CDE Dataset

California Department of Education Dataset

Original sources
The original sources of the raw data are public school districts, county offices of education, and charter schools. The information was collected and reported by school administrators, teachers and staff.

How the data was generated

- The school districts and offices of education submit their student, staff and school information into two CDE systems which put together the downloadable files:
    - **California Longitudinal Pupil Achievement Data System:** This collects individual level data on students and staff, including demographics, enrollment, program participation, discipline, course completion and statewide assessment results.
    - **California Basic Educational Data System**: This collects data at the school and district level, such as staff counts, estimated teacher hires, school calendars and kindergarten program types.
- The CDE processes the data from these two systems, verifies it for state and federal reporting, aggregates it to protect student privacy, then releases it to the public as downloadable files.

Funding for the Data Set
Primary funding comes from the State of California through the annual state budget for the operation of the California Department of Education. The U.S. Department of Education may also award grants like the Statewide Longitudinal Data Systems Grant Program.

Information Left Out of the Spreadsheet
Individual student data is not included in order to protect student privacy. The accessible data through these files is aggregated and can only be viewed through counts/rates/averages of certain demographics. The data is primarily quantitative which means there is not a lot of attention paid to the quality of instruction, the narratives and feelings students may have towards the school, or teachers' experiences.

Ideological Effects and Ontology
One key issue we noticed about this dataset is that it includes and sorts by categories such as "absenteeism" and "discipline." Dividing students by these categories can create a certain stigma surrounding their behavior and without their backgrounds to contextualize their situations, it can imply that lower numbers for these categories mean that the school is "succeeding", while higher numbers imply that the school is "failing."

Another core idea that this dataset is missing is the reason why certain demographics are higher than others. For example, we might be able to see that one school has a 10% chronic absenteeism rate, but we would not know the context behind that statistic without doing further research into these students' home life, housing security, etc.

This dataset also fails to address the students' experience at these schools. For example, some schools may seem more diverse in certain categories or have lower rates of absenteeism, but this does not necessarily mean that all students feel safe to attend the school or engage with the community.

ACS Dataset

American Community Survey Dataset

How the data was generated
The American Community Survey is a federal survey drawing samples across the US via internet, mail, telephone, and personal visits.

An accurate Master Address File (MAF) which is updated continuously by the federal government is combined with a Topographically Integrated Geographic Encoding and Referencing (TIGER) database. The MAF contains records which have geographic codes, mailing addresses, physical characteristics, and other geographic features.

ACS uses a filtered subset of the MAF (given that the MAF has hundreds of millions of entries) to try to approximate the most accurate representation of the US. This includes filtering out new construction units, some of which may not even exist yet, and also includes housing units that may not be geocoded (not linked to a census tract yet). ACS also includes units that are "excluded from delivery statistics" to ensure the best coverage. This both ensures statistics are accurate and also that as many underrepresented groups are included in the data as possible.

Housing Units included in the data are randomly selected of 3.54 million which are allocated for the 12 month sampling period. The data is then categorized by race in the resulting data which is downloadable from the census bureau.

Funding for the data set
As the census bureau is a federal entity, it is funded by tax-payer money allocated by the federal government. This also means the census bureau is affected by federal shutdowns.

Information left out of the spreadsheet
Since the census bureau only collects data on responders to the survey, those who don't respond to the survey may be underrepresented unintentionally.

This may happen for example for low income neighborhoods. People may have not have the time to spend on surveys so the data may be underrepresented.

Month specific information is not included, as the ACS database is for 3 yr or 5 yr estimates and are not fine grained temporally.

Ideological effects and ontology
One problem with the census data is that it is purely information essentially of "respondents to ACS median reported income for a 3-5 yr span of time." This excludes important information like how this affects their average life experience, what correlations their geographic location

have with their income, how their income is spent per month (housing, transportation, food, etc.) and how their occupation affects it.

The data also doesn't include any information about their social mobility, but instead we have to try to use the data in combination with other sources to draw conclusions. For example, we may know what the average income is for a particular neighborhood, but determining whether that is because of the cost of rent or because of how close it is to their work is unknown.

BTS Dataset

Bureau of Transportation Statistics Dataset

How the data was generated
Established in 2016, the National Transit Map aggregates fixed route transit service data across the US into a centralized database. Participation in this map is voluntary and in 2016, they had 270 transit agencies submitting their fixed route services to the map.

For our specific use case, the city of Los Angeles has contributions from Los Angeles city, SOCAL Regional Rail Authority, and more. The data contains GTFS files with routes and stops.

Funding for the data set
As the bureau of transportation statistics is a federal entity, it is funded by tax-payer money allocated by the federal government. This also means the bureau of transportation statistics is affected by federal shutdowns.

Information left out of the spreadsheet
Since participation is voluntary, information is not necessarily entirely accurate, as there may be organizations that do not participate in the national transit map and do not submit their transit routes to BTS. The NTM also does not contain information about usage of the public transit lines. This could include info such as how often people use it, how frequent the transit comes and goes, how affected it is by traffic, and other useful statistics.

It also doesn't include information other than mostly geographic information and who owns the lines. This means that things like how safe the lines are for the average rider are not included. In addition, the planning/history of how the routes were developed is not included. For example, the absence of lines through Beverly Hills are caused by pushback from the wealthier neighborhood, none of which is captured in the pure transit data.

Ideological effects and ontology
Since the transit information is essentially just geographic, this information would provide very little insight into the wealth gap and would only serve to say, "this is where most of the public transit is in Los Angeles."

Further, seeing that a transit line exists does not include information such as how much it costs, the infrastructure or additional transit around the transit line, and other important information. The lack of use also affects how useful the information is, since most riders would not use public transit that is slow or heavily affected by traffic, even if the routes exist and are close and nearby.

Fed Reserve Dataset

Federal Reserve Dataset

How the data was generated
This data was collected by the Board of Governors of the Federal Reserve System. It combines two main sources:

- **Survey of Consumer Finances (SCF)**: collects detailed data from U.S. households on income, assets, and debt every three years.
- **Financial Accounts of the United States (FAUS**): records national balance-sheet data from institutions and government sources.

The Federal Reserve merges SCF and FAUS data to estimate quarterly wealth distribution by race, income percentile, and asset type. This process produces consistent national estimates that allow long-term tracking of wealth inequality.

Original sources
The original sources for this data are based on survey responses from households about financial and demographic information, as well as compiled institutional and government financial data to represent the entire U.S. economy. Together, these sources provide both micro-level (household) and macro-level (national) data.

Funding for the dataset
This data is funded and maintained by the Federal Reserve, a publicly financed federal institution. It uses taxpayer dollars and supports the Fed's mission to monitor financial stability and inequality. This data is intended for economists, researchers, and policymakers to understand economic trends.

Information left out of the spreadsheet
This information leaves out geographic data, so regional differences cannot be analyzed. Additionally, there is no data about the subject's occupation, education or family structure to explain why disparities exist, nor breakdowns about what specific types of debt each individual might have accrued. Lastly, the dataset uses very broad racial categories and leaves out groups like Asian Americans, Native Americans and Pacific Islanders.

Ideological effects and ontology
Some of the ideological effects of this data include the fact that inequality is framed as a statistical problem with not much information provided about the historical or social aspects that go into creating the wealth gap. It also excludes historical causes of inequality such as segregation, redlining and labor discrimination. If this dataset were used as the only source, it

would suggest that inequality is purely financial, and not structural due to some of the marginalized categories it excludes in this way.