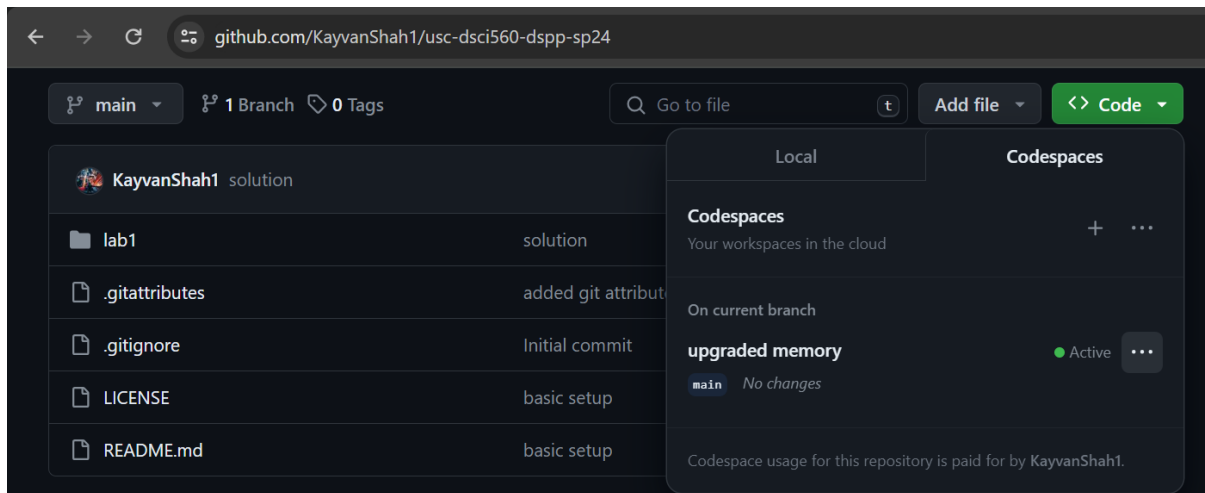


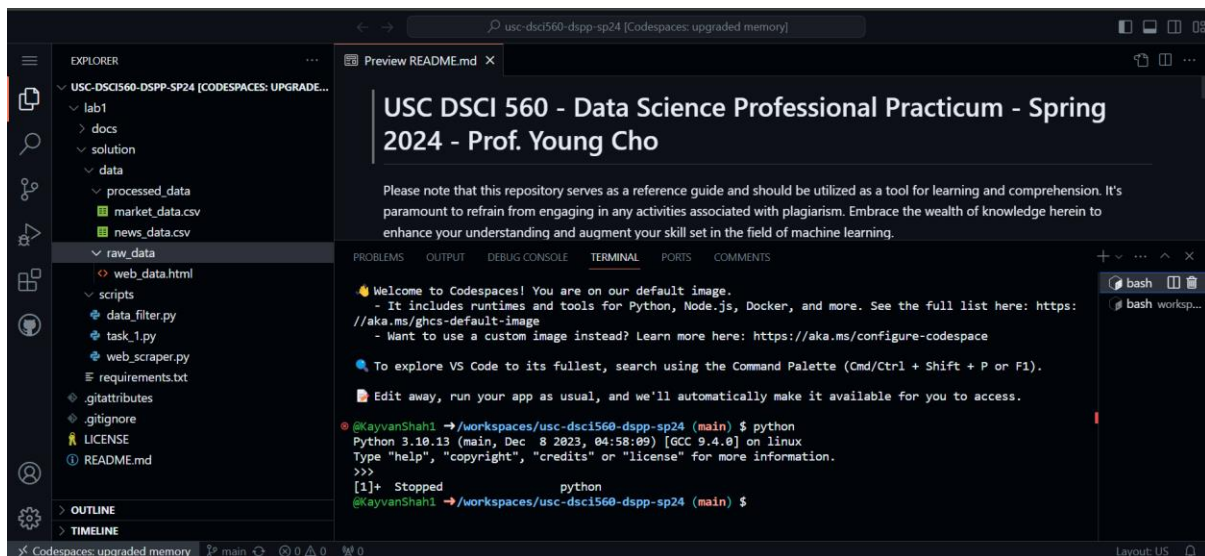
Lab 1 Solution

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

```

@KayvanShah1 → /workspaces $ mkdir kayvan_1106650685
@KayvanShah1 → /workspaces $ ls
kayvan_1106650685
@KayvanShah1 → /workspaces $ cd kayvan_1106650685
@KayvanShah1 → /workspaces/kayvan_1106650685 $ mkdir data
@KayvanShah1 → /workspaces/kayvan_1106650685 $ mkdir scripts
@KayvanShah1 → /workspaces/kayvan_1106650685 $ ls
data  scripts
@KayvanShah1 → /workspaces/kayvan_1106650685 $ cd scripts
@KayvanShah1 → /workspaces/kayvan_1106650685/scripts $ touch task_1.py
@KayvanShah1 → /workspaces/kayvan_1106650685/scripts $ ls
task_1.py
@KayvanShah1 → /workspaces/kayvan_1106650685/scripts $

```

```

task_1.py
@KayvanShah1 → /workspaces/kayvan_1106650685/scripts $ nano task_1.py
@KayvanShah1 → /workspaces/kayvan_1106650685/scripts $ cat task_1.py
# Prompt the user for input
user_name = input("Enter your name: ")

# Display the greeting in the terminal
print(f"Hello, {user_name}!")
@KayvanShah1 → /workspaces/kayvan_1106650685/scripts $

```

A basic Python Script

```

@KayvanShah1 → /workspaces/kayvan_1106650685/scripts $ python task_1.py
Enter your name: Kayvan Shah
Hello, Kayvan Shah!
@KayvanShah1 → /workspaces/kayvan_1106650685/scripts $

```

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.

```

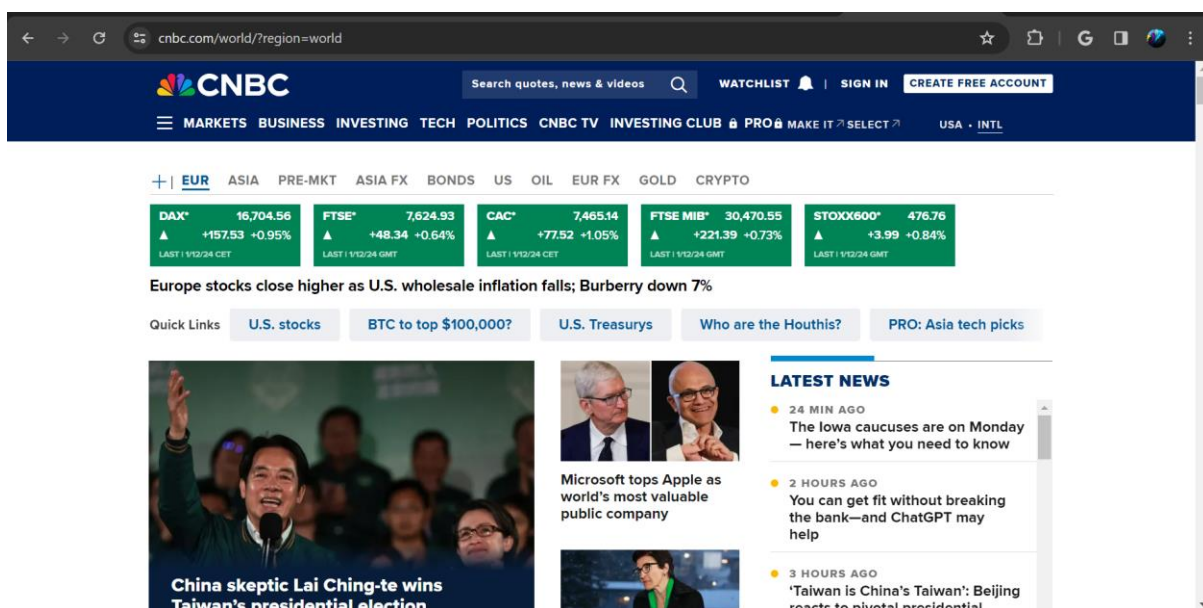
(venv) @KayvanShah1 → /workspaces/usc-dsci560-dspp-sp24/lab1/solution (main) $ which python
/workspaces/usc-dsci560-dspp-sp24/venv/bin/python
(venv) @KayvanShah1 → /workspaces/usc-dsci560-dspp-sp24/lab1/solution (main) $ pip install -r requirements.txt
Collecting beautifulsoup4
  Downloading beautifulsoup4-4.12.2-py3-none-any.whl (142 kB)
    143.0/143.0 kB 986.5 kB/s eta 0:00:00
Collecting chromedriver-py
  Downloading chromedriver_py-120.0.6099.109-py3-none-any.whl (40.1 MB)
    40.1/40.1 MB 8.6 MB/s eta 0:00:00
Collecting lxml
  Downloading lxml-5.1.0-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (8.0 MB)
    8.0/8.0 MB 35.3 MB/s eta 0:00:00
Collecting numpy
  Downloading numpy-1.26.3-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (18.2 MB)
    18.2/18.2 MB 29.3 MB/s eta 0:00:00
Collecting pandas
  Downloading pandas-2.1.4-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (12.3 MB)
    12.3/12.3 MB 34.4 MB/s eta 0:00:00
Collecting pydantic
  Downloading pydantic-2.5.3-py3-none-any.whl (381 kB)
    381.9/381.9 kB 18.5 MB/s eta 0:00:00
Collecting pydantic-settings
  Downloading pydantic_settings-2.1.0-py3-none-any.whl (11 kB)
Collecting requests
  Downloading requests-2.31.0-py3-none-any.whl (62 kB)
    62.6/62.6 kB 3.4 MB/s eta 0:00:00
Collecting selenium
  Downloading selenium-4.16.0-py3-none-any.whl (10.0 MB)
    10.0/10.0 MB 34.4 MB/s eta 0:00:00
Collecting soupsieve>1.2

```

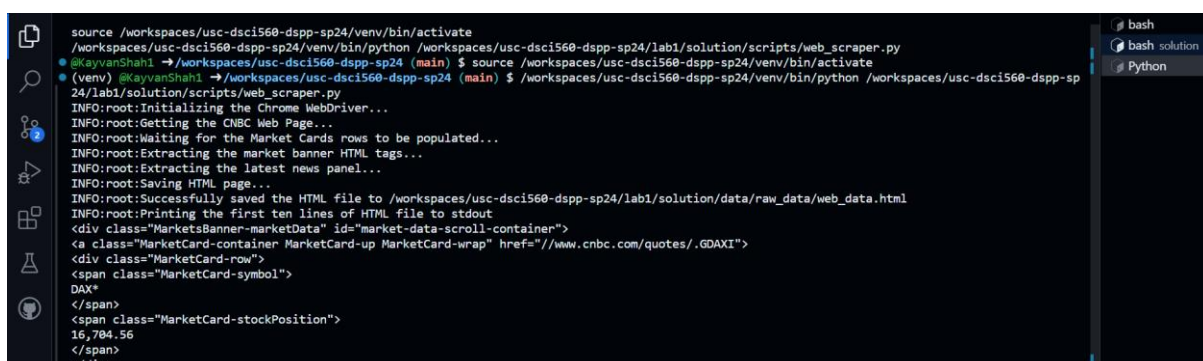
```
Installing collected packages: sortedcontainers, pytz, chromedriver-py, urllib3, tzdata, typing-extensions, soupsieve, sn
iffio, six, python-dotenv, pysocks, numpy, lxml, idna, h11, exceptiongroup, charset-normalizer, certifi, attrs, annotated
-types, wsproto, requests, python-dateutil, pydantic-core, outcome, beautifulsoup4, trio, pydantic, pandas, trio-websocket
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
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.

```
(venv) @KayvanShah1 → /workspaces/usc-dsci560-dspp-sp24 (main) $ /workspaces/usc-dsci560-dspp-sp24/venv/bin/python /workspaces/usc-dsci560-dspp-sp24/lab1/solution/scripts/data_filter.py
INFO:root:Reading and parsing the raw HTML file...
INFO:root:Filter the latest news feed...
INFO:root:Saving the filtered latest news feed data to CSV...
INFO:root:Filtering the market banner data...
INFO:root:Saving the filtered market data to CSV...
```

Output Files

- Raw data output HTML file

```
lab1 > solution > data > raw_data > web_data.html > div#market-data-scroll-container.MarketBanner-marketData
1 <div class="MarketsBanner-marketData" id="market-data-scroll-container">
2 <a class="MarketCard-container MarketCard-up MarketCard-wrap" href="//www.cnbc.com/quotes/.GDAXI">
3 <div class="MarketCard-row">
4 <span class="MarketCard-symbol">
5 DAX*
6 </span>
7 <span class="MarketCard-stockPosition">
8 16,704.56
9 </span>
10 </div>
11 <div class="MarketCard-row">
12 <span aria-hidden="true" class="MarketCard-triangle-up">
13 </span>
14 <div class="MarketCard-changeData">
15 <span class="MarketCard-changesPts">
16 +157.53
17 </span>
18 <span class="MarketCard-changesPct">
19 +0.95%
20 </span>
21 </div>
22 </div>
23 <div class="MarketCard-row">
24 <span class="MarketCard-lastTime">
25 LAST | 1/12/24 CET
```

- Processed data CSV files
 - Market data

```
lab1 > solution > data > processed_data > market_data.csv
1 symbol,stock_position,change_pts
2 DAX*,16704.56,157.53
3 FTSE*,7624.93,48.34
4 CAC*,7465.14,77.52
5 FTSE MIB*,30470.55,221.39
6 STOXX600*,476.76,3.99
7
```

- Latest news

The screenshot shows a VS Code editor window with a project named "USC-DSCI560-DSPP-SP24 [CODESPACES: sturdy yodel]". The Explorer sidebar on the left displays the project structure:

- USC-DSCI560-DSPP-SP24 [CODESPACES: sturdy yodel]
 - lab1
 - docs
 - solution
 - data
 - processed_data
 - market_data.csv
 - news_data.csv (selected)
 - raw_data
 - web_data.html (selected)
 - scripts
 - data_filter.py
 - task_1.py
 - web_scraper.py
 - report.docx
 - requirements.txt
 - venv
 - .gitattributes
 - .gitignore
 - LICENSE
 - README.md
 - OUTLINE
 - TIMELINE

The main editor area shows the content of "news_data.csv" with the following data:

```
1 timestamp,title,link
2 2 Hours Ago,Wall Street made bold calls on 4 of our stocks this week. Here's where we stand,https://w
3 2 Hours Ago,The Iowa caucuses are on Monday - here's what you need to know,https://www.cnbc.com/2024/
4 4 Hours Ago,You can get fit without breaking the bank-and ChatGPT may help,https://www.cnbc.com/2024/
5 5 Hours Ago,'Taiwan is China's Taiwan': Beijing reacts to pivotal presidential election,https://www.c
6 6 Hours Ago,'This year, I'm doing less-it's a 'great idea for a resolution,' expert says",https://www
7 6 Hours Ago,The 3 most important things we're watching in the stock market this coming week,https://w
8 6 Hours Ago,"38-year-old went from $218,000 in debt to restaurant bringing in $739,000 a year",https:
9 7 Hours Ago,"The 5 levels of financial freedom, according to married money coaches",https://www.cnbc.
10 7 Hours Ago,"In 1997, Jeff Bezos revealed why books were the 'best product' to sell on Amazon",https:
11 7 Hours Ago,"Market has cooled to start 2024, but it's still checking most of the bull boxes",https:/
12 7 Hours Ago,Investing in hobbies can be money well spent - just don't take on debt,https://www.cnbc.c
13 8 Hours Ago,How the Rubik's Cube captures hearts and market share 50 years on,https://www.cnbc.com/20
14 8 Hours Ago,Bitcoin ETFs could open floodgates to $30 trillion wealth management market,https://www.c
15 8 Hours Ago,China skeptic Lai Ching-te wins Taiwan's presidential election,https://www.cnbc.com/2024/
16 9 Hours Ago,Investor shares his top 5 picks to play an uncertain backdrop in 2024,https://www.cnbc.cc
17 9 Hours Ago,"Goldman's top stocks for 2024, including this solar company",https://www.cnbc.com/2024/e
18 21 Hours Ago,Kansas City Chiefs CEO says audience grew after player's high profile romance,https://ww
19 21 Hours Ago,Jim Cramer compares NFL playoff teams to stock picks,https://www.cnbc.com/2024/01/12/jim
20 21 Hours Ago,Cramer's Lightning Round: No to Altria,https://www.cnbc.com/2024/01/12/cramers-lightning
21 21 Hours Ago,Cramer's week ahead: Keep an eye on banks' earnings reports,https://www.cnbc.com/2024/01
22 22 Hours Ago,Microsoft tops Apple as world's most valuable public company,https://www.cnbc.com/2024/01
23 24 Hours Ago,"Flight cancellations pile up as weather, 737 Max 9 groundings disrupt travel",https://w
24 "January 12, 2024","Bitcoin's post-ETF launch decline may continue, but all-time highs are in sight,
25 "January 12, 2024","Investors selling Wells Fargo stock after earnings are focusing on the wrong things
```