

## **Certificación profesional, caso práctico “Bellabeat”.**

### **Contexto:**

Soy analista de datos junior que trabaja en el equipo de analistas de marketing de Bellabeat. Bellabeat, es una empresa de alta tecnología que fabrica productos inteligentes focalizados en el cuidado de la salud.

Los productos con los cuales trabaja Bellabeat son los siguientes:

- **Aplicación Bellabeat:** La aplicación Bellabeat proporciona a los usuarios datos de salud relacionados con su actividad física, sueño, estrés, ciclo menstrual y hábitos de conciencia plena. Estos datos pueden ayudar a los usuarios a comprender sus hábitos actuales y adoptar decisiones saludables. La aplicación Bellabeat se conecta a su línea de productos de bienestar inteligentes.
- **Leaf:** Dispositivo de seguimiento clásico de bienestar de Bellabeat que se puede usar como pulsera, collar o clip. El dispositivo Leaf se conecta a la aplicación Bellabeat para hacer un seguimiento de la actividad física, el sueño y el estrés.
- **Time:** Este reloj de bienestar combina el aspecto intemporal de un reloj clásico con la tecnología inteligente para hacer el seguimiento de la actividad física, el sueño y el estrés del usuario. El reloj Time se conecta a la aplicación Bellabeat para proporcionar información sobre el bienestar diario.
- **Spring:** Es una botella de agua que hace el seguimiento diario del consumo de agua mediante el uso de tecnología inteligente para garantizar la hidratación adecuada a lo largo del día. La botella Spring se conecta a la aplicación Bellabeat para hacer el seguimiento de los niveles de hidratación.

### **1.PREGUNTAR:**

#### **1.1 Problemática.**

Sršen pide que analicen los datos de uso de los dispositivos inteligentes para saber cómo usan los consumidores los dispositivos inteligentes que no son de Bellabeat. Después, quiere que seleccionen un producto Bellabeat para aplicar estos conocimientos en tu presentación. Estas preguntas orientarán tu análisis:

#### **1.2 Preguntas:**

- 1. ¿Cuáles son algunas tendencias de uso de los dispositivos inteligentes?
- 2. ¿Cómo se podrían aplicar estas tendencias a los clientes de Bellabeat?
- 3. ¿Cómo podrían ayudar estas tendencias a influir en la estrategia de marketing de Bellabeat?

### 1.3 Interesados:

Principales actores, que buscan escuchar los resultados del análisis de datos..

- **Urška Sršen:** Fundadora y directora creativa de Bellabeat
- **Sando Mur:** Matemático y fundador de Bellabeat, miembro clave del equipo ejecutivo de Bellabeat.

## 2. PREPARAR:

Los datos, que se utilizarán en este análisis, son del conjunto “**Datos de seguimiento de actividad física de fitbit**”. Este siendo un dominio público, conjunto de datos disponibles a través de Mobius): Este conjunto de datos de Kaggle contiene el seguimiento de la actividad física personal en treinta usuarios de Fitbit. Treinta usuarios elegibles de Fitbit prestaron su consentimiento para el envío de datos personales de seguimiento que incluyen rendimiento de la actividad física en minutos, ritmo cardíaco y monitoreo del sueño. Incluye información sobre la actividad diaria, pasos y ritmo cardíaco que se puede usar para explorar los hábitos de los usuarios.

### 2.1 Característica de los datos:

Los datos están almacenados en formato CSV y se dividen en dos carpetas que contienen información sobre el seguimiento de la actividad física de treinta usuarios.

Estos datos se organizan con variables relacionadas a la actividad física, como pasos diarios, calorías quemadas, distancia recorrida y duración de la actividad en minutos, etc. Cada usuario está identificado mediante un número único que facilita el análisis individual y grupal, respetando la privacidad de los datos al no incluir informaciones personales como nombres personales o direcciones.

En términos de la calidad e integridad de los datos, estos mantienen uniformidad y estructura, lo que son necesarios para realizar un análisis de series de tiempos o patrones de comportamiento, aun así por el tamaño de muestra de 30 usuarios, los resultados de este análisis pueden no garantizar un resultado correcto aplicado en la realidad ya que no son lo suficientemente amplios para denominarse muestra representativa de la comunidad.

En cuanto a la confiabilidad de este conjunto de datos, los dispositivos Fitbit se consideran instrumentos precisos y confiables para la medición de actividad física y monitoreo de salud.

Al evaluar el conjunto bajo el criterio ROCCC, se puede decir lo siguiente de los datos de fitbit:

- **Confiables** debido a su precisión de dispositivos.
- **Originales:** provienen directamente de la fuente de monitoreo personal de cada usuario.
- **Íntegros:** cubren una amplia gama de variables relacionadas con la actividad física.
- **Actuales,** siempre y cuando se haya accedido a los datos en un período cercano a la fecha de análisis.

- **Citados**, con las fuentes claramente documentadas y disponibles en plataformas como Kaggle.

En conclusión, este conjunto de datos son adecuados para realizar el análisis de hábitos de salud y actividad física. Sin embargo, es fundamental considerar las limitaciones en cuanto a la representatividad del tamaño de muestra y el posible sesgo de selección.

En esta etapa, he decidido el uso de Posit-Cloud como medio de análisis, utilizando “Rstudio”.

RStudio, es un entorno de desarrollo integrado (IDE) directamente en la nube. Usar R en el proyecto de análisis de datos es mi selección debido a las capacidades analíticas y de visualización.

La integridad de los datos es primordial es por ello que comenzamos instalando algunos comandos en la consola de R para manipular y limpiar los datos.

Los comandos a utilizar son los siguientes:

- Tidyverse
- Dplyr
- ggplot2
- Tidy
- Here
- Skimr
- Janitor
- Readr

“ Documentación del proceso de instalación y cargado a la consola R”

```
> install.packages("tidyverse")
```

```
Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.4'
(as 'lib' is unspecified)
```

```
> install.packages("dplyr")
```

```
Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.4'
(as 'lib' is unspecified)
```

```
install.packages("ggplot2")
```

```
Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.4'
(as 'lib' is unspecified)
```

```
> install.packages("tidyr")
```

```
Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.4'
(as 'lib' is unspecified)
```

```

> install.packages("here")
Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.4'
(as 'lib' is unspecified)
> install.packages("skimr")
Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.4'
(as 'lib' is unspecified)

> install.packages("janitor")
Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.4'
(as 'lib' is unspecified)

> install.packages("readr")
Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.4'
(as 'lib' is unspecified)

> 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 3.5.1  ✓ tibble 3.2.1
✓ lubridate 1.9.3 ✓ tidyr 1.3.1
✓ purrr 1.0.2
— Conflicts — tidyverse_conflicts() —
X dplyr::filter() masks stats::filter()
X dplyr::lag() masks stats::lag()

Attaching package: 'janitor'

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

> library(scales)

Attaching package: 'scales'

The following object is masked from 'package:purrr':
  discard

The following object is masked from 'package:readr':
  col_factor

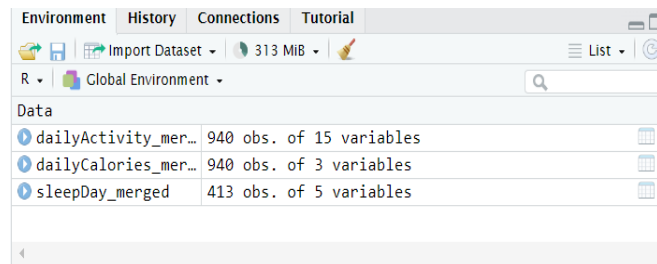
```

Estos comandos, son herramientas clave para estructurar información, detectar patrones y así podremos comunicar los resultados de una forma mucho mas clara y reproducible.

## 2. PROCESAR

2.1 Centraremos nuestro análisis en los siguientes elementos CSV:

1. DailyActivity\_merged.csv
2. SleepDay\_merged.csv
3. DailyCalories\_merged.csv



Environment	History	Connections	Tutorial
R   Global Environment			
Data			
dailyActivity_merged	940 obs. of 15 variables		
dailyCalories_merged	940 obs. of 3 variables		
sleepDay_merged	413 obs. of 5 variables		

2.2 Ingresamos una función para detectar duplicados:

```
> sum(duplicated(dailyActivity_merged))
[1] 0
> sum(duplicated(dailyCalories_merged))
[1] 0
> sum(duplicated(sleepDay_merged))
[1] 3
\
```

Esta fórmula suma los duplicados de las tablas y muestra la cantidad encontrada, en este caso la tabla "sleepDay\_merged" contiene tres elementos duplicados.

2.3 Ingresamos una función que elimine los duplicados encontrados:

```
> sleepDay_sin_duplicados <- sleepDay_merged %>%
+   distinct() %>%
+   drop_na()
```

2.4 Verificamos en nuestra tabla que no existan duplicados:

```
> sum(duplicated(sleepDay_sin_duplicados))
[1] 0
```

2.5 Usamos la función Str():

Esta función proporciona una visual general de la estructura del data frame mostrando los tipos de datos.

```
> str(dailyActivity_merged)
spc_tbl_ [940 × 15] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ Id                : num [1:940] 1.5e+09 1.5e+09 1.5e+09 1.5e+09 1.5e+09 ...
 $ ActivityDate       : chr [1:940] "4/12/2016" "4/13/2016" "4/14/2016" "4/15/2016"
 ...
 $ TotalSteps         : num [1:940] 13162 10735 10460 9762 12669 ...
 $ TotalDistance      : num [1:940] 8.5 6.97 6.74 6.28 8.16 ...
 $ TrackerDistance    : num [1:940] 8.5 6.97 6.74 6.28 8.16 ...
 $ LoggedActivitiesDistance: num [1:940] 0 0 0 0 0 0 0 0 0 ...
 $ VeryActiveDistance : num [1:940] 1.88 1.57 2.44 2.14 2.71 ...
 $ ModeratelyActiveDistance: num [1:940] 0.55 0.69 0.4 1.26 0.41 ...
 $ LightActiveDistance : num [1:940] 6.06 4.71 3.91 2.83 5.04 ...
 $ SedentaryActiveDistance : num [1:940] 0 0 0 0 0 0 0 0 0 ...
 $ VeryActiveMinutes  : num [1:940] 25 21 30 29 36 38 42 50 28 19 ...
 $ FairlyActiveMinutes : num [1:940] 13 19 11 34 10 20 16 31 12 8 ...
 $ LightlyActiveMinutes : num [1:940] 328 217 181 209 221 164 233 264 205 211 ...
 $ SedentaryMinutes   : num [1:940] 728 776 1218 726 773 ...
 $ Calories           : num [1:940] 1985 1797 1776 1745 1863 ...
```

```
> str(dailyActivity_merged)
spec_tbl_ [940 × 15] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ Id                : num [1:940] 1.5e+09 1.5e+09 1.5e+09 1.5e+09 1.5e+09 ...
 $ ActivityDate       : chr [1:940] "4/12/2016" "4/13/2016" "4/14/2016" "4/15/2016"
 ...
 $ TotalSteps         : num [1:940] 13162 10735 10460 9762 12669 ...
 $ TotalDistance      : num [1:940] 8.5 6.97 6.74 6.28 8.16 ...
 $ TrackerDistance    : num [1:940] 8.5 6.97 6.74 6.28 8.16 ...
 $ LoggedActivitiesDistance: num [1:940] 0 0 0 0 0 0 0 0 0 ...
 $ VeryActiveDistance : num [1:940] 1.88 1.57 2.44 2.14 2.71 ...
 $ ModeratelyActiveDistance: num [1:940] 0.55 0.69 0.4 1.26 0.41 ...
 $ LightActiveDistance : num [1:940] 6.06 4.71 3.91 2.83 5.04 ...
 $ SedentaryActiveDistance : num [1:940] 0 0 0 0 0 0 0 0 0 ...
 $ VeryActiveMinutes  : num [1:940] 25 21 30 29 36 38 42 50 28 19 ...
 $ FairlyActiveMinutes : num [1:940] 13 19 11 34 10 20 16 31 12 8 ...
 $ LightlyActiveMinutes : num [1:940] 328 217 181 209 221 164 233 264 205 211 ...
 $ SedentaryMinutes   : num [1:940] 728 776 1218 726 773 ...
 $ Calories           : num [1:940] 1985 1797 1776 1745 1863 ...

> str(sleepDay_sin_duplicados)
tibble [410 × 5] (S3: tbl_df/tbl/data.frame)
 $ Id                : num [1:410] 1.5e+09 1.5e+09 1.5e+09 1.5e+09 1.5e+09 ...
 $ SleepDay          : chr [1:410] "4/12/2016 12:00:00 AM" "4/13/2016 12:00:00 AM" "4/15/2016 12:00:00 AM" "4/16/2016 12:00:00 AM" ...
 $ TotalSleepRecords : num [1:410] 1 2 1 2 1 1 1 1 1 1 ...
 $ TotalMinutesAsleep: num [1:410] 327 384 412 340 700 304 360 325 361 430 ...
 $ TotalTimeInBed    : num [1:410] 346 407 442 367 712 320 377 364 384 449 ...
```

2.6 Deseamos cambiar el formato “chr” ya que las cadenas de caracteres pueden no seguir un formato estándar, lo que puede llevar a confusiones. Por ejemplo, "05-10-2024" puede interpretarse como el 5 de octubre de 2024 o el 10 de mayo de 2024, dependiendo de las convenciones de fecha.

```
> str(dailyActivity_merged)
spec_tbl_ [940 × 15] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ Id                : num [1:940] 1.5e+09 1.5e+09 1.5e+09 1.5e+09 1.5e+09 ...
 $ ActivityDate       : Date[1:940], format: "2016-04-12" "2016-04-13" ...
 $ TotalSteps         : num [1:940] 13162 10735 10460 9762 12669 ...
 $ TotalDistance      : num [1:940] 8.5 6.97 6.74 6.28 8.16 ...
 $ TrackerDistance    : num [1:940] 8.5 6.97 6.74 6.28 8.16 ...
 $ LoggedActivitiesDistance: num [1:940] 0 0 0 0 0 0 0 0 0 ...
 $ VeryActiveDistance : num [1:940] 1.88 1.57 2.44 2.14 2.71 ...
 $ ModeratelyActiveDistance: num [1:940] 0.55 0.69 0.4 1.26 0.41 ...
 $ LightActiveDistance : num [1:940] 6.06 4.71 3.91 2.83 5.04 ...
 $ SedentaryActiveDistance : num [1:940] 0 0 0 0 0 0 0 0 0 ...
 $ VeryActiveMinutes  : num [1:940] 25 21 30 29 36 38 42 50 28 19 ...
 $ FairlyActiveMinutes : num [1:940] 13 19 11 34 10 20 16 31 12 8 ...
 $ LightlyActiveMinutes : num [1:940] 328 217 181 209 221 164 233 264 205 211 ...
 $ SedentaryMinutes   : num [1:940] 728 776 1218 726 773 ...
 $ Calories           : num [1:940] 1985 1797 1776 1745 1863 ...
 attr(* "spec" )=

> head(dailyCalories_merged)
# A tibble: 6 × 3
      Id ActivityDay Calories
  <dbl> <date>     <dbl>
1 1503960366 2016-04-12    1985
2 1503960366 2016-04-13    1797
3 1503960366 2016-04-14    1776
4 1503960366 2016-04-15    1745
5 1503960366 2016-04-16    1863
6 1503960366 2016-04-17    1728

> str(dailyCalories_merged)
spec_tbl_ [940 × 3] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ Id                : num [1:940] 1.5e+09 1.5e+09 1.5e+09 1.5e+09 1.5e+09 ...
 $ ActivityDay       : Date[1:940], format: "2016-04-12" "2016-04-13" ...
 $ Calories          : num [1:940] 1985 1797 1776 1745 1863 ...
```

```

> head(sleepDay_sin_duplicados)
# A tibble: 6 × 5
      Id SleepDay   TotalSleepRecords TotalMinutesAsleep TotalTimeInBed
  <dbl> <date>         <dbl>             <dbl>             <dbl>
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
> str(sleepDay_sin_duplicados)
tibble [410 × 5] (S3: tbl_df/tbl/data.frame)
 $ Id      : num [1:410] 1.5e+09 1.5e+09 1.5e+09 1.5e+09 1.5e+09 ...
 $ SleepDay: Date[1:410], format: "2016-04-12" "2016-04-13" ...
 $ TotalSleepRecords: num [1:410] 1 2 1 2 1 1 1 1 1 1 ...
 $ TotalMinutesAsleep: num [1:410] 327 384 412 340 700 304 360 325 361 430 ...
 $ TotalTimeInBed    : num [1:410] 346 407 442 367 712 320 377 364 384 449 ...

```

Esta conversión de datos fue a través de la función **as.date**. As.date se utiliza para convertir objetos (generalmente cadenas de caracteres o números) a un formato de fecha.

```

> dailyActivity_merged$ActivityDate <- as.Date (dailyActivity_merged$ActivityDate, format =
"%m/%d/%Y")

> dailyCalories_merged$ActivityDay <- as.Date (dailyCalories_merged$ActivityDay, format =
"%m/%d/%Y")

> sleepDay_sin_duplicados$SleepDay <- sleepDay_sin_duplicados$SleepDay %>%
+   as.POSIXct(format = "%m/%d/%Y %I:%M:%S %p") %>%
+   as.Date(format = "%m/%d/%Y %I:%M:%S %p")

```

2.7 Corroboramos la participación de usuarios en los archivos CSV, La función `n_distinct()` permite contar los valores únicos en una columna o conjuntos de columnas de un dataframe.

```

> n_distinct(dailyActivity_merged$Id)
[1] 33
> n_distinct(dailyCalories_merged$Id)
[1] 33
> n_distinct(sleepDay_sin_duplicados$Id)
[1] 24

```

Al corroborar la participación, la tabla `sleepDay` tiene información de sueño solo para 24 usuarios. Aun así seguiremos utilizando estos datos para nuestro informe.

### 3. ANALIZAR

Describiremos las variables básicas, como media, medianas,mínimos, máximos con la función summary() para cada variable.

```
> summary(dailyActivity_merged)
   Id      ActivityDate      TotalSteps      TotalDistance
Min.   :1.504e+09   Min.   :2016-04-12   Min.    :    0   Min.    : 0.000
1st Qu.:2.320e+09   1st Qu.:2016-04-19   1st Qu.: 3790   1st Qu.: 2.620
Median :4.445e+09   Median :2016-04-26   Median : 7406   Median : 5.245
Mean   :4.855e+09   Mean   :2016-04-26   Mean    : 7638   Mean    : 5.490
3rd Qu.:6.962e+09   3rd Qu.:2016-05-04   3rd Qu.:10727   3rd Qu.: 7.713
Max.   :8.878e+09   Max.   :2016-05-12   Max.    :36019   Max.    :28.030
TrackerDistance   LoggedActivitiesDistance   VeryActiveDistance   ModeratelyActiveDistance
Min.    : 0.000   Min.    :0.0000   Min.    : 0.000   Min.    :0.0000
1st Qu.: 2.620   1st Qu.:0.0000   1st Qu.: 0.000   1st Qu.:0.0000
Median : 5.245   Median :0.0000   Median : 0.210   Median :0.2400
Mean    : 5.475   Mean    :0.1082   Mean    : 1.503   Mean    :0.5675
3rd Qu.: 7.710   3rd Qu.:0.0000   3rd Qu.: 2.053   3rd Qu.:0.8000
Max.    :28.030   Max.    :4.9421   Max.    :21.920   Max.    :6.4800
LightActiveDistance   SedentaryActiveDistance   VeryActiveMinutes   FairlyActiveMinutes
Min.    : 0.000   Min.    :0.000000   Min.    : 0.00   Min.    : 0.00
1st Qu.: 1.945   1st Qu.:0.000000   1st Qu.: 0.00   1st Qu.: 0.00
Median : 3.365   Median :0.000000   Median : 4.00   Median : 6.00
Mean    : 3.341   Mean    :0.001606   Mean    : 21.16   Mean    : 13.56
3rd Qu.: 4.782   3rd Qu.:0.000000   3rd Qu.: 32.00   3rd Qu.: 19.00
Max.    :10.710   Max.    :0.110000   Max.    :210.00   Max.    :143.00
LightlyActiveMinutes   SedentaryMinutes   Calories
Min.    : 0.0   Min.    : 0.0   Min.    : 0
1st Qu.:127.0   1st Qu.: 729.8   1st Qu.:1828
Median :199.0   Median :1057.5   Median :2134
Mean    :192.8   Mean    : 991.2   Mean    :2304
3rd Qu.:264.0   3rd Qu.:1229.5   3rd Qu.:2793
Max.    :518.0   Max.    :1440.0   Max.    :4900
```