W200 - Project 2 (Dec 2022)

Andrew Higgins, Emily Kenney, Alberto Lopez Rueda

GitHub repository

- https://github.com/UC-Berkeley-I-School/Project2_Kenney_Rueda_Higgins

Dataset

- U.S. College data by Institution. The US Department of Education releases data for U.S. universities to provide insights into the performance of such institutions.
- Our primary dataset will be the "Most Recent Institution-Level Data" which includes the most recent data available for every variable

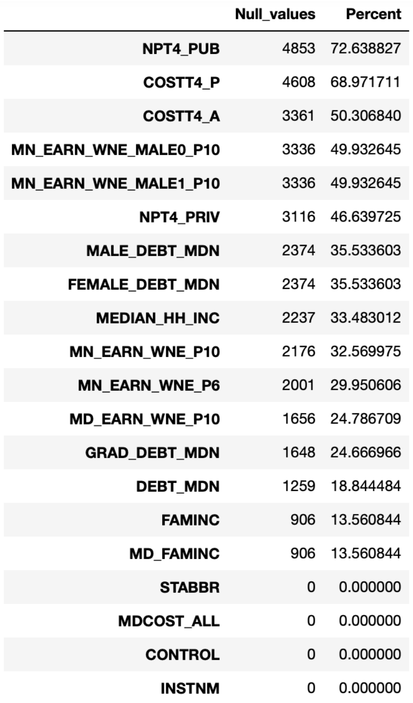
Data structure

- The dataset contains 2,989 variables for 6,681 U.S. universities
- The data variables fall into ten different categories: Academic, Admissions, Aid, Completion, Cost, Earnings, Repayment, Root, School, Student
- Some of the variables we intend to explore include:

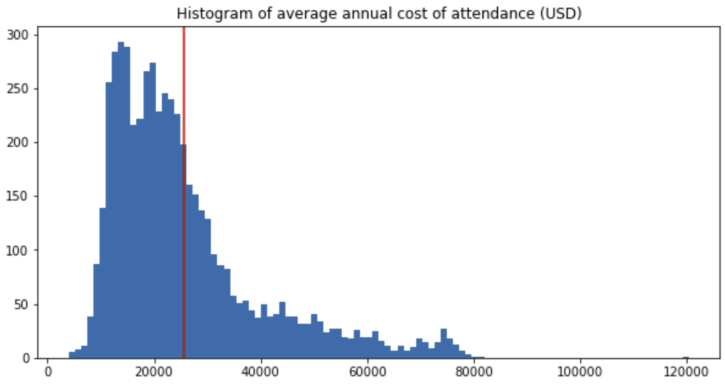
Variable name	Description	Example
INSTNM	Institution name (string)	Alabama A & M University
STABBR	State (string)	AL (Alabama)
CONTROL	Ownership of institution (integer)	1 (Public), 2 (Private)
COSTT4_A, COSTT4_P	Average cost of attendance (integer)	23445
NPT4_PUB, NPT4_PRIV	Average net price for Title IV institutions (integer)	15529
MN_EARN_WNE_P10	Mean earnings of students working and not enrolled 10 years after entry (int)	35500
DEBT_MDN	The median original amount of the loan principal upon entering repayment (int)	15250
GRAD_DEBT_MDN	The median debt for students who have completed (integer)	31000
MALE_DEBT_MDN, FEMALE_DEBT_MDN	The median debt for male or female students (integer)	16500
FAMINC / MD_FAMINC	Average/median family income (float)	32362.826114
MEDIAN_HH_INC	Median household income (float)	49720.22
UGDS_WHITE, UGDS_BLACK & OTHERS	A number of student demographic variables (integers and floats)	0.0159

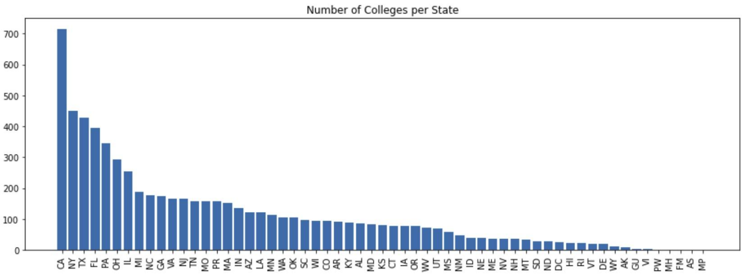
Initial exploration

- Almost all of the variables we intend to use have less than 35% of null values which we consider reasonable. The only exception is the earnings of female and male students after graduation which we keep pending further EDA given the data is still available for ~3,000 institutions. When combining both NPT4 variables, the percentage of null values is c20%. Likewise for COSTT4 variables.
- The distribution of the average annual cost of attendance is right-skewed, with a fatter tail towards expensive universities.
- There is a good distribution of universities by state. The highest concentration of universities is in California, representing only c10% of total.



	Null_values	Percent
NPT4_PUB	4853	72.638827
COSTT4_P	4608	68.971711
COSTT4_A	3361	50.306840
MN_EARN_WNE_MALE0_P10	3336	49.932645
MN_EARN_WNE_MALE1_P10	3336	49.932645
NPT4_PRIV	3116	46.639725
MALE_DEBT_MDN	2374	35.533603
FEMALE_DEBT_MDN	2374	35.533603
MEDIAN_HH_INC	2237	33.483012
MN_EARN_WNE_P10	2176	32.569975
MN_EARN_WNE_P6	2001	29.950606
MD_EARN_WNE_P10	1656	24.786709
GRAD_DEBT_MDN	1648	24.666966
DEBT_MDN	1259	18.844484
FAMINC	906	13.560844
MD_FAMINC	906	13.560844
STABBR	0	0.000000
MDCOST_ALL	0	0.000000
CONTROL	0	0.000000
INSTNM	0	0.000000





What we plan to cover in the final report:

We aim to provide insights into the current state of the U.S. educational system and particularly into two very salient topics: cost of education and student debt. We will do so by exploring question such as:

- **University cost vs. student earnings**
  - Which are the states with the most/least expensive universities? *Variables: STABBR,* COSTT4*, NPT4*
  - Which institutions offer the best return on investment? (i.e., earnings after graduation over university cost). *Variables: INSTNM,* COSTT4*, NPT4, MN_EARN_WNE_P10, MN_EARN_WNE_P6*

        - Does it pay off to attend a private-for-profit university? *Variables: CONTROL,* COSTT4*, NPT4, MN_EARN_WNE_P10, STABBR*
    - Do male students earn more than female students who graduate from the same university? *Variables: INSTNM, MN_EARN_WNE_MALE0_P10, MN_EARN_WNE_MALE1_P10*

- **Student debt**
    - What is the relationship between the level of student debt and the ownership of the university? (e.g., private, public). *Variables: CONTROL, DEBT_MDN,* GRAD_DEBT_MDN, MALE_DEBT_MDN, FEMALE_DEBT_MDN
    - In which states students have the highest level of debt relative to household income? *Variables: STABBR, DEBT_MDN, GRAD_DEBT_MDN, FAMINC, MEDIAN_HH_INC*
    - How are different demographics impacted by debt? *Variables: DEBT_MDN, GRAD_DEBT_MDN and a number of demographic variables such as race (e.g., UGDS_BLACK), sex (e.g., FEMALE), age (e.g., UG25ABV), origins (e.g., UG_NRA, FIRST_GEN) and others (e.g., MARRIED)*