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. Because I sometimes have to wait for the particular computer at the corner of the lab. In this post, I will explore if I can have a good estimate when a user will leave given how long he/she has been using a 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:

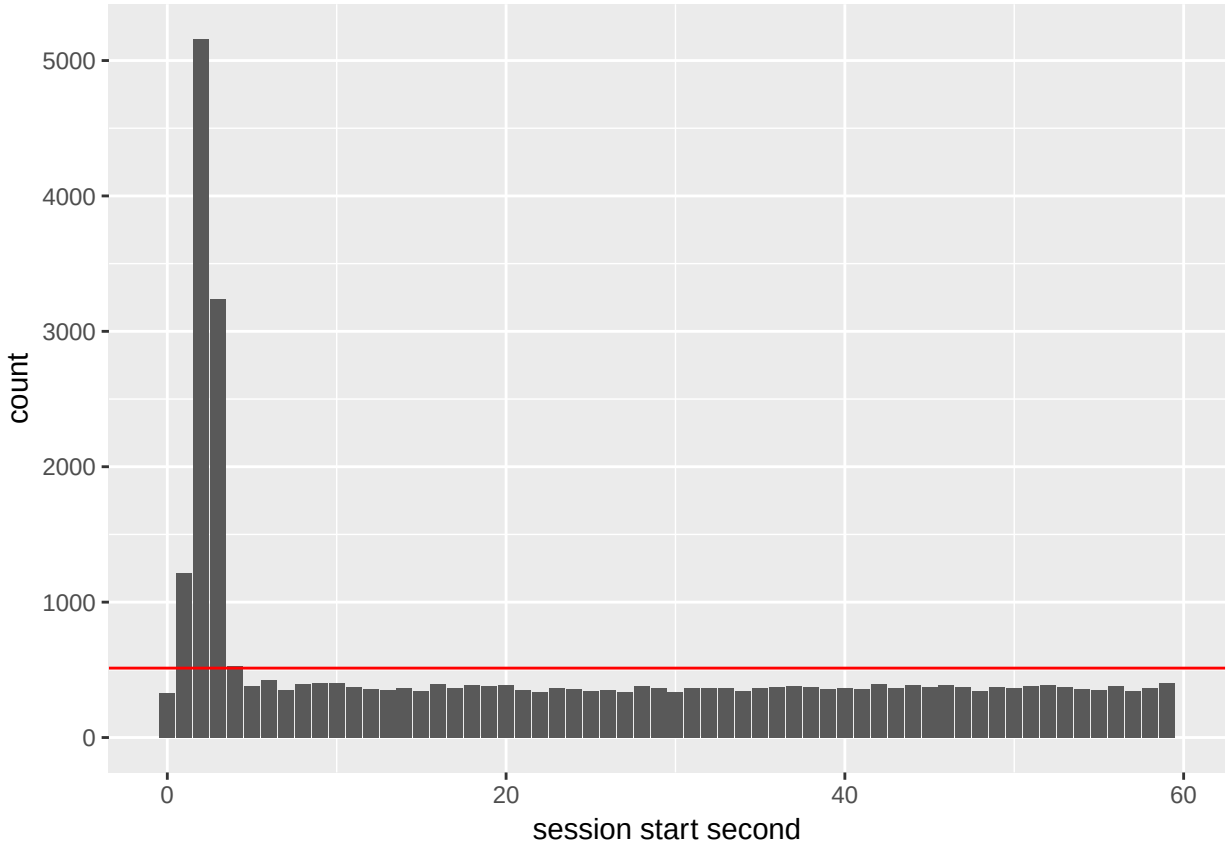
1. Exclude all sessions from the volunteer staff because lab volunteer staff members often have longer session durations and login after lab open hours.
2. Exclude sessions that have 0 or negative durations. Because they are not accurate and extremely rare (6 out of 33k sessions this semester).
3. Exclude sessions from host "blizzard" and "eruption" because they are desktops exclusive to front desk staff members and volunteer staff memebers.
4. 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

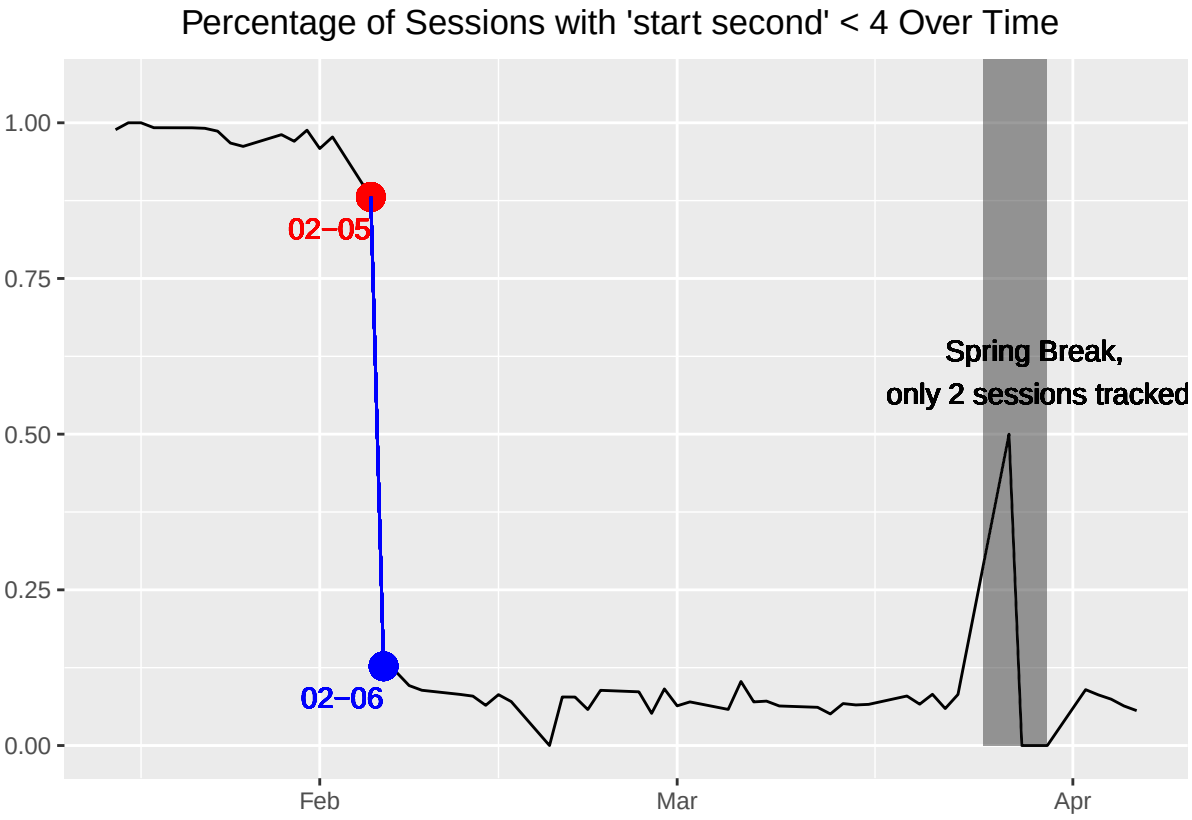
There used to be a bug in the lab's session tracking infrastructure: we could only record a session's start time at the beginning of the minute (know more about the bug from one of my previous post. For instance, if a session started at 12:**30:45**, it would appear in our data as 12:**31:03**.

See figure below about the distribution of session start seconds.



Notice due to the bug, the counts of sessions that "started" at second 1, 2, 3 are abnormally high, 2 to 10 times higher than seconds 4 to 59. Also because second 4 to 59 all have similar counts, I assume the sessions start second should be uniformly distributed (i.e each second should have 1/60 of the total counts).

To find out when sessions data was corrupted, I plots the percentage of sessions with "start second" < 4 against time.



As the graph below shows, after Feb 5th, the session tracking system returned to normal again.

Thus, for all the session before 2018-02-05, I adjust the duration by adding X seconds, where X is a random variable that is uniformly distributed from the set $\{0, 1, 2, \ldots 55 + S_{session\ start\ second}\}$. For instance, if the session's recorded start second is 3, given our tracking system was malfunctioning, the real session start time can be any time from 0 to 3 seconds, or 4 to 59 seconds from the previous minute. The actual duration for the inaccurate session data should be somewhere between 0 to 3, or 3 + 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

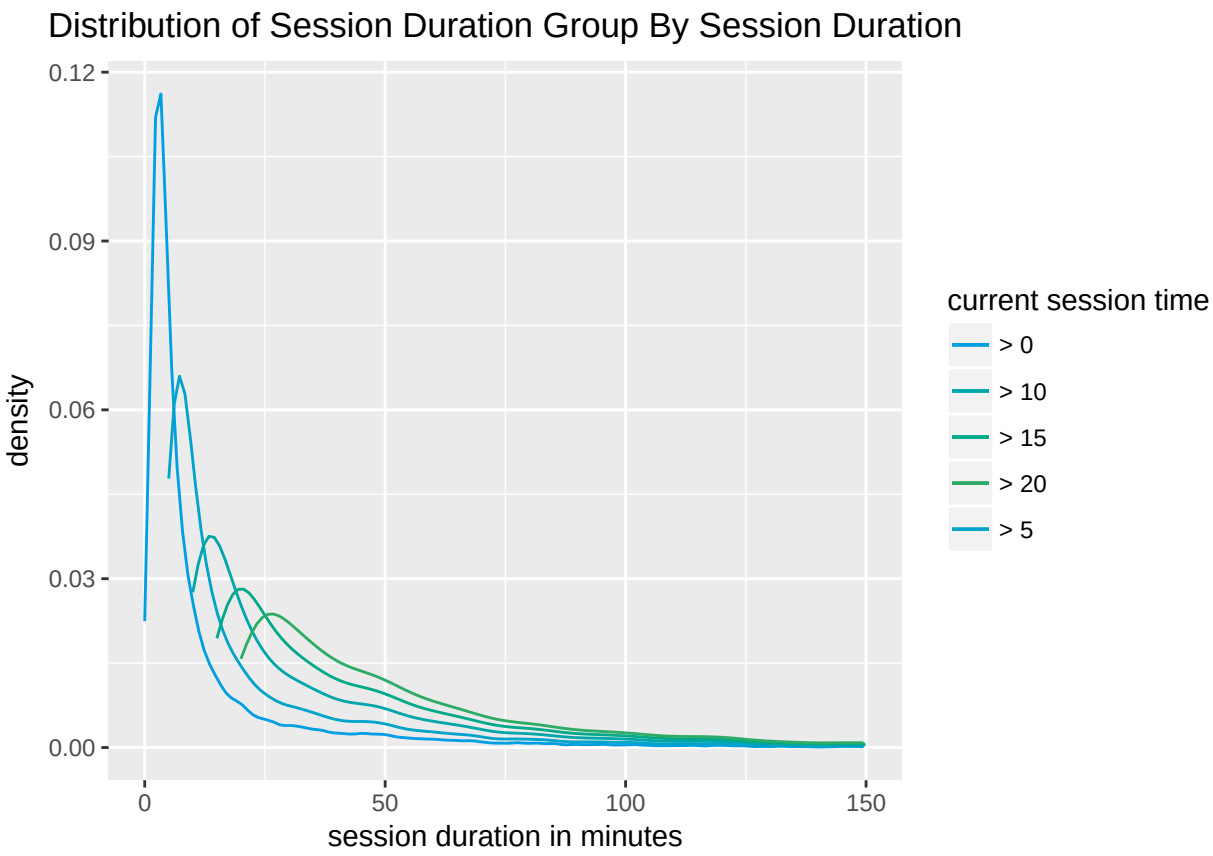
Overall each quantile and mean is increased by about 15 seconds. This is significant because the wait time of a computer when the lab is full is often under 100 seconds. Unless we discard the inaccurate data, our model to predict wait time may suffer a high percent irreducible errors.

From the summary statistics, we can also see 50% of the sessions are under 6 minutes because many users leave the lab after printing their papers and documents. However, sometimes I have to wait for a non-printing user to get the corner computer, I wonder if I can have a good guess when will that user leave given how long he/she has been there.

Basic Session Survival Analysis

Below are the distributions of the session duration when a user has been on a computer for 0, 10, 15, and 20min respectively.

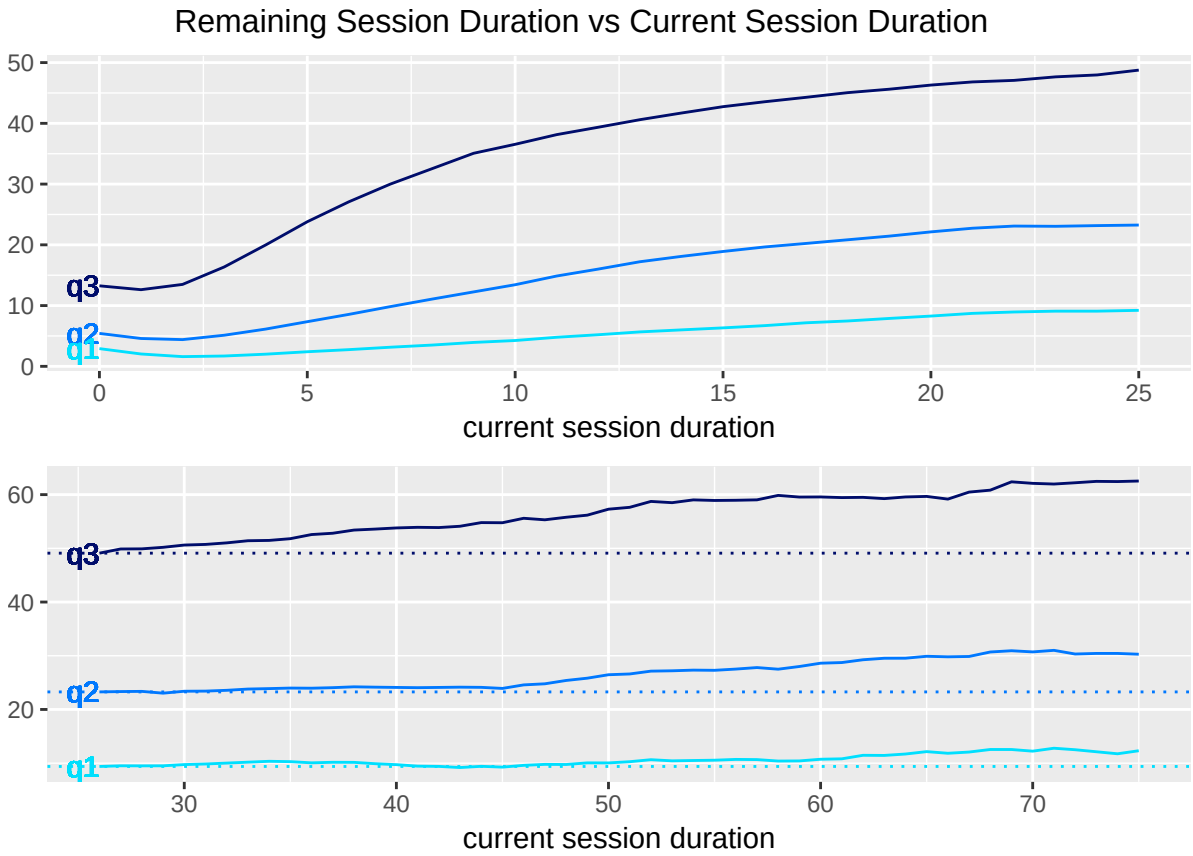
Expected Wait Time Given Current Session Duration



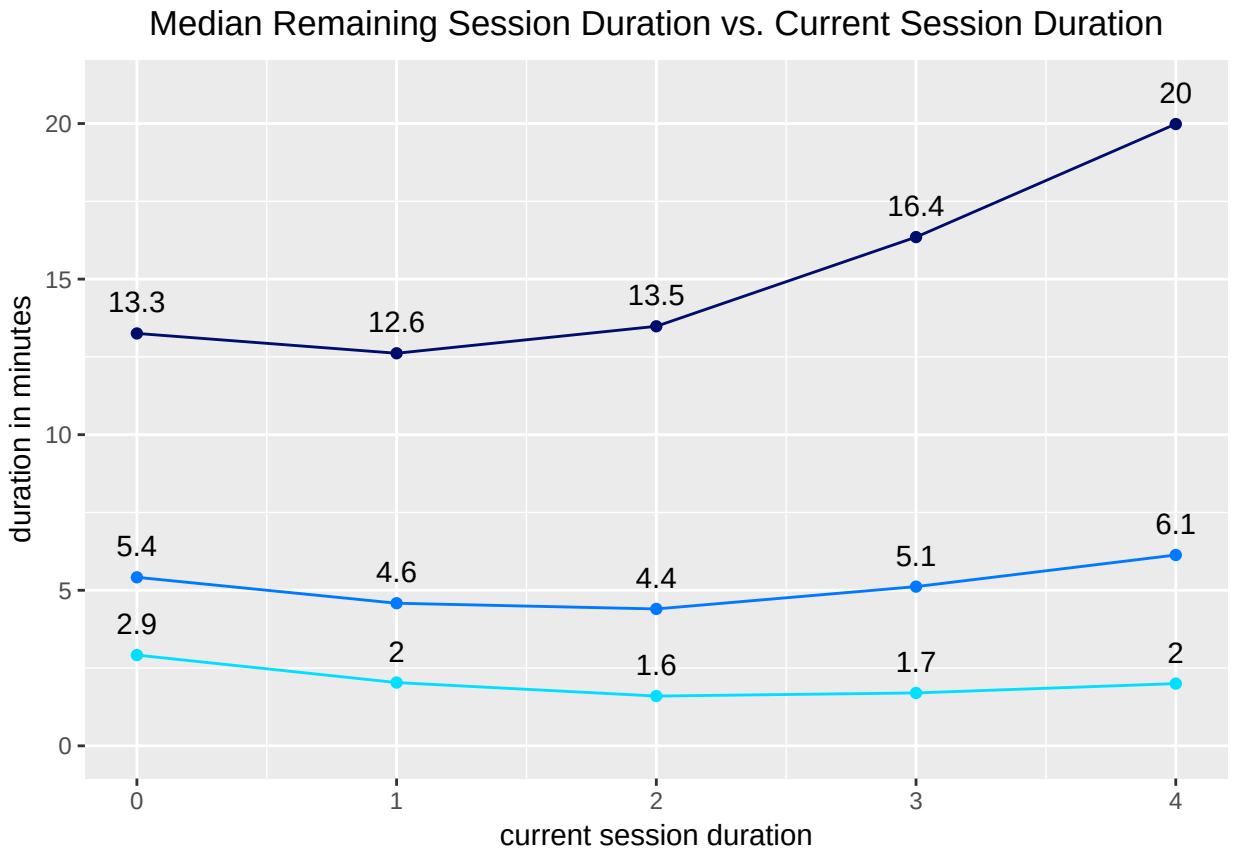
Notice nearly all distribution have a peak at the beginning. This means that at any given time, sessions have the highest probability (i.e pdf) of leaving within the first few minutes. However, the distribution is getting flatter and flatter as current session time increases. In other words, we are less certain about when a user will leave as he stays in the lab longer.

Here are the steps I use to estimate the wait time of a computer.
1. Find the how long the user has been using that computer, denote **current session time** X.
2. Select all the sessions with duration longer than X.
3. Find the 25%, 50% and 75% quantiles of those sessions.
3. Subtract the quantiles by X to get the quantiles of **remaining session time**.

Below are two graphs showing **current session time** X from 0 to 25 and 25 to 75 against **remaining session time**.



As **current session time** increases, almost all quantiles increases. This increasing pattern slows down after 25 minutes or so (check the second graph). Also, there is a small dip in the first graph during the first few points, let's zoom in the first graph and take a closer look.



We can see the **remaining session time** is lowest at the 2 min mark for both the 25% and 50% quantile. Even for the 75% quantile, the lowest waiting time is when the session is at 1 min mark, instead of 0. Thus, the waiting time conditioned on **current session duration** first experiences a small decreases, then increases rapidly until the 25min mark, then increases slowly afterward. The overall lesson is, if somebody has been on a computer for a longer time, don't expect him or her to leave any time soon.

Future Directions

In part 1, we learn that a majority of the sessions are short. But the distribution of session duration is extremely right skew with some unreasonable outlier. The maximum session length of our user this semester is 9+ hours! We also 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 25 minutes or so.

In the next part, I will examine additional variables such as **the computer used**. Students may favor computers closer to the printer; or students may avoid using double monitor computers from the misconception that those computers are for staff members only...