Introducción.

Este documento recoge las respuestas a las distintas preguntas planteadas en la práctica 1 de la asignatura de *Tipología y ciclo de vida de los datos* del *Máster Universitario en Ciencia de Datos* impartido por la Universitat Oberta de Catalunya.

Contexto.

Se decide realizar un estudio sobre la evolución de precios en los terminales móviles. Para ello, se elige Mediamarkt (https://www.mediamarkt.es/), que es una cadena de establecimientos de grandes superficies alemana, dedicada a la venta de electrodomésticos, informática y electrónica de consumo y específicamente el apartado de smartphones (https://www.mediamarkt.es/es/category/_smartphones-701189.html). El sitio web da acceso a todos los productos ofrecidos por la cadena en cartas organizadas en múltiples páginas y se ofrece la posibilidad de filtrado por ámbito o tipo de dispositivo. Esta decisión se toma porque dicho comercio es uno de los referentes en España para la venta de de telefonía, por ello puede darnos una idea muy aproximada sobre los precios en el mercado de móviles. Además, dada la cantidad de información técnica que proporciona el sitio web acerca de cada producto, podremos tener en cuenta las especificaciones técnicas de los productos a la hora de realizar las comparativas y observar así, por ejemplo, cómo afecta al precio la cantidad de RAM de un dispositivo.

Dataset.

El **título** elegido ha sido “Phones_Mediamarkt” y por ello la imagen identificactiva del mismo consta de un conjunto de smartphones. El dataset representa los terminales móviles vendidos en la web de Mediamarkt, incluyendo su nombre, características principales y precio. Está compuesto por 319 registros organizados en 26 páginas, que son los móviles en venta en el momento de realizar la captura de datos (07/04/21) y los atributos que incluye son los siguientes:

- **name:** nombre comercial del dispositivo.
- **color:** color del equipo.
- **storage**: almacenamiento interno del que dispone el móvil.
- **memory:** memoria RAM disponible en en el terminal.
- **screen:** tamaño y tecnología de la pantalla.
- **cpu:** procesador incluido en el equipo.
- **battery:** tamaño de la batería, medida en mAh.
- **os:** sistema operativo del dispositivo.

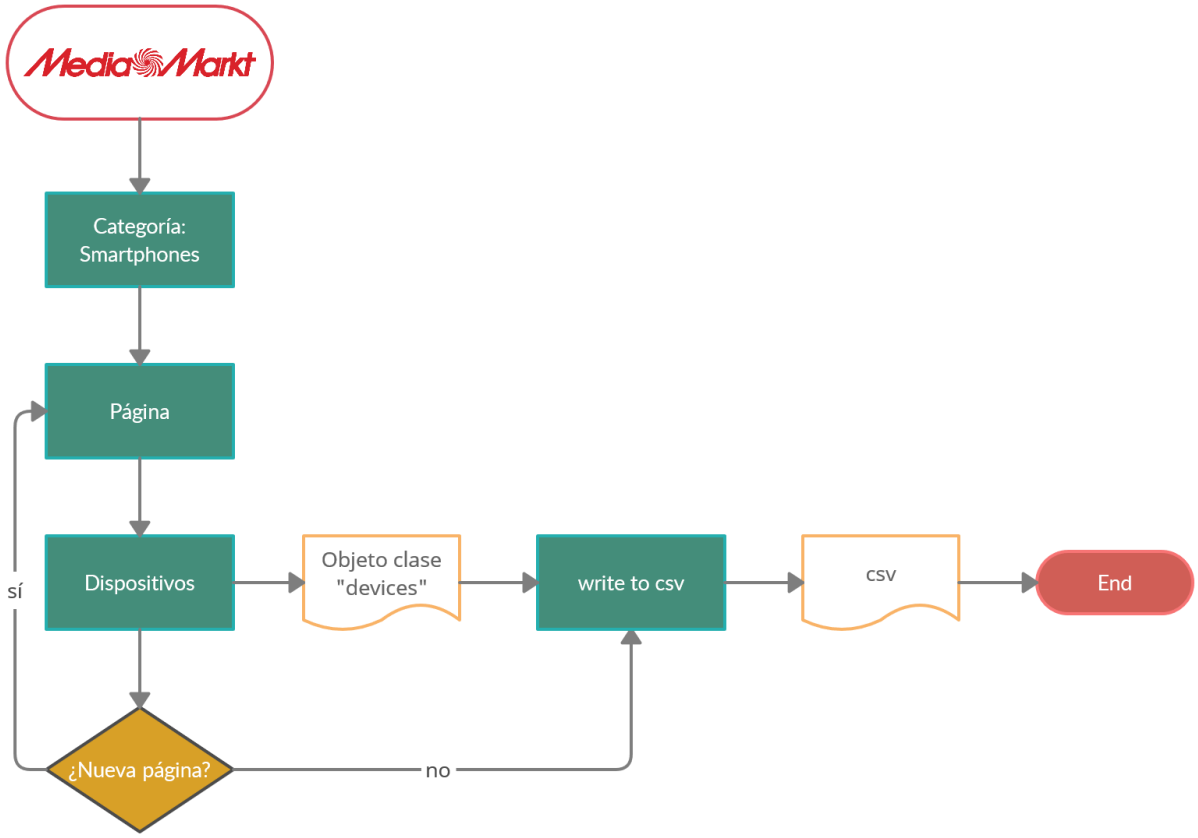
- **cpu_speed:** velocidad del procesador.

- **price**: precio final, en euros.

Se han tratado de obtener todos los atributos mencionados para cada uno de los dispositivos, aunque en ciertos casos no han sido proporcionados por lo que el dataset contiene campos vacíos en algunos casos y por lo que para algunos estudios se podría hacer necesaria una limpieza del dataset. Este dataset se proporciona en formato .csv de modo que tendrá la siguiente forma:

	A	B	C	D	E	F	G	H	I	J
1	name	color	storage	memory	screen	cpu	battery	os	cpu_speed	price
2	Xiaomi Mi 11 Lite	Negro	128GB	6GB RAM	6.55" FHD+	Qualcomm Snapdragon 732G	4250 mAh	Android	2.40 GHz	349.0
3	Xiaomi Mi 11 Lite	Azul	128GB	6GB RAM	6.55" FHD+	Qualcomm Snapdragon 732G	4250 mAh	Android	2.40 GHz	349.0
4	Xiaomi Mi 11 Lite		128GB	6GB RAM	6.55" FHD+	Qualcomm Snapdragon 732G	4250 mAh	Android	2.40 GHz	349.0
5	TCL 20 SE	Negro	64 GB	4 GB RAM	6.82" HD+	Qualcomm® Snapdragon™ 460	5000 mAh	Android	1.80 GHz	127.0
6	TCL 20 SE	Verde	64 GB	4 GB RAM	6.82" HD+	Qualcomm® Snapdragon™ 460	5000 mAh	Android	1.80 GHz	127.0
7	TCL 20 5G	Azul	6 GB		6.67"	Snapdragon 690 5G	4500 mAh	Android	256 GB	299.0
8	TCL 20 5G	Gris	6 GB		6.67"	Snapdragon 690 5G	4500 mAh	Android	256 GB	299.0
9	Xiaomi Redmi Note 10	Gris	128 GB	4 GB RAM	6.43" Full HD+	Snapdragon™ 678	5260 mAh	Android	2.2 GHz	199.0
10	Xiaomi Redmi Note 10	Blanco	128 GB	4 GB RAM	6.43" Full HD+	Snapdragon™ 678	5260 mAh	Android	2.2 GHz	199.0
11	realme 8 Pro	Negro	128 GB	8 GB RAM	6.43" Full HD+	Qualcomm SM7125 Snapdragon 720G	4500 mAh	Android	2.30 GHz	299.0
12	Xiaomi Redmi 9AT	Gris	32 GB	2 GB RAM	6.53" HD+	MediaTek Helio G25	5000 mAh	Android	2.00 GHz	99.0
13	Samsung Galaxy A12	Blanco	32 GB	3 GB RAM	6.5" HD+	Octa-Core	5000 mAh	Android	1.80 GHz	129.0
14	Samsung Galaxy A12	Azul	32 GB	3 GB RAM	6.5" HD+	Octa-Core	5000 mAh	Android	1.80 GHz	129.0
15	Xiaomi MI 10 T	Plata	6 GB		6.67" Full HD+	Qualcomm Snapdragon 865	5000 mAh	Android	2.84 GHz	329.0
16	Xiaomi Redmi Note 9S		6GB		6.67" Full HD+	Qualcomm® Snapdragon™ 720G	Snapdragon™ 720G,5020 mAh,Android	Android	2.3 GHz	179.0
17	Apple iPhone 11	Blanco	64 GB		6.1" Liquid Retina HD	Chip A13 Bionic		iOS	64 GB	649.0
18	Xiaomi Redmi Note 9S	Blanco Glaciar	6GB		6.67" Full HD+	Qualcomm® Snapdragon™ 720G	5020 mAh	Android	2.3 GHz	179.0
19	Xiaomi Redmi 9	Gris	4 GB		6.53" Full HD+	2 GHz	5020 mAh	Android	2 GHz	129.0
20	Apple iPhone 11	Blanco	128 GB		6.1" Liquid Retina HD	Chip A13 Bionic		iOS	128 GB	738.0

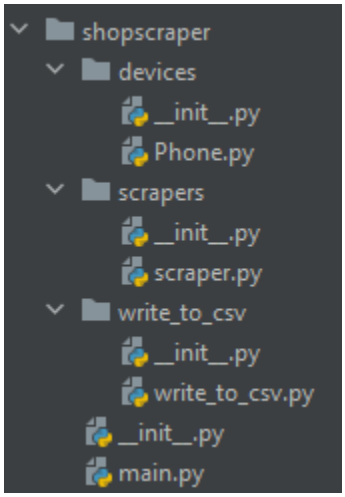
En cuanto al flujo de vida que han seguido los datos, se trata de una extracción, mediante técnicas de scraping, del apartado web de Mediamarkt destinado a dispositivos de telefonía móvil. Dicha extracción se lleva a cabo de forma secuencial, página a página hasta encontrar una que no contenga dispositivos, y que con cada iteración organiza los datos en un array de estructuras (objetos “Phone”) que contienen todos los datos de interés para el estudio. Una vez terminada la recolección y estructuración de datos se pasa a un proceso de escritura de los mismos, de modo que mediante el recorrido “dispositivo a dispositivo” del conjunto de datos se va formando un fichero de tipo .csv. Este comportamiento se puede observar en el siguiente **diagrama de flujo de datos**.



La recolección de datos actualmente se resume en un punto temporal de las ofertas de Mediamarkt, permitiendo el análisis del dataset de forma comparativa entre distintos dispositivos en el momento de la ejecución de la aplicación, o si se desean conservar los distintos CSV formados se podría llevar a cabo una comparación en múltiples puntos temporales. El proyecto está estructurado de tal forma que en un futuro se puedan incorporar nuevos tipos de dispositivos mediante la creación de nuevos scrapers, nuevos objetos y nuevos exportadores, y la incorporación de un factor histórico. Mediante el almacenamiento de los datos en un sistema de bases de datos podríamos almacenar cada ejecución de la aplicación en almacenamiento persistente y estructurado, de modo que sería muy sencillo realizar análisis acerca de la evolución (en el tiempo) de los dispositivos y sus precios. Sería sencillo elaborar una base de datos relacional donde se mantengan los datos descriptivos de un dispositivo (aquellos que no varían), junto con un identificador único, y en cuanto a los precios sería recomendable el uso de un sistema de bases de datos no relacional donde se almacene el precio obtenido de cada ejecución de la aplicación junto con una fecha y el identificador único del dispositivo.

Estructura del proyecto.

Como se ha mencionado anteriormente, el proyecto, está estructurado con vistas a su crecimiento y evolución de modo que identificamos 3 módulos diferenciados y un script de ejecución de la aplicación:



Como podemos observar, los módulos existentes constan de:

- **devices**: este módulo está destinado a la definición de objetos que contengan la información que define a un producto. Actualmente se cuenta solo con el objeto “Phone” que recoge la información técnica y el precio de un dispositivo de telefonía móvil.

- **scrapers**: módulo que contendrá los distintos scripts que realizan scraping de las distintos apartados de la web de Mediamarkt u otras webs. Como punto de partida se define un scraper que forma un dataset con los datos de smartphones ofrecidos por Mediamarkt.

- **write_to_csv**: en este módulo se recogerán los distintos archivos que llevan a cabo la exportación de datos a formato CSV. En este caso contamos con un script que lleva a cabo la transformación de objetos “Phone” a formato CSV.

Agradecimientos.

El conjunto de datos ha sido extraído de la tienda online de Mediamarkt cuyo objetivo es la venta y distribución de productos electrónicos y accesorios, por lo que muestran un listado de

sus productos, junto con la información técnica de los mismos y su precio. La información ha sido extraída mediante el uso de herramientas de *web scraping* y el escrutinio del código HTML del sitio web.

Inspiración.

Existen múltiples finalidades para el conjunto de datos recolectado. Como podrían ser:

- Apoyo en la elección de un producto: actualmente, dada la cantidad de oferta del mercado, se hace esencial la búsqueda de la mejor opción calidad/precio entre los distintos distribuidores, por lo que podría utilizarse el conjunto de datos para llevar a cabo una comparativa con conjuntos similares extraídos de otros distribuidores, de modo que se podría averiguar de forma rápida quien ofrece el producto al menor precio.

- Identificación de sucesos: mediante la aplicación de modelos de predicción se podría tratar de llegar a determinar cómo afectan los distintos atributos de un dispositivo a su precio final de modo que, por ejemplo, se podría determinar el aporte de la capacidad de almacenamiento del dispositivo de telefonía móvil al precio final del producto.

Licencia.

La licencia escogida para este proyecto sería **CC BY-NC-SA 4.0** puesto que:

- Implica el **reconocimiento** de los autores del contenido original y la mención de los cambios realizados sobre el mismo.

- **Impide** el uso del proyecto con **objetivos comerciales.**

- Los productos llevados a cabo con este proyecto como base deberán ser distribuidos con la misma licencia.

Publicación de los datos.

Para la publicación del dataset se ha escogido la plataforma Zenodo (zenodo.org) y se ha publicado bajo el **título** "*Phones_Mediamarkt* dataset" con el **DOI** "*10.5281/zenodo.4678303*" y con la url "https://zenodo.org/record/4678303#.YHGYBegzYkk".

Participación.

Contribuciones	Firma
Investigación previa	DSR, XPP
Redacción de las respuestas	DSR, XPP
Desarrollo código	DSR, XPP