Creator: Akanksha Singh Tomar

// SAU Id num:9999 03732

//Course name: MCIS6273_030242S- S25 Data Mining

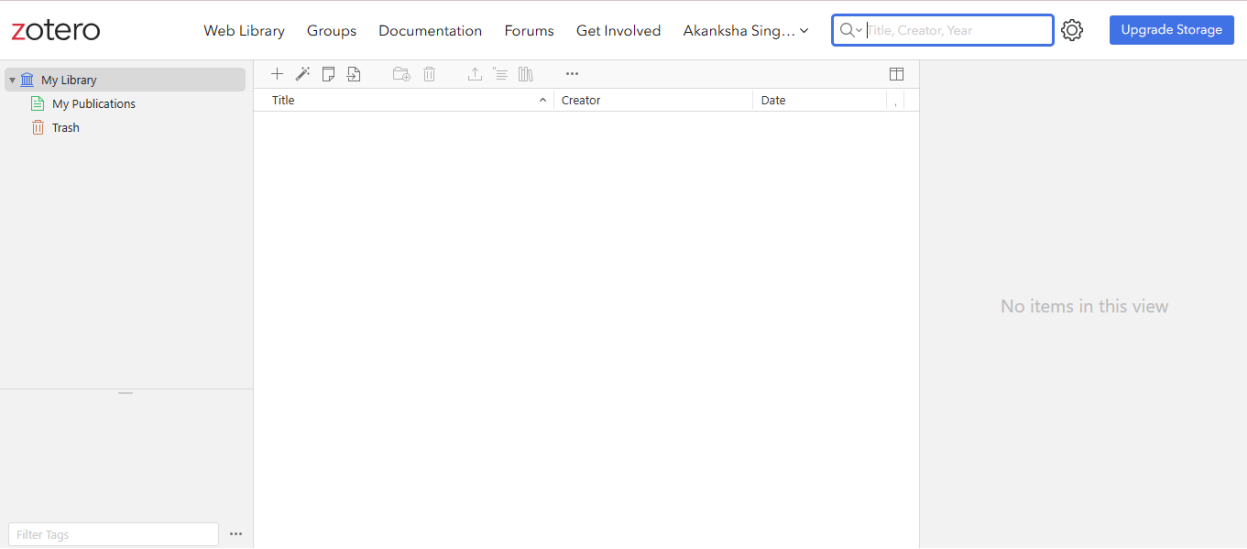
//Instructor name: Keith Maull

Date:2/11/2025 MCIS6273_030242S- S25 Data Mining_AKANKSHASINGHTOMAR_999903732_H1

Homework: -1

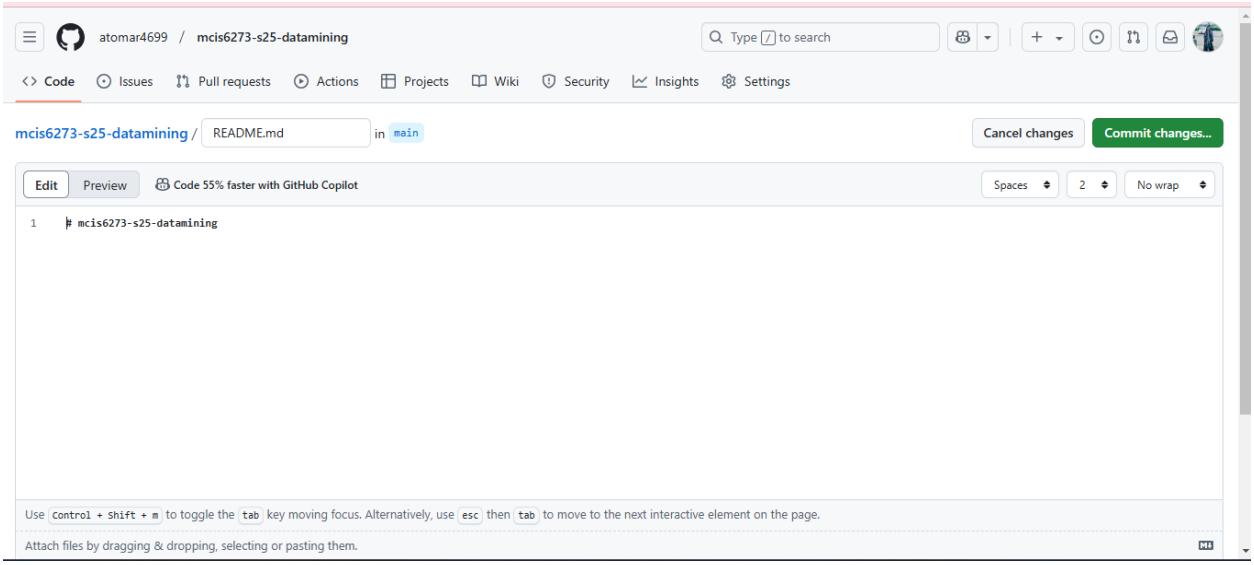
Step 1: Set Up GitHub and Zotero

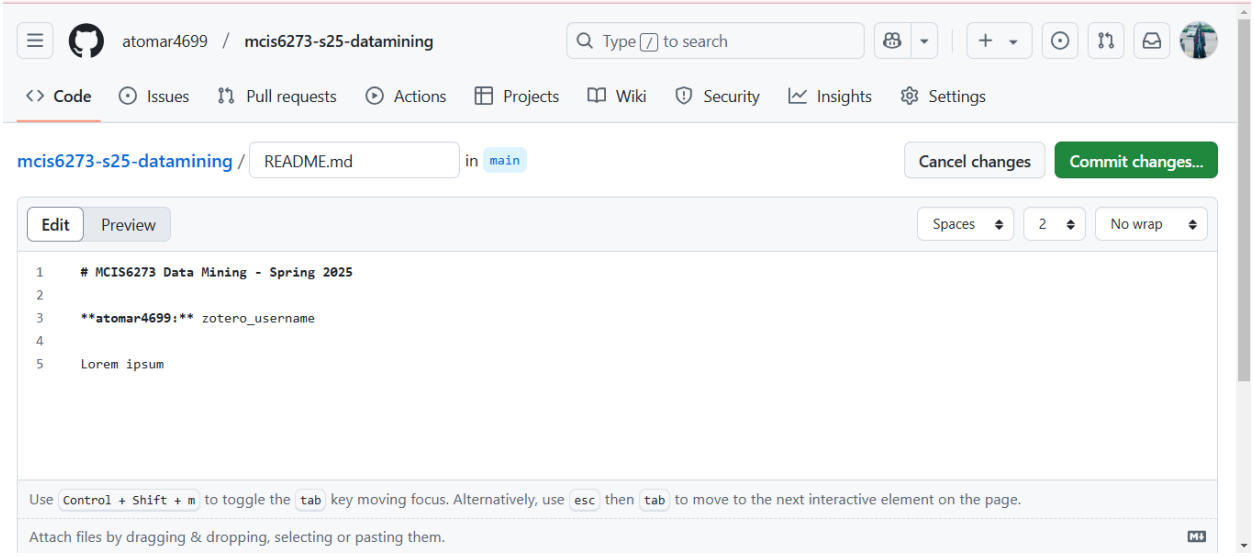
1. **Create a Zotero Account**



2. **Create a GitHub Repository**
    - Go to GitHub and create a new public repository named: **mcis6273-s25-datamining**
3. **Go to your repository**
    - Visit GitHub and navigate to your repository (mcis6273-s25-datamining).
    - **Create a new file**
    - Click on **"Add file" > "Create new file"**.
    - In the **file name** field, type: README.md.
    - **Add Content**
    - Enter the following content inside the file.
        - Your Zotero username.

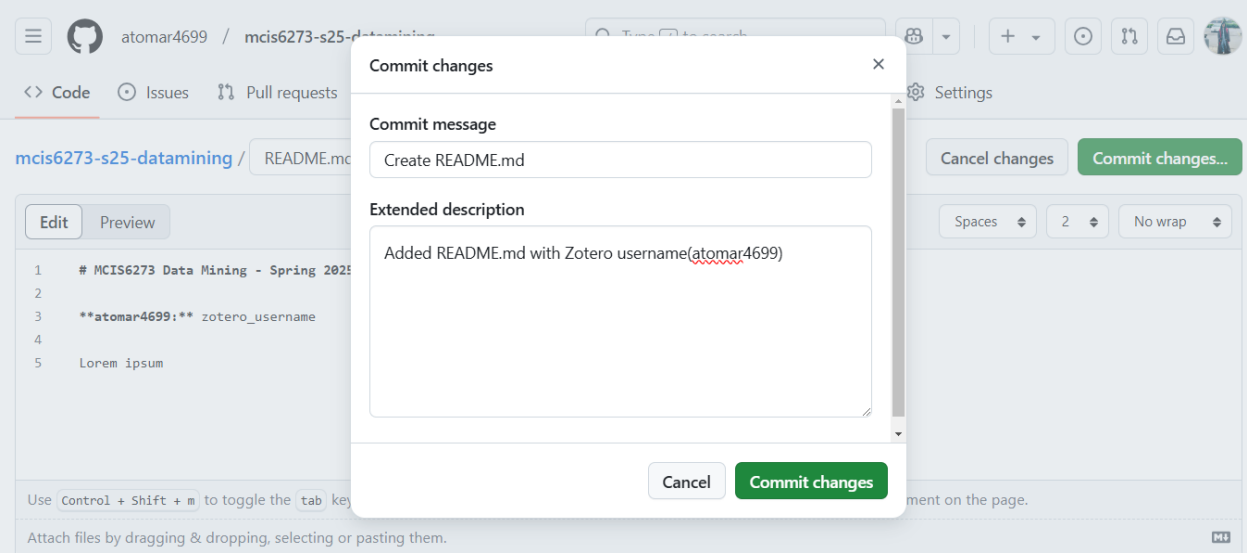
- Any additional text like (e.g., "Lorem Ipsum" placeholder text).



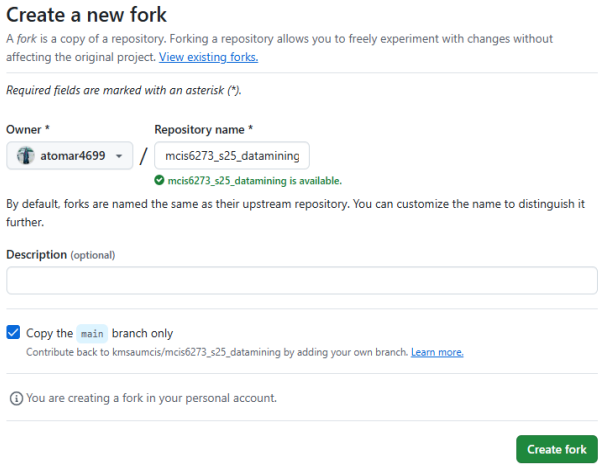


4. **Commit the File**

- Scroll down and add a commit message (e.g., "Added README.md with Zotero username").
- Click **"Commit new file"**.



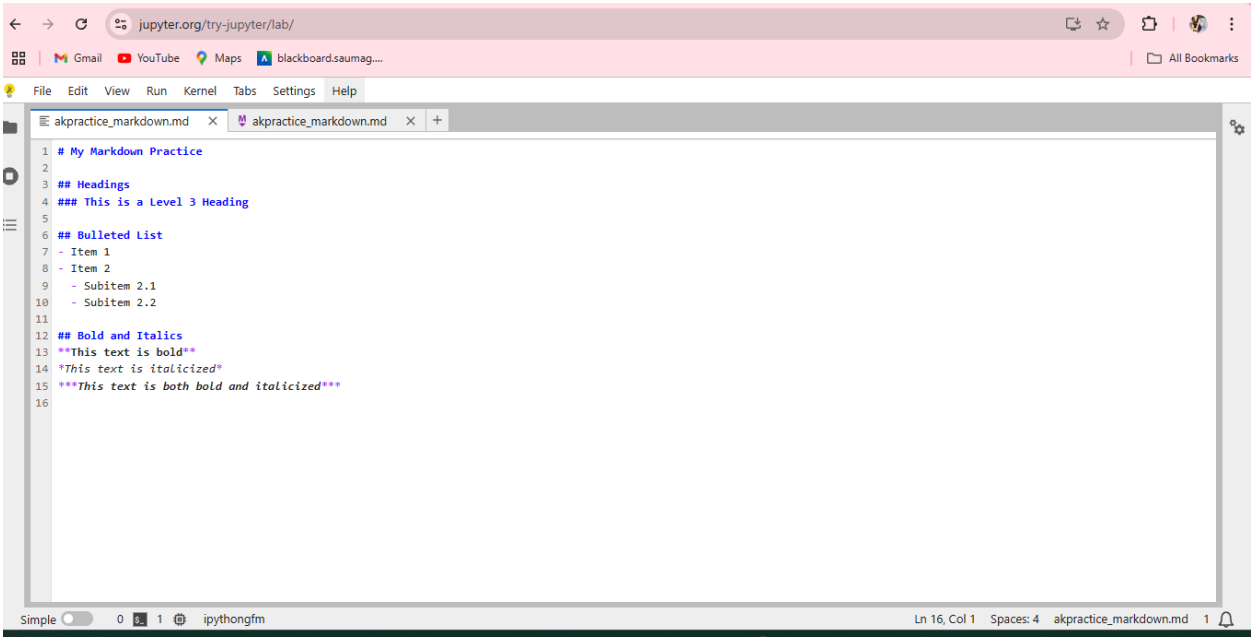
5. **Fork the Course Repository**
    - Visit [MCIS6273 Course Repository].
    - Click the **Fork** button in the top right corner.



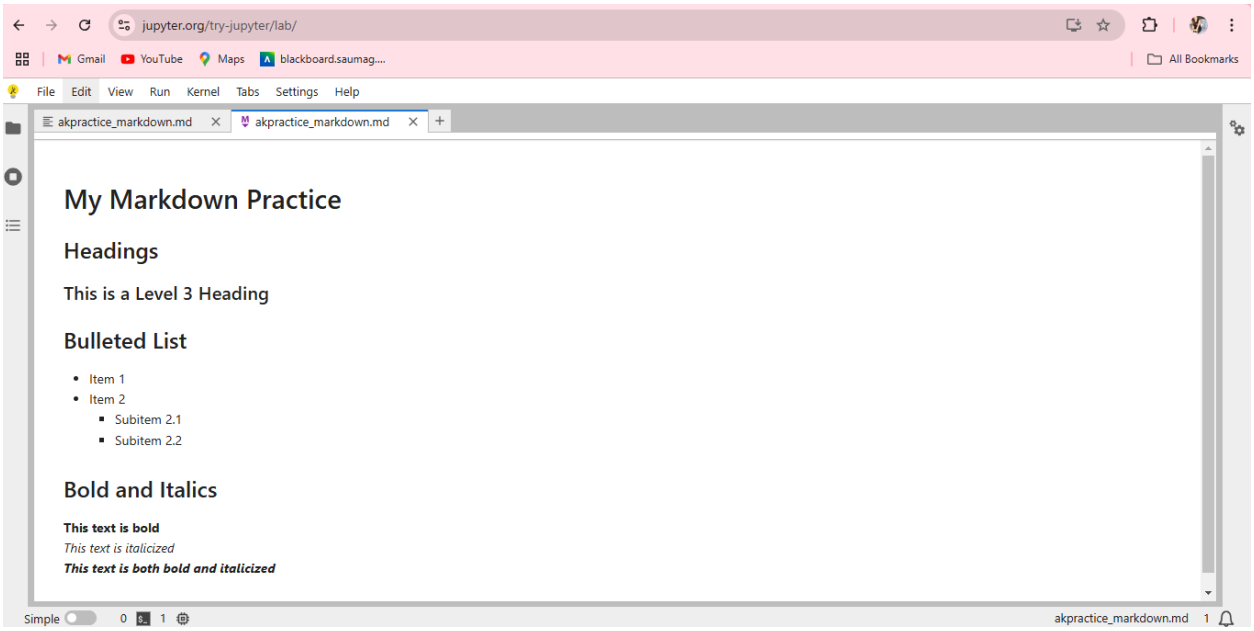
Step 2: Get Familiar with JupyterLab

1. Access the JupyterLab environment set up for the course.
2. **Option 1: Using a .md File in JupyterLab**
3. **Open JupyterLab** (from your course setup).
4. **Create a new text file**:
5. Click **File > New > Text File**.

6. Create a Markdown document and practice:
    - **Headings** (# Heading)
    - **Bullets** (- Item 1)
    - **Bold/Italics** (**bold** or *italics*)
7. Click **File > Save As** and name it **akpractice_markdown.md**.
8. **Add Markdown**



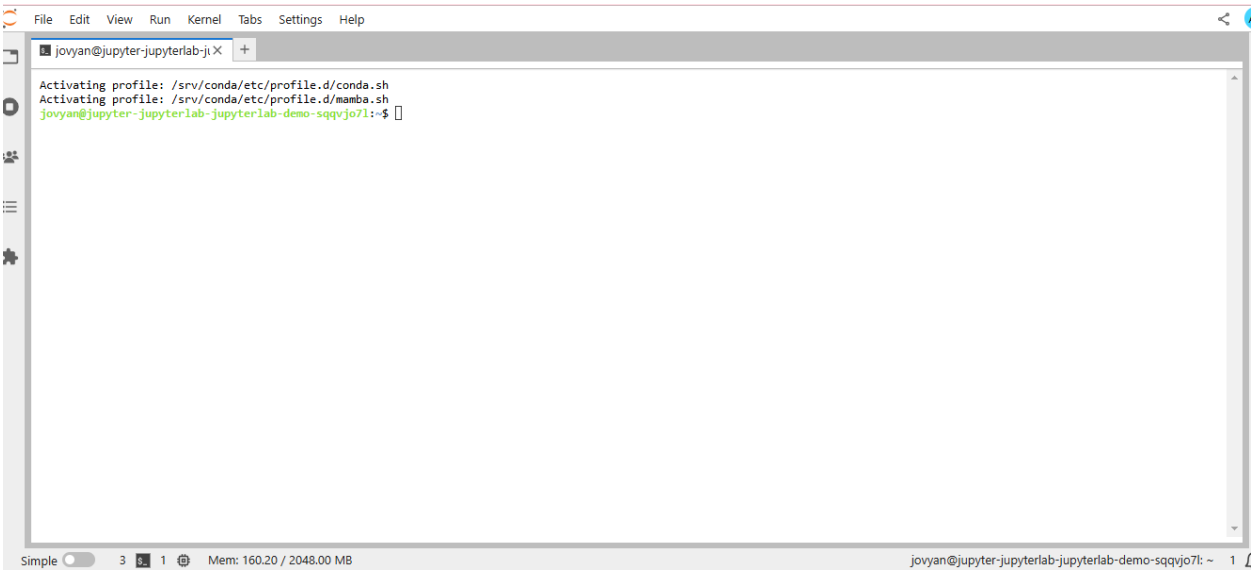
Open in a Markdown Viewer: Right-click the file and **open with "Markdown Preview**



Step 3 Clone Your GitHub Repository
Open a JupyterLab terminal and run:

- **Step 1: Open the JupyterLab Terminal**

1. **Log in to your JupyterLab environment.**
2. Click **File > New > Terminal** to open a Linux command-line console.



- **Step 2: Clone Your GitHub Repository**

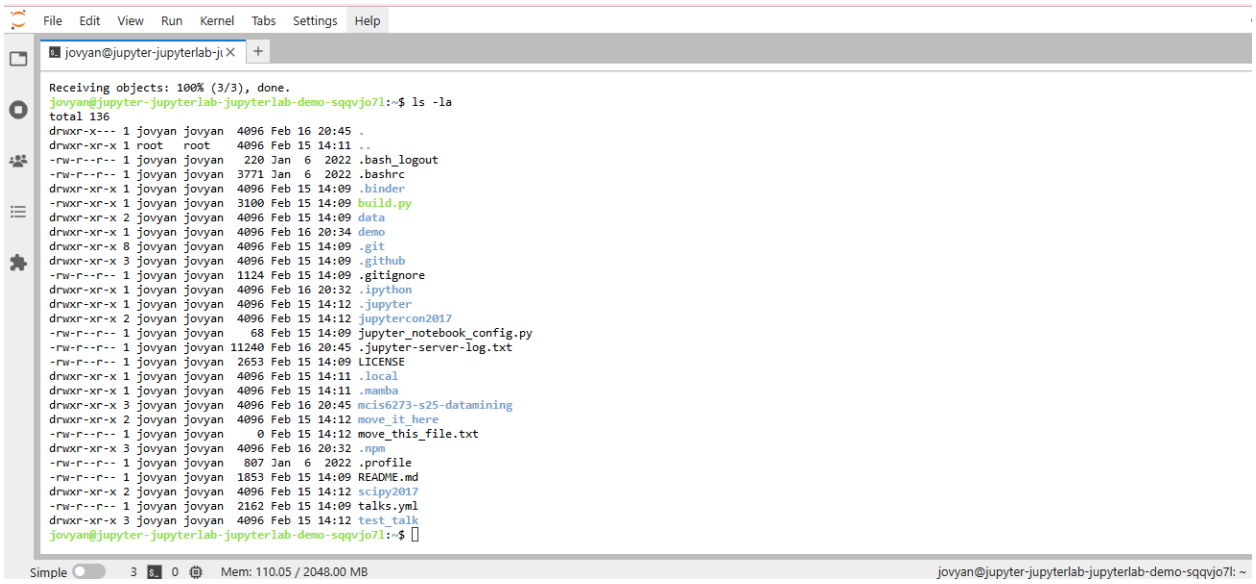
Run the following command in the

>>git clone https://github.com/atomar4699/mcis6273-s25-datamining

This will create a local copy of your repository.

To verify the repository was cloned successfully, list the files:

>>**ls -la**



- **Step 3: Navigate to the Cloned Repository**

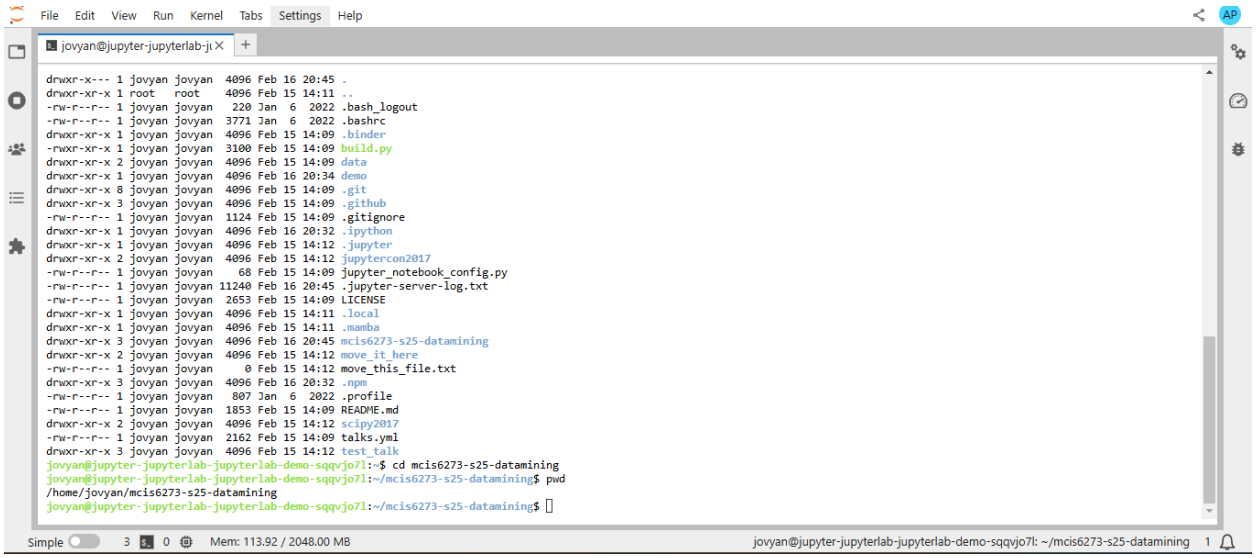
Change into the cloned repository directory:

>>**cd mcis6273-s25-datamining**

Check the current directory:

>>**pwd**

It will show a path ending with /mcis6273-s25-datamining.



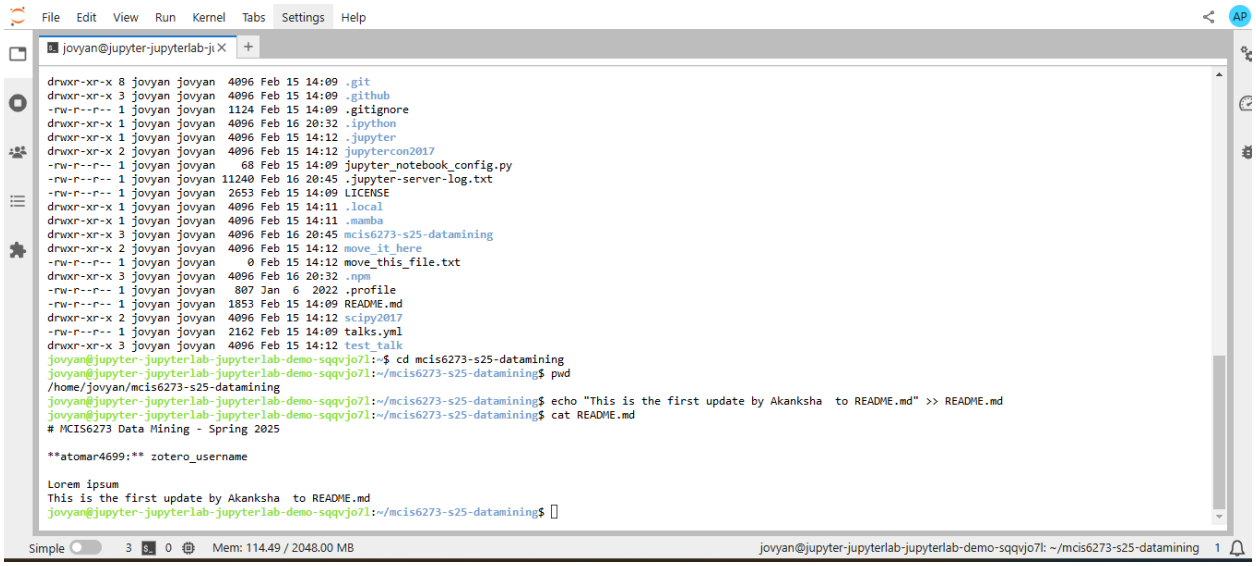
- **Step 4: Modify the README.md File**

Append a new line to the README.md file:

>>echo "This is the first update by Akanksha to README.md" >> README.md

Verify the changes:

>>cat README.md



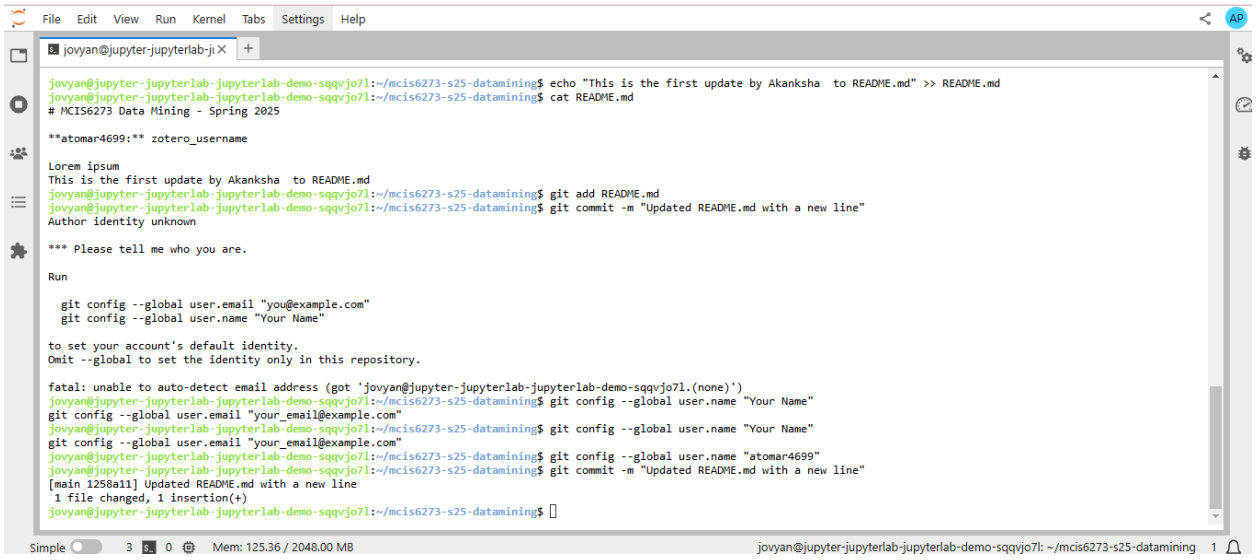
- **Step 5: Add and Commit the Changes**

Stage the modified file:

>>git add README.md

Commit the changes:

git commit -m "Updated README.md with a new line"



Performing Basic Data Engineering in Python Using Chocolate Review Data

1. Check Your Current Directory

>>pwd

This will show your current directory.

If you are not in the correct location, navigate to your home directory first:

>>cd ~

Then run pwd again to confirm.

2. Create the Required Directories

If the homework/hw0 directory does not exist, create it:

>>mkdir -p homework/hw0/data

3. Navigate to the Directory

Now, enter the hw0 directory:

>>cd homework/hw0

Then go into the data folder:

>>cd data

4. Verify Directory Structure

>>pwd

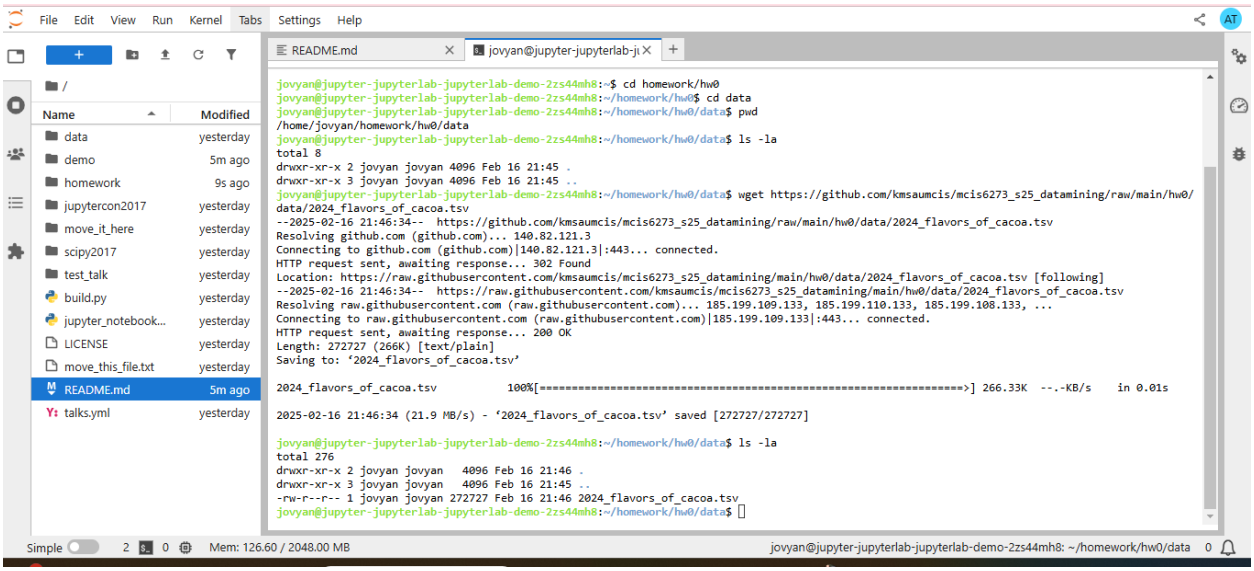
It should return:

>>/home/jovyan/homework/hw0/data

Then, check the contents:

>>ls -la

6. Download the Chocolate Data File



Now, run:

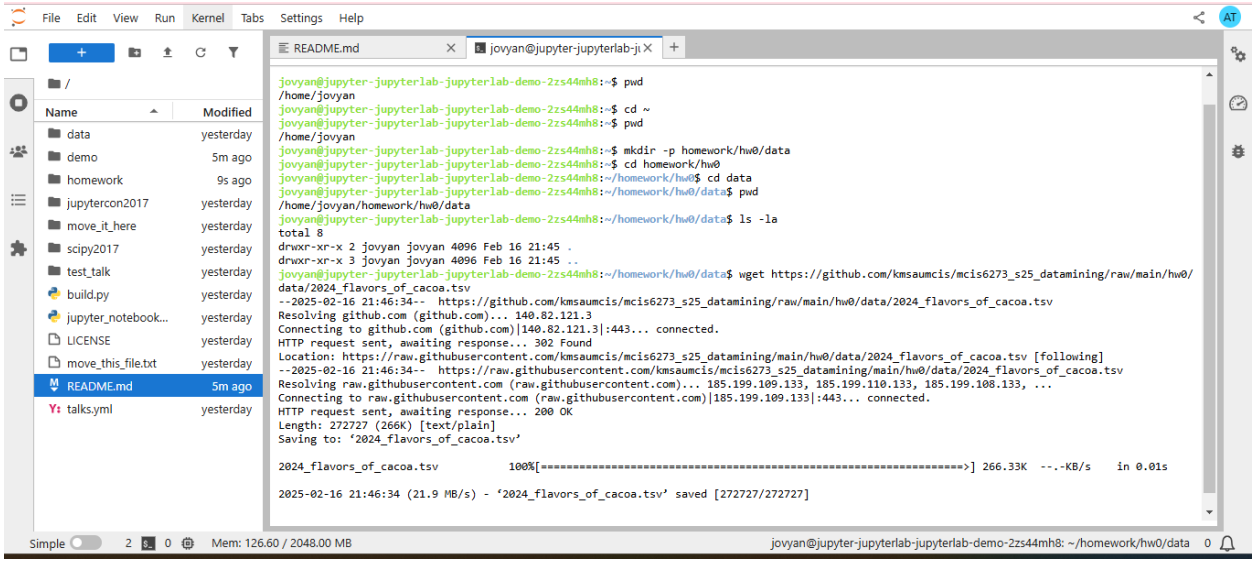
wget https://github.com/kmsaumcis/mcis6273_s25_datamining/raw/main/hw0/data/2024_flavors_of_cacoa.tsv

6. Confirm the Download

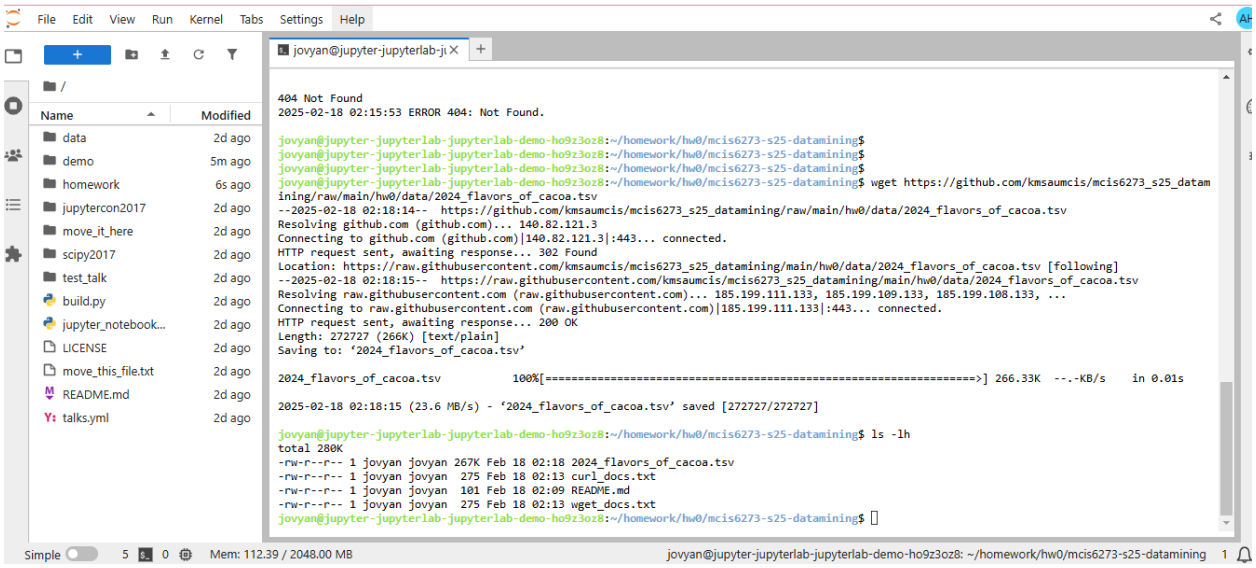
ls -la

You should see:

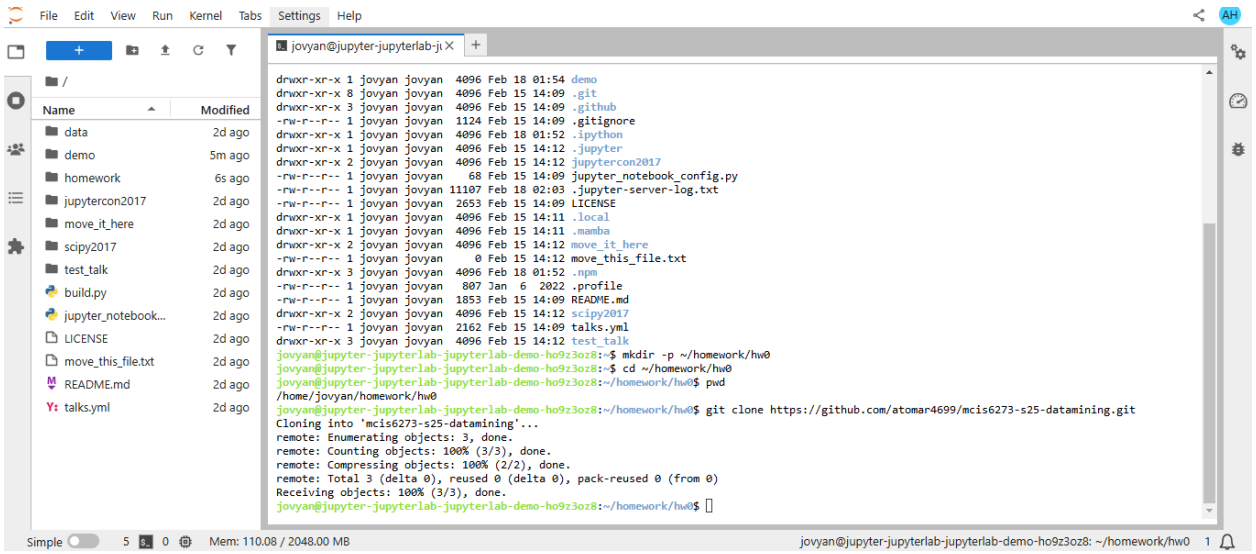
2024_flavors_of_cacoa.tsv



7. **Verify the File:**



Clone Your Repository



Commit and Push the Changes

1. **Check Git User Configuration** (If prompted for identity, set your details):

>>git config --global user.email "atomar4699@muleriders.saumag.edu"

>>git config --global user.name "atomar4699"

2. **Stage the changes**:

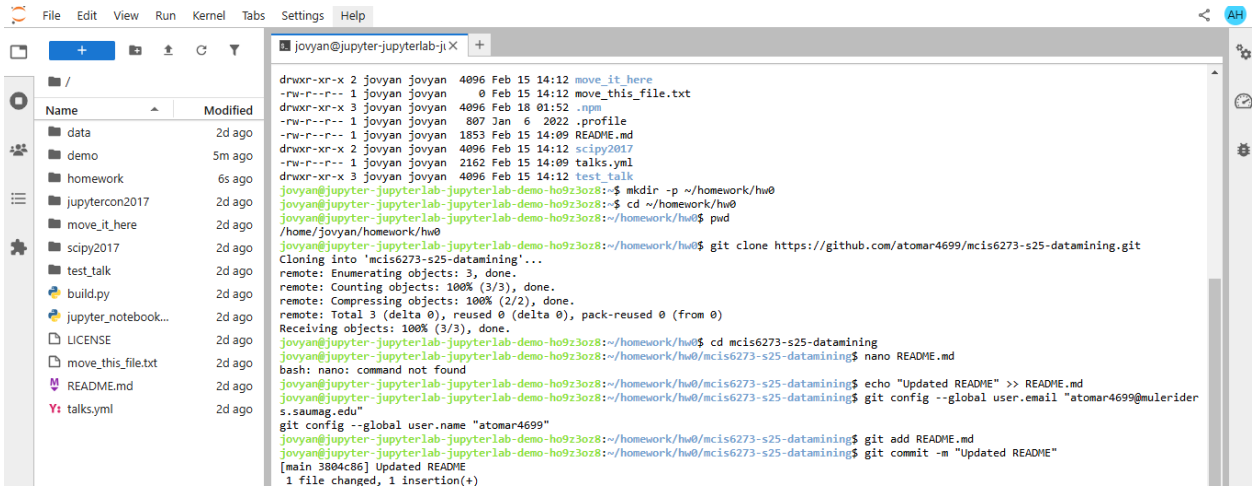
>>git add README.md

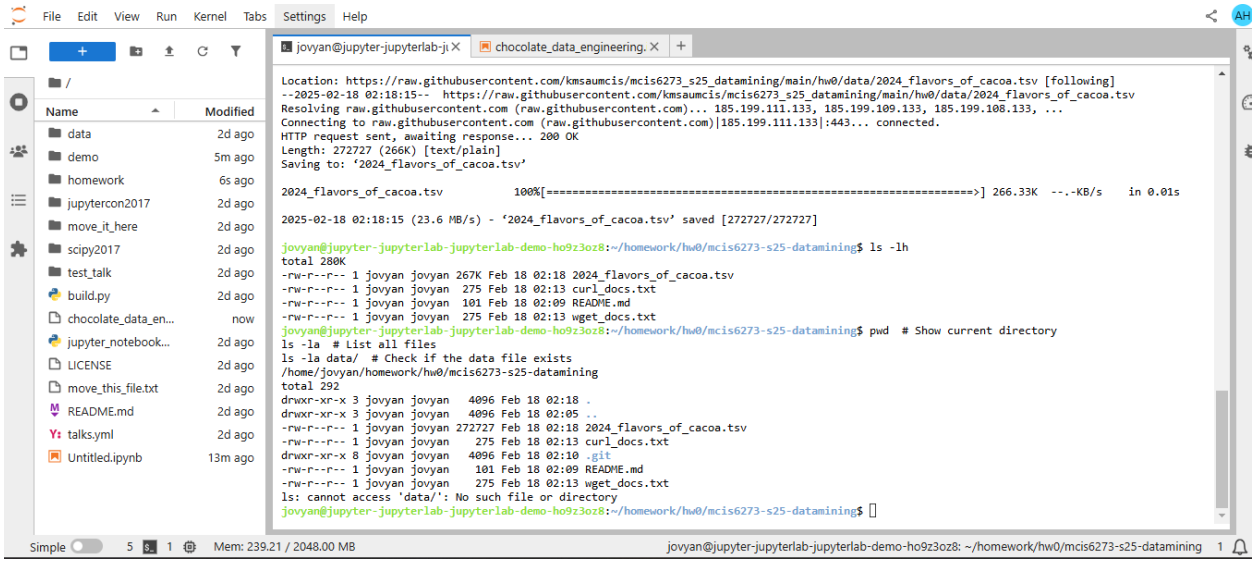
3. **Commit the changes**:

>>git commit -m "Updated README"

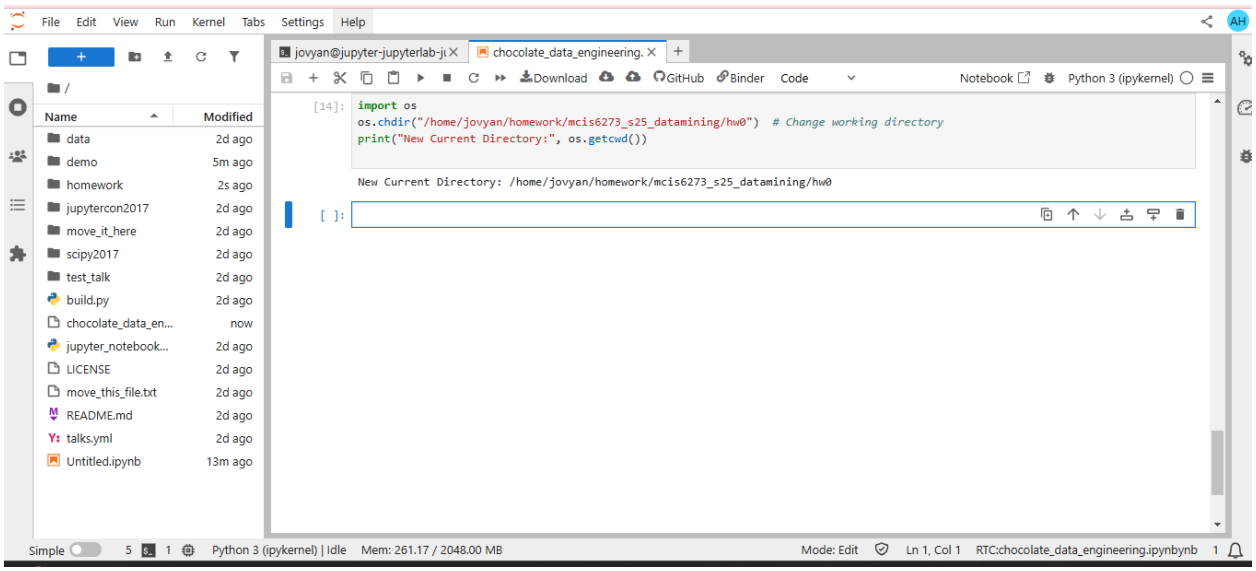
4. **Push the changes to GitHub:**

>>git push origin main

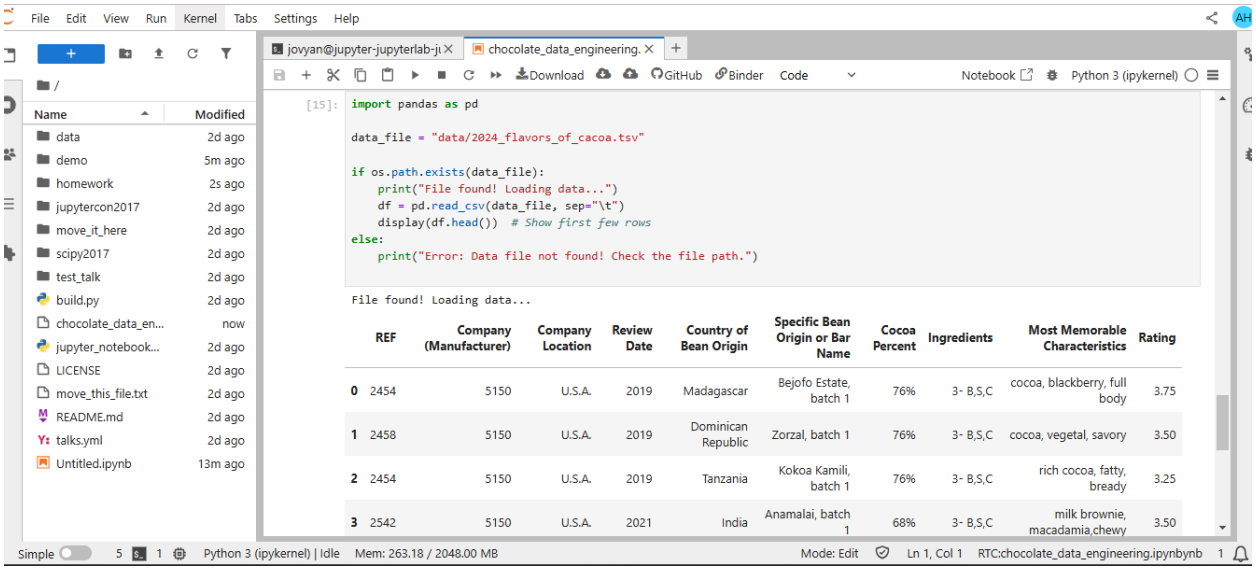


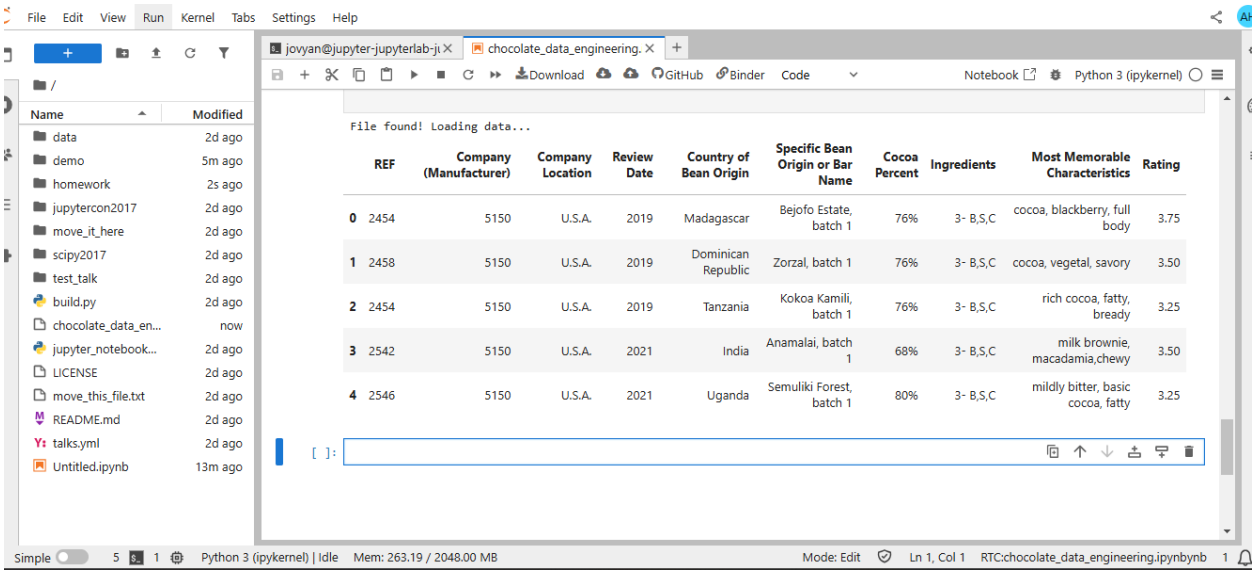


Check Your Current Working Directory



Now run your data loading code:

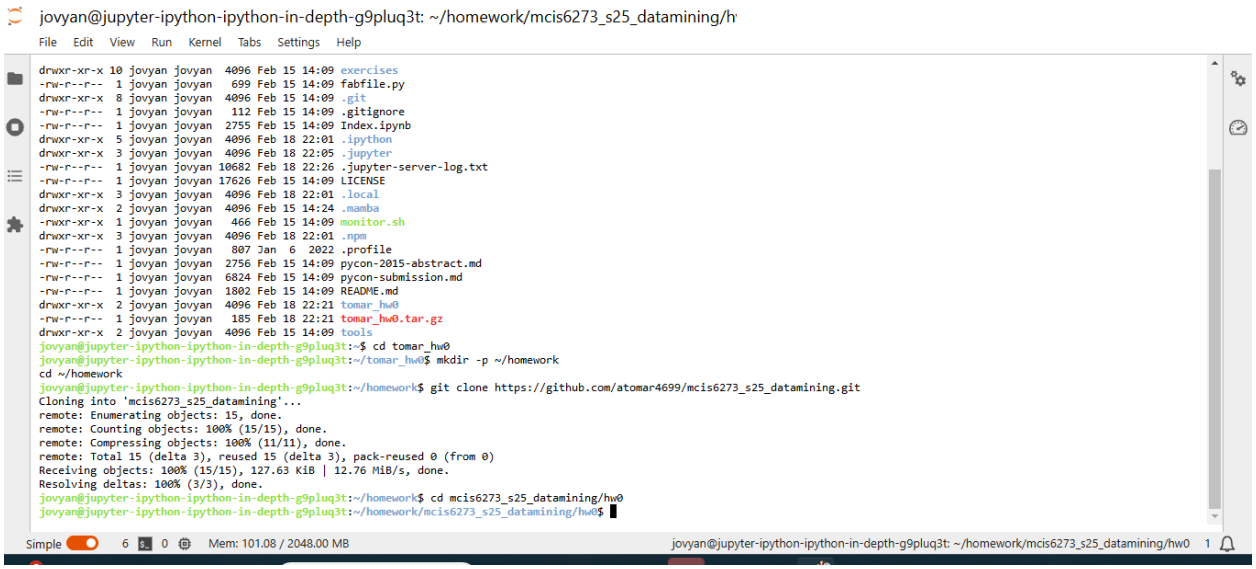




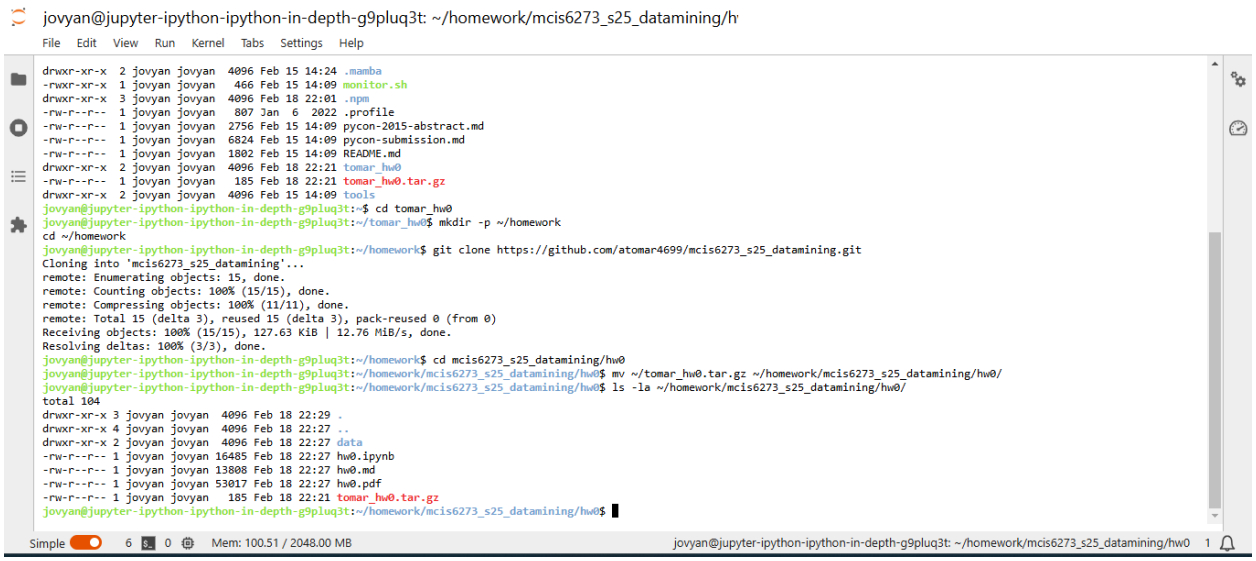
	REF	Company (Manufacturer)	Company Location	Review Date	Country of Bean Origin	Specific Bean Origin or Bar Name	Cocoa Percent	Ingredients	Most Memorable Characteristics	Rating
0	2454	5150	U.S.A.	2019	Madagascar	Bejofo Estate, batch 1	76%	3- B,S,C	cocoa, blackberry, full body	3.75
1	2458	5150	U.S.A.	2019	Dominican Republic	Zorzal, batch 1	76%	3- B,S,C	cocoa, vegetal, savory	3.50
2	2454	5150	U.S.A.	2019	Tanzania	Kokoa Kamili, batch 1	76%	3- B,S,C	rich cocoa, fatty, bready	3.25
3	2542	5150	U.S.A.	2021	India	Anamalai, batch 1	68%	3- B,S,C	milk brownie, macadamia,chewy	3.50
4	2546	5150	U.S.A.	2021	Uganda	Semuliki Forest, batch 1	80%	3- B,S,C	mildly bitter, basic cocoa, fatty	3.25

[ ]:

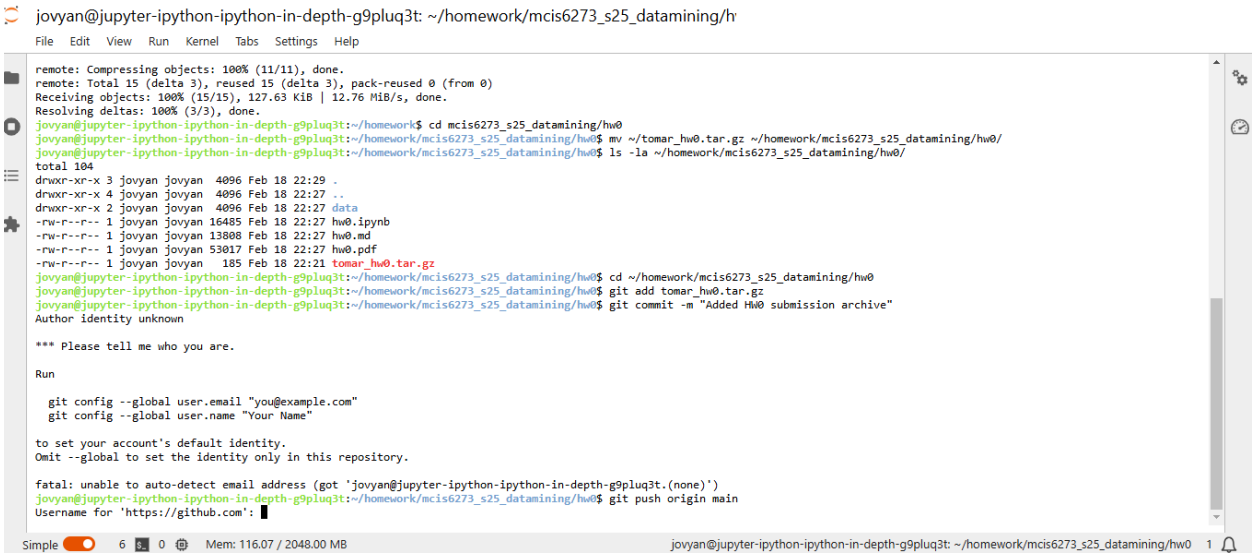
Step 1: Clone the GitHub Repository



Step 2: Move the ZIP/TAR File to the Repository

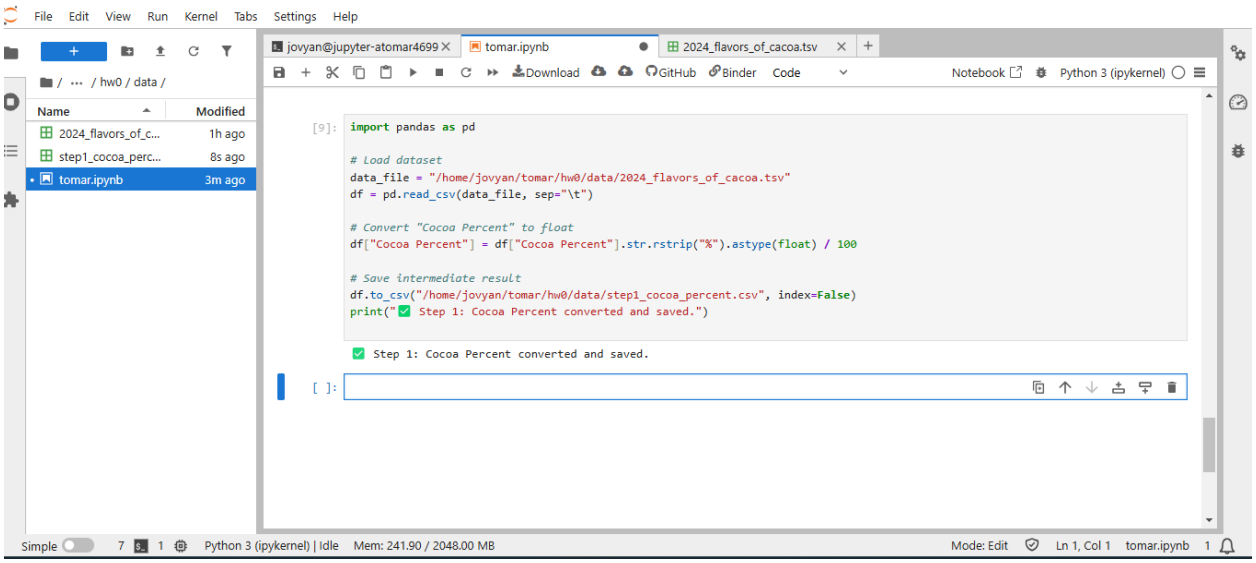


Step 3: Upload the File to GitHub

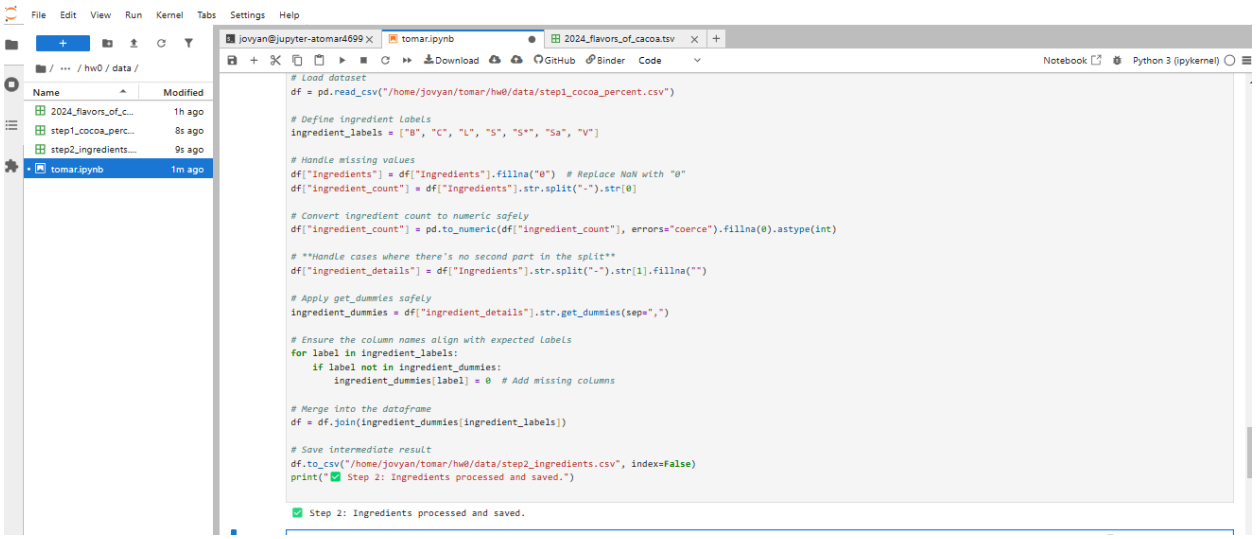


Data Transformation:

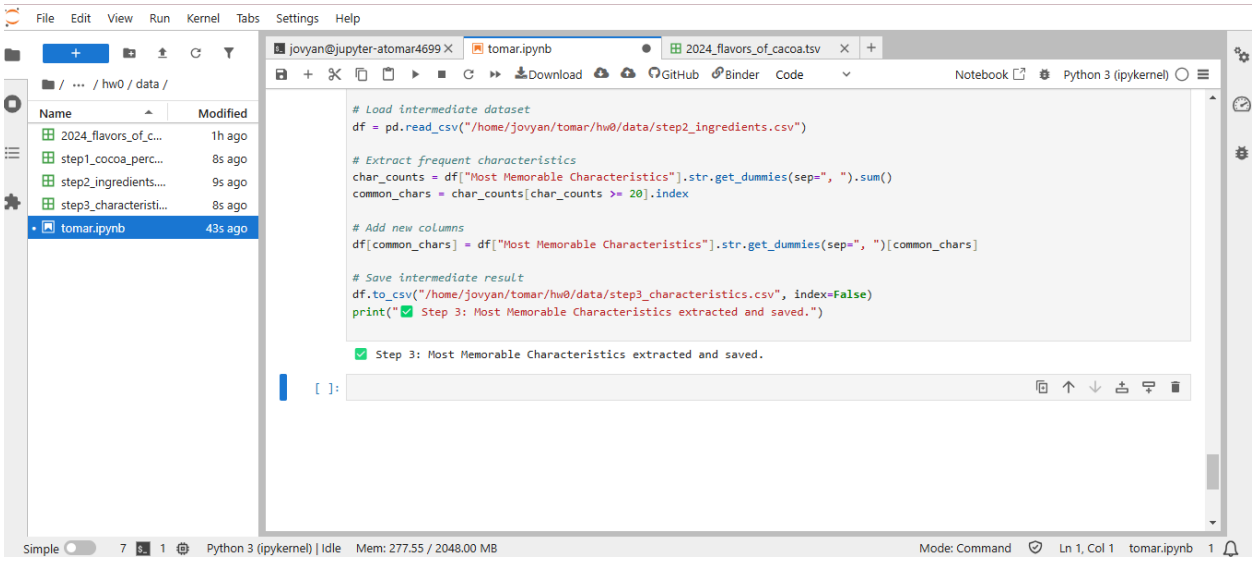
1 Convert "Cacao Percent" to Float



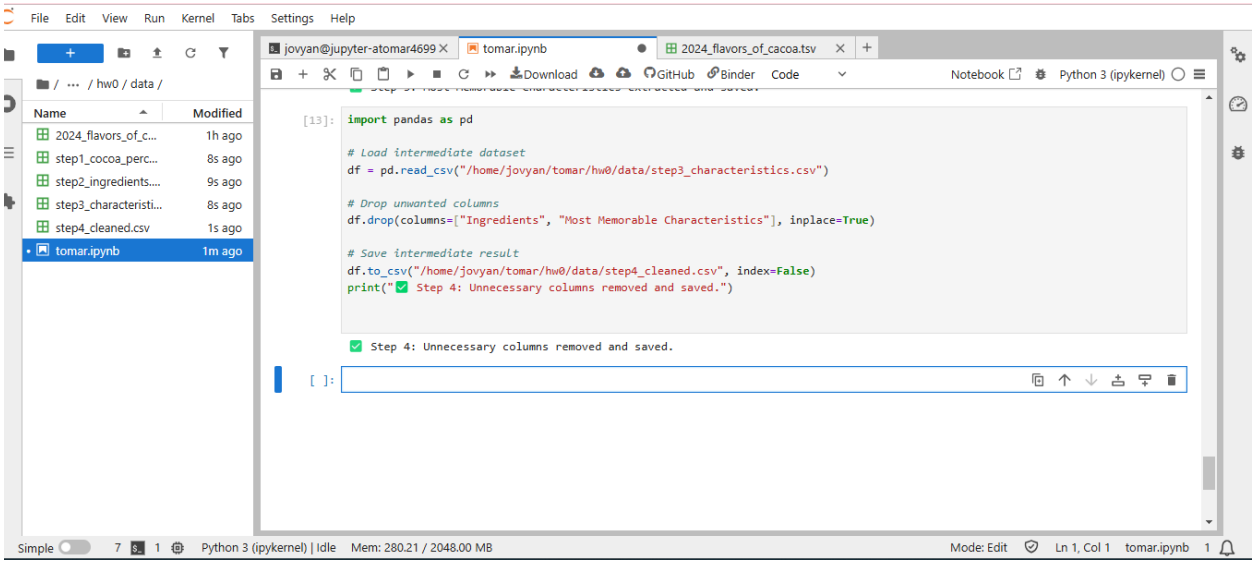
2 Split "Ingredients" into Columns



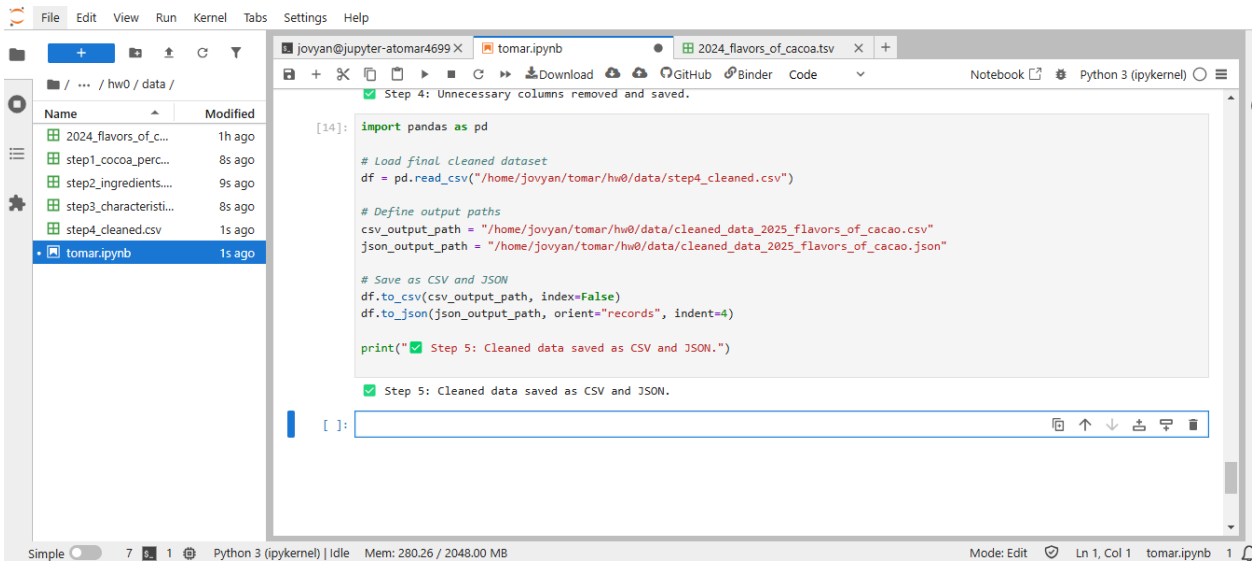
3 Extract "Most Memorable Characteristics"



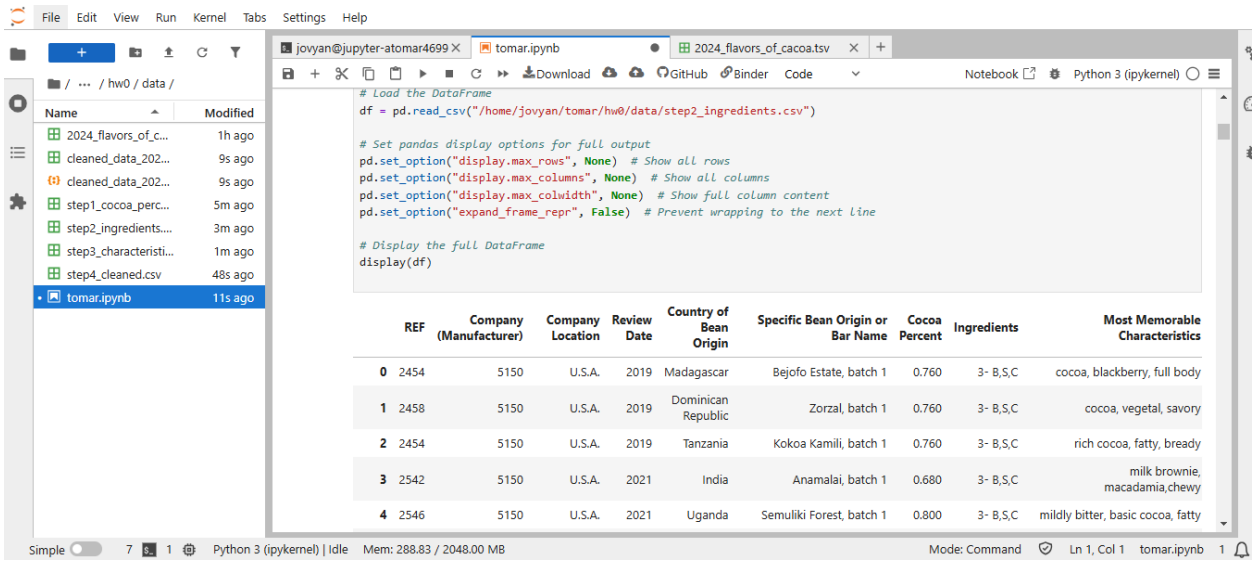
4 Remove Unnecessary Columns



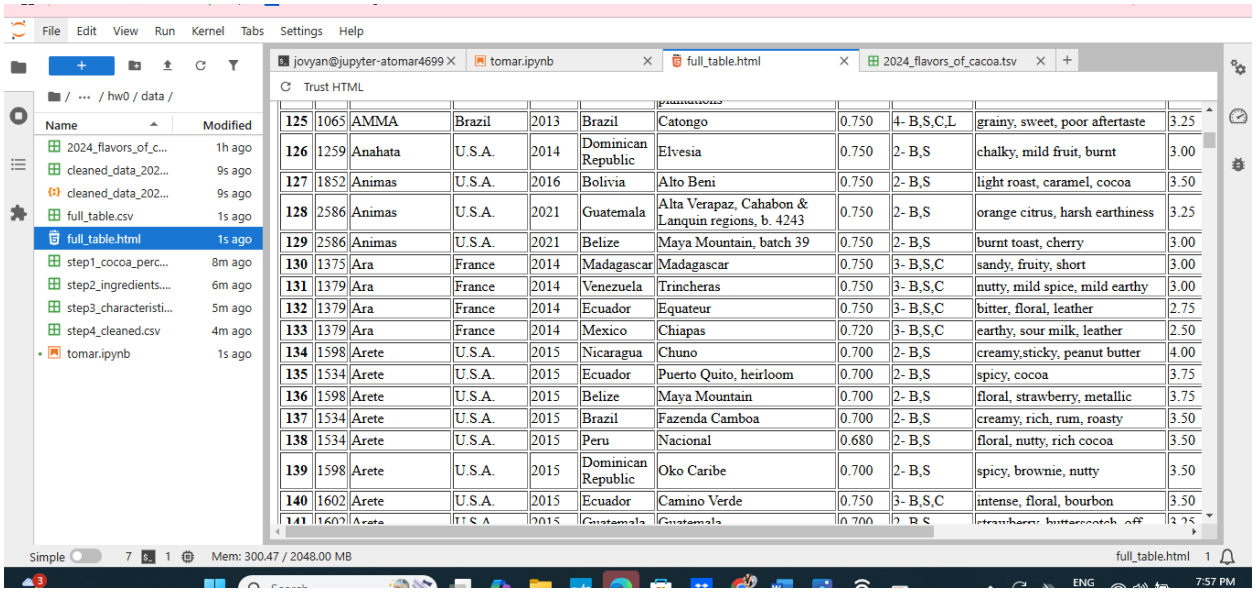
5 Save Cleaned Data



To display the complete DataFrame without truncation in Jupyter Notebook,



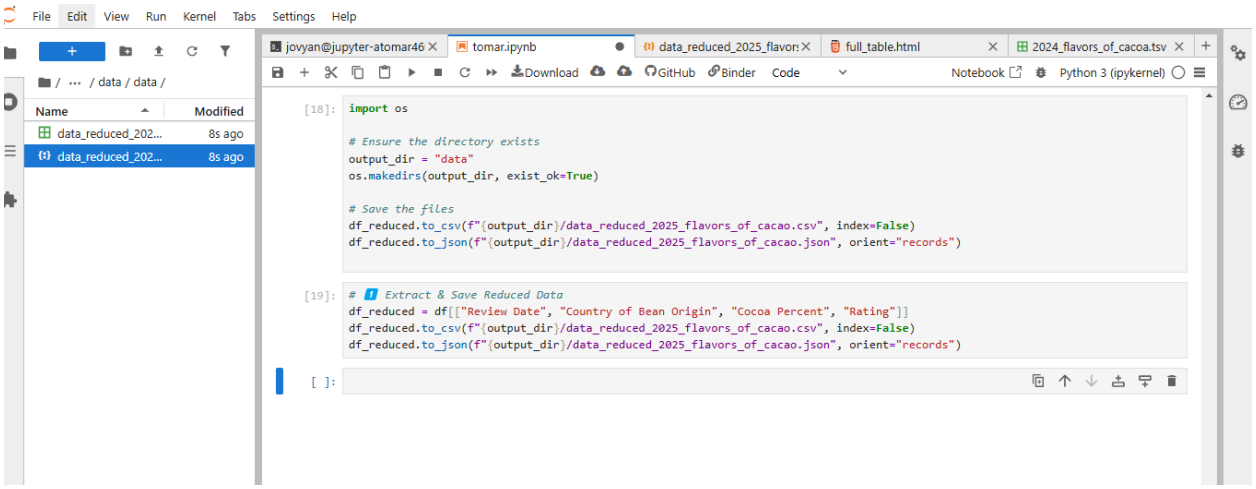
	REF	Company (Manufacturer)	Company Location	Review Date	Country of Bean Origin	Specific Bean Origin or Bar Name	Cocoa Percent	Ingredients	Most Memorable Characteristics
0	2454	5150	U.S.A.	2019	Madagascar	Bejofo Estate, batch 1	0.760	3- B,S,C	cocoa, blackberry, full body
1	2458	5150	U.S.A.	2019	Dominican Republic	Zorzal, batch 1	0.760	3- B,S,C	cocoa, vegetal, savory
2	2454	5150	U.S.A.	2019	Tanzania	Kokoa Kamili, batch 1	0.760	3- B,S,C	rich cocoa, fatty, bready
3	2542	5150	U.S.A.	2021	India	Anamalai, batch 1	0.680	3- B,S,C	milk brownie, macadamia,chewy
4	2546	5150	U.S.A.	2021	Uganda	Semuliki Forest, batch 1	0.800	3- B,S,C	mildly bitter, basic cocoa, fatty

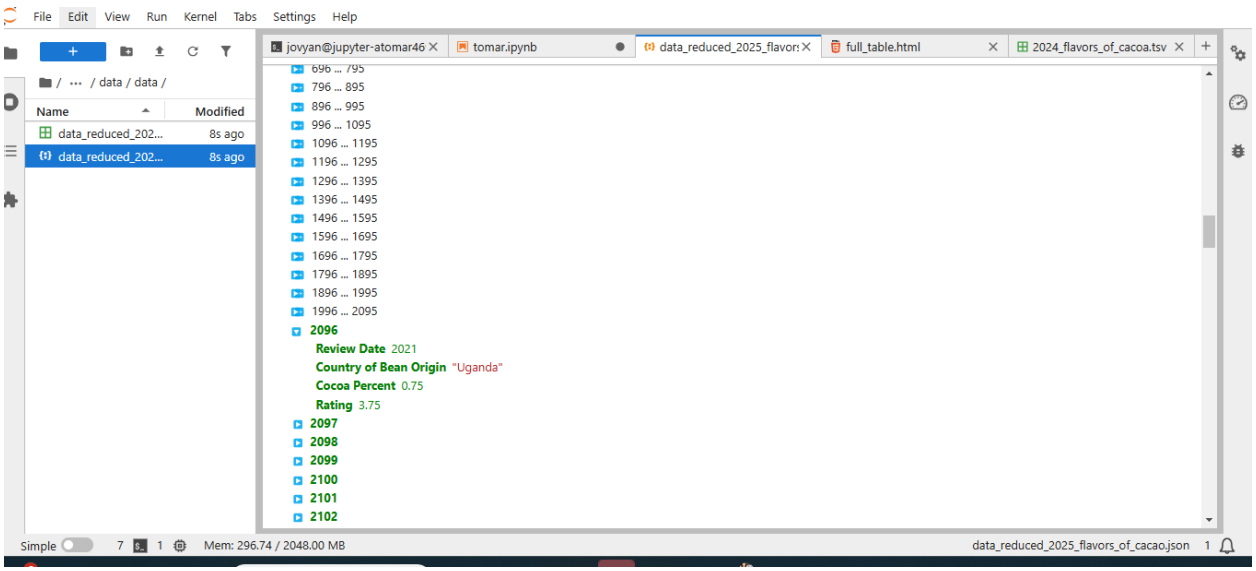


125	1065	AMMA	Brazil	2013	Brazil	Catongo	0.750	4- B,S,C,L	grainy, sweet, poor aftertaste	3.25
126	1259	Anahata	U.S.A.	2014	Dominican Republic	Elvesia	0.750	2- B,S	chalky, mild fruit, burnt	3.00
127	1852	Animas	U.S.A.	2016	Bolivia	Alto Beni	0.750	2- B,S	light roast, caramel, cocoa	3.50
128	2586	Animas	U.S.A.	2021	Guatemala	Alta Verapaz, Cahabon & Lanquin regions, b. 4243	0.750	2- B,S	orange citrus, harsh earthiness	3.25
129	2586	Animas	U.S.A.	2021	Belize	Maya Mountain, batch 39	0.750	2- B,S	burnt toast, cherry	3.00
130	1375	Ara	France	2014	Madagascar	Madagascar	0.750	3- B,S,C	sandy, fruity, short	3.00
131	1379	Ara	France	2014	Venezuela	Trincheras	0.750	3- B,S,C	nutty, mild spice, mild earthy	3.00
132	1379	Ara	France	2014	Ecuador	Equateur	0.750	3- B,S,C	bitter, floral, leather	2.75
133	1379	Ara	France	2014	Mexico	Chiapas	0.720	3- B,S,C	earthy, sour milk, leather	2.50
134	1598	Arete	U.S.A.	2015	Nicaragua	Chuno	0.700	2- B,S	creamy,sticky, peanut butter	4.00
135	1534	Arete	U.S.A.	2015	Ecuador	Puerto Quito, heirloom	0.700	2- B,S	spicy, cocoa	3.75
136	1598	Arete	U.S.A.	2015	Belize	Maya Mountain	0.700	2- B,S	floral, strawberry, metallic	3.75
137	1534	Arete	U.S.A.	2015	Brazil	Fazenda Camboa	0.700	2- B,S	creamy, rich, rum, roasty	3.50
138	1534	Arete	U.S.A.	2015	Peru	Nacional	0.680	2- B,S	floral, nutty, rich cocoa	3.50
139	1598	Arete	U.S.A.	2015	Dominican Republic	Oko Caribe	0.700	2- B,S	spicy, brownie, nutty	3.50
140	1602	Arete	U.S.A.	2015	Ecuador	Camino Verde	0.750	3- B,S,C	intense, floral, bourbon	3.50

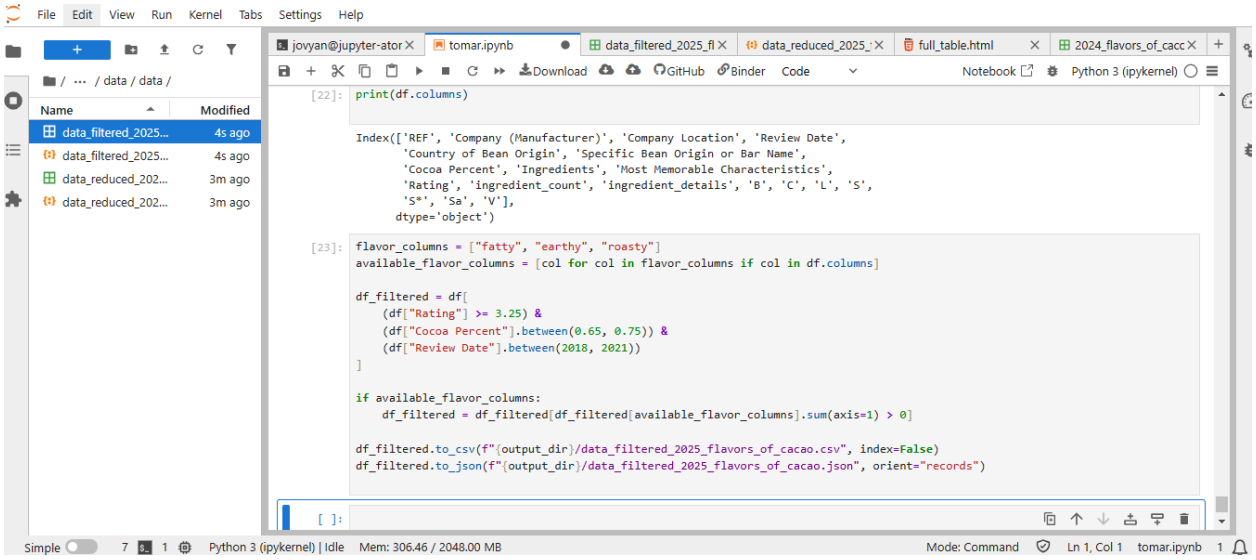
Step 6: Export CSV & JSON

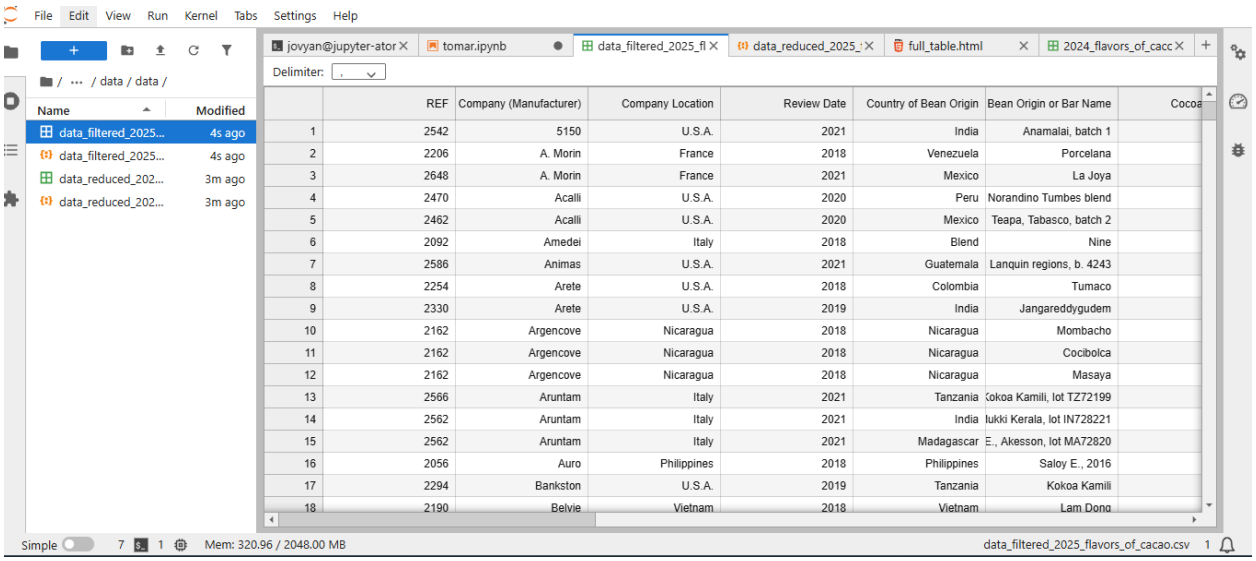
1 Extract & Save Reduced Data





2 Filter & Save Data





3 Create ZIP/TAR File

>> tar -czvf atomar_hw0_files.tar.gz -C ~/tomar/homework hw0