Note: I read the instructions (first tab on the spreadsheet) and also skimmed the data to check the formatting and what the core concept behind the data storage is. Then, I gave the whole spreadsheet to chatGPT and asked it to explain the project to me. This way, I can ask it questions as I am trying to understand what exactly to do and bounce off ideas from it as if it were a real person.

What do I use the donations tab for? All required data for the 2 output tabs is in the other 2 input tabs.

Donation History is out of scope for outputs
→ But, I can use it to validate existing tags?

Sometimes names and emails are missing. Are those the anonymous cases?
→ There's no flag for this in the sheet, so I'll assume it's not safe to assume.

ChatGPT said: "Could two reasonable engineers disagree about this value?"

If yes → **leave it blank**."

Lol. I'll also leave blank the rows where first and last names are missing along with the emails. But it still has a patronID so it can't be discarded. It'll just contribute to unclean data that nothing can be done about to make it prettier.
→ document this in the README

Alright alright alright

Inputs to use
- Input Constituents → primary
- Input Emails → supporting
- Input Donation History → QA only

Outputs to produce
- CueBox Constituents CSV
- CueBox Tags CSV

The primary key is the PatronID.

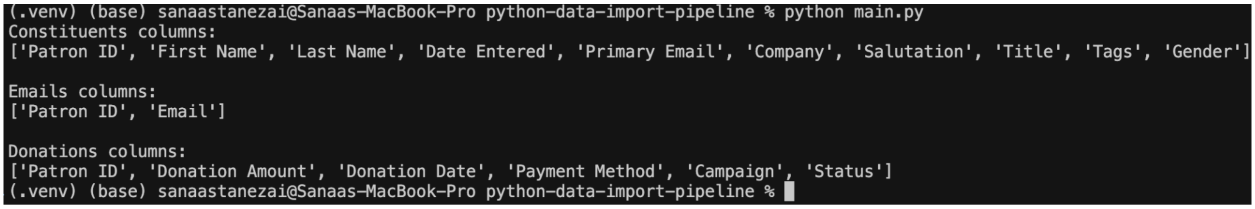
Rules:
Missing names → leave blank
Missing email → leave blank
Multiple emails → deterministic choice
Tags → split, trim, deduplicate
Donation inconsistencies → logged, not fixed

Below table is from chatGPT after I decided on business rules

Output Column	Source Column(s)	Rule	Notes
first_name	Input Constituents.First Name	normalize case	blank if missing
last_name	Input Constituents.Last Name	normalize case	blank if missing
Email 1	Input Constituents.Primary Email	deterministic pick	documented
Email 2	Input Emails.Email	deterministic pick	documented
is_company	Input Constituents.Company	boolean logic	see README
title	Input Constituents.Title	trim	blank allowed

Made a repository. Starting code now in VSCode.

Stuff is printing!!!



```
(.venv) (base) sanaastanezai@Sanaas-MacBook-Pro python-data-import-pipeline % python main.py
Constituents columns:
['Patron ID', 'First Name', 'Last Name', 'Date Entered', 'Primary Email', 'Company', 'Salutation', 'Title', 'Tags', 'Gender']

Emails columns:
['Patron ID', 'Email']

Donations columns:
['Patron ID', 'Donation Amount', 'Donation Date', 'Payment Method', 'Campaign', 'Status']
(.venv) (base) sanaastanezai@Sanaas-MacBook-Pro python-data-import-pipeline % 
```

That’s all for today. I’m going to sleep. 12:39am

I’m back.

Clean up the data. Emails are already lowercase. Some first names have 2 names like __ & __.
What to do there?.
Chat said to leave it because “You are doing a **migration**, not data correction”. 🫡

There’s a bunch of emails with domains like .biz or something unusual.
I’m not correcting or rejecting these as long as the format is valid (has @ and a domain)

“Date Entered” column is little ugly. Not in the same date format and some have time stamps.

OMG wait, 1288 shows up twice with different names. Whattt
→ I’m gonna go with Rebecca on this because it has a more complete row (has tags). I’ll mention Brandon’s row in the README but not include it in the output file bc ID is key value.

Mr. and Mrs. AND Rev salutations not allowed. Will print empty strings for this and document it.

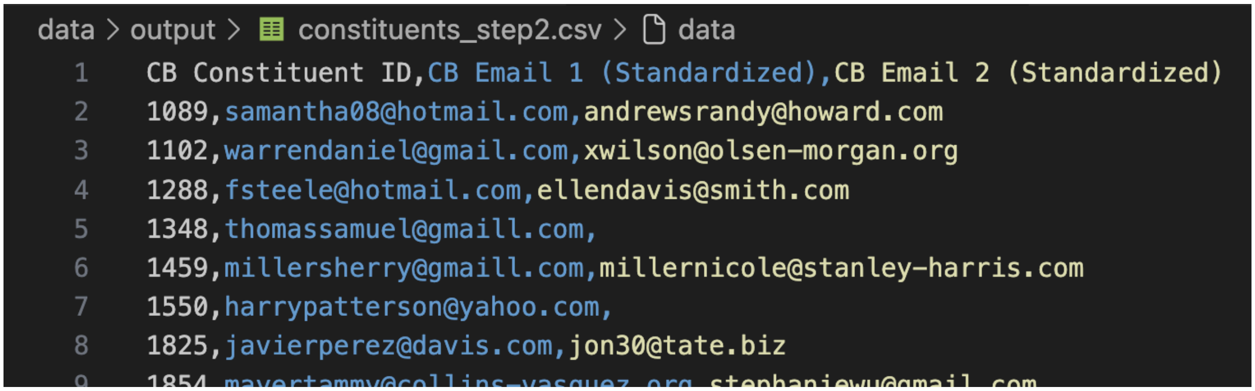
Idk why marital status column is labeled as “Gender” LOL.
→ This’ll be used for the Background Information field.

Oh wait, Donations tab is needed.
- CB Lifetime Donation Amount
- CB Most Recent Donation Date
- CB Most Recent Donation Amount

Sometimes, the same tag is written twice for one Constituent. Count that once.

Okay, now I generated a partial constituents CSV that deduplicates Patron ID, picks email 1 and email 1 (sometimes it’s blank because no email was given in the input file)

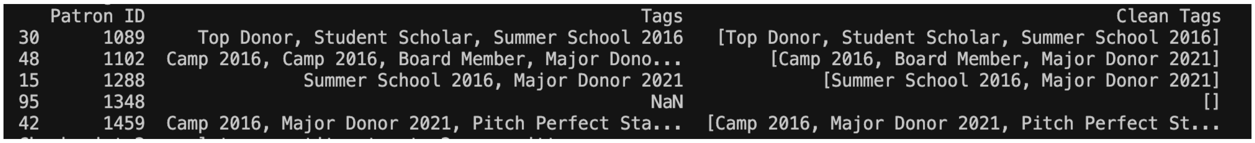
I validated the output with the primary email given and the alternate email given in the emails tab. It matched up :)



```
CB Constituent ID,CB Email 1 (Standardized),CB Email 2 (Standardized)
1089,samantha08@hotmail.com,andrewsrandy@howard.com
1102,warrendaniel@gmail.com,xwilson@olsen-morgan.org
1288,fsteele@hotmail.com,ellendavis@smith.com
1348,thomassamuel@gmaill.com,
1459,millersherry@gmaill.com,millernicole@stanley-harris.com
1550,harrypatterson@yahoo.com,
1825,javierperez@davis.com,jon30@tate.biz
1854,mayertammy@collins-vasquez.org,stephaniewu@gmail.com
```

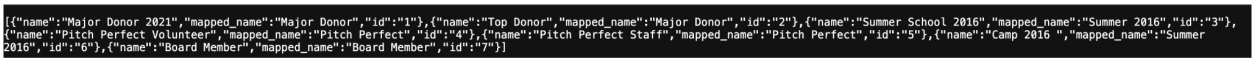
100 patron IDs and corresponding emails are in the temp file. There are 101 in teh given input file but that’s because there was one repeated ID so after removing the duplicate, there are 100 entries.

Now, all other necessary information for the CueBox constitutes output file will be easily accessed except for the tags. So, I need to Split + clean tags per Patron ID. Done below. Basically, duplicates are removed, and random “NaN” stuff is changed to empty strings.



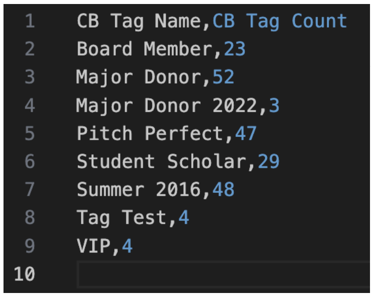
```
   Patron ID                                        Tags                                        Clean Tags
30      1089     Top Donor, Student Scholar, Summer School 2016   [Top Donor, Student Scholar, Summer School 2016]
48      1102  Camp 2016, Camp 2016, Board Member, Major Dono...       [Camp 2016, Board Member, Major Donor 2021]
15      1288                Summer School 2016, Major Donor 2021           [Summer School 2016, Major Donor 2021]
95      1348                                             NaN                                                []
42      1459  Camp 2016, Major Donor 2021, Pitch Perfect Sta...  [Camp 2016, Major Donor 2021, Pitch Perfect St...
```

I shall now apply the tag-mapping API given in the assignment to normalize tag names.



```
[{"name":"Major Donor 2021","mapped_name":"Major Donor","id":"1"},{"name":"Top Donor","mapped_name":"Major Donor","id":"2"},{"name":"Summer School 2016","mapped_name":"Summer 2016","id":"3"},
{"name":"Pitch Perfect Volunteer","mapped_name":"Pitch Perfect","id":"4"},{"name":"Pitch Perfect Staff","mapped_name":"Pitch Perfect","id":"5"},{"name":"Camp 2016 ","mapped_name":"Summer
2016","id":"6"},{"name":"Board Member","mapped_name":"Board Member","id":"7"}]
```

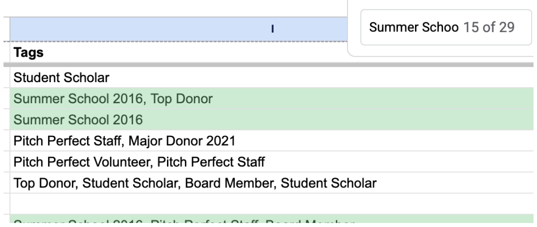
Both “Major Donor 2021” and “Top Donor” are mapped to “Major Donor”. Should have 7 mapped tags. BUT, we also have “Major Donor 2022” in the sheet, so that stays as-is since it’s not mentioned in the API. That gives me a total of 8 tags.



```
CB Tag Name,CB Tag Count
Board Member,23
Major Donor,52
Major Donor 2022,3
Pitch Perfect,47
Student Scholar,29
Summer 2016,48
Tag Test,4
VIP,4

```

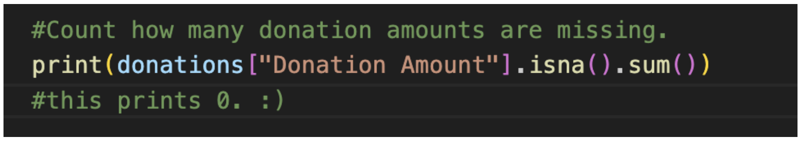
Now, I checked to see if the tag count is correct by Ctrl f-ing in the sheet. Checks out



Tags
Student Scholar
Summer School 2016, Top Donor
Summer School 2016
Pitch Perfect Staff, Major Donor 2021
Pitch Perfect Volunteer, Pitch Perfect Staff
Top Donor, Student Scholar, Board Member, Student Scholar
Summer School 2016, Pitch Perfect Staff, Board Member

“CB Lifetime Donation Amount” means I need to only count donations with a “paid” status. Don’t count the refunded amounts.

validate that all paid donations have amounts before aggregating, so the rollups wouldn’t be affected by missing data.



```python
#Count how many donation amounts are missing.
print(donations["Donation Amount"].isna().sum())
#this prints 0. :)
```

Note: the output tab wants a timestamp of the constituent's most recent donation but most of the donations in the input tab don’t have a time so pandas fills the missing time as “00:00:00” when converting to a timestamp.

Okay, I think I’m doneee

Some names are in all Caps. who’s yelling. I’m gonna fix this; it doesn’t count as data correction, just normalization.

All that work and this is how the output file looks lol —>

I am once again humbled by how time consuming the coding portion was. It’s not the cleanest, but it works.

Will complete the read me and check how to submit.

Big note: the output file rules say that first name and last name are required if the type is a person. BUT, some rows (like ID 2976) are missing first and last names AND missing company names. So, in my code, that case is labeled as “person” but it is missing names. I couldn't just come up with names nor could I come up with a company name and label it as type “company” I hope that’s okay.