Video Game Reviews

What characteristics make a good review?

David Bierer, Danny Moncada, Andrea Montez
MBSA 6330 Team 3
Date: Aug 22,2019

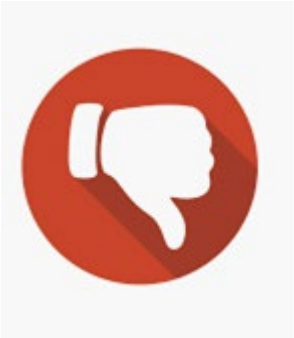
Agenda

- Problem Setup
- Approach
- Data Transformation
- Data Analysis: Setup and Results
- Summary

Problem Setup

Users rely on product reviews to make purchases, but the helpfulness of a review can vary

***Amazon mission statement**: Our vision is to be earth's most customer-centric company; to build a place where people can come to find and discover anything they might want to buy online.*



If Amazon can learn what distinguishes helpful reviews from unhelpful reviews, it can:

- encourage users to leave better reviews
- find the best reviews and place them towards the top, helping users get good information faster

All while enhancing their customer-centricity focus.

Some Key Questions

Before beginning any analysis, it's important to understand the objective.

Here are some key questions that shaped the analysis:

- ✓ Are there specific words that are likely to indicate whether a rating is good or bad?
- ✓ Are there specific words that are likely to indicate if a review was considered helpful by other users?
- ✓ Can we build models to predict rating or helpfulness?
- ✓ Are there certain users who leave a lot of reviews? Are they consistently helpful?
- ✓ Are there products that get reviewed noticeable more than others?
- ✓ Do any of these variables change over time

Approach

Approach

- **Data:**
  - 231,780 video game reviews collected over 15 years

- **Objective:**
  - **Exploratory:** what insights about product, time, and users can be gleaned from the data-set?
  - **Predictive:** what are characteristics associated with a good rating or a helpful review?

- **Methodology:**
  - Turn user-reviews into helpful information using ***natural language processing***
  - Create models, using ***logistic regression*** , to understand the impact of certain words
  - Calculate ***share of reviews*** per product to see which products had a large portion of reviews
  - Utilize different metrics ***over time*** to understand the shifts in trends
  - Visualize the information, using ***Seaborn***, to make the information easy to digest

Description of original data

Column	Type	Description
asin	string	product identifier
helpful	array	includes **helpfulness numerator:** number of users who found the review helpful & **helpfulness denominator:** number if users who indicated whether they found the review helpful or not
overall:	double	product rating between 1 & 5
reviewText:	string	Review text (the body of the review)
reviewTime	string	month, day and year of review
reviewerID:	string	Reviewer ID
reviewerName	string	username of reviewer
summary	string	Review Title
unixReviewTime	long	Review time in unix

Data Transformation

Natural Language Processing

- The words from each review must be extracted &counted; this was the basis of many of our different analyses

- The review title and text were combined to form a single text column, as there was unique information in both. By combining, it is more efficient as you don't have to run the analysis on multiple columns

- In order to analyze, all words within the summary and review needed to be a list of words
  - "Tokenizer" splits all words into a single array
  - "Stop-Words" removes often-used, but meaningless words (the, that, a, at, etc.)

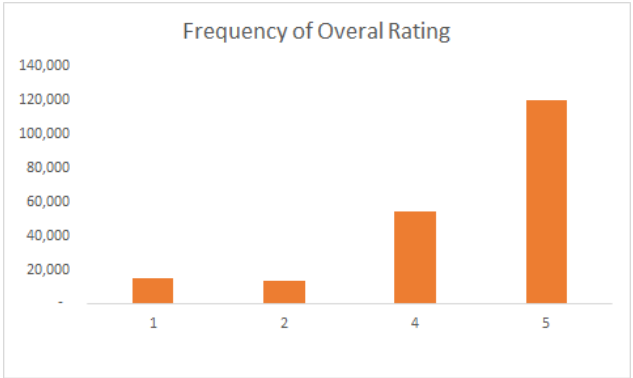
Understanding Star -Rating

- In creating a regression model, we will be able to use the model to predict whether a product got a good/ rating based on the review alone

- A logistic regression predicts a binary result.
  - Therefore, overall rating was converted to a 0/ 1 variables
    - 4-5 Star reviews were converted into a good rating (or 1)
    - 1-2 Start reviews were converted into a bad rating (or 0)
    - 3-Star ratings were excluded, as they were neutral, neither good nor bad

- Ratings skew left, but provide enough variation for a model to learn from

overall	label	count
1.0	0.0	14853
2.0	0.0	13663
4.0	1.0	54804
5.0	1.0	120185



Understanding Helpfulness

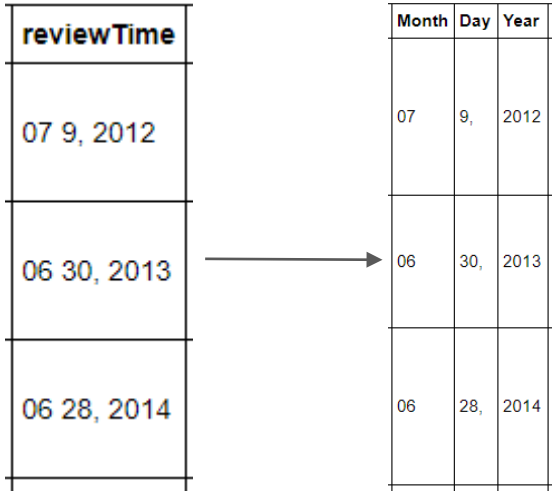
- Helpfulness tracks the thumbs-up/ thumbs-down response from users
- Initially given in an array, helpfulness was split into three distinct columns
  - helpful votes
  - total votes
  - helpfulness %
- We also passed this to a regression model, creating a need for binary variables:
  - 60%+=helpful review (or 1)
  - 40%or less =not helpful (or 0)
  - 40%- 60%excluded from analysis

	asin	helpful
0	0700099867	[8, 12]
1	0700099867	[0, 0]
2	0700099867	[0, 0]

helpful_votes	total_votes	perc_helpful
8	12	66.666667
0	0	NaN
0	0	NaN

Extracting Time from the data -set

- Review Time was captured in an array with month day year
- Review_Time was split to create 3 distinct columns and then cast as double in order to perform mathematical functions



Logistic Regression

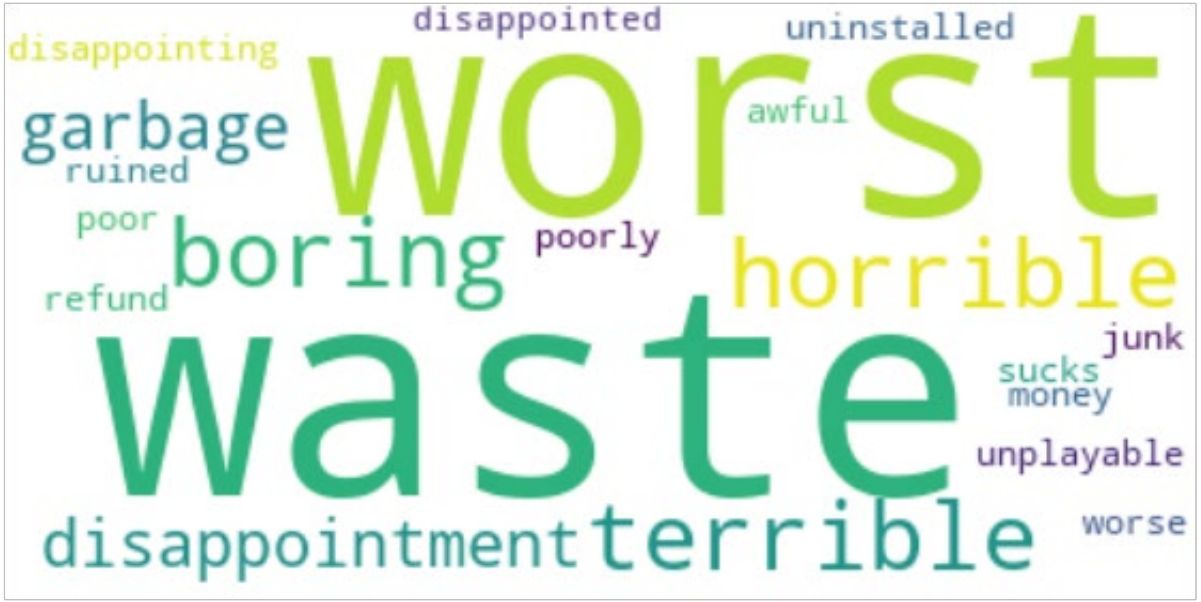
Brief Outline of the Process

- Create a training and test data set
  - We use the training data set to "train" our model
  - We use the test data set to see how well our model does
- Convert all the review text into array of words
  - Remove all "stop" words that don't provide any additional context
  - Find the number of times that each one of those words appears in the text
    - This is what we use as our "features" column for doing our predictions
  - "Down-weight" our features column for any of those words that might appear frequently and skew our results; we can now measure the *significance* of the word in the review.
- We create two models:
  - Overall Rating
  - Review Helpfulness
- The goal of our models is to predict the overall rating of product and the helpfulness of a review based on the review text

Rating: Top 20 Wordcloud

"love", "great", "fun" were all among the top words, and are not surprising words to appear in a positive review

Rating: Bottom 20 Wordcloud



No surprises in the words generated from poor reviews either, i.e. “worst”, “waste”, and several forms of “disappointed”

Overall Rating: Evaluation Results

AreaUnderROC = 0.92
f1 = 0.87
Accuracy = 0.90
weightedPrecision = 0.90
weightedRecall = 0.90

Interpretation: Overall, this is a reliable model as AreaUnderROC is 92%, which is considered an excellent (A) score, meaning it is able to tell the two groups (good/ bad reviews) apart. Precision and recall are nearly identical which means the model is pretty balanced. The model finds about 90% of positive reviews (recall) and guesses correctly 90% of the time (precision)

Helpfulness: Top 20 contributing words

- The words that contribute most toward a helpful review are all positive adjectives

- This is somewhat surprising, as "bad" words (i.e. "disappointing" or "garbage") are also considered helpful, but just in deterring purchases, rather than helpful in making them

	word	coefficients
10	fun	0.085536
343	highly	0.085198
9	great	0.084925
0	game	0.071756
25	best	0.071030
303	puzzles	0.055701
337	challenging	0.054488
801	addictive	0.051192
1408	hooked	0.050746
85	easy	0.047211
32	little	0.046590
247	simple	0.045256
42	different	0.044453
235	excellent	0.043981
199	definitely	0.043520
3444	must-have	0.043091
1300	fun!	0.043019
3	games	0.042793
479	including	0.042267
241	perfect	0.041736

Helpfulness: Evaluation Results

- AreaUnderROC = 0.74
- f1 = 0.62
- Accuracy = 0.70
- weightedPrecision = 0.70
- weightedRecall = 0.70

Interpretation: Overall, this is a moderate model as AreaUnderROC is only 74%, which is considered an fair (C) score. It is not able to distinguish between helpful/ not helpful reviews, as easily as our other model out with star-ratings. Precision and recall are nearly identical which means the model is pretty balanced. The model finds about 70% of positive reviews (recall) and guesses correctly 70% of the time (precision). While it can produce helpful results, making large decisions off of this model should be questioned and further investigation taken.

Leads to More Questions

Why does this model having lower scores than the Star-Rating model? And why are there no negative words in the Top 20 Helpful Words list? Is there some bias in the dataset? Are only good or only bad products are getting reviews? Are users not voting on helpfulness?

	good_or_bad_review	num_reviews	total_num_votes	avg_helpfullness	votes_per_review
0	0.0	28516	451524	44.043587	15.834058
1	1.0	174989	814542	68.450023	4.654818

We can see that bad reviews get significantly more votes per review than good reviews (15.8 votes on a bad review, compared to 4.7 votes on good reviews). We can also see that bad reviews have an overall sub-par avg_helpfulness score. Below is a sample review. You can see how it can be confusing as they are using very poliarizing words, but yet received no helpful votes.

|1.0 Star | Not Helpful |The most hated videogame of all time and greatest betrayal of a fanbase in gaming history The greatest bait and switch betrayal of a fanbase in video gaming history and most hated game of all time.

User Analysis

Approach & Summary

- With such a large data set, we want to get a sense of who our "top" performing reviewers were, were there any that stuck out for leaving a lot of reviews, or any that consistently left helpful reviews (or not)

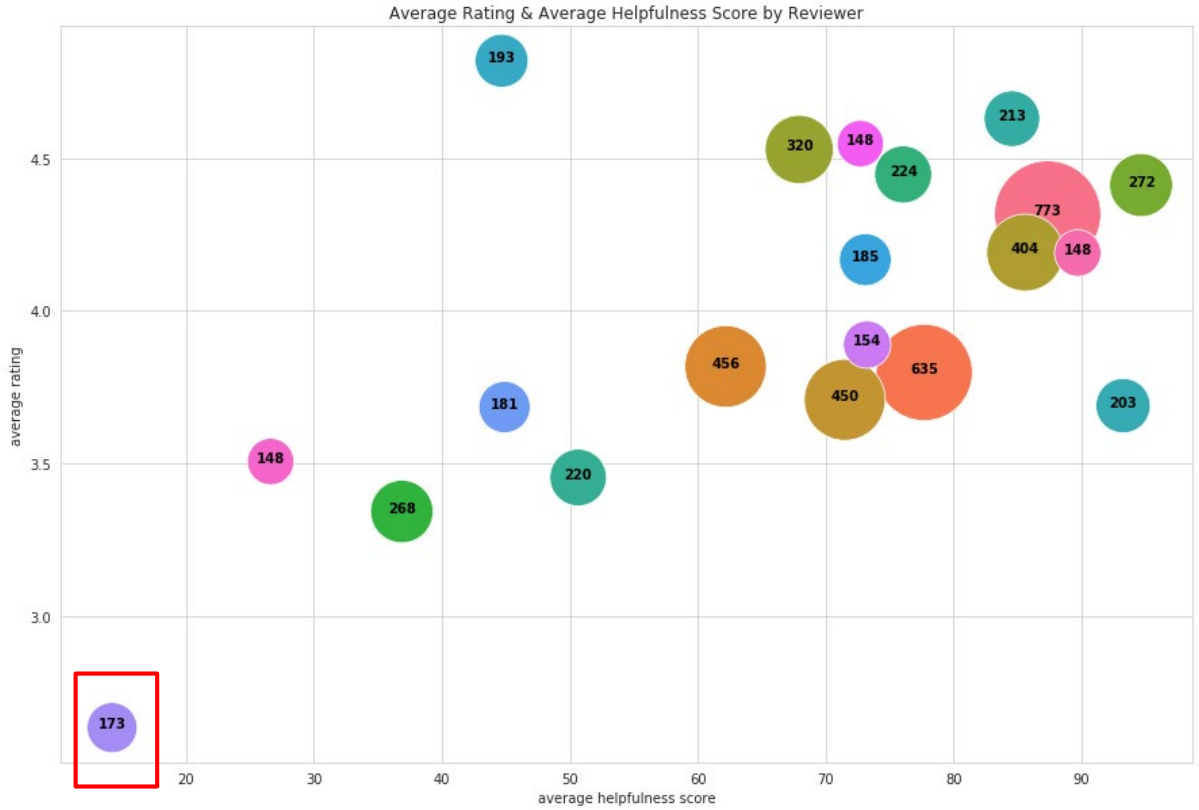
- We get our top twenty reviewers based on the total number of reviews submitted
    - For the most part, our consistent reviews are ones leaving high-star rating reviews
    - Most of them are also helpful majority of the time, with the top using being helpful 94%+ of the time
    - But then there is 1 that stands out for leaving 173 most-negative reviews and is only considered helpful 14% of the

	reviewer	average helpfulness score	average rating	total number of reviews
0	A3V6Z4RCDGRC44	87.373554	4.316947	773
1	AJKWF4W7QD4NS	77.716366	3.798425	635
2	A3W4D8XOGLWUN5	62.186783	3.817982	456
3	A2QHS1ZCIQOL7E	71.500577	3.708889	450
4	A29BQ6B90Y1R5F	85.610524	4.190594	404
5	AFV2584U13XP3	67.942264	4.528125	320
6	A2TCG2HV1VJP6V	94.675931	4.411765	272
7	A20DZX38KRBIT8	36.864941	3.343284	268
8	A1AISPOIIHTHXX	76.071715	4.446429	224
9	A2582KMXLK2P06	50.653141	3.454545	220
10	A3GKMQFL05Z79K	84.576069	4.629108	213
11	A74TA8X5YQ7NE	93.270532	3.689655	203
12	A8NHN9UPML858	44.667119	4.818653	193
13	AQMWZIH22R6LE	73.112936	4.167568	185
14	A3J8ABVGK7ZL6H	44.897563	3.685083	181
15	A21GTH20R33D6B	14.203548	2.635838	173
16	A2GBBDNZLYC4A9	73.245431	3.889610	154
17	A2IGEPJJYKMOWK	72.722546	4.547297	148
18	A3PASG15BRR40D	26.605045	3.506757	148
19	A36UKFV79879MD	89.724026	4.189189	148

For the most active users, overall rating and helpfulness are highly correlated

Insight: The top 20 users write more helpful reviews, and are slightly more critical of the games/ consoles they review

	rating	Perc_helpul
top 20 users	4	66.4%
Total	4.1	62.2%



Product Analysis

Approach & ASIN look up

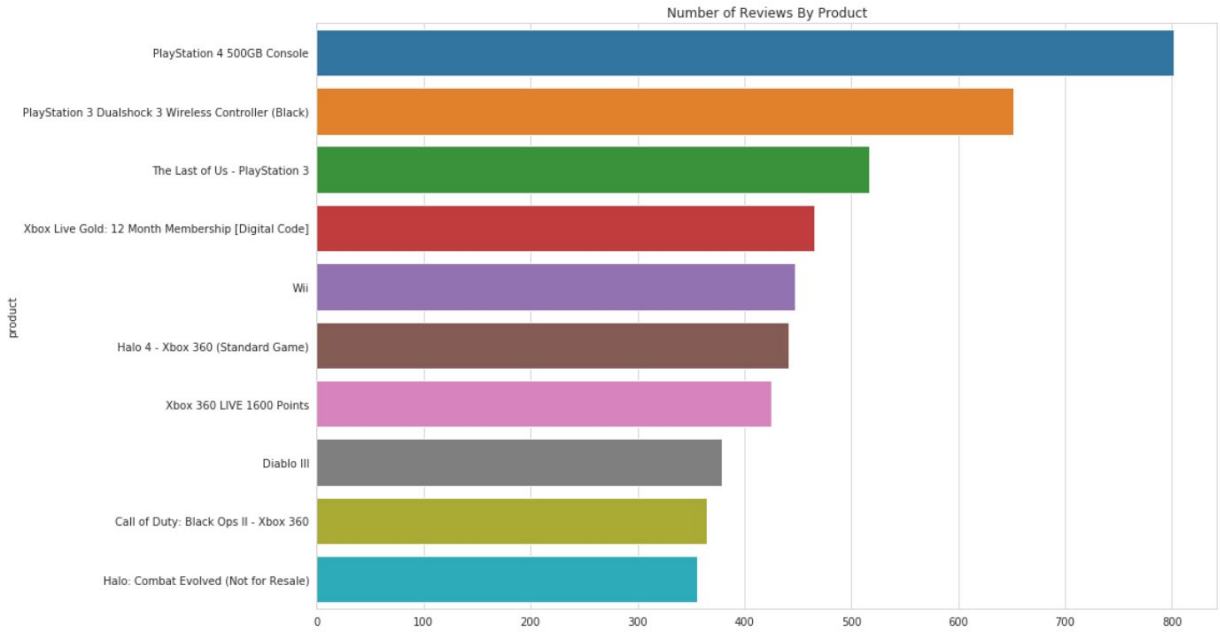
- We selected the top ten products with the most reviews
- Amazon has a helpful website for looking up ASIN, which we use to get the product name

asin	num_reviews	percent_of_total	product
B00BGA9WK2	802	0.346018	PlayStation 4 500GB Console
B0015AARJI	652	0.281301	PlayStation 3 Dualshock 3 Wireless Controller ...
B007CM0K86	517	0.223056	The Last of Us - PlayStation 3
B002VBWIP6	465	0.200621	Xbox Live Gold: 12 Month Membership [Digital C...
B0009VXBAQ	447	0.192855	Wii
B0050SYX8W	441	0.190267	Halo 4 - Xbox 360 (Standard Game)
B000B9RI14	425	0.183364	Xbox 360 LIVE 1600 Points
B00178630A	379	0.163517	Diablo III
B007XVTR3K	365	0.157477	Call of Duty: Black Ops II - Xbox 360
B00005NZ1G	356	0.153594	Halo: Combat Evolved (Not for Resale)

Top 10 Products

- The majority of the top items are each consoles;
  - This makes sense as they have a broader appeal

- The top 10 products account for only 3% of total reviews

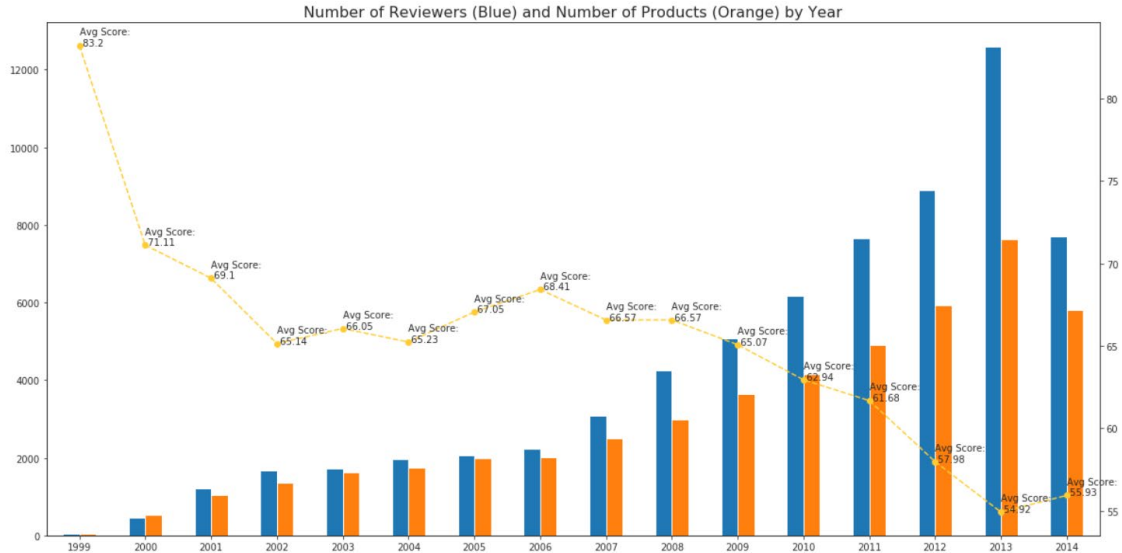


Time Series Analysis

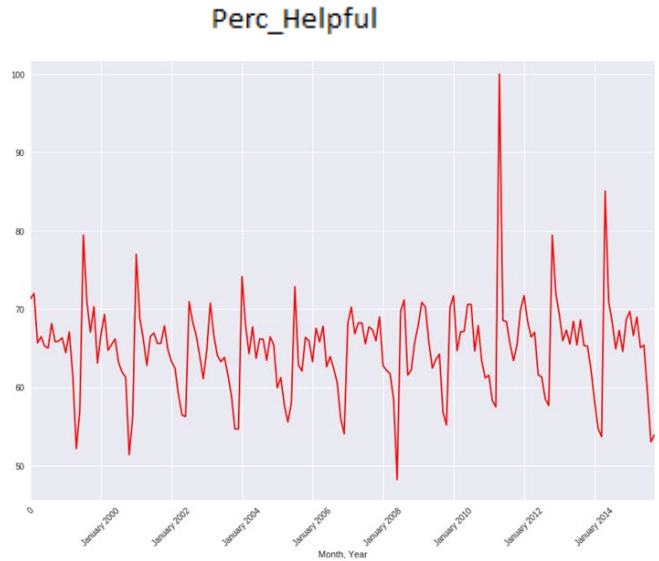
Insights by Year

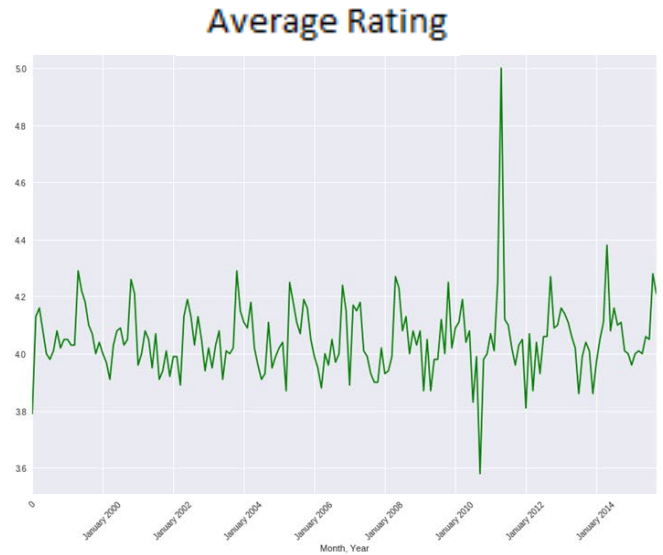
- As Amazon becomes more mainstream, the number of unique reviewers and products increase at similar rates

- At the same time, however, the ratings become much less helpful; falling from ~70% → 55%



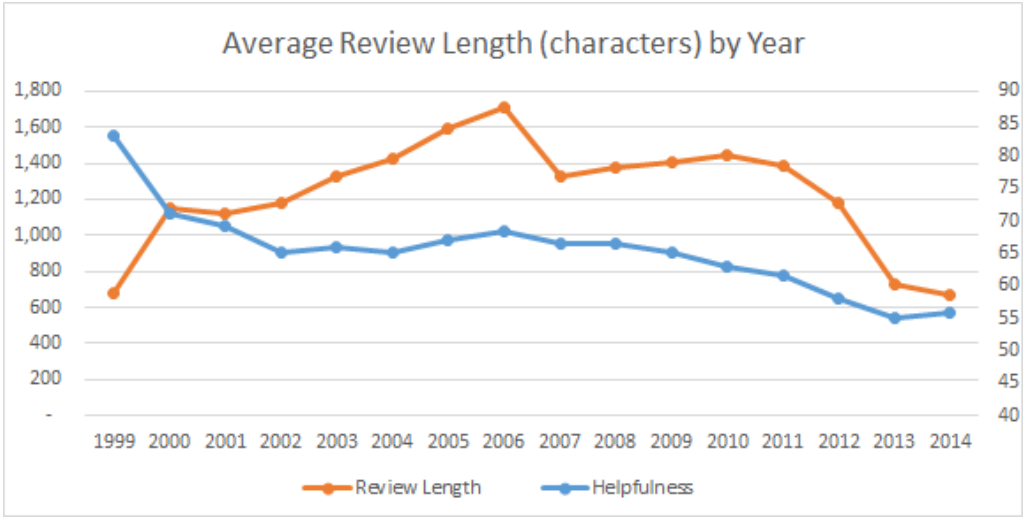
Insights by Month





- Perc_Helpful and Avg_Rating trends are similar to each other in volume
- Both are seasonal &and cyclical
    - There is a large spike in July - likely due to Amazon Prime Day

Review Length



One possible explanation for this is that the early adopters of Amazon (in the early 2000s) were the most passionate gamers. They left lengthy (and helpful) reviews. As Amazon has become more mainstream, the site has been flooded with less sophisticated users

Summary

Our Findings

Logistic Regression

Based on the words in a review, we can predict whether a review was **positive** with 90% accuracy and whether or not a review was **helpful** with 70% accuracy

Exploratory

- **Users:** the top users give more helpful reviews than the total, but are critical of the games/ consoles they review
- **Products:** the most common products are consoles since they have the broadest reach
- **Helpfulness** : helpfulness has declined as more users crowd the platform

Thank You!