Introduction

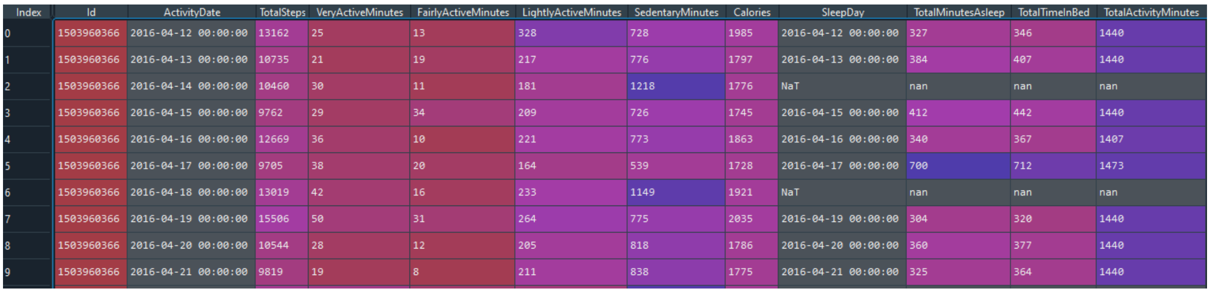
With the popularity of wear devices, a huge amount of physical-/lifestyle-related data can be easily and continuously collected. An exploration on these data will be helpful for monitoring physical conditions and understanding lifestyle patterns of the users so that personalised approaches can be developed to boost fitness.

This analysis is performed based on FitBit Fitness Tracker Data in Kaggle (ref: https://www.kaggle.com/datasets/arashnic/fitbit). I try to understand the users' lifestyle and Fitbit usage habit. Moreover, I am also interested in finding out if longer physical activity can increase sleep time.

1. Data Exploration

In total, there are activity data of 33 users and sleep data of 24 users. Since 32 users have activity data of at least one week, weekly activity times were summarised based on the 32 users.

Let's left join the activity data with the sleep data on Id and date, and take a look at the duration of each recorded activity.

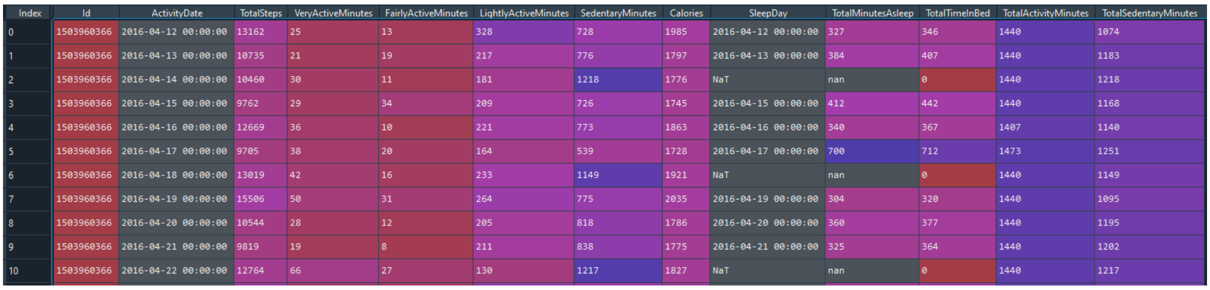


Index	Id	ActivityDate	TotalSteps	VeryActiveMinutes	FairlyActiveMinutes	LightlyActiveMinutes	SedentaryMinutes	Calories	SleepDay	TotalMinutesAsleep	TotalTimeInBed	TotalActivityMinutes
0	1503960366	2016-04-12 00:00:00	13162	25	13	328	728	1985	2016-04-12 00:00:00	327	346	1440
1	1503960366	2016-04-13 00:00:00	10735	21	19	217	776	1797	2016-04-13 00:00:00	384	407	1440
2	1503960366	2016-04-14 00:00:00	10460	30	11	181	1218	1776	NaT	nan	nan	nan
3	1503960366	2016-04-15 00:00:00	9762	29	34	209	726	1745	2016-04-15 00:00:00	412	442	1440
4	1503960366	2016-04-16 00:00:00	12669	36	10	221	773	1863	2016-04-16 00:00:00	340	367	1407
5	1503960366	2016-04-17 00:00:00	9705	38	20	164	539	1728	2016-04-17 00:00:00	700	712	1473
6	1503960366	2016-04-18 00:00:00	13019	42	16	233	1149	1921	NaT	nan	nan	nan
7	1503960366	2016-04-19 00:00:00	15506	50	31	264	775	2035	2016-04-19 00:00:00	304	320	1440
8	1503960366	2016-04-20 00:00:00	10544	28	12	205	818	1786	2016-04-20 00:00:00	360	377	1440
9	1503960366	2016-04-21 00:00:00	9819	19	8	211	838	1775	2016-04-21 00:00:00	325	364	1440

I create a "TotalActivityMinutes" column by summing durations of all activities (i.e., Very active, Fairly active, Lightly active, Sedentary and In bed). From the results, there are two parts that get my attention:

1) For rows with missing value in "TotalTimeInBed", the "TotalSedentaryMinutes" value is nan. However, if you add the durations of other activities together, you will find that the summed duration is 1440 mins (i.e., 24 hrs). This tells us that in some cases, although total time in bed is not reported, the actual time might still be included in the "SedentaryMinutes". Therefore, it's better for us to impute missing value in "TotalTimeInBed" by 0 and combine "SedentaryMinutes" and "TotalTimeInBed" together as a new "TotalSedentaryMinutes" column.

2) In some cases, the "TotalActivityMinutes" are not exactly 1440. It is understandable if a value is below 1440 since the user might need to remove the fitness device in some circumstances like security check, taking exams, BIA evaluation or battery charging. However, for a value above 1440, I cannot figure out a reasonable explanation. Actually, if you take a look at the newly created "TotalActivityMinutes", you will find that 155 out of 940 observations have values above 1440.



Index	Id	ActivityDate	TotalSteps	VeryActiveMinutes	FairlyActiveMinutes	LightlyActiveMinutes	SedentaryMinutes	Calories	SleepDay	TotalMinutesAsleep	TotalTimeInBed	TotalActivityMinutes	TotalSedentaryMinutes
0	1503960366	2016-04-12 00:00:00	13162	25	13	328	728	1985	2016-04-12 00:00:00	327	346	1440	1074
1	1503960366	2016-04-13 00:00:00	10735	21	19	217	776	1797	2016-04-13 00:00:00	384	407	1440	1183
2	1503960366	2016-04-14 00:00:00	10460	30	11	181	1218	1776	NaT	nan	0	1440	1218
3	1503960366	2016-04-15 00:00:00	9762	29	34	209	726	1745	2016-04-15 00:00:00	412	442	1440	1168
4	1503960366	2016-04-16 00:00:00	12669	36	10	221	773	1863	2016-04-16 00:00:00	340	367	1407	1140
5	1503960366	2016-04-17 00:00:00	9705	38	20	164	539	1728	2016-04-17 00:00:00	700	712	1473	1251
6	1503960366	2016-04-18 00:00:00	13019	42	16	233	1149	1921	NaT	nan	0	1440	1149
7	1503960366	2016-04-19 00:00:00	15506	50	31	264	775	2035	2016-04-19 00:00:00	304	320	1440	1095
8	1503960366	2016-04-20 00:00:00	10544	28	12	205	818	1786	2016-04-20 00:00:00	360	377	1440	1195
9	1503960366	2016-04-21 00:00:00	9819	19	8	211	838	1775	2016-04-21 00:00:00	325	364	1440	1202
10	1503960366	2016-04-22 00:00:00	12764	66	27	130	1217	1827	NaT	nan	0	1440	1217

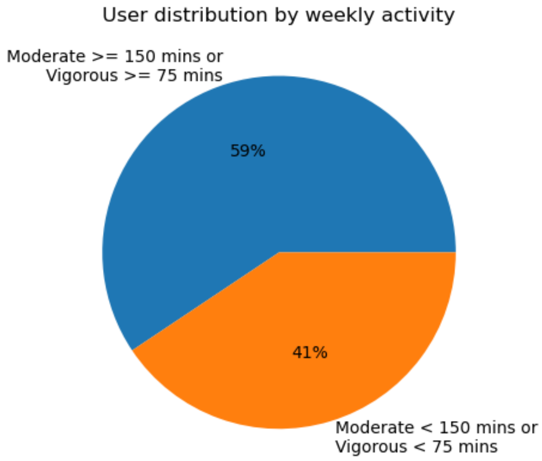
Since the data were directly collected by fitness devices, I assume that the records themselves are correct. There must be some algorithms in defining or aggregating the “SedentaryMinutes” and “TotalTimeInBed” that I fail to figure out. Therefore, I will continue to use original data for analyses on single activity. Meanwhile, in order to present a reliable result, I dropped observations with “TotalActivityMinutes” above 1440 for analyses that combined all activities such as user habit analysis and activity-wise correlation analysis.

2. Lifestyle Evaluation

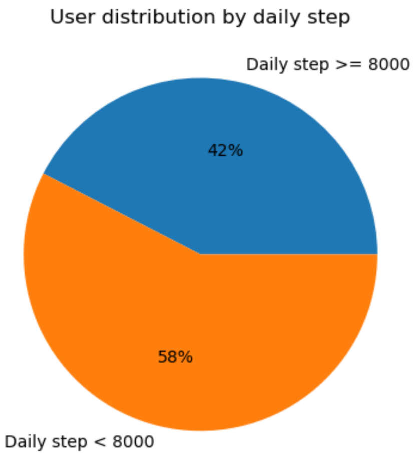
One of the fundamental functions of wearable fitness device is to help users check if they are having a healthy lifestyle. There are several guidelines regarding health maintenance based on the metrics presented in this dataset, e.g., weekly activity, daily steps and daily sleep. According to the physical activity guideline of WHO, an adult is recommended to perform at least 150–300 minutes of moderate-intensity aerobic physical activity or at least 75–150 minutes of vigorous-intensity throughout the week (ref: https://www.who.int/news-room/fact-sheets/detail/physical-activity). The US CDC recommends adults to have over 7 hours of sleep (ref: https://www.cdc.gov/sleep/about_sleep/how_much_sleep.html) for health maintenance and over 8,000 steps per day are beneficial for lowering risk of all-cause mortality (ref: https://www.cdc.gov/media/releases/2020/p0324-daily-step-count.html).

Therefore, let’s first check the proportion of users that lead a suggested healthy lifestyle.

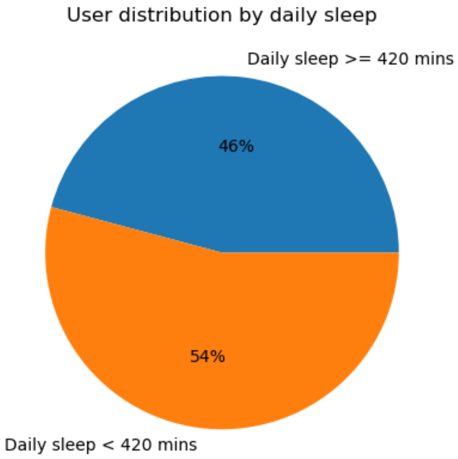
From the aspect of weekly physical activity, let’s assume “fairly active” is “moderate intensive” and “very active” is “vigorous intensive”, then 59% (n = 19) of the users met the suggested level.



In respect to daily steps, 42% (n=14) of the users met the suggested level.



For daily sleep time, 46% (n=11) of the users met the suggested level.



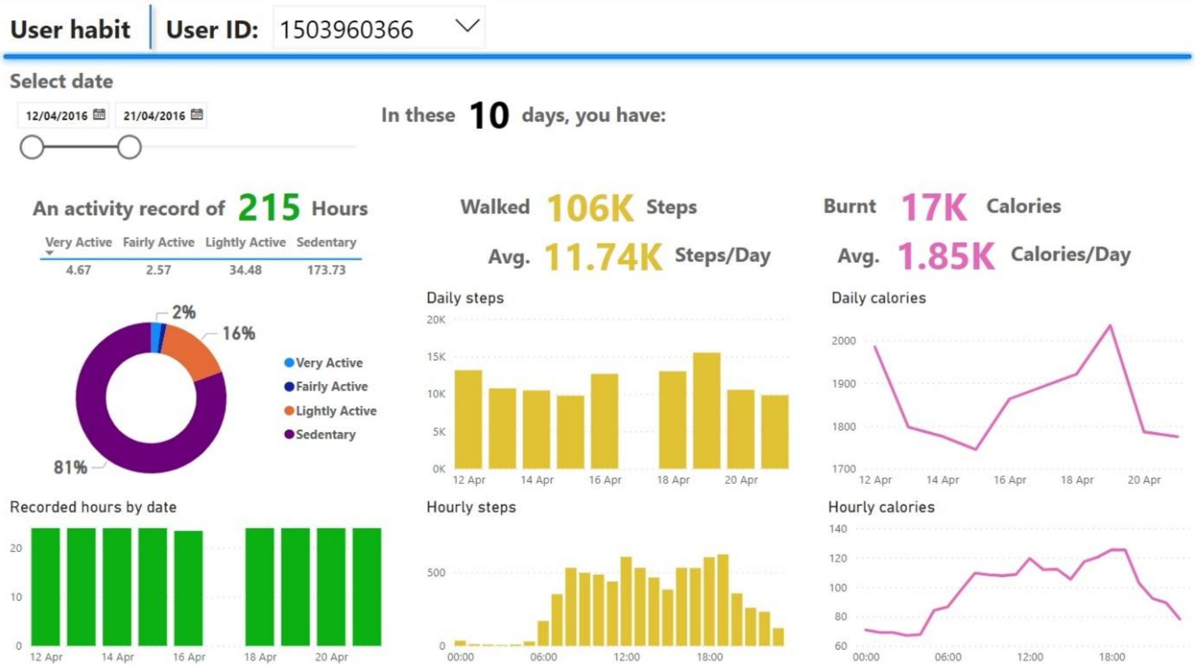
We can also generate personalised feedback report using the lifestyle_feedback() function in src/feedback.py

3. User Habit

ATTENTION: As discussed before, this part of analysis is based on users with “TotalActivityMinutes” no greater than 1440 mins.

The filtered dataset covers a timespan of 31 days (from 2016-04-12 to 2016-05-12). To check the usage of Fitbit by each user, we can call the usage_bar_chart () function in src/visualization.py.

Meanwhile, I create a PowerBI file (UserDescription.pbix) to present a more intuitive and interactive visualisation. We can easily check the Fitbit usage of a user over selected time periods and look into the data of a single day by highlighting the corresponding bar.



User habit | User ID: 1503960366

Select date

12/04/2016 | 21/04/2016

In these 10 days, you have:

An activity record of 215 Hours

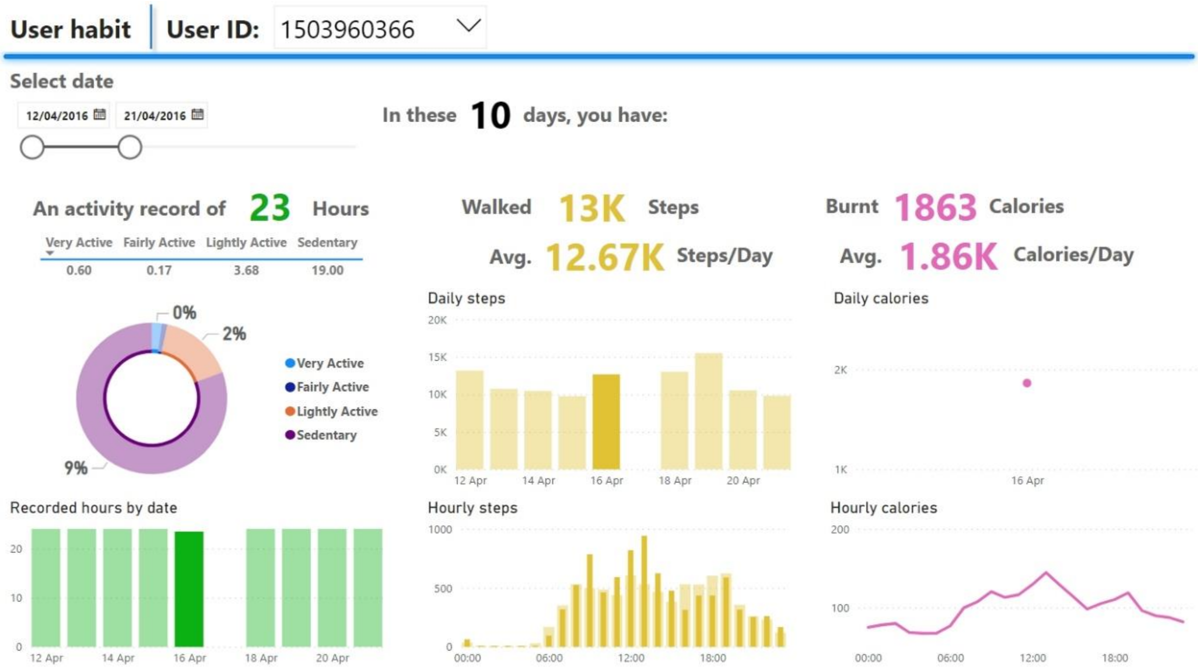
Very Active	Fairly Active	Lightly Active	Sedentary
4.67	2.57	34.48	173.73

Walked 106K Steps

Avg. 11.74K Steps/Day

Burnt 17K Calories

Avg. 1.85K Calories/Day



User habit | User ID: 1503960366

Select date

12/04/2016 | 21/04/2016

In these 10 days, you have:

An activity record of 23 Hours

Very Active	Fairly Active	Lightly Active	Sedentary
0.60	0.17	3.68	19.00

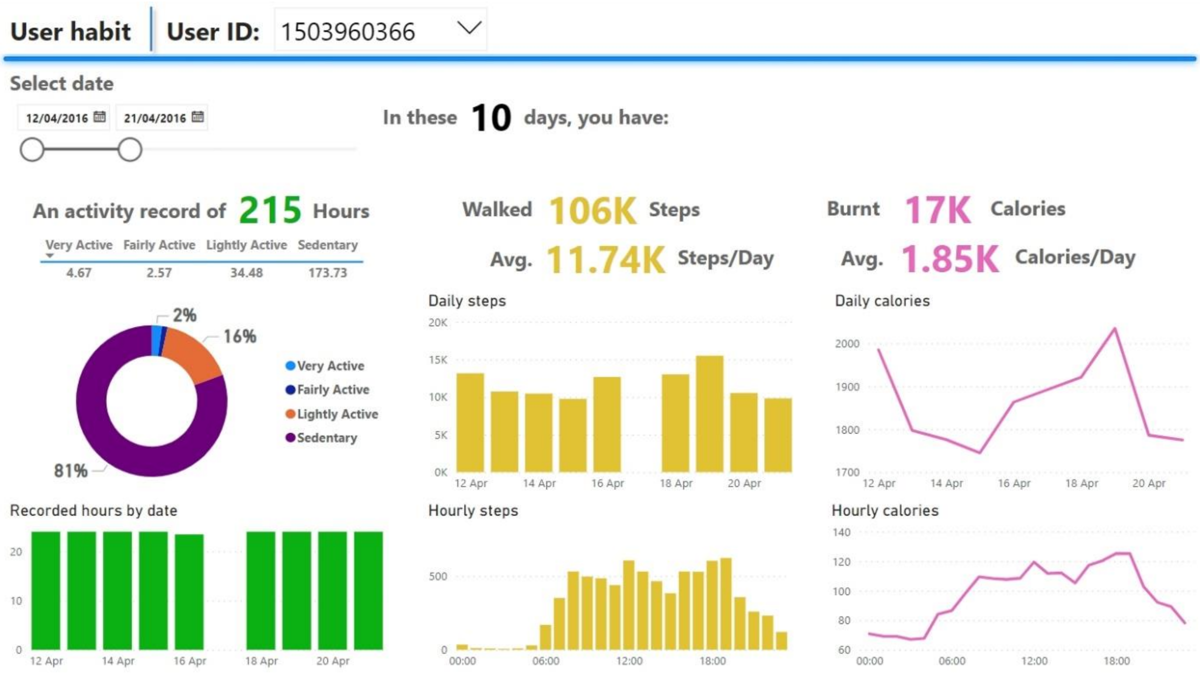
Walked 13K Steps

Avg. 12.67K Steps/Day

Burnt 1863 Calories

Avg. 1.86K Calories/Day

Now comes the question: are people who often use Fitbit physically more active than those who occasionally wear it? Let me first present four plots:



User habit | User ID: 1503960366

Select date

12/04/2016 21/04/2016

In these **10** days, you have:

An activity record of **215** Hours

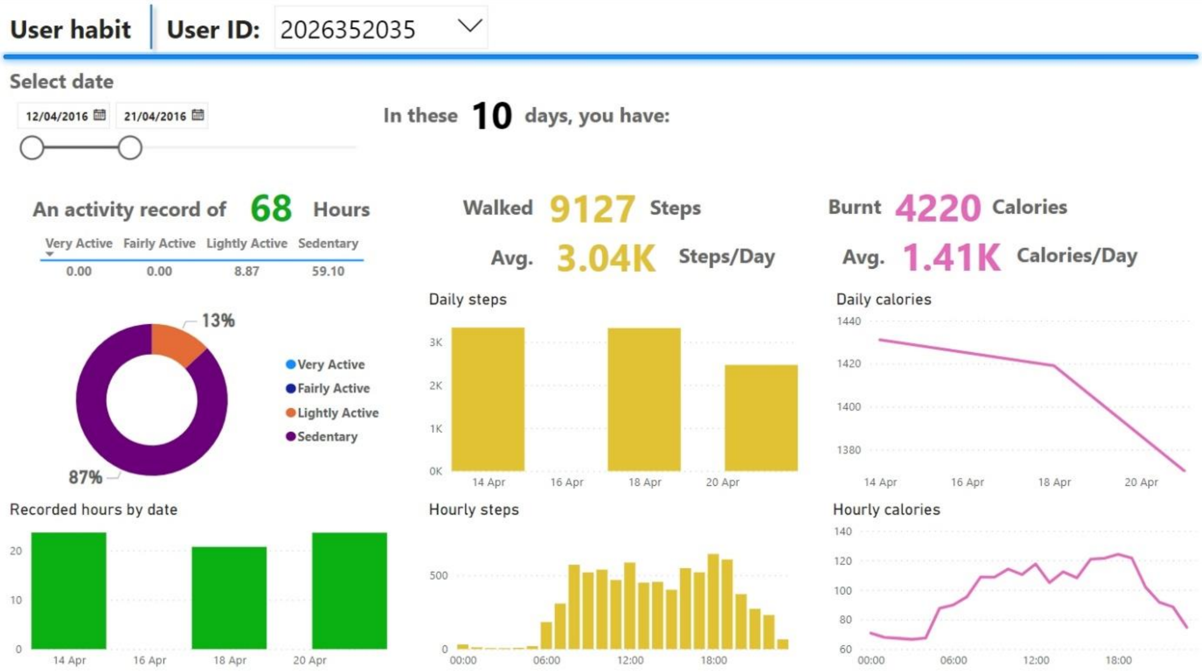
Very Active	Fairly Active	Lightly Active	Sedentary
4.67	2.57	34.48	173.73

Walked **106K** Steps

Avg. **11.74K** Steps/Day

Burnt **17K** Calories

Avg. **1.85K** Calories/Day



User habit | User ID: 2026352035

Select date

12/04/2016 21/04/2016

In these **10** days, you have:

An activity record of **68** Hours

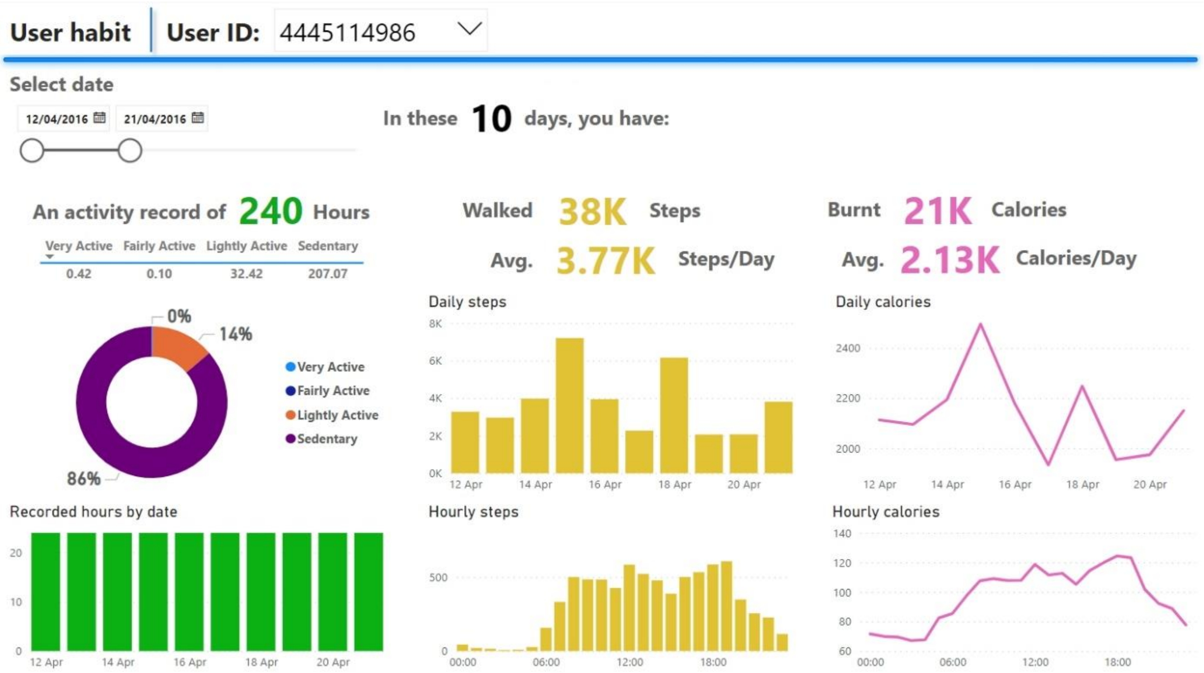
Very Active	Fairly Active	Lightly Active	Sedentary
0.00	0.00	8.87	59.10

Walked **9127** Steps

Avg. **3.04K** Steps/Day

Burnt **4220** Calories

Avg. **1.41K** Calories/Day



User habit | User ID: 4445114986

Select date

12/04/2016 – 21/04/2016

In these 10 days, you have:

An activity record of 240 Hours

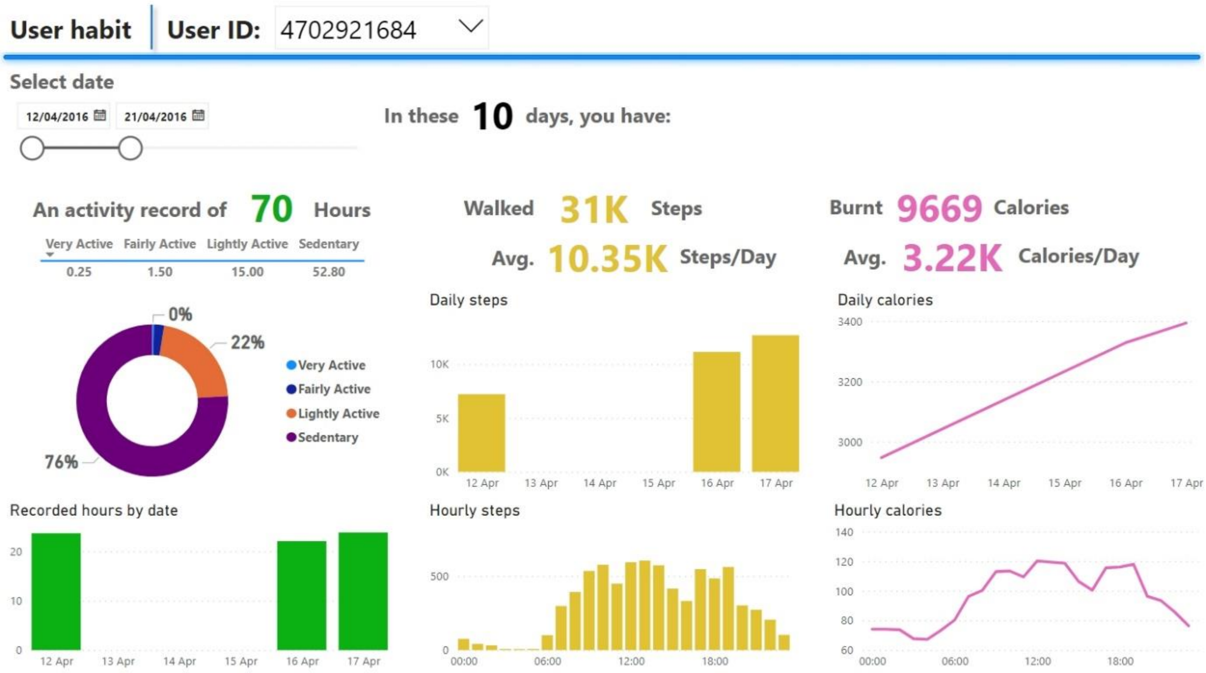
Very Active	Fairly Active	Lightly Active	Sedentary
0.42	0.10	32.42	207.07

Walked 38K Steps

Avg. 3.77K Steps/Day

Burnt 21K Calories

Avg. 2.13K Calories/Day



User habit | User ID: 4702921684

Select date

12/04/2016 – 21/04/2016

In these 10 days, you have:

An activity record of 70 Hours

Very Active	Fairly Active	Lightly Active	Sedentary
0.25	1.50	15.00	52.80

Walked 31K Steps

Avg. 10.35K Steps/Day

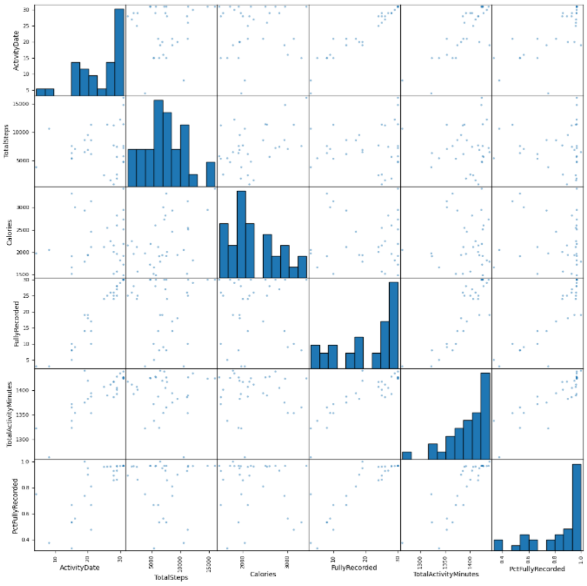
Burnt 9669 Calories

Avg. 3.22K Calories/Day

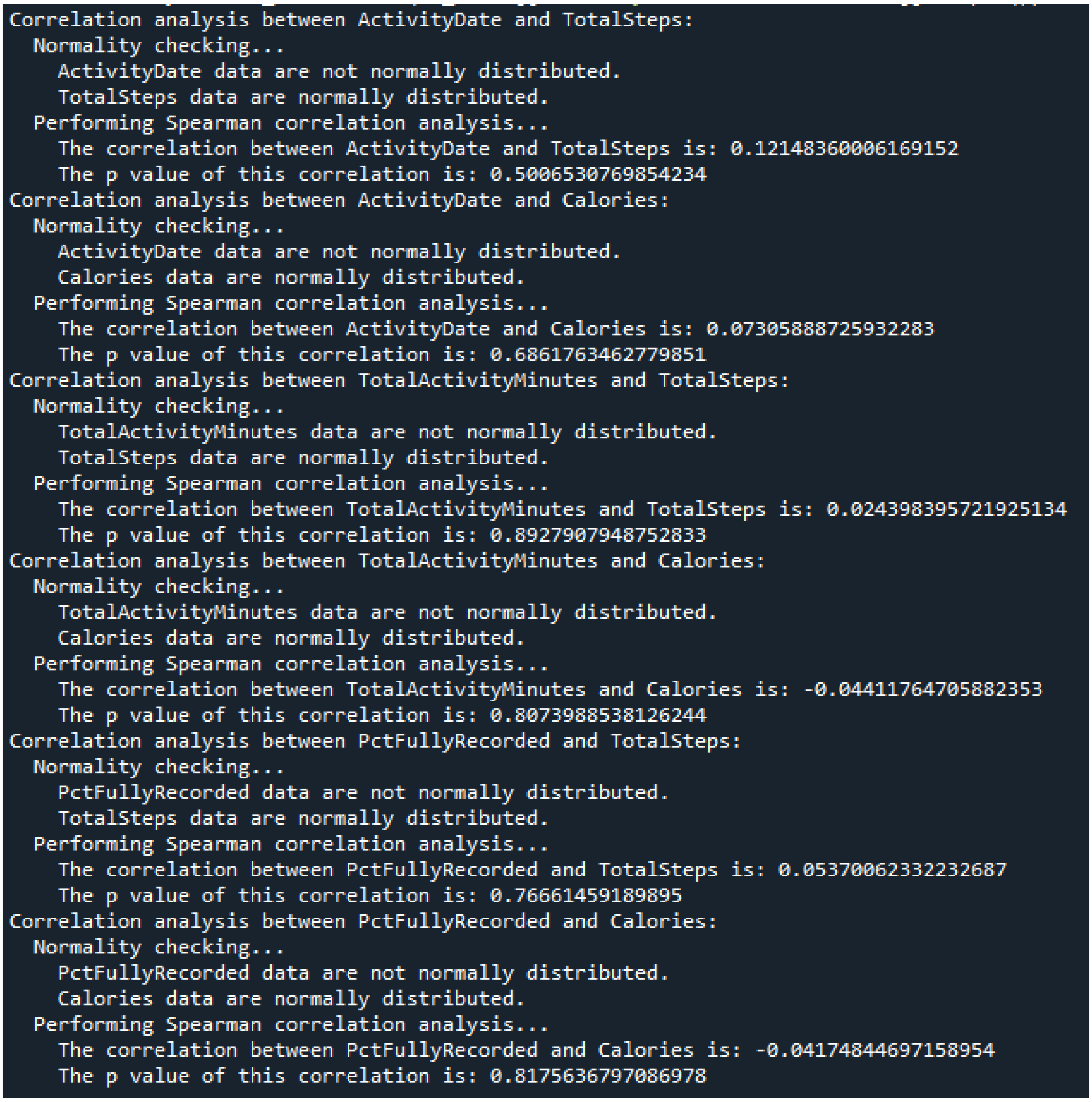
From the four plots above, it is a bit hard to tell if the frequency of device usage is related to daily steps or calories as we can see that users with a long device usage can still have fewer steps and calories than those who do not wear it so often. Yet, I would like to see if there is any trend in general.

To analyse the device usage, I decide to use two metrics: wearing days in the 31 days timespan and percentage of fully recorded days among wearing days. Since Fitbits take around 2 hours to be fully charged (ref: https://help.fitbit.com/manuals/charge_5/Content/manuals/Topics/Set%20Up/Charge%20Device.htm). I assume that a day with wearing time no less than 22 hours as being fully recorded.

From the scatter matrix, we cannot tell any clear linear relation between device usage and daily steps or calory expenditure.



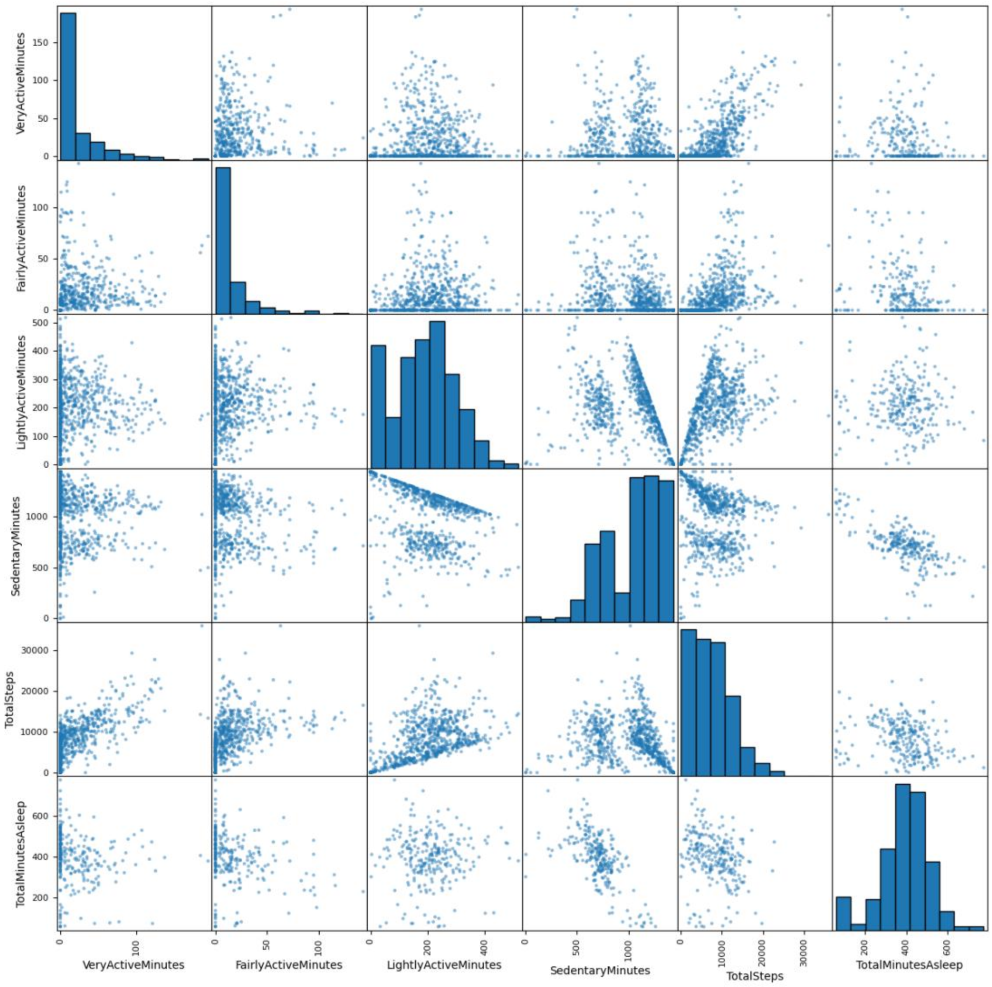
Correlation analyses further confirm that no significant correlation exits. Therefore, we cannot tell the activity level of a user by simple looking at the device usage habit.



```
Correlation analysis between ActivityDate and TotalSteps:
  Normality checking...
    ActivityDate data are not normally distributed.
    TotalSteps data are normally distributed.
  Performing Spearman correlation analysis...
    The correlation between ActivityDate and TotalSteps is: 0.12148360006169152
    The p value of this correlation is: 0.5006530769854234
Correlation analysis between ActivityDate and Calories:
  Normality checking...
    ActivityDate data are not normally distributed.
    Calories data are normally distributed.
  Performing Spearman correlation analysis...
    The correlation between ActivityDate and Calories is: 0.07305888725932283
    The p value of this correlation is: 0.6861763462779851
Correlation analysis between TotalActivityMinutes and TotalSteps:
  Normality checking...
    TotalActivityMinutes data are not normally distributed.
    TotalSteps data are normally distributed.
  Performing Spearman correlation analysis...
    The correlation between TotalActivityMinutes and TotalSteps is: 0.024398395721925134
    The p value of this correlation is: 0.8927907948752833
Correlation analysis between TotalActivityMinutes and Calories:
  Normality checking...
    TotalActivityMinutes data are not normally distributed.
    Calories data are normally distributed.
  Performing Spearman correlation analysis...
    The correlation between TotalActivityMinutes and Calories is: -0.04411764705882353
    The p value of this correlation is: 0.8073988538126244
Correlation analysis between PctFullyRecorded and TotalSteps:
  Normality checking...
    PctFullyRecorded data are not normally distributed.
    TotalSteps data are normally distributed.
  Performing Spearman correlation analysis...
    The correlation between PctFullyRecorded and TotalSteps is: 0.05370062332232687
    The p value of this correlation is: 0.76661459189895
Correlation analysis between PctFullyRecorded and Calories:
  Normality checking...
    PctFullyRecorded data are not normally distributed.
    Calories data are normally distributed.
  Performing Spearman correlation analysis...
    The correlation between PctFullyRecorded and Calories is: -0.04174844697158954
    The p value of this correlation is: 0.8175636797086978
```

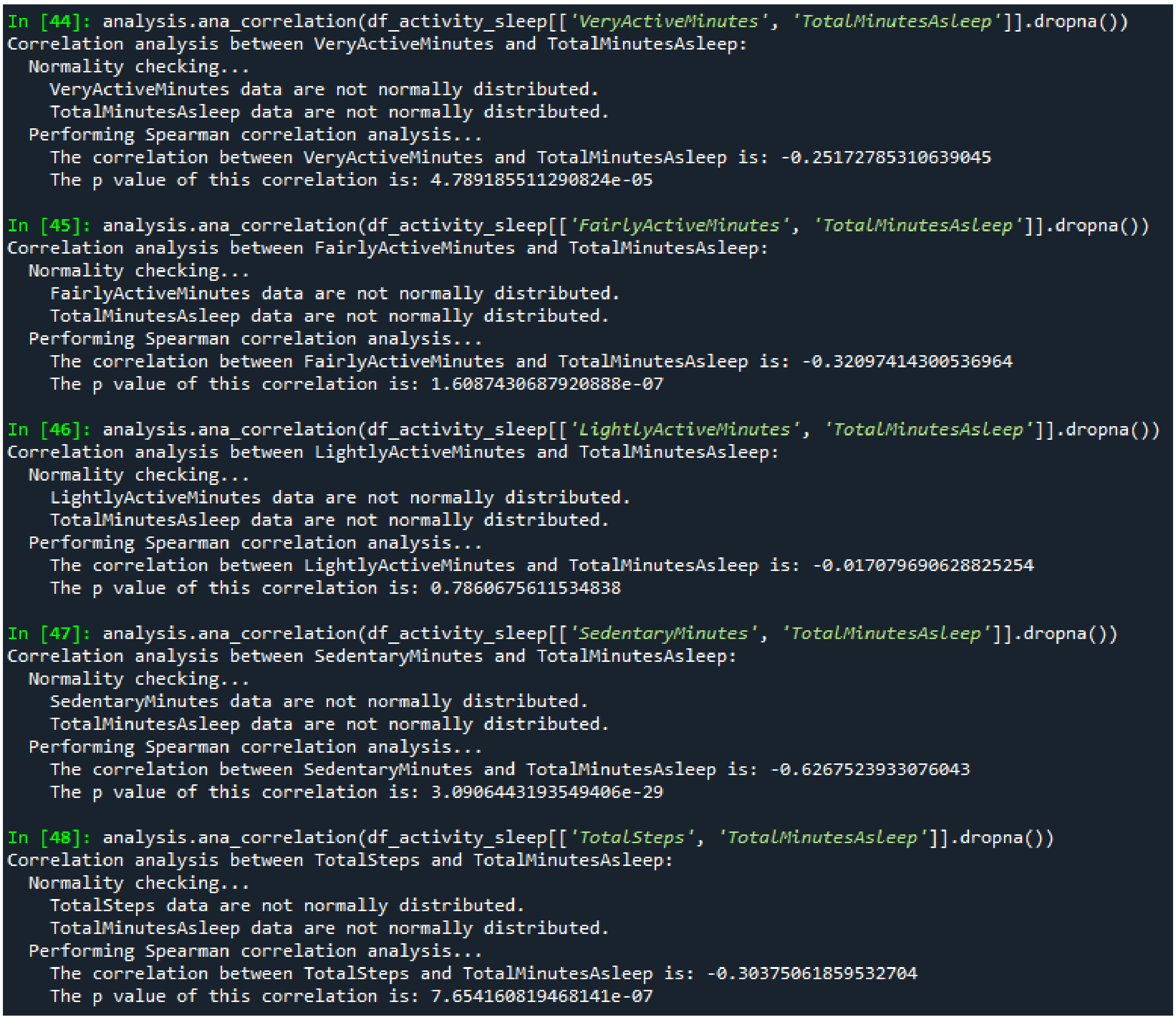
4. Activity and Sleep Analysis

Now comes the last question: are physical activity and daily steps related to sleep time? Before answering this question, let’s take a peek at the correlations between these variables by a scatter matrix.



From the graph, we can see that there are not clear linear correlations between sleep time and other variables, except for “sedentary minutes”, in which we can clear tell a negative correlation between these two variables.

Now let’s analyse the linear associations between sleep time and other variables one by one using the ana_correlation() function in src/analysis.py. The function first performs a normality test to check if the input data are normally distributed, if so, Pearson correlation analysis is performed, otherwise, Spearman correlation analysis is used. Results of the current data are presented as follows:



```
In [44]: analysis.ana_correlation(df_activity_sleep[['VeryActiveMinutes', 'TotalMinutesAsleep']].dropna())
Correlation analysis between VeryActiveMinutes and TotalMinutesAsleep:
  Normality checking...
    VeryActiveMinutes data are not normally distributed.
    TotalMinutesAsleep data are not normally distributed.
  Performing Spearman correlation analysis...
    The correlation between VeryActiveMinutes and TotalMinutesAsleep is: -0.25172785310639045
    The p value of this correlation is: 4.789185511290824e-05

In [45]: analysis.ana_correlation(df_activity_sleep[['FairlyActiveMinutes', 'TotalMinutesAsleep']].dropna())
Correlation analysis between FairlyActiveMinutes and TotalMinutesAsleep:
  Normality checking...
    FairlyActiveMinutes data are not normally distributed.
    TotalMinutesAsleep data are not normally distributed.
  Performing Spearman correlation analysis...
    The correlation between FairlyActiveMinutes and TotalMinutesAsleep is: -0.32097414300536964
    The p value of this correlation is: 1.6087430687920888e-07

In [46]: analysis.ana_correlation(df_activity_sleep[['LightlyActiveMinutes', 'TotalMinutesAsleep']].dropna())
Correlation analysis between LightlyActiveMinutes and TotalMinutesAsleep:
  Normality checking...
    LightlyActiveMinutes data are not normally distributed.
    TotalMinutesAsleep data are not normally distributed.
  Performing Spearman correlation analysis...
    The correlation between LightlyActiveMinutes and TotalMinutesAsleep is: -0.017079690628825254
    The p value of this correlation is: 0.7860675611534838

In [47]: analysis.ana_correlation(df_activity_sleep[['SedentaryMinutes', 'TotalMinutesAsleep']].dropna())
Correlation analysis between SedentaryMinutes and TotalMinutesAsleep:
  Normality checking...
    SedentaryMinutes data are not normally distributed.
    TotalMinutesAsleep data are not normally distributed.
  Performing Spearman correlation analysis...
    The correlation between SedentaryMinutes and TotalMinutesAsleep is: -0.6267523933076043
    The p value of this correlation is: 3.0906443193549406e-29

In [48]: analysis.ana_correlation(df_activity_sleep[['TotalSteps', 'TotalMinutesAsleep']].dropna())
Correlation analysis between TotalSteps and TotalMinutesAsleep:
  Normality checking...
    TotalSteps data are not normally distributed.
    TotalMinutesAsleep data are not normally distributed.
  Performing Spearman correlation analysis...
    The correlation between TotalSteps and TotalMinutesAsleep is: -0.30375061859532704
    The p value of this correlation is: 7.654160819468141e-07
```

As presented in the results, weak negative associations are found between sleep time and three variables: very active minutes, fair active minutes and daily steps. Moderate negative association is found between sleep time and sedentary minutes. Such moderate association is not a surprise considering the mystery between time in bed and sedentary minutes discussed at the beginning of this report.

5. Summary

In summary, we have:

1) evaluated the lifestyle of the users based on thresholds recommended by health organisations, we can nudge users to be more active if they fail to meet these thresholds.

2) created an interactive dashboard to present user habit. It will be helpful for us to understand the activity pattern of the user and give personalised suggestions according to day of week (not presented but created in the PowerBI file as it is based on the unfiltered dataset) and hour. Additionally, for users who occasionally wear the device, we can encourage them to wear the device more often by showing them the importance of continuous monitoring on physical status and lifestyle patterns in health maintenance. For users who often use the device, we can give them more challenging tasks to help them stay active.

3) not found any significant linear correlation between device usage and physical activity level. Weak associations are found between sleep time and three activity variables: very active minutes, fair active minutes and daily steps. However, we should treat these findings with caution due to the limited sample size and the weird negative association.