Data Journal Case Study 2: How can a wellness technology company play it smart?

El objetivo principal de este caso de estudio es dar respuesta a la pregunta planteada por la jefa de la oficina creativa y cofundadora Urška Sršen, la cual fue: ***Analizar la información proporcionada por los dispositivos inteligentes para sacar conclusiones sobre el uso por parte de los usuarios para orientar una mejor estrategia de marketing.***

- **Ask**

Durante está primera etapa se busca encontrar el objetivo y el porqué de realizar este análisis. Esto para orientar de mejor manera los resultados.

Las preguntas que se buscan responder con este análisis son:

- Cuáles son algunas tendencias del uso de dispositivos inteligentes
- Como esas tendencias pueden aplicar sobre los usuarios de Bellabeat
- Como esas tendencias ayudarían a influenciar la estrategia de marketing de Bellabeat.

Ahora con el fin de acotar este análisis, la directora de la oficina creativa nos indico orientarnos en analizar información de uso de productos NO Bellabeat y aplicar estas conclusiones sobre las posibles estrategias de marketing.

Está pregunta y el objetivo planteado permiten decir que el tipo de problema es uno orientado a la realización de predicciones y de realizar conexiones. Desde cierto punto de vista puede verse la búsqueda de patrones entre los usuarios, también con fines de predicción.

Summary Table:

Step	**ASK**
Guiding Questions	**Cual es el problema que se trata de resolver** El problema que se busca resolver es que a traves del analisis de datos de fuentes externas no relacionadas a la compañía se extraigan insights que ayuden a mejorar la campaña de marketing de está y enfocar hacia los usuarios según sus necesidades y los mayores usos que ellos puedan darle a los productos de Bellabeat. **Cómo pueden sus conocimientos impulsar decisiones comerciales** Los insights que se puedan extraer desde este análisis podrían ayudar a guiar la toma de decisiones de marketing hacia los usuarios con un mayor potencial de utilizar los servicios, aumentar el número de miembros o incrementar las ventas de los productos. Esto a través del uso eficiente y localizados de los recursos de marketing de la compañía.

Key tasks	Identificar la **business task** Considerar a los **key stakeholders** Los principales stakeholders son: la jefa del área creativa y el equipo de marketing
Deliverables	**A clear statement of the business task** Realizar un análisis cuantitativo sobre datos de salud obtenidos por dispositivos no relacionados a la marca Bellabeat y aplicar las conclusiones obtenidas para una mejor orientación de los recursos de marketing y la generación de nuevas campañas publicitarias enfocadas en las categorías más consultadas por los usuarios de esos servicios y enfocar todo esto a uno de los productos de Bellabeat. Además, como enfoque secundario, el aumento de membresías y el incremento de la venta de productos inteligentes.

- **Prepare**

En está etapa se procesa la información procedente de la base de datos.
Los datos obtenidos provienen de una base de datos de acceso público. Está cuenta con datos provenientes de 30 sujetos seleccionados, identificados a través de un ID único para cada uno. Se registraron diversas mediciones de ejercicios, tales como; actividad diaria, ritmo cardiaco, intensidad del ejercicio, seguimiento del sueño, entre otras. Los sujetos fueron muestreados en un transcurso de 2 meses. Por eficiencia, los datos fueron separados en dos directorios abarcando un mes cada uno.

La información se descargó y se encuentra almacenada para su procesamiento en un directorio en Google Drive y, además, se almacenó una copia del archivo original en un servidor local.
La veracidad y confiabilidad de la información se sustenta en que la data fue obtenida mediante una base de datos anonimizada y de acceso público verificada. Está cumple con los requisitos **ROCCC (Reliable, Original, Comprehensive, Current, Cited).**

Al revisar cada directorio de la base de datos se obtiene que:

- El directorio comprendido entre Marzo y Abril, está conformado por 11 tablas.
- El directorio comprendido entre Abril y Mayo está conformado por 18 tablas.

Si bien, la jefa de la oficina creativa indicó que la información en estas tablas podría no ser suficiente y que se necesita aumentar la cantidad de información, con el propósito de acotar el análisis y que este sea más eficiente se utilizara solo está información.
Para este análisis se utilizaran los datos de:

- Daily Activity
- Intensity (hourly)
- Steps (hourly)
- Sleep (minute)

Se usarán tablas de ambos directorios para luego realizar una comparación entre los resultados obtenidos.
Para ordenar y filtrar se decidió subir la información a la plataforma Google BigQuery para realizar estas acciones.

Summary Table:

Step	PREPARE
Guiding Questions	**Donde está localizada la data** La data estaba localizada en una base de datos de público acceso. Además, se guardaron copias de estas en Google Drive y en un servidor local. Con el fin de ser procesadas también se subió a la plataforma de Google BigQuery. Todas las tablas están en formato .CSV **Cómo está organizada la data y en qué formato se encuentra** Se encuentran organizadas en tablas CSV en formato de columnas, donde cada sujeto tiene una fila con sus respectiva información. Se puede considerar una base de datos *externa, continua y estructurada*. Se pueden apreciar que existen ***primary y foreign key*** lo que permitirá conectar la información de las tablas. **Existe algún asunto con respecto a la credibilidad o sesgo** No **Cómo se trabaja con la licencia, privacidad, seguridad y accesibilidad a la data** La base de datos posee una licencia de acceso libre, la base de datos no está anonimizada pero los datos de los usuarios se encuentran protegidos ya que la identidad de ellos está bajo un ID que solo la empresa originaria de los datos y los usuarios conocen. **Cómo se verificó la veracidad de la data** Se verificó la fuente de la información y la veracidad del sitio web de donde se extrajo la información. Se determinó que es un sitio válido y de acceso público.Además, en el sitio web se informa como se obtuvo la información y el nivel de acceso que se puede tener a ella. **Como ayuda a responder la pregunta planteada** Es de gran ayuda ya que permite generar patrones de uso entre los diferentes usuarios y ver las tendencias de ellos. Otro foco es el poder ver sus preferencias y si estas se pueden luego aplicar sobre algún producto de Bellabeat. **Existe algún problema en la data** Si, existen vacíos de información que deben ser filtrados para un análisis más expedito.
Key tasks	Descargar la data y almacenarla correctamente. Identificar cómo se encuentra organizada. Ordenar y filtrar la data. Determinar el nivel de credibilidad de ella.
Deliverables	**A description of all data sources used** Las fuentes de información que se utilizaron fueron la base de datos de carácter público, donde los usuarios son identificados utilizando un ID.

- **Process**

En está etapa del análisis el enfoque se centrará en el de procesar la data con el fin de filtrar y ordenar más en detalle.

Para la realización de está etapa se continúa utilizando la herramienta de Google BigQuery, esto se justifica debido a la enorme cantidad de información contenida en la base de datos de ambos periodos mensuales, y utilizar SQL facilita el análisis de los datos.

Se realizó una revisión más completa de la información en la búsqueda de datos repetidos o información que no estuviera en línea con lo esperado o que estuviera fuera de lugar bajo el contexto que se está trabajando.

Para la optimización del trabajo se trabajaron las 8 tablas, 4 por cada periodo mensual, en grupos de dos. En el caso de las tablas “DailyActivity” se realizaron los siguientes etapas de procesado:

- Se buscaron filas repetidas para ambas tablas, en ambos casos no se encontraron filas con información repetida.
- Basados en está información el total de filas para cada una de ellas y con el propósito de realizar un análisis óptimo y utilizando solo los valores con datos completos, además, se buscaron los valores NULL y 0 en parámetros como tiempo de actividad, pasos y distancia recorrida. Ya que las tablas están categorizadas por días, la presencia de estos valores implica que los sujetos no registraron actividad esos días. En este casos las filas encontradas fueron:
  - Tabla Marzo - Abril:
    - Total de filas: 457 filas
    - Filas NULL o 0: 63 filas
    - **Total de filas completas: 394 filas**
  - Tabla Abril - Mayo:
    - Total de filas: 940 filas
    - Filas NULL o 0: 78 filas
    - **Total de filas completas: 862 filas**
- Finalmente, se crearon dos tablas nuevas, una con los datos sin valores NULL y otra que los tuvieran. Además, con el fin de optimizar y dirigir los resultados de estos análisis se utilizaron solo las columnas originales de: Id, date, tiempo de ejercicio, tiempo total, pasos, distancia, calorías y se calculó el porcentaje de actividad del día y se agregó como una nueva columna.

Por otro lado, las tablas “Intensity”, “Sleep”, “Steps”, ya que están destinadas a medir cada variable durante periodos de tiempo, lo que se realizó con ellas fue realizar resúmenes de ellas etiquetando para cada Id de cada sujeto.

- Para la tabla “Intensity” se calculó la intensidad promedio diaria para cada uno de los sujetos
- Para la tabla “Sleep” se calculó el tiempo que permaneció en cada uno de los estados de sueño cada sujeto en cada dia
- Para la tabla “Steps” se calculó el total de pasos dados por cada sujeto diario y el total de tiempo de actividad que estos pasos representan.

Independiente del procesamiento, en estos tres últimos casos se utilizaron funciones que permiten no considerar los valores NULL en los cálculos realizados.

STEP	PROCESS
Guiding Questions	**¿Qué herramienta se usó y por qué?** Se utilizó Google BigQuery como herramienta para la limpieza y filtrado de la base de datos. Esto se justificó debido a la gran cantidad de información contenida en ellas que de haber sido procesada en Google SpreadSheet o Microsoft Excel hubiera tardado más tiempo y a costa de una mayor cantidad de recursos digitales. **Se aseguró la integridad de la data?** Si, esto mediante el filtrado y limpieza de la data **¿Qué pasos se tomaron para asegurar la limpieza de la data?** Una re-revisión de los datos repetidos y los datos faltantes, los cuales no serán considerados para el análisis. Además, se realizó un resumen de datos de tablas donde la información se encontraba más en detalle lo que permitirá contrastar los resultados y tener un mejor entendimiento de los datos y las posibles conclusiones que se puedan obtener. **¿Cómo se verificó que la data está limpia y lista para su uso?** Lo primero es verificar si existen términos *null,* y ordenar la información para determinar si las etapas de filtrados fueron útiles. También se utiliza la data original con el fin de comparar el número de filas de información original y las filas de información luego del filtrado, viendo una reducción en el número de ellas. Este filtrado se realizó debido a que esa información no estaba completa o abarcaba días donde el ejercicio era considerado nulo, por lo cual no será utilizada está información dado el objetivo de este análisis. **¿Se aseguro de documentar su proceso de limpieza para que sea revisado y compartido?** Si, todos los procesamientos y cambios realizados se documentaron y se encuentran comentados en el código de cada una de las tablas de información.
Key tasks	Revisar la data en busca de errores Elegir la herramienta correcta para el procesamiento Transformar la data para que el trabajo sea eficiente Documentar el proceso de limpieza
Deliverables	**Documentation of any cleaning or manipulation of data** Todo lo realizado fue descrito en términos generales al inicio de este encabezado y las funciones utilizadas así como los pasos realizados, se encuentran comentados en el código correspondiente a cada tabla de datos. Además, de presentar este documento con descripciones generales y las principales conclusiones de está etapa.

- **Analyze**

Es en está etapa donde se realizan los análisis propiamente tal. Aquí se realizan cálculos basados en la información ordenada, recolectada y los primeros cálculos realizados en las etapas anteriores.

Se continúa el trabajo de la información en Google BigQuery.

Lo primero que se realiza sobre la data es unir las 4 tablas de cada periodo en una sola donde se muestre toda la información para luego realizar cálculos más complejos y determinación de variables. Para unir estas tablas se utilizó la función INNER JOIN. Al finalizar está etapa se generaron 2 tablas con toda la información, una para cada periodo de tiempo. Un dato importante a mencionar es que luego de las etapas de filtrado y la unión de estas tablas los usuarios elegibles bajaron a **18 usuarios** que sus datos pueden ser luego comparados entre los periodos descritos.

Luego con el fin de obtener una mayor comprensión de los datos obtenidos se calculó una puntuación de actividad que corresponde a la suma ponderada de los pasos realizados por cada sujeto y el tiempo de actividad diaria realizada. La ecuación que se utilizó para ello fue: $pasos * 0.6 + actividad * (0.4/60)$, se divide en 60 dado que el tiempo de actividad se encuentra en minutos. Está puntuación se utilizó para obtener los días de mayor actividad física de los usuarios.

Otros valores previamente calculados fueron el porcentaje de tiempo de actividad física. Se calculó utilizando el tiempo de actividad (intenso, medio y liviano) y la cantidad de tiempo total medida para cada día.

Adicionalmente, se calcularon los valores mínimo, máximo, suma, promedio, desviación estándar, de todos los datos de las tablas calculadas para obtener un panorama general y establecer recomendaciones basadas en datos provistos por entidades especializadas.

Todos estos análisis fueron realizados en la plataforma de Google Bigquery, en específico las funciones diferenciadores utilizadas para cada una de las situaciones fueron:

- Cálculo de los días con mayor frecuencia de actividad física: ROW NUMBER() y PARTITION()
- Cálculo de estadísticas básicas: UNPIVOT(), CAST(), UNION ALL()
- Cálculo del porcentaje de tiempo de actividad física: SAFE_DIVIDE () y UNNEST()

Posterior a este análisis, las tablas fueron exportadas a Google Sheet donde se analizaron y prepararon para ser procesadas para la generación de visualizaciones. Es en está etapa donde se resumieron los resultados para obtener mayores insight. En está oportunidad se utilizaron diversas funciones de Google Sheet con este fin. Las funciones utilizadas fueron: SUMAR.SI, CONTAR.SI.CONJUNTO, TEXTO, DIASEM.

STEP	**ANALYZE**
Guiding Questions	**¿Cómo deberías organizar tus datos para realizar análisis sobre ellos?** Todos los datos utilizados se organizaron siempre manteniendo la business task en mente, es decir, sacar insight con el fin de extraer información del uso de los

	dispositivos para orientar la campaña de marketing y así enfocar los recursos disponibles. **¿Tus datos están correctamente formateados?** En general si tenían el formato correcto, salvo que fue necesario realizar un CAST sobre algunas variables para que fueran utilizables para el cálculo de las variables estadísticas básicas. Además, se utilizó solo la fecha y no la hora en las columnas que contenían está información, por que se utilizó la función DATE para extraer está parte de la información. **¿Qué sorpresas descubriste en los datos?** A raíz de las estadísticas básicas podemos extraer algunas conclusiones generales en términos promedios: ➔ Existe un aumento en el número de pasos promedios, de 446 pasos aprox, de un periodo a otro. ➔ La intensidad de los ejercicios presentó un leve aumento, de 0.217 a 0.228 ➔ El tiempo de actividad presentó una disminución, de aproximadamente 400 min a 265 min. En ambos casos se encuentran en línea con lo recomendado por la OMS (link). ➔ Las calorías quemadas igualmente tienen una leve disminución, de 2337 a 2261. Otra conclusión es la establecida a los días de semana con mayor tráfico de datos. En términos generales son días de semana, pero más específicamente: ➔ Periodo Marzo - Abril: Viernes y Martes. ➔ Periodo Abril - Mayo: Miércoles, Lunes y Viernes. Además, podemos concluir que los fines de semana tienden a tener una disminución en el registro de actividad física. **¿Qué tendencias o relaciones encontró en los datos?** Las tendencias que se identifican son: ➔ Existe una tendencia hacia la baja en los tiempos de actividad a pesar de un leve aumento de la intensidad, no fue considerado significativo. ➔ Esto también se puede explicar con la tendencia a la baja en calorías quemadas. Una explicación es que a pesar del aumento de pasos dados la intensidad de ellos puede que haya disminuido. ➔ Los datos muestran que los periodos donde se presenta un mayor numero de registro de actividad son los días viernes. Tal Vez con motivo de jornadas laborales más cortas que permiten comenzar antes el entrenamiento. ➔ El bajo registro los fines de semana, probablemente asociados al descanso de los usuarios, evitando el uso de estos dispositivos de registro. **¿Cómo le ayudarán estos conocimientos a responder las preguntas de su empresa?** .En términos generales, creo que serían útiles ya que nos permite tener un panorama general del uso de estos dispositivos por los usuarios y orientar las campañas de marketing en los días de mayor uso, así como desarrollar más los productos asociados a las características que necesitan una mayor precisión en su medición como lo son la intensidad de la actividad, pasos y calorías quemadas.

Key tasks	Agregue sus datos para que sean útiles y accesibles Organiza y formatea la data Realizar cálculos o procedimientos Identificar tendencias y relaciones
Deliverables	**Resumen del análisis** En términos generales, se obtuvieron varios resultados a los cálculos realizados, se generó información útil y eficaz para la toma de decisiones y la información obtenida se considera valiosa para la toma de ellas.

- **Share**

En está etapa se generan las visualizaciones necesarias para la correcta y eficiente muestra de resultados del análisis realizado previamente.

En está oportunidad las visualizaciones fueron realizadas en Google Sheet, mediante la utilización de tablas dinámicas para luego ser exportadas a Google Slides y luego descargadas en formato pdf para ser presentadas de la mejor manera posible.

Las visualizaciones generadas tienen como fin el mostrar los resultados que avalan y sostienen las conclusiones mostradas en la sección anterior. Estas visualizaciones están conformadas por tablas y gráficos.

Finalmente, se generan 3 set de gráficas. El primer set resume los cambios promedios experimentados en las variables básicas asociadas al tiempo de actividad, calorías quemadas, intensidad y pasos realizados, además se muestran insights asociados a está gráficas. El segundo y el tercer set muestra la relación que se establece entre los tiempos de actividad, la cantidad de actividad registrada y la puntuación que se generó entre las variables para que sean comparables. Aquí también se mencionan algunos insight que se consideraron relevantes.

Con el fin de tener un respaldo, en el repositorio se subieron las tablas que respaldan la información mostrada en los gráficos.

STEP	**SHARE**
Guiding questions:	**¿Fue posible dar respuesta a la pregunta de negocios?** Considerando la business question ha responder, considero que la información proporcionada en este análisis ayudará a responder está pregunta. Esto se puede destacar en los patrones de uso de los usuarios, los días de mayor afluencia y las conclusiones que se pueden extraer a partir de las métricas básicas calculadas. **¿Qué historia cuentan sus datos?** La historia postrada en los datos se basa en la idea de determinar los patrones de uso de los usuarios de aplicaciones móviles que les ayuda a registrar su actividad física. Los datos muestran que en un inicio los usuarios presentan un gran entusiasmo en mediar su actividad física, pero con el tiempo eso disminuye pero no necesariamente la intensidad de la actividad. Una posible explicación de ello es que los usuarios se enfocan en ejercicios puntuales, o la forma en que se mide los datos no puede ser tan intuitiva.

	Por otro lado, la utilización de estos dispositivos en los diferentes días de semana tienden a ser mayores los días viernes. Una posible explicación es el inicio del fin de semana o una menor jornada laboral que implicaría una mayor disponibilidad de ejercitación. La historia que muestran los datos hace que este análisis sea interesante y aplicable con el objetivo de este. Pero, contrario a lo pensado originalmente, podemos coincidir con la jefa de la gerencia creativa en la necesidad de mayor información o la incorporación de una base de datos más detallada y orientada al uso específico de dispositivos smart. **¿Cómo se relacionan sus hallazgos con su pregunta original?** Se relacionan directamente, este análisis nos permite tener una mayor perspectiva de la situación de los usuarios y permitirá orientar los recursos al desarrollo de marketing o el desarrollo de nuevos dispositivos smart. Se sugiere orientar los recursos en dispositivos asociados a medir de mejor manera la intensidad y la medición de pasos y su asociación con las calorías quemadas durante está actividad. **¿Quién es su audiencia? ¿Cuál es la mejor manera de comunicarse con ellos?** La audiencia es la jefa del área creativa y el equipo de marketing. La mejor manera de comunicar estas conclusiones es mediante una reunión informativa y un posterior reporte con las conclusiones y principales insights de este análisis. **¿Puede la visualización de datos ayudarlo a compartir sus hallazgos?** Si ya que es una presentación intuitiva y fácil de entender. Está posee graficas, datos numericos y representaciones graficas de los datos **¿Su presentación es accesible para su audiencia?** Si, fue construida pensando en los diferentes espectros de color y como se asocian cada uno de ellos.
Key Task	Determinar la mejor manera de compartir lo obtenido Crear visualizaciones efectivas Presentar los hallazgos Asegurarse de que el trabajo es accesible
Deliverable	**Supporting visualizations and key findings** Presentación de los resultados y las principales conclusiones de ellos se muestran al final de la presentación

- **Act**

En está última etapa es donde se presentan las conclusiones finales y se realizan las sugerencias de la implementación de la business task asociada al análisis.

Con fines de facilitar el acceso a la información y a todo el contenido asociado a este análisis, la información, tablas, códigos, visualizaciones, entre otros datos, fueron subidos a GitHub para que puedan ser estudiados, comentados y revisados por otros usuarios.

STEP	ACT
Guiding questions:	**¿Cuál es su conclusión final en función de su análisis?** Las conclusiones generales que se pueden obtener del análisis son: los días de mayor actividad registrada suelen ser días de semana pero con puntuaciones bajas, esto debido a la disminución en tiempos de actividad. Esto puede ser asociado a los horarios laborales de los días de semana pero, se puede extraer de está situación que los usuarios utilizan con mayor regularidad sus dispositivos esos días más que otros días. Otra conclusión a resaltar es el hecho de la disminución de los tiempos de actividad de un periodo a otro pero, con un leve aumento de intensidad y quema de calorías. Esto tal vez explicado por la orientación de los usuarios a ejercicios de menor duración pero con un aumento en la intensidad y gasto energético. **¿Cómo podrían su equipo y su empresa aplicar sus conocimientos?** La mejor manera de aplicar estos resultados es a través de dos estrategias. La primera consistirá en orientar los recursos de marketing los días que presenten mayor registro de actividades independiente del tiempo de uso esto ya que los usuarios revisan sus dispositivos y estarán más atento a la publicidad esos días. La segunda es orientar el desarrollo e investigación de productos smart orientados al registro de actividades de alta intensidad mejorando su registro y su facilidad para los usuarios. **¿Qué próximos pasos darían usted o sus partes interesadas en función de sus hallazgos?** Los próximos pasos a realizar están dirigidos a basar el marketing de la compañía los días en que los usuarios presentan un mayor uso de los dispositivos y esperar feedback de ellos para ir adaptando está campaña para tornarla siempre más efectiva. Y aumentar el desarrollo de dispositivos smart de medición de ejercicios de alta intensidad y quema de calorías. **¿Existen datos adicionales que podría utilizar para ampliar sus hallazgos?** Si, otros datos que podrían ser útiles para la mejor toma de decisiones están las edades de los usuarios del servicio, el propósito con el que ellos lo utilizan y el tipo de dispositivo que utilizan para realizar la medición.
Key Task	Crear un portafolio Agregar el caso de estudio Practicar la presentación de los resultados obtenidos
Deliverable	**Your top three recommendations based on your analysis** Basado en los datos analizados, mis tres principales recomendaciones son: ● Enfocar el marketing los días de mayor utilización de los dispositivos, en este caso: miércoles y viernes. Una vez implementada, esperar feedback de los usuarios y adaptarla. ● Orientar en la mejora y desarrollo de nuestros dispositivos diseñados para actividades de alta intensidad ya que estos ejercicios parecen ir al alza. ● Recabar una mayor información para mejorar el análisis y proyectar nuevas formas de mejorar el marketing y el desarrollo de tecnologías.