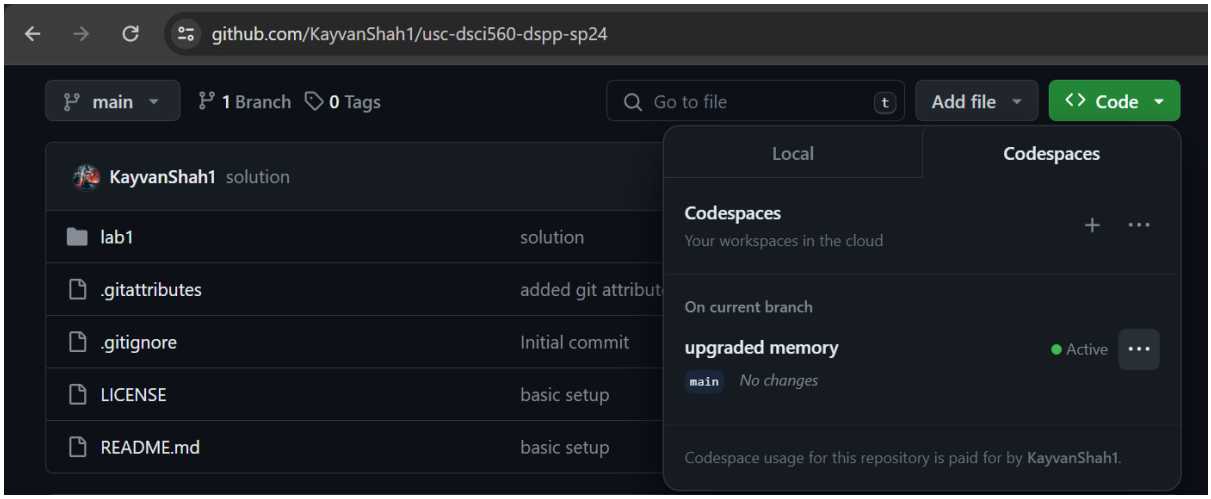
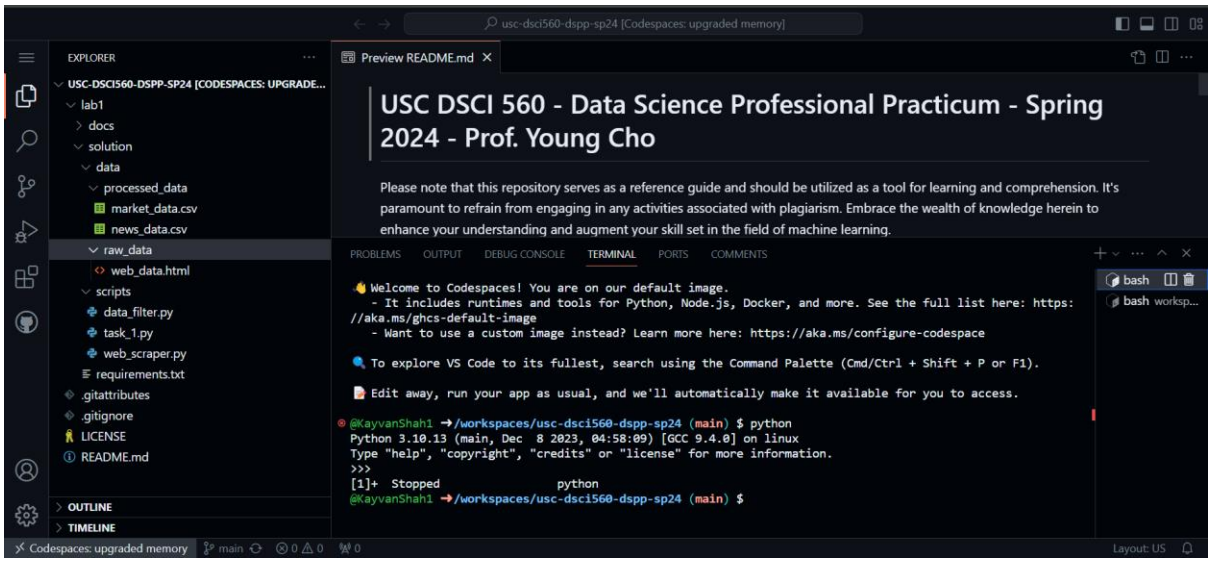
Lab 1 Solution

Installation & Setup

- Development done locally in a local git repository on a windows machine.
- The code is synced with a remote git repository.

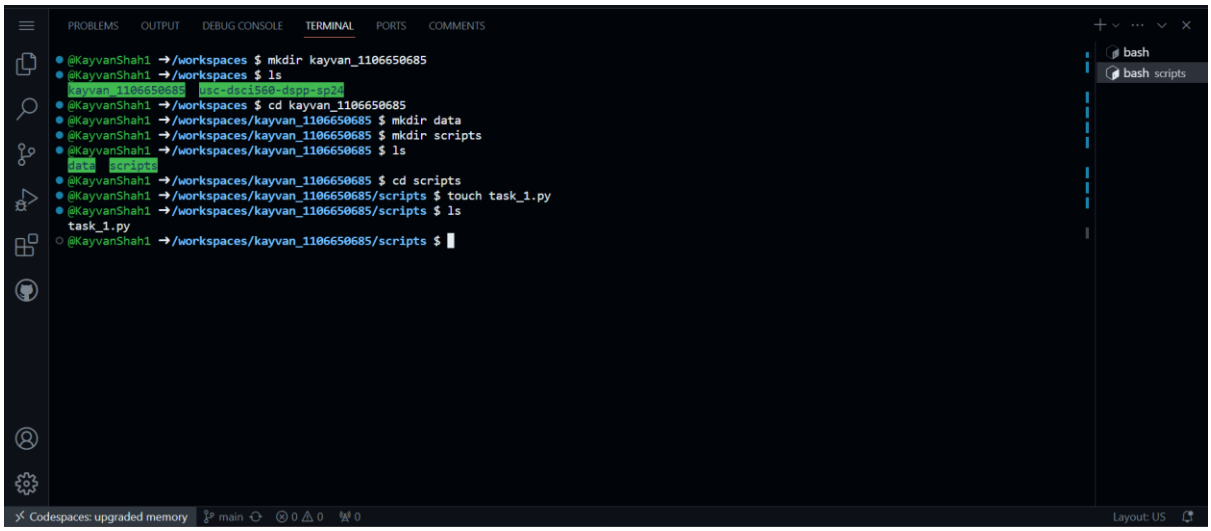


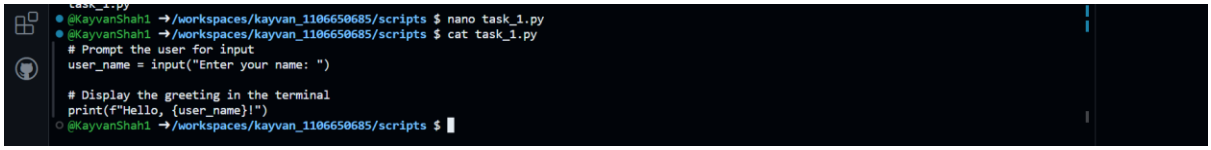
- A Linux Ubuntu instance is spin up using GitHub Codespace for the repository.



Get Familiar with Linux and Python

Playing around with Linux Terminal





A basic Python Script

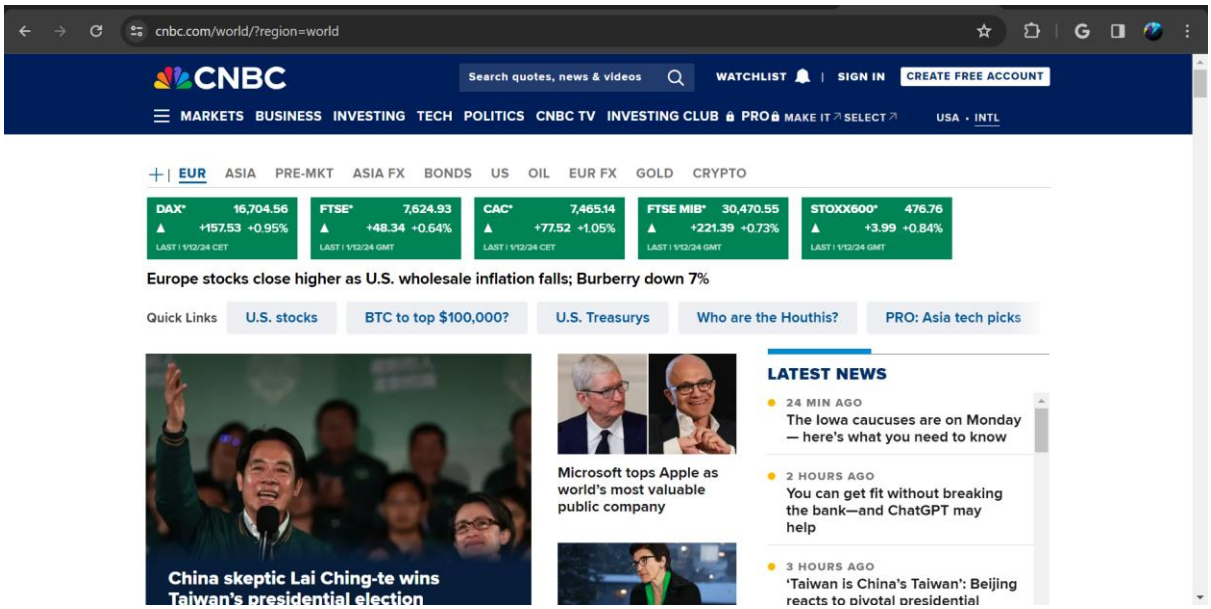
Web Scraping Task

- Create a virtual environment and install the dependencies using the ***requirements.txt***.
  - *venv* – Created a virtual environment using command $python - m\ venv\ venv$
  - Linux has a built-in support for Chrome and using a library **chromedriver-py** we automatically manage the webdriver downloads and retrieval of the executable path while initializing the web driver.

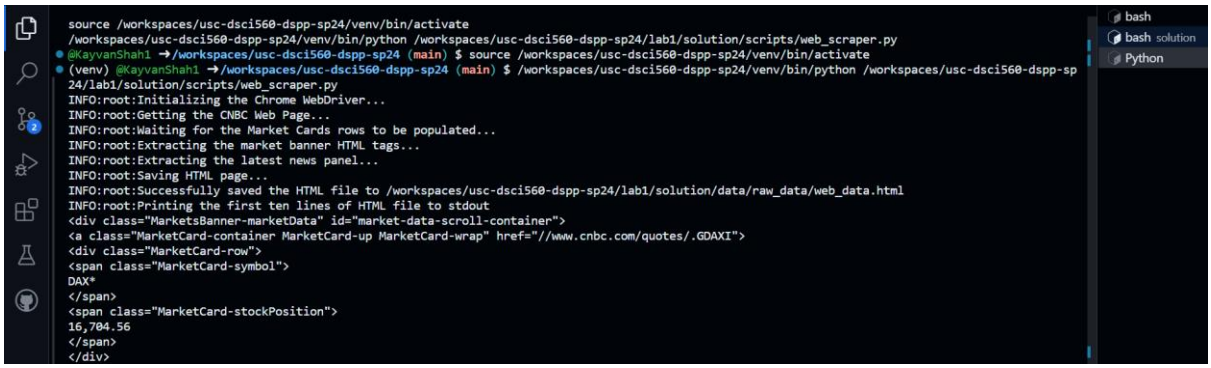
```
Installing collected packages: sortedcontainers, pytz, chromedriver-py, urllib3, tzdata, typing-extensions, soupsieve, sn
iffio, six, python-dotenv, pysocks, numpy, lxml, idna, h11, exceptiongroup, charset-normalizer, certifi, attrs, annotated
-types, wsproto, requests, python-dateutil, pydantic-core, outcome, beautifulsoup4, trio, pydantic, pandas, trio-websocke
t, pydantic-settings, selenium
Successfully installed annotated-types-0.6.0 attrs-23.2.0 beautifulsoup4-4.12.2 certifi-2023.11.17 charset-normalizer-3.3
.2 chromedriver-py-120.0.6099.109 exceptiongroup-1.2.0 h11-0.14.0 idna-3.6 lxml-5.1.0 numpy-1.26.3 outcome-1.3.0.post0 pa
ndas-2.1.4 pydantic-2.5.3 pydantic-core-2.14.6 pydantic-settings-2.1.0 pysocks-1.7.1 python-dateutil-2.8.2 python-dotenv-
1.0.0 pytz-2023.3.post1 requests-2.31.0 selenium-4.16.0 six-1.16.0 sniffio-1.3.0 sortedcontainers-2.4.0 soupsieve-2.5 tri
o-0.24.0 trio-websocket-0.11.1 typing-extensions-4.9.0 tzdata-2023.4 urllib3-2.1.0 wsproto-1.2.0

[notice] A new release of pip is available: 23.0.1 -> 23.3.2
[notice] To update, run: pip install --upgrade pip
(venv) @KayvanShah1 ➜/workspaces/usc-dsci560-dspp-sp24/lab1/solution (main) $ pip install -U pip
Requirement already satisfied: pip in /workspaces/usc-dsci560-dspp-sp24/venv/lib/python3.10/site-packages (23.0.1)
Collecting pip
  Downloading pip-23.3.2-py3-none-any.whl (2.1 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 2.1/2.1 MB 6.3 MB/s eta 0:00:00
Installing collected packages: pip
  Attempting uninstall: pip
    Found existing installation: pip 23.0.1
    Uninstalling pip-23.0.1:
      Successfully uninstalled pip-23.0.1
Successfully installed pip-23.3.2
```

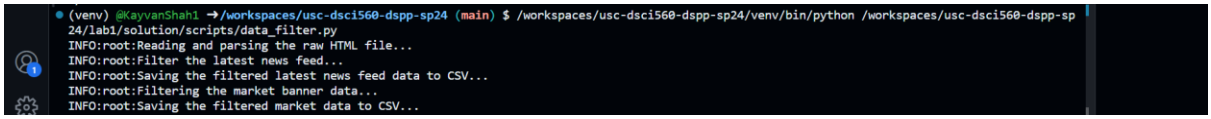
- A snapshot of the webpage from where the data is to be scraped.
    - The webpage loads a few fields dynamically. In this case the market banner is loaded using a JavaScript.
    - To capture the data from dynamically populated HTML elements we must use selenium with a web driver suited for the installed browser on the system.
    - While the latest news isn't loaded dynamically hence, we can get the static snapshot of the HTML tree of the scrape that its section's data.



- Scraping the raw data

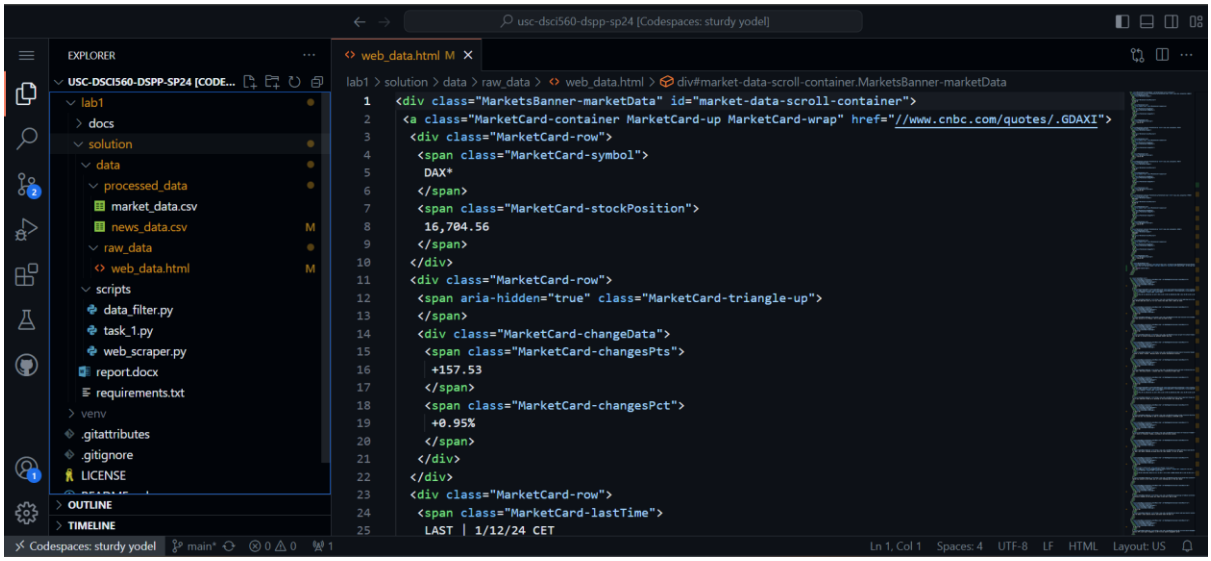


- Filtering the data to extract the fields of interest.

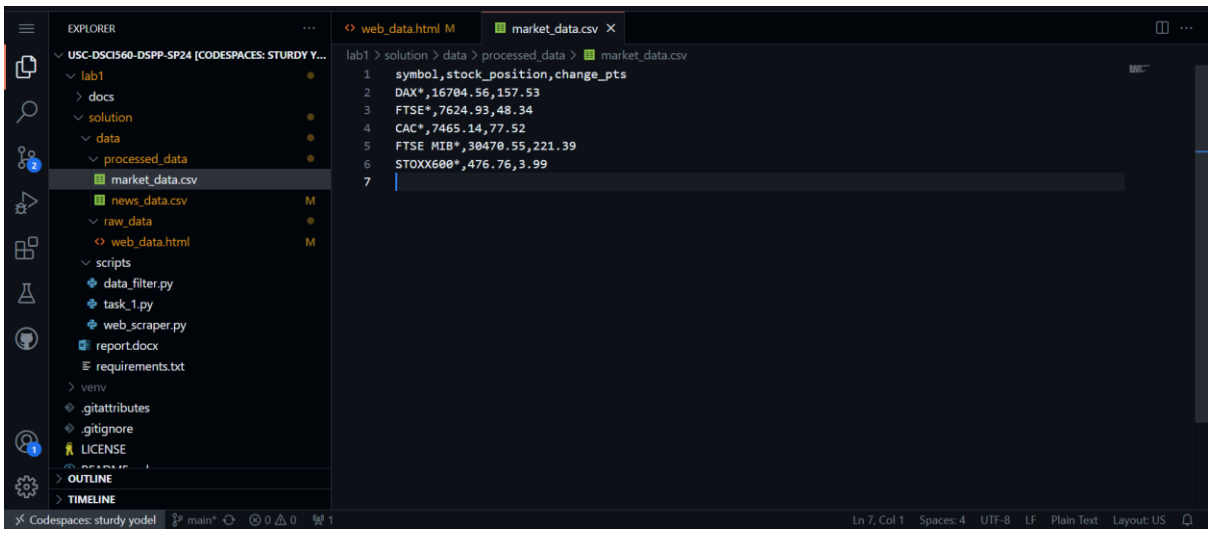


Output Files

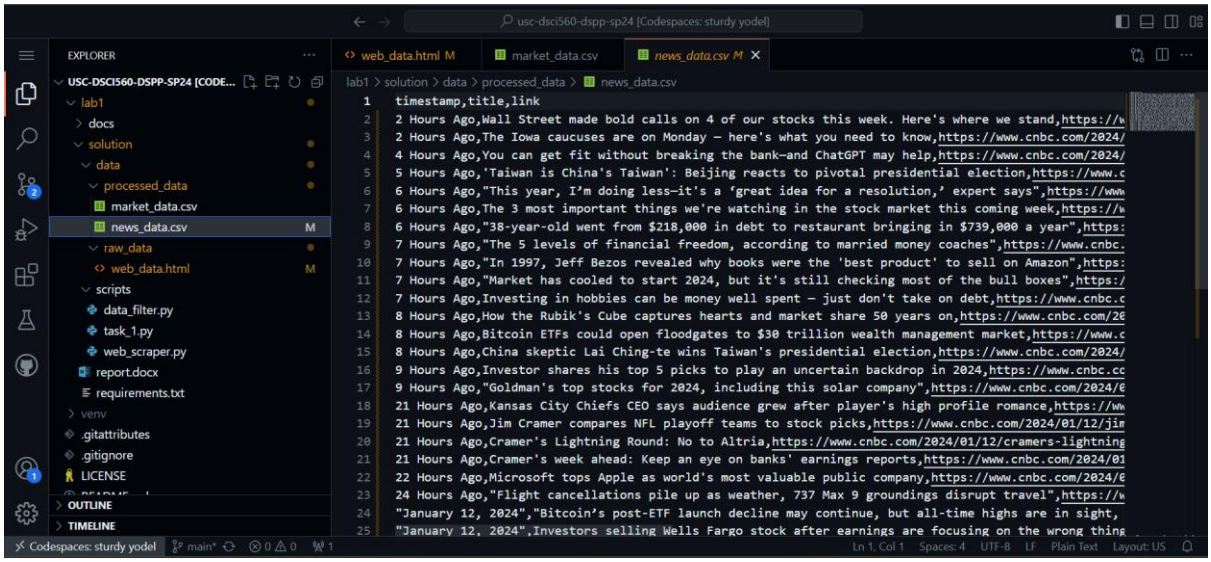
- Raw data output HTML file



- Processed data CSV files
    - Market data



    - Latest news


usc-dsci560-dspp-sp24 [Codespaces: sturdy yodel]
EXPLORER
web_data.html M
market_data.csv
news_data.csv M
lab1 > solution > data > processed_data > news_data.csv
timestamp,title,link
2 Hours Ago,Wall Street made bold calls on 4 of our stocks this week. Here's where we stand,https://w
2 Hours Ago,The Iowa caucuses are on Monday — here's what you need to know,https://www.cnbc.com/2024/
4 Hours Ago,You can get fit without breaking the bank—and ChatGPT may help,https://www.cnbc.com/2024/
5 Hours Ago,'Taiwan is China's Taiwan': Beijing reacts to pivotal presidential election,https://www.c
6 Hours Ago,"This year, I'm doing less—it's a 'great idea for a resolution,' expert says",https://www
6 Hours Ago,The 3 most important things we're watching in the stock market this coming week,https://w
6 Hours Ago,"38-year-old went from $218,000 in debt to restaurant bringing in $739,000 a year",https:
7 Hours Ago,"The 5 levels of financial freedom, according to married money coaches",https://www.cnbc.
7 Hours Ago,"In 1997, Jeff Bezos revealed why books were the 'best product' to sell on Amazon",https:
7 Hours Ago,"Market has cooled to start 2024, but it's still checking most of the bull boxes",https:/
7 Hours Ago,Investing in hobbies can be money well spent — just don't take on debt,https://www.cnbc.c
8 Hours Ago,How the Rubik's Cube captures hearts and market share 50 years on,https://www.cnbc.com/20
8 Hours Ago,Bitcoin ETFs could open floodgates to $30 trillion wealth management market,https://www.c
8 Hours Ago,China skeptic Lai Ching-te wins Taiwan's presidential election,https://www.cnbc.com/2024/
9 Hours Ago,Investor shares his top 5 picks to play an uncertain backdrop in 2024,https://www.cnbc.cc
9 Hours Ago,"Goldman's top stocks for 2024, including this solar company",https://www.cnbc.com/2024/0
21 Hours Ago,Kansas City Chiefs CEO says audience grew after player's high profile romance,https://ww
21 Hours Ago,Jim Cramer compares NFL playoff teams to stock picks,https://www.cnbc.com/2024/01/12/jim
21 Hours Ago,Cramer's Lightning Round: No to Altria,https://www.cnbc.com/2024/01/12/cramers-lightning
21 Hours Ago,Cramer's week ahead: Keep an eye on banks' earnings reports,https://www.cnbc.com/2024/01
22 Hours Ago,Microsoft tops Apple as world's most valuable public company,https://www.cnbc.com/2024/0
24 Hours Ago,"Flight cancellations pile up as weather, 737 Max 9 groundings disrupt travel",https://w
"January 12, 2024","Bitcoin's post-ETF launch decline may continue, but all-time highs are in sight,
"January 12, 2024",Investors selling Wells Fargo stock after earnings are focusing on the wrong thing
lab1
docs
solution
data
processed_data
market_data.csv
news_data.csv
raw_data
web_data.html
scripts
data_filter.py
task_1.py
web_scraper.py
report.docx
requirements.txt
venv
.gitattributes
.gitignore
LICENSE
OUTLINE
TIMELINE
Codespaces: sturdy yodel
main*
Ln 1, Col 1
Spaces: 4
UTF-8
LF
Plain Text
Layout: US