DATA SCIENCE PROJECT REPORT

WORLD HEALTH ORGANIZATION SUICIDE STATISTICS

RollNo: K16-3986 (Gr-1)

Submitted To : DR. Zeeshan Ahmed

Introduction

Close to 800 000 people die due to suicide every year, which is one person every 40 seconds. Suicide is a global phenomenon and occurs throughout the lifespan. Effective and evidence-based interventions can be implemented at population, sub-population and individual levels to prevent suicide and suicide attempts. There are indications that for each adult who died by suicide.

Research Goal:

The project aims to identify which countries are more affected, or which of them has high rate of suicides, based on the vastness of the country. The Problem is to predict the number of suicides based on country's population, year , age and gender.

The data of Suicide statistics is collected from WHO, which includes the fields :

1) Country : Country name to which person belong

2) Year : year in which he/she committed suicide

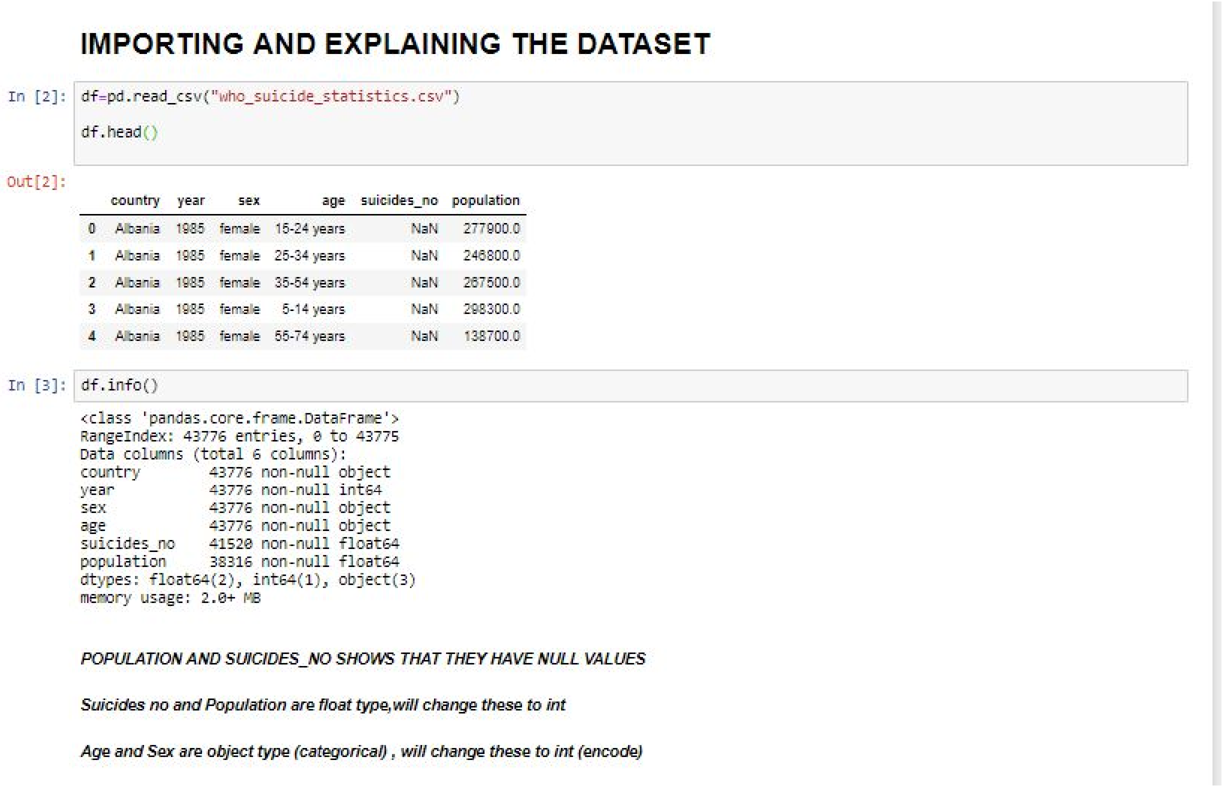
3) Sex : male /female

4) Age : age group to which the person belong

5) Suicides_no : number of suicides in year

6) Population : total population of country

With this data, we will go through the process of data science, showing results of each stage.



IMPORTING AND EXPLAINING THE DATASET

```
In [2]: df=pd.read_csv("who_suicide_statistics.csv")

        df.head()
```

Out[2]:

	country	year	sex	age	suicides_no	population
0	Albania	1985	female	15-24 years	NaN	277900.0
1	Albania	1985	female	25-34 years	NaN	246800.0
2	Albania	1985	female	35-54 years	NaN	267500.0
3	Albania	1985	female	5-14 years	NaN	298300.0
4	Albania	1985	female	55-74 years	NaN	138700.0

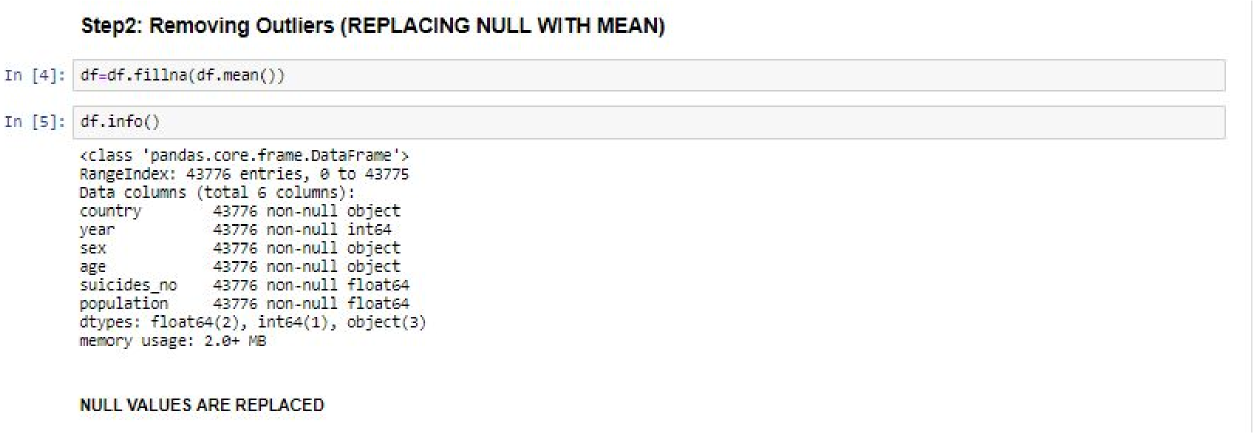
```
In [3]: df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 43776 entries, 0 to 43775
Data columns (total 6 columns):
country        43776 non-null object
year           43776 non-null int64
sex            43776 non-null object
age            43776 non-null object
suicides_no    41520 non-null float64
population     38316 non-null float64
dtypes: float64(2), int64(1), object(3)
memory usage: 2.0+ MB
```

POPULATION AND SUICIDES_NO SHOWS THAT THEY HAVE NULL VALUES

Suicides no and Population are float type,will change these to int

Age and Sex are object type (categorical) , will change these to int (encode)

This stage shows the Suicides_no and Population contains null values. Also, the datatype of population and suicides_no is float64, in order to apply machine learning model, we must change these to int. The country , age and sex data is in object from, we will encode these too.



Step2: Removing Outliers (REPLACING NULL WITH MEAN)

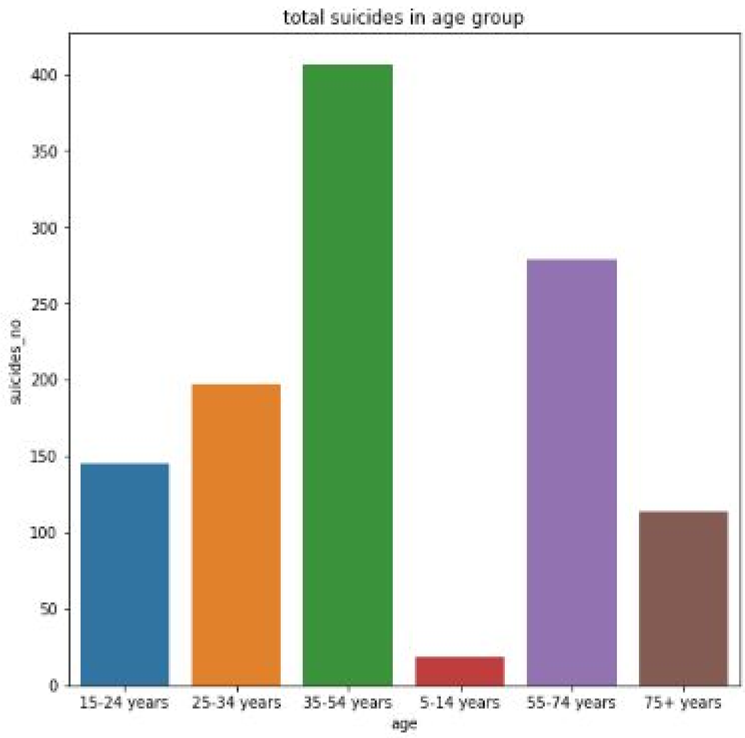
```
In [4]: df=df.fillna(df.mean())

In [5]: df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 43776 entries, 0 to 43775
Data columns (total 6 columns):
country        43776 non-null object
year           43776 non-null int64
sex            43776 non-null object
age            43776 non-null object
suicides_no    43776 non-null float64
population     43776 non-null float64
dtypes: float64(2), int64(1), object(3)
memory usage: 2.0+ MB
```

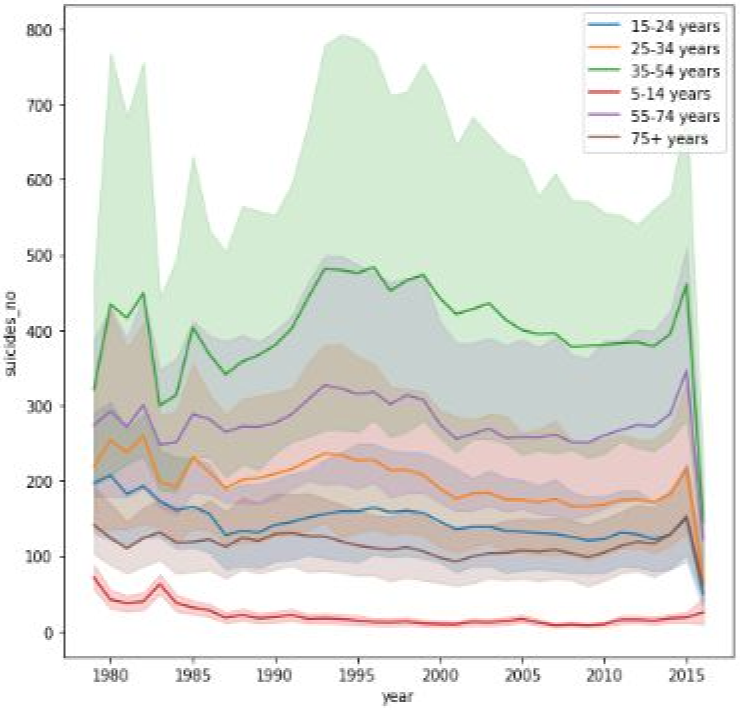
NULL VALUES ARE REPLACED

We can clearly see the non-null count, all contain same number of non-null rows now.

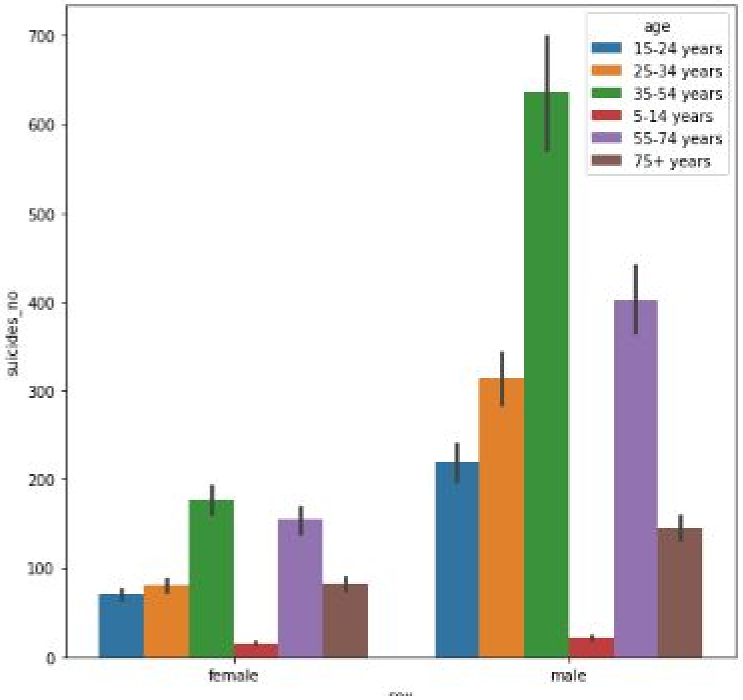
STAGE 2: REMOVING OUTLIERS AND EDA

In EDA, we first identified the total number of suicides committed in each age group



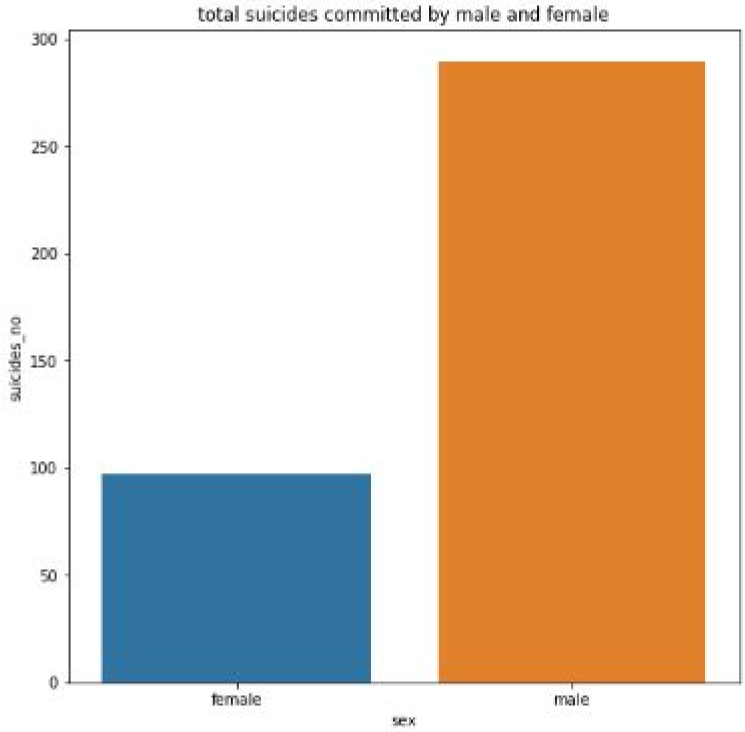


Here,the modal age group is 35-55 years old. To further elaborate, the suicides committed by each age group separated by genders is plotted.



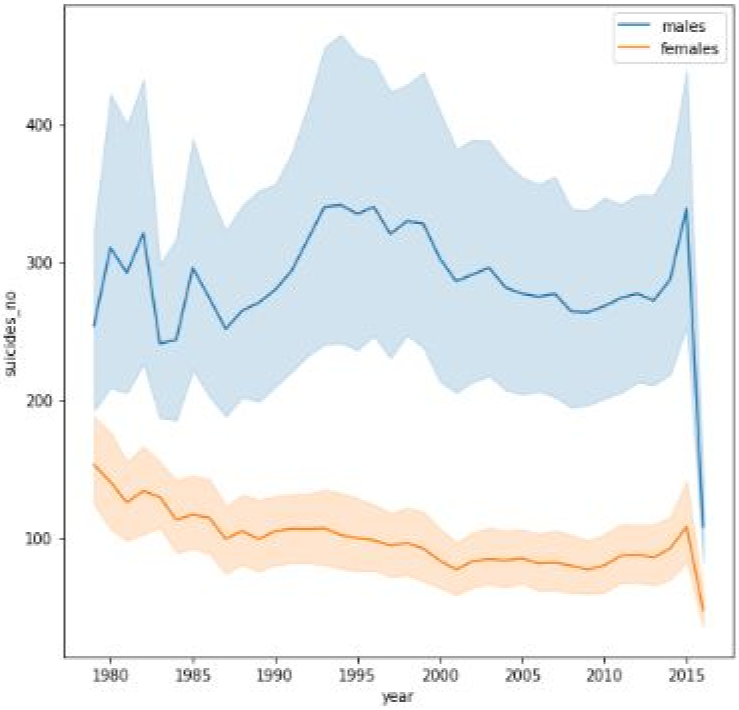
In both groups i-e. Male and Female, the modal classes is the age group 35-54 years. Also, there is an evident age effect that as the age increases , suicide rates increase until 55 years.

While the total number of suicides committed by each group is as follows:



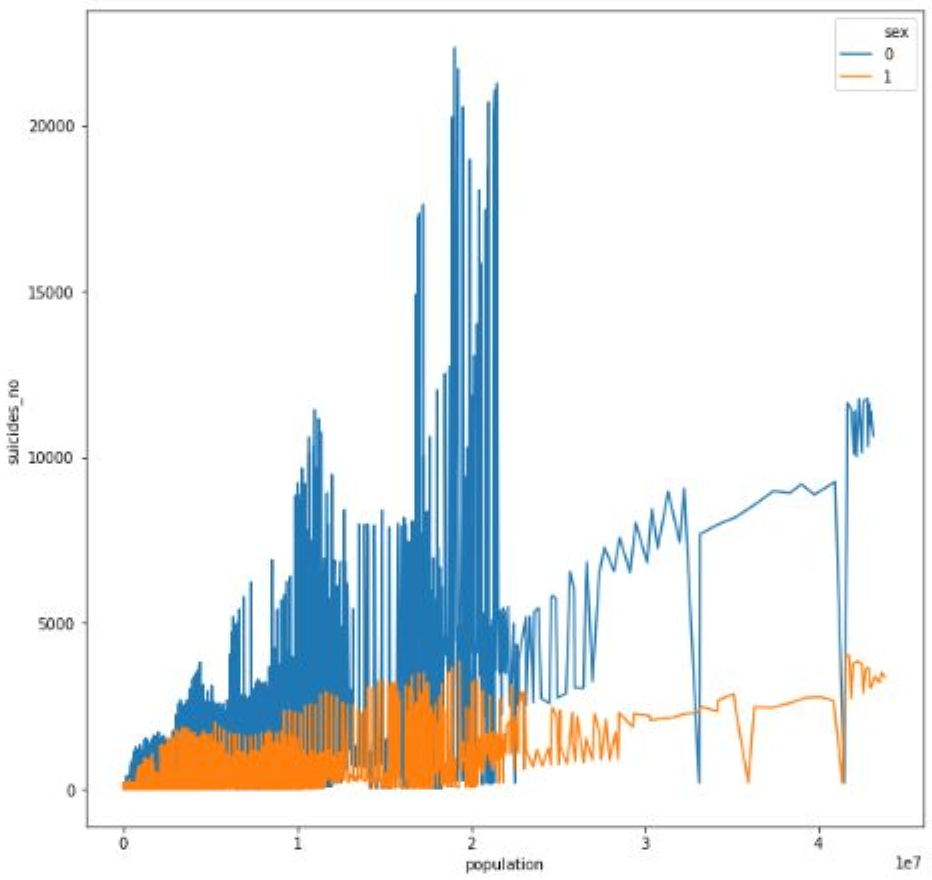
Here, Males committed twice as much of suicides than the females.

Now we check the trend of suicide over the years. Line graph gives a good illustration of it.



The trend have some fluctuation with the peak number of suicides in year 1995. For Female , there is a decreasing trend, and both male and female hits the peak in 2014 and then there is a drastic decrease in the rate. The reason for this decrease might be the economy of countries have stabilized .

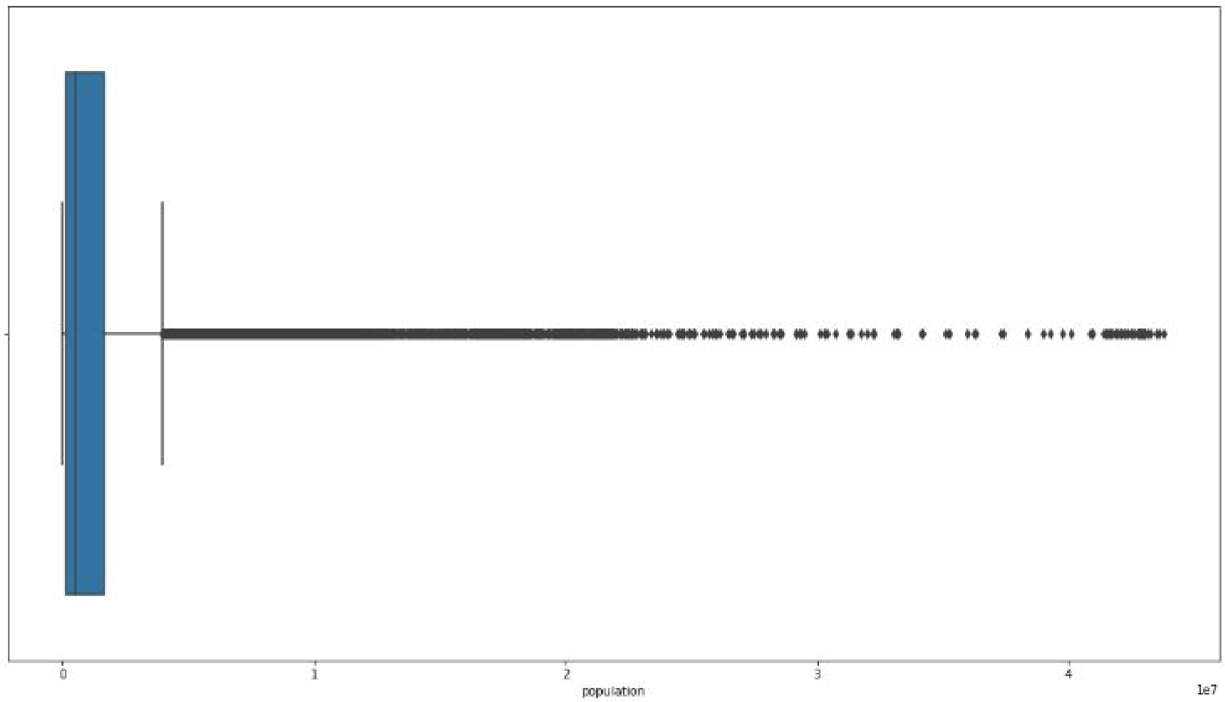
We then compare the suicide rate with population of a country.



Here the blue line represents male, and orange line represent female. And with the increase in population, there in an overall increasing trend, while at higher population, there are some sharp decreases which might be because the country is more urbanized, or the GDP per capita of the country might be greater.

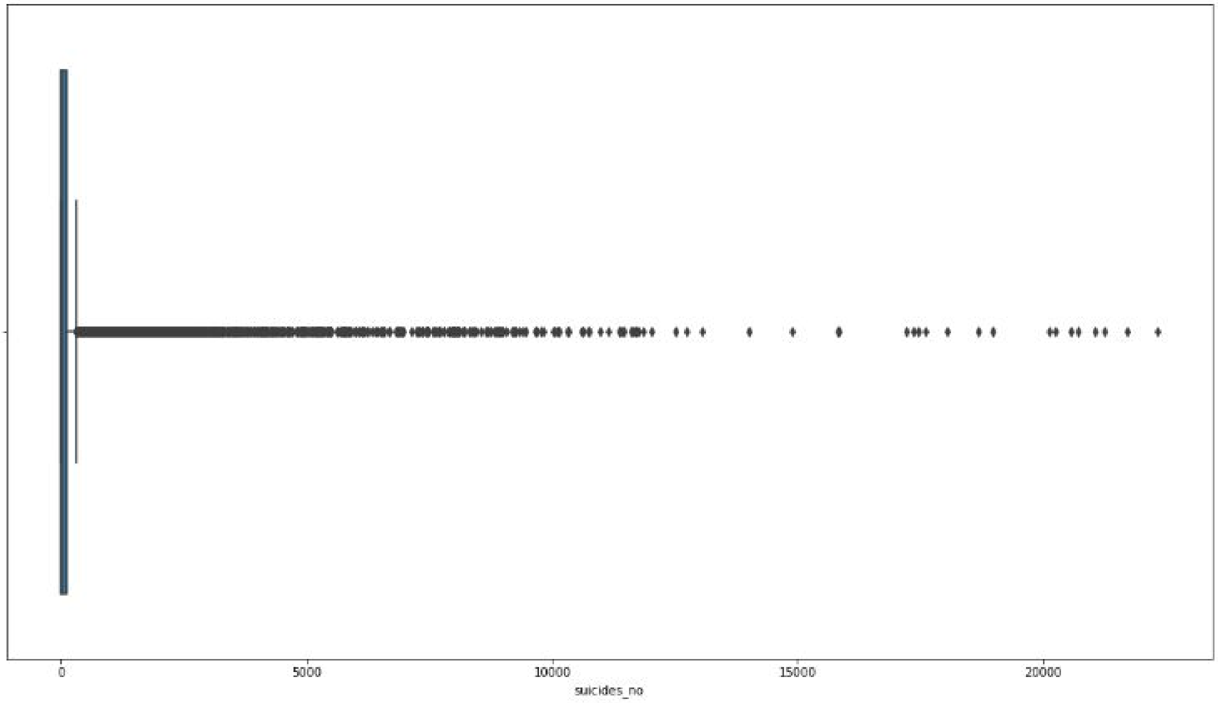
To check for outliers

1) Outliers in population



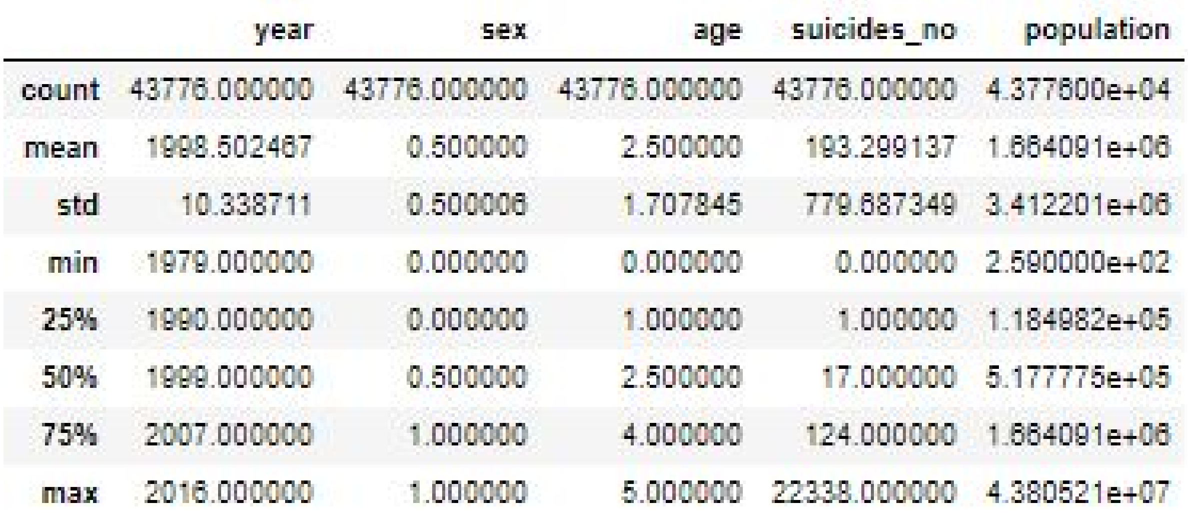
Note that the Population ran over 40,000,000.

2) Outliers in Suicide_no



Note that the suicide number ran over 20,000

Description of Columns:

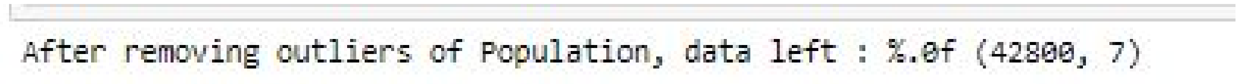


	year	sex	age	suicides_no	population
count	43776.000000	43776.000000	43776.000000	43776.000000	4.377600e+04
mean	1998.502467	0.500000	2.500000	193.299137	1.664091e+06
std	10.338711	0.500006	1.707845	779.687349	3.412201e+06
min	1979.000000	0.000000	0.000000	0.000000	2.590000e+02
25%	1990.000000	0.000000	1.000000	1.000000	1.184982e+05
50%	1999.000000	0.500000	2.500000	17.000000	5.177775e+05
75%	2007.000000	1.000000	4.000000	124.000000	1.664091e+06
max	2016.000000	1.000000	5.000000	22338.000000	4.380521e+07

Here the suicide have upper quartile of 124 while maximum number of suicide attempted 22338. For population, upper quartile is almost 1,660,000 while the maximum population is 43,800,000. So we can exclude high population countries. The total number of observations are: 43776.

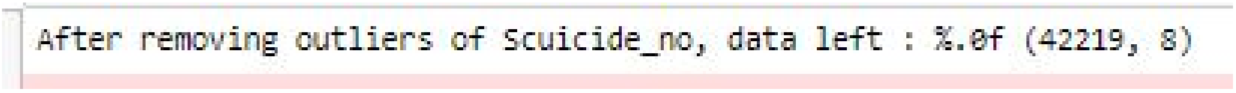
To remove these outlying data, we use Z-standardization.

Removing outlier of Population:



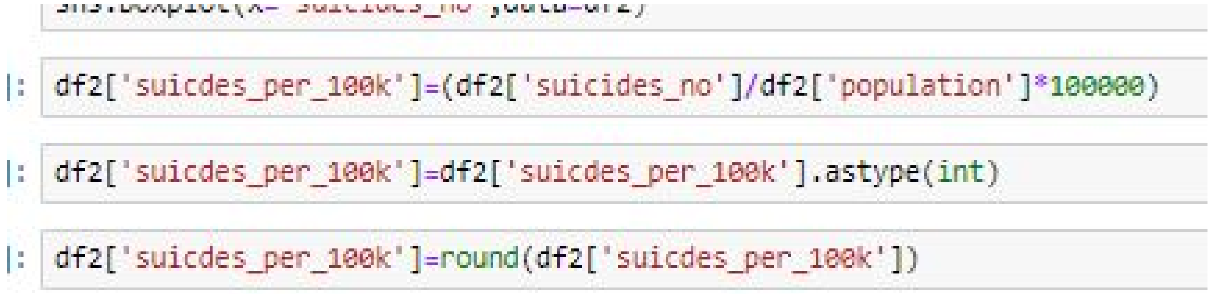
```
After removing outliers of Population, data left : %.0f (42800, 7)
```

Removing outlier of Suicide_no:



```
After removing outliers of Scuicide_no, data left : %.0f (42219, 8)
```

In further preprocessing, we added a column, SuicideRate per 100k of Population, and removed the outlier from it too.



```
sns.boxplot(x='suicides_no',data=df2)
```

```
]: df2['suicdes_per_100k']=(df2['suicides_no']/df2['population']*100000)
```

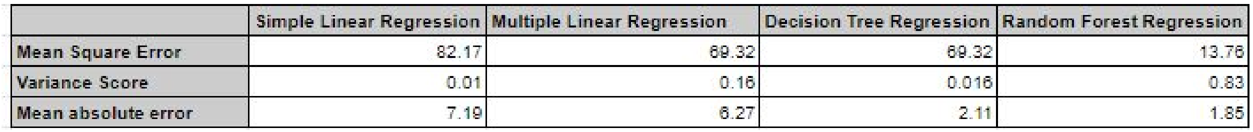
```
]: df2['suicdes_per_100k']=df2['suicdes_per_100k'].astype(int)
```

```
]: df2['suicdes_per_100k']=round(df2['suicdes_per_100k'])
```

Now our data is ready.

Stage 3: Machine Learning Model

We split the train set and test set in ratio 70%: 30%. The results of each model is summarized in the table below:



	Simple Linear Regression	Multiple Linear Regression	Decision Tree Regression	Random Forest Regression
Mean Square Error	82.17	69.32	69.32	13.76
Variance Score	0.01	0.16	0.016	0.83
Mean absolute error	7.19	6.27	2.11	1.85

And as expected, Random forest gives the best result with n_estimator: 50. It is clear from above result as the model consider more & more relevant attributes(in this case all are relevant) the accuracy increases.

Conclusion

We predicted the results from the data, where each attribute, country, year, age , gender and population , contributes significantly to the result. The rate of suicide per 100k of population is predicted for each country by using Random forest Regression. These results are helpful for the country to implement measure to prevent such self directed violence.