Date-20/8/2020

**Content flow:-**

1- Introduction

2- About the dataset

3- Advantage of the model

4- Problem Statement

5- Importing the required libraries

6- Importing the dataset and checking the dataset shape

7- Checking the data type

8- Checking & imputing missing values

9- EDA(Exploratory data analysis)

10- Feature Engineering and data preprocessing

11-Divideing training and testing set as we don’t want to scale the target variable.

12-skewness removal

13-Data Scaling

14-Model Building

15-Choosing and saving the final model

16-Testing the test data on the final saved model

17-Saving the test results in a csv format in local drive

18-Conclusion.

**An Article on global pandemic covid-19 with machine learning.**

**Introduction:-**

**Viral pandemics**are a serious threat to our world.COVID-19 is not the first, and it might not be the last

There have been many pandemics before in past and have caused a massive loss of human life.

Example- Spanishflu

The Spanish flu, also known as the 1918 flu pandemic, was an unusually deadly influenza pandemic caused by the H1N1 influenza A virus. Lasting from February 1918 to April 1920, it infected 500 million people–about a third of the world's population at the time–in four successive waves. [Wikipedia](https://en.wikipedia.org/wiki/Spanish_flu).

In 1918 there was no proper collection of data. These data about spanishflu are just estimated data. Hence we could not do much about the spread of the disease neither find a cure on time.

But now we are collecting the data more efficiently. Hundreds of research teams around the world are combining their efforts to the collect data and develop solutions which can be very helpful in future if such a pandemic situation arises again.

With the collected data, Machine learning is helping us to:

* Identify who is most at risk,
* Diagnose patients,
* Develop drugs faster,
* Finding existing drugs that can help
* Predict the spread of the disease,
* Understand viruses better,
* Map where viruses come etc…

Machine learning is playing a very vital role in this time of the pandemic and I am trying to bring in my knowledge of Machine learning to give some insight on the data collected by the researchers which might be useful for the future references.

**About the data:-**

The data is of 213 different countries for a span of 31 days(18/4/2020 to 18/5/2020).

There are total of 7(excluding date & country) different features described for each country on particular date. The features are:-

| **1->Date**  **2->Country** |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |

**3->Population**

**4->Total Tests**

**5->Total Cases**

**6->Total Deaths**

**7->Total Recovered**

**8->Serious or Critical**

**9->Active Cases**

**Problem Statement:-**

My goal is to do a predictive analysis on the dataset and predict the ‘Total Deaths’ in a country on a particular date.

**Advantage of the model:-**

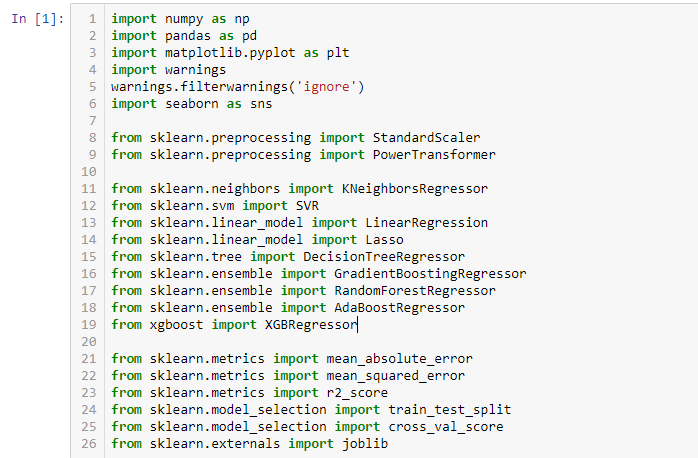
This predictive model can be used as a case study in a country where currently the no. of deaths are not so high .Hence preventive measures can be taken by country to stop the spread of the virus by comparing the preventive measures taken by the more infected countries as a benchmark.

Modification in preventive measures can be also decided taking the preventive measures of more infected countries as a benchmark.

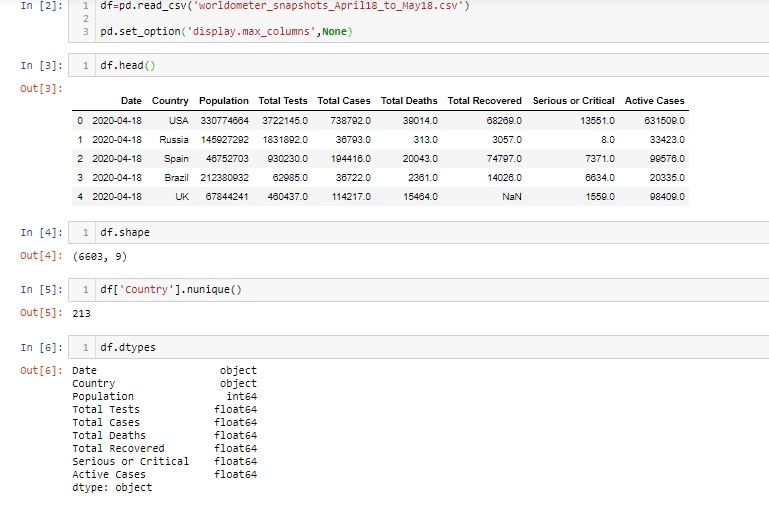
**Importing the required libraries:-**

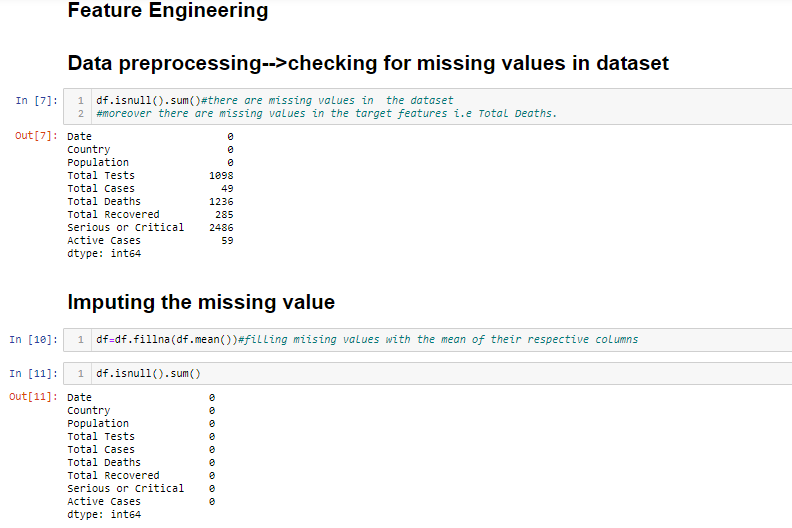
This is a regression problem where we will predict ‘Total Deaths’ from given features.

We will consider r2\_score as our evaluation metric and rmse score as error measurement .

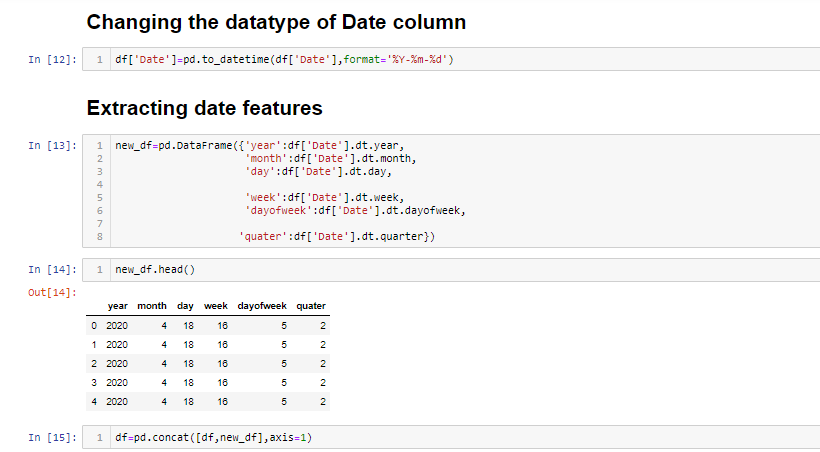


**Importing the dataset and checking the dataset shape and dtypes:-**

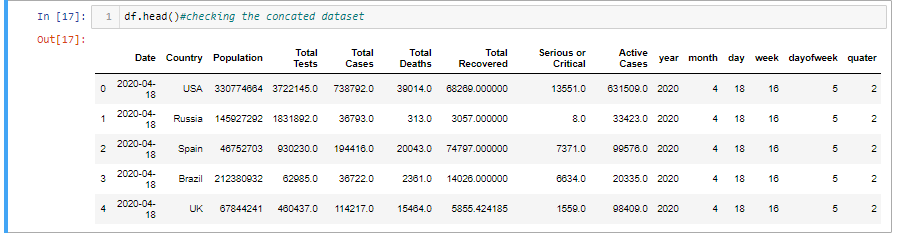
****

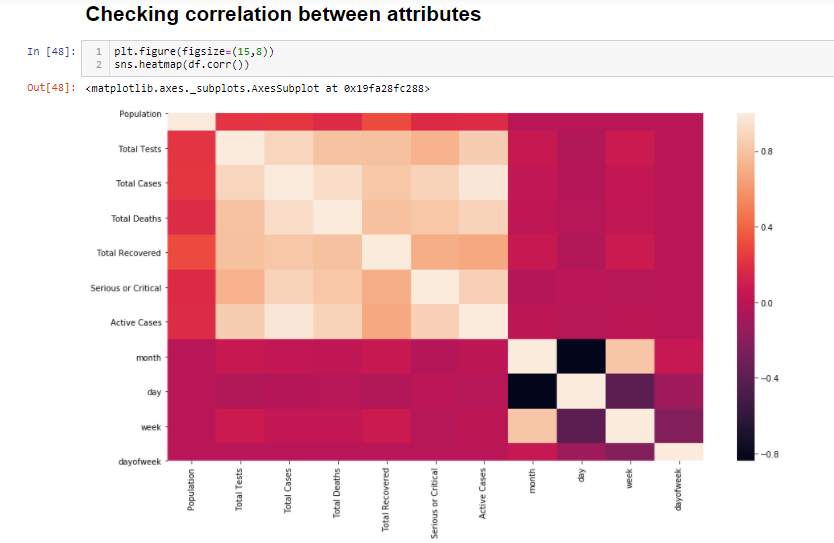


Data type of Date column is in object format which cannot be used to extract features, hence changing the format and data type of the data to carry out preprocessing of the date column.

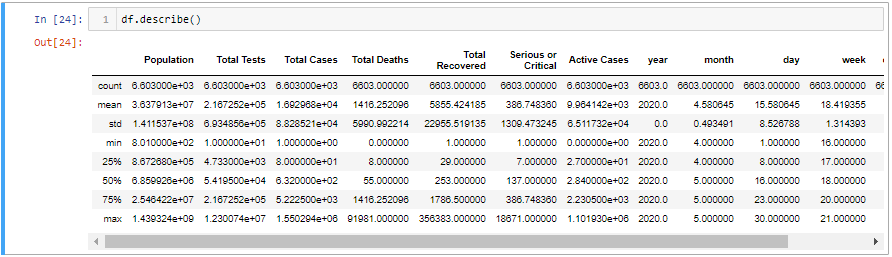


**Checking the dataset after concatenation:-**





Lighter the color shade higher is the correlation between the features as per the scale on the right.

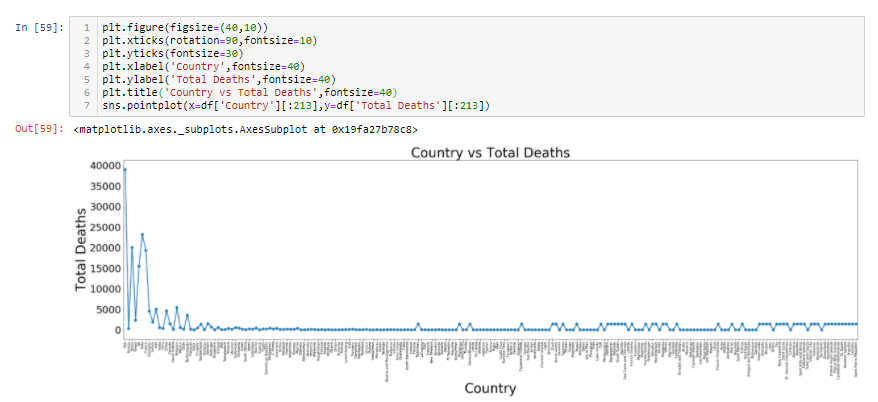


Outliers are present in some columns as difference between mean value and max value is very high.

**EDA(Exploratory data analysis)**

**Bivariate analysis(233 counties) on 18/4/2020**

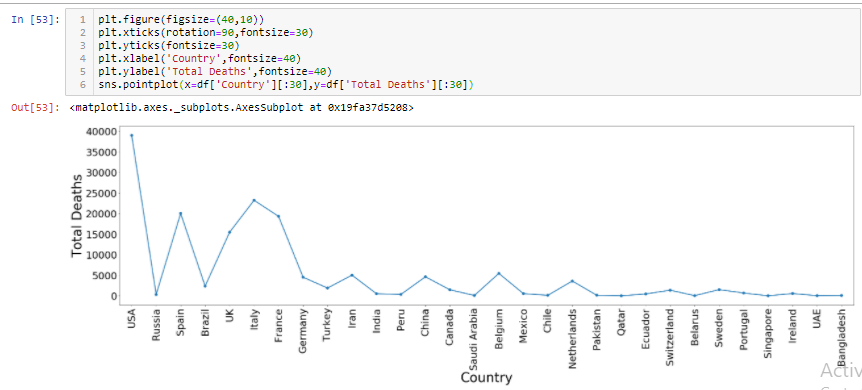
1-->**Country vs Total Deaths**



We an see from the graph that, except the starting few countries in the dataset there are few deaths in other countries.

**Lets analyze the starting 30 countries in the dataset and find some insight from it which can be helpful to the countries which are still not affected much by the virus.**

2--> **Country vs Total Deaths( starting 30 countries) on 18/4/2020**

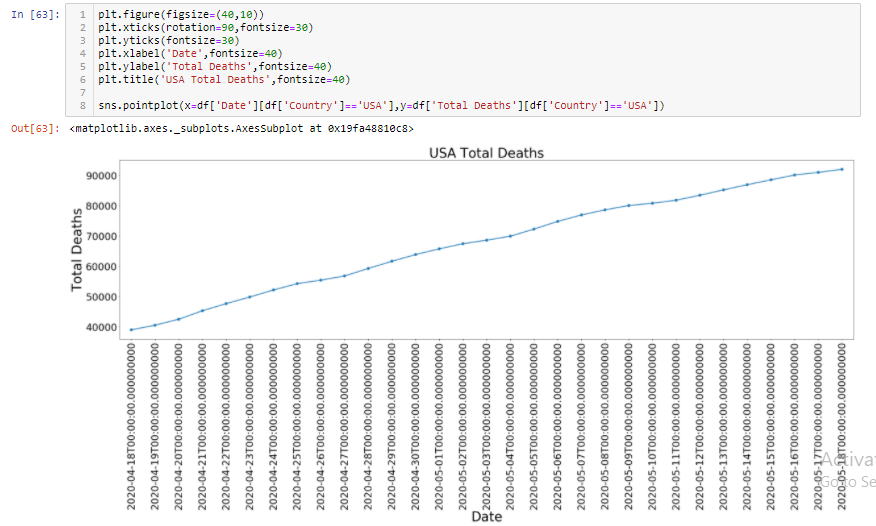


We can coclude from the graph that USA has the highest death as on date 18/4/2020.India is in initial stages.

Hence we can infer that India should start taking serious preventive measures.

Now we will do a detail analysis on the countries which have high Death numbers for a span of 31 days (18/4/2020 to 18/5/2020)and see the pattern how the no. of deaths have been controlled by the preventive measures taken by the respective country government.

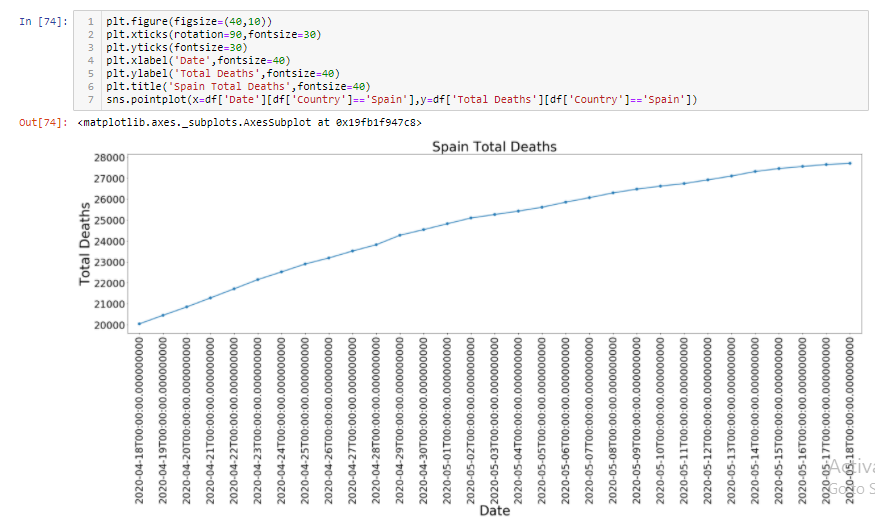
**USA**



From the graph we can say that the number of deaths have increased gradually with time and whatever preventive measures were taken by the government during this period was not effective in controlling the death numbers in the country.

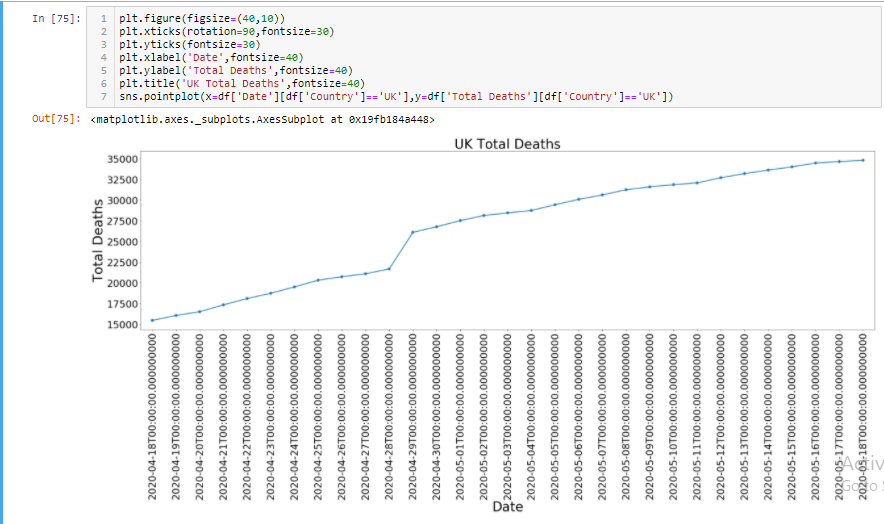
The strategy of lockdown,shutdown etc dint have any effect in controlling the no. of deaths in the country.The stratergy has to be changed or the period of lockdown,shutdown has to be increased so as to control the situation in the country.

**Spain**

****

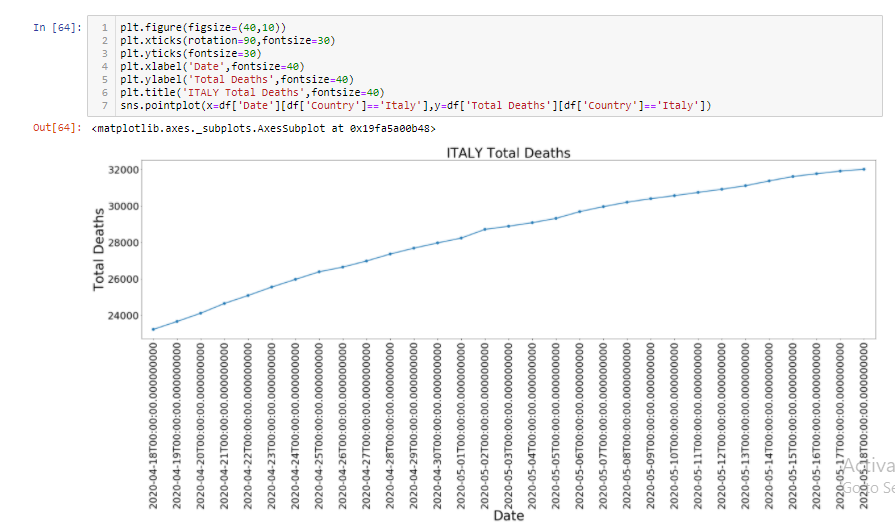
Situation is same as USA,There is a rise in number of deaths, strategy not working.

**UK**

****

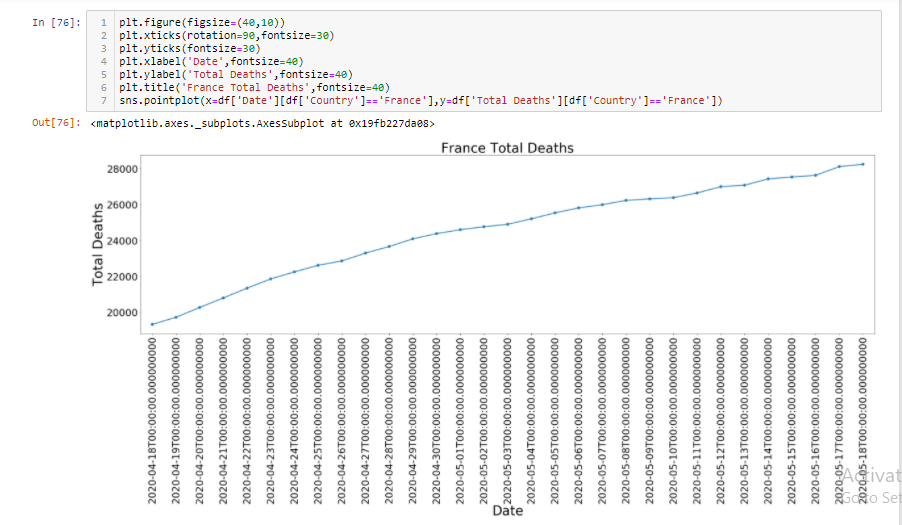
Here the situation got became more challenging after 28th april as we can see from graph there is a sudden rise in number of deaths on 29th april.Here also the strategy implemented did not work rather there is a spike in number of deaths instead od being linear as in USA,Spain.

**Italy**

****

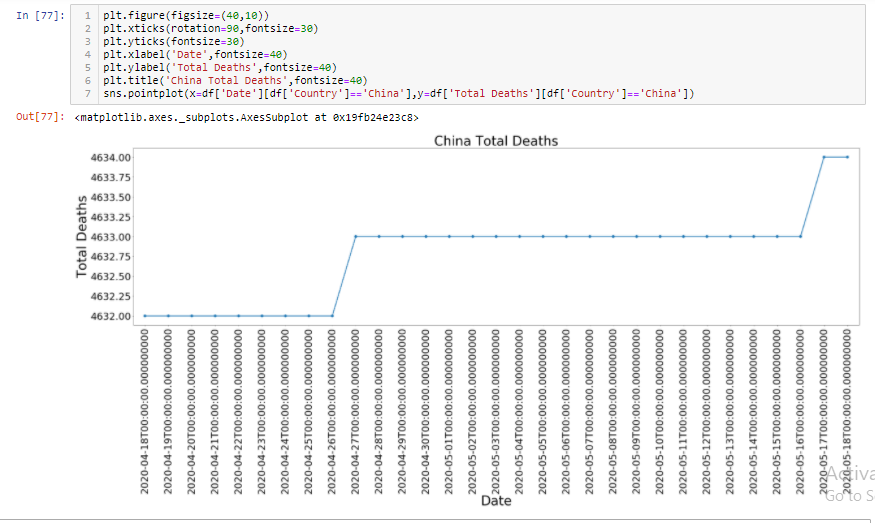
Situation is same as USA,Spain.There is a rise in number of deaths, strategy not working.

**France**

****

Situation is same as USA,Spain,Italy.There is a rise in number of deaths, strategy not working.

**China**

****

China has controlled the situation very well as from the graph we can say there is no rise in number of deaths.

**The graph has flattened after a rise on26th april till 19th may.**

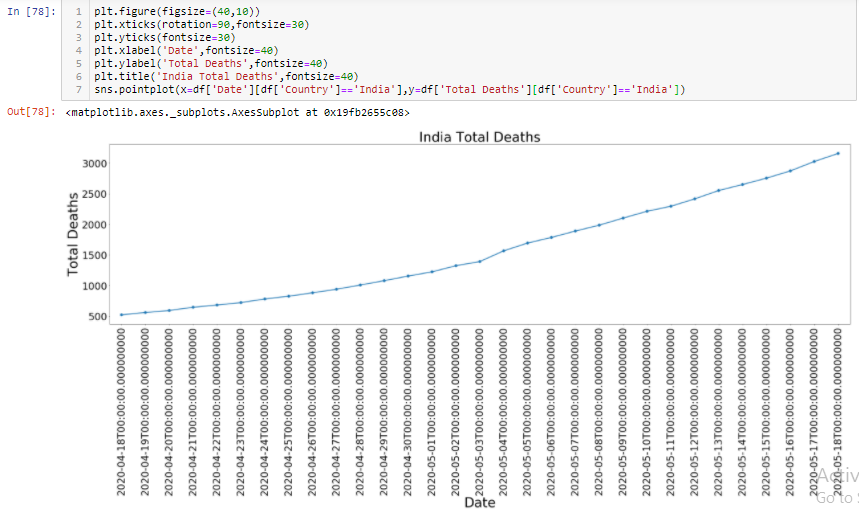
Whatever stratergy china applied was successful in controlling the increasing number of deaths in their country till 16th may.

Again on 17th may there is asteep rise in number of deaths but again we can see the graph to be flattening as per our data available.

China must share these strategies to the whole world so that everybody can apply in their respective countries and stop the spread of the virus as well the number of deaths.

Finall lets check for our country INDIA

**India**

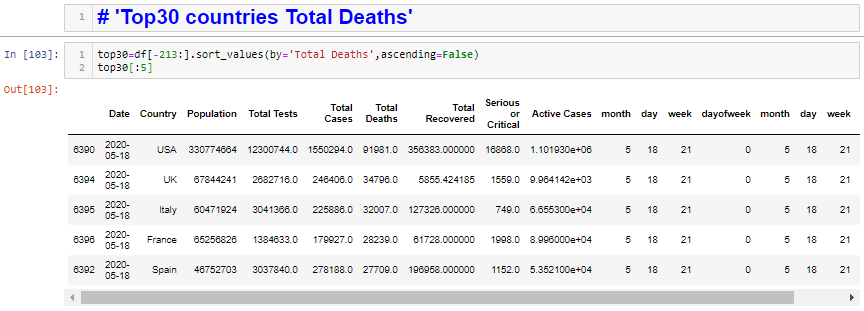
****

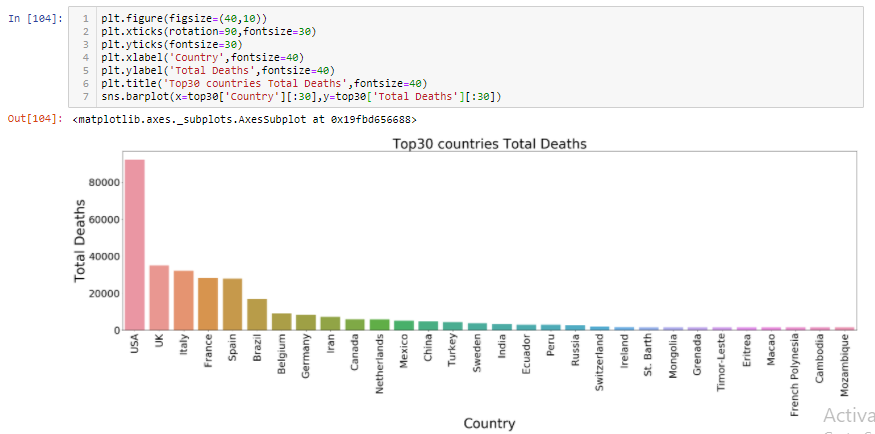
Though the number of deaths are not so high in comparision to other countries,still we can see the numbers increasing day by day as was in case of USA,Italy,Spain.

This is high time to take some serious preventive measures in india as India is more densely populated than the countries which are showing gradual increase in death numbers.

The situation if not controlled from now then the situation might get worse in coming future.

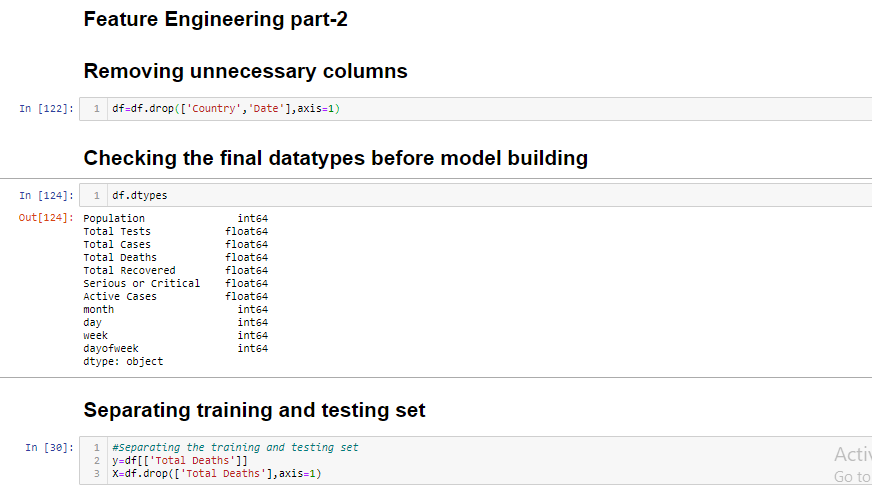
**Now lets checkout the Top30 countries Total Death**

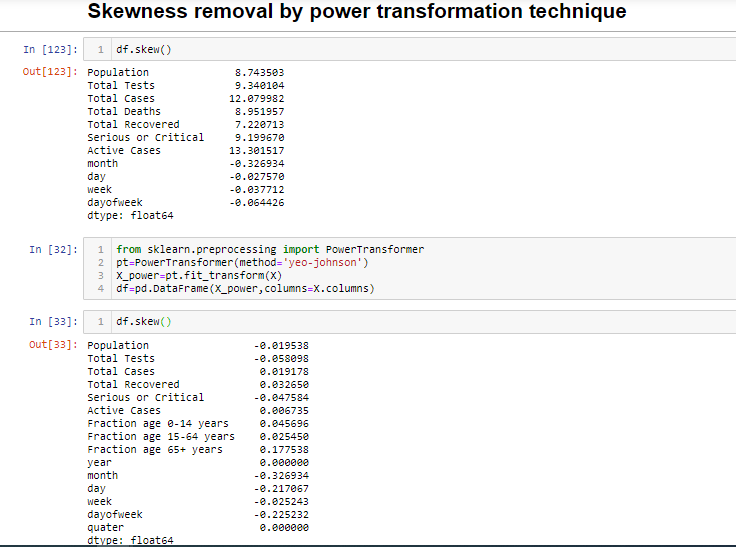




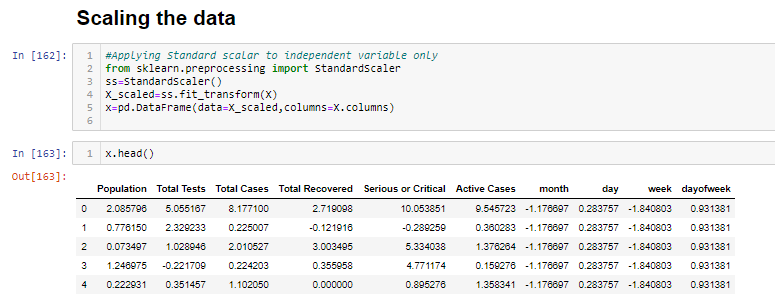
So we can see in the graph the Country wise Total deaths as on 18/5/2020.

Its high time in India we start taking preventive measures seriously and make strategy if the situation gets out of control.Its always better to have contingency plan as a super power like USA is not able to control the outbreak of the virus we should be more careful and ready for the worst case scenario.



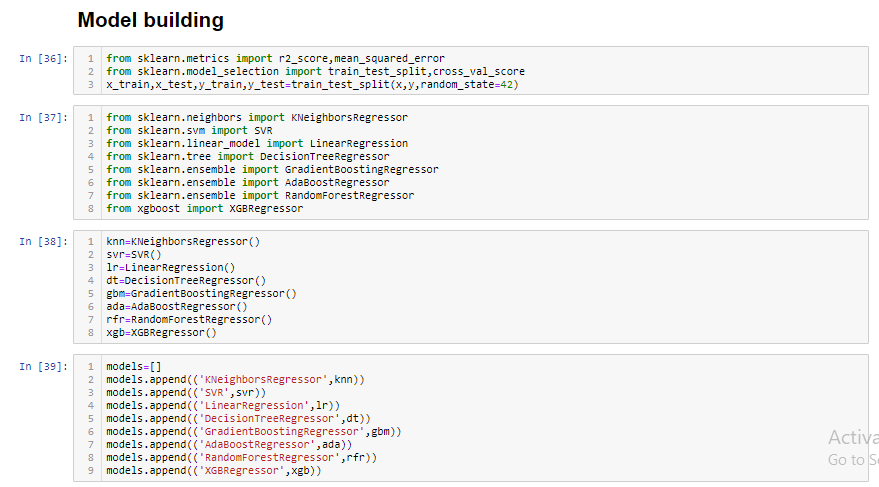


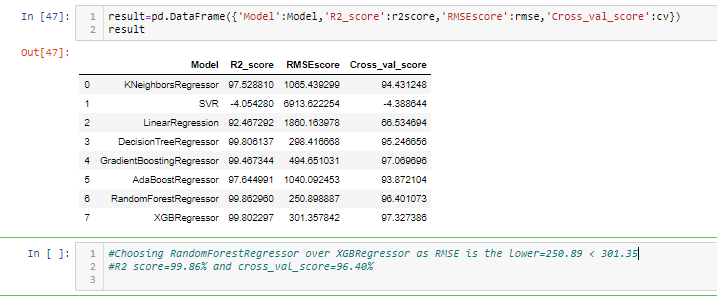
Data scaling to a common base unit is one of the important process to be performed in a machine learning process as all the features in the dataset does not have the same units.



Checking the data we can see that all the data have been scaled and in numerical form.

Hence the data is ready to feed to the model building process as we know our machine learning algorithms only can process numerical data.

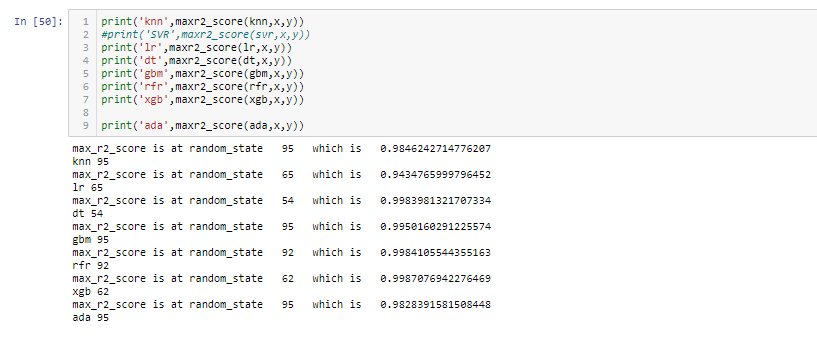




Finding the best random\_state will give us the random state at which the individual model gives the best performance which we will be using in later stages to determine the final model.

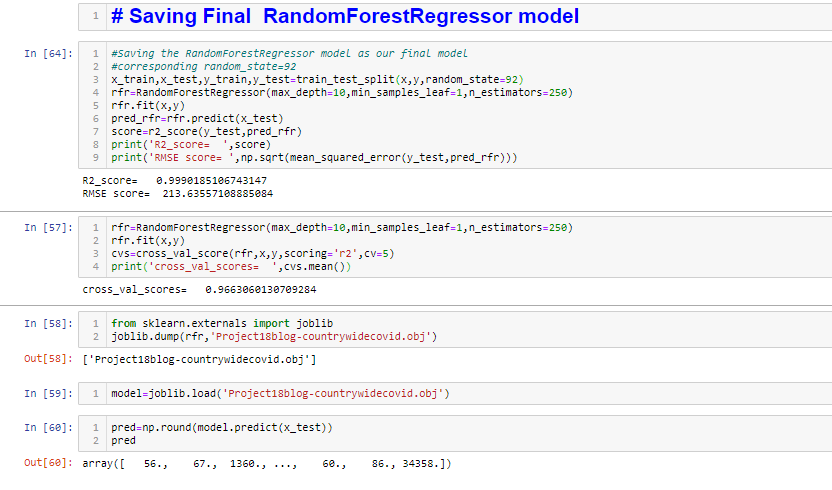
We are neglecting the SVR model here because our r2\_score for the SVR model is very low i.e negative also.

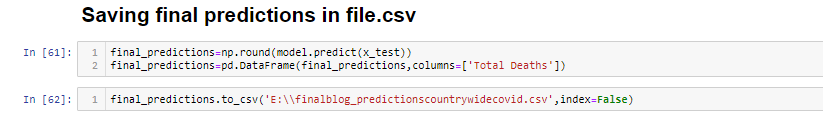




GridsearchCV at best Random\_state=92:-







Conclusion:-

-Total Deaths of a country on a particular date can be predicted using the model above.

-By using sklearn(XGBRegressor)we have build a machine learning model that will help us predict Total Deaths due to covid in different countries.

-This predictive model mainly can be used as a case study in a country where currently the no. of deaths are not so high .

-Hence preventive measures can be taken by country to stop the spread of the virus by comparing the preventive measures taken by the more infected countries as a benchmark.

-Modification in preventive measures can be also decided taking the preventive measures of more infected countries as a benchmark.

**Thoughts for India**

This is high time to take some serious preventive measures in India as India is more densely populated than the countries like USA,Italy,Spain etc which are showing gradual increase in death numbers.

The situation if not controlled from now then the situation might get worse in coming future.

If u like my work please contact me on [sourov.sahoo@gmail.com](mailto:sourov.sahoo@gmail.com) to take the project towards perfectness.

Two things I want to do with this model next:-

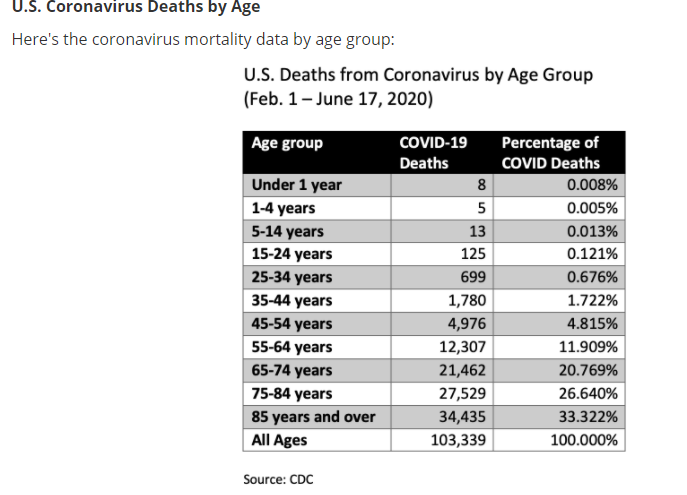
1🡪 Testing the model with 30 days India data.

2🡪As India is a more populated country I would like to predict the number of deaths grouped by age of each country.But the data of all the 213 countries,is not available with me currently .

I currently have the data of USA only from the source link given below:-

<https://www.acsh.org/news/2020/06/23/coronavirus-covid-deaths-us-age-race-14863>

Please feel free to contribute the data and help take the project further. I have just started the project and don’t know where I will end.This is my first step towards the Data world.Trying to contribute something toward mankind.Hopefully this situation never comes in in future.But we will always have to be ready with a base level contingency plan.



Keeping the restriction of word limits to 2000 in mind I am ending my Article here. I have some exams going on with the esteemed institute( ) from where I started my data science carrier past 5-6 months. So currently I am not able to collect the required data to take the project further. In few days I will upgrade the project again after collecting the data and preprocessing the data.

**Please do revert me back if I can or// cannot take the name of the institute and teachers while posting this content on different social media platforms.**

**Courtesy:-**

**1.Deepika madam**

**2.Vishal sir**

**Yours sincerely,**

**Sourov Sahoo**

-