Bellabeat\_summary\_of\_analysis

Ikechukwu Ugbo

2024-01-04

# Summary of Analysis

## Setting up my environment

library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.4 ✔ readr 2.1.4  
## ✔ forcats 1.0.0 ✔ stringr 1.5.1  
## ✔ ggplot2 3.4.4 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.3 ✔ tidyr 1.3.0  
## ✔ purrr 1.0.2   
## ── 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

library(ggplot2)  
library(dplyr)  
library(skimr)  
library(lubridate)  
library(stringr)

## Importing the relevant spreadsheet files

dailyActivity = read\_csv("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.

sleepDay = read\_csv("sleepDay\_merged.csv")

## Rows: 413 Columns: 5  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (1): SleepDay  
## dbl (4): Id, TotalSleepRecords, TotalMinutesAsleep, TotalTimeInBed  
##   
## 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.

weightLog = read\_csv("weightLogInfo\_merged.csv")

## Rows: 67 Columns: 8  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (1): Date  
## dbl (6): Id, WeightKg, WeightPounds, Fat, BMI, LogId  
## lgl (1): IsManualReport  
##   
## 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.

WeightLog contains demographics data

## Cleaning: formatting data type: converting “character” date types to date

dailyActivity$ActivityDate = mdy(dailyActivity$ActivityDate)  
sleepDay = sleepDay %>%   
 separate(SleepDay, into=c('Date', 'Time'), sep=" ") %>%   
 select(-Time)  
weightLog = weightLog %>%   
 separate(Date, into=c('Date', 'Time'), sep=" ") %>%   
 select(-Time)  
sleepDay$Date = mdy(sleepDay$Date)  
weightLog$Date = mdy(weightLog$Date)

### Removing white spaces;

dailyActivity = dailyActivity %>%  
 mutate\_if(is.character, str\_trim)  
weightLog = weightLog %>%  
 mutate\_if(is.character, str\_trim)  
sleepDay = sleepDay %>%  
 mutate\_if(is.character, str\_trim)

## Data is ready for Analysis!

### Determination of the number of unique participants in each data frame

n\_distinct(dailyActivity$Id)

## [1] 33

n\_distinct(weightLog$Id)

## [1] 8

n\_distinct(sleepDay$Id)

## [1] 24

### Determination of the number of observations in each data frame

nrow(dailyActivity)

## [1] 940

nrow(weightLog)

## [1] 67

nrow(sleepDay)

## [1] 413

### Quick summary statistics

dailyActivity %>%   
 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

weightLog %>%   
 select(WeightKg,  
 BMI) %>%   
 summary()

## WeightKg BMI   
## Min. : 52.60 Min. :21.45   
## 1st Qu.: 61.40 1st Qu.:23.96   
## Median : 62.50 Median :24.39   
## Mean : 72.04 Mean :25.19   
## 3rd Qu.: 85.05 3rd Qu.:25.56   
## Max. :133.50 Max. :47.54

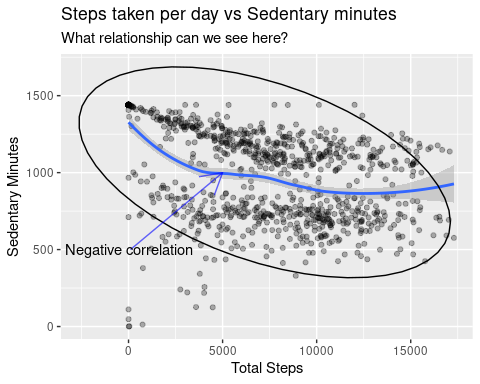
sleepDay %>%   
 select(TotalSleepRecords,   
 TotalMinutesAsleep,  
 TotalTimeInBed) %>%   
 summary()

## 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

### Relationship between steps taken per day and sedentary minutes

dailyActivity %>%   
 filter(TotalSteps < 17500) %>%   
 ggplot() +  
 geom\_point(mapping=aes(x = TotalSteps, y = SedentaryMinutes), alpha=0.3, position = position\_jitter()) +  
 geom\_smooth(mapping=aes(x = TotalSteps, y = SedentaryMinutes)) + stat\_ellipse(mapping=aes(x = TotalSteps, y = SedentaryMinutes)) +  
 labs(x = "Total Steps", y = "Sedentary Minutes", title="Steps taken per day vs Sedentary minutes", subtitle = "What relationship can we see here?") +  
 annotate("segment", x=100, xend=5000, y=500, yend=1000, color = "blue", alpha=0.6, arrow=arrow()) +  
 annotate("text", x=0, y=500, label = "Negative correlation") # There is a weak negative correlation between total steps taken and number of sedentary minutes

## `geom\_smooth()` using method = 'loess' and formula = 'y ~ x'



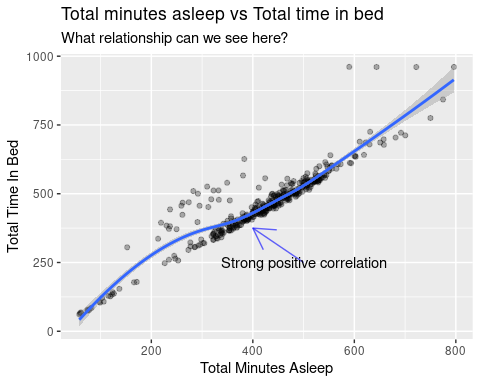
ggsave("steps\_taken\_vs\_sedentary\_minutes.png")

## Saving 5 x 4 in image  
## `geom\_smooth()` using method = 'loess' and formula = 'y ~ x'

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

ggplot(data = sleepDay) +  
 geom\_point(mapping=aes(x=TotalMinutesAsleep, y = TotalTimeInBed), alpha=0.3, position = position\_jitter()) +  
 geom\_smooth(mapping=aes(x=TotalMinutesAsleep, y = TotalTimeInBed)) +   
 labs(x = "Total Minutes Asleep", y = "Total Time In Bed", title="Total minutes asleep vs Total time in bed", subtitle = "What relationship can we see here?") +  
 annotate("segment", x=500, xend=400, y=250, yend=375, color = "blue", alpha=0.6, arrow=arrow()) +  
 annotate("text", x=500, y=250, label = "Strong positive correlation") # There is a strong positive correlation between the total number of minutes asleep and the total number of minutes spent in bed.

## `geom\_smooth()` using method = 'loess' and formula = 'y ~ x'



ggsave("totalMinutesAsleep\_vs\_totalTimeInBed.png")

## Saving 5 x 4 in image  
## `geom\_smooth()` using method = 'loess' and formula = 'y ~ x'

### Merging data sets together

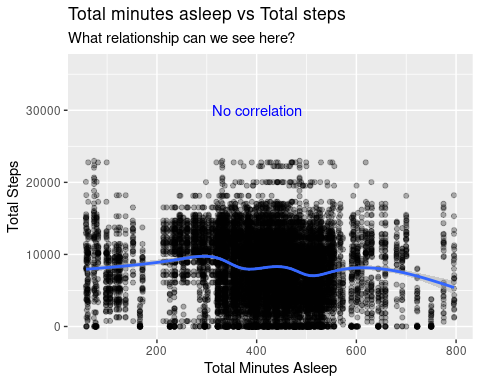
combined\_df = dailyActivity %>%   
 full\_join(sleepDay, by="Id")  
  
total\_df = combined\_df %>%   
 full\_join(weightLog, by="Id")  
  
n\_distinct(combined\_df$Id)

## [1] 33

### Relationship between minutes asleep and number of steps;

ggplot(data = combined\_df) +   
 geom\_point(mapping=aes(x=TotalMinutesAsleep, y = TotalSteps), alpha=0.3, position = position\_jitter()) +  
 geom\_smooth(mapping=aes(x=TotalMinutesAsleep, y = TotalSteps)) +  
 labs(x = "Total Minutes Asleep", y = "Total Steps", title="Total minutes asleep vs Total steps", subtitle = "What relationship can we see here?") +  
 annotate("text", x=400, y=30000, color = "blue", label = "No correlation") # There is almost no correlation between total number of minutes of sleep and the total steps taken per day.

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

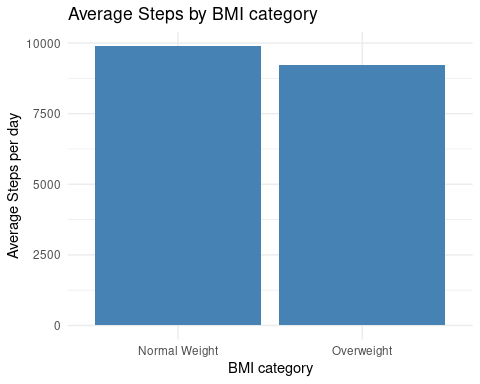


ggsave("totalMinutesAsleep\_vs\_totalSteps.png")

## Saving 5 x 4 in image  
## `geom\_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'

### Relationship between bmi and number of steps;

total\_df$bmi\_category = cut(total\_df$BMI, breaks = c(-Inf, 18.5, 24.9, 29.9, Inf), labels = c("Underweight", "Normal Weight", "Overweight", "Obese"), right = FALSE)  
average\_steps = total\_df %>%   
 group\_by(bmi\_category) %>%   
 summarise(Average\_steps = mean(TotalSteps))  
  
average\_steps = average\_steps %>%   
 filter(!is.na(bmi\_category)) %>%   
 filter(!bmi\_category == "Obese") # Observations for obesity were removed to ensure data fairness. Also, observation was just one and could be considered as an outlier.  
ggplot(average\_steps, aes(x=bmi\_category, y = Average\_steps), alpha=0.3, position = position\_jitter()) +  
 geom\_bar(stat = "identity", fill = "steelblue") +  
 labs(x = "BMI category", y = "Average Steps per day", title="Average Steps by BMI category") +  
 theme\_minimal() # There is almost no correlation between the BMI and the number of steps taken per day.



ggsave("average\_steps\_by\_bmi\_category.png")

## Saving 5 x 4 in image