Data Analysis Documentation

1. Project Overview

Project Name: Leveraging Bellabeat trends and products using wearable device customer data
Date: March 5th, 2025
Author: Farhan Hauzan Nadhir
Version: [v1.0]

1.1 Business Understanding

Business Objective:
To analyze smart device usage data to gain insight into how people are already using their smart devices. And applying data insights towards one of the Bellabeat existing products, and the consumer to leverage the user trends and satisfaction.

Key Business Questions:

1. What are some trends in smart device usage?
2. How could these trends apply to Bellabeat customers?
3. How could these trends help influence Bellabeat marketing strategy?

Stakeholders:
Bellabeat's co-founder and Chief Creative Officer, Bellabeat executive team, Marketing Team.

Key Performance Indicators (KPIs):
High-level recommendations for how these trends can inform Bellabeat marketing strategy.

Initial Hypotheses:
This data might give insights into the trend in current lifestyle shifts towards modern culture of using smart devices. Analysts focus on the data that serves the information about the user's physical activity.

Limitation:
The data provided is gathered in 2016, which is about ~10 years old, and only covers ~30 users that are not equipped with user demographic information.

2. Data Understanding

Deliverables: A description of all data sources used.

2.1 Data Sources

Data Provenance: These datasets were generated by respondents to a distributed survey via Amazon Mechanical Turk between 03.12.2016-05.12.2016. Thirty eligible Fitbit users consented to submitting personal tracker data, including minute-level output for physical activity, heart rate, and sleep monitoring. Individual reports can be parsed by export session ID (column A) or timestamp (column B). Variation between output represents the use of different types of Fitbit trackers and individual tracking behaviours/preferences.

Dataset Authors: Furberg, R., Brinton, J., Keating, M., & Ortiz, A. (2016). Crowd-sourced Fitbit datasets 03.12.2016-05.12.2016 [Data set]. Zenodo. https://doi.org/10.5281/zenodo.53894

Collection Methodology: Pre-processed

Format: .csv

Dimensions format: narrow, wide

Dataset Name	Description	Source	Update Frequency	Owner
FitBit Fitness Tracker Data	This dataset contains thirty eligible Fitbit users who consented to submitting their personal tracker data.	Kaggle https://www.kaggle.com/datasets/arashnic/fitbit/data	Last update a year ago	MÖBIUS

2.2 Data Dictionary

dailyActivities_merged.csv:

Column Name	Description	Data Type	Potential Issues
Id	Unique identifier for user	Numeric	Some users might need to be dropped due to insufficient data
ActivityDate	The date of activity “4/5/2016” m/dd/yyyy format	Date	Date format needs to be standardized, Several dates are out of time frame exist
TotalSteps	Total of user steps taken by day	Numeric	Outliers possible
TotalDistance	Total distance travelled by day	Numeric	
TrackerDistance	Total distance tracked by day	Numeric	Redundancy with TotalDistance
LoggedActivitiesDistance		Numeric	90% of the entries are zero value
VeryActiveDistance	Distance travelled during a very active day in km	Numeric	“Right skewed data”
ModeratelyActiveDistance	Distance travelled during a moderately active day in km	Numeric	“Right skewed data”
LightActiveDistance	Distance travelled during a light active day in km	Numeric	“Right skewed data”
SedentaryActiveDistance	Distance travelled during a sedentary active day in km	Numeric	“Right skewed data”
FairlyActiveMinutes	Period during a very active day in minutes	Numeric	“Right skewed data”

FairlyActiveMinutes	Period during a fairly active day in minutes	Numeric	“Right skewed data”
LightlyActiveMinutes	Period during a lightly active day in minutes	Numeric	“Right skewed data”
SedentaryMinutes	Period during a sedentary day in minutes	Numeric	
Calories	Number of calories spent in a day	Numeric	

dailyData_merged.csv (generated from sleep, steps, intensity, and MET table)

Column Name	Description	Data Type	Potential Issues
user_id	Unique identifier for user	Numeric	Some users might need to be dropped due to insufficient data
date	The date of activity in %m-%d-%Y	Date	
calories_expenditure	The total number of calories burned by users in that day	Numeric	
total_steps	The total number of steps taken by a user in that day	Numeric	
total_intensity	A cumulative measure of users’ activity intensity level in that day	Numeric	
average_intensity	Average measure of users’ activity intensity level in that day	Numeric	

energy_expenditure	The total amount of energy used by the user in that day	Numeric	
sleep_mins	The total duration of sleep in minutes	Numeric	
drowsy_mins	The total amount of time (in minutes) spent in a drowsy state	Numeric	
awake_mins	The total number of minutes spent awake during the recorded period, including interruptions in sleep.	Numeric	
totalBed_time	The total time (in minutes) spent in bed from each user	Numeric	

2.3 Data Quality Issues

Overall, we found that the main limitations are the absence of user demographic information and a rather small sample size for every table that we have, which might even become smaller after dropping ineffective records. After doing basic statistical analysis in the four datasets given, there is variance in the date count for every user. Though most of the data shows almost zero and na value, we need to reconsider which row to drop, because it might indicate a user using the device partially in their day, or even almost not using it at all. We are going to complete the data first by force filling the date column, so we can drop some users that might have missing key values. And we are going to fill them in with the NA for manageable further EDA analysis as indicators they are not using the device.

3. Process

Deliverable: Documentation of any manipulation or cleaning of the data.

dailyActivities_merged.csv

For data preprocessing, I am using both SQL and R to identify all the data presented. First, we are going to merge two sets of data of *dailyActivities_merge* to get the full range of periods. Looking at the daily activities data, we acknowledge there are some cleaning needed to gain insight on how Fitbit users use their devices. There are some limitations regarding the data, such as stale data from 2016 and insufficient samples throughout the period.

These are the processes and cleaning steps applied to the dataset;

- Merged two datasets from two periods.
- Drop LoggedActivitiesDistance column due to huge numbers of 0 values.
- Dropped TrackerDistance column due to data redundancy to TotalDistance column.
- Removed data entries that are not in the main period.
- Removed user_id which has low and insufficient data to eliminate bias.
- Change Incorrect data types: e.g., date.
- Change and standardize column names implementing snake cases.
- Dropped/removed rows that have NA / 0.0 of steps column value, considered as not wearing the device for the whole day.

dailyData_merged.csv

To generate this complete dataset regarding user data, there are several processes taken from smaller datasets. Through querying and merging the tracker's data in the minute into a complete daily tracker dataset. These are the datasets that are processed and merged into one complete tracker data set: METs, Calories, Steps, and Sleep.

These are the processes and cleaning steps applied to the dataset;

- Generate daily datasets by using GROUP BY to date in every hourly dataset.
- Force Filling those datasets to smoothen the JOIN process.
- Standardized missing value and impossible value to NULL.
- Change Incorrect data types: e.g., date.
- Change and standardize column names implementing snake cases.
- Dropped/removing rows that have NA / =< 1.0 of MET value, considered as not wearing the device for the whole day.
- Removed user_id which has insufficient sample to eliminate bias. Example: user_id “2891001357” has only 1 record.

skim_variable	**n_missing**	**complete_rate**	**numeric.mean**	**numeric.sd**	**numeric.p0**	**numeric.p100**	**numeric.hist**
user_id	0	1	4882201218.531	2468399281.12391	1503960366	8877689391	▇▅▃▅▅
date_count	0	1	51.625	11.49123	24	62	▂▁▂▃▇

Here's the list of smaller datasets that have been cleaned and generated:

- dailyActivities_dataset,
  Generated from;

  ✓ dailyActivities_merged.csv (3.12.16-4.11.16)
  ✓ dailyActivities_merged.csv (4.12.16-5.12.16)

- dailyData_dataset,

  Generated from;

  ✓ dailySleep_merged.csv,
  ✓ dailySteps_merged.csv,
  ✓ dailyCalories_merged.csv,
  ✓ dailyMETs_merged.csv,
  ✓ hourlySleep_merged.csv,
  ✓ hourlySteps_merged.csv,
  ✓ hourlyCalories_merged.csv,
  ✓ hourlyMETs_merged.csv

4. Analyze

Deliverables: A summary of your analysis.

4.1 Summary Statistics

dailyActivities_datasets.csv:

using the str() function, we can observe some statistical metrics that prompt further investigation.

skim_type	skim_variable	n_missing	complete_rate	Date.n_unique	numeric.mean	numeric.sd	numeric.p0	numeric.p25	numeric.p50	numeric.p75	numeric.p100	numeric.hist
Date	date	0	1	62	NA	NA	NA	NA	NA	NA	NA	NA
numeric	user_id	0	1	NA	4777439620.52	2403518679.11	1503960366	2320127002.00	4445114986.00	6962181067.00	8877689391.00	▇▅▂▅▅
numeric	total_steps	0	1	NA	8096.65	4867.39	4	4507.50	7656.00	10997.50	36019.00	▇▇▂▁▁
numeric	total_distance	0	1	NA	5.80	3.79	0	3.10	5.41	7.80	28.03	▇▆▁▁▁
numeric	veryActive_distance	0	1	NA	1.56	2.71	0	0.00	0.33	2.13	21.92	▇▁▁▁▁
numeric	moderatelyActive_distance	0	1	NA	0.60	0.90	0	0.00	0.27	0.83	6.48	▇▁▁▁▁
numeric	lightActive_distance	0	1	NA	3.55	1.93	0	2.19	3.53	4.87	12.51	▆▇▂▁▁

numeric	sedentaryActive_distance	0	1	*NA*	0.00	0.01	0	0.00	0.00	0.00	0.11	▇▁▁▁▁
numeric	veryActive_minutes	0	1	*NA*	21.82	32.70	0	0.00	5.00	33.00	210.00	▇▂▁▁▁
numeric	fairlyActive_minutes	0	1	*NA*	14.40	20.53	0	0.00	8.00	20.00	143.00	▇▁▁▁▁
numeric	lightlyActive_minutes	0	1	*NA*	205.19	99.33	0	141.00	206.00	270.00	586.00	▂▇▆▁▁
numeric	sedentaryActive_minutes	0	1	*NA*	951.49	289.30	0	720.00	1009.00	1189.00	1440.00	▁▁▇▆▆
numeric	calories_expenditure	0	1	*NA*	2337.93	731.32	50	1848.25	2201.50	2819.25	4900.00	▁▆▇▂▁

After we clean and remove users that have insufficient records, <90% of the data we have of the total **33 sample users**. This data mainly tracks users' activity, we can see that users' average time in the sedentary state of **951 minutes** significantly outweighs their active time of **241 minutes**, while mostly these users remain lightly active in their day. The mean distance users travelled in a day was **5.41 kilometers** throughout the day with a standard deviation of 3.8 kilometres.

dailyData_datasets.csv

By briefly examining the merged calculated tracker data using the str() function, we can observe some statistical metrics that prompt further investigation.

skim_type	skim_variable	n_missing	complete_rate	Date.n_unique	numeric.mean	numeric.sd	numeric.p0	numeric.p25	numeric.p50	numeric.p75	numeric.p100	numeric.hist
Date	date	0	1.0000000	62	NA	NA	NA	NA	NA	NA	NA	NA
numeric	user_id	0	1.0000000	NA	4816593701.3934164	2417454679.9300656	1503960366.000000	2320127002.0000000	4558609924.0000000	6962181067.0000000	8877689391.000	
numeric	calories_expenditure	0	1.0000000	NA	2389.8347458	705.3202218	247.000000	1888.7500000	2250.0000000	2832.0000000	5453.000	
numeric	total_steps	2	0.9987893	NA	8412.3769697	4919.2523093	14.000000	4735.0000000	7904.5000000	11263.0000000	37322.000	
numeric	total_intensity	0	1.0000000	NA	317.0944310	151.8918238	1.000000	205.7500000	313.0000000	412.0000000	991.000	
numeric	average_intensity	0	1.0000000	NA	0.3391383	0.1512441	0.016667	0.2376644	0.3257985	0.4209491	1.275	
numeric	average_metabolicEquivalent	0	1.0000000	NA	1.5183111	0.2697249	1.010000	1.3200000	1.5100000	1.6700000	2.690	
numeric	sleep_mins	764	0.5375303	NA	398.4549550	152.0590292	0.000000	333.7500000	419.0000000	490.0000000	1108.000	

numeric	drowsy_mins	764	0.5375303	*NA*	4.5259009	8.8973480	0.000000	1.0000000	3.0000000	5.0000000	129.000	▇▁▁▁▁
numeric	awake_mins	764	0.5375303	*NA*	30.3479730	35.9975803	0.000000	13.0000000	21.0000000	33.0000000	306.000	▇▁▁▁▁
numeric	totalBed_time	764	0.5375303	*NA*	433.3288288	160.1997625	4.000000	377.0000000	450.5000000	522.2500000	1149.000	▂▇▇▁▁

After we clean and remove users that have insufficient records <90% of data, the number of users results in **32 sample** users. The duration period of the device recording still goes out **62 days**, but Actual usage may vary based on individual preferences and habits, seen by the variance of date records for each user. Average calories expenditure spent by these users is **2389 kcal/day,** which is categorized as moderately active people. Total steps taken by users in a day averaged at **8412 steps/day,** which is also considered as a moderately active person. And the median sleep time for users is around **419 minutes,** or they take ~7 hours of sleep daily.

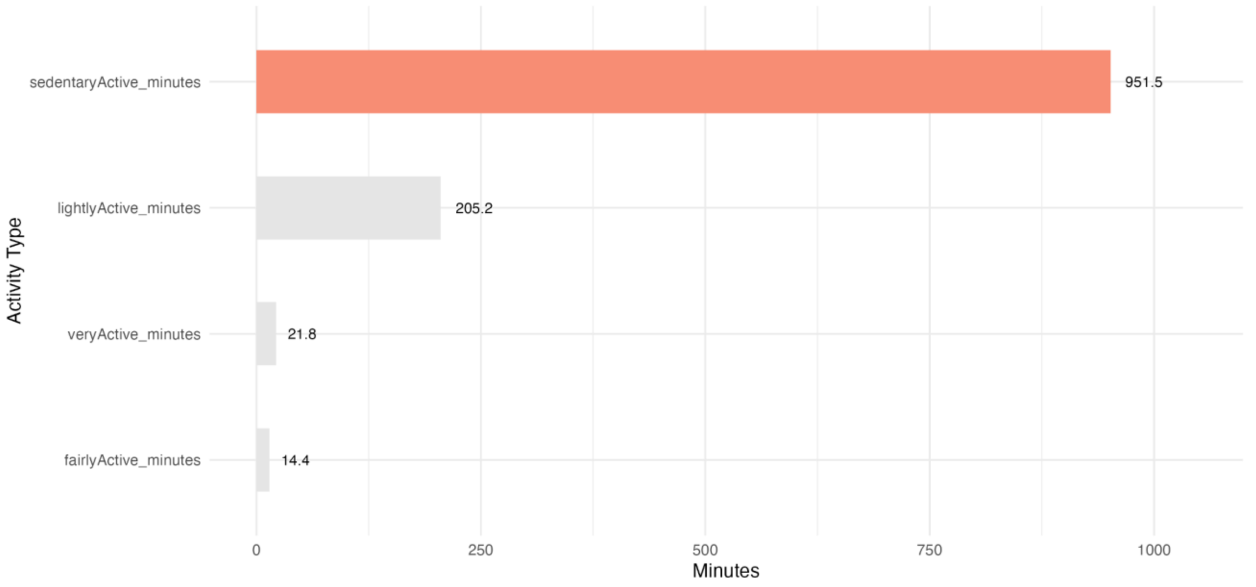
Another fact that we can see from this data is NA value for sleeping from active records (another column is filled) is missing of **764 of entries** which 48% of the time the smart devices failed to record or almost half of users are not preferred using this device while sleeping, though this assumption will need further investigation. Data captured presents a long right tail distribution that might come from occasional heavy physical activity.

4.2 Data Visualizations

- ✓ Now we are going to categorize these users' level of physical activity by seeing their number of steps and METs or metabolic equivalent throughout the day.
- ✓ And seeing the distributions of these users in terms of their level of physical activity.
- ✓ We also can see the average data of total steps, sleep, and calories in a weekly manner each day.
- ✓ We can also see their activity throughout the day by making a heat map on metabolic equivalent values based on user category.
- ✓ We can also see how often they wear the device. And making a user profiling based on their engagement using the device.
- ✓ See the day of highest calories expenditure, are they having longer sleep time (rest)

Understanding when and how the trend of smart devices and user behaviour

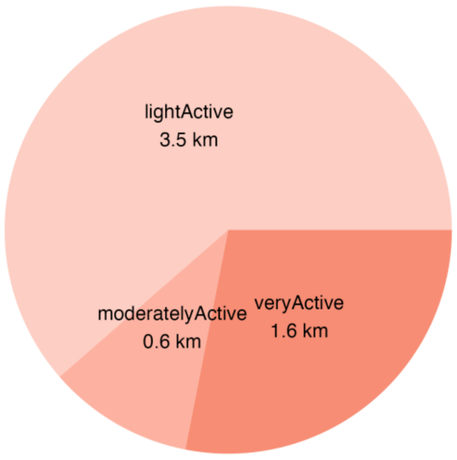
User average activity types (minutes)



The chart describes a significant number of minutes that users spend two-thirds of their day being sedentary (80% of total time in a day). Then they do light physical activity such as walking, standing, and simple household tasks.

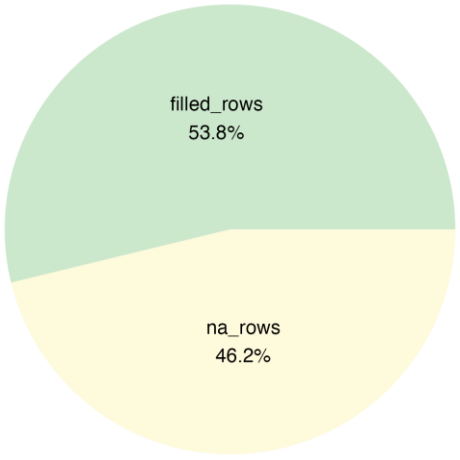
We can assume these users are modern individuals who lead a sedentary lifestyle in an urban environment.

Average Activity Distance Breakdown



Avg distance distribution of activities

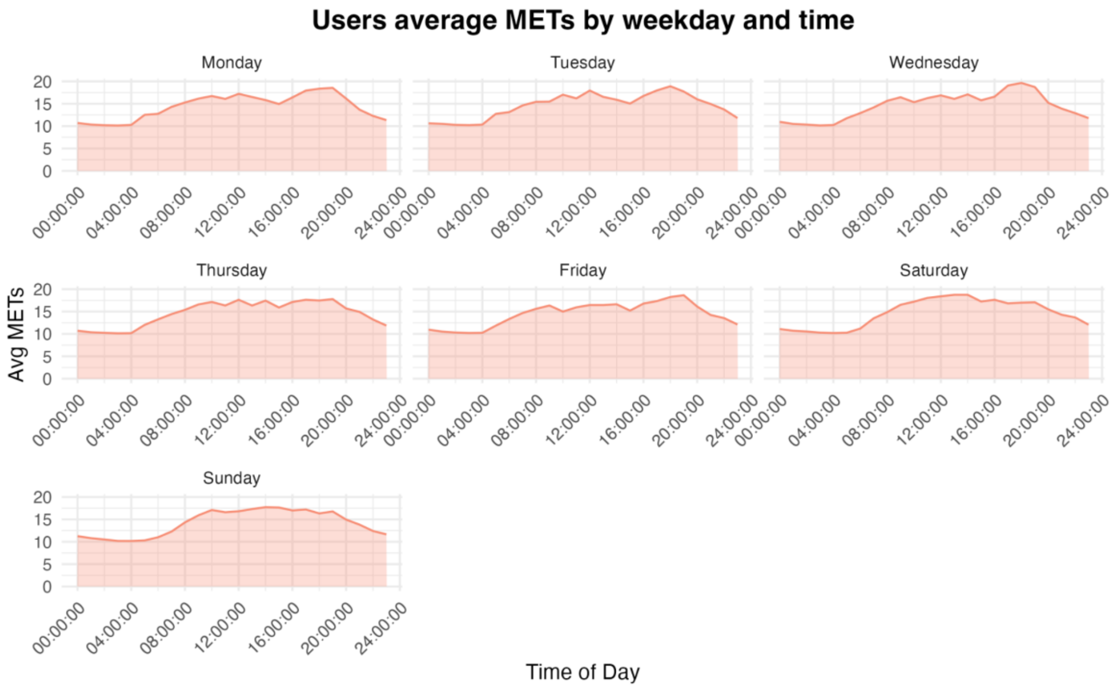
Distribution of Filled Sleep records



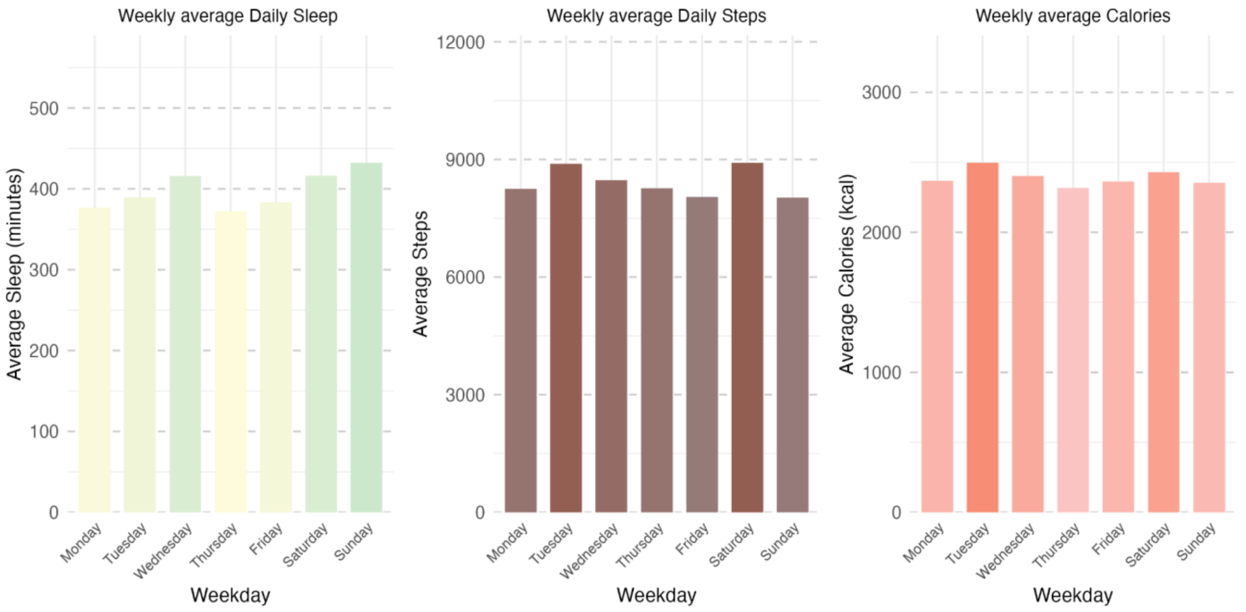
Sleep Records Percentage

From the bar chart above, we see that most distances these users travelled are spent lightly, while the very active period is only 1.6 kilometres/day, considered as really low in physical exercise.

The existence of high records of na_rows in sleeping records (46.2%) in the data can be assumed as the users are not wearing the device on them. This might imply the uncomfortable feeling of using a device while they are sleeping.

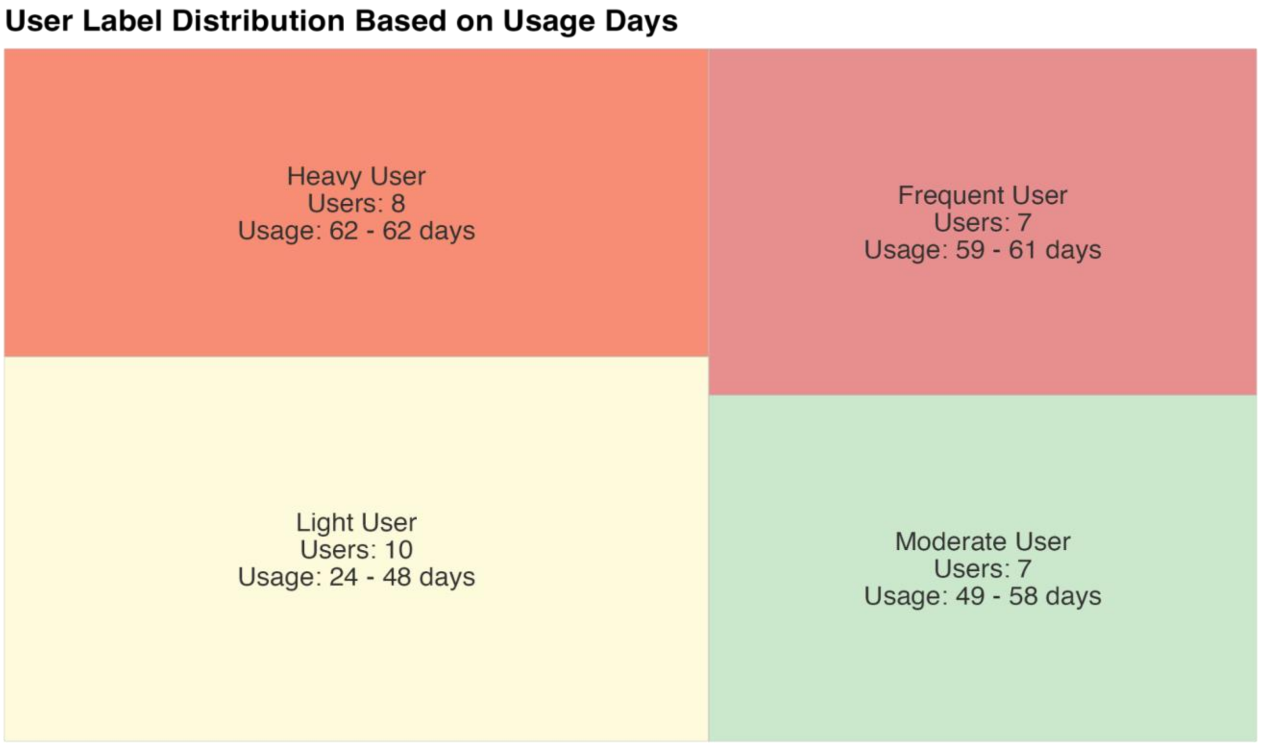


MET (Metabolic Equivalent of Task) is a unit that measures the energy cost of physical activities, reflecting the intensity of a user's physical exertion. On average, users engage in higher-intensity activities, especially at night. As shown in the chart, activity levels decline around 4:00 PM but rise again around 7:00 PM, likely after work hours. However, the level of intensity remains within a light activity range (<2.0).



The weekly average of daily steps closely follows the trend of calorie expenditure among smart device users, showing a notable increase in Tuesday and Saturday. These two metrics exhibit a positive correlation. Additionally, sleep duration tends to be shorter on weekdays compared to weekends, except on Wednesday.

Smart device users segmentation



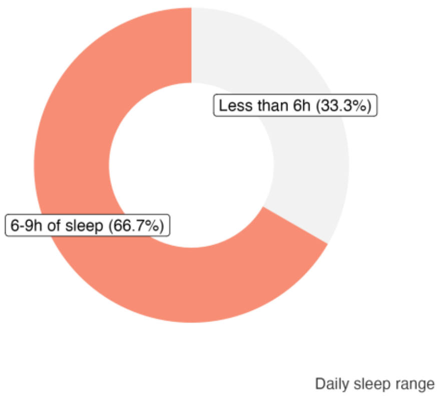
Most of the smart device users (68%) use the device pretty frequently, around 62 to 49 days, throughout 2 months of data taken.

The other 32% of users use around 24 to 48 days in total, considered as lightly using smart devices, but still considered a good frequency of use.

This shows that they possess the device intending to track their health and physical activity, mostly on a daily basis.

Heavy Users

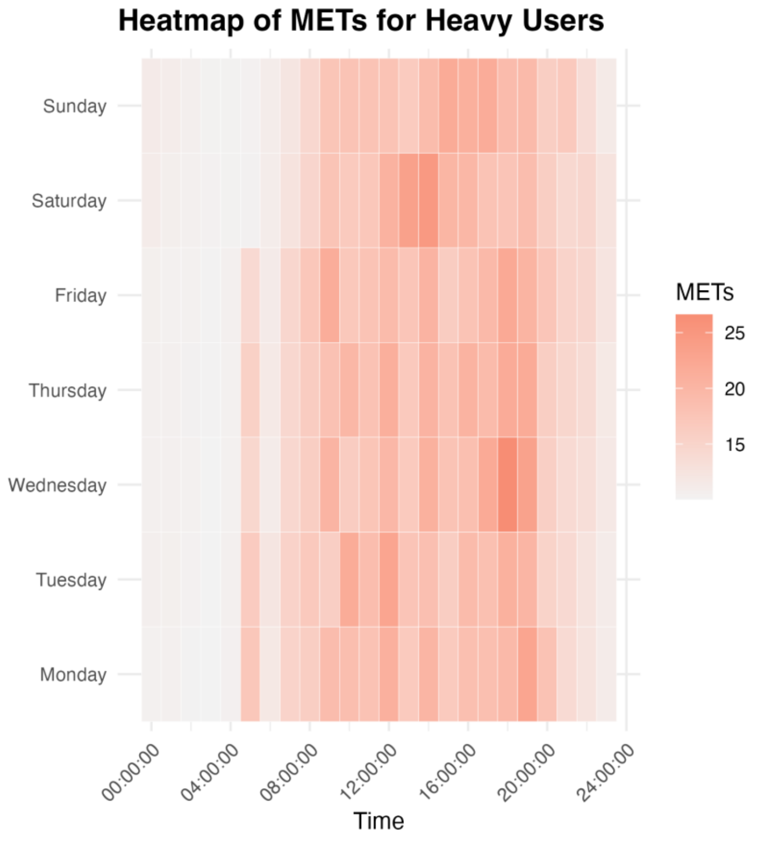
62 out of 62 days



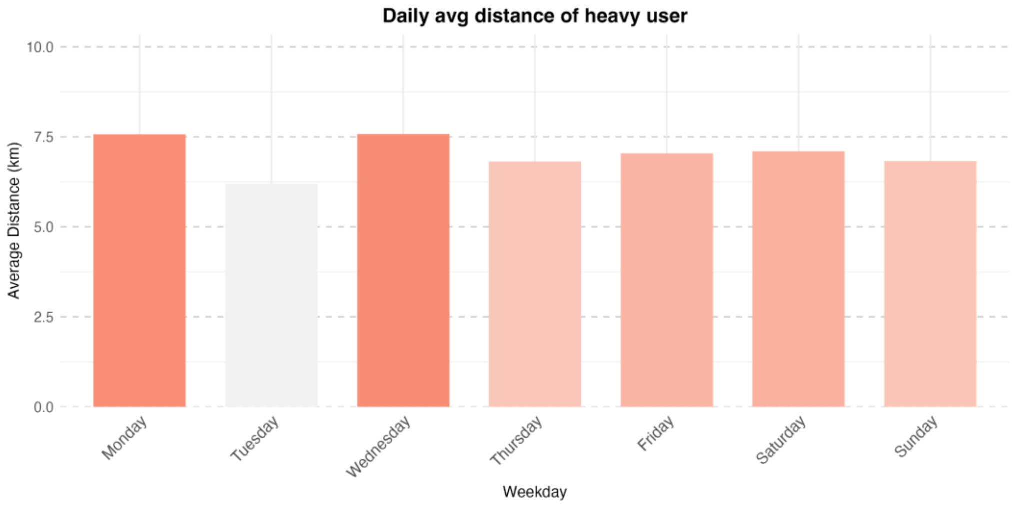
“Heavy users” seemed to be **most active in their physical activity** among other types of users. Having a higher intensity of activity by seeing their **metabolic equivalent peak at 2.5 on Wednesday, around 7:00 pm,** and on Saturday. It is seen that **they start their day earlier than other users, assuming taking a light exercise at 5:00 AM** and taking an hour break after before continuing their activity.

Average data of steps, distance travelled, active minutes, and sedentary time seemed to be the highest among other users. We can speculate that higher activity levels of these users make them get enough sleep at night.

Appeared that they owned the device to track their activity level daily, to get constant information about their physical well-being.

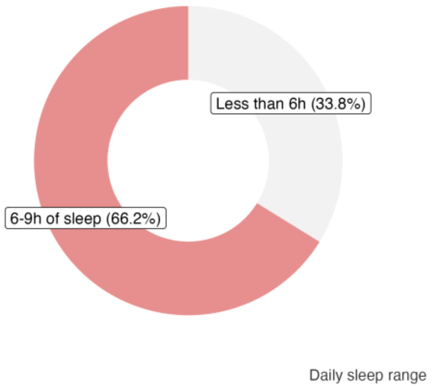


9461 steps on average	6.9 km distance	247 mins active time	937 mins sedentary time



Frequent Users

59-61 out of 62 days



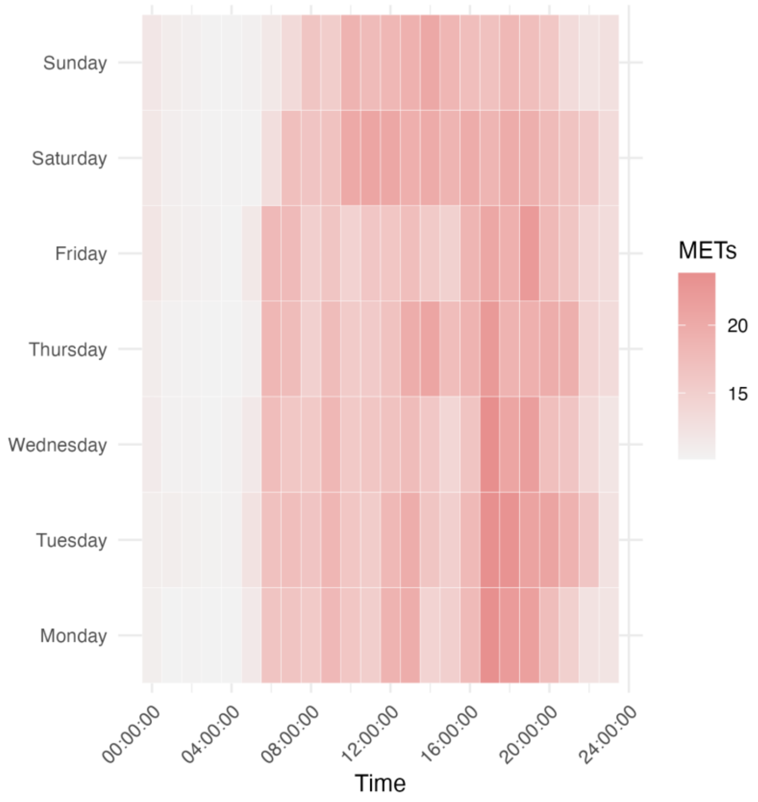
Daily sleep range

This type of user is the second highest of their physical activity throughout the day., But also **peaked at their weekday night around 6:00 PM after work, this indicates they become active after work, assuming socializing, going for groceries, or doing physical exercise at night.**

Average data of steps, distance travelled, active minutes, and sedentary time seemed similar with moderate users. We can speculate they have similar intensity throughout the week as ordinary users.

But the amount of distance travelled varies for these users. **They walk mostly on Tuesday and Saturday.**

Heatmap of METs for Frequent Users



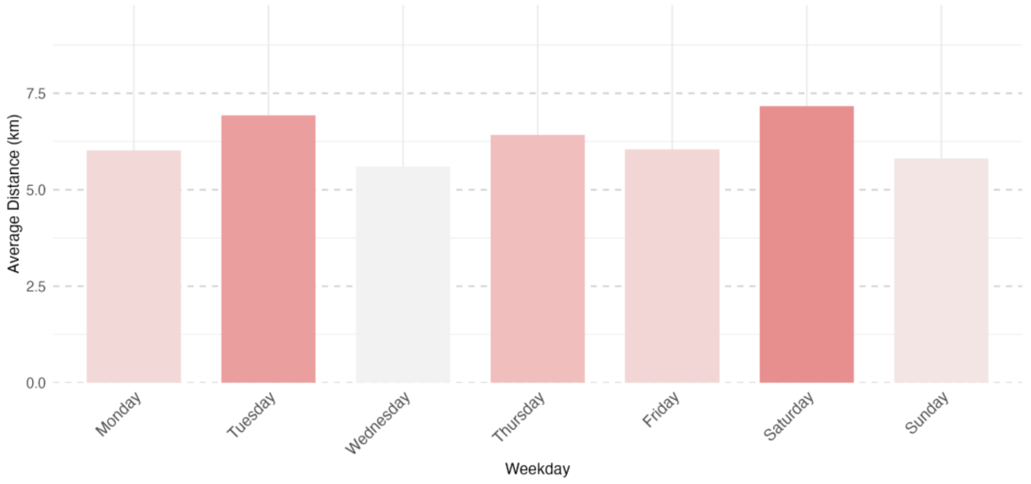
8968
steps
on average

6.2
km
distance

262
mins
active time

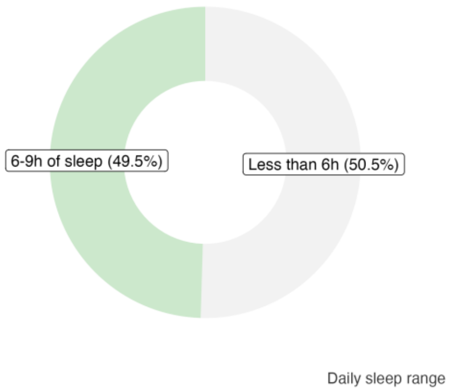
849
mins
sedentary time

Daily avg distance of frequent user



Moderate Users

58-49 out of 62 days

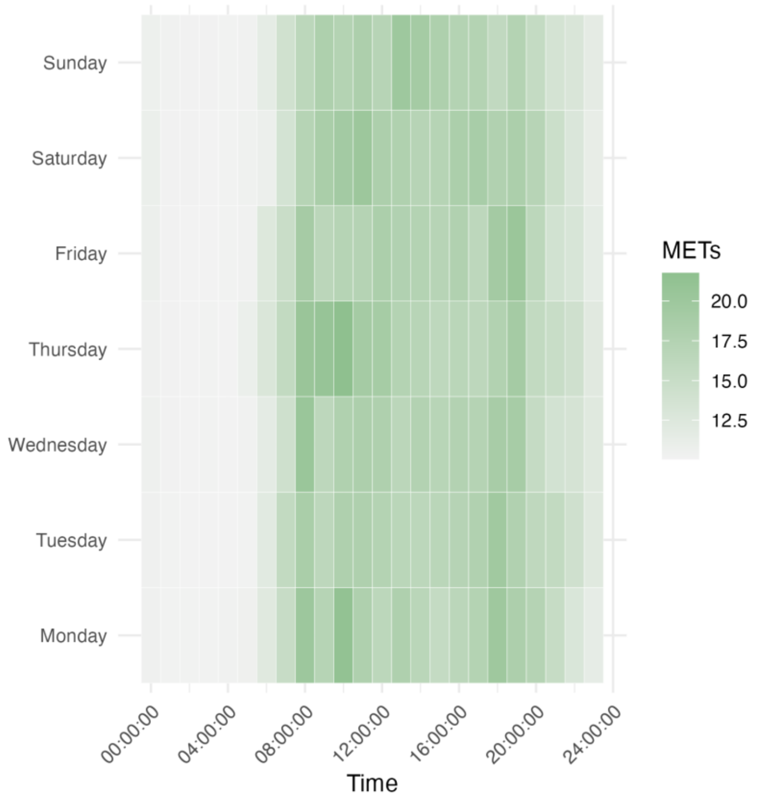


This type of user is the third highest of their physical activity throughout the day., **Seemed to have a well-distributed level of physical activity throughout the day**, but they preferred to do higher physical activity in the morning, around 8:00 to 10:00 AM in the morning. And increase normally in the after work hours.

Average data of steps, distance travelled, active minutes, and sedentary time seemed similar with moderate users. We can speculate they have similar intensity throughout the week as ordinary users.

The distance of their walk remains quite high from Monday to Saturday. **And they have significant rest from walking on Sunday.**

Heatmap of METs for Moderate Users



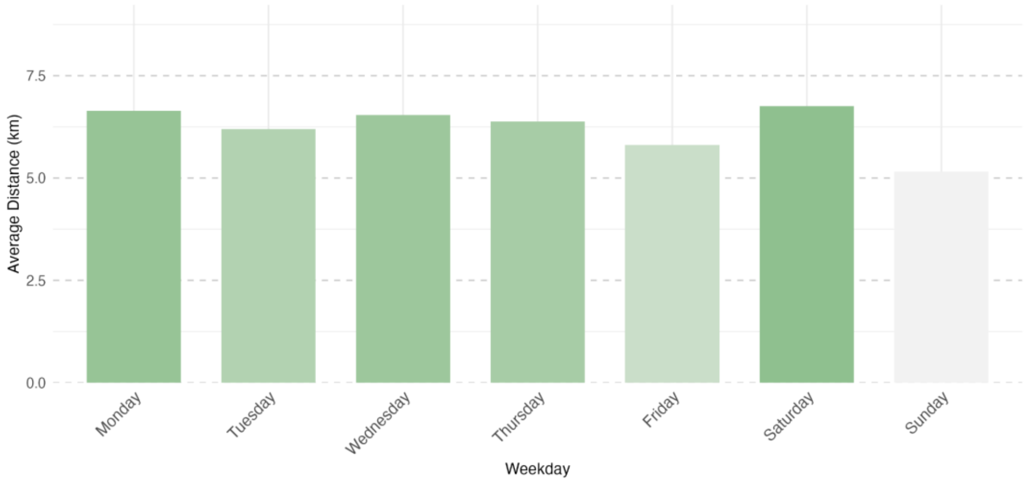
8978
steps
on average

6.2
km
distance

302
mins
active time

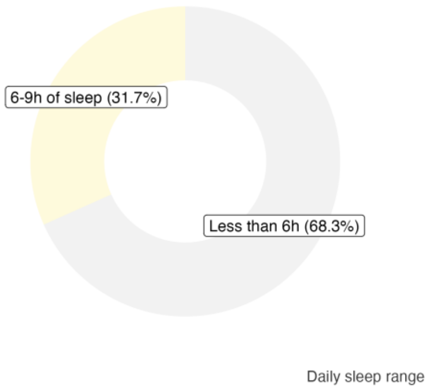
862
mins
sedentary time

Daily avg distance of moderate user



Light Users

58-49 out of 62 days



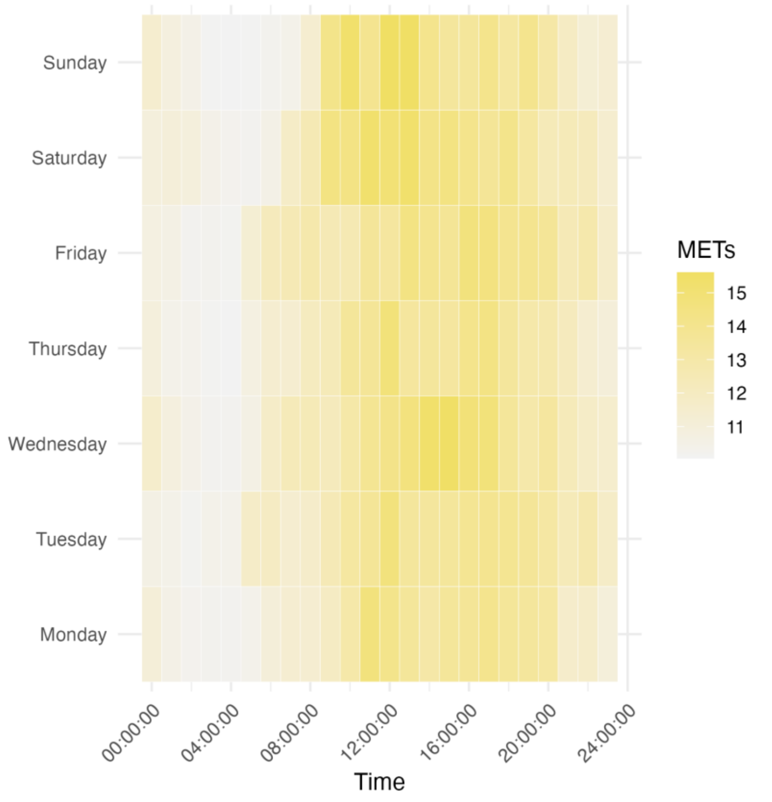
10 out of 32 users who wear the device less than 48 out of 62 days show noticeable patterns of user. They have a way less physical activity level than the other 3 types of users, and **they are likely to spend two-thirds just being sedentary.**

The metabolic equivalent value is only peaked just above 1.5, meaning they spend their time only on light intensity, like reading, watching TV, etc. Unique patterns that **they start to be physically active at 12:00 PM, and on Sunday to Saturday morning.**

They sleep less, doing lighter intensity activity till past midnight, presumably using their phones and watching television.

We can assume this type of user spends their time at home, such as a household mother, a young adolescent, and students.

Heatmap of METs for Light Users



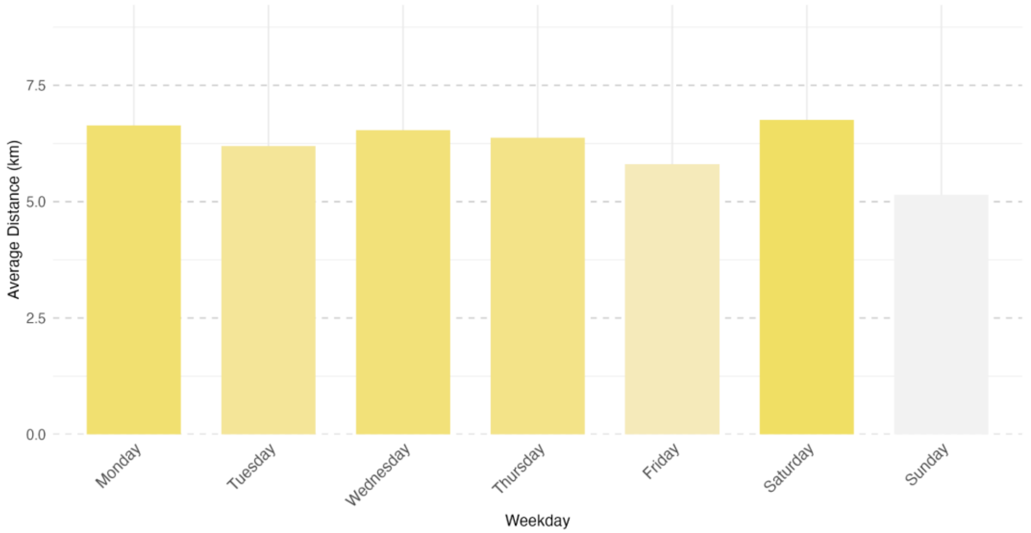
5398
steps
on average

3.9
km
distance

177
mins
active time

1109
mins
sedentary time

Daily avg distance of light user



5. Share

Deliverables: High-level recommendations for how these trends can inform Bellabeat marketing strategy.

5.1 Key Insights and Observation

- Smart devices users observed seemed to have a sedentary lifestyle in urban cities, they spend the most daytime as being sedentary like deskbound lifestyle, increased level of activity mostly after working hours, probably due to socializing, physical exercise or individual matters. Though they seemed to be highly interested in tracking their level of physical wellbeing.

- Almost half of the users appear to opt not to wear the devices during sleeping time, as indicated by the absence of tracking records at night.

- According to 4 types of users, we can sum up to 3 segments of users;

  Category 1: We can classify this as a highly active/enthusiastic group of users who wear the device 100% of the time, yet still maintain high physical activity. Compared to the other group, they tend to be early risers. We suggest that their exercise routine is followed by an hour of rest before they resume their daily activities. This pattern shows that owning the smart devices to support them in tracking their physical activity data, while motivating them to maintain physical wellbeing. Generally, they also have good portions of sleep to balance their high physical activity during the day.

  Category 2: This is a grouped class from frequent and moderate users who wear the device around 88% of the time, so we can call them usual users. They have a considerably high level of physical intensity, furthermore, the highest durations of active time during the day. Mostly active in the morning and late in the evening, however, they tend to rest more on Sunday. These 14 users get relatively enough sleep during the weekday due to their busy activity.

  Category 3: We have an interesting class of smart devices users who are the contrast to those two active classes. Although the trends in the smart devices market cater to physically active people, this class refers to a group of highly stationary individuals. They start to get physically active at midday, and get even more energetic during weekends compared to weekdays. They sleep relatively less to other categories of users by spending their late nighttime doing sedentary activities such as reading books, sitting, lying down, or working at a desk. We are reluctant to conclude that they acquire smart devices primarily to continuously monitor their physical activity, but rather than for casual checking or simply following the latest wearable tech trends. This group portion is about a third of the total users, might come from people who spend most time being sedentary, such as a household mother, a young adolescent, and students.

5.2 Recommended Actions

- Communicates these general insights and particular category of user behaviour towards using the wearable smart devices to get a general idea about Bellabeat users and choosing their marketing channels personalised by their type of user segmentation, how they can benefit from owning particular or multiple Bellabeat devices.

- Due to the limited data gathering of sleeping records in available data, we can take this chance as a marketing campaign as Bellabeat is a leading brand that offers reliable health and wellbeing recommendations as a result of data gathering across multiple comfortable smart devices.

- Communicates clearly to stakeholders that the trends of smart devices are concentrated among people in big cities, while some groups are motivated by maintaining or improving their physical well-being. However, a significant portion of users purchase these devices casually, while maintaining a sedentary lifestyle. Further customer surveys may be needed to uncover their motivations for wearing these devices. This insight could help expand market opportunities by developing new designs and features that better align with their interests.

5.3 Limitations & Future Considerations

These trends might change over time due to current trends of human behaviour mental health matters rather than maintaining their physical health over psychological well-being.