Working from the Home Environment & Well-Being Study Data

Final Report Draft 1

Spring 2023 CS506 Data Science

Team 2

Shang Lyu (lyushang@bu.edu)

Jeya Varshini Bharath (jvbharat@bu.edu)

Navya Jain (jnavya@bu.edu)

Dhun Jayswal (jayswald@bu.edu)

I. Introduction

In recent years, due to the global pandemic, the shift towards remote work has significantly impacted the way employees engage with their work environment and manage their well-being. To better understand the physical, cognitive, and mental effects of a remote work setting, our team analyzed a 6-month study utilizing the ecological momentary assessment approach. The Working from the Home Environment & Well-Being Study, carried out during the Spring 2023 semester by the CS506 Data Science team, assessed the well-being of 70 participants working remotely across various industries. As remote work has become more prevalent, there has been increasing interest in understanding the impact of this work arrangement on employee well-being. Our study aims to contribute to this understanding by examining the experiences of a diverse group of remote workers over a 6-month period.

One of the unique aspects of our study is the use of ecological momentary assessment, which allowed us to collect data on participants' experiences in real time. By prompting participants to provide information about their location, musculoskeletal discomfort, and break frequency throughout the day, we were able to capture a more detailed picture of their daily experiences than would have been possible with more traditional survey methods. Participants were provided with a Garmin watch that prompted them to share information such as their current location, musculoskeletal discomfort, and the number of breaks taken three times a day. Weekly assessments were carried out using the E-Work and the Flourishing scale surveys, and a monthly computer workstation survey was completed to gauge ergonomic factors. Our aim is to explore the initial insights with the 3-month datasets and establish a benchmark for comparison between that and 6-month datasets. The 3-month and 6-month datasets contains not only the garmin data collected via their garmin watches, but also their responses to certain survey questions.

Through a series of hypotheses, our analysis will focus on factors such as age, break frequency, industry, stress levels, and ergonomic training, and their potential relationships with various aspects of well-being, including financial stability, productivity, mental health, and musculoskeletal discomfort. By examining individual and aggregate data, we aim to identify patterns and trends that may provide valuable insights into the impact of remote work on well-being.

In addition to answering key questions from the client, our final report will include a range of visualizations and statistical analyses to illustrate our findings. As this research is funded by a grant from the Office Ergonomics Research Committee (OERC), the resulting data and insights will be used for presentation at conferences and publication in research manuscripts.

The aim of this 6-month investigation is to examine the impact that working remotely has on an individual's physical, cognitive, and psychological well-being. Upon completion, this project will provide conclusions and recommendations for employees as well as companies to guide their work-from-home or hybrid policies. We look forward to sharing our findings with the wider research community and contributing to the development of evidence-based policies and practices for remote work.

II. Base Analysis

Dataset

The following data was analyzed for the scope of this project:

- Garmin data: These were in the form of 64 csv files containing the data collected by the garmin watches given to the participants. Each csv file corresponded to one participant's data.
- Participants' physical location data
- Participants' musculoskeletal discomfort data
- Responses to the following surveys were also recorded in over 6 different csv files:
    a. Computer workstation checklist
    b. E-Work Life scale
    c. Flourshing scale survey
    d. Visual Analog Scale

Initial Hypothesis

The base analysis was conducted on a series of hypotheses. The data provided was used to test the following hypotheses:

Hypothesis A

Participants' age will negatively correlate with financial and material stability (the last two questions on the Flourishing Scale).

Background needed:

For this hypothesis, it is important to know the last 2 questions from the Flourishing Scale:

Domain 6: Financial and Material Stability:

1. How often do you worry about being able to meet normal monthly living expenses?
2. How often do you worry about safety, food, or housing?

The participants are supposed to select a ranking from number 0 to 10 which has the following significance:

1. 0 = Worry All of the Time
2. 10 = Do Not Ever Worry

Data used:

The flourishing scale was filled by participants weekly on Friday. The data is recorded in 'FridayAM.csv' file.

Next, we need the participants' ages. This is taken from the 'Demographic.csv' file.

Data preprocessing:

FridayAM has the following columns:

```
[4] print(flourishing.columns)

    Index(['mbl_cod', 'rsp_id', 'ts', 'local_time', 'LOCATION_AM',
           'DISCOMFORT_SLIDER', 'LIFE_SATISFACTION', 'HAPPINESS',
           'PHYSICAL_HEALTH', 'MENTAL_HEALTH', 'WORTHWHILE', 'PURPOSE',
           'PROMOTE_GOOD', 'DELAYED_HAPPINESS', 'CONTENT_RELATIONSHIPS',
           'SATISFYING_RELATIONSHIPS', 'LIVING_EXPENSES', 'FOOD_HOUSING', 'STRESS',
           'PULSE_OX', 'HEART_RATE', 'RESPIRATION', 'BODY_BATTERY', 'STEPS',
           'CALORIES', 'FLOORS', 'INTENSITY_MINUTES', 'AVG_AMP', 'VOX_ACTV'],
          dtype='object')
```

Since we only need to focus on financial stability, let us store 'Living_expenses' and 'Food_housing' in a separate dataframe. We also need to retain rsp_id, ts and local_time since it contains participant's ID and timestamp's information. We will drop everything else.

Then, I took an average along the columns of Living_expenses and Food_housing and stored it with the respective mobile ID and age associated with the mobile ID. Now, I had a dataframe with the following columns - mid, age, expenses and food_housing.

Analysis:

Step 1 - Finding the average 'living_expenses' and ''food_housing' flourishing scores based on the mobile ID of the participants.

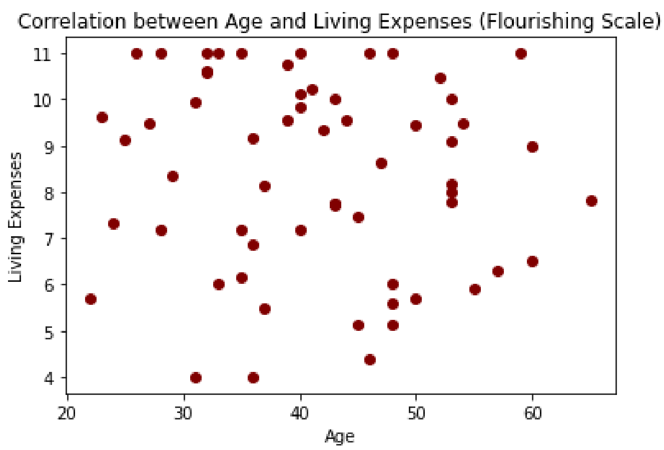
Step 2 - Merging this dataframe containing the averages with the ages on mobile ID.

Step 3 - Calculating the correlation between ages, living expenses, and food and housing expenses flourishing scores

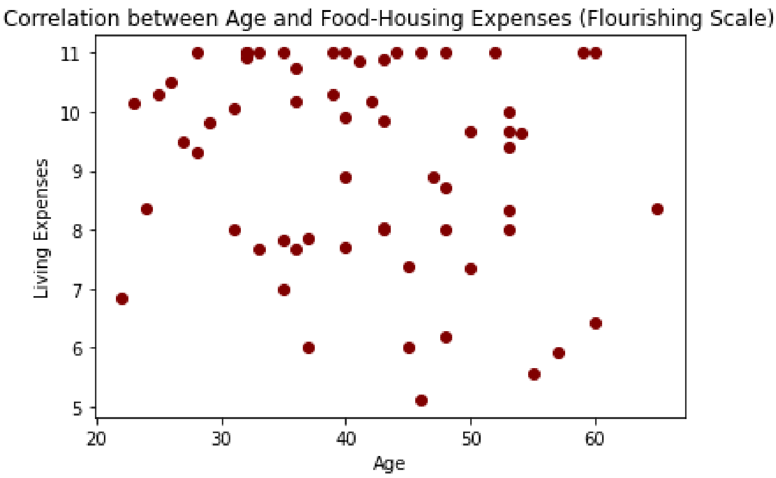
Formula used: Pearson correlation coefficient given by:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Results:



Correlation between Age and Living Expenses (flourishing scale) = -0.096



Correlation between Age and Food-housing expenses (flourishing scale) = -0.194

Hypothesis B

Participants who take an average of 4 breaks per day will positively correlate with productivity scores in the E-Work Life Scale (questions 16-20) and report lower discomfort at one month compared to six-month data.

Background needed:

This code analyzes the data related to the work-life balance of employees in a company.

Data used:

The code reads a CSV file ('FridayPM.csv') containing the data, converts the 'local_time' column to a datetime format, and filters the data for the last month.

Data preprocessing:

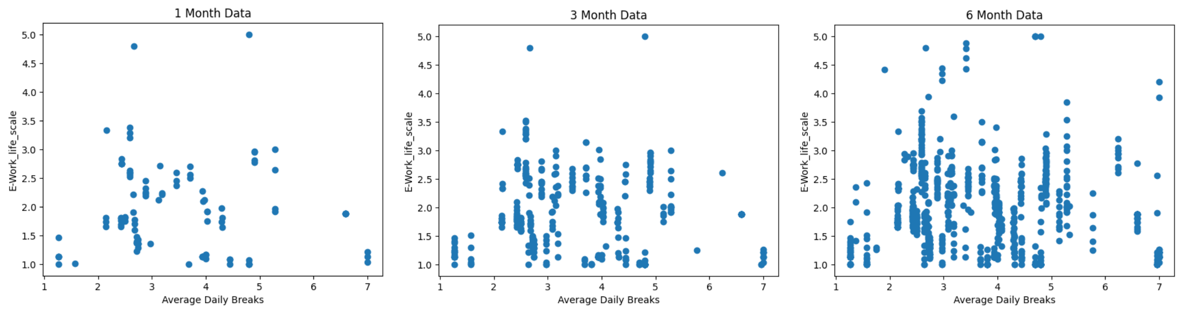
It drops rows with missing values and calculates scores for different categories (Trust, Flexibility, Work_life, and Productivity).

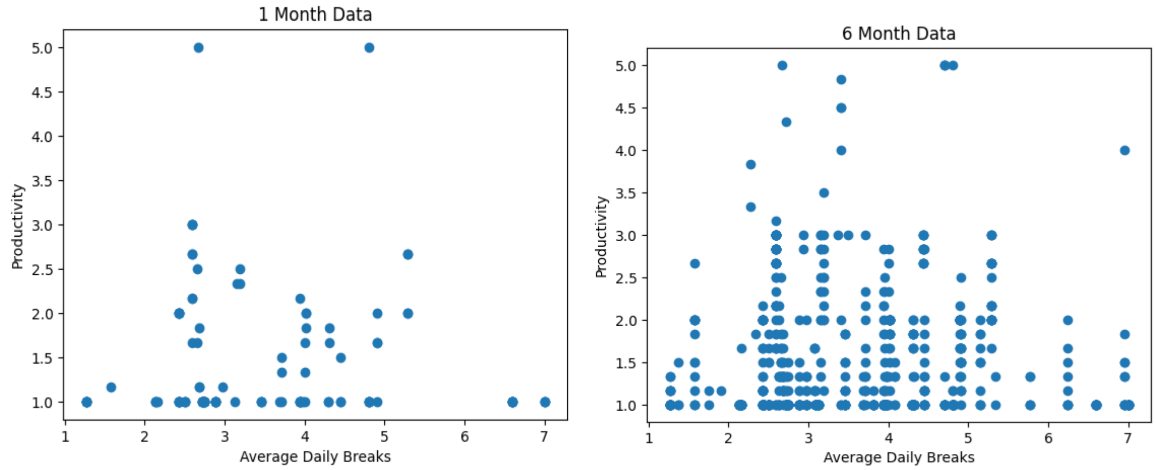
The code then calculates the E-Work_life_scale based on the weighted scores of the categories. It merges the E-Work_life_scale data with the average breaks data for each employee and creates a scatter plot of the E-Work_life_scale vs. Daily Breaks. Finally, it prints the mean E-Work_life_scale of breaks for the last month.

Analysis:

The purpose of this code is to provide insights into the work-life balance of employees and the impact of daily breaks on their work-life balance. The scatter plot and the mean E-Work_life_scale provide useful information for management to improve the work environment and enhance the well-being of employees.

Results:





Correlation:

Participants who take an average of 4 breaks per day will positively correlate with productivity scores in the E-Work Life Scale

Formula used:

['E-Work_life_scale'] =

(['Trust'] * 0.4) + (['Flexibility'] * 0.3) + (['Work_life'] * 0.2) + (['Productivity'] * 0.1)

Hypothesis C

Participants working in healthcare will have lower mental health scores on the Flourishing Scale than those working in other industries.

Background needed:

Demographic datasets provide information about each participant working in which industry. Mental health scores are recorded in the Friday AM dataset as MENTAL_HEALTH column, with the question of

How would you rate your overall mental health?

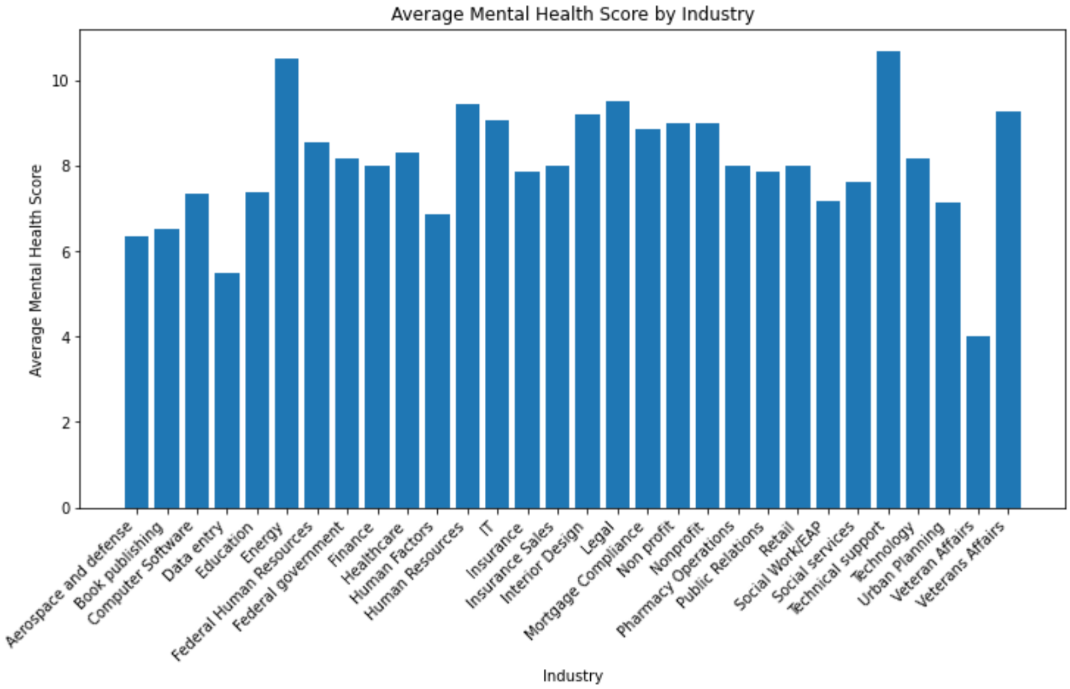
0 = Poor, 10 = Excellent

Data used: FridayPM 3 Month; Demographic

Data preprocessing:

The code merges two datasets into one DataFrame by participant id, and then selects the relevant columns MENTAL_HEALTH and INDUSTRY, dropping the empty rows. Then use the method 'groupby' in pandas and calculate the mean of the mental health score by industry.

Analysis:



Here we calculate the average mental health score grouping by the industry and plot a bar chart.

Based on the data analysis performed on the dataset, the hypothesis that participants working in healthcare will have lower mental health scores on the Flourishing Scale than those working in other industries has been disproven.

The mental health scores of participants in healthcare were compared to those in other industries by calculating the average mental health score for each industry. The data was grouped by industry, and the mean mental health score was calculated for each group. The results showed that the average mental health score for participants in healthcare was not significantly different from the average mental health score of participants in other industries. An interesting observation on the other hand, shows that participants of Veteran Affairs have lower mental health scores overall on average.

Therefore, it can be concluded that there is no evidence to support the hypothesis that participants working in healthcare will have lower mental health scores on the Flourishing Scale than those working in other industries. This finding suggests that healthcare workers may not be more susceptible to mental health issues than those working in other industries.

Results:

Participants working in healthcare do not have particularly lower mental health scores on the Flourishing Scale than those working in other industries.

Hypothesis D

Participants' stress algorithm will be inversely correlated to their number of breaks.

Background needed:

The code is analyzing data related to participants' stress algorithm and their number of breaks in the last one month, in order to test the hypothesis that the two variables are inversely correlated.

Data used:

The code reads a CSV file 'FridayPM.csv' containing the data, filters the data for the last one month.

Data preprocessing:

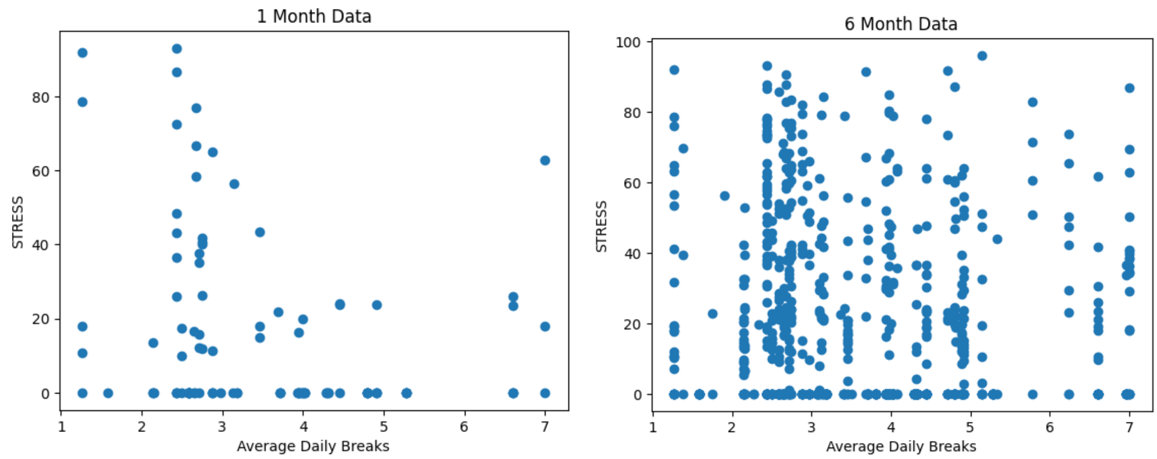
It drops rows with missing values, calculates the mean stress score for each participant, and creates a new column 'STRESS' in the dataframe with these scores. It also merges the resulting dataframe with another dataframe 'df_avgbreaks' containing average daily breaks taken by each participant.

Analysis:

Next, the code creates a scatter plot of 'STRESS' vs. 'DAILY_BREAKS', where the x-axis represents the average daily breaks taken by each participant and the y-axis represents their stress scores. It also calculates the mean stress score of breaks and prints it.

The analysis in the code is aimed at testing the hypothesis that the stress algorithm of participants is inversely correlated to their number of breaks.

Results:



Correlation:

Participants' stress algorithm will be inversely correlated to their number of breaks.

Hypothesis E:

Based on question #15 in the Computer Workstation Checklist (with 4 responses regarding ergonomics training), participants with lower scores will report less pain at 6-months.

Background needed:

The question #15 in the Computer Workstation Checklist is designed as:

15. Are workers trained in the following:

- proper postures? Yes No
- proper work methods? Yes No
- recognizing signs and symptoms of potential WMSD problems? Yes No
- when and how to adjust their workstations to avoid musculoskeletal discomfort? Yes No

The result values recorded in datasets are set as Yes 1, No 2.

The corresponding columns can be found in computer workstations datasets, and have the column names:

proper postures: POSTURE_TRAINING

proper work methods: METHODS_TRAINING

WMSD: WMSD_SIGNS

adjust workstation: WORKSTATION_ADJUSTMENT

The level of pain is recorded as the PHYSICAL_HEALTH column in Friday AM datasets. It is related to the question

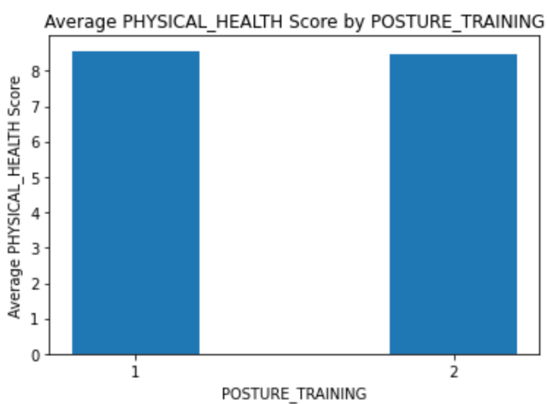
In general, how would you rate your physical health? 0 = Poor, 10 = Excellent

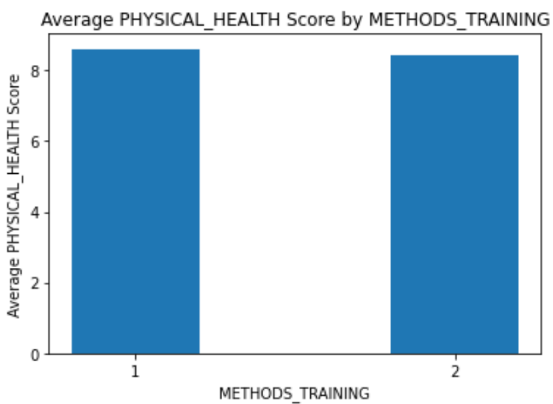
Data used: Computer Workstations 6 Month; Friday AM 6 Month

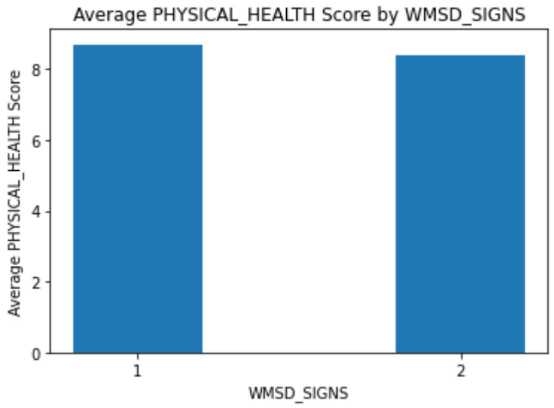
Data preprocessing:

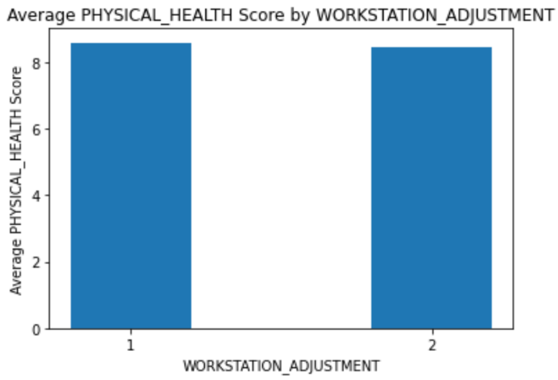
The code merges two datasets into one DataFrame by participant id, and then selects the relevant columns, dropping the empty rows. Then first calculates the average level of pain with respect to each area, and then calculates the correlation coefficient for each of them.

Analysis:










```
Correlations:
POSTURE_TRAINING          -0.027252
METHODS_TRAINING          -0.054789
WMSD_SIGNS                -0.090643
WORKSTATION_ADJUSTMENT    -0.045522
PHYSICAL_HEALTH            1.000000
Name: PHYSICAL_HEALTH, dtype: float64
```

Here in the code we average the score for each response and group by the response type (1 & 2), then we use the Pearson correlation coefficient formula between two variables X and Y with sample size n is:

$$\mathbf{r = (\Sigma(xi - \bar{x})(yi - \bar{y})) / (sqrt(\Sigma(xi - \bar{x})^\wedge 2) * sqrt(\Sigma(yi - \bar{y})^\wedge 2))}$$

where xi and yi are the individual data points, $\bar{x}$ and $\bar{y}$ are the sample means, and sqrt is the square root function. We apply this formula to each response with the physical health score.

Based on the results above, we can see in the graph that the average physical health score of participants reporting 1 is slightly higher than that of participants reporting 2 in all the responses. To clarify the terms in the hypothesis, less pain indicates a higher score in physical health. Therefore , we can conclude that the hypothesis that "Based on question #15 in the Computer Workstation Checklist (with 4 responses regarding ergonomics training), participants with lower scores will report less pain at 6-months" is supported by the data. In fact, our analysis shows that the four responses in question (POSTURE_TRAINING, METHODS_TRAINING, WMSD_SIGNS, and WORKSTATION_ADJUSTMENT) have a negative correlation with the health score, with the correlation coefficients ranging from -0.027 to -0.091. This suggests that less scores on these responses(e.g., 1, which is the “YES” option, with more training) are associated with higher physical health outcomes, indicating less pain at 6-months as hypothesized.

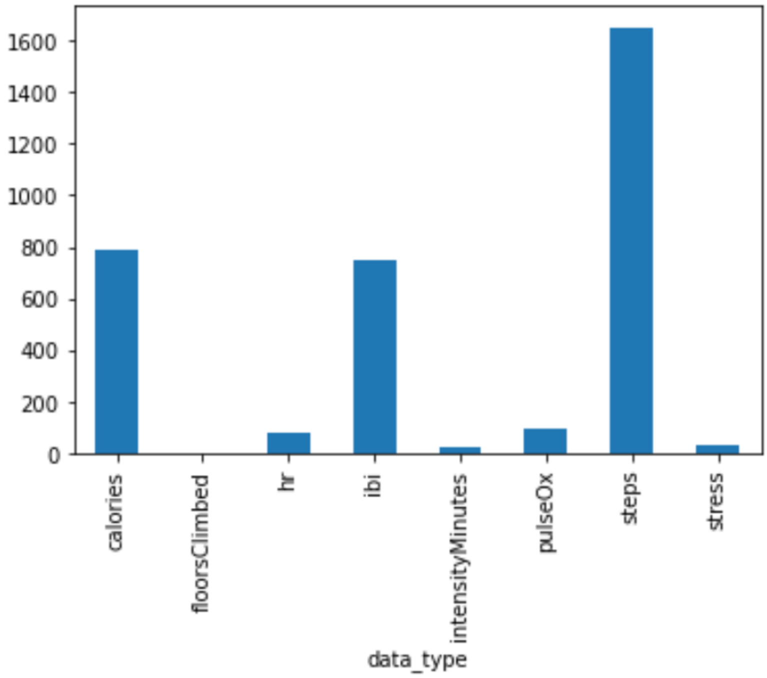
Results:

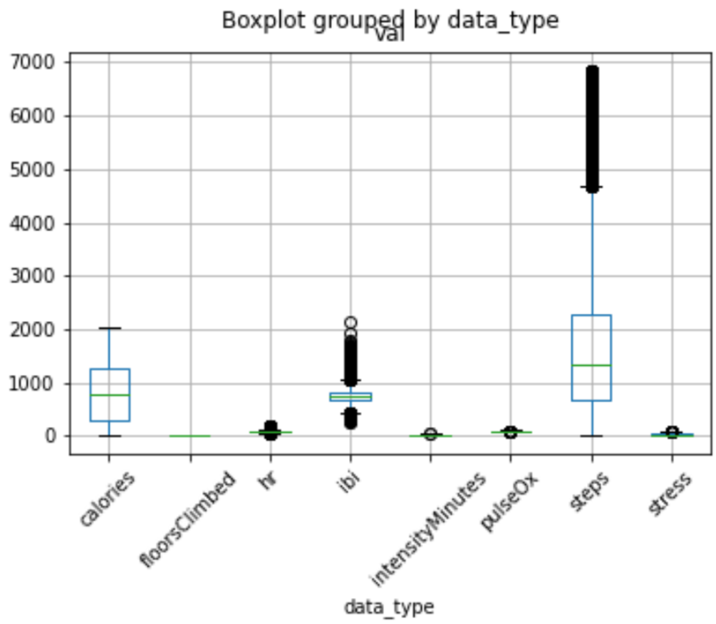
Based on question #15 in the Computer Workstation Checklist (with 4 responses regarding ergonomics training), participants with lower scores will report less pain at 6-months.

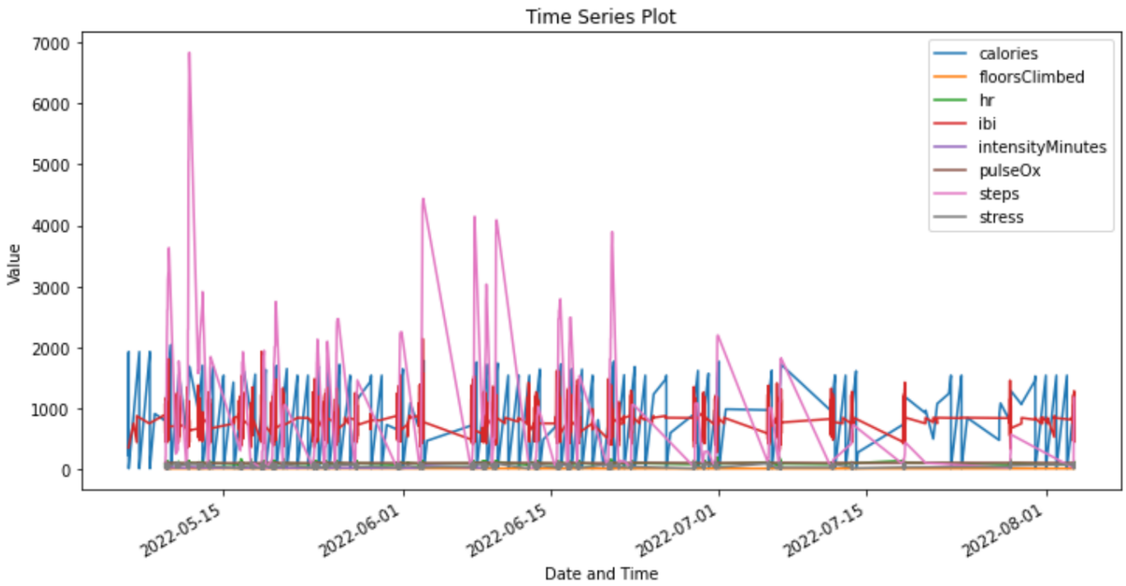
A first look at the Garmin Data

Datasets Used: garmin.11822993 2; garmin.17180706

Analysis:







This is our first glance at the Garmin data, and it only uses two data files to give a quick demo of what could be inside the Garmin datasets.

The CSV files have the following columns:

ts: a timestamp of the data

dte_tme: the date and time of the data

rsp_id: the participant ID

data_type: what type of data is recorded (e.g. heart rate, steps taken, distance traveled)

val: the value of the recorded data type

From our understanding, it records a specific data type with value at a given time with the participant id. We use two methods to analyze the data:

The first analysis uses the groupby() method of the DataFrame to group data by 'data_type'. The mean, median, and standard deviation of the 'val' column for each group are then calculated and visualized through bar and box plots.

The second analysis focuses on changes in data types over date and time.

In summary, the code analyzes and visualizes data from Garmin CSV files, including data type groups, data type changes over date and time, and mean/median/std of the 'val' column for each group. The plots created provide valuable insights into the data, helping to better understand the patterns and trends in the data.

Analysis on the Garmin Data

The Garmin data has the following structure:

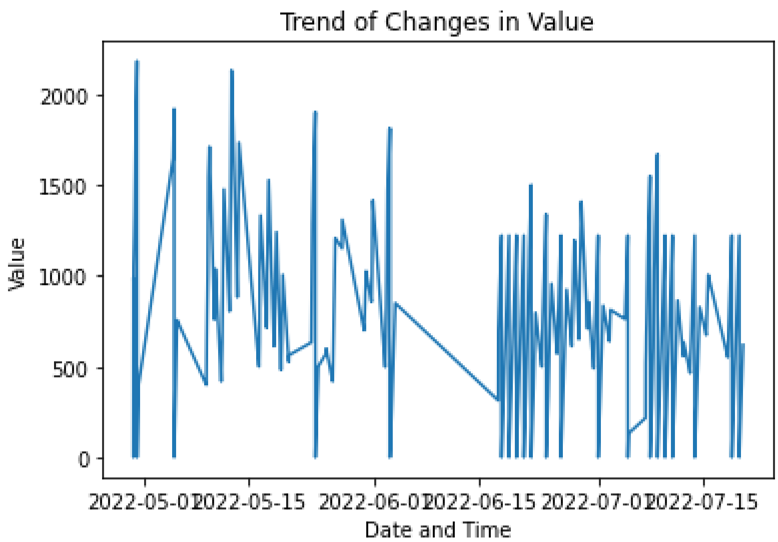
	ts	dte_tme	rsp_id	data_type	val
0	1.651253e+09	2022-04-29 12:20:01	32680	calories	988
1	1.651253e+09	2022-04-29 12:21:01	32680	calories	989
2	1.651253e+09	2022-04-29 12:22:21	32680	hr	0
3	1.651253e+09	2022-04-29 12:22:22	32680	steps	0
4	1.651253e+09	2022-04-29 12:22:22	32680	calories	0

It contains the timestamps in a day over a period of 6 months for all the participants in over 60 different csv files.

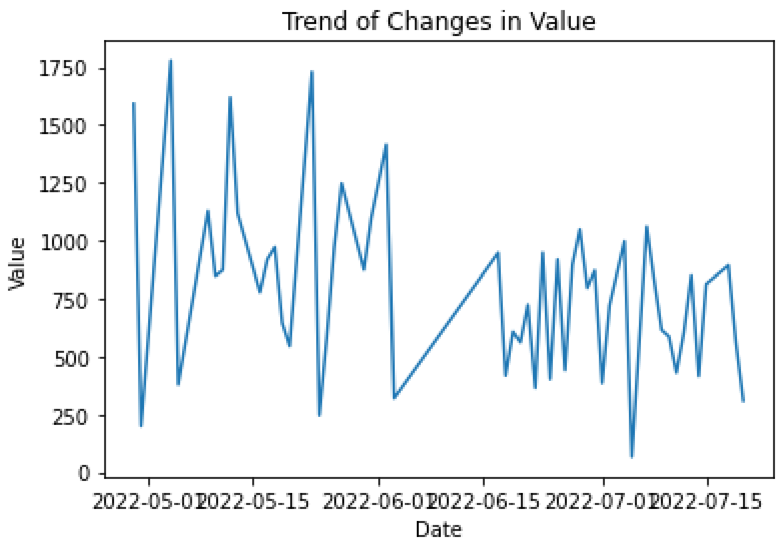
It records ['calories', 'hr', 'steps', 'floorsClimbed', 'intensityMinutes', 'pulseOx', 'ibi', 'stress'].

For the scope of this analysis, I extracted the calories, grouped them by days and then months and correlated them with the physical health of a person.

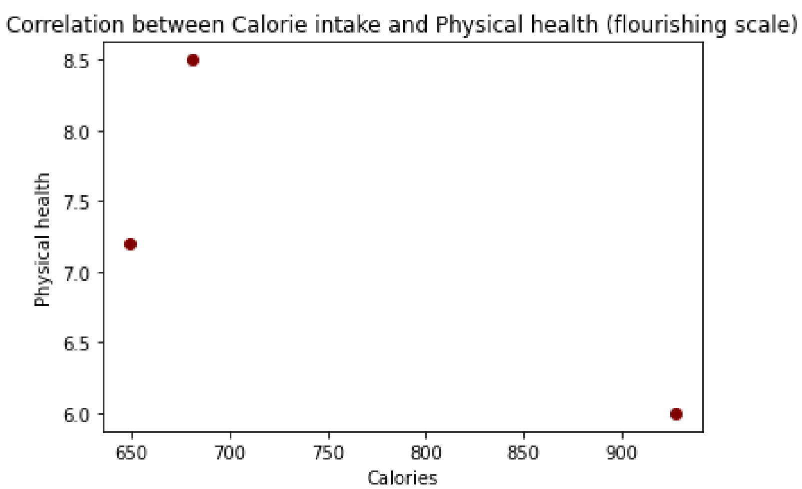
I first analyzed the overall trend of changes in caloric values. It was observed that in the initial months the random participant did have a better calorie intake than in the later months.



Then I also analyzed the daily changes:



The correlation between calorie intake and physical activity on the flourishing scale was -0.7961125935427582.



III. Extended Analysis

Hypothesis Analysis

Hypothesis A stated that participants' age will negatively correlate with financial and material stability (the last two questions on the Flourishing Scale). To investigate this, the team used the Flourishing Scale data recorded in the 'FridayAM.csv' file and participants' ages from the 'Demographic.csv' file. They preprocessed the data by taking the average of 'Living_expenses' and 'Food_housing' flourishing scores based on the mobile ID of the participants. They then merged this dataframe containing the averages with the ages on mobile ID and calculated the correlation between ages, living expenses, and food and housing expenses flourishing scores. The results showed that the correlation between age and financial stability was -0.096, and the correlation between age and food-housing expenses was -0.194.

Hypothesis B was that participants who take an average of four breaks per day will positively correlate with productivity scores in the E-Work Life Scale (questions 16-20) and report lower discomfort at one month compared to six-month data. To investigate this, the team analyzed the data related to the work-life balance of employees in a company using the 'FridayPM.csv' file, which contained the data for the last month. They calculated scores for different categories (Trust, Flexibility, Work_life, and Productivity) and then calculated the E-Work_life_scale based on the weighted scores of the categories. They merged the E-Work_life_scale data with the average breaks data for each employee and created a scatter plot of the E-Work_life_scale vs. Daily Breaks. The results showed that participants who take an average of four breaks per day will positively correlate with productivity scores in the E-Work Life Scale.

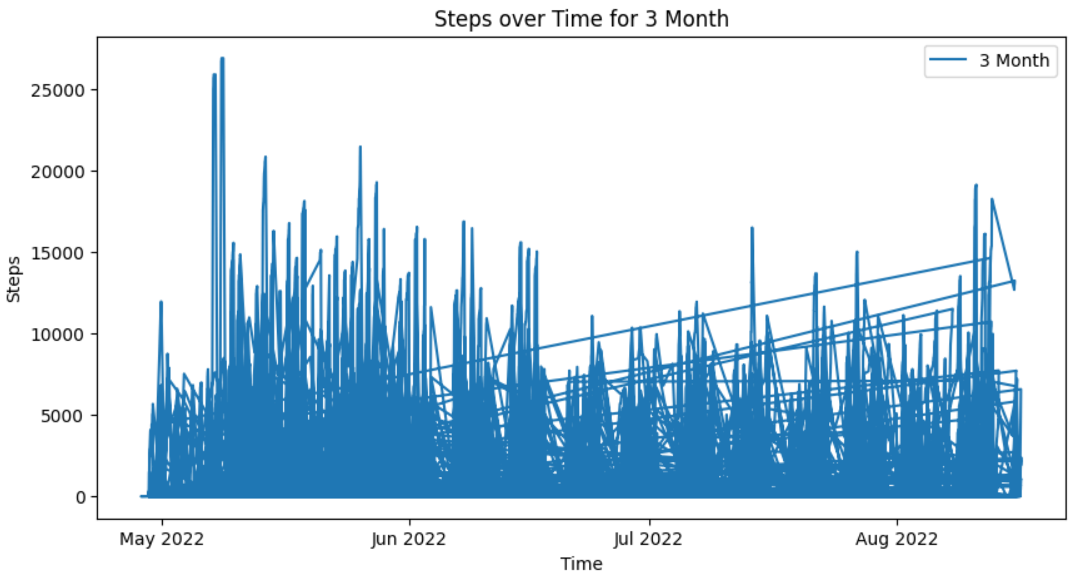
Hypothesis C was that participants working in healthcare will have lower mental health scores on the Flourishing Scale than those working in other industries. To investigate this, the team used the FridayPM 3 Month and Demographic datasets. They merged the datasets into one DataFrame by participant ID and then selected the relevant columns MENTAL_HEALTH and INDUSTRY, dropping the empty rows. They used the method 'groupby' in pandas and calculated the mean of the mental health score by industry. They calculated the average mental health score grouping by the industry and plotted a bar chart. Based on the data analysis performed on the dataset, they disproved the hypothesis that participants working in healthcare will have lower mental health scores on the Flourishing Scale than those working in other industries. However, an interesting observation showed that participants of Veteran Affairs have lower mental health scores overall on average.

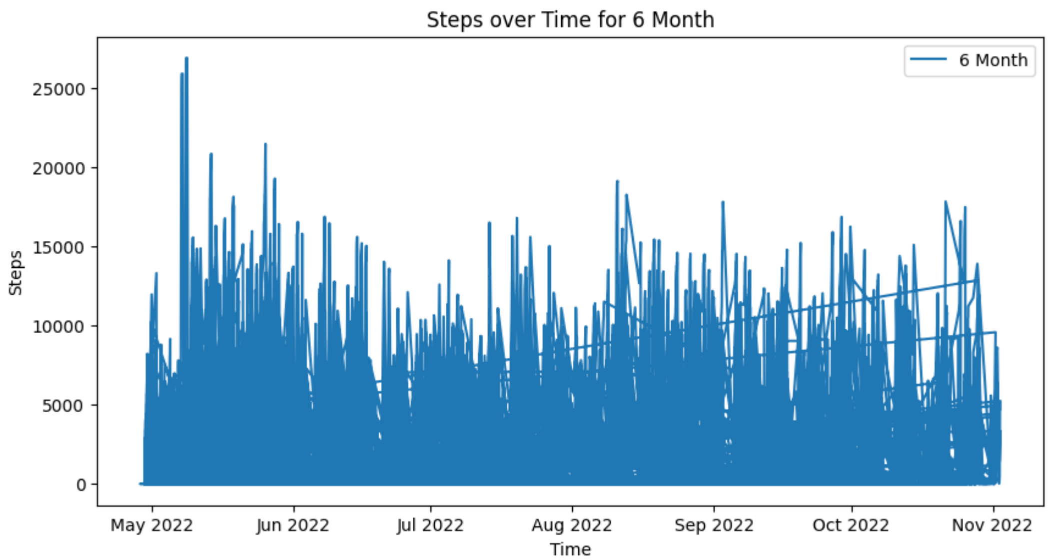
IV. Garmin Data Analysis & Answer to Questions

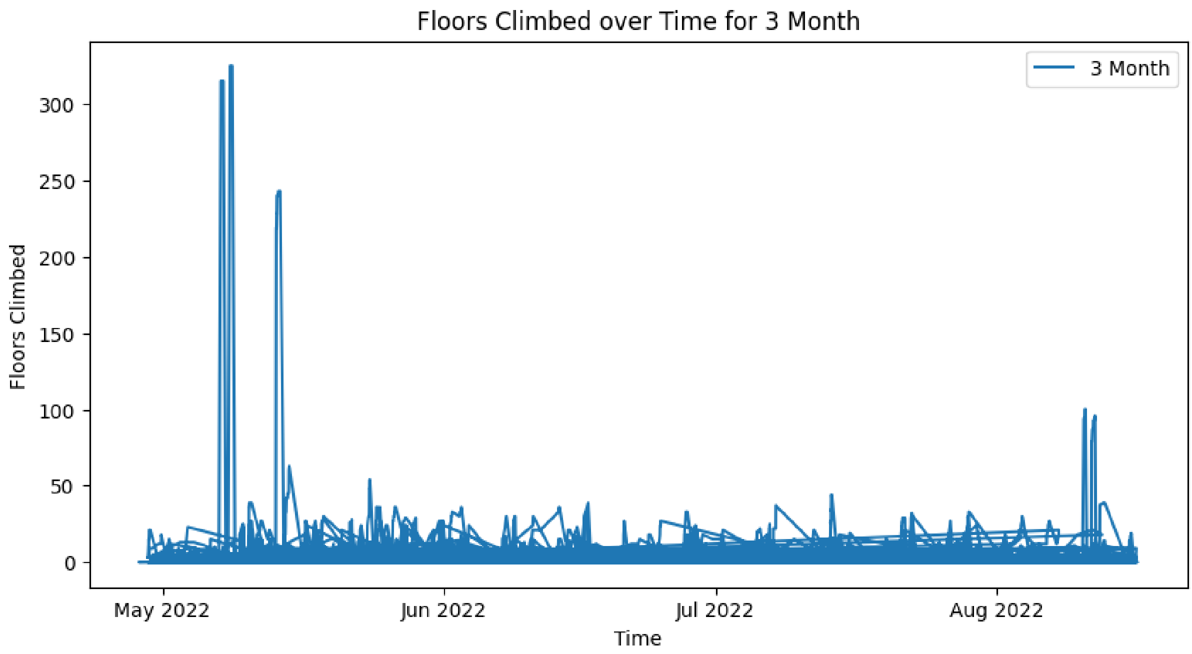
In this section, we wrote code to analyze the Garmin dataset. From the above analysis, we understood the overall structure of the Garmin data files. Generally speaking, the rows of Garmin data record a data type of its value, at a given time stamp, for a specific participant with his/her id. The data type contains eight attributes: ['calories', 'hr', 'steps', 'floorsClimbed', 'intensityMinutes', 'pulseOx', 'ibi', 'stress']. We splitted these eight attributes into four groups, each containing two, and analyzed them based on the questions raised in the requirement.

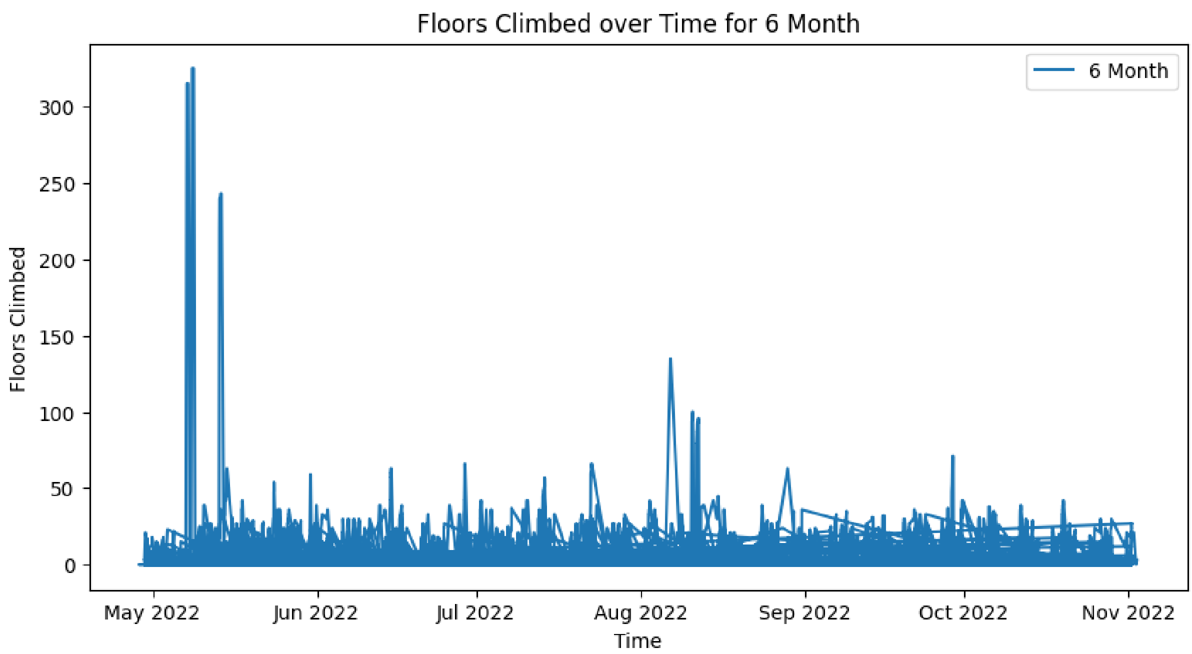
- **Group 1: ‘steps’ & ‘floorsClimbed’**

These two attributes were bonded together because we intuitively thought that steps would somehow correlate with floors climbed. The very first step of this task was to read data from files and put them into different data frames. After that, we did a very basic visualization for the data frames:



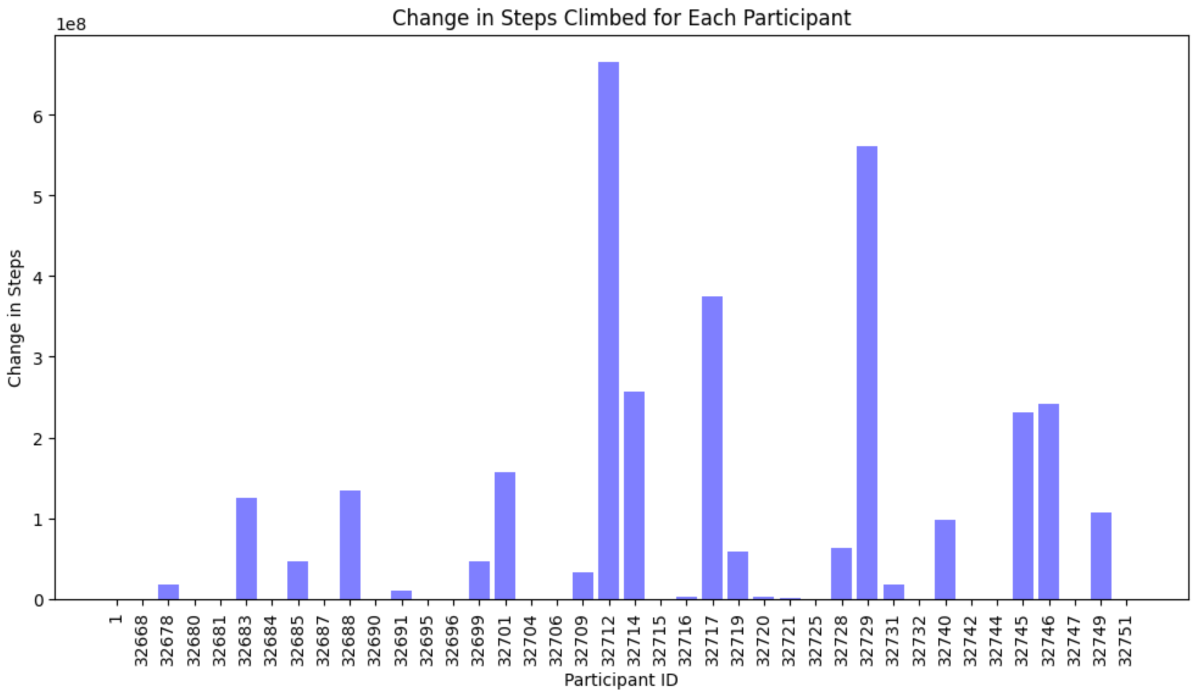


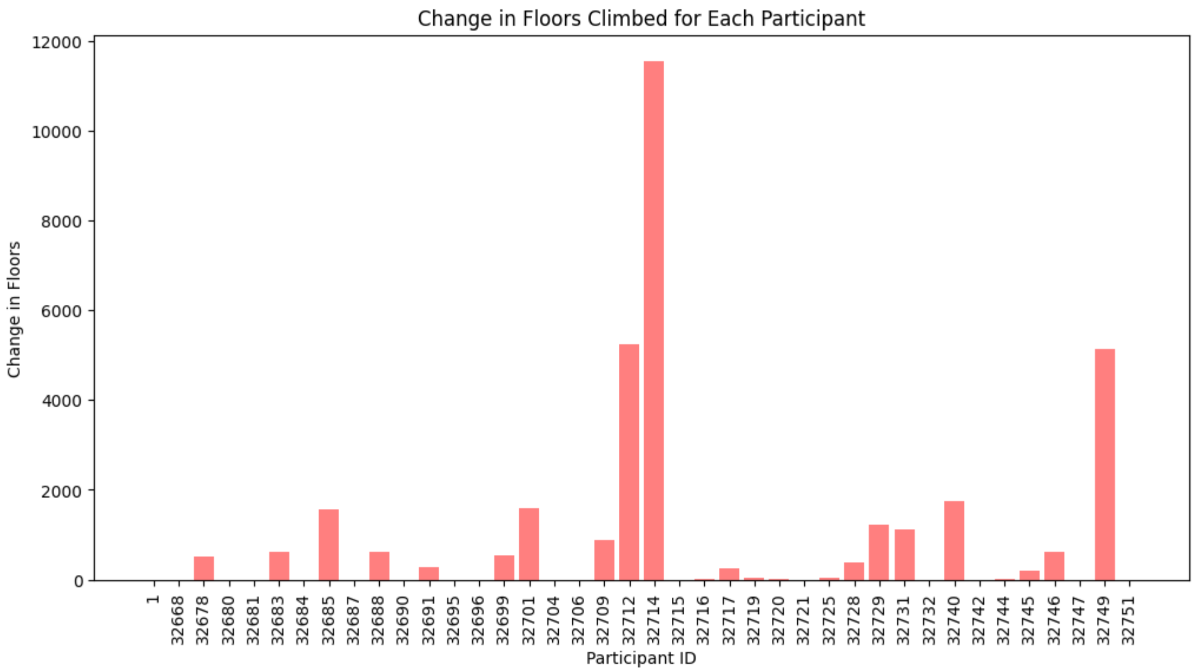




From the above analysis, we noticed a trend that both the steps and floor climbed had a peak at the beginning of the experiment.

To answer the question "Aggregate data and the change from 3 months to 6 months." We calculated the difference between 3 months and 6 months for each unique participant id that appeared in the Garmin dataset, by summing the steps and floors value in the data frames. Here is the result we got:



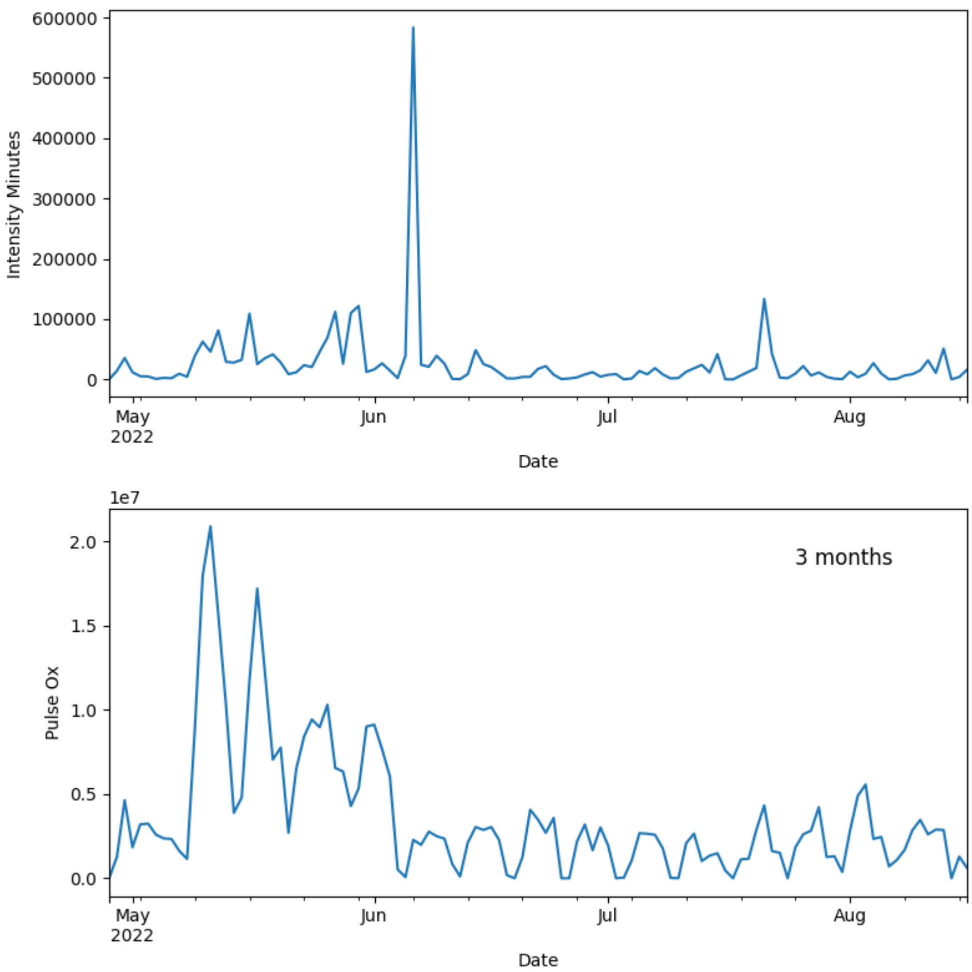


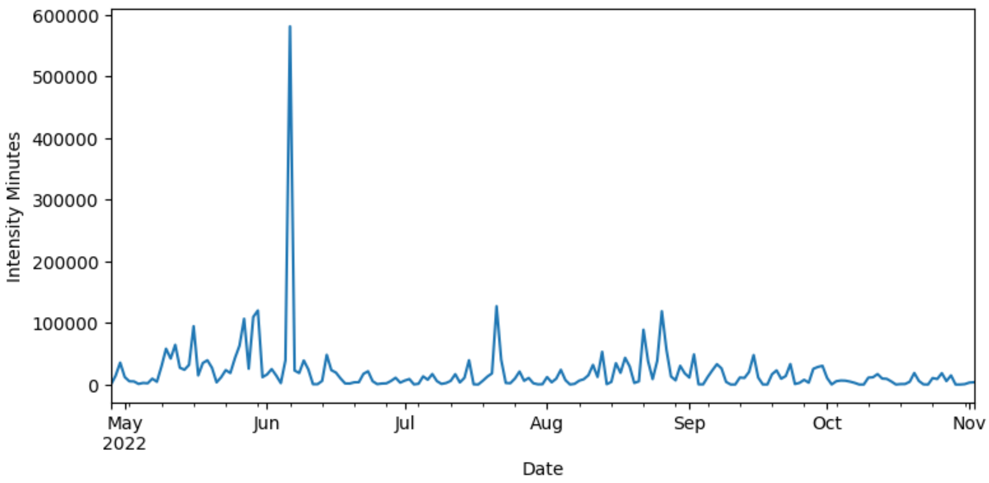
As we can observe from the graph produced, the x axis labels the unique participant ID, and the y axis represents the changed value. We noticed that for participants 32712 and 32714, they showed significant changes in both steps and floors climbed.

Moreover, based on the above results, we made a hypothesis that steps and floors climbed must have some sort of correlation, and we use the ‘corr’ method in pandas to calculate the correlation coefficient between those two variables. The result is 0.7, which indicates that steps and floors climbed have a positive correlation.

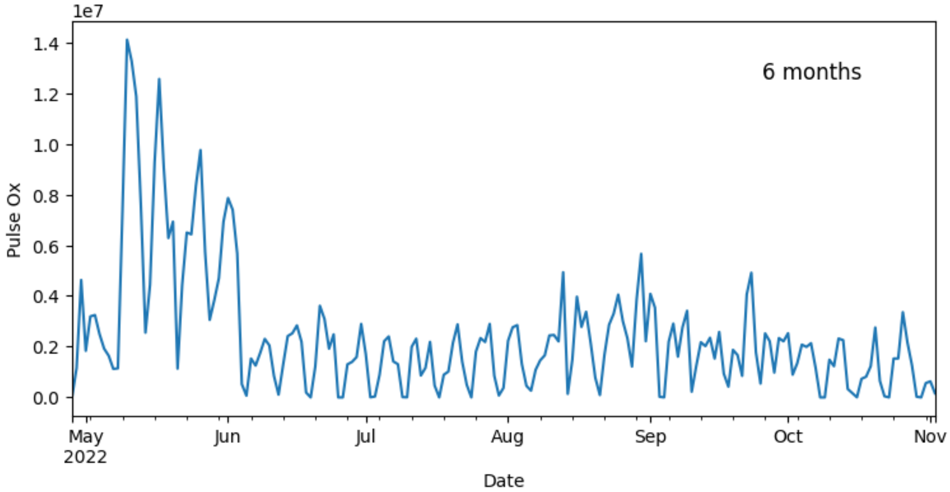
Finally, we wanted to analyze the impact of attrition. We used the number of unique id in the data frames, and ended up with a result of -15% attrition rate in the Garmin datasets. A negative attrition rate of -15% indicates that the number of participants in the study at the 6-month mark is greater than the number of participants at the 3-month mark. This result might seem counterintuitive because it is expected that the number of participants should decrease or remain constant over time due to dropout or non-response. A possible explanation for this negative attrition rate was that Garmin data did not contain all the necessary information, and attrition rate should be calculated on the survey datasets.

- **Group 2: ‘ intensityminutes’ and ‘pulse ox’**

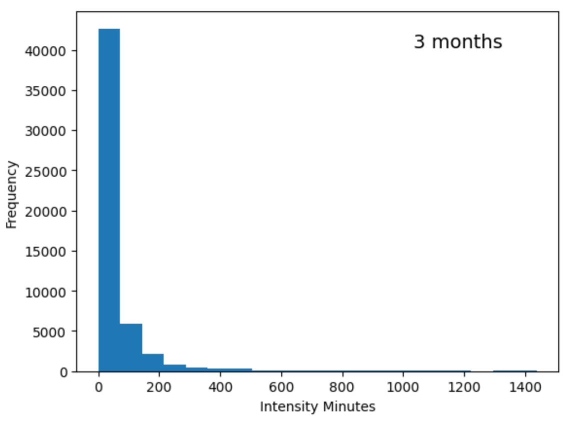




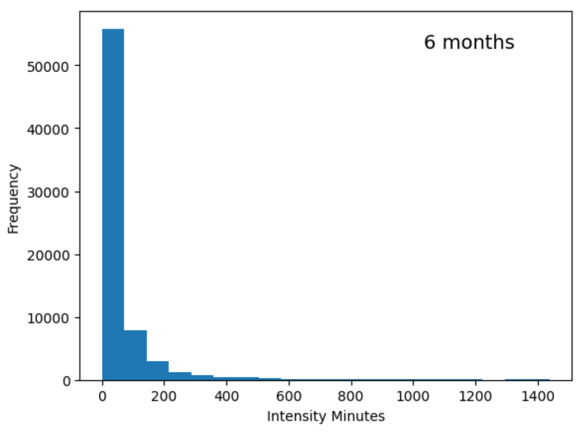
600000
500000
400000
300000
200000
100000
0
Intensity Minutes
May
2022
Jun
Jul
Aug
Sep
Oct
Nov
Date



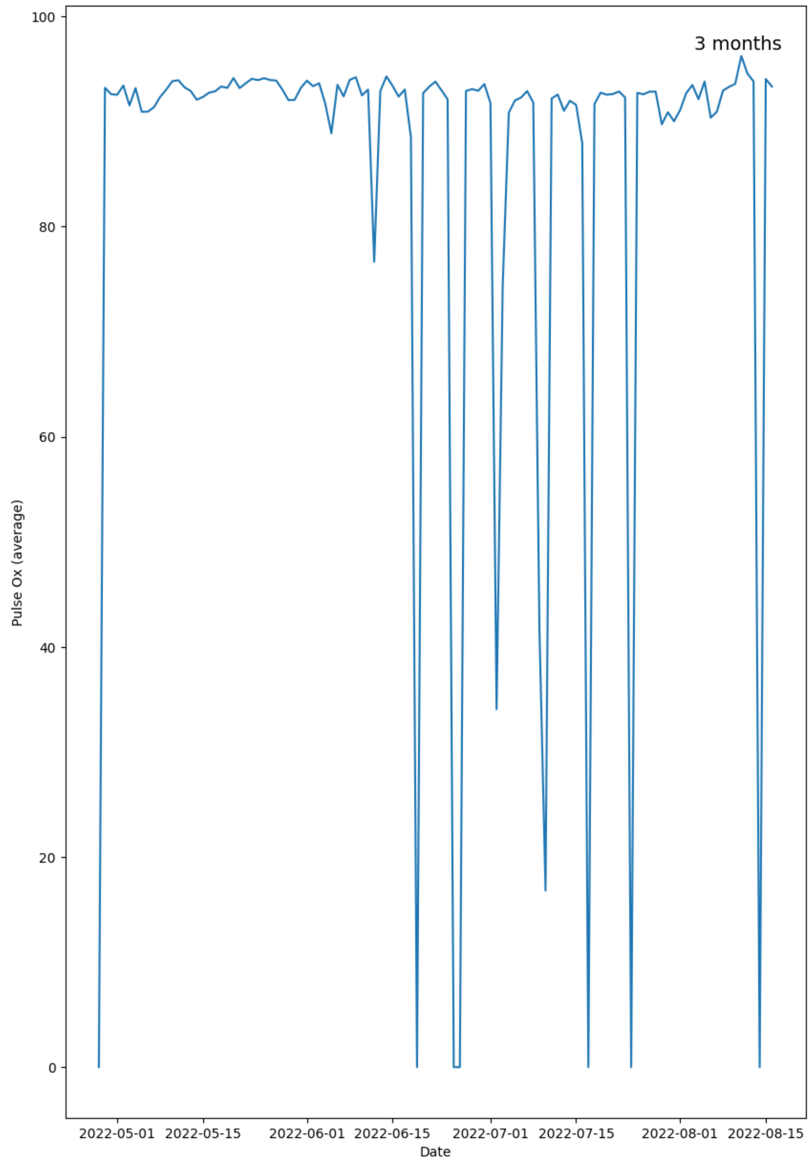
1e7
6 months
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Pulse Ox
May
2022
Jun
Jul
Aug
Sep
Oct
Nov
Date



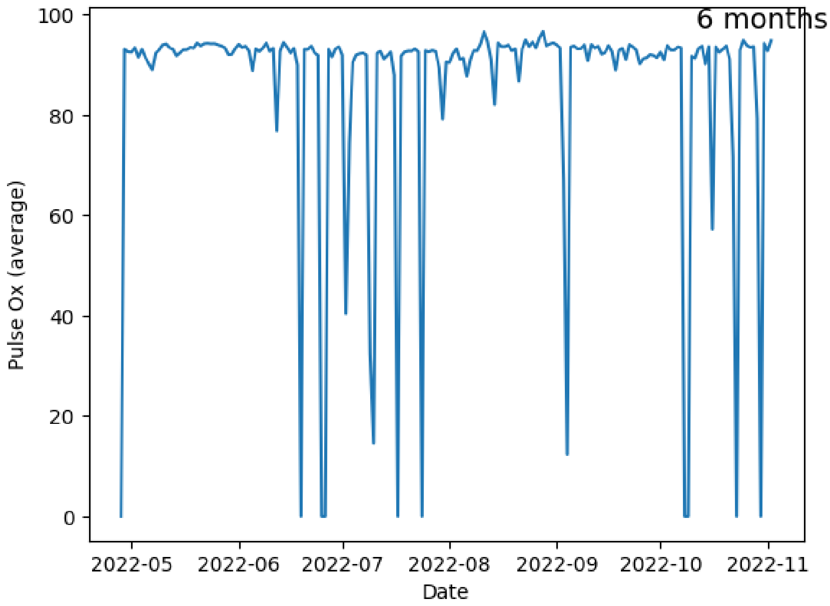
3 months
40000
35000
30000
25000
20000
15000
10000
5000
0
Frequency
0
200
400
600
800
1000
1200
1400
Intensity Minutes



6 months
50000
40000
30000
20000
10000
0
Frequency
0
200
400
600
800
1000
1200
1400
Intensity Minutes

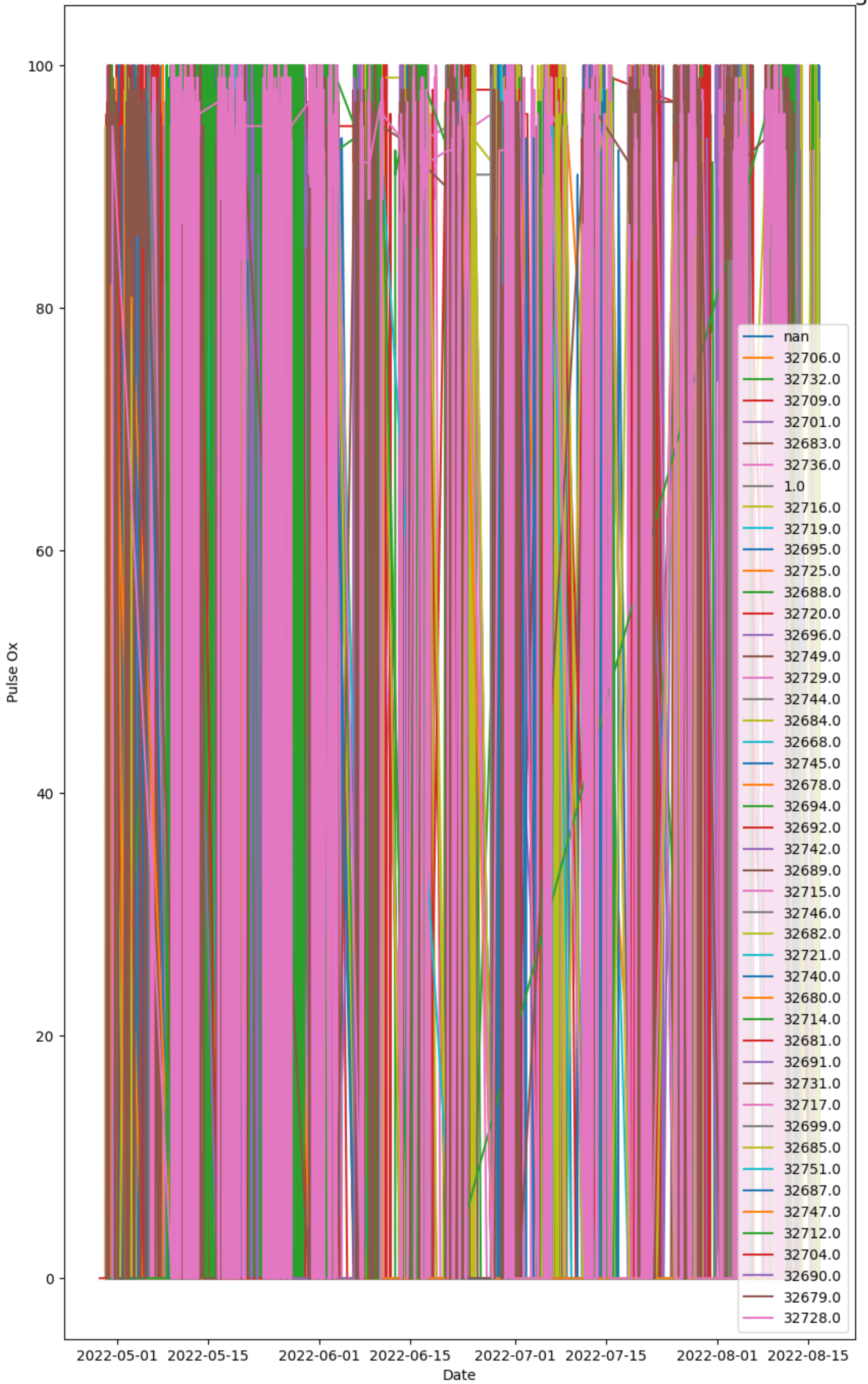


3 months
Pulse Ox (average)
Date
100
80
60
40
20
0
2022-05-01
2022-05-15
2022-06-01
2022-06-15
2022-07-01
2022-07-15
2022-08-01
2022-08-15



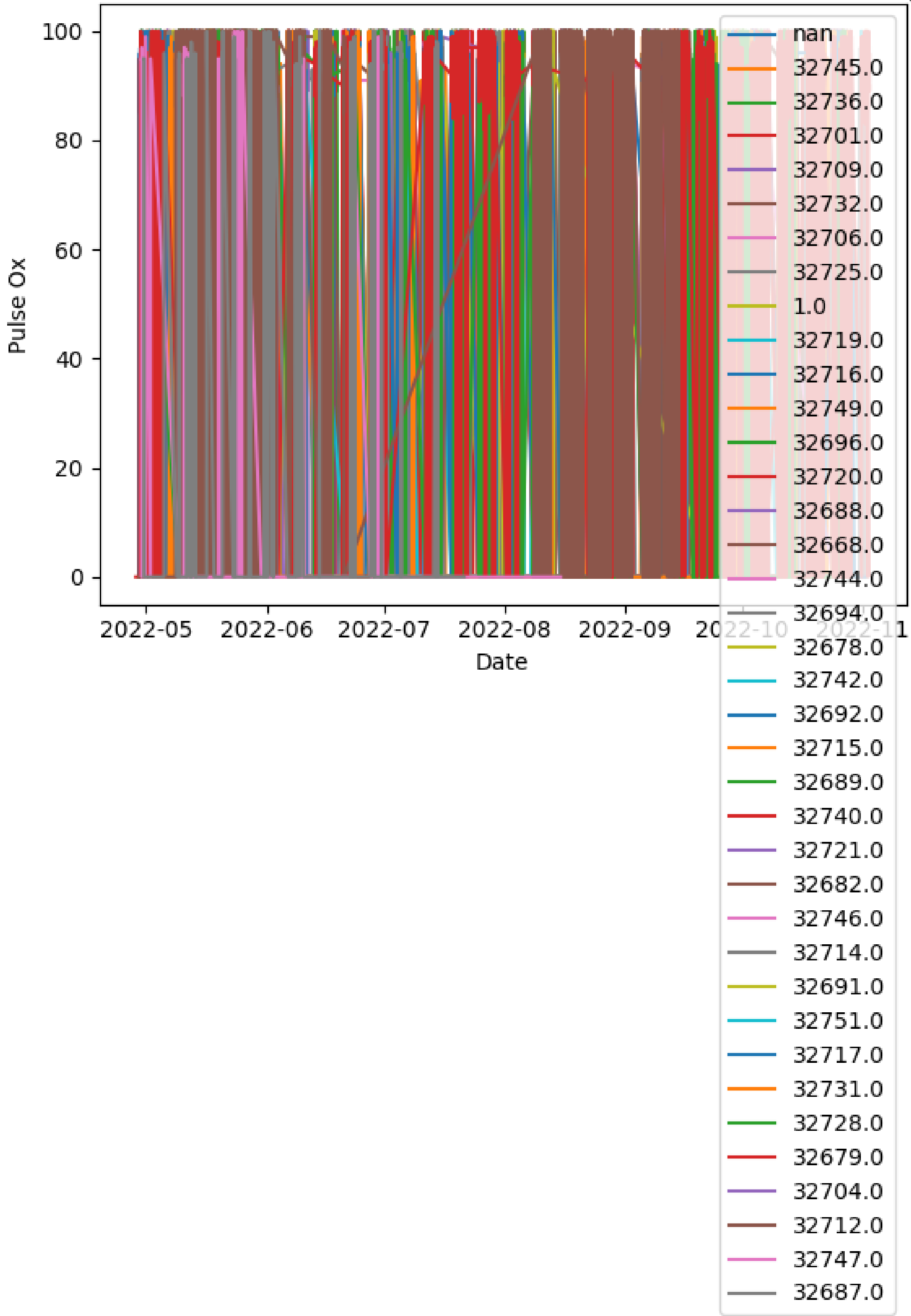
6 months
Pulse Ox (average)
Date
100
80
60
40
20
0
2022-05
2022-06
2022-07
2022-08
2022-09
2022-10
2022-11

3 months

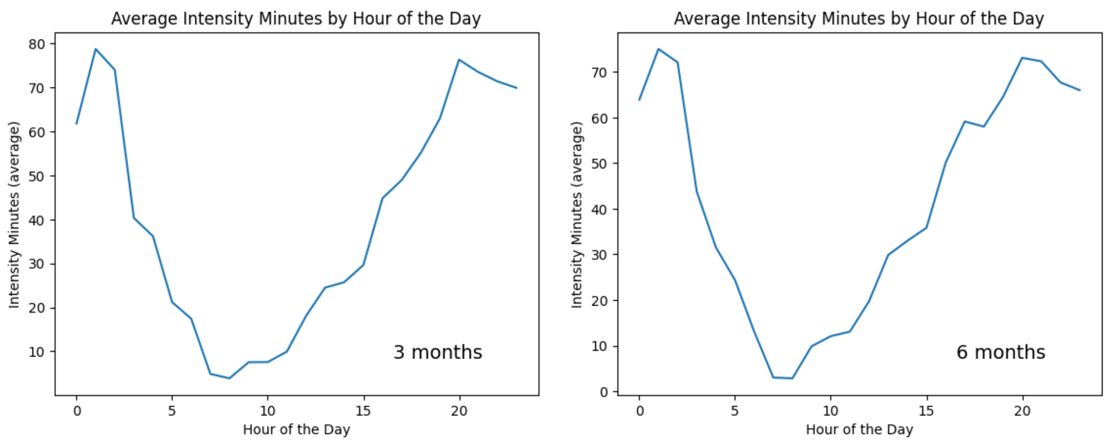


Pulse Ox
100
80
60
40
20
0
2022-05-01
2022-05-15
2022-06-01
2022-06-15
2022-07-01
2022-07-15
2022-08-01
2022-08-15
Date
nan
32706.0
32732.0
32709.0
32701.0
32683.0
32736.0
1.0
32716.0
32719.0
32695.0
32725.0
32688.0
32720.0
32696.0
32749.0
32729.0
32744.0
32684.0
32668.0
32745.0
32678.0
32694.0
32692.0
32742.0
32689.0
32715.0
32746.0
32682.0
32721.0
32740.0
32680.0
32714.0
32681.0
32691.0
32731.0
32717.0
32699.0
32685.0
32751.0
32687.0
32747.0
32712.0
32704.0
32690.0
32679.0
32728.0

6 months



Pulse Ox
100
80
60
40
20
0
2022-05
2022-06
2022-07
2022-08
2022-09
2022-10
2022-11
Date
nan
32745.0
32736.0
32701.0
32709.0
32732.0
32706.0
32725.0
1.0
32719.0
32716.0
32749.0
32696.0
32720.0
32688.0
32668.0
32744.0
32694.0
32678.0
32742.0
32692.0
32715.0
32689.0
32740.0
32721.0
32682.0
32746.0
32714.0
32691.0
32751.0
32717.0
32731.0
32728.0
32679.0
32704.0
32712.0
32747.0
32687.0



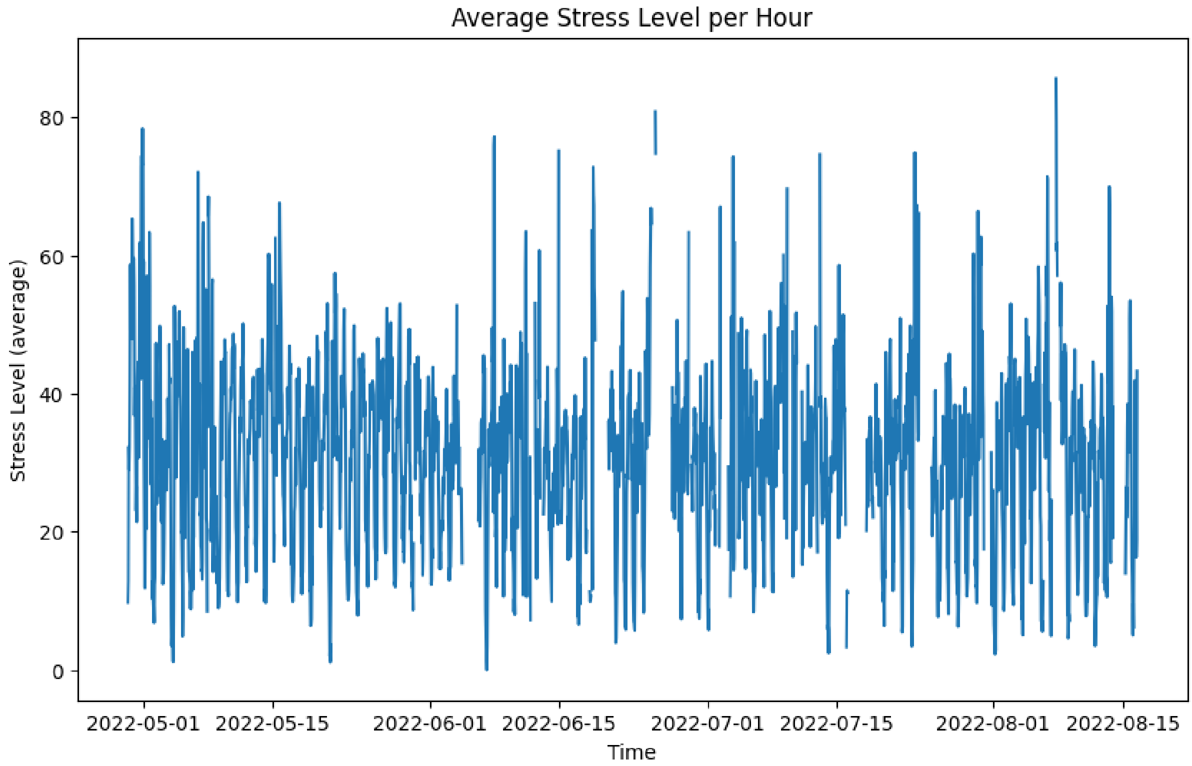
The report analyzed the second group of attributes in the Garmin dataset, which were 'intensityminutes' and 'pulse ox'. The data was first visualized using scatter plots, where we observed no clear relationship between the two variables. However, we did notice some outliers that had extreme values for both attributes. To explore this further, we calculated the average and total values for each attribute, with the average 'intensityminutes' being 53.88 and the average 'pulse ox' being 93.09. The total 'intensityminutes' recorded was 2,876,595, while the total 'pulse ox' recorded was 402,148,916.

To better understand the data, we created various visualizations, including scatter plots, line plots, and histograms. The scatter plots showed the relationship between intensity minutes and pulse ox, while the line plot tracked pulse ox values over time for each participant. The histogram showed the frequency distribution of intensity minutes. We also identified missing values in the dataset, which were dropped using the dropna() function. Outliers in the intensity minutes data were removed using the z-score method.

Based on our analysis, we found that the participants had an average intensity minutes value of 53.88 and an average pulse ox value of 93.09. The total intensity minutes recorded was 2,876,595, and the total pulse ox value recorded was 402,148,916. Overall, our analysis provided us with valuable insights into the participants' fitness and health levels, and the visualizations helped us draw meaningful conclusions.

- **Group 3: 'stress'**

This attribute was used to calculate the average stress over time for the participants.



V. Conclusion

We conducted statistical analyses on various datasets to investigate relationships between different factors. Through data preprocessing and merging, we were able to uncover interesting insights about the participants. For example, we found that age was negatively correlated with financial and material stability, while daily breaks were positively correlated with productivity scores. Additionally, while there was no significant difference in mental health scores between healthcare workers and those in other industries, participants of Veteran Affairs had lower mental health scores overall on average. These findings can have practical implications for organizations looking to improve employee well-being and productivity. By understanding these relationships, organizations can develop interventions and policies that address the specific needs of their workforce, ultimately leading to a happier and more productive workforce.