The Hodinkee Online Community

A Closer Look At Some Numbers

April 30, 2020

Hamza Masood

Data Science - Flatiron School

Hodinkee User Profile | Email | LinkedIn

Contents

- Why is this interesting?
- High level overview

- All comments over time
- Most prolific commenters
- User engagement over time
- Future explorations

- Methodology details
- Assumptions
- Some random dead ends

Why is this interesting?

To Hodinkee

- Better user engagement → more time spent on hodinkee.com → Higher conversion rate for the Hodinkee Shop
- Inform new user engagement feature design
- Build a better community than competitors
- Better understand the overlap between the online and in-person Hodinkee communities

To Me

- I love watches
- I'm learning new data science techniques and needed a test subject
- I needed to build a dataset for the capstone project for my bootcamp (more details here). Once finished in May, I would hope this can be considered for deployment on hodinkee.com

Some high level numbers

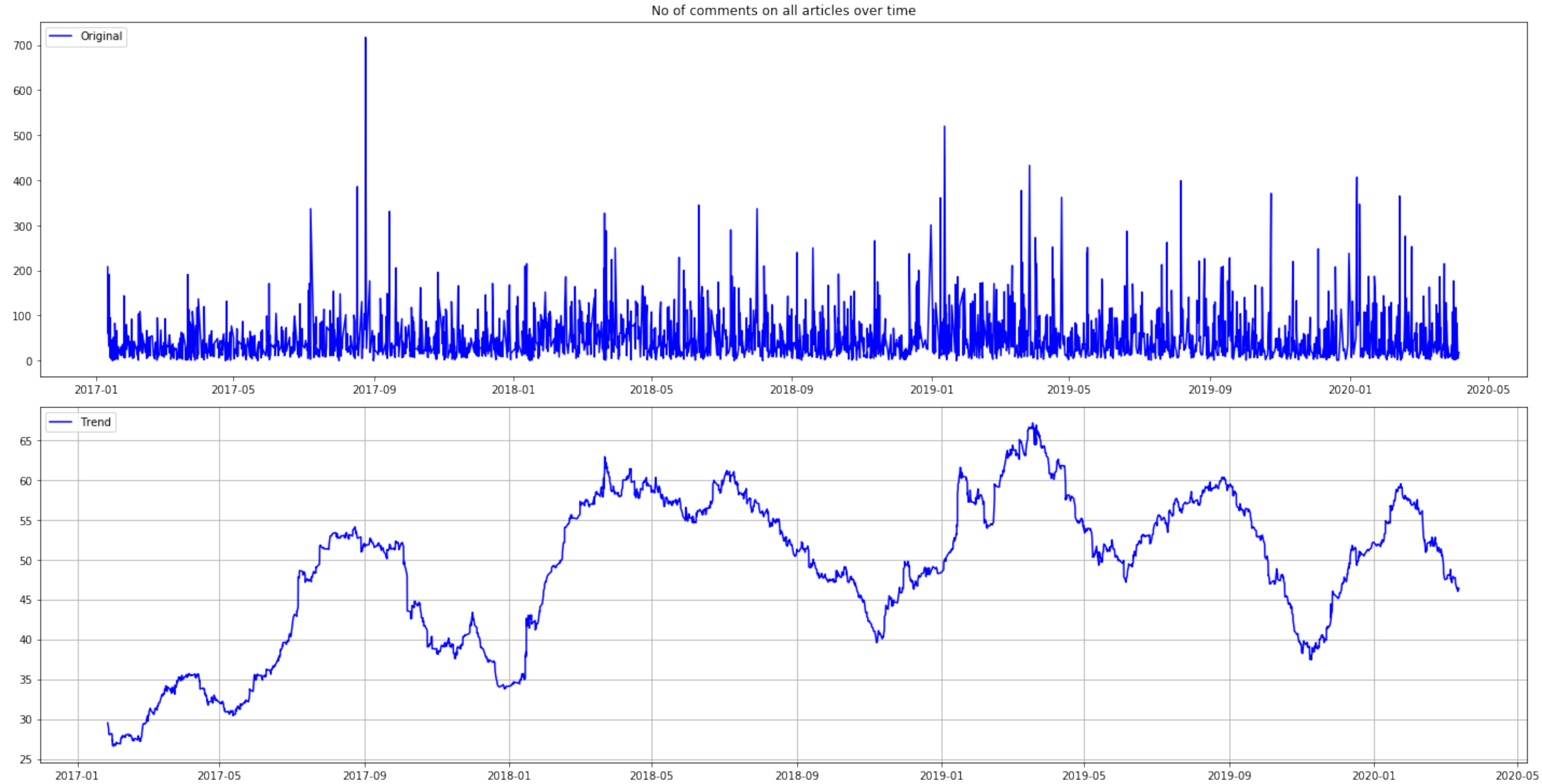
	Articles	Comments
Disqus Comments System	3,636	
New Comments System	2,865	137,495

Article Categories (not specified counted as one category)	66
Median article word count	457
90th percentile article word count	1,381

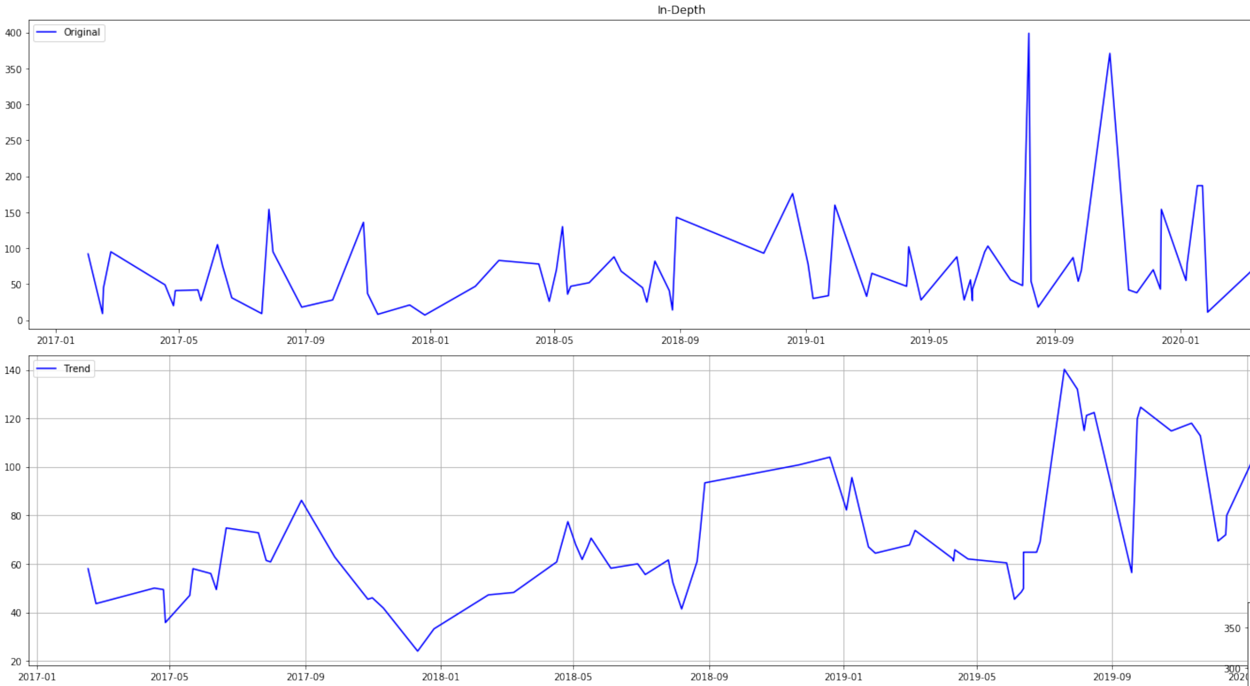
Comments posted in reply to an existing comment	36,183
Comments by Hodinkee staff (identified using profile flag…sorry James)	3,129
Duplicate comments posted by user in error	315
Most commented article: “Friday Live Episode 16: What Three Watches Would You Buy With $15,000?”	717
Median number of comments per article	32
Articles with zero comments under the new system	10

(Website snapshot taken on Apr 5)

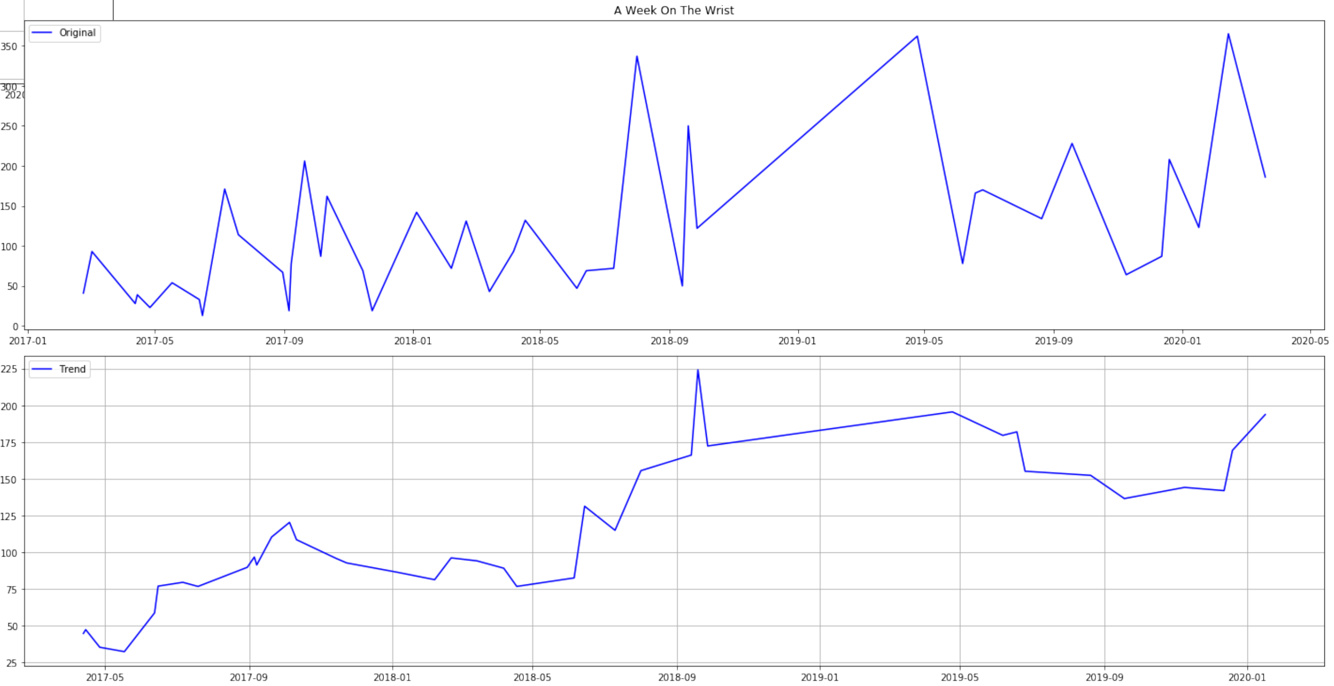
All comments over time: seasonal, cyclical trend



Categories with growing engagement



In-Depth



A Week On The Wrist

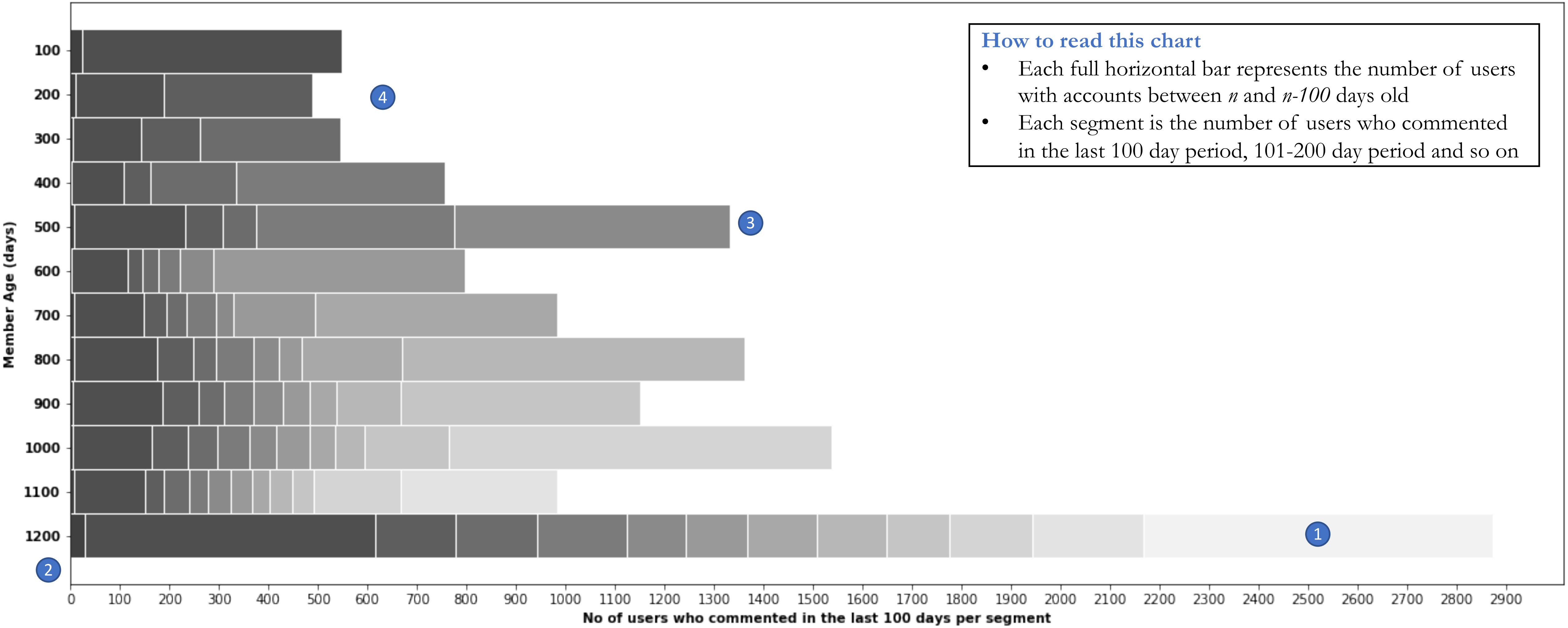
Most prolific commenters (normalized by account age)

Rank	Username	Prolificness Score	Comment Count	Account Age (days)	No. of Watches in Profile
1	Shatners	1.38	721	521	0
2	ICH	1.38	1392	1009	1
3	Gav	1.35	499	370	5
4	GreatScot	1.34	127	95	0
5	JackForster	1.11	1324	1190	0
6	Bside	0.98	1171	1190	0
7	Boman	0.97	625	644	5
8	wkf	0.90	1072	1190	12
9	PaulMiller	0.85	1008	1190	0
10	ripwatch	0.81	720	886	0
11	Oliver_H	0.77	27	35	5
12	Yev	0.77	703	917	5
13	CynicalBastard	0.76	626	825	0
14	TheOmegaMan	0.75	853	1131	10
15	Putito	0.73	136	187	1
16	AJ117	0.72	244	340	0
17	ThicknessMatters	0.71	762	1070	0
	…	…	…	…	…
20	Cole	0.63	271	429	9
	…	…	…	…	…
82	BenClymer	0.27	318	1190	3
	…	…	…	…	…
124	CaraBarrett	0.22	264	1190	2
	…	…	…	…	…
133	ghariyaan	0.21	131	613	7

Prolificness is calculated as $\frac{Comment\ Count}{Account\ Age\ in\ days}$

Included some Hodinkee Staff and myself for reference (I’m surprised and disappointed my score isn’t higher ☹)

User age and engagement pyramid



How to read this chart

- Each full horizontal bar represents the number of users with accounts between *n* and *n-100* days old
- Each segment is the number of users who commented in the last 100 day period, 101-200 day period and so on

1. 700 of the oldest users on the site haven't commented in 1,200 days
2. There is a continuously engaged cohort of users across all account ages
3. Users picked up around December 2018 continue to be more active
4. Fewer new users have signed up in the last year than in the previous two years

Key Takeaways: Most users who signed up didn't remain engaged and fewer users have been signing up over time

Future explorations

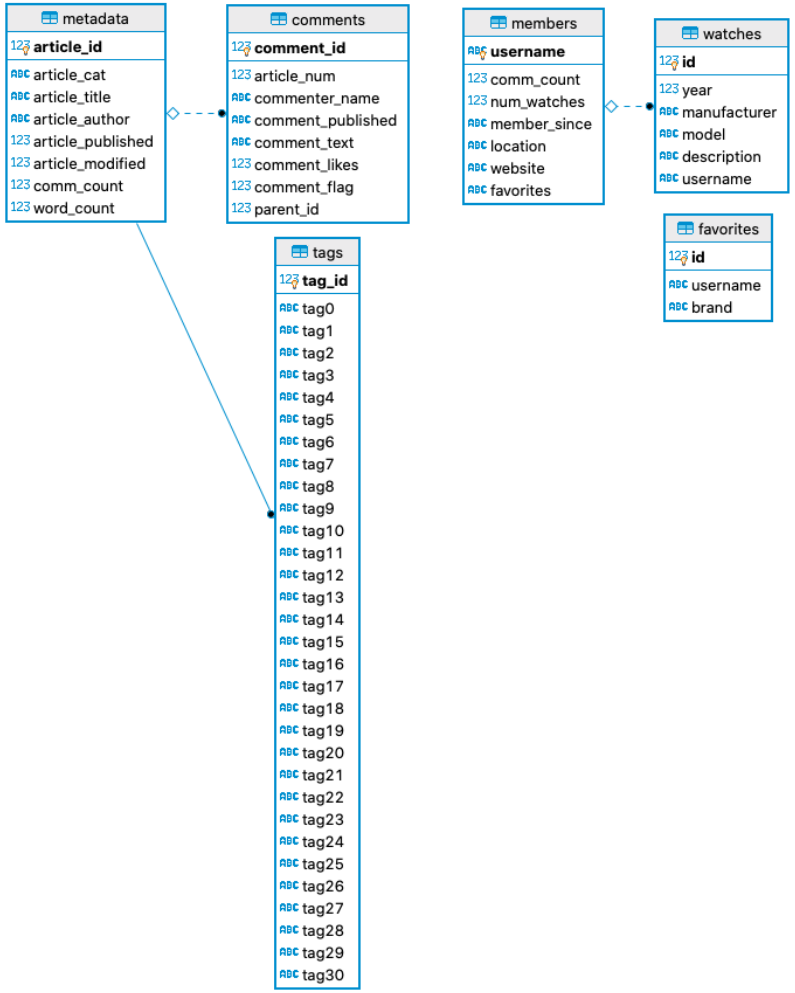
Comments

- Proportion of comments posted in reply to other comments
- Longest conversations (most comments in reply to the same top-level parent comment)
- Most liked comments
- Likelihood of comments being left on older articles
- Comments distribution by brand

Users

- Watch collection analysis (number of watches, types, etc)
- Favorite brands
- User location distribution
- Online profile

Methodology Details



metadata
- 123 **article_id**
- ABC article_cat
- ABC article_title
- ABC article_author
- 123 article_published
- 123 article_modified
- 123 comm_count
- 123 word_count

comments
- 123 **comment_id**
- 123 article_num
- ABC commenter_name
- ABC comment_published
- ABC comment_text
- 123 comment_likes
- 123 comment_flag
- 123 parent_id

members
- ABC **username**
- 123 comm_count
- 123 num_watches
- ABC member_since
- ABC location
- ABC website
- ABC favorites

watches
- 123 **id**
- 123 year
- ABC manufacturer
- ABC model
- ABC description
- ABC username

favorites
- 123 **id**
- ABC username
- ABC brand

tags
- 123 **tag_id**
- ABC tag0
- ABC tag1
- ABC tag2
- ABC tag3
- ABC tag4
- ABC tag5
- ABC tag6
- ABC tag7
- ABC tag8
- ABC tag9
- ABC tag10
- ABC tag11
- ABC tag12
- ABC tag13
- ABC tag14
- ABC tag15
- ABC tag16
- ABC tag17
- ABC tag18
- ABC tag19
- ABC tag20
- ABC tag21
- ABC tag22
- ABC tag23
- ABC tag24
- ABC tag25
- ABC tag26
- ABC tag27
- ABC tag28
- ABC tag29
- ABC tag30

SQLite database schema diagram

Python libraries used:

- pandas
- numpy
- beautifulsoup4
- sqlite
- matplotlib
- statsmodels
- datetime
- colour

metadata	6,501 rows
tags	6,501 rows
comments	137,495 rows
members	13,355 rows
favorites	17090 rows
watches	13490 rows

Total scraped data >1.75GB

Code is privately stored on GitHub, please let me know if you'd like access

Assumptions

- That this was as interesting for you as it was for me ☺
- My expertise is in data analysis and I looked for things that seemed interesting to me. I'm new to the world of horology and media so I'm not sure if the Hodinkee staff would have asked different questions

- Disqus comments not analyzed
- Articles only, no special coverage included; articles with more than one author in the byline default to taking the first one as the only author
- For comments in reply to a deleted comment, the parent comment is assumed to be the next comment up in the tree, or no parent if the deleted comment was itself a parent comment.

- Website snapshot was taken at end of day, Saturday April 4, 2020; all new content since then not included in the analysis
- User profile snapshots taken on Friday April 17, 2020

Random dead ends

- Time to First Comment after article is published: ‘datePublished’ in the HTML for the page is not the time the article is actually visible to users

- Member locations: difficult to analyze since this field is free-form

- Tags-based analysis: highly skewed assignment on articles, mixed tag types (article categories, events, watch types, watch brands, etc)