****

**Bellabeat App Case Study**

*Analysis to identify ways to improve Bellabeat’s smart devices and identify trends that can improve marketing strategies for these devices*

**Summary**

Bellabeat, a high-tech manufacturer of health-focused products for women they offer devices that collect data on activity, sleep, stress, menstrual cycle, and mindfulness habits. This has allowed Bellabeat to empower women with knowledge about their health and habits.

Bellabeat is interested in opportunities for growth by studying smart device fitness data in order to inform this endeavor. This study will focus on one of Bellabeat’s products “Bellabeat App”.

The Bellabeat app connects and works hand in hand with other products like:

* **Leaf**: Bellabeat’s classic wellness tracker which can be worn as a bracelet, necklace, or clip to track activity, sleep, and stress.
* **Time Watch**: This wellness watch combines the timeless look of a classic timepiece with smart technology to track user activity, sleep, and stress
* **Spring**: This is a water bottle that tracks daily water intake using smart technology to ensure that you are appropriately hydrated throughout the day.

1. **Ask Phase**

**Business Task:** Analyze consumer habits in relation to smart device usage in order to gain insights that will help guide the marketing strategy for the company

Stakeholders

* Urška Sršen - Bellabeat cofounder and Chief Creative Officer
* Sando Mur - Bellabeat cofounder and key member of Bellabeat executive team
* Bellabeat Marketing Analytics team

1. **Prepare Phase**

**2.1 Dataset**

The dataset used was the “Fitbit Fitness Tracker Data” stored on Kaggle and made available through Mobius.

**2.1.1** **Accessibility and privacy of data**

Verifying the metadata of our dataset we can confirm it is open-source. The owner has dedicated the work to the public domain meaning it can copied, modified, distributed and without asking permission.

**2.1.2 Information about our dataset**

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 the submission of personal tracker data, including minute-level output for physical activity, heart rate, and sleep monitoring. Variation between output represents use of different types of Fitbit trackers and individual tracking behaviors / preferences.

**2.1.3 Data Organization and verification**

Available to us are 18 CSV documents. Each document represents different quantitative data tracked by Fitbit. The data is considered long since each subject will have data in multiple rows. Every user has a unique ID and different rows since data is tracked by day and time.

The tables were sorted and filtered in R . I was able to verify attributes and observations of each table and relations between tables. Counted sample size (users) of each table and verified time length of analysis - 21 days.

**2.1.3 Limitations of Data**

* There was no information on where the data was gathered meaning there might be some sampling bias
* Thirty participants make up a small sample size
* Data gathered over a two-month period which we believe is not enough time to give us optimized insights

For my analysis I focused on three datasets namely:

* daily\_activity\_merged
* sleep\_day\_merged
* hourly\_steps\_merged

1. **Process Phase**

* 1. **Loading Libraries**

The following packages were use this endeavor:

* tidyverse
* skimr
* janitor
* lubridate

````

library(tidyverse)

library(janitor)

library(skimr)

library(lubridate)

``````

**3.2 Imported datasets into R.**

``````

hourly\_steps\_merged <- read\_csv("Desktop/Case Study Wellness/fitabase\_data /hourly\_steps\_merged.csv")

daily\_activity\_merged <- read\_csv("Desktop/Case Study Wellness/fitabase\_data /daily\_activity\_merged.csv",

+ col\_types = cols(ActivityDate = col\_date(format = "%m/%d/%Y")))

hourly\_steps\_merged <- read\_csv("Desktop/Case StudyWellness/fitabase\_data / sleep\_day\_merged.csv")

```````

I renamed my datasets to make them easier to work with

````````

daily\_activity <- daily\_activity\_merged

sleep\_records <- sleep\_day\_merged

hourly\_steps <- hourly\_steps\_merged

,,,,,,,

**3.2.1 Viewing the datasets**

I used the view function to view my datasets and the str function to summarize them

**`````**

view (daily\_activity)

str(daily\_activity)

view (sleep\_records)

str(sleep\_records)

view (hourly\_steps)

str(hourly\_steps)

**3.3 Cleaning Data**

**3.3.1 Verifying number of unique users per selected data frame**

It is important to know the number of unique users we are dealing with in each data frame

`````

n\_unique(daily\_activity$Id)

n\_unique(sleep\_records$Id)

n\_unique(hourly\_steps$Id)

`````

**3.3.2 Checking for duplicates**

I checked for duplicates in my data sets

`````

sum(duplicated(daily\_activity))

sum(duplicated(sleep\_records))

sum(duplicated(hourly\_steps))

`````

No duplicates were found in these datasets

**3.3.3 Cleaning columns**

Since we will be combining columns, we need to make sure the column name formats are consistent across datasets

,,,,

clean\_names(daily\_activity)

daily\_activity<- rename\_with(daily\_activity, tolower)

clean\_names(sleep\_records)

sleep\_records <- rename\_with(sleep\_records, tolower)

clean\_names(hourly\_steps)

hourly\_steps <- rename\_with(hourly\_steps, tolower)

````

**3.3.4 Date formats**

We will also make sure our date formats are the same across datasets we are interested in the date only so we will use the as\_date function

````

daily\_activity <- daily\_activity %>%

rename(date = activitydate) %>%

mutate(date = as\_date(date, format = "%m/%d/%Y"))

sleep\_records <- sleep\_records %>%

rename(date = sleepday) %>%

mutate(date = as\_date(date,format ="%m/%d/%Y %I:%M:%S %p" , tz=Sys.timezone()))

`````

Because we are considering steps taken hourly we will include time for the “hourly steps” dataset.

``````

hourly\_steps<- hourly\_steps %>%

rename(date\_time = activityhour) %>%

mutate(date\_time = as.POSIXct(date\_time,format ="%m/%d/%Y %I:%M:%S %p" , tz=Sys.timezone()))

`````

Separating date and time values for hourly steps

````````

hourly\_steps <- hourly\_steps %>%

separate(date\_time, into = c("date", "time"), sep= " ") %>%

mutate(date = ymd(date))

````````

**3.4 Merging Datasets**

We will merge the daily\_activity and sleep\_records datasets to see any correlation between variables by using id and date as their primary keys.

```

daily\_activity\_sleep <- merge(daily\_activity, sleep\_records , by=c ("id", "date"))

view(daily\_activity\_sleep)

````

We will analyze and share the results using Microsoft excel. We will start by exporting the file to excel.

``````

write\_xlsx(daily\_activity\_sleep,"Desktop\\daily\_activity\_sleep.xlsx")

write\_xlsx(hourly\_steps,"Desktop\\hourly\_steps.xlsx")

``````

1. **Analyse and Share Phase**

During this phase we will scrutinize the cleaned data for trends that can benefit Bellabeat’s marketing strategy. We will work with the data we have.

**Types of users based on activity level**

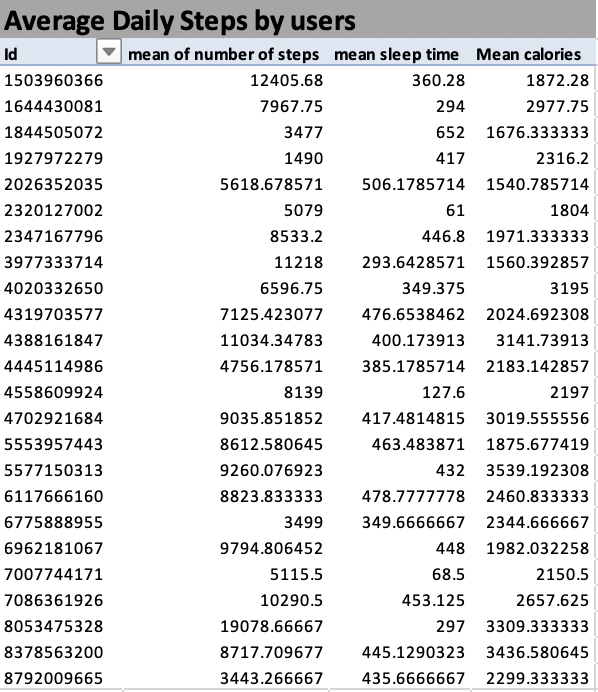
We can classify users based on the daily number of steps. We will group users based on the following:

* Very active - More than 10000 steps a day.
* Fairly active - Between 7500 and 9999 steps a day.
* Lightly active - Between 5000 and 7499 steps a day.
* Sedentary - Less than 5000 steps a day.

Our classification was guided using the following article https://www.10000steps.org.au/articles/counting-steps/

*We will first calculate the daily average steps per user*

*This was done by creating a pivot table with “id” in row and “Total Steps, Sleep Time, Calories” in the values section. We than changed the function to calculate the average of these values.*



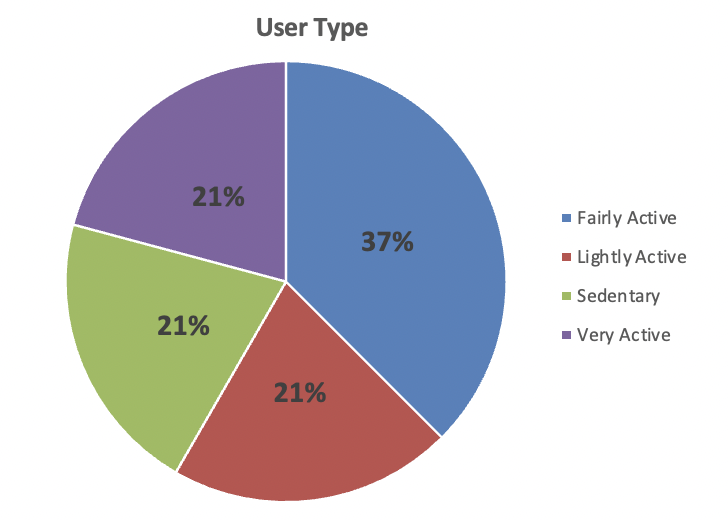
*Next, we will group these users under the above classification and determine the percentage of users under each type*

*We attached a user type to each user in a separate column using the formular below*

“=IF(AND(H5<5000),"Sedentary",IF(AND(H5>=5000,H5<=7499),"Lightly Active",IF(AND(H5>=7500,H5<=9999),"Fairly Active",IF(AND(H5>=10000),"Very Active"))))”

*We found the percentage out of a whole of each user type by dividing each user type by the total number of users and then multiplied our results by 100 to get a percentage value*

*We then created a pie chart with our result*



From this chart we can see that users are fairly distributed considering their number of steps. **We can determine from this that all kinds of users wear smart devices.**

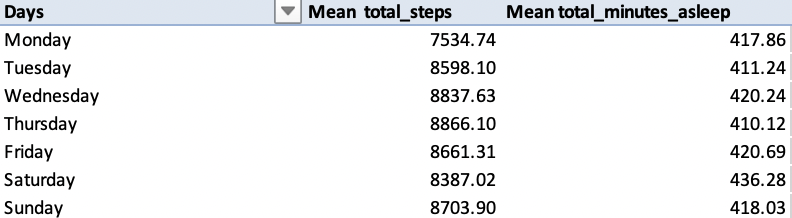
**Investigating activity throughout the week**

*To achieve this, we decided to create a new column and then represent each day of the month with their corresponding day of the week. We did that by using the following formular below*

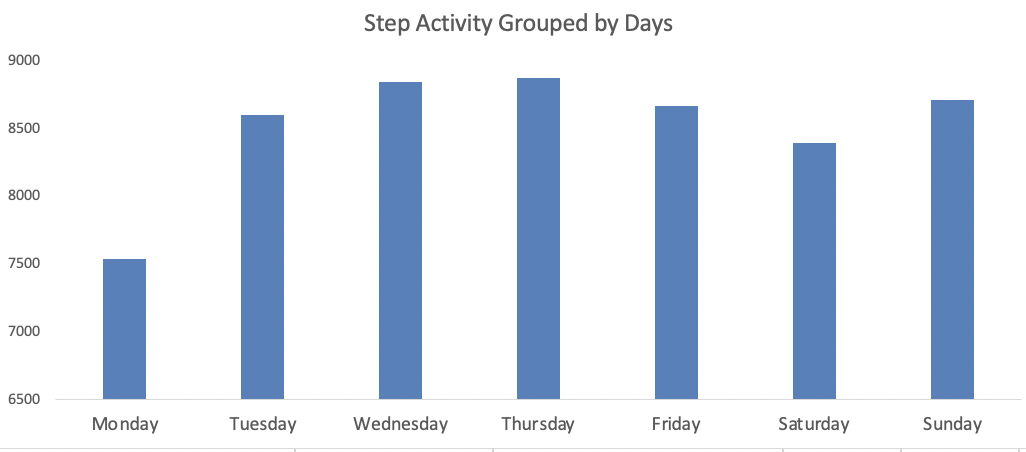
‘’ =TEXT([@date],"dddd") ”

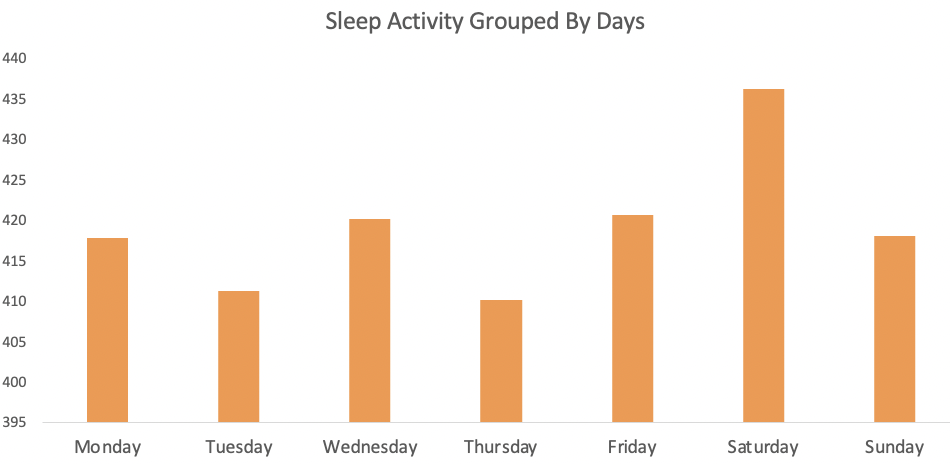
*We then created a pivot table with “Days” in the row column as seen below*

We will first take a look at the number of steps by users throughout the week.



We can further visualize this information using a bar chart to make it easier to identify the average total steps and average minutes asleep throughout the week.



****

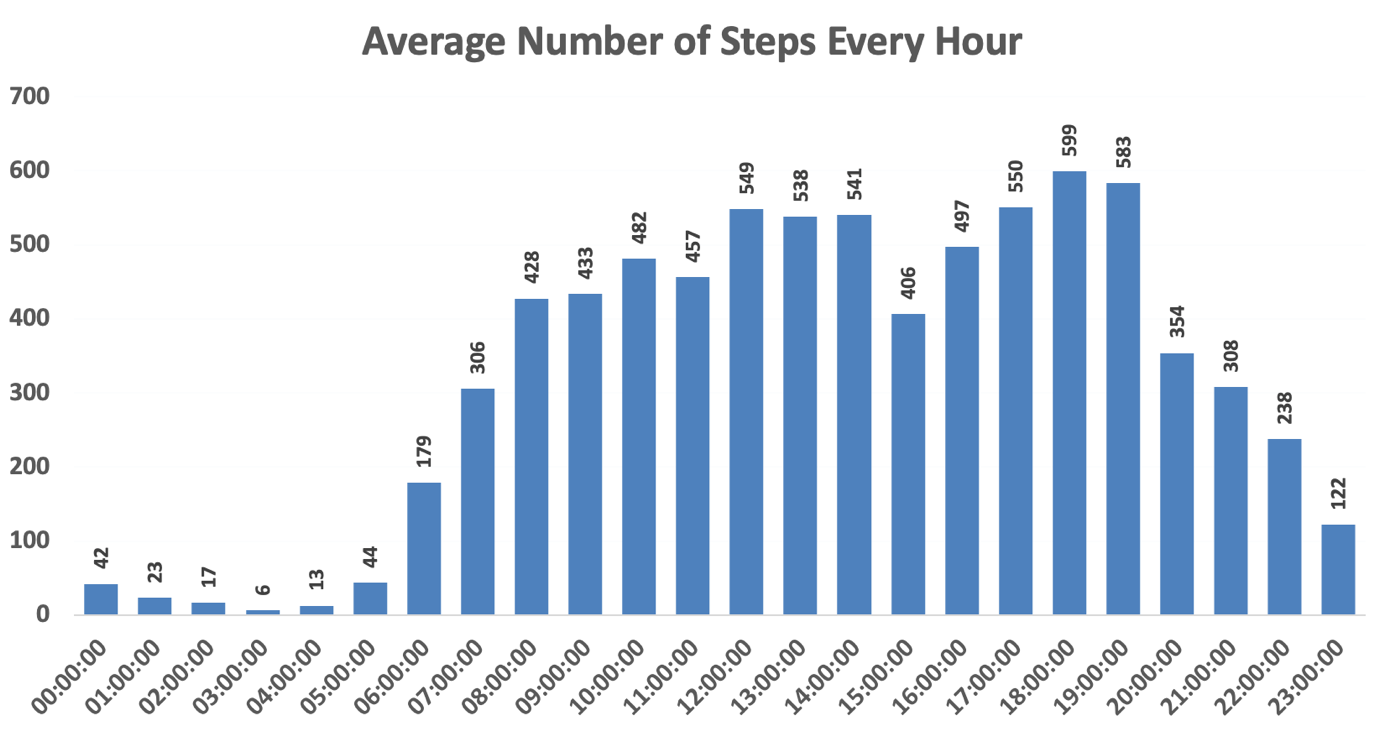
From the tables and chart above we can determine that all **users walk the recommended minimum number of steps (7500) daily.**

The healthy minimum amount of sleep is 8hours which is 480minuites. The table and chart show that **none of the users meet the minimum daily sleep requirement.**

**Hourly Steps**

We want to find when users are more active during the day so we find the average number of steps for every hour.

*This was done by creating a pivot table with “Hour” in the row section*

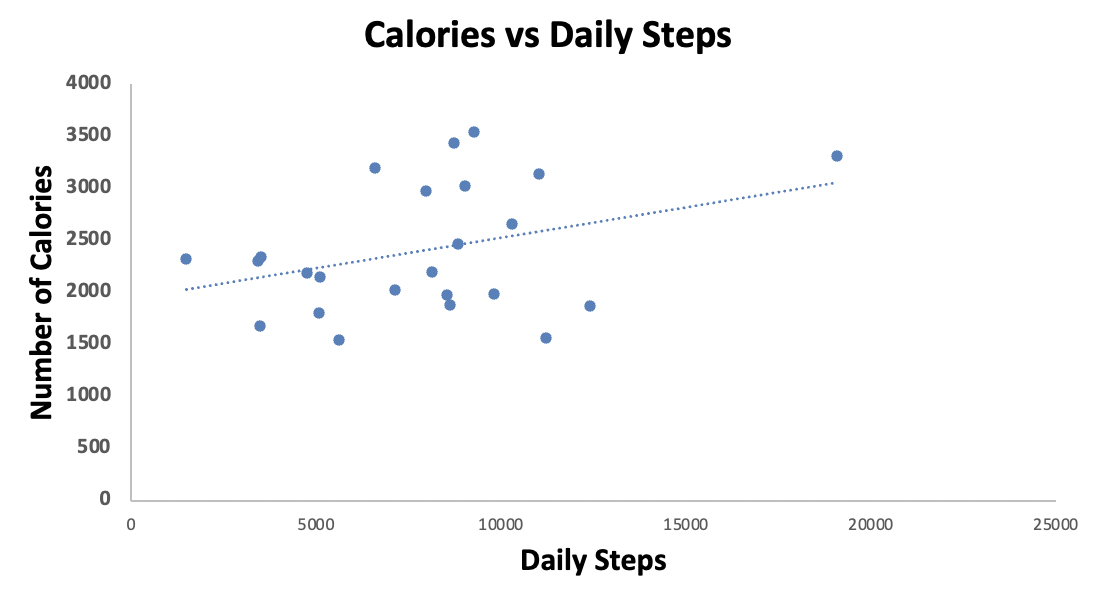


Most steps are from **12pm to 2pm** and from **5pm to 7pm** these time are usually lunch and supper times

**Correlation between Values**

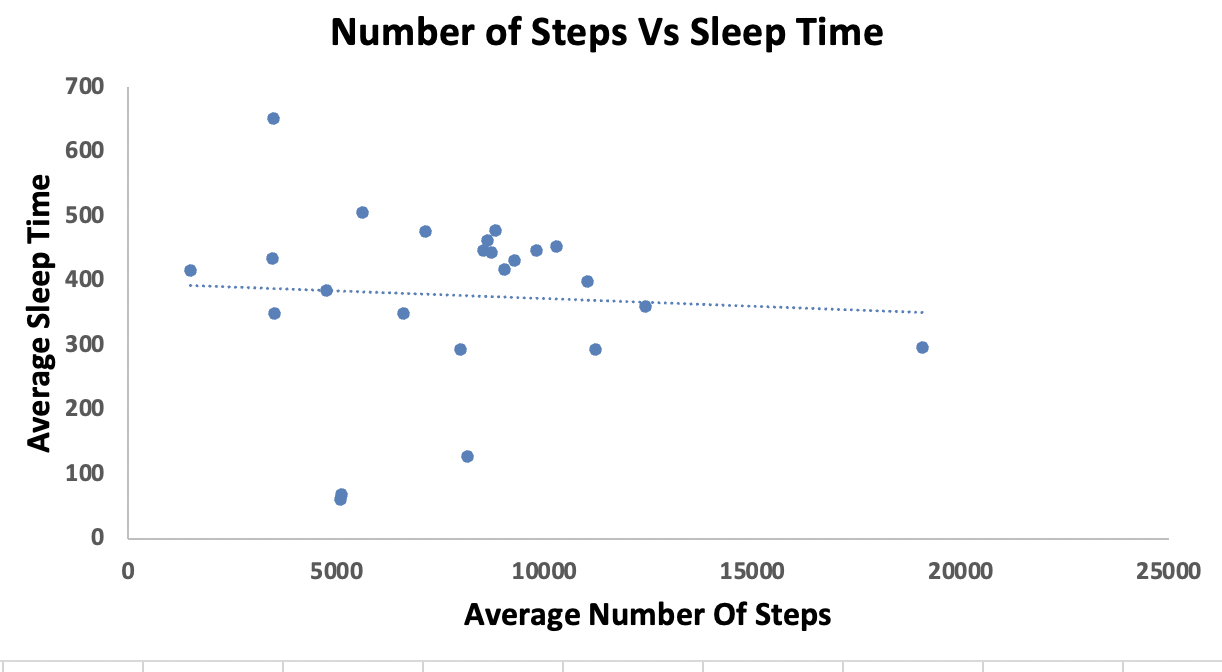
We go on to test the correlation between Steps and Calories

*This was done by creating a scatter plot with the variables*



We notice **there is a correlation between calories and daily steps**

We then go ahead to test the correlation between number of steps and minutes spent asleep



From the chart above we can see that

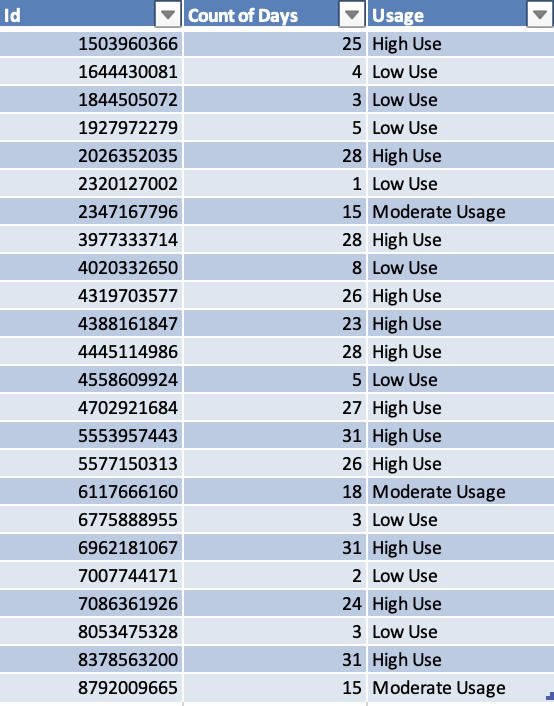
**Usage of Smart Devices**

We calculated usage of the smart device by users and grouped then into three categories

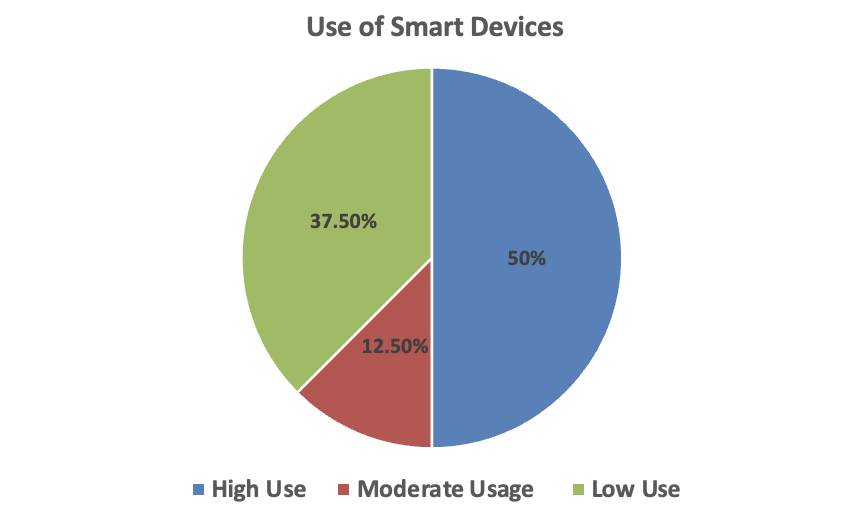
* High Usage(21-31 days)
* Moderate use(11-20days)
* Low use(0-10days)

*A pivot table was created with “id” as the row and a count of the various days the device was used in the value section. The formular below was used to classify the frequency of usage*

“=IF(AND(F7>=21, F7<=31 )," High Usage ",IF(AND(F7>=11, F7<=20)," Moderate use ",IF(AND(F7>=0, F7<=10)," Low use “ )))



We then find the percentage of each category and put them in a chart



From this chart we can see that:

* 50% use their device regularly (21-31 days)
* 12.5% use their device moderately(11-20days)
* 37.5% rarely use their device (0-10days)

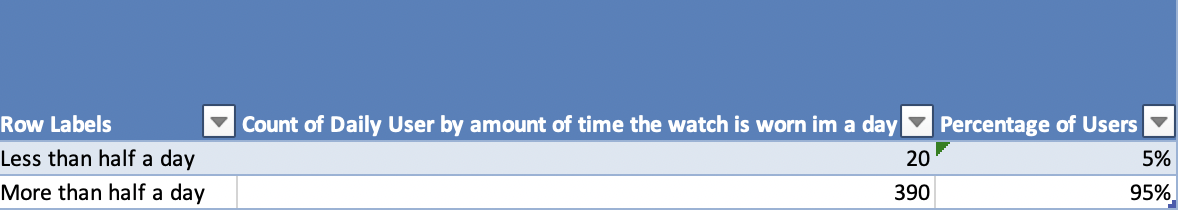
**Amount of time the device is worn throughout the day**

We further investigated how often users wore the device throughout the day. Data showed that none of the users wore the device the whole day. Hence users were grouped on whether they wore the watches for **more than half a day** or **less than half a day**

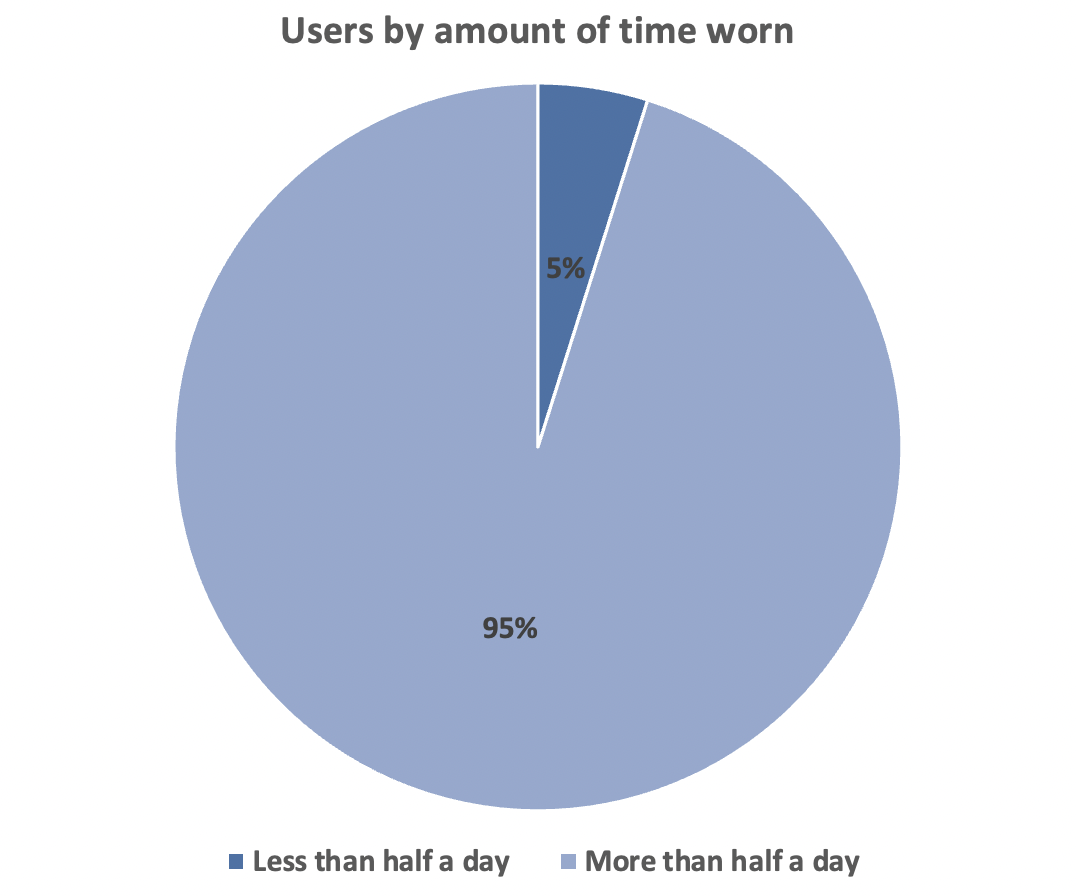
*This was done by first calculating the total active minutes for the device and dividing it by the total amount of minutes in day and then finding the percentage of a whole day it was used. The formular below was used to group the devices based on the amount of time it was worn.*

*“=IF(AND(U5=100%),"All Day",IF(AND(U5<100%,U5>=50%),"More than half a day",IF(AND(U5<50%,U5>0),"Less than half a day")))”*

*This gave us the following table*

**

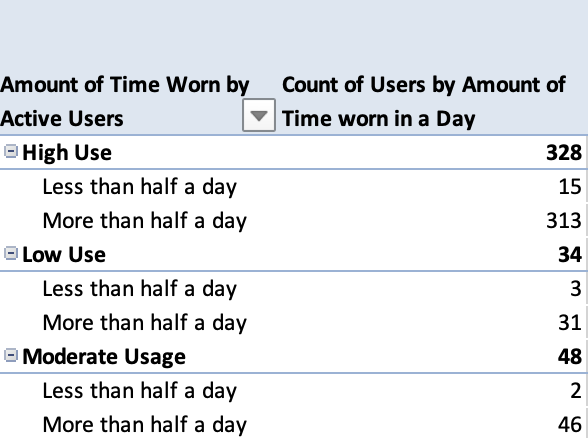
*A chart was then calculated from the table*

****

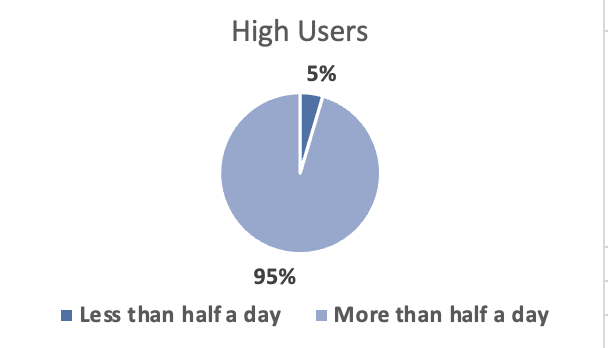
**From this chart we can see that most users wore the device for more than half a day.**

**We then decided to classify the time worn by the usage of the device**

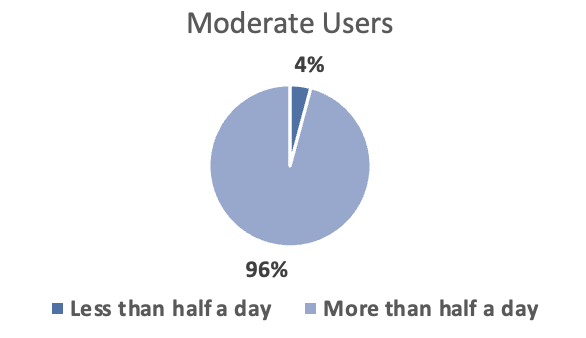
*The a pivot table was created to show how frequently the various classification of users wear their device throughout the day.*

**

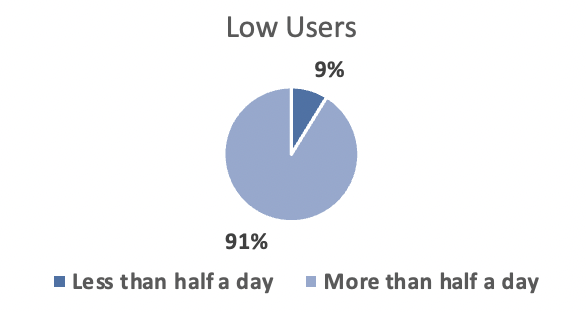
*A chart was then created from the above information*

****

This chart shows that **95%** of people who used the device regularly wore them for more than half a day and **5%** wore them for less than half a day



**96%** of Moderate users wore the device for more than a day and **4%** wore the device for less than half a day



**91%** of users who rarely use the device wear it for more than half a day and **9%** wear it for less than half a day

**5.ACT PHASE**

**Recommendations**

* From the chart showing the usage of the device the highest percentage was from fairly active users 38% as compared to 21% from active users. Marketing campaigns can be organized to show the benefits of tracking your activities in order to increase the appeal to the active users.
* Articles or blogs can be added to the Bellabeat app to educate users on benefits of living a healthy life and how being able to keep track of such data can help. This endeavor is aimed at reducing the percentage of low users of smart devices which is currently at 37.5%
* Also, the Bellabeat app can be designed to set targets in terms of steps walked, water drank, etc. This will motivate users to meet the targets hence increasing activity and use of the device
* From our analysis the needed sleep target of 480minutes (8 hours) wasn’t reached on any of the days. Bellabeat’s app can have an alarm that serves as a sleep reminder as well as wake users up. Users can set a particular time they have to wake up and the app will then calculate a recommended interval with alarm reminders for time to sleep and wake up.
* *From our analysis we found out that none of the users wore the device the all day even though a high percentage (95%) wore it for more than half a day.* Suggestions to remedy this include:
* *A strong battery that can last the whole day*
* *Fashionable devices that can be worn to events so that it can be on the user at all times*
* *Water and dust resistant products for durability*