3rd July 2025

Final dataset descriptions:  
dataset name: final\_course\_evals\_one\_row\_per\_prof\_per\_course\_per\_term.xlsx

|  |  |
| --- | --- |
| Column name | Description |
| professor\_uni | Map of professor uni that was already in the course evals data, or a rough mapping which may be inaccurate. |
| professor\_fullname | Fullname field in the course evals data. Is not clean- sometimes has names reversed, or alternate names, or multiple prof names |
| lowercase | Converted to lowercase for better matching (standardizing) |
| enrollment\_total | From course evals data- original data itself- no manipulations. |
| eval\_respondents | No manipulations-- this was used to evaluate which bie\_course or bie\_professor should be kept based on the matches in tcdb |
| bie\_course | If multiple different values existed for same unique course id and same prof, this was manually chosen as to what best matched the tcdb. |
| bie\_professor | If multiple different values existed for same unique course id and same prof, this was manually chosen as to what best matched the tcdb. |
| average\_workload\_rating | No manipulations |
| response\_rate | No manipulations |
| clean\_course\_number | Cleaned the course\_number field in order to standardize it across datasets: Make sure it's a string, strip whitespace, and convert to uppercase; Extract the first letter and a 4-digit number at the end; removing special chars etc. |
| new\_unique\_course\_id | Re-did unique course\_id as the original unique\_course\_id field was sometimes messy—freshly combined (as strings) term\_number+ clean\_course\_number + section\_number |
| name\_1 | Professor fullname column was split into four possible names based on spaces between names. Names were cleaned to standardize: lowercase, no special characters, no whitespace |
| name\_2 | “ |
| name\_3 | “ |
| name\_4 | “ |
| fixed\_name | UNI was mapped based on the four name parts—and the professors’ full names as in SIS (or tcdb) were added to “fixed names”—this column has clean professor names. Instances where there were multiple profs: this has each professor names for each course (the ‘multiple’ names wee exploded to give each prof their own row). This column best matches names as in SIS or tcdb |
| correct\_map | This is the correct and final UNI map for each professor based on all models and manual mapping. |

27 June 225

course\_evals = pd.read\_excel("final\_course\_evals\_with\_Uni.xlsx")

#THis is the excel I created after adding UNIs to every prof.

#Some rows have multiple profs still-- labelled 'multiple';

#HOWEVER, the rows for each courses with individual profs already exists--

#but for some reason, when we remove 'multiple' there are fewer rows than the original dataset

course\_evals\_indiv = pd.read\_excel("final\_course\_evals\_with\_Uni.xlsx", "final\_map multiple removed")

SIS = pd.read\_excel("SIS Course Instructors.xlsx")

c

# count number of rows with same 'unique\_course\_id'

unq\_course\_counts = course\_evals\_indiv['unique\_id'].value\_counts().reset\_index()

unq\_course\_counts.columns = ['unique\_id', 'count']

# filtering for courses with >=2 values

multi\_unq\_courses = unq\_course\_counts[unq\_course\_counts['count'] >= 2]

#getting full data of these courses from course\_evals\_indv

multi\_courses\_df = course\_evals\_indiv[course\_evals\_indiv['unique\_id'].isin(multi\_unq\_courses['unique\_id'])]

#Now, we need to make sure the PROF Final\_unis are diff-- and only retain rows where the unique\_id and 'final uni map' are the same

# Step 1: Count occurrences of each (unique\_id, final\_uni\_map) pair

pair\_counts = multi\_courses\_df.groupby(['unique\_id', 'final uni map']).size().reset\_index(name='count')

# Step 2: Filter for combinations that occur more than once

duplicate\_pairs = pair\_counts[pair\_counts['count'] > 1]

# Step 3: Filter the original dataframe to keep only those (unique\_id, final\_uni\_map) pairs

duplicate\_pairs\_course\_evals = multi\_courses\_df.merge(duplicate\_pairs[['unique\_id', 'final uni map']],

on=['unique\_id', 'final uni map'], how='inner')

FOUND THAT THERE ARE STILL SOME UNIS WHICH ARE MISATTRIBUTED TO A DIFF PROF TEACHING THE SAME COURSE. IM GOING TO GO BACK TO final\_course\_evals\_with\_Uni.xlsx AND FIX THIS THEN RE-RUN (Remember to UDPDATE THE HIVE XL).--- in the final excel—changing the map for profs who have wrong uni to the correct ones manually—then fixing the main sheet with XLOOKUP.

New column called ‘correct\_remap’

import pandas as pd

course\_evals = pd.read\_excel("final\_course\_evals\_with\_Uni.xlsx")

#THis is the excel I created after adding UNIs to every prof.

#Some rows have multiple profs still-- labelled 'multiple';

#HOWEVER, the rows for each courses with individual profs already exists--

#but for some reason, when we remove 'multiple' there are fewer rows than the original dataset

course\_evals\_indiv = pd.read\_excel("final\_course\_evals\_with\_Uni.xlsx", "final\_map multiple removed")

SIS = pd.read\_excel("SIS Course Instructors.xlsx")

# count number of rows with same 'unique\_course\_id'

unq\_course\_counts = course\_evals\_indiv['unique\_id'].value\_counts().reset\_index()

unq\_course\_counts.columns = ['unique\_id', 'count']

# filtering for courses with >=2 values

multi\_unq\_courses = unq\_course\_counts[unq\_course\_counts['count'] >= 2]

#getting full data of these courses from course\_evals\_indv

multi\_courses\_df = course\_evals\_indiv[course\_evals\_indiv['unique\_id'].isin(multi\_unq\_courses['unique\_id'])]

#Now, we need to make sure the PROF Final\_unis are diff-- and only retain rows where the unique\_id and 'correct remap' are the same

# Step 1: Count occurrences of each (unique\_id, correct remap) pair

pair\_counts = multi\_courses\_df.groupby(['unique\_id', 'correct remap']).size().reset\_index(name='count')

# Step 2: Filter for combinations that occur more than once

duplicate\_pairs = pair\_counts[pair\_counts['count'] > 1]

# Step 3: Filter the original dataframe to keep only those (unique\_id, correct remap) pairs

duplicate\_pairs\_course\_evals = multi\_courses\_df.merge(duplicate\_pairs[['unique\_id', 'correct remap']],

on=['unique\_id', 'correct remap'], how='inner')

AAAH I NOTICED AS KIERSTEN POINTED OUR THAT SOME unique\_id in COURSE\_EVALS RAW DATA DON’t HAVE 13 digits!! (they are missing the term\_number!!!USE new\_unique\_course\_id COLUMN WHICH I HAD ALREADY CREATED for name\_uni mapping!!!

Now I will export the courses which are duplicates (‘duplicate\_pairs\_course\_evals,csv’) and then manually look up enrollment total for that unique course Id to decide what to do with the course eval numbers!

* Added column ‘kepp\_manual’ --- 0 to drop, 1 to keep; 2 to average.
* Manual duplicate removal notes:
* course ID 20203B9659001 seems to have 23 enrolled students in the SIS dataset, but in the course\_evals-- there's two rows for this course: one says enrollment is 24, the other says enrollment is 23. which one do I keep, and should I update it in the SIS?
* Can’t find course 20213B6502005 in SIS
* Can’t find course 20213B8767001 in SIS
* Cant find 20221B9334001 in SIS:
  + Profs

|  |
| --- |
| Razvan Popescu |
| Ritt Keeratiwutthikul |

Don’t exist=-- on tcdb: it’s

|  |  |
| --- | --- |
| Harry | Mamaysky |

* Can’t find 20223B6302003 in SIS
* Cant find 20223B6302006 in SIS
* Cant find 20243B7103001in SIS

Finding missing SIS courses on tcdb and deciding on retaining.

1 July 2025

Re-ran code for new final\_prof\_uni course evals: “final\_course\_evals\_with UNI.xlsx”

Got waayyy more—but some are just duplicates not identified as duplicates because the name col in the splitting process is diff. cleaning this up using excel formula. In retained rows:

=INDEX(FINAL\_duplicate\_pairs\_course\_ev!A2:Y629,

MATCH(UNIQUE(FINAL\_duplicate\_pairs\_course\_ev!Z2:Z629),

FINAL\_duplicate\_pairs\_course\_ev!Z2:Z629,

0),

SEQUENCE(1, COLUMNS(FINAL\_duplicate\_pairs\_course\_ev!A2:Y2)))

When I do this, there’s 320 rows, with just the ones from the “OLD duplicates” version as multiple profs—which were already tagged.

I WILL ALSO NEED TO ADDRESS SAME UNIQUE COURSE ID BUT DIFF PROF- WHO IS THE EVAL FOR????--- check on tcdb.