Project: Bellabeat Data Analysis | R

Contents

1. Company Background ¶

Bellabeat (https://bellabeat.com) is high tech company that manufactures health-focused smart products.It was found in 2013 that Bellabeat has grown rapidly & positioned itself as a tech-driven wellness company for women. Bellabeat had opened offices around the world by 2016.Products are available through a growing number of online retailers in addition to their own e-commerce channel on their website.

Bellabeat has launched 5 products:

- Bellabeat app (https://bellabeat.com/health/): The Bellabeat app provides users with health data related to their activity, sleep, stress, menstrual cycle, and mindfulness habits. It helps users better understand their current habits and make healthy decisions.
- Leaf (https://bellabeat.com/leaf-urban/): Bellabeat's classic wellness tracker can be worn as a bracelet, necklace, or clip. It connects to the Bellabeat app to track activity, sleep, and stress.
- Time (https://bellabeat.com/time/): This wellness watch combines the timeless look of a classic timepiece with smart technology to track user activity, sleep, and stress. It connects to the Bellabeat app to provide you with insights into your daily wellness.
- Spring (https://bellabeat.com/spring/): This is a water bottle that tracks daily water intake using smart technology to ensure that you are appropriately hydrated throughout the day. It connects to the Bellabeat app to track your hydration levels.
- Bellabeat membership (https://bellabeat.com/coach/): Bellabeat also offers a subscription-based membership program for users. It gives users 24/7 access to fully personalized guidance on nutrition, activity, sleep, health and beauty, and mindfulness based on their lifestyle and goals.

2. Ask

2.1 The Business Task

As a team member of a marketing analytics team, I have to focus on a bellabeat product and have to gain insight of its usage by customers.Also, discovering trends from the sample data of users of fitbit. Lastly, coming up with recommendation for how these trends can help bellabeat improve marketing strategies.

2.2 Key Stakeholders

- Urška Sršen: cofounder and Chief Creative Officer.
- Sando Mur: cofounder and a key member of the Bellabeat executive team.

3. Prepare

In total, data has 18 CSV files which are available from FitBit Fitness Tracker Data (https://www.kaggle.com/arashnic/fitbit) (CC0: Public Domain, dataset made available through Mobius). This Kaggle data set contains personal fitness tracker from 33 Fitbit users consented to the submission of personal tracker data.Out of 18, 4 CSV files were used basically for this analysis. Reason for this is that other CSV files were included in this 4 CSV files already.

Also,We are using R as a major tool. We will run SQL queries in R for cleaning, analysis & visualizing. We will using RStudio for whole process.Initially inspection of data will be done by Excel, then importing data into RStuduio & doing all process at one platform.

Data Limitations

1. Data is from 2016 & not current.
2. Sample size is small for proper data analysis.
3. Bellabeat launch products for women wellness but data is recorded by fitbit & not sure about gender of the customers

3.1 Loading packages

Here are the packages used in this case study:

```
{r}

install.packages("tidyverse")
library(tidyverse)
library(readr)
install.packages("sqldf")
library(sqldf)
library(janitor)
library(skimr)
library(lubridate)
library(ggplot2)
```

3.2 Importing the data

Now when I have loaded all required packages, now importing the data is next step.For this, I used 'Upload' option to import Zip folder from PC into R.Once data is in R, now creating a new folder for this project & importing all files into the folder.

```
{r}
bellabeat <- "/cloud/project/bellabeat_data"
setwd(bellabeat)
all_files <- dir(bellabeat,pattern ="*.csv")
all_files
```

3.3 Naming files

Data has been loaded from the csv files into new data frames created. All the tables that were required are loaded into new data frames. As d_activity table is having information from all tables, so not all tables were loaded.

```
{r}
d_activity <- read_csv("dailyActivity_merged.csv")
d_calories <- read_csv("dailyCalories_merged.csv")
d_intensity <- read_csv("dailyIntensities_merged.csv")
d_steps <- read_csv("dailySteps_merged.csv")
sleep_day <- read_csv("sleepDay_merged.csv")
weight_info <- read_csv("weightLogInfo_merged.csv")
```

3.4 Inspecting data

In this step, I checked the data for it sample size, unique Ids & total number of entries by unsing JOIN in SQL. For this we need to load sqldf() first, & then run queries.

```
{r}
library(sqldf)
sqldf("SELECT COUNT()
      FROM d_activity
      JOIN d_calories ON
      d_activity.Id = d_calories.Id AND
      d_activity.ActivityDate = d_calories.ActivityDay AND
      d_activity.Calories = d_calories.Calories")
sqldf("SELECT COUNT()
      FROM d_activity
      JOIN d_steps  ON
      d_activity.Id = d_steps.Id AND
      d_activity.ActivityDate = d_steps.ActivityDay AND
      d_activity.Totalsteps = d_steps.StepTotal")

sqldf("SELECT COUNT()
      FROM d_activity
      LEFT JOIN d_intensity  ON
      d_activity.Id = d_intensity.Id AND
      d_activity.ActivityDate = d_intensity.ActivityDay
     ")
```

Data is verified to proceed for analysis. d_activity data is collection of data from other tables,So, d_intesity data will be out main focus for analysis.Table sleep_day is having dates in different format, so we will try to change that in next step.

Also to check out sample size, I used following code:

```
{r}
sqldf("SELECT * FROM d_activity GROUP BY Id")
```

This showed that we are having 33 unique Ids in our data. This is small sample size for analysis. As Stakeholder have provided this data, we need to focus on these 33 to get to some conclusion.

4. Process

First step in this phase is data cleaning. We will check out for missing values, null values, making format consistent in columns & try to transform data as per out need for analysis.

4.1 Converting date and time format

Here we will convert character format to date format. First, we make consistent date format in `d_activity` and `sleep_days`

```{r eval=FALSE, include=FALSE} d_activity ← d_activity %>% rename(date= ActivityDate) %>% mutate(date= as_date(date, format= "%m/%d/%Y")) sleep_days ← sleep_days %>% mutate(date= as_date(date, format= "%m/%d/%Y %I:%M:%S %p", tz= Sys.timezone())) head(d_activity) head(sleep_days)

```
4.2 Removing Duplicates & NULL values

First to find out duplicates in data , I used following code & later try to remove duplicates & NULL values.

```{r eval=FALSE, include=FALSE}
sum(duplicated(sleep_days))
sum(duplicated(d_activity))
```

Output showed 3 duplicates in sleep_days table, so to remove it:

```
{r}
d_activity <- d_activity %>%
  distinct() %>%
  drop_na()
```

```{r eval=FALSE, include=FALSE} sleep_days ← sleep_days %>% distinct() %>% drop_na()
```

~~~
## <span style="color:#007BA7"> 5. Analyze </span> <a class="anchor" id="analyze_share"></a>



### 5.1 Data summary<a class="anchor" id="daily"></a>

First, summarizing the data before checking for any trends & correlations. Used 'skim_without_charts' commands also for better view of tables.

```{r eval=FALSE, include=FALSE}
d_activity %>%
  select(TotalSteps, Calories,
         VeryActiveMinutes, FairlyActiveMinutes, LightlyActiveMinutes, SedentaryMinutes)%>%
  drop_na()%>%
  summary()

skim_without_charts(d_activity)
skim_without_charts(weight_info)
~~~

```{r eval=FALSE, include=FALSE} skim_without_charts(sleep_days)
```

~~~
Observations from Summary:
1. SedentaryMinutes Average is 991.2 minutes that are around 16.5 hrs of the da
y.
2. For VeryActiveMinutes, mean is only 21.16 minutes.

So it seems people are not using the product to its full potential to help them
 for living a healthy life. Other reason could be that people are not aware abou
t all feature of the products.

### 5.2 Used Spreadsheets for Filtering & Sorting

Before analyzing data in R, I used Excel for it. I tried to filter the data & ch
ecked the recording for various variable that is zero i.e customer has recorded
 it zero.

Observation from Spreadsheets
 1. Noticed that in d_activty table, 'loggedActiveDistance' is 0 for 908 entries
out of total 940 enteries.
 2. Similarly, 'VeryActiveDistance' & 'VeryActiveMinutes' are zero for 413 & 409
enteries respectively out of 940.

###5.3 Creating custom tables for trend & correaltion

First, I checked people with most calories so that I can track their sleep patte
rn & their activity to know trends & relations between calories, Total distance
 & sleep hours. During this process, created various tables by using JOINS in SQ
L.
Reason for new tables for better visualization later in 'Share' step of analysi
s.
```{r eval=FALSE, include=FALSE}

total_calories_id <- sqldf("SELECT Id, SUM(Calories) As Total_Cal
                           FROM d_activity
                           GROUP BY Id ")
max_calories_id <-
  sqldf ("SELECT * FROM total_calories_id ORDER BY Total_cal DESC")
~~~

```
max_distance_Id <- sqldf("SELECT DISTINCT Id AS Unique_id, SUM(TotalDistance)
                         AS total_distance FROM d_activity GROUP BY Id")

trend_dis_cal <-sqldf("SELECT * FROM max_distance_Id
      JOIN max_calories_id ON
      max_distance_Id.unique_id=max_calories_id.Id")



total_sleep_id <-sqldf("SELECT DISTINCT Id, SUM(TotalMinutesAsleep/60) AS total_
sleep_hrs
      FROM sleep_days
      GROUP BY Id
       ORDER BY total_sleep_hrs DESC")

#merging the related data in one table-distance,calories & sleep time

merged_data <- merge(total_sleep_id,trend_dis_cal,by="Id")
View(merged_data)

sqldf("SELECT DISTINCT(Id) FROM sleep_days")
#now it shows its only 24 people for sleep, so data is not complete for 33 peopl
e.
#lets compare very active distance & sedentary active distance

sed_active_dis <- sqldf("SELECT Id, SUM(VeryActiveDistance), SUM(SedentaryActive
Distance)
      FROM d_activity
      GROUP BY Id
      ORDER BY VeryActiveDistance DESC")
View(sed_active_dis)
```

```{r eval=FALSE, include=FALSE} ggplot(sleep_days, aes(x=TotalMinutesAsleep)) + geom_histogram(aes(y=..density..), binwidth=50,alpha=0.2)+ geom_density(alpha=0.2, fill="#FF0000") + geom_vline(aes(xintercept=mean(TotalMinutesAsleep, na.rm=T)), color="blue", linetype="dashed")+ labs(title="Total Minutes Asleep", x= "Total Minutes Asleep", y="Density")
```

~~~
The data shows that the sleep time is normally distributed among all participants. The average sleep time is 418 minutes, which means on average, participants have adequate sleep.

```{r}
ggplot(d_activity, aes(x=SedentaryMinutes)) +
  geom_histogram(aes(y=..density..), binwidth=100,alpha=0.2)+
  geom_density(alpha=0.2, fill="red") +
  geom_vline(aes(xintercept=mean(SedentaryMinutes, na.rm=T)), color="red", linetype="longdash")+
  labs(title="Sedentary Minutes Distribution", x= "Sedentary Minutes", y="Density")
~~~

As mentioned earlier also, Graph shows SedentaryMinutes Average is 991.2 minutes that are around 16.5 hrs of the day. This is be reduced for better results.

```
{r}
ggplot(d_activity, aes(x = VeryActiveMinutes, y = Calories))+
  geom_point()+
  geom_smooth()+
  labs(title="Daily Activity vs. Calories Burned", x= "Active Minutes", y="Calories Burned")+
  theme(plot.title = element_text(size=12), text = element_text(size=10))
```

Graph shows clearly that more active minutes lead to more burned calories.

6. Share

Now, data analysis is completed and its time to share observations & putting forward some recommendations.

Few Observations

- People using product is not logging all of the data due to which trends seen in data can be not so accurate.
- Around 16.5 Hrs were seen for sedentary activities that should be reduce for better results.
- Weight information was not enough to conduct any analysis.
- Clearly graphs shows that more activity leads to more burning of calories.
- Sleeping minutes average is 418 which is almost 7 hours of good sleep.

7. Act

So now when I have shared the observation, now its time to share some strategies that will help marketing team of Bellabeat to imporve their sales as well as better experience for their customers.

1. Notification for the weight logging. As I came across that people have not logged the weight, if the app have some feature of sending notification for logging weight information after the activity for the day is done, that might help to gather more data related to weight.
2. Sending daily motivational message to customer in app will motivate them to do more activity & eventually loose calories. This will also help to reduce sedentary minutes that are too high as per given data.
3. Monthly Challenges is another recommendation as this will boost their morale to complete the challenge & get better results