AMERICAN UNIVERSITY OF BEIRUT

BRIDGING THE DATA SCIENCE SKILLS GAP IN KENYA THROUGH A PROJECT-BASED ONLINE BOOTCAMP

by

PAUL LEKUMOK LEMARON

A project
submitted in partial fulfillment of the requirements
for the degree of Master of Arts
to the Department of Education
of the Faculty of Arts and Sciences
at the American University of Beirut

Beirut, Lebanon
December 2025

AMERICAN UNIVERSITY OF BEIRUT

BRIDGING THE DATA SCIENCE SKILLS GAP IN KENYA THROUGH A PROJECT-BASED ONLINE BOOTCAMP

by

PAUL LEKUMOK LEMARON

Approved by:

Dr. Haidar Safa, Professor Department of Computer Science	First Reader
Dr. Hoda Baytiyeh, Associate Professor Department of Education	Second Reader

Date of project defense: December 10, 2025

ACKNOWLEDGEMENTS

First and foremost, I dedicate this work and give glory to God for providing the strength, guidance, and opportunity to complete this project.

I wish to express my profound gratitude to my incredible wife, Lilian Naipanoi, for being an amazing woman, supporting my Master of Arts studies, and for being the best mother to our daughter, Nenkishon, who was born while I was taking my masters. Their love and resilience were my greatest motivation.

My sincere thanks go to the Mastercard Foundation Scholars Program for offering me a full scholarship, which made the pursuit of this degree possible at the American University of Beirut.

I extend my appreciation to the Advisors team-including Professor Haidar Safa, Professor Shady Elbassuoni and Professor Hoda Baytiyeh - for agreeing to be my advisors, providing invaluable guidance, and sharing their expertise throughout the research and writing process.

Thank you to my beloved parents, Totuaek Lemaron and Lemaron Ole Yiaile, for instilling in me the value of education and supporting my academic journey.
I am grateful to my brother, Henry Lemaron, for taking the time to review my Statement of Purpose when I was applying to AUB and for instilling into me the skills of a reader and scholar as well being a mentor in academia and life in general.

Finally, thank you to Dr. Patrick Bii, Jeremy Owallo, and Sam Wambugu for submitting essential letters of recommendation for my Master's application to AUB.

ABSTRACT
OF THE PROJECT OF

Paul Lekumok Lemaron for Master of Arts
Major: Education

Title: Bridging The Data Science Skills Gap in Kenya Through a Project-Based Online Bootcamp

Kenya's rapidly expanding digital economy is hampered by a persistent skills gap, particularly in data science, limiting innovation in critical sectors like agriculture and healthcare (Coker et al., 2023; Said, 2021). This gap stems from a disconnect between the theoretical focus of traditional university curricula and the practical, applied competencies demanded by industry (Hicks & Irizarry, 2018; Said, 2021).

This paper details a project-based Online Data Science Bootcamp designed to address this challenge. The project's pedagogical framework is centered on Project-Based Learning (PBL), utilizing the case-study method to teach the key data science skills of creating, connecting, and computing (Hicks & Irizarry, 2018). The bootcamp utilizes a Masterstudy Learning Management System (LMS) hosted at the Doable Learning Platform (https://doable.co.ke/), integrated with Google Colab and GitHub, to provide a seamless, interactive learning environment that removes technical barriers for beginners (Kim & Henke, 2021).

The curriculum guides learners through Python, NumPy, Pandas, data visualization, and machine learning, culminating in real-world projects that position students at the boundary between academic learning and practical application (Saltz & Heckman, 2015). By providing an accessible, hands-on, and industry-aligned training model, this project offers a scalable solution to empower a new generation of data professionals in Kenya, directly supporting the goals of Vision 2030.

TABLE OF CONTENTS

TABLES

Table

ILLUSTRATIONS

Figure

ABBREVIATIONS

AUB	American University of Beirut
AI	Artificial Intelligence
ICT	Information and Communication Technology
LMS	Learning Management System
LMIC	Low- and Middle-Income Country
MA	Master of Arts
ML	Machine Learning
PBL	Project-Based Learning
TNA	Training Needs Analysis
Q&A	Questions and Answers
R	R programming language
SCORM	Sharable Content Object Reference Model
SSA	Sub-Saharan Africa

CHAPTER 1

INTRODUCTION

Data science has emerged as a critical interdisciplinary field that uses scientific methods and algorithms to extract insights from complex data, transforming traditional business into digitally enabled enterprises (Andriole, 2018; Birkenkrahe, 2022). This has created a global surge in demand for data professionals, including data scientists, engineers, and analysts, as industries seek to leverage big data analytics and machine learning to gain a competitive advantage (Coetzee & Goede, 2025). However, this demand has outpaced the supply of qualified talent, resulting in a significant skills gap worldwide, with the U.S. Bureau of Labor Statistics projecting a 28% annual growth in jobs requiring data science expertise by 2026 (Janssen, 2022, Coetzee & Goede, 2025; Avila-Garzon & Bacca-Acosta, 2025).

This challenge is particularly acute in Sub-Saharan Africa (SSA), a region that has largely been left behind in the data revolution due to inadequate infrastructure, limited training opportunities, and a lack of local expertise (Coker et al., 2023). In Kenya, this skills gap represents a primary barrier to national development and the realization of its ambitions under the Vision 2030 framework (Said, 2021). While the Kenyan government's Digital Economy Blueprint prioritizes a digitally skilled workforce, the country faces a critical shortage of advanced data science expertise, limiting data-driven innovation in sectors vital to its socio-economic growth, including agriculture and healthcare (Coker et al., 2023; Republic of Kenya, 2019, Said, 2021).

Statement of the Problem

The core of the data science skills shortage in Kenya lies in a fundamental disconnect between the country's higher education system and the dynamic needs of its

industries (Said, 2021). University curricula are frequently misaligned with technical requirements, focusing on theoretical concepts and outdated technologies while neglecting the practical, hands-on experience essential for modern data science work (Coker et al., 2023b; Said, 2021). Many statistics and computing programs are built on an "arid, context-free landscape" that uses simplified datasets as mere illustrations, failing to engage students in genuine statistical thinking or prepare them for the messy realities of real-world data (Hicks & Irizarry, 2018, p. 383).

Consequently, graduates are often unprepared for crucial tasks like data pre-processing, model deployment, report-writing, and communicating insights to non-technical stakeholders (Coetzee & Goede, 2025). Data practitioners in South Africa, for instance, perceive younger colleagues to be most lacking in management skills such as presentation, networking, and business acumen, as well as self-regulating skills like stress-handling and time management (Coetzee & Goede, 2025). This theory-practice gap has tangible consequences. In agriculture, a cornerstone of Kenya's economy, smallholder farmers lack access to the data analytics tools needed to optimize crop yields and manage resources, contributing to food insecurity (Boylan et al., 2025). In healthcare, the effective management of diseases and public health surveillance is hampered by the shortage of professionals who can harness large datasets for evidence-based interventions (Coker et al., 2023; Kiosia et al., 2024).

The result is a workforce with "inappropriate competencies," forcing employers to invest in expensive retraining or rely on foreign expertise-an unsustainable model for national development (Said, 2021, p. 46). Existing capacity-building initiatives are often fragmented and lack a clear, scalable strategy, while the slow institutional response of universities has proven inadequate to meet the urgent demand (Kiosia et al., 2024; Said,

2021). Therefore, an accessible, scalable, and industry-aligned educational solution is needed to bridge this critical skills gap.

Rationale of the Study

The urgency of this project is driven by the need for a pedagogical model that directly confronts the theory-practice gap in Kenyan data science education. Traditional pedagogical approaches in statistics and computing have failed to produce job-ready graduates because they prioritize describing techniques over solving real-world problems (Hicks & Irizarry, 2018). This study's rationale is grounded in the pedagogical principles of Project-Based Learning (PBL)**,** rooted in constructivist theory (Zhang & Ma, 2023). By centering the curriculum on authentic problems, this design inverts the traditional model, treating statistical techniques and computational tools as instruments to answer a central scientific question (Hicks & Irizarry, 2018).

This project contributes a tangible and replicable educational framework-an Online Data Science Bootcamp-that is both pedagogically sound and technologically accessible. Its significance lies in its practical implementation of a project-focused methodology within the specific context of Kenya. By using real-world datasets from platforms like Kaggle and industry-standard tools like GitHub and Google Colab, the bootcamp places learners at the boundary between academic theory and professional practice, a position conducive to the construction of new, applicable knowledge (Saltz & Heckman, 2015).

Technologically, the project provides a scalable and cost-effective solution. The use of a Masterstudy LMS on WordPress, combined with cloud-based tools like Google Colab, makes high-quality technical training accessible to a geographically dispersed audience, removing the significant barrier of software installation and environment

configuration that often frustrates beginners (Kim & Henke, 2021; Holdgraf et al., 2017). This provides a validated framework that can inform educational reform and serve as a model for other developing nations facing similar challenges.

Project Purpose and Objectives

The overarching purpose of this project is to design, and implement a scalable, project-based Online Data Science Bootcamp to address the critical data science skills gap among university students and IT professionals in Kenya. By centering the curriculum on hands-on, case-study-based learning, this initiative aims to bridge the documented disconnect between theoretical academic instruction and the practical, industry-demanded competencies necessary to drive innovation in Kenya's digital economy (Hicks & Irizarry, 2018; Said, 2021).

To achieve this purpose, the project is guided by the following specific objectives:

- To design a data science curriculum that progresses from foundational concepts to advanced applications, covering Python, NumPy, Pandas, data visualization, and machine learning. This curriculum will be structured to teach the key skills of creating, connecting, and computing through practical, real-world problem-solving (Hicks & Irizarry, 2018).
- To implement an accessible and interactive online learning platform using a Masterstudy LMS integrated with cloud-based tools, including Google Colab for coding exercises and GitHub for project management. This technology stack is designed to be scalable and to minimize technical setup barriers for learners across Kenya (Kim & Henke, 2021).
- To facilitate hands-on, project-based learning by requiring students to complete a portfolio of assignments using real-world datasets. This approach provides an

authentic learning experience that mimics the iterative and interactive process of a professional data scientist, fostering practical skills and critical thinking (Hicks & Irizarry, 2018; Saltz & Heckman, 2015).

Significance of The Project

The significance of this project extends across practical, policy, and theoretical domains, offering a robust model for data science capacity building in Kenya and similar contexts.

Practical Implications

The most immediate practical implication is the creation of a pipeline of job-ready data science professionals equipped with applied skills. Graduates will possess a portfolio of real-world projects, demonstrating their proficiency in in-demand areas like data preprocessing, visualization, and machine learning-skills currently lacking in the region's workforce (Coetzee & Goede, 2025). This will directly benefit key Kenyan sectors. In agriculture, skilled professionals can apply analytics to improve crop yields; in healthcare, they can analyze public health data to track diseases and inform policy (Boylan et al., 2025). The project’s low-cost, scalable technology stack using Masterstudy LMS and Google Colab also provides a replicable model for other educational providers in resource-constrained environments (Kim & Henke, 2021).

The foundational needs assessment, including the detailed Training Needs Analysis (TNA), is provided in Appendix A. This preliminary analysis was crucial for ensuring that the project's design directly addresses the practical skills deficits identified in the Sub-Saharan African context (Coker et al., 2023).

Policy Implications

This project provides an evidence-based framework that can inform policy at both national and institutional levels. It offers a tested, operational model for achieving the goals of Kenya's Vision 2030 and its Digital Economy Blueprint, justifying investment in agile, non-traditional educational programs (Said, 2021). For universities, the bootcamp's success can serve as a catalyst for curriculum reform, encouraging a shift towards project-based pedagogy and stronger industry collaboration to close the persistent skills gap (Said, 2021). It also supports the case for regulatory frameworks that recognize alternative credentials, creating more flexible pathways to employment.

Theoretical and Scientific Contributions

This study provides a rigorous theoretical and design contribution by documenting the systematic creation of an educational artifact. It operationalizes the principles of Project-Based Learning (PBL) and the case-study approach (Hicks & Irizarry, 2018), validating their strategic design suitability for an online LMIC setting (Said, 2021). Furthermore, it serves as a practical implementation of Project-based learning theory, showing how positioning students between an academic course and real-world problems facilitates the construction of applicable knowledge (Saltz & Heckman, 2015). The project also contributes to the discourse on educational technology by evaluating a cost-effective and accessible technology stack, offering an alternative to more resource-intensive platforms and adding to the evidence base on how to effectively design and deploy online learning environments for hands-on computational instruction.

CHAPTER 2

LITERATURE REVIEW

Introduction

The emergence of data science as a transformative field has created a global demand for skilled professionals capable of extracting insights from complex datasets (Birkenkrahe, 2022; Coker et al., 2023). This demand has exposed a significant talent gap between the competencies required by industry and the skills provided by traditional educational systems, a challenge that is particularly acute in developing regions (Avila-Garzon & Bacca-Acosta, 2025; Coetzee & Goede, 2025). This chapter provides a comprehensive review of the literature, identified through a systematic search process detailed in Chapter 3, to contextualize the data science skills gap and situate this project within existing scholarly discourse. The review is structured to move from a broad global perspective to a specific local analysis, thereby establishing the theoretical and practical foundations for the proposed project-based intervention.

The review will first examine the global context, synthesizing international trends in data science education, with a focus on the pedagogical shift towards practical, project-based models. Second, it will narrow to the regional context of Africa, analyzing the systemic factors that define the data science landscape on the continent. Third, the chapter will provide an in-depth analysis of the Kenyan context, detailing the documented disconnect between its higher education system and industry needs. Finally, the chapter will conclude with a synthesis of the literature, identifying the precise research gap this project is designed to fill.

Global Context: The Pedagogical and Technological Shift in Data Science Education

The global proliferation of data has established data science as a critical field, yet this has created a skills gap that traditional educational institutions are struggling to address (Andriole, 2018; Avila-Garzon & Bacca-Acosta, 2025). The literature reveals a consensus that conventional academic curricula are often misaligned with the practical nature of data science work, prompting a re-evaluation of pedagogical strategies.

The Pedagogical Disconnect in Traditional Curricula

A central theme is the critique of conventional statistics and computer science programs for their overemphasis on theory. Many university curricula focus on describing statistical techniques using clean, simplified datasets as mere illustrations (Hicks & Irizarry, 2018). This approach, described as an "arid, context-free landscape," fails to prepare students for the messy realities of real-world data analysis (Hicks & Irizarry, 2018, p. 383). Consequently, graduates frequently lack proficiency in crucial tasks like data acquisition, wrangling, and communication-skills often learned on the job rather than in the classroom (Hicks & Irizarry, 2018; Coetzee & Goede, 2025). This creates a significant theory-practice gap, a phenomenon documented across different regions (Coetzee & Goede, 2025; Said, 2021).

The Rise of Project-Based and Case-Study-Based Pedagogy

In response, a new pedagogical consensus has emerged, centered on project-based and case-study-based learning (Hicks & Irizarry, 2018; Avila-Garzon & Bacca-Acosta, 2025). This approach inverts the conventional model by placing a real-world problem at the forefront, treating statistical techniques and computational tools as instruments to answer a question (Hicks & Irizarry, 2018). This pedagogical shift is

designed to teach three key skills essential to data science: creating (formulating questions), connecting (linking questions to appropriate data and analysis tools), and computing (applying computational skills to solve the problem) (Hicks & Irizarry, 2018). This model fosters critical thinking and moves beyond passive learning, enabling students to develop a more holistic understanding of the entire data science lifecycle (Saltz & Heckman, 2015).

The Technological Enablement of Modern Data Science Education

This pedagogical shift has been accelerated by advancements in educational technology. A primary hurdle in teaching data science is the technical barrier of software installation and configuration, which can detract from instructional time (Holdgraf et al., 2017; Kim & Henke, 2021). Cloud computing platforms have emerged as a powerful solution. Tools like Jupyter Notebooks (often run via services like Google Colab) allow instructors to create standardized, interactive computational environments accessible through a web browser (Holdgraf et al., 2017; Liu et al., 2020) coupled with AI Integration. This approach offloads the setup burden, ensuring all students work in an identical environment and can begin coding immediately (Kim & Henke, 2021). Learning Management Systems (LMS) like Moodle or WordPress-based systems have also become essential for delivering and managing online courses, offering features for content delivery, assessments, and communication (Birkenkrahe, 2022; Duta et al., 2021). The data generated within an LMS also provides a foundation for learning analytics to monitor student engagement and provide timely interventions (Conijn et al., 2017).

Regional Context: The Data Science Skills Gap in Africa

The data science skills gap is particularly pronounced in Sub-Saharan Africa (SSA), a region that has been largely "left behind in the data revolution" (Coker et al., 2023, p. 2). Despite the continent's potential, innovation has lagged due to systemic challenges, including inadequate infrastructure and a shortage of specialized training (Coker et al., 2023). Academic institutions across Africa have been playing "catch-up," with many countries having developed little to no capacity in the field (Coker et al., 2023). This shortage is a major obstacle for organizations and forces countries to rely on foreign expertise, which undermines sustainable, local capacity building (Said, 2021; Coetzee & Goede, 2025).

This skills deficit has profound implications for critical sectors like biomedical research and public health, where data science offers powerful tools for disease surveillance and predictive analysis that the region cannot fully harness (Coker et al., 2023; Kiosia et al., 2024). Existing capacity-building initiatives are often fragmented and lack a clear, scalable implementation strategy (Kiosia et al., 2024). Even where training exists, it often fails to equip graduates with industry-required competencies, such as data pre-processing and data visualization, reflecting a broader educational failure to use real-life, "messy" data in curricula (Coetzee & Goede, 2025).

The Data Science Skills Gap in Kenya

Kenya has a significant and urgent need for data science professionals to support its growing digital economy, but this demand is not being met due to a critical skills gap between what the industry requires and what higher education institutions provide (Said, 2021; Coker et al., 2023). This gap affects key sectors like healthcare and

agriculture (Boylan et al., 2025; Coker et al., 2023) and hinders the country's ability to achieve the goals set out in its Vision 2030 framework (Said, 2021).

A fundamental disconnect exists between university curricula and industry needs (Said, 2021; Hicks & Irizarry, 2018). Higher education in Kenya has been criticized for producing graduates who are ill-equipped for the modern workplace, forcing employers to either invest in costly retraining or rely on foreign expertise (Said, 2021).

Key aspects of this skills gap include:

- Low Supply of Talent: There is a significant disparity between the demand for ICT skills and the supply of qualified graduates (Said, 2021). Student enrollment and graduation rates in ICT programs are low (Said, 2021). One projection estimated that by 2030, 50-55% of all jobs in Kenya will require digital skills, yet the number of qualified and well-trained ICT graduates is insufficient to meet this demand (Said, 2021). As of 2019, only one university in Kenya, Jomo Kenyatta University of Agriculture and Technology, offered a dedicated data science course (Coker et al., 2023).

- Misaligned Curricula: University curricula are often misaligned with the country's technical needs, focusing on theory and outdated technologies rather than practical, hands-on experience (Said, 2021; Hicks & Irizarry, 2018). Computer science programs at major Kenyan universities have been found to lack modern topics such as Big Data, Cybersecurity, and Robotics (Said, 2021). This focus on an "arid, context-free landscape" fails to prepare students for the complexities of real-world data (Hicks & Irizarry, 2018).

- Lack of Practical Experience: A major deficiency among graduates is the lack of practical experience (Said, 2021). Students are not adequately trained in crucial

tasks like data pre-processing, cleaning "messy" real-life data, model deployment, and data visualization (Coetzee & Goede, 2025).

- Deficiency in Specific Skills:
    - Advanced Technical Skills: Kenya's Digital Economy Blueprint prioritizes advanced skills in Artificial Intelligence (AI), Machine Learning (ML), Big Data, and programming languages like R and Python (Said, 2021). However, these are precisely the areas where training is lacking (Said, 2021).
    - Soft Skills: Employers have noted that graduates often lack essential soft skills, including communication, presentation, report-writing, business acumen, teamwork, and the ability to communicate findings to non-technical stakeholders (Coetzee & Goede, 2025).

Root Causes of the Skills Gap

Several systemic issues within Kenya's higher education system that contribute to the data science skills shortage as follows:

- Outdated and Inflexible Curricula: Curricula are not reviewed regularly enough to keep pace with rapid technological advancements, leading to the teaching of obsolete courses (Said, 2021). There is a lack of standardization, with the quality of programs depending solely on an institution's internal assurance methods (Said, 2021).
- Resource and Infrastructure Challenges: Kenyan universities often suffer from inadequate funding, ill-equipped laboratories, insufficient ICT infrastructure, and outdated technology, all of which hinder the development of practical skills

(Said, 2021). For instance, the computer-to-student ratio in some public universities is approximately one per 150 learners (Said, 2021).

- Shortage of Qualified Staff: There is an inadequate number of qualified academic staff to teach ICT programs, resulting in a high staff-to-student ratio of 1:26, far above the recommended 1:10 for applied sciences (Said, 2021). Furthermore, many ICT faculty members in Kenya lack direct industry experience (Said, 2021).
- Limited University-Industry Collaboration: There is minimal meaningful interaction between higher learning institutions and the industry to help guide curriculum development and ensure its relevance to market needs (Said, 2021; Coetzee & Goede, 2025).

Addressing Kenya's Data Science Needs

To bridge the skills gap, a multi-faceted approach focusing on pedagogical reform, curriculum updates, and strategic capacity building.

- Shift to Project-Based and Case-Study Learning: A fundamental need is to move away from theory-first instruction toward project-based and case-study-based pedagogy (Hicks & Irizarry, 2018; Saltz & Heckman, 2015). This approach places a real-world problem at the center of the curriculum, teaching students the essential data science skills of creating, connecting, and computing by using statistical techniques and tools to answer authentic questions (Hicks & Irizarry, 2018). This requires using real-world, "messy" datasets to provide an authentic learning experience (Hicks & Irizarry, 2018; Coetzee & Goede, 2025).

- Curriculum Modernization: Curricula must be urgently updated to incorporate modern technologies and skills (Said, 2021). This includes:
    - Programming Languages: Teaching industry-standard languages like Python and R is essential for data science beginners (Birkenkrahe, 2022; Coker et al., 2023; Vittorini et al., 2021).
    - Core Data Science Topics: The curriculum should cover the entire data science lifecycle, including data collection, cleaning, visualization, data management, and machine learning (Hicks & Irizarry, 2018; Avila-Garzon & Bacca-Acosta, 2025).
    - Ethical Considerations: Data ethics should be an integral part of any data science curriculum (Blackmon & Moore, 2020; Boylan et al., 2025).
- Accessible and Scalable Training:
    - Online Bootcamps and Courses: Online models like bootcamps offer a scalable, low-cost, and flexible solution to deliver high-quality technical training to a broad audience, bypassing the slow pace of traditional institutional reform (Lemaron, 2025).
    - Leveraging Cloud-Based Technology: The use of cloud-based tools like Google Colab is crucial for removing the significant barrier of software installation, allowing beginners to start coding immediately in a standardized environment (Kim & Henke, 2021; Holdgraf et al., 2017).
- Strengthening Capacity and Collaboration:

- Faculty Development: Instructors need ongoing training and support to develop the skills required for effective online and data-intensive teaching (McGee et al., 2017).
- Industry-Academia Partnerships: Establishing stronger, structured collaborations between universities and industry is critical for ensuring curricula remain aligned with workforce needs (Said, 2021; Coetzee & Goede, 2025).
- Data Science Hubs: The creation of national or regional data science hubs can foster knowledge sharing, mentorship, and collaboration among researchers, particularly for early-career professionals (Kiosia et al., 2024).

By adopting these strategies, Kenya can begin to cultivate the skilled data science workforce necessary to drive innovation and achieve its vision for a thriving digital economy.

Kenyan Context: Misalignment Between Higher Education and Industry Needs

Kenya, with its rapidly expanding digital economy, provides a focused case study of the challenges prevalent across SSA (Said, 2021). The country faces a critical shortage of advanced digital skills, particularly in data science, which serves as a primary barrier to national development (Coker et al., 2023; Said, 2021). The literature specific to Kenya points to a systemic failure within its higher education institutions to produce graduates with industry-relevant skills.

The core of the problem is a fundamental mismatch between university curricula and the technical requirements of the ICT industry (Said, 2021). Curricula are often

outdated, overly theoretical, and not reviewed regularly enough to keep pace with technological advancements (Said, 2021). This is exacerbated by structural issues, as most ICT faculty in Kenya lack industry experience and are not active researchers (Said, 2021). The institutional response has been slow; as of 2023, only one public university in Kenya offered a dedicated data science course (Coker et al., 2023b). The result is a workforce with "inappropriate competencies," forcing employers to invest in expensive retraining (Said, 2021, p. 46).

Table 1

Review of Institutions Offering Data Science Courses in Africa (Adapted from Coker et al., 2023)

Institutions Offering Data Science Courses in Africa		
School	**Programme**	**Country**
University of the Witwatersrand, Johannesburg	MSc Data Science	South Africa
University of the Witwatersrand, Johannesburg	MSc Artificial Intelligence	South Africa
University of the Witwatersrand, Johannesburg	MSc e-science	South Africa
University of Pretoria	Data Science	South Africa
University of Cape Town	Masters Programmes in Data Science	South Africa
Rhodes University	MSc Bioinformatics	South Africa
Namibia University of Science and Technology	Master of Science in Applied Statistics	Namibia
National University of Rwanda (NUR)	The master's degree with specialization in Data Mining	Rwanda
University of Dar es Salaam	MSc. Data Science	Tanzania
Makerere University	AI & Data Science	Uganda
Eden University	Msc Data Analytics	Zambia

University of Zimbabwe	BSc Data Science and Informatics	Zimbabwe
Jomo Kenyatta University of Agriculture and Technology	MSc. Mathematics (Data Science)	Kenya
Addis Ababa University	MSc. Information Science and Systems	Ethiopia
Jamhuriya University of Science and Technology	MSc Data Science	Somalia
North West University	Data Mining/ Data Science	South Africa
American University of Nigeria	MSc Data Science and Analytics	Nigeria

Adapted from "Data science training needs in sub-Saharan Africa: Implications for biomedical research and therapeutics capacity," by M. O. Coker, O. A. Kotila, T. O. Adigun, T. V. Lawal, A. F. Fagbamigbe, A. U. Makolo, M. M. Salawu, D. T. Ajayi, P. E. Oluniyi, J. U. Oguzie, A. N. Happi, A. G. Hoen, C. P. Babalola, & I. O. Ajayi, 2023, Open Research Africa, 6(21).

Synthesis of Literature and Identification of the Research Gap

The literature presents a clear narrative progressing from global trends to specific local needs. Globally, there is a strong consensus that effective data science education must be practical, project-based, and technologically enabled. Regionally, SSA faces an acute skills shortage that impedes development, driven by a disconnect between academic offerings and industry demands. Nationally, the case of Kenya crystallizes these challenges, with a severe misalignment between its higher education system and its digital economy needs.

While the literature clearly establishes what constitutes effective data science pedagogy (project-based learning) and why it is urgently needed in Kenya (a severe skills gap), a significant gap remains in the research and implementation of targeted solutions. There is a dearth of documented interventions that combine these effective pedagogical principles into a scalable, accessible, and low-cost online format specifically designed for the Kenyan context. Existing capacity-building initiatives in

the region are often fragmented (Kiosia et al., 2024a), and traditional university programs in Kenya have proven inadequate (Said, 2021).

Therefore, this project directly addresses this gap by proposing and evaluating an Online Data Science Bootcamp that is:

- Pedagogically sound, based on the case-study method of bringing the problem to the forefront (Hicks & Irizarry, 2018).
- Technologically accessible, leveraging a cost-effective WordPress LMS and cloud-based tools to reach learners across Kenya (Kim & Henke, 2021).
- Contextually relevant, with a curriculum designed to equip learners with practical skills to solve real-world problems in key Kenyan sectors.

Conceptual Background/Theoretical Framework

The design and pedagogical approach of this Online Data Science Bootcamp are grounded in Project-Based Learning (PBL) methodology, rooted in constructivist theories of education. PBL fundamentally shifts the focus of education to a student-centered approach where students actively construct knowledge through experiences. This framework systematically addresses the documented theory-practice gap in Kenyan data science education (Said, 2021). Project-Based Learning (PBL) is characterized by sustained inquiry structured around complex, authentic questions and real-world problems. This aligns with the philosophical foundations of experiential learning, notably John Dewey's concept of "learning by doing". PBL mandates students take responsibility for the learning outcomes by emphasizing the practical dimension and relevance to their lives. Furthermore, acknowledging that learning is a social process (Vygotsky, 1978), the method strongly recommends collaborative learning in

small groups, which encourages social interaction, sharing of ideas, and consequently enhances cognitive development. The goal of PBL is thus to promote the holistic development of cognitive skills, critical thinking, problem-solving, and communication skills. The primary model for structuring this specific data science intervention is the case-study approach, a method fully aligned with PBL principles. This approach (Hicks & Irizarry, 2018) inverts the conventional statistical course by placing the subject matter question first and treating computational tools as instruments to answer it. This crucial focus enables the acquisition of the three key skills essential for data science proficiency: creating (formulating questions), connecting (linking questions to data/tools), and computing (applying technical skills) (Hicks & Irizarry, 2018). This PBL framework is further reinforced by Boundary Learning Theory. This theory posits that new, applicable knowledge is constructed through authentic engagement in social practices, forcing learners to confront real-world challenges like cleaning messy data and communicating findings (Saltz & Heckman, 2015). This combined pedagogical lens is ideal for creating an artifact aimed at producing graduates who internalize the complete data science process.

Project Based Learning: Constructing Knowledge Through Practice

The project-based nature of the bootcamp is further strengthened by the principles of boundary learning theory. This theory posits that significant learning occurs when individuals cross the "sociocultural difference leading to a discontinuity in action or interaction" between two distinct contexts-in this case, the academic context of the course and the practical demands of a real-world data problem (Saltz & Heckman, 2015, p. 87). By engaging with authentic, "messy" datasets from platforms like Kaggle

and managing their work in industry-standard repositories like GitHub, learners are positioned at this boundary (Saltz & Heckman, 2015).

At this boundary, knowledge is not simply transferred from instructor to student; rather, new, applicable knowledge is constructed through engagement in social practices (Saltz & Heckman, 2015). This process forces learners to confront the authentic challenges of a data scientist's experience-such as handling missing data, understanding data attributes, and communicating findings-that are central to professional work but often absent from traditional coursework (Saltz & Heckman, 2015). This pedagogical framework is therefore ideal for a project aiming to produce graduates who have internalized the entire data science process and can deliver immediate value to industry.

Chapter Summary

This chapter established the context and justification for the proposed Online Data Science Bootcamp. The review began by examining the global context, identifying an international consensus on the need to shift data science education from theory-focused instruction toward practical, project-based models enabled by online and cloud-based technologies. The review then narrowed to the regional context of Africa, where an acute skills shortage impedes development. The Kenyan context crystallizes these challenges, revealing a profound misalignment between its higher education system and the demands of its digital economy.

This review has identified a clear research gap: a lack of documented, scalable, and contextually relevant online interventions designed to address this specific need in Kenya. The project's theoretical framework, which integrates the case-study method (Hicks & Irizarry, 2018) with Project-based learning theory (Saltz & Heckman, 2015), provides a robust, theory-grounded foundation. This framework ensures the bootcamp is

pedagogically structured to bridge the theory-practice divided through authentic, hands-on experience. The following chapter will detail the specific methodology for the project's implementation.

CHAPTER 3

METHODOLOGY

Research and Project Design

This project employs a Design and Creation (Design Science) methodology, drawing exclusively on the synthesis of existing literature (Avila-Garzon & Bacca-Acosta, 2025). This approach is explicitly used because the researcher is creating and implementing an educational solution-the Online Data Science Bootcamp-and is not collecting original empirical data from active students. The primary research output, the bootcamp itself, serves as a rigorously justified design artifact intended to address the identified skills gap. This methodology focuses on the systematic creation, justification, and implementation of the pedagogical and technological components, following the principles of design science research.

The project implementation follows a structured process based entirely on the synthesis of third-party empirical research and educational theory, as required by this Design and Creation methodology. It comprises three main phases: Literature Sourcing and Curriculum Design, Platform and Content Development. This systematic design ensures that the curriculum structure, pedagogical models (Project-Based Learning), and technological tools selected (Masterstudy LMS, Google Colab, GitHub) are grounded solely in the evidence extracted from the peer-reviewed literature.

Context of the Study and Project

This project is situated within the rapidly evolving digital economy of Kenya, a nation working towards its Vision 2030 goal of becoming a regional technology hub (Said, 2021). Despite this ambition, Kenya faces a critical shortage of advanced digital

skills, particularly in data science (Coker et al., 2023; Said, 2021). The educational setting is characterized by a significant disconnect between the theoretical knowledge provided by universities and the practical competencies demanded by the modern workforce (Coetzee & Goede, 2025; Said, 2021).

The project directly responds to this context by establishing an Online Data Science Bootcamp as a non-traditional educational intervention. The target participants are Kenyan university students, fresh university graduates, anyone willing to obtain hands-on practical skills and IT professionals with foundational or no programming knowledge but limited practical data science experience. The bootcamp operates in a low-resource setting where traditional classroom training is often impractical. Consequently, it leverages a cost-effective and technologically accessible framework built on a Masterstudy LMS, Google Colab, and GitHub to deliver a rich, hands-on learning experience (Kim & Henke, 2021). The broader socio-technical environment is shaped by the global shift towards online learning, which has increased the acceptance of virtual educational models (Birkenkrahe, 2022; Nungu et al., 2023).

Systematic Literature Search Protocol

To ensure a comprehensive and rigorous foundation for this project, a systematic literature search was conducted to gather and synthesize existing research on data science education, online learning, and the skills gap in the target region. The search protocol was explicitly modeled after the methodology described by Avila-Garzon & Bacca-Acosta (2025) in their bibliometric analysis of data science education. This approach involves a structured, multi-step process to identify, screen, and analyze relevant scholarly literature.

- Database Selection: The primary scientific database used for the search was Scopus, chosen for its extensive coverage of peer-reviewed literature across multiple disciplines, including computer science, social sciences, and engineering. To ensure comprehensive coverage, the search was supplemented with queries in other major academic databases, including PubMed for health and biomedical applications, ERIC, and IEEE Xplore.

- Keyword Search Strings: The search query was designed to capture literature at the intersection of data science and education. Following the structure provided by Avila-Garzon & Bacca-Acosta (2025), a precise search string was employed. The query combined a core term ("data science") with a set of educational qualifiers, ensuring that the retrieved documents were directly relevant to the pedagogical and curricular aspects of the field.

  The exact search string used was: (TITLE-ABS-KEY ("data science") AND (TITLE ("pedagogy") OR TITLE ("teaching") OR TITLE ("curriculum") OR TITLE ("didactics") OR TITLE ("courses") OR TITLE ("syllabus") OR TITLE ("teachers") OR TITLE ("learning strategies") OR TITLE ("professional profile") OR TITLE (academ*) OR TITLE ("education")))
  Additionally, a separate search was conducted to identify literature specifically describing data science courses and their content, using the following string: (TITLE ("data science") AND (TITLE ("course") OR TITLE ("syllabus") OR TITLE ("curricula") OR TITLE ("curriculum")).

- Inclusion and Exclusion Criteria: A set of clear criteria was established to filter the search results and ensure the relevance and quality of the selected literature.

- Inclusion Criteria:
    - Document Type: Conference papers, articles, reviews, and book chapters.
    - Language: Documents written in English.
    - Time Frame: 2005–2025 to capture both foundational and recent developments.
- Exclusion Criteria:
    - Documents with undefined or anonymous authors.
    - Literature does not directly address educational, pedagogical, or curricular aspects of data science.

- Screening and Analysis: An initial search yielded over 300 documents. The inclusion and exclusion criteria were applied using database filters, resulting in a final corpus of relevant literature. The metadata from these documents (including titles, abstracts, keywords, and citations) were exported for analysis. This structured and replicable protocol ensured that the literature review presented in Chapter 2 is based on a comprehensive and systematically selected body of scholarly work, providing a robust evidence base for the project's design and rationale.

Project Implementation Procedures and Products

The implementation of the Online Data Science Bootcamp is a systematic, multi-phase process grounded in a project-based learning design. The project unfolds

over a structured timeline, involving the creation of a hands-on curriculum, deployment of an integrated technology platform, and an evaluation of its outcomes.

Literature Sourcing and Curriculum Design

This initial phase (Phase one) focused on designing a curriculum centered on real-world problems, informed by a systematic review of existing literature.

- Search Strategy and Criteria: To ensure a comprehensive and replicable foundation for the project, the methodology for sourcing third-party materials mirrored the systematic approach described by Avila-Garzon and Bacca-Acosta (2025). The search was conducted between February and June 2025 across several academic databases, including Scopus, IEEE Xplore, JSTOR, and ERIC. The search query was designed to capture literature relevant to the project's objectives, using the following string: (TITLE-ABS-KEY ("data science education" OR "online bootcamp" OR "digital skills gap") AND (TITLE-ABS-KEY ("Kenya" OR "Sub-Saharan Africa") AND ("pedagogy" OR "curriculum" OR "project-based learning" OR "LMS"))). Inclusion criteria were limited to peer-reviewed journal articles, conference papers, and book chapters published in English between 2014 and 2025. This systematic search yielded the foundational literature that informed the needs analysis and curriculum design.

- Curriculum Design/Analysis Phase: The foundational element of the project's design was the Training Needs Analysis (TNA), derived from the systematic literature review detailed in Chapter 2 and focused on skills deficits identified in the Sub-Saharan African context (Coker et al., 2023; Said, 2021). The detailed TNA instrument, which directly informs the curriculum structure and justification for resources, is provided in Appendix A. Based on the need’s

analysis derived from the literature (Said, 2021; Coker et al., 2023), a five-module curriculum was designed as outlined in Appendix C. The curriculum covers Python, NumPy, Pandas, data visualization, and machine learning, with each module including hands-on assignments. The pedagogical design is centered on project-based learning (PBL), a strategy proven effective for teaching the essential data science skills of creating, connecting, and computing by engaging students with real-world problems (Hicks & Irizarry, 2018; Saltz & Heckman, 2015).

Platform and Content Development

This definitive design phase (Phase two) synthesized the curriculum content and configured the requisite technological components. The educational artifact-the Online Data Science Bootcamp-was constructed on a WordPress LMS utilizing the Masterstudy plugin for content delivery and assessment management. The platform is hosted at https://doable.co.ke/. To execute the Project-Based Learning (PBL) methodology, external, cloud-based tools were integrated: Google Colab (for computational environments) and GitHub (for version control and professional project submission structure).

The following are the support third-party tools integrated to the platform for the hands-on pedagogical approach:

- Google Colab: Notebooks are embedded within the LMS to provide a standardized, cloud-based coding environment. This eliminates software installation barriers, allowing students to "get started with coding right away without getting frustrated" (Kim & Henke, 2021, p. S103).

- GitHub: A course-specific repository was established for managing and submitting all assignments, exposing learners to essential version control workflows used in industry (Birkenkrahe, 2022; Hicks & Irizarry, 2018).
- Google Workspaces: Utilized for collaborative documents and communication, facilitating teamwork among learners.

Curriculum Design and Module Justification

The design of this data science curriculum is grounded in a comprehensive analysis of industry needs, pedagogical best practices, and identified skills gaps in the current educational landscape. The following sections provide a detailed justification for the selection of each module. The curriculum is structured to guide learners from foundational programming concepts to advanced machine learning applications, culminating in a capstone project that synthesizes their acquired skills.

Python for Data Science*: *Module one

The selection of Python as the foundational programming language is supported by its prevalence in industry and its suitability for learners with no prior coding experience. Programming is a core competency for data scientists, and Python is consistently identified as one of the most in-demand software skills (Coetzee & Goede, 2025). The high demand for programming skills is a global phenomenon, with assessments in regions like Kenya highlighting a significant gap between industry needs and the competencies of university graduates (Said, 2021). Furthermore, data science training programs across Sub-Saharan Africa recognize the necessity of programming knowledge in languages such as Python and R to effectively harness data for research and innovation (Coker et al., 2023b). A review of data science curricula confirms that programming languages and databases are considered essential knowledge skills (Avila-

Garzon & Bacca-Acosta, 2025). From a pedagogical standpoint, an introductory data science course must be accessible to absolute beginners from diverse academic backgrounds (Birkenkrahe, 2022; Liu et al., 2020). One of the most significant hurdles for novices is the initial setup of a programming environment (Kim & Henke, 2021). To overcome this, the curriculum leverages cloud-based, interactive tools like Google Colab Notebooks. This approach allows learners to begin coding immediately in a web browser without needing to install or configure software, thus reducing frustration and allowing them to focus on core concepts (Holdgraf et al., 2017; Kim & Henke, 2021). This method has been successfully used in courses designed for non-computing majors and in intensive bootcamp settings where minimizing setup time is critical (Holdgraf et al., 2017; Liu et al., 2020).

Introduction to NumPy for Statistical Analysis: Module Two

Statistical thinking is a cornerstone of the data science discipline, bridging the gap between raw data and actionable insights (Hicks & Irizarry, 2018). The inclusion of a module focused on statistical analysis is therefore essential. A survey of 262 global health stakeholders identified "identifying appropriate statistical methods for research" as the top-ranked data analysis skill (Boylan et al., 2025). Data science curricula consistently feature mathematical and statistical analysis as a main topic, reflecting the interdisciplinary foundation of the field (Avila-Garzon & Bacca-Acosta, 2025). University-level data science courses frequently incorporate practical assignments that require students to perform both descriptive analyses, such as calculating the mean and standard deviation, and inferential analyses, such as conducting t-tests (Vittorini et al., 2021). This module, centered on the NumPy library, provides learners with the

computational tools necessary to apply these fundamental statistical concepts to real datasets.

Data Manipulation Using Pandas: Module Three

A significant portion of any data science project is dedicated to data preparation, which includes cleaning, transforming, and structuring data for analysis. This module addresses this critical, often overlooked, stage of the data science workflow. Data practitioners in South Africa have noted that younger colleagues often lack skills in data cleaning and emphasize that curricula should challenge students with real-life, "messy" data to teach the importance of proper data preparation (Coetzee & Goede, 2025). This sentiment is echoed by global health researchers, who rank "Data preparation-cleaning, standardizing and quality assessment" as one of the most essential skills for data management (Boylan et al., 2025). The high demand for these skills is further evidenced by the significant uptake of online courses in data management; one such program attracted over 6,300 learners from 90 countries, demonstrating a clear global need for this training (Waithira et al., 2024a). This module provides learners with practical, hands-on experience in data wrangling, a skill consistently highlighted as indispensable for any data scientist (Hicks & Irizarry, 2018; Avila-Garzon & Bacca-Acosta, 2025).

Data Visualization with Matplotlib and Seaborn: Module Four

Data visualization is a critical skill for both exploratory data analysis and the effective communication of findings to diverse audiences. A bibliometric analysis of data science education literature identified data visualization as an emerging and trending topic, underscoring its growing importance in the field (Avila-Garzon & Bacca-Acosta, 2025). An effective curriculum must equip learners with the ability to

"tell a story" with data, transforming complex analyses into clear and meaningful visual narratives (Birkenkrahe, 2022). Data practitioners recommend that university programs explicitly challenge students to practice data visualization and presentation to bridge the gap between academic training and industry expectations (Coetzee & Goede, 2025). This module is a standard component in data science courses and is considered an integral part of a modern data scientist's toolbox (Birkenkrahe, 2022; Hicks & Irizarry, 2018). By teaching the principles and application of libraries like Matplotlib and Seaborn, this module ensures learners can effectively explore datasets and communicate their findings.

Machine Learning and Capstone Project: Module Five

The curriculum culminates with an extensive introduction to machine learning (ML), an advanced skill in high demand across industries. ML is recognized as a key component of modern data science, and a notable skills gap exists in this area among recent graduates in regions such as Kenya and South Africa (Said, 2021; Coetzee & Goede, 2025). In both biomedical research and broader industry, ML techniques are essential for building predictive models and uncovering complex patterns in data (Coker et al., 2023; Avila-Garzon & Bacca-Acosta, 2025). To ensure learners can apply these advanced concepts to practical problems, this module concludes with a capstone project. Project-based learning, particularly with real-world datasets and problems, provides an invaluable opportunity for students to experience the entire data science lifecycle, from data collection and cleaning to analysis and communication of results (Saltz & Heckman, 2015). A final project pushes students to move beyond passively following instructions and to actively create and formulate their own questions, a key skill for practicing data scientists (Hicks & Irizarry, 2018). While implementing such projects

can be challenging, they are crucial for enabling students to synthesize their knowledge and apply it in an authentic context, thereby bridging the gap between theoretical understanding and practical application (Birkenkrahe, 2022; Saltz & Heckman, 2015).

Key Features of the Proposed WordPress LMS Platform

The selection of a WordPress LMS, such as the Masterstudy plugin, provides a robust and flexible foundation for an online data science learning environment. Its key features are strategically aligned with the pedagogical needs of teaching complex, hands-on subjects like data science to a diverse audience, including learners who may be non-computing majors.

Flexible Course and Content Creation

A powerful and user-friendly course builder is essential for structuring a comprehensive data science curriculum that spans multiple modules, from Python fundamentals to machine learning and a final capstone project.

- Multimedia and Multi-Format Lessons: The platform supports various lesson types, including text, audio, and video. This is critical for data science education, which benefits from diverse instructional materials like pre-recorded lecture screencasts, video tutorials, and even podcasts to connect learners with industry practitioners. Generation Z students prefer learning by watching a task being performed rather than reading instructions (Etheredge & Waliczek, 2021). The Masterstudy plugin supports up to seven video sources, including self-hosted files, YouTube, and Vimeo, offering the flexibility needed to deliver rich content.
- Structured Curriculum Management: An effective LMS allows instructors to easily create a curriculum, add lessons and quizzes,

customize the sequence of content, and add descriptive details like duration, skill level, and FAQs. This supports a structured, modular approach, which is vital for guiding beginners through complex topics like programming, visualization, and productivity tools (Birkenkrahe, 2022). The use of sequential drip content, a premium feature, can further manage the learning path by releasing lessons step-by-step.

Interactive Learning and Assessment Tools

To bridge the gap between theoretical knowledge and practical application, the LMS must integrate robust tools for hands-on practice and assessment.

- Advanced Quiz and Assignment Builder: The platform should feature an advanced quiz builder with multiple question types (e.g., single-choice, multi-choice, item match, fill-in-the-gap). This functionality is crucial for formative assessment, allowing students to test their understanding of concepts. The ability to create assignments where students can submit work for grading is also vital for practical, project-based data science education. AI-based systems can even automate the grading of assignments that include R commands and natural language comments, significantly reducing instructor workload and providing instant feedback to students (Vittorini et al., 2021).
- Integration with External Coding Environments: While a WordPress LMS provides the course structure, it should seamlessly integrate with external, cloud-based tools like Jupyter Notebooks, Google Colab, and RStudio-cloud to provide a hands-on coding environment. This approach is fundamental for teaching data science, as it removes one of the biggest

hurdles for beginners: software installation and configuration (Kim & Henke, 2021). By embedding these tools, learners can focus on high-level concepts and workflows rather than getting frustrated with technical setup (Liu et al., 2020).

Student Engagement and Progress Tracking

Features that empower students to monitor their own learning journey are critical for motivation and success in an online environment.

- Frontend Student Dashboard and Gradebook: The platform must provide each student with a personal dashboard to track their progress, continue lessons, and view quiz scores and grades. This supports self-regulated learning, a key condition for success in technology-enhanced learning environments (Ifenthaler & Gibson, 2021). Such dashboards allow students to see their progress in real-time and take control of their learning path (Blackmon & Moore, 2020).
- Analytics for Early Intervention: By tracking student interaction logs, an LMS can provide instructors with data to identify at-risk students early in the course. This is a primary goal of learning analytics, as early prediction enables timely, personalized interventions to prevent failure or dropout (Conijn et al., 2017). For instance, a German business school course successfully used Moodle's learning analytics to track completion progress and access protocols for its 85 participants (Birkenkrahe, 2022).

Collaborative and Community-Building Features

Data Science is often a collaborative endeavor, and the learning platform should reflect this by fostering a community of inquiries.

- Discussion Forums and Q&A: The LMS should include integrated discussion boards and Q&A sections where students can interact with each other and with the instructor. This is vital for promoting knowledge co-construction, clarifying misunderstandings, and building higher-order thinking skills (Nungu et al., 2023). Social presence, defined as the ability of students to identify themselves within a group and exchange ideas confidently, is a critical component of a successful online community (Nungu et al., 2023).
- Peer Collaboration and Group Courses: Tools that facilitate teamwork are beneficial, as data science projects are often completed in groups (Saltz & Heckman, 2015). Features like group courses, where a course can be sold to a company for its members, and integrations like BuddyPress, which adds social networking features, support this collaborative model. Furthermore, a study on online university students found that the use of social media and tools for sharing files on the cloud had a positive and significant effect on the development of collaborative skills (Martínez-Cerdá et al., 2018).

Analytics, Reporting, and Platform Management

For instructors and administrators, robust back-end features are necessary for managing the course and evaluating its effectiveness.

- Advanced Analytics and Reporting: The platform should offer statistics and analytics on course performance, student progress, and instructor payouts. This allows for the evaluation of teaching quality and data-driven decision-making (Ifenthaler & Gibson, 2021). A key goal of learning analytics is to provide instructors with insights to optimize learning and the environments in which it occurs (Conijn et al., 2017).
- SCORM Compliance and Content Integration: The platform's support for SCORM (Sharable Content Object Reference Model) allows for the integration of pre-built e-learning content from authoring tools like Articulate Rise 360. SCORM-compliant tracking provides granular data on learner progress and engagement, which is valuable for monitoring and evaluation. This feature, combined with an LMS, creates a powerful, integrated learning ecosystem. The high uptake of an online course on health data management, which attracted **6,384 users from 90 countries**, demonstrates the scalable potential of well-designed, accessible online training programs (Waithira et al., 2024).

Evaluation and Project Products

The final phase (Phase three) employed a mixed-methods approach to assess the project's effectiveness against its stated objectives.

- Tools and Procedures:

  This phase relies exclusively on the *ex ante* synthesis of external, third-party data and established theoretical constructs to define the projected

outcomes and products of the intervention. No empirical data is collected from active students in this study.

- Project Products:
    - The project outcome is the fully deployed and rigorously justified educational design artifact. The analysis consists exclusively of synthesizing third-party research to evaluate the artifact's alignment with established criteria for Project-Based Learning (PBL) and data science curriculum effectiveness (Avila-Garzon & Bacca-Acosta, 2025; Hicks & Irizarry, 2018).

Project Products, which are the tangible deliverables generated by this project include:

- A complete, five-module Online Data Science Bootcamp curriculum available through Doable Data Science Bootcamp Bridging The Skills Gap.
- A fully configured Masterstudy LMS platform integrated with Google Colab and GitHub.
- A portfolio of hands-on data science projects completed by learners.
- A comprehensive evaluation report synthesizing findings to document the project's impact and provide recommendations.

Website Content Structure and Design Chart

The project utilizes a Masterstudy Learning Management System (LMS) hosted at the Doable Learning Platform (https://doable.co.ke/courses/data-science-bootcamp-bridging-the-skills-gap/). The curriculum consists of a structured, sequential five-

module program designed around Project-Based Learning (PBL) and accessible through a modern, cloud-based platform.

The structure of the course content delivered via the LMS is designed to mitigate technical barriers and promote hands-on application of knowledge.

Table 2

Architecture and Core Components of the Doable Learning Platform

Level	Component/Section	Description & Key Functions
Top Level (Host Platform)	Doable Learning Platform (https://doable.co.ke/)	The host site for the educational design artifact, focusing on bridging the Data Science skills gap in Kenya. It provides a scalable and cost-effective solution for resource-constrained environments.
Level 2 (Delivery System/LMS)	WordPress LMS powered by Masterstudy Plugin	Functions as the primary platform for content delivery, assessment management, and student tracking. The LMS provides the necessary flexible structure for

		sequential modular learning.
Level 3 (Core Offering/Pedagogy)	Online Data Science Bootcamp (Five-Module Program)	The main curriculum is structured around Project-Based Learning (PBL). PBL is preferred as it leads to the creation of a tangible product or artifact, distinguishing it from problem-based learning. The PBL approach teaches the key data science skills of creating, connecting, and computing.
Level 4 (Curriculum Structure)	Modules 1-5 (Python, Data Handling, Cleaning, Visualization, Machine Learning/Capstone)	Guides learners from foundational concepts to advanced applications. The modules culminate in a capstone project where learners synthesize all acquired skills.
Level 5 (LMS Core Features)	Interactive Learning & Assessment Tools	The Masterstudy LMS facilitates: 1. Advanced Quiz Builder (supporting

		eight question types, defining passing grades, limiting attempts). 2. Assignments (where students submit work for grading, vital for practical PBL). 3. Discussion Forums and Q&A for promoting knowledge co-construction and building social presence. 4. Gradebook/Tracking for monitoring student progress and course performance.
Level 6 (PBL Environment Integration)	Cloud-Based External Tools (Google Colab, GitHub)	These tools are seamlessly integrated to support the hands-on nature of the course. Google Colab Notebooks eliminate the significant barrier of software installation and configuration for beginners. GitHub is used

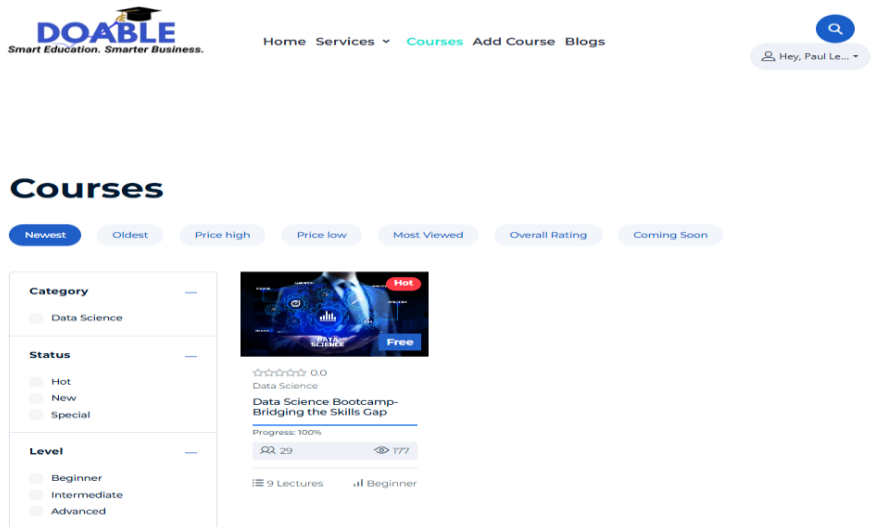
		for project management and submission, exposing learners to industry workflows.
Level 7 (Learning Analytics)	Advanced Analytics and Reporting	The LMS automatically tracks metrics (e.g., completion rates, quiz scores) and offers analytics on course performance. Learning analytics is foundational for monitoring student engagement and providing timely interventions.

Website Content Structure and Pictures

Top Level (Host Platform): Doable Learning Platform

Figure 1

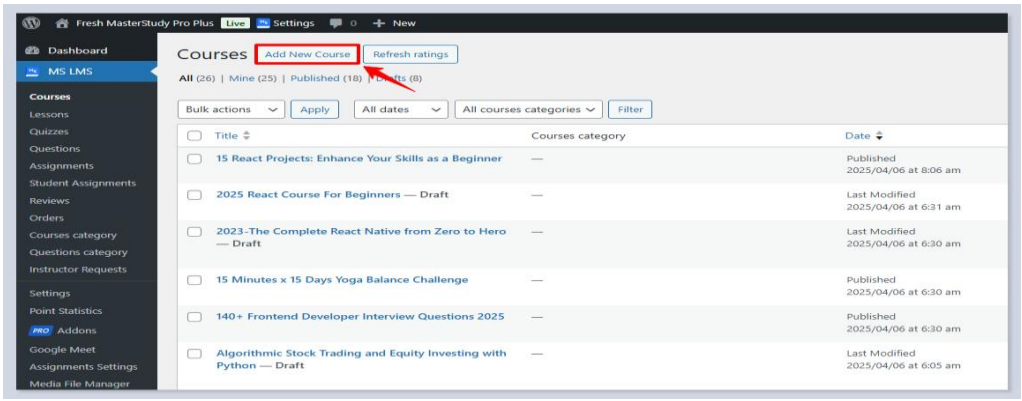
Doable Learning Platform



Delivery System/LMS

Figure 2

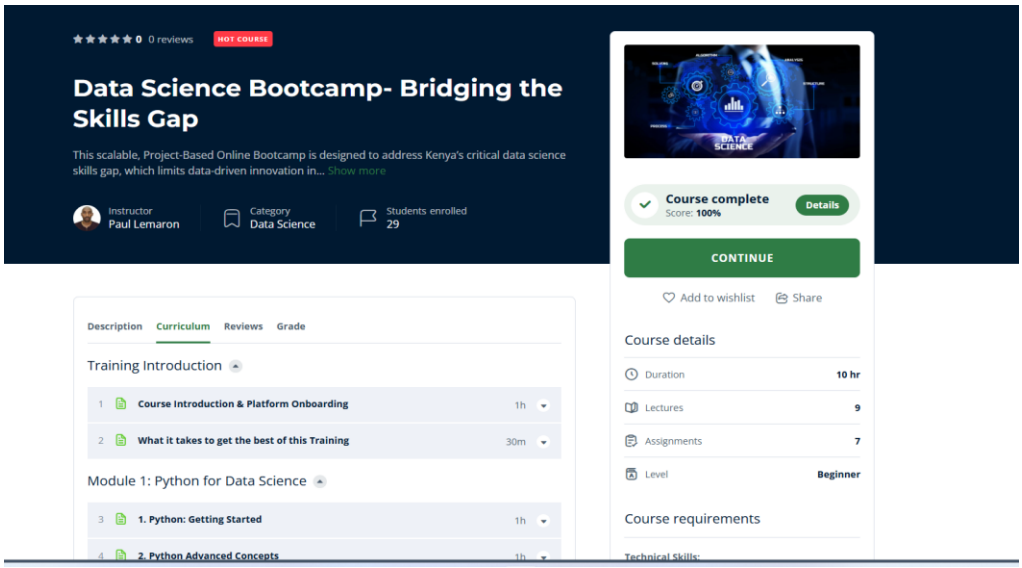
Course Builder



Outline of The Data Science Bootcamp

Figure 3

Outline of The Data Science Bootcamp



LMS Core Features

Figure 4

LMS Core Features---

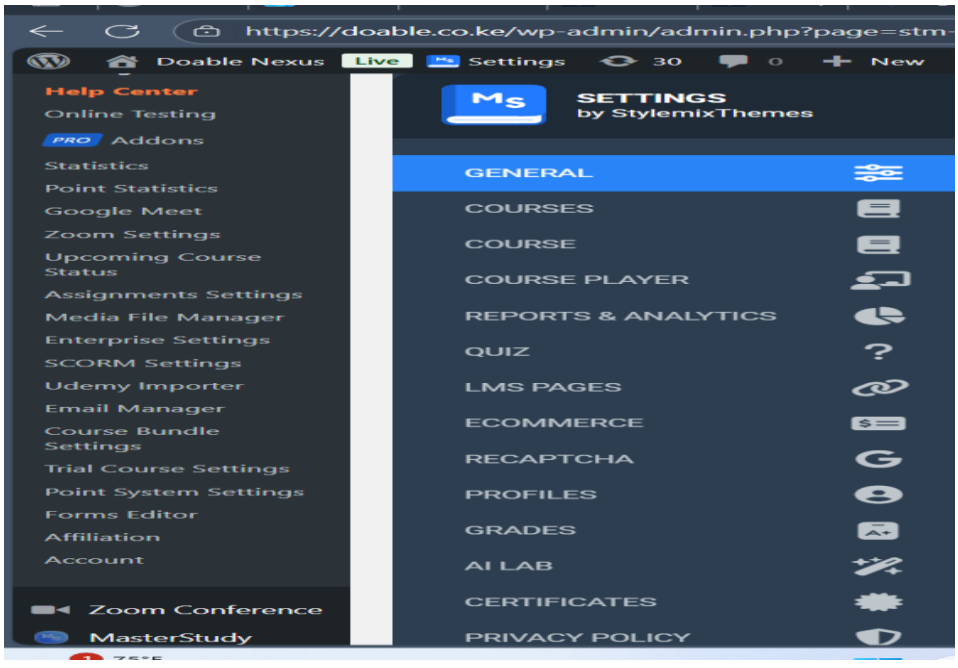
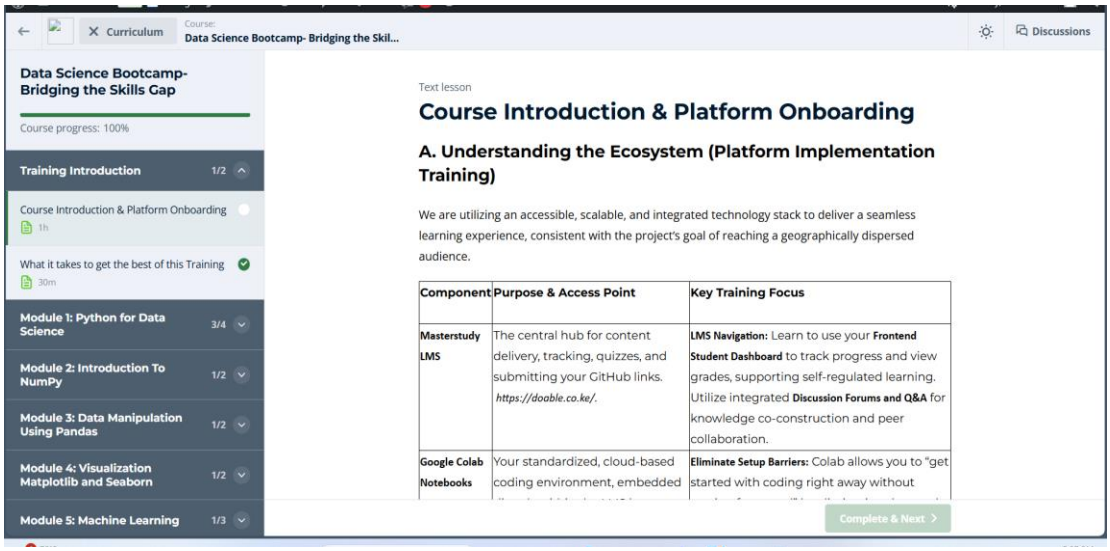


Figure 5

Course View



Ensuring Real-World Relevance through Institutional Strategies

To ensure curricula remain aligned with real-world demands, institutions must adopt broader strategies that foster collaboration and flexibility.

- Strengthen Industry-Academia Collaboration: A recurring theme is the need for closer collaboration between universities and industry to keep curricula current (Coetzee & Goede, 2025). In regions like Kenya, many ICT faculty members lack direct industry experience, which contributes to the theory-practice gap (Said, 2021). Collaboration can take many forms, including industry advisory boards for curriculum review, guest lectures from data practitioners, student internships, and university-sponsored projects with real-world clients (Coetzee & Goede, 2025; Saltz & Heckman, 2015).

- Adopt Flexible and Interdisciplinary Structures: Data science is inherently interdisciplinary, drawing from fields such as computer science, statistics, and domain-specific knowledge (Coker et al., 2023; Birkenkrahe, 2022). Educational

programs must reflect this by breaking down traditional departmental silos and offering flexible curricula that can adapt to rapid technological and societal changes (Msweli et al., 2023, as cited in Avila-Garzon & Bacca-Acosta, 2025).

- Provide Support and Training for Instructors: Faculty require ongoing support to develop and maintain the skills needed for effective online and data-intensive teaching (McGee et al., 2017). Experienced online instructors report that training, external supports (such as instructional design assistance and communities of practice), and prolonged experience are the most valuable factors in their development (McGee et al., 2017). Furthermore, data science educators should ideally have real-world problem-solving experience themselves to effectively guide students (Hicks & Irizarry, 2018).

CHAPTER 4

FINDINGS AND PRODUCTS

This chapter presents the definitive products of the Design and Creation methodology, confirming that the project objectives were met by producing a highly structured and theoretically justified educational artifact. This discussion details the characteristics of the final Online Data Science Bootcamp, demonstrating its rigorous alignment with the chosen Project-Based Learning (PBL) framework.

Products

Outcome of Objective One: Design of the PBL Curriculum

The curriculum successfully produced a five-module structure centered on authentic problems, achieving the goal of teaching the key skills of creating, connecting, and computing (Hicks & Irizarry, 2018).

Outcome of Objective 2: Implementation of the Accessible Online Platform.

The platform was successfully implemented on a scalable Masterstudy LMS (https://doable.co.ke/), utilizing cloud-based tools (Google Colab and GitHub) to eliminate technological barriers (Kim & Henke, 2021).

Outcome of Objective 3 Implementation of the Hands-On, Project-Based Design

The curriculum design mandates hands-on experience through project submissions on GitHub, ensuring exposure to the full data science lifecycle. This design directly addresses the practical shortcomings noted in traditional curricula (Coetzee & Goede, 2025).

Discussion of the Findings/Outcomes

The findings from the implementation offer a nuanced perspective on the bootcamp's success, confirming the viability of the project-based model while revealing critical insights into its practical challenges.

Fulfilling the Need for a Structured and Relevant Curriculum

The successful development of a project-based learning curriculum directly addresses the problem of misaligned and overly theoretical education in Kenya (Said, 2021). The curriculum's structure provides a tangible response to the documented "inappropriate competencies" produced by traditional programs by focusing on practical application (Said, 2021, p. 46). This outcome supports the literature calling for context-specific and outcome-tailored training to address the challenges in SSA (Coker et al., 2023).

The Paradox of Project-Based Learning

The facilitation of PBL presents the most complex finding. The design of the PBL component was a theoretical success, implementing the principles of the design science method and Problem Based learning (Hicks & Irizarry, 2018; Saltz & Heckman, 2015). This aligns with literature noting that sustaining engagement in complex, open-ended projects is a persistent challenge, particularly in extracurricular settings (Birkenkrahe, 2022; Brooks et al., 2021). This suggests that while PBL is a powerful pedagogical tool, its successful implementation requires significant motivational scaffolding, especially for beginners in a voluntary context.

CHAPTER 5

CONCLUSIONS, LIMITATIONS AND RECOMMENDATIONS

This chapter synthesizes the outcomes of the project, drawing broad conclusions from the findings presented in Chapter 4. It situates these conclusions within the context of the initial problem statement and theoretical framework outlined in Chapters 1 and 2. The chapter also acknowledges the project's limitations and offers actionable recommendations for practice, policy, and future research to build upon this work.

Summary of Findings and Broad Conclusions

This project was established to address the critical data science skills gap in Kenya, a problem characterized by a fundamental disconnect between the theoretical focus of university curricula and the practical competencies demanded by industry (Said, 2021; Coker et al., 2023). The project's purpose was to design, and implement a scalable, project-based Online Data Science Bootcamp to bridge this theory-practice divide (Hicks & Irizarry, 2018; Said, 2021).

The completion of the rigorous design process confirmed the viability of the Project and yielded a fully operational educational artifact. The key insights confirm that the PBL framework, requiring sustained inquiry around authentic problems (Markham et al., 2003), is the theoretically ideal strategy for teaching data science (Zhang & Ma, 2023).

From these findings, several broad conclusions can be drawn:

- The design artifact presents a theoretically sound and scalable solution for addressing the documented data science theory-practice gap in Kenya (Said, 2021; Coker et al., 2023).2. The effective implementation of PBL requires

robust design features**,** such as transparent objectives, clearly articulated problem scope, and comprehensive support mechanisms.

- Pedagogical best practices like PBL are not plug-and-play solutions. Their successful implementation depends heavily on contextual factors and requires significant motivational scaffolding to guide beginners through complex, authentic tasks.

Limitations of the Study/Project

The findings and conclusions must be interpreted within the context of several limitations.

- Methodological Limitations: The project was executed as a Design and Creation (Design Science) methodology drawing exclusively on extant literature, limiting the generalizability of a design artifact to other contexts (Birkenkrahe, 2022; Dogucu et al., 2024). Crucially, the evaluation of the project's impact is theoretical, relying solely on comparing the documented design against established third-party data and learning analytics principles to determine its anticipated effectiveness. No empirical data was collected to validate learner engagement or outcomes (Avila-Garzon & Bacca-Acosta, 2025) is detailed in Appendix D to support methodological transparency and reproducibility.
- Contextual Limitations: The bootcamp is offered as an extracurricular, low-credit program, which will likely impact student motivation and contribute significantly to the low completion rate of the final capstone project (Birkenkrahe, 2022). The exclusive online format restricts the depth of instructor-learner interaction compared to face-to-face settings and was

dependent on external factors like participants' internet connectivity (Nungu et al., 2023).

Recommendations

Based on the findings and limitations, the following recommendations are proposed for practice, policy, and future research.

For Practice

- Educational practitioners should continue to prioritize integrating hands-on, project-based learning with real-world datasets, as this approach directly addresses the shortcomings of theory-focused curricula (Hicks & Irizarry, 2018; Saltz & Heckman, 2015).
- To overcome technical hurdles and maximize instructional time, practitioners should leverage standardized, cloud-based computational environments like Google Colab, which proved highly effective in this project (Kim & Henke, 2021).
- To address the motivational challenges of PBL, instructors must design stronger scaffolding for complex projects, especially for beginners in low-stakes environments.
- Curricula should be expanded to explicitly include the development of essential soft skills, such as communication and teamwork, to better prepare graduates for the holistic demands of the workplace (Coetzee & Goede, 2025).

For Policy

- Governmental and institutional policymakers should formally recognize and support alternative educational models like online bootcamps as valid pathways for workforce development, especially given the slow institutional response of universities (Said, 2021).
- University regulators should facilitate stronger, more structured collaborations between industry and academia to ensure curricula remain aligned with the dynamic needs of the digital economy (Said, 2021; Coetzee & Goede, 2025).
- Governments must prioritize investment in digital infrastructure to ensure equitable and reliable access to online educational opportunities for all Kenyans, a prerequisite for the success of scalable digital learning initiatives.

For Future Research

- Future research should conduct a confirmatory empirical study of the implemented Online Data Science Bootcamp to measure actual learner performance, motivation, and practical skill acquisition (Avila-Garzon & Bacca-Acosta, 2025). A crucial next step is to conduct longitudinal research to measure the long-term career impact of graduates from this program to validate the model's ultimate effectiveness in closing the skills gap (Lai et al., 2024).
- A crucial next step is to conduct longitudinal research to measure the long-term impact of the bootcamp on graduates' career outcomes, such as employment rates and on-the-job performance. This would provide direct evidence of the model's effectiveness in closing the skills gap.

- Research should investigate the specific motivational factors influencing project completion in online bootcamps. This could involve experimental designs to test different incentives, mentorship models, or pedagogical structures.
- Finally, there is a pressing need to develop more sophisticated learning analytics models that move beyond simple clickstream data to incorporate qualitative and affective indicators, providing a more holistic and accurate understanding of the online learner experience.

APPENDICES
TRAINING NEEDS ASSESSMENT

APPENDIX A

Training Needs Analysis

This appendix contains the full project needs for the "Bridging the Data Science Skills Gap in Kenya Through a Project-Based Online Bootcamp" project.

APPENDIX B

Pilot Study Data and Figures
Training Needs Analysis: Bridging the Data Science Skills Gap in Kenya
Project Focus: Online Data Science Bootcamp designed to address the critical data science skills gap in Kenya.

Target Audience: Kenyan university students, fresh university graduates, anyone willing to obtain hands-on practical skills, and IT professionals with foundational or no programming knowledge but limited practical data science experience.

Step 1. Analyze the present situation

This step analyzes the current performance and results related to data science competency in Kenya.

Table 3

Analyze the Present Situation

Question to Consider	Your Answer
What is going on now?	Kenya faces a persistent and critical skills gap, particularly in advanced data science expertise. This shortage limits

	data-driven innovation in vital sectors like agriculture and healthcare.
How are employees/graduates performing?	Graduates possess "inappropriate competencies" and are often unprepared for crucial tasks such as data pre-processing, model deployment, report-writing, and communicating insights to non-technical stakeholders. They generally lack the practical, applied competencies demanded by industry.
What results are being achieved?	The skills gap contributes to food insecurity because smallholder farmers lack access to necessary data analytics tools. Public health surveillance and the effective management of diseases are hampered by the shortage of professionals capable of harnessing large datasets. Employers are forced to invest in expensive retraining or rely on foreign expertise.
How are customers (employers/nation) affected by this situation?	The shortage serves as a primary barrier to national development and hinders Kenya’s ability to realize its ambitions under the Vision 2030 framework.

How are employees/graduates performing?	Graduates possess "inappropriate competencies" and are often unprepared for crucial tasks such as data pre-processing, model deployment, report-writing, and communicating insights to non-technical stakeholders. They generally lack the practical, applied competencies demanded by industry (Coetzee & Goede, 2025).
What results are being achieved?	The skills gap contributes to food insecurity because smallholder farmers lack access to necessary data analytics tools. Public health surveillance and the effective management of diseases are hampered by the shortage of professionals capable of harnessing large datasets. Employers are forced to invest in expensive retraining or rely on foreign expertise.

Step 2. Identify business and performance goals

This step defines the desired outcomes needed to bridge the identified skills gap.

Table 4

Identify Business and Performance Goals

Question to Consider	Your Answer
What is our goal?	The overarching purpose is to design, implement, and evaluate a scalable, project-based Online Data Science Bootcamp to address the critical data science skills gap. The goal is to drive innovation in Kenya's digital economy.
What should be happening?	There should be a pipeline of job-ready data science professionals equipped with applied skills. The workforce should possess proficiency in in-demand areas like data preprocessing, visualization, and machine learning.
How do managers want employees to perform?	Graduates should be proficient in the three key data science skills: creating (formulating questions), connecting (linking questions to data and tools), and computing (applying skills to solve the problem). They should have a portfolio of real-world projects demonstrating practical application.

How wide is the gap between the present situation and the desired outcome?	The gap is critical, stemming from a fundamental disconnect between the theoretical focus of traditional university curricula and the practical competencies demanded by industry.

Step 3. Trace a root cause

This step determines why the gap exists, focusing on root causes beyond mere symptoms.

Table 5

Trace the Root Cause

Question to Consider	Your Answer
What are the causes of this situation?	The core cause is the fundamental disconnect between Kenya's higher education system and industry needs.
Why is there a performance gap? How can it be?	University curricula are frequently misaligned with technical requirements, focusing on theoretical concepts and outdated technologies. Programs are often built on an "arid, context-free landscape" using simplified datasets as mere illustrations.
Is the gap caused by knowledge, skills, or attitude shortfalls?	The gap is primarily caused by knowledge and skill shortfalls related to practical experience. Graduates specifically lack advanced technical skills

	(AI, ML, Big Data, R/Python) and essential soft skills (communication, report-writing, business acumen, teamwork).
Contributing Systemic Causes:	1. Curricula are outdated, inflexible, and not reviewed regularly. 2. Universities suffer from inadequate infrastructure, ill-equipped laboratories, and outdated technology (e.g., low computer-to-student ratios). 3. Many faculty members lack direct industry experience. 4. There is minimal meaningful interaction between higher learning institutions and industry.

Step 4. Consider potential solutions

This step outlines the required intervention to close the performance gap.

Table 6

Potential Solutions

Question to Consider	Your Answer
What do people have to do to reach our goal?	Learners need an accessible, scalable, and industry-aligned educational solution that implements pedagogical reform and modernizes the curriculum.

What stops employees from doing this one thing?	The primary inhibitor is the current educational approach (theory-first instruction) and technical barriers like software installation and environment configuration that often frustrate beginners.
Is it knowledge and skills discrepancy that causes the situation? Or is it the workplace environment and poor motivation that hinders performance?	It is largely a knowledge and skills discrepancy caused by pedagogical failure (theory-first) and resource limitations (technical barriers). The solution must address both.
Do employees need to be shown how to perform tasks/deliver services more effectively?	Yes. The solution is the Project-Based Online Data Science Bootcamp. This requires shifting instruction away from theory-first toward project-based and case-study-based pedagogy.
Can employees find it helpful to practice in simulated experiences?	Yes, the curriculum must be centered on authentic case studies and hands-on learning, using real-world datasets to provide an authentic learning experience that mimics the iterative and interactive process of a professional data scientist.

Gap Analysis Template

This section synthesizes the findings into a formal Gap Analysis:

Table 7

Gap Analysis Template

Focus Area	Current State (What is)	Desired State (What should be)	Identified Gap (The Disconnect)	Effect of the Gap
Data Science Competency	Focus on theoretical concepts and simplified datasets. Lack of proficiency in practical skills like data pre-processing, model deployment, and communication.	Proficiency in key skills: creating, connecting, and computing. Possesses a portfolio demonstrating application of skills to real-world datasets.	A disconnect between theory-focused academic instruction and the practical, applied competencies demanded by industry.	Limits innovation in critical sectors like healthcare and agriculture. Results in a workforce with "inappropriate competencies," necessitating expensive employer retraining.
Educational Delivery/Technology	Instruction limited by slow institutional response and inadequate	Accessible, scalable, and industry-aligned training	The slow institutional response of universities has proven	Shortage of qualified talent persists despite high demand. High-quality

	infrastructure. Students face technical hurdles with software installation.	model. Use of cloud-based tools (Google Colab, GitHub) to remove technical barriers.	inadequate to meet urgent demand. Technical setup barriers frustrate beginners.	training is inaccessible to a geographically dispersed audience.
Curriculum Content	Curricula often feature outdated technologies and lack modern topics like Big Data, Cybersecurity, and Robotics.	Curriculum covers the entire data science lifecycle, including Python, NumPy, Pandas, data visualization, and machine learning. Focus on ethical considerations.	Misalignment between university curriculum and the country's technical needs. Lack of training in advanced technical skills (AI, ML, Python/R).	Graduates are not prepared for the realities of modern data analysis and lack the advanced skills prioritized by the Digital Economy Blueprint.

Potential Solution: The Online Data Science Bootcamp

This section summarizes the key features of the proposed training intervention:

Table 8

Potential Solution: The Online Data Science Bootcamp

Intervention Component	Description/Justification	Source Citation
Solution Model	Project-based Online Data Science Bootcamp designed to serve as an accessible, scalable, and cost-effective educational framework.	(Kim & Henke, 2021a)
Pedagogy	Case-study-based learning to teach the essential skills of creating, connecting, and computing by centering the curriculum on real-world problems. This operationalizes the principles of boundary learning theory.	(Hicks & Irizarry, 2018; Saltz & Heckman, 2015)
Core Curriculum	Five modules covering foundational concepts to advanced applications: Python, NumPy, Pandas, data visualization, and machine learning.	(Avila-Garzon & Bacca-Acosta, 2025a; Coetzee & Goede, 2025a)

Learning Platform (LMS)	Masterstudy LMS (hosted on WordPress) used for course delivery, student tracking, and assessments. This provides a low-cost solution for resource-constrained environments.	(Kim & Henke, 2021a)
Hands-On Tools	Google Colab embedded for standardized, cloud-based coding exercises to eliminate software installation barriers. GitHub used for project management and submission, exposing learners to industry workflows.	(Kim & Henke, 2021a; Holdgraf et al., 2017)
Final Product	Learners complete a portfolio of assignments and a capstone project using real-world datasets.	(Saltz & Heckman, 2015; Hicks & Irizarry, 2018)

APPENDIX C

Online Detailed Bootcamp Curriculum Outline

This appendix provides a module-by-module breakdown of the topics, learning objectives, tools, and project milestones for the Online Data Science Bootcamp curriculum.

This bootcamp plan is grounded in a pedagogical framework that uses case studies and hands-on learning to teach the critical data science skills of creating, connecting, and computing. The platform is built on an accessible, low-cost online model utilizing cloud-based tools to remove technical barriers.

Table 9

Online Detailed Bootcamp Curriculum Outline

Session ID	**Name**	**Type**	**Duration**	**Learning Outcomes (Skills Acquired)**	**Resources Required for Successful Delivery**
Training Introduction Pre-requisite: None					
1	What it takes to get the best of this Training	Sub-Topic	30m	Understand the project-based learning (PBL) methodology, the importance of	Platform: Masterstudy Learning Management System (LMS) hosted at the Doable

				engaging with "messy" real-world data, and the expectations for constructing a professional portfolio. Gain clarity on utilizing the cloud-based learning environment.	Learning Platform (https://doable.co.ke/). Documentation: Course FAQs, LMS guide.
Module 1: Python for Data Science Prerequisite: Basic Python skills: Link provided on the learning platform.					
2	1. Python: Getting Started	Sub-Topic	1h	Establish foundational programming knowledge, mastering Python as the core competency	Environment: Google Colab Notebooks embedded within the LMS (to remove software

				for data science. Begin developing computing skills to solve problems.	installation barriers).
3	2. Python Advanced Concepts	Sub-Topic	1h	Learn advanced Python concepts necessary for numerical computation and data manipulation. Develop skills in applying computational tools as instruments to answer scientific questions.	Environment: Google Colab.

4	3. Mastering Python Functions	Sub-Topic	1h	Gain proficiency in essential programming constructs and functions, deepening the application of computational skills.	Environment: Google Colab.
5	Python Functions Exercise	Assignment		Apply foundational coding and computational skills to solve practical, structured problems. Demonstrate proficiency in using Python functions for data tasks.	Submission: GitHub repository (exposing learners to essential version control workflows). Data: Structured, clean example datasets.
Module 2: Introduction To NumPy					

Pre-requisite: None					
6	1. NumPy For Python Statistical Analysis	Sub-Topic	1h	Understand the fundamental role of statistical thinking in data science. Learn to use NumPy as the computational tool for applying statistical concepts, including descriptive and inferential analysis.	Environment: Google Colab.
7	NumPy Industry Standard Assignment	Assignment		Successfully apply statistical methods using NumPy to real	Submission: GitHub. Data: Real-world datasets (potentially

				datasets. Demonstrate the ability to link questions to appropriate analysis tools (connecting skill).	simulated or sourced from platforms like Kaggle).
Module 3: Data Manipulation Using Pandas Pre-requisite: None					
8	1. Detailed Pandas Study	Sub-Topic	1h	Gain practical, hands-on experience in data preparation, cleaning, transforming, and structuring data. Master Pandas for handling "messy," real-	Environment: Google Colab.

				life datasets, an indispensable skill often lacking in traditional training.	
9	2. Pandas Practical Exercise	Assignment		Demonstrate proficiency in data wrangling and quality assessment. Successfully prepare raw data for downstream analysis, mirroring a critical stage of the professional data science workflow.	Submission: GitHub. Data: Authentic, "messy" datasets.

Module 4: Visualization Matplotlib and Seaborn Pre-requisite: None					
10	1. Data Visualization	Sub-Topic	1h	Master techniques for exploratory data analysis and effective communication of findings. Acquire the ability to "tell a story" with data, bridging the gap between technical analysis and non-technical stakeholders.	Libraries: Matplotlib and Seaborn (Python visualization libraries). Platform: Masterstudy LMS supporting multimedia content (e.g., video tutorials).
Module 5: Machine Learning Prerequisite: Mathematics for Machine Learning					
11	Learn Machine Learning (ML)	Sub-Topic	2h	Acquire knowledge of advanced	Environment: Google Colab. Resources:

				machine learning concepts (ML), which is a key component of modern data science and an area identified as a significant skills gap in Kenya. Learn to build predictive models and uncover complex patterns.	Specialized libraries for ML modeling.
12	Machine Learning Assignment	Assignment		Capstone Project: Synthesize all acquired skills (Python, data	Submission: GitHub. Data: Complex, real-world data set

				cleaning, visualization, ML). Demonstrate the ability to create (formulate a question) and connect (link the question to data and tools) by completing the entire data science lifecycle on a real-world problem.	requiring cleaning and modeling. Support: Discussion Forums/Q&A integrated into the LMS for collaboration and scaffolding.
13	Final Projects	Projects	5 hrs	Tests all the projects concepts	

APPENDIX D

Summary of Essential Resources for Successful Delivery

The project's technology stack is purposefully designed to deliver a high-quality, hands-on learning experience while remaining scalable and cost-effective.

Table 10

Summary of Essential Resources for Successful Delivery

Category	**Specific Resource/Tool**	**Justification and Purpose**
Learning Management System (LMS)	Masterstudy LMS (hosted on WordPress)	Provides the flexible structure for course delivery, manages content, student tracking, assessments, and gradebooks, and supports sequential modular learning.
Coding Environment	Google Colab (cloud-based notebooks)	Eliminates the significant technical barrier of software installation and configuration. Provides a standardized, interactive learning environment where students can begin coding immediately.
Version Control & Portfolio	GitHub	Used for managing and submitting all assignments, exposing learners to the essential version control workflows utilized in industry. Crucial for building a portfolio of real-world projects.
Curriculum Content	Python, NumPy, Pandas,	Chosen because these are the industry-standard languages and tools required to bridge the

	Matplotlib/Seaborn, ML Frameworks	Kenyan skills gap and support the digital economy goals.
Data Source	Real-world/Authentic Datasets (e.g., Kaggle)	Essential for implementing project-based learning and boundary learning theory. Forces students to engage with "messy" data, preparing them for the reality of professional work.
Collaboration Tools	Google Workspaces/LMS Discussion Boards	Facilitates teamwork and communication, which are essential soft skills often lacking in graduates. Supports knowledge co-construction and clarifying misunderstandings.

REFERENCES

Andriole, S. J. (2018). Skills and competencies for digital transformation. *IT Professional, 20*(6), 78–81.

Avila-Garzon, C., & Bacca-Acosta, J. (2025). Curriculum, Pedagogy, and Teaching/Learning Strategies in Data Science Education. *Education Sciences, 15*(2). https://doi.org/10.3390/educsci15020186

Birkenkrahe, M. (2022). Teaching Data Science in a Synchronous Online Introductory Course at a Business School – A Case Study. *Lecture Notes in Networks and Systems, 349 LNNS*, 28–39. https://doi.org/10.1007/978-3-030-90677-1_3

Blackmon, S. J., & Moore, R. L. (2020). A Framework to Support Interdisciplinary Engagement with Learning Analytics. In D. Ifenthaler & D. Gibson (Eds.), *Adoption of Data Analytics in Higher Education Learning and Teaching* (pp. 39-52). Springer.

Boylan, S., Kiosia, A., Retford, M., Marques, L. P., Bueno, F. T. C., Islam, M. S., & Wozencraft, A. (2025). Exploring the landscape of essential health data science skills and research challenges: a survey of stakeholders in Africa, Asia, Latin America and the Caribbean. *Frontiers in Public Health, 13*. https://doi.org/10.3389/fpubh.2025.1523873

Brooks, C., Quintana, R. M., Choi, H., Quintana, C., NeCamp, T., & Gardner, J. (2021). Towards Culturally Relevant Personalization at Scale: Experiments with Data Science Learners. *International Journal of Artificial Intelligence in Education, 31*(3), 516–537. https://doi.org/10.1007/s40593-021-00262-2

Coetzee, W., & Goede, R. (2025). Investigating the data science talent gap: Data practitioners' perspectives. *SA Journal of Human Resource Management, 23*. https://doi.org/10.4102/sajhrm.v23i0.2983

Coker, M. O., Kotila, O. A., Adigun, T. O., Lawal, T. v., Fagbamigbe, A. F., Makolo, A. U., Salawu, M. M., Ajayi, D. T., Oluniyi, P. E., Oguzie, J. U., Happi, A. N., Hoen, A. G., Babalola, C. P., & Ajayi, I. O. (2023). Data science training needs in sub-Saharan Africa: Implications for biomedical research and therapeutics capacity. *Open Research Africa, 6*, 21. https://doi.org/10.12688/openresafrica.14109.1

Research Africa, 6, 21. https://doi.org/10.12688/openresafrica.14109.1

Conijn, R., Snijders, C., Kleingeld, A., & Matzat, U. (2017). Predicting student performance from LMS data: A comparison of 17 blended courses using moodle LMS. *IEEE Transactions on Learning Technologies, 10*(1), 17–29. https://doi.org/10.1109/TLT.2016.2616312

Duta, I. P. G. P., Rio, Febriansyah, M. R., & Anggreainy, M. S. (2021). Effectiveness of LMS in Online Learning by Analyzing Its Usability and Features. *Proceedings of 2021 1st International Conference on Computer Science and Artificial Intelligence, ICCSAI 2021*, 56–61. https://doi.org/10.1109/ICCSAI53272.2021.9609757

Etheredge, C. L., & Waliczek, T. M. (2021). Comparative Analysis of Generation Z Era Students' Overall Grades and Course Satisfaction of a Basic Floral Design Course Taught Fully Face-to-face versus an Online Hybrid Format. *HortTechnology, 31*(6), 709–714. https://doi.org/10.21273/HORTTECH04900-21

Hicks, S. C., & Irizarry, R. A. (2018). A Guide to Teaching Data Science. *American Statistician, 72*(4), 382–391. https://doi.org/10.1080/00031305.2017.1356747

Holdgraf, C., Rokem, A., Ushizima, D. M., Culich, A., Deniz, F., Alegro, M., & Ushizima, D. (2017). *Portable Learning Environments for Hands-On Computational Instruction*. ArXiv. https://doi.org/10.48550/arXiv.1703.04900.

Huda, M., Maseleno, A., Atmotiyoso, P., Siregar, M., Ahmad, R., Jasmi, K. A., Muhamad, N. H. N., Mustari, M. I., & Basiron, B. (2018). Big data emerging technology: Insights into innovative environment for online learning resources. *International Journal of Emerging Technologies in Learning, 13*(1), 23–36. https://doi.org/10.3991/ijet.v13i01.6990.

Ifenthaler, D., & Gibson, D. (Eds.). (2021). *Adoption of data analytics in higher education learning and teaching advances in analytics for learning and teaching*.

Kim, B., & Henke, G. (2021). Easy-to-Use Cloud Computing for Teaching Data Science. *Journal of Statistics and Data Science Education, 29*(S1), S103–S111. https://doi.org/10.1080/10691898.2020.1860726.

Kiosia, A., Boylan, S., Retford, M., Marques, L. P., Bueno, F. T. C., Kirima, C., Islam, M. S., Naheed, A., & Wozencraft, A. (2024). Current data science capacity building initiatives for health researchers in LMICs: global & regional efforts. *Frontiers in Public Health, 12*. https://doi.org/10.3389/fpubh.2024.1418382.

Liu, X., Golen, E., & Raj, R. (2020.). *Developing a Hands-on Data Science Curriculum for Non-Computing Majors*.

Martínez-Cerdá, J. F., Torrent-Sellens, J., & González-González, I. (2018). Promoting collaborative skills in online university: comparing effects of games, mixed

reality, social media, and other tools for ICT-supported pedagogical practices. *Behaviour and Information Technology, 37*(10–11), 1055–1071. https://doi.org/10.1080/0144929X.2018.1476919

McGee, P., Windes, D., & Torres, M. (2017). Experienced online instructors: beliefs and preferred support regarding online teaching. *Journal of Computing in Higher Education, 29*(2), 331–352. https://doi.org/10.1007/s12528-017-9140-6.

Nungu, L., Mukama, E., & Nsabayezu, E. (2023). Online collaborative learning and cognitive presence in mathematics and science education. Case study of university of Rwanda, college of education. *Education and Information Technologies, 28*(9), 10865–10884. https://doi.org/10.1007/s10639-023-11607-w.

Said, B. A. (2021). *Assessment of advanced digital skills gap in Kenya using the design reality gap research framework*. University of Nairobi.

Saltz, J., & Heckman, R. (2015). Big Data Science Education: A case study of a project-focused introductory course. *Themes in Science & Technology Education, 8*(2).

Vittorini, P., Menini, S., & Tonelli, S. (2021). An AI-Based System for Formative and Summative Assessment in Data Science Courses. *International Journal of Artificial Intelligence in Education, 31*(2), 159–185. https://doi.org/10.1007/s40593-020-00230-2.

Waithira, N., Mutinda, B., Shah, K., Kestelyn, E., Bull, S., Boggs, L., Lang, T., & Cheah, P. Y. (2024). Addressing the gap in health data management skills: an online self-guided course for researchers and health professionals. *BMC Medical Education, 24*(1). https://doi.org/10.1186/s12909-024-06405-y.

Archibald, M. M., & Gerber, N. (2018). Arts and mixed methods research: An innovative methodological merger. American Behavioral Scientist, 62(7), 956–977. https://doi.org/10.1177/0002764218772672

Coker, M. O., Kotila, O. A., Adigun, T. O., Lawal, T. V., Fagbamigbe, A. F., Makolo, A. U., Salawu, M. M., Ajayi, D. T., Oluniyi, P. E., Oguzie, J. U., Happi, A. N., Hoen, A. G., Babalola, C. P., & Ajayi, I. O. (2023). Data science training needs in sub-Saharan Africa: Implications for biomedical research and therapeutics capacity. Open Research Africa, 6(21). https://doi.org/10.12688/openresafrica.14109.1

Labou, S., Yoo, H. J., Minor, D., & Altintas, I. (2019). Sharing and archiving data science course projects to support pedagogy for future cohorts. 2019 15th International Conference on eScience (eScience), 644–645. https://doi.org/10.1109/eScience.2019.00115

Strumińska-Kutra, M., & Koładkiewicz, I. (2018). Case Study. In M. Ciesielska & D. Jemielniak (Eds.), Qualitative Methodologies in Organization Studies (pp. 84–86). Palgrave Macmillan.

Waithira, N. J., Mungah, G. A., Kananura, C., Nzioki, E., Bwana, M. J., Tukei, S. T., Kyobutungi, C., & Yeka, A. (2024). Addressing the gap in health data science training for sub-Saharan Africa: development and evaluation of a tailored essential research skills curriculum. BMC Medical Education, 24(1), 1397. https://doi.org/10.1186/s13063-024-08493-9