MAKERERE UNIVERSITY

COLLEGE OF COMPUTING AND INFORMATION SCIENCES

DATA SCIENCE PROJECT RECESS REPORT

GROUP Y

ARUBE JOSHUA	**21/U/1102**
LUBEMBE COLIN WAFULA	**21/U/05828/PS**
LUYIMBAZI TIMOTHY	**21/U/18300/PS**
KATO TREVOR THOMAS	**21/U/13085/PS**
NANSUMBA MARY VANESSA	**21/U/10131/PS**

1. INTRODUCTION ABOUT PROJECT AND DATASET

This report analyzes a dataset containing statistics related to the Malaria disease, public health, water, sanitation, and demographic indicators for all countries in Africa from the year 2007 to 2017. The dataset helps to provide insights into the state of health and living conditions across these nations.

The dataset has 594 rows and 27 columns containing numerical values, objects and geographical values. It contains 0 duplicates and quite a number of null values. The dataset comprises of information from different countries, each represented by specific columns. Here are some key findings:

1. **Malaria concerned statistics:**

- Incidence of Malaria (per 1,000 population at risk): This column indicates the rate of Malaria incidence per 1,000 population at risk. It's a key metric for measuring the prevalence of malaria in a country and can help identify regions with higher disease burden.
- Malaria Cases Reported: The number of reported Malaria cases in a given country and year. This information provides insight into the magnitude of the Malaria problem in different regions.

2. **Healthcare Measures:**

- Use of Insecticide-Treated Bed Nets (% of under-5 population): This percentage indicates the proportion of children under the age of 5 who are using insecticide-treated bed nets, which are a common preventive measure against Malaria.
- Children with Fever Receiving Antimalarial Drugs (% of children under age 5 with fever): This percentage represents the portion of children under 5 years old who have a fever and are receiving antimalarial treatment. It reflects the effectiveness of healthcare interventions.
- Intermittent Preventive Treatment (IPT) of Malaria in Pregnancy (% of pregnant women): This column gives the percentage of pregnant women receiving intermittent preventive treatment for Malaria. It's essential for maternal health and preventing Malaria-related complications during pregnancy.

3. **Water and Sanitation:**

- People Using Safely Managed Drinking Water Services (% of population): The percentage of the population with access to safely managed drinking water services. This measure is related to overall public health and disease prevention.
- People Using Safely Managed Drinking Water Services, Rural (% of rural population): Similar to the previous column, this percentage specifically represents rural population access to safe drinking water.
- People Using Safely Managed Drinking Water Services, Urban (% of urban population): Similarly, this percentage represents urban population access to safe drinking water.

- People Using Safely Managed Sanitation Services (% of population): The percentage of the population with access to safely managed sanitation services. This is important for hygiene and disease prevention.
- People Using Safely Managed Sanitation Services, Rural (% of rural population): Represents rural population access to safe sanitation services.
- People Using Safely Managed Sanitation Services, Urban (% of urban population): Represents urban population access to safe sanitation services.
- People Using at Least Basic Drinking Water Services (% of Population): The percentage of the population with access to basic drinking water services, which is essential for public health.
- People Using at Least Basic Drinking Water Services, Rural (% of Rural Population): Represents rural population access to basic drinking water services.
- People Using at Least Basic Drinking Water Services, Urban (% of Urban Population): Represents urban population access to basic drinking water services.
- People Using at Least Basic Sanitation Services (% of Population): The percentage of the population with access to basic sanitation services.
- People Using at Least Basic Sanitation Services, Rural (% of Rural Population): Represents rural population access to basic sanitation services.
- People Using at Least Basic Sanitation Services, Urban (% of Urban Population): Represents urban population access to basic sanitation services.

4. **Population and Demographics:**

- Rural Population (% of Total Population): The proportion of the total population residing in rural areas. This demographic information can influence disease distribution.
- Rural Population Growth (Annual %): The annual growth rate of the rural population. Population growth can impact healthcare resource allocation.
- Urban Population (% of Total Population): The proportion of the total population residing in urban areas. Similar to rural population, this demographic information can influence disease distribution.
- Urban Population Growth (Annual %): The annual growth rate of the urban population.

5. **Geographical Information:**

- Country Name: The name of the African country for which the statistics are reported. This column will help identify individual countries and group them for analysis based on their malaria-related statistics.
- Year: The year in which the data was recorded. This temporal information is important for tracking trends and changes in malaria-related statistics over time.
- Country Code: The unique code assigned to each country.
- Latitude: The geographical latitude coordinate of the country's location. This information is important for spatial analysis and visualization.
- Longitude: The geographical longitude coordinate of the country's location. Also crucial for spatial analysis and visualization.
- Geometry: Spatial geometry information that could represent the geographical shape of the country. This is useful for creating maps and conducting spatial analyses.

2. KEY OBJECTIVES FOR ANALYZING THE DATA

GENERAL OBJECTIVES

- To analyze the trends in the number of Malaria cases in African countries from the year 2007 to 2017.

- To analyze the changes in the Incidence of Malaria in African countries from the year 2007 to 2017.

- To study and compare the impact of the columns of the dataset towards the number of Malaria cases reported.

- To determine which country had the highest average of reported cases of Malaria from the year 2007 to 2017.

- To determine the year with the highest number of reported Malaria cases in Africa.

- To determine the region with the highest number of Malaria cases reported.

- To analyze the changes in the Rural and Urban population rates in African countries over the years 2007 to 2017.

- To compare the water and sanitation services between the Urban and Rural population in African countries.

- To determine the year with the highest number of children receiving antimalarial drugs

- To analyze the relationship between the number of reported Malaria cases and the percentage of the Rural and Urban population in African countries.

3. THE FEATURES ANALYZED

- Incidence of malaria (per 1,000 population at risk): Analyzing this column can help identify countries with higher rates of Malaria transmission, allowing for targeted interventions and resource allocation.

- Malaria cases reported: Understanding reported cases will provide a sense of the disease burden in each country.

- People using at least basic drinking water services, rural (% of rural population) and People using at least basic drinking water services, rural (% of urban population): This percentage reflects the levels of access to the basic drinking water services in the rural and urban areas which allows us to compare between the two sections for each country.

- People using at least basic sanitation services, urban (% of urban population) and People using at least basic sanitation services, rural (% of rural population): These percentages reflect access to proper sanitation facilities, contributing to health and hygiene in the urban and rural areas in the African countries.

- Rural population (% of total population) and Urban population (% of total population): Analyzing these columns offers insights into the level of urbanization and population distribution in each country and how it relates to the Malaria concerned Statistics in the dataset.

- Rural population growth (annual %) and Urban population growth (annual %): These growth rates provide an understanding of demographic shifts and urbanization trends.

- Latitude and Longitude: Geographical coordinates are used for mapping and spatial analysis to visualize how these indicators vary geographically.

4. PROCESSES AND TECHNIQUES

a. Data Exploration

This involved analyzing the dataset to identify the features and familiarize with them such that we can do data cleaning and preparation.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 594 entries, 0 to 593
Data columns (total 27 columns):
 #   Column                                                                    Non-Null Count  Dtype
---  ------                                                                    --------------  -----
 0   Country Name                                                              594 non-null    object
 1   Year                                                                      594 non-null    int64
 2   Country Code                                                              594 non-null    object
 3   Incidence of malaria (per 1,000 population at risk)                       550 non-null    float64
 4   Malaria cases reported                                                    550 non-null    float64
 5   Use of insecticide-treated bed nets (% of under-5 population)             132 non-null    float64
 6   Children with fever receiving antimalarial drugs (% of children under age 5 with fever)  122 non-null  float64
 7   Intermittent preventive treatment (IPT) of malaria in pregnancy (% of pregnant women)  106 non-null  float64
 8   People using safely managed drinking water services (% of population)     99 non-null     float64
 9   People using safely managed drinking water services, rural (% of rural population)  88 non-null  float64
 10  People using safely managed drinking water services, urban (% of urban population)  176 non-null  float64
 11  People using safely managed sanitation services (% of population)         132 non-null    float64
 12  People using safely managed sanitation services, rural (% of rural population)  110 non-null  float64
 13  People using safely managed sanitation services, urban  (% of urban population)  132 non-null  float64
 14  Rural population (% of total population)                                  588 non-null    float64
 15  Rural population growth (annual %)                                        588 non-null    float64
 16  Urban population (% of total population)                                  588 non-null    float64
 17  Urban population growth (annual %)                                        588 non-null    float64
 18  People using at least basic drinking water services (% of population)    588 non-null    float64
 19  People using at least basic drinking water services, rural (% of rural population)  566 non-null  float64
 20  People using at least basic drinking water services, urban (% of urban population)  566 non-null  float64
 21  People using at least basic sanitation services (% of population)         588 non-null    float64
 22  People using at least basic sanitation services, rural (% of rural population)  566 non-null  float64
 23  People using at least basic sanitation services, urban  (% of urban population)  566 non-null  float64
 24  latitude                                                                  594 non-null    float64
 25  longitude                                                                 594 non-null    float64
 26  geometry                                                                  594 non-null    object
dtypes: float64(23), int64(1), object(3)
memory usage: 125.4+ KB
```

Figure 1 Datatypes of columns in Malaria dataset

From Figure 1 above, we see that a majority of our data is numerical including twenty-three columns with float values, one column with integers and three columns with the datatype object.

8]:

	count	mean	std	min	25%	50%	75%	max
Year	594.0	2.012000e+03	3.164943e+00	2007.000000	2009.000000	2012.000000	2.015000e+03	2.017000e+03
Incidence of malaria (per 1,000 population at risk)	550.0	1.900875e+02	1.630545e+02	0.000000	30.857500	174.775000	3.476375e+02	5.855400e+02
Malaria cases reported	550.0	1.068330e+06	2.192802e+06	0.000000	2211.750000	113026.000000	1.154808e+06	1.682113e+07
Use of insecticide-treated bed nets (% of under-5 population)	132.0	4.253030e+01	2.015706e+01	1.000000	26.675000	42.900000	5.932500e+01	9.550000e+01
Children with fever receiving antimalarial drugs (% of children under age 5 with fever)	122.0	3.020164e+01	1.890320e+01	0.500000	17.275000	29.300000	4.262500e+01	7.690000e+01
Intermittent preventive treatment (IPT) of malaria in pregnancy (% of pregnant women)	106.0	1.501396e+01	1.238917e+01	0.000000	5.763285	11.500000	2.185000e+01	5.960000e+01
People using safely managed drinking water services (% of population)	99.0	3.347899e+01	2.667832e+01	5.770000	8.975000	28.390000	4.389000e+01	9.266000e+01
People using safely managed drinking water services, rural (% of rural population)	88.0	1.247057e+01	1.007837e+01	0.930000	4.185000	10.675000	1.688750e+01	3.993000e+01
People using safely managed drinking water services, urban (% of urban population)	176.0	5.154955e+01	2.415742e+01	11.200000	34.125000	51.365000	7.074750e+01	8.954000e+01
People using safely managed sanitation services (% of population)	132.0	2.878894e+01	1.863151e+01	6.370000	16.532500	25.410000	3.572500e+01	7.812000e+01
People using safely managed sanitation services, rural (% of rural population)	110.0	1.436173e+01	7.088038e+00	2.300000	7.200000	15.950000	2.031500e+01	2.554000e+01
People using safely managed sanitation services, urban (% of urban population)	132.0	3.217462e+01	2.169346e+01	7.950000	18.262500	22.755000	3.922500e+01	8.829000e+01
Rural population (% of total population)	588.0	5.683865e+01	1.808633e+01	11.020000	43.057500	58.445000	7.120500e+01	9.014000e+01
Rural population growth (annual %)	588.0	1.389371e+00	1.199693e+00	-3.450000	0.410000	1.675000	2.130000e+00	7.090000e+00
Urban population (% of total population)	588.0	4.316412e+01	1.808512e+01	9.860000	28.795000	41.560000	5.694500e+01	8.898000e+01
Urban population growth (annual %)	588.0	3.523081e+00	1.458244e+00	-4.650000	2.512500	3.730000	4.450000e+00	7.400000e+00
People using at least basic drinking water services (% of population)	588.0	6.599491e+01	1.728336e+01	28.960000	52.375000	64.470000	7.916500e+01	9.987000e+01
People using at least basic drinking water services, rural (% of rural population)	566.0	5.144958e+01	1.892787e+01	17.050000	37.075000	50.435000	6.224500e+01	9.983000e+01
People using at least basic drinking water services, urban (% of urban population)	566.0	8.426850e+01	9.307286e+00	52.010000	78.080000	85.420000	9.008250e+01	9.992000e+01
People using at least basic sanitation services (% of population)	588.0	3.946980e+01	2.630493e+01	4.990000	18.197500	32.555000	5.481000e+01	1.000000e+02
People using at least basic sanitation services, rural (% of rural population)	566.0	2.807721e+01	2.404673e+01	1.890000	8.842500	18.815000	3.808250e+01	9.518000e+01
People using at least basic sanitation services, urban (% of urban population)	566.0	4.808837e+01	2.180213e+01	12.580000	30.775000	44.695000	5.884500e+01	9.830000e+01
latitude	594.0	2.828798e+00	1.567823e+01	-30.559482	-6.369028	6.744051	1.286281e+01	3.388692e+01
longitude	594.0	1.734255e+01	2.004126e+01	-24.013197	0.824782	18.611308	3.146587e+01	5.755215e+01

Figure 2 Brief description of the values in the dataset

From the figure above, we see the maximum and minimum values, the mean, the standard deviation, the first, second and third quartiles of each column. This basically provides summary statistics for each column in the data frame. The dataset was seen to have null values, no duplicates, 594 rows and 27 columns containing numerical values, objects and geographical values. It was also categorized from year 2007 to 2017 across different African countries.

b. Data Cleaning

	Country Name	Year	Country Code	Incidence of malaria (per 1,000 population at risk)	Malaria cases reported	Use of insecticide-treated bed nets (% of under-5 population)	Children with fever receiving antimalarial drugs (% of children under age 5 with fever)	Intermittent preventive treatment (IPT) of malaria in pregnancy (% of pregnant women)	People using safely managed drinking water services (% of population)	People using safely managed drinking water services, rural (% of rural population)	People using safely managed drinking water services, urban (% of urban population)	People using safely managed sanitation services (% of population)	People using safely managed sanitation services, rural (% of rural population)	n s s u pop
0	Algeria	2007	DZA	0.01	26.0	NaN	NaN	NaN	NaN	NaN	NaN	18.24	19.96	
1	Angola	2007	AGO	286.72	1533485.0	18.0	29.8	1.5	NaN	NaN	NaN	NaN	NaN	
2	Benin	2007	BEN	480.24	0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
3	Botswana	2007	BWA	1.03	390.0	NaN	NaN	NaN	NaN	NaN	83.96	NaN	NaN	
4	Burkina Faso	2007	BFA	503.80	44246.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
589	Togo	2017	TGO	278.20	1755577.0	69.7	31.1	41.7	NaN	NaN	NaN	NaN	NaN	
590	Tunisia	2017	TUN	NaN	NaN	NaN	NaN	NaN	92.66	NaN	NaN	78.12	NaN	
591	Uganda	2017	UGA	336.76	11667831.0	NaN	NaN	NaN	7.07	4.46	15.70	NaN	NaN	
592	Zambia	2017	ZMB	160.05	5505639.0	NaN	NaN	NaN	NaN	NaN	46.25	NaN	NaN	
593	Zimbabwe	2017	ZWE	108.55	467508.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	

594 rows × 27 columns

Figure 3 Dataset before cleaning

For sections of the data with low percentage of null values we decided to use a descriptive statistic of mean to fill in for the missing values in order to preserve the overall distribution of the data. For sections of the data with high percentage of null values we decided to fill in with the value 0 because of data type compatibility and also 0 can be used to represent nonexistent data which will allow us do necessary computations and visualizations. We didn't use descriptive statistics in this case as it wouldn't be able to maintain the overall distribution of data due to the large number of null values. This method of cleaning allowed us to come up with clean data without changing a lot of meaning to it thus allowing proper analysis and visualizations. Further on, we shortened the column names in order to reduce on the amount of code to be written when referring to the columns in cases of generating visualizations. We also converted the data type of the "Malaria cases reported" column from Float to Integer.

	cntry_name	yr	cntry_code	malaria_incidence	malaria_cases	ITNs_usage(% <5)	antimalarial_Rx(% <5)	IPT_in_pregnancy(% Pregnant)	safe_drinking_water(% Total)	safe_drinki
0	Algeria	2007	DZA	0.010000	26	0.0	0.0	0.0	0.00	
1	Angola	2007	AGO	286.720000	1533485	18.0	29.8	1.5	0.00	
2	Benin	2007	BEN	480.240000	0	0.0	0.0	0.0	0.00	
3	Botswana	2007	BWA	1.030000	390	0.0	0.0	0.0	0.00	
4	Burkina Faso	2007	BFA	503.800000	44246	0.0	0.0	0.0	0.00	
...	...	...	...	...	...	...	...	...	...	
589	Togo	2017	TGO	278.200000	1755577	69.7	31.1	41.7	0.00	
590	Tunisia	2017	TUN	190.087491	1068330	0.0	0.0	0.0	92.66	
591	Uganda	2017	UGA	336.760000	11667831	0.0	0.0	0.0	7.07	
592	Zambia	2017	ZMB	160.050000	5505639	0.0	0.0	0.0	0.00	
593	Zimbabwe	2017	ZWE	108.550000	467508	0.0	0.0	0.0	0.00	

594 rows × 27 columns

Removing 0's from the column of reported malaria cases

Figure 4 Dataset after cleaning

c. Data Visualization

Under Data visualization, we analyze and view the components of the dataset in order to generate meaningful conclusions from the various visualizations of the data. The main categories of diagrams used were;

Heatmaps

With heat maps, we are able to study the Correlation among the features of the dataset which allows us to analyze the strength of the relationships among them.

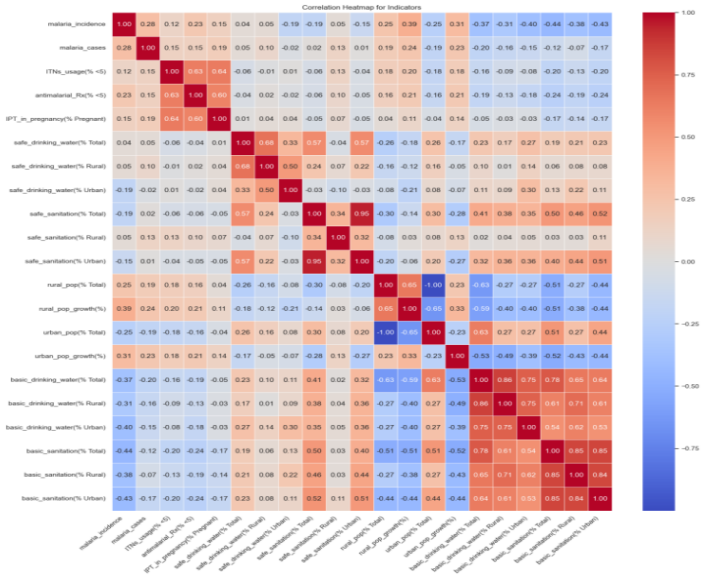


Figure 5 Heatmap showing the features of the dataset

From the figure above, we get to see the variations in the strengths of relationships among the features of the dataset. Focusing on the malaria_cases column, we see that the features malaria_incidence, urban_pop_growth(%) and rural_pop_growth(%) have the strongest correlations with the malaria_cases column.

Bar Charts

With bar charts, we are able to study the variations in figures and compare one component of a feature with another. The colorful bar charts not only help us generate conclusions but also provide a foundation for the creating of ideas that arise from the analysis of the diagrams.

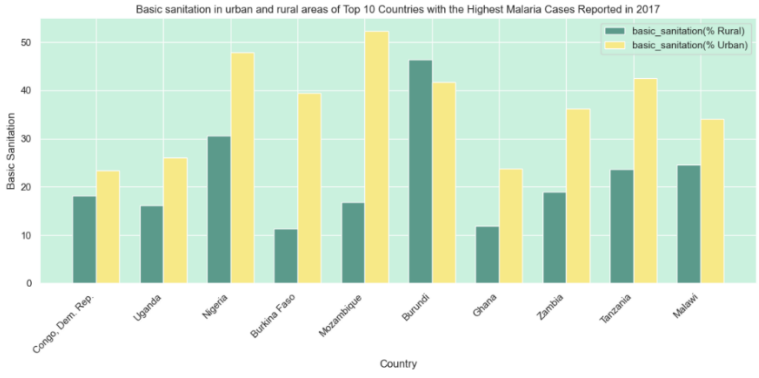


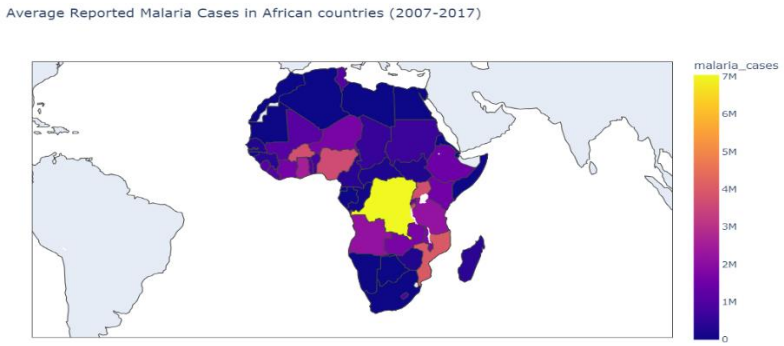
Figure 6 Bar chart comparing sanitation between the rural and urban population

From the figure above, we are able to make comparisons between the basic sanitation in the urban and rural areas. We focus on the countries with the highest number of Malaria cases in

the year 2017 and we see that basic sanitation in the urban areas takes larger percentages than that of rural areas.

Choropleth Maps

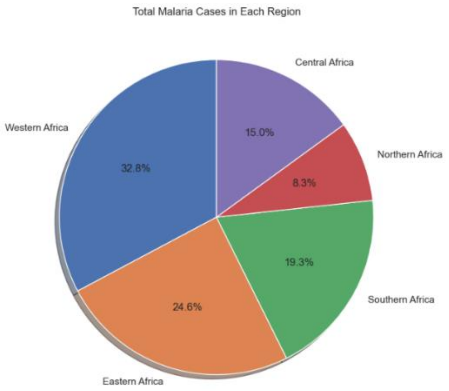
Choropleth maps allow us to visualize the African countries according to their respective locations. And with the use of colors, we can analyze the differences among each of the countries.



From the choropleth map above, the colors range according to the average number of malaria cases in such a way that the brighter the color, the larger the value. The Democratic Republic of Congo appears to be the country that has the highest Average number of reported Malaria cases from the years 2007 to 2017.

Pie Charts

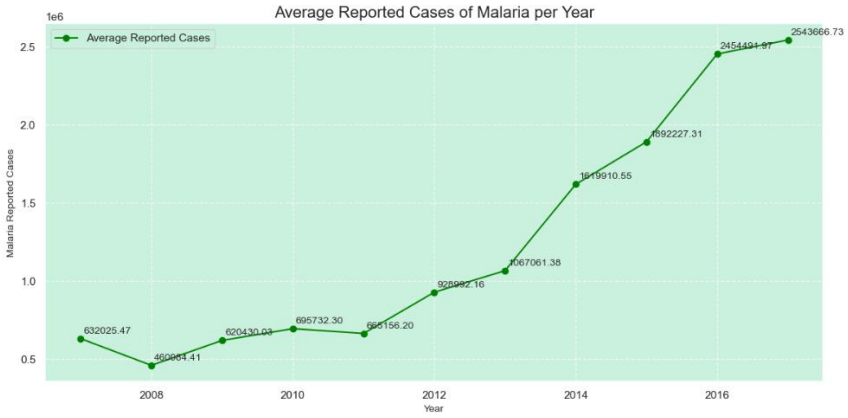
Pie charts can play a crucial role in visually representing the Malaria dataset for African countries by providing a concise overview of the differences in the features across different countries.



From the figure above, the region with the highest number of Malaria cases is Western Africa. Despite the fact that the country with the highest average number of Malaria cases that is to say Democratic Republic of Congo is in Central Africa, Western Africa appeared to have the largest total of Malaria cases mostly due to the fact that it consists of the highest number of countries.

Line Charts

Line charts offer a valuable tool for analyzing the Malaria dataset for African countries as they depict temporal trends in malaria concerned statistics over a specific period.



From the Line chart above we see that the average number of Malaria cases continues to increase over time which is a foundation to the conclusion that the year 2017 had the highest number of Malaria cases reported from the Africa from 2007 to 2017.

d. Model Building

Basing on the dataset we make a predictive model by selecting the best algorithm to do informed predictions, through training the model and evaluation of its accuracy

DEDUCTION AND CONCLUSIONS FOR OUR DATA

- We see that the number of malaria cases reported has a general increase from the year 2007 to 2017.
- There was a general decrease in the average of the incidences of malaria per year.
- We established that there was a low correlation for most of the columns.
- The Democratic Republic of Congo was found to dominate in the average malaria cases reported from the year 2007 to 2017 with Mozambique coming in second.
- 2017 was registered with the highest number of malaria cases.
- There was a population decrease in rural areas while there was a population rise in urban areas as time passed from 2007 to 2017
- The percentage of population growth is decreasing over time from 2007 to 2008
- Overall, we see that even if measures have been put in place to reduce malaria which facilitated reduction in the risk of malaria, the overall malaria cases have increased through out the years from 2007 to 20017