Paso 1: Preguntar

En primer lugar y como parte del proceso de análisis de la tarea empresarial y los interesados claves procederemos a resumir algunos elementos que ayudarán en la construcción de nuestro caso. En primer lugar, pasemos a definir la tarea empresarial un paso súper importante pues nos permite: trabajar con enfoque y claridad, sirve como brújula a todo el equipo, aumenta la eficiencia en la recopilación de datos, mejora la calidad del análisis, facilita la comunicación efectiva y se asegura que el proyecto esté alineado con los objetivos organizacionales.

Tarea Empresarial:

 Analizar los datos de uso de dispositivos inteligentes para orientar la estrategia de marketing de Bellabeat.

Un segundo paso interesante sería realizar un pequeño resumen de los principales interesados y los productos de la empresa que utilizaremos como base:

Principales Interesados:

Interesado	Rol
Urška Sršen	Cofundadora y directora creativa de Bellabeat.
Sando Mur	Matemático y cofundador de Bellabeat, miembro clave del equipo ejecutivo.
Equipo de análisis computacional de datos de marketing de Bellabeat	Analistas de datos encargados de recopilar, analizar e informar datos para la estrategia de marketing.

Productos:

Producto	Descripción
Aplicación Bellabeat	Proporciona a los usuarios datos de salud relacionados con actividad física, sueño, estrés, ciclo menstrual y hábitos de conciencia plena. Conectada a la línea de productos de bienestar inteligentes de Bellabeat.
Leaf	Dispositivo de seguimiento clásico de bienestar que se puede usar como pulsera, collar o clip. Conectado a la aplicación Bellabeat para hacer un seguimiento de actividad física, sueño y estrés.

Producto	Descripción
Time	Reloj de bienestar que combina el aspecto de un reloj clásico con tecnología inteligente para hacer el seguimiento de actividad física, sueño y estrés. Conectado a la aplicación Bellabeat.
Spring	Botella de agua que hace el seguimiento diario del consumo de agua mediante tecnología inteligente. Conectada a la aplicación Bellabeat para hacer el seguimiento de los niveles de hidratación.
Membresía de Bellabeat	Programa de membresía mediante suscripción que ofrece acceso 24/7 a orientación personalizada sobre nutrición, actividad física, sueño, salud y belleza, y conciencia plena según el estilo de vida y las metas del usuario.

Paso 2: Preparar

A continuación, se detallan los pasos a seguir en la etapa de preparación de los datos.

Paso	Acción
1	Descargar los datos del conjunto de actividad física de Fitbit y almacenarlos en una PC local.
2	Realizar una primera exploración de los datos de acuerdo con la descripción proporcionada.
3	Modelar sus datos en hojas de cálculos (si es necesario) para incorporarlos a trabajar con SQL.

Se procede a extraer los datos de una base de datos existentes en Kaggle (1). Una primera aproximación de los datos brinda los siguientes detalles de su estructura e historia:

- El conjunto de datos fue generado a través de una encuesta distribuida en Amazon Mechanical Turk entre el 03.12.2016 y el 05.12.2016.
- Treinta usuarios elegibles de Fitbit consintieron en compartir datos personales de seguimiento, incluyendo actividad física, ritmo cardíaco y monitoreo del sueño a nivel de minuto.
- Los informes individuales pueden ser analizados por ID de sesión de exportación (columna A) o marca de tiempo (columna B), y la variación en la salida refleja el uso de diferentes tipos de rastreadores Fitbit y comportamientos de seguimiento individuales/preferencias.
- La inspiración detrás del conjunto de datos se centra en el análisis del comportamiento temporal humano y el reconocimiento de patrones.
- Se mencionan lecturas adicionales sobre cómo analizar los datos de FitBit para mejorar la salud general y cómo obtener y analizar datos de rastreadores de fitness con un enfoque forense.
- Etiquetas asociadas al conjunto de datos incluyen negocios y ejercicio.
- Se recopilan opiniones sobre el conjunto de datos, con la mayoría de los usuarios describiéndolo como bien documentado, bien mantenido, con datos limpios y de alta calidad.

Los datos serán descargados e incorporados a un PC para su posterior preparación y procesamiento con el objetivo de identificar: posibles valores faltantes, incoherencias en la estructura, sesgos y valorar la pertinencia (o no) de utilizar fuentes de datos adicionales que complementen el análisis.

A continuación, se resumen algunas características generales de las observadas a primera vista cuando se interactúan con los datos:

- Según Kaggle (fuente muy confiable para la obtención de datos) estos están valorados de forma perfecta en las categorías: etiquetado, credibilidad y compatibilidad.
- La base de datos consta de 18 archivos en total cada uno de ellos contiene un atributo diferente en función del individuo y el atributo que se esté pretendiendo delimitar.
- Como clave única que conecta todos los datos se utilizó un identificador numérico para cada individuo reflejado en cada tabla bajo el nombre: "Id".
- Adicionalmente se observa que el monitoreo se realizó por un total de 31 días (desde el 12 de abril de 2016 al 12 de mayo de 2016)
- La base de datos tiene un formato largo al estar indicado para cada Id su respectiva medición.
- Los datos parecen ser confiables, originales, integrales, actuales y citados

Proceso de carga para su almacenamiento en BigQuery:

- Al introducir los datos en BigQuery se presentó un error en un grupo de tablas relacionado al formato de una de sus columnas (la que combinaba día y hora) por lo que se procedió a realizar una modificación de los datos en hojas de cálculo y RStudio que separare en dos columnas diferentes: Día y Hora.
- Para esta tarea se utilizaron Google Sheets en los archivos más pequeños y para aquellos donde fuera más complicada la manipulación de datos se utilizó Rstudio dejándose reflejado todo el código en el siguiente archivo RMarkdown (2).
- Luego de la manipulación se logró uniformidad entre todas las tablas en lo que a fecha y hora se refiere posibilitando que todos estos documentos fueran admitidos por BigQuery para posteriores análisis.

Paso 3: Procesar

Paso	Acción
2	Cargar los datos en BigQuery para almacenamiento en la nube.
3	Realizar una primera exploración con lenguajes de consulta como SQL.
4	Evaluar la credibilidad de los datos:
	a. Revisar metadatos.
	b. Realizar exploración preliminar.
	c. Verificar consistencia de datos.
	d. Validar cruzadamente con otras fuentes.
5	Considerar agregar datos adicionales si es necesario y documentar todas las fuentes utilizadas.

A continuación, se procede a realizar una exploración de los datos en BigQuery donde están ya cargados todos los documentos. Para ello se revisarán metadatos, se realizará una investigación preliminar y se verificará la consistencia de los datos.

En primer lugar, se procede a revisar los datos de muestra para ello se utiliza la consulta:

```
SELECT * FROM [nombre_del_dataset].[nombre_de_la_tabla] LIMIT 10 para revisar datos de muestra y verificar que la información tenga sentido.
```

Luego se buscan identificar valores extremos o anómalos, con este objetivo se utilizan consultas con funciones agregadas como MIN(), MAX(), AVG().

De momento no se observan datos anómalos luego de la exploración como podrían ser datos negativos o ilógicos en lo que respecta a duración, pasos, etc.

Como algunos datos fueron modificados en hoja de cálculo y otros en Rstudio se observa que hay inconsistencia en que algunas tablas reconocen la Hora como cadena de texto por lo que se procederá a formatearlo en R para que todos los datos sean uniformes.

Continuando con el proceso de revisión de datos se procede a contar la cantidad de elementos diferentes de cada tabla en las columnas "Id" y Activity Date" al saber que en cada columna debe haber un máximo de 31 elementos. Para ello se utiliza la siguiente consulta para cada tabla.

```
SELECT COUNT(DISTINCT Id) AS CountId,

COUNT(DISTINCT ActivityDate) AS CountActivityDate

FROM nombre_de_la_tabla;
```

Un conteo de todas las tablas refleja que la mayoría de las tablas poseen 33 elementos diferentes en la columna "Id" y 31 elementos en la columna "Activity Date". Resulta de especial relevancia que una de las tablas posee 32 días por lo que se procede a revisarla para eliminar el posible dato aberrante.

```
SELECT

ActivityDate,

COUNT(*) AS RecordCount

FROM `proyecto-final-bellabet.proyecto_final_bellabet.minuteSleep_merged`

GROUP BY ActivityDate

ORDER BY RecordCount DESC;
```

Luego de ejecutar la consulta se verifica que esa tabla posee datos del día 11 de abril siendo la única tabla que lo hace por lo tanto se procede a eliminar los datos de ese día con el objetivo de asegurar la uniformidad, esto se realiza con la consulta siguiente:

```
DELETE FROM `proyecto-final-
bellabet.proyecto_final_bellabet.minuteSleep_merged`
WHERE ActivityDate = '2016-04-11';
```

Igualmente es llamativo que la tabla que refleja el peso de los individuos posee solamente 8 ld diferentes. Cuando se obtiene una visualización de esta tabla se observa que al parecer solamente 8 individuos midieron su peso durante el estudio, algo que no nos servirá de mucho por ser muy pocos datos, pero también nos percatamos que los valores de pesos en kg y libras están evidentemente incorrectos y se deben arreglar infiriendo a partir de los valores correctos (en los que el peso en kg es correcto se inferirá el peso en lb y viceversa).

Luego de la manipulación de datos se logra inferir el peso en Kilogramos y Libras a partir de los datos existentes, así mismo se lograron formatear esos datos a número:

```
CREATE OR REPLACE TABLE `proyecto-final-
bellabet.proyecto final bellabet.weightLogInfo merged updated` AS
SELECT
  Id,
  ActivityDay,
  Hour,
  WeightKg,
  WeightPounds,
  Fat,
  BMI,
  IsManualReport,
  LoaId,
  Hour new,
  CASE
    WHEN SAFE CAST(SUBSTR(WeightPounds, 0, STRPOS(WeightPounds, '.')) AS
FLOAT64) > 100 THEN ROUND (SAFE CAST (SUBSTR (WeightPounds, 0,
STRPOS (WeightPounds, '.')) AS FLOAT64) / 2.20462, 2)
    WHEN SAFE CAST (WeightKg AS FLOAT64) <= 150 THEN SAFE CAST (WeightKg AS
FLOAT64)
    ELSE NULL
  END AS InferredWeightKg,
    WHEN SAFE CAST(SUBSTR(WeightPounds, 0, STRPOS(WeightPounds, '.')) AS
FLOAT64) > 100 THEN SAFE CAST(SUBSTR(WeightPounds, 0,
STRPOS (WeightPounds, '.')) AS FLOAT64)
    WHEN SAFE CAST (Weightkg AS FLOAT64) < 150 THEN
ROUND (SAFE CAST (Weightkg AS FLOAT64) * 2.20462, 2)
    ELSE NULL
  END AS InferredWeightPounds
FROM `proyecto-final-
bellabet.proyecto final bellabet.weightLogInfo_merged`;
```

De momento no consideramos agregar datos adicionales debido a que no tenemos muy claro como se relacionan las variables y que datos enriquecerían nuestro análisis. No obstante, en las conclusiones abordaremos con que datos se potenciarían nuestras conclusiones.

Paso 4: Analizar

Paso	Acción
1	Descargar los datos del conjunto de actividad física de Fitbit y almacenarlos en una ubicación específica.
1	Importar todos los datos ya modificados a Rstudio para realizar ahí el procesamiento
2	Preparar en Rstudio todos los paquetes necesarios para mi proceso de análisis
4	Construir al cliente promedio de acuerdo con los datos disponibles.
5	Identificar correlaciones de variables importantes
6	Estratificar datos en dos tipos de clientes objetivos

Con vistas a realizar el análisis de datos lo primero que se buscaba era importar las bases de datos que fueran necesarias para el procesamiento. Para eso se procedió a ubicarlas en RStudio. Primero se instalaron todos los paquetes que iban a ser necesarios para el trabajo.

```
install.packages("tidyverse")
install.packages("lubridate")
  install.packages("dplyr")
  install.packages("ggplot2")
  install.packages("tidyr")
install.packages("corrplot")
```

Luego se procedió a activar todas las librerías que debían estar disponibles:

```
library(tidyverse)
library(lubridate)
  library(dplyr)
library(ggplot2)
  library(tidyr)
```

```
library(corrplot)
```

Se importaron los archivos CSV seleccionados a mi ecosistema de estudio y se incorporaron a las bases de datos de ese proyecto:

```
dailyActivity <- read_csv("dailyActivity_merged.csv")
    Calories <- read_csv("hourlycalories_merged.csv")
Intensities <- read_csv("hourlyIntensities_merged.csv")
    Sleep <- read_csv("sleepDay_merged.csv")
Weight <- read csv("weightLogInfo merged.csv")</pre>
```

Posteriormente se construyó una nueva tabla que reflejara el promedio de cada Id lo que posteriormente nos permitiría analizar a cada miembro de la muestra:

Listo, ahora existe los datos de como se comportaron los elementos de la muestra a lo largo de esos 31 días. Este dato nos será de mucha relevancia porque nos permitiría caracterizar a cada individuo e inferir grupos objetivos de acuerdo con sus hábitos medios.

De igual forma me resultaba relevante analizar los promedios de cada variable así vemos como se comporta el grupo de control de conjunto. Para ello usamos el siguiente comando:

```
promedio_por_variable <- promedio_por_id %>%
    summarise_all(mean, na.rm = TRUE) %>%
    select(-Id, -ActivityDate)
```

Una matriz de correlaciones nos ayudaría a darle un enfoque muy visual a las relaciones que existen entre algunos elementos lo que posteriormente nos ayudaría en la etapa de visualización a encontrar que variables tienen una relación fuerte y por tanto habría indicios de causalidad en los que se podría incidir. Con el objetivo de construir esta matriz ejecutamos:

La matriz nos sugiere algunas relaciones que luego comprobaremos en la etapa de visualización. De momento vamos a continuar calculando algunos datos promedios como podría ser los minutos de sueño y de tiempo en la cama:

```
promedio_minutos_sueno <- Sleep %>%
summarise(promedio_minutos_sueno = mean(TotalMinutesAsleep),
    promedio_tiempo_en_cama = mean(TotalTimeInBed))
```

Y por último vamos a realizar una estratificación en función de la cantidad de calorías quemadas como promedio por cada individuo. Este proceso nos va a permitir segmentar a la muestra con el objetivo de enfocar nuestros esfuerzos, productos o servicios específicos en un grupo en función de sus hábitos. Para ello ejecutamos el siguiente código:

Descubrimientos luego de la exploración:

- La persona promedio de este estudio recorre 5.4 km por día, quema casi 2300 calorías por día y tiene alrededor de 224 minutos activos al día (casi 3.7 horas). En contraposición, mantiene una actividad sedentaria 999 minutos como promedio al día (16.65 horas), duerme unos 419 minutos como promedio (6.98 horas) y pasa en la cama 458 minutos como promedio (7.6 horas).
- El promedio de calorías quemadas se encuentran en el rango aceptado por la USDA (https://www.myplate.gov/) que lo ubica entre 2000 y 2500 calorías al día.
- La estratificación por la cantidad de calorías quemadas como promedio refleja que la mayoría de las personas se ubican en los extremos. El 40% se podrían calificar como sedentarios donde no se queman la cantidad de calorías necesarias, 36% se encontrarían en el lado opuesto sobre cumpliendo la cantidad de calorías recomendadas y el 24% restante está en el rango aceptado.
- La estratificación nos permite identificar dos grupos objetivos importantes relacionados con personas que no se ejercitan lo suficiente y otro grupo de personas que sí lo hacen.
- Por otro lado, el análisis descubrió que existen como promedio 39 minutos que las personas permanecen en la cama sin dormir.
- Como promedio las personas caminan 7519 pasos al día bastante lejos de los 10000 que como promedio recomiendan los especialistas. Según estudios realizados los adultos que dan 10000 o más pasos al día como promedio tienen 31% menos de riesgo de mortalidad y 25% menos de desarrollar enfermedades cardiovasculares (Lee, 2018) (Wen, 2019) Sin duda algo que tener en cuenta para enfocarse.

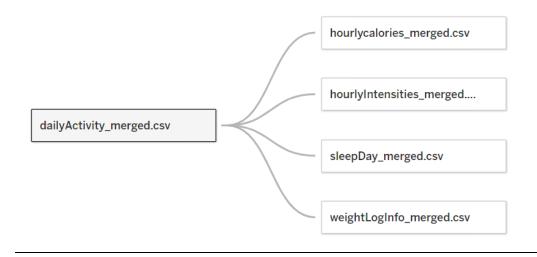
Paso 5: Visualizar

Paso	Acción
1	Ubicar todas las tablas ya procesadas en Tableau
1	Establecer las relaciones entre ellas utilizando una clave principal
2	Graficar relaciones directas entre variables
4	Observar la distribución del ejercicio a lo largo de las horas del día
5	Observar la distribución del ejercicio a lo largo de los días de la semana

Lo primero que era necesario para llevar mis datos a Tableau era exportar todos aquellos datos que ya había manipulado en RStudio para que mi información fuera consistente. Para esto utilicé el siguiente código:

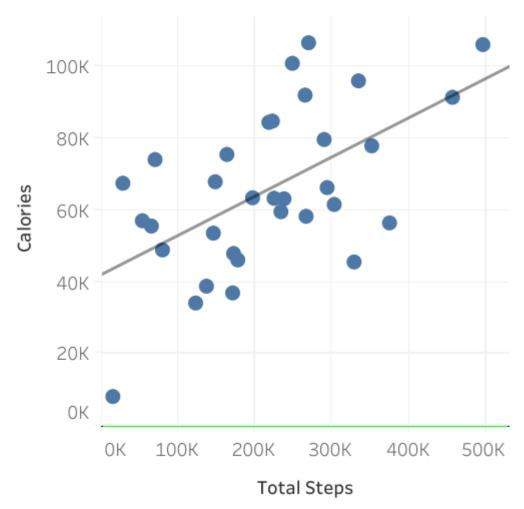
```
write.csv(dailyActivity, "dailyActivity.csv", row.names = FALSE)
    write.csv(Calories, "Calories.csv", row.names = FALSE)
    write.csv(Intensities, "Intensities.csv", row.names = FALSE)
    write.csv(Sleep, "Sleep.csv", row.names = FALSE)
    write.csv(Weight, "Weight.csv", row.names = FALSE)
```

Posteriormente se procede a cargar los datos en Tableau y establecer las relaciones lógicas entre las tablas utilizando como clave principal las columnas Id y Activity Date existentes en cada tabla producto de nuestra manipulación previa:



Ahora procederemos a graficar algunas relaciones entre variables para corroborar que cómo nos había indicado la matriz de correlación efectivamente la existencia de una variable (o no) está relacionada con otra.

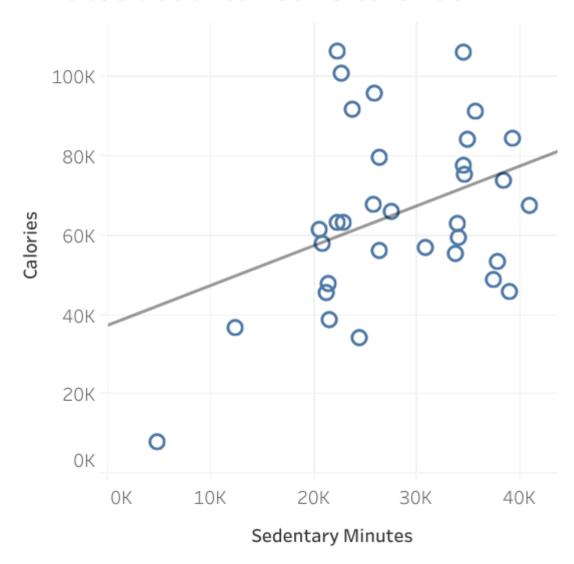
Relación entre la cantidad de pasos y las calorías



Acá se evidencia algo bastante obvio: a mayor cantidad de pasos se logran quemar más calorías por lo que una buena alternativa para la aplicación de Bellabeat sería recordarles a las personas que no han cumplido con su objetivo de pasos diarios y como eso se relaciona con sus calorías quemadas.

A continuación, se observa el gráfico que relaciona la cantidad de minutos sedentarios con las calorías quemadas. La lógica indicaría que una mayor cantidad de minutos sin realizar actividad física causaría que tu metabolismo se ralentizara y que por tanto quemaras menos calorías luego. Veamos que nos dicen los datos:

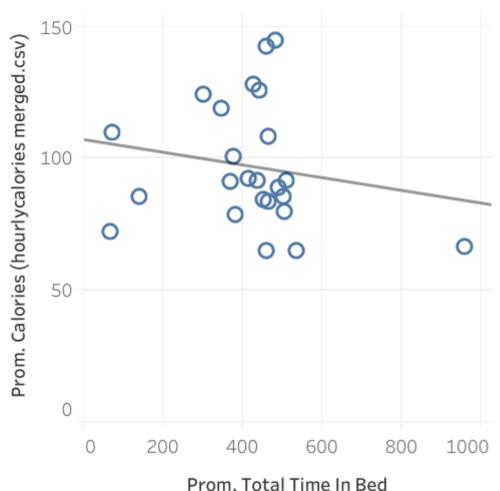
Minutos Sedentarios vs Calorias



El gráfico se observa como sorprendente. Muchas personas con un alto volumen de minutos sentados logran quemas altos volúmenes de calorías por lo que se podría llegar a la conclusión de que el tener un trabajo mayormente sedentario no debe ser una excusa para no realizar actividad física o alcanza tu objetivo de calorías.

Luego analizaremos la relación entre el tiempo promedio en cama con el promedio de calorías quemadas. La lógica indica que personas que pasan más tiempo en la cama quemarán menos calorías. Vamos a ver que se puede observar.

Sueño vs Calorías

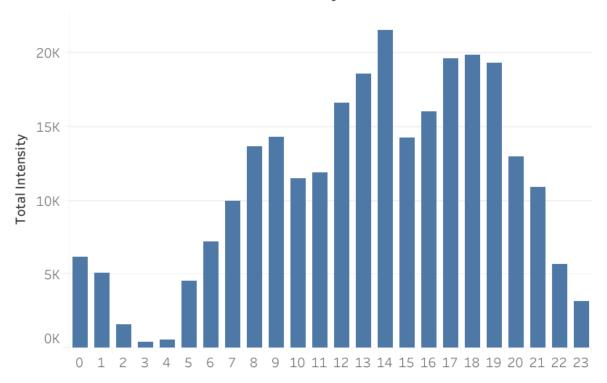


Prom. Total Time in Bed

Efectivamente existe una relación negativa entre los datos. Por lo general las personas que están más tiempo en la cama queman menos calorías, pero no creo que deba ser tomado como causalidad sino como una caracterización completa de un subgrupo menos activo físicamente y que por tanto prefiere pasar más tiempo en cama.

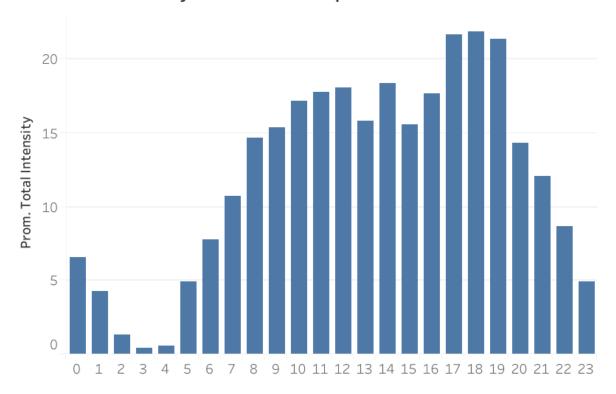
A continuación, sería interesante ver a qué hora nuestros sujetos concentran su ejercicio y en que días de la semana. Para ello vamos a usar un histograma y un gráfico de barras para observar como se distribuye el ejercicio.

Distribución de las horas de mayor intensidad



La distribución de intensidad total demuestra un valor máximo importante a las 14PM y una concentración de 17PM a 19PM. Observemos la distribución media para poder sacar más conclusiones.

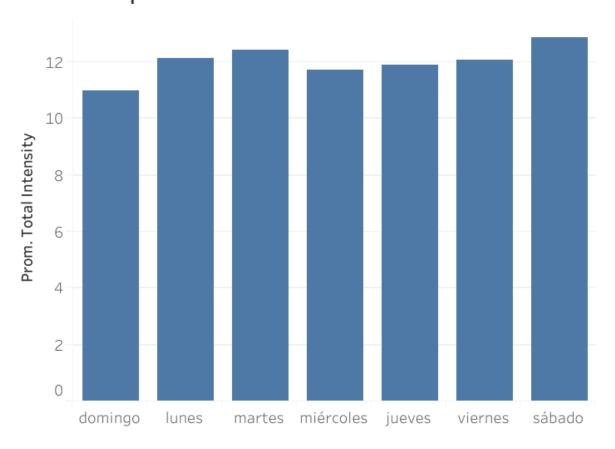
Promedio de Mayor intensidad por Hora



Cuando promediamos la intensidad para obtener un valor suavizado observamos que efectivamente la mayor cantidad de ejercicio se concentra en ese rango horario de la tarde. Esto tiene mucho sentido debido a que normalmente a esta hora se concentran las salidas del trabajo y las personas suelen ir al gimnasio. Así mismo vemos que no hay mucha actividad entre las 6AM y las 7AM (otro horario usado a menudo como bueno para asistir al gimnasio). De esto se deslindan dos estrategias para la empresa: notificar a sus clientes cuando se esté acercando el horario de la tarde y que recuerden hacer su ejercicio o incentivarlos a aprovechar también el horario de la mañana.

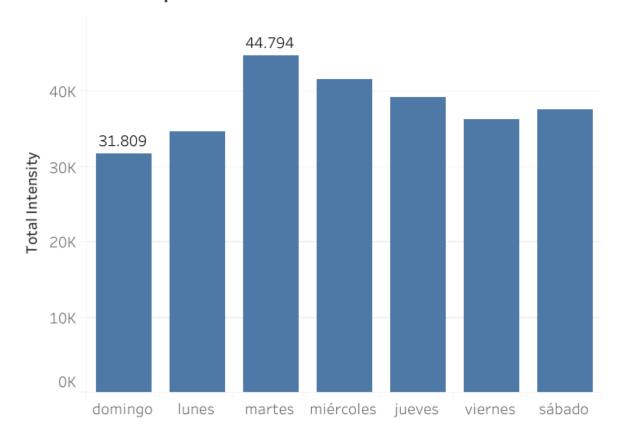
A continuación, vamos a ver como se distribuye la intensidad del ejercicio entre los días de la semana. Lo esperado sería observar una concentración total y promedio entre semana con el domingo como día menos activo (Algo similar a una campana de Gauss).

Promedio por día de la semana



A primera vista observamos que el domingo es el día que menos activo como promedio y el sábado el que más. Veamos que se observa con el gráfico de valores totales y luego saquemos conclusiones.

Distribución por día de la semana



Acá observamos que el domingo sigue siendo el día con menor intensidad total pero acá el martes el que más (algo con un poco más de lógica). Sorprende ver el lunes como el segundo peor día y definitivamente algo que la empresa podría explotar. Se podría lanzar un mensaje a sus clientes similar a: "Ayer (domingo) tu intensidad fue de x, normalmente este es el día menos activo para ti. ¡Y si rescatamos el lunes y vamos a por nuestro objetivo semanal desde pronto!". De esta forma los clientes se sentirían motivados a comenzar su ejercicio desde el lunes gracias a las recomendaciones personalizadas de su aplicación de Bellabeat.

Paso 6: Actuar

Luego de realizar todo el análisis exploratorio del caso, manipular los datos, encontrar correlaciones y visualizar los resultados se pueden llegar a conclusiones que sirvan al equipo de desarrollo de Beallabeat a identificar su público objetivo, segmentar sus clientes y encontrar factores que hagan sus servicios más personalizados y útiles:

 De forma general se pueden utilizar la cantidad de pasos y las calorías para segmentar a los clientes de acuerdo con su actividad física. Así mismo, dos buenos puntos de referencia para llamar la atención de sus clientes pueden ser cumplir

- con su objetivo de 10000 pasos al día y de 2500 calorías. Esto puede representar un llamado de atención a los clientes para que enfoquen sus esfuerzos y contribuyan a lograr una vida más saludable de la mano de los productos de Bellabeat.
- Existe un promedio de 39 minutos que los clientes están en la cama y no están durmiendo. Esto se podría explotar desde varias vías: se podría recordar a los clientes con una notificación que es hora de dormir en un momento determinado, si esa persona pasa tiempo en la cama sin dormir porque dedica ese espacio a leer o estudiar se podría incorporar en la aplicación un servicio de lectura que motive a las personas a estudiar mientras hacen ejercicio contribuyendo así a simultanear tareas o se les podría recordar periódicamente la importancia de dormir bien y premiarlos de alguna forma si cumplen este objetivo.
- Se podría dar un tratamiento diferenciado a clientes categorizados como sedentarios y a clientes categorizados como muy activos.
- Existen relaciones marcadamente directas entre variables y sería interesante determinar causalidad (Ejemplo: sueño y calorías quemadas).
- La mayoría de los clientes concentran su actividad física en la tarde (17PM-19PM) por lo que focalizar la atención ahí sería beneficioso para ellos al recordarle que deben realizar su ejercicio diario.
- El otro rango horario comúnmente utilizado para hacer ejercicio (6AM-8AM) no está tan explotado por lo que se podría premiar de alguna forma a los clientes que realicen su ejercicio a esa hora.
- Sería interesante complementar el caso práctico con datos de nutrición que enriquezcan el análisis y nos de una idea de cuántas calorías ingieren esos clientes y en consecuencia cuál es su balance calórico del día.
- El lunes es un día que se ejercita poco (probablemente por la pereza que provoca el comenzar la semana) por lo que si se aumentaran los minutos de ejercicio ese día se podría contribuir en gran medida a elevar los promedios de calorías quemadas y pasos.
- Al ser el público objetivo de la empresa las mujeres sería interesante saber que parte de la muestra de la base de datos correspondía a ese género y por tanto realizar un análisis mucho más preciso.

Mensaje Clave para la Campaña Online de Bellabeat

Como empresa entendemos tus necesidades y estamos dispuestos a acompañarte en tu camino hacia una vida más sana. Nuestra aplicación no es una aplicación de seguimiento de ejercicio más pues nos enfocamos en empoderar a las mujeres a balancear una vida profesional exitosa con hábitos de vida más saludables. Para ello nos valemos de recomendaciones personalizadas, un coach de salud online y una interrelación entre todos nuestros productos enfocados 100% en la mujer del siglo XXI.

Referencias

- Lee, I.-M. S. (2018). Estudio sobre la relación entre los pasos diarios y la mortalidad por todas las causas. Association of Daily Steps With All-Cause Mortality in Healthy Adults: A Meta-Analysis. JAMA Internal Medicine, 178(1), 112-119.
- Wen, C.-P. L. (2019). Estudio sobre la relación entre los pasos diarios y la incidencia de enfermedades cardiovasculares. *Association of Daily Steps With Incident Cardiovascular Disease: A Meta-Analysis of Prospective Cohort Studies. Circulation*, 140(10), 852-862.