ISYE6414 Group Project - Student Debt Analysis Plan

Kaitlyn Goldman, Pui Yee Ng, Wilson Hsu, Christopher J. Haut, Sheikh Fahad

June 2025

1 Problem Description

Higher education in the U.S. has become more accessible, but declining state support and rising tuition have led more students to rely on loans and graduate with substantial debt (Baum, 2017; Heller, 2002). This financial burden often hinders young adults from achieving their financial goals. Although the Higher Education Act of 1965 aimed to make college more affordable, the escalating costs have shifted reliance toward federal and private loans, contributing to the burgeoning student debt crisis.

Despite the financial implications, a college degree is increasingly seen as essential due to evolving job market demands (Robb et al., 2012). Students view education as an investment, weighing its long-term costs against future benefits. However, the perceived financial burden can dissuade students from continuing or completing their studies (Leppel, 2005). Financial stress, influenced by rising college costs and a shift from need-based to merit-based aid, disproportionately affects lower and middle-income families (Heller, 2002; Robb et al., 2012). Demographic factors, such as race, also play a role, with Black students more likely to incur debt due to factors like credit constraints (Braga, 2016). Student loan debt can reduce degree completion rates across various demographic and socioeconomic lines (Zimmerman et al., 2025). The increasing costs can push students towards more affordable institutions (Kang & García Torres, 2021) and influence borrowing patterns across income strata (Looney, 2021). Notably, a significant portion of student debt stems from advanced degrees, with federal loans constituting the majority of graduate student aid(Baum, 2017; College Board, 2024)

A significant concern is the lack of transparency in student aid processes, leading to unawareness regarding loan accumulation, payment schedules, and interest accrual (Johnson, 2012; Shi, 2021). Some students resort to credit cards for daily expenses, compounding their debt (Macy & Terry, 2007). Working while studying, though a means to alleviate costs, can add further burden (Macy & Terry, 2007). High student loan levels are directly correlated with increased financial stress (Zimmerman et al., 2025).

The long-term implications of student debt are profound. Graduating with significant debt can burden individuals, particularly those in lower-paying professions like arts, and teaching (Baum, 1998; Macy & Terry, 2007). Some even alter career plans, opting for higher-paying jobs to accelerate debt repayment (Robb et al., 2012; Shi, 2021). Loan defaults are common among older adults returning to community college or those who do not complete their degrees (Baum, 2017; Zimmerman et al., 2025). Accumulated debt can persist for decades, leading to suboptimal financial decisions and worse overall financial outcomes (Shi, 2021).

Recent research indicates shifts in student debt trends. The College Board (2024) reported a drop in total outstanding federal student loan balance from $1.87 trillion to $1.62 trillion since 2020. Undergraduate reliance on federal loans has decreased, now comprising 24% of all undergraduate financial aid, down from 40% 15 years ago (College Board, 2024). Additionally, student debt varies by field of study, with behavioral sciences majors having the highest median debt ($42,822) and biological/physical sciences majors the lowest ($7,591) for bachelor's degrees (Hanson, 2024). Despite debt, a degree remains advantageous: in 2023, college graduates were half as likely to

be unemployed and earned $1.2 million more over their lifetime than non-degree holders (Federal Reserve Bank of New York, n.d.; U.S. Bureau of Labor Statistics, n.d.).

The student debt crisis is a trillion-dollar issue. Identifying factors affecting total student loan trends can aid in predicting future expenses and assessing repayment affordability. By analyzing debt segmented by state, this research can inform institutional and federal financial planning, helping them adapt to the rising costs of higher education.

2 Review of Previous Efforts

Several prior studies have explored student debt, each offering valuable insights but with specific limitations.

Robb, Moody, and Abdel-Ghany (2012) conducted logistic regression study involving 2,258 survey responses from two universities. They identified student race, school year, and parental income as significant predictors of perceived difficulty in finishing degree due to debt. However, the study's generalizability was limited by its small sample size and institutional specificity, which was too inclusive.

Braun (2016) utilized multiple linear regression to investigate demographic factors influencing total accrued federal student loan for 992 first-time, full-time undergraduates at Bowling Green State University. Outliers with family incomes exceeding $300,000 removed using Mahalanobis distance. The final model found family income, minority status, and parental education to be significant. Despite an overall model significance, its R-squared of 0.072 (adjusted R-squared 0.069), later noted as $\tilde{0}$.007 in critiques, indicated poor explanatory power compared to other studies (0.30-0.40 range). This suggests that while useful for factor identification, the model's predictive performance was weak.

Macy and Terry (2007) analyzed determinants of average student debt using data from nearly 200 schools, categorizing variables as finance, institutional, and demographic. A stepwise-reduced model improved the adjusted R-squared from 0.372 to 0.385. Key predictors included graduation with debt, fees, institutional size, endowment, class size, alumni giving, and Hispanic ethnicity (linked to lower debt). The model was explainable but limited by its narrow set of predictors.

Previous studies highlight the complexity of student debt and the need for a more comprehensive approach due to limitations in scope, generalizability, and model performance. Lessons learned from these studies, regarding outlier elimination (Braun, 2016) and stepwise feature processing for improved explainability (Macy & Terry, 2007), that helped our analysis.

3 Analysis Plan

Our analysis combines elements from the reviewed literature to create a hybrid analytical study. We take advantage of multiple linear regression for prediction and integrate insights from clustering approach by segmenting data by state. The integrated approach aims to identify different factors, including state, that predict overall student debt. The resulting model will be used to predict student loan trends, offering insight for institutional and government financial planning regarding future borrowing affordability. Our primary focus will be on data from 2019 to 2023 to ensure relevance to the current environment.

3.1 Approach #1: Outcome prediction with Multiple Linear Regression and Feature Selection with K-means Clustering

The core of our analysis will involve multiple linear regression. We will employ common regression metrics such as Mean Squared Error (MSE) and R-squared to evaluate model performance. As this dataset has shown the debt amount per state, we can run k-mean clustering to group states with similar loan amount to reduce the noises and run principal component analysis to group certain amount bracket as high-risk, which will bring data insights when we run regression models with other variables. By understanding debt variations across states, we can refine our data segmentation, potentially reduce model generality, and improve precision for our predictive analysis.

3.1.1 Data Sources:

Primary data sources will be the US Census Bureau (United States Census Bureau, 2025)for demographic, income level, and social variables, the College Scorecard (College Scorecard, 2025) and Federal Student Aid (Federal Student Aid, 2025) for student debt amounts. The outcome variable will be the total amount of student debt upon graduation.

Feature Engineering and Selection: To optimize model performance, we will utilize feature engineering and Variance Inflation Factor (VIF) analysis. This will help us identify predictors that offer the greatest signal while mitigating noise and multicollinearity. Specific techniques will include:

- **Binning:** Effectively grouping continuous predictors, such as household income, into discrete categories.
- **Encoding:** Appropriately transforming categorical variables, such as race or gender, into a format suitable for regression analysis.
- **Outlier Treatment:** Leveraging industry-standard practices, such as the Mahalanobis distance (Braun, 2016), to identify and address influential outliers, thus improving the robustness and performance of the model.
- **Stepwise Feature Processing:** Employing methods such as stepwise elimination (Macy & Terry, 2007) to select the most significant variables, balancing model complexity with explanatory power.
- **Model Assumptions:** Our Exploratory Data Analysis (EDA) will be critical for understanding the behavior of the data and ensuring alignment with linear modeling assumptions: linearity, normality of residuals, homoscedasticity (constant variance of residuals) and independence of errors. The anomalies identified during EDA will be corrected or eliminated.

3.1.2 Rationale

1. We think our approach will succeed because the previous efforts did not implement a K-means clustering, where we are reducing the noise and also focusing on the clusters generated instead of focusing on the entire U.S. data. State-level segmentation enhances the precision and context of our findings. We also analyze 2019–2023 data to ensure socio-economic relevance and distinguish our work from prior studies.

2. The main challenge would be data cleaning due to the size and span of our datasets. Considering large size and temporal span of our datasets, we will need to conduct thorough tidying. We will perform data cleansing and exploratory data analysis to isolate and retain only the variables that are truly relevant for our predictive modeling. We also anticipate some computational challenges, again, due to the size and span of the datasets.
3. Alternatives were considered: 1) Logistic regression was ruled out due to the lack of binary variable, such as repayment status. 2) Simple linear regression was not suitable, as student debt is influenced by multiple factors rather than single predictor. So we chose multiple linear regression (MLR) to better capture the complex relationships between student debt amounts, state-level variations, and relevant variables. This approach allows for more comprehensive analysis of the factors contributing to student debt in a broader scale.

3.1.3 Data Wrangling

Data wrangling will be a crucial preliminary step:

- **Column Renaming:** Columns with unique or less readable field mappings in the raw data will be renamed for clarity and consistency.
- **Missing Data Handling:** Missing values will be addressed by imputation or removal, depending on the specific data column and variable type (e.g. handling string incorrectly parsed as floats).
- **Removing duplicates:** Identifying and removing duplicate entries in the datasets.
- **Data Grouping/Categorization:** Variables will be grouped or categorized to enhance model performance. For example, gender may be a static factor that does not directly impact debt value but is useful for segmentation, while household income may require binning.

3.1.4 Exploratory Data Analysis (EDA)

Initial EDA will provide a holistic understanding of the datasets, identifying patterns, relationships, and potential issues. We will create histograms, box plots, scatter plots, multivariate analysis, and correlation analysis to help determine if the chosen features are truly relevant to our predictive modeling with Multiple Linear Regression.

3.2 Limitation and Assumptions

Assumptions: We assume that students studied and worked in the same location. Furthermore, this research assumes that the cost of education is relatively similar for students within the same state. This simplification aims to manage edge cases related to diverse student backgrounds and reduce the complexity introduced by external factors not explicitly provided in the dataset.

Limitations: The primary limitation of this analysis is that the dataset may not directly reflect all granular social and economic factors pertinent to each individual student. Additionally, like many student debt research efforts, the inherent complexity and variety of factors contributing to the student debt crisis mean that any single analysis may be limited in providing a fully conclusive solution. We aim to mitigate this by employing robust feature processing and selection.

References

Baum, S. (1998). Life after debt: Results of the national student loan survey [Publisher unknown].

Baum, S. (2017). What colleges should know about students' borrowing patterns. *The Chronicle of Higher Education*, *63*(27), A47–A47.

Braga, B. (2016). *Racial and ethnic differences in family student loan debt*. Urban Institute.

Braun, T. P. (2016). *Demographic predictors of accrued undergraduate federal student loan debt* [Ph.D. Dissertation]. ProQuest Dissertations & Theses.

College Board. (2024). *Trends in student aid 2024* [Accessed June 20, 2025]. College Board Research. https://research.collegeboard.org/media/pdf/Trends-Student-Aid-2024-presentation.pdf

College Scorecard. (2025). *Data home — college scorecard* [Accessed June 20, 2025]. U.S. Department of Education. https://collegescorecard.ed.gov/data

Federal Reserve Bank of New York. (n.d.). *The labor market for recent college graduates* [Accessed June 20, 2025]. https://www.newyorkfed.org/research/college-labor-market#--:explore:wages

Federal Student Aid. (2025). *Federal student aid* [Accessed June 20, 2025]. https://studentaid.gov/data-center/student/portfolio#servicer-portfolio-by-loan-status

Hanson, M. (2024). *Student loan debt by major [2024]: Highest + lowest average debt* [Education Data Initiative. Accessed June 20, 2025]. https://educationdata.org/student-loan-debt-by-major

Heller, D. E. (2002). *Who should we help? the negative social consequences of merit scholarships* (Publisher unknown).

Johnson, C. L. (2012). *Do new student loan borrowers know what they are signing? a phenomenological study of the financial aid experiences of high school seniors and college freshmen* [Ed.D. Dissertation]. ProQuest Dissertations & Theses.

Kang, C., & García Torres, D. (2021). College undermatching, bachelor's degree attainment, and minority students. *Journal of Diversity in Higher Education*, *14*(2), 264–277. https://doi.org/10.1037/dhe0000145

Leppel, K. (2005). College persistence and student attitudes toward financial success. *College Student Journal*, *39*(2), 223–241.

Looney, A. (2021). *Biden is right: A lot of students at elite schools have student debt* [Brookings Institution. Accessed June 20, 2025]. https://www.brookings.edu/opinions/biden-is-right-a-lot-of-students-at-eliteschools-have-student-debt/

Macy, A., & Terry, N. (2007). The determinants of student college debt. *Southwestern Economic Review*, *34*, 15–25.

Robb, C. A., Moody, B., & Abdel-Ghany, M. (2012). College student persistence to degree: The burden of debt. *Journal of College Student Retention: Research, Theory & Practice*, *13*(4), 431–456. https://doi.org/10.2190/CS.13.4.b

Shi, M. (2021). *Essays on debt aversion* [Ph.D. Dissertation]. ProQuest Dissertations & Theses A&I [Order No. 30540334]. https://www.proquest.com/dissertations-theses/essays-on-debt-aversion/docview/2838438872/se-2

United States Census Bureau. (2025). *Acs demographic and housing estimates* [Accessed June 20, 2025]. https://data.census.gov/table?q=DP05&g=010XX00US$0500000

U.S. Bureau of Labor Statistics. (n.d.). *Unemployment rates for persons 25 years and older by educational attainment* [Accessed June 20, 2025]. https://www.bls.gov/charts/employment-situation/unemployment-rates-for-persons-25-years-and-older-by-educational-attainment.htm

Zimmerman, T. S., Jones, F. R., Davis, A. M. D., & Dear, C. (2025). Loans in the long game: How student debt affects financial stress post-graduation. *The Journal of Student Financial Aid*, *53*(3). https://doi.org/10.55504/0884-9153.1809