Tipología y ciclo de vida de lo datos

Práctica 1

Alumnas:

Erika Paola Martínez Soria
Eugenia Bezek

Enunciados

1. **Contexto**. Explicar en qué contexto se ha recolectado la información. Explicar por qué el sitio web elegido proporciona dicha información.

Con el objetivo de crear un dataset a través de recolectar los datos por medio de web scraping, y luego de analizar distintas opciones, elegimos la página web https://arriendayvende.com, la cual proporciona información sobre propiedades que están en alquiler en Ecuador. Analizando el código fuente de la misma pudimos identificar cómo estaban diferenciados los bloques que contienen datos sobre cada propiedad, y dentro de cada uno pudimos diferenciar cada dato que necesitábamos para el dataset final.

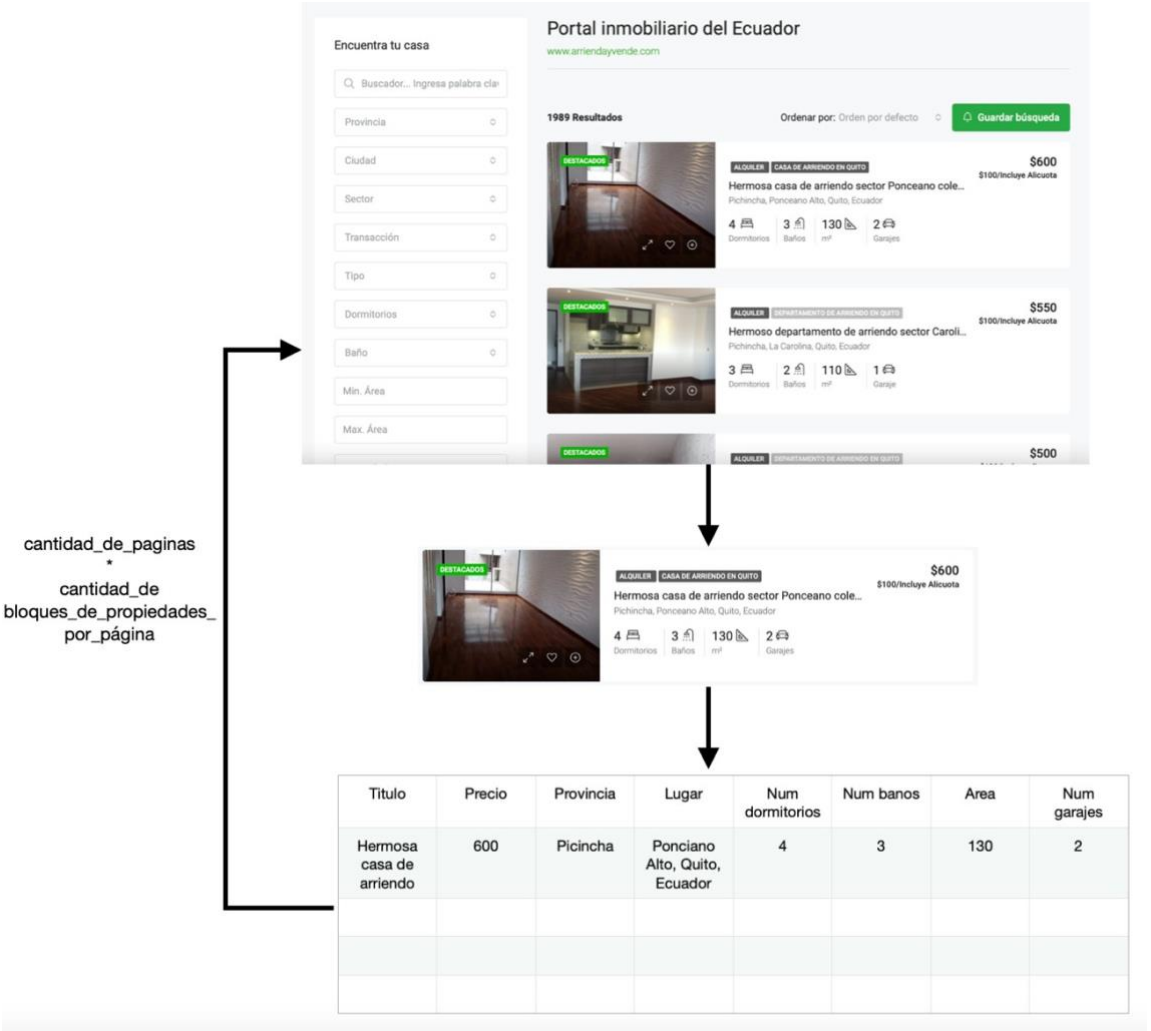
2. **Título**. Definir un título que sea descriptivo para el dataset.

El nombre que elegimos para el dataset es: real_state_ecuador_dataset.csv

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

El conjunto de datos contiene las características de las propiedades que están en alquiler en Ecuador. Entre estas características se encuentra la ubicación, el precio mensual, la superficie, la cantidad de baños y la cantidad de garajes.

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



5. **Contenido.** Explicar los campos que incluye el dataset, el periodo de tiempo de los datos y cómo se han recogido.

El dataset incluye los siguientes campos:

- Título: una breve descripción de la propiedad
- Precio: El precio del alquiler mensual
- Provincia: Provincia donde está ubicada la propiedad
- Lugar: Contiene más detalles de la ubicación de la propiedad, como la Ciudad y el país.
- Num. Dormitorios: Cantidad de dormitorios que posee la propiedad
- Num. banos: Cantidad de baños que posee la propiedad
- Area: Superficie de la propiedad
- Num. Garajes: En caso de que la propiedad posea garaje, indica cuántos son.

En cuanto al periodo de tiempo, los datos que obtuvimos fueron de las propiedades que hoy están activamente publicadas en la web, cuando una propiedad deja de estar disponible, no aparecerá más.

Web scraping fue la herramienta que utilizamos para extraer estos datos.
Para ello generamos un código en lenguaje python, para el cual requerimos importar librerías como ´requests´, ´BeautifulSoup´, ´writer, ´csv´y ´re´.
Como primer paso, definimos la url de la página web de la que extraeríamos la información.
Luego, utilizando la librería ´csv´, comenzamos a escribir en el archivo que nombramos real_state_ecuador_dataset.csv.
Para encontrar los datos, primero ubicamos en el código de la página, la clase que se utilizaba en los bloques de cada propiedad publicada, en este caso esa clase era: ´item-listing-wrap´.
Una vez dentro de cada bloque, identificando las clases de cada objeto, obtuvimos el título, el precio, la ubicación, la cantidad de baños, garajes y la superficie.
Cada bloque que fuimos encontrando en la página web corresponde a una fila del dataset. En todos los casos, cuando no había datos cargados, utilizamos la palabra ´NaN´ para analizar más adelante cómo trataremos estos registros faltantes.
Una vez que recorrimos las 28 páginas del sitio web, el archivo se cierra y queda con la información guardada.

6. **Agradecimientos.** 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.

Para conocer al propietario del conjunto de datos utilizamos la librería whois, como se muestra a continuación:

```
1 import whois
2 print(whois.whois('https://arriendayvende.com'))
```


```
{
  "domain_name": [
    "ARRIENDAYVENDE.COM",
    "arriendayvende.com"
  ],
  "registrar": "GoDaddy.com, LLC",
  "whois_server": "whois.godaddy.com",
  "referral_url": null,
  "updated_date": [
    "2021-08-12 23:43:43",
    "2021-08-12 18:43:41"
  ],
  "creation_date": [
    "2018-10-19 23:04:27",
    "2018-10-19 18:04:27"
  ],
  "expiration_date": [
    "2023-10-19 23:04:27",
    "2023-10-19 18:04:27"
  ],
  "name_servers": [
    "NS8912.BANAHOSTING.COM",
    "NS8913.BANAHOSTING.COM"
  ],
  "status": [
    "clientDeleteProhibited https://icann.org/epp#clientDeleteProhibited",
    "clientRenewProhibited https://icann.org/epp#clientRenewProhibited",
    "clientTransferProhibited https://icann.org/epp#clientTransferProhibited",
    "clientUpdateProhibited https://icann.org/epp#clientUpdateProhibited"
  ],
```

```
  "emails": "abuse@godaddy.com",
  "dnssec": "unsigned",
  "name": null,
  "org": "Arrienda y Vende",
  "address": null,
  "city": null,
  "state": "Pichincha",
  "zipcode": null,
  "country": "EC"
}
```


Podemos ver que el propietario es “Arrienda y Vende” y nos brinda su email.
La fecha de creación del dominio es el 19/10/2018 y la de caducidad es el 19/10/2023.

Los pasos que hemos seguido para actuar de acuerdo a los principios éticos y legales en el contexto del proyecto son:

1- Leer el archivo robots.txt de la página web que elegimos, el objetivo de leer este documento es conocer si hay alguna restricción para realizar web scraping sobre ese sitio web, de forma de minimizar las posibilidades de ser bloqueados.
   A continuación se muestra un fragmento del robots.txt donde se detallan los User-agent bloqueados para esta página web.
   También vemos que al final, establece un retardo para algunos User-agent específicos, como 50 segundos para noxtrumbot.

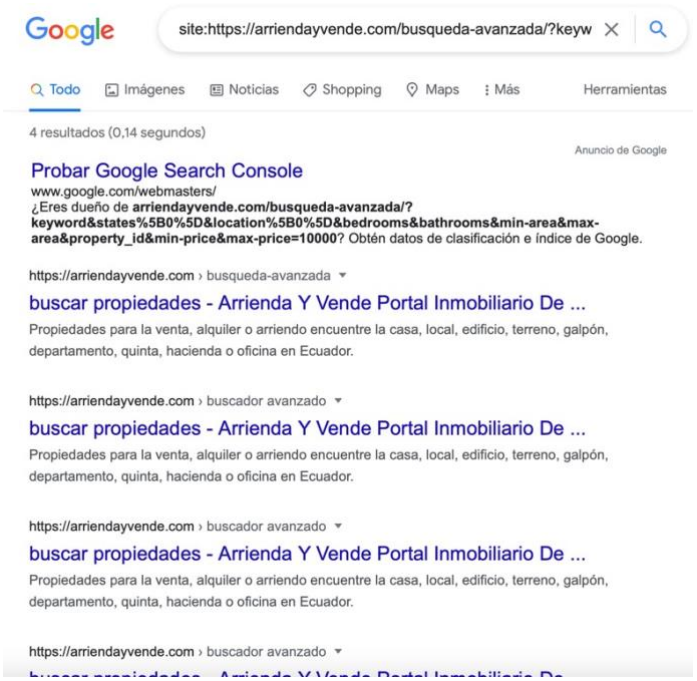

```
Disallow: /agencias/

Disallow: /wp-content/plugins/
Disallow: /wp-content/themes/
Disallow: /wp-includes/
Disallow: /wp-admin/
Disallow: /wp-
Disallow: /?s=
Disallow: /search
Allow: /feed/$
Disallow: /feed
Disallow: /comments/feed
Disallow: /*/feed/$
Disallow: /*/feed/rss/$
Disallow: /*/trackback/$
Disallow: /*/*/feed/$
Disallow: /*/*/feed/rss/$
Disallow: /*/*/trackback/$
Disallow: /*/*/*/feed/$
Disallow: /*/*/*/feed/rss/$
Disallow: /*/*/*/trackback/$
User-agent: MSIECrawler
Disallow: /
User-agent: WebCopier
Disallow: /
User-agent: HTTrack
Disallow: /
User-agent: Microsoft.URL.Control
Disallow: /
User-agent: libwww
Disallow: /
User-agent: noxtrumbot
Crawl-delay: 50
User-agent: msnbot
Crawl-delay: 30
User-agent: Slurp
Crawl-delay: 10
```


2- Como segundo paso, revisamos el sitemap, este también lo encontramos en el robots.txt como podemos ver en el siguiente fragmento:

```
# This virtual robots.txt file was created by the Virtual Robots.txt WordPress plugin:
https://www.wordpress.org/plugins/pc-robotstxt/
User-Agent: *
Sitemap: https://arriendayvende.com/sitemaps.xml
https://arriendayvende.com/sitemaps.xml
```

En el sitemap pudimos leer un listado con los links a las páginas web que contiene el sitio. Pudimos ver que la fecha de actualización de algunos links fue el 2022-04-05T23:34:04-05:00, por lo que entendemos que se mantiene al día.



3- Como tercer paso realizamos la estimación del tamaño de la página web que elegimos. Para esto utilizamos la búsqueda avanzada del buscador Google: www.google.com/advanced_search, y filtramos únicamente por la URL que utilizamos y nos dio el siguiente resultado:

4 resultados (0,14 segundos)

Anuncio de Google

Probar Google Search Console
www.google.com/webmasters/
¿Eres dueño de **arriendayvende.com/busqueda-avanzada/?keyword&states%5B0%5D&location%5B0%5D&bedrooms&bathrooms&min-area&max-area&property_id&min-price&max-price=10000**? Obtén datos de clasificación e índice de Google.

https://arriendayvende.com › busqueda-avanzada
buscar propiedades - Arrienda Y Vende Portal Inmobiliario De ...
Propiedades para la venta, alquiler o arriendo encuentre la casa, local, edificio, terreno, galpón, departamento, quinta, hacienda o oficina en Ecuador.

https://arriendayvende.com › buscador avanzado
buscar propiedades - Arrienda Y Vende Portal Inmobiliario De ...
Propiedades para la venta, alquiler o arriendo encuentre la casa, local, edificio, terreno, galpón, departamento, quinta, hacienda o oficina en Ecuador.

https://arriendayvende.com › buscador avanzado
buscar propiedades - Arrienda Y Vende Portal Inmobiliario De ...
Propiedades para la venta, alquiler o arriendo encuentre la casa, local, edificio, terreno, galpón, departamento, quinta, hacienda o oficina en Ecuador.

https://arriendayvende.com › buscador avanzado

Vemos que sólo devolvió 4 resultados, por lo que pudimos realizar el web scraping a través de descargas secuenciales sin problemas.

4- Como último paso del análisis inicial, analizamos la tecnología utilizada en el diseño del sitio web, utilizando el comando builtwith:

```
import builtwith
builtwith.parse('https://arriendayvende.com/busqueda-avanzada/?keyword&states%5B0%5D&loc
```

```
{'font-scripts': ['Google Font API'],
 'tag-managers': ['Google Tag Manager'],
 'web-frameworks': ['Twitter Bootstrap'],
 'ecommerce': ['WooCommerce'],
 'cms': ['WordPress'],
 'programming-languages': ['PHP'],
 'blogs': ['PHP', 'WordPress'],
 'javascript-frameworks': ['jQuery']}
```

Podemos ver que el sitio web utiliza como framework: ‘Twitter Bootstrap’, el e commerce para permitir las transferencias a través de su página utiliza ‘WooCommerce’, como cms (Content Management System), para la creación de la página web, usa WordPress, el lenguaje de programación es PHP y algunas librerías de JavaScript.

En cuanto al análisis de trabajos similares, a continuación mostramos algunos que nos han servido como referencia para llevar adelante el proyecto.
En todos los casos realizan web scraping a sitios web de 'Real State' y utilizan Python como lenguaje de programación. Los primeros dos utilizan las librerias BeautifulSoup y requests para leer el código del sitio. En el tercer ejemplo utiliza selenium.

https://github.com/guiaramos/webscraping-realestate/blob/master/webscraping-realestate.ipynb

https://github.com/rahulguptagzb09/Scraping-Real-Estate-Property-Data-From-Web-Using-Python/blob/master/century.py

https://github.com/iamparbonaaa/property-scraping/blob/master/get_data.py

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.

El dataset de alquileres es interesante e importante para varios tipos de personas. En el caso de los agentes de bienes raíces es esencial saber los precios de cierta área para poder ser competitivo y eficaz al arrendar una propiedad. Para una persona que desea alquilar, el dataset provee información de arriendos en diferentes zonas y así tomar una decisión mejor informada. Incluso podría ser de utilidad para investigadores, que quieren conocer más sobre el estado del sector inmobiliario en el país.

Mediante un EDA (Exploratory data analysis), por ejemplo, el dataset nos puede proveer con información de la distribución de precios según ubicación, tamaño y otras características de las propiedades en alquiler en Ecuador. Siendo algunas de las preguntas de interés:

¿Cuál es la ciudad y/o provincia con mayores precios?
¿Cómo se relaciona el tamaño de la propiedad con el precio del alquiler?
¿Aquellas propiedades que poseen garaje, son más costosas que las que no poseen?
¿Cuál es la superficie promedio de las propiedades en alquiler por ciudad o provincia?

También podríamos utilizar técnicas de Minería de datos para elaborar modelos predictivos que permitan predecir el precio de alquiler de propiedades basados en ciertas características.

Adicionalmente, para realizar el proyecto elegimos hacerlo con las librerías BeautifulSoup y requests, a diferencia del tercer ejemplo que pusimos en el punto 6, que utiliza la librería selenium, específicamente el módulo de webdriver. En nuestro caso, como no fue necesario movernos por la página a través de hacer click en

botones, no utilizamos selenium. En comparación con los dos primeros ejemplos, el trabajo es similar, utilizamos las mismas librerías para recolectar los datos, pero en nuestro caso, incluimos todo el código dentro de una clase, lo cual lo hace más robusto y organizado.

8. **Licencia**. Seleccionar una de estas licencias para el dataset resultante y justificar el motivo de su selección.

La licencia que escogimos para el dataset generado es **Released Under CC BY-NC-SA 4.0 License**, la cual permite que otros compartan el material en cualquier medio y formato y puedan adaptarlo y transformarlo según su necesidad, pero bajo los siguientes términos:

- *Atribución*: En cualquier explotación de la obra autorizada por la licencia será necesario reconocer la autoría (obligatoria en todos los casos)
- *No comercial:* La explotación de la obra queda limitada a usos no comerciales.
- *Compartir Igual:* La explotación autorizada incluye la creación de obras derivadas siempre que mantengan la misma licencia al ser divulgadas.

Creemos que es importante compartir la información que generamos para colaborar con todos aquellos interesados en realizar análisis sobre estos datos, permitiéndoles realizar las transformaciones que crean necesarias, pero reconociendo la autoría de quien facilitó esa información.
Además la explotación de este archivo quedaría excluida del uso comercial, ya que el objetivo principal es colaborar con la comunidad de analistas de datos y no con fines de generar ganancias.

9. **Código.** Adjuntar en el repositorio Git el código con el que se ha generado el dataset, preferiblemente en Python o, alternativamente, en R.

https://github.com/eugeniabezek/Realstate_dataset_webscraping

10. **Dataset.** Publicar el dataset obtenido en formato CSV en Zenodo con una breve descripción. Obtener y adjuntar el enlace del DOI.

Link al enlace del DOI en Zenodo: https://zenodo.org/record/6429586#.YlNp3i-xDu1

11. **Vídeo.** Se debe hacer entrega de un vídeo explicativo de la práctica en donde cada uno de los integrantes del grupo explique con sus propias palabras tanto las respuestas del proyecto como el código utilizado para llevar a cabo la extracción. El vídeo debe ser enviado a través de un enlace a Google Drive que deben proporcionar, junto con el enlace al repositorio Git, al momento de entregar la práctica.

Link de enlace de video:
https://drive.google.com/file/d/1D-TH-0_v3z5s8l4y52J1l_aRDdggFw0O/view?usp=sharing

Tabla de contribuciones:

Contribuciones	Firma
Investigación previa	Erika Paola Martínez Soria Eugenia Bezek
Redacción de las respuestas	Erika Paola Martínez Soria Eugenia Bezek
Desarrollo del código	Erika Paola Martínez Soria Eugenia Bezek