M2.851 - Tipología y ciclo de vida de los datos

Práctica 1: ¿Cómo podemos capturar los datos de la web?

Nombres de los estudiantes:

- **Félix Antonio Mucha Morales**
- **Jose Carlos Enriquez Lira**

Aula 2

Criterios de evaluación generales de la Practica.

- Idoneidad de las respuestas (deberán ser claras y completas).
- Complejidad del sitio web elegido para la extracción de datos. Es importante tener en cuenta que la complejidad será un factor que se evaluará y dependerá tanto del sitio elegido como del análisis realizado en la práctica.
- Síntesis y claridad, a través del uso de comentarios, del código resultante.
- Presentación adecuada de los datos.
- Organización y claridad de los documentos de entrega final.
- Completitud de los documentos requeridos para la entrega final.
- Seguimiento de recomendaciones para el buen uso del web scraping.

Profesor Responsable: María Isabel Guitart Hormigo

Profesor Colaborador: Daniel Romero Pérez

Índice de contenidos

1. Contexto:

Para el desarrollo de la práctica se ha buscado identificar un negocio que ofrezca y muestren las características del producto a buscar, esto nos dará la certeza que podremos realizar el análisis de la información. En ese marco decidimos explorar en la búsqueda de empresas dedicadas a la venta y alquiler de departamentos, observamos que este rubro es muy amplio así que decidimos focalizarnos en el alquiler de departamentos. Asimismo, la industria inmobiliaria genera una gran cantidad de datos que pueden ser manipulados por los analistas debido a que la industria está ligada a diversos factores económicos y sociales. Actualmente, la mayor fuente de datos inmobiliarios son los sitios web que muestran la disponibilidad de propiedades para la compra, la venta y el alquiler. Debido a la gran cantidad de datos y al flujo continuo de datos en tiempo real, optamos por trabajar con una de las empresas más grandes del sector inmobiliario peruano, Properati.

La empresa inmobiliaria PROPERATI fue fundada en Argentina 2012 y tiene operaciones en Sudamérica (Argentina, Colombia, Ecuador, Perú y Uruguay) y se encuentra posicionada en el top 5 de las mejores empresas inmobiliarias. Inicialmente tenían operaciones en países como Brasil, México y Chile, pero con la adquisición de OLX Group en junio de 2018, estas operaciones fueron cerradas y se establecieron en Argentina, Colombia, Ecuador, Perú y Uruguay. En enero 2022, la firma japonesa Lifull Connect completó la adquisición total de Properati de OLX Group.

Properati, recopila y analiza datos de su sitio web en tiempo real para brindar a los clientes la información más reciente sobre precios, ubicaciones y otras características. También brinda servicios de marketing para promocionar las propiedades de manera analítica y basada en datos.

El portal de empresas PROPERATI en Perú muestra una variedad de propiedades con las que puede interactuar con los clientes, tales como la ubicación, el precio y otros criterios. Este portal también permite mostrar fotografías y descripción detallada a los posibles clientes o inquilinos. Properati proporciona algunas estadísticas sobre la cobertura en Perú aquí.

Enlace: https://www.properati.com.pe/

2. Título

- “Información del alquiler de inmuebles de la ciudad de Lima-Perú a través de la plataforma Properati”.

3. Descripción del dataset

El dataset contiene información de alquiler inmobiliario en la ciudad de Lima que nos va a permitir realizar análisis, investigación, predicción y toma de decisiones para una mejor selección del inmueble a alquilar. Las principales variables que se recojan serán:

- Precio de alquiler.
- Cantidad de habilitaciones.
- Ubicación.
- Antigüedad del inmueble.
- La inmobiliaria que ofrece el inmueble.

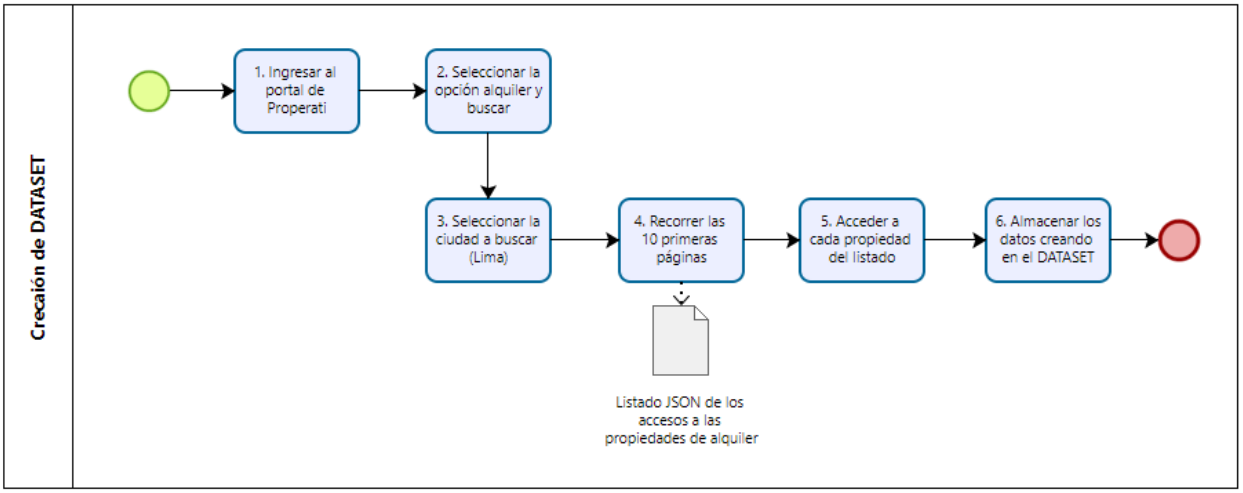
Analizando las variables podemos obtener:

- Análisis de precios. - Ayuda a los propietarios o corredores de alquiler de propiedades asignar precios adecuados y competitivos.
- Análisis de preferencias. - Permite a los propietarios realizar adecuaciones a los inmuebles teniendo en cuenta la temporalidad y condiciones económicas.
- Análisis de la oferta. - Ayuda a tener la información al propietario de la disponibilidad de inmuebles a alquilar de otros inmuebles similares.
- Análisis de la demanda de propiedades. - Nos permite determinar la demanda de propiedades en distintos distritos de Lima e identificar aquellos con mayor demanda.
- Análisis de marketing. – Permite evaluar el alcance y la efectividad de los canales de marketing y publicidad.

Cabe mencionar que la información que se almacenará en el dataset no ha pasado por un proceso de limpieza de información por ello es probable que se encuentren datos nulos o valores atípicos. El archivo resultante será un archivo de texto CSV para que pueda ser importado por cualquier herramienta de análisis de datos.

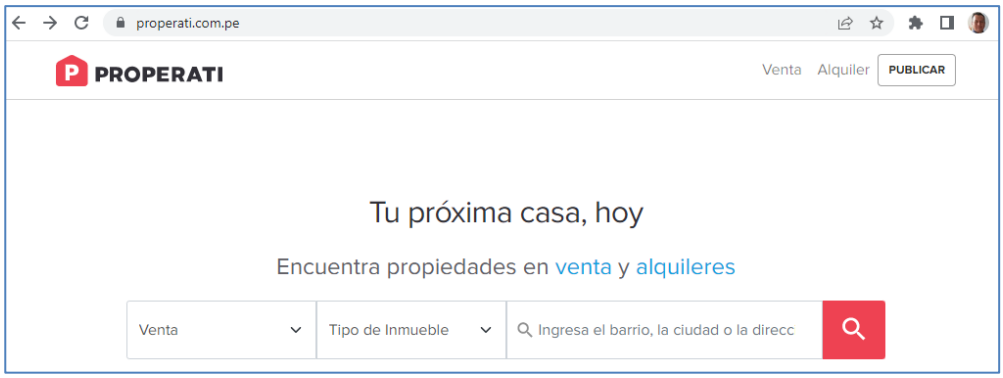
4. Representación gráfica

En la imagen siguiente se muestra el flujo de procesos como se ingresa al portal de PROPERTI-Perú y los pasos a seguir para poder obtener el DATASET que se requiere.

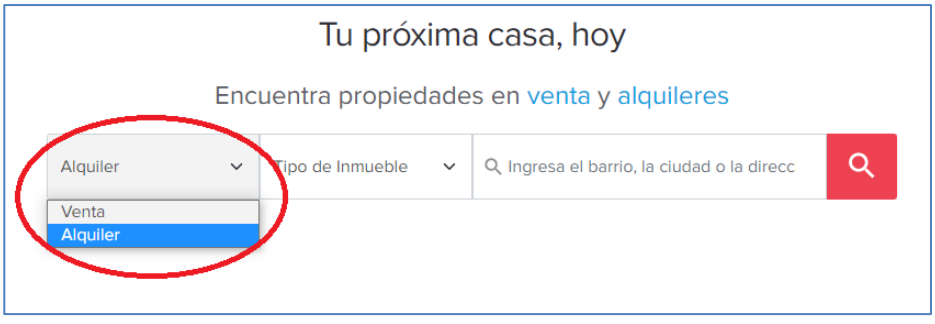


Las capturas de las pantallas son las siguiente con las cuales se generará el dataset:

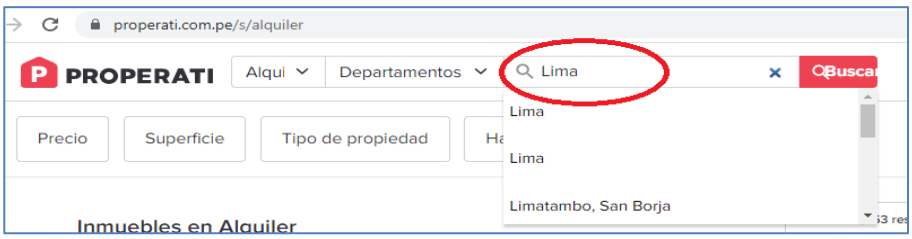
a) Ingresar al portal de Properati



b) Seleccionar la opción alquiler y buscar



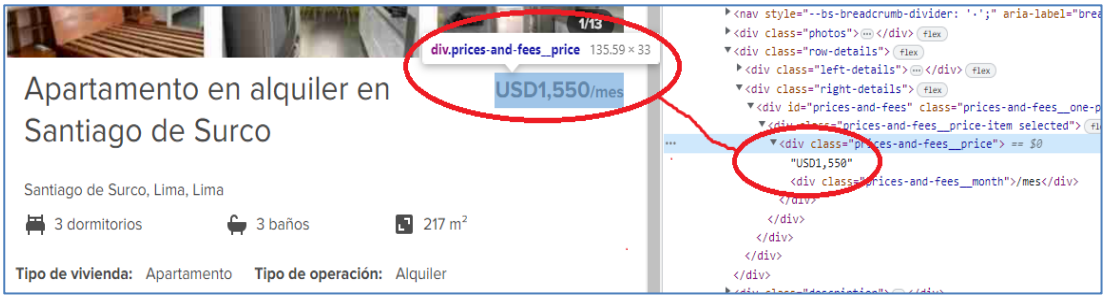
c) Seleccionar la ciudad a buscar (Lima)



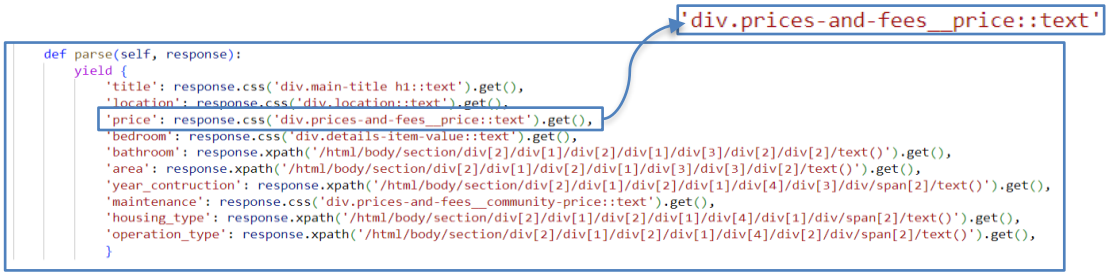
d) Recorrer las 10 primeras páginas. - en la parte inferior tiene la opción para recorrer página a página.

e) Acceder a cada propiedad del listado. - en cada accesos del listado podemos observar a detalle la información del inmueble que se desea alquilar.

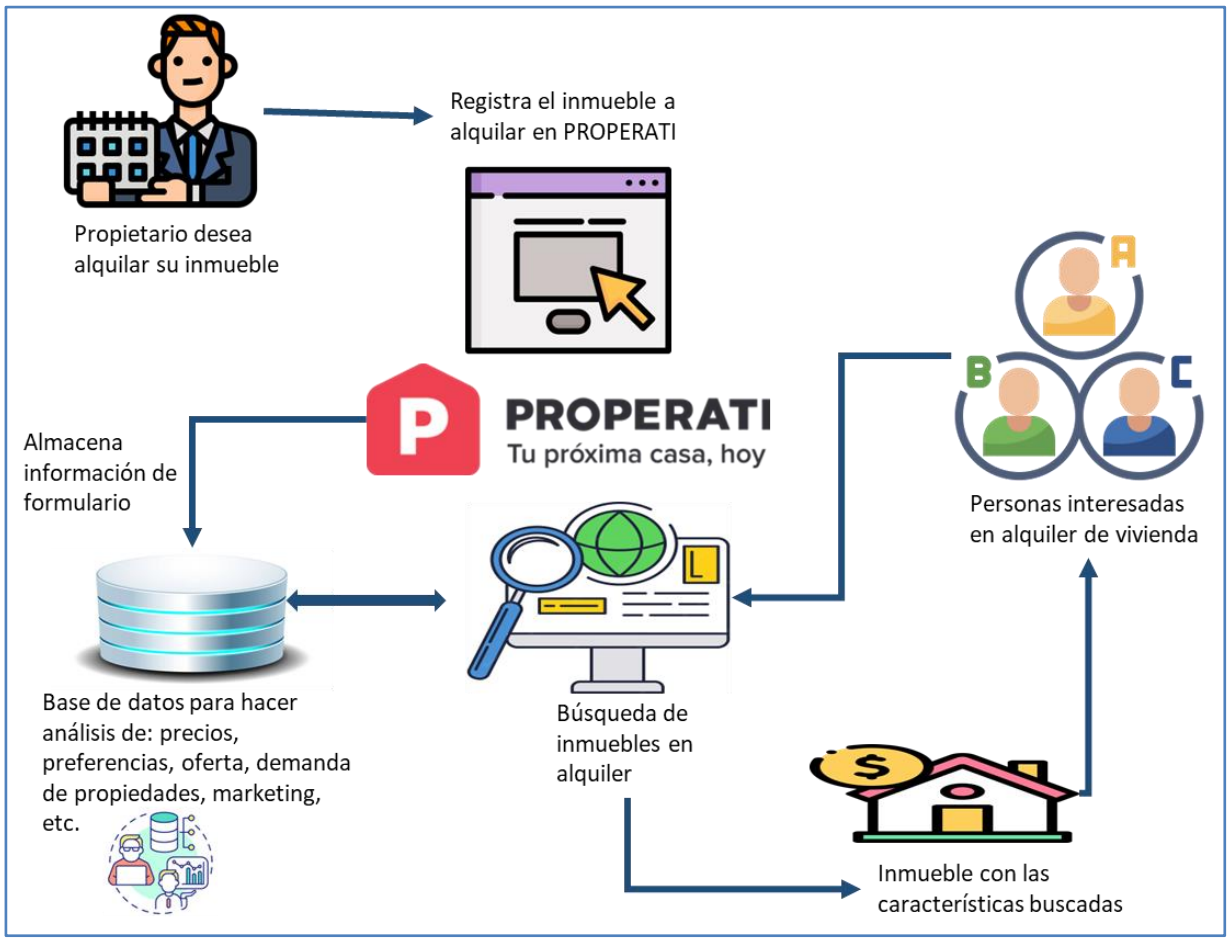
f) Almacenar los datos creando en el DATASET. - con la opción de inspeccionar se inicia hacer la búsqueda del contenido, para identificar en que tag HTML o CSS se encuentra ubicado la información que requerimos capturar.



g) Posteriormente a la identificación del tag HTML o CSS se pasará a digitar en el script de Python, para su posterior ejecución.



En el diagrama siguiente representamos la interacción de la plataforma PROPERATI:



5. Contenido.

Los atributos del contenido del DATASET son los siguientes:

- ✓ title: título del anunció de alquiler.
- ✓ location: dirección del inmueble.
- ✓ price: pecio de alquiler.
- ✓ bedroom: número de habitaciones.
- ✓ bathroom: número de baños.
- ✓ area: área del inmueble.
- ✓ year_contruction: año de construcción.
- ✓ maintenance: costo de mantenimiento en el edificio del inmueble.
- ✓ housing_type: tipo de inmueble.
- ✓ operation_type: tipo de operación alquiler o venta. En nuestro caso solo es alquiler.
- ✓ date_pub: fecha de publicación del anuncio.
- ✓ url: link para acceder al inmueble.

6. Propietario.

El dueño de la empresa PROPERATI es la empresa japonesa LIFULL Connect, el CEO en Latinoamérica es Mauricio Silber.

El propietario de los datos es de los dueños de los inmuebles la empresa PROPERATI es el contenedor de la información que se registra en su portal, en su página principal indica que todos los derechos son reservados de la empresa.

Por otra parte, se ha encontrado que en Argentina (Sudamérica) se ha realizado un análisis con información similar de la empresa denominado “Análisis mercado inmobiliario de la Ciudad de Buenos Aires” [1] de acuerdo al enunciado de este ítem esta sería la justificación del uso de información de la empresa PROPERATI. Entre otros estudios realizados con Properati tenemos:

- Publicaciones realizadas por el mismo Properati[2].
- Análisis exploratorio de un dataset de precios de propiedades[3].
- Análisis exploratorio e introducción a Regresión lineal[4].

[1] *Link de acceso a dataset de kaggle https://www.kaggle.com/code/federicorichardok/primeros-pasos-datascience-properati/notebook.*
[2] *Link de acceso: https://blog.properati.com.pe/category/properati-data/data/*
[3] *Link de acceso: https://github.com/mauriciomem/DS_desafio_1_Properati*
[4] *Link de acceso: https://rstudio-pubs-static.s3.amazonaws.com/537891_b15d23a5635e4602af06057104cd525c.html*
[5] *Link de acceso: https://www.kaggle.com/code/jazmin/analisis-explorarorio-properati/notebook*

- Análisis Exploratorio Properati[5].

En el proyecto se ha tenido en cuenta lo siguiente:

- Rastrear sólo información pública: se ha accedido a información libre, sin restricción de accesos a los datos, que soliciten la creación de una cuenta o aceptar las políticas de uso.
- No causar daño: no saturar el servidor con peticiones, no tratar de acceder a los servidores donde se encuentra la información.
- Utilizar la información de manera justa: no se va usar la información con fines comerciales, sólo son con fines académicos.

7. Inspiración.

Con la búsqueda de la información de inmuebles de alquiler se desea responder las siguientes preguntas:

- ¿Cómo va variando el precio del alquiler en el tiempo?
- ¿Cuáles son las diferencias de precios de alquiler de los inmuebles que se encuentran ubicados en avenidas principales y cuales en calles no principales?
- ¿Existe una variación del precio del alquiler en base a un mes en particular? Deseamos identificar temporalidad.
- ¿Cuál es la relación de metros cuadros por el precio de alquiler?
- ¿Cuál es la ratio del precio de alquiler respecto al número de habitaciones?
- ¿Cómo influye la antigüedad del inmueble en el costo de mantenimiento?
- ¿Cuáles son los distritos con mayor demanda en el alquiler inmuebles?

El análisis que se puede realizar para responder las preguntas planteadas son:

- Análisis de precios.
- Análisis de temporalidad de alquiler.
- Análisis de ubicación del inmueble.
- Análisis del área del inmueble.
- Análisis de antigüedad de los inmuebles.
- Análisis de demanda y oferta de propiedades.

8. Licencia.

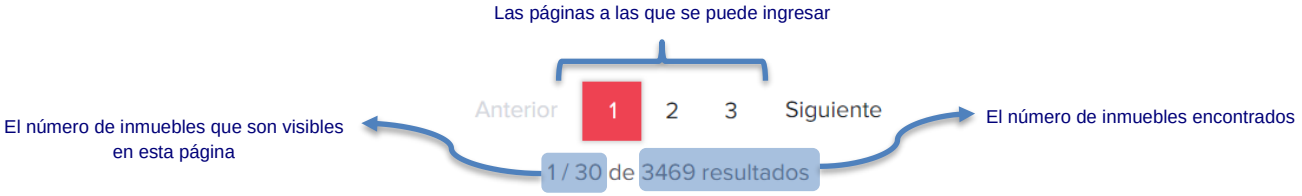
El tipo de licencia es “Database released under Open Database License, individual contents under Database Contents License” si bien es cierto el portal PROPERTI es el contenedor, pero esta información se alimenta de dueños de inmuebles que ponen a disposición los inmuebles que desean alquilar de manera que ellos permiten el acceso de dicha información.

No obstante, PROPERATI podrá en determinadas circunstancias limitadas, conceder el acceso al control de sus datos, lo cual se acuerda, a través de un acuerdo legal que garantice el cumplimiento de todas las finalidades pactadas.

9. Código.

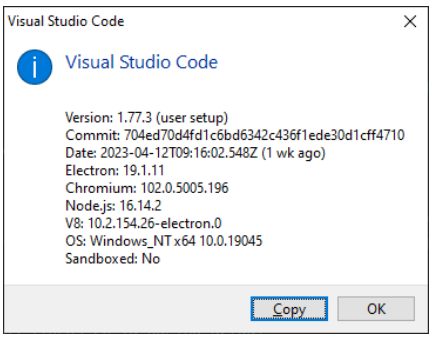
Entre los principales retos que se presentaron en el sitio web, tenemos:

- El primer desafío al que nos enfrentamos fue generar un código que nos permitiera navegar de una página del sitio web a otra. Es decir, el sitio web cuenta con miles de propiedades que se dividen en páginas que contienen registros parciales. En este caso, tenemos registros de 30 propiedades en cada página, otras 30 propiedades en la página siguiente y así sucesivamente. El desafío es profundizar en cada página y obtener la información sobre cada propiedad.



- El segundo desafío fue poder ingresar a los detalles de cada propiedad. La página generalmente muestra algunas descripciones de las propiedades, pero si se desea obtener más información sobre una propiedad, como costos de mantenimiento, fechas de construcción, etc., deberá ingresar a la página de la propiedad donde se puede encontrar mayor información de la propiedad.

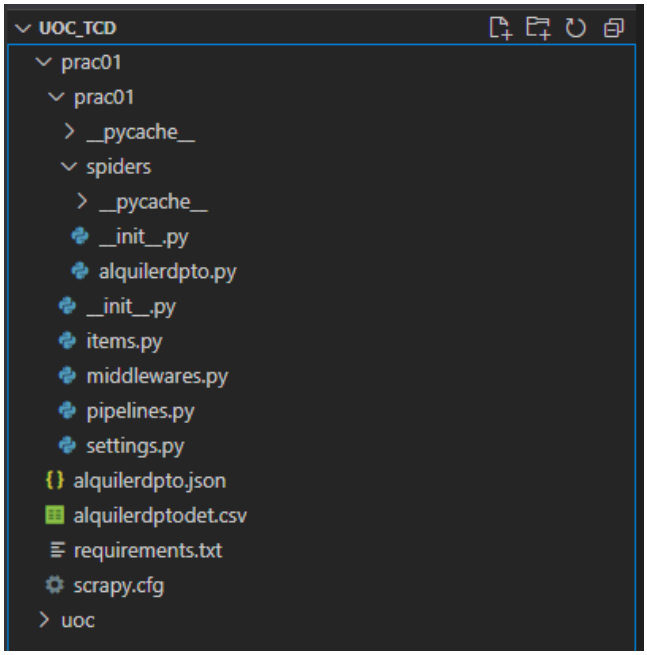
El desarrollo del proyecto se realizó en el Visual Studio Code en el leguaje Python.



Para crea el proyecto se siguió los pasos siguientes:

- Paso 1: Creación de la carpeta “uoc_tcd”
- Paso 2: En dicha carpeta se creo el entorno virtual uoc: python -m venv uoc
- Paso 3: Se instaló scrapy: pip install scrapy
- Paso 4: Se generó un proyecto scrapy denominado prac01: scrapy startproject prac01
- Paso 5: Se creo el spider denominado “alquilerdpto”: scrapy genspider alquilerdpto www.properati.com.pe
- Paso 6: Se procedió a codificar las sentencias necesarias para scrapear la información requerida.

Imagen de estructura de carpetas y archivos del proyecto en el Visual Studio Code



Se está subiendo al GITHUB el proyecto generado denominado PRAC01. Se sube 2 repositorios, uno privado y el otro público, esto debido a que se pueden presentar problemas al brindar los permisos.

Link GITHUB de la ruta del proyecto scrapy (public): https://github.com/fmucham/source
Link GITHUB de la ruta del proyecto scrapy (private): https://github.com/JoseC468/P1-Web-Scraping

10. Dataset

DOI: 10.5281/zenodo.7846211
Link de ruta: https://doi.org/10.5281/zenodo.7846211

11. Vídeo

Para poder ingresar a la ruta del video es necesario iniciar sesión con una cuenta de la institución (UOC).

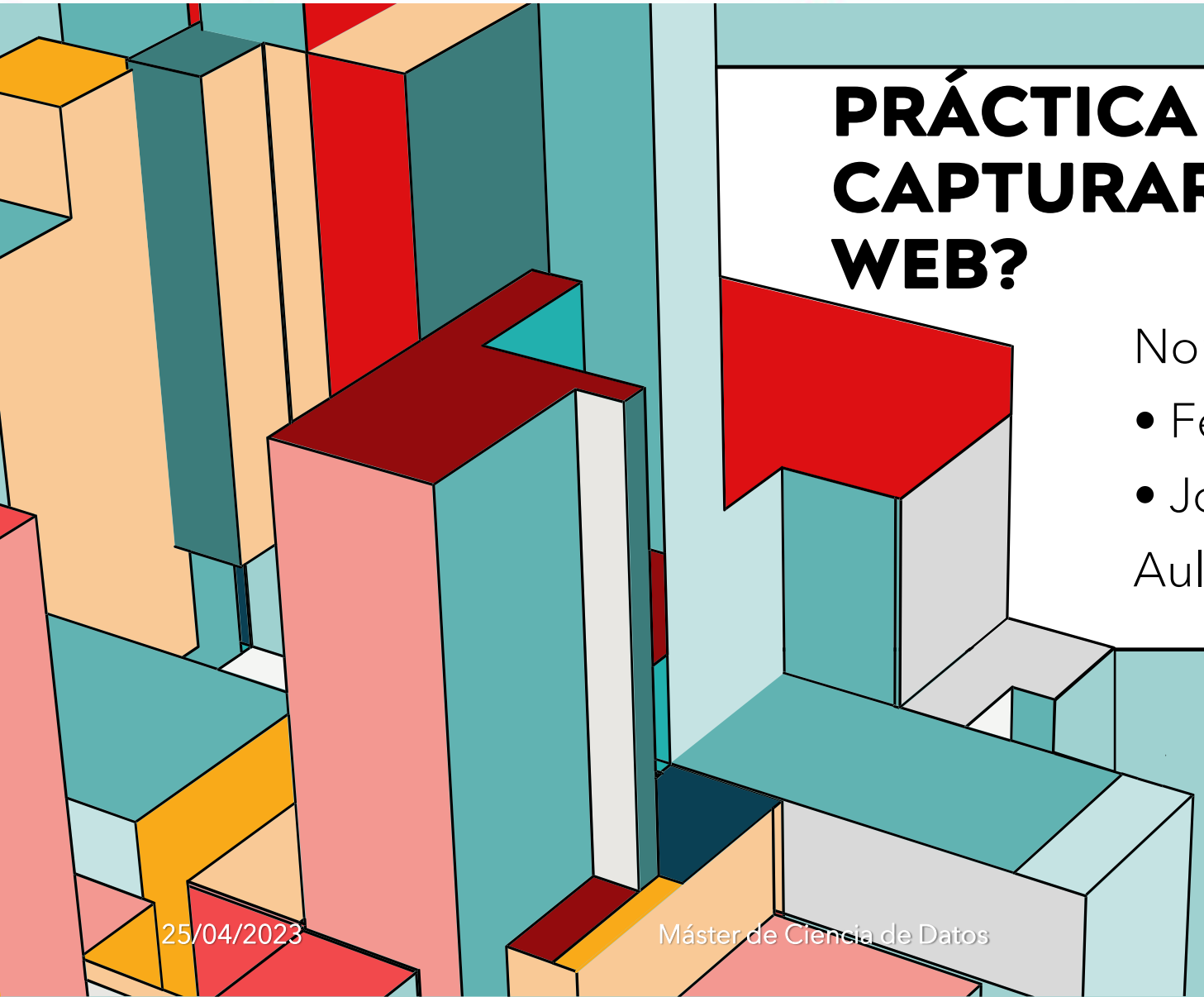
Link de la ruta: https://drive.google.com/drive/folders/1s1J4rf-N2MMlXh9XCSoZXOheoR5NJ5Nd

12. Tabla de contribuciones

Contribución	Firma 1	Firma 2
Investigación previa	**FAMM**	**JCEL**
Redacción de las respuestas	**FAMM**	**JCEL**
Desarrollo del código	**FAMM**	**JCEL**
Participación en el video	**FAMM**	**JCEL**

Referencia Bibliográfica

- Subirats, L., Calvo, M. (2018). Web Scraping. Editorial UOC.
- Masip, D. (2019). El lenguaje Python. Editorial UOC.
- Lawson, R. (2015). Web Scraping with Python. Packt Publishing Ltd. Chapter 2. Scraping the Data.
- Simon Munzert, Christian Rubba, Peter Meißner, Dominic Nyhuis. (2015). Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining. John Wiley & Sons.



PRÁCTICA 1: ¿CÓMO PODEMOS CAPTURAR LOS DATOS DE LA WEB?

Nombres de los estudiantes:

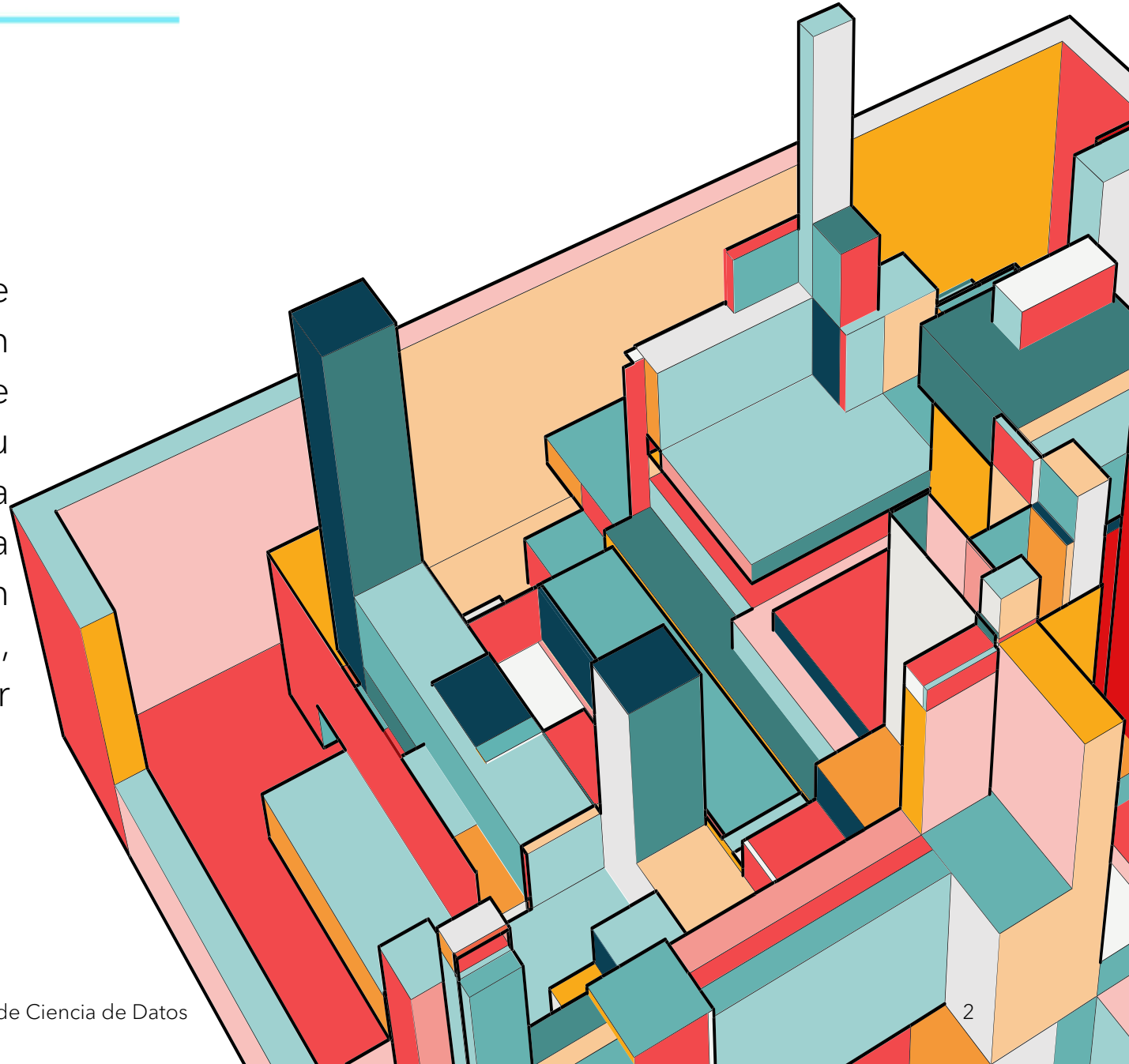
- Félix Antonio Mucha Morales
- José Carlos Enriquez Lira

Aula 2

1. CONTEXTO

Para la práctica se busco un negocio que ofrezca y muestren las características de un producto, que nos de la certeza que podremos realizar el análisis de su información, por ello se selecciono a la empresa PROPERATI que muestra una variedad de bienes inmuebles con información de ubicación, precio, habitaciones y otros, que permiten tomar decisiones de compra o alquiler.

PROPERATI
Tu próxima casa, hoy



2. TÍTULO

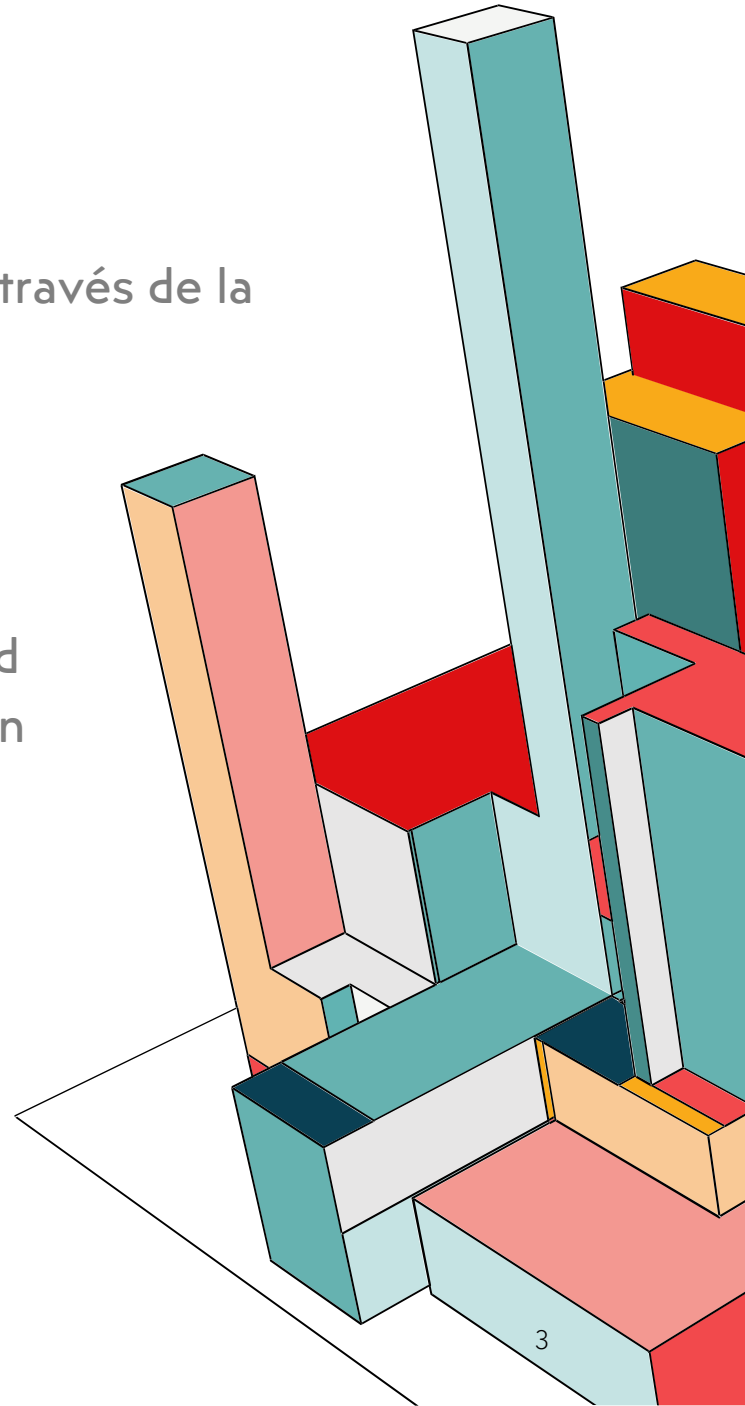
Información del alquiler de inmuebles de la ciudad de Lima-Perú a través de la plataforma PROPERATI

3. DESCRIPCIÓN DEL DATASET

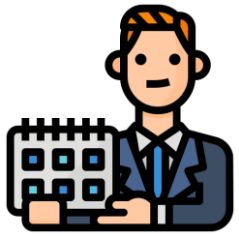
El dataset contiene información de alquiler inmobiliario de la ciudad de Lima que nos permitirá realizar análisis, investigación, predicción y toma de decisiones.

Las principales variables principales son:

- Precio de alquiler.
- Cantidad de habilitaciones.
- Ubicación.
- Antigüedad del inmueble.
- La inmobiliaria que ofrece el inmueble.



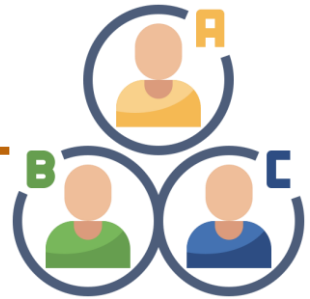
4. REPRESENTACION GRÁFICA



1. Propietario desea alquilar su inmueble

2. Registra el inmueble a alquilar en PROPERATI

3. Almacena información de formulario



4. Personas interesadas en alquiler de vivienda

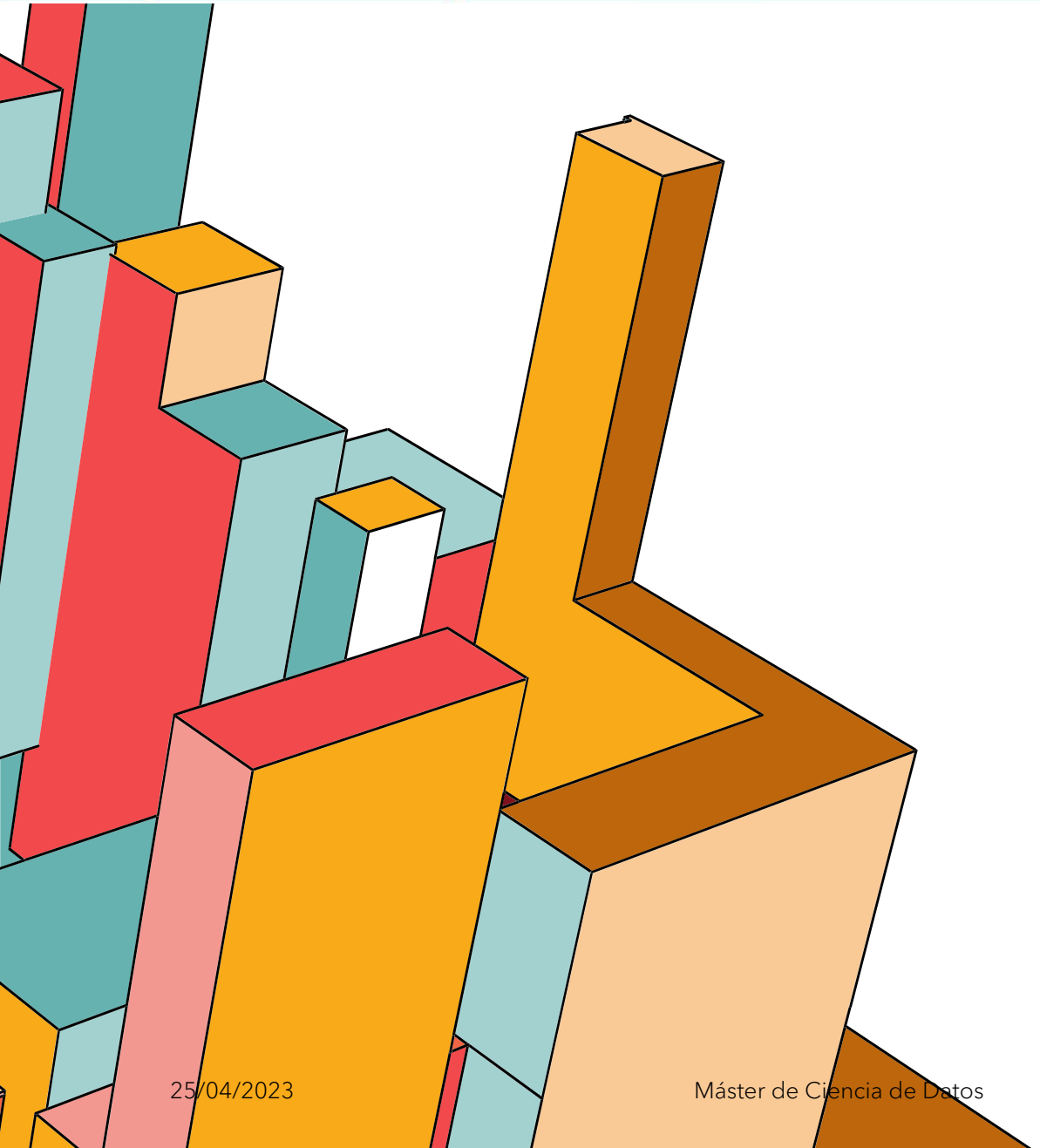
5. Búsqueda de inmuebles en alquiler

6. Inmueble con las características buscadas

7. Base de datos para hacer análisis de: precios, preferencias, oferta, demanda de propiedades, marketing, etc.

5. CONTENIDO

- title: título del anunció de alquiler.
- location: dirección del inmueble.
- price: pecio de alquiler.
- bedroom: número de habitaciones.
- bathroom: número de baños.
- area: área del inmueble.
- year_contruction: año de construcción.
- maintenance: costo de mantenimiento.
- housing_type: tipo de inmueble.
- operation_type: tipo de operación alquiler o venta. En nuestro caso solo es alquiler.
- date_pub: fecha de publicación del anuncio.
- url: link para acceder al inmueble.



6. PROPIETARIO

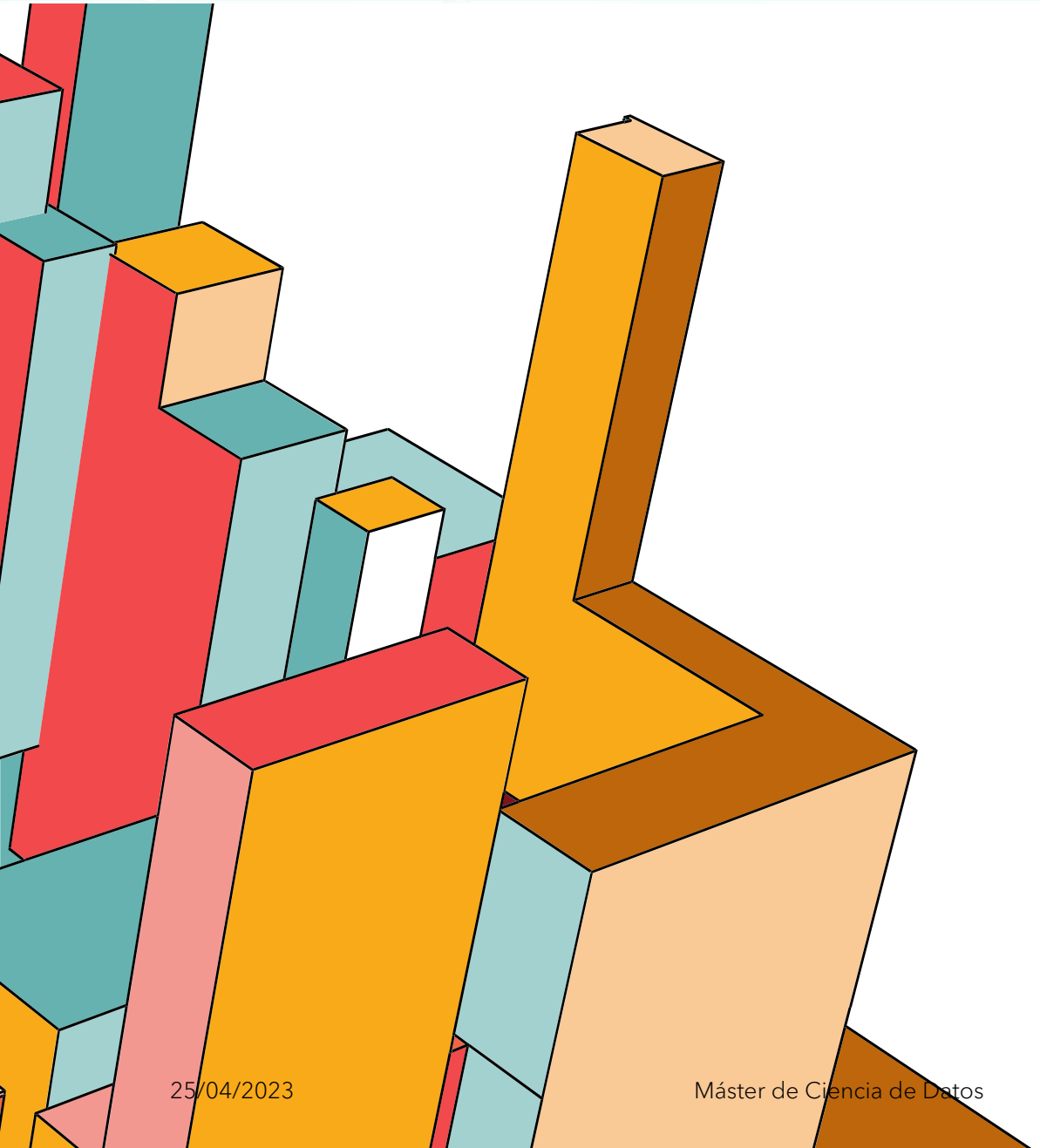
Mauricio Silber, CEO de LIFULL Connect

Otros estudios de la información:

- Análisis mercado inmobiliario de la Ciudad de Buenos Aires
- Publicaciones realizadas por el mismo PROPERATI
- Análisis exploratorio e introducción a Regresión lineal
- Análisis Exploratorio PROPERATI

Consideraciones durante el desarrollo:

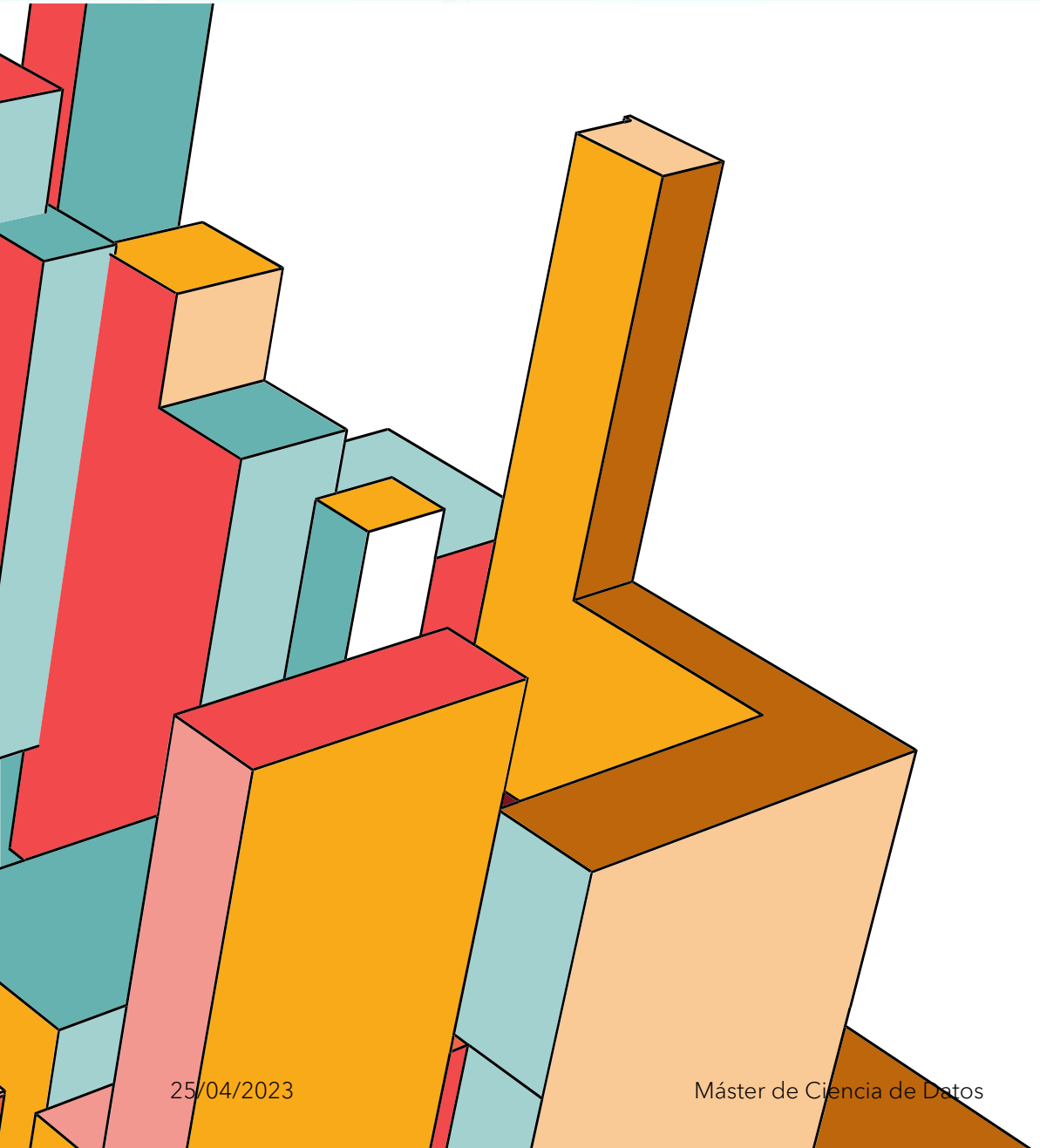
- Rastrear sólo información pública
- No causar daño
- Utilizar la información de manera justa



7. INSPIRACIÓN

Algunas preguntas que se desea responder:

- ¿Cómo va variando el precio del alquiler en el tiempo?
- ¿Existe una variación del precio del alquiler en base a un mes en particular? Deseamos identificar temporalidad.
- ¿Cuál es la ratio del precio de alquiler respecto al número de habitaciones?
- ¿Cómo influye la antigüedad del inmueble en el costo de mantenimiento?



8. LICENCIA

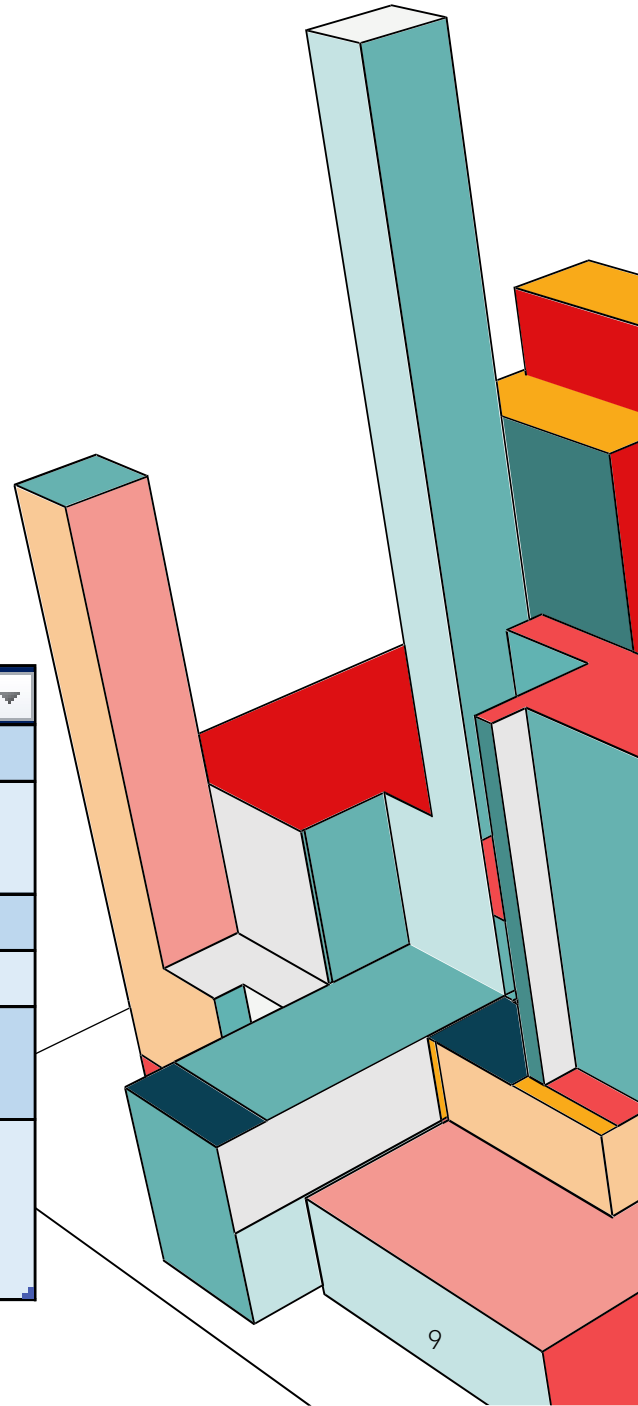
"Database released under Open Database License, individual contents under Database Contents License"

El portal PROPERATI es un contenedor de información de inmuebles de venta y alquiler de acceso libre. Dichos inmuebles son registrados por los mismo propietarios o corredores inmobiliarios en este portal.

9. CÓDIGO

Pasos para creación del proyecto:

Pasos	Sentencia
1. Crear carpeta del proyecto	mkdir uoc_tcd
2. Crear entorno virtual para el proyecto	python -m venv uoc
3. Instalar scrapy	pip install scrapy
4. Generar proyecto scrapy	scrapy startproject prac01
5. Crear spider denominado “alquilerdpto”	scrapy genspider alquilerdpto www.properati.com.pe
6.Se procedió a codificar las sentencias necesarias para scrapear la información requerida.	import scrapy from scrapy.spiders import CrawlSpider, Rule from scrapy.linkextractors import LinkExtractor



MUCHAS GRACIAS

Félix Antonio Mucha Morales

José Carlos Enriquez Lira