Education Meets Opportunity Platform (EMOP)

Florida Department of Education

Comparing Florida Postsecondary Career Technical Education (CTE) Programs

Data Pipeline Document

Authors:

Dr. A. Yonah Meiselman

Dr. Joe Long

October 20, 2023

RIPL | \ri'-pəl\ | **Research Improving People's Lives**
One Park Row, Suite 401, Providence, RI 02903
https://ripl.org

Purpose

This document serves to explain how the datasets used in the Education Meets Opportunity Platform (EMOP) Return on Investment (ROI) dashboard postsecondary CTE program analysis are constructed from the raw data, describing each script used, its location, runtime, inputs, and outputs. The intended audience is an analyst who seeks to replicate the analysis done previously or make edits/updates to the underlying data and needs to generate the analysis datasets.

Outline

The data pipeline consists of five child scripts that generate intermediate datasets, as well as one script that takes intermediate datasets and merges them together to produce two more intermediate datasets, which we call “merged datasets.” These are called by one parent script to run the entire pipeline.

Child Scripts

The data pipeline is set up to create the following five intermediate datasets from the raw datasets, which are listed below each intermediate dataset (with the unique identifier used to join them in parentheses):

1. Individual FCS dataset (joined by fl_anonymous_id)
    - Script: s3://code/derived_non_scons/fcs_individual.R
    - Runtime: 1-2 minutes
    - Level of observation: Individual, by program
    - Input datasets:
        - s3://CCTCMIS/FCS_Demographic
        - s3://CCTCMIS/FCS_Program
        - s3://CCTCMIS/FCS_Completion
    - Description: This script joins the FCS datasets together by merging on individual identifier (fl_anonymous_id) into an individual-level cross-sectional dataset
    - Output: s3://code/scratch/derived/fcs_individual.csv
    - Purpose of inclusion: For all analyses (to identify program of study)
2. Individual WDIS dataset (joined by fl_anonymous_id)
    - Script: s3://code/derived_non_scons/wdis_indiviudal2.R
    - Runtime: 4-5 minutes
    - Level of observation: Individual, by program
    - Input datasets:
        - s3://WDIS/WDIS_Demographics
        - s3://WDIS/WDIS_End_Term_Status
        - s3://WDIS/WDIS_Supplement
        - s3://WDIS/WDIS_CTE_Course
        - s3://WDIS/WDIS_Adult_Education_Course
    - Description: This script joins the WDIS datasets together by merging on individual identifier (fl_anonymous_id) into an individual-level cross-sectional dataset
    - Output: s3://code/scratch/derived/wdis_individual2.csv
    - Purpose of inclusion: For all analyses (to identify program of study)
3. Wage dataset (joined by emp_num/empnum)
    - Script: s3://code/derived_non_scons/wage_individual.R
    - Runtime: 3-4 minutes per quarter
    - Level of observation: Individual, by employer, by quarter
    - Input datasets:

        - All quarters of wage files in the s3://FETPIP-Wage folder from 2016 Q1 to 2021 Q3
        - s3://FETPIP_Employers/FETPIP_Employer
    - Description: This script joins each wage dataset with the employer file by merging on employer identifier (emp_num/empnum)
    - Output: A set of files with the pattern "*s3://code/scratch/derived/wage_individual_{quarter}.csv,*" where *quarter* is the quarter corresponding to the wage in YYQ format
    - Purpose of inclusion: All outcome analyses related to wages and employment outcomes
4. SNAP/TANF dataset (appended)
    - Script: s3://code/derived_non_scons/snap_tanf.R
    - Runtime: 1-2 minutes
    - Level of observation: Individual, by quarter
    - Input datasets:
        - s3://SNAP_TANF/FCS_EMOP_PA_MATCH_THRU2022Q2.csv
        - s3://SNAP_TANF/WDIS_EMOP_PA_MATCH_THRU2022Q2.csv
    - Description: This script appends the FCS and WDIS SNAP/TANF datasets together into an individual-by-quarter level panel dataset
    - Output: s3://code/scratch/derived/snap_tanf.csv
    - Purpose of inclusion: All analyses related to SNAP/TANF outcomes
5. Continuing education dataset (appended)
    - Script: s3://code/derived_non_scons/continuing_ed.py
    - Runtime: 1-2 minutes
    - Level of observation: Individual, by quarter, by program
    - Input datasets:
        - All postsecondary files in the s3://Continuing_Education folder with the format "RIPL_Postsecondary_cont_edu_{????}.data", where ???? is the school year, for 2017-18 through 2021-2022.
    - Description: This script appends the continuing education datasets together into an individual-quarter-program level panel dataset
    - Output: s3://working/continuing_ed.csv
    - Purpose of inclusion: Continuing education outcome analysis

From the intermediate datasets, two merged datasets are also created: fcs_wages.csv and wdis_wages.csv:

6. FCS + WDIS analysis dataset
    - Script: s3://code/derived_non_scons/fcs_wdis_wages.R
    - Runtime: 15-20 minutes
    - Level of observation: Individual, by quarter
    - Input datasets:
        - s3://code/scratch/derived/fcs_individual.csv
        - s3://code/scratch/derived/wdis_individual2.csv
        - s3://code/scratch/derived/wage_individual_{quarter}.csv for each quarter
        - s3://code/scratch/derived/snap_tanf.csv
    - Description: This script joins each of the individual FCS and WDIS datasets to the wage dataset and the SNAP/TANF dataset
    - Outputs: s3://working/fcs_wages.csv & s3://working/wdis_wages.csv
    - Purpose of inclusion: All analyses

The demographics files CCTCMIS/FCS_Demographic and WDIS/WDIS_Demographics contain both the fl_anonymous_id and emop_uuid variables; fl_anonymous_id allows for joining individual data with the SNAP/TANF data, and emop_uuid allows for joining individual data with the wage data.

Runbook

The R script s3://code/derived_non_scons/compile.R runs all of the cleaning, calling the six child scripts detailed in the previous section in the order in which they appear below.

- 1. fcs_individual.R
- 2. wdis_individual2.R
- 3. wage_individual.R
- 4. snap_tanf.R
- 5. continuing_ed.py
- 6. fcs_wdis_wages.R

In compile.R, every script's inputs and outputs are specified, and then passed to the child scripts. The input file paths can be updated as new data comes in. As described above, the final outputs are the two merged intermediate datasets, s3://working/fcs_wages.csv & s3://working/wdis_wages.csv, plus the continuing education intermediate dataset s3://working/continuing_ed.csv. These are the three datasets used in the analysis (as detailed in the analysis documentation).

Note: After generating the files for the data pipeline, they must be copied into the folder

For ease of updating the files and/or filepaths, in the compile.R tab of the Excel spreadsheet runbook-supp.xlsx, all of the lines where there are filepaths are listed, as well as descriptions. For example, if the FETPIP Employer dataset location changes, then you can go to line 45 of compile.R, and update the filepath.

The data pipeline makes use of the following R packages:

- data.table
- fst
- dplyr
- dtplyr
- lubridate
- stringr
- assertthat
- readr

and the following Python packages:

- pandas
- math
- os
- sys

Appendix: Variables

Intermediate dataset	Raw dataset source	Variable name	Definition
FCS individual	*CCTCMIS/FCS_Demographic*	FI_anonymous_id	Unique ID
FCS individual	*CCTCMIS/FCS_Demographic*	Entry_quarter	Earliest term quarter
FCS individual	*CCTCMIS/FCS_Demographic*	College	College name
FCS individual	*CCTCMIS/FCS_Demographic*	Uuid	Wage – individual ID
FCS individual	*CCTCMIS/FCS_Demographic*	Disability	Disability status
FCS individual	*CCTCMIS/FCS_Demographic*	Gender	Gender
FCS individual	*CCTCMIS/FCS_Demographic*	Birth_date	Birth date
FCS individual	*CCTCMIS/FCS_Demographic*	Grad_code_hschl	Highest schooling completed
FCS individual	*CCTCMIS/FCS_Demographic*	First_time	First time student
FCS individual	*CCTCMIS/FCS_Demographic*	Mil_status	Military status
FCS individual	*CCTCMIS/FCS_Demographic*	Race	Race
FCS individual	*CCTCMIS/FCS_Demographic*	Ethnicity	Ethnicity
FCS individual	*CCTCMIS/FCS_Program*	Program_quarter	Program quarter
FCS individual	*CCTCMIS/FCS_Program*	Program_cip	Program CIP
FCS individual	*CCTCMIS/FCS_Completion*	Completion_quarter	Completion quarter
FCS individual	*CCTCMIS/FCS_Completion*	Completion_cip	Completion CIP
FCS individual	*CCTCMIS/FCS_Completion*	Degree	Degree
FCS individual	*CCTCMIS/FCS_Completion*	Cip	CIP code
WDIS individual	*WDIS/WDIS_Demographics*	FI_anonymous_id	Unique ID
WDIS individual	Derived from *quarter* in *WDIS/WDIS_Demographics*	Min_quarter	Earliest quarter
WDIS individual	Derived from *quarter* in *WDIS/WDIS_Demographics*	Max_quarter	Latest quarter
WDIS individual	*WDIS/WDIS_CTE_Course* *WDIS/WDIS_Adult_Education_Course*	Program_code	Program code
WDIS individual	*WDIS/WDIS_CTE_Course* *WDIS/WDIS_Adult_Education_Course*	Disability	Disability
WDIS individual	Derived from *WDIS/WDIS_CTE_Course, WDIS/WDIS_Adult_Education_Course*	N	Number of courses in most common program code
WDIS individual	*WDIS/WDIS_End_Term_Status*	Completion_quarter	Completion quarter
WDIS individual	*WDIS/WDIS_End_Term_Status*	Diploma_type	Diploma type
WDIS individual	*WDIS/WDIS_End_Term_Status*	Certificate_type	Certificate type
WDIS individual	*WDIS/WDIS_Demographics*	Uuid	Wage - individual ID

Intermediate dataset	Raw dataset source	Variable name	Definition
WDIS individual	*WDIS/WDIS_Demographics*	District_id	District number
WDIS individual	*WDIS/WDIS_Demographics*	Gender	Gender
WDIS individual	*WDIS/WDIS_Demographics*	Birth_date	Birth date
WDIS individual	*WDIS/WDIS_Demographics*	Grad_code_hschl	Highest schooling completed
WDIS individual	*WDIS/WDIS_Demographics*	First_time	First time student
WDIS individual	*WDIS/WDIS_Demographics*	Mil_status	Military status
WDIS individual	*WDIS/WDIS_Demographics*	Asian	Indicator for Asian
WDIS individual	*WDIS/WDIS_Demographics*	Black	Indicator for Black
WDIS individual	*WDIS/WDIS_Demographics*	Native	Indicator for Native American
WDIS individual	*WDIS/WDIS_Demographics*	White	Indicator for White
WDIS individual	*WDIS/WDIS_Demographics*	Ethnicity	Ethnicity
Wage	All files in the *FETPIP-Wage* folder	Emop_uuid	Wage - individual ID
Wage	All files in the *FETPIP-Wage* folder	Emp_num	Employer ID
Wage	All files in the *FETPIP-Wage* folder	Year_quater	Year and quarter
Wage	All files in the *FETPIP-Wage* folder	wages	Wages
SNAP/TANF	*SNAP_TANF/FCS_EMOP_PA_MATCH_THRU2022Q2.csv, SNAP_TANF/FCS_EMOP_PA_MATCH_THRU2022Q2.csv*	snap_yq	Year and quarter
SNAP/TANF	*SNAP_TANF/FCS_EMOP_PA_MATCH_THRU2022Q2.csv, SNAP_TANF/FCS_EMOP_PA_MATCH_THRU2022Q2.csv*	tanf_amt	TANF amount
SNAP/TANF	*SNAP_TANF/FCS_EMOP_PA_MATCH_THRU2022Q2.csv, SNAP_TANF/FCS_EMOP_PA_MATCH_THRU2022Q2.csv*	snap_amt	SNAP amount
Continuing Education	All postsecondary files in the *Continuing_Education* folder	Emop_uuid	Wage - individual ID
Continuing Education	All postsecondary files in the *Continuing_Education* folder	FI_anonymous_id	Unique ID

Intermediate dataset	Raw dataset source	Variable name	Definition
Continuing Education	All postsecondary files in the *Continuing_Education* folder	Ce_year	Continuing Education year
Continuing Education	All postsecondary files in the *Continuing_Education* folder	Ce_term	Continuing Education term
Continuing Education	All postsecondary files in the *Continuing_Education* folder	Ce_school_year	Continuing Education school year
Continuing Education	All postsecondary files in the *Continuing_Education* folder	Ce_quarter	Continuing Education quarter
Continuing Education	All postsecondary files in the *Continuing_Education* folder	Ce_pgm_title	Continuing Education program title
Continuing Education	All postsecondary files in the *Continuing_Education* folder	Ce_cip_code	Continuing Education CIP code
Continuing Education	All postsecondary files in the *Continuing_Education* folder	Cip_code	CIP code
Continuing Education	All postsecondary files in the *Continuing_Education* folder	Continuing_edu_flag	Continuing Education flag
Continuing Education	All postsecondary files in the *Continuing_Education* folder	Completion_year	Completion year
Continuing Education	All postsecondary files in the *Continuing_Education* folder	Full_program_completer	Full program completion flag