Working from the Home Environment & Well-Being Study Data

Final Report Draft

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)

1. Task

The goal of this project is to analyze the data generated in a study conducted with 70 volunteers on work-from-home well-being. We aim to find meaningful relationships between the various types of data collected and provided for our client. Our client urged us to focus more on the Garmin data and the relationship between various data points rather than the initial hypothesis mentioned in the project expectations document.

2. Introduction

Due to the global pandemic, many companies had to shift from a work-from-office model to a work-from-home model. This remained prevalent even after the national lockdown was lifted in either complete or hybrid forms. The importance of this study lies in the well-being of this new way of working and can be imperative for forming new policies for employees.

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 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 the volunteers.

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.

Overall, throughout the scope of this project, we aim to identify relationships between the various data points provided for a better understanding of the work-from-home well-being of employees.

3. Data Used

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

1. 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.
2. Participants' physical location data
3. Participants' musculoskeletal discomfort data
4. Responses to the following surveys were also recorded in over 6 different CSV files:
    a. Computer workstation checklist
    b. E-Work Life scale
    c. Flourishing scale survey
    d. Visual Analog Scale

4. Base Analysis

The data and background information were thoroughly reviewed and the initial hypotheses were answered. The initial proposed questions regarding the data were also answered. It was confirmed that all the subjects had an ID number associated with them, and the changes from 3 to 6 months were aggregated during the preprocessing stages.

Hypothesis	Observations	Accepted/rejected
Participants' age will negatively correlate with financial and material stability (the last two questions on the Flourishing Scale)	Corr. Age & Living Exp = -0.096 Corr. Age & Food-housing expenses = -0.194	Accepted
Emp who take an average of 4 breaks/day will positively correlate with productivity scores and report lower discomfort at 1-month compared	They corr positively but there was not much significant difference between 1-month and 6-month data results	Rejected

to 6-month data		
Emp working in healthcare will have lower mental health scores than those in other industries	The average mental health score for healthcare was not significantly different from the average mental health score in other industries.	Rejected
Participants stress algorithm will be inversely correlated to their number of breaks.	The correlation score was -0.1169	Accepted

Table 4.1: An overview of the base analysis

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

Background information

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:
0 = Worry All of the Time
10 = Do Not Ever Worry

Data used:

The flourishing scale was filled by participants weekly on Friday. The data is recorded in the 'FridayAM.xlsx' file. Next, we need the participants' ages. This is taken from the 'Demographic.csv' file.

Data preprocessing:

For this hypothesis, we only need **LIVING_EXPENSES** and **FOOD_HOUSING**. So, a separate data frame was created from **FridayAM.xlsx**. Then, their average was calculated. Ages were then appended to the data frame after loading them from **Demographic.csv**.

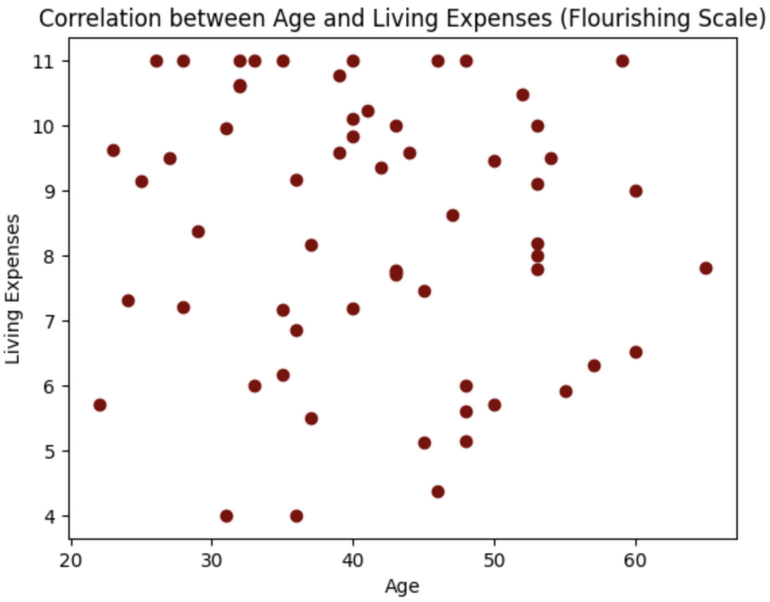
Methods and results:

To approve or disapprove of the hypothesis, the correlation between ages, living expenses, and food and housing expenses flourishing scores were calculated using the Pearson correlation coefficient given by:

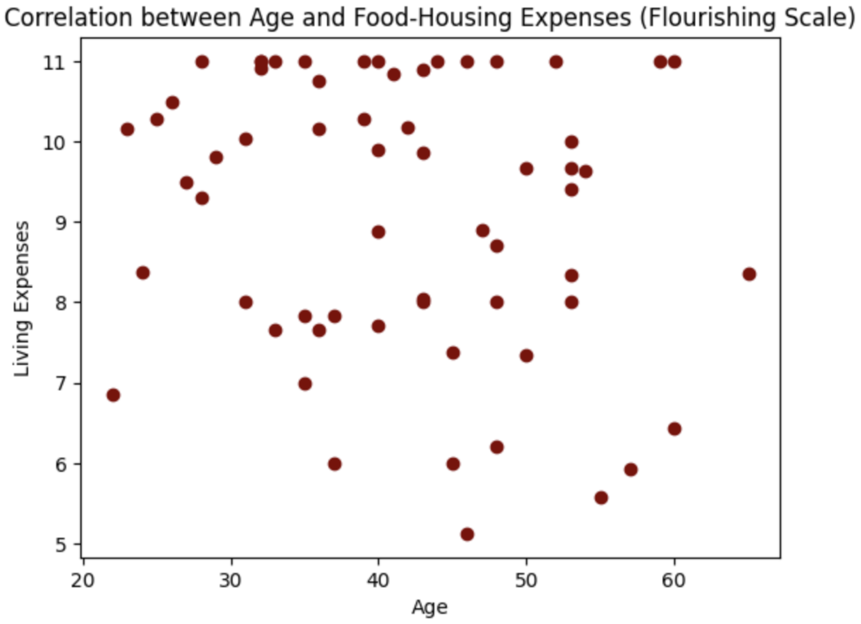
$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

The following results were obtained:

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



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



This hypothesis is accepted.

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 information

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:

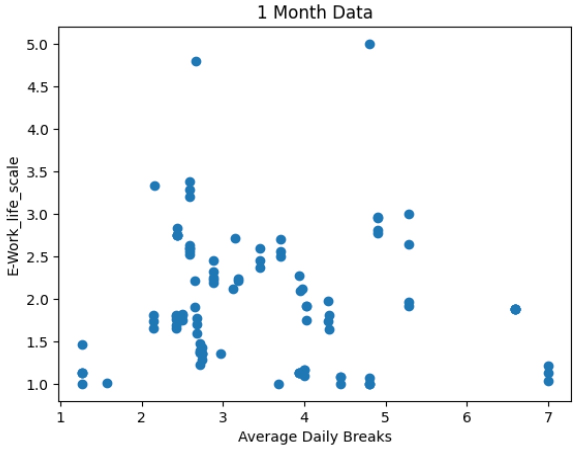
It drops rows with missing values and calculates scores for different categories (Trust, Flexibility, Work_life, and Productivity). The code then calculates the E-Work_life_scale based on the weighted scores of the categories. It merges the **E-Work_life_scale data** with the average breaks data for 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.

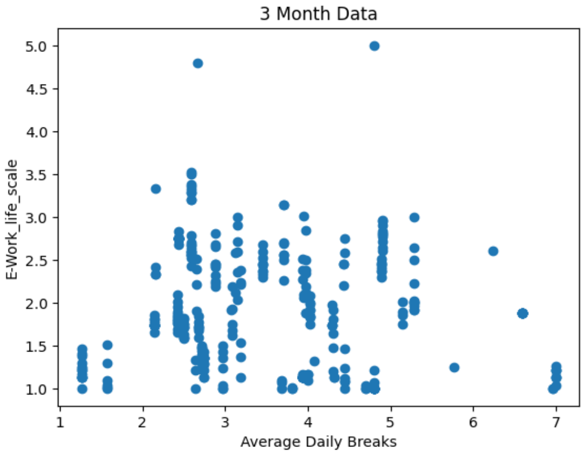
Methods and results:

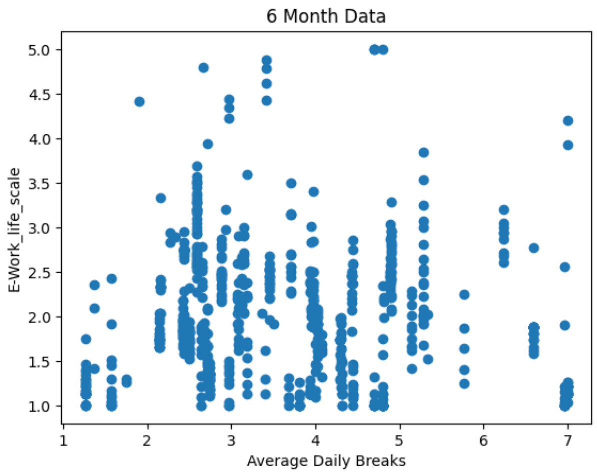
The purpose of this code is to provide insights into the work-life balance of employees and the impact of daily breaks on their work-life balance. The scatter plot and the mean E-Work_life_scale provide useful information for management to improve the work environment and enhance the well-being of employees. To approve or disapprove of the hypothesis, the correlation was calculated using the Pearson correlation coefficient as done previously.

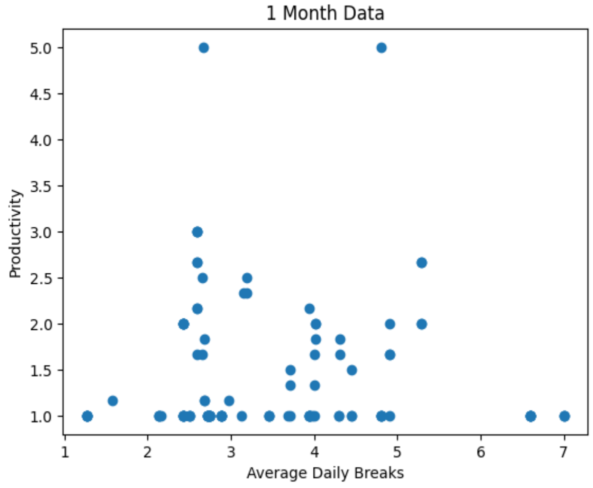
['E-Work_life_scale'] = (['Trust'] * 0.4) + (['Flexibility'] * 0.3) + (['Work_life'] * 0.2) + (['Productivity'] * 0.1)

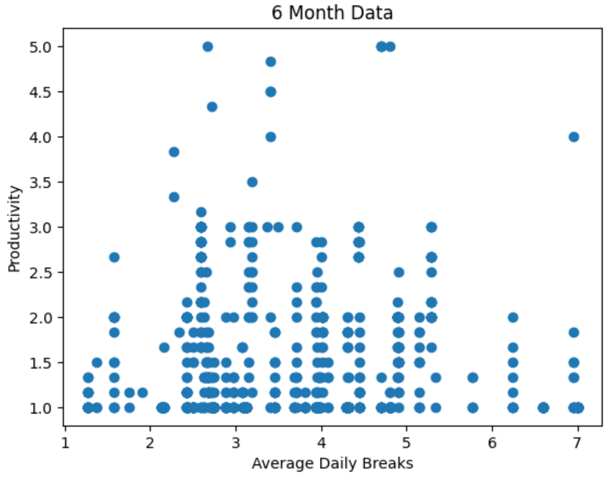
The following results were obtained:


1 Month Data


3 Month Data







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

Background information

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

How would you rate your overall mental health?

0 = Poor, 10 = Excellent

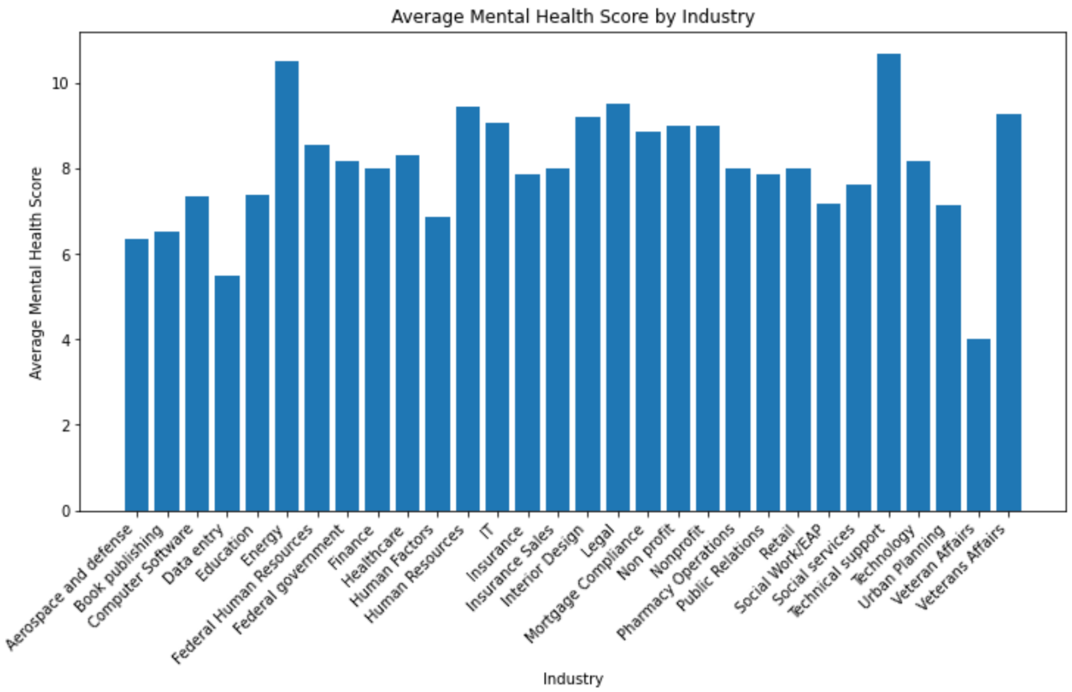
Data used:

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.

Methods and results:



Here we calculate the average mental health score grouping by 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 were 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.

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

This hypothesis is rejected.

Hypothesis D: Participants' stress algorithm will be inversely correlated to their number of breaks.
Hypothesis: 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 information

The code is analyzing data related to participants' stress algorithm and their number of breaks in the last 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 and filters the data for the last month.

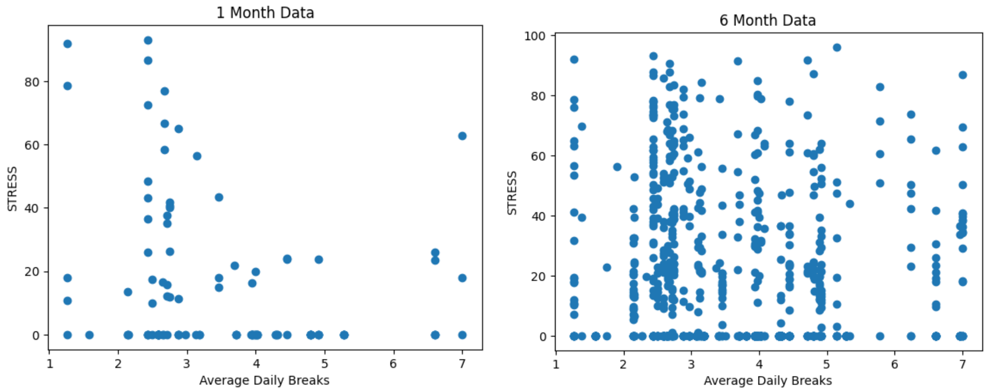
Data preprocessing:

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

Methods and results:

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.



Correlation:

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

The hypothesis is accepted.

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 information:

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

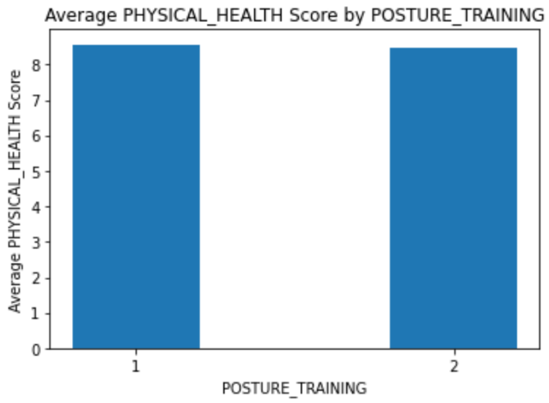
Data used:

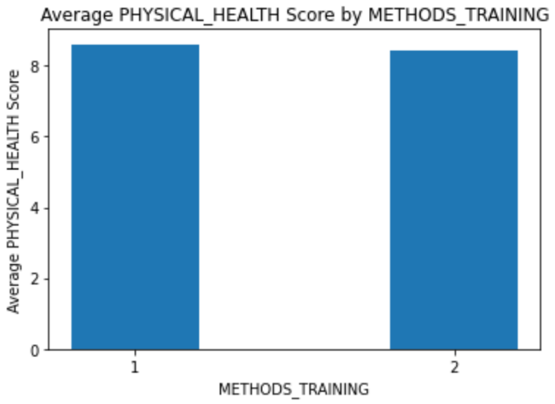
Computer Workstations 6 Month; Friday AM 6 Month

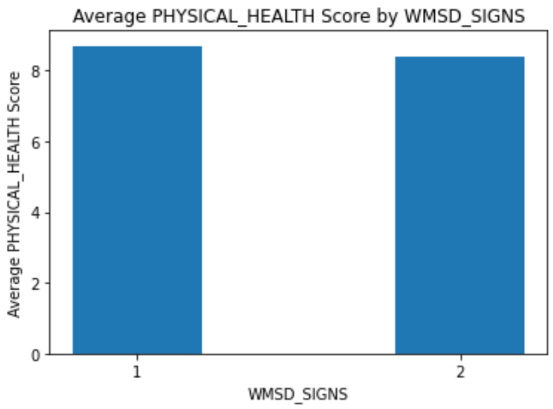
Data preprocessing:

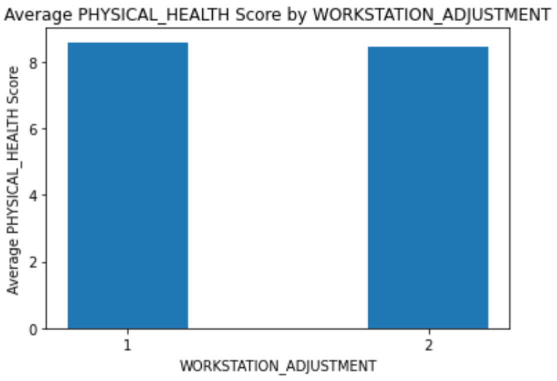
The code merges two datasets into one DataFrame by participant id, and then selects the relevant columns, dropping the empty rows. The first calculates the average level of pain with respect to each area, and then calculates the correlation coefficient for each of them.

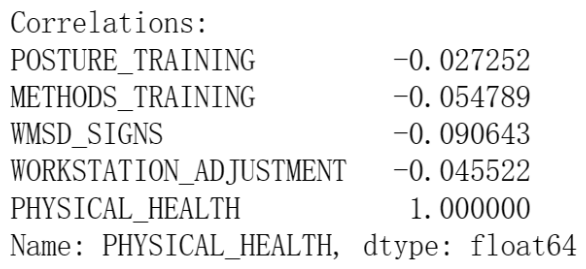
Methods and results:











```
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:

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

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 fewer 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.

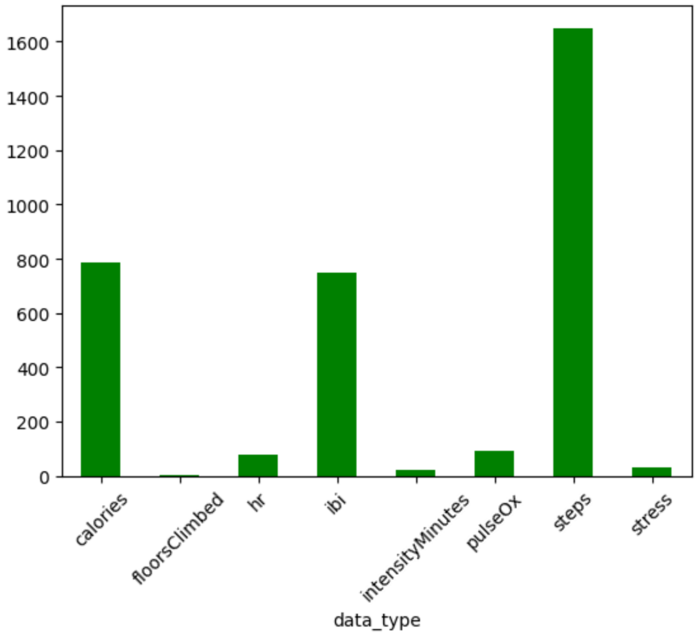
Based on question #15 in the Computer Workstation Checklist (with 4 responses regarding ergonomics training), participants with lower scores will report less pain at 6 months.

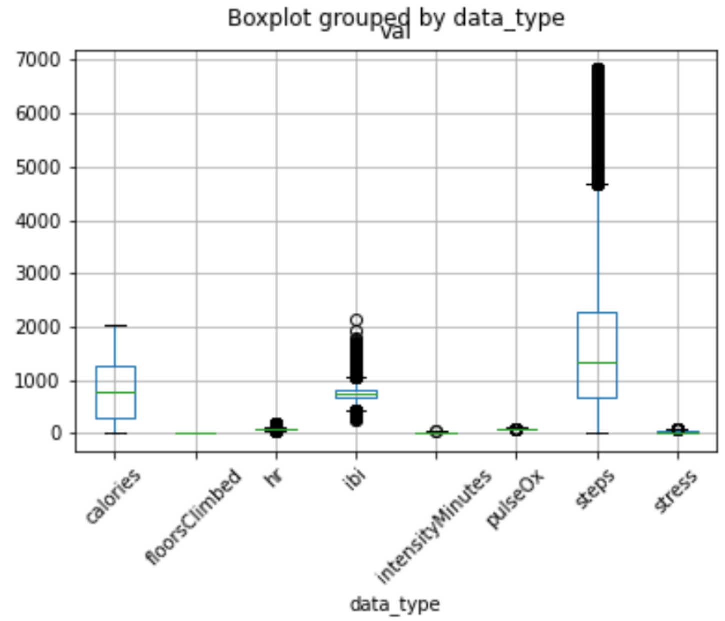
The hypothesis is accepted.

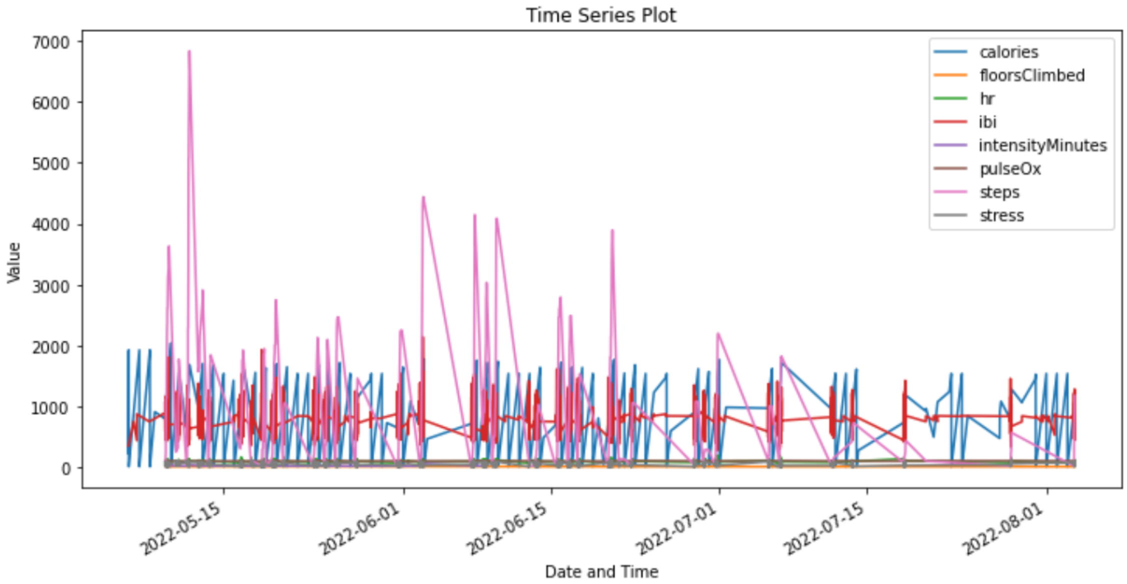
A first look at the Garmin Data

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:

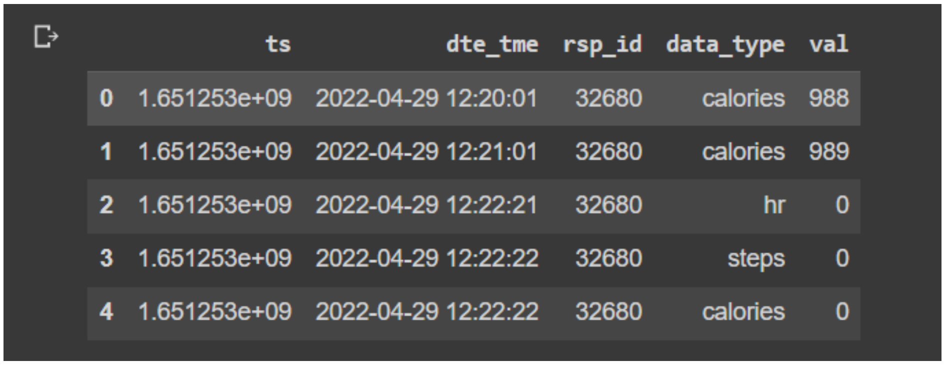
The first analysis uses the groupby() method of the DataFrame to group data by 'data_type'. The mean, median, and standard deviation of the 'val' column for each group are then calculated and visualized through bar and box plots.

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 of the Garmin Data

The Garmin data has the following structure:



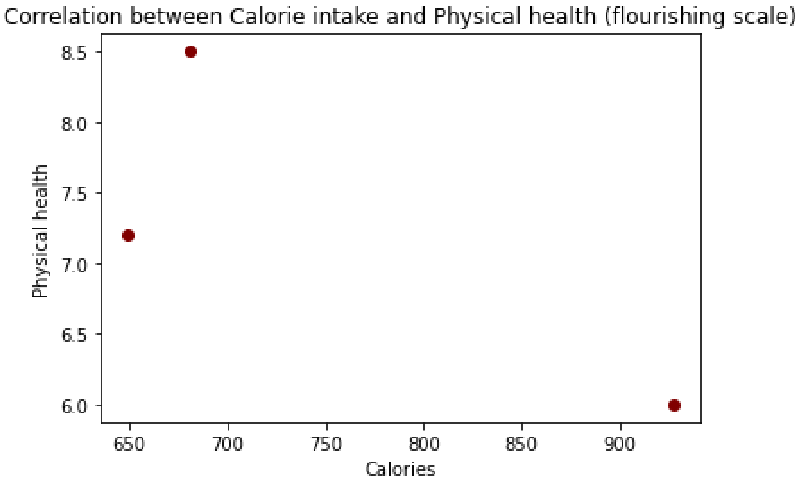
	ts	dte_tme	rsp_id	data_type	val
0	1.651253e+09	2022-04-29 12:20:01	32680	calories	988
1	1.651253e+09	2022-04-29 12:21:01	32680	calories	989
2	1.651253e+09	2022-04-29 12:22:21	32680	hr	0
3	1.651253e+09	2022-04-29 12:22:22	32680	steps	0
4	1.651253e+09	2022-04-29 12:22:22	32680	calories	0

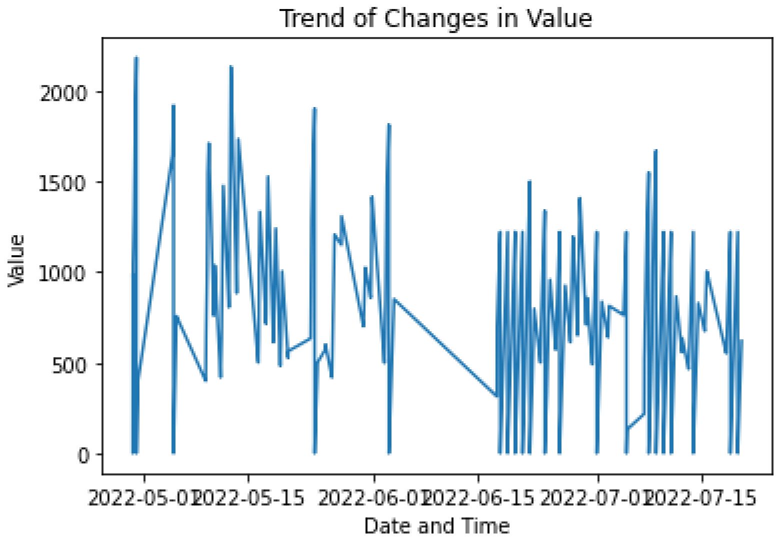
It contains the timestamps in a day over a period of 6 months for all the participants in over 60 different CSV files.

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

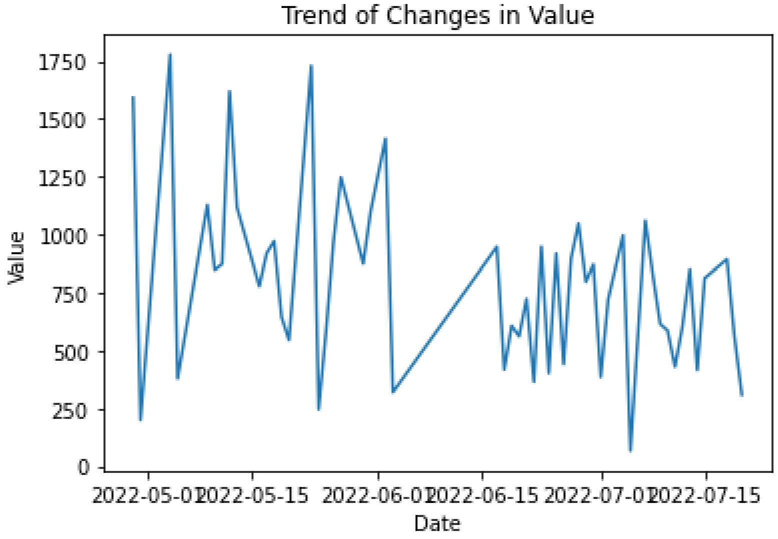
For the scope of this analysis, I extracted the calories, grouped them by days and then months, and correlated them with the physical health of a person.

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.

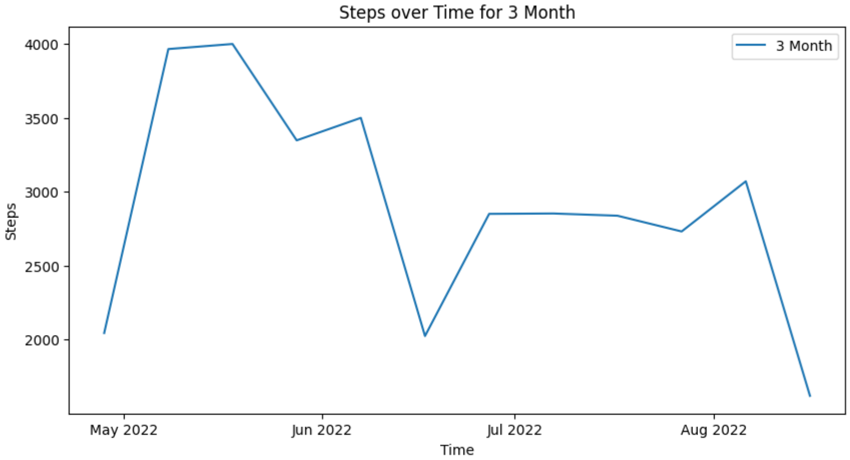
5. Extended Analysis

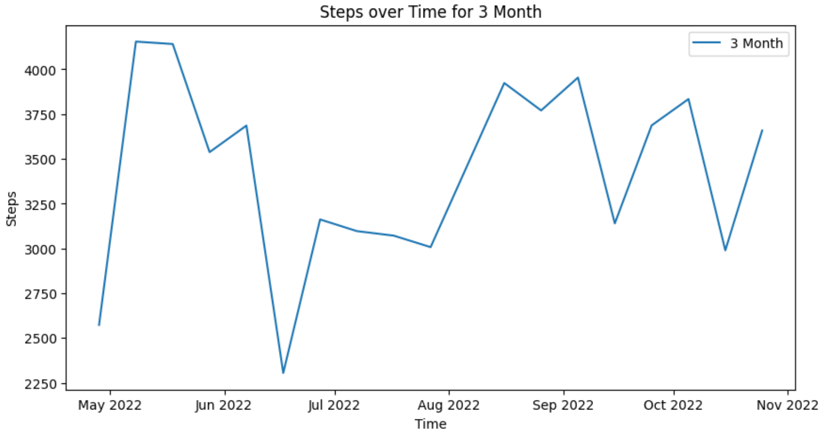
The extended analysis was performed as per the client's requirements to analyze the Garmin data in great detail. An overview of our approach was to analyze the various types of data collected by the Garmin watches and see if any correlations were observed from them or if we could find any major conclusions based on the various data types.

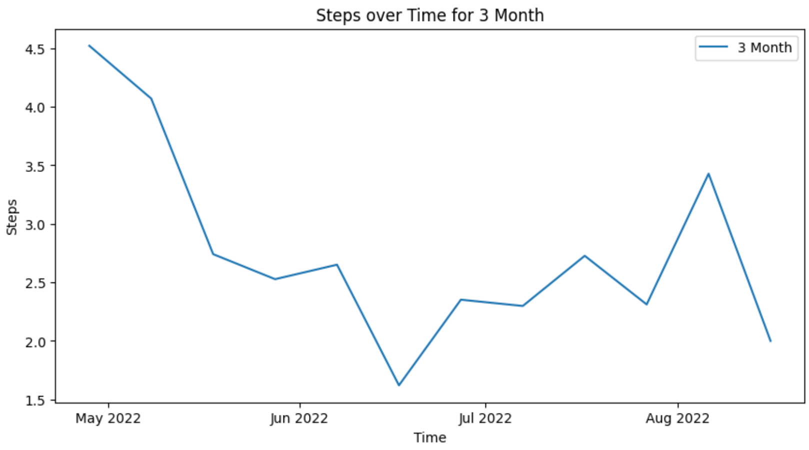
Generally speaking, the rows of Garmin data record a data type of its value, at a given time stamp, for a specific participant with his/her id. The data type contains eight attributes: ['calories', 'hr', 'steps', 'floorsClimbed', 'intensityMinutes', 'pulseOx', 'ibi', 'stress']. We split these eight attributes into four groups, each containing two, and analyzed them based on the questions raised in the requirement. The splitting of these groups was based on logic since the attributes in the groups complimented each other. We then also found the overall correlations between them.

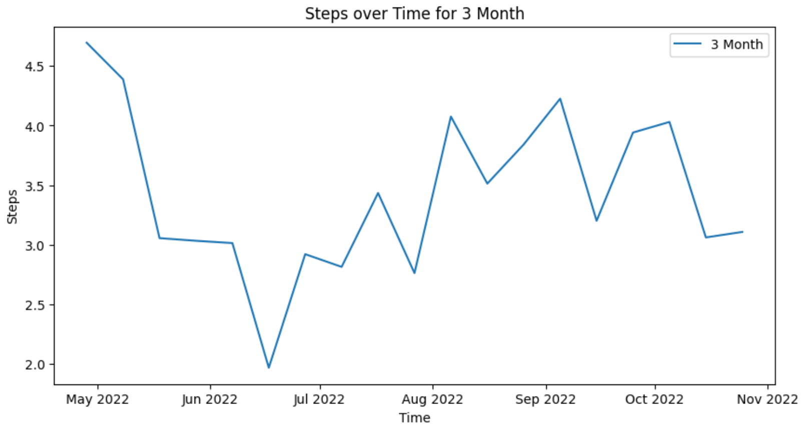
- **Group 1: 'steps' & 'floorsClimbed'**

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 of the data frames:



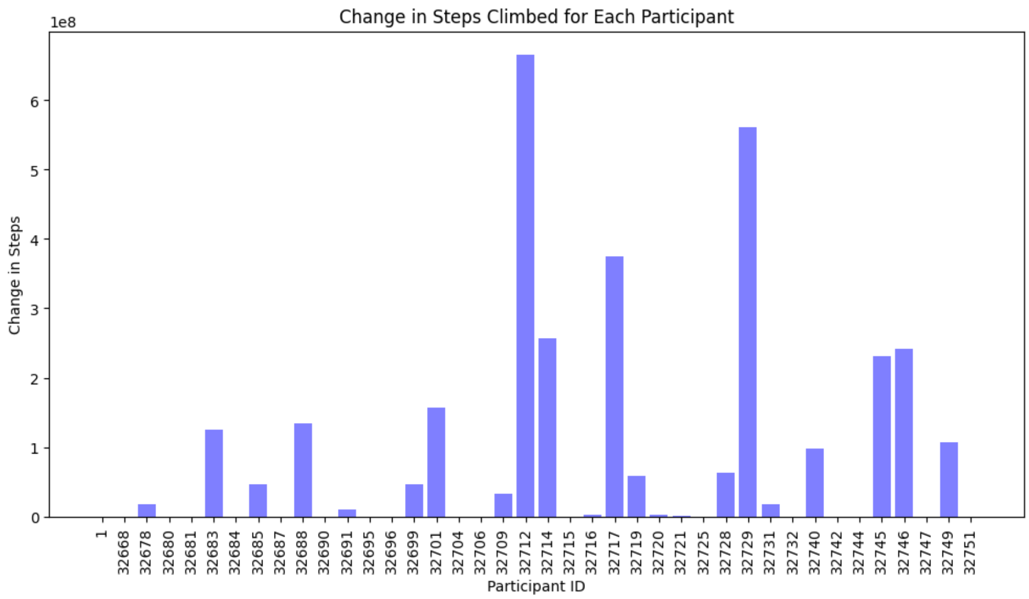


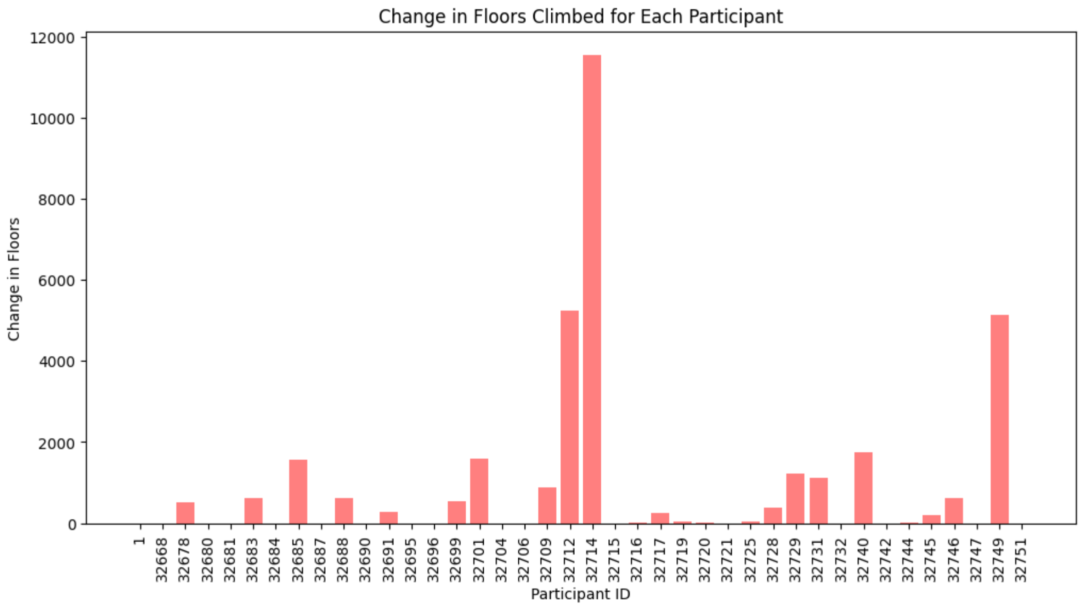




From the above analysis, we noticed a trend that both the steps and floor climb had a peak at the beginning of the experiment.

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:


Change in Steps Climbed for Each Participant


Change in Floors Climbed for Each Participant

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 participants 32712 and 32714, 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.

$$\mathbf{corr(x, y) = cov(x, y) / (std(x) * std(y))}$$

where cov is the covariance between x and y, std is the standard deviation of x and y.
The result is 0.7, which indicates that steps and floors climbed to 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'**

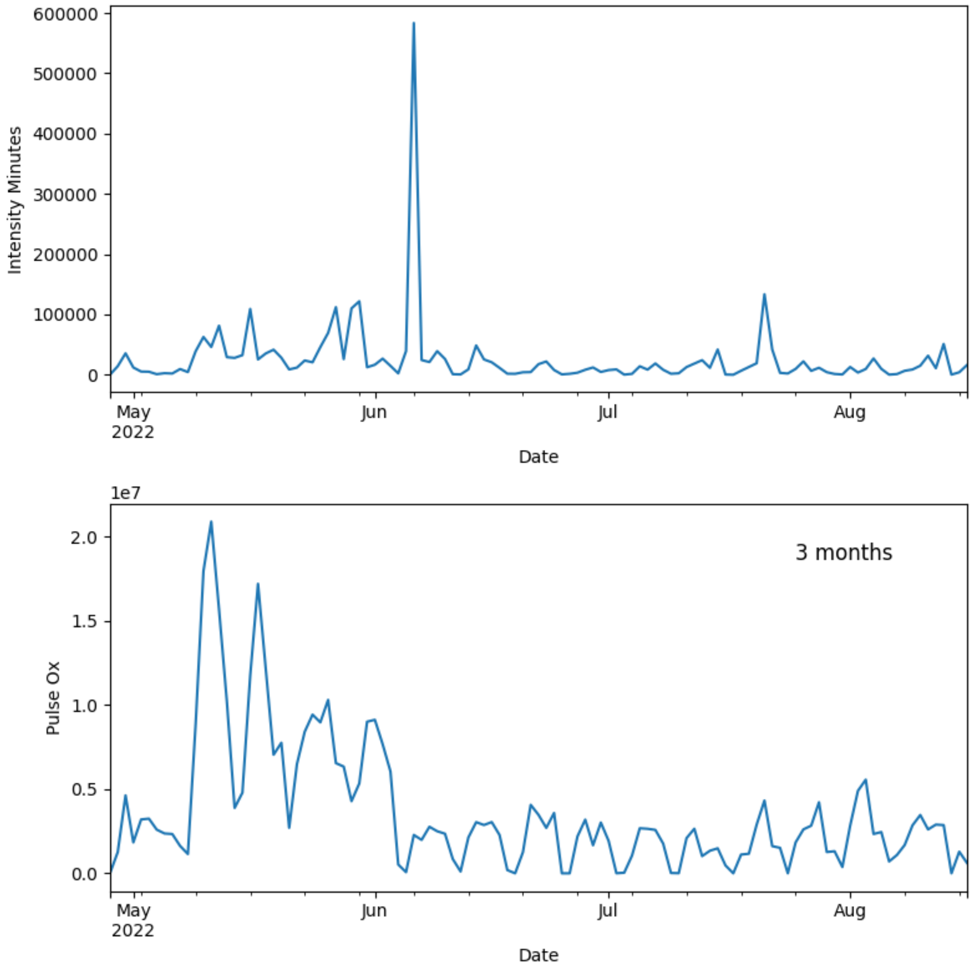


Fig: Intensity minutes and Pulse Ox data throughout the 3 months

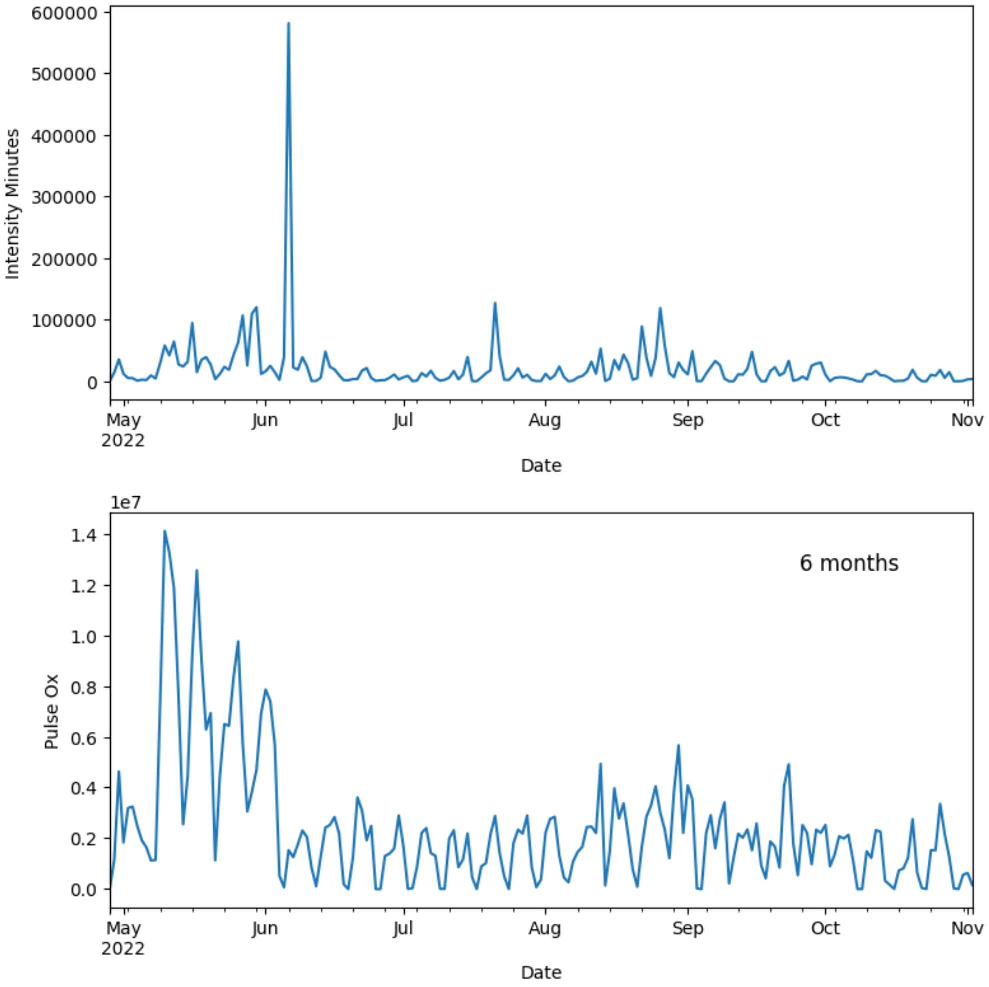


Fig: Intensity minutes and Pulse Ox data throughout the 6 months

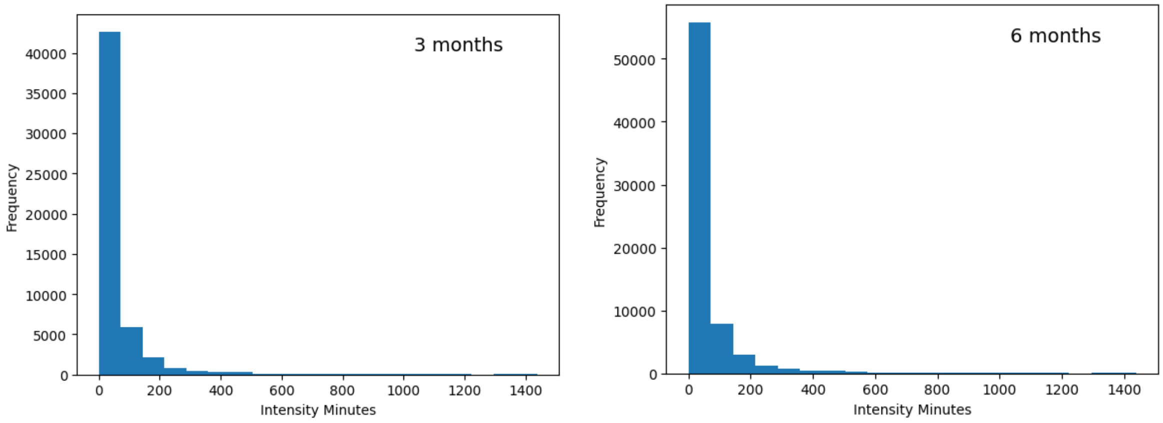


Fig: Frequency vs Intensity Minutes for 3 and 6 month data

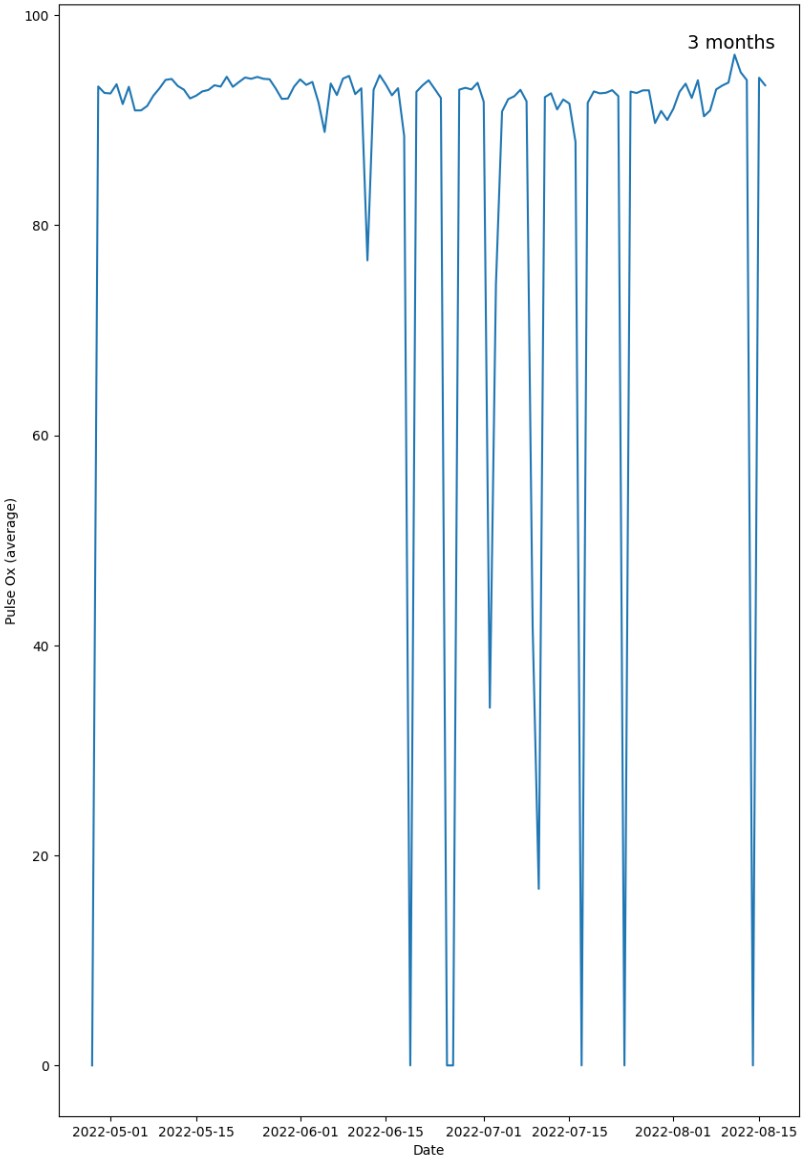


Fig: Average Pulse Ox throughout the 3 months

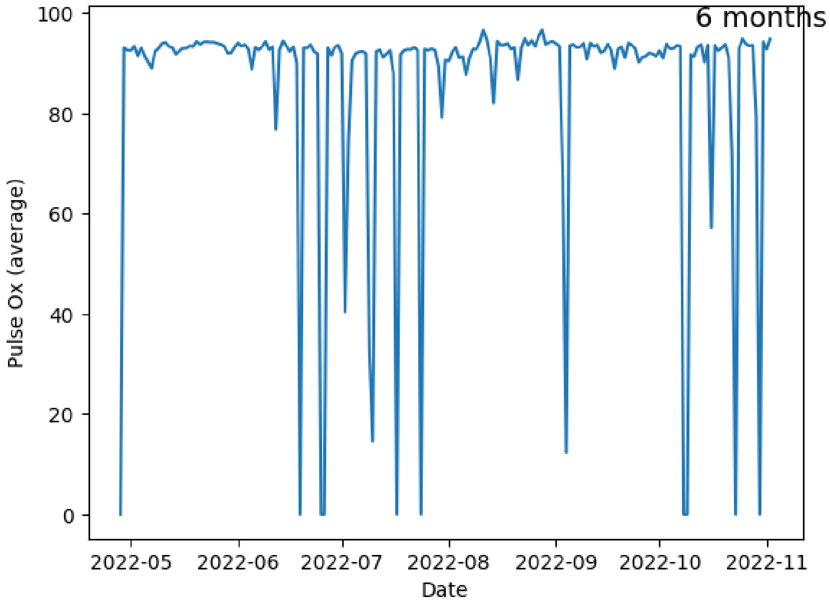


Fig: Average Pulse Ox throughout the 6 months

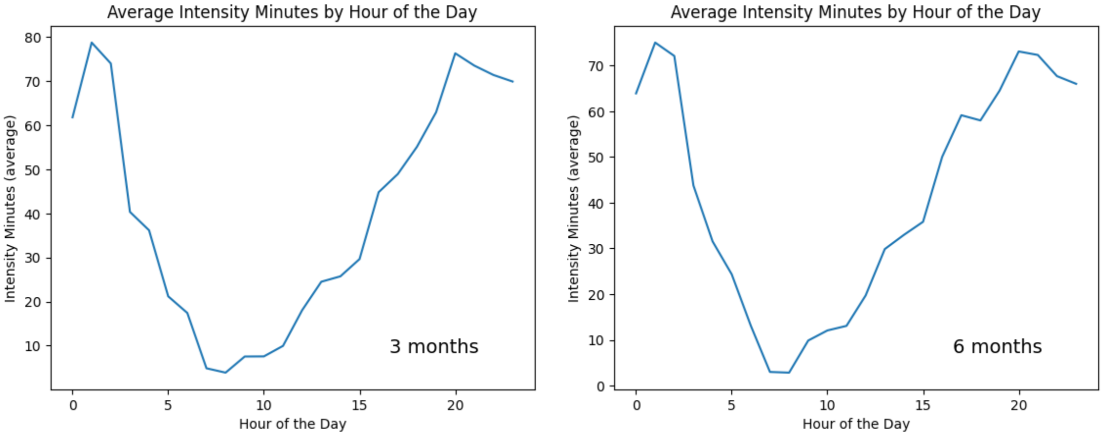


Fig: Average Intensity minutes throughout the day for 3 and 6 months

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.

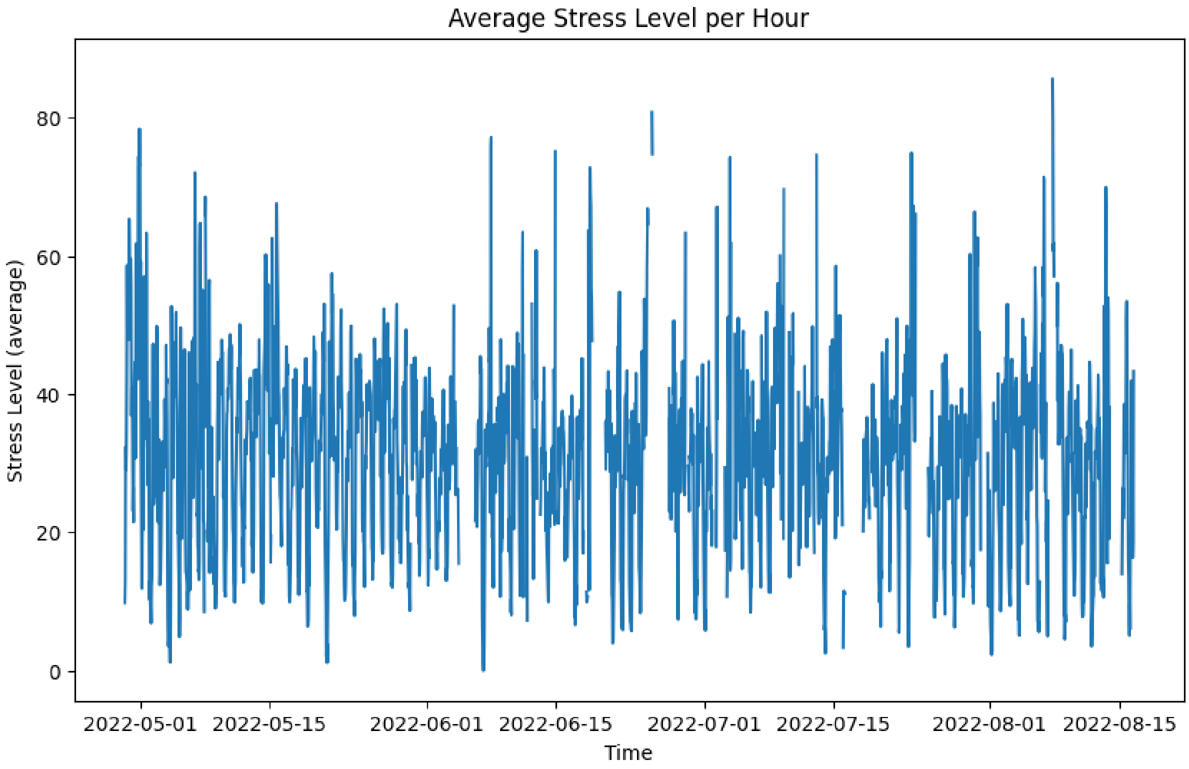


Fig: Average Stress levels per hour

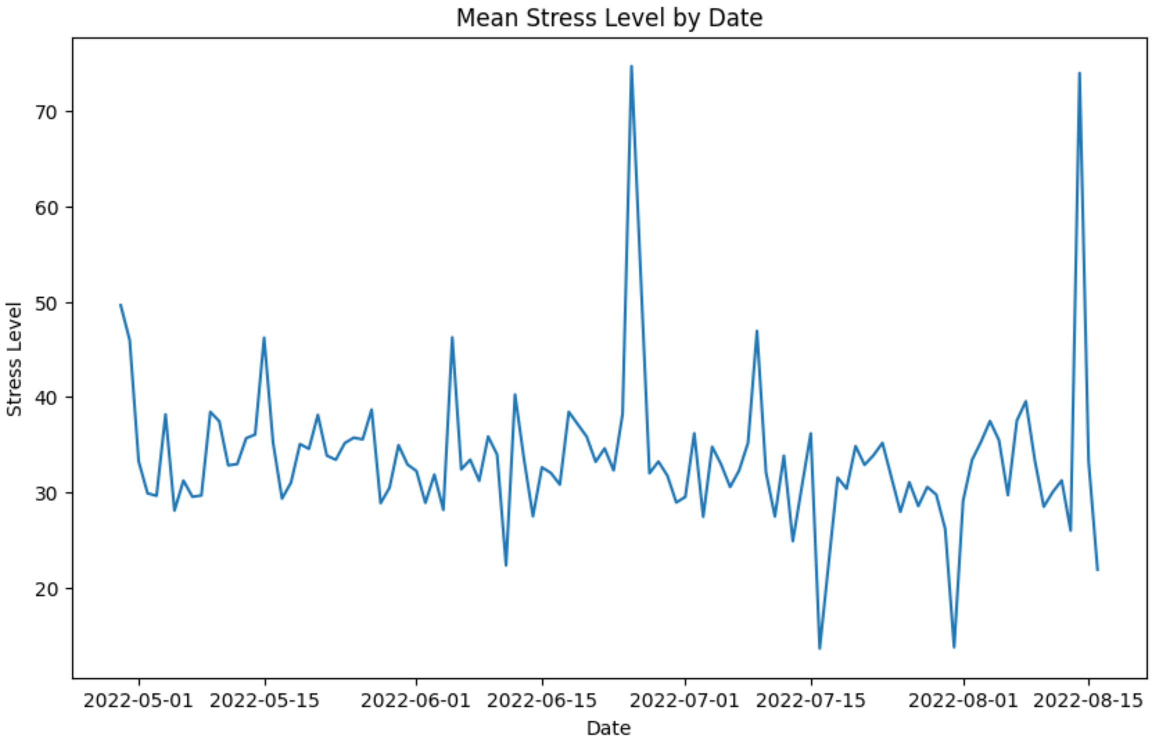


Fig: Average stress levels by date

The report analyzed the ‘stress’ attribute in the Garmin dataset. On analysis of the plots, it was found that the stress levels on average had increased 2 months into the experiment. There was a gradual decrease following which there was a steep increase towards the end of the experiment.

Correlation between stress and discomfort -

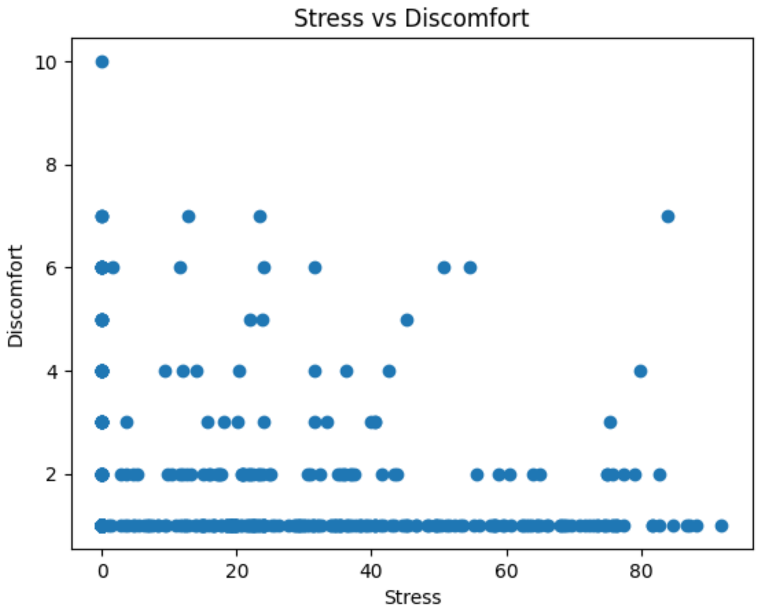


Fig: Average Stress vs Discomfort levels

On analysing the ‘stress’ and ‘discomfort levels’ attributes from the survey data, we found that the correlation score was -0.10199. This shows a negative correlation score. Indicating that level of discomfort does not necessarily relate to stress levels.

6. 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.

Some key points concluded:

- Calories burned and steps taken are typically positively correlated, as both are indicators of physical activity.
- Intensity in minutes is also positively correlated with calories burned and steps taken, as it measures the amount of time spent in moderate to high-intensity activity.
- Heart rate can be positively correlated with physical activity, as the heart rate increases with exercise. However, it can also be influenced by other factors, such as stress or illness.
- Stress levels may be negatively correlated with physical activity and work-life balance, as stress can make it more difficult to prioritize exercise or maintain a healthy work-life balance.
- Floors climbed may be positively correlated with physical activity, as it can be an indicator of taking the stairs instead of using elevators or escalators.

7. Presentation Recording Link

https://bostonu.zoom.us/rec/share/pRlrKofBdzP6YAuoc2w4_2vuIU_IUTZgZnpCTb9MaTK_3KsNnf46rlTCm7iMgq37.Jh9TiPJ_4z7FyxEt