OCF Lab Session Analysis Part 1: How Long Does It Take To Get a Computer

Shicheng Huang

April 8, 2018

Introduction

I am a volunteer staff member for the Open Computing Facility (OCF) at the University of California, Berkeley, where we provide free computer access to all students. Additionally, we also let students print a maximum of 10 pages per day and 100 pages per semester. As a staff member who spends an average of 7 hours per day in the lab, I often see people waiting for a computer. I wonder if I can have a decent estimate of when people have to wait for a computer. To break down the question, I first try to estimate the wait time for a single computer, since I sometimes have to wait for the particular computer at the corner of the lab. In this post, I will explore how does a desktop session remaining duration distribution changes conditioned on the current session duration. We define current session duration the time a student has been on a computer, and the remaining session duration how long it will take for him to leave to computer.

Session Dataset

The dataset we use is the lab session data this semester. Below is a snippet of the session data. The field "host" represents each desktop. The field "duration" measures the duration of a session by minutes.

	id	host	start	end	duration
1	82360	acid.ocf.berkeley.edu	2015−07−18 21:01:02	2015−07−18 21:05:01	4.0
2	82423	acid.ocf.berkeley.edu	2015−08−25 15:02:30	2015−08−25 15:04:43	2.2
3	82426	acid.ocf.berkeley.edu	2015−08−25 15:08:14	2015−08−25 15:08:29	0.2
4	82459	acid.ocf.berkeley.edu	2015−08−26 09:13:27	2015−08−26 09:24:07	10.7
5	82470	acid.ocf.berkeley.edu	2015−08−26 09:31:58	2015−08−26 09:54:10	22.2
6	82483	acid.ocf.berkeley.edu	2015−08−26 09:55:29	2015−08−26 10:00:37	5.1

Basic Data processing

Here are the procedures I use to clean the data:

1. Because the lab volunteer staff often uses the desktops much longer than regular users and have very different login patterns, I exclude all sessions from the volunteer staff.
2. I exclude sessions that have 0 or negative durations. This is mostly a data engineering issue because it is physically very difficult and rare that some user logins and logouts within 1-2 seconds to have a 0 session duration.
3. I exclude sessions from host "blizzard" and "eruption" because they are desktops exclusive to front desk staff members and volunteer staff memebers.

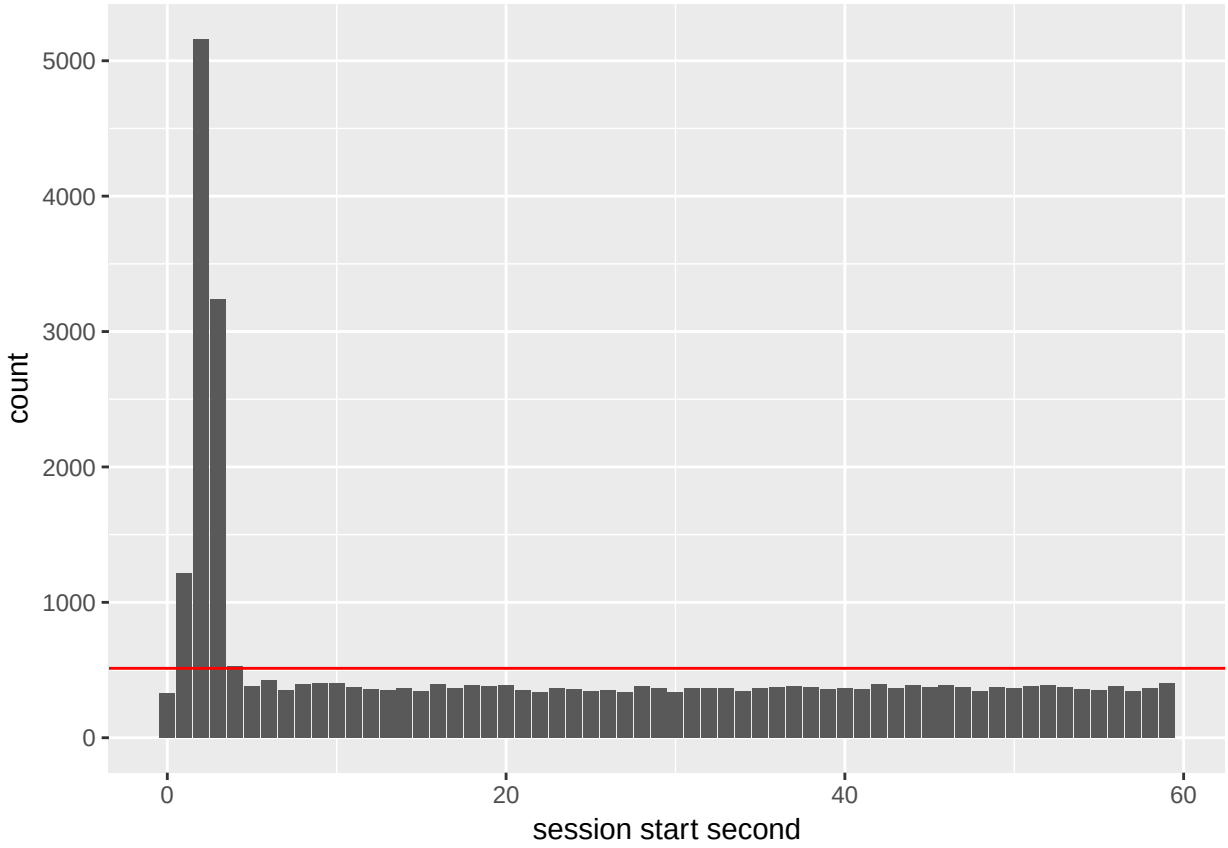
4. I exclude sessions during the weekends because weekend sessions are often longer than the weekdays'. The table below shows each weekday's mean and median session duration. The weekends are highlighted.

	day_of_week	mean_duration	median_duration
1	Saturday	24.92549	7.033333
2	Sunday	23.26227	6.475000
3	Friday	15.44429	5.450000
4	Thursday	14.87453	5.433333
5	Monday	13.54434	5.233333
6	Tuesday	14.55749	5.166667
7	Wednesday	14.49821	5.066667

Data Adjustments

There used to be a bug in the lab's session tracking infrastructure: we could only record a session's start time at the beginning of the minute (know more about the bug from one of my previous post. As a result, lots of sessions appear to "start" around the beginning of the minute, but they actually started the minute before.

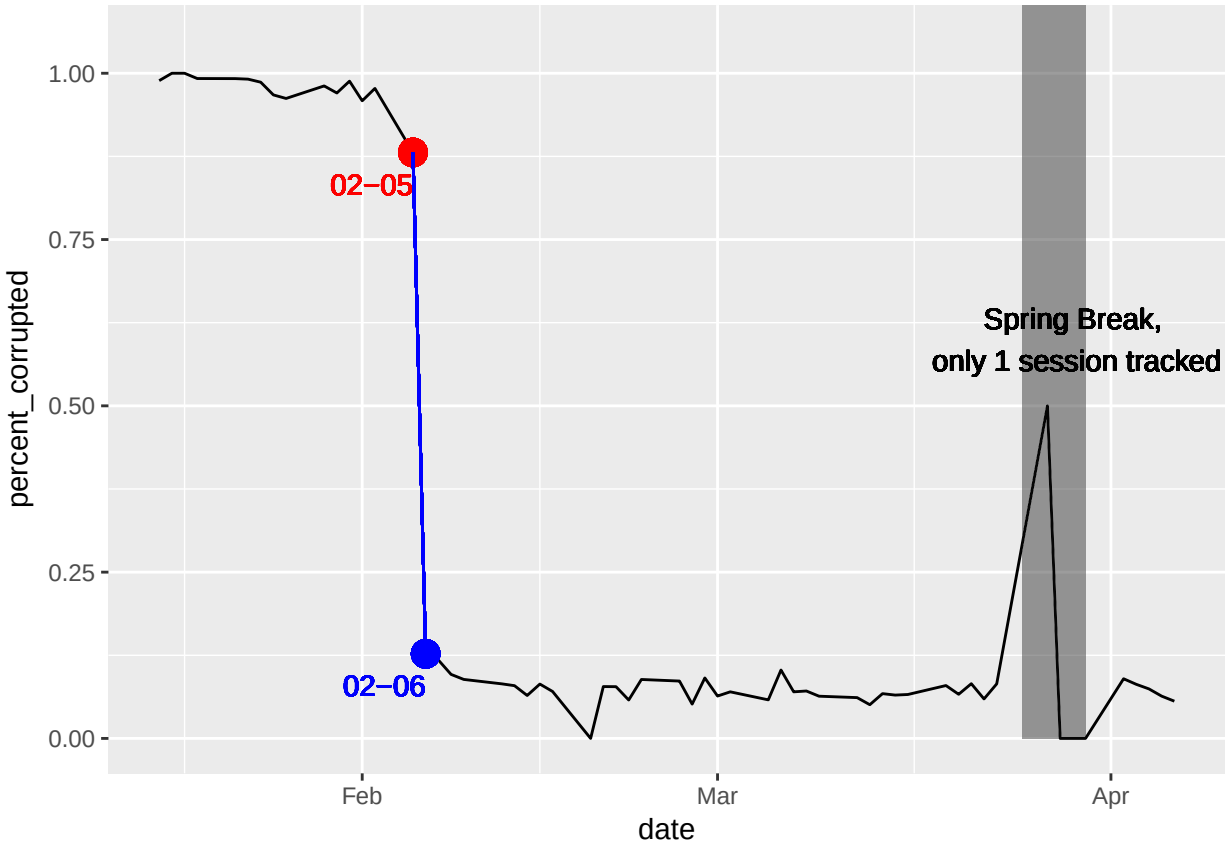
See figure below about the distribution of session start seconds.



To find out the time interval when sessions data that are corrupted, I make a plot shows the percentage of

sessions with start seconds < 4 against time. The darker and bigger the dot is, the more sessions are there in the day.

We can see the percentage is abnormally high until early February. The "peak" in the end of March is Spring break. As the graph below show, after Feb 6th, the session tracking system goes back to normal again. ## not clear ##



Thus, for all the session before 2018-02-05, I will adjust the duration by adding a random variable that is uniformly distributed from the set $\{0, 1, 2, \ldots 55 + S_{session\ start\ second}\}$.
Because if the session's recorded start time is X given our tracking system was malfunctioning, the real session start time could be from 0 to X or 5 to 55 from the previous minute. So the real session duration should be anywhere between 0 to $X + 1 + 55$ seconds longer.

Let's have a rough look at the difference before and after the adjustment through some summary statistics.

	Min.	X1st.Qu.	Median	Mean	X3rd.Qu.	Max.
adjusted_duration	0	2.917	5.4	14.71	13.22	571.5
original_duration	0	2.783	5.25	14.58	13.05	571.5

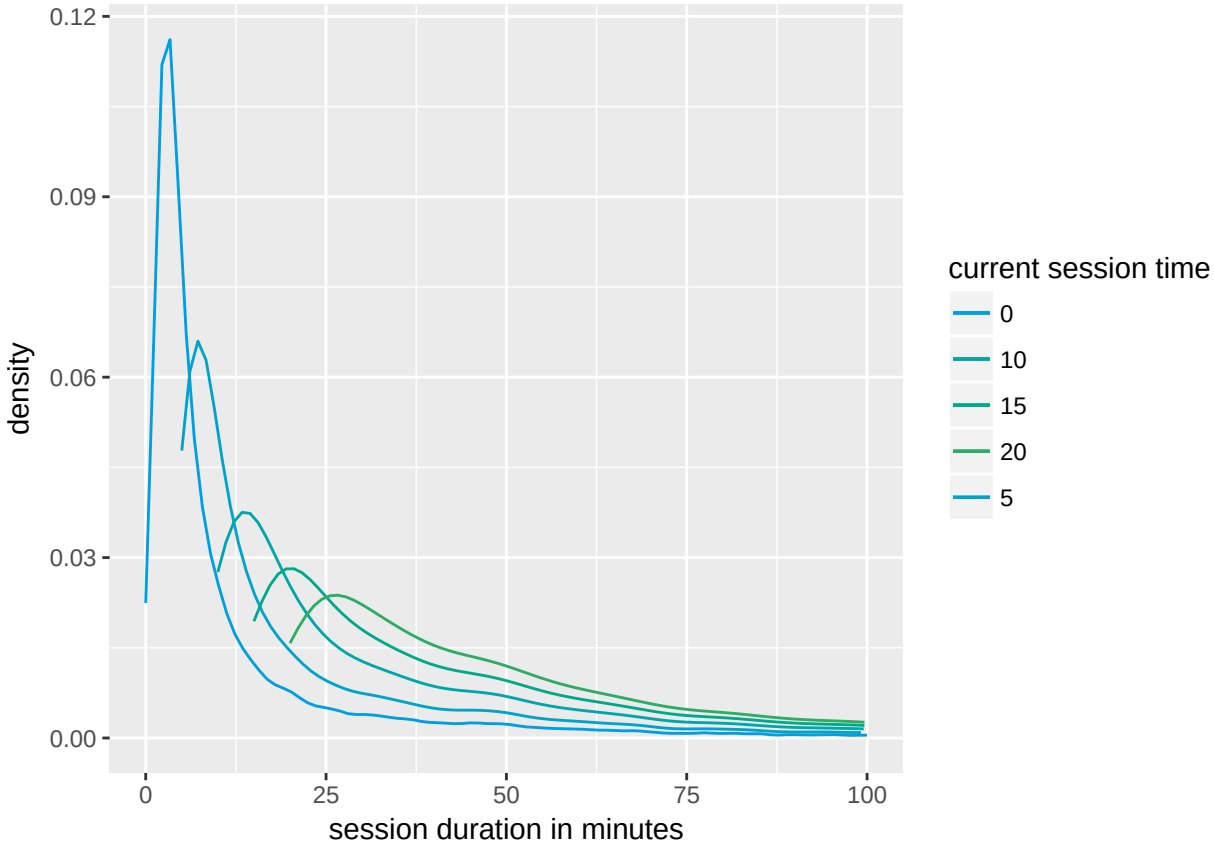
There isn't much visible difference but I think it is still important to take good care of the data inaccuracy issue.
We can see 75% of the sessions are under 14 minutes. And 95% of the sessions are under 65 min. In fact, 99% of the sessions are shorter than 145min. Maybe most of the users come to the lab just to print (I will investigate further in part 2).

Session Analysis

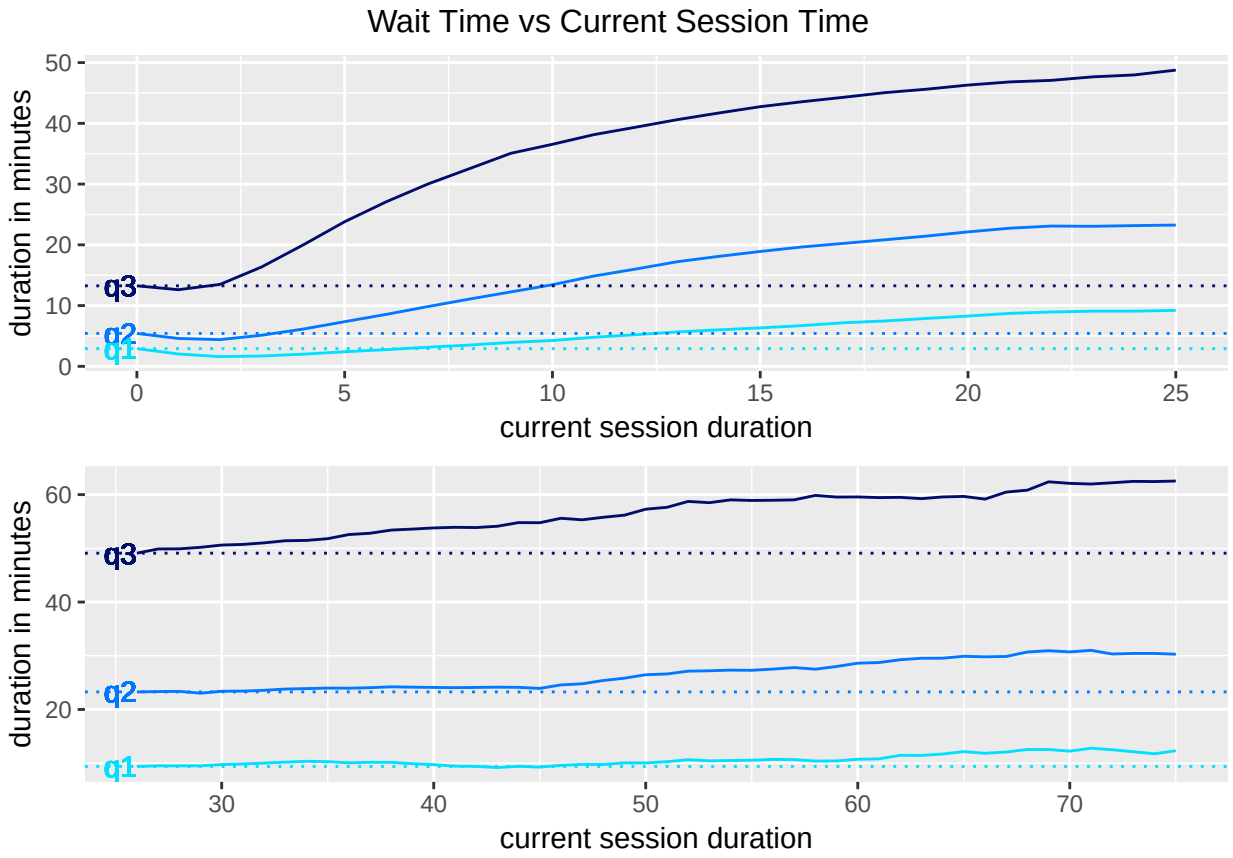
Lets first look at the distribution of the session duration. Because the raw histogram is extremely skewed, I make two other histograms with x axis limit (0, 500) and (0, 50) respectively.

While taking a closer look, it seems like lots of sessions are under 100 minutes. To visualize a skewed distribution, we can also see its different quantiles.

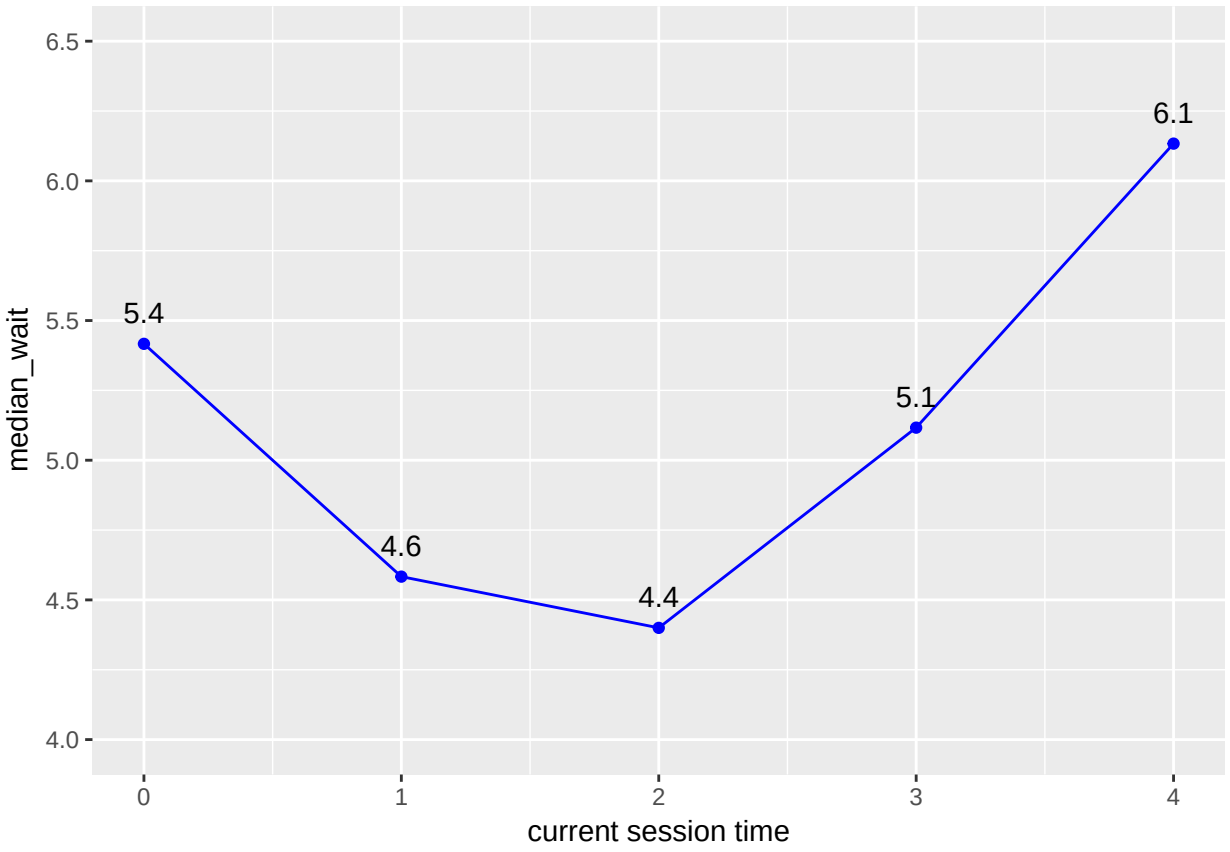
Expected Wait Time Given Current Session Duration



mean and median has 0.9956232 correlation. So they are only different in scale.



As current session increases, expected wait time increases. This increasing pattern slows down after 25 minutes or so (check the second graph). Also, there is a small dip in the first few minutes, let's take a closer look.



since many duration are cluster at the 2 - 3 min mark, we can anticipate a session will last only around 2-3min or so. As a result, remaining session time (both mean and median) is the lowest when the current session time is 2-3 min.

Future Directions

In part 1, we learn that a majority of the sessions are short but its distribution is extremely right skew with some unreasonable outliers (700+ minutes). Therefore, it is better to look at different quantiles instead. And we find that knowing how long a session has been (i.e the current session time) can help us infer how long the session will last from the current time (i.e remaining session time), but this heuristic's effectiveness will decrease after current session time exceeds 20 minutes or so.

In the next part, I will examine additional variables **day-of-the-week** and **the computer used**. I do expect sessions during the weekends will be longer because it seems like more people come to the lab just to study instead of printing; and students may favor computers differently because of their locations in the lab.