Bellabeat Case Study

Rajadurga

9/27/2021

Bellabeat is a high-tech manufacturer of health focused products for women. This case study involves analyzing the smart device data to gain insight into how customers are using the smart devices.

Stakeholders: - Urška Sršen: Bellabeat's cofounder and Chief Creative Officer - Sando Mur: Mathematician and Bellabeat's cofounder; key member of the Bellabeat executive team - Bellabeat marketing analytics team : A team of data analysts

Business Task: To identify the trends in the non Bellabeat smart device usage and to use those insights to inform Bellabeat marketing strategy.

Data source Description: This Kaggle data set contains personal fitness tracker from thirty fitbit users. The 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.

Installing and loading the packages

I will be analyzing this dataset using R studio. The packages needed for this analysis are the tidyverse and the janitor. Lets install and load the necessary packages.

```
install.packages("tidyverse")
install.packages("janitor")
library(tidyverse)
library(janitor)
```

Importing the dataset into R studio

The dataset is in .csv format. There are many files containing information about daily activity, steps, heart rate, calories and sleep in this dataset. For the analysis ,the dailyActivity_merged and sleepDay_merged datasets are used.

Importing the dailyActivity_merged and sleepDay_merged files into a dataframe.

```
daily_activity <- read.csv("dailyActivity_merged.csv")
sleep <- read.csv("sleepDay_merged.csv")
```

Lets view some data from daily_activity dataframe to have a glimpse of what details are available, like how many columns and rows are there and what are the datatypes of the variables.

```
head(daily_activity)
glimpse(daily_activity)
colnames(daily_activity)
```

Lets check out the sleep dataframe to know details about its rows and columns.

```
head(sleep)
glimpse(sleep)
colnames(sleep)
```

From the initial observation, the date column in both the dataframes are found to be of string datatype. That can be converted to date datatype in the data cleaning process.

Data Cleaning

Removing duplicates:

In order to clean the dataframe, lets identify and remove the duplicate records. First the duplicate records are identified and the row count is taken .

```
get_dupes(sleep)
nrow(sleep)
```

Once it is observed that the sleep dataframe has 3 duplicate records, they can be removed. After removing them the row count of the sleep dataframe is again taken to ensure that the duplicate records have been removed. The resulting records are stored in sleep_df dataframe for further analysis.

```
distinct(sleep)
sleep_df <- distinct(sleep)
```

Checking for duplicates in daily_activity dataframe.

```
get_dupes(daily_activity)
```

There are no duplicate records in daily_activity dataframe.

Removing empty rows:

The dataset may contain some empty rows when imported from excel which can be removed using the janitor package. Both the dataframes are checked for empty rows and if present empty rows are removed.

```
remove_empty(sleep_df,c("rows"))
remove_empty(daily_activity,c("rows"))
```

Changing the date format:

Date is in string format in both the daily_activity and sleep data frames. Lets convert it into proper date format. First lets load the lubridate package for date conversion. A new column is created with the proper date format.

```
library(lubridate)

daily_activity_final <- daily_activity %>%
       mutate(activity_date = mdy(ActivityDate)) %>%
       select(-ActivityDate)

head(daily_activity_final)
```

A new column is created in sleep_activity dataframe to hold the proper date format.

```
sleep_activity <- sleep_df %>%
        separate(SleepDay, into = c("sleep_date","sleep_time"), sep = " ")

sleep_cleaned <- sleep_activity %>%
        mutate(sleepdate = mdy(sleep_date)) %>%
        select(-sleep_date)
```

```
head(sleep_cleaned)
```

Analyzing and Visualizing the data

In order to analyze the sleep data, I created a column TotalHoursAsleep to know the time slept in hours instead of minutes and rounded it to 2 digits for convenience.

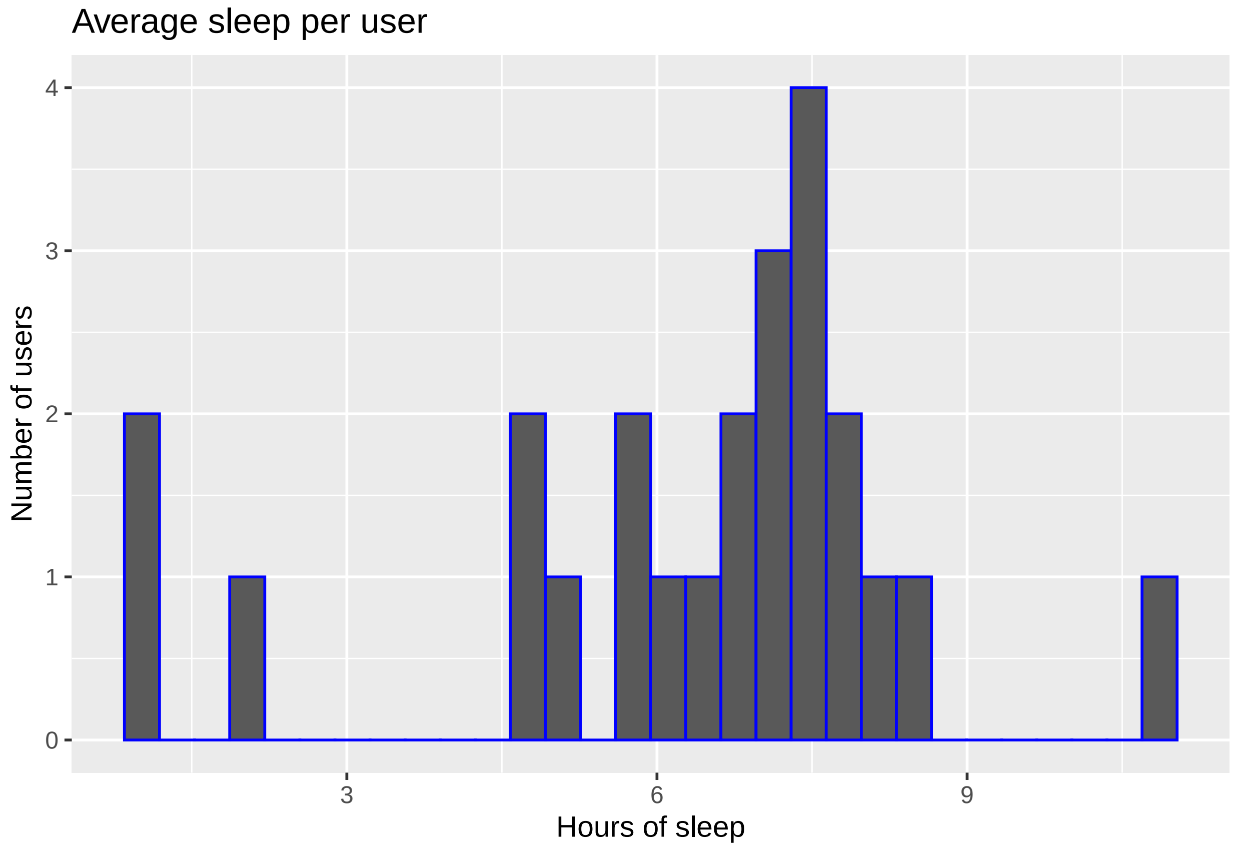
```
sleep_final <- sleep_cleaned %>%
        mutate(TotalHoursAsleep = round((TotalMinutesAsleep/60),2))
```

I wanted to see how much sleep a user gets on average. Average recommended sleep as per CDC is 7 hours .

```
sleep_graph <- sleep_final %>%
        select(Id,TotalHoursAsleep) %>%
        group_by(Id) %>%
        summarize(average_sleep = mean(TotalHoursAsleep))
head(sleep_graph)
```

Lets visualize it to get a clear picture.

```
ggplot(data = sleep_graph) +
        geom_histogram(mapping = aes(x=average_sleep), color = "blue") +
        labs(title = "Average sleep per user", y ="Number of users",
             x= "Hours of sleep")
```



Graph shows that several users dont get the recommended amount of sleep. Some users get dangerously low amount of sleep like,less than 3 hours, which should be definitely addressed.

Since both dataframes have column ‘Id’ in common,lets find out how many distinct Id’s are there in both dataframes

```
distinct(sleep,Id)
distinct(daily_activity,Id)
```

From this we can infer that sleep dataframe has 24 id's and daily__activity has 33 ids which may be because some of the users did not track the sleep. For further analysis both the dataframes has to be merged so we have the details about activity and sleep in a single dataframe. When merged we will have the Id's that are common in both dataframes.

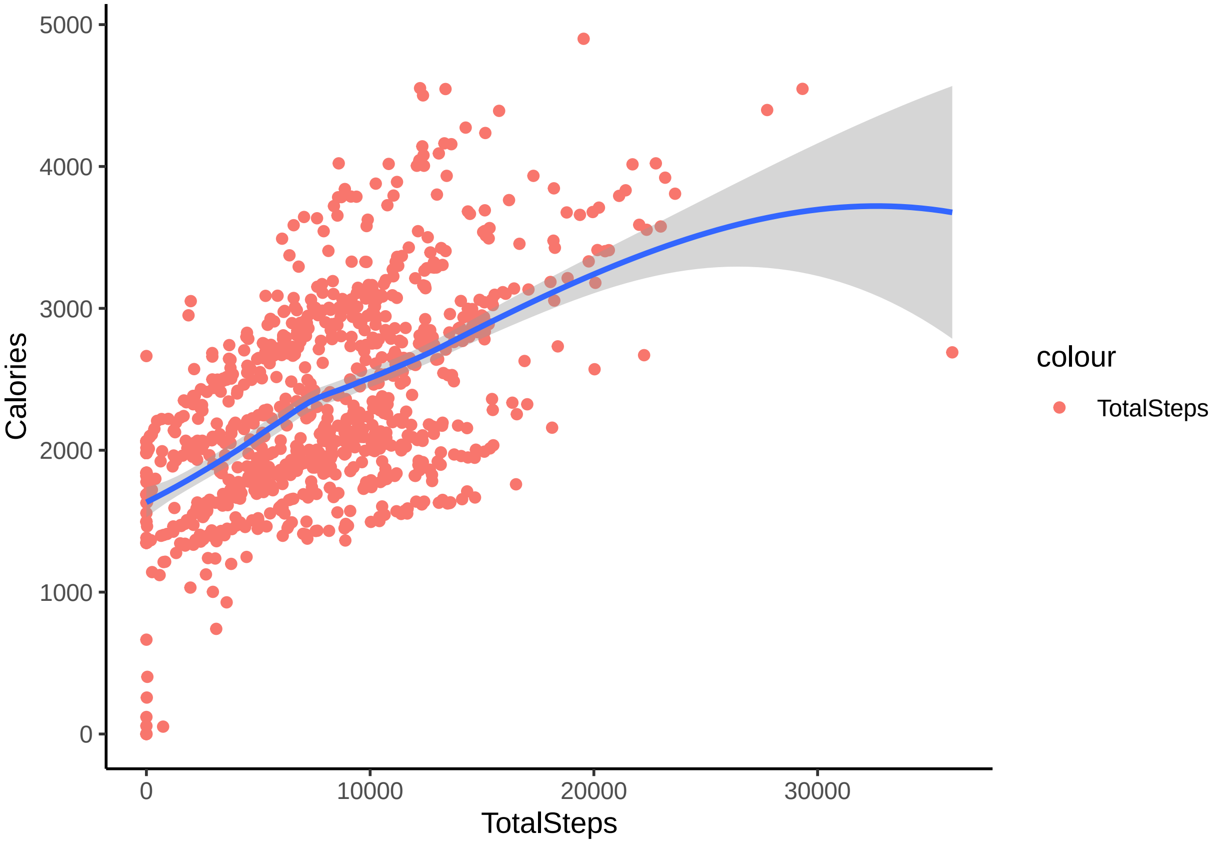
I wanted to see the relation between calories burned and the total steps taken by the user.

```
daily_activity_graph <- daily_activity_final %>%
        select(Id,activity_date,TotalSteps,Calories,SedentaryMinutes,
               TotalDistance) %>%
        group_by(Id) %>%
        summarize(avg_calories = mean(Calories),avg_steps = mean(TotalSteps),avg_sedentary_min = mean(S
```

we can see that more the steps taken more the calories burnt. But even though the users have taken more steps the sedentary minutes is still high. I assume may be that is due to the nature of the work .

Lets see the relation between steps and calorie through visualization.

```
ggplot(data = daily_activity_final) +
        geom_point(mapping = aes(x = TotalSteps, y= Calories,color="TotalSteps")) +
        geom_smooth(mapping = aes(x=TotalSteps, y=Calories)) +
        theme_classic()
```



It shows that the calories burnt increases with the number of steps.They have a positive linear correlation.we can see that more the steps taken more the calories burnt. But even though the users have taken more steps the sedentary minutes is still high. I assume may be that is due to the nature of the work .

```
df <- merge(sleep_final, daily_activity_final, by= "Id")
head(df)
distinct(sleep_final,Id)
distinct(daily_activity_final,Id)
```

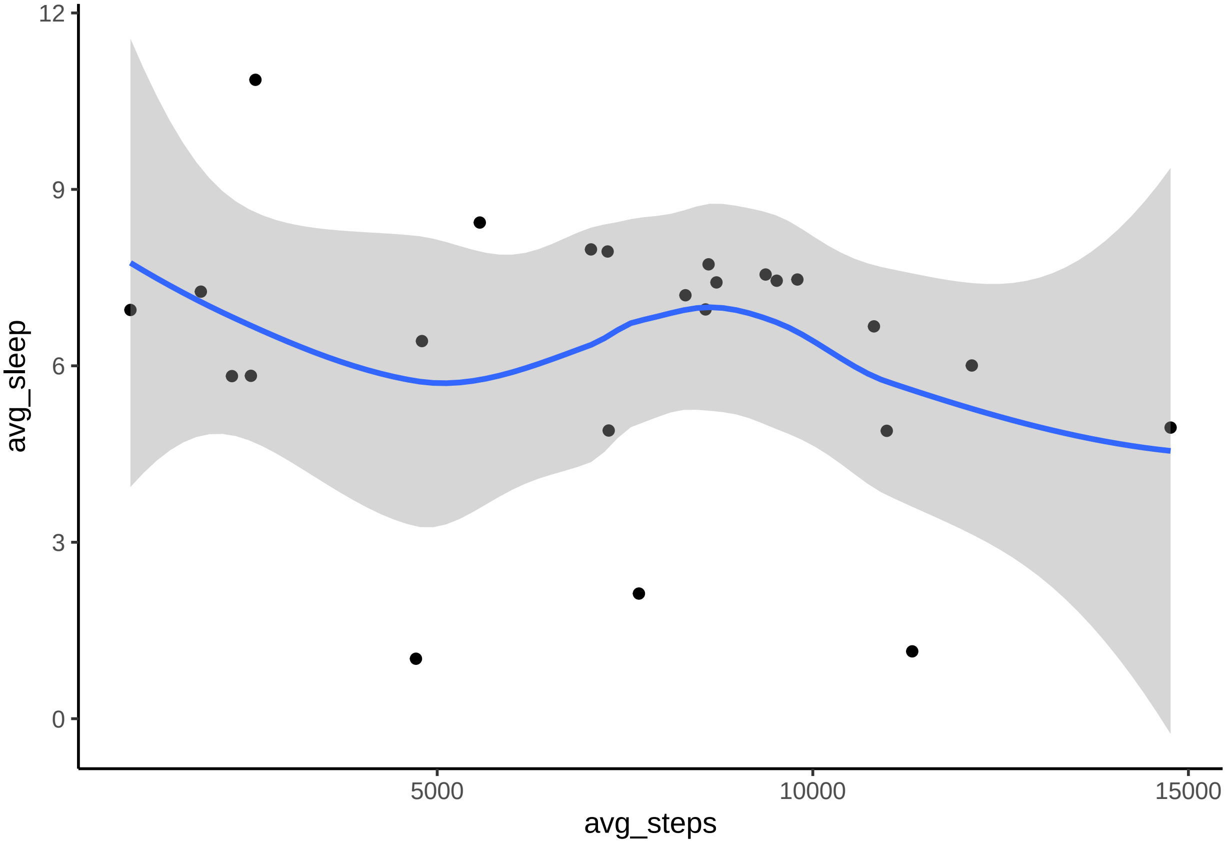
There are 33 distinct users in daily_activity_final dataframe and 24 distinct users in sleep_final dataframe. we want the records that are matching from both dataframes, so we group by Id's that are common in both the dataframes.

```
df_graph <-  df %>%
        select(Id,TotalHoursAsleep,TotalSteps) %>%
        group_by(Id) %>%
        summarize(avg_sleep = mean(TotalHoursAsleep),
                  avg_steps = mean(TotalSteps))

distinct(df_graph,Id)
```

I wanted to see if there is any relation between steps taken by the user and the time slept.

```
ggplot(data = df_graph,mapping = aes(x=avg_steps,
                                      y=avg_sleep),
       color="blue" ) +
        geom_point() +
        geom_smooth() +
        theme_classic()
```

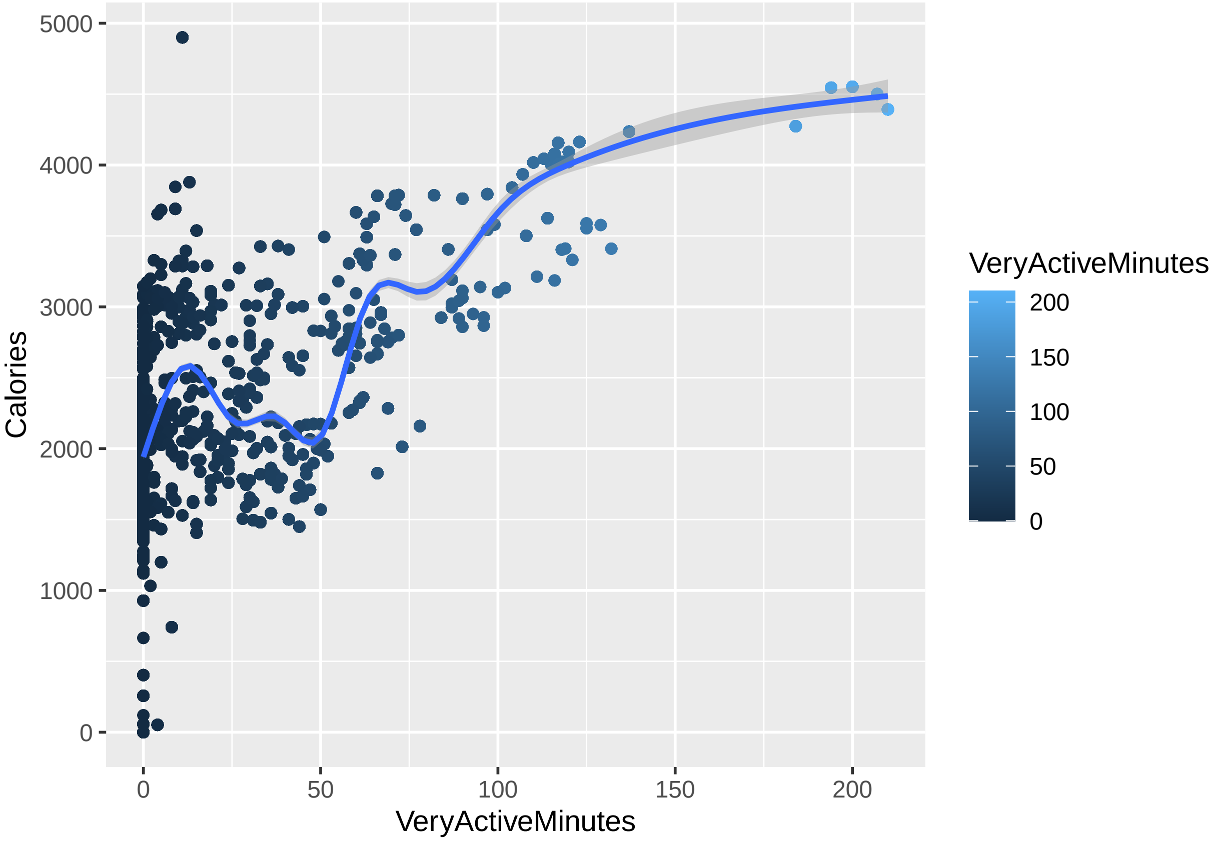


There is no correlation between time slept and the steps taken.

I wanted to find out if there is any relation between calories burnt and the very active zone minutes. So I created a dataframe conataining only the relevant columns.

```
calories <- df %>%
        select(Id,VeryActiveMinutes,Calories) %>%
        group_by(Id)
```

```
ggplot(data = calories, mapping = aes(x=VeryActiveMinutes,y=Calories,
                                      color=VeryActiveMinutes))+
        geom_point() +
  geom_smooth()
```



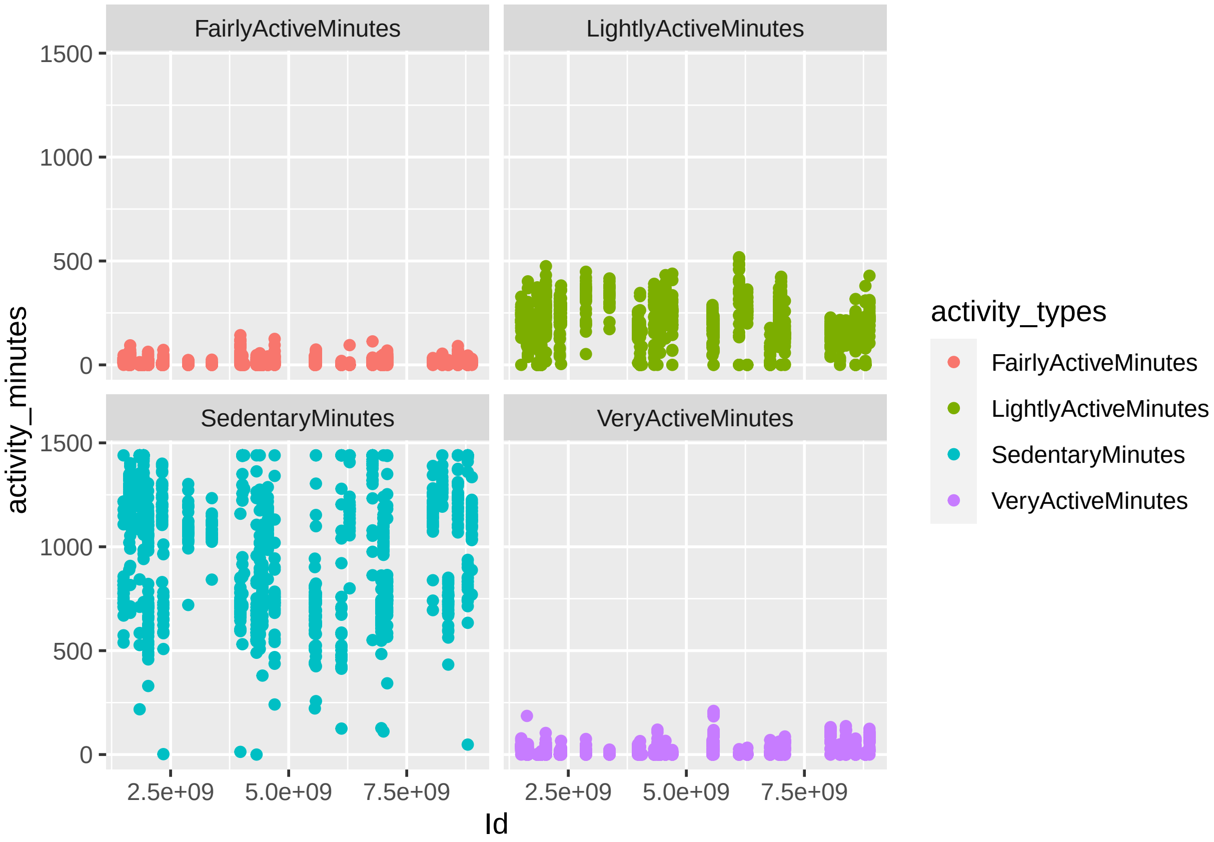
When the time spent in very active minutes is greater than 50 minutes then the calories burnt will be more. Even when time spent in Very active zone is less than 25 minutes the calories burnt is still more.

I wanted to find out how much time is spent in lightly active, fairly active, very active and sedentary zones.

```
active_minutes <- daily_activity %>%
        select(Id,VeryActiveMinutes,FairlyActiveMinutes,LightlyActiveMinutes,
               SedentaryMinutes)

activity_type <- active_minutes %>%
        pivot_longer(names_to = "activity_types", values_to = "activity_minutes",
                     VeryActiveMinutes:SedentaryMinutes)
```

```
ggplot(data = activity_type,mapping = aes(y=activity_minutes,x= Id,color = activity_types)) +
        geom_point() +
        facet_wrap(~activity_types)
```



From the visualization it is clear that the Sedentary minutes are very high compared to the active minutes for all the users.Even for users who have very active minutes, sedentary minutes are still high.I assume may be this is because we are active when we exercise and other times we will not be moving much, or may be the nature of the work is sedentary.

Recommendations:

After the analysis of the daily activity and the sleep files from Bellabeat dataset, I have identified some data, that Bellabeat can use to improve their customer experience and to design their marketing strategy.

- It has been observed that the number of users getting recommended amount of sleep is less. There are some users who get very little sleep which is less than 3 hours per day.Bellabeat can notify the users of their sleep trends in the weekly emails and remind them to sleep at 10pm daily night by gentle vibration so that the users will be encouraged to think about sleeping, 2 hours before their usual logged bedtime of 12am.
- There is a positive linear correlation between the steps taken and the calories burnt which means more the steps taken , more the calories burnt. So Bellabeat can encourage users to take more steps by giving them some reward points for crossing each 1000,5000 or 10,000 steps which will keep the users motivated.
- The amount of calories burnt is more, even when less than 25 minutes per day is spent on very active zone, which can be used to emphasize the importance of spending time on very active zones consistently by the user. Bellabeat can provide some digital badges to honor the time spent in very active zone.
- When analyzing the time spent in each zone minutes(Sedentary,Lightly active, Fairly active and very active),it has been noted that most of the time is spent in sedentary zone. When no steps has been taken for an hour by the user , then the Bellabeat device can nudge the users to mov,e by gentle vibration.

All these recommendations entice the user for a more healthier choice of lifestyle which Bellabeat can provide through their smart devices.