2020-03

Data visualization tool for student dropouts in Tanzania

Kayanda, Angelika Mark

NM-AIST

https://doi.org/10.58694/20.500.12479/901

DATA VISUALIZATION TOOL FOR STUDENT DROPOUTS IN TANZANIA

Angelika Mark Kayanda

A Dissertation Submitted in Partial Fulfilment of the Requirements for the Degree of Master's in Information and Communication Science and Engineering of the Nelson Mandela African Institution of Science and Technology

Arusha, Tanzania

March, 2020

ABSTRACT

Education is a crucial sector and key component of many governments' agenda. Despite that, students dropout has been among the persisting challenges in education, being experienced from basic education levels to colleges and universities. This work presents a study on data visualization in education and a glimpse of data visualization in other domains, and suggests a web based data visualization tool for student dropouts in Tanzania, targeting primary and secondary schools. It also presents users' feedback regarding the developed web-based tool.

This study collected data from the Government Basic Statistics Portal and the President's Office - Regional Administration and Local Government (PO-RALG). Rapid prototyping was employed to develop the proposed web-based data visualization tool for interactive visualization of the prepared data. Moreover, focus group discussions and questionnaires were used to gather feedback from the users, whereby majority agreed that data visualization is useful for understanding data and providing insights for reporting and decision making. Most users further agreed that the developed tool was easy to use, useful and recommendable.

The various challenges that came up during the course of this study enlightens on the need for sound data collection practices and the need for good visualization literacy. The results from this study will be beneficial for decision makers in education domain by providing a wide range of options to visually compare and observe variations in the student dropouts trends, thus facilitating informed decisions. Moreover, this study will be useful source of information for further research in dropout prediction and modelling.

DECLARATION

I, Angelika Mark Kayanda, do declare hereby to the Senate of the Nelson Mandela African Institution of Science and Technology that, this dissertation is my own original work and that it has neither been submitted nor is it being concurrently submitted for a degree award in any other institution.

Angelika Mark Kayanda 25/03/2020

Name and signature of candidate **Date**

The above declaration is confirmed

Dr. Dina Machuve 25/03/2020

Name and signature of Supervisor **Date**

COPYRIGHT

CERTIFICATION

I, the undersigned certify that I have read and found the dissertation titled "Data Visualization Tool for Student Dropouts in Tanzania" conforming to the standard and format acceptable by the Nelson Mandela African Institution of Science and Technology.

Dr. Dina Machuve 25/03/2020

Name and signature of Supervisor **Date**

ACKNOWLEDGEMENTS

I am first and foremost very grateful to Almighty God for His great love, care and blessings He has given me, without which I would not have been able to successfully pursue this Master's degree at the Nelson Mandela African Institution of Science and Technology (NM-AIST).

My heartfelt gratitude goes to my supervisor, Dr. Dina Machuve for her invaluable guidance and support in ensuring that this work is successfully accomplished.

I sincerely thank the African Development Bank (AfDB) for the financial support from the start to the finish of my Master's study.

I especially thank all my classmates, friends and colleagues at NM-AIST for their endless support throughout the time of my study.

Deep gratitude goes to my family for always being there for me, with love, care, solemn wishes and prayers.

Thank you so much all.

DEDICATION

To my loved family.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
API	Application Programming Interface
BEST	Basic Education Statistics
D3	Data Driven Documents
DOM	Document Object Model
HTTP	HyperText Transfer Protocol
ICT	Information and Communication Technology
JSX	JavaScript XML
MOOCs	Massive Open Online Courses
NBS	National Bureau of Statistics
npm	Node Package Manager
OECD	Organization for Economic Cooperation and Development
ORM	Object-Relational Mapping
PDF	Portable Document Format
PEO	Principal Education Officer
PO-RALG	President’s Office - Public Administration and Local Government
SDLC	Software Development Life Cycle
SVG	Scalable Vector Graphics
TAMISEMI	Tawala za Mikoa na Serikali za Mitaa
UML	Unified Modelling Language
UNDP	United Nations Development Programme
XML	Extensible Markup Language

CHAPTER ONE

INTRODUCTION

1.1 Background of the Study

Education plays a key role in providing the required skills and competencies for ensuring civilized living, community engagement, increased productivity, and supporting creativity and innovation in solving the everyday global challenges (OECD, 2016). In Article 26 of the Universal Declaration of Human Rights, education is declared as every individual's right (United Nations, 1948). Education is among the highly prioritized sectors in Tanzania and a key component of government's development agenda.

For a number of years now, the education sector budget has been higher than all other priority sectors accounting for 20% of the total national budget, and declining to 15% in the fiscal year 2017/2018 (Ministry of Finance and Planning, 2018a). Significant efforts are being continually made to expand, improve and ensure universal access to education, one of them being the implementation of free basic education policy in 2016. Currently, the Government pays 20.8 billion Tanzanian shillings monthly to ensure free basic education, which has resulted to increased enrollments in the basic education levels (Ministry of Finance and Planning, 2018b). Nevertheless, the Tanzania education sector face a number of persistent challenges including poor learning outcomes, equity issues, rapid changes in technology, quality of teaching, linkage between education system and the labor market, student retention and completion, to mention but a few (HakiElimu, 2015).

As student retention and completion of school remain a challenge, large numbers of students who were enrolled and supposed to graduate are out of school. A significant amount of the nation's current and future intellect is lost due to student dropouts. The United Republic of Tanzania through the Government Open Data Portal estimated about 56 573 primary school dropouts for both boys and girls in the year 2013. The numbers significantly increased to about 85 985 pupils in 2015 and 117 927 pupils in 2016 (Datasets - Basic Statistics Tanzania, 2018). Secondary school dropouts were about 94 990 in 2012; 61 488 in 2015 and 63 903 in 2016 (Datasets - Basic Statistics Tanzania, 2018). A study on student dropouts in three regions of Tanzania revealed that socio-economic and political factors, together with parents' views and government contribute to school dropouts (Kalinga, 2013). The study results were presented in percentages, tables and simple text. Similarly, a study focused on the artisanal mining areas of Geita region in Tanzania presented the data regarding the dropouts causes in mostly text (Ouma, 2017). This kind of presentation provide limited interactivity and difficulty in interpreting the information or revealing trends which are helpful in decision making.

On the other hand, data visualization techniques make it possible to present data in visual format which leverages on the human visual processing capabilities, thus easily perceived by the human mind (Alexandre & Tavares, 2010). Visualization works best with the human cognitive system, which is known to have active, quick and strong capabilities in visual processing (Pasternak, Bisley & Calkins, 2003). Humans are known to best interpret visual information than numbers or text. Visualization graphically presents data and intensifies its storytelling or persuasive function which helps create arguments and reasonable explanations of the events occurring in the world around us (Williamson, 2016). Additionally, countless number of works show that data visualization is important in providing insights into complex datasets, by communicating key aspects in a more intuitive and meaningful way (Li *et al.*, 2018). Visualization has become vital in decision making for almost every knowledge area.

In education domain, data visualization has been less explored compared to other domains like health and commerce (Chen, Härdle, Unwin & Friendly, 2008). The United Republic of Tanzania Government Open Data Portal has visualized maps of primary schools pupil teacher ratio, leaving examination performance, pupil classroom ratio and form four examination performance for the year 2014 but lacking visualizations for student dropouts ("Dashboard - Basic Statistics Tanzania," 2019). Moreover, TAMISEMI publishes yearly statistics of student dropouts in Tanzania, which contain a number of visual presentations, but they are static and at a very high level ("Allbest | PO-RALG," 2019). The National Bureau of Statistics (NBS) visualizes Tanzania's statistics on various issues including education, however, for the case of dropouts, very little information can be found (National Bureau of Statistics, 2019).

Nonetheless, various studies on visualizing student dropouts have been conducted in different learning environments. Predictive analytics and data visualization have been used on longitudinal data for 6^{th} to 12^{th} grade from school information system for assisting educators in individual-level decision making and intervention (Lacefield & Applegate, 2018). A visualization tool known as DropoutSeer was developed for Dropout Reasoning and Prediction for the Massive Open Online Courses (Chen *et al.*, 2017). Another study used a visualization tool known as DataWrapper to analyze the social issues leading to girl-child school dropout in India (Samantaray & Dash, 2018). To assist in proactively giving personalized guidance to pupils, a data driven system to predict course grades and intention of dropping out was developed (Rovira, Puertas & Igual, 2017). For identification of students who are at risk of dropping out and intervene early, a recent study employed machine learning techniques develop a predictive model for dropouts in secondary schools in Tanzania. This study also embedded a visualization module. However, the module visualized school data from only five regions in Tanzania (Mduma, Kalegele & Machuve, 2019). A variety of visualizations were used in order to provide early help to students by analyzing the causes of dropout that are related to curriculum, using data from computer science degree students' progress in Germany (Askinadze, Liebeck & Conrad, 2019). Both of these studies have indicated that visualization plays a key part in identifying key causes and their variations in data, thus assisting in decision making.

1.2 Statement of the Problem

The increasingly generated data being stored in massive datasets bring about difficulty and infeasibility to using the standard tools that have been utilized in the past decades for data analysis (Wang, Zhang, Ma, Xia & Chen, 2016). Currently, data visualization has become important in uncovering trends and patterns in datasets belonging to various domains and a number of tools have been developed for that purpose. In education domain, visualization tools which have been developed are mostly for e-learning platforms and the Massive Open Online Courses (MOOCs) (Kuosa *et al.*, 2016; Chen *et al.*, 2017). Furthermore, most visualizations for education either do not present student dropouts data or present only primary schools data, only secondary school data or only girls' dropouts data (Samantaray & Dash, 2018; Mduma *et al.*, 2019). Thus little is found about tools for visualizing data concerned with student dropouts in the regular school and learning environment which is common in Tanzania. By building upon the already done works on student dropouts and developing a data visualization tool in that context, the student dropouts problem can be regarded in a new way. Interactive data visualizations will give the decision makers the possibility of digging deeper into the data, and thus get dynamic results of what they require in a simple, meaningful and intuitive manner (Roels, Baeten & Signer, 2017).

1.3 Rationale of the Study

The way data is presented matters a lot to decision makers in the respective domain since data is critical for informed decision making. Compared to other ways of information presentation such as tables or plain text, visualization takes advantage of the mind's visual processing capability making it easy for users to quickly and intuitively understand data (Alexandre & Tavares, 2010). Consequently, this leads to making of sound decisions based on data. Apart from that, data visualization is useful for exploratory data analysis before applying machine learning models. This study then will be useful source of information for AI and machine learning researchers who are interested in dropout prediction and modelling.

1.4 Objectives

1.4.1 Main Objective

This study aims to develop a data visualization tool for student dropouts in Tanzania.

1.4.2 Specific Objectives

(i) To identify datasets with student dropouts data and process them into relevant datasets for visualization.

(ii) To develop a tool that interactively visualizes student dropouts data.

(iii) To validate the developed tool.

1.5 Research Questions

(i) How will the datasets with student dropouts data be identified and turned into relevant datasets for visualization?

(ii) How will the dropout visualization tool be developed?

(iii) How will the data visualization tool be validated?

1.6 Significance of the Study

Visualizations are easily and intuitively understood with less cognitive load. Thus, data visualization will support informed decision making, by providing a wide range of options for decision makers to visually compare and observe variations in the student dropouts trends. They will hence draw meaningful conclusions with less effort compared to other computational means.

1.7 Delineation of the Study

The study is focused on student dropout, thus it will use a very small portion of education data. The study has also visualized the student dropouts data up to council level. Furthermore, the available data is annual compiled thus the study can only report annual visualization results.

CHAPTER TWO

LITERATURE REVIEW

2.1 Data Visualization

Data visualization is a technique for transforming data into various graphical representations, with the aim of intuitively incorporating human intelligence into analysis process, thereby enhancing the human ability of understanding and exploring datasets (Wang *et al.*, 2016). Currently, data visualization is a rapidly growing field that has drawn the interest of many data scientists and researchers. The deep roots of information visualization can be traced back to the earliest graphic presentations such as the early maps and statistical visual depictions (Chen *et al.*, 2008). Over the course of time, improvements and innovations in data visualization tools have facilitated the uncovering of patterns and trends hidden in massive datasets and at the same time the invisible rendering of the underlying algorithmic and statistical techniques performed on them (Williamson, 2016).

As the field of data visualization grows and awaken more interest, various scholars have proposed frameworks, models and tools for data visualization. In 2005, an information visualization toolkit was developed using the Java 2D graphics library of Java programming language with the aim of reducing the difficulty in building visualizations by providing a user interface for creating interactive visualizations (Heer, Card & Landay, 2005). In 2009, a four-level nested model for design and evaluation of visualizations was proposed. In this model, visualization design was split such that the output form the upper level became the input to the lower level (Munzner, 2009). A nine stage linear framework for visualization design study was later proposed, which insisted on collaboration between visualization researchers and the domain experts throughout the design study (Sedlmair, Meyer & Munzner, 2012). A design activity framework for visualization which provides actionable guidance and advice for informing designers the stage they are in, methods they can use and decisions they can make at each stage was proposed (Mckenna, Mazur, Agutter & Meyer, 2014). This actionable guidance was lacking in the previously suggested nine-stage framework for visualization design study. Visual embedding was proposed as a model for construction of visualizations, whereby it was argued that a good visualization ensures that the structures present in the data domain are preserved by the embedded visual elements (Demiralp, Scheidegger, Kindlmann, Laidlaw & Heer, 2014). A study in 2018 proposed a structural model which represents the relationship between data, its representation and interactive visualization (Tamayo, Hernández & Gómez, 2018). A declarative framework for users with no programming skills was developed, to enable rapid and easy construction of interactive, web-based visualizations (Li *et al*., 2018).

In various domains, a number of data visualization tools have been developed for meaningful and intuitive data representation. Most of these tools employ tables, plots, histograms, various kind of charts or a combination of them, Venn diagrams and timelines to visualize data. A few use treemaps, semantic networks, parallel coordinates and cone trees which are still less known and less used techniques (Wang, Wang & Alexander, 2015). Text visualization has also been widely applied through word clouds, word trees, and phrase-nets for conveying relationships in occurrences among terms; while river visualizations have been used to present dynamic textual information (Liu, Cui, Wu & Liu, 2014). Recently, scholars have pushed visualization further to effectively visualizing errors and uncertainties using hypothetical outcome plots (Kale, Nguyen, Kay & Hullman, 2019).

In visualization design, consideration of how viewers perceive and interpret data plays a key part in informing visual encoding decisions. Consequently this leads to creation of effective visualizations (Demiralp, Bernstein & Heer, 2014). Therefore, as far as information visualization is concerned, it is important to evaluate user performance on how they interpret information on graphical displays (Liu & Shen, 2015).

2.2 Visualization and Education Domain

Within the education domain, data visualization has been employed in various aspects. A visualization tool was developed for Dropout Reasoning and Prediction for the MOOCs. This tool aimed at informing education domain experts of the reasons for dropout and help researchers in machine learning improve their dropout predictive models (Chen *et al.*, 2017). The Pearson's Learning Curve provides dynamic, interactive and easy-to-use data visualization tools to enable researchers and policymakers derive meaning from global education datasets (Williamson, 2016). Furthermore, visualization was found to unearth interesting findings and more detailed depiction of learners' behaviors in an adaptive learning environment, which would otherwise remain hidden (Liu, Kang, Pan, Zou & Lee, 2017). For improved teaching and learning online, two interactive visualization tools were developed at the Unitelma Sapienza University and the IISLab of Tampere University of Technology for the Moodle Learning Management System (LMS) as plug-ins. These tools assist both students and teachers to monitor, evaluate and make decisions to improve learning outcomes (Kuosa *et al.*, 2016). To explore courses interdependencies and their corresponding relationships in university programs, a visualization tool known as EduVis was developed (Jordão, Gonçalves & Gama, 2014).

Similarly, discrete graphs have been used to visualize learning logs, a method that collected data from e-text and learning management systems. The graphs assisted in observing learning activities features for each grade by visualizing combinations of achievements and failures, thus helping to discover the learning activities that should be avoided by students (Okubo, Shimada, Yin & Ogata, 2015). A web platform named SWATShare, designed to enable collaborative hydrology research and education online, makes use of visualizations such as dynamic time-series plots and spatial maps, to help users understand the mutual variability of various hydrological processes through time (Rajib *et al.*, 2016). Another study explains that education can greatly benefit by integrating teaching strategies with interactive visualization techniques and moreover insists on enhancing visualization literacy to children in early grades (Alper, Riche, Chevalier, Boy & Sezgin, 2017). It further points out that, even though visualizations have recently become common, the ability of youths and adults to interpret them is still relatively low, which can be a serious handicap to learning and making informed decisions.

In education then, data visualization has been shown to play a key part in identifying and analyzing key causes and their variations in various issues. This has also been found to assist decision makers to enforce more effective schemes in teaching, monitoring learner’s behavior and improve learner’s understanding and learning outcomes. Moreover, the usefulness and success of interactive visualizations in online learning environments suggests that it could also be useful in non-online educational environment, which is the target for this study.

2.3 Visualization in Other Domains

In health, a work which investigated the prevalence of visualization for electronic health records data from 1996 to 2013 found out that, although electronic health records are increasingly being visualized, not many techniques have been found to display the complex data in these records effectively and efficiently (West, Borland & Hammond, 2014). Another study developed a visualization plugin that equipped the translational research open data warehouse with visual analytical workflows which are dynamic, by using modern web technologies such as AngularJS and D3 (Herzinger *et al.*, 2017). In health domain, visualizing connectomics which have massive datasets of brain tissue details made it easier and convenient for neuroscientists to focus on the insights provided by connectomes data, rather than management of these massive datasets (Haehn *et al.*, 2017). These dynamic and interactive data visualizations provide easier access to data and supports post-processing exploration for easier hypothesis generation.

Furthermore, a web application called NGL Viewer was developed for visualization of macromolecular structures. Embracing the modern web technologies, the viewer can assist life scientists to easily access 3D structural data (Rose & Hildebrand, 2015). A direct-manipulation approach for non-programmers to understand neural networks has been illustrated through an interactive open source visualization called TensorFlow Playground. This visualization intuitively gives its users a hands-on feel of how neural nets work without any coding (Smilkov, Carter, Sculley, Viégas & Wattenberg, 2017). In medical physics, a 3D visualization technique named Cinematic Rendering has been used to show photorealistic representation of 3D images from traditional Computed Tomography (CT) and Magnetic Resonance Imaging (MRI) data (Eid *et al.*, 2017). All these works bring to light the data visualization tools in other domains and the technologies employed to create them, whereby some of the technologies could also be applied in this study.

2.4 Web-based Visualizations in Tanzania

Recently, a number of visualizations and interactive dashboards are seen in many web-based systems and portals. The National Bureau of Statistics has a data visualization section with interactive visualizations for Tanzania's statistics on population, economy, health, environment, agriculture and education. However, for the case of dropouts, only school dropout status of population above 5 years is shown in numbers but no visualization can be found (National Bureau of Statistics, 2019). The Tanzania National Health Portal of the Ministry of Health Community Development Gender Elderly and Children has a data statistics menu which contains interactive graphs, charts and maps for various health statistics (Tanzania National Health Portal, 2019). The Government of Tanzania and the World Bank have developed dashboards for three sectors; health, water and education, aiming to support reporting and decision making through interactive mechanisms. Similarly, visualizations for school dropouts are missing (Basic Statistics Tanzania Dashboard, 2019). The United Nations Development Programme (UNDP) in Tanzania has a crisis risk dashboard with visually compelling graphs and images, aiming to establish a viable process for risk identification and tracking on an ongoing basis. The tool is further useful for conflict analysis, regional trends and profiles reporting and efficient design of programs (UNDP, 2018).

2.5 Visualization Challenges

The most challenging task for visualization designers is the means to perform abstractions rightly in order to facilitate efficient algorithm development (Childs *et al.*, 2013) and how to abstract tasks in a specific domain to create visual encodings with implications that are far-reaching and more generalizable (Borkin, 2014). Furthermore, it is challenging to develop visualizations that can scale to multiple platforms and guarantee real-time performance (Haehn *et al.*, 2017). While it is normally challenging to integrate and analyze heterogeneous data, the main challenge in the visualization of big data is the high dimensionality and its large size. Moreover, visualization tools for big data have performed poorly in scalability, response time and functionality (Chen & Zhang, 2014). Success in solving these challenges will enable researchers transfer findings beyond a single domain, thus contributing to the visualization research community (Brehmer, 2016).

2.6 Research Gap

Even though data visualization has been widely adopted, most visualization tools are domain specific and therefore inapplicable to other domains. Furthermore, visualization tools which have been developed in education domain are mostly for e-learning platforms and the MOOCs. Literature also shows that, most visualizations for education either do not present student dropouts data or presents only girls' dropout or only dropout in either primary schools alone or secondary schools alone. The little information given in the existing visualizations limits decisions required to be made based on data. This study then aims to develop a data visualization tool for student dropouts in Tanzania, whereby student dropouts data for both primary and secondary schools in the whole country will be presented to the council level. The study will employ interactive visualizations to unearth trends hidden in datasets of student dropouts for better reporting and support in student retention decision making.

CHAPTER THREE

MATERIALS AND METHODS

3.1 Study Area and Scope of the Research

The study targeted the education domain of the United Republic of Tanzania, specifically dropout-related datasets for primary and secondary schools. Both government and non-government schools in Tanzania mainland have been included in this study. The basic education levels were chosen for this study since they exhibit larger numbers of student dropouts than the other levels.

3.2 Data Collection Methods

3.2.1 Interview

During data collection, a short interview was conducted with the principal education officer and the head of education statistics at the Education Department of PO-RALG. The aim was to understand their procedure of data collection, analysis, visualization and publishing of student dropout statistics. The interview style was semi-structured so as to allow flexibility and freedom for more clarifications in a natural conversational manner (Mahat-Shamir, Neimeyer & Pitcho-Prelorentzos, 2019).

3.2.2 Secondary Data

Secondary data is increasingly being utilized for research due to advances in technology which lead to huge amounts of data being generated, collected and made accessible for researchers (Johnston, 2017). This study used secondary data gathered from the Government Basic Statistics Portal and the PO-RALG. The datasets gathered were comma delimited (.csv) files and excel (.xls) files. Some data were provided as PDF files which were converted to csv files in order to make them in a format that can be further processed.

3.3 Data Preprocessing and Analysis Methods

Python was used to do pre-processing of data, which involved exploration, cleaning, merging and feature selection. The originally collected data was first loaded to Jupyter notebook and explored using various Pandas dataframe exploration functions such as summary statistics description and missing value checks. Some tuples had few missing values, to which case the missing values were replaced with zeros while some tuples had all missing values, to which case they were dropped; these for instance included those named 'Pregnancy-Male'. Figure 1 shows the missing value check done for primary school region-level dataset.

```
In [17]: #check for missing values
         primary_rrs_16.isnull().sum()

Out[17]: region                0
         Death-Male            0
         Death-Female          0
         Death-Total           0
         Indiscipline-Male     0
         Indiscipline-Female   0
         Indiscipline-Total    0
         Pregnancy             0
         Pregnancy-Total       0
         Truancy-Male          0
         Truancy-Female        0
         Truancy-Total         0
         Total-Male            0
         Total-Female          0
         Grand_Total           0
         dtype: int64
```

Figure 1: Checking Missing Values in the Region Level Dataset for Primary Schools

After the initial exploration of the data and checking for missing values, there was a need to merge the different datasets for different years but which had similar features. It was important to do so since the initially collected datasets were not combined; for each year, there was a separate dataset in both region-level data and council level data. There was also a need to introduce new features such as the category of the data, whether they were in primary school or secondary school category; and the year of the data. This was necessary so as to enable development of visualizations with various user options such as drop-down menus for selecting a year and proper naming of the graph titles by automatically showing the year for each visualization.

To learn more about the newly created datasets, various visualizations including choropleth maps and heatmaps were created using python libraries such as bokeh and seaborn. This helped in getting a quick understanding of the data and identifying anomalies in the datasets before finalizing them to be fed into the visualization tool. Figure 2 presents the data preprocessing pipeline in the context of this work.

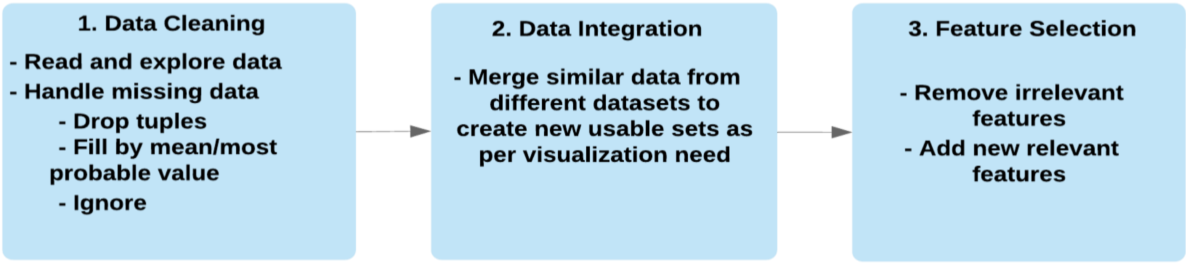


Figure 2: The Data Preprocessing Flow

3.4 Architectural Design

3.4.1 Visualization Pipeline

Visualization pipeline consists of a number of stages from data preprocessing to the creation of the graphical representations of the data. Figure 3 illustrates the overall process followed in this study, which is modified from Heer (2005) and Tamayo *et al.* (2018). Data transformations included data analysis, cleaning and feature selection from the relevant datasets. The clean data is then changed into visual form by mapping the data elements to corresponding visual abstractions or graphical displays. The last part of the process allows the user to view visualizations interactively through the user interface. The whole process is repetitive. The pipeline structure was chosen and modified to include various elements from different frameworks that have been proposed by various researchers, so as to come up with a simple and comprehensive one suitable for this study.

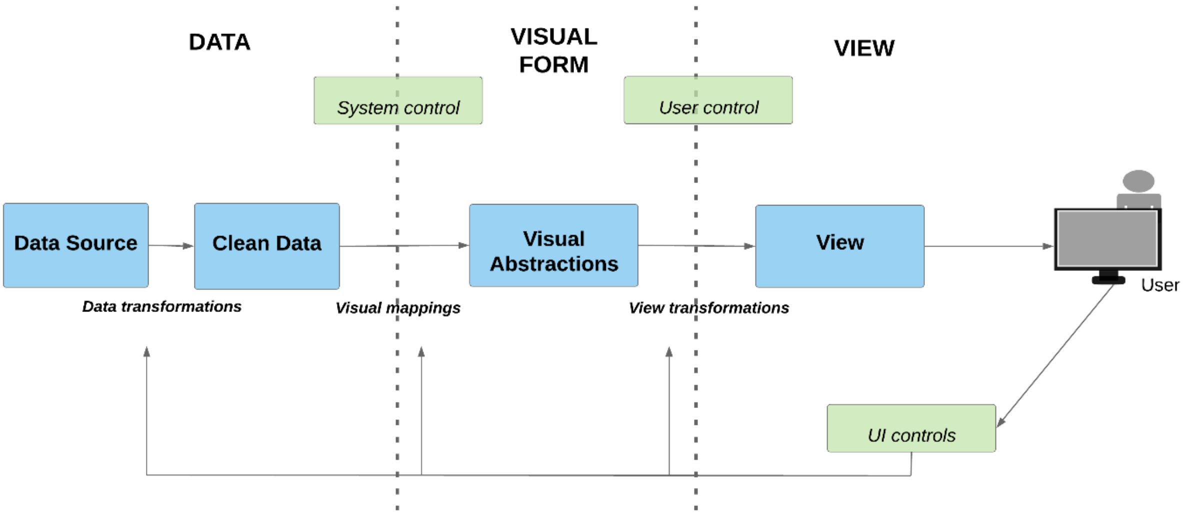


Figure 3: Visualization Pipeline

3.4.2 Conceptual Framework

The web based data visualization tool will display data in different graphical representations such as maps and graphs/charts. The users of the tool will include decision makers in education domain such as education officers, education statistics experts, heads of schools, teachers, parents, students, researchers in education and any citizen interested in education information. Through a web browser, the users will be able to access and view various visualizations using various menus provided in the user interface. Users can further upload data and view the resulting visualizations. Figure 4 shows the conceptual framework for the proposed tool.

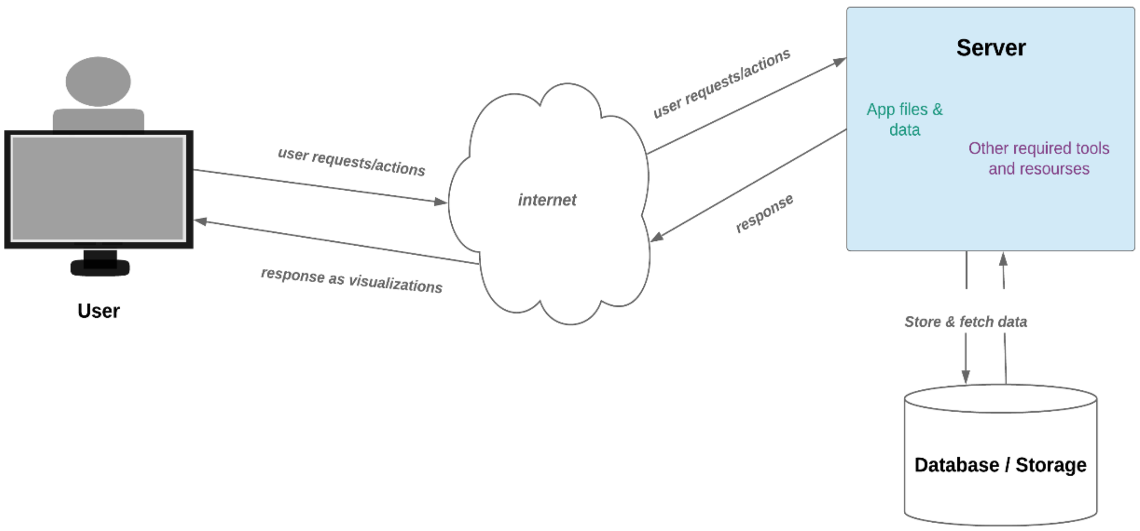


Figure 4: Conceptual Framework for the Proposed Tool

3.4.3 Use Case Diagram

A use case diagram in Fig. 5 indicates user interactions in the tool. Use case diagramming is one of the fundamental techniques in Unified Modelling Language (UML) used to summarize a set of use cases for a given system (Dennis, Wixom & Roth, 2012). In the shown figure, the user actions which precede the main functionalities such as selecting a category and year before viewing a visualization are depicted as include candidates for the main action. A description of each use case as the user interacts with the tool is defined in Table 1.

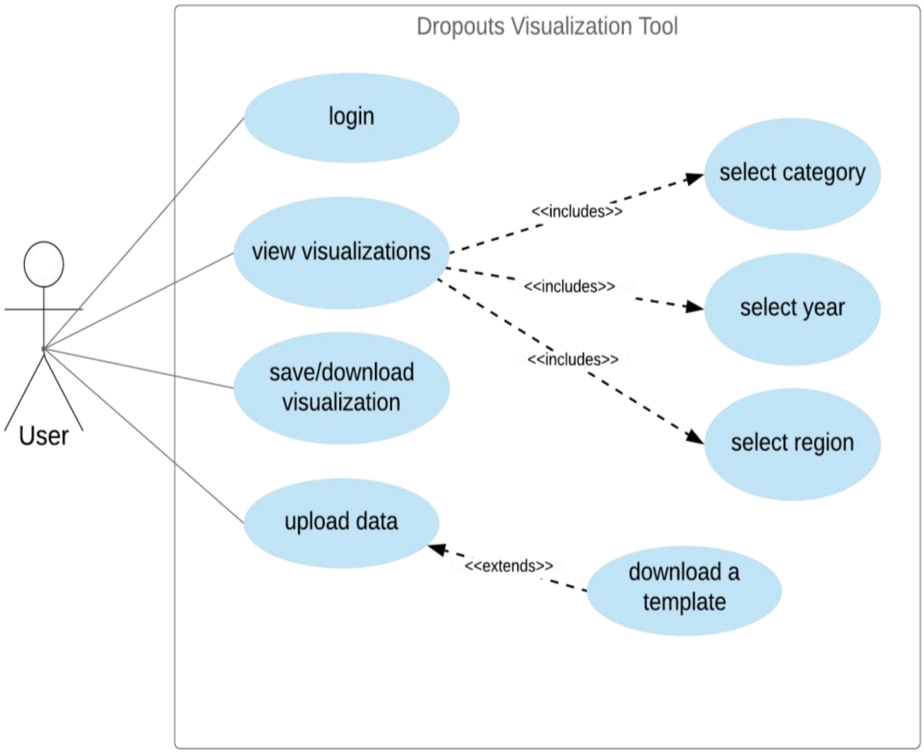


Figure 5: Use Case Diagram

Table 1: Description of the Use Cases

Use case	Description
View visualizations	A user will be able to view general visualizations just after accessing the link to the visualization tool. For viewing visualizations, a user will not need to log in.
Select category	To view specific visualizations, a user will need first to select a category, either primary school or secondary school
Select year	To tool will allow the user to select a year after selecting a category and view visualization for the selected year
Select region	The user will be able to view visualizations for district/council after selecting a region
Save/download visualizations	The tool will provide an option of saving/downloading a visualization in .pdf, .jpeg or .svg format.
Login	A user wanting to upload data to the tool will be required to login
Upload data	A user will be able to upload data after formatting them as in the template provided in the tool

3.5 Development Approach

A number of stages are involved in the software development life cycle (SDLC), including but not limited to planning, analysis, design and implementation. Different approaches have been applied in realizing these stages including waterfall approach, agile development, rapid application development; utilizing techniques like parallel development, prototyping and iterative development (Dennis *et al.*, 2012). This study employed rapid prototyping, where by an initial prototype was developed, then evaluated and continuously improved through a number of cycles or iterations until the final prototype was completed. Rapid prototyping was selected because of its flexibility and ability to allow high levels of user involvement, thus offering high guarantee of success for software project (Kendall & Kendall, 2011). Figure 6 presents the development stages followed in this study and Table 2 shows a comparison of prototyping and other software development techniques.

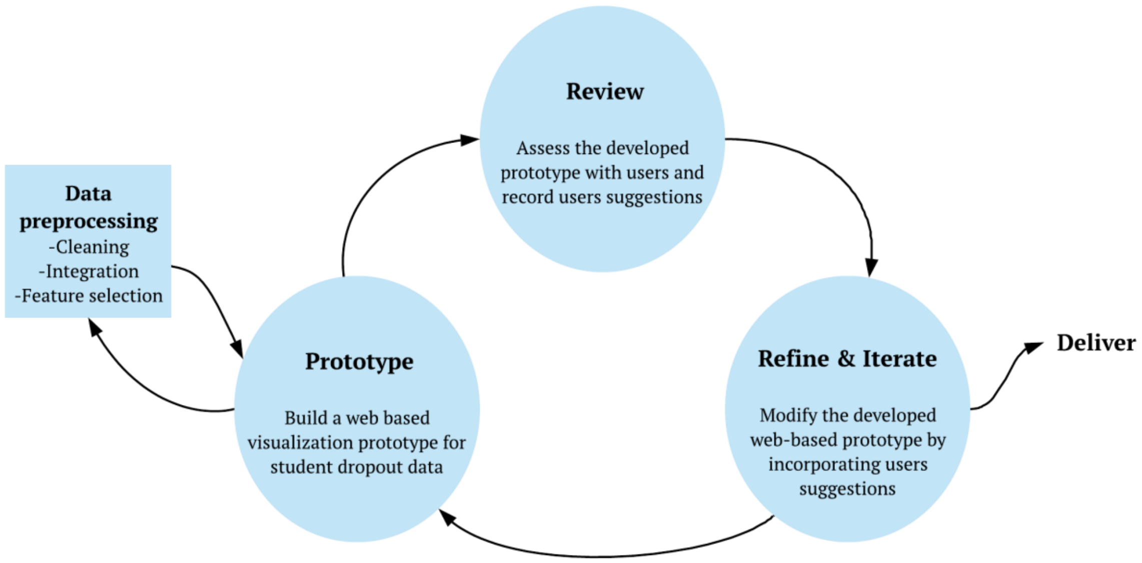


Figure 6: Development Stages for Visualization Tool

Table 2: Comparison of Prototyping and other Software Development Methodologies

Method	Knowledge of requirements	User/customer involvement	Flexibility	Guarantee of success	Project risk
Waterfall	High	Low	Low	High, for non-changing requirements	High
Iterative development	High	Medium	High	High	Low - medium
Prototyping	Low - medium	High	High	High	Low

Aziz (2012)

3.5.1 The Development Architecture for the Proposed Tool

The architecture presented in Fig. 7 is made up of two main parts; the front end, and the back end part. The front end carries the development tools that runs the web application dynamically in the user's browser. The back end consists of modules and tools required to run the full business logic of the web tool, provide the front end with access to resources that it needs, including networking protocols such as HyperText Transfer Protocol (HTTP), the database and the necessary utilities to manage the reading and writing of data thereof.

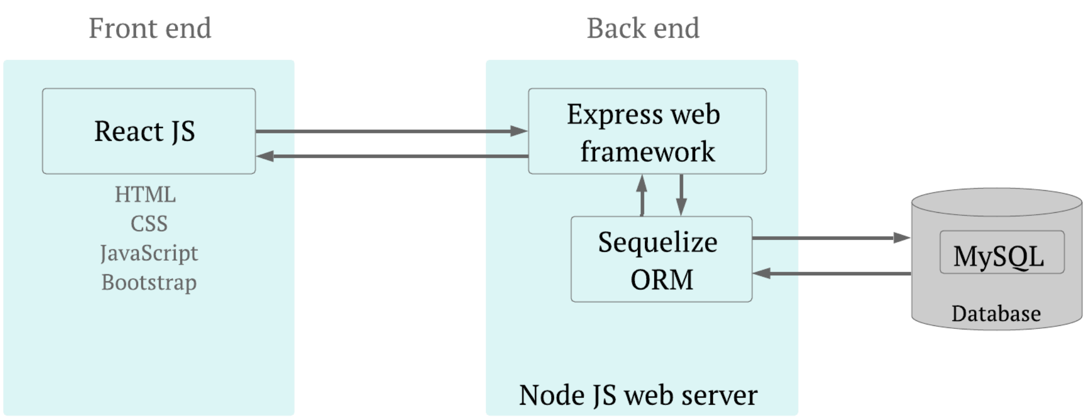


Figure 7: Development Architecture for the Proposed Web Based Visualization Tool

3.5.2 Technologies and Tools

The programming language used for the development of the web based data visualization tool was JavaScript. It was chosen in order to match with the interactive and dynamic nature of the visualizations. Other tools and technologies employed are as explained below:

(i) Node.js

A runtime environment built on Chrome's V8 engine, used to develop full-stack JavaScript applications. It was used to host the web pages and JavaScript libraries. Node.js was chosen because, apart from being JavaScript-based, its single threaded, asynchronous, non-blocking and event-driven nature makes it suitable for data-intensive applications. Since it has a HTTP server library built in, no separate webserver program is required (Bangare, Gupta, Dalal & Inamdar, 2016). The Node Package Manager (npm) which is installed when installing Node.js, was used for package installation and managing dependencies. Yarn as an alternative to npm, was also used as a package installer and dependency manager**.**

(ii) React.js

A JavaScript library which was used for building the user interface. Created by Facebook for large-scale, data-driven websites, it was chosen because of its simplicity, scalability, its ability to create reusable components and allowing an application to alter data with no need of reloading a page. Together with React, other JavaScript libraries such as D3.js and Highmaps were used for graphs, maps design and querying data. These libraries are all JavaScript based and suitable for creation of interactive visualizations.

(iii) MySQL

A structural database used in the backend to capture and store users' data. MySQL was suitable for the development of the web-based visualization tool because of its structured nature which would fit the users' data, high reliability, availability and data consistency qualities. Sequelize, was used as an Object-Relational Mapping (ORM) tool due to its ability to provide easy access to MySQL database and handling it as a JavaScript object and method.

(iv) Express.js

A fast, minimal, web application framework and a middleware for Node.js used for handling API's. With Express.js, developer does not need to worry on small technical details such as managing application routes, requests and views but concentrate on the application functionalities. Created for Node.js, Express.js was suitable since Node.js was used as the runtime environment during development.

(v) CSV to JSON converter

An online tool used to change the preprocessed data to JSON format, which were further formatted and saved as JavaScript files inside the project directory. This allowed the data to be imported to the application components easily. By expressing data as objects in name/value pairs, JSON provides a format of lightweight data exchange which is easy for humans to understand and for machines to break down and generate.

3.6 Tool Testing Approach

3.6.1 Component and System Testing

Verification and evaluation of the capabilities of a software product is an important task in attaining a well-functioning software. Different strategies are employed in testing which include unit or component testing, integration testing, system testing and acceptance testing; utilizing methodologies such as black box, white box and grey box testing (Gaur, Goyal, Choudhury & Sabitha, 2016). During the development phase of this study, unit testing was done to find and fix errors within each component. Moreover the whole tool was tested as a complete system to check if it performs the overall functionality as a whole.

3.6.2 Acceptance Testing

User acceptance testing is intended to check whether a software satisfies the needs of the intended audience (Otaduy & Díaz, 2017). Focus group discussions are among useful methods in user acceptance testing for finding out what users think of developed systems, in terms of benefits, drawbacks and other input for improvements (Aittoniemi, Penttinen, Rämä & Rech, 2014). This study used focus group discussions to solicit feedback from thirteen (13) chosen participants regarding the usefulness of the tool, the ease of use and what suggestions for improvements. Additionally, questionnaires were administered to supplement on the feedback from the focus group discussion. Questionnaires were used due to their ability to collect much information in a cost effective manner, while encouraging anonymity and thus reducing bias in answers (Marshall, 2005). Forty one (41) responses were collected from questionnaires, the number being slightly higher than the default recommended sample size (Perneger, Courvoisier, Hudelson & Gayet-Ageron, 2015)

3.7 Operating Environment

The developed prototype being web-based required the following environment for its proper operation:

(i) Server computer providing the hardware, operating system, programming language runtime and framework libraries on which web based tool will run; and a web server.

(ii) Internet connection

(iii) A web browser

CHAPTER FOUR

RESULTS AND DISCUSSION

4.1 Data Collection Results

From the short interview conducted during data collection, it was found that student dropouts information is collected via annual school census, through questionnaires sent and filled at school level. The information is then entered in the database at council level and exported to excel when analysis needs to be done. This study used secondary data from Tanzania mainland, whereby a total of 16 relevant datasets were collected described as follows: 9 datasets for region-level data including 5 for primary schools and 4 for secondary schools and six 6 datasets for council-level data including 4 for primary schools and two 2 secondary schools.

4.2 Data Preprocessing and Analysis Results

The originally collected data had more than 18 features. However, the preprocessing of data resulted into creation of five datasets with less features than they were initially present i.e. nine features. The created datasets are listed below:

(i) A primary school region-level dataset containing region-wise data for all the available years (2012 to 2017).

(ii) A primary school council-level dataset containing council-wise data for all the available years (2016 to 2017).

(iii) A secondary school region-level dataset containing region-wise data for all the available years (2012 to 2016).

(iv) A secondary school council-level dataset containing council-wise data for all the available years (2016 to 2017).

(v) A reasons dataset containing reasons for dropout for the years 2012 to 2017 for primary and secondary schools

The features retained after preprocessing generally comprised of region name, council name, year, enrollment, dropout, Pupil-Teacher Ratio (PTR), Pupil-Qualified Teacher Ratio (PQTR), Pupil-Classroom Ratio (PCR), Pupil-Latrine Ratio (PLR). These features were selected to enable making comparisons between dropouts and the listed indicators, aiming to give more meaningful insights than when dropouts are visualized alone. The dataset for dropout reasons included such reasons as death, truancy, indiscipline and pregnancy. The following sections further show the results from exploration and analysis of the preprocessed data.

4.2.1 Primary School Dataset Exploration and Analysis Results

A total of 396 505 pupils dropped out from school in the years 2012 to 2017, with the exception of 2014. This number comprises of 215 412 (54.44%) boys and 181 093 (45.56%) girls, with an average value of 1683 boys and 1415 girls annual dropouts in each region. Even though for the same years more girls (50.62%) were enrolled than boys (49.38%), an 8.88% more boys dropped out from primary schools than girls. The summary statistics from the created primary school dataset are shown in Table 3.

Table 3: Summary Statistics for Primary Schools Dropouts Data, 2012 to 2017 Except 2014

	Year	**Enrollment male**	**Enrollment female**	**Enrollment total**	**Dropout male**	**Dropout female**	**Dropout total**
count	128.0000	98.0000	98.000000	124.0000	128.0000	128.0000	128.0000
mean	2014.6328	16 7184.000	17 1350.4184	34 6875.1532	1682.9063	1414.7891	3097.7188
std	1.85 6553	6 8629.3904	7 0288.3247	14 0133.3535	1441.2960	1307.5806	2735.3608
min	2012.0000	4 6277.0000	4 6105.0000	9 2382.0000	48.0000	27.0000	75.0000
25%	2013.0000	11 9717.2500	12 4060.0000	24 6241.7500	751.0000	594.5000	1377.2500
50%	2015.0000	15 5575.5000	16 0425.0000	32 4467.0000	1339.5000	991.5000	2326.0000
75%	2016.0000	21 2936.2500	21 3244.5000	43 5511.5000	2137.5000	1791.5000	3920.2500
max	2017.0000	43 5165.0000	44 3394.0000	87 8559.0000	7872.0000	7425.0000	1 5297.0000

In Fig. 8, dropouts in primary schools for year 2017 have been presented in a map, revealing that more students dropped-out from central and north-western parts of Tanzania. This could be due to factors such as small-scale mining activities in the gold-mining areas of Geita (Ouma, 2017) and fishing activities for the lake regions. On the other hand, Fig. 9 shows that 2016 experienced higher number of dropouts in primary schools than all years between 2012 and 2017, with Geita, Tabora and Kagera leading in dropouts.

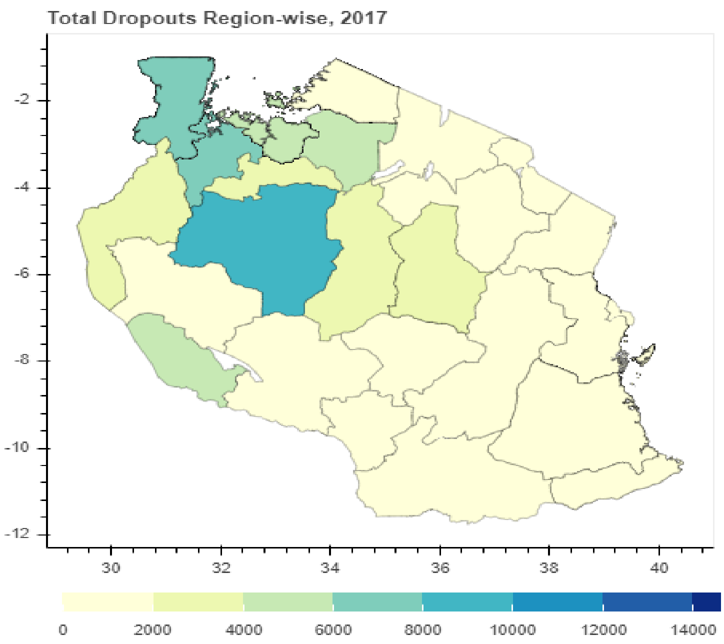


Figure 8: A Map showing Dropouts in Regions for Primary Schools, 2017

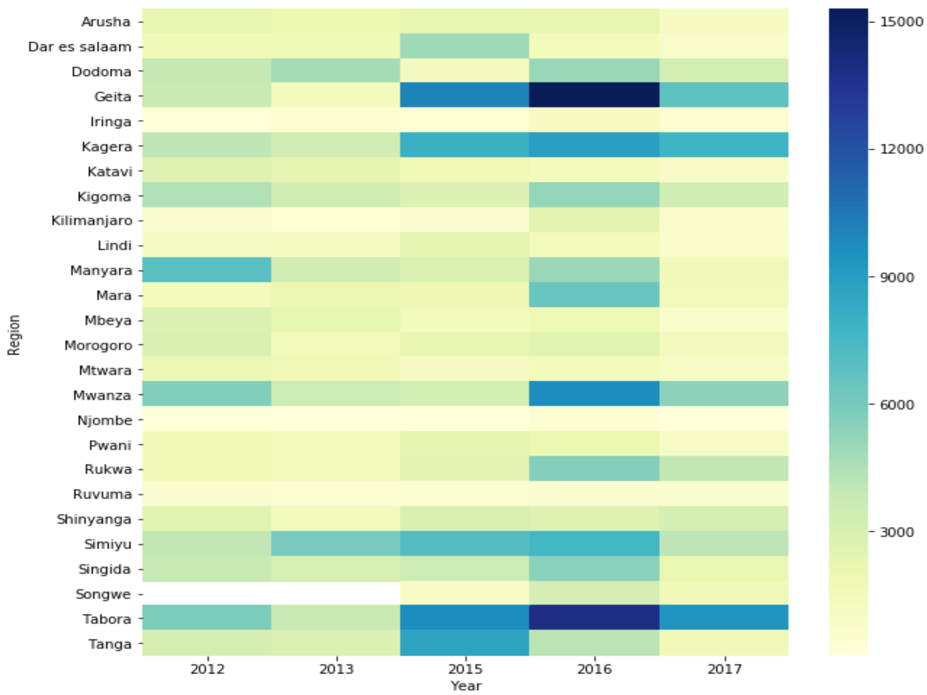


Figure 9: A Heatmap for Dropouts in Primary Schools in Each Region for Five Years

4.2.2 Secondary School Dataset Exploration and Analysis Results

A total of 294 986 students dropped out from secondary schools in the years 2012 to 2016, with the exception of 2014. This number is comprised of 159 764 (54.16%) boys and 135 222 (45.84%) girls, with an average value of 1582 boys and 1339 girls annual dropouts in each region. Table 4 shows the summary statistics from the exploration of the created secondary school dataset.

Table 4: Summary Statistics for Secondary School Regional Data, 2012 to 2016 Except 2014

	Year	Enrolment male	Enrolment female	Enrolment total	Dropout male	Dropout female	Dropout total
count	101.00000	100.0000	100.0000	101.0000	101.0000	101.0000	101.0000
mean	2014.0198	3 7552.5900	3 5298.9000	7 3127.1881	1581.8217	1338.8316	2920.6534
std	1.5936	2 0742.2188	2 1537.3042	4 1697.9996	783.8609	629.3098	1391.4188
min	2012.0000	6176.0000	4461.0000	1 0745.0000	325.0000	295.0000	620.0000
25%	2013.0000	2 3399.2500	2 1246.0000	4 5636.0000	1052.0000	892.0000	1923.0000
50%	2015.0000	3 1756.5000	2 8985.5000	5 9567.0000	1416.0000	1218.0000	2642.0000
75%	2016.0000	4 8838.7500	4 3559.7500	9 1998.0000	1904.0000	1639.0000	3564.0000
max	2016.0000	9 6692.0000	9 7524.0000	19 4216.0000	4746.0000	3502.0000	8248.0000

Table 4 also shows that, contrary to primary schools enrollment, about 3.1% more boys were enrolled for secondary school education than girls. Moreover, similar to primary schools, about 8.32% more boys dropped out than girls. Additionally, a choropleth map in Fig. 10, present secondary school data for 2013 showing Mwanza and Mbeya as leading regions experiencing student dropouts.

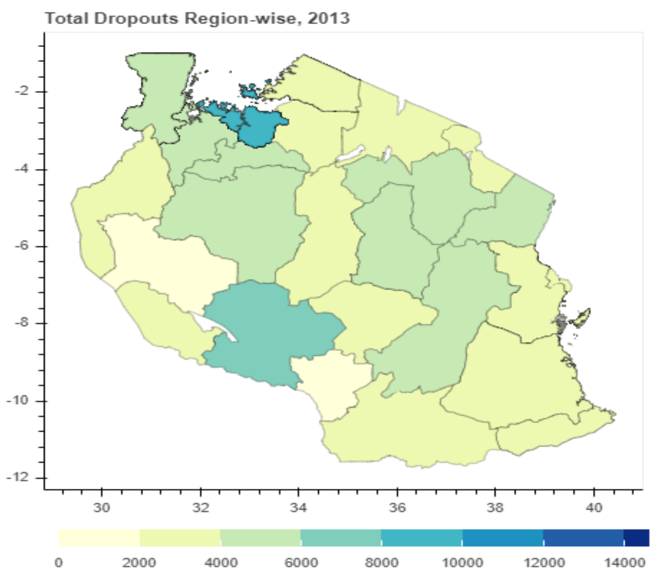


Figure 10: A Map showing Dropouts in Regions for Secondary Schools, 2013

Figure 11 shows that for secondary schools, 2013 had the highest number of dropouts in most regions, with Mwanza, Mbeya, Tanga and Tabora leading in dropouts. It also shows that, Mwanza has experienced a consistently high number of dropouts than all other regions for all years. Songwe region has missing data for years 2012, 2013 and 2015 since it was a council of Mbeya region until 2016 when it was instituted as a region.

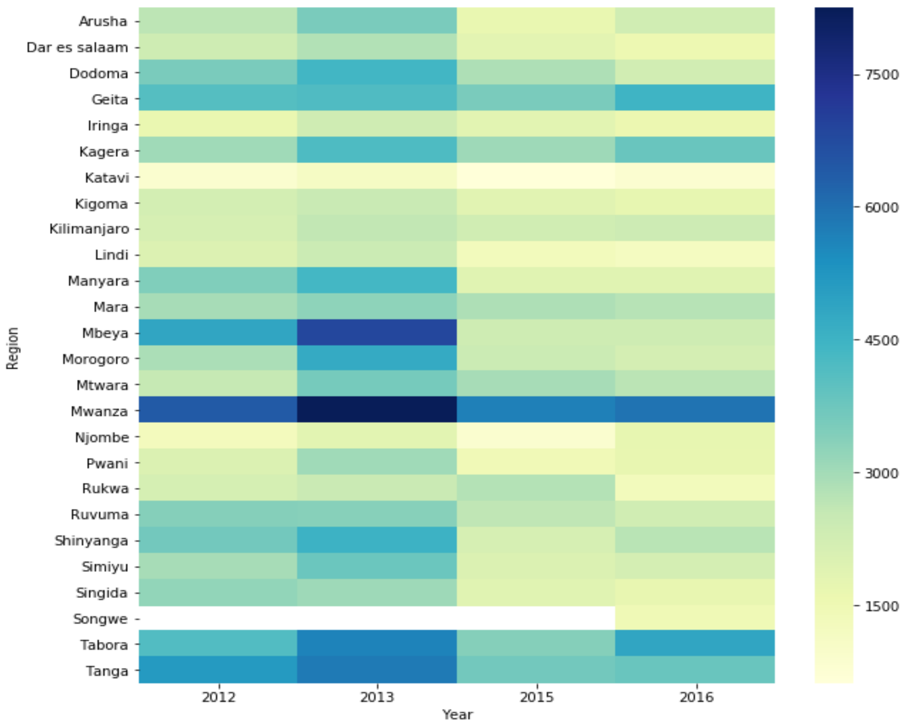


Figure 11: A Heatmap for Dropouts in Secondary Schools in Each Region for Four Years

Generally, the present data show higher numbers of dropouts for boys compared to girls, which may be connected to larger numbers of enrolment for boys than girls in secondary schools but still remaining a puzzle for primary school, since enrolment rates are higher for girls than boys. Moreover, truancy was found to be the leading known reason of student dropouts, which is caused by various factors including lack of basic needs; livestock keeping; works such as industrial, farm, mining, fishing and domestic; peer pressure; parents separation; illness; unfriendly learning environment; and other unknown reasons.

4.3 The Developed Web-based Data Visualization Tool

After the preparation and exploration of data was complete, a web based data visualization tool was developed for further presentation and exploration of the data. The first page of the developed web tool displays visualizations for student dropout statistics for the years 2012 to 2017 in maps and donut charts as shown in Fig. 12. From this page a user can navigate through the tool using the tabs provided at the top of the page, enabling to select the category in which he/she wants to see the visualizations. The tabs lead the user into four main menus in the tool which are Home, Secondary Schools, Primary Schools and Upload data. The following sections provide a description of the developed prototype.

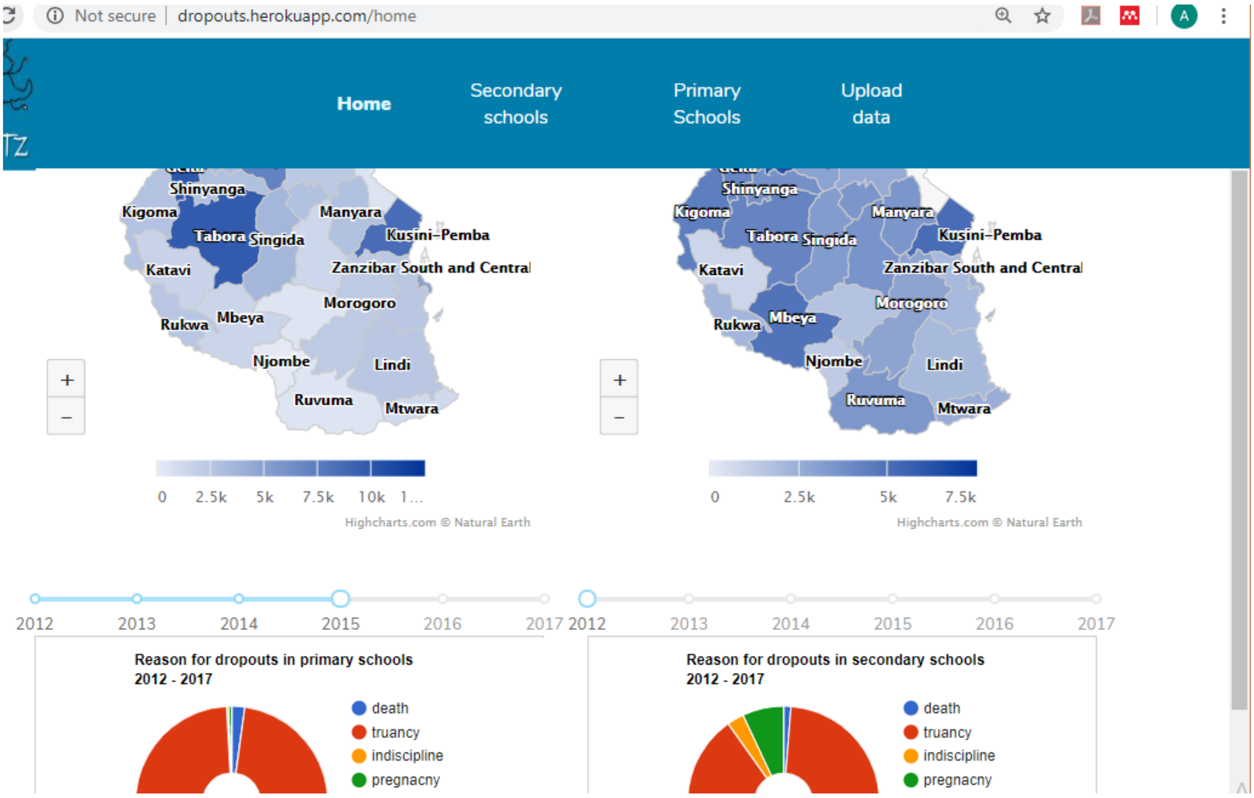


Figure 12: The Home Page of the Developed Tool

4.3.1 Visualizations for Primary and Secondary Schools Dropouts

The developed prototype visually presents both primary school and secondary schools data. Depending on user's selection, various charts are displayed showing dropouts in regions and districts. Grouped bar charts have been used to visualize dropouts in region by gender. Figure 13 shows a bar graph which appears after a user selects 'Secondary schools' category and year 2015 in the drop down menu.

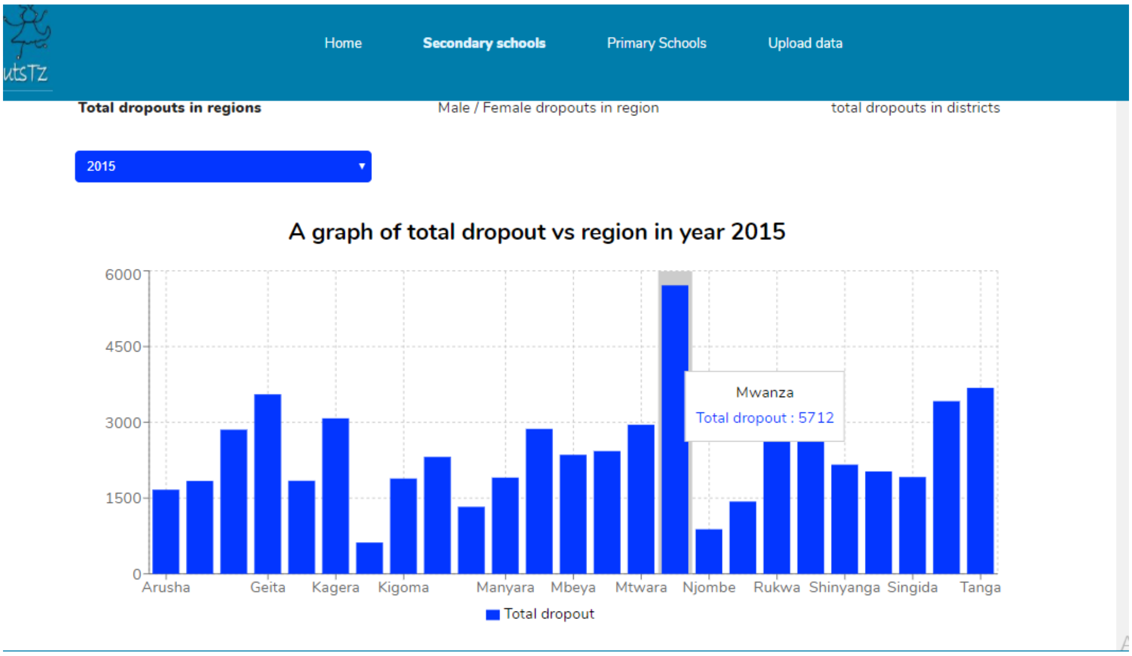


Figure 13: Total Number of Dropouts in all Regions for Year 2015. The Name and Exact Figure for each Region is shown when a User hovers over a Bar Representing that Region.

Figure 14 shows a grouped bar chart showing dropouts by gender in different regions. In this case, the user has selected the primary school category and year 2016 from the dropdown menu. Exact values are displayed when a specific bar is hovered over as shown in the chart.

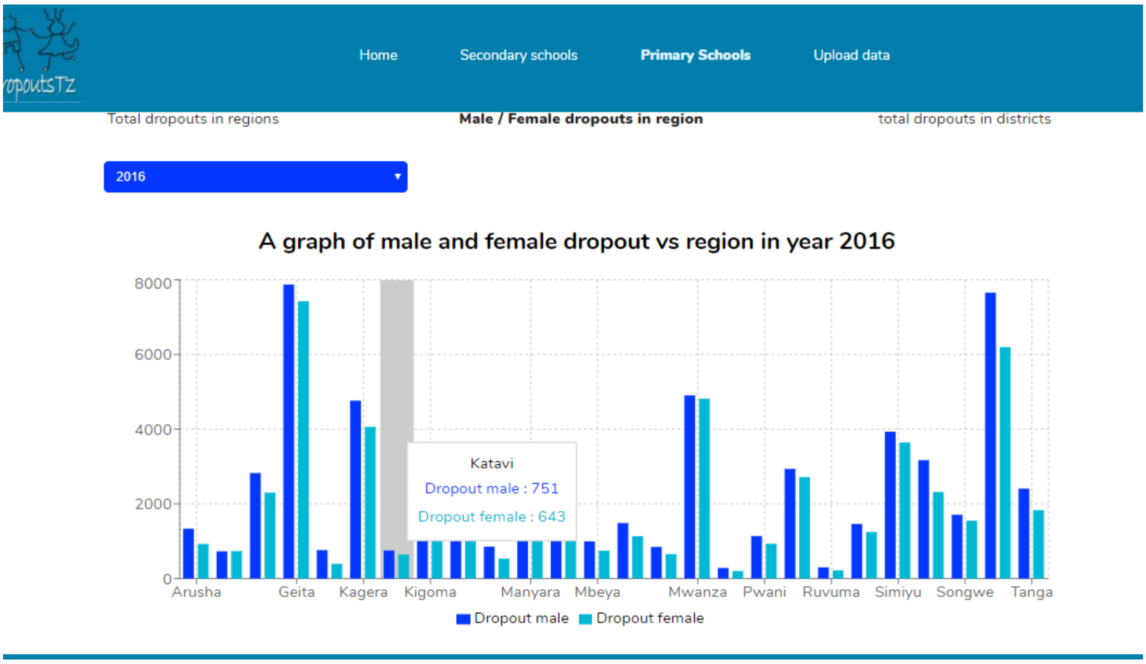


Figure 14: Dropouts by Gender in Different Regions

Figure 15 displays the dropout in districts/councils in relation to enrollment in a bubble chart. This happens after selecting a region and year from the dropdown menus and clicking the refresh button. In the figure, the dropout rate for each council is displayed in a bar that appears above the bubble chart. Moreover, hovering over a specific bubble displays more information about dropout and enrollment in a district.

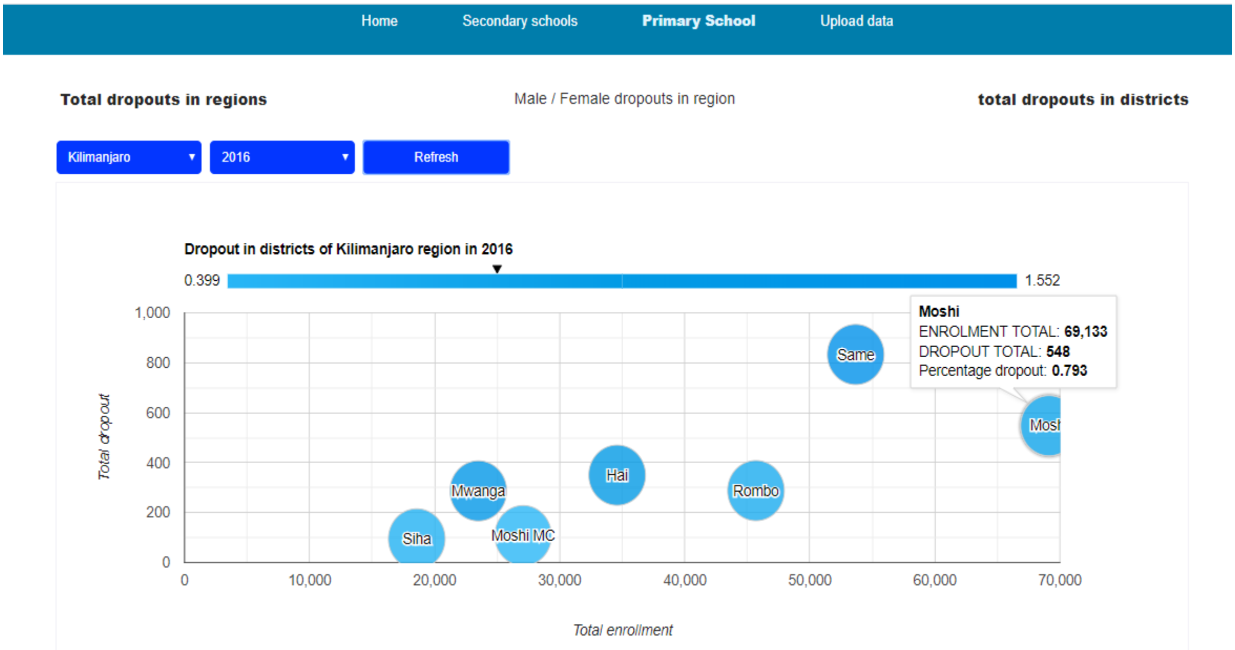


Figure 15: Dropout in Districts/Councils in Relation to Enrollment

4.3.2 Uploading Data

To upload data, a user is required to log in by providing a username and password, otherwise, the application is available for any user wanting to just see the visualizations. To successfully login, a user is firstly required to register an account using the sign-up menu in the login interface. Figure 16 shows the login interface.

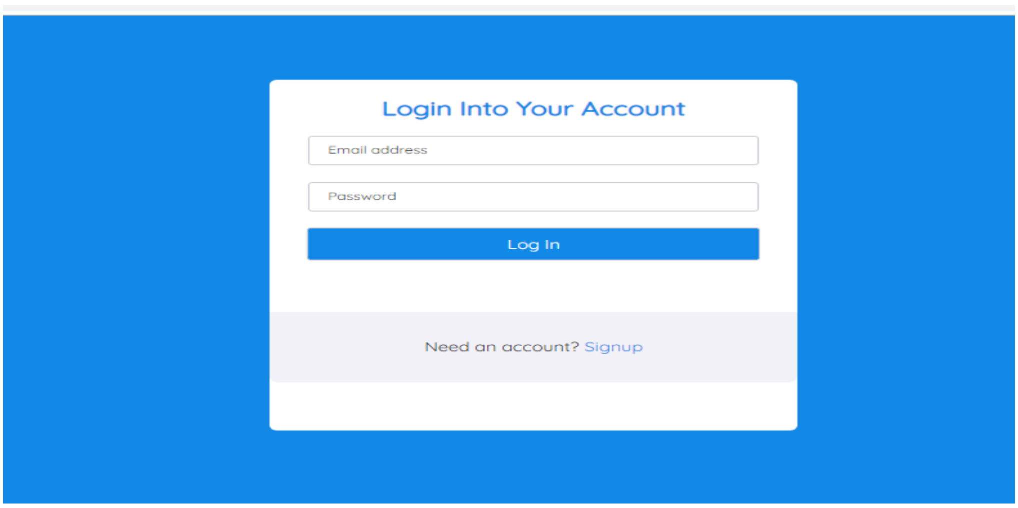


Figure 16: The Login Interface

After a successful login, the upload data section is allows a user to download a template from the tool. The provided template is purposed to guide users in structuring their data into a proper format before uploading them. The template is a csv file with the required column headings for the data to be uploaded. Figure 17 shows the upload data page and Fig. 18 shows the template.

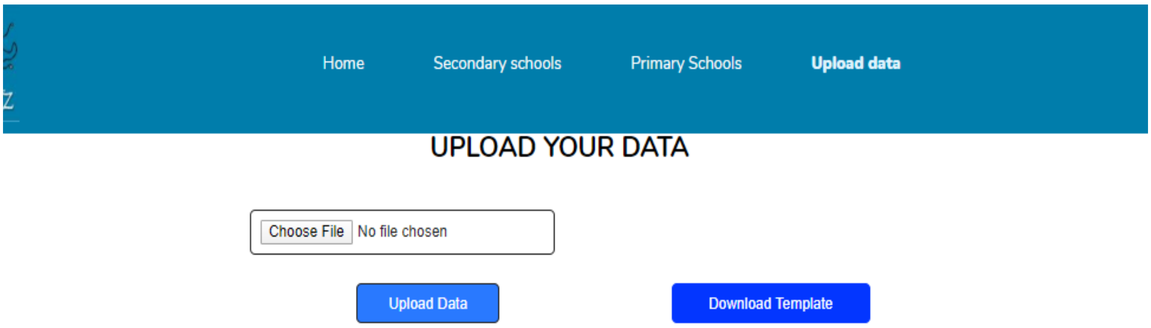


Figure 17: The Upload Data Section

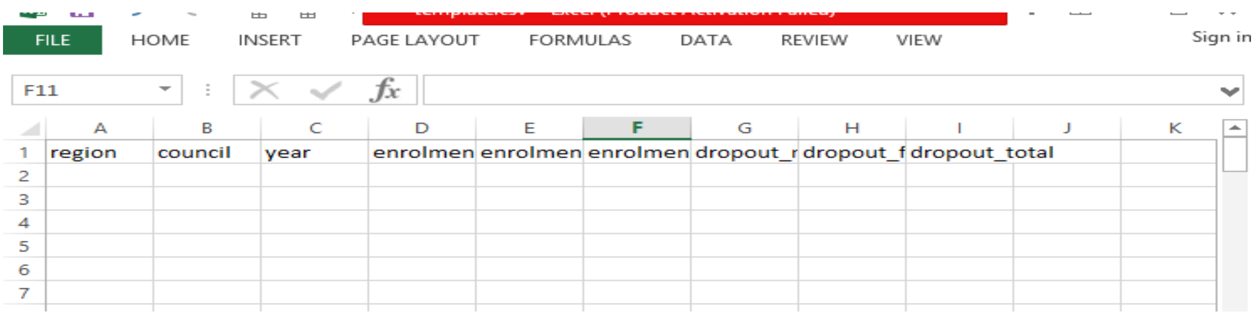


Figure 18: The Provided Template

After filling and formatting data according to the provided template, users can upload their data. The tool allows a user to choose a file and upload it for visualization by clicking the choose file button, selecting the file from his/her device and then click the upload data button. After successful upload, the user is given an option of selecting a graph type and the year as in Fig. 19 and Fig. 20.

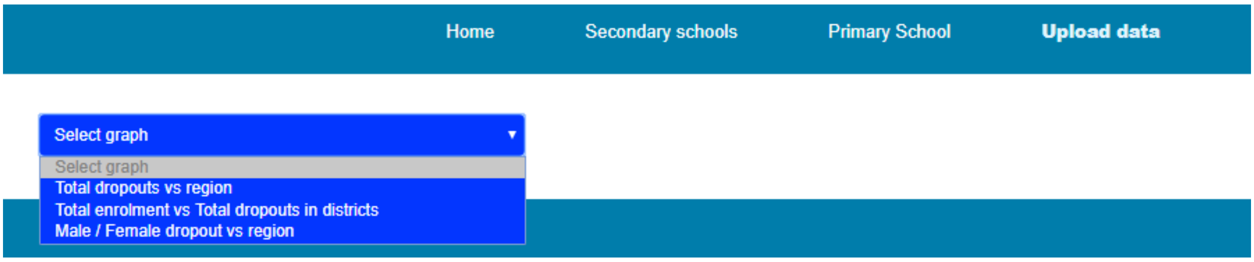


Figure 19: The Dropdown Menu for Selecting the Chart Type which Appears after User Uploads a File of Data in the Tool.

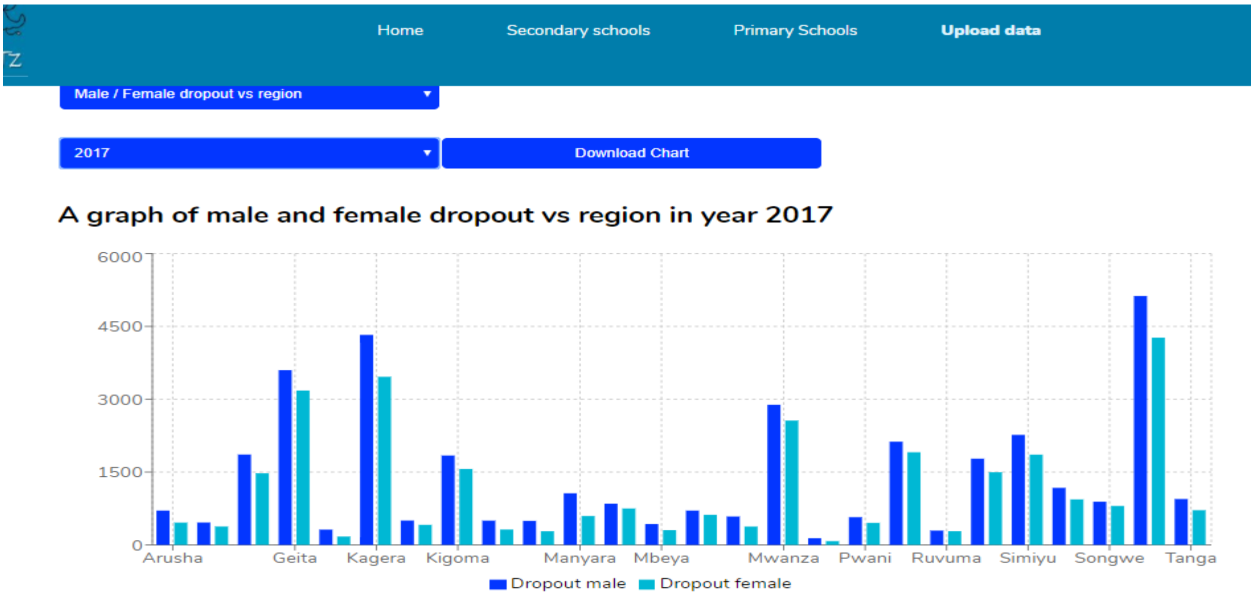


Figure 20: The Grouped Bar Chart showing Dropouts by Sex in 2017. This Chart Appears after the User Selected Male/Female Dropout vs Region as the Type of Graph and 2017 as the Year.

4.4 Component and System Testing Results

The components of the developed tool were tested and found to be functioning properly. Moreover each component performed well within the overall developed web application. The results of the tests are summarized in Table 5.

Table 5: Results for System Testing

Function	Description	Status
View visualizations	User wanting to only see visualizations can view them without being required to log in.	Pass
Select category	User can successfully select a category, either primary school or secondary school and view visualizations of the selected category.	Pass
Select year	Users can successfully select a year after selecting a category and view visualization for the selected year	Pass
Select region	User can successfully select a region and view visualizations for district/council	Pass
Save/download visualizations	User can save/download a visualization	Pass
Login	User can successfully login after registration	Pass
Upload data	Users are able to upload data after formatting them as in the template provided in the tool	Pass

4.5 Validation Testing Results

Results from a focus group discussion involving thirteen (13) participants and questionnaire responses from 41 respondents were collected. The study used R statistical software to analyze the collected data from the questionnaires. The following sections describe the results obtained from both cases.

4.5.1 Focus Group Discussion Results

During the focus group discussion, all participants pointed out that the data visualization tool is useful for getting quick impression of the student dropouts data. Majority liked the way data was presented in bar, pie and bubble charts except a few who pointed out that the bubble charts were difficult to understand since they were not as familiar as the other charts. The participants also liked the way the tool presented data in various levels for regions and councils, though a majority suggested that more features should be added to the tool, especially going deeper to school level and including other education data apart from student dropouts. Table 6 shows the participants of the focus group discussion.

Table 6: Description of the Focus Group Discussion Participants

Title	Count
Principal Education Officers (PEO)	2
Principal Statisticians	1
Heads of School	4
ICT Officers	4
Statisticians	2
Total	**13**

To gather more specific information on the developed tool, the discussion was led by a number of questions, which asked the participants whether they have used a similar tool before; which features they liked or disliked most; when and how they would use the tool; their suggestions for improvements and what they thought was the most important thing about the developed tool. A few participants stated that they have used a similar tool while majority replied that they have not. Some further commented that, even though they have not used a tool like this before, they think that it is very helpful in making student dropouts data understandable and that it was very informative.

On the feature preference, majority liked the way data was presented in bar, pie and bubble charts. One participant commented that "*The use of bubble charts helps to quickly compare the dropout rates among districts".* On the contrary another participant commented: "*Why bubble charts? They are difficult to understand and they need someone to be familiar with them, unlike the other charts such as pie charts with numbers and percentages". A* few others suggested the use of '*bubbles over maps*' instead of just bubble charts. This brings to light that, though data visualization has become inescapably important, there is a significant number of people who are unaware of most current and uncommon visualization techniques such as bubble charts.

On using the tool, one participant commented "*The tool is very promising in visualization and can further be extended to present data about other situations apart from dropouts and be included in every system*". Another noted that: *"It is very useful especially when researching on issues related to education".* A further comment was that *"It can be used when dealing with challenges in education"* and *"in planning and budgeting"* and *"to intervene dropouts and find a solution".*

For the improvement of the tool for reporting and decision making, the participants suggested that the tool incorporate other data, and present them down to school level. One participant noted that: *"The tool need to incorporate feedback to the actions taken, based on the data perceived in the presented visualizations".* The participants were also asked what was the most important thing about the visualization tool; where by apart from majority saying that the tool was simple and useful, one specifically said that: "*The most important issue that I have come up with is that, it is easy to make one get information or data easily if tools like this are made or discovered*".

4.5.2 Results from Questionnaires

The questionnaires were filled by 41 respondents on a five point Likert scale of strongly agree, agree, neutral, disagree and strongly disagree. To get the users' views on the tool, the respondents were asked how they found the developed tool easy to use; how useful was it for reporting and decision making in their line of work and other organizations in the government; whether they will recommend the tool to their colleagues and if they thought more features need to be added to or removed from the tool. Table 7 shows the description of the respondents. Furthermore, a summary of all responses is presented in Fig. 21.

Table 7: Respondents Description

Title	Count
Principal Education Officers (PEO)	2
Principal Statisticians	1
Heads of School	28
ICT Officers	4
Statisticians	2
Researchers	4
Total	**41**

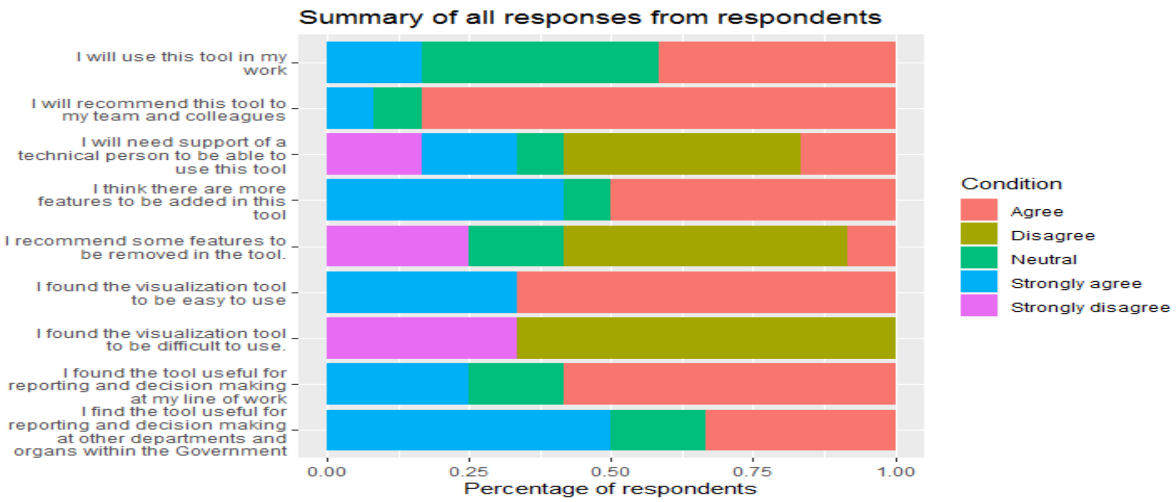


Figure 21: Percentage-wise Summary of all Responses from the Respondents

In using the tool, 20% strongly agreed that they will use the tool in their work, 37% agreed that thy will use it while the remaining 44% were neutral. On the ease of use, majority (83%) agreed that the web based tool was easy to use and the rest (17%) strongly agreed that the tool was easy to use. Figure 22 shows the responses for using the tool and Fig. 23 shows the responses for the ease of use, describing also the position (title) of the respondents in a stacked bar chart.

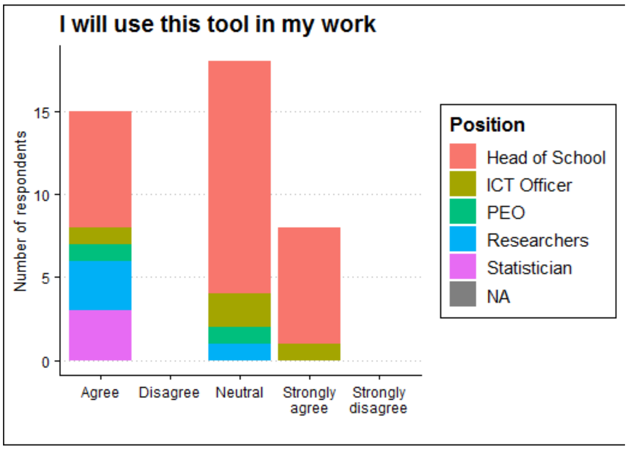


Figure 22: Responses on Using the Tool

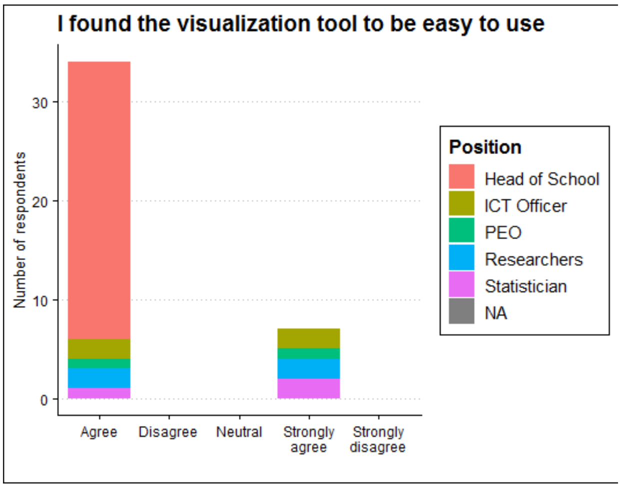


Figure 23: Responses on the Ease of Use

For the usefulness of the tool, 22% of respondents strongly agreed that the tool was useful at their' line of work, 56% agreed to it while the rest 22% were neutral. This discloses that even though interactive visualizations have less been used in student dropout issues in the context of non-online studying environment, they can yet be applied in this area to inform not only the decision makers but also the society in general. Figure 24 shows the user responses on the usefulness of the tool.

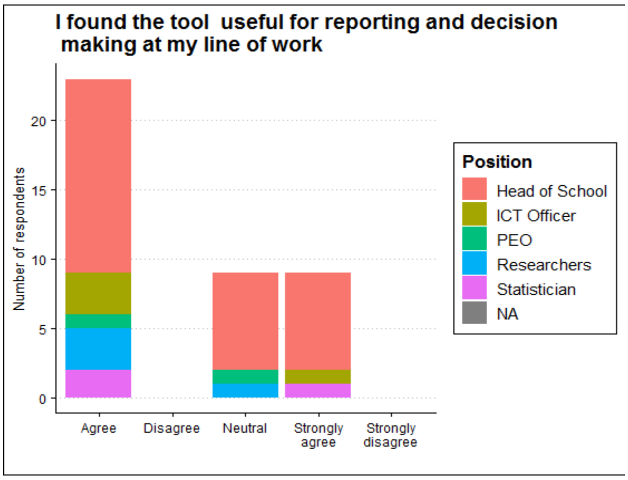


Figure 24: The Usefulness of the Tool at Respondents' Line of Work

Moreover, users were asked if they thought the developed tool would be useful at other organs or departments. Forty nine percent (49%) of the respondents strongly agreed to it while 32% agreed and 19% were neutral as in Fig. 25. The responses show that, visualizations play an important role for informing all domains of operation.

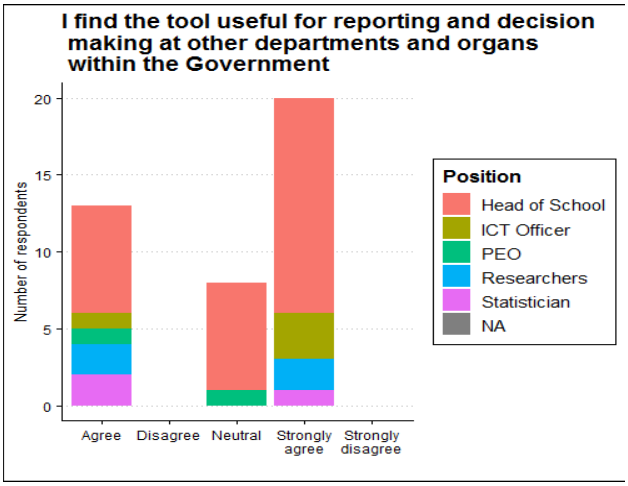


Figure 25: The Usefulness of the Tool at other Organs and Departments

In checking if the tool would be recommendable, 7% of respondents strongly agreed that they would recommend it to their colleagues, 88% agreed that it was recommendable and the remaining 5% were neutral. On feature addition, majority (59%) strongly agreed that more features be added to the tool and the remaining ones agreed by 39% while 2% were neutral. These responses are shown in Fig. 26 and Fig. 27.

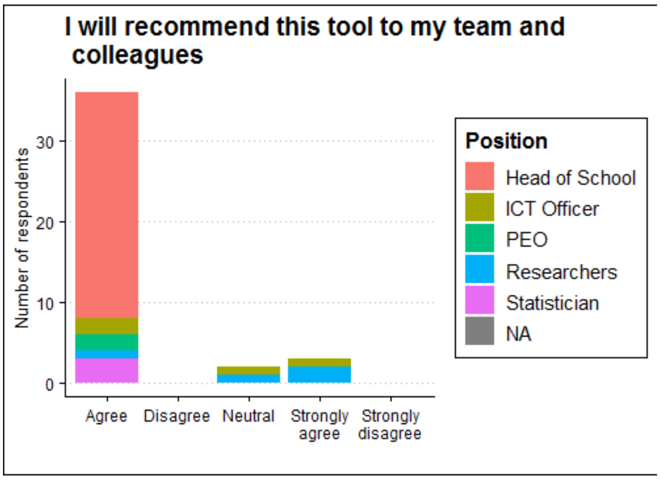


Figure 26: Responses on Recommending the Tool

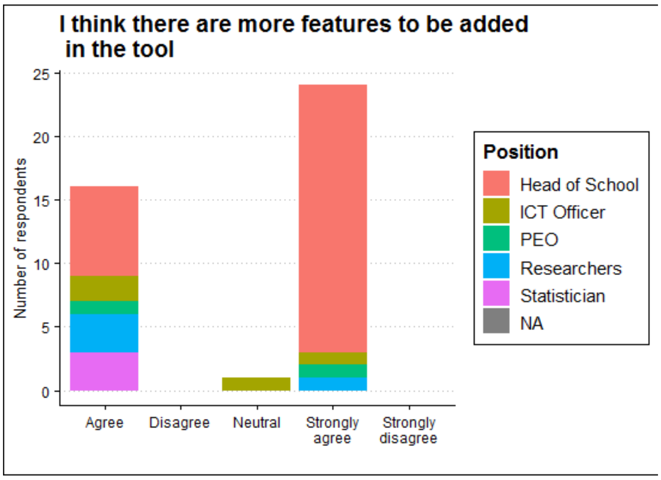


Figure 27: Responses on Features Addition

4.6 Further Discussion of the Results

The results from preprocessing and analysis of the collected data reveal that more girls are enrolled in primary school education than boys while more boys are enrolled in secondary schools than girls. They further show that, in both primary and secondary education, boys drop out more than girls, the main reason being truancy. The smaller number of girls in secondary schools could be attributed to lack of basic needs, early pregnancies, cultural prejudices and the boys' dropout could be attributed to truancy, death and also lack of basic needs (Tamisemi Regional Data, 2016). Moreover, the challenges encountered in data preprocessing due to disparity of data and lack of continuity in the existing datasets enlightens on the need for continuous and comprehensive data collection practices for purposes of enabling good visualizations which will accurately inform the respective domain.

The results from both focus group discussion and questionnaires confirm the usefulness of data visualization for data presentation and informing decisions not only in student dropout issues but also in other areas of administration. Moreover, the results apart from emphasizing on the importance of data visualization, also show that different users have different knowledge, understanding and preferences on various kinds of visualization techniques, and that visualization literacy is still low. Therefore, in order to exploit the benefits of data visualization, not only are data needed but also awareness of the users on the visualization techniques.

CHAPTER FIVE

CONCLUSION AND RECOMMENDATIONS

5.1 Conclusion

Visualization is an easy, understandable and an effective way of presenting large amounts of information. With the current technology and the massive amounts of data generated from everyday activities, data visualization has become crucial for informing decision makers in many areas of operation. The education domain, owning huge amounts of data is particularly one of those areas. This study sought to develop a web based data visualization tool for student dropouts in Tanzania, for better reporting and decision making. The developed tool, was found to be useful and necessary in presenting not only student dropouts information but also other education data. It was further agreed to be simple, user friendly and recommendable.

Furthermore, the study shows a possibility of interactively visualizing the currently published data on education issues in Tanzania, which mostly remain in pdf documents containing huge amounts of numeric data but only summary visualizations. Nevertheless, the tool is focused on student dropout data which is a very small subset of all education data and presents them only up to council level and do only annual reporting. In this light then, further improvement is required to include deeper and wider levels of data and users in order to inform all areas in education domain.

5.2 Recommendations

The following recommendations were put forth in order to improve this study:

Widening the scope of the study from Tanzania to regional and global levels in order to make the study applicable to common global challenges and also more informative. This could be done by including worldwide data about persistent global challenges in the education sector such as poor learning outcomes, equity issues, infrastructure and learning materials, quality of teaching and more so and thus make regional and worldwide comparisons. Moreover, it was recommended to include more variables in the data and provide future prediction of student dropouts. This will make the tool more useful for policy makers for early decision making and planning.

Direct linking of the developed tool with the existing PO-RALG data capturing system so that it captures and visualizes the already present data in many areas of education, thus easily present both student dropout data and other education data. It was also suggested to include the feedback from the actions taken after decisions based on the data presented in the tool have been made, and presenting data in shorter intervals such as in quarterly basis instead of only annually. This will also be useful for early decision making.

REFERENCES

Aittoniemi, E., Penttinen, M., Rämä, P., & Rech, J. (2014). Evaluating User Acceptance of C2X Systems by Focus Group Discussions. 10th ITS European Congress. ERTICO-ITS Europe.

Alexandre, D., & Tavares, J. (2010). Introduction of Human Perception in Visualization. *International Journal of Imaging*, *4*(10).

Allbest | PO-RALG. (2019). Retrieved November 12, 2019, from http://tamisemi.go.tz/allbest.

Alper, B., Riche, N. H., Chevalier, F., Boy, J., & Sezgin, M. (2017). Visualization Literacy at Elementary School. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 5485–5497. ACM.

Askinadze, A., Liebeck, M., & Conrad, S. (2019). Using Venn, Sankey, and UpSet Diagrams to Visualize Students' Study Progress Based on Exam Combinations. *9th International Conference on Learning Analytics and Knowledge,* 1–5. ACM.

Aziz, A. (2012). A brief Comparison of SDLC Models. Retrieved September 14, 2019, from https://asma-aziz.blogspot.com/search?q=A+brief+comparison+of+SDLC+Models.

Bangare, S., Gupta, S., Dalal, M., & Inamdar, A. (2016). Using Node.Js to Build High Speed and Scalable Backend Database Server. *National Conference NCPCI-2016, (4)*19.

Borkin, M. A. (2014). Perception, Cognition and Effectiveness of Visualizations with Applications in Science and Engineering. Doctoral Dissertation, Harvard University. Retrieved from http://nrs.harvard.edu/urn-3:HUL.InstRepos:12274335.

Brehmer, M. M. (2016). Why Visualization?: Task Abstraction for Analysis and Design. Doctoral Dissertation, University of British Columbia. https://doi.org/10.14288/1.0228806

Chen, C., Härdle, W., Unwin, A., & Friendly, M. (2008). A Brief History of Data Visualization. *Handbook of Data Visualization*, 15–56. https://doi.org/10.1007/978-3-540-33037-0_2

Chen, C. L. P., & Zhang, C. Y. (2014). Data-intensive Applications, Challenges, Techniques and Technologies: A survey on Big Data. *Information Sciences*, *275*, 314–347. https://doi.org/10.1016/j.ins.2014.01.015

Chen, Y., Chen, Q., Zhao, M., Boyer, S., Veeramachaneni, K., & Qu, H. (2017). DropoutSeer: Visualizing learning patterns in Massive Open Online Courses for dropout reasoning and prediction. *IEEE Conference on Visual Analytics Science and Technology, VAST 2016 - Proceedings*, 111–120. https://doi.org/10.1109/VAST.2016.7883517

Childs, H., Geveci, B., Schroeder, W., Meredith, J., Moreland, K., Sewell, C., … Bethel, E. W. (2013). Research Challenges for Visualization Software. *Computer*, *46*(5), 34-42. https://doi.org/10.1109/MC.2013.179.

Dashboard - Basic Statistics Tanzania. (2019). Retrieved October 26, 2019, from http://opendata.go.tz/dash

Demiralp, Ç., Bernstein, M. S., & Heer, J. (2014). Learning Perceptual Kernels for Visualization Design. *IEEE Transactions on Visualization and Computer Graphics*, *20*(12), 1933–1942. https://doi.org/10.1109/TVCG.2014.2346978.

Demiralp, Ç., Scheidegger, C. E., Kindlmann, G. L., Laidlaw, D. H., & Heer, J. (2014). Visual Embedding: A Model for Visualization. *IEEE Computer Graphics and Applications*, *34*(1), 10–15. https://doi.org/10.1109/MCG.2014.18.

Dennis, A., Wixom, B. H., & Roth, R. M. (2012). *System Analysis and Design* (5th Edition). John Wiley & Sons, Inc.

Dropout in Primary Schools - Datasets - Basic Statistics Tanzania. (2018). Retrieved October 22, 2019, from http://opendata.go.tz/dataset/wanafunzi-wa-shule-za-msingi-walioachishwa-au-kuacha-shule

Dropout in Secondary Schools. - Datasets - Basic Statistics Tanzania. (2018). Retrieved October 22, 2019, from http://opendata.go.tz/dataset/idadi-ya-wanafunzi-walioacha-au-kuachishwa-shule-sekondari-kwa-mkoa

Eid, M., De Cecco, C. N., Nance Jr, J. W., Caruso, D., Albrecht, M. H., Spandorfer, A. J., …

Schoepf, U. J. (2017). Cinematic Rendering in CT: A Novel, Lifelike 3D Visualization technique. *American Journal of Roentgenology*, *209*(2), 370–379. https://doi.org/ 10.2214/AJR.17.17850.

Gaur, J., Goyal, A., Choudhury, T., & Sabitha, S. (2016). A Walk Through of Software Testing Techniques. *2016 International Conference System Modeling & Advancement in Research Trends (SMART)*, 103–108. IEEE.

Haehn, D., Hoffer, J., Matejek, B., Suissa-Peleg, A., Al-Awami, A., Kamentsky, L., … Schalek, R. (2017). Scalable Interactive Visualization for Connectomics. *Informatics, 4*(3), 29. https://doi.org/10.3390/informatics4030029.

HakiElimu (2015). HakiElimu Education Position Paper. Retrieved from http://hakielimu.org/ files/publications/HakiElimu_Education_PostitionPaper_2014.pdf.

Heer, J., Card, S. K., & Landay, J. A. (2005). Prefuse : A Toolkit for Interactive Information Visualization. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 421–430. https://doi.org/10.1145/1054972.1055031.

Herzinger, S., Gu, W., Satagopam, V., Eifes, S., Rege, K., Barbosa-Silva, A., … Consortium, eTRIKS. (2017). SmartR: An Open-source Platform for Interactive Visual Analytics for Translational Research Data. *Bioinformatics*, *33*(14), 2229–2231. https://doi.org/10.1093/ bioinformatics/btx137.

Johnston, M. P. (2017). Secondary Data Analysis: A Method of Which the Time Has Come. *Qualitative and Quantitative Methods in Libraries*, *3*(3), 619–626.

Jordão, V., Gonçalves, D., & Gama, S. (2014). EduVis: Visualizing Educational Information. *Proceedings of the 8th Nordic Conference on Human-Computer Interaction: Fun, Fast, Foundational*, 1011–1014.

Kale, A., Nguyen, F., Kay, M., & Hullman, J. (2019). Hypothetical Outcome Plots Help Untrained Observers Judge Trends in Ambiguous Data. *IEEE Transactions on Visualization and Computer Graphics*, *25*(1), 892–902. https://doi.org/10.1109/ TVCG.2018.2864909

Kalinga, T. S. (2013). Causes of Drop- Out in Secondary Schools in Tanzania: The Case

Study Of Mbeya, Dar es salaam and Kilimanjaro Regions. Masters Thesis, The Open University of Tanzania.

Kendall, K. E., & Kendall, J. E. (2011). *Systems Analysis and Design* (9th Edition). Pearson Education, Inc.

Kuosa, K., Distante, D., Tervakari, A., Cerulo, L., Fernández, A., Koro, J., & Kailanto, M. (2016). Interactive Visualization Tools to Improve Learning and Teaching in Online Learning Environments. *International Journal of Distance Education Technologies*, *14*(1), 1–21. https://doi.org/10.4018/IJDET.2016010101.

Lacefield, W. E., & Applegate, E. B. (2018). Data Visualization in Public Education: Longitudinal Student Intervention - School and District-Level Performance Modeling. *2018 Annual Meeting of the American Education Research Association.* New-York.

Li, D., Mei, H., Shen, Y., Su, S., Zhang, W., Wang, J., … Chen, W. (2018). ECharts: A Declarative Framework for Rapid Construction of Web-based Visualization. *Visual Informatics*, *2*(2), 136–146. https://doi.org/10.1016/j.visinf.2018.04.011

Liu, M., Kang, J., Pan, Z., Zou, W., & Lee, H. Y. (2017). Exploring Data Visualization as an Emerging Analytic Technique. *Proceedings of E-Learn: World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education*, 1681-1690. Retrieved from https://www.learntechlib.org/primary/p/181326/.

Liu, S., Cui, W., Wu, Y., & Liu, M. (2014). A Survey on Information Visualization: Recent Advances and Challenges. *The Visual Computer*, *30*(12), 1373–1393. https://doi.org/10.1007/s00371-013-0892-3.

Liu, X., & Shen, H.W. (2015). The Effects of Representation and Juxtaposition on Graphical Perception of Matrix Visualization. *Proceedings of the 33rd annual ACM conference on Human factors in computing systems*, 269–278. ACM.

Mahat-Shamir, M., Neimeyer, R. A., & Pitcho-Prelorentzos, S. (2019). Designing in-depth semi-structured Interviews for Revealing Meaning Reconstruction After Loss. *Death Studies*, 1–8. https://doi.org/10.1080/07481187.2019.1617388.

Marshall, G. (2005). The Purpose, Design and Administration of a Questionnaire for Data

Collection. *Radiography*, *11*(2), 131–136. https://doi.org/10.1016/j.radi.2004.09.002.

Mckenna, S., Mazur, D., Agutter, J., & Meyer, M. (2014). Design Activity Framework for Visualization Design. *IEEE Transactions on Visualization and Computer Graphics*, 20(12), 2191 - 2200. https://doi.org/10.1109/TVCG.2014.2346331.

Mduma, N., Kalegele, K., & Machuve, D. (2019). An Ensemble Predictive Model Based Prototype for Student Drop-out in Secondary Schools. *Journal of Information Systems Engineering & Management*, *4*(3). https://doi.org/10.29333/jisem/5893.

Munzner, T. (2009). A Nested Model for Visualization Design and Validation. *IEEE Transactions on Visualization and Computer Graphics*, *6*(15), 921–928. https://doi.org/ 10.1109/TVCG. 2009.111.

National Bureau of Statistics. (2019). Tsed_Final. Retrieved October 26, 2019, from http:// www.tsed.go.tz/dashboard/Dashboard#/

OECD. (2016). *Trends Shaping Education 2016*. Paris: OECD Publishing. https://dx.doi.org/ 10.1787/trends_edu-2016-en

Okubo, F., Shimada, A., Yin, C., & Ogata, H. (2015). Visualization and Prediction of Learning Activities by Using Discrete Graphs. *23rd International Conference on Computers in Education, ICCE 2015*, 739–744. Asia-Pacific Society for Computers in Education.

Otaduy, I., & Díaz, O. (2017). User Acceptance Testing for Agile-developed Web-based Applications: Empowering Customers Through Wikis and Mind Maps. *Journal of Systems and Software*, *133*, 212–229.

Ouma, D. H. (2017). Analysis of the Socio-Economic Factors that Contribute to Children School Dropout in Artisanal Small-Scale Gold Mining Communities of Tanzania. *Journal of Education and Practice*, *8*(14), 71–78.

Pasternak, T., Bisley, J. W., & Calkins, D. (2003). Visual Processing in the Primate Brain. *Handbook of Psychology*, 139–185.

Perneger, T. V., Courvoisier, D. S., Hudelson, P. M., & Gayet-Ageron, A. (2015). Sample Size for Pre-tests of Questionnaires. *Quality of Life Research*, *24*(1), 147–151.

Ministry of Finance and Planning. (2018). Speech by the Minister for Finance and Planning. Retrieved from http://www.mof.go.tz/docs/english budget speech.pdf

Ministry of Finance and Planning. (2018). A Simplified Version of Government Budget for Financial Year 2017/2018. Retrieved from http://www.mof.go.tz/mofdocs/ budget/Citizens Budget/citizens budget 2017_18 (Eng_Final).pdf.

Rajib, M. A., Merwade, V., Kim, I. L., Zhao, L., Song, C., & Zhe, S. (2016). SWATShare–A Web Platform for Collaborative Research and Education Through Online Sharing, Simulation and Visualization of SWAT Models. *Environmental Modelling & Software*, *75*, 498–512.

Roels, R., Baeten, Y., & Signer, B. (2017). Interactive and Narrative Data Visualisation for Presentation-based Knowledge Transfer. *Communications in Computer and Information Science*, *739*, 237–258. https://doi.org/10.1007/978-3-319-63184-4_13

Rose, A. S., & Hildebrand, P. W. (2015). NGL Viewer: A Web Application for Molecular Visualization. *Nucleic Acids Research*, *43*(W1), W576–W579. https://doi.org/10.1093/ nar/gkv402

Rovira, S., Puertas, E., & Igual, L. (2017). Data-driven System to Predict Academic Grades and Dropout. *PLoS ONE*, *12*(2). https://doi.org/doi:10.1371/ journal.pone.0171207

Samantaray, A., & Dash, S. R. (2018). Demographic Analytical Study of Girl Child Dropout from Schools in India. *International Journal of Engineering Technology Science and Research*, *4*(10), 921-926.

Sedlmair, M., Meyer, M., & Munzner, T. (2012). Design Study Methodology : Reflections from the Trenches and the Stacks. *IEEE Transactions on Visualization and Computer Graphics*, *18*(12), 2431-2440. https://doi.org/ 10.1109/TVCG.2012.213.

Smilkov, D., Carter, S., Sculley, D., Viégas, F. B., & Wattenberg, M. (2017). Direct-Manipulation Visualization of Deep Networks. Retrieved from http://arxiv.org/ abs/1708.03788

Tamayo, J. L., Hernández, M., & Gómez, H. (2018). Digital Data Visualization with Interactive and Virtual Reality Tools. Review of Current State of the Art and Proposal of a Model. *Revista ICONO14 Revista Científica de Comunicación y Tecnologías Emergentes*, *16*(2), 40–65. https://doi.org/10.7195/ri14.v16i2.1174

Tanzania National Health Portal. (2019). Data Statistics. Retrieved October 26, 2019, from https://hmisportal.moh.go.tz/hmisportal/#/datastatistics/hwJNMC3LrPd/hwJNMC3LrPd.

UNDP. (2018). UNDP Tanzania Early Warning and Response System. Retrieved from https://www.undp.org/content/dam/tanzania/docs/Project%20Documents/Inclusive-Democratic-Governance/Crisis%20Risk%20Dashboard_Fact%20Sheet.pdf.

United Nations. (1948). Universal Declaration of Human Rights | United Nations. Retrieved November 17, 2019, from https://www.un.org/en/universal-declaration-human-rights/.

Wang, L., Wang, G., & Alexander, C. A. (2015). Big Data and Visualization: Methods, Challenges and Technology Progress. *Digital Technologies*, *1*(1), 33–38. https://doi.org/ 10.12691/dt-1-1-7.

Wang, X.M., Zhang, T.Y., Ma, Y.X., Xia, J., & Chen, W. (2016). A Survey of Visual Analytic Pipelines. *Journal of Computer Science and Technology*, *31*(4), 787–804. https://doi.org/ 10.1007/s11390-016-1663-1.

West, V. L., Borland, D., & Hammond, W. E. (2014). Innovative Information Visualization of Electronic Health Record Data: A Systematic Review. *Journal of the American Medical Informatics Association*, *22*(2), 330–339. https://doi.org/10.1136/amiajnl-2014-002955.

Williamson, B. (2016). Digital Methodologies of Education Governance : Pearson plc and the Remediation of Methods. *European Educational Research Journal*, *15*(1), 34–53. https:// doi.org/10.1177/1474904115612485.

APPENDICES

Appendix 1: Interview Questions for Data Collection

General Information

I am Angelika Mark Kayanda, currently a student at the Nelson Mandela African Institution of Science and Technology, taking a Master's degree in Information and Communication Science and Engineering. I am conducting a research titled ***"Data Visualization Tool for Student Dropouts in Tanzania"***, aiming to come up with a web-based tool that visualizes student dropouts information for primary and secondary schools so as to support informed decision making.

I am kindly requesting for assistance with some information that I need; your input will be very valuable and important not only for successful development of this tool but also successful completion of my research. All the information collected will be used solely for the purpose of this research and not otherwise. Thank you very much in advance for your cooperation.

Please help us get information for our tool by responding to these few questions:

1. When making decisions related to student dropouts and retention issues, what kind of information do you use?
2. How do you get this information?
3. What tools do you prefer in analyzing the information? Why?
4. How is the information presented?
5. At which level is this information presented?
    a. School level
    b. District level
    c. Region level
6. For visual presentation, which ones would you prefer to be used?
    a. Pie charts
    b. Line graphs
    c. Bar charts
    d. Other (please state)
7. Please tell us the kind of reports that you produce regarding student dropouts information.
8. How often are these reports produced?
    a. Monthly

    b. Quarterly
    c. Semi-annually
    d. Annually
9. Is there any information on student dropouts that you would like to present, but is missing from the information that you have?
10. Can we get information on what are the most contributing factors to student dropouts?
11. How do you monitor student dropouts?
12. What more do you suggest is to be done to monitor student dropouts?

Appendix 2: Focus Group Discussion Questions

I am Angelika Mark Kayanda a student pursuing a Master's Degree in Information and Communication Science and Engineering at Nelson Mandela African Institution of Science and Technology (NM-AIST). My research is titled "**Data Visualization Tool for Student Dropouts in Tanzania**" and I have developed a web-based tool which visualizes student dropouts in primary and secondary schools. To evaluate the tool, please assist us by giving your views, and your feedback will be used to improve the developed tool.

1. Have you used a tool like this before?
   (Please put a tick in the box on the right of the most likely option)

Yes	
No	
Other response:	

2. What do you think of the tool?

Response:

3. Which feature did you like the most in the tool? Please tell us more about it.

Response:

4. Which feature did you dislike the most in the tool? Please tell us more about it.

Response:

5. As an expert in education issues, how and when do you suppose you might use the tool?

Response:

6. Since you have a chance to influence decisions in Tanzania's education issues, how do you suggest that the tool be improved to benefit decision makers in education domain?

Response:

7. Of all the things we have talked about today, what is the most important to you?

Response:

8. Is there anything else you would like to comment about the tool?

Response:

Thank you for your time and feedback!

Appendix 3: Questionnaire for Tool Acceptance Testing

I am Angelika Mark Kayanda a student pursuing a Master's Degree in Information and Communication Science and Engineering at Nelson Mandela African Institution of Science and Technology (NM-AIST). My research is titled "**Data Visualization Tool for Student Dropouts in Tanzania**" and I have developed a web-based tool which visualizes student dropouts in primary and secondary schools. To evaluate the tool, please assist us by giving your views, and your feedback will be used to improve the developed tool.

Note: Tick (√) where applicable

Title/Position:
……………………………………………………………………………………………

Please select the best answer on the evaluation of the developed Data Visualization Tool for Student Dropouts.

S/N	**Question**	Strongly agree	Agree	Neutral	Disagree	Strongly disagree
2.	I will use this tool in my work					
3.	I found the visualization tool to be difficult to use.					
4.	I found the visualization tool to be easy to use					
5.	I will need support of a technical person to be able to use this tool					
6.	I found the tool useful for reporting and decision making at my line of work					
7.	I find the tool useful for reporting and decision making at other departments and organs within the Government					
8.	I will recommend this tool to my team/colleagues					
9.	I recommend some features to be removed in the tool.					
10.	I think there are more features to be added in this tool					

11. Is there anything you would like to comment on this tool?

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

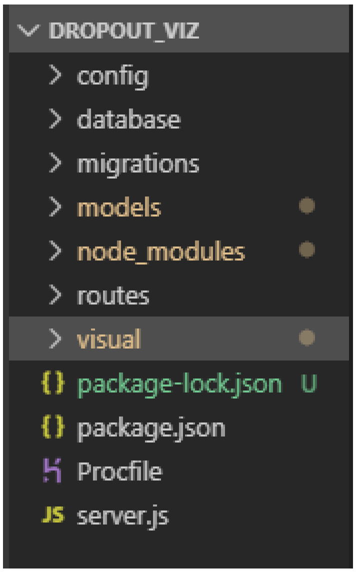
…………………………………………………………………………………………………

……………

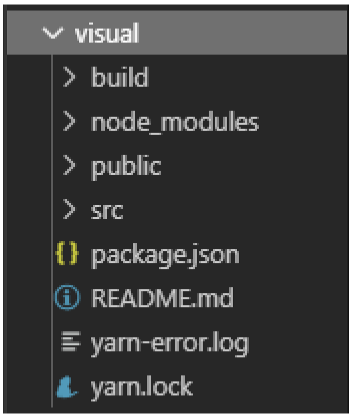
Thank you very much for your feedback!

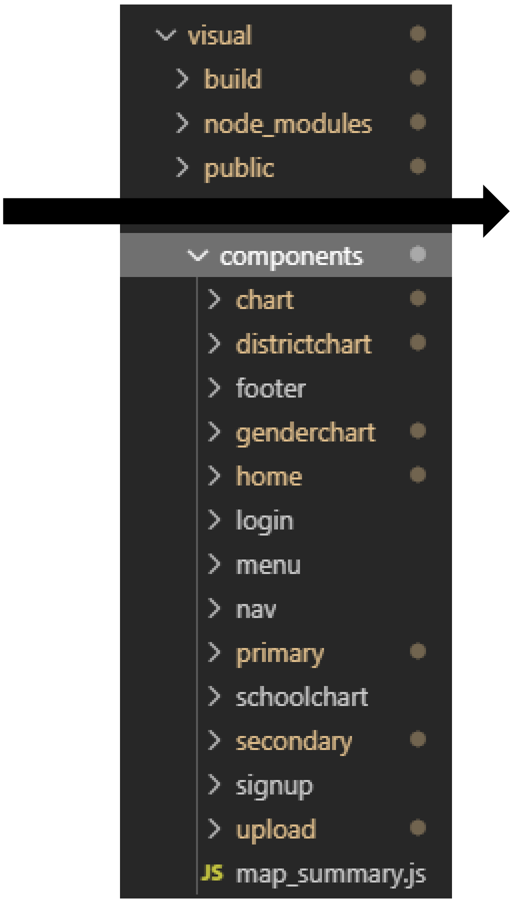
Appendix 4: Application structure

A. Overall Application structure



B. Application Front end structure





C. Application Back end structure