DOCUMENTATION FOR THE MALARIA

Table of Contents

[Problem Characterization 1](#_Toc127709766)

[Data Collection 1](#_Toc127709767)

[Data Preparation and Exploration 2](#_Toc127709768)

[Data Preparation 2](#_Toc127709769)

[Data Exploration 4](#_Toc127709770)

[Data Cleaning 4](#_Toc127709771)

[Feature Engineering 5](#_Toc127709772)

[Model Development and Optimization 5](#_Toc127709773)

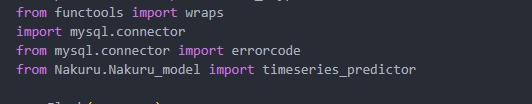
[Model Evaluation 8](#_Toc127709774)

[Model Deployment 8](#_Toc127709775)

[Model Deployment 9](#_Toc127709776)

[Challenges. 9](#_Toc127709777)

[References 10](#_Toc127709778)

MODEL

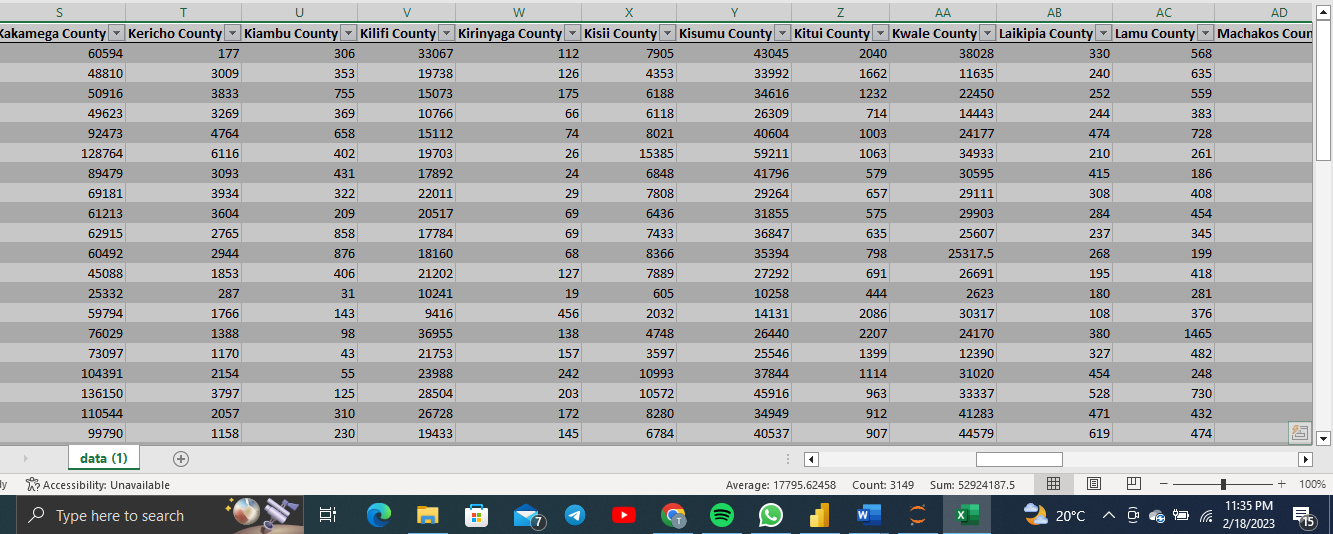
# Problem Characterization

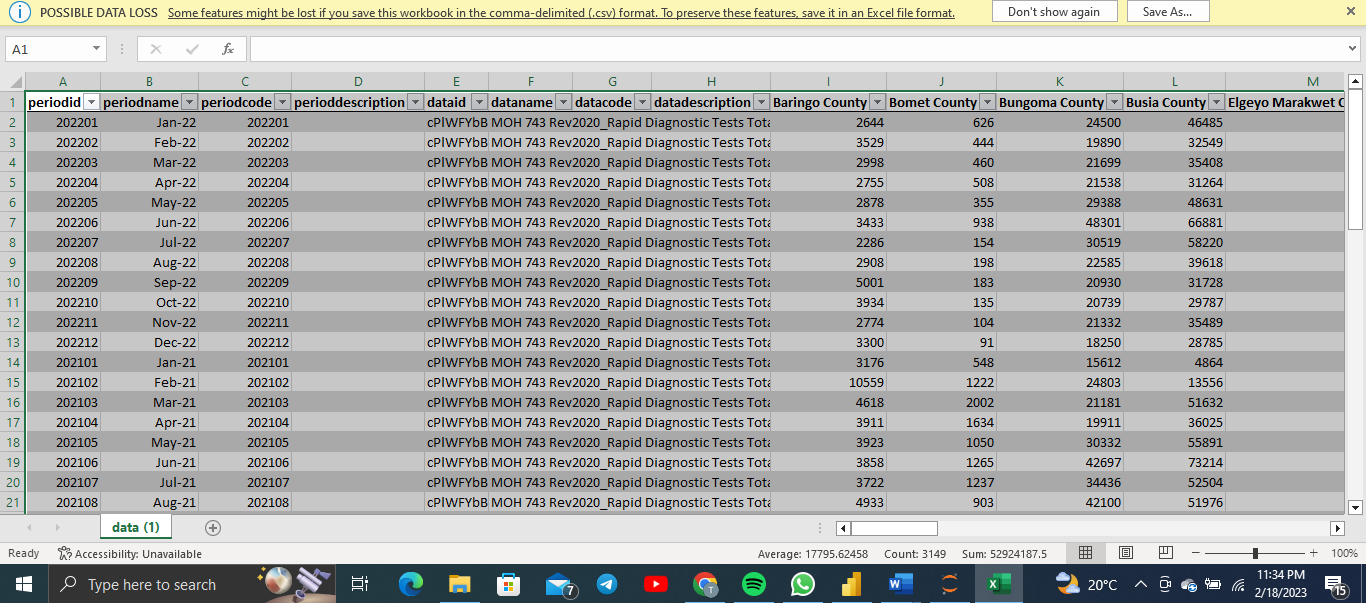
The problem was presented to us by the client (KHIS), through scheduled meet-ups

The client proceeded to explain that they are in need of a model that does demand prediction of malaria commodities in order to make accurate restocking decision. Previously, the client stated that they had been using six-month averages to use to make restocking decisions. This created a problem as it wasn’t capturing the factors such as season and region that affected the demand for malaria commodities, therefore the restocking decision was unreliable

# Data Collection

The data was provided to us by our client (KHIS).

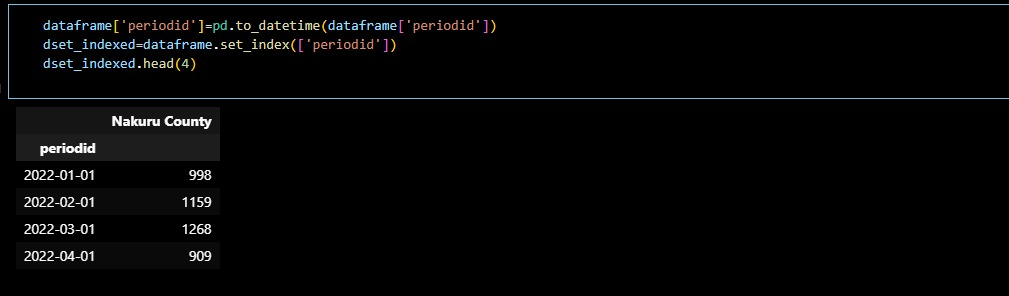
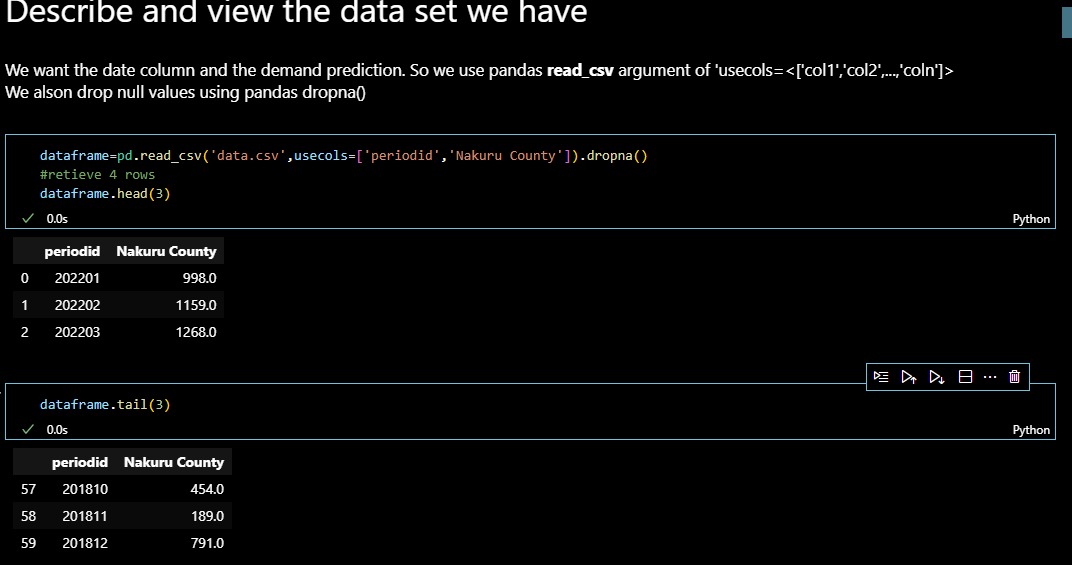
The data was provided to us in the format of a table that contained records of data. The table compromised of each month of the years (2018-2022) and the supply that has been done during that time period in each county.

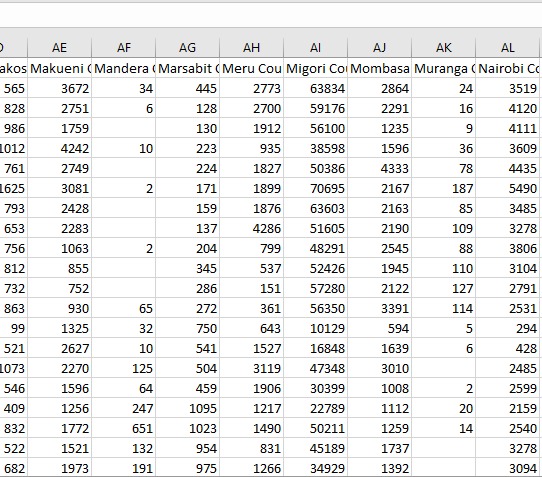


# Data Preparation and Exploration

## Data Preparation

The data as provided needed shaping from the previous table format.

Let’s take an example of data preparation that was directed to Nakuru County region as shown below

A problem that was identified was cases of missing values in certain counties such as Mandera and Murang’a 

The missing values were eventually dropped.

## Data Exploration

The variables used is the periodid that was a representation of the month and year put together against each county

The format of the variables was then changed from float data type to arrays

## Data Cleaning

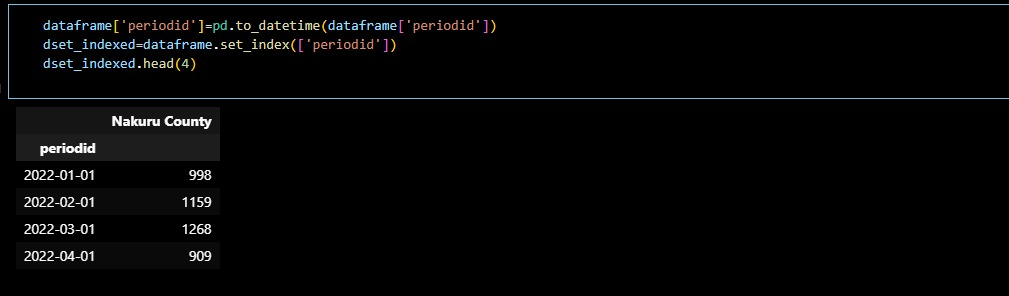
There was no much data cleaning that was performed since there were no redundance values.

Counties with missing values were not of reflection in the data.

## Feature Engineering

The format of the variables was then changed from float data type to arrays.

The data was categorized as time related feature engineering. The periodid column was then changed to be time-based

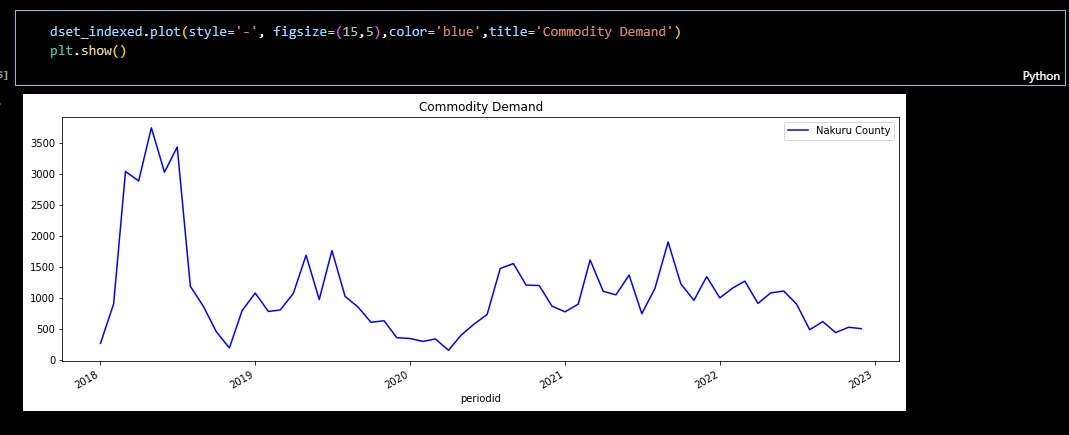


The county column was converted to an array only.

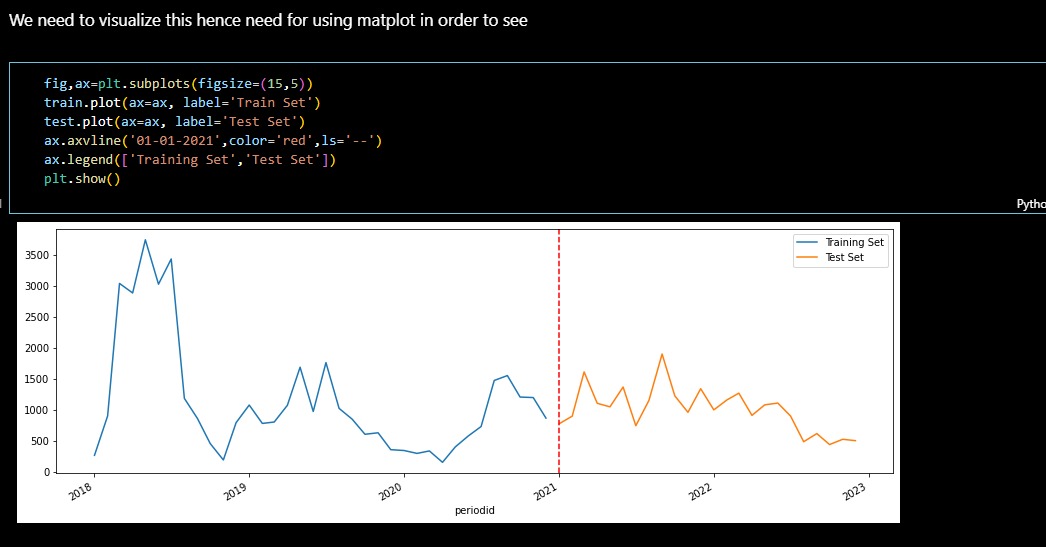
# Model Development and Optimization

Given that the data was a time series based it was converted the same as shown in the data exploration segment

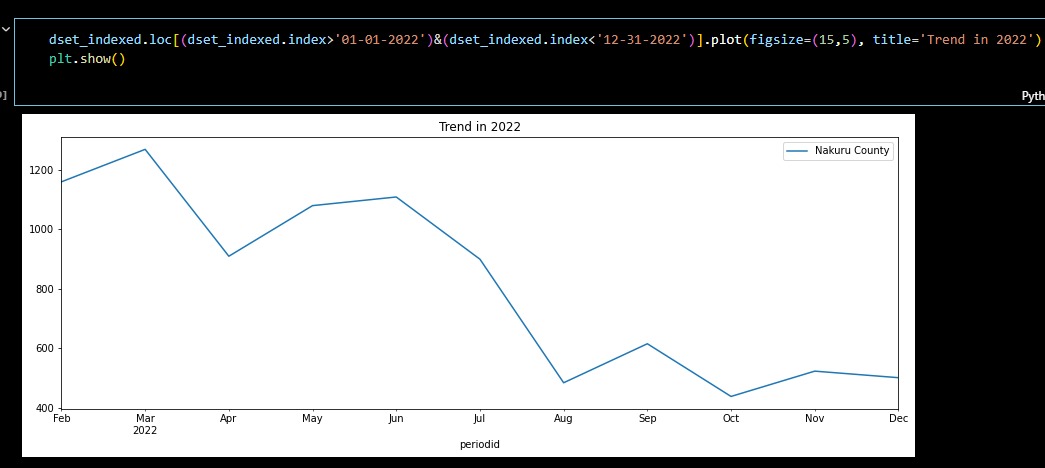
The graph obtained was in terms of year against commodity demand that was particular to Nakuru County as seen below



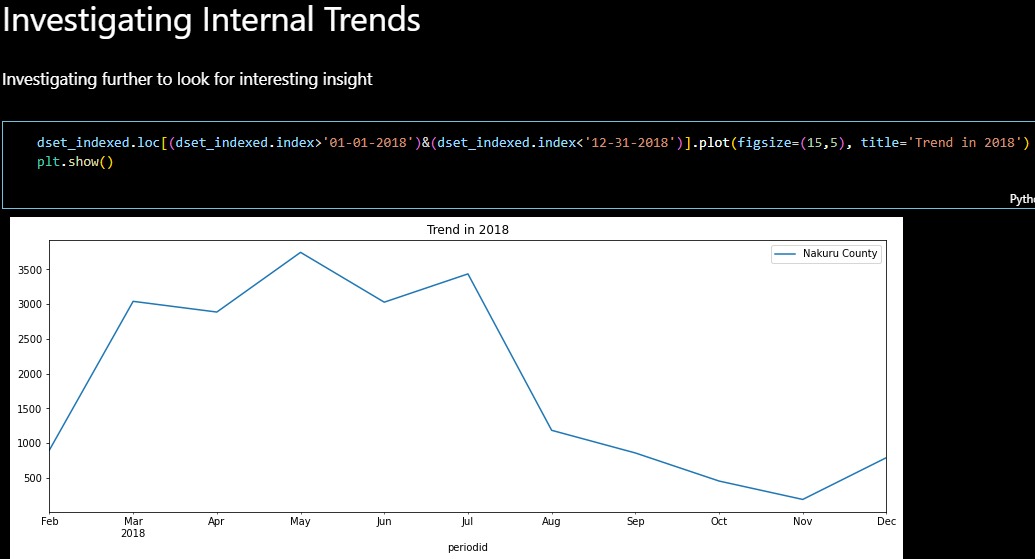
The following image shows a visualization the training and testing sets which have been split using periodid and plotted against the demand of commodities during the course of the years



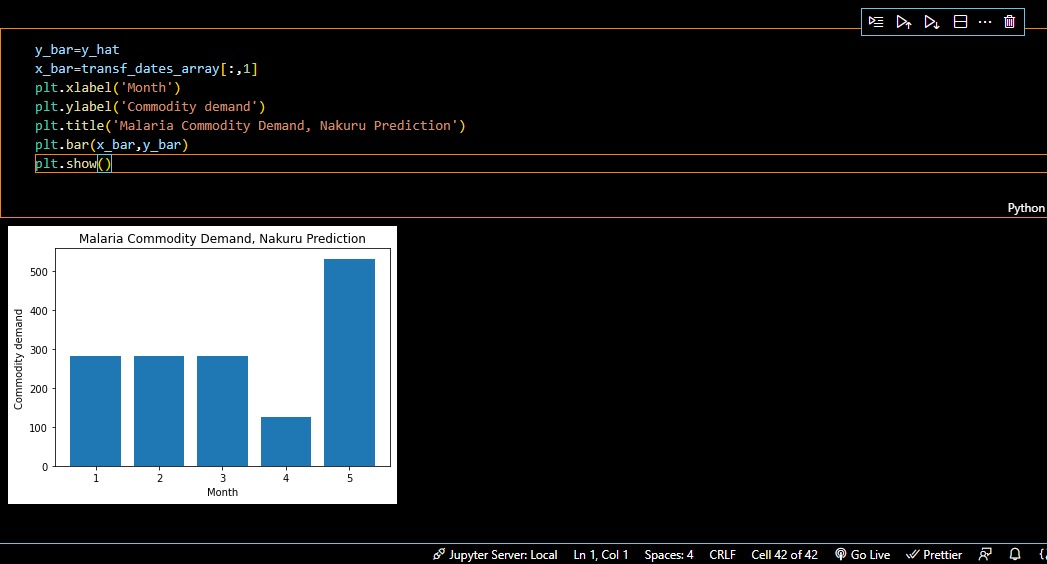
The graph below shows the trend in Nakuru County in the year 2022, in each month of the year



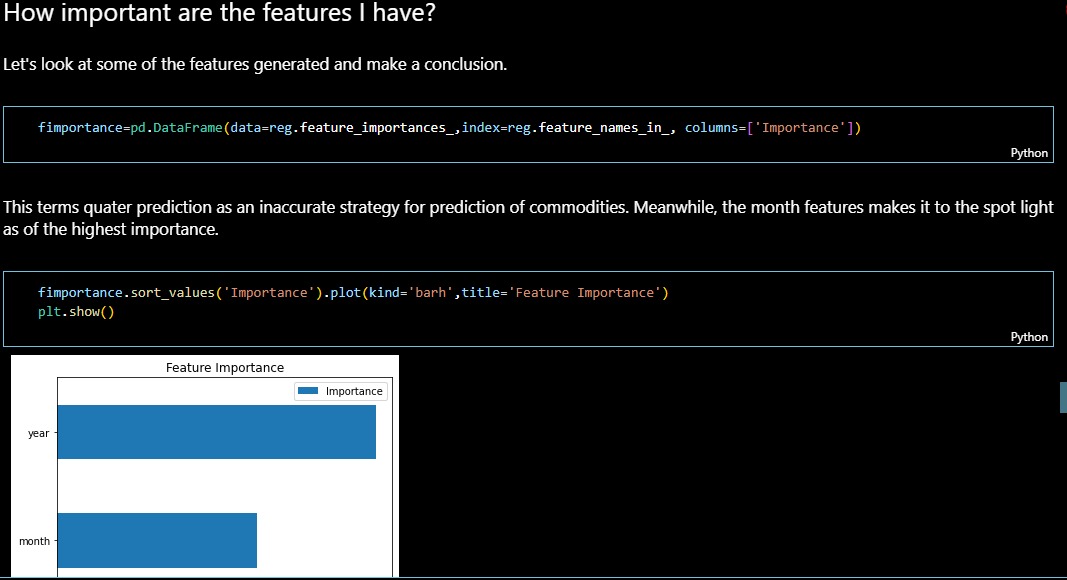
This graph shows the trend of Nakuru County in the year 2018, in each month against the demand of malaria commodities



This is a graph that plots a prediction of the malaria demand commodities for the first five months of 2022 in Nakuru County



In accordance with the features, this terms quarter prediction as an inaccurate strategy for prediction of commodities. Meanwhile, the month features make it to the spot light as the highest importance



# Model Evaluation

The metrics that was used for our model was the Root Mean Squared Error(RMSE) and Mean Squared Error(MSE) ,important thing to note is that the Regressor uses RMSE only by default because of the inconsistencies.

In accordance with the features, this terms quarter prediction as an inaccurate strategy for prediction of commodities. Meanwhile, the month features make it to the spot light as the highest importance.

We used a pandas feature that allows us to compute the importance of the features that we are to use for plotting.

XGBR is a Regression Model that works by building ensemble of decision trees, where each tree is trained to make prediction based on a subset of the available data. The final prediction is made by taking the average of the prediction from all the trees in the ensemble.

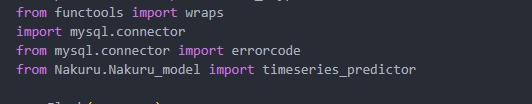
It was the used model as it worked better as compared to the others and the over-fitting was easier to identify using the cross-validation function.

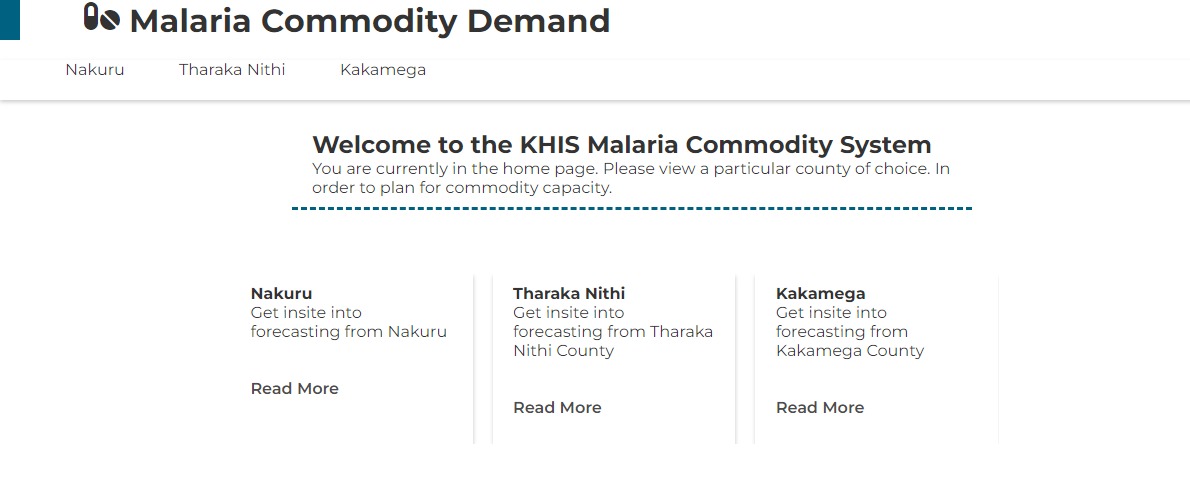
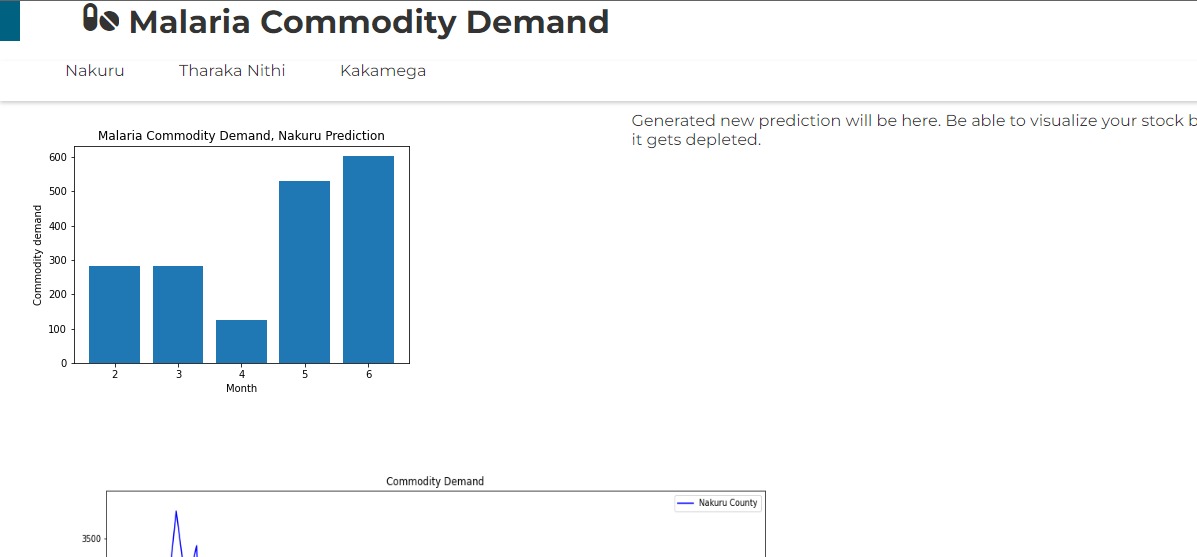
# Model Deployment

We used local deployment technique to deploy our model which means that we will be integrating our model with the application that we have built on our personal computer.

The process of deployment;

1. Extract the relevant code
2. Create a python module under a specific County
3. Initiate a Flask Application
4. Import the module in the main application file.
5. In order to do a prediction we derive our the last date record from the database based on the user input, where the input from the database itself is used as argument for the model that we have.
6. The prediction is made and stored





# Challenges.

* The data lacked detailed features to train the model effectively for example the exact date and time were not given, only the month and year.
* Inadequate time to create a full functioning system.

# References

https://scikit-learn.org/stable/auto\_examples/applications/plot\_cyclical\_feature\_engineering.html