bellabeat Case Study

Tushar Kumar

2025-06-21

## Installed and loaded common packages and libraries

options(repos = c(CRAN = "https://cloud.r-project.org"))  
install.packages('tidyverse')

## Installing package into 'C:/Users/s/AppData/Local/R/win-library/4.5'  
## (as 'lib' is unspecified)

## package 'tidyverse' successfully unpacked and MD5 sums checked  
##   
## The downloaded binary packages are in  
## C:\Users\s\AppData\Local\Temp\RtmpCa6mzZ\downloaded\_packages

library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.4 ✔ readr 2.1.5  
## ✔ forcats 1.0.0 ✔ stringr 1.5.1  
## ✔ ggplot2 3.5.2 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.4 ✔ tidyr 1.3.1  
## ✔ purrr 1.0.4

## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

## Loaded CSV files from the dataset

### Created a dataframe named ‘daily\_activity’, ‘sleep\_day’ and read in as the CSV files from the dataset.

daily\_activity <- read.csv("C:/Users/s/Desktop/Case Study/Portfolio Projects/bellabeat/Fitabase Data 4.12.16-5.12.16/dailyActivity\_merged.csv")

sleep\_day <- read.csv("C:/Users/s/Desktop/Case Study/Portfolio Projects/bellabeat/Fitabase Data 4.12.16-5.12.16/sleepDay\_merged.csv")

## Explored a few key tables

head(daily\_activity)

## Id ActivityDate TotalSteps TotalDistance TrackerDistance  
## 1 1503960366 4/12/2016 13162 8.50 8.50  
## 2 1503960366 4/13/2016 10735 6.97 6.97  
## 3 1503960366 4/14/2016 10460 6.74 6.74  
## 4 1503960366 4/15/2016 9762 6.28 6.28  
## 5 1503960366 4/16/2016 12669 8.16 8.16  
## 6 1503960366 4/17/2016 9705 6.48 6.48  
## LoggedActivitiesDistance VeryActiveDistance ModeratelyActiveDistance  
## 1 0 1.88 0.55  
## 2 0 1.57 0.69  
## 3 0 2.44 0.40  
## 4 0 2.14 1.26  
## 5 0 2.71 0.41  
## 6 0 3.19 0.78  
## LightActiveDistance SedentaryActiveDistance VeryActiveMinutes  
## 1 6.06 0 25  
## 2 4.71 0 21  
## 3 3.91 0 30  
## 4 2.83 0 29  
## 5 5.04 0 36  
## 6 2.51 0 38  
## FairlyActiveMinutes LightlyActiveMinutes SedentaryMinutes Calories  
## 1 13 328 728 1985  
## 2 19 217 776 1797  
## 3 11 181 1218 1776  
## 4 34 209 726 1745  
## 5 10 221 773 1863  
## 6 20 164 539 1728

colnames(daily\_activity)

## [1] "Id" "ActivityDate"   
## [3] "TotalSteps" "TotalDistance"   
## [5] "TrackerDistance" "LoggedActivitiesDistance"  
## [7] "VeryActiveDistance" "ModeratelyActiveDistance"  
## [9] "LightActiveDistance" "SedentaryActiveDistance"   
## [11] "VeryActiveMinutes" "FairlyActiveMinutes"   
## [13] "LightlyActiveMinutes" "SedentaryMinutes"   
## [15] "Calories"

head(sleep\_day)

## Id SleepDay TotalSleepRecords TotalMinutesAsleep  
## 1 1503960366 4/12/2016 12:00:00 AM 1 327  
## 2 1503960366 4/13/2016 12:00:00 AM 2 384  
## 3 1503960366 4/15/2016 12:00:00 AM 1 412  
## 4 1503960366 4/16/2016 12:00:00 AM 2 340  
## 5 1503960366 4/17/2016 12:00:00 AM 1 700  
## 6 1503960366 4/19/2016 12:00:00 AM 1 304  
## TotalTimeInBed  
## 1 346  
## 2 407  
## 3 442  
## 4 367  
## 5 712  
## 6 320

colnames(sleep\_day)

## [1] "Id" "SleepDay" "TotalSleepRecords"   
## [4] "TotalMinutesAsleep" "TotalTimeInBed"

#### Insight: Both datasets have Id which can be used to merge the datasets.

## 

## Summary statistics

### Number of unique participants are there in each dataframe?

n\_distinct(daily\_activity$Id)

## [1] 33

n\_distinct(sleep\_day$Id)

## [1] 24

### 

### Number of observations are there in each dataframe?

nrow(daily\_activity)

## [1] 940

nrow(sleep\_day)

## [1] 413

### Summary statistics about each Data Frame?

### For the daily\_activity data frame:

daily\_activity %>%   
 select(TotalSteps, TotalDistance, SedentaryMinutes) %>%   
 summary()

## TotalSteps TotalDistance SedentaryMinutes  
## Min. : 0 Min. : 0.000 Min. : 0.0   
## 1st Qu.: 3790 1st Qu.: 2.620 1st Qu.: 729.8   
## Median : 7406 Median : 5.245 Median :1057.5   
## Mean : 7638 Mean : 5.490 Mean : 991.2   
## 3rd Qu.:10727 3rd Qu.: 7.713 3rd Qu.:1229.5   
## Max. :36019 Max. :28.030 Max. :1440.0

### Insights from daily\_activity Data Frame:

#### 1. Half of the users walk fewer than ~7,400 steps per day — below the commonly recommended 10,000 steps.

#### 2. Distance traveled is closely correlated with steps. The average user travels ~5.5 km/day.

#### 3. Users are sedentary for most of the day. This highlights a health concern — long sedentary periods ~1,057 minutes (~17.6 hours)

### For the sleep\_day data frame:

sleep\_day %>%   
 select(TotalSleepRecords,  
 TotalMinutesAsleep,  
 TotalTimeInBed) %>%  
 summary()

## TotalSleepRecords TotalMinutesAsleep TotalTimeInBed   
## Min. :1.000 Min. : 58.0 Min. : 61.0   
## 1st Qu.:1.000 1st Qu.:361.0 1st Qu.:403.0   
## Median :1.000 Median :433.0 Median :463.0   
## Mean :1.119 Mean :419.5 Mean :458.6   
## 3rd Qu.:1.000 3rd Qu.:490.0 3rd Qu.:526.0   
## Max. :3.000 Max. :796.0 Max. :961.0

### Insights from sleep\_day dataset:

#### 1. Most users have only one sleep session per day, but a few may nap or track segmented sleep (up to 3 sessions).

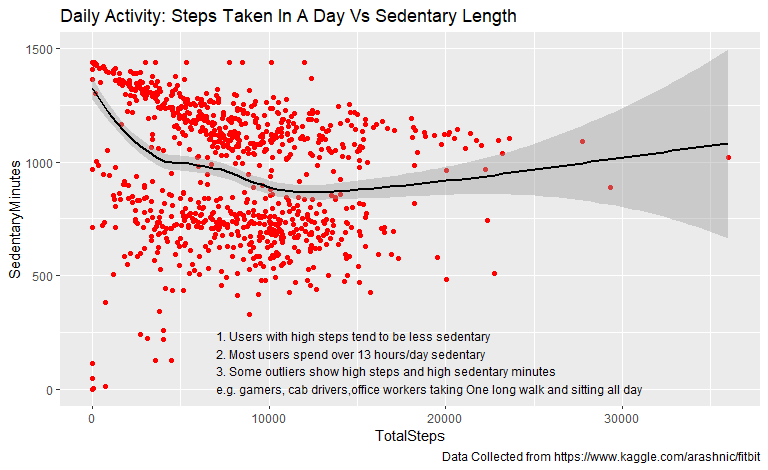
#### 2. Most users get around 419.5 minutes (~7 hours) of sleep, which aligns with health guidelines (but some sleep less than 1 hour — possible outliers).

#### 3. There’s a gap between time in bed and actual sleep, suggesting some time spent awake (e.g., ~30–40 min).

## Plotting a few explorations

### Relationship between steps taken in a day and sedentary minutes?

ggplot(data=daily\_activity, aes(x=TotalSteps, y=SedentaryMinutes)) +   
 geom\_point(colour = 'red') +  
 geom\_smooth(method = 'loess', formula = y~x, color='black') +  
 labs(title = "Daily Activity: Steps Taken In A Day Vs Sedentary Length ",  
 caption ="Data Collected from https://www.kaggle.com/arashnic/fitbit") +  
 annotate("text", x=7000, y=120, hjust=0,size=3.2,  
 label =paste0("1. Users with high steps tend to be less sedentary\n2. Most users spend over 13 hours/day sedentary\n3. Some outliers show high steps and high sedentary minutes\ne.g. gamers, cab drivers,office workers taking One long walk and sitting all day"))



### Insights:

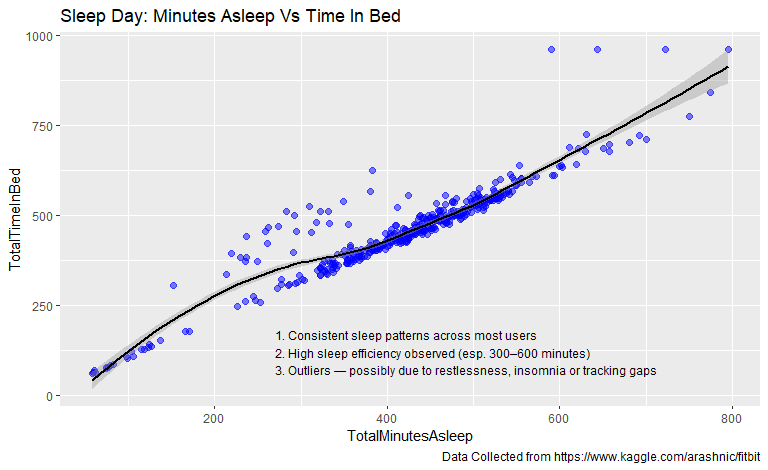
#### 1. Inverse Trend: More steps generally mean fewer sedentary minutes, especially up to ~12,000 steps.

#### 2. Sedentary Lifestyle: Most users remain inactive for over 13 hours/day.

#### 3. Outliers: Some users show high steps and high sedentary time—likely due to long walks followed by prolonged sitting (e.g., gamers, drivers, office workers).

### Relationship between minutes asleep and time in bed?

ggplot(data = sleep\_day, aes(x = TotalMinutesAsleep, y = TotalTimeInBed)) +   
 geom\_point(color = "blue", alpha = 0.5, size =2) +  
 geom\_smooth(method = 'loess', formula = y ~ x, color = "black") +  
 labs(  
 title = "Sleep Day: Minutes Asleep Vs Time In Bed",  
 caption = "Data Collected from https://www.kaggle.com/arashnic/fitbit"  
 ) +  
 annotate(  
 "text", x = 270, y = 120, hjust = 0, size = 3.2,  
 label = paste(  
 "1. Consistent sleep patterns across most users",  
 "2. High sleep efficiency observed (esp. 300–600 minutes)",  
 "3. Outliers — possibly due to restlessness, insomnia or tracking gaps", sep = "\n"  
 )  
 )



### Insights:

#### 1. Strong Positive Correlation - More time in bed generally results in more sleep.

#### 2. High Sleep Efficiency - Most users sleep efficiently, especially in the 300–600 minute range.

#### 3. Visible Outliers - Some users spend a long time in bed but sleep less — possibly due to restlessness, insomnia, or device tracking gaps.

## Merging these two datasets together

combined\_data <- merge(sleep\_day, daily\_activity, by="Id", all = TRUE)

### How many participants are in the combined data set and what attributes can be explored?

n\_distinct(combined\_data$Id)

## [1] 33

colnames(combined\_data)

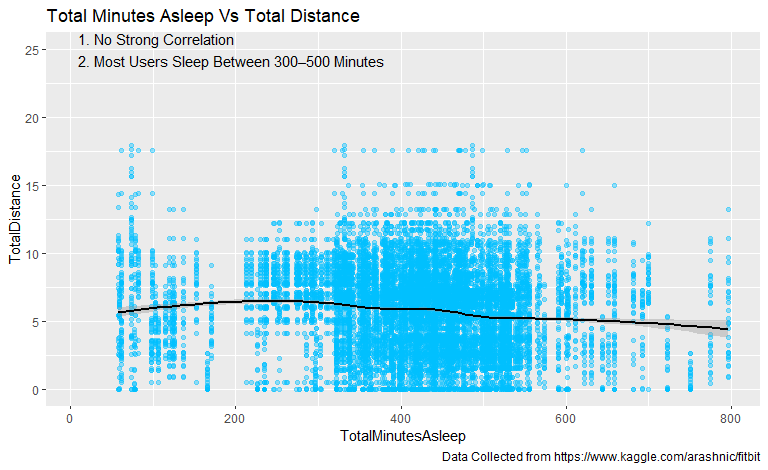
## [1] "Id" "SleepDay"   
## [3] "TotalSleepRecords" "TotalMinutesAsleep"   
## [5] "TotalTimeInBed" "ActivityDate"   
## [7] "TotalSteps" "TotalDistance"   
## [9] "TrackerDistance" "LoggedActivitiesDistance"  
## [11] "VeryActiveDistance" "ModeratelyActiveDistance"  
## [13] "LightActiveDistance" "SedentaryActiveDistance"   
## [15] "VeryActiveMinutes" "FairlyActiveMinutes"   
## [17] "LightlyActiveMinutes" "SedentaryMinutes"   
## [19] "Calories"

### 

### Do the participants who sleep more take more steps or fewer steps per day? Is there a relationship at all?

library(dplyr)  
clean\_combined\_data <- combined\_data %>%  
 filter(  
 !is.na(TotalMinutesAsleep),  
 !is.na(TotalTimeInBed),  
 is.finite(TotalMinutesAsleep),  
 is.finite(TotalTimeInBed)  
 )

ggplot(data = clean\_combined\_data, aes(x = TotalMinutesAsleep, y = TotalDistance)) +  
 geom\_point(color = "deepskyblue", alpha = 0.4) +  
 geom\_smooth(method = "loess", color = "black", formula = y~x) +  
 labs(title = "Total Minutes Asleep Vs Total Distance",  
 caption = "Data Collected from https://www.kaggle.com/arashnic/fitbit") +  
 annotate("text", x = 10, y = 25, hjust = 0, label =   
 "1. No Strong Correlation\n2. Most Users Sleep Between 300–500 Minutes")



### Insights:

#### 1. Low Sleep, High Distance - A few users covered long distances (10–20 km) with very little sleep (< 200 minutes). Could indicate highly active users with irregular sleep — e.g., shift workers or athletes.

#### 2.High Sleep, Low Distance - Some users slept over 600 minutes (10+ hours) but covered very little distance, suggesting sedentary behavior or rest days.

#### 3.Scattered Points Outside Core Cluster - These indicate inconsistent activity-sleep patterns that deviate from the typical 300–500 minute sleep + moderate activity range.