

Source Files Download and Preparation Instructions

Data preparation Jupyter notebook : 6414_project_data_preparation.ipynb

The raw source files used in this project are very large and come from multiple federal data portals (College Scorecard, Federal StudentAid, and ACS microdata). Because these datasets are too big and sparse to store directly with the project, they must be downloaded from the official sources by following the instructions below.

After downloading, each source is cleaned, filtered, and aggregated in the Jupyter notebook 6414_project_data_preparation.ipynb. The cleaning code produces compact state-year CSV files:

- scorecard_state_year.csv — state-level financial and sector variables derived from College Scorecard
- acs_state_year_controls.csv — demographic and socioeconomic controls constructed from the ACS microdata extract
- fsa_state_year_annualized.csv — FAFSA application volumes aggregated across the quarterly state-level files

These three cleaned datasets are then merged into a single analysis-ready panel:

- state_year_joined_2011_2020.csv

This joined file is the only dataset used for further analysis.

Dataset 1: College Scorecard

- From US Department of Education
- Includes median debt
- **Time frame:** Annual, 1996–present.

Download link:

<https://collegescorecard.ed.gov/data>[¶]

Download the data that appear on the College Scorecard, as well as supporting data on student completion, debt and repayment, earnings, and more.

This data was last updated April 23, 2025

All Data Files

[Download \(.zip, 390 MB\)](#)

- Institution-level data files for 1996-97 through 2022-23 containing aggregate data for each institution. Includes information on institutional characteristics, enrollment, student aid, costs, and student outcomes.
- Field of study-level data files for the pooled 2014-15, 2015-16 award years through the pooled 2018-19, 2019-20 award years containing data at the credential level and 4-digit CIP code combination for each institution. Includes information on cumulative debt at graduation and earnings one year after graduation.
- Crosswalk files for 2000-01 through 2022-23 that link the Department's OPEID with an IPEDS UNITID for each institution.

Most Recent Institution-Level Data

[Download \(.zip, 22 MB\)](#)

Most Recent Data by Field of Study

[Download \(.zip, 13 MB\)](#)

Steps:

1. click the dark green button with "download(.zip,390MB)" (this file size number changes by time) next to "All Data Files", zip file name is College_Scorecard_Raw_Data_10032025.zip.
2. merge all the MERGEDYYYY_YY_PP.csv files as our combined college scorecard data.

Dataset 2 - Federal StudentAid

Download link:

<https://studentaid.gov/data-center/student/application-volume/fafsa-school-state>

On This Page

FAFSA Data by Demographic Characteristics

FAFSA by Postsecondary School:

FAFSA Data by State

FAFSA Report Definitions

FAFSA Data by State

2025-2026



- [2025-2026 Q3](#)
- [2025-2026 Q2](#)
- [2025-2026 Q1](#)

2024-2025



- [2024-2025 Q7](#)
- [2024-2025 Q6](#)

Steps:

1. Go to page via above link.

2. On the page, find "**FAFSA by State**"
3. Under each year range, such as "2021-2022", click the blue link such as " 2021-2022 Q7", then xls will be downloaded automatically.
4. Click all the links of the desired years to download the data
5. Then use the code below to merge to 1 single csv file.

Dataset 3 :

ACS (American Community Survey)

- Run by the **U.S. Census Bureau**.
- A **large annual survey** of U.S. households (about 3.5 million per year).
Contains: demographics (age, race, education, migration), housing, income, and employment (occupation, industry, wages, commute, telework).
- **Microdata (IPUMS)** = person-level records you can analyze directly (instead of just summary tables).
- Data can be tailored and downloaded upon submitting the request.
- Downloaded version covers the variables as below and range(2010-2023)
- Variables as below:

In cart	Variable	Variable Label	Type	2023 2022 2021 2020 2019 2018 2017 2016 2015 2014 2013 2012 2011 2010											
				Codes	acs										
<input checked="" type="checkbox"/>	<u>YEAR</u>	Census year [<u>preselected</u>]	H	<u>codes</u>	X	X	X	X	X	X	X	X	X	X	X
<input checked="" type="checkbox"/>	<u>HHWT</u>	Household weight [<u>preselected</u>]	H	<u>codes</u>	X	X	X	X	X	X	X	X	X	X	X
<input checked="" type="checkbox"/>	<u>GQ</u>	Group quarters status [<u>preselected</u>]	H	<u>codes</u>	X	X	X	X	X	X	X	X	X	X	X
<input checked="" type="checkbox"/>	<u>STATEFIP</u>	State (FIPS code)	H	<u>codes</u>	X	X	X	X	X	X	X	X	X	X	X
<input checked="" type="checkbox"/>	<u>HHINCOME</u>	Total household income	H	<u>codes</u>	X	X	X	X	X	X	X	X	X	X	X
<input checked="" type="checkbox"/>	<u>PERNUM</u>	Person number in sample unit [<u>preselected</u>]	P	<u>codes</u>	X	X	X	X	X	X	X	X	X	X	X
<input checked="" type="checkbox"/>	<u>PERWT</u>	Person weight [<u>preselected</u>]	P	<u>codes</u>	X	X	X	X	X	X	X	X	X	X	X
<input checked="" type="checkbox"/>	<u>SEX</u>	Sex	P	<u>codes</u>	X	X	X	X	X	X	X	X	X	X	X
<input checked="" type="checkbox"/>	<u>AGE</u>	Age	P	<u>codes</u>	X	X	X	X	X	X	X	X	X	X	X
<input checked="" type="checkbox"/>	<u>RACE</u>	Race	P	<u>codes</u>	X	X	X	X	X	X	X	X	X	X	X
<input checked="" type="checkbox"/>	<u>HISPAN</u>	Hispanic origin	P	<u>codes</u>	X	X	X	X	X	X	X	X	X	X	X
<input checked="" type="checkbox"/>	<u>SCHOOL</u>	School attendance	P	<u>codes</u>	X	X	X	X	X	X	X	X	X	X	X
<input checked="" type="checkbox"/>	<u>EDUC</u>	Educational attainment	P	<u>codes</u>	X	X	X	X	X	X	X	X	X	X	X
<input checked="" type="checkbox"/>	<u>EMPSTAT</u>	Employment status	P	<u>codes</u>	X	X	X	X	X	X	X	X	X	X	X
<input checked="" type="checkbox"/>	<u>INCTOT</u>	Total personal income	P	<u>codes</u>	X	X	X	X	X	X	X	X	X	X	X
<input checked="" type="checkbox"/>	<u>POVERTY</u>	Poverty status	P	<u>codes</u>	X	X	X	X	X	X	X	X	X	X	X

Download Link:

<https://usa.ipums.org/usa.acs.shtml>

Steps:

1. register with GT email, and login.

2. On homepage, click "select data" on the top bar.
3. click "select samples", go to its page.
4. check the year box for ACS (1996 – 2023 this is the range I selected. But we only use 2011 to 2020 in the end) and uncheck the content not desired. then click "submit sample selections" button at the bottom of the page.
5. select the variables, click the "plus" sign in front of the desired variable, then it will be added to the cart.
6. The quick way is using the "search" to find the desired variables and add to cart:
`'YEAR'"HHWT'"STATEFIP'"GQ'"HHINCOME'"PERNUM'"PERWT'"SEX'"AGE'"RACE'"RACED'"HISPAN'"HISPAND'"SCHOOL'"EDUC'"EDUCD'"EMPSTAT'"EMPSTATD'"INCTOT'"POVERTY'`
7. After selecting the wanted variables, click "view cart" button.
8. Review the select vars, then click "create data extract"
9. IPUMS.org will review this extract request shortly.
10. Download the dataset once it's got approved.

All cleaning and joining steps are implemented in 6414_project_data_preparation.ipynb