



**AFRICAN CENTER OF EXCELLENCE
IN DATA SCIENCE**



**Use of Data Mining Techniques to Create Reference Pricing Approach for the
Rwanda Public Procurement**

By

Marie Josee UMUGWANEZA

Registration Number: 219013888

**A Dissertation Submitted in Partial Fulfilment of the Requirement for the Degree of
Master of Science in Data Science**

**University of Rwanda College of Business and Economic
The African Center of Excellence in Data Science (ACE-DS)**

Supervisor: Dr RUHARA MULINDABIGWI Charles

September 2020

Declaration

I declare that this thesis was written solely by myself and that it has not been submitted, in whole or in part, in any previous application for a degree. Except where stated otherwise by reference the work presented is entirely my own.

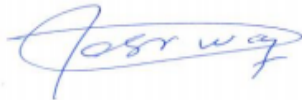
Submitted by: Marie Josee UMUGWANEZA

Signature:



Supervisor: Dr RUHARA MULINDABIGWI Charles

Signature:



Dedication

I dedicate my dissertation to God Almighty, my creator, my strong pillar, my source of inspiration, wisdom, knowledge, and understanding. I also dedicate my work to my whole family for the endless support and encouragement throughout my studies. Thank You. My love for you all can never be qualified.

God Bless you.

Acknowledgements

Foremost, I would like to express my gratitude to my academic supervisor Dr. Charles RUHARA MULINDABIGWI for the continuous support of my research, for his patience, motivation, enthusiasm, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. Many thanks to Rwanda Public Procurement for every useful information, guidance, and entrusting this study with essential data from the online e-procurement system. Special thanks go to Mr. Jean Luc ZIRAVUGA for his warm welcome, career guidance, and knowledge transfer during my industrial attachment period. I also acknowledge family and friends who played their roles for the successful completion of this dissertation. Be Blessed All.

Abstract

In order to support e-procurement operation, Rwanda Public Procurement Authority (RPPA) initiated and implemented policies of using reference prices as tool to strengthen public procurement practice. The reference prices are used by procuring entities (public institutions) in evaluation of tenders. Currently, reference price is calculated manually by gathering information from different suppliers on the market place located in Kigali and other districts of Rwanda. Due to the manual process, the update of the previous reference price long time. The main purpose of this study was to develop a digitalized approach of creating reference prices by using data from online e-procurement system for Rwanda. Additionally, the study purpose of this study was to show how k-means clustering with text vectorization method used to create a reliable reference price for similar items. Text clustering was applied to identify groups of similar services or products. Text vectorization methods, namely Bag of words and tf-idfVectoriser are investigated in clustering similar item using k-means clustering. Tf-idfVectoriser fitted k-means clustering method with robustness and optimal clusters of 13 obtained from elbow method. Thus, tf-idfVectoriser was the best method in creation of reference price per similar item. Finally the average reference price per item were computed, the results proved that there exist large standard deviation between prices of similar services or products. This deviation resulted in abnormally higher reference price per item. To avoid the abnormality, the median was measured for the purpose of comparison with the obtained average price. In this case, the median prices per similar item performed well than the average price based on the existing market price plus other related costs. Therefore, for products and services category of items, the study accepted the median prices as the reliable reference prices per item.

Key words: Reference price, e-procurement, clustering, vectorization, public procurement, item approach, and digitalization.

Table of contents

Declaration	i
Dedication	ii
Acknowledgements	iii
Abstract	iv
List of tables	vii
List of figures	viii
1.0 GENERAL INTRODUCTION	1
1.1 Background	1
1.2 Problem statement	3
1.3 Objectives	4
1.3.1 General objectives	4
1.3.2 Specific objectives	4
1.4 Significance of the study	4
1.5 Structure of the thesis	5
2.0 LITTERATURE REVIEW	6
2.1 Introduction.....	6
2.2 Digitalization of public procurement	6
2.3 Reference price	7
3.0 METHODOLOGY	9
3.1 Introduction.....	9
3.2 Data selection and variables discussion	9
3.3 Data pre-processing.....	11
3.3.1 Dimensionality reduction	11
3.4 Clustering and text vectorization methods	12
4.0 PRESENTATION OF THE RESULTS.....	14
4.1 Introduction.....	14
4.2 Data pre-processing and visualization of procured items.....	14
4.2.1 Distribution of the standardized features	15
4.2.2 Clustering	16
4.2.3 Using text vectorization to classify item clusters	18
4.3 Results	20

4.3.1 Price trends	20
4.3.2 Text clustering and text vectorization for creating the reference price	22
4.3.3 Computing reference prices	23
5.0 CONCLUSIONS AND RECOMMENDATIONS	26
5.1 Conclusions.....	26
5.2 Recommendations and future research.....	26
REFERENCES.....	27

List of tables

Table 1: Sample dataset of items purchased through the online e-procurement system...	10
Table 2: Frequent words in each cluster	19
Table 3: Randomly selected items in products	21
Table 4: Reference Price for items in the product category.....	23
Table 5: Reference price for item categorized in services	24

List of figures

Figure 1: Visualization of procured items	14
Figure 2: Distribution of standardized features.....	15
Figure 3: Plot of Cumulative Explained Variance	16
Figure 4: Elbow for k-means clustering	18
Figure 5: Elbow Curve for the optimal number of clusters	20

1.0 GENERAL INTRODUCTION

1.1 Background

Public procurement is the acquisition of goods, services or construction materials on behalf of public entities (Young, 2019). Public procurement accounts for a substantial part of the global economy, in the sense that, in 2018, it accounted for 12 % of global GDP (Bosio and Djankov, 2020). Public procurement entails the use of public funds to ensure value for money, achievement, and realization of public programs and projects at minimum cost. Governments around the world constantly buy goods and services from the private sector, from small expenses to large infrastructure projects, keeping the procurement process simple and cost-efficient in competitive situations (OECD, 2020).

To minimize complexity in procurement processes, countries worldwide have implemented a digitized process to make it easier, transparent, and less costly. Digital procurement is linked to e-Government as one of the key drivers toward the implementation of a single market strategy (European Commission, 2020). In the age of big data, digital procurement is crucial in enabling the government to make a data-driven decision on public spending. With digital instruments, public spending has become more transparent, evidence-oriented, optimized, streamlined, and integrated with market conditions. The adoption of digital procurement offered a range of benefits such as significant savings for all parties, shortened processes, increased transparency, and greater innovation (European Commission, 2020). The digital transformation of public procurement has affected all levels of the public sector of the governments where many countries moved towards administrative decentralization with local authorities by leading the way to invent and implement better policies. In addition, digitalization in public procurement helped public institutions to accelerate growth and pillow their modernization process (Salazar and Harper, 2018).

Governments around the world considered digital procurement to improve their performance in public procurement. Innovation in government procurement is recognized to be important for the enhancement of the competitiveness of government operation and performance (Gun Lim et al., 2008). The digitalization of public procurement triggered

the implementation of reference prices in many countries. Data from the e-procurement portal was shown to be a tool that proves how essential information about prices of the product can be retrieved and refined to improve social control and provide solid accountability for the government (Paiva et al., 2014).

In the process of replacing traditional public procurement activities with electronic means, the Government of Rwanda launched a full-functioning online e-procurement system, for attracting the maximum number of bidders, achieving the objective of quality at the best price, and also for fighting against corruption (World Bank Group, 2016). In order to support e-procurement operation, Rwanda Public Procurement Authority (RPPA) initiated and implemented policies of using reference prices as a tool to strengthen public procurement practice. The reference prices are used by procuring entities (public institutions) in the evaluation of tenders. The reference prices document is prepared and published by RPPA, comprising of prices collected on the domestic market from suppliers in Kigali and other selected Districts. In the preparation of reference prices, exchange rate and price of raw materials are considered in determining variation between previous prices and current prices of a product on the market (Rwanda Public Procurement Authority, 2019)

After the implementation of reference prices in public procurement of Rwanda, it was shown that some entities do not refer to reference prices when they evaluate tenders (Rwanda Public Procurement Authority, 2017) . On the other side, bidders declared that reference prices are not accurate, for being either above or below the market prices. This means that reference prices need to have frequent updates. In addition, the Auditor General Report (2016) pointed out the existing reference price is inaccurate and is a source of losses because of overspending (Auditor general report, 2016; Transparency Rwanda, 2016). In fact, the Government paid higher amounts for purchased goods or services than estimated because of outdated reference prices, originated from abnormally high prices provided during the submission of bids at the tendering stage. Given the above, there are drawbacks and flaws identified with the existing reference price methodology. Thus, there is a need for the new approach of developing average reference prices per product purchased by the government of Rwanda through an online e-procurement system.

1.2 Problem statement

The E-Procurement system for Rwanda was established to strengthen and accelerate RPPA activities. This is an online platform where tendering related activities for all government institutions are published and monitored by a system. Though the public procurement procedures were digitalized, there are still issues identified in relation to reference price being used in public procurement (Mulugeta, 2016). Currently, reference prices are prepared manually by gathering information from different suppliers in Kigali and other districts. In addition, RPPA takes more than six months to publish a new version of reference prices, and this results in outdated reference prices. Also the use of outdated or completely ignoring reference prices causes differences in prices of similar items being procured by government institutions on a daily basis.

Awarding tenders based on prices that are above or below the market prices cause failures and incompleteness of government projects. This is where the government encounters significant losses (Transparency International Rwanda(TIR), 2016) The TIR reports mentioned that public procurement has weaknesses that cost the government of Rwanda billions of losses. Most of the losses came from overspending made while government procures goods, services, or construction materials.

The existing approach of preparing reference prices manually is subject to bias from the inflation rate. For instance, the inflation rate between the end of 2019 to February 2020 lied between 7.7 and 8.7 (NBR, 2020). Thus, waiting for 6 months to adjust reference prices results in referring to price levels that are different from market prices. To address the issue, this study proposes an approach of determining accurate reference prices using data mining and that minimizes human errors. The approach allows frequent updates of reference prices, therefore help procurement decision makers awarding bids at the accurate prices per product or service. The new approach proposes reference prices that will be used by government institutions in the evaluation and auditing of tenders.

1.3 Objectives

1.3.1 General objectives

This main objective of this study is to develop a digitalized approach for creating a reference price for Rwanda public procurement using data mining techniques such as clustering methods together with text mining techniques.

1.3.2 Specific objectives

The specific objectives of this study include:

- To propose a digitalized reference prices for items procured through online e-procurement system for Rwanda
- To propose an appropriate text clustering and text vectorization methods reliable in creating reference price per item.
- Formulate recommendations based on findings.

1.4 Significance of the study

The result of this study rebound on to the benefit of the society considering that prices of products and services are an important selection criterion which plays a major role in the procurement process of today. Also, the increase of government expenditures justifies the need for more reliable reference pricing approach for items purchased through online e-procurement system. Thus, the government applies the result of this study, failure in execution of tenders or the number of abandoned projects due to insufficient resources will be reduced. In fact, price of a given item will be analyzed based on the reference prices before awarding a contracts. For Rwanda Public Procurement, the proposed digitalized approach will help in developing a reliable reference prices that require a short period of time and less human efforts. In addition, auditors will benefits from the results of this study because it will be easier to investigate whether the government procurement made at normal prices compared to the market prices. For suppliers, the availability of updated reference prices will be used in quotation of appropriate prices based on the reference prices proposed on the similar item.

1.5 Structure of the thesis

This thesis propose digitalized approach of creating reference price for item procured through online e-procurement system for Rwanda. The general structure of this thesis is as follows. Chapter 1 represent the introductory features; general concepts to help in understanding the context of the research. It also highlight the background, research problem, and study objectives. Chapter 2 concentrates on reviewing other related work, especially literatures in line with reference price and digital procurement. Chapter 3 provides a brief description of the methodology which relies on data mining techniques applied to identify similar items in each purchase to obtain a reliable reference prices. Chapter 4 emphasizes on the prior analysis on the reference prices and discuss the final results related to the proposed reference prices approach. Chapter 5 draws the conclusion, and defines the recommendations based on findings.

2.0 LITTERATURE REVIEW

2.1 Introduction

This chapter reviews the literature that is relevant to the understanding of the concepts and development in the area of e-procurement and reference price. Apart from the introduction, the chapter is divided into two sections. Section two describes the introduction of digitalized public procurement worldwide discussing the advantages and disadvantages associated with it. The third section highlights the concept of reference prices in public procurement and encloses various methodologies used to determine reference prices worldwide.

2.2 Digitalization of public procurement

One of the areas that attracted research in public procurement is procurement digitalization. Bobowski and Gola (2018) argued that procurement digitalization is an important trigger of sustainable growth based on innovation due to the expected increase in efficiency and transparency of public spending. Market research reports show that e-procurement practices can save between 10% and 50% of government spending (Peria, 2003). The savings are believed to be derived mainly from better sourcing decisions and reduced administrative costs (AGESHIN, 2001; Baker and m.Sinkula, 1999). The public sector in some countries is characterized by high purchasing volume, maverick buying, and the lack of transparency, standing to benefit significantly from e-procurement. The private sector driven by the desire to maintain a competitive advantage and by the need to maintain profitability has taken rapid strides in using e-procurement (Frank and Daniela , 2003). E-procurement in the public sector is being implemented worldwide and much money is spent to build up and implement e-procurement solutions. In general, the digitalization of public procurement proved to be a greater contribution to the achievement of enhanced efficiency, accountability, transparency, and participation of small and medium enterprises in tenders (McCue and Roman, 2012).

A recurring issue in e-procurement is the analysis of its potential advantages. Chopra et al., (2001) state that they can derive from reduced transaction charges, improved market efficiency, and enhanced supply chain benefits. Essig and Arnold (2001) showed how

digital procurement deeply empowers the buyer's position because of both greater actual and existing forecasted information it can guarantee in the purchase of exploring, experience, and reliance goods. Presutti (2003) pointed out the usefulness of e-procurement in addressing a wide spectrum of various issues in social and environmental sectors, it addressed the legal framework of e-procurement in the suggestion of further provisions and solutions that should be delivered to enhance deeper standardization and interoperability in the e-procurement market (Thai, 2017).

However, despite the advantages provided by e-procurement digitally, other papers found that there exist downfalls. In 2017, more than 60% of procurement procedures used the lowest price as the only award criteria without consideration of innovation and fair competition between suppliers (Somasundaram and Damsgaard, 2005). Asenso-Boakye (2014) raised issues of prices as one of the determinants that trigger public procurement irregularities in different countries. Gouveia (2002) wrote that normal prices should be set as technical criteria by which contracting authorities shall base the award of the contract. The study also mentioned that there are still contracts that fall through because little attention was given to the price of contracts being awarded.

2.3 Reference price

Several authors reviewed the introduction of reference prices in procurement and showed that there are significant impacts of using reference prices in procurement processes. Though some authors defended the importance of reference prices in the reduction of hospitals' inefficiency in the procurement process, it also raised several questions about calculation and incorporation of reference prices in daily procurement routine (Marchi, 2016). Other studies were conducted to find a relevant approach to classify and determine the average reference prices per product. Aurelio (2018) worked on anthology based text mining and clustering techniques to consolidate reference prices per medication. Apart from providing reference price per medication, this study presented a way of discovering procurement cases that merit further investigations by comparing reference prices and purchasing prices. Adani (2016) defined reference prices of medical devices by using non-parametric method testing whether the introduction of reference price reduces inefficiencies in procurement, and how much the observed paid prices departs from

estimated optimal prices. Findings show that inefficiency without reference prices dominates inefficiency with the presence of reference prices.

With regards to reference prices in public procurement, other studies investigated the methodology of setting price criteria in public procurement. Fuentes-Bargues (2015) proposed a methodology as a tool for control and price justification for public contractors. The methodology showcased how abnormal prices could be identified in tendering. The study further concluded that tenders below market prices cause problems during implementation such as delays, claims, contradictory pricing, complementary projects, and even paralysis and non-completion of government projects. Some other studies such as Carvalho et al., (2014) did research on the use of data mining techniques to determine average reference prices per product. But, due to unstructured data, the study concluded as unsuccessful in the definition of average reference price per product.

In public procurement, prices are always considered as an important criterion. Most studies presented in this literature showed gaps in the selection of appropriate methodologies and the availability of suitable data to be used. Based on the literature review, there is no study conducted in Rwanda relating to the approach of creating reference prices by using data mining techniques. Thus, conducting a similar study will support policy formulation.

3.0 METHODOLOGY

3.1 Introduction

This chapter discusses the methodology used to develop the approach of creating reference price for the Rwanda Public Procurement. Apart from the introduction, the chapter is composed by other three main sections. Section two presents data collection and variables discussion. In the same section, we deeply explain the origin of the data, techniques used to obtain the data and the importance of each variables on the final results. The third section explain in details steps involved in data preprocessing as well as exploratory data analysis. The last section of this chapter highlight the use of text clustering and text vectorization methods in computing reference prices.

3.2 Data selection and variables discussion

The study uses Public Procurement administrative data collected from an online e-procurement system. These are secondary data stored in the Oracle Database management system to be retrieved from the database with Structured Query Language (SQL). We focused on data that defines items purchased in the year 2018 and 2019 because this is the range that defines significance of data needed.

	CONT_NO	CL_ID	CL_NM	ITEM_IDENT_NO	ITEM_IDENT_NM	UNIT_PRC_AMT	QUANT	TOTL_ITEM_AMT
0	C1020432832018000010	78181507	Automotive and light truck maintenance and repair	10016473	Automotive and light truck maintenance and rep...	5000.0	1.0	5000.0
1	C1020432832018000010	78181507	Automotive and light truck maintenance and repair	10016472	Automotive and light truck maintenance and rep...	2500.0	1.0	2500.0
2	C1020432832018000010	78181507	Automotive and light truck maintenance and repair	10016565	Automotive and light truck maintenance and rep...	59000.0	1.0	59000.0
3	C1020432832018000010	78181507	Automotive and light truck maintenance and repair	10016417	Automotive and light truck maintenance and rep...	2500.0	1.0	2500.0
4	C1018884712018000024	55101520	Instruction sheets or booklets	10003275	Examination booklet per page	59.0	1.0	59.0
5	C1013389972018000056	73131904	vegetable oil	10017768	vegetable oil, N/A, Refined cooking oil, Jerry...	1451.4	1.0	1451.0
6	C1074296092018000016	50202206	Spirits or liquors	10011138	Spirits or liquors, NA, Whisky (bottle) 100 cl	300000.0	1.0	300000.0
7	C1024201942018000208	56111513	Conference or non modular room packages	10011415	Conference or non modular room packages, NA, f...	25000.0	1.0	25000.0
8	C1006004072018000010	81112306	Printer, scanner and multifunctional equipment...	10016878	Printer, scanner and multifunctional equipment...	5000.0	1.0	5000.0
9	C1006004072018000010	43222612	Network switches	10016871	Network switches, N/A, Switch 3com	35000.0	1.0	35000.0
10	C1016114842018000032	72101517	Portable generator maintenance and or repair s...	10008320	Portable generator maintenance and or repair s...	150000.0	1.0	150000.0
11	C1016114842018000032	72101517	Portable generator maintenance and or repair s...	10008320	Portable generator maintenance and or repair s...	8000.0	1.0	8000.0

Table 1: Sample dataset of items purchased through the online e-procurement system

The table 1 above represents the sample structure of the whole dataset which is composed by 100826 observations and 8 attributes of items purchased through the online e-procurement system. The dataset structure is composed by purchase identification code noted as CONT_NO, it identifies a single transaction in the dataset. Class identification number, noted as CL_ID, and uniquely classify individual types of similar items. Class name per purchase is denoted by CL_NM, it is a description of a single class of items. Item identification number noted as ITEM_IDENT_NO and identifies a unique item in the dataset. Identification name per item is denoted by ITEM_IDENT_NM, and briefly explains an item to be procured. Unit price amount per item is noted as UNIT_PRC_AMT in the dataset, it also highlight the amount paid per item in Rwandan Francs. Quantity per item noted as QUANT, it denotes quantity purchased per transaction and is measured based on the type of item. Total price amount per item is noted as TOTL_ITEM_AMT in the dataset, it is the product of quantity and unit price per item expressed in Rwandan Francs. It should be noted that item identification name and unit price per item are targeted variables in the creation of the reference price.

3.3 Data pre-processing

Before starting text mining and clustering data analysis, we applied data preprocessing steps. These steps helped to enhance the quality of the data and to promote the extraction of meaningful insight from the data. This section referred to the techniques of cleaning and organizing the data to make it suitable for analysis. The process involved acquiring the data, importing all crucial libraries, importing the dataset, and identifying and handling missing data. We applied data cleaning by removing all null values identified in the dataset. We also explored the data by using a boxplot. Boxplot is a function used in exploratory data analysis to visually show the distribution of the numerical data (McLeod, 2019) , and skewness through displaying the data quartiles (or percentiles) and averages. Also, we applied feature normalization, feature scaling and dimensionality reduction as explained in the next sections

3.3.1 Dimensionality reduction

Principal Component Analysis (PCA) was used to further reduce the higher dimension of the large dataset into lower dimension dataset (He et al., 2013). As previously explained, the dataset contains 100826 observations and 8 variables, and this is the reason for transforming large sets of variables into smaller one that still contains most of the information from the original set. The main purpose with dimensionality reduction is to enable much easier, faster and accurate driven approach of analyzing data without inessential variables to process.

We perform dimensionality reduction by following the following steps:

- i. We first standardize the features using the SkLearn Library – StandardScaler. Standardizing our data makes it to be distributed around a mean of zero and a standard deviation of one. This is an important requirement for many of the machine learning models before fitting the model.

- ii. Next, we select the number of principal components to use for fitting the PCA model. This is done by plotting the cumulative explained variance against the number of components. This helps to indicate the minimum number of components required to explain the most variance in our model.
- iii. We then proceeded to fit a PCA model with the selected number of components. We fitted and transform with the standardized features to obtain the principal components to be used for further clustering analysis.

3.4 Clustering and text vectorization methods

In clustering and text vectorization we focused on item identification name and unit price per item as our targeted variables. The aim was retrieving unique information of an item being procured by using the identification name column, then cluster similar items together to find their reference prices.

Clustering was used to divide the group of purchases into many groups such that items in the same groups are more similar to other items than those in other different groups(example: fuel diesel, a bottle of mineral water(0.5L), meeting room or banquet, etc...). Simply the aim is to segregate groups with traits and assign them into clusters (Kaushik, 2016).

The clustering method defines how consolidated prices of similar items can be found. This is exactly what clustering techniques did, as its main objective is combining similar items in one cluster and dissimilar items in a separate cluster (Mining, 2011). We examined the approach by the k-means clustering algorithm as it represents each cluster by a single mean vector (Pradeep, 2015).In this analysis k-means clustering was applied in its respective six steps of specifying the desired number of clusters, randomly assigning each data point to a cluster, computing cluster centroids, re-assigning each item to the closest cluster, re-computing each cluster centroids, and finally repeating 4th and 5th point until no other improvement possible.

After similar items grouped together, the next part was understanding the exact item being described in each cluster. The text mining technique was used to retrieve information from

each cluster by using the item identification name. To apply k-means clustering on text data, we converted text into mutually comparable vectors.

To achieve this task, we represented the item identification name using Tf-idf (Term Frequency-Inverse document Frequency) algorithm. The Tf-Idf was used as a weight that ranks the importance of a term in a contextual document. Each term was calculated as a normalized frequency. A ratio of the number of occurrences of a word in its document to the total number of the word in its document (Zong, 2013). In general, we transformed text into numbers, then fed Tf-idf score to the k-means clustering model. Finally, we evaluated average reference prices on a cluster that define similar item purchased in public procurement. Note that, clustering and text mining analysis compiled in a python programming language.

4.0 PRESENTATION OF THE RESULTS

4.1 Introduction

This chapter underlines and discusses the results. Apart from the introduction, section 2 visualizes for better understanding of procured items between the Year 2018 and 2019. It covers steps involved in the application of the k-means clustering and text vectorization method. Section 3 highlights the discussion of the reference prices based on the findings.

4.2 Data pre-processing and visualization of procured items

The figure 1 below highlights the overview of procured items between 2018 and 2019. It was shown that the topmost procured items are meeting or banquet rooms accommodation, Single room, dining servers or buffet, conference rooms, tea break with snacks per person, and garage services. The top procured items vary between 100 and 250 transactions per year. Meeting or banquet rooms has 247 transactions, single room has 148 transactions, dining servers has 134, tea break with snacks per one person has 119 transaction and garage services has 119 transactions. The above mentioned number of transactions computed by grouping similar items based on information retrieved from identifications name from the dataset.

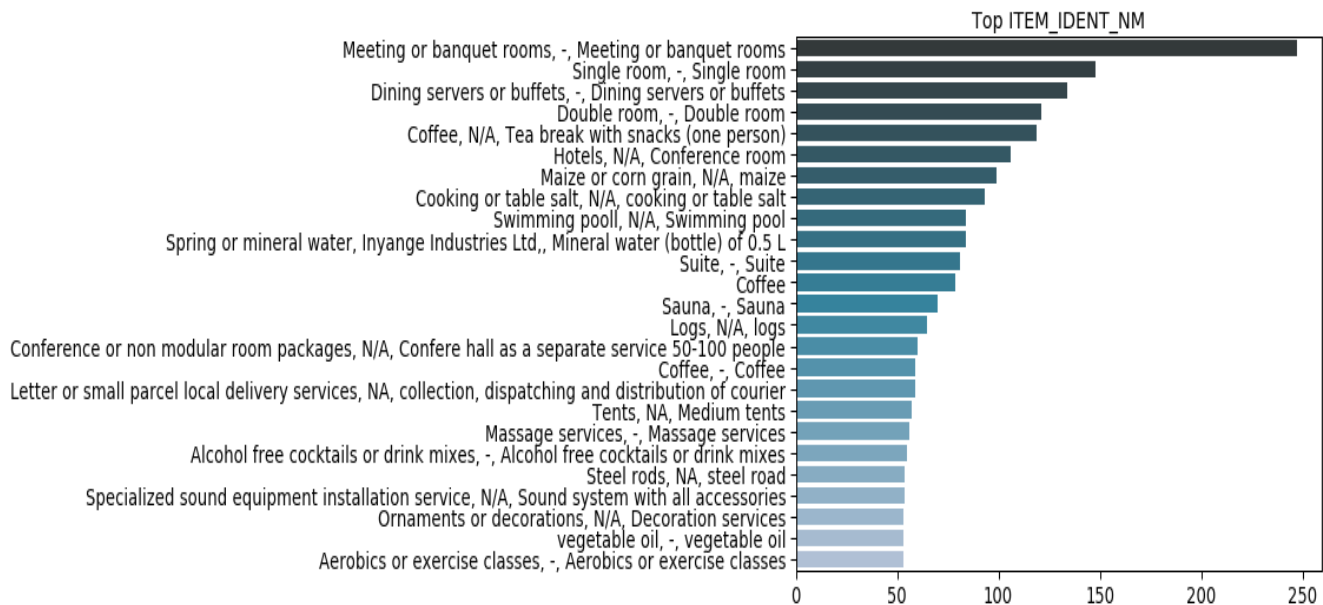


Figure 1: Visualization of procured items

4.2.1 Distribution of the standardized features

Before clustering, we applied dimensionality reduction to the returned feature so that we can project the data onto the lower dimension with just few principal components. For example, unit price per item has higher dimension data than other components. So, we standardized to get the normally distributed unit prices of mean 0 and standard deviation of 1 before fitting the PCA model. Distribution of standardized features is shown in the below figure 2.

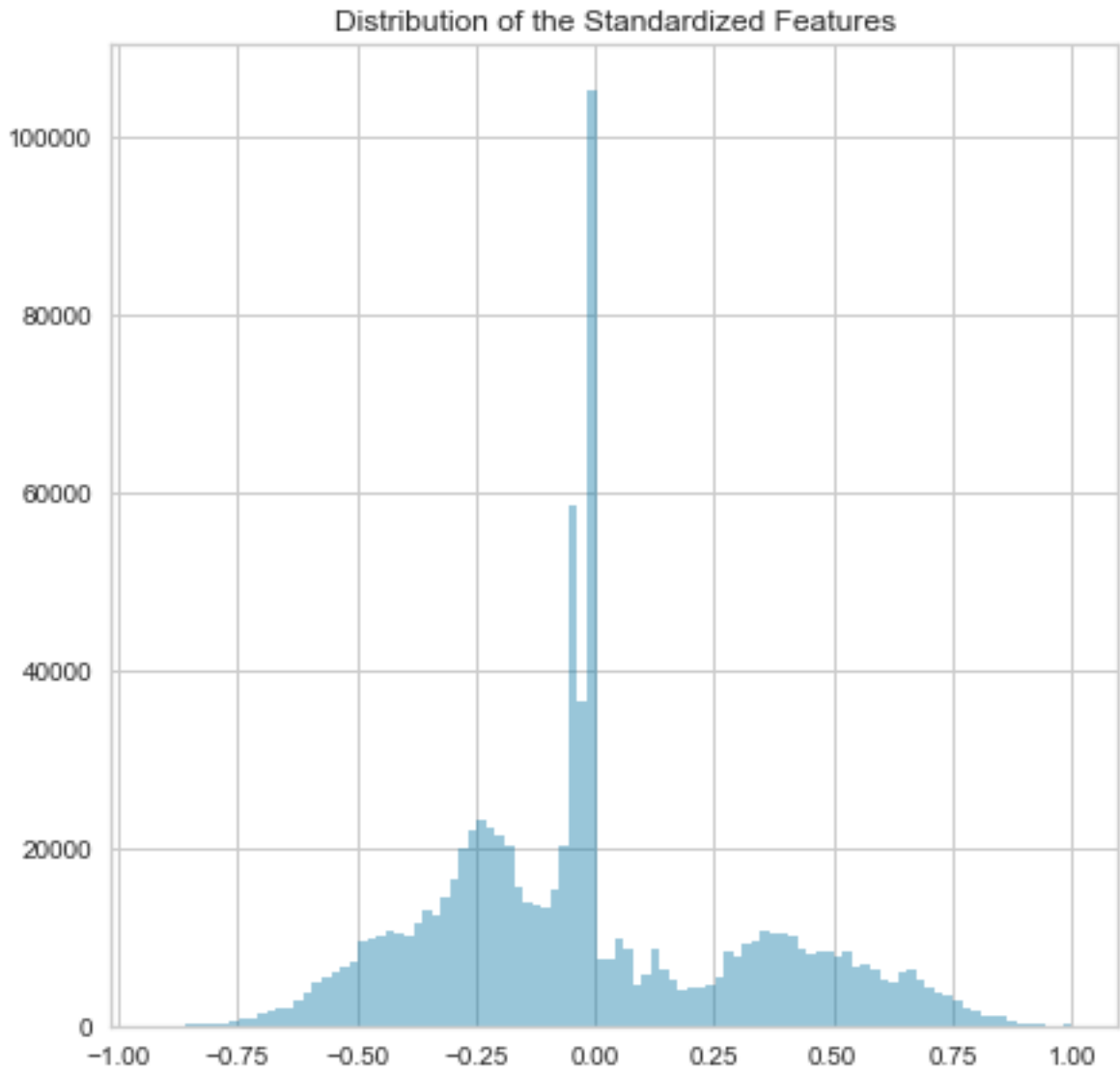


Figure 2: Distribution of standardized features

The following figure 3 describes the cumulative explained variance plotted against to the number of principal components. It also highlights the number of initial variables to be used while fitting the PCA (Principal Components Analysis) model.

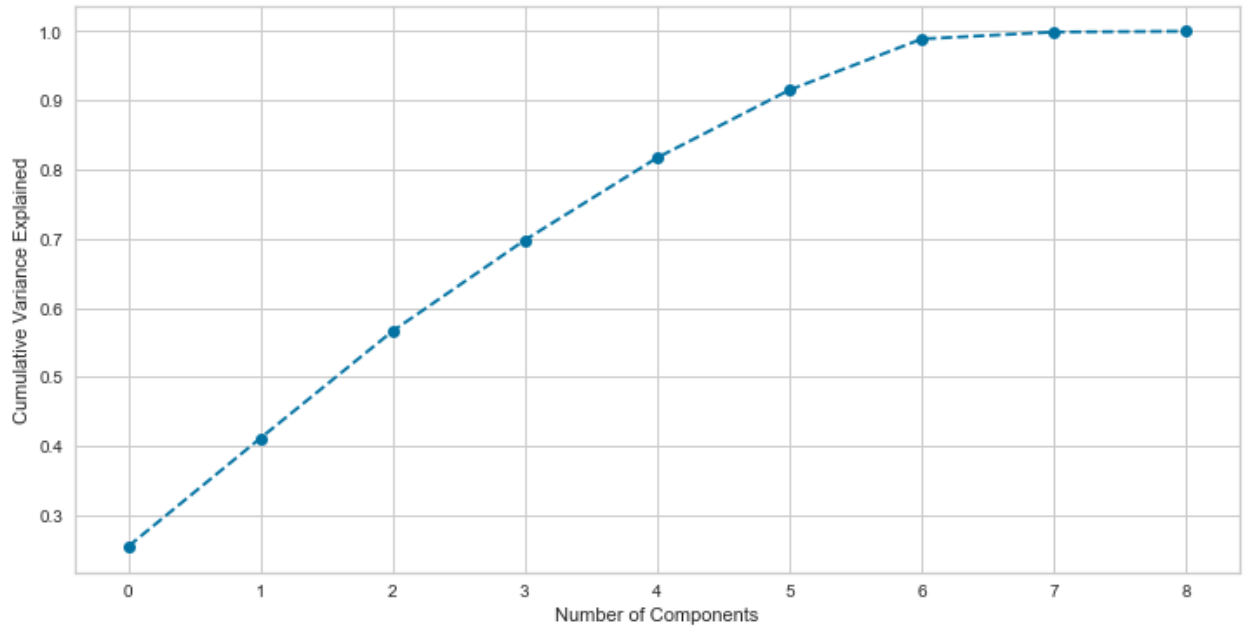


Figure 3: Plot of Cumulative Explained Variance

From figure 3 above, we can see that almost 100% of the variance can be explained by just about six principal components and as such, we proceed to fit our PCA model with 6 components to represent the important features for clustering analysis.

4.2.2 Clustering

K-means clustering was implemented to find an optimal cluster to be used in identification of reference price. The core idea was to select k centers, one for each cluster. There are many ways to initialize those centers. It can be done randomly, pass certain points that we believe are the center, or smartly place them (e.g. as far away from each other as possible). Then, we calculated the Euclidean distance between each point and the cluster centers. We assigned the points to the cluster center where the distance is minimum. After that, we recalculated the new cluster center. We select the point that is in the middle of each cluster as the new center.

And we started again to calculate the distance assigned to the cluster, calculated new centers, and stopped when the centers could not move anymore.

We needed to select the optimal number of clusters to get a good within cluster score. To realize that we iterated through different K values and plotted the total within cluster distances for each K value. We selected the K value that caused a sharp drop in total within cluster distance. This drop usually resembles an “Elbow”.

4.2.2.1 Choosing the right K

The way to evaluate the choice of K is made using a parameter known as WCSS. WCSS stands for Within Cluster Sum of Squares. Here’s the formula representation, for example when $K = 3$. The K-Means algorithm clusters data by trying to separate samples in n groups of equal variances, minimizing a criterion known as inertia, or within-cluster sum-of-squares Inertia, or the within-cluster sum of squares criterion, can be recognized as a measure of how internally coherent clusters are.

4.2.2.2 The Elbow Method

K-means clustering was executed by selecting a range of 20 clusters and results were displayed in Elbow Visualizer using the elbow method. We created clusters considering K values from one to twenty. In the depiction below we can see that after 3 there's no significant decrease in WCSS so 3 is the best here. Therefore, there's an elbow shape that forms and it is usually a good idea to pick the number where this elbow is formed. There would be many times when the graph wouldn't be this intuitive but with practice, it becomes easier. When analyzing the clusters, it was observed that the procurement reference prices decrease as the number of clusters increases.

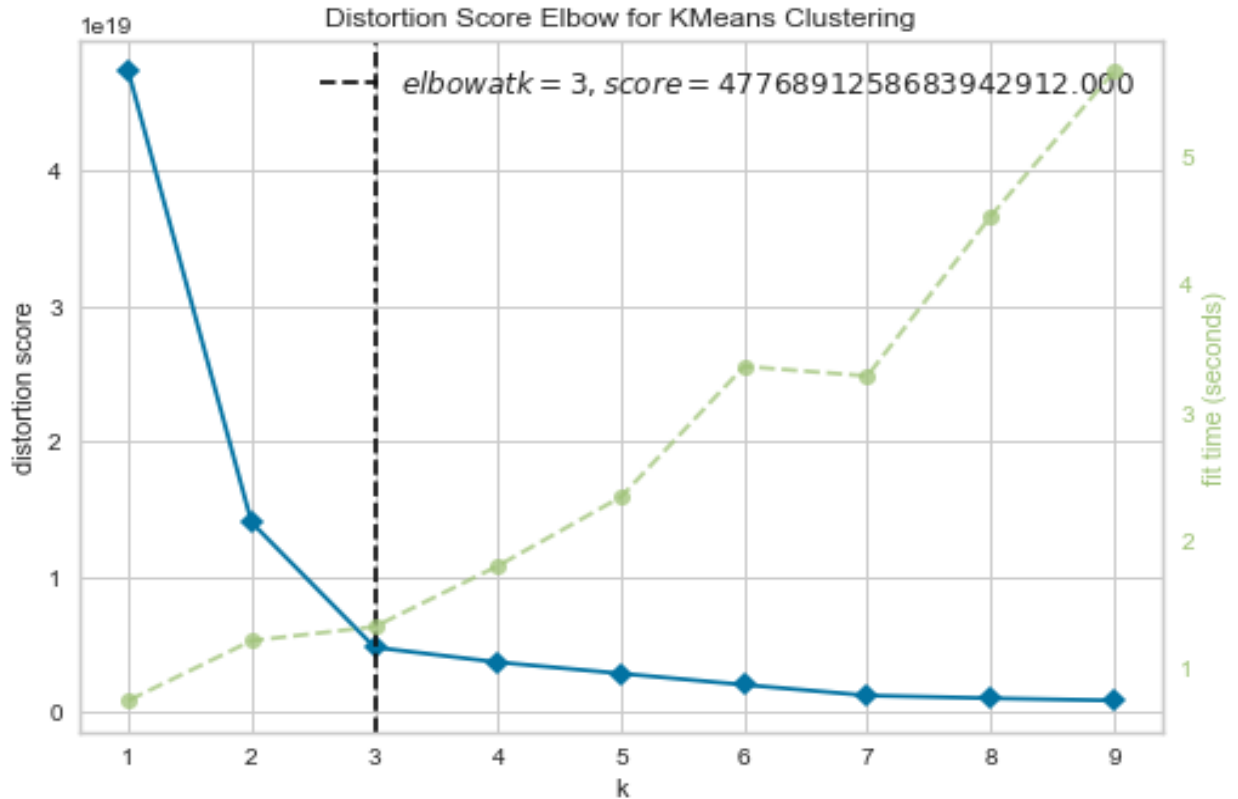


Figure 4: Elbow for k-means clustering

Figure 4 above explains that, the lesser the model inertia (score), the better the model fit. Even though the K was optimal at 3, we can see that the model has a very high score (inertia) at this point. So, this is not a good model fit to the data. Due to the high value of inertia the data points within clusters may not be static, such randomness of data points would impact negatively the computation of reference prices. Hence to improve model fit on data vectorization (feature extraction) was used.

4.2.3 Using text vectorization to classify item clusters

Text vectorization refers to the process of converting text data into numerical representation suitable for k-means clustering. This useful information was found from statistical pattern learning. After applying some text mining techniques, like creating the corpus, which represents a collection of text documents, preprocessing (e.g., stripping white spaces, removing stop words) the corpus, and creating the term-document matrix, we were able to find the most frequent words in each cluster.

Finally, with these most frequent words per cluster, we were able to compute the words that better define the cluster, which we interpret as the cluster classification.

4.2.3.1 Vectorization

CountVectorizer was used to transform the review to the token count matrix. It tokenizes the review and according to the number of occurrences of each token, a sparse matrix is created. For instance, in cluster 1 the most frequent words were as follows; ‘rooms’, ‘banquet’, ‘conference’, ‘room’ among others.

Cluster number	Most frequent words within a cluster
Cluster 1	Rooms, banquet, conference, person, for, room, hall
Cluster 2	Maintenance, repair, light, automotive, truck, service
Cluster 3	Water, inyange, mineral, 0.5l
Cluster 4	Day, and, with, room, per, system, hall, half
Cluster 5	With, and, room, in, single
Cluster 6	Building, insurance, printer, hp, ltd, paper
Cluster 7	Pvc, galvanized, plastic, commercial, known, steel, iron

Table 2: Frequent words in each cluster

Tf-idfVectorizer (Term frequency-Inverse frequency document vectorizer) was used to transform text to feature vectors that could be utilized as input to the estimator. Tf-idf Vectorizer has a vocabulary which is a dictionary that converts each word to feature index in the matrix, each unique word gets a feature index.

Tf-idf Vectorizer aids in improving model performance and reduce the inertia score. After fitting Tf-idf Vectorizer the elbow curve was used to find the optimal cluster. In the figure 6 below we can see that after 13 clusters there's no significant decrease in WCSS (Within Cluster Sum of Squares) so, 13 cluster is the best in this case. Moreover, we can see the value of inertia has reduced compared to the preceding one. Hence using Tf-idf Vectorizer fitted cluster is more robust for finding optimal clusters for this dataset.

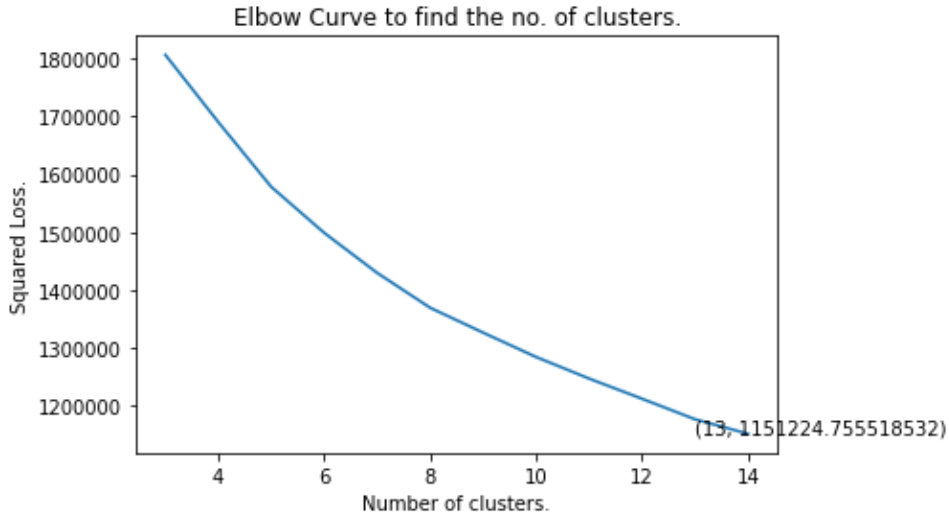


Figure 5: Elbow Curve for the optimal number of clusters

4.3 Results

4.3.1 Price trends

The items defined in the below table are in two different categories. The first category is products and the second category is services. We selected only three different transactions for each category. Thereafter, we examined the percentage of variation to prove how similar transacted item differ from the unit price. For product category, the variation percentage for the first, the second, and the third transaction were evaluated and the percentages of variation were compared. It was noted that, the first item price increased from 5% to 113%, on the second items, prices increased from 20% to 100% and the third item decreased from 150% to 14%. For the second category of services, percentage of variation presented with an increase of 40% to 42% on its first item. Another increase varied from 13% to 110% on the second item, finally, the third item was significantly increased from -84.8% to 39%. Thus, the above observations explain the need for harmonization of prices per similar item by creating a digitalized reference prices for Rwanda Public Procurement.

		Unit Price Per Transaction (Rwandan Francs)			
		1 st transaction	2 nd transaction	3 rd transaction	Percentage of Variation
Item Categorized as products	Mineral water 0.5L	500	700	1000	5% to 113%
	Maize or corn grain	168	190	400	20% to 1100%
	Diesel fuel	963	1078	1500	150% to 14%
Item Categorized as services	Meeting or banquet rooms	100000	150000	320000	40% to 42%
	Dining servers or buffets including the soft drink	4050	6000	8000	13% to 110%
	Coffee or tea with snacks	2000	5000	12000	-84.8% to 39%

Table 3: Randomly selected items in products

After pinpointing specific items, also the author realized that the unit price still had a large variance, and this discrepancy would have a huge effect when calculating the reference prices. The reason for such discrepancy was that the unit price feature had some outliers and other subtleties that may not have been captured when first looking at the data. Since the subtleties of unit price were difficult to catch by hand, the authors retrieved unit price range which could be thought as the rationale for the type of similar items in the unit of measure they were thinking about plus a large margin of error to account for unexpected values such as overprice and errors. Lastly, the price average was computed (see table 3). Nonetheless, once the average price was computed, the authors found out that the result was still not quite correct for some items as far as a large standard deviation is concerned especially for service categorized items. The challenge was that finding these price ranges was a hard task and some outliers had a huge impact on the average. Hence, the median was preferred to be used as the reference price.

4.3.2 Text clustering and text vectorization for creating the reference price

As it was previously explained, one of the objectives of this study is to show how clustering and text vectorization methods are appropriate in creating the reference price per item. The text clustering method and two different types of text vectorization methods were tested together. K-means clustering algorithm was the best model due to its effectiveness in segregating related elements into different combined groups based on the semantics similarity (Singh and Sashi, 2019).

Firstly, the data normalization was applied to the item identification names to find the optimal number of clusters that are reliable in the calculation of reference prices. This was achieved by executing k-means clustering with a range of 20 clusters and the within-cluster sum of squares (WSS) inertia scores were measured to investigate how coherent clusters are. The final decrease of WSS inertia is $4.776891258683954e+18$ the third cluster, so 3 was the best cluster, hence the smaller the better, but the higher WSS inertia is the indication of how not internally coherent clusters are. So, text normalization was not reliable in clustering similar items and in the computation of reference price.

Alternatively, we investigated text vectorization namely `CountVectorizer` and `TfidfVectorizer` so as to find appropriate method. The `CountVectorizer` applied to the text data, specifically on item identification name, using encoded text data from `CountVectorizer`, k-means clustering model fitted by iterating through a range of 15 clusters. The silhouette score computed again to interpret and validate consistency within a cluster of data. Having a silhouette score of 0.0151 shows that consistency between item clusters are still not consistent to segregate groups of similar items.

To improve the model performance and reduce the model inertia. `TfidfVectorizer` with k-means clustering introduced. The `TfidfVectorizer` produces scores that increase proportionally by the count of a particular word appearing in a given document (term frequency) and is neutralized by the count (inverse-document frequency) of the total number of the documents in the corpus (Kumar Singh and Sashi, 2019). After applying `TfidfVectorizer`, we computed the within a cluster of data to validate the consistency of the model. Based on the silhouette score of 0.0175, there exists a significant increase in the

model performance compared to CountVectorizer method. After, we executed the elbow method to determine the optimal number of clusters. It was clear that at 13th cluster there is no significant decrease in WSS (within cluster sum of squares error). So, 13 is the best in this experiment. Therefore, we confirmed that, the tf-idfVectorizer with k-means clustering is more robust for finding reliable reference price for similar services or products.

4.3.3 Computing reference prices

As explained beforehand, the main problem is to find a group of the item that describes the same item (e.g. Mineral water (0.5 L)) and in the same unit. Instinctively, we used price range for such tasks, though, selecting the correct range and extenuating the reason for that was not trivial. The item identification name and unit price for procurement from 2018 to 2019 depicted how clustering was used.

4.3.3.1 Reference prices for product category

Product categorized items						
Cluster Representation	Price range	Standard Deviation(Rwf)	Minimum price(Rwf)	Maximum price(Rwf)	Average Price (Rwf)	Reference prices
Mineral water(0.5L)	[500, 1700]	378.852757	500	1700	1000	966.7
Maize or corn grain per kg	[168, 2000]	213.43619	168	2000	273.8	225
Diesel Fuel	[960, 1500]	150.875522	960	1500	1087.38	1037
Steel road	[2500, 165000]	24839.659403	2500	165000	21333.3	15000
Medium tents	[7500, 1250000]	263696.96	7500	1250000	282911.2	180000

Table 4: Reference Price for items in the product category

Table 4 above, describes the analysis done on clusters of purchased items in the product category. The first column of the table is a cluster representation part represented by words that better define each cluster. The next column is the price ranges obtained from items that belong in the same cluster (example. Purchases of maize or corn grain per kg). The Standard deviation shows the deviation between the minimum and maximum price for a

similar clustered item. The above items vary in their initial standard deviations especially cases where maximum prices are higher than the minimum price (example: medium tents) the standard deviation is 263696.96 compared to 378.852757 of mineral water (0.5L). Considering quite a large standard deviation between prices of similar items, there is need for consolidated reference prices to reduce the standard deviation to 0 or closer to 0. It is in this regards, we computed average based on the price range of similar items, the aim was to test whether it is the right metrics to be adopted while creating reference prices for other items procured through an online e-procurement system. The average price turns out abnormal compared to the existing market price plus other associated costs. Due to the above-mentioned abnormality, we examined the median price to make a comparison. Afterward, we found that the median performs better than the average price, especially cases where large standard deviation is concerned.

4.3.3.2 Reference prices for service category

Service categorized items						
Cluster Representation	Price range	Standard Deviation(Rwf)	Minimum price(Rwf)	Maximum price(Rwf)	Average Price (Rwf)	Reference prices (Rwf)
Meeting or banquet rooms	[35000, 6000000]	711855.57541	35000	6000000	390855	180000
Dining servers or buffets	[3500, 10000000]	1.359230e+06	3500	10000000	375702	12000.01
Single room	[10000, 1200000]	152746.38102	10000	1200000	79714.6	38000
Coffee or tea break with snacks	[1050, 30000000]	2.758692e+06	1050	30000000	294614.2	3500

Table 5: Reference price for item categorized in services

Similarly, we selected items from the service category to apply the same methodologies. It was shown that services related items have a very large standard deviation than product related items. The standard deviation mainly caused by the greater difference between minimum prices and maximum prices for items in the same cluster. In the application of the proposed methodology, we considered overpriced items as outliers.

Then, we computed the average reference prices based on the range of similar items. The result was not quite reliable as described previously. In table 5, coffee or tea break with snacks has a maximum price of 30000000 Rwf which resulted in the average price of 294614.2 Rwf. This is the main reason we evaluated the median of 3500 Rwf per this item. Based on the good result obtained from the median price for both products and services related items, we confirmed that, the reference price for similar items procured through the e-procurement are quite reliable when the median is used. Based on the above results, we confidently confirm that the use of k-means clustering and text vectorization methods are significantly reliable while creating digitalized references price for Rwanda Public Procurement. In addition, the proposed approach allows flexible updates on the reference prices quarterly. This means that a situation where an item varied in the price of raw materials or affected by the inflation rate, the digitalized approach will allow easy updates after a given quarter. Based on the information available from the online e-procurement system, approximately, 4045 items were procured within a quarter. This number proves the availability of data required while checking the deviation from the proposed reference prices.

5.0 CONCLUSIONS AND RECOMMENDATIONS

5.1 Conclusions

This research aimed to develop an approach of creating digitalized reference prices for Rwanda public procurement. It also intended to propose an appropriate text clustering and text vectorization techniques reliable in creation of reference price approach for Rwanda Public Procurement. The efficiency of text vectorization known as tf-idfVectorizer and countVectorizer, for taking the similarity of items was investigated by clustering the vectorized items using k-means algorithm. The results obtained upon computation of reference prices by using data retrieved from online e-procurement system database are in favor of tf-idfVectorizer, because it fitted k-means clustering with more robust for finding optimal clusters required in computation of reference prices. The average price per cluster was calculated to be the reference price. However, the average prices had an abnormality caused by large standard deviation between ranges of prices. To avoid the abnormality in prices, we computed the median price per cluster of similar items and we found that it is reliable when compared to the market prices.

In general, data obtained from the online e-procurement system indicate to be a tool that proves how essential information about reference prices can be retrieved and refined to improve social control and provide solid accountability of Rwanda Public Procurement. We also believe that, this research contributed to Rwanda public procurement in terms of digitalization. And its findings can serve as the basis for future research projects and the continuation of enhancement in public procurement practices.

5.2 Recommendations and future research

The research revealed the reference prices as the important tool to be adopted in evaluation and auditing of public tenders. On this basis, RPPA should build on the findings of this study to implement digitalized reference price database. The author of this study investigated the efficiency of using text data mining techniques known as k-means clustering and text vectorization that group items in the same cluster to obtain associated reference prices. As future extension to this research, the authors propose to apply supervised machine learning algorithm such as feature engineering and predictive analytics to extract more insights from e-procurement data.

REFERENCES

- Essig and Arnold. (2001). Electronic Procurement in Supply Chain Management: An Information Economics-Based Analysis of Electronic Markets. *Journal of Supply Chain Management*, 43-49.
- Gun Lim et al. (2008). PUBLIC E-PROCUREMENT: THE KOREAN ON-LINE EPROCUREMENT SYSTEM (KONEPS). *3rd INTERNATIONAL PUBLIC PROCUREMENT CONFERENCE PROCEEDINGS* (pp. 744-757). Seoul: unpcdc.
- World Bank Group. (2016). *Rwanda: Pioneering e-procurement in Africa*. Malaysia: Global knowledge and research hub in Malaysia.
- Adani, R. C. (2016). *Reference prices in the Italian procurement for*. ROME: Editorial Express.
- AGESHIN, E. A. (2001). *E-PROCUREMENT AT WORK*. Chicago: Northeastern Illinois University.
- Bobowski and Gola. (2018). E-Procurement in the European Union. *Asia-Pacific Journal of EU Studies*, 13-35.
- Bosio and Djankov. (2020, February 05). *How large is Public Procurement?* Retrieved from World Bank Blog: <https://blogs.worldbank.org/developmenttalk/how-large-public-procurement>
- Carvalho et al. (2014). Using Clustering and Text Mining to Create a Reference Price Database*. *Learning and Nonlinear Models - Journal of the Brazilian Society on Computational Intelligence (SBIC)*, 38-52.
- Chaitrong, W. (2017). Project reference prices up for review. *The Nation*, 1-2.
- Chopra et al. (2001). B2B e-Commerce Opportunities. *SUPPLYCHAINMANAGEMENTREVIEW*, 50-58.
- E. Baker and m.Sinkula. (1999). The synergistic effect of market orientation and learning orientation on organizational performance. *Academy of Marketing Science Journal*, 411.
- Eduardo de Paiva et Al. (2014). Using Clustering and Text Mining to Create a Reference Price Database. *Journal of the Brazilian Society on Computational Intelligence (SBIC)*, 38-52.
- European Commission. (2020, June 15). *Internal Market, Industry, Entrepreneurship, and SMEs*. Retrieved from European Commission: https://ec.europa.eu/growth/single-market/public-procurement/digital_en

- Frank and Daniela. (2003). Production, consumption, and general equilibrium with physical constraints. *Journal of Environmental Economics and Management*, 513-538.
- Gouveia, R. d. (2002). The Price Factor in EC Public Tenders. *Public Contract Law Journal*, 679-693.
- Hennig, C. (2002). Fixed Point Clusters for Linear Regression: Computation and Comparison. *Journal of Classification*, 249-276.
- Marchi, R. C. (2016). *Reference prices in the Italian procurement for medical devices*. Rome: EditorialExpress.
- Maxwell Asenso-Boakye, D. E. (2014). Irregularities in Ghana's Public Sector Procurement and Their Possible Reinforcers: A Study of the Auditor General's Report. *International Journal of Economics, Commerce, and Management, United Kingdom Vol. II, Issue 2, 2014*, 15.
- McCue and Roman. (2012). E-PROCUREMENT: MYTH OR REALITY. *JOURNAL OF PUBLIC PROCUREMENT*, 212-238.
- McLeod. (2019, July 19). *What does a box plot tell you?* Retrieved from simplypsychology.org: <https://www.simplypsychology.org/boxplots.html>
- Mulugeta Dinka. (2016, My 12). *Breaking the glass ceiling in Africa: Rwanda E-Government Procurement System*. Retrieved from The World Bank Procurement Framework: <https://wbnpf.procurementinet.org/featured/breaking-glass-ceiling-africa-rwanda-e-government-procurement-system>
- National bank of Rwanda. (2020, May 23). *Consumer Price Index, February 2020*. Retrieved from National Bank of Rwanda: www.nbr.rw
- OECD. (2020, July 11). *Public Procurement*. Retrieved from OECD: <https://www.oecd.org/gov/public-procurement/>
- Peria. (2003). Foreign Bank Entry: Experience, Implications for Developing Economies, and Agenda for Further Research. *The World Bank Research Observer* (pp. 25-59). Oxford: The world bank Research observer.
- Presutti. (2003). Supply management and e-procurement: creating value-added in the supply chain. *Industrial marketing management*, 219-226.
- Rwanda Public Procurement Authority. (2017). *Survey Report on the RPPA Services, Law on Public Procurement, and the use of Reference Price*. Kigali: Rwanda Public Procurement Authority.
- Rwanda Public Procurement Authority. (2019). *Reference prices*. Kigali: RPPA.

- Salazar and Harper. (2018, SEPTEMBER 07). *Public Procurement: A Journey Towards the Digital Frontier*. Retrieved from RECAUDANDO BIENESTAR: <https://blogs.iadb.org/gestion-fiscal/en/public-procurement-digitalization/>
- Simon Croom et Al. (2000). Supply chain management: an analytical framework for a critical literature review. *European Journal of Purchasing & Supply Management*, 67-83.
- Thai, K. V. (2017). *Global Public Procurement Theories and Practices*. Florida: Springer International Publishing.
- Transparency International Rwanda(TIR). (2016). *Analysis of Auditor General Report*. Kigali: Transparency International Rwanda(TIR).
- Wikipedia. (2020, August 15). *Bag of Words Model*. Retrieved from www.wikipedia.org: https://en.wikipedia.org/wiki/Bag-of-words_model
- Yixin Fang, J. W. (2012). Selection of the number of clusters via the bootstrap method. *Computational Statistics & Data Analysis*, 468-477.
- Zong. (2013, February 02). *K Means Clustering with Tf-idf Weights*. Retrieved from jonathanzong.com: <https://jonathanzong.com/blog/2013/02/02/k-means-clustering-with-tfidf-weights>
- He, Y., Fataliyev, K. and Wang, L., 2013, November. Feature selection for stock market analysis. In *International conference on neural information processing* (pp. 737-744). Springer, Berlin, Heidelberg.

Plagialism_Check_Final

ORIGINALITY REPORT

18%

SIMILARITY INDEX

16%

INTERNET SOURCES

4%

PUBLICATIONS

7%

STUDENT PAPERS

PRIMARY SOURCES

1	www.kaggle.com Internet Source	3%
2	ejeg.com Internet Source	1%
3	www.inlm-sbic.org Internet Source	1%
4	towardsdatascience.com Internet Source	1%
5	www.lbd.dcc.ufmg.br Internet Source	1%
6	ippa.org Internet Source	1%
7	ec.europa.eu Internet Source	1%
8	thesai.org Internet Source	1%
9	jonathanzong.com Internet Source	1%

10	docplayer.net Internet Source	1%
11	Submitted to essex Student Paper	1%
12	hdl.handle.net Internet Source	<1%
13	scholar.mzumbe.ac.tz Internet Source	<1%
14	Submitted to Mount Kenya University Student Paper	<1%
15	Submitted to University for Development Studies Student Paper	<1%
16	Submitted to American Sentinel University Student Paper	<1%
17	Tripathy, Abinash, Ankit Agrawal, and Santanu Kumar Rath. "Classification of Sentimental Reviews Using Machine Learning Techniques", <i>Procedia Computer Science</i> , 2015. Publication	<1%
18	Submitted to Ghana Technology University College Student Paper	<1%
19	Submitted to Zambia Centre for Accountancy Studies Student Paper	<1%

20	Submitted to Chulalongkorn University Student Paper	<1%
21	Submitted to National University of Singapore Student Paper	<1%
22	Submitted to Atlantic International University Student Paper	<1%
23	link.springer.com Internet Source	<1%
24	www.coursehero.com Internet Source	<1%
25	core.ac.uk Internet Source	<1%
26	Submitted to University of East London Student Paper	<1%
27	www.ijrte.org Internet Source	<1%
28	citeseerx.ist.psu.edu Internet Source	<1%
29	www.theijbm.com Internet Source	<1%
30	www.slideshare.net Internet Source	<1%
31	Submitted to Southampton Solent University Student Paper	<1%

32 pt.scribd.com Internet Source <1%

33 blogs.worldbank.org Internet Source <1%

34 pdfs.semanticscholar.org Internet Source <1%

35 epubl.luth.se Internet Source <1%

36 www.ijcaonline.org Internet Source <1%

37 repository.out.ac.tz Internet Source <1%

38 bada.hb.se Internet Source <1%

39 www.rug.nl Internet Source <1%

40 ir.tum.ac.ke Internet Source <1%

41 E COX. "Fuzzy Clustering", Fuzzy Modeling and Genetic Algorithms for Data Mining and Exploration, 2005
Publication <1%

42 www.scribd.com Internet Source <1%

43 Rika Koch. "Green Public Procurement under WTO Law", Springer Science and Business Media LLC, 2020 <1%
Publication

44 wdsinet.org <1%
Internet Source

45 www.grin.com <1%
Internet Source

46 etd.aau.edu.et <1%
Internet Source

47 "Social Networks: A Framework of Computational Intelligence", Springer Science and Business Media LLC, 2014 <1%
Publication

Exclude quotes On

Exclude matches Off

Exclude bibliography On