College Scorecard Institution-Level Data Set

Summary:

- Data Source: This is an external data source. It is provided by the US government, specifically the US Department of Education. The data is trustworthy because it comes from the government.
- Data Collection: This data is collected through federal reporting from institutions, federal financial aid, and tax information. Most comes from data reported to the Integrated Postsecondary Education Data System (IPEDS), which is part of the US Department of Education's National Center for Education Statistics (NCES). It is collected annually from surveys by the NCES. The Higher Education Act requires all institutions that participate in federal student aid programs to complete these surveys. The data about federal student aid comes from the National Student Loan Data System (NSLDS), also a part of the US Department of Education. Finally, earnings data comes from tax records at the US Department of the Treasury.
- Data Contents: This data contains information about US postsecondary institutions, including their characteristics, enrollment, student aid, costs, and student outcomes. The most recent dataset was released July 17, 2020, and covers the Academic Year 2018-19 from the IPEDS data collection year 2018-19, or more recent data (Academic Year 2019-20 from the IPEDS data collection year 2019-20) if available for certain columns. It's likely that data collection was interrupted during the pandemic year, so that's why most of the data covers 2018-19 instead. Note that some columns have data from academic years prior to 2018-19.
- Resources: https://collegescorecard.ed.gov/data/, https://data.ed.gov/dataset/college-scorecard-all-data-files-through-6-2020, as well as the Documentation PDF file (available under Data Documentation section of the first link) and the Data Dictionary Excel file (available under Resources section of the second link)

Data Profiles

Most-Recent-Cohorts-All-Data-Elements.csv → scorecard_data_cleaned.pkl

Data cleaning & consistency checks:

- The original data set contains 2,384 columns, so in order to narrow down my focus, I only imported columns that had the following category in the data dictionary: root, school, admissions, aid, cost, and student. So, the data set I imported had a total of 375 columns and 6,806 rows.

- Dropping columns:

    - I dropped 59 columns having to do with parent PLUS loans because I did not want to use that data.
    - I dropped 2 columns of URLs and 13 columns of old demographic classifications that are no longer used.
    - I dropped 1 column that was only relevant for the 2000-01 academic year.
    - I dropped 15 columns about the costs of program-year institutions that only had missing values.
    - I dropped 15 columns about the costs of other academic year institutions that only had missing values.
    - I dropped 3 columns of squares and average logarithms that only had missing values, as well as 6 columns having to do with number of FAFSAs sent that was no longer updated.
    - I dropped the LOCALE2 column that only had missing values
    - Final number of columns: 260.

- Renaming columns:
    - As much as I would love to have renamed all of the columns for them to make more intuitive, it just wasn't worth it to rename all 260 columns. So I only renamed ones that I thought might be used a lot.
    - Renamed 'INSTNM' (Institution Name) to 'NAME'
    - Renamed 'STABBR' (State Abbreviation) to 'STATE'

- Fixing mixed-type columns:
    - 105 columns were mixed type, likely because there were a lot of 'PrivacySuppressed' observations. I changed all 'PrivacySuppressed' to NaN (missing values), and then changed the columns data types to either string (5) or to float (100). I was unable to change the data type to integer because of all of the NaN observations.

- Missing values:
    - There are a lot of missing values in this data set, with the total count being 668,356. That ends up being about 38% of the data set. However, there are many columns that represent different categories. For example, the data set includes both public and private institutions, and instead of having one column for 'Average net price for $0-$30,000 family income', there are two columns (one for public institutions and one for private institutions). Therefore, all public institutions would have missing values for all columns only relevant to private institutions. Also, there were a number of observations that came from the NSLDS or the Treasury were suppressed for privacy reasons. The creators of this data set also created separate columns for certain elements where they suppressed data involving fewer than 30 students. Also, although institutions are required by law to answer the survey that results in this data set, they are clearly allowed to not provide answers to certain questions.

- Duplicate values:
    - There were no full duplicate values.

Basic descriptive statistics

Rows: 6,806
Columns: 260
Total record count: 1,769,560

Descriptive statistics were calculated for every column in the Jupyter notebook; because there are 260 columns, I will not replicate all of them here. Instead, I will only include certain ones here.

	Min	Max	Mean	Count	Freq. Table
UNITID	100654	49146400	2126810	6806	See Jupyter notebook: all values have frequency of 1
TUITFTE (Net tuition revenue per full-time equivalent student)	0	455440	11099.01	6304	See Jupyter notebook
INEXPFTE (Instructional expenditures per full-time equivalent student)	0	542922	8477.34	6304	See Jupyter notebook
ADM_RATE (Admission Rate)	0	1	0.67	2006	See Jupyter notebook
SAT_AVG	785	1566	1141.17	1298	See Jupyter notebook
DEBT_MDN (Median original amount of loan principal upon entering repayment)	1834	39375	11325.35	5740	See Jupyter notebook
COSTT4_A (Average cost of attendance, academic year institutions)	3990	96375	26956.9	3431	See Jupyter notebook

Limitations and ethical considerations

The biggest limitation is the variation in years being reported on. The data set I'm using was collected in 2020 and supposed to cover the 2019-20 academic year. However due to the pandemic, it seems as though the Department of Education let institutions not respond to their

survey (even though they are legally required to do so). It makes sense because clearly there were a lot more important things to worry about, and also, it's possible that with so many people working remotely from these institutions, they might not have had access to the information needed to complete the surveys. Therefore, according to the documentation, a lot of the columns provide information about academic year 2018-19, or more recent if available. This timeliness issue affects the accuracy of the data and my analysis.

Another limitation has to do with family income levels. That data comes from the NSLDS (National Student Loan Data System), which gets that information from FAFSA applications (Free Application for Federal Student Aid). Parents and students who fill out this application are incentivized to show as little income as legally possible in order to receive the largest amount of aid. Therefore, it's likely that family income level data will appear to be lower than it should be. Also, only parent's income information goes on the FAFSA; no income information about grandparents, foster parents, legal guardians, older siblings, widowed stepparents, or aunts and uncles is available (unless the student was legally adopted by that person, thus technically counting as a parent). The family income level data will be distorted because it doesn't take into account any circumstances where students are raised by someone other than their parents who have a different income level. This is an accuracy issue that affects the data and the analysis.

One smaller limitation is that not all higher education institutions are part of this data set. Only ones that participate in federal aid or student loan programs are required to provide this information. It seems as though there are only around 20 colleges in the US that don't participate and thus are not part of this data set. All of them seem to be rather small, and most of them have strong religious affiliations. Because they are so small, it's likely that most of their data would have been suppressed for privacy reasons if they had been included in this data set.

In terms of ethics, I think it's important to note that the population of students in higher education tends to be whiter and richer than the general population. When looking at demographic and income level data, I need to be aware of that. I also need to be careful about making generalizations based on that data because there are a lot more complicated factors at play that aren't represented here. It would probably make more sense to avoid prescriptions based on my analysis, but rather focus on any issues or abnormalities that should be investigated further by people with more knowledge and experience.

scorecard_data_cleaned.pkl → public_data.pkl

<u>Data cleaning & consistency checks:</u>

- I needed to create a subset with only data about public institutions. The newly created subset had a total of 2,102 rows and 260 columns.

- Dropping columns:
    - Only the following 16 columns were kept: UNITID, NAME, CITY, STATE, ZIP, REGION, LOCALE, COSTT4_A, DEBT_MDN, FAMINC, ADM_RATE, SAT_AVG, UGDS, RET_FT4_POOLED, UGDS_WHITE, COST_CATEGORY

- Missing values:
    - COSTT4_A: 488 rows removed
    - DEBT_MDN: 139 rows removed
    - FAMINC: Imputed median for 4 rows
    - ADM_RATE: Left 916 rows in dataset
    - SAT_AVG: Left 971 rows in dataset
    - RET_FT4_POOLED: Left 843 rows in dataset

- Outliers removed:
    - COSTT4_A: 1 outlier removed
    - DEBT_MDN: 1 outlier removed
    - FAMINC: 2 outliers removed
    - ADM_RATE: 2 outliers removed
    - SAT_AVG: 2 outliers removed
    - UGDS: 16 outliers removed
    - RET_FT4_POOLED: 2 outliers removed

- New columns derived:
    - COST_CATEGORY: Lower cost (COSTT4_A < 10000), Middle cost (COSTT4_A >= 10000 & <20000), or Higher cost (COSTT4_A >= 20000)

Basic information

Rows: 1,449
Columns: 16
Total record count: 23,184

Limitations

This is a small data set, made even smaller by removing so many missing values and outliers.

scorecard_data_cleaned.pkl → private_data_incl_outliers.pkl

Data cleaning & consistency checks:

- I needed to create a subset with only data about private, non-profit institutions. The newly created subset had a total of 2,006 rows and 260 columns.

- Dropping columns:
    - Only the following 17 columns were kept: UNITID, NAME, CITY, STATE, ZIP, REGION, LOCALE, COSTT4_A, NPT4_PRIV, DEBT_MDN, FAMINC, ADM_RATE, SAT_AVG, UGDS, RET_FT4_POOLED, UGDS_WHITE, COST_CATEGORY

- Missing values:
    - Removed 233 rows with missing values for all continuous variables
    - COSTT4_A: removed 397 rows

- New columns derived:
    - COST_CATEGORY: Lower cost (COSTT4_A < 20000), Middle cost (COSTT4_A >= 20000 & < 55000), or Higher cost (COSTT4_A >= 55000)

Basic information

Rows: 1,376
Columns: 17
Total record count: 23,392

Limitations

This is a small data set, made even smaller by removing so many missing values.

scorecard_data_cleaned.pkl → private_data.pkl

Data cleaning & consistency checks:

- I needed to create a subset with only data about private, non-profit institutions. The newly created subset had a total of 2,006 rows and 260 columns.

- Dropping columns:
    - Only the following 17 columns were kept: UNITID, NAME, CITY, STATE, ZIP, REGION, LOCALE, COSTT4_A, NPT4_PRIV, DEBT_MDN, FAMINC, ADM_RATE, SAT_AVG, UGDS, RET_FT4_POOLED, UGDS_WHITE, COST_CATEGORY

- Missing values:
    - Removed 233 rows with missing values for all continuous variables
    - UGDS: removed 153 rows

    - COSTT4_A: removed 63 rows where COSTT4_A is missing and UGDS < 50, imputed median for 170 rows
    - NPT4_PRIV: removed 110 rows (total of missing values and < 0)
    - DEBT_MDN: imputed median for 137 rows
    - FAMINC: imputed median for 37 rows
    - ADM_RATE: left 280 rows in dataset
    - SAT_AVG: left 646 rows in dataset
    - RET_FT4_POOLED: imputed median for 164 rows

- Removing outliers:
    - UGDS: removed 10 rows with 0 students, removed 7 outliers
    - NPT4_PRIV: removed 6 outliers
    - DEBT_MDN: removed 3 outliers
    - ADM_RATE: removed 2 outliers
    - SAT_AVG: removed 3 outliers
    - RET_FT4_POOLED: removed 13 outliers

- New columns derived:
    - COST_CATEGORY: Lower cost (COSTT4_A < 25000), Middle cost (COSTT4_A >= 25000 & < 55000), or Higher cost (COSTT4_A >= 55000)

Basic information

Rows: 1,402
Columns: 17
Total record count: 23,834

Limitations

This is a small data set, made even smaller by removing so many missing values and outliers.

Private_data.pkl & Most-Recent-Cohorts-All-Data-Elements.csv → private_clusters.xlsx

Data cleaning & consistency checks:

- I needed to create an Excel file with the cluster results for private, non-profit institutions, and import latitude and longitude data. The private institution data set had a total of 1,402 rows and 17 columns. Geographic data set import had 6,806 rows and 3 columns.

- Renaming columns:
    - COSTT4_A: AVG_COST
    - NPT4_PRIV: AVG_NET_PRICE
    - UGDS: NUM_STUDENTS
    - RET_FT4_POOLED: RET_RATE
    - UGDS_WHITE: PCT_WHITE

- Missing values:
    - SAT_AVG: removed 634 rows

- New columns derived:
    - Clusters: derived from results of clustering (0, 1, 2)
    - Cluster: derived from clusters column (0 = pink, 1 = light purple, 2 = dark purple)

- Merging data sets:
    - Performed left merge of geographic data onto private dataset.

Basic information

Rows: 768
Columns: 21
Total record count: 16,128

Limitations

This is a very small data set.

Most-Recent-Cohorts-All-Data-Elements.csv & 17 datasets (MERGED20**_**_PP.csv) → nyu_sat_time_series.pkl

Data cleaning & consistency checks:

- I needed to create a data set with SAT_AVG for NYU from 18 different years (18 data sets). Three columns were imported from each data set (UNITID, INSTNM, and SAT_AVG), and only the row for NYU (UNITID: 193900) in each data set was kept and appended to create the main data set.

- New columns derived:
    - DATE: created new column and input relevant start date of academic year

- Dropping columns: once main data set had all relevant data, dropped UNITID and INSTNM columns

Basic information

Rows: 18
Columns: 2
Total record count: 36

Limitations

This is an extremely small data set.