Amazon Helpfulness: Milestone Report I

Amazon is one of the largest ecommerce marketplaces in the world with millions of users making transactions daily. These users can review the products they buy which help others make their decision whether to also buy. The reviewer gets points in the form of votes depending on how helpful the review was to others. With enough points that reviewer may qualify for free items - as long as you have more helpful votes than actual reviews. What actually makes a review helpful for other users? Knowing the answer to this and becoming a top reviewer could result in free products.

Data Wrangling:

The data comes from UCSD with reviews up to 2018. The data includes reviews (ratings, text, helpfulness votes), product metadata (descriptions, category information, price, brand, and image features). The data was divided into 29 categories - however I was unable to include all of the categories due to my lack of memory. I was limited to datasets below ~10 Million reviews, leaving me with 21 categories to process.

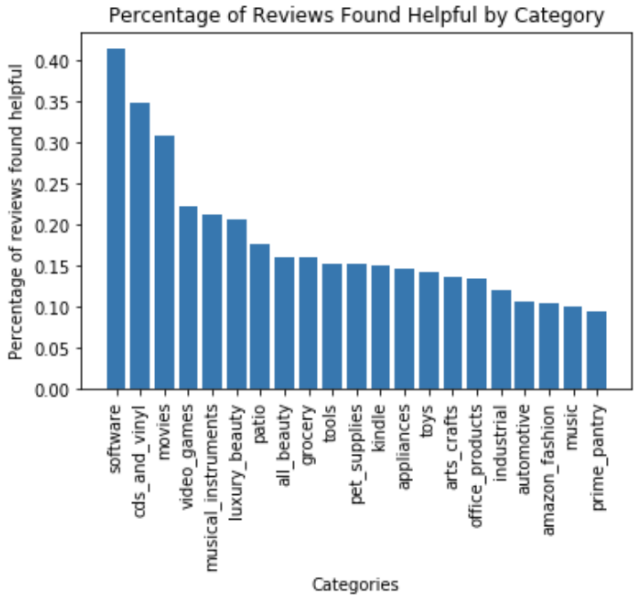
The goal was to take equal random samples from each category and combine them into one large dataset. For each category a sample of 3000 random products were selected through the ASIN (Product ID) that have at least 5 reviews. From these products only 5 reviews were selected randomly- resulting in a dataframe with 15000 rows. I then added a column indicating which category the product came from. These were merged with the metadata and saved as a csv file - repeating until there was a dataframe for each category.

All csv files had only relevant columns selected. I concatenated all of the dataframes into one final dataframe.

Data Story:

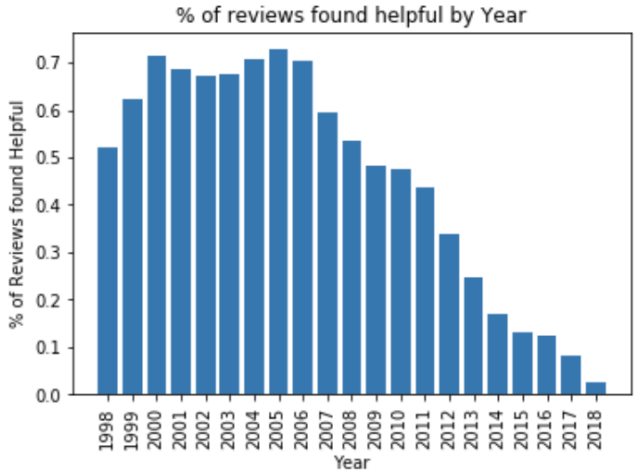
After cleaning the data I added a column to indicate whether or not a review was found helpful. This was done by simply seeing if it had any helpful votes and creating a separate binary column. I took the sum of helpful reviews and found that the percentage of reviews that are found helpful are about 17% of all reviews.

I split the data and found how many reviews are helpful by category:



We see that each category is different in their share of helpful reviews. The categories with the highest percentage of helpful reviews are Software (41.4%), CD's and Vinyl (34.7%) and Movies (30.7%) followed by Video Games (22.1%). The categories with the lowest percentage of helpful reviews are Prime Pantry (9.46%), Digital Music (10%) and Amazon Fashion (10.3%). There is a clear distinction between the categories for the amount of helpful reviews.

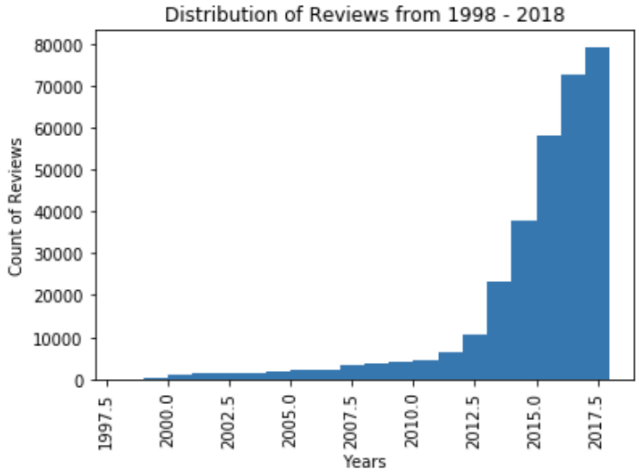
I also plotted time against the percentage of reviews found helpful:



Year of posting appears to have a dramatic effect on the % of reviews that are found helpful. This appears to indicate that the length of time that a review is available on Amazon is a large factor in determining whether or not it gets helpful votes, with effects plateauing out after around 12 years.

We see that older reviews are found more helpful than newer reviews. This is likely because it gives more consumers exposure to the review. There appears to be a decline in reviews found helpful after 2006.

I plotted the distribution of reviews based on the date posted:



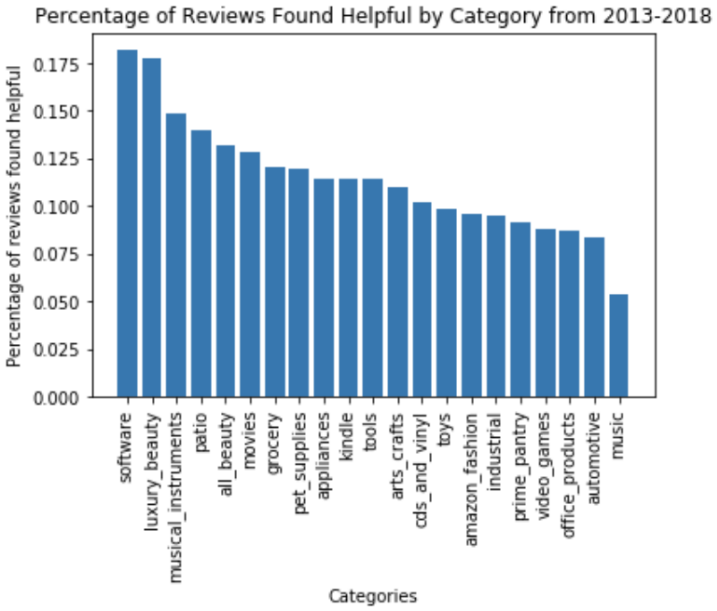
The distribution of reviews are skewed to the left, with a majority of reviews from the dataset posted recently. I queried the data for reviews posted after 2013 and found that this contains roughly 79% of the data.

I looked at the quantity of reviews for each category:

prime_pantry	14874
amazon_fashion	14299
industrial	13686
arts_crafts	13576
automotive	13394
appliances	13338
all_beauty	13273
patio	13168
luxury_beauty	13058
tools	12940
pet_supplies	12727
office_products	12517
grocery	12417
toys	12189
musical_instruments	11771
kindle	11669
music	10354
video_games	8485
movies	8189
software	6278
cds_and_vinyl	5689

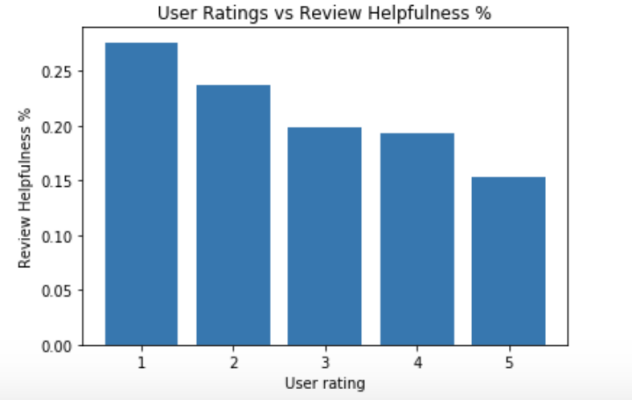
The four categories with the lowest amount of reviews are video games, movies, software and cds. Interestingly these categories had the highest share of helpful reviews before the older dates were filtered out. This underlines the importance that time has on review helpfulness. These categories may not have had the highest share of helpful reviews because they were digital, but because they had the highest amount of older reviews. I removed the older years (2012 and prior) to observe what effect it would have on the percentage of reviews found helpful.

I plotted the share of helpful reviews by category after 2013:



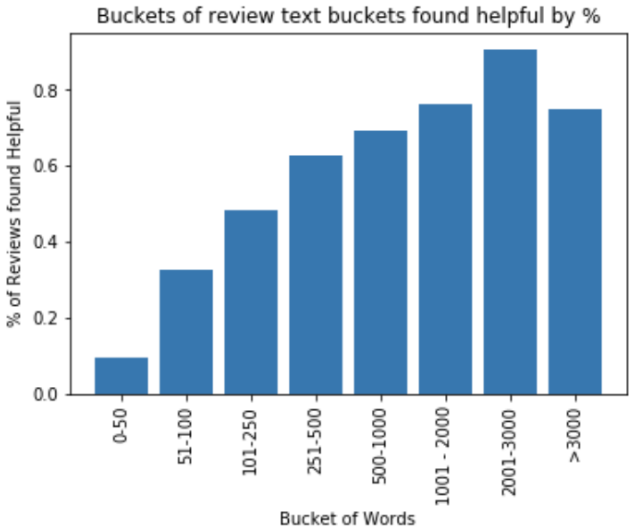
Without the older reviews - software is still on top with the highest percentage of helpful reviews however the next categories are luxury beauty, musical instruments and patio. The other 3 previous most helpful categories (movies, cds and video games) have all fallen out of the top 4, further highlighting the importance that time has on helpful reviews. We also see that the highest percentage of helpful reviews decreased from 41% to 18%.

After analyzing the effect that time has on review helpfulness, I plotted the percentage of reviews that are found helpful by the rating that each reviewer had given.



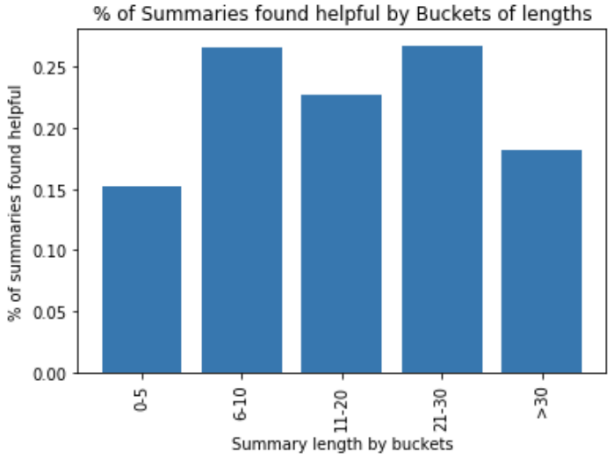
From the chart we see that a higher percentage of reviews that rate the product lower are found helpful. Reviews that are more critical of the product are helpful to consumers proportionately more than reviews that just praise the product.

I plotted buckets of review lengths to see if longer reviews resulted in a higher share in helpful reviews:



We see that longer reviews are typically found more helpful. A majority of reviews over 250 words are found helpful (62% and above). However, this peaks at between 2001-3000 word reviews, with helpfulness declining modestly after that.

I plotted buckets of summary lengths to see if there was a similar effect:



Unlike for review text, it appears that a longer summary length is not exactly indicative of a more helpful review. There appears to be a sweet spot between 6 and 30 words, with very long and very short summary lengths having lower helpfulness ratings.

Statistical Inference:

I conducted a chi-squared test to see if the frequencies of helpful and unhelpful reviews are different. For this I took a random sample of 10,000 helpful and 10,000 unhelpful reviews and used the review length buckets for comparison. When conducted on review lengths, the chi-squared test yielded a p-value of 0.0 which makes it clear that helpful reviews are longer than unhelpful ones.

I also conducted a chi-squared test for the frequencies of helpful and unhelpful summaries. I used the same sample of 10,000 helpful and 10,000 unhelpful summaries and used the summary length buckets for comparison. When conducted on summary lengths, the chi-squared test yielded a p-value of 6.610e-98. Helpful summaries are significantly longer than unhelpful summaries.

Helpful/Unhelpful Words:

I used a Naive Bayes classifier to predict helpful reviews and helpful summaries then found words that had both the highest and lowest probabilities of being in a helpful review. The results varied for each category, review text and summary text.

I listed the words with the highest probability of being helpful with their probabilities and the words with the lowest probability of being helpful with their probabilities.

Helpful Probability by Word: Review Text

Helpful words	P(Helpful | word)
scherzo	0.89
sonata	0.88
seventh	0.88
brahms	0.87
bartok	0.86
richter	0.86
quartets	0.84
frontpage	0.83
schubert	0.83
poser	0.81
Unhelpful words	**P(Helpful | word)**
manifold	0.02
historia	0.02
drumstick	0.02
carburetor	0.02
goodie	0.02
reeds	0.02
libro	0.02
peavey	0.02
producto	0.01
sheath	0.01

We see that words that are listed in languages other than english (libro, producto) have low probabilities of being considered helpful. The helpful words are typically terms or names that are associated with the specific product (scherzo and sonata are latin words used in classical music). There are an unproportional amount of musical terms and brands in this list. I've then listed the helpful and unhelpful words associated with helpfulness for the summary text.

Helpful Probability by Word: Review Summary

```
Helpful words        P(Helpful | word)
           acquaint 0.85
           baritone 0.85
         challenged 0.85
                 96 0.80
          collision 0.71
               crop 0.67
             coffee 0.66
             crunch 0.66
             dances 0.65
              crazy 0.64
Unhelpful words      P(Helpful | word)
            confirm 0.04
             cooked 0.04
          categorize 0.04
           abilities 0.04
            creation 0.04
               cakes 0.04
            bedtime 0.03
             ballad 0.03
        compression 0.03
               babe 0.03
```

The summary text words overall are very different from the review text. I don't really see a pattern in the summary words.

Next, I performed the same analysis but with a breakdown by category:

Helpful Reviews: Music

	Helpful_words	Helpful_prob	Unhelpful_words	Unhelpful_prob	category
0	op	0.938433	chant	0.018605	music
1	adagio	0.833471	whenever	0.018605	music
2	di	0.812117	sea	0.018605	music
3	ii	0.803728	hearing	0.017530	music
4	walter	0.803728	blessing	0.017199	music
5	andante	0.784481	remix	0.015990	music
6	verdi	0.761050	cassette	0.014019	music
7	concerto	0.761050	lol	0.013206	music
8	cobham	0.761050	oldie	0.013206	music
9	liszt	0.747313	relate	0.012481	music

Helpful Reviews: Automotive

	Helpful_words	Helpful_prob	Unhelpful_words	Unhelpful_prob	category
0	thule	0.856139	opening	0.020454	automotive
1	lbs	0.789856	chain	0.019200	automotive
2	dump	0.652696	yes	0.019200	automotive
3	charcoal	0.632719	miles	0.019200	automotive
4	buffer	0.632719	thread	0.018629	automotive
5	65	0.610303	toyota	0.017750	automotive
6	tv	0.610303	described	0.016873	automotive
7	fasten	0.610303	2000	0.014696	automotive
8	crv	0.610303	amaze	0.014037	automotive
9	weigh	0.610303	explorer	0.013435	automotive

Helpful Reviews: Appliances

	Helpful_words	Helpful_prob	Unhelpful_words	Unhelpful_prob	category
0	iron	0.700845	exactly	0.018365	appliances
1	warn	0.670747	shipment	0.016693	appliances
2	cooker	0.670747	perfect	0.016621	appliances
3	capacity	0.670747	took	0.016036	appliances
4	alert	0.647074	received	0.015429	appliances
5	cascade	0.647074	saved	0.015429	appliances
6	claim	0.647074	promised	0.015429	appliances
7	soda	0.647074	fits	0.015236	appliances
8	queen	0.633914	worked	0.012809	appliances
9	utensil	0.619734	described	0.008331	appliances

I looped through the dataframe to find the most predictive words per category for review text and found that the exact words are very different across each category, however the type of words across categories have similarities. Words that describe a family member are common for bad words across categories (nephew, grandson, granddaughter). It appears that reviews for products that were purchased as a gift for someone are rarely viewed as helpful. For good words I noticed that the words that are helpful are those that pertain to the specifics of each category. These words are usually specific brands or terms for each category.

Next I looped through the summaries for all categories:

Helpful Summaries: Music

	Helpful_words	Helpful_prob	Unhelpful_words	Unhelpful_prob	category
0	performances	0.735191	use	0.038149	music
1	important	0.689542	happy	0.035696	music
2	power	0.689542	want	0.033540	music
3	read	0.689542	awesome	0.032749	music
4	beyond	0.624876	artist	0.031629	music
5	search	0.624876	song	0.029442	music
6	type	0.624876	four	0.015617	music
7	under	0.624876	three	0.014402	music
8	sad	0.624876	stars	0.007227	music
9	loves	0.624876	five	0.006528	music

Helpful Summaries: Automotive

	Helpful_words	Helpful_prob	Unhelpful_words	Unhelpful_prob	category
0	compact	0.692397	way	0.030315	automotive
1	scanner	0.628005	job	0.028766	automotive
2	inside	0.628005	installation	0.028766	automotive
3	mind	0.628005	tool	0.027367	automotive
4	mine	0.628005	bad	0.022910	automotive
5	give	0.584517	fine	0.020417	automotive
6	weak	0.529516	them	0.019035	automotive
7	repair	0.529516	stars	0.016598	automotive
8	base	0.529516	five	0.015468	automotive
9	lol	0.529516	four	0.009092	automotive

Helpful Summaries: Appliances

	Helpful_words	Helpful_prob	Unhelpful_words	Unhelpful_prob	category
0	stainless	0.751827	fits	0.036487	appliances
1	avoid	0.751827	fit	0.033618	appliances
2	food	0.707907	arrive	0.032564	appliances
3	reliability	0.707907	replace	0.031712	appliances
4	beautiful	0.707907	original	0.029404	appliances
5	blue	0.707907	item	0.026802	appliances
6	stop	0.707907	four	0.024482	appliances
7	steel	0.707907	three	0.020125	appliances
8	cool	0.679551	stars	0.016704	appliances
9	garbage	0.645097	five	0.016694	appliances

For summaries I found in every category that the most common unhelpful words are numbers and stars (i.e. “Five Stars”).

After querying by dates later than 2013, I found the helpful and unhelpful words for each category:

Helpful words after 2013: Prime Pantry

	Helpful_words	Helpful_prob	Unhelpful_words	Unhelpful_prob	category
0	regardless	0.225628	stuff	0.000238	prime_pantry
1	hemp	0.172268	some	0.000235	prime_pantry
2	sesame	0.142732	flavor	0.000208	prime_pantry
3	gram	0.142732	taste	0.000203	prime_pantry
4	consume	0.099908	good	0.000186	prime_pantry
5	nutrient	0.097654	them	0.000182	prime_pantry
6	hardly	0.088525	these	0.000168	prime_pantry
7	seed	0.086875	try	0.000159	prime_pantry
8	wave	0.076850	love	0.000151	prime_pantry
9	fatty	0.076850	great	0.000043	prime_pantry

Helpful words after 2013: Software

	Helpful_words	Helpful_prob	Unhelpful_words	Unhelpful_prob	category
0	aftershot	0.513592	easy	0.001616	software
1	32	0.363130	works	0.001528	software
2	plugin	0.232875	hard	0.001335	software
3	color	0.197672	great	0.001271	software
4	black	0.191968	love	0.001226	software
5	transcription	0.174357	found	0.001226	software
6	spot	0.174357	dvd	0.001146	software
7	written	0.174357	computer	0.000858	software
8	proficient	0.174357	game	0.000837	software
9	lightroom	0.174357	learn	0.000621	software

Helpful words after 2013: Patio

	Helpful_words	Helpful_prob	Unhelpful_words	Unhelpful_prob	category
0	3rd	0.138617	hair	0.000408	luxury_beauty
1	warning	0.138617	bottle	0.000348	luxury_beauty
2	physical	0.118246	cream	0.000313	luxury_beauty
3	blackhead	0.107694	product	0.000302	luxury_beauty
4	coola	0.091386	love	0.000297	luxury_beauty
5	burnt	0.074470	good	0.000291	luxury_beauty
6	opt	0.074470	from	0.000259	luxury_beauty
7	mainly	0.074470	scent	0.000235	luxury_beauty
8	newborn	0.074470	great	0.000182	luxury_beauty
9	chipping	0.074470	smell	0.000134	luxury_beauty

While the predictive power of helpful words are not very strong, the most interesting pattern emerged after removing the reviews prior to 2013; words that are subjective to the user have are unhelpful. Words like “great”, “love” and “perfect” are all unhelpful words while the helpful words are objective. They’re primarily words that are used to describe the product, not

the user's sentiment. This pattern also emerged in the summaries for reviews posted in 2012 and later.

Conclusion:

I've found that there are patterns in how reviews are deemed helpful vs not helpful. There are several differences between the two groups.

- The categories with the highest percentage of helpful reviews are those that typically deal with technology - software, cds/vinyl, movies.
- Reviews that reference the product with terms that pertain to the category specifically are found more helpful - i.e. musical terms such as sharp, flat and adagio have a higher probability of being helpful in the music category.
- Words with a bad probability of being helpful are those that pertain to a family member. These are reviews for products that were purchased for someone else (grandson, granddaughter, nephew).
- Words that are subjective and share the reviewers sentiment are much more likely to be found unhelpful than words that are objective which describe the product.
- For summaries in every category, the words denoting a number and "stars" showed up as bad predictors for helpfulness.
- Reviews that are more critical of the product are also found more helpful than those that praise the product.
- Older reviews have a higher rate of being helpful - Time has a dramatic effect on the helpfulness of reviews.
- Helpful reviews and summaries are longer than unhelpful reviews and summaries.