**Name: Yiyang Jia, USC ID: 3882757155**

**Github link: https://github.com/YiyangJiaR/DSCI-560/tree/main**

**Lab 1 Solution:**

1. Installation & Setup

* Using Mac and M chips so download the VMware fusion and Ubuntu ARM version to create the new virtual machine.
* Create new public repo of DSCI 560

A screenshot of a computer

Description automatically generated

1. Playing Around with Linux Terminal

* Open the Linux Terminal and write the commands below to achieve the requirements. Just use the mkdir to create new directory and nano to get into the .py file to edit and cat to show what the files got after run it.

A screenshot of a computer

Description automatically generated

1. A basic python script

A computer screen shot of a computer code

Description automatically generated

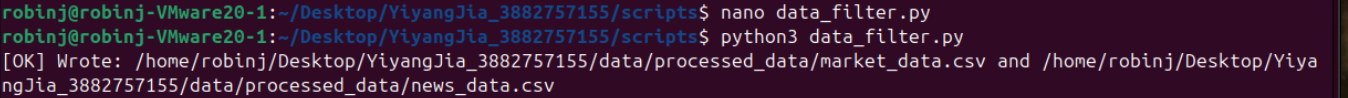
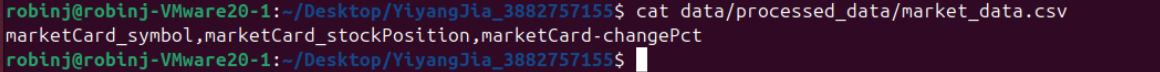
1. Python Web Scraping Task

I first installed the requests and beautifulsoup4 libraries and created the web\_scraper.py file inside the scripts and the raw\_data and processed\_data inside the web\_scraper.py file when nano inside to edit it. When first write the web\_scraper.py I just use the request.get without headers inside the function, however, it does not work so I need to add the real browser headers then it works. As requested, fetching CNBC world page HTML and saves it as web\_data.html inside the raw\_data which also inside the data directory. The results show the 10 headlines of the web\_data.html file. You will see the web\_scraper.py in the GitHub repo link.

A computer screen shot of a computer code

Description automatically generated

1. Data Filtering Task



I did the data filtering and when get to the market\_data.csv and the news\_data.csv it shows just the headers, and I think it’s the html problem and as shown below my web\_data.html only shows the CSS class names but not the data-symbol attributes and I try to redefine the headers, but I does not work out. I do not know how to fix it, but I think the logic of the code is correct and match the lab requirements. I will try to fix out. A screen shot of a computer program

Description automatically generated