Bella beat analysis

Kelani

2022-09-02

Bella beat fitness app analysis

Scenario

You are a junior data analyst working on the marketing analyst team at Bellabeat, a high-tech manufacturer of health-focused products for women. Bellabeat is a successful small company, but they have the potential to become a larger player in the global 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. You 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 insights you discover will then help guide marketing strategy for the company. You will present your analysis to the Bellabeat executive team along with your high-level recommendations for Bellabeat's marketing strategy

```
  library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.7     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
  library(skimr)
  library(janitor)
```

```
##
## Attaching package: 'janitor'
##
## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

Start by importing all necessary data for analysis coming from an african country I notice that the date are in *MM/DD/YYYY* so i decided to transform the date to *DD/MM/YYYY* using excel and then I decided to import the *dailyActivity_merged.csv*, *hourlyCalories_merged.csv*, *hourlyIntensities_merged.csv*, *hourlySteps_merged.csv*, *sleepDay_merged.csv*,*weightLoginfo_merged.csv* why I chose these data is because they are the reason people love to purchase their smart devices. let's start by renaming each dataset

```r
daily_activity <- read_csv("E:/Fitabase Data/dailyActivity_merged.csv")
```

```
## Rows: 940 Columns: 15
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (1): ActivityDate
## dbl (14): Id, TotalSteps, TotalDistance, TrackerDistance, LoggedActivitiesDi...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
hourly_calories <- read_csv("E:/Fitabase Data/hourlyCalories_merged.csv")
```

```
## Rows: 22099 Columns: 5
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (2): ActivityHour, Date
## dbl  (2): Id, Calories
## time (1): Time
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
hourly_intensities <- read_csv("E:/Fitabase Data/hourlyIntensities_merged.csv")
```

```
## Rows: 22099 Columns: 6
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (2): ActivityHour, Date
## dbl  (3): Id, TotalIntensity, AverageIntensity
## time (1): Time
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
hourly_steps <- read_csv("E:/Fitabase Data/hourlySteps_merged.csv")
```

```
## Rows: 22099 Columns: 5
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (2): ActivityHour, Date
## dbl  (2): Id, StepTotal
## time (1): Time
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
sleep_day <- read_csv("E:/Fitabase Data/sleepDay_merged.csv")
```

```
## Rows: 413 Columns: 7
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (2): SleepDay, Date
## dbl (5): Id, TotalSleepRecords, TotalMinutesAsleep, TotalTimeInBed, Time
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
weight_log <- read_csv("E:/Fitabase Data/weightLoginfo_merged.csv")
```

```
## New names:
## Rows: 67 Columns: 10
## -- Column specification
## -------------------------------------------------------- Delimiter: "," chr
## (2): Date...2, Date...9 dbl (6): Id, WeightKg, WeightPounds, Fat, BMI, LogId
## lgl (1): IsManualReport time (1): Time
## i Use `spec()` to retrieve the full column specification for this data. i
## Specify the column types or set `show_col_types = FALSE` to quiet this message.
## * `Date` -> `Date...2`
## * `Date` -> `Date...9`
```

Checking the number of partipant in each data set Id

```
n_distinct(daily_activity$Id)
```

```
## [1] 33
```

```
n_distinct(hourly_calories$Id)
```

```
## [1] 33
```

```
n_distinct(hourly_intensities$Id)
```

```
## [1] 33
```

```
n_distinct(hourly_steps$Id)
```

```
## [1] 33
```

```
n_distinct(sleep_day$Id)
```

```
## [1] 24
```

```
n_distinct(weight_log$Id)
```

```
## [1] 8
```

I got the total number of each participants in the datasets daily_activities 33, hourly_calories 33, hourly_intensities 33, hourly_steps 33, sleep_day 24 and weigth_log 8 it goes without saying that 8 participants is not a proper sample size.

Taking the average of each total steps , total distance and sedentary active distance shows

```
daily_activity %>%
  summarise(avg_total_steps =mean(TotalSteps),
            avg_calories = mean(Calories),
            avg_Total_distance = mean(TotalDistance),
            avg_sedentaryMinutes = mean(SedentaryMinutes))
```

```
## # A tibble: 1 x 4
##   avg_total_steps avg_calories avg_Total_distance avg_sedentaryMinutes
##             <dbl>        <dbl>              <dbl>                <dbl>
## 1           7638.        2304.               5.49                 991.
```

```
# taking average time people sleep per day and average time people stay_in bed
sleep_day %>% summarise(avg_TotaltimeInBed = mean(TotalTimeInBed),
                        avg_Totaltimeasleep = mean(TotalMinutesAsleep))
```

```
## # A tibble: 1 x 2
##   avg_TotaltimeInBed avg_Totaltimeasleep
##                <dbl>               <dbl>
## 1               459.                419.
```

```
hourly_steps %>% summarise(avg_totalsteps = mean(StepTotal))
```

```
## # A tibble: 1 x 1
##   avg_totalsteps
##            <dbl>
## 1           320.
```

```
weight_log %>%
  summarise(avg_weight_kg = mean(WeightKg),
            avg_BMi = mean(BMI),
            avg_fat = mean(Fat, na.rm=TRUE))
```

```
## # A tibble: 1 x 3
##   avg_weight_kg avg_BMi avg_fat
##           <dbl>   <dbl>   <dbl>
## 1          72.0    25.2    23.5
```

- Average total steps of 7,638 is okay but it will be better if people can walk an average of 10,000 steps which is recommended by the WHO
- the average calory intake is 2304 which seems to not be a bad thing for people who are doing physical activity everyday while it is quite a lot id you are doing less physical activities the average a day should be 2000
- average sedentary time seems to be around 16 hours (991 minutes) whch is alot compared to 7-10 hours recommended by WHO

- Majority of people are lightly active
- The average total time in Bed may 7 hours (419 minute)also contribute to the reason why people average sedentary time is very high
- while average time asleep is 6 hours (419 minutes)

visualization

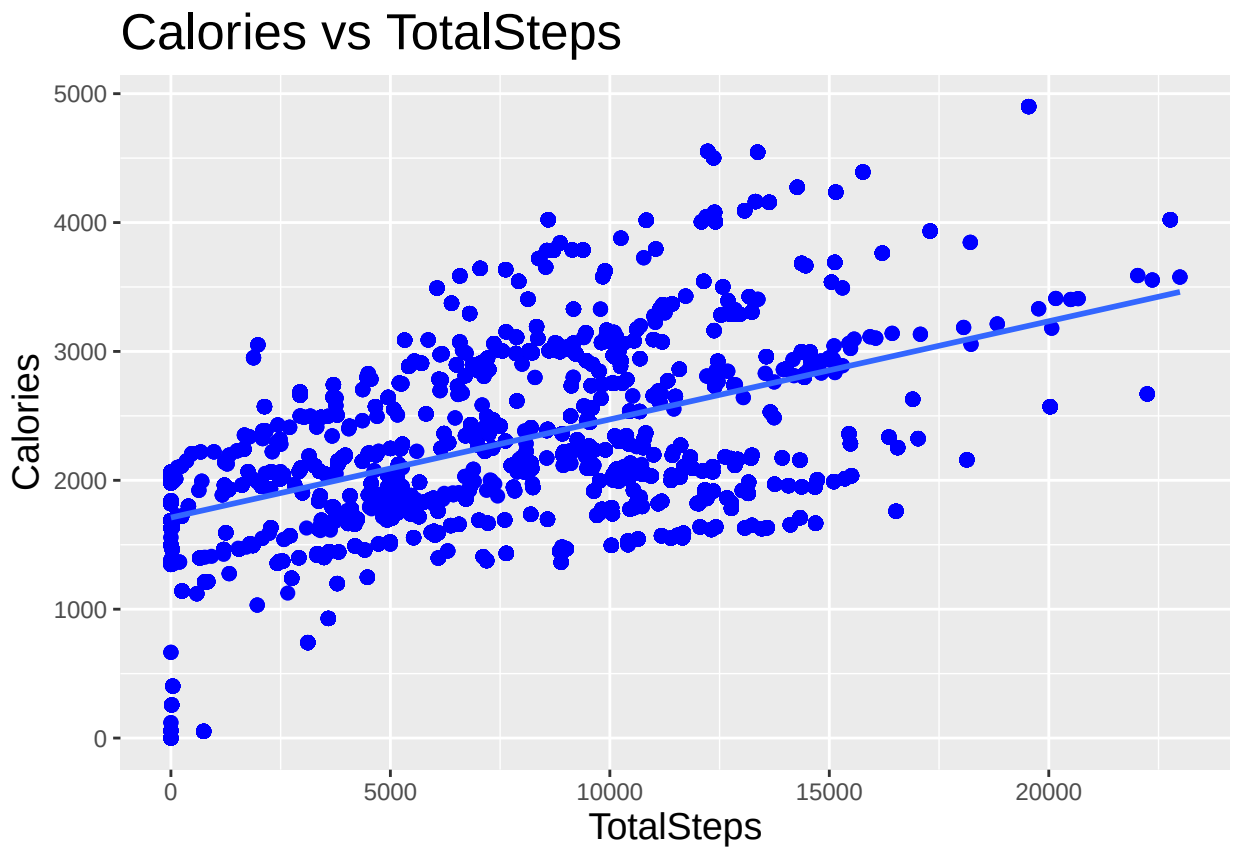
I think it is now time to visualize our data

let's start by merging our data

```
merged_data <-  merge(daily_activity, sleep_day, by=c('Id'))
```

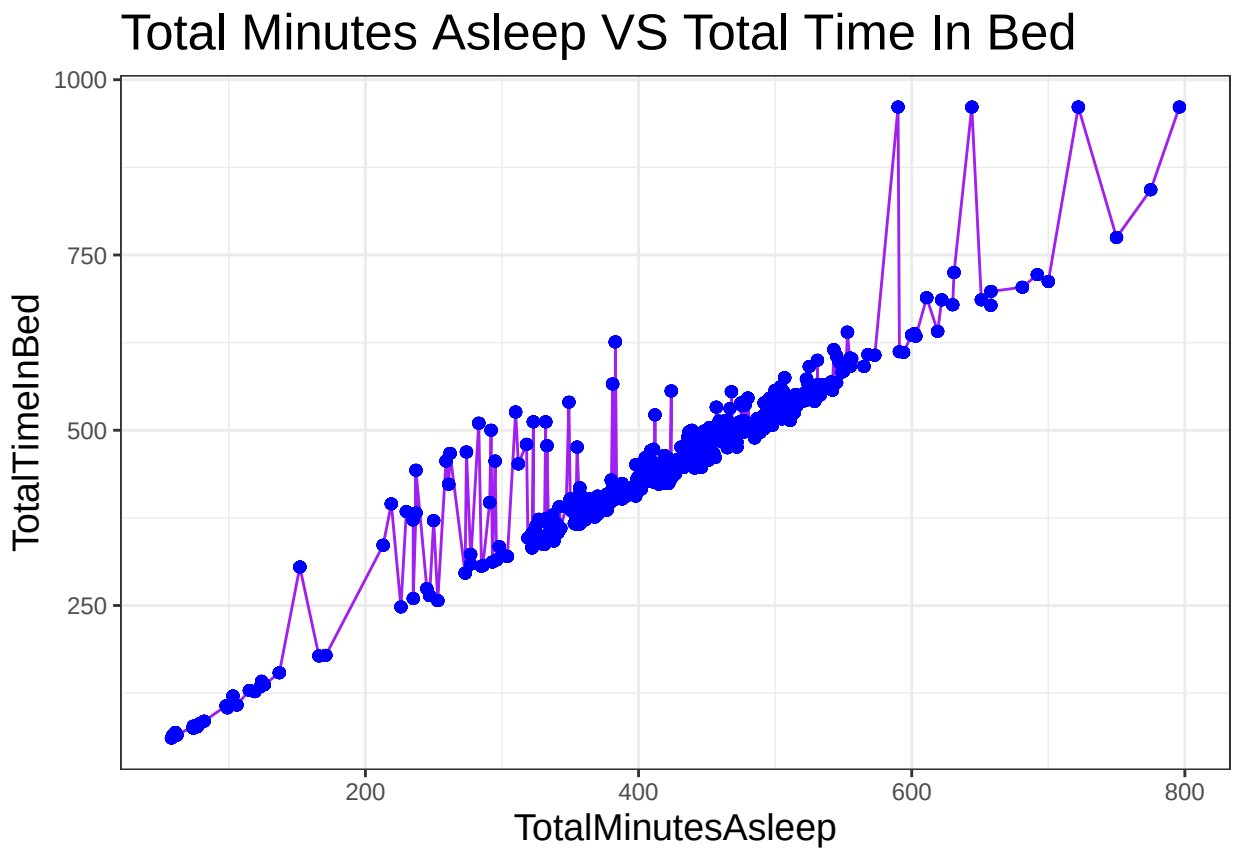
```
ggplot(data = merged_data, aes(x=TotalSteps, y=Calories)) +
  geom_point(color = "Blue", size= 2) +
  geom_smooth(method = lm, se = FALSE) +
  labs(title = "Calories vs TotalSteps",)+
  theme(axis.title.y = element_text(size = 14),
        axis.title.x = element_text(size = 14),
        title = element_text(size = 16))
```

```
## `geom_smooth()` using formula 'y ~ x'
```



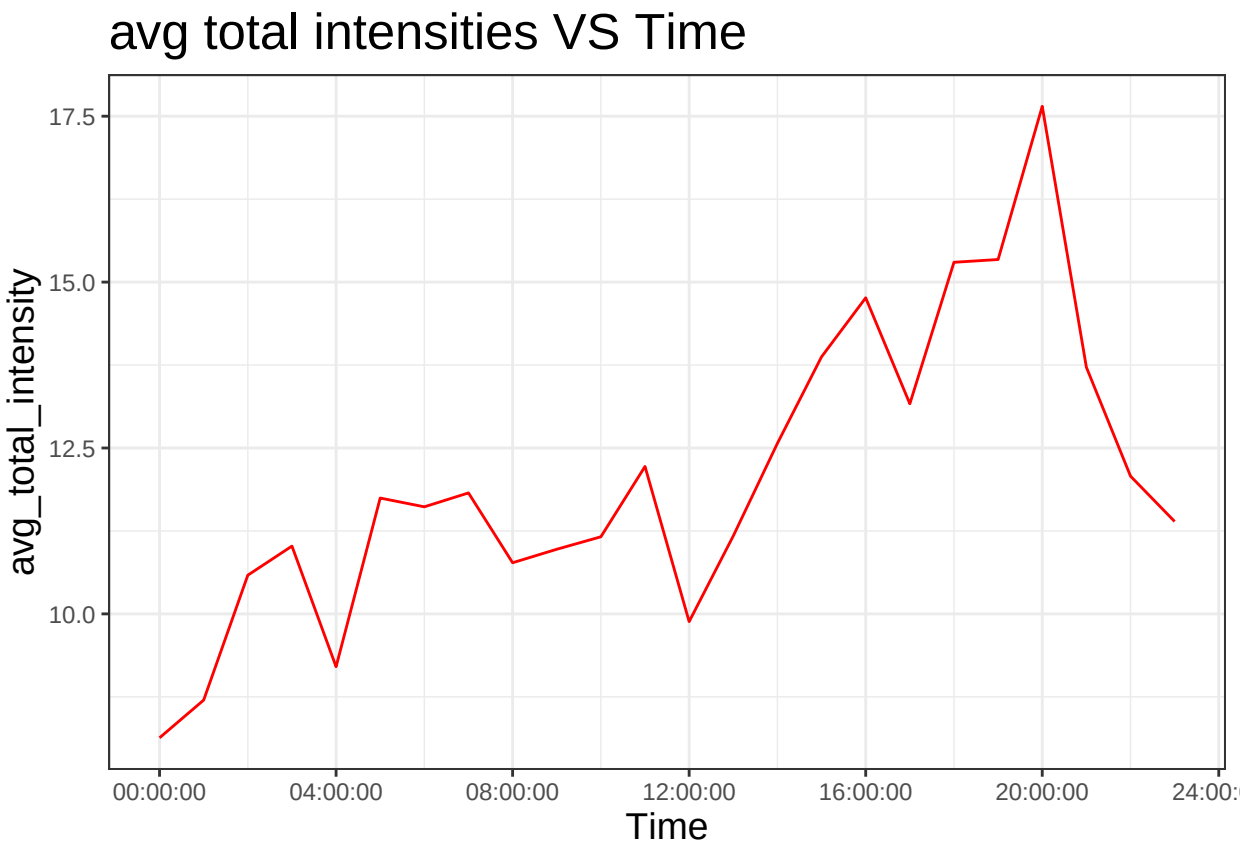
Here we could see a positive relationship between walking and calories which means the more you walk the more calories you burn

```
ggplot(data = merged_data, aes(x=TotalMinutesAsleep, y= TotalTimeInBed)) +
  geom_line(color = "purple")+
  geom_jitter(color = "Blue")+
  labs(title = "Total Minutes Asleep VS Total Time In Bed")+
  theme_bw()+
  theme(axis.title.x = element_text(size = 14),
        axis.title.y = element_text(size = 14),
        title = element_text(size= 16))
```



There's a linear trend in the total time spent sleeping compared to the total time spent on Bed this is to show that Bella beat can help it customer by notifying them on when to sleep and advising customers to sleep an average of 7 hours a day

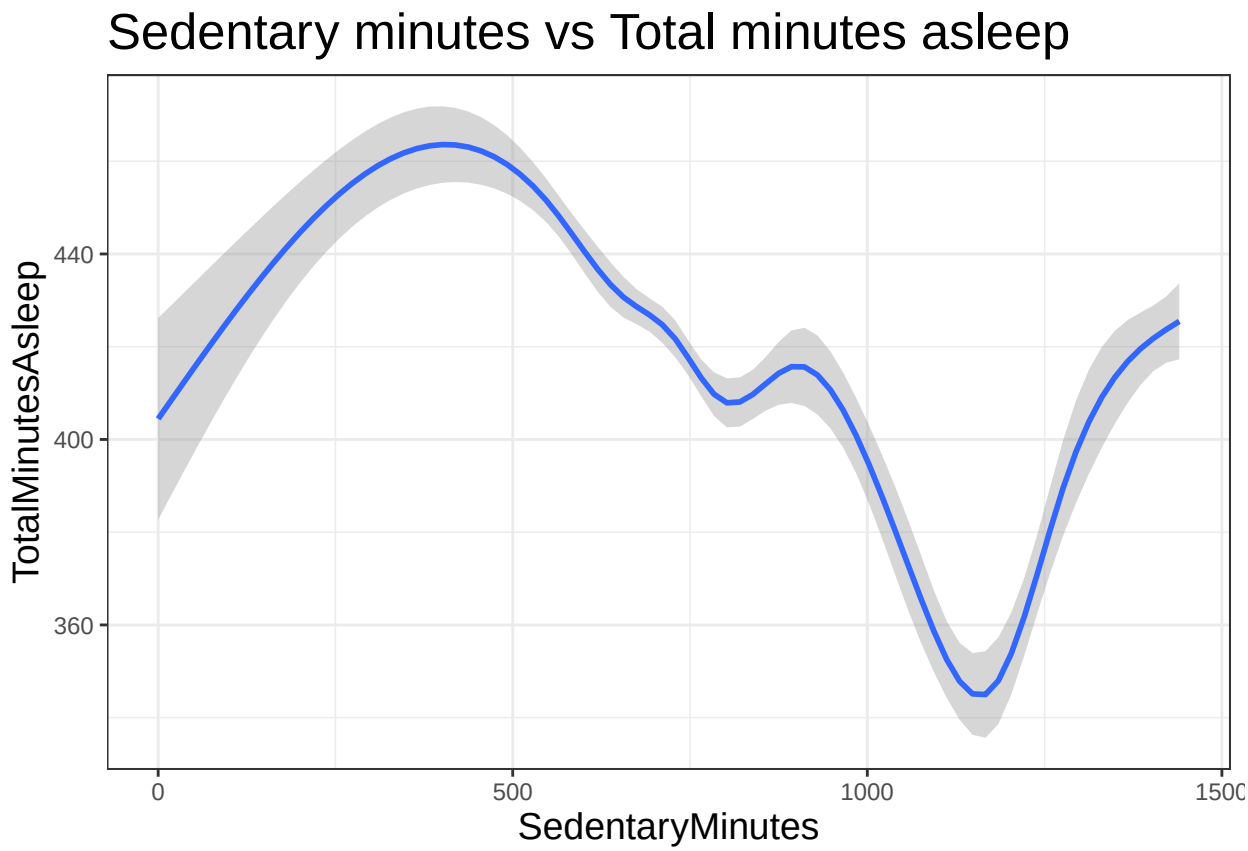
```
avg_hourly_intensities <- hourly_intensities %>%
  group_by(Time) %>%
  summarise(avg_total_intensity = mean(TotalIntensity))

ggplot(data = avg_hourly_intensities, aes(x= Time, y=avg_total_intensity)) +
  geom_line( color= "Red") +
  theme_bw()+
  labs(title = "avg total intensities VS Time")+
  theme(axis.title.x = element_text(size = 14),
        axis.title.y = element_text(size = 14),
        title = element_text(size= 16))
```



Here we see that people seems to be more active in the afternoon the graph shows us that from 12pm to 8 pm is when people become more intense which means they just have their office lunch time or are just coming back from work and wants to visit the gym aroound this time they could be notified to do some sort of fitness exercise to help them loose fatigue

```
ggplot(merged_data, aes(x=SedentaryMinutes, y= TotalMinutesAsleep)) +
  geom_smooth()+
  labs(title = "Sedentary minutes vs Total minutes asleep")+
  theme_bw()+
  theme(axis.title.x = element_text(size = 14),
        axis.title.y = element_text(size = 14),
        title = element_text(size= 16))
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



This graph made us understand that the less sedentary we are the more we sleep and also vise versa because the more we sleep the more energy we could gather Bella beat could create a notification for users to remind them on when to sleep

This are the insight that I was able to draw from the analysis * and also not to make mistake that correlation does not lead to causation*

Conslusion

In conclusion after analyzing the data I think that I have been able to Identify pattern that can help Bella beat in their marketing strategy

- Bella beat target audience are office people this means they spend alot of time wasted at work but are always active during office lunch time or 5 pm when they've close fro work up until 8 pm in the night
- We also found out that customers do more of light activity

To sum up my conclusion

- Bella beat can help improve customer sleep by notifying them on when to sleep and also advising them to take an adverage sleep of up to 7 hours
- Bella beat could help notify customers around 12 pm noon to take an exercise so as to reduce being sedent
- Bella beat could advise customer take an average of 10,000 walk per day so as to reduce fatigue and