Deliverable 2

Team 2

Link to presentation with the client: WFH Presentation with Karen Jacobs

Timestamp: 0:36:26

Deliverable 2 consists of a comprehensive report of all the hypotheses and key questions given, followed by further steps taken to find more key findings of the project.

Hypothesis A

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

Background needed:

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

Domain 6: Financial and Material Stability:

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

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

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

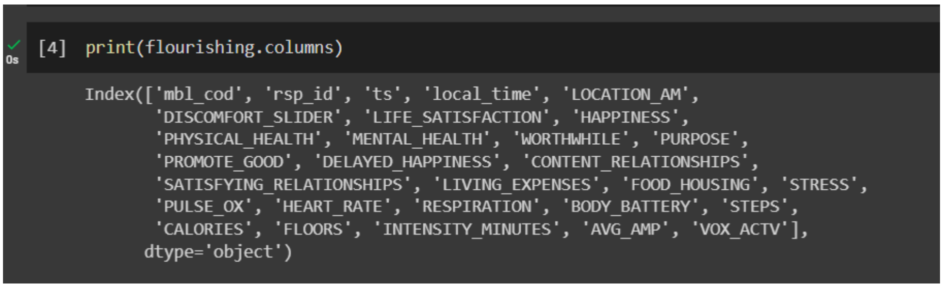
Data used:

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

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

Data preprocessing:

FridayAM has the following columns:



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

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

Since we only need to focus on financial stability, let us store ‘Living_expenses’ and ‘Food_housing’ in a separate dataframe. We also need to retain rsp_id, ts and local_time since it contains participant’s ID and timestamp’s information. We will drop everything else. Then, I took an average along the columns of Living_expenses and Food_housing and stored it with the respective mobile ID and age associated with the mobile ID. Now, I had a dataframe with the following columns - mid, age, expenses and food_housing.

Analysis:

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

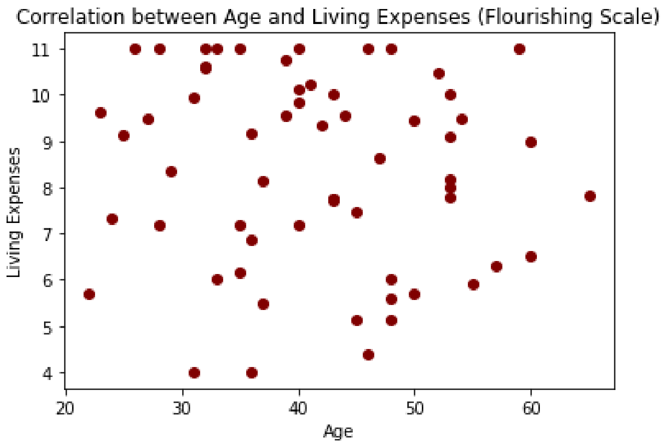
Step 2 - Merging this dataframe containing the averages with the ages on mobile ID.

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

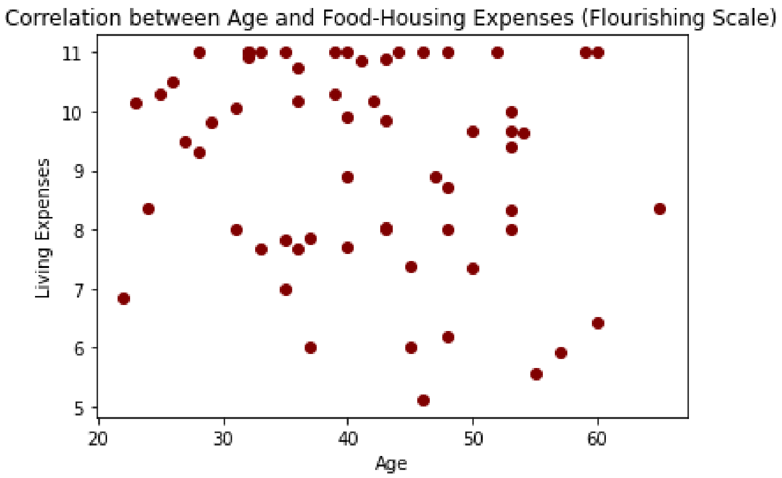
Formula used: Pearson correlation coefficient given by:

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

Results:



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



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

Hypothesis B

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

Background needed:

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

Data used:

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

Data preprocessing:

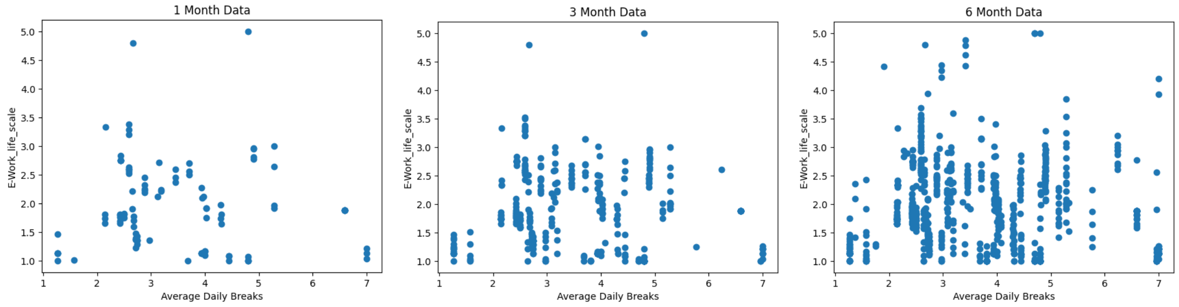
It drops rows with missing values and calculates scores for different categories (Trust, Flexibility, Work_life, and Productivity).

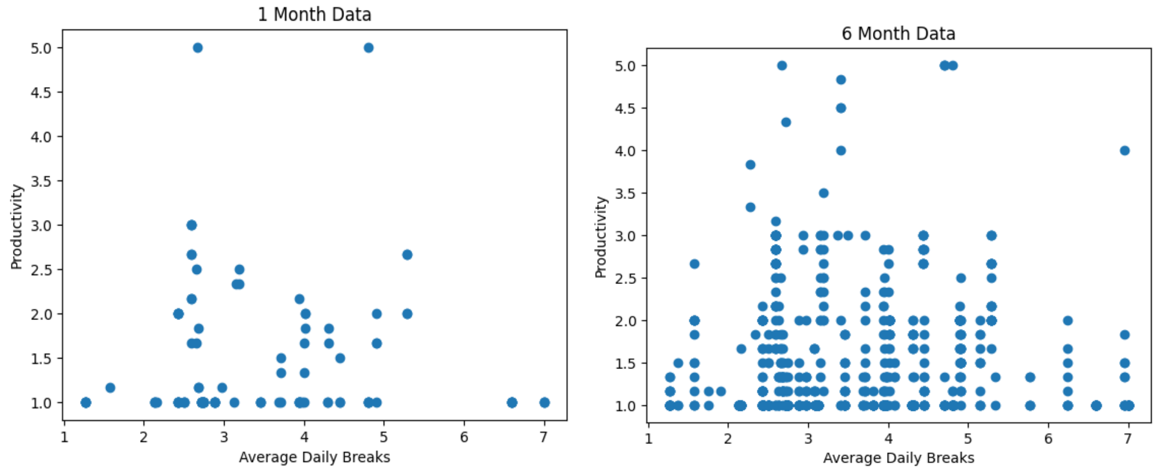
The code then calculates the E-Work_life_scale based on the weighted scores of the categories. It merges the E-Work_life_scale data with the average breaks data for each employee and creates a scatter plot of the E-Work_life_scale vs. Daily Breaks. Finally, it prints the mean E-Work_life_scale of breaks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

Analysis:

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

Results:





correlation:

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

Formula used:

['E-Work_life_scale'] =

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

Hypothesis C

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

Background needed:

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

How would you rate your overall mental health?

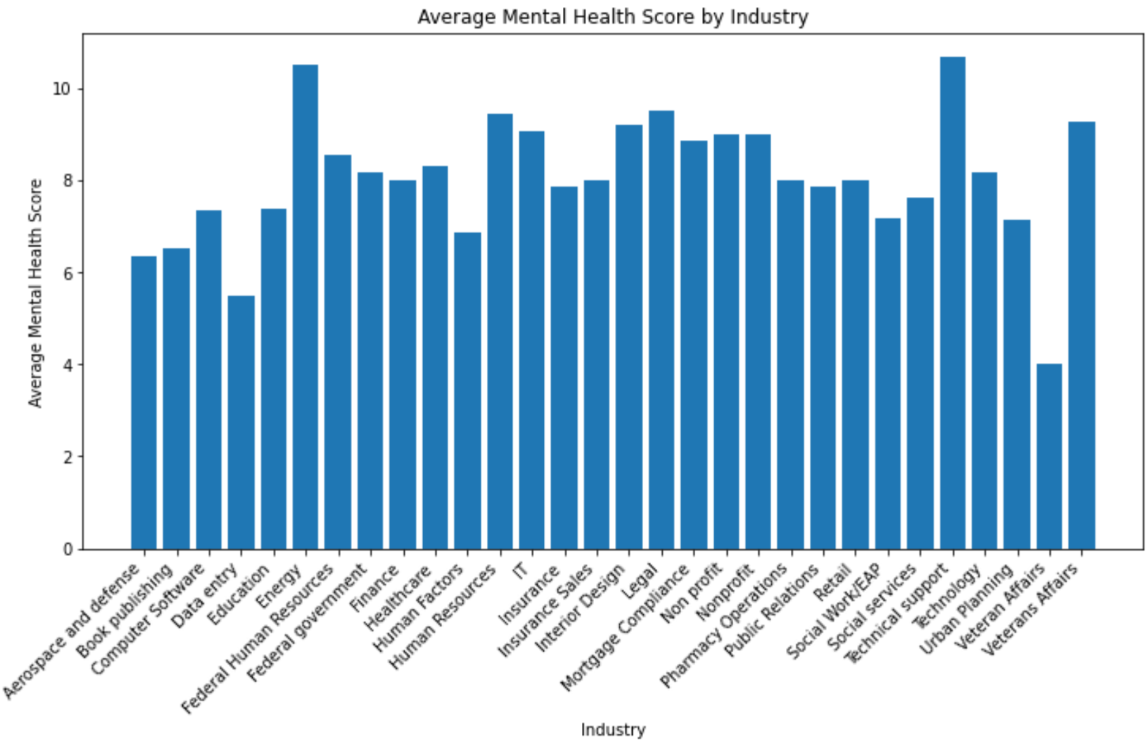
0 = Poor, 10 = Excellent

Data used: FridayPM 3 Month; Demographic

Data preprocessing:

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

Analysis:



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

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

The mental health scores of participants in healthcare were compared to those in other industries by calculating the average mental health score for each industry. The data was grouped by industry, and the mean mental health score was calculated for each group. The results showed that the average mental health score for participants in healthcare was not significantly different from the average mental health score of participants in other industries.

An interesting observation on the other hand, shows that participants of Veteran Affairs have lower mental health scores overall on average.

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

Results:

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

Hypothesis D

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

Background needed:

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

Data used:

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

Data preprocessing:

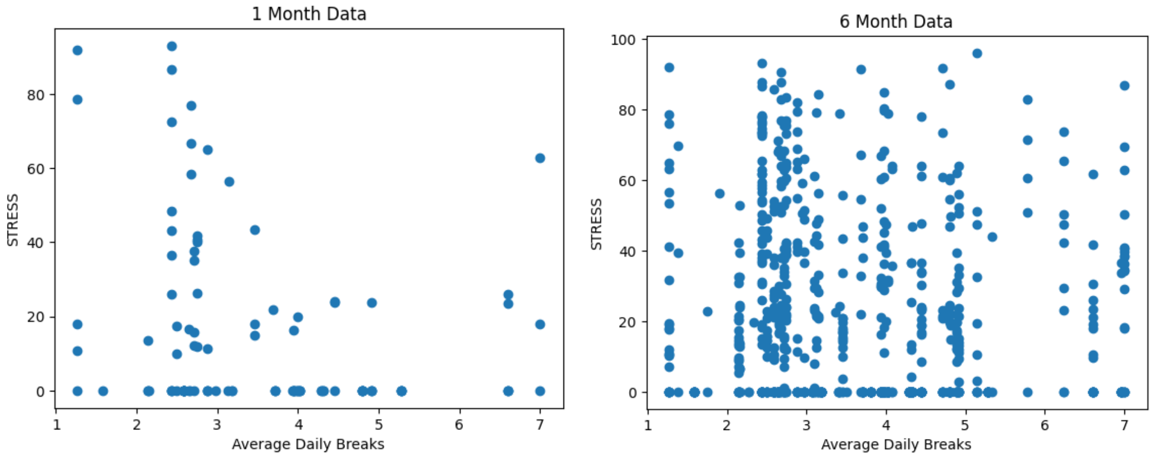
It drops rows with missing values, calculates the mean stress score for each participant, and creates a new column 'STRESS' in the dataframe with these scores. It also merges the resulting dataframe with another dataframe 'df_avgbreaks' containing average daily breaks taken by each participant.

Analysis:

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

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

Results:



Correlation:

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

Hypothesis E:

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

Background needed:

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

15. Are workers trained in the following:

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

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

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

proper postures: POSTURE_TRAINING

proper work methods: METHODS_TRAINING

WMSD: WMSD_SIGNS

adjust workstation: WORKSTATION_ADJUSTMENT

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

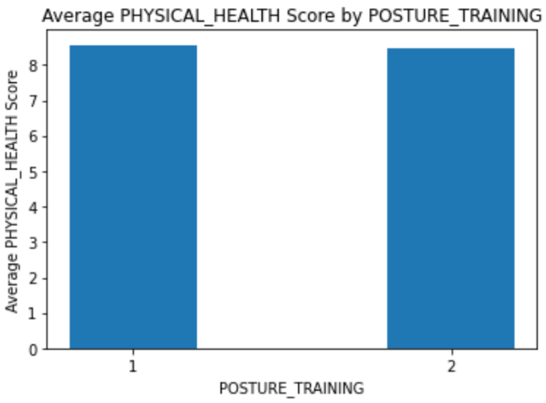
In general, how would you rate your physical health? 0 = Poor, 10 = Excellent

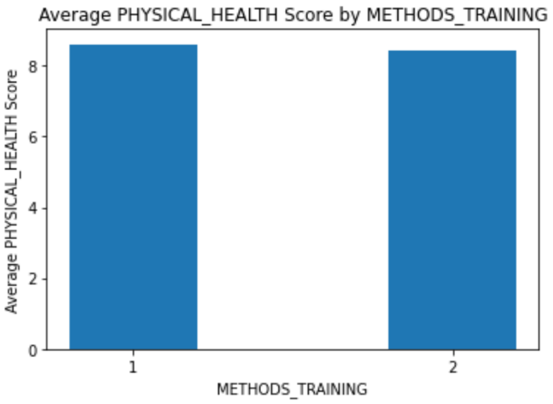
Data used: Computer Workstations 6 Month; Friday AM 6 Month

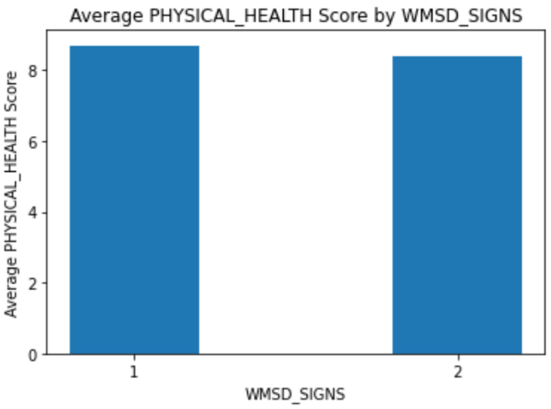
Data preprocessing:

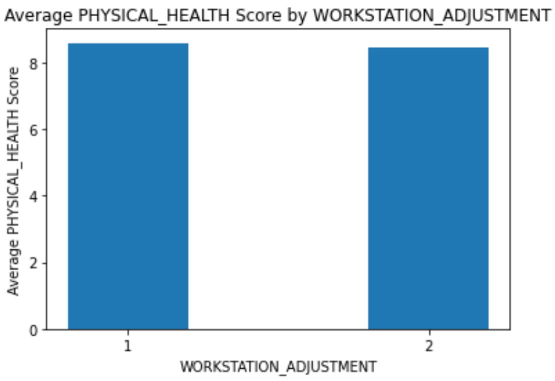
The code merges two datasets into one DataFrame by participant id, and then selects the relevant columns, dropping the empty rows. Then first calculates the average level of pain with respect to each area, and then calculates the correlation coefficient for each of them.

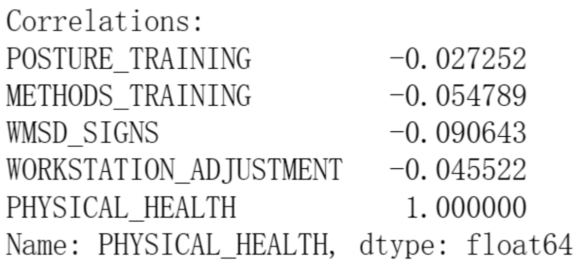
Analysis:











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

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

r = (Σ(xi - x̄)(yi - ȳ)) / (sqrt(Σ(xi - x̄)^2) * sqrt(Σ(yi - ȳ)^2))

where xi and yi are the individual data points, x̄ and ȳ are the sample means, and sqrt is the square root function. We apply this formula to each response with the physical health score.

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

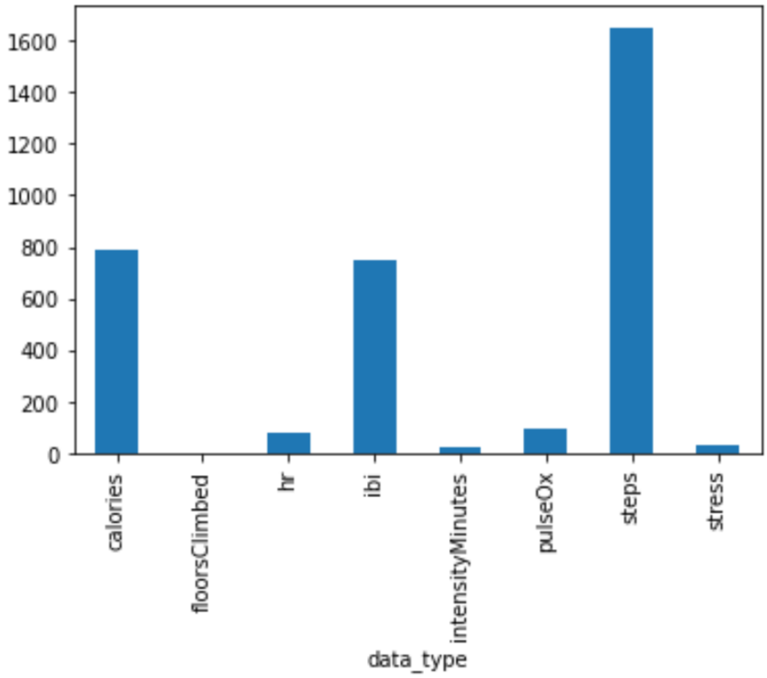
Results:

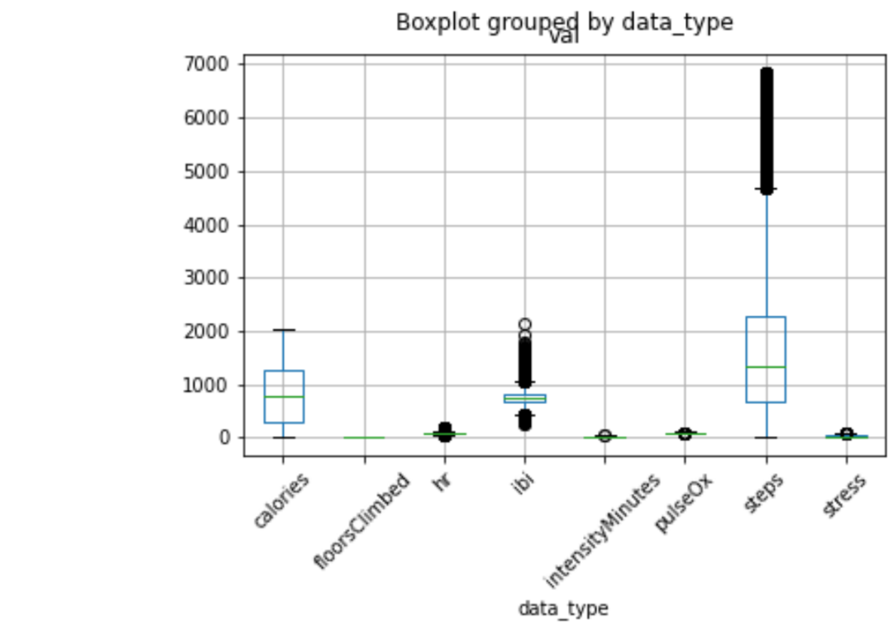
Based on question #15 in the Computer Workstation Checklist (with 4 responses regarding ergonomics training), participants with lower scores will report less pain at 6-months.

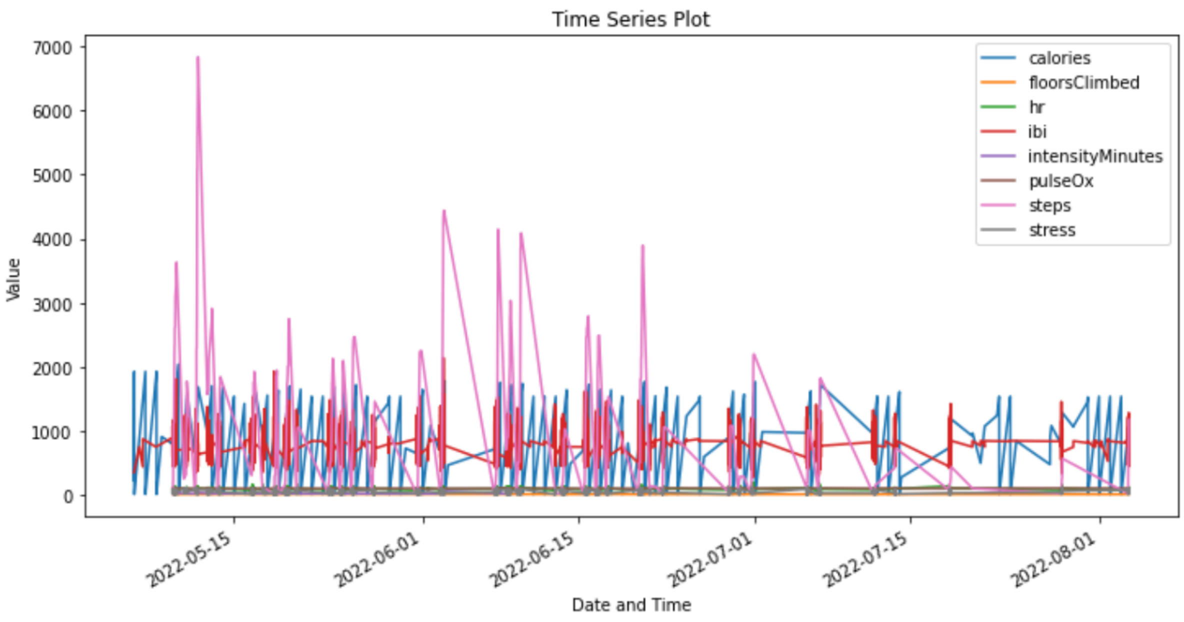
A first look at the Garmin Data

Datasets Used: garmin.11822993 2; garmin.17180706

Analysis:







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

The CSV files have the following columns:

ts: a timestamp of the data

dte_tme: the date and time of the data

rsp_id: the participant ID

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

val: the value of the recorded data type

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

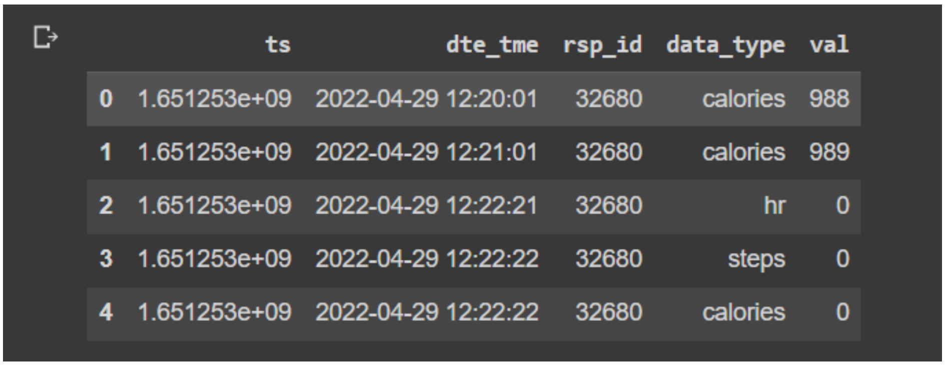
The first analysis uses the groupby() method of the DataFrame to group data by 'data_type'. The mean, median, and standard deviation of the 'val' column for each group are then calculated and visualized through bar and box plots.

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

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

Analysis on the Garmin Data

The Garmin data has the following structure:



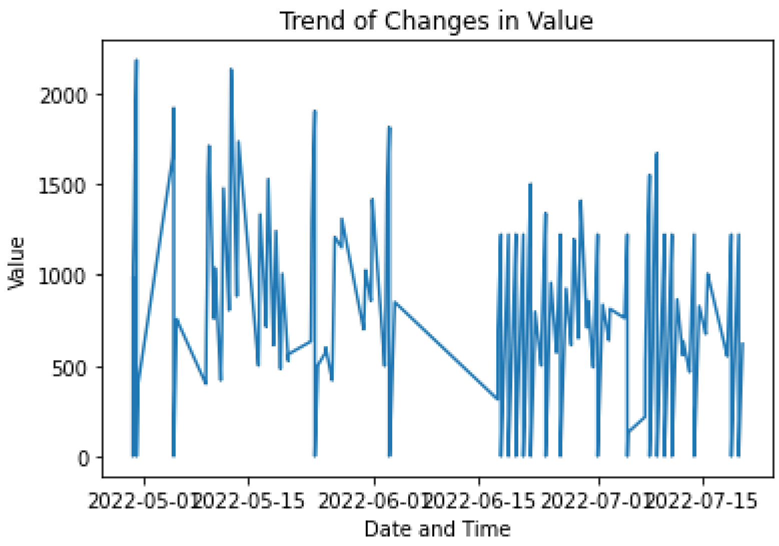
	ts	dte_tme	rsp_id	data_type	val
0	1.651253e+09	2022-04-29 12:20:01	32680	calories	988
1	1.651253e+09	2022-04-29 12:21:01	32680	calories	989
2	1.651253e+09	2022-04-29 12:22:21	32680	hr	0
3	1.651253e+09	2022-04-29 12:22:22	32680	steps	0
4	1.651253e+09	2022-04-29 12:22:22	32680	calories	0

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

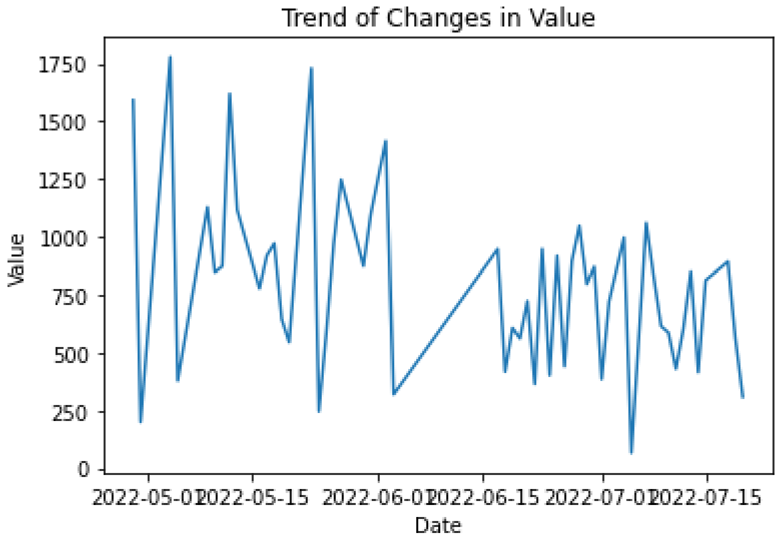
It records ['calories', 'hr', 'steps', 'floorsClimbed', 'intensityMinutes', 'pulseOx', 'ibi', 'stress'].

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

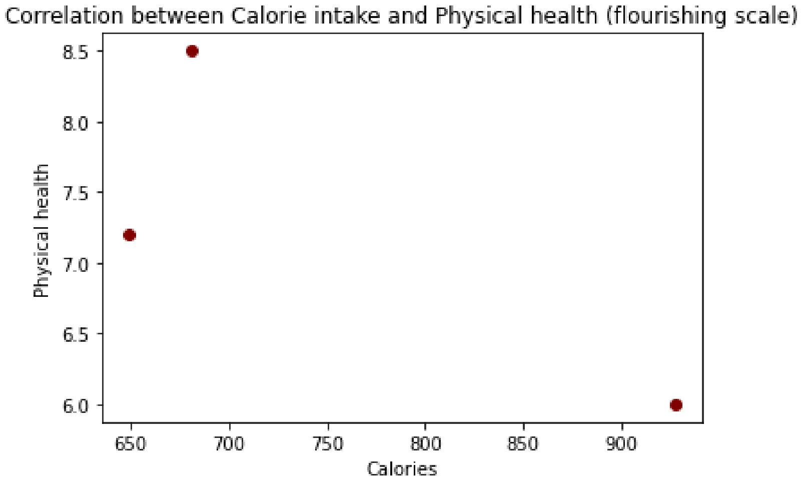
I first analyzed the overall trend of changes in caloric values. It was observed that in the initial months the random participant did have a better calorie intake than in the later months.



Then I also analyzed the daily changes:



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


Correlation between Calorie intake and Physical health (flourishing scale)
Physical health
Calories