CURNEU TASK - 1

MALARIAL DISEASE ANALYSIS

DATE : 6/3/2021 **NAME : REKHA V S**

1.INTRODUCTION

1.1 PROBLEM OVERVIEW :

Malaria *is more commonly seen in rural areas in India. In these areas, hygienic conditions are not good. One invariably finds places which do not have proper drains to drain water, and that causes stagnation of water. The ill-ventilated and ill-lit houses provide ideal indoor resting places for mosquitoes. Malaria is acquired in most instances by mosquito bites within the houses. Many of the people in rural areas sleep outdoors, not using mosquito nets makes them more prone to mosquito bites. The spraying of insecticides, if not undertaken on a regular basis, also favors the spread of malaria.*

1.2 TASK ANALYSIS

The objective of this task is to analyze and predict the number of deaths compared to the number of cases affected by people on yearly basis. And further analysis can also be done to find the relationship between cases and deaths in WHO countries so that further precautions or preventive measures can be undertaken in order to reduce the number of cases and to take some extra care and sanitation. Furthermore, processing and analysis with the data can be done by grouping the number of cases and death cases with the year so that we can have a keen check whether there is any decline in the death rate or how the trend behaves yearly based on the countries.

1.3 DATA SET – KNOWLEDGE DISCOVERY

For this task there were 3 sort of data sets which are :

1. **Incidence data**
2. **Reported numbers**
3. **Estimated numbers**

From my analysis, estimated numbers just resemble the **analysis done on the reported data**, and also it doesn't make any impact on the prediction part, the features in estimated data are much like in the range format which tells us that in that range the estimated value for our instance will rely on.

To my understanding of the data, estimated data is for the understanding of the reported data since the features in the particular dataset tells about minimum value, median, maximum deaths, minimum deaths, etc. **hence we can have this data for reference purpose.**

The main processing and prediction is to be done with the Reported and incidence data.

REPORTS DATA

	Country	Year	No. of cases	No. of deaths	WHO Region
0	Afghanistan	2017	161778.0	10.0	Eastern Mediterranean
1	Algeria	2017	0.0	0.0	Africa
2	Angola	2017	3874892.0	13967.0	Africa
3	Argentina	2017	0.0	1.0	Americas
4	Armenia	2017	0.0	NaN	Europe

INCIDENCE DATA

	Country	Year	No. of cases	WHO Region
0	Afghanistan	2018	29.01	Eastern Mediterranean
1	Algeria	2018	0.00	Africa
2	Angola	2018	228.91	Africa
3	Argentina	2018	0.00	Americas
4	Armenia	2018	0.00	Europe

From the above two data samples one thing we can find out is that **reports data** depicts like a **TRAINING DATA** and **incidence data** is much identical to **TEST DATA,** because in the reports data there is a feature called No. of deaths and in the incidence data there is no such feature and **the main aim is to predict the no. of deaths in the incidence data.**

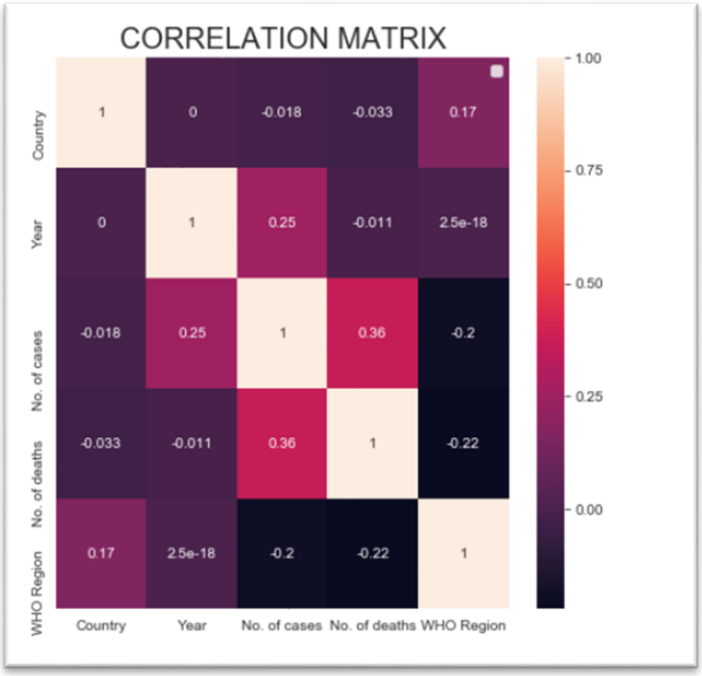
2.DATA ANALYSIS :

2.1 NULL VALUE ANALYSIS:

```
Country            0
Year               0
No. of cases     234
No. of deaths    269
WHO Region         0
dtype: int64
```

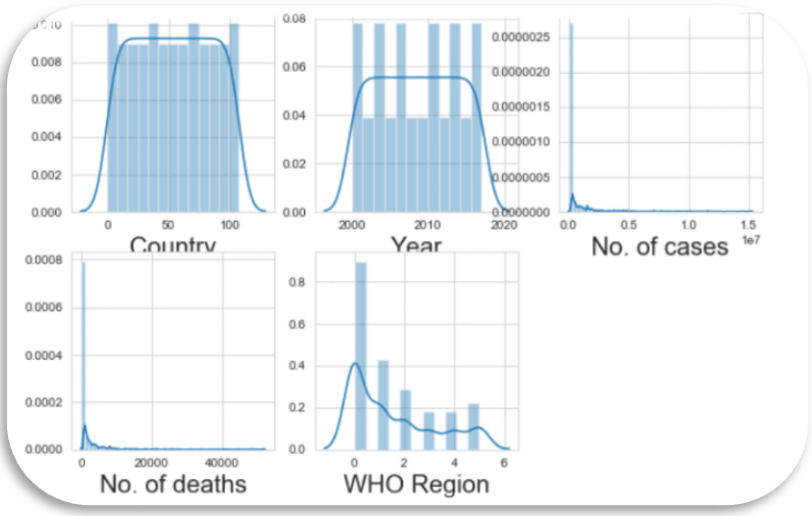
Here we can notice that No. of cases and No. of deaths have nearly 200 missing values, here we can apply hit a try on imputing with **Mean, Zero, KNN** imputation or if the 200 data doesn't make any impact on future prediction we can just **drop** it and which ever method yield a good accuracy can be chosen for further processing.

2.2 **CORRELATION ANALYSIS:**



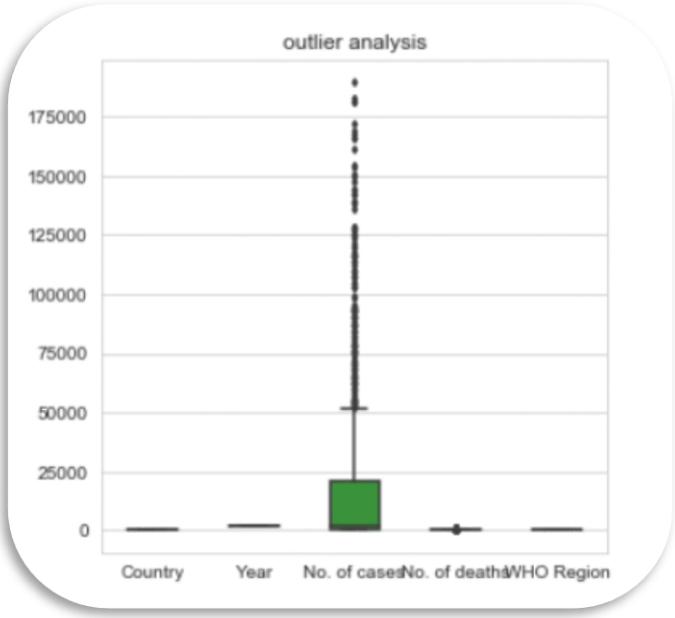
Here we can see that all the **features are independent** and there no relation between these features, hence we can include all the features for modeling, if suppose any features are highly correlated they must me removed because the presence of both features doesn't impact more in turn it leads **to very high computational power and it may also leads to over fitting of the model.**

2.3 DATA DISTRIBUTION:



We can here note that all the features follows normal distribution, except WHO region this is because in certain WHO region the impact of malarial disease and cases are more when compared with other WHO countries.

2.4 ANAMOLY DETECTION:

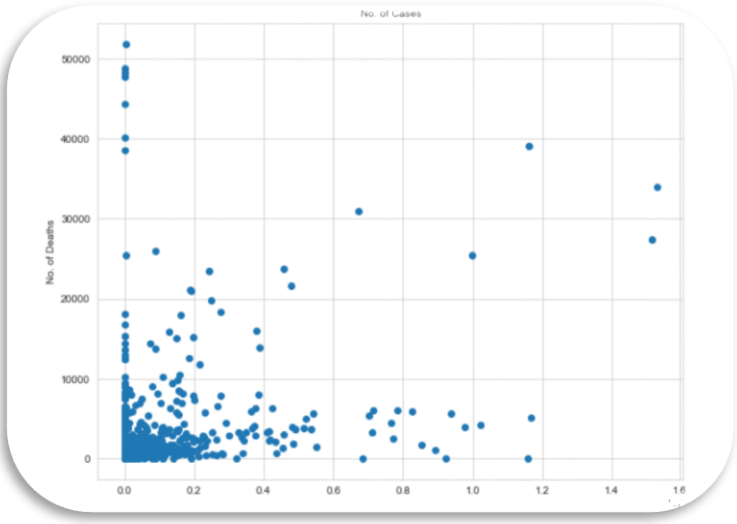


Here we can note that in No. of cases we have outlier and by using **DBSCAN** clustering it is found that about 300 instances are above 3rd standard deviation, hence these data should be removed by using variuos technique like **Zscore, IQR, DBSCAN** etc.

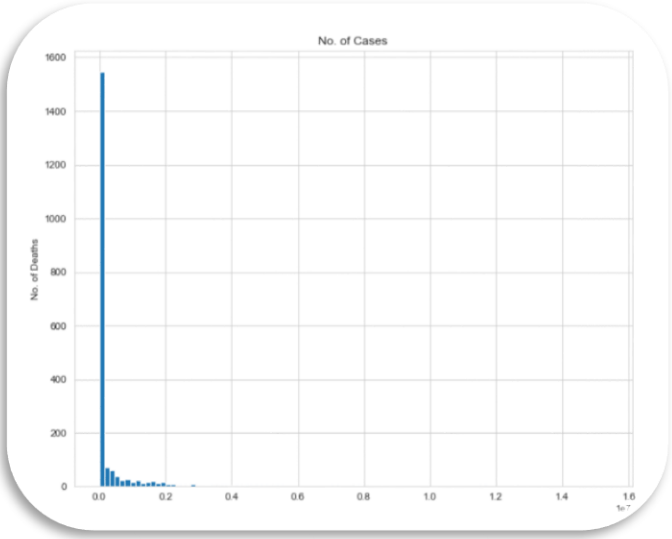
2.5 SCATTER PLOT VS HISTOGRAM VS PROBABILITY PLOT:

2.5.1 NUMBER OF CASES VS NUMBER OF DEATHS

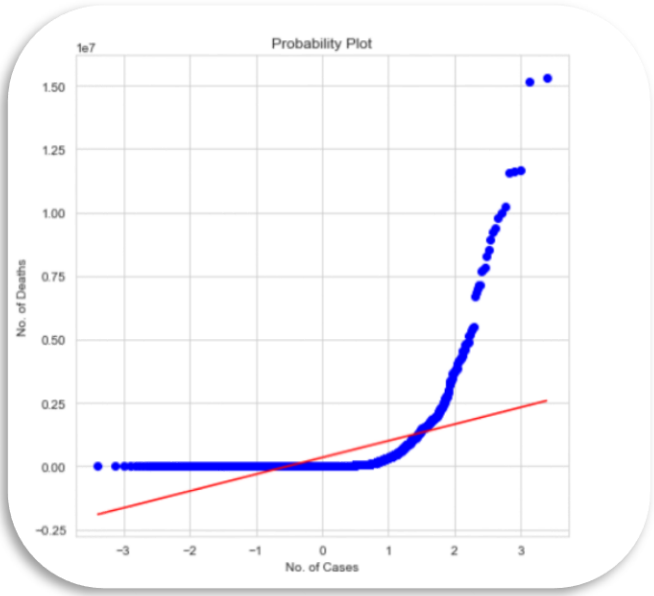
SCATTER PLOT



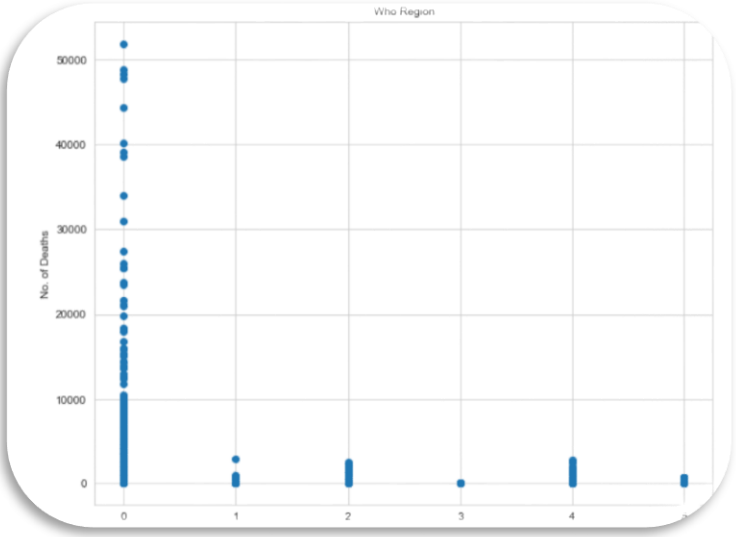
HISTOGRAM PLOT



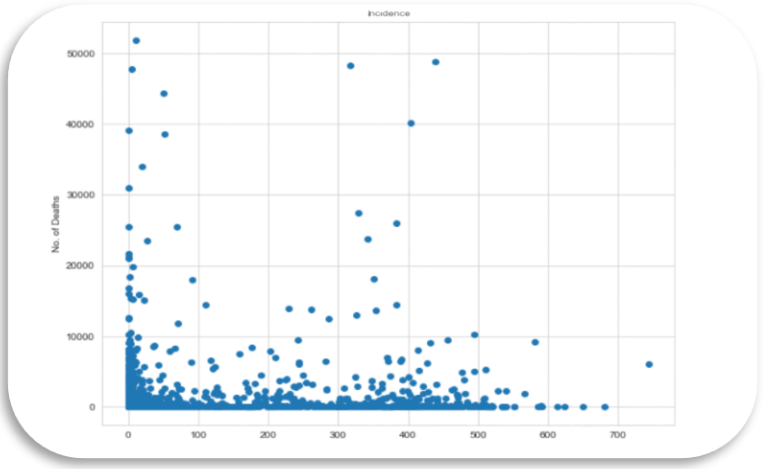
PROBABLITY PLOT



2.5.2 WHO REGION VS NO. OF CASES

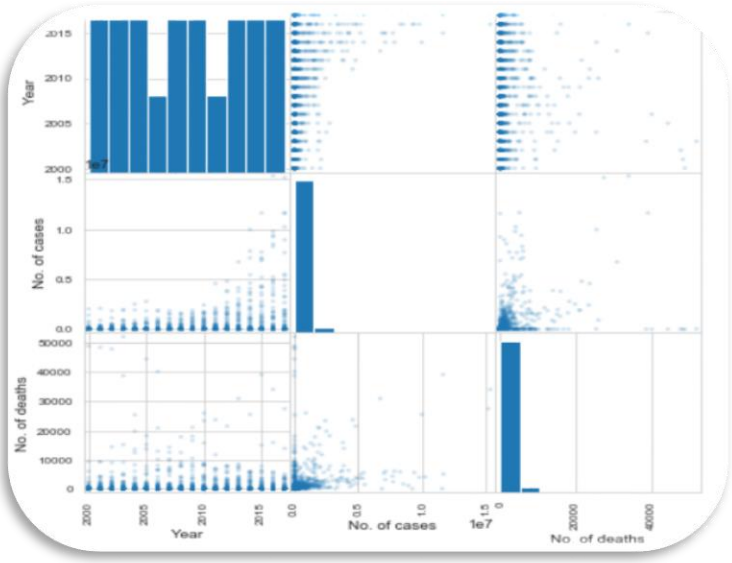


2.5.3 INCIDENCE VS NO. OF CASES

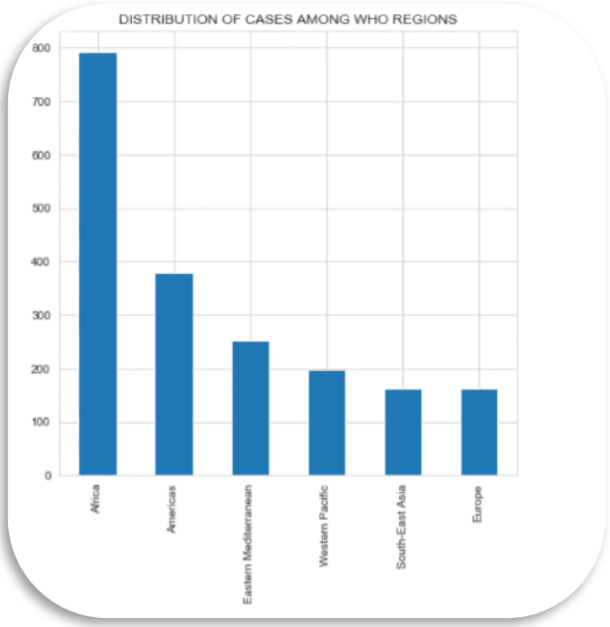


Here no. of cases is clusterd around incidence value between 0-500, so this can be taken as margin and all other scatter points can be bring down to this range.

2.6 SCATTER MATRIX:

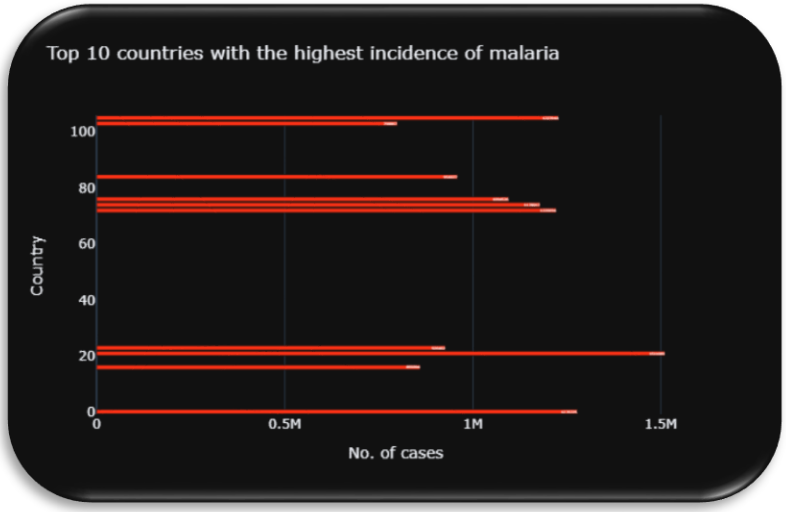


2.7 DISTRIBUTION OF CASES AMONG WHO REGIONS

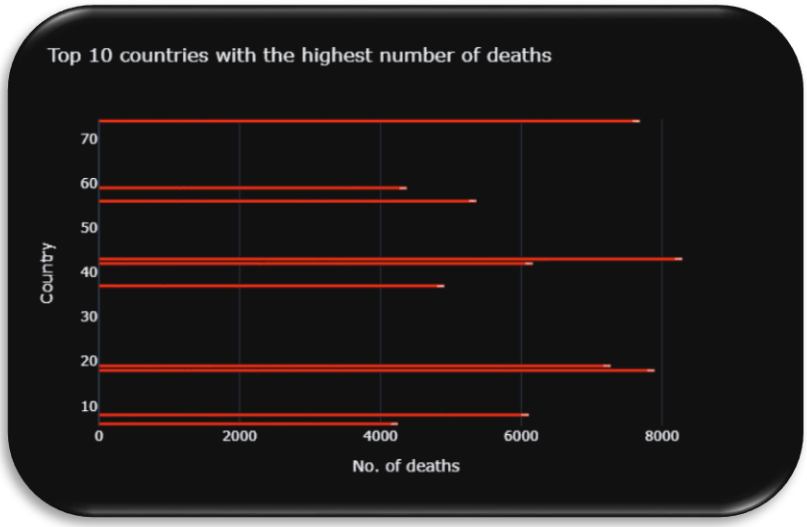


No. of cases in Africa is much higer when compared to all other WHO regions, it is also noted that No. of death is also higer in this region.

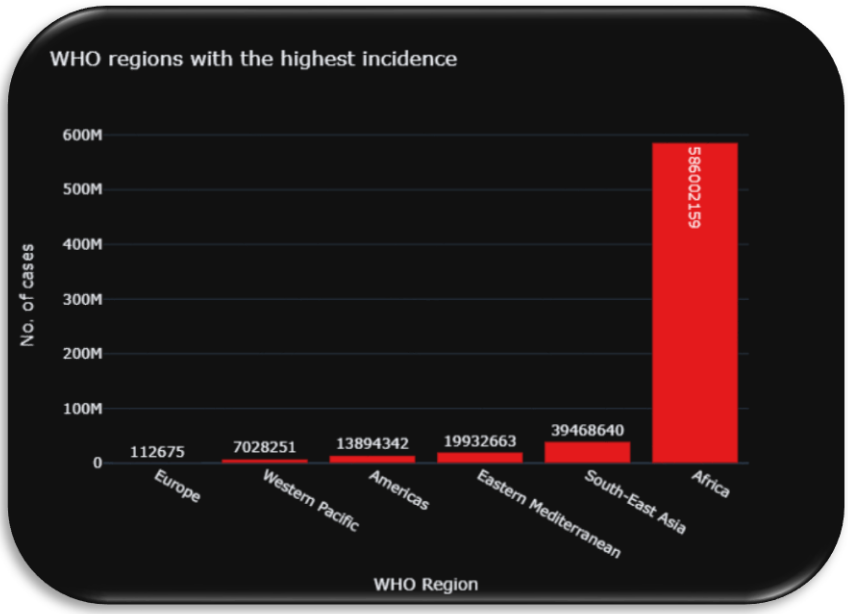
2.8 Top 10 countries with the highest incidence of malaria



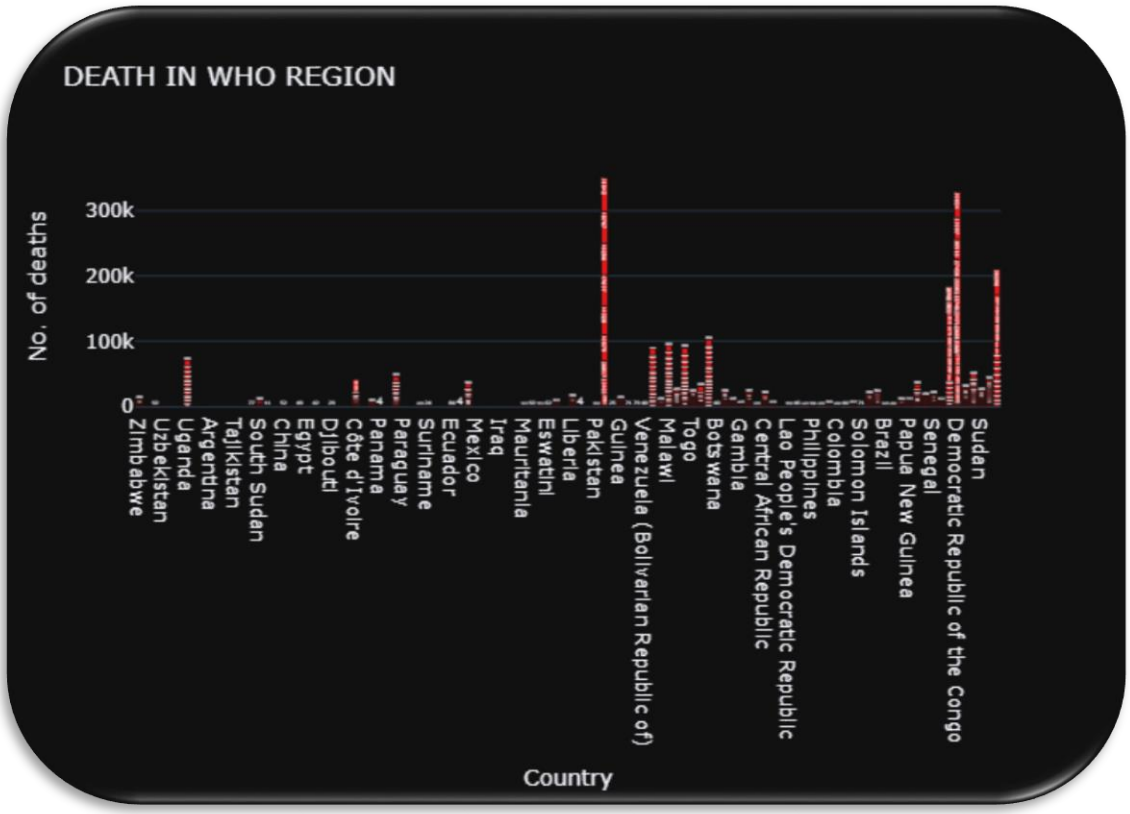
2.9 Top 10 countries with the highest death due to malaria



2.10 WHO REGION WITH HIGHEST INCIDENCE OF CASES



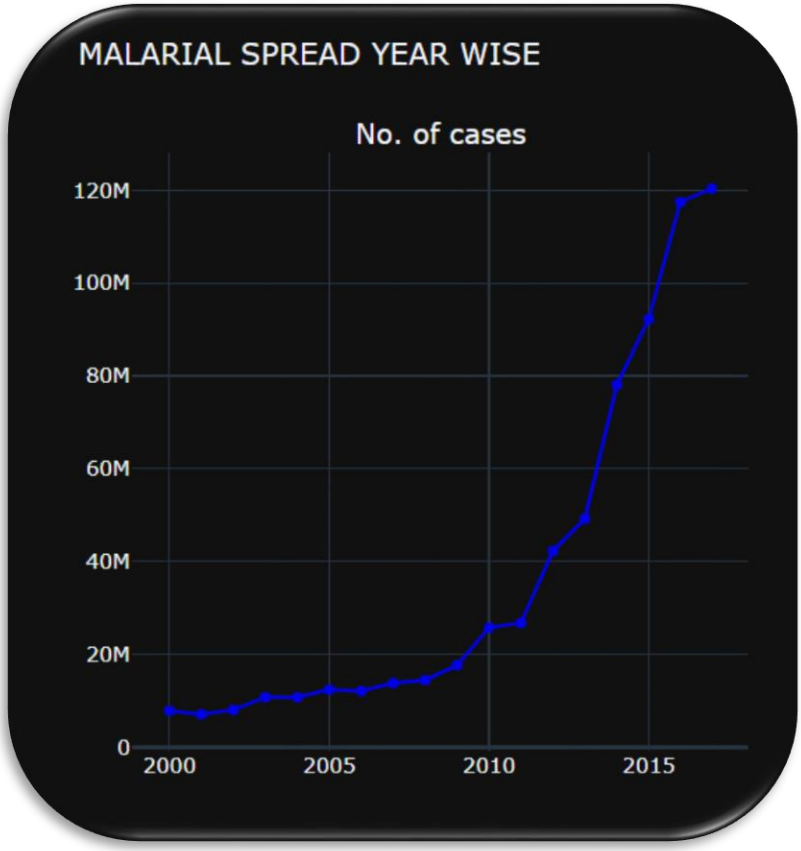
2.11 **DEATH IN WHO REGION**



2.12 GROUP OF YEAR WITH NO. OF CASES AND NO. OF DEATHS

	Year	No. of cases	No. of deaths
0	2000	7816830.0	88220.0
1	2001	7026451.0	112612.0
2	2002	7971965.0	119336.0
3	2003	10732686.0	161577.0
4	2004	10714285.0	122020.0

2.13 MALARIAL SPREAD YAER WISE WITH NO. OF CASES



2.14 MALARIAL SPREAD YEAR WISE WITH NO. OF DEATHS

3.FEATURE ENGINEERING

In the data given feature engineering technique can also be implemented but due to lack of size of data, as well features I am limiting this to imputation, anamoly detection and removal and scaling.

3.1 IMPUTATION:

In report's dataset, there were nearly 200 missing values in No. of cases and No. of death hence to impute those fields, first **mean imputation** is done but It doesn't make any impact then **KNN imputation** is implemented but this also doesn't contribute much, hence finally the missing value is implemented by filling those instances with **ZERO.**

3.2 DATA ENCODING:

In the malarial data, there are 2 fetures like **COUNTRY and WHO REGION which are the categorical data**, hence need to encode them, first **LABEL encoding** is applied on the data but the overall performance of the model declined to 20%, hence **DUMMY VARIABLE** encoding[**one hot]** is done.

3.3 OUTLIER TREATMENT:

The data above **3rd standard deviation** are considered as outliers, because they are much deviated from normal distribution of the data, by building model with outlier may result in **skewness of the data**, and to get a **robust model** there is a need to remove such outlier from the data.

In this task with the help of **DBSCAN** number of outlier present in the data is analyzed then

by using IQR techinque the data above **75th percentile** and data below **25th percentile** is removed.

3.4 SCALING THE DATA :

In analysis phase it is noticed that many features follow **gaussian distribution,**but features like No. of cases,No. of deaths are **right skewed** hence to remove skewness, here scaling is implemented on top of processed data.

4) MODEL BUILDING:

Since the task here is to predict **no. of deaths in incidence data** by learning from the relations in the report's data. Multiple regression model is fitted and from that most accurate and robust model is used for further prediction, here a **custom machine learning model** can also be implemented but here I haven’t done so due to time and resource constraints.

4.1 LINEAR REGRESSION:

A simple regression model is fitted on the processed data, which yields an r2 score of **60% which** is very much less to be used for further prediction.

OLS Regression Results

Dep. Variable:	2	R-squared:	0.579
Model:	OLS	Adj. R-squared:	0.554
Method:	Least Squares	F-statistic:	23.16
Date:	Mon, 08 Mar 2021	Prob (F-statistic):	1.87e-269
Time:	18:21:12	Log-Likelihood:	-1917.0
No. Observations:	1944	AIC:	4054.
Df Residuals:	1834	BIC:	4667.
Df Model:	109		
Covariance Type:	nonrobust		

4.2 DECISION TREE REGRESSOR

Since linear regression produces only 60% accuracy so tree based such as Decision tree model is fitted which also produces very poor r2 score like only 78%, hence tuning is done which also doesn't make any impact on the score.

```
1 r2_score(y_test, dt_pred)
```

```
0.798374375916277
```

4.3 RANDOM FOREST REGRESSOR

The previous 2 models doesn't provide better score hence, hence in this case ensemble model is preferred and model is built, which gave 85% r2 score which is much higher than the previous 2 models. after tuning the score has raised by 5% hence random forest

best fits and it can also be fine tuned using **grid search** by due to **high computation,** here tuning is done using **randomized search** CV.

```
1 r2_score(y_test, pr)
```

```
0.8740527116319945
```

```
1 tarinpred = rf.predict(X_train)
```

```
1 r2_score(y_train, tarinpred)
```

```
0.8939702763327111
```

4.4 XG BOOST:

Bagging produced 89% as r2 score hence to even more get production grade score of 90% and above boosting is implemented which produced 92% accuracy even without tuning, after tuning the accuracy of the model is 95%.

```
1 r2_score(y_test,pxgb )
```

```
before tuning 0.92458002398733932
```

```
1 r2_score(y_test, y_pred_xgb)
```

```
0.9458002398733932
```

CONCLUSION:

During Data Analysis it is found that No. of cases and No. of death increases, as year increases. And understanding of data took much time and at last found the real usable data, hence model is built using reported data and it is tested with incidence data. Clearly it is found that as year increases death and cases ratio also increases but one thing it is witnessed that No. of death is much lower when compared to cases, i.e. all the people who have got symptoms aren't died. and the WHO countries which has no. of cases high is found and for those countries proper sanitation can be done and preventive measures to be taken in order to reduce the number of cases.