PROYECTO PARA LA EMPRESA: BELLABEAT

Juan Mario Moreno Chaparro
Proyecto de analista de datos

Índice

Introducción

Realizaremos un análisis de datos, mediante los datos públicos de Kaggle, en base a esto utilizaremos las herramientas siguientes, Excel, Google Sheets, BigQuery SQL, Rstudio, Tableu, con ello podremos limpiar, organizar y transformar si así lo requiere el conjunto de datos que hemos descargado para su análisis.
La empresa nos pedirá analizar estos datos para responder preguntas o averiguar tendencias en cuanto a sus productos.

Acerca de la empresa:

"Urška Sršen y Sando Mur fundaron Bellabeat, una empresa de alta tecnología que fabrica productos inteligentes focalizados en el cuidado de la salud. Sršen usó su experiencia como artista para desarrollar una tecnología con un bonito diseño que informará e inspirará a las mujeres de todo el mundo. Recopilar 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".

"En 2016, Bellabeat ya había inaugurado oficinas en todo el mundo y lanzado múltiples productos. Los productos Bellabeat pasaron a estar disponibles en línea a través de un creciente número de comerciantes minoristas además del canal de comercio electrónico propio de Bellabeat en su sitio web. La empresa invirtió en medios publicitarios tradicionales, como radio, cartelería en la vía pública, prensa gráfica y televisión, pero se centra mayormente en el marketing digital. Bellabeat invierte todo el año en Google Search, mantiene activas las páginas de Facebook e Instagram e interactúa de manera constante con los consumidores en Twitter. A su vez, Bellabeat publica anuncios por video en YouTube y avisos publicitarios en Red de Display de Google para apoyar las campañas en fechas de marketing claves".

“Analices 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 selecciones un producto Bellabeat para aplicar estos conocimientos en tu presentación”.

Preguntas:

¿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?

¿Cuál es el problema que intentas resolver?

El punto donde las tendencias están a favor de la empresa, con ayuda de las herramientas básicas para un analista de datos.

¿Cómo tus conocimientos pueden impulsar las decisiones empresariales?

Sabiendo que los datos son por el uso de productos de la empresa, y que se han ido recolectado por un tiempo, los datos pueden decir tendencias, puntos clave, desastres en el futuro, productos que no son muy usados por alguna causa, etc.

- Los interesados son Urška Sršen: Cofundadora y directora creativa de Bellabeat. Sando Mur: Matemático y cofundador de Bellabeat, miembro clave del equipo ejecutivo de Bellabeat..

- Ella le pidió al equipo de análisis computacional de datos de marketing que se concentrara en un producto Bellabeat y analizara los datos de uso de dispositivos inteligentes para conocer cómo las personas están usando sus dispositivos inteligentes. Después, con esta información, le gustaría recibir recomendaciones de alto nivel sobre cómo estas tendencias pueden colaborar en la estrategia de marketing de Bellabeat.

- En resumen, pide que se demuestren tendencias que ayuden a las estrategias de marketing

Preparación de los datos

Los datos vienen en archivos separado por comas, lo cual hay que limpiar debido a que el tipo de dato fecha no es del todo reconocido en su columna, ya que hay datos de tipo fecha y datos de numero entero, que asimilar ser una fecha, pero el software de Excel no reconoce.

Así que usamos la herramienta en la pestaña de datos llamada “Texto en columnas” y organizamos como debe de ser, ósea en el formato “M/D/A” (Mes/Día/Año).

El siguiente problema es el formato de fecha con hora, ya que los datos no cumplen un formato, hay que separarlos, para poder hacer un análisis correcto.

Teniendo fecha y hora en dos columnas distintas.

1503960366	05/12/2016 13:00
1503960366	05/12/2016 14:00
1503960366	05/12/2016 15:00
1503960366	05/12/2016 16:00
1503960366	05/12/2016 17:00
1503960366	05/12/2016 18:00
1503960366	05/12/2016 19:00
1503960366	05/12/2016 20:00
1503960366	05/12/2016 21:00
1503960366	05/12/2016 22:00
1624580081	4/13/2016 12:00:00 AM
1624580081	4/13/2016 1:00:00 AM
1624580081	4/13/2016 2:00:00 AM
1624580081	4/13/2016 3:00:00 AM
1624580081	4/13/2016 4:00:00 AM
1624580081	4/13/2016 5:00:00 AM
1624580081	4/13/2016 6:00:00 AM

Imagen 1: Propia

Los datos de fecha vienen de esta manera, lo cual es incorrecto y nos puede generar problemas en el futuro, por el echo de tener el espacio, el orden, y el tipo de nomenclatura con la hora y minutos.

Limpiando y ordenando los datos se verían de la siguiente forma:

Id	ActivityDate	ActivityMinutes	Intensity
1503960366	12/04/2016	12:00:00 p. m.	0
1503960366	13/04/2016	12:01:00 p. m.	0
1503960366	14/04/2016	12:02:00 p. m.	0
1503960366	15/04/2016	12:03:00 p. m.	0
1503960366	16/04/2016	12:04:00 p. m.	0
1503960366	17/04/2016	12:05:00 p. m.	0
1503960366	18/04/2016	12:06:00 p. m.	0
1503960366	19/04/2016	12:07:00 p. m.	0
1503960366	20/04/2016	12:08:00 p. m.	0
1503960366	21/04/2016	12:09:00 p. m.	0
1503960366	22/04/2016	12:10:00 p. m.	0
1503960366	23/04/2016	12:11:00 p. m.	0
1503960366	24/04/2016	12:12:00 p. m.	0
1503960366	25/04/2016	12:13:00 p. m.	0
1503960366	26/04/2016	12:14:00 p. m.	0
1503960366	27/04/2016	12:15:00 p. m.	0
1503960366	28/04/2016	12:16:00 p. m.	0
1503960366	29/04/2016	12:17:00 p. m.	0

Imagen 2: Propia

Ya que los dato de fecha son más fáciles de entender, separándolo con el tiempo del día, ya que se puede medir mejor las tendencias que arrojan cada usuario.

- ¿Dónde se almacenan tus datos?

Los datos están almacenados en hojas de cálculo.

¿Cómo están organizados los datos? ¿Están en formato largo o ancho? Algunas hojas de cálculo tienen formato largo y ancho.

Están en su mayor parte en formato largo, y los archivos que sobran en formato ancho.

- ¿Hay problemas con el sesgo o la credibilidad de estos datos? ¿Tus datos son confiables, originales, integrales, actuales y citados (ROCCC)?

Ningún problema, todos son datos cuantitativos, con un id que identifica a una persona del conjunto sin importar su tiempo de actividad, descanso o calorías quemadas. Ósea que no hay un limite en cuanto a sus datos. Ni limitaciones como el sexo, nombre, o alguna discapacidad.

Sitio de los datos: "Möbius. (2016). FitBit Fitness Tracker Data. Retrieved January 12, 2023, from Kaggle.com website: https://www.kaggle.com/datasets/arashnic/fitbit"

Con la licencia siguiente: Creative Commons — CC0 1.0 Universal. (2023). Retrieved January 12, 2023, from Creativecommons.org website: https://creativecommons.org/publicdomain/zero/1.0/

- ¿Cómo estás abordando la autorización, la privacidad, la seguridad y la accesibilidad?

Todo en un servidor local, ósea mi zona de estudio, hasta que termine el trabajo, y si lo pide el proyecto al final se eliminaran los datos usados.

- ¿Cómo verificaste la integridad de los datos?

Excel dice de donde vienen los datos, en este caso de una base de datos SQL, por parte de Kaggle y del equipo de Google.

- ¿De qué manera te ayuda a responder tu pregunta?

Los datos me dirán las tendencias de sus productos, en cuanto a calorías, tiempo, intensidad, sueño, etc.

- ¿Existe algún problema con los datos?

El tipo de fecha, sin un formato específico.

Procesar

- ¿Qué herramientas eliges y por qué?

  SQL, R, Excel, Tableu.

  Debido a que son de las mejores herramientas para hacer un análisis de datos bien hecho.

  SQL para operar con el gran volumen de datos y hacer combinaciones entre tablas y limpieza a gran escala, Excel es para dar una mínima limpiada o revisar datos de manera más pequeña, R para análisis más estadístico y verificar tendencias, tableu para mostrar visualmente a los gerentes o lideres de Bi.

- ¿Has garantizado la integridad de los datos?

  Sí, se ha almacenado esa información en la base de datos creada en BigQuery, los datos de Excel que fueron cambiados a formatos de fecha para su correcto funcionamiento.

- ¿Qué pasos seguiste para garantizar que tus datos están limpios?

  1.Descargar el conjunto de datos para su análisis

  2. Abrirlo en un editor como lo es Excel, para checar la forma en la que están los datos.

  3. Limpiamos en Excel los datos de tipo fecha ya que están en un formato no legible por BigQuery.

  4. Hacemos unas consultas en BigQuery para usar la información que necesitamos y poder pasar a un formato como Excel y poder hacer visualizaciones estadísticas en R y visualizaciones Bi para Tableu.

- ¿Cómo puedes verificar que tus datos están limpios y listos para analizar?

  Al hacer los modelos estadísticos darán un resultado, los datos deberán ser de un tipo de dato en las columnas.

La fuente de datos es confiable, directos de Kaggle, así que no hay mucho problema con usar esos datas de forma pública.

- ¿Documentaste tu proceso de limpieza para poder revisar y compartir estos resultados?

Los datos los subiremos a una base de datos. En Big Query tuvimos un problema el cual no reconocía el tipo de dato de la fecha, pero se arregló acomodando y transformando las hojas de cálculo con columnas nuevas, separando los tipos de dato date y date time..
Solo así se pudo resguardar la información en una base de datos.
Adjunto imagen:

Imagen 3: propia

Tenemos la base de datos en la cual hacemos las siguientes consultas para verificar que los datos cargados son los correctos.

```
SELECT * FROM `portafolio2023.dailyActivity_merged.dailyActivity_merged` LIMIT 1000
```

Está consulta arroja lo siguiente:

Fila	Id	ActivityDate	TotalSteps	TotalDistance	TrackerDistanc	LoggedActivitie	VeryActiveDista	ModeratelyActi	LightActiveDist	SedentaryA
1	1624580081	2016-05-01	36019	28.0300006...	28.0300006...	0.0	21.9200000...	4.19000005...	1.90999996...	0.01999999
2	1644430081	2016-04-14	11037	8.02000045...	8.02000045...	0.0	0.36000001...	2.55999994...	5.09999990...	
3	1644430081	2016-04-19	11256	8.18000030...	8.18000030...	0.0	0.36000001...	2.52999997...	5.30000019...	
4	1644430081	2016-04-28	9405	6.84000015...	6.84000015...	0.0	0.20000000...	2.31999993...	4.30999994...	
5	1644430081	2016-04-30	18213	13.2399997...	13.2399997...	0.0	0.62999999...	3.14000010...	9.46000003...	
6	1644430081	2016-05-03	12850	9.34000015...	9.34000015...	0.0	0.72000002...	4.09000015...	4.53999996...	
7	2022484408	2016-04-20	15112	10.6700000...	10.6700000...	0.0	3.33999991...	1.92999994...	5.40000009...	

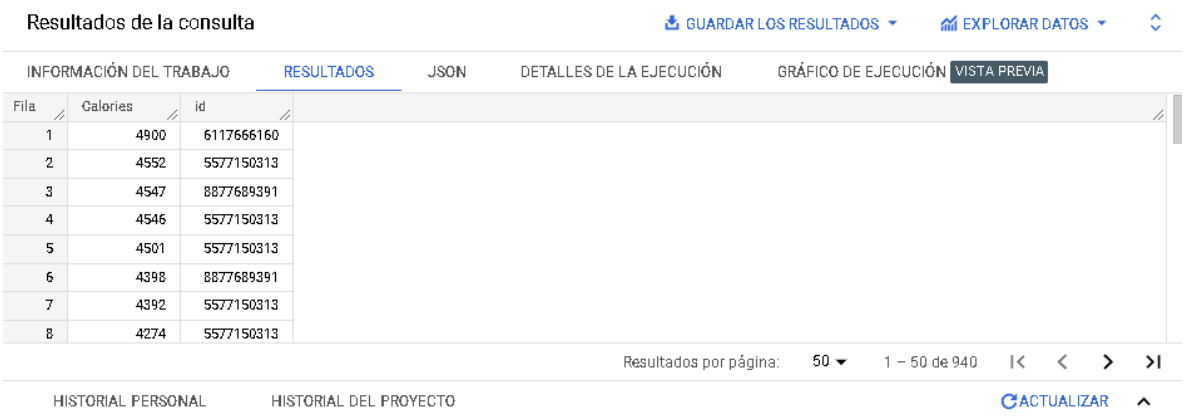
Imagen 4: propia

En la siguiente consulta usamos la función “distinct” para poder ver los id de los usuarios sin que se repita.

```
SELECT  distinct id From `portafolio2023.dailyActivity_merged.dailyActivity_merged`
```

En la siguiente consulta vemos las máximas calorías, por id la cual nos indica que usuario quemo más calorías y con que producto.

```
select Calories, id from `portafolio2023.dailyActivity_merged.dailyActivity_merged` order
 by Calories desc
```



Fila	Calories	id
1	4900	6117666160
2	4552	5577150313
3	4547	8877689391
4	4546	5577150313
5	4501	5577150313
6	4398	8877689391
7	4392	5577150313
8	4274	5577150313

Imagen 5: Propia

Ahora consultaremos varias tablas para tener un mejor resumen de los datos y poder graficar de manera conjunta, en los programas de visualización.

Consulta:

```
SELECT
A.Id,
A.Calories,
* EXCEPT(Id,
Calories,
```

```
ActivityDay,SleepDay,
SedentaryMinutes,
LightlyActiveMinutes,
FairlyActiveMinutes,
VeryActiveMinutes,
SedentaryActiveDistance,
LightActiveDistance,
ModeratelyActiveDistance,
VeryActiveDistance),
I.SedentaryMinutes,
I.LightlyActiveMinutes,
I.FairlyActiveMinutes,
I.VeryActiveMinutes,
I.SedentaryActiveDistance,
I.LightActiveDistance,
I.ModeratelyActiveDistance,
I.VeryActiveDistance
FROM
`portafolio2023.dailyActivity_merged.dailyActivity_merged` A
LEFT JOIN
 `portafolio2023.dailyActivity_merged.dailyCalories_merged` C
ON
A.Id = C.Id
AND A.ActivityDate=C.ActivityDay
AND A.Calories = C.Calories
LEFT JOIN
 `portafolio2023.dailyActivity_merged.dailyIntensities_merged` I
ON
A.Id = I.Id
AND A.ActivityDate=I.ActivityDay
AND A.FairlyActiveMinutes = I.FairlyActiveMinutes

AND A.LightActiveDistance = I.LightActiveDistance
AND A.LightlyActiveMinutes = I.LightlyActiveMinutes
AND A.ModeratelyActiveDistance = I.ModeratelyActiveDistance
AND A.SedentaryActiveDistance = I.SedentaryActiveDistance
AND A.SedentaryMinutes = I.SedentaryMinutes
AND A.VeryActiveDistance = I.VeryActiveDistance
AND A.VeryActiveMinutes = I.VeryActiveMinutes
LEFT JOIN
`portafolio2023.dailyActivity_merged.dailySteps_merged` S
ON
A.Id = S.Id
AND A.ActivityDate=S.ActivityDay
LEFT JOIN
`portafolio2023.dailyActivity_merged.sleepDay_merged` Sl
ON
A.Id = Sl.Id
AND A.ActivityDate=Sl.SleepDay;
```

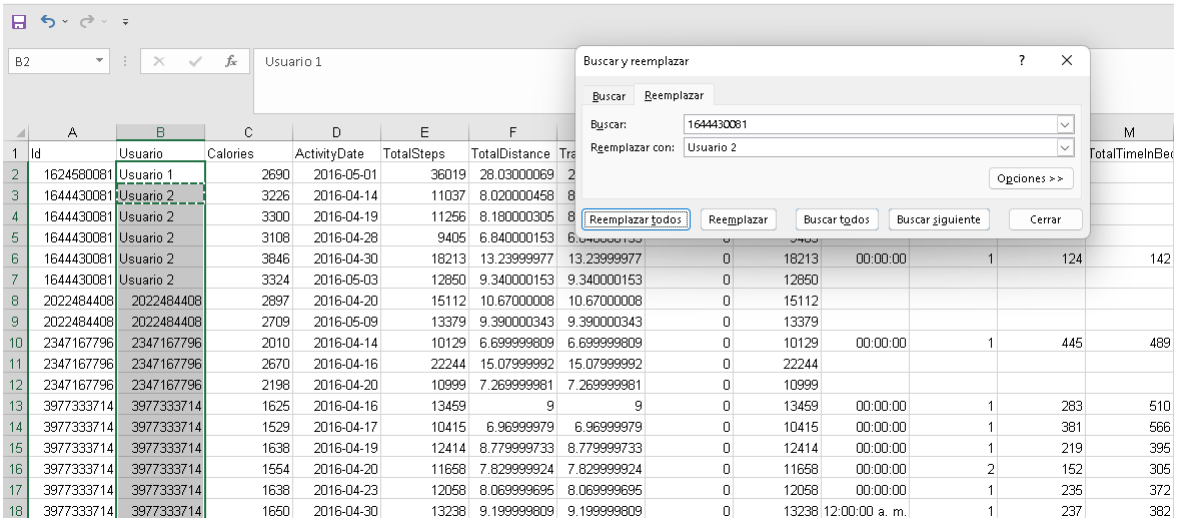
Agregamos Left join porque queremos los datos de la izquierda, agregando un On para poner las coincidencias de las demás tablas, los iguales son porque los datos de otras tablas son los mismos a las de las tablas consultadas, las letras mayúsculas a lado de las variables y conjunto de datos, son un sub nombre para poder diferenciar y decir que contienen los mismos datos en algunas tablas.

Después creamos una nueva columna en la tabla registrada para poder identificar de manera más fácil el id, como “Usuario x” y así poder mejores visualizaciones a los datos.
En otra hoja copiamos la columna de Id, y eliminamos todos los id repetidos dejando solo los id que son únicos.
Y así le damos un usuario a cada Id.

	A	B
1	Id	Usuario
2	1624580081	Usuario 1
3	1644430081	Usuario 2
4	2022484408	Usuario 3
5	2347167796	Usuario 4
6	3977333714	Usuario 5
7	4319703577	Usuario 6
8	4388161847	Usuario 7
9	4702921684	Usuario 8
10	5577150313	Usuario 9
11	6775888955	Usuario 10
12	6962181067	Usuario 11
13	7007744171	Usuario 12
14	7086361926	Usuario 13
15	8253242879	Usuario 14
16	8583815059	Usuario 15
17	8792009665	Usuario 16
18	1844505072	Usuario 17
19	1927972279	Usuario 18
20	2026352035	Usuario 19
21	2320127002	Usuario 20
22	2873212765	Usuario 21

Imagen 6: propia

Ahora con ctrl + b, podemos buscar y remplazar datos, ponemos el id y lo remplazamos en una nueva columna con su igual que seria “Usuario 1”. Y así para todos los datos, en automático los id que se busquen remplazara a todos como “Usuario 1”.



	Id	Usuario	Calories	ActivityDate	TotalSteps	TotalDistance								TotalTimeInBe
2	1624580081	Usuario 1	2690	2016-05-01	36019	28.03000069	2							
3	1644430081	Usuario 2	3226	2016-04-14	11037	8.020000458	8							
4	1644430081	Usuario 2	3300	2016-04-19	11256	8.180000305	8							
5	1644430081	Usuario 2	3108	2016-04-28	9405	6.840000153	6							
6	1644430081	Usuario 2	3846	2016-04-30	18213	13.23999977	13.23999977	0	18213	00:00:00	1	124	142	
7	1644430081	Usuario 2	3324	2016-05-03	12850	9.340000153	9.340000153	0	12850					
8	2022484408	2022484408	2897	2016-04-20	15112	10.67000008	10.67000008	0	15112					
9	2022484408	2022484408	2709	2016-05-09	13379	9.390000343	9.390000343	0	13379					
10	2347167796	2347167796	2010	2016-04-14	10129	6.699999809	6.699999809	0	10129	00:00:00	1	445	489	
11	2347167796	2347167796	2670	2016-04-16	22244	15.07999992	15.07999992	0	22244					
12	2347167796	2347167796	2198	2016-04-20	10999	7.269999981	7.269999981	0	10999					
13	3977333714	3977333714	1625	2016-04-16	13459	9	9	0	13459	00:00:00	1	283	510	
14	3977333714	3977333714	1529	2016-04-17	10415	6.96999979	6.96999979	0	10415	00:00:00	1	381	566	
15	3977333714	3977333714	1638	2016-04-19	12414	8.779999733	8.779999733	0	12414	00:00:00	1	219	395	
16	3977333714	3977333714	1554	2016-04-20	11658	7.829999924	7.829999924	0	11658	00:00:00	2	152	305	
17	3977333714	3977333714	1638	2016-04-23	12058	8.069999695	8.069999695	0	12058	00:00:00	1	235	372	
18	3977333714	3977333714	1650	2016-04-30	13238	9.199999809	9.199999809	0	13238	12:00:00 a. m.	1	237	382	

Imagen 7: Propia

Analizar

Los datos tienen el formato correcto, lo único que faltan son datos, ya que hay datos sin existir en algunas de las columnas, y eso cambiarlo con algún otro tipo de dato de Int a double no es correcto. Si cambiara a otro tipo de dato en alguna columna daría error en los análisis estadísticos.

Por lo general hubiera servido si no faltarán datos en más de tres columnas en las misma posición, siento que la falta de datos es un peligro.

- ¿Qué tendencias o relaciones encontraste en los datos?

Primero las relaciones más sencillas, los pasos totales en relación con las calorías, entre más caminas más quemas.
Después tenemos el rastreador y su medidor de distancia, por conteo de pasos.
El sedentarismo por Usuario, todo está identificado por Usuario, nos muestra pequeños gráficos como en Excel en tendencias de ganancias o líneas, podemos ver que, muchos tienen altos niveles de sedentarismo otros nulos.
Algo que verifica el sedentarismo son el tiempo en cama, algunos tienen altas horas en cama y algunos en nulo, esto se debe a que los datos son fueron bien registrados.

En los datos estadísticos tenemos que la media de calorías para cada usuario ronda entre los 1900.
Y la media de pasos anda entre 6000 y 7000 pasos al día, por Usuario.

Aunque puede variar por los tipos de distancia.

Después tenemos que las calorías tienen mucha relación con los pasos, aunque no se sabe si el usuario está corriendo, trotando o caminando normalmente, con las calorías podemos deducir que el usuario esta corriendo.

El sedentarismo tiene un poco de relación con la poca actividad , y lo podemos ver con un diagrama de dispersión y la pendiente de tendencia, que nos indican los puntos donde hay más relación.

Todo este proceso analítico puede revisarlo en el R Markdown o en "RPubs

BELLABEAT_PROYECTJ_GOOGLE. (2023, January 19). Retrieved January 19, 2023, from Rpubs.com website: https://rpubs.com/Marioouo/993512"

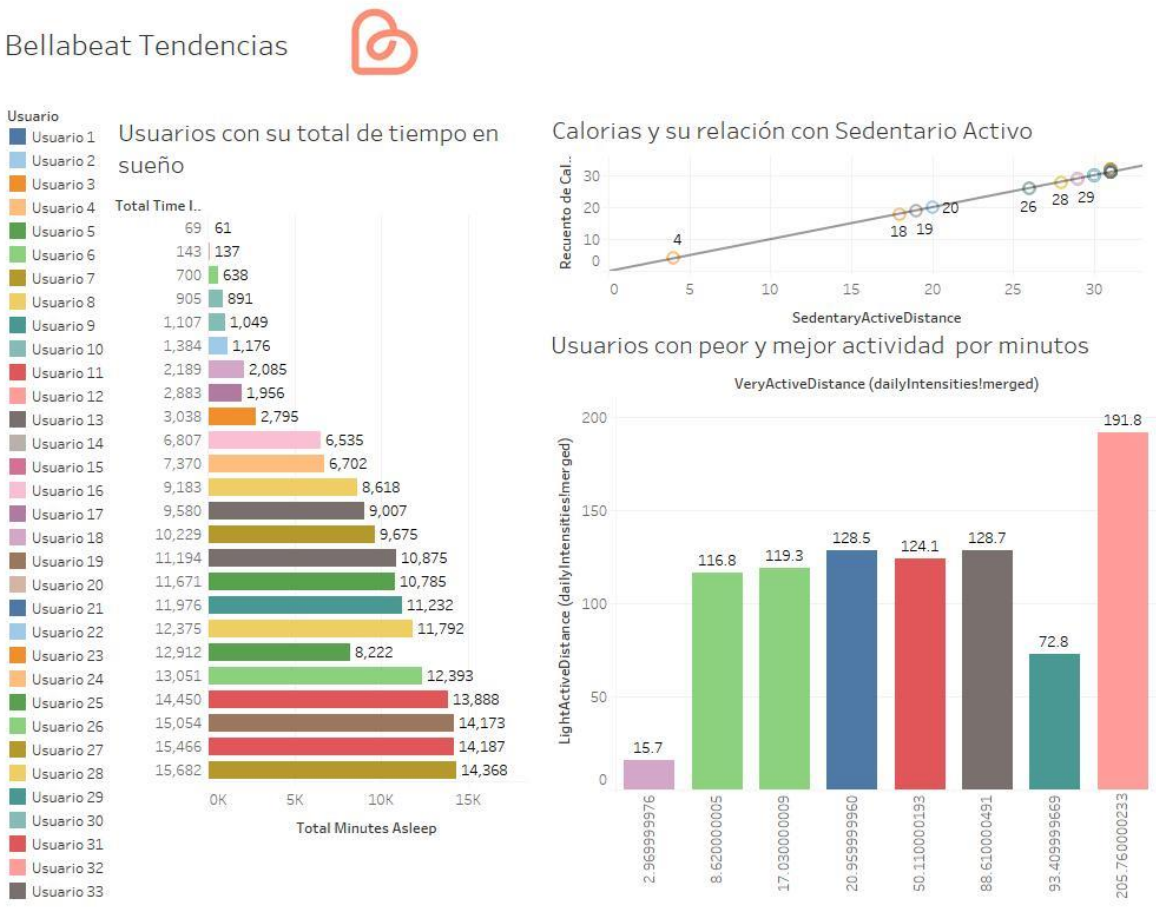
Compartir

Teniendo los datos y su limpieza crearemos los gráficos en tableu y por ende algunos Dashboard que nos digan de mejor manera que esta pasando con los productos de la empresa y su resultado con los usuarios.

Tenemos el siguiente Dashboard:

Donde podemos ver las tendencias y gráficas de los usuarios y su comportamientos con los productos de la empresa.

33 usuarios fueron analizados para su tendencias, podemos ver los títulos y el significado de cada gráfico, con sus subtítulos por variables.



Para más gráficos visita mi perfil de GitHub:

https://github.com/JuanMario0/Proyectj_google_bellabeat

Conclusiones:

Los productos de Bellabeat han sido de alta calidad brindando mediciones cercanas a lo esperado, debido a que varias variables mantienen una relación coherente y constante, los retos más complejos fue el cambio de tipo de dato, combinaciones de tablas, cálculos estadísticos y repasos de funciones.

Podemos definir que las tendencias apuntan a un buen uso de los productos, tenemos las calorías que se pueden verificar si son correctas por medio de las actividades, el sedentarismo si tiene relación o error por las horas de sueño, o estar en cama. El ritmo cardiaco si es bajo alto dependiendo la intensidad de la actividad, y podemos definir si el usuario esta corriendo o

caminando, en función de los pasos.

Usuarios con peor actividad, que es la mayoría, pero solo apuntamos a los peores para poder visualizar mejor las tendencias de los productos para apoyar a una mejor vida.

Ha salió de la mejor manera posible espero que estos análisis ayuden a otros a analizar los datos de distintas empresas, organizaciones, grupos , etc.

Bibliografía

1. Möbius. (2016). FitBit Fitness Tracker Data. Retrieved January 12, 2023, from Kaggle.com website: https://www.kaggle.com/datasets/arashnic/fitbit
2. Con la licencia siguiente: Creative Commons — CC0 1.0 Universal. (2023). Retrieved January 12, 2023, from Creativecommons.org website: https://creativecommons.org/publicdomain/zero/1.0/
3. Analisis de datos de Google | Coursera. (2023). Retrieved January 19, 2023, from Coursera website: https://www.coursera.org/programs/data-analytics-vebri?collectionId=A6y9c¤tTab=CATALOG&productId=hXUHfDgkEeylXgqHwJpmyQ&productType=s12n&showMiniModal=true
4. ThianCode. (2022). Como Crear un Repositorio y Subir Proyectos en GitHub paso a paso 2022 [YouTube Video]. Retrieved from https://www.youtube.com/watch?v=cGL8nH9HOoE
5. RPubs - BELLABEAT_PROYECTJ_GOOGLE. (2023, January 19). Retrieved January 19, 2023, from Rpubs.com website: https://rpubs.com/Marioouo/993512
6.