

To complete Objective one I used GCP cloud.

First you create vm machine on GCP.

I follow below documents to create vm and install jupyter in vm

<https://sinanartun.medium.com/how-to-setup-jupyter-notebook-on-google-cloud-platform-42751e585fc7>

Then I create a folder

`mkdir scraping` (make folder)

`Nano file.py` (create python file)

Copy code from file.py (which I give to you) and paste here

Press `ctrl+s` (for save)

`Ctrl+x` (exit)

The run command

`Python file.py`

I used tmux to run script to keep script in running even my system is off.

I follow below video to do above :

<https://www.youtube.com/watch?v=IEKp2O7MTfY&t=1s>

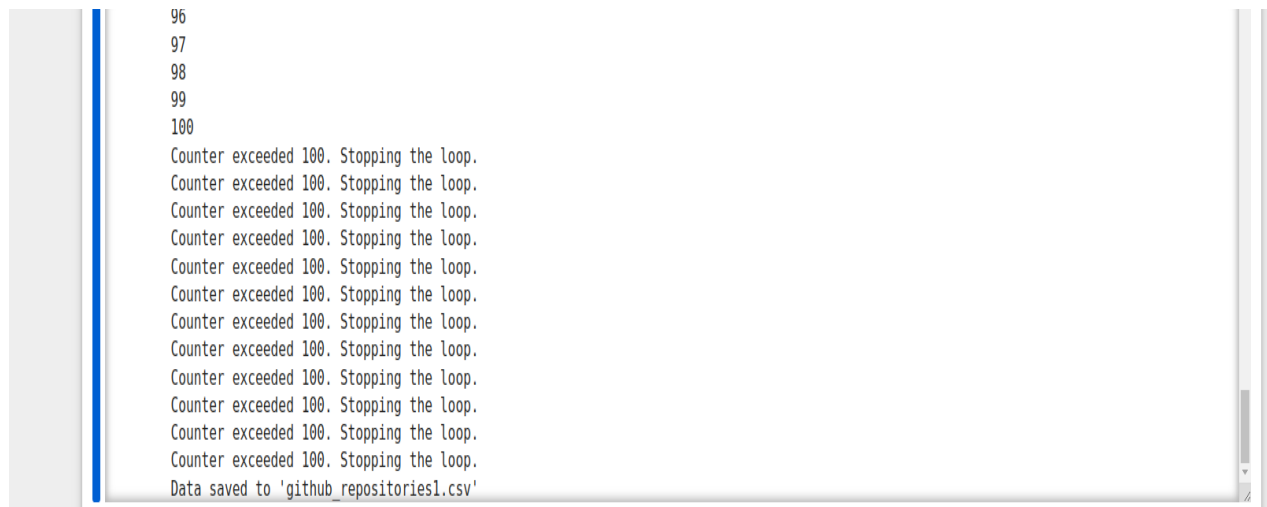
`tmux new -s demo`

(to create session, here session name is demo)

`demotmux ls` (to check how many session is running)

`demotmux a -t demo` (to check the script is running or not)

After scraping all data,



```
96
97
98
99
100
Counter exceeded 100. Stopping the loop.
Counter exceeded 100. Stopping the loop.
Counter exceeded 100. Stopping the loop.
Counter exceeded 100. Stopping the loop.
Counter exceeded 100. Stopping the loop.
Counter exceeded 100. Stopping the loop.
Counter exceeded 100. Stopping the loop.
Counter exceeded 100. Stopping the loop.
Counter exceeded 100. Stopping the loop.
Counter exceeded 100. Stopping the loop.
Counter exceeded 100. Stopping the loop.
Counter exceeded 100. Stopping the loop.
Counter exceeded 100. Stopping the loop.
Counter exceeded 100. Stopping the loop.
Data saved to 'github_repositories1.csv'
```

The reason of these extra lines : counter exceed 100.stop....

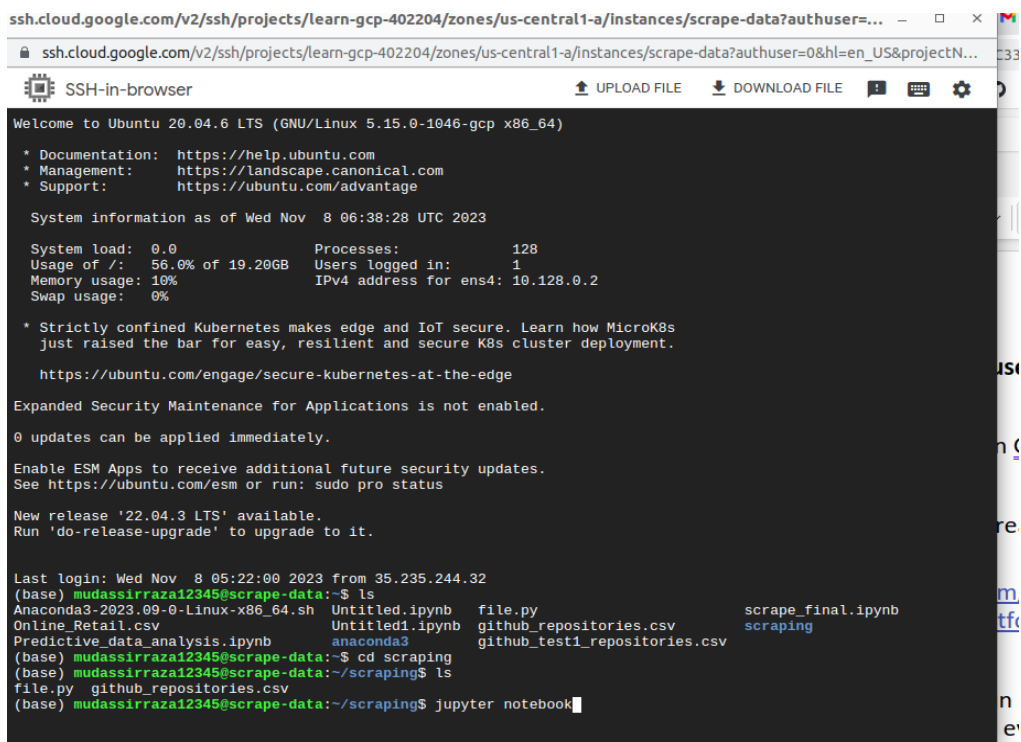
The thread is waiting till all data fatching then it will shutdown

That why come but it will stop after all fetching by threads

I will start exploratory data analysis.

run command `jupyter-notebook`

In the scrapping folder



The screenshot shows a terminal window titled "SSH-in-browser" with a URL bar at the top. The terminal output includes system information for Ubuntu 20.04.6 LTS, system load, usage statistics, and a list of files in the current directory. The user has navigated to the 'scrapping' directory and run the 'jupyter notebook' command.

```
ssh.cloud.google.com/v2/ssh/projects/learn-gcp-402204/zones/us-central1-a/instances/scrape-data?authuser=...
ssh.cloud.google.com/v2/ssh/projects/learn-gcp-402204/zones/us-central1-a/instances/scrape-data?authuser=0&hl=en_US&projectN...
SSH-in-browser
Welcome to Ubuntu 20.04.6 LTS (GNU/Linux 5.15.0-1046-gcp x86_64)

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

System information as of Wed Nov  8 06:38:28 UTC 2023

System load:  0.0               Processes:    128
Usage of /:   56.0% of 19.20GB   Users logged in: 1
Memory usage: 10%              IPv4 address for ens4: 10.128.0.2
Swap usage:   0%

* Strictly confined Kubernetes makes edge and IoT secure. Learn how MicroK8s
  just raised the bar for easy, resilient and secure K8s cluster deployment.

https://ubuntu.com/engage/secure-kubernetes-at-the-edge

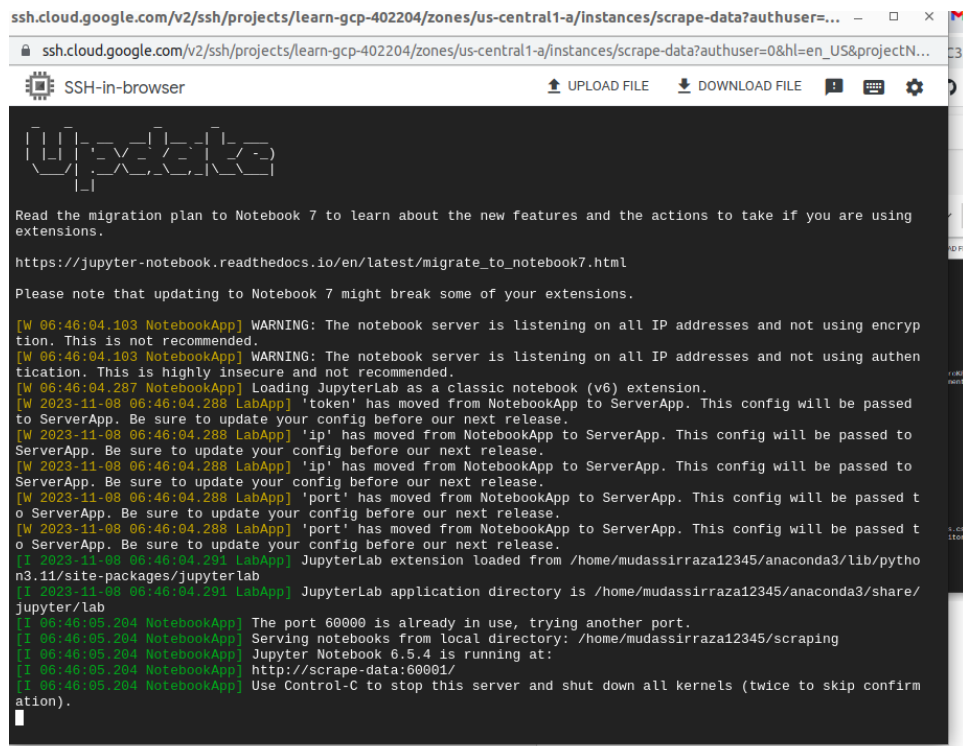
Expanded Security Maintenance for Applications is not enabled.

0 updates can be applied immediately.

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

New release '22.04.3 LTS' available.
Run 'do-release-upgrade' to upgrade to it.

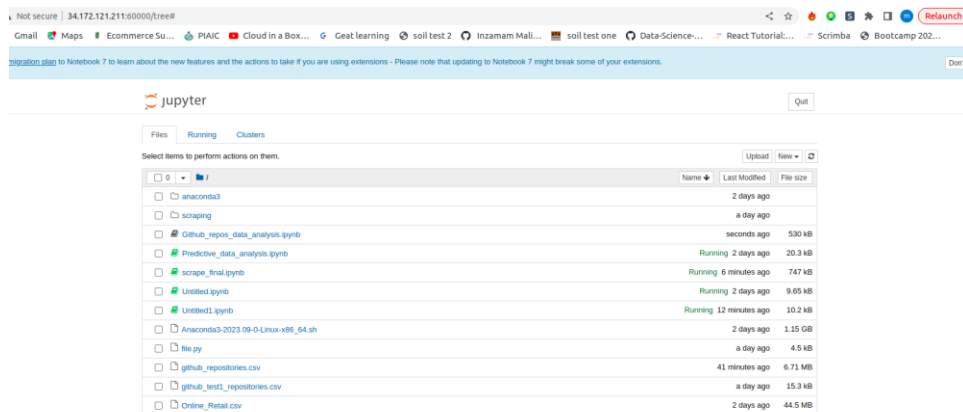
Last login: Wed Nov  8 05:22:00 2023 from 35.235.244.32
(base) mudassirraza12345@scrape-data:~$ ls
Anaconda3-2023.09-0-Linux-x86_64.sh  Untitled1.ipynb  file.py  github_repositories.csv  scrape_final.ipynb
Online_Retail.csv                  anaconda3        github_test1_repositories.csv  scrapping
Predictive_data_analysis.ipynb
(base) mudassirraza12345@scrape-data:~$ cd scrapping
(base) mudassirraza12345@scrape-data:~/scrapping$ ls
file.py  github_repositories.csv
(base) mudassirraza12345@scrape-data:~/scrapping$ jupyter notebook
```



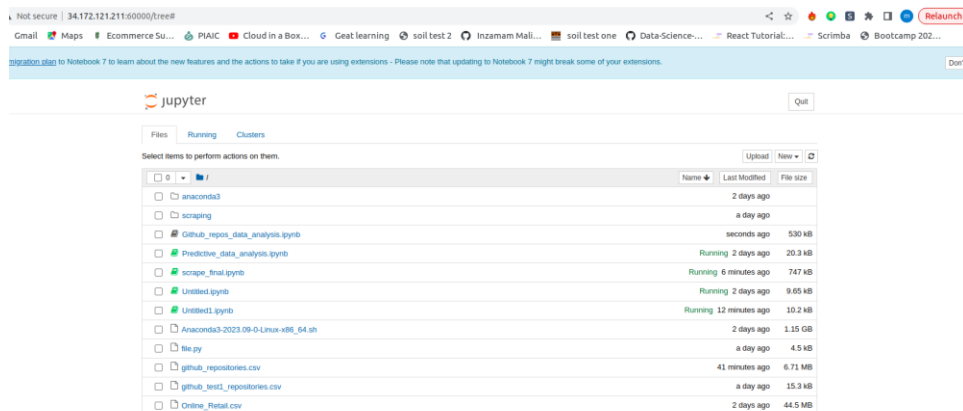
Then go to google

<http://<your vm ip>:60000/>

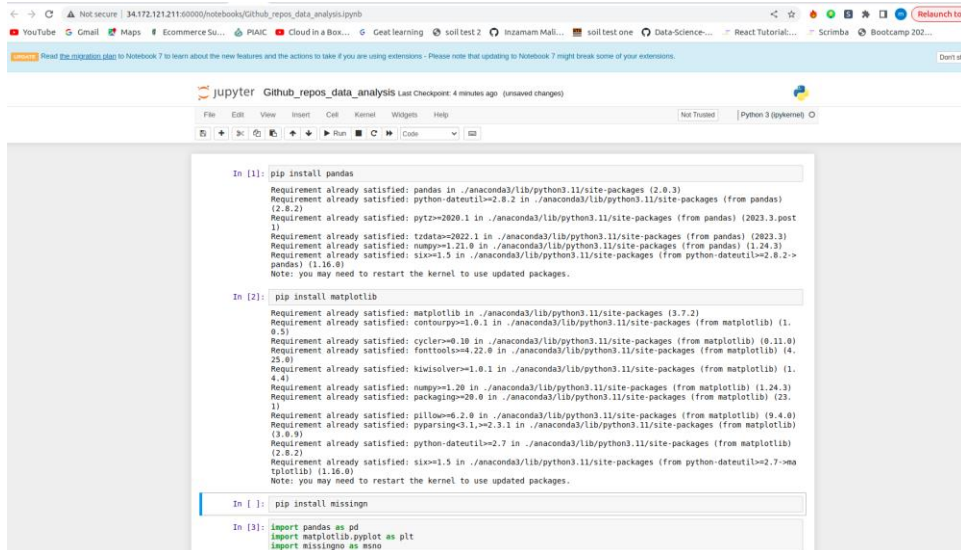
This jupyter terminal is open



click to upload and select the scrape data analysis notebook from system which I send you and upload.



Then click to [Github_repos_data_analysis.ipynb](#)



The screenshot shows a Jupyter Notebook interface with a browser window at the top. The notebook title is "Github_repos_data_analysis" and it was last checked 4 minutes ago. The interface includes a menu bar (File, Edit, View, Insert, Cell, Kernel, Widgets, Help) and a toolbar with icons for file operations, running cells, and code execution. The notebook content consists of three code cells. The first cell runs "pip install pandas" and shows the output of the pip command, indicating that pandas and its dependencies are already installed. The second cell runs "pip install matplotlib" and shows the output of the pip command, indicating that matplotlib and its dependencies are already installed. The third cell runs "pip install missingno" and shows the output of the pip command, indicating that missingno is already installed. The fourth cell runs "import pandas as pd", "import matplotlib.pyplot as plt", and "import missingno as mso".

```
In [1]: pip install pandas

Requirement already satisfied: pandas in ./anaconda3/lib/python3.11/site-packages (2.0.3)
Requirement already satisfied: python-dateutil<=2.8.2 in ./anaconda3/lib/python3.11/site-packages (from pandas) (2.8.2)
Requirement already satisfied: pytz<=2020.1 in ./anaconda3/lib/python3.11/site-packages (from pandas) (2023.3.post1)
Requirement already satisfied: tzdata<=2022.1 in ./anaconda3/lib/python3.11/site-packages (from pandas) (2023.3)
Requirement already satisfied: numpy<=1.21.0 in ./anaconda3/lib/python3.11/site-packages (from pandas) (1.24.3)
Requirement already satisfied: six<=1.5 in ./anaconda3/lib/python3.11/site-packages (from python-dateutil<=2.8.2->pandas) (1.16.0)
Note: you may need to restart the kernel to use updated packages.

In [2]: pip install matplotlib

Requirement already satisfied: matplotlib in ./anaconda3/lib/python3.11/site-packages (3.7.2)
Requirement already satisfied: contourpy<=1.0.1 in ./anaconda3/lib/python3.11/site-packages (from matplotlib) (1.0.5)
Requirement already satisfied: cycler<=0.10 in ./anaconda3/lib/python3.11/site-packages (from matplotlib) (0.11.0)
Requirement already satisfied: fonttools<=4.22.0 in ./anaconda3/lib/python3.11/site-packages (from matplotlib) (4.25.0)
Requirement already satisfied: kiwisolver<=1.0.1 in ./anaconda3/lib/python3.11/site-packages (from matplotlib) (1.4.4)
Requirement already satisfied: numpy<=1.20 in ./anaconda3/lib/python3.11/site-packages (from matplotlib) (1.24.3)
Requirement already satisfied: packaging<=20.0 in ./anaconda3/lib/python3.11/site-packages (from matplotlib) (23.1)
Requirement already satisfied: pillow<=6.2.0 in ./anaconda3/lib/python3.11/site-packages (from matplotlib) (9.4.0)
Requirement already satisfied: pyparsing<3.1.>=2.3.1 in ./anaconda3/lib/python3.11/site-packages (from matplotlib) (3.0.9)
Requirement already satisfied: python-dateutil<=2.7 in ./anaconda3/lib/python3.11/site-packages (from matplotlib) (2.8.2)
Requirement already satisfied: six<=1.5 in ./anaconda3/lib/python3.11/site-packages (from python-dateutil<=2.7->matplotlib) (1.16.0)
Note: you may need to restart the kernel to use updated packages.

In [ ]: pip install missingno

In [3]: import pandas as pd
import matplotlib.pyplot as plt
import missingno as mso
```

This screen come and then click to cell and run each cell by shift+Enter

