

**Creator: Akanksha Singh Tomar**

**// SAU Id num:9999 03732**

**//Course name: MCIS6273\_030242S- S25 Data Mining**

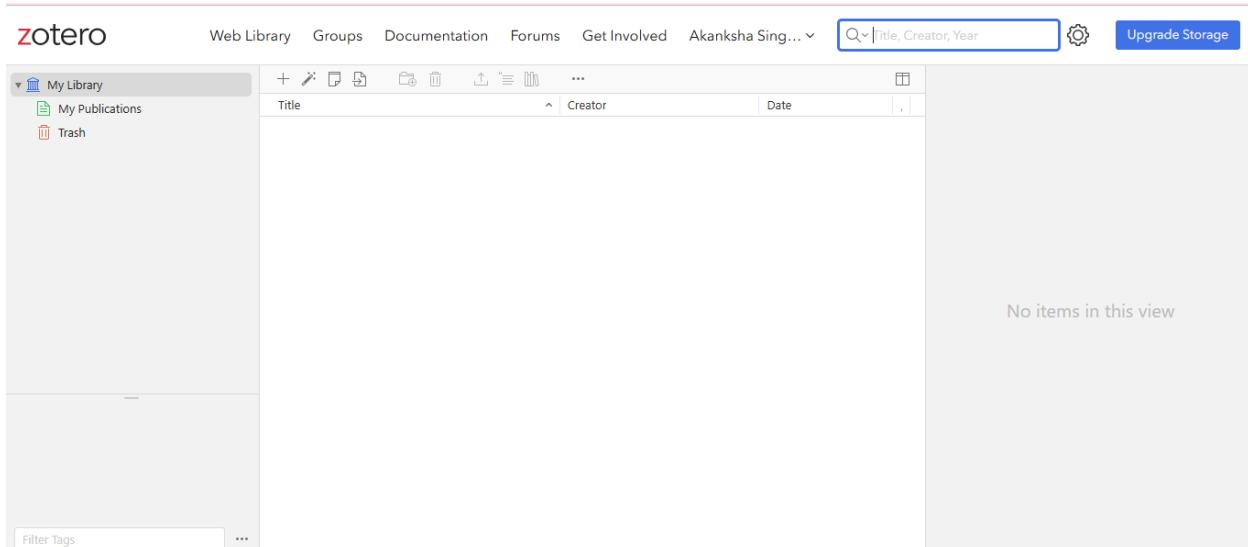
**//Instructor name: Keith Maull**

**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



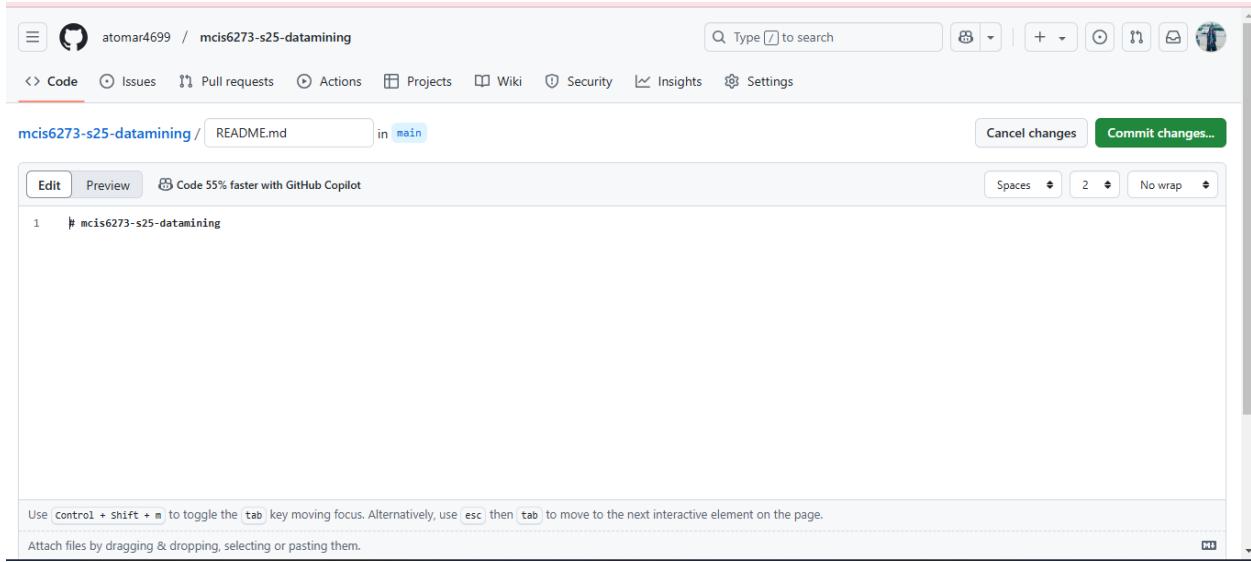
#### 2. Create a GitHub Repository

- 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.

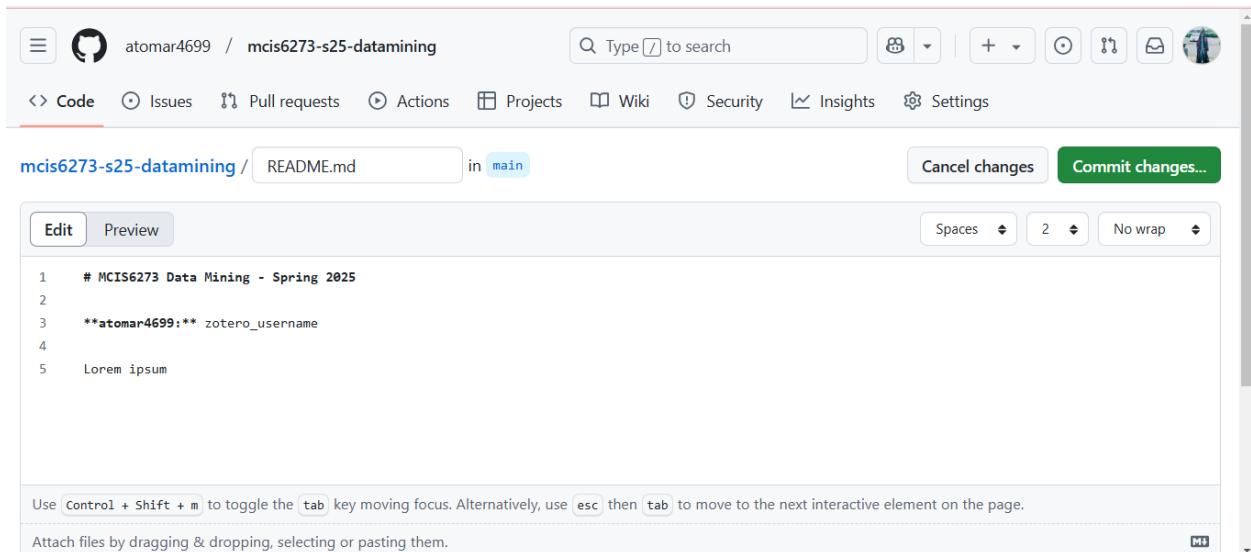
- Any additional text like (e.g., "Lorem ipsum" placeholder text).



The screenshot shows a GitHub repository interface. The top navigation bar includes links for Code, Issues, Pull requests, Actions, Projects, Wiki, Security, Insights, and Settings. The main content area shows a file named README.md in the main branch. The code editor displays the following content:

```
1 # mcis6273-s25-datamining
```

Below the code editor, there is a note: "Use `Control + Shift + m` to toggle the `tab` key moving focus. Alternatively, use `esc` then `tab` to move to the next interactive element on the page." There is also a placeholder text "Attach files by dragging & dropping, selecting or pasting them."



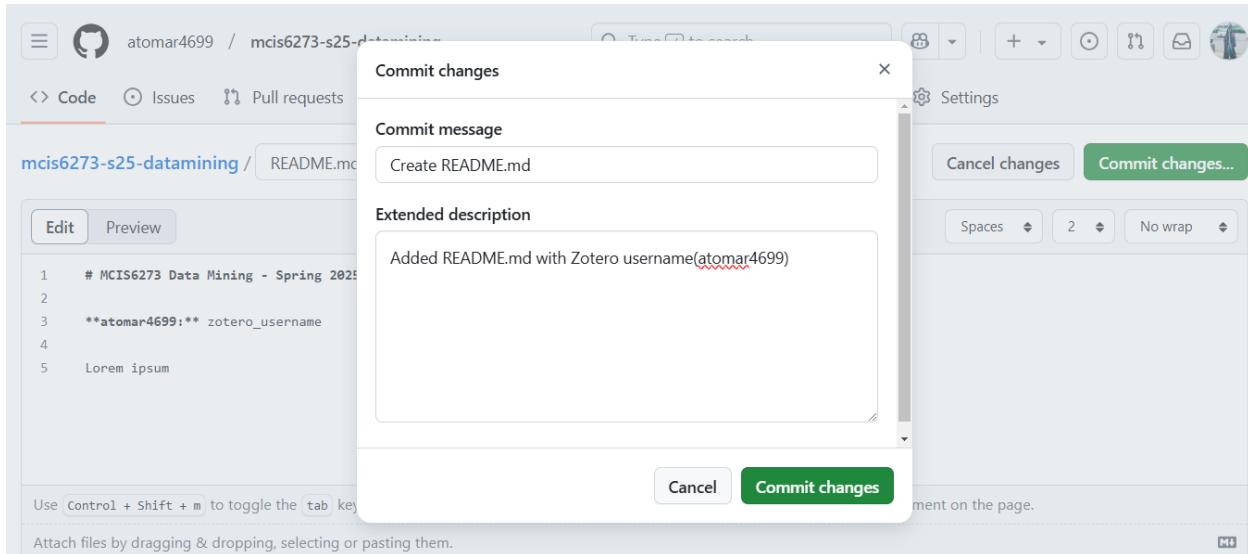
The screenshot shows the same GitHub repository interface. The README.md file now contains the following content:

```
1 # MCIS6273 Data Mining - Spring 2025
2
3 **atomar4699:** zotero_username
4
5 Lorem ipsum
```

Below the code editor, there is a note: "Use `Control + Shift + m` to toggle the `tab` key moving focus. Alternatively, use `esc` then `tab` to move to the next interactive element on the page." There is also a placeholder text "Attach files by dragging & dropping, selecting or pasting them."

#### 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.

**Create a new fork**

A *fork* is a copy of a repository. Forking a repository allows you to freely experiment with changes without affecting the original project. [View existing forks](#).

Required fields are marked with an asterisk (\*).

Owner \*      Repository name \*

 atomar4699 / mcis6273\_s25\_datamining  
 mcis6273\_s25\_datamining is available.

By default, forks are named the same as their upstream repository. You can customize the name to distinguish it further.

Description (optional)

Copy the `main` branch only  
Contribute back to [kmsaucmcis/mcis6273\\_s25\\_datamining](#) by adding your own branch. [Learn more](#).

ⓘ You are creating a fork in your personal account.

**Create fork**

## Step 2: Get Familiar with JupyterLab

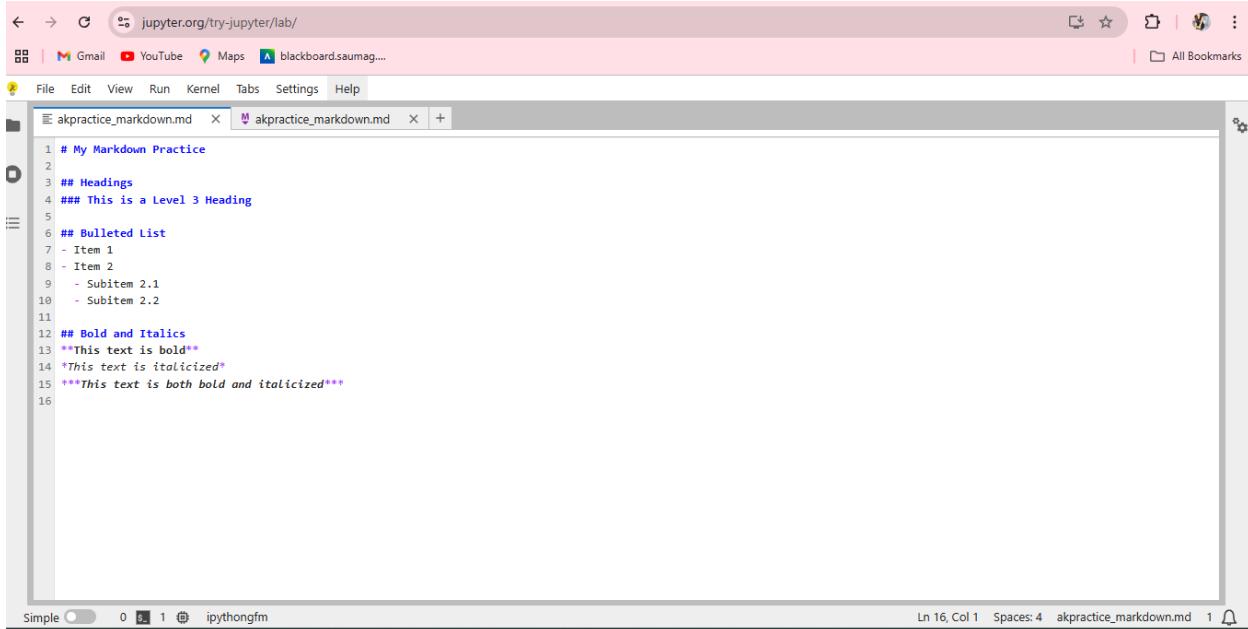
1. Access the JupyterLab environment set up for the course.
2. **Option 1: Using a .md File in JupyterLab**
3. **Open JupyterLab** (from your course setup).
4. **Create a new text file:**
5. Click **File > New > Text File**.

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**.

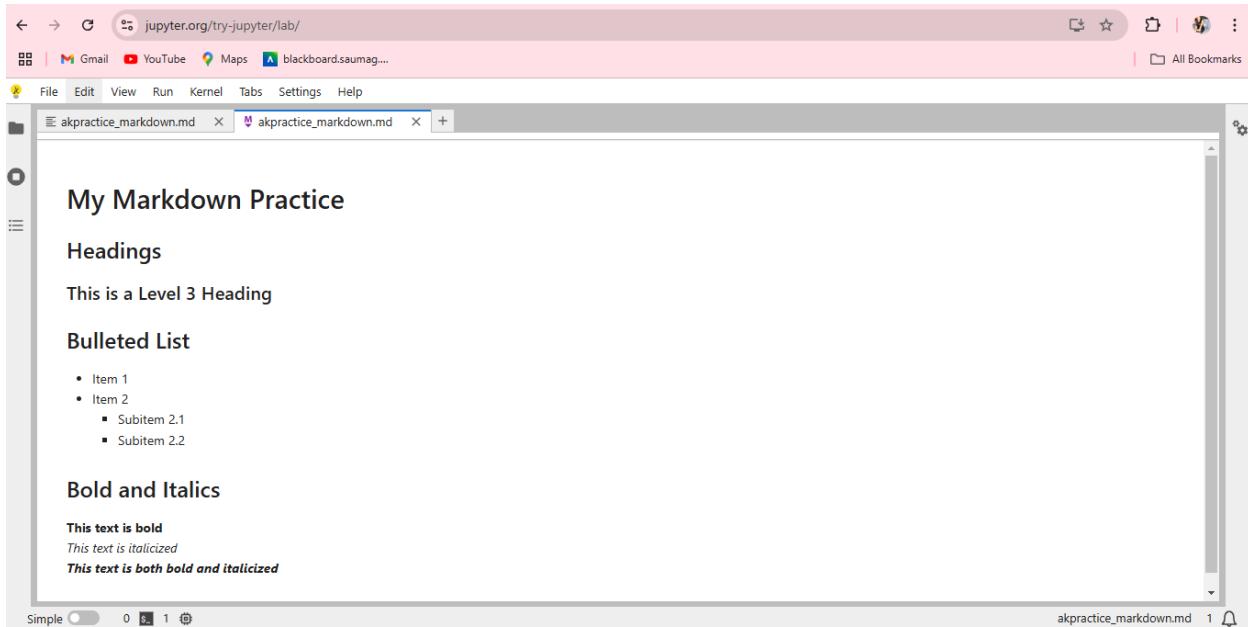
## 8. Add Markdown



The screenshot shows a Jupyter Notebook interface with two tabs open: 'akpractice\_markdown.md' and 'akpractice\_markdown.md'. The left tab contains the following Markdown code:

```
1 # My Markdown Practice
2
3 ## Headings
4 ### This is a Level 3 Heading
5
6 ## Bulleted List
7 - Item 1
8 - Item 2
9   - Subitem 2.1
10  - Subitem 2.2
11
12 ## Bold and Italics
13 **This text is bold**
14 *This text is italicized*
15 ***This text is both bold and italicized***
```

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



The screenshot shows a Jupyter Notebook interface with two tabs open: 'akpractice\_markdown.md' and 'akpractice\_markdown.md'. The left tab contains the following Markdown code, which is rendered in the preview:

```
My Markdown Practice

Headings

This is a Level 3 Heading

Bulleted List

- Item 1
- Item 2
  - Subitem 2.1
  - Subitem 2.2

Bold and Italics



This text is bold



This text is italicized



This text is both bold and italicized

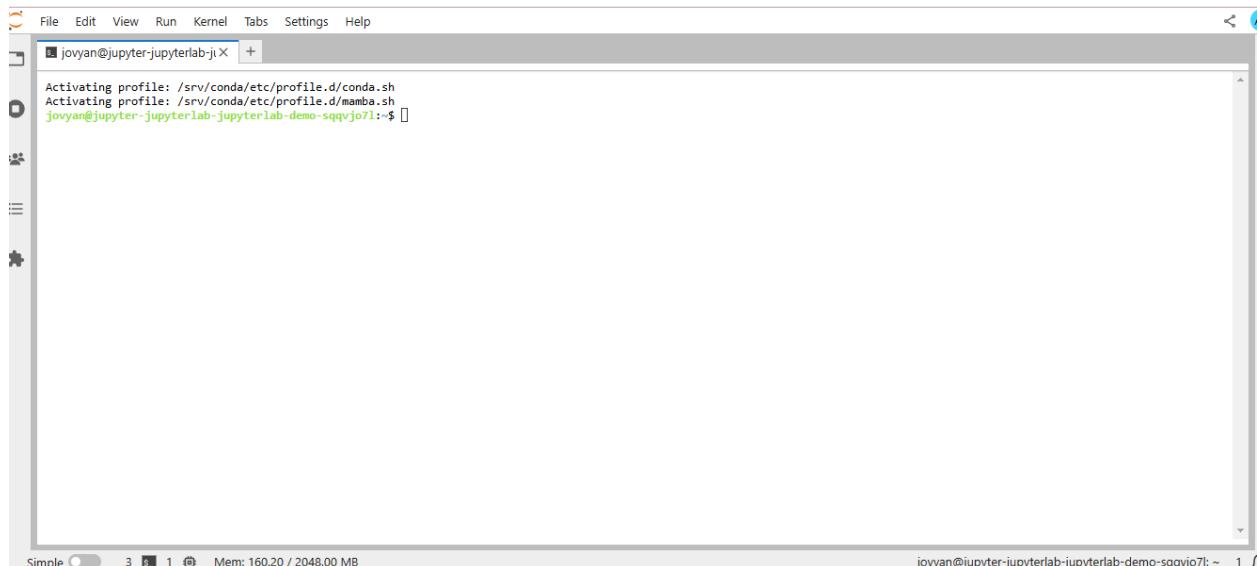

```

### 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.

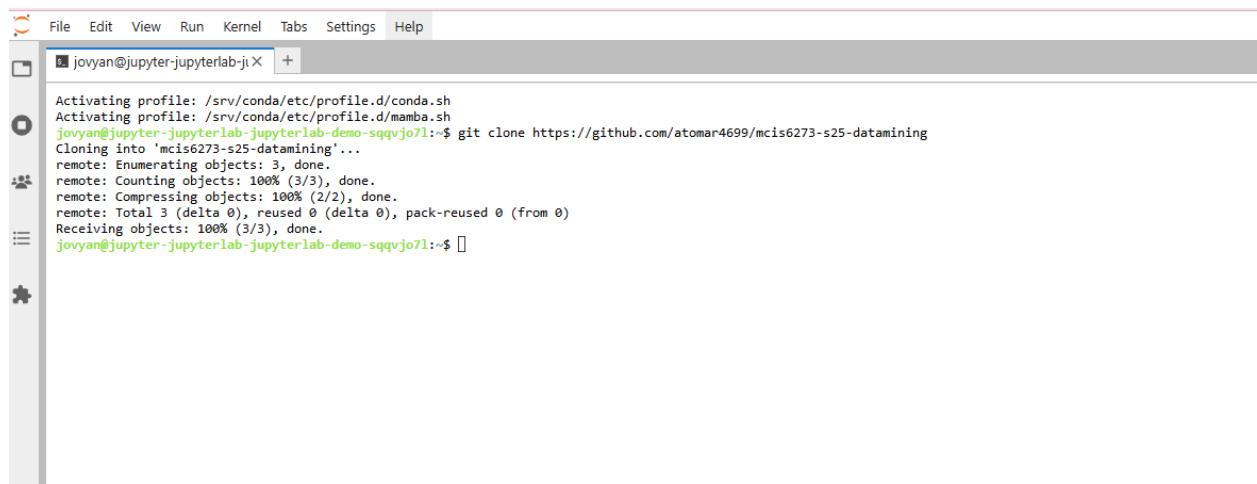


A screenshot of a JupyterLab terminal window. The title bar says "jovyan@jupyter-jupyterlab-jupyterlab-ji X". The terminal pane shows the following text:  
Activating profile: /srv/conda/etc/profile.d/conda.sh  
Activating profile: /srv/conda/etc/profile.d/mamba.sh  
jovyan@jupyter-jupyterlab-jupyterlab-demo-sqqvjo71:~\$  
The bottom status bar shows "Simple" mode, 3 tabs, 1 file, and "Mem: 160.20 / 2048.00 MB".

#### ◆ Step 2: Clone Your GitHub Repository

Run the following command in the

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



A screenshot of a JupyterLab terminal window. The title bar says "jovyan@jupyter-jupyterlab-jupyterlab-ji X". The terminal pane shows the following text:  
Activating profile: /srv/conda/etc/profile.d/conda.sh  
Activating profile: /srv/conda/etc/profile.d/mamba.sh  
jovyan@jupyter-jupyterlab-jupyterlab-demo-sqqvjo71:~\$ git clone https://github.com/atomar4699/mcis6273-s25-datamining  
Cloning into 'mcis6273-s25-datamining'...  
remote: Enumerating objects: 3, done.  
remote: Counting objects: 100% (3/3), done.  
remote: Compressing objects: 100% (2/2), done.  
remote: Total 3 (delta 0), reused 0 (delta 0), pack-reused 0 (from 0)  
Receiving objects: 100% (3/3), done.  
jovyan@jupyter-jupyterlab-jupyterlab-demo-sqqvjo71:~\$  
The bottom status bar shows "Simple" mode, 3 tabs, 1 file, and "Mem: 160.20 / 2048.00 MB".

This will create a local copy of your repository.

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

```
>>ls -la
```

```

Receiving objects: 100% (3/3), done.
jovyan@jupyter-jupyterlab-jupyterlab-demo-sqqvjo71:~$ ls -la
total 136
drwxr-x--- 1 jovyan jovyan 4096 Feb 16 20:45 .
drwxr-xr-x 1 root root 4096 Feb 15 14:11 ..
-rw-r--r-- 1 jovyan jovyan 220 Jan 6 2022 .bash_logout
-rw-r--r-- 1 jovyan jovyan 3771 Jan 6 2022 .bashrc
drwxr-xr-x 1 jovyan jovyan 4096 Feb 15 14:09 .binder
-rw-r--r-- 1 jovyan jovyan 3180 Feb 15 14:09 build.py
drwxr-xr-x 2 jovyan jovyan 4096 Feb 15 14:09 data
drwxr-xr-x 1 jovyan jovyan 4096 Feb 16 20:34 demo
drwxr-xr-x 8 jovyan jovyan 4096 Feb 15 14:09 .git
drwxr-xr-x 3 jovyan jovyan 4096 Feb 15 14:09 .github
-rw-r--r-- 1 jovyan jovyan 1124 Feb 15 14:09 .gitignore
drwxr-xr-x 1 jovyan jovyan 4096 Feb 16 20:32 .ipython
drwxr-xr-x 1 jovyan jovyan 4096 Feb 15 14:12 .jupyter
drwxr-xr-x 2 jovyan jovyan 4096 Feb 15 14:12 jupytercon2017
-rw-r--r-- 1 jovyan jovyan 68 Feb 15 14:09 jupyter_notebook_config.py
-rw-r--r-- 1 jovyan jovyan 11240 Feb 16 20:45 jupyter_server_log.txt
-rw-r--r-- 1 jovyan jovyan 2653 Feb 15 14:09 LICENSE
drwxr-xr-x 1 jovyan jovyan 4096 Feb 15 14:11 .local
drwxr-xr-x 1 jovyan jovyan 4096 Feb 15 14:11 .mamba
drwxr-xr-x 3 jovyan jovyan 4096 Feb 16 20:45 mcis6273-s25-datamining
drwxr-xr-x 2 jovyan jovyan 4096 Feb 15 14:12 move_it_here
-rw-r--r-- 1 jovyan jovyan 0 Feb 15 14:12 move_this_file.txt
drwxr-xr-x 3 jovyan jovyan 4096 Feb 16 20:32 .npm
-rw-r--r-- 1 jovyan jovyan 807 Jan 6 2022 .profile
-rw-r--r-- 1 jovyan jovyan 1853 Feb 15 14:09 README.md
drwxr-xr-x 2 jovyan jovyan 4096 Feb 15 14:12 scipy2017
-rw-r--r-- 1 jovyan jovyan 2162 Feb 15 14:09 talks.yml
drwxr-xr-x 3 jovyan jovyan 4096 Feb 15 14:12 test_talk
jovyan@jupyter-jupyterlab-jupyterlab-demo-sqqvjo71:~$ 

```

Simple 3 0 Mem: 110.05 / 2048.00 MB jovyan@jupyter-jupyterlab-jupyterlab-demo-sqqvjo71: ~

### ◆ 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.

```

drwxr-x--- 1 jovyan jovyan 4096 Feb 16 20:45 .
drwxr-xr-x 1 root root 4096 Feb 15 14:11 ..
-rw-r--r-- 1 jovyan jovyan 220 Jan 6 2022 .bash_logout
-rw-r--r-- 1 jovyan jovyan 3771 Jan 6 2022 .bashrc
drwxr-xr-x 1 jovyan jovyan 4096 Feb 15 14:09 .binder
-rw-r--r-- 1 jovyan jovyan 3180 Feb 15 14:09 build.py
drwxr-xr-x 2 jovyan jovyan 4096 Feb 15 14:09 data
drwxr-xr-x 1 jovyan jovyan 4096 Feb 16 20:34 demo
drwxr-xr-x 8 jovyan jovyan 4096 Feb 15 14:09 .git
drwxr-xr-x 3 jovyan jovyan 4096 Feb 15 14:09 .github
-rw-r--r-- 1 jovyan jovyan 1124 Feb 15 14:09 .gitignore
drwxr-xr-x 1 jovyan jovyan 4096 Feb 16 20:32 .ipython
drwxr-xr-x 1 jovyan jovyan 4096 Feb 15 14:12 .jupyter
drwxr-xr-x 2 jovyan jovyan 4096 Feb 15 14:12 jupytercon2017
-rw-r--r-- 1 jovyan jovyan 68 Feb 15 14:09 jupyter_notebook_config.py
-rw-r--r-- 1 jovyan jovyan 11240 Feb 16 20:45 jupyter_server_log.txt
-rw-r--r-- 1 jovyan jovyan 2653 Feb 15 14:09 LICENSE
drwxr-xr-x 1 jovyan jovyan 4096 Feb 15 14:11 .local
drwxr-xr-x 1 jovyan jovyan 4096 Feb 16 20:45 mcis6273-s25-datamining
drwxr-xr-x 2 jovyan jovyan 4096 Feb 15 14:12 move_it_here
-rw-r--r-- 1 jovyan jovyan 0 Feb 15 14:12 move_this_file.txt
drwxr-xr-x 3 jovyan jovyan 4096 Feb 16 20:32 .npm
-rw-r--r-- 1 jovyan jovyan 807 Jan 6 2022 .profile
-rw-r--r-- 1 jovyan jovyan 1853 Feb 15 14:09 README.md
drwxr-xr-x 2 jovyan jovyan 4096 Feb 15 14:12 scipy2017
-rw-r--r-- 1 jovyan jovyan 2162 Feb 15 14:09 talks.yml
drwxr-xr-x 3 jovyan jovyan 4096 Feb 15 14:12 test_talk
jovyan@jupyter-jupyterlab-jupyterlab-demo-sqqvjo71:~$ cd mcis6273-s25-datamining
jovyan@jupyter-jupyterlab-jupyterlab-demo-sqqvjo71:~/mcis6273-s25-datamining$ pwd
/home/jovyan/mcis6273-s25-datamining
jovyan@jupyter-jupyterlab-jupyterlab-demo-sqqvjo71:~/mcis6273-s25-datamining$ 

```

Simple 3 0 Mem: 113.92 / 2048.00 MB jovyan@jupyter-jupyterlab-jupyterlab-demo-sqqvjo71: ~/mcis6273-s25-datamining 1

### ◆ 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

```
File Edit View Run Kernel Tabs Settings Help
jovyan@jupyter-jupyterlab-ji-X: ~
drwxr-xr-x 8 jovyan jovyan 4096 Feb 15 14:09 .git
drwxr-xr-x 3 jovyan jovyan 4096 Feb 15 14:09 .github
-rw-r--r-- 1 jovyan jovyan 1124 Feb 15 14:09 .gitignore
drwxr-xr-x 1 jovyan jovyan 4096 Feb 16 20:32 .ipython
drwxr-xr-x 1 jovyan jovyan 4096 Feb 15 14:12 .jupyter
drwxr-xr-x 2 jovyan jovyan 4096 Feb 15 14:12 jupytercon2017
-rw-r--r-- 1 jovyan jovyan 68 Feb 15 14:09 jupyter_notebook_config.py
-rw-r--r-- 1 jovyan jovyan 11240 Feb 16 20:45 jupyter-server-log.txt
-rw-r--r-- 1 jovyan jovyan 2653 Feb 15 14:09 LICENSE
drwxr-xr-x 1 jovyan jovyan 4096 Feb 15 14:11 .local
drwxr-xr-x 1 jovyan jovyan 4096 Feb 15 14:11 .mamba
drwxr-xr-x 3 jovyan jovyan 4096 Feb 16 20:45 mcis6273-s25-datamining
drwxr-xr-x 2 jovyan jovyan 4096 Feb 15 14:12 move_it_here
-rw-r--r-- 1 jovyan jovyan 0 Feb 15 14:12 move_this_file.txt
drwxr-xr-x 3 jovyan jovyan 4096 Feb 16 20:32 .npm
-rw-r--r-- 1 jovyan jovyan 807 Jan 6 2022 .profile
-rw-r--r-- 1 jovyan jovyan 1853 Feb 15 14:09 README.md
drwxr-xr-x 2 jovyan jovyan 4096 Feb 15 14:12 scipy2017
-rw-r--r-- 1 jovyan jovyan 2162 Feb 15 14:09 talks.yml
drwxr-xr-x 3 jovyan jovyan 4096 Feb 15 14:12 test_talk
jovyan@jupyter-jupyterlab-jupyterlab-demo-sqvyjo71: ~ cd mcis6273-s25-datamining
jovyan@jupyter-jupyterlab-jupyterlab-demo-sqvyjo71: ~/mcis6273-s25-datamining$ pwd
/home/jovyan/mcis6273-s25-datamining
jovyan@jupyter-jupyterlab-jupyterlab-demo-sqvyjo71:~/mcis6273-s25-datamining$ echo "This is the first update by Akanksha to README.md" >> README.md
jovyan@jupyter-jupyterlab-jupyterlab-demo-sqvyjo71:~/mcis6273-s25-datamining$ cat README.md
# MCIS6273 Data Mining - Spring 2025

**atamar4699:** zotero_username

Lorem ipsum
This is the first update by Akanksha to README.md
jovyan@jupyter-jupyterlab-jupyterlab-demo-sqvyjo71:~/mcis6273-s25-datamining$ [
```

#### ◆ 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"

File Edit View Run Kernel Tabs Settings Help AP

jovyan@jupyter-jupyterlab-jupyterlab-demo-sqqvjo7l:~/mcis6273-s25-datamining$ echo "This is the first update by Akanksha to README.md" >> README.md
jovyan@jupyter-jupyterlab-jupyterlab-demo-sqqvjo7l:~/mcis6273-s25-datamining$ cat README.md
# MCIS6273 Data Mining - Spring 2025

**atomar4699:** zotero_username

Lorem ipsum
This is the first update by Akanksha to README.md
jovyan@jupyter-jupyterlab-jupyterlab-demo-sqqvjo7l:~/mcis6273-s25-datamining$ git add README.md
jovyan@jupyter-jupyterlab-jupyterlab-demo-sqqvjo7l:~/mcis6273-s25-datamining$ git commit -m "Updated README.md with a new line"
Author identity unknown

*** Please tell me who you are.

Run

git config --global user.email "you@example.com"
git config --global user.name "Your Name"

to set your account's default identity.
Omit --global to set the identity only in this repository.

fatal: unable to auto-detect email address (got 'jovyan@jupyter-jupyterlab-jupyterlab-demo-sqqvjo7l.(none)')
jovyan@jupyter-jupyterlab-jupyterlab-demo-sqqvjo7l:~/mcis6273-s25-datamining$ git config --global user.name "Your Name"
git config --global user.email "your_email@example.com"
jovyan@jupyter-jupyterlab-jupyterlab-demo-sqqvjo7l:~/mcis6273-s25-datamining$ git config --global user.name "Your Name"
git config --global user.email "your_email@example.com"
jovyan@jupyter-jupyterlab-jupyterlab-demo-sqqvjo7l:~/mcis6273-s25-datamining$ git config --global user.name "atomar4699"
jovyan@jupyter-jupyterlab-jupyterlab-demo-sqqvjo7l:~/mcis6273-s25-datamining$ git commit -m "Updated README.md with a new line"
[main 125fa11] Updated README.md with a new line
1 file changed, 1 insertion(+)
jovyan@jupyter-jupyterlab-jupyterlab-demo-sqqvjo7l:~/mcis6273-s25-datamining$ 
```

## **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**

```

File Edit View Run Kernel Tabs Settings Help
 README.md x jovyjan@jupyter-jupyterlab-ji X +
/ README.md
Name Modified
data yesterday
demo 5m ago
homework 9s ago
jupytercon2017 yesterday
move_it_here yesterday
scipy2017 yesterday
test_talk yesterday
build.py yesterday
jupyter_notebook... yesterday
LICENSE yesterday
move_this_file.txt yesterday
 README.md 5m ago
Y: talksyml yesterday

jovyjan@jupyter-jupyterlab-jupyterlab-demo-2zs44mh8:~/homework/hw0$ cd homework/hw0
jovyjan@jupyter-jupyterlab-jupyterlab-demo-2zs44mh8:~/homework/hw0$ cd data
jovyjan@jupyter-jupyterlab-jupyterlab-demo-2zs44mh8:~/homework/hw0/data$ pwd
/home/jovyjan/homework/hw0/data
jovyjan@jupyter-jupyterlab-jupyterlab-demo-2zs44mh8:~/homework/hw0/data$ ls -la
total 8
drwxr-xr-x 2 jovyjan jovyjan 4096 Feb 16 21:45 .
drwxr-xr-x 3 jovyjan jovyjan 4096 Feb 16 21:45 ..
jovyjan@jupyter-jupyterlab-jupyterlab-demo-2zs44mh8:~/homework/hw0/data$ wget https://github.com/kmsaumcis/mcis6273_s25_datamining/raw/main/hw0/2024_flavors_of_cacao.tsv
Resolving github.com (github.com) ... 140.82.121.3
Connecting to github.com (github.com)|140.82.121.3|:443... connected.
HTTP request sent, awaiting response... 302 Found
Location: https://raw.githubusercontent.com/kmsaumcis/mcis6273_s25_datamining/main/hw0/data/2024_flavors_of_cacao.tsv [following]
--2025-02-16 21:46:34-- https://raw.githubusercontent.com/kmsaumcis/mcis6273_s25_datamining/main/hw0/data/2024_flavors_of_cacao.tsv
Resolving raw.githubusercontent.com (raw.githubusercontent.com) ... 185.199.109.133, 185.199.110.133, 185.199.108.133, ...
Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|185.199.109.133|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 272727 (266K) [text/plain]
Saving to: '2024_flavors_of_cacao.tsv'

2024_flavors_of_cacao.tsv      100%[=====] 266.33K --.-KB/s   in 0.01s

2025-02-16 21:46:34 (21.9 MB/s) - '2024_flavors_of_cacao.tsv' saved [272727/272727]

jovyjan@jupyter-jupyterlab-jupyterlab-demo-2zs44mh8:~/homework/hw0$ ls -la
total 276
drwxr-xr-x 2 jovyjan jovyjan 4096 Feb 16 21:46 .
drwxr-xr-x 3 jovyjan jovyjan 4096 Feb 16 21:45 ..
-rw-r--r-- 1 jovyjan jovyjan 272727 Feb 16 21:46 2024_flavors_of_cacao.tsv
jovyjan@jupyter-jupyterlab-jupyterlab-demo-2zs44mh8:~/homework/hw0$ 

```

Now, run:

```
wget
https://github.com/kmsaumcis/mcis6273_s25_datamining/raw/main/hw0/data/2024_flavors_o
f_cacao.tsv
```

## 6. Confirm the Download

**ls -la**

**You should see:**

**2024\_flavors\_of\_cacao.tsv**

```

File Edit View Run Kernel Tabs Settings Help
 README.md x jovyjan@jupyter-jupyterlab-ji X +
/ README.md
Name Modified
data yesterday
demo 5m ago
homework 9s ago
jupytercon2017 yesterday
move_it_here yesterday
scipy2017 yesterday
test_talk yesterday
build.py yesterday
jupyter_notebook... yesterday
LICENSE yesterday
move_this_file.txt yesterday
 README.md 5m ago
Y: talksyml yesterday

jovyjan@jupyter-jupyterlab-jupyterlab-demo-2zs44mh8:~$ pwd
/home/jovyjan
jovyjan@jupyter-jupyterlab-jupyterlab-demo-2zs44mh8:~$ cd ~
jovyjan@jupyter-jupyterlab-jupyterlab-demo-2zs44mh8:~$ pwd
/home/jovyjan
jovyjan@jupyter-jupyterlab-jupyterlab-demo-2zs44mh8:~$ mkdir -p homework/hw0/data
jovyjan@jupyter-jupyterlab-jupyterlab-demo-2zs44mh8:~$ cd homework/hw0
jovyjan@jupyter-jupyterlab-jupyterlab-demo-2zs44mh8:~/homework/hw0$ cd data
jovyjan@jupyter-jupyterlab-jupyterlab-demo-2zs44mh8:~/homework/hw0/data$ pwd
/home/jovyjan/homework/hw0/data
jovyjan@jupyter-jupyterlab-jupyterlab-demo-2zs44mh8:~/homework/hw0/data$ ls -la
total 8
drwxr-xr-x 2 jovyjan jovyjan 4096 Feb 16 21:45 .
drwxr-xr-x 3 jovyjan jovyjan 4096 Feb 16 21:45 ..
jovyjan@jupyter-jupyterlab-jupyterlab-demo-2zs44mh8:~/homework/hw0/data$ wget https://github.com/kmsaumcis/mcis6273_s25_datamining/raw/main/hw0/2024_flavors_of_cacao.tsv
Resolving github.com (github.com) ... 140.82.121.3
Connecting to github.com (github.com)|140.82.121.3|:443... connected.
HTTP request sent, awaiting response... 302 Found
Location: https://raw.githubusercontent.com/kmsaumcis/mcis6273_s25_datamining/main/hw0/data/2024_flavors_of_cacao.tsv [following]
--2025-02-16 21:46:34-- https://raw.githubusercontent.com/kmsaumcis/mcis6273_s25_datamining/main/hw0/data/2024_flavors_of_cacao.tsv
Resolving raw.githubusercontent.com (raw.githubusercontent.com) ... 185.199.109.133, 185.199.110.133, 185.199.108.133, ...
Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|185.199.109.133|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 272727 (266K) [text/plain]
Saving to: '2024_flavors_of_cacao.tsv'

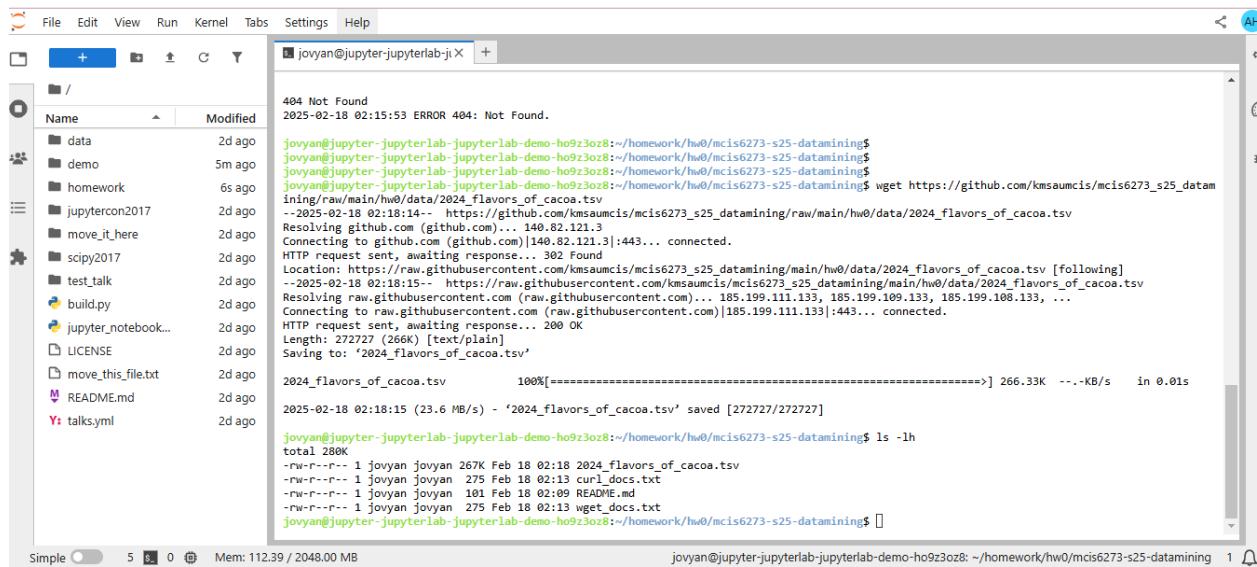
2024_flavors_of_cacao.tsv      100%[=====] 266.33K --.-KB/s   in 0.01s

2025-02-16 21:46:34 (21.9 MB/s) - '2024_flavors_of_cacao.tsv' saved [272727/272727]

jovyjan@jupyter-jupyterlab-jupyterlab-demo-2zs44mh8:~/homework/hw0$ 

```

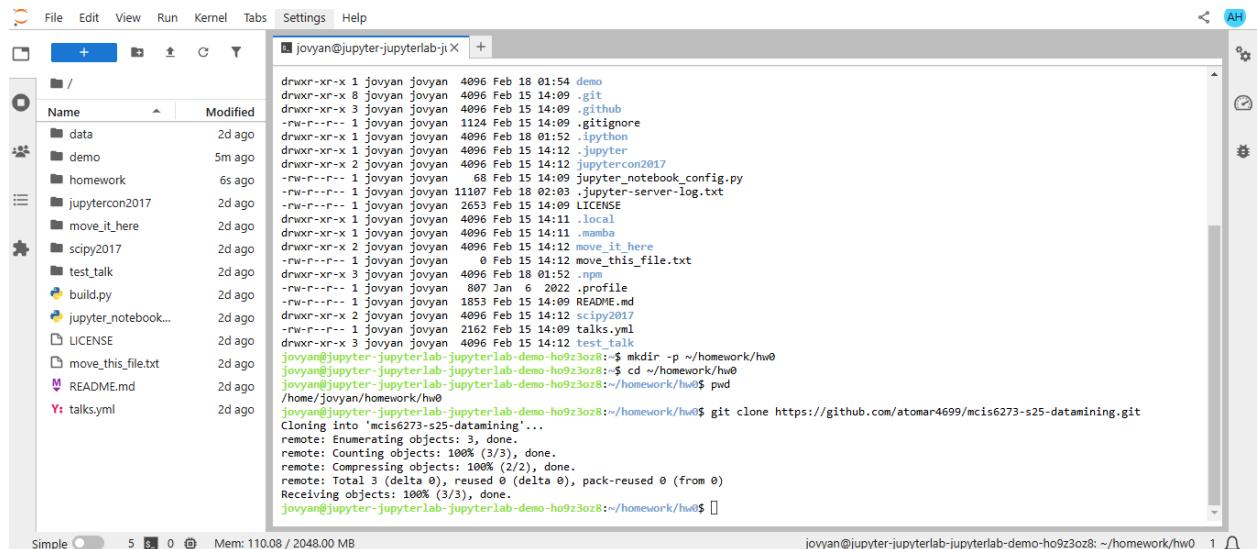
## 7. Verify the File:



```
joyyan@jupyter-jupyterlab-ji: ~
```

404 Not Found  
2025-02-18 02:15:53 ERROR 404: Not Found.  
joyyan@jupyter-jupyterlab-jupyterlab-demo-ho9z3oz8:~/homework/hw0/mcis6273-s25-datamining\$ wget https://github.com/kmsaumcis/mcis6273\_s25\_datamining/raw/main/hw0/data/2024\_flavors\_of\_cacao.tsv  
--2025-02-18 02:18:14- https://github.com/kmsaumcis/mcis6273\_s25\_datamining/main/hw0/data/2024\_flavors\_of\_cacao.tsv  
Resolving github.com (github.com)... 140.82.121.3  
Connecting to github.com (github.com)|140.82.121.3|:443... connected.  
HTTP request sent, awaiting response... 302 Found  
Location: https://raw.githubusercontent.com/kmsaumcis/mcis6273\_s25\_datamining/main/hw0/data/2024\_flavors\_of\_cacao.tsv [following]  
--2025-02-18 02:18:15- https://raw.githubusercontent.com/kmsaumcis/mcis6273\_s25\_datamining/main/hw0/data/2024\_flavors\_of\_cacao.tsv  
Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 185.199.111.133, 185.199.109.133, 185.199.108.133, ...  
Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|185.199.111.133|:443... connected.  
HTTP request sent, awaiting response... 200 OK  
Length: 272727 (266K) [text/plain]  
Saving to: '2024\_flavors\_of\_cacao.tsv'  
  
2024\_flavors\_of\_cacao.tsv 100%[=====] 266.33K --.-KB/s in 0.01s  
2025-02-18 02:18:15 (23.6 MB/s) - '2024\_flavors\_of\_cacao.tsv' saved [272727/272727]  
joyyan@jupyter-jupyterlab-jupyterlab-demo-ho9z3oz8:~/homework/hw0/mcis6273-s25-datamining\$ ls -lh  
total 280K  
-rw-r--r-- 1 joyyan joyyan 267K Feb 18 02:18 2024\_flavors\_of\_cacao.tsv  
-rw-r--r-- 1 joyyan joyyan 275 Feb 18 02:13 curl\_docs.txt  
-rw-r--r-- 1 joyyan joyyan 101 Feb 18 02:09 README.md  
-rw-r--r-- 1 joyyan joyyan 275 Feb 18 02:13 wget\_docs.txt  
joyyan@jupyter-jupyterlab-jupyterlab-demo-ho9z3oz8:~/homework/hw0/mcis6273-s25-datamining\$

## Clone Your Repository



```
joyyan@jupyter-jupyterlab-ji: ~
```

drwxr-xr-x 1 joyyan joyyan 4096 Feb 18 01:54 demo  
drwxr-xr-x 3 joyyan joyyan 4096 Feb 15 14:09 .git  
drwxr-xr-x 3 joyyan joyyan 4096 Feb 15 14:09 .github  
-rw-r--r-- 1 joyyan joyyan 1124 Feb 15 14:09 .gitignore  
drwxr-xr-x 1 joyyan joyyan 4096 Feb 15 14:12 .ipython  
drwxr-xr-x 1 joyyan joyyan 4096 Feb 15 14:12 .jupyter  
drwxr-xr-x 2 joyyan joyyan 4096 Feb 15 14:12 jupytercon2017  
-rw-r--r-- 1 joyyan joyyan 68 Feb 15 14:09 jupyter\_notebook\_config.py  
-rw-r--r-- 1 joyyan joyyan 1107 Feb 18 02:03 .jupyter-server-log.txt  
-rw-r--r-- 1 joyyan joyyan 2653 Feb 15 14:09 LICENSE  
drwxr-xr-x 1 joyyan joyyan 4096 Feb 15 14:11 .local  
drwxr-xr-x 1 joyyan joyyan 4096 Feb 15 14:11 .mamba  
drwxr-xr-x 2 joyyan joyyan 4096 Feb 15 14:12 move\_it\_here  
-rw-r--r-- 1 joyyan joyyan 0 Feb 15 14:12 move\_this\_file.txt  
drwxr-xr-x 3 joyyan joyyan 4096 Feb 18 01:52 .npn  
-rw-r--r-- 1 joyyan joyyan 807 Jan 6 2022 .profile  
-rw-r--r-- 1 joyyan joyyan 1853 Feb 15 14:09 README.md  
drwxr-xr-x 2 joyyan joyyan 4096 Feb 15 14:12 scipy2017  
-rw-r--r-- 1 joyyan joyyan 2162 Feb 15 14:09 talks.yml  
drwxr-xr-x 3 joyyan joyyan 4096 Feb 15 14:12 test\_talk  
joyyan@jupyter-jupyterlab-jupyterlab-demo-ho9z3oz8:~\$ mkdir -p ~/homework/hw0  
joyyan@jupyter-jupyterlab-jupyterlab-demo-ho9z3oz8:~\$ cd ~/homework/hw0  
joyyan@jupyter-jupyterlab-jupyterlab-demo-ho9z3oz8:~/homework/hw0\$ pwd  
/home/joyyan/homework/hw0  
joyyan@jupyter-jupyterlab-jupyterlab-demo-ho9z3oz8:~/homework/hw0\$ git clone https://github.com/atomar4699/mcis6273-s25-datamining.git  
Cloning into 'mcis6273-s25-datamining'...  
remote: Enumerating objects: 3, done.  
remote: Counting objects: 100% (3/3), done.  
remote: Compressing objects: 100% (2/2), done.  
remote: Total 3 (delta 0), reused 0 (delta 0), pack-reused 0 (from 0)  
Receiving objects: 100% (3/3), done.  
joyyan@jupyter-jupyterlab-jupyterlab-demo-ho9z3oz8:~/homework/hw0\$

## 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

The screenshot shows a Jupyter Notebook interface with a terminal tab open. The terminal output is as follows:

```
joyvan@jupyter-jupyterlab-ji:~ +  
drwxr-xr-x 2 joyvan joyvan 4096 Feb 15 14:12 move_it_here  
-rw-r--r-- 1 joyvan joyvan 0 Feb 15 14:12 move_this_file.txt  
drwxr-xr-x 3 joyvan joyvan 4096 Feb 18 01:52 .npm  
-rw-r--r-- 1 joyvan joyvan 807 Jan 6 2022 .profile  
-rw-r--r-- 1 joyvan joyvan 1853 Feb 15 14:09 README.md  
drwxr-xr-x 2 joyvan joyvan 4096 Feb 15 14:12 scipy2017  
-rw-r--r-- 1 joyvan joyvan 2162 Feb 15 14:09 talks.yml  
drwxr-xr-x 3 joyvan joyvan 4096 Feb 15 14:12 test_talk  
joyvan@jupyter-jupyterlab-jupyterlab-demo-ho9z3o8:~$ mkdir -p ~/homework/hw0  
joyvan@jupyter-jupyterlab-jupyterlab-demo-ho9z3o8:~$ cd ~/homework/hw0  
joyvan@jupyter-jupyterlab-jupyterlab-demo-ho9z3o8:~/homework/hw0$ pwd  
/home/joyvan/homework/hw0  
joyvan@jupyter-jupyterlab-jupyterlab-demo-ho9z3o8:~/homework/hw0$ git clone https://github.com/atomar4699/mcis6273-s25-datamining.git  
Cloning into 'mcis6273-s25-datamining'...  
remote: Enumerating objects: 1000+, done.  
remote: Counting objects: 1000+, done.  
remote: Compressing objects: 1000+, done.  
remote: Total 3 (delta 0), reused 0 (delta 0), pack-reused 0 (from remote)  
Receiving objects: 1000+ (3/3), done.  
joyvan@jupyter-jupyterlab-jupyterlab-demo-ho9z3o8:~/homework/hw0$ cd mcis6273-s25-datamining  
joyvan@jupyter-jupyterlab-jupyterlab-demo-ho9z3o8:~/homework/hw0$ nano README.md  
bash: nano: command not found  
joyvan@jupyter-jupyterlab-jupyterlab-demo-ho9z3o8:~/homework/hw0$ echo "Updated README" >> README.md  
joyvan@jupyter-jupyterlab-jupyterlab-demo-ho9z3o8:~/homework/hw0$ git config --global user.email "atomar4699@mulerider.s.ausmag.edu"  
git config --global user.name "atomar4699"  
joyvan@jupyter-jupyterlab-jupyterlab-demo-ho9z3o8:~/homework/hw0$ git add README.md  
joyvan@jupyter-jupyterlab-jupyterlab-demo-ho9z3o8:~/homework/hw0$ git commit -m "Updated README"  
[main 3804c86] Updated README  
1 file changed, 1 insertion(+)
```

## Check Your Current Working Directory

A screenshot of a Jupyter Notebook interface. On the left, there is a file tree showing various files and folders. In the center-right, a code cell [14] contains Python code to change the working directory and print the new current directory. The output shows the new current directory is /home/jovyan/homework/mcis6273\_s25\_datamining/hw0.

```
[14]: import os  
os.chdir("/home/jovyan/homework/mcis6273_s25_datamining/hw0") # Change working directory  
print("New Current Directory:", os.getcwd())
```

New Current Directory: /home/jovyan/homework/mcis6273\_s25\_datamining/hw0

Now run your data loading code:

A screenshot of a Jupyter Notebook interface. On the left, there is a file tree. In the center-right, a code cell [15] contains Python code to import pandas, specify a data file, and handle its existence. If the file exists, it prints "File found! Loading data..." and reads the CSV file. If it doesn't exist, it prints an error message. The output shows the file was found and the data was loaded, displaying a Pandas DataFrame with columns: REF, Company (Manufacturer), Company Location, Review Date, Country of Bean Origin, Specific Bean Origin or Bar Name, Cocoa Percent, Ingredients, Most Memorable Characteristics, and Rating. The data includes four rows of chocolate bean information.

```
[15]: import pandas as pd  
  
data_file = "data/2024_flavors_of_cacao.tsv"  
  
if os.path.exists(data_file):  
    print("File found! Loading data...")  
    df = pd.read_csv(data_file, sep="\t")  
    display(df.head()) # Show first few rows  
else:  
    print("Error: Data file not found! Check the file path.")
```

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	Zorral, 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

File Edit View Run Kernel Tabs Settings Help

jovyan@jupyter-jupyterlab-jx chocolate\_data\_engineering.x +

Download GitHub Binder Code

Notebook Python 3 (ipykernel)

Name Modified

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	Zorزال, 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

Simple 5 1 Python 3 (ipykernel) | Idle Mem: 263.19 / 2048.00 MB Mode: Edit Ln 1, Col 1 RTC:chocolate\_data\_engineering.ipynb 1

## Step 1: Clone the GitHub Repository

```
jovyan@jupyter-ipython-ipython-in-depth-g9pluq3t: ~/homework/mcis6273_s25_datamining/h
File Edit View Run Kernel Tabs Settings Help

drwxr-xr-x 10 jovyan jovyan 4096 Feb 15 14:09 exercises
-rw-r--r-- 1 jovyan jovyan 699 Feb 15 14:09 fabfile.py
drwxr-xr-x 8 jovyan jovyan 4096 Feb 15 14:09 .git
-rw-r--r-- 1 jovyan jovyan 112 Feb 15 14:09 .gitignore
-rw-r--r-- 1 jovyan jovyan 2755 Feb 15 14:09 Index.ipynb
drwxr-xr-x 5 jovyan jovyan 4096 Feb 18 22:01 .ipython
drwxr-xr-x 3 jovyan jovyan 4096 Feb 18 22:05 .jupyter
-rw-r--r-- 1 jovyan jovyan 10682 Feb 18 22:26 .jupyter-server-log.txt
-rw-r--r-- 1 jovyan jovyan 17626 Feb 15 14:09 LICENSE
drwxr-xr-x 3 jovyan jovyan 4096 Feb 18 22:01 .local
drwxr-xr-x 2 jovyan jovyan 4096 Feb 15 14:24 .mamba
-rwrxr-xr-x 1 jovyan jovyan 466 Feb 15 14:09 monitor.sh
drwxr-xr-x 3 jovyan jovyan 4096 Feb 18 22:01 .profile
-rw-r--r-- 1 jovyan jovyan 6 22:01 .profile
-rw-r--r-- 1 jovyan jovyan 2756 Feb 15 14:09 pycon-2015-abstract.md
-rw-r--r-- 1 jovyan jovyan 6824 Feb 15 14:09 pycon-submission.md
-rw-r--r-- 1 jovyan jovyan 1892 Feb 15 14:09 README.md
drwxr-xr-x 2 jovyan jovyan 4096 Feb 18 22:21 tomar_hw0
-rw-r--r-- 1 jovyan jovyan 185 Feb 18 22:21 tomar_hw0.tar.gz
drwxr-xr-x 2 jovyan jovyan 4096 Feb 15 14:09 tools
jovyan@jupyter-ipython-in-depth-g9pluq3t:~$ cd tomar_hw0
jovyan@jupyter-ipython-in-depth-g9pluq3t:~/tomar_hw0$ mkdir -p ~/homework
cd ~/homework
jovyan@jupyter-ipython-in-depth-g9pluq3t:~/homework$ git clone https://github.com/atomar4699/mcis6273_s25_datamining.git
Cloning into 'mcis6273_s25_datamining'...
remote: Enumerating objects: 15, done.
remote: Counting objects: 100% (15/15), done.
remote: Compressing objects: 100% (11/11), done.
remote: Total 15 (delta 3), reused 15 (delta 3), pack-reused 0 (from 0)
Receiving objects: 100% (15/15), 127.63 Kib | 12.76 MiB/s, done.
Resolving deltas: 100% (3/3), done.
jovyan@jupyter-ipython-in-depth-g9pluq3t:~/homework$ cd mcis6273_s25_datamining/hw0
jovyan@jupyter-ipython-in-depth-g9pluq3t:~/homework/mcis6273_s25_datamining/hw0$
```

Simple 6 0 Mem: 101.08 / 2048.00 MB jovyan@jupyter-ipython-in-depth-g9pluq3t: ~/homework/mcis6273\_s25\_datamining/hw0 1

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

```

jovyan@jupyter-ipython-ipython-in-depth-g9pluq3t: ~/homework/mcis6273_s25_datamining/hw0$ cd tomor_hw0
jovyan@jupyter-ipython-ipython-in-depth-g9pluq3t:~/homework$ mkdir -p ~/homework
jovyan@jupyter-ipython-ipython-in-depth-g9pluq3t:~/homework$ git clone https://github.com/atomar4699/mcis6273_s25_datamining.git
Cloning into 'mcis6273_s25_datamining'...
remote: Enumerating objects: 15, done.
remote: Counting objects: 100% (15/15), done.
remote: Compressing objects: 100% (11/11), done.
remote: Total 15 (delta 3), reused 15 (delta 3), pack-reused 0 (from 0)
Receiving objects: 100% (15/15), 127.63 KiB | 12.76 MiB/s, done.
Resolving deltas: 100% (3/3), done.
jovyan@jupyter-ipython-ipython-in-depth-g9pluq3t:~/homework$ cd mcis6273_s25_datamining/hw0
jovyan@jupyter-ipython-ipython-in-depth-g9pluq3t:~/homework/mcis6273_s25_datamining/hw0$ mv ~/tomor_hw0.tar.gz ~/homework/mcis6273_s25_datamining/hw0/
jovyan@jupyter-ipython-ipython-in-depth-g9pluq3t:~/homework/mcis6273_s25_datamining/hw0$ ls -la ~/homework/mcis6273_s25_datamining/hw0/
total 104
drwxr-xr-x 3 jovyan jovyan 4096 Feb 18 22:29 .
drwxr-xr-x 4 jovyan jovyan 4096 Feb 18 22:27 ..
drwxr-xr-x 2 jovyan jovyan 4096 Feb 18 22:27 data
-rw-r--r-- 1 jovyan jovyan 16485 Feb 18 22:27 hw0.ipynb
-rw-r--r-- 1 jovyan jovyan 13808 Feb 18 22:27 hw0.md
-rw-r--r-- 1 jovyan jovyan 53017 Feb 18 22:27 hw0.pdf
-rw-r--r-- 1 jovyan jovyan 185 Feb 18 22:21 tomor_hw0.tar.gz
jovyan@jupyter-ipython-ipython-in-depth-g9pluq3t:~/homework/mcis6273_s25_datamining/hw0$ 

```

Simple 6 0 Mem: 100.51 / 204.00 MB jovyan@jupyter-ipython-ipython-in-depth-g9pluq3t: ~/homework/mcis6273\_s25\_datamining/hw0 1

### Step 3: Upload the File to GitHub

```

jovyan@jupyter-ipython-ipython-in-depth-g9pluq3t: ~/homework/mcis6273_s25_datamining/hw0$ cd mcis6273_s25_datamining/hw0
File Edit View Run Kernel Tabs Settings Help
remote: Compressing objects: 100% (11/11), done.
remote: Total 15 (delta 3), reused 15 (delta 3), pack-reused 0 (from 0)
Receiving objects: 100% (15/15), 127.63 KiB | 12.76 MiB/s, done.
Resolving deltas: 100% (3/3), done.
jovyan@jupyter-ipython-ipython-in-depth-g9pluq3t:~/homework$ cd mcis6273_s25_datamining/hw0
jovyan@jupyter-ipython-ipython-in-depth-g9pluq3t:~/homework/mcis6273_s25_datamining/hw0$ mv ~/tomor_hw0.tar.gz ~/homework/mcis6273_s25_datamining/hw0/
jovyan@jupyter-ipython-ipython-in-depth-g9pluq3t:~/homework/mcis6273_s25_datamining/hw0$ ls -la ~/homework/mcis6273_s25_datamining/hw0/
total 104
drwxr-xr-x 3 jovyan jovyan 4096 Feb 18 22:29 .
drwxr-xr-x 4 jovyan jovyan 4096 Feb 18 22:27 ..
drwxr-xr-x 2 jovyan jovyan 4096 Feb 18 22:27 data
-rw-r--r-- 1 jovyan jovyan 16485 Feb 18 22:27 hw0.ipynb
-rw-r--r-- 1 jovyan jovyan 13808 Feb 18 22:27 hw0.md
-rw-r--r-- 1 jovyan jovyan 53017 Feb 18 22:27 hw0.pdf
-rw-r--r-- 1 jovyan jovyan 185 Feb 18 22:21 tomor_hw0.tar.gz
jovyan@jupyter-ipython-ipython-in-depth-g9pluq3t:~/homework/mcis6273_s25_datamining/hw0$ cd ~/homework/mcis6273_s25_datamining/hw0
jovyan@jupyter-ipython-ipython-in-depth-g9pluq3t:~/homework/mcis6273_s25_datamining/hw0$ git add tomor_hw0.tar.gz
jovyan@jupyter-ipython-ipython-in-depth-g9pluq3t:~/homework/mcis6273_s25_datamining/hw0$ git commit -m "Added Hw0 submission archive"
Author identity unknown

*** Please tell me who you are.

Run

git config --global user.email "you@example.com"
git config --global user.name "Your Name"

to set your account's default identity.
Omit --global to set the identity only in this repository.

fatal: unable to auto-detect email address (got 'jovyan@jupyter-ipython-ipython-in-depth-g9pluq3t.(none)')
jovyan@jupyter-ipython-ipython-in-depth-g9pluq3t:~/homework/mcis6273_s25_datamining/hw0$ git push origin main
Username for 'https://github.com': 


```

Simple 6 0 Mem: 116.07 / 204.00 MB jovyan@jupyter-ipython-ipython-in-depth-g9pluq3t: ~/homework/mcis6273\_s25\_datamining/hw0 1

## Data Transformation:

### 1 Convert "Cacao Percent" to Float

The screenshot shows a Jupyter Notebook interface with the following details:

- File Bar:** File, Edit, View, Run, Kernel, Tabs, Settings, Help.
- Toolbar:** Includes icons for file operations like new, open, save, and run cells.
- File Explorer:** Shows a directory structure under "/ ... / hw0 / data /". Files listed include "2024\_flavors\_of\_c...", "step1\_cocoa\_perc...", and "tomar.ipynb".
- Code Cell:** Contains Python code for reading a CSV file, converting the "Cocoa Percent" column to float, and saving the result. A checkmark indicates the step was successful: "Step 1: Cocoa Percent converted and saved."
- Output Cell:** Displays the status message from the previous cell.
- Bottom Status Bar:** Shows "Simple" mode, Python 3 (ipykernel) | Idle, Mem: 241.90 / 2048.00 MB, Mode: Edit, Ln 1, Col 1, tomar.ipynb, and a cell number indicator (1).

## 2 Split "Ingredients" into Columns

The screenshot shows a Jupyter Notebook interface with the following details:

- File Bar:** File, Edit, View, Run, Kernel, Tabs, Settings, Help.
- Toolbar:** Includes icons for file operations like new, open, save, and run cells.
- File Explorer:** Shows a directory structure under "/ ... / hw0 / data /". Files listed include "2024\_flavors\_of\_c...", "step1\_cocoa\_perc...", and "step2\_ingredients...".
- Code Cell:** Contains Python code for reading a CSV file, defining ingredient labels, handling missing values, splitting the "Ingredients" column into multiple columns, applying get\_dummies, and merging the dataframes. A checkmark indicates the step was successful: "Step 2: Ingredients processed and saved."
- Output Cell:** Displays the status message from the previous cell.
- Bottom Status Bar:** Shows "Simple" mode, Python 3 (ipykernel) | Idle, Mem: 241.90 / 2048.00 MB, Mode: Edit, Ln 1, Col 1, tomar.ipynb, and a cell number indicator (1).

## 3 Extract "Most Memorable Characteristics"

A screenshot of a Jupyter Notebook interface. The left sidebar shows a file tree with files like '2024\_flavors\_of\_cacao.csv', 'step1\_cocoa\_percents.csv', 'step2\_ingredients.csv', 'step3\_characteristics.csv', and 'tomar.ipynb'. The main area contains Python code for extracting frequent characteristics from a dataset and saving it. A status message at the bottom right indicates 'Step 3: Most Memorable Characteristics extracted and saved.'

```
# Load intermediate dataset
df = pd.read_csv("/home/jovyan/tomar/hw0/data/step2_ingredients.csv")

# Extract frequent characteristics
char_counts = df["Most Memorable Characteristics"].str.get_dummies(sep=", ").sum()
common_chars = char_counts[char_counts >= 20].index

# Add new columns
df[common_chars] = df["Most Memorable Characteristics"].str.get_dummies(sep=", ")[common_chars]

# Save intermediate result
df.to_csv("/home/jovyan/tomar/hw0/data/step3_characteristics.csv", index=False)
print("Step 3: Most Memorable Characteristics extracted and saved.")
```

## 4 Remove Unnecessary Columns

A screenshot of a Jupyter Notebook interface. The left sidebar shows a file tree with files like '2024\_flavors\_of\_cacao.csv', 'step1\_cocoa\_percents.csv', 'step2\_ingredients.csv', 'step3\_characteristics.csv', 'step4\_cleaned.csv', and 'tomar.ipynb'. The main area contains Python code for dropping unnecessary columns ('Ingredients' and 'Most Memorable Characteristics') and saving the cleaned data. A status message at the bottom right indicates 'Step 4: Unnecessary columns removed and saved.'

```
[13]: import pandas as pd

# Load intermediate dataset
df = pd.read_csv("/home/jovyan/tomar/hw0/data/step3_characteristics.csv")

# Drop unwanted columns
df.drop(columns=["Ingredients", "Most Memorable Characteristics"], inplace=True)

# Save intermediate result
df.to_csv("/home/jovyan/tomar/hw0/data/step4_cleaned.csv", index=False)
print("Step 4: Unnecessary columns removed and saved.")
```

## 5 Save Cleaned Data

The screenshot shows a Jupyter Notebook interface. On the left, a file tree displays files in the directory `/hw0 / data /`, including `2024_flavors_of_cacao.csv`, `step1_cocoa_perc...`, `step2_ingredients...`, `step3_characteristi...`, `step4_cleaned.csv`, and `tomar.ipynb`. The main notebook cell contains Python code for loading a dataset, defining output paths for CSV and JSON, and saving the data. A status message at the bottom of the cell indicates "Step 5: Cleaned data saved as CSV and JSON." The bottom status bar shows "Python 3 (ipykernel) | Idle" and memory usage.

```

import pandas as pd

# Load final cleaned dataset
df = pd.read_csv("/home/jovyan/tomar/hw0/data/step4_cleaned.csv")

# Define output paths
csv_output_path = "/home/jovyan/tomar/hw0/data/cleaned_data_2025_flavors_of_cacao.csv"
json_output_path = "/home/jovyan/tomar/hw0/data/cleaned_data_2025_flavors_of_cacao.json"

# Save as CSV and JSON
df.to_csv(csv_output_path, index=False)
df.to_json(json_output_path, orient="records", indent=4)

print("Step 5: Cleaned data saved as CSV and JSON.")

```

To display the complete DataFrame without truncation in Jupyter Notebook,

The screenshot shows a Jupyter Notebook interface. On the left, a file tree displays files in the directory `/hw0 / data /`, including `2024_flavors_of_cacao.csv`, `cleaned_data_2025...`, `step1_cocoa_perc...`, `step2_ingredients...`, `step3_characteristi...`, `step4_cleaned.csv`, and `tomar.ipynb`. The main notebook cell contains Python code for reading a CSV file, setting pandas display options to show all rows, columns, and content, and then displaying the DataFrame. The resulting DataFrame is shown below, with all columns and rows visible. The bottom status bar shows "Python 3 (ipykernel) | Idle" and memory usage.

```

# Load the DataFrame
df = pd.read_csv("/home/jovyan/tomar/hw0/data/step2_ingredients.csv")

# Set pandas display options for full output
pd.set_option("display.max_rows", None) # Show all rows
pd.set_option("display.max_columns", None) # Show all columns
pd.set_option("display.max_colwidth", None) # Show full column content
pd.set_option("expand_frame_repr", False) # Prevent wrapping to the next line

# Display the full DataFrame
display(df)

```

	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	Zorza, 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

The screenshot shows a Jupyter Notebook interface with the following details:

- File Browser:** On the left, a sidebar shows a directory structure under `hw0 / data /` containing files like `2024\_flavors\_of\_cacao.csv`, `cleaned\_data\_202...`, `full\_table.csv`, etc.
- Open Tabs:**
  - tomar.ipynb
  - full\_table.html
  - 2024\_flavors\_of\_cacao.tsv
- Content Area:** The `full\_table.html` tab displays a table titled "Cacao Flavors" with 141 rows of data. The columns include ID, Name, Country, Year, Origin, Ratings, and Descriptions.
- Bottom Status Bar:** Shows "Simple" mode, memory usage (300.47 / 2048.00 MB), and the date/time (7:57 PM).

## Step 6: Export CSV & JSON

### 1 Extract & Save Reduced Data

```
[18]: import os
       # Ensure the directory exists
       output_dir = "data"
       os.makedirs(output_dir, exist_ok=True)

       # Save the files
       df_reduced.to_csv(f"{output_dir}/data_reduced_2025_flavors_of_cacao.csv", index=False)
       df_reduced.to_json(f"{output_dir}/data_reduced_2025_flavors_of_cacao.json", orient="records")

[19]: # Extract & Save Reduced Data
       df_reduced = df[["Review Date", "Country of Bean Origin", "Cocoa Percent", "Rating"]]
       df_reduced.to_csv(f"{output_dir}/data_reduced_2025_flavors_of_cacao.csv", index=False)
       df_reduced.to_json(f"{output_dir}/data_reduced_2025_flavors_of_cacao.json", orient="records")
```

A screenshot of a Jupyter Notebook interface. The top navigation bar includes File, Edit, View, Run, Kernel, Tabs, Settings, and Help. Below the navigation bar is a file browser showing a directory structure with files like 'data\_reduced\_202...', '2096', 'Review Date 2021', 'Country of Bean Origin "Uganda"', 'Cocoa Percent 0.75', 'Rating 3.75', and several other numerical entries. The bottom status bar shows 'Simple' mode, 7 tabs open, 1 kernel, Mem: 296.74 / 2048.00 MB, and a file path 'data\_reduced\_2025\_flavors\_of\_cacao.json'.

## 2 Filter & Save Data

A screenshot of a Jupyter Notebook interface. The top navigation bar includes File, Edit, View, Run, Kernel, Tabs, Settings, and Help. Below the navigation bar is a file browser showing a directory structure with files like 'data\_filtered\_2025...', 'data\_reduced\_202...', and others. The main area shows a code cell with the following Python code:

```
[22]: print(df.columns)
Index(['REF', 'Company (Manufacturer)', 'Company Location', 'Review Date',
       'Country of Bean Origin', 'Specific Bean Origin or Bar Name',
       'Cocoa Percent', 'Ingredients', 'Most Memorable Characteristics',
       'Rating', 'ingredient_count', 'ingredient_details', 'B', 'C', 'L', 'S',
       'S*', 'Sa', 'V'],
      dtype='object')

[23]: flavor_columns = ["fatty", "earthy", "roasty"]
available_flavor_columns = [col for col in flavor_columns if col in df.columns]

df_filtered = df[
    (df["Rating"] >= 3.25) &
    (df["Cocoa Percent"].between(0.65, 0.75)) &
    (df["Review Date"].between(2018, 2021))
]

if available_flavor_columns:
    df_filtered = df_filtered[available_flavor_columns].sum(axis=1) > 0

df_filtered.to_csv(f"{output_dir}/data_filtered_2025_flavors_of_cacao.csv", index=False)
df_filtered.to_json(f"{output_dir}/data_filtered_2025_flavors_of_cacao.json", orient="records")
```

The bottom status bar shows 'Python 3 (ipykernel) | Idle' and 'Mem: 306.46 / 2048.00 MB'.

File Edit View Run Kernel Tabs Settings Help

joyyan@jupyter-ator X tomor.ipynb data\_filtered\_2025\_fl X data\_reduced\_2025\_X full\_table.html 2024\_flavors\_of\_cacao X +

Delimiter: . ,

Name Modified

	REF	Company (Manufacturer)	Company Location	Review Date	Country of Bean Origin	Bean Origin or Bar Name	Cocoa
1	2542	5150	U.S.A.	2021	India	Anamalai, batch 1	
2	2206	A. Morin	France	2018	Venezuela	Porcelana	
3	2648	A. Morin	France	2021	Mexico	La Joya	
4	2470	Acalli	U.S.A.	2020	Peru	Norandino Tumbes blend	
5	2462	Acalli	U.S.A.	2020	Mexico	Teapa, Tabasco, batch 2	
6	2092	Amadel	Italy	2018	Blend	Nine	
7	2586	Animas	U.S.A.	2021	Guatemala	Lanquin regions, b. 4243	
8	2254	Arete	U.S.A.	2018	Colombia	Tumaco	
9	2330	Arete	U.S.A.	2019	India	Jangareddygudem	
10	2162	Argencove	Nicaragua	2018	Nicaragua	Mombacho	
11	2162	Argencove	Nicaragua	2018	Nicaragua	Cocibolca	
12	2162	Argencove	Nicaragua	2018	Nicaragua	Masaya	
13	2566	Aruntam	Italy	2021	Tanzania	Kokoa Kamili, lot TZ72199	
14	2562	Aruntam	Italy	2021	India	Tukki Kerala, lot INT78221	
15	2562	Aruntam	Italy	2021	Madagascar	E., Akesson, lot MA72820	
16	2056	Auro	Philippines	2018	Philippines	Saloy E., 2016	
17	2294	Bankston	U.S.A.	2019	Tanzania	Kokoa Kamili	
18	2190	Belvie	Vietnam	2018	Vietnam	Lam Dong	

Simple 7 Mem: 320.96 / 2048.00 MB data\_filtered\_2025\_flavors\_of\_cacao.csv 1

### 3 Create ZIP/TAR File

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

File Edit View Run Kernel Tabs Settings Help

joyyan@jupyter-jupyterlab-ji X tomor\_hw0.ipynb cleaned\_data\_2025\_flavors\_o X README.md +

On branch main  
Your branch is up to date with 'origin/main'.

Untracked files:  
(use "git add <file>..." to include in what will be committed)  
2024\_flavors\_of\_cacao.tsv  
Newreadme-Copy1.md

nothing added to commit but untracked files present (use "git add" to track)

```
joyyan@jupyter-jupyterlab-jupyterlab-demo-plu0r2jb:~/tomar/homework/mcis6273-s25-datamining$ pip install pandas
Requirement already satisfied: pandas in /srv/conda/envs/notebook/lib/python3.12/site-packages (2.2.3)
Requirement already satisfied: numpy>=1.26.0 in /srv/conda/envs/notebook/lib/python3.12/site-packages (from pandas) (2.2.3)
Requirement already satisfied: python-dateutil>=2.8.2 in /srv/conda/envs/notebook/lib/python3.12/site-packages (from pandas) (2.9.0.post0)
Requirement already satisfied: pytz>=2020.1 in /srv/conda/envs/notebook/lib/python3.12/site-packages (from pandas) (2024.1)
Requirement already satisfied: tzdata>=2022.7 in /srv/conda/envs/notebook/lib/python3.12/site-packages (from pandas) (2025.1)
Requirement already satisfied: six>=1.5 in /srv/conda/envs/notebook/lib/python3.12/site-packages (from python-dateutil>=2.8.2>pandas) (1.17.0)
joyyan@jupyter-jupyterlab-jupyterlab-demo-plu0r2jb:~/tomar/homework/mcis6273-s25-datamining$ tar -czvf atomar_hw0_files.tar.gz -C ~/tomar/homew
ork/hw0/
hw0/
hw0/data/
hw0/data/tomar_hw0.ipynb
hw0/data/data/
hw0/data/data/reduced_2025_flavors_of_cacao.csv
hw0/data/data/reduced_2025_flavors_of_cacao.json
hw0/data/data/filterd_2025_flavors_of_cacao.csv
hw0/data/data/.ipynb_checkpoints/
hw0/data/data/cleaned_data_2025_flavors_of_cacao.csv
hw0/data/data/filterd_2025_flavors_of_cacao.json
hw0/data/2024_flavors_of_cacao.tsv
hw0/data/.ipynb_checkpoints/
hw0/data/.ipynb_checkpoints/tomar_hw0-checkpoint.ipynb
joyyan@jupyter-jupyterlab-jupyterlab-demo-plu0r2jb:~/tomar/homework/mcis6273-s25-datamining$ []
```

Simple 1 0 Mem: 128.01 / 2048.00 MB joyyan@jupyter-jupyterlab-jupyterlab-demo-plu0r2jb: ~ /tomar/homework/mcis6273-s25-datamining 1