DEI: Take Home Exercise Processing Guide (WI college-going rates)

2021 Wisconsin (WI) Percent Enrolled in College Immediately Following High School

The Product team has asked us to load new college-going rates for the state of Wisconsin for the College Success Awards. Unfortunately, the source data from the Wisconsin Department of Education does not come in a readily loadable format for the data project. The Data Engineering team has been tasked with creating a scriptable solution for transforming the underlying data files into a state that can be ingested by our data warehouse (e.g. example output attached in file).

Please use the below write-up/documentation to inform your scriptable solution.

Data File(s):

File for use in numerator creation:
postsecondary_enrollment_current_2020_21.csv

File for use in denominator creation:
hs_completion_certified_2020-21.zip > hs_completion_certified_2020-21.csv

Join "postsecondary_enrollment_current_{year}.csv" and "hs_completion_certified_{year}.csv" files:

Fields to match on across files
SCHOOL_YEAR, DISTRICT_CODE, SCHOOL_CODE, GROUP_BY_VALUE

Identifying entity levels and how to construct `state_id` field:

> **State_id** is a unique foreign key provided by the state Department of Education to uniquely identify a school or district.

Level	Path	State ids
State	postsecondary_enrollment_current_{year}.csv > SCHOOL_NAME is "[Statewide]"	N/A format as string "state"
District	postsecondary_enrollment_current_{year}.csv > SCHOOL_NAME is "[Districtwide]"	DISTRICT_CODE
School	postsecondary_enrollment_current_{year}.csv > SCHOOL_NAME **is neither** "[Districtwide]" **nor** "[Statewide]"	CONCATENATE: DISTRICT_CODE, SCHOOL_CODE

Columns to filter before aggregation:

`breakdown`: `GROUP_BY_VALUE`

Column	Keep Values	Skip Values	Notes
GROUP_BY_VALUE	"All Students" "Amer Indian" "Asian "Black" "Econ Disadv" "ELL/LEP" "Eng Prof" (Non-LEP) "Female" "Hispanic" "Male" "Not Econ Disadv" "Not Migrant" "Pacific Isle" "SwD" "SwoD" (students without disabilities) "Two or More" (two or more races) "White" "Migrant"	"[Data Suppressed]" "Unknown"	

postsecondary_enrollment_current_2020_21.csv: `INITIAL_ENROLLMENT`

Column	Keep Values	Skip Values	Notes
`INITIAL_ENROLLMENT`	“First Fall”	“Later Enrollment”, “Second Fall”, “*”	We are not interested in calculating college enrollment outside of “First Fall” per the data type definition

hs_completion_certified_2020-21.csv: `TIMEFRAME`, `COMPLETION STATUS`, `COHORT`

Column	Keep Values	Skip Values	Notes
`TIMEFRAME`	“4-year rate”	“5-Year rate”, “6-Year rate”, “7-Year rate”	Per our business logic, we are only interested in college going rates for 4-year graduates
`COMPLETION STATUS`	“Completed - Regular High School Diploma”, “Completed - High School Equivalency Diploma”, “Completed - Other High School Completion Credential”	“Not Completed - Continuing Toward Completion”, “Not Completed - Not Continuing”, “Not Completed - Not Known to be Continuing Toward Completion”, “Not Completed - Reached Maximum Age”, “*”	Per data definition, college going rates are calculated out of high school students who did graduate /complete.
`COHORT`	“2021”	”2018”, “2019”, “2020”	

`value_numerator`: sum the `STUDENT_COUNT` values in the postsecondary_enrollment_current_2020_21.csv file, for each `GROUP_BY_VALUE`, across `INSTITUTION_LOCATION`, `INSTITUTION_LEVEL`, `INSTITUTION_TYPE`

`value_denominator`: sum the `STUDENT_COUNT` values in the hs_completion_certified_2020-21.csv file, for each `GROUP_BY_VALUE`

`value`: calculated college-going rate percent:

`value_numerator` / `value_denominator` * 100

Column	Keep Values	Skip Values	Notes
value	Where % rate was able to be calculated	“*” in calculation	This is the `calculated college-going rate` value, which represents the percent of students enrolled in college immediately following High School (first Fall after graduating)