A black rectangle with a black background

Description automatically generated with low confidenceLogo, company name

Description automatically generated

DAEN 690

Project Report

James Moll

Josh Kloc

Mouli Chittabathina

Pragathi Tummala

Poorna Sai Veera Harish Pedapudi

Sai Charan Doniparthi

Spring 2024

Decrypting the Cost of Attendance

**About the Cover**

Professor Berlin is an instructor at the George Mason University College of Engineering and Computing, Volgenau School of Engineering, MS Data Analytics Engineering (DAEN) program. He began working with the DAEN program as an adjunct faculty member in 2012 and became a fulltime faculty member in 2016. He is a passionate contributor to the program and a devoted mentor to his students.

His passion for new value creation is built on over 50 years of professional experience – innovating and advocating for innovators applying leading-edge digital solutions to mission challenges. He has served with outstanding teams in various roles, including senior strategy executive, consultant, and mentor; applied information and systems technologist; collaborative leader; computer scientist, and public policy entrepreneur.

He serves as a strategy advisor and mentor to public and private sector innovators and entrepreneurs and as a public speaker (emerging challenges, innovation opportunities, and ethics). His core interests include public policy, high-performance computing, cyber, emerging big data, health informatics, and digital economy and governance challenges.

In addition to teaching and mentoring, Professor Berlin seeks new engagements with high-quality, core-value-centered innovation teams – collaborating to address societal and market challenges with cyber-physical and policy innovation. Specifically, sustainable solutions can be delivered at the intersection of innovative value creation, human aspiration, and strategic vision.

Professor Berlin is a graduate of the US Air Force Academy with a Bachelor of Science in Computer Science and Mathematics as well as a graduate of the University of Texas at Austin with a Master of Arts in Computer Science. He is also attended the USAF Air Command and Staff College for leadership and strategy training.

Contents

Table of Contents

[Section 1: Problem Definition 3](#_Toc165409053)

[1.1 Background 3](#_Toc165409054)

[1.2 Problem Space 8](#_Toc165409055)

[1.3 Research 9](#_Toc165409056)

[1.4 Solution Space 11](#_Toc165409057)

[1.4.1 Solution Approach 11](#_Toc165409058)

[1.4.2 Solution Space 11](#_Toc165409059)

[1.5 Project Objectives 11](#_Toc165409060)

[1.6 Primary User Stories 12](#_Toc165409061)

[1.7 Product Vision 12](#_Toc165409062)

[1.7.1 Scenario 1 12](#_Toc165409063)

[1.7.2 Scenario 2 12](#_Toc165409064)

[1.7.3 Scenario 3 12](#_Toc165409065)

[Section 2: Datasets 13](#_Toc165409066)

[2.1 Overview 13](#_Toc165409067)

[2.2 Field Descriptions 13](#_Toc165409068)

[2.3 Data Context 16](#_Toc165409069)

[2.4 Data Conditioning 18](#_Toc165409070)

[2.5 Data Quality Assessment 18](#_Toc165409071)

[2.6 Other Data Sources 21](#_Toc165409072)

[2.7 Storage Medium 21](#_Toc165409073)

[2.8 Storage Security 21](#_Toc165409074)

[2.9 Storage Costs 22](#_Toc165409075)

[Section 3: Algorithms & Analysis 22](#_Toc165409076)

[3.1 Solution Approach 22](#_Toc165409077)

[3.1.1 Systems Architecture 22](#_Toc165409078)

[3.1.2 Systems Security 23](#_Toc165409079)

[3.1.3 Systems Data Flows 24](#_Toc165409080)

[3.1.4 Algorithms & Analysis 25](#_Toc165409081)

[Section 4: Visualizations 28](#_Toc165409082)

[4.1 Overview 28](#_Toc165409083)

[4.2 Visualizations 29](#_Toc165409084)

[Section 5: Findings 30](#_Toc165409085)

[Section 6: Summary 34](#_Toc165409086)

[Section 7: Future Work 34](#_Toc165409087)

[Appendix A: Glossary 37](#_Toc165409088)

[Appendix B: GitHub Repository 38](#_Toc165409089)

[Overview 38](#_Toc165409090)

[GitHub Repository Link 38](#_Toc165409091)

[GitHub Repository Contents 38](#_Toc165409092)

[Appendix C: Risks 39](#_Toc165409093)

[Sprint 1 Risks 39](#_Toc165409094)

[Sprint 2 Risks 40](#_Toc165409095)

[Sprint 3 Risks 41](#_Toc165409096)

[Sprint 4 Risks 42](#_Toc165409097)

[Sprint 5 Risks 43](#_Toc165409098)

[Appendix D: Agile Development 45](#_Toc165409099)

[Scrum Methodology 45](#_Toc165409100)

[Sprint 1 Analysis 45](#_Toc165409101)

[Sprint 2 Analysis 45](#_Toc165409102)

[Sprint 3 Analysis 46](#_Toc165409103)

[Sprint 4 Analysis 46](#_Toc165409104)

[Sprint 5 Analysis 46](#_Toc165409105)

[Works Cited 47](#_Toc165409106)

Table of Figures

[Figure 1 – Median Earnings by Educational Attainment, 2022 3](#_Toc165409107)

[Figure 2 - Unemployment Rate by Educational Attainment, 2022 4](#_Toc165409108)

[Figure 3 - Total Undergraduate Enrollment 5](#_Toc165409109)

[Figure 4 - College Enrollment Rate 5](#_Toc165409110)

[Figure 5 - Non-Housing Consumer Debt 6](#_Toc165409111)

[Figure 6 - Student Loan Debt vs. Problematic Drinking and Anxiety by SES Instability 7](#_Toc165409112)

[Figure 7 - After Loans & Aid, Pell Students Face Gap of One-Third of Cost of Attendance 9](#_Toc165409113)

[Table 1: Data Quality Assessment 21](#_Toc165409114)

[Figure 8 - System Architecture 23](#_Toc165409115)

[Figure 9 - Data Flow Diagram 25](#_Toc165409116)

[Figure 10 - EPI Estimate Methodology 26](#_Toc165409117)

[Figure 11 - EPI Median Costs 26](#_Toc165409118)

[Figure 12 - EPI Data - Unadjusted 27](#_Toc165409119)

[Table 2: Adjustments to EPI Costs 28](#_Toc165409120)

[Figure 13 - Distribution of Institutions by State 29](#_Toc165409121)

[Figure 14 - Tableau Dashboard 30](#_Toc165409122)

[Figure 15 - Living Expense Variance by Percent 31](#_Toc165409123)

[Figure 16 - Living Expense Grading by Percent 31](#_Toc165409124)

[Figure 17 - Living Expense Variance by Dollars 32](#_Toc165409125)

[Figure 18 - Living Expense Grading by Dollars 33](#_Toc165409126)

[Figure 19 - Living Expense Variance by Institution Type 33](#_Toc165409127)

[Table 3: Glossary Table 37](#_Toc165409128)

[Table 4: Sprint 1 Risks 39](#_Toc165409129)

[Table 5: Sprint 2 Risks 41](#_Toc165409130)

[Table 6: Sprint 3 Risks 42](#_Toc165409131)

[Table 7: Sprint 4 Risks 43](#_Toc165409132)

[Table 8: Sprint 5 Risks 44](#_Toc165409133)

[Figure 20 - Sprint Project Dates 45](#_Toc165409134)

**This page intentionally left blank**

Abstract

There are more costs for higher education than tuition and related scholastic fees for an incoming student. These additional costs are difficult to quantify, poorly communicated, and can result in a mismatch between financial aid and the true costs of attendance. Insufficient financial aid can lead to a student quitting school, while too much aid can burden the student with unnecessary debt. Previous research has explored the lack of meaningful federal policy guidance concerning cost estimations and how they vary widely between institutions. This paper attempts to find the true cost of attendance to better align financial aid to student needs. We developed an independent cost estimate based on EPI cost of living data and compared that to IPEDS data reported by institutions to explore the cost disparities. The data was visualized with a Tableau dashboard which will be made available to the public. Our findings show institution estimations vary widely, even between similar schools within the same state. This paper and its accompanying dashboard introduce a methodology to estimate living expenses designed to offer decision-makers—whether students, parents, policy makers, or financial aid officers—a more accurate calculation that transparently outlines all associated costs. With this knowledge, students can plan appropriately for the financial impact of their education, financial aid officers can tune their federally mandated cost estimates to better aid incoming students, and policy makers can investigate the discrepancies highlighted to adjust legislation addressing cost estimation accuracy.

**This page intentionally left blank**

Report

# Problem Definition

## Background

Historically, college was seen as a path to a more stable professional and economic future for students rather than entering the workforce immediately after finishing high school. While a college education is not a guarantee of a more prosperous and financially stable life, the data shows there are considerable socioeconomic benefits linked to higher education.

In May of 2023, the Bureau of Labor Statistics (BLS), whose mission is to measure various economic and labor related activity within the U.S., published a study [1] that reported median weekly earnings by education level. This study highlights the considerable link between education level and higher levels of compensation. Additionally, this study also examined the unemployment rates against these same levels of education. The BLS study shows favorable relationships exist; higher education levels correspond to higher pay and lower rates of unemployment as seen in Figure 1 – Median Earnings by Educational Attainment, 2022 and Figure 2 - Unemployment Rate by Educational Attainment, 2022. Each of these economic and educational elements factor into and boost socioeconomic status (SES), a concept that captures the holistic interplay between income, education, economic security, along with more subjective elements related to social class and strata.

A graph with green and white text

Description automatically generated

Figure – Median Earnings by Educational Attainment, 2022

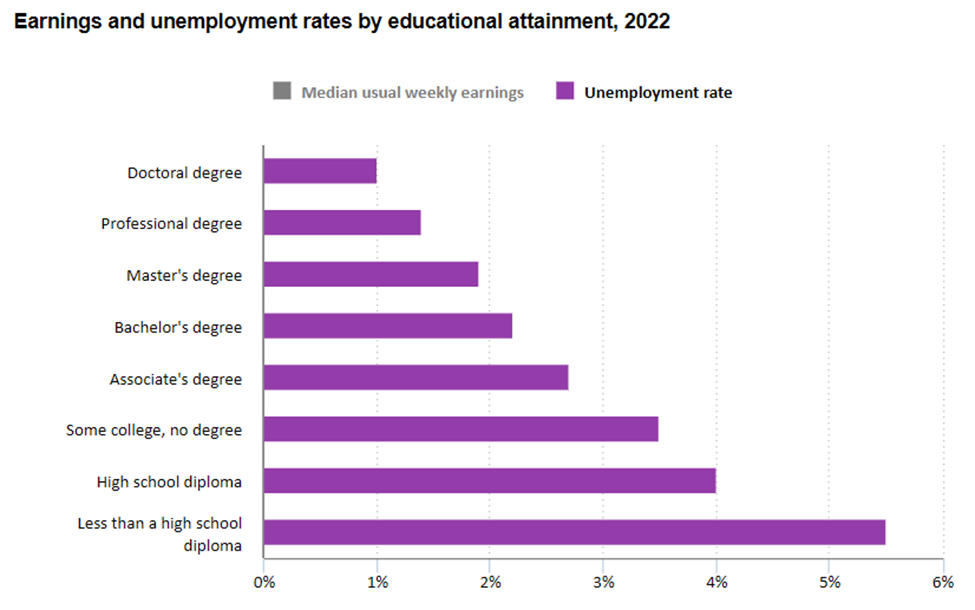


Figure - Unemployment Rate by Educational Attainment, 2022

These benefits do not come without costs. Since the early 1990’s, the listed costs for higher education have nearly doubled for private four-year colleges and more than doubled for four-year public schools, even after adjustments for inflation are made [2]. The costs for schooling are considerable and as Paul Tough writes “the average net price for private-college students is about $33,000 a year; at public institutions, it is about $19,000.” The increase in costs corresponds to changes in the perceived value of pursuing an advanced degree. Polls in the 2010’s show strong support for the value of continued schooling. In one poll, 86% of college graduates considered their education a good investment. Another survey of young adults revealed that 74% thought college education was very important [2]. More recent polls and surveys reflect a significant negative shift in opinion. A survey from the early 2020’s found only 41% of those surveyed felt that a college degree was very important [2]. In another survey from March 2023, conducted by The Wall Street Journal and The University of Chicago, respondents were asked if a four-year college degree was worth the cost. 56% said that it was not, with the primary reason being that students graduate with a large amount of debt.

These changes of perception of higher education can also be seen in enrollment trends, where fewer students enter college. This is the case both for the number of students enrolled, and for the proportion of students who start college after graduating from high school. 2010 saw the highest enrollment numbers with more than 21 million undergraduate students. From that peak, the numbers have steadily declined to 18.7 million in 2021. Likewise, the ratio of students starting college directly after high school was highest in 2009, at more than 70%. As of 2021 data, that ratio sits below 62%, a value not seen since 2001. The data is visualized in the following plots, Figure 3 - Total Undergraduate Enrollment and Figure 4 - College Enrollment Rate [3].

A graph showing the fall of enrollment

Description automatically generated

Figure 3 - Total Undergraduate Enrollment

A graph showing the growth of the company's income

Description automatically generated

Figure 4 - College Enrollment Rate

Even though enrollments have trended downward, college borrowing, and related debt has exploded over the past several years. Student loan debts were estimated to be $500 billion in 2007 [2]. The value of outstanding loans has grown by more than three times, estimated to be more than $1.7 trillion as of 2022 [4]. For student loans originating between 2010 and 2019, more than half of borrowers owe more than the initial amount [2]. 52% of borrowers with student loan debt expressed the feeling that their education was not worth the cost [4]. The growth of this loan debt, along with other types, can be seen in Figure 5 - Non-Housing Consumer Debt [5].

An even more devastating scenario is one where students have taken loans but ultimately do not graduate. Some estimates suggest that around 40% of students who start college do not graduate [2]. Another study measured the population of individuals who have left post-secondary education without any credentials at nearly 41 million [6]. People that fall into this ‘some college but no degree’ category are often the hardest hit; they have no credentials to reflect their time in school but still must deal with the costs of education. Federal Reserve data indicate that this group of people end up in a worse financial situation than those that never went to college [2].

A graph of different colored lines

Description automatically generated

Figure - Non-Housing Consumer Debt

Beyond the connection between education, earnings, socioeconomic status, and student loans, there are also relationships between one’s economic condition and overall well-being. Many groups of researchers [7] [8] [9] have studied and acknowledged the relationship that is present between SES and health. Mackenbach et al. [7] find that lower levels of health quality “were substantially higher in groups of lower socioeconomic status.” Other researchers [9] state “those of low ‘socio-economic status’ have been found to have increased risk of poor mental health, depression, poor physical health, and even death.”

These researchers expressly studied the link between health and debt. More specifically, their focus was on a certain category of debt, those that are unsecured. Unsecured debt refers to types of debt that do not have any possessable collateral against it. Examples include credit card debt, medical debt, and student loans. In contrast, car or home loans are examples of secured debt as there is some real asset or property or backing the loan.

The authors [9] find “…the results suggest that unsecured debt increases the risk of poor health, with some studies showing a dose–response effect with more severe debts being related to more severe health difficulties” but they also caution that “Most current studies simply show a relationship between health and debt, though which effects which is unclear. It might be that, for example, debt induces symptoms of depression. However, it might also be that those who are depressed are more prone to debt due to greater levels of unemployment or poor financial management.” In other studies, [4] [10] researchers specifically investigated the connections between student loan debt and physical and mental health, broken out by different SES groupings. The authors [4] identified that “Student loan debt also negatively affects well-being and mental health, even after controlling for other types of debt, assets, income, and demographic factors.” In reference to less stable SES conditions the findings [4] also show that such instability “had a consistent link with problematic drinking and mental health symptoms: even after controlling for student debt, SES, and monthly income.” The relationships between student loan debt, SES instabilities, and either alcohol use or anxiety is conveyed in Figure 6 - Student Loan Debt vs. Problematic Drinking and Anxiety by SES Instability. These plots indicate that a less stable SES is linked to more negative alcohol use and anxiety scores. Alcohol use is captured by the AUDIT score, an acronym of Alcohol Use Disorder Identification Test. This test, developed by the World Health Organization (WHO), helps to identify unsafe drinking behavior by assessing multiple factors, such frequency of consumption and levels impairment [11].

In other studies, [4] [10] researchers specifically investigated the connections between student loan debt and physical and mental health, broken out by different SES groupings. The authors [4] identified that “Student loan debt also negatively affects well-being and mental health, even after controlling for other types of debt, assets, income, and demographic factors.” In reference to less stable SES conditions the findings [4] also show that such instability “had a consistent link with problematic drinking and mental health symptoms: even after controlling for student debt, SES, and monthly income.” The relationships between student loan debt, SES instabilities, and either alcohol use or anxiety is conveyed in Figure 6 - Student Loan Debt vs. Problematic Drinking and Anxiety by SES Instability. These plots indicate that a less stable SES is linked to more negative alcohol use and anxiety scores. Alcohol use is captured by the AUDIT score, an acronym of Alcohol Use Disorder Identification Test. This test, developed by the World Health Organization (WHO), helps to identify unsafe drinking behavior by assessing multiple factors, such frequency of consumption and levels impairment [11].

A graph of student loan debt

Description automatically generated

Figure - Student Loan Debt vs. Problematic Drinking and Anxiety by SES Instability

A college education can provide improved career prospects, earning potential, and other socioeconomic status benefits to students. These benefits are often touted by institutions as reasons to attend and study in their programs. However, there are costs of pursuing higher education that are not nearly as advertised and are often unclear. Some costs are unambiguous, such as tuition, books, and related fees. Then there are costs that are more nebulous such as housing, meals, transportation, health care, supplies, and other personal related expenses. On average, these additional cost categories can make up 60% of the total cost to attend school [12]. Even when colleges publish approximations of attendance costs, they may not reflect the true costs that students will face. An influential study in 2017 [13] highlighted a concerning disparity: almost half of all colleges established living-cost allowances that deviated by at least 20% from local expenses estimated by the Massachusetts Institute of Technology (MIT) Living Wage Calculator [14].

Beyond traditional costs, there are also risks. Some students are unable to continue school due to insufficient aid, but still have the burden of debt from school. Another related example is graduating with too much debt. Both situations have long-term economic consequences that can affect the success and happiness of the individual, as reflected by the data and various polls.

In short, a college education is likely to benefit students but by no means is it a sure bet, in part due to unclear cost of attendance information from these institutions.

## Problem Space

The awards for most of the FSA programs rely on some form of financial need. This is unlike scholarship programs that may award funds based on how well the student performs academically or what they study. “Need-based” grants, loans, and work-study depend on the student’s demonstrated financial need for support. The COA is the key factor in determining a student’s financial need, as it defines the maximum amount of aid that a student can get for the Campus-Based, TEACH Grant, and Direct Loan programs, and is one of the main elements of the Pell Grant computation. [15]

The guidelines by which an academic institution derives a COA are only vaguely addressed by the Department of Education. “There are a variety of methods to arrive at average costs for your students, such as conducting periodic surveys of your student population, assessing local housing costs or other pertinent data, or other reasonable methods you may devise which generate accurate average costs for various categories of students.” [15] This leaves a lot of room for interpretation and allows for a variety of methods that do not accurately reflect the cost of attending a 2 or 4 year college.

The crux of our initiative lies in addressing a pivotal issue within higher education finance—the complexity and confusion of the true COA. While tuition and fees take center stage, the equally crucial COA, encompassing accommodation, meals, and supplies, often remains in the ambiguous. Notably, tuition and fees make up a mere 30% of the COA at public community colleges and 40% at public 4-year institutions [12]. Figure 7 - After Loans & Aid, Pell Students Face Gap of One-Third of Cost of Attendance shows how Pell Grant students have an average of 34% of their costs that “needed to be made up via work earnings, private or federal parent loans, savings, or other sources” [16].

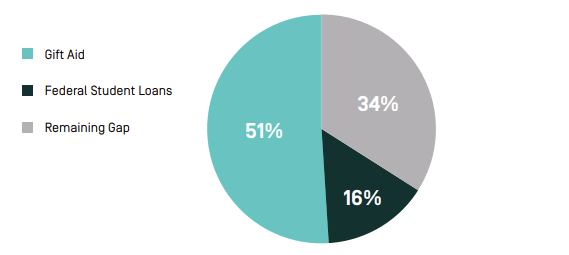


Figure - After Loans & Aid, Pell Students Face Gap of One-Third of Cost of Attendance

It is important to understand that both overestimating and underestimating the COA can negatively impact a prospective student. In the case of underestimation, the student could find themselves without sufficient funds to live while attending which may require working additional jobs, doing without necessities, or even leaving school. With overestimation, the student will have sufficient funds while in school, but will find themselves burdened with an oversized debt.

The issue with COA inconsistencies can even happen within the institution itself. An example is American University’s customer facing pricing calculator, which is based on the Federal EFC methodology; however, the school’s financial aid officer uses the Institutional EFC methodology for awards. Due to the distinct approaches these two methodologies take towards certain inputs, such as income and home equity, the estimated awards generated by the Net Price Calculator (NPC) may ultimately be considerably higher than the amounts presented in the official award letters received by families [17].

Institutions that inflate living-cost allowances inadvertently show favoritism towards more affluent student bodies, perpetuating an existing disparity. Conversely, underestimating these costs places an unjust burden on low-income students. Our initiative will provide a mechanism for policy makers to explore the interrelationship of COA with economic background and how financial biases may contribute to and perpetuate inequality within higher education. Armed with a solid data-backed foundation, our goal is to provide advocates with the tools required for essential policy reforms that can create a level playing field for all prospective students.

As legislation and rules are updated, the COA has fluctuated over the years. Our project will only focus on the latest year of data and will not explore or account for any past changes or trends.

## Research

This section is a review of the existing literature that provides the foundation for this study. The team performed extensive research into the problem by reviewing articles, discussing with educators, and reaching out to institutions geared to educate policymakers on the issues.

There are several sources that detail and summarize Department of Education rules and regulations concerning COA. The Department of Education Higher Education Opportunity Act of 2008 [18] required that by October 29, 2011, every postsecondary institution in the United States participating in Title IV student aid programs must display a net price calculator on its website. This calculator utilizes institutional data to offer estimated net price information to both current and prospective students and their families. A. Weisman [19] provides a summary of changes to requirements of the FAFSA Simplification Act for the year 2024-2025. These changes will modify the formulas and how the Pell Grants are calculated. The Department of Education Federal Student Aid Handbook [15] provides guidance to institutions on how to accumulate and calculate the costs associated with the COA.

The issues created by the net price calculators at the heart of COA estimations were explored by K. Degnan [17]. While NPCs are required by law, there is no enforceable guidance on their use other than that they exist on the institution’s web site. As a result, they are often poorly maintained or linked to out-of-date information which creates a false view of the costs involved.

The descrepancies in the NPCs provided by university websites were investigated by L. A. Davis, G. C. Wolniak, C. E. George and G. R. Nelson [20]. They examined the quality of information available on institutional websites regarding tuition and costs of attendance, with a focus on the potential impact on individuals seeking such information. The study identifies four key themes: significant variation in transparency and information scope across institutions, challenges in navigation and information continuity, discrepancies in page design and language use among departmental offices, and the potential for institutional websites to contribute to misconceptions about college costs. They ultimately recommend the redesign of institutional websites to better cater to diverse audiences, using clear language and consistent visual displays.

The variations in the cost-of-living allowances were examined by R. Kelchen, S. Goldrick-Rab and B. Hosch [13]. They identify systematic disparities across institutional sectors, particularly within the less-than-2-year for-profit sector, indicating potential influences like policy incentives and efforts to lower net prices. Additionally, the study advocates for clearer guidance from Federal Student Aid to ensure uniformity in cost of living calculations.

The difficulties in financing a college education were not entirely caused by the COA estimation as financial aid letters also create misinformation for students and parents. S. Burd, R. Fishman, L. Keane and J. Habbert [16] researched issues created by the lack of transparency and incomplete information within financial aid award letters. Inadequate information in these letters led to students and families making poorly informed decisions while financing the cost of attending college. Within their paper, they propose seven recommendations for policy makers and institutions of higher learning to create a more student-centric system.

With limited research available on student costs and expenses, it was necessary to explore living wages for working families. A.K. Glasmeier [14] created the Living Wage Calculator with the aim of assisting individuals, communities, employers, and various stakeholders in estimating the local wage necessary for a full-time worker to meet the basic needs of their family in their specific geographical location. This tool provides information on the living wage for 12 distinct family types, offering insights into the financial requirements in by county, metro area, or state. Another resource of living wages was the EPI Family Budget Calculator [21] which is very similar in function to the MIT Living Wage Calculator.

These issues contributed to the ballooning of student college debt which has become a national crisis and even resulted in a Presidential student loan forgiveness plan. B. Miller, C. Campbell and B. J. Cohen [22] explore the expanding student debt and how to address the issue. They review six options to tackle student debt with consideration towards equity of the student population.

The impact of student debt on Black individuals were discussed by A. M. Perry, M. Steinbaum and C. Romer [23] and challenges the notion that a college degree is a guaranteed path to economic prosperity. It highlights that despite the emphasis on higher education as a means to escape poverty, Black students often face financial challenges and accumulate debt, contributing to the fragility of the Black middle class. The article argues that the existing student debt system reinforces racial wealth gaps, as Black households carry more debt, impacting creditworthiness and homeownership rates. The author's conclusion was to advocate for student debt cancellation to achieve economic equity.

To understand the important shifts in college tuition through the years. J. Ma [24] investigated the trends and broke down tuition costs by sector (public, private, two-year, four-year) and by state. The paper further details enrollment patterns by sector and race/ethnicity. Finally, there is an extensive study into the various financial aid instruments available by sector and state.

The method by which colleges and universities receive their funds was the subject of a paper by A. Stauffer, et al, [25]. They provide detailed data on state and federal funding to colleges and universities through 2017 and show a shift over the years from primarily state funding to increasing federal funding. The state money largely funds the operation of the college while the federal money will mainly go directly to the student as aid or used for research project.

## Solution Space

### Solution Approach

Our end goal is to develop an interactive dashboard that will display cost estimates from schools against cost estimates generated from an independent source. This will validate the accuracy of school-generated estimates and provide an independent and accurate COA metric. Our approach will utilize the College Scorecard and IPEDS data to establish the cost estimates that higher education institutions have developed themselves. The EPI Family Budget Calculator data will form the basis of our independent cost estimate. Both sets of data will need to be cleaned, transformed, and organized to facilitate a comparative analysis. A series of acceptability thresholds will then be implemented to provide generalized insights into the veracity of the cost estimates from the schools. These thresholds will be informed by the Sponsors and their criteria for appropriate levels of accuracy.

### Solution Space

Our system delivers value by shedding light on discrepancies in college living-cost allowances, thereby fostering accountability, equity, and informed decision-making within higher education finance. Users, including prospective students and their families, benefit from accessing transparent and understandable comparisons of living expenses across institutions, helping them navigate financial aid considerations and budgetary planning effectively. Moreover, by highlighting institutions with accurate living-cost allowances and advocating for policy changes to address disparities, our initiative aims to mitigate educational inequality and promote fairness in access to higher education. Through its dual role as an advocacy tool and a resource for stakeholders, our solution contributes to a more transparent and equitable landscape in college financing, ultimately enhancing the educational journey for all involved parties.

## Project Objectives

1. The team will learn the features that are used to create COA estimates. As a part of that, we will explore the most impactful elements and how those costs can be adjusted for a more accurate result. We will also learn about the policies that govern financial aid for higher education.
2. The team will develop a web-accessible dashboard to provide meaningful and accurate COA to students, parents, and policy makers.
3. The team will gain an understanding of the complexities and nuances involved in estimating both costs of higher education and costs of living.
4. This dashboard will allow for easy, fast, and meaningful comparisons between institution-developed and independently developed cost of attendance estimates. Armed with this information, students, policy makers, and other interested parties can make more information decisions.

## Primary User Stories

A user story is a natural language explanation of a software feature written from the perspective of the end user that explains how value will be provided to that user. The users identified were a student or parent, policy maker, and university financial administrator. By addressing the perspectives of each user group, we will create a tool that informs various individuals involved in the higher education ecosystem about the COA of colleges and universities nationwide.

The following were the primary user stories created and reviewed for development of the interactive dashboard.

1. As a student, I want to compare estimated living expenses with the allowances set by different colleges, so that I can make informed decisions about my educational expenses.
2. As a parent, I want to have an easy-to-understand website that details the expenses in my child's education so that I can budget for my child's education without being surprised by hidden costs.
3. As a policy maker, I want to see a filtered list of educational institutes with the best and worst estimations within the past year, so that I know which institutions to evaluate further and how to approach policy advocation.
4. As a university financial administrator, I want to have an easy-to-understand breakdown of my institution's COA, so I can work with the school administration to improve accuracy.

## Product Vision

### Scenario 1

For the policy maker who needs to understand which institutions are delivering acceptable levels of estimates and which are out of bounds. Our product will quickly deliver such information to identify schools with poor estimates so that remediation may take place.

### Scenario 2

For the student who is considering enrolling in college. The dashboard will help assess the true cost of a school, allowing the student to better plan and prepare for the level of financial support needed to attend. Our product will allow students to compare true costs easily and quickly between schools to make a more informed decision for their academic and personal future.

### Scenario 3

For the Financial Aid administrator who wants to understand how the college compares to the living wage to better advise students and the administration on financial aid. Our product will identify discrepancies in the school’s COA calculation and give the administration a view of how they compare to their peers and help them improve their NPC.

# Datasets

## Overview

Three datasets were planned for this project: MIT Living Wage Calculator [14], Integrated Postsecondary Education Data System (IPEDS) [26], and College Scorecard Cost of Attendance [27]. The MIT Living Wage Calculator was unavailable and replaced by the EPI Family Budget Calculator [21]. These datasets fall into two general categories: datasets that capture COA estimates from colleges and universities, and cost of living datasets that provide an estimate of minimum realistic costs for different geographic locations throughout the United States.

IPEDS and College Scorecard contain COA estimates that schools participating in Title IV federal aid programs are mandated to submit to the Department of Education. These datasets provide a wealth of important information, namely socioeconomic, demographic, and financial data, related to the students that attend the respective schools and the related educational costs.

The cost-of-living dataset comes from the Economic Policy Institute, or EPI. The EPI data aims to investigate and measure the costs an individual would likely face when living in a specified geographic area. Their estimates are broken into several categories such as housing, food, transportation, childcare, and healthcare. The estimates reflect various life and living scenarios, ranging from single individuals with no children to two-parent families with multiple children.

## Field Descriptions

**Dataset 1 – EPI Family Budget Calculator**

1. Case\_ID (type: Integer): An identifier determining different possibilities for family sizes and states.

2. State\_ABV (type: String): Abbreviations for all 50 states in the United States, aiding in data organization and state identification.

3. County\_Fips (type: Integer): Unique identifiers for U.S. counties according to Federal Information Processing Standards, essential for geographic referencing and data analysis.

4. County (type: String): County names categorized by states, providing regional context for expenditure data.

5. Family (type: String): Information about family sizes and corresponding expense variations, facilitating analysis based on household composition.

6. Housing\_Monthly (type: Integer): Monthly housing expenses for each family size and state, providing insights into housing affordability.

7. Food\_Monthly (type: Integer): Monthly food expenses for each family size and state, offering insights into dietary spending patterns.

8. Transportation\_Monthly (type: Integer): Monthly transportation expenses for each family size and state, aiding in understanding commuting costs.

9. Healthcare\_Monthly (type: Integer): Monthly healthcare expenses for each family size and state, providing insights into healthcare affordability.

10. Other Necessities\_Monthly (type: Integer): Monthly miscellaneous expenses for each family size and state, offering insights into general living expenses.

11. Childcare\_Monthly (type: Integer): Monthly expenses for children in families with children across states, aiding in understanding childcare costs.

12. Taxes\_Monthly (type: Integer): Monthly tax expenses for each family size and state, providing insights into tax burdens.

13. Total\_Monthly (type: Integer): Total monthly expenses including housing, food, transportation, healthcare, necessities, childcare, and taxes, offering a comprehensive view of family expenditure patterns.

14. Housing\_Annual (type: Integer): Annual housing expenses for each family size and state, providing insights into long-term housing affordability.

15. Food\_Annual (type: Integer): Annual food expenses for each family size and state, offering insights into annual dietary spending patterns.

16. Transportation\_Annual (type: Integer): Annual transportation expenses for each family size and state, aiding in understanding annual commuting costs.

17. Healthcare\_Annual (type: Integer): Annual healthcare expenses for each family size and state, offering insights into annual healthcare affordability.

18. Other Necessities\_Annual (type: Integer): Annual miscellaneous expenses for each family size and state, providing insights into annual general living expenses.

19. Childcare\_Annual (type: Integer): Annual expenses for children in families with children across states, aiding in understanding annual childcare costs.

20. Taxes\_Annual (type: Integer): Annual tax expenses for each family size and state, providing insights into annual tax burdens.

21. Total\_Annual (type: Integer): Total annual expenses including housing, food, transportation, healthcare, necessities, childcare, and taxes, offering a comprehensive view of annual family expenditure patterns.

22. Median\_family\_income\_Rankings (type: Integer): Rankings among all family sizes across different counties and states, providing insights into income distribution.

23. Num\_counties\_in\_st\_Rankings (type: Integer): Total number of counties in each state, facilitating state-level comparative analysis.

24. St\_cost\_rank\_Rankings (type: Integer): Rankings of counties by total budget cost, offering insights into expenditure disparities.

25. St\_med\_aff\_rank\_Rankings (type: Integer): Affordability rank based on median family income as a share of total annual cost, aiding in understanding affordability disparities.

26. St\_income\_rank\_Rankings (type: Integer): Rankings of counties by median family income, providing insights into income distribution across counties.

**Dataset 2 –IPEDS**

1. UNITID (type: Integer): A unique identification number assigned to each educational institution, facilitating data linkage across datasets, and enabling institutional identification.

2. XCHG1AY3 (type: String): Indicates attempted changes or modifications for in-district published tuition and required fees for the academic year 2022-23, offering insights into pricing adjustments.

3. CHG1AY3 (type: String): Represents in-district published tuition and required fees for the academic year 2022-23, providing data on educational expenses for students residing within the institution's district.

4. XCHG2AY3 (type: String): Reflects attempted changes or modifications for in-state published tuition and required fees for the academic year 2022-23, offering insights into pricing adjustments.

5. CHG2AY3 (type: String): Displays in-state published tuition and required fees for the academic year 2022-23, offering data on educational expenses for in-state students.

6. XCHG3AY3 (type: String): Represents attempted changes or modifications for out-of-state published tuition and required fees for the academic year 2022-23, offering insights into pricing adjustments.

7. CHG3AY3 (type: String): Indicates out-of-state published tuition and required fees for the academic year 2022-23, providing data on educational expenses for out-of-state students.

8. XCHG4AY3 (type: String): Indicates attempted changes or modifications for books and supplies expenses for the academic year 2021-22, offering insights into pricing adjustments.

9. CHG4AY3 (type: String): Represents books and supplies expenses for the academic year 2021-22, providing data on educational materials costs.

10. XCHG5AY3 (type: String): Reflects attempted changes or modifications for on-campus room and board charges for the academic year 2022-23, offering insights into pricing adjustments.

11. CHG5AY3 (type: String): Represents on-campus room and board charges for the academic year 2022-23, providing data on housing and dining expenses for on-campus students.

12. XCHG6AY3 (type: String): Indicates attempted changes or modifications for on-campus other expenses for the academic year 2022-23, offering insights into pricing adjustments.

13. CHG6AY3 (type: String): Represents on-campus other expenses for the academic year 2022-23, providing data on additional expenses incurred by on-campus students.

14. XCHG7AY3 (type: String): Reflects attempted changes or modifications for off-campus (not with family) room and board charges for the academic year 2022-23, offering insights into pricing adjustments.

15. CHG7AY3 (type: String): Represents off-campus (not with family) room and board charges for the academic year 2022-23, providing data on housing and dining expenses for off-campus students.

16. XCHG8AY3 (type: String): Indicates attempted changes or modifications for off-campus (not with family) other expenses for the academic year 2022-23, offering insights into pricing adjustments.

17. CHG8AY3 (type: String): Represents off-campus (not with family) other expenses for the academic year 2022-23, providing data on additional expenses incurred by off-campus students.

18. XCHG9AY3 (type: String): Reflects attempted changes or modifications for off-campus (with family) other expenses for the academic year 2022-23, offering insights into pricing adjustments.

19. CHG9AY3 (type: String): Indicates off-campus (with family) other expenses for the academic year 2022-23, providing data on additional expenses incurred by students living off-campus with their families.

**Dataset 3 – College Scorecard Cost of Attendance**

1. UNIT\_ID (type: integer): Numerical identifier uniquely assigned to each educational institution for precise differentiation and organization within datasets.

2. INSTNM (type: object): Official name of an educational institution.

3. CITY (type: object): Name of the city where an educational institution is located.

4. STABBR (type: object): Postal code associated with the state where an educational institution is located.

5. ZIP (type: object): Postal code associated with the location of an educational institution.

6. INSTURL (type: object): Hyperlink address directing to the official website of an educational institution.

7. NPCURL (type: object): Web address leading to the net price calculator tool provided by an educational institution.

8. LATITUDE (type: float): Geographical latitude coordinate of an educational institution's location.

9. LONGITUDE (type: float): Geographical longitude coordinate of an educational institution's location.

10. MENONLY (type: integer): Flag indicating if the institution is men-only.

11. WOMENONLY (type: integer): Flag indicating if the institution is women-only.

12. COSTT4\_A (type: Integer): Average cost of attendance for academic year institutions.

13. COSTT4\_P (type: Integer): Average cost of attendance for program-year institutions.

14. TUITIONFEE\_IN (type: Integer): In-state tuition and fees.

15. TUITIONFEE\_OUT (type: Integer): Out-of-state tuition and fees.

16. TUITIONFEE\_PROG (type: Integer): Tuition and fees for program-year institutions.

17. TUITFTE (type: Integer): Net tuition revenue per full-time equivalent student.

18. BOOKSUPPLY (type: Integer): Cost of attendance for estimated books and supplies.

19. ROOMBOARD\_ON (type: Integer): Cost of attendance for on-campus room and board.

20. OTHEREXPENSE\_ON (type: Integer): Cost of attendance for on-campus other expenses.

21. ROOMBOARD\_OFF (type: Integer): Cost of attendance for off-campus room and board.

22. OTHEREXPENSE\_OFF (type: Integer): Cost of attendance for off-campus other expenses.

23. OTHEREXPENSE\_FAM (type: Integer): Cost of attendance for residing with-family other expenses.

24. ADDR (type: String): Address of the institution.

## Data Context

**1. EPI Family Budget Calculator**

Origin: The data originates from the Economic Policy Institute's Family Budget Calculator; a comprehensive tool designed to analyze and understand the financial needs of families across the United States.

Purpose: The primary purpose of this dataset is to provide detailed information on family budget estimates across different states and family sizes in the United States. It encompasses various expenses such as housing, food, transportation, healthcare, childcare, and taxes, offering insights into expenditure patterns and affordability.

Collection Method: The data collection method involves researching multiple publicly accessible government datasets such as the Bureau of Labor Statistics or Social Security Administration statistics as an example.

Temporal Context: While past years are available, the dataset used for this report was for the year 2024. The data for this dataset is in 2023 dollars.

Geographical Context: The data covers family budgets across all 3,143 U.S. counties and county equivalents, providing insights into regional variations in expenditure patterns and affordability.

Institutional Context: Each record in the dataset includes identifiers for different family sizes and states, facilitating data analysis and comparison.

Quality and Reliability: The quality and reliability of the data are influenced by factors such as the accuracy of reporting by participants, data collection methodologies employed by the Economic Policy Institute, and any validation processes implemented to ensure data accuracy.

**2. IPEDS Cost Dataset 2022**

Origin: The data originates from the Integrated Postsecondary Education Data System (IPEDS), managed by the National Center for Education Statistics (NCES) within the U.S. Department of Education.

Purpose: The primary purpose of this dataset is to offer detailed information on the cost of attendance (COA) estimates for various educational institutions throughout the United States. These estimates encompass tuition, fees, and other related expenses for different types of students (in-district, in-state, out-of-state).

Collection Method: Data is collected through surveys and reporting mechanisms mandated by federal regulations. Educational institutions are required to report this information to IPEDS, ensuring consistency and standardization in data collection.

Temporal Context: The dataset used was for 2022 which covers the 2022-2023 academic year, reflecting the cost estimates for that period.

Geographical Context: The data covers educational institutions across the United States and territories, providing insights into regional variations in educational expenses.

Institutional Context: Each record in the dataset is associated with a unique identification number (UNITID) assigned to each educational institution, enabling data linkage and institutional identification.

Quality and Reliability: The quality and reliability of the data are influenced by factors such as the accuracy of reporting by educational institutions, data collection methodologies, and data validation processes implemented by IPEDS.

**3. College Scorecard**

Origin: The dataset originates from surveys and data collection efforts undertaken by the Department of Education and academic institutions.

Purpose: The purpose of collecting this dataset is to provide comprehensive information on the cost of attendance at various educational institutions, aiding students, parents, and policymakers in making informed decisions.

Collection Method: Data is collected through surveys, academic libraries information centers, and the calculation of average institutional education net price.

Temporal Context: The dataset dated October 2023 was used and reflects the cost of attendance data for the academic year 2021-22.

Geographical Context: The dataset covers educational institutions across the United States and territories, providing information on their costs of attendance.

Institutional Context: The dataset includes information on various aspects of educational institutions, such as their names, locations, tuition fees, room and board expenses, and other related costs.

Quality and Reliability: The dataset is sourced from the Department of Education and is subjected to quality checks and validation processes to ensure accuracy and reliability.

## Data Conditioning

**Dataset 1:** **EPI Family Budget Calculator:**

FIPS code data was added to enable geographical analysis and link to the IPEDS dataset. Rows for family scenarios other than 1 adult with 0 children and 1 adult with 1 child (specifically for calculating shared housing costs) were removed. We validate the accuracy of cost-related columns, ensuring the absence of negative values in both monthly and annual costs.

**Dataset 2:** **IPEDS Cost Dataset 2022:**

The column headers were changed into a more understandable format.

**Dataset 3: College Scorecard** **Cost of Attendance**

This dataset has over 3000 columns, most of which are not required for this study, so the first step was to remove all but 53 columns. Next, only schools that grant a two-year degree or higher were kept, which removed approximately 40% of the 6543 institutions listed. The remaining cleaning resulted in finer adjustments, such as removal of schools in US territories, currently shut down, or institutions missing values in critical columns like 'INSTURL,' 'NPCURL,' 'LATITUDE,' or 'LONGITUDE.'

The datasets were then linked into a single spreadsheet using a numeric school ID shared between IPEDS and College Scorecard and then a County FIPS code between EPI and College Scorecard.

## Data Quality Assessment

The data quality for each dataset was assessed based on the following criteria [28]:

* Completeness – tests that a particular column, element, or class of data is populated and does not feature null values or values in place of nulls (e.g., N/A).
* Uniqueness – tests whether all the entities or attributes within a dataset are unique.
* Accuracy – tests whether the data values stored for an object are the correct values.
* Integrity – tests if all the relationships are populated for a particular entity. [e.g., the inability to link related records together may introduce duplication across systems]
* Conformity – tests if the data conforms to right conventions and standards.

**Dataset 1:** **EPI Family Budget Calculator**

1. **Completeness:**

* The dataset exhibits a high level of completeness, with no missing values in most columns.
* Only the 'median\_family\_income' column has 10 missing values, constituting a negligible percentage of missing data (0.03%).
* The completeness percentage for each column ranges from 0% to 0.03%, indicating a comprehensive dataset.

1. **Uniqueness:**

* All rows in the dataset are unique, as evidenced by the absence of duplicate entries.
* The uniqueness check resulted in 31,430 'False' values, indicating no duplicate records.

1. **Accuracy:**

* The accuracy check confirms the absence of negative values in both monthly and annual cost columns.
* All columns related to costs, both monthly and annual, show no instances of negative values.

1. **Integrity:**

* The integrity checks for 'Total\_Monthly' reveal discrepancies, where transactions may not be treated as indivisible units. This highlights the need for thorough investigation and corrective actions to uphold data integrity standards.
* Further investigation is recommended to identify and rectify the discrepancies in the 'Total\_Monthly' column.

1. **Conformity:**

* Non-numeric values have been checked for in numerical columns ('Housing\_Monthly' to 'Total\_Annual').
* All columns have consistent numeric values, ensuring conformity for numerical analyses.

1. **Overall Quality:**

* The dataset demonstrates high overall quality, with completeness, uniqueness, and accuracy being notable strengths.
* The atomicity check reveals a specific area for improvement in the 'Total\_Monthly' column, warranting further attention.
* Conformity checks indicate consistent numeric values, enhancing the reliability of the dataset for quantitative analyses.

**Dataset 2:** **IPEDS Cost Dataset 2022**

1. **Completeness:**

* The completeness check for the "IPEDS Cost Dataset 2022" reveals missing values in various columns.
* Specifically, indicator columns such as 'Published in-district tuition and fees 2022-23,' 'Published in-state tuition and fees 2022-23,' 'Published out-of-state tuition and fees 2022-23,' 'Books and supplies 2022-23,' 'On campus, room and board 2022-23,' 'On campus, other expenses 2022-23,' 'Off campus (not with family), room and board 2022-23,' 'Off campus (not with family), other expenses 2022-23,' and 'Off campus (with family), other expenses 2022-23' have some missing values. The total number of missing values varies for each column.

1. **Uniqueness:**

* The uniqueness check indicates that there are no duplicate rows in the "IPEDS Cost Dataset 2022," as all values are unique across the dataset.

1. **Accuracy:**

* The accuracy checks for numeric values, represented by columns 'Published in-district tuition and fees 2022-23' to 'Off campus (with family), other expenses 2022-23,' passes successfully.
* There are no negative values present in these columns, ensuring the accuracy of the dataset.

1. **Integrity:**

* The integrity check confirms that the dataset adheres to predefined conditions. Specifically, it validates that when the indicator columns contain 'A,' the corresponding 'CHG' columns have null values.
* Additionally, when indicator columns contain 'R' or 'C,' the corresponding 'CHG' columns have non-null values.

1. **Conformity:**

* The conformity check assesses the presence of non-numeric values in columns that should contain numeric data. In this dataset, columns 'Published in-district tuition and fees 2022-23' to 'Off campus (with family), other expenses 2022-23' have only numeric values, meeting the conformity criteria.

1. **Overall Quality:**

* The "IPEDS Cost Dataset 2022" demonstrates good overall data quality. It has limited missing values, unique rows, accurate numeric values, and maintains integrity and conformity according to predefined conditions.
* However, addressing missing values in certain indicator columns may further enhance the dataset's completeness.

**Dataset 3:** **College Scorecard Cost of Attendance**

1. **Completeness:**

* The dataset exhibits varying levels of completeness across different columns. Notable observations include:
* 'INSTURL' has 17 missing values.
* 'NPCURL' shows 540 missing values.
* Columns like 'LATITUDE,' 'LONGITUDE,' 'MENONLY,' 'WOMENONLY' have 497 missing values.
* Multiple other columns throughout the dataset also contain missing values.

1. **Uniqueness:**

* The dataset contains 6543 unique rows, indicating the absence of duplicate entries.

1. **Accuracy:**

* No negative values have been identified in any numerical columns, ensuring the accuracy of cost-related data. The overall accuracy check has passed successfully.

1. **Integrity:**

* The dataset undergoes thorough checks to ensure all relationships are properly populated for each entity. Failure to link related records may lead to duplication across systems, potentially impacting data consistency and reliability.

1. **Conformity:**

* The dataset adheres well to the expected data types, with successful conversion of 'LATITUDE' and 'LONGITUDE' columns to the object data type. Additionally, numerical columns such as 'COSTT4\_A,' 'COSTT4\_P,' 'TUITIONFEE\_IN,' and 'TUITIONFEE\_OUT' show conformity, with no non-numeric values identified.

1. **Overall Quality:**

* The overall quality of the dataset is satisfactory, considering the completeness, uniqueness, accuracy, and conformity checks. However, there are missing values that may need attention, and the dataset's quality could be improved through further cleaning and validation processes.

The overall data quality for all three datasets was good with the results of the quality assessments summarized in Table 1.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **EPI** | **IPEDS** | **College Scorecard** |
| **Completeness** | Good | Good | Good |
| **Uniqueness** | Good | Good | Good |
| **Accuracy** | Good | Good | Good |
| **Integrity** | Good | Good | Good |
| **Conformity** | Good | Good | Good |

Table : Data Quality Assessment

## Other Data Sources

The project team planned to use the MIT Living Wage dataset; however, The Living Wage Institute rejected our request for use of the data. The team switched to the EPI Family Budget Calculator as a replacement to gather a living wage for localities across the United States. This dataset provides the same information required by our project and was freely available.

## Storage Medium

Personal computers were used as the initial storage medium during data collection, exploration, cleaning, preprocessing, and dashboard development. This enabled a flexible and inexpensive solution for the project team to access and manipulate the data quickly.

## Storage Security

The data used in this project and any findings are all publicly available. With that in mind, the focus on storage security would involve data integrity and availability. To protect against unauthorized changes or data corruption, the students should implement basic security practices such as keeping their operating system updated and employing antivirus software. Additionally, the students should employee some form of backup solution such as cloud storage or an external hard drive to safeguard against data loss.

## Storage Costs

The first phase of the project utilized a sponsor provided Google drive and student-owned personal computers for data storage. While there are costs associated with both, for the purposes of this project, those costs are negligible as the sponsors would pay any costs associated with the Google drive and the data set sizes were small enough to not require any enhancement (e.g. external hard drive) to the student-owned computers.

# Algorithms & Analysis

## Solution Approach

The approach for this project can be broken into three distinct phases, with each phase reliant upon the previous, and all are critical to project success.

The initial phase is data collection, cleaning, and curation. Various data sets were collected that report the cost estimates from colleges and universities. Data was also collected from EPI to form the basis of the independent estimate. Both data sources collect various, and often disparate, information. Cleaning and organizing is necessary to remove unneeded data. Examples of this include stripping away data not relevant to the research, removing features where ‘null’ made up a significant portion of the records, and removing institutions that are listed as inactive. Additionally, external geolocation data was utilized to join the multiple datasets that otherwise did not have a common identifier.

The next phase entails fine tuning of the independent cost estimates. This requires understanding the different elements of the cost estimates and developing appropriate approaches to model costs that are reasonable and realistic for college students. This process can broadly be separated into two halves: one for deeply understanding the methods, approaches, and assumptions made that form the basis of the cost estimate from EPI and the other for adjustment factors and assumptions to refine the EPI data to reflect college related costs. This second half of the phase requires seeking out research and other forms of data to substantiate the rationale for making given adjustments.

Visualizing the data and developing an intuitive, user-friendly dashboard rounds out the final phase. This is the most visible and forward-facing part of the solution as it is what the end-user will see and interact with. The dashboard takes the data from the previous phases and presents easily understandable and digestible visualizations, along with numeric and textual data to address the questions about college cost of attendance estimates.

### Systems Architecture

The system architecture in Figure 8 leans heavily on the established Tableau architecture with the primary source of the data coming from downloaded Excel spreadsheets.



Figure - System Architecture

The system has the following components:

* Data Source: The data source will be composed of one or more excel worksheets that will be linked together within Tableau Desktop.
* Tableau Desktop: Tableau Desktop will connect to the data sources and create the worksheets and dashboards that will make up the visualizations of the data.
* Visualizations: These will be charts, maps, and graphs to visually represent the data in an easily accessible manner for the user. A subset of the visualizations will be user interactive.
* Tableau Server: Cloud server hosted by Tableau Public where the data source and visualizations are published.
* User System: Users will connect via the web to the Tableau Public server to consume the dashboards.

### Systems Security

The security considerations for developing the interactive dashboard on Tableau Public involve a combination of measures to ensure data integrity, confidentiality, and compliance with Tableau Public's platform security features.

* Data Encryption:
  + Tableau Public employs secure connections through HTTPS, ensuring encrypted data transmission between the users' browsers and the Tableau Public servers.
  + This fundamental encryption layer provides confidentiality and prevents unauthorized access during data transit.
* User Authentication:
  + Tableau Public utilizes Tableau Server authentication mechanisms.
  + While Tableau Public itself does not support user authentication, the public dashboards can be embedded into secure environments or accessed through secure links, allowing users to authenticate through those channels if needed.
* Data Sensitivity Considerations:
  + Given that Tableau Public is a public platform, it's essential to be cautious about the sensitivity of the data being visualized.
  + Avoid displaying personally identifiable information (PII) or any confidential data directly on the public dashboards.
  + If sensitive information is involved, consider using anonymized or aggregated data for public-facing visualizations.
* Regular Monitoring and Updates:
  + Tableau Public benefits from the security measures implemented by Tableau as a service provider.
  + Regular updates and patches applied by Tableau help address any security vulnerabilities that may arise.
  + It is crucial to stay informed about Tableau's updates and promptly implement them in the interactive dashboard.
* Privacy Compliance:
  + Ensure compliance with Tableau Public's terms of service and any relevant data protection regulations.
  + Avoid including data that violates privacy standards and be transparent about the nature of the data being presented.
  + If necessary, consider creating multiple versions of the dashboard, one for public consumption and another for more sensitive data accessible through secure channels.
* Access Controls:
  + While Tableau Public dashboards are public facing, ensure that the underlying data is appropriately curated and shared. Any data connected to Tableau should be reviewed for public consumption.
  + If sensitive information is involved, explore ways to implement additional access controls on the hosting platform where the dashboard is embedded.
* Incident Response Planning:
  + Have a well-defined incident response plan in place to address any unforeseen security incidents.
  + This includes protocols for identifying and responding to security breaches and communication plans to inform stakeholders about incident resolution.

Developing the interactive dashboard on Tableau Public involves leveraging the security features inherent in the platform. By adhering to best practices and being mindful of data sensitivity, the initiative aims to provide a secure and reliable tool for enhancing transparency in college living costs while maintaining user trust on the Tableau Public platform.

### Systems Data Flows

The data is collected from three datasets, the first one is taken from the Department of Education, the second from the National Institute for Education Statistics, and third from the Economic Policy Institute. These are downloaded as CSV files for cleaning and pre-processing using Jupyter Notebook, Python, and Excel. All three datasets are merged into a single dataset using the common attributes from each dataset. Once complete, the data is adjusted Excel using algorithms developed in 3.1.4. and uploaded to Tableau Public. Using both the web interface and the desktop application of Tableau, the data is manipulated to produce visualizations for analysis by users.

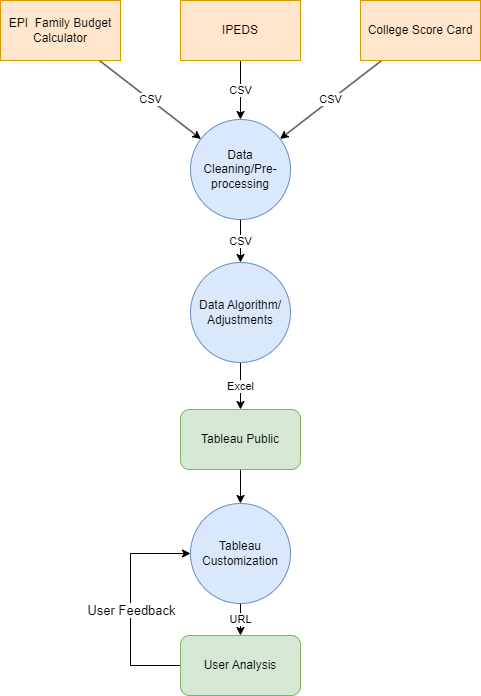


Figure - Data Flow Diagram

It is expected that the user analysis could generate additional requirements for new visualizations. There are areas of interest outlined in Future Work that highlight potential new customizations.

### Algorithms & Analysis

One of the primary goals of this project was to compare Institutional COA with an independent cost estimate. Due to a lack of data and related research that specifically addresses student costs, it was necessary to create new estimates based on available living wage calculators such as MIT’s Living Wage Calculator or EPI’s Family Budget Calculator. As the MIT data was unavailable, the EPI data was selected as the starting baseline for our independent estimate.

The EPI Family Budget Calculator estimates budgets for 10 different family types. “These families include a single person with no children; a married couple with no children; single-parent families with one, two, three, or four children; and a married couple with one, two, three, or four children.” [21] Each family is assumed to have a working adult who is considered the head of household for tax purposes. This report will assume a single adult with no children with an average age of 25 years old who is a full-time student. When calculating housing, EPI data for 1 adult and 1 child was used as these generated costs for a 2-bedroom household with the minimum number of individuals. EPI assumes 2 adults would be married and share a room. Two bedrooms are required for our analysis as all housing costs assume a roommate.

Five of the seven costs captured in the EPI Family Budget Calculator were used in this report: Housing, Food, Transportation, Healthcare, and Other. Childcare was excluded as this study assumes a single adult with no children and Taxes were excluded because it was assumed the individual would be below a taxable income level. Figure 10 is a summary of the baseline assumptions and methodologies used for EPI’s cost estimates. This captures how the Economic Policy Institute developed their estimates and what data they used to support such estimates. This is important for our work as the EPI estimates, with adjustments, will form the basis of our independent cost estimate.

|  |  |
| --- | --- |
| EPI Estimates | |
| **Cost Element** | **EPI Methodology/Assumption** |
| Housing | Utilized a 2023 U.S. Department of Housing and Urban Development report on fair market rents. 40th percentile is used. |
| Food | Developed per 2023 report - Official USDA Food Plans: Cost of Food at Home at Four Levels. Estimates use the Low-Cost food plan |
| Transportation | Developed using data from Center for Neighborhood Technology (CNT), which in turn used data from multiple Government sources |
| Healthcare | Premiums: 40yo non-smoker, bronze plans per healthcare.gov Out of Pocket Costs: HHS Medical Expenditure Panel Survey (MEPS) |
| Other | A function of Housing and Food costs. EPI makes use of Bureau of Labor Statistics (BLS) Consumer Expenditure Survey (CEX) data and uses the second fifth (20th - 40th) percentile. |

Figure - EPI Estimate Methodology

Aggregating all the cost estimate data for every county as seen Figure 11 reveals the median, or 50th percentile.

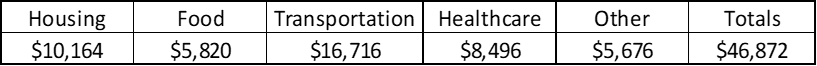


Figure - EPI Median Costs

This shows that Transportation is the largest of any of the cost elements considered in this analysis when making no adjustments to the data. Housing accounts for the next largest share of estimated living costs. When combined, these two categories make up 58% of total costs. The remaining cost elements of Healthcare, Food, and Other round out the remaining share of costs.

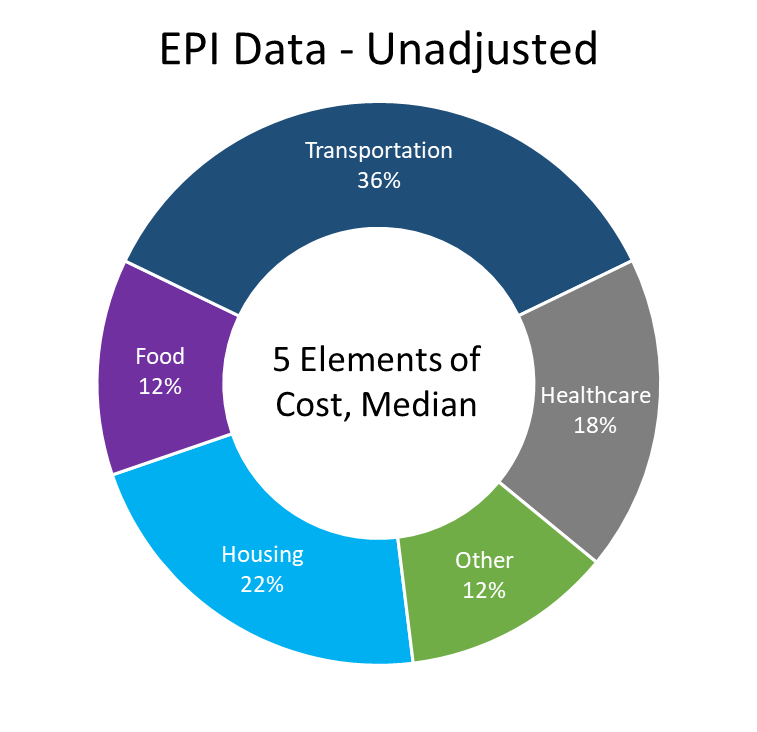


Figure - EPI Data - Unadjusted

The total costs and allocations as seen in Figure 12 were not reasonable as compared to an average college student. It was necessary to develop an adjustment methodology to better align the costs to what an average college student would be expected to experience. In Table 2 – Adjustments to EPI Costs, adjustment factors for each of the cost elements are shown.

**Housing:**

Housing costs assumes a 2-bedroom apartment shared by 2 people, reducing the costs by 50%. This follows a similar assumption made within the IPED data where students are assumed to be living with another student, sharing common living spaces while having their own bedroom. The costs were further adjusted to account for a 9-month academic year. The project team assumed that living with relatives provided an additional 85% reduction in costs.

**Food:**

Food costs were adjusted based on a 9-month academic year. The project team assumed living with relatives provided an additional 20% reduction in costs.

**Transportation:**

Transportation costs were based on the 2022 Department of Energy study [29] [30] which found transportation was 16% of an average household’s total yearly expenditures. Since this was a fixed percentage, no further adjustments were required. Food, Housing, Healthcare, and Other are totaled and multiplied by 0.190476 to achieve 16% of the total household expenditures.

**Healthcare:**

The EPI healthcare assumptions and methodology are based on premiums for 40-year-old non-smokers sourced from healthcare.gov, along with out-of-pocket costs obtained from the HHS Medical Expenditure Panel Survey (MEPS). For our analysis, we gathered ACA insurance quotes for a 25-year-old and a 40-year-old and left all other EPI assumptions unchanged. The two sets of data samples were compared to understand how quotes varied with age. This was calculated by dividing the average aggregated insurance totals of the 25-year-old scenario by the 40-year-old scenario. The resulting ratio produced our adjustment factor. Please note that due to the MEPS data being restricted and inaccessible, we also assume that the out-of-pocket costs proportionally scale with insurance premiums.

Through further discussions and subsequent research performed by the Sponsors at Hildreth Institute to test these assumptions, the decision was made to remove Healthcare from the cost estimations based on feedback from various industry partners. The rationale for this decision rests with the expectation that most colleges and universities include healthcare costs within their tuition and fee schedules.

**Other:**

This cost element captures other necessities as stated in the Bureau of Labor Statistics (BLS) Consumer Expenditure Survey (CEX) that the EPI data is built on. Costs for things such as apparel, personal care, household supplies, and telephone services fall into this category. This factor follows the same method that EPI uses, which determined this cost as a static ratio of housing and food expenditures, equal to 35.4% of those cost elements. No adjustments were required as housing and food costs were already adjusted.



Table : Adjustments to EPI Costs

# Visualizations

## Overview

One of the key goals of this project was to visualize the data into an easy-to-understand dashboard. Tableau Public was selected as the dashboarding tool due to its versatility and being free to use. Given all the data collected started as CSV files, the project team performed most of the preprocessing and calculations in Excel prior to uploading to the Tableau Public servers. We wanted to highlight the Institutional derived cost estimates and compare them to the independently developed cost estimates created by our project team. To this end, the dashboard needed to show the breakdown of how a given school’s COA was calculated and provide comparisons to schools across a state or the entire nation. This would provide the data at a macro level and enable a decision maker to drill down as desired.

## Visualizations

The map in Figure 13 examines the distribution of colleges and universities within the U.S. as reported in the institutional data from IPEDS. More than 3,000 higher education institutions are captured and represented here. Those states with a larger share of schools are shaded in the darker hues, while the lighter colors reflect fewer institutions in those respective states. While not necessarily in the same order, it should be no surprise that the states with the most institutions are also the states with higher populations levels. California, New York, and Texas make up the top 3 states with respect to number of schools. Note that due to small state size, Delaware and Rhode Island do not show values on the map but have 7 and 13 schools each, respectively.

A map of the united states

Description automatically generated

Figure - Distribution of Institutions by State

The Hildreth Dashboard offers an interactive visualization of the adjusted data to enable informed decision-making and analysis. The dashboard, built using Tableau Public, consists of two primary worksheets: one for out-of-state data and one for in-state data. Figure 14 illustrates an unfiltered view of the out-of-state worksheet. Several parameters are available across the top that allow the user to drill into more detailed scenarios, specifically the Living Scenario, State, Institution, Institution Type and Degree Type.

Each scenario directly compares IPEDS living expense estimates to the Hildreth estimate and provides a cost breakdown by category. By leveraging data from both the EPI family budget calculator and IPEDS and adjusting the parameters as required, users can analyze variations within the cost estimation methodologies employed by different institutions.

Graphs across the bottom highlight institutions with the highest overestimations and underestimations as well as the schools that were closest to the Hildreth estimate. These visualizations help policy makers focus analysis on the most problematic estimates as well as the best.

A screenshot of a computer

Description automatically generated

Figure - Tableau Dashboard

# Findings

The goal of this research is to examine institution developed cost of living estimates against independently developed values, spurred by concern that colleges and universities may not produce realistic, and thus meaningful, living expense estimates. The analysis offers additional insight that builds on prior research, which found notable percentage differences between estimates and actual costs.

For this research, two approaches were taken to examine the reasonableness and realism of school cost of living estimates. The first is a proportional evaluation, as was used in the 2017 paper by Kelchen, where costs deltas are expressed as a percentage difference from the independent estimates. This provides an opportunity for an ‘apple-to-apple’ comparison to the initial research and allows for normalized assessments across various colleges and universities. The other metric that we are introducing for this research is expressing differences in absolute dollar terms. Expressing differences this way may provide a more meaningful perspective to potential users, namely students or parents, as they learn about the non-academic costs of attending college.

From the perspective of the first approach where cost variances are proportional, the findings from this research generally fall in line with the 2017 research results. The analysis reveals that a sizeable portion of schools produce estimates that vary considerably and thus puts into question the reliability of their estimates. As shown in Figure 15, we find that half of school estimates for an Off Campus scenario vary by at least 20%. The variances center around zero and appear to be roughly normally distributed. When looking at the With Family scenario, the proportion of schools that exhibit a variance of at least 20% jumps to more than 80%. The distribution for this scenario detectably shifted to the left or underestimated size of zero, an indication that most of the estimates are underestimated when compared to the estimates developed for this research.



Figure - Living Expense Variance by Percent

An additional layer to this analysis is categorizing how these distributions fall into five distinct groupings, providing an adjectival description and unique color for each category with the expectation that this provides an easier way to assess the quality of an estimate. The five categories, their intervals, and the proportion of estimates that fall into each is captured with the plots in Figure 16.

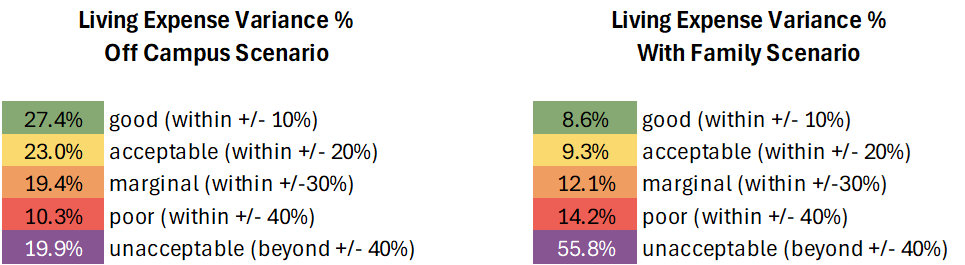


Figure - Living Expense Grading by Percent

This shows that 27% of estimates for the Off Campus scenario are classified as good while about half are considered to have marginal or worse estimates. Estimates for the With Family scenario skew much more negative, with more than half of them considered unacceptable. Less than 10% of estimates fall into the good category.

The next perspective looks at distributions not of variance by percentage difference but rather the absolute dollar difference. As demonstrated by the distributions in Figure 17, the change to dollar comparisons is considerable. Similar trends exist as before, with Off Campus estimate differences centering around zero. The distribution for the With Family scenario also follows, with the bulk of the distribution hovering on the negative side of zero, signaling that most estimates for students living at home are underestimating the true costs they are likely to encounter.



Figure - Living Expense Variance by Dollars

As was done when exploring variance by percentage difference, the findings were classified into one of five categories. Here, the thresholds are incremented in units of $1,000 until $4,000, after which all other variances are combined. This demonstrates that 37% and 32% of estimates are acceptable or better, for the Off Campus and With Family scenarios, respectively.

Additionally, a considerable issue needs to be highlighted for the With Family scenario. As it is understood by this research team, the current reporting system for IPEDS only allows colleges to enter values in the Other Expenses cost category. The Room and Board field is not accessible when modeling costs when the student is living with relatives. This shortcoming creates an issue when schools model costs for housing and food. The schools are faced with assuming, de facto, that these respective costs are costs are zero or it requires those cost estimates to be entered into the Other Expense category, which then misclassifies those costs and inflates all other expenses. This necessarily causes estimation issues across the entirety of this scenario as schools are forced to decide if they misclassify housing and food costs or do not include them at all.



Figure - Living Expense Grading by Dollars

Another question addressed by the findings focuses on if there are significant differences between cost estimates from different types of institutions. One might expect that for profit schools, referred to as Proprietary within the IPEDS data, would have worse estimates than other school types, as underestimating costs makes them look more affordable. That notion is not supported within the data. Calculations show that, using both mean and median measures, there are not out outsized differences between school types. As demonstrated in Figure 19, the Proprietary schools displayed better estimates, that is, closer to zero, than Public or Private Nonprofit institutions.



Figure - Living Expense Variance by Institution Type

With these findings and related research, we have identified several areas where changes to processes or policy would likely yield considerable benefit. We recommend the following:

* A standardized set of guidelines for institutions to develop estimates, to include the basis and rationale for such estimates
* Break the IDEPS cost reporting structure down to more specific levels; it currently aggregates many cost elements together and schools only have two reporting categories: Room and Board, and Other Expenses
* Institute a regular review of the assumptions and collegiate landscape such that the estimating philosphy is a reflection of the current economic environment and also mirrors the habits and trends of the current student body
* Consider developing a range instead of a singular point estimate; for instance, a series of Low, Likely, and High estimates may prove to be more useful while conveying a degree of variation for living costs

# Summary

The project aims to address the disparity between institution-reported estimates and independently calculated living expenses for higher education. It identifies that these discrepancies can significantly impact students' financial planning and aid distribution, leading to potential barriers to accessing and completing higher education. By leveraging datasets from sources like the EPI Family Budget Calculator, IPEDS, and College Scorecard, the project developed an interactive dashboard to compare and visualize these cost disparities across institutions.

The project's solution approach encompasses three key phases: data collection and cleaning, fine-tuning independent cost estimates, and visualization dashboard development. Through meticulous data conditioning and adjustment factors, the project ensures accuracy and reliability in comparing living expenses across different living scenarios and institutions. The algorithm specification outlines the process of harmonizing and visualizing the data using tools like Excel and Tableau Public, facilitating seamless integration and user-friendly visualization.

User stories cater to various stakeholders, including students, parents, policy makers, and university financial administrators, ensuring the dashboard's relevance and usability across the higher education ecosystem. Scenarios illustrate how the dashboard empowers users to make informed decisions, advocate for policy changes, and improve institutional cost estimation accuracy.

The findings underscore the disparities between institution-reported estimates and independently calculated values, particularly concerning living expenses. The analysis reveals that most institution estimates fall short and underestimate true living costs, highlighting the need for policy adjustments and process improvements in higher education finance.

Overall, the project's findings contribute to fostering transparency, equity, and informed decision-making in higher education finance. By providing stakeholders with actionable insights and tools to navigate cost disparities, the project aims to mitigate financial barriers and promote accessibility and affordability in higher education.

# Future Work

The biggest areas for future work lie within the data and assumptions used to develop the cost estimates.

Most of the estimates for this analysis were built using data that was reflective of a generalized adult, their lifestyle, and their consumer habits. It is very possible, and even likely, that college students exhibit distinctly different habits such that some subsets of the data used for this analysis may not be a strong proxy for the demongraphic being modeled. Not having to rely upon tangential data would help remove possible errors tied to fundamental differences between the general population and students. While every reasonable action was taken to normalize and adjust the data that formed the basis for our estimates, having data that explicitly examines college students would be very beneficial in developing and validating such cost estimates.

Another area that we welcome future effort and focus towards is the revision of the assumptions that were made for this analysis. Due to the lack of direct college expense data, various estimating assumptions were required to make use of the available data. While there is logic to support each assumption made, challenging those assumptions with newer data or different approaches may enhance the overall estimates and thus result in substantial improvement of the analysis.

Finally, allow a user to modify various cost assumptions such that those assumptions are more specific to their own criteria. This should result in a more personalized estimate and thus result in a more accurate and meaningful estimate.

Appendix

Appendix A: Glossary

|  |  |
| --- | --- |
| Term | Definition |
| Bureau of Labor Statistics (BLS) | An agency within the U.S. that measures, studies, and publishes economic and labor related activity |
| FAFSA | Free Application for Federal Student Aid |
| Net Price Calculator | Tool integrated into college websites that calculates the net cost of attendance, after grants and scholarships |
| Socioeconomic Status (SES) | Generally used by social scientists, it describes a person’s economic, financial, educational, and social status |
| YouTrack | Agile Management tool used for tracking progress, scrums, reporting, etc. |
| World Health Organization (WHO) | A United Nations agency that works to promote public health and safety. |
| EPI | Economic Policy Institute, a nonprofit, nonpartisan think tank created to include the needs of low- and middle-income workers in economic policy discussions |
| COA | Cost of Attendance, an estimate of both academic and living costs that a student would likely face while attending a given college or university |

Table : Glossary Table

Appendix B: GitHub Repository

Overview

Documentation associated with the project was uploaded to a GitHub repository.

GitHub Repository Link

https://github.com/jsmoll/DAEN690-Financial-Navigators

GitHub Repository Contents

README.MD

* Introduction
* Link to dashboard

Project Report

Data Dictionary

Data File

Appendix C: Risks

Sprint 1 Risks

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Risk | Description | Probability | Impact | Mitigation |
| Incomplete Data​ | Data sets may be incomplete or lacking in content​ | Medium​ | Medium​ | Conduct a thorough review of available data sources and discuss data mitigation strategies with team​ |
| Tool Availability​ | Tools and associated licensing may not be available​ | Medium​ | High​ | Review tooling and licensing requirements with Program and Sponsor​ |
| Team Skills​ | The project team may have mis-matched skills for creating the desired outcomes​ | Low​ | High​ | Review project problem space and desired outcome with customer to ensure alignment​ |
| Abundance of Data Features ​ | Main data set contains thousands of features, many of which are likely to be unimportant for our task​ | High​ | Medium​ | Spend time and better understand data features, select assumed important feature subset – share and discuss with sponsors to leverage their knowledge and elicit feedback regarding feature selection ​ |
| Undefined or Evolving Sponsor Goals ​ | Sponsors may not have a very well-defined scope, or their idea of the deliverable may materially change over time​ | Medium​ | High​ | Weekly Sponsor meetings where progress and ideas are shared to guide and fortify project deliverable. Additional meetings or correspondence when necessary.  ​ |
| Not getting MIT data (in time)​ | We requested the MIT data but are unsure of the turn-around time nor certain what data will be available to us​ | Medium​ | High​ | Need to discuss with Sponsor to determine what other data sources can be used should we not have access to the data in time.​ |

Table : Sprint 1 Risks

The risks identified in Table 2 were discussed and identified in the first week of the sprint. Additional risks were added later as the problem space and problem goals were better understood through research and discussion with the project sponsors. The initial set of risks was meant to capture standard project risks that affect cost, schedule, and people. In this case, it concerned the team composition, the tool availability, and the data that formed the basis of the project.

We were correct in identifying these risks as both the tool availability and data availability remain at risk at the end of Sprint 1. Specifically, the project sponsors would like to use Tableau as the data visualization tool, but there are no licenses available. We plan to explore other products, to include the free-to-use version of Tableau. The backend of the database is also in flux as it was initially stated to be AWS and may now be an Excel spreadsheet. Finally, one of the primary data sets required to complete this project was the MIT Living Wage Calculator. The team has requested the data in accordance with MIT requirements, but there has been no response.

Sprint 2 Risks

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Risk | Description | Probability | Impact | Mitigation |
| Incomplete Data​ | Data sets may be incomplete or lacking in content​ | Low | Medium​ | Conduct a thorough review of available data sources and discuss data mitigation strategies with team​ |
| Tool Availability​ | Tools and associated licensing may not be available​ | Low​ | High​ | Review tooling and licensing requirements with Program and Customer​ |
| Undefined or Evolving Sponsor Goals ​ | Sponsors may not have a very well-defined scope, or their idea of the deliverable may materially change over time​ | Low | High​ | Weekly Sponsor meetings where progress and ideas are shared to guide and fortify project deliverable. Additional meetings or correspondence when necessary.  ​ |
| Cost Adjustment Factors | Finding data to support reasonable cost element adjustments may prove to be either difficult or incongruent | Medium | High | Rely upon the 2017 Kelchen paper as well as independent research to help guide and inform assumptions |
| Tableau Development | There may be limitations (software, data, or skill) to what features and visualizations are published | Medium | High | Open, honest lines of communication with Sponsors to ensure requests are feasible. Additionally, leverage various training and other sources to ensure successful end-deliverable |
| Transfer of Tableau Product | Tableau Public is linked to a user account and the transfer of the dashboard to another entity may be an issue | Low | Medium | Research best practices for dashboard hand-off to Sponsor at project end |
| Team Skills​ | The project team may have mis-matched skills for creating the desired outcomes​ | Closed | | |
| Abundance of Data Features ​ | Main data set contains thousands of features, many of which are likely to be unimportant for our task​ | Closed | | |
| Not getting MIT data (in time)​ | We requested the MIT data but are unsure of the turn-around time nor certain what data will be available to us​ | Closed | | |

Table : Sprint 2 Risks

The team continued to monitor and adjust the risk register throughout this Sprint. It was determined that the datasets appear to be largely complete; however, there continue to be data manipulation tasks required. As a result, the probability of the ‘Incomplete Data’ risk was changed to ‘Low’ to reflect the change. The same adjustments were made to ‘Tool Availability’ and ‘Evolving Sponsor Goals’ as Tableau Public appears to meet the needs of the project for data visualization and the Sponsor meetings have been very productive and organized.

There were three additional risks that were contemplated throughout this Sprint. Two of these risks relate to the use of Tableau, with ‘Tableau Development’ identifying potential issues with the dashboard construction and ‘Transfer of Tableau Product’ highlighting concerns with the transfer of the finalized product to the Sponsors. The third risk speaks to the uncertainty and magnitude to which cost adjustments will be made. These risks will continue to be monitored going forward for any changes or adjustments to mitigation actions.

Three risks were closed. Team skills were determined to be sufficient to meet the needs of the project. The ‘Abundance of Data Features’ risk was closed once a thorough analysis of the data features was completed and approximately 100 relevant features were identified. Finally, the MIT data was not available and will not be available going forward. However, the project Sponsors identified the EPI dataset which was a direct replacement for the information.

Sprint 3 Risks

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Risk | Description | Probability | Impact | Mitigation |
| Incomplete Data​ | Data sets may be incomplete or lacking in content​ | Low | Medium​ | Conduct a thorough review of available data sources and discuss data mitigation strategies with team​ |
| Undefined or Evolving Sponsor Goals ​ | Sponsors may not have a very well-defined scope, or their idea of the deliverable may materially change over time​ | Medium | High​ | Weekly sponsor meetings where progress and ideas are shared to guide and fortify project deliverable. Additional meetings or correspondence when necessary.  ​ |
| Cost Adjustment Factors | Finding data to support reasonable cost element adjustments may prove to be either difficult or incongruent | Medium | High | Rely upon the 2017 Kelchen paper as well as independent research to help guide and inform assumptions |
| Tableau Development | There may be limitations (software, data, or skill) to what features and visualizations are published | Medium | High | Open, honest lines of communication with Sponsors to ensure requests are feasible. Additionally, leverage various training and other sources to ensure successful end-deliverable |
| Transfer of Tableau Product | Tableau Public is linked to a user account and the transfer of the dashboard to another entity may be an issue | Low | Medium | Research best practices for dashboard hand-off to Sponsor at project end |

Table : Sprint 3 Risks

There were no new risks added during this sprint; however, several risks did trigger:

* Incomplete data was discovered as Penn State University did not have cost data in the College Scorecard Dataset. In this case, the school will be ignored. The data is available from the prior year, but there is not a practical solution for incorporating data based on this problem.
* Evolving Sponsor goals triggered when they requested a fundamental change in how a typical student was calculated. The team will schedule a meeting with the sponsors to discuss the scenarios and hopefully come to an agreement on a workable way forward.
* Cost adjustment factors continue to trigger as the team has been unable to find any research or data that specifically addresses the use cases of student costs. This involved transportation costs as well as a student living with relatives. In both cases, several assumptions were required based on the best data available and then vetted through the sponsors.

The team continues to actively review the risks weekly and address via mitigation plans as appropriate.

Sprint 4 Risks

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Risk | Description | Probability | Impact | Mitigation |
|  |  |  |  |  |
| Incomplete Data​​ | Data sets may be incomplete or lacking in content​​ | Low​ | Medium​​ | Conduct a thorough review of available data sources and discuss data mitigation strategies with team​​ |
| Undefined or Evolving Sponsor Goals ​​ | Sponsors may not have a very well-defined scope, or their idea of the deliverable may materially change over time​​ | **Low**​  **(was Medium**)​ | High​​ | Weekly sponsor meetings where progress and ideas are shared to guide and fortify project deliverable. Additional meetings or correspondence when necessary.  ​​ |
| Cost Adjustment Factors​  (Closed)​ | ~~Finding data to support reasonable cost element adjustments may prove to be either difficult or incongruent~~​ | ~~Medium~~​ | ~~High~~​ | ~~Rely upon the 2017 Kelchen paper as well as independent research to help guide and inform assumptions~~​ |
| Tableau Development​ | There may be limitations (software, data, or skill) to what features and visualizations are published ​ | Medium​ | High​ | Open, honest lines of communication with Sponsors to ensure requests are feasible. Additionally, leverage various training and other sources to ensure successful end-deliverable​ |
| Transfer of Tableau Product​ | Tableau Public is linked to a user account and the transfer of the dashboard to another entity may be an issue ​ | Low​ | Medium​ | Research best practices for dashboard hand-off to Sponsor at project end​ |

Table : Sprint 4 Risks

There were no new risks added during this Sprint; however, there were two adjustments to existing risks.

* Undefined or Evolving Sponsor Goals probability was adjusted to low based on the weekly Sponsor meetings and the generally good rapport that the project team has established with the Sponsors.
* Cost Adjustment Factors was closed as the project team established all the relevant adjustment factors and validated them with the Sponsors.

The team continues to actively review the risks weekly and address via mitigation plans as appropriate.

Sprint 5 Risks

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Risk | Description | Probability | Impact | Mitigation |
| Incomplete Data​​ | Data sets may be incomplete or lacking in content​​ | Low​ | Medium​​ | Conduct a thorough review of available data sources and discuss data mitigation strategies with team​​ |
| Undefined or Evolving Sponsor Goals ​​ | Sponsors may not have a very well-defined scope, or their idea of the deliverable may materially change over time​​ | Low​ | High​​ | Weekly Sponsor meetings where progress and ideas are shared to guide and fortify project deliverable. Additional meetings or correspondence when necessary.  ​​ |
| Tableau Development​ | There may be limitations (software, data, or skill) to what features and visualizations are published ​ | Medium​ | High​ | Open, honest lines of communication with Sponsors to ensure requests are feasible. Additionally, leverage various training and other sources to ensure successful end-deliverable​ |
| Transfer of Tableau Product​ | Tableau Public is linked to a user account and the transfer of the dashboard to another entity may be an issue ​ | Low​ | Medium​ | Research best practices for dashboard hand-off to Sponsor at project end​ |

Table : Sprint 5 Risks

There were no new risks added during this Sprint. Tableau Development did trigger three issues:

* + Tableau does not allow multi-user, simultaneous updates which complicates the ability of the project team to fully utilize all their members. As a result, Tableau development was a serial process. It was difficult to overcome this inherent limitation of the software as each team member has different schedules and skills, making it very difficult to even schedule work periods between team members. While this did slow progress, it did not stop it. Ideally the project would have more time to complete comprehensive dashboards, or the team was assigned a Tableau expert capable of visualizing the data quickly.
  + Tableau Public does not allow connection to a database which greatly limits the flexibility to update or modify the data. Tableau Public essentially copies the spreadsheets up to the servers and any dashboard created will be tied to the copied instance. While updating the data is possible, it is more complicated than using a linked database and will likely break several elements within the dashboard.
  + The Sponsors continued to request changes after the last class and before the final presentation. The requests were reasonable; however, time is very limited at the end of a project. As an example, Hildreth asked for Healthcare costs to be removed from all calculations which required a rework of the dashboard, report, etc.

The team continues to actively review the risks weekly and address via mitigation plans as appropriate.

Appendix D: Agile Development

Scrum Methodology

Figure 20 - Sprint Project Dates

The project required the use of agile and scrum methodologies. Due to the varied work, home, and school commitments on the team, it was only possible to meet 5 times a week in the evenings. Task monitoring was accomplished using cards in the YouTrack tool which promoted visibility and accountability. This approach not only facilitated organized task management but also ensured a balanced distribution among team members, resulting in timely and successful completion of assignments.

YouTrack is a project management and issue tracking tool designed by JetBrains capable of supporting teams practicing Scrum methodologies. As concerns this project, YouTrack serves to record and organize user stories, tasks, and other work items in the form of cards. These cards can be moved through various stages on a kanban board, reflecting the progress of the work. YouTrack was easy to use and helpful prior to and during scrum meetings to assign and track tasks within the team.

Sprint 1 Analysis

During this Sprint, the team's approach to identifying user stories involved an exploration of the problem statement to better understand the various users and their needs. There was a review of the research papers and articles provided by the sponsors followed by follow-on research as the team better understood the problem space. Additionally, e team met with both the professor and the project sponsor to gather their perspectives and refine the collective understanding of the users’ requirements. During this process, tasks were generated to help focus the team’s efforts and practice scrum, which is an important part of new team formation.

The team performed well during Sprint 1. They rapidly agreed on roles, scrum cadence, and transitioned into utilizing YouTrack for task management. The Scrum meetings provided an opportunity to review progress, answer questions, and address any blockers encountered by team members. Overall, the team was successful in creating the required work products and briefing the class and professors weekly.

In terms of areas for improvement, the team is still in the formation stage and would benefit from improved communication and teamwork. As the team is comprised of individuals from diverse backgrounds and work experience, ongoing efforts to foster a collaborative and communicative environment will be crucial. Additionally, the team might consider implementing more structured feedback mechanisms or retrospectives to identify and address challenges proactively. Embracing lessons learned from this initial sprint, the team is well-positioned to iterate on their strategies for subsequent phases of the project, contributing to a more cohesive and efficient working dynamic.

Sprint 2 Analysis

During this Sprint, the team shifted their approach to card management within YouTrack. Specifically, in the last Sprint, each interim week, a new card was created for updating reports, slides and such. While this worked, it created a lot of overhead that cluttered the board with overlapping tasks. As an example, there would be an ‘Update Risks’ for interim 1 and an ‘Update Risk’ for interim 2. This also meant that notes entered into the cards by various team members would be spread across the 3 week Sprint. In order to get a clear picture of why changes were made, this required bringing up all three cards associated with that item.

For this Sprint, a single card was created for each slide and project section and was moved between ‘In Progress’ and ‘Review’ with ‘Done’ being the final state of truly being done and ready for final submission at the conclusion of the Sprint. This allowed for all reviewer comments to be consolidated in the same card which made it much easier to update the deliverable in question.

Sprint 3 Analysis

The team has settled into a cadence of reviewing cards and performing tasks. Several members continue to struggle with properly documenting task status within the cards, which the team will collectively work on into the next sprint. There were no new User Stories.

The primary challenge to more efficient Sprints and Scrums revovle arount students having day jobs. It is difficult to find common times to meet with some students unavailable in the mornings or evenings on given days.

Sprint 4 Analysis

The team presented a draft Tableau board to the Sponsors as a part of the sprint to gain feedback and concurrence of the direction of the development. This was a very productive discussion and helped both the team and the Sponsors understand what a final product would look like given the time remaining. In particular, it was determined that the project team should focus on the policy maker role more than on a student role as the policy makers would be the first users.

The team continues to struggle with day jobs and other responsibilities while meeting regularly for scrum.

Sprint 5 Analysis

The team presented an updated draft of the Tableau dashboard to the Sponsors to iterate on the development. There were two fundamental shifts: remove the “on-campus” scenerio as it did not provide enough value and removed healthcare as a cost as industry partners indicated that would be part of fees. Additionally, the Sponsors provided written feedback on specific items that required adjustment. A lesson learned in this Sprint was to use a paid version of Tableau to facilitate data changes.

As with previous Sprints, the team had challenges balancing full time jobs and differing availability schedules.

Reference

Works Cited

|  |  |
| --- | --- |
| [1] | "Education pays, 2022," Bureau of Labor Statistics, May 2023. [Online]. Available: https://www.bls.gov/careeroutlook/2023/data-on-display/education-pays.htm. |
| [2] | P. Tough, "Americans Are Losing Faith in the Value of College. Whose Fault Is That?," *New York Times,* 5 September 2023. |
| [3] | L. Welding, R. Long and M. Rose, "College Enrollment Statistics in the U.S.," 8 August 2023. [Online]. Available: https://www.bestcolleges.com/research/college-enrollment-statistics/. |
| [4] | K. P. Lindgren, T. Tristao and C. Neighbors, "The Association Between Student Loan Debt and Perceived Socioeconomic Status and Problematic Drinking and Mental Health Symptoms: A Preliminary Investigation," *Addictive Behaviors,* vol. 139, 2022. |
| [5] | A. Ford and L. Egan, "Student loan debt in America, in four charts," NBC News, 24 August 2022. [Online]. Available: https://www.nbcnews.com/data-graphics/student-loan-debt-america-charts-rcna44439. |
| [6] | National Student Clearinghouse, "Some College, No Credential Student Outcomes," 25 April 2023. [Online]. Available: https://nscresearchcenter.org/some-college-no-credential/. |
| [7] | J. P. Mackenbach, I. Stirbu, A.-J. R. Roskam, M. M. Schaap, G. Menvielle, M. Leinsalu and A. Kunst, "Socioeconomic Inequalities in Health," *The New England Journal of Medicine,* vol. 358, pp. 2468-2481, 2008. |
| [8] | H. Bosma, C. Schrijvers and J. P. Mackenbach, "Socioeconomic Inequalities in Mortality and Importance of Perceived Control: Cohort Study," *British Medical Journal,* vol. 319, no. 7223, pp. 1469-1470, 1999. |
| [9] | T. Richardson, P. Elliot and R. Roberts, "The Relationship Between Personal Unsecured Debt and Mental and Physical Health: A Systematic Review and Meta-Analysis," *Clinical Psychology Review,* vol. 33, pp. 1148-1162, 2013. |
| [10] | K. M. Walsemann, G. C. Gee and D. Gentile, "Sick of Our Loans: Student Borrowing and the Mental Health of Young Adults in the United States," *Social Science & Medicine,* vol. 124, pp. 85-93, 2015. |
| [11] | T. F. Babor, J. C. Higgins-Biddle, J. B. Saunders and M. G. Monteiro, *AUDIT - The Alcohol Use Disorders Identification Test,* World Health Organization, 2001. |
| [12] | Hildreth Institute, *Transparency in Cost of College Proposal.* |
| [13] | R. Kelchen, S. Goldrick-Rab and B. Hosch, "The Costs of College Attendance: Examining Variation and Consistency in Institutional Living Cost Allowances," *The Journal of Higher Education,* pp. 947-971, 2017. |
| [14] | A. K. Glasmeier, "Living Wage Calculator," Massachusetts Institute of Technology, 2024. [Online]. Available: https://livingwage.mit.edu. |
| [15] | Department of Education, "2023-2024 Federal Student Aid Handbook," 2024. [Online]. Available: https://fsapartners.ed.gov/knowledge-center/fsa-handbook/2023-2024/vol3/ch2-cost-attendance-budget. |
| [16] | S. Burd, R. Fishman, L. Keane and J. Habbert, "Decoding the Cost of College, The Case for Transparent Financial Aid Award Letters," June 2018. [Online]. Available: https://d1y8sb8igg2f8e.cloudfront.net/documents/Decoding\_the\_Cost\_of\_College\_Final\_6218.pdf. |
| [17] | K. Degnan, "Net Price Calculators: Part of the Problem, Not the Solution," 10 October 2022. [Online]. Available: https://collegeaidpro.com/net-price-calculators-part-of-the-problem-not-the-solution/. |
| [18] | Department of Education, "Higher Education Opportunity Act - 2008," 11 December 2023. [Online]. Available: https://www2.ed.gov/policy/highered/leg/hea08/index.html. |
| [19] | A. Weisman, "(GEN-23-11) FAFSA Simplification Act Changes for Implementation in 2024-25," 04 August 2023. [Online]. Available: https://fsapartners.ed.gov/knowledge-center/library/dear-colleague-letters/2023-08-04/fafsa-simplification-act-changes-implementation-2024-25. |
| [20] | L. A. Davis, G. C. Wolniak, C. E. George and G. R. Nelson, "Demystifying Tuition? A Content Analysis of the Information Quality of Public College and University Websites," *Aera Open,* 5 August 2019. |
| [21] | Z. Mokhiber, E. Gould and K. deCourcy, "The Economic Policy Institute's Family Budget Calculator," 30 January 2024. [Online]. Available: https://www.epi.org/publication/family-budget-calculator-documentation/. [Accessed April 2024]. |
| [22] | B. Miller, C. Campbell and B. J. Cohen, "Addressing the $1.5 Trillion in Federal Student Loan Debt," 12 June 2019. [Online]. Available: https://www.americanprogress.org/article/addressing-1-5-trillion-federal-student-loan-debt/. |
| [23] | A. M. Perry, M. Steinbaum and C. Romer, "Student loans, the racial wealth divide, and why we need full student debt cancellation," 23 June 2021. [Online]. Available: https://www.brookings.edu/articles/student-loans-the-racial-wealth-divide-and-why-we-need-full-student-debt-cancellation/. |
| [24] | J. Ma, "Trends in College Pricing and Student Aid 2023," 2023. [Online]. Available: https://research.collegeboard.org/media/pdf/Trends%20Report%202023%20Updated.pdf. |
| [25] | A. Stauffer, P. Oliff, C. Foard, R. Thiess, M. Reese, B. Samms, M. Bryant, L. Pontari, E. Spence and J. Einsiedler, "Two Decades of Change in Federal and State Higher Education Funding," Oct 2019. [Online]. Available: https://www.pewtrusts.org/-/media/assets/2019/10/fedstatefundinghigheredu\_chartbook\_v1.pdf. |
| [26] | National Center for Educational Statistics, "IPEDS Integrated Postsecondary Education Data System (Use of the Data)," [Online]. Available: https://nces.ed.gov/ipeds/use-the-data#SurveyData. [Accessed April 2024]. |
| [27] | Department of Education, "College Scorecard," 23 October 2023. [Online]. Available: https://collegescorecard.ed.gov/data/. [Accessed April 24]. |
| [28] | George Mason University, College of Engineering and Computing, "Supplementary Course Data," [Online]. Available: https://mymasonportal.gmu.edu/ultra/courses/\_508689\_1/cl/outline. [Accessed 04 2024]. |
| [29] | O. o. E. E. &. R. Energy, "FOTW #1248, July 25, 2022: The Average U.S. Household Spent Nearly $10,000 on Transportation in 2020," 25 July 2022. [Online]. Available: https://www.energy.gov/eere/vehicles/articles/fotw-1248-july-25-2022-average-us-household-spent-nearly-10000. [Accessed March 2024]. |
| [30] | U. B. O. L. STATISTICS, "Consumer Expenditure Surveys," 25 October 2023. [Online]. Available: https://www.bls.gov/cex/tables.htm. [Accessed March 2024]. |

**This page intentionally left blank**