Predicting Skill Demand Shifts Over time: Exploring Demographic Changes and Future Workforce Trends using Machine Learning

By

Jane Jepkemboi

Student No - 150961

A Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Science in Data Science and Analytics at Strathmore University

Institute of Mathematical Sciences

Strathmore University

Nairobi, Kenya

June, 2025

Declaration and Approval

Declaration

I hereby declare that neither this university nor any other university has previously accepted my work as a submission for a degree. To the best of my knowledge and belief, the dissertation does not contain any previously published or written works by other authors, with the exception of the references properly cited within the dissertation.

Name: Jane Jepkemboi

Signature  Date: 19/05/2025

Approval

The dissertation of Jane Jepkemboi was reviewed and approved for examination by the following:

Dr John Olukuru,

Lecturer

Institute of Mathematical Science,

Strathmore University

Dr. Godfrey Achono Madigu ,

Dean, Institute of Mathematical Science,

Strathmore University

Prof. Bernard Shibwabo,

Director of Graduate Studies,

Strathmore University

Abstract

Understanding the differences in skill requirements between age groups is crucial for strategic workforce planning and education. This study examines how demographic trends and machine learning can predict shifts in skill demand across generations. Focusing on Kenya, from 2020 to 2024, it analyzes job postings, industry trends, job functions, and required skills This study fills a gap in African research by using advanced machine learning to analyze generational skill demands, a perspective often overlooked in previous studies. This is significant as it fills a gap in existing research and provides critical insights for the African job market. In this study, we used KMeans clustering, hierarchical clustering, and DBSCAN to predict future patterns in skill demand by examining past job market data and demographic trends. We used the silhouette score and the elbow method to assess the models' performance. KMeans silhouette score = 0.49, Agglomerative Clustering Silhouette Score = 0.44, and Silhouette Score for DBSCAN = -0.24. A KMeans model of 4 centroids, as indicated by the elbow curve, was used to predict skill demand shifts. The model revealed changes in skills over time, with visible clusters showing previous and current market requirements. There is a notable demand for digital skills, data-related skills, and accounting, with remote skills being highly sought after. This study has several potential applications. Educational institutions can tailor curricula to future skill demands, while employers align workforce plans for upcoming transitions. Policymakers and industry leaders can address skill gaps, fostering a resilient job market. This study provides a data-driven approach to workforce dynamics, promoting adaptability in the labor ecosystem. In an increasingly dynamic job landscape, combining demographic insights with machine learning developments holds significant potential for informed decision-making.

Keywords: skill demand shifts, demographic changes, generational trends, machine learning, predictive analytics, workforce planning, education alignment, job market dynamics.

Table of Contents

List of Figures

Chapter One

1.1 Background

Technological improvements, demographic shifts, and changing socioeconomic dynamics are all contributing to a paradigm shift in the modern worker (Draskovic, D., Drinic, A., Kylymnyk, I., Seyidzade, L., Tilavov, M., & Tuna, G. (2021). The needs for skills, abilities, and knowledge are always changing as generations enter and leave the profession. In order to inform workforce planning and education initiatives, a thorough investigation is required.

The gap between the speed of change and the immobility of our institutions is clearly evident. Throughout history, education has always lagged behind technical advancement (Schleicher, 2015). Currently, educators acknowledge a "40-year gap" between experts speculating on where the workplace and educational landscape will need to be in 15 years, on-the-ground workers, and parents, whose notions of what makes "good" education are formed by their own earlier experiences. The ultimate result is a structure that resembles sedimentary rock since each layer is based on its own set of assumptions. However, there is little that holds the layers together, and once in place, they can limit future possibilities and policy changes.

According to the 2022 analysis of labor market imbalances in the EU27, Norway, and Switzerland, based on data provided by EURES National Coordination Offices, 29 nations—primarily from Southern, Eastern, Western, and Northern Europe—are experiencing labor shortages, while 24 have labor surpluses (EURES National Coordination Offices, 2023). The report highlighted widespread shortages were related to software, healthcare and construction and engineering craft occupations. The report also highlighted that in most cases, the imbalances in the labor market were a result of differences in the level of economic development within the education institutions in the different regions.

Kenya Ministry of Labor, in 2023, during their Skill Up Africa Expo, themed “Skilling Teachers, Trainers, and Youth for a Transformative Future,” emphasized the wealth in the labor market that has been produced over time. This event highlighted the critical role of skill development in creating wealth within the labor market. By focusing on equipping teachers, trainers, and the youth with necessary skills, the Expo emphasized the importance of preparing the workforce for future challenges and opportunities (Kenya Ministry of Labor, 2023).

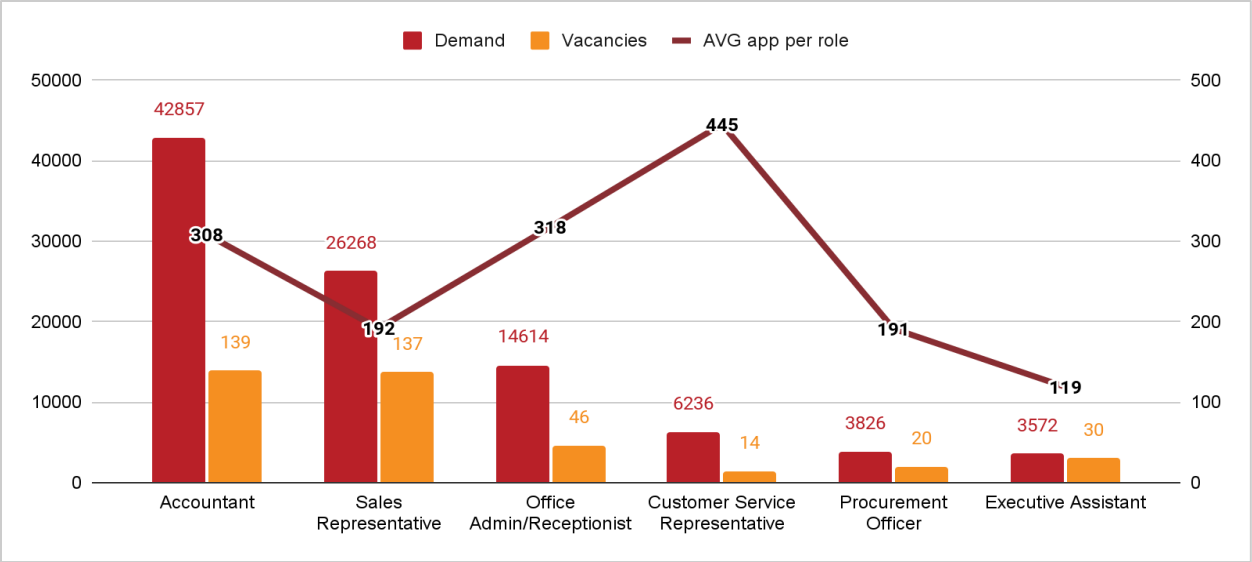


Figure 1.1.1: Top 6 Job Functions by Demand, Vacancies, and Average Applications

From figure 1.1.1 we can see an evident gap in Kenya's labor market, BrighterMonday Kenya (October 2023). This highlights a mismatch between job demand, vacancies, and applicant supply, revealing critical gaps in the labor market. High-demand roles like accounting and sales have significantly fewer vacancies, leading to intense competition, while oversaturated fields like customer service and administration see far more applicants per role than available jobs. Procurement and executive assistant roles have fewer applicants, indicating potential opportunities in less competitive fields. This misalignment suggests that education and training programs may not be fully aligned with market needs. Employers, policymakers, and job seekers must adapt to these insights to bridge skill gaps and create a more balanced workforce.

1.2 Problem Statement

Like other markets, the labor market is being impacted by computerization (Carl Benedikt Frey, Michael A. Osborne, 2017), globalization and outsourcing (Gereffi, G. (2018)), aging population (Bloom, D. E., Canning, D., & Fink, G. (2011), among other structural changes, which is upsetting the equilibrium between supply and demand for skills. Misalignment is common throughout the business cycle, but if persistent mismatches go unaddressed, the costs can be high: This also prevents businesses from innovating, adopting new technologies and relocating labor from less productive to more productive activities.

The continual imbalance between skill supply and employment demand creates major economic, social, and technological issues. Labour shortages in high-demand industries make it difficult for businesses to innovate, embrace automation, and expand, while oversupply in certain professions causes underemployment and pay stagnation (Autor, 2019). The rapid pace of technology innovation, including AI and automation, is further displacing workers whose skills are becoming obsolete, yet education and training institutions frequently fail to adapt in time to meet changing industrial needs (World Economic Forum, 2020). According to studies, automation has the potential to displace up to 30% of occupations by 2030, disproportionately harming low-skilled individuals and exacerbating economic inequities. Furthermore, globalization and outsourcing have transformed workforce demand, necessitating ongoing skill development to remain competitive (Gereffi, 2018). This imbalance not only impedes productivity and economic progress, but it also contributes to income inequality and workforce inefficiencies. Addressing these gaps through targeted policies, industry engagement, and adaptive education systems is crucial to developing a resilient, future-ready workforce capable of sustaining economic growth despite continuous structural changes.

1.3 Research Aim

The aim of this study was to develop and deploy a predictive model that was used to predict the skill demand shifts and uncovering the demographic factors affecting current job market

1.4 Research Question

a. How have skill demands evolved across different generations in the workforce over time?

b. What are the key demographic attributes that contribute to skill demand shifts among generational cohorts?

c. To what extent can machine learning models effectively predict future skill demand trends based on historical data and demographic changes?

1.5 Research Objective

a. Identify changes in skill demand across different generations and categories in the workforce overtime using machine learning and descriptive analysis.

b. Identify key demographic attributes that contribute to skill demand shifts among generational cohorts using machine learning for exploratory data analysis.

c. To evaluate the effectiveness of machine learning models in predicting future skill demand trends based on historical data and demographic changes.

1.6 Research Scope and Limitations

The scope of this dissertation encompasses an examination of the evolving skill demands within the workforce across various generations, focusing primarily on Africa, specifically Kenya. The temporal range under investigation spanned from 2020 to 2024 and included an analysis of job postings, industry trends, job functions, and the skills required within these regions. The research faced several limitations. Firstly, while certain skills such as Python programming, statistical machine learning, and applied machine learning remained consistently popular across various regions, the analysis might not have fully captured the nuances of skill demand variations within specific contexts. Additionally, the widespread use of specific tools such as Microsoft Excel, Tableau, and JavaScript might have influenced skill demand but could have been challenging to comprehensively account for. Furthermore, external factors such as political events, pandemics, and global factors like war were unpredictable and could have significantly impacted both the demand for and supply of skills within the workforce, potentially affecting the accuracy and generalizability of the findings. Moreover, the available data primarily focused on white-collar jobs and positions advertised online, potentially limiting the scope of the analysis and overlooking skill demand trends in other sectors or regions not covered by online job postings.

1.7 Significance of the Study

The results of this study have significance for a variety of parties, including businesses, policymakers, educators, and job seekers. Organizations can match existing talent strategies with anticipated future demands by using accurate projections of skill demand shifts to guide proactive workforce planning. In order to ensure that graduates have the skills required by the labor market, educational institutions can customize their curricula, improving graduates' employability. The

information can be used by policymakers to address any skill gaps and guarantee that all generations have equal access to opportunities.

Chapter 2: Literature Review

2.1 Overview

This chapter gives a thorough literature analysis on the topic of forecasting jobs and abilities. It provides a thorough examination of pertinent academic papers, theories, techniques, and empirical findings that advance our knowledge of how jobs and skills are expected to correspond with changing workforce patterns.

2.2 Theoretical Review

2.2.1 Shifting skills Demand

This section examines the historical development of the labor market, charting its journey from traditional labor markets to the modern, technologically advanced landscape. The key forces behind this transformation are mentioned as being globalization, automation, digitization, and changes in consumer preferences (Smith, 2018; Autor, 2015). This section creates a framework for comprehending the complex dynamics of skill demand anticipation by looking at past trends and patterns.

The literature on the shifting skills needed is relevant to this paper. According to conventional thinking, these changes are a result of how well high-skilled labor and technology complement one another. In other words, as technology advances, the demand for skills increases, and as a result, so does investment in skills. The distribution of wages and employment across advanced economies has seen significant changes over time, and this paradigm has shown to be a workhorse for economists in its ability to explain these shifts. (Goldin and Katz, 2009)

A growing body of work emphasizes the importance of "noncognitive" skills, such as social and leadership abilities. This stems from Heckman's (1995) key insight that labor market outcomes, such as earnings, are likely shaped by a variety of skills insofar as measured cognitive ability accounts for only a small portion of the variation in such outcomes.

Psychologists and neuroscientists have also supported the validity of social intelligence tests (Poropat, 2009; Woolley et al., 2010). In fact, the distinction between cognitive and non-cognitive skills has been rejected by more contemporary thought. Some contend that reason is considerably more diverse and opportunistic than is often believed and that it has evolved solely to aid humans in justifying their actions and influencing others, which is essential for communication and cooperation. When seen as adaptive reactions to the conundrums of social interaction, thought patterns that look "irrational" from a purely cognitive standpoint become beneficial (Mercier and Sperber, 2017).

2.2.2 Demographic and Socioeconomic Influences

This section looks at how socioeconomic and demographic trends affect trends in skill demand. It is investigated how shifts in population demographics, such as generational shifts and workforce diversity, affect skill demand (Fry & Taylor, 2012; Fry, 2018). In connection to skill demand patterns, the impact of socioeconomic factors such as educational attainment, income inequality, and technology adoption rates is examined (Katz & Goldin, 2008; Brynjolfsson & McAfee, 2014).

A comprehensive understanding of the dynamics of skill demand is obtained by examining how these elements interact with skill anticipation.

A European Institute of Public Administration article published in 2012 by Danielle Bossaert in cooperation with Christoph Demmke and Timo Moilanen research emphasizes the imminent challenges faced by European public sectors in light of a diminishing and aging workforce, compounded by the baby-boom generation's retirement. A reevaluation of human resource management (HRM) policies is crucial to ensure the continued health, productivity, and motivation of an aging workforce. While raising the retirement age is important, relying solely on legislative measures is inadequate; a comprehensive approach is essential. This approach involves implementing age-conscious personnel policies, providing training, combating age discrimination, promoting health initiatives, enhancing job satisfaction, managing careers, and improving working conditions. The study emphasizes that maintaining employability in an aging workforce necessitates ongoing training and preventing the obsolescence of knowledge and skills. The findings underscore the importance of tailoring training methods to accommodate generational differences. Regarding career management, the study recommends establishing flexible career paths and aligning compensation with competencies to avoid unhealthy competition with younger colleagues. Ultimately, the research underscores the importance of a proactive and comprehensive approach to health and safety for sustained workability. Case studies from Finland, France, and Germany demonstrate the practical implementation of age-sensitive initiatives, showcasing the benefits of holistic and proactive HRM strategies in addressing the challenges posed by an aging workforce in European public sectors.

2.3 Empirical Review

2.3.1 Skill Demand Analysis and Prediction

This section explores several strategies and procedures used to analyze and forecast skill demand. It includes common methods including trend extrapolation, expert surveys, and labor market analysis (Estrin et al., 2016; Smith & Kalleberg, 2019). It also examines the advent of cutting-edge techniques that have transformed skill demand forecasting, including machine learning, data mining, and predictive modeling (Nguyen et al., 2020; Acemoglu & Autor, 2012). Examining the benefits, drawbacks, and uses of various techniques sheds light on how skill demand analysis has developed through time.

A more forward-thinking strategy can aid in avoiding these traps. This is demonstrated by Frey and Osborne (2017), who evaluate the viability of automating current occupations under the presumption that new technologies are adopted on a greater scale across industries. In this study, machine learning experts manually classified a sample of jobs as either strictly automatable or not automatable. They utilized a standardized set of nine O*NET features of occupations, which assess three key bottlenecks to automation: perception and manipulation, creative intelligence, and social intelligence. Employing a classifier algorithm, they generated a "probability of computerization" for all jobs. Their analysis revealed that 47% of jobs held by US workers face a high risk of automation within the next two decades.

2.3.2 Machine Learning in Labor Market Analysis:

Nikola Dawson's 2020 study, conducted in the context of the Australian labor market, applied a machine learning framework to predict temporal skill shortages across 132 standardized occupations using data spanning 2012 to 2018. The study utilized the Boost classifier to forecast annual shortage categories and found that features derived from job advertisements and employment statistics were particularly effective in capturing year-to-year fluctuations in skill demand. While Dawson's findings reinforce the value of digital labor market signals in anticipating workforce needs, the study also reveals key limitations that affect its generalizability. Critically, Dawson's model relied on assumptions to compensate for sparse or non-representative occupation-level data, a challenge common in labor market analytics. However, the study did not fully interrogate the implications of these assumptions for model reliability or fairness — particularly whether prediction accuracy varied across sectors or job types. Moreover, while the use of Boost added predictive power, the study did not benchmark its performance against simpler or alternative models, limiting transparency around model interpretability and robustness.

In contrast to Dawson's approach, the current study incorporates unsupervised clustering methods to explore latent skill groupings rather than forecasting predefined shortage categories. This shift reflects a methodological departure that favors structural pattern discovery over predictive classification. Additionally, by comparing multiple clustering algorithms (KMeans, DBSCAN, hierarchical), this study places greater emphasis on evaluating model fit, internal consistency, and the nuanced structure of job-skill relationships — areas Dawson's work did not address.

Ultimately, while Dawson (2020) offers a valuable foundation for skill forecasting using machine learning, its lack of methodological triangulation and limited treatment of data bias constrain the broader applicability of its conclusions.

LUNA-ROMERA, et al., draw attention to Spain's status as having the highest unemployment rate in Europe. Leveraging this observation, the researchers take the initiative to utilize unsupervised machine learning techniques on publicly available data sourced from the Spanish Ministry of Employment, Migration, and Social Security. The analysis employs clustering, utilizing two different algorithms: the average linkage algorithm within Stata and the K-means algorithm. A subsequent comparative assessment of the outcomes is undertaken to evaluate the precision and efficacy of each algorithm. The model results suggested that changes in the Spanish labor market had an impact on the physiognomy of some jobs. The analysis shows how the labor market has changed: the crisis has increased geographic mobility and encouraged the appearance of new employment opportunities. Some clusters, however, continue to exist and are stable in spite of the crisis.

2.3.3 Skill Mismatch and Labor Market Inefficiencies

Edwin Leuven conducted an extensive examination of the economics literature regarding over schooling and labor market mismatch in his research. This investigation was prompted by the increasing number of college graduates in the US during the early 1970s, alongside a decline in the returns to college education, as highlighted in Freeman's publication "The Overeducated American." Despite subsequent evidence that should have alleviated these concerns, Duncan and Hoffman's influential paper introduced a new subfield within economics, focusing on overeducation. This body of literature primarily employed an extended version of Mincer's wage equation, which dissected actual years of schooling into over schooling, required schooling, and under schooling. However, the estimates derived from this specification were deemed less informative for both individual and aggregate schooling investment decisions. The efficiency

implications of skill mismatch remained largely unexplored, hindered by challenges in separating mismatch impact from unobserved ability, measurement errors, and omitted variable bias. Originating from Duncan and Hoffman's work, the micro over schooling literature lacked substantive hypotheses and relied on loose interpretations within existing theories. The overall assessment was pessimistic, highlighting the literature's limitations in contributing to a nuanced understanding of past labor markets and education investment. Despite numerous estimates, issues such as omitted variable bias and measurement errors were not adequately addressed, rendering reported estimates non-causal. The chapter urged for forthcoming contributions to meticulously address these issues and seamlessly incorporate the over schooling/mismatch literature into the broader framework of contemporary labor economics and the economics of education.

2.4 Research Gap

There is a significant gap in applying unsupervised and semi-supervised learning techniques to predict skill demand in Kenya and Africa at large, with most existing research focusing on developed countries. Exploring Kenya's labor market through machine learning is crucial for aligning educational curricula with industry needs, addressing skills mismatches, and bridging the gap between labor supply and demand. This research is vital for adapting to global trends and preparing a workforce that can meet emerging challenges.

Chapter 3: Research Methodology

3.1 Research Design

The research utilizes the Cross-Industry Standard Process for Data Mining (CRISP-DM) model, a widely recognized methodology used by data mining experts. CRISP-DM offers a structured and systematic approach to addressing various data mining tasks, enabling a comprehensive understanding of the data. It consists of six major phases. Business Understanding is the first phase, it involves defining the project's objectives and requirements from a business standpoint. It is the most critical as it sets foundation for the entire data mining process. Once the business objectives are defined, Data Understanding is the second phase, key focus is on acquiring and exploring the data. This includes collecting relevant datasets, understanding the characteristics of the data, its

relevance to objective and identifying any potential issues or inconsistencies. The insights gathered in this step are critical for informing the subsequent Data Preparation and Modeling processes.

The third phase involves Data Preparation for modeling, focusing on pre-processing and transforming the data to ensure its quality and suitability for modeling. This phase involves tasks such as data cleansing, data transformation, feature engineering, and data integration, addressing issues like missing, inconsistent, or duplicate data to create a clean and structured dataset. The quality of the data prepared in this stage directly influences the accuracy and efficiency of the subsequent modeling process. The fourth phase, Modeling, involves selecting and applying appropriate data mining techniques to solve the business problem. This stage involves selecting the appropriate modeling techniques, splitting the data into training and test sets for model validation, training the model on the training dataset, evaluating the model performance using evaluation metrics and fine tuning the model parameters to optimize performance.

After the models have been built, the evaluation phase focuses on assessing their effectiveness in addressing the original research objectives. Key tasks in this phase involve evaluating the models against predefined success criteria and using the relevant evaluation metrics. Model results are evaluated to ensure it aligns with research objective, this is also the decision stage for deployment or if further refinement is needed. The sixth phase, Deployment, involves integrating the model into the businesses processes and systems, this ensures that insights gained from data mining are used to support business decisions.

In summary, the CRISP-DM provides structured framework that guides the entire data mining process, ensuring that each phase is aligned with the projects objectives and that the results are actionable for business decision making.

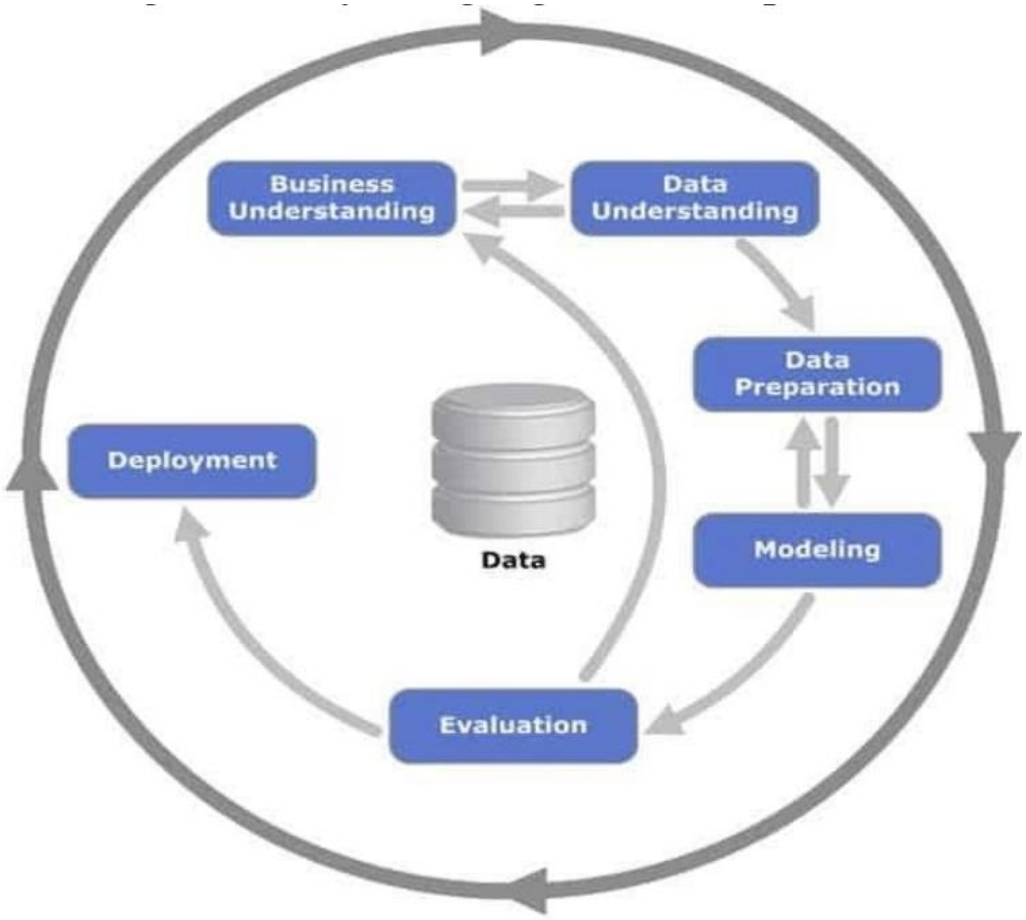


Figure 3.1: CRISP-DM Methodology (Saltz J. & Hotz N., 2018).

3.2 Data Collection

Data was collected from The African Talent Company database. The African Talent Company (TATC) is a collective of pan-African businesses dedicated to addressing the talent gap in Africa

through innovative, locally-developed solutions. They operate comprehensive career development and recruitment solution companies in Ghana and Nigeria under the Jobberman brand, and in Kenya and Uganda under the BrighterMonday brand. The data is broken down into three parts, part 1 of the data contains job posted information, i.e. job posted, time, sites posted, location etc., the second part of the data has skills required for the job posted and third has the job performance in terms of application, reach and qualification data. The first dataset with scraped jobs contains 183158 records and 27 variables. The second data has skills required for different jobs available, with 66 records and 21 variables. The data variables are explained in the table below

Table 3.2.1: Dataset 1 Variables

No.	Variable Name	Description
1	CompanyName	Advertising company
2	Industry	Business sector or field classification
3	Title	Position or role identifier in employment
4	Job Function	Role or responsibility in employment.
5	Joblocation	Workplace or job site location.
6	Monthposted	Month of job posting.
7	Year month	Combination of year and month.
8	Dateposted	Posting date of job openings.
9	Job Boards	Online platform that posted the job
10	DateApplyUntil	Application deadline for job openings.
11	Top Categories	Primary job classifications or sectors.
12	Salary	Compensation for employment.
13	Source1BackLink	Reference link for Source 1.

14	expired	No longer valid or active.
15	Joblevel	Position hierarchy within an organization.
16	joblevel.1	Secondary position hierarchy within an organization.
17	joblevel.2	Tertiary position hierarchy within an organization
18	Source3Name	Name of Source 3
19	Source3BackLink	Reference link for Source 3.
20	Source4Name	Name of Source 4.
21	Source4BackLink	Reference link for Source 4.
22	expired.1	No longer valid or active (additional information provided by ".1").
23	joblevel.3	Quaternary position hierarchy within an organization.
24	joblevel.4	Quinary position hierarchy within an organization
25`	Skill 1	Skills required for the role listed
26	Skill 2	Skills required for the role listed
27	Skill 3	Skills required for the role listed

Table 3.2.2: Dataset 2 Variables

No.	Variable name	Description
1	Job title	brief description that identifies the position or role
2	Skills 1 - skills 21	the abilities, knowledge, competencies, and expertise

		that individuals possess and can apply effectively to perform tasks

3.3 Data Processing

The process of preparing the dataset for analysis involved cleaning it to ensure accuracy and reliability of the results. This encompassed various procedures aimed at addressing different issues within the dataset. These procedures included handling missing data entries, which involved dealing with any empty or null values present. Additionally, outliers, which are data entries significantly different from other values in the dataset, were identified and managed to prevent skewing the analysis results. Furthermore, duplicate entries, or repeated data points, were identified and removed to avoid redundancy in the dataset. By employing these cleaning procedures, the dataset was refined and optimized for subsequent analysis, facilitating more accurate and meaningful insights.

3.3.1 Missing Data Imputation

The presence of missing data within a dataset can be categorized into three distinct scenarios, each posing unique challenges for analysis and interpretation. Firstly, when missing data occurs uniformly across all cases, it falls under the category of missing completely at random (MCAR). This implies that the reasons for missing data are unrelated to the observed variables, simplifying the analysis process despite the loss of information. For instance, our data scraper rebooting/updating could lead to missing data due to chance of occurrences. On the other hand, missing at random (MAR) describes a situation where the probability of data being missing depends on known properties within the dataset. For example, if the likelihood of missing data on a spool varies depending on the day/season it is placed on, it falls under the MAR category. Unlike MCAR, MAR encompasses a broader range of scenarios and serves as the foundation for modern missing data methodologies. Lastly, missing not at random (MNAR) occurs when the probability of data being missing is influenced by factors unknown to the analyst. This presents significant challenges as the missingness may distort the analysis results, leading to biased estimates. Strategies to address MNAR include gathering additional data to understand the reasons behind

missingness or conducting sensitivity analyses to assess the impact of various scenarios on the results. Understanding these distinctions is crucial for selecting appropriate methods to handle missing data and ensure the validity of statistical inferences drawn from the analysis.

3.3.1.1 Dropping Columns with Missing Values

To remove columns with missing values, we used missingno bar plot to inspect columns with all or over 90% missing values.

3.3.1.2 Dropping rows with Missing Values

Given the size of our dataset, dropping rows with missing values in specific columns i.e. Industry, Joblocation, dateposted, title, year-month, skill 1, skill 2, skill 3 was ideal. This is because imputing the missing values would introduce bias into the dataset given the values are not missing at random hence would distort the analysis results (Advised industry experts).

3.3.1.3 Forward filling missing values

This method was useful for some columns, i.e. Job boards. This is because the missing values occur sequentially over time (Consult and advise from the Data Engineers (TATC). To maintain the continuity of our data, we filled the missing data using the above-mentioned method. Resulting data shape (90936,11)

3.4 Machine Learning Model

3.4.1 Feature Transformation

In order to improve their suitability for a machine learning model, input features—also referred to as independent variables or predictors—are subjected to mathematical changes. Feature transformation can involve various techniques, such as scaling, normalization, encoding, and creating new features. In our data, encoding was implemented. Encoding involves converting categorical features into numerical representations that can be used by machine learning algorithms. Common encoding techniques include one-hot encoding, label encoding, and ordinal encoding. Label Encoding was implemented, this is assigning unique integer labels to categorical data, i.e. Industry, job function, skills 1, skills 2, skills 3, location.

3.4.2 Feature Scaling

The process of scaling is converting the numerical feature range to a standard range, usually between 0 and 1 or -1 and 1. This is crucial for algorithms that depend on the size of the features, such as distance-based algorithms (like K-means clustering) and gradient descent-based algorithms (like linear and logistic regression). Z-score scaling (Standardization) and Min-Max scaling are two popular scaling methods. StandardScaler from the sklearn. preprocessing module in scikit-learn was used to perform standardization on the dataset. The standard scaler standardizes features by removing the mean and scaling to unit variance. Standard score of a sample x is calculated as shown below.

$$z = \frac{(x - \mu)}{\sigma} \quad (1)$$

Where

μ - mean of training samples, 0 when sample mean is false

σ- standard deviation of the sample mean, 1 when sample std is false

x - original value

z - the standardized value

3.4.3 Trained Machine Learning Models

Next steps after data preprocessing and preparation was training unsupervised machine learning models from Sklearn. The models trained were

3.4.3.1 KMeans clustering

This is a technique used to group a set of observations in a d-dimensional real vector (x1, x2, ..., xn) into k clusters k (≤ n) sets. The goal is to minimize the within-cluster sum of squares (WCSS) (i.e. variance). Mathematically, the objective is to find a partition S = {S1, S2, ..., Sk} of the observations such that the total variance within each cluster is minimized.

$$arg_S min \sum_{i=1}^{k} \sum_{x \epsilon s_i} ||x - \mu_i||^2 = arg_S min \sum_{i=1}^{k} |S_i| Var\, S_i \quad (2)$$

μ_i - is the mean/ centroid of points in S_i

$$\mu_i = \frac{1}{|s_i|} \sum_{x \epsilon S_i} x, \quad (3)$$

The algorithm uses an iterative technique in assigning observations to a cluster. Centroids are obtained and observations are assigned to the nearest centroid, with the least Euclidean distance. Centroids/ means can be recalculated and the observations reassigned. The two steps are repeated until the assignments converge, i.e. observations assignments do not change.

3.4.3.2 Hierarchical clustering

This method seeks to build a hierarchy of clusters. It has two categories, agglomerative and divisive. Agglomerative approach was used. Each observation begins in its own cluster, and pairs of clusters merge as one moves up the hierarchy and centroid linkage.

$$||\mu_a - \mu_b||^2 \quad (4)$$

where μa and μb are the centroids of A resp. B.

3.4.3.3 Density-Based Spatial Clustering of Applications with Noise- DBSCAN

Density-based spatial clustering of applications with noise (DBSCAN) does not require the number of clusters to be specified in advance. Instead, it groups together closely packed points while marking points in low-density regions as outliers or noise.

The best performing model, KMeans clustering, was deployed on streamlit.

Model development process

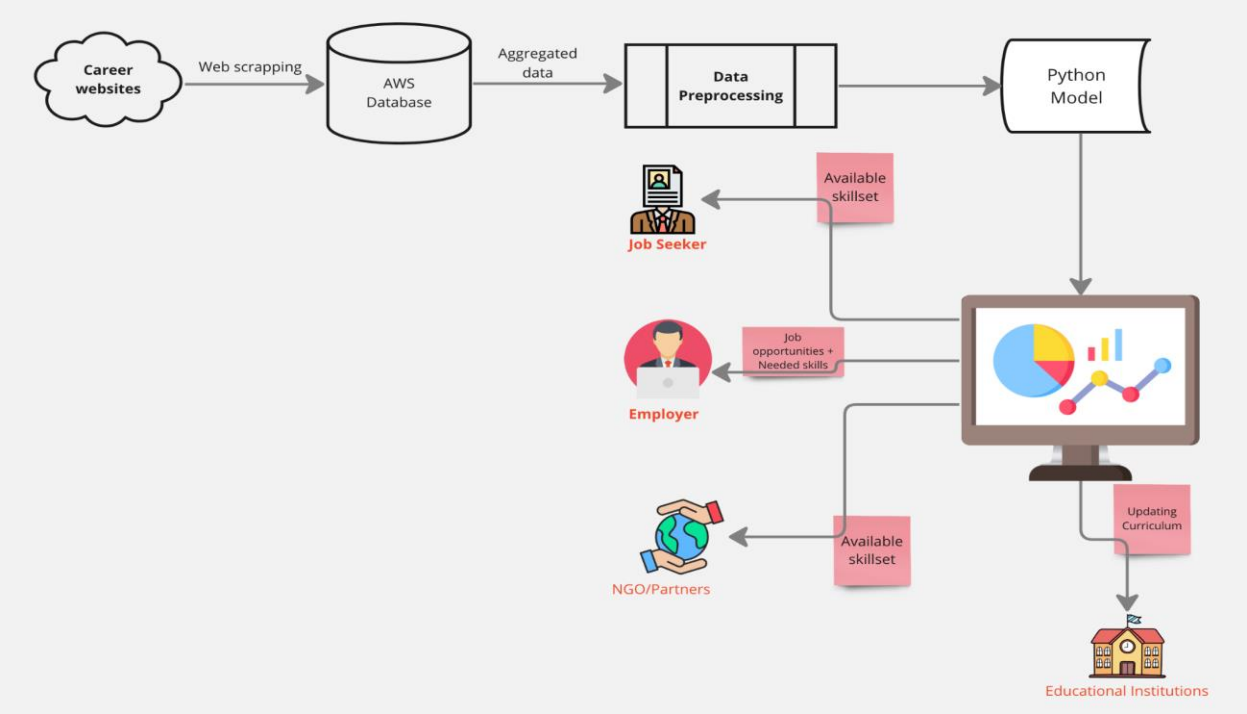


Figure 3.4: Project end to end workflow

3.4.3.4 Model Performance Evaluation Criteria

To evaluate the developed models, two metrics were used to evaluate the machine learning models, Silhouette score and the elbow curve for optimum number of clusters.

Elbow curve - visible bend at k = 4

3.5 Model Deployment

To integrate our machine learning model into real-world application, we used streamlit as a deployment tool, with focus on creating an interactive data product that allows users to easily interact with the model and visualize the results.

3.5.1 Deployment Process for Streamlit App

3.5.1.1 Model Preparation

After training the model and evaluating its performance, the best performing model, Kmeans clustering was saved in a .pkl file format. This model will be loaded by the streamlit app to make real time predictions

3.5.1.2 Development of Streamlit Application

A streamlit app was built, providing a platform that is interactive for users to input data and receive predictions from the trained model. The app includes sliders, text boxes and file upload options, enabling users to customize inputs and view model predictions. The user is able to explore the data using exploratory data section.

Chapter 4: Results

This chapter discusses the descriptive statistics, diagnostic tests, results from the models fitted, and the interpretation of the findings.

4.1 Model Optimization

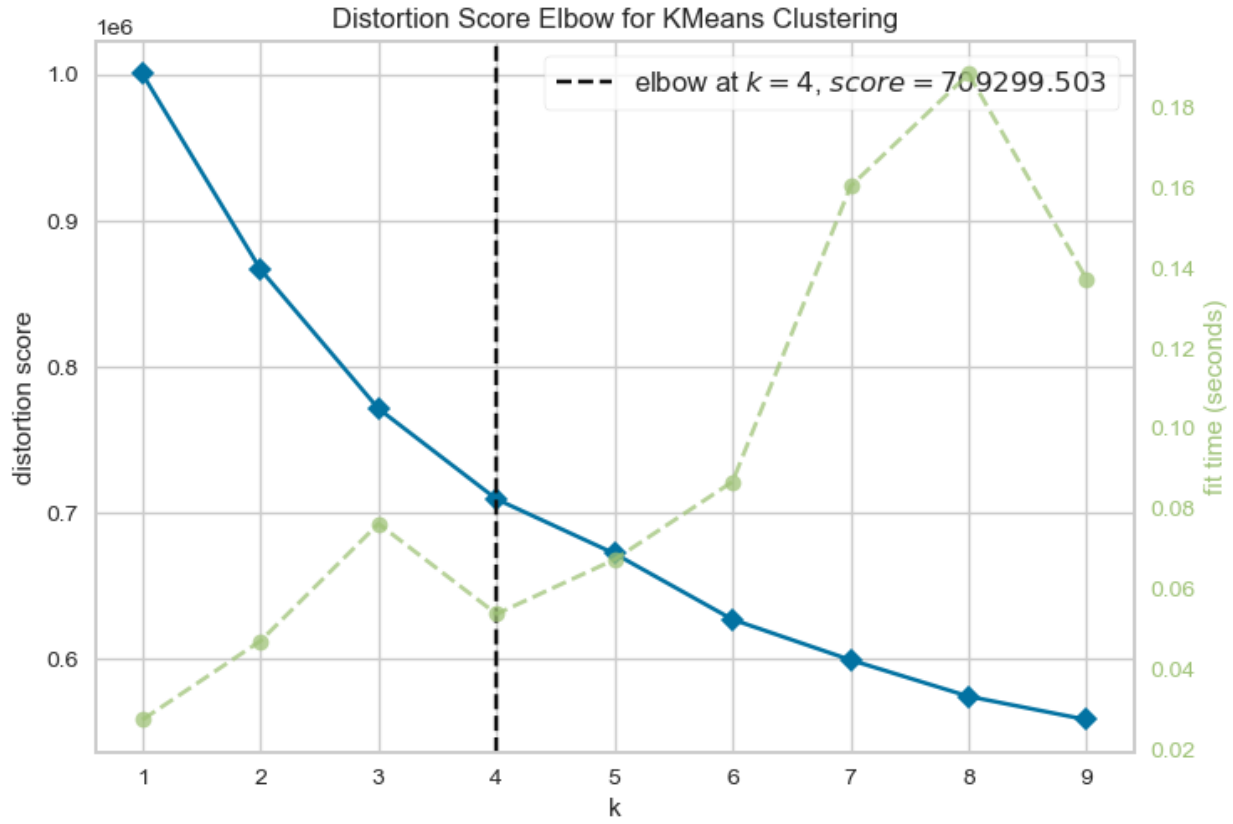


Figure 4.1: Distortion Score elbow for K Means Clustering

Silhouette Score

$$a(i) = \frac{1}{|C_1|-1} \sum_{j \epsilon C_1, i \neq j} \quad d(i,j) \qquad (5)$$

A score close to +1 indicates that data points are well-clustered and far from neighboring clusters.

A score close to 0 indicates overlapping clusters.

A score close to -1 indicates that data points may have been assigned to the wrong clusters.

Comparing silhouette scores between two models can help determine which model produces better-defined clusters. The model with a higher silhouette score is generally preferred as it indicates better separation between clusters and more cohesive clusters.

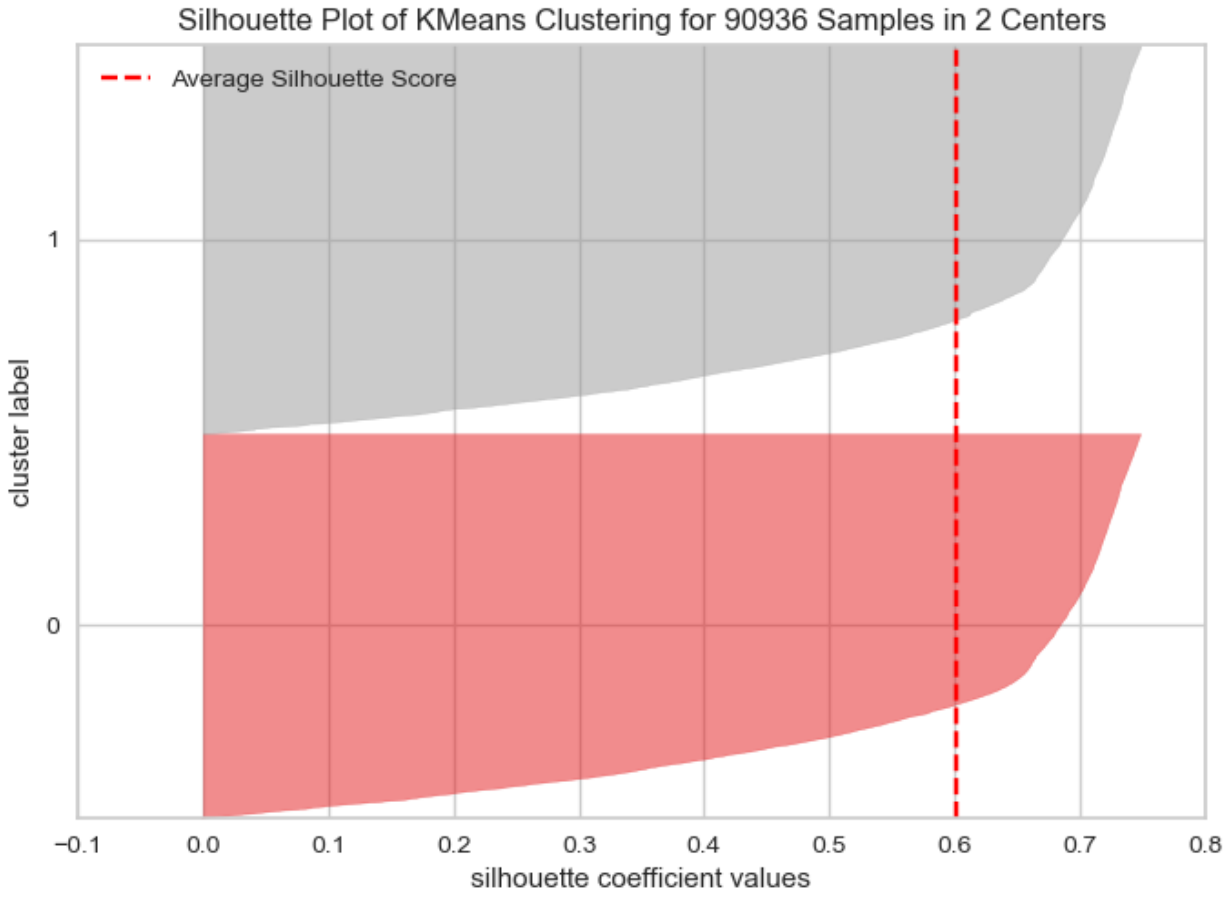


Figure 4.2: Silhouette plot of K Means clustering for 90936 samples in 2 centers.

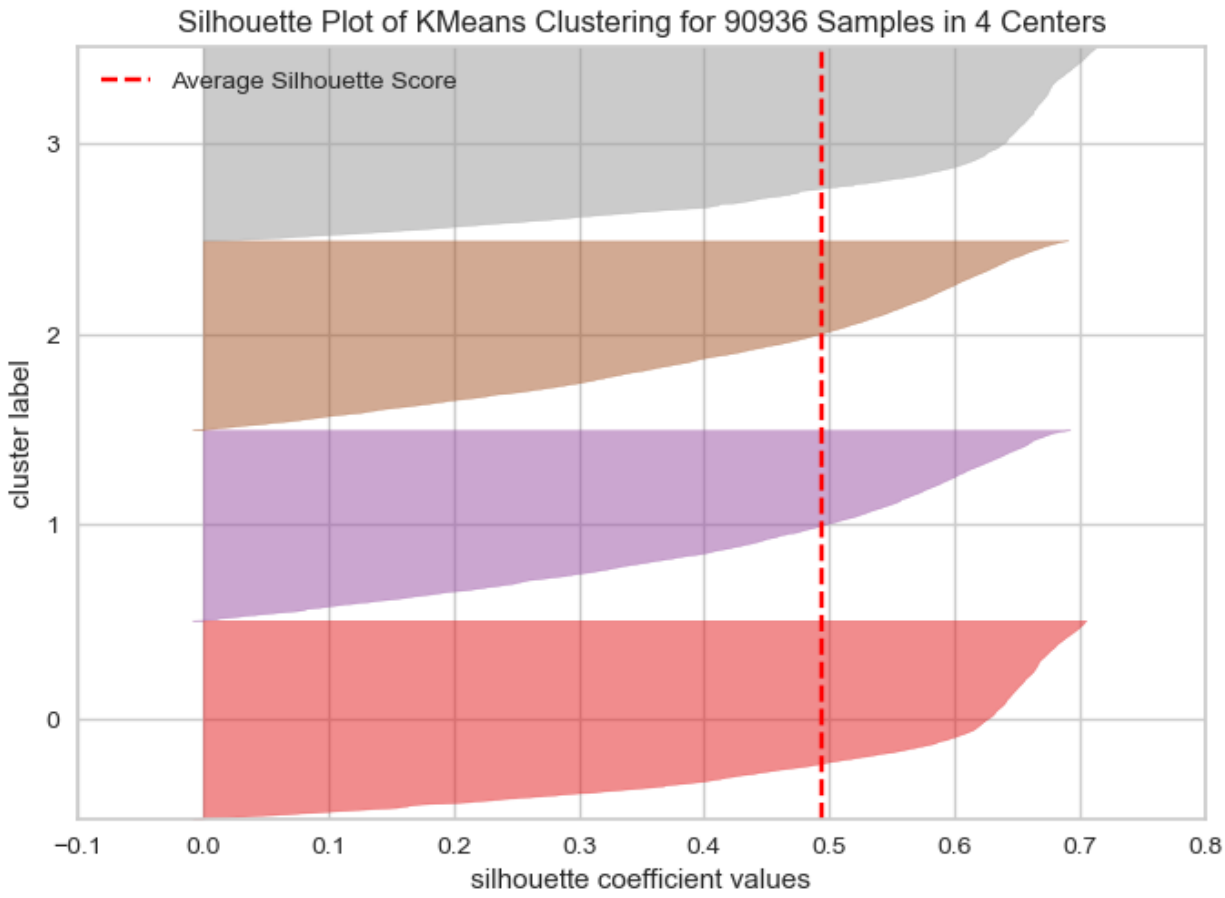


Figure 4.3: Silhouette Plot of K Means Clustering for 90936 samples in 4 Centers

Silhouette Score for KMeans: 0.49266616808147425

Agglomerative Clustering Silhouette Score: 0.4380507964220889

Silhouette Score for DBSCAN: -0.24234402976113067

4.2 Initial Data Exploration Results

From the chart below, we can see that there have been increased jobs in the markets over time. (missing data in March 2023 due to a change in scraping strategy). This shows a positive indicator of economic vitality, employment opportunities, and overall prosperity. It reflects a dynamic and growing economy that benefits both businesses and individuals.

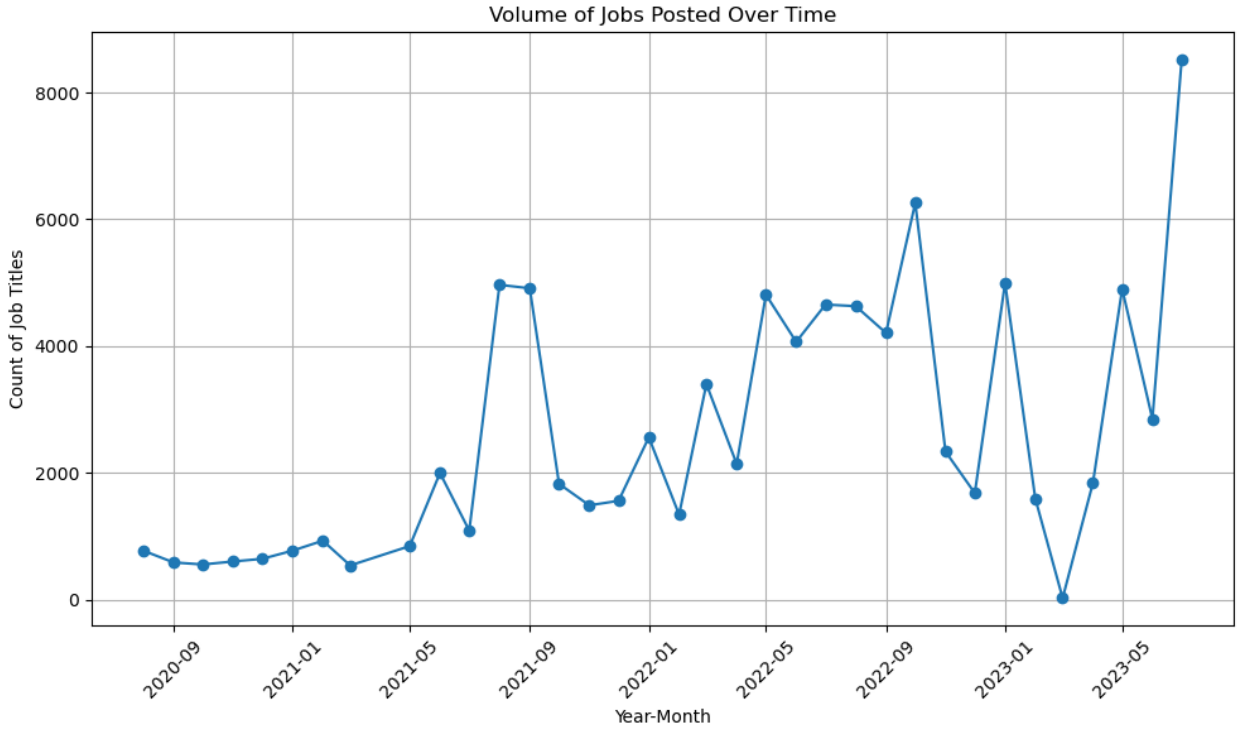


Figure 4.4: Volumes of Jobs Posted Over Time

Figure 4.4 illustrates a noticeable increase in the number of available jobs within the job market, accompanied by clear seasonal fluctuations. The data reveals a consistent upward trend, indicating a rising demand for workers and highlighting a growing need for a skilled workforce over time.

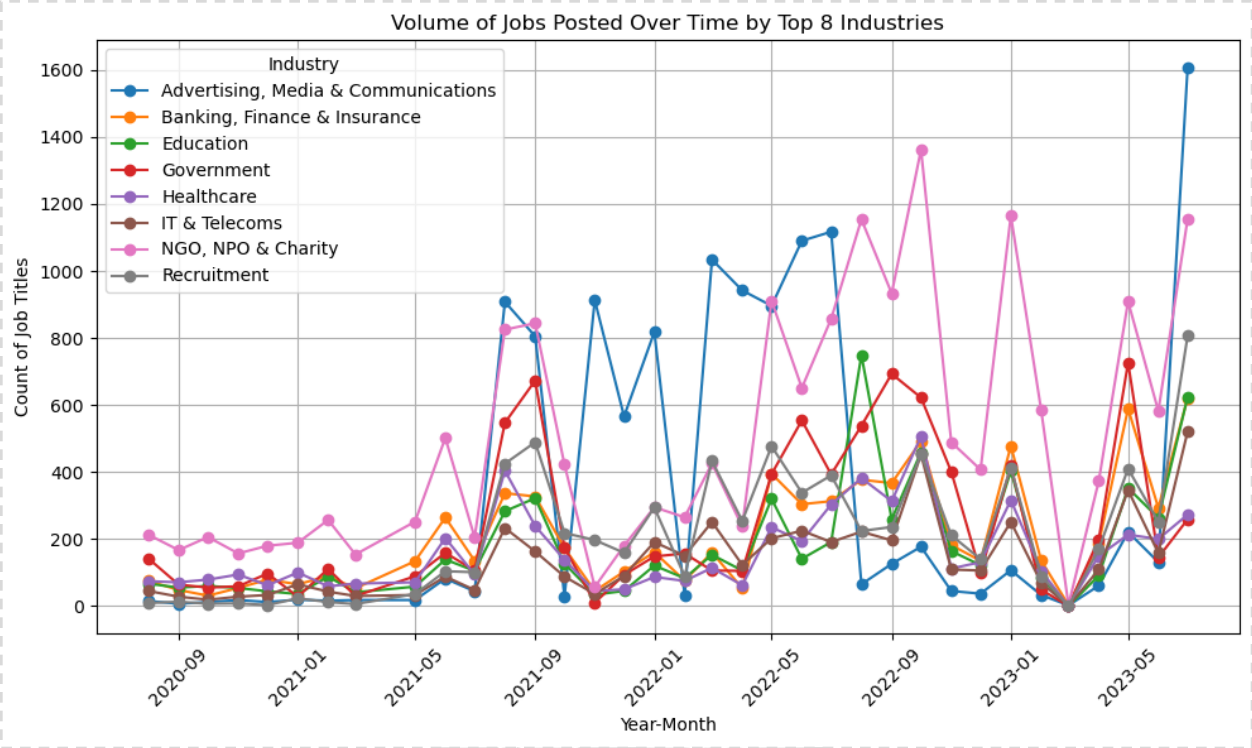


Figure 4.5: Volume of jobs posted over time by top 8 industries

Figure 4.5 shows the top industries that are hiring over time. Despite fluctuations in jobs in the market, there has been increased demand for different types of skills within, NGO, NPO & charity sector, Advertising, media and Communications and in the Government sector.

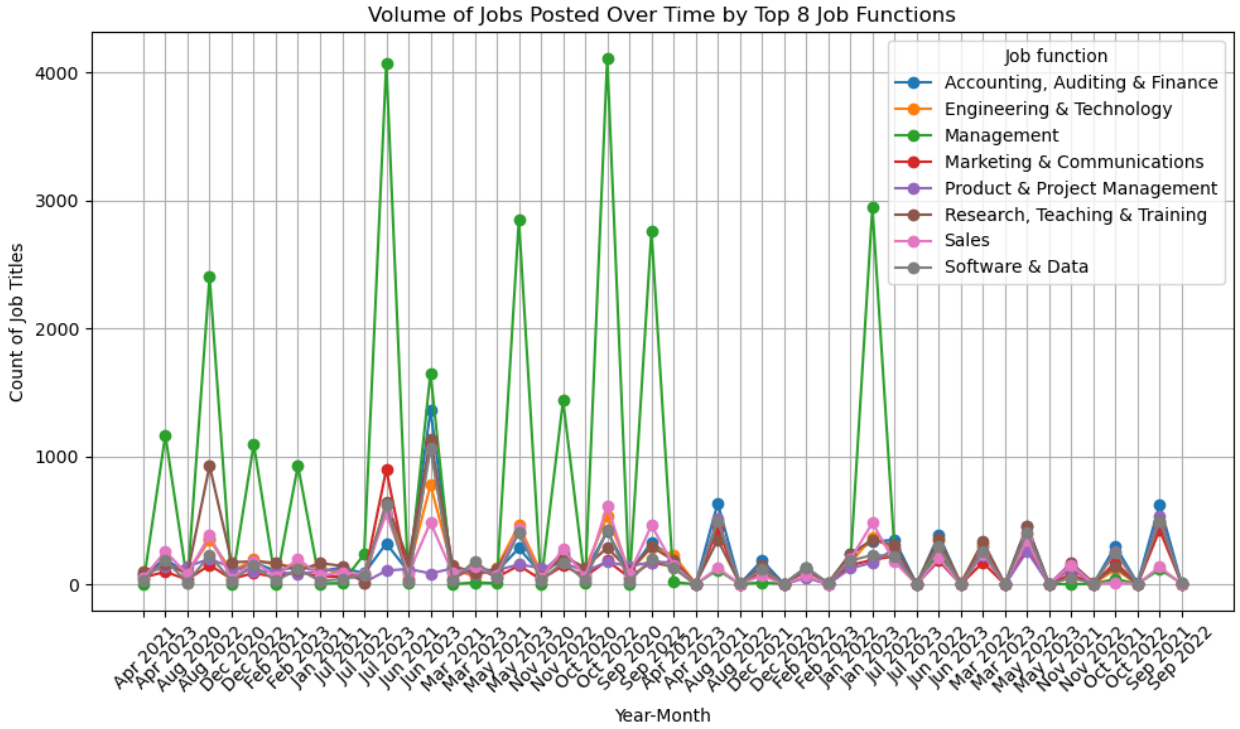


Figure 4.6: Volume of jobs posted over time by top 8 Job functions

Figure 4.6 shows respective functions that have demand for different types of skills, high demand seen in management functions, accounting roles, Engineering & technology, product and project management, sales and marketing and software and data.

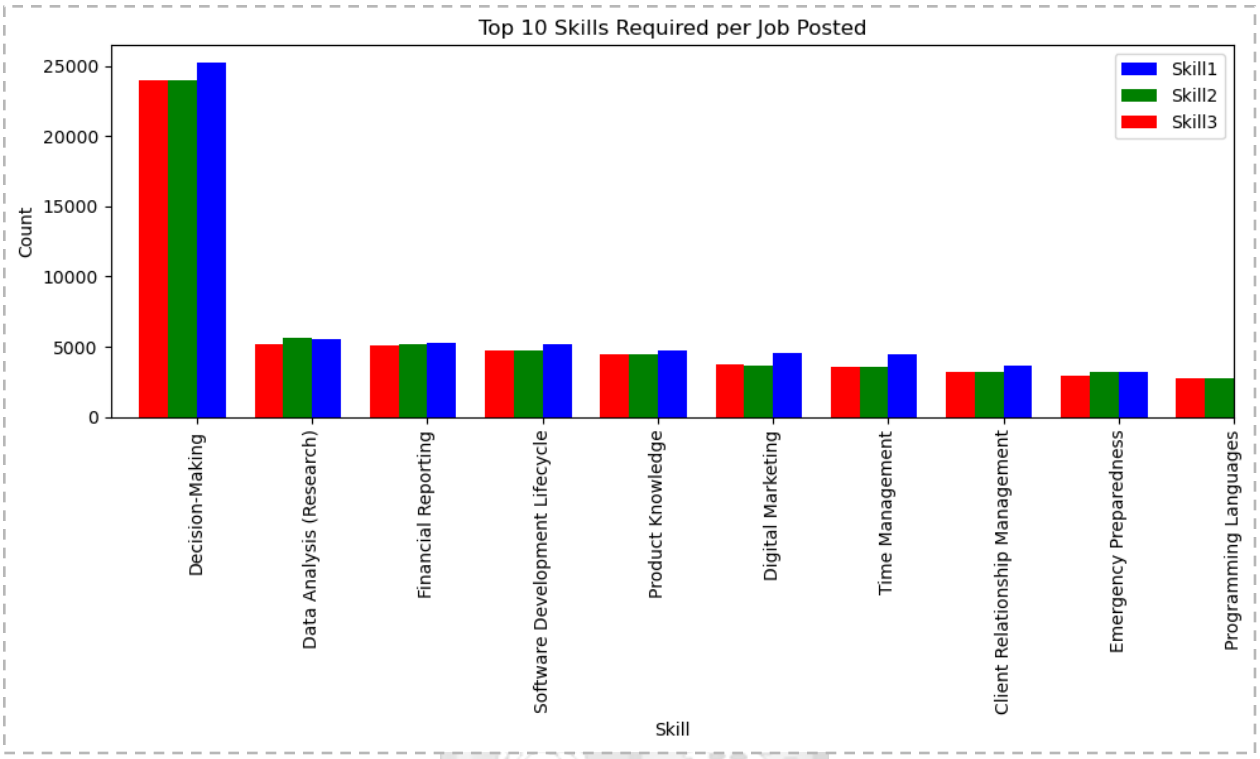


Figure 4.7: Top 10 skills required per job posted

Figure 4.7 highlights the most in-demand skills for the top job postings, with decision-making, data analytics, and financial reporting ranking at the top. Additionally, the data emphasizes the importance of digital skills and transferable skill sets, which are increasingly sought after by employers across various industries.

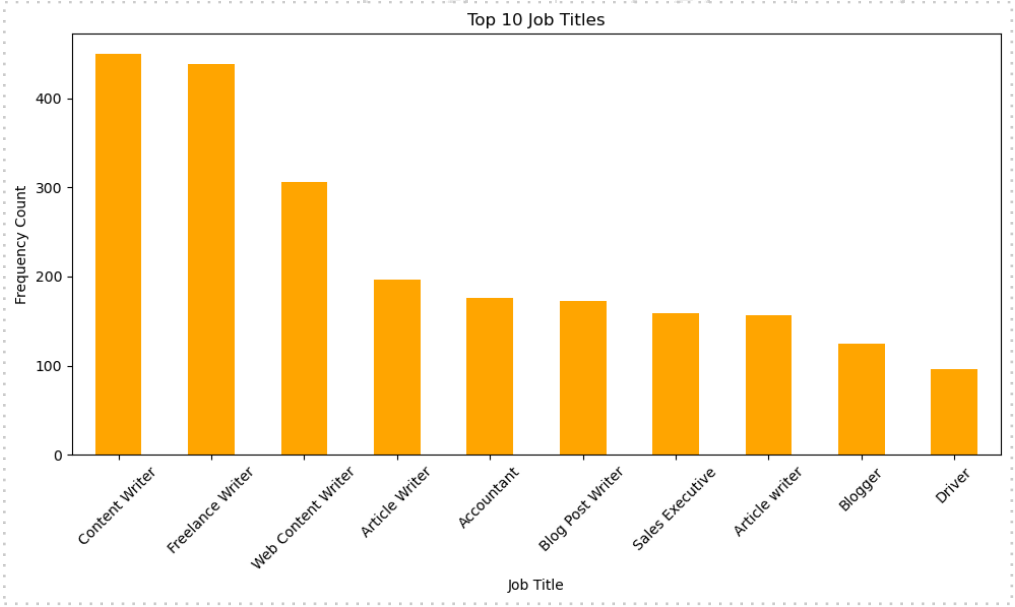


Figure 4.8: Top 10 job titles posted

Figure 4.8 shows current demographic trends in the job market. This is a composition of current active seekers across sub-Saharan. (TATC)

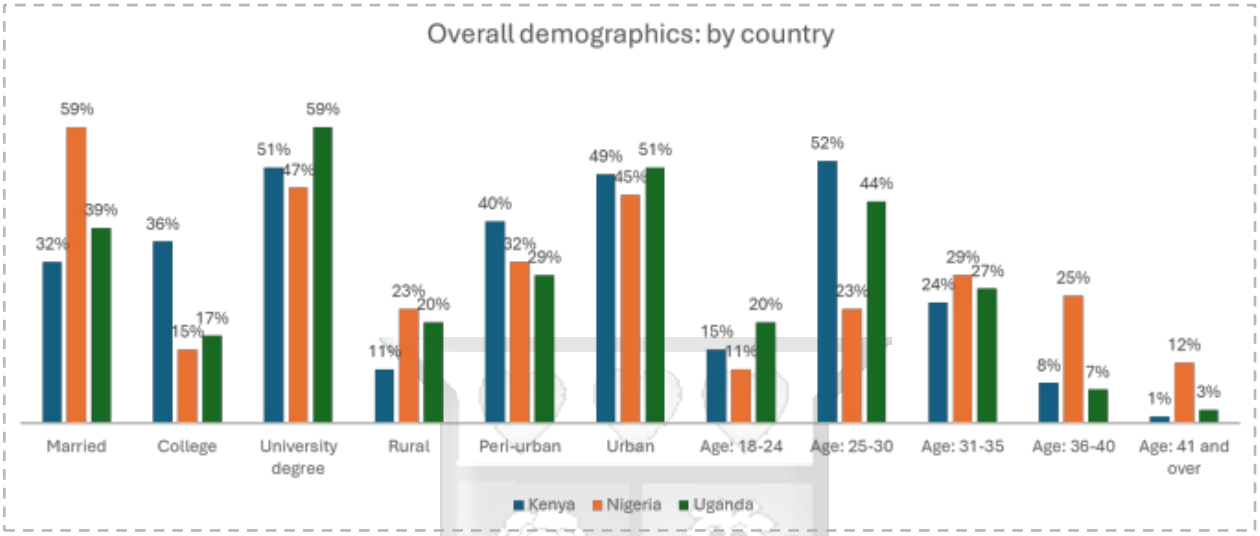


Figure 4.8: Overall demographics by Country

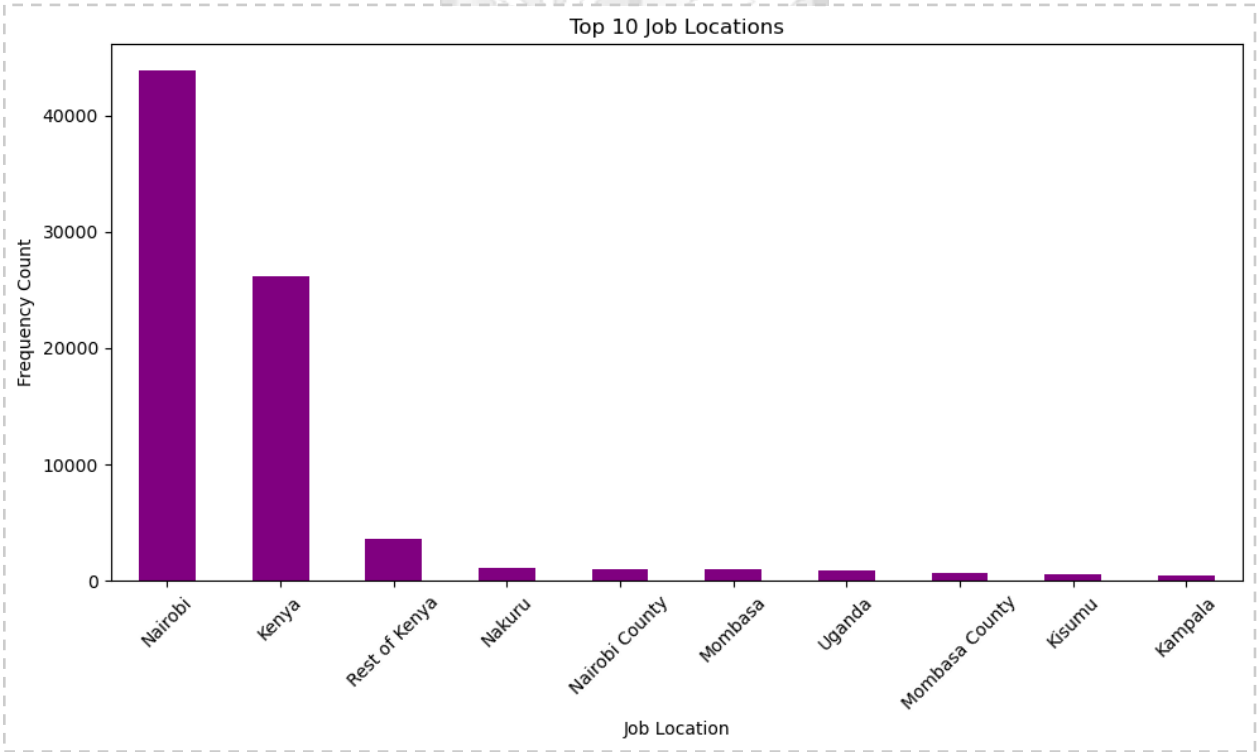


Figure 4.9: Top 10 Job locations

Figure 4.9 shows location for listed jobs. (Location is not explicitly provided by employers due to rise in remote working). These are online jobs, with most demand in Nairobi, overall in Kenya.

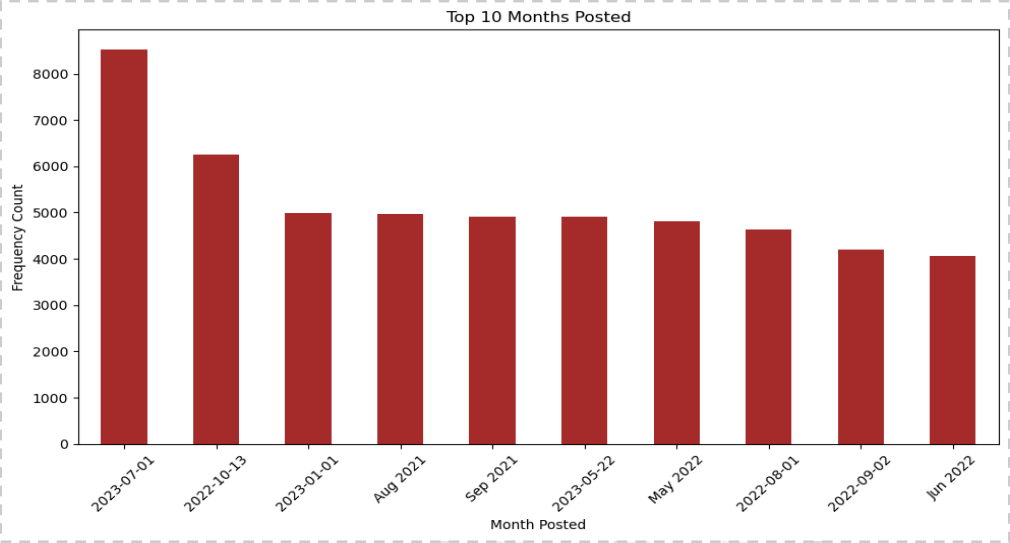


Figure 4.10: Top 10 months posted

Figure 4.10 illustrates the months with the highest levels of job market activity, with July, October, and January standing out as the peak periods for hiring. This pattern suggests that these months experience a significant surge in recruitment, potentially influenced by seasonal factors, budget cycles, or business planning periods.

4.3 Machine Learning Predictions / Results

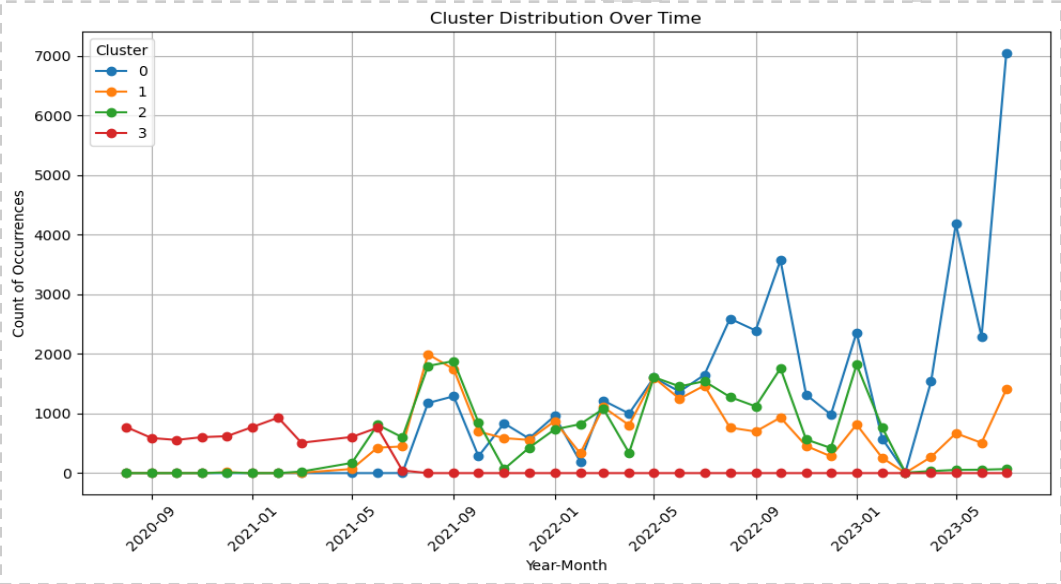


Figure 4.11: Cluster Distribution over time

Figure 4.11shows the model output; optimum clusters were 4. From above, we can see that Cluster 0 shows skills that are currently in high demand and cluster 1 and cluster 2 are moderate. Cluster 3 shows that the demand was high before but subsided with time.

Unique skills in cluster 0

Financial Analysis, Underwriting, Financial Reporting, Cybersecurity , Clinical Skills, Organization, Content Creation, Discipline, Inventory Management, Problem Solving, Research Skills, Training and Development , Legal Research, Volunteer Coordination, Patient Care, Market Research, Communication, Negotiation, Leadership, Team Management, Analytical Thinking, Programming, Curriculum Development, Data Analysis, Economic Analysis, Verbal Communication, Renewable Energy, Public Service, Crop Management, Food Preparation,
Guest Relations, Construction Management, Piloting, Makeup Artistry, Production Management, Acting, Research, Risk Assessment, Sales, Artistic Skills, Nutritional Analysis, Cross-Cultural Communication, Project Management, Extraction Techniques, Mechanical Skills, Crisis Management

Unique skills in cluster 3

Programming Languages, Research Methodology, Technical Problem-Solving, Project Planning, Strategic Planning, Occupational Health and Safety, Lean Manufacturing, Risk Assessment, Supply Chain Management, Sales Management, Destination Knowledge, Customer Service, Variable depending on industry,Recruitment,Quality Management Systems, Policy Analysis, Technical Skills, Medical Knowledge, Scientific Method, Real Estate Law, Fundraising, Mineral Extraction, Social Work, Marketing Strategy, Defensive Driving, Legal Research, Business Strategy,Initiative,Volunteer Coordination.

Figure 4.12: Industry vs Job functions

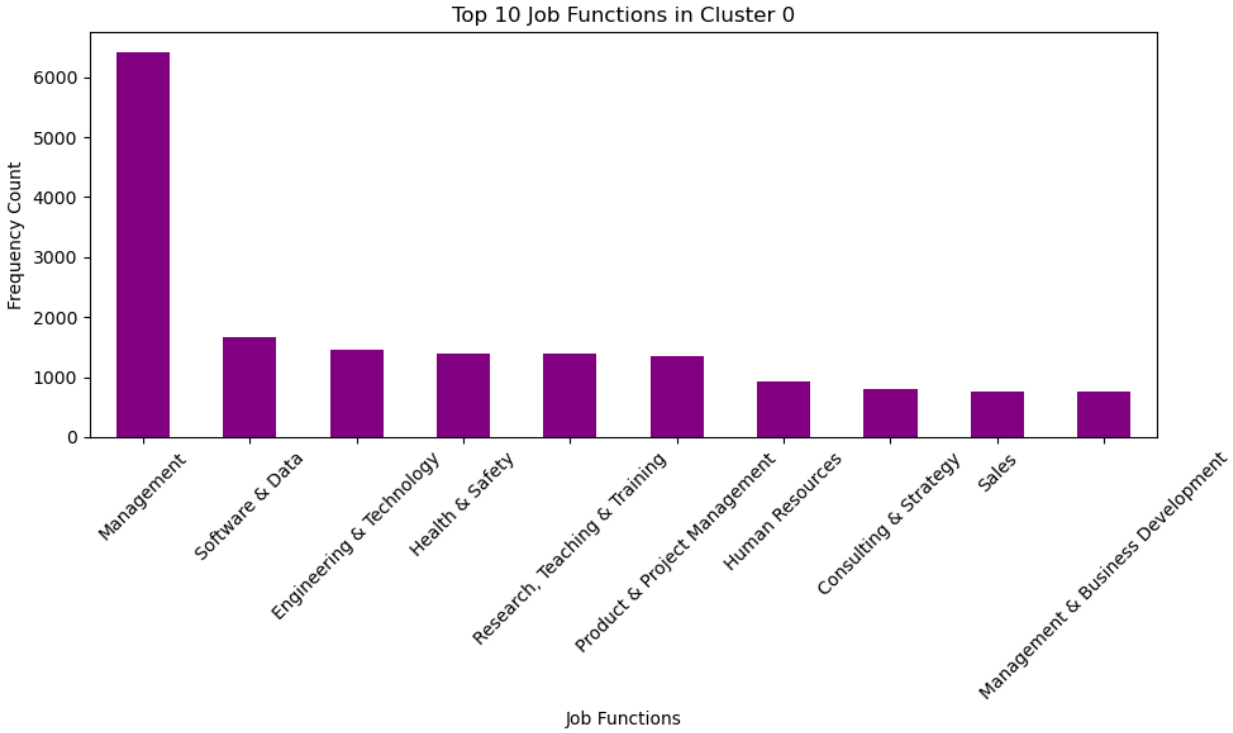


Figure 4.13. Visual representation of Top 10 job functions in Cluster 0

Figure 4.13 presents the top 10 job functions within Cluster 0, with management, engineering, and technology emerging as the leading sectors. Additionally, research, software development, and data-related industries rank highly, reflecting the growing demand for expertise in these fields.

This distribution highlights the prominence of sectors that require specialized skills, especially in the context of technological advancement and innovation.

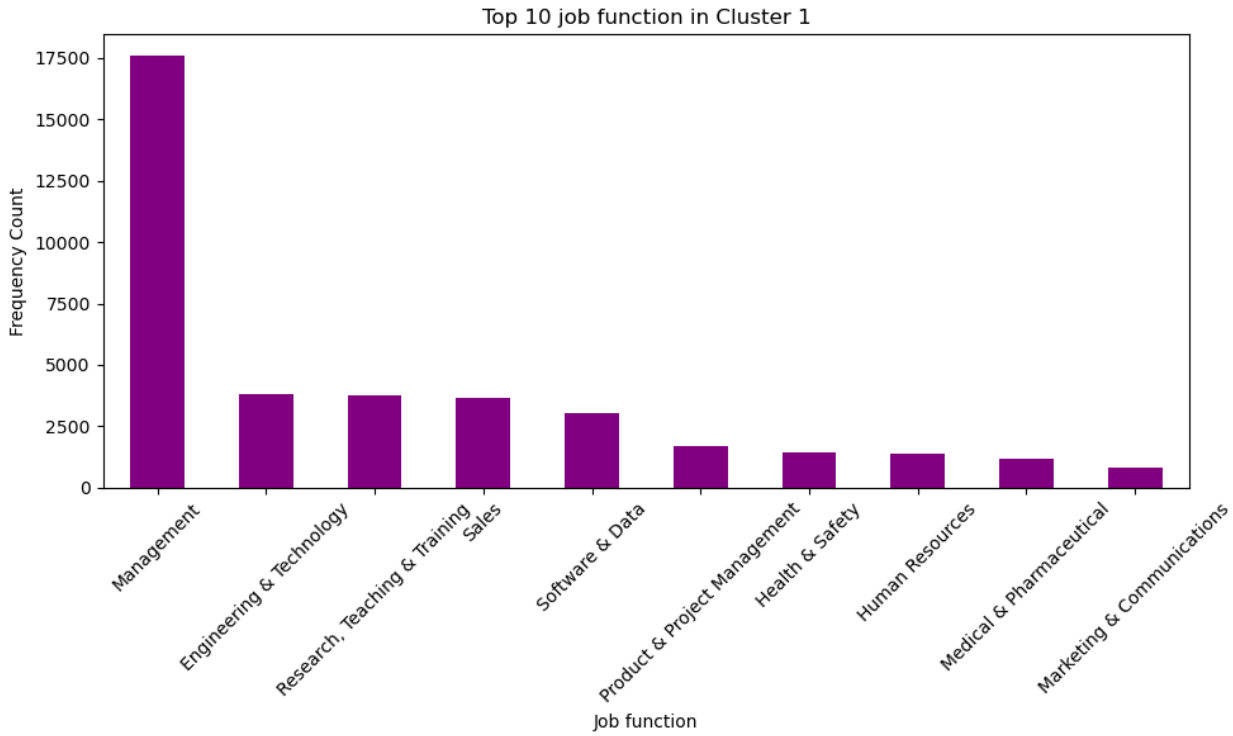


Figure 4.14. Visual representation of Top 10 job functions in Cluster 1

4.4 Model Deployment Results

The streamlit mode has 4 sections in the navigation panel, landing page/home page, the EDA, data clustering input and model section.

4.4.1 Home Page

The home page introduces the research background and objectives, helping users understand the app's purpose. It outlines the research goals and methodology, giving users context on the problem being addressed. This section ensures users are familiar with the app's functionality and its intended use. Overall, it helps users navigate the app effectively by aligning their expectations with its objectives.

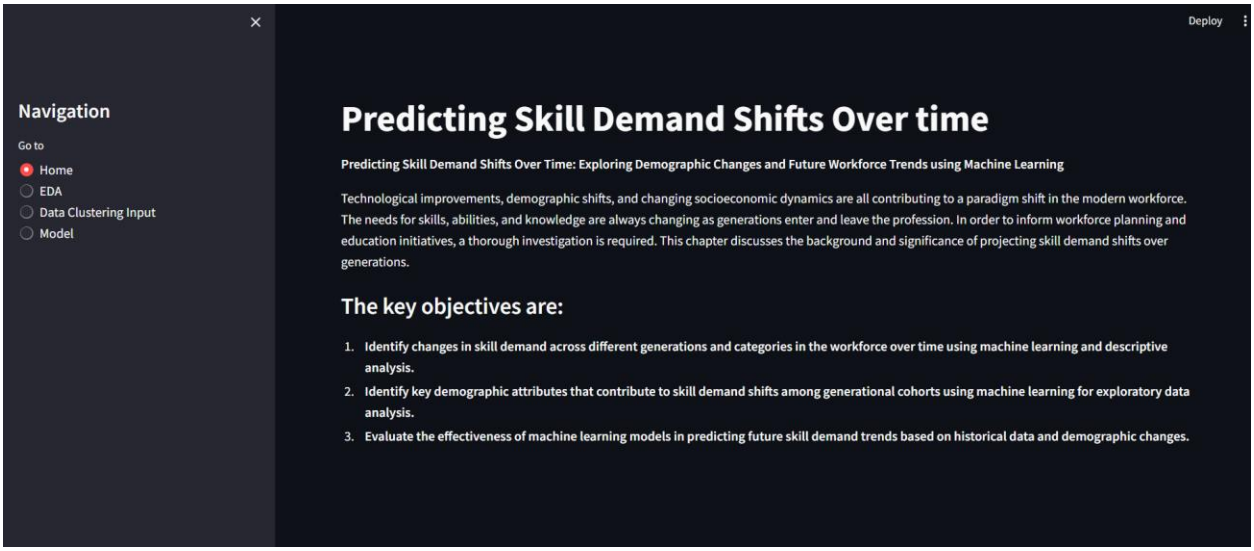


Figure 4.15: Streamlit App home page screenshot

4.4.2 EDA

The EDA section allows users to interact with the training data in real time, enabling them to add or modify fields and apply filters based on different metrics. It provides dynamic visualizations, such as histograms, scatter plots, and heatmaps, to uncover key patterns and trends in the data. Users can also visualize and interpret the results of machine learning models, helping to understand feature importance and model performance. This interactive approach empowers users to explore the data, make informed decisions, and gain deeper insights into the model's behavior.

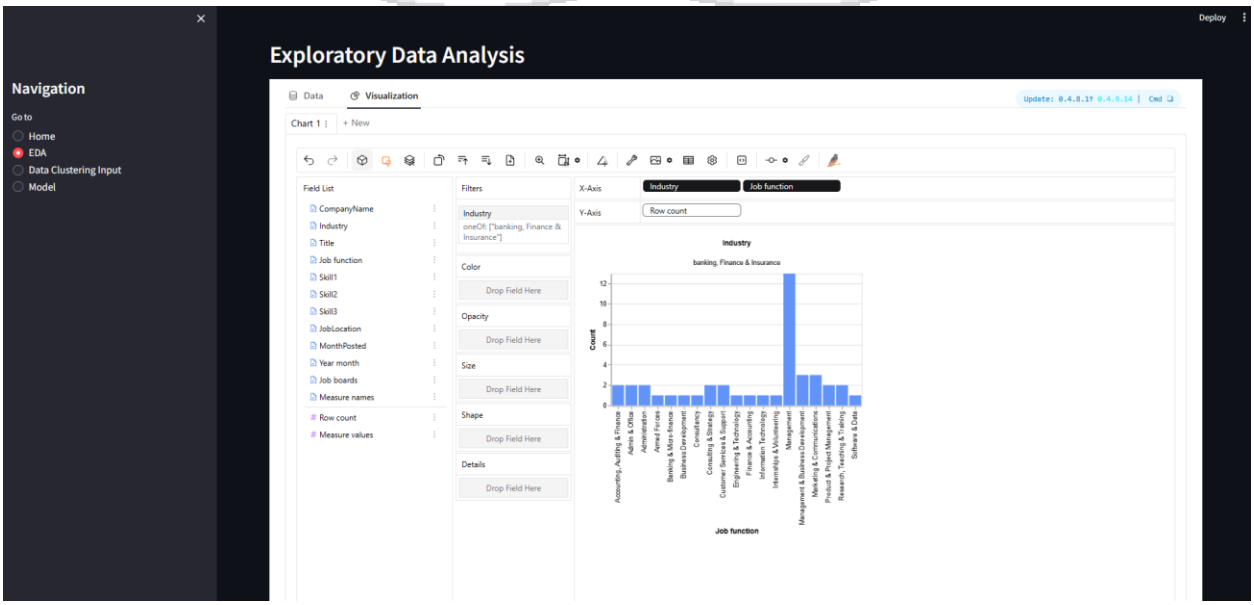


Figure 4.16: Streamlit app EDA page screenshot

4.4.3 Data Clustering Input

The Data Clustering Input section is where users can interact with the model to generate predictions. It allows users to input specific features through the provided data fields. After entering the necessary information, users can click the submit button located at the bottom of the form, which triggers the processing of their input. The trained clustering model is then applied to the submitted data, and the resulting predictions are displayed at the bottom of the form for the user to review. This section seamlessly integrates user input with the predictive capabilities of the model, providing an intuitive way to generate and visualize clustering outcomes based on real-time data.

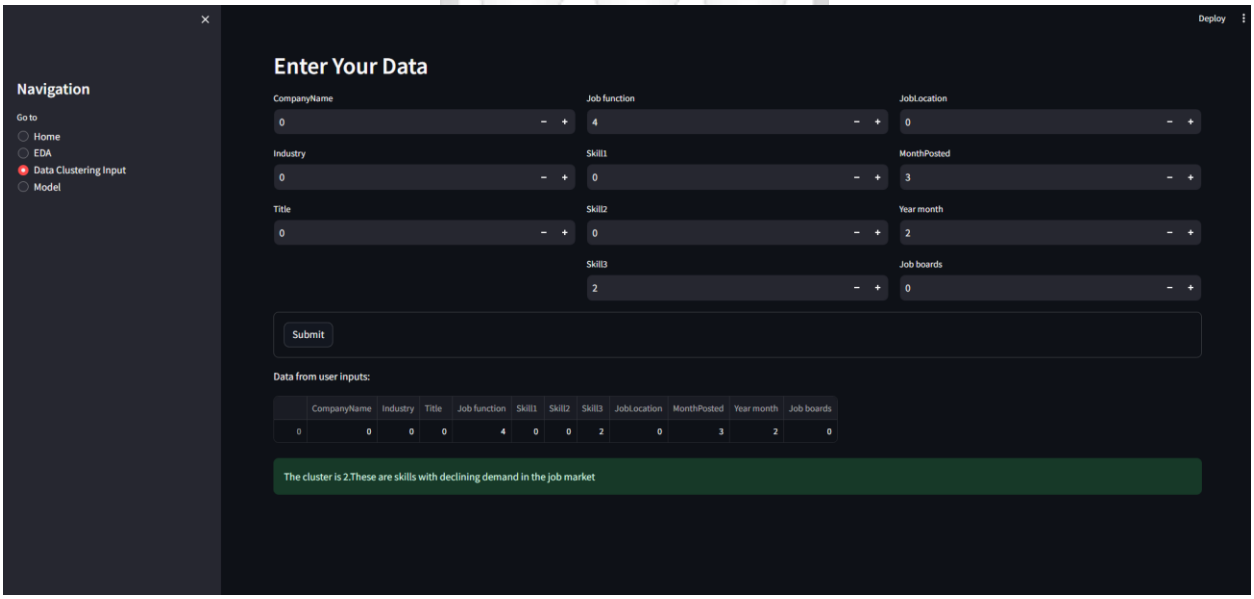


Figure 4.17 Streamlit app data clustering input screenshot

4.4.4 Model

The Training Data Preview section allows users to view the dataset used for model training, providing a clear and organized display of its structure. Users can explore the data to understand its features, variables, and relationships. This preview helps users familiarize themselves with the data, ensuring transparency and better-informed decision-making.

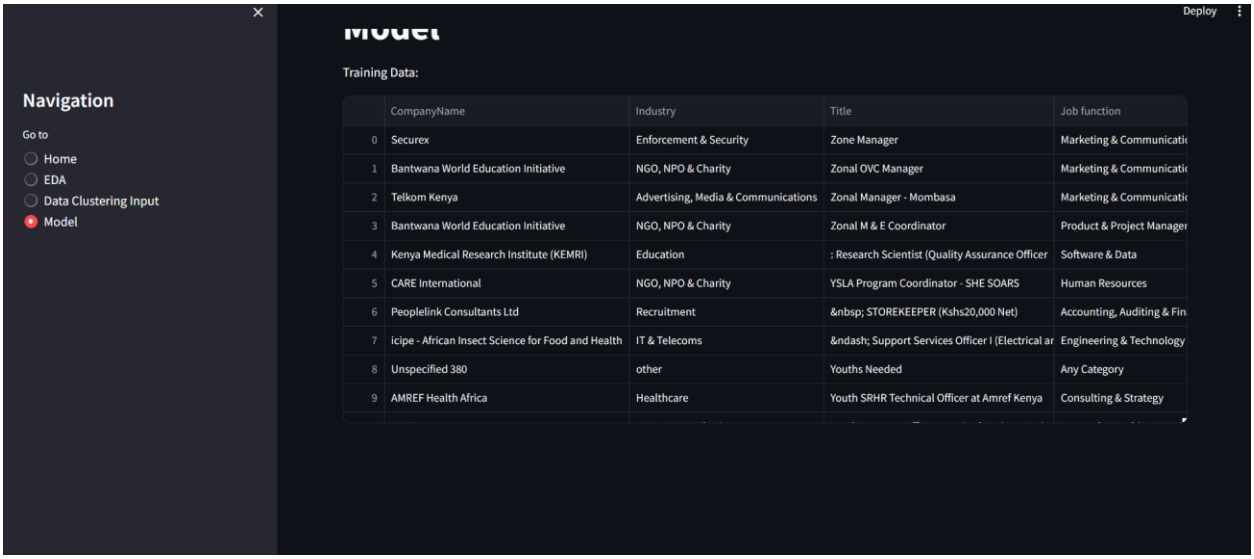


Figure 4.18 Streamlit app model training screenshot

Chapter 5: Discussion

5.1 Discussion

The clustering analysis revealed significant shifts in skill demand across generational and occupational categories between 2020 and 2024, fulfilling the objective to identify changes in skill demand across different generational cohorts and occupational categories over time using machine learning and descriptive analysis. Cluster 0 exhibited high and diverse demand for skills such as financial analysis, cybersecurity, leadership, and data analytics signaling a growing emphasis on hybrid and adaptable talent. Clusters 1 and 2 showed steady demand for core professional capabilities, while Cluster 3 captured declining relevance of once-prominent technical skills. These changes reflect evolving labor market needs, industry realignments, and generational transitions in skill preferences.

While the current analysis did not integrate demographic features into clustering, it sets the groundwork for future studies to uncover key demographic attributes that influence skill demand shifts among different generations using exploratory data analysis and clustering techniques. Incorporating variables such as age, education level, and work experience could enrich insights into how generational traits shape labor demand. Nevertheless, the current findings offer valuable implications: policymakers can use these patterns to guide reskilling initiatives, educators can align training with high-demand clusters, and employers can optimize hiring strategies by focusing on functional skill sets rather than static job roles. These practical takeaways also reinforce the study's theoretical grounding in human capital theory and labor market segmentation.

Finally, this study contributes to the objective to evaluate the effectiveness of machine learning models in forecasting future skill demand trends based on historical labor market data and demographic patterns. The use of unsupervised models like KMeans successfully uncovered relevant clusters, aligning with real-world trends and validating the use of machine learning for exploratory labor analysis. However, the study did not perform sensitivity tests or apply time-series forecasting models. Future work should evaluate the robustness of cluster results under different parameters and compare outcomes against traditional methods such as ISCO classification. These additions would strengthen the predictive utility of machine learning in understanding and anticipating labor market evolution.

These findings provide valuable insights for job seekers, employers, and policymakers, helping them understand the changing dynamics of the job market and adapt their strategies accordingly. Additionally, they underscore the importance of continuous skills development and staying abreast of industry trends to remain competitive in today's job market.

5.2 Limitations and Future Research

5.2.1 Limitations

This study has several limitations that should be considered. Firstly, the analysis is restricted to data from job postings and industry trends within Kenya, Uganda, Nigeria, and Ghana, which may not fully represent the broader African context. Secondly, the machine learning models used, such as KMeans, hierarchical clustering, and DBSCAN, have their own limitations in terms of accuracy and applicability to different types of data. Additionally, the study relies on historical data from 2020 to 2024, which may not capture the most recent trends in skill demand. Lastly, the focus was primarily on white-collar jobs, leaving blue-collar roles and skills underexplored. The model relies primarily on structured data from online job postings, which may not capture unmeasured or latent variables such as informal labor dynamics, socioeconomic inequalities, evolving education quality, or sudden policy or economic shifts. These factors may influence observed skill demand trends but are not directly represented in the features used

5.2.2 Future Research

Future research could address these limitations by expanding the scope of the study to include a more diverse set of African countries. More sophisticated machine learning models and techniques could be employed to improve the accuracy and relevance of predictions. There is also a need to include more recent data to ensure the findings remain current. Finally, future studies should explore blue-collar roles and skills to provide a more comprehensive understanding of the job market dynamics.

5.2.3 Dataset Bias and Representativeness

While the dataset used in this study provides a rich source of information from job postings, it inherently reflects the biases of online job markets. In many African countries, formal-sector and urban employers are more likely to post jobs online, which may lead to the underrepresentation of informal, rural, or blue-collar job opportunities. This introduces a form of selection bias, where certain sectors (e.g., tech, finance, education) are overrepresented. Additionally, no formal sampling strategy was applied during data collection — postings were scraped as they appeared, which limits the control over representativeness across industries, locations, and company sizes. As a result, the dataset may not fully reflect the broader dynamics of the labor market in the four countries studied. Future work should consider stratified sampling or complementary data sources (e.g., government labor statistics or survey data) to mitigate this bias and improve generalizability.

5.3 Recommendations

This study can be expanded to:

Collaborate across academia, research, and industry to benchmark course intakes, industry uptake, and skill supply in the market.

Leverage APIs or web scraping techniques to collect real-time data from job boards, professional networking sites, or government databases.

Apply NLP techniques such as topic modeling, sentiment analysis, or named entity recognition to extract deeper insights from textual data sources such as job descriptions or social media posts.

Explore blue-collar roles/skills/jobs, which are still an unexplored area in data recording, collection, and tracking by relevant bodies.

References

Acemoglu, D., & Autor, D. H. (2012). What Does Human Capital Do? A Review of Goldin and Katz's The Race between Education and Technology. Journal of Economic Literature, 50(2), 426-463.

Autor, D. H. (2019). *The labor market and the economy: The persistent mismatch between skill supply and job demand*. Harvard University Press.

Bloom, D. E., Canning, D., & Fink, G. (2011). Implications of population aging for economic growth. Oxford Review of Economic Policy, 27(4), 583-600. https://doi.org/10.1093/oxrep/grr025

Bossaert, D., Demmke, C., & Moilanen, T. (2021). The impact of demographic change and its challenges for the workforce in the European public sectors: Three priority areas to invest in future HRM.

Brynjolfsson, E., & McAfee, A. (2014). The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies. W. W. Norton & Company.

Carnevale, A., Smith, N., & Strohl, J. (2010). Help wanted: Projections of jobs and education requirements through 2018. (Tech. Rep.). Georgetown Center on Education and the Workforce.

Draskovic, D., Drinic, A., Kylymnyk, I., Seyidzade, L., Tilavov, M., & Tuna, G. (2021). The changing nature of work: 30 signals to consider for a sustainable future. UNDP. Retrieved from https://www.undp.org/sites/g/files/zskgke326/files/migration/acceleratorlabs/NEW_The-Changing-Nature-of-Work_15-June-2021_FINAL.pdf

Duncan, G. and Hoffman, S. (1981). The incidence and wage effects of overeducation. Economics of Education Review, 1(1):75–86.

Estrin, S., Stephan, U., & Vujić, S. (2016). Do Women Earn Less Even as Social Entrepreneurs? Labour Economics, 41, 363-371.

European Labour Authority, (2023). Report on labour shortages and surpluses: 2022, Publications Office of the European Union. https://data.europa.eu/doi/10.2883/50704

Frey, C. B., & Osborne, M. A. (2017). The future of employment: How susceptible are jobs to computerization? Technological Forecasting and Social Change, 114, 254-280. https://doi.org/10.1016/j.techfore.2016.08.019

Fry, R., & Taylor, P. (2012). The Rise of the Second Generation: Changing Patterns in Population Growth and Geographic Distribution. Pew Research Center.

Gereffi, G. (2018). Globalization, development, and the global value chains: The role of the global south. World Development, 108, 1-18. https://doi.org/10.1016/j.worlddev.2018.03.009

Government of Kenya. (2023, July 11). Skill Up Africa Expo Conference - World Youth Skills Day 2023. Kenya Labour Market Information System. https://www.labourmarket.go.ke/news-itme/skill-up-africa-expo-conference-world-youth-skills-day-2023/11/

Hotz, N. (2018, September 10). What is CRISP DM? *Data Science Process Alliance*. https://www.datascience-pm.com/crisp-dm-2/

J. M. Luna-Romera, F. Nuñez-Hernández, M. Martinez-Ballesteros, J. C. Riquelme, & C. Usabiaga Ibáñez. (2019). Analysis of the Evolution of the Spanish Labour Market Through Unsupervised Learning. IEEE Access, 7, 121695-121708.

Leuven, Edwin; Oosterbeek, Hessel (2011) : Overeducation and mismatch in the labor market, IZA Discussion Papers, No. 5523, Institute for the Study of Labor (IZA), Bonn.

McKinsey Global Institute. (2017). Jobs lost, jobs gained: Workforce transitions in a time of automation. McKinsey & Company. https://www.mckinsey.com/featured-insights/future-of-work/jobs-lost-jobs-gained-workforce-transitions-in-a-time-of-automation

N. Dawson, M.-A. Rizoiu, B. Johnston, & M.-A. Williams. (2020). Predicting Skill Shortages in Labor Markets: A Machine Learning Approach. In 2020 IEEE International Conference on Big Data (Big Data) (pp. 3052-3061). Atlanta, GA, USA. DOI: 10.1109/BigData50022.2020.9377773.

Nguyen, D. A., Nguyen, Q. V., Nguyen, A. T., & Dinh, T. T. A. (2020). A Machine Learning Approach to Predict Skill Demand in the Labor Market. In Proceedings of the 11th Asian Conference on Intelligent Information and Database Systems (pp. 93-101). Springer.

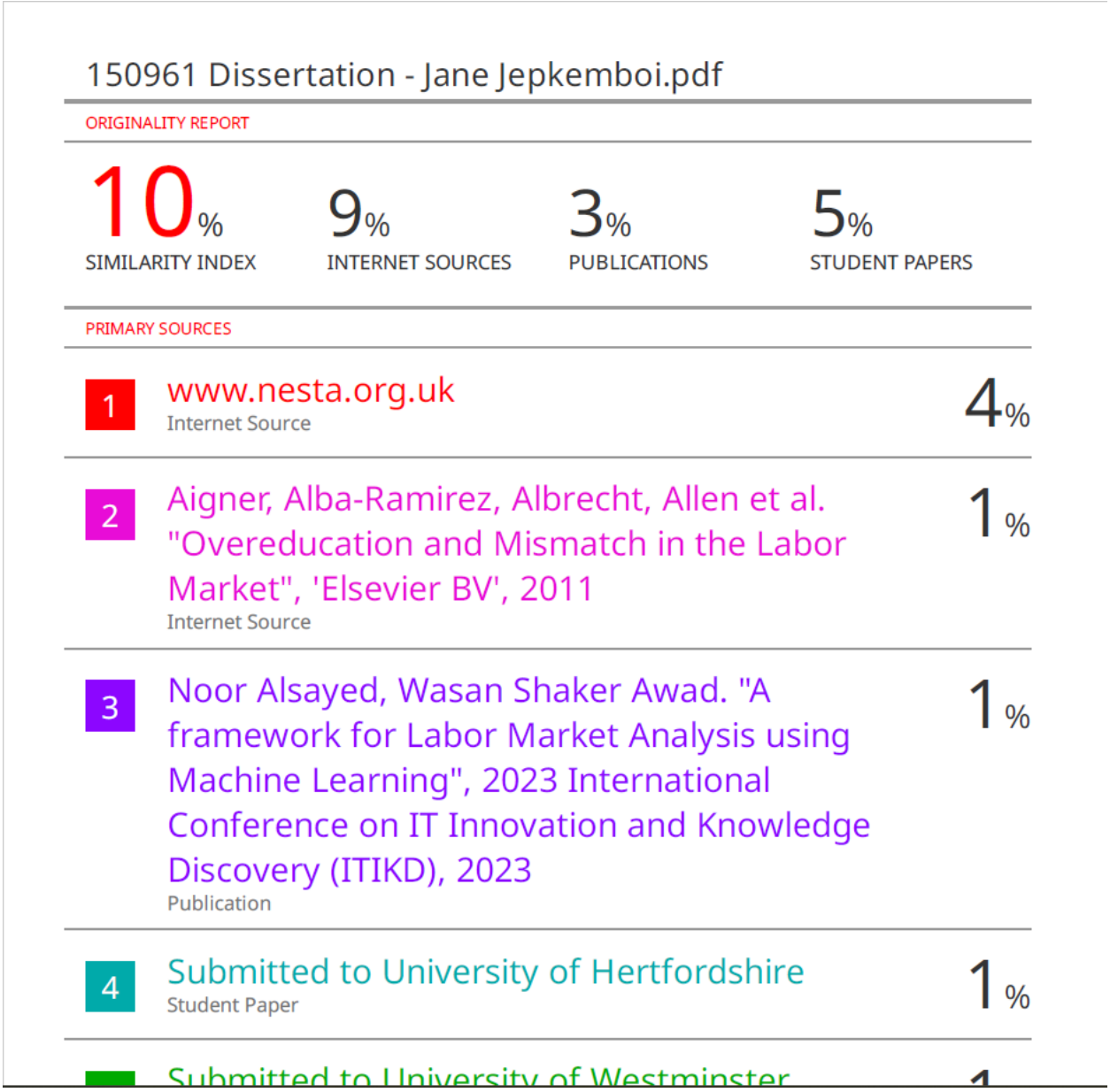
Schleicher, A. (2015, December 17). How can we equip the future workforce for technological change? World Economic Forum Agenda (blog).

World Economic Forum. (2020). The Future of Jobs Report 2020. World Economic Forum.

https://www.weforum.org/reports/the-future-of-jobs-report-2020

Appendices

Appendix A: Similarity Report



150961 Dissertation - Jane Jepkemboi.pdf

ORIGINALITY REPORT

10%	9%	3%	5%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.nesta.org.uk Internet Source	4%
2	Aigner, Alba-Ramirez, Albrecht, Allen et al. "Overeducation and Mismatch in the Labor Market", 'Elsevier BV', 2011 Internet Source	1%
3	Noor Alsayed, Wasan Shaker Awad. "A framework for Labor Market Analysis using Machine Learning", 2023 International Conference on IT Innovation and Knowledge Discovery (ITIKD), 2023 Publication	1%
4	Submitted to University of Hertfordshire Student Paper	1%
	Submitted to University of Westminster	1

Appendix B: Ethical Clearance Confirmation

Strathmore UNIVERSITY

24th April 2024

Ms Jepkemboi Jane,
jepkemboi.jane@strathmore.edu

Dear Ms Jepkemboi,

RE: Predicting Skill Demand Shifts Over time: Exploring Demographic Changes and Future Workforce Trends using Machine Learning

This is to inform you that SU-ISERC has reviewed and **approved** your above **SU-masters** research proposal. Your application reference number is **SU-ISERC2153/24.** The approval period is from **24th April 2024 to 23rd April 2025.**
This approval is subject to compliance with the following requirements:

i. Only approved documents including (informed consents, study instruments, MTA) will be used.
ii. All changes including (amendments, deviations, and violations) are submitted for review and approval by SU-ISERC.
iii. Death and life-threatening problems and serious adverse events or unexpected adverse events whether related or unrelated to the study must be reported to SU-ISERC within 72 hours of notification.
iv. Any changes anticipated or otherwise that may increase the risks or affected safety or welfare of study participants and others or affect the integrity of the research must be reported to SU-ISERC within 72 hours.
v. Clearance for the export of biological specimens must be obtained from relevant institutions.
vi. Submission of a request for renewal of approval at least 60 days prior to the expiry of the approval period. Attach a comprehensive progress report to support the renewal.
vii. Submission of an executive summary report within 90 days of completion of the study to SU-ISERC.

Before commencing your study, you will be expected to obtain a research license from National Commission for Science, Technology, and Innovation (NACOSTI) https://research-portal.nacosti.go.ke/ and obtain other clearances needed.

Yours sincerely,

Mr Ambrose Rachier,
Chairperson; SU-ISERC

STRATHMORE UNIVERSITY INSTITUTIONAL SCIENTIFIC AND ETHICAL REVIEW COMMITTEE
(SU-ISERC)
24-Apr-2024
Email:ethicsreview@strathmore.edu
P.O BOX 59857-00200
NAIROBI-KENYA

VT OMNES
VNVM SINT