HOW CAN A WELLNESS COMPANY PLAY IT SMART?

A CASE STUDY ON BELLABEAT

About the company

Urška Sršen and Sando Mur founded Bellabeat, a high-tech company that manufactures health-focused smart products. Sršen used her background as an artist to develop beautifully designed technology that informs and inspires women around the world.

Collecting data on activity, sleep, stress, and reproductive health 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 2016, Bellabeat had opened offices around the world and launched multiple products. Bellabeat products became available through a growing number of online retailers in addition to their own e-commerce channel on their website. 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. Sršen knows that an analysis of Bellabeat's available consumer data would reveal more opportunities for growth.

She has asked the marketing analytics team to focus on a Bellabeat product and analyze smart device usage data in order to gain insight into how people are already using their smart devices. Then, using this information, she would like high-level recommendations for how these trends can inform Bellabeat marketing strategy.

Introduction

Welcome to the Bellabeat analysis case study! In this case study, I take on the role of a junior data analyst in Bellabeat's marketing and analytics team. Bellabeat, a dynamic small company that designs health-focused products for women.

Bellabeat is a successful small company, but they have the potential to become a larger player in the global smart device market. Urška Sršen, cofounder and Chief Creative Officer of Bellabeat, believes that analyzing smart device fitness data could help unlock new growth opportunities for the company.

I have been asked to focus on one of Bellabeat's products and analyze smart device data to gain insight into how consumers are using their smart devices.

The primary goal of this case study is to perform many real-world tasks as a data analyst and to answer the key business questions using the six phases of data analysis process which are: ask, prepare, process, analyze, share, and act.

The insights I uncover will inform the company's marketing strategy. I will share my analysis and offer high-level recommendations to the Bellabeat executive team.

Goals :

Three questions will guide my analysis:

a. What are some trends in smart device usage?

b. How could these trends apply to Bellabeat customers?

c. How could these trends help influence Bellabeat marketing strategy?

Bellabeat's cofounder and Chief Creative Officer Urška Sršen want me to analyze i.e., above question no. (a). smart device usage data in order to gain insight into how consumers use non-Bellabeat smart devices. She then wants me to select one Bellabeat product to apply these insights to in our presentation. Then, I will produce a report with the following deliverables :

i. A clear summary of the business task

ii. A description of all data sources used

iii. Documentation of any cleaning or manipulation of data 4.

iv. A summary of our analysis

v. Supporting visualizations and key findings

vi. And last one is our top high-level content recommendations based on our analysis.

Business Task :

Bellabeat, a wellness and technology company dedicated to empowering women to reach their full potential, needs support in marketing their product lineup. Bellabeat's portfolio includes the Leaf, Ivy, and Time devices, which track a variety of health metrics such as hydration, heart rate, menstrual cycles, sleep, and activity levels.

In this case study, we're focusing on a comparative analysis with competitor data, specifically examining user trends from Fitbit to identify opportunities in the wellness smart device market. The insights gleaned from this analysis will help identify growth opportunities for Bellabeat. I will also analyze how consumers use their smart devices and present these findings along with strategic marketing recommendations to the Bellabeat executive team.

To tackle this case study, I am follow the data analytics six phases as outlined by Google. i.e., Ask, Prepare, Process, Analyze, Share and Act.

1. Ask Phase :

The Ask Phase includes a clear summary of the business goal, which focuses on analyzing smart device data to uncover insights into how Bellabeat's smart devices are being used.

My objective is to study FitBit smart device user data to uncover valuable insights about trends, patterns, and connections within various health-related metrics. This analysis will help pinpoint potential growth opportunities and assist in crafting targeted marketing recommendations and strategies for the Marketing department.

We must first ask ourselves: Who are our main stakeholders? We have the following stakeholders in this situation:

Urška Sršen: Bellabeat’s co-founder and Chief Creative Officer Sando

Sando Mur: Mathematician and Bellabeat’s co-founder; key member of the Bellabeat executive team.

The Bellabeat analytics team for marketing: a group of data analysts whose job is is to gather, examine, and report data that aids in determining Bellabeat’s marketing plan.

2. Prepare Phase :

Urška Sršen recommended that I utilize publicly available data to analyze the daily habits of smart device users. She pointed me to a specific dataset, The datasets for this case study can be viewed through this **link**

This dataset comes from the FitBit fitness tracker and has been released into the public domain through Mobius on Kaggle. It includes data from thirty Fitbit customers who consented to share their personal tracker data. This collection offers minute-level details on heart rate, sleep, and physical activity, plus daily records of activity, steps, and heart rate, which can provide insights into user routines. There are 18 CSV files available on Kaggle containing this FitBit tracker data.

During the Prepare phase, we assess the data and its limitations:

The data was collected in 2016, making it seven years old. Changes in daily routines, diet, exercise, and sleep patterns over the years might make this data less relevant or outdated.

With data from only 30 FitBit users, the sample size is too small to accurately reflect the broader fitness market.

Since the data comes from a survey, we cannot fully vouch for its accuracy or integrity.

Is the data ROCCC?

ROCCC stands for Reliable, Original, Comprehensive, Current, and Cited, which are criteria for assessing data quality.

Reliability: LOW — The data, sourced from just 30 respondents, is not robust enough to be considered reliable.

Originality: LOW — The data was collected by a third-party provider, Amazon Mechanical Turk.

Comprehensiveness: MEDIUM — The parameters covered largely align with those tracked by most Bellabeat devices.

Current: LOW — The data is 8 years old and may not reflect current trends or behaviors.

Citation: LOW — The data comes from an unspecified third party.

Overall, the quality of the dataset is considered low, making it unsuitable for basing critical business decisions on it.

3. Process Phase :

Process Phase includes processing the data by cleaning and ensuring that it is correct, relevant, complete and error free.

To manage this data effectively, I employed one tools. I utilized RStudio for data cleaning, transformation, analysis, and visualization.

To start, we must install and load the necessary packages for analysis. I already have all the required packages installed, so I load them all at once.

Cleaning Process:

I started by setting up RStudio, ensuring that all necessary library packages were installed for data processing and cleaning.

```
# Install Packages

install.packages("tidyverse")
install.packages("lubridate")
install.packages("ggplot2")
install.packages("ggplot2")
install.packages("dplyr")
install.packages("janitor")
install.packages("plotrix")
install.packages("skimr")
install.packages("sqldf")

# Import necessary libraries

library(tidyverse)
library(lubridate)
library(ggplot2)
library(dplyr)
library(janitor)
library(plotrix)
library(skimr)
library(sqldf)
```

After installing the necessary libraries, I imported the relevant datasets into RStudio.

```
# Importing the data

daily_activity <- read_csv("dailyActivity_merged.csv")

sleep_per_day <- read_csv("sleepDay_merged.csv")

weight_log_info <- read_csv("weightLogInfo_merged.csv")
```

Quick Analysis of Data :

```
# Daily Activity Dataset

str(daily_activity)
skim(daily_activity)
head(daily_activity)
glimpse(daily_activity)

# Sleep Per Day Dataset

str(sleep_per_day)
skim(sleep_per_day)
head(sleep_per_day)
glimpse(sleep_per_day)

# Weight Log Info Dataset

str(weight_log_info)
skim(weight_log_info)
head(weight_log_info)
glimpse(weight_log_info)
```

After running the commands, we gathered the following information:

- Total number of records and columns.

- Count of null and non-null values.

- Data types of each column.

From this, we discovered there are 67 entries in the weight log info, 413 in the sleep per day data, and 940 in the daily activity log. We noted some null values in the "Fat" variable within the weight log info, which we can address by either removing the variable or imputing the missing values with the most frequent ones. Apart from this, the dataset requires minimal cleaning. We also need to convert the date column from character format to datetime64 type for proper analysis. Additionally, I created columns for the year, month, day, and day of the week to facilitate our analysis.

```
# The date column is in character format, so we need to convert it into datetime64 type.

daily_activity$Rec_Date <- as.Date(daily_activity$ActivityDate,"%m/%d/%y")

daily_activity$month <- format(daily_activity$Rec_Date,"%B")

daily_activity$day_of_week <- format(daily_activity$Rec_Date,"%A")

# Fixing NA values

# Check the current count of 'Fat' values
table(weight_log_info$Fat)

# Replace missing 'Fat' values with random integers between 22 and 25
set.seed(123)  # Set seed for reproducibility
weight_log_info <- weight_log_info %>%
  mutate(Fat = if_else(is.na(Fat), sample(22:25, 1, replace = TRUE), Fat))

# Check the updated counts of 'Fat' values
table(weight_log_info$Fat)
```

I will address the missing values in the 'Fat' variable by imputing them based on the available data. Alternatively, we could consider removing the 'Fat' variable altogether if it does not provide any significant insights.

We will also count the number of unique IDs to verify if the data indeed contains 30 IDs as reported in the survey.

```
# count unique IDs

n_distinct(daily_activity$Id)
```

After executing the above command, we found 33 unique IDs instead of the expected 30. This discrepancy could be due to some participants creating additional IDs during the survey period.

With the data cleaning and manipulation complete, the data is now prepared and ready for analysis.

4. Analyze Phase :

Now, we need to summarize the data. So that we can find some insights about the data.

```
# Now, we need to summarize the data. So that we can find some insights about the data.

daily_activity %>%
select(TotalSteps,TotalDistance,SedentaryMinutes,VeryActiveMinutes) %>%
summary()

# Data Preparation
daily_activity$Rec_Date <- as.Date(daily_activity$ActivityDate,"%m/%d/%y")
daily_activity$month <- format(daily_activity$Rec_Date,"%B")
daily_activity$day_of_week <- format(daily_activity$Rec_Date,"%A")


# Data Summary
daily_activity %>%
select(TotalSteps,TotalDistance,SedentaryMinutes,VeryActiveMinutes) %>%
summary()
weight_log_info %>%  select(WeightKg,BMI) %>% summary()


# Find the average sleeping time in minutes
Avg_minutes_asleep <- sqldf("SELECT
SUM(TotalSleepRecords),SUM(TotalMinutesAsleep)/SUM(TotalSleepRecords) as avg_sleeptime
FROM sleep_per_day")
Avg_minutes_asleep

# Find the average bed time in minutes
Avg_TimeInBed <- sqldf("SELECT
SUM(TotalTimeInBed)/SUM(TotalSleepRecords) as avg_timeInBed
FROM sleep_per_day")
Avg_TimeInBed

n_distinct(sleep_per_day$Id)
n_distinct(weight_log_info$Id)
```

Findings:

Daily Activity Dataset:

a. The average distance traveled per day is 5.490 km, which falls short of the recommended 8 km, and the average number of steps taken is 7,638, below the recommended 10,000 steps.

b. Users spend an average of 991.2 minutes, or about 16.52 hours, in sedentary activities daily, which exceeds the recommended maximum of 7 hours.

c. The average daily caloric intake is 2,304 kcal.

d. The goal of 30 minutes of highly active time per day is not met, with an average of only 21.16 minutes.

Sleep Per Day Dataset:

a. On average, users sleep for about 419 minutes, or approximately 7 hours per night.

b. The typical duration spent in bed each night is 458 minutes, or 7 hours and 30 minutes.

c. This indicates that users typically spend about 30 minutes awake in bed.

Weight Log Info Dataset:

a. It is not possible to assess a person's health solely based on their weight; factors like height and body fat percentage also influence health.

b. The average body fat percentage among users is 23.5%.

c. While a healthy BMI range is considered to be between 18 and 24.9, the average BMI recorded is 25.19, slightly above the upper limit.

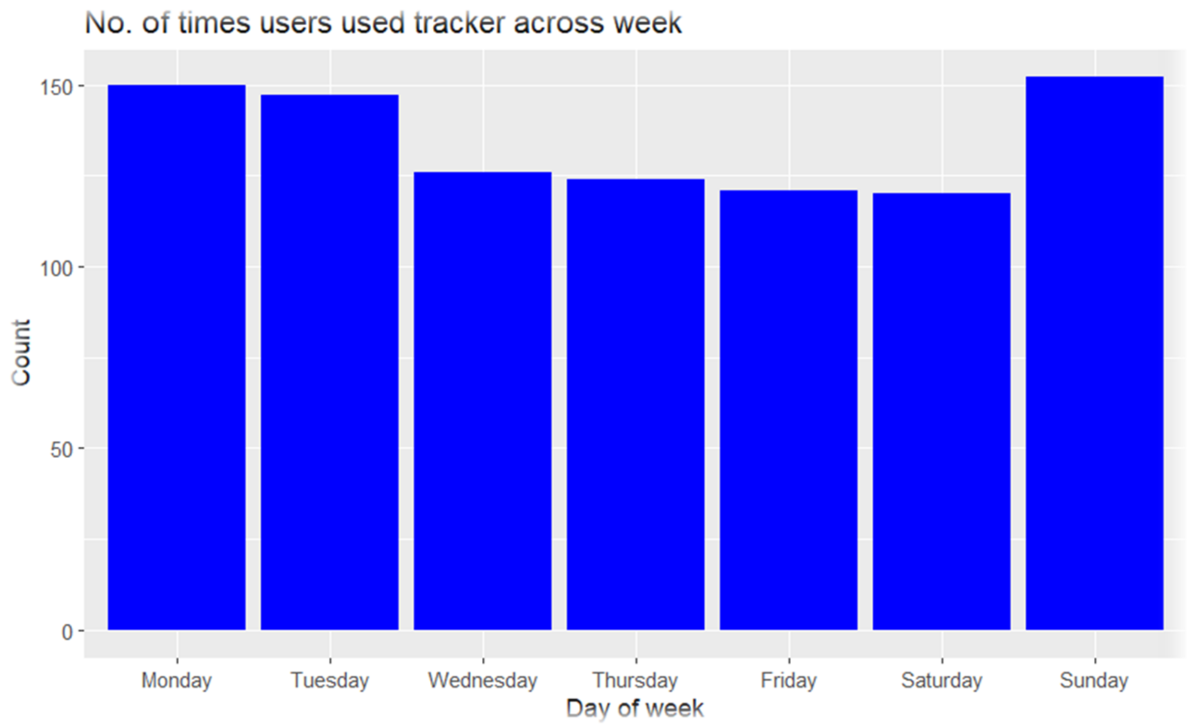
d. The average weight of users is 72 kg, or 158.8 pounds.

5. Share Phase :

In this phase, we will develop several visualizations to illustrate the findings from our analysis and align them with the objectives of our project.

```
# --------------------- Data Visualizations ---------------------
daily_activity$day_of_week <-
ordered(daily_activity$day_of_week,levels=c("Monday","Tuesday","Wednes
day","Thursday","Friday","Saturday","Sunday"))

ggplot(data=daily_activity) + geom_bar(mapping =
aes(x=day_of_week),fill="Blue") +
  labs(x="Day of week",y="Count",title="No. of times users used tracker
across week")
```



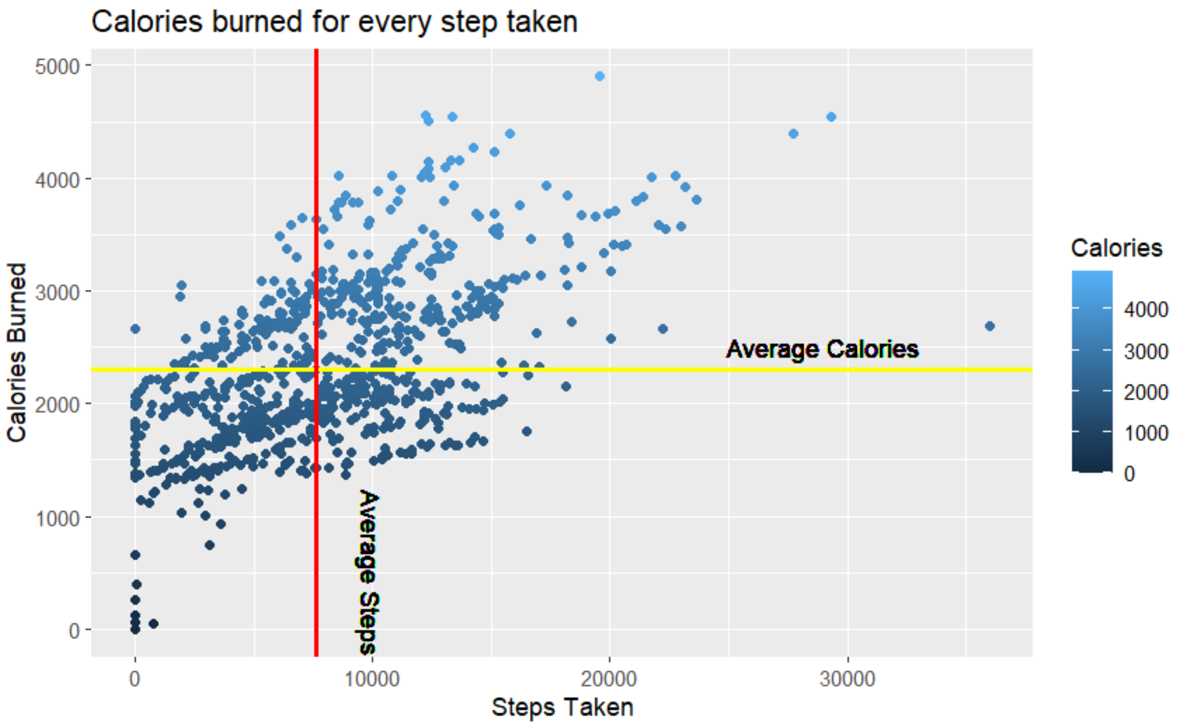
Visualization : 1

From our analysis, it's evident that usage rates for the FitBit fitness tracker app are highest on Sunday, Monday, and Tuesday compared to other days of the week. This pattern likely stems from individuals having busier schedules during the weekend, limiting their time to monitor their activities. Consequently, people tend to be more active on Sunday and at the start of the week.

```
mean_steps <- mean(daily_activity$TotalSteps)
mean_calories <- mean(daily_activity$Calories)
mean_steps
mean_calories
```

```
# Calories burned for every step taken

ggplot(data=daily_activity) + geom_point(mapping=aes(x=TotalSteps,
y=Calories, color=Calories)) +
  geom_hline(mapping =
aes(yintercept=mean_calories),color="yellow",lwd=1.0) +
  geom_vline(mapping = aes(xintercept=mean_steps),color="red",lwd=1.0) +
  geom_text(mapping = aes(x=10000,y=500,label="Average Steps",srt=-90))
+
  geom_text(mapping = aes(x=29000,y=2500,label="Average Calories")) +
  labs(x="Steps Taken",y="Calories Burned",title = "Calories burned for every
step taken")
```

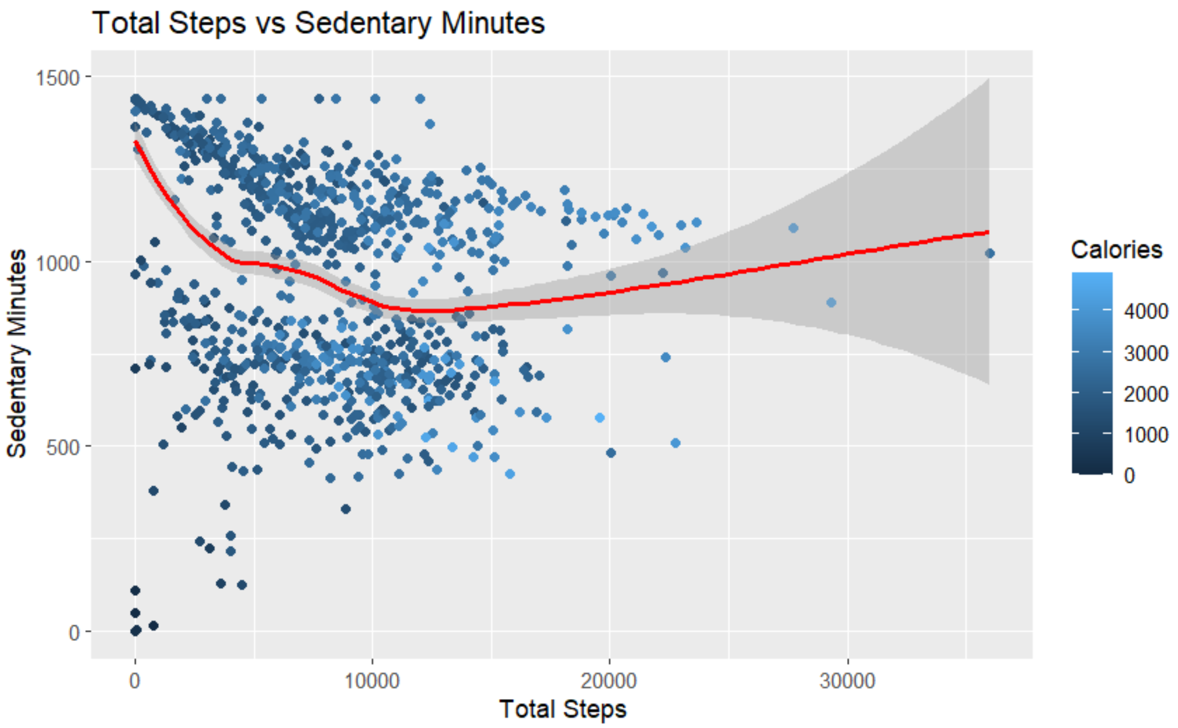


Visualization : 2

The scatter plot shows a positive correlation between the number of steps taken and the calories burned, with some outliers present at both the lower and upper ends. The data clearly indicates that as the number of steps increases, the intensity of calories burned also rises.

```
# Total Steps vs Sedentary Minutes

ggplot(data=daily_activity, aes(x=TotalSteps, y=SedentaryMinutes, color =
Calories)) + geom_point() +
  geom_smooth(method = "loess",color="red") +
  labs(x="Total Steps",y="Sedentary Minutes",title="Total Steps vs Sedentary
Minutes")
```



Visualization : 3

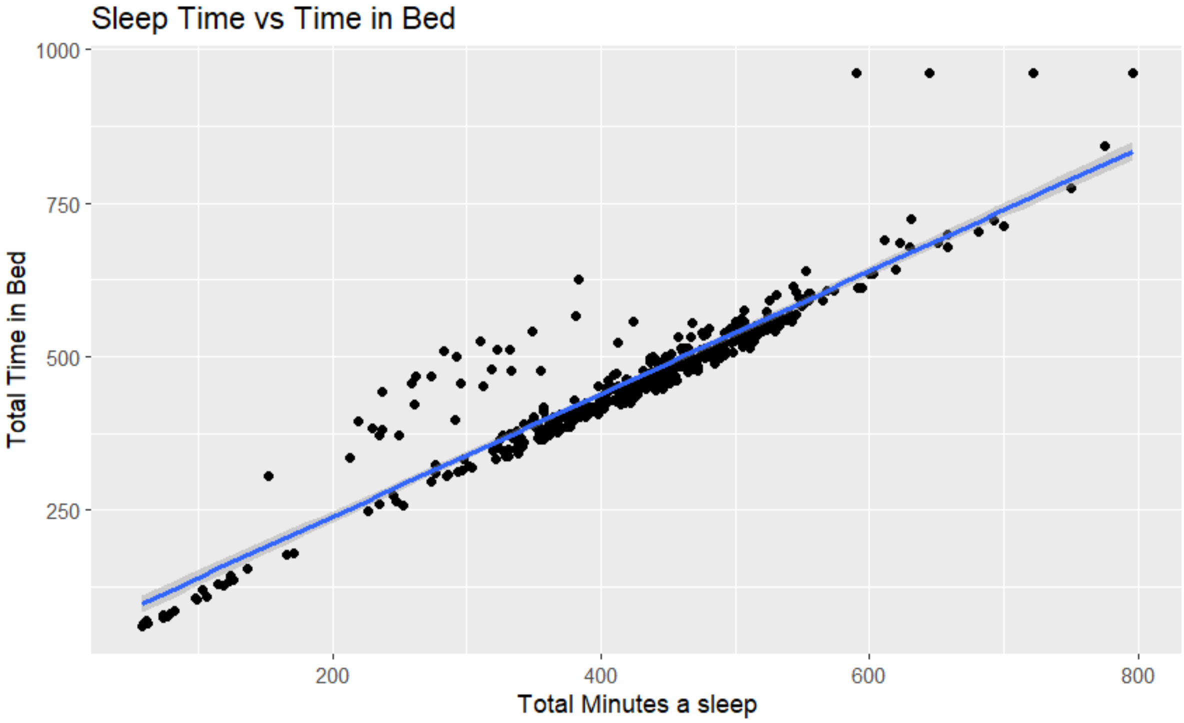
Total Steps Taken vs. Sedentary Minutes :

I anticipated a completely inverse relationship between the number of steps taken and sedentary minutes.

Initially, when the steps are fewer than 10,000, there is an inverse relationship, but once the step count exceeds 10,000, the relationship does not change dramatically. It was surprising to observe that after 15,000 steps, the relationship between steps and sedentary minutes becomes slightly positive.

```
# Sleep Time vs Time in Bed

ggplot(data=sleep_per_day, aes(x=TotalMinutesAsleep, y=TotalTimeInBed))
+ geom_point() + stat_smooth(method = lm) +
  labs(x="Total Minutes a sleep", y="Total Time in Bed", title = "Sleep Time
vs Time in Bed")
```



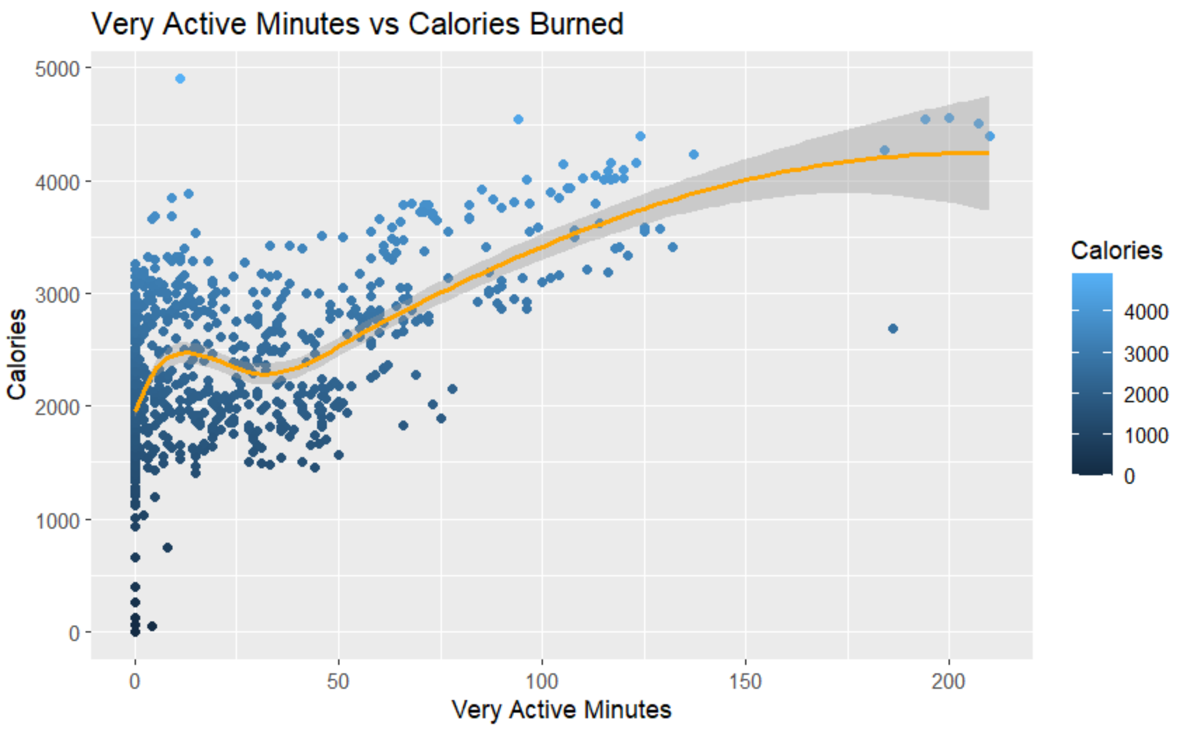
Visualization : 4

Relationship Between Sleep and Time in Bed:

There is a strong positive correlation between the total minutes asleep and the total time in bed, although some outliers are evident in the middle and upper portions of the plot. These outliers represent individuals who spend a significant amount of time in bed but do not actually sleep much. Various factors could explain this behavior.

```
# Very Active Minutes vs Calories Burned

ggplot(data=daily_activity,aes(x = VeryActiveMinutes, y = Calories, color =
Calories)) + geom_point() +
  geom_smooth(method = "loess",color="orange") +
  labs(x="Very Active Minutes",y="Calories",title = "Very Active Minutes vs
Calories Burned")
```



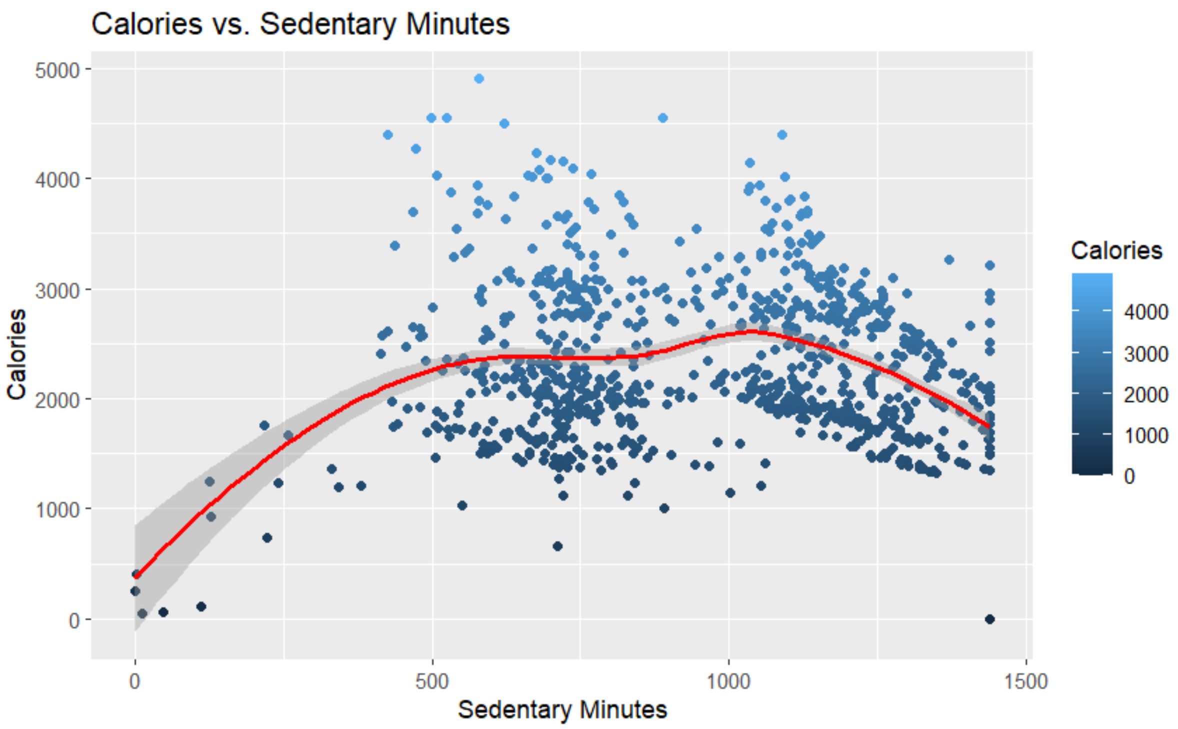
Visualization : 5

Relationship Between Very Active Minutes and Calories Burned:

There is a strong correlation between very active minutes and calories burned, as observed in the data. However, some outliers are present, notably in the bottom left and top left of the plot, which deviate from the general trend.

```
# Calories vs. Sedentary Minutes

ggplot(data=daily_activity,aes(x=SedentaryMinutes,y=Calories,color=Calories)) + geom_point() +
  geom_smooth(method="loess",color="red") +
  labs(y="Calories", x="Sedentary Minutes", title="Calories vs. Sedentary Minutes")
```



Visualization : 6

Relationship Between Sedentary Minutes and Calories Burned:

I anticipated a completely inverse relationship between sedentary minutes and calories burned.

Interestingly, the data shows a positive correlation for up to 1000 sedentary minutes. However, beyond 1000 sedentary minutes, the relationship becomes inverse, aligning with my expectations.

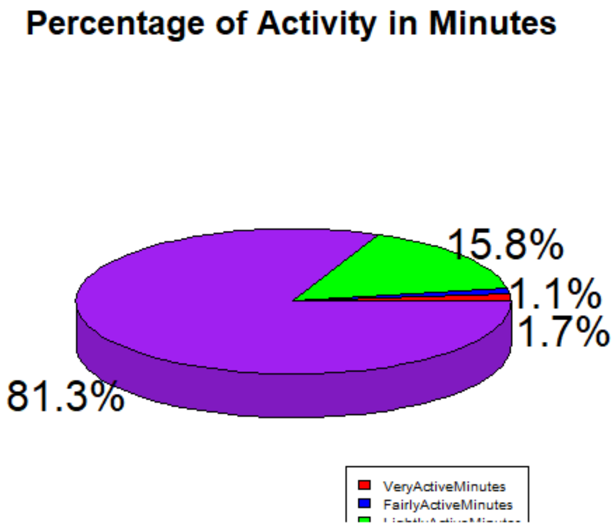
```
# Calculating the sum of individual minute

activity_min <- sqldf("SELECT
SUM(VeryActiveMinutes),SUM(FairlyActiveMinutes),
    SUM(LightlyActiveMinutes),SUM(SedentaryMinutes)
    FROM daily_activity")
activity_min

# Percentage of Activity in Minutes

x <- c(19895,12751,181244,931738)
piepercent <- round(100*x / sum(x), 1)
colors = c("red","blue","green","purple")
pie3D(x,labels = paste0(piepercent,"%"),col=colors,main = "Percentage of
Activity in Minutes")
```

```
legend("bottomright",c("VeryActiveMinutes","FairlyActiveMinutes","Lightly
ActiveMinutes","SedentaryMinutes"),cex=0.5,fill = colors)
```



Visualization : 7

Percentage of Activity in Minutes:

The data reveals that sedentary minutes comprise a substantial portion, accounting for 81.3% of the total activity, indicating that people are inactive for most of the time.

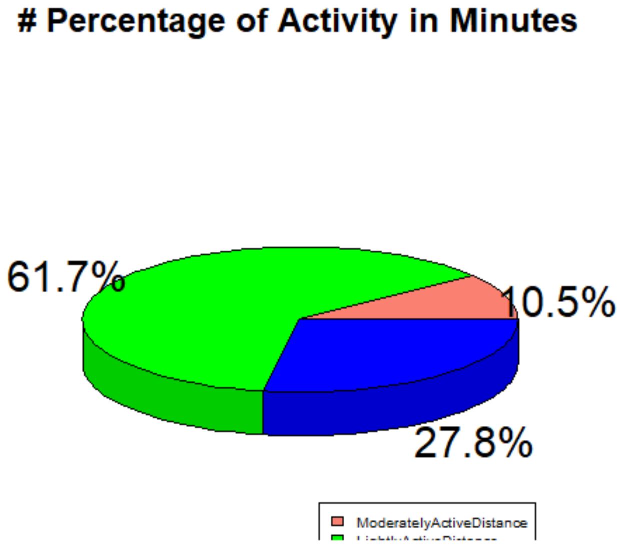
Conversely, the percentages for very active and fairly active minutes are quite low, at 1.7% and 1.1% respectively, which are significantly smaller in comparison to other activity levels.

```
# Calculating the sum of different distance values

activity_dist <- sqldf("SELECT
SUM(ModeratelyActiveDistance),SUM(LightActiveDistance),
     SUM(VeryActiveDistance),SUM(SedentaryActiveDistance)
     FROM daily_activity")
activity_dist
```

```
# Percentage of Activity in Minutes

y <- c(533.49,3140.37,1412.52)
piepercent <- round(100*y / sum(y), 1)
colors = c("salmon","green","blue")
pie3D(y,labels = paste0(piepercent,"%"),col=colors,main = "# Percentage of
Activity in Minutes
")
legend("bottomright",c("ModeratelyActiveDistance","LightlyActiveDistance"
,"VeryActiveDistance"),cex=0.5,fill = colors)
```



Visualization : 8

Percentage of Activity by Distance:

The data shows that lightly active distance accounts for the majority at 61.7%, while moderately active distance represents 10.5%.

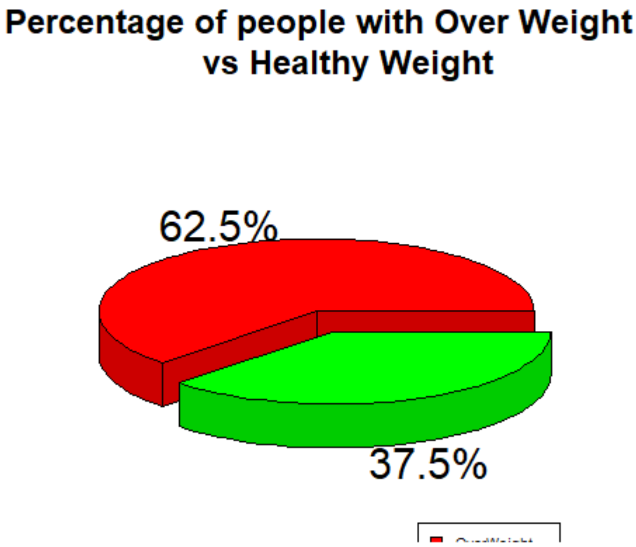
The percentage for very active distance stands at 27.8%, which is commendable. However, there is room for improvement to help individuals reach their fitness goals more effectively.

```
# calculating the count of people with over weight

count_overweight <- sqldf("SELECT COUNT(DISTINCT(Id))
                    FROM weight_log_info
                    WHERE BMI > 24.9")
count_overweight

# Percentage of people with Over Weight vs Healthy Weight

z <- c(5,3)
piepercent <- round(100*z / sum(z),1)
colors = c("red","green")
pie3D(z,labels=paste0(piepercent,"%"),explode=0.1,col=colors,radius=1,main
="Percentage of people with Over Weight
    vs Healthy Weight")
legend("bottomright",c("OverWeight","HealthyWeight"),cex=0.5,fill=colors)
```



Visualization : 9

Although the dataset contains a limited number of records, it reveals a significant discrepancy in weight categories. Specifically, 62.5% of individuals are categorized as overweight, compared to 37.5% who are within a healthy weight range. This imbalance presents a substantial opportunity to enhance the proportion of the population that achieves a healthy weight.

6. Act Phase :

The final step involves presenting my observations and offering recommendations based on my analysis. To do this effectively, I will revisit the initial business questions.

a. What are some trends in smart device usage?

Based on our analysis, we've found that the majority of people utilize apps primarily to monitor their activity levels and calories burned, while fewer individuals use them to track their sleep, and even fewer to monitor their weight. Therefore, I recommend enhancing the focus on tracking steps, calories, and sleep within the application.

The average sedentary time for users is approximately 16 hours per day. The typical daily step count is 7,638, which falls short of the recommended 10,000 steps for adults.

The average user sleeps for about 419 minutes, or roughly 7 hours per night. Additionally, users typically spend 458 minutes in bed each night, which is about 7 hours and 30 minutes. From these observations, it can be inferred that the average user spends around 30 minutes awake in bed.

People tend to track their activities more on Sunday, Monday, and Tuesday compared to other weekdays. This pattern may stem from increased work pressures toward the end of the week, which leaves less time for activity tracking. Consequently, individuals are more active and engaged in tracking their activities on Sunday and the first two days of the week.

b. How could these trends apply to Bellabeat customers?

We have obtained very detailed information, particularly from female users. Fitbit, in contrast, provides products that help customers gain a deeper understanding of their sleep, activity, and overall health habits, while also granting them access to this data to promote healthier practices. These insights are relevant given that Bellabeat's user base exhibits similar characteristics.

A significant portion of fitness tracker users (62.5%) are categorized as obese. This presents an opportunity to encourage individuals to adopt healthier lifestyles.

c. How could these trends help influence Bellabeat marketing strategy?

Effectively communicate across all channels about Bellabeat's wearable technology and instruct users on how to maximize its benefits. Tailor communications to cater to different user demographics.

To provide users with a comprehensive view of their health, it's crucial for them to be able to monitor their sleep. Investigating the reasons why users might not wear the device while sleeping should be a priority. This could be addressed through focus groups or customer surveys.

Introduce guided goal-setting programs that allow users to establish fitness and wellness objectives. Assist users in accessing their data and visualizing their progress toward these goals. Clearly explain the various types of metrics available, what they represent, and how they can be utilized to track health and fitness progress.

Name : Jitu Kumar

Email : jitukumar9387@gmail.com

Contact Me : Blog | LinkedIn | GitHub