outputs.R

tanma

2023-02-06

```
library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse
1.3.2 ──
## ✓ ggplot2 3.4.0      ✓ purrr   0.3.5
## ✓ tibble  3.1.8      ✓ dplyr   1.0.10
## ✓ tidyr   1.2.1      ✓ stringr 1.5.0
## ✓ readr   2.1.3      ✓ forcats 0.5.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()

library(janitor)

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

activity <- 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.

heartrate <- read_csv("heartrate_seconds_merged.csv")

## Rows: 2483658 Columns: 3
## ── Column specification
──────────────────────────────────────────────────────
## Delimiter: ","
## chr (1): Time
```

```
## dbl (2): Id, Value
##
## 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.

sleep <- 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.

n_distinct(activity$ActivityDate)

## [1] 31

n_distinct(sleep$SleepDay)

## [1] 31

n_distinct(heartrate$Time)

## [1] 961274

heartrate <- separate(heartrate, col="Time", into=c("Date", "Time"),
sep="\\s")

## Warning: Expected 2 pieces. Additional pieces discarded in 2483658 rows
[1, 2,
## 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, ...].

n_distinct(heartrate$Date)

## [1] 31

#Identifying number of users

n_distinct(activity$Id)

## [1] 33

n_distinct(sleep$Id)

## [1] 24

n_distinct(heartrate$Id)
```

```
## [1] 14
```

```
#Cleaning
```

```
activity %>% distinct()
```

```
## # A tibble: 940 × 15
##            Id Activity…¹ Total…² Total…³ Track…⁴ Logge…⁵ VeryA…⁶ Moder…⁷
Light…⁸
##         <dbl> <chr>        <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
<dbl>
##  1 1503960366 4/12/2016    13162    8.5     8.5        0    1.88   0.550
6.06
##  2 1503960366 4/13/2016    10735    6.97    6.97       0    1.57   0.690
4.71
##  3 1503960366 4/14/2016    10460    6.74    6.74       0    2.44   0.400
3.91
##  4 1503960366 4/15/2016     9762    6.28    6.28       0    2.14   1.26
2.83
##  5 1503960366 4/16/2016    12669    8.16    8.16       0    2.71   0.410
5.04
##  6 1503960366 4/17/2016     9705    6.48    6.48       0    3.19   0.780
2.51
##  7 1503960366 4/18/2016    13019    8.59    8.59       0    3.25   0.640
4.71
##  8 1503960366 4/19/2016    15506    9.88    9.88       0    3.53   1.32
5.03
##  9 1503960366 4/20/2016    10544    6.68    6.68       0    1.96   0.480
4.24
## 10 1503960366 4/21/2016     9819    6.34    6.34       0    1.34   0.350
4.65
## # … with 930 more rows, 6 more variables: SedentaryActiveDistance <dbl>,
## #   VeryActiveMinutes <dbl>, FairlyActiveMinutes <dbl>,
## #   LightlyActiveMinutes <dbl>, SedentaryMinutes <dbl>, Calories <dbl>,
and
## #   abbreviated variable names ¹ActivityDate, ²TotalSteps, ³TotalDistance,
## #   ⁴TrackerDistance, ⁵LoggedActivitiesDistance, ⁶VeryActiveDistance,
## #   ⁷ModeratelyActiveDistance, ⁸LightActiveDistance
```

```
sleep %>% distinct()
```

```
## # A tibble: 410 × 5
##            Id SleepDay              TotalSleepRecords TotalMinutesAsleep
Total…¹
##         <dbl> <chr>                             <dbl>              <dbl>
<dbl>
##  1 1503960366 04-12-2016 0.00                      1                327
346
##  2 1503960366 4/13/2016 12:00:00 AM                2                384
407
##  3 1503960366 4/15/2016 12:00:00 AM                1                412
```

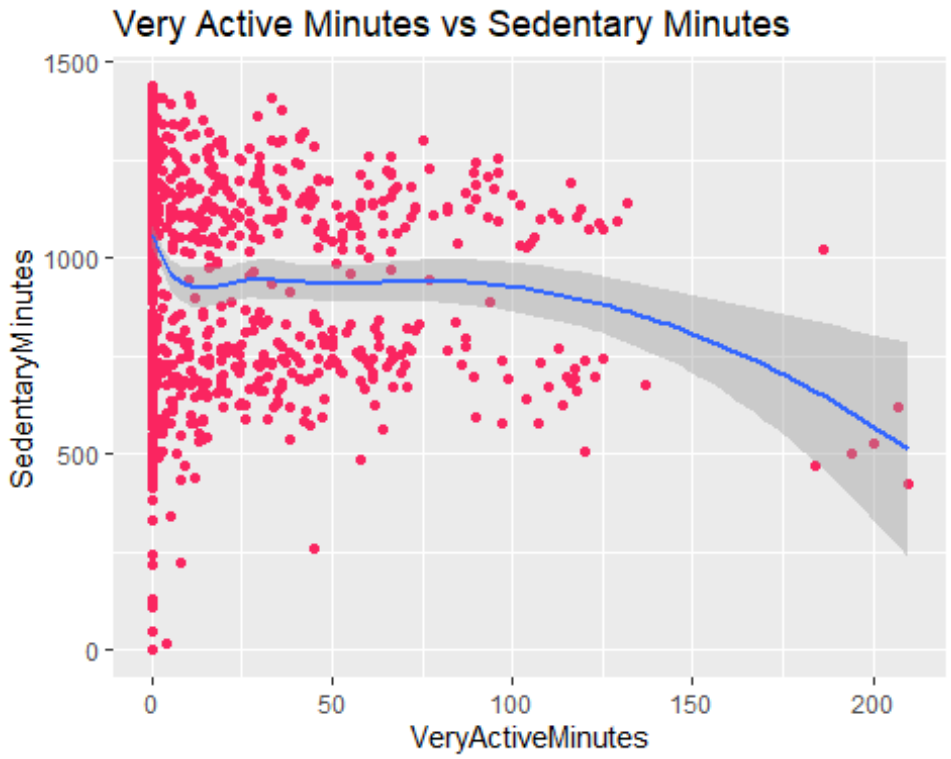
```
442
##  4 1503960366 4/16/2016 12:00:00 AM                2                340
367
##  5 1503960366 4/17/2016 12:00:00 AM                1                700
712
##  6 1503960366 4/19/2016 12:00:00 AM                1                304
320
##  7 1503960366 4/20/2016 12:00:00 AM                1                360
377
##  8 1503960366 4/21/2016 12:00:00 AM                1                325
364
##  9 1503960366 4/23/2016 12:00:00 AM                1                361
384
## 10 1503960366 4/24/2016 12:00:00 AM                1                430
449
## # … with 400 more rows, and abbreviated variable name ¹TotalTimeInBed
```

```
heartrate %>% distinct()
```

```
## # A tibble: 2,474,324 × 4
##            Id Date      Time    Value
##         <dbl> <chr>     <chr>   <dbl>
##  1 2022484408 4/12/2016 7:21:00    97
##  2 2022484408 4/12/2016 7:21:05   102
##  3 2022484408 4/12/2016 7:21:10   105
##  4 2022484408 4/12/2016 7:21:20   103
##  5 2022484408 4/12/2016 7:21:25   101
##  6 2022484408 4/12/2016 7:22:05    95
##  7 2022484408 4/12/2016 7:22:10    91
##  8 2022484408 4/12/2016 7:22:15    93
##  9 2022484408 4/12/2016 7:22:20    94
## 10 2022484408 4/12/2016 7:22:25    93
## # … with 2,474,314 more rows
```

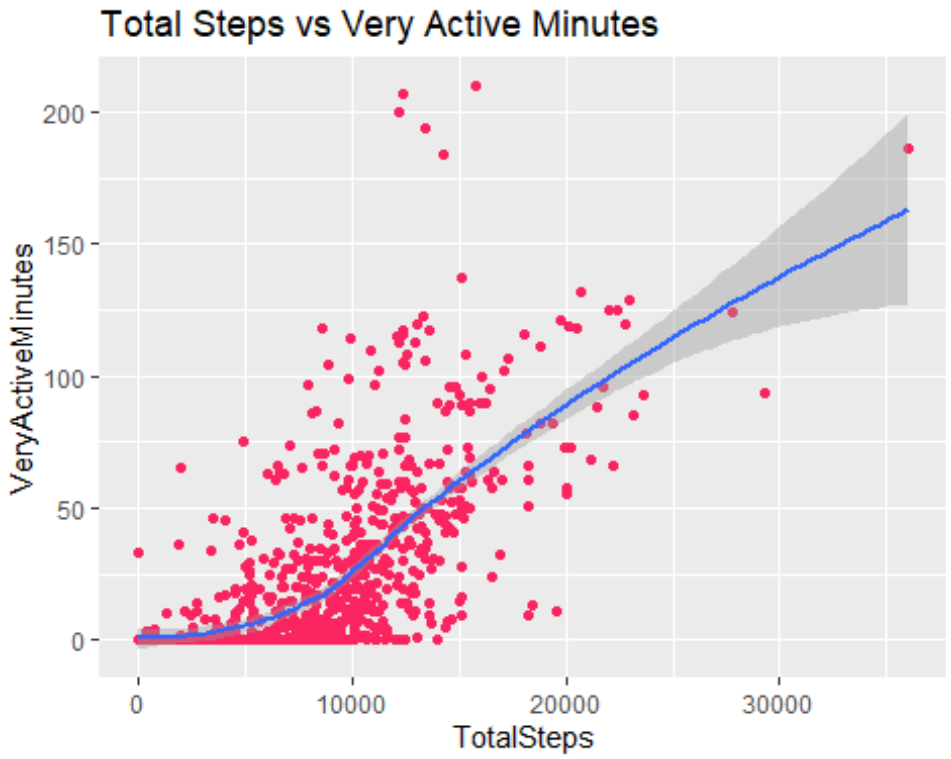
```
#plots
ggplot(activity, mapping = aes(x=VeryActiveMinutes, y=SedentaryMinutes)) +
  geom_point(color="#FA2560") + geom_smooth() +
  labs(title="Very Active Minutes vs Sedentary Minutes")
```

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



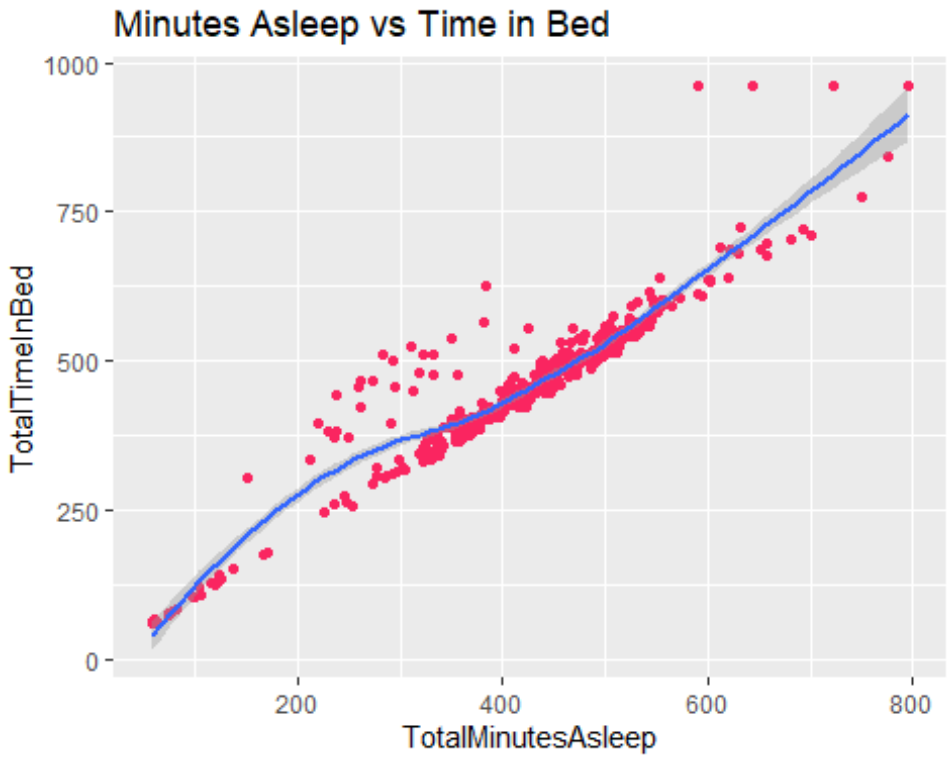

```
ggplot(activity, mapping = aes(x=TotalSteps, y=VeryActiveMinutes)) +
  geom_point(color="#FA2560") + geom_smooth() +
  labs(title="Total Steps vs Very Active Minutes")

## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```




```
ggplot(sleep, aes(x=TotalMinutesAsleep, y=TotalTimeInBed)) +
  geom_point(color="#FA2560") + geom_smooth() +
  labs(title="Minutes Asleep vs Time in Bed")

## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```


Minutes Asleep vs Time in Bed
TotalTimeInBed
1000
750
500
250
0
200
400
600
800
TotalMinutesAsleep