
Data Analysis Sprint

Andrew Green

ENGR112-0001

4/8/22

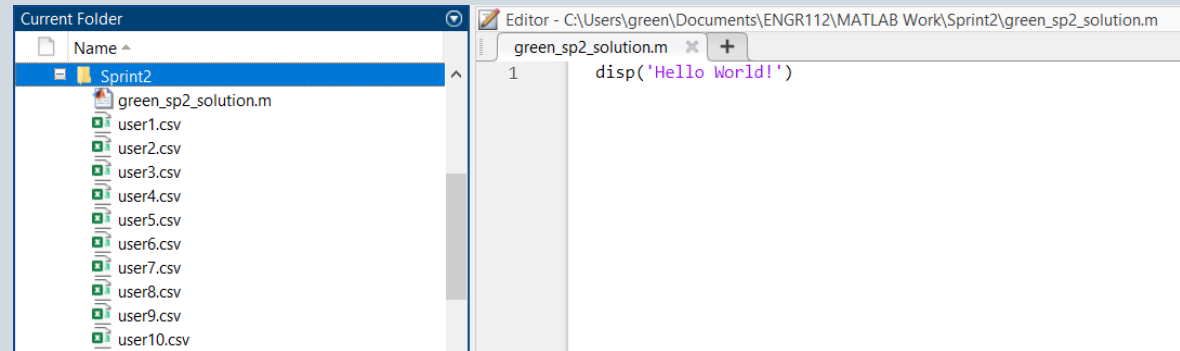
In accordance with JMU honor code policy.

Task #1

Hello World!

And

Following the honor code



Current Folder

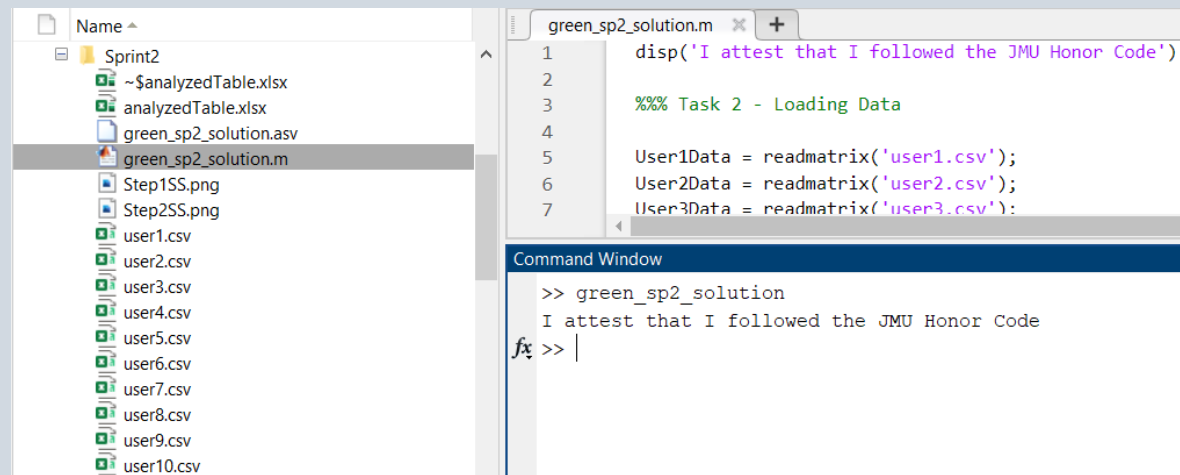
Name ^

- Sprint2
 - green_sp2_solution.m
 - user1.csv
 - user2.csv
 - user3.csv
 - user4.csv
 - user5.csv
 - user6.csv
 - user7.csv
 - user8.csv
 - user9.csv
 - user10.csv

Editor - C:\Users\green\Documents\ENGR112\MATLAB Work\Sprint2\green_sp2_solution.m

green_sp2_solution.m

```
1 disp('Hello World!')
```



Name ^

- Sprint2
 - ~\$analyzedTable.xlsx
 - analyzedTable.xlsx
 - green_sp2_solution.asv
 - green_sp2_solution.m
 - Step1SS.png
 - Step2SS.png
 - user1.csv
 - user2.csv
 - user3.csv
 - user4.csv
 - user5.csv
 - user6.csv
 - user7.csv
 - user8.csv
 - user9.csv
 - user10.csv

green_sp2_solution.m

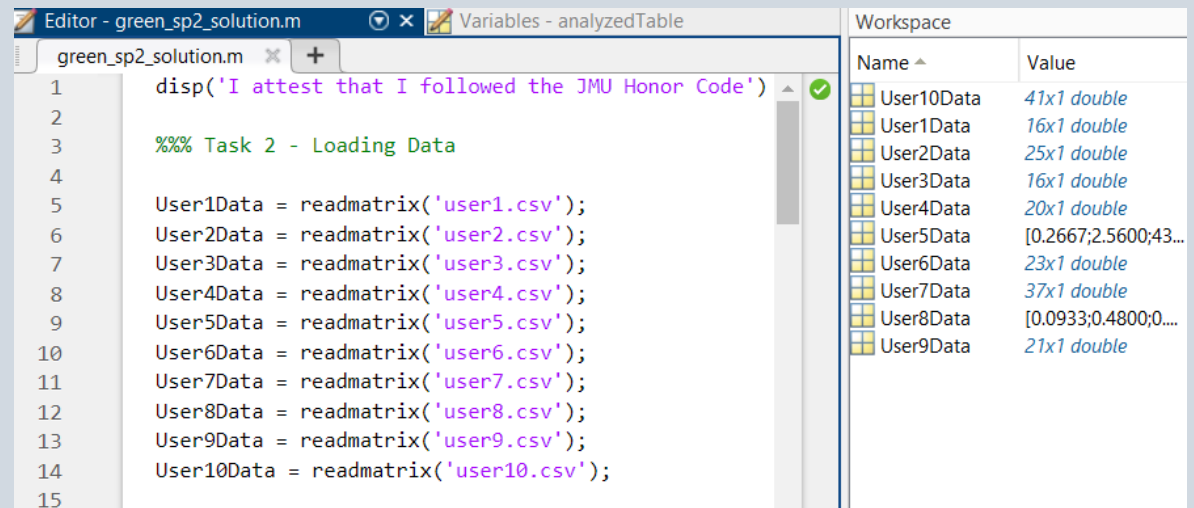
```
1 disp('I attest that I followed the JMU Honor Code')
2
3 %%% Task 2 - Loading Data
4
5 User1Data = readmatrix('user1.csv');
6 User2Data = readmatrix('user2.csv');
7 User3Data = readmatrix('user3.csv');
```

Command Window

```
>> green_sp2_solution
I attest that I followed the JMU Honor Code
fx >> |
```

Task #2

Data: Loaded



The screenshot shows the MATLAB environment. The Editor window displays a script named 'green_sp2_solution.m' with the following code:

```
1 disp('I attest that I followed the JMU Honor Code')
2
3 %% Task 2 - Loading Data
4
5 User1Data = readmatrix('user1.csv');
6 User2Data = readmatrix('user2.csv');
7 User3Data = readmatrix('user3.csv');
8 User4Data = readmatrix('user4.csv');
9 User5Data = readmatrix('user5.csv');
10 User6Data = readmatrix('user6.csv');
11 User7Data = readmatrix('user7.csv');
12 User8Data = readmatrix('user8.csv');
13 User9Data = readmatrix('user9.csv');
14 User10Data = readmatrix('user10.csv');
15
```

The Variables window shows a table of workspace variables:

Name ^	Value
User10Data	41x1 double
User1Data	16x1 double
User2Data	25x1 double
User3Data	16x1 double
User4Data	20x1 double
User5Data	[0.2667;2.5600;43...
User6Data	23x1 double
User7Data	37x1 double
User8Data	[0.0933;0.4800;0....
User9Data	21x1 double

Task #3

Table with labels

	A	B	C	D	E
1	User Name	Gap Count	Mean	Min	Max
2	User 1	16	10.21501	0.1467	67.0667
3	User 2	25	18.3456	0.2667	95.84
4	User 3	16	6.770836	0.0267	32.6667
5	User 4	20	15.51666	0.1067	93
6	User 5	6	13.08223	0.2667	43
7	User 6	23	33.40172	0.0267	226.333
8	User 7	37	23.5917	0.1333	219.333
9	User 8	6	19.51777	0.0933	99.9733
10	User 9	21	3.113653	0.0933	14.5467
11	User 10	41	14.38244	0.1067	155.8

Task #4

Good Data

Example

An example of good data is the data from User 3. Compared to all the other data sets provided except from User 9, their average gap time is much shorter. They also have a low number of gaps, and their max gap time is one of the lowest. That means that the interference caused by motion for that data set should be low, and hopefully won't affect the usability of it for detecting stress.

Task #4

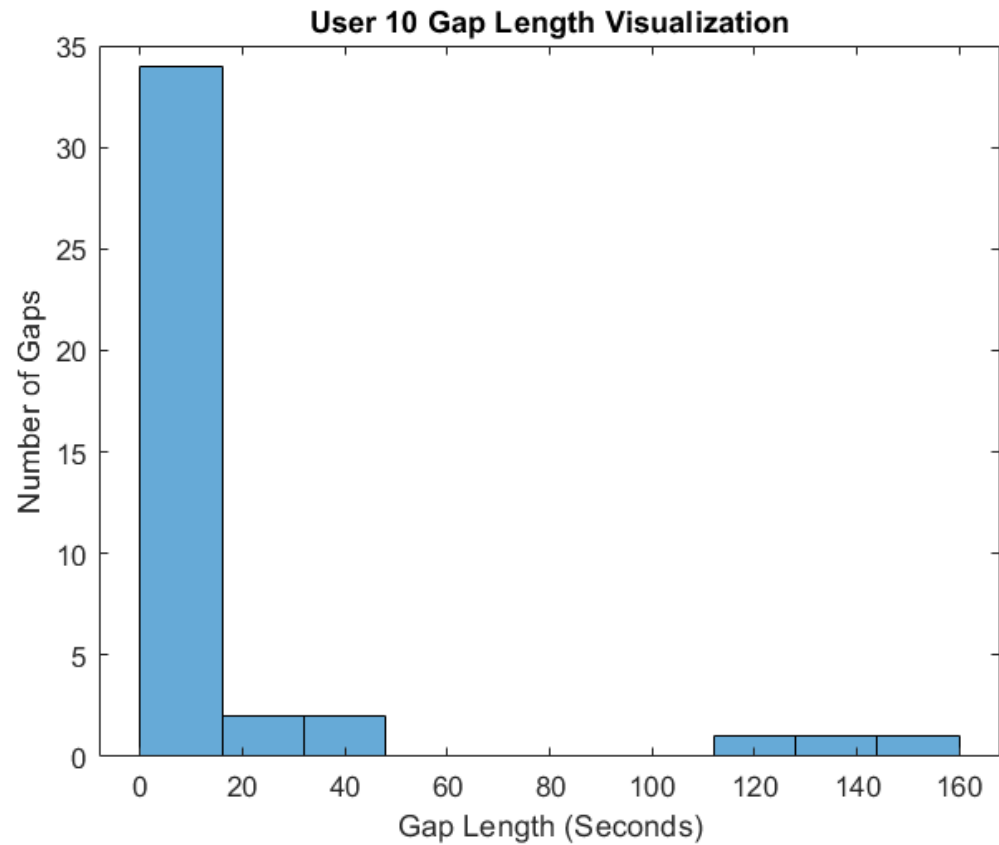
Bad Data

Example

An example of bad data is the data from User 7. The data provided from them has a massive number of gaps, the max gap length is huge, and the mean gap length is second highest in all the data provided. This means that data is likely inconsistent, will be difficult to analyze without a significant amount of editing, and could introduce a large amount of error to a larger data set if not caught.

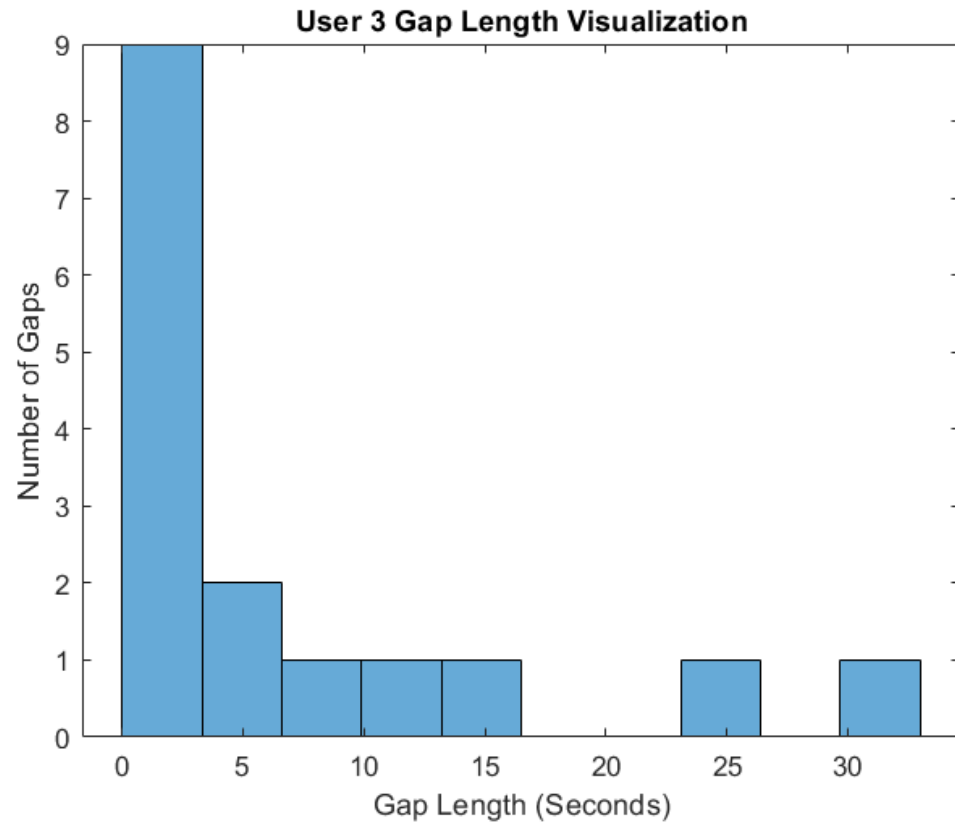
Task #5

User 10



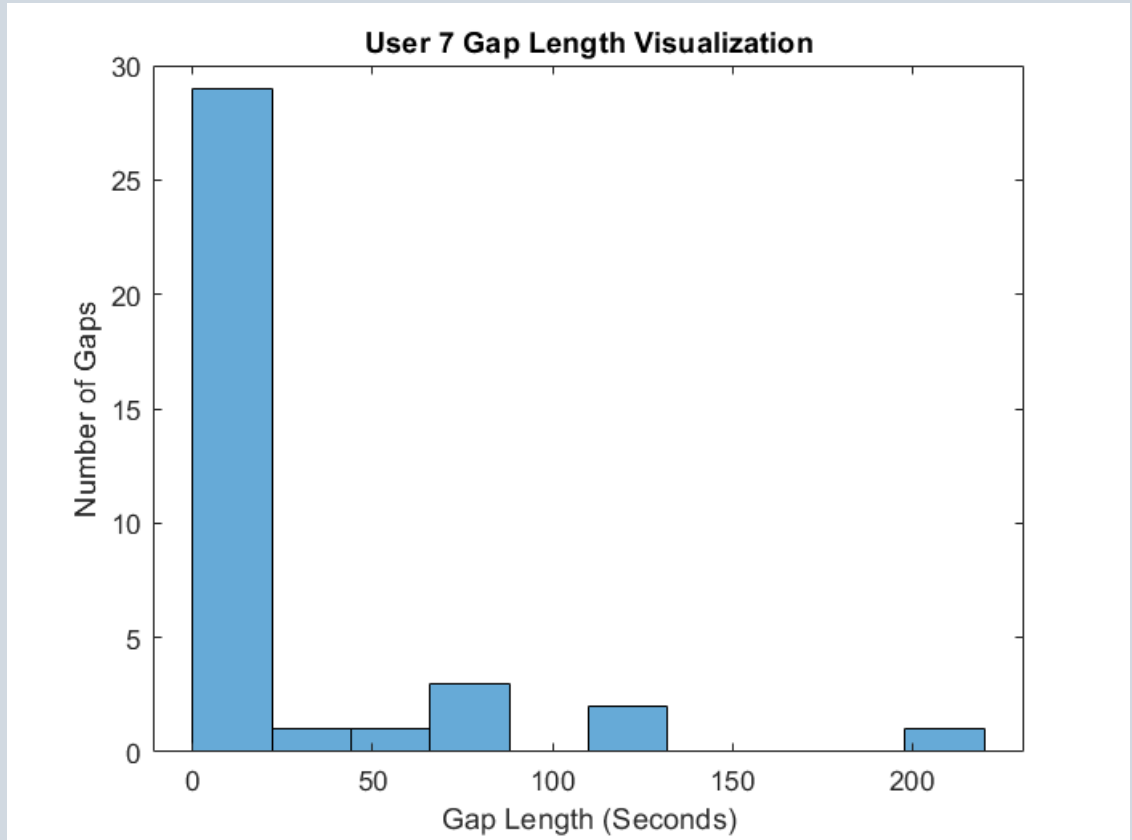
Task #5

User 3



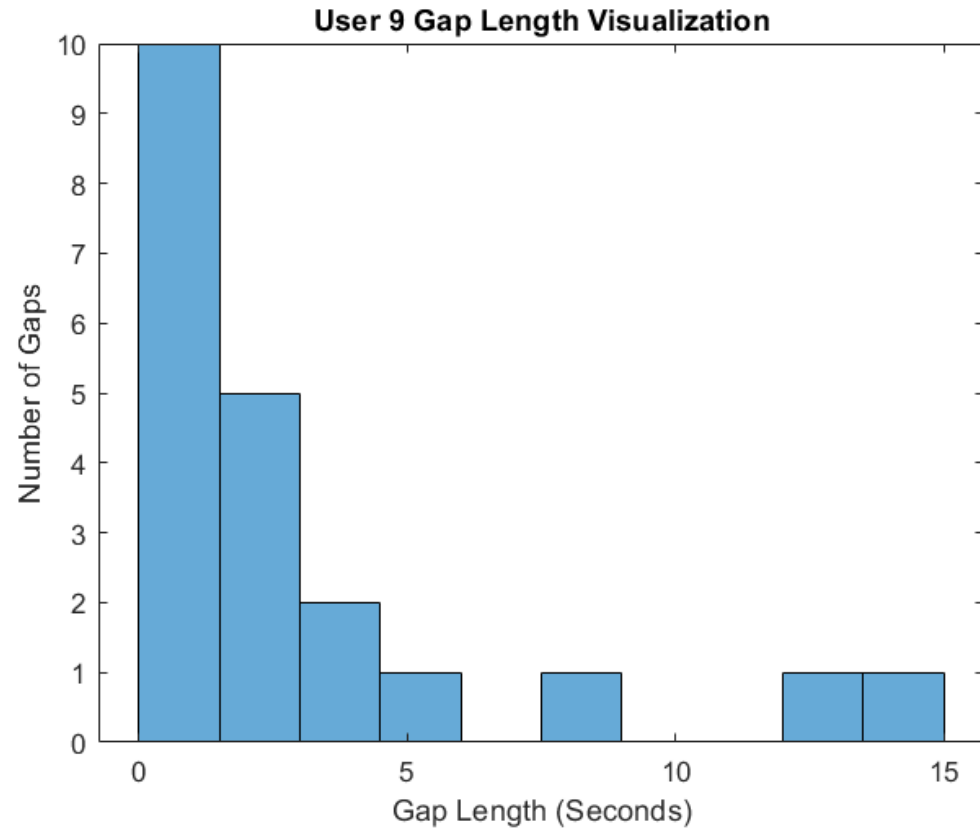
Task #5

User 7



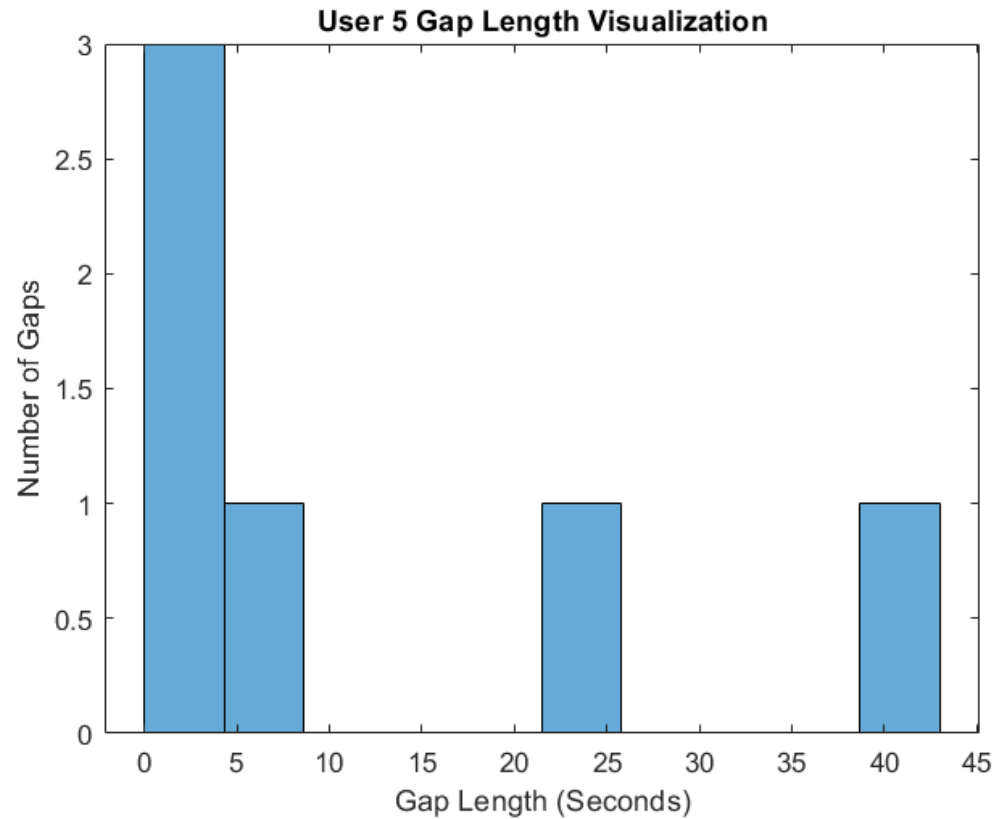
Task #5

User 9



Task #5

User 5



Task #6

Visualization Reflections

The visualization of the data didn't change my understanding of a "good" or "bad" user, but it did introduce a few questions.

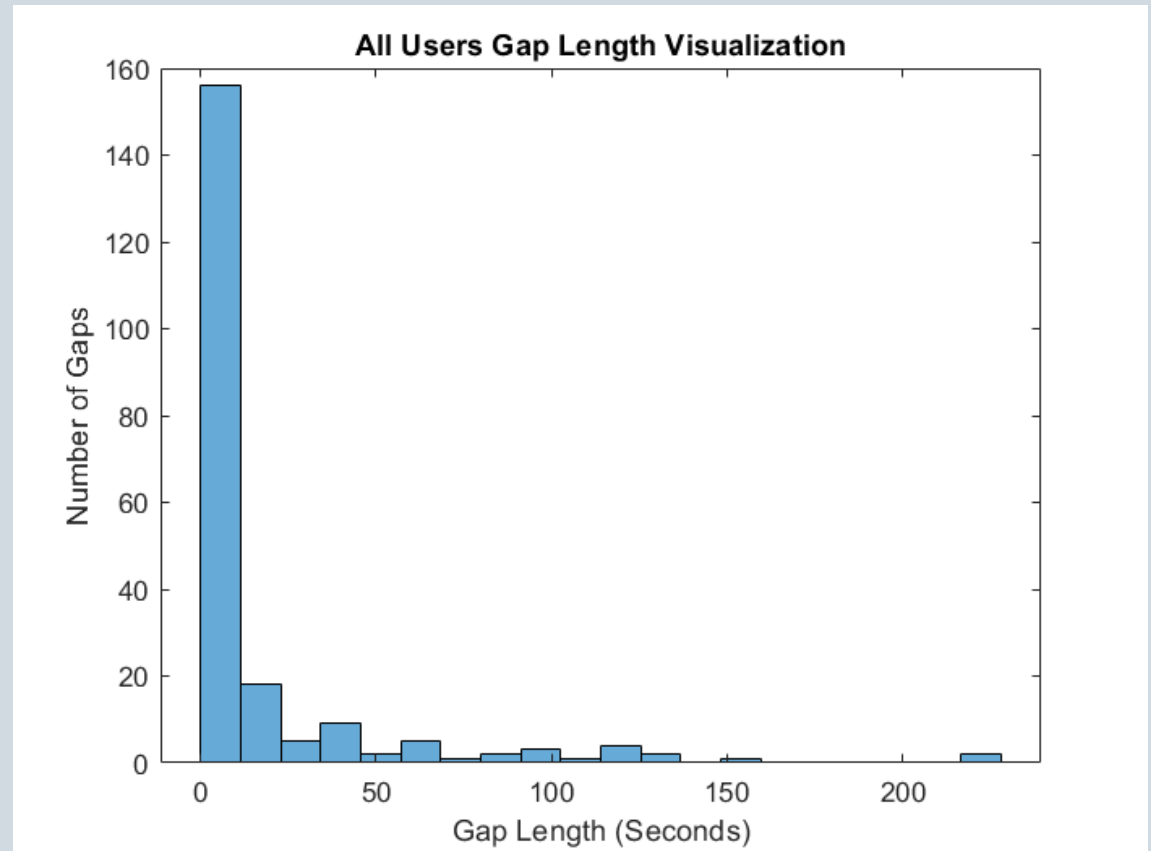
Are all the samples taken over the same time period? If so, User 5 and User 8 have the best data, because it has the fewest gaps. If the data needs to be cleaned up and have the gaps removed to be useful for the stress measuring, is it harder to remove one longer gap or multiple shorter gaps? This is important for determining which user's data is "worse" as it attaches more significance to the numbers provided.

The "good" users' data have a smaller x and y axis and the area taken up by their bars is located closer to zero on the x axis.

This visual representation of the data helps my understanding of the data. Previously, I was only thinking about the averages, which gave me a good general understanding, but lacked the granularity and detail the graphs provide.

Task #7

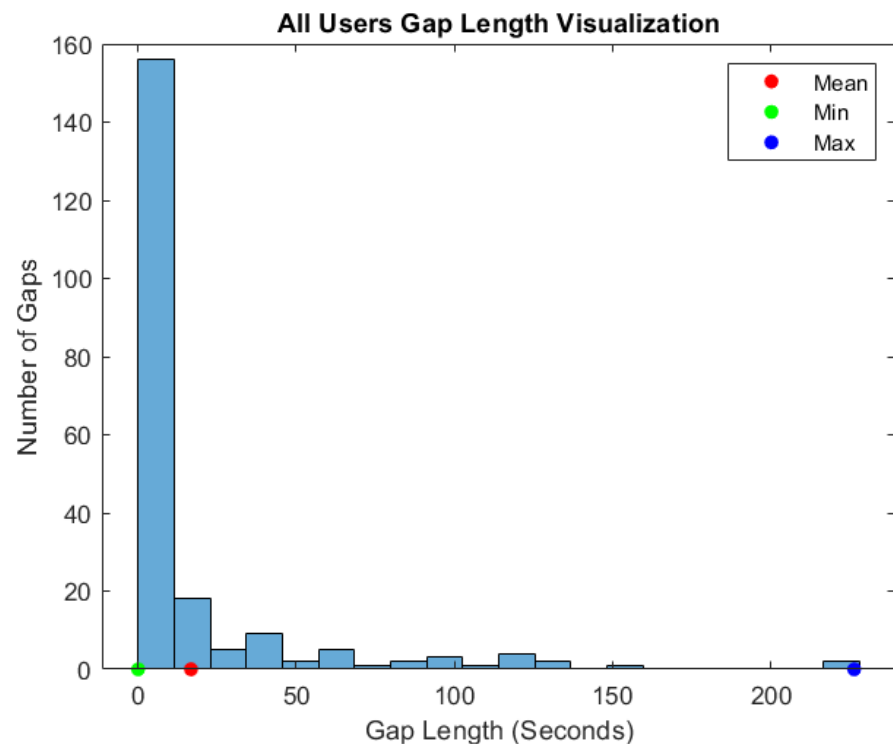
All the data in one graph



Task #8

A reflection on task #7's results.

The general distribution of the gaps shows that they are much more likely to be “short” than “long”. The minimum gap size is 0 seconds, the mean is 16.7419 seconds, and the max is 226.333 seconds. The distribution is long tailed, due to the gap lengths generally decaying exponentially as they go towards the max. I believe that in the application of monitoring stress, we should be more concerned about the more frequent, shorter gaps. This is because stress can change very rapidly, and with small gaps missing from our data, we could miss a lot of peaks or changes in stress levels that could show important stress responses.



Task #9

The answer I got for the probability of a gap greater than 60 seconds occurring is 0.135%. This seems possible, due to fitting the data to a lognormal function, which goes towards zero as time increases, but I do not think that it is correct given the data provided. Not using lognormal or logncdf functions, I would estimate the percentage of gaps over 60 seconds to be closer to 3%.

I wish I could express more confidence in my work, but I do not understand how the logncdf function works, and as a result cannot claim with any confidence that the results are what I might expect. I tried to better understand it, but all the documentation I could find assumed a much stronger understanding of CDF functions than I have.

Task #10

If the system is down, it would not be providing usable data, so it would not be usable in practice.

~~For any device in the medical field, downtime should be less than 0.001% of operating time, in my opinion.~~

During the MATLAB modules in ENGR112, I learned a lot about data manipulation and become much more comfortable with programming in MATLAB. My gains stemmed mostly from the data collection sprint, as I pushed myself to use more complex data and ask questions that demanded much more work in order to get answers from the data. I was really excited about the graphs and such that I pulled from the stream of numbers that the accelerometer gave us to work with. My biggest struggle was using logncdf on this assignment.

I felt that this assignment was fairly easy till step 9, when it became painful, then it was over. The information is somewhat interesting, but the work was pretty repetitive and I don't get much from it.

The tech described in the assignment would be very interesting if applied to classroom environments in my opinion. Being able to see stress levels of different students during tests, or stress induced by different topics, would be a very interesting bit of data to read about.