Cleaning and transformation.

2023-06-01

Data transformation and cleaning

First, we load the packages that will be used for importing, cleaning, transforming, and analyzing the data.

```
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✓ dplyr     1.1.1     ✓ readr     2.1.4
## ✓ forcats   1.0.0     ✓ stringr   1.5.0
## ✓ ggplot2   3.4.1     ✓ tibble    3.2.1
## ✓ lubridate 1.9.2     ✓ tidyr     1.3.0
## ✓ purrr     1.0.1
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the ]8;;http://conflicted.r-lib.org/conflicted package]8;; to force all conflict
s to become errors
```

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

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

```
library(here)
```

```
## here() starts at G:/Otros ordenadores/Mi Portátil/Gaspar Facultad/Analisis de datos/Caso p
ractico/R
```

```
library(chron)
```

```
##
## Attaching package: 'chron'
##
## The following objects are masked from 'package:lubridate':
##
##     days, hours, minutes, seconds, years
```

The databases are imported.

```
dailyActivity_merged <- read.csv("G:/Otros ordenadores/Mi Portátil/Gaspar Facultad/Analisis d
e datos/Caso practico/Base de datos/dailyActivity_merged.csv")
```

```
sleepDay_merge <- read.csv("G:/Otros ordenadores/Mi Portátil/Gaspar Facultad/Analisis de dato
s/Caso practico/Base de datos/sleepDay_merged.csv")
```

Selection of variables of interest.

dailyActivity_merge

The daily_activity data frame is created based on the selection of variables of interest from the dailyActivity_merged dataset.

```
daily_activity <- dailyActivity_merged %>%
  select(-TrackerDistance, -LoggedActivitiesDistance)
daily_activity <- daily_activity %>%
  select(-Calories)
```

The name of the FairlyActiveMinutes column is changed to ModeratelyActiveMinutes.

```
daily_activity <- daily_activity %>%
  rename(ModeratelyActiveMinutes = FairlyActiveMinutes)
```

dailySleep_merge

The data in the SleepDay column is separated into two columns: one for date and another for time.

```
sleep_day <- sleepDay_merge %>%
  separate(SleepDay, into = c("date", "hour"), " ")
```

```
## Warning: Expected 2 pieces. Additional pieces discarded in 413 rows [1, 2, 3, 4, 5, 6,
## 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, ...].
```

The time in the hour column is always the same. Therefore, we will only keep the date column.

```
sleep_day %>%
  select(hour) %>%
  filter(hour != "12:00:00")
```

```
## [1] hour
## <0 rows> (or 0-length row.names)
```

```
sleep_day <- sleep_day %>%
  select(-hour)
```

Data type transformation.

The data in the necessary columns is transformed into date data type.

```
daily_activity <- daily_activity %>%
  mutate(ActivityDate = as.Date(ActivityDate, format = "%m/%d/%Y" ))
```

```
sleep_day <- sleep_day %>%
  mutate(date = as.Date(date ,format = "%m/%d/%Y"))
```

Cleaning

The data outside the specified date range is searched for and removed.

```
daily_activity %>%
  filter(!between(ActivityDate, as.Date("2016-03-12"), as.Date("2016-05-12")))
```

```
##  [1] Id                       ActivityDate             TotalSteps
##  [4] TotalDistance            VeryActiveDistance       ModeratelyActiveDistance
##  [7] LightActiveDistance      SedentaryActiveDistance  VeryActiveMinutes
## [10] ModeratelyActiveMinutes  LightlyActiveMinutes     SedentaryMinutes
## <0 rows> (or 0-length row.names)
```

In the daily_activity database, the total activity sum is checked to ensure it is not greater than 24 hours, which is equivalent to 1440 minutes.

```
daily_activity %>%
  filter(VeryActiveMinutes > 1440 |
           ModeratelyActiveMinutes > 1440 |
           LightlyActiveMinutes > 1440 |
           SedentaryMinutes > 1440)
```

```
##  [1] Id                       ActivityDate             TotalSteps
##  [4] TotalDistance            VeryActiveDistance       ModeratelyActiveDistance
##  [7] LightActiveDistance      SedentaryActiveDistance  VeryActiveMinutes
## [10] ModeratelyActiveMinutes  LightlyActiveMinutes     SedentaryMinutes
## <0 rows> (or 0-length row.names)
```

Summary of the data.

```
glimpse(daily_activity)
```

```
## Rows: 940
## Columns: 12
## $ Id                       <dbl> 1503960366, 1503960366, 1503960366, 150396036…
## $ ActivityDate             <date> 2016-04-12, 2016-04-13, 2016-04-14, 2016-04-…
## $ TotalSteps               <int> 13162, 10735, 10460, 9762, 12669, 9705, 13019…
## $ TotalDistance            <dbl> 8.50, 6.97, 6.74, 6.28, 8.16, 6.48, 8.59, 9.8…
## $ VeryActiveDistance       <dbl> 1.88, 1.57, 2.44, 2.14, 2.71, 3.19, 3.25, 3.5…
## $ ModeratelyActiveDistance <dbl> 0.55, 0.69, 0.40, 1.26, 0.41, 0.78, 0.64, 1.3…
## $ LightActiveDistance      <dbl> 6.06, 4.71, 3.91, 2.83, 5.04, 2.51, 4.71, 5.0…
## $ SedentaryActiveDistance  <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …
## $ VeryActiveMinutes        <int> 25, 21, 30, 29, 36, 38, 42, 50, 28, 19, 66, 4…
## $ ModeratelyActiveMinutes  <int> 13, 19, 11, 34, 10, 20, 16, 31, 12, 8, 27, 21…
## $ LightlyActiveMinutes     <int> 328, 217, 181, 209, 221, 164, 233, 264, 205, …
## $ SedentaryMinutes         <int> 728, 776, 1218, 726, 773, 539, 1149, 775, 818…
```

```
head(daily_activity)
```

```
##           Id ActivityDate TotalSteps TotalDistance VeryActiveDistance
## 1 1503960366   2016-04-12      13162          8.50               1.88
## 2 1503960366   2016-04-13      10735          6.97               1.57
## 3 1503960366   2016-04-14      10460          6.74               2.44
## 4 1503960366   2016-04-15       9762          6.28               2.14
## 5 1503960366   2016-04-16      12669          8.16               2.71
## 6 1503960366   2016-04-17       9705          6.48               3.19
##   ModeratelyActiveDistance LightActiveDistance SedentaryActiveDistance
## 1                     0.55               6.06                        0
## 2                     0.69               4.71                        0
## 3                     0.40               3.91                        0
## 4                     1.26               2.83                        0
## 5                     0.41               5.04                        0
## 6                     0.78               2.51                        0
##   VeryActiveMinutes ModeratelyActiveMinutes LightlyActiveMinutes
## 1                25                      13                  328
## 2                21                      19                  217
## 3                30                      11                  181
## 4                29                      34                  209
## 5                36                      10                  221
## 6                38                      20                  164
##   SedentaryMinutes
## 1              728
## 2              776
## 3             1218
## 4              726
## 5              773
## 6              539
```

```
glimpse(sleep_day)
```

```
## Rows: 413
## Columns: 5
## $ Id                 <dbl> 1503960366, 1503960366, 1503960366, 1503960366, 150…
## $ date               <date> 2016-04-12, 2016-04-13, 2016-04-15, 2016-04-16, 20…
## $ TotalSleepRecords  <int> 1, 2, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, …
## $ TotalMinutesAsleep <int> 327, 384, 412, 340, 700, 304, 360, 325, 361, 430, 2…
## $ TotalTimeInBed     <int> 346, 407, 442, 367, 712, 320, 377, 364, 384, 449, 3…
```

```
head(sleep_day)
```

```
##           Id       date TotalSleepRecords TotalMinutesAsleep TotalTimeInBed
## 1 1503960366 2016-04-12                 1                327            346
## 2 1503960366 2016-04-13                 2                384            407
## 3 1503960366 2016-04-15                 1                412            442
## 4 1503960366 2016-04-16                 2                340            367
## 5 1503960366 2016-04-17                 1                700            712
## 6 1503960366 2016-04-19                 1                304            320
```