Tanya Charan
USC ID: 6316469892
YOUTUBE LINK: https://youtu.be/SS-X1uPgndg
GITHUB LINK: https://github.com/tanyacharan/dsci560-lab1

Overview:

The assignment required us to:

1. Setup the environment: Ubuntu VM, Python, pip.

2. Go over Linux and Python essential basics: terminal navigation, understanding a simple script.

3. Do web scraping: fetch CNBC World News webpage.

4. Do Data Filtering: extract specific fields into CSVs.

5. Update Github: commit/push daily with meaningful commit messages.

This is how I set up my files in Ubuntu:

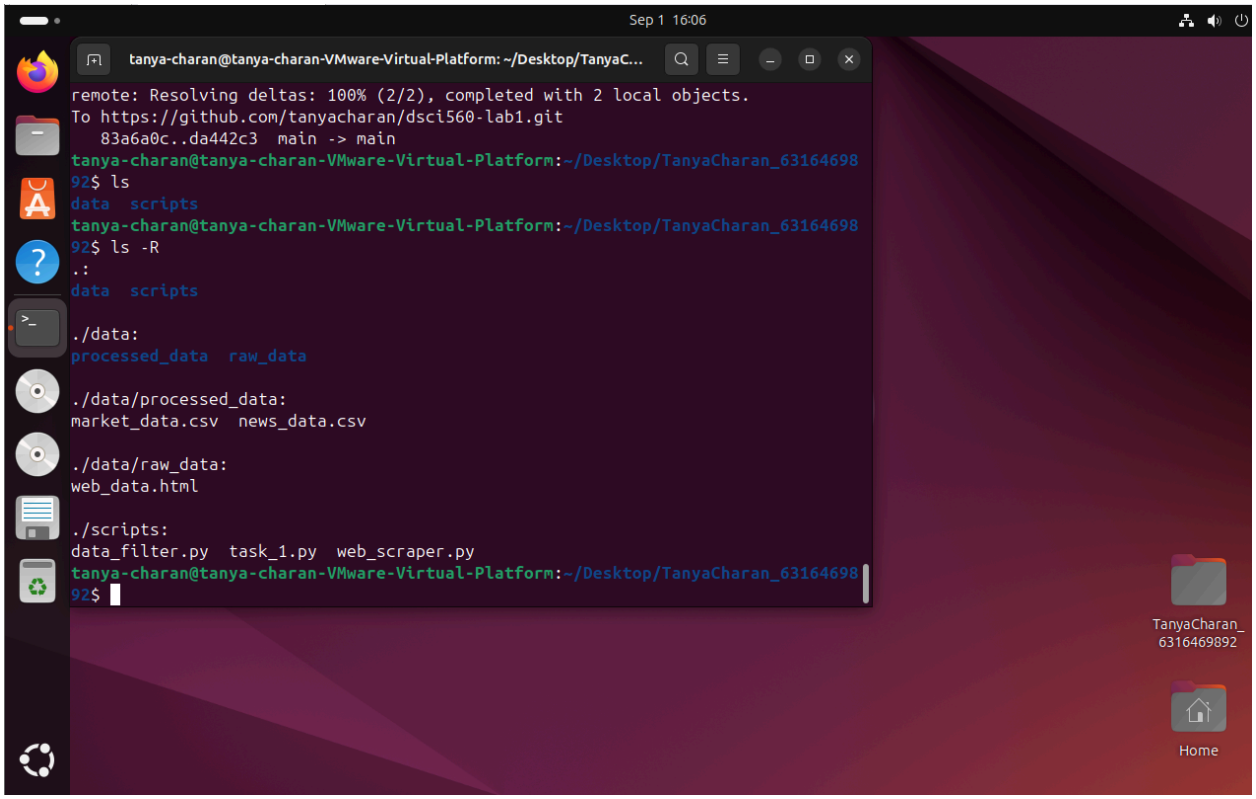
```
TanyaCharan_6316469892/
├── data/
│   ├── raw_data/
│   │   └── web_data.html
│   └── processed_data/
│       ├── market_data.csv
│       └── news_data.csv
├── scripts/
│   ├── task_1.py
│   ├── web_scraper.py
│   └── data_filter.py
└──
```

For the basic python script I was able to prompt a user to put in a name and then give out the greeting.

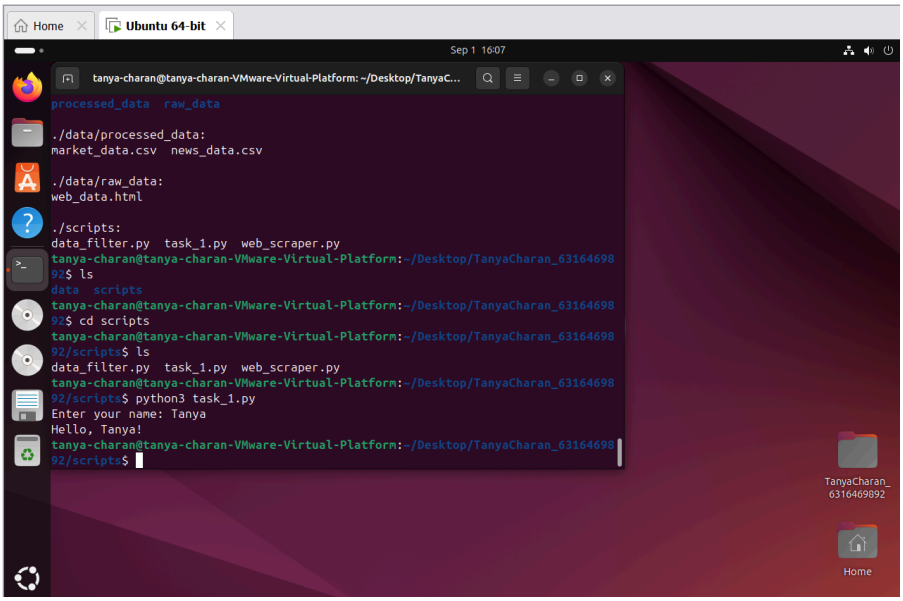
For the web scraping task I was able to use web_scraper.py to implement requests and BeautifulSoup to collect raw HTML from the CNBC website and then save that to an html file (web_data.html). Then I looked at the file contents to make sure there was information loaded in the html file.

For the data filtering task I had to create data_filter.py. For Latest News I saved the information to news_data.csv. I ran into an issue with the Market Banner. Though I wrote the code, I noticed that the website uses JavaScript for this information so the script I wrote does not extract the information how I wanted it to. To help debug and understand this issue in the script I included console messages and print statements in between each step.

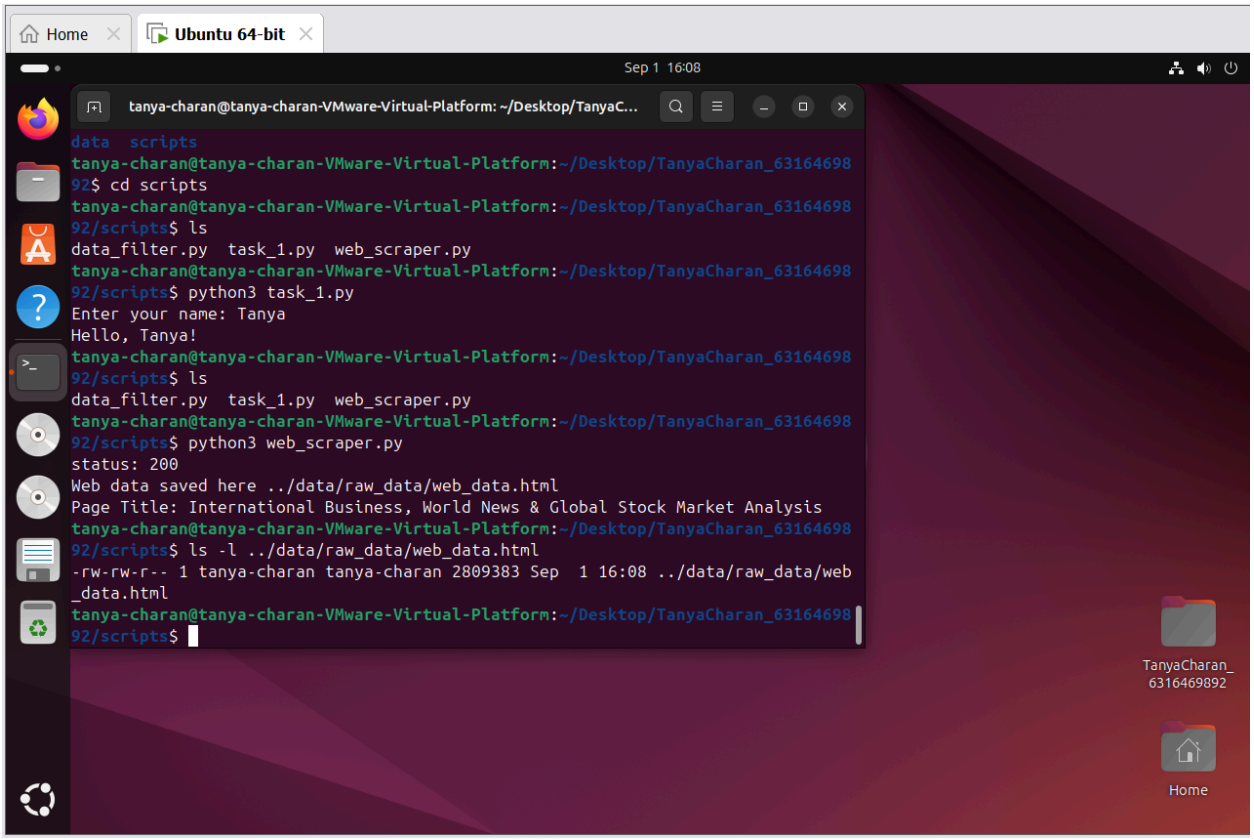
Screenshots of my lab:
Here is the file hierarchy as shown in the terminal:



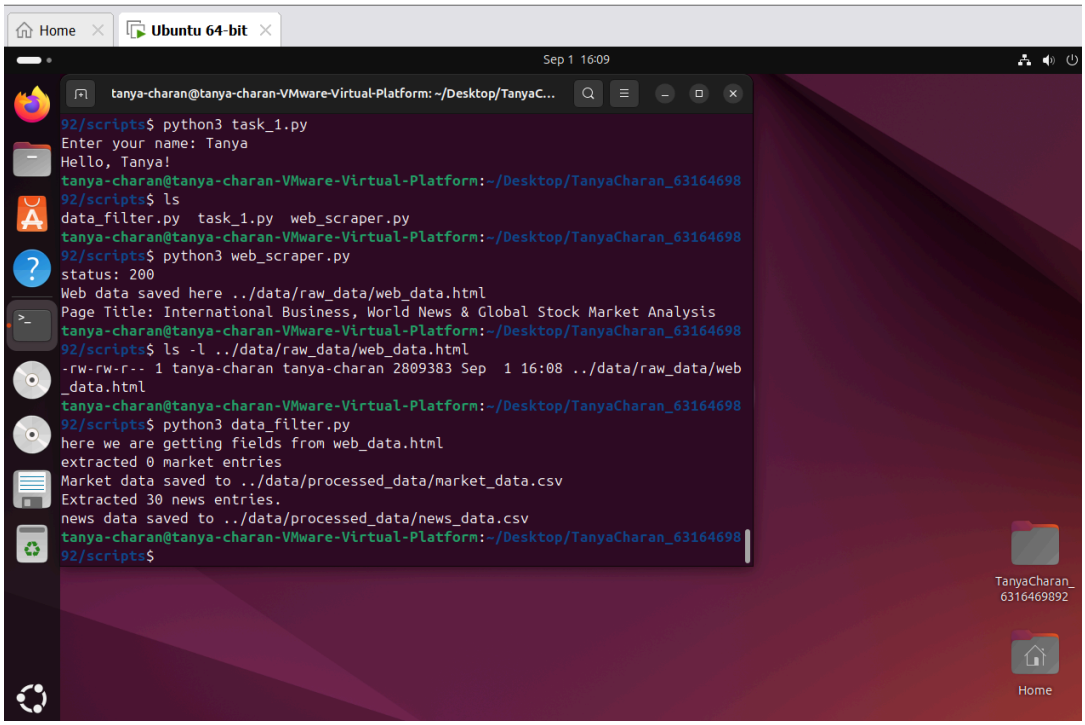
Here is an example of task_1.py running in the terminal:

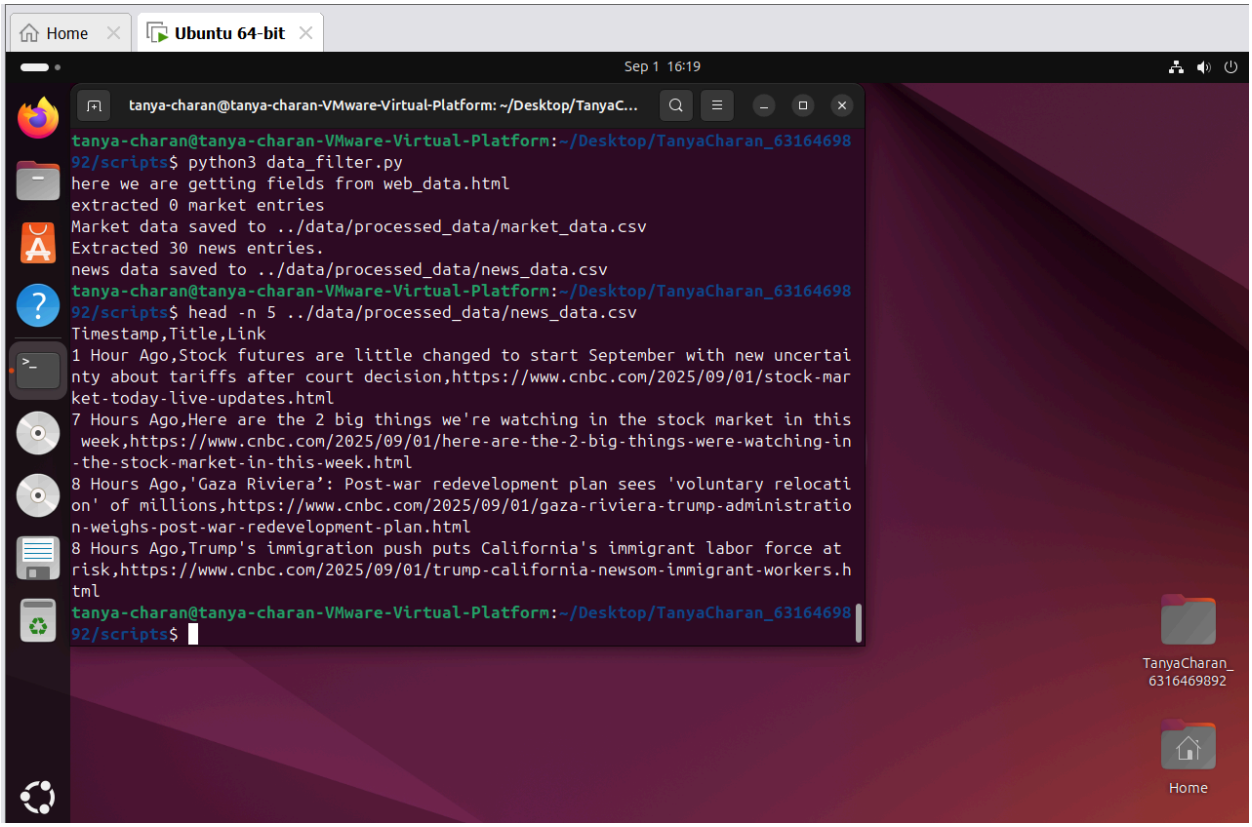



Here is an example of what running web_scraper.py looks like:



Here I am running data_filter.py:




Looking at news_data.csv: