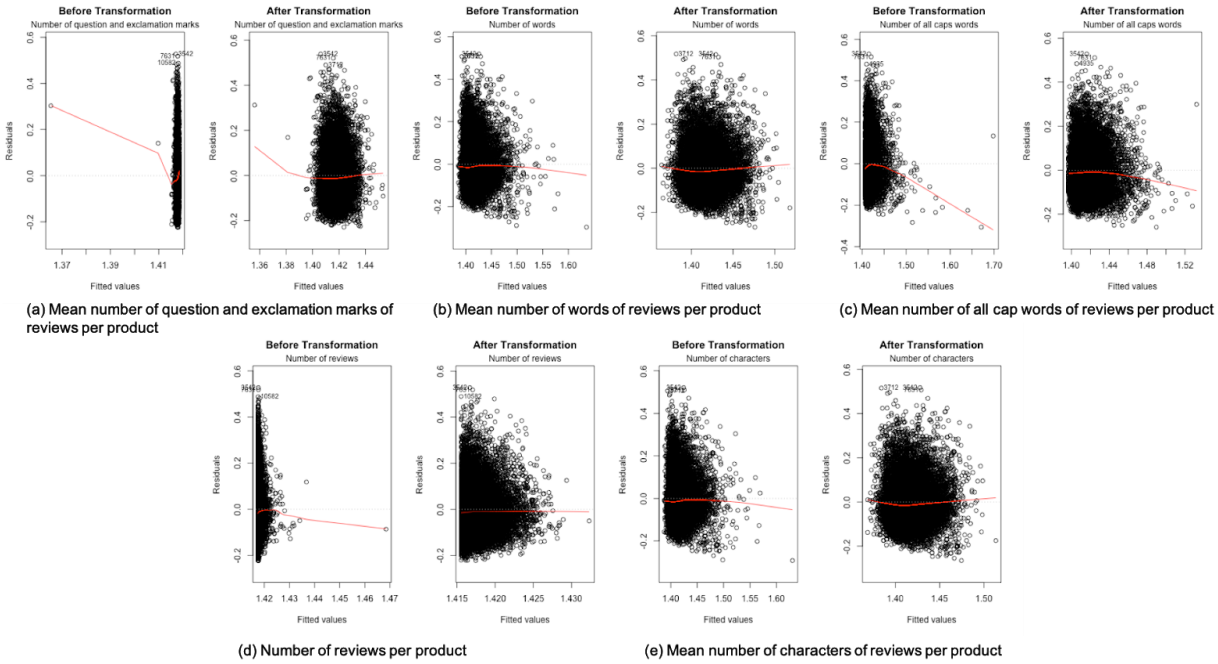


Figure 6: The transformation of the predictors

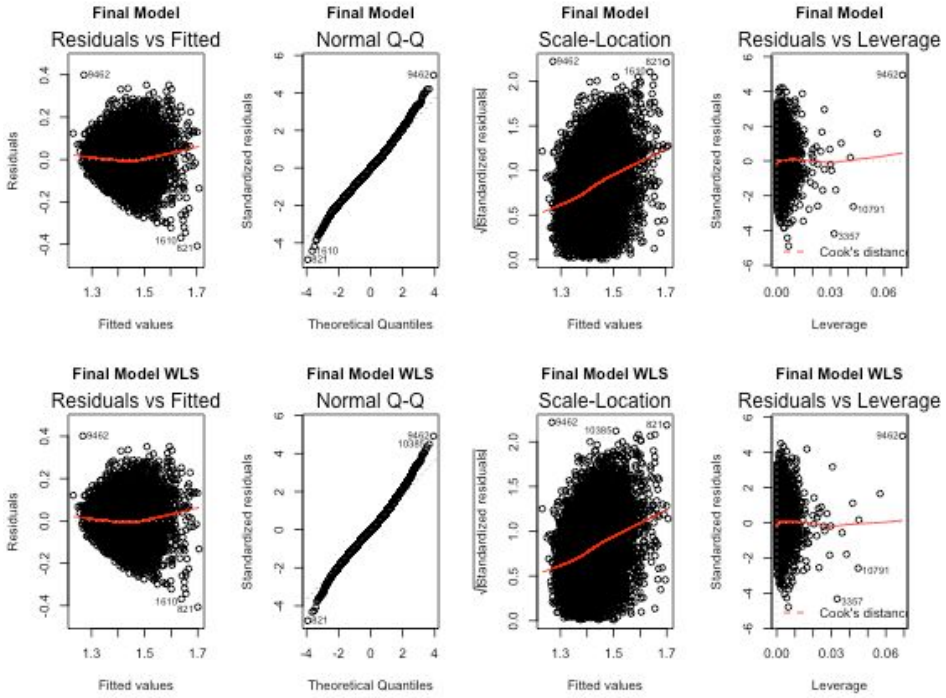


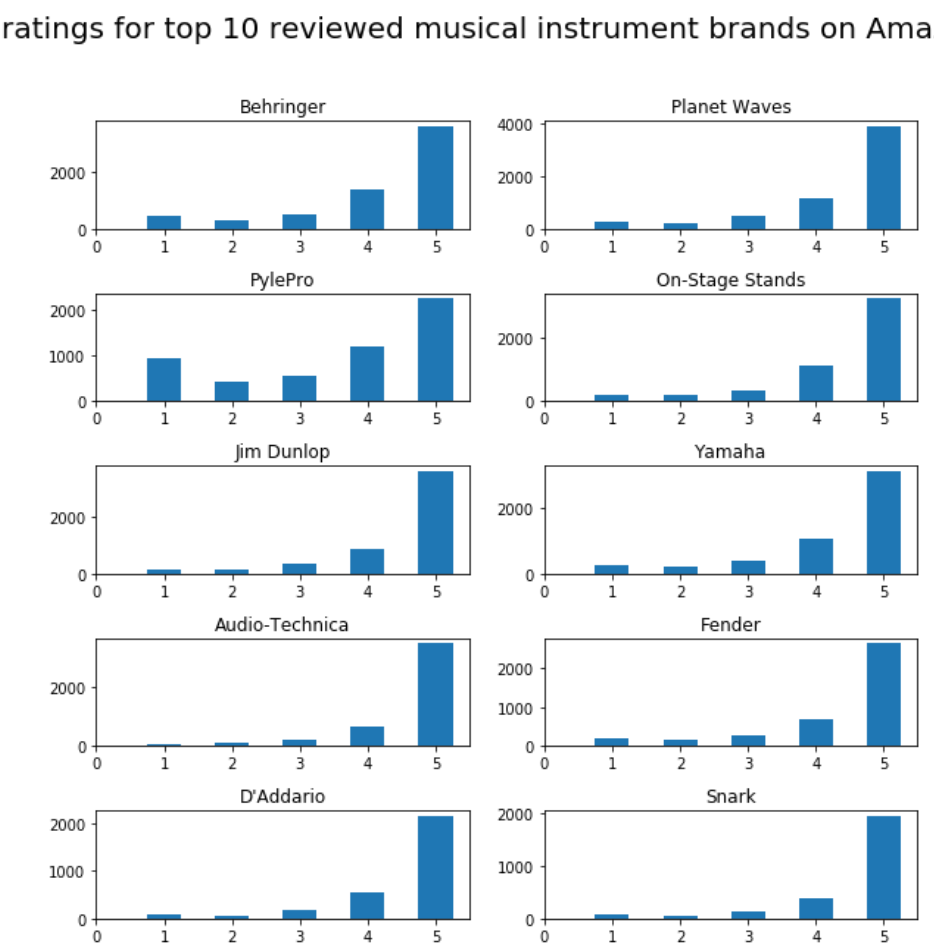
Figure 7: The Assumption Check with the final model

	Model 1: Simple Linear Regression	Model2: All main effects
price.ed	-0.015*** (.001)	-0.010*** (.001)
num_char.ed		-0.639*** (.042)
num_words.ed		0.628*** (.043)
num_allcaps.ed		0.029*** (.003)
num_allmark.ed		0.001 (.002)
num_review.ed		-0.004*** (.001)
sentiment		-0.121*** (.002)
Intercept		2.685*** (.068)
R-squared	0.016	0.331

*Notes: Standard errors in parentheses.

Sig: <0.001 '***', <0.01 '**', <0.05 '*', <0.1 '.'

Table 2: Coefficients of model 1 and 2, with standard error and significance



A check of assumptions on the ANOVA test across musical instrument brands

The histograms prompt us that PylePro, Snark and Audio-Technica might have different variances and so they are. Therefore, we not only performed tests on the ten companies, but also tried to see if the rest seven companies receive the same comments. The seven groups have variances from 0.8 to 1.4, which is in good compliance to equal variance assumptions. They are all left skewed in similar fashion, but we have large sample sizes from 3000 to 6000 each, therefore the law of large numbers will come into play, give us a reasonably good normality assumption. Therefore, ANOVA and also Tukey HSD assumptions are satisfied