Análisis para Bellabeat

Isabel Colorado López

2023-08-16

A cerca de la empresa

Es una empresa de alta tecnología que fabrica productos inteligentes focalizados en el cuidado de la salud. Recopila datos sobre la actividad física, el sueño, el estrés y la salud reproductiva le ha permitido a Bellabeat proporcionar a las mujeres conocimientos sobre su propia salud y sus hábitos. Desde su fundación, en 2013, Bellabeat creció a un ritmo vertiginoso y rápidamente se posicionó como empresa de bienestar impulsada por la tecnología para las mujeres.

Preguntas para el analisis

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

Tarea empresarial

Identificar patrones y tendencias entre los datos de usuarios que utilizan dispositivos inteligentes que no son de Bellabeat, para ayudar a la estrategia de marketing y al crecimiento de la empresa.

Los datos

La fuente de los datos utilizados para este analisis son:

- ***Datos de seguimiento de actividad física de Fitbit*** (CC0: 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. En el año de 2016
- Un ***Estudio*** realizado por la Universidad del Estado de Arizona, Estados Unidos, establece una escala de cuántos pasos se necesitan dar al día.
- Esta ***página*** contiene las horas recomendadas de sueño por edad.

Análisis

Las bases de datos utilizadas las limpié con google sheets antes de subirlas a R. Al explorar por los datos me di cuenta que el formato de las fechas me las reconocía como caracteres, y al usar la función de cambiar formato no tuve éxito, por lo que utlicé una función llamada `FECHA` y pude corregir el error. También quité nulls y duplicados si existiesen, al igual que verifiqué si había algún tipo de sesgo en los datos. Después de que todo estaba correcto, los cargué a R.

Códigos de análisis en R

Lo primero que hice fue cargar los archivos y los paquetes que iba a necesitar para mi análisis.

1. Daily activity

```
DailyActivity <- read.csv("C:\\Users\\isabe\\Downloads\\dailyActivity_merged.csv")
```

2. Sleep Day

```
sleepDay <- read.csv("C:\\Users\\isabe\\Downloads\\sleepDay_merged.csv")
```

3. Hourly Steps

```
hourlySteps <- read.csv("C:\\Users\\isabe\\Downloads\\hourlySteps_merged.csv")
```

Paquetes necesarios para el análisis

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.2     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(dplyr)
library(rstatix)
```

```
##
## Attaching package: 'rstatix'
##
## The following object is masked from 'package:stats':
##
##     filter
```

```
library(ggplot2)
```

Limpieza de los datos

Al explorar estos datos en R me dí cuenta que el formato de fecha y hora lo renocía como caracteres, a pesar de haber cambiado el formato en las hojas de cálculo, por lo que cambié. También, los usuarios que están escritos como números (Id's), estaban en formato *integer*, entonces los cambié a formato *character* para que mi análisis fuera basados en ellos.

```
DailyActivity$Id= as.character(DailyActivity$Id)
DailyActivity$ActivityDate=as.POSIXct(DailyActivity$ActivityDate,
                                      format = "%d/%m/%y",tz=Sys.timezone())
```

Hice lo mismo con los otros dos archivos que presentaban el mismo problema de formato. Y al archivo que contenía formato *fecha-hora* también le apliqué el cambio.

```
sleepDay$Id=as.character(sleepDay$Id)
sleepDay$SleepDay=as.POSIXct(sleepDay$SleepDay,
                             format= "%m/%d/%y")
```

Con los datos de *hourlySteps* separé la fecha y la hora en columnas distintas para poder facilitar mi análisis. También cambié el formato de 12hrs al de 24hrs para facilitar las visualizaciones.

```
hourlySteps$Id=as.character(hourlySteps$Id)
hourlySteps$ActivityHour=as.POSIXct(hourlySteps$ActivityHour,
                                    format ="%m/%d/%Y %H:%M:%S")
hourlySteps$fecha <- date(hourlySteps$ActivityHour)
hourlySteps$hora <- format(hourlySteps$ActivityHour,
                           format= "%H:%M:%S")
hourlySteps<- separate(hourlySteps,col = ActivityHour,
                       into = c("fecha","hora","pmoam"),sep = " ")
hourlySteps<- unite(hourlySteps,
                    "hora","pmoam",col="Hora",sep = ":",remove = FALSE)
```

Análisis

Primero revisé la cantidad de usuarios que participaron en cada estudio.

```
count(distinct(DailyActivity,Id))
```

```
##    n
## 1 33
```

```
count(distinct(sleepDay, Id))
```

```
##    n
## 1 24
```

```
count(distinct(hourlySteps,Id))
```

```
##    n
## 1 33
```

Luego creé un nuevo marco de datos con la suma de todos los tiempos de los tipos de actividad para tener el *total*. Esto me ayudará más adelante en mi análisis.

```
DailyActivityV2 <- DailyActivity %>%
  mutate(TotalActivity=VeryActiveMinutes+FairlyActiveMinutes+
           LightlyActiveMinutes)
```

Hallazgos que los datos mostraron

Para adultos sanos, caminar menos de 5.000 pasos diarios es equivalente a un estilo de vida sedentario.

- Si se dan entre 5.000 y 7.500 pasos diarios determina una actividad baja/moderada.
- Entre 7.500 y 10.000 pasos diarios equivale a un estilo algo activo.
- Una persona tiene una vida activa cuando supera los 10.000 pasos.

Por lo tanto, andar más siempre es una buena idea. Así lo demuestra este estudio que observó cómo las tasas de mortalidad disminuyeron progresivamente antes de nivelarse aproximadamente a los 7.500 pasos/día.

Al crear un resumen de los datos me dí cuenta que el promedio de pasos entre los usuarios es de 7671, lo que quiere decir que las personas que participaron en este estudio son medianamente activas.

También se muestra que el promedio de la cantidad en minutos de sedentarismo es mayor sobre la cantidad de actividad total (989/228), lo cual quiere decir que los usuarios deberían ser más activos para tener una vida más saludable.

```
DailyActivityV2 %>% select(TotalSteps,SedentaryMinutes,
                           TotalActivity) %>% summary()
```

```
##    TotalSteps    SedentaryMinutes TotalActivity
##  Min.   :    0   Min.   :   0.0   Min.   :  0.0
##  1st Qu.: 3790   1st Qu.: 729.8   1st Qu.:146.8
##  Median : 7406   Median :1057.5   Median :247.0
##  Mean   : 7638   Mean   : 991.2   Mean   :227.5
##  3rd Qu.:10727   3rd Qu.:1229.5   3rd Qu.:317.2
##  Max.   :36019   Max.   :1440.0   Max.   :552.0
```

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

```
##  TotalMinutesAsleep TotalTimeInBed
##  Min.   : 58.0      Min.   : 61.0
##  1st Qu.:361.0      1st Qu.:403.0
##  Median :433.0      Median :463.0
##  Mean   :419.5      Mean   :458.6
##  3rd Qu.:490.0      3rd Qu.:526.0
##  Max.   :796.0      Max.   :961.0
```

```
hourlySteps %>% group_by(hora) %>%
  select(StepTotal) %>% summary()
```

```
## Adding missing grouping variables: `hora`
```

```
##      hora             StepTotal
##  Length:22099       Min.   :    0.0
##  Class :character   1st Qu.:    0.0
##  Mode  :character   Median :   40.0
##                     Mean   :  320.2
##                     3rd Qu.:  357.0
##                     Max.   :10554.0
```

Creé un nuevo marco de datos para resumir el promedio de cada variable por usuario y poder visualizar mejor las tendencias de los datos

```
promedioTotalA<-DailyActivityV2 %>% group_by(Id) %>%
  summarise_all(mean)
```

Al crear el nuevo marco de datos me encuentro que los valores que me interesan, son decimales, por lo que decido cambiarlos a número enteros.

```
promedioTotalAV2<-promedioTotalA %>%
  mutate(RoundSteps=as.integer(round(TotalSteps)))
```

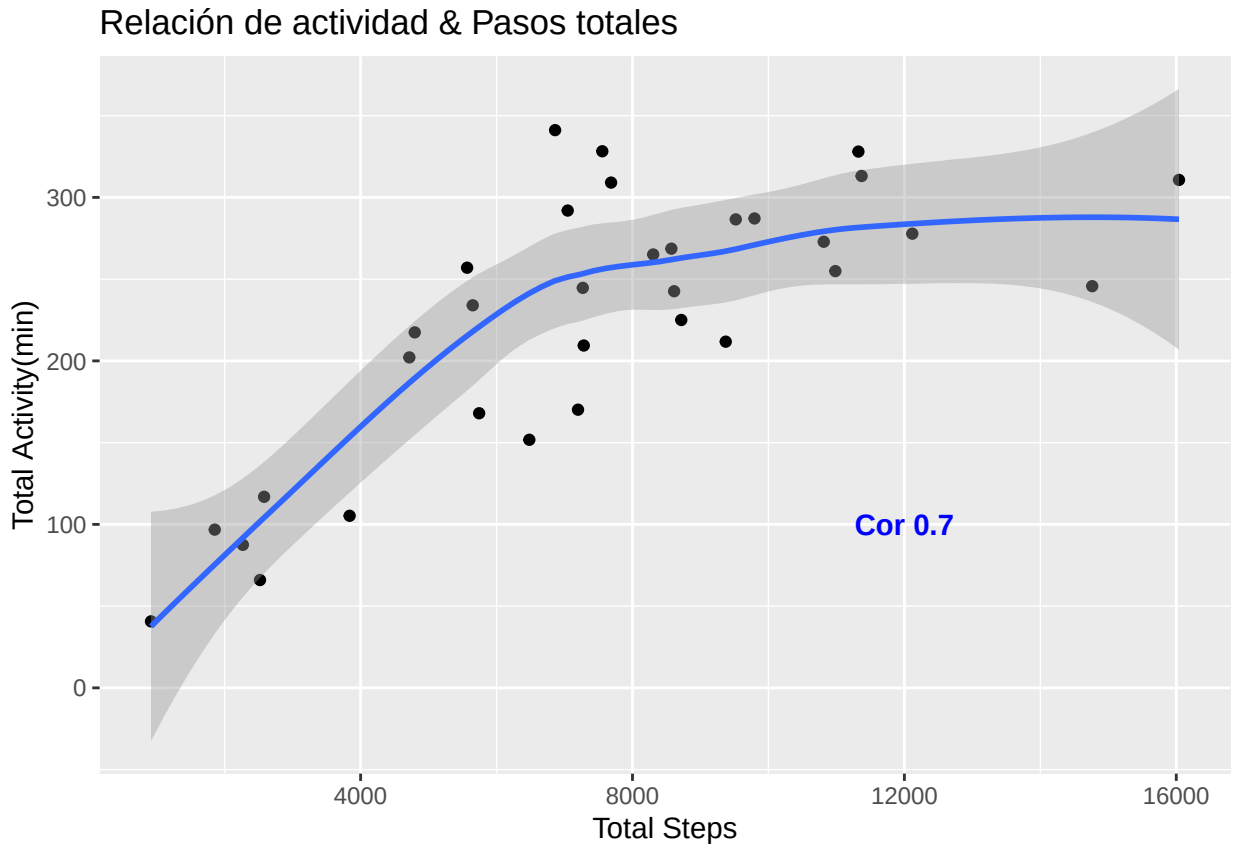
Nivel de actividad por usuario

Después de saber lo que se considera activo, moderado o sendentario en cuanto a pasos diarios, lo que hice fue clasificar los usuarios según su nivel de actividad, para así poder tener una idea más clara de qué tan activos son estas personas.

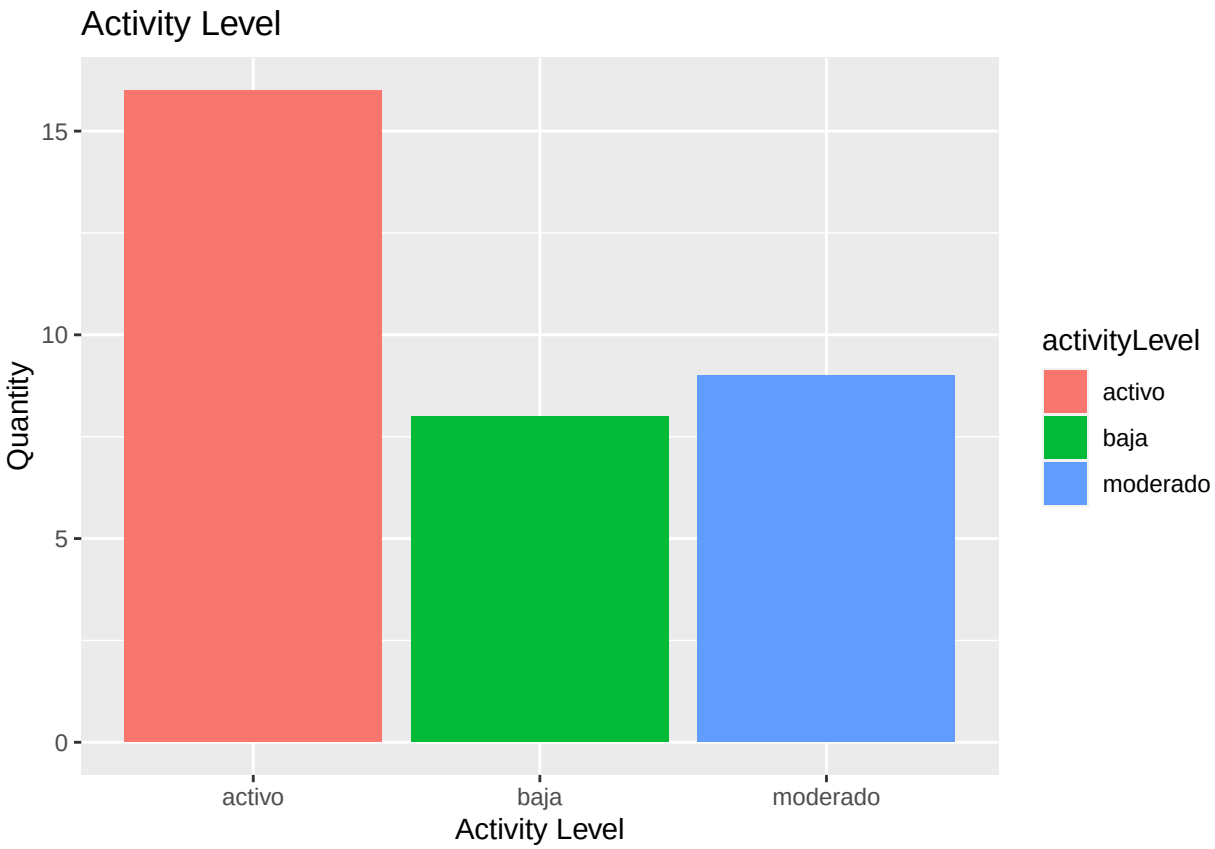
Visualizaciones de los datos

Primero creé un diagrama de dispersión para ver la relación que hay entre el promedio de pasos totales por persona y el promedio de actividad. Me encontré que hay una correlación positiva del 0.7, lo que quiere decir que mientras más pasos se da, hay mayor nivel de actividad general. Caminar más siempre es mejor.

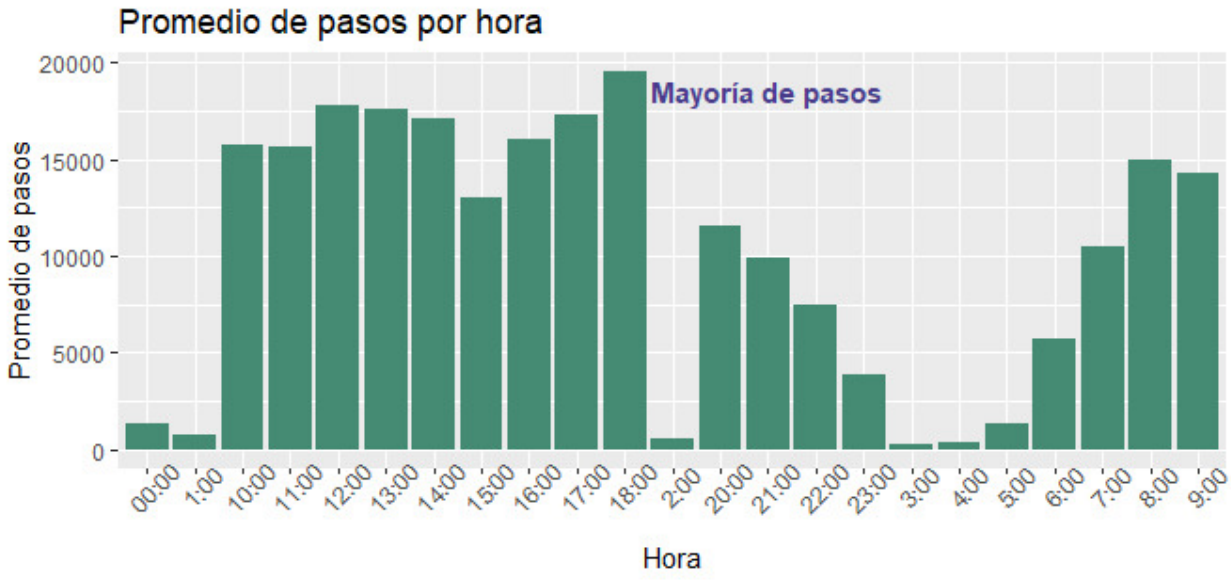
```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



Después creé un gráfico de barras para ver qué tan activas son las personas, y se puede ver que la mayoría se clasifican como **activos**. Esta clasificación puede ser de gran ayuda para los dispositovos inteligentes, ya que se pueden implementar alertas según el nivel de actividad para ayudarle a las personas a tener un día más activo.



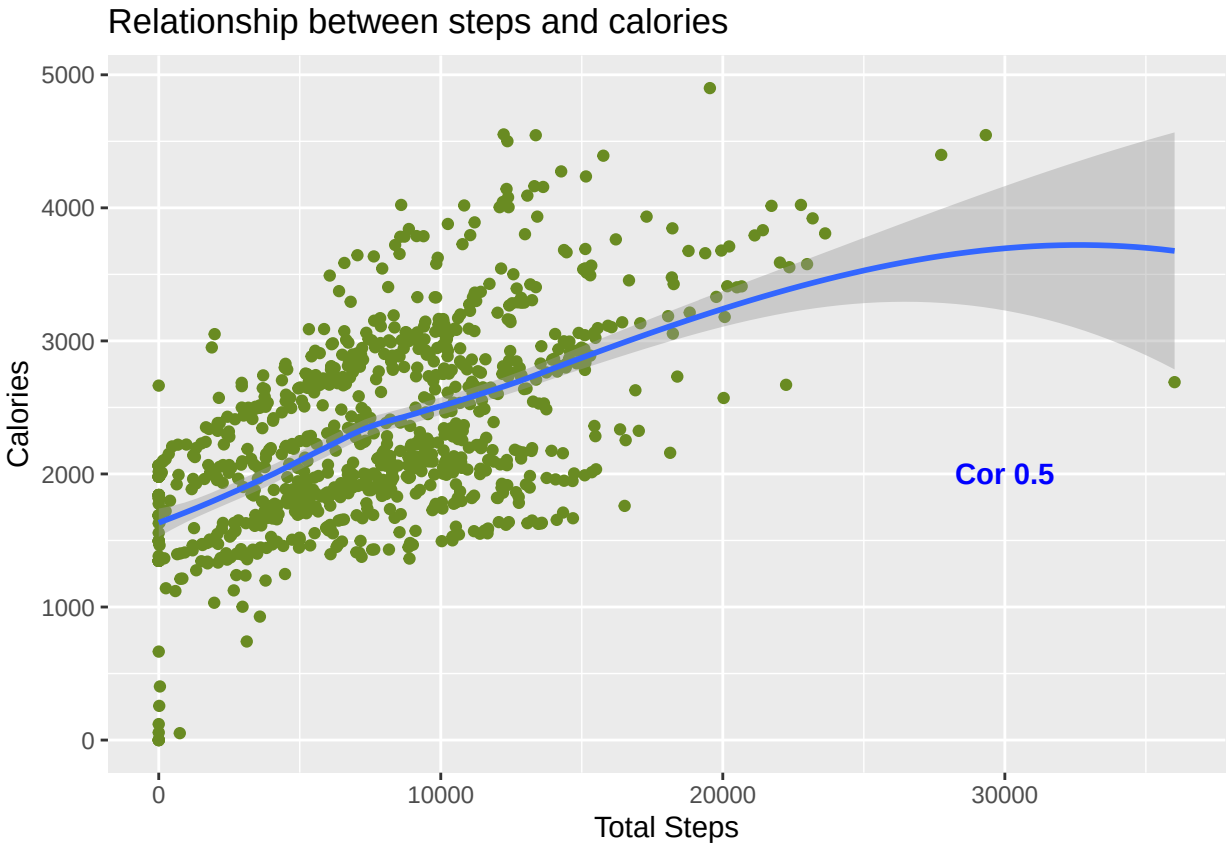
Luego de esto hice un gráfico de columnas ya que quería ver el promedio de pasos por hora. Lo que mostraron los datos fue que la hora en donde se realizan más pasos es a las 18:00, puede ser una hora en la que las personas tienen más tiempo para salir a caminar y porque el clima es más fresco, sin embargo si quisiera ahondar más, haría falta otro análisis dedicado específicamente en esto.



Seguido a esto quería comprobar que a mayor pasos dados, mayor calorías se gastan durante el día, por medio de un diagrama de dispersión. Esto lo digo porque se puede dar el caso en que no se camine mucho, pero

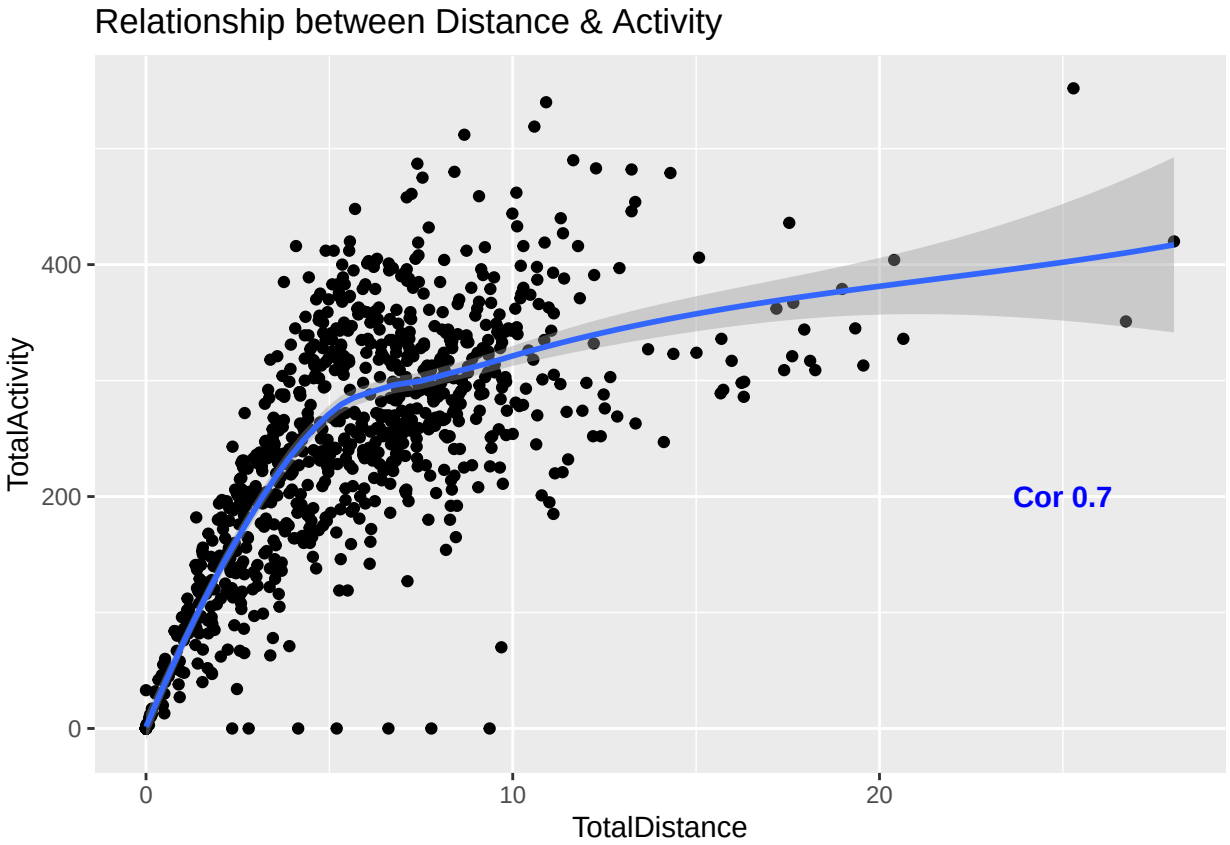
se gasten calorías de otras maneras. Dado a los datos se puede ver que efectivamente existe una correlación positiva.

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



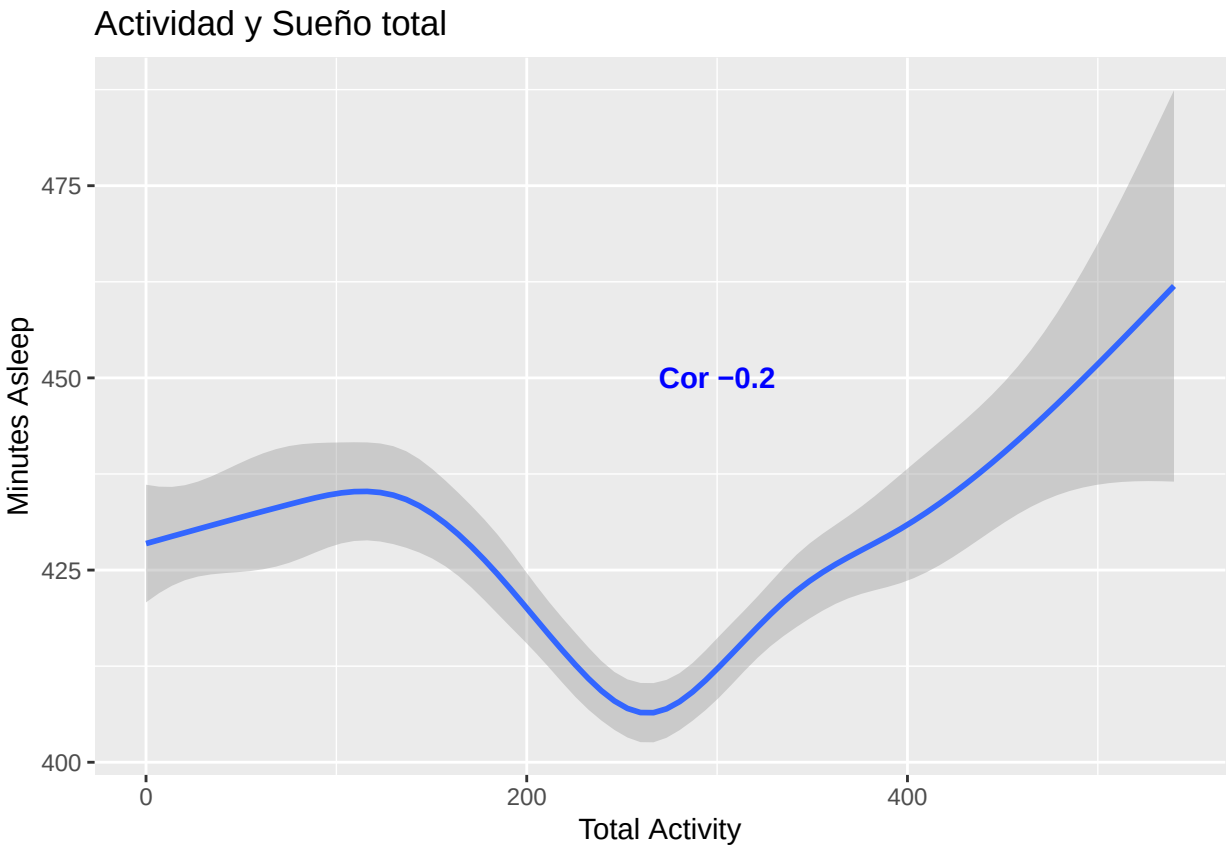
Hice lo mismo pero con las variables **distancia y actividad**. El resultado fue que a mayor distancia, mayor actividad general durante el día. Esto quiere decir que la mayoría del tiempo invertido en ejercitarse, corresponde a los pasos dados. Y esto una vez más nos dice que se debe incentivar más a las personas a caminar más, para tener una mejor salud.

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



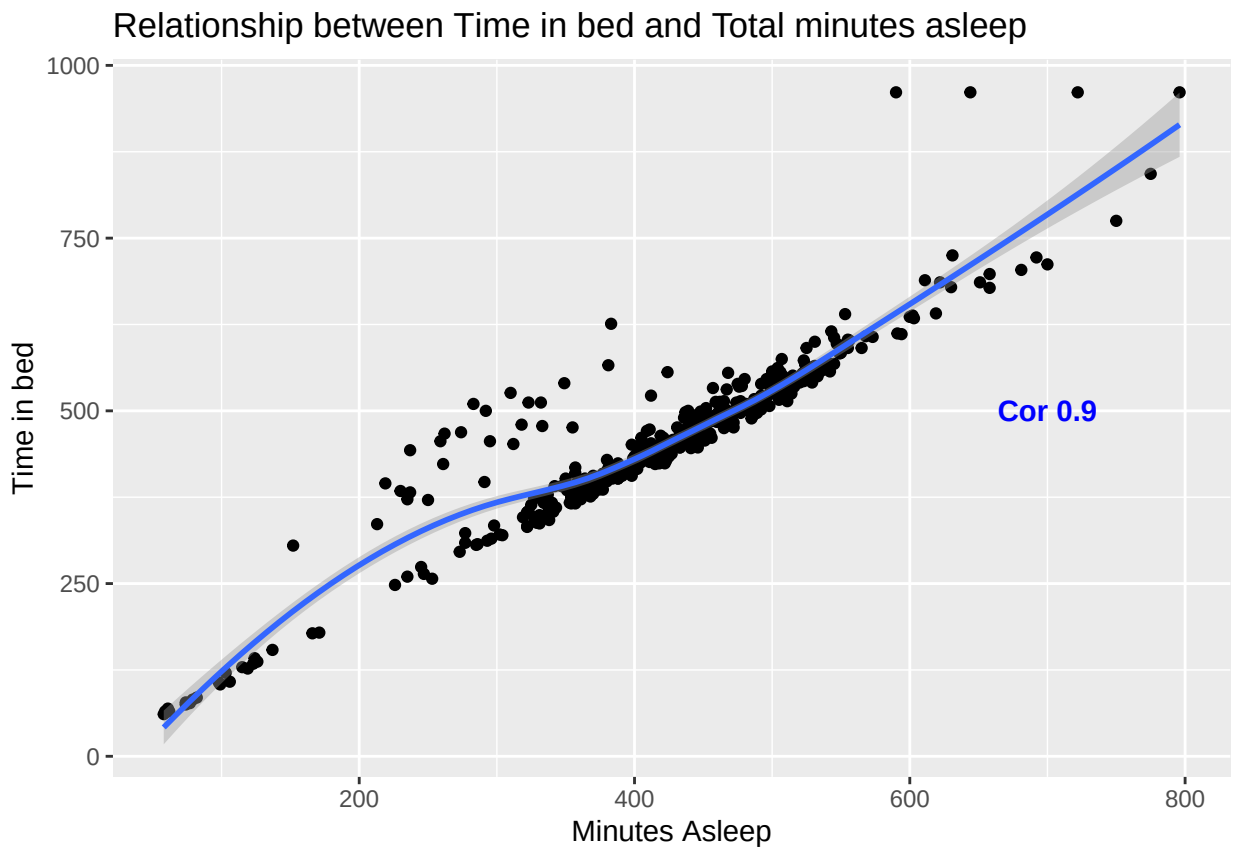
Dado a que ya sabemos que la actividad durante el día se debe en mayor medida a los pasos dados, entonces creé un gráfico que me mostrara la relación que hay entre el sueño y la actividad total. Mi teoría era que si las personas lograban dormir más, iban a tener más energía durante el día, pero los datos mostraron lo contrario. Con una correlación negativa del -0.2, se comprueba que no por dormir más, tienden a tener más actividad.

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

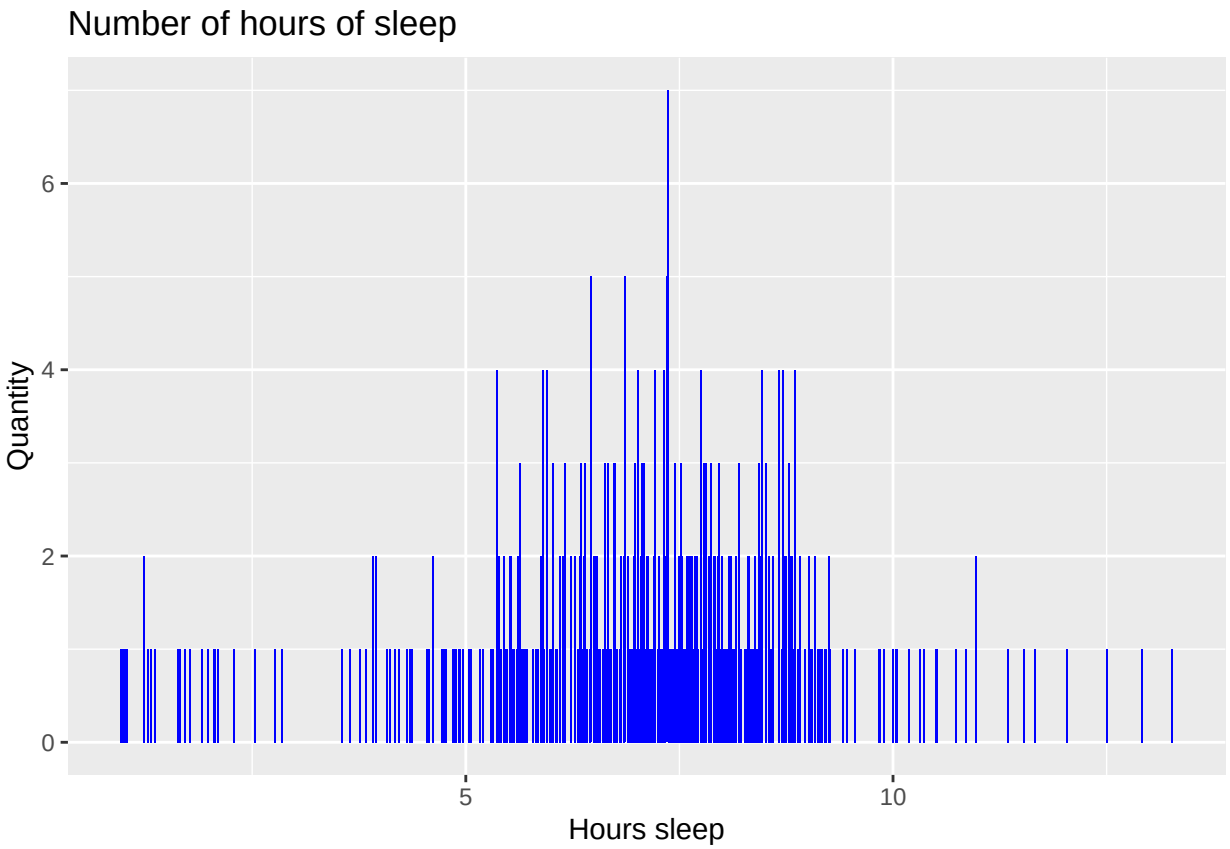


Otro punto que quise visualizar fue el tiempo en cama de los usuarios y la minutos de sueño. Se puede apreciar bastante claro que la mayoría del tiempo que pasan en su cama es dedicada exclusivamente al sueño, cosa que es realmente positiva.

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



Para finalizar, deseaba ver las horas promedio que duermen los usuarios de fitbit,por medio de un gráfico de barras, ya que lo recomendado para adultos y adultos jóvenes es dormir entre 7 y 9hrs. Lo que descubrí fue que el promedio de horas dormidas es de 6.9, lo cual está casi en el rango recomendado. También pude ver que el máximo de horas dormido es de 13hrs, cosa que no es del todo bueno, más no es mejor. Por lo que se debería incentivar a los usuarios a dormir las horas recomendadas, con recordatorios y alarmas.



Recomendaciones para Bellabeat

La mayoría de los usuarios tienden a ser más activos, aunque no todos, es por esto que se debe alentar a los usuarios a tener estilos de vida más saludables, y así tener una mejor vida en general.

Por lo que:

- Mandar alertas cuando se tiene un nivel de actividad bajo o moderado para animar a las personas a caminar más.
- Enviar notificaciones cuando el usuario haya alcanzado un nivel de más de 10000 diarios, lo cual se considera **muy activo**, esto para premiar la actividad física y promoverla.
- Implementar un sistema de *vibración* cuando la persona pasa más tiempo sedentaria.
- Recordatorios de las horas de sueño recomendadas para cada usuario, dependiendo de su edad. Esto ayudaría a que las personas logren mejor calidad de sueño.