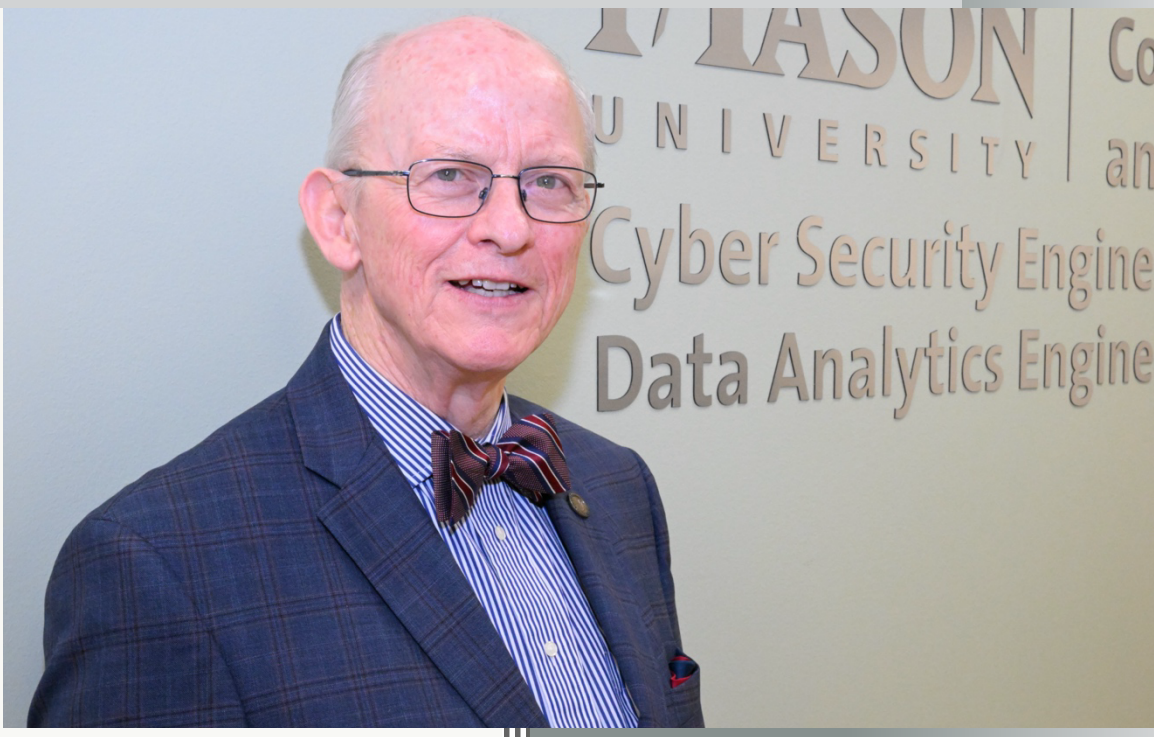
GEORGE MASON UNIVERSITY

Data Analytics Engineering

Spring 2024

Balancing the Scales: Navigating College Affordability in Contrast to U.S. Median Household Income



DAEN 690

Project Report

Chaitanya Pallamreddy
Keerthi Reddy Chintha Reddy
Lakshmi Prasanna Sai Nelakuditi
Narasimha Krishna Amruth Vemuganti
Sahith Kumar Patlori
Sai Krishna Marikukala

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

Table of Figures

This page intentionally left blank

Abstract

This research tackles the urgent problem of rising tuition prices and how they affect accessibility in US higher education. Rising tuition has outpaced increase in household income over the last three decades, causing social mobility to be hampered and economic inequality to worsen. The suggested remedy is creating a sophisticated interactive dashboard that compares the growth of tuition and fees with median household income across different geographic locations by utilizing cloud-based infrastructure and analytics. With the use of transparent data visualizations and real-time analytics, this user-friendly platform seeks to empower stakeholders, including researchers, politicians, and students, to support policy changes, make informed decisions, and create a more accessible and affordable higher education environment. The goal of the project is to offer a special degree of in-depth research and openness, acting as a useful tool for students and families navigating the challenges of affording college as well as a potent lobbying tool for legislative change.

This page intentionally left blank

Report

Section 1: Problem Definition

1.1 Background

Higher education is largely acknowledged as the primary factor influencing social and economic growth in the United States. The chance of students to enroll in and complete their college education has been severely limited over the course of the last thirty years by increasing educational expenses and the allocation of financial stress to students via different charges. Adding to previous inequities caused by race and socioeconomic status, this trend has caused significant obstacles for students, particularly those from low-income homes and those with limited opportunity.

Educational expenses have significantly risen in recent years. Tuition and fee hikes are directly caused by cuts in governmental financing for public educational institutions. However, private non-profit universities have also seen cost increases because of this trend, thus it is not limited to public institutions. Public universities are more cost-effective than private ones. Regarding a public university, the average full-time student at a private non-profit college spent $48,965 in 2019–2020, which is $21,035 higher.

The latest findings from the Georgetown University Center for Education and the Workforce reveal a stark reality - the cost of obtaining a college degree has exponentially risen in the past four decades. From past four decades, the median expenses for tuition, housing, and other associated costs have climbed by a whopping 180%. Inflation has also played a significant factor, with an item that previously cost $10,000 now equivalent to a staggering an approximate of $29,000. Income for those aged 22-27 have only grown by 19% during a 40-year period, which is insufficient to keep up with the rising cost of education. The results highlight the considerable financial obstacles faced by persons pursuing a higher degree.

It might be difficult to guarantee the advantages of a college education because of the diverse ethnic and socioeconomic backgrounds of students nowadays. Financial limitations, increased tuition fees, and stagnating family incomes are hindrances that make it harder to fully benefit from higher education.

The increasing tuition costs create an urgent threat to the cost and availability of higher education. Therefore, students and their families, especially those from different cultures and with little financial resources, struggle to pay for college or end up overburdened with large debt. Due of the considerable obstacles that low-income, non-traditional, and students of color have historically encountered when pursuing a college degree, these groups will be most badly impacted by this. Furthermore, the increase in costs endangers not just the individual prospects of these students but also the future of entire governments and towns, which depend more and more on a workforce with higher levels of education.

1.2 Problem Space

Analysis:
Economic Disparities and Barriers to Accessibility in Higher Education

The complex web of economic disparities plaguing higher education in America greatly impacts the accessibility for a sizable portion of the population. In the past four decades, the cost of attending college

has risen at a significantly higher rate than the average income growth, creating major barriers for aspiring students.

Growing Expenses in Contrary to Stalling Wages

Over the past four decades, the costs associated with attending a four-year college have skyrocketed by a staggering 180%. These expenses, which encompass housing and board, tuition, and fees, have significantly outpaced the modest 19% increase in salaries for young workers aged 22 to 27 during the same time frame. As a result, students and their families are burdened with a substantial financial strain while striving to obtain a higher education.

What Is the True Cost of An Undergraduate Four-Year Degree?

Since the 1980s, the cost of attending college has steadily risen. Interestingly, the National Center for Education Statistics reported that during the 2015-2016 academic year, the typical annual expense for attending a public four-year university was a staggering $19,000. Private universities, on the other hand, come with a hefty price tag of around $40,000, which includes not only tuition and fees but also room and board.

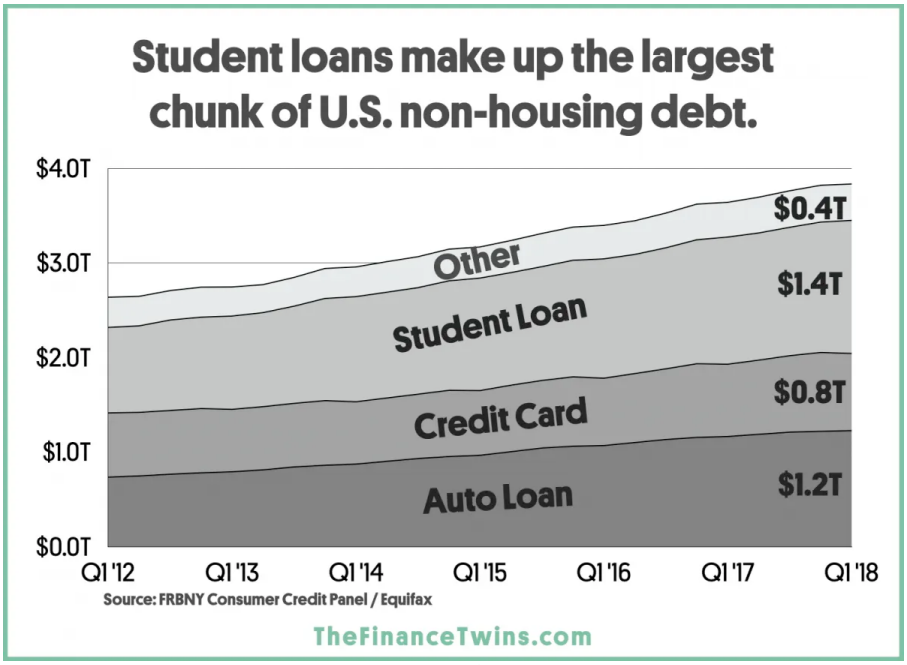


Figure 1. Student loans account for $1.4T of U.S. Household Debt

Obtaining a college degree has become increasingly costly. As a result, most of the non-mortgage debt in the United States is now made up of student loans, exceeding the combined amount of personal loans and credit card debt.

What could be the reason behind the hikes?

One contributing factor is the policies of colleges and universities, as well as reductions in state funding. As state financial support for public institutions decreases, students and their families are left with a larger financial burden, leading to higher tuition fees. This trend is not limited to public institutions, as private non-profit colleges also have a high cost of attendance, further worsening the issue of affordability.

The Requirement for Competent Solutions

To effectively combat the challenges within this problem domain, a data-driven approach is essential. We must have access to high-quality tools and platforms that are capable of comprehensively evaluating, visually presenting, and effectively communicating the disparities between struggling household incomes and rising college costs. These invaluable resources will play a vital role in advocating for policy change, informing institutional improvements, and ultimately promoting greater accessibility to higher education for all members of our society.

1.3 Research

Analysis of tuition growth rates based on clustering and regression models, explores the factors affecting tuition growth rates in higher education. It critiques existing studies for their lack of comprehensive analysis and limited data, proposing a detailed examination through clustering and regression models using extensive authentic data. Key factors considered include university ranking, class size, faculty composition, and macro-economic indicators like unemployment rates and government policies. Utilizing a dataset from the National Center for Education Statistics, the study segments schools and analyzes tuition growth trends, applying various quantitative methods to understand the dynamics of tuition fees in the United States. The findings highlight the complex interplay of in-school and external factors on tuition rates, offering insights into the broader implications for policy and educational financing.

Demystifying Tuition? A Content Analysis of the Information Quality of Public College and University, investigates how public colleges and universities in the United States present tuition and cost information on their websites. It focuses on analysing the quality, clarity, and accessibility of financial information, examining how the way this information is conveyed might influence prospective students and families' understanding and decisions regarding higher education. The study aims to uncover patterns in the presentation of this crucial information and assess its impact on the audience's perceptions of tuition and costs.

1.4 Solution Space

The central focus of the project's solution approach revolves around developing an interactive dashboard that addresses two key components: the issue of escalating tuition costs, which are not tied to median household income, and their impact on the accessibility of higher education in the United States. Additionally, it includes an examination of tuition fees in both profit and non-profit educational institutions in relation to the cost of living.

Analytics and Discrepancy Detection:

Develop a dedicated analytics module within the dashboard to process the integrated data efficiently. Calculate and analyse the discrepancies between the growth rates of tuition and fees and median household income.

Cloud-Based Infrastructure:

Create an effective, cloud-based framework for streamlined management of data storage, retrieval, and real-time analytics. Ensure that this framework can effectively manage large volumes of data, conduct rapid analysis, and scale as required to meet future demands.

User-Friendly visualizations:

Ensure the dashboard is available to a wide range of users, including policymakers, educators, students, and parents, irrespective of their technical or financial knowledge. Promote user involvement by enabling them to personalize data displays, implement filters, and monitor past trends.

Interactive Dashboard Features:

Develop an intuitive web-based dashboard that offers users the ability to seamlessly explore and interact with the data. Implement advanced filtering and customization functionalities that empower users to personalize their queries, considering factors such as distinctions between public and private higher education institutions, geographical regions, income categories, and other relevant data attributes.

1.5 Project Objectives

Our preliminary focus is to create an immersive and comprehensive dashboard that compares the rise of tuition and fees to the growth of median household income in different areas of the United States. The dashboard acts as a powerful advocacy tool besides being a bolster to the common parents and students who are uncovering the complicated tuition and living fee structure of the institutions. This advocacy tool can be utilized by many student-cantered affiliates and professionals looking to reform institutional policies making financial aid available for the deserving students. The tool will gather data from reputable sources such as the U.S. Census Bureau, Bureau of Economic Analysis, U.S. Bureau of Labour Statistics, Federal Reserve Economic Data, and National Centre for Education Statistics. Its purpose is to facilitate a thorough analysis, allowing users to identify trends, patterns, and discrepancies in the cost of education compared to income at a county level. Special attention will be given to organizing data for both public and private institutions of higher education, providing valuable insights for stakeholders.

1.6 Primary User Stories

Non-profit Organization

Non-profit Organization Scene: Emily, a passionate member of a non-profit organization dedicated to promoting educational equity, skilfully utilizes the dashboard to conduct in-depth research on the intricate relationship between race, income, and college affordability. Through thorough analysis of the disparities in tuition fees and average incomes within marginalized communities, Emily is empowered to effectively advocate for necessary resources and support services that cater to the specific needs of underserved student populations.

Student Decision-Making

Scenario: As a high school senior, Sarah is facing the important decision of selecting which colleges to apply to. She turns to the interactive dashboard to compare the varying tuition costs and median household incomes of her top college choices. With the aid of this data, Sarah can make well-informed decisions about which college will best suit her academic and financial needs.

Parent Financial Planning

Scenario: Mark and Lisa, parents of two college-bound children, are deeply troubled by the soaring expenses of pursuing higher education. In their quest for solutions, they turn to the dashboard for insights into the evolution of tuition costs in relation to household incomes. With these findings in hand, they devise a strategic financial strategy to prepare for their children's future college expenditures.

1.7 Product Vision

Our project aims to develop a highly advanced and user-friendly interactive dashboard that will play a central role in supporting student-centered advocates at the Hildreth Institute. These advocates are actively involved in reforming college policies to tackle the urgent challenges of affordability and accessibility in higher education.

This dashboard will provide a unique level of in-depth analysis and transparency regarding the growing disparity between the rising costs of higher education and the slow growth in household incomes across the United States by utilizing cutting-edge cloud-based infrastructure and data analytics.

This project aims to serve as both a powerful advocacy tool for policy change in higher education financing and a valuable resource for students and families, offering clear, accessible visualizations and insights to navigate college affordability complexities and promote equitable access to education.

1.7.1 Scenario #1

For: Advocacy groups and policymakers within the Hildreth Institute's team.

Who: Strive to rectify discrepancies in the affordability of higher education

The: Solution is a highly advanced interactive dashboard

Is a: Tool for analysing and visualizing data related to the disparity between rising cost of attendance and stagnant household incomes.

The dashboard compares tuition and fees to median household incomes, allowing advocacy groups and lawmakers to highlight critical higher education affordability challenges.

Visualizations can highlight discrepancies and promote specific approaches for distinct demographic groups, and even execute trends over time to determine efficiency.

This method, which relies on evidence, gives policymakers the ability to support changes, encouraging teamwork and involvement from the community to make significant enhancements in financing for higher education.

1.7.2 Scenario #2

For: High School Students and Parents

Who: Are highly motivated individuals aiming for college attendance, yet worried about financial constraints

The: user-friendly interactive dashboard

Is a: A revolutionary tool providing in-depth financial analysis, serving as a resource for understanding and managing the costs of higher education and making well-informed decisions regarding college choices.

Steps to Utilize the College Affordability Dashboard Effectively

1.Access the dashboard via web browser or app, then navigate the user-friendly interface featuring key data categories like tuition, costs, and income.

2.Select your state, county, or city to customize data to your area.

3.Use filters to refine your view, selecting between public/private institutions, degrees, or urban/rural settings.

4.Visualize affordability by comparing college costs to local median income with clear graphs and maps.

5.By comparing institutions and predicting expenses within your budget, you may make wise decisions.

Section 2: Datasets

2.1 Overview

The dataset overview highlights the importance of collecting reliable data from respected sources like the Integrated Postsecondary Education Data System (IPEDS) and the Federal Reserve Economic Data (FRED). IPEDS offers comprehensive data on higher education, including state-specific price information that is essential for analysing tuition patterns. FRED provides detailed financial indicators, including median family income data for each state, which is crucial for calculating college expenses.

The summary outlines the precise data variables gathered, including unique identity numbers, tuition fees, living costs, and extra expenses. This thorough method guarantees a detailed examination of developments in higher education affordability and accessibility. The initiative intends to provide significant insights into the intricate relationship between educational expenditures and economic issues by utilizing data, with the goal of aiding informed decision-making and policy development in higher education.

2.2 Field Descriptions

Collecting data is critical for modern research to offer valuable information, with IPEDS and FRED providing as essential sources for studying education and economic trends in the United States.

IPEDS is an extensive database that contains statistics regarding postsecondary education in the US. A wealth of information is offered on all facets of university education, including enrollment, graduation rates, expenses, and institutional characteristics. Our study issue is highly relevant to the full picture of state-by-state educational pricing provided by IPEDS, which offers tuition and fee information for both public and private universities.

In addition to IPEDS, the Federal Reserve Bank of St. Louis curates FRED, which is the leading source of financial data in the US. FRED's provision of a wide range of economic indicators facilitates the analysis of patterns and trends at the local, national, and worldwide levels. Specifically, FRED provides information on median household income for each state in the US for this study, which helps put the financial context of college expenditures in perspective.

1. **Unit_ID (**Type: Integer) : Unique Identification number

2. **Institution_name** (Type: Character): Name of the University

3. **State** (Type: Character): State in which university is located.

4. **Year** (Type: Integer): Year in which students attend an educational institution.

5. **Median Income** (Type: Integer): Income of the family living in a particular state.

6. **Degree Type** (Type: Character): Classification of academic degrees, such as associate or bachelor's degrees, based on the duration and depth of the educational program.

7. **Sector** (Type: Character): Institutions are classified as public, private not-for-profit, or private for-profit depending on how they operate and who funds them.

8. **In_district_tuition** (Type: Integer): The cost of attendance for students who are reside within the geographic area served by the institution.

9. **In_State_tuition** (Type: Integer): The cost of attendance for students who are residents of the state in which the institution is located.

10. **Out_of_state_tuition** (Type: Integer): The cost of attendance for students who do not reside within the state in which the institution is located.

11. **On_campus_room_and_board** (Type: Integer): The total expenses associated with residing on campus, including room, board, and other living costs.

12. **Off_campus_room_and_board_without_family** (Type: Integer): The estimated expenses for students living independently off-campus.

13. **On_campus_other_expenses** (Type: Integer): Indicates the estimated expenses for personal requirements such as laundry, transport, entertainment, and household goods, not covering costs associated with living accommodations, meals, or educational fees.

14. **Off_campus_other_expenses_without_family** (Type: Integer): Represents the expenses related to studying at the institution for those living with family off-campus, not including tuition or housing costs.

15. **Off_campus_other_expenses_with_family** (Type: Integer): Represents the expenses for living and studying away from family off-campus, covering personal and transportation costs but not including tuition or housing/meals.

16. **Books_and_supplies** (Type: Integer): Represents the estimated cost of textbooks and necessary educational materials for a given academic period.

2.3 Data Context

Data has been sourced from official government websites, specifically the National Center for Education Statistics (NCES) and the Federal Reserve Economic Data (FRED) by Federal Reserve Bank of St. Louis. FRED is used to extract the median household income of a person living in a particular state every year from 2002 to 2022. The NCES provides extensive statistics on every aspect related to education, including tuition and fees. The Integrated Post -secondary Education Data System (IPEDS) is particularly helpful in collecting valid data about post-secondary colleges in the United States.

Data has been compiled on more than 4,000 universities in the United States that qualify for Title IV, with a specific focus on the differing tuition and fees for in-district, in-state, and out-of-state scenarios. Following this, the data helps to break down the cost of living, books, supplies, and other expenses, sorted by whether the student is in-district, in-state, or out-of-state. The data format includes integers and text, with each university assigned a unique Unit_Id.

IPEDS data is a precious tool for prospective students and their families, providing them to examine and compare tuition, living expenses, and financial aid prospects at other colleges and universities. This information helps them decide where to apply, which financial assistance alternatives to pursue, and how to budget for higher education expenditures. This evidence-based strategy enables policymakers to argue for reforms while encouraging cooperation and community participation to generate significant increases in higher education financing. This project aims to provide a web-based dashboard that allows users to interactively examine data on tuition expenses and median family earnings.

2.4 Data Conditioning

Data conditioning capabilities allow large businesses and cloud-based data storage facilities to substantially enhance their system's performance and efficiency, resulting in better applications' execution and lower operational and initial expenses.

1. Data Sorting: - Ensure that all monetary values, such as tuition, are expressed consistently throughout the datasets. – Arranging the all-years data starting with fees and including all other data fields necessary for analysis.
2. Variable Re-naming: - Conditioning the variables to the most affective names which are SQL readable format.
3. Column Alignment for Data Merging: - - Prior to merging multiple datasets, ensure that the column structures match and are consistent. -Utilize Excel's MATCH and INDEX functions to rearrange columns in a specific order based on a master column structure. -This approach helps align and organize data elements for seamless merging operations.

It's essential to gather and arrange tuition and income information from different years to create a comprehensive dataset. This dataset presents challenges due to its structure, having specific columns for yearly room and board expenses and median income.

Consolidation approach:
The current data format makes it difficult to analyse trends over time, as data for different years is spread across separate columns. To address this, we propose a consolidation strategy that changes the data into a

format that makes it easier to study. This is done by converting the data into a table (data frame) with two levels of indexing: the first for the institution and the second for the year.

This approach offers several advantages are:

Facilitated Time-Series Analysis: It prepares data by arranging it chronologically within each institution's history. This format enables the use of time-series analysis methods to explore temporal trends, growth rates, and cyclical patterns in the data over time.

Improved Data Integrity: By combining data, consolidation reduces the likelihood of inconsistent information. This ensures that the results for each year are correctly associated with the corresponding institution.

Streamlined Data Integration: By utilizing a structured multi-index approach, data handling becomes simpler. It allows for easier filtering, slicing, and aggregating of data, making it more efficient to extract meaningful insights.

2.5 Data Quality Assessment

Evaluating the quality of the data collected in Sprint 2 is essential, in accordance with best practices for data analysis initiatives. This is an assessment of our datasets using the given criteria:

1. Completeness: Evaluate the proportion of missing values for every variable that is pertinent to the study topic using the IPEDS data. To learn about desired data completeness, get in touch with IPEDS or review the documentation further enhances quality of the data accumulated. Analyzing the completeness of each state's median household income statistics using the FRED dataset and look for any states where data is missing and investigate possible causes (such as restrictions on data access).

2. Individuality:

IPEDS Data: Examine the dataset for entries that are duplicates. This may include figuring out unique identifiers (such university IDs) and making sure there aren't any duplicates. Further reduction of data repetition is achieved by arranging the data attributes in columns.

FRED Data: Determine whether state identities in the dataset are unique. Make sure every state is represented just once and has a unique identification.

3. Precision: Apply data validation approaches for the two datasets to evaluate the correctness of the values in each. This could consist of checking values for anomalies and discrepancies or comparing data to outside sources.

4. Atomicity: When we talk about this quality, we usually mean making sure that data is intact when we manipulate it. Atomicity is not a top priority at this point because there was no data alteration done in Sprint 2 other than merging.

5. Adherence: Both Datasets determine if the pertinent variables' data formats (such as data types and units) are the same in both datasets. For a subsequent study, data transformation or standardization may be necessary due to inconsistencies.

6. Overall Quality: The raw data achieved after merging is intact to the initial necessities of the research addressing various attributes needed to answer the problem statement. The team designed a new nomenclature for attributes for easy understanding of the data, The missing values may be omitted or nullified based on the client's requirement. The summarized project architecture designed oversees the initial hurdles of data sourcing, enabling better ways for future data sourcing.

2.6 Other Data Sources

The following organizations were given to us to supplement our existing data sources: The United States Census Bureau, The Bureau of Economic Analysis (BEA), The United States Bureau of Labor Statistics (BLS), and various academic institutions and research groups, such as The Stanford Center on Poverty and Inequality and the Institute for Research on Poverty at the University of Wisconsin–Madison. Despite their usefulness, these sources fall short in providing the detailed explanations and specific data that we need. Data about the typical income of families in a state is available from the US Census Bureau, the BEA, and the BLS. However, the data is often presented in graphs and charts, which can be hard to understand and apply. Whereas academic institutions and research groups data set includes information regarding tuition fee, however it has a lot of null values that make it difficult for us to use them.

2.7 Storage Medium/ Storage Security

Leveraging AWS for storing substantial datasets to analyse university tuition fees relative to median household income is a practical strategy, especially when considering the ongoing expansion of cloud-based data year over year. Below, you'll find a detailed guide outlining the steps you can follow to proceed with this endeavour.

- Set up an AWS Account: Register at aws.amazon.com to utilize AWS. As instructed, select the region in which your analysis will be conducted, and enter correct billing information.
- Choose a Storage Service: For this data analysis, consider using Amazon S3 (Simple Storage Service) as your storage solution. This service provides scalable, reliable, and secure object storage, which is tailored to meet the needs of storing datasets, documents, and analysis findings pertaining to university tuition fees and median household income.
- Create an S3 Bucket: Proceed to Amazon S3 service after logging into the AWS Management Console. The next step is to establish a new bucket, giving it a unique name (such as "tuition-fees-analysis") and selecting the region that best suits your needs.
- Set Bucket Permissions: Establish access control rules to safeguard your S3 data. To control who can access, modify, or remove files from the S3 bucket, use IAM policies. To grant services or apps restricted access to the data, create IAM roles if needed.
- Upload Data: Use your Amazon S3 storage to keep any relevant papers and data. This contains information for your analysis on research, fiscal statistics, family income, historical tuition trends, census numbers, and other topics. Amazon offers software development kits (SDKs) and a command-line interface (CLI) for automating the process of adding files, or you can add files manually through the Amazon Management Console.
- Access Control: To limit access to sensitive data inside the S3 bucket, define policies for access control. To restrict who can access, modify, or remove anything kept in the bucket, use IAM policies. If your applications or services require access to the data, you should think about adding IAM roles.
- Security: Use best practices for security to safeguard the information you save on S3. Use SSL/TLS encryption to safeguard data in transit and enable server-side encryption to encrypt

data at rest. Control access to your S3 bucket by implementing AWS Identity and Access Management (IAM) controls and turning on logging to keep an eye on access activity.

- Monitoring and Logging: Use AWS CloudTrail and Amazon CloudWatch to monitor and log the actions related to your S3 bucket. With CloudWatch, you can set up alerts for events and keep an eye on important bucket metrics. To further serve as an audit trail for security and regulatory reasons, CloudTrail also generates a comprehensive log of all API actions made on the bucket.
- Backup and Disaster Recovery: Put strong backup and disaster recovery procedures in place to protect your data. To save several item revisions and prevent accidental removal or manipulation, enable versioning in your S3 bucket. Cross-region replication, which distributes data across multiple AWS regions, can be used to improve resilience.
- Cost Management: Use several S3 storage classes if you want to keep prices down. Data that is not regularly needed can be moved to less expensive storage solutions like Amazon S3 Glacier or S3 Glacier Deep Archive by using lifecycle controls and data analysis to determine how frequently you use the data. You can reduce expenses without sacrificing the amount of data you can store and retrieve by routinely checking and modifying your storage configurations.
- Integrate with Analysis Tools: Link your Amazon Athena, Redshift, or Glue analytic platforms and tools to your S3 storage. This allows you to analyse information on household income and college tuition without having to move the data. Data from your S3 bucket can be directly queried, modified, and analysed.
- When assessing college tuition costs in connection to median household income, these techniques will assist you in using AWS S3 as a scalable, affordable, and secure storage medium for your data analysis and storage needs.

2.8 Storage Costs

When handling massive datasets like household income versus tuition fees, storage costs play a critical role in data management. AWS offers scalable and adaptable pricing options that companies can customise to meet their specific storage requirements. There are several storage solutions with different pricing structures that depend on data transfer, frequency of retrieval, and type of storage. Users can reduce expenses by only paying for the storage they use by utilizing AWS's pay-as-you-go concept. In addition to guaranteeing cost-effectiveness, this enables effective data management.

AWS gives customers access to tools and services that optimize storage expenses. S3 Standard, S3 Intelligent-Tiering, S3 Glacier, and S3 Glacier Deep Archive are just a few of the storage class options available; each has a distinct price, durability, and accessibility. AWS provides lifecycle policies, which use pre-established criteria to automatically migrate data to less expensive storage tiers to further cut expenses. By using cost-saving techniques and choosing the right storage classes, businesses may efficiently manage storage costs while guaranteeing the accessibility and durability of data for continuing analysis and insights.

Section 3: Algorithms & Analysis / ML Model Exploration & Selection

3.1 Solution Approach

3.1.1 Systems Architecture

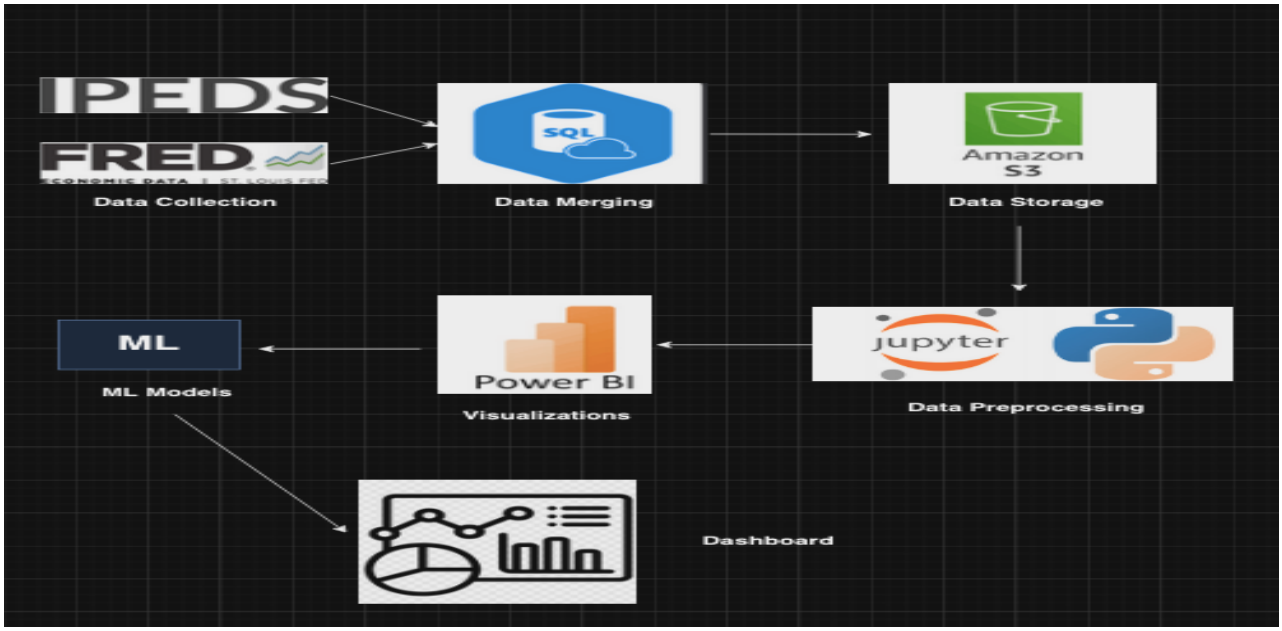


Figure 1. Systems Architecture

A thorough data analysis process is depicted in the picture, starting with data collection from the IPEDS and FRED websites and ending with data merging in MySQL, where the data is combined to create a single, cohesive dataset. The S3 bucket system, which is intended to effectively organize and maintain massive volumes of data, is then used to store the data in the AWS cloud. Data pre-processing is done in a Jupyter notebook to clean, normalize, and convert the data before analysis to make sure it is in the right format. After that, Power BI visualizations are made to find trends, patterns, and anomalies in the data, enabling it to be understood and used for practical purposes. In addition, appropriate algorithms are chosen, pre-processed data is used to train them, and parameters are adjusted to improve performance to create ML Models. The insights obtained from the ML model outputs and visualizations are then combined into a dashboard, which offers an interactive platform where users can examine data, comprehend analysis findings, and make responsible choices based on real-time and predictive analytics. Using cutting-edge machine learning and data processing methods to extract useful insights from unprocessed data, this architecture is an excellent example of a comprehensive approach to data management and analysis.

Data Architecture:

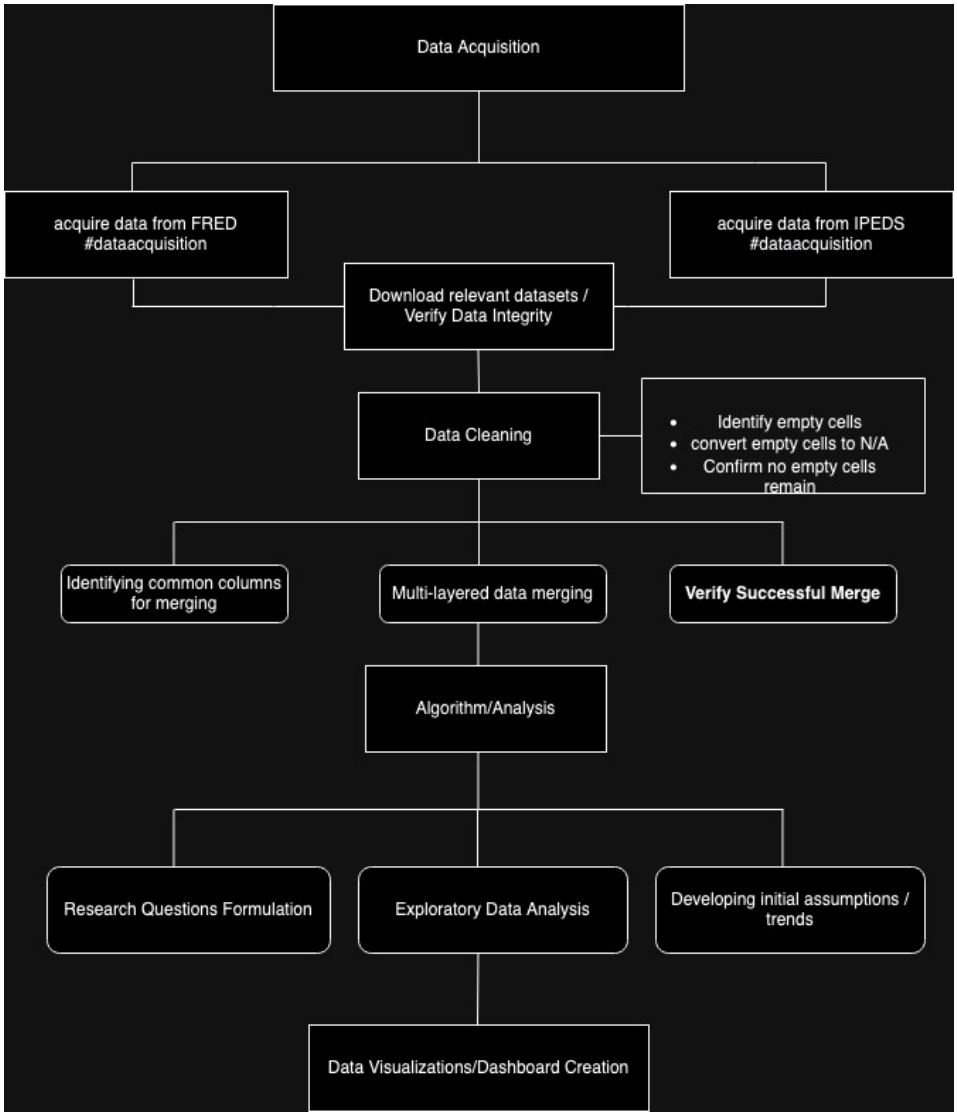


Figure 3. data Architecture

1. Data Acquisition

Acquire data from FRED:

Get pertinent data sets: Go to www.FRED, the Federal Reserve Economic Data website. Look for the datasets that pique your interest, such as GDP, inflation rates, or employment statistics. To get the data in the format of your choice (such as CSV or Excel), use the download options.

Check for data integrity. Verify the accuracy and completeness of the data after downloading. Make sure that the data looks fair, the time period covered by the data matches your expectations, and there are no corrupted files.

Acquire data from IPEDS:

Get pertinent data sets: Go to the website of the Integrated Postsecondary Education Data System (IPEDS). Data about higher education in the US are available from IPEDS. Determine the datasets—such as enrolment figures, graduation rates, or financial data—you require, then download them.
Check for data integrity. As with the FRED data, make sure the downloaded files are exact and full. Look for any irregularities or contradictions in the data that might point to problems.

2. Data Cleaning

Recognize vacant cells: Find any cells that are empty or have missing values by scanning the datasets with your favourite data analysis program. (Excel, pandas, Python, etc.).
Set all blank cells to N/A: To indicate that the data is not available, enter "N/A" in these empty cells. By doing this, you can clean up your datasets and avoid mistakes during subsequent phases of analysis.
Verify that no empty cells are left: To make sure that "N/A" has been inserted in place of any empty cells, rescan the datasets.

3.Data Merging and Integration

Using Python and SQL, combine and integrate data:
Determine which columns should be combined: In both datasets, look for columns that can be used as merger keys, like time periods or institution IDs.
Combine data sets: Tables based on shared columns can be combined using SQL's JOIN methods. Similar outcomes can be obtained in Python by using the merge function of the pandas library.
Check for a successful merger: Verify the consistency and completeness of the resultant dataset. Make that the merge process doesn't result in any lost or duplicated data.

4. Multi-layered Data Merging

Combine all the institutional data, including Inside the district: Combining data from several sources, such as institutional or geographic data, may be necessary to achieve this. The objective is to compile a comprehensive dataset that offers a thorough image of the concerned educational institutions, including details unique to each district.

5. Exploratory Data Analysis

Gaining additional insight into the factors that lead to the rising trends in tuition: Determine the important factors (inflation, government funding, institutional costs) that could affect tuition rates first. Create research questions that will help you navigate your investigation to comprehend these connections.
Forming preliminary beliefs: Construct theories regarding the ways in which various factors could impact tuition patterns based on a review of the literature and early data. Your exploratory data analysis will follow these presumptions.

6. Data Visualizations

Python: To build visual representations of your data, use libraries such as Plotly, Seaborn, or Matplotlib. Finding trends, patterns, and outliers can be aided by these illustrations. R-using A strong tool for producing intricate data visualisations in R is the ggplot2 package. Numerous graphs that aid in dataset exploration can be plotted with it.

With a solid foundation for any additional in-depth analysis or modelling work, you may perform a comprehensive data analysis process by adhering to these procedures, which cover everything from acquisition and cleaning to merging, analysis, and visualization.

Tableau / Power BI: Tableau and Power BI are two of the top business intelligence and data visualization solutions; each offers different advantages. Tableau is preferred by users requiring custom dashboards and deep data analysis, as it is great for creating dynamic and complex graphics. Its simplicity of use appeals to both technical and non-technical people. Power BI easily interfaces with other Microsoft products and is well regarded for being user-friendly and economically priced. Several tools are provided for real-time processing of data, data networking, and countless tools for data modeling. Both platforms become indispensable resources for companies that want to use their data for strategic decision-making because they have active communities and offer a lot of support provided by forums, tutorials, and documentation.

3.1.2 Systems Security

To guarantee strong security measures for the interactive dashboard, it is imperative to adopt a complete strategy.

Initially, it is imperative to encrypt all data communications with SSL/TLS protocols to protect against interception and uphold data integrity. This encryption guarantees the confidentiality of sensitive information while it is being transmitted.
It is necessary to establish authentication and authorization methods to authenticate the identities of users and regulate access according to predetermined roles and permissions. Implementing robust password restrictions and incorporating supplementary security measures like multi-factor authentication provides an additional level of safeguard against illegal entry.

The security of Amazon S3 buckets is comprised of multiple layers. ACLs and bucket policies govern the authorization of anyone to access the bucket and its contents. Data secrecy is ensured using encryption alternatives such as SSE-S3, SSE-KMS, and client-side encryption. Versioning serves as a safeguard against inadvertent removals or alterations.

Multi-factor authentication (MFA) enhances the security of sensitive processes by introducing an additional layer of protection. The management of user permissions and responsibilities is facilitated by AWS Identity and Access Management (IAM).

AWS CloudTrail and Amazon Macie are monitoring solutions that aid in identifying and addressing instances of unauthorized access or data breaches. Routine audits and compliance assessments are implemented to ensure strict adherence to established security protocols. SSL encryption safeguards data transmission from interception.

In addition, S3 has the capability to provide precise access control by utilizing bucket policies and Access Control Lists (ACLs). Amazon Web Services (AWS) provides various technologies, such as Amazon S3 Block Public Access and VPC endpoint restrictions, which serve the purpose of mitigating unintentional exposure to the public internet.

The implementation of a secure cloud architecture, incorporating powerful access restrictions, encryption, and auditing tools, establishes a strong basis for ensuring the security of the dashboard. Regular security audits and penetration testing play a crucial role in the proactive identification and mitigation of vulnerabilities.

The utilization of monitoring and logging technologies facilitates the systematic observation of user actions and the identification of potentially illicit conduct. This facilitates prompt reaction to security

incidents and quick countermeasures against possible threats.

Ensuring data availability and integrity in the case of a security breach or system failure is achieved through the implementation of regular data backups and crisis recovery strategies. The provision of security awareness training to employees is to educate staff members about potential dangers and optimal strategies for upholding security measures.

Comprehensive security assessments of third-party suppliers are conducted to ensure their compliance with established security protocols and standards, thereby mitigating potential risks associated with external dependencies.

By implementing a range of robust security protocols, the interactive dashboard can effectively uphold the principles of data confidentiality, integrity, and availability. This, in turn, fosters a sense of assurance among users regarding the safeguarding of their personal information.

3.1.3 Systems Data Flow

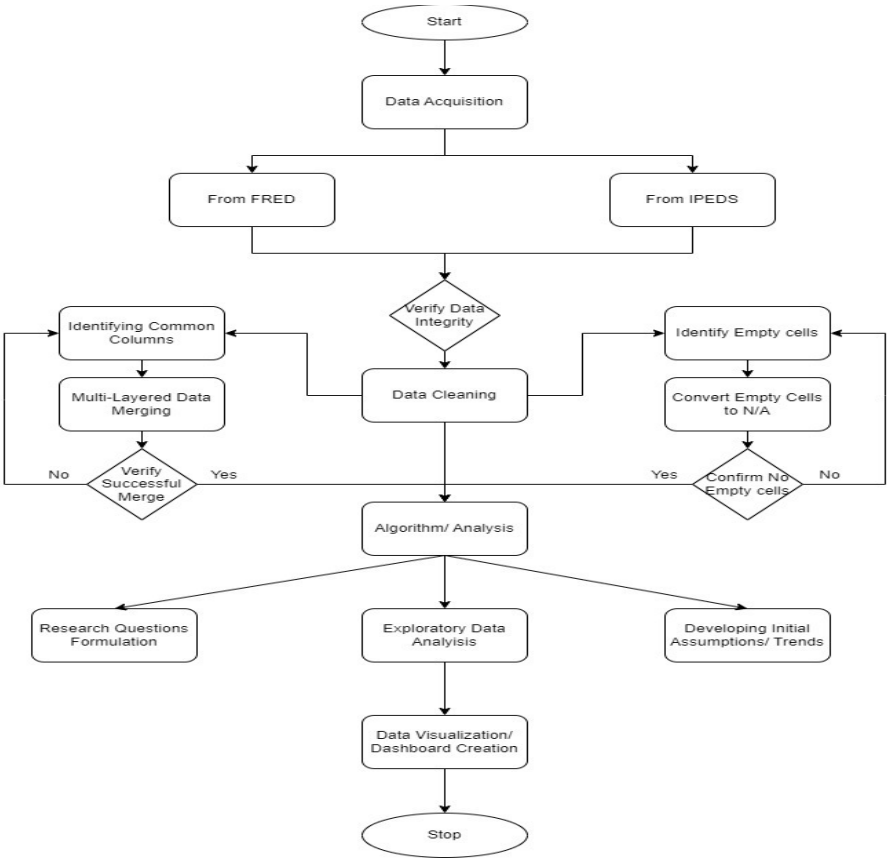


Figure 4. Systems Data Flow

Obtaining Data:

First, information is gathered from two reliable sources: FRED and IPEDS. The term FRED most likely refers to the Federal Reserve Economic Data, which provides time series data on the economy and finance. The name "IPEDS" implies the data source for higher education institutions: Integrated Postsecondary Education Data System. The pertinent datasets are downloaded after being obtained. Next, data integrity is confirmed, which might entail looking for mistakes or discrepancies in the data.

Data Cleaning:

The data cleaning step includes ensuring that the data is in an analysis-ready state.

- First, empty cells are located.
- After that, these empty cells are changed to a predetermined value, usually "N/A."
- To make sure there are no more empty cells, the data is examined once again.

Data Combination:

After that, common columns between the two datasets are found, making it possible to combine them into a single dataset. The next step involves a multi-layered data merging procedure, the specifics of which are not depicted in the picture. Lastly, confirmation is given that the data fusion was successful.

Method/Analysis:

A set of questions can be developed when the data has been combined and cleansed. The next step is exploratory data analysis, which entails compiling the data and spotting preliminary trends and patterns. More thorough analyses will probably come after this, albeit the precise algorithms employed will depend on the study topics being looked at. The analysis's findings are subsequently shown using dashboards for data visualization.

All things considered; this data flow diagram shows how to manage research project data in an organized manner. It provides a detailed flowchart of the processes involved in gathering data, analysing it, and visualizing it.

3.1.4 Algorithms & Analysis

The information under examination includes tuition fees collected across many states over several years, including in-state, out-of-state, and in-district tuition expenses. A first study indicates the following major features and possible challenges:

- Temporal Nature: The data is time series, with annual observations serving as the basis for trend analysis, seasonality detection, and forecasting.
- Multiple Variables: The inclusion of in-state, out-of-state, and in-district tuition fees provides a multidimensional picture of tuition expenses, allowing for comparative and correlational analysis across these categories.
- Completeness and Consistency: Preliminary tests reveal some missing data, notably in the median income column, which may demand imputation or exclusion procedures depending on their scope and analytic aims.

We use machine learning models and statistical forecasting approaches to estimate future tuition prices while analysing tuition fee data, which included in-state, out-of-state, and in-district fees. Because of the dataset's temporal character, we focused mostly on time series forecasting approaches.
Algorithm Selection

We began by establishing a baseline using Linear Regression, which is simple and easy to read. Despite its simplicity, Linear Regression's predictive power was restricted by the assumption of a linear connection between time and tuition fees.

3.2 Machine Learning

A machine learning model is a program that can identify patterns or make judgments based on previously unknown data. A machine learning algorithm is a mathematical tool are often drawn from the statistics, calculus, and linear algebra. Most machine learning approaches may be classed as supervised, unsupervised, or reinforcement learning.

3.2.1 Model Exploration

The major goal is to use this dataset to create machine learning models that can provide insights and projections on financial indicators for educational institutions. This might involve forecasting future tuition expenses, assessing affordability, and determining the influence of the surrounding area on costs.

In the model exploration phase for our dataset, we will emphasize the temporal aspect of financial measures across institutions. Specifically, we will use fields like “Tuition_fees “, “room_and_board_without_family_[Year]" and "Books_and_supplies_[Year]" as our major variables for forecasting models.

ARIMA and SARIMA models will be used to forecast future expenditures, with historical data used to identify underlying patterns and seasonality in expenses such as room and board and book supply. These models are useful because they produce accurate short-term projections, which are critical for financial planning and budgeting in educational institutions. For more complicated patterns, particularly those with long-term dependencies (e.g., the influence of policy changes or economic shifts on tuition costs over time), LSTM models will be investigated.

This investigation will include breaking down the information into distinct time series for each organization, allowing for more detailed, institution-specific projections that may guide specialized financial management and policy implementation tactics.

3.2.2 Model Selection

Model Selection and Application (Preliminary Analysis)

ARIMA/SARIMA (Autoregressive Integrated Moving Average) / (Seasonal ARIMA) Models

- Applicability: Time series forecasting for tuition fees. To anticipate future expenses, use ARIMA for non-seasonal trends and SARIMA for seasonal data, allowing for more exact budgeting and planning.
- Benefit: These models can capture linear correlations and seasonality in financial patterns, resulting in accurate short-term projections needed for yearly budgeting and financial assistance planning.

XGBOOST or LIGHTGBM

- Applicability: If we intend to use other characteristics (e.g., economic indicators, states or specific geographic data) in conjunction with past tuition costs to forecast future expenses, gradient boosting models such as XGBoost or LightGBM may be effective. These models can handle both numerical and categorical data, as well as complicated feature relationships.
- Benefit: They are very versatile, capable of capturing complicated nonlinear interactions, and can manage missing data by default, making them suitable for forecasting with extra variables.

Multiple Linear Regression

- Applicability: It is appropriate for linear connections between the dependent variable (tuition expenses) and several independent variables (year, room and board costs, book and supply costs, etc.)

- Benefit: Provides a clear knowledge of how each component (year, room and board, etc.) affects tuition expenses, allowing for easy interpretation and implementation.

Classification of Universities based on Cost and Affordability

- Models: K-Means Clustering or Hierarchical Clustering.
- Purpose: Divide universities into categories based on cost, enabling students to choose colleges that fit their budget.

Logistic Regression or Decision Trees

- Applicability: Effect of Costs on Enrolment Rates.
- Benefit: Understand how changes in tuition, housing and board, and other expenses affect enrolment rates, allowing schools to establish competitive and sustainable prices.

Consider the following factors while deciding if these models are appropriate for your dataset.

Data Structure and Quality: Confirm that the time series data is full and consistent across time points. Missing values or major abnormalities may need pre-processing.

Analysis Goals: Define what we want to predict or evaluate. What will be our outcome requirement by this analysis.

Seasonality and Dependencies: Determine if the data has seasonal patterns or long-term dependencies that may impact the decision between SARIMA and LSTM models.

Section 4: Visualizations / ML Model Training, Evaluation, & Validation

4.1 Overview

Using tuition prices and median income levels from 2002 to 2022 as its main points of focus, this thorough analysis offers an interesting look at higher education institutions across the country. By employing diverse data visualization techniques such as line graphs, bar plots, and histograms, the study proficiently depicts the financial metrics and distribution of colleges. The ability to customize views according to state, degree level, and sector through an interactive dashboard that provides both broad and in-depth perspectives improves user engagement. Some of the most important conclusions are that private nonprofit organizations predominate, out-of-state tuition has increased disproportionately when compared to median income, and geographical differences in tuition offer useful data for scholars, politicians, and students.

This forecast section describes the development and assessment of two alternative machine-learning techniques designed to anticipate tuition costs based on historical data. In the beginning, a Linear Regression framework was created via pandas for data management and sklearn for analysis. The regression model was validated on a dataset of tuition prices across time, with 'Year' serving as the variable that is independent with 'Out_of_State_Tuition' as the variable that is dependent. The model performed well on a test set, with an R^2 value of 0.88 showing good explanatory power. However, errors such as RMSE and MAE indicated probable flaws caused by anomalies or breaches of the linear model's assumptions.

Subsequently, an ARIMA framework was used to forecast prospective expenditures on tuition with training determined by the discovery of optimum characteristics using ACF and PACF plots. The sequence

was evaluated for linearity using the Augmented Dickey-Fuller test to ensure that the model hypotheses were valid. The above framework was used to project tuition for the decade to come, with forecasts including average tuition prices and confidence intervals. The model's performance varied; in-state tuition forecasting were relatively accurate with a lower MAPE, whereas tuition from other states forecasts had high errors, suggesting the need for model reassessment, the incorporation of additional variables, or the exploration of more complex models such as SARIMA to improve prediction accuracy.

4.2 Visualizations

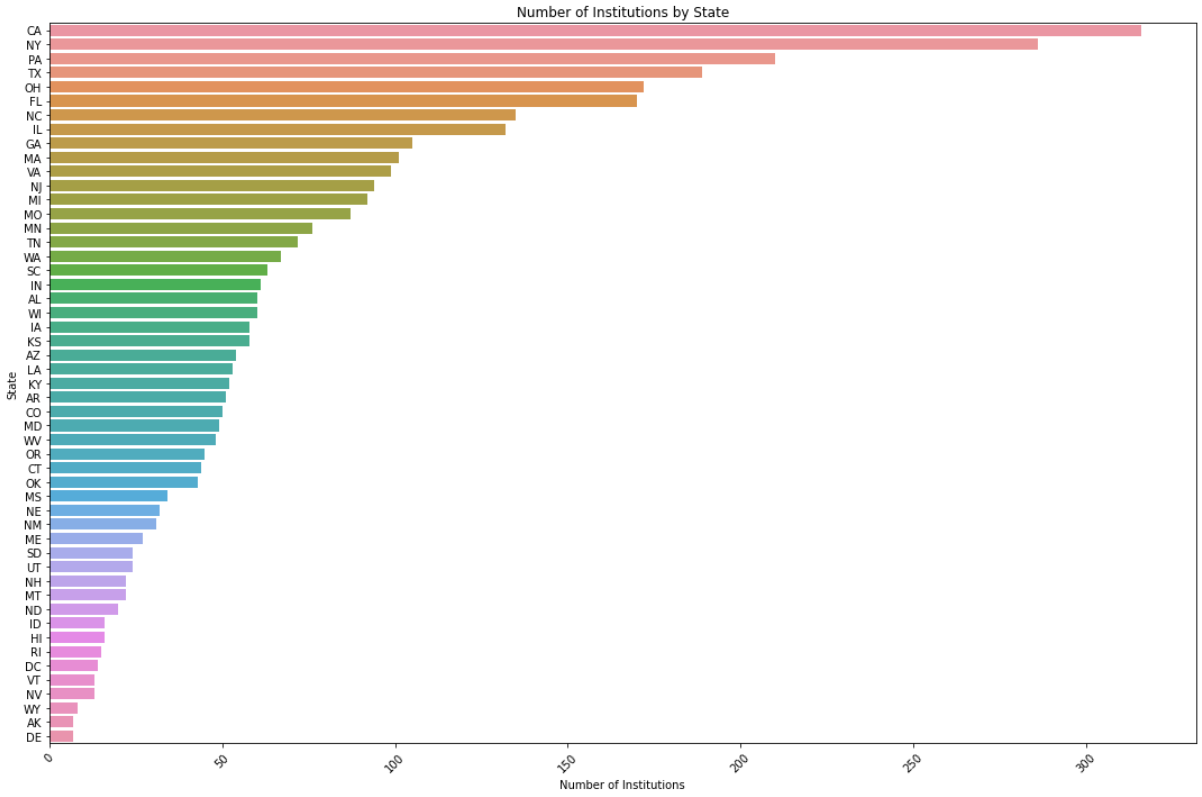


Figure 5. Number of institutes by state.

The histogram above illustrates the total number of institutions in each state, with California having the most institutions and Delaware having the fewest. We chose a histogram for this representation because it works well with large data sets.

No of institutes Private vs Public:

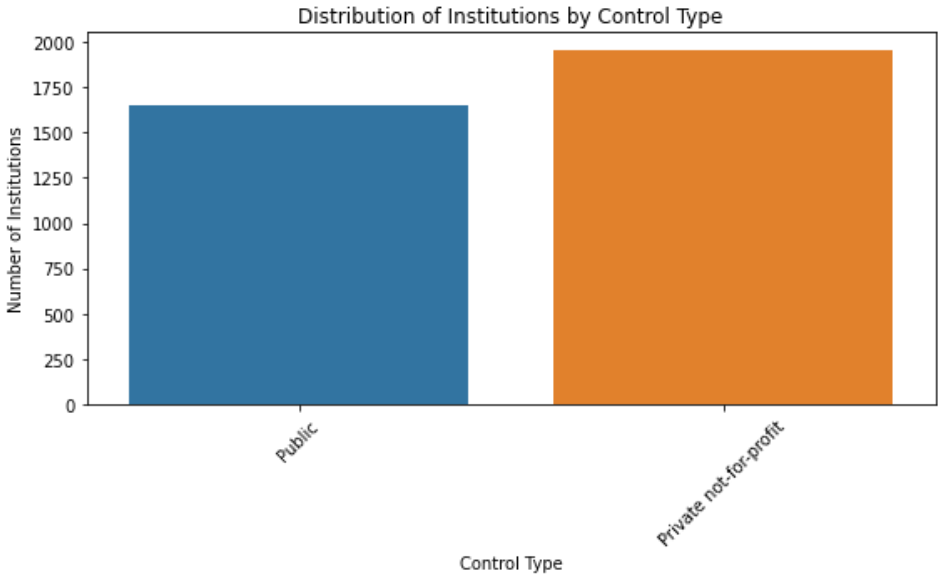


Figure 6. No of institutes Private vs Public.

The above bar plot depicts the total number of public and private nonprofit entities. When comparing the two, the United States has more private nonprofit entities. A bar plot allows us to quickly compare data sets from different groups.

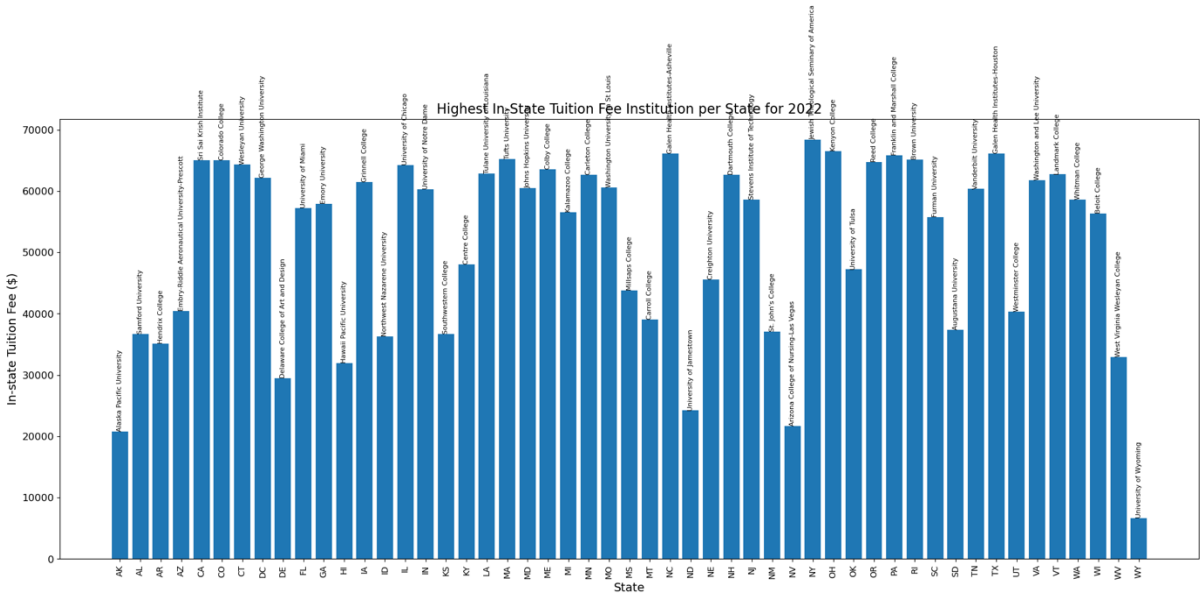


Figure 7. University with top tuition fee with each state.

The above plot depicts the universities with the highest tuition fees in each state for the year 2022. New York has the highest tuition fee, but California and Colorado will have the same tuition fee for their institute in 2022.

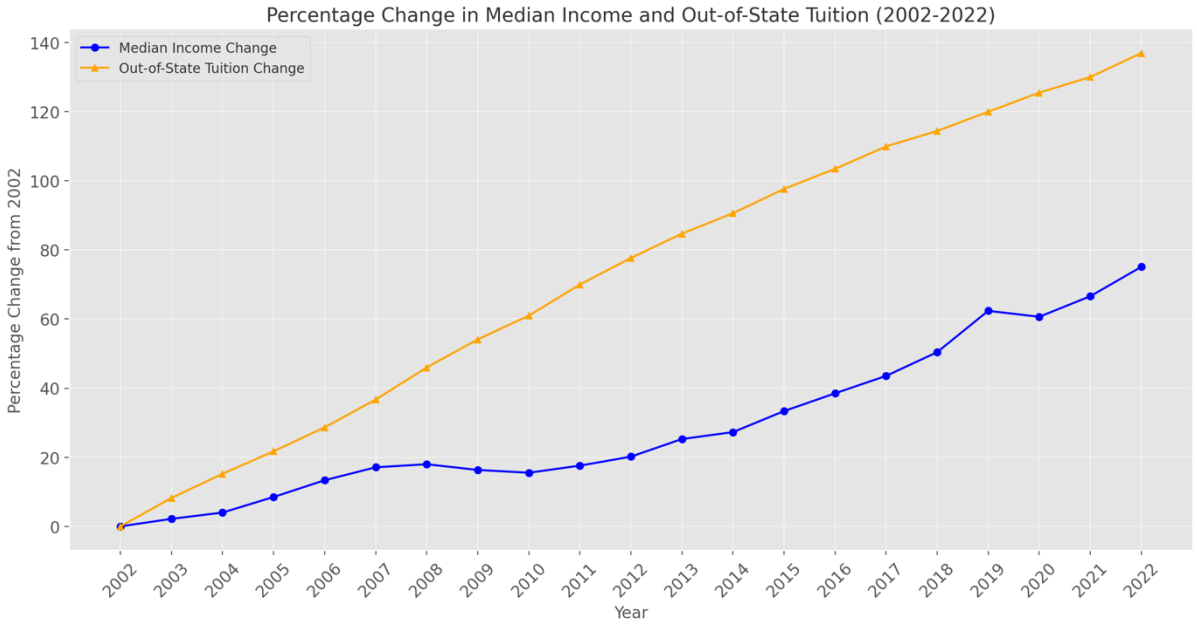


Figure 8. Median income vs tuition

The visualization shows the percentage increases in median income and out-of-state tuition costs from 2002 to 2022, with each year's change computed to 2002.

From the graph:

- Median Income Change (%): The blue line shows that median income has steadily increased since 2002, with insignificant fluctuation. There are instances where the growth rate slows, but there is a general increasing tendency.
- Out-of-State Tuition Change (%): The orange line shows a more consistent and steeper upward trend in out-of-state tuition fees over the same period. The rate of increase appears to be higher for out-of-state tuition compared to median income.

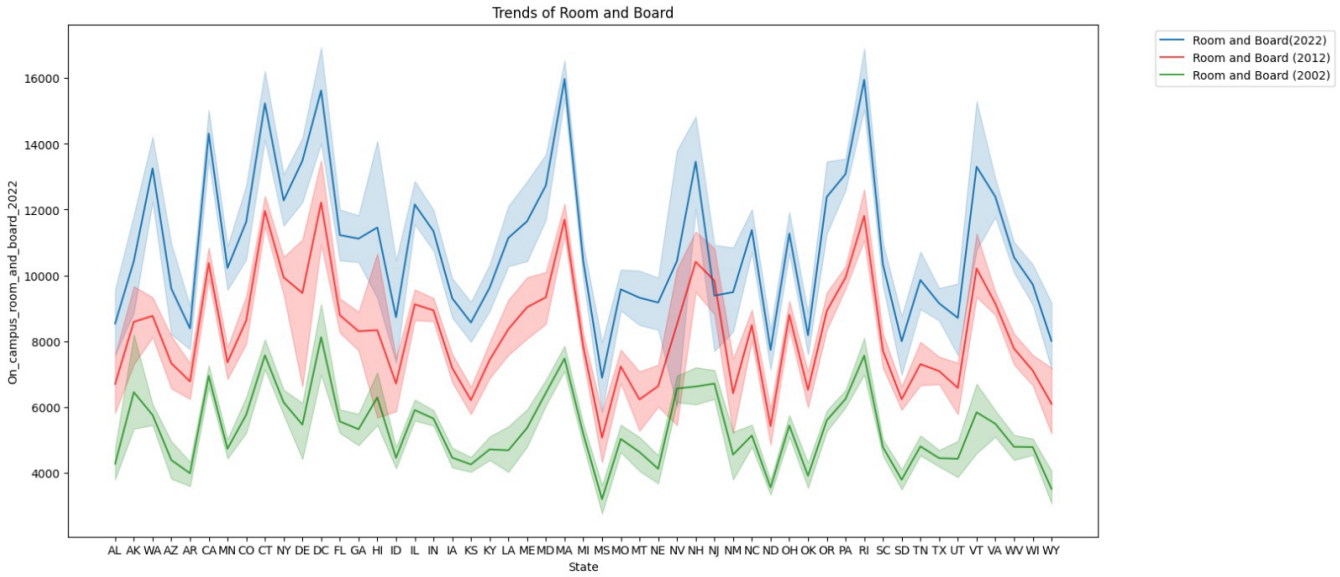


Figure 9. Median income vs tuition

The developments in room and board expenses in different states of the United States from 2002 to 2022 are depicted in this graph. Annual U.S. dollars are used to represent the costs. It is evident that room and

board expenses have increased substantially over the past two decades, with notable surges occurring in the same states between 2002 and 2022. This indicates that regional disparities in costs continue to persist. The variability and growth of costs are represented by the shaded regions between the lines for each year. States with steeper increases, such as New York and California, demonstrate broader gaps, suggesting that costs in these regions have escalated more significantly in comparison to others.

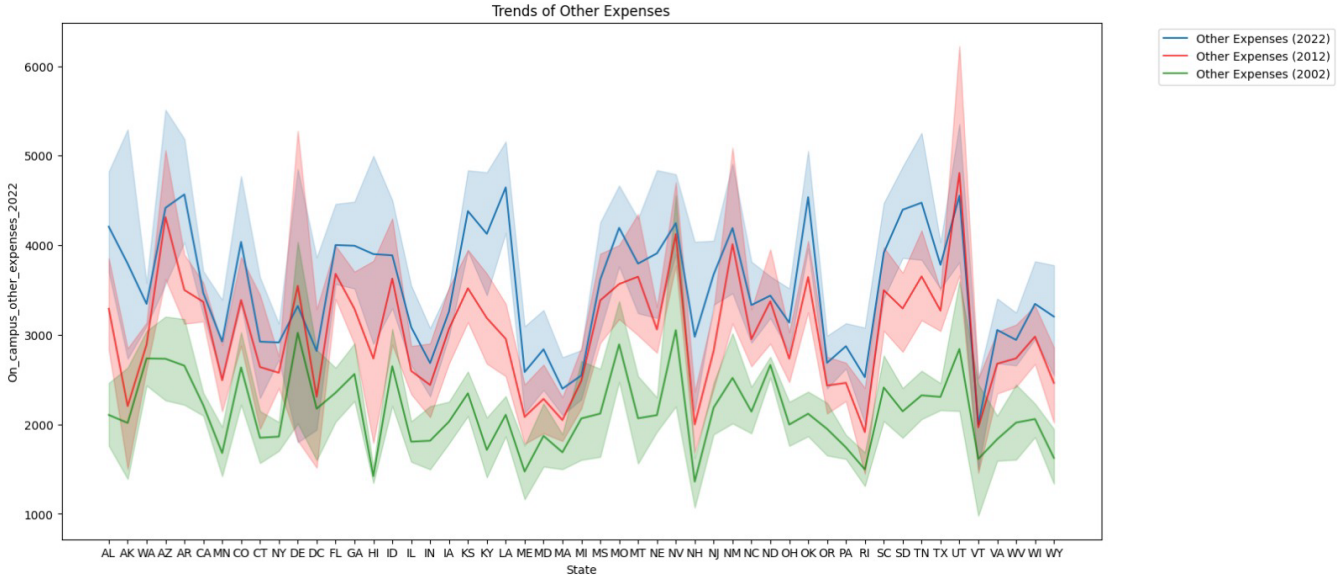


Figure 10. Median income vs tuition

The US state-specific on-campus office costs for 2002, 2012, and 2022 are shown in the "Trends of Other Expenses" graph. The expenditures are shown on the y-axis and the representation of each state on the x-axis. The costs are shown as three color-coded lines, green for 2002, red for 2012, and blue for 2022; each line is darkened to show fluctuation or uncertainty. Though the rates of increase and the precise amounts differ throughout each state, the graph clearly shows that most have seen a general increase in spending throughout time. Moreover, the obvious variations among the states within each year imply that geographic elements have a major role in the expenses. Sometimes the very angular trend lines show abrupt changes between subsequent stages. This could be the result of the fact that each state has its own budget, set of administrative rules, or set of regional economic circumstances.

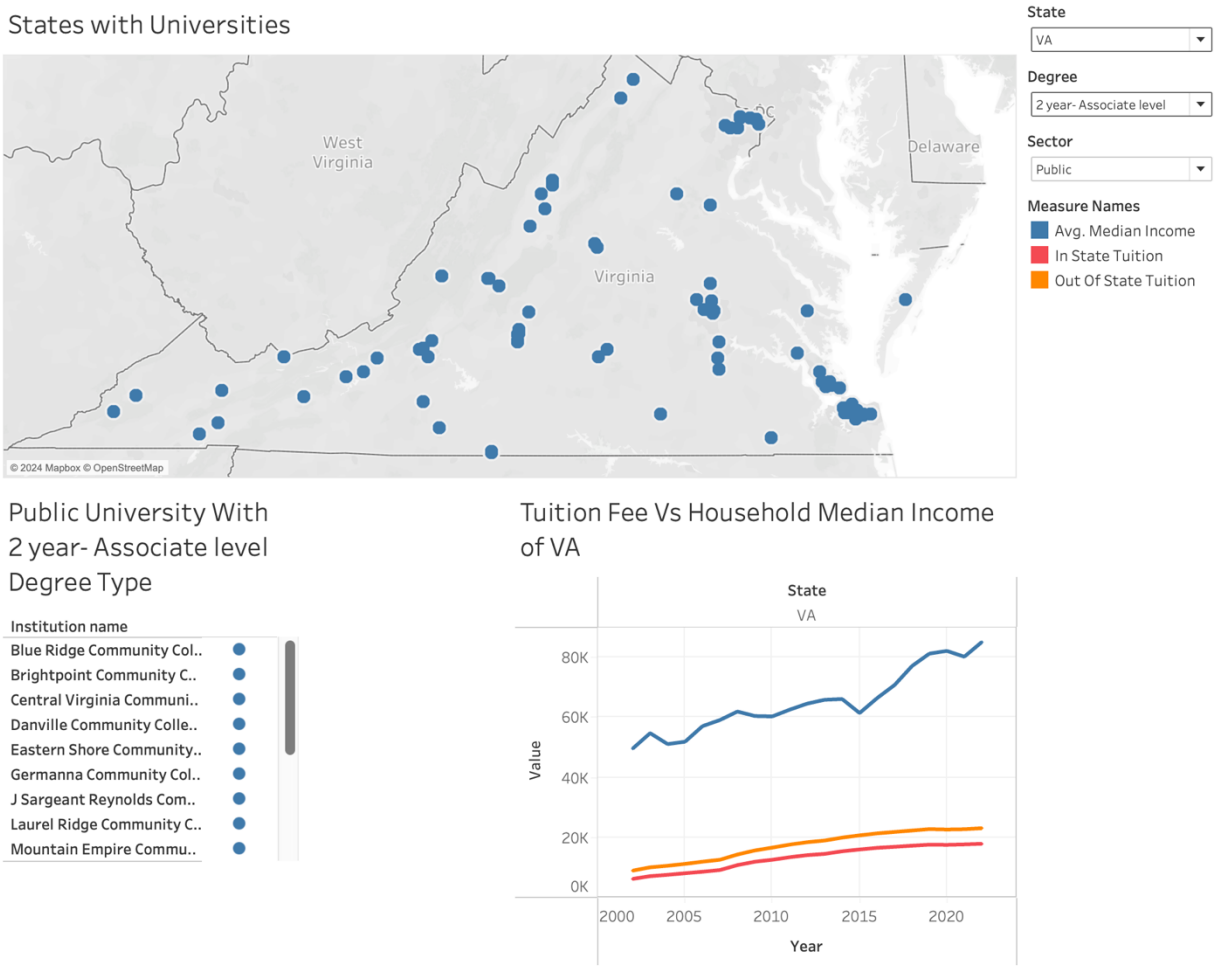


Figure 11. Tableau Dashboard 1

This interactive dashboard is designed to offer insights into higher education institutions across the United States with an emphasis on tuition costs and median income levels. The dashboard features two main filters: 'Degree Level' with options for 'Bachelor Level' and 'Associate Level', and 'Sector' with choices among 'Public', 'Private', and 'Proprietary' institutions.

Upon selecting a state, Degree Level, and Sector, the dashboard dynamically updates to list only those universities that meet the specified criteria. Alongside this, a line graph presents a comparative analysis of the average median income, average out-of-state tuition, and average in-state tuition fees for the chosen state. Further interaction is possible: clicking on any university in the list generates a specific line graph reflecting the financial metrics pertinent to that university.

This tool is particularly useful for prospective students, educational researchers, and policymakers looking to understand the financial landscape of higher education in various states, offering both a broad statewide perspective and a detailed institutional analysis.

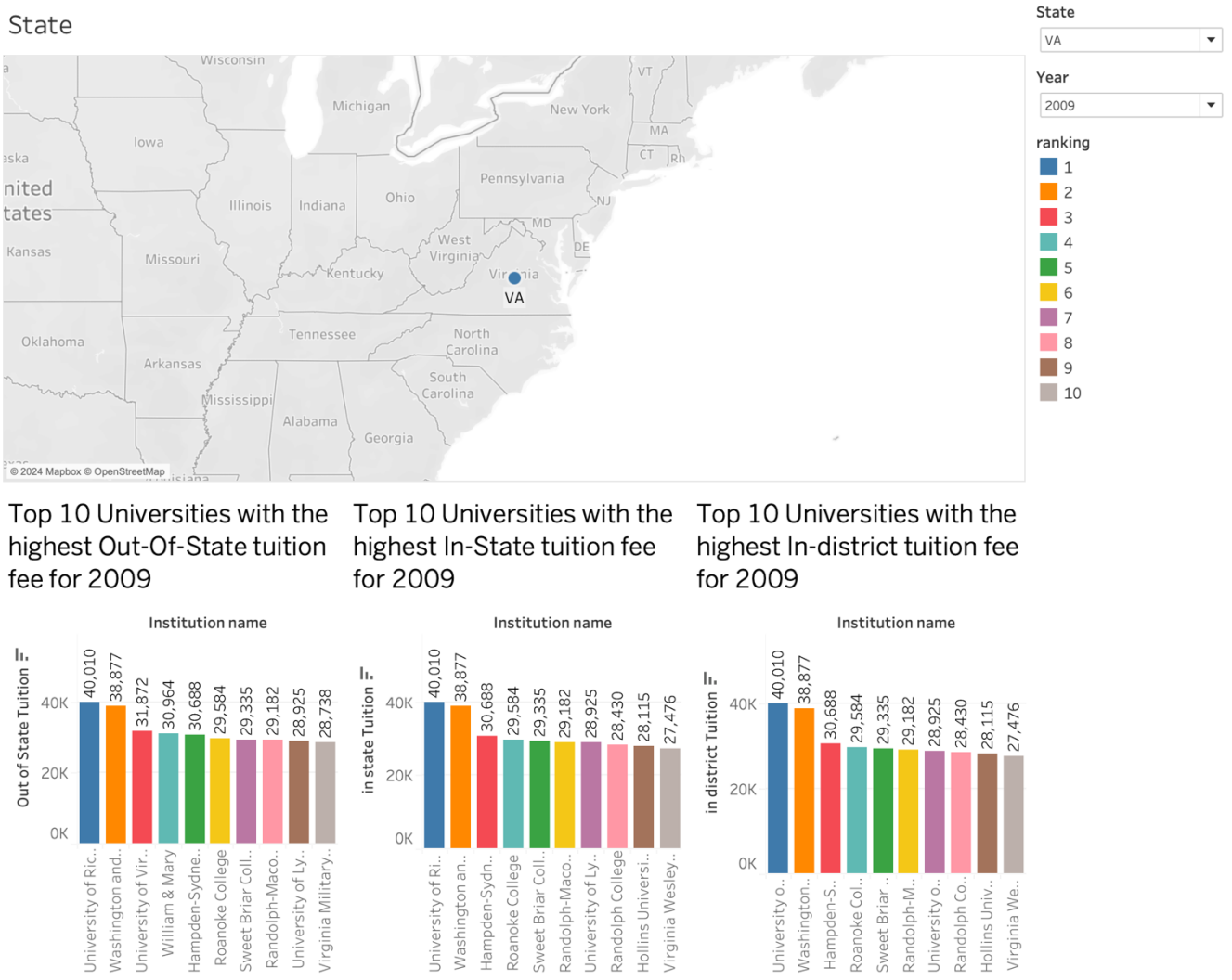


Figure 12. Tableau Dashboard 2

The dashboard provides a comparison of the top 10 universities within a selected state, categorized by various tuition fee levels. The selection can be refined by state and year to offer a customized view. The color-coded ranking system enhances the user experience by visually representing each university's standing in terms of tuition fees, allowing for quick and easy comparison. This intuitive visualization helps users identify not only the most expensive institutions but also how they rank relative to each other within different tuition fee categories, including Out-Of-State, In-State, and In-district fees for the selected year.

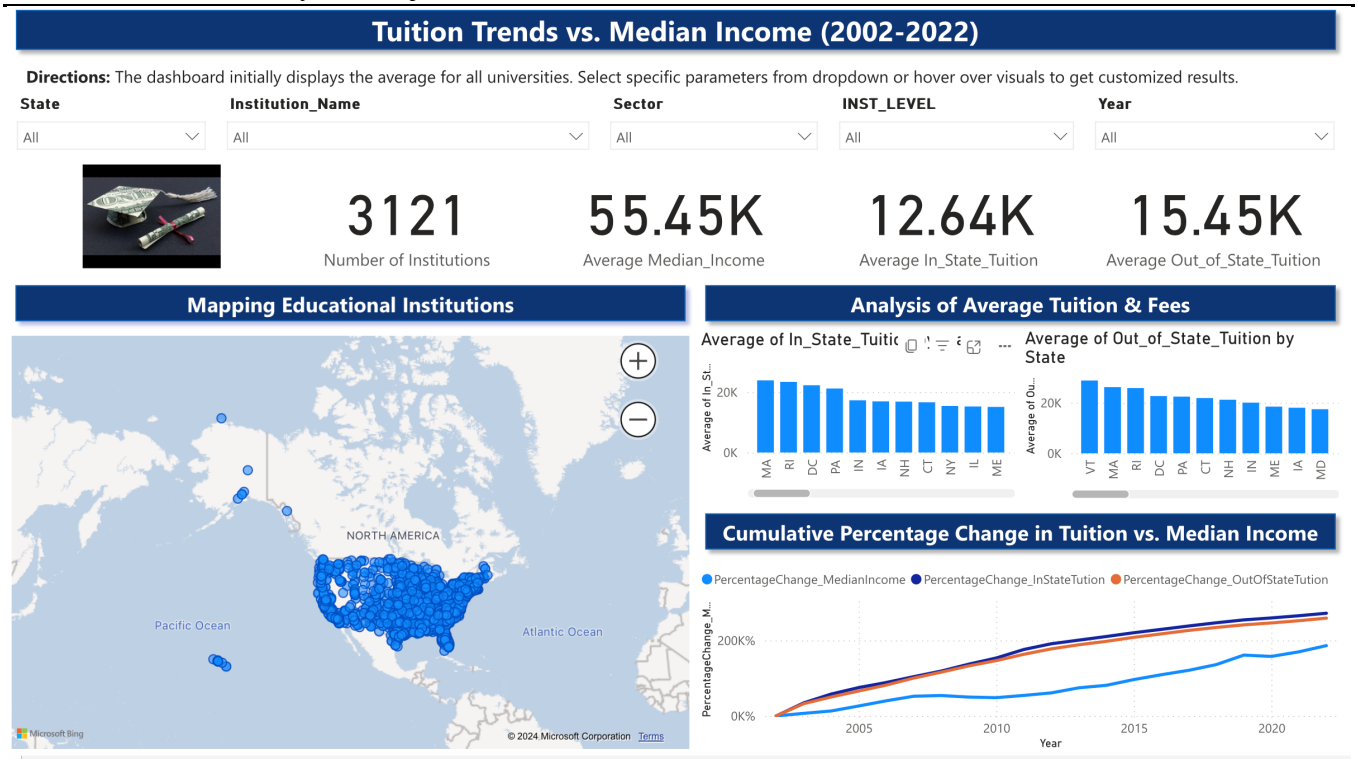


Figure 13. Power BI Dashboard 2

The interactive dashboard offers a detailed exploration of the interplay between tuition costs and median income in the U.S. over a 20-year period. It includes a geographical map marking the dispersion of 3,121 educational institutions, providing a visual sense of scale and distribution. Vital statistics such as average median income ($55.45K), average in-state tuition ($12.64K), and out-of-state tuition ($15.45K) are prominently displayed. Insightful bar charts present state-by-state comparisons of both in-state and out-of-state tuition averages, along with a focused view of the states with the most significant tuition increases. The central feature is a line graph that eloquently illustrates the percentage changes in tuition fees against the backdrop of median income shifts, highlighting the economic dynamics affecting education affordability. A detailed examination of educational economics by state, institution name, institution level, sector and academic year is made possible by the customisable parameters that enable users to go deeply into the data.

4.3 Machine Learning

4.3.1 Model Training

Data Preparation

The model training started out with the guidance of historical tuition price data, loaded from an Excel dataset (transformed_out_state_tuition.Xlsx). The 'Year' served as the impartial variable, even as 'Out_of_State_Tuition' became the based variable focused for prediction. The instruction became facilitated by the pandas library (pd.read_excel feature).

Model Training

A Linear Regression model from the sklearn.linear_model library was applied due to its effectiveness in capturing linear relationships. The version become trained on a delegated education set comprised of the 'Year' and corresponding tuition costs.

python
model = LinearRegression()
model.fit(X_train, y_train)

Model Evaluation:

Testing Set:

The version's overall performance was assessed on a distinct testing set that it had not encountered all through training. This technique guarantees an objective evaluation of the forecasting abilities of the model.

Performance Metrics:

The model's predictions on the testing data were carefully evaluated with the usage of numerous key performance metrics:

- Mean Absolute Error (MAE): 467.02
- Mean Squared Error (MSE): 414,360.62
- Root Mean Squared Error (RMSE): 643.71
- R-squared (R^2): 0.88

A vast portion of the variance in tuition prices is explained by means of the model, which seems to have a very sturdy suit in terms of R^2. Nonetheless, the errors shown via the RMSE and MAE advise that even though the version generally operates well, there are instances whilst it makes sizable mistakes. This is maximum likely the result of outliers, odd outcomes, or specific conditions where the assumptions of the linear version are broken.

Visual Evaluation:

The actual as opposed to predicted tuition expenses have been visualized through scatter and line plots with the use of matplotlib.pyplot. The visualization absolutely represented the model's predictive overall performance, where the proximity of the predicted values (red crosses) to the actual values (green dots) on the testing set indicated a terrific model fit.

4.3.2 Model Validation

Cross validation

Cross-validated RMSE scores: [546.83134496 191.2557568 208.37057913 469.48942524 177.609182]
Mean RMSE: 318.71125762514623

‘The mean RMSE, calculated as the average of the five folds, was 318.71. This value represents the model's typical error magnitude when predicting out-of-state tuition fees.
The spread of RMSE rankings throughout the folds suggests some variability in model's overall performance, which is not unusual in real-world situations due to statistics heterogeneity. The mean RMSE affords a consolidated view of model overall performance, suggesting that, on average, the model's predictions are within approximately $318.71 of the actual tuition costs.

Model Diagnostics and Improvement

For additional validation, the model's assumptions and overall performance, a residual plot will be employed, which would contain examining the differences between real and anticipated values for patterns that would advocate enhancements.

Conclusion and Recommendations

With an R^2 of 0.88, the linear regression model indicated a great degree of variance explained. The RMSE, however, counselled that extra development could be made to the predictions. Possible movements to improve the model consist of:

• Looking into possible characteristic engineering opportunities to seize more complex relationships.
• Taking into account extra sophisticated models consisting of polynomial regression or machine learning strategies capable of nonlinear modelling.
• Employing time-series forecasting models like ARIMA if tendencies, seasonality, or cycles extensively affect tuition prices.

ARIMA (Autoregressive Integrated Moving Average)

Model Training

The Autoregressive Integrated Moving Average (ARIMA) model is a frequently utilized statistical technique for estimating time series data. Employing past information from 2001 to 2022, we trained an ARIMA model to foresee future tuition prices. The method of training included figuring out the optimal sequence of differencing (d), the number of autoregressive terms (p), and the number of moving average terms (q). These characteristics were found using the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots to find substantial delays and stationarity in the time series.

Check for stationarity:
The underlying data must be steady in order for ARIMA models to function correctly. Stationarity suggests the statistical properties, including variance, average, and autocorrelation, are consistent throughout time. To check for stationarity, we adapted the Augmented Dickey-Fuller (ADF) test to our tuition cost data. The ADF test findings suggested that the time series was not stationary, thus we differentiated the data to achieve stationarity. This phase is crucial because non-stationary data might provide false findings, reducing the trustworthiness of the model's projections.

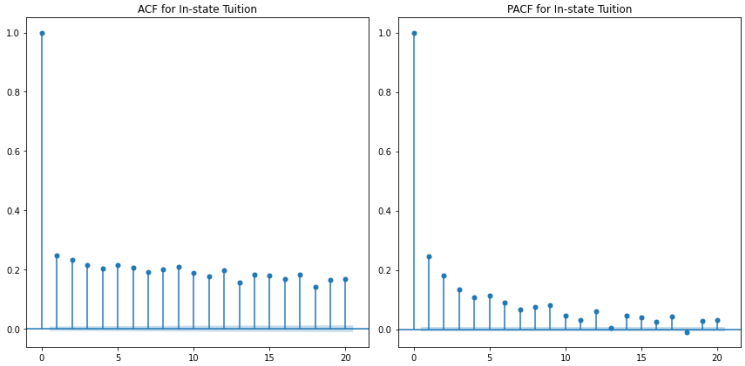


Figure 14. ACF, PACF for In_state_Tuition

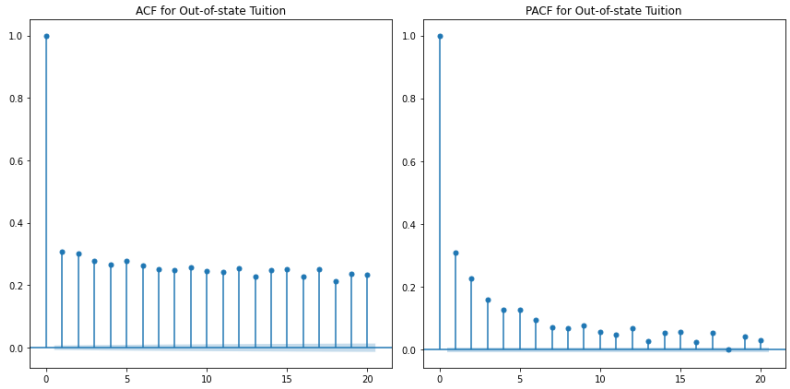


Figure 15. ACF, PACF for Out_of_state_Tuition

ADF Test (stationary analysis)

For out-of-state tuition, the ADF Statistic of -12.628 surpasses (in a negative direction) the critical values for all generally used significance thresholds (1%, 5%, and 10%), and the very tiny p-value (basically 0) shows that we may reject the null hypothesis of a unit root. Similar findings were observed for in-state tuition, with an ADF Statistic of -13.831 and a correspondingly modest p-value. These findings give strong evidence that both tuition cost time series are stable, which is a necessary assumption for ARIMA modeling.

Model Specification

```
Out-of-State Tuition:
ADF Test
ADF Statistic: -12.628385888727811
p-value: 1.523790862665388e-23
Critical Value (1%): -3.430436647177462
Critical Value (5%): -2.8615782965577967
Critical Value (10%): -2.5667903839415236

In-State Tuition:
ADF Test
ADF Statistic: -13.831493199182189
p-value: 7.577810800131181e-26
Critical Value (1%): -3.430436647177462
Critical Value (5%): -2.8615782965577967
Critical Value (10%): -2.5667903839415236
```

```
Out-of-State Tuition after differencing:
ADF Test
ADF Statistic: -50.984694258231784
p-value: 0.0
Critical Value (1%): -3.430436987194358
Critical Value (5%): -2.8615784468371968
Critical Value (10%): -2.566790463930467

In-State Tuition after differencing:
ADF Test
ADF Statistic: -53.833457357985395
p-value: 0.0
Critical Value (1%): -3.430436987194358
Critical Value (5%): -2.8615784468371968
Critical Value (10%): -2.566790463930467
```

Figure 16. ADF Test for in sate an out of state tuition

After the parameters were identified, the ARIMA model was designed for both in-state and out-of-state tuition costs. This specification comprises merging the model structure with the parameters we identified:

- AR terms (p=2): The model would use the tuition fees from the previous two years to forecast the future year's charge, allowing for any trends or cycles in tuition adjustments over time.
- Differencing order (d=0): No differencing was used since the data were already stationary, suggesting that the series' mean and variance remained constant across time.
- MA terms (q=1): A single moving average term was used to represent the random shocks or blips in tuition charge data from year to year.

Model Fitting

The models were fitted to the data based on past tuition charge levels. This method included determining the parameters that best reflected the patterns seen in the historical data and creating the models to generate projections. The model fitting step is critical because a well-fitted model generates more accurate predictions. The strength of this fit has a direct influence on the model's capacity to anticipate future tuition fees, which helps institutions and policymakers with budgeting and financial strategy.

Model Evaluation

Forecasts: The models were used to anticipate tuition fees for the following 10 years (2023-2032). The forecast contained the estimated tuition fee mean as well as the standard errors, which provided information about the projections' expected accuracy and confidence ranges.

	Year	In-State Tuition Forecast	Out-of-State Tuition Forecast
0	2023	22582.633856	24374.881401
1	2024	23117.784853	24780.282832
2	2025	23067.137782	24745.878240
3	2026	22986.089494	24688.572769
4	2027	22904.003414	24630.366949
5	2028	22822.428046	24572.420151
6	2029	22741.441157	24514.791857
7	2030	22661.042777	24457.483514
8	2031	22581.228856	24400.493513
9	2032	22501.995158	24343.820096

Figure 17. Forecast Output

Visual Analysis: A graph depicted predicted tuition fees for the forecast period. This graphic depiction aided in comprehending the trend and unpredictability in the anticipated fees. Here is the graph:

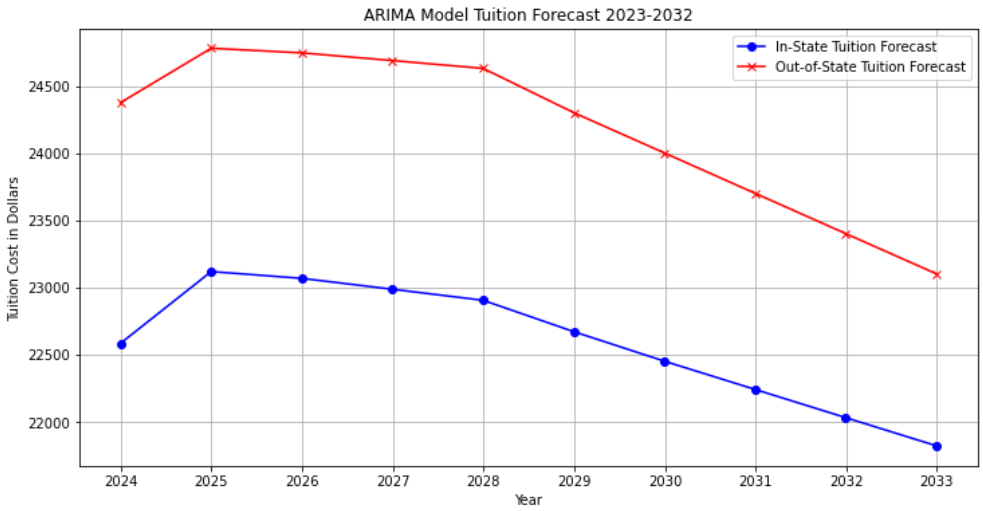


Figure 18. ARIMA Model Tuition Forecast.

The blue line depicts the in-state tuition estimate, while the orange line reflects the out-of-state tuition forecast. The dip in the conclusion reflects a decrease in the anticipated values based on the ARIMA model's output. Remember the uncertainty in these forecasts, as evidenced by the large mean standard errors and broad confidence ranges from your ARIMA output.

Model Validation

```
In-State Tuition Forecast Accuracy:
MAE: 1449.4369496847041
MSE: 2770113.458615241
RMSE: 1664.3657826977942
MAPE: 6.354766404485265%

Out-of-State Tuition Forecast Accuracy:
MAE: 7699.1008677026275
MSE: 61608909.38673949
RMSE: 7849.1343081093655
MAPE: 31.385124242024144%
```

Figure 19. in-state, out-of-state tuition Forecast.

The accuracy measures for ARIMA model projections show a significant difference between in state and out-of-state tuition estimates. For in-state tuition, the model performs rather well, with a Mean Absolute Error (MAE) of around $1,449 and a Mean Absolute Percentage Error (MAPE) of 6.35%, meaning that the model's predictions are often within 6.35% of actual tuition prices. This degree of accuracy reflects a good match, indicating that the model well represents the key dynamics of in-state tuition fluctuations. However, the Root Mean Squared Error (RMSE) of $1,664, although higher, indicates decent predictability.

Out-of-state tuition estimates, on the other hand, show substantially bigger mistakes, with an MAE of around $7,699 and a particularly high MAPE of 31.39%, indicating that the model's predictions are wrong by more than 30% on average. The RMSE for out-of-state forecasts is a staggering $7,849. This highlights significant mistakes that might be caused by outliers, structural data difficulties, or model deficiencies. These results clearly indicate the need for model evaluation, either including advanced modelling approaches such as SARIMA or integrating new variables that might better capture the factors impacting out-of-state tuition.

Implications and recommendations:

- The lower MAPE for in-state tuition suggests a strong model fit for this segment, indicating that the model well reflects the key dynamics. However, the model for out-of-state tuition is less dependable, as seen by a high MAPE.
- Given the high RMSE and MAPE for out-of-state tuition, it may be helpful to reassess the model's parameters, add additional explanatory variables if feasible, or experiment with other modeling methods such as SARIMA or machine learning approaches.
- The considerable differences in out-of-state tuition forecasts recommend looking for outliers, data input mistakes, or structural breaks in the series that might have impacted the model's performance.

By adopting these insights, you may improve the accuracy and dependability of your forecasting models, resulting in more successful tuition planning and financial strategy formulation.

Section 5: Findings

The project's data and research highlight many key aspects of rising tuition and accessibility issues in US higher education:

Over the past 40 years, college tuition has outpaced household income growth, especially for 22-27-year-olds. Students and their families suffer economically from this.

Different attributes of public and private institutions: Although public schools are cheaper than private ones, tuition and fees have risen significantly. State financing for public institutions has decreased, adding to higher expenses.

Student loans make up many US non-mortgage debt due to growing tuition. Student debt worsens inequality for low-income, non-traditional, and minority students.
The research emphasizes data-driven solutions to these issues.

Stakeholders can understand tuition rise and accessibility inequalities by using large datasets and advanced analytics like clustering and regression models.

The interactive dashboard empowers politicians, educators, students, and parents by providing a central solution. This software provides clear data visualizations and real-time analytics to compare tuition fee increases to median household incomes in different regions. It's a strong tool for policy reform and helps students and families navigate college affordability.

The project emphasizes the need for higher education financing reforms to ensure fair and equal education for all. Stakeholders can make higher learning easier and less costly by addressing the main causes of rising tuition and inconsistent access.

Section 6: Summary

The research examines the vital problem of rising tuition prices and the adverse effects that they have on access to higher education in the United States for the past two decades. Within the last two decades, tuition costs have increased at a faster pace than family income growth, leading to the expanding of economic inequality. The suggested answer is to develop an intricate and dynamic dashboard that analyzes tuition and fee increases with median family income in various regions. To reach its objective of offering transparent data visualizations and real-time insights, the platform employs visualization tool infrastructure and analytics. Its goal is to provide numerous stakeholders, such as academics, legislators, and students, with the ability to campaign for policy changes and make educated decisions. In addition to working as a lobbying tool for legislative reform, the project's purpose is to provide students and families with comprehensive information, transparency, and accessibility to assist them in overcoming the hurdles to higher education.

Section 7: Future Work

The interactive dashboard that has been suggested seeks to shed light on the discrepancy between growing college costs and median household income in different US regions, with a particular emphasis on how this affects educational disparities. Future developments might include integrating predictive analytics to predict tuition and income trends, as well as improving data integration by adding more cost components like housing and textbooks. Students and families preparing for college could gain a great deal from improved user interaction through customisable scenarios and personalised reports. The development of a mobile application would improve accessibility, particularly for those residing in underprivileged areas. By working together with policymakers and educational institutions, the dashboard's impact and reach can be increased and its data will be used to support legislative decision-making. More thorough insights might

be obtained by modifying the dashboard to accommodate usage in various regions and countries and adding indicators like employment results and graduation rates. The dashboard's usefulness and relevancy can also be increased by incorporating user feedback mechanisms and integrating instructional modules on financial literacy. The dashboard's evolution in these directions has the dual benefit of actively supporting initiatives to improve economic mobility and lessen income gaps among disadvantaged groups, in addition to acting as a crucial instrument for drawing attention to educational discrepancies.

Appendix

Appendix A: Glossary

Term	Definition
Academic Year	Described as the academic year, in accordance with the Board of Regents convention, as the duration of an educational institution's yearly sessions based on the summer, fall, and spring seasons for most institutional research reasons.
AWS Cloud Trail	CloudTrail allows for auditing, security monitoring, and operational troubleshooting by recording user activity an API usage.
Big Data	The data which has all these characteristics: Volume, Variety, Velocity, Veracity. It mostly varies depending on the source of the data.
Bucket Policies	Bucket policies are an Identity and Access Management (IAM) method that controls access to resources.
Cost of Attendance (COA)	The entire anticipated cost of attending a college for a single year (tuition and living), which includes accommodation and board, books, and other incidentals.
Dashboard	An interactive visual interface that displays the data in multiple visual formats that allows users to interact and draw insights from the data.
Data Integration	Bringing together information from several sources to create a single, cohesive data location.
Data Type Conversion	Converting the data type of a column or variable in a dataset involves altering values from one format to another, such transforming a text into a numeric type or a timestamp into a datetime format.
Federal Reserve Economic Data (FRED)	A database maintained by the Research division of the Federal Reserve Bank of St. Louis that has more than 500,000 economic time series from 87 sources.
Imputation	Imputation is a technique used to replace missing data with estimated values. Common ways include using the mean, median, or mode of the column, as well as more advanced techniques such as regression or K-nearest neighbours (KNN).
Integrated Postsecondary Education Data System (IPEDS)	A series of surveys carried out yearly by the National Centre for Education Statistics (NCES) of the U.S. Department of Education, which collects data from all colleges, universities, and technical and vocational institutions that take part in federal student financial assistance programs.
Interactive Filters	UI components that let users dynamically change the data set that's shown on a dashboard based on criteria like institution type, date period, or geography

K Means Clustering	Is a vector quantization method, originating from signal processing, that tries to partition N observations into K clusters, with each observation belonging to the cluster with the nearest mean.
Linear Regression	Linear regression analysis predicts the value of one variable depending on the value of another.
Machine learning	A branch of artificial intelligence that deals with creating algorithms that are not explicitly designed for a given purpose, but rather have the ability to learn and make predictions or judgments.
Mean Absolute Error (MAE)	Matric that measures the average magnitude of errors in a set of predictions, without considering their direction.
Median Household Income	The average household's income is represented by the middle-income value, which is the result of sorting all households' incomes from lowest to highest
Root Mean Squared Error (RMSE)	Standard way to measure the error of a model in predicting quantitative data, representing the square root of the average squared differences between predicted and observed values.
Scenario Modelling	The dashboard to provide features that allow to simulate possible results from different financial aid policy improvements, such as how assistance modifications could lessen the impact of rising tuition on students from different socioeconomic backgrounds.
Stakeholder Engagement	The practice of sharing information and collaborating with relevant parties such as parents, teachers, students and legislators, to facilitate the creation of supportive coalition for change.
Time Series Analysis	A statistical method for handling time series data, also known as trend analysis, which examines distinct data points gathered over an extended period.

Table 1: Glossary Table

Appendix B: GitHub Repository

Overview

This research goal to create a dashboard that can properly identify the tuition trends in the United States of America.

The system's goal is to create a dashboard these are some of the aims and goals of project:

1. How have tuition fees evolved over the past two decades across different types of institutions (public, private non-profit) in the United States?
2. What are the geographical variations in college affordability and access across different counties and states in the United States?
3. What are the top five universities with the highest tuition fees in each state?
4. How do trends in median household income correlate with the rising costs of university?

GitHub Repository Link

https://github.com/cpallamr/Balancing-the-Scales-Navigating-College-Affordability-in-Contrast-to-U.S.-Median-Household-Income

GitHub Repository Contents

At present, Balancing the Scales Navigating College Affordability in Contrast to U.S. Median Household Income and Action the Readme file, the main component of the GitHub repository, provides a quick overview and instructions for anyone interested in the project. Following are the topics covered in the Readme file: Problem description.

Appendix C: Risks

Sprint 1 Risks

Risk	Description	Probability	Impact	Mitigation
Data Availability (before 2000)	Difficulty in obtaining historical data from IPEDS.	High	High	Establish early partnerships with data providers; explore alternative data sources.
Data Integrity	Inaccuracies in the data due to reporting errors or inconsistencies.	Medium	High	Implement rigorous data validation and cleaning processes.
Regulatory Compliance	Ensuring all data usage complies with privacy laws and institutional policies.	High	High	Consult with legal experts to ensure full compliance with all relevant regulations.

Table 2: Sprint 1 Risks

A data processing and analysis project's three main hazards are listed in the risk table. The difficulty in acquiring historical data from IPEDS makes data availability the first risk, which has a high probability and impact. Early collaborations with data providers and searching for alternate data sources are key components of the mitigation plan. A medium chance and high impact are associated with the second risk, data integrity, which is related to probable inaccuracies brought on by reporting errors or inconsistencies. Strict procedures for data cleaning and validation are advised in order to lessen this. Regulatory compliance, in terms of following institutional policies and privacy regulations, has a significant risk in terms of both probability and impact. As part of the mitigation plan, legal professionals are consulted to guarantee that all applicable requirements are fully followed.

Sprint 2 Risks

Risk	Description	Probability	Impact	Mitigation
Data Availability	Difficulty in obtaining historical data from IPEDS.	Mitigated	low	Establish early partnerships with data providers.
Data Integrity	Inaccuracies in the data due to reporting errors or inconsistencies.	low	low	Implement rigorous data validation and cleaning processes.
Regulatory Compliance	Ensuring all data usage complies with privacy laws and institutional policies.	Mitigated	low	Data obtain from legalized website
Missing values	It refers to instances within an observation where a variable lacks a data value.	Medium	low	We have a special instruction from the client to replace the missing data.

Table 3: Sprint 2 Risks

The four main risks that come with undertaking a data handling project. Since acquiring historical data from IPEDS might be challenging, the Data Availability risk is considered low effect and is currently managed by forming early collaborations with data suppliers. With rigorous data validation and cleaning procedures as part of mitigation techniques, the Data Integrity risk—which is defined by probable inaccuracies owing to reporting errors or inconsistencies—also has a low probability and impact. The Regulatory Compliance risk, which is now minimized and regarded as low impact due to data being gathered from websites that have been legalized, concerns making sure that all data usage complies with institutional policies and privacy regulations. The last risk category is lacking Values, which describes situations in which data variables are lacking values. This risk is assessed as medium in probability but low in impact, and it may be managed with a client-directed approach to replace the missing data.

Sprint 3 Risks

Risk	Description	Probability	Impact	Mitigation
Incomplete Data Acquisition	Data sources may not provide all the required data for analysis.	Medium	High	Identify alternative data sources during data acquisition phase (Week 1). Explore data imputation techniques to address missing values if necessary.

Data Cleaning Challenges	Unexpected complexities in data cleaning may arise, extending the timeline.	Medium	Medium	Allocate buffer time in Sprint 3 for potential data cleaning challenges. Utilize profiling and visualization techniques to identify data inconsistencies early.
System Design Errors	Errors in system architecture design may lead to security vulnerabilities or data flow issues.	Low	High	Conduct thorough code reviews and testing of the system architecture (Week 2). Consult with security experts if needed.
Model Selection Issues	The chosen machine learning model(s) may not be suitable for the data or problem.	Medium	Medium	Conduct comprehensive exploratory data analysis (Week 3) to understand data characteristics before model selection. Evaluate and compare performance of multiple machine learning models.
Communication Breakdown	Lack of clear communication within the team or with stakeholders could lead to misunderstandings and delays.	Low	Medium	Maintain clear and concise documentation of decisions and progress. Hold regular team meetings (daily scrums) to facilitate open communication.

Table 4: Sprint 3 Risks

The risk management strategy lists the most important weaknesses and possible problem-solving techniques throughout the duration of the project. Since incomplete data acquisition has a substantial impact on analysis, it is a serious concern that can be lessened by finding alternate sources and using data imputation techniques. Less severe Data Cleaning Challenges are handled by setting up buffer time and using tools for early identification in profiling and visualization.
Although less common, system design errors have a significant impact and necessitate extensive code reviews, testing, and professional advice to guarantee system resilience. Moderately risky model selection issues require thorough data analysis prior to choosing machine learning models, with a focus on performance comparison and evaluation. Even if it is less common, communication breakdowns can still impede development and stakeholder alignment. These can be avoided by carefully documenting everything and holding frequent team meetings to promote open communication. By combining these strategies, the project is strengthened against a range of possible obstacles, leading to more efficient project management and alignment with goals.

Sprint 4 Risks

Risk	Description	Probability	Impact	Mitigation
Geolocation Data Inaccuracy	Data acquired for latitude and longitude may be inaccurate or incomplete.	Medium	Medium	Utilize data validation techniques to assess geolocation data accuracy (Week 1). Explore alternative sources for geolocation data if necessary.
Overfitting of ML Models	Machine learning models may overfit the training data, leading to poor performance on unseen data.	Medium	High	Implement techniques like cross-validation and regularization during model development (Week 2). Evaluate model performance on a separate hold-out test set.
Limited Model Generalizability	Chosen machine learning models may not generalize well to future data.	Medium	Medium	Utilize a diverse dataset for training the models to improve generalizability. Consider incorporating domain knowledge into the modeling process.
Dashboard Design Issues	The developed dashboard may not meet the client's needs or expectations.	Low	Medium	Maintain clear communication with the client throughout development (all weeks). Conduct regular demonstrations and incorporate client feedback (Week 3).
Data Visualization Misinterpretation	Data visualizations may be misinterpreted by users due to poor design or unclear presentation.	Low	Medium	Apply best practices for data visualization design (Weeks 1 & 3). Ensure visualizations are clear, concise, and aligned with the intended message.

Table 4: Sprint 4 Risks

Several risks have been observed and addressed during the machine learning-based solution development process to guarantee the project's success. One possible issue that could impact the model's dependability is the precision and comprehensiveness of the geolocation data. Techniques for validating data will be used to reduce this risk, and if needed, other sources of geolocation information will be investigated. Overfitting machine learning models to training data might result in subpar performance on untrained data, which is a serious issue. To address this problem, methods like regularization and cross-validation will be used when developing the model, and its performance will be assessed on a different set of hold-out tests.

Moreover, a medium-level risk is presented by the selected machine learning models' restricted generalizability. To improve the models' capacity to generalize to new data, a varied dataset will be used for training. To improve model performance, domain expertise will also be included into the modeling process. Issues with dashboard design and misinterpretation of data display provide possible problems from the user interface perspective. The client will be kept informed at every stage of the development process to reduce these risks. Clear and consistent communication with the intended message will be ensured by applying best practices for dashboard design and data visualization, as well as regular demonstrations to solicit customer feedback. All in all, the project seeks to reduce potential setbacks and offer a strong machine learning solution that satisfies the client's goals and expectations by proactively identifying and addressing these risks through targeted mitigation measures deployed throughout the development schedule.

Sprint 5 Risks

Risk	Description	Probability	Impact	Mitigation
Incomplete Data Insights	Extracted insights from data visualizations may be incomplete or lack depth.	Low	Medium	Conduct thorough review of data visualizations and ensure all relevant insights are captured (Week 1). Utilize brainstorming techniques to explore potential deeper interpretations of the data.
Dashboard Misinterpretation by Audience	The intended message of the dashboards may be misunderstood by the target audience (classmates in this case).	Low	Medium	Conduct user testing with representatives from the target audience to gather feedback on dashboard clarity (Week 1). Refine dashboards based on user feedback to improve clarity and communication (Week 2).
Presentation Shortcomings	The final presentation may not effectively communicate the project findings or be engaging for the audience.	Low	Medium	Rehearse the presentation beforehand to ensure clarity, flow, and timing (Week 2). Incorporate visuals and storytelling elements to enhance audience engagement.
Project Report Issues	The final project report may contain errors, inconsistencies, or lack clarity.	Low	Medium	Conduct thorough peer review of the report to identify and address any issues (Week 2). Ensure the report adheres to the required format and style guidelines.
Time Pressure for Completion	The team may face time constraints in finalizing deliverables (presentations, reports) within the sprint timeframe.	Low	Medium	* Prioritize tasks effectively and allocate sufficient time for finalization (Week 2). * Communicate any potential delays to stakeholders proactively.

Table 5: Sprint 5 Risks

This risk-management plan for a data visualization project lists several potential hazards together with the implications, probability, and mitigation techniques associated with each. The risks cover a wide range of topics, such as inadequate data insights, audience misinterpretation of the dashboard, poor presentation quality, problems with the project report, and deadline pressure. Every risk is ranked as having a medium impact but a low chance, emphasizing the value of preventative mitigation strategies. To mitigate these risks, a few techniques are suggested, including comprehensive evaluations, user testing, presentation rehearsing, peer review of reports, and efficient time management. The project team intends to provide accurate and compelling data insights to the target audience while reducing the possibility of problems and guaranteeing the project's successful completion within the allotted timeframe by putting these mitigation techniques into practice.

Appendix D: Agile Development

Scrum Methodology

Data analytics projects, which frequently contain uncertainty and need for flexibility, are ideally suited for Scrum, an incremental and iterative Agile project management approach. To handle this research project, our team used the scrum methodology.

Adapting to Scrum:

The scrum technique proved to be quite straightforward for the team to adjust to. The scrum structure is easy to comprehend and use, and the group members were already educated with some of the fundamental ideas, like sprints and backlog grooming. Though the daily scrum sessions demand a high degree of discipline and attention, the team did discover that it takes some time to become used to them.

In Person Meetings:

The team occasionally had in person brainstorming sessions twice or thrice a week, where we used to discuss individual problems faced and try to mitigate them immediately and prepare the in-class presentation for updates and summarize each week's work goals into a report.

You Track Tool & Usage

Daily Scrum Meetings:

Every day, the team held scrum meetings, which lasted for fifteen to twenty minutes on average. Each weekday began with these sessions, which gave team members a chance to discuss their accomplishments, pinpoint any obstacles, and modify their daily objectives. The team stayed focused on its objectives and on course thanks to the daily scrum meetings. Every day, the squad met in scrums, which usually lasted fifteen to twenty minutes. Team members had the chance to discuss their progress, pinpoint any obstacles, and modify their daily goals at these sessions, which started each workday. The team was able to stay focused and on task thanks to the daily scrum meetings.

Sprint 1 Analysis:

Team Analysis: Our team project, "Balancing the Scales: Navigating College Affordability in Contrast to U.S. Median Household Income," focuses on comparing university fees across all states in the USA with median income. For this sprint, our team initially collaborated on breaking down tasks and dividing them among us, including user stories, stakeholders, and problem definition.

To gather data for this study on student tuition costs, we examined actual cases of students who were having trouble selecting a college due to a variety of criteria, including living costs, tuition costs, and state

income. This dashboard helps students choose which school, college, or university to attend by giving them a thorough overview of living expenses, income, and tuition costs in each state.

This dashboard will give parents with necessary information regarding tuition prices and living expenditures in different states of the USA, allowing them to prepare ahead financially for their children's education. Reformers can utilize the dashboard's capabilities to evaluate the disparity between the rise in tuition prices and the stagnant or slow increase in median household earnings in the past twenty years. This would assist in identifying specific days or trends when the disparity significantly rose, indicating a higher demand for financial aid and student support services. They utilize this data to advocate for policy changes aimed at addressing deficiencies and enhancing education accessibility when presenting to institutional leaders.Managing activities for this sprint was not challenging for our team because we all scheduled our daily tasks and allocated time to complete tasks, making it easy to manage our activities efficiently.

To enhance the study, the team could enhance the data collection process by exploring additional data sources beyond IPEDS and FRED. Increasing the scope of data collection could have provided a more comprehensive insight into the variables influencing fluctuations in tuition fees and household incomes. A more iterative method for selecting and validating models might have resulted in more precise adjustments and improvements to the predictive capabilities of the dashboard, ensuring its capacity to adapt to new trends and data.

Sprint 2 Analysis:

Team Analysis: Our team project, "Balancing the Scales: Navigating College Affordability in Contrast to U.S. Median Household Income," focuses on comparing university fees across all states in the USA with median income. For this sprint, our team initially collaborated on breaking down tasks and dividing them among us, including extracting Tuition and Fees data from IPEDS and Median Household income families of each state from FRED, Data cleaning, merging the data to a raw file, Naming the attributes, handling missing values.

The team concentrated on data preparation and extraction for Sprint 2. Since the goal of this sprint was to provide the foundation for further investigation, user stories were not specifically specified. But the main objective was to produce a clear and useful dataset for additional research.

The group successfully downloaded data from FRED and IPEDS, demonstrating good resource management and teamwork. The team's cooperative spirit was seen during the fruitful data merging brainstorming session. But cleaning and merging took longer than anticipated, suggesting that the operation's difficulty was overestimated. Preliminary exploratory data analysis demonstrates an understanding of the data exploration process.

It became challenging to oversee the sprint since the team underestimated the amount of time needed for data purification. As a result, the focus switched from adding new datasets to completing primary EDA and data purification. Accurate time estimation is crucial for data cleansing tasks. Using techniques like story-pointing or timeboxing might be helpful. While maintaining concentration on the sprint goal is essential, flexibility is also required to get over unforeseen challenges.

Cooperation and honest communication are crucial throughout changes. This sprint highlights the iterative nature of data analysis activities. Even though the team's early estimations weren't perfect, they demonstrated their ability to adjust and advance during the sprint.

To further improve the study, Segment large, complex jobs, such as data cleansing, into smaller, more manageable subtasks for better estimation and progress tracking. Maintain open channels of communication with team members and stakeholders to resolve issues and make any required plan revisions. Examine tools and techniques for data wrangling to speed up the process of cleaning and merging data.

Sprint 3 Analysis:

Team Analysis This section describes the team's third research project sprint's use of the Scrum methodology. By Sprint 3, the team found it easier to adjust to Scrum. The early difficulties in comprehending roles and rituals had been solved. Still, continuous process improvement is always essential. During this Sprint, the team most likely convened daily scrum sessions. These quick sessions (15–20 minutes) are crucial for presenting one's progress on the day's work, locating any obstacles or hurdles preventing advancement, modifying preparations for the following day to make sure objectives are achieved.

YouTrack most likely turned out to be a useful tool for tracking progress and managing the backlog of projects. Its ease of usage ought to have enabled Giving team members duties to do, monitoring the state of job completion and determining the relationships among the jobs.
displaying the project backlog's progress visually.

Lessons from Sprint 3, Continuous Refinement to adapt Scrum procedures to the unique requirements of the team by regularly evaluating and improving them. Daily scrums are essential for keeping communication open, staying focused, and quickly resolving issues. As a Facilitator, YouTrack Make use of YouTrack's features to enhance teamwork and expedite project management.

Extra thoughts on the tasks of Sprint 3, focusing on Data Architecture the first week on data architecture emphasizes the significance of laying a strong foundation for data analysis. For System Architecture Exploration, the team appears to be thinking about possible deployment choices and data security measures based on their investigation of system architecture in week two. For model discovery and early EDA, the emphasis placed in Week 3 on these two topics demonstrates a comprehensive strategy that balances comprehending the data with the discovery of analytical methods.

For further improvement of study, Accept Continuous Learning throughout the project, keep learning and modifying Scrum techniques. Make sure the sprint reviews and daily scrums are targeted and fruitful. Record the most important lessons learned from each sprint for your future use and project enhancement.

Sprint 4 Analysis

By Sprint 4, the team should be at ease adjusting to Scrum. Most of the early difficulties in comprehending roles and rituals would have been resolved. Still, constant improvement is necessary to achieve the best outcomes.

During Sprint 4, the team most likely kept up their regular scrum meetings. These quick sessions (15–20 minutes) are crucial for presenting one's progress on the day's work, locating any obstacles or hurdles preventing advancement. modifying preparations for the following day to make sure objectives are achieved.

YouTrack ought to have remained an important resource for tracking development and managing the backlog of projects. Its ease of usage ought to have enabled giving team members duties to do, monitoring the state of job completion, determining the relationships among the jobs, displaying the project backlog's progress visually.

Learnings from the Fourth Sprint include continuous Improvement in adapting Scrum procedures to the team's workflow by regularly assessing and improving them. Daily scrums are essential for keeping communication open, staying focused, and quickly resolving issues. YouTrack as a Collaboration Tool and its features were helpful to enhance teamwork and expedite project management.

Remarks on the Activities of Sprint 4 are mainly, data acquisition for geolocation emphasizing the collection of latitude and longitude in Week 1, the team is enhancing the data in preparation for additional analysis. Data Visualization and EDA, the team appears to be improving their comprehension of the data and its linkages based on their continued work on these two areas.
Machine Learning and Prediction, this week's highlights include the group's advancements in developing models and making future projections. Development and Improvement of the Dashboard by team show's a commitment to iterative improvement and responsiveness to customer requests.

Sprint 5 Analysis

By Sprint 5, the team should have gotten used to implementing Scrum. Most of the early difficulties in comprehending Scrum roles and rituals would have been solved. But it's crucial to keep in mind that Scrum is an iterative process, and improvement is something that should never stop.
It's probable that the team kept up their regular scrum sessions during the project. These quick (15–20 minute) sessions are crucial for sharing updates on work completed the day before, identifying obstacles or blockages that need to be removed, and modifying plans for the next day to guarantee that objectives are accomplished.

YouTrack ought to have shown itself to be a useful instrument for managing the backlog of projects and monitoring development over the course of sprints. Its ease-of-use ought to have made it easier to assign tasks to team members, monitor task completion, find linkages between tasks, and visualize project backlog progress.

Among the lessons this sprint taught us are Scrum gave the team a structure for iterative development, enabling them to modify and improve their strategy as the sprints went by. Daily scrums made sure that everyone on the team stayed on task, spoke clearly, and dealt with problems quickly. YouTrack made project management more effective and team member collaboration better.

Remarks on the Activities of Sprint 5 are the first week's emphasis on completing data visualizations and deriving insights shows how the data analysis work has end. The team was getting ready to share its results, as seen by their continued work on dashboards and presentations in Week 2.

References

1. *Bucket policies and user policies - Amazon Simple Storage Service*. (n.d.). https://docs.aws.amazon.com/AmazonS3/latest/userguide/using-iam-policies.html
2. *Cloud computing* services - *Amazon Web Services (AWS)*. (n.d.). Amazon Web Services, Inc. https://aws.amazon.com/
3. C3.ai. (2022, March 31). *Mean absolute error*. C3 AI. https://c3.ai/glossary/data-science/mean-absolute-error/#:~:text=In%20the%20context%20of%20machine,errors%20for%20the%20entire%20group.
4. Cheng, L., & You, C. (2016). Analysis of tuition growth rates based on clustering and regression models. International Journal of Data Mining & Knowledge Management Process, 6(4), 01–17. https://doi.org/10.5121/ijdkp.2016.6401
5. CDATA ARC- Secure Data Integration & Managed File Transfer (MFT). (n.d.). CData Software. https://arc.cdata.com/?kw=data%20integration&cpn=17853316056&utm_source=google&utm_medium=cpc&utm_campaign=ArcESB_-_Search_-_General&utm_content=General&utm_term=e|data%20integration&kw=data%20integration&cpn=17853316056&gad_source=1&gclid=CjwKCAiA8sauBhB3EiwAruTRJiNxU3oNObf4EfBQ8xquXWpXxNBAgLmSRE9tyMotmuM1P69380QZ-hoC8XYQAvD_BwE
6. Davis, L., Wolniak, G. C., George, C. E., & Nelson, G. (2019). Demystifying Tuition? A content analysis of the information quality of public college and university websites. AERA Open, 5(3), 233285841986765. https://doi.org/10.1177/2332858419867650
7. Federal Reserve Economic Data | FRED | St. Louis Fed. (n.d.). https://fred.stlouisfed.org/
8. Federal Student Aid. (n.d.). https://studentaid.gov/help-center/answers/article/what-does-cost-of-attendance-mean
9. Gerasymov, O. (2024, February 24). Machine learning for time series forecasting. *CodeIT*. https://codeit.us/blog/machine-learning-time-series-forecasting#
10. Hildreth. (2024, January 10). Hildreth. https://www.hildrethinstitute.org/.
11. Helhoski, A. (2024, January 25). What is the median household income? NerdWallet. https://www.nerdwallet.com/article/finance/median-household-income
12. Hernandez, H. (2022, July 8). What is Data Conditioning and Cleaning? JANA, Inc. https://janacorp.com/what-is-data-conditioning-and-cleaning/
13. IPEDS. (n.d.). https://nces.ed.gov/ipeds/
14. Maldonado, C. (2018, July 24). Price of college increasing almost 8 times faster than wages. *Forbes*. https://www.forbes.com/sites/camilomaldonado/2018/07/24/price-of-college-increasing-almost-8-times-faster-than-wages/?sh=38197d6c66c1
15. Nair, M. (2023, January 6). *How the Meaning of Tuition is Changing*. University of the People. https://www.uopeople.edu/blog/how-the-meaning-of-tuition-is-changing/
16. Optilogic. (2023, August 21). *Optilogic | What's the difference between scenario modeling and simulation?*Optilogic. https://www.optilogic.com/resources/blog/whats-the-difference-between-scenario-modeling-and-simulation/
17. Simplilearn. (2023, August 16). *Introduction to data imputation*. Simplilearn.com. https://www.simplilearn.com/data-imputation-

article#:~:text=Data%20imputation%20is%20a%20method,from%20a%20dataset%20each%20time.
18. *Time Series Analysis: Definition, Types & Examples | SiGMA Computing*. (n.d.). https://www.sigmacomputing.com/resources/learn/what-is-time-series-analysis
19. *What is FRED? | Getting To Know FRED*. (n.d.). https://fredhelp.stlouisfed.org/fred/about/about-fred/what-is-fred/
20. *What is FRED? | Getting To Know FRED*. (n.d.). https://fredhelp.stlouisfed.org/fred/about/about-fred/what-is-fred/
21. *What is Machine Learning? | IBM*. (n.d.). https://www.ibm.com/topics/machine-learning
22. Wikipedia contributors. (2024, April 26). *Root-mean-square deviation*. Wikipedia. https://en.wikipedia.org/wiki/Root-mean-square_deviation

This page intentionally left blank