Grammarly Exercise
Charles Wang

Tools I Used

I am most comfortable using Python, so that was what I used. I could also have imported the data into a SQL-style relational database, as I am conversant in SQL as well. However, the local PostgreSQL instance I have doesn't lend itself easily to the kinds of data manipulations and visualizations I wanted.

The Python packages I wound up relying on the most included Pandas and matplotlib.

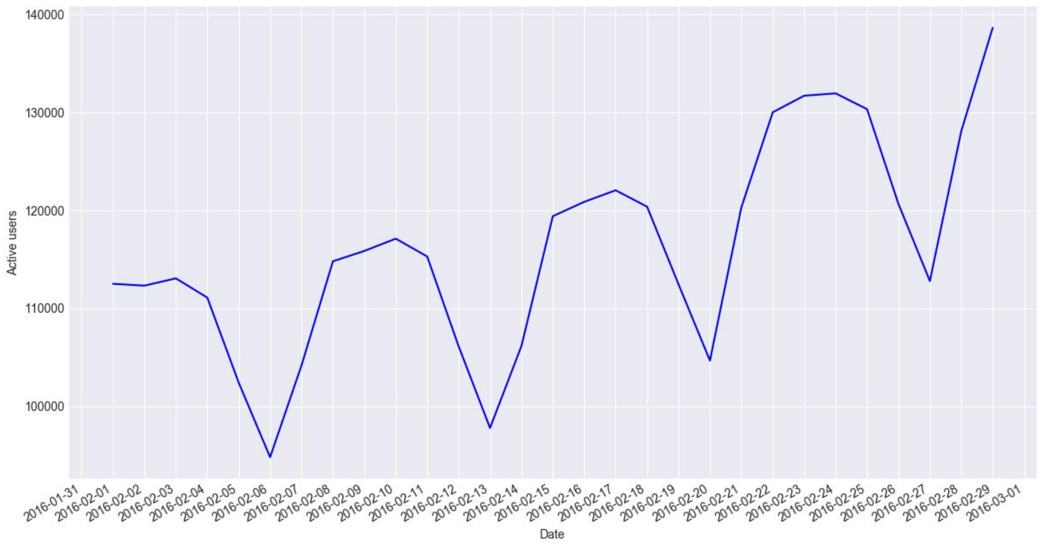
Reformatting and Cleaning the Data

JSON doesn't lend itself readily to easy manipulation or human-friendly reading, so I decided to convert the file, with its ~4.5 million records, into a CSV file with six columns: "date," "timestamp," "utmSource," "utmSource_cleaned," "uid," and "isFirst.".

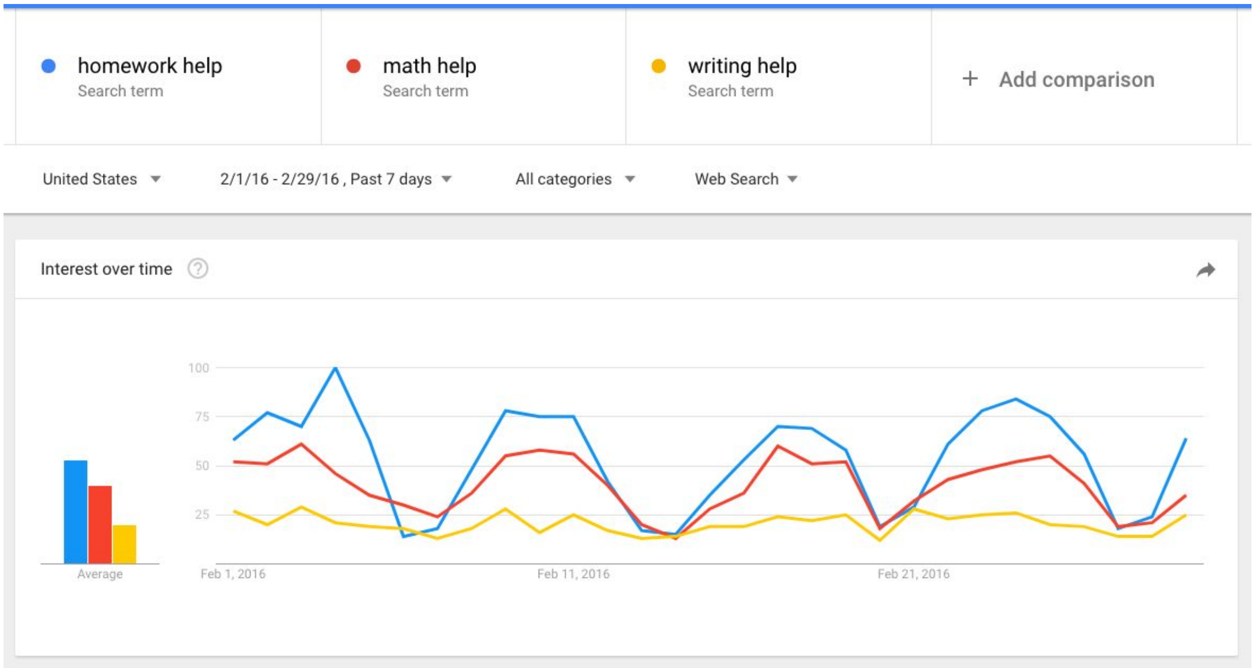
I cleaned the field "utmSource" by combining records that were obvious duplicates, such as "bing" and "Bing." These cleaned records were placed under the field name "utmSource_cleaned" in order to preserve the original values of the field "utmSource" in case they were subsequently needed. I also categorized all records where "utmSource" was "None" into its own category, under the label of "other." In all other cases, I assumed that distinct names were, in fact, distinct.

Daily Active Users

(See Table 1)



There is a clear weekly pattern to daily use, as measured by pings by unique users. Usage peaks on every Wednesday (February 3, 10, 17, 24), though it is generally high from Monday to Thursday. Out of curiosity, I did a similar search on Google Trends for the same period, and found very similar results for a selection of phrases such as “homework help,” “math help,” and “writing help.”



● homework help Search term

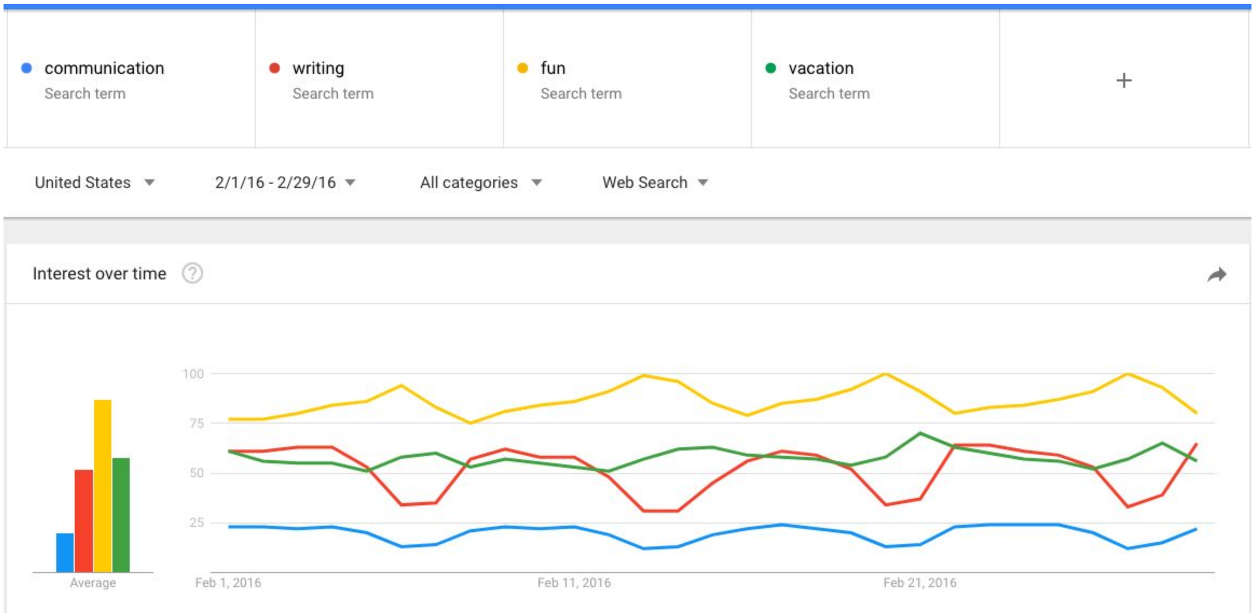
● math help Search term

● writing help Search term

United States ▾ 2/1/16 - 2/29/16 , Past 7 days ▾ All categories ▾ Web Search ▾

(https://trends.google.com/trends/explore?date=2016-02-01%202016-02-29&geo=US&q=homework%20help,math%20help,writing%20help)

A first impression suggests that Grammarly is used in an academic capacity, though both professional and academic activity might peak midweek. I decided to check this against several other queries:



● communication Search term

● writing Search term

● fun Search term

● vacation Search term

United States ▾ 2/1/16 - 2/29/16 ▾ All categories ▾ Web Search ▾

(https://trends.google.com/trends/explore?date=2016-02-01%202016-02-29&geo=US&q=communication,writing,fun,vacation)

It turns out that both professional and academic queries do tend to peak between Tuesday and Wednesday, while recreational queries tend to peak between Saturday and Sunday. The immediate takeaway is that Grammarly is most heavily used in the context of academic and professional writing and that such usage peaks midweek. People find Grammarly useful, and in February of 2016, the number of such people increased by the week, to the tune of 4- to 5-thousand. It seems that users do not, in general, use Grammarly for fun.

In order to more fully determine whether Grammarly is driven by academic or professional writing, it would be necessary to see trends by time of year. High academic usage would be represented by a lull in the number of daily active users in the summer months, specifically July and August.

Since we roughly know which days are high-activity and low-activity, possible implications for different methods of outreach and marketing might be as follows:

Blogging, email, and other published content

If the goal is to put a piece of writing in front of as many eyeballs as possible, then the content should be published from Monday to Thursday. Ideally, there would be an easy way for users to move from the Grammarly blog, Facebook page, or other web presence to the Grammarly app or web app; such a feature does not yet exist.

At the same time, publishing during times of peak traffic risks diluting the quality of engagement. It might, for instance, be advisable to publish on Fridays if it turns out that the chances of reaching and engaging a smaller number of higher-quality users are better. If Grammarly is anything like other products, then it is likely that a Pareto rule of some sort applies - that a relative minority of highly prolific users drive a plurality or majority of activity on the app or site.

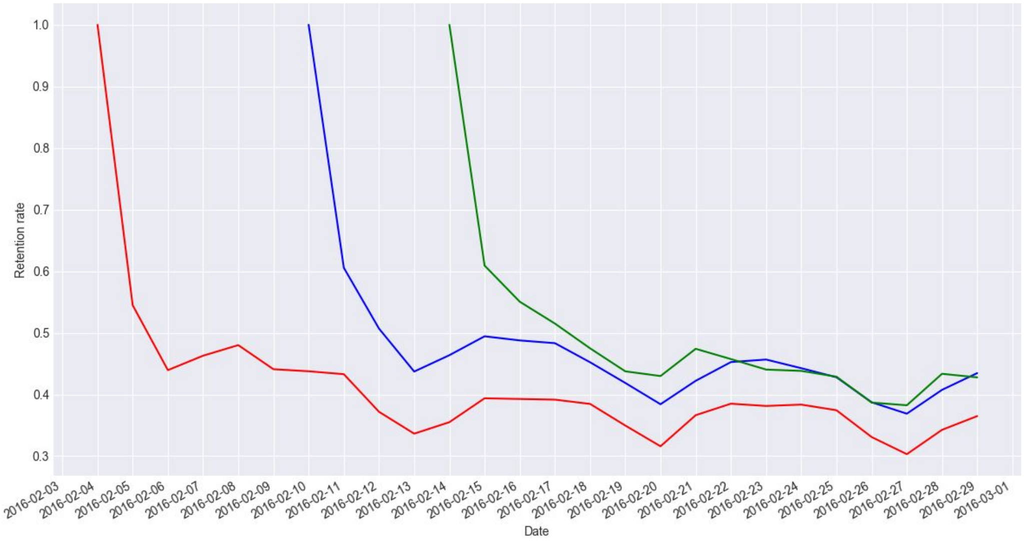
From what I have seen of Grammarly’s blog posts, most of the publishing is indeed from Monday to Thursday, with the occasional weekend piece. On a longer-term note, it would be of great personal interest to me to analyze Grammarly’s blog, email marketing, and social media data.

Updates

Updates and new features should be released on days where they will cause the least personal disruption, yet have enough activity to be evaluated for acceptance. Since the highest usage often lasts from Monday to Thursday, Friday is probably a reasonable time of the week.

Retention Rates

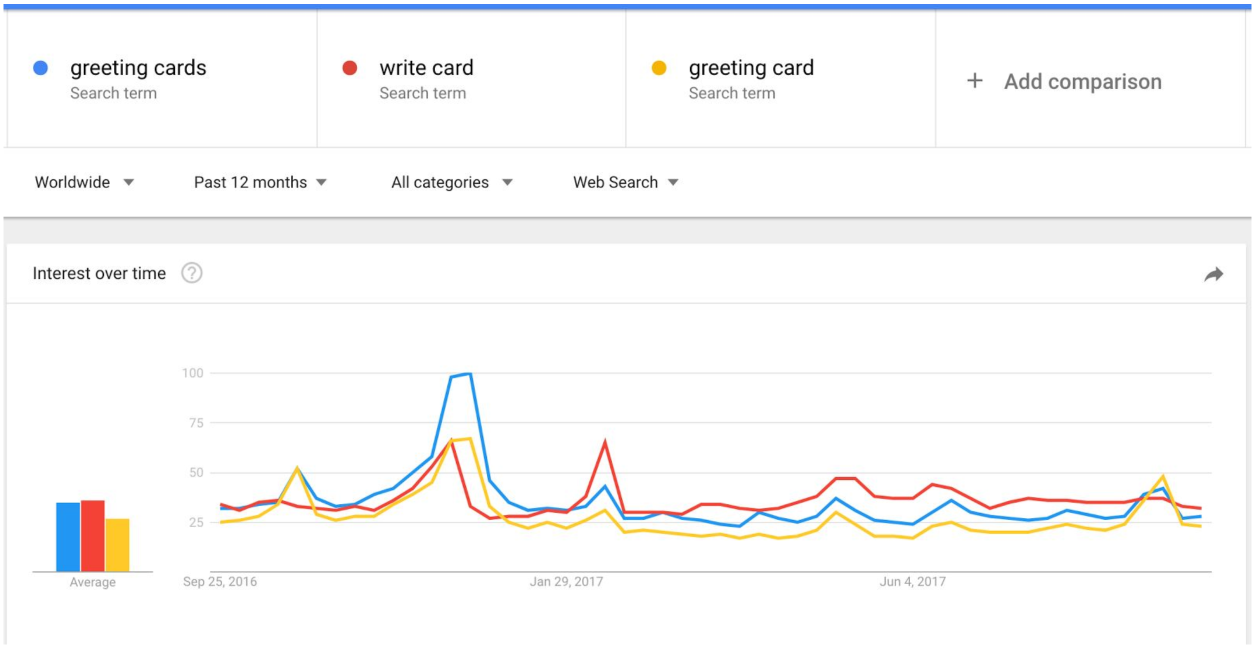
(See Table 2, Table 3)



At 16 days from first use, 42.8% of users who started on February 10 returned, as did 42.8% of users who started on February 14. A mere 34.9% of users who started on February 4 returned to the app. As evidenced by the chart and Table 3, retention for any *n* number of days from the first day was significantly and consistently worse for the cohort starting on February 4 than on the other two dates. The difference between the February 10 and February 14 cohorts, by contrast, is extremely slight, with a very slightly higher retention for the February 10 cohort.

One perhaps speculative explanation is that users who first used the product in the days up to and including Valentine's Day (February 14) are more likely to have used the product in a memorable, loyalty-inspiring way than users who did not.

Indeed, Valentine's Day is one of several holidays for which greeting cards (and gifts) are customary. The others, of course, are Christmas, Easter, Father's Day, and Mother's Day. Here again Google Trends helps as a sanity check. Note that the local peaks for the search terms "greeting card(s)" and "write card" coincide almost exactly with the holidays listed:



● greeting cards Search term

● write card Search term

● greeting card Search term

Worldwide ▾ Past 12 months ▾ All categories ▾ Web Search ▾

(https://trends.google.com/trends/explore?q=greeting%20cards,write%20card,greeting%20card)

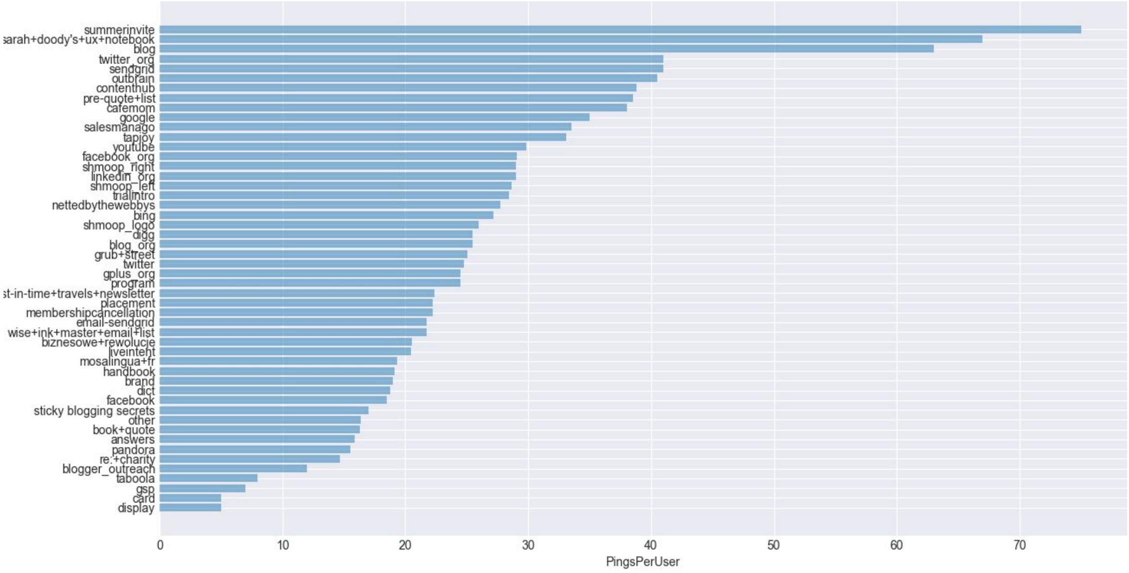
The implication for Grammarly's marketing is that first-time users (and returning users, too) should be targeted before events with relatively high stakes and emotional salience for people's personal lives. Widely recognized holidays where greeting cards are customary are one example. If possible, personal events such as birthdays, anniversaries, weddings, baby showers, and other events should also be targeted.

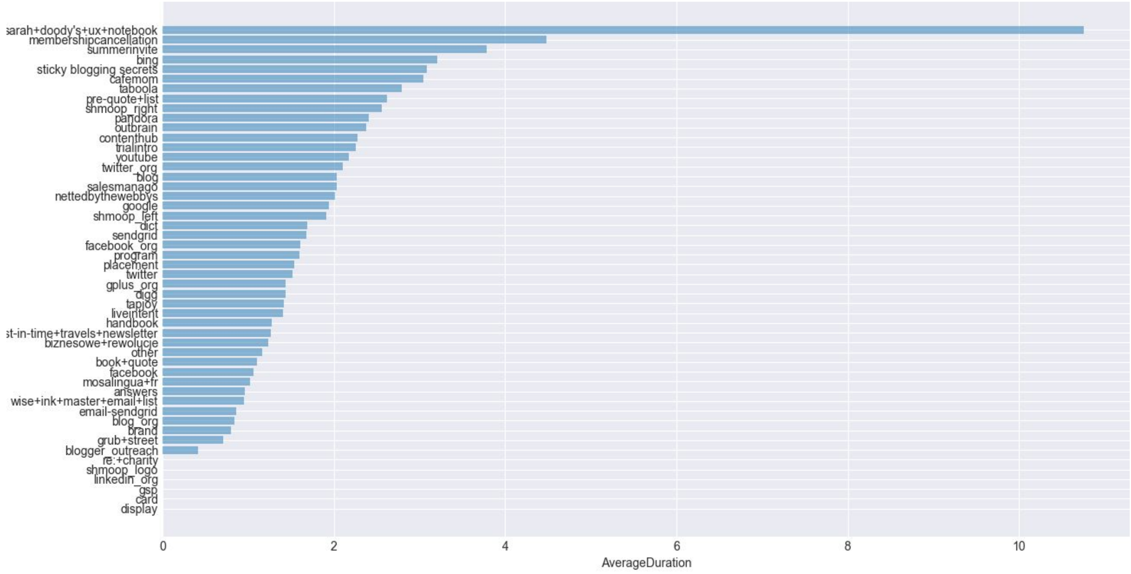
From the perspective of someone writing a highly personal message, it is likely far preferable to consult an impartial machine than a friend or family member (or nosy colleague).

These findings also suggest that Grammarly's reference database should be stocked not only with academic papers in order to detect plagiarism, but also a corpus of literature in order to suggest poetic expressions of affection and other kinds of personal regard.

Different Sources of Traffic

(See Table 4)





I decided to try several approaches to determine the quality of users from each source. I eventually settled on two. The first was the number of pings per user. The second was the average duration of a user’s session in hours, based on their first and final pings each day. I also used the number of unique users to inform my analysis.

There are of course conceivable disadvantages to either approach. A multitude of pings could very well represent multiple, frustrated attempts at using the app without success. A long duration between first and final pings per day could easily indicate long periods of idleness between uses of the app. The source “membershipcancellation” is likely one such example.

I decided to assume that, for the most part, frustrated users will simply disengage rather than attempt repeated interactions with the app. There are, after all, many sources of distraction on the internet, as well as several other services for checking spelling and grammar.

There were a number of sources for whom only one, two, or a handful of users materialized. Excluding those, the most successful source of traffic was the blog, which produced 63 pings per user and about 2 hours’ duration of activity. Following were ContentHub, Google, SALESmanago. Tapjoy, and Youtube. Each also produced between 30-39 pings and also about 2 hours each.

By far, the highest-quality traffic comes from the blog. It may be that the blog readers are both more likely to read as well as to write, or at least seriously consult writing advice. However, only 30 users were brought to the app from the blog, so, in the immediate term, relying on blog readers alone is probably not a good marketing strategy. The other sources mentioned - ContentHub, Google, etc - represent a good tradeoff between quality and raw numbers.

Longer-term, the question is whether it’s possible to drive more engagement using the blog. This could take the form of more frequent (and maybe shorter) updates, weekly newsletters/roundups, or any number of other approaches that can be tested empirically. Curiosity and literacy are, after all, social goods that should be encouraged.

General Thoughts on Grammarly’s Product

I used Grammarly’s product to check this document. I have also installed it as a Chrome extension, so every email and Facebook message of mine is now checked by the app (I would love to see Google Drive compatibility, though!). As a “dogfooding” matter I am definitely becoming a keen user of the app, and that is unlikely to change regardless of the outcome of my application. Grammarly is definitely many cuts above the usual services for checking spelling and grammar, and over the course of this exercise, it definitely helped belay some of my less comprehensible grammatical tics.

For personal amusement, I tested Grammarly out with non-Anglo-American phrases such as “Let’s prepone the meeting” and “Please do the needful.” Grammarly did not consider those phrases invalid, though I have no idea whether this is by design or not. As a somewhat tangential note, I do wonder how and whether Grammarly will evolve as usage of English changes through the adoption of neologisms, especially ones of foreign origin.

There are many data analysis and machine learning projects I can think of that will become practical with additional data. I would love, for instance, to examine Grammarly's blogging and social media outreach data in detail. I am also eager to learn more about the higher-quality "power" users, as well as the proportion of users who are non-native English speakers.

Longer-term, I can think of several applications of machine learning and natural language processing. The length, lexical complexity, and overall diction of users' writings, for instance, can be used to infer attributes such as age, education level, and even personality. If we can predict the attributes that might lead someone to become a high-quality user (like the blog readers), then perhaps we can nudge people to write both more frequently and more legibly. Maybe in some unlikely future, we could smooth out the curve for daily active users so that people write a great deal on weekends, too.

Finally, I think that Grammarly has a great deal of potential as an English learning and literacy aid if products and services are specifically tailored in accordance with Common Core reading and writing guidelines, and with special consideration for non-native speakers. A simple and concrete application could be automatically evaluating whether someone is writing at "grade level" or not.

Appendix

Table 1

Daily Active Users

date	count
2016-02-01	112504
2016-02-02	112317
2016-02-03	113062
2016-02-04	111098
2016-02-05	102420
2016-02-06	94807
2016-02-07	104175
2016-02-08	114802
2016-02-09	115859
2016-02-10	117113
2016-02-11	115290

2016-02-12	106132
2016-02-13	97788
2016-02-14	106158
2016-02-15	119409
2016-02-16	120874
2016-02-17	122053
2016-02-18	120374
2016-02-19	112452
2016-02-20	104665
2016-02-21	120306
2016-02-22	130017
2016-02-23	131707
2016-02-24	131947
2016-02-25	130334
2016-02-26	120658
2016-02-27	112789
2016-02-28	128096
2016-02-29	138644

Table 2

Retention Rates

date	retention_rate_04	retention_rate_10	retention_rate_14
2016-02-01			
2016-02-02			
2016-02-03			
2016-02-04	1		
2016-02-05	0.5447897623		
2016-02-06	0.439488117		
2016-02-07	0.4628884826		
2016-02-08	0.4800731261		
2016-02-09	0.4409506399		
2016-02-10	0.4376599634	1	
2016-02-11	0.4329067642	0.6056248003	

2016-02-12	0.3718464351	0.5068712049	
2016-02-13	0.3363802559	0.4372003835	
2016-02-14	0.3550274223	0.4637264302	1
2016-02-15	0.3937842779	0.4944071588	0.6091995654
2016-02-16	0.3926873857	0.4876957494	0.550525172
2016-02-17	0.3915904936	0.4832214765	0.5150307859
2016-02-18	0.3846435101	0.4522211569	0.4748279609
2016-02-19	0.3495429616	0.4186641099	0.4375226367
2016-02-20	0.315904936	0.3841482902	0.4299166968
2016-02-21	0.3663619744	0.4221796101	0.4741035857
2016-02-22	0.3850091408	0.4525407478	0.4574429555
2016-02-23	0.3813528336	0.4566954299	0.4404201376
2016-02-24	0.3835466179	0.4426334292	0.438247012
2016-02-25	0.3744058501	0.4279322467	0.428830134
2016-02-26	0.3308957952	0.3873441994	0.3868163709
2016-02-27	0.3031078611	0.3688079259	0.3824701195
2016-02-28	0.3425959781	0.4074784276	0.433538573
2016-02-29	0.3648994516	0.4346436561	0.4277435712

Table 3
Retention rate, indexed by number of days

day	**retention_rate_04**	**retention_rate_10**	**retention_rate_14**
1	1	1	1
2	0.5447897623	0.6056248003	0.6091995654
3	0.439488117	0.5068712049	0.550525172
4	0.4628884826	0.4372003835	0.5150307859
5	0.4800731261	0.4637264302	0.4748279609
6	0.4409506399	0.4944071588	0.4375226367
7	0.4376599634	0.4876957494	0.4299166968
8	0.4329067642	0.4832214765	0.4741035857
9	0.3718464351	0.4522211569	0.4574429555
10	0.3363802559	0.4186641099	0.4404201376
11	0.3550274223	0.3841482902	0.438247012
12	0.3937842779	0.4221796101	0.428830134

13	0.3926873857	0.4525407478	0.3868163709
14	0.3915904936	0.4566954299	0.3824701195
15	0.3846435101	0.4426334292	0.433538573
16	0.3495429616	0.4279322467	0.4277435712
17	0.315904936	0.3873441994	
18	0.3663619744	0.3688079259	
19	0.3850091408	0.4074784276	
20	0.3813528336	0.4346436561	
21	0.3835466179		
22	0.3744058501		
23	0.3308957952		
24	0.3031078611		
25	0.3425959781		
26	0.3648994516		

Table 4

Different metrics by source

Source	UniqueUsers	Pings	PingsPerUser	AverageDuration
summerinvite	1	75	75	3.787822953
sarah+doody's+ux+notebook	1	67	67	10.75348822
blog	30	1890	63	2.033660784
twitter_org	1	41	41	2.104490067
sendgrid	1	41	41	1.67709869
outbrain	2	81	40.5	2.373781816
contenthub	182	7060	38.79120879	2.277462391
pre-quote+list	2	77	38.5	2.618756696
cafemom	1	38	38	3.041081331
google	28	980	35	1.939388635
salesmanago	1122	37633	33.54099822	2.029603054
tapjoy	347	11488	33.10662824	1.421504287
youtube	458	13666	29.83842795	2.17856972
facebook_org	66	1921	29.10606061	1.604891185
shmoop_right	1	29	29	2.556205224
linkedin_org	1	29	29	0

shmoop_left	6913	197982	28.63908578	1.911291038
trialintro	9	256	28.44444444	2.26055103
nettedbythewebbys	4	111	27.75	2.016836708
bing	5	136	27.2	3.202465323
shmoop_logo	1	26	26	0
digg	2	51	25.5	1.436373271
blog_org	15	382	25.46666667	0.8412096209
grub+street	2431	60868	25.03825586	0.7038662838
twitter	5151	127426	24.73810911	1.51452991
gplus_org	2	49	24.5	1.439062271
program	16493	403909	24.48972291	1.601317772
just-in-time+travels+newsletter	515	11531	22.39029126	1.259698429
placement	264	5865	22.21590909	1.54092952
membershipcancellation	5	111	22.2	4.477485897
email-sendgrid	4	87	21.75	0.8582538177
wise+ink+master+email+list	31	674	21.74193548	0.9510773518
biznesowe+rewolucje	86287	1774190	20.56149826	1.235139208
liveintent	89	1819	20.43820225	1.411063603
mosalingua+fr	6345	122887	19.36753349	1.023836858
handbook	1158	22119	19.10103627	1.276022343
brand	1	19	19	0.803612338
dict	342	6415	18.75730994	1.685910277
facebook	74	1367	18.47297297	1.059661347
sticky blogging secrets	1	17	17	3.082315778
other	107393	1754850	16.34045049	1.167201282
book+quote	3	49	16.33333333	1.102123611
answers	14979	237701	15.86894986	0.9586223264
pandora	4	62	15.5	2.408915682
re:+charity	3	44	14.66666667	0.000374870801
blogger_outreach	1	12	12	0.4196349537
taboola	1	8	8	2.795745397
gsp	1	7	7	0
card	1	5	5	0

display	1	5	5	0