Project Report - College Education

Amanda Balker, Haile Ghebrendrias, James Moua, Ryan Mannon

California State University, Fresno

Database Systems

April 27th, 2025

Introduction to Data:

For our project, we wanted to explore whether attending a university is financially worthwhile given today's job market—a pressing question for many young adults. To analyze this, we selected three interconnected datasets from a Kaggle project by Jesse Mostipak, which examines college tuition, diversity, and graduate salaries. We chose to obtain our datasets from this project as they were incredibly extensive, providing data for almost three thousand colleges in the United States. The specific datasets we chose were "diversity_school," "salary_potential," and "tuition_costs."

Diversity Dataset:

In this dataset, the attributes were name, total_enrollment, state, category, and enrollment. There are 4,575 unique values in this file. This dataset provides key information on the demographics from each college, the total enrollment of students in each college, and the enrollment of the students in different categories. This data will help us explore how the demographics of certain colleges play a part in our topic of choice, allowing examination of potential correlations between institutional demographics and either tuition costs or postgraduate salaries.

Salary Potential Dataset:

There are 934 unique values in the salary potential dataset. The attributes are rank, name, state_name, early_career_pay, mid_career_pay, make_world_better_percent, and stem_percent are included. However, we chose to only use name, state, early_career_pay, mid_career_pay, and stem_percent. The rank showed the potential salary rank within the state, and make_world_better_percent showed the percentage of alumni who believe they're making the world better, both statistics that are unnecessary for the topic we wish to delve into. However, we

did use early_career_pay and mid_career_pay, which represented early career (0-5 years post-graduation) and mid-career (10+ years) median salaries, respectively. Additionally, we used the STEM graduate percentage, an important attribute when analyzing differences in salaries between institutions. Although it does include less information than the tuition costs or diversity datasets, it still provides valuable, in-depth statistics.

Tuition Cost Dataset:

The tuition costs dataset provided the most data out of the three, including 2,938 unique values, as well as ten attributes. The attributes were name, state, state_code, type, degree_length, room_and_board, in_state_tuition, in_state_total, out_of_state_tuition, and out_of_state_total. For this dataset, we chose to exclude state_code and room_and_board. The state_code and room_and_board were unnecessary for our topic since the total tuition cost, as well as the tuition cost without room and board, were provided. This dataset was structured to analyze each college, where the college is located, the type of college it is (such as public, private, or non-profit), the length of the degree (four or two year degree), the cost of tuition in state, tuition and housing in state, the cost of tuition out of state, and the tuition and housing out of state cost. The data is well formatted to consider the costs of attending college in different circumstances, such as the length of the degree or the location.

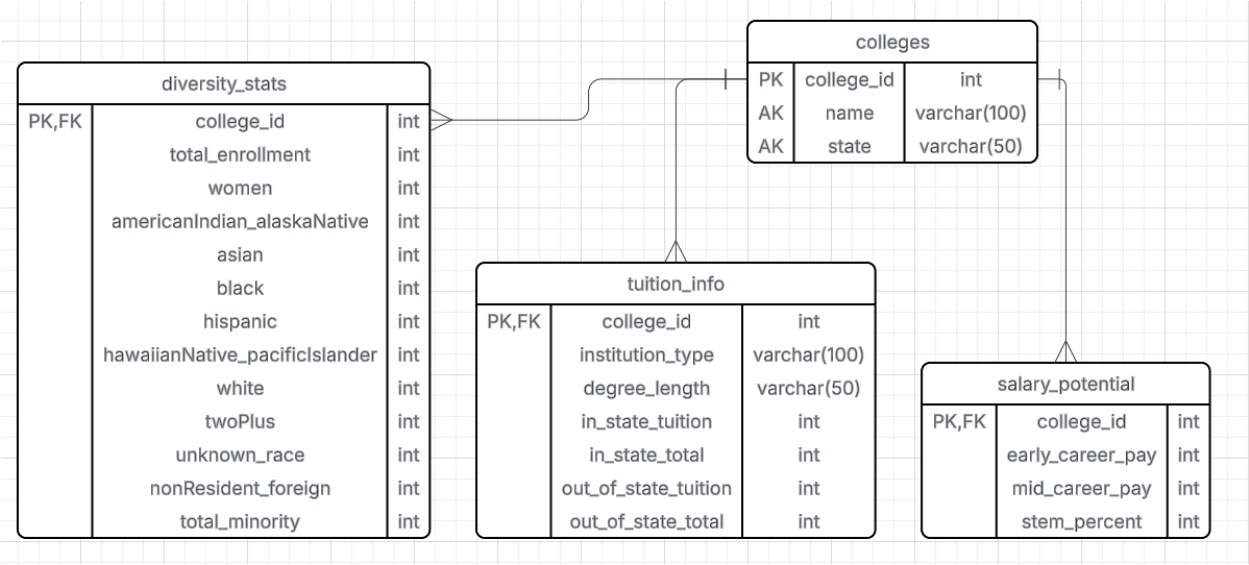
Potential Questions or Insights:

We had a solid starting point for the topic we wanted to explore for our project. After finding these datasets, we were able to think of potential questions we could answer supported by the data. One of our biggest questions was, “Do colleges with higher tuition lead to jobs with better pay?” We wanted to see if the statistics support the idea that a college that is worth more to attend is advantageous over a college that is much cheaper.

Other questions we considered were:

- Which colleges offer the best salary-to-tuition ratios?
- Is there a salary difference between colleges with higher enrollment rates for specific minority groups?
- Which states have the most affordable high-salary colleges?
- Which colleges have the largest salary growth from early to mid-career?
- And many more.

E/R Diagram:



An E/R (Entity-Relationship) diagram is a visual representation of a database's structure. For our project, we designed one to store and organize information about colleges in the United States. The database consists of four entities (main tables): colleges, diversity_stats, tuition_info, and salary_potential. The colleges table serves as the central table, containing basic identifying information about each institution, including a unique college_id, the name of the college, and the state in which it is located. The college_id is set as the primary key (PK), ensuring that each

college record is unique. Additionally, an alternate key (AK) is placed on the combination of name and state to enforce uniqueness for college names with each state.

Connected to the colleges table through foreign keys (FK) are the other three tables, each serving a different aspect of college data. The diversity_stats table stores information regarding student demographics, including the total enrollment numbers and breakdowns across categories such as women, various racial and ethnic groups, and total minority enrollment. Each entry in diversity_stats is linked to a college via college_id, which is both a primary key and a foreign key, ensuring that there is a one-to-one relationship with colleges.

The tuition_info table holds financial information related to the cost of attending each college. This includes data on in-state and out-of-state tuition, total costs for in-state and out-of-state students, as well as the type of institution (public or private) and the length of the degree program (two-year or four-year degrees). Furthermore, in the diversity_stats table, tuition_info uses college_id as its primary key and foreign key, maintaining data integrity across the system.

The salary_potential table records the outcomes that are related to the graduates' earnings, including fields for early career pay, mid-career pay, and the percentage of students graduating in a STEM field. In addition, this table connects to the main entity (colleges) using college_id as both a primary and foreign key. THe inclusion of salary data allows the database to represent institutional characteristics, which provides insight into the long-term value of attending a specific college.

Overall, the E/R diagram displays the well-normalized (3NF) structure of the database, avoiding any instance of redundancy by separating the data into logically distinct tables. With this, it maintains the relationships through properly defined keys. Our approach allows for

consistency of the data extracted from Kaggle, scalability, and flexibility of querying across multiple datasets of our college data (diversity, tuition, career outcomes).

Normalization:

To apply normalization principles to our combined data structure, we needed to make sure we eliminated redundancies, ensure data consistency, and enable complex joins for analysis. We noticed each dataset contained a name (the name of the college) and a state (the state the college is located in) attribute. To reduce redundancies, we created a table "colleges" that would create a college_id for each of the colleges, as well as contain the attributes name and state. This means we can use the college id to define the colleges, instead of creating new name and state attributes to each table.

We also altered the original diversity dataset. We realized that the dataset was showing multiples of the same college to show different diversity statistics. To change this, we created two tables in SQLite Studio, simply to alter the data as its import speed was much faster than MySQL. The first table was diversity_stats, in which we imported the data from the original csv file. The second table was new_diversity_stats, in which we used attributes that represented each category, rather than an attribute called category. Instead of having repeated rows, we added the attributes women, americanIndian_alaskaNative, asian, black, hispanic, nativeHawaiian_pacificIslander, white, twoPlus, unknown, nonResident_foreign, and total_minority. Although this does add more attributes, it avoids redundancy in the data. After completing this, we simply exported the table's data as a csv file, so we could import the data later to MySQL.

We then made sure the datasets passed the first three normalized forms. For the first form, this means that each cell contains an atomic value, or a single, indivisible value. We also ensured

that each table had a primary key. For each of the tables, the primary key was simply the college id, since each college appears only once in each child table, and it uniquely identifies the record in both tables. For the second normalization, we needed to ensure that non-key attributes depend on the whole primary key. We found this to be true in the tables that we created, as none of the non-prime attributes depended on only a part of the primary key. For the third normal form, we needed to ensure that no attributes depend on another non-primary attribute. A simpler way to visualize it is ensuring that none of the columns depend on another column that isn't the college id. We found our tables to fit this criteria as well, successfully applying normalization principles to our schema!

Data Exploration/Conclusions:

After running our queries we came across some interesting results. For example, to answer the question of which college would be good for new graduates, salary-wise, we took a join between the early career salary with the name of the college from the college table. We also ran this with a join for the tuition so we could see the cost relative to the potential benefits. To do this, we joined the `salary_potential` and `tuition_info` tables on `college_id`, then calculated `(early_career_pay - in_state_tuition)` as a new field to determine value. The data shows that SUNY Maritime College has the best ratio with early pay being $74,900 and in-state tuition being only $8,283. Keep in mind that this does not mean that the college has the best early career pay, but instead has the best pay-to-tuition ratio. The best early career pay overall belonged to Samuel Merritt University at $91,200.

This process was repeated for all the queries to give us answers that we found interesting. To find the results for highest pay growth from early to mid-career, we created a new field by subtracting `early_career_pay` from `mid_career_pay`, then ranked colleges by greatest

increase. We found the usual suspects that we expected to be there, like Harvard, Stanford, and other Ivy League schools, but we also noticed some smaller colleges that aren't as well-known, such as SUNY Maritime College with only around two thousand students, and Lehigh University with around 7.6 thousand students.

In addition, we explored colleges based on diversity statistics, joining enrollment data with minority population percentages to find the most diverse institutions. We aggregated minority enrollment percentages by state from the `diversity_stats` table, grouping by public/private status. Public colleges tended to dominate the diversity rankings, with states like California and Texas showing especially high averages. This made sense given their overall population demographics, but it was still interesting to see how much more diverse public universities were compared to private ones in a lot of cases.

As we kept analysing, we found that some two-year colleges actually showed strong mid-career salary numbers too, which was a bit surprising. We filtered to focus specifically on two-year colleges by using the `degree_length` field, then ordered the results by **mid-career pay** in descending order. It challenged the assumption that only four-year degrees lead to high-paying jobs. Another thing we noticed was that private colleges, even though they usually have higher tuition, don't always offer better starting salaries compared to public schools. We joined the **early career pay** data from the salary_potential table with the **college name** and **institution type** from the tuition_info table. After that, we compared **private** vs. **public** colleges, considering that private colleges often have higher tuition we expected them to completely fill the top of the table. However, some affordable public colleges, like SUNY Maritime, clearly held their own or even outperformed much more expensive private schools when it came to early career earnings.

This all leads us back to the question that we first asked in this report, “Do colleges with higher tuition lead to jobs with better pay?” We can definitely say that although most colleges that have high tuition *often* lead to higher salaries, it’s not a guaranteed outcome. There are definitely public colleges and lower-cost options that offer just as good, if not better, starting pay for graduates. Reputation and prestige definitely matter in some cases, but cost should absolutely be part of the decision-making process for students trying to maximize their return on investment.

Overall, our exploration showed that while big-name schools and high tuition can sometimes offer an advantage, like prestige and connections, they aren't the only path to financial success after graduation. Students willing to look beyond the Ivy League and well-known private colleges can find affordable options that offer just as strong career payoffs, and in some cases, even better ones.

Future Work:

While our analysis provided valuable insights into college tuition costs, salary outcomes, and diversity statistics, there are several directions for future work that could help deepen the findings and create even more accurate and usable results.

Major-Specific Comparisons

One major limitation of our current analysis is that we treated each college as a single entity without breaking down differences between majors. In reality, a student’s chosen field of study can have a huge impact on their salary potential after graduation. For example, a computer science graduate and an English major from the same university might have very different early career and mid-career salaries. Future work could involve collecting and analyzing salary data by major within each institution. This would allow for much more personalized recommendations

for prospective students and help them make better-informed decisions about both where and what to study.

Additionally, comparing major-specific outcomes could reveal which colleges offer strong programs in specific fields, even if their overall salary averages are lower. This could change the perception of certain colleges that might be hidden gems for particular industries or career paths.

More Metrics

While we primarily focused on tuition costs, early career pay, and mid-career pay, there are many other factors that contribute to a college's overall value. Future analyses could pull in additional metrics like graduation rates, employment rates after graduation, student satisfaction scores, or even average student debt upon graduation. These extra layers would give a more complete picture of the overall return on investment for a college education.

Including metrics like graduate school enrollment rates, number of research opportunities, or internship placement programs could also add depth to our understanding. It would allow students not only to judge colleges by starting salary but also by longer-term success indicators.

More Data Normalization

Although we joined and analyzed data from multiple sources, there is always room to improve the cleanliness and consistency of the data. In the future, we could focus on more sophisticated normalization techniques. For example, adjusting salary data to account for differences in regional cost of living would allow for fairer comparisons across states. A $70,000 salary in a rural area might go a lot further than $80,000 in a major city like San Francisco or New York City. Normalizing tuition prices based on state, financial aid opportunities, or average

student debt per college would also create a clearer picture of real costs. Improving our data cleaning methods and standardizing missing or inconsistent values would make the results even more reliable.

Trend Analysis

Our current analysis gives a good overview of salary outcomes at one point in time, but salaries, tuition, and demographics can all change significantly over the years. Take the technology industry as an example. In recent years we have seen many layoffs and less job opportunities in our specific field. Tracking these changes over time would reveal important trends that could influence a student's college decision. We could analyze how tuition inflation has impacted different regions, how diversity enrollment percentages have changed over the past decade, or whether salary growth for certain majors is speeding up or slowing down. We could also evaluate the impact of major policy changes, like free community college initiatives or student loan forgiveness programs, on overall outcomes.

Trend analysis would help answer questions like: “Are private colleges becoming a worse value over time?” “Are STEM salaries continuing to outpace liberal arts fields?” “Which states are making the most progress toward affordability and diversity?” These insights would be extremely valuable for future students and policymakers alike, especially if someone chose to run for some office on one or more of these issues.

Closing Statement:

Overall, this project gave us valuable insight into the relationships between college costs, salary potential, and diversity. While we faced some challenges, particularly with cleaning incomplete data and organizing everything in a way that is easy to access and that makes sense, our findings highlighted how thoughtful data analysis can reveal important trends. We believe

studying colleges and the data surrounding education is very important and that it is imperative that we keep gathering data so that we may make even greater insights into such an important field as education.

Team Summary:

As a team, we collectively decided to base our project on college education data, focusing on information related to colleges and universities. Throughout the project, we collaborated closely to ensure all aspects were well-organized and thoroughly covered.

Amanda created the shared drive for the team and organized all the necessary materials needed to begin our project. She was primarily responsible for the implementation of the E/R diagram and contributed significantly to writing the project report, completing a majority of the document. Haile and James focused mainly on developing the presentation slides and determining the queries to run based on the research questions we formulated about college education. Haile contributed more heavily to the introduction and data design slides, while James worked mainly on the data implementation and query results slides. Both also contributed a smaller section to the written report. Ryan worked on the exploration, future work, and closing sections of the project report as well, and played a key role in completing the presentation slides, particularly the design, conclusions and future work.

Overall, through strong collaboration and teamwork, we successfully completed our project focused on college and university information.