Python para la ciencia de datos – Análisis del fichero Fitness_Trackers

El objetivo de esta práctica es utilizar los conocimientos adquiridos en la asignatura para realizar un análisis de distintas características de productos tecnológicos (concretamente, fitnessbands y smartwatches) basándonos en la información recogida en el fichero *Fitness_trackers.csv*. La práctica se realizará en **grupos de 2 personas** y corresponde a **un 50% de la nota de la asignatura**. La fecha límite de entrega será el **08/01/2021** a las **23:59**. Se permite la entrega hasta dos días después (**10/01/2020**), pero las prácticas que se entreguen en fechas **posteriores** al **08/01** tendrán como **máximo** una puntuación de **8 puntos.** La entrega consistirá en **un notebook de Python en el que se incluyan las respuestas a las preguntas planteadas, así como el código desarrollado para responderlas.**

El fichero consta de una tabla con 11 columnas y 565 filas que contienen datos referentes a varias características de productos tecnológicos vendidos en el mercado indio. Cada una de estas columnas representa la siguiente información:

Brand Name: Indica el fabricante del producto
Device Type: Tiene dos categorías: FitnessBand y Smartwatch
Model Name: Indica la variante/nombre del product
Color: incluye el color de la correa/cuerpo del rastreador de fitness
Selling Price: Esta columna contiene el precio de venta de un dispositivo (puede tener un descuento apolicado)
Original Price: Incluye el precio original del producto (sin descuentos).
Display: Esta variable categórica muestra el tipo de pantalla del dispositivo: AMOLED, LCD, OLED, etc.
Rating (Out of 5): Valoración media de los clientes en una escala de 5.
Strap Material: Detalles del material utilizado para la correa del dispositivo
Average Battery Life (in days): La duración media de la batería citada por el fabricante, basada en las instrucciones individuales del producto. (No se trata de datos reales recogidos)
Reviews: recuento de las reseñas recibidas sobre el producto.

Se pide al alumno que, una vez leídos los datos, realice las siguientes operaciones con ellos:

1.- Uno de los primeros pasos para realizar un buen análisis de datos es familiarizarnos con los datos que contiene el fichero a analizar. Para ello, calcularemos los estadísticos descriptivos elementales de las variables del fichero. Una vez cargados los datos en nuestro programa Python (utilizando la librería Pandas), calcula los siguientes valores **para cada una de las variables**:

- Número de muestras (valores distintos de *missing*)
- Media y desviación estándar de aquellas variables en las que tenga sentido (numéricas)

- Valor mínimo y valor máximo de aquellas variables en las que tenga sentido (numéricas)

2.- Hay datos que nos interesa analizar basándonos en agrupaciones, para darle un sentido a nuestro análisis en base a esa agrupación. Basándonos en las siguientes agrupaciones:

- Por tipo de dispositivo
- Por precio de venta. Estableceremos cuatro grupos en base a la media del precio de venta de cada tipo de dispositivo:
    - Smartwatches con un precio inferior o igual a la media de precios de venta de estos dispositivos
    - Smartwatches con un precio superior a la media de precios de venta de estos dispositivos
    - Fitnessbands con un precio inferior o igual a la media de precios de venta de estos dispositivos
    - Fitnessbands con un precio superior a la media de precios de venta de estos dispositivos
- Por marca

Calcula los siguientes estadísticos en base a cada una de las agrupaciones definidas previamente con respecto a las variables selling price, original price, rating, average battery life in days y reviews:

- Número de observaciones
- Número de valores ausentes (*missing*)
- Mediana
- Varianza
- Valores máximo y mínimo

¿Qué conclusiones podemos sacar de estos cálculos? Comenta los resultados.

3.- Selecciona los dispositivos en los que la ratio de la duración de la batería con respecto al precio sea superior a la media de ratio de duración, que también debe calcularse. Comenta los resultados obtenidos.
NOTA: El cálculo de la ratio de duración se calculará con la fórmula:

$$ratio\_duración\_precio = \frac{\text{nº de días de batería} * 1000}{\text{precio dispositivo}}$$

4.- Ordena los dispositivos en base a la ratio calculada (de **mayor a menor**). ¿En qué dispositivos hay una mayor relación duración / precio? ¿Cuál es la marca que más relaciona el precio de venta con la duración de la batería?

5.- Obtén el total de ingresos por la venta de estos dispositivos para cada una de las marcas que aparecen en el conjunto de datos

6.- Representa la información obtenida en el ejercicio 5 mediante un diagrama de barras, pero sólo para las **5 marcas con los porcentajes de ingresos más elevados**.

7.- Estudia la correlación entre las variables y represéntala de la forma que consideres más precisa (*swarmplots,* mapas de calor…). ¿Existe alguna correlación que llame especialmente la atención? Comenta los resultados.

NOTA: Puedes investigar sobre el método corr() de pandas para analizar la correlación de las variables.

8.- Representa mediante un *boxplot* el precio de los dispositivos por marca. Comenta los resultados.

9.- Selecciona dos o más variables que te llamen la atención y analiza los datos mediante las gráficas o mediciones estadísticas que consideres oportunas y comenta los resultados que obtengas.