CURNEU MEDTECH INNOVATIONS PRIVATE LIMITED

-Nikinprasad V R (1832038)

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	F	G	H	I	J	K	L	M	N	O
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 seaborn as sns
import matplotlib.pyplot as plt
import warnings
warnings.filterwarnings('ignore')
```

Code:

```
rdata=pd.read_csv(r'D:\#PROJECTS\Curneu\SD03Q016\reported_numbers.csv')
edata=d.read_csv(r'D:\#PROJECTS\Curneu\SD03Q016\estimated_numbers.csv')
idata=pd.read_csv(r'D:\#PROJECTS\Curneu\SD03Q016\incidence_per_1000_pop_at_risk.csv')
Importing data.
```

Code:

```
print(rdata.head() , rdata.info())

print(edata.head() , edata.info())

print(idata.head() , idata.info())
```

Output:

```
Data columns (total 5 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   Country        1944 non-null   object
 1   Year           1944 non-null   int64
 2   No. of cases   1710 non-null   float64
 3   No. of deaths  1675 non-null   float64
 4   WHO Region     1944 non-null   object
dtypes: float64(2), int64(1), object(2)
```

```
Data columns (total 11 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   Country               856 non-null    object
 1   Year                  856 non-null    int64
 2   No. of cases          856 non-null    object
 3   No. of deaths         856 non-null    object
 4   No. of cases_median   856 non-null    int64
 5   No. of cases_min      544 non-null    float64
 6   No. of cases_max      544 non-null    float64
 7   No. of deaths_median  856 non-null    int64
 8   No. of deaths_min     524 non-null    float64
 9   No. of deaths_max     524 non-null    float64
 10  WHO Region            856 non-null    object
dtypes: float64(4), int64(3), object(4)
```

```
Data columns (total 4 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   Country       2033 non-null   object
 1   Year          2033 non-null   int64
 2   No. of cases  2033 non-null   float64
 3   WHO Region    2033 non-null   object
dtypes: float64(1), int64(1), object(2)
```

Checking the info and few data of reported data, estimated data, incidence data.

Code:

```
def getUnique(df):
    for col in df.columns:
        print(col + " : " + str(df[col].nunique()))
getUnique(rdata)
getUnique(edata)
getUnique(idata)
```

Output:

```
Country : 108
Year : 18
No. of cases : 1426
No. of deaths : 708
WHO Region : 6


Country : 107
Year : 8
No. of cases : 695
No. of deaths : 519
```

```
No. of cases_median : 694
No. of cases_min : 448
No. of cases_max : 481
No. of deaths_median : 447
No. of deaths_min : 255
No. of deaths_max : 336
WHO Region : 6


Country : 107
Year : 19
No. of cases : 1432
WHO Region : 6
```

Here using this function we have found the number of unique values in a column of a data frame.

Reported Data:

Code:

```
reported = rdata.groupby(['WHO Region']).agg({'No. of cases' :
'sum'}).reset_index()

x = rdata['No. of cases'].sum()

reported['Percentage'] = (reported['No. of cases']/x)*100

reported
```

Output:

	WHO Region	No. of cases	Percentage
0	Africa	586002159.0	87.930388
1	Americas	13894342.0	2.084864
2	Eastern Mediterranean	19932663.0	2.990922
3	Europe	112675.0	0.016907
4	South-East Asia	39468640.0	5.922321
5	Western Pacific	7028251.0	1.054598

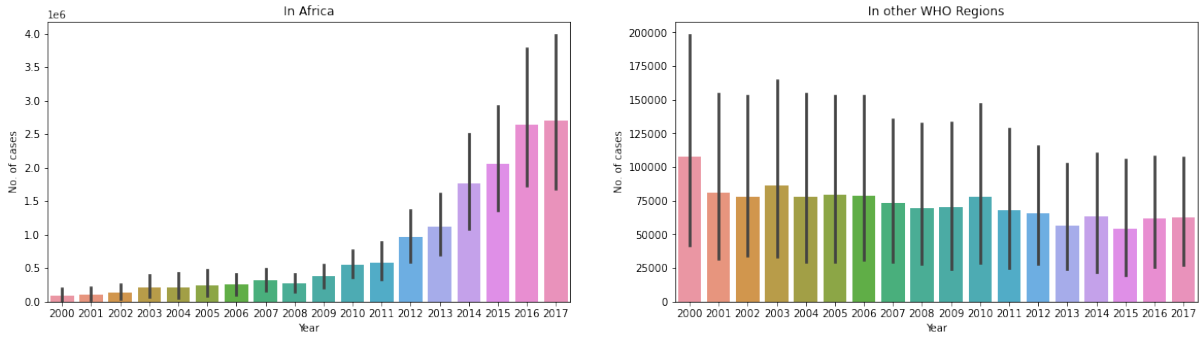
Here we can see that AFRICA constitutes 88% of the reported data. This means, if we include this, most of our data would be biased, so the analysis can be done in two groups, one including AFRICA, and another excluding AFRICA.

Code:

```
fig_dims = (20, 5)

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

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

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

Output:

```
Text(0.5, 1.0, 'In other WHO Regions')
```



Estimated Data:

Code:

```
estimated = edata.groupby(['WHO Region']).agg({'No. of
cases_median' : 'sum'}).reset_index()

x = edata['No. of cases_median'].sum()

estimated['Percentage']=(estimated['No. of
cases_median']/x)*100

estimated
```

Output:

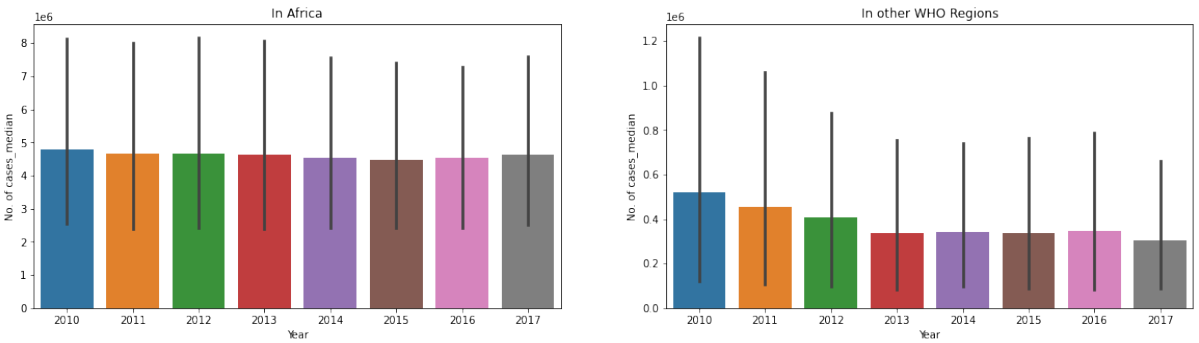
	WHO Region	No. of cases_median	Percentage
0	Africa	1587152540	89.056616
1	Americas	13025802	0.730890
2	Eastern Mediterranean	34456176	1.933368
3	Europe	262	0.000015
4	South-East Asia	132725734	7.447365
5	Western Pacific	14823234	0.831746

Here we can clearly see that, again, Africa holds 89% of the estimated data. Plotting two graphs.

Code:

```
fig_dims = (20, 5)

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

sns.barplot(x = 'Year' , y = 'No. of cases_median'

            , data = edata[edata['WHO Region'] == 'Africa'],
ax= axes[0]).set_title("In Africa")

sns.barplot(x = 'Year' , y = 'No. of cases_median'

            , data = edata[edata['WHO Region'] != 'Africa'],
ax= axes[1]).set_title("In other WHO Regions")
```

Output:



The estimated data shows a stangnant number of increase in cases, not much. Where as for the other regions, it is declining.

Incidence data:

Code:

```
incidence = idata.groupby(['WHO Region']).agg({'No. of cases' :
'sum'}).reset_index()

x = idata['No. of cases'].sum()

incidence['Percentage'] = (incidence['No. of cases']/x)*100

incidence
```

Output:

	WHO Region	No. of cases	Percentage
0	Africa	192303.47	86.632035
1	Americas	9666.15	4.354566
2	Eastern Mediterranean	4196.10	1.890328
3	Europe	107.84	0.048582
4	South-East Asia	2824.23	1.272306
5	Western Pacific	12879.53	5.802183

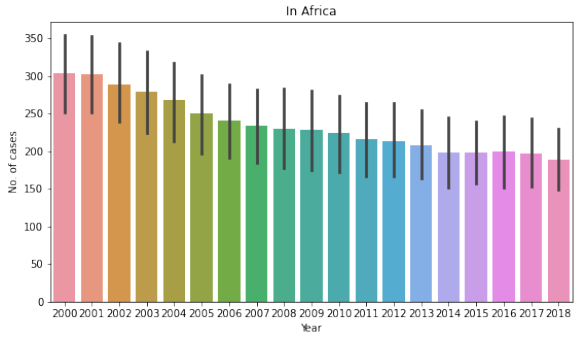
Here we see, Africa holds 87% data.

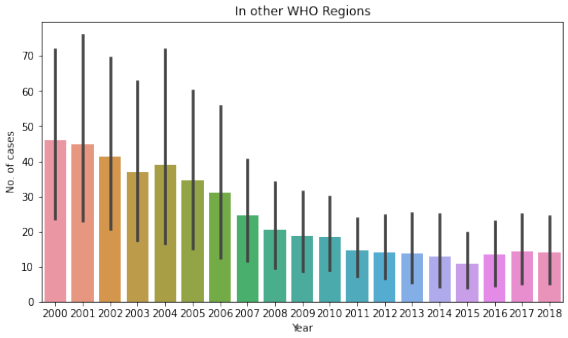
Code:

```
fig_dims = (20, 5)

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

sns.barplot(x = 'Year' , y = 'No. of cases'

            , data = idata[idata['WHO Region'] == 'Africa'], ax=
axes[0]).set_title("In Africa")

sns.barplot(x = 'Year' , y = 'No. of cases'

            , data = idata[idata['WHO Region'] != 'Africa'], ax=
axes[1]).set_title("In other WHO Regions")
```

Output:

Text(0.5, 1.0, 'In other WHO Regions')





Africa holds 87% data.

Observing the graph, the trends show that the numbers are declining for both. Its small yearly decrease for Africa region, whereas, for other regions, it's a good yearly decrease.

Conclusion:

Finally we have achieved the plot that answers the problem statement. Observing all the graphs, the following conclusions can be drawn. **For all the regions, except Africa, the cases are declining yearly. Africa shows a steady decrease, but not a huge decrease.**