Preparación y Limpieza

Tratamiento de datos

Miguel Angel Cubillas C.

2024-07-13

Introducción

Este documento solo reflejará el proceso de tratamiento de los datos, limpieza, procesamiento, transformación etc. El documento de análisis, conclusiones, recomendaciones y demás será aparte, pero en el mismo repositorio.

Acerca de los datos

Los datasets que se proveyeron representan la información cuantificada respecto a las calorías, actividad física, ritmo cardiaco etc. de 30 clientes, el cual fue distribuido via Amazon Mechanical Turk, entre las fechas del 03/12/2016 hasta el 05/12/2016 (un registro de 61 días).

Almacenamiento y organización

Se usó el lenguaje de programación R para tratar con la información que se proveyó, se usarán los siguientes datasets que fueron almacenados en archivos CSV:

- dailyActivity_merged (Formato Ancho)
- sleepDay_merged (Formato Largo)
- heartrate_seconds_merged (Formato Largo)
- hourlyIntensities_merged (Formato Largo)

El resto de los datasets que fueron provistos contienen información más detallada y granular sobre las actividades y mediciones registradas.

Librerías

Las siguientes librerías nos proveerá suficientes herramientas para tratar con la información al menos en los primeros pasos.

```
library(tidyverse)
library(lubridate)
library(tibble)
library(skimr)
```

Lectura y transformación

Para la lectura correspondiente de los distintos csv elegidos a analizar, se usa la librería dplyr, que viene incluida con en la librería de tidyverse.

Las fechas están en un formato "mdy" (mes/dia/año), por lo tanto se ajusta en todos los datasets una fecha más estándar, también se ajustará a un formato más correcto los Id de los clientes, pues al ser una cadena de número se tomará como número, pero se dejará como caracteres, puesto que es un tipo de variable calitativa categórica. ##### Dataset de Actividades Diarias

```
daily_activity_3_12 <- read.csv("mturkfitbit_export_3.12.16-4.11.16/Fitabase Data 3.12.16-4.11.16/da
daily_activity_4_12 <- read.csv("mturkfitbit_export_4.12.16-5.12.16/Fitabase Data 4.12.16-5.12.16/da

daily_activity_union <- rbind(mutate(daily_activity_3_12, ActivityDate=mdy(ActivityDate)), mutate(da

daily_activity_union$Id <- as.character(daily_activity_union$Id)

# Guardar el dataframe en un archivo CSV
write.csv(daily_activity_union, file = "data_cleaned/daily_activity_union.csv", row.names = FALSE)
```

La asignación a los nombres de variables que terminan con "union", son simplemente la combinación entre los datasets con un mismo contexto de información pero de distintas fechas y datos, pero mismas variables.

Dataset de Registros de Sueño o Descanso

Solo hubo un dataset que registrara el descanso de los usuarios, que es del mes 4 hasta el mes 5, se tendrá en cuenta de todas maneras.

```
sleep_day_4_12 <- read.csv("mturkfitbit_export_4.12.16-5.12.16/Fitabase Data 4.12.16-5.12.16/sleepDay
sleep_day_4_12 <- mutate(sleep_day_4_12, TotalNoSleepMinutes= TotalTimeInBed-TotalMinutesAsleep, Sle

sleep_day_4_12$Id <- as.character(sleep_day_4_12$Id)

write.csv(sleep_day_4_12, file = "data_cleaned/sleep_activity.csv", row.names = FALSE)
```

Dataset de Ritmo Cardiaco

```
heartrate_4_12 <- read.csv("mturkfitbit_export_4.12.16-5.12.16/Fitabase Data 4.12.16-5.12.16/heartra
heartrate_3_12 <- read.csv("mturkfitbit_export_3.12.16-4.11.16/Fitabase Data 3.12.16-4.11.16/heartra

heartrate_union <- rbind(mutate(heartrate_4_12, Time = mdy_hms(Time)),mutate(heartrate_3_12, Time = m

heartrate_union$Id <- as.character(heartrate_union$Id)

write.csv(heartrate_union, file = "data_cleaned/heartrate_union.csv", row.names = FALSE)
```

Dataset de Intensidad

Este dataset hace referencia a cuánta intensidad hay registrada por las actividades realizada por la persona, puede variar desde medir una caminata normal, hasta deporte de alta intensidad.

```
hourly_intensities_3_12 <- read.csv("mturkfitbit_export_3.12.16-4.11.16/Fitabase Data 3.12.16-4.11.16
hourly_intensities_4_12 <- read.csv("mturkfitbit_export_4.12.16-5.12.16/Fitabase Data 4.12.16-5.12.16

hourly_intensities_union <- rbind(mutate(hourly_intensities_3_12, Date=mdy_hms(ActivityHour)),mutate

hourly_intensities_union$Id <- as.character(hourly_intensities_union$Id)
hourly_intensities_union <- hourly_intensities_union %>%
  select(-ActivityHour)
write.csv(hourly_intensities_union, file = "data_cleaned/intensities_union.csv", row.names = FALSE)
```

Confiabilidad de los datos

En este punto se observan que hay en varios datasets más de 30 clientes que proveyeron información a la empresa, y algunos datasets con menos de 30 clientes, contrario a los 30 exactos que se comentan en la recopilación de la información, por lo tanto quizás no sea la recopilación de información más hecha cuidadosamente, por ende la credibilidad sería menor. Sin embargo, las fuentes de la información sí es confiable, por lo que es información de primera fuente. Con el defecto de que no es información actual, pues al fin y al cabo es del 2016.

```
length(unique(daily_activity_union$Id))
```

```
## [1] 35
```

```
length(unique(hourly_intensities_union$Id))
```

```
## [1] 35
```

```
length(unique(sleep_day_4_12$Id))
```

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

```
length(unique(heartrate_union$Id))
```

```
## [1] 15
```

```
length(unique(hourly_intensities_union$Id))
```

```
## [1] 35
```

Limpieza

Datos nulos

No se hallaron datos nulos. Las fechas fueron corregidas a un formato manipulable.

```
skim_without_charts(daily_activity_union)
skim_without_charts(hourly_intensities_union)
skim_without_charts(sleep_day_4_12)
skim_without_charts(heartrate_union)
```

Datos duplicados

Eliminiación de duplicados:

```
daily_activity_union <- distinct(daily_activity_union)
hourly_intensities_union <- distinct(hourly_intensities_union)
sleep_day_4_12 <- distinct(sleep_day_4_12)
heartrate_union <- distinct(heartrate_union)
```