Project Deliverable 3:
Working from the Home Environment & Well-Being Study Data

Author: Team 3

Semester: 2022 Fall

Date: 5 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 will have higher stress algorithms than people who use an average of 4 or more locations per week. **”**

I. Data Collection

We now have the second month data, the third month data, and the sixth month data collected using surveys.

We have looked at the basic structure of the dataset on AM, PM and the END OF DAY daily data in a 3 month dataset. In particular, we worked on the AM and PM data and thoroughly analyzed these datasets.

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.

In addition to these columns being filled with 0s, we noticed that the DISCOMFORT_SLIDER was almost filled with only 1s (aside from 3 rows which had values 2,3, and 4). Thus, this column also may not provide us with too much information as the majority of data is exactly the same aside from a very small chunk. We noticed that the heart rate data and stress data which had valuable information had many rows with missing values. Thus, it is essential to figure out a way to handle these missing values whether it be dropping them, replacing them with zero or other methods.

For the second hypothesis that we were asked to explore, a similar data collection process was collected. In this, the Garmin Watch data was used and specifically the stress values. On top of that, the AM, PM, and END OF DAY data was used in this hypothesis as well.

II. Preliminary Analysis

Preprocessing the data

For starters, we simply remove all the columns that have values being zero. due to the various columns simply only being values of 0; we believe they provide no new information. However, this implies that we have less covariates to analysis, which can be a potential limitation.

Another limitation is the NaN values in the dataset, which will affect predictions that are less accurate. However, it is risky to handle these NaN values; If we are not careful, it is possible to reach conclusions that are not true due to skewing the data. In the end, we decided to discard all NaN values as the number of NaN values was very minimal.

DATA CLEANING:

Clean Null Values:

As mentioned above, we discarded all NaN values. Due to the fact that there were barely any NaN values while working with the data, it made the most sense to simply discard them as we had a lot of data either way.

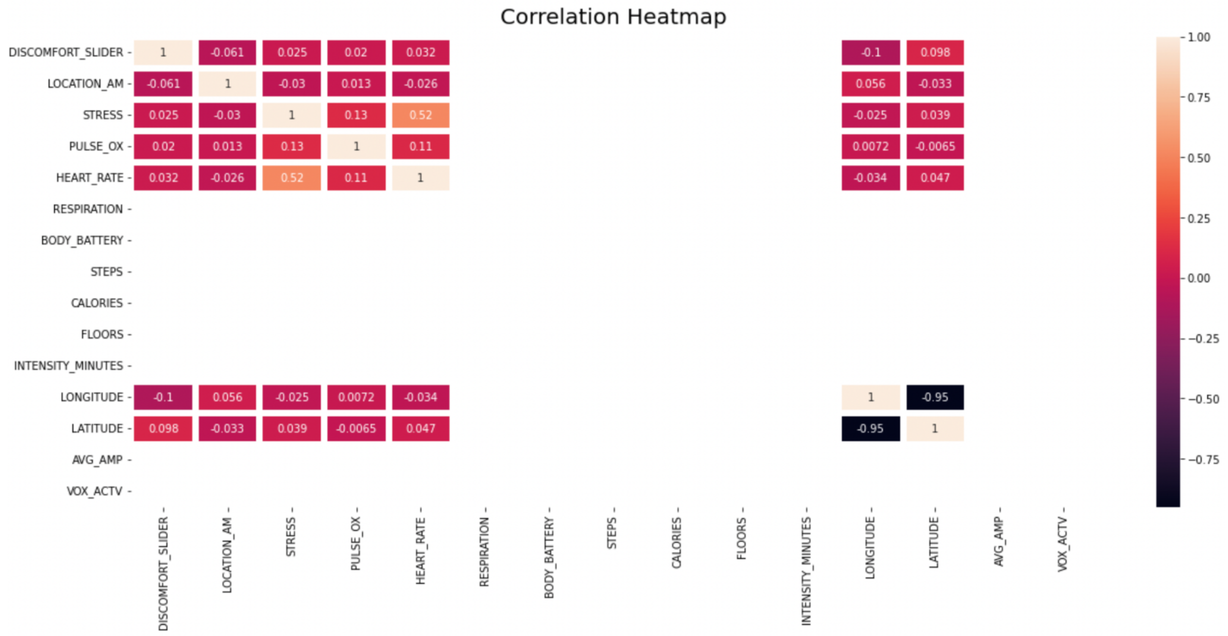
Uniforming Time Format and Merge table:

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.

DATA CORRELATION MATRIX

We performed a heatmap matrix 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 it does not give too much information we were looking for. On the other hand, the correlation between Heart_rate and Stress is around 0.52. Even though it is not very significant, we still choose to do a linear regression analysis



LINEAR REGRESSION MODEL AND INTERPRETATION

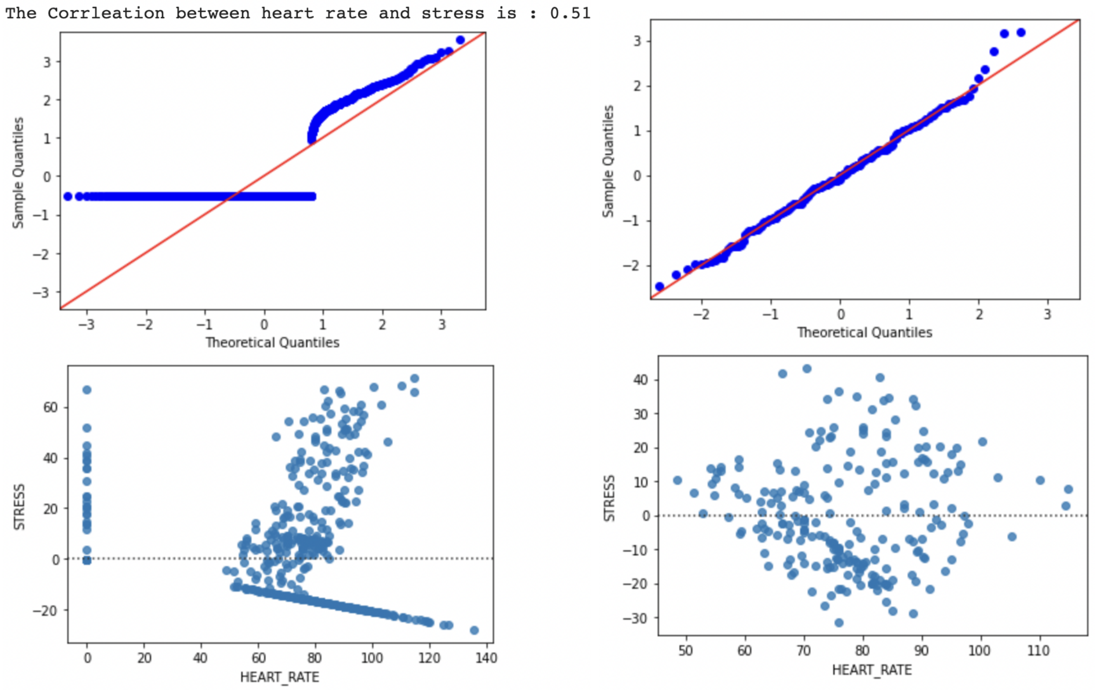
To examine the assumptions of doing a linear regression, we have to check the following assumptions are fully met:

1. The data needs to be normally distributed
2. The standard deviation is randomly distributed.

If the two assumptions are met, we can say that the linear regression model is valid.

<u>Heart Rate vs Stress:</u>

To begin with, we draw the QQ plot and Residual plot to see whether the raw data satisfied the assumption. However, they are not satisfied which means that we remove the data located at value 0.


```
The Corrleation between heart rate and stress is : 0.51
```

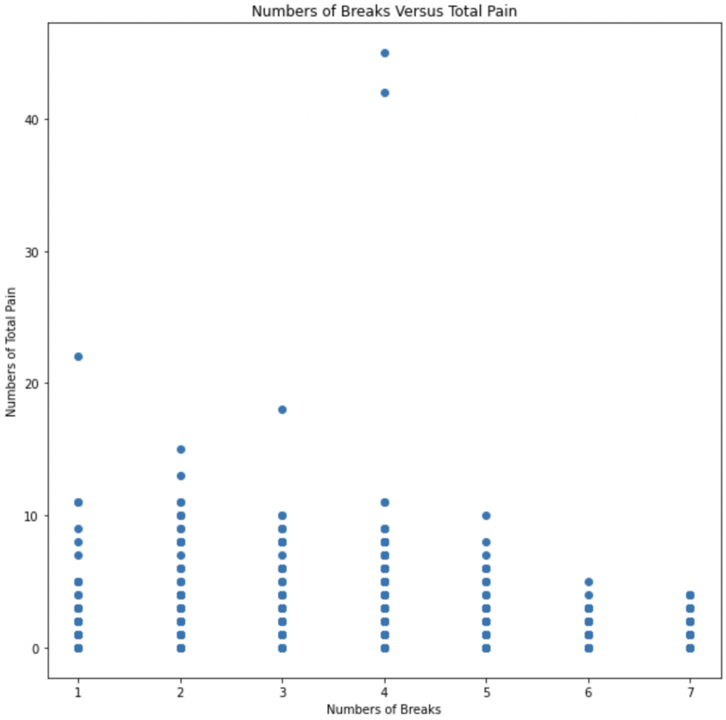
Therefore, we delete the data at value 0, and plot the QQ plot and residual plot. Here, the data follows the assumption and we can finally make linear models between heart rate and stress.

After performing the regression, we found there exists a positive linear relationship between the HEART RATE and STRESS. And we also calculated the R^2 which is 0.51, and adjusted R^2 which is 0.5. And this means that the data can interpret 50% of the situation that we predict.
Limitations and Refined

First Hypothesis:

First Hypothesis to test: "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 ".
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. Therefore, we first approached by summing up all the pain data for each individual and this was columned as "Total Pain". After that, we formed two groups: 2 or fewer breaks and 4 or more breaks by averaging the total pain for every individual in each group.

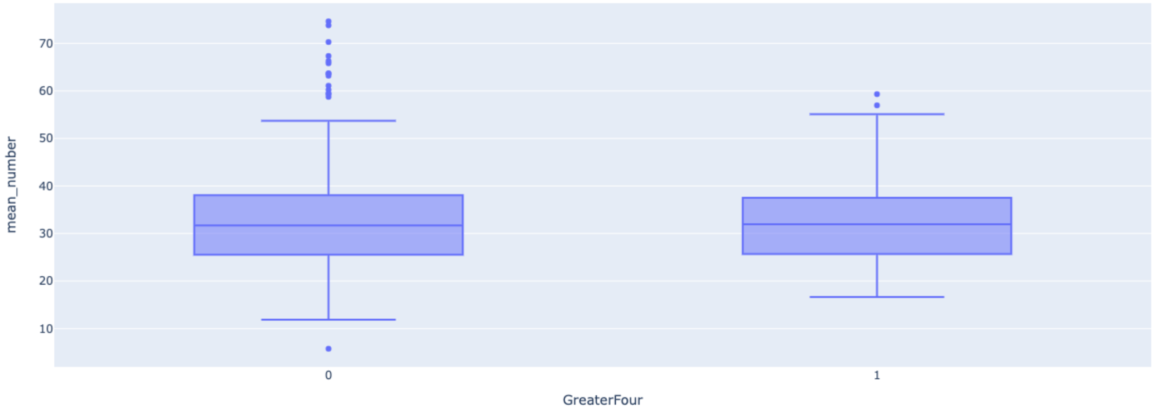
The "Total Pain" for two or fewer was recorded as 1.37 whereas the "Total Pain" for four or more was recorded as 1.54. It seems negligible at first, hence further analysis was required to continue in later iterations of the deliverable. In addition, KMeans was run to see if the pain factors could predict the number of breaks by clustering. Out of the 2,308,880 pairs formed there were 574189 that disagreed which simply shows that KMeans was relatively successful in forming the clusters. In the future, a possible approach is to use K Nearest Neighbors and some logistical regression to accurately cluster and predict the data.



Our scope has changed to do the following. For starters, we will find alternative ways to see if we can use the correlation between heart rate and stress to effectively predict some meaningful predictions. However, more importantly, we will focus on acquiring better results for our hypothesis. Our primary goal is to answer the hypothesis and potentially find a way to predict the pain people are feeling with all of the data, including the number of breaks.

Second Hypothesis:

For recollection purposes, the second hypothesis was “Participants who have an average of 3 locations per week will have higher stress algorithms than people who use an average of 4 or more locations per week”. Below is the box-and-whisker plot of the distribution of the data.



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 or less than four.

From a simple eye test, we can see that there seems to be no significant difference between the two distributions and this relationship will further be analyzed through the hypothesis testing. For the future modifications, we will look at more months of data to get a better understanding of the relationship as we will have more data. In addition, we want to use the number of locations along with the actual location to predict the stress so that we can provide even better results to the client.

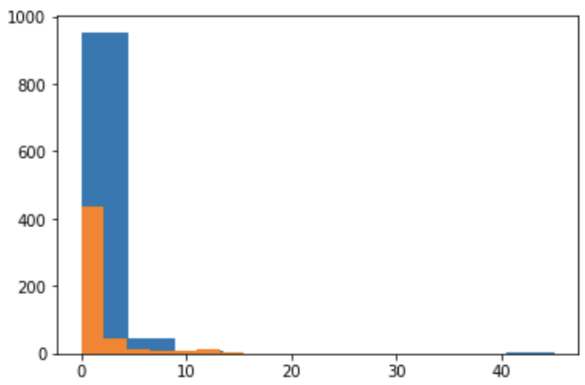
Hypothesis Test for Hypothesis I

Naive Approach

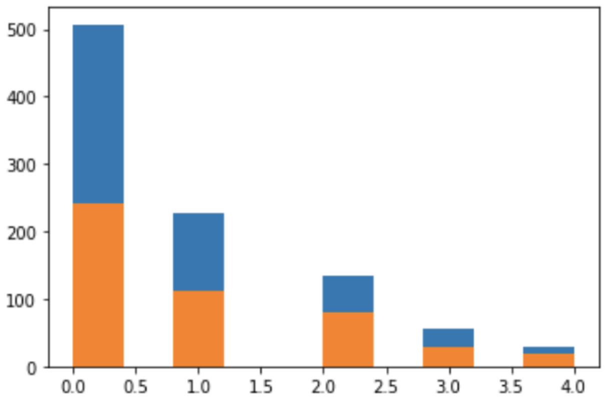
For Naive Approach, we computed and compared the average pain in a naive approach and found a difference between that of people who take break less than 2 times and those who take break more than 4 times a day. For example, in the fourth month data, we have the average pain for people who took four or more breaks to be 1.207, and the average pain for people who
took two or less breaks to be 1.6155. Therefore, from a naive approach perspective, there might exist a difference between the pain of people who took an average of 2 or fewer breaks per day and that of people who took an average of 4 or more breaks per day.

Paired t-tests(Without Outliers)

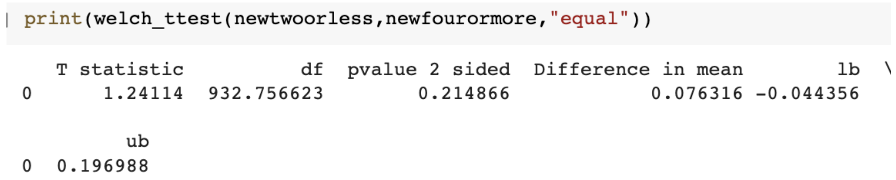
We are using the 2th, 3th,4th data to conduct the hypothesis testing process. We removed the outliers according to their interquartile range multiplied by 1.5. Using the fourth month data as an example, we plot the number of pain reported in the histogram. We found that people who take breaks four or more times a day presented as blue have a higher proportion of outliers, and another group of people has less proportion of outliers.



Then we utilize the interquartile range to remove the outliers. And we replot the number of pain for two groups of people in the histogram as below. We noticed that after removing outliers, as pain increases more, the proportion of people who take breaks less than 2 times a day increases.



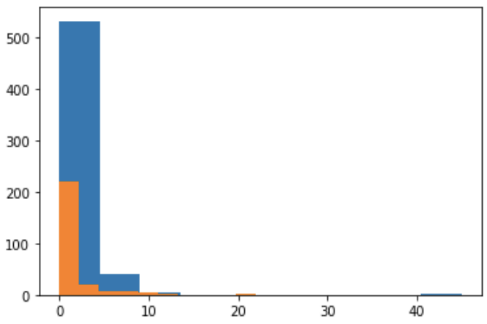
In this case, we decide to conduct the paired t-test and we set the alpha value to be 0.05. We noticed that the number of two groups are different. Therefore, we can only use the Welch t-test. Here we obtained a p value of 0.21 and we can not reject the null hypothesis and there is no significant difference between them in the fourth month data.

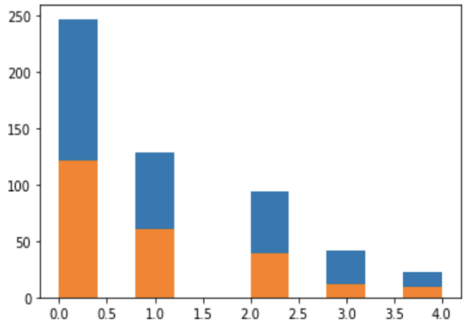


```
print(welch_ttest(newtwoorless,newfourormore,"equal"))

   T statistic          df  pvalue 2 sided  Difference in mean        lb  \
0      1.24114  932.756623        0.214866            0.076316 -0.044356

         ub
0  0.196988
```

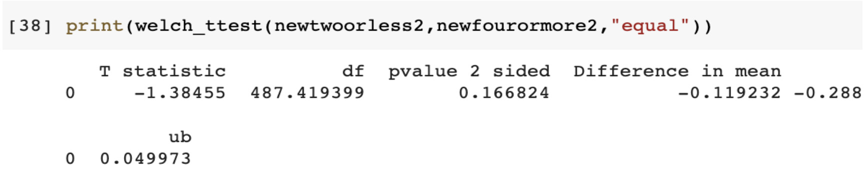
By using the same method described above, we decided to conduct the hypothesis test for the 2nd, 3rd month.

For the second month, we found that after removing outliers, most of the outliers are from people who take four or more breaks a day since we can not see the blue part at the pain level of 40 after we remove the outlier.



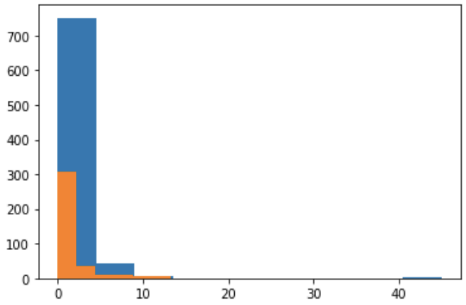


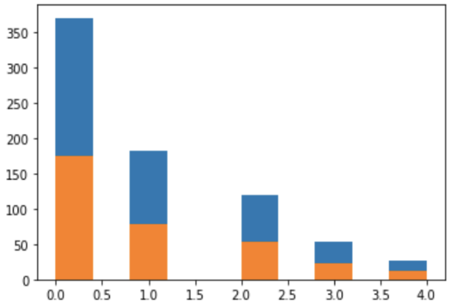
We used Welch t-test and found a p value to be 0.16. And here we can not reject the null hypothesis and there is no significant difference between them in the second month data.



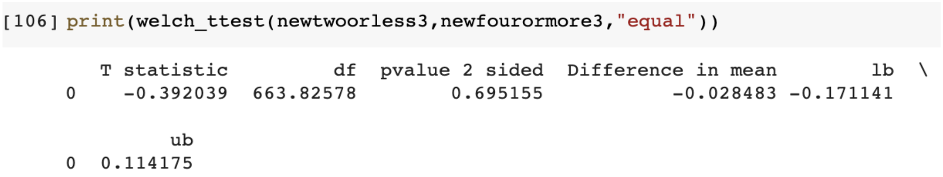
```
[38] print(welch_ttest(newtwoorless2,newfourormore2,"equal"))

       T statistic           df  pvalue 2 sided  Difference in mean
0     -1.38455  487.419399         0.166824            -0.119232 -0.288

            ub
0   0.049973
```

For the third month data, we also found that after removing outliers, most of the outliers are from people who take four or more breaks a day since we can not see the blue part at the pain level of 40 after we remove the outlier. And from here, we wonder why we can notice that most of the outliers are from the blue part, which is the group of people who take four or more breaks per day.





For the Welch t-test, we obtained a p-value that equal to 0.695. And here we can not reject the null hypothesis and there is no significant difference between them in the second month data.

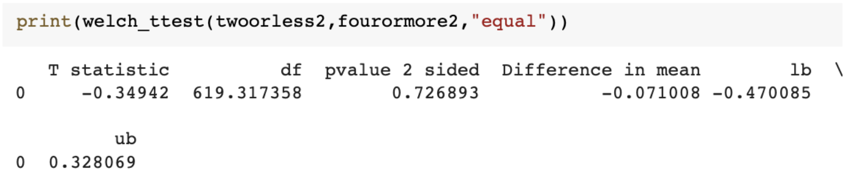


```
[106] print(welch_ttest(newtwoorless3,newfourormore3,"equal"))

       T statistic          df  pvalue 2 sided  Difference in mean         lb  \
0    -0.392039   663.82578        0.695155            -0.028483 -0.171141

            ub
0   0.114175
```

Paired-tests (With Outliers)

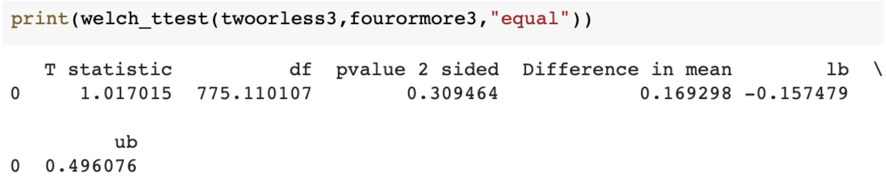
The reason why we redo the Paired-tests with outliers is that we think maybe those outliers are contributing to our project. In other words, they are not outliers with bad influence. If we can reject the null hypothesis with those good outliers, we can further explore what those outliers are and for what reason they exist.

For the second month, without removing outliers, we obtain a p value to be 0. 7, and we could not reject the null hypothesis.



```
print(welch_ttest(twoorless2,fourormore2,"equal"))

   T statistic          df  pvalue 2 sided  Difference in mean        lb  \
0     -0.34942  619.317358        0.726893           -0.071008 -0.470085

         ub
0  0.328069
```

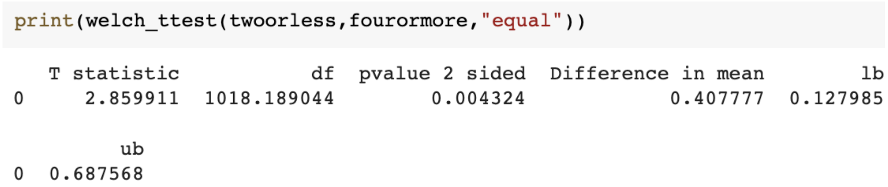
For the third month, without removing outliers, we obtained a p value of 0.3, and even though we could not reject the null hypothesis, we found that it performs better towards the result!



```
print(welch_ttest(twoorless3,fourormore3,"equal"))

   T statistic          df  pvalue 2 sided  Difference in mean        lb  \
0     1.017015  775.110107        0.309464            0.169298 -0.157479

         ub
0  0.496076
```

For the fourth month, without removing outliers, we obtain a p value of 0.004, and here we could reject the null hypothesis and claim there exists a significant difference between the painness of the two groups people.

With functional programming, we found that both lower confidence intervals and higher confidence intervals are both larger than 0. Which means that the mean pain for people who rest four or more times is larger than that of people who rest two or less times. Therefore, we wonder if those outliers imply a causal relationship. Therefore, we decided to explore the Causal-Relationship.



```
print(welch_ttest(twoorless,fourormore,"equal"))

   T statistic           df  pvalue 2 sided  Difference in mean        lb
0     2.859911  1018.189044        0.004324            0.407777  0.127985

         ub
0  0.687568
```

Causal-Relationship Risks

The interesting thing is that before we clear out the outliers, there is a significant difference between them in the fourth month data In other words, there is a significant difference between the pain for people who rest less than 2 and people who rest more than 3 times because some people are originally painful even before they rest for longer times. And when we exclude those people who are originally painful, we are analyzing the other group of people who strategically decided to avoid pain by taking a large number of breaks.

Here we assume there are two types of participants in our group. And here they are:

1. People who used to have pain and have to take more breaks told by their doctor
2. People who do not have illness before and take breaks to avoid the potential pain.

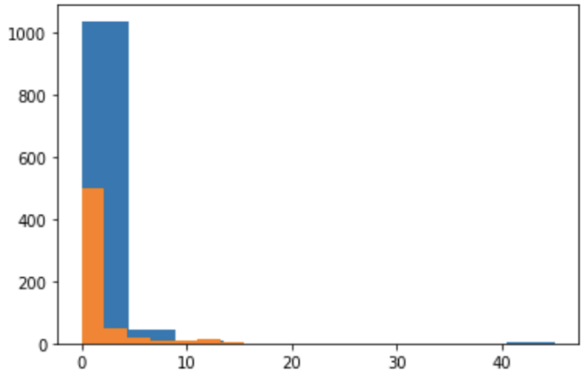
Our assumptions can explain why there are always outliers associated with people who take four or more times a day. Since there are some people who originally have so much pain and they have to take more breaks told by their doctors, they belong to the blue part and serve as outliers. However, do the outliers belong to bad outliers or good outliers?

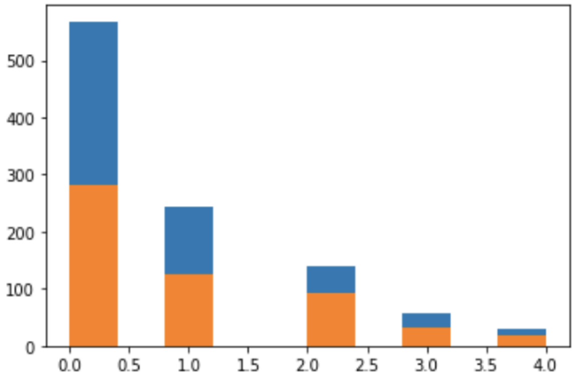
In our understanding, those outliers can not be identified as bad outliers. Since our project is to analyze the effects, both physically and mentally, of remote work over a long period of time (6 months), we need to take all kinds of workers into consideration.

We wonder that as time goes by, the effect of outliers will gradually decrease.

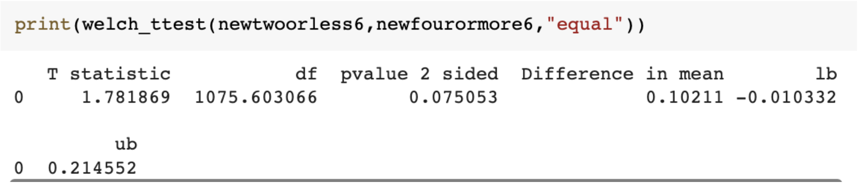
Paired t-tests for the sixth month

In the sixth month data, we still have outliers.

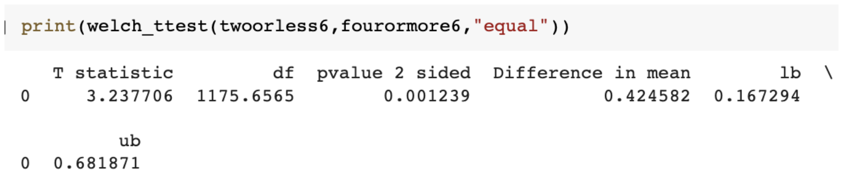




And we compare both the effect with and without outliers. We found that both with and without outliers, we can conclude that we can reject the null hypothesis.



```
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
```



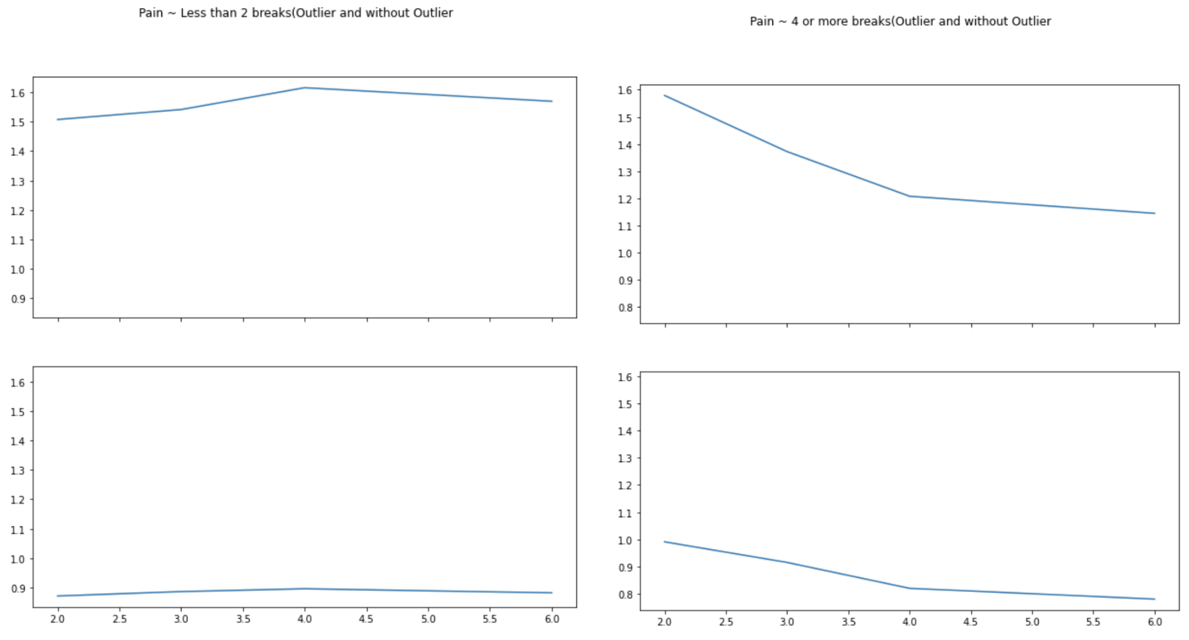
```
| print(welch_ttest(twoorless6,fourormore6,"equal"))

    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
```

Here, we noticed that before dropping outliers, people who take four or more breaks will have less pain than people who take two or less breaks. And even after dropping outliers, we still have small p values. \

To further explore the Cause- Effect Risks Analysis, we combine all month data into one data list. According to the left graph, we found that for people who take breaks less than 2 times per day, their painness does not change during the whole six month scale.

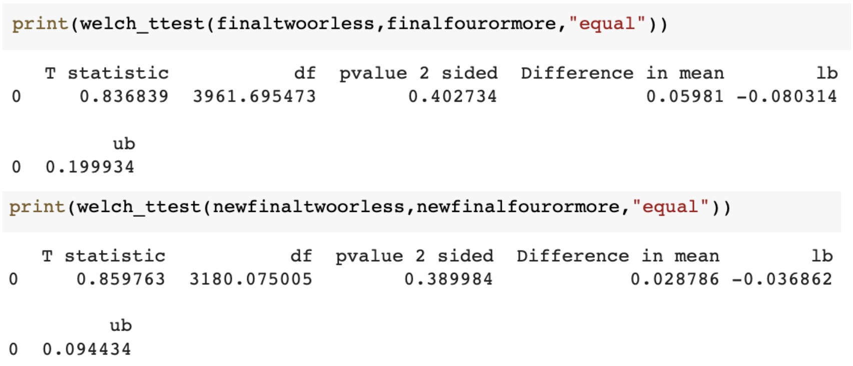
However, for people who take breaks four or more times per day, we found that people who take four or more breaks per day will gradually lower their pain report during the five months both with or without outliers. And for people who take two or three breaks per day, they will gradually feel more pain at the third month but feel less pain at six months.

According to the trendy graph. We found that the difference between people who rest four or more times and people who rest less than 2 times is affected more by people who take four or more breaks. And we can decide whether people who rest more times do increase their pains despite what we previously said about the confusing cause-effect relationship.



Here, we redo the hypothesis both and without outliers. Therefore, we found that we could not reject the null hypothesis using the dataset as a whole.

However, we do not limit ourselves to the current dataset. So by combining them together we found that it is not significant. However, when we seperate them and conduct hypothesis tests one by one, it shows a trend that even though some people rest four or more times because of their original pain, they feel less and less pain as time goes by. And we can see the effects of outliers became less and less also because we see the p-value of two tests became convergent. Therefore, we assume in the future dataset, the effect of outliers will gradually decrease. And we can also notice that for people who have to take breaks, their painness also decreases along their frequent breaks in such a long time period.



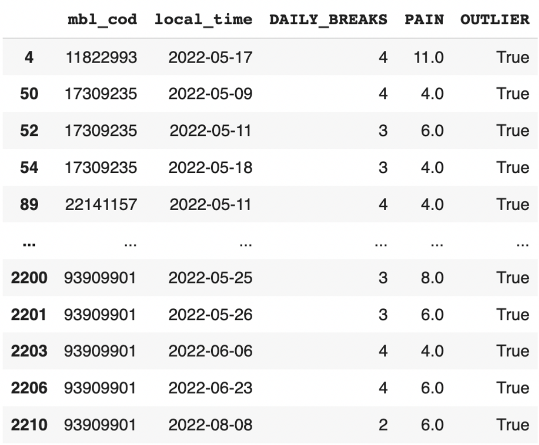
```
print(welch_ttest(finaltwoorless,finalfourormore,"equal"))

   T statistic           df  pvalue 2 sided  Difference in mean        lb
0     0.836839  3961.695473        0.402734             0.05981 -0.080314

         ub
0  0.199934
```

```
print(welch_ttest(newfinaltwoorless,newfinalfourormore,"equal"))

   T statistic           df  pvalue 2 sided  Difference in mean        lb
0     0.859763  3180.075005        0.389984            0.028786 -0.036862

         ub
0  0.094434
```

We faced many times that our null hypothesis could not be rejected. As we all know, work from home has been there for a long time, so ideally there should be no change when comparing month on month. However, we highly expect that if we compare them year by year there will be a more obvious change.

New Finding to Include Outliers

Given our Outlier testing for the entire patient group, we concluded the finding of the outlier based on interquartile range test.



	mbl_cod	local_time	DAILY_BREAKS	PAIN	OUTLIER
4	11822993	2022-05-17	4	11.0	True
50	17309235	2022-05-09	4	4.0	True
52	17309235	2022-05-11	3	6.0	True
54	17309235	2022-05-18	3	4.0	True
89	22141157	2022-05-11	4	4.0	True
...	...	...	...	...	...
2200	93909901	2022-05-25	3	8.0	True
2201	93909901	2022-05-26	3	6.0	True
2203	93909901	2022-06-06	4	4.0	True
2206	93909901	2022-06-23	4	6.0	True
2210	93909901	2022-08-08	2	6.0	True

The above printout shows the description of the outliers. In total, there are 204 outliers. Based on the data description result, we realized that there exists a significant problem, where patients who reported a pain level larger than 3 also gets included in the outlier range. This significantly reduces the range for pain reporting, and it shows that removing these outliers will cause a dataset with a lot of data condensed between a pain rating of 0 to 3.

In other words, someone who has pain level reported larger than 3 will contribute significantly towards relational analysis based on the the size of the pain number range of the whole group and its data distributions. Given that we are doing a interquartile range approach to figure out outliers, it is natural that it does not give us the outliers that we desired. Hence, we decided to re-study the entire dataset based on individual hypothesis testing with current "outliers" included.

Individual Identification With Hypothesis Testing

According to our project, they want to know exactly which individual failed at this Hypothesis Testing. Therefore, we conducted a list and composed individual data into each person's id. Then for each ID, we conducted t-hypothesis testing for them one by one.

Among total 69 participants, there were 31 participants that lacked data so we could
not properly conduct hypothesis testing for them. There were 21 participants that failed to reject a null hypothesis, and there were 17 participants that could reject our hypothesis. To better present, we build a column below to present participants IDs

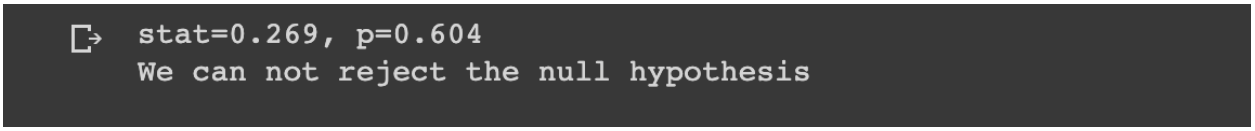
Therefore, with the individual identification, scientists could use the id provided below to further track individuals groups for further research.

Participants ID(no data)	Participants ID (reject null hypothesis test)	Participants ID (could not reject null hypothesis test)
13392141,	11822993,	23916703,
17309235,	17180706,	32937810,
20126808,	20763027,	33075391,
22141157,	22541511,	34865333,
25230030,	27148444,	37720972,
26141560,	34633705,	49164240,
27361835,	36505757,	49669568,
32455277,	38876664,	55508636,
35549180,	47443793,	58805130,
38656882,	54042771,	60404747,
38886354,	57026233,	61881920,
47685985,	58601340,	66958688,
51755925,	64811087,	66999191,
55448394,	70975009,	69497234,
56954906,	71552354,	71681441,
58395682,	81875100,	73262082,
58500979,	91556555	77253909,

58947714, 61307863, 62547526, 68415107, 68454890, 70398973, 79316883, 81914178, 82420964, 82878753, 86570707, 87350835, 88676885, 96243591		80515680, 81862952, 86548395, 93909901

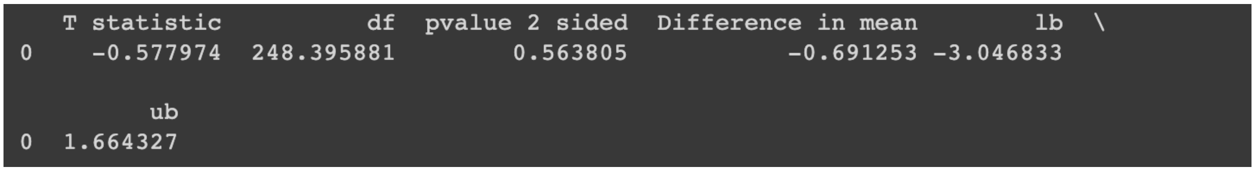
Hypothesis Test for Hypothesis II

For the second hypothesis, a one-way ANOVA test was conducted to see if the means were statistically significant. Below were the results of the hypothesis



```
stat=0.269, p=0.604
We can not reject the null hypothesis
```

We can clearly see that it is not statistically significant at the 0.05 level of significance nor the 0.1 level of significance. Hence, we would fail to reject the null. As discussed for Hypothesis 1, the Welch t-test was performed and the results are shown below.



```
   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
```

We notice the same issue here that it is not statistically significant. Hence, we can not reject the null hypothesis which means that we are unable to determine if the two means are indeed different.

Individual Identification With Hypothesis Testing

Similar to hypothesis 1, we conducted a one-sided ANOVA t-test on every single individual to account for individual differences. The table below shows which individuals indicated behavior that was statistically significant, behavior that was not statistically significant, and behavior that could not be analyzed due to lack of data.

Participants ID(no data)	Participants ID (reject null hypothesis test)	Participants ID (could not reject null hypothesis test)
17180706.0, 17309235.0, 20763027.0, 23916703.0, 25230030.0, 27361835.0, 36505757.0, 49164240.0, 54042771.0, 55448394.0, 55508636.0, 58805130.0, 61881920.0, 70975009.0, 71552354.0, 71681441.0, 73262082.0, 80515680.0, 81875100.0, 88676885.0, 91556555.0	11822993.0, 38656882.0, 58601340.0, 60404747.0, 64811087.0, 77253909.0	22541511.0, 27148444.0, 32455277.0, 32937810.0, 33075391.0, 34865333.0, 35549180.0, 37720972.0, 38876664.0, 51755925.0, 56954906.0, 57026233.0, 61307863.0, 66958688.0, 66999191.0, 69497234.0, 79316883.0, 81862952.0, 86548395.0

We notice that there are much more statistically insignificant individuals than statistically significant individuals. This may further our proof that there does not seem to be much of a difference at all between the number of locations in a week and the stress of that individual.

Challenges Faced and Suggestions

In terms of challenges, the process of cleaning the Garmin data was incredibly challenging. This is due to the sheer volume of data that is collected and as a result the computing power required to actually analyze the data was immense. Other challenges that were faced were in the areas of merging tables while also retaining the information required.

For the future, it is highly recommended that intermediate datasets are saved so that later on computing time can be saved. In addition, it is highly recommended to thoroughly look at all the data at hand and understand each variable. This will significantly help cut time while analyzing the data.