CURNEW MEDTECH INNOVATIONS PRIVATE LIMITED

SD03Q01

-Selva Vignesh M (1832045)

Problem Statement:

To explore whether the no. of cases and deaths of malaria increases every year.

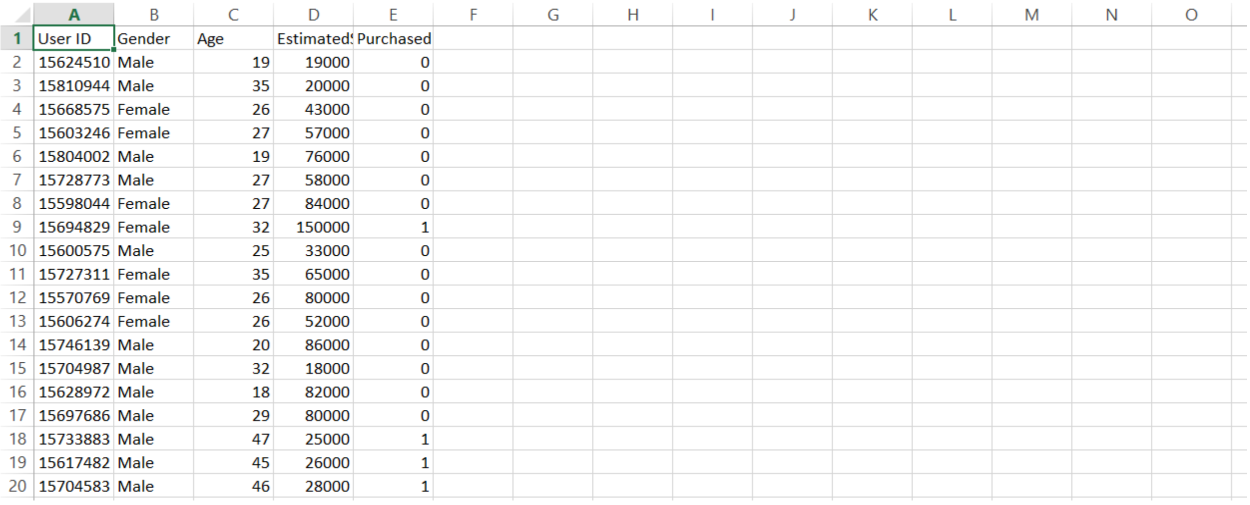
Abstract:

The given problem‘s motive is to find whether the cases and deaths of malaria increases each year. The tool used here is Python. Here our only way to solve the problem is by doing an Exploratory Data Analysis. For the given dataset, we first explore the data and analyse each and every variable, and by exploring more about the data we will come into a conclusion to the given problem statement

About the Dataset:

There are 3 datasets totally. That is estimated, reported and incidence cases. The common attributes among these datasets are No. of deaths and cases, WHO regions, countries and the year.

- Reported_numbers.csv - Reported no. of cases across the world
- Estimated_numbers.csv - Estimated no of cases across the world
- Incidenceper1000popat_risk.csv - Incidence per 1000 people at risk area



	A	B	C	D	E
1	User ID	Gender	Age	Estimated!	Purchased
2	15624510	Male	19	19000	0
3	15810944	Male	35	20000	0
4	15668575	Female	26	43000	0
5	15603246	Female	27	57000	0
6	15804002	Male	19	76000	0
7	15728773	Male	27	58000	0
8	15598044	Female	27	84000	0
9	15694829	Female	32	150000	1
10	15600575	Male	25	33000	0
11	15727311	Female	35	65000	0
12	15570769	Female	26	80000	0
13	15606274	Female	26	52000	0
14	15746139	Male	20	86000	0
15	15704987	Male	32	18000	0
16	15628972	Male	18	82000	0
17	15697686	Male	29	80000	0
18	15733883	Male	47	25000	1
19	15617482	Male	45	26000	1
20	15704583	Male	46	28000	1

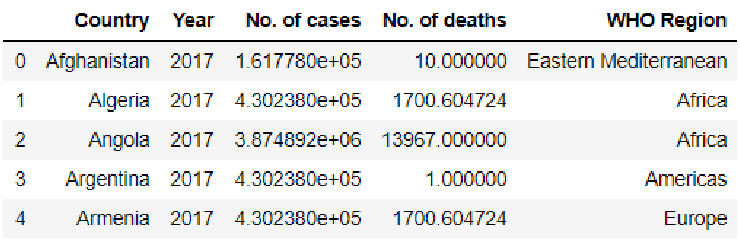
Exploratory Data Analysis:

In order to find whether the no. of deaths and cases increase every year we are using the reported cases dataset. We use different types of graphs and finally come into a conclusion.

Code:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import plotly.express as px
import plotly.io as pio
import plotly.graph_objects as go
from plotly.subplots import make_subplots
import warnings
warnings.filterwarnings("ignore")
%matplotlib inline
df = pd.read_csv(r"C:\Users\Selva Vignesh
M\Desktop\reported_numbers.csv", na_values=[None, 0.0])
df['No. of deaths'].fillna(value=df['No. of deaths'].mean(),
inplace=True)
df['No. of cases'].fillna(value=df['No. of cases'].mean(),
inplace=True)
df.head(5)
```

Output:



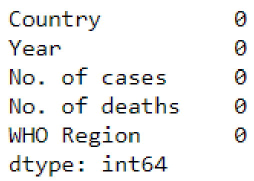
	Country	Year	No. of cases	No. of deaths	WHO Region
0	Afghanistan	2017	1.617780e+05	10.000000	Eastern Mediterranean
1	Algeria	2017	4.302380e+05	1700.604724	Africa
2	Angola	2017	3.874892e+06	13967.000000	Africa
3	Argentina	2017	4.302380e+05	1.000000	Americas
4	Armenia	2017	4.302380e+05	1700.604724	Europe

Here we have done some data pre-processing techniques that is data imputation. We have replaces Nan and null values with the column means and this is now our cleaned dataset.

Code:

```
df.isnull().sum()
```

Output:



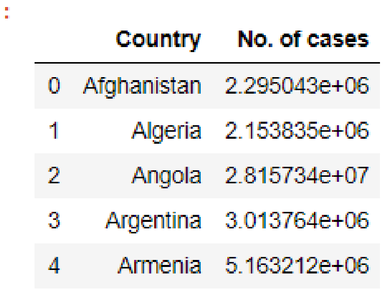
```
Country          0
Year             0
No. of cases     0
No. of deaths    0
WHO Region       0
dtype: int64
```

Now we check for null values in the dataset. More the null values lesser the accuracy. Here we can see that there are no null values as we earlier pre-processed the data.

Code:

```
df_group =df.groupby('Country')["No. of cases","No. of
deaths"].sum().reset_index()

df_group.head()

df_cases = df_group[["Country","No. of cases"]]

df_cases.head()
```

Output:



:

	Country	No. of cases
0	Afghanistan	2.295043e+06
1	Algeria	2.153835e+06
2	Angola	2.815734e+07
3	Argentina	3.013764e+06
4	Armenia	5.163212e+06

Here we have grouped the countries and the number of cases to understand better and to use them in the upcoming plots.

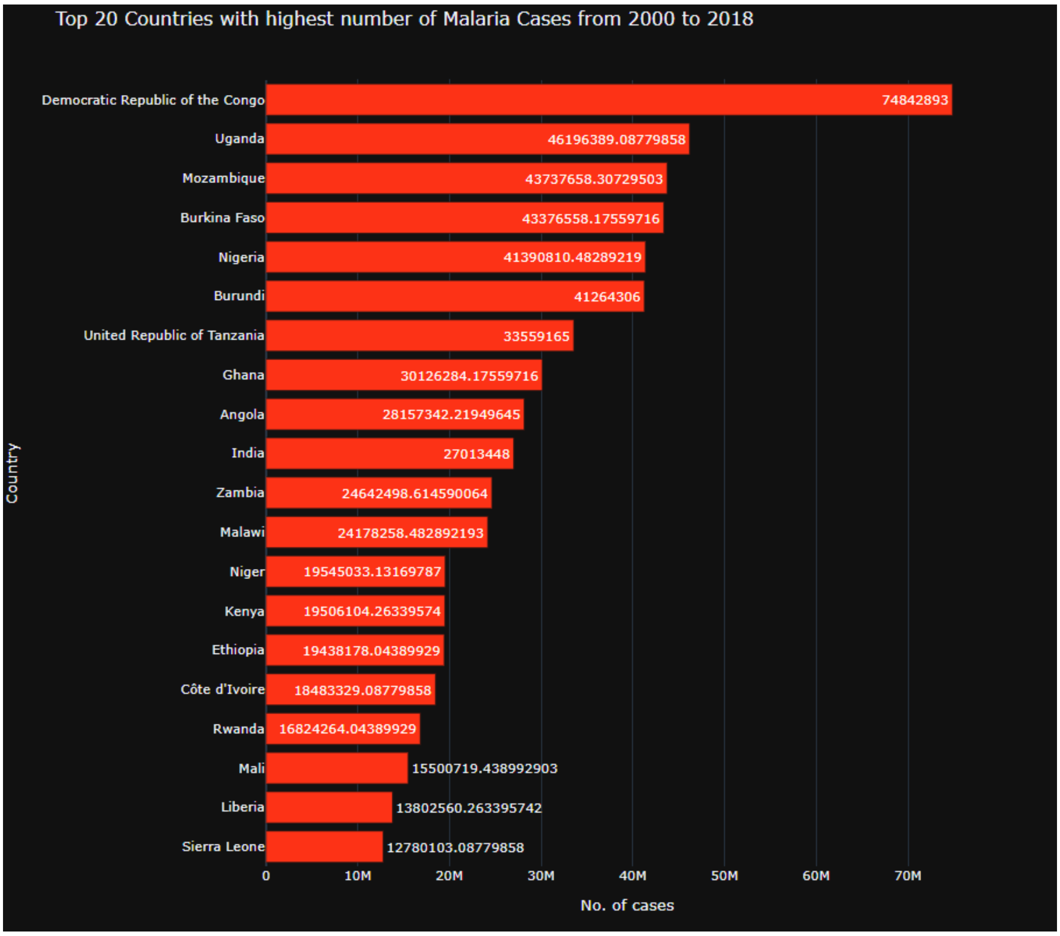
Top 20 Countries with highest number of Malaria Cases from 2000 to 2018:

Code:

```
pio.templates.default ='plotly_dark'
```

```
fig = px.bar(df_cases.sort_values("No. of
cases",ascending=False)[:20][::-1],x="No. of cases",y
="Country",text="No. of cases",

                 title="Top 20 Countries with highest number of Malaria
Cases from 2000 to 2018",

                 color_discrete_sequence=
px.colors.qualitative.Light24,height=900,orientation="h")#

fig.show()
```

Output:



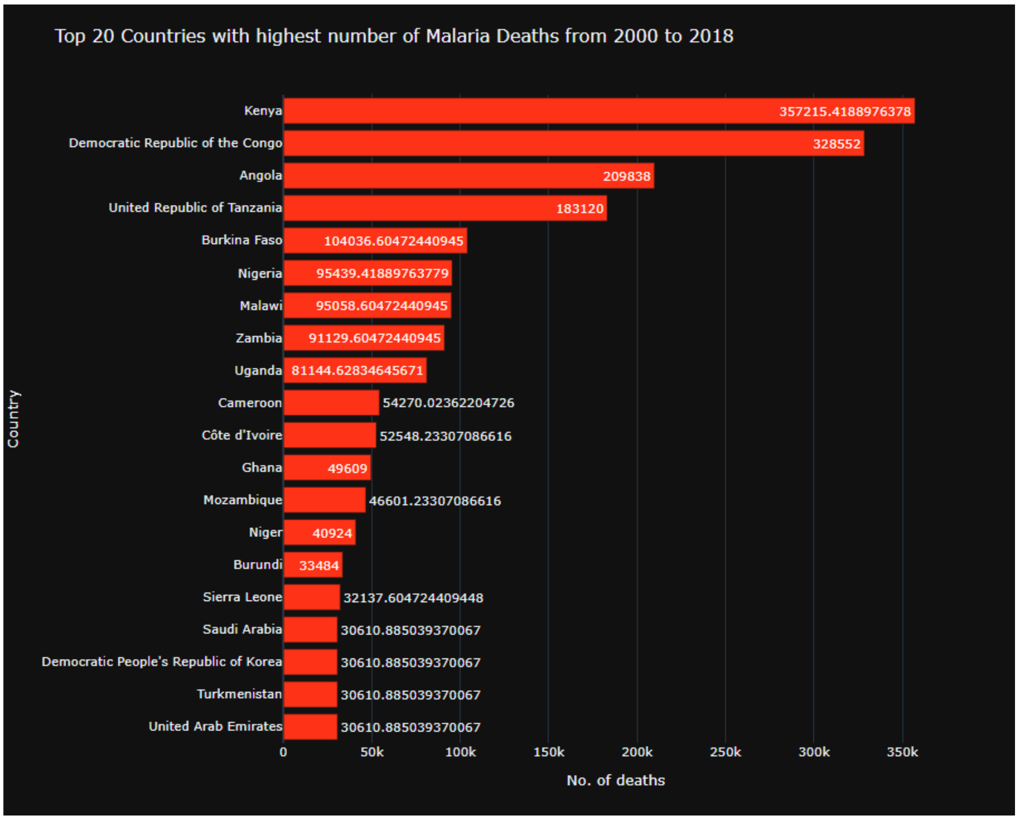
Here we have a plot that clearly shows the top 20 countries with highest number of cases from the period of 2008 to 2018. The Democratic Republic of Congo has the highest number of cases (74842893 cases).

Top 20 Countries with highest number of Malaria Deaths from 2000 to 2018:

Code:

```
df_death = df_group[["Country","No. of deaths"]]

pio.templates.default ='plotly_dark'

fig = px.bar(df_death.sort_values("No. of
deaths",ascending=False)[:20][::-1],x="No. of deaths",y
="Country",text="No. of deaths",

             title="Top 20 Countries with highest number of
Malaria Deaths from 2000 to 2018",

             color_discrete_sequence=
px.colors.qualitative.Light24,height=800,orientation="h")

fig.show()
```

Output:



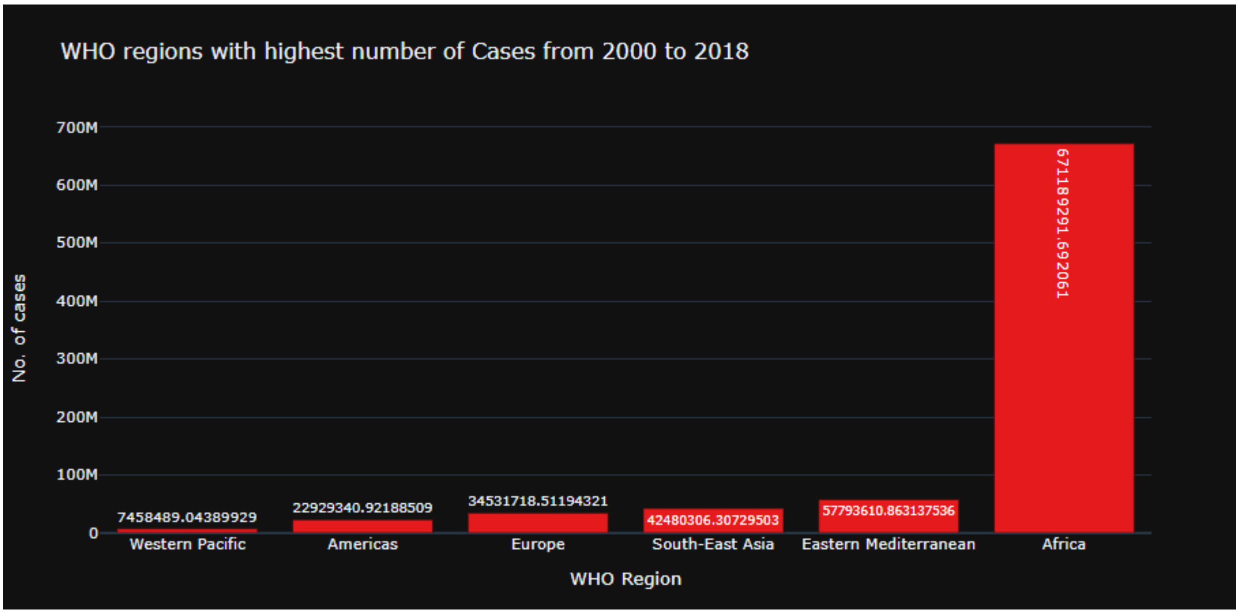
Here we have a plot that clearly shows the top 20 countries with highest number of deaths from the period of 2008 to 2018. Kenya has the highest number of deaths (357215 deaths).

WHO regions with highest number of cases from 2000 to 2018:

Code:

```
who_group =df.groupby('WHO Region')["No. of cases","No. of deaths"].sum().reset_index()

who_group.head().style.background_gradient(cmap ='Reds')

pio.templates.default = "plotly_dark"

fig = px.bar(who_group.sort_values("No. of cases",ascending=False)[::-1],y="No. of cases",x ="WHO Region",text="No. of cases",
             title="WHO regions with highest number of Cases from 2000 to 2018",
             color_discrete_sequence= px.colors.qualitative.Set1,height=500,orientation="v")

fig.show()
```

Output:



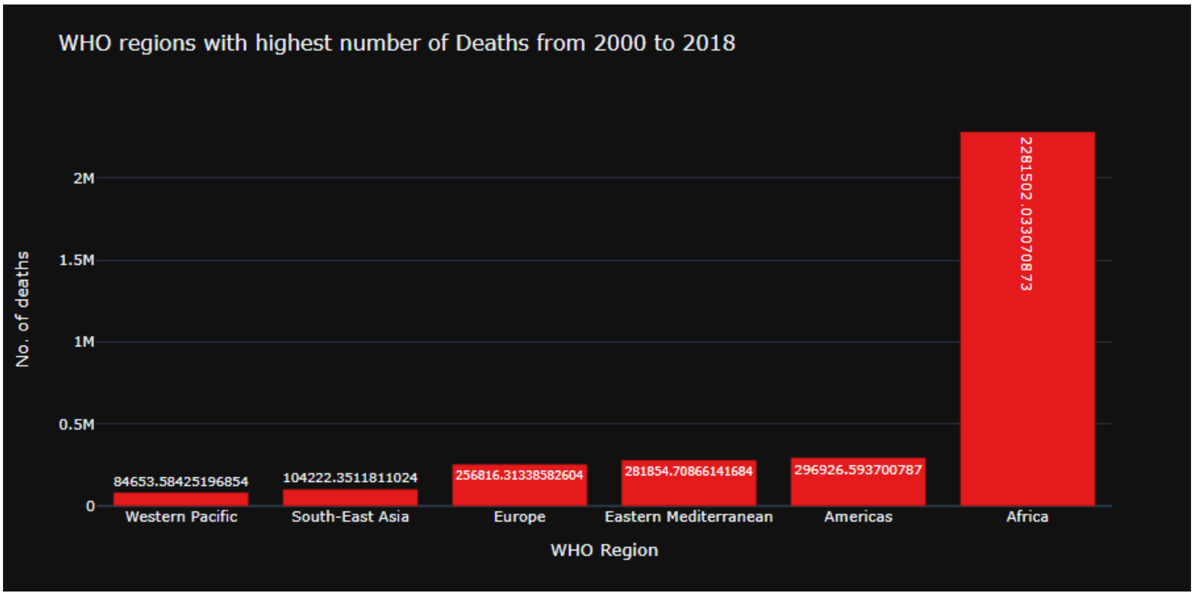
Here we have a plot that clearly shows the WHO regions with highest number of cases from the period of 2008 to 2018. Africa has the highest number of cases (total of 671189291 cases).

WHO regions with highest number of deaths:

Code:

```
pio.templates.default = "plotly_dark"

fig = px.bar(who_group.sort_values("No. of
deaths",ascending=False)[::-1],y="No. of deaths",x ="WHO
Region",text="No. of deaths",

             title="WHO regions with highest number of Deaths from
2000 to 2018",

             color_discrete_sequence=
px.colors.qualitative.Set1,height=500,orientation="v")

fig.show()
```

Output:

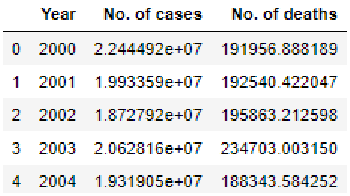


Here we have a plot that clearly shows the WHO regions with highest number of deaths from the period of 2008 to 2018. Africa has the highest number of deaths (total of 2281502 deaths).

Code:

```
year_group= df.groupby("Year")[["No. of cases","No. of
deaths"]].sum().reset_index()

year_group.head()
```

Output:



	Year	No. of cases	No. of deaths
0	2000	2.244492e+07	191956.888189
1	2001	1.993359e+07	192540.422047
2	2002	1.872792e+07	195863.212598
3	2003	2.062816e+07	234703.003150
4	2004	1.931905e+07	188343.584252

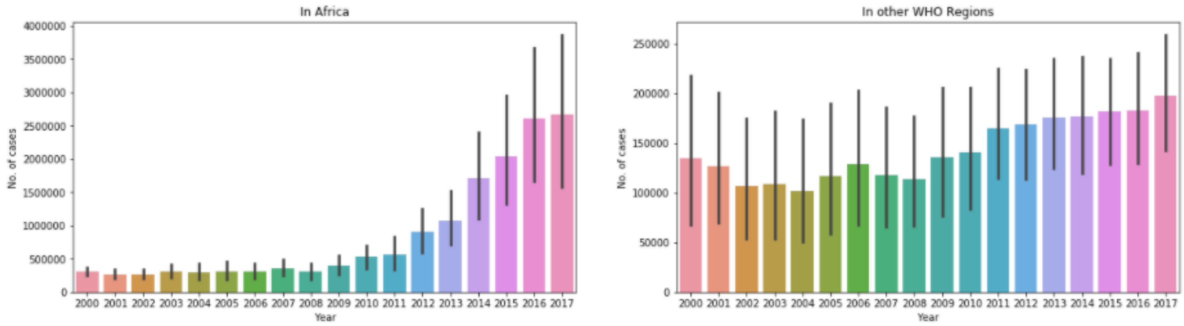
Here we have grouped the number of deaths and the number of cases with respect to year, to understand better and to use them in the upcoming plots.

Comparing Africa with other WHO regions:

Code:

```
fig_dims = (20, 5)

fig, axes = plt.subplots(1, 2, figsize=fig_dims)

sns.barplot(x = 'Year' , y = 'No. of cases' , data = df[df['WHO
Region'] == 'Africa'], ax= axes[0]).set_title("In Africa")

sns.barplot(x = 'Year' , y = 'No. of cases' , data = df[df['WHO
Region'] != 'Africa'], ax= axes[1]).set_title("In other WHO
Regions")
```

Output:

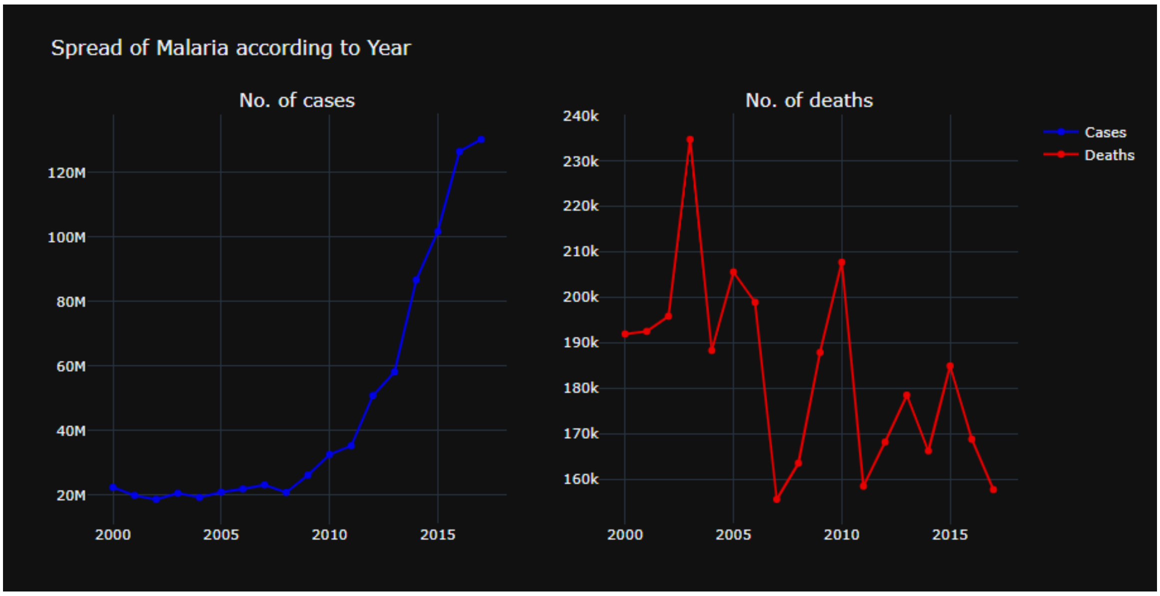


The first plot is the rise of cases in Africa versus the other WHO regions. We can clearly see that there is a gradual increase in the number of cases in Africa and the other regions have some deviations among them.

Spread of malaria according to the year:

```
fig = make_subplots(rows=1,cols=2,column_titles = ('No. of cases','No. of deaths'))

trace_1 = go.Scatter(x=year_group['Year'],y=year_group['No. of cases'],name='Cases',opacity=0.9,mode='lines+markers',line_color='blue')

trace_2 = go.Scatter(x=year_group['Year'],y=year_group['No. of deaths'],name='Deaths',opacity=0.9,mode='lines+markers',line_color='red')

fig.append_trace(trace_1,1,1)

fig.append_trace(trace_2,1,2)

fig.update_layout(title_text="Spread of Malaria according to Year")

fig.show()
```

Conclusion:



Finally we have achieved the plot that answers the problem statement. By analysing the earlier plots we have concluded them with this final plot. **Hence we can clearly see that the number of malaria cases increases every year and the number of deaths has it rise and fall for each year which depends upon the cases that increases every year.**