**MALARIA IN AFRICA DATA SCIENCE PROJECT, GROUP Z**

1. **Introduction**

Malaria, a deadly vector-borne disease caused by Plasmodium parasites, remains a significant public health concern in many regions across the world. Among these regions, Africa bears the heaviest burden, with a disproportionate share of the global malaria cases and deaths.

Efforts to combat malaria in Africa encompass a range of strategies, from vector control (using insecticide treated bed nets) to preventative treatments for high-risk groups (Intermittent Preventative treatment for pregnant women).

In this project, we delve into the multifaceted issue of malaria in Africa, exploring its epidemiology, prevention strategies, and ongoing efforts to combat this pervasive threat.

1. **Dataset description**

The dataset contains the following columns;

* “**Country Name**”: The name of the African country
* “**Year**”: The year of data collection.
* “**Country code**”: A unique code representing each country.
* “**Incidence of malaria (per 1000 population at risk)**”: This metric calculates the rate of malaria cases in relation to the population at risk. It’s expressed as the number of malaria cases per 1000 individuals in the population who are at the risk of contracting malaria. The “population at risk” typically refers to those living in areas with active malaria transmission. This measure helps standardize and compare the malaria burden between countries with different population sizes.
* “**Malaria cases reported**”: This refers to the total number of malaria cases officially documented and reported by health authorities or healthcare facilities in a specific country or region. It is a raw count of malaria cases and does not take into consideration the size of the population or the level of risk. The reported cases may be influenced by factors such as healthcare infrastructure, surveillance systems, and reporting practices.
* “**Use of insecticide-treated bed nets (% of under-5 population)**” : This provides a measure of the percentage of children under the age of 5 who have access to and use insecticide-treated bed nets in a given country.
* “**Children with fever receiving antimalarial drugs (% of children under the age of 5 with fever)**”: This provides a measure of the percentage of children under the age of 5 receiving antimalarial drugs.
* “**Intermittent preventative treatment (IPT) of malaria in pregnancy (% of pregnant women)**”: This a measure of the percentage of pregnant women who receive intermittent preventative treatment for malaria in a given country. This indicator is specifically focused on pregnant women as they are at the risk of malaria-related complications and adverse outcomes during pregnancy.
* “**Rural population (% of total population)**”: Percentage of total population living in rural areas.
* “**Rural population growth (annual %)**”: Annual growth rate of the rural population.
* “**Urban population (% of total population)**”: Percentage of the total population living in urban areas.
* “**Urban population growth (annual %)**”: Annual growth rate of the urban population.
* “**People using at least basic drinking water services (% of population)**”: This indicator is a measure of the percentage of the total population using basic water services as well as using safely managed water services. Basic drinking water services is defined as drinking water from and improved source, provided collection time is not more than 30 minutes for a round trip. Improved water sources include piped water, boreholes, protected dug wells, protected springs, and packaged or delivered water.
* “**People using at least basic drinking water services, rural (% of rural population)**”: This indicator is a measure of the percentage of the rural population using basic water services as well as using safely managed water services. Basic drinking water services is defined as drinking water from and improved source, provided collection time is not more than 30 minutes for a round trip. Improved water sources include piped water, boreholes, protected dug wells, protected springs, and packaged or delivered water.
* “**People using at least basic drinking water services, urban (% of urban population)**”: This indicator is a measure of the percentage of the urban population using basic water services as well as using safely managed water services. Basic drinking water services is defined as drinking water from and improved source, provided collection time is not more than 30 minutes for a round trip. Improved water sources include piped water, boreholes, protected dug wells, protected springs, and packaged or delivered water.
* “**People using at least basic sanitation services (% of population)**”: This indicator is a measure of the percentage of the total population using at least basic sanitation services, that is, improved sanitation facilities that are not shared with other households. It encompasses both people using basic sanitation services as well as those using safely managed sanitation services. Improved sanitation facilities include flush toilets, piped sewer systems, septic tanks or pit latrines, ventilated or improved pit latrines, compositing toilets or pit latrines with slabs.
* “**People using at least basic sanitation services, rural (% of rural population)**”: This indicator is a measure of the percentage of the rural population using at least basic sanitation services, that is, improved sanitation facilities that are not shared with other households. It encompasses both people using basic sanitation services as well as those using safely managed sanitation services. Improved sanitation facilities include flush toilets, piped sewer systems, septic tanks or pit latrines, ventilated or improved pit latrines, compositing toilets or pit latrines with slabs.
* “**People using at least basic sanitation services, urban (% of urban population)**”: This indicator is a measure of the percentage of the urban population using at least basic sanitation services, that is, improved sanitation facilities that are not shared with other households. It encompasses both people using basic sanitation services as well as those using safely managed sanitation services. Improved sanitation facilities include flush toilets, piped sewer systems, septic tanks or pit latrines, ventilated or improved pit latrines, compositing toilets or pit latrines with slabs.
* “**Latitude**”: The latitude coordinate of the country’s geographical location.
* “**Longitude**”: The longitude coordinate of the country’s geographical location.
* “**Geometry**”: Geographical geometry data for mapping purposes.

1. **Project Objectives**

The main objectives of this data science project are as follows;

* Identify countries with the highest and lowest malaria incidence rates, aiding in targeting interventions and allowing international organizations prioritize resource allocation.
* Compare the burden of malaria across different African countries, helping prioritize regions for preventative measures and healthcare infrastructure.
* Assess the impact of geographical location on malaria incidence.
* Track and visualize the trend of malaria incidences over the years and evaluate the effectiveness of existing malaria control and prevention measures.
* Assess the effectiveness of malaria prevention campaigns and programs promoting bed net usage for children under the age of 5 years.
* Identify countries with low bed net usage targeting them for interventions to increase usage and reduce malaria transmission.
* Investigate the impact of Insecticide-treated bed nets on the percentages of children with fever under the age of 5 receiving antimalarial drugs.
* Explore the relationship between the use of insecticide-treated bed nets in children under the age of 5 and the percentage of children under the age of 5 with fever receiving antimalarial drugs
* Examine the coverage and implementation of Intermittent Preventative treatment for pregnant women, crucial for preventing malaria related complications during pregnancy.
* Assess the adoption of Intermittent Preventative treatment over the years as a preventative measure against malaria among pregnant women.
* Identify regions with lower IPT coverage, targeting them for interventions to improve maternal and child health outcomes.
* Explore the relationship between the percentage of pregnant women using Intermittent Preventative treatment and the number of malaria incidences.
* Showcase the distribution of malaria incidences across the African countries.
* Explore the relationship between the use of basic sanitation services and malaria incidences.
* Compare the reporting and surveillance systems for malaria across the African countries.
* Assess the extent to which reported cases accurately reflect the true burden of malaria in the country with the highest number of reported cases.
* Assess the impact of population distribution on the number of reported cases in the country with the highest number of reported cases.
* Assess the impact of population distribution on the number of Incidences per 1000 population at risk in the country with the highest number of incidences per 1000 population at risk.
* Assess the impact of annual growth rates on the number of Incidences per 1000 population at risk in the country with the highest number of incidences per 1000 population at risk.
* Investigate the impact of access to basic drinking water and sanitation services on the number of Incidences per 1000 population at risk in the country with the highest number of Incidences per 1000 population at risk.
* Provide a visual representation of the malaria burden across African countries, highlighting malaria hotspots.
* Create a model to predict the number of Incidences per 1000 population at risk using the most influential factors.

1. **Methodology**
2. **Data Cleaning**

Data cleaning is a crucial step in preparing our raw data for analysis and modelling. In this section, we detail the steps taken to clean the data, handle missing values, remove any duplicates and address any inconsistencies.

**Data sources**

Raw data source: “<https://www.kaggle.com/datasets/lydia70/malaria-in-africa>”

Data format: CSV

**Data cleaning steps**

* **Renaming columns.**

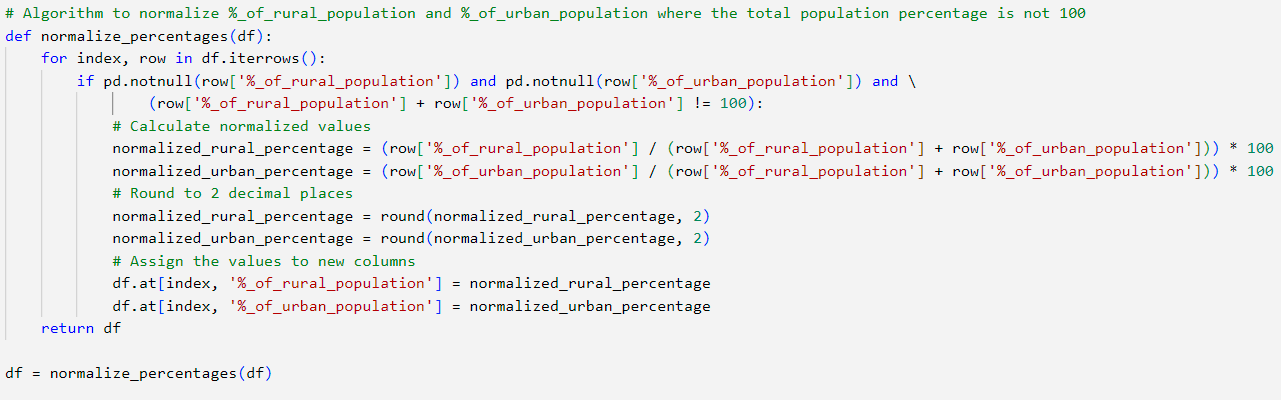
We shortened certain column names to enhance readability and maintain consistency.

|  |  |
| --- | --- |
| **Original column name** | **Shortened column name** |
| Country Name | Country |
| Malaria Cases Reported | Reported\_cases |
| Incidence of malaria (per 1000 population at risk) | Incidences\_per\_1000\_population\_at\_risk |
| Intermittent Preventative treatment (IPT) of malaria in pregnancy (% of pregnant women) | %\_of\_pregnant\_women\_using\_IPT |
| Use of insecticide-treated bed nets (% of under-5 population) | %\_using\_IBNs |
| Children with fever receiving antimalarial drugs (% of children under age 5 with fever) | %\_of\_children\_under\_age\_5\_with\_fever |
| Rural population (% of total population) | %\_of\_rural\_population |
| Urban population (% of total population) | %\_of\_urban\_population |
| Rural population growth (annual %) | annual\_%\_growth\_of\_rural\_population |
| Urban population growth (annual %) | annual\_%\_growth\_of\_urban\_population |
| People using safely managed sanitation services (% of population) | %\_using\_safe\_sanity\_services |
| People using safely managed sanitation services, urban (% of urban population) | %\_of\_urban\_using\_safe\_sanity\_services |
| People using safely managed sanitation services, rural (% of rural population) | %\_of\_rural\_using\_safe\_sanity\_services |
| People using at least basic sanitation services (% of population) | %\_using\_atleast\_basic\_sanity\_services |
| People using at least basic sanitation services, urban (% of urban population) | %\_of\_urban\_using\_atleast\_basic\_sanity\_services |
| People using at least basic sanitation services, rural (% of rural population) | %\_of\_rural\_using\_atleast\_basic\_sanity\_services |
| People using at least basic drinking water services (% of population) | %\_using\_atleast\_basic\_drinking\_water\_services |
| People using at least basic drinking water services, urban (% of urban population) | %\_of\_urban\_using\_atleast\_basic\_drinking\_water\_services |
| People using at least basic drinking water services, rural (% of rural population) | %\_of\_rural\_using\_atleast\_basic\_drinking\_water\_services |
| People using safely managed drinking water services (% of population) | %\_using\_safe\_drinking\_water\_services |
| People using safely managed drinking water services, urban (% of urban population) | %\_of\_urban\_using\_safe\_drinking\_water\_services |
| People using safely managed drinking water services, rural (% of rural population) | %\_of\_rural\_using\_safe\_drinking\_water\_services |

* **Addressing inconsistencies.**

We identified rows where the total population percentage was not 100.

We created an algorithm to normalize the values.

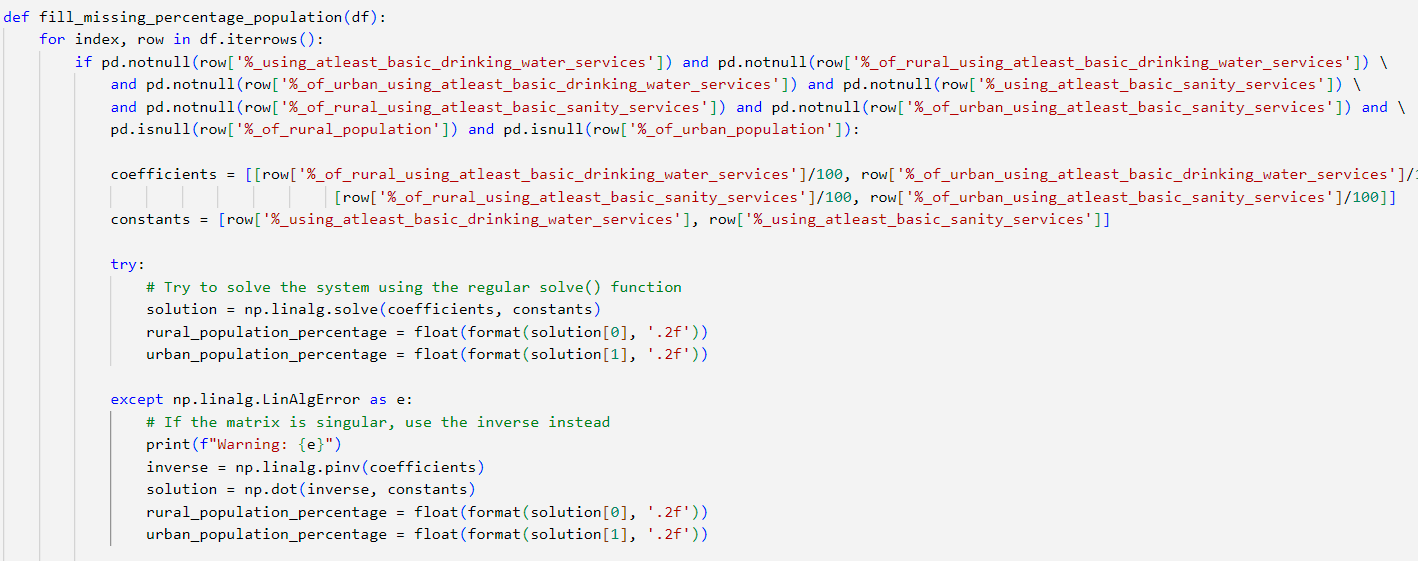


* **Handling missing values**

We discovered that some of the columns are related by mathematical formulas listed below,

1. %\_using\_safe\_sanity\_services = (%\_of\_rural\_using\_safe\_sanity\_services \* %\_of\_rural\_population) + (%\_of\_urban\_using\_safe\_sanity\_services \* %\_of\_urban\_population)/100
2. %\_using\_atleast\_basic\_drinking\_water\_services = (%\_of\_rural\_using\_atleast\_basic\_drinking\_water\_services \* %\_of\_rural\_population) + (%\_of\_urban\_using\_atleast\_basic\_drinking\_water\_services \* %\_of\_urban\_population)/100
3. %\_using\_atleast\_basic\_sanity\_services = (%\_of\_rural\_using\_atleast\_basic\_sanity\_services \* %\_of\_rural\_population) + (%\_of\_urban\_using\_atleast\_basic\_sanity\_services \* %\_of\_urban\_population)/100
4. %\_using\_safe\_drinking\_water\_services = (%\_of\_rural\_using\_safe\_drinking\_water\_services \* %\_of\_rural\_population) + (%\_of\_urban\_using\_safe\_drinking\_water\_services \* %\_of\_urban\_population)/100

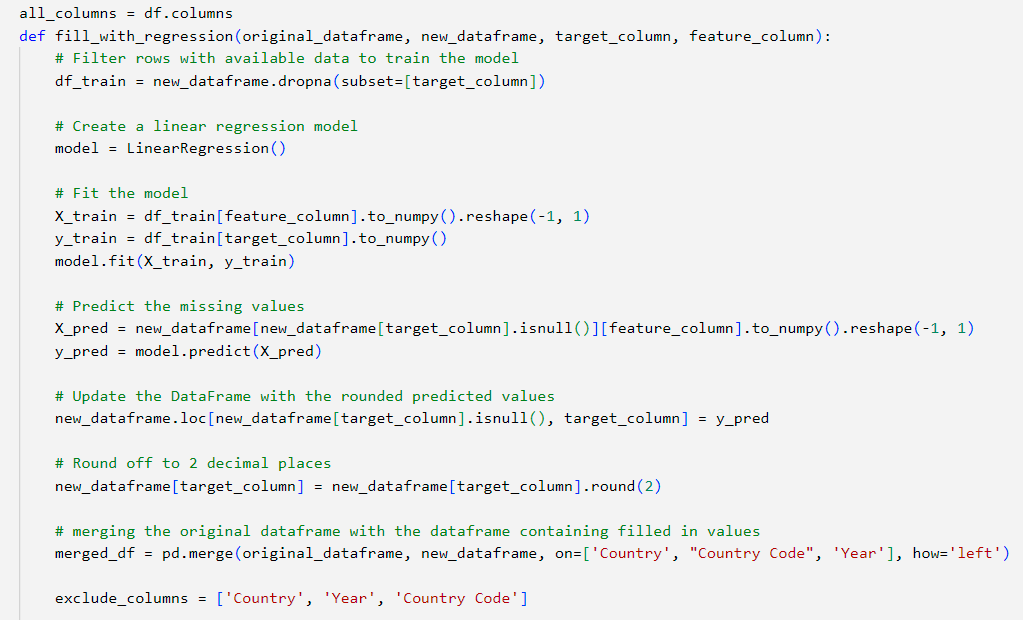
We created an algorithm, “fill\_missing\_percentage\_columns()”, to fill in the missing values for theses columns using already existing data, while utilizing these 4 mathematical equations.

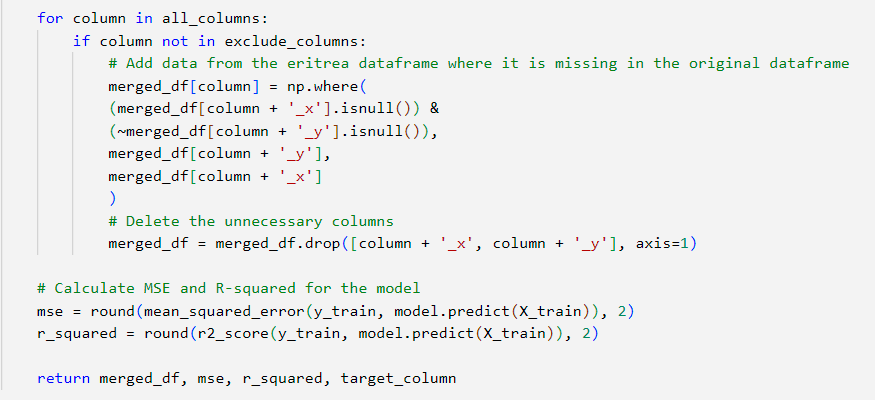




The missing values, not filled in with the first algorithm

“fill\_missing\_percentage\_columns()”, where filled in by the second regression algorithm “fill\_with\_regression()” that uses existing values to predict missing values.

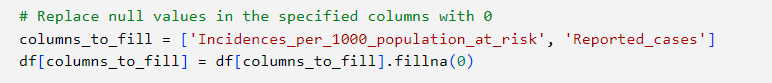




Mean and mode imputation were not used because the data is time series data, which implies that, the data has trends or patterns hence using mean and mode imputation would distort the underlying trends.

The regression algorithm provides the **Mean square error values**, as well as, the **r-squared value** for each model created.

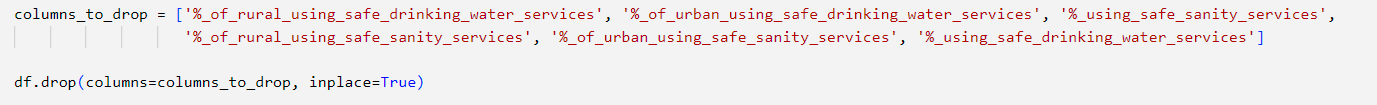
The missing values for 'Incidences\_per\_1000\_population\_at\_risk', 'Reported\_cases' were filled with 0, because filling missing values for one country using other countries data, was not logical.



* **Dropping columns**

'%\_of\_rural\_using\_safe\_drinking\_water\_services', '%\_of\_urban\_using\_safe\_drinking\_water\_services', '%\_using\_safe\_sanity\_services', '%\_of\_rural\_using\_safe\_sanity\_services' and '%\_of\_urban\_using\_safe\_sanity\_services' were dropped because;

1. High percentage of missing values.
2. They could be represented by other columns within the same dataset. According to the World Bank Open Data Source, “The percentage of people using at least basic services encompasses both people using basic services, as well as, those using safely managed services”.



1. **Pattern discovery**

The following tools and Libraries were used;

* Python
* Matplotlib
* Pandas

The visualizations are explained below;

1. **Bar graph showing Malaria incidences per Country.**

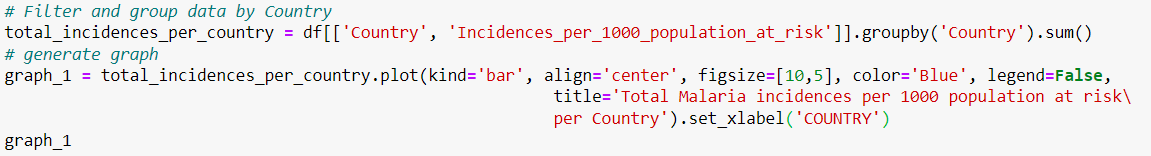
**Objective**:

Compare the burden of malaria across the African countries, helping prioritize regions for preventative measures and healthcare infrastructure.

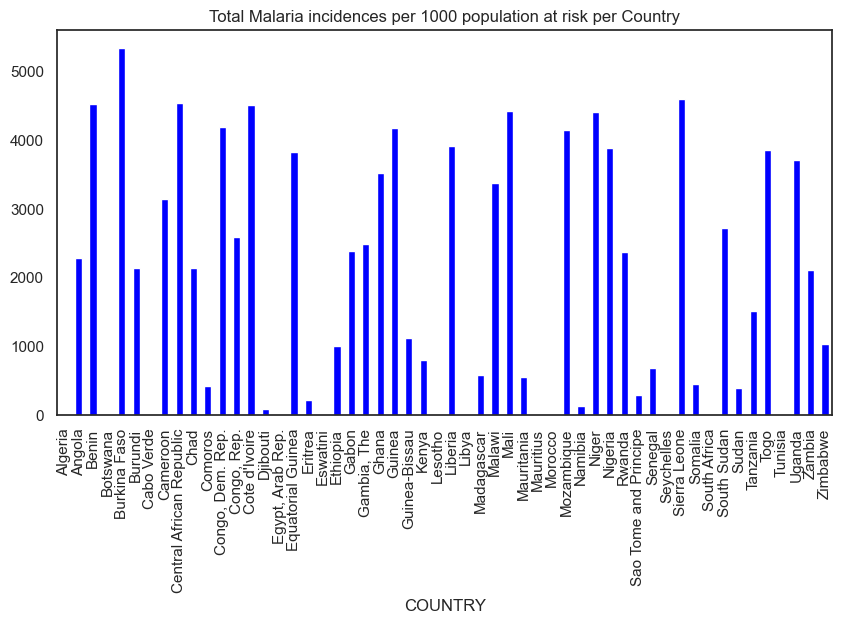
**Observation**:

The varying heights of the bars signify a diverse burden of malaria across different African countries.

**Code snippet**:



**Image**:



1. **Horizontal bar graph showing the top 10 countries with the highest number of incidences per 1000 population at risk.**

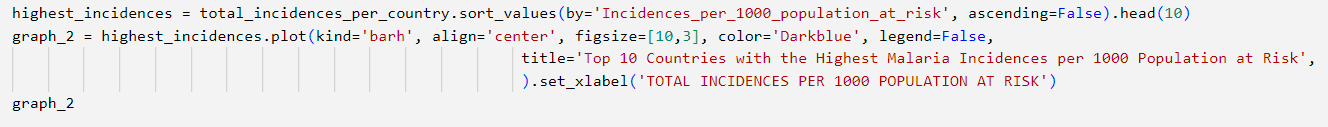
**Objective**:

Identify countries with the highest number of incidences, aiding in targeting interventions and allow international organizations and donors prioritize resource allocation.

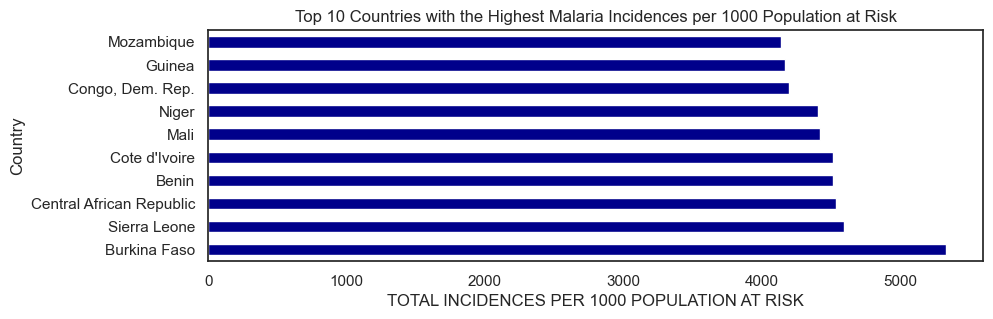
**Observation**:

Mozambique, Guinea, Democratic Republic of Congo, Niger, Mali, Benin, Central African Republic, Sierra Leone, Cote d’Ivoire and Burkina Faso have the highest incidences.

**Code Snippet**:



**Image**:



1. **Horizontal bar graph showing the countries with the lowest number of incidences per 1000 population at risk.**

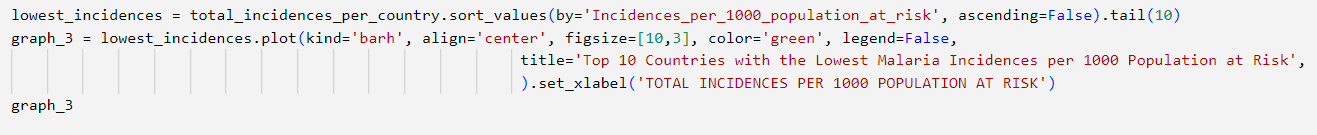
**Objective:**

Identify countries with the highest number of incidences, aiding in targeting interventions and allow international organizations and donors prioritize resource allocation.

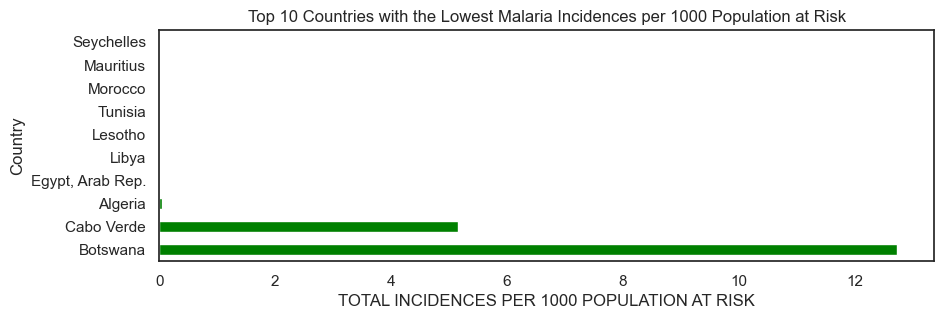
**Observation:**

Algeria, Tunisia, Lesotho, Libya, Morocco, Mauritius, Seychelles, Egypt, Botswana and Cabo Verde have the lowest incidences.

**Code Snippet**:



**Image**:



1. **Map of Africa showing the location of the top 10 countries with the highest number of incidences per 1000 population at risk.**

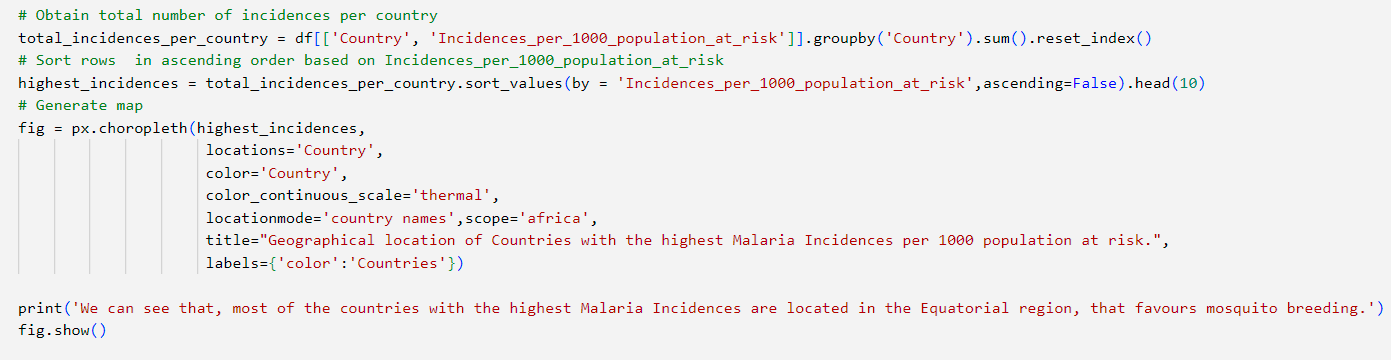
**Objective:**

Assess the impact of geographical location on malaria incidence.

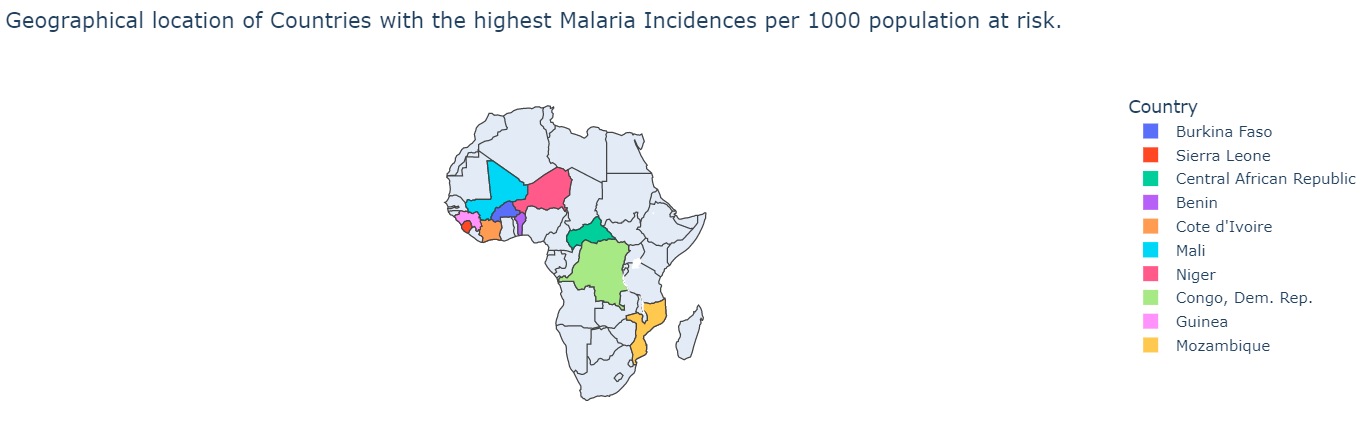
**Observation:**

They are located in the Equatorial region.

**Code snippet**:



**Image**:



1. **Map of Africa showing the location of the top 10 countries with the least number of incidences per 1000 population at risk.**

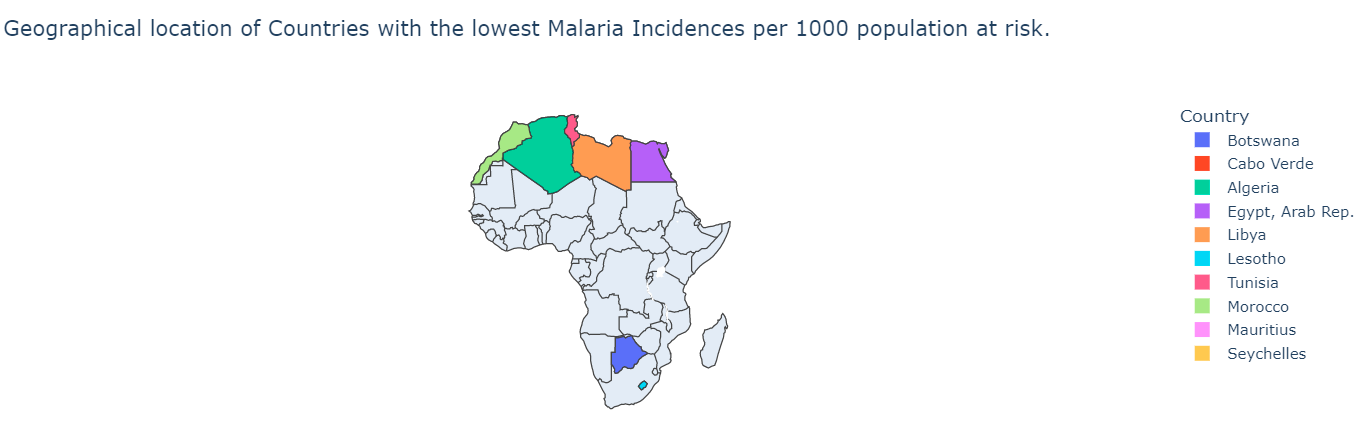
**Objective:**

Assess the impact of geographical location on malaria incidence.

**Observation:**

They are located in Northern Africa.

**Code snippet**:



1. **Line chart showing the trend in the number of Incidences per 1000 population at risk over the years.**

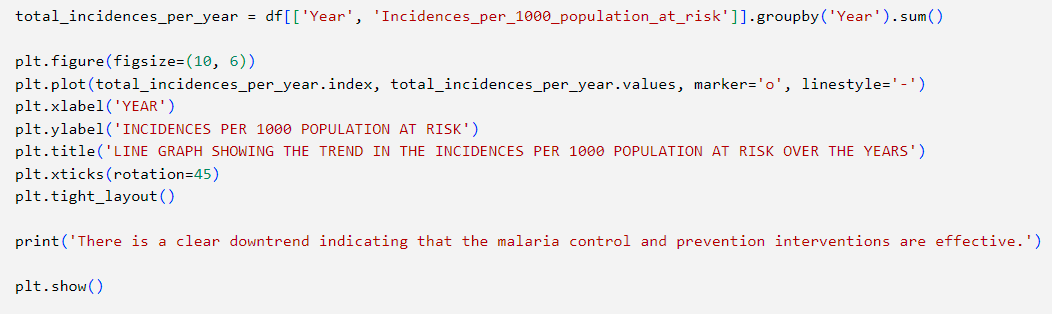
**Objective:**

Evaluate the effectiveness of existing malaria prevention and control interventions.

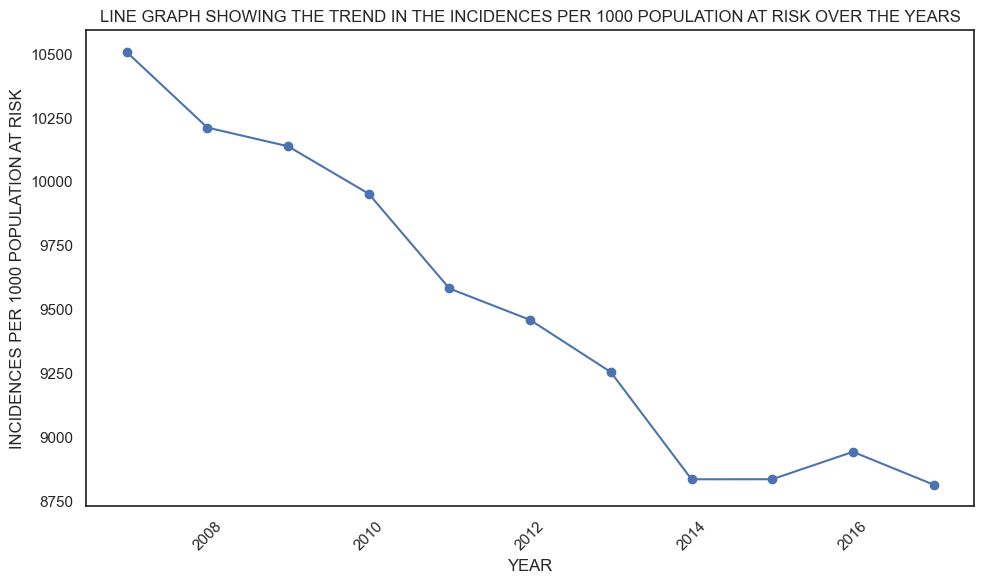
**Observation**:

The interventions have been a success.

**Code Snippet**:



**Image**:



1. **Line chart showing the trend in the usage of insecticide-treated bed nets among children under 5 years over the years.**

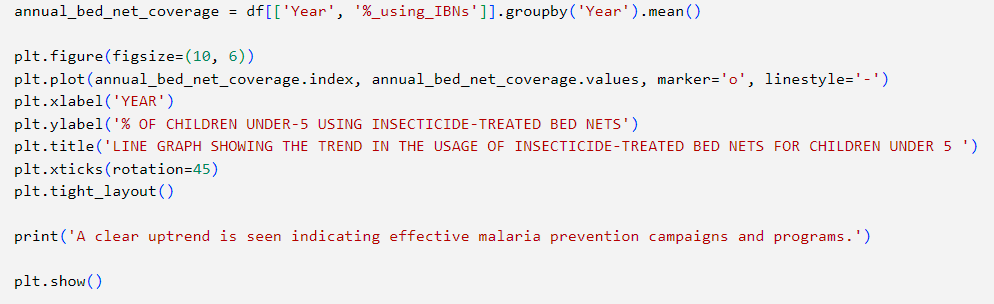
**Objective:**

Assess the effectiveness of malaria prevention campaigns and programs promoting bed net usage for children under the age of 5 years.

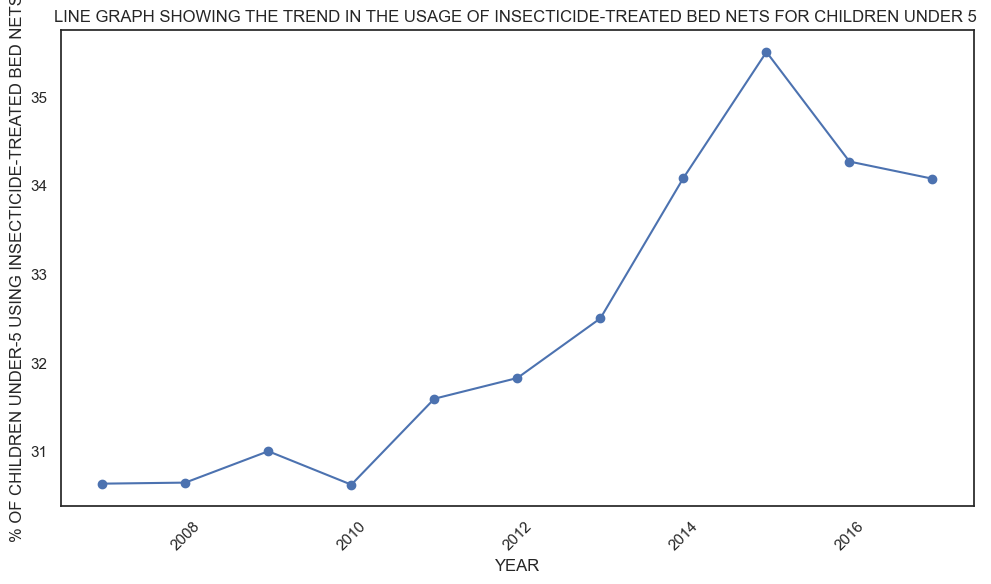
**Observation**:

We see an uptrend in the usage of insecticide-treated bed nets hence the campaigns and programs promoting the usage were successful.

**Code snippet**:



**Image**:



1. **Horizontal bar graph showing the top 10 countries with the lowest percentage of children under the age of 5 years using insecticide treated bed nets.**

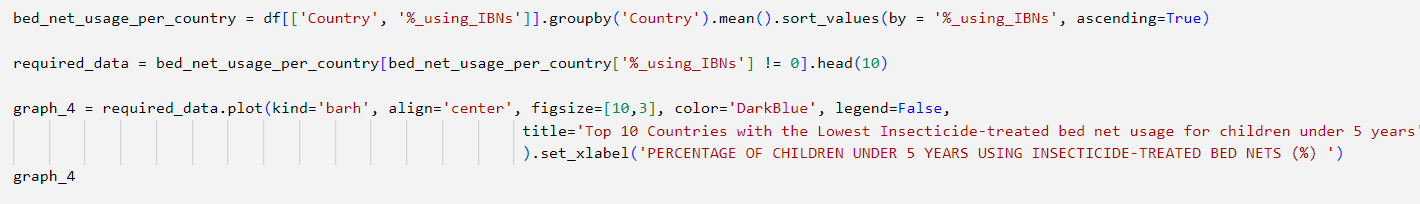
**Objective:**

Identify countries with low bed net coverage targeting them for interventions to increase usage and reduce malaria transmission.

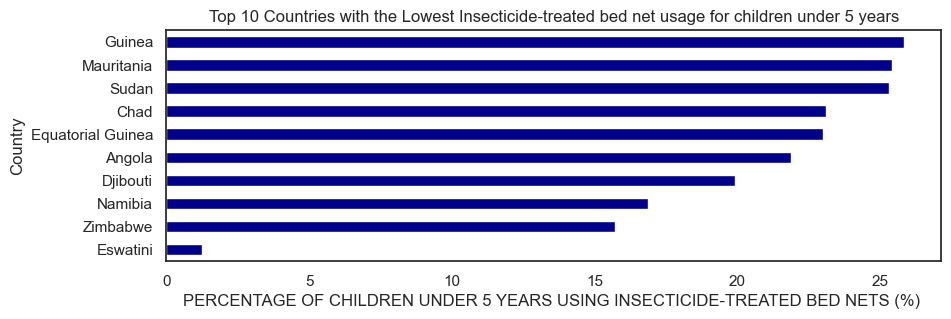
**Observation**:

Sudan, Chad, Mauritania, Guinea, Angola, Namibia, Equatorial Guinea, Zimbabwe, Djibouti, Eswatini have the lowest percentage of children under 5 years using insecticide treated bed nets.

**Code Snippet**:



**Image**:



1. **Line graph showing the trend of Insecticide-treated bed net usage in children under 5 versus the trend of the percentage of children with fever under the age of 5 years receiving antimalarial drugs.**

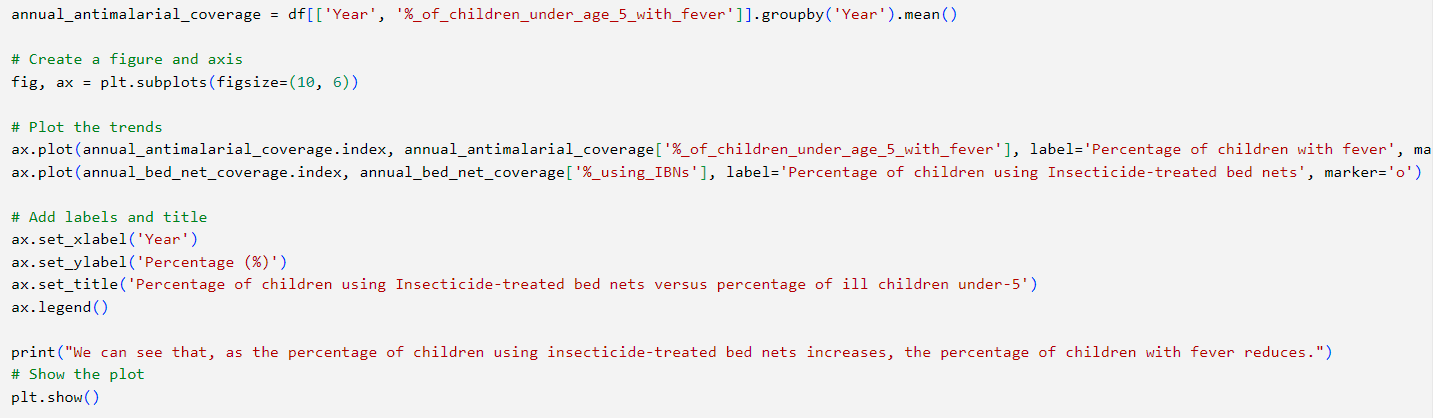
**Objective**:

Investigate the impact of insecticide-treated bed net usage in children under 5 on the percentages of sick children receiving antimalarials

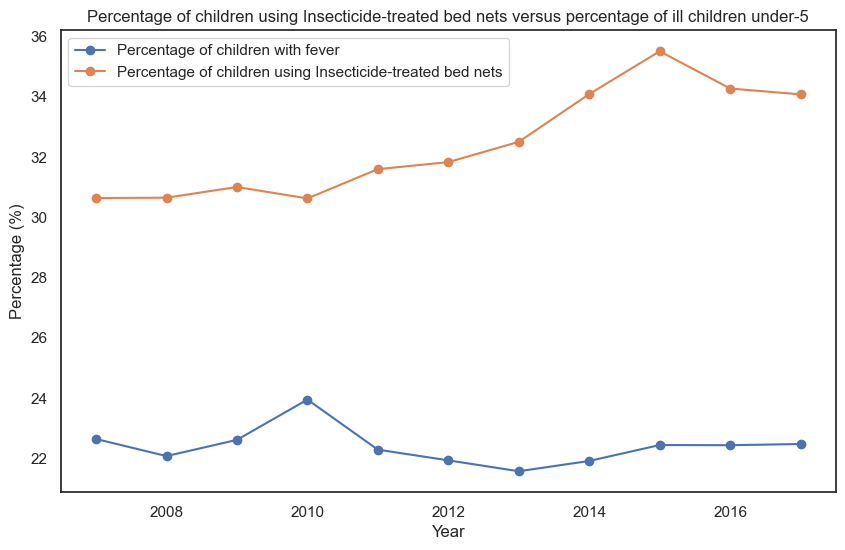
**Observation**:

As the percentage of children under 5 using the insecticide treated bed nets increases, the number of sick children under 5 receiving antimalarials reduces.

**Code Snippet**:



**Image**:



1. Correlation heatmap showing the relationship between the percentages of children under 5 using insecticide treated bed nets and the percentage of children with fever receiving antimalarials

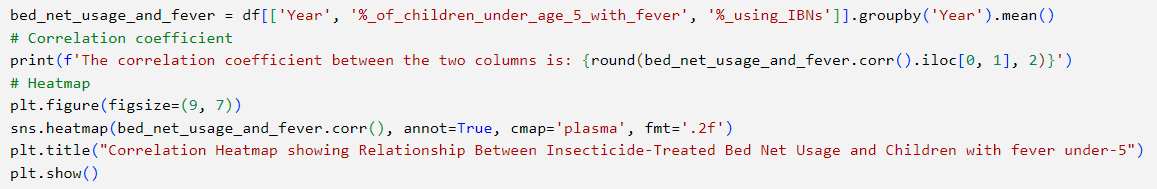
**Objective**:

Explore the relationship between the usage of insecticide-treated bed nets and the percentages of sick children getting antimalarials.

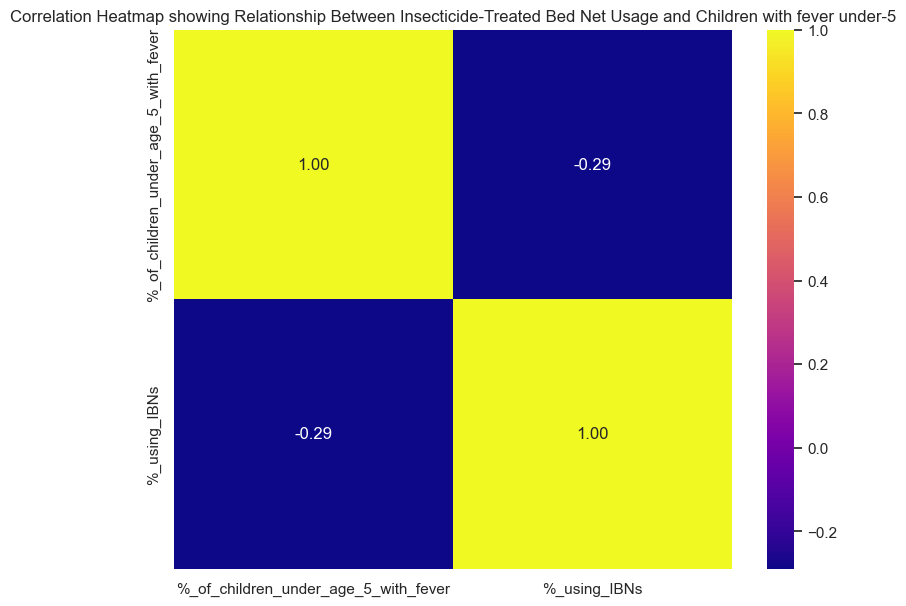
**Observation**:

Weak negative correlation.

**Code Snippet**:



**Image**:



1. **Bar graph showing the percentages of pregnant women using IPT per Country.**

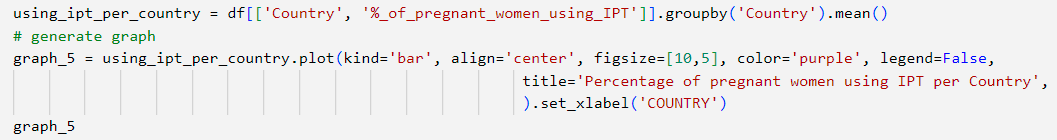
**Objective**:

Examine the coverage and implementation of Intermittent preventative treatment for pregnant women, crucial for preventing malaria related complications during pregnancy.

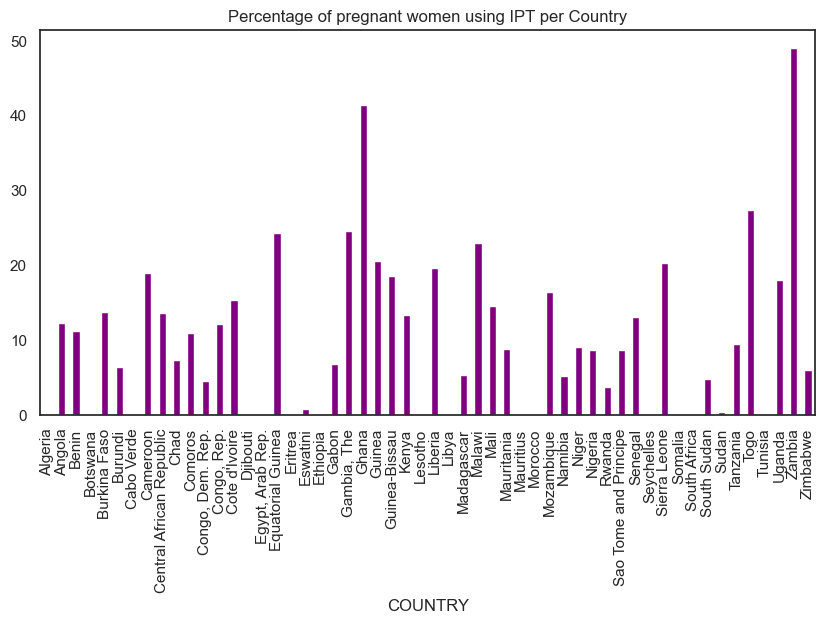
**Observation**:

The varying bar lengths emphasize disparities in IPT coverage and the need for targeted interventions.

**Code Snippet**:



**Image**:



1. **Line chart showing the trend in the usage of Intermittent Preventative treatment among pregnant women over the years.**

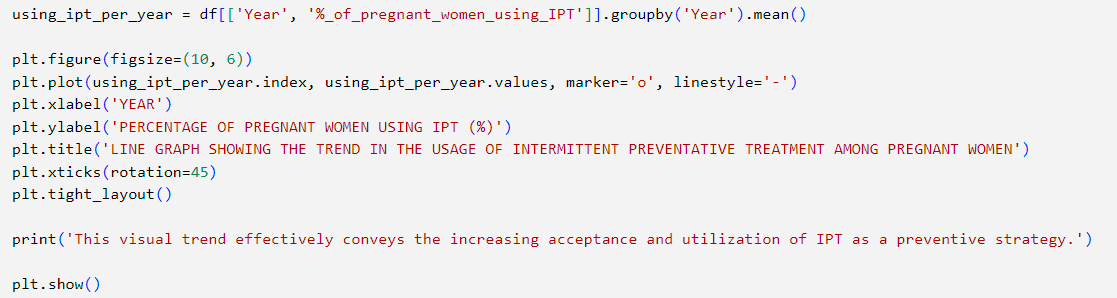
**Objective**:

Assess the adoption of IPT over the years as a preventative measure against malaria among pregnant women.

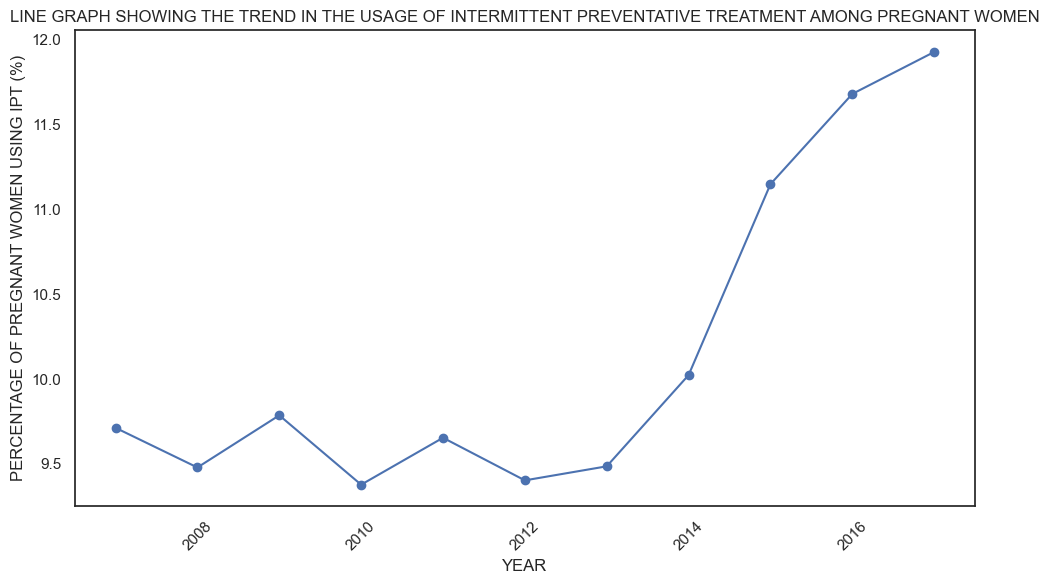
**Observation**:

The uptrend suggests that efforts to promote and increase the utilization of IPT have been effective.

**Code snippet**:



**Image**:



1. **Bar graph showing the top 10 countries with the lowest levels of Intermittent preventative treatment usage among pregnant women.**

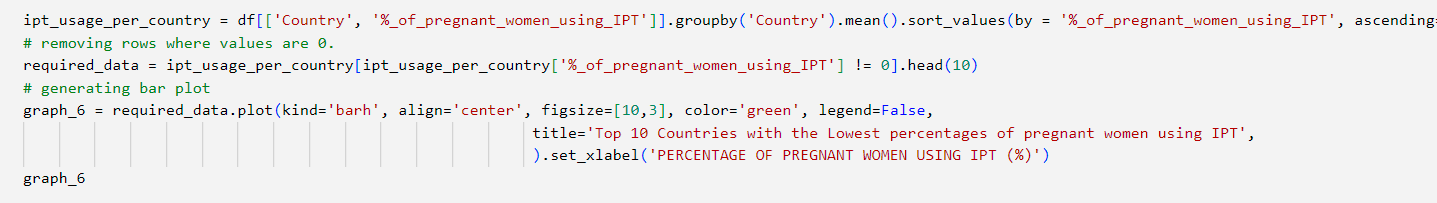
**Objective**:

Identify regions with low IPT coverage, requiring targeted efforts to improve maternal and child health outcomes.

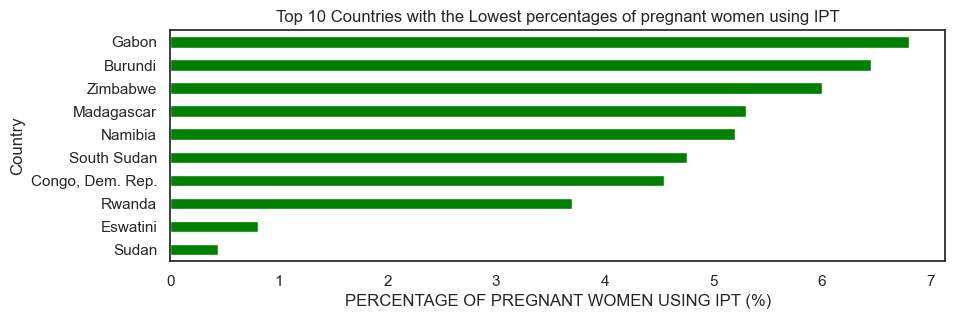
**Observation**:

Gabon, Burundi, Namibia, Madagascar, Democratic republic of Congo, South Sudan, Rwanda, Eswatini, Sudan

**Code Snippet**:



**Image**:



1. **Correlation heatmap showing the relationship between the percentage of pregnant women using IPT and Incidences per 1000 population at risk.**

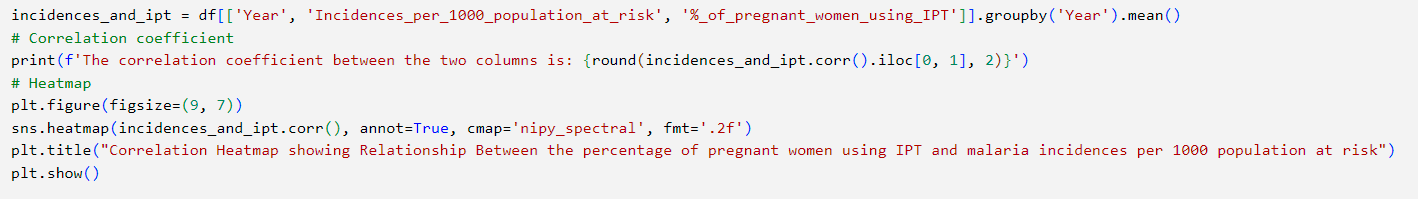
**Objective**:

Explore the relationship between the percentage of pregnant women using IPT and Incidences per 1000 population at risk.

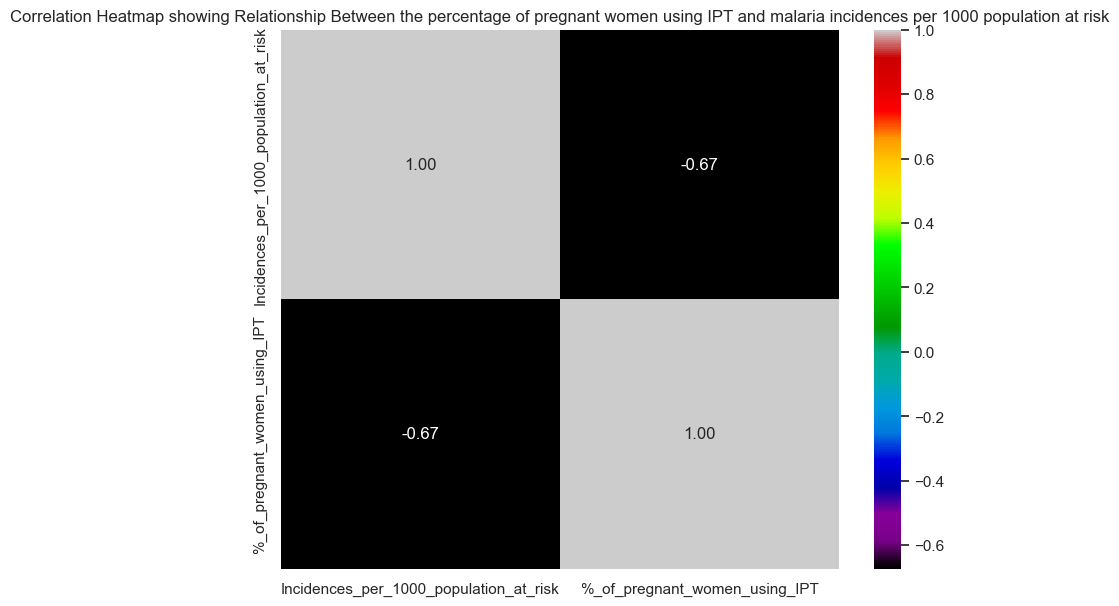
**Observation**:

Moderate negative correlation.

**Code Snippet**:



**Image**:

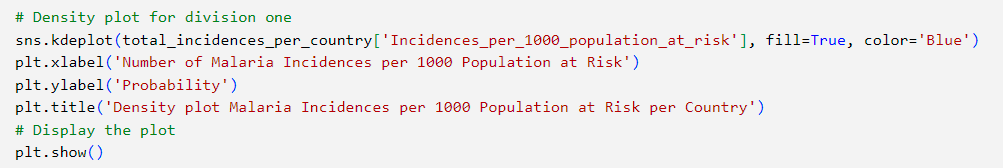


1. **Kernel density Distribution plot to visualize the probability distribution of malaria incidences across African countries.**

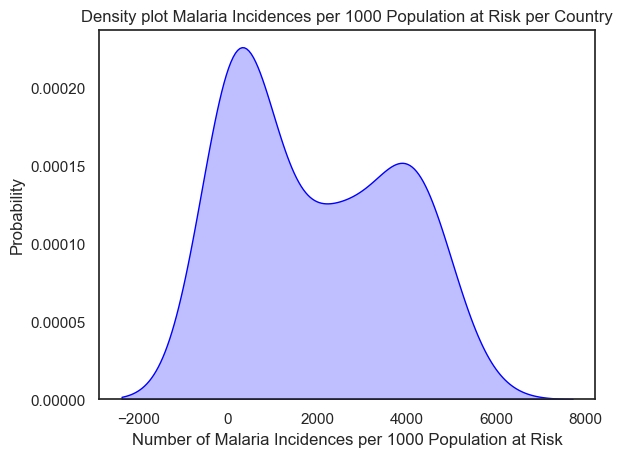
**Observation**

Bimodal distribution seen highlights the fact that malaria incidence rates vary significantly among African countries.

**Code Snippet**:



**Image**:



1. **Correlation heatmap showing the relationship between the percentage using basic sanitation services and Incidences per 1000 population at risk**

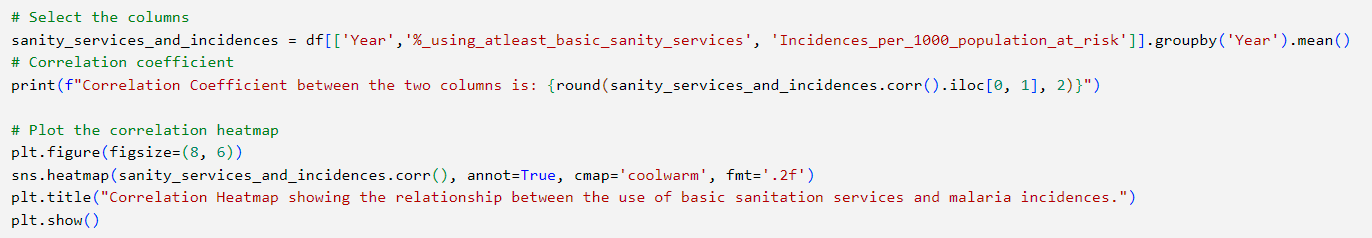
**Objective**:

Exploring the relationship between the use Basic sanitation services and malaria incidences.

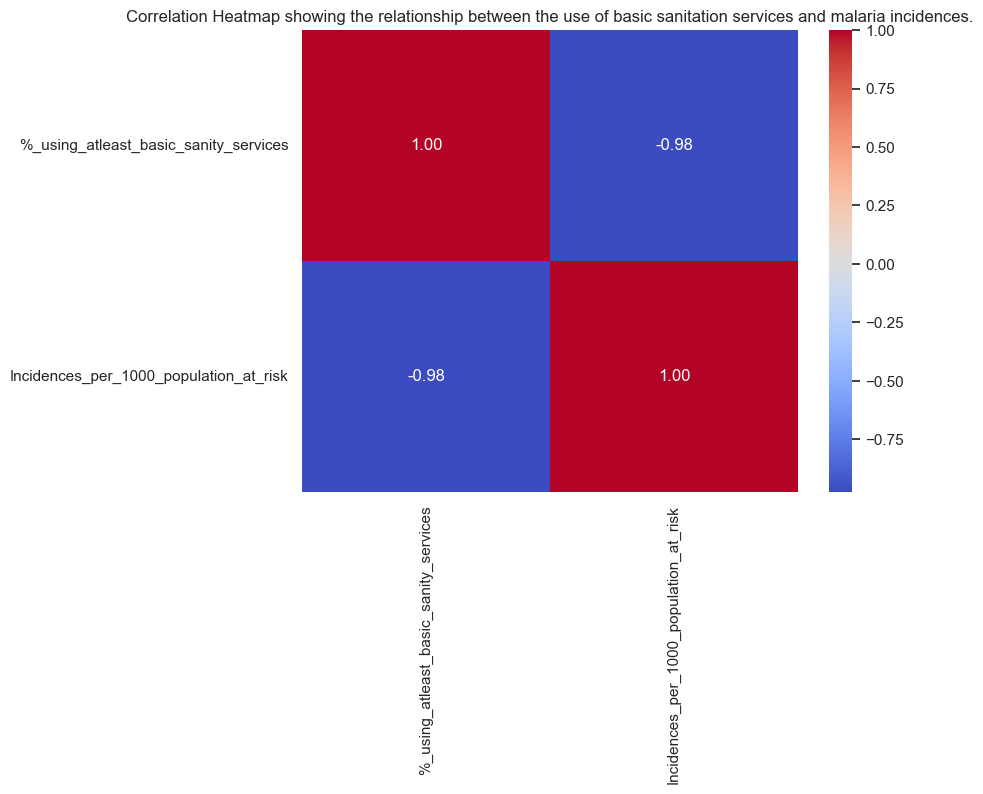
**Observation**

The value of -0.98 suggests that the is strong negative correlation between them.

**Code Snippet**:



**Image**:



1. Correlation heatmap showing the relationship between the percentage of people using basic drinking water services and the number of incidences per 1000 population at risk

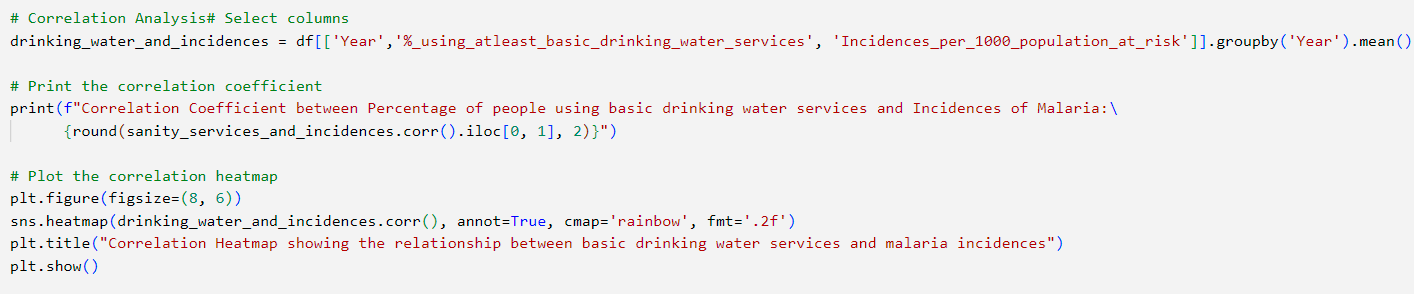
**Objective**:

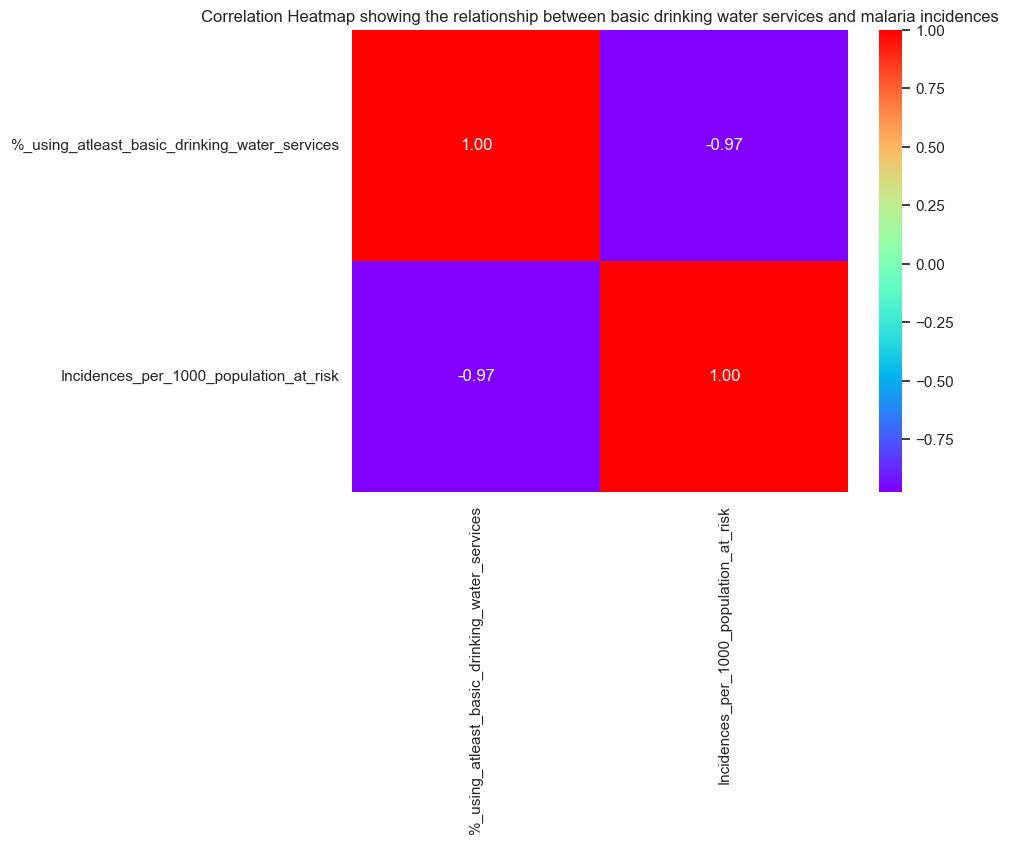
Explore the relationship between the use of basic drinking water services and malaria incidences per 1000 population at risk.

**Observation**

Value of -0.98 shows strong negative correlation between them

**Code Snippet**:





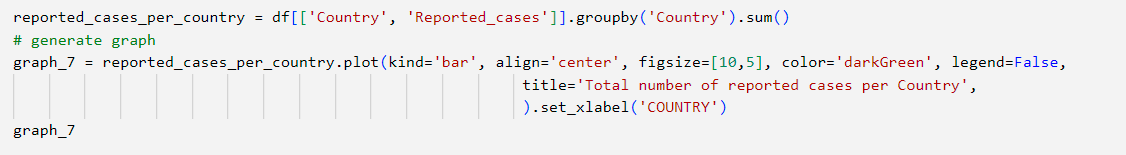
1. **Bar graph showing reported cases per country**

**Objective**:

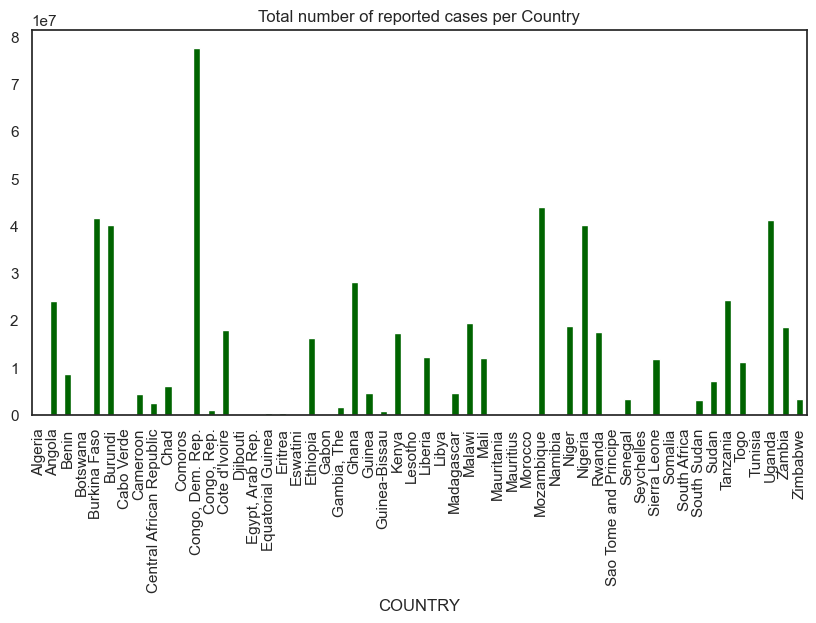
Compare the reporting and surveillance systems for malaria across the African country

**Observation**:

**Code Snippet**:



**Image:**

****

1. **Line graph showing the trend of reported cases versus Incidences per 1000 population at risk in the country with the highest number of Incidences.**

**Objective:**

Assess the extent to which reported cases accurately reflect the true burden of malaria in the country with the highest number of incidences

**Observation**:

As the number of Incidences per 1000 population at risk reduces, the number of reported cases increases.

**Code snippet**:

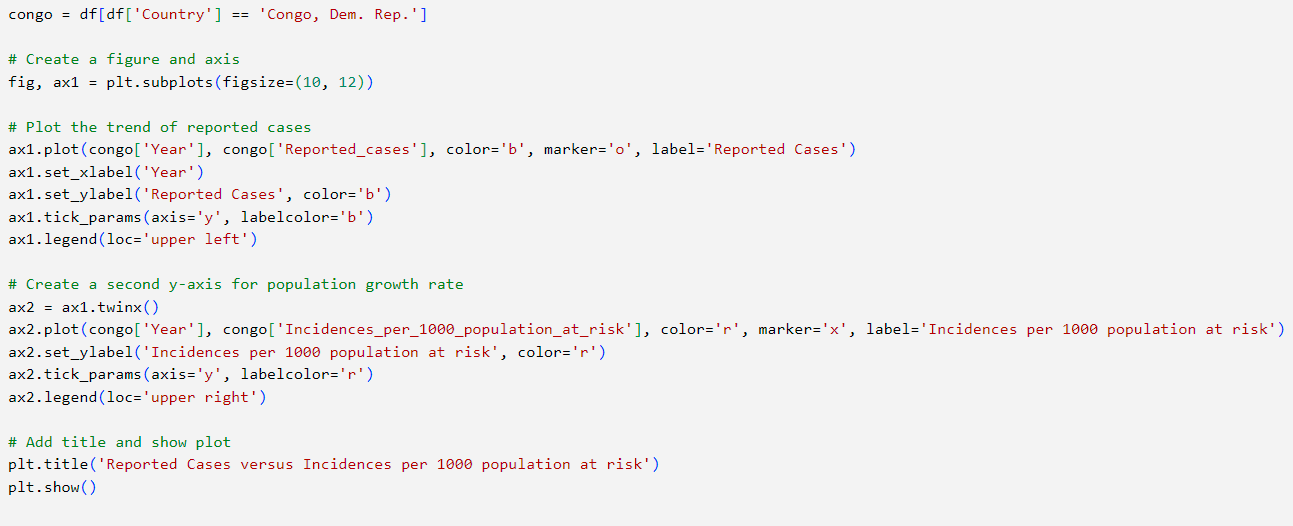
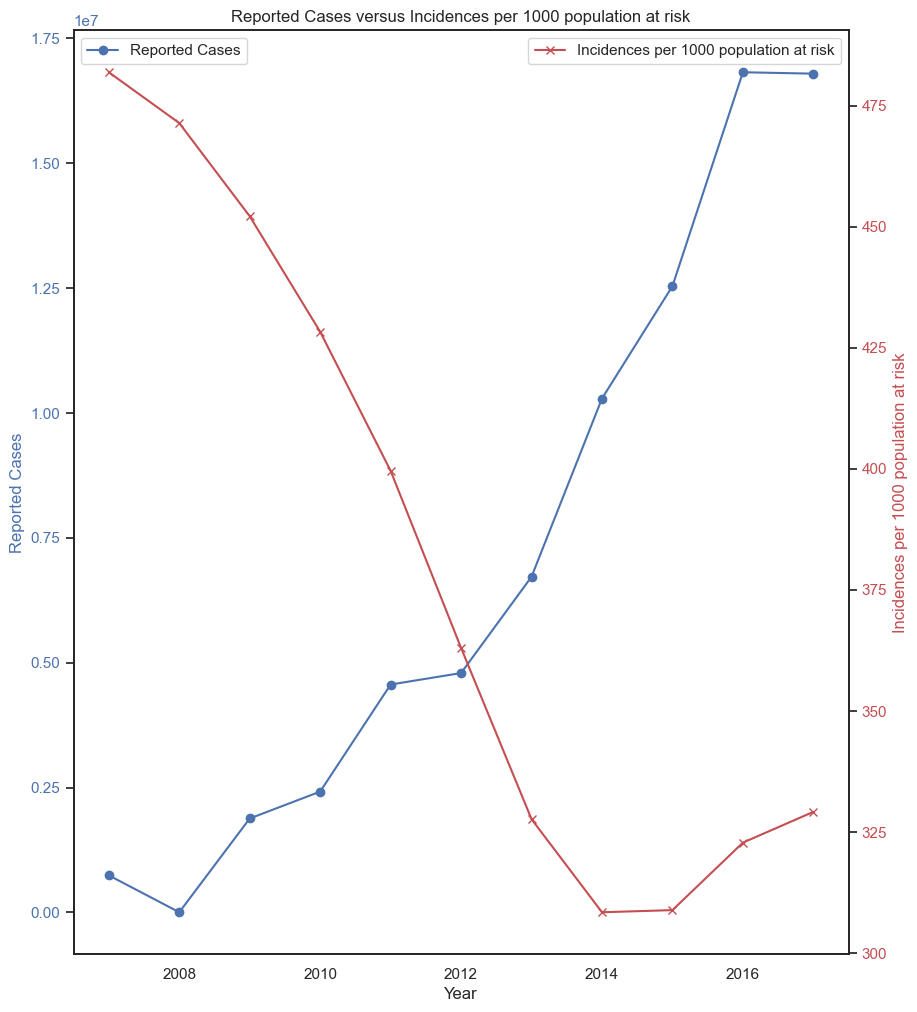


Image:

1. **A line graph showing the trend in the number of reported cases versus the percentages of urban and rural population in the country with the highest number of reported cases**

**Objective:**

Assess the impact of population distribution on the number of reported cases in the country with the highest number of reported cases.

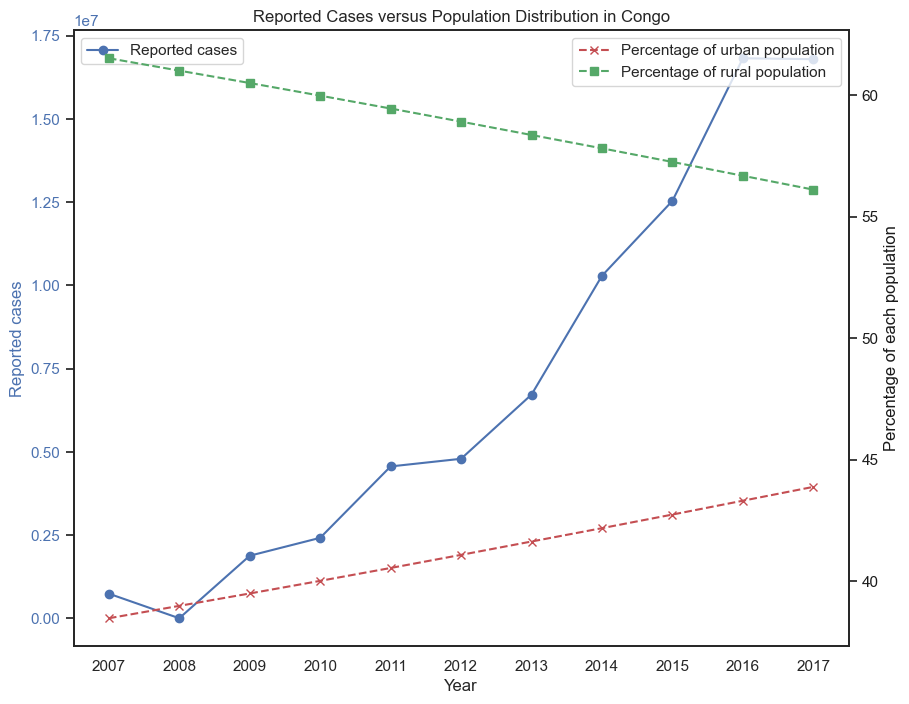
**Observation**:

The number of reported cases increases with increase in the percentage of the urban population.

**Code Snippet**;



Image:



1. **A line graph showing the trend in the Incidences per 1000 population a risk and the trend in the rural and urban population percentages in the country with the highest number of incidences.**

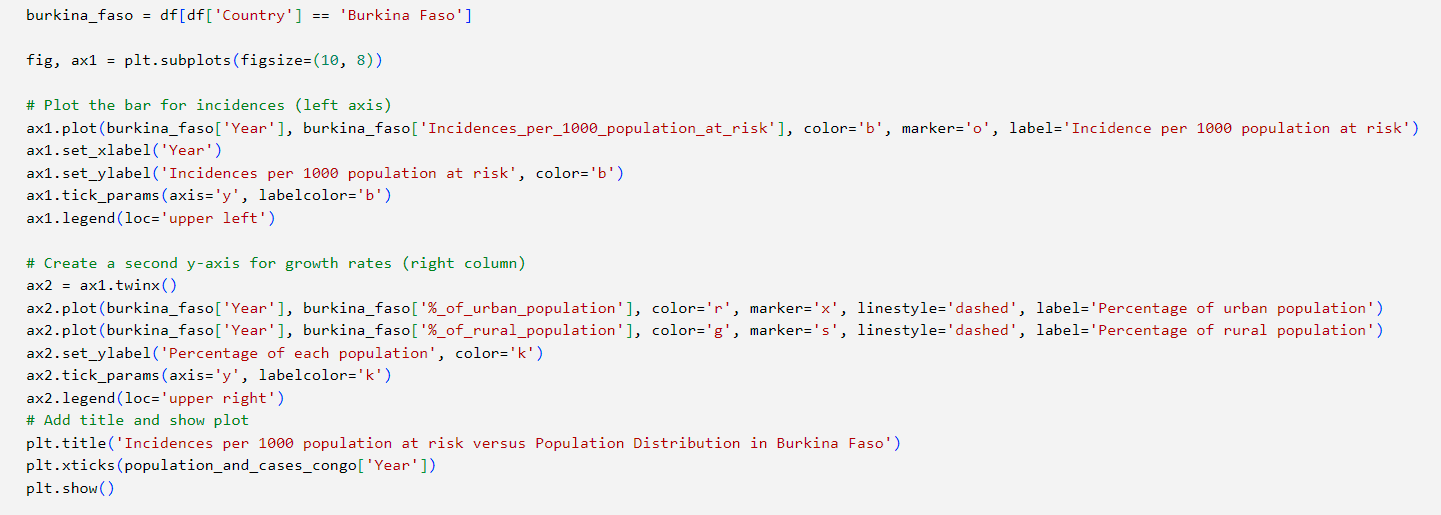
**Objective**:

Assess the impact of population distribution on the number of incidences per 1000 population at risk in the country with the highest number of incidences per 1000 population at risk

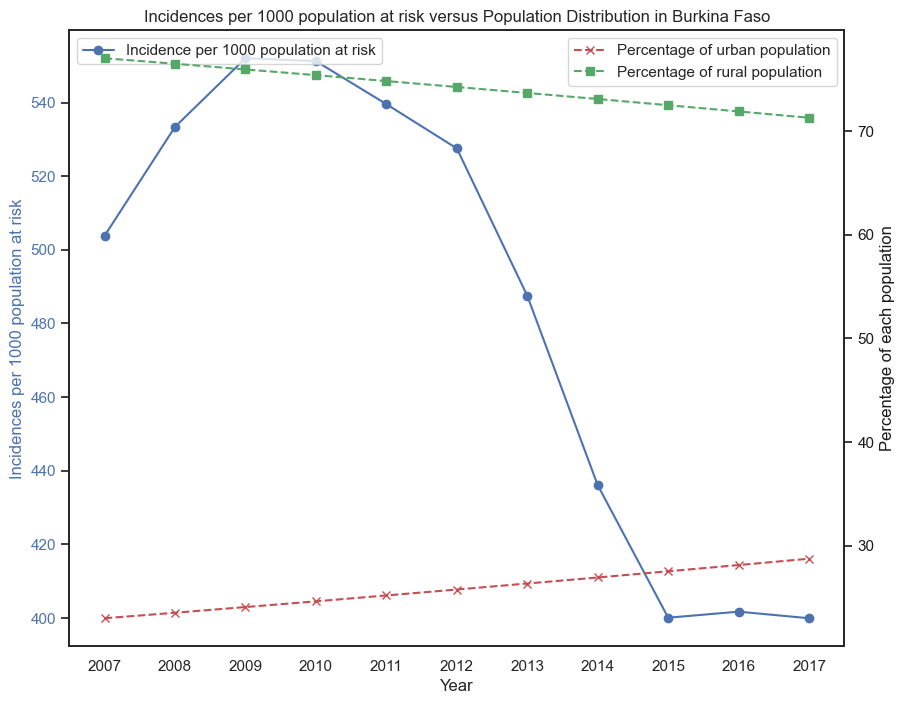
**Observation**:

As the percentage of the urban population increases, the number of Incidences\_per\_1000 population at risk decreases

**Code Snippet**:



**Image**:



1. **A line graph showing the trend in the Incidences per 1000 population a risk and the trend in the percentage annual growth rates in the country with the highest number of incidences.**

**Objective**:

Assess the impact of percentage annual growth rates on the number of incidences per 1000 population at risk in the country with the highest number of incidences per 1000 population at risk

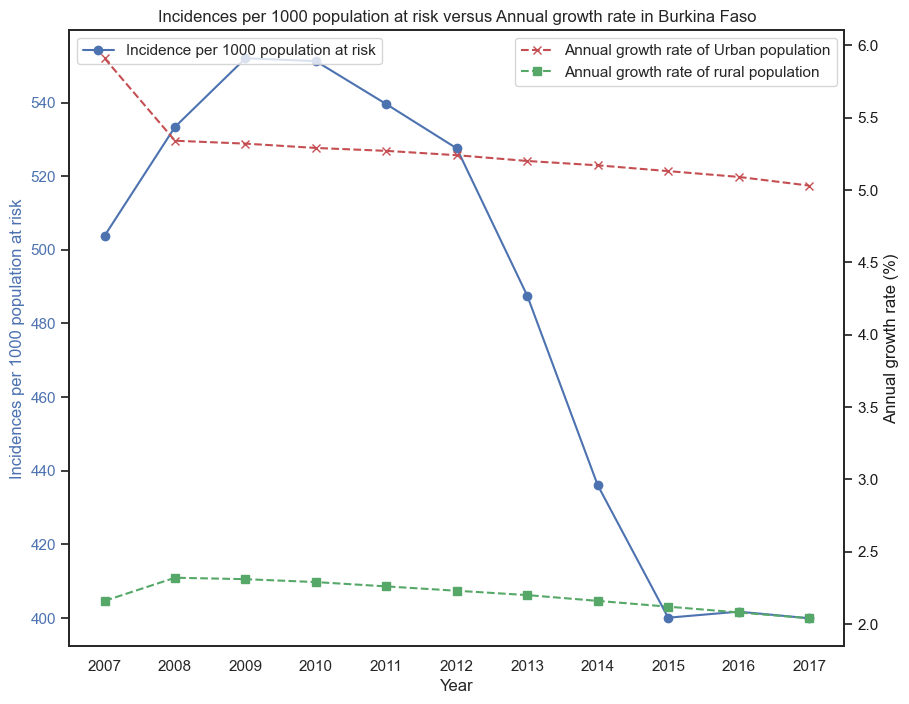
**Observation**:

As the growth rates reduce, the number of incidences per 1000 population at risk also reduces.

**Code Snippet**:



**Image**:



1. **Line graph showing the trend in the percentage of people with access to basic drinking water and sanitation services versus the trend of the number of Incidences per 1000 population at risk.**

**Objective**:

To investigate the impact of access to basic drinking water and sanitation services on the number of incidences per 1000 population at risk in the country with the highest number of incidences.

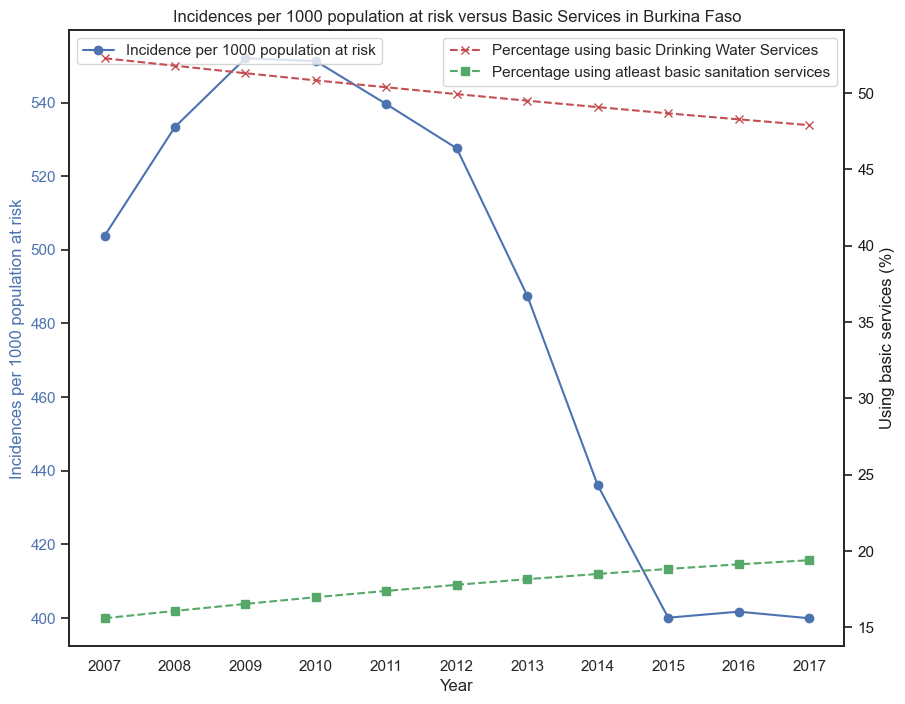
**Observation:**

As the percentage of the population using basic sanitation services increases, the number of incidences per 1000 population at risk reduces.

**Code Snippet**:

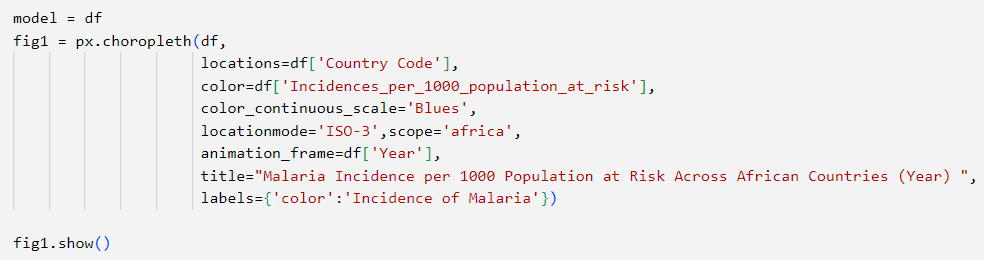


**Image:**

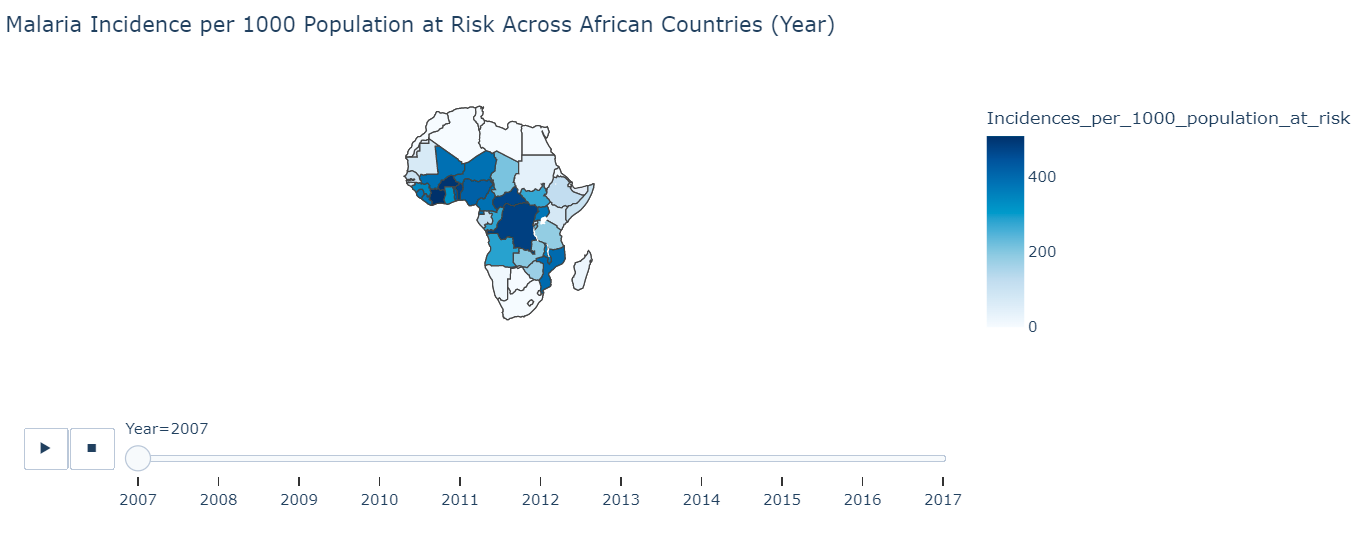
****

1. **Map of Africa showing a visual representation of the malaria burden across African countries per year.**

**Code Snippet:**

****

**Image:**

****

1. **Predictive Modelling**

In this section, we present the application of a multivariate linear regression model to predict the number of Incidences per 1000 population at risk. This model is built based on factors from the exploratory data analysis that are believed to influence the dependent variable.

**Predictors and dependent variable**

The following predictors were selected to build the regression model.

* **annual\_%\_growth\_of\_urban\_population** : Annual percentage growth of urban population.
* **%\_using\_atleast\_basic\_drinking\_water\_services**: Percentage of population using at least basic drinking water services.
* **%\_using\_atleast\_basic\_sanity\_services**: Percentage of population using at least basic sanitation services.
* **%\_of\_urban\_population**: Percentage of population living in urban areas.
* **Latitude**: Geographical latitude of the region.
* **Longitude**: Geographical longitude of the region.

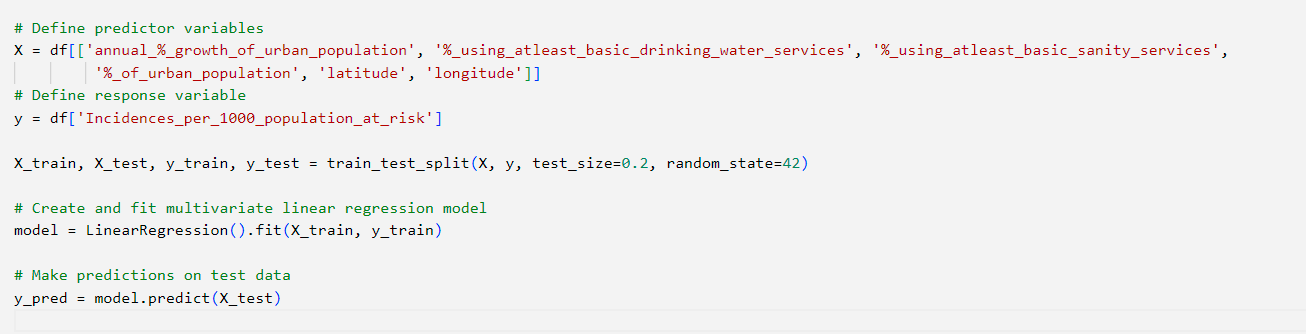
The response/dependent variable is:

* **Incidences\_per\_1000\_population\_at\_risk:** Number of incidences per 1000 population at risk.

**Data splitting, model creation and fitting.**

The data was split into training and test sets using a testing size of 20%. The random state was 42 for reproducibility.

**Code Snippet**:



**Model evaluation**

An r-squared value of 0.48 tells us that only 48% of the variability in the dependent variable is accounted for by the model. This also tells us that the model’s predictions are reasonable but not accurate and that there may be other factors or sources of variation that are not accounted for by the model.

**Code Snippet**:

