PRA1- Tipología y ciclo de vida de los datos:

¿Cómo podemos capturar los datos de la web?

Nombre estudiante:

Jeisson David Chavarro Torres
Sebastian Meneses Oliveros

Práctica-1
Tipología y ciclo de vida de los datos.

1. Contexto.

Explicar en qué contexto se ha recolectado la información. Explicar por qué el sitio web elegido proporciona dicha información. Indicar la dirección del sitio web.

Para esta práctica, el equipo decidió enfocarse en el mercado de las criptomonedas y la obtención de un dataset que permita visualizar la información más relevante de este mercado en los últimos años.

Para lograr esto, se escogió como fuente para el proceso de scraping y la creación de un datasets, la página coinmarketcap (https://coinmarketcap.com/). La cual usando el lenguaje de programación Python, y librerías como Beautifulsoup se logró obtener los datos almacenados en la sección de Cryptocurrency Historical Data Snapshot.

2. Título. Definir un título que sea descriptivo para el dataset.

El título del dataset, llevaría el mismo de la fuente **Cryptocurrency Historical Data Snapshot**, ya que la data obtenida es básicamente una instantánea de la situación del mercado en diferentes días, teniendo como intervalo de tiempo desde el 1 de enero de 2013 y hasta el 13 de noviembre de 2022.

3. Descripción del dataset. Desarrollar una descripción breve del conjunto de datos que se han extraído. Es necesario que esta descripción tenga sentido con el título elegido.

Partiendo del nombre dado al dataset **(Cryptocurrency Historical Data Snapshot),** podemos observar que los datos obtenidos no solo nos permiten identificar diferentes cryptocurrencies bajo datos como el nombre, precio, la cantidad de unidades en circulación, symbol y date.

En una fase posterior podríamos identificar y modelar variables como precio, intensidad de los movimientos del mercado en diferentes días y volumen de transacciones, creando así un histórico que nos permite ver lo que pasó y está pasando.

4. Representación gráfica. Dibujar un esquema o diagrama que identifique el dataset visualmente y el proyecto elegido.

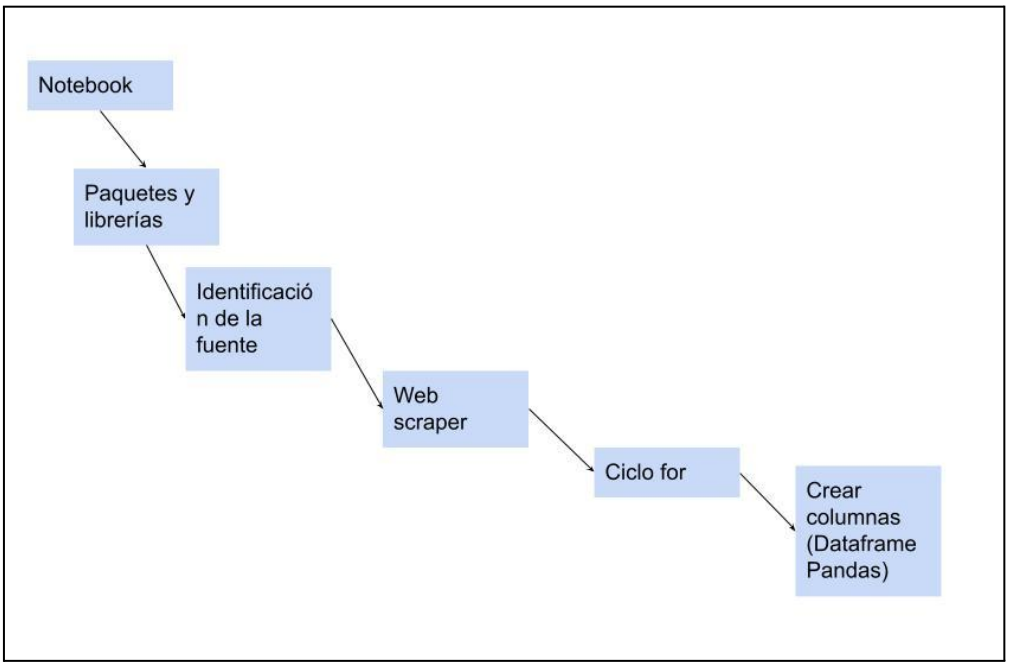
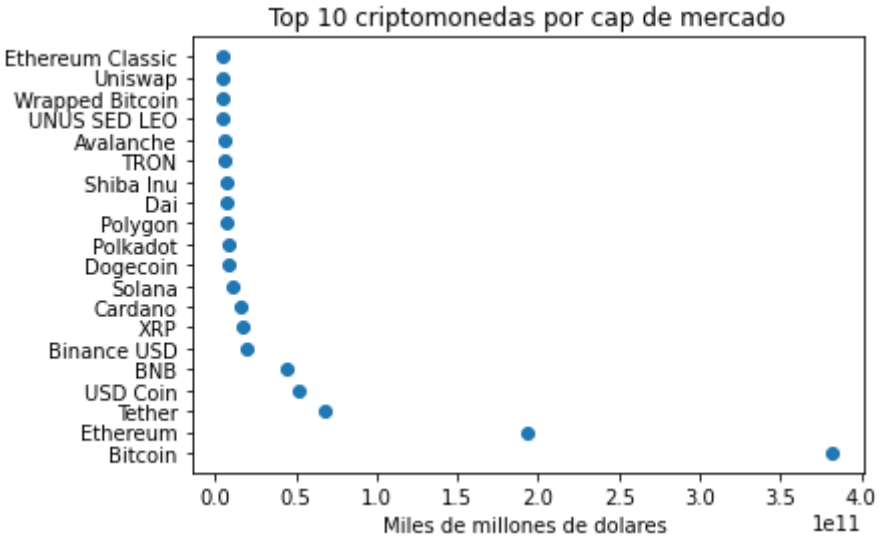


Gráfico del proceso.

Capitalización de mercado por cryptocurrency



5. Contenido. Explicar los campos que incluye el dataset y el periodo de tiempo de los datos.

- Name
  - Nombre de la criptomoneda.
- Symbol
  - Símbolo o ticket con el cual se identifica a la criptomoneda en el mercado.
- Price
  - El precio del cripto activo en el momento de la captura de datos.
- Market Cap
  - Es el valor del cripto activo, basado en la cantidad de unidades en circulación y el valor que el mercado da a este.
- Circulating Supply
  - Cantidad de unidades en circulación.

6. Propietario. Presentar al propietario del conjunto de datos. Es necesario incluir citas de análisis anteriores o, en caso de no haberlas, justificar esta búsqueda con análisis similares. Justificar qué pasos se han seguido para actuar de acuerdo a los principios éticos y legales en el contexto del proyecto.

El propietario de estos datos es Coin Market Cap, el cual es un sitio de seguimiento de precios del mercado de los cripto activos que lleva activo desde 2013, siendo uno de los sitios de seguimiento de precio más antiguo y por consiguiente quien tiene las mejores bases de datos.

Según su propia página web y las condiciones del servicio (Monitoreo de precios y otros datos del mercado), el servicio debe ser usado para actividades personales y no se autoriza su uso para actividades NO comerciales, y si se usa para actividades comerciales debería realizarse otro tipo de acuerdo con la compañía.

Por otra parte, al consultar el archivo robots.txt, encontraremos la siguiente información:

```
User-agent: *
Allow: /

Disallow: */currencies/*/social/$
Disallow: */currencies/*/onchain-analysis/$
Disallow: */currencies/*/wallets/$
Disallow: */currencies/*/ratings/$
```

Disallow: */currencies/*/price-estimates/$

Disallow: /*/headlines/*$

En cual no se indica que este prohibido entrar o consultar el apartado de https://coinmarketcap.com/historical/ , el cual se esta usando para obtener la data necesaria.

También, dentro de los términos del servicio no se indica que este prohibido usar métodos como el scraping para obtener datos publicados en la página, solamente se especifica que se "otorga una licencia limitada, personal, no exclusiva, no sublicenciable e intransferible para usar el Contenido y usar este Servicio, en cada caso únicamente para su uso personal" y que cualquier uso comercial o explotación de los datos obtenidos con fines comerciales, será causante de cancelación de la licencia".

Teniendo en cuenta que la información obtenida a través de esta página, no será usada de manera comercial, sino solamente para hacer análisis y pruebas de manera académica, no se tendría ningún problema para usar esta técnica.

7. Inspiración. Explicar por qué es interesante este conjunto de datos y qué preguntas se pretenden responder. Es necesario comparar con los análisis anteriores presentados en el apartado 6.

Este conjunto de datos es interesante, debido a que:

- Permite concentrar los datos de los principales cripto activos en un solo dataset.
- Existen casi 10 años de datos (desde el momento en que Coinmarketcap inició el monitoreo) de las principales criptomonedas en el mercado.
- Muestra la evolución del mercado, y cómo algunas se consolidaron y otras se estancaron y otras se perdieron con el tiempo. (Precio y capitalización de mercado).
- Muestra datos de variación en diferentes momentos de bonanza así como de crisis del mercado, y de todos los cripto activos que lo componen.

8. Licencia. Seleccionar una licencia adecuada para el dataset resultante y justificar el motivo de su elección. Ejemplos de licencias que pueden considerarse:

El sitio usa un tipo de licencia llamado (limited license personal), en el cual la información mostrada en el sitio es solo para uso de la persona que creó una cuenta en

el servicio, no puede transferir dicha capacidad a otras personas, y no puede hacer explotación comercial de dichos recursos, ya que de ser detectada esta actividad, la cuenta y el acceso serán revocados.

9. Código. Código con el que se ha obtenido el dataset, preferiblemente en Python o, alternativamente, en R.

El código está hecho en lenguaje de programación python, el cual nos permite a través de librerias como BeautifulSoup, hacer scraping de la información alojada en una página web, en este caso coinmarketcap.

Posteriormente a través de un ciclo for, se extraer las columnas identificadas en la tabla html del sitio web.

Por otra parte, por medio de la librería de pandas, el código puede crear columnas para alojar la información en un nuevo data frame, que nos permita crear un dataset para ser analizado en posteriores fases del proyecto.

10. Dataset. Publicar el dataset obtenido en formato CSV en Zenodo, incluyendo una breve descripción. Obtener y adjuntar el enlace del DOI del dataset (https://doi.org/…). El dataset también deberá incluirse en la carpeta /dataset del repositorio.Si existe alguna circunstancia que impida publicar abiertamente el dataset real en Zenodo, se deberá: (1) comentar esta circunstancia y justificar el motivo en este apartado; (2) generar un dataset simulado y publicarlo en Zenodo, obteniendo el enlace del DOI; y (3) comunicar al profesor el dataset real de forma privada (p. ej., utilizando un repositorio privado).

- **Link**
  - https://zenodo.org/record/7342586#.Y3ugSXbMK5c
- **DOI**
  - **[![DOI](https://zenodo.org/badge/DOI/10.5281/zenodo.7342586.svg)](https://doi.org/10.5281/zenodo.7342586)**

11. Vídeo. Realizar un breve vídeo explicativo de la práctica (máximo 10 minutos), que deberá contar con la participación de los dos integrantes del grupo. En el vídeo se deberá realizar una presentación del proyecto, destacando los puntos más relevantes, tanto de las respuestas a los apartados como del código

utilizado para extraer los datos. Indicar el enlace del vídeo (https://drive.google.com/…), que deberá ubicarse en el Google Drive de la UOC.

- https://drive.google.com/file/d/1N40AeU4hIrITcVNYhkLoekh-mcustaTm/view?usp=sharing

12. Tabla de contribuciones del equipo .

Contribuciones	Firma
Investigación previa	Jeisson, Sebastian
Redacción de las respuestas	Jeisson, Sebastian
Desarrollo del código	Jeisson, Sebastian
Participación en el vídeo	Jeisson, Sebastian