



ESCUELA POLITÉCNICA NACIONAL

ESCUELA DE FORMACIÓN DE TECNÓLOGOS



ANÁLISIS DE DATOS

ASIGNATURA:

Análisis de Datos

PROFESOR:

Ing. Lorena Chulde

PERÍODO ACADÉMICO:

2023-B

LECCIÓN

TÍTULO:

WEB SCRAPING con MySQL

ESTUDIANTE

Marcelo Pinzón



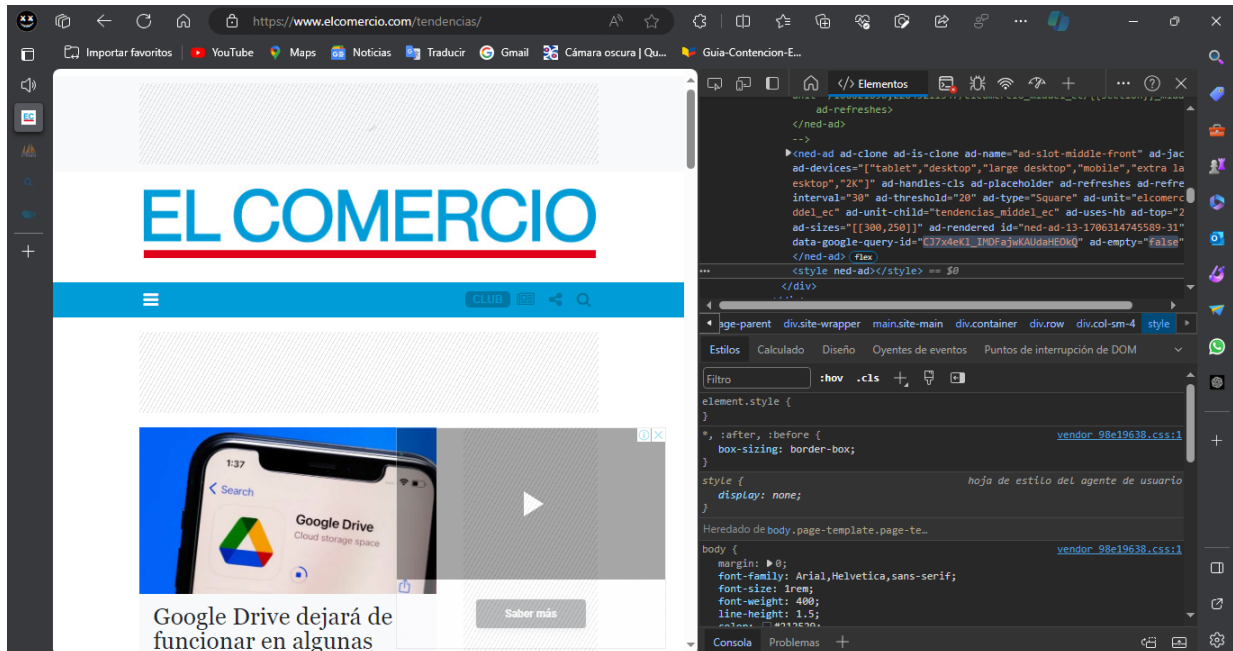
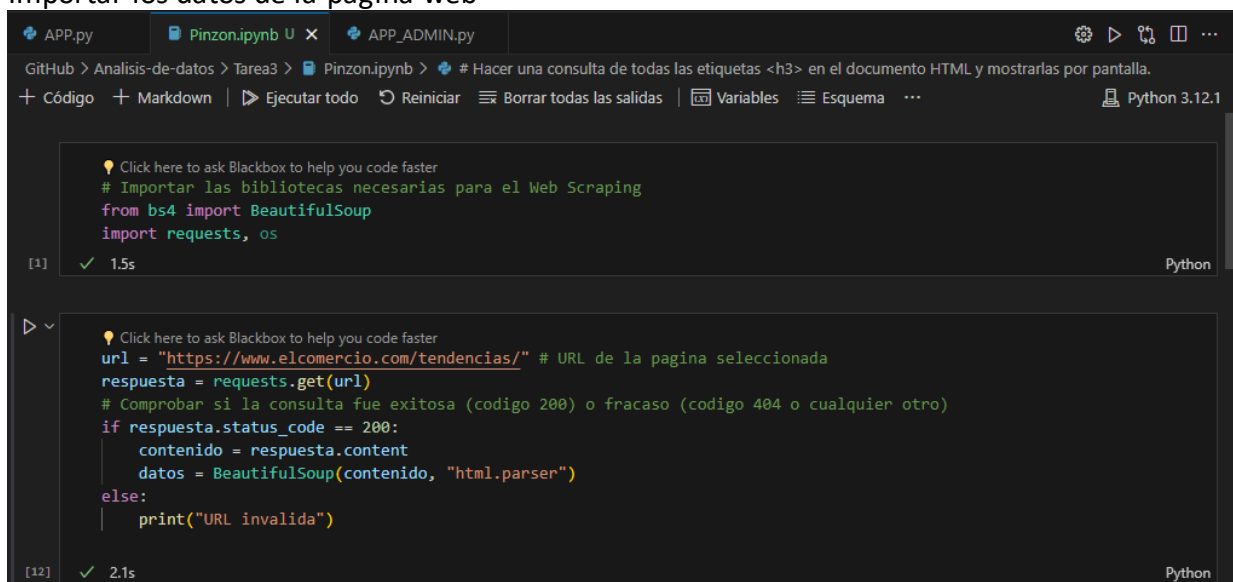
2023 - B

INDICACIONES:

El sitio que usted elija.

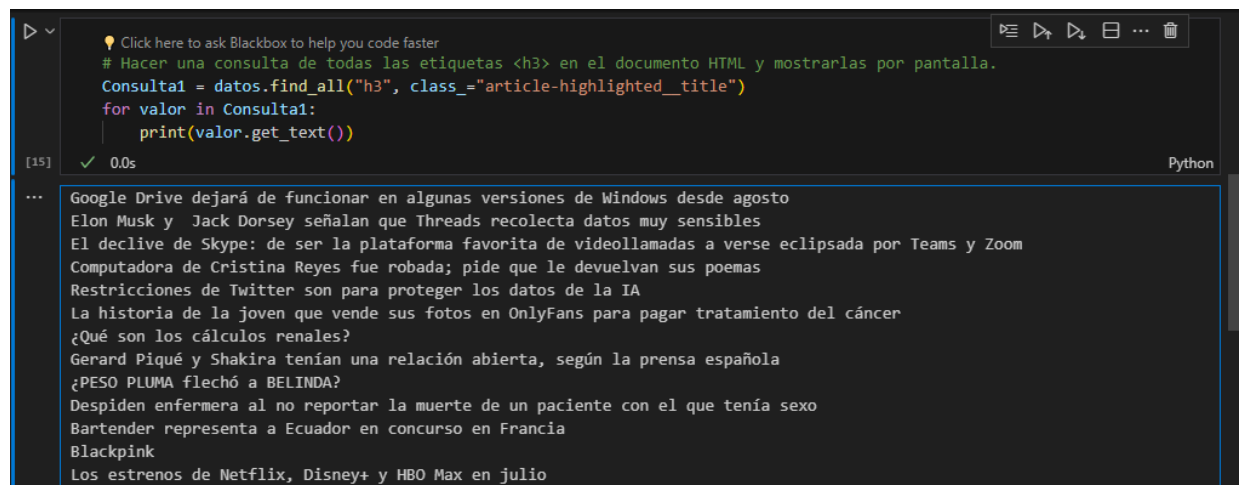
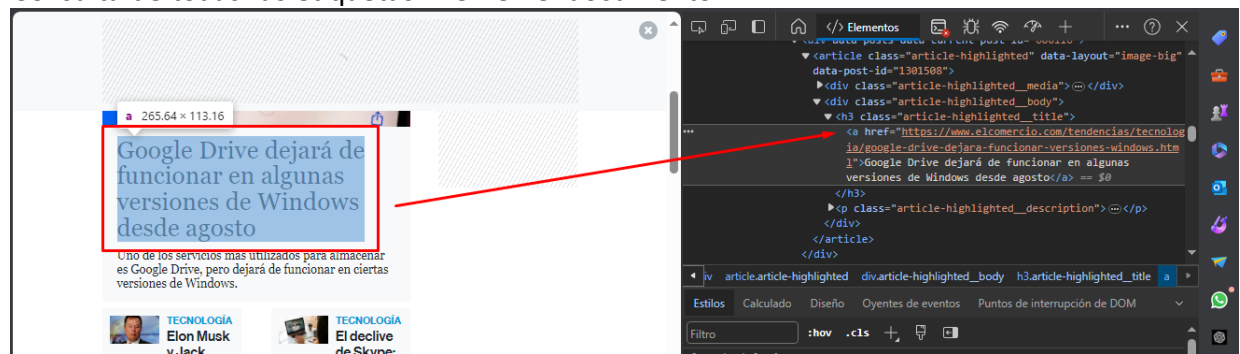
1. Importe la data de toda la página

Página web escogida: [Tendencias - El Comercio](https://www.elcomercio.com/tendencias/)

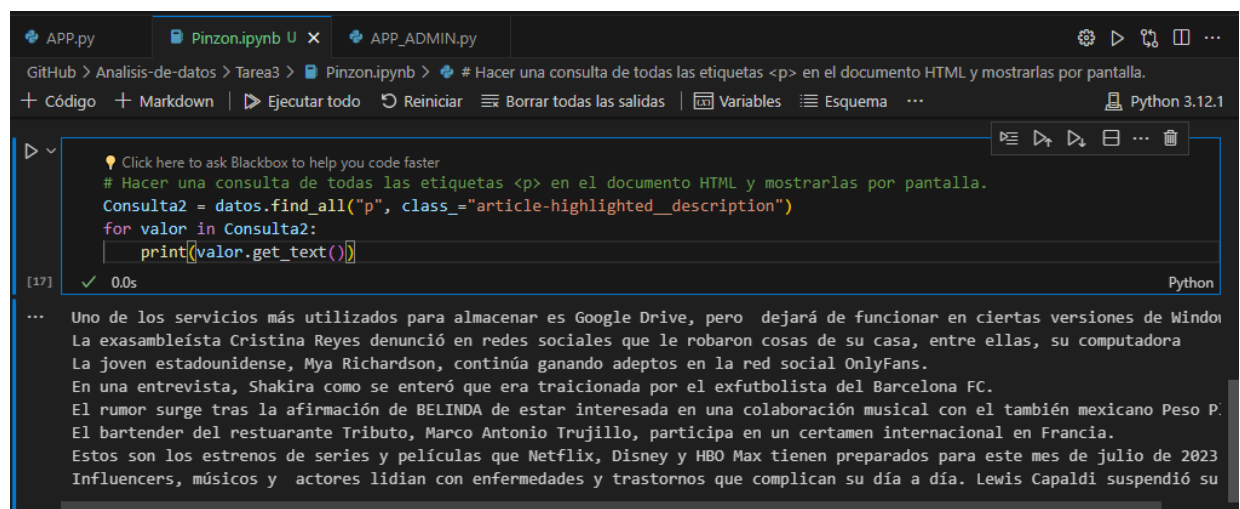
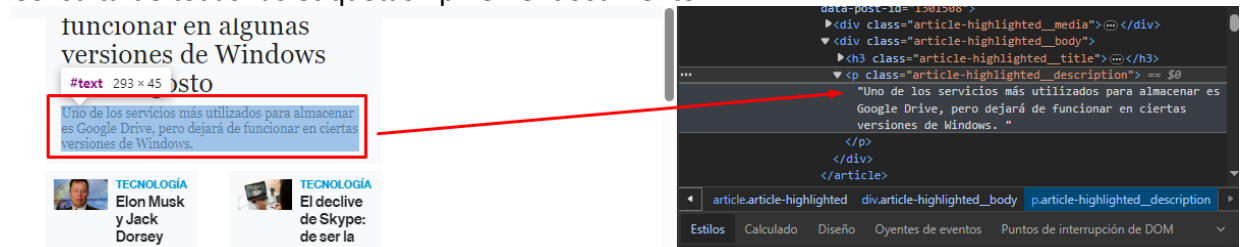
**Importar los datos de la página web**

2. Importe 8 elementos de HTML y obtenga la información de cada uno de ellos

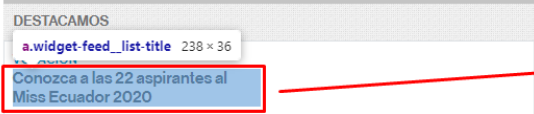
Consulta de todas las etiquetas <h3> en el documento



Consulta de todas las etiquetas <p> en el documento



Consulta de una etiqueta <a> presente en el documento



DESTACAMOS

[Conozca a las 22 aspirantes al Miss Ecuador 2020](#)

```

<div class="p-2 w-50">
  <div class="widget-feed_list-category text-primary">
    Votación</div>
    <a href="/tendencias/candidatas-miss-ecuador-2020-votacion.html" class="widget-feed_list-title">Conozca a las 22 aspirantes al Miss Ecuador 2020</a> == $0
  </div>

```

```

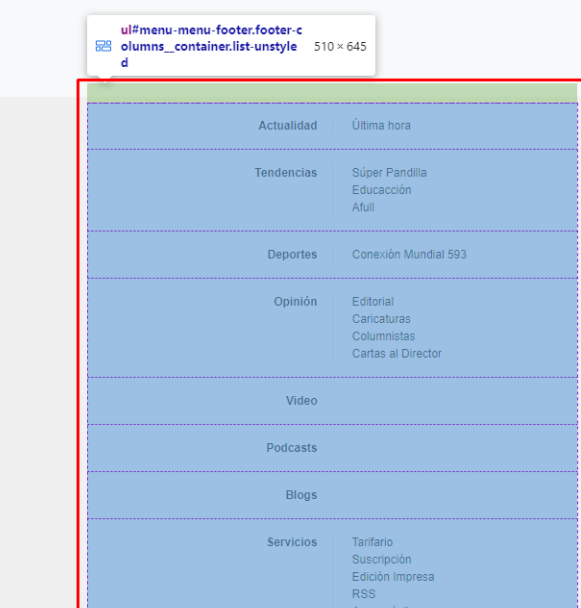
# Hacer una consulta de una de las etiquetas <a> en el documento HTML y mostrarlas por pantalla.
Consulta3 = datos.find("a", class_="widget-feed_list-title")
for valor in Consulta3:
    print(valor.get_text())

```

[18] ✓ 0.0s Python

Conozca a las 22 aspirantes al Miss Ecuador 2020

Consulta de la etiqueta presente en el documento



ul#menu-menu-footer.footer-columns__container.list-unstyled

Actualidad	Última hora
Tendencias	Súper Pandilla Educación Afull
Deportes	Conexión Mundial 593
Opinión	Editorial Caricaturas Columnistas Cartas al Director
Video	
Podcasts	
Blogs	
Servicios	Tarifario Suscripción Edición Impresa RSS Apps móviles

```

<div class="site-footer">
  <div class="container">
    <div class="products">
      <div class="columns">
        <div class="container">
          <ul id="menu-menu-footer" class="footer-columns__container list-unstyled">
            <li></li>
            <li class="current-menu-item"></li>
            <li></li>
            <li></li>
            <li></li>
          </ul>
        </div>
      </div>
    </div>
  </div>

```

```

# Hacer una consulta de una de las etiquetas <ul> en el documento HTML y mostrarlas por pantalla.
Consulta4 = datos.find("ul", class_="footer-columns__container list-unstyled", id="menu-menu-footer")
for valor in Consulta4:
    print(valor.get_text(strip=True, separator=" | "))

```

[22] ✓ 0.0s Python

Actualidad | Última hora

Tendencias | Súper Pandilla | Educación | Afull

Deportes | Conexión Mundial 593

Opinión | Editorial | Caricaturas | Columnistas | Cartas al Director

Video

Podcasts

Blogs

Servicios | Tarifario | Suscripción | Edición Impresa | RSS | Apps móviles | Ediciones Anteriores | Síntesis noticiosa | Fa

```

APP.py Pinzon.ipynb Python: APP APP_ADMIN.py
GitHub > Analisis-de-datos > Tarea3 > Pinzon.ipynb > # Hacer una consulta de una de las etiquetas <ul> en el documento HTML y mostrarlas por pantalla.
+ Código + Markdown | Ejecutar todo | Reiniciar | Borrar todas las salidas | Variables | Esquema ... Python 3.12.1

```

```

# Hacer una consulta de una de las etiquetas <ul> en el documento HTML y mostrarlas por pantalla.
Consulta4 = datos.find("ul", class_="footer-columns__container list-unstyled", id="menu-menu-footer")
for valor in Consulta4:
    print(valor.get_text(strip=True, separator=" | "))

```

[22] ✓ 0.0s Python

Actualidad | Última hora

Tendencias | Súper Pandilla | Educación | Afull

Deportes | Conexión Mundial 593

Opinión | Editorial | Caricaturas | Columnistas | Cartas al Director

Video

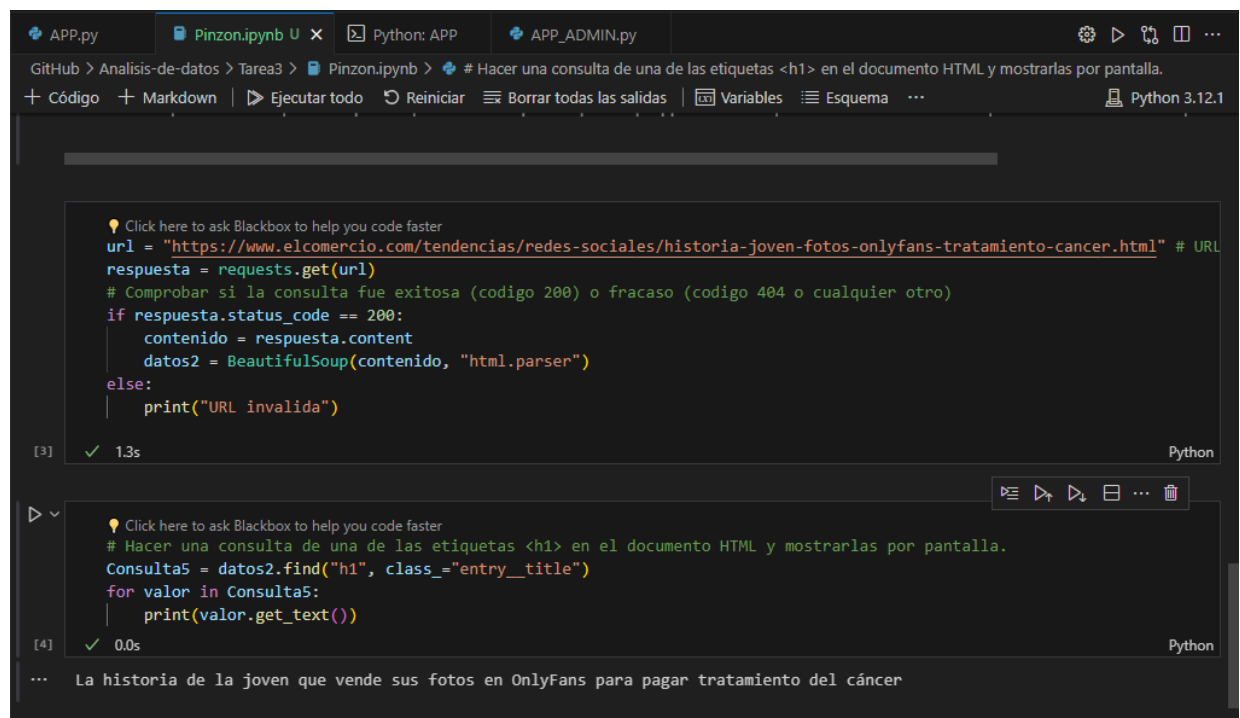
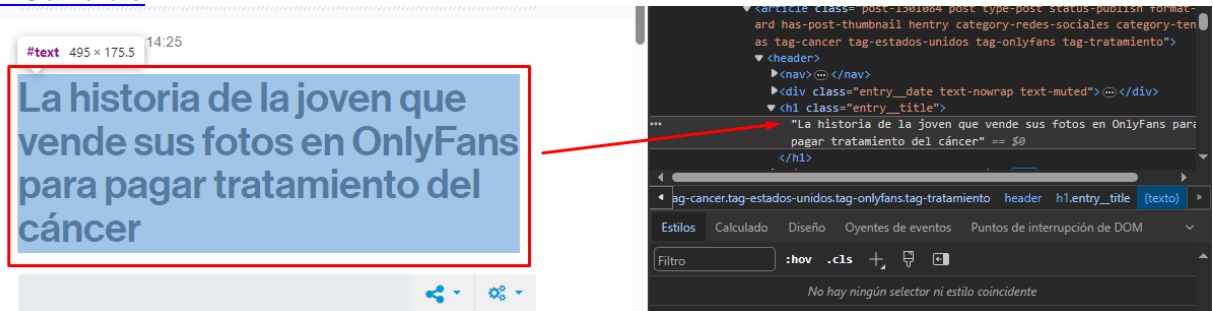
Podcasts

Blogs

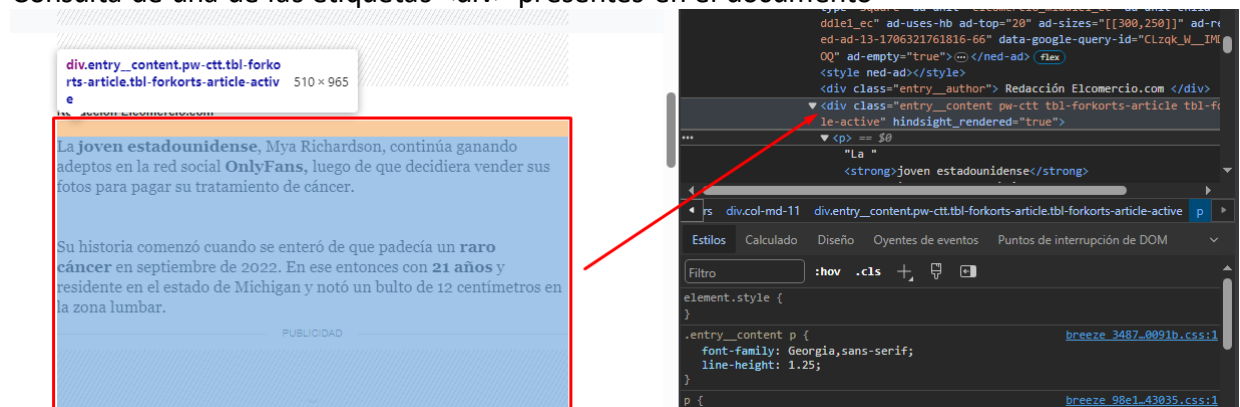
Servicios | Tarifario | Suscripción | Edición Impresa | RSS | Apps móviles | Ediciones Anteriores | Síntesis noticiosa | Fa

Consulta de la etiqueta <h1> presente en el documento con el enlace

[La historia de la joven que vende sus fotos en OnlyFans para pagar tratamiento del cáncer - El Comercio](https://www.elcomercio.com/tendencias/redes-sociales/historia-joven-fotos-onlyfans-tratamiento-cancer.html)



Consulta de una de las etiquetas <div> presentes en el documento



The screenshot shows a Jupyter Notebook interface with a Python script and its output. The script is designed to find a specific `<div>` element in an HTML document and print its text content.

```

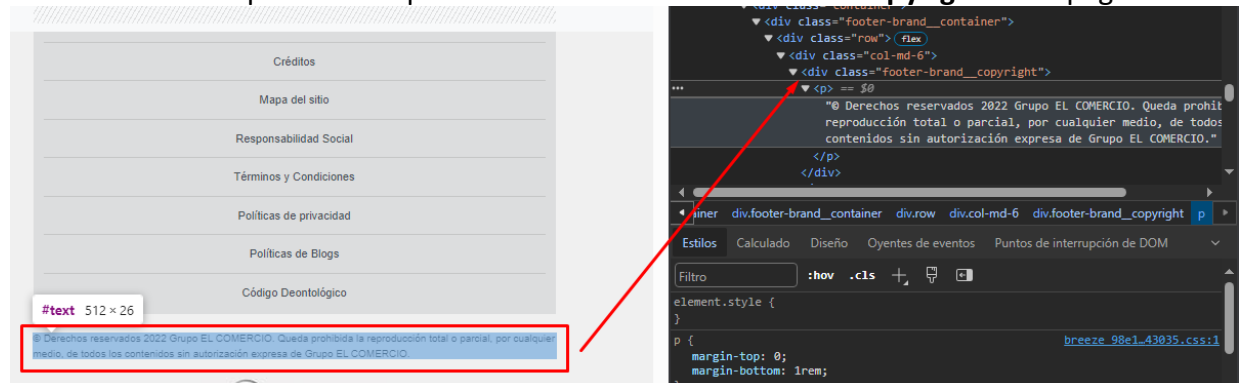
# Hacer una consulta de una de las etiquetas <div> en el documento HTML y mostrarlas por pantalla.
Consulta6 = datos2.find_all("div", class_="entry_content")
for valor in Consulta6:
    print(valor.get_text(strip=True, separator="\n"))

```

The output of the script is a block of text, which is a news article snippet about a young man named Mya Richardson.

La joven estadounidense, Mya Richardson, continúa ganando adeptos en la red social OnlyFans, luego de que decidiera vender sus fotos para pagar su tratamiento de cáncer. Su historia comenzó cuando se enteró de que padecía un raro cáncer en septiembre de 2022. En ese entonces con 21 años y residente en el estado de Michigan y notó un bulto de 12 centímetros en la zona lumbar. La joven acudió al especialista y le diagnosticando sarcoma de células fusiformes, un tumor que puede desarrollarse en el hueso o en los tejidos blandos. Gana 80 000 dólares en Onlyfans. Con la intención de pagar las sesiones de quimioterapia y radioterapia que le correspondían, decidió recurrir a vender sus desnudos en OnlyFans. Aunque ya tenía presencia en esta red social antes de ser diagnosticada, mantuvo la noticia en secreto y usaba pelucas. Hasta que se dio cuenta que a sus seguidores les gustaba verla sin pelo. En apenas unos meses, Mya consiguió 80 000 dólares, con los que se ha costado cuatro sesiones de quimioterapia y 25 sesiones de radioterapia. Más noticias en:

Consulta de la etiqueta `<div>` que contiene información sobre el **copyright** de la pagina



The screenshot shows a Jupyter Notebook interface with a Python script and its output. The script is designed to find a specific `<div>` element in an HTML document and print its text content.

```

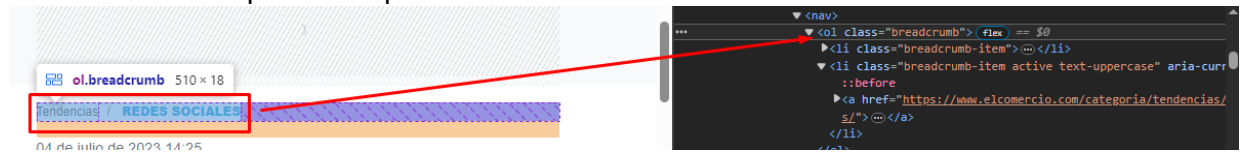
# Hacer una consulta de una de las etiquetas <div> en el documento HTML y mostrarlas por pantalla.
Consulta7 = datos2.find("div", class_="footer-brand__copyright")
for valor in Consulta7:
    print(valor.get_text(strip=True, separator="\n"))

```

The output of the script is the copyright notice from the website footer.

© Derechos reservados 2022 Grupo EL COMERCIO. Queda prohibida la reproducción total o parcial, por cualquier medio, de todos los contenidos sin autorización expresa de Grupo EL COMERCIO.

Consulta de la etiqueta presente en el documento



04 de julio de 2023 14:25

```

<ol class="breadcrumb">
  <li class="breadcrumb-item">
    <li class="breadcrumb-item active text-uppercase">
      <a href="https://www.elcomercio.com/categoria/tendencias/s/">
    </li>
  </li>
</ol>

```

APP.py Pinzon.ipynb Python: APP APP_ADMIN.py

GitHub > Analisis-de-datos > Tarea3 > Pinzon.ipynb > # Hacer una consulta de una de las etiquetas en el documento HTML y mostrarlas por pantalla.

+ Código + Markdown | Ejecutar todo Reiniciar Borrar todas las salidas Variables Esquema Python 3.12.1

Click here to ask Blackbox to help you code faster

Hacer una consulta de una de las etiquetas en el documento HTML y mostrarlas por pantalla.

```

Consulta8 = datos2.find("ol", class_="breadcrumb")
for valor in Consulta8:
    print(valor.get_text(strip=True, separator="\n"))

```

[12] ✓ 0.0s Python

Tendencias
Redes sociales

3. Exporte a la base de datos MySQL

Importar las extracciones a la base de datos "análisis_bd" en MySQL

```

from mysql.connector import MySQLConnection, Error
try:
    conn = MySQLConnection(
        host="localhost",
        user="root",
        password="*****",
        database="analisis_bd"
    )
    if conn.is_connected():
        cursor = conn.cursor()

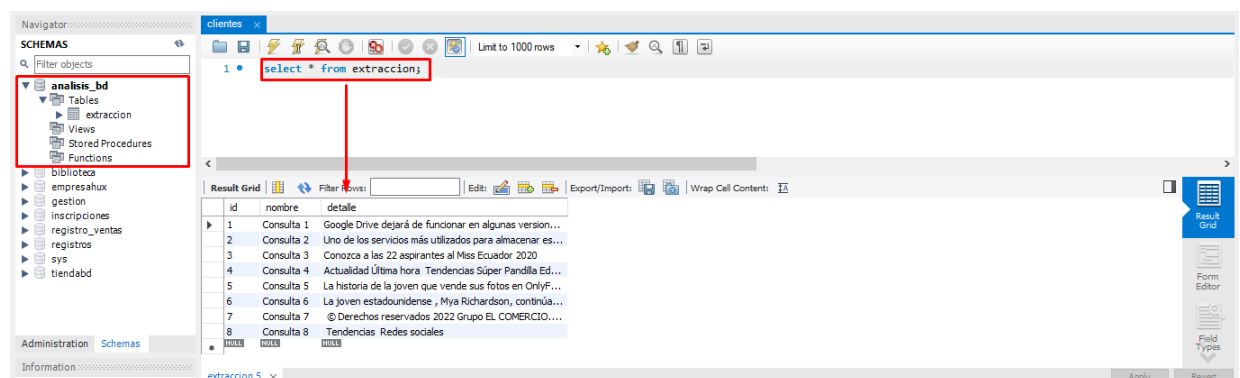
        cursor.execute("""CREATE TABLE IF NOT EXISTS Extraccion(
            id INT PRIMARY KEY AUTO_INCREMENT,
            nombre VARCHAR(50) NOT NULL,
            detalle TEXT)""")

        conn.commit()

        consultas = [Consulta1, Consulta2, Consulta3, Consulta4, Consulta5, Consulta6, Consulta7, Consulta8]
        for i, elementos in enumerate(consultas, 1):
            subelem=str()
            for value in elementos:
                subelem+=f'{value.get_text(strip=True, separator="\n")}\n'
            consulta = "INSERT INTO Extraccion (nombre, detalle) VALUES (%s, %s)"
            cursor.execute(consulta, (f"Consulta {i}", subelem))
            conn.commit()
        conn.close()
    else:
        print("Conexion no establecida!")
except Error as error:
    print("Fallo la conexion debido a:", error)
finally:
    conn.close()

```

[72] ✓ 0.4s Python



Navigator: clientes

Schemas

analisis_bd

Tables

extraccion

Views

Stored Procedures

Functions

biblioteca

empresahux

gestion

inscripciones

registro_ventas

registros

sys

tiendabd

Administration Schemas Information

extraccion 5

Limit to 1000 rows

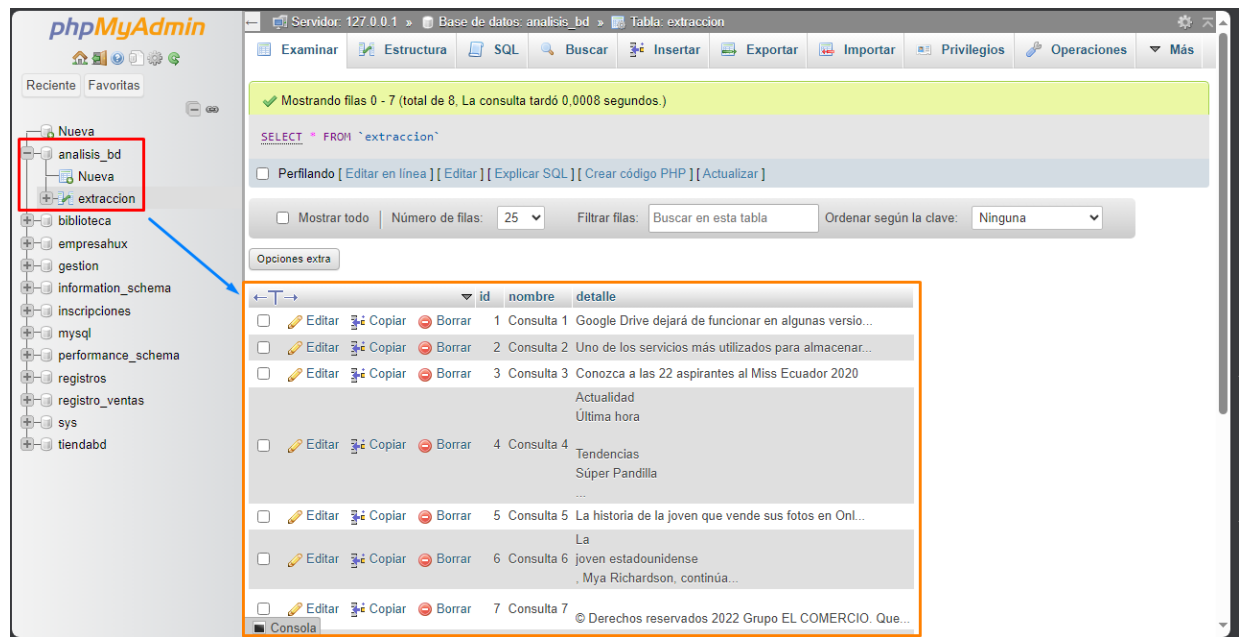
1 select * from extraccion;

Result Grid

id	nombre	detalle
1	Consulta 1	Google Drive dejará de funcionar en algunas version...
2	Consulta 2	Uno de los servicios más utilizados para almacenar es...
3	Consulta 3	Conozca o les 22 aspirantes al Miss Ecuador 2020
4	Consulta 4	Actualidad Última hora Tendencias Súper Pandilla Ed...
5	Consulta 5	La historia de la joven que vende sus fotos en OnlyF...
6	Consulta 6	La joven estadounidense, Mya Richardson, continúa...
7	Consulta 7	© Derechos reservados 2022 Grupo EL COMERCIO...
8	Consulta 8	Tendencias Redes sociales

extraccion 5

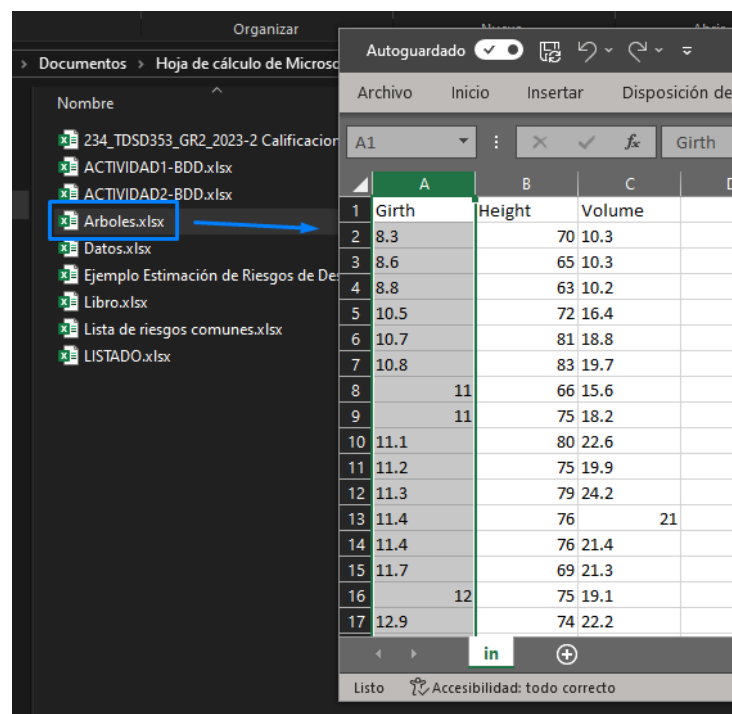
Apply Revert



4. Consulte 8 funciones para graficar y realice un ejercicio de cada una.

Para los datos se usará un set de datos perteneciente a un archivo.xlsx llamado "Arboles.xlsx" con Pandas usando `pd.read_excel()`.

Nota: En caso de que salga un error sobre "dependencia no encontrada" el comando para instalar una dependencia para abrir un archivo.xlsx en Python es "**pip install openpyxl**".

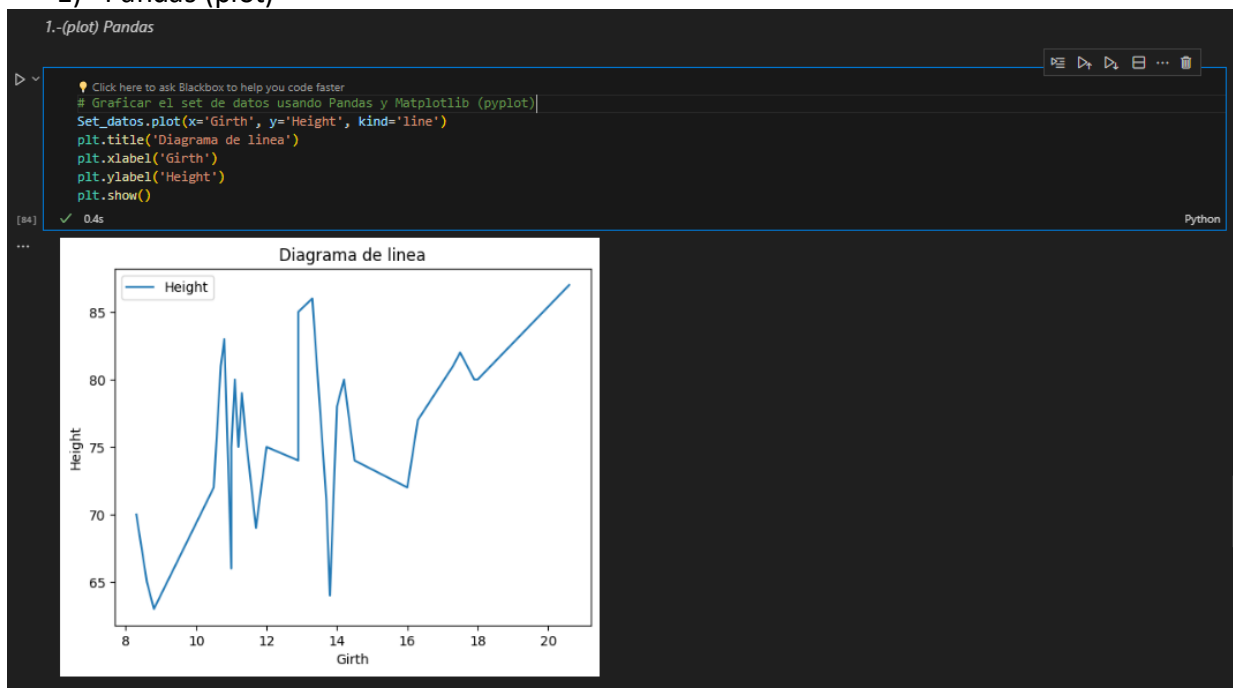



```
Click here to ask Blackbox to help you code faster
import pandas as pd
Set_datos = pd.read_excel(r"C:\Users\pinzo\OneDrive\Documentos\Arboles.xlsx", sheet_name="in")
Set_datos
```

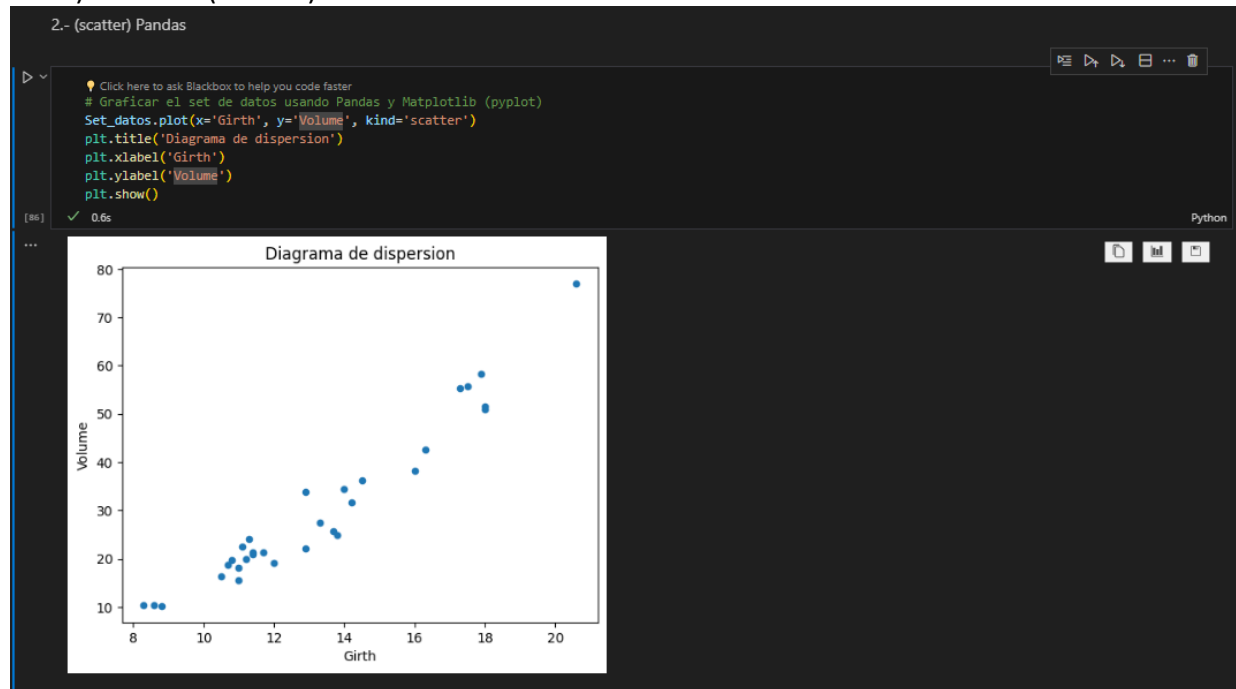
[77] ✓ 2.7s Python

	Girth	Height	Volume
0	8.3	70	10.3
1	8.6	65	10.3
2	8.8	63	10.2
3	10.5	72	16.4
4	10.7	81	18.8
5	10.8	83	19.7
6	11.0	66	15.6
7	11.0	75	18.2
8	11.1	80	22.6
9	11.2	75	19.9
10	11.3	79	24.2
11	11.4	76	21.0
12	11.4	76	21.4
13	11.7	69	21.3
14	12.0	75	19.1
15	12.9	74	22.2
16	12.9	85	33.8
17	13.3	86	27.4
18	13.7	71	25.7
19	13.8	64	24.9
20	14.0	78	34.5

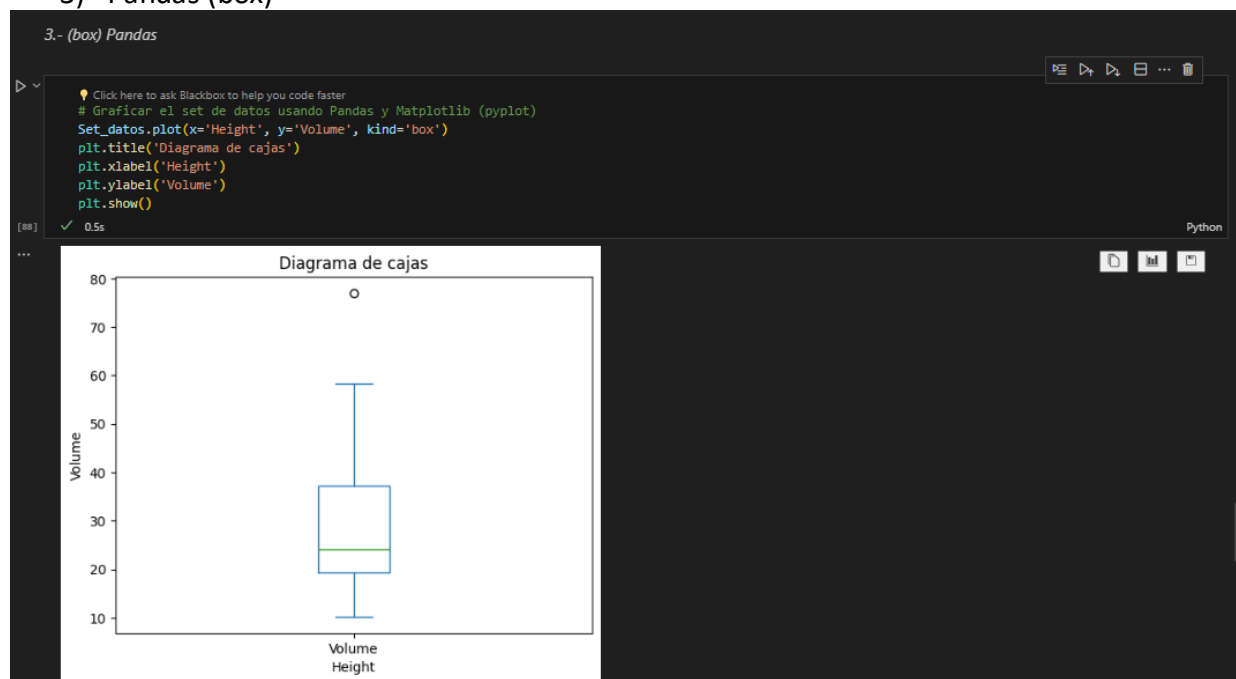
1) Pandas (plot)



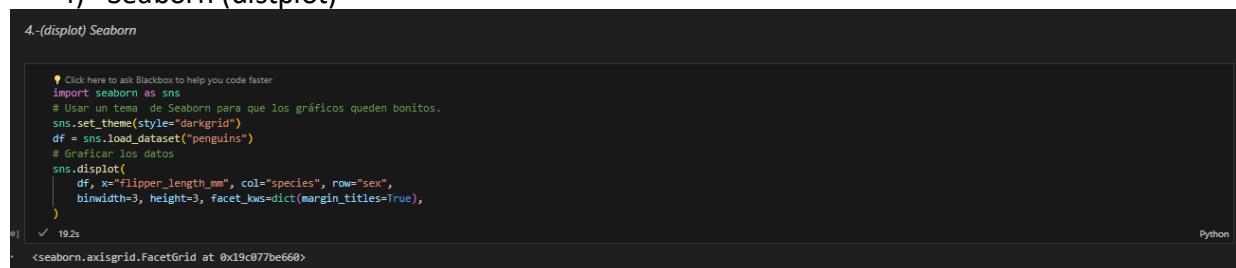
2) Pandas (scatter)

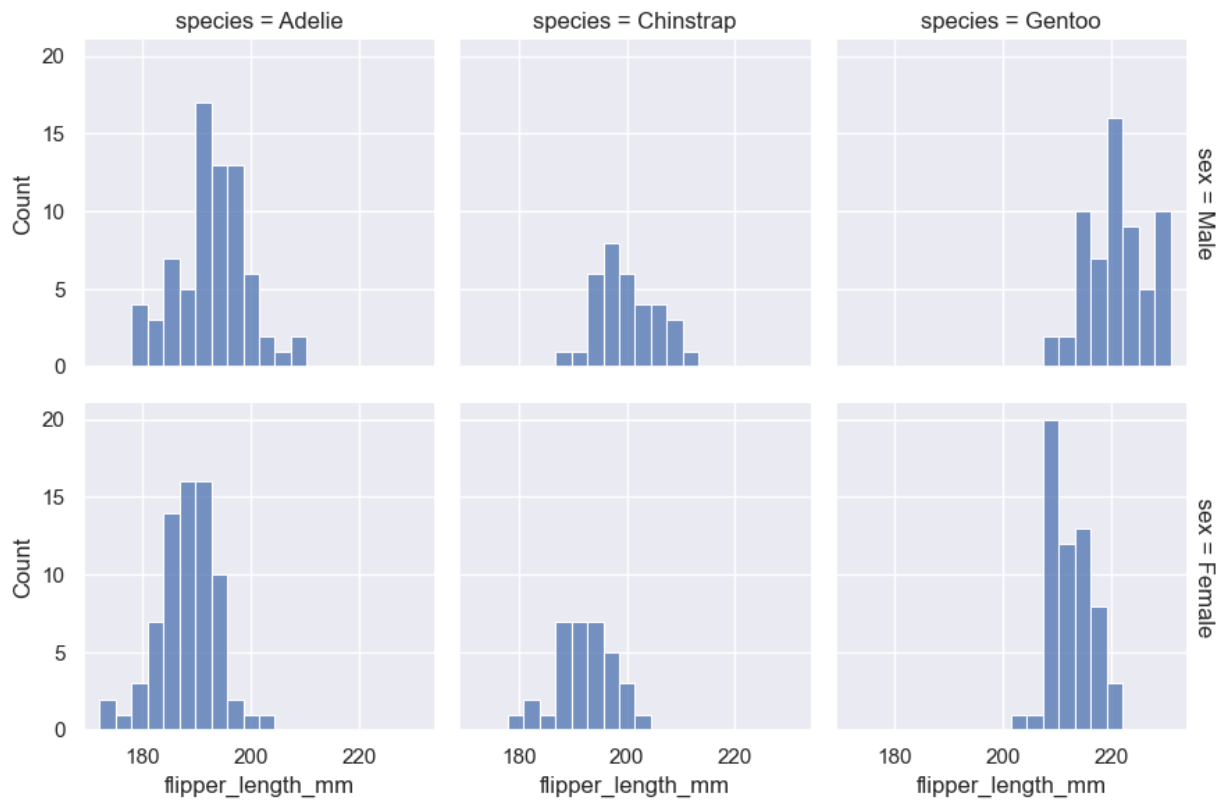


3) Pandas (box)

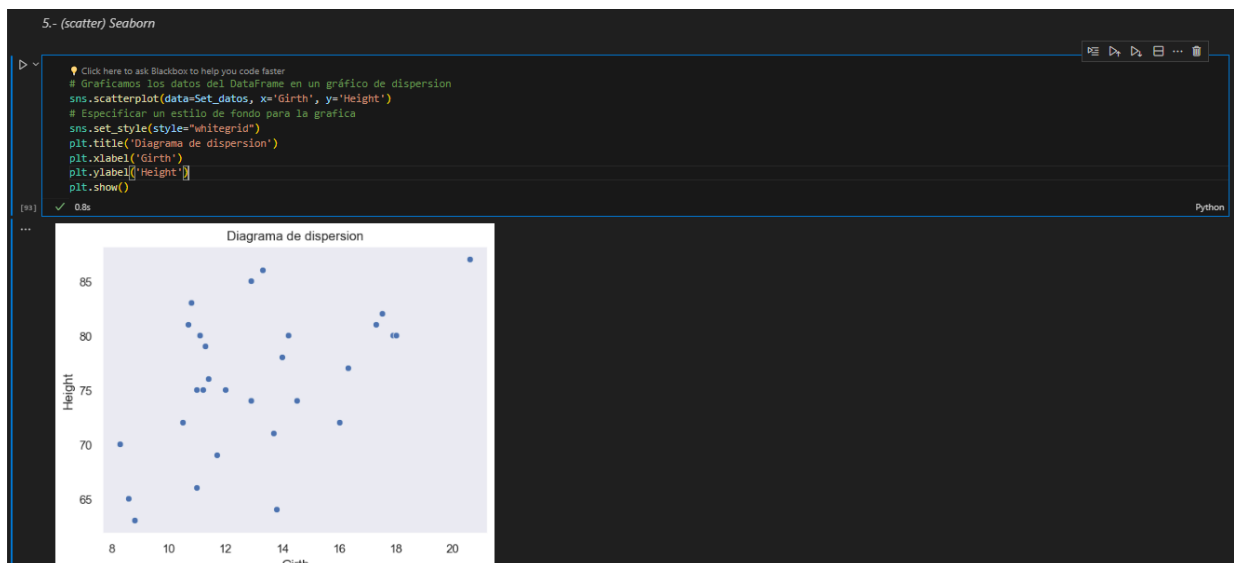


4) Seaborn (distplot)

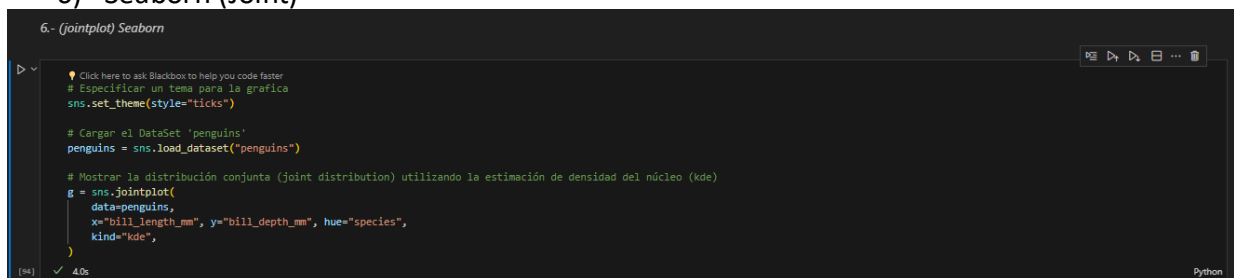


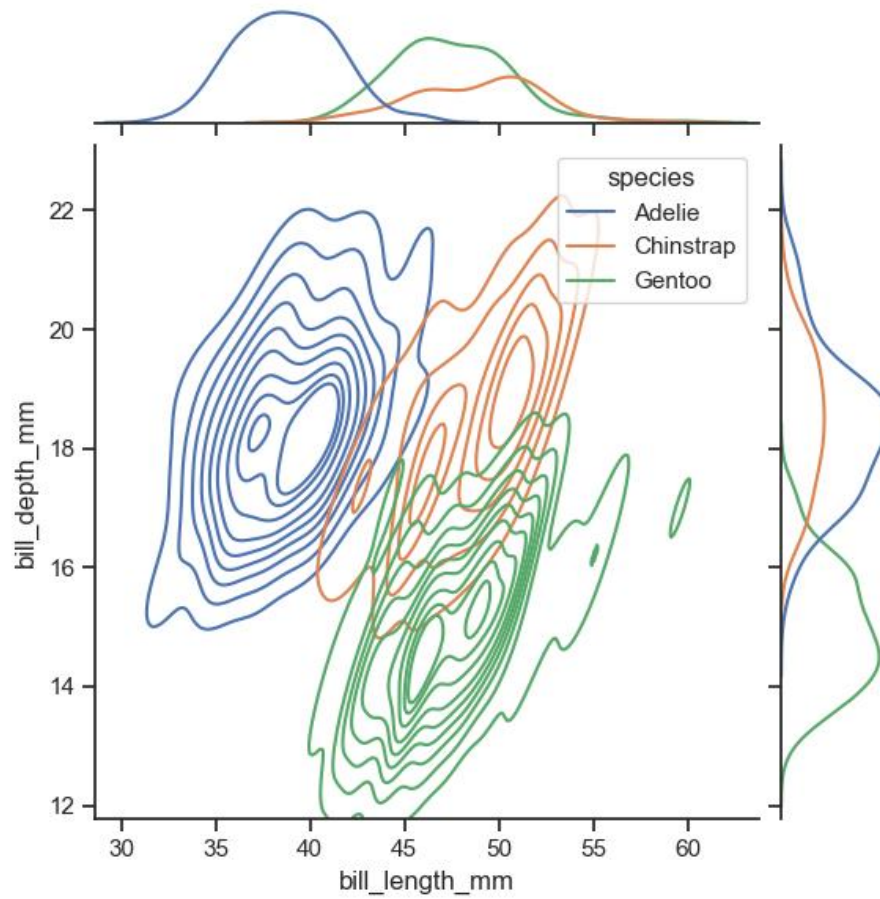


5) Seaborn (scatterplot)

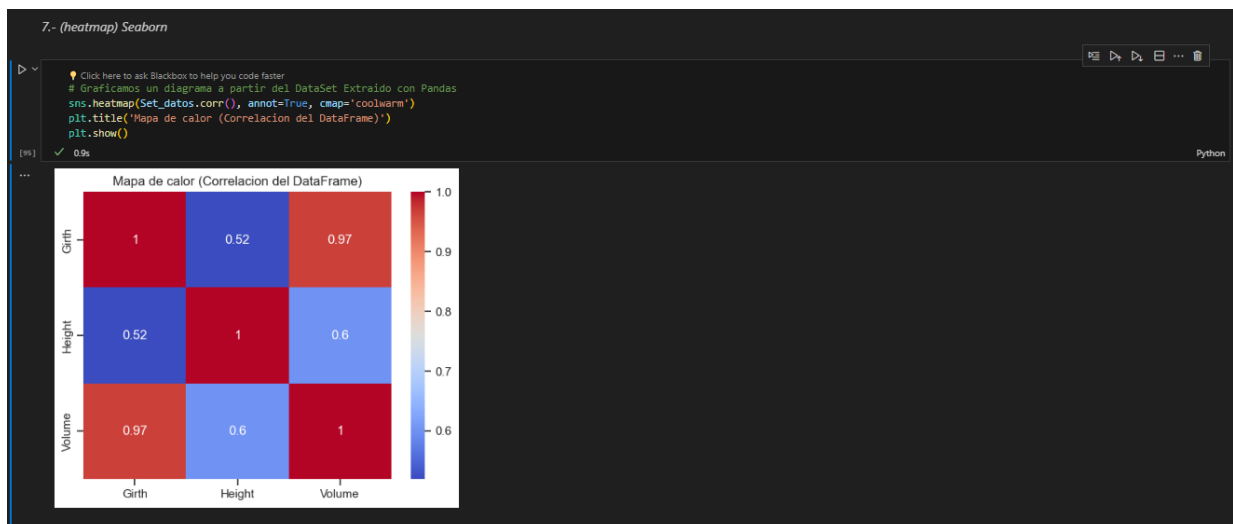


6) Seaborn (Joint)

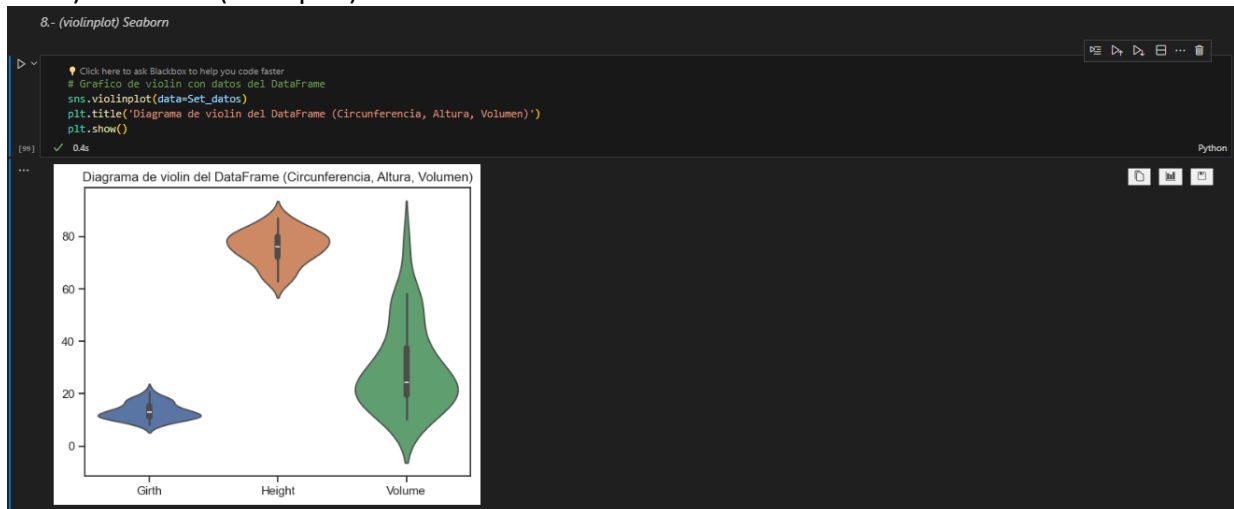




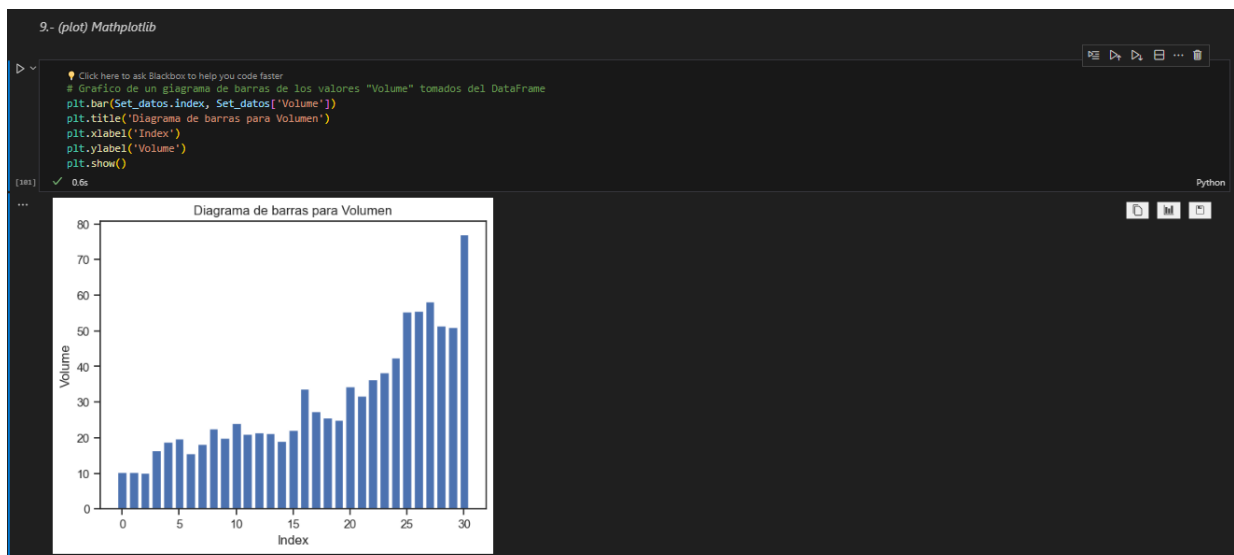
7) Seaborn (heatmap)



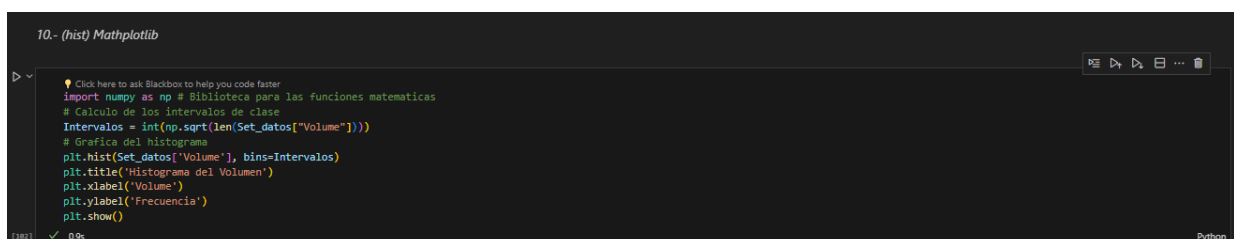
8) Seaborn (violinplot)

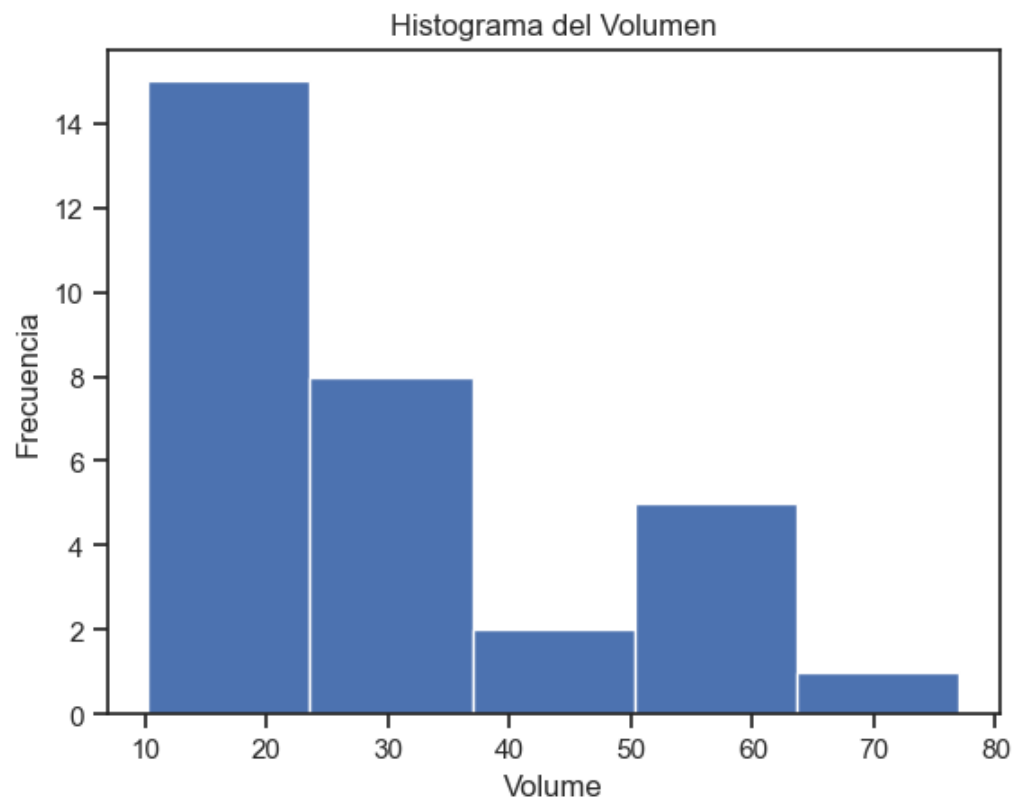


9) Matplotlib (bar)



10) Matplotlib (hist)





RECURSOS NECESARIOS

- Material de clase