2022

Leveraging Data Analytics To Study The Impacts Of State Budget Cuts On Public Higher Education Institutions In The United States

Praveen Kumar Guraja
Indiana State University

LEVERAGING DATA ANALYTICS TO STUDY THE IMPACTS OF STATE BUDGET CUTS ON PUBLIC HIGHER EDUCATION INSTITUTIONS IN THE UNITED STATES

A

Dissertation

Presented to

The College of Graduate and Professional Studies

The College of Technology

Indiana State University

Terre Haute, Indiana

In Partial Fulfillment

of the Requirements for the Degree

Ph.D. in Technology Management

---

by

Praveen Kumar Guraja

July 2022

Keywords: Data analytics, graduation rate, budget cuts, enrollment, state appropriations

VITAE

PRAVEEN KUMAR GURAJA

EDUCATION

Ph.D.	Indiana State University, Terre Haute, Indiana, 2022 Major: Technology Management
M.S.	Pittsburg State University, Pittsburg, Kansas, 2017 Major: Technology
B.Com.	Acharya Nagarjuna University, Guntur, Andhra Pradesh, 2009 Major: Business Management

PROFESSIONAL AND ACADEMIC EXPERIENCE

2021 – Current	University of Pikeville, Pikeville, KY Assistant Professor - Information Technology Management, College of Business
2019 – 2021	Pittsburg State University, Pittsburg, KS Adjunct Faculty, College of Technology and College of Business
2017 – 2021	Pittsburg State University, Pittsburg, KS Database Administrator, University Advancement
2016 – 2017	Pittsburg State University, Pittsburg, KS Graduate Teaching Assistant, College of Education
2015 – 2015	Automatic Data Processing (ADP) Pvt. Ltd., Hyderabad, India Senior Technical Associate
2014 – 2015	Unisys Pvt., Ltd., Hyderabad, India Senior Support Associate-2A
2010 – 2014	Genpact India Pvt., Ltd., Hyderabad, India Technical Specialist

CERTIFICATIONS

2021 – Current	IBM Cybersecurity Analyst International Business Machines (IBM)
2020 – Current	Certified IT Project Manager Indian School of Business (ISB)
2018 – Current	Lean Six Sigma Green Belt Certified Six Sigma Global Institute (SSGI)

COMMITTEE MEMBERS

Committee Chair: M. Affan Badar, Ph.D.

Professor of Applied Engineering and Technology Management

Indiana State University, Terre Haute, Indiana, USA

Committee Member: Farman Moayed, Ph.D.

Professor of Built Environment

Indiana State University, Terre Haute, Indiana, USA

Committee Member: Christopher Kluse, Ph.D.

Associate Professor of Quality Systems

Bowling Green State University, Bowling Green, Ohio, USA

ABSTRACT

Public higher education in the United States (US) is funded through two primary forms: one is through state higher education appropriation funds, and the other is student financial aid that is directly given to students. Increasing postsecondary full-time equivalent (FTE) enrollment and graduation rate are becoming a crucial economic priority in the US. However, only a limited study is available about whether state investment in higher education and increasing tuition charges can impact FTE student enrollment (FTEE) and graduation rate (GR) at 4-year public universities in the US. A systematic literature review was conducted for the present research to comprehend the literature gap and identify factors or variables that may affect FTEE and GR. Five independent variables (IVs): state higher education appropriations per FTE (SHEA), average undergraduate charges per FTE (AUGC), student tuition share as a percentage of per capita income (STSPCI), state higher education appropriations as a percentage of GDP (SHEAGDP), and state financial aid (SFA) per FTE were selected. The dependent variables (DVs) were full-time equivalent enrollment (FTEE) and graduation rate (GR) at 4-year US public higher education institutions. Historical US public higher education data for 50 years (each year as one dataset, n=50) between 1971 and 2020 were collected and analyzed. The multiple linear regression tool of the open-source data analytics and machine learning software was used to test the hypotheses if the independent variables were significantly related to the dependent variables. Hypothesis 1 was about FTEE, and hypothesis 2 was

about GR.

For FTEE, three variables were found to be significant: SHEA, AUGC, and SFA. For GR, two variables were found to be significant: SHEA and STSPCI. Hence, two data analytical models were developed involving the significant IVs: one for FTEE and the other for GR. Findings from the first model revealed that when state higher education appropriation (SHEA) funds increase, average undergraduate tuition charge (AUGC) decreases, and more student financial aid (SFA) is awarded, FTE enrollment (FTEE) increases. The second model results indicated that when state higher education appropriation (SHEA) funds increase and student tuition share as a percentage of per capita income (STSPCI) decreases, there is an increase in graduation rate (GR). These findings show how state budget cuts could impact students enrolling and graduating at public 4-year institutions in the US. State policymakers, higher education administrators, and other stakeholders could use this study to develop their customized data analytics and machine learning models and analyze their past data to better prepare themselves for future uncertainties.

This study did not investigate how state funding or budget cuts a) impact full-time faculty vs part-time faculty at public higher education institutions, as data were unavailable for some of the years, and b) if any specific year impact the corresponding year of cohort. These can be investigated in future work.

ACKNOWLEDGMENTS

My first and foremost thanks to God for his countless blessings, knowledge, and opportunities. My journey has been a long one, made possible only with the help of my family, friends, and colleagues along the way, who've encouraged me throughout my PhD journey. At the very top of what will be a genuinely fabulous list, pride of place goes to my dissertation committee.

To Dr. M. Affan Badar, my advisor: you have supported and accompanied me in this program. I'll be forever grateful for the learning, growing, and working with you. You always kept our careers in mind and encouraged us to do more academic research and publish more. Without his encouragement and guidance, this dissertation would not have materialized. As my committee member, I want to show my greatest appreciation to Dr. Farman Moayed for the valuable inputs to be included in this dissertation and for generously spending precious time giving guidance, encouragement, suggestion, and comments. As my committee member, I want to thank Dr. Christopher Kluse for his invaluable advice, continuous support, and patience during my research. Once again, I want to offer my sincere appreciation to all my committee members for the learning opportunities.

I want to extend my appreciation to wonderful people at my current workplace and a few special people at my previous workplace who supported me and helped me to grow professionally and personally. Thank you, my friend, Pawan Bhandari, for your friendship and support as we navigated our Indiana State University (ISU) journey. Several people in my life, starting from my school education until my university education, who supported and helped me study, without whom I would not have had the privilege of being educated. I offer my sincere appreciation to the leaders at ISU who helped me navigate this Ph.D. program over the last three and half years.

Finally, to my family, who have always been there for me: especially my beautiful wife Keerthi, for her love and constant support, I want to thank her very much for all the late nights and busy weekends she spent with me. Most importantly, I want to thank you for being my best friend. My parents, sister, and brother-in-law: thank you for your love, support, and unwavering belief in me. I want to thank my cousin Ratna IPS for her constant support and confidence in me since my childhood. I also would like to thank my school friend Harish for his encouragement and guidance for me in coming to the United States for higher education. Each of you who helped me be here today earned this degree right along with me.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Chapter Overview

The essence of this research study was clearly defined as problem statements, research questions, and hypotheses to address a statement of need and a statement of purpose. Assumptions, limitations, and delimitations were identified. Key terms were defined.

Background

Public higher education institutions and state governments always have a mutualistic relationship where they together play a crucial role in improving the US economy, where states primarily fund public higher education institutions, and these institutions, in turn, create educated youth. The unprecedented economic challenges caused by the COVID-19 pandemic could possibly result in reduced tax revenues, and it could directly affect higher education funding in the years to come. Public higher education was funded by state governments to improve their residents' economic and social well-being (Williams, 2016). According to Zhao (2018), higher education is one of the fields that contributes to several social and fiscal benefits. However, the money given to the public higher education sector remained 6% and 14.6% below the 2008 and 2001 levels, respectively (Jimmerson, 2021).

Amid state budget cuts to higher education and negative economic impacts on state budgets due to the COVID-19 pandemic, public higher education in the US is going through turbulent times and probably heading toward another possible recession. After the great recession in 2008, public higher education was one of the sectors that took huge cuts, and the states were not able to recover the funding back to earlier numbers. According to the State Higher Education Executive Officers Association (2021), during the 2008 Great Recession, there was more than $2,000 funding reduced per student, and by 2019, the student education appropriations per full-time equivalent (FTE) student were only $8,196, which is 8.7% below the prerecession level.

According to Enders and Jongbloed (2007), state governments are one of the main and crucial funding sources that provide for public higher education. Over the last two decades, higher education funding in the US has experienced a decline in state funding that could impact public institutions directly. More than ever before, less state funding is challenging public higher education institutions in the US to reduce waste and use resources efficiently. This research study addressed the issues associated with the distribution of education funds and the student graduation rate. The time has come for these public institutions to use technology to their advantage to make data-driven decisions using insights from advanced data analytical models and machine learning algorithms to manage their resources better.

According to Hanushek and Woessmann (2008), college-educated youth with postsecondary degree attainment play a crucial role in the economic growth of a country. More than ever before, less state funding is challenging public higher education institutions in the US, and more investments in higher education are needed. Ferlie et al. (2008) believe the state's role is expected to become stronger as higher education becomes bigger. Several authors discussed the connection between the higher education system and economic growth. Hence, this research study analyzes the impacts of state funding reductions on public institutions in the US on enrollment and student postsecondary attainment, i.e., graduation rate.

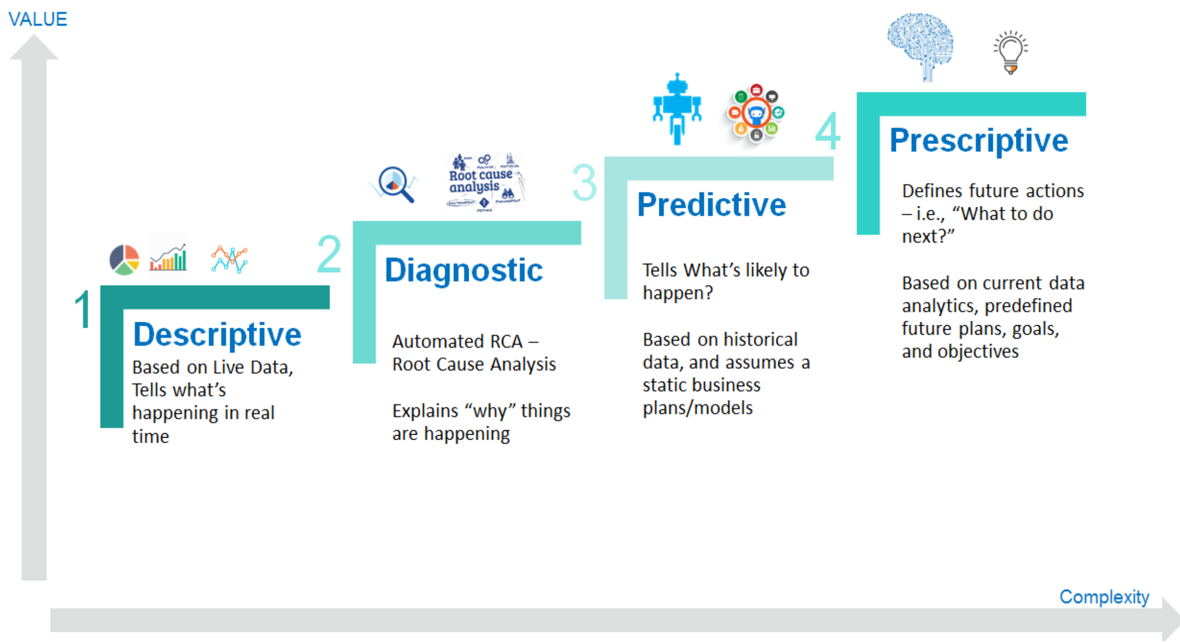
Data Analytics – Types and Steps

Data analytics is the science that helps in analyzing data to gain insights; it can be used to improve things and make inferences from the information. Data analytics tools using various models and algorithms can be used to analyze a variety of datasets. Through data analytical modeling techniques, data analysts can determine the best analytical solutions for an organization and make insightful decisions. Many data analytical techniques can be automated into algorithms that can be applied to different datasets. Data analytics helps reveal metrics and trends through data visualization, which otherwise is difficult to capture in the large stack of information (Cote, 2021).

Data analytics is an important tool because it can also forecast future trends based on historical data, and it helps optimize performance while reducing costs. This research not only studies the trends of past higher education data but also analyzes whether there is a significant relationship between higher education budget cuts and student outcomes based on historical information. According to Cote (2021), there are four major types of data analytics: descriptive analytics, diagnostic analytics, predictive analytics, and prescriptive analytics, as shown in Fig. 1. Descriptive analytics, after analyzing the data, helps in giving insights into what happened over a period; diagnostic analytics give insights into why something happened, and it involves hypothesizing; predictive analytics helps in giving a forecast of what is likely going to happen. Prescriptive analytics guides future actions based on current data analytics, project goals and objectives.

Figure 1

Four Types of Data Analytics Every Analyst Should Know-Descriptive, Diagnostic, Predictive, and Prescriptive



Note: Four Types of Data Analytics Every Analyst Should Know-Descriptive, Diagnostic, Predictive, and Prescriptive. Source: Kachchi & Kothiya, 2021

Most of the COVID-19 models built today to forecast the number of infections and hospital occupancy are based on predictive analytics. This is an important data analytical type that was used in this research to help higher education public institutions in the US gain insight into future trends. Prescriptive analytics is the last of the four types; it mostly deals with suggesting a course of action for better results.

Data Analytics and Higher Education

Advancements in technology bring both good and bad to how technology management practices are handled today. To solve the current problems in the world, technology, management, and data play an equally important role. In particular, higher education institutions hold a special place for data and technologies around data, as traditional degree programs are evolving to fit current workplace requirements (Choudhury, 2021). Data analytics is one such technology management tool that could help higher education institutions optimize their efficiency and improve their performance in this increasingly competitive environment. More data are contributed today by higher education systems than earlier because of digital transformation. Universities and colleges need advanced analytical tools to study the large demographic datasets at higher education institutions (Bonderud, 2020). Hence, academic institution leaders are showing interest in data analytics to improve their organizations. However, the education sector is still behind in using data analytics technology compared to other sectors, such as healthcare, which takes advantage of the data-driven approach (Nda, Tasmin, & Hamid, 2020).

According to Spear (2019), data analytics in higher education are instrumental in tracking trends that help better plan an institution's success. Data analytics in higher education could provide incredible insights into admission rates, enrollment counts, classroom utilization, persistence rates, and graduation rate, to name a few variables. It provides access to customized reports and data trends per the institution's needs and provides research departments with higher-value research. Data analytics could help investigate institutional barriers that lead students to leave an institution or stop their college journey. Kiu (2018) conducted a study to predict student's performance based on student's background and social activities using data analytics tools and techniques. It could provide insights to improve student retention that encourage campus leaders to move their institutions toward data insights and from data insights to practical action, ultimately leading to student success practices (Carmean, Kil, & Baer, 2021).

Statement of Problem

State funding to public higher education institutions in the US has been decreasing steadily for the past two decades, and with COVID-19, it is expected to become even worse in the coming years. Public higher education is primarily supported by local state governments through state higher education funding. That funding is again divided into two types: the first type of funding is sent to public higher education institutions as general state higher education operating support that may help in reducing the burden of tuition rate increases for all students, and the second type of funding is directed to students through state financial aid programs that target low-income students. Currently, there is a shortage of youth with postsecondary degrees and people with high-quality skills. Even the wage gap between the postsecondary degree holder and a person without a degree is widening in the US (Hershbein, Kearney, & Pardue, 2020). As the graduation rate is directly related to US economic growth, this study aims to apply data analytics methods and machine learning algorithms to gain a deeper understanding of how state investment in higher education impacts variables such as graduation rate, enrollment, student tuition share, and other important variables.

Most public higher education institutions in the US today are under severe pressure with diminishing financial support from state government and trying to maintain enrollment numbers. This study expands on the present literature available in two important ways. First, it employs the latest data analytics techniques and machine learning algorithms to analyze the relationship between variables impacting public postsecondary institutions in the US. This analysis provides insights and trends for state decision-makers. Second, we analyze the predictive relationship between independent and dependent variables based on historical information to better prepare public higher education institutions in the US for future uncertainties.

Research Questions and Hypotheses

This research study used the following research questions based on a quantitative study framework. Note that FTE is full-time equivalent.

RQ1. Is there a significant predictive relationship between the five independent variables (IVs): state higher education appropriations per FTE (SHEA), average undergraduate charges per FTE (AUGC), student tuition share as a percentage of per capita income (STSPCI), state higher education appropriations as a percentage of GDP (SHEAGDP), and state financial aid (SFA) per FTE, and dependent variable (DV): full-time equivalent enrollment (FTEE) at public higher education institutions?

RQ2. Is there a significant predictive relationship between the five IVs: state higher education appropriations per FTE (SHEA), average undergraduate charges per FTE (AUGC), student tuition share as a percentage of per capita income (STSPCI), state higher education appropriations as a percentage of GDP (SHEAGDP), and state financial aid (SFA) per FTE, and DV: graduation rate (GR) at public higher education institutions?

The associated null hypothesis statements are as follows:

H_{01}: There is no significant predictive relationship between the five independent variables (IVs): state higher education appropriations per FTE (SHEA), average undergraduate charges per FTE (AUGC), student tuition share as a percentage of per capita income (STSPCI), state higher education appropriations as a percentage of GDP (SHEAGDP), and state financial aid (SFA) per FTE, and dependent variable (DV): full-time equivalent enrollment (FTEE) at public higher education institutions.

H_{02}: There is no significant predictive relationship between IVs: state higher education appropriations per FTE (SHEA), average undergraduate charges per FTE (AUGC), student tuition share as a percentage of per capita income (STSPCI), state higher education appropriations as a percentage of GDP (SHEAGDP), and state financial aid (SFA) per FTE, and DV: graduation rate (GR) at public higher education institutions.

Statement of Purpose

The purpose of the paper is to examine the impacts of state budget cuts on public higher education institutions in terms of FTE student enrollment and student graduation rate. The focus was on analyzing higher education appropriations and how they impacted student enrollment and the student graduation rate. More extensive insights and the significance of the impacts of state budget cuts on the enrollment and graduation rate of students help these institutions survive in the years to come, with more budget cuts expected due to the impact of the COVID-19 pandemic on tax revenues.

Statement of Need

Students and families have been helpless with the rise of college costs for the past 30 years. However, the economic value of attaining a postsecondary degree has increased. Hence, student enrollment and graduation rate have become crucial metrics for the US economy, as a future workforce with a postsecondary degree plays a vital role. A reduction in higher education appropriations and the impacts it could have on various variables, such as enrollment, graduation rate, student tuition share, and other variables, has become a concern. Especially with the COVID-19 pandemic and the future budget cuts that are very likely due to fewer tax revenues, more pressure is placed on public higher education institutions in the US (SHEF Report, 2021). Hence, more than ever before, there is a need for deeper research to investigate the correlation between these variables and analyze their relationship to better prepare these institutions.

Most of the studies emphasized funding cuts and increased education and higher education costs. However, there is only limited research available on state appropriations and their impact on student enrollment and graduation rate using data analytics. This research adds significant new knowledge in the area along with the use of the latest technologies, such as data analytics and machine learning, which could help in better decision making. As states could struggle to resolve the problem of affording to provide adequate funding to higher education institutions amidst a pandemic, this research study needs states to think critically about the cost of failing to do so.

Statement of Assumptions

The main assumption of this study was that the effect of change in state appropriations is linear. The data analytics model and machine learning algorithms developed in this research are assumed to be applicable in any higher education setup. The higher education data collected are assumed to be accurately collected, reported, and updated to the website. The software used in this research study for data analytics is assumed to provide accurate results. The assumptions of a regression model are made in terms of linearity between dependent and independent variables, homoscedasticity, multivariate normality that assumes residuals are normally distributed, and lack of multicollinearity in the data. Due to the nature of the software, the results are assumed to be subject to the known reliability and validity of the software.

Statement of Limitations and Delimitations

This research study is subject to the following limitations and delimitations:

Limitations

1. Data analytics libraries and models may be impacted by computer system configuration and their ability to run large datasets on those computers.
2. This study does not include the effect of state budget cut or increase on the number of full-time faculty vs part-time faculty at public higher education institutions due to data being unavailable for some of the years.

Delimitations

1. The major delimitation of this study was that this study only selected full-time equivalent students at public higher education institutions in the US.
2. This study only considers public institutions that receive state funds for higher education.
3. This study analyzed the data of public higher education institutions in the US from 1971 to 2020.
4. This research did not investigate the impact of any specific year's budget cut or budget increase on the corresponding year of cohort.

Statement of Terminology

Data Analytics: Data analytics is the science of analyzing raw data to make conclusions about that information. Basically, it helps individuals and organizations make sense of data. Most of its algorithms could be automated and used over raw datasets to gain a deeper understanding of the data.

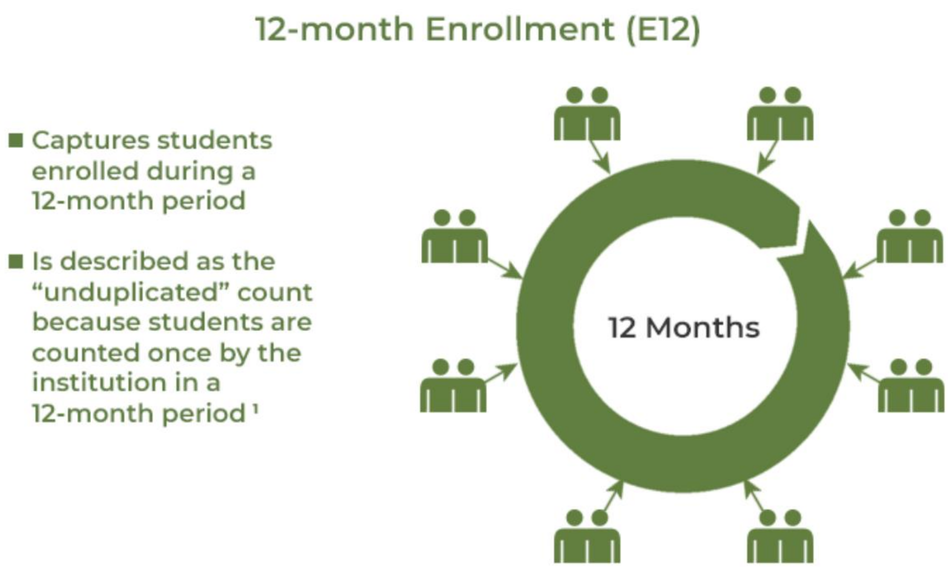
Algorithm: This algorithm can be defined as a step-by-step procedure that defines a set of instructions to be executed in a certain order to obtain the desired output.

Predictive Analytics: Predictive analytics helps answer questions about what will happen in the future. These techniques use historical data to identify trends and determine if they are likely to recur.

Full-Time Equivalent (FTE): The number of FTE students is calculated based on fall student headcounts as reported by the institution on the IPEDS Enrollment component. The full-time equivalent of students is a single value providing a meaningful combination of full-time and part-time students. Data products currently have two calculations of FTE students, using fall student headcounts and the other using 12-month instructional activity, as referred to in Fig. 2.

Figure 2

FTE 12-Month Enrollment



Note. The 12-month enrollment unduplicated headcount of students. Source: IPEDS, 2021

SHEEO: Refers to State Higher Education Executive Officers Association.

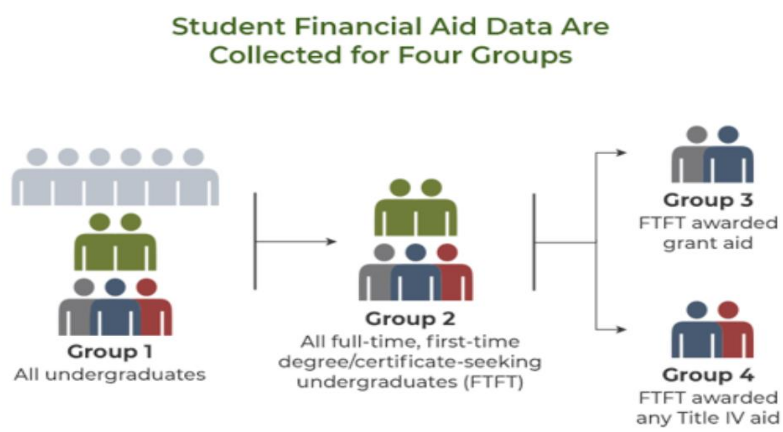
SHEF: State Higher Education Finance

General Operating Support: Direct funds given to institutions by the states. It is the portion of state and local support appropriated directly to public institutions for the purposes of general operations.

State Financial Aid (SFA): Allocations to state scholarships or other states financial aid for students attending public in-state institutions. Student financial aid data groups are shown in Fig. 3.

Figure 3

Student financial aid data groups



Note. Student financial aid data are collected for four groups. Source: IPEDS, 2021

State Higher Education Appropriations (SHEA): State and local support available for public higher education operating expenses. It does not include spending for medical education, research, and agriculture because they vary substantially across states, and excluding them helps in comparisons at the state level and per student level (National Center for Education Statistics, 2020).

Average Undergraduate Charge (AUGC): It is the average undergraduate cost to attend public 4-year institutions in the United States (National Center for Education Statistics, 2020). AUGC includes tuition cost and other costs like fees, room, and board rates.

Student Tuition Share (STS): The student share is the proportion of total education revenue at public institutions that comes from students and their families. It is calculated as a percentage of net tuition revenue over total education revenue (National Center for Education Statistics, 2020).

Student Tuition Share as a percentage of per capita income (STSPCI): It is the average annual cost per student to attend public 4-year institutions in the United States as a percentage of per capita personal income (National Center for Education Statistics, 2020). STS is only the cost paid for attending classes and does not include any other additional costs. While per capita income is the amount of money earned per person on an average in the US.

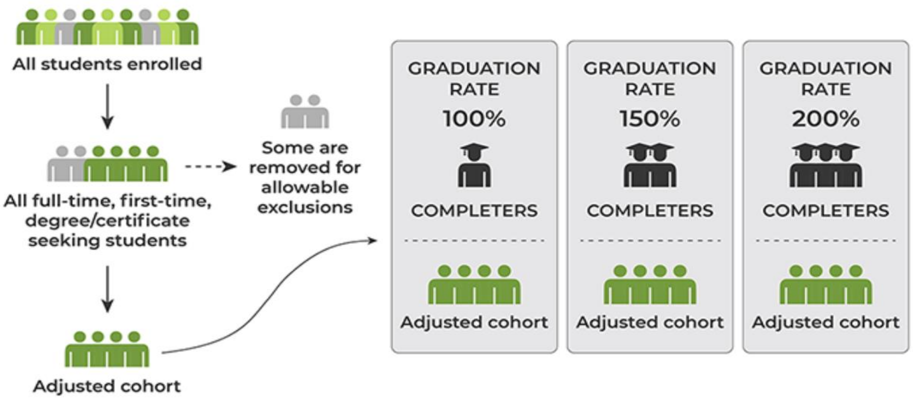
State higher education appropriations as a proportion of gross domestic product (SHEAGDP): This indicator represents the state higher education appropriations as a proportion of its gross domestic product (GDP). State higher education appropriations data were drawn from the annual state higher education finance report produced by SHEEO (National Center for Education Statistics, 2020).

Full-time Equivalent Enrollment (FTEE): Full-time equivalent enrollment is a calculation showing how many students would be attending the university if all were enrolled full time.

Graduation rate (GR): Graduation rate of all first-time, full-time enrolled students who completed their degree in 4 years - within 100% of normal time, 6 years - within 150% of normal time, and 8 years – within 200% of normal time. The explanation of the graduation rate is shown in Fig. 4.

Figure 4

IPEDS graduation rate



Note. IPEDS graduation rate data were collected for all undergraduate students who were first-time degree, full-time, and certificate-seeking students. Source: IPEDS, 2021

CHAPTER 2

REVIEW OF LITERATURE

Chapter Overview

This chapter provides a comprehensive explanation of the existing literature related to this research study. The systematic literature review study starts with a graphical flowchart of review criteria of prior literature followed by a discourse on how state budget cuts impact public universities in the US. The review of the literature is divided into two major sections to understand the impacts of state budget cuts on public institutions. The first section reviews the studies related to the effects of state appropriations on public institutions. The second section reviews the studies related to the effects of state financial aid on student enrollment and graduation rate. This study delves into a discussion about the effects of state appropriations on public institutions, the effects of state financial aid, education expenditures, institutional expenditures, and changes in tuition on graduation rate and enrollment numbers.

Background

The twenty-first-century challenges of producing more educated citizens and skilled labor remain a problem due to the lack of state investments in public higher education institutions. There are very limited research studies that have conducted data analytics and quantified the effects of state budget cuts on public institutions in the United States. Using the latest information technologies, such as data analytics, along with machine learning algorithms could provide deeper insights into data to better prepare public institutions in the US for future challenges. Especially after the COVID-19 pandemic, it became very important for public higher education institutions to analyze their historical data using the latest information technologies to prepare for future financial hurdles.

According to Johnson (2013), higher education institutions started using analytics to monitor student progress, recommend courses, predict student behavior, and predict student outcomes. He calls those techniques and methods of extracting data 'academic analytics'. The use of data technologies in higher education institutions could show unseen data relationships. Technology management plays a crucial role in promoting technologies such as data analytics to improve decision-making and face complex situations. Hoseini, Badar, Shahhosseini, and Kluse (2021) presented a review of the application of machine learning and data analytics in quality issues in the service, manufacturing, food, software and information technology, healthcare, and health insurance industries. Their study highlighted the importance of data analytics and data science to study vast amounts of datasets generated because of advancements in the technology field. Hoseini (2020) applied machine learning and data analytics to find fraud in Medicaid claims. He adopted machine learning algorithms and data analytical methods to predict fraudulent claims.

Data analytics could be used for operational efficiency in a variety of fields. It could help organizations make profits and allow companies to identify potential problems and eliminate them. A study conducted by Anggrahini, Kurniati, and Sukma (2021) used data analytics to obtain insights into supplier relationship management at a large-scale manufacturing company. The researcher’s study helped the manufacturing company assess supplier performance and responsiveness to produce high-quality products. The data analytical models and analysis helped the manufacturing company in strategic decision-making. Data analytics could benefit various industries: higher education, healthcare, supply chain, manufacturing, automotive, information technology, retail, banking and finance, construction, communications and entertainment, construction, energy, and many more. For example, a study by Summerfield, Zhang, Motiwalla, Mai, and Mazza (2018) showed that it could predict project management success by creating analytical models that could provide insights into crucial predictors such as cost overruns, overscheduling, employee utilization, and expected end date. Another study conducted by Parimi and Babu (2020) to explore and analyze various software security vulnerabilities at a large information technology enterprise in the cybersecurity domain observed a correlation between security levels and vulnerability scores.

Review Criteria

We systematically reviewed the related literature to assess the existing knowledge and gaps on state budget cuts and their impact on public institutions in the US to further develop the knowledge base. This systematic literature review helps in collecting related articles that answer the research questions. Previous research studies studying the effects of state funding cuts on public institutions using the latest information technologies, such as data analytics and machine learning algorithms, were not well developed. Significant advancements in data analytical techniques and machine learning algorithms along with higher-quality datasets allow for a more credible study on the effects of state budget cuts on public institutions in the US. The researcher performed a systematic literature review that allowed a collection of relevant evidence on state budget cuts' impact on public institutions using a literature review flowchart. The graphical flowchart in Fig. 5 shows the review criteria of existing literature on the topic that helped in performing a systematic literature review study.

In this research study, we focused on articles that study the relationship between state budget cuts and public higher education institutions in the US. This literature review study is primarily divided into two parts: one that shows the effects of state appropriations on public institutions and student enrollment and graduation rate and two that shows the effects of state financial aid on college enrollment and graduation rate. Hence, the primary focus was on papers that have state appropriations or state financial aid as their independent variables.

The inclusion criteria for this study for the search terms ‘State Budget cuts’, ‘Public Higher Education’, or ‘State Appropriations’ in the titles, keywords, or abstracts in three databases, EBSCO, JSTOR, and Google Scholar, in the English language resulted in N = 1692. This study did not include other databases, such as SCOPUS and Web of Science, because they are very exclusive. The researcher selected the three databases EBSCO, JSTOR, and Google Scholar, which include all publications, including anything indexed in SCOPUS and Web of Science. A publication in SCOPUS and Web of Science would most likely be indexed in Google Scholar, whereas a Google Scholar-indexed article may not find indexing in SCOPUS and Web of Science. This research study was filtered to have only studies that were published in 2000 or later, which resulted in the exclusion of 808 papers, N = 884. Later, this research study searched for studies that have either state appropriations or state financial aid as one of the independent variables, which resulted in the exclusion of 687 studies, N =197. The researcher performed a skim reading of the abstract and main body of the papers to find only experimental or quasi-experimental studies, which resulted in the exclusion of 81 papers (N=116). After 116 studies were identified, a critical appraisal tool was developed, as shown in Table 2, with a set of 10 questions to identify reliable studies that could answer the research questions. Finally, an appraisal of articles using the critical appraisal tool identified studies closely related to state budget cuts to public higher education in the US, excluding 99 papers, N=17. Hence, the final review criteria included 17 studies. The flow chart of the inclusion and exclusion criteria is shown in Figure 5.

Figure 5

Systematic Literature Review Graphical Flowchart

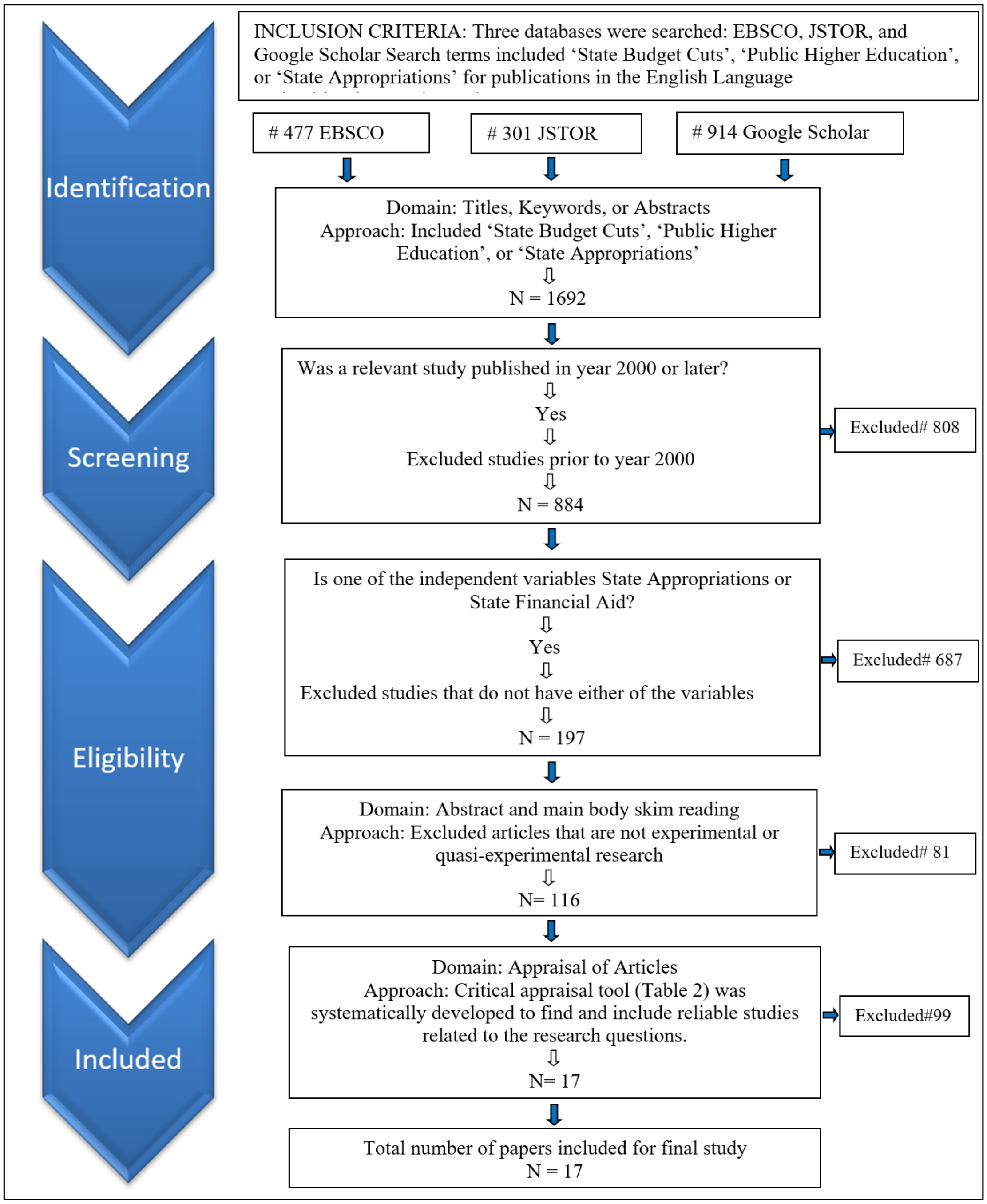


Table 1

Literature study selection with inclusion and exclusion criteria from the EBSCO, JSTOR, and Google Scholar databases.

CRITERIA	DECISION
Search terms included ‘State Budget Cuts’, ‘Public Higher Education’, or ‘State Appropriations’ in the titles, keywords, or abstracts.	Inclusion
Papers written in the English language	Inclusion
Studies conducted outside of the United States	Exclusion
Consider studies published between 2000-2021 and exclude prior studies	Exclusion
Studies that do not have independent variables of State Appropriations or State Financial Aid?	Exclusion
Articles that are not experimental or quasi-experimental research	Exclusion
Appraisal of articles related to the research questions was conducted and articles were screened based on critical appraisal tool	Exclusion
Number of papers that are included in final study N = 17	

The inclusion and exclusion criteria of articles were based on Table 1, and an appraisal of articles was conducted for 116 articles to answer the research questions that allowed the researcher to identify reliable, methodologically appropriate, and unbiased research studies. The quality of the study could be analytically evaluated using critical appraisal tools, and they help minimize biases in a research study (Katrak, Bialocerkowski, Massy-Westropp, Kumar, &

Grimmer, 2004). The researcher could not find a critical appraisal tool specific to this research; hence, the researcher developed it, as seen in Table 2. The articles were divided based on the research design, and the details are shown in Table 3. Additionally, the total number of articles is presented as a bar chart based on the year of publication, which can be seen in Figure 6. From a total of 116 research studies, 17 articles that were closely related to state budget cuts to public higher education institutions in the United States were included for final review after conducting an appraisal using the critical appraisal tool from Table 2. Hence, the number of studies included for final review was N=17.

Table 2

Critical Appraisal Tool

		Yes	No	Unclear	Not applicable
1.	Were the research questions and method clearly stated?	☐	☐	☐	☐
2.	Was the need for the research study adequately established?	☐	☐	☐	☐
3.	Were the statistical analysis methods described in detail?	☐	☐	☐	☐
4.	Was the study design appropriate for the research question? *Note:* Appropriate study design explains clearly how to collect, analyze, and interpret data to provide an answer to the question.	☐	☐	☐	☐
5.	Was the study limited to only public higher education institutions?	☐	☐	☐	☐
6.	Was the data collected and used for analysis address the research question?	☐	☐	☐	☐
7.	Was appropriate statistical analysis used for the study?	☐	☐	☐	☐

Note: Appropriate statistical analysis are studies that are linear, ordinal, or multinomial regressions.

8. Was the data extracted from reliable sources? ☐ ☐ ☐ ☐

Note: Reliable sources are institutions that are approved by the US Department of Education to publish postsecondary data.

9. Were multiple variables used in the analyses of the research study? ☐ ☐ ☐ ☐

10. Were specific directions for new research initiatives proposed? ☐ ☐ ☐ ☐

Table 3

Research design and number of studies in each design

Research Design	**Count of Research Design**
Difference-in-differences	47
Regression Discontinuity	36
Randomized Control Trial	8
Difference-in-differences, Regression Discontinuity	6
Instrumental variables estimation	6
Fixed effects panel model	5
Event History Analysis	3
Instrumental Variables	2
Dynamic fixed effects panel model	1
Two-way Fixed Effects	1
Instrumental variables estimation & Fixed effects panel model	1
Grand Total	**116**

Figure 6

Number of articles appraised by publication year

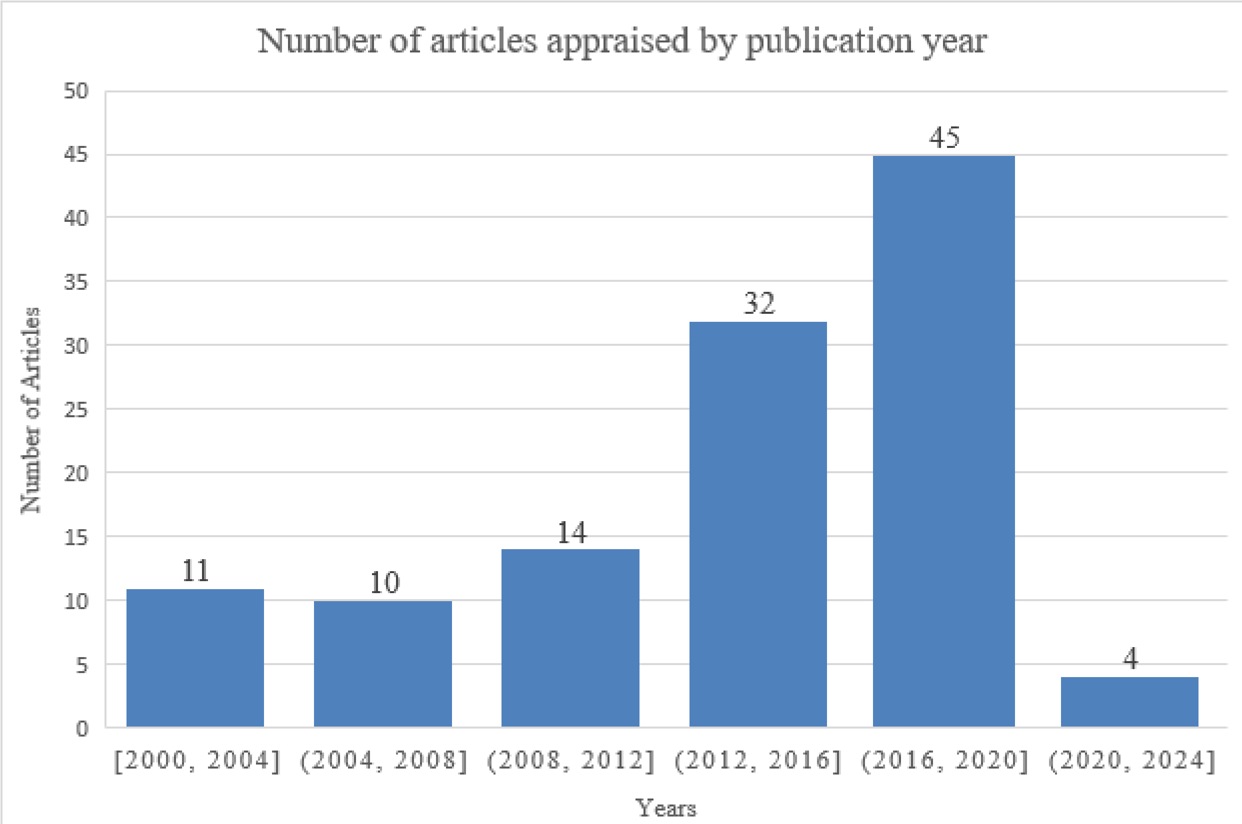


Table 4a

Intra-rater reliability test

	Time 2: Yes	Time 2: No	Time 2: Unclear	Time 2: N/A	
Time 1: Yes	15	2	0	0	17
Time 1: No	2	94	0	0	96
Time 1: Unclear	0	1	1	0	2
Time 1: N/A	0	0	0	1	1
	17	97	1	1	116

Table 4b

Intra-rater reliability test

	Time 2: Yes	Time 2: No	Time 2: Unclear	Time 2: N/A
Time 1: Yes	2.4913793			
Time 1: No		80.275862		
Time 1: Unclear			0.017241379	
Time 1: N/A				0.00862069

Table 4c

Intra-rater reliability test

Total Observed	111
Total expected for the agreement	82.7931
Grand Total	116

Cohen’s Kappa Score = 0.849428868

An intrarater reliability test (Cohen's kappa coefficient - κ) was calculated with a gap of three weeks to determine if there was an agreement between 'article appraisal time 1' and 'article appraisal time 2'. The score could range from -1 to +1. The calculation of Cohen's kappa test (κ) resulted in a score of 0.849, which means there was an 85% measure of agreement between the appraisal study from time 1 and time 2. The interrater reliability (κ) test score suggests that there was a very significant strength of agreement between ‘time 1’ and ‘time 2’.

Prior Literature on the Effects of State Appropriations on Public Institutions

Literature related to the effects of state appropriations on tuition cost, institutional expenditure, research activity, enrollment, and graduation rate was reviewed. According to Delaney and Doyle (2011), state appropriations at public higher education institutions have been decreasing significantly compared with other budget categories for the past two decades. Public institutions, in response, either increase tuition or decrease institutional expenditure when state appropriations decline. A Delta Cost Project for a dataset from academic years 2000 to 2010 conducted to examine the impact of declining state appropriations on domestic students (In-state tuition) showed that a 10% reduction in state appropriations resulted in a 1.1% increase in tuition and fees and a 0.7% increase in the total yearly cost of college education (Goodman and Volz 2020). According to Webber (2017), an additional $257 of tuition and fees on average was required for a $1000 reduction in state appropriations per FTE student. He also mentions that most public institutions greatly rely on tuition revenue to avoid disruption to their core functions when there are state budget cuts. Hyman (2017) study found that there is a direct correlation between state appropriations and student outcomes. The author recommended the state governments to consider the findings when preparing or modifying education policies.

With each recession since 1980, the state support per FTE student has declined at a greater rate, and recovery has become slower. In 2001, public institutions enrolled 8.7 million students, and per FTE, students received $9,547 as a general operating expenditure. However, in 2019, public institutions were only provided $7,388 per FTE student for a total of 10.9 million enrolled students (Laderman, & Tandberg, 2021). Institutions respond to these changes by making changes to their spending categories. The most affected categories are academic support, student services, and instructional spending (Deming and Walters 2018). As per the study conducted by Frye (2015), declining state funding has shown a negative impact on the academic workforce at public institutions. A regression model conducted using IPEDS data from 1994 to 2013 has shown that when there are state budget cuts, public higher education institutions respond by decreasing the number of tenure track faculty and increasing part-time faculty. For every 10% decrease in the state, appropriations resulted in a 0.23% increase in part-time faculty. Husted and Kenny's (2018) study provided evidence that state appropriations could also negatively impact public research universities in conducting research activities. Their analysis of 152 public universities showed a positive relationship between state appropriations and research productivity. A reduction of 10% of state appropriations reduces the number of patents awarded by 8.4%.

A study conducted by Goodman and Voltz (2020) found that a 10% decrease in state appropriations resulted in a 3% decrease in enrollments at public institutions. The IPEDS enrollment data from 1900 to 2013 found that there was a positive relationship between the increase in the state budget to public institutions on current and future student enrollments. These studies present an issue that might arise due to declining state appropriations; that is, they may enroll less underrepresented students, as they may not have the ability to pay most tuition (Jaquette and Curs, 2015). According to Rothstein and Schanzenbach (2021), courts in many states found that state higher education resources matter and that states are constitutionally required to provide schools with funds. Students are impacted by the changes at higher education institutions, most importantly the way they are funded. Stress is one of the factors reported by students based on the financial burden laid on them (Robotham & Julian, 2006).

Previous studies find significant evidence of the impact of state budget cuts on student outcomes. The student outcomes are a combination of the graduation rate and the number of credentials awarded. Zhang (2009) conducted a research study to find a correlation between the graduation rate at public four-year institutions and state appropriations. He found that a 10% increase in state appropriation per FTE resulted in a 0.75% increase in graduation rate, and simultaneously, a 10% decrease in state funding resulted in a 0.56% decrease in graduation rate. This shows that prior studies identified a significant relationship between state appropriations and graduation rate.

Prior Literature on the Effects of State Financial Aid

Literature related to the effects of state financial aid on college enrollment, college persistence, and graduation rate was reviewed. Postsecondary investments made by students in higher education became a concern as state support declined steadily (Chakrabarti et al., 2020). There was no significant evidence suggesting that state financial aid influences overall college enrollment. It only helped students attend more costly universities where they could receive aid, and financial aid packages that are less than realized packages significantly impact a student's likelihood of enrolling (Laderman, & Tandberg, 2021; Dynarski, 2003). However, the effects of financial aid, state education appropriations, student tuition cost, and institutional expenditure on persistence to stay in college are evident. According to Avery et al. (2014), suggest that the quality and quantity of college counseling on different aspects, including the financial aid application process, could impact college enrollment. Most studies have suggested that financial aid helps students graduate at higher rates than nonaided students. However some studies found no significant relationship between financial aid and student enrollment (Bruce & Carruthers, 2014; Gurantz & Odle, 2021).

Anderson and Zaber (2021) studied the potential effects of financial aid on students and their decision to enroll in college. Their study shows that most students are enrolled in a college by the time they apply for financial aid, and they are at a point in deciding whether they should keep going forward and pay for college or not. Students benefit greatly by discussing the potential options for financial aid and expenses related to college, as they could be very complex to understand (Oreopoulos & Petronijevic, 2013). Students entering college for the first time find it hard to see or predict what it takes to complete a college degree until they receive financial aid (Deming and Dynarski, 2010).

In summary, a decrease in state education appropriations results in an increase in tuition rates at public four-year institutions and enrolls more out-of-state and international students. It also negatively impacts expenditures on instruction, academic support, and student services at these institutions (Ehrenberg, 2004). There was a concern with a decrease in in-state undergraduate students enrolled at public institutions and, finally, a decrease in degrees and certificates awarded. Another study finds that funding to students in the form of financial aid was highly positively correlated with college enrollment (Deming and Walters 2017). According to Shin (2010), US institutions did not experience much change in their institutional performance by adapting to a performance-based funding model. The performance-based funding model is a model that proposes more funding to institutions that have higher graduation rate. According to Monarrez et al. (2022), state spending decisions related to state financial aid programs have a great impact on public institutions. If public institutions' funds are cut by the states, then it results in a decline in enrollment and the number of degrees awarded. A study conducted to identify the causal effect of financial aid on students’ persistence toward degree attainment showed a positive correlation if the aid is awarded as a need-based aid (Alon 2011). According to Stevenson (2006), poverty could play a significant role and be an important factor in terms of student performance.

Despite concerns expressed by policymakers and scholars that the declines in state support have reduced the return to education investment for public sector students, little evidence exists that can identify the causal effect of these funds on long-run student outcomes.

Chapter Summary

After conducting a systematic literature review, the researcher found that the literature related to the research study examined several impacts of state budget cuts on higher education. However, no single data analytical and machine learning model has been developed to study the impacts of state budget cuts on FTE enrollment (FTEE) and graduation rate (GR). The two dependent variables of FTEE and GR are crucial factors for the survival of 4-year public higher education institutions in the US. Hence, the researchers developed a multiple linear regression model using data analytics and machine learning techniques to find the predictive relationship between the independent variables and the dependent variable.

CHAPTER 3

METHODOLOGY

Chapter Overview

This chapter reviewed the problem statement, research questions and hypotheses, data collection methods and procedures, and research design and procedures. This chapter concludes with findings from the preliminary study and a summary of the chapter.

Statement of the problem

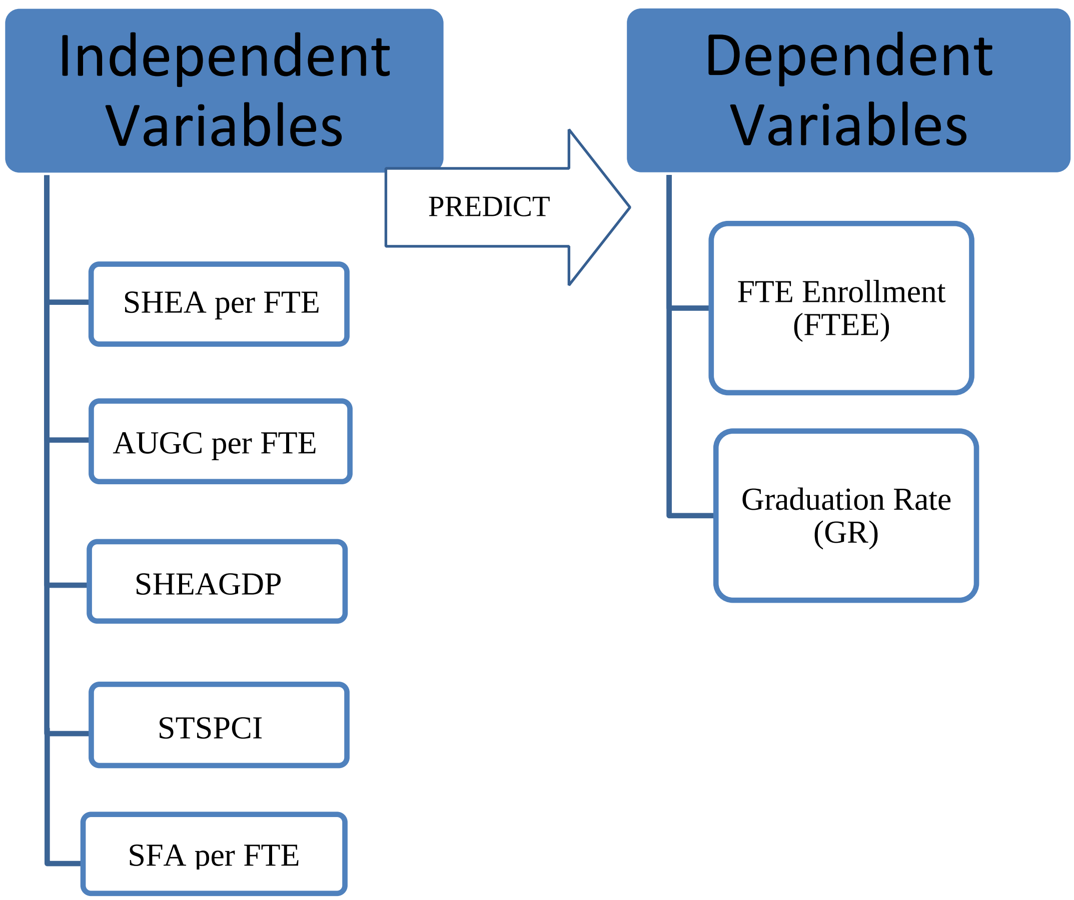
State budget cuts to public higher education institutions in the US had an increasing trend for several years, and it is expected that it could become even worse if the trend continues. Higher education institutions play a crucial role in providing the workforce with skill sets to face a rapidly changing economy and global competition. This requires states to invest in public higher education institutions. However, the current trend is the opposite. Hence, this study aims to leverage data analytics to study the impact of state budget cuts on 4-year public higher education institutions.

Research Design

A quantitative research methodology was used to investigate the effects of state budget cuts on 4-year public higher education institutions in the United States. The researcher's objective was to find a relationship, if any, between declining state appropriations and student FTE enrollment and graduation rate at public institutions using data analytics. The variables of the study are summarized in Fig. 7.

Figure 7

Variables used in this research study



Note. The independent and dependent variables used in this research study are presented in the figure.

The study examined whether there exists any significant predictive relationship between the five independent variables (IVs): state higher education appropriations per FTE (SHEA), average undergraduate charges per FTE (AUGC), student tuition share as a percentage of per capita income (STSPCI), state higher education appropriations as a percentage of GDP

(SHEAGDP), and state financial aid (SFA) per FTE, and dependent variables (DV): full-time equivalent enrollment (FTEE) and graduation rate (GR) at public higher education institutions. Similarly, the following two null hypotheses were tested.

H_{01}: There is no significant predictive relationship between the five independent variables (IVs): state higher education appropriations per FTE (SHEA), average undergraduate charges per FTE (AUGC), student tuition share as a percentage of per capita income (STSPCI), state higher education appropriations as a percentage of GDP (SHEAGDP), and state financial aid (SFA) per FTE, and dependent variable (DV): full-time equivalent enrollment (FTEE) at public higher education institutions.

$$H_{01}: \beta_{SHEA} = \beta_{AUGC} = \beta_{STSPCI} = \beta_{SHEAGDP} = \beta_{SFA} = 0 \tag{1}$$

$$H_{A1}: \text{At least one in } (\beta_{SHEA}, \beta_{AUGC}, \beta_{STSPCI}, \beta_{SHEAGDP}, \beta_{SFA}) \neq 0 \tag{2}$$

H_{02}: There is no significant predictive relationship between IVs: state higher education appropriations per FTE (SHEA), average undergraduate charges per FTE (AUGC), student tuition share as a percentage of per capita income (STSPCI), state higher education appropriations as a percentage of GDP (SHEAGDP), and state financial aid (SFA) per FTE, and DV: graduation rate (GR) at public higher education institutions.

$$H_{02}: \beta_{SHEA} = \beta_{AUGC} = \beta_{STSPCI} = \beta_{SHEAGDP} = \beta_{SFA} = 0 \tag{3}$$

$$H_{A2}: \text{At least one in } (\beta_{SHEA}, \beta_{AUGC}, \beta_{STSPCI}, \beta_{SHEAGDP}, \beta_{SFA}) \neq 0 \tag{4}$$

The above hypotheses H_{01} *and* H_{02} from equations (1) and (3) were tested using multiple linear regression (Montgomery, 2012). Generally, the dependent variable y may be related to n independent variables. Multiple linear regression was used to find correlations between independent variables and dependent variables and to determine a regression model. Here, the regression model from equation (5) is:

$$y = \beta_0 + \beta_1 \times x_1 + \beta_2 \times x_2 + \beta_3 \times x_3 + \beta_4 \times x_4 + \beta_5 \times x_5 \quad (5)$$

where y is a dependent variable, β_0 is the y-intercept, and x_1, x_2, x_3, x_4, x_5 are multiple independent variables (GeeksforGeeks, 2021; Rich, Shahhosseini, Badar, & Kluse, n.d.).

The models in this multiple linear regression study for FTEE and GR were fit using ordinary least squares (OLS). OLS tries to minimize the sum of squares of the differences between actual and predicted values and estimate the unknown values from the given dataset. The differences between actual and predicted values for the training dataset are called residuals, and for the test set, they are called prediction errors. After the regression model was trained and tested, the model was evaluated based on the R^2 score, root mean squared error, mean absolute error, and mean squared error. The primary goal of prediction model studies was to generalize beyond the examples in the training dataset irrespective of the variety of data in the training set (Killada, 2017). The regression analysis was performed using an open-source software programming language called Python, which is commonly used in data analytics and machine learning. This was explained further in the subsequent section titled the research procedure.

Data Gathering Procedure

The data needed in this study were collected from the State Higher Education Executive Officers Association (SHEEO) website and Integrated Postsecondary Education Data System (IPEDS) website, and they are publicly available. These publicly available data were retrieved directly from the respective SHEEO and IPEDS websites. The data required for this study were downloaded as multiple small datasets starting from 1971 to 2020. Later, the researcher combined all the small datasets collected into a single large dataset. Then, the data were cleaned and enriched to avoid having unwanted or inaccurate data from the dataset, which makes it more valuable. The missing values from the dataset were mean imputed for the variables highlighted in yellow as shown in Appendix P. Two values are missing in SHEA, 1 value in STSPCI, 2 values in SHEAGDP, and 2 values in SFA. They all are replaced by the mean values.

Research Procedure

The data were analyzed using a multiple linear regression model, data analytics libraries, and machine learning algorithms to establish the effect of state budget cuts on public higher education institutions. For research purposes, data analytics is an area focused on providing patterns and meaning from data by performing robust analyses on datasets. Data analytics libraries are a huge collection of programming codes that are open to the public. They could be used in Python programming to analyze the research data, which could help the researcher avoid writing the programming code from scratch. They significantly reduce the time required by the researcher to code in Python. The various data analytics libraries used in this research study are shown in Table 4. The regression analysis and analytical data models were operated by an open-source software programming language called Python programming. Python software version 3.8 was used in this research project. The researcher selected Python programming over traditional statistical software because it is a free software that runs on code compared to statistical software that is licensed and expensive. The greatest advantage of using Python over other software programs, such as SPSS or Minitab, is that the data analytical model built in Python can be automated and shared with the public.

This research study attempts to analyze data to identify trends, correlations, and insights from data. Data analytics algorithms are a set of calculations that would be deployed to uncover the historical unknown trends in the data. Python, with the help of data analytical libraries and algorithms, could provide better insights from datasets because of its built-in data analytics tools that could identify trends and correlations and produce interactive dashboards. Finally, data analytics modeling was conducted to examine the hypotheses. This methodology was accomplished through the analytical data model shown in Fig. 8.

Python Vs Other Statistical Tools

There are several reasons to choose Python programming over other statistical tools, such as SPSS, Minitab, SAS, and R. One of the most important reasons is that Python is the fastest-growing software in the world today that is completely free to use. Some world-class technology organizations that use Python programming as their official programming language are Google, Netflix, Facebook, NASA, IBM, Intel, JP Morgan Chase, Quora, and Spotify. Today, it has even become an introductory programming language in many reputable universities worldwide (BrainStation, 2021). Advantages that Python has over other statistical software are that it could handle large datasets at ease, and it could also provide interactive graphs that could be customized and optimized as per user needs compared to the traditional graphs and charts that other tools offer. Python libraries make the tool even more powerful in crunching data.

The researcher also reviewed various Python software libraries and data modeling applications required for this study, which are easy to use and completely free. Python libraries and their functionalities are shown in Table 4. For example, Pandas is a software library written for the Python programming language for data manipulation and analysis. It helps slice the data frame and change the index, conversion, concatenation, joining, and merge data. Pandas have simple, powerful, and efficient functionality for performing resampling operations during frequency conversion. With the help of other libraries in Python, pandas can be used to plot series objects or data frames. NumPy is the core library for scientific computing in Python. It provides a high-performance multidimensional array object and tools for working with these arrays. It only works with numbers and helps in processing and storing large datasets at a swift pace. Matplotlib is a Python package used for 2D graphs such as bar graphs, histograms, scatterplots, pie plots, and area plots. Plotly is an advanced data visualization library used for 3D graphs, interactive graphs, and heatmaps with colored cells to represent data. (Badole, 2021).

Table 5

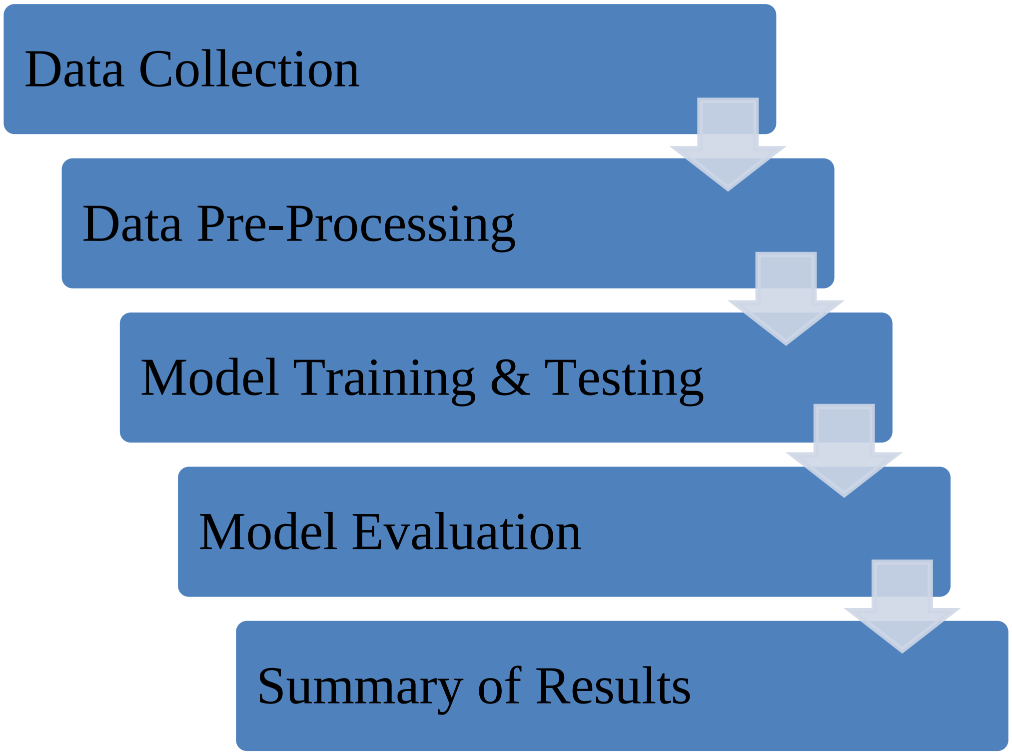
Python Libraries and their functionalities

Library Name	Functionality
Pandas	Data Manipulation and Analysis
NumPy and SciPy	Fundamental Scientific Computing
Matplotlib	Visualization and Plotting
Scikit-learn	Machine Learning and Data Mining
Seaborn	For Statistical Data Visualization
Stats Models	Statistical Modeling, Testing, and Analysis
Plotly	Advanced Data Visualization

This dissertation study, with the help of data analytics tools and methods, focused on finding the impacts of state budget cuts on public institutions. To achieve that, the researcher developed a five-step data analytics model and research approach to collect, process, train, test, evaluate and summarize the results, as shown in Fig. 8.

Figure 8

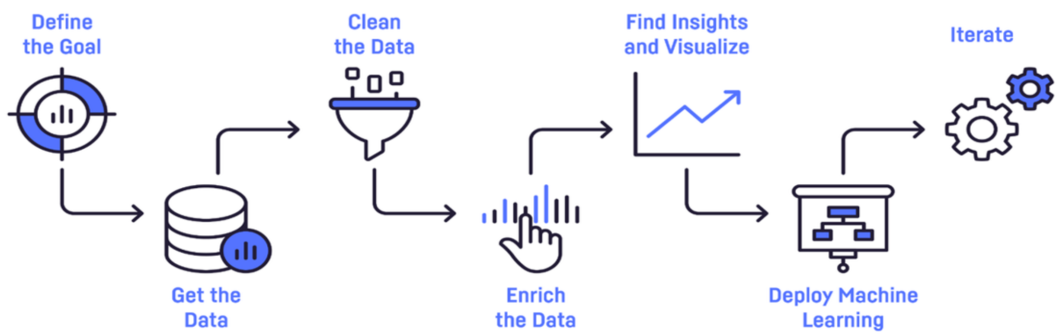
Five-Step Data Analytics Model and Research Approach



The data analytics model presented in Fig. 8 involves seven steps to work on the model, as shown in Fig. 9, starting with defining the goal, obtaining the data, cleaning it up, enriching the data, finding insights and visualization, deploying machine learning algorithms, and iterating. The data analyzed through the multiple linear regression model in Python starts with preprocessing the data, which involves importing the libraries, importing the dataset, dummy variable encoding, and splitting the dataset into a training set and test set. After that, fitting multiple linear regression to the training set was performed, and the test set results were predicted and reported to determine the accuracy of the model.

Figure 9

Seven fundamental steps to complete a Data Analytics Project

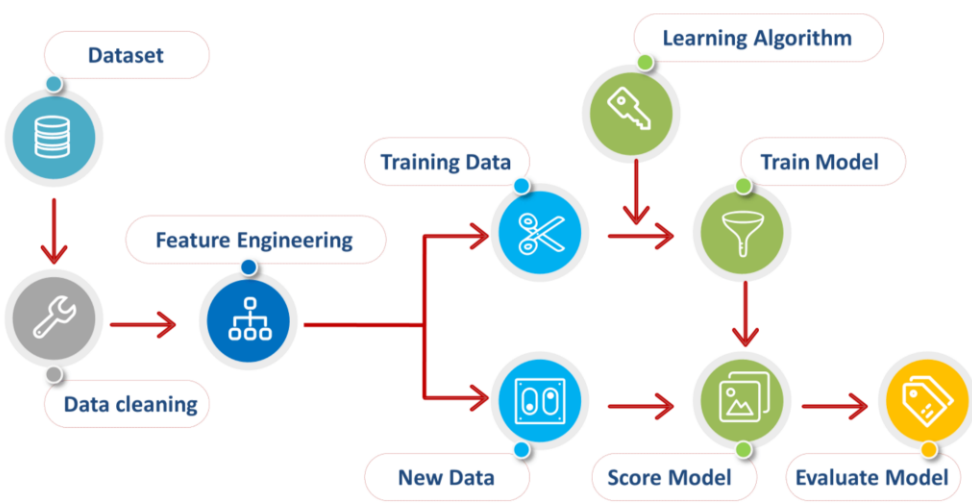


Note: Seven fundamental steps to complete a Data Analytics Project. Source: Smith, 2019

Figure 10 shows the in-depth process of how the datasets are processed in Python. This flowchart helps to understand how the process flows in a data analytical model in Python.

Figure 10

Data Processing in Python



Note: Data pprocessing in Python. Source: Chakure, 2021

Chapter Summary

In summary, this data analytics research project's objective is to develop a model for FTEE and GR to find the impacts of state budget cuts on public higher education institutions in the US. A set of two research questions and two related hypotheses lead the project to its completion. The five-step data analytics model, the seven steps of the data analytical model, and the data processing steps involved in completing the project in Fig. 8, Fig. 9, and Fig. 10, respectively, served as a foundation that helped in answering the research questions and testing the hypotheses.

CHAPTER 4

RESULTS

The purpose of this study was to determine whether and to what extent student FTE enrollment and graduation rate at public higher education institutions in the US are impacted by, and if yes, can be predicted by various components related to state budget cuts to higher education. State budget funds are an essential source of revenue for public universities in the US. Suppose that public higher education administrators and policymakers can ascertain the variables that impact student FTE enrollment (FTEE) and the graduation rate (GR) at public higher education institutions. In that case, they could make data-driven decisions and policies that could affect the state budget cuts to public higher education.

The first hypothesis tested the association between the predictive relationship between independent variables (IVs) related to state budget cuts and the dependent variable: (DV) FTE enrollment (FTEE) of 4-year undergraduate students at public universities. The second hypothesis tested the association between independent variables (IVs) related to state budget cuts and the dependent variable (DV) graduation rate (GR) of 4-year undergraduate students at public universities. The results of the study are presented in this chapter. This study was guided by two main research questions from chapter 1 under the subheading research questions. The two main research questions guiding this chapter are as follows:

Note: FTE is full-time equivalent.

RQ1. Is there a significant predictive relationship between five independent variables (IVs): state higher education appropriations per FTE (SHEA), average undergraduate charges per FTE (AUGC), student tuition share as a percentage of per capita income (STSPCI), state higher education appropriations as a percentage of GDP (SHEAGDP), and state financial aid (SFA) per FTE, and dependent variable (DV): full-time equivalent enrollment (FTEE) at public higher education institutions?

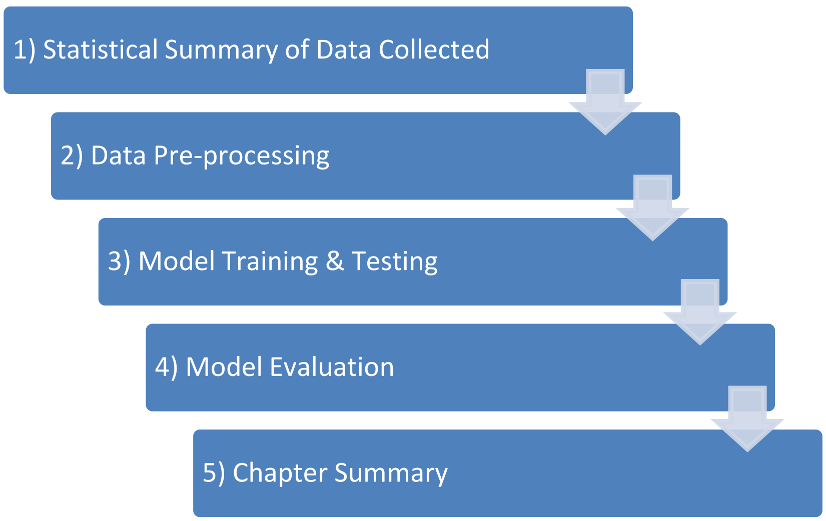
RQ2. Is there a significant predictive relationship between IVs: state higher education appropriations per FTE (SHEA), average undergraduate charges per FTE (AUGC), student tuition share as a percentage of per capita income (STSPCI), state higher education appropriations as a percentage of GDP (SHEAGDP), and state financial aid (SFA) per FTE, and DV: graduation rate (GR) at public higher education institutions?

This data analytics model was used to predict the relationship between outcome variables (DV): FTE enrollment (FTEE) and graduation rate (GR) from the five independent variables (IVs). The study developed a data analytics model that was trained to predict the known data initially and later tested using test data and applied to generalize other nontrained data. Test data are used to test the prediction ability, which also considers the accuracy of the model. The training data of independent and dependent variables are used to fit the regression model to make a linear model.

This chapter was primarily divided into five major sections, and each of the sections explains the implementation of the data analytics algorithms step by step. Since there are two research questions and two hypotheses, the algorithm steps were repeated separately for both dependent variable (DV) models. This means that the steps completed to determine whether there is any relationship between dependent variable (DV): FTE enrollment (FTEE) and independent variables (IVs) were again repeated as a separate model for dependent variable (DV): graduation rate (GR). Figure 11 shows the five major sections of this chapter:

Figure 11

Five major sections of the results section



The statistical summary of data collected for all variables, including means and standard deviations, is presented and discussed. The data preprocessing step includes importing the libraries, importing the dataset, feature scaling, handling the missing data, and finally splitting the dataset into training and test datasets. In the next section, the researcher, with the help of a wide range of data analytical algorithms, trained and tested linear regression models using the dataset. A model evaluation process was conducted to ensure that the best prediction model was selected. Finally, a summary of the multiple linear regression analysis results was presented for the models. Additionally, the results are discussed in detail in this section.

The data required for this study were collected and stored as a comma-separated values (CSV) file. Once the data were collected, they were analyzed using an open-source programming software called Python 3.9.7 version. The Python programming code written for this research project was run in a web-based application called Jupyter Notebook. Jupyter Notebook is an open-source web application that allows the creation and execution of Python code. Jupyter Notebook provided a platform to leverage data analytical tools, including live code, visualizations, and statistical computational output. It helped to explore the data with Python libraries such as pandas, scikit learn, NumPy, statsmodels, plotly, ggplot2, and more. This research study used the pandas library module to load comma-delimited (CSV) files, the NumPy module to convert the data into arrays, and scikit_learn to compute multiple linear regression.

Statistical Summary of Data Collected

The data collection process was conducted through several steps, and the data required for this study were collected from publicly available sources. The target population in this study is 4-year undergraduate students at public universities in the US. The dataset compiled in this study shows the impact of state budget cuts and their components on student FTE enrollment (FTEE) and the graduation rate (GR) of 4-year undergraduate students at public universities. The total number of years of data selected for the study is 50 from 1971 to 2020 (N = 50). Table 6 presents the information about the dataset, Table 7 shows a statistical summary of the research study, and Table 8 presents the correlation matrix of independent and dependent variables. The independent and dependent variables are coded as shown in Table 6 for statistical analysis:

Table 6

Summary of Variables for statistical analysis

Variable Code	Name of the Variable	Type of Variable	Measurement
SHEA	State Higher Education Appropriations per FTE	Independent	$ (Dollars)
AUGC	Average Undergraduate Charges per FTE	Independent	$ (Dollars)
STSPCI	Student Tuition Share as a percentage of Per-Capita Income	Independent	Percentage
SHEAGDP	State Higher Education Appropriations as a percentage of GDP	Independent	Percentage
SFA	State Financial Aid per FTE	Independent	$ (Dollars)
FTEE	Fulltime Equivalent Enrollment	Dependent	Number of Students
GR	Graduation rate	Dependent	Percentage

The original datasets had data of all 50 states in the US from 1971 to 2020 as separate files. Later, the datasets were combined to obtain the US average values. Detailed information on the coded variables is presented in Table 7.

Table 7

Dataset Information

#	Column	Non-Null Count	Dtype
0	SHEA	50 non-null	float64
1	AUGC	50 non-null	float64
2	STSPCI	50 non-null	float64
3	SHEAGDP	50 non-null	float64
4	SFA	50 non-null	float64
5	FTEE	50 non-null	float64
6	GR	50 non-null	float64

The descriptive statistics summary of the dataset shows the count, mean, standard deviation, quartiles, and minimum and maximum of the variables. The count in the data are the annual data of 50 years, i.e., from 1971 to 2020. 25, 50, and 75 from Table 8 represent the interquartile range starting with the 25^{th} percentile (first quartile), 50^{th} percentile (second quartile), and 75^{th} percentile (third quartile), respectively.

Table 8

Statistical Summary of Data Collected

	SHEA	**AUGC**	**STSPCI**	**SHEAGDP**	**SFA**	**FTEE**	**GR**
Count	50	50	50	50	50	50	50
Mean	5810	9459	37	48	548	148351	51
Std	806	5343	5	7	294	43664	5
Min	4494	4171	30	41	239	86126	40
25	5345	5417	33	42	274	107346	47
50	5657	6952	36	43	420	142956	52
75	6351	13099	41	53	813	184052	54
Max	7928	20698	46	62	1278	220022	61

Table 9

Correlation matrix

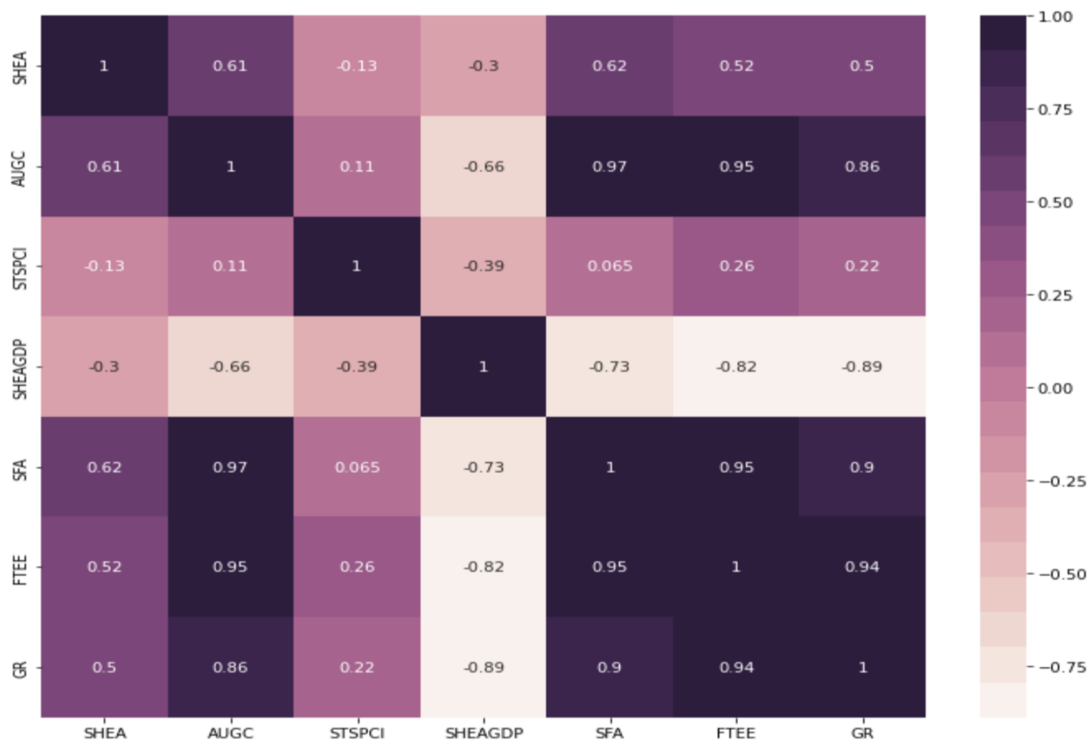
Independent Variables Dependent Variables

	SHEA	AUGC	STSPCI	SHEAGDP	SFA	FTEE	GR
SHEA	1.00	0.61	-0.13	-0.30	0.62	0.52	0.50
AUGC	0.61	1.00	0.11	-0.66	0.97	0.95	0.86
STSPCI	-0.13	0.11	1.00	-0.39	0.06	0.26	0.22
SHEAGDP	-0.30	-0.66	-0.39	1.00	-0.73	-0.82	-0.89
SFA	0.62	0.97	0.06	-0.73	1.00	0.95	0.90
FTEE	0.52	0.95	0.26	-0.82	0.95	1.00	0.94
GR	0.50	0.86	0.22	-0.89	0.90	0.94	1.00

The correlation analysis matrix in table 9 examined the degree of relationship between all the variables. The correlation coefficients between independent variables showed a moderate to low significant relationship between the independent variables except AUGC and SFA, which showed a statistically significant correlation. There was a statistically significant relationship between most of the independent variables and dependent variables. This analysis was conducted to understand the dependency between independent continuous variables. This process helped to identify and select significant and nonredundant variables. The below correlational matrix heatmap in Fig. 12 shows a two-dimensional correlation between two discrete dimensions with the help of colored cells to represent data. The plot shows a 7 × 7 matrix and color cells based on the correlation coefficient of the variables.

Figure 12

Correlation Matrix Heatmap



Note: The correlation between two discrete dimensions is explained by colored cells based on the scale shown in the figure.

Data Preprocessing

In this section, data preprocessing steps were used to transform the data into a suitable format for the data analytical process. Data preprocessing is basically a data mining technique that performs data quality assessment, data cleaning, data integration, data transformation, and data reduction. The steps used for the application of data preprocessing techniques using Python programming are as follows:

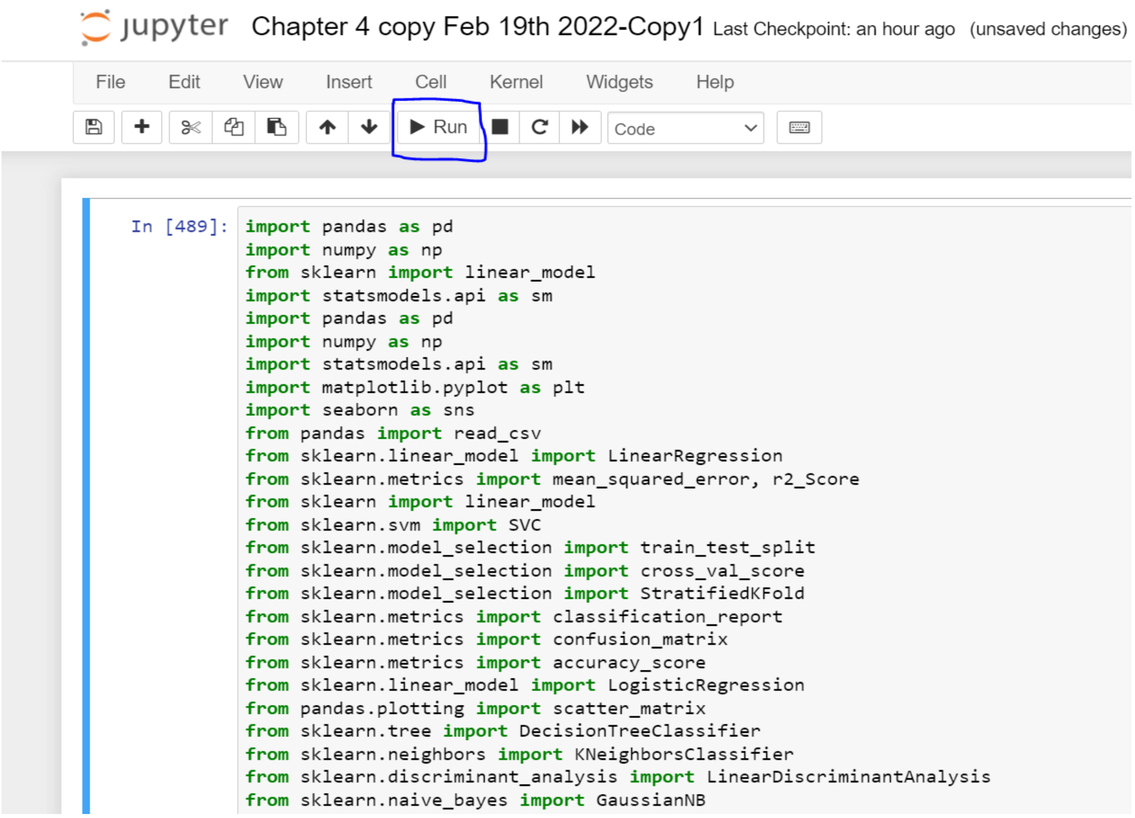
- Importing the libraries
- Importing the dataset
- Feature scaling
- Handling missing data
- Splitting the dataset into training and test datasets

Creating a multiple regression model requires importing the required libraries. Hence, various libraries imported and used in this study along with their functionalities are available in Table 5. Below is the Python code for importing the libraries required in this research study.

Importing the libraries

Figure 13

Importing the required libraries



The Python code from Fig. 13 was selected and run to import the libraries required for this study. As marked in Fig. 13, the Python code entered in the console was selected first, and then the run command was executed. This step imported all the required libraries needed in this study.

Importing the dataset

The dataset used for this study was imported into Python using different Python libraries. The following commands in Fig. 14 were executed to import the dataset from the data source CSV file, followed by two quality check commands to ensure that the dataset was imported correctly.

Figure 14

Importing the dataset

```
import plotly.express as px
import plotly.graph_objects as go
from sklearn.linear_model import LinearRegression
df = pd.read_csv (r'C:\Users\guraj\Downloads\bookpython2.csv')
df.head(5).round(2)
```

Feature scaling

In machine learning feature scaling is one of the most important steps during data preprocessing before creating a data analytics model. This standardization step makes a weak machine learning model a better one by transforming the data to have zero mean and variance of 1. StanadardScaler function under feature scaling process standardizes the input variables from the dataset. It scales each feature in such a way that the distribution is centered around 0, with a standard deviation of 1. That means 68% of the values will fall between -1 and 1. (Roy, 2021). This scaling feature works best when data is normally distributed and this step in the research study brings all the variables on the same scale because, for example, the graduation rate could range from 40 to 61 and state higher education appropriation per FTE could range from 4494 to 7928. This process standardized the values and converted them to the same scale with a standard deviation of 1 to avoid problems with variables not having the same scale of feature scaling code as shown in Fig. 16; its code is shown in Fig. 15. After the models were developed and the results were obtained, the standardized values under the feature scaling process were converted back to normal values using the python code: ‘scaler.inverse_transform’.

Figure 15

Standardizing Values as Variables

```
import sklearn
from sklearn.datasets import load_iris
from sklearn import preprocessing
from sklearn.preprocessing import StandardScaler
data = load_iris()
X_data = data.data
target = data.target
standard = preprocessing.scale(X_data)
print(standard)
scale= StandardScaler()
X_data = data.data
target = data.target
scaled_data = scale.fit_transform(X_data)
print(scaled_data)
```

Figure 16

Results from Feature Scaling

	SHEA	AUGC	STSPCI	SHEAGDP	SFA	FTEE	GR
0	-0.38	-0.99	0.05	1.80	-1.05	-1.10	-2.14
1	-1.01	-0.97	-0.94	1.67	-1.02	-1.07	-1.94
2	1.48	-0.89	-0.35	1.26	-1.01	-1.05	-1.74
3	-1.31	-0.96	-0.94	1.94	-1.00	-1.02	-1.54
4	-0.98	-0.89	-0.75	1.67	-0.98	-1.01	-1.34

The statistical summary was rerun after the values were standardized, and the results are presented in Fig. 17. The descriptive statistics summary of the dataset shows the count, mean, standard deviation, quartiles, and minimum and maximum of the variables. The count in the data are the annual data of 50 years, i.e., from 1971 to 2020. 25, 50, and 75 from Table 8 represent the interquartile range starting with the 25th percentile (first quartile), 50th percentile (second quartile), and 75th percentile (third quartile), respectively.

Figure 17

Updated Statistical Summary of Data after Feature Scaling

	SHEA	AUGC	STSPCI	SHEAGDP	SFA	FTEE	GR
Count	50.00	50.00	50.00	50.00	50.00	50.00	50.00
Mean	0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00
Std	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Min	-1.63	-0.99	-1.34	-0.89	-1.05	-1.10	-2.14
25	-0.58	-0.76	-0.75	-0.76	-0.93	-0.79	-0.74
50	-0.19	-0.47	-0.25	-0.62	-0.43	-0.39	0.26
75	0.67	0.68	0.84	0.72	0.90	1.06	0.66
Max	2.63	2.10	1.83	1.94	2.48	1.78	2.06

Handling the missing data

The researcher replaced the missing values in the data by taking the mean value of the columns. The following code was executed to impute the missing value.

Figure 18

Code for Handling Missing Values in Python

```
imputer = SimpleImputer(missing_values=np.nan, strategy='mean') imputer = imputer.fit(X[:, 1:])
X[:, 1:] = imputer.transform(X[:, 1:])
```

Splitting the dataset into training and test datasets

In this section, the data were split into training and test data using the algorithms built in Fig. 19 and Fig. 20 for FTEE and GR. The dataset was randomly split into training and test data by the software, which helped in determining the performance of the model and the algorithms developed for this research study.

Figure 19

Splitting the dataset into training and test datasets for FTE Enrollment

```
X = df[['SHEA','AUGC', 'STSPCI', 'SHEAGDP', 'SFA']]

Y = df['FTEE']

X_train, X_test, Y_train, Y_test = sklearn.model_selection.train_test_split(X, Y, test_size = 0.20, random_state = 0)

sc_X = StandardScaler()

X_train = sc_X.fit_transform(X_train)

X_test = sc_X.transform(X_test)
```

Figure 20

Splitting the dataset into training and test datasets for the Graduation rate

```
X = df[['SHEA','AUGC', 'STSPCI', 'SHEAGDP', 'SFA']]

Y = df['GR']

X_train, X_test, Y_train, Y_test = sklearn.model_selection.train_test_split(X, Y, test_size = 0.20, random_state = 0)

sc_X = StandardScaler()

X_train = sc_X.fit_transform(X_train)

X_test = sc_X.transform(X_test)
```

Model Training & Testing

The researcher first loaded the dataset into Python programming using the pandas library to train the model and test it. Then, the dataset was split into independent (X) variables and dependent (Y) variables. This would help predict values of dependent (Y) variables using independent (X) variables. The dataset was randomly divided into two parts by Python algorithms: the first part comprised 80% of the dataset called the training dataset, and the second part comprised 20% of the dataset called the test dataset. Python's data analytical model was deployed to train the model on the training dataset. Then, the model was tested on the testing dataset to evaluate the model's accuracy. An essential feature of all data analytical models is determining their prediction accuracy levels. Hence, to assess the model's accuracy, we trained with 80% of the dataset and predicted the output values from the test dataset. The advantage of training and testing the dataset was to evaluate how well the model worked. The code shown in Fig. 21 helped in developing the regression model. Various Python libraries were imported to develop the multiple regression model.

Figure 21

Model Development

```
import sklearn
from sklearn import linear_model
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split

X_train, X_test, Y_train, Y_test = sklearn.model_selection.train_test_split(X, Y, test_size = 0.20, random_state = 0)

model = sm.OLS(Y_train, sm.add_constant(X_train)).fit()

Y_pred = model.predict(sm.add_constant(X_test))

print_model = model.summary()
```

Figure 22
OLS Regression Results for FTE Enrollment (Scaled Values)

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                   FTEE   R-squared:                       0.978
Model:                            OLS   Adj. R-squared:                  0.974
Method:                 Least Squares   F-statistic:                     302.3
Date:                Mon, 14 Mar 2022   Prob (F-statistic):           3.18e-23
Time:                        19:45:37   Log-Likelihood:                 8.4239
No. Observations:                  40   AIC:                            -2.735
Df Residuals:                      34   BIC:                             5.490
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.1314      0.035      3.754      0.000       0.060       0.202
SHEA           0.6794      0.171      3.972      0.000       0.332       1.027
AUGC          -0.1045      0.043     -2.430      0.023      -0.192       0.017
STSPCI         0.3025      0.199      1.523      0.137      -0.101       0.706
SHEAGDP       -0.0663      0.065     -1.015      0.317      -0.199       0.067
SFA            0.1333      0.051      2.613      0.015       0.030       0.237
==============================================================================
Omnibus:                        0.770   Durbin-Watson:                   1.928
Prob(Omnibus):                  0.792   Jarque-Bera (JB):                1.567
Skew:                           0.179   Prob(JB):                        0.702
Kurtosis:                       2.558   Cond. No.                         13.3
==============================================================================
```

This study was conducted with five different independent attributes at the 0.05 significance level. When the initial model was run within five attributes, an R squared value or a coefficient of determination value of 0.978 and an adjusted R^2 value of 0.974 were obtained, as shown in Fig. 22. However, out of five independent variables (IVs), two of them (STSPCI and SHEAGDP) have p values greater than 0.05. Therefore, the results indicate that the two variables with higher p values are not statistically significant in the model. Hence, the researcher removed the nonsignificant variables from the initial model and reran the multiple linear regression model with significant independent variables (IVs): SHEA, AUGC, and SFA, as shown in Fig. 23.

Figure 23

OLS Regression Results for FTE enrollment (FTEE) with Significant IVs (Scaled Values)

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                   FTEE   R-squared:                       0.961
Model:                            OLS   Adj. R-squared:                  0.955
Method:                 Least Squares   F-statistic:                     295.7
Date:                Mon, 14 Mar 2022   Prob (F-statistic):           4.78e-25
Time:                        19:55:52   Log-Likelihood:                 6.3954
No. Observations:                  40   AIC:                            -4.791
Df Residuals:                      36   BIC:                             1.965
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.1314      0.035      3.754      0.000       0.060       0.202
SHEA           0.6296      0.149      4.227      0.000       0.328       0.932
AUGC          -0.1915      0.049     -3.908      0.000      -0.291       0.092
SFA            0.4124      0.150      2.755      0.009       0.109       0.716
==============================================================================
Omnibus:                        0.843   Durbin-Watson:                   2.001
Prob(Omnibus):                  0.208   Jarque-Bera (JB):                2.216
Skew:                           0.217   Prob(JB):                        0.830
Kurtosis:                       2.768   Cond. No.                         9.76
==============================================================================
```

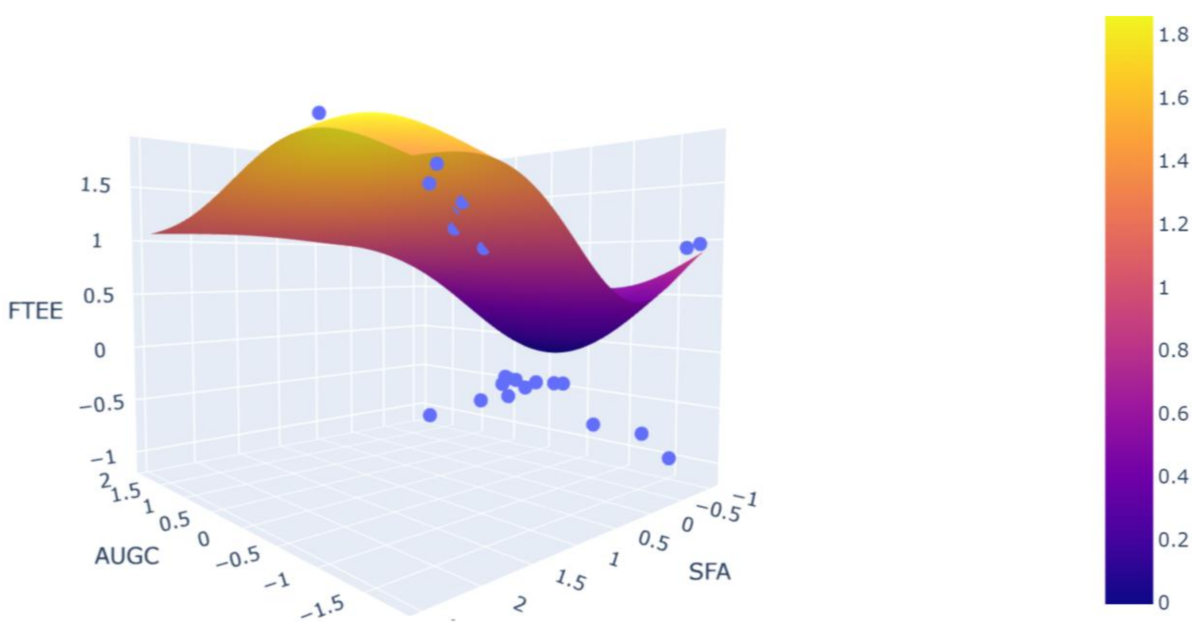
This study was conducted with three different attributes at the 0.05 significance level. When the model was run within three attributes, an R squared value or a coefficient of determination value of 0.961 and an adjusted R^2 value of 0.955 were obtained, as shown in Fig. 23. This indicates that 96.1% of the variation in the dependent variable FTEE can be explained by the independent variables SHEA, AUGC, and SFA. The F-statistic for the regression model is 295.7. The p value associated with the overall F-statistic for this model, the Prob (F-statistic), is 4.78e-25. This shows that this regression model is statistically significant. This means that the three predictor variables SHEA, AUGC, and SFA combined have a statistically significant association with the response variable FTE enrollment.

Figure 23 shows the positive and negative variation in the dependent variable FTEE.

The coefficient in Figure 23, which is the coefficient for each predictor variable, shows the average expected change in FTE enrollment, assuming that the other predictor variables remain constant. The FTE enrollment at 4-year public universities is expected to increase 0.63 students for every $1 more provided than the annual average to state higher education appropriations (SHEA) per FTE. For example, public higher education institutions are receiving $5000 per FTE annually as SHEA dollars for 100 FTE students. Then, with $5001 dollars awarded annually to public higher education institutions, FTE enrollment (FTEE) reaches 100.63 students. Likewise, an increase in average undergraduate tuition (AUGC) per FTE by $1 annually results in a decrease in FTE enrollment (FTEE) by 0.19 students. Finally, FTE enrollment (FTEE) increases 0.41 students for every additional $1 given annually to student financial aid (SFA) per FTE, assuming the other predictor variables remain constant.

Figure 24

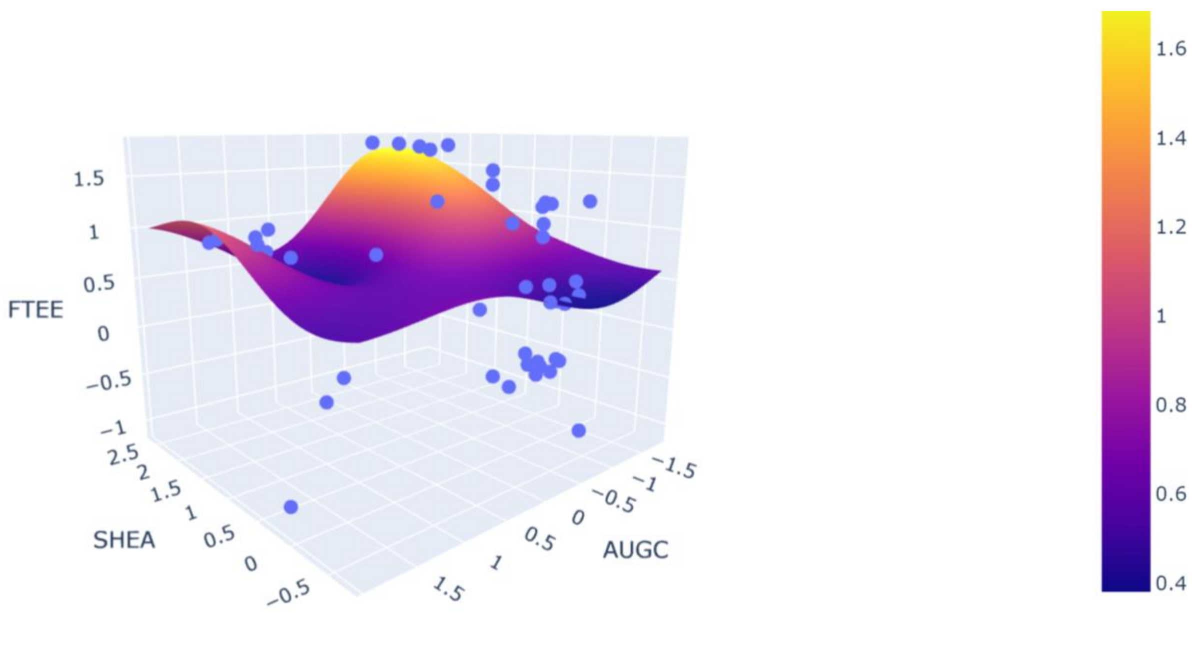
3D Regression Surface for FTE Enrollment, AUGC, and SFA (Scaled Values)



Three-dimensional regression surfaces were constructed with three significant independent variables with the dependent variable FTEE. This helped in understanding how independent variables affected the dependent variable FTEE. Since a 3D model accommodates only one dependent variable and two independent variables at a time, three different 3D regression surfaces were presented using two independent variables per 3D regression surface. The first surface is shown in Figure 24. Here, the three-dimensional regression surface was constructed to show the variation in the two independent variables, AUGC and SFA, on the FTEE at each grid point. It helps in viewing the relationship between the variables. This is created as an interactive 3D model, and the link to interacting with the 3D regression surface is provided in the appendix.

Figure 25

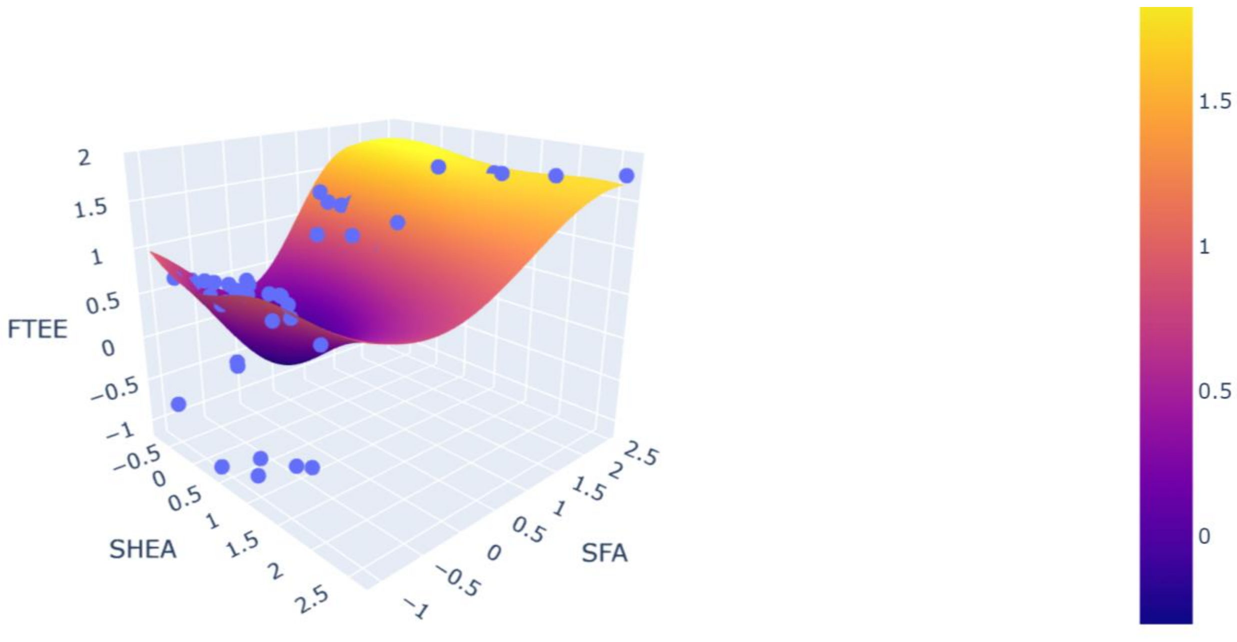
3D Regression Surface for FTE Enrollment, SHEA, and AUGC (Scaled Values)



In Figure 25, the three-dimensional regression surface is constructed to show the variation in two independent variables, SHEA and AUGC, on the FTEE at each grid point. It helps in viewing the relationship between the variables. This is created as an interactive 3D model, and the link to interacting the 3D regression surface is provided in the appendix.

Figure 26

3D Regression Surface for FTE Enrollment, SHEA, and SFA (Scaled Values)



In Figure 26, the three-dimensional regression surface is constructed to show the variation of two independent variables, SHEA and SFA, on FTEE at each grid point. It helps in viewing the relationship between the variables. This is created as an interactive 3D model, and the link to interacting the 3D regression surface is provided in the appendix.

Based on the regression analysis, the predicted values for FTE enrollment are shown in Figure 27. The dataset was split into training and test data to obtain a good estimate of the performance of the model. To evaluate the supervised machine learning algorithms and help implement this model on other datasets. Most of the predictions were very close to the actual value. This shows that the model developed by this research study using 80% of the training data of the three significant independent variables and a dependent variable of FTE enrollment

(FTEE) was able to predict 20% of the test data. The model performance could be considered good because 80% of the training data were able to predict 20% of the test data. This is very significant because there are no abnormal predictions in the model; However, as seen in fig. 27, for the 2nd and 38th FTEE values, there was a 10.43% and 11.76% difference between the actual value and the predicted value, respectively. The 1973-74 oil crisis and the great recession during 2008-09 could have caused these issues. It was an interesting finding from the study, and the US historical events align with it. The scaled values of variables under the feature scaling process were converted back to normal values using the python code: 'scaler.inverse_transform'. The values shown in fig. 27 are actual values from the dataset, not the standardized values. The dataset's actual values and scaled values can be seen in Appendix P. The model performance was evaluated in the next section of this chapter, called model evaluation.

Figure 27

Multiple Linear Regression Model Test Results for FTEE

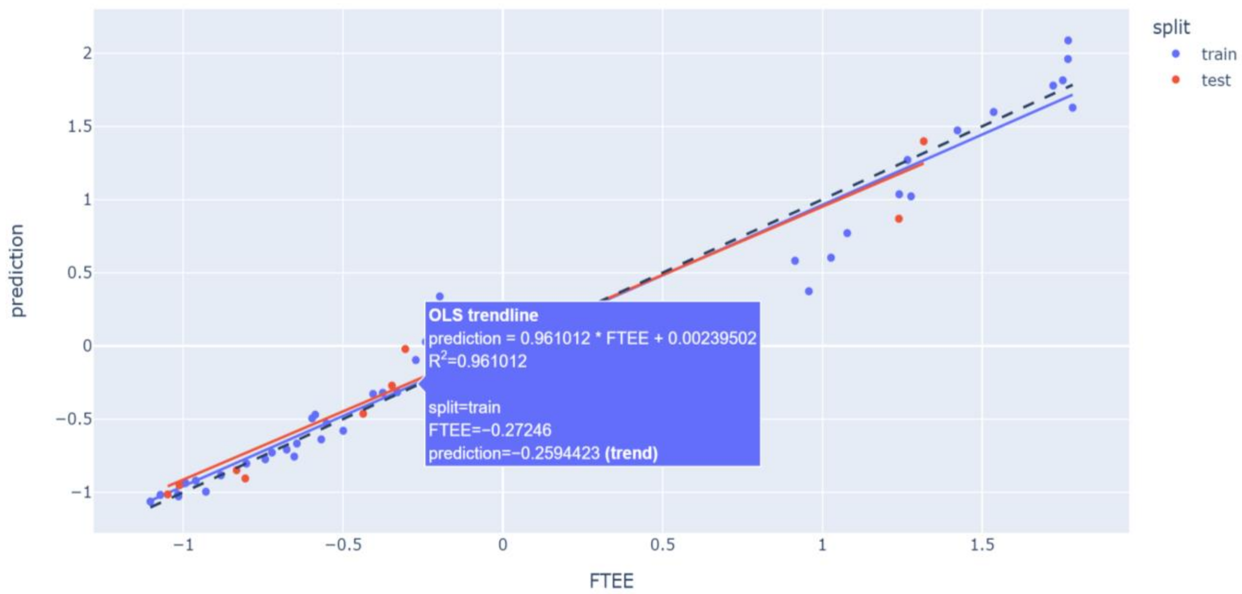
```
mlr_diff = pd.DataFrame({'Actual value': Y_test, 'Predicted value': Y_pred, 'Residual': Y_test - Y_pred})
mlr.diff.head()
print (mlr_diff.round())
```

```
    Actual value  Predicted value  Residual
28        155321         155668.0    -347.0
11        105791         102134.0    3657.0
10        103989         104145.0    -156.0
41        210111         213278.0   -3167.0
2          89716          98310.0   -8594.0
27        148893         147253.0    1640.0
38        198893         181983.0   16910.0
31        153179         154010.0    -831.0
22        149627         144563.0    5064.0
4          92127          96835.0   -4708.0
```

In figure 28, the prediction error plot for FTE enrollment (FTEE) shows the actual training and test values from the dataset against the predicted values generated by the research model. This shows how much variance is present in the model. This model shows that the predicted values line fits through the FTE enrollment's training and test values.

Figure 28

Prediction Error Analysis for FTEE



In Fig. 29, the residuals are on the vertical axis, and the predicted values of FTE enrollment are on the horizontal axis. The residual plot in Fig. 29 shows a random pattern. Some of the residuals are positive, and some are negative. This pattern indicates that a linear model provides an excellent fit for the data.

Figure 29

Prediction Residuals for FTEE

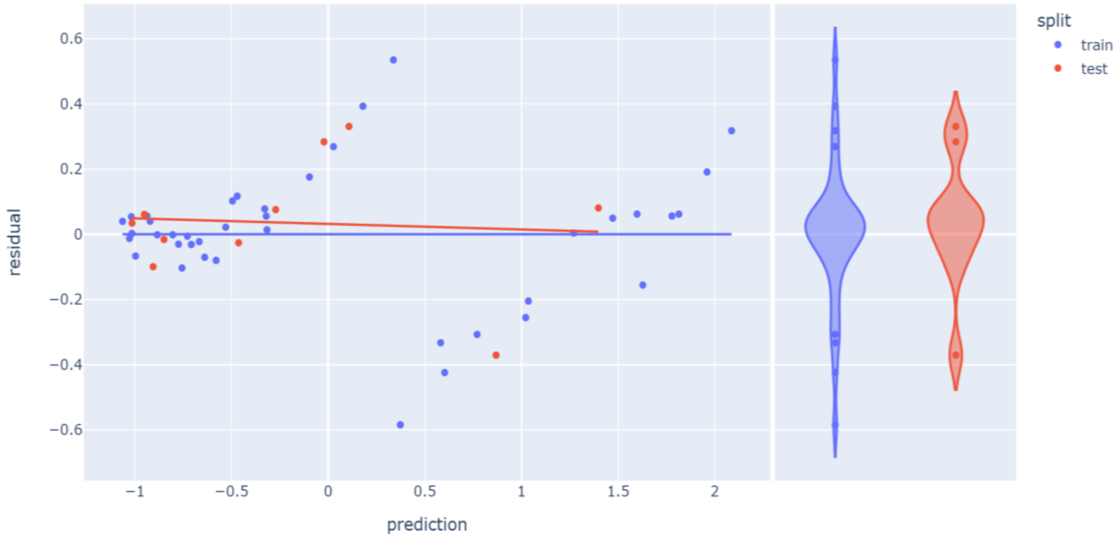


Figure 30 shows the comparison of actual output and predicted output. Since the scatter dots are close to the diagonal black line, this indicates that it is a good model. The values between 0 and 1 are mostly missing because this plot shows the predicted value in comparison to the actual value. As the dataset values were scaled to be within a standard deviation of 1, most of the values are close to -1 or +1. Since we don't have any abnormal values, most of them were seen to be close to the predicted value. It could be seen from fig. 30 that the value from the X-axis to its corresponding value from Y-axis matches closely. Since there are no null values in the dataset, the FTEE data was spread out to be close to -1 and +1.

Figure 30

Predicted Output vs. The Actual Output for FTEE

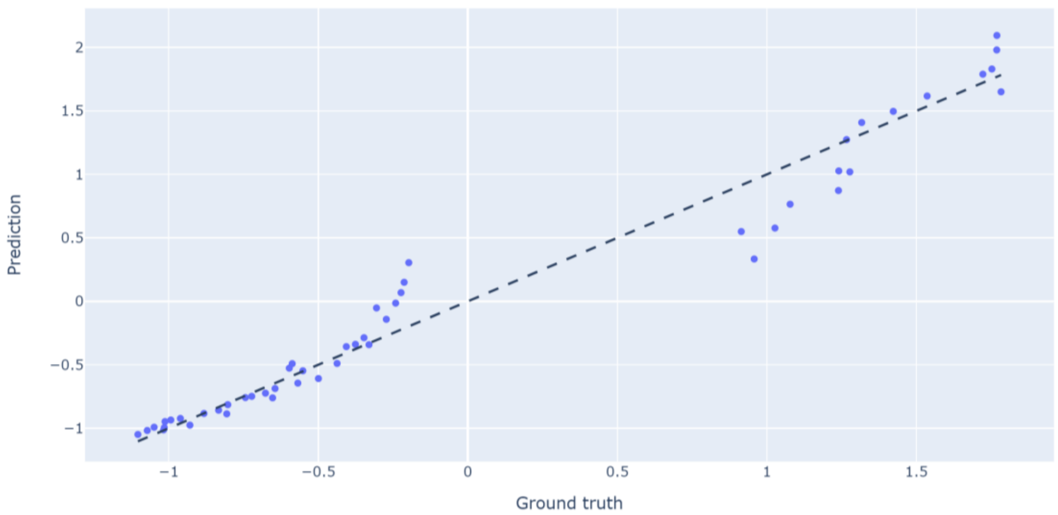


Figure 31

OLS Regression Results for Graduation rate (Scaled Values)

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                     GR   R-squared:                       0.965
Model:                            OLS   Adj. R-squared:                  0.959
Method:                 Least Squares   F-statistic:                     187.48
Date:                Mon, 14 Mar 2022   Prob (F-statistic):           5.83e-19
Time:                        20:39:27   Log-Likelihood:                -1.5703
No. Observations:                  40   AIC:                             15.52
Df Residuals:                      34   BIC:                             25.65
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.1705      0.043      3.965      0.000       0.083       0.257
SHEA           0.1917      0.064      2.995      0.005       0.062       0.322
AUGC           0.3092      0.215      1.440      0.159      -0.127       0.745
STSPCI        -0.5732      0.082     -6.988      0.000      -0.740      -0.407
SHEAGDP       -0.0314      0.054     -0.585      0.562      -0.141       0.078
SFA            0.1308      0.249      0.525      0.603      -0.376       0.637
==============================================================================
Omnibus:                        0.713   Durbin-Watson:                   2.646
Prob(Omnibus):                  0.700   Jarque-Bera (JB):                0.486
Skew:                          -0.267   Prob(JB):                        0.784
Kurtosis:                       2.919   Cond. No.                         13.3
==============================================================================
```

This study was conducted with five different independent attributes at the 0.05 significance level. When the initial model was run within five attributes, an R squared value or a coefficient of determination value of 0.965 and an adjusted R^2 value of 0.959 were obtained, as shown in Fig. 31. However, out of five independent variables (IVs), three of them (AUGC, SHEAGDP, and SFA) have p values greater than 0.05. Therefore, the results indicate that the three variables with higher p values are not statistically significant in the model. Hence, the researcher removed the nonsignificant variables from the initial model and reran the multiple linear regression model with significant independent variables (IVs): SHEA and STSPCI, as shown in Fig. 32.

Figure 32

OLS Regression Results for Graduation rate (GR) with Significant Variables (Scaled Values)

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                     GR   R-squared:                       0.937
Model:                            OLS   Adj. R-squared:                  0.928
Method:                 Least Squares   F-statistic:                     275.15
Date:                Mon, 14 Mar 2022   Prob (F-statistic):           3.34e-18
Time:                        20:57:15   Log-Likelihood:                -14.685
No. Observations:                  40   AIC:                             35.37
Df Residuals:                      37   BIC:                             40.44
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.1705      0.043      3.965      0.000       0.083       0.257
SHEA           0.3293      0.062      5.314      0.000       0.204       0.455
STSPCI        -0.7747      0.062    -12.501      0.000      -0.900      -0.649
==============================================================================
Omnibus:                        0.734   Durbin-Watson:                   1.131
Prob(Omnibus):                  0.693   Jarque-Bera (JB):                1.808
Skew:                           0.197   Prob(JB):                        0.668
Kurtosis:                       2.426   Cond. No.                         1.48
==============================================================================
```

This study was conducted with two different attributes at the 0.05 significance level.

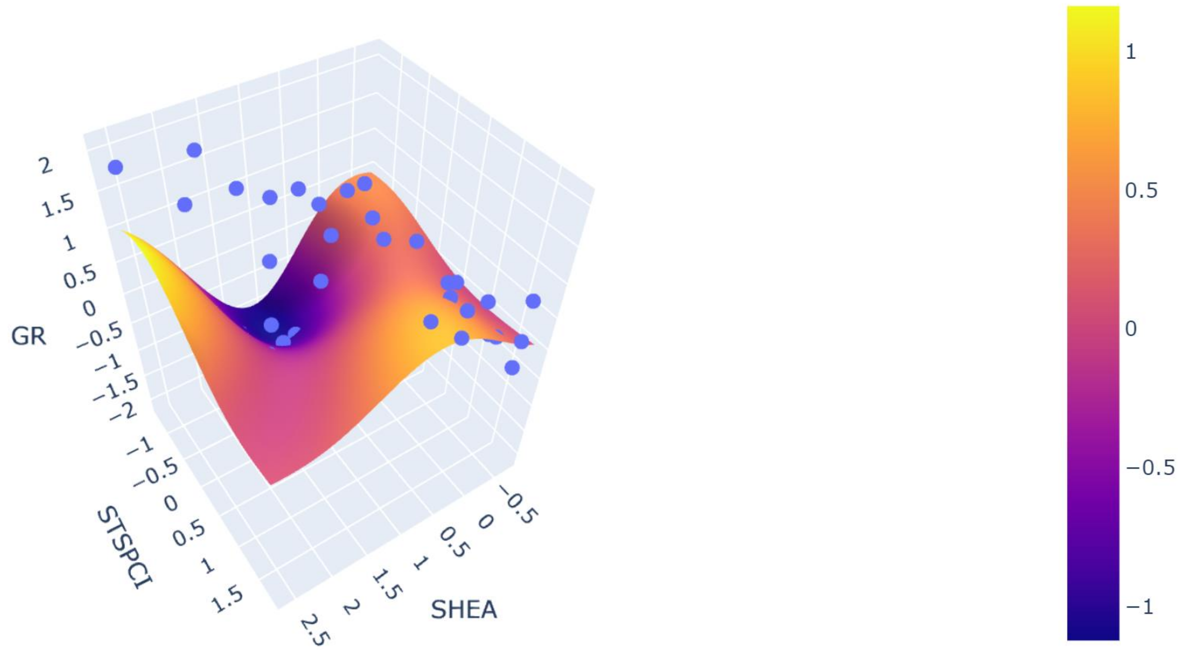
When the model was run within two attributes, an R squared value or a coefficient of determination value of 0.937 and an adjusted R squared value of 0.928 were obtained, as shown in Fig. 32. This indicates that 93.7% of the dependent variable GR variation can be explained by the independent variables SHEA and STSPCI. The F-statistic for the regression model is 275.1. The p value associated with the overall F-statistic for this model, which is the Prob (F-statistic), is 3.34e-18. This shows that this regression model is statistically significant. This means that the two predictor variables SHEA and STSPCI combined have a statistically significant association with the response variable FTE enrollment. The coefficient in Fig. 32, which is the coefficient for each predictor variable, shows the average expected change in graduation rate, assuming that the other predictor variables remain constant.

Figure 32 shows the positive and negative variation in the dependent variable GR. The coefficient in Figure 32, which is the coefficient for each predictor variable, shows the average expected change in graduation rate (GR), assuming that the other predictor variables remain constant. The graduation rate at 4-year public universities is expected to increase 0.33% for every $1 more provided than the annual average to state higher education appropriations (SHEA) per FTE. For example, public higher education institutions are receiving $5000 per FTE annually as SHEA dollars for a 50 percent annual graduation rate. Then, with $5001 dollars awarded to public higher education institutions, the graduation rate (GR) is 50.33%. Likewise, an increase in student tuition share as a percentage of per capita income (STSPCI) by 1 percent annually results in a decrease in graduation rate (GR) by 0.77%, assuming the other predictor variables remain constant.

In Figure 33, the three-dimensional regression surface is constructed to show the variation in the two independent variables SHEA and STSPCI on GR at each grid point. It helps in viewing the relationship between the variables. This is created as an interactive 3D model, and the link to interacting with the 3D regression surface is provided in the appendix.

Figure 33

3D Regression Surface for Graduation rate, SHEA, and STSPCI (Scaled Values)



Based on the regression analysis, the predicted values for the graduation rate (GR) are shown in Figure 34. The data were split into training and test data to obtain a good estimate of the performance of the model. To evaluate the supervised machine learning algorithms and to help implement this model on other datasets. Most of the predictions were very close to the actual value. This shows that the model developed by this research study using 80% of the training data of the two significant independent variables SHEA and STSPCI and a dependent variable of graduation rate (GR) was able to predict 20% of the test data. This is very significant because there are no abnormal predictions in the model, and the model was evaluated in the next section of this chapter called model evaluation. Figure 34 values suggest that the performance of the model was very good. However, as seen in fig. 34, for the 2nd value, there was a 9% difference between the actual and predicted values. The 1973-74 oil crisis could have caused this issue. The study found an interesting finding, and the US historical event aligned with it. The scaled values of variables under the feature scaling process were converted back to normal values using the python code: 'scaler.inverse_transform'. The values are shown in fig. 34 are actual values from the dataset, not the standardized values. The dataset's actual and scaled values can be seen in Appendix P.

Figure 34

Multiple Linear Regression Model Test Results for Graduation rate

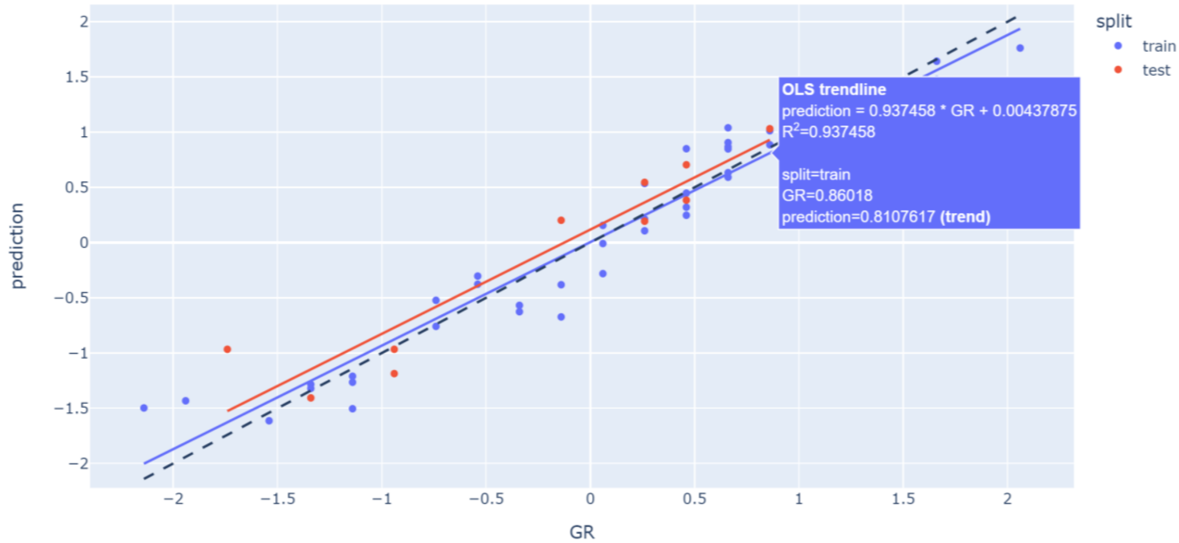
```
mlr_diff = pd.DataFrame({'Actual value': Y_test, 'Predicted value': Y_pred, 'Residual': Y_test - Y_pred})
mlr.diff.head()
print (mlr_diff.round())
```

```
    Actual value  Predicted value  Residual
28            53             53.0       0.0
11            46             46.0       0.0
10            46             45.0       1.0
41            55             56.0      -1.0
2             42             46.0      -4.0
27            52             52.0       0.0
38            53             54.0      -1.0
31            52             53.0      -1.0
22            50             52.0      -2.0
4             44             44.0       0.0
```

In Figure 35, the prediction error plot for the graduation rate (GR) shows the actual training and test values from the dataset against the predicted values generated by the research model. This shows how much variance is present in the model. This model shows that the predicted values line fits through the training and test values of the graduation rate.

Figure 35

Prediction Error Analysis for GR



Here, the residuals are on the vertical axis, and the predicted values of the graduation rate are on the horizontal axis. The residual plot in Fig. 36 shows a random pattern. Some of the residuals are positive, and some are negative. This pattern indicates that a linear model provides a good fit for the data.

Figure 36

Prediction Residuals for GR

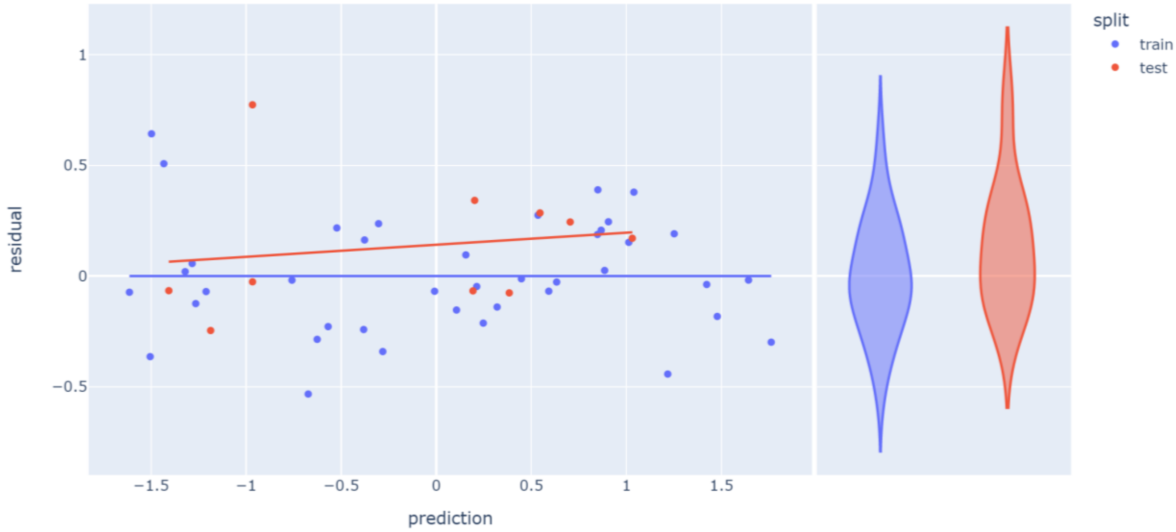
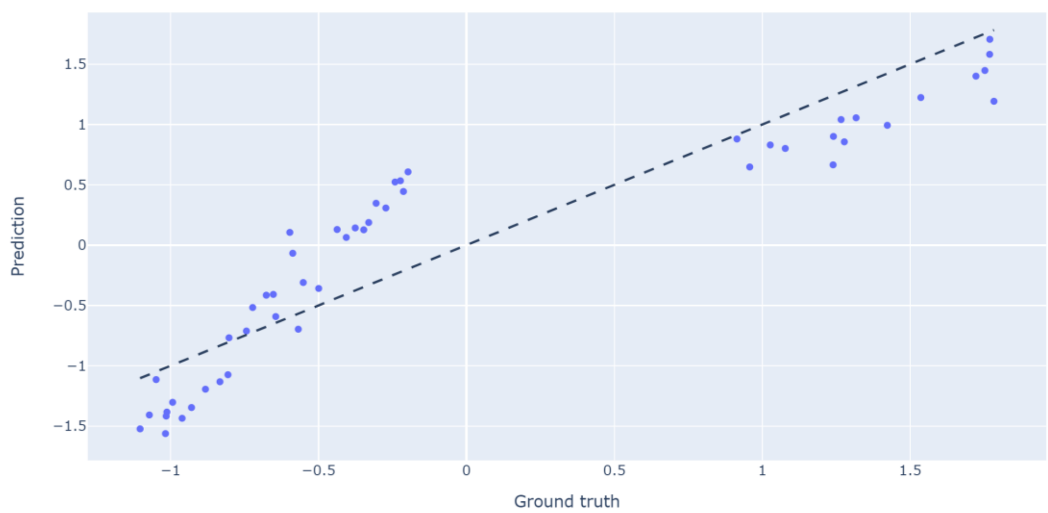


Figure 37 shows the comparison of actual output and predicted output. Since the scatter dots are close to the diagonal black line, this indicates that this is a good model. The values between 0 and 1 are mostly missing because this plot shows the predicted value in comparison to the actual value. As the dataset values were scaled to be within a standard deviation of 1, most of the values are close to -1 or +1. Since we don't have any abnormal values, most of them were seen to be close to the predicted value. It could be seen from fig. 37 that the value from the X-axis to its corresponding value from Y-axis matches closely. Since there are no null values in the dataset, the GR data was spread out to be close to -1 and +1.

Figure 37

Predicted Output vs. The Actual Output for the Graduation rate (GR)



Model Evaluation

In this section, both the FTE enrollment (FTEE) and graduation rate (GR) models were evaluated. To evaluate the FTEE and GR model's performance for this research study, we imported the sklearn library, which provides metrics for model evaluation. The first evaluation starts with FTE enrollment. In Figure 38, the evaluation model developed is presented along with scores reported for FTE enrollment.

Figure 38

R^2 score, Mean Absolute Error, Mean Square Error, Root Mean Square Error for FTEE

```
from sklearn import metrics
from sklearn.linear_model import LinearRegression
LR = LinearRegression()
LR = LinearRegression()
LR.fit(X_train,Y_train)
Y_pred = LR.predict(X_test)
meanAbErr = metrics.mean_absolute_error(Y_test, Y_pred)
meanSqErr = metrics.mean_squared_error(Y_test, Y_pred)
rootMeanSqErr = np.sqrt(metrics.mean_squared_error(Y_test, Y_pred))
print('R squared: {:.2f}'.format(LR.score(X,Y)*100))
print('Mean Absolute Error:', meanAbErr.round(2))
print('Mean Square Error:', meanSqErr.round(2))
print('Root Mean Square Error:', rootMeanSqErr.round(2))
```

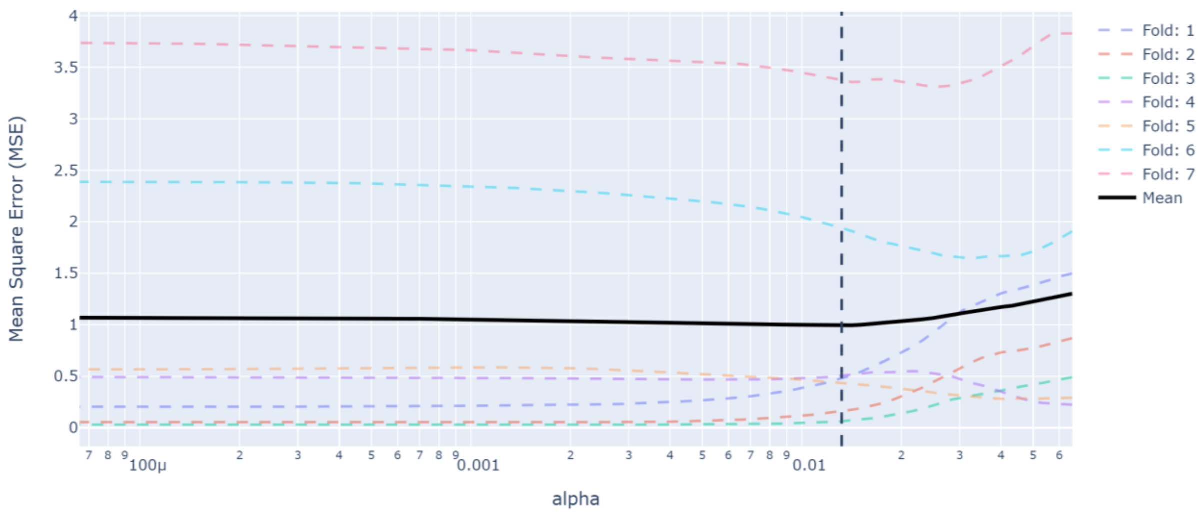
```
R squared: 96.14
Mean Absolute Error: 0.14
Mean Square Error: 0.04
Root Mean Square Error: 0.19
```

To evaluate the performance of the model developed for this research study dataset, the study measured how well the predictions made by the model matched the observed data. K-fold cross-validation is a commonly used method for validation. From the output in Figure 38, it can be seen that the mean absolute error was 0.14, and the mean square error was 0.04. The mean absolute error is the average absolute error between the model prediction and the actual observed data, and for this model, it was 0.14. Mean square error is the average squared difference between the predicted values and the true value of the observed data. For this model, the mean square error is 0.04. The root mean square error for the model was 0.19. The lower the root mean square error was, the better the model was able to predict the actual values. Therefore, this model with the lowest test error rates was the best model to use for FTE enrollment (FTEE). Additionally, the R squared value for this model is 0.961, which means that 96.1% of the variation in the dependent variable FTE enrollment (FTEE) is explained by the independent variables SHEA, AUGC, and SFA.

To show that this model produces the lowest rate of error rates and is the best model for this study, cross-validation results for mean square error were conducted. To show the cross-validation results in different folds, this study chose to use K=7-fold. Studies typically select between 5-fold and 10-fold to see if the study could produce reliable test error rates. This cross-validation process was conducted in Python, and Figure 39 shows 7 different mean square error cross-validation folds for FTE enrollment (FTEE).

Figure 39

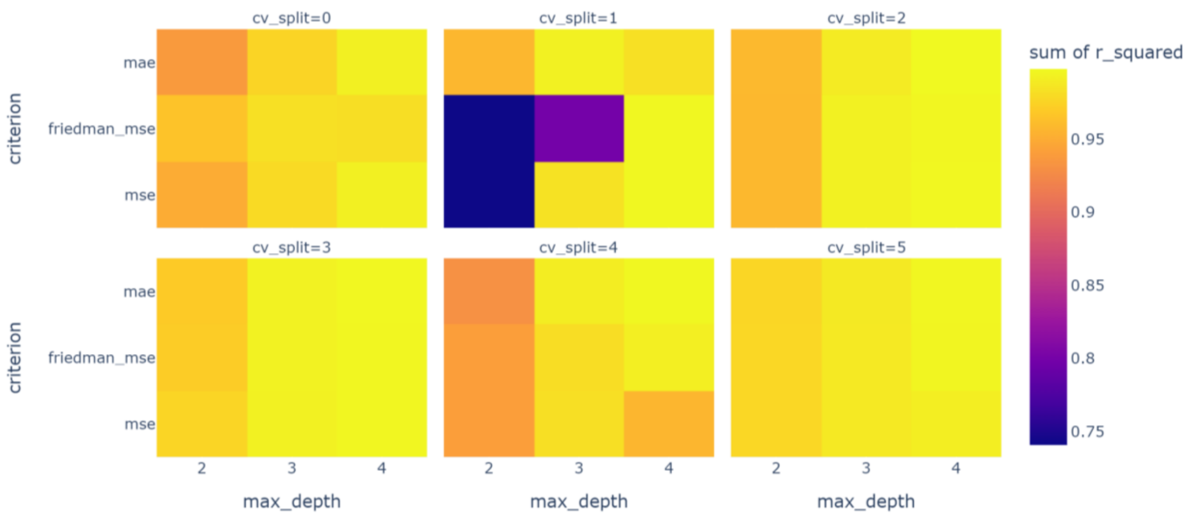
Optimal Alpha Variation Across Cross-Validation Folds for FTEE



In Fig. 40, a two-dimensional graphical visualization of grid results shows the density heatmap using a decision tree regressor for the response variable FTE enrollment (FTEE). A heatmap is a visual representation of data that shows selected values in a matrix represented as colors. The color bar on the right side shows what values various colors represent in the heatmap. In this heatmap, the researchers selected the mean absolute error, Friedman mean square error, and mean square error on the R squared value to evaluate the model. This map was developed as an interactive map to display values when an individual fold is selected in the map. The link to accessing the interactive heatmap was added to the appendix.

Figure 40

Density Heatmap using Decision Tree Regressor for FTEE



A box plot was used to evaluate the model by graphically depicting the error data for FTE enrollment (FTEE) through quartiles. Figure 41 shows the upper and lower quartiles in a box-and-whisker plot.

Figure 41

Box Plot using Decision Tree Regressor for FTEE

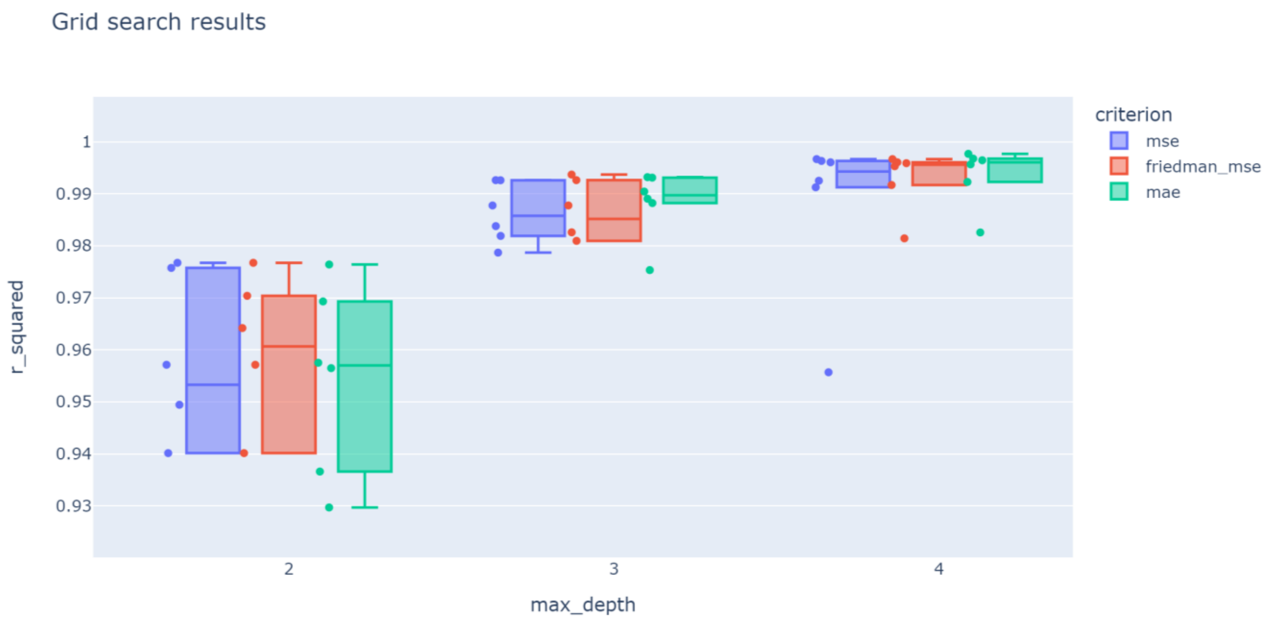


Figure 42

R^2 score, Mean Absolute Error, Mean Square Error, Root Mean Square Error for GR

```
from sklearn import metrics
from sklearn.linear_model import LinearRegression
LR = LinearRegression()
LR = LinearRegression()
LR.fit(X_train,Y_train)
Y_pred = LR.predict(X_test)
meanAbErr = metrics.mean_absolute_error(Y_test, Y_pred)
meanSqErr = metrics.mean_squared_error(Y_test, Y_pred)
rootMeanSqErr = np.sqrt(metrics.mean_squared_error(Y_test, Y_pred))
print('R squared: {:.2f}'.format(LR.score(X,Y)*100))
print('Mean Absolute Error:', meanAbErr.round(2))
print('Mean Square Error:', meanSqErr.round(2))
print('Root Mean Square Error:', rootMeanSqErr.round(2))
```

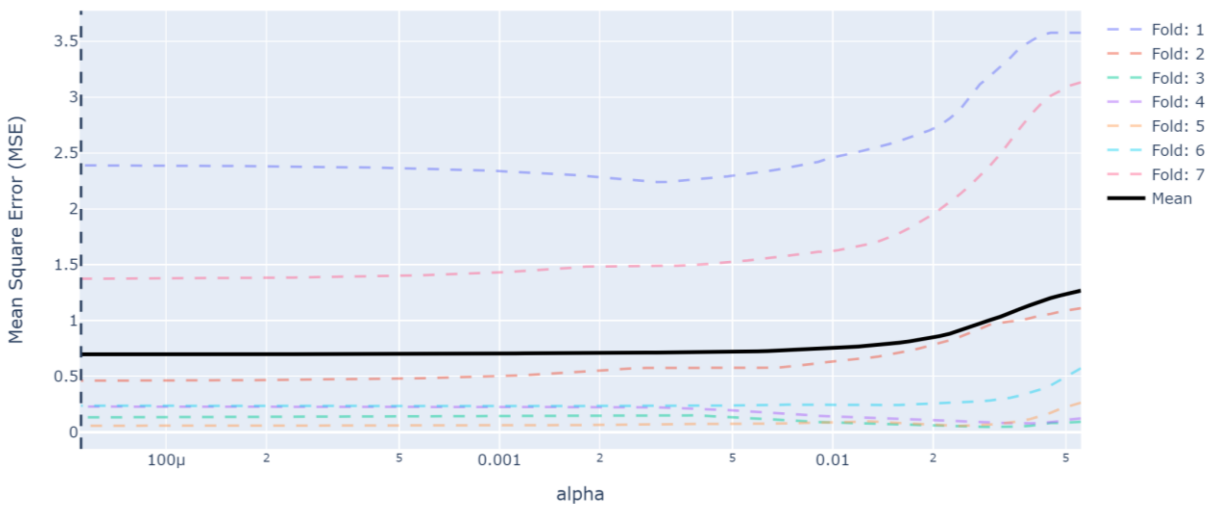
```
R squared: 93.77
Mean Absolute Error: 0.23
Mean Square Error: 0.1
Root Mean Square Error: 0.31
```

To evaluate the performance of the model developed for this research study dataset, the study measured how well the predictions made by the model matched the observed data. K-fold cross-validation is a commonly used method for validation. From the output in Figure 42, it can be seen that the mean absolute error was 0.23, and the mean square error was 0.1. The mean absolute error is the average absolute error between the model prediction and the actual observed data, and for this model, it was 0.23. Mean square error is the average squared difference between the predicted values and the true value of the observed data. For this model, the mean square error is 0.1. The lower the root mean squared error is, the better the model can predict the actual values. The root mean squared error for this model was 0.31. Therefore, this model with the lowest test error rates was the best model for the graduation rate (GR). Additionally, the R squared value for this model is 0.937, which means that 93.7% of the variation in the dependent variable GR is explained by the independent variables SHEA and STSPCI.

To show that this model produces the lowest rate of error rates and is the best model for this study, cross-validation results for mean square error were conducted. To show the cross-validation results in different folds, this study used K=7-fold. Studies typically choose between 5-fold and 10-fold to see if the study could produce reliable test error rates. This cross-validation process was conducted in Python, and Figure 43 shows 7 different mean square error cross-validation folds for the graduation rate.

Figure 43

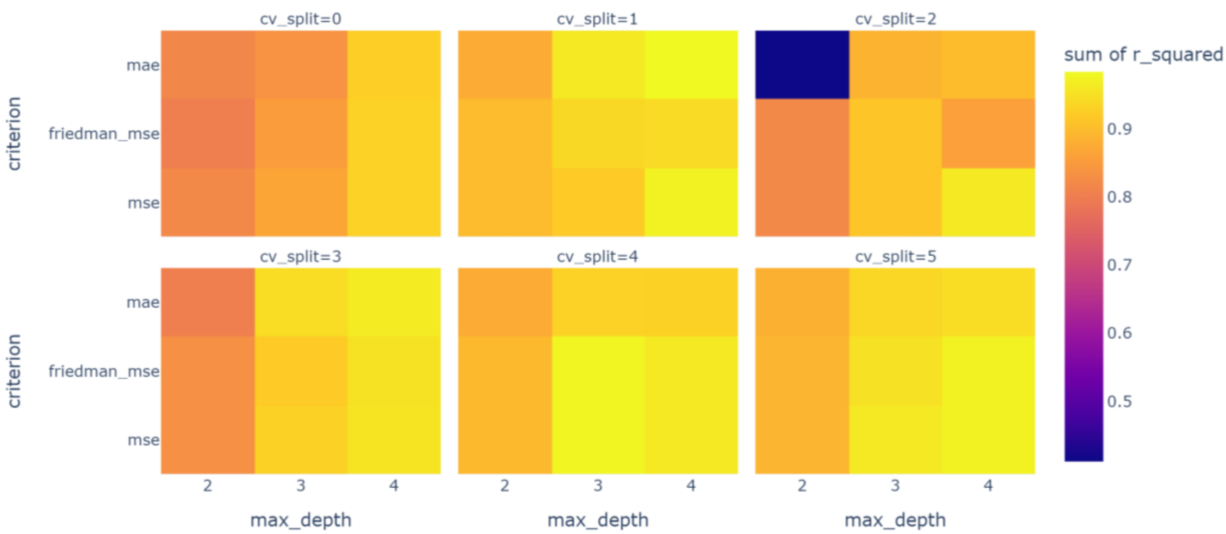
Optimal Alpha Variation Across Cross-Validation Folds for FTEE



In Fig. 44, a two-dimensional graphical visualization of grid results shows the density heatmap using a decision tree regressor for the response variable graduation rate. A heatmap is a visual representation of data that shows selected values in a matrix represented as colors. The color bar on the right side shows what values various colors represent in the heatmap. In this heatmap, the researchers selected the mean absolute error, Friedman mean square error, and mean square error on the R squared value to evaluate the model. This map was developed as an interactive map to display values when an individual fold is selected in the map. The link to accessing the interactive heatmap was added to the appendix.

Figure 44

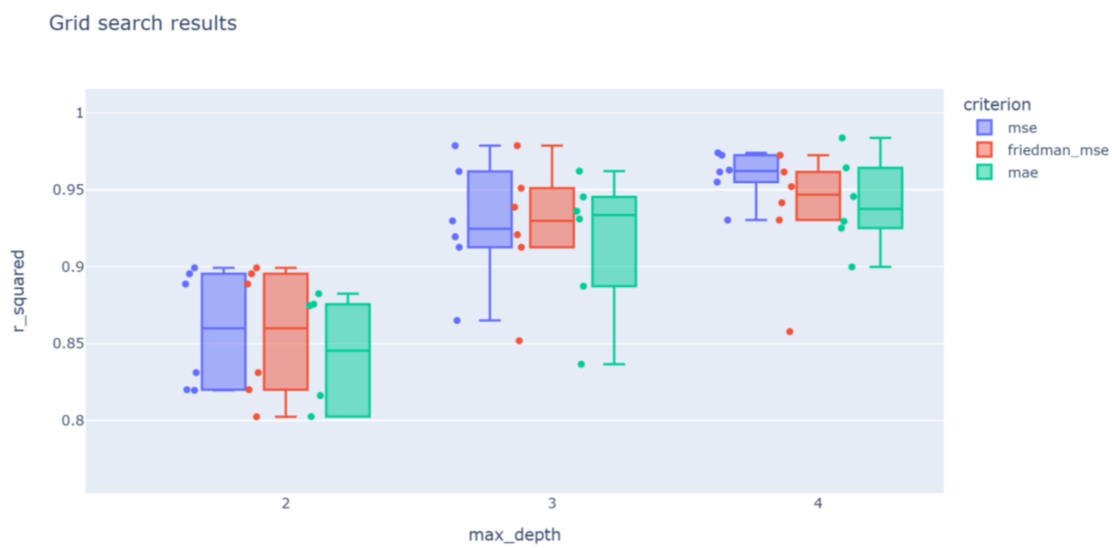
Density Heatmap using Decision Tree Regressor for GR



A box plot was used to evaluate the model by graphically depicting the error data for graduation rate through quartiles. Figure 45 shows the upper and lower quartiles in a box-and-whisker plot.

Figure 45

Boxplot using Decision Tree Regressor for Graduation rate



Chapter Summary

In this chapter, two main research questions were reviewed, and the results of the hypotheses were presented. By developing a multiple linear regression model using data analytical methods and machine learning algorithms, this study was able to find the variation in dependent variables FTEE and GR values based on changes to significant independent variables. Both null hypotheses were rejected at the $\alpha = 0.05$ significance level because the models developed by this research study showed that the three predictor variables (SHEA, AUGC, and SFA) together could predict the response variables of FTE enrollment (FTEE) statistically significantly. The second model with two predictor variables (SHEA and STSPCI) together was able to predict the response variable of graduation rate (GR) statistically significantly. The detailed explanation of rejecting the null hypotheses is presented in chapter 5 under the subsection model summary for FTEE and GR.

CHAPTER 5

MODEL SUMMARY, CONCLUSION, DISCUSSION, AND RECOMMENDATIONS

Chapter Overview

This chapter is primarily divided into four major sections: model summary, conclusion, discussion, and recommendations. The purpose of this chapter was to discuss these quantitative research study models in more detail and summarize the impact of state budget cuts on public 4-year higher education universities in the US. This chapter provides a summary of the research models with conclusion, discussion, and recommendations for future study based on the objectives of the research study.

Model Summary

In chapter 4, multiple linear regression models were developed to predict dependent variables (DVs) from independent variables (IVs), and the summary of the findings from the data analytical study was reviewed and discussed. The two models for FTEE and GR are explained in detail as follows:

Multiple Regression Model for FTEE

Initially, in this FTEE model, there were a total of five IVs related to state budget cuts to predict FTE enrollment, and their p value determines their significance in the model. Out of five variables, two of them (STSPCI, SHEAGDP) had higher p values. Therefore, the results from Fig. 22 indicate that the two variables with higher p values are not statistically significant in the model. Hence, the estimated regression equation for the model after removing nonsignificant variables is:

$$\mathbf{FTEE = 0.1314 + 0.6296 \times (SHEA) + \text{-}0.1915 \times (AUGC) + 0.4124 \times (SFA)} \quad (6)$$

$$\mathbf{R\text{-}\ Square = 96.1\%} \quad (7)$$

The model output from equation (6) indicates that if other variables are held constant, FTEE at 4-year public universities is expected to increase 0.63 students for every $1 more provided than the annual average to SHEA per FTE. FTEE will decrease by 0.19 students following an annual rise of $1 per FTE in AUGC. The model also shows that FTEE will increase 0.41 students if SFA is increased by $1 per FTE. For example, public higher education institutions are receiving $5000 per FTE annually as SHEA dollars for 100 FTE students. Then, with $5001 dollars awarded annually to public higher education institutions, FTE enrollment (FTEE) reaches 100.63 FTE students. Likewise, an increase in average undergraduate tuition (AUGC) per FTE by $1 annually results in a decrease in FTE enrollment (FTEE) by 0.19 students. Finally, FTE enrollment (FTEE) increases 0.41 students for every additional $1 given annually to student financial aid (SFA) per FTE, assuming the other predictor variables remain constant. Here, 96.1% of the variation in the dependent variable FTEE, as shown in equation (7), is explained by the independent variables SHEA, AUGC, and SFA.

The main null hypothesis (H_{01}) in this multiple regression study states that there was no significant predictive relationship between the five independent variables (IVs): SHEA, AUGC, STSPCI, SHEAGDP, and SFA, and the dependent variable (DV): FTEE at public higher education institutions. The results from Fig. 23 indicate that three IVs added together are statistically significant to the prediction, $p < .05$. Since the p value for the final FTEE model is less than 0.05, this study rejects the null hypothesis (H_{01}).

Multiple Regression Model for GR

Initially, in this GR model, there were a total of five IVs related to state budget cuts to predict FTE enrollment, and their p value determines their significance in the model. Out of five variables, Fig. 31 indicates that three of them (AUGC, SHEAGDP, and SFA) had higher p values. Therefore, the results indicate that the three variables with higher p values are not statistically significant in the model. Hence, the estimated regression equation for the model after removing nonsignificant variables is:

$$\mathbf{GR = 0.1705 + 0.3293 \times (\mathit{SHEA}) + \text{-}0.7747 \times (\mathit{STSPCI})} \tag{8}$$

$$\mathbf{R\text{- } Square = 93.7\%} \tag{9}$$

The multiple linear regression model from equation (8) suggests that IVs: SHEA and STSPCI variables added together are statistically significant to the prediction, $p < .05$. The model output indicates that if other variables are held constant, GR at 4-year public universities is expected to increase 0.32 for every $1 more provided than the annual average to SHEA per FTE. The model also shows that GR will decrease 0.77% if the STSPCI increases by 1%. For example, if public higher education institutions receive $5000 per FTE annually as SHEA dollars for a 50% graduation rate, the other predictor variables remain constant. Then, with $5001 dollars awarded annually to public higher education institutions per FTE, the graduation rate (GR) reaches 50.33%. The model also shows that if the STSPCI increases by 1%, i.e., assuming it went from 30% to 31%, then the GR rate goes down from 50% to 49.23%. Here, 93.7% of the variation in the dependent variable GR, as shown in equation (9), is explained by the independent variables SHEA and STSPCI.

The second null hypothesis (H_{02}) in this multiple regression study states that there is no significant predictive relationship between the five independent variables (IVs): SHEA, AUGC, STSPCI, SHEAGDP, and SFA, and the dependent variable (DV): GR at public higher education institutions. However, Fig. 32 indicates that there is a statistically significant predictive relationship between the two IVs: SHEA and STSPCI added together and DV: GR. Since the p value of the final model for GR is less than 0.05, this study rejects the null hypothesis (H_{02}).

Conclusion

In this quantitative study using data analytical methods, we evaluated the impacts of state budget cuts with multiple criteria considering multicollinearity and homoscedasticity. As a result, we built two multiple linear regression models at the 0.05 significance level with five independent variables and two dependent variables, resulting in 0.961 and 0.937 R^2 values, respectively, for the two regression models. Although multicollinearity could not be eliminated completely in these models, the study was successful by combining variables to produce one single variable instead of two individual variables. For example, student tuition share (STS) and personal per capita income (PCI) could be two variables in this study, but student tuition share was taken as a percentage of personal per capita income, becoming one variable, STSPCI. One of the explanations that we could not eliminate multicollinearity completely is that most of the independent variables are connected to each other in the higher education field.

One of the interesting things in the study is the fact that the independent variables STSPCI and SHEAGDP did not significantly predict the impacts on FTE enrollment (FTEE), which may be because prospective undergraduate students planning to study at 4-year public higher education institutions did not consider STSPCI and SHEAGDP as important factors for enrolling at a university. This study has added significant new knowledge in leveraging modern technologies such as data analytics and machine learning algorithms to study the long-standing problems of state budget cuts to public 4-year higher education institutions in the US. This research is critical because the COVID-19 pandemic has dramatically affected state economies and budgets. The states are struggling to prioritize their budget allocations in these uncertain times. The institutions that produce higher education in the public sector has gone through dramatic changes for the past 50 years (Baum et al., 2013). Unfortunately, higher education has been the biggest target to budget cuts whenever the financial situations of states become bad. Hence, this study adds more significance to helping decision-makers at the state and institutional levels make wise choices.

Discussion

The results from this study show that state funding for 4-year public higher education institutions impacts outcome variables: student enrollment (FTEE) and graduation rate (GR). However, some of the variables are more significant than others. Amid the COVID-19 pandemic, state finances were negatively affected due to reduced tax collections; hence, this study is crucial to understand what variables related to state budget cuts could significantly predict the two critical outcomes of FTE enrollment (FTEE) and graduation rate (GR) at 4-year public universities in the US. If previous economic downturns are any indication, we should expect many state governments to decrease higher education financing for at least two to three more years to save money.

The findings from this study show if SHEA decreases, FTEE and GR decrease. These results are consistent with previous studies (Zhang, 2009; Hyman, 2017; Goodman & Voltz, 2020). Public institutions rely on increased tuition and other fees to make up for state budget cuts. This study did not study the reasons for a tuition increase, but the results showed that an increase in AUGC decreases FTEE. Although some studies (Bruce & Carruthers, 2014; Gurantz & Odle, 2021) did not find any significant impact of SFA on FTEE and GR, but many previous studies (Alon, 2011; Deming & Walters, 2017; Monarrez et al., 2022; Deming & Dynarski, 2010; Anderson & Zaber, 2021) have reported that SFA affects FTEE and GR. The present study agrees with these previous studies, i.e., an increase in SFA increases FTEE and GR.

This study measured the predictability of the impacts of state appropriations on both institutional and student outcomes. FTE enrollment (FTEE) impacts the institution in terms of tuition revenue and the amount of state funds they receive. The graduation rate (GR) is more of a student outcome that could be attributed to student success. The data analytical and machine learning models developed by this study found that a decrease in state funding variables could cause enrollment declines and lower graduation rate. State appropriations feed directly into revenues for public 4-year higher education institutions. The results from this study show that state budget cuts to public 4-year higher education institutions could predict changes to tuition increases, which in turn result in an increase in average undergraduate charges per FTE (AUGC).

The statistical analysis conducted in chapter 4 showed that state higher education appropriations (SHEA), average undergraduate charges per FTE (AUGC), and student financial aid (SFA) are significant factors for predicting changes in FTE enrollment (FTEE). The results from chapter 4 also showed that state higher education appropriations per FTE (SHEA) and increasing student tuition share as a percentage of per capita income (STSPCI) negatively affected the graduation rate (GR).

Assumptions

- *Assumption 1*: There was a linear relationship between the predictor variables SHEA, AUGC, STSPCI, SHEAGDP, and SFA and the response variables of FTE enrollment (FTEE) and Graduation rate (GR).

- *Assumption 2*: Most of the predictor variables for this study were not highly correlated, which can be seen in table 9 from chapter 4. Hence, multicollinearity does not exist among the predictor variables except between SFA and AUGC.
- *Assumption 3*: The independence of residuals was checked, and there was no significant correlation between the residuals. A Durbin-Watson test was conducted in Python, and the results, as shown in Fig. 21 for FTE enrollment and Fig. 30 for graduation rate, show scores of 2.001 and 1.131, respectively.
- *Assumption 4*: Normality of residuals was checked for both FTEE and GR models, and they have values closer to 1, which means the null hypothesis could be rejected at the 0.05 significance level.

Limitations

This study has its own limitations. The primary limitation of this study is that it is currently based on many unmeasured factors, such as student characteristics, changes in government policies, institutional factors, and historical factors. This study is limited to studying public 4-year undergraduate universities in the United States. This study is limited to the data available with IPEDS postsecondary higher education data. The missing values were replaced with the mean values. There is a lack of previous studies using data analytical models and machine learning algorithms to study the impact of state budget cuts on public higher education institutions. This research study has the limitation of not being able to conduct a deeper analysis of the cohort year graduation rate, as the data could only be obtained six years after the student joins the university. This study is limited to using open-source software tools for this research study, and it is limited to the services available on the platform.

Recommendations

This research study provides an essential direction for 4-year public higher education institutions in the US to make data-driven decisions regarding FTE enrollment (FTEE) and graduation rate (GR). Future studies should consider declining state appropriations that could force institutions to cut spending on student services, academic support, and instruction, leading to a negative impact on student enrollment and graduation rate. Therefore, more research quantifying the effects of these cuts on high school graduates pursuing college needs to be conducted. Further research is required to establish personal factors deterring students from enrolling in college. Impacts of state budget cuts on various student demographics would provide insight into knowing specifically who the target group is to focus on precisely. Once the COVID-19 pandemic ends, a research study is recommended to determine its impact on dependent variables (DV): FTE enrollment (FTEE) and graduation rate (GR).

An increase in state investment in public higher education institutions will help states meet the post pandemic need for a strong workforce, leading to an increase in state income tax revenue and providing several societal benefits (Chakrabarti et al., 2020). The results from this study highlight the importance of state support for public higher education in driving the US economy and future workforce.

State policy decisions often depend on the total amount of funds available. These funds primarily come from taxes collected. The states might be struggling to allocate sufficient budget resources to higher-education funding; however, to achieve the post higher education goal, they must keep it a priority to reduce state budget cuts to higher education. Increasing state support could help achieve a better graduation rate and encourage more FTE enrollment. This study recommends that states make strategic decisions to look for alternative revenues rather than solely depending on uncertain state appropriations. Dougherty and Reddy (2011) recommend improving institutional capacity by having enough human resources and fiscal resources to provide good quality education. Graduation rates and low attrition rates in colleges have become very crucial to keep up with the increased labor needs of the US. Hence, it is time for innovation in the higher education system in the US, and seeking smart growth is essential (Douglass, 2011).

This study recommends for a study on state budget cuts impact on a) full-time faculty vs part-time faculty at public higher education institutions as data were unavailable for some of the years and b) of any specific year impact the corresponding year of cohort. These can be investigated in future work.

REFERENCES

Anggrahini, D., Kurniati, N., & Sukma, P. A. (2021). *Data Analytics for Transforming Toward Smart Supplier Relationship Management (A Case Study in Manufacturing Company).* IEOM Proceedings of the 2nd South American Conf. on Industrial Engr and Operations Mgmt, San Paulo - Virtual, Apr 5-8, pp. 1722-1728. http://www.ieomsociety.org/brazil2020/papers/640.pdf

Avery, C., Howell, J. S., & Page, L. (2014). A Review of the Role of College Counseling, Coaching, and Mentoring on Students' Postsecondary Outcomes. https://paperzz.com/doc/8776439/research-brief---college-board

Badole, M. (2021, May 1). Multiple Linear Regression Using Python and Scikit-learn. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/05/multiple-linear-regression-using-python-and-scikit-learn/

Baum, S., Kurose, C., & McPherson, M. (2013). An overview of American higher education. *The future of children 23*(1), 17-39. https://doi.org/10.1353/foc.2013.0008

Bonderud, D. (2020, June 24). *How Higher Ed Can Use Data Analytics to Boost Admissions.* Technology Solutions That Drive Education. https://edtechmagazine.com/higher/article/2020/06/how-higher-ed-can-use-data-analytics-boost-admissions

Bruce, D.J., & Carruthers, C.K. (2014). Jackpot? The impact of lottery scholarships on enrollment in Tennessee. *Journal of Urban Economics*, *81*, 30-44.

Carmean, C., Kil, D., & Baer, L. (2021, August 10). Why Data Matters for Student Success in a PostPandemic World. *EDUCAUSE*. https://er.educause.edu/articles/2021/8/why-data-matters-for-student-success-in-a-post-pandemic-world

Chakrabarti, R., Gorton, N., & Lovenheim, M. (2020, September). State Investment in Higher Education: Effects on Human Capital Formation, Student Debt, and Long-Term Financial Outcomes of Students. *Federal Reserve Bank of New York Staff Report* (No. 941). https://www.newyorkfed.org/research/staff_reports/sr941

Chakure, A. (2021, December 10). *Data Preprocessing in Python - DataDrivenInvestor. Medium*. https://medium.datadriveninvestor.com/data-preprocessing-3cd01eefd438

Choudhury, A. (2021, October 7). Technology, Management or Data? How To Choose Your Career. *Analytics India Magazine.* https://analyticsindiamag.com/technology-management-or-data-how-to-choose-your-career/

Cote, C. (2021, October 19). 4 Types of Data Analytics to Improve Decision-Making. *Business Insights - Blog.* https://online.hbs.edu/blog/post/types-of-data-analysis

Delaney, J. A., & Doyle, W. R. (2011). State spending on higher education: Testing the balance wheel over time. *Journal of Education Finance*, 36(4), 343-369.

Deming, D. J., & Dynarski, S. M. (2010). Into College, Out of Poverty? Policies to Increase the Postsecondary Attainment of the Poor. In Targeting Investments in Children: Fighting Poverty When Resources are Limited. *National Bureau of economic research,* 10, 283–302.

Dougherty, K. J., & Reddy, V. T. (2011). The impacts of state performance funding systems on higher education institutions: Research literature review and policy recommendations. (Working Paper No. 37). https://eric.ed.gov/?id=ED527751

Douglass, J. A. (2011). Can We Save the College Dream? The death and life of California's public universities. *Boom: A Journal of California*, *1*(2), 25–42. https://doi.org/10.1525/boom.2011.1.2.25

Dynarski, S.M. (2003). Does aid matter? Measuring the effect of student aid on college attendance and completion. *American Economic Review, 93* (1), 279-288. https://doi.org/10.1257/000282803321455287

Ehrenberg, R. G. (2004). Econometric studies of higher education. *Journal of econometrics*, *121*(1-2), 19-37. https://doi.org/10.1016/j.jeconom.2003.10.008

Ferlie, E., Musselin, C., & Andresani, G. (2008). The steering of higher education systems: A public management perspective. *Higher education*, *56*(3), 325. https://doi.org/10.1007/s10734-008-9125-5

Frye, J.R. (2015). *Flexibility for survival: State funding and contingent faculty employment at public higher education institutions.* (Doctoral dissertation, University of Michigan). https://deepblue.lib.umich.edu/bitstream/handle/2027.42/116724/jrfrye_1.pdf?sequence=1&isAllowed=y

GeeksforGeeks. (2021, September 14). *ML | Multiple Linear Regression using Python.* https://www.geeksforgeeks.org/ml-multiple-linear-regression-using-python/

Goodman, S., & Volz, A. H. (2020). Attendance Spillovers between Public and For-Profit Colleges: Evidence from Statewide Variation in Appropriations for Higher Education. *Education Finance and Policy, 15*(3), 428–456. https://doi.org/10.1162/edfp_a_00281

Gurantz, O., & Odle, T. K. (2022). The Impact of Merit Aid on College Choice and Degree Attainment: Reexamining Florida's Bright Futures Program. *Educational Evaluation and Policy Analysis, 44*(1), 79–104. https://doi.org/10.3102/01623737211030489

Hanushek, E. A., & Woessmann, L. (2008). The Role of Cognitive Skills in Economic Development. *Journal of Economic Literature, 46*(3), 607–668. https://doi.org/10.1257/jel.46.3.607

Hoseini, C. (2020). *Leveraging machine learning to identify quality issues in the Medicaid claim adjudication process* (Doctoral dissertation, Indiana State University). https://www.proquest.com/openview/6e8b80a1aae0b0e8ce6d0e8013aa571d/1?pq-origsite=gscholar&cbl=18750&diss=y.

Hoseini, C., Badar, M.A., Shahhosseini, A.M., & Kluse, C.J. (2021). *A Review of Machine Learning Methods Applicable to Quality Issues.* IEOM Proceedings of the 11th Int. Conf. on Industrial Engr and Operations Mgmt, Singapore - Virtual, Mar 7-10, pp. 1225-1240. http://www.ieomsociety.org/singapore2021/papers/250.pdf.

Husted, T.A., & Kenny, L.W. (2018). The effect of reduced support from state government on research in state universities. *Journal of Education Finance, 44*(2), 164-174.

Hyman, J. (November 2017). Does money matter in the long run? Effects of school spending on educational attainment. *American Economic Journal: Economic Policy 9*(4), 256-80. https://doi.org/10.1257/pol.20150249.

Jaquette, O., & Curs, B. R. (2015). Creating the Out-of-State University: Do Public Universities Increase Nonresident Freshman Enrollment in Response to Declining State Appropriations? *Research in Higher Education, 56*(6), 535–565. https://doi.org/10.1007/s11162-015-9362-2

Jimmerson, A. (2021, May 26). New Report Finds Despite Eight Years of Increases in State Support, Public Colleges Entered the 2020 Recession with Historically Low Funding. *State Higher Education Executive Officers Association.* https://sheeo.org/new-report-finds-despite-eight-years-of-increases-in-state-support-public-colleges-entered-the-2020-recession-with-historically low-funding/

Johnson, J. A. (2013). Ethics of Data Mining and Predictive Analytics in Higher Education. *SSRN Electronic journal*. https://doi.org/10.2139/ssrn.2156058

Kachchi, V., & Kothiya, Y. (2021, October 1). *4 Type of Data Analytics Every Analyst Should Know-Descriptive, Diagnostic, Predictive, Prescriptive*. Medium. https://medium.com/colearning-lounge/types-of-data-analytics-descriptive-diagnostic-predictive-prescriptive-922654ce8f8f

Katrak, P., Bialocerkowski, A. E., Massy-Westropp, N., Kumar, V. S., & Grimmer, K. A. (2004). A systematic review of the content of critical appraisal tools. *BMC Medical Research Methodology, 4*(1). https://doi.org/10.1186/1471-2288-4-22

Killada, P. (2017). *Data Analytics using Regression Models for Health Insurance Market Place Data* (Master Thesis, University of Toledo). https://etd.ohiolink.edu/apexprod/rws_etd/send_file/send?accession=toledo1501721348961437&disposition=inline

Kiu, C. C. (2018). *Data Mining Analysis on Student's Academic Performance through Exploration of Student's Background and Social Activities*. 2018 Fourth International Conference on Advances in Computing, Communication & Automation (ICACCA), 1–5. https://doi.org/10.1109/ICACCAF.2018.8776809

Laderman, S., & Tandberg, D. (2021). SHEEO analysis of fiscal year 2021 state funding for higher education. *State Higher Education Executive Officers Association*. https://sheeo.org/wpcontent/uploads/2021/03/SHEEO_Analysis_FiscalYear2021_State_Funding.pdf

Monarrez, T., Brian K., & Matthew C. (2022). The Effect of Charter Schools on School Segregation. *American Economic Journal: Economic Policy, 14* (1): 301 - 40. https://doi.org/10.1257/pol.20190682

Montgomery, D. C. (2012). *Design and analysis of experiments.* (8th ed.). John Wiley & Sons.

National Center for Education Statistics (2020). Integrated Postsecondary Education Data System (IPEDS), Graduation rate component final data (2002 - 2019) and provisional data. *U.S. Department of Education.* https://nces.ed.gov/ipeds/TrendGenerator/app/buildtable/7/19

Oreopoulos, P., & Petronijevic, U. (2013). *Making college worth it: A review of research on the returns to higher education.* http://www.nber.org/papers/w19053

Parimi, R. M., & Babu, S. (2020). *Critical Analysis of Software Vulnerabilities through Data Analytics.* IEOM Proceedings of the 10th Int. Conf. on Industrial Engr and Operations Mgmt, Dubai – Virtual, Mar 10 -12, pp. 923-934. http://www.ieomsociety.org/ieom2020/papers/810.pdf

Rich, F.A., Shahhosseini, A.M., Badar, M.A., & Kluse, C.J. (n.d.). A Study on Factors Affecting the Wear of Steel Track Undercarriage, *Journal of Quality in Maintenance Engineering*, submitted.

Robotham, D., & Julian, C. (2006). Stress and the higher education student: a critical review of the literature. *Journal of further and higher education, 30*(02), 107-117. https://doi.org/10.1080/03098770600617513

Rothstein, J., & Schanzenbach, D. W. (2021). Does Money Still Matter? Attainment and Earnings Effects of Post-1990 School Finance Reforms. *SSRN Electronic Journal.* https://doi.org/10.2139/ssrn.3909622

Roy, B. (2021, December 14). *All about Feature Scaling - Towards Data Science*. Medium. https://towardsdatascience.com/all-about-feature-scaling-bcc0ad75cb35

SHEF Report. (2021, July 23). State Higher Education Finance Report 2021. *State Higher Education Executive Officers Association*. https://shef.sheeo.org/report/

Spear, E. (2019, October 30). Data Analytics in Higher Ed: 4 Tips to Find the Best Software. *Precision Campus*. https://precisioncampus.com/blog/data-analytics-higher-education/

Stevenson, K. R. (2006). School Size and Its Relationship to Student Outcomes and School Climate: A Review and Analysis of Eight South Carolina State-Wide Studies. *National Clearinghouse for Educational Facilities*. https://eric.ed.gov/?id=ED495953

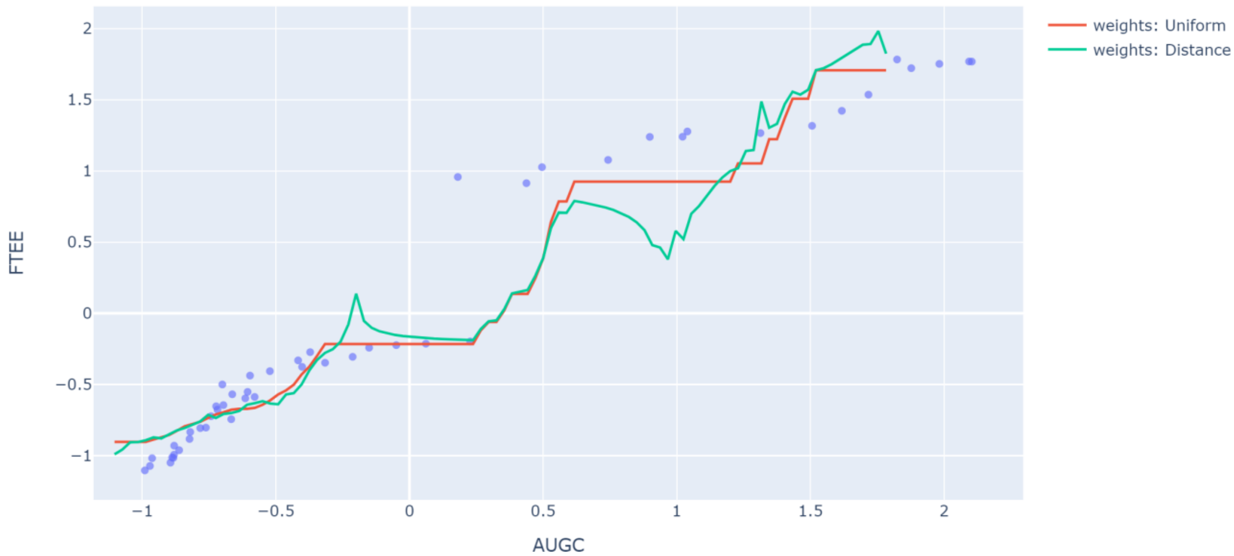
Summerfield, N., Zhang, J., Motiwalla, L., Mai, T., & Mazza, K. (2018). *Using Analytics to Predict Project Management Success*. IEOM Proceedings of the 3rd North American conf. on Industrial Engr and Operations Mgmt, Washington DC, Sep 27 -29, pp. 1414-1419. http://ieomsociety.org/dc2018/papers/387.pdf

Webber, D. A. (2017). State divestment and tuition at public institutions. *Economics of Education Review*. https://doi.org/10.1016/j.econedurev.2017.07.007

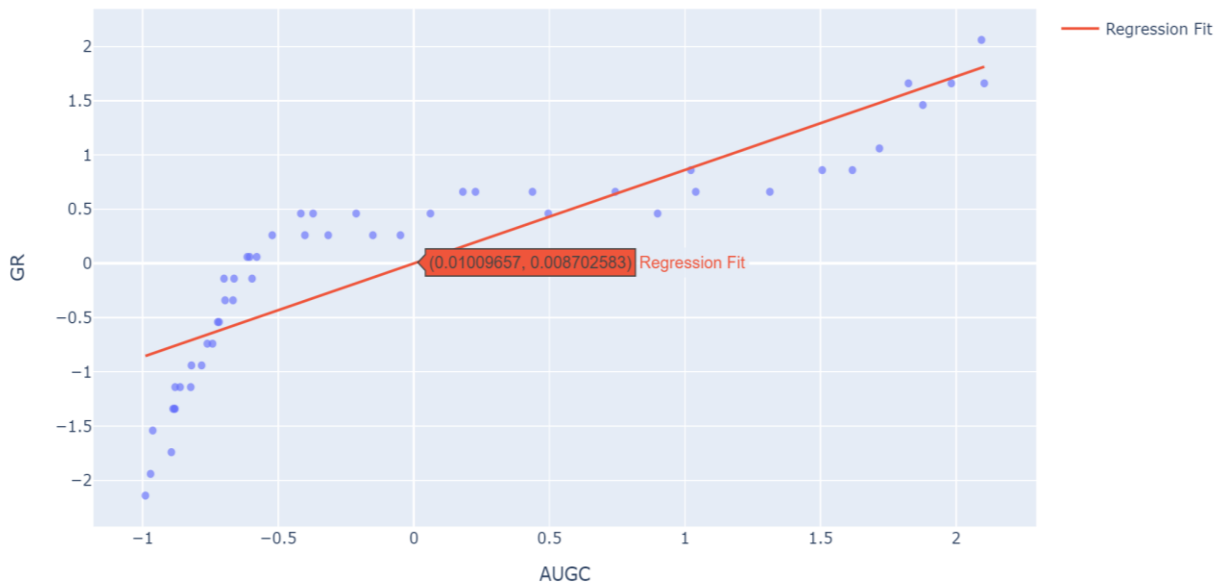
Zhang. L. (2009). Does State Funding Affect Graduation rate at Public Four-Year Colleges and Universities? *Educational Policy, 23*(5), 714–731. https://doi.org/10.1177/0895904808321270

Zhao, B. (2018). Disinvesting in the future? A comprehensive examination of the effects of state appropriations for public higher education. *SSRN Electronic Journal*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3197605

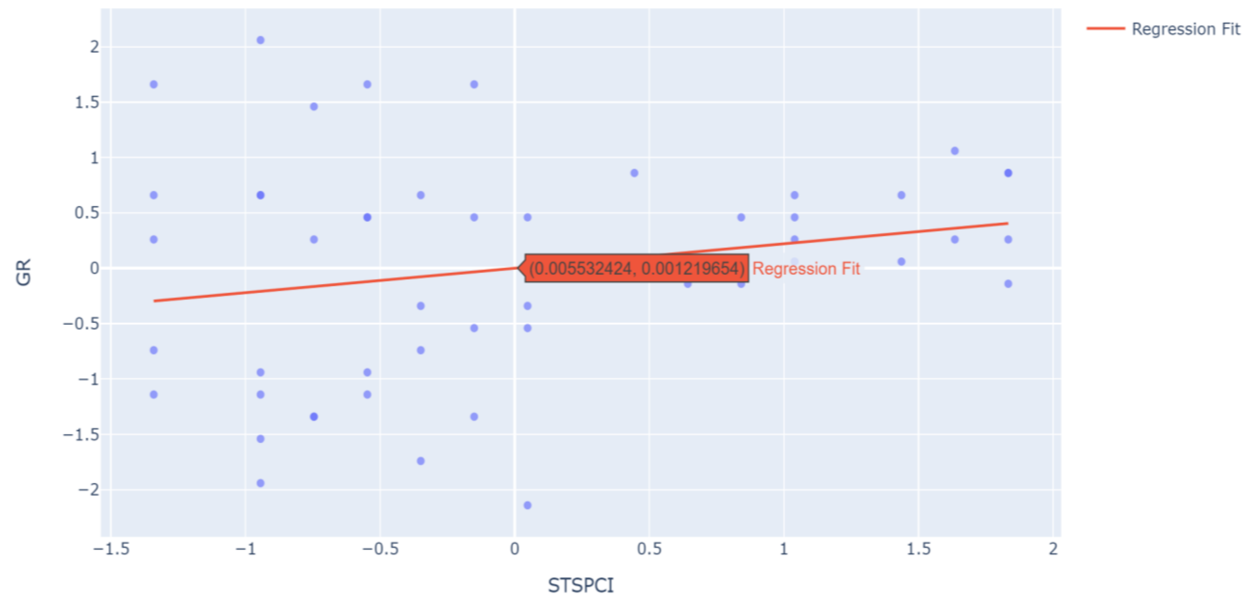
APPENDIX A: FTEE vs AUGC Expected vs Actual Values



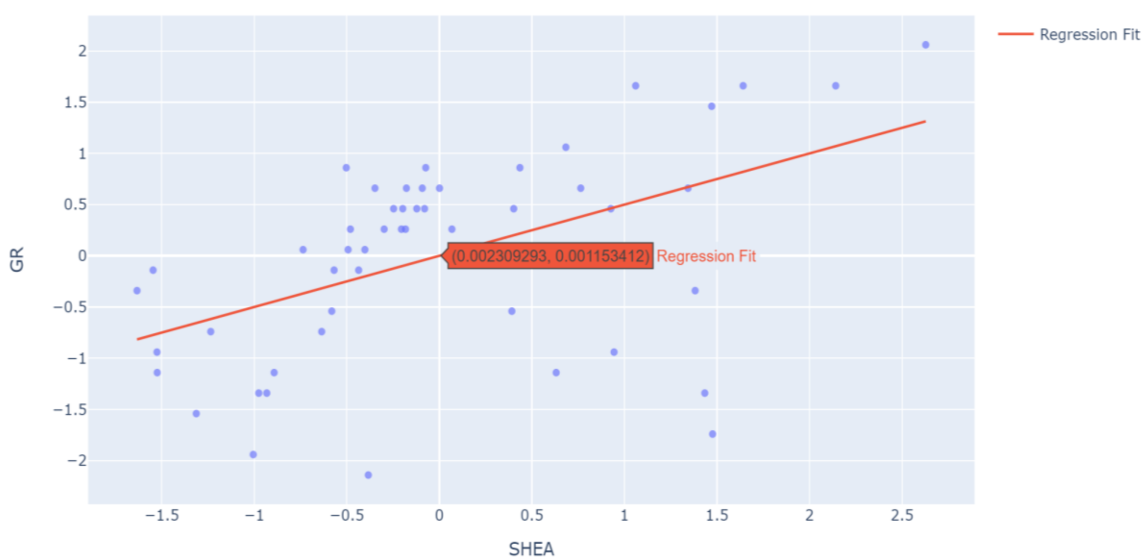
APPENDIX B: Regression Fit for GR and AUGC



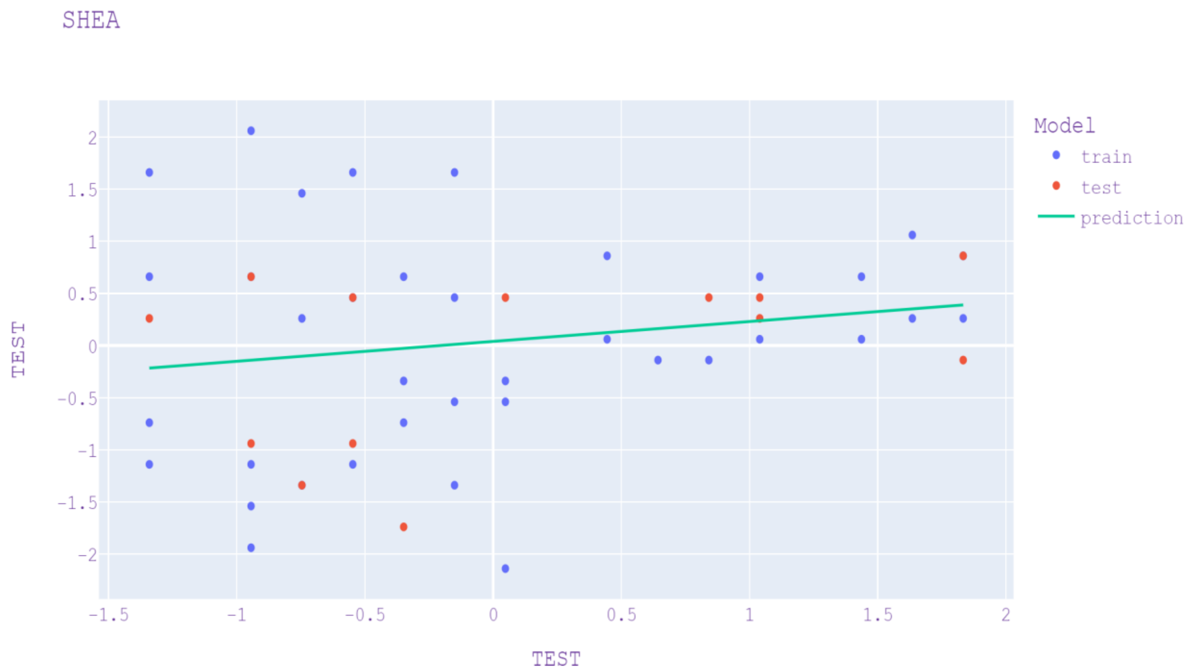
APPENDIX C: Regression Fit for GR and STSPCI



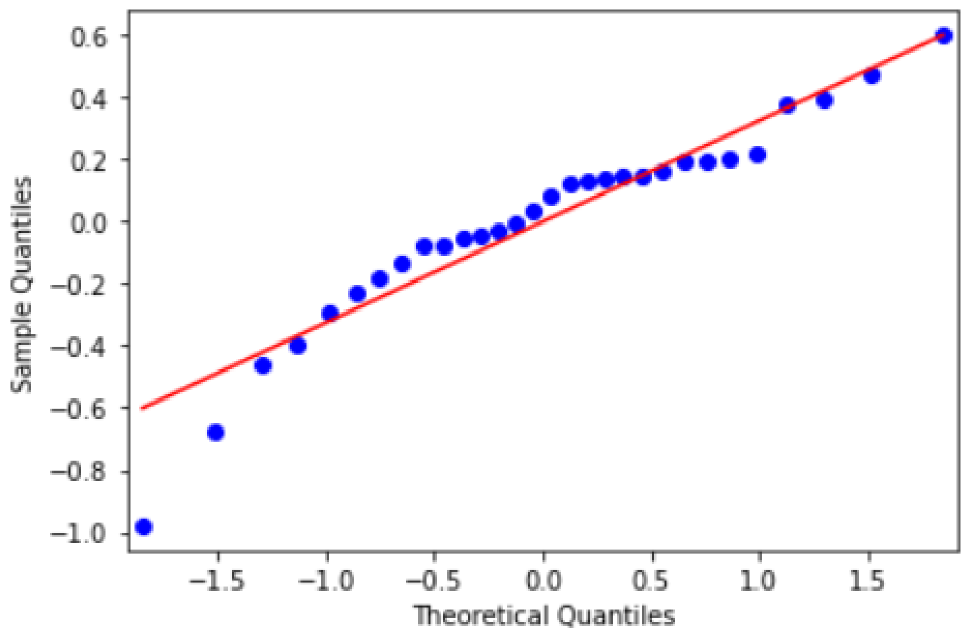
APPENDIX D: Regression Fit for GR and SHEA



APPENDIX E: SHEA Prediction Scatter Plot



APPENDIX F: Q-Q Plot



APPENDIX G: Python Library

```
import pandas as pd
import numpy as np
import statsmodels.api as sm
import matplotlib.pyplot as plt
import seaborn as sns
from pandas import read_csv
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, r2_Score
from sklearn import linear_model
from sklearn.svm import SVC
from sklearn.model_selection import train_test_split
from sklearn.model_selection import cross_val_score
from sklearn.model_selection import StratifiedKFold
from sklearn.metrics import classification_report
from sklearn.metrics import confusion_matrix
from sklearn.metrics import accuracy_score
from sklearn.linear_model import LogisticRegression
from pandas plotting import scatter_matrix
from sklearn.tree import DecisionTreeClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
from sklearn.naive_bayes import GaussianNB
```

APPENDIX H: Standardization of Data

```
import sklearn
from sklearn.preprocessing import StandardScaler
sc_X = StandardScaler()
X_train = sc_X.fit_transform(X_train)
X_test = sc_X.transform(X_test)
```

APPENDIX I: Python Commands executed

Data Standard Scaling

```
import sklearn
from sklearn.datasets import load_iris
from sklearn import preprocessing
from sklearn.preprocessing import StandardScaler
data = load_iris()
X_data = data.data
target = data.target
standard = preprocessing.scale(X_data)
print(standard)
scale= StandardScaler()
X_data = data.data
target = data.target
scaled_data = scale.fit_transform(X_data)
print(scaled_data)
```

Handling the Missing Data

```
imputer = SimpleImputer(missing_values=np.nan, strategy='mean') imputer = imputer.fit(X[:, 1:])
X[:, 1:] = imputer.transform(X[:, 1:])
```

Training and Testing the FTEE Model

```
X = df[['SHEA','AUGC','SFA']]
Y = df['FTEE']
X_train, X_test, Y_train, Y_test = sklearn.model_selection.train_test_split(X, Y, test_size = 0.20, random_state = 0)
sc_X = StandardScaler()
X_train = sc_X.fit_transform(X_train)
X_test = sc_X.transform(X_test)
model = sm. OLS(Y_train, sm.add_constant(X_train)).fit()
Y_pred = model.predict(sm.add_constant(X_test))
print_model = model.summary()
print(print_model)
```

Training and Testing the GR Model

```
X = df[['SHEA','STSPCI']]
Y = df['GR']
X_train, X_test, Y_train, Y_test = sklearn.model_selection.train_test_split(X, Y, test_size = 0.20, random_state = 0)
sc_X = StandardScaler()
X_train = sc_X.fit_transform(X_train)
X_test = sc_X.transform(X_test)
model = sm. OLS(Y_train, sm.add_constant(X_train)).fit()
Y_pred = model.predict(sm.add_constant(X_test))
print_model = model.summary()
print(print_model)
```

```
ks = sm. OLS(X_train, Y_train)
ks_res =ks.fit()
from sklearn.linear_model import LinearRegression
LR = LinearRegression()
LR.fit(X_train,Y_train)
Y_prediction = LR.predict(X_test)
print (Y_prediction)
from sklearn.metrics import r2_score
from sklearn.metrics import mean_squared_error
score=r2_score(Y_test,Y_prediction)
print(score)
mean_squared_error=mean_squared_error(Y_test,Y_prediction)
print(mean_squared_error)
rgr = RandomForestRegressor(n_estimators=100)
svr = LinearSVR()

rgr.fit(X_train, Y_train)
svr.fit(X_train, Y_train)

print("Mean squared error with 95% confidence interval:")
print("")
```

APPENDIX J: Python Code for heatmap and box plots

```
import numpy as np
import pandas as pd
import plotly.express as px
import plotly.graph_objects as go
from sklearn.model_selection import GridSearchCV
from sklearn.tree import DecisionTreeRegressor

N_FOLD = 6

# Load and shuffle dataframe

df = pd.read_csv (r'C:\Users\guraj\Downloads\bookpython2.csv')
df = df.sample(frac=1, random_state=0)
X = df[['SHEA','AUGC','SFA']]
y = df['FTEE']

# Define and fit the grid
model = DecisionTreeRegressor()
param_grid = {
    'criterion': ['mse', 'friedman_mse', 'mae'],
    'max_depth': range(2, 5)
}
grid = GridSearchCV(model, param_grid, cv=N_FOLD)
grid.fit(X, y)
grid_df = pd.DataFrame(grid.cv_results_)

# Convert the wide format of the grid into the long format
# accepted by plotly.express
melted = (
    grid_df
    .rename(columns=lambda col: col.replace('param_', ''))
    .melt(
        value_vars=[f'split{i}_test_score' for i in range(N_FOLD)],
        id_vars=['mean_test_score', 'mean_fit_time', 'criterion', 'max_depth'],
        var_name="cv_split",
        value_name="r_squared"
    )
)

# Format the variable names for simplicity
melted['cv_split'] = (
    melted['cv_split']
    .str.replace('_test_score', '')
    .str.replace('split', '')
)

# Single function call to plot each figure
fig_hmap = px.density_heatmap(
    melted, x="max_depth", y='criterion',
    histfunc="sum", z="r_squared",
    title='Grid search results on individual fold',
    hover_data=['mean_fit_time'],
    facet_col="cv_split", facet_col_wrap=3,
    labels={'mean_test_score': "mean_r_squared"}
)

fig_box = px.box(
    melted, x='max_depth', y='r_squared',
    title='Grid search results ',
    hover_data=['mean_fit_time'],
    points='all',
    color="criterion",
    hover_name='cv_split',
    labels={'mean_test_score': "mean_r_squared"}
)

# Display
fig_hmap.show()
fig_box.show()
```

APPENDIX K: Python Code for K=7 Folds

```
import numpy as np
import pandas as pd
import plotly.express as px
import plotly.graph_objects as go
from sklearn.linear_model import LassoCV

N_FOLD = 7

# Load and preprocess the data
df = pd.read_csv (r'C:\Users\guraj\Downloads\bookpython2.csv')
X = df.drop(columns=['FTEE', 'GR'])
X = pd.get_dummies(X, columns=['SHEA','STSPCI'])
y = df['GR']

# Train model to predict GR
model = LassoCV(cv=N_FOLD, normalize=True)
model.fit(X, y)
mean_alphas = model.mse_path_.mean(axis=-1)
from sklearn.pipeline import make_pipeline


fig = go.Figure([
    go.Scatter(
        x=model.alphas_, y=model.mse_path_[:, i],
        name=f"Fold: {i+1}", opacity=.5, line=dict(dash='dash'),
        hovertemplate="alpha: %{x} <br>MSE: %{y}"
    )
    for i in range(N_FOLD)
])
fig.add_traces(go.Scatter(
    x=model.alphas_, y=mean_alphas,
    name='Mean', line=dict(color='black', width=3),
    hovertemplate="alpha: %{x} <br>MSE: %{y}",
))

fig.add_shape(
    type="line", line=dict(dash='dash'),
    x0=model.alpha_, y0=0,
    x1=model.alpha_, y1=1,
    yref='paper'
)

fig.update_layout(
    xaxis_title='alpha',
    xaxis_type="log",
    yaxis_title="Mean Square Error (MSE)"
)
fig.show()
```

APPENDIX L: Python Code for Residual vs Prediction for FTEE

```
import numpy as np
import plotly.express as px
import plotly.graph_objects as go
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split

df = pd.read_csv (r'C:\Users\guraj\Downloads\bookpython2.csv')

# Split data into training and test splits
train_idx, test_idx = train_test_split(df.index, test_size=.20, random_state=0)
df['split'] = 'train'
df.loc[test_idx, 'split'] = 'test'

X = df[['SHEA','AUGC','SFA']]
y = df['FTEE']
X_train = df.loc[train_idx, ['SHEA','AUGC','SFA']]
y_train = df.loc[train_idx, 'FTEE']

# Condition the model on SHEA, AUGC, and SFA, predict the FTEE
model = LinearRegression()
model.fit(X_train, y_train)
df['prediction'] = model.predict(X)
df['residual'] = df['prediction'] - df['FTEE']

fig = px.scatter(
    df, x='prediction', y='residual',
    marginal_y='violin',
    color='split', trendline='ols'
)
fig.show()
```

APPENDIX M: Code for Residual vs Prediction for GR

```
import numpy as np
import plotly.express as px
import plotly.graph_objects as go
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split

df = pd.read_csv (r'C:\Users\guraj\Downloads\bookpython2.csv')

# Split data into training and test splits
train_idx, test_idx = train_test_split(df.index, test_size=.20, random_state=0)
df['split'] = 'train'
df.loc[test_idx, 'split'] = 'test'

X = df[['SHEA','STSPCI']]
y = df['GR']
X_train = df.loc[train_idx, ['SHEA','STSPCI']]
y_train = df.loc[train_idx, 'GR']

# Condition the model on SHEA and STSPCI, predict the GR
model = LinearRegression()
model.fit(X_train, y_train)
df['prediction'] = model.predict(X)
df['residual'] = df['prediction'] - df['GR']

fig = px.scatter(
    df, x='prediction', y='residual',
    marginal_y='violin',
    color='split', trendline='ols'
)
fig.show()
```

APPENDIX N: Python Code for prediction vs dependent variable FTEE

```
import plotly.express as px
import plotly.graph_objects as go
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split

df = pd.read_csv (r'C:\Users\guraj\Downloads\bookpython2.csv')

# Split data into training and test splits
train_idx, test_idx = train_test_split(df.index, test_size=.20, random_state=0)
df['split'] = 'train'
df.loc[test_idx, 'split'] = 'test'

X = df[['SHEA','AUGC','SFA']]
y = df['FTEE']
X_train = df.loc[train_idx, ['SHEA','AUGC','SFA']]
y_train = df.loc[train_idx, 'FTEE']

# Condition the model on SHEA, AUGC, and SFA, predict the FTEE
model = LinearRegression()
model.fit(X_train, y_train)
df['prediction'] = model.predict(X)

fig = px.scatter(
    df, x='FTEE', y='prediction',
    color='split', trendline='ols'
)
fig.update_traces(histnorm='probability', selector={'type':'histogram'})
fig.add_shape(
    type="line", line=dict(dash='dash'),
    x0=y.min(), y0=y.min(),
    x1=y.max(), y1=y.max()
)

fig.show()
```

APPENDIX O: Python Code for prediction vs dependent variable GR

```
import plotly.express as px
import plotly.graph_objects as go
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split

df = pd.read_csv (r'C:\Users\guraj\Downloads\bookpython2.csv')

# Split data into training and test splits
train_idx, test_idx = train_test_split(df.index, test_size=.20, random_state=0)
df['split'] = 'train'
df.loc[test_idx, 'split'] = 'test'

X = df[['SHEA','STSPCI']]
y = df['GR']
X_train = df.loc[train_idx, ['SHEA','STSPCI']]
y_train = df.loc[train_idx, 'GR']

# Condition the model on SHEA and STSPCI, predict the GR
model = LinearRegression()
model.fit(X_train, y_train)
df['prediction'] = model.predict(X)

fig = px.scatter(
    df, x='GR', y='prediction',
    color='split', trendline='ols'
)
fig.update_traces(histnorm='probability', selector={'type':'histogram'})
fig.add_shape(
    type="line", line=dict(dash='dash'),
    x0=y.min(), y0=y.min(),
    x1=y.max(), y1=y.max()
)

fig.show()
```

APPENDIX P: Actual Data and Scaled Data

Actual Data before preprocessing

SHEA	AUGC	STSPCI	SHEAGDP	SFA	FTEE	GR
5810	4171	37	48	548	86126	40
5810	4273	34	48	548	89716	42
4577	4683	35	57	251	91803	41
4752	4316	32	62	253	92127	44
4987	4716	33	60	259	92757	44
4567	4741	36	61	265	93479	44
5110	4754	33	59	261	93408	44
5236	4856	30	61	266	95529	45
6319	4757	34	60	261	97633	45
6091	5064	32	58	269	103989	46
5581	5078	32	57	272	105791	46
6571	5279	34	57	270	109917	45
5298	5391	30	57	273	110023	47
5764	5494	35	54	275	113114	47
6126	5596	37	52	263	115835	48
6324	5623	36	52	291	119357	48
5458	5741	37	54	300	123477	49
6924	5897	35	53	302	125916	49
6352	5918	40	52	357	127518	50
6234	5920	41	47	362	126074	52
7103	6225	42	47	375	137630	51
5926	6178	39	45	381	149627	50
7928	6273	46	41	387	150204	51
5362	6365	44	42	399	150274	51
6532	6670	46	43	395	152253	52
7208	7314	45	42	416	154258	52
6357	7334	41	42	435	148893	52
5784	7773	42	43	424	155321	53
6194	8324	42	41	475	138929	53
6281	8956	36	43	645	141103	53
6292	9521	33	41	604	143162	52
5987	9196	30	42	638	144358	52
6288	9787	34	43	634	145047	53
6272	9656	32	42	586	146072	54
5418	9821	32	42	745	148095	54
5364	10075	34	41	722	150789	55
4556	10791	30	41	805	177524	54
4494	13429	35	43	825	198893	53
4897	14262	37	43	745	199521	54
5139	15014	42	42	732	205342	54
5246	15918	39	42	847	210111	55
4146	17505	46	41	860	209543	55
5447	17474	44	41	825	217249	54
5733	18100	46	43	910	219406	55
6581	18632	45	41	961	217357	56
6088	19204	36	43	959	219936	59
5494	19488	33	42	1082	215873	58
6053	20049	30	42	1057	217231	59
5816	20698	34	41	1126	219946	59
6624	20643	32	41	1278	220022	61

Scaled Data after Preprocessing

```
[[-9.00681170e-01 1.01900435e+00 -1.34022653e+00 -1.31544430e+00]
 [-1.14301691e+00 -1.31979479e-01 -1.34022653e+00 -1.31544430e+00]
 [-1.38535265e+00 3.28414053e-01 -1.39706395e+00 -1.31544430e+00]
 [-1.50652052e+00 9.82172869e-02 -1.28338910e+00 -1.31544430e+00]
 [-1.02184904e+00 1.24920112e+00 -1.34022653e+00 -1.31544430e+00]
 [-5.37177559e-01 1.93979142e+00 -1.16971425e+00 -1.05217993e+00]
 [-1.50652052e+00 7.88807586e-01 -1.34022653e+00 -1.18381211e+00]
 [-1.02184904e+00 7.88807586e-01 -1.28338910e+00 -1.31544430e+00]
 [-1.74885626e+00 -3.62176246e-01 -1.34022653e+00 -1.31544430e+00]
 [-1.14301691e+00 9.82172869e-02 -1.28338910e+00 -1.44707648e+00]
 [-5.37177559e-01 1.47939788e+00 -1.28338910e+00 -1.31544430e+00]
 [-1.26418478e+00 7.88807586e-01 -1.22655167e+00 -1.31544430e+00]
 [-1.26418478e+00 -1.31979479e-01 -1.34022653e+00 -1.44707648e+00]
 [-1.87002413e+00 -1.31979479e-01 -1.51073881e+00 -1.44707648e+00]
 [-5.25060772e-02 2.16998818e+00 -1.45390138e+00 -1.31544430e+00]
 [-1.73673948e-01 3.09077525e+00 -1.28338910e+00 -1.05217993e+00]
 [-5.37177559e-01 1.93979142e+00 -1.39706395e+00 -1.05217993e+00]
 [-9.00681170e-01 1.01900435e+00 -1.34022653e+00 -1.18381211e+00]
 [-1.73673948e-01 1.70959465e+00 -1.16971425e+00 -1.18381211e+00]
 [-9.00681170e-01 1.70959465e+00 -1.28338910e+00 -1.18381211e+00]
 [-5.37177559e-01 7.88807586e-01 -1.16971425e+00 -1.31544430e+00]
 [-9.00681170e-01 1.47939788e+00 -1.28338910e+00 -1.05217993e+00]
 [-1.50652052e+00 1.24920112e+00 -1.56757623e+00 -1.31544430e+00]
 [-9.00681170e-01 5.58610819e-01 -1.16971425e+00 -9.20547742e-01]
 [-1.26418478e+00 7.88807586e-01 -1.05603939e+00 -1.31544430e+00]
 [-1.02184904e+00 -1.31979479e-01 -1.22655167e+00 -1.31544430e+00]
 [-1.02184904e+00 7.88807586e-01 -1.22655167e+00 -1.05217993e+00]
 [-7.79513300e-01 1.01900435e+00 -1.28338910e+00 -1.31544430e+00]
 [-7.79513300e-01 7.88807586e-01 -1.34022653e+00 -1.31544430e+00]
 [-1.38535265e+00 3.28414053e-01 -1.22655167e+00 -1.31544430e+00]
 [-1.26418478e+00 9.82172869e-02 -1.22655167e+00 -1.31544430e+00]
 [-5.37177559e-01 7.88807586e-01 -1.28338910e+00 -1.05217993e+00]
 [-7.79513300e-01 2.40018495e+00 -1.28338910e+00 -1.44707648e+00]
 [-4.16009689e-01 2.63038172e+00 -1.34022653e+00 -1.31544430e+00]
 [-1.14301691e+00 9.82172869e-02 -1.28338910e+00 -1.31544430e+00]
 [-1.02184904e+00 3.28414053e-01 -1.45390138e+00 -1.31544430e+00]
 [-4.16009689e-01 1.01900435e+00 -1.39706395e+00 -1.31544430e+00]
 [-1.14301691e+00 1.24920112e+00 -1.34022653e+00 -1.44707648e+00]
 [-1.74885626e+00 -1.31979479e-01 -1.39706395e+00 -1.31544430e+00]
 [-9.00681170e-01 7.88807586e-01 -1.28338910e+00 -1.31544430e+00]
 [-1.02184904e+00 1.01900435e+00 -1.39706395e+00 -1.18381211e+00]
 [-1.62768839e+00 -1.74335684e+00 -1.39706395e+00 -1.18381211e+00]
 [-1.74885626e+00 3.28414053e-01 -1.39706395e+00 -1.31544430e+00]
 [-1.02184904e+00 1.01900435e+00 -1.22655167e+00 -7.88915558e-01]
 [-9.00681170e-01 1.70959465e+00 -1.05603939e+00 -1.05217993e+00]
 [-1.26418478e+00 -1.31979479e-01 -1.34022653e+00 -1.18381211e+00]
 [-9.00681170e-01 1.70959465e+00 -1.22655167e+00 -1.31544430e+00]
 [-1.50652052e+00 3.28414053e-01 -1.34022653e+00 -1.31544430e+00]
 [-6.58345429e-01 1.47939788e+00 -1.28338910e+00 -1.31544430e+00]
 [-1.02184904e+00 5.58610819e-01 -1.34022653e+00 -1.31544430e+00]
 [ 1.40150837e+00 3.28414053e-01 5.35408562e-01 2.64141916e-01]
 [ 6.74501145e-01 3.28414053e-01 4.21733708e-01 3.95774101e-01]
```

```
[ 1.28034050e+00 9.82172869e-02 6.49083415e-01 3.95774101e-01]
[-4.16009689e-01 -1.74335684e+00 1.37546573e-01 1.32509732e-01]
[ 7.95669016e-01 -5.92373012e-01 4.78571135e-01 3.95774101e-01]
[-1.73673948e-01 -5.92373012e-01 4.21733708e-01 1.32509732e-01]
[ 5.53333275e-01 5.58610819e-01 5.35408562e-01 5.27406285e-01]
[-1.14301691e+00 -1.51316008e+00 -2.60315415e-01 -2.62386821e-01]
[ 9.16836886e-01 -3.62176246e-01 4.78571135e-01 1.32509732e-01]
[-7.79513300e-01 -8.22569778e-01 8.07091462e-02 2.64141916e-01]
[-1.02184904e+00 -2.43394714e+00 -1.46640561e-01 -2.62386821e-01]
[ 6.86617933e-02 -1.31979479e-01 2.51221427e-01 3.95774101e-01]
[ 1.89829664e-01 -1.97355361e+00 1.37546573e-01 -2.62386821e-01]
[ 3.10997534e-01 -3.62176246e-01 5.35408562e-01 2.64141916e-01]
[-2.94841818e-01 -3.62176246e-01 -8.98031345e-02 1.32509732e-01]
[ 1.03800476e+00 9.82172869e-02 3.64896281e-01 2.64141916e-01]
[-2.94841818e-01 -1.31979479e-01 4.21733708e-01 3.95774101e-01]
[-5.25060772e-02 -8.22569778e-01 1.94384000e-01 -2.62386821e-01]
[ 4.32165405e-01 -1.97355361e+00 4.21733708e-01 3.95774101e-01]
[-2.94841818e-01 -1.28296331e+00 8.07091462e-02 -1.30754636e-01]
[ 6.86617933e-02 3.28414053e-01 5.92245988e-01 7.90670654e-01]
[ 3.10997534e-01 -5.92373012e-01 1.37546573e-01 1.32509732e-01]
[ 5.53333275e-01 -1.28296331e+00 6.49083415e-01 3.95774101e-01]
[ 3.10997534e-01 -5.92373012e-01 5.35408562e-01 8.77547895e-04]
[ 6.74501145e-01 -3.62176246e-01 3.08058854e-01 1.32509732e-01]
[ 9.16836886e-01 -1.31979479e-01 3.64896281e-01 2.64141916e-01]
[ 1.15917263e+00 -5.92373012e-01 5.92245988e-01 2.64141916e-01]
[ 1.03800476e+00 -1.31979479e-01 7.05920842e-01 6.59038469e-01]
[ 1.89829664e-01 -3.62176246e-01 4.21733708e-01 3.95774101e-01]
[-1.73673948e-01 -1.05276654e+00 -1.46640561e-01 -2.62386821e-01]
[-4.16009689e-01 -1.51316008e+00 2.38717193e-02 -1.30754636e-01]
[-4.16009689e-01 -1.51316008e+00 -3.29657076e-02 -2.62386821e-01]
[-5.25060772e-02 -8.22569778e-01 8.07091462e-02 8.77547895e-04]
[ 1.89829664e-01 -8.22569778e-01 7.62758269e-01 5.27406285e-01]
[-5.37177559e-01 -1.31979479e-01 4.21733708e-01 3.95774101e-01]
[ 1.89829664e-01 7.88807586e-01 4.21733708e-01 5.27406285e-01]
[ 1.03800476e+00 9.82172869e-02 5.35408562e-01 3.95774101e-01]
[ 5.53333275e-01 -1.74335684e+00 3.64896281e-01 1.32509732e-01]
[-2.94841818e-01 -1.31979479e-01 1.94384000e-01 1.32509732e-01]
[-4.16009689e-01 -1.28296331e+00 1.37546573e-01 1.32509732e-01]
[-4.16009689e-01 -1.05276654e+00 3.64896281e-01 8.77547895e-04]
[ 3.10997534e-01 -1.31979479e-01 4.78571135e-01 2.64141916e-01]
[-5.25060772e-02 -1.05276654e+00 1.37546573e-01 8.77547895e-04]
[-1.02184904e+00 -1.74335684e+00 -2.60315415e-01 -2.62386821e-01]
[-2.94841818e-01 -8.22569778e-01 2.51221427e-01 1.32509732e-01]
[-1.73673948e-01 -1.31979479e-01 2.51221427e-01 8.77547895e-04]
[-1.73673948e-01 -3.62176246e-01 2.51221427e-01 1.32509732e-01]
[ 4.32165405e-01 -3.62176246e-01 3.08058854e-01 1.32509732e-01]
[-9.00681170e-01 -1.28296331e+00 -4.30827696e-01 -1.30754636e-01]
[-1.73673948e-01 -5.92373012e-01 1.94384000e-01 1.32509732e-01]
[ 5.53333275e-01 5.58610819e-01 1.27429511e+00 1.71209594e+00]
[-5.25060772e-02 -8.22569778e-01 7.62758269e-01 9.22302838e-01]
[ 1.52267624e+00 -1.31979479e-01 1.21745768e+00 1.18556721e+00]
[ 5.53333275e-01 -3.62176246e-01 1.04694540e+00 7.90670654e-01]
[ 7.95669016e-01 -1.31979479e-01 1.16062026e+00 1.31719939e+00]
[ 2.12851559e+00 -1.31979479e-01 1.61531967e+00 1.18556721e+00]
```

```
 [-1.14301691e+00 -1.28296331e+00 4.21733708e-01 6.59038469e-01]
 [ 1.76501198e+00 -3.62176246e-01 1.44480739e+00 7.90670654e-01]
 [ 1.03800476e+00 -1.28296331e+00 1.16062026e+00 7.90670654e-01]
 [ 1.64384411e+00 1.24920112e+00 1.33113254e+00 1.71209594e+00]
 [ 7.95669016e-01 3.28414053e-01 7.62758269e-01 1.05393502e+00]
 [ 6.74501145e-01 -8.22569778e-01 8.76433123e-01 9.22302838e-01]
 [ 1.15917263e+00 -1.31979479e-01 9.90107977e-01 1.18556721e+00]
 [-1.73673948e-01 -1.28296331e+00 7.05920842e-01 1.05393502e+00]
 [-5.25060772e-02 -5.92373012e-01 7.62758269e-01 1.58046376e+00]
 [ 6.74501145e-01 3.28414053e-01 8.76433123e-01 1.44883158e+00]
 [ 7.95669016e-01 -1.31979479e-01 9.90107977e-01 7.90670654e-01]
 [ 2.24968346e+00 1.70959465e+00 1.67215710e+00 1.31719939e+00]
 [ 2.24968346e+00 -1.05276654e+00 1.78583195e+00 1.44883158e+00]
 [ 1.89829664e-01 -1.97355361e+00 7.05920842e-01 3.95774101e-01]
 [ 1.28034050e+00 3.28414053e-01 1.10378283e+00 1.44883158e+00]
 [-2.94841818e-01 -5.92373012e-01 6.49083415e-01 1.05393502e+00]
 [ 2.24968346e+00 -5.92373012e-01 1.67215710e+00 1.05393502e+00]
 [ 5.53333275e-01 -8.22569778e-01 6.49083415e-01 7.90670654e-01]
 [ 1.03800476e+00 5.58610819e-01 1.10378283e+00 1.18556721e+00]
 [ 1.64384411e+00 3.28414053e-01 1.27429511e+00 7.90670654e-01]
 [ 4.32165405e-01 -5.92373012e-01 5.92245988e-01 7.90670654e-01]
 [ 3.10997534e-01 -1.31979479e-01 6.49083415e-01 7.90670654e-01]
 [ 6.74501145e-01 -5.92373012e-01 1.04694540e+00 1.18556721e+00]
 [ 1.64384411e+00 -1.31979479e-01 1.16062026e+00 5.27406285e-01]
 [ 1.88617985e+00 -5.92373012e-01 1.33113254e+00 9.22302838e-01]
 [ 2.49201920e+00 1.70959465e+00 1.50164482e+00 1.05393502e+00]
 [ 6.74501145e-01 -5.92373012e-01 1.04694540e+00 1.31719939e+00]
 [ 5.53333275e-01 -5.92373012e-01 7.62758269e-01 3.95774101e-01]
 [ 3.10997534e-01 -1.05276654e+00 1.04694540e+00 2.64141916e-01]
 [ 2.24968346e+00 -1.31979479e-01 1.33113254e+00 1.44883158e+00]
 [ 5.53333275e-01 7.88807586e-01 1.04694540e+00 1.58046376e+00]
 [ 6.74501145e-01 9.82172869e-02 9.90107977e-01 7.90670654e-01]
 [ 1.89829664e-01 -1.31979479e-01 5.92245988e-01 7.90670654e-01]
 [ 1.28034050e+00 9.82172869e-02 9.33270550e-01 1.18556721e+00]
 [ 1.03800476e+00 9.82172869e-02 1.04694540e+00 1.58046376e+00]
 [ 1.28034050e+00 9.82172869e-02 7.62758269e-01 1.44883158e+00]
 [-5.25060772e-02 -8.22569778e-01 7.62758269e-01 9.22302838e-01]
 [ 1.15917263e+00 3.28414053e-01 1.21745768e+00 1.44883158e+00]
 [ 1.03800476e+00 5.58610819e-01 1.10378283e+00 1.71209594e+00]
 [ 1.03800476e+00 -1.31979479e-01 8.19595696e-01 1.44883158e+00]
 [ 5.53333275e-01 -1.28296331e+00 7.05920842e-01 9.22302838e-01]
 [ 7.95669016e-01 -1.31979479e-01 8.19595696e-01 1.05393502e+00]
 [ 4.32165405e-01 7.88807586e-01 9.33270550e-01 1.44883158e+00]
 [ 6.86617933e-02 -1.31979479e-01 7.62758269e-01 7.90670654e-01]]
[[-9.00681170e-01 1.01900435e+00 -1.34022653e+00 -1.31544430e+00]
 [-1.14301691e+00 -1.31979479e-01 -1.34022653e+00 -1.31544430e+00]
 [-1.38535265e+00 3.28414053e-01 -1.39706395e+00 -1.31544430e+00]
 [-1.50652052e+00 9.82172869e-02 -1.28338910e+00 -1.31544430e+00]
 [-1.02184904e+00 1.24920112e+00 -1.34022653e+00 -1.31544430e+00]
 [-5.37177559e-01 1.93979142e+00 -1.16971425e+00 -1.05217993e+00]
 [-1.50652052e+00 7.88807586e-01 -1.34022653e+00 -1.18381211e+00]
 [-1.02184904e+00 7.88807586e-01 -1.28338910e+00 -1.31544430e+00]
 [-1.74885626e+00 -3.62176246e-01 -1.34022653e+00 -1.31544430e+00]
 [-1.14301691e+00 9.82172869e-02 -1.28338910e+00 -1.44707648e+00]
```

```
[-5.37177559e-01 1.47939788e+00 -1.28338910e+00 -1.31544430e+00]
[-1.26418478e+00 7.88807586e-01 -1.22655167e+00 -1.31544430e+00]
[-1.26418478e+00 -1.31979479e-01 -1.34022653e+00 -1.44707648e+00]
[-1.87002413e+00 -1.31979479e-01 -1.51073881e+00 -1.44707648e+00]
[-5.25060772e-02 2.16998818e+00 -1.45390138e+00 -1.31544430e+00]
[-1.73673948e-01 3.09077525e+00 -1.28338910e+00 -1.05217993e+00]
[-5.37177559e-01 1.93979142e+00 -1.39706395e+00 -1.05217993e+00]
[-9.00681170e-01 1.01900435e+00 -1.34022653e+00 -1.18381211e+00]
[-1.73673948e-01 1.70959465e+00 -1.16971425e+00 -1.18381211e+00]
[-9.00681170e-01 1.70959465e+00 -1.28338910e+00 -1.18381211e+00]
[-5.37177559e-01 7.88807586e-01 -1.16971425e+00 -1.31544430e+00]
[-9.00681170e-01 1.47939788e+00 -1.28338910e+00 -1.05217993e+00]
[-1.50652052e+00 1.24920112e+00 -1.56757623e+00 -1.31544430e+00]
[-9.00681170e-01 5.58610819e-01 -1.16971425e+00 -9.20547742e-01]
[-1.26418478e+00 7.88807586e-01 -1.05603939e+00 -1.31544430e+00]
[-1.02184904e+00 -1.31979479e-01 -1.22655167e+00 -1.31544430e+00]
[-1.02184904e+00 7.88807586e-01 -1.22655167e+00 -1.05217993e+00]
[-7.79513300e-01 1.01900435e+00 -1.28338910e+00 -1.31544430e+00]
[-7.79513300e-01 7.88807586e-01 -1.34022653e+00 -1.31544430e+00]
[-1.38535265e+00 3.28414053e-01 -1.22655167e+00 -1.31544430e+00]
[-1.26418478e+00 9.82172869e-02 -1.22655167e+00 -1.31544430e+00]
[-5.37177559e-01 7.88807586e-01 -1.28338910e+00 -1.05217993e+00]
[-7.79513300e-01 2.40018495e+00 -1.28338910e+00 -1.44707648e+00]
[-4.16009689e-01 2.63038172e+00 -1.34022653e+00 -1.31544430e+00]
[-1.14301691e+00 9.82172869e-02 -1.28338910e+00 -1.31544430e+00]
[-1.02184904e+00 3.28414053e-01 -1.45390138e+00 -1.31544430e+00]
[-4.16009689e-01 1.01900435e+00 -1.39706395e+00 -1.31544430e+00]
[-1.14301691e+00 1.24920112e+00 -1.34022653e+00 -1.44707648e+00]
[-1.74885626e+00 -1.31979479e-01 -1.39706395e+00 -1.31544430e+00]
[-9.00681170e-01 7.88807586e-01 -1.28338910e+00 -1.31544430e+00]
[-1.02184904e+00 1.01900435e+00 -1.39706395e+00 -1.18381211e+00]
[-1.62768839e+00 -1.74335684e+00 -1.39706395e+00 -1.18381211e+00]
[-1.74885626e+00 3.28414053e-01 -1.39706395e+00 -1.31544430e+00]
[-1.02184904e+00 1.01900435e+00 -1.22655167e+00 -7.88915558e-01]
[-9.00681170e-01 1.70959465e+00 -1.05603939e+00 -1.05217993e+00]
[-1.26418478e+00 -1.31979479e-01 -1.34022653e+00 -1.18381211e+00]
[-9.00681170e-01 1.70959465e+00 -1.22655167e+00 -1.31544430e+00]
[-1.50652052e+00 3.28414053e-01 -1.34022653e+00 -1.31544430e+00]
[-6.58345429e-01 1.47939788e+00 -1.28338910e+00 -1.31544430e+00]
[-1.02184904e+00 5.58610819e-01 -1.34022653e+00 -1.31544430e+00]
[ 1.40150837e+00 3.28414053e-01 5.35408562e-01 2.64141916e-01]
[ 6.74501145e-01 3.28414053e-01 4.21733708e-01 3.95774101e-01]
[ 1.28034050e+00 9.82172869e-02 6.49083415e-01 3.95774101e-01]
[-4.16009689e-01 -1.74335684e+00 1.37546573e-01 1.32509732e-01]
[ 7.95669016e-01 -5.92373012e-01 4.78571135e-01 3.95774101e-01]
[-1.73673948e-01 -5.92373012e-01 4.21733708e-01 1.32509732e-01]
[ 5.53333275e-01 5.58610819e-01 5.35408562e-01 5.27406285e-01]
[-1.14301691e+00 -1.51316008e+00 -2.60315415e-01 -2.62386821e-01]
[ 9.16836886e-01 -3.62176246e-01 4.78571135e-01 1.32509732e-01]
[-7.79513300e-01 -8.22569778e-01 8.07091462e-02 2.64141916e-01]
[-1.02184904e+00 -2.43394714e+00 -1.46640561e-01 -2.62386821e-01]
[ 6.86617933e-02 -1.31979479e-01 2.51221427e-01 3.95774101e-01]
[ 1.89829664e-01 -1.97355361e+00 1.37546573e-01 -2.62386821e-01]
[ 3.10997534e-01 -3.62176246e-01 5.35408562e-01 2.64141916e-01]
```

```
[-2.94841818e-01 -3.62176246e-01 -8.98031345e-02 1.32509732e-01]
[ 1.03800476e+00 9.82172869e-02 3.64896281e-01 2.64141916e-01]
[-2.94841818e-01 -1.31979479e-01 4.21733708e-01 3.95774101e-01]
[-5.25060772e-02 -8.22569778e-01 1.94384000e-01 -2.62386821e-01]
[ 4.32165405e-01 -1.97355361e+00 4.21733708e-01 3.95774101e-01]
[-2.94841818e-01 -1.28296331e+00 8.07091462e-02 -1.30754636e-01]
[ 6.86617933e-02 3.28414053e-01 5.92245988e-01 7.90670654e-01]
[ 3.10997534e-01 -5.92373012e-01 1.37546573e-01 1.32509732e-01]
[ 5.53333275e-01 -1.28296331e+00 6.49083415e-01 3.95774101e-01]
[ 3.10997534e-01 -5.92373012e-01 5.35408562e-01 8.77547895e-04]
[ 6.74501145e-01 -3.62176246e-01 3.08058854e-01 1.32509732e-01]
[ 9.16836886e-01 -1.31979479e-01 3.64896281e-01 2.64141916e-01]
[ 1.15917263e+00 -5.92373012e-01 5.92245988e-01 2.64141916e-01]
[ 1.03800476e+00 -1.31979479e-01 7.05920842e-01 6.59038469e-01]
[ 1.89829664e-01 -3.62176246e-01 4.21733708e-01 3.95774101e-01]
[-1.73673948e-01 -1.05276654e+00 -1.46640561e-01 -2.62386821e-01]
[-4.16009689e-01 -1.51316008e+00 2.38717193e-02 -1.30754636e-01]
[-4.16009689e-01 -1.51316008e+00 -3.29657076e-02 -2.62386821e-01]
[-5.25060772e-02 -8.22569778e-01 8.07091462e-02 8.77547895e-04]
[ 1.89829664e-01 -8.22569778e-01 7.62758269e-01 5.27406285e-01]
[-5.37177559e-01 -1.31979479e-01 4.21733708e-01 3.95774101e-01]
[ 1.89829664e-01 7.88807586e-01 4.21733708e-01 5.27406285e-01]
[ 1.03800476e+00 9.82172869e-02 5.35408562e-01 3.95774101e-01]
[ 5.53333275e-01 -1.74335684e+00 3.64896281e-01 1.32509732e-01]
[-2.94841818e-01 -1.31979479e-01 1.94384000e-01 1.32509732e-01]
[-4.16009689e-01 -1.28296331e+00 1.37546573e-01 1.32509732e-01]
[-4.16009689e-01 -1.05276654e+00 3.64896281e-01 8.77547895e-04]
[ 3.10997534e-01 -1.31979479e-01 4.78571135e-01 2.64141916e-01]
[-5.25060772e-02 -1.05276654e+00 1.37546573e-01 8.77547895e-04]
[-1.02184904e+00 -1.74335684e+00 -2.60315415e-01 -2.62386821e-01]
[-2.94841818e-01 -8.22569778e-01 2.51221427e-01 1.32509732e-01]
[-1.73673948e-01 -1.31979479e-01 2.51221427e-01 8.77547895e-04]
[-1.73673948e-01 -3.62176246e-01 2.51221427e-01 1.32509732e-01]
[ 4.32165405e-01 -3.62176246e-01 3.08058854e-01 1.32509732e-01]
[-9.00681170e-01 -1.28296331e+00 -4.30827696e-01 -1.30754636e-01]
[-1.73673948e-01 -5.92373012e-01 1.94384000e-01 1.32509732e-01]
[ 5.53333275e-01 5.58610819e-01 1.27429511e+00 1.71209594e+00]
[-5.25060772e-02 -8.22569778e-01 7.62758269e-01 9.22302838e-01]
[ 1.52267624e+00 -1.31979479e-01 1.21745768e+00 1.18556721e+00]
[ 5.53333275e-01 -3.62176246e-01 1.04694540e+00 7.90670654e-01]
[ 7.95669016e-01 -1.31979479e-01 1.16062026e+00 1.31719939e+00]
[ 2.12851559e+00 -1.31979479e-01 1.61531967e+00 1.18556721e+00]
[-1.14301691e+00 -1.28296331e+00 4.21733708e-01 6.59038469e-01]
[ 1.76501198e+00 -3.62176246e-01 1.44480739e+00 7.90670654e-01]
[ 1.03800476e+00 -1.28296331e+00 1.16062026e+00 7.90670654e-01]
[ 1.64384411e+00 1.24920112e+00 1.33113254e+00 1.71209594e+00]
[ 7.95669016e-01 3.28414053e-01 7.62758269e-01 1.05393502e+00]
[ 6.74501145e-01 -8.22569778e-01 8.76433123e-01 9.22302838e-01]
[ 1.15917263e+00 -1.31979479e-01 9.90107977e-01 1.18556721e+00]
[-1.73673948e-01 -1.28296331e+00 7.05920842e-01 1.05393502e+00]
[-5.25060772e-02 -5.92373012e-01 7.62758269e-01 1.58046376e+00]
[ 6.74501145e-01 3.28414053e-01 8.76433123e-01 1.44883158e+00]
[ 7.95669016e-01 -1.31979479e-01 9.90107977e-01 7.90670654e-01]
[ 2.24968346e+00 1.70959465e+00 1.67215710e+00 1.31719939e+00]
```

```
[ 2.24968346e+00 -1.05276654e+00 1.78583195e+00 1.44883158e+00]
[ 1.89829664e-01 -1.97355361e+00 7.05920842e-01 3.95774101e-01]
[ 1.28034050e+00 3.28414053e-01 1.10378283e+00 1.44883158e+00]
[-2.94841818e-01 -5.92373012e-01 6.49083415e-01 1.05393502e+00]
[ 2.24968346e+00 -5.92373012e-01 1.67215710e+00 1.05393502e+00]
[ 5.53333275e-01 -8.22569778e-01 6.49083415e-01 7.90670654e-01]
[ 1.03800476e+00 5.58610819e-01 1.10378283e+00 1.18556721e+00]
[ 1.64384411e+00 3.28414053e-01 1.27429511e+00 7.90670654e-01]
[ 4.32165405e-01 -5.92373012e-01 5.92245988e-01 7.90670654e-01]
[ 3.10997534e-01 -1.31979479e-01 6.49083415e-01 7.90670654e-01]
[ 6.74501145e-01 -5.92373012e-01 1.04694540e+00 1.18556721e+00]
[ 1.64384411e+00 -1.31979479e-01 1.16062026e+00 5.27406285e-01]
[ 1.88617985e+00 -5.92373012e-01 1.33113254e+00 9.22302838e-01]
[ 2.49201920e+00 1.70959465e+00 1.50164482e+00 1.05393502e+00]
[ 6.74501145e-01 -5.92373012e-01 1.04694540e+00 1.31719939e+00]
[ 5.53333275e-01 -5.92373012e-01 7.62758269e-01 3.95774101e-01]
[ 3.10997534e-01 -1.05276654e+00 1.04694540e+00 2.64141916e-01]
[ 2.24968346e+00 -1.31979479e-01 1.33113254e+00 1.44883158e+00]
[ 5.53333275e-01 7.88807586e-01 1.04694540e+00 1.58046376e+00]
[ 6.74501145e-01 9.82172869e-02 9.90107977e-01 7.90670654e-01]
[ 1.89829664e-01 -1.31979479e-01 5.92245988e-01 7.90670654e-01]
[ 1.28034050e+00 9.82172869e-02 9.33270550e-01 1.18556721e+00]
[ 1.03800476e+00 9.82172869e-02 1.04694540e+00 1.58046376e+00]
[ 1.28034050e+00 9.82172869e-02 7.62758269e-01 1.44883158e+00]
[-5.25060772e-02 -8.22569778e-01 7.62758269e-01 9.22302838e-01]
[ 1.15917263e+00 3.28414053e-01 1.21745768e+00 1.44883158e+00]
[ 1.03800476e+00 5.58610819e-01 1.10378283e+00 1.71209594e+00]
[ 1.03800476e+00 -1.31979479e-01 8.19595696e-01 1.44883158e+00]
[ 5.53333275e-01 -1.28296331e+00 7.05920842e-01 9.22302838e-01]
[ 7.95669016e-01 -1.31979479e-01 8.19595696e-01 1.05393502e+00]
[ 4.32165405e-01 7.88807586e-01 9.33270550e-01 1.44883158e+00]
[ 6.86617933e-02 -1.31979479e-01 7.62758269e-01 7.90670654e-01]]
```