Final Deliverable:
Working from the Home Environment & Well-Being Study Data

Author: Team 3
Semester: Fall 2022
Date: 12 December 2022

Table of Contents

Hypothesis 1: “Participants who take an average of 2 or fewer breaks per day will report more pain and less comfort than those who take an average of 4 or more breaks per day“.

Hypothesis 2: “Participants who have an average of 3 locations per week or less will have higher stress algorithms than people who use an average of 4 or more locations per week. **”**

Section 1: Data Collection and Manipulation

1.1: Data Collection

Throughout the four months of analysis, we were able to continue receiving new data as the Work From Home study is ongoing. At the time of beginning this project, we had up to four months of data but by the end we were able to receive all of the data possible from the study (six months of data). For the purposes of this paper, we used the six-month data for all of our analysis as it included all of the data from the beginning of the study.

While new data arrived, the structure and format of the data remained the same. There was Garmin data which was data from a watch which tracked the individual's movement, heart rate, stress etc.

Then, there was data in the format of surveys that each participant took. Every weekday, each participant was subjected to completing three surveys a day: one in the morning, one in the afternoon, and one at the end of the day. The results of these surveys were given in the form of csv files. These daily surveys gave us the location of the individuals, their discomfort level, and the areas they felt pain.

In the AM dataset, we noticed that there were several columns filled with only 0s. These columns were RESPIRATION, BODY_BATTERY, STEPS, CALORIES, FLOORS, INTENSITY_MINUTES. As these would provide no additional information when performing data analysis, it makes sense to drop these columns for future analysis. We also noticed that the heart rate data and stress data also had many rows with missing values.

1.2: Preprocessing Data

In order to preprocess the data, we looked at many different methods. In the beginning, we replaced all NaN values with 0 but this heavily affected our analysis as it would drag down the average of various columns. We then tried substituting the values with the average of the columns, however, this also did not make sense as it rendered all individuals the same value despite them having distinct nuances. We then finally settled on simply removing all NaN values as it was non-invasive. The drawbacks of such an approach is that we lose a lot of data. However, due to the sheer amount of data that we had, the data lost was negligible and did not affect our analysis at all.

1.3: Data Manipulation

For the second hypothesis, a great deal of manipulation for the data is required. Due to the nature of the hypothesis, it was required to get the average stress value for a certain individual every week of the study. Hence, the Garmin data was used to get over sixteen million

data points relating to the stress of all the individuals over the six month period. In order to properly preprocess, we had to ensure that we separated the data into weekly parts for each individual. In order to do this, the local_time field was used which was a timestamp for the data. Using that, we separated into the date which we then were able to assign a week number. This process only highlighted the process to get the stress portion. In order to get the number of locations, the AM, PM, END, FRIDAY_AM, FRIDAY_PM, and FRIDAY_END datasets were used. These datasets had to all be merged by the id of the person and the week number. Afterwards, we counted the number of locations used in that week throughout the six datasets and this number was a new column in the dataset. Finally, the garmin data and the six datasets were merged into one final dataframe which gave us the two fields we needed: number of locations and stress values.

1.4: Data Visualization

We performed a heatmap matrix, as displayed in figure 1, that describes the AM data set by displaying the correlation between each data column. The result shows that column has the most significant correlation: Longitude vs latitude and heart rate vs Stress.

For the longitude and latitude column, the result correlation is 0.95, which is very significant. However, these are insignificant because it is expected that location variables will have correlation with one another. On the other hand, the correlation between Heart_rate and Stress is around 0.52. This is also expected.

Figure 1

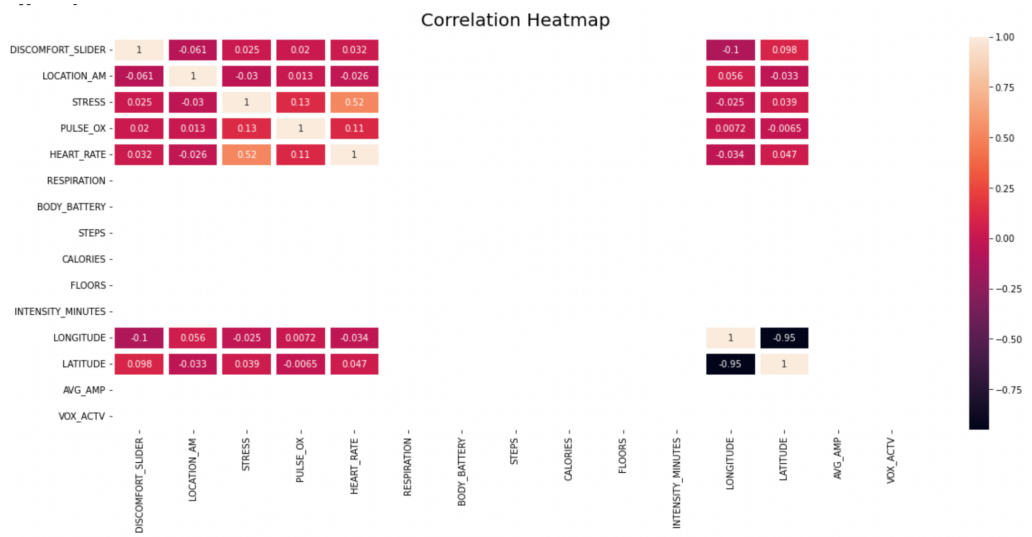


Figure 1: This figure simply shows the heatmap representation of all the variables in the AM dataset (which is very similar to PM and END datasets).

Section 2: Hypothesis 1

2.1: Hypothesis Declaration

Hypothesis: "Participants who take an average of two or fewer breaks per day will report more pain and less comfort than those who take an average of four or more breaks per day"

2.2: Preliminary Hypothesis Analysis: Naive Approach

After conducting preliminary data processings, we determined the differences between those that take more than four and those that take less than two. The results were extremely minor. We first approached by summing up all the pain data for each individual and this was columned as "Total Pain". Figure 2 shows a scatter plot of all the individuals in the study and their average number of breaks vs the total summed pain that they felt.

Figure 2

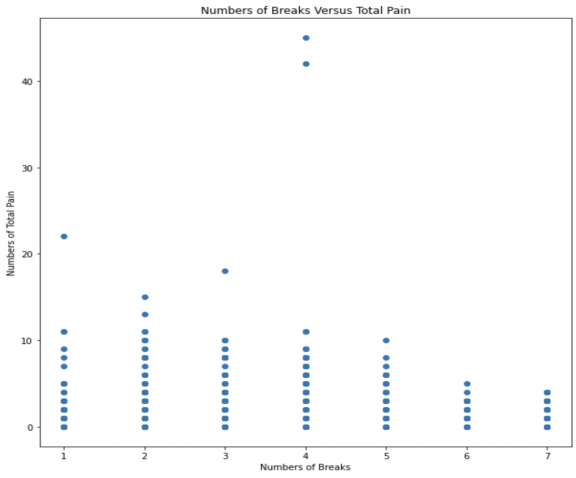


Figure 2: This figure shows a scatter plot where the x-axis represents the number of breaks and the y-axis represents the total summed pain.

After that, we formed two groups: two or fewer breaks and four or more breaks by averaging the total pain for every individual in each group.

The "Total Pain" for two or fewer was recorded as 1.5695 whereas the "Total Pain" for four or more was recorded as 1.145. This value did not seem negligible at first but further analysis in the form of hypothesis testing was required.

2.3: Paired t-tests

In order to further our analysis regarding the difference between the two groups, we utilized paired t-tests**.** Paired t-tests are statistical tests that note if there is a significant difference between two groups. In this case, the two groups are whether one took two or fewer breaks or

whether one took four or more breaks. Figure 3 below shows the two group distributions separated in groups.

Figure 3

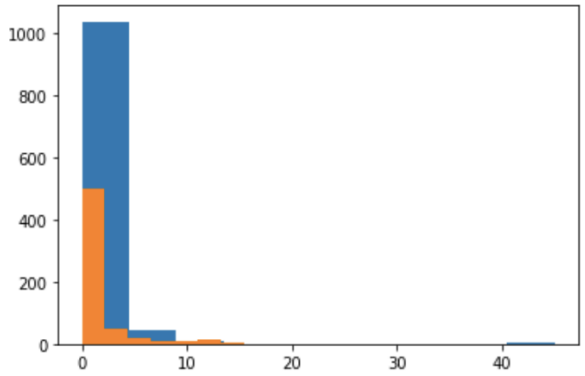


Figure 3: This shows a bar graph of the distribution of data for both groups. Blue represents four or more breaks while orange represents two or less.

Due to the fact that there was more data for four or more breaks, we need to use the Welch t-test as that is primarily used when there are differences in sample sizes/variances. The hypothesis we would be testing is whether or not the two groups have equal population means. In other words, do people who take two or less breaks have the same amount of pain as those who take four or more breaks. This will be our null hypothesis. If there is sufficient evidence to reject the null hypothesis, we will be able to conclude that there is a significant difference between the two groups in terms of pain. We will use an alpha level of 0.05. After conducting the Welch t-test, the results we obtained are shown in figure 4.

Figure 4

```
print(welch_ttest(twoorless6,fourormore6,"equal"))
```

```
   T statistic         df  pvalue 2 sided  Difference in mean        lb  \
0     3.237706  1175.6565        0.001239            0.424582  0.167294

         ub
0  0.681871
```

Figure 4: Results of Welch t-test

We note that we reject the null hypothesis at the alpha level we specified of 0.05. This is because the p-value is 0.001239 which is significantly less than 0.05. Thus, we would conclude that the difference in pain is significant and since the group with two or less breaks had much more pain, we would conclude that those who take two or less breaks experience more pain than those that take four or more breaks.

2.4: Outlier Analysis

When taking a closer look at figure three, we notice that there are some outliers that are skewing the distribution to the right. Thus, figure 5 shows the updated distribution of data after removing outliers by utilizing the 1.5 IQR test.

Figure 5

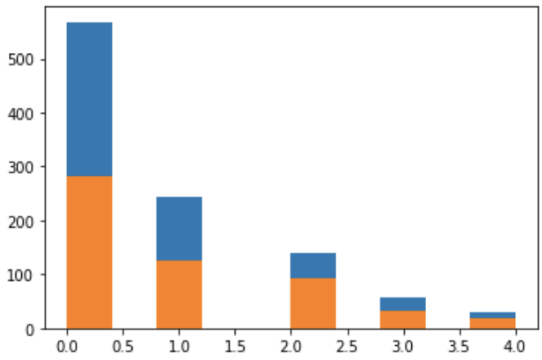


Figure 5: This shows a bar graph of the distribution of data for both groups after removing outliers. Blue represents four or more breaks while orange represents two or less.

Now, after re-running the tests, the results of the Welch t-test are given in figure 6.

```
print(welch_ttest(newtwoorless6,newfourormore6,"equal"))
```

```
   T statistic           df  pvalue 2 sided  Difference in mean        lb
0     1.781869  1075.603066        0.075053             0.10211 -0.010332

         ub
0  0.214552
```

Figure 6: Results of Welch t-test after removing outliers

Here, we note that we would not reject the null hypothesis at an alpha level of 0.05. However, by removing outliers, we notice that we are significantly reducing the range of pain levels reported. The max pain level that can be reported now is simply 4 whereas there were several candidates who reported significantly more pain. As a result, it is our opinion that we should not remove the outliers in this case as doing so will significantly reduce the range of pain levels reported.

2.5: Individual Participant Analysis

Now, a next level analysis would be conducting these significance tests for each individual. By doing so, we are able to eliminate more variance as we are now conducting a t-test for each individual to determine which individuals are actually statistically significant. The results of these individual tests are shown in table 1 below.

Table 1

Statistically Significant IDs	Statistically Insignificant IDs
11822993	23916703

Statistically Significant IDs	Statistically Insignificant IDs
17180706	32937810
20763027	33075391
22541511	34865333
27148444	37720972
34633705	49164240
36505757	49669568
38876664	55508636
47443793	58805130
54042771	60404747
57026233	61881920
58601340	66958688
64811087	66999191
70975009	69497234
71552354	71681441
81875100	73262082
91556555	77253909
	80515680
	81862952
	86548395
	93909901

Table 1: This table shows the statistically significant and insignificant individuals

The individuals not shown on table 1 did not have sufficient information to conduct a valid t-test. As a result, those have been left off the list. We notice that there are many individuals who are statistically significant. What this means is that these individuals experienced a difference in pain that was statistically significant when they took two or less breaks or four or more breaks.

2.6: Similarities between Statistically Significant Individuals

The main use-case for having the statistically significant individuals would be to note any key differences between these individuals and those that were not statistically significant. This analysis was done on the following variables: Discomfort, Life Satisfaction, Happiness, Physical Health, Mental Health, Worthwhile, Purpose, Promote_Good, Content Relationships, and

Satisfying Relationships. After conducting the tests for each of these individuals between the two groups, the covariates which were statistically significant were Physical Health, Mental Health, Worthwhile, and Purpose. What this highlights are the key differences between the two groups of interests: those that were statistically significant and those that were not.

2.7: Conclusion

Through these various tests, we noted several key findings. For starters, when we didn't remove outliers we noticed there was a statistical difference between the group that took two or fewer breaks as opposed to those that took four or more breaks. In addition, we took note of the IDs that were statistically significant and were able to give key areas to look at within the statistically significant individuals as those were the key similarities to look for in the future.

Section 3: Hypothesis 2

3.1: Hypothesis Declaration

Hypothesis: "Participants who have an average of 3 locations per week or less will have higher stress algorithms than people who use an average of 4 or more locations per week. "

3.2: Preliminary Hypothesis Analysis

After conducting some data visualization techniques, we formed a box-and-whisker plot of both groups: three and below locations and four and above locations. Below is the box-and-whisker plot of the distribution of the data.

Figure 7

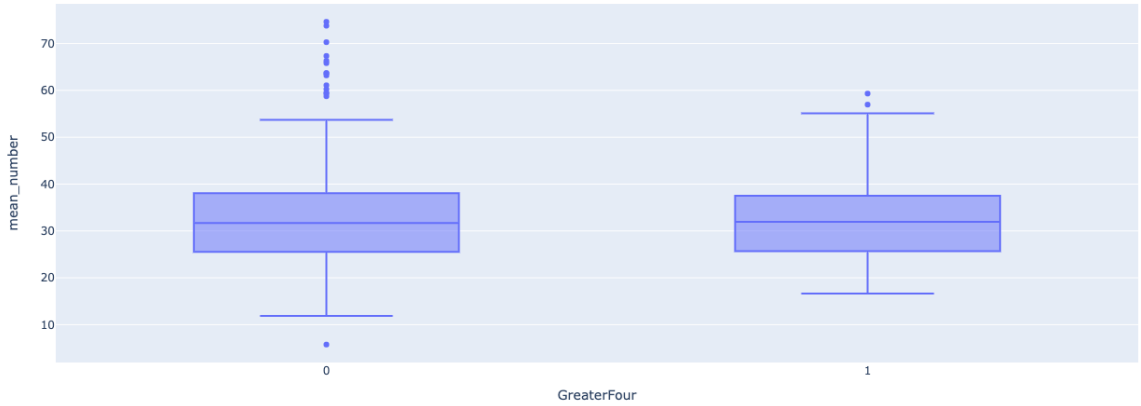
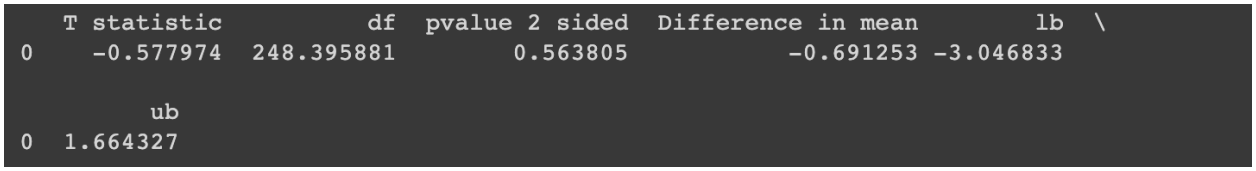


Figure 7: The following box-and-whisker plot has the mean stress values for each week on the y axis and has the two different buckets: greater than or equal to four breaks (1) or less than four (0).

3.3: Paired t-tests

Similar to the previous hypothesis, we utilized Welch t-tests in order to prove statistical significance. In this case, the null hypothesis is that the two groups' populations are exactly the same. In other words, it does not matter whether one worked in three or less locations in a week or four or more locations. The results of the t-test are given in figure 8 below. We note that we were unable to reject the null hypothesis. This means that there was not sufficient evidence to prove that the two population means were statistically different.

Figure 8



```
   T statistic          df  pvalue 2 sided  Difference in mean        lb  \
0    -0.577974  248.395881        0.563805           -0.691253 -3.046833

         ub
0  1.664327
```

Figure 8: Results of Welch t-test

3.4: Individual Participant Analysis

Similar to the previous hypothesis, we conducted a Welch t-test for each individual to note statistically significant individuals. The results of this test are shown in Table 2.

Table 2

Statistically Significant IDs	Statistically Insignificant IDs
11822993	22541511
38656882	27148444
58601340	32455277
60404747	32937810
64811087	33075391
77253909	34865333
	35549180
	37720972
	38876664
	51755925
	56954906

	57026233
	61307863
	66958688
	66999191
	69497234
	79316883
	81862952
	86548395

Table 2: This table shows the statistically significant and insignificant individuals for the second hypothesis

The individuals not shown on table 2 did not have sufficient information to conduct a valid t-test. As a result, those have been left off the list. We notice that there are many individuals who are statistically significant. What this means is that these individuals experienced a difference in stress that was statistically significant when they worked in three or less locations or four or more locations.

3.5: Similarities between Statistically Significant Individuals

Similar to the first hypothesis, the main reason that we found statistically significant individuals is to note any key differences between these individuals and those that were statistically insignificant. The analysis was done on the following variables: Discomfort, Life Satisfaction, Happiness, Physical Health, Mental Health, Worthwhile, Purpose, Promote_Good, Content Relationships, and Satisfying Relationships.

3.6: Conclusion

Through these various tests, we noted several key findings. For starters, we noticed that the two groups (three or less locations and four or more locations) were not statistically different when it came to the dependent variable stress. In addition, we took note of the IDs that were statistically significant and were able to give key areas to look at within the statistically significant individuals as those were the key similarities to look for in the future.

Section 3: Reproducibility

In order to reproduce our results for hypothesis one, simply navigate to the Ipython Notebook labeled Final Deliverable Code P1.ipynb. Then, navigate to the six month data and run all these cells. Similarly, to reproduce our results for hypothesis two, simply navigate to the Ipython Notebook labeled Final Deliverable Code P2.ipynb. Then, simply run all the cells to reproduce the results.