# DS2002 Final Project Report

Rachel Leach, Elizabeth Wu, Adair Hancock, Kendle Schooler, & Anisha Sapkota

## Data Selection

Focusing on income and educational attainment, we analyzed two datasets: institution-level College Scorecard data from the U.S. Department of Education and Social Explorer's American Community Survey data on median earnings from the U.S. Census Bureau. We chose these datasets based on source credibility, large sample size, personal relevance, and potential to produce valuable insights.

The first dataset details characteristics of each institution related to enrollment, student aid, costs, student demographics, and financial outcomes for colleges and universities in the United States during 2022-2023. It also categorizes institutions by state, predominant degree type, region, and institution type.

The second dataset is based on 5-year estimates from the American Community Survey for 2018-2022, which includes demographic information from all 50 states such as age, sex, employment status, and our variables of interest: education and income. The dataset is sufficiently large with a sample size of about 3.5 million and survey information collected almost every day of the year. We filtered this large dataset for specific variables of interest.

Using these datasets, we explored whether a Bachelor's degree is worth the investment by analyzing the impact of education level on income. Specifically, we investigated how the state and institution attended affected median earnings, and how this translates to a return on investment for college education. Comparisons were made between state median earnings for each education level. We also assessed the role of institutional variables such as student population, acceptance rate, and cost of attendance in post-graduation earnings.

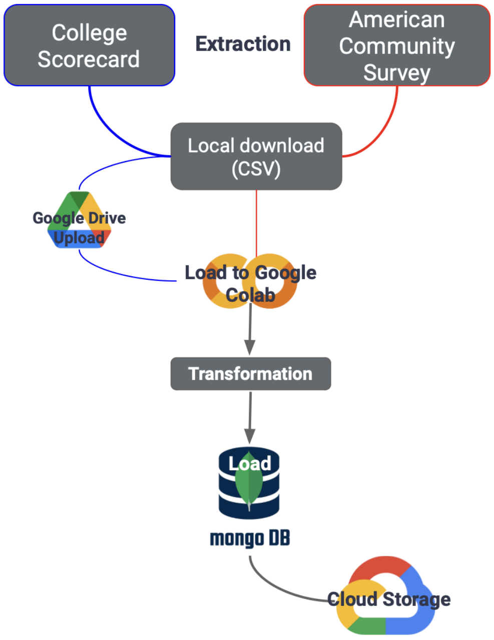
## ETL Process

Both of our data sets are publicly available csv's which were downloaded from their respective websites. The original college scorecard data was too large to store in Github or Excel, so we uploaded it to Google Drive then mounted it from Drive to Collab. For the ACS data, the website has built-in tools to select what variables you want to include in extracted data. We used this to download median earnings by state and education level as a csv file, then upload locally in the Python file.

We used Pandas to load both csv's and convert them to data frames. From there, we transformed the columns to include only the relevant variables. For the institution data, the easiest method was to add relevant columns to a new dataframe then change the column names to make them more intuitive. We included columns for acceptance rate, % of students part-time, median debt after graduation, predominant degree type, % of students receiving Pell Grants, student-faculty ratio, % of students with federal loans, school id, school name, state, zip code, graduation rate, yearly cost of attendance, undergraduate population, institution type, region, median ACT score, percent female students, average family income, and median earnings of employed, not-enrolled graduates 4 years after graduation. Finally, we filtered the data to only include institutions that primarily offer Bachelor's degrees. The data from Social Explorer required minimal cleaning and transformation because the site allowed us to pre-select data. However, we did rename the columns.

We used code to upload the cleaned data as collections of dictionaries to a MongoDB database and also save as new CSV files. To analyze our data, we can now easily access it through MongoDB or the downloaded CSV's.

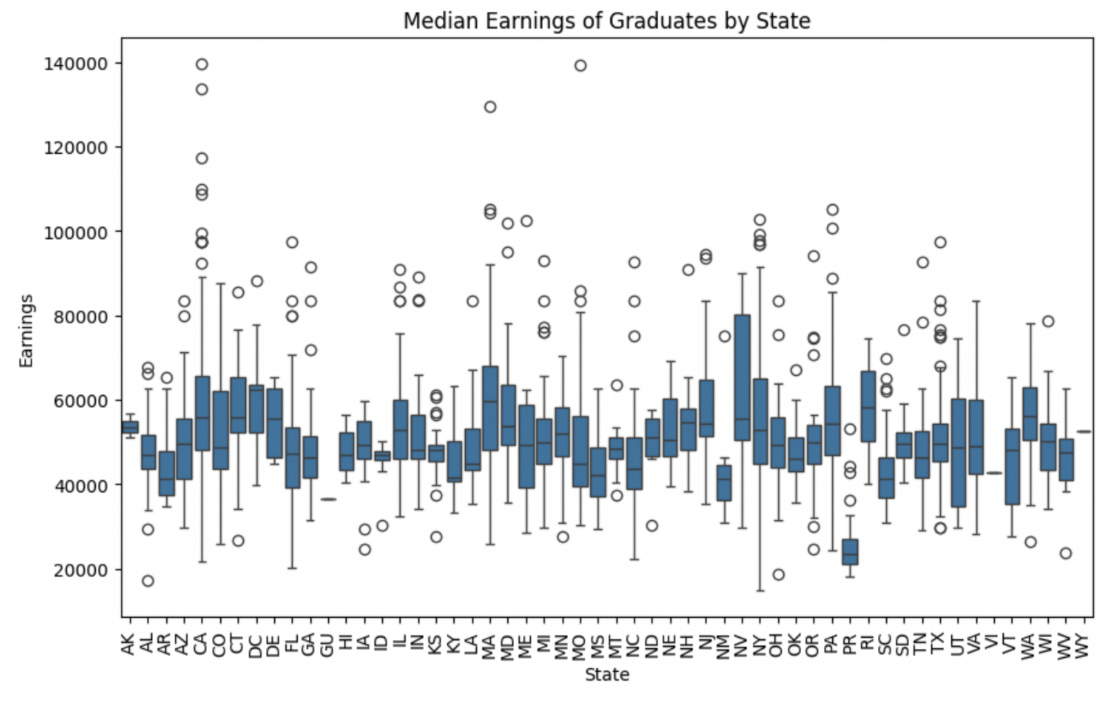
The data's cloud storage location will ideally be highly accessible with low latency and high availability across the US. We will want to limit credentials barriers as much as possible for ease of access. This should not pose any security risks since the data is already public and does not contain sensitive information. When uploading to Google Cloud, we will choose settings accordingly.



Analysis

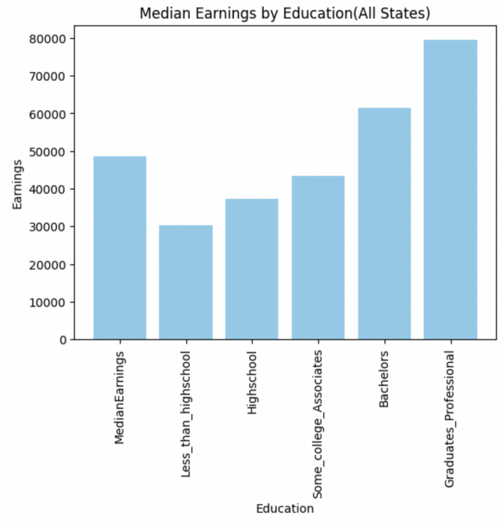
To analyze the data and identify relationships, we produced visualizations and calculated descriptive statistics. In particular, we computed averages, medians, and correlation coefficients to draw inferences. Visualizations included bar graphs, scatterplots, box plots, and heatmaps. Most of these visualizations depicted state data, providing understanding of how income varies by education and location.

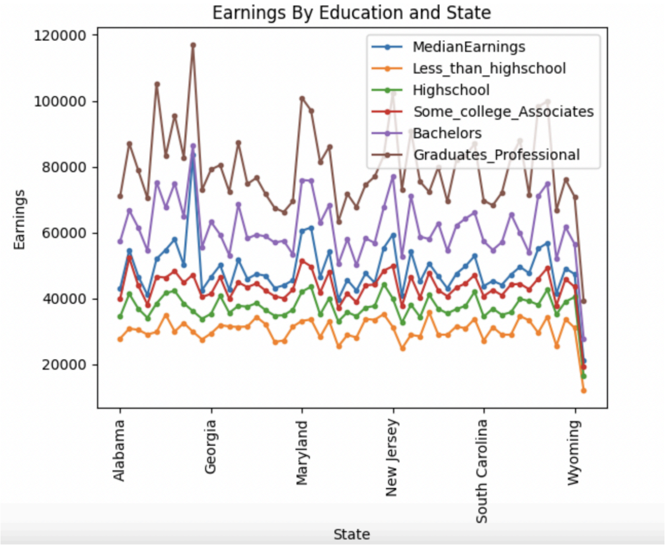
**Median graduate earnings by state:**



We created a box plot using earnings of working graduates (4 years after graduation) by state from the institutional data. The maximums, minimums, and ranges fluctuate vastly by state though the median earnings were largely situated between $40-60,000. Puerto Rico had the lowest median at around $30,000 while DC was double that amount, around $65,000.

**Median earnings by education level:**



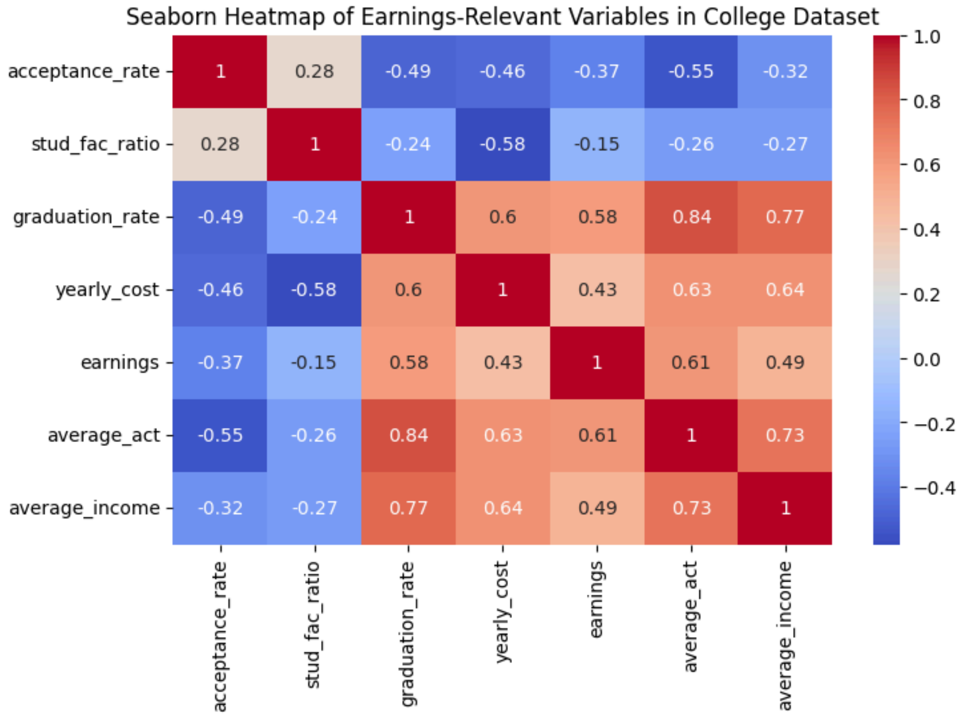


The bar plot was created by taking the average of all state data on earnings at different education levels. As expected, earnings increase nearly linearly with more education, and there is a pronounced jump from Associate’s to Bachelor’s as well as from Bachelor’s to

Graduate/Professional degree. Median earnings at the Bachelor's level surpass overall median earnings, highlighting the financial benefit of obtaining a 4-year degree. Earnings for those with some college education or an Associate's degree were lower than the overall median.

Using the state data, we also created a line plot showing earnings at different education levels for every state. This allowed us to identify similarities and differences across states. We see that median earnings are between Associate's and Bachelor's for every state. However, the exact value and its proximity to Bachelor's vs Associate's varies significantly, likely due to state differences in the population's education. Earnings at each education level vary similarly across most states. We confirmed this by calculating correlations, which showed a positive moderate to strong correlation between all categories. DC has the highest earnings for median, Bachelor's, and Graduates/Professional. New Hampshire has the highest earnings for those with only high school education while Alaska has the highest earnings for those with an Associate's or some college education. Puerto Rico has the lowest earnings for all categories.

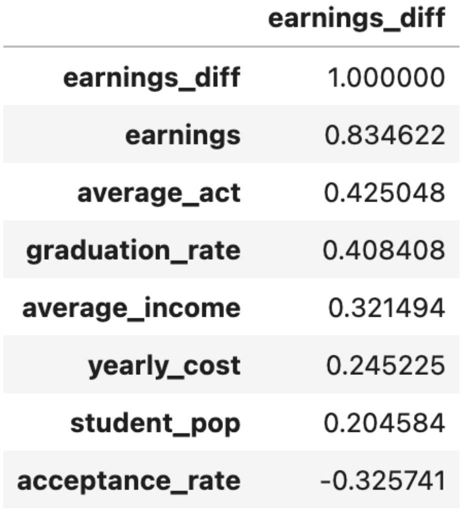
**Institutional factors in graduate earnings:**



This heatmap allowed us to analyze the correlations between multiple factors and graduate earnings by institution. Average ACT score and graduation rate have moderately positive correlation with earnings, reasonably implicating that institutions with good academic performance and higher rates of completion correspond with higher earnings after graduation. Yearly cost and average family income have a weaker but still positive correlation with earnings. This suggests that there could be a possible relationship between expensive colleges and higher earnings of their graduates. Financial background and family support may have an impact on ability to secure higher earnings post-graduation.

More expensive institutions that value academic performance and support students in completing their degree seem to provide the best opportunity for students to attain higher earnings after graduation.

**Graduate earnings by institution compared to state Bachelor's median:**



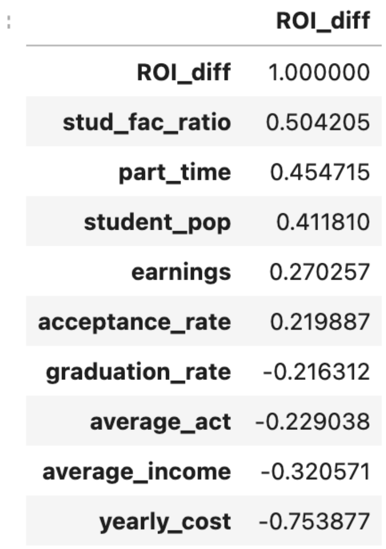
| | earnings_diff |
|---|---|
| **earnings_diff** | 1.000000 |
| **earnings** | 0.834622 |
| **average_act** | 0.425048 |
| **graduation_rate** | 0.408408 |
| **average_income** | 0.321494 |
| **yearly_cost** | 0.245225 |
| **student_pop** | 0.204584 |
| **acceptance_rate** | -0.325741 |

We merged the two datasets and created a new column calculating the difference between institutional graduate earnings and the respective state Bachelor's median earnings. The University for Health Sciences and Pharmacy in St. Louis had the largest difference. Gallaudet University in DC had the smallest difference.

We then calculated correlation coefficients for these values and institutional variables. Earnings had a strong positive correlation, indicating that institutions with higher earnings are typically further from the respective state median for Bachelor's. Unsurprisingly, other relationships thus resembled those we identified earlier in the heatmap. Average ACT score and graduation rate exhibit moderate to weak positive correlations, suggesting that more academically rigorous institutions [as indicated by higher ACT scores] with a higher completion rate generally produce graduates with higher earnings compared to the state median at the Bachelor's level. It would be helpful to incorporate data on what % of these graduates obtained further education, which would contribute to this higher than average income.

In contrast, acceptance rate has a weak negative correlation, suggesting that institutions with higher acceptance rates tend to produce earnings closer to the median. More selective institutions tend to produce higher earnings relative to the state Bachelor's median. This could be due to the characteristics of students these institutions accept, the preparatory skills they instill, or the apparent prestige of a degree from such an institution. It is difficult to determine the impact of confounding variables in this relationship.

**Difference between cost and post-graduation earnings:**



| | ROI_diff |
|---|---|
| ROI_diff | 1.000000 |
| stud_fac_ratio | 0.504205 |
| part_time | 0.454715 |
| student_pop | 0.411810 |
| earnings | 0.270257 |
| acceptance_rate | 0.219887 |
| graduation_rate | -0.216312 |
| average_act | -0.229038 |
| average_income | -0.320571 |
| yearly_cost | -0.753877 |

The data shows the correlation between institutional variables and the return on investment (ROI), or the difference between graduate earnings and the cost of attending an institution. The United State Merchant Naval Academy had the highest return on investment and Oberlin College in Ohio had the lowest.

Yearly cost has a strong negative correlation, showing that less expensive institutions typically confer higher return on investment. This makes sense because lower cost creates greater return on investment if income is the same or higher. This relationship suggests that lower price does not necessarily decrease the benefit of a degree for post-graduation earnings. In contrast, student-faculty ratio has a moderate positive correlation meaning that institutions with a higher student to faculty ratio tend to provide better return on investment to graduates. Looking back at our heat map, there was a moderate negative correlation between yearly cost and student-faculty ratio. This indicates that cheaper institutions tend to have larger class sizes, which would explain the relationship between student-faculty ratio and return on investment.

Cloud Storage

To store our data in Google Cloud, we first created a bucket to hold both datasets. After choosing a name, we set the bucket's settings as follows:

Location: Multi-region(US)

Storage class: standard (default)

Access control: uniform; public access prevention off

We chose multiple US regions for data location because this would provide the highest availability and prevent access issues if someone were in another part of the country. Storage class was left at the default, which is best for short-term storage and frequent access. Public access is permitted so that we can use web hosting and easily access the data, since none of the information is sensitive. Access control was left uniform as we have no need to restrict access by object. All users (public) were given reading permission for objects and datasets so that we, as well as anybody else, can easily access the data from any account.

After setting up the storage bucket, we directly uploaded the CSV's for the cleaned state and college data. This data can now be easily accessed from any Google account and downloaded to use for analysis.