Heavy User Analysis

Kris Thayer

March 4, 2019

Executive Summary

Based on this exploratory data analysis, I decided that heavy users can be broken down on three dimensions: time, intensity, and revenue impact. I propose that heavy users for a given dimension be defined as the 80th percentile or higher of a seven day average of that dimension.

Introduction

Motivation

Mozilla would like to understand the attributes and usage patterns of heavy users of the Firefox browser. This can help drive engineering and marketing decisions about browser features to promote retention and growth of heavy users.

Prior Work

Brendan Colloran from the Strategy and Insights team did this analysis of Heavy Users in 2016: Heavy users - Proposed definition of heavy users for intra-year work (2016-03). He wanted to use page views as the measure of heavy users, but this data was not available then. He originally proposed using session hours, with the following concerns: “Session hours are less than ideal for a number of reasons -- they include idle time (FF running in the background) and can even include time that the computer is asleep, and there is a known anomalies in the measurement that can cause more than 24 session hours to be counted for a single calendar date.” He settled on defining Heavy Users on a day as those users that are in the top active 10% of total active ticks for a 28 day window ending on that day. If a user had a total of 0 active ticks in the 28 days they were not considered active and were not used in the calculation of the top 10%. He chose 28 days because it aligns with Monthly Active Users (MAU), it is not so long as to introduce undue lag, and it is not so short as to introduce undue churn and bounciness.

The product data science team proposed in 2017 a similar definition of Heavy Users where a user is a Heavy User as of day N if, for the 28 day period ending on day N, the sum of their active ticks is in the 90th percentile (or above) of all clients during that period. The full analysis is in Operationalizing "Heavy Users" OKR.

In December of 2018, Rosanne Scholl of the Firefox Strategy & Insights team presented an analysis in Attitudes of Heavy Desktop Users combining survey and telemetry data to determine attitudes of Heavy Users. They used the following definition: Heavy users are identified by their hours of use, URI count, and search count, but the range of heaviness is retained- not artificially sorted into “heavy” and “not heavy”. The hours of use was determined by subsession hours which indicates how long a client computer was running Firefox. This measure is subject to measurement error due to ping reporting, among other issues. All 3 daily measurements were used from the date the user clicked through to the survey and from the 6 days prior to that day, then averaged. Averaging over a week helps to make the heaviness measures less sensitive to weekly patterns and general volatility. They averaged across the week among days which had values for the metric, which ignores the fact that many users have 0 or missing values on some of the days. A missing value does not necessarily mean there was 0 usage that day, so they used the average across days with values, rather than sum.

Methods

Terminology

Heavy users are characterized by usage patterns that go beyond a normal, average use of online media, websites, apps or other digital services. The idea is traced back to Dik Warren Twedt who coined the term “heavy-half” in 1964 to describe the market segment that accounted for a major proportion of a product’s sales. This is the 80/20 rule - 80% of the volume of a product is consumed by 20% of its consumers.

Heavy users will be considered based on time, intensity and revenue impact. For a time measurement we will consider active hours which indicate the time a user was actively using the browser. We will also consider subsession hours as a time measurement which includes time the browser is idle. The intensity measurement is generated from Uniform Resource Identifier (URI) counts. The revenue impact measurement is from search counts in the search_clients_daily table adding sap for "searches issued from Mozilla's UI", and organic for "searches made in content".

Active ticks are a count of 5 second increments during which the browser is actively being used. This is converted to active hours by multiplying by 5 to convert to seconds and dividing by 3600 seconds per hour.

An Active Daily User (aDAU) is defined as a client who has total daily URI >= 5 for a given date.

Monthly Active Users (MAU) is the number of unique clients who have been aDAU on any day in the last 28 days.

The sample id can be used to get data from 1% of clients to make the queries manageable.

This analysis uses:

- Individual ping data from the main_summary table
- Daily totals summed from individual pings by client_id
- Weekly averages of daily totals, averaged over days with data (not 7 days)

Procedure

My initial task was to determine a cutoff for “heavy users”. I started by exploring the data in the main_summary table. Since this is an enormous amount of data and the queries can take a long time, I did my exploration through Databricks notebooks. I chose a week of data in September as a baseline to study since it would avoid summer and major holidays. I used 1% of the client data by selecting sample_id 42.

Based on the previous analyses, I considered URI, subsession hours, active hours and search count as the metrics of interest in determining a cutoff. For URI, subsession hours and active hours I looked at individual pings, daily totals from adding the pings for a client and weekly averages of daily totals averaged over the days of the week with data. For search counts, the data came from the search_clients_daily table so there were no individual ping values. I joined the daily search counts with the daily totals from main_summary, and also averaged them over a week.

To determine if there were outliers or abnormal values that should be excluded from the cutoff consideration, I looked at the summary statistics for the 4 variables. Because of the max values and standard deviation for the time variables, I did an analysis of active hours compared to subsession hours to determine which of those two variables to use as my time indication. From this analysis I decided to use active hours (further discussed in the Discussion).

Next I considered the segments of data to use for determining the cutoffs which could minimize the effects of outliers and abnormal values. I looked at the distribution and quantiles for 4 different segments of data - all the records, records where the URI and active hours were both greater than 0, records where URI and active hours were both greater than 0 and URI and search counts were less than extreme values (described in Summary Stats), and aDAU users.

To determine if I could define one cutoff value for a heavy user I looked at the correlation between the 3 variables. Since the three variables were not very correlated, I used the quantile information for the 4 segments of data above to choose cutoffs for each of the 3 variables.

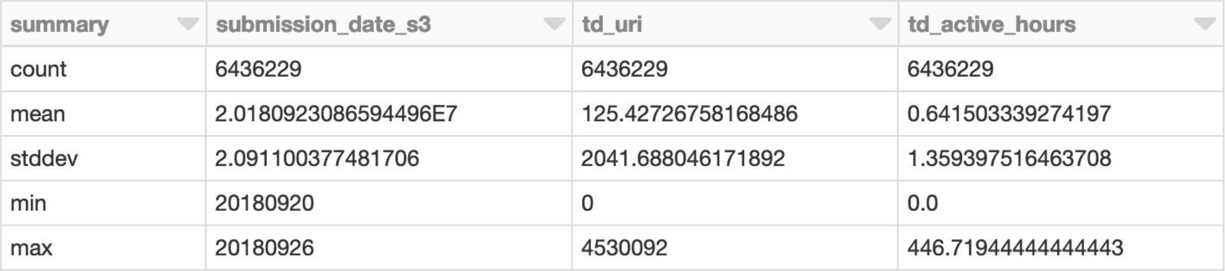
To validate these cutoff values for each type of heavy user (URI, search count and active hours), I looked at the percentages of clients who would be classified as heavy for a day and a week. I generated a histogram to see the distribution of heavy users. I checked the correlation of users who were heavy in any of the 3 categories. Then I checked the percentage of clients who were heavy in all 3 categories.

Having gained confidence in my cutoffs, I investigated heavy use over time. I looked at the fluctuation of the cutoff values over a week. I plotted the number of heavy users for each day of a week to see variation between weekday and weekend usage. I also plotted the number of days in a week that clients had heavy usage. Since these plots seemed contradictory, I plotted the day of the week that clients had heavy use for clients with only 1 day in the week of heavy use. Finally I looked at heavy user retention over 6 weeks.

I used one day of data to look at the attributes of heavy users. Specifically, I looked at the frequencies and percentages of heavy users for tab count, window count, active addons count, sessions started on this day, normalized channel, os, is default browser, country, locale, and default search engine.

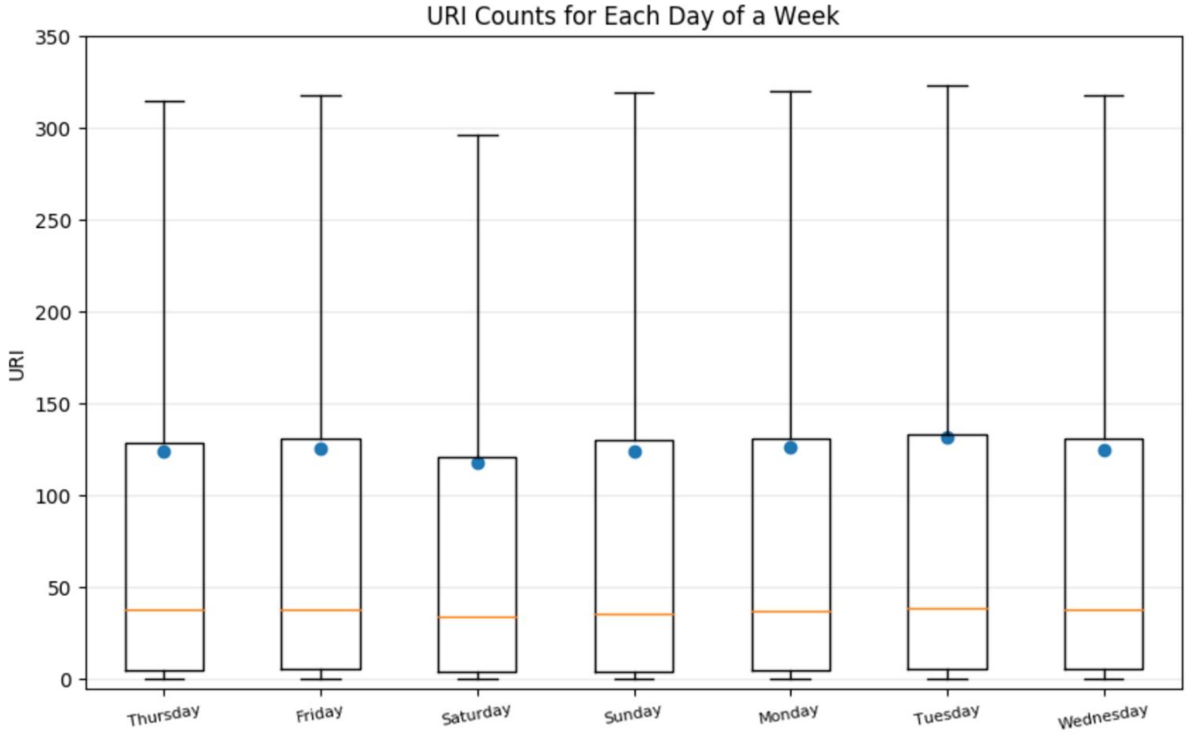
Summary stats

The summary statistics for all the daily client records for a week are shown below.



summary	submission_date_s3	td_uri	td_active_hours
count	6436229	6436229	6436229
mean	2.0180923086594496E7	125.42726758168486	0.641503339274197
stddev	2.091100377481706	2041.688046171892	1.359397516463708
min	20180920	0	0.0
max	20180926	4530092	446.71944444444443

The distribution of the URI counts over a week are shown below. The mean of the data is shown with the blue dot. Since there is some variation for each day of the week, for the rest of the analysis when I look at one day of data I've chosen Wednesday.



The segments of data used to determine the cutoff can compensate for daily variability along with reducing the effects of outliers and inactive days. I looked at the following 4 segments of data:

- all the records
- records where the URI count and active hours were both greater than 0 to try to capture active users
- records where URI and active hours were both greater than 0 with URI and search counts less than extreme values to try to capture “typical” active users
- records for aDAU users where total daily URI >= 5

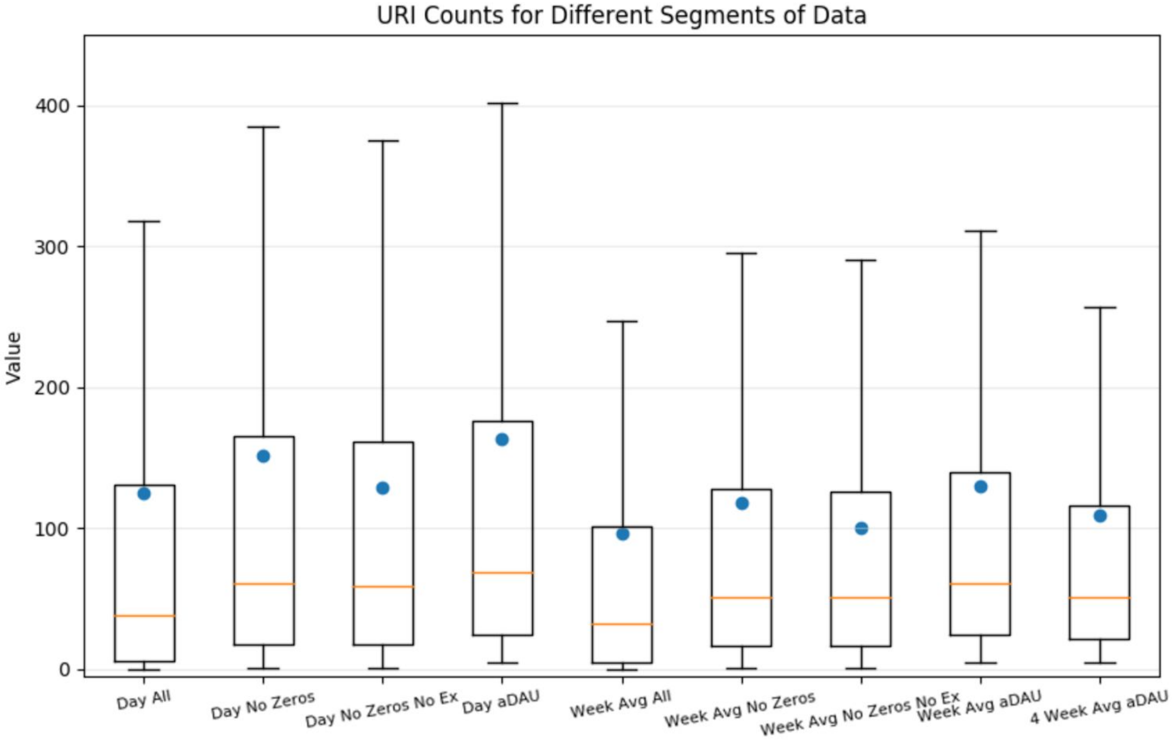
For one week, 8.7% of the records had both zero URI count and active hours representing 237,144 distinct clients.

For extreme values for URI and search count I selected numbers above the 99.5th percentile. I chose 1500 for URI count. For one week 0.67% of all records had URI count or search count at or above that threshold representing 27,411 clients.

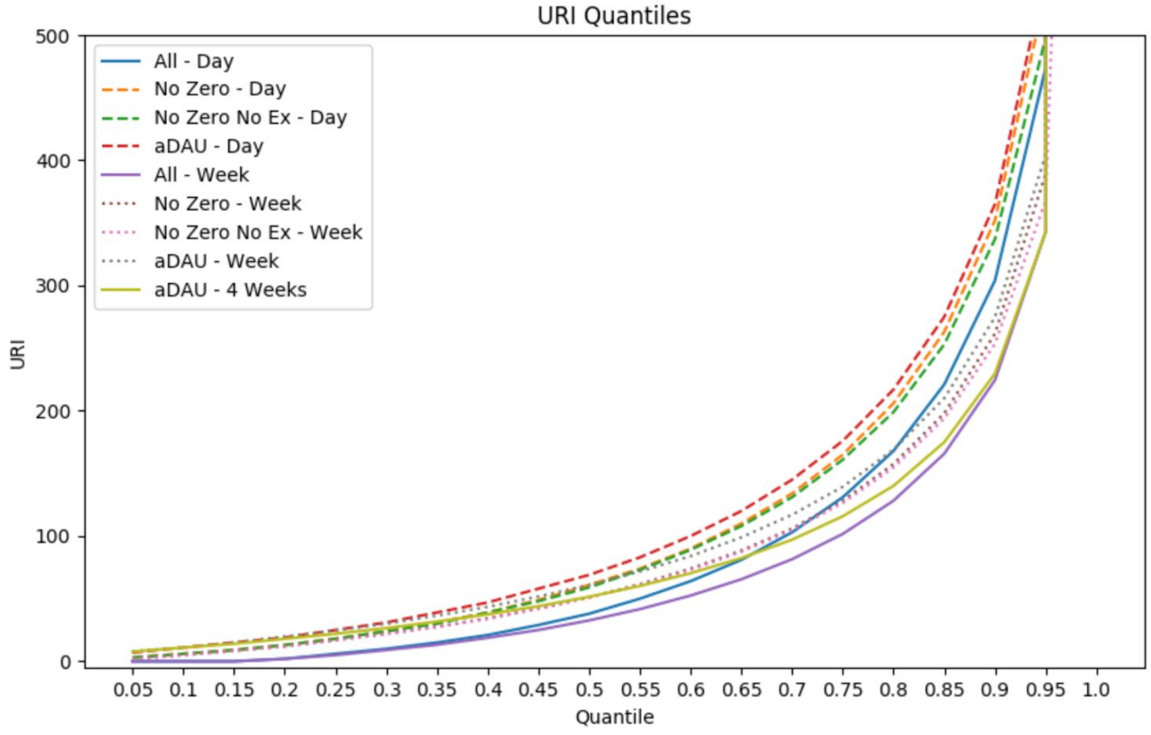
The aDAU records for the week are 75% of all the records for the week representing 1,432,289 clients.

I looked at these 4 segments of data for 1 day and averaged over a week, and the aDAU records averaged over 4 weeks.

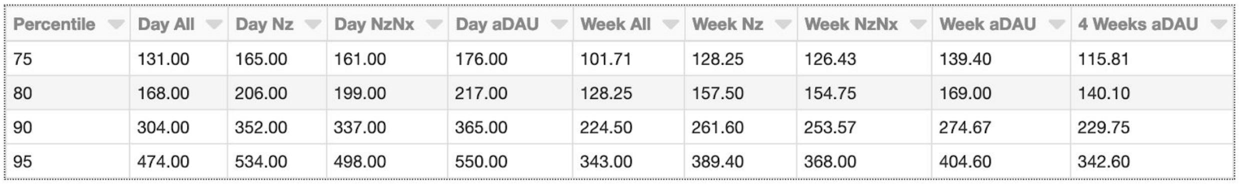
The distribution of the URI counts for different segments of data are shown below. The mean of the data is shown with the blue dot. Search counts and active hours followed the same patterns for the different segments although with a different range of values. As expected, averaging over a week lowers the variability and averaging over 4 weeks lowers the variability even more.



The URI quantiles for the different data segments are shown below. The URI values for the dashed line are higher because those segments eliminate the zero values for 1 day. The dotted lines are lower because they eliminate the zero values but are averaged over a week. The aDAU values averaged over 4 weeks start out higher but end slightly below all the values averaged over 1 week. All of the values increase sharply at the 95th percentile. Search counts and active hours followed the same patterns for the different segments although with a different range of values.

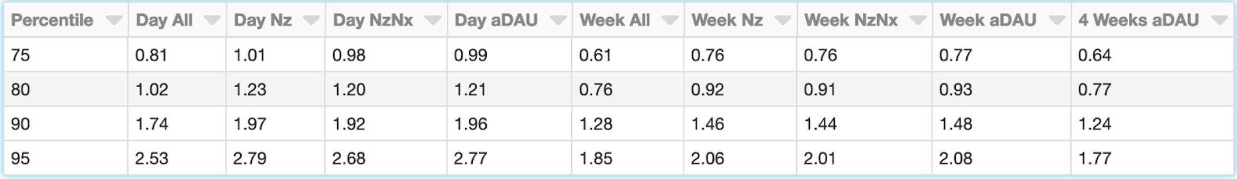


The URI quantile values for the 9 segments of data are shown below. This is the set of most likely URI heavy user cutoffs.



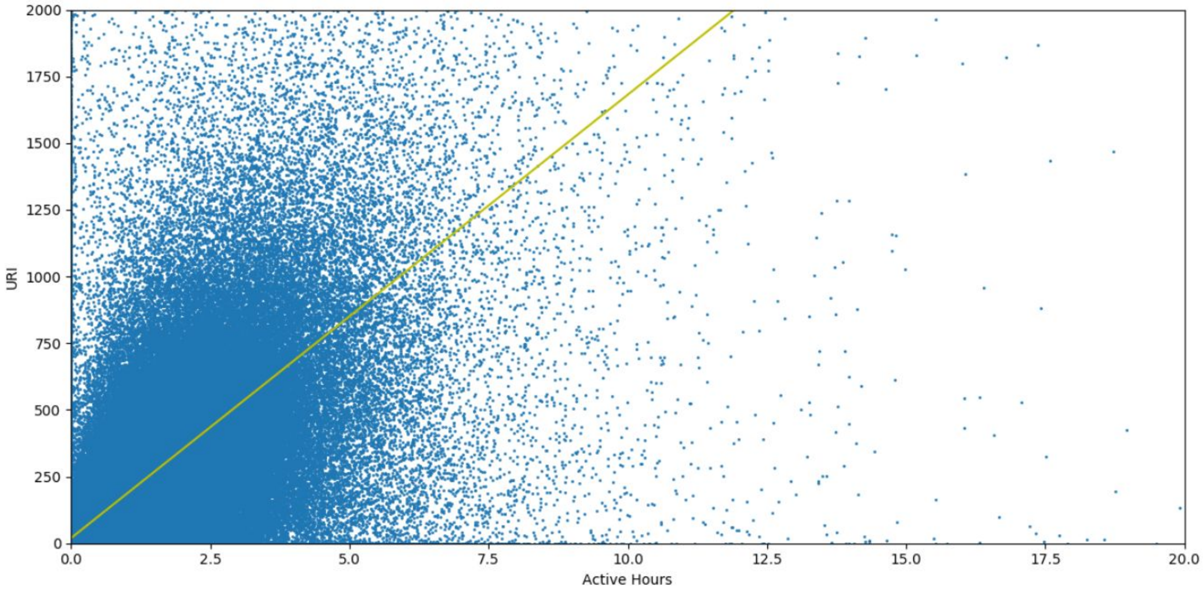
Percentile	Day All	Day Nz	Day NzNx	Day aDAU	Week All	Week Nz	Week NzNx	Week aDAU	4 Weeks aDAU
75	131.00	165.00	161.00	176.00	101.71	128.25	126.43	139.40	115.81
80	168.00	206.00	199.00	217.00	128.25	157.50	154.75	169.00	140.10
90	304.00	352.00	337.00	365.00	224.50	261.60	253.57	274.67	229.75
95	474.00	534.00	498.00	550.00	343.00	389.40	368.00	404.60	342.60

The active hours quantile values for the 9 segments of data are shown below. This is the set of most likely active hours heavy user cutoffs.



Percentile	Day All	Day Nz	Day NzNx	Day aDAU	Week All	Week Nz	Week NzNx	Week aDAU	4 Weeks aDAU
75	0.81	1.01	0.98	0.99	0.61	0.76	0.76	0.77	0.64
80	1.02	1.23	1.20	1.21	0.76	0.92	0.91	0.93	0.77
90	1.74	1.97	1.92	1.96	1.28	1.46	1.44	1.48	1.24
95	2.53	2.79	2.68	2.77	1.85	2.06	2.01	2.08	1.77

These scatter plots for all records from 1 day show that the 3 cutoff variables are not very correlated.



The correlation of search counts to URI is 0.12, the correlation of active hours to URI is 0.25, and the correlation of active hours to search counts is 0.35.

Based on the distribution and quantiles for the 3 variables, I chose the cutoffs from the 80th percentile of the weekly average aDAU. Using the weekly average will reduce volatility and daily fluctuations and using aDAU records will ensure that we are considering active use.

The cutoffs for this date and sample were compared to other samples from the same date, and the same sample at later dates to determine the variability of the cutoffs. The cutoffs were very similar across different samples for the same date, and increased slightly for later dates. The results are summarized below.

Date	Sample ID	URI Cutoff	Active Hours Cutoff
09-26-2018	42	169.00	0.93
09-26-2018	35	168.71	0.93
09-26-2018	78	169.00	0.93
10-24-2018	42	171.00	0.95
1-16-2019	42	172.33	0.97

Discussion

Usage Patterns

I found some unusual usage patterns when analyzing pings for 1 day which pointed out the need to select the data for determining heavy user cutoffs carefully.

There are 20% of pings in 1 day with 0 URI counts and 0 active hours. These are for 165,413 distinct clients. When the pings are totaled over a day, there are still 83,621 clients with 0 URI counts and 0 active hours.

There is a client with 0 URI counts and 24 active hours for 7 days in a week for 4 weeks, and for most days of 12 weeks. There are 18 clients in a week with 0 URI counts and 24 active hours.

It is a known usage pattern that multiple users can be using the same client_id. I found multiple ping with the same client_id, submission_date_s3 and profile_subsession_counter. These could be duplicate pings or they could be multiple users. One client_id has 1157 pings with the same profile_subsession_counter on the same day, some with Windows and some with Linux. One client_id has 37,964 pings in a week from 10 different countries.

There are also clients with pings with more than 1000 URI counts and 0 active ticks. There are 1,656 distinct client_ids in a week with this usage pattern.

Main summary vs clients daily

To determine if the clients daily table can be used instead of the main summary table, I checked the values in each table for the same date for 1% of users. The summary stats for URI count and subsession hours is the same in both tables, but the summary stats for active hours in clients_daily are higher than for active hours in main_summary. The summary stats for active hours from the two tables are shown below.

	Daily Total Active Hours	
	Clients Daily Table	Main Summary
Count	1045624	1045624
Mean	0.81745	0.64331

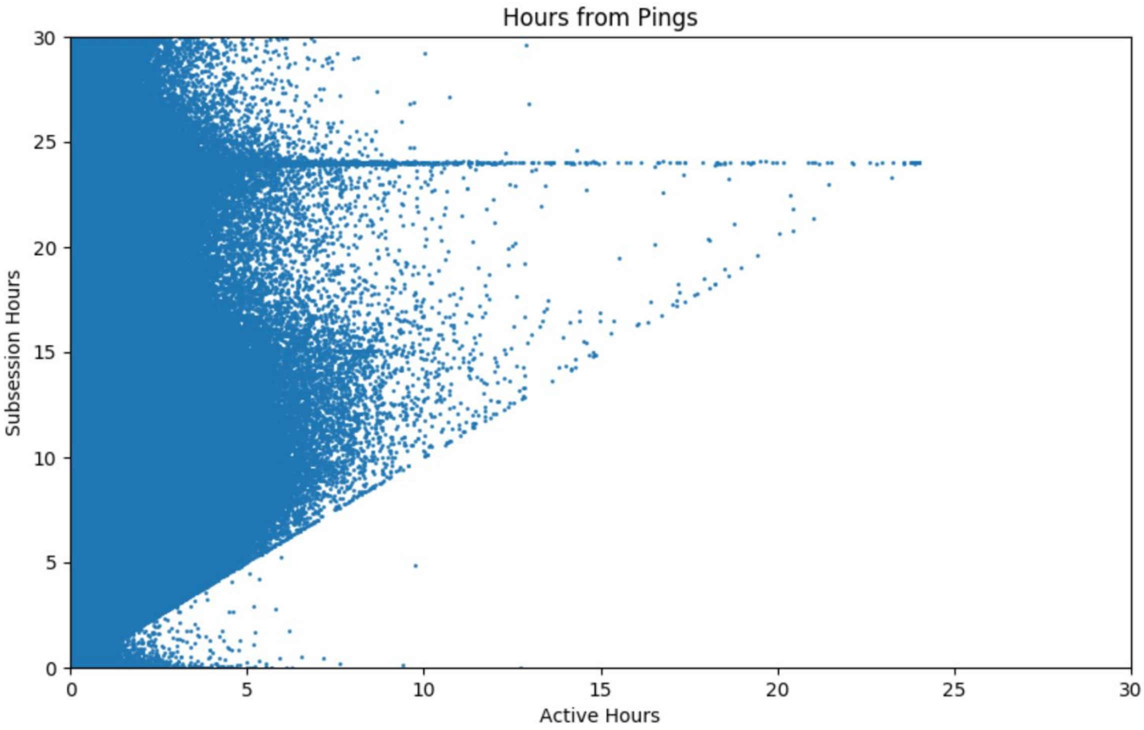
Standard Deviation	16.1003	1.4651
Minimum	0.0	0.0
Maximum	14137.22	436.93

There are 26 records in clients_daily where the active hours were greater than the maximum of the daily total from main_summary. Looking at individual pings for these 26 clients, the pings have 0 active hours. I could find no reason for this discrepancy between active hours in the two tables.

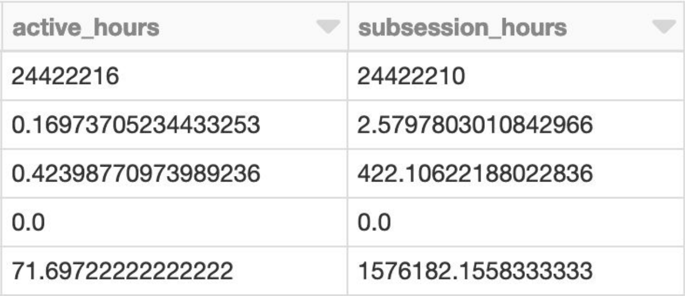
I also discovered data that looked like duplicate pings based on client_id and subsession_id, but had different client_submission_dates. There were 8.5% of the pings for the day that looked like duplicates. Removing the duplicate-looking pings modified the mean and standard deviation of the day's total URI and active hours very slightly but not enough to be meaningful.

Active Hours vs Subsession Hours

To analyze if the heavy user cutoff should be based on active hours or subsession hours, I checked the correlation between the two values. For 1 week of pings, the scatter plot of the two time variables is shown below. The correlation of these values is 0.004.



The summary stats for the 2 time variables for 1 week of pings is shown below. This shows count, mean, standard deviation, minimum and maximum. The maximum value for both should be 24 hours. The active hours have a maximum of 71 hours, but the subsession hours have a maximum of 1,576,182 hours. The standard deviation for subsession hours is 422.



active_hours	subsession_hours
24422216	24422210
0.16973705234433253	2.5797803010842966
0.42398770973989236	422.1062218802836
0.0	0.0
71.69722222222222	1576182.1558333333

Looking at the individual pings, there are only 4 pings in a week with active hours > 24 hours. There are 198,466 pings with subsession hours > 25 hours in one week, representing 131,870, unique client_ids. Active hours indicate when someone is actively using the browser, but will undercount times when watching a video or reading a long article. Subsession hours will capture times when the browser is being passively used. Based on all this information, I chose active hours for the heavy user cutoff.

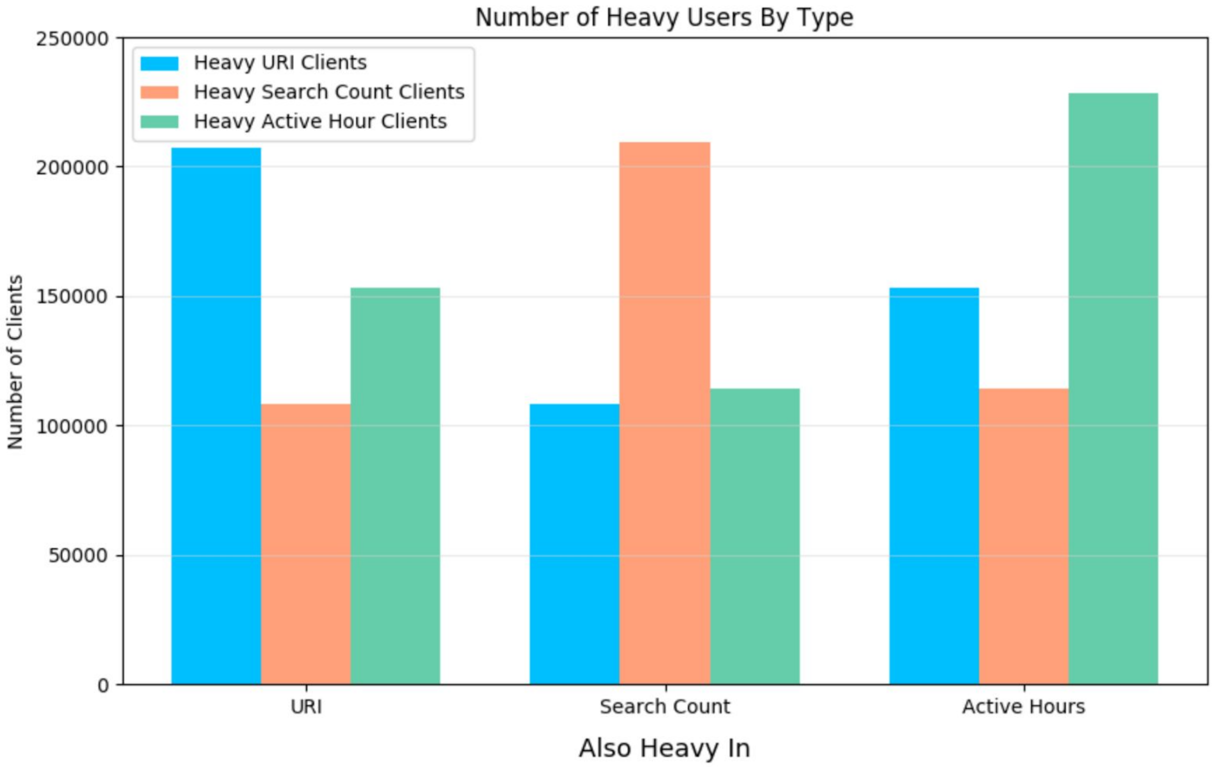
Heavy Users

Now that the heavy user cutoffs have been determined, we can learn more about the heavy users.

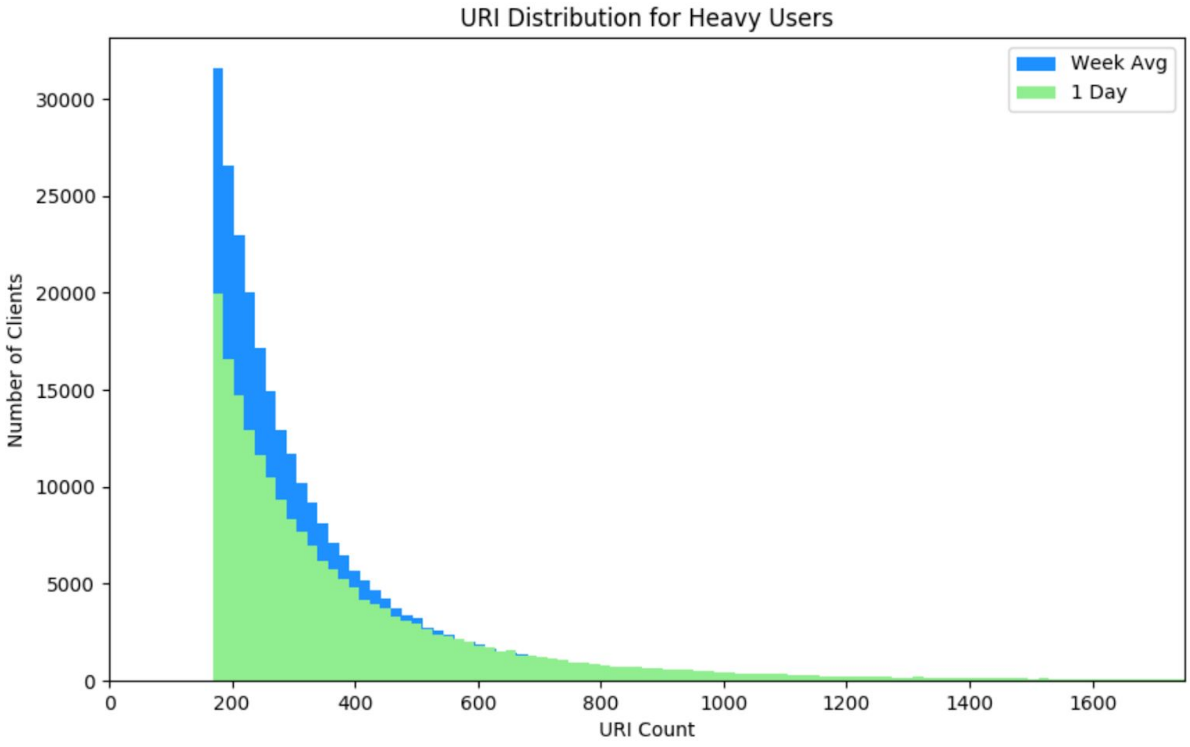
The following table shows the percentage of heavy users of each type for 1 day and for the weekly average. The weekly average of all users is less than 20% because the cutoff was determined from the weekly average of aDAU.

Heavy User Type	**1 Day**	**Weekly Average**
URI	19.93%	14.66%
Search Count	20.16%	16.61%
Active Hours	21.94%	15.69%
Any of the 3	34.47%	27.55%
All of the 3	8.54%	5.86%

The following chart shows the number of heavy users who are also heavy in another type. The first blue bar is the number of heavy URI users who are also heavy in URI, so this is all the heavy URI users. The second blue bar is the number of heavy URI users who are also heavy in search counts. The third blue bar is the number of heavy URI users who are also heavy in active hours. The height of the first green bar is the same as the height of the third blue bar because these are the users who are heavy in both URI and active hours.

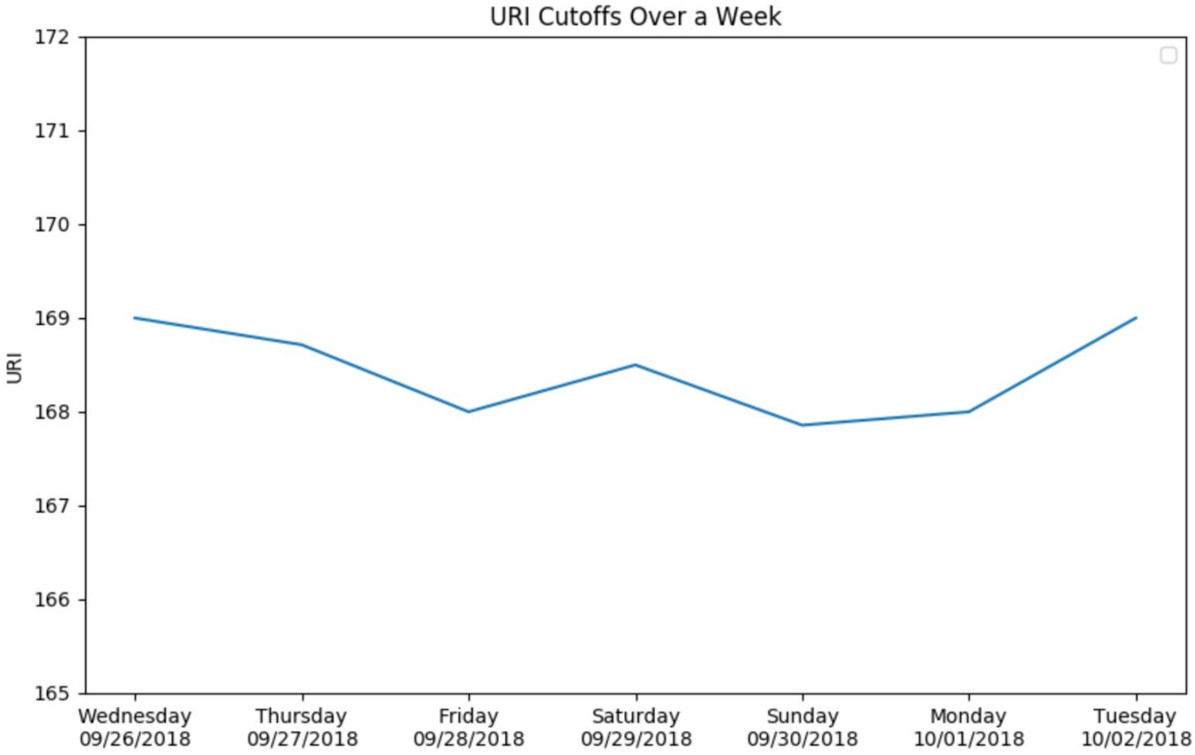


The distribution of heavy URI users is shown below for 1 day of URI values and for the weekly average of URI values. The search count and active hours distributions are similar.

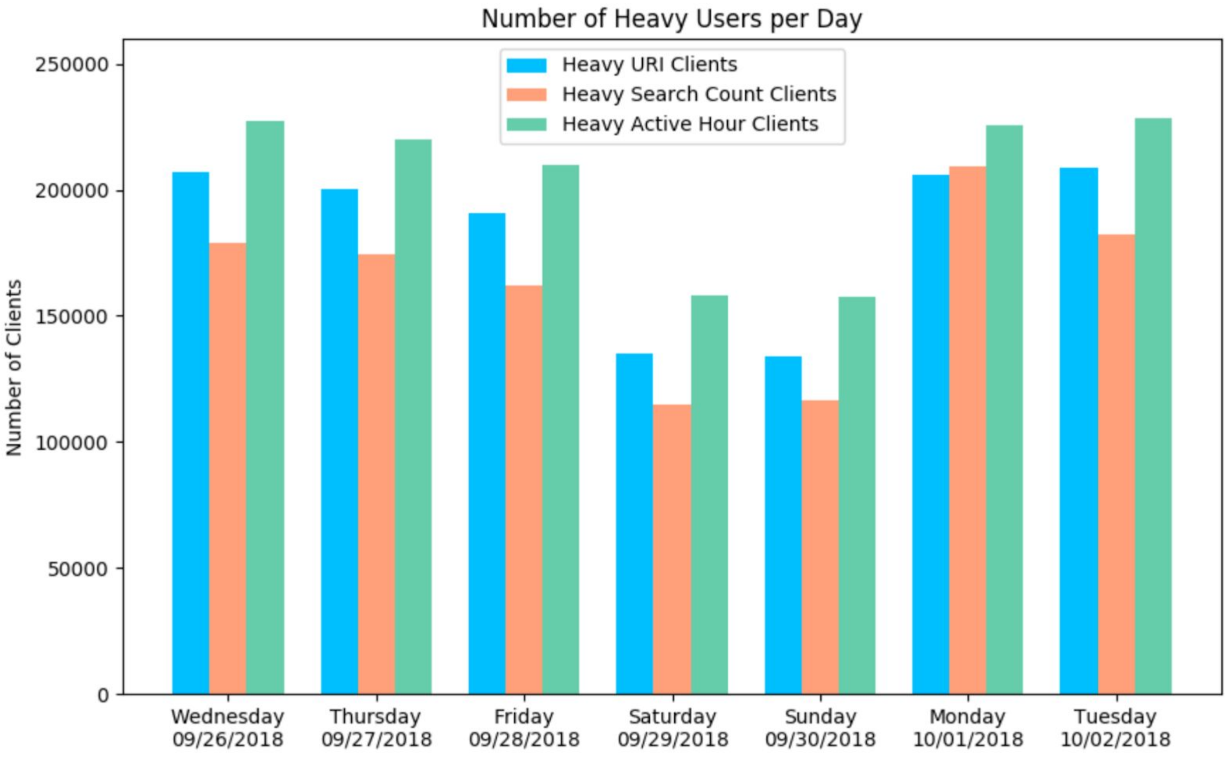


Heavy Users over Time

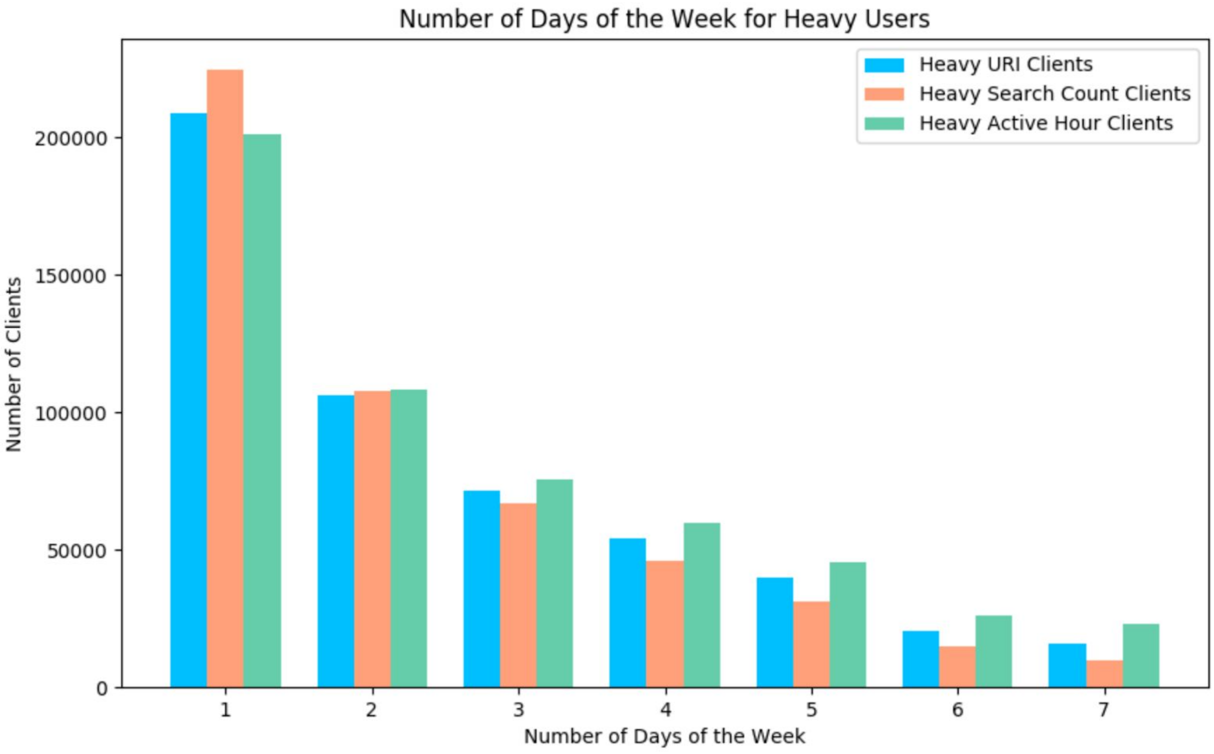
The following plot shows the variation of the URI heavy user cutoff over a week. There is small variation since the values are based on the average of the week ending on that day.



Even though the cutoff is very similar for Saturday and Sunday, the number of heavy users decreases on the weekend as seen on the following bar chart.



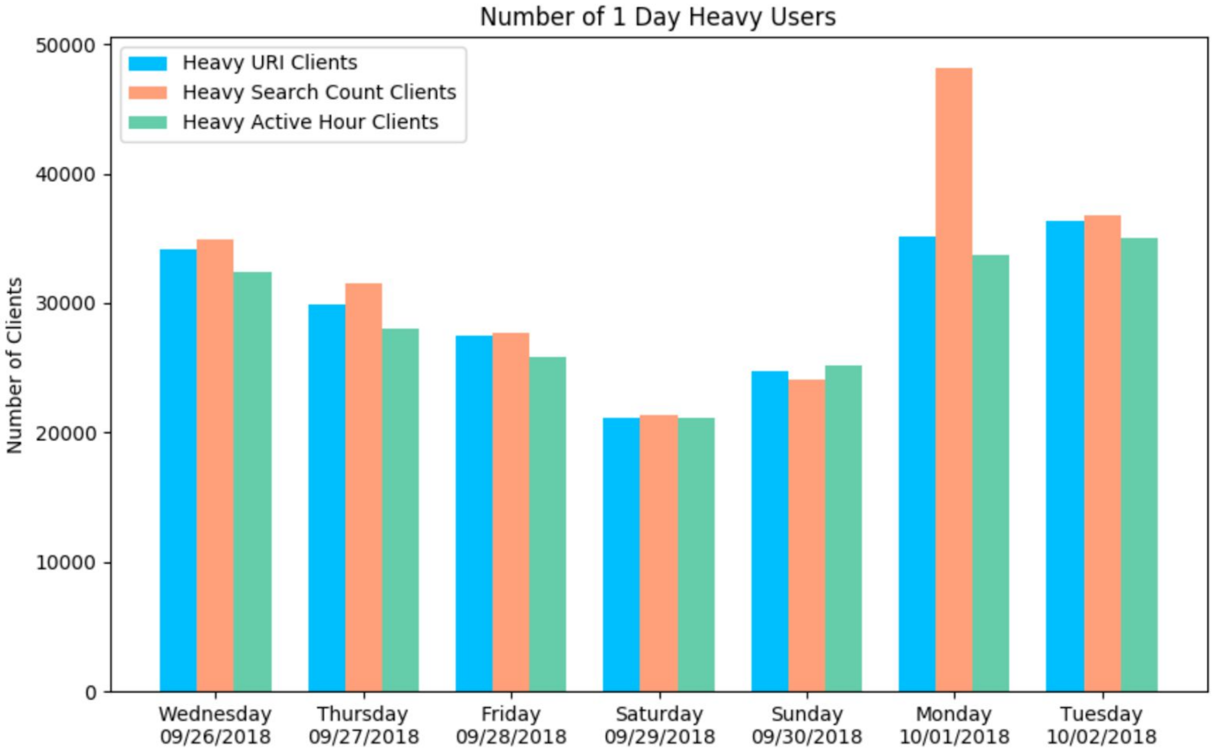
The number of days a week that users are heavy is shown below. The number of heavy user clients decreases for each additional day.



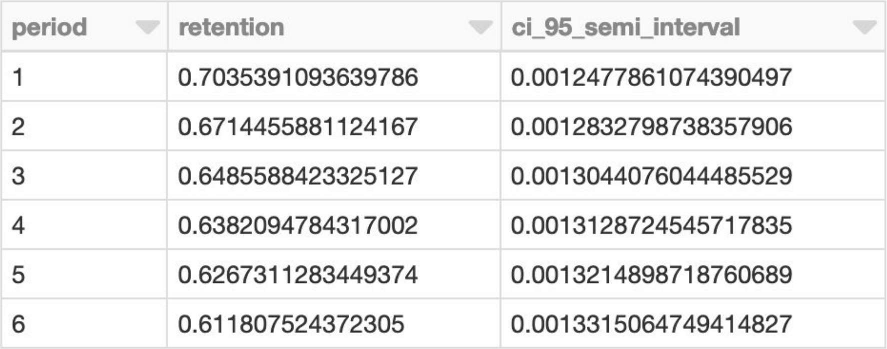
These two plots seem to be contradictory. If the number of heavy user clients is about the same for 5 days of the week, why does the number of heavy users for more than 1 day drop off so much? This is because there is a large percentage of different users on each day. For example, for Wednesday and Thursday, there are 106,420 users who are heavy on both days. But the table below shows the percentage that are heavy on only 1 of the days.

	Wednesday	Thursday
Total Heavy Users for the day	207,344	200,380
Heavy Users only on this day	100,924	93,960
Percentage Heavy Users only on this day	48%	47%

The following plot shows the distribution of heavy clients throughout the week for clients who are only heavy users 1 day of the week.



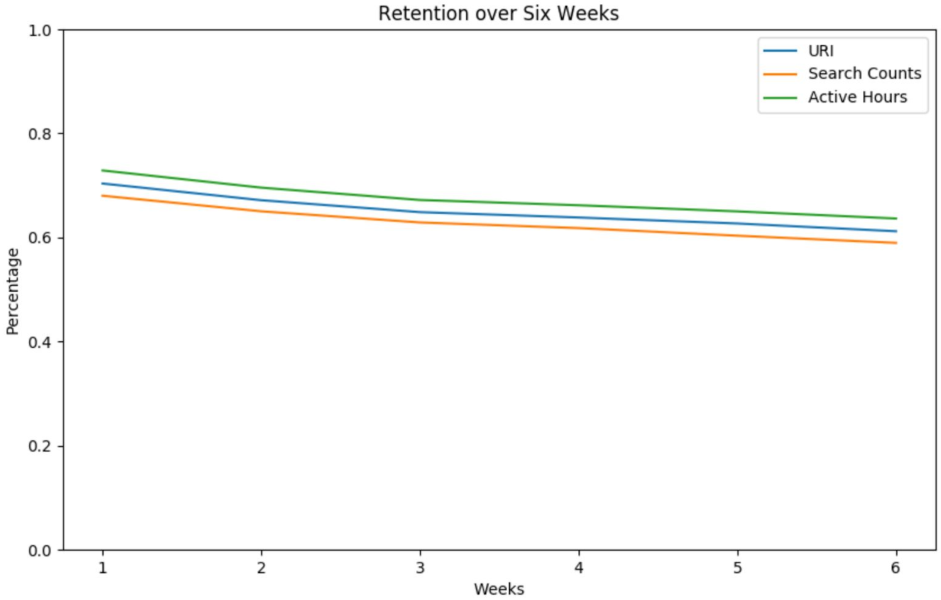
The retention rate shows heavy users for a base week who are still heavy users in later weeks. The 95% CI spans the range `retention ± ci_95_semi_interval`.



period	retention	ci_95_semi_interval
1	0.7035391093639786	0.0012477861074390497
2	0.6714455881124167	0.0012832798738357906
3	0.6485588423325127	0.0013044076044485529
4	0.6382094784317002	0.0013128724545717835
5	0.6267311283449374	0.0013214898718760689
6	0.611807524372305	0.0013315064749414827

URI Heavy Users

The following plot shows that the retention over 6 weeks for all 3 types of heavy users are similar.

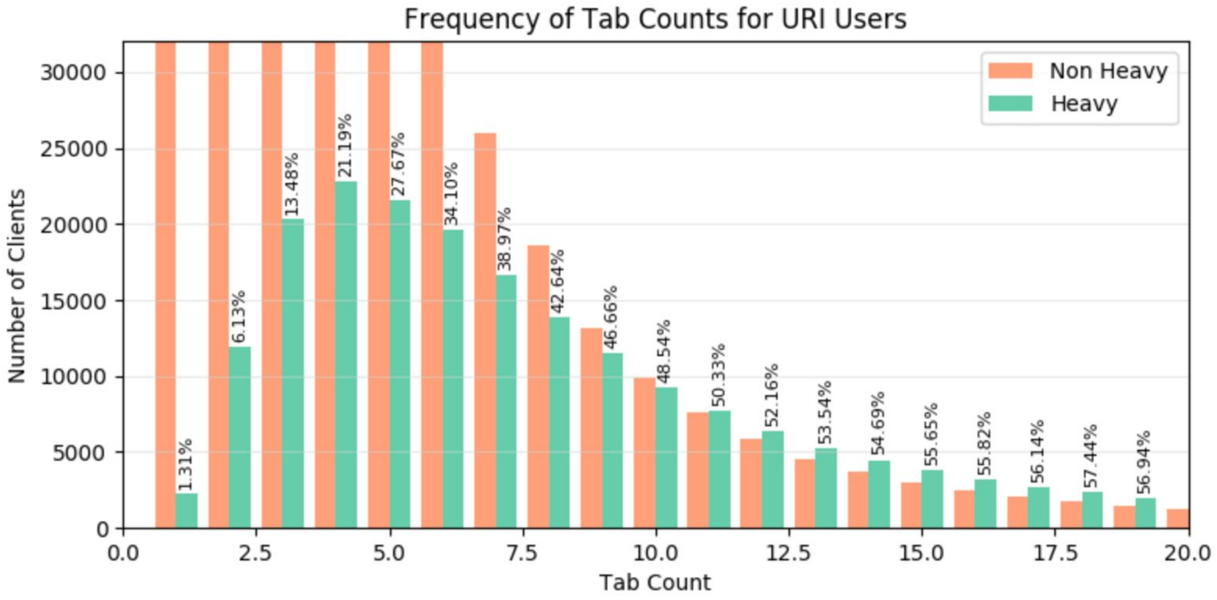


Heavy User Attributes

The frequency of a type (URI, search count or active hours) of heavy and non-heavy user for a specific attribute, such as tab count, is calculated for 1 day of data. For the heavy users on that day, I group by the attribute and count the number of distinct clients. I then repeat that for non-heavy users. Since there is only one record for each client_id on a day, each user should fall into either the heavy user or non-heavy user category. I only consider those clients with a non-null value for the attributes. The percentage above the heavy user bar indicates the percentage of users who are heavy for that specific quantity or type of attribute. For example, 6.13% of all users with 2 tab counts are heavy users, and 16.19% of all users in the beta channel are heavy users.

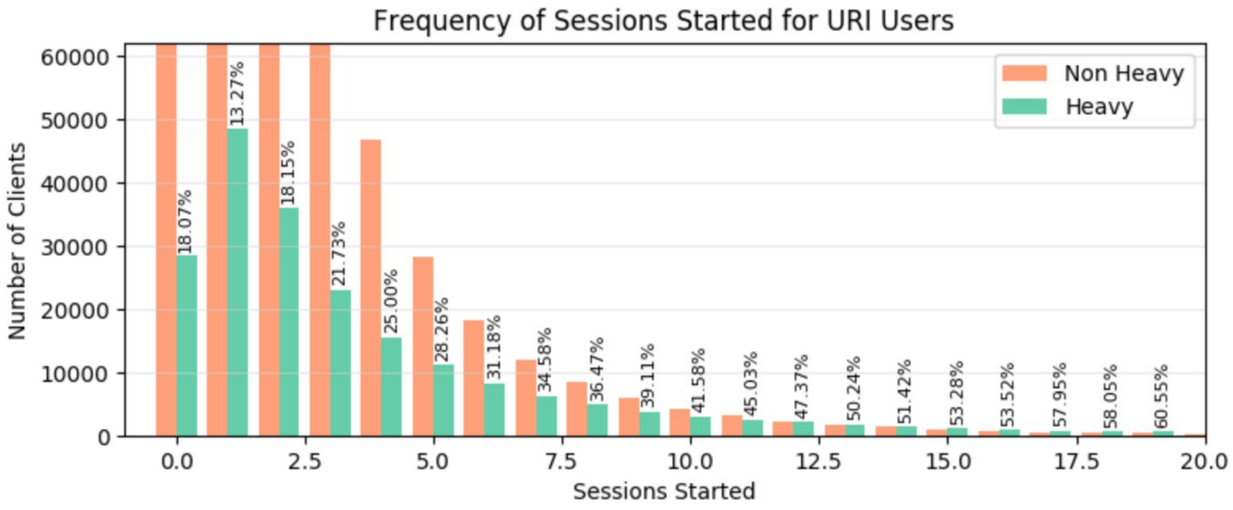
Tab Counts

The frequency of tab counts for 1 day is shown below indicating both the number of clients for a given number of tab counts, and the percentage of clients for a given number of tab counts who are heavy users. The tab count is the maximum tab count for a client id in 1 day, which means this data is for 1 actual user and is not an effect of having multiple users using the same client_id. For higher tab counts, there is a higher percentage of users who are heavy. This same pattern holds for all three heavy user types for tab counts and window counts.



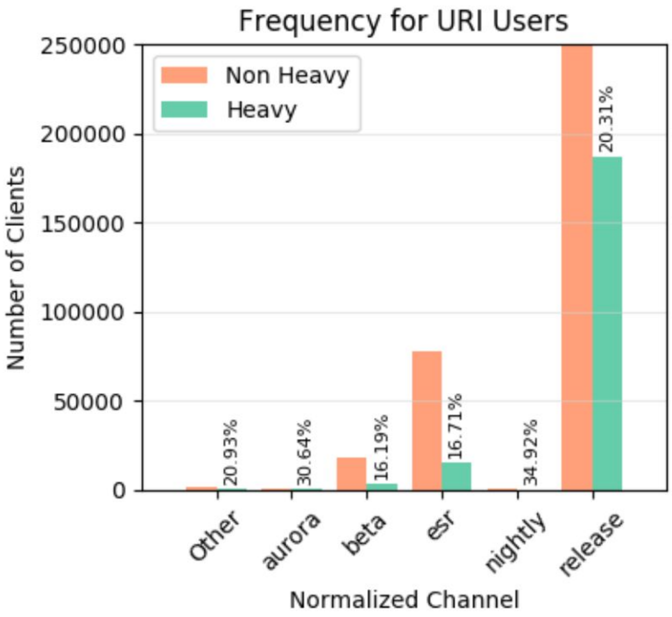
Sessions Started

The frequency of sessions started for 1 day is shown below indicating both the number of clients for a given number of sessions started and the percentage of clients for a given number of sessions started who are heavy users. As the number of sessions started increases, the percentage of heavy users increases. This may be caused by multiple users using the same client_id.



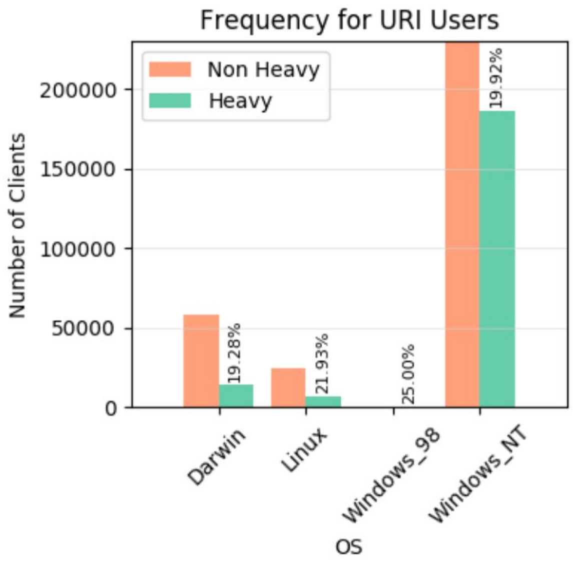
Normalized Channel

The frequency of normalized channel for 1 day is shown below indicating both the number of clients for a given channel and the percentage of clients for a given channel who are heavy users. If the heavy users were distributed evenly among all the channels, we would expect to see about 20% heavy users for each channel. The percentage of heavy users are high for nightly and aurora and low for beta and esr, but the number of users in those channel are low.



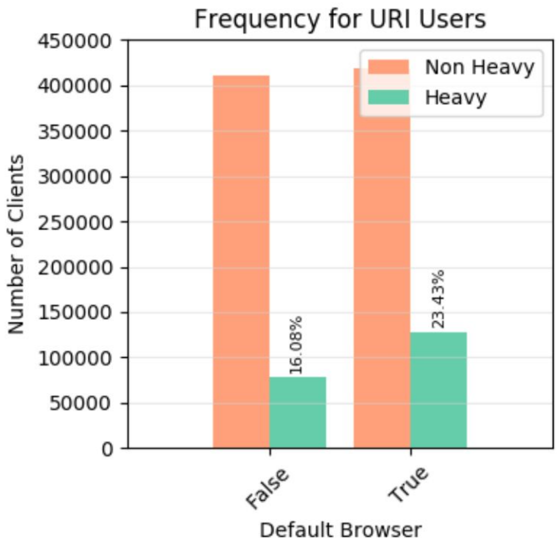
OS

The frequency of OS for 1 day is shown below indicating both the number of clients for a given os and the percentage of clients for a given OS who are heavy users. The heavy users are fairly evenly distributed.



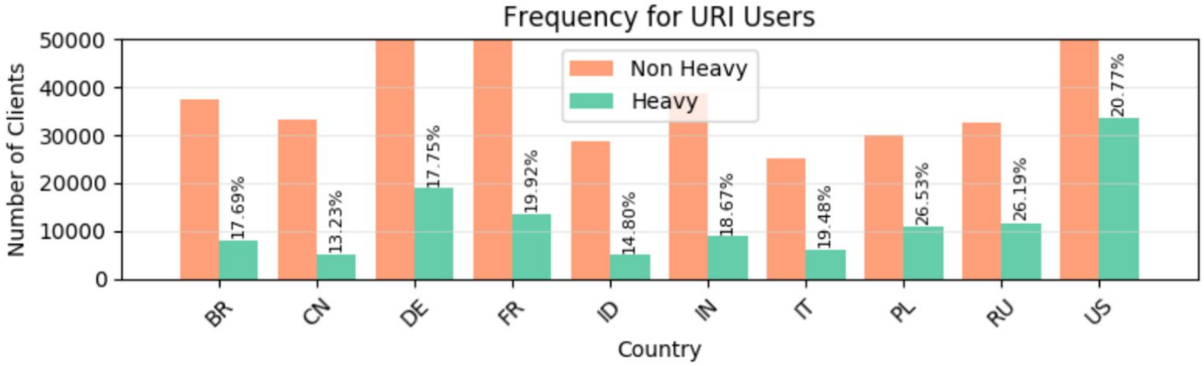
Default Browser

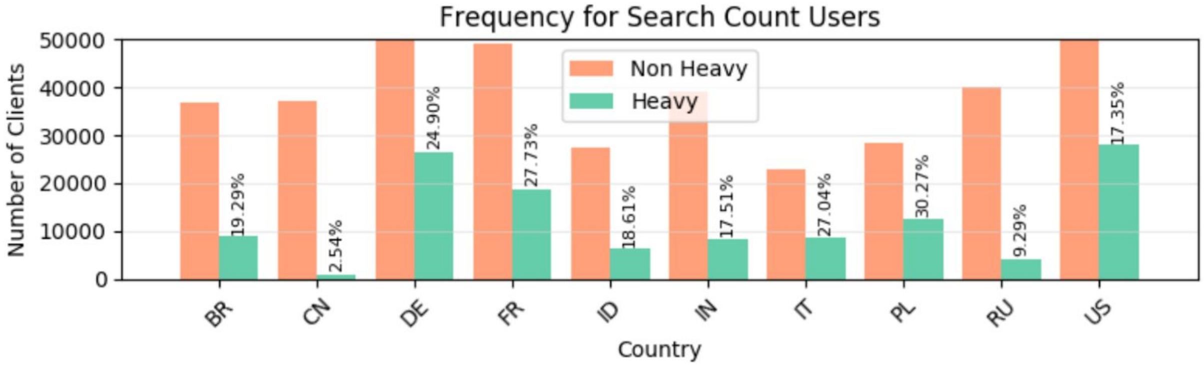
The frequency of default browser for 1 day is shown below indicating both the number of clients for a default browser setting and the percentage of clients for a default browser setting who are heavy users. A higher percentage of clients who have chosen Firefox as their default browser are heavy users.

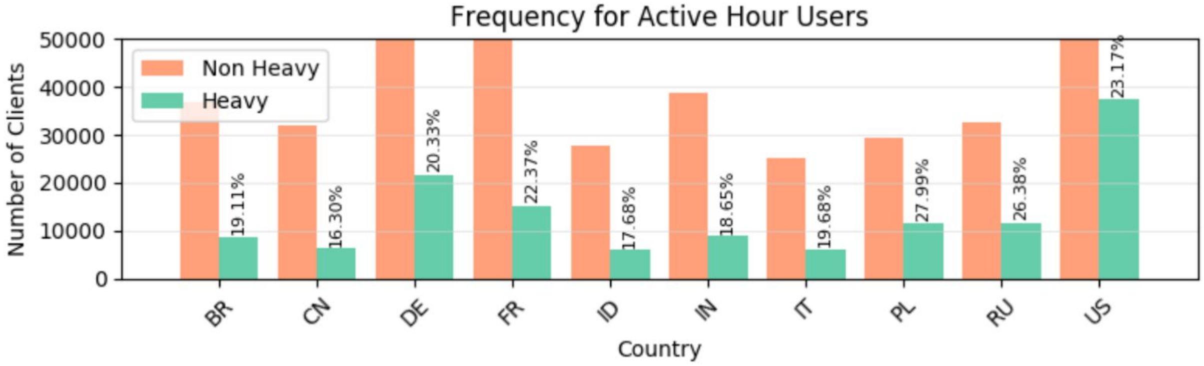


Country

The frequency of countries for the top 10 countries for 1 day is shown below indicating both the number of clients from a country and the percentage of clients from a country who are heavy users. When gathering the data for a day, the first country in all the pings for the day is the one selected. In cases where multiple users in different countries are using the same client_id, this can be hiding information. This is the first attribute where the patterns are different between the heavy user types. China and Russia have low percentages of heavy users for search counts, and France and Italy have high percentages of heavy users for search counts. Poland has a high percentage of heavy users in all three categories.







Key Findings

There is not one cutoff value for heavy users, but one cutoff each for URI, search count and active hours.

- The variables of URI, search count and active hours are not closely correlated.
- The cutoff values will change over time as user populations change.

I propose this definition for heavy user cutoff values on a given day D for the variables URI, search count and active hours:

The 80th percentile of the variable for 1 week ending on day D of daily values where the total daily URI count >= 5, averaged over the week for days that have data.

Averaging over a week helps to make the heaviness measures less sensitive to weekly patterns and general volatility.

Drawbacks to this approach:

- The URI counts do not include private browsing.
- The active hours do not include time watching videos or reading a long article.

Each attribute such as tab count, sessions started on a day, os and country has to be analyzed for each heavy user type.

Future Directions

The next things I would explore are:

1. Research differences in active hours in the clients_daily table. Further analysis would be faster and easier using that table instead of recreating it from main_summary.
2. Look at further attributes such as dev tool use or specific add ons.
3. Apply the cutoff and attribute analysis to different segments of clients, such as for a particular country or locale.
4. Apply this exploration to mobile data.
5. Determine if there is a way to separate multiple users with the same client_id.

Appendices

Previous Analysis:
Heavy users - Proposed definition of heavy users for intra-year work (2016-03)

Operationalizing "Heavy Users" OKR

Attitudes of Heavy Desktop Users

My Analysis:
Databricks Notebook for Heavy User Cutoffs

Databricks Notebook for Heavy Users Over Time

Databricks Notebook for Heavy User Attributes

Databricks Notebook for Firefox Usage Patterns

Databricks Notebook for Main Summary vs Clients Daily

Databricks Notebook for Active Hours vs Subsession Hours