STATISTICAL ANALYSIS OF MALARIA

Individual Project in Statistical Methods and Applications I
(Spring 2023)

By
Devyani Srivastava
Student ID: 110620647

EXECUTIVE SUMMARY

Malaria is still a significant global health issue, particularly in developing countries with poor healthcare infrastructure. We conducted analysis on malaria to better understand the disease and develop effective prevention and treatment strategies using various statistical testing methods, time series analysis and regression analysis.

TABLE OF CONTENTS

I. INTRODUCTION

Malaria is still a serious global health problem. The World Health Organization (WHO) estimates that there were 229 million cases of malaria worldwide in 2019 and 409,000 fatalities as a result of the illness, with sub-Saharan Africa hosting the bulk of cases. However, there has been improvement in recent years in lowering malaria occurrences and fatalities.

In order to comprehend the present malaria situation and create practical prevention and treatment plans, statistical analysis is pertinent and helpful. Researchers can determine regions where the disease is most prevalent and monitor the development of methods for lessening the burden of malaria by examining patterns in malaria cases and fatalities over time. Identification of malaria risk factors and evaluation of the efficacy of therapies can both be assisted by statistical analysis.

In the context of malaria, statistical analysis aims to offer perceptions and evidence-based suggestions for public health policy and practice. Researchers can find patterns and links in the data by analyzing huge datasets and performing statistical modeling, which may not be obvious from a cursory visual review. This may result in strategies for malaria prevention and treatment that are more precise and successful.

Therefore, the main objective of statistical analysis of malaria is to understand the patterns and trends of the disease, identify risk factors, and evaluate the effectiveness of interventions. Statistical analysis can provide insights into the prevalence and distribution of malaria, as well as the factors that contribute to the disease burden in different regions. By analyzing large datasets, researchers can identify patterns and relationships in the data that can help inform public health policy and practice.

Furthermore, this project also applies regression analysis that can help identify risk factors for malaria and understand how changes in these factors may impact the prevalence of the disease. It also employs Time series analysis that can help predict future trends in malaria incidence and inform public health interventions. The goal of this project is to offer an invaluable chance to investigate a variety of statistical techniques and methodologies, including data exploration, hypothesis testing, and regression analysis, which may be utilized to solve problems in a variety of business sectors.

The big-picture questions this project is trying to answer with its data are:

- Is there differences in malaria incidence and mortality rates between different species of malaria (e.g. pfalciparum vs pvivax)?
- Which countries had increment or decrement in cases and deaths due to malaria over time?

- Is there any correlation between number of tests performed and the number of malaria cases ?

II. DATA SOURCE

The dataset used in this analysis was sourced from the World Health Organization. They are 6 datasets i.e. Imported Cases, Indigenous_pfalciparum_cases, Indigenous_pvivax_cases, estimated_deaths, microscopy_tests and rdt_tests. By cleaning and manipulating these datascources was able to build dataframe merged_df .

Variables

Variable Name	Variable Type	About
Indicator Code	Chr vector	Type of Malaria
Region	Chr vector	Various regions of the world
Country	Chr vector	Name of the country
Period	Numeric vector	Year
Confirmed Cases	Numeric Vector	Confirmed Cases of malaria
Confirmed Deaths	Numeric Vector	Confirmed Deaths due to malaria

table (1): Variables in the merged_df dataset with their descriptions

Variables

Variable Name	Variable Type	About
Indicator Code	Chr vector	Type of Test microscopy or RDT
Region	Chr vector	Various regions of the world
Country	Chr vector	Name of the country
Period	Numeric vector	Year
Estimated Cases	Numeric Vector	Estimated Cases of malaria

table (1): Variables in the tests_cases_df dataset with their descriptions

Variable Distributions

There are only 3 numeric variables in this dataset, whose distributions are as follows:

Confirmed Deaths: The distribution of quantity is slightly skewed to the left, indicating that most transactions involve a less number of deaths.

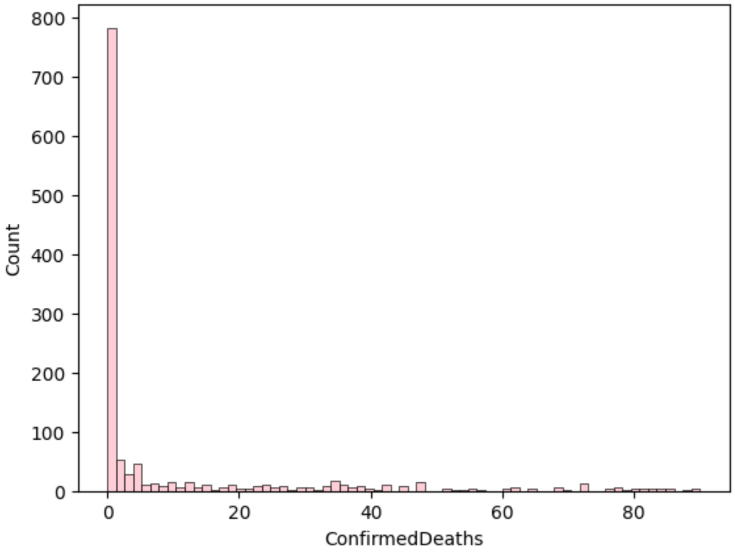


fig : Distribution of Confirmed Deaths

Confirmed Cases: The distribution of unit confirmed cases is also skewed to the left, indicating low number of confirmed cases.

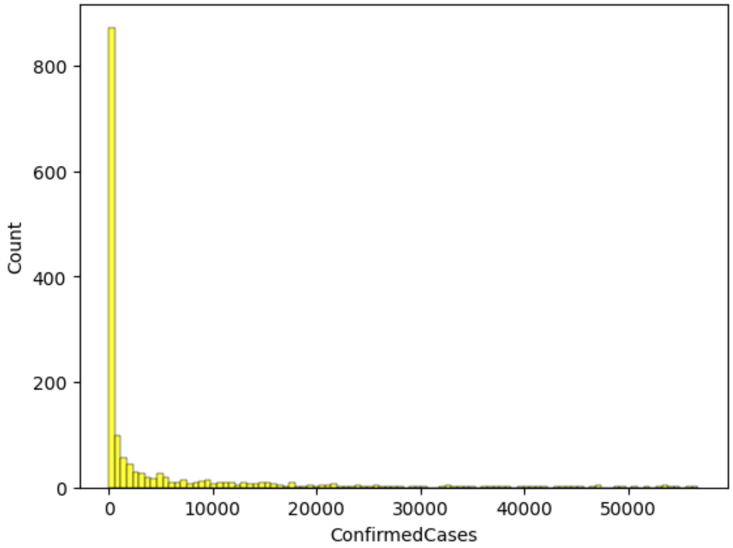


fig : Distribution of Confirmed Cases

Period: It includes for the years 2010 to 2020. It is a integer type variable.

Variable Relationships

To explore the relationship between numerical variables, we created a heatmap of the correlation matrix between Confirmed Cases, Confirmed Deaths and Period. The heatmap shows a correlation coefficient of around 0.48, which indicates a moderate relationship between confirmed cases and confirmed deaths. There is weak correlation between period and confirmed cases.

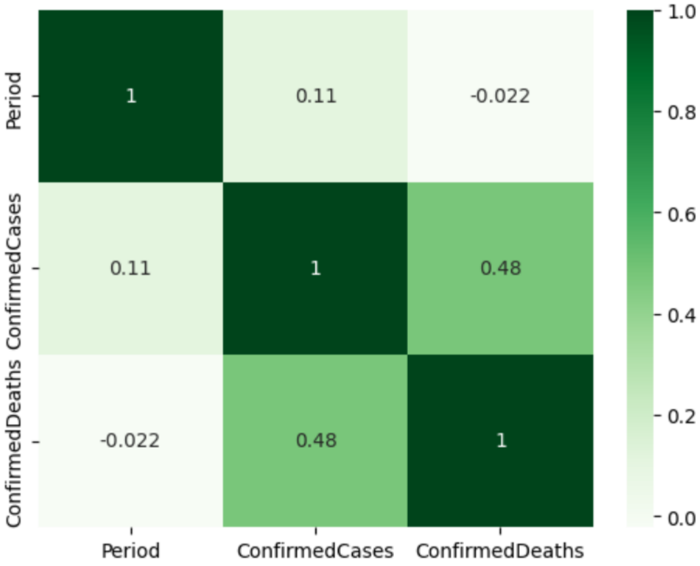


fig (4): Correlation heatmap of Period, Confirmed Cases and Confirmed Deaths

The heatmap also shows that there is no significant correlation between Quantity and any other variable in the dataset except a weak negative correlation between Period and Confirmed Cases, which suggests that number of deaths is not affected by which year it occurred.

III. MAIN METHODOLOGY

The steps followed to complete this project are listed in the diagram below:

1. *Data Collection & Preprocessing*
    a. Duplicate and missing values: To improve data quality and ensure accurate results when connected together, all databases were checked for mistakes and missing data patterns before being exposed to standardization algorithms. Among these are reducing duplicate entries, standardizing column names, deleting unnecessary columns, upgrading data types, and carefully examining and addressing null values.

    b. Outliers: Outlier detection is a crucial step in the analysis of malaria data because it ensures correctness, spots errors, provides insights, helps with decision-making, and enables comparisons. By identifying outliers, we can confirm that the analysis is reliable and that the decisions made using it are well-informed real instances, and a thorough analysis is needed to ascertain the fundamental reasons behind their recurrence. By looking more closely at the outliers, we may be able to better understand the problem and identify any underlying patterns or trends that might be contributing to the high number of cases in this place. This can help direct targeted strategies and actions to address the issue and reduce the malaria burden in Africa. It is important to remember that just eliminating the outliers might not be the best course of action because these instances may include important information that can guide the analysis and decision-making process. To fully understand the problem and suggest acceptable remedies, a detailed analysis of the outliers is also essential.

c. Normalization: Since we prefer not to remove the outliers, we refrained from normalizing the data. Outliers can distort the distribution of the data and make normalization inappropriate as it is sensitive to extreme values.

2. *Statistical Analysis*

a. Central and dispersion tendencies: Key patterns and trends in malaria dispersion were discovered through statistical research. The initial phase was gathering certain fundamental central and dispersion tendencies. Following that, non-numeric columns were removed to produce summary statistics including minimum, maximum, mean, median, mode, range, variance, and standard deviation. The final summary statistics were then given a new name and made into a data frame.

b. Trend and pattern analysis: To find patterns and trends in the data, exploratory data analysis was carried out. The regions with most cases and deaths were identified for years 2012 and 2020. The regions that prefer microcopy or RDT were discovered.

.

c. Hypothesis Testing: Four hypotheses regarding region, period, countries, test conducted, type of test, type of malaria, confirmed cases and confirmed deaths. The null and alternative hypotheses for each of the hypotheses were determined, and the following steps were taken to test them: deciding on a significance level of 0.05; gathering pertinent data; calculating the test statistic; calculating the p-value; and making a decision based on whether the p-value was below or equal to the significance level.

3. *Correlation and Regression Analysis*

    a. Correlation Analysis: To understand the relationship between various features correlation analysis was performed. Seaborn library was used to plot correlation matrix using corr method.

    b. Regression Analysis: A linear regression model was fitted using Quantity as the response variable and Price as the predictor variable. The regression analysis showed the p-value and the relationship successfully.

4. *Time Series Analysis*

    To understand the trends of tests conducted over period of time seasonality test was employed whereas stationarity test was used to determine if the time series variables are stationary or not using ADF (Augmented Dickey Fuller) test.

5. *Forecasting*

    By examining historical data on incidence and death rates, researchers may create models to predict future trends in the severity and spread of malaria. This information can be used to better allocate resources in disease-endemic areas and to guide policy decisions regarding disease preventive and treatment initiatives. In order to lessen the spread and effects of malaria, forecasting models can also help in identifying potential risk factors for the illness and designing targeted treatments. ARIMA (Autoregressive Integrated Moving Average) and SARIMA (Seasonal Autoregressive Integrated Moving Average) was applied to forecast the malaria cases.

IV. FINDINGS

The insights are listed section-wise in the same order as they were in the methodology section.

1. *Data Collection & Preprocessing*

An examination of the dataset revealed that they were several columns that had missing values those columns were removed. The occurrence of several outliers in the African region when compared to other nations shows that there might be underlying patterns or trends that are particular to this region. These outliers are true examples, and a thorough analysis is necessary to ascertain the underlying factors that keep happening to them.

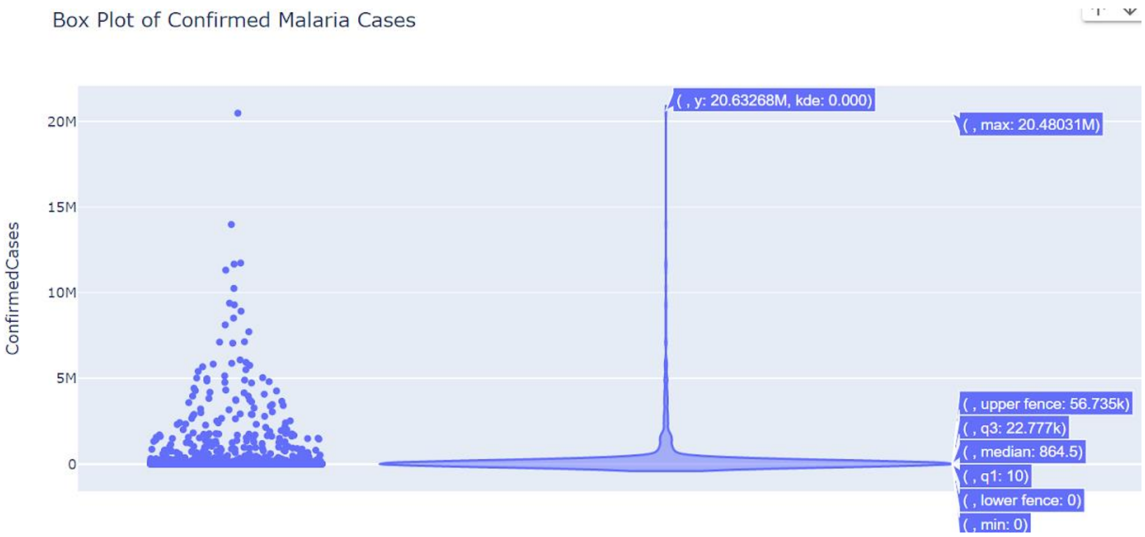


fig (8): Outliers in Confirmed Cases

By looking more closely at the outliers, we may be able to better understand the problem and identify any underlying patterns or trends that might be contributing to the high number of cases in this place. This can help direct targeted strategies and actions to address the issue and reduce the malaria burden in Africa.

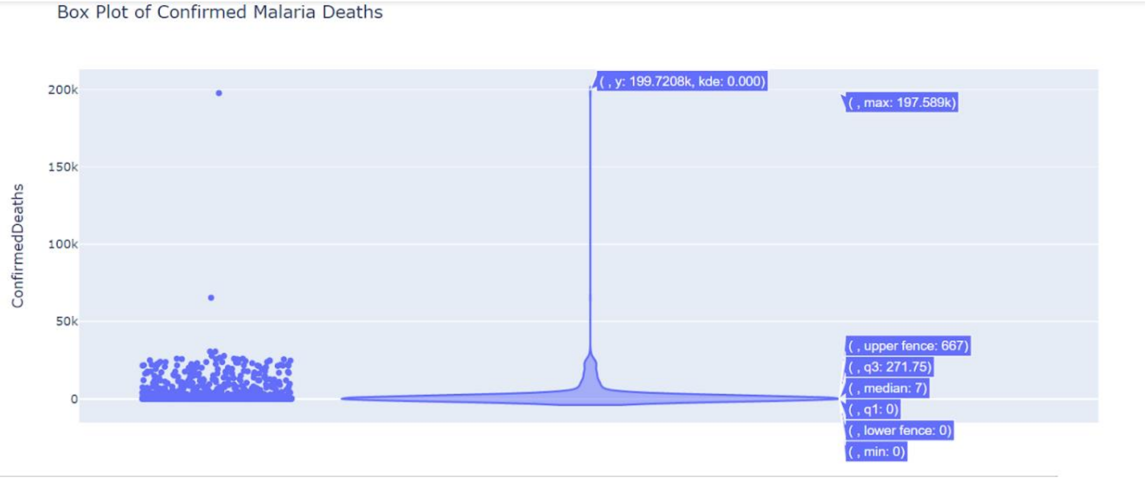


fig (8): Outliers in Confirmed Deaths

It is important to remember that just eliminating the outliers might not be the best course of action because these instances may include important information that can guide the analysis and decision-making process. To fully understand the problem and suggest acceptable remedies, a detailed analysis of the outliers is also essential.

2. *Statistical Analysis*

a. Central and dispersion tendencies: Minimum, maximum, mean, median, mode, range, variance, and standard deviation were ultimately the central and dispersion statistics obtained. The tendencies data set contained these.

fig (9) : Statistical summary of all numerical values of the dataset (Customer ID was not considered)

b. Trend analysis: Exploratory data analysis was performed to identify patterns and trends in the data.

i) The number of malaria cases and deaths is higher in certain Regions.

ii) The number of confirmed malaria cases and deaths has decreased over time.

iii) They are more microscopy test than RDT test

vi) There are differences in malaria incidence and mortality rates between different species of malaria (e.g. pfalciparum vs pvivax).

Numerous findings that are pertinent to the study topics were obtained from the analysis that was done.

Firstly, Total cases over the years were plotted against region. Africa has the most malaria cases over the years due to various factors such as the presence of the Anopheles mosquito, which is the primary carrier of the malaria parasite, and the high population density in many parts of the continent.

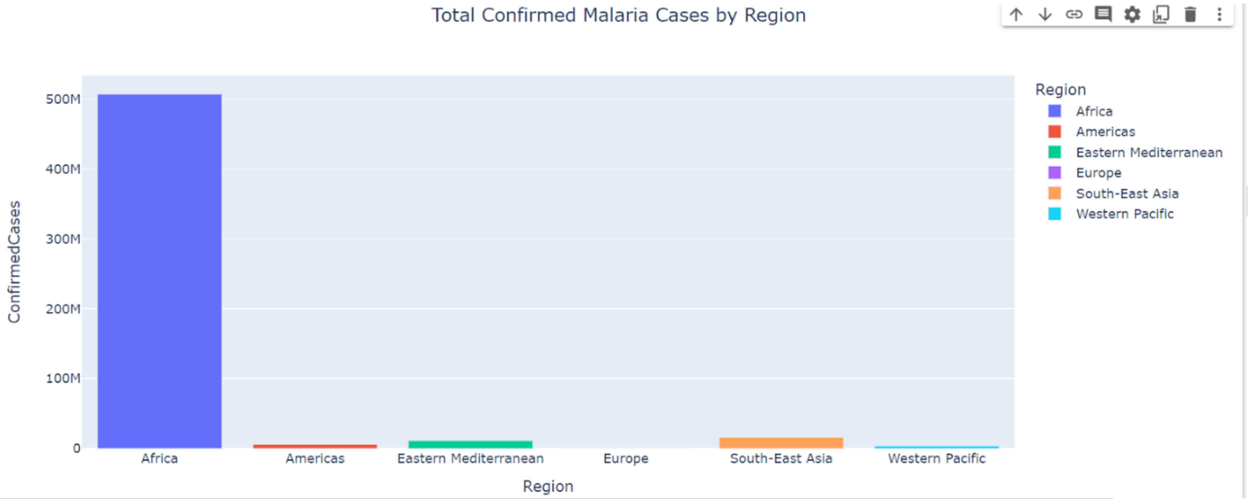


fig: Comparison of various regions as per total confirmed cases

Additionally, poverty, poor sanitation, and limited access to healthcare and effective anti-malarial drugs contribute to the high malaria burden in Africa. Climate change and environmental factors such as deforestation and water storage also play a role in creating suitable breeding grounds for mosquitoes.

Secondly, Number of cases were plotted against years for various regions. Similarly Total number of deaths were as well plotted to understand the trend.

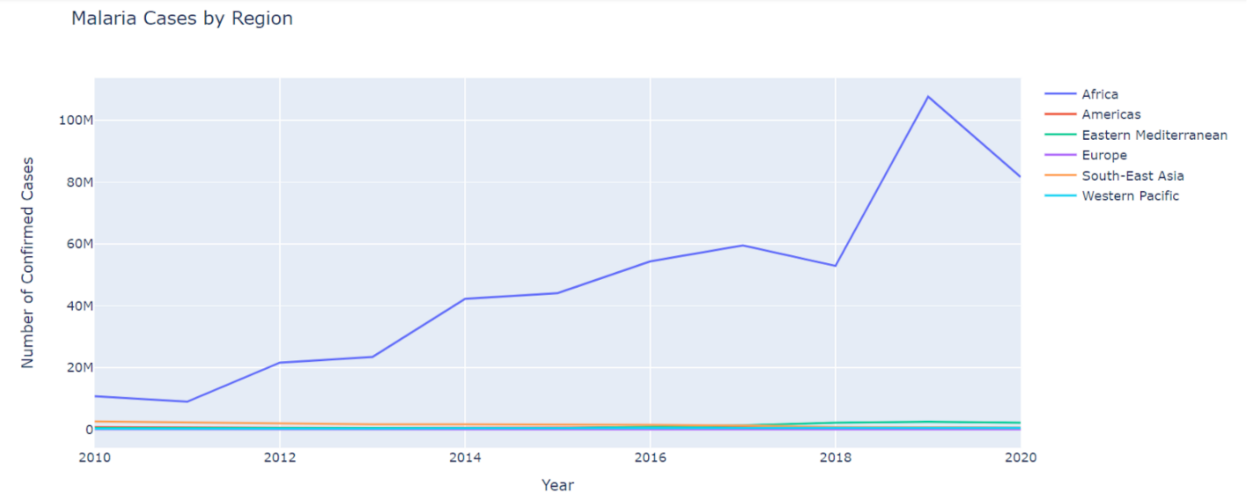


fig (8):Confirmed cases over time for different regions

The results show that throughout time, there has been a significant rise in the number of malaria cases in Africa and the Eastern Mediterranean region. Contrarily, there have been fewer confirmed instances in Southeast Asia, which is related to more awareness and improved medical care.

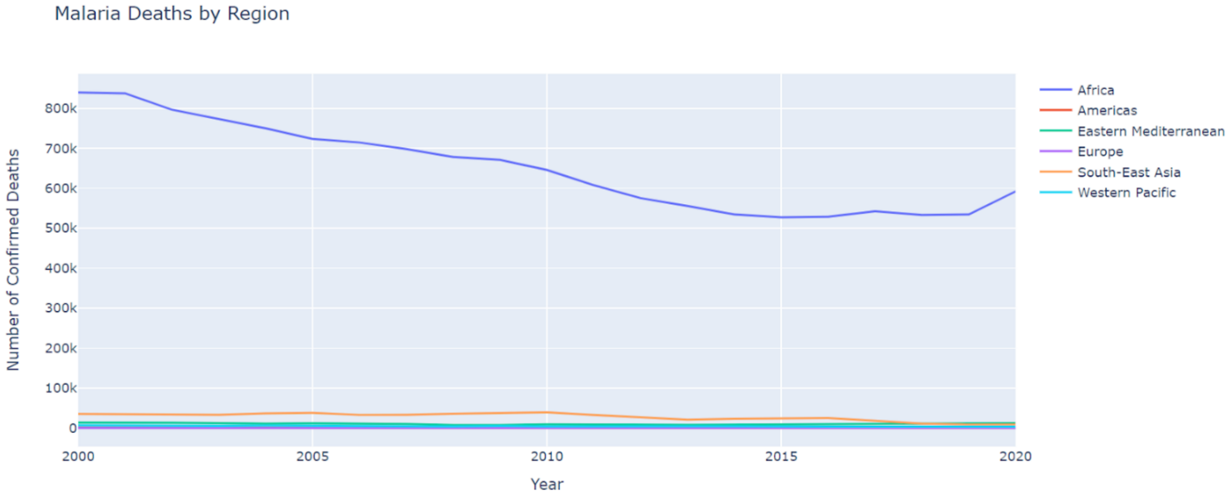


fig :Confirmed deaths over years for different regions

Despite the rise in instances, medical support services and healthcare facilities have improved, which has reduced the fatality rate. Any report on the prevalence of malaria and the efforts being undertaken to control it in these areas must include this information.

The fatality rate has reduced over time, despite an increase in the number of malaria cases that have been reported, because to improvements in medical care and healthcare infrastructure. Medical professionals now have the resources and knowledge necessary to provide prompt and effective care to people suffering from the illness. There have also been fewer deaths as a result of earlier diagnosis and treatment due to increasing public awareness and education programs.

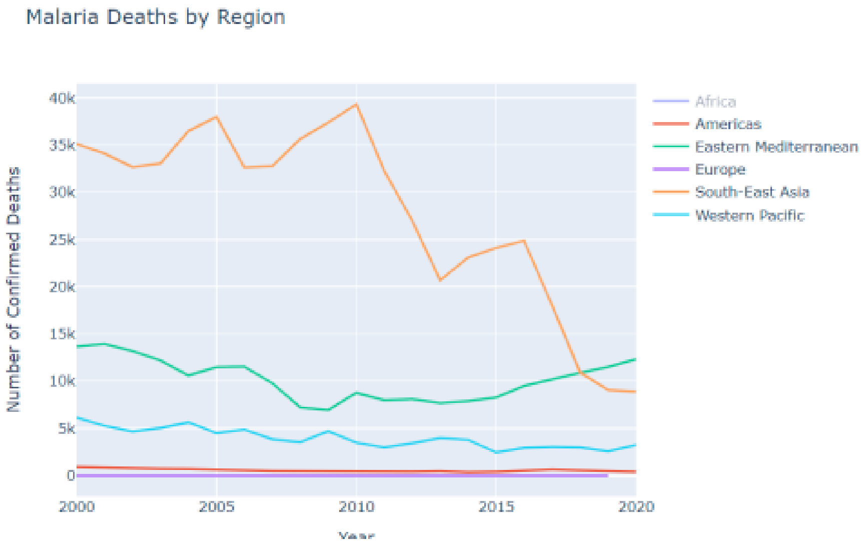
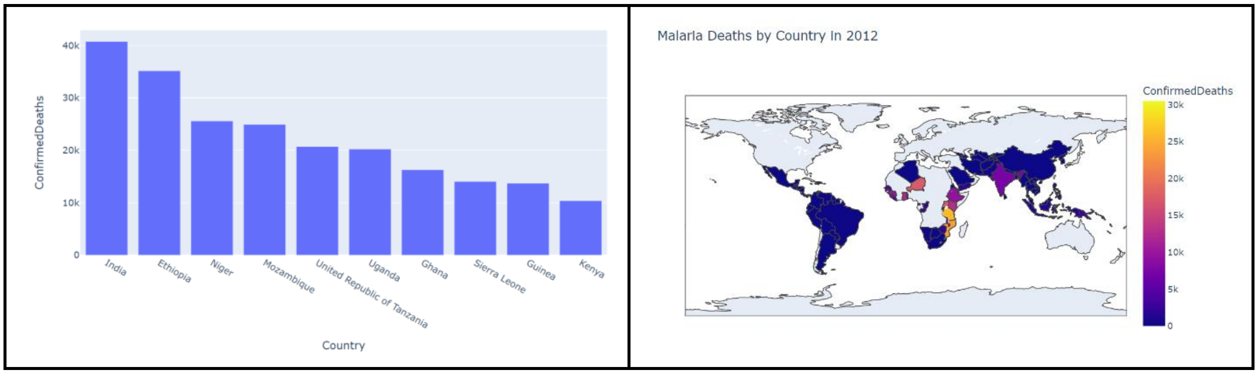


fig: Confirmed deaths over years for different regions sans Africa

Despite these positive advances, there is still a large burden of malaria throughout Africa and the Eastern Mediterranean, underscoring the significance of ongoing investments in preventive and control strategies. As a result, continued initiatives and collaboration are necessary to eradicate malaria and lessen its influence on the world.



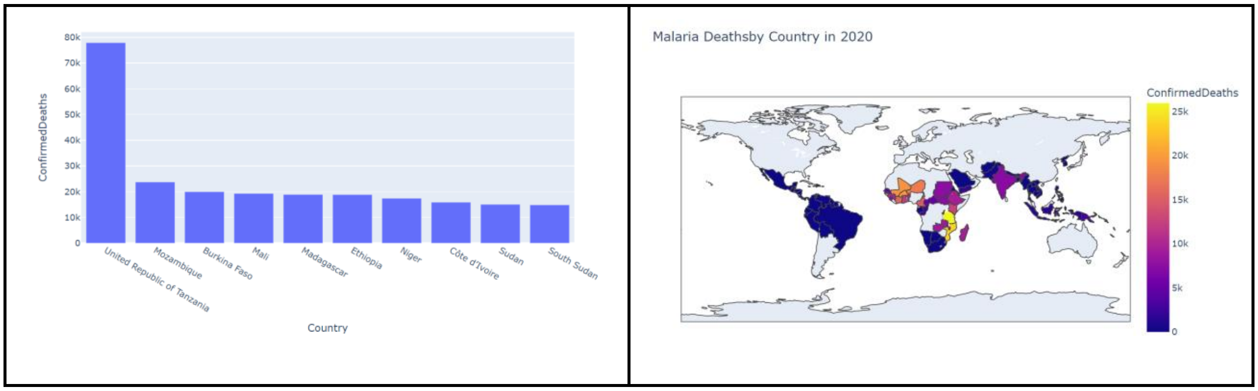


fig: Global deaths due to malaria for years 2012 (Top) and 2020 (Bottom)

A comparative study for year 2012 and 2020 depicts that India used to have the highest global malaria death followed by Ethiopia, Niger and Mozambique. The situation of these countries improved with exception of Mozambique where deaths rather increased.

Malaria deaths in Mozambique may have increased due to several factors such as a lack of access to healthcare, limited resources to implement preventative measures such as insecticide-treated bed nets and indoor residual spraying, and the emergence of drug-resistant strains of the malaria parasite.

In contrast, malaria deaths in India may have decreased due to successful implementation of various malaria control measures such as the distribution of insecticide-treated bed nets, effective case management through the use of artemisinin-based combination therapy, and targeted indoor residual spraying in high-risk areas. Additionally, India has made significant progress in improving access to healthcare services and strengthening its health systems over the years, which may have contributed to the decline in malaria deaths.

Thirdly, to understand which kind of test are mostly done count plot was employed.

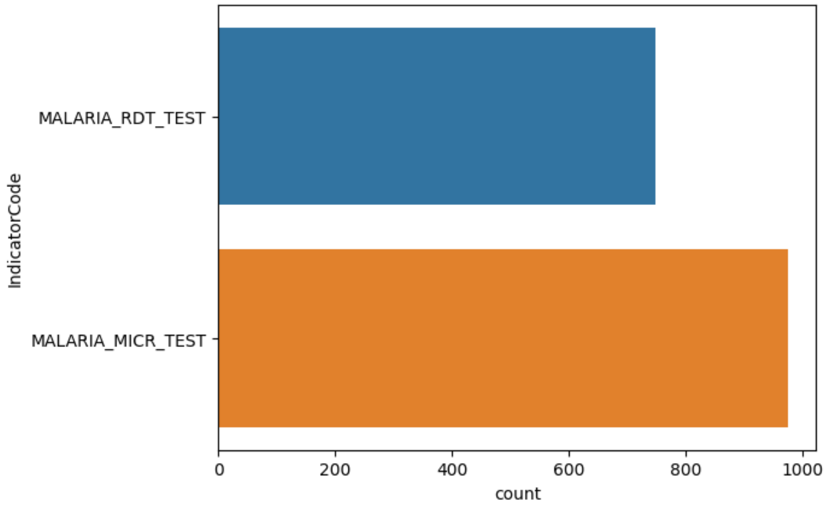


fig : Count plot of various tests

Tree map was used to understand which countries conducts most tests and of what kind. Because of their vast populations and high malaria loads, India, Pakistan, Ethiopia, Kenya, and China have some of the greatest numbers of microscopy tests for malaria. In many areas of these nations, malaria is an endemic disease, and the high population density increases the danger of transmission. As a result, the governments of these nations have put in place significant malaria control programs that comprise active case detection and microscopy tests for diagnosis.

One of the most popular techniques for diagnosing malaria is microscopy testing. In these procedures, blood samples are examined under a microscope to look for malaria parasites. Microscopy tests are a common diagnostic technique in situations with low resources because they are accessible, affordable, and can produce results quickly.

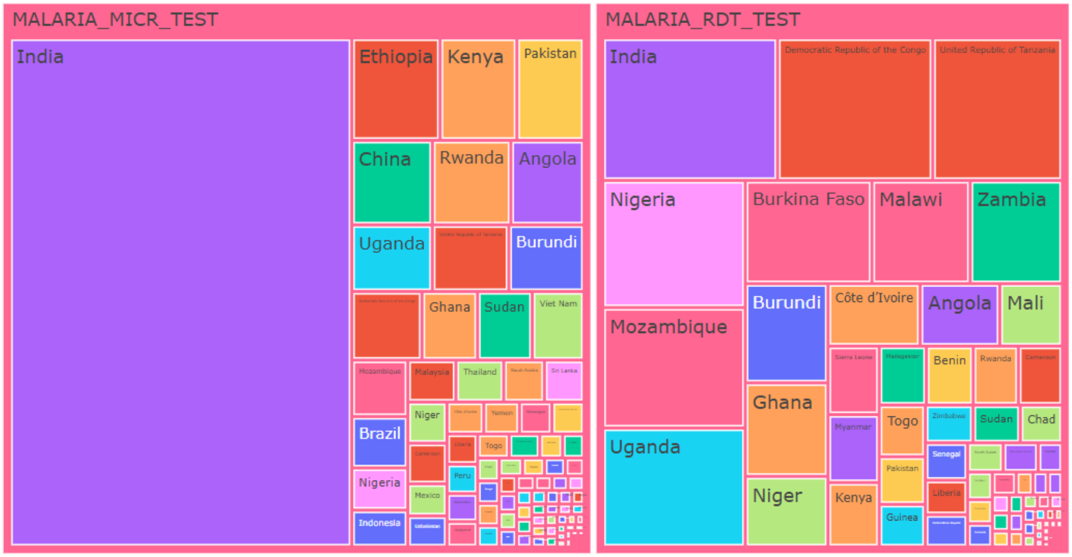


fig : Tree map of Microscopy Test (Left) and RDT Test (Right)

These nations frequently use skilled health professionals to conduct microscopy testing at the community level, and patients with positive test outcomes receive antimalarial drug treatment. These nations can lower the prevalence of malaria and avoid serious sequelae from the disease by doing routine microscopy tests and treating cases right away.

RDT (Rapid Diagnostic Test) is one of the most widely used diagnostic tests for malaria. India, Tanzania, Congo, Nigeria, and Uganda are among the nations with the highest burden of malaria in the world. These nations have high rates of malaria transmission, and their healthcare systems frequently depend on RDTs to swiftly identify and treat malaria cases, particularly in remote or resource-constrained locations where other diagnostic tools like microscopy may not be accessible.

RDTs are a useful and important tool in the diagnosis and treatment of malaria since they are inexpensive, quick, and generally simple to use. The expansion of malaria control and eradication activities in these nations have also received major funding from international organizations and donor agencies, which may have helped to the rise in RDT usage in recent years.

The bar plot shows that the number of deaths brought on by locally acquired malaria variants is significantly higher than the number of deaths brought on by imported variants, most notably Malaria_Imported. This implies that people who get malaria at home face a higher chance of dying than those who do so while traveling to endemic regions. Malaria is divided into three types: imported, PV imported, and PF imported. Imported malaria refers to cases in which the disease is spread by a foreign variant.

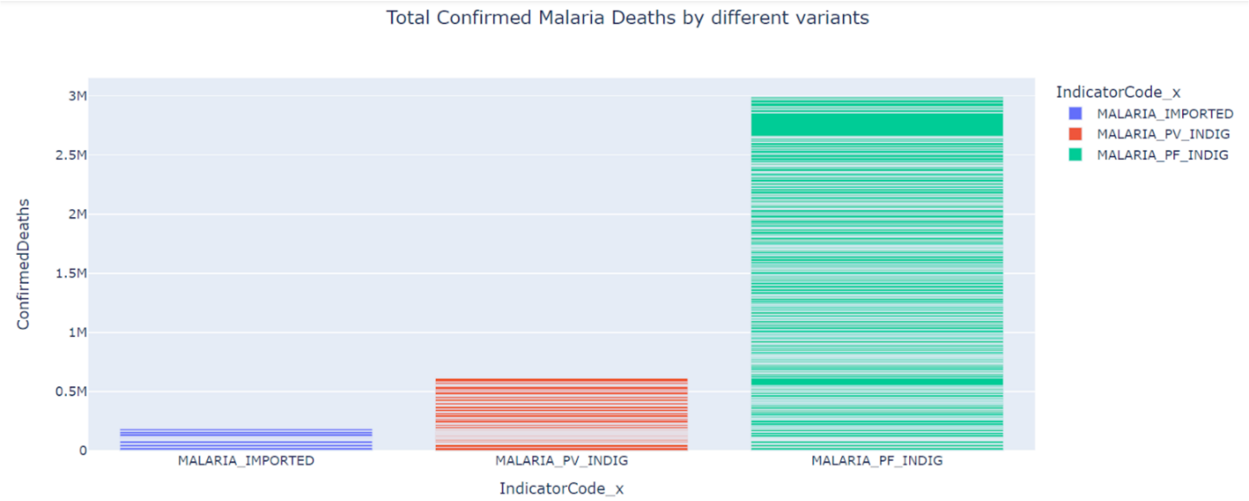


fig : Count plot of various sources of Malaria

Pfalciparum is the most frequent cause of malaria fatality, accounting for more than 3 million deaths, according to the statistics. Pvivax is the second-leading cause of death, contributing to almost 0.7.

million deaths. These startling statistics highlight the importance of effective malaria prevention and treatment efforts, especially in endemic areas. There is an urgent need for more funding for research to create new treatments and methods to tackle the disease given its major impact on world health.

In summary, visualizations helped identify patterns and trends in malaria data, such as the number of cases and deaths, the distribution of the disease across regions, and the effectiveness of malaria control measures. This information can inform decision-making for malaria prevention and control programs, leading to more targeted and effective interventions.

c. Hypothesis Testing: Hypothesis were presented:

i. The mean number of confirmed malaria cases in countries with malaria imported from other regions is significantly different from the mean number of confirmed cases in countries with indigenous malaria (either P. falciparum or P. vivax).

ii. There is significant association between variables region and indicator code

iii. The average confirmed indigenous cases in 2010 has same mean as the average of total confirmed indigenous cases.

iv. There is casual relationship between years and cases

Each one's null and alternate hypotheses were created. Seven different types of statistical tests were used to verify the 7 hypotheses. These were the t Test, Chi squared test, Anova, F test, causality test and z test

From the visualization we interpret that it is indeed true that the malaria cases imported from other countries are significantly different in number than the imported cases. The p value was 3.44e-23 which is less than significance level 0.05. Hence we reject null hypothesis. The kernel density curve of all three cases depicts the same below.

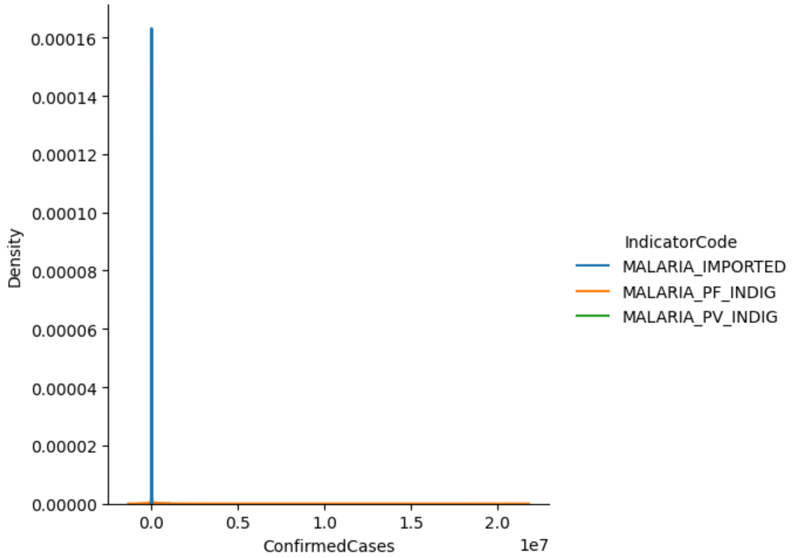


fig : Kernal density plot for first hypothesis

Same hypothesis is used to perform one way anova. The p value was approximately 6 e-05 and is less than significance value 0.05.

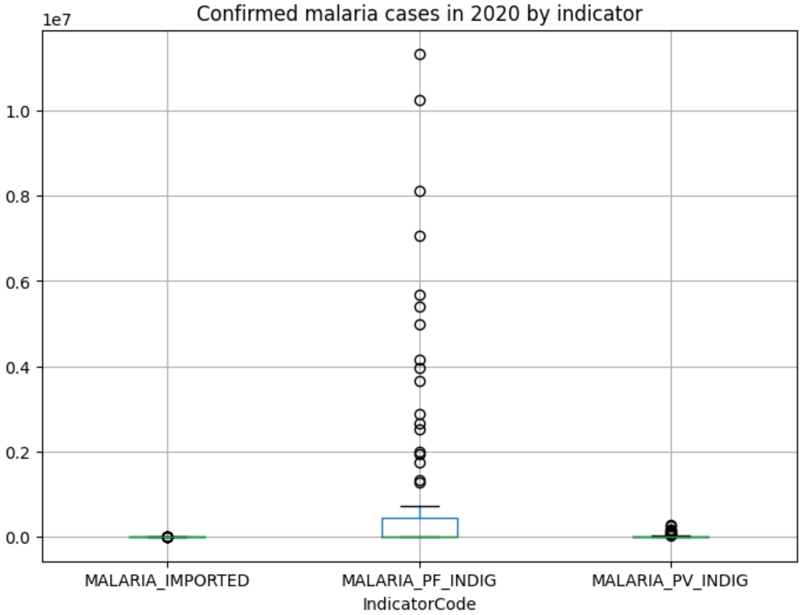


fig : Box plot for first hypothesis

The boxplot also confirmed this result by showing that the median and interquartile range for the the number of cases by each category. These insights suggest that there is significant difference in the mean of imported and indigenous cases.

Chi squared test is used to test second hypothesis. The p value was 6.1e-09 which is less than the significance value 0.05. Therefore the null hypothesis is rejected.The below correlation plot clearly depicts the same.

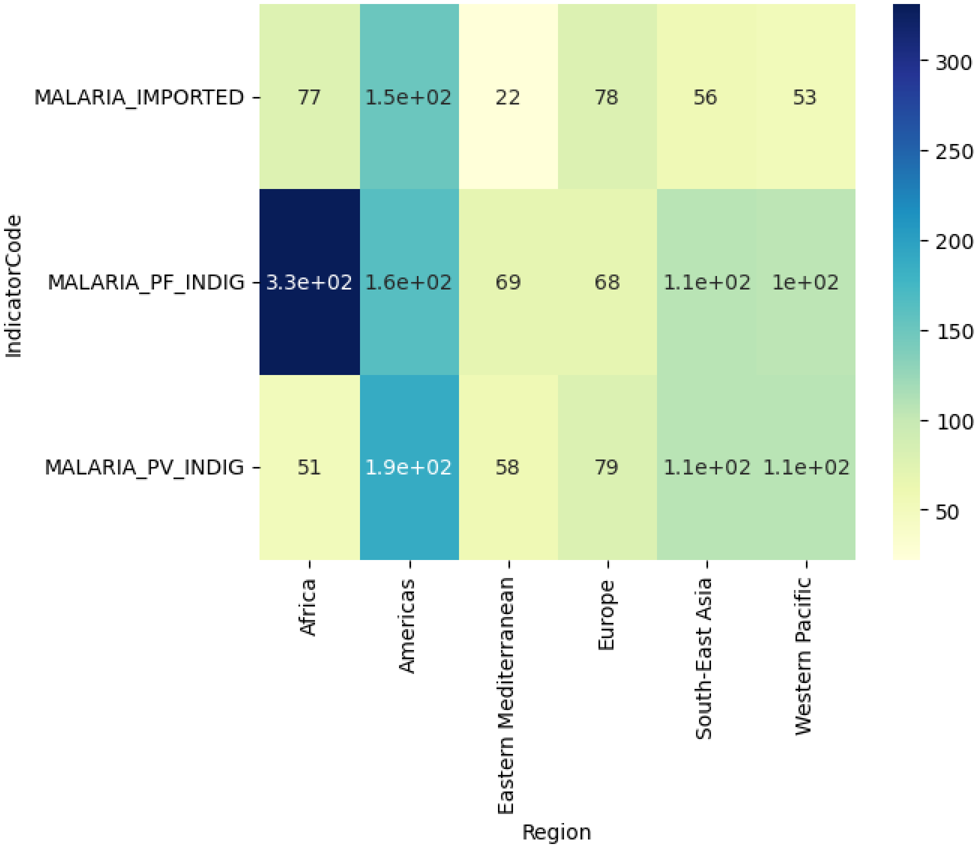


fig : Correlation plot for third hypothesis

Z Test was used to test third hypothesis which was if indigenous confirmed cases have same mean as the indigenous confirmed cases over the time. The p value was 0 and smaller than significance value 0.05. Hence our hypothesis is rejected.

To inspect fourth hypothesis, we employed granger causality test and F test. The p value was 0.85 and 0.98. Both values are higher than significance value 0.05 and hence fail to reject null hypothesis. Both scatter plot and boxplot visualizations conform to the same conclusion.

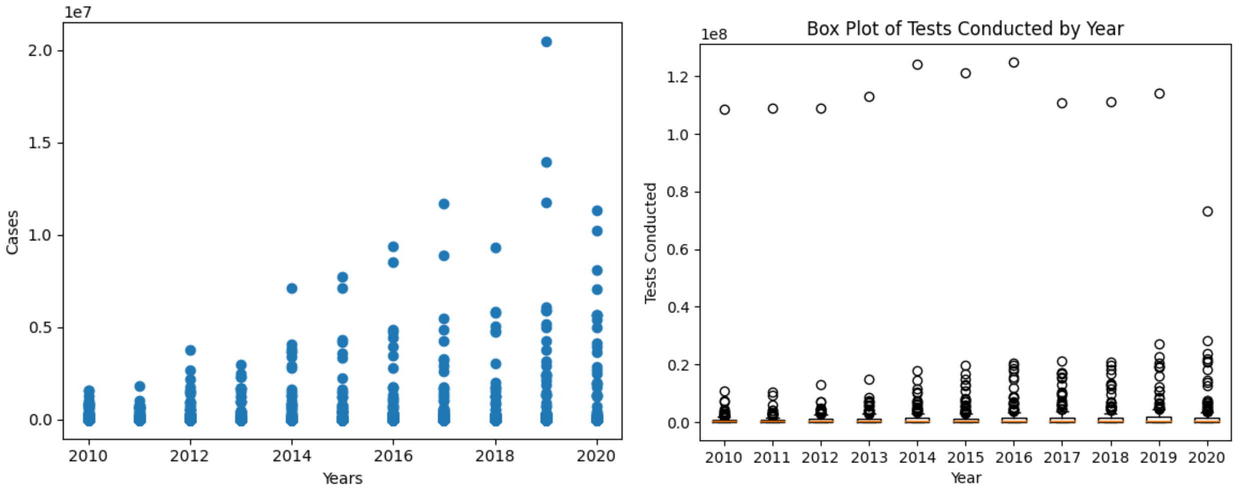


fig : Scatter and box plot for fourth hypothesis

3. *Correlation and Regression Analysis*

Correlation, which has a coefficient range of -1 to +1, is a statistical technique used to determine the link between two or more variables. There is no linear relationship when the correlation is zero. However, a correlation does not prove a cause and effect. We can comprehend and predict correlations between variables thanks to correlation, which is important. Research in the economy and the medical field both benefit from this.

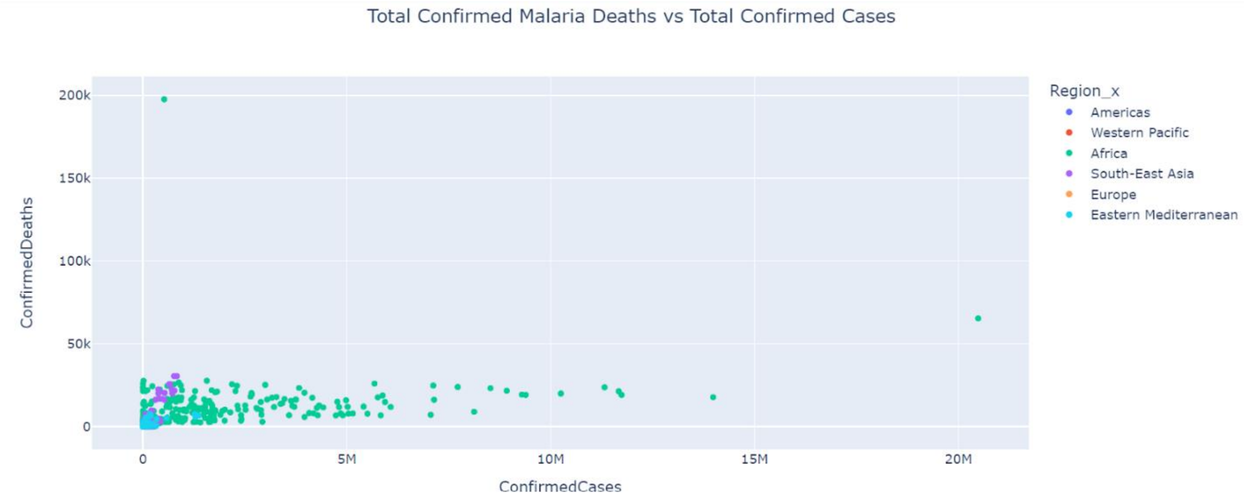


fig : Scatter Plot between Confirmed Deaths and cases

The scatter plot shows a strong correlation between the number of confirmed cases and confirmed fatalities, indicating that if one variable increases, the other climbs as well. It's a widely used infectious illnesses have a pattern. It should be noted, however, that the association might not hold true everywhere and might alter depending on other elements including medical assistance, population density, and demographic characteristics. It was executed using matplotlib.

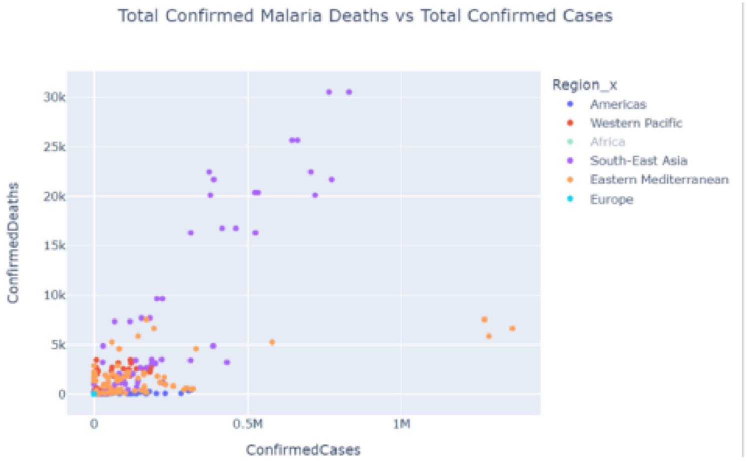


fig : Scatter plot between Confirmed Deaths and Cases sans Africa

A statistical method for simulating the relationship between one or more independent variables and a dependent variable is called linear regression. It forecasts the slope and intercept of the line that most closely matches the data and makes the assumption that the independent and dependent variables are connected linearly. When comparing observed and anticipated values, linear regression looks for the line that fits the data the best while minimizing the sum of squared residuals. The value of the dependent variable can therefore be predicted using this line using the value of the independent variable as a base.

The average absolute difference between the predicted and actual values is represented by the model's mean absolute error (MAE), which is 579084.9877. Better forecast accuracy from the model is shown by a lower MAE. The R-squared value for this model, however, is just 0.0191, indicating that only 1.91% of the variability in the dependent variable (Confirmed Cases), is explained by the independent variables (Tests Conducted and Period), despite the model having a low MAE. This shows that based on the quantity of tests performed and the time period, the model is ineffective at forecasting the number of confirmed cases.

5. Time Series Analysis and Forecasting

A seasonality test is a statistical method used to determine whether a time series data set exhibits seasonal patterns, which are variations that occur at regular intervals of time, such as daily, weekly, or yearly.

In the context of malaria, seasonality tests can help identify patterns in the number of malaria cases that occur throughout the year, which can be used to inform interventions such as the timing of distribution of insecticide-treated bed nets or the scheduling of indoor residual spraying. This information can help optimize the use of resources and target interventions to the times of the year when malaria transmission is most likely to occur.

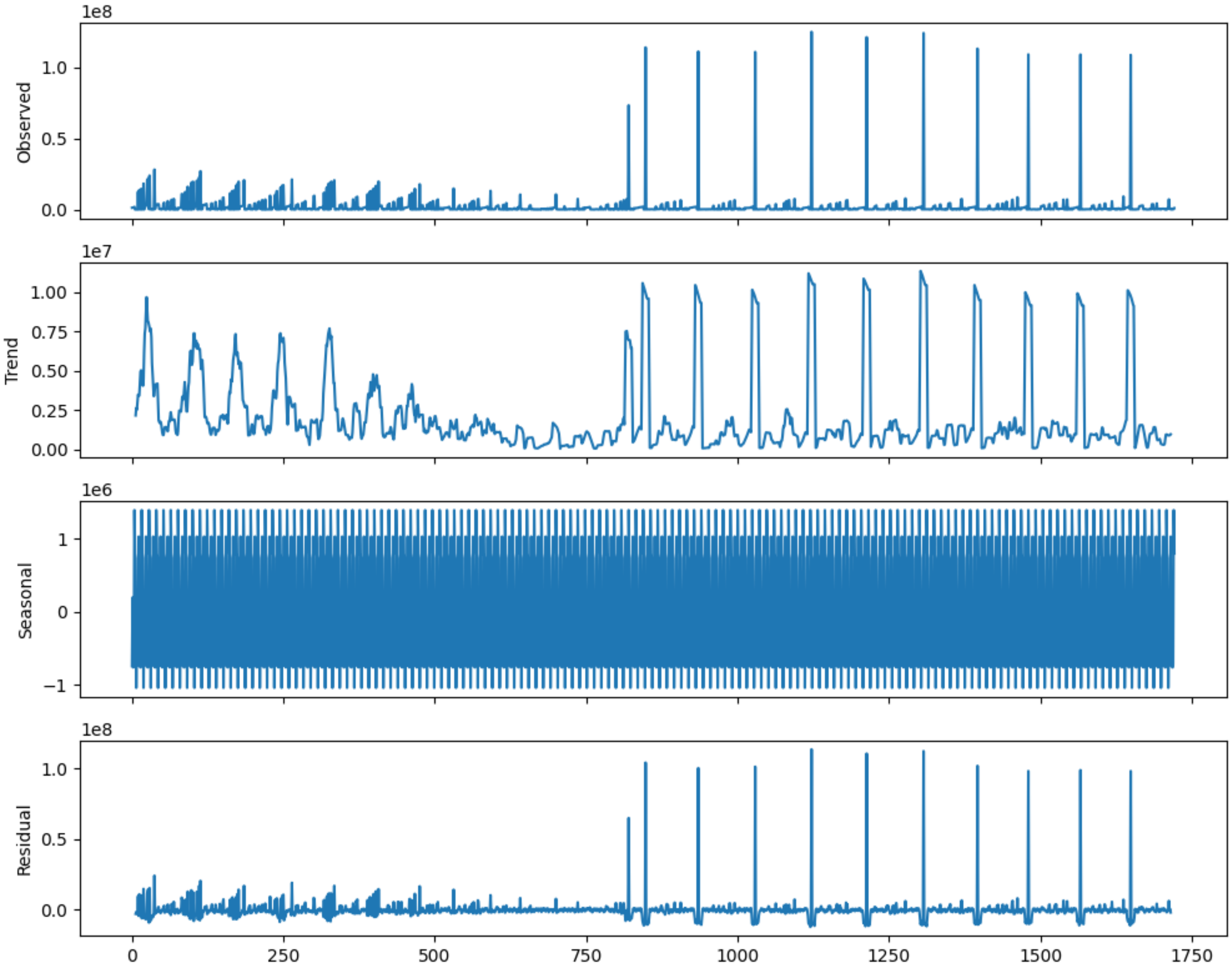


fig : Seasonality test for test conducted variable

First, the Period column was converted to datetime format. Then, it sets Period as the index of the data. After that, it performs seasonal decomposition on the TestsConducted column using an additive model with a period of 12 (since the data is yearly and we expect seasonal patterns to repeat each year). Finally, it plots the observed values, trend component, seasonal component, and residual component of the decomposition.

The null hypothesis for this test is that there is no seasonality in the TestsConducted variable. Since the p value is 0 which less than significance value 0.05. Hence null hypothesis is rejected, there is seasonality in the test conducted variable.

To perform stationarity test Augmented Dickey Fuller method is used. The null hypothesis is that the time series is non-stationary, and the alternative hypothesis is that the time series is stationary.

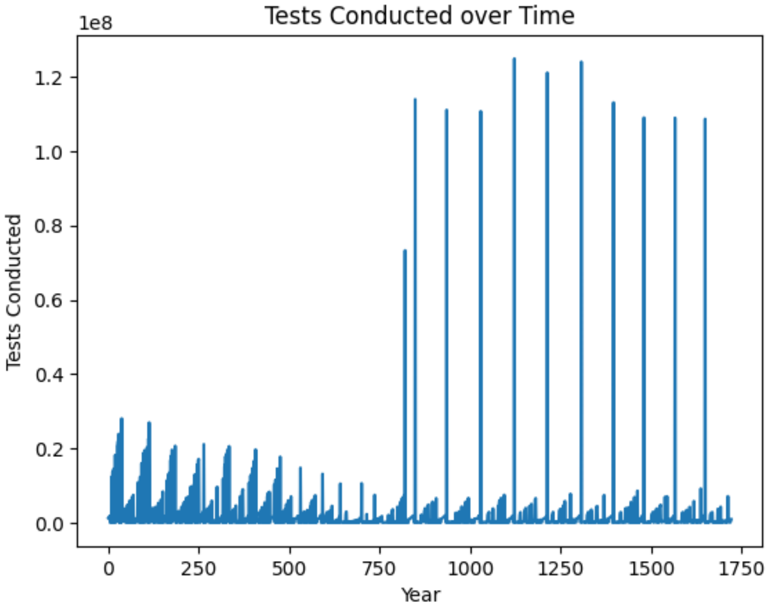


fig : Stationarity test for test conducted variable

The p value of significance value 5% is -2.8. Therefore we reject null hypothesis hence time series is stationary.

Cases Forecasting

The information from forecasting can be used to better allocate resources in disease-endemic areas and to guide policy decisions regarding disease preventive and treatment initiatives. In order to lessen the spread and effects of malaria, forecasting models can also help in identifying potential risk factors for the illness and designing targeted treatments.

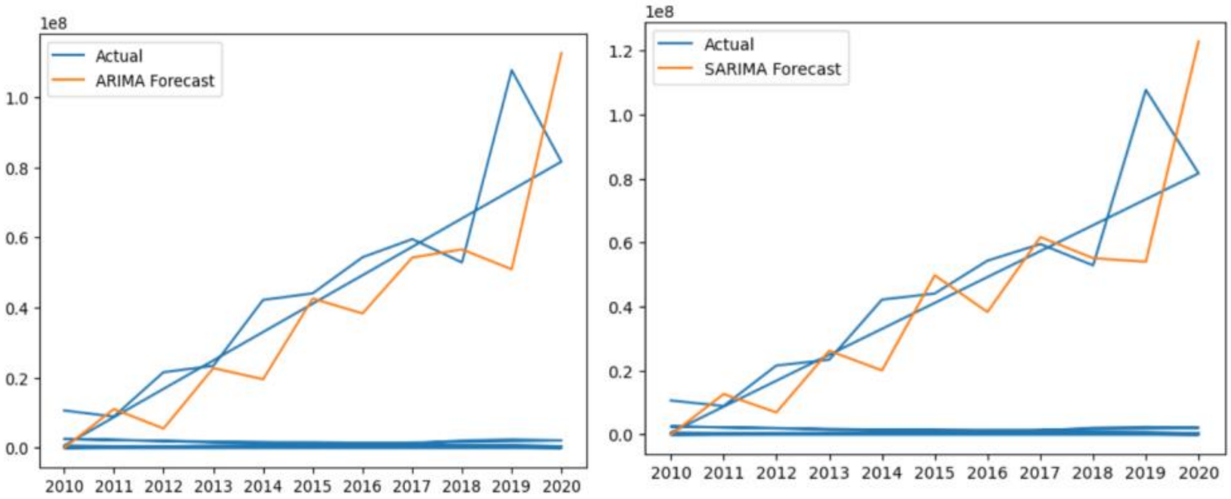


fig) : ARIMA and SARIMA model for case forecasting

The number of confirmed cases of malaria were projected using the ARIMA and SARIMA models. The prediction models demonstrated their dependability and accuracy in predicting future values by closely following the actual value trends.The basis of ARIMA models include time-series analysis, integrated moving average, and autoregression. On the other hand, SARIMA models are seasonal variations of ARIMA models that take into account seasonal variations in data. These models can aid in the allocation of healthcare resources, the selection of sickness prevention and treatment strategies, and other informed decisions.

V. CONCLUSIONS

The statistical analysis concluded that there is casual relationship between years and cases . There is no significance association between indicator code and region. The average mean of total confirmed cases is different form 2010. The average mean of indigenous cases is different from imported cases. The time series analysis indicated that there is seasonality and stationarity for tests conducted variable. The scatter figure reveals a slight but positive association between the quantity of tests and confirmed cases, with the largest numbers in the African region. An increase in instances is shown in Africa and the Eastern Mediterranean region, whereas a drop is seen in South-East Asia, according to the line plot that shows the development of confirmed cases over time. Despite a rise in cases recorded, the line plot also demonstrates a drop in malaria mortality rates over time, which has been linked to improvements in healthcare infrastructure and medical support.

While other measures of model performance should be taken into consideration, the ARIMA and SARIMA models were both shown to be accurate in forecasting future values for the number of confirmed malaria cases over time.

VI. REFERENCES

[1] Adelaja Oluwaseun, Oluwayemisi Nyaaku, Mani Shanker Chaubey, “ANALYSIS OF MALARIA DIAGNOSIS ON PATIENTS USING DATA MINING CLUSTERING TECHNIQUES”, April 2020.

[2] Dr Margaret Chan, “WHO Global Malaria Programme, World Malaria Report”, WHO Press, Geneva, Switzerland, 2013; 1-286.

[3] The World Health Report, “Rolling Back Malaria (The Challenge Of Malaria)”, 1999; 49-63.

[4] Chen Ke, “Machine Learning K-means clustering”, University of Manchester, 1-22.

[5] David Kosiur. 2001. Understanding Policy-Based Networking (2nd. ed.). Wiley, New York, NY..