HarvardX: PH125.9x: Data Science: Capstone: Predict-Spreading-Coronavirus(COVID19)

Tarun Ch. Bordoloi

3/30/2020

1.0 Executive summary:

1.1 Background:

The Project : Predict the Spreading of Coronavirus(COVID19)

As advised under the 'Project Overview: Choose Your Own!' section Of the 'HarvardX: PH125.9x: Data Science: Capstone' I have choosen the project – "Predict the Spreading of Coronavirus" from Kaggle considering its very critical importance and contemporary relevance. I am aware that being an absolute beginner in the field of data analysis with my very basic knowledge it will be a challenging task. Further, due to very little historical as well as epidemiological data available at this point and also in the absence of adequate study on this novel Corona virus till date it is virtually impossible at this stage to offer a credible prediction of the nature and degree of its spread.Yet, I have volunteered to accept the challenge with lot of excitement.

COVID-19 Novel Coronavirus :

The 2019–20 coronavirus pandemic is an ongoing pandemic of coronavirus disease 2019 (COVID-19) caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2).The outbreak was first identified in Wuhan, Hubei, China, in December 2019 and recognised as a pandemic by the World Health Organization (WHO) on 11 March 2020, as the first known pandemic that can be controlled.As of 27th March, over 5,50,000 cases of COVID-19 have been reported in more than 176 countries and territories, with major outbreaks in United States, mainland China, Europe, Iran, and South Korea, among others,which include the cruise ship Diamond Princess.More than 25000 people have died from the disease and over 1,27,000 have recovered. The virus primarily spreads between people in a way similar to influenza, via respiratory droplets from coughing or sneezing.The time between exposure and symptom onset is typically five days, but may range from two to fourteen days.Symptoms are most often fever, dry cough, and shortness of breath.Complications may include pneumonia and acute respiratory distress syndrome. There is no vaccine or specific antiviral treatment, but research is ongoing. Efforts are aimed at managing symptoms and supportive therapy. Public health responses have included national pandemic preparedness and response plans,travel restrictions, quarantines, curfews, event postponements and cancellations, and facility closures. Effects of the pandemic include social and economic instability, xenophobia and racism, and the online spread of misinformation and conspiracy theories about the virus.(Source :Wikipedia)

1.2 Summary goals:

The outbreak of Covid-19 is developing into a major international crisis, and it's starting to influence important aspects of daily life. For example: • Travel: Bans have been placed on hotspot countries, corporate travel has been stopped/reduced. • Supply chains: International manufacturing operations have often had to throttle back production and many goods solely produced in China have been halted altogether. • Grocery stores: In highly affected areas, people are starting to stock up on essential goods.

A strong model that predicts how the virus could spread across different countries and regions may be able to help mitigation efforts. The goal of this task is to build a model that predicts the progression of the virus throughout March 2020

1.3 The Data set :

On 31 December 2019, WHO(World Health Organisation) was alerted to several cases of pneumonia in Wuhan City, Hubei Province of China. The virus did not match any other known virus. This raised concern because when a virus is new, it is not known how it affects people.

So daily level information on the affected people can give some interesting insights when it is made available to the broader data science community.

Johns Hopkins University has made an excellent dashboard using the affected cases data. Data is extracted from the google sheets associated and made available here.

These data are now taken from the Johns Hopkins Github repository where it is available as csv files .

1.4 Key steps performed :

- Downloaded the datasets $ Ensured that the required packages and libraries are installed
- Carried out the exploration of the data and performed feature engineering as required $ Included data visualization tools as required $ Incorporated insights gained
- Algorithm for 2 Models, namely 'Base Line' and 'FB Prophet Forecast', were developed and those were evaluated $ Results tabulated - with relevant section of the report
- Conclusion stated

1.5 Installing the required packages, loading required libraries & downloading data :

Let us now have a glimpse of the data sets. Both 'covid_19_confirmed' and 'covid_19_deaths' data sets have 249 observations and 4 variables as follows , while the 'covid_19_recovered' data set has 235 observations and 4 varibles as mentioned above :

'Province/State' 'Country/Region' 'Latitude' 'Longitude'

Starting from column 5 of these data sets, each column corresponds to a single day these are .

We need to check the time frame of the data now.

```
## [1] "2020-01-22" "2020-03-29"
```

```
## [1] "22 January 2020"
```

```
## [1] "29 March 2020"
```

It would appear that the data was last updated on the 27 March 2020 UTC at the point of compilation of this report finally. All the stats and charts in this report are based on that data.

2.0 Data Preparation :

2.1 Data Cleaning & Feature engineering :

We will now be carrying out the following operations – I. Three data sets will be converted from wide to long format. II. They will be aggregated by country. III. Then they will be merged into a one single data set.

We shall now be cleaning the 3 data sets

We shall now be merging above 3 data sets into one by country and date

It would appear that in case of China where(Wuhan) the first instances were detected and it was spreading uncotrollably like wild fire , although the recorded number of total confirmed cases have risen to 81897

Table 1: Raw data, First 10 Columns only)

	country	date	confirmed	deaths	recovered
59	China	2020-03-20	81250	3253	71266
60	China	2020-03-21	81305	3259	71857
61	China	2020-03-22	81435	3274	72362
62	China	2020-03-23	81498	3274	72814
63	China	2020-03-24	81591	3281	73280
64	China	2020-03-25	81661	3285	73773
65	China	2020-03-26	81782	3291	74181
66	China	2020-03-27	81897	3296	74720
67	China	2020-03-28	81999	3299	75100
68	China	2020-03-29	82122	3304	75582

as of 27th March 2020 the remarkable observation is that the number of recovered cases has now been recorded at 74720 and very encouragingly death cases have more or less remained at 3296 during the same period.This would suggest that China has decisively managed to get the situation in her grip.This has raised and reaffirmed the hope that COVID19 is not invincible and it is important that China shares their experience with the rest of the world most of which are still grappling with the scourge of the virus so that they too can avail the benefit of their model.After all it is an issue where the entire humanity is getting threatened.
Data for the whole world : The raw data above provide the number of cases every day for every country. Those figures will now be aggregated to obtain the statistics of the whole world

```
##        country       date confirmed deaths recovered current.confirmed
## 1  Afghanistan 2020-01-22         0      0         0                 0
## 2  Afghanistan 2020-01-23         0      0         0                 0
## 3  Afghanistan 2020-01-24         0      0         0                 0
## 4  Afghanistan 2020-01-25         0      0         0                 0
## 5  Afghanistan 2020-01-26         0      0         0                 0
## 6  Afghanistan 2020-01-27         0      0         0                 0
## 7  Afghanistan 2020-01-28         0      0         0                 0
## 8  Afghanistan 2020-01-29         0      0         0                 0
## 9  Afghanistan 2020-01-30         0      0         0                 0
## 10 Afghanistan 2020-01-31         0      0         0                 0
```

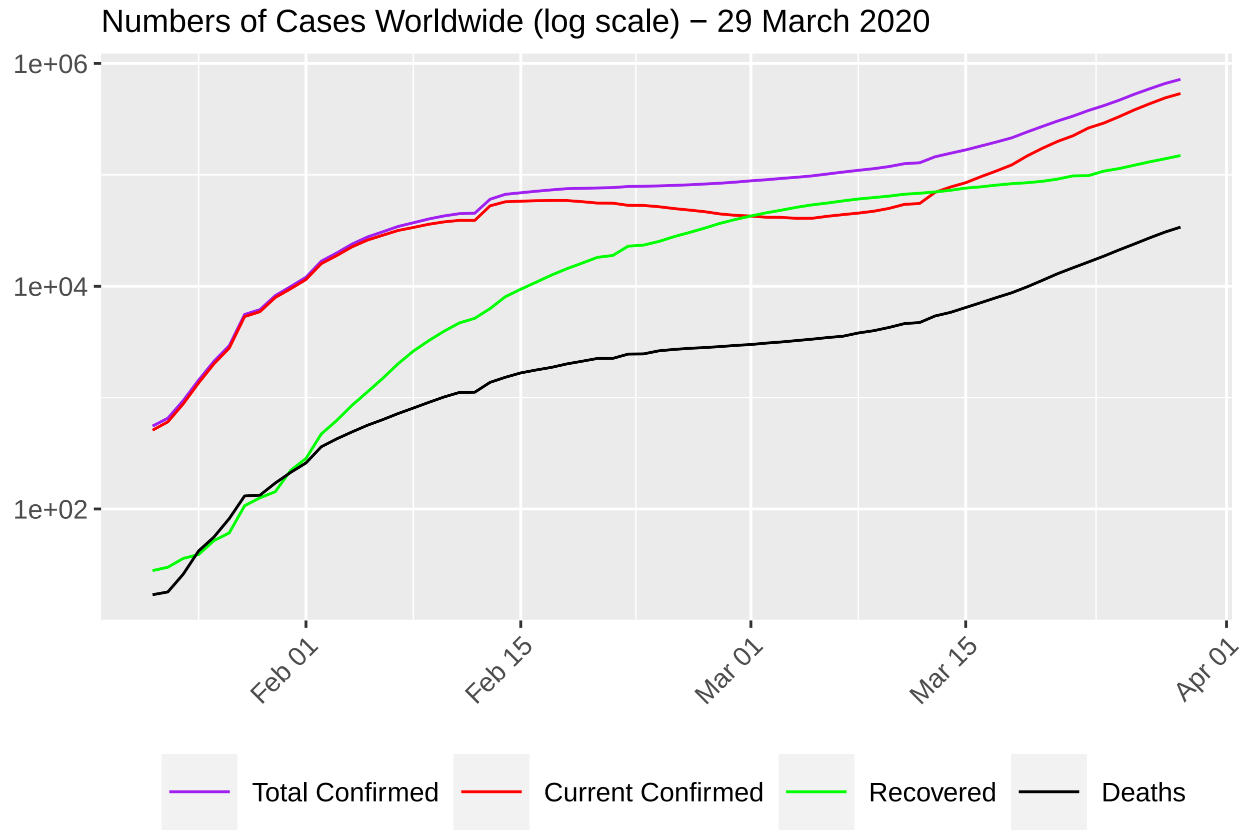
```
##       country       date confirmed deaths recovered current.confirmed
## 12095   World 2020-03-20    272035  11299     87420            173316
## 12096   World 2020-03-21    304396  12973     91692            199731
## 12097   World 2020-03-22    336953  14651     97899            224403
## 12098   World 2020-03-23    378235  16505     98351            263379
## 12099   World 2020-03-24    418045  18625    108000            291420
## 12100   World 2020-03-25    467653  21181    113787            332685
## 12101   World 2020-03-26    529591  23970    122150            383471
## 12102   World 2020-03-27    593291  27198    130915            435178
## 12103   World 2020-03-28    660706  30652    139415            490639
## 12104   World 2020-03-29    720117  33925    149082            537110
```

If we now look at the whole world scenarion it is indeed alarming. Recorded confirmed cases have risen to 593291 and recorded recovered number has only been 22.07% at 130915 while number of deaths has been 4.60% at 27198 as on 27th Marcg 2020.However , it will be noteworthy that 73.34% of the confirmed cases at 435178 is remainining confirmed which would suggest that hopefully substanial number, if not all, of these cases too would recover with the advent of developing management/ treatment protocols.

3.0 Data Visualisation :

Let us now visualise the data that we have tidied above

3.1 Whole World Scenario :

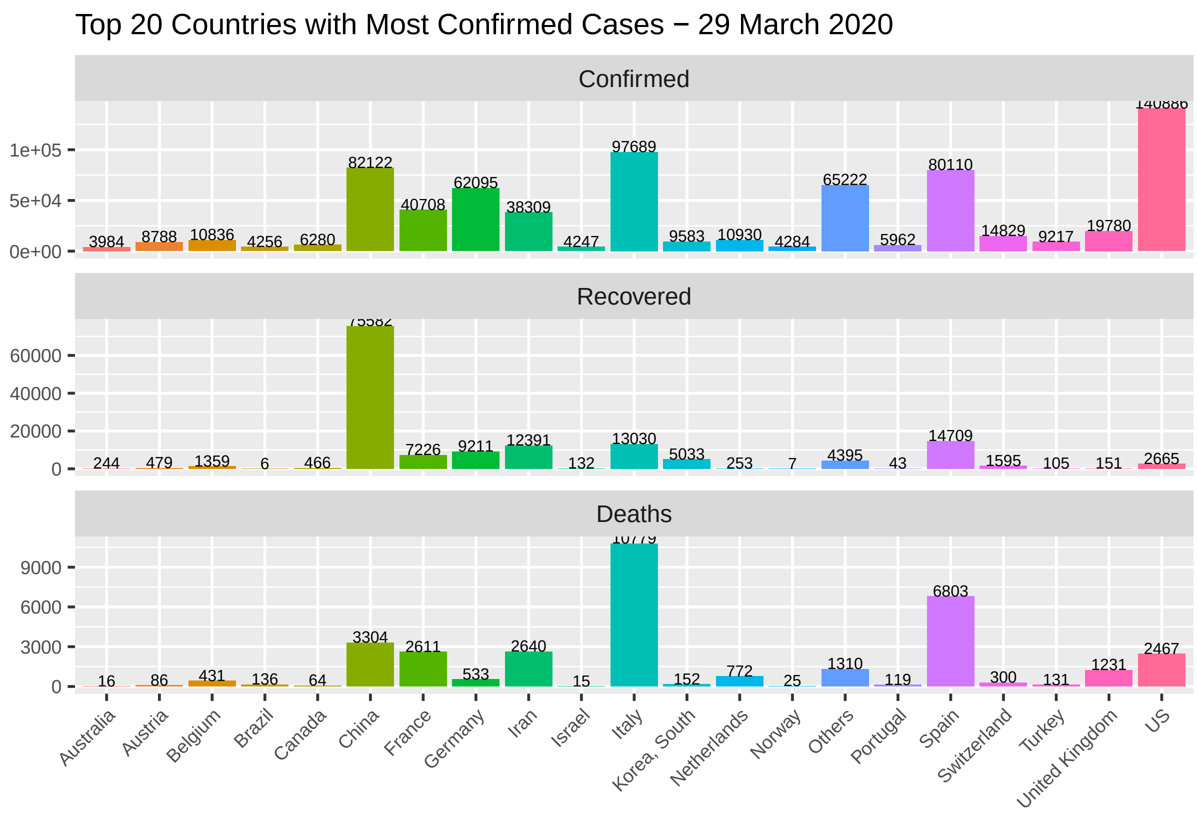


3.2 Top countries :

Following are the top 20 countries in confirmed cases

```
##  [1] "US"              "Italy"           "China"           "Spain"
##  [5] "Germany"         "France"          "Iran"            "United Kingdom"
##  [9] "Switzerland"     "Netherlands"     "Belgium"         "Korea, South"
## [13] "Turkey"          "Austria"         "Canada"          "Portugal"
## [17] "Norway"          "Brazil"          "Israel"          "Australia"
```

```
## Warning: Unknown or uninitialised column: 'txt'.
```

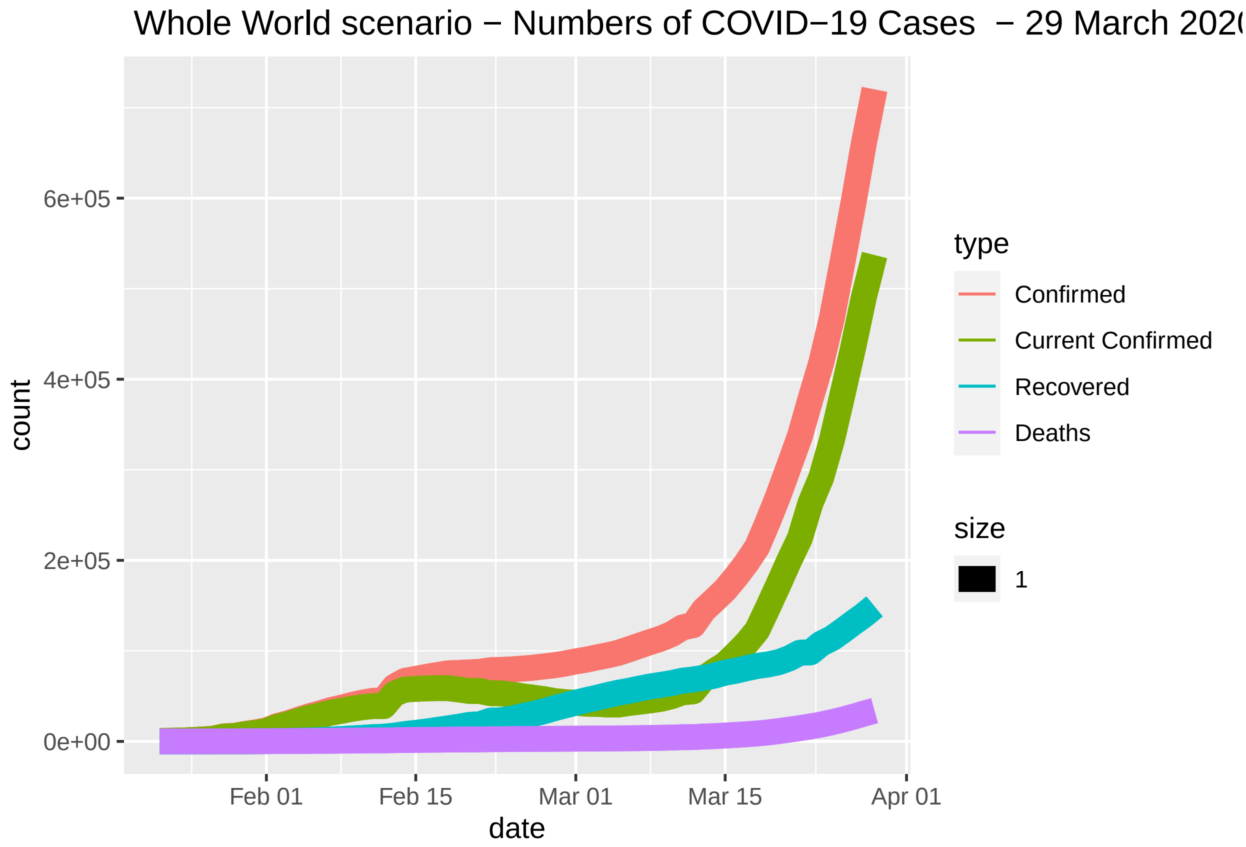


Let us now look at the whole world scenarion again , generally , as it is developing.

Table 2: Cases in Top 20 Countries - 29 March 2020

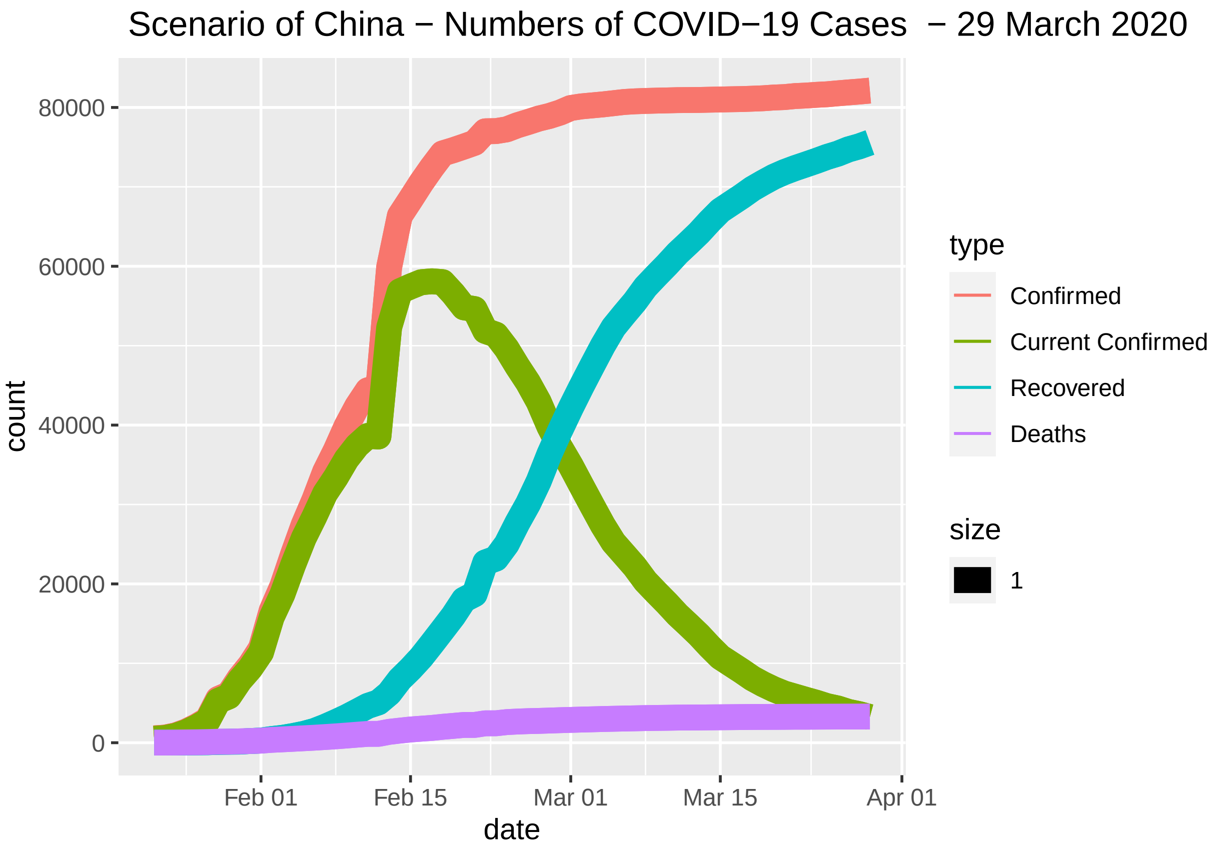
country	type	count
Australia	Confirmed	3984
Austria	Confirmed	8788
Belgium	Confirmed	10836
Brazil	Confirmed	4256
Canada	Confirmed	6280
China	Confirmed	82122
France	Confirmed	40708
Germany	Confirmed	62095
Iran	Confirmed	38309
Israel	Confirmed	4247
Italy	Confirmed	97689
Korea, South	Confirmed	9583
Netherlands	Confirmed	10930
Norway	Confirmed	4284
Others	Confirmed	65222
Portugal	Confirmed	5962
Spain	Confirmed	80110
Switzerland	Confirmed	14829
Turkey	Confirmed	9217
United Kingdom	Confirmed	19780
US	Confirmed	140886
Australia	Recovered	244
Austria	Recovered	479
Belgium	Recovered	1359
Brazil	Recovered	6
Canada	Recovered	466
China	Recovered	75582
France	Recovered	7226
Germany	Recovered	9211
Iran	Recovered	12391
Israel	Recovered	132
Italy	Recovered	13030
Korea, South	Recovered	5033
Netherlands	Recovered	253
Norway	Recovered	7
Others	Recovered	4395
Portugal	Recovered	43
Spain	Recovered	14709
Switzerland	Recovered	1595
Turkey	Recovered	105
United Kingdom	Recovered	151
US	Recovered	2665
Australia	Deaths	16
Austria	Deaths	86
Belgium	Deaths	431
Brazil	Deaths	136
Canada	Deaths	64
China	Deaths	3304
France	Deaths	2611
Germany	Deaths	533
Iran	Deaths	2640
Israel	Deaths	15
Italy	Deaths	10779
Korea, South	Deaths	152
Netherlands	Deaths	772

3.3 Visualising World scenario to add few more observations :



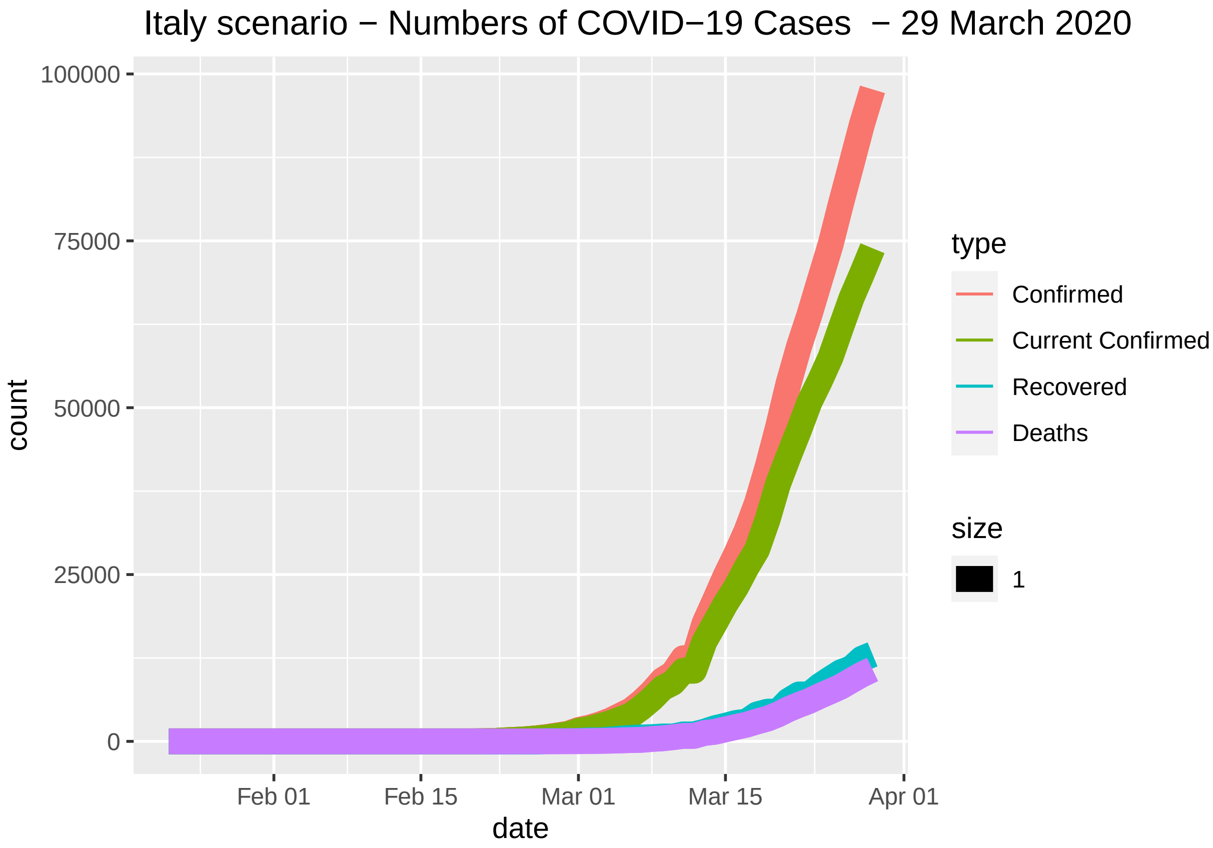
It would appear that since 1st Feb.2020 number of confirmed cases world wide has been surging up ,quite significantly at that from around the 1st March 2020. The remaining confirmed cases would seem to have also been surging up quite significantly since about the 10th March 2020 after a declining trend during the period 15th February 2020 and the 7th March 2020.That could be due to the fact that the number of cases recovering would seem to have been increasing since the 15th Feb.2020. The number of deaths has been showing a constant to gradually increasing trend since about the 1st week of Feb. 2020. In my overall assesment these observations would suggest that the COVID 19 has been assuming a global pandemic proportion since about the early March 2020.

3.4 Let us have a look at the China scenario , as that is where(Wuhan) the epicenter of this pandemic has been.



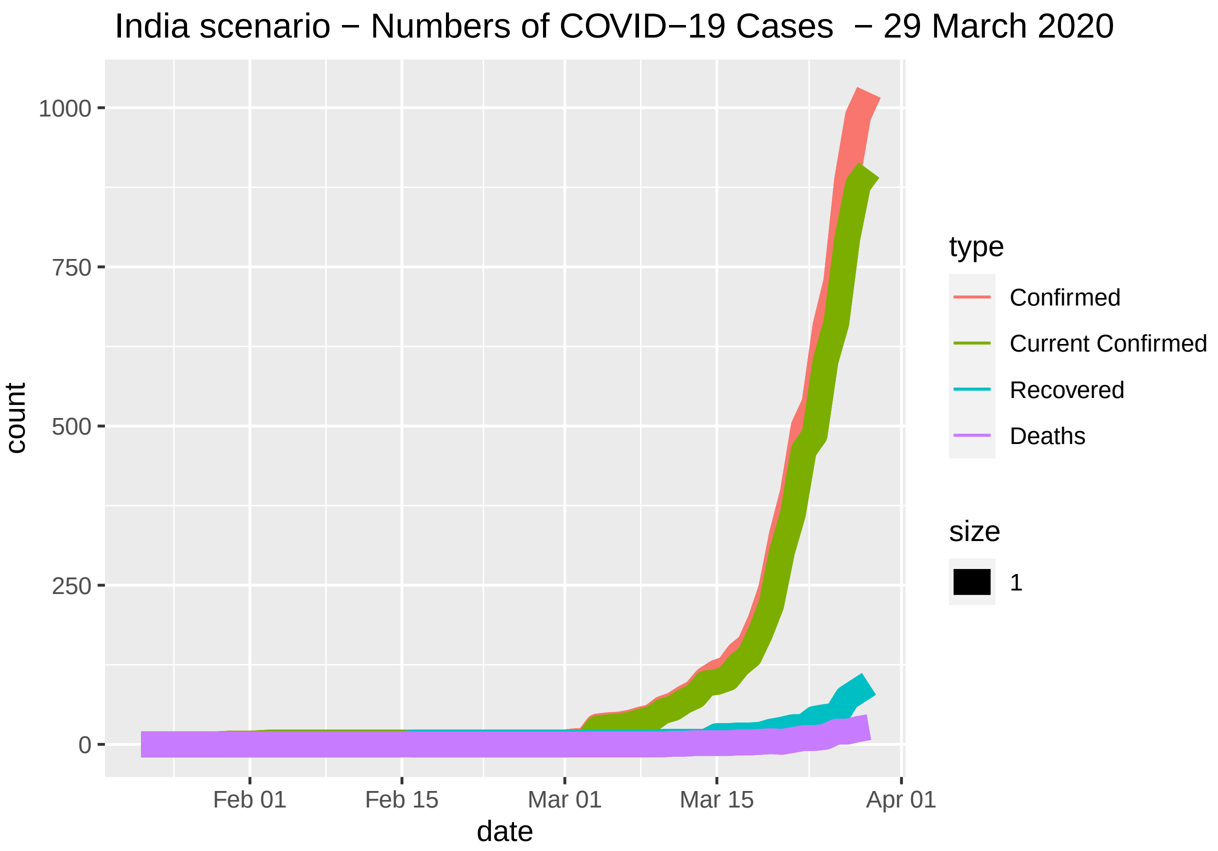
So far as China is concerned the number of confirmed cases seem to have plateued since about the first week of March 2020, after a steep increase since about the early Feb. 2020. Remaining confirmed cases too have started declining since about the 10th Feb. 2020.There has been an encouraging increase in cases recovered since about the 10th Feb 2020.The number of deaths although showing a increasing trend ,the rate of increase would seem to be small.This would suggest that China has been succeeding in getting the situation under their grip and COVID 19 is indeed not invincible.When the rest of the world is still grappling with a seriously emerging scenario in order to deal with this mammoth challenge posed to the entire humanity China perhaps could offer to share their experience.

3.5 Italy Scenario :



So far Italy is concerned the number of confirmed cases as well as the remaining confirmed cases these seem to have been increasing quite rapidly since about the last week of Feb. 2020 and almost in similiar prportion. Further , the number of deathcases would seem to have been increasing significantly since about the first week of March 2020 and so is the case in respect of cases recovering, though not in comfortable numbers.

3.6 Developing India Story :



So far as India is concerened the number of confirmed cases as well as remaining con firmed cases seem to have been surging up rapidly since about the 7th March 2020 and that too in somewhat similiar numbers with cases of deaths and recovery in small numbers . There seem to have been some encouraging improvement in caseses recovered since about the 12th March 2020. This would suggest that although the inrese in rapidity of increase in confirmed cases as well as remaining confirmed cases has been a matter of serious concern the trend of improvement in number of cases recovering since about the 12th March 2020 is encouraging. ### 4.0 Building and evaluating model for prdicting the spread of COVID 19:

We shall be considering the China case since the developments there has been in somewhat advanced stage.

4.1 : Base Line Model :

In its simplest form this model is generated by considering the same value for 'new.confirmed' cases(every day addition of fresh cases) for all the days irrespective of different possible causes.All the differences explained by random variation. The formula would look like this: $Y = \hat{\mu} + \varepsilon$ With $\hat{\mu}$ is the mean and ε is the independent errors sampled from the same distribution centered at 0

```
## Warning in set.seed(1, sample.kind = "Rounding"): non-uniform 'Rounding' sampler
## used
```

```
##    country       date confirmed deaths recovered current.confirmed
## 2    China 2020-01-23       643     18        30               595
## 3    China 2020-01-24       920     26        36               858
## 4    China 2020-01-25      1406     42        39              1325
## 5    China 2020-01-26      2075     56        49              1970
## 6    China 2020-01-27      2877     82        58              2737
## 7    China 2020-01-28      5509      131       101               5277
## 9    China 2020-01-30      8141      171       135               7835
## 10   China 2020-01-31      9802      213       214               9375
## 11   China 2020-02-01     11891      259       275              11357
## 12   China 2020-02-02     16630      361       463              15806
##    new.confirmed new.deaths new.recovered
## 2             95          1             2
## 3            277          8             6
## 4            486         16             3
## 5            669         14            10
## 6            802         26             9
## 7           2632         49            43
## 9           2054         38            15
## 10          1661         42            79
## 11          2089         46            61
## 12          4739        102           188
```

Base Line model:

```
## [1] 1273.186
```

If we predict all the values of ‘new.confirmed’ cases of the validation set with $\hat{\mu}$ our RMSE will be as follows

RMSE_BASE

```
## [1] 1150.534
```

RMSE prediction is 1157.907. We shall see how our rest of the models fare incomparison to this Base Line Model.

4.2 FB Prophet Forecast Model:

Prophet is a procedure for forecasting time series data based on an additive model where non-linear trends are fit with yearly, weekly, and daily seasonality, plus holiday effects. It works best with time series that have strong seasonal effects and several seasons of historical data. Prophet is claimed to be robust to missing data and shifts in the trend, and typically handles outliers well.Prophet is again claimed to be robust to outliers, missing data, and dramatic changes in time series.The Prophet procedure includes many possibilities for users to tweak and adjust forecasts. One can use human-interpretable parameters to improve forecast by adding ones domain knowledge.One disadvantage in our this particular project could be absence of long term historical data. Even one years data would have helped.Althogh Prophet is meant for business operations for forecasts for planning and goal setting its features and capabilities prompts me to believe that it holds good promise of being a good tool for this kind of project .

```
## Warning: package 'prophet' was built under R version 3.6.3
```

```
## Loading required package: Rcpp
```

```
## Loading required package: rlang
```

```
## Warning: package 'rlang' was built under R version 3.6.3
```

```
##
## Attaching package: 'rlang'
```

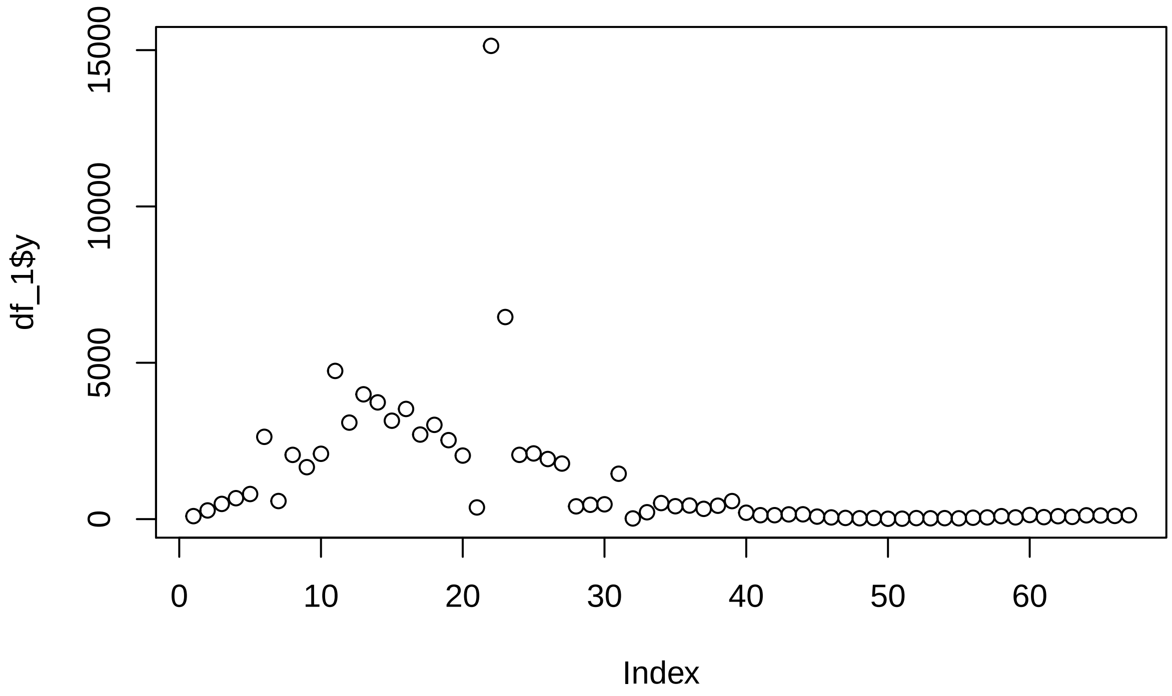
```
## The following object is masked from 'package:magrittr':
##
##     set_names
```

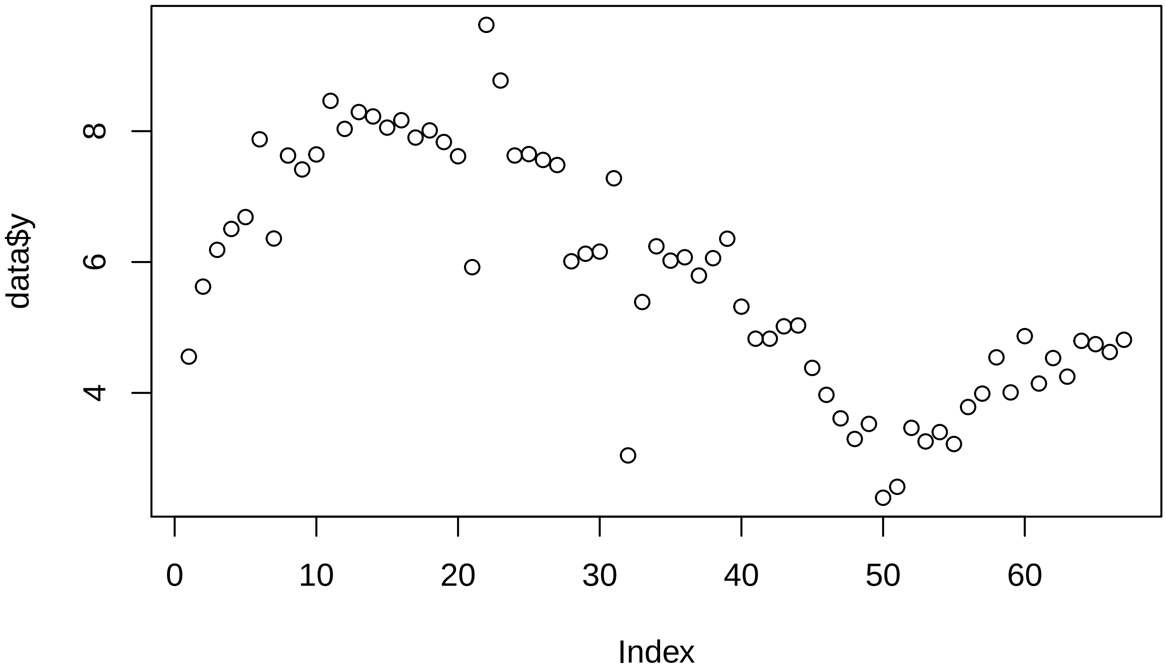
```
## The following object is masked from 'package:data.table':
##
##     :=
## The following objects are masked from 'package:purrr':
##
##     %@%, as_function, flatten, flatten_chr, flatten_dbl, flatten_int,
##     flatten_lgl, flatten_raw, invoke, list_along, modify, prepend,
##     splice
```

```
## 'data.frame':    67 obs. of  2 variables:
##  $ date         : Date, format: "2020-01-23" "2020-01-24" ...
##  $ new.confirmed: num  95 277 486 669 802 ...
```

```
##   country       date confirmed deaths recovered current.confirmed new.confirmed
## 1   China 2020-01-23       643     18        30               595            95
## 2   China 2020-01-24       920     26        36               858           277
## 3   China 2020-01-25      1406     42        39              1325           486
## 4   China 2020-01-26      2075     56        49              1970           669
## 5   China 2020-01-27      2877     82        58              2737           802
## 6   China 2020-01-28      5509    131       101              5277          2632
##   new.deaths new.recovered         ds    y
## 1          1             2 2020-01-23   95
## 2          8             6 2020-01-24  277
## 3         16             3 2020-01-25  486
## 4         14            10 2020-01-26  669
## 5         26             9 2020-01-27  802
## 6         49            43 2020-01-28 2632
```

Let us plot





Let us fit the model on the data - ‘df_1’

```
## Disabling yearly seasonality. Run prophet with yearly.seasonality=TRUE to override this.
```

```
## Disabling daily seasonality. Run prophet with daily.seasonality=TRUE to override this.
```

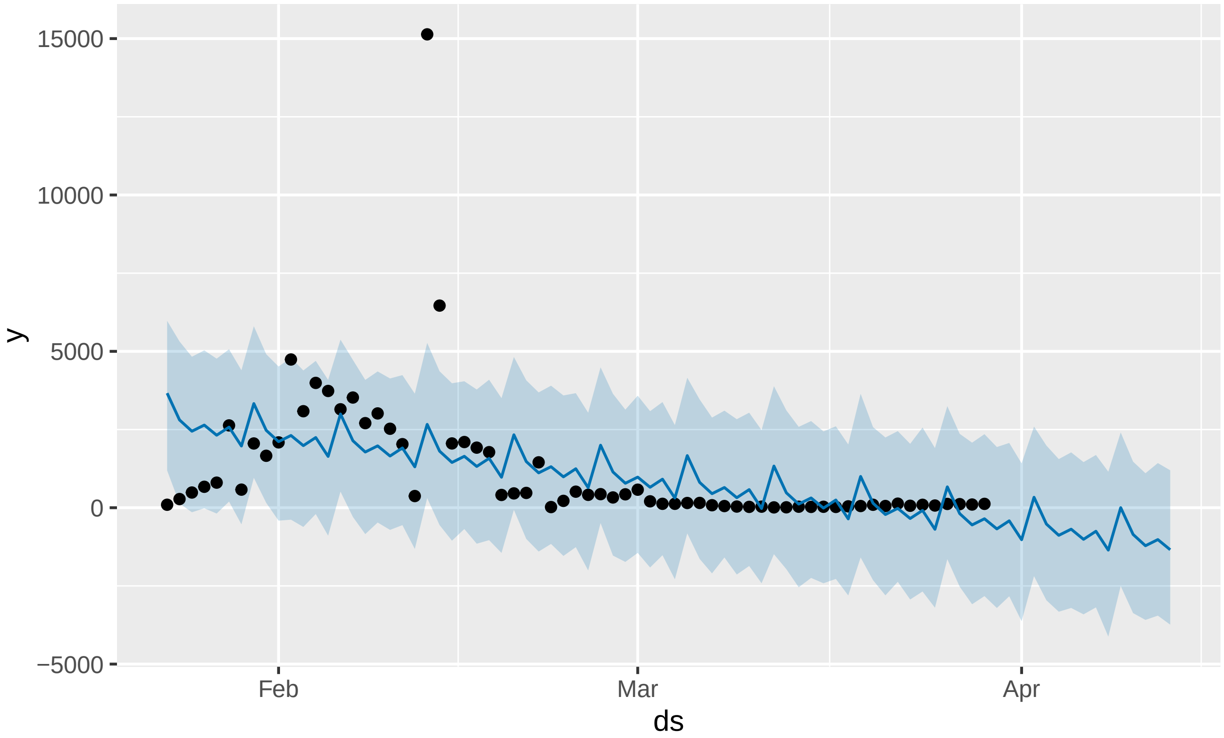
Predictions can now be made on a data frame containing the dates for the forecast.

```
##           ds
## 1 2020-01-23
## 2 2020-01-24
## 3 2020-01-25
## 4 2020-01-26
## 5 2020-01-27
## 6 2020-01-28
```

```
##            ds
## 63 2020-03-25
## 64 2020-03-26
## 65 2020-03-27
## 66 2020-03-28
## 67 2020-03-29
## 68 2020-03-30
## 69 2020-03-31
## 70 2020-04-01
## 71 2020-04-02
## 72 2020-04-03
## 73 2020-04-04
## 74 2020-04-05
## 75 2020-04-06
## 76 2020-04-07
## 77 2020-04-08
## 78 2020-04-09
## 79 2020-04-10
## 80 2020-04-11
## 81 2020-04-12
## 82 2020-04-13
```

Lets predict.

```
##            ds      trend additive_terms additive_terms_lower
## 77 2020-04-08  -839.2883     -516.92406           -516.92406
## 78 2020-04-09  -886.8540      888.11360            888.11360
## 79 2020-04-10  -934.4198       79.87282             79.87282
## 80 2020-04-11  -981.9855     -234.21664           -234.21664
## 81 2020-04-12 -1029.5512       10.36284             10.36284
## 82 2020-04-13 -1077.1170     -266.70595           -266.70595
##    additive_terms_upper     weekly weekly_lower weekly_upper
## 77           -516.92406 -516.92406   -516.92406   -516.92406
## 78            888.11360  888.11360    888.11360    888.11360
## 79             79.87282   79.87282     79.87282     79.87282
## 80           -234.21664 -234.21664   -234.21664   -234.21664
## 81             10.36284   10.36284     10.36284     10.36284
## 82           -266.70595 -266.70595   -266.70595   -266.70595
##    multiplicative_terms multiplicative_terms_lower multiplicative_terms_upper
## 77                    0                          0                          0
## 78                    0                          0                          0
## 79                    0                          0                          0
## 80                    0                          0                          0
## 81                    0                          0                          0
## 82                    0                          0                          0
##    yhat_lower yhat_upper trend_lower trend_upper         yhat
## 77  -4121.478   1151.510   -839.3301   -839.2513 -1356.212358
## 78  -2504.081   2405.680   -886.9027   -886.8106     1.259565
## 79  -3369.463   1473.099   -934.4733   -934.3706  -854.546939
## 80  -3587.905   1102.021   -982.0439   -981.9309 -1216.202133
## 81  -3450.423   1427.350  -1029.6177  -1029.4895 -1019.188386
## 82  -3739.521   1196.438  -1077.1910  -1077.0473 -1343.822906
```



Plots and data above cleaarly shows the declining trend in the predictions upto 12th April 2020.Relatively speaking this looks like the most preferable model.

5.0 CONCLUSION :

On completion of my analysis I have arrived at the following conclusions : 1. These are still very early days to design an effective model for prediction of spread of COVID 19. We hope ongoing research will provide us some useful lead to pursue the analysis. 2. Precisely for the above reason there is hardly any predictor. Expect that in the coming days there will be several of those to enable a meaningful analysis. 3. Despite limitations the analysis leads us to believe that two scenarios , China and Italy, requires deeper study. The interesting question is how China got the situation under grip relatively sooner while in Italy it is getting bad to worse. There is further scope of analysis. 4. Under the constraint of the circumstances FB Prophet model seems to have performed better in terms of predicting the scenario till the 12th April 2020. However , its accuracy would be tested only after the event. Future work would entail testing few more models with more predictors and data. It would be interesting to work on the ARIMA model too. it may please be borne in mind that my analysis is based on the data available till the 27th March 2020. But , while checking different sections you will see that these data are getting regularly updated at the source. Stay safe and healthy !