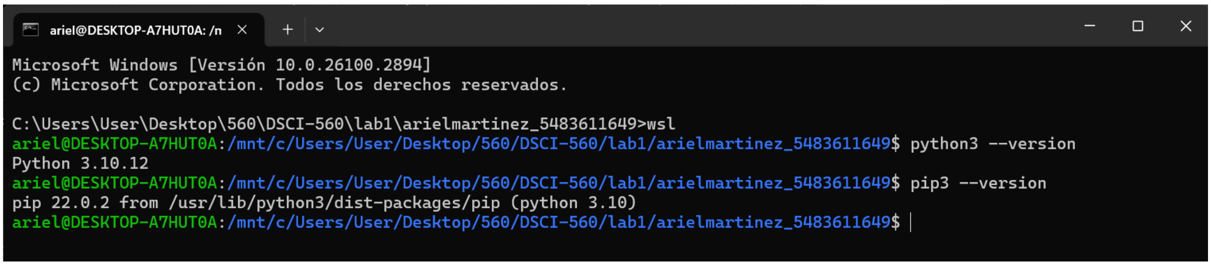
Ariel Martinez Birlanga USC_ID: 5483611649

DSCI 560 - Lab 1 Documentation

1. Installation and Setup

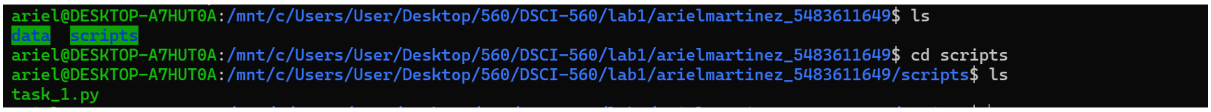
- To keep things simple, **WSL** (Windows Subsystem for Linux) was used instead of VMware for the linux environment.



```
Microsoft Windows [Versión 10.0.26100.2894]
(c) Microsoft Corporation. Todos los derechos reservados.

C:\Users\User\Desktop\560\DSCI-560\lab1\arielmartinez_5483611649>wsl
ariel@DESKTOP-A7HUT0A:/mnt/c/Users/User/Desktop/560/DSCI-560/lab1/arielmartinez_5483611649$ python3 --version
Python 3.10.12
ariel@DESKTOP-A7HUT0A:/mnt/c/Users/User/Desktop/560/DSCI-560/lab1/arielmartinez_5483611649$ pip3 --version
pip 22.0.2 from /usr/lib/python3/dist-packages/pip (python 3.10)
ariel@DESKTOP-A7HUT0A:/mnt/c/Users/User/Desktop/560/DSCI-560/lab1/arielmartinez_5483611649$
```

2.1. Playing around with Linux Terminal

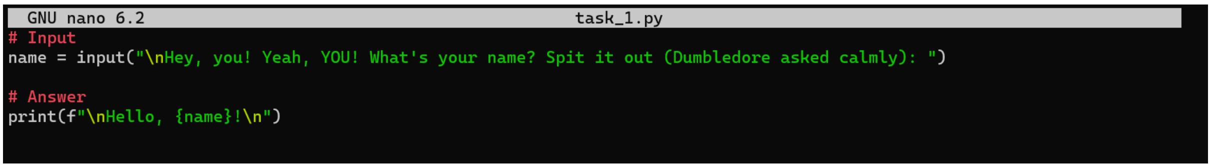
- Python3 and pip were successfully installed and the basic files and repositories were created using **touch** and **mkdir** respectively.



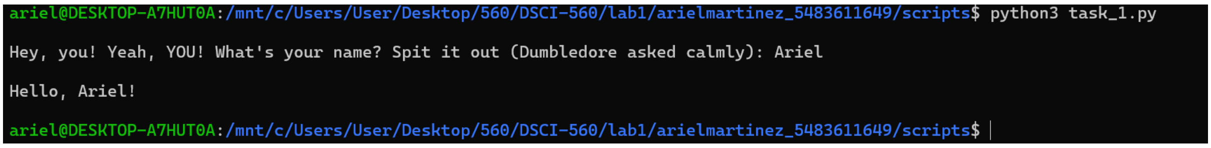
```
ariel@DESKTOP-A7HUT0A:/mnt/c/Users/User/Desktop/560/DSCI-560/lab1/arielmartinez_5483611649$ ls
data  scripts
ariel@DESKTOP-A7HUT0A:/mnt/c/Users/User/Desktop/560/DSCI-560/lab1/arielmartinez_5483611649$ cd scripts
ariel@DESKTOP-A7HUT0A:/mnt/c/Users/User/Desktop/560/DSCI-560/lab1/arielmartinez_5483611649/scripts$ ls
task_1.py
```

2.2. A basic Python Script

- The **nano** command was used to write in the first python file.
- The file asks for your name (albeit a bit aggressively) and greets you in return as intended.



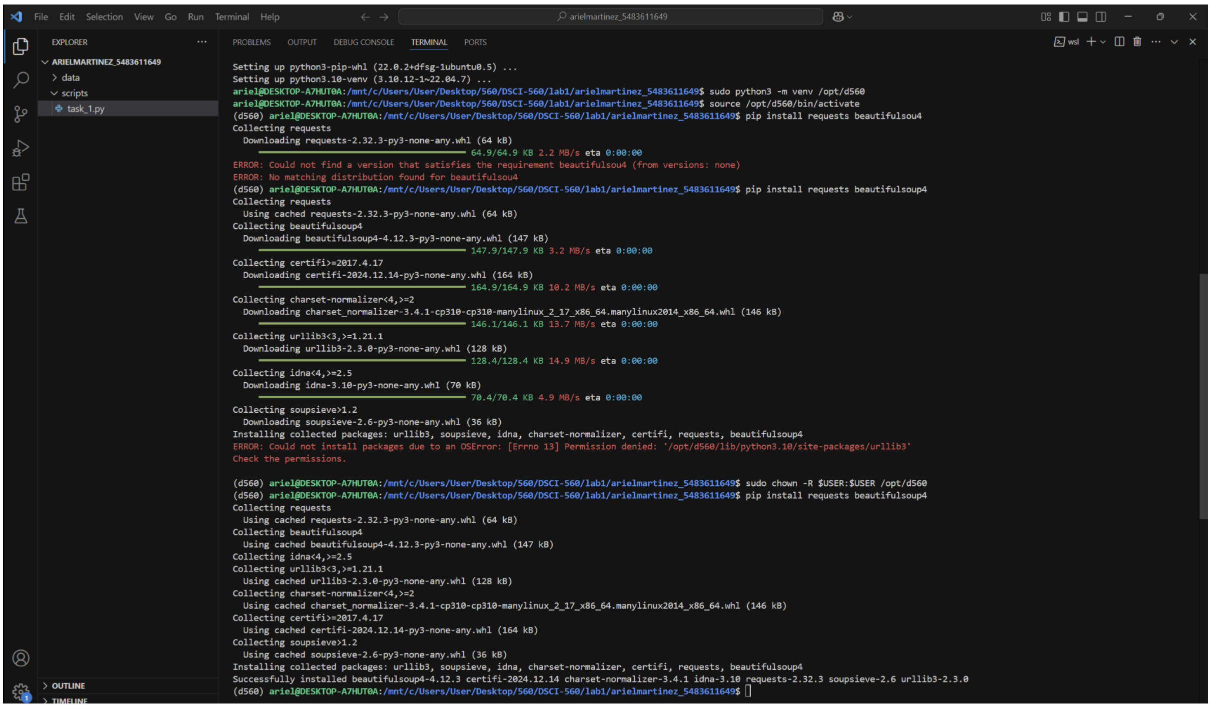
```
  GNU nano 6.2                                        task_1.py
# Input
name = input("\nHey, you! Yeah, YOU! What's your name? Spit it out (Dumbledore asked calmly): ")

# Answer
print(f"\nHello, {name}!\n")
```



```
ariel@DESKTOP-A7HUT0A:/mnt/c/Users/User/Desktop/560/DSCI-560/lab1/arielmartinez_5483611649/scripts$ python3 task_1.py

Hey, you! Yeah, YOU! What's your name? Spit it out (Dumbledore asked calmly): Ariel

Hello, Ariel!

ariel@DESKTOP-A7HUT0A:/mnt/c/Users/User/Desktop/560/DSCI-560/lab1/arielmartinez_5483611649/scripts$
```

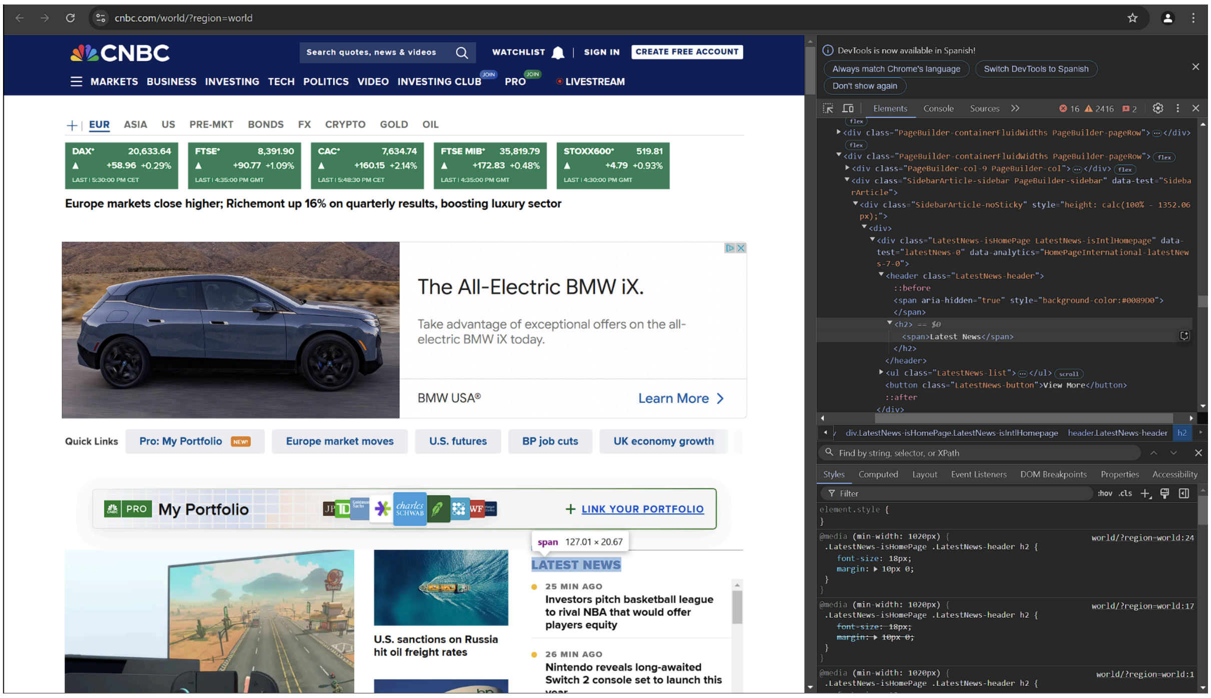
2.3. Python Web-scraping Task

- Command **venv** was used to create an exclusive environment (d560) for this subject which most likely will be used in future labs (inside WSL).
- Requests and BeautifulSoup were installed correctly after solving some minor problems with user rights.



```
Setting up python3-pip-whl (22.0.2+dfsg-1ubuntu0.5) ...
Setting up python3.10-venv (3.10.12-1~22.04.7) ...
ariel@DESKTOP-A7HUT0A:/mnt/c/Users/User/Desktop/560/DSCI-560/lab1/arielmartinez_5483611649$ sudo python3 -m venv /opt/d560
ariel@DESKTOP-A7HUT0A:/mnt/c/Users/User/Desktop/560/DSCI-560/lab1/arielmartinez_5483611649$ source /opt/d560/bin/activate
(d560) ariel@DESKTOP-A7HUT0A:/mnt/c/Users/User/Desktop/560/DSCI-560/lab1/arielmartinez_5483611649$ pip install requests beautifulsou4
Collecting requests
  Downloading requests-2.32.3-py3-none-any.whl (64 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 64.9/64.9 KB 2.2 MB/s eta 0:00:00
ERROR: Could not find a version that satisfies the requirement beautifulsou4 (from versions: none)
ERROR: No matching distribution found for beautifulsou4
(d560) ariel@DESKTOP-A7HUT0A:/mnt/c/Users/User/Desktop/560/DSCI-560/lab1/arielmartinez_5483611649$ pip install requests beautifulsoup4
Collecting requests
  Using cached requests-2.32.3-py3-none-any.whl (64 kB)
Collecting beautifulsoup4
  Downloading beautifulsoup4-4.12.3-py3-none-any.whl (147 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 147.9/147.9 KB 3.2 MB/s eta 0:00:00
Collecting certifi>=2017.4.17
  Downloading certifi-2024.12.14-py3-none-any.whl (164 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 164.9/164.9 KB 10.2 MB/s eta 0:00:00
Collecting charset-normalizer<4,>=2
  Downloading charset_normalizer-3.4.1-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (146 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 146.1/146.1 KB 13.7 MB/s eta 0:00:00
Collecting urllib3<3,>=1.21.1
  Downloading urllib3-2.3.0-py3-none-any.whl (128 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 128.4/128.4 KB 14.9 MB/s eta 0:00:00
Collecting idna<4,>=2.5
  Downloading idna-3.10-py3-none-any.whl (70 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 70.4/70.4 KB 4.9 MB/s eta 0:00:00
Collecting soupsieve>1.2
  Downloading soupsieve-2.6-py3-none-any.whl (36 kB)
Installing collected packages: urllib3, soupsieve, idna, charset-normalizer, certifi, requests, beautifulsoup4
ERROR: Could not install packages due to an OSError: [Errno 13] Permission denied: '/opt/d560/lib/python3.10/site-packages/urllib3'
Check the permissions.

(d560) ariel@DESKTOP-A7HUT0A:/mnt/c/Users/User/Desktop/560/DSCI-560/lab1/arielmartinez_5483611649$ sudo chown -R $USER:$USER /opt/d560
(d560) ariel@DESKTOP-A7HUT0A:/mnt/c/Users/User/Desktop/560/DSCI-560/lab1/arielmartinez_5483611649$ pip install requests beautifulsoup4
Collecting requests
  Using cached requests-2.32.3-py3-none-any.whl (64 kB)
Collecting beautifulsoup4
  Using cached beautifulsoup4-4.12.3-py3-none-any.whl (147 kB)
Collecting idna<4,>=2.5
Collecting urllib3<3,>=1.21.1
  Using cached urllib3-2.3.0-py3-none-any.whl (128 kB)
Collecting charset-normalizer<4,>=2
  Using cached charset_normalizer-3.4.1-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (146 kB)
Collecting certifi>=2017.4.17
  Using cached certifi-2024.12.14-py3-none-any.whl (164 kB)
Collecting soupsieve>1.2
  Using cached soupsieve-2.6-py3-none-any.whl (36 kB)
Installing collected packages: urllib3, soupsieve, idna, charset-normalizer, certifi, requests, beautifulsoup4
Successfully installed beautifulsoup4-4.12.3 certifi-2024.12.14 charset-normalizer-3.4.1 idna-3.10 requests-2.32.3 soupsieve-2.6 urllib3-2.3.0
(d560) ariel@DESKTOP-A7HUT0A:/mnt/c/Users/User/Desktop/560/DSCI-560/lab1/arielmartinez_5483611649$
```

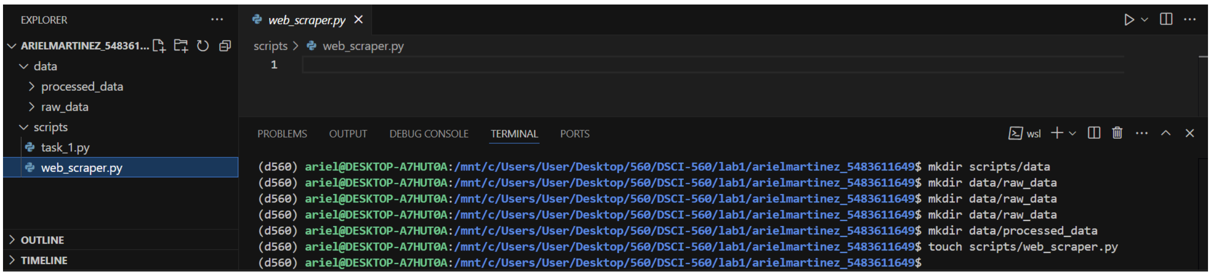
- The CNBC website (https://www.cnbc.com/world/?region=world) was analyzed to find the Market Banner (`MarketsBanner-main`) and the Latest News (`LatestNews-list`) sections.


cnbc.com/world/?region=world
CNBC
Search quotes, news & videos
WATCHLIST
SIGN IN
CREATE FREE ACCOUNT
MARKETS BUSINESS INVESTING TECH POLITICS VIDEO INVESTING CLUB JOIN PRO JOIN LIVESTREAM
EUR ASIA US PRE-MKT BONDS FX CRYPTO GOLD OIL
DAX* 20,633.64 ▲ +58.96 +0.29% LAST | 5:30:00 PM CET
FTSE* 8,391.90 ▲ +90.77 +1.09% LAST | 4:35:00 PM GMT
CAC* 7,634.74 ▲ +160.15 +2.14% LAST | 5:48:30 PM CET
FTSE MIB* 35,819.79 ▲ +172.83 +0.48% LAST | 4:35:00 PM GMT
STOXX600* 519.81 ▲ +4.79 +0.93% LAST | 4:30:00 PM GMT
Europe markets close higher; Richemont up 16% on quarterly results, boosting luxury sector
The All-Electric BMW iX.
Take advantage of exceptional offers on the all-electric BMW iX today.
BMW USA®
Learn More
Quick Links
Pro: My Portfolio NEW
Europe market moves
U.S. futures
BP job cuts
UK economy growth
PRO My Portfolio
LINK YOUR PORTFOLIO
span 127.01 × 20.67
LATEST NEWS
25 MIN AGO
Investors pitch basketball league to rival NBA that would offer players equity
U.S. sanctions on Russia hit oil freight rates
26 MIN AGO
Nintendo reveals long-awaited Switch 2 console set to launch this year
DevTools is now available in Spanish!
Always match Chrome's language
Switch DevTools to Spanish
Don't show again
Elements Console Sources
16 2416 2
div.LatestNews-isHomePage.LatestNews-isIntlHomepage header.LatestNews-header h2
Find by string, selector, or XPath
Styles Computed Layout Event Listeners DOM Breakpoints Properties Accessibility
Filter :hov .cls

```
flex
▶<div class="PageBuilder-containerFluidWidths PageBuilder-pageRow">…</div>
flex
▼<div class="PageBuilder-containerFluidWidths PageBuilder-pageRow"> flex
  ▶<div class="PageBuilder-col-9 PageBuilder-col">…</div> flex
  ▼<div class="SidebarArticle-sidebar PageBuilder-sidebar" data-test="SidebarArticle">
    ▼<div class="SidebarArticle-noSticky" style="height: calc(100% - 1352.06px);">
      ▼<div>
        ▼<div class="LatestNews-isHomePage LatestNews-isIntlHomepage" data-test="latestNews-0" data-analytics="HomePageInternational-latestNews-7-0">
          ▼<header class="LatestNews-header">
              ::before
              <span aria-hidden="true" style="background-color:#0089D0">
              </span>
            ▼<h2> == $0
                <span>Latest News</span>
              </h2>
            </header>
          ▶<ul class="LatestNews-list">…</ul> scroll
            <button class="LatestNews-button">View More</button>
            ::after
          </div>
```

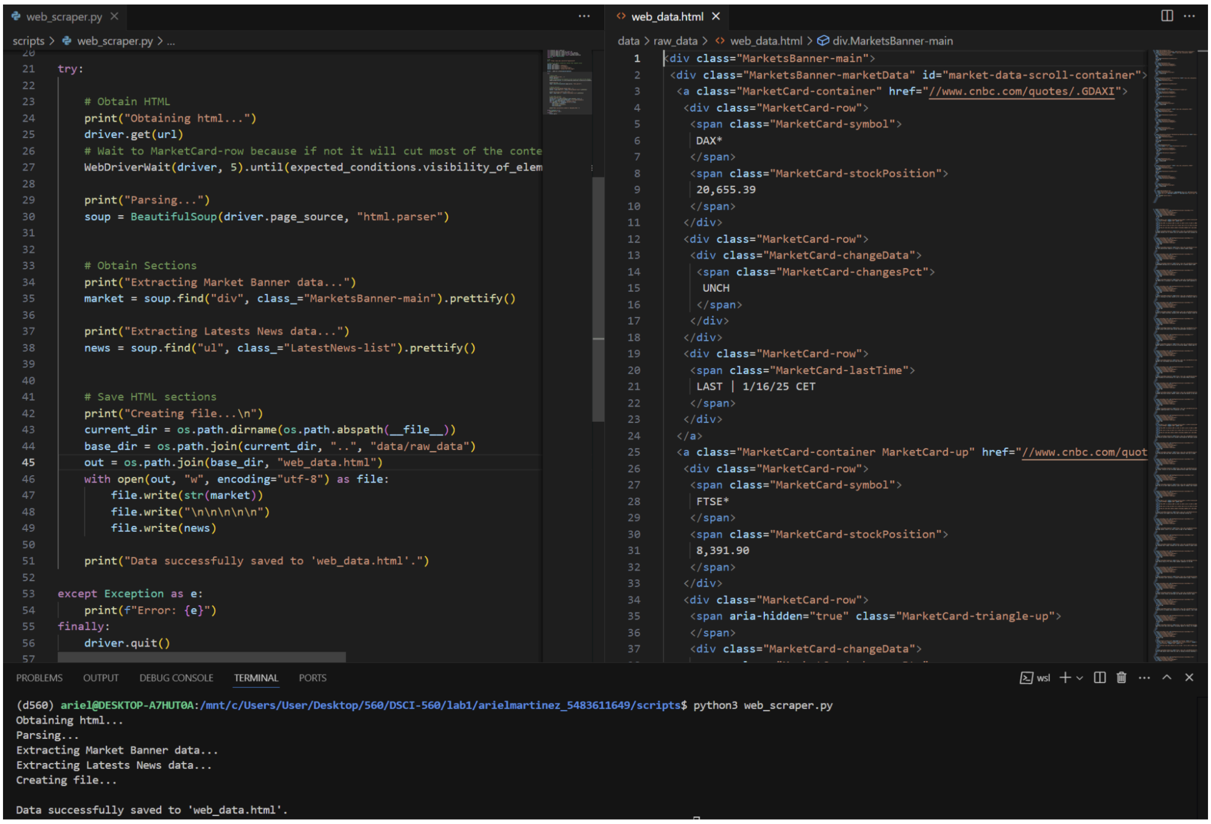
```
element.style {
}
@media (min-width: 1020px) {                                world/?region=world:24
 .LatestNews-isHomePage .LatestNews-header h2 {
    font-size: 18px;
    margin: ▶ 10px 0;
 }
}
@media (min-width: 1020px) {                                world/?region=world:17
 .LatestNews-isHomePage .LatestNews-header h2 {
    ~~font-size: 18px;~~
    ~~margin: ▶ 10px 0;~~
 }
}
@media (min-width: 1020px) {                                world/?region=world:1
 .LatestNews-isHomePage .LatestNews-header h2 {
```

- The new folders and files were successfully created like the first time, just that this time from the VScode terminal.



```
(d560) ariel@DESKTOP-A7HUT0A:/mnt/c/Users/User/Desktop/560/DSCI-560/lab1/arielmartinez_5483611649$ mkdir scripts/data
(d560) ariel@DESKTOP-A7HUT0A:/mnt/c/Users/User/Desktop/560/DSCI-560/lab1/arielmartinez_5483611649$ mkdir data/raw_data
(d560) ariel@DESKTOP-A7HUT0A:/mnt/c/Users/User/Desktop/560/DSCI-560/lab1/arielmartinez_5483611649$ mkdir data/raw_data
(d560) ariel@DESKTOP-A7HUT0A:/mnt/c/Users/User/Desktop/560/DSCI-560/lab1/arielmartinez_5483611649$ mkdir data/raw_data
(d560) ariel@DESKTOP-A7HUT0A:/mnt/c/Users/User/Desktop/560/DSCI-560/lab1/arielmartinez_5483611649$ mkdir data/processed_data
(d560) ariel@DESKTOP-A7HUT0A:/mnt/c/Users/User/Desktop/560/DSCI-560/lab1/arielmartinez_5483611649$ touch scripts/web_scraper.py
(d560) ariel@DESKTOP-A7HUT0A:/mnt/c/Users/User/Desktop/560/DSCI-560/lab1/arielmartinez_5483611649$
```

- web_scraper.py successfully extracted the desired parts of the html.
- **Selenium** was used instead of **request** due to the fact that request wasn't able to extract Markets Banner data correctly due to how its data was integrated. **WebDriverWait** in particular was needed to avoid this problem by waiting to receive information on Markets Banner Row sections.
- **BeautifulSoup** was used to parse and filter the html sections and the **os** library was used to obtain the correct file path (while maintaining replicability) for the output (web_data.html).
- Prints were used to mark every section of the code.


web_scraper.py
scripts > web_scraper.py > ...
web_data.html
data > raw_data > web_data.html > div.MarketsBanner-main
PROBLEMS
OUTPUT
DEBUG CONSOLE
TERMINAL
PORTS
wsl

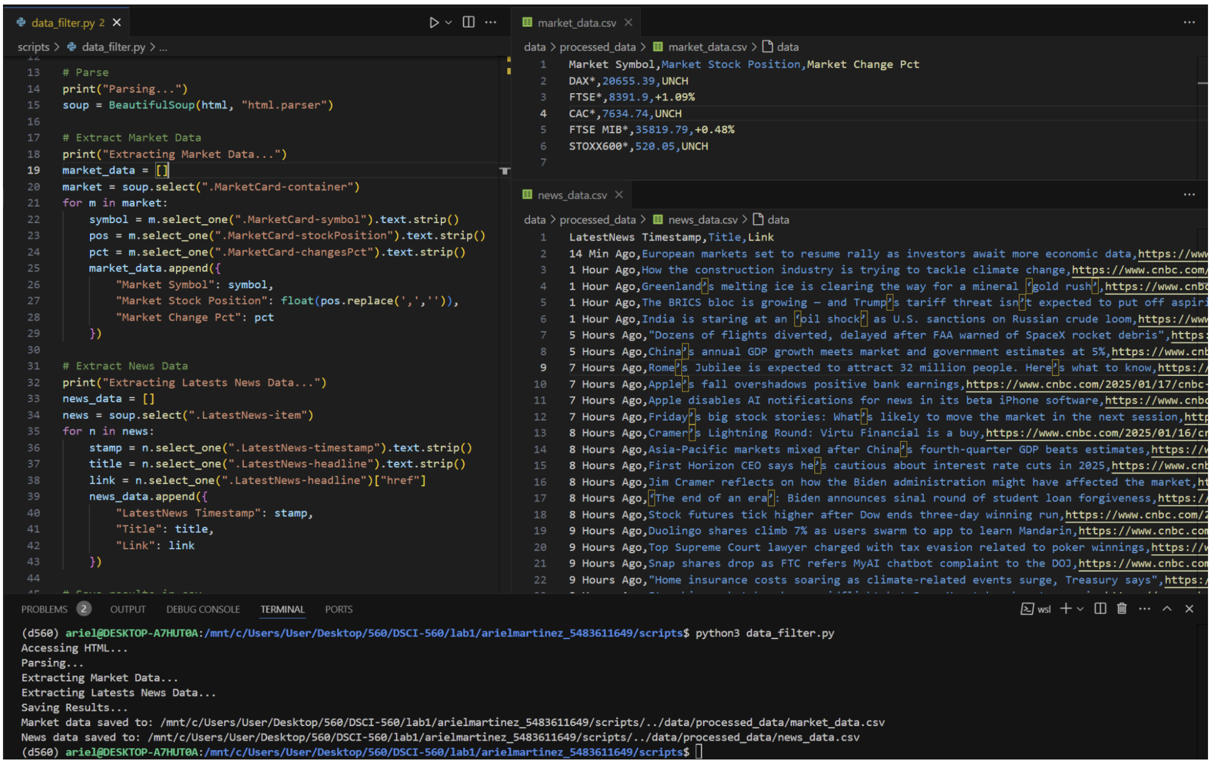
```python
try:

    # Obtain HTML
    print("Obtaining html...")
    driver.get(url)
    # Wait to MarketCard-row because if not it will cut most of the conte
    WebDriverWait(driver, 5).until(expected_conditions.visibility_of_elem

    print("Parsing...")
    soup = BeautifulSoup(driver.page_source, "html.parser")


    # Obtain Sections
    print("Extracting Market Banner data...")
    market = soup.find("div", class_="MarketsBanner-main").prettify()

    print("Extracting Latests News data...")
    news = soup.find("ul", class_="LatestNews-list").prettify()


    # Save HTML sections
    print("Creating file...\n")
    current_dir = os.path.dirname(os.path.abspath(__file__))
    base_dir = os.path.join(current_dir, "..", "data/raw_data")
    out = os.path.join(base_dir, "web_data.html")
    with open(out, "w", encoding="utf-8") as file:
        file.write(str(market))
        file.write("\n\n\n\n\n")
        file.write(news)

    print("Data successfully saved to 'web_data.html'.")

except Exception as e:
    print(f"Error: {e}")
finally:
    driver.quit()
```

```html
<div class="MarketsBanner-main">
 <div class="MarketsBanner-marketData" id="market-data-scroll-container">
  <a class="MarketCard-container" href="//www.cnbc.com/quotes/.GDAXI">
   <div class="MarketCard-row">
    <span class="MarketCard-symbol">
     DAX*
    </span>
    <span class="MarketCard-stockPosition">
     20,655.39
    </span>
   </div>
   <div class="MarketCard-row">
    <div class="MarketCard-changeData">
     <span class="MarketCard-changesPct">
      UNCH
     </span>
    </div>
   </div>
   <div class="MarketCard-row">
    <span class="MarketCard-lastTime">
     LAST | 1/16/25 CET
    </span>
   </div>
  </a>
  <a class="MarketCard-container MarketCard-up" href="//www.cnbc.com/quot
   <div class="MarketCard-row">
    <span class="MarketCard-symbol">
     FTSE*
    </span>
    <span class="MarketCard-stockPosition">
     8,391.90
    </span>
   </div>
   <div class="MarketCard-row">
    <span aria-hidden="true" class="MarketCard-triangle-up">
    </span>
    <div class="MarketCard-changeData">
```

```
(d560) ariel@DESKTOP-A7HUT0A:/mnt/c/Users/User/Desktop/560/DSCI-560/lab1/arielmartinez_5483611649/scripts$ python3 web_scraper.py
Obtaining html...
Parsing...
Extracting Market Banner data...
Extracting Latests News data...
Creating file...

Data successfully saved to 'web_data.html'.
```

2.4. Data Filtering Task

- Script data_filter.py was successful in creating a structured csv for both sections of web_data.html.
- Similarly to the task before, **BeautifulSoup** was used to parse the html and **os** was used to correctly manage the input and output files in a replicable way.
- The Market data extraction was straightforward since it had specific classes for the 3 desired columns, but in the case of the Latest News section, the **link** was found in the href (as usual in a html) of the headline (title) class.
- Prints were used to mark every section of the code.


data_filter.py 2
scripts > data_filter.py > ...
market_data.csv
data > processed_data > market_data.csv > data
news_data.csv
data > processed_data > news_data.csv > data
PROBLEMS 2
OUTPUT
DEBUG CONSOLE
TERMINAL
PORTS
wsl

```python
# Parse
print("Parsing...")
soup = BeautifulSoup(html, "html.parser")

# Extract Market Data
print("Extracting Market Data...")
market_data = []
market = soup.select(".MarketCard-container")
for m in market:
    symbol = m.select_one(".MarketCard-symbol").text.strip()
    pos = m.select_one(".MarketCard-stockPosition").text.strip()
    pct = m.select_one(".MarketCard-changesPct").text.strip()
    market_data.append({
        "Market Symbol": symbol,
        "Market Stock Position": float(pos.replace(',','')),
        "Market Change Pct": pct
    })

# Extract News Data
print("Extracting Latests News Data...")
news_data = []
news = soup.select(".LatestNews-item")
for n in news:
    stamp = n.select_one(".LatestNews-timestamp").text.strip()
    title = n.select_one(".LatestNews-headline").text.strip()
    link = n.select_one(".LatestNews-headline")["href"]
    news_data.append({
        "LatestNews Timestamp": stamp,
        "Title": title,
        "Link": link
    })

```

```
Market Symbol,Market Stock Position,Market Change Pct
DAX*,20655.39,UNCH
FTSE*,8391.9,+1.09%
CAC*,7634.74,UNCH
FTSE MIB*,35819.79,+0.48%
STOXX600*,520.05,UNCH

```

```
LatestNews Timestamp,Title,Link
14 Min Ago,European markets set to resume rally as investors await more economic data,https://www
1 Hour Ago,How the construction industry is trying to tackle climate change,https://www.cnbc.com/
1 Hour Ago,Greenland’s melting ice is clearing the way for a mineral ‘gold rush’,https://www.cnbc
1 Hour Ago,The BRICS bloc is growing — and Trump’s tariff threat isn’t expected to put off aspiri
1 Hour Ago,India is staring at an ‘oil shock’ as U.S. sanctions on Russian crude loom,https://www
5 Hours Ago,"Dozens of flights diverted, delayed after FAA warned of SpaceX rocket debris",https:
5 Hours Ago,China’s annual GDP growth meets market and government estimates at 5%,https://www.cnb
7 Hours Ago,Rome’s Jubilee is expected to attract 32 million people. Here’s what to know,https://
7 Hours Ago,Apple’s fall overshadows positive bank earnings,https://www.cnbc.com/2025/01/17/cnbc-
7 Hours Ago,Apple disables AI notifications for news in its beta iPhone software,https://www.cnbc
7 Hours Ago,Friday’s big stock stories: What’s likely to move the market in the next session,http
8 Hours Ago,Cramer’s Lightning Round: Virtu Financial is a buy,https://www.cnbc.com/2025/01/16/cr
8 Hours Ago,Asia-Pacific markets mixed after China’s fourth-quarter GDP beats estimates,https://w
8 Hours Ago,First Horizon CEO says he’s cautious about interest rate cuts in 2025,https://www.cnb
8 Hours Ago,Jim Cramer reflects on how the Biden administration might have affected the market,ht
8 Hours Ago,‘The end of an era’: Biden announces sinal round of student loan forgiveness,https://
8 Hours Ago,Stock futures tick higher after Dow ends three-day winning run,https://www.cnbc.com/2
9 Hours Ago,Duolingo shares climb 7% as users swarm to app to learn Mandarin,https://www.cnbc.com
9 Hours Ago,Top Supreme Court lawyer charged with tax evasion related to poker winnings,https://w
9 Hours Ago,Snap shares drop as FTC refers MyAI chatbot complaint to the DOJ,https://www.cnbc.com
9 Hours Ago,"Home insurance costs soaring as climate-related events surge, Treasury says",https:/
```

```
(d560) ariel@DESKTOP-A7HUT0A:/mnt/c/Users/User/Desktop/560/DSCI-560/lab1/arielmartinez_5483611649/scripts$ python3 data_filter.py
Accessing HTML...
Parsing...
Extracting Market Data...
Extracting Latests News Data...
Saving Results...
Market data saved to: /mnt/c/Users/User/Desktop/560/DSCI-560/lab1/arielmartinez_5483611649/scripts/../data/processed_data/market_data.csv
News data saved to: /mnt/c/Users/User/Desktop/560/DSCI-560/lab1/arielmartinez_5483611649/scripts/../data/processed_data/news_data.csv
(d560) ariel@DESKTOP-A7HUT0A:/mnt/c/Users/User/Desktop/560/DSCI-560/lab1/arielmartinez_5483611649/scripts$
```