Data Analysis Task Report

Executive Summary

This report describes the findings of analyzing data from an eight-day test of search algorithms "a" and "b" run by the Search Platform Team. Measuring clickthrough rate and zero results rate for the two approaches, we find that approach "a" is clearly the preferred approach. It has a vastly superior clickthrouh rate (66.8%, compared to 17.5% for "b"), while having a slightly lower but comparable zero results rate (18.3% and 18.7%, respectively). We also analyze user preference in search results and find that the majority of the time, users choose the first (top ranked) search result. This preference is consistent across all eight days in the dataset. Lastly, we examine circadian patterns in session length and find that there is no relationship between session length and hour of the day. Based on these findings, it is recommended that search approach "a" is deployed for all users as soon as possible.

0. Dataset assumptions

I will assume that the supplied eventlogging dataset contains session data for an A/B test of two search systems or algorithms given the reference to the Search Platform team in the task's background section, and that the goal is to figure out if one of those two systems/algorithms is the preferred solution. The data will be inspected for issues such as errors and outliers where applicable, but advanced heuristis or filtering to determine "valid sessions" or repairing session data is regarded as outside the scope of this assignment. In other words, I will regard the dataset as generally trustworthy.

1: Daily clickthrough rate

The definition of clickthrough rate is defined by the Search Platform team as "the proportion of search sessions where the user clicked on one of the results displayed". It is worth noting that search sessions can span date boundaries, and I will in this analysis assume that the number of occurrences of these are small enough to be ignored. I will define a "search session" as a session with at least one "searchResultPage" action, and a "clickthrough" as a search session with at least one "visitPage" action. The task is to find the "daily overall clickthrough rate"", which I will interpret as the average clickthrough rate per day over the eight days of data in the given dataset.

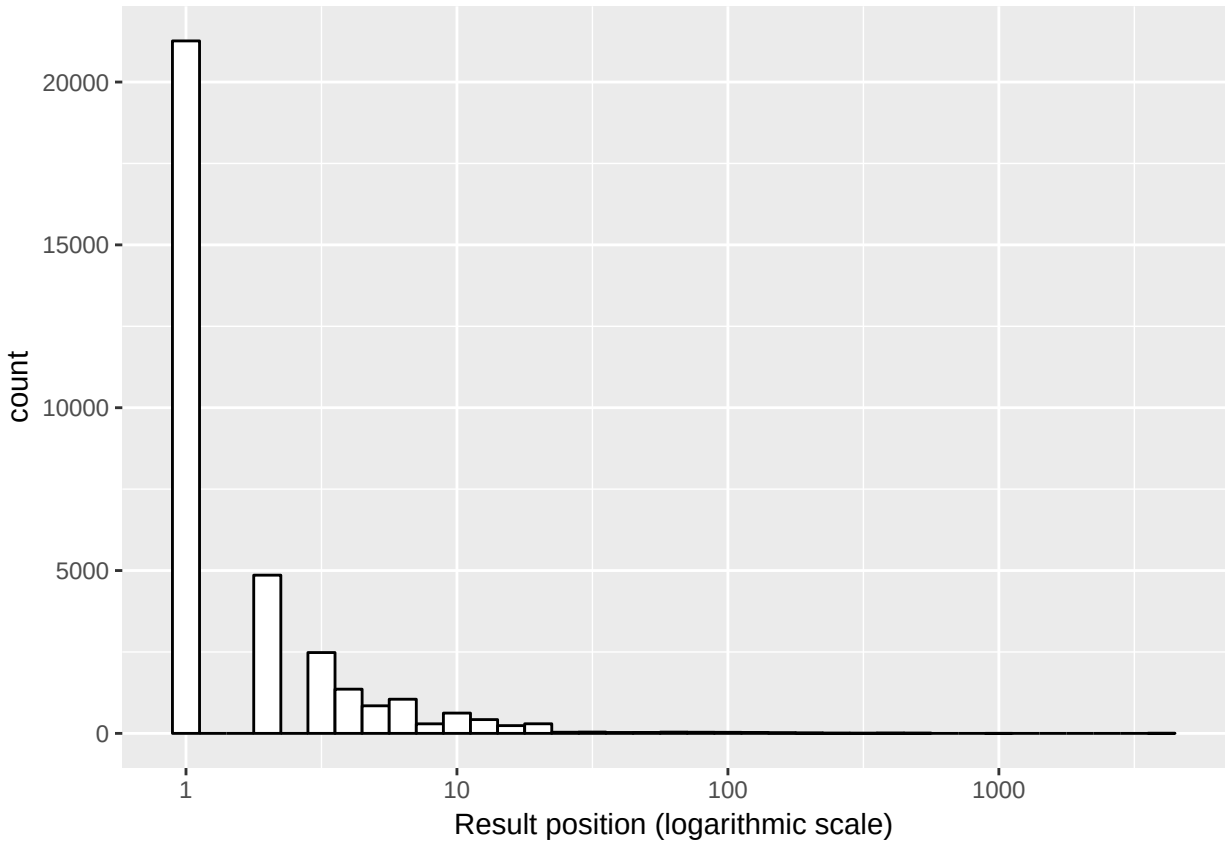
Given these definitions and assumptions, I find that the daily overall clickthrough rate is 38.8%.

How does this rate vary between the two groups ("a" and "b")? First, I verified that no session IDs are assigned to both groups. Knowing that all sessions belong to a single group, I can alter the previous calculation to also take group into account.

I find that group "a" has an average daily clickthrough rate of 66.8%, while group "b" has a much lower rate of 17.5%. Given the size of the dataset and the large difference in proportions it would come as no surprise if further analysis found this to be a statistically significant difference. In other words, that approach "a" is highly preferred given the higher clickthrough rate.

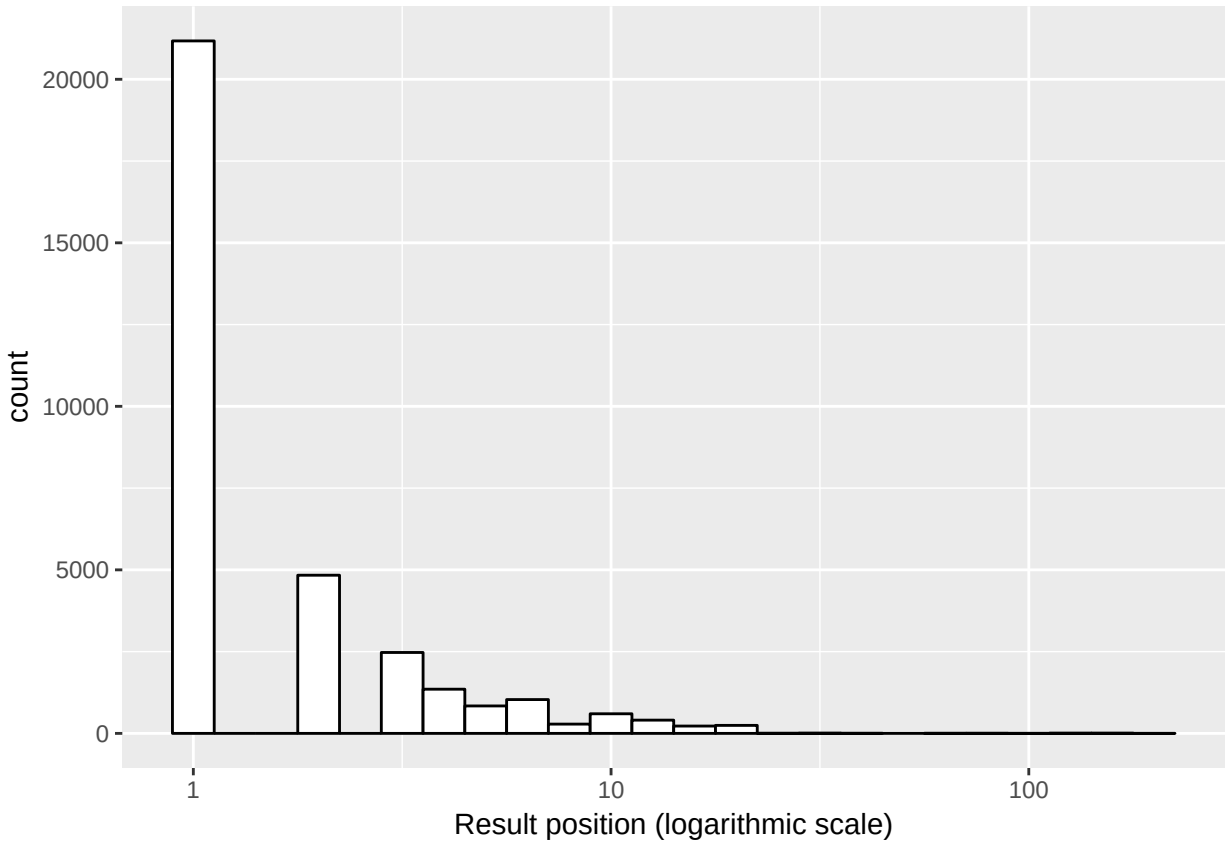
2: Preference in result selection

Which results do people tend to try first? In order to answer this, I first examined the overal distribution of position on the SERP, plotted in the histogram below.



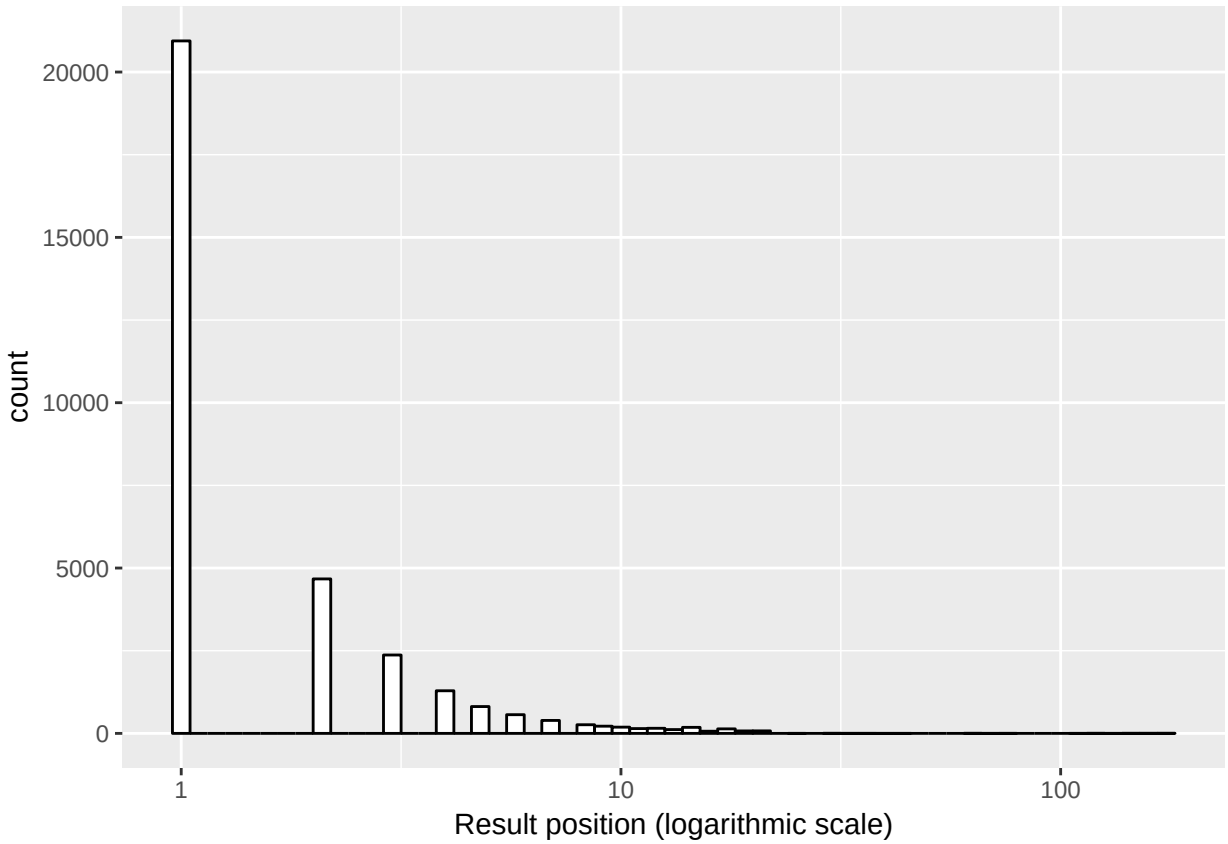
The maximum position is alarmingly large, is it really possible to get more than four thousand search results? Turns out that it is not, the maximum number of search results is 500.

There are several sessions with 500 search results. Without knowing more about the system, I conclude that it's a reasonable maximum that could be available. That number is much lower than the visited page position, suggesting that I should apply a check for whether the position reported in a session is lower than the number of search results. A reasonable approach should in this case be to only accept sessions where the highest search result position is lower than or equal to the maximum number of search results in that session. I apply that heuristic and check the data again, finding that the maximum of those sessions is 179, and it results in the following histogram:



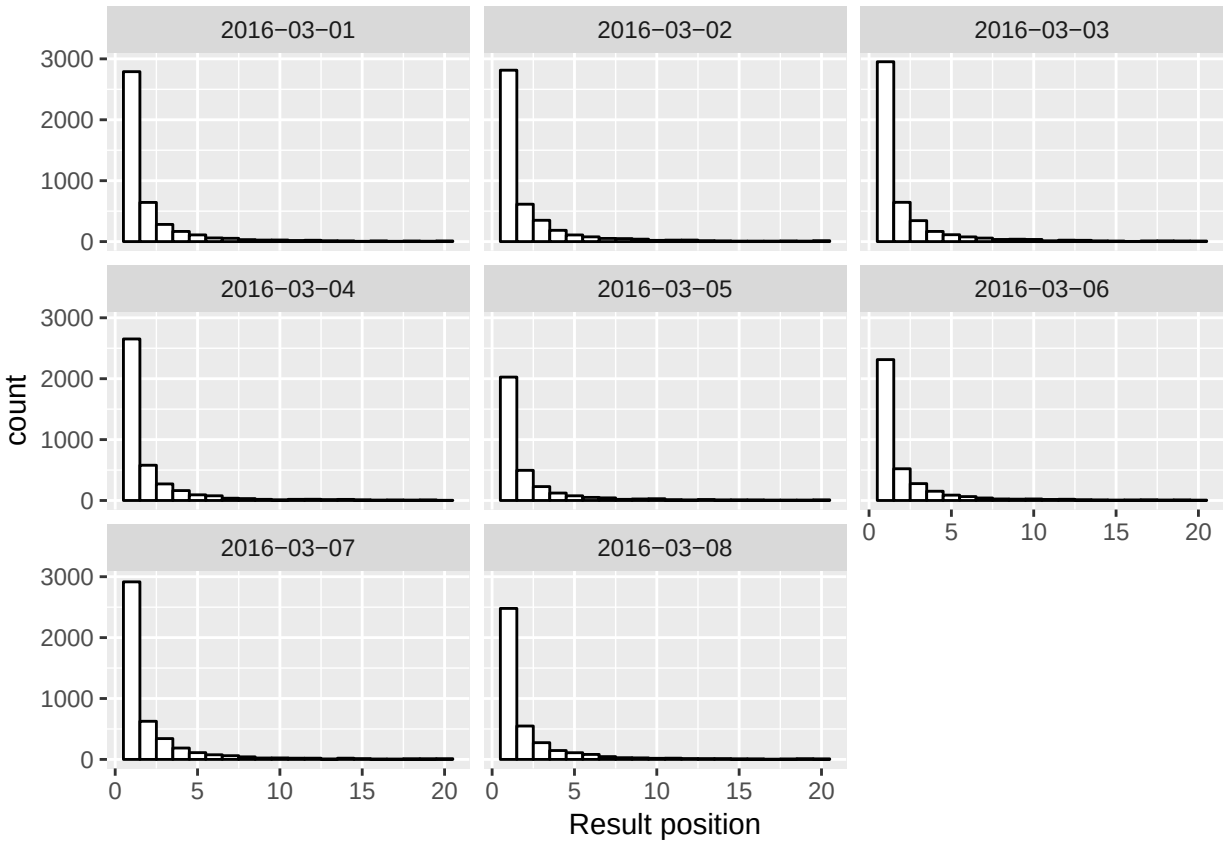
A maximum position of 179 seems more believable given the data suggests that a maximum of 500 search results could be returned. The filter could be further improved by making sure that search position refers to the most recent search instead of the in-session maximum, something I implement below.

The histogram suggests that the first search result is very popular, it's visited more than half the time, but the question is which position does users visit *first*. To do that, I utilize the heuristic above to identify "valid" sessions, and process each session to identify the page visited directly after a SERP. This leads to the following histogram:



We can see from the histogram that the first search result is by far the most popular choice, more than four times as popular as the second position.

Next I look into how this distribution varies across the eight days in the dataset. Based on the histogram above, it is reasonable to limit the analysis to the first 20 results as it is unlikely that we will see much meaningful variation in the data above that threshold. This also allows a switch from the logarithmic X-axis back to a linear axis where each bar in the histogram is equal in width. The plot of a histogram for each day is as follows:



In the plot above, we can see that the first search result is consistently the preferred choice. There is some reduction in volume on March 5 and 6 of 2016, likely due to those being weekend days where activity on Wikipedia is lower (ref for example "Circadian patterns of wikipedia editorial activity: A demographic analysis" by Yasseri, Sumi, and Kertész, 2012) . While it is difficult to ascertain from the plot, the ratio between selection of the first and second result might change during the weekends, e.g. that users are more likely to choose the second result on the weekends. These types of patterns could be further studied in a follow-up research project.

3: Daily zero results rate

The "zero results rate" is defined as the proportion of searches that yielded zero results. I will again interpret "daily overall zero results rate" to ask for a daily average across the week-long dataset. I can calculate that using an approach similar to what I did for clickthrough rate, although in this case we are only interested in "searchResultPage" events. I find that the daily overall zero results rate is 18.4%%.

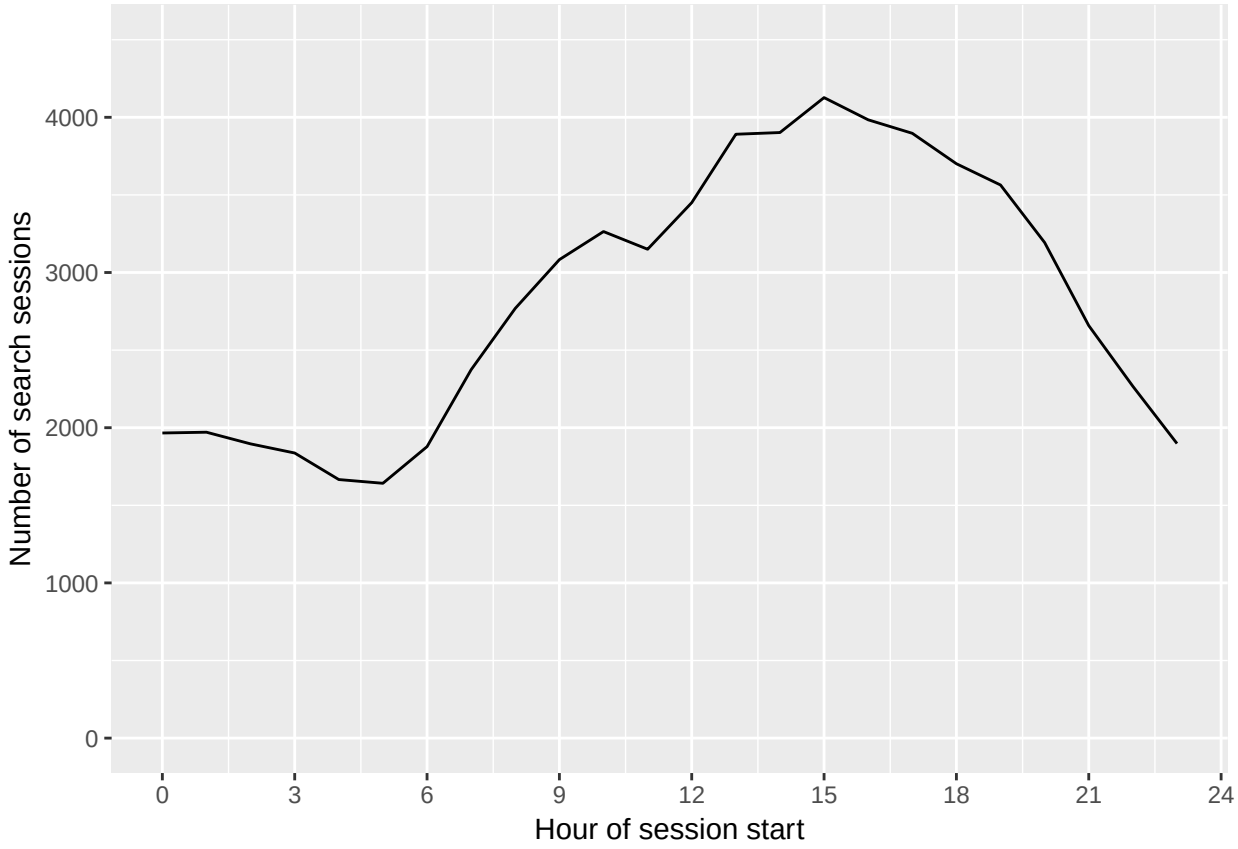
How does this rate vary between the two groups ("a" and "b")? Similarly as we saw for clickthrough rate, the calculation can be adapted to average across the dataset for each group separately.

The zero results rate for group "a" is 18.3%, while for group "b" it is 18.7%. Further analysis might reveal that this is a statistically significant difference, but it is also not a particularly large difference. What is worth noting is that approach "a" has again the better result with a lower average zero results rate, further substantiating that it should be the recommended approach.

4: Circadian patterns in session length

Per the task, "session length" is defined as approximately the time between the first and last event in a session. I am interested in understanding whether there is a relationship between time of day and how long sessions last. Previously, I referred to research on circadian patterns in Wikipedia editor activity, and it might be that this shows up in our search session data as well.

First, I'd like to investigate whether the appears to be a pattern in search volume (number of searches) depending on the time of day. If there is no such pattern, then it would be surprising if we happened to find a pattern in session length. Let's plot number of search sessions per hour of the day:



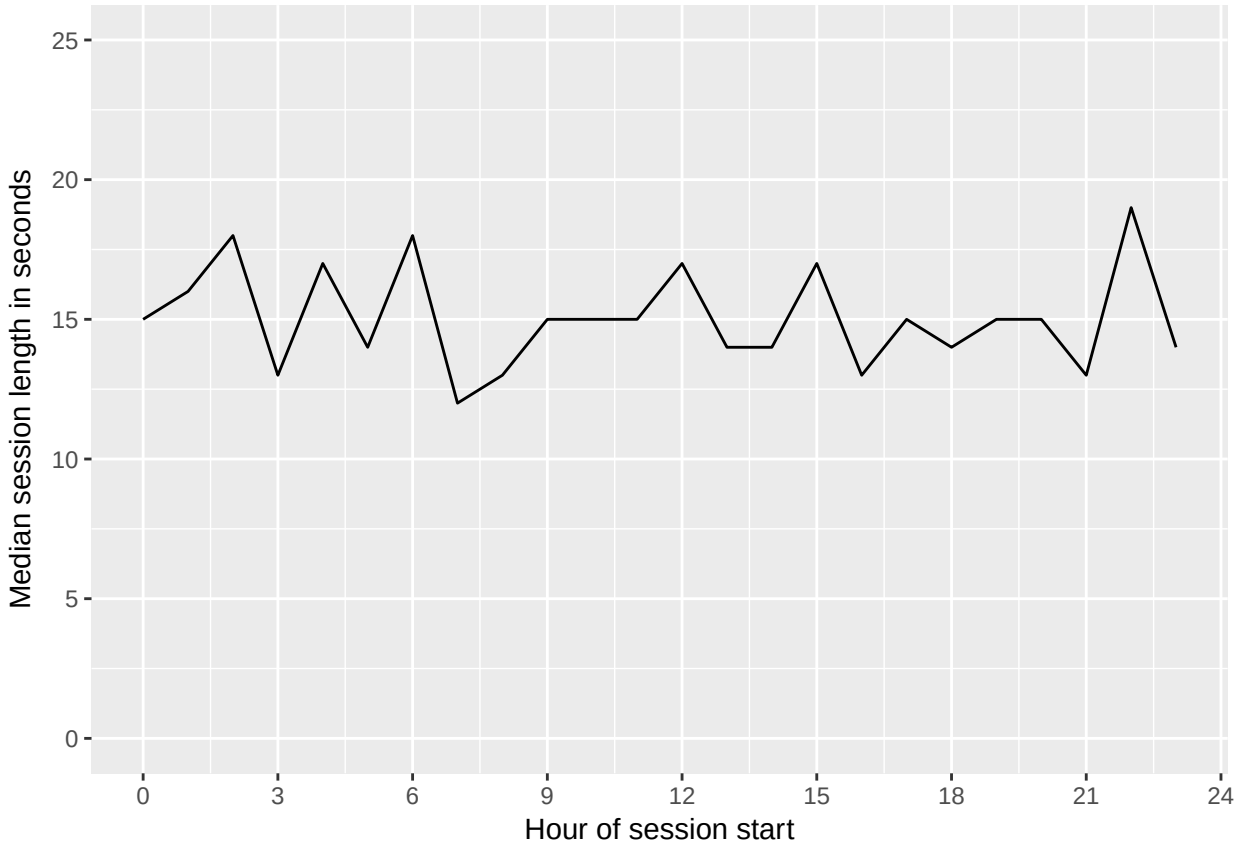
Based on the plot above, it appears that search activity has a strong daily pattern with more activity during what would be daytime in Europe. Since there appears to be a strong pattern in search activity, it would be interesting to see if time of day also affects session length. First, I calculated session length and sanity checked the data, finding that there are many sessions of very short length (0 seconds), and the maximum session length is almost six days.

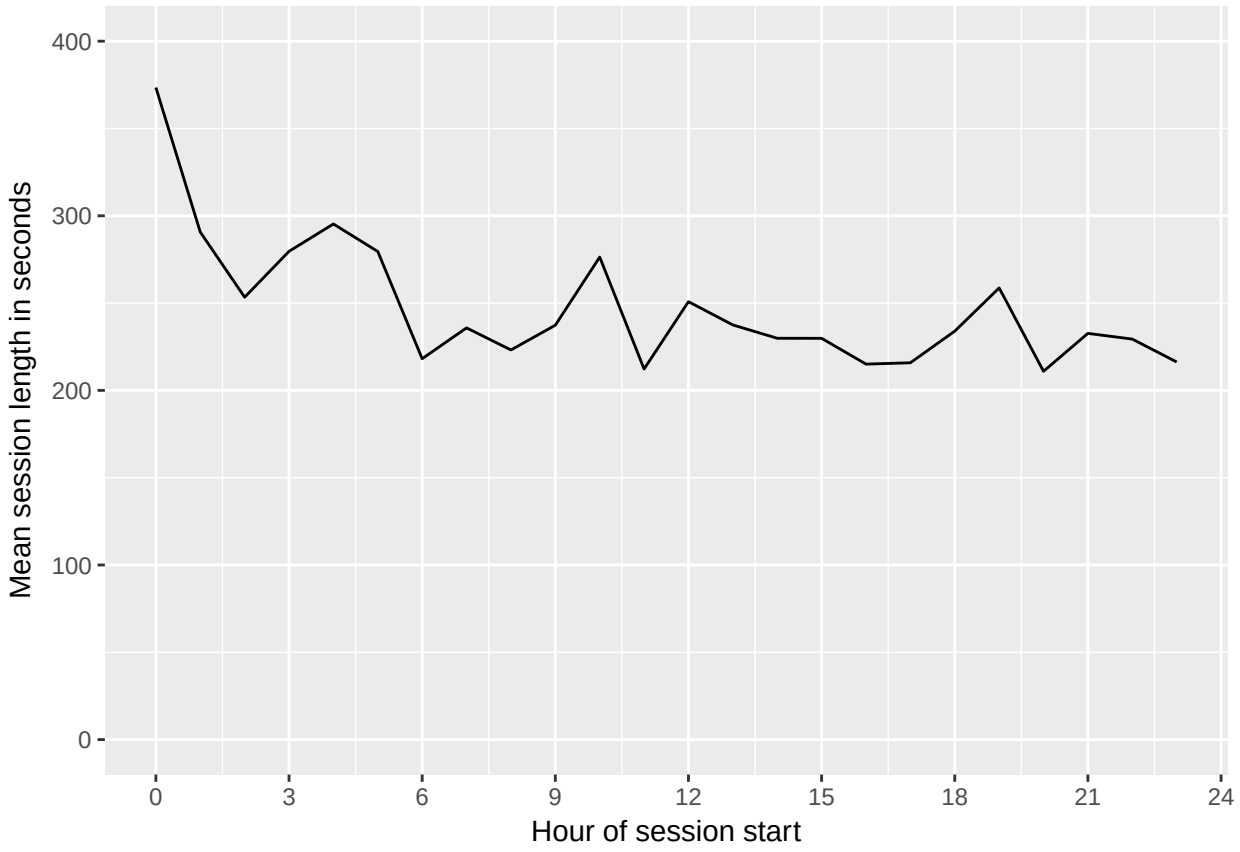
The zero-length sessions are all sessions that only include a single search, and we see that more than 25% of the sessions are like this. Since they make up a substantial number of sessions, I will include them in the analysis and use a statistic that accounts for a skewed distribution. Further work in this area could aim to define an "average session length" for these types of sessions (similar to Geiger and Halfaker, "Using edit sessions to measure participation in Wikipedia", 2013), but for now that is outside the scope of this assignment.

When it comes to long sessions, I find that there are not many of them. Only two sessions are longer than 24 hours, and 29 are longer than four hours. An inspection of the sessions that are longer than 24 hours suggests that they should have been split up into multiple sessions. Instead of doing that (using for example the approach suggested by Geiger and Halfaker cited above, or Halfaker et al. in "User Session Identification

Based on Strong Regularities in Interactivity Time", 2015), I choose to label them as erroneous outliers and discard them from the analysis.

The two plots below show the relationship between hour of session start and session length. The first plot includes all zero-length sessions and calculates the median session length due to the skewed distribution. The second plot exludes all the zero-length sessions and calculates the mean session length.





From the plots above, we can see that there does not appear to be a circadian pattern in average session length. When including sessions that only contain a single search, we see that median session length stays fairly consistent around 15 seconds across the day. Exluding zero-length sessions and instead measuring mean session length, we can see that average session length generally varies somewhat between three and a half and five minutes, but there is not a particular pattern correlating with the hour of the day.