

Laboratory Assignment 1

1 Installation and Setup

```
Ubuntu 24.04.3 LTS ubuntu tty1
ubuntu login: csian
Password:
Welcome to Ubuntu 24.04.3 LTS (GNU/Linux 6.8.0-90-generic aarch64)

 * Documentation:  https://help.ubuntu.com
 * Management:    https://landscape.canonical.com
 * Support:       https://ubuntu.com/pro

System information as of Thu Jan 15 02:55:45 AM UTC 2026

System load:  0.0           Temperature:           29.9 C
Usage of /:   26.1% of 9.75GB Processes:             230
Memory usage: 7%           Users logged in:       0
Swap usage:   0%           IPv4 address for enp2s0: 192.168.126.128

Expanded Security Maintenance for Applications is not enabled.

55 updates can be applied immediately.
To see these additional updates run: apt list --upgradable

Enable ESM Apps to receive additional future security updates.
See https://ubuntu.com/esm or run: sudo pro status

The programs included with the Ubuntu system are free software;
the exact distribution terms for each program are described in the
individual files in /usr/share/doc/*/copyright.

Ubuntu comes with ABSOLUTELY NO WARRANTY, to the extent permitted by
applicable law.

To run a command as administrator (user "root"), use "sudo <command>".
See "man sudo_root" for details.

csian@ubuntu:~$ _
```

ubuntu-24.04.3-live-server-arm64.iso install on VMware Fusion (Apple M4 pro)

```
csian@ubuntu:~$ which python3
/usr/bin/python3
csian@ubuntu:~$ python3 --version
Python 3.12.3
csian@ubuntu:~$ which pip3
/usr/bin/pip3
csian@ubuntu:~$ pip3 --version
pip 24.0 from /usr/lib/python3/dist-packages/pip (python 3.12)
csian@ubuntu:~$ |
```

installed Python3 and pip3

```
(venv) csian@ubuntu:~/dsci-560/lab1/JeongHoonChoi_5023184813/scripts$ df -h
Filesystem      Size  Used Avail Use% Mounted on
tmpfs           391M  1.3M  390M   1% /run
efivarfs        256K   34K  223K  14% /sys/firmware/efi/efivars
/dev/mapper/ubuntu--vg-ubuntu--lv 9.8G  3.2G  6.1G  35% /
tmpfs           2.0G    0  2.0G   0% /dev/shm
tmpfs           5.0M    0  5.0M   0% /run/lock
/dev/nvme0n1p2  1.7G  102M  1.5G   7% /boot
/dev/nvme0n1p1  952M   6.4M  945M   1% /boot/efi
tmpfs           391M   12K  391M   1% /run/user/1000
vmhgfs-fuse     927G  609G  319G  66% /mnt/hgfs
```

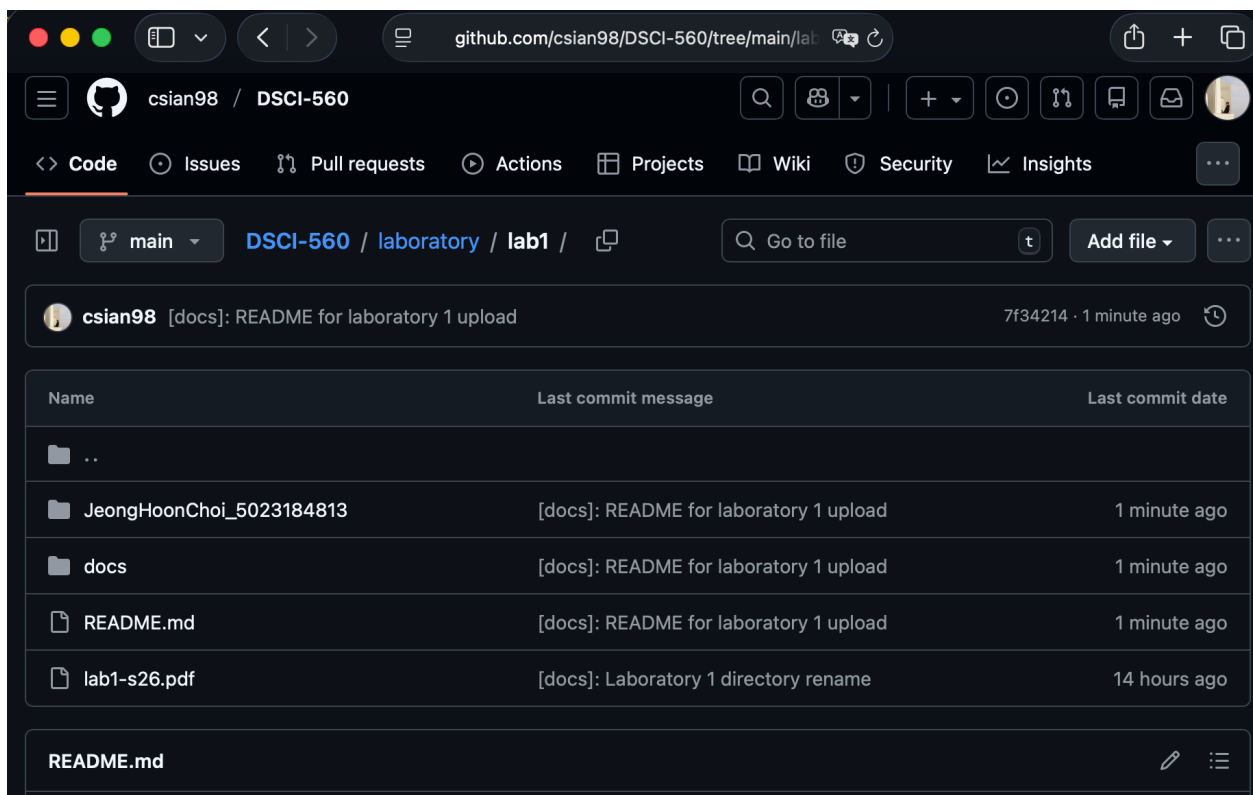
created and used a Python virtual environment using venv, and mounted a folder on the host PC using vm-tools.

2 Get Familiar with Linux and Python

2.1 Playing around with Linux Terminal

```
csian@ubuntu:~$ mkdir JeongHoonChoi_5023184813
csian@ubuntu:~$ cd JeongHoonChoi_5023184813/
csian@ubuntu:~/JeongHoonChoi_5023184813$ mkdir data scripts
csian@ubuntu:~/JeongHoonChoi_5023184813$ touch scripts/task_1.py
csian@ubuntu:~/JeongHoonChoi_5023184813$ ls scripts/
task_1.py
csian@ubuntu:~/JeongHoonChoi_5023184813$ |
```

creating directories and files using built-in Linux commands



The screenshot shows a GitHub repository page for the user 'csian98' and repository 'DSCI-560'. The page is displaying the 'main' branch and the 'laboratory/lab1' directory. The file list includes:

Name	Last commit message	Last commit date
..		
JeongHoonChoi_5023184813	[docs]: README for laboratory 1 upload	1 minute ago
docs	[docs]: README for laboratory 1 upload	1 minute ago
README.md	[docs]: README for laboratory 1 upload	1 minute ago
lab1-s26.pdf	[docs]: Laboratory 1 directory rename	14 hours ago

The 'README.md' file is selected, and its content is displayed below the file list.

uploaded codes on the GitHub user account csian98

<https://github.com/csian98/DSCI-560/tree/main/laboratory/lab1>

2.2 A basic Python Script

```
csian@ubuntu:~/JeongHoonChoi_5023184813/scripts$ vim task_1.py
csian@ubuntu:~/JeongHoonChoi_5023184813/scripts$ cat task_1.py
# store the input in the 'name' variable
# name = "Jeong Hoon Choi"
print("Enter your name:")
name = input()

# formatted string output
print(f"Hello, {name}!")

csian@ubuntu:~/JeongHoonChoi_5023184813/scripts$ python3 task_1.py
Enter your name:
Jeong Hoon Choi
Hello, Jeong Hoon Choi!
csian@ubuntu:~/JeongHoonChoi_5023184813/scripts$ |
```

Python standard input (stdin) and standard output(stdout)

2.3 Python Web-scraping Task

I tried to read the file statically using the `requests` library, but the JavaScript file containing the market data was loaded as is. Dynamic web scraping was performed using a browser engine.

```
<div class="MarketsBanner-main" id="HomePageInternational-MarketsBanner-1-panel">
  <div class="MarketsBanner-marketData" id="market-data-scroll-container">
    <a class="MarketCard-container" href="//www.cnbc.com/quotes/.STOXX">
      <div class="MarketCard-row">
        <span class="MarketCard-symbol">
          STOXX600*
        </span>
        <span class="MarketCard-stockPosition">
          611.56
        </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 | 1/14/26 GMT
        </span>
      </div>
    </a>
  </div>
</div>
```

All MarketCard-row data is contained within the div.MarketBanner-main element.

```
<ul class="LatestNews-list">
  <li class="LatestNews-item" id="HomePageInternational-latestNews-7-0">
    <div class="LatestNews-container">
      <div class="LatestNews-headlineWrapper">
        <span class="LatestNews-wrapper">
          <time class="LatestNews-timestamp">
            17 Min Ago
          </time>
        </span>
        <a class="LatestNews-headline" href="https://www.cnbc.com/2026/01/15/european-markets-live-updates-sto
tse-cac.html" title="European markets head for higher open as traders track Greenland, Iran news">
          European markets head for higher open as traders track Greenland, Iran news
        </a>
      </div>
    </div>
  </li>
  <li class="LatestNews-item" id="HomePageInternational-latestNews-7-1">
    <div class="LatestNews-container">
      <div class="LatestNews-headlineWrapper">
```

All LatestNews-item data is contained within the ul.LatestNews-list element.

```
(venv) csian@ubuntu:~/dsci-560/lab1/JeongHoonChoi_5023184813$ cat data/raw_data/web_data.html | head -n 10
<div class="MarketsBanner-marketData" id="market-data-scroll-container">
  <a class="MarketCard-container" href="//www.cnbc.com/quotes/.STOXX">
    <div class="MarketCard-row">
      <span class="MarketCard-symbol">
        STOXX600*
      </span>
      <span class="MarketCard-stockPosition">
        611.56
      </span>
    </div>
  </a>
(venv) csian@ubuntu:~/dsci-560/lab1/JeongHoonChoi_5023184813$
```

Parsing web_data.html, printing the first 10 lines

2.4 Data Filtering Task

```
(venv) csian@ubuntu:~/DSCI-560/lab1/JeongHoonChoi_5023184813/scripts$ python3 data_filter.py
2026-01-15 21:27:30,610 - root - DEBUG - Reading web_data.html
2026-01-15 21:27:30,614 - root - DEBUG - Filtering market data: #1
2026-01-15 21:27:30,615 - root - DEBUG - Filtering market data: #2
2026-01-15 21:27:30,615 - root - DEBUG - Filtering market data: #3
2026-01-15 21:27:30,615 - root - DEBUG - Filtering market data: #4
2026-01-15 21:27:30,615 - root - DEBUG - Filtering market data: #5
2026-01-15 21:27:30,615 - root - DEBUG - Filtering news data: #1
2026-01-15 21:27:30,615 - root - DEBUG - Filtering news data: #2
2026-01-15 21:27:30,615 - root - DEBUG - Filtering news data: #3
2026-01-15 21:27:30,615 - root - DEBUG - Filtering news data: #4
2026-01-15 21:27:30,615 - root - DEBUG - Filtering news data: #5
2026-01-15 21:27:30,615 - root - DEBUG - Filtering news data: #6
2026-01-15 21:27:30,615 - root - DEBUG - Filtering news data: #7
2026-01-15 21:27:30,616 - root - DEBUG - Filtering news data: #8
2026-01-15 21:27:30,616 - root - DEBUG - Filtering news data: #9
2026-01-15 21:27:30,616 - root - DEBUG - Filtering news data: #10
2026-01-15 21:27:30,616 - root - DEBUG - Filtering news data: #11
2026-01-15 21:27:30,616 - root - DEBUG - Filtering news data: #12
2026-01-15 21:27:30,616 - root - DEBUG - Filtering news data: #13
2026-01-15 21:27:30,616 - root - DEBUG - Filtering news data: #14
2026-01-15 21:27:30,616 - root - DEBUG - Filtering news data: #15
2026-01-15 21:27:30,616 - root - DEBUG - Filtering news data: #16
2026-01-15 21:27:30,616 - root - DEBUG - Filtering news data: #17
2026-01-15 21:27:30,616 - root - DEBUG - Filtering news data: #18
```

execute data_filter.py, including logging

```
(venv) csian@ubuntu:~/dsci-560/lab1/JeongHoonChoi_5023184813/data/processed_data$ cat market_data.csv | head -n 10
STOXX600*,611.56,UNCH
DAX*,25,286.24,UNCH
FTSE*,10,184.35,+0.46%
CAC*,8,330.97,UNCH
FTSE MIB*,45,647.40,+0.27%
(venv) csian@ubuntu:~/dsci-560/lab1/JeongHoonChoi_5023184813/data/processed_data$ |
```

```
(venv) csian@ubuntu:~/dsci-560/lab1/JeongHoonChoi_5023184813/data/processed_data$ cat news_data.csv | head -n 10
35 Min Ago,https://www.cnbc.com/2026/01/15/european-markets-live-updates-stoxx-600-dax-ftse-cac.html,European markets he
ad for higher open as traders track Greenland, Iran news
1 Hour Ago,https://www.cnbc.com/2026/01/15/tsmc-q4-profit-record-ai-chip-demand-nt1-trillion.html,TSMC fourth-quarter pr
ofit beats estimates, soaring 35%, as AI chip demand stays strong
2 Hours Ago,https://www.cnbc.com/2026/01/15/goldman-sachs-gs-q4-2025-earnings.html,Goldman Sachs is set to report earnin
gs - here's what the Street expects
2 Hours Ago,https://www.cnbc.com/2026/01/15/-gene-yu-us-special-forces-cybersecurity-startup-blackpanda.html,This former
U.S. special forces officer raised $22 million for his startup
2 Hours Ago,https://www.cnbc.com/2026/01/15/us-greenland-trump-takeover-iceland-president.html,U.S. seizing Greenland ri
sks 'monumental' fallout, ex-Iceland president warns
3 Hours Ago,https://www.cnbc.com/2026/01/15/us-stops-immigrant-visas-for-75-countries-see-the-full-list.html,U.S. freeze
s new immigrant visas for 75 countries: See the full list
4 Hours Ago,https://www.cnbc.com/2026/01/15/china-ai-chips-ipos-huawei-hisilicon-nvidia-rivals-biren-metax-moore-threads
-smic.html,How Huawei's dominance hangs over China's AI chip IPO boom
4 Hours Ago,https://www.cnbc.com/2026/01/15/tripcom-shares-plunge-as-china-opens-antitrust-probe-into-company.html,Trip.
com shares plunge over 20% as China opens antitrust probe into Asia's largest online travel firm
5 Hours Ago,https://www.cnbc.com/2026/01/15/iran-briefly-closes-airspace-as-us-tensions-rise-flights-rerouted-across-reg
ion.html,Iran reopens airspace after hours-long shutdown spooks airlines
5 Hours Ago,https://www.cnbc.com/2026/01/14/musk-xai-blocks-grok-chatbot-from-creating-sexualized-images-of-people.html,
Musk's xAI limits Grok's ability to create sexual images of people on X after backlash
(venv) csian@ubuntu:~/dsci-560/lab1/JeongHoonChoi_5023184813/data/processed_data$ |
```

The data for marketCard_stockPosition, LatestNews-title itself contains commas. Substitution and string processing are required for the CSV format.

```
(venv) csian@ubuntu:~/DSCI-560/lab1/JeongHoonChoi_5023184813/data/processed_data$ cat *.csv
STOXX600*,611.56,UNCH
DAX*,25286.24,UNCH
FTSE*,10184.35,+0.46%
CAC*,8330.97,UNCH
FTSE MIB*,45647.40,+0.27%
35 Min Ago,https://www.cnbc.com/2026/01/15/european-markets-live-updates-stoxx-600-dax-ftse-cac.html,"European markets head for higher open as traders track Greenland, Iran news"
1 Hour Ago,https://www.cnbc.com/2026/01/15/tsmc-q4-profit-record-ai-chip-demand-nt1-trillion.html,"TSMC fourth-quarter profit beats estimates, soaring 35%, as AI chip demand stays strong"
2 Hours Ago,https://www.cnbc.com/2026/01/15/goldman-sachs-gs-q4-2025-earnings.html,"Goldman Sachs is set to report earnings - here's what the Street expects"
2 Hours Ago,https://www.cnbc.com/2026/01/15/-gene-yu-us-special-forces-cybersecurity-startup-blackpanda.html,"This former U.S. special forces officer raised $22 million for his startup"
2 Hours Ago,https://www.cnbc.com/2026/01/15/us-greenland-trump-takeover-iceland-president.html,"U.S. seizing Greenland risks 'monumental' fallout, ex-Iceland president warns"
3 Hours Ago,https://www.cnbc.com/2026/01/15/us-stops-immigrant-visas-for-75-countries-see-the-full-list.html,"U.S. freezes new immigrant visas for 75 countries: See the full list"
4 Hours Ago,https://www.cnbc.com/2026/01/15/china-ai-chips-ipos-huawei-hisilicon-nvidia-rivals-biren-metax-moore-threads-smic.html,"How Huawei's dominance hangs over China's AI chip IPO boom"
```