

MCIS6273 Data Mining (Prof. Maull) / Spring 2025 / HW0

Points Possible	Due Date	Time Commitment (estimated)
20	Monday February 17 @ Midnight	<i>up to</i> 18 hours

- **GRADING:** Grading will be aligned with the completeness of the objectives.
- **INDEPENDENT WORK:** Copying, cheating, plagiarism and academic dishonesty *are not tolerated* by University or course policy. Please see the syllabus for the full departmental and University statement on the academic code of honor.

OBJECTIVES

- Familiarize yourself with Github and basic git
- Familiarize yourself with the JupyterLab environment, Markdown and Python
- Explore JupyterHub Linux console integrating what you learned in the prior parts of this homework
- Perform basic data engineering in Python using Chocolate Review Data

WHAT TO TURN IN

You are being encouraged to turn the assignment in using the provided Jupyter Notebook. To do so, make a directory in your Lab environment called `homework/hw0`. Put all of your files in that directory. Then zip or tar that directory, rename it with your name as the first part of the filename (e.g. `maull_hw0_files.zip`, `maull_hw0_files.tar.gz`), then download it to your local machine, then upload the `.zip` to Blackboard.

If you do not know how to do this, please ask, or visit one of the many tutorials out there on the basics of using zip in Linux.

If you choose not to use the provided notebook, you will still need to turn in a `.ipynb` Jupyter Notebook and corresponding files according to the instructions in this homework.

ASSIGNMENT TASKS

(0%) Familiarize yourself with Github and basic git

Github (<https://github.com>) is the *de facto* platform for open source software in the world based on the very popular [git](https://git-scm.org) (<https://git-scm.org>) version control system. Git has a sophisticated set of tools for version control based on the concept of local repositories for fast commits and remote repositories only when collaboration and remote synchronization is necessary. Github enhances git by providing tools and online hosting of public and private repositories to encourage and promote sharing and collaboration. Github hosts some of the world's most widely used open source software.

If you are already familiar with git and Github, then this part will be very easy!

\$ Task: Create a Zotero account.

Learn about Zotero and if you haven't already, create a free account:

- <https://zotero.org>

\$ Task: Create a public Github repo named "mcis6273-s25-datamining" and place a README.md file in it.

Create your first file called `README.md` at the top level of the repository.

Please put your Zotero username in the file. Aside from that you can put whatever text you like in the file (If you like, use something like [Lorem Ipsum](#) to generate random sentences to place in the file.). Please include the link to **your** Github repository that now includes the minimal `README.md`. You don't have to have anything elaborate in that file or the repo.

\$ Task: Fork the course repository.

Learn to use Github workflows and fork the class repo:

- https://github.com/kmsaumcis/mcis6273_s25_datamining/

(0%) Familiarize yourself with the JupyterLab environment, Markdown and Python

As stated in the course announcement [Jupyter](https://jupyter.org) (<https://jupyter.org>) is the core platform we will be using in this course and is a popular platform for data scientists around the world. We have a JupyterLab setup for this course so that we can operate in a cloud-hosted environment, free from some of the resource constraints of running Jupyter on your local machine (though you are free to set it up on your own and seek my advice if you desire).

You have been given the information about the Jupyter environment we have setup for our course, and the underlying Python environment will be using the [Anaconda](https://anaconda.com) (<https://anaconda.com>) distribution. It is not necessary for this assignment, but you are free to look at the multitude of packages installed with Anaconda, though we will not use the majority of them explicitly.

As you will soon find out, Notebooks are an incredibly effective way to mix code with narrative and you can create cells that are entirely code or entirely Markdown. Markdown (MD or md) is a highly readable text format that allows for easy documentation of text files, while allowing for HTML-based rendering of the text in a way that is style-independent.

We will be using Markdown frequently in this course, and you will learn that there are many different “flavors” or Markdown. We will only be using the basic flavor, but you will benefit from exploring the “Github flavored” Markdown, though you will not be responsible for using it in this course – only the “basic” flavor. Please refer to the original course announcement about Markdown.

\$ Task: THERE IS NOTHING TO TURN IN FOR THIS PART.

Play with and become familiar with the basic functions of the Lab environment given to you online in the course Blackboard.

\$ Task: THERE IS NOTHING TO TURN IN FOR THIS PART.

Please *create a markdown document* and read the documentation for basic Markdown [here](#). Learn to use all of the following:

- headings (one level is fine),
- bullets,
- bold and italics

Again, the content of your document can be whatever you like, just learn some of the basic functionality of Markdown.

(0%) Explore JupyterHub Linux console integrating what you learned in the prior parts of this homework

The Linux console in JupyterLab is a great way to perform command-line tasks and is an essential tool for basic scripting that is part of a data scientist’s toolkit. Open a console in the lab environment and familiarize yourself with your files and basic commands using git as indicated below.

1. In a new JupyterLab command line console, run the `git clone` command to clone the new repository you created in the prior part. You will want to read the documentation on this command (try here <https://www.git-scm.com/docs/git-clone> to get a good start).
2. Within the same console, modify your README.md file, check it in and push it back to your repository, using `git push`. Read the [documentation about git push](#).
3. The commands `wget` and `curl` are useful for grabbing data and files from remote resources off the web. Read the documentation on each of these commands by typing `man wget` or `man curl` in the terminal. Make sure you pipe the output to a file or use the proper flags to do so.

\$ Task: THERE IS NOTHING TO TURN IN FOR THIS PART.

(100%) Perform basic data engineering in Python using Chocolate Review Data

We have learned that data engineering is an important task and ultimately the *initial* process which most data mining begins.

In this assignment you will get your hands with *real* data that have relevant and practical application.

As you may have different feelings about it, but *many* humans love chocolate. I am personally a dark chocolate lover – 70% cocoa or greater, thank you. But I cannot call myself a chocolate Connoisseur, though maybe one day I’ll have enough time and money to be one. But there are indeed true connoisseurs and there is even an interesting site out there dedicated to it:

- <https://flavorsofcacao.com/>

On this site, you might learn a lot about chocolate tasting, evaluation and even another group's evaluation of chocolate from around the world – and the source of our data for this part of the assignment.

You will notice there is a “database” of over 2800 plain dark chocolate bars that have been reviewed over several years. These bars hail from all over the world and are subjectively rated, though it is unclear who actually did the rating.

In any event, while it is called a database, to you and I it is a big Javascript table. I have made it easier for the assignment by including that file on Github here:

- https://github.com/kmsaumcis/mcis6273_s25_datamining/blob/main/hw0/data/2024_flavors_of_cocoa.tsv

Data engineering as you have learned from the readings is about transforming data from one form to another so that it can be used in the appropriate analysis contexts.

In this part of the homework you will prepare data for analysis. This dataset is rather small, but it is a great warmup set for introductory data engineering and usage of Pandas.

Your code must be implemented in Jupyter as a notebook and you will be required to turn in a .ipynb file.

§ Task: Use Python/Pandas to load, transform and store a CSV file.

A lot of times we would like to clean up data in a file. In this case, we have a lot of it, and the reality is that we only need *some* of the data in the file.

You will first need to grab the .tsv listed above into your Jupyter environment.

You will need to place that file into a folder called data and then load the CSV into a DataFrame using the following code:

```
df = pd.read_csv( "./data/2024_flavors_of_cocoa.tsv", sep='\t')
```

There is a starter notebook on Github with the first cells running these steps for you. See:

- https://github.com/kmsaumcis/mcis6273_f25_datamining/hw0/hw0_starter_nb.ipynb

When you inspect the DataFrame you will notice that some of the data will need transformation and must do so using Pandas: <https://pandas.pydata.org>.

DATA TRANSFORMATION

There are three things we are going to do to the data:

1. We are going to transform the "Cacao Percent" column to a floating between 0 and 1, so that the original string "76%" is transformed to the number 0.76. Later this will facilitate machine learning algorithms.
2. We are going to split the "Ingredients" column into multiple columns. You will notice that the column currently might look something like "3- B,S,C". This means there are 3 ingredients, "B", "S" and "C". Above the table on the site is a description of what these are. You will need the following tools: `DataFrame.apply()`, `Series.str.split()` and `Series.str.strip()`. Once your munging is done, make sure the column type is set to int using `pd.astype()` and that the resulting columns are merged into the main DataFrame. *This step should add 8 new columns to your DataFrame: "B", "C", "L", "S", "S*", "Sa", "V" and "ingredient_count".*
3. We are going to split out the "Most Memorable Characteristics" into **all** of the characteristics listed which occur **20 or more times**. To do this, you may need to study `Series.str.get_dummies()`. You might also find the `Series.agg("sum")`, `Series.sort_values()`, `Series.where()` or `DataFrame.query()` of great use. There are other ways to accomplish the goal, but these are accessible and easily implemented in one or two lines. You will end up with 71 new columns.

DATA FINALIZATION

To finalize your data, you will need to drop the "Ingredients" and "Most Memorable Characteristics" columns. They are no longer needed (steps 2 and 3 above solved that problem for us).

Your final DataFrame should have 86 total columns.

MINIMUM EXPECTATIONS IN THE ANSWER FOR THIS TASK PART

Your answer must include:

- the cleaned files need to go into a 'data/' folder (created by you)

- the first row of your newly exported CSV file must include ALL the headers of ALL columns
- all the data rows should have data and nothing else (e.g. there should be no rows with headers or junk data, etc.)
- your new cleaned output file should be called 'data/cleaned_data_2025_flavors_of_cacao.csv'

§ Task: Extract, transform and export JSON data.

Now that you have a CSV file, we will transform and deploy it to CSV and now JSON. The file needs to have the following characteristics:

- the columns will be restricted to 4 columns, Review Date, Country of Bean Origin, Cacao Percent and Rating
- store the finalized data set in two files (use `Pandas.to_csv()` and `Pandas.to_json()`):
 - a CSV file 'data/data_reduced_2025_flavors_of_cacao.csv'
 - a JSON file 'data/data_reduced_2025_flavors_of_cacao.json'

§ Task: Filter, transform and export CSV data.

You will now loop over all the data to produce a filtered file with the following:

- *at or above* rating 3.25,
- a cacao percent between 65 and 75,
- reviewed between 2018 and 2021
- and have “fatty”, “earthy” or “roasty” characteristics.

Reduce the data to only those rows with a filters provided and store it in a file. Use the enriched DataFrame you produced from the first part, NOT the source original file.

You must store the data table in CSV and JSON files called 'data/data_filtered_2025_flavors_of_cacao.csv' and 'data/data_filtered_2025_flavors_of_cacao.json' in the 'data/' folder

§ Task: Complete the online HW0 assessment.

Once you are done with the coding part of the assignment, you will need to complete the online assessment for the final 6 points of your grade.