MAKERERE UNIVERSITY

**SEMESTER ONE 2024/2025 ACADEMIC YEAR**

**SCHOOL COMPUTING AND IMFORMATICS TECHNOLOGY**

**DEPARTMENT OF COMPUTER SCIENCE**

**MASTER OF SCIENCE IN COMPUTER SCIENCE**

**MCS 7103**

**MACHINE LEARNING**

**STEVEN MAGEZI**

**2400721935**

**2024/HD05/21935U**

## Introduction

Energy consumption has become increasingly critical for utilities and policymakers as they strive to optimize energy distribution, enhance energy efficiency, and improve reliability of electricity. The exploration in this report looks to identify relationships between energy consumption, energy revenues, customer characteristics and energy revenue collection channels by focusing on identifying patterns and trends that can help energy utilities better understand the consumption behavior.

The data set used in this exploration is billing data from the largest energy distribution company in Uganda called UMEME UGANDA covering the one of the busiest energy consuming electoral district called Wandegeya district for a period running between 13th December 2023 to 02nd January 2024.

The analysis considers multiple factors including customer category, billing channel, cash tendered, Cost of energy, Total units, Lifeline Units and above lifeline units.

## Questions before EDA

1. How much data do I need to have accurate findings?
2. What are the energy consumption categories and how many are they?
3. How many channels are open to receive energy revenue from customers?
4. What is the difference between cash tendered and cost of energy
5. What are life line units?

## Questions after EDA

1. How does service charge affect energy consumed?
2. Which parameters are most important in improving energy consumption?
3. Can increasing lifeline units increase energy revenue and consumption
4. Which period of a calendar month is energy revenue highest
5. What is the correlation between energy consumed and customer category

## Data Wrangling

This process involved various stages at which the data was exposed to as listed below;

Imported libraries pandas, numpy, matplotlib and seaborn

Read the data from the its source in google colab into the data frame

I described the structure of my data set to identify null values, data type of the data set and naming convention of the column names

listed the first 5 data rows in my data set to have visibility of how the data appears in the data frame

Using the describe command, I gained a statistical overview of the major numerical columns in my dataset to help me identify the underlying patterns and anomalies in the data set

I also identifed null values and duplicates in my data set to either correct the data set by removing data with null values and deleting duplicates or filling null values with dummy test data that is closer to the actual data.

I lastly analyzed the categorical data columns and their frequency of occurrences.

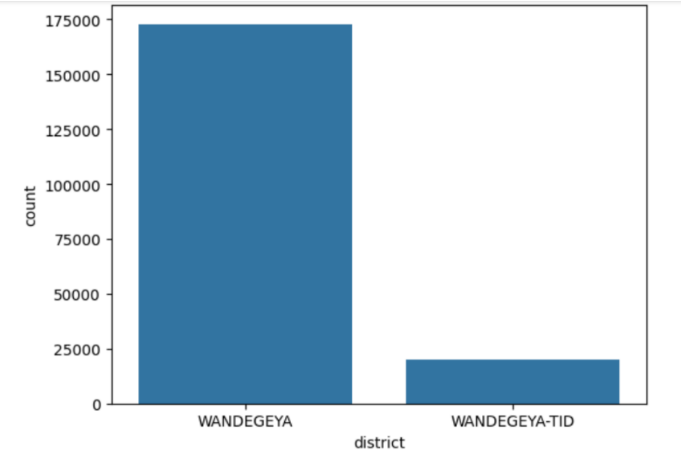
## Exploratory Data Analysis

EDA analysis was performed in in 3 categories below;

### Univariate Analysis

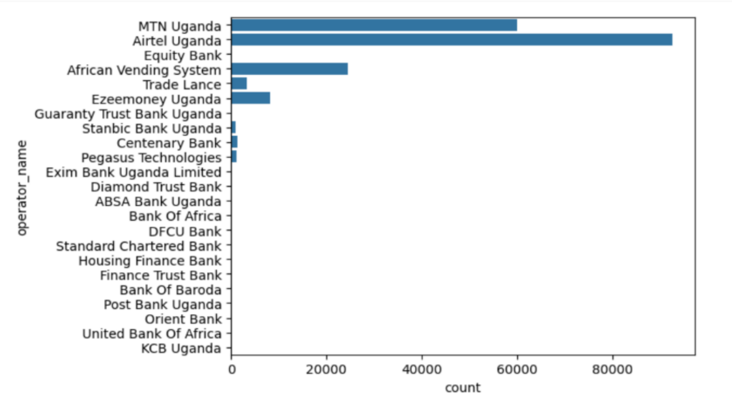
Using this anaylsis I reviewdd the distribution of the main influencial columns in my data set to identify the patterns and behavior both the categorical columns and numerical columns

Fig 1



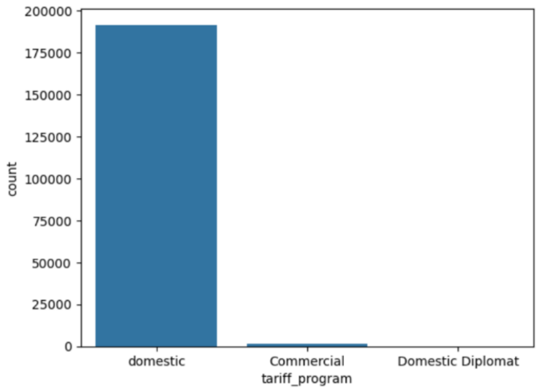
From the histogram fig 1, it is observed that Wandegeya district contributes more to the energy consumed compared the its sister district Wandegeya-TID which is a district that has just been created by umeme to help curd power shortages

Fig 2



From the histogram in fig 2 , it is observed that customer use Airtel Network to purchase energy which is followed by MTN

Fig 3

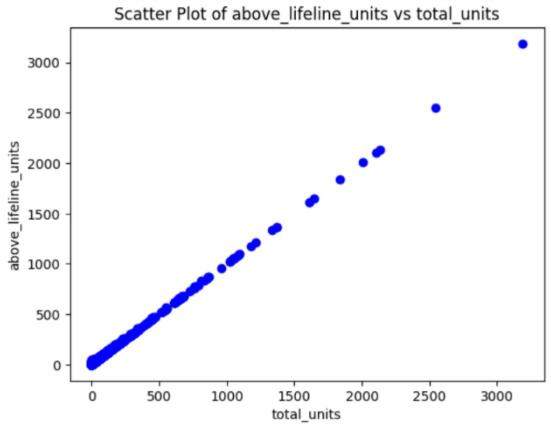


From figure 3, it is observed that domestic customers in Wandegeya district consume more energy and generate the highest revenue.

**Bivariate analysis**

This analysis was done to identify the relationships between the different parameters of the data set. As shown in figures below

Fig 1



The scatter plot in fig 1, indicates a very close relationship between the total number of units purchased by a customer and the number of units this customer receives above the life line campaign units.

Fig 2

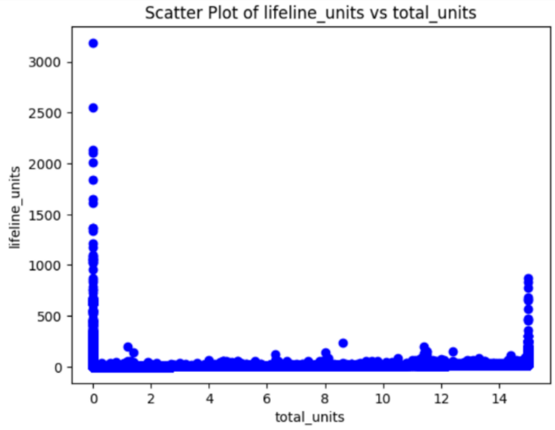


Fig 2 indicates a relationship between the total units purchased by a customer to the lifeline units given to this customer. This shows that life line units are more close to the number of total units purchased by the customer which indicates that customers purchase little energy since lifeline units are 15units a month.

Fig 3

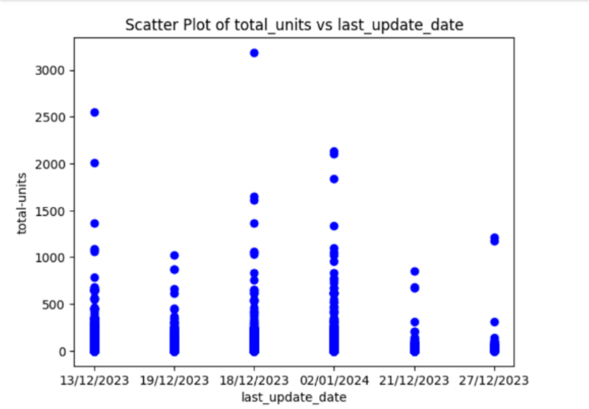
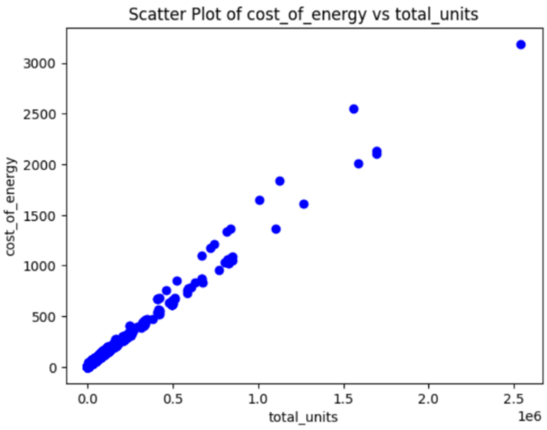


Fig 3 it is observed that that more energy is purchased by customers at the beginning of the month and mid months days.

Fig 4

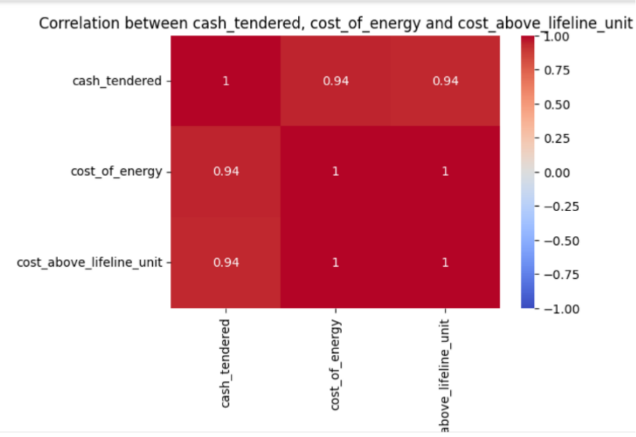


Form fig 4, it is observed that there is a close relations hip between total units received by customer with the cost of energy collected from this customer, but the correlation separates wider as both total units and cost of energy increases.

**Multivariate Analysis**

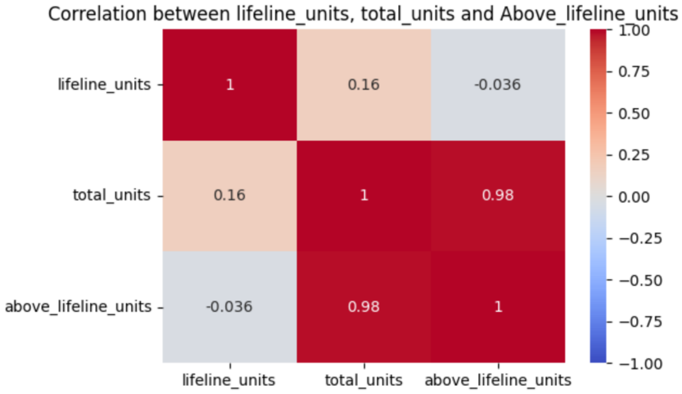
This analysis was done to identify the relationships and patterns in all revenue parameters as a group and all energy consumed or purchased parameters as a group from the data set as shown in the figures below;

Fig 1,



From Fig 1, it is observed that all cash parameters are strongly connected and dependent on each other from the heat map colours in fig 1. Thay all lie in the positive area of the map which indicates a close correlation.

Fig 2



From Fig 2, it is observed that the correlation between energy parameters is weak which may result from the small amounts of energy purchased by customers in relation to the set lifeline units of 15.

## Conclusions

1. The energy purchased by most customers is almost equal or less than the lifeline units reserved for a discount
2. Domestic customers category contributes to the highest count of purchases
3. Airtel network is the most used network by customers to purchase electricity
4. Electricity is most purchased at the beginning of the month and mid-month days