# **Google Data Analytics Capstone Project**

**Created: October 23, 2023 Start Time: 23/10/2023, 09:20**

**Author: Aniekan Ekanem End time: 30/10/2023, 16:23**

**Client: Bellabeat Modified: 31/10/2023**

**2nd Modification: 03/11/2023**

**3rd Modification: 17/11/2023**

As part of my Google Data Analytics Capstone Project, Track 1, Case Study 2, I will be conducting an analysis of datasets for Bellabeat. This project entails a comprehensive examination of the data, including data cleaning, analysis, and visualization using a range of tools. I will present a thorough data summary and provide final recommendations for the business based on my findings.

**Quick Links:**

[Github R Code for Analysis and Visualization](https://github.com/aniekanekanem/Bellabeat-Product-Analysis-Case-Study/blob/main/bellabeat-product-analysis-case-study-i%20(2).ipynb)

Below is a table of contents in case you want to go to a specific section.

### **Analysis and Tableau Visualization**

### **Table of Contents:**

1. [Background](#kix.1tmqd1azq79m)
2. [Analysis Process](#kix.htzb39xtxur0)

2.1 [Ask](#kix.7sudt8wk7f1z)

2.2 [Prepare](#kix.y1g0qtg2i1hi)

2.3 [Process](#kix.4s00qkvvymt6)

2.4 [Analyze](#kix.fo7ez6rgtq8g)

1. [Summary of Visualization](#kix.euxi0mtjwoqf)

3.1 [Summary](#kix.kjl6b6axc18h)

1. [Business Suggestions](#kix.i1288zwf6kbs)
2. [Observations](#kix.6tsda98f6hm2)
3. [Links](#kix.qfz80llccmk)
4. [What I learnt](#kix.wyu1yj4p4hso)
5. [Conclusion](#kix.pt0xguxjjkth)
6. [Resources](#kix.z52l66ffcca4)

## 

## **1.** **BACKGROUND**

Bellabeat is a high-tech wellness company based in San Francisco, United States of America that manufactures health-focused smart products. Its founders are **Urška Sršen** and **Sando Mur**. **Urška Sršen** as a co-founder has a background as an artist and used her artistic skills to develop beautifully designed technology that informs and inspires women around the world. This shows that the majority of their customers are women, other customers could include husbands, siblings, sons, uncles, etc that may want to buy items for their loved ones as a gift, etc.

What this technology does is to collect data on activity, sleep, stress, and reproductive health and this has allowed Bellabeat to empower women with knowledge about their own health and habits. Since it was founded in 2013, Bellabeat has grown rapidly and quickly positioned itself as a tech-driven wellness company for women.

By the year 2016, different offices have been opened around the world and multiple products have been launched by the company such as Bellabeat app, Leaf, Time, Spring and the Bellabeat membership, a subscription-based membership program for users.

Bellabeat products became available through a growing number of online retailers in addition to their own e-commerce channel on their website (https://bellabeat.com/). The company

has invested in traditional advertising media, such as radio, out-of-home billboards, print, and television, but focuses on digital marketing extensively. Bellabeat invests year-round in Google Search, maintaining active Facebook and Instagram pages, and consistently engages consumers on Twitter. Additionally, Bellabeat runs video ads on Youtube and display ads on the Google Display Network to support campaigns around key marketing dates.

**My Role**: In this scenario, I am a Junior Data Analyst at Bellabeat and my team has been tasked with the overall goal of investigating the company's product usage and analyzing smart device usage in order to gain insight into how people are already using their smart devices, and then using the outcome of this investigation to make recommendations for how these trends can be used to build upon the company's marketing strategy.

**Overall Goal**: To build on Bellabeat’s marketing strategy by finding out how Bellabeat’s products are being used by users and to analyze other smart device usage for gaining insight into how people are already using their smart devices.

**Business Question**: “What are some trends in smart device usage?”

I will now outline the sequential procedure I followed to accomplish this project. If you prefer to jump directly to the business recommendations, please proceed to the section titled "[**business suggestions**](#kix.i1288zwf6kbs)".

**Overview**: I first answered a key question assigned to me by the Co-founder (**Urška Sršen**) plus other questions that could relate towards solving the business task by inspecting the dataset downloadable from [FitBit Fitness Tracker Data](https://www.kaggle.com/arashnic/fitbit). Then I prepared the data after the download.

Next, was to prepare the datasets while taking note of the following:

* Location of the data
* How organize the data is
* Checking for any issues with bias or credibility of the data
* If the data meets the credibility criteria of being reliable, original, comprehensive, complete, and if it is cited.
* Addressing licensing, privacy, security, and accessibility of the data
* Verification of the integrity of the data
* How the data help in answering questions
* Checking if there are any problems with the data

Next, I processed the data using R via Kaggle, then I analyze the data separately using necessary health metrics. With these analyses, I was able to come out with implementable marketing strategies ready to be shared before the necessary action could be taken.

## **2.** **ANALYSIS PROCESS**:

**2.1** **ASK**

During this stage, the business and its operatives should be studied and understood, then relevant questions can be asked.

**Getting to know more about the business**:

1. **Understanding the business task**: Bellabeat as a company offers manufactured wellness products that benefits women. Products offered by the company are smart products which includes bellabeat app, leaf, time, spring and a membership subscription service which offers a more personalized health services.

What Bellabeat is trying to do is to examine how users interact with smart products from other vendors or companies and then use the insight to be applied to its own products. With a request to select one of the products, I have decided to select the Leaf product for further analysis.

What I need to do first of all, is to access Bellabeat top competitors’ datasets and compare it with its own datasets. During this process, I will need to examine why the trend is the way it is, outlining out factors responsible, etc. The result of this will lead to the business suggestions needed to be adopted to keep the company moving.

Below are ten list of bellabeat competitors:

* Clue.
* Natural Cycles.
* Ava.
* Ovia Health.
* Kindara.
* Thalman Health.
* Ovatemp.
* Alt12 Apps.
* Glow
* Groove

However, I don’t have access rights to the datasets of any of the above listed companies. Hence, for the sake of this project, I will be making use of the dataset provided [here](https://www.kaggle.com/code/aniekanekanem/bellabeat-product-analysis-case-study-i?scriptVersionId=151520918&cellId=10).

1. **Key Stakeholders:**

* Lily Moreno (Director of Marketing/Manager)
* Cyclistic marketing analytics team
* Cyclistic executive team

1. **Key Questions:**

i. **"How do our customers make use of our smart devices?"**

Analyzing the historical data can provide valuable insights into the activity patterns of fitness enthusiasts/subscribers. By comparing factors such as heart rate, BMI, weight, distance travelled, steps taken, user type categories, etc, one can identify how subscriptions can be influenced thereby paving way for a new marketing strategy.

**2.2** **Prepare**

In this stage, I carried out inspection of the datasets using the **R programming language**. With the inspection going through, I was able to attend to the following questions:

* **Where is the Location of the data?**: The data selected can be gotten from [Fitbit Fitness Tracker Data](https://www.kaggle.com/datasets/arashnic/fitbit/).
* **How organized is the data?**: The data in study are in cohorts/segments which at one point may be merged. Some of these data to be anchored by the Ids for instance are incomplete meaning that not all Ids were accounted for by virtue of some health metrics. For instance not all Ids have their BMIs accounted for which will result in “NA” in the merging process and if eventually dropped will result in fewer data for analysis. However, the data all share variable names which are enough to address the business challenge.
* **Did you notice any issue with bias or credibility of the data?**: Bias may set in in the areas of analyzing data were rows with missing observations were dropped. However, using logics and real life happenings in comparison with the outcome after analyses, the biases may pose no significance.
* **Does the data meet the credibility criteria of being reliable, original, comprehensive, complete, and cited?**: Apart from being incomplete, all other credibility criteria are met. The summary based on my humble observation is stated below:
* **R**: Data is reliable since it is coming from a reliable source which is [Fitbit Fitness Tracker Data](https://www.kaggle.com/datasets/arashnic/fitbit/).
* **O**: Data is original as it tends to be gotten from a first party being [Mobius](https://www.kaggle.com/arashnic).
* **C**: The data is comprehensive enough to answer the business question.
* **C**: The data is expected to be complete but not complete based on my observation.
* **C**: The data is cited making it more credible.

In conclusion, the question raised is, “*Does the data exhibit the complete*

*characteristics of credible data*? May be Yes, May be No. To clarify

completeness in credibility, this is where communicating with the stakeholders is very

Important to avoid making wrong decisions. Assumptions and taking a step further

without any consultation is not advisable. Hence, in the real sense, before I go

further, I will need to confirm the credibility of the datasets.

However, for the sake of this project, I will take it that the data meets these criteria.

* **How have you been able to address the aspect of licensing, privacy, security, and accessibility of the data?**: I have taken note of the various terms associated with the data and have observed and will continue to observe the agreements as long as I am using the data for practice.
* **How did you verify the data’s integrity?**: I have been able to verify the integrity of the data since it comes from a reliable source.
* **How does it help you answer your question?**: The data helps in answering the business question as it contains the necessary data comprehensive enough to answer the question.
* **Are there any problems with the data?**: No, except for the fact that the data is incomplete based on my observation.

**2.3** **Processing**

Kindly click [here](https://github.com/aniekanekanem/Bellabeat-Product-Analysis-Case-Study/blob/main/bellabeat-product-analysis-case-study-i%20(2).ipynb) to my Github to view my code for the step by step data processing for this capstone project. The following processing activities were carried out on the tables. See details about the processing [here](https://github.com/aniekanekanem/Bellabeat-Product-Analysis-Case-Study/blob/main/bellabeat-product-analysis-case-study-i%20(2).ipynb):

While using R for this processing, find below, links displaying packages and datasets loaded for the project:

Link1: [**Importing the libraries**](https://www.kaggle.com/code/aniekanekanem/bellabeat-product-analysis-case-study-i?scriptVersionId=151520918&cellId=8)

Link2: [**Loading the dataset**](https://www.kaggle.com/code/aniekanekanem/bellabeat-product-analysis-case-study-i?scriptVersionId=151520918&cellId=10)

**Data Inspection and cleaning**

1. **Processing using R**

**Data inspection**

The data were checked to ensure that they were clean using R.

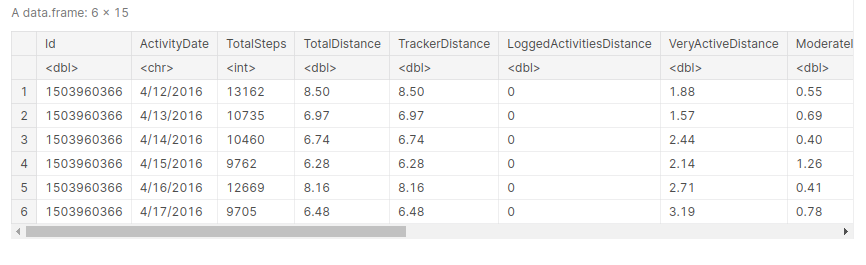
The following datasets were viewed for inspection as well as checked for issues and cleaned accordingly:

* ***DAILYACTIVITY\_MERGED DATASET***
* ***HOURLYSTEPS\_MERGED DATASET***
* ***HEARTRATE\_SECONDS\_MERGED DATASET***
* ***SLEEPDAY\_MERGED DATASET***
* ***WEIGHTLOGINFO\_MERGED DATASET***

* 1. Each dataset was checked and inspected by general view, structure, datatype uniqueness for each column, and finally summary to ensure the dataset is in the expected format. For instance for the “dailyactivity\_merged” dataset, the following were performed and the respective result displayed as shown below:

* **Viewing the dataset**

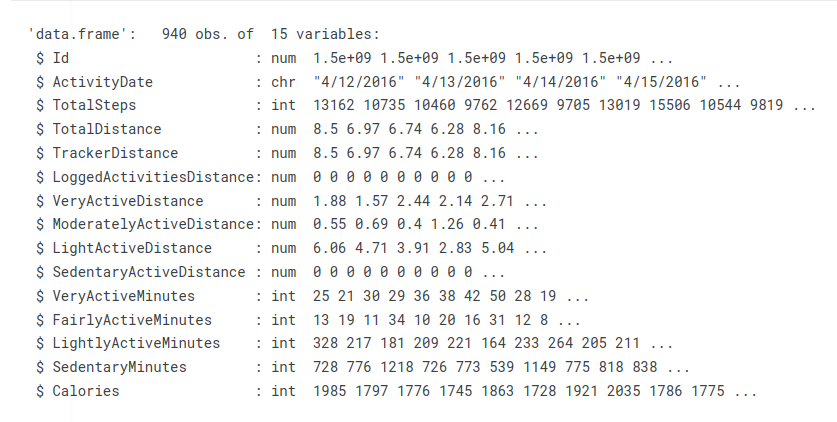
head(dailyactivity\_merged)

Result:

* **Checking the structure of the dataset**

str(dailyactivity\_merged)

Result:



The above result shows that the Id and ActivityDate have datatype of number and character respectively which may need to be changed to character and date format respectively as shown below:

“

*# Converting the Id and ActivityDate datatype to 'string' and 'date format' data type respectively*

dailyActivity\_merged$Id <- as.character(dailyActivity\_merged$Id)

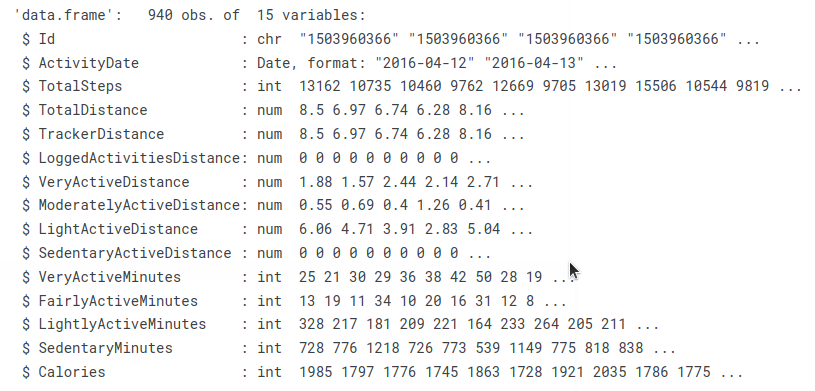
dailyActivity\_merged <- dailyActivity\_merged %>%

mutate(ActivityDate = as.Date(ActivityDate, format = "%m/%d/%Y")) *# The date format as observed in the original dataset is in the form 'm/d/Y'. This is the format that will be used for the conversion*

*# Checking to view the respective data types*

*str(dailyActivity\_merged)*

“

Final result:

As seen above the Id and ActivityDate datatype are now in their needed format.

* **Checking for unique datatype**

*# Checking for unique datatype*

*# Function to count unique data types in a column and return column name*

count\_unique\_data\_types <- function(col) {

data\_type\_count <- length(unique(class(col)))

return(data\_type\_count)

}

*# Count the number of unique data types in each column and store the result in a named vector*

result <- sapply(dailyActivity\_merged, count\_unique\_data\_types)

*# Print the result with column names*

cat("Column Name: Count of Unique Data Types\n")

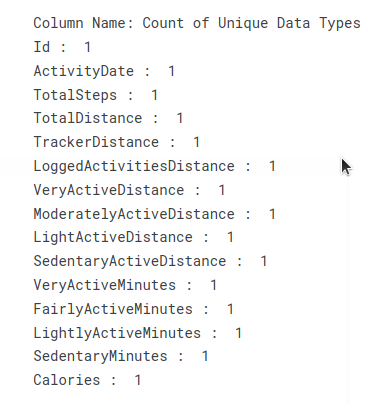
for (i in 1:length(result)) {

col\_name <- names(result)[i]

col\_count <- result[i]

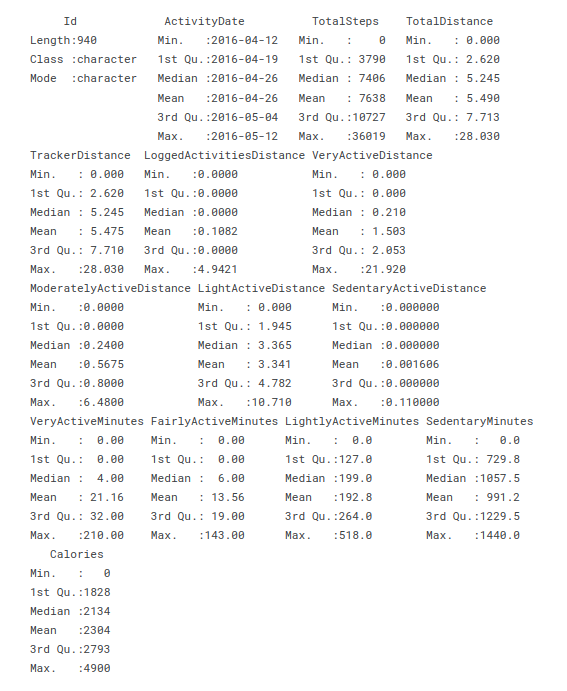
cat(col\_name, ": ", col\_count, "\n")

}

Result:

* *Checking the summary of the dataset*

summary(dailyActivity\_merged)

Result:

* 1. The data was checked to ensure there are no duplicate observations. The result is shown below:

The above initial data checks were also performed for the rest of the dataset cohorts.

**Data cleaning:**

The next thing was to start the data cleaning process.

Each of the dataset was checked for missing data. For instance, the **“dailyActivity\_merged”** was checked and confirmed for missing data as shown below:

*# Checking for missing data*

*# Counting the number of rows with missing data*

num\_rows\_with\_missing\_data <- sum(rowSums(is.na(hourlySteps\_merged)) > 0)

num\_rows\_with\_missing\_data

*# confirming if truly there are no missing data in each row*

missing\_data\_in\_hourlySteps\_merged <- hourlySteps\_merged %>%

filter\_all(any\_vars(is.na(.))) %>%

complete(Id, nesting(ActivityDate, StepTotal))

missing\_data\_in\_hourlySteps\_merged

All the above mentioned dataset reported no missing data except for the **"weightLogInfo\_merged" dataset** which reported **65** missing data. This missing data came from the “Fat” column which may not be in use during the analysis process. A more suitable column will be used as an alternative for any further analysis.

**2.4.** **Analyze:** Analysis was carried out using Kaggle Notebook because of their ease of usage and aesthetics they provide. Another tool that can be used is Python and Tableau for the data processing and visualization respectively.

1. **Identifying metrics enough for answering the business questions**

The following are metrics that will be used to make analysis on the datasets in study:

**i. usertype distribution**

* Concentration of Usertype

**ii. Pattern observation by movement**

* TotalSteps v distanceCovered

**iii.. Pattern observation by date**

* Weekdays v distanceCovered
* Weekdays v TotalSteps

**iv. Pattern observation by aerobic activities**

* TotalSteps v Calories
* DistanceCovered v Calories
* HeartRate v Calories
* Calories v Sleep
* Sleep v HeartRate
* Weight vs distance covered
* Weight vs steps

**v. Pattern observation between user type and activities**

* TotalSteps v usertype\_by\_distance
* TotalSteps v usertype\_by\_minutes (two in one plot)
* DistanceCovered v usertype\_by\_distance
* DistanceCovered v usertype\_by\_minutes
* Calories v usertype\_by\_distance
* Calories v usertype\_by\_minutes
* Sleep efficiency v usertype\_by\_distance
* Sleep efficiency v usertype\_by\_minutes

**vi.. BMI logs pattern**

* BMI v Weight
* BMI v SquaredMeanHeight
* BMI v WeightHealth
* BMI v usertype\_by\_distance
* BMI v usertype\_by\_minute
* BMI v TotalDistance
* BMI v TotalSteps
* BMI classification distribution

**vii. Sleep activities**

* Sleep quality distribution
* Sleep efficiency vs distance covered
* sleep efficiency vs steps

**viii. Subscriber activities**

* Number of subscribers
* Reporting activities

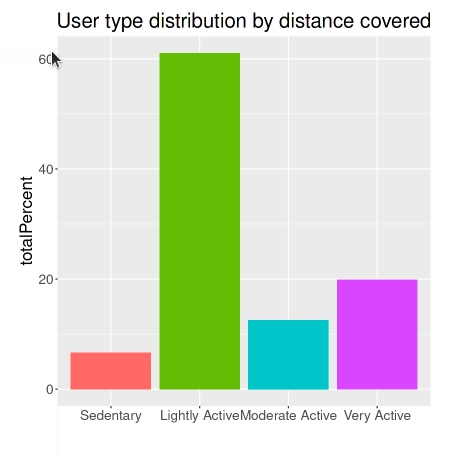
**Analysis using R in Kaggl****e**

* 1. **Usertype distribution**: During this analysis, the concentration of usertype by virtue of the distance they cover and by the virtue of the minutes they spend during fitness activities are observed.

Concentration of Usertype: The User type distribution by distance covered was deducted as shown in the table below:

| **userTypeDistance** | **totalPercent** |
| --- | --- |
| Sedentary | 6.617647 |
| Lightly Active | 61.029412 |
| Moderate Active | 12.500000 |
| Very Active | 19.852941 |

**Table 1: Number of usertype categories by distance covered**

Below is the visual representation of the table:

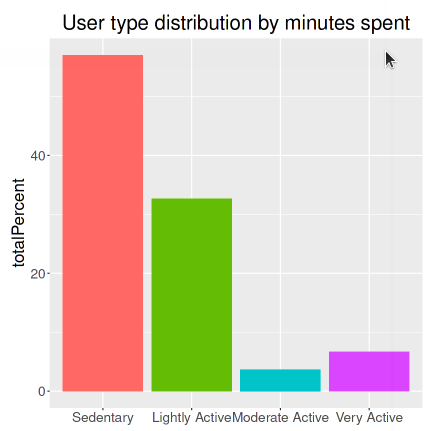
**Fig. 1: Bar chart of user type distribution by distance covered**

Here, there are more lightly active users based on the distance they covered followed by the very active ones. The least are the sedentary users. As much as individuals have subscribed to this service, for some reasons which may be personal and other factors coming in, some do have difficulties meeting up with the fitness demands. However, there is a positive that more will gradually grow into becoming more active users in the long run.

The User type distribution by minutes spent was deducted as shown in the table below:

| **userTypeMinute** | **totalPercent** |
| --- | --- |
| Sedentary | 57.012195 |
| Lightly Active | 32.621951 |
| Moderate Active | 3.658537 |
| Very Active | 6.707317 |

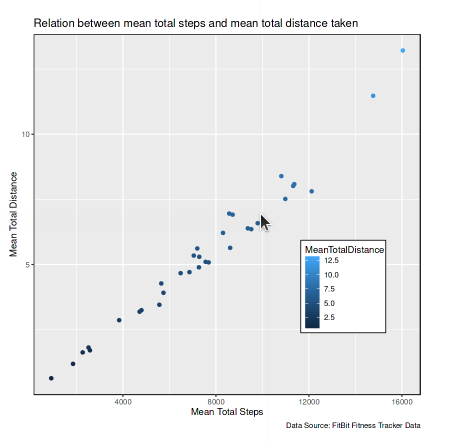
**Table 2: Number of usertype categories by minutes spent**

Below is the visual representation of the table:

**Fig. 2: Bar chart of user type distribution by minutes spent**

From the above visualization, there are more sedentary users based on the minutes spent followed by the lightly active ones. The least are the moderate active users. This is an opportunity to engage more users into spending more active time during fitness activities.

* 1. **Pattern observation by movement**: During this analysis, the relationship between the total steps taken and the distance covered is observed.

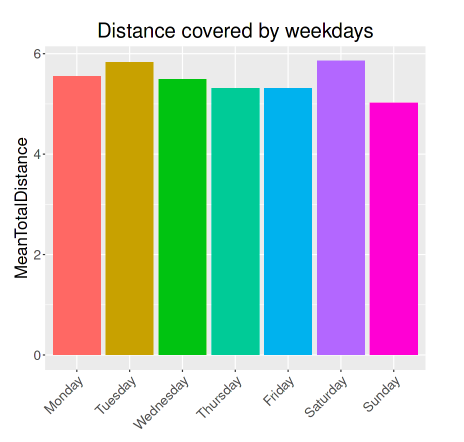
TotalSteps v distanceCovered: Below is the visualization of this relationship.

**Fig. 3: Graph plot of steps and distance relationship**

The above visualization confirms that the more steps one takes the more distance covered. However, distance doesn’t necessarily mean a long linear path but can be within a localized space. In this visualization, it is seen that a maximum of 16040.0 steps were recorded over a distance of 13.2129 Miles.

* 1. **Pattern observation by date**: During this analysis, the following relationships were observed:
* Weekdays v distanceCovered
* Weekdays v TotalSteps

Weekdays v distanceCovered: Below is the visualization relationship between Weekdays and distanceCovered.

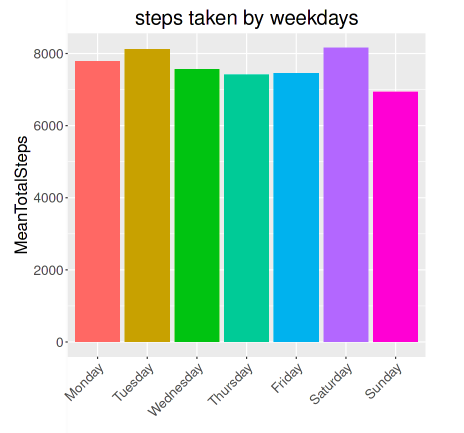


**Fig. 4: Bar plot of total distance and weekdays**

From the visualisation above, as much as each week day experiences fitness activities, there seems to be a high influx of activities on Saturdays being that this is the period where a lot of individuals have enough free time to engage in fitness activities since they are free and not engaged in any work activities.

Bellabeat can capitalize on this trend to engage more users in various fitness activities.

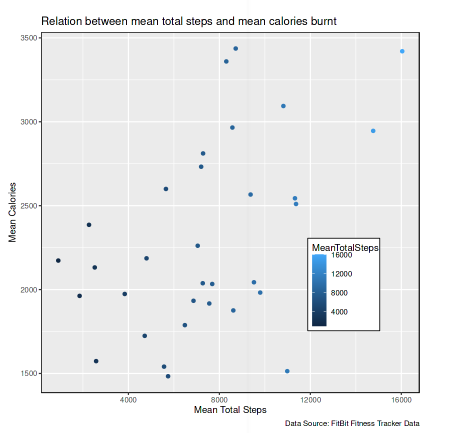
Weekdays v TotalSteps: Below is the visualization relationship between Weekdays and TotalSteps.



**Fig. 5: Bar plot of total steps and weekdays**

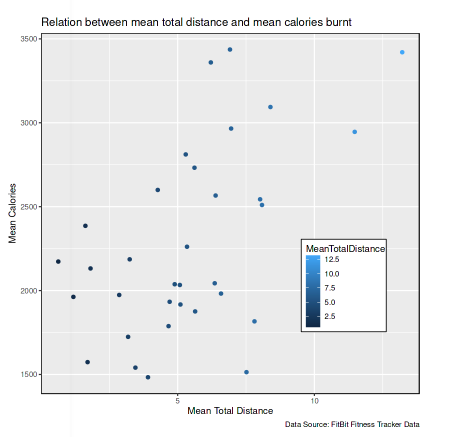
Both steps and distance have expected positive correlation which makes the steps taken on weekdays to approximately follow the same pattern as that of the distance covered on weekdays. This overall behaviour in this context suggests individuals tend to be more physically engaged on weekends.

**iv. Pattern observation by aerobic activities**: Below are the following relationships observed and their visualizations:

* Visualization of TotalSteps v Calories: 

**Fig. 6: Graph plot of calories and steps relationship**

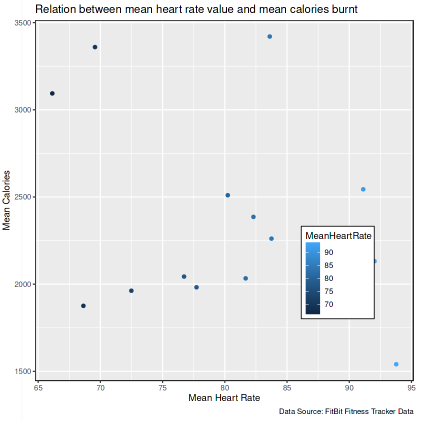
From the visualization above, there is an understandable relationship between calories burnt and the total steps taken as this indicates that more calories are burnt as steps are taken. However, the scattered nature of the graph could be as a result of individual preferences, discipline, and other factors which could have a greater effect on this relationship.

* Visualization of DistanceCovered v Calories

**Fig. 7: Graph plot of calories and distance relationship**

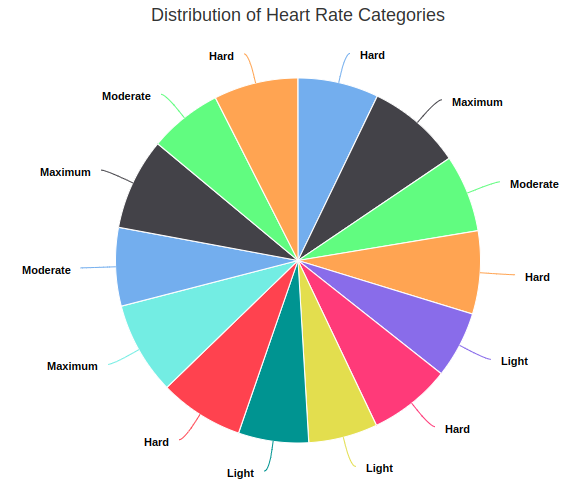
Just like the visualization for “calories” against “steps”, the visualization for “calories” against “distance” indicates burning of calories as individuals move over a distance.

* Visualization of HeartRate v Calories



**Fig. 8: Graph plot of calories and heart rate relationship**

The report above shows that engaging in fitness/aerobic activities is a good fit for maintaining a good heart rate especially if an individual is able to maintain a healthy lifestyle.

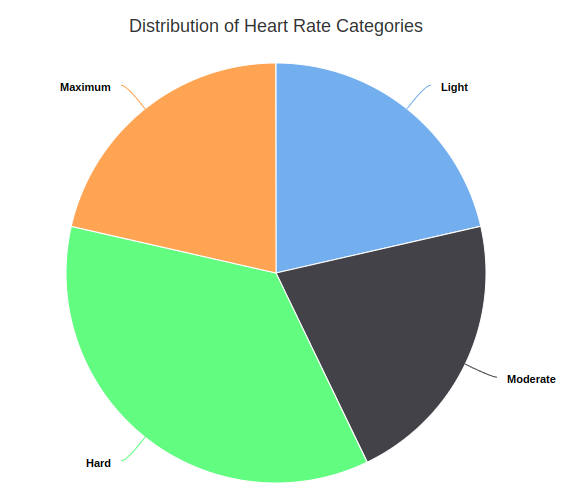


**Fig. 9: Pie chart of heart rate categories I**

The heart health range for a healthy heart rate is from 60 to 100 beats and are grouped according to “**The American Heart Association (AHA)**” general guideline for target heart rate zones during exercise as shown below:

* Very Light (50-60%)
* Light (60-70%)
* Moderate (70-80%)
* Hard (80-90%)
* Maximum (90-100%)

The above visualization shows that the captured individual’s heart health are in the safe zone. From this capture, it was affirmed that 66.13 was the lowest and the highest being 93.78.

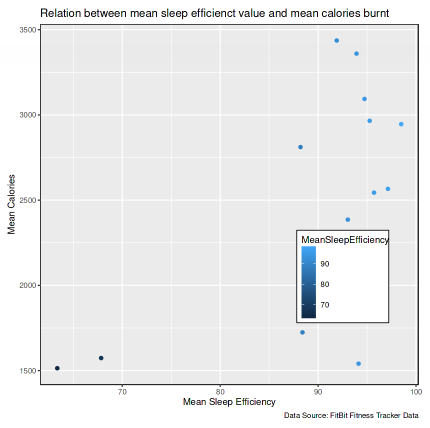


* Visualization of Calories v Sleep
* Visualization of Sleep v HeartRate

**Fig. 10: Pie chart of heart rate categories II**

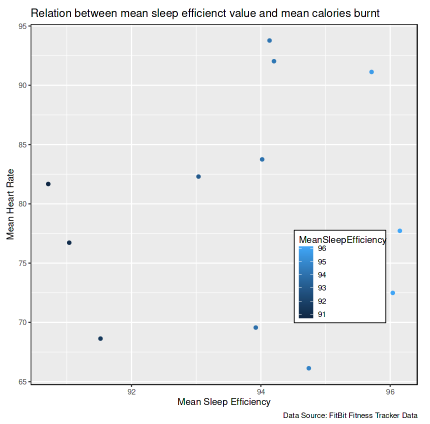
Still on the heart rate categorization, the above distribution for the captured individuals shows that heart rate of “Hard” accounts for the highest among these individuals.

* Visualization of Calories v Sleep



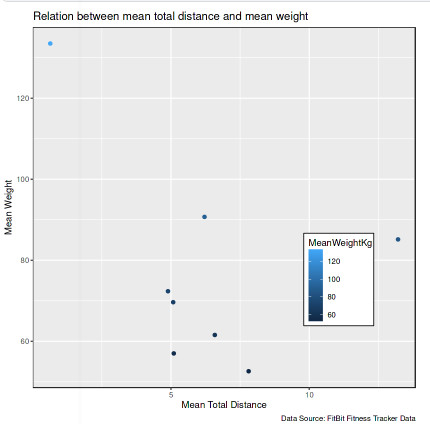
**Fig. 10: Graph plot of calories and sleep efficiency**

There is a likelihood that burning more calories can improve sleep efficiency although this has not been scientifically confirmed. However in the grand scheme of things, this should be the case in the majority of some individuals.

* Visualization of Sleep v HeartRate

**Fig. 11: Graph plot of heart rate and sleep efficiency**

It is expected that a good heart rate leads to improved sleep efficiency. However, factors affecting individuals have to be considered.

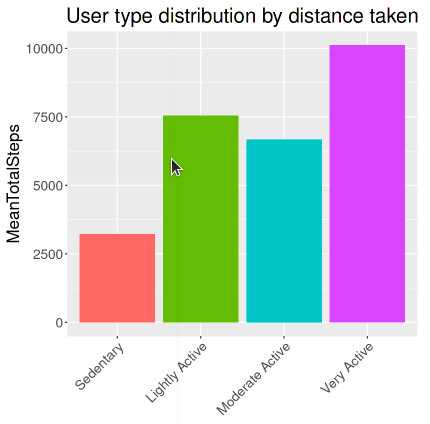
* Visualization of Weight vs distance covered

**Fig. 12: Graph plot of weight and distance**

Under normal circumstances, in the context of gaining fitness, it is expected for individuals with more weight to cover more distance compared to those with less weight but interfering factors are likely to set in which can distort these movements. It is then left for the company to find a way to encourage massive participation.

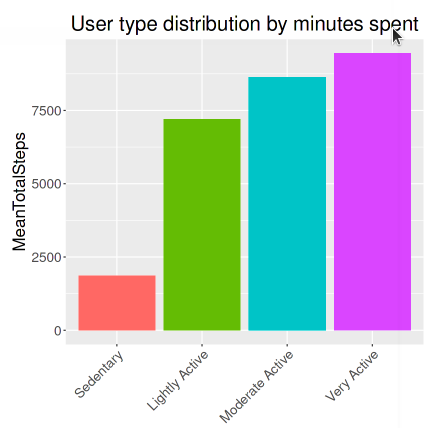
**v. Pattern observation between user type and activities:** Below are the following relationships observed and their visualizations:

* TotalSteps v usertype\_by\_distance



**Fig. 13: Bar plot of steps and user type distribution by distance taken**

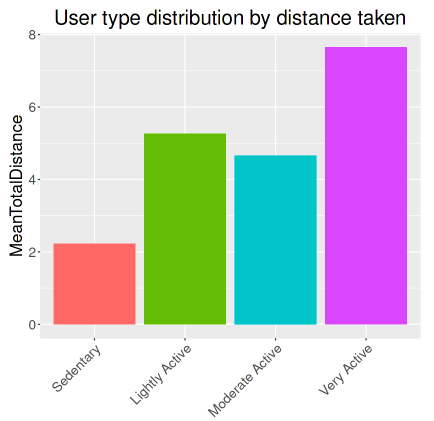
* TotalSteps v usertype\_by\_minutes



**Fig. 14: Bar plot of steps and user type distribution by minutes spent**

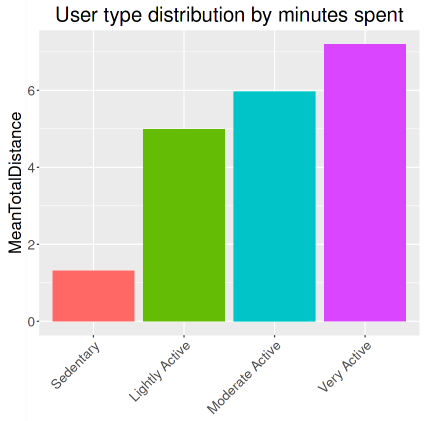
Regarding the usertype distribution by distance covered and by minutes spent, the two visualizations above depict that users who tend to traverse more distance used up more steps which of course is the normal case and these are the "Very Active" users while the "Sedentary" users used up the least steps. And these "Sedentary" users may have used up more time during this fitness activity.

* DistanceCovered v usertype\_by\_distance



**Fig. 15: Bar plot of distance and user type distribution by distance covered**

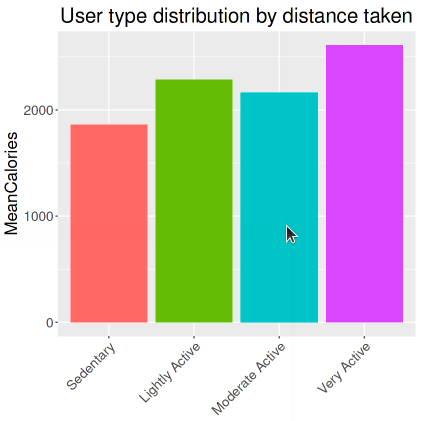
* DistanceCovered v usertype\_by\_minutes



**Fig. 16: Bar plot of distance and user type distribution by minutes spent**

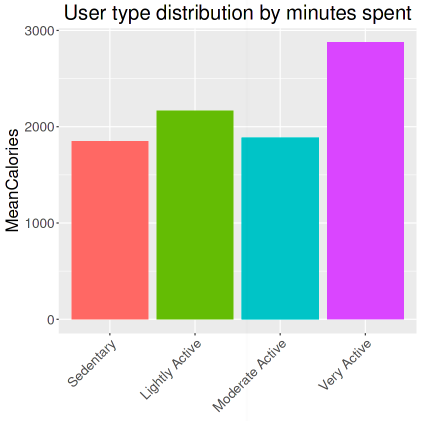
From the two visualizations above, there is a likelihood that users who traversed long distances did so while spending more time during fitness activities. These were the “**Active users**”.

* Calories v usertype\_by\_distance



**Fig. 17: Bar plot of calories and user type distribution by distance taken**

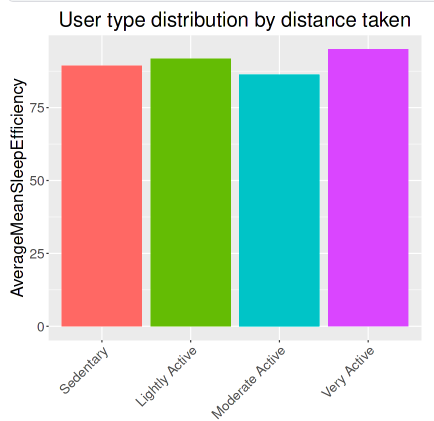
* Calories v usertype\_by\_minutes



**Fig. 18: Bar plot of calories and user type distribution by minutes spent**

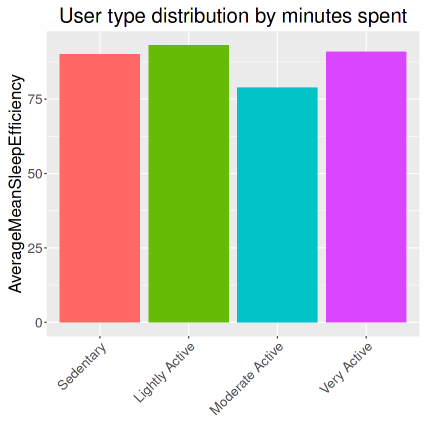
The two visualizations above depict that users who spent more time and traversed longer distances, were indeed the "**Very Active**" ones and those were the ones that burn more calories.

* Sleep efficiency v usertype\_by\_distance



**Fig. 19: Bar plot of sleep efficiency and user type distribution by distance taken**

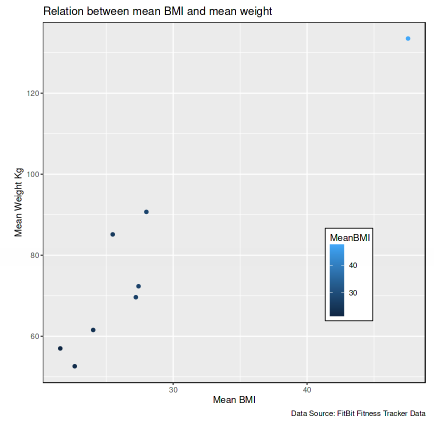
* Sleep efficiency v usertype\_by\_minutes



**Fig. 20: Bar plot of sleep efficiency and user type distribution by minutes spent**

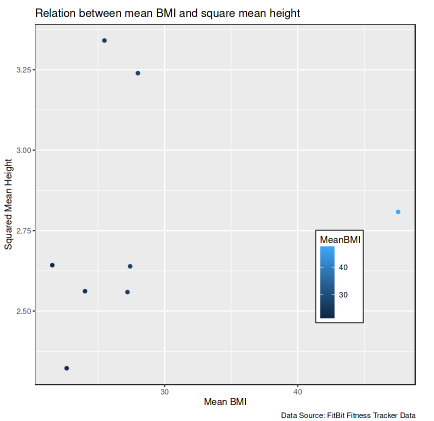
It appears as seen from the two visualizations above that users who tend to traverse more distances and spend more time have the tendency of experiencing better sleep efficiency and these are the "**Very Active**" users.

**vi.. BMI logs pattern:** Below are the following relationships observed and their visualizations:

* BMI v Weight

**Fig. 21: Graph plot of weight and BMI**

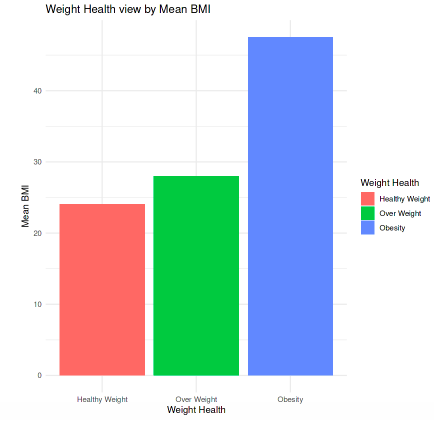
* BMI v SquaredMeanHeight



**Fig. 21: Graph plot of squared mean height and BMI**

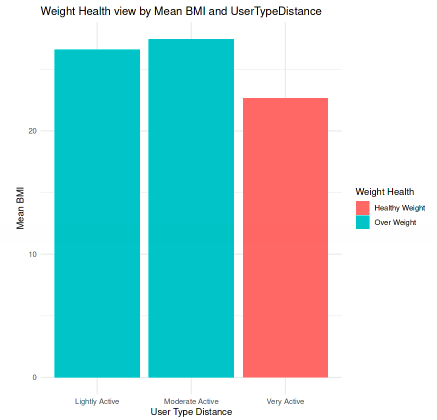
Regarding the above visualizations showing the relationships between "WeightKg" and "BMI" as well as the relationship between "squared mean height" and "BMI", the weight of an individual has a greater significance in BMI maintenance. Hence users are expected to engage in fitness activities.

* BMI v WeightHealth



**Fig. 22: Bar plot of BMI and Weight Health**

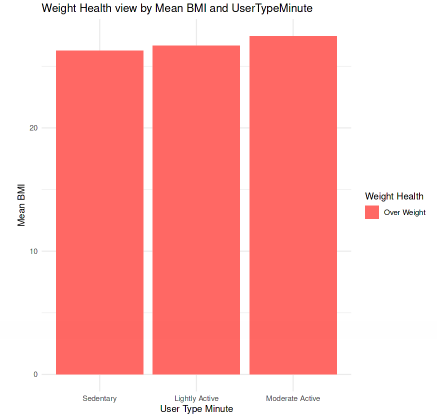
* BMI v usertype\_by\_distance



**Fig. 23: Bar plot of BMI and user type by distance covered**

The above visualizations show that there are users that are not in the safe region of height health. Any weight health that is not in the "Healthy Weight" category is not in the safe region. And it is noted that the users in the safe region are the "Very Active" users.

* BMI v usertype\_by\_minute



**Fig. 24: Bar plot of BMI and user type by distance covered**

The above visualization shows that none of the “**Very Active**” users fall within the safe weight health region. This shows the importance of such users to be actively involved in fitness activities.

* BMI v TotalDistance

The data to back up the claims here is very little but it does suggest a normal case as regards the relationship between the distance covered by an individual involved in fitness activity and his/her BMI.

| **Id** | **MeanBMI** | **MeanTotalDistance** |
| --- | --- | --- |
| 1503960366 | 22.65 | 7.809677 |
| 4319703577 | 27.45 | 4.892258 |

**Table 3: BMI and Total distance relationship**

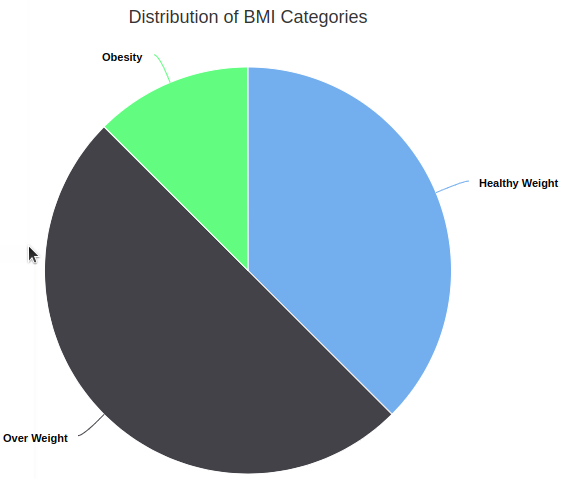
* BMI v TotalSteps

| **Id** | **MeanBMI** | **MeanTotalSteps** |
| --- | --- | --- |
| 1503960366 | 22.65 | 12116.742 |
| 4319703577 | 27.45 | 7268.839 |

**Table 4: BMI and Total steps relationship**

The two tables above show that physical fitness activities do help maintain the BMI of an individual.

* BMI classification distribution

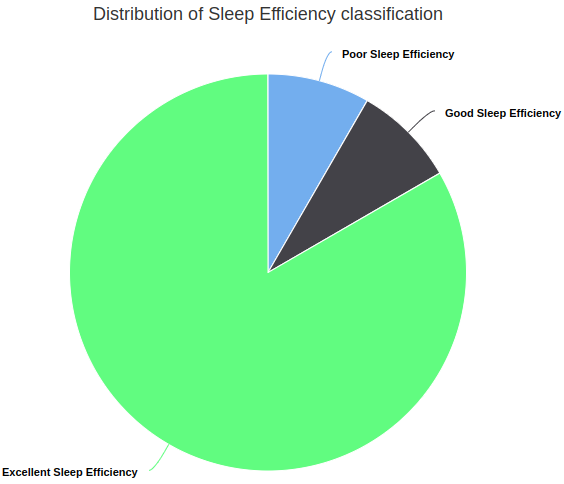


**Fig. 25: Pie chart of BMI category distribution**

The above visualization shows that there are users whose BMIs are not in the safe region. Special attention can be placed on such users using them as representative samples to other users not captured and to non subscribers as well.

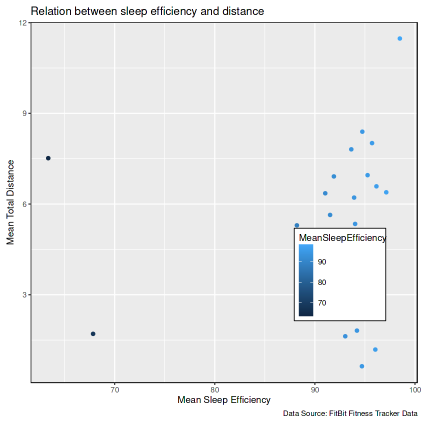
**vii. Sleep activities:** Below are the following relationships observed and their visualizations:

* Sleep quality distribution

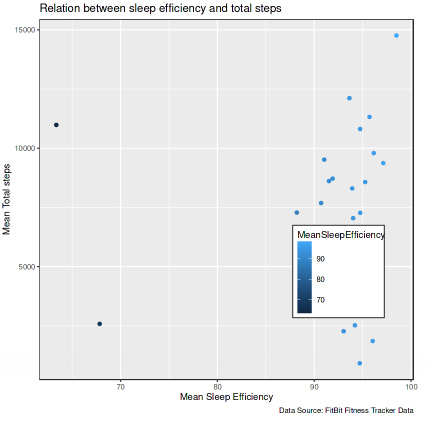


**Fig. 26: Pie chart of sleep efficiency classification distribution**

The above visualization shows that some users do experience poor sleep efficiency. The company can partner with health professionals or health centers to address challenges faced.

* Sleep efficiency vs distance covered

**Fig. 27: Graph plot of distance and sleep efficiency**

* sleep efficiency vs steps

**Fig. 28: Graph plot of steps and sleep efficiency**

Overall fitness activities in a normal state should have a great significance on the sleep efficiency of an individual.

**viii. Subscriber activities:** Below are the following relationships observed and their visualizations:

* Number of subscribers

Out of the numerous records, only 33 subscribers with complete data were assessed. The rest were dropped due to incomplete information.

However, if I am to address the overall performance of the resulting data, it approximately agrees with the various health factors and their relationships and should be able to address the business question “**How do our customers make use of our smart devices?**”.

* Reporting activities

| **IsManualReport** | **total** | **percent** |
| --- | --- | --- |
| False | 26 | 38.80597 |
| True | 41 | 61.19403 |

**Table 5: Reporting activities of subscribers**

**1.** **Summary**

Recall that the business asks the question “How do our customers make use of our smart devices?" and the goal is to design marketing strategies aimed at discovering how customers make use of the Bellabeat product (Leaf Product) which can be used to get more users over time.

After going through the data provided, processing and then analysing it, I came out with the following conclusions;

1. There are two major highlights connecting the health status of any individual:

a. Internal factors

- BMI

- Sleep Efficiency

- Overall Weight

- Heart Rate

- General body Fitness

b. External factors

- Dieting

- Consistency

- Disciplinary routines

- Behavioural pattern

2. Being consistent with daily fitness routine while adhering to health professional advice will promote better health in the areas of one's BMI, Sleep Efficiency, Weight, Heart Rate, and general body fitness.

3. Dieting while on a fitness activity is very important.

4. Mind conditioning towards fitness and general health awareness is very important. If an individual is the type that is very passive in terms of his/her health, it will be difficult for such a person to maintain and pull through during a fitness journey.

5. Apart from passive aerobic activities, every individual can set a time frame for a real active fitness activity.

And then further questions were asked such as:

* What are the reasons why some people are not in the safe weight health region?
* What could be responsible for poor sleep efficiency?
* Why aren’t some individuals’ statistics reported for a better analysis?
* How can Bellabeat improve on getting massive participation in her fitness activities and other fitness activity programs?

After going through the business case scenario once again, the answers to these business questions may center on the following areas:

* Health related factors: The reason for individuals not being in the safe weight health region could be as a result of a number of health factors which includes:
  + dieting and lifestyle.
  + Genetics
  + Medications
  + Psychological factors
  + Health conditions
  + Environment
  + Family factors
  + Food addiction
* Method of reporting activities: Data participation is very important as this is what the company will use in making necessary post activity reports especially for marketing and training/awareness purposes. It is easier if the report is done using an automated method as it will capture subscribers details. If reporting is done manually, there could be issues of error in entering customer details by the users themselves.
* Customer inclusive projects: These are projects that will attract non subscribers into participating in any fitness activities and other health related programs.
* Company marketing activities: This has to do with the marketing efforts of the company which could play a big role in customer subscribing/paying/buying behaviours.
* Other practices and policies influencing customer buying preferences or behaviours: There are other factors influencing customer buying behaviour which will be noted as time goes on.

The company can use these answers to design marketing strategies with the aim of getting more subscribers into her network.

**Business suggestions**

With careful questions asked, data processed and analysed, the following are my suggestions for Bellabeat :

* Creating a community where subscribers can get periodic health talks and awareness and get to learn about the experiences of other participants. Here non subscribers also invited to the community can perceive the need to become a subscriber to the company.
* Creating sponsorship adverts targeted to individuals who are more concerned about their health as well as the general populace where the ads can have existing members share their stories about how using Bellabeat products has been of great benefit to them and how it has changed their life.
* Offering a trial period or discounted membership rates to encourage casual riders to experience the benefits of Cyclistic membership firsthand.
* Implementing a loyalty program where subscribers earn points or rewards based on their consistency and if possible more based on their results. These rewards could be redeemed for discounts, free subscription periods, or other incentives. Such programs can create a sense of value and encourage subscribers to remain active and non subscribers to become regular members.

**Observations**

a. Working with the data

1. The datasets worked on are cohorts and some of these datasets don’t carry some information for some Ids. Also, some observations contain missing data and as a result for the purpose of visualizing the results, these observations were dropped.

b. Working with the tools

I. When using Notebook on Kaggle, the following should be noted:

* Packages are already installed, you only need to load them into memory for further use.
* It can take some minutes for the code to be loaded before the result is displayed. Merging depending on the volume of datasets involved can take a lot of memory and if overloaded can restart the kernel.
* While creating your notebook, the current contents are saved as draft. Once you are done with completing the notebook, save the version, connect to your Github account, make a commit message and save. The issue here is that Github as at the time of writing this report cannot display the full content.
* You can download the notebook which can be uploaded as a file to your Github account. With this method, you can view the content from your Github account.
* For reasons unknown to me, pie charts could not be displayed after saving the notebook and viewing the results.

c. Surprises

No surprises.

**Links**

* Click [here](https://github.com/aniekanekanem/Bellabeat-Product-Analysis-Case-Study/blob/main/bellabeat-product-analysis-case-study-i%20(2).ipynb) to access my Github to view the code used for data cleaning, processing and analysis

**What I learnt**

Below is what I learned/practised from over 80 hours spent on this project:

* Charts using Kaggle Notebook
* Practice data cleaning, processing and analysis using R for comprehensive data reports using the group\_by, summarise, mutate, piping (%>%), read.csv(), str, head, summary, and other keywords for data analysis.
* Combining multiple tables into one unit using merge in R
* Working with cohorts/segmented datasets.

**Conclusion**

It is important to note that whether we make assumptions or a claim is confirmed, further analysis and data exploration would still be valuable to gain more insights and assess the potential impact of any given pattern based on data gathered regarding the overall business operations of the company.

This data analysis report is for individual learning and development purposes and was conducted from a limited data source and as such cannot fully be relied on but can provide solutions in a given direction and aid in learning and understanding on carrying out analysis of data. However, this report and its content is subject to being updated from time to time.

**Resources**

“Datasets” by Motivate International Inc <https://divvy-tripdata.s3.amazonaws.com/index.html>

"Google capstone project V1." Kelly, J Adams, October 18, 2021

, <https://www.kellyjadams.com/post/google-capstone-project#viewer-41dce>

"Change Scale on Y-Axis." Chan Tan, April 1, 2020, <https://community.tableau.com/s/question/0D54T00000C6omSSAR/change-scale-on-yaxis>

"how to have 3 or more variables on one graph." Tableau Community, March 19, 2013, <https://community.tableau.com/s/question/0D54T00000C5eA6SAJ/how-to-have-3-or-more-variables-on-one-graph>