### **Working from the Home Environment & Well-Being Study Data**

Spring 2023 CS506 Data Science

Team 1

Vani Singhal([vanising@bu.edu](mailto:vanising@bu.edu))

Ming-Han Hsieh([jeff0987@bu.edu](mailto:jeff0987@bu.edu))

Wei-Tse Kao([kaowt@bu.edu](mailto:kaowt@bu.edu))

Jessica Woo ([jesswoo@bu.edu](mailto:jesswoo@bu.edu))

**Introduction**

One of the impacts of COVID-19 is the shift from in-person work to working from home. The goal of this project is to determine how transitioning to remote work has affected employees physically and mentally. The study collected data for 6 months from a group of 70 participants, all of whom were full time employees that worked remotely 90 percent of the time and used a computer for at least 4 hours a day. Each participant was given a Garmin watch which was paired with the Ecological Momentary Assessment app (meMA). The study collected physiological data from the Garmin watch such as heart rate, inter-beat interval, stress levels, calories burned, number of steps and floors climbed, and the number of minutes the participant was active for (i.e. intensityMinutes). The meMA app was used to collect data about the participant’s environment and how they behave in that environment by making them answer questions daily, weekly or monthly depending on the survey.

Our role in this project is to find how the physiological data given by the Garmin watch relates to the data collected from the Ecological Momentary Assessment app. In addition, we were responsible for testing a series of hypotheses given to us by the researchers.

**Garmin Data Analysis**

**Cleaning the Garmin Data**

The Garmin data consists of data types - Heart rate, calories, steps, floors Climbed, intensity in Minutes, ibi, pulseOx, stress. In the 3 month folder there are 62 files, one for each participant. Similarly, in the 6 Month folder there are 63 files, one for each participant.

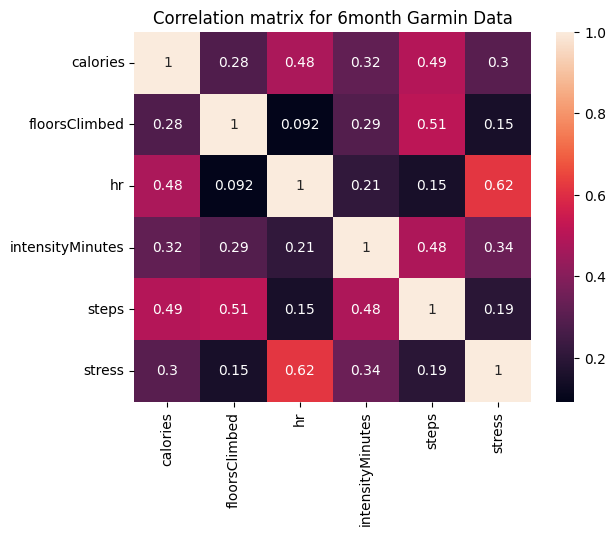
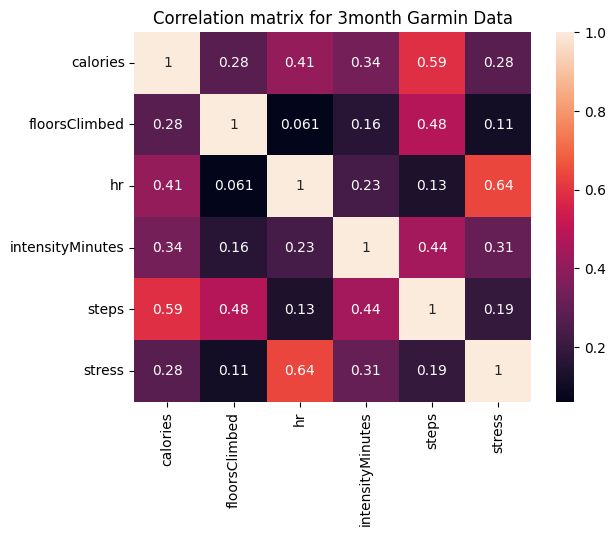
These files also have a lot of 0 values. So it is important to clean them. Firstly we comprehended the Garmin data and noticed that data types - calories, steps, floors climbed and intensity in Minutes was increasing along with the day. So for these data types we took the last value(max value). For data type - heart rate there were 0 values which is practically not possible so we replaced the 0 values with values between 60-100 (normal heart rate). Then we took the average of the values for each day. Along with this we also took the average of Stress values.

Now we have cleaned the data. The last step was to combine each file to one csv file.

**The link of cleaned data is** - <https://drive.google.com/drive/folders/1gHvvxFBsXZCGluVb0R--hfh2K6jsPTsB?usp=share_link>

**Data Exploration on Garmin data types**

We generated a correlation matrix for the 3 month data and the 6 month data to look for potential relationships between the Garmin data types, which are calories, floorsClimbed, heart rate (hr), intensityMinutes, steps, and stress.



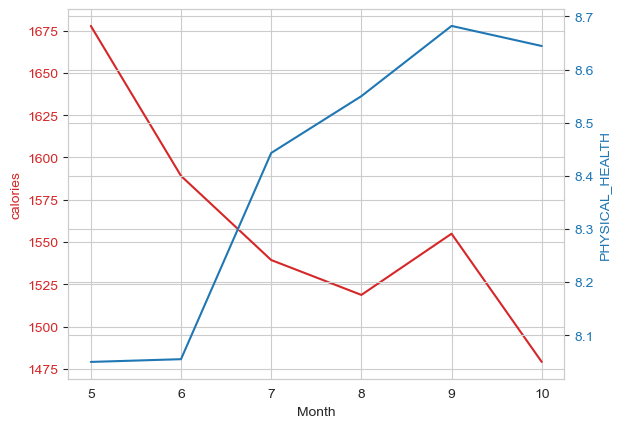
From the correlation matrix, we can see that stress and heart rate have the highest correlation. This makes sense since the Garmin watch determines stress levels based on the inter-beat interval of the wearer’s heart rate.

**Potential Relationships between garmin data & survey data.**

Once we finished cleaning and preprocessing the garmin data, we would like to combine it with the survey flourishing scale and see if there is any potential correlation between them. By doing so, we can know how the participants change from the first month to the sixth month.

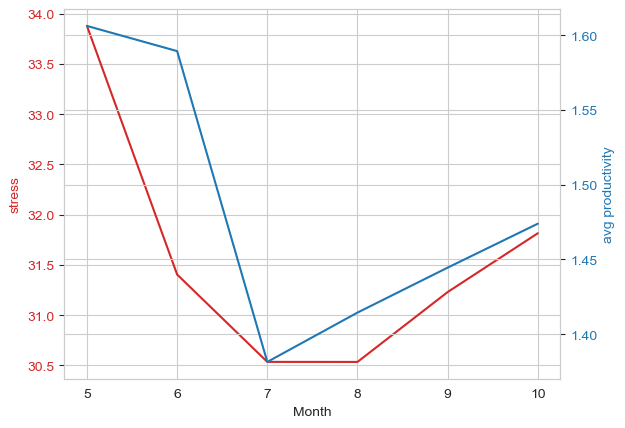
For checkpoint A, we just tested some dataset and made new hypotheses. We are sure we would use more models such as Linear regression t-test in next week’s deliverable 3. Here are some new hypotheses we made.

1. **The calories burned per day has an inverse correlation with physical health.**



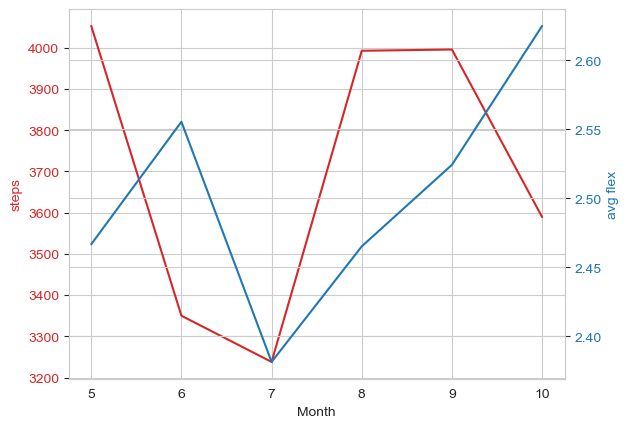
On the right side of the graph (red), the calories data is from Garmin watch. The physical health is from the survey. By observing the graph above, we made an assumption that the calories burned per day has an inverse correlation with physical health.

1. **The average stress is correlated to the Productivity**



On the right side of the graph (red), the stress data is from Garmin watch. The Productivity is the flourishing scale from the survey. By observing the graph above, we made an assumption that stress is correlated to the Productivity.

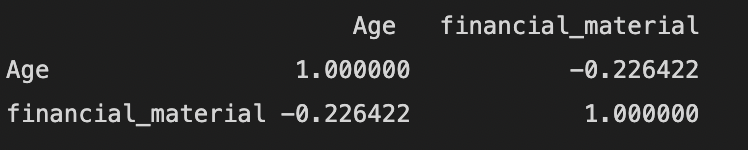
1. **Steps per day is correlated to the Flexibility**



On the right side of the graph (red), the steps data is from Garmin watch. The Flexibility is the flourishing scale from the survey. By observing the graph above, we made an assumption that steps per day is correlated to the Flexibility

**Hypothesis Testing**

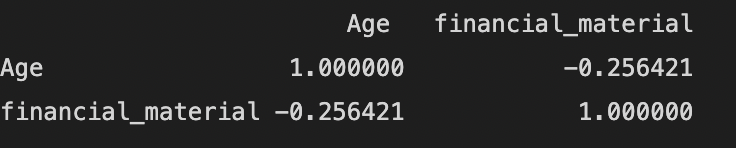
**Hypothesis A**: Participants’ age will negatively correlate with financial and material stability (the last two questions on the Flourishing Scale)



**3 month data (Correlation used here is pearson correlation)**

It can be seen that there is negative correlation between participants age and financial and material stability

With correlation value = -0.22



**6 month data**

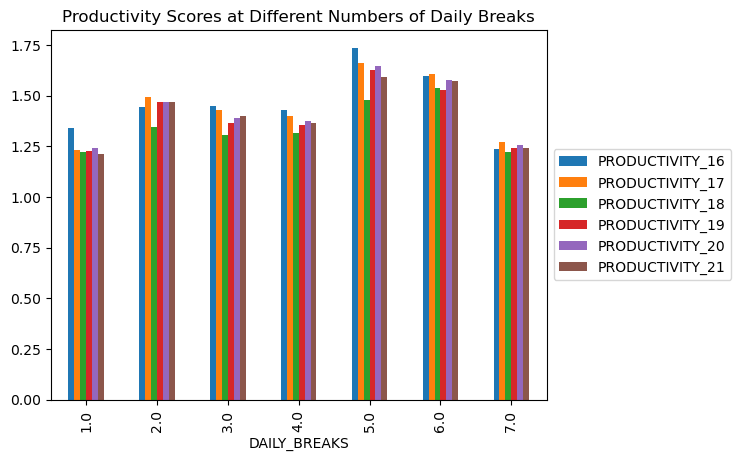
It can be seen that there is negative correlation between participants age and financial and material stability

With correlation value = -0.25

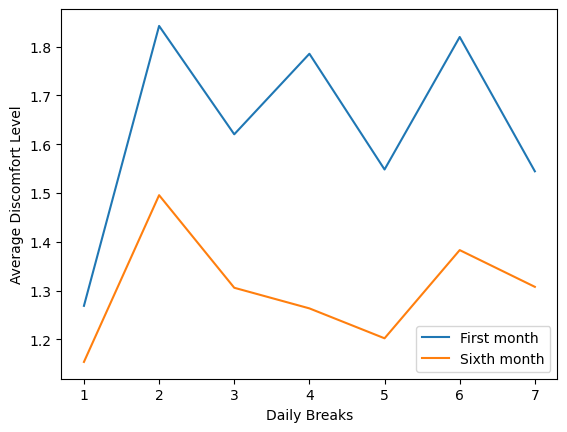
**Conclusion**

This hypothesis is correct for 3 months as well as 6 months data

**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.

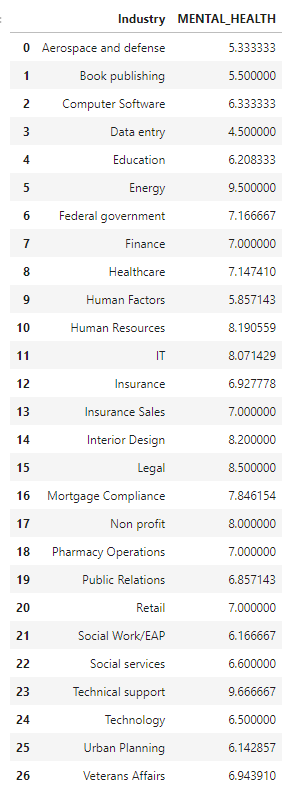
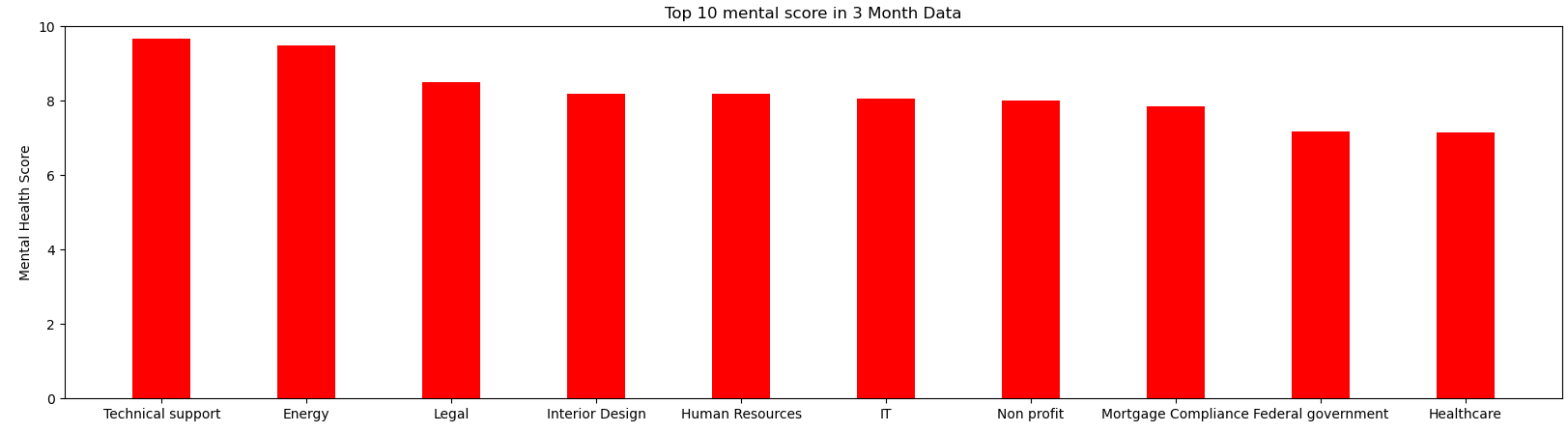
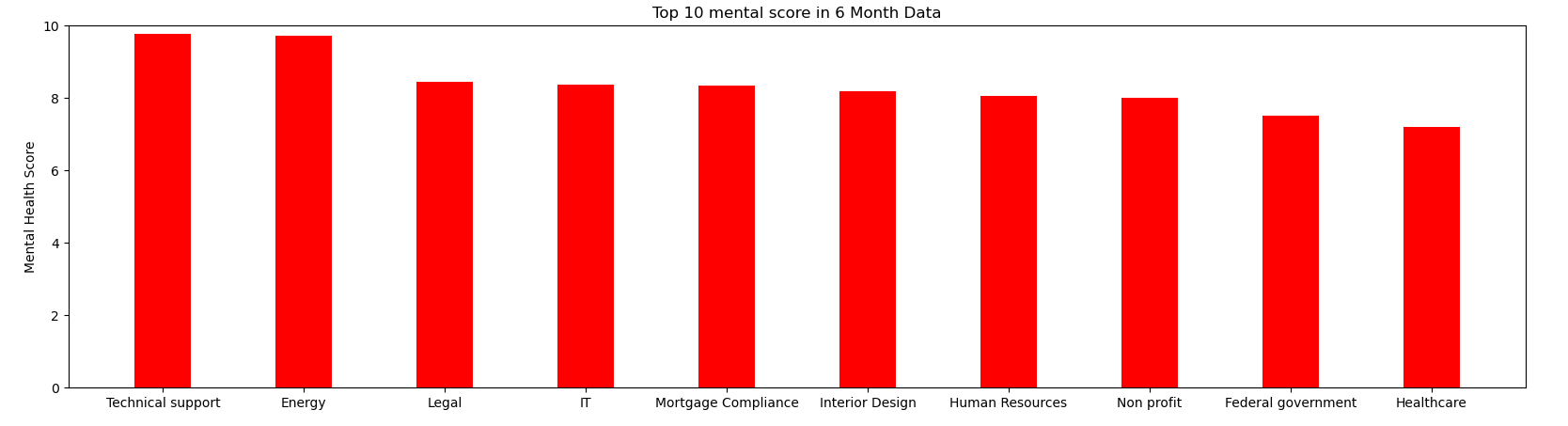


There is a positive correlation between taking an average of 4 daily breaks and productivity scores. The bar graph shows when participants take 4 daily breaks, the average productivity scores are between 1.0 and 1.5, which means that most participants reported better productivity (lower scores mean better productivity, higher scores mean worse productivity).

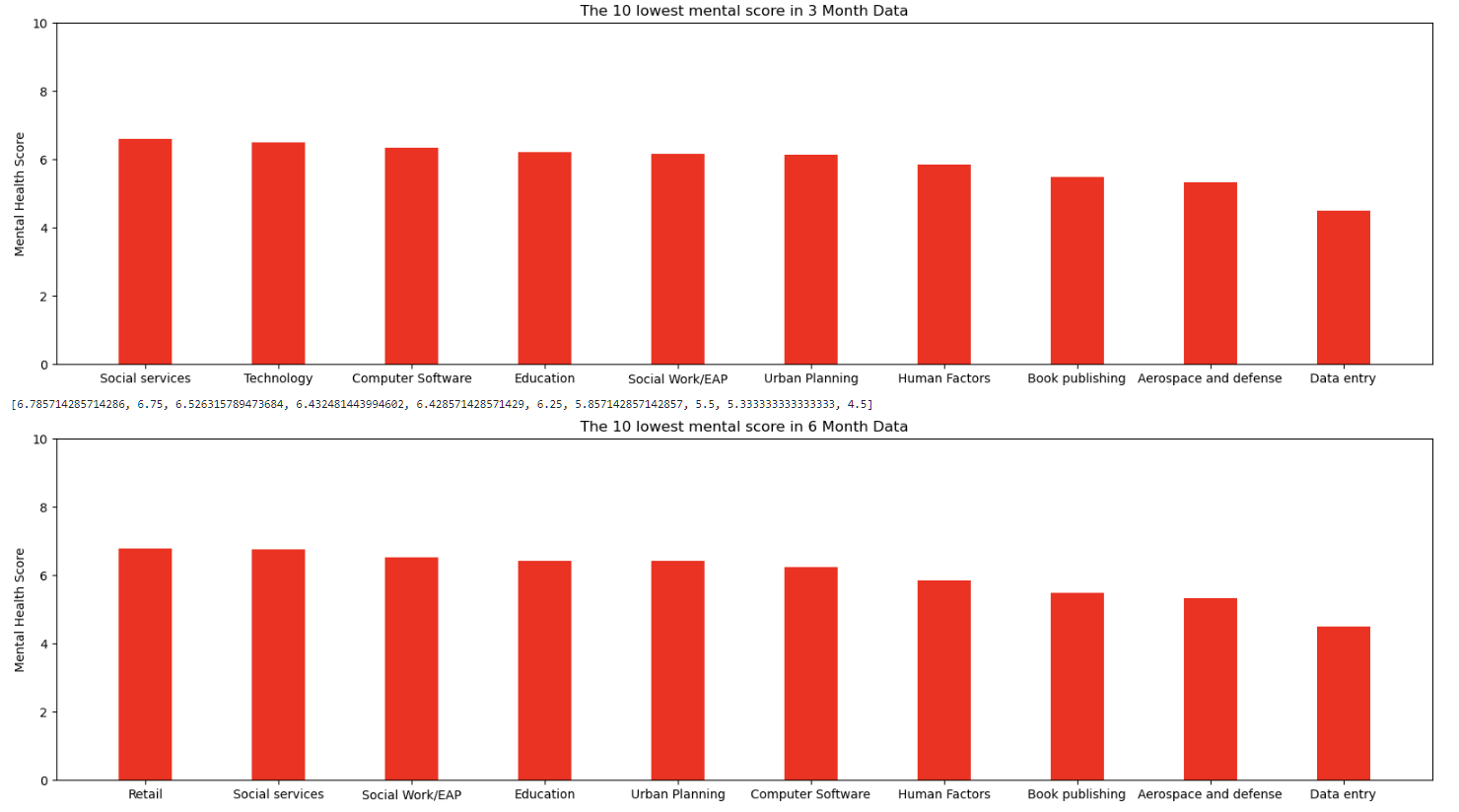


Participants who take 4 daily breaks experienced a higher discomfort during the first month, but had a lower discomfort level during the sixth month. The graph also shows that there was an overall decrease in discomfort from the first month of working from home to the sixth month, which suggests that the participants got used to working remotely. Participants who took 2 daily breaks and 6 daily breaks experienced the most discomfort for each month.

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

**

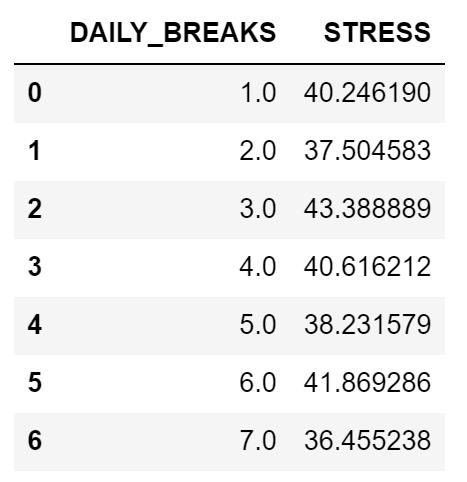
Totally, there are 26 unique industries in our dataset. We plot the top 10 industries that have the highest mental score. However, in both 6 month and 3 month data, we can observe that Healthcare has the 10th highest mental score in 26 industries. Therefore, we can confirm that the hypothesis is not valid. So maybe we take a look at what kind of jobs would really have the lowest mental score:



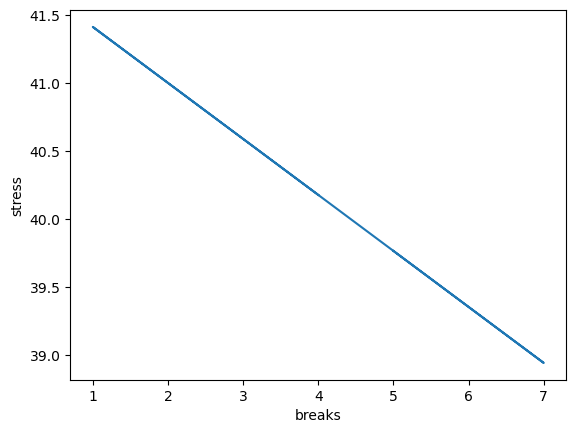
Here are the industries that have the lowest mental score. We can observe that the lowest mental score industry is Data Entry in both 3 month data and 6 month data.

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

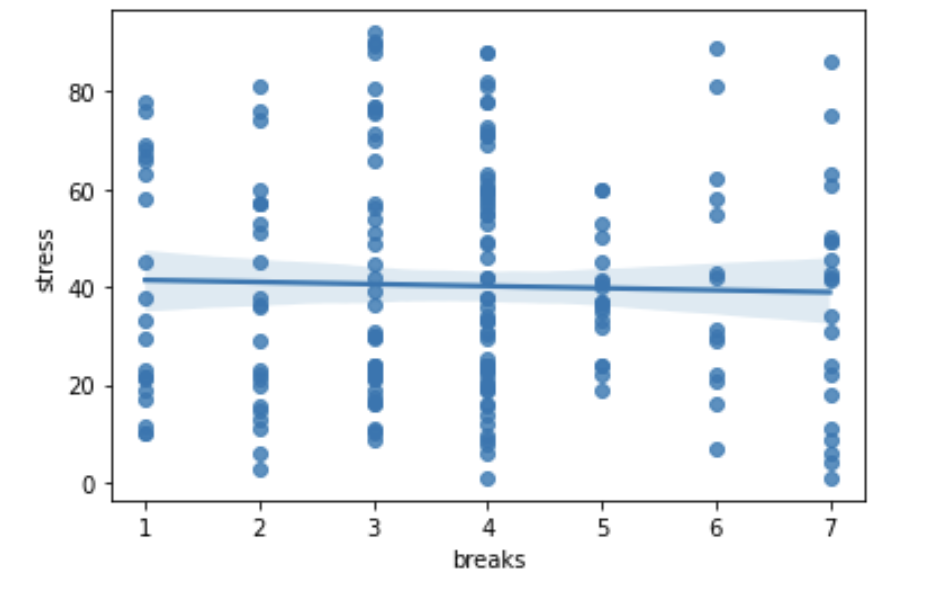
**Preprocessing:**

We dropped all the 0 stress value since we can’t confirm that the value is actually 0 or is missing. 

Then we plot the linear regression line when daily breaks in the x axis and stress in y axis.



We can clearly see the inverse relationship between daily breaks and stress value. However, if we take a deeper look at the data, we would find out that the data is pretty distributed. That is, the correlation is not as strong as we thought.



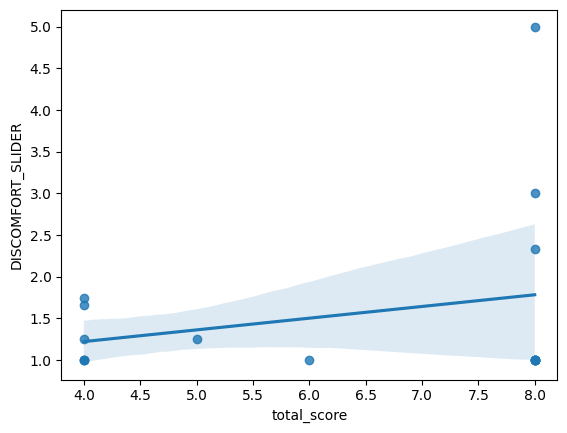
**The conclusion:**

The correlation between participants’ stress levels and the number of breaks is slightly inverse.

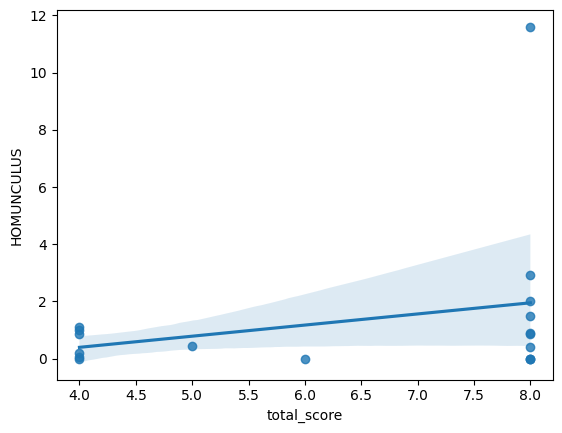
**Hypothesis e**: Based on question #15 in the Computer Workstation Checklist (with 4 responses regarding ergonomics training), participants with lower scores will report less pain at 6-months.

Pre-processing steps:

1. Extracted the six month data from *Daily PM, Friday AM,* and *Computer Workstation Checklist*
2. Extracted the columns that represent question 15 in *Computer Workstation Checklist*
3. For *Daily PM*:
   1. Categorized the entries in “OTHER\_DISCOMFORT” to locations the participants experienced pain. For example, if the entry was “Stomach and chest” it would have the value True under the columns “stomach” and “chest”.
   2. In the “HOMUNCULUS” column, I extracted all of the integers into a list and got the sum of that list.
   3. Dropped rows if all of the custom pain location categories had NaN values or if the “HOMUNCULUS” column had a NaN value.
   4. Merged with the question-15 data
4. For *Friday AM*:
   1. Grouped the data by “mbl\_cod” and calculated the mean discomfort level for each participant (since some participants had more than one entry for “DISCOMFORT”)
   2. Dropped NaN values
   3. Merged with the question-15 data



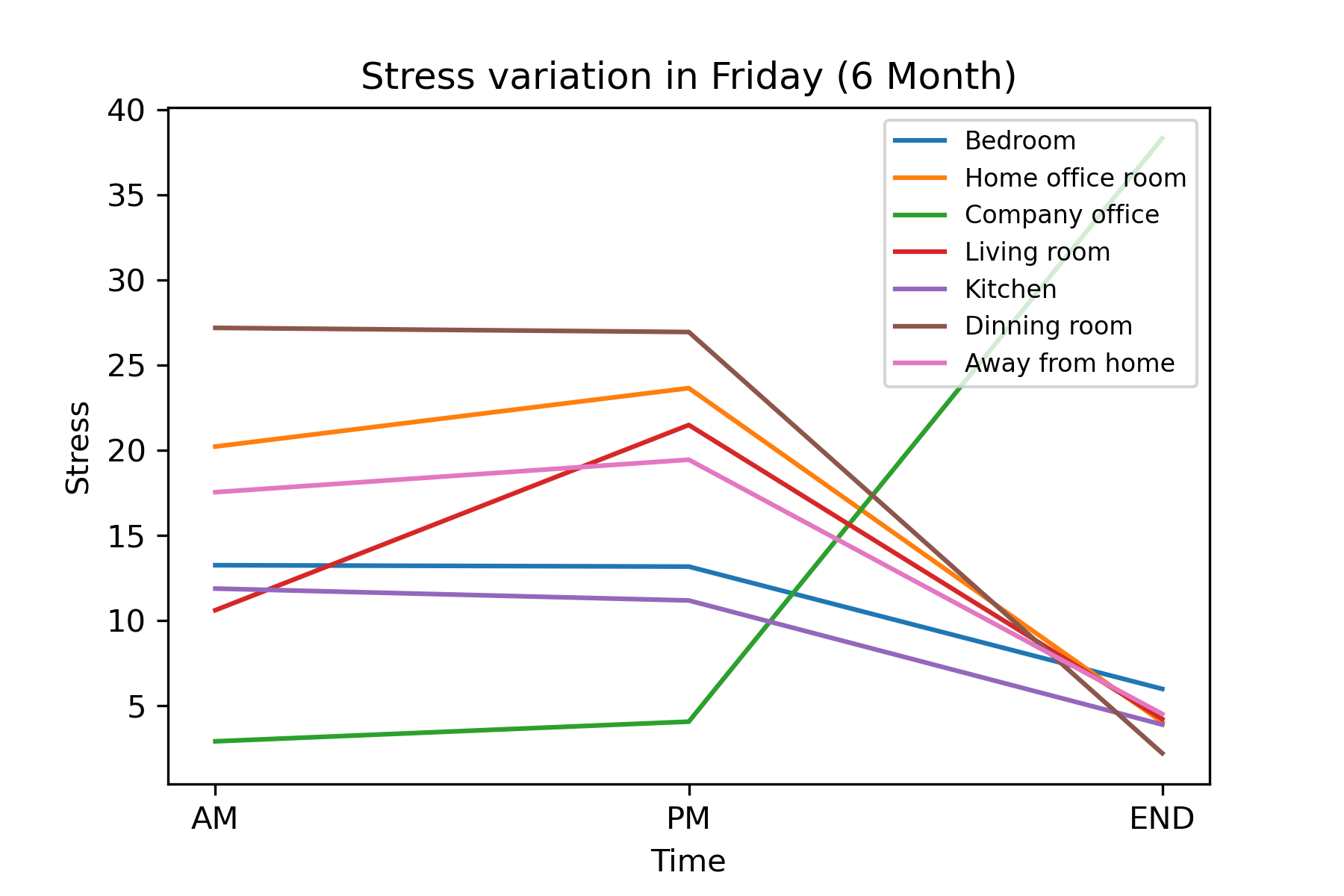
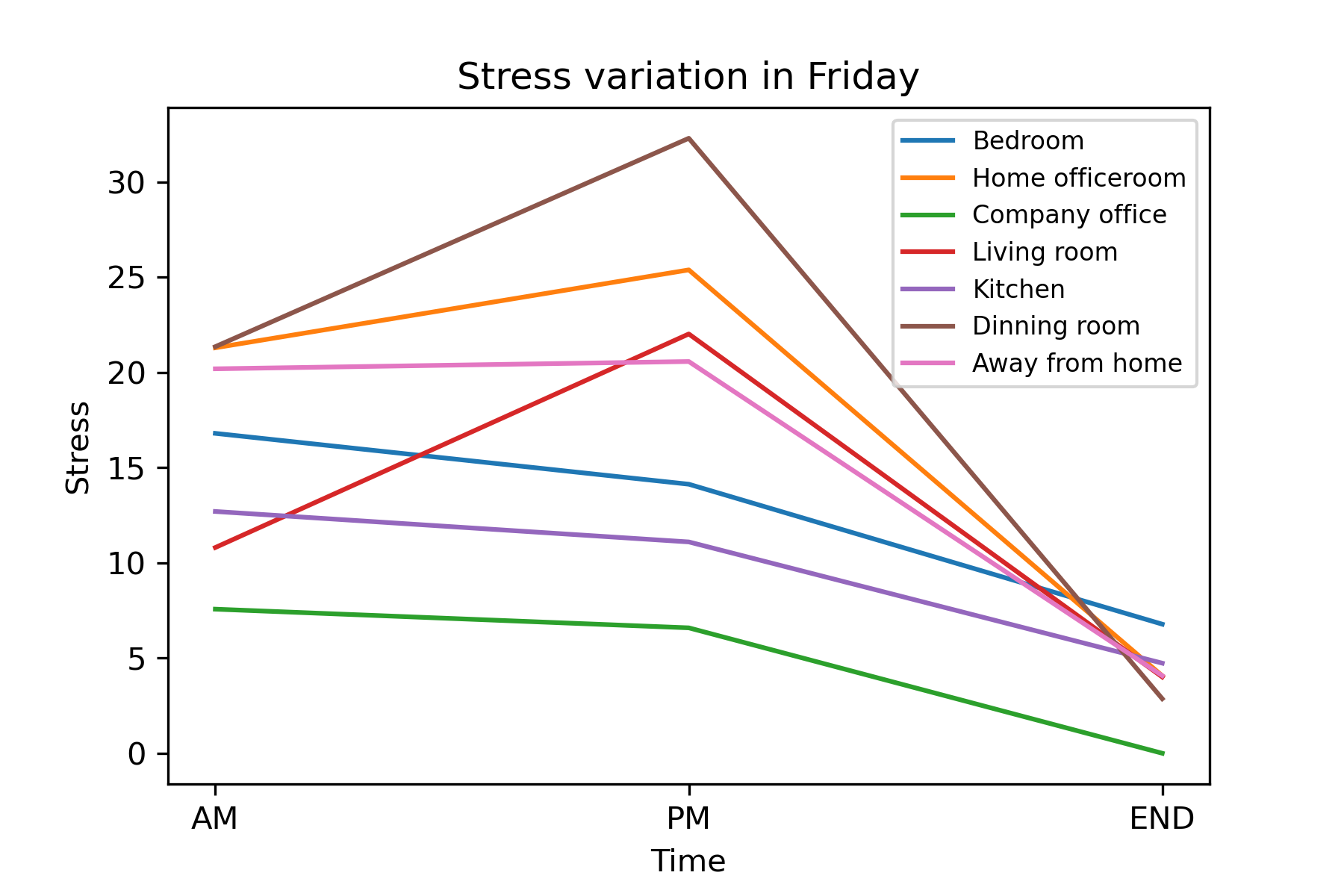
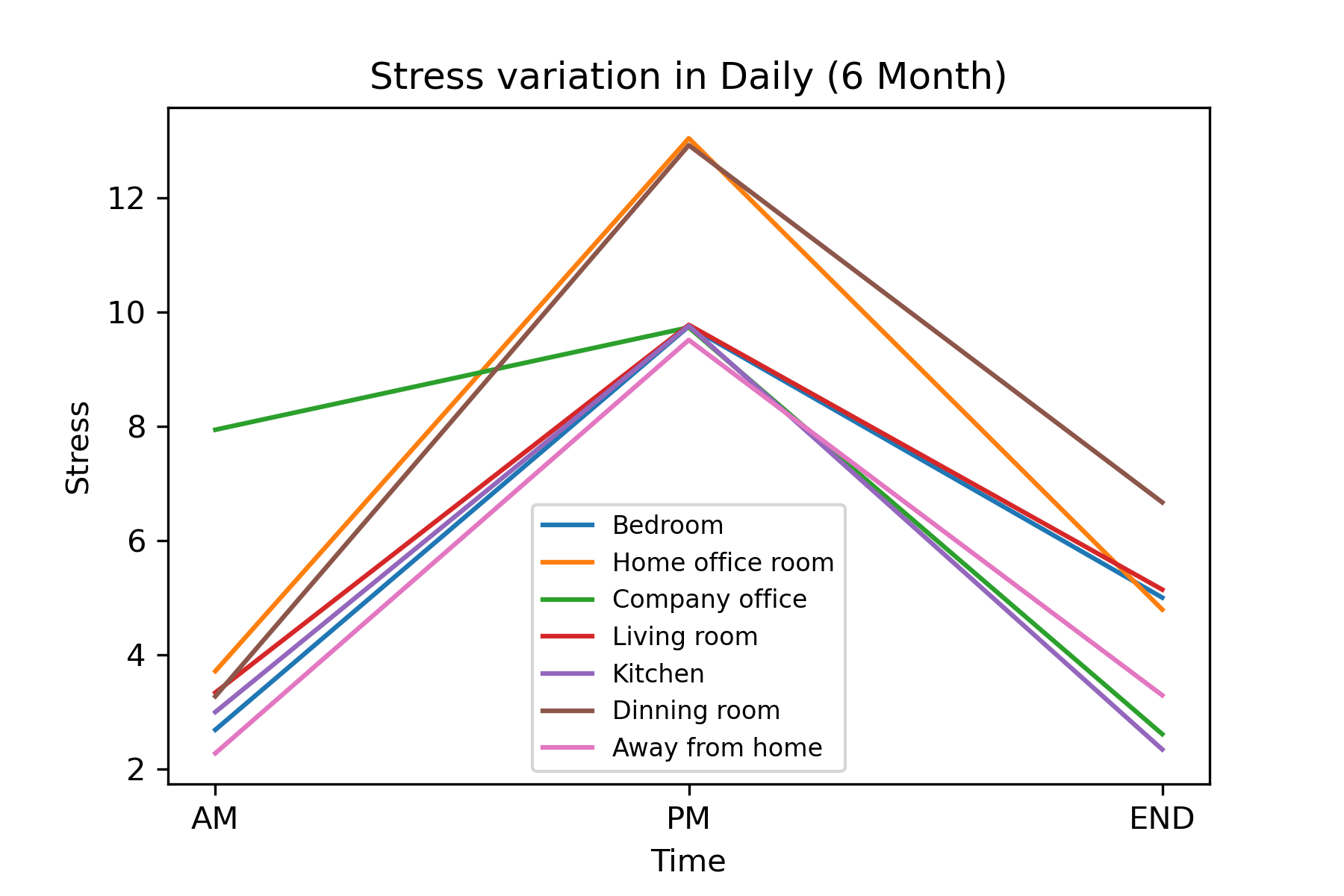
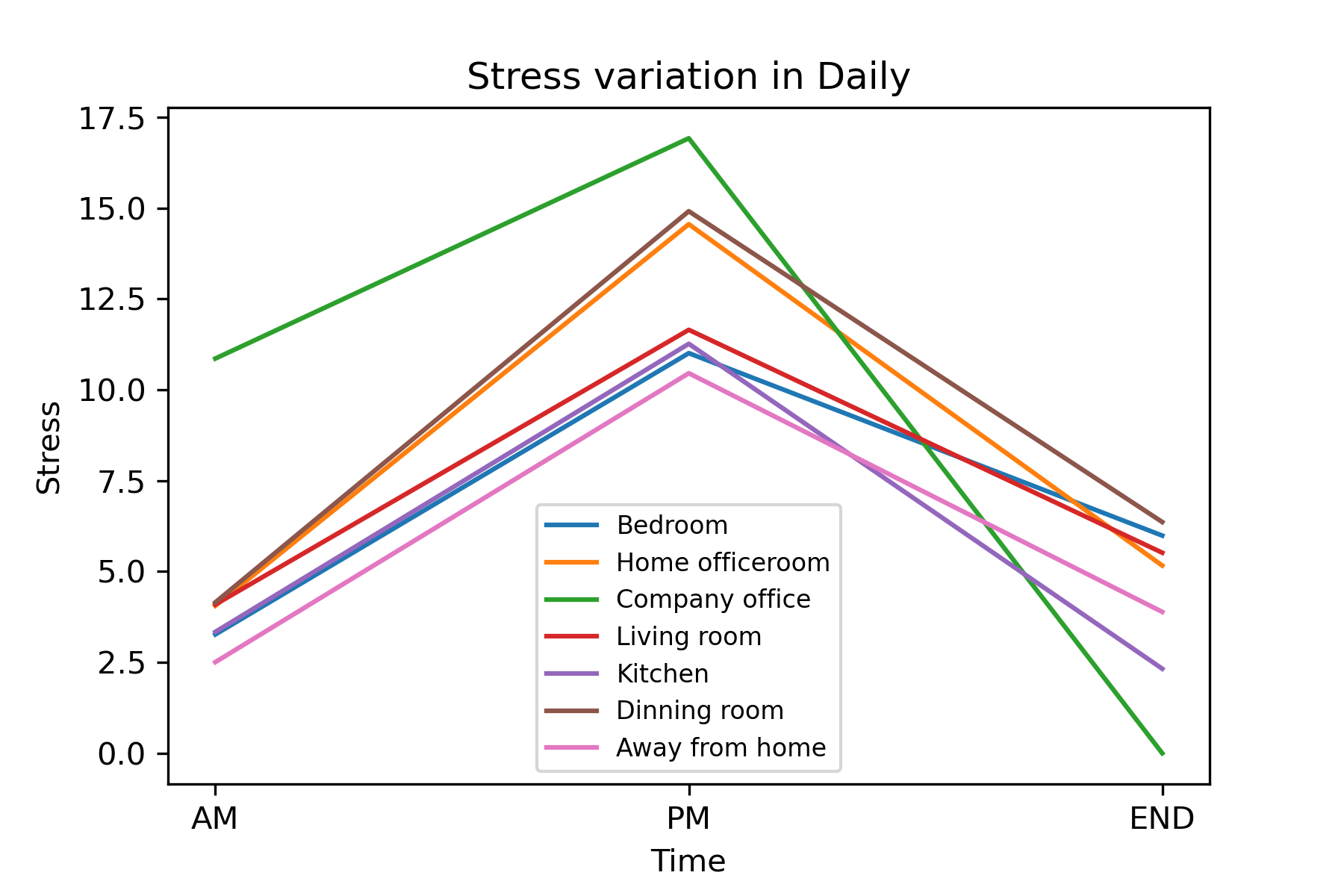
The regression line shows that there is a direct relationship between scores on question 15 and average discomfort level. Therefore, participants with lower scores will report a lower discomfort level.



The regression line shows that there is a direct relationship between question-15 scores and average homunculus scores. Therefore, participants with lower scores will report a lower homunculus score (i.e. less pain)

**Analyze and Answer**

iv. Where are people with their stress levels? Using the data from the Garmin watches as well as the survey data, are stress levels Increased, decreased, no change?



I utilized the “3 month” and “6 month” data in the day(“AM”, “PM”, “END”) including “Daily” and “Friday”, and analyzed them to plot line charts. We can observe the stress level of these locations.

In a day, the stress level will increase at “PM”, and decrease at the “END” of a day.

Comparing daily data in “3 month” and “6 month”, we can find that the stress level at “Home office room” decreases significantly, but the stress level of other locations doesn’t change apparently.

In “6 month data, we can observe that the stress level in Friday data doesn’t change significantly between “AM” and “PM”, and the stress level of all locations decrease at the “END” of the day except the stress level at “Company” office. However, I find that there is only one stress data at “Company office” in the dataset, so the result of the stress level at “Company office” in “6 month” day seems to be uneven.