Understanding the COVID-19 Pandemic as a Data Analytics Issue
Data Analysis Report
Saad Naeem i19-1207

This report covers detailed analysis of the results and findings that were uncovered during the analysis of the data. The report contains the following sections:
Results and Findings: Most of our report is comprised of this part, here we elaborate our results that we got after performing the tasks given.
Observations and limitations: This Section Contains some observations relating to the data and the limitations that are there.
Future Directions: This section will contain some recommendations or suggestions on how the current system can be improved.

1. Results and Findings:

1.1 In our first task we found the top 20 countries with the most confirmed, most death and most recovered cases. In order to do so we first separated out the data for a given day and on that given day we performed a 'groupby' on countries by getting the maximum value for a country on that given day then sorted this data in descending order and finally output 20 records and got the following results on 2020-04-10:

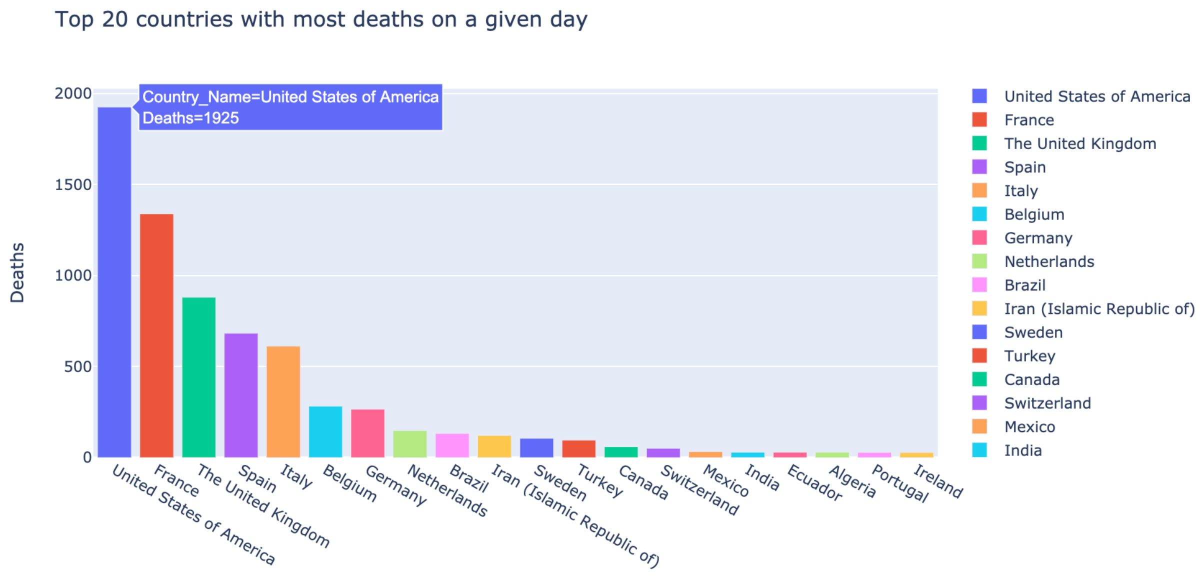


Figure 1.1

Table 1.1 shows the number of deaths on a given day here the input was 2020-04-10 it is important to note that these are not cumulative deaths i.e. these are number of people who died in a single day. As we start to go back in time, we see that the order of the countries change. for example, on 2020-03-20 Italy had most deaths (625) and United States of America stood at fifth place (with 51 deaths).

The same approach was applied to the Confirmed cases with the following results:

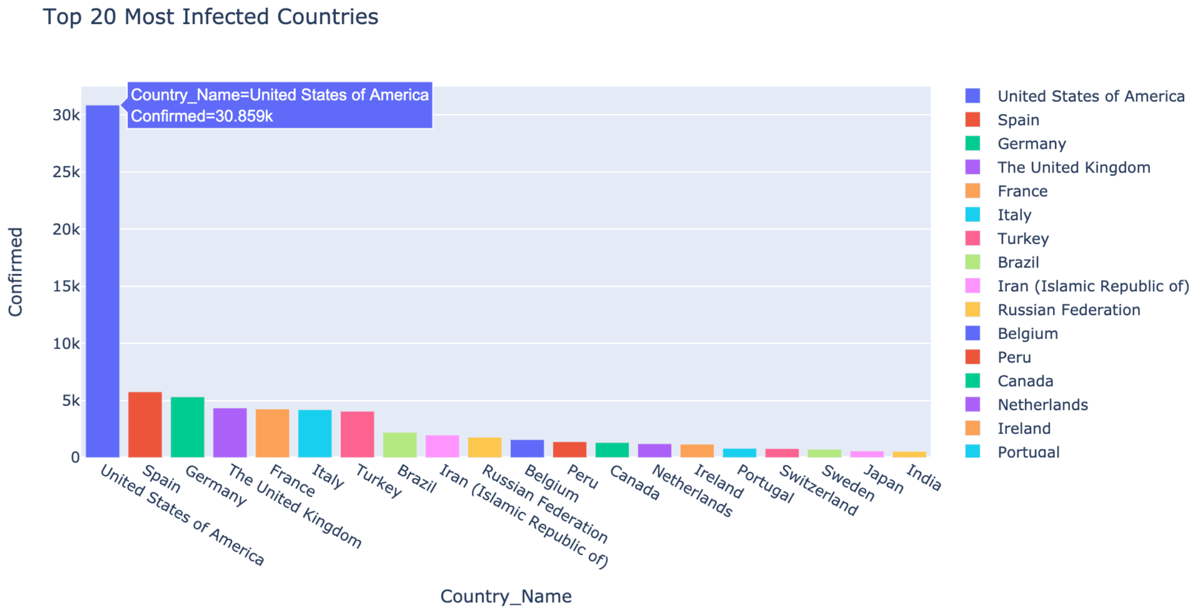


Figure 1.2

Figure 1.1 shows the number of most confirmed cases on 2020-04-10 sorted by the Confirmed Cases from left to right and it can be seen that United States of America had the greatest number of confirmed cases i.e. 30859 on 2020-04-10. The confirmed cases also follow the same pattern as of deaths i.e. as we move backward in time, we see that United States of America had relatively less cases as compared to other countries.

As for the recovered cases the case was different because the recovered cases had to be inferred from the data provided. Since subtracting the Deaths from Confirmed cases does not mean that the remaining patients have recovered it just means they are sick and if we look for any missing people by adding the cases for previous days to the cumulative cases to find out whether any patients are missing it would still not mean that the missing patients have recovered so we adopted a probabilistic approach i.e. if we calculate the death rate by $\frac{Deaths}{Confirmed} \times 100$ We get the fatality rate i.e. out total infected what percentage of patients have a fatal outcome so we assume that the remaining i.e.

$$1 - Death\ Rate = Recovery\ Rate$$

Gives us the recovery rate. Based on this assumption we multiplied the recovery rate with the total number of confirmed cases to get the Recovered cases or those that will eventually recover this method was also adopted by **[1]** and got the following results:

Country Name	Deaths	Confirmed	Death_Rate	Recovery_Rate	Recovered
United States of America	1925	30859	0.062381	0.937619	28934
Spain	683	5756	0.118659	0.881341	5073
Germany	266	5323	0.049972	0.950028	5057
Turkey	96	4056	0.023669	0.976331	3960
Italy	612	4204	0.145576	0.854424	3592

Table 1.1

Table 1.1 contains the first five records out of the total of 20 records that shows how the recovered were calculated and the figure below shows the visual representation of the results.

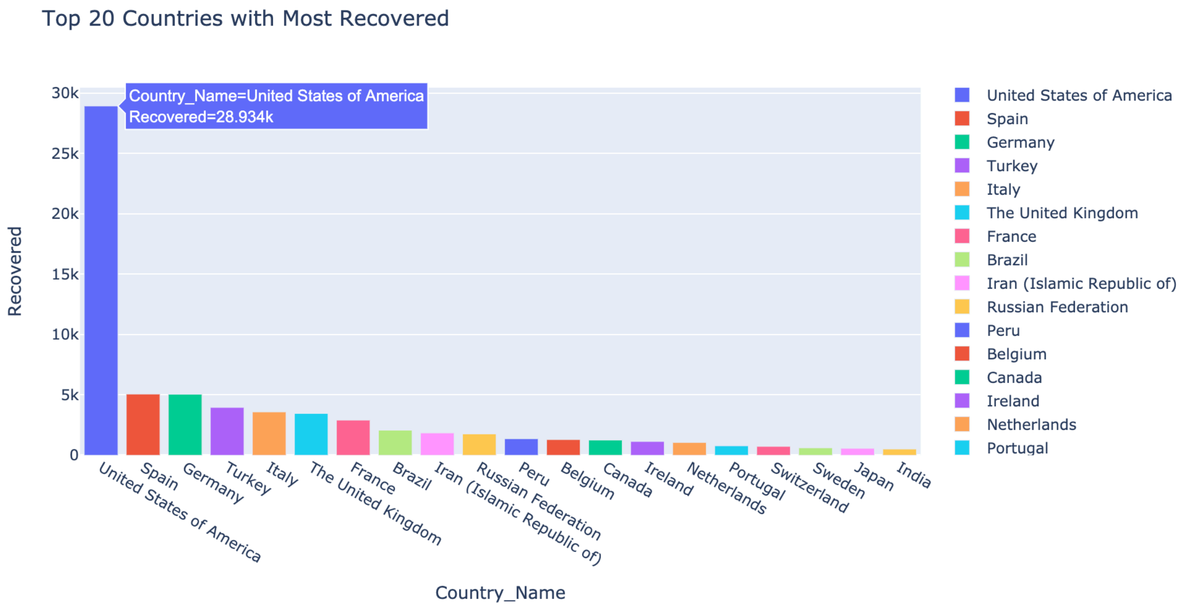


Figure 1.3

Figure 1.3 complements Figure 1.1 and Figure 1.2 i.e. there is an obvious trend here that shows that the country with the highest number of infections has the highest number of recovered and consequently the highest number of deaths as well.

1.2 In our second task we found the countries with the highest number of new cases and highest number of deaths between two given dates. In order to find the countries with highest cases and deaths, we first separated out all the data for the given inputs i.e. the starting date and the ending date then further filtered out deaths for each country between those dates and using ‘groupby’ along with ‘sum()’ and ‘sort()’ identified the countries with the highest deaths between the two dates and got the following results:

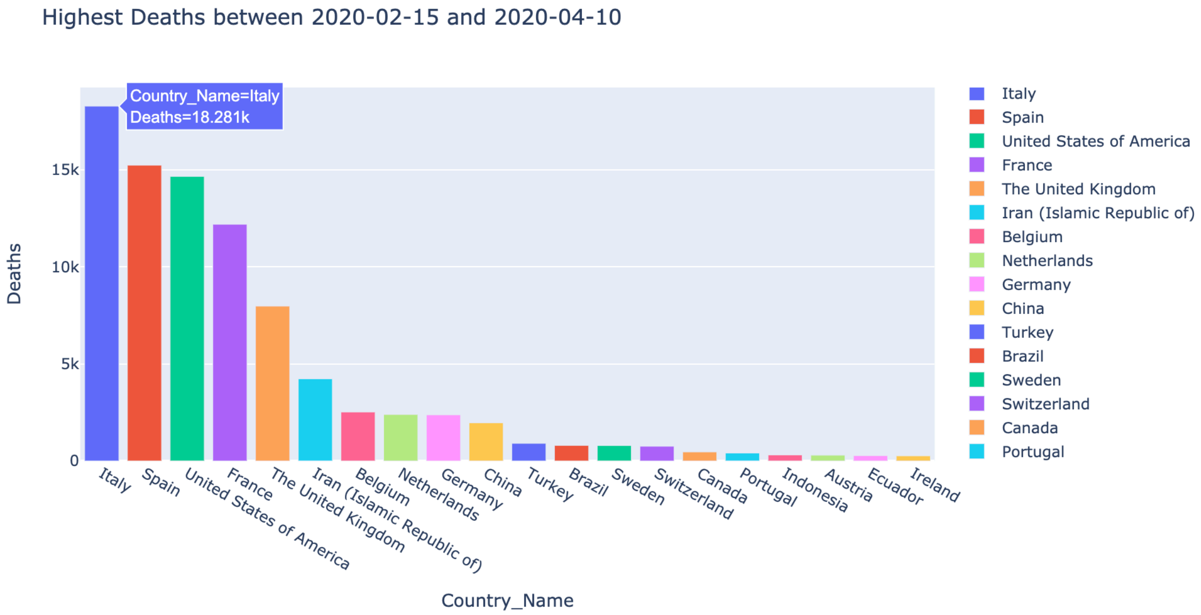


Figure 1.2.1

Figure 1.2.1 shows that Italy had the highest number of deaths between 2020-02-15 and 2020-04-10 and Spain had the second highest number i.e. 15238 it is important to note that these numbers represent the combined deaths that were reported between these two days and does not reflect the current situation on the ending date. for example, a country could have faced the highest death toll in the prior weeks but saw recent decline in deaths.

The same approach was followed for new cases and gave the following results:

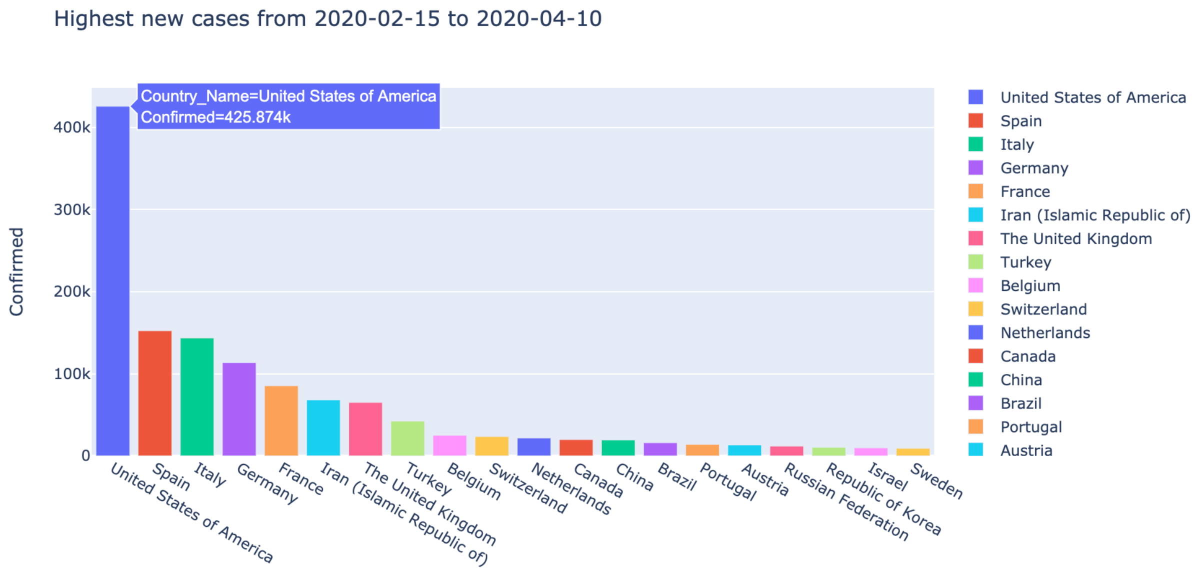


Figure 1.2.2

Figure 1.2.2 shows some interesting results. The country with the highest number of deaths does not have the highest number of new cases between the same dates which is Italy in this case (having the highest number of deaths as shown in Figure 1.2.1). Whereas United States of America despite having the highest number of cases has relatively fewer deaths than Italy. This could be either due to the medical systems these countries have in place or it could be due to the ratio of older population between these two countries.

1.3 In our third task we had to identify the starting and the ending days of the longest spread for a given country which is defined as the period where the new cases tend to increase and can also contains those days where the new cases were relatively low or none at all. To accomplish this task following approach was used:

We first identified all those consecutive days where the new cases were increasing and it also included those days where the new cases were relatively low or none at all. These consecutive spread periods were grouped by according to their timeline, the group with the most consecutive days is selected and the final elements are output. To clarify this further we show the longest spread period for Italy having the following original curve:

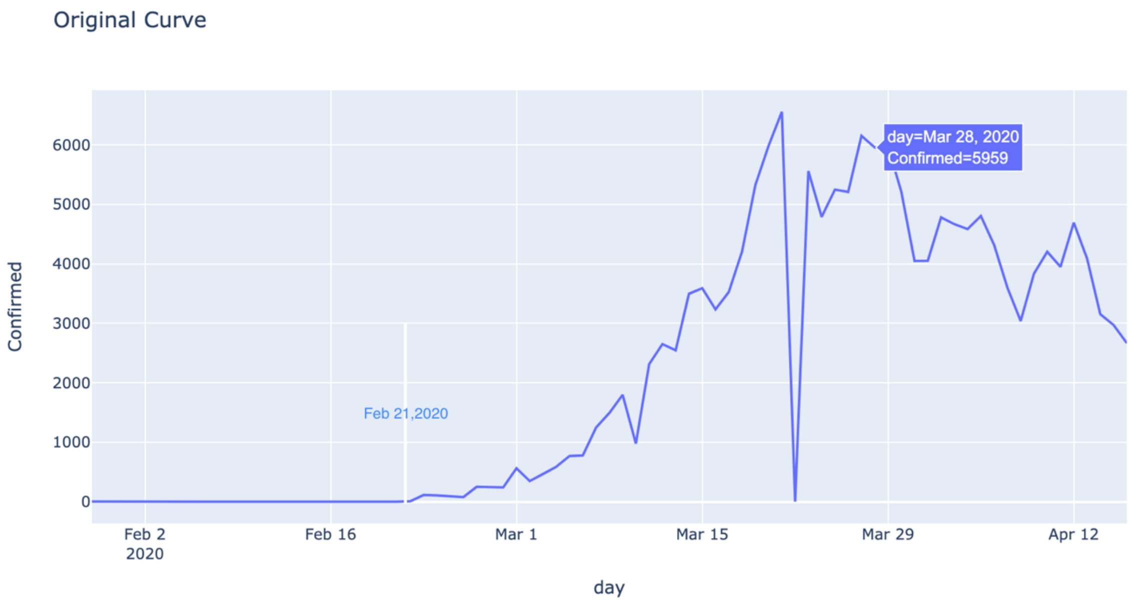


Figure 1.3.1

Figure 1.3.1 shows that the new cases started to increase on 21st Feb with 1 case and continuously increased until 28th Mar 2020 having 22nd March where there were no cases and 23rd March again saw a spike with 5560 cases which again went down the next day and then went back up. After 28th March there was a continuous decline in new cases until 9th of April where they started increasing again. This shows us that the longest spread period was from 21st Feb to 28th Feb

[

1, 7, 113, 105, 93, 78, 250, 238, 240, 561, 347, 466, 587, 769, 778, 1247, 1492, 1797, 977, 2313, 2651, 2547, 3497, 3590, 3233, 3526, 4207, 5322, 5986, 6557, 0, 5560, 4789, 5249, 5210, 6153, 5959

]

and is elapsing 36 days starting from 22nd Feb day 2(7) with the following curve:

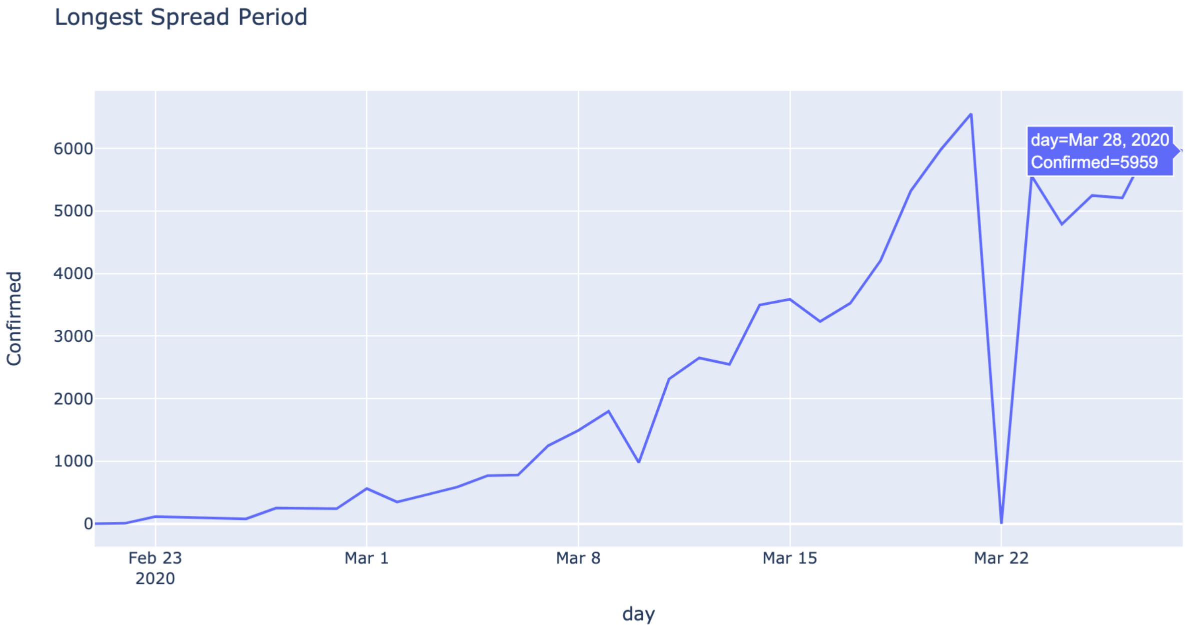


Figure 1.3.2

Figure 1.3.2 shows the identified longest spread period. This curve can be compared with the curve in Figure 1.3.1 (the original curve out of which we had to identify the longest spread period).

And the elements are:
[
1, 7, 113, 250, 561, 587, 769, 778, 1247, 1492, 1797, 2313, 2651, 3497, 3590, 4207, 5322, 5986, 6557
]
It can be seen that the final elements array ends at the highest point in that curve.

1.4 In the fourth task we had to use a prediction method to chart the progression of COVID-19 cases in the top 5 countries for the next 7 and 30 days.
The top 5 countries with most COVID-19 cases were identified to be:

Country Name	COVID-19 Cases
United States of America	604070
Spain	177633
Italy	165155
Germany	130450
France	105155

Table 1.4.1

The prediction method chosen for the task was **"FbProphet"** which works well for forecasting time series data.

It is an additive regression model having main components:

- It can model the data using piecewise linear or logistic growth curve trend i.e. how fast the time series is growing or declining.
- It automatically detects changes in trends by selecting the change points from within the data itself. The change points are those points where the data changes its trajectory.
- It has a yearly seasonal component that is modeled using Fourier series which can be further decomposed into Monthly, Weekly, Daily, and hourly basis in our case we set the seasonality to monthly due to lack of year seasonality in our data. Seasonality means something that happens in cycles over time.
- It implements weekly seasonal component using dummy variables where it stores the local trends i.e. weekly trends that are useful to model the effects of weekdays and weekends.
- It also provides user with the option to input the holidays for a given country that could be extremely useful for modeling special events (events are those days where the spike in trend occurs on yearly basis) but we won't be using this feature due to lockdown and special circumstances. This feature would be more useful when the lockdown is lifted and the life goes back to normal.

The reason for choosing this model includes the following:

- The ability to address the outliers that can have severe consequences if left unaddressed.
- Easily interpretable results.
- Robust Model Evaluation
- Automatically Interpolates the missing data

Since every country has its own dynamics, environmental variability, progression stage, medical systems in place, and political landscape for the lockdown policy, we trained a separate model for each of these top 5 countries.

The results for United States of America are shown below:

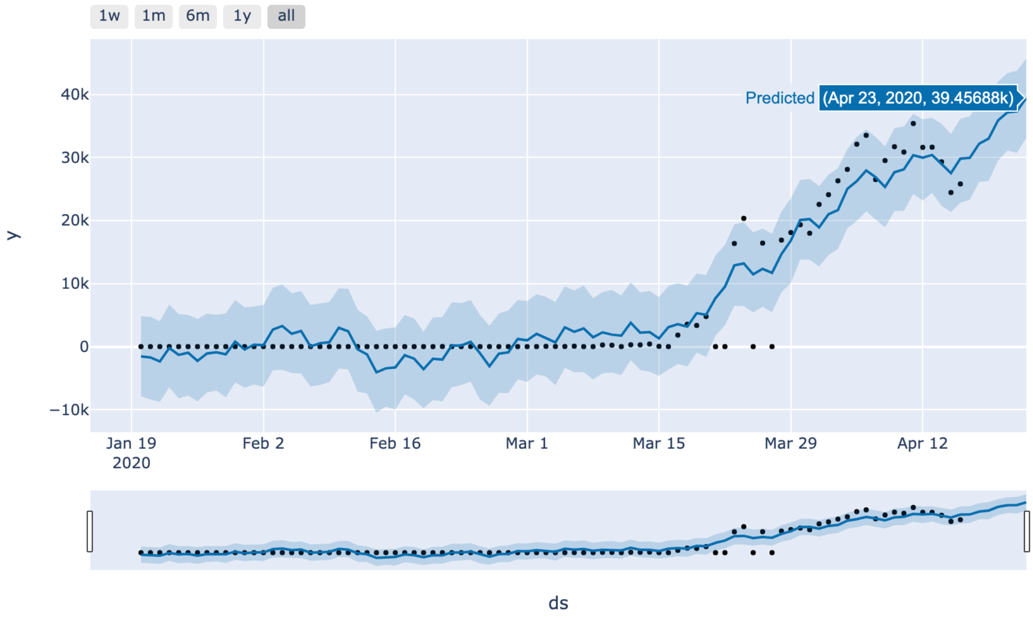


Figure 1.4.1
United States 7 days forecast

Figure1.4.1 charts the progression of United States of America. Y-axis represent the number of cases and ds on X-axis is representing the dates. The data for US was available from 20th Jan 2020 until 16th April 2020 and an additional forecast of 7 days is shown in the figure.

The black dots represent the actual datapoints that were available to us and the blue line represents the curve that was fitted to the data while the light blue shaded regions around the line represents the uncertainty interval. The greater the variability in the data the greater this shaded region is. The model is predicting 39,456 cases on 23rd Apr 2020.

The chart for next 30 days progression for United States is shown in the figure below:

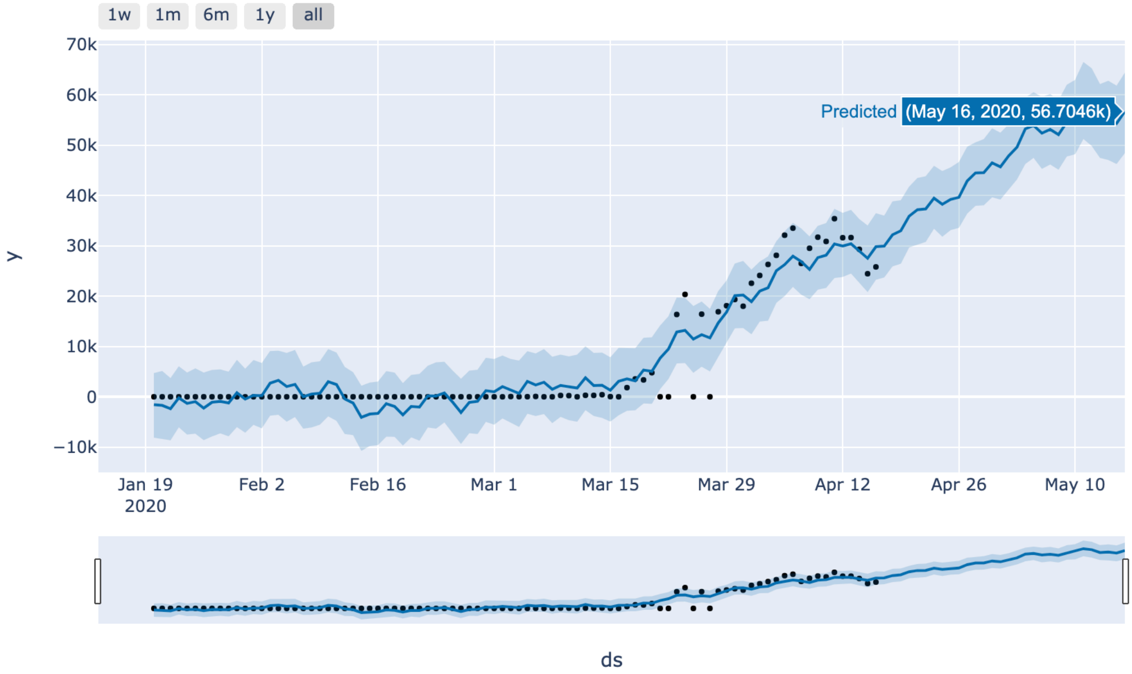


Figure 1.4.2
United States 30 days forecast

Figure 1.4.2 shows the progression of COVID-19 for the next 30 days. The model is predicting 56,704 cases on 16th May 2020.

1.4.1 Trend Analysis of United States (Using the Model):

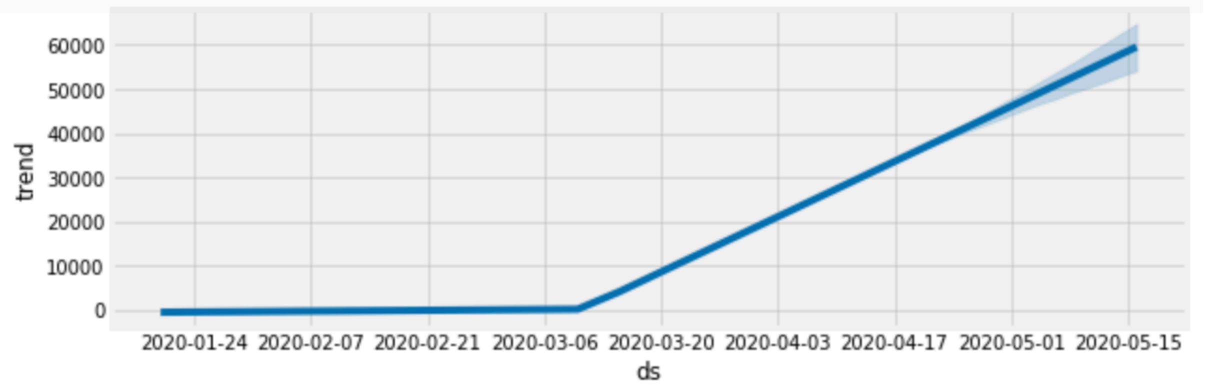


Figure 1.4.3

Figure 1.4.3 show the overall trend of progression in United Stated. The Y-axis represents the number of cases, ds on the X-axis represents the time line. The light blue region towards the end is representing the uncertainty region. It can be observed that the most spike in cases occurred in March and continued to increase towards the end.

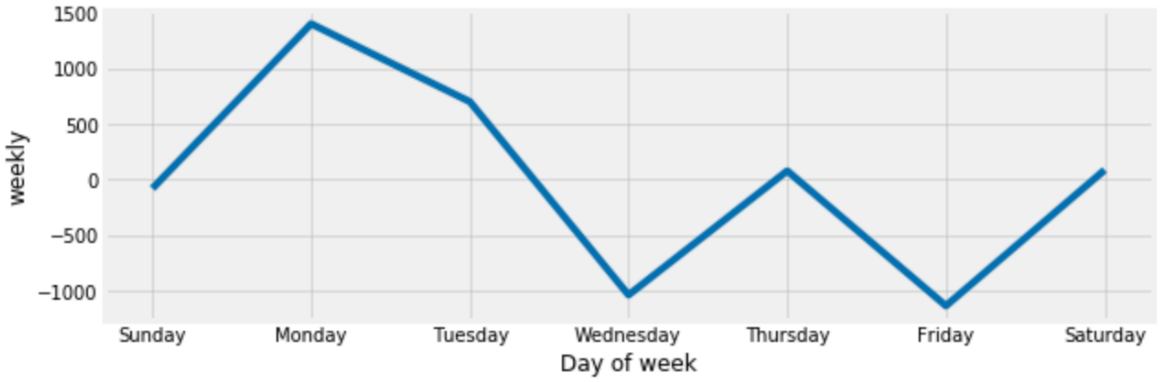


Figure 1.4.4

Figure 1.4.4 represents the number of cases reported on the weekly basis. The trend shows a sudden dip in the cases on Wednesday where no cases are being reported. There were several points in the original data where the number of cases were zero. This could be due to the way the data is entered.

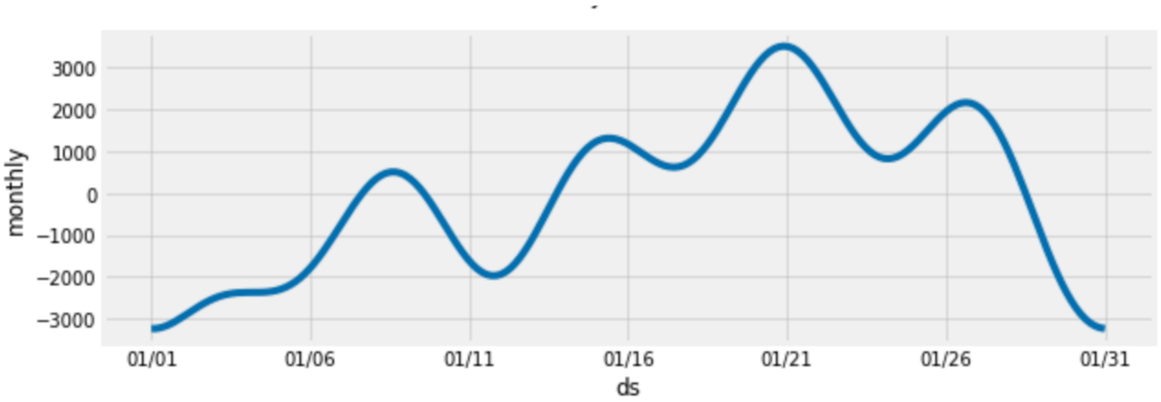


Figure 1.4.5

Figure 1.4.5 shows the monthly trend that our model picked up. It can be seen that the cases start to rise at the start of the month and begins to go down towards the end. Again, this behavior could be due to the way the cases are reported or the data is entered.

1.4.2 United States Model Evaluation:

While training the model the width of uncertainty interval was set to 0.95 using the interval_width parameter of the model so the overall model evaluation gets effected by this interval width. For model evaluation we used cross_validation and performance_metrics libraries from fbprophet.diagnostics some of the results are shown in the table below:

	ds	yhat	yhat_lower	yhat_upper	y	cutoff
1	2020-04-11	29475.757	23986.587	35821.050	35386	2020-04-10
2	2020-04-12	29663.716	23403.874	35864.119	31606	2020-04-10
3	2020-04-13	31435.527	25387.051	37585.603	31633	2020-04-10
4	2020-04-12	33116.046	27300.736	38925.635	31606	2020-04-11
5	2020-04-13	34920.030	29421.599	40598.380	31633	2020-04-11

Table 1.4.2

Table 1.4.2 shows the predicted values after two cutoff dates i.e. 2020-04-10 and 2020-04-11. The cutoff date is the point after which the model made the predictions. For example, the first row in table 1.4.2, the cutoff date is 2020-04-10 and ds i.e. the date for which the prediction is made is 2020-04-11 which is one day into the future.

yhat_lower and yhat_upper represents the light blue shaded region boundary (Figure 1.4.2) i.e. the uncertainty interval. The time line selected for cross validation is where the variability in data is maximum for United States i.e. from 2020-04-05 to 2020-04-13 (this can be verified from the original curve shown in Figure 1.4.6).

It can be seen from the above table that yhat_upper for first entry is 35,821 and the actual value is 35,386 which is quite close but as we move further in time, the difference between the predicted values and the actual values starts to increase.

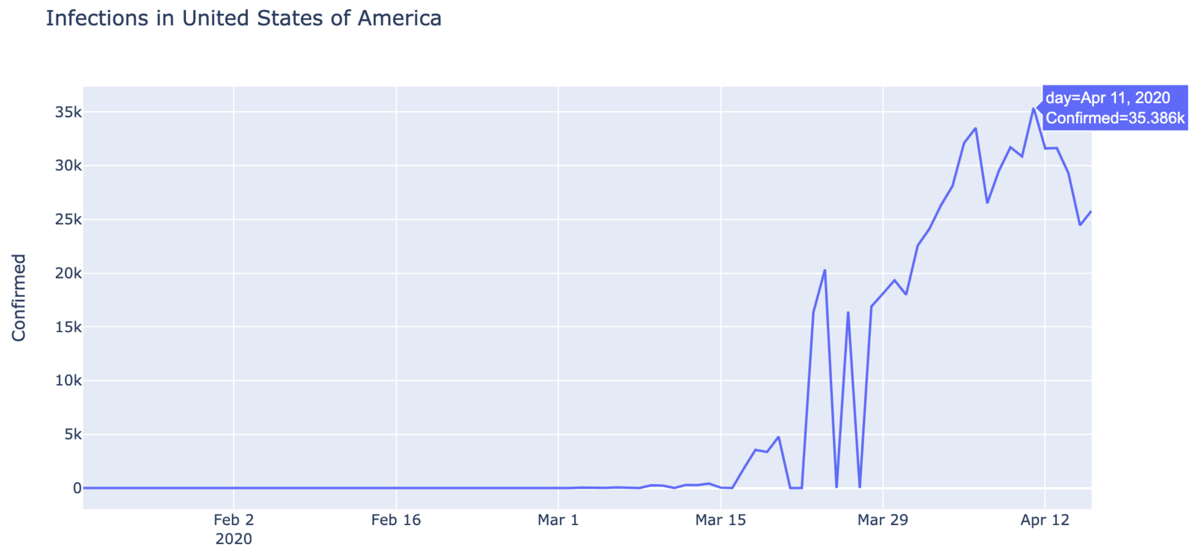
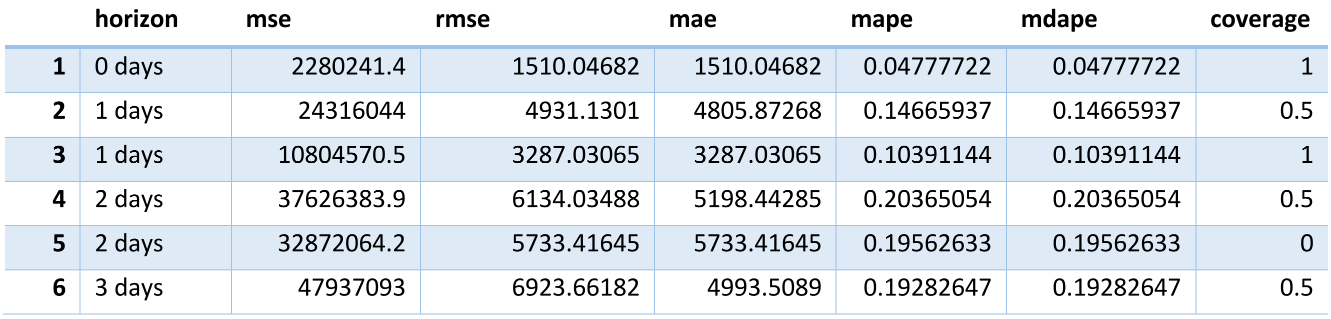


Figure 1.4.6

Figure 1.4.6 is showing the actual data/curve against which, the predictions were made.

1.4.3 United States Model Error:



	horizon	mse	rmse	mae	mape	mdape	coverage
1	0 days	2280241.4	1510.04682	1510.04682	0.04777722	0.04777722	1
2	1 days	24316044	4931.1301	4805.87268	0.14665937	0.14665937	0.5
3	1 days	10804570.5	3287.03065	3287.03065	0.10391144	0.10391144	1
4	2 days	37626383.9	6134.03488	5198.44285	0.20365054	0.20365054	0.5
5	2 days	32872064.2	5733.41645	5733.41645	0.19562633	0.19562633	0
6	3 days	47937093	6923.66182	4993.5089	0.19282647	0.19282647	0.5

Table 1.4.3

Table 1.4.3 contains various error metrics for the United States model these metrics provide useful insight into our prediction performance. These metrics include Root Mean Square Error (RMSE) and Mean Absolute Percent Error (MAPE). Horizon represents the number of days into the future and is calculated by $horizon = date - cutoff\ date$. The coverage column represents the coverage of yhat_lower and yhat_upper estimates.

As for error terms perhaps, the simplest metric is MAPE which is expressing the accuracy as a percentage of error. So, for example a MAPE value of 0.04 would mean that on average our forecast is off by 4% and as the Horizon is increased so is the error.

It can be observed that as we move further into the future dates the errors start to increase. When horizon is 3 days the MAPE is 19% i.e. our forecast is off by 19 percent. This is due to the volatility in the data i.e. the sudden spikes and dips towards the end. The Plot for MAPE and RMSE terms is shown in the figures below.

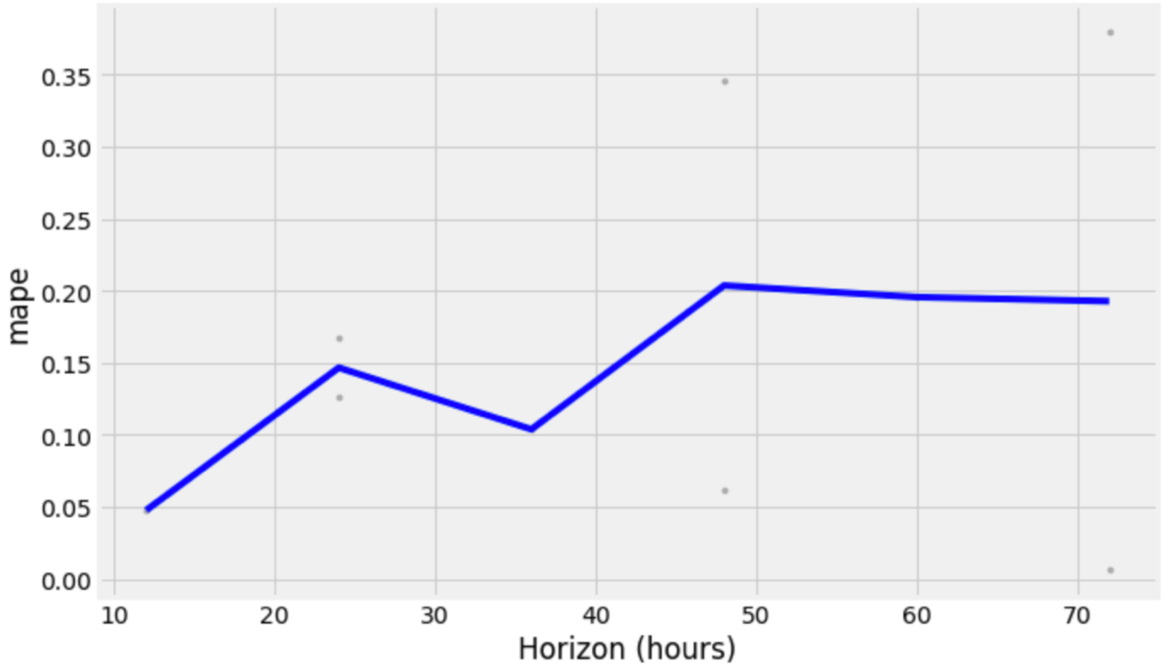


Figure 1.4.7

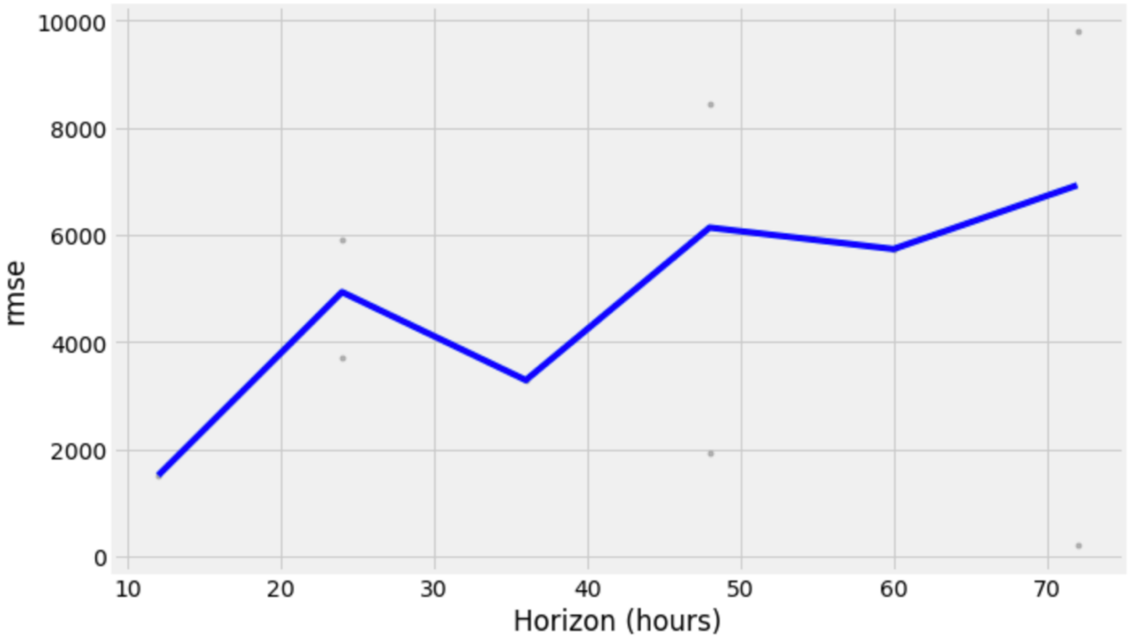


Figure 1.4.8

Figure 1.4.7 and Figure 1.4.8 shows the MAPE and RMSE error respectively. The black dots are showing the absolute percent error and root mean square error for each prediction respectively.

It can be observed that with increasing horizon these error terms increase as well. The label Horizon(hours) is actually representing the days in hours. This is still an open issue in **FbProphet** and will be fixed in the next version.

This concludes the United State model. The Models for the remaining four countries i.e. Spain, Italy, Germany, and France are processed in the same fashion and can be found in code file provided with this report. It contains the same level of details for each country.

1.5 Detailed Exploratory Analysis:

Most Infected Country:

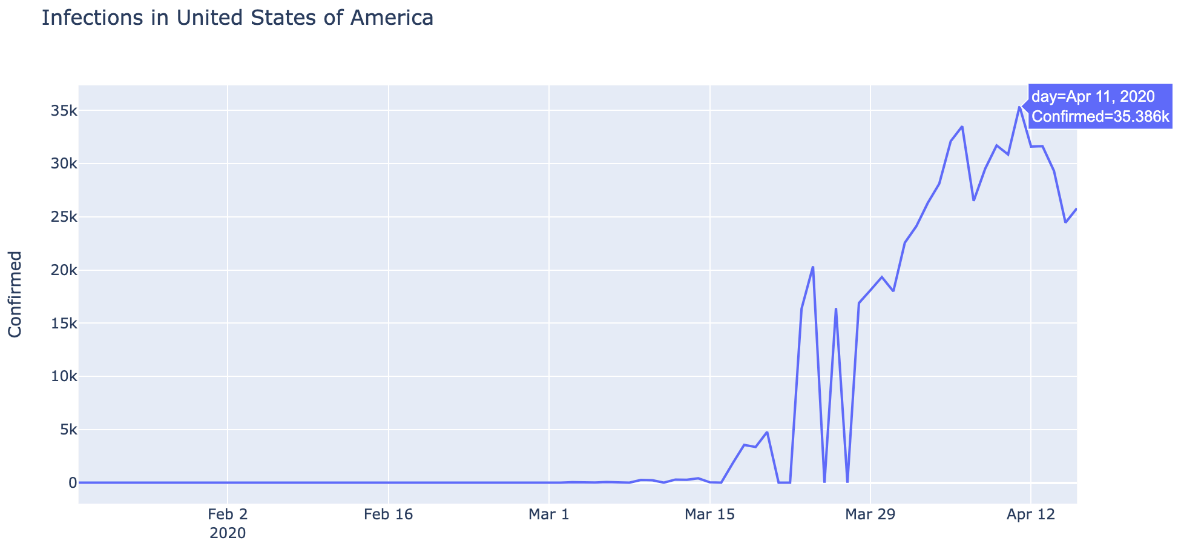


Figure 1.5.1

Figure 1.5.1 shows the country with the greatest number of cases. The United States of America had 60,4070 until 16th April 2020 with 25,871 death and recorded its highest number on 11th April 2020 where 35,386 confirmed cases were reported in a single day.

Identifying the trend in Figure 1.5.1 using first degree polynomial:

We tried to identify the trend by fitting a 1st degree polynomial to the United States data using NumPy's polyfit function and output the slope which came out to be **351.5** this shows that the trend is increasing with 351.5 as its magnitude. The code for this can be found in the Jupyter Notebook provided with this report under the Exploratory Data Analysis section.

Worldwide Confirmed Cases and Deaths (Summary):

	Confirmed	Deaths
mean	19959.83	1310.37
std	29406.78214	2194.545597
min	0	0
25%	1038	43
50%	2782	107
75%	30358.25	1625.75
max	90778	7911

Table 1.5.1

Table 1.5.1 shows different stats for Infections and Deaths in all the affected countries **mean** represents mean of Confirmed cases and deaths all over the world in a single day. **25%, 50%, 75%** are the percentiles. **max** is the maximum number of Confirmed cases and Deaths reported across the world in a single day.

Worldwide Confirmed Cases and Deaths (Plot):

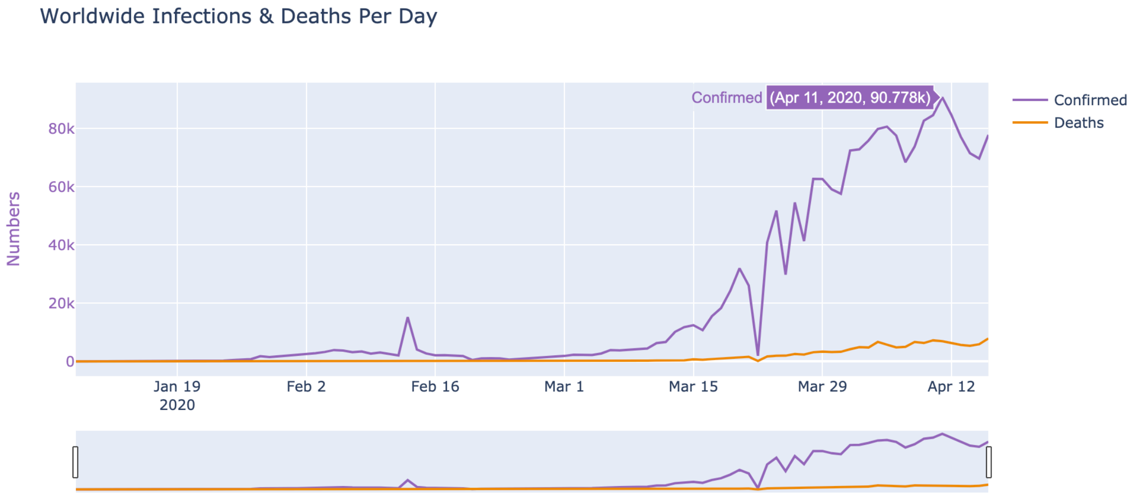


Figure 1.5.2

As seen in the Figure 1.5.2 world saw its maximum number of cases i.e. 90,778 in a single day on 11th April 2020. And the number of Deaths on that day were 6,946. This can be seen in the figure below

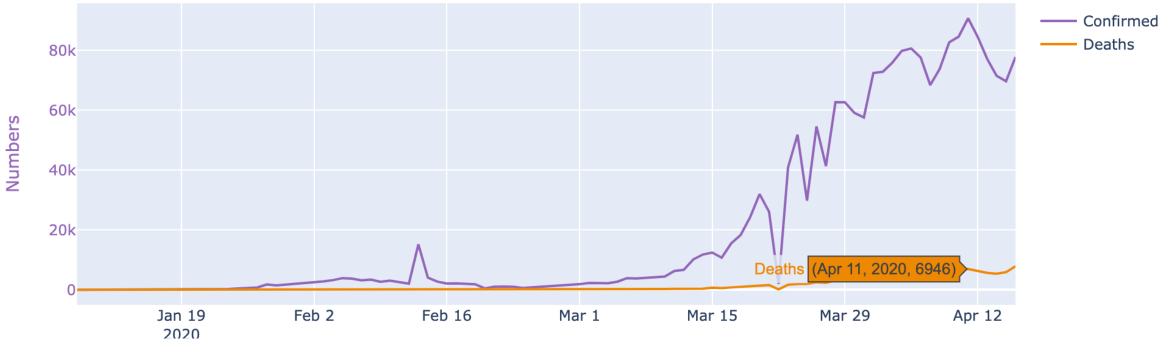


Figure 1.5.3

Infection Fatality Rate in Top 5 Countries:

Country Name	Deaths	Confirmed	Death Rate
United States of America	25871	604070	4.28
Spain	18579	177633	10.45
Italy	21647	165155	13.10
Germany	3569	130450	2.73
France	17146	105155	16.30

Table 1.5.2

Table 1.5.2 reveals an interesting fact. The country with the highest number of cases i.e. United States of America has far lesser death rate as compared other European countries that are having 3 times fewer confirmed cases.

Elapsed number of days until the top 5 countries reported their highest number of cases in a single day:

Country	Days Elapsed	Cases Reported	Day
United States of America	82	35386	2020-04-11
Spain	61	9222	2020-04-01
Italy	52	6557	2020-03-21
Germany	52	7324	2020-03-20
France	68	7500	2020-04-01

Table 1.5.2

Table 1.5.2 shows the number of days that passed before the highest number of the cases were reported in a single day. It can be seen that in United States of America the highest number of cases were reported on 11th April 2020 which is 35,386 after which they start declining.

The average number of days for these countries becomes 63 which means that on average these countries reported their highest number on the 63rd day since the pandemic started. And the cases started decreasing after that. This stat only holds until mid of April since the data provided to us is only until 16th April 2020.

Worldwide Cumulative Confirmed Cases (Plot):

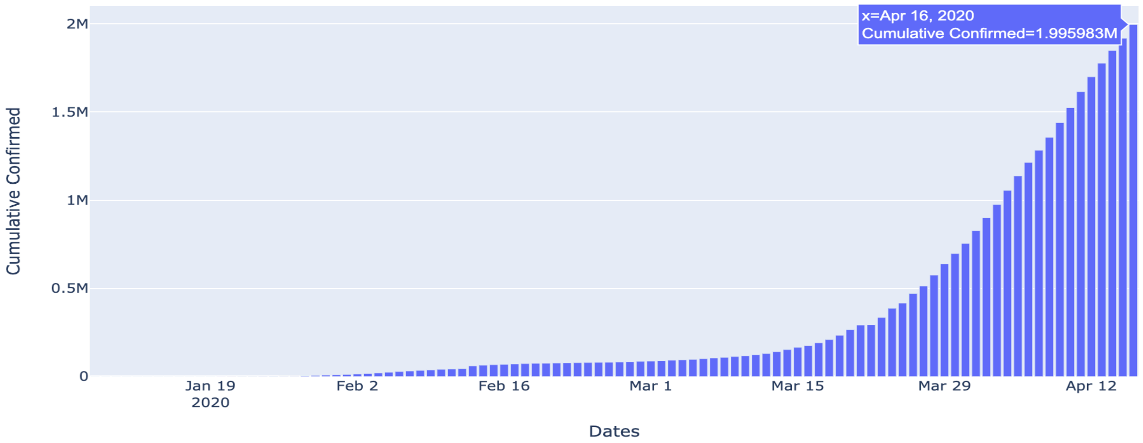


Figure 1.5.4

Figure 1.5.4 shows the cumulative confirmed cases across the world. The cumulative confirmed means that the cases of the preceding days have been added to the next day.

Worldwide Cumulative Deaths (Plot):

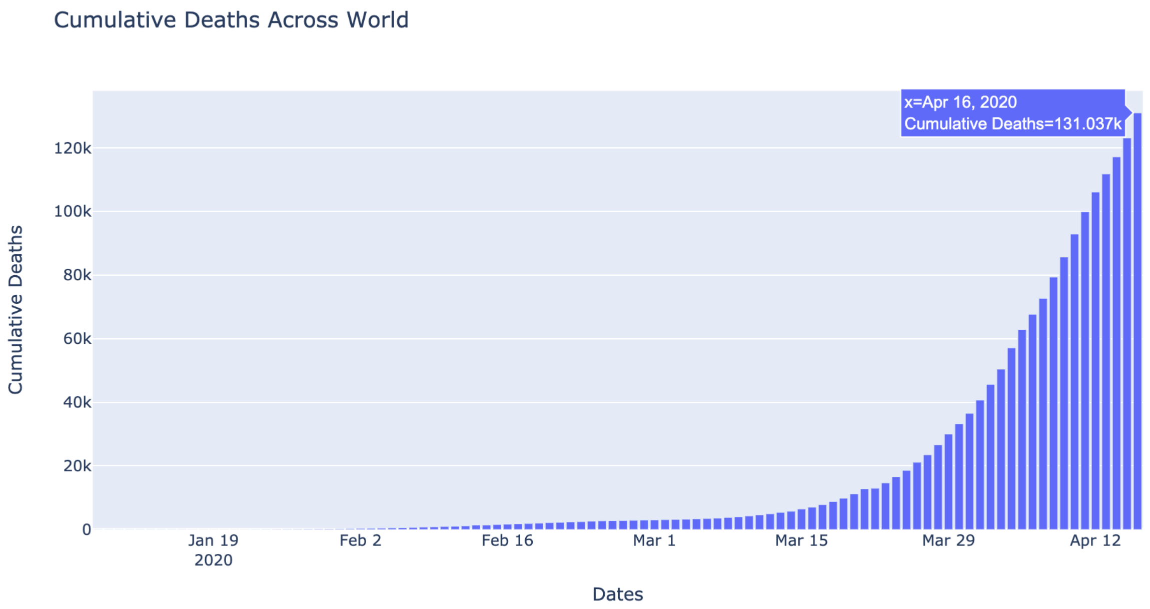


Figure 1.5.5

Figure 1.5.5 shows the cumulative deaths across the world which stood at 131,037 as of 16th April 2020

Interactive plot for Confirmed Cases:

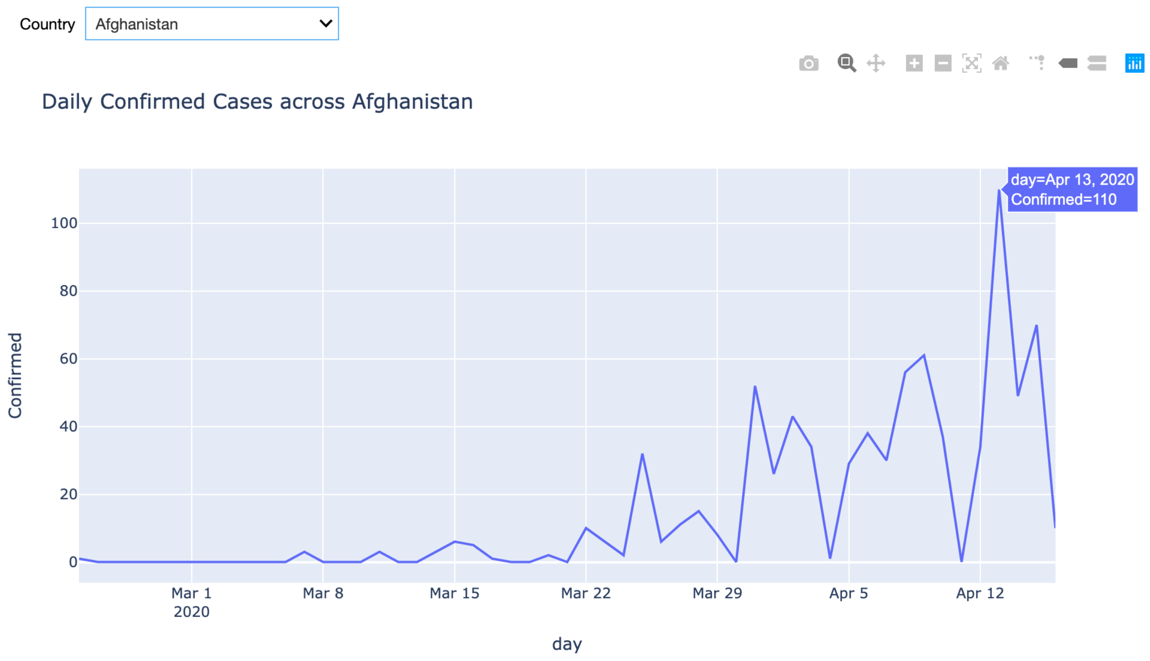
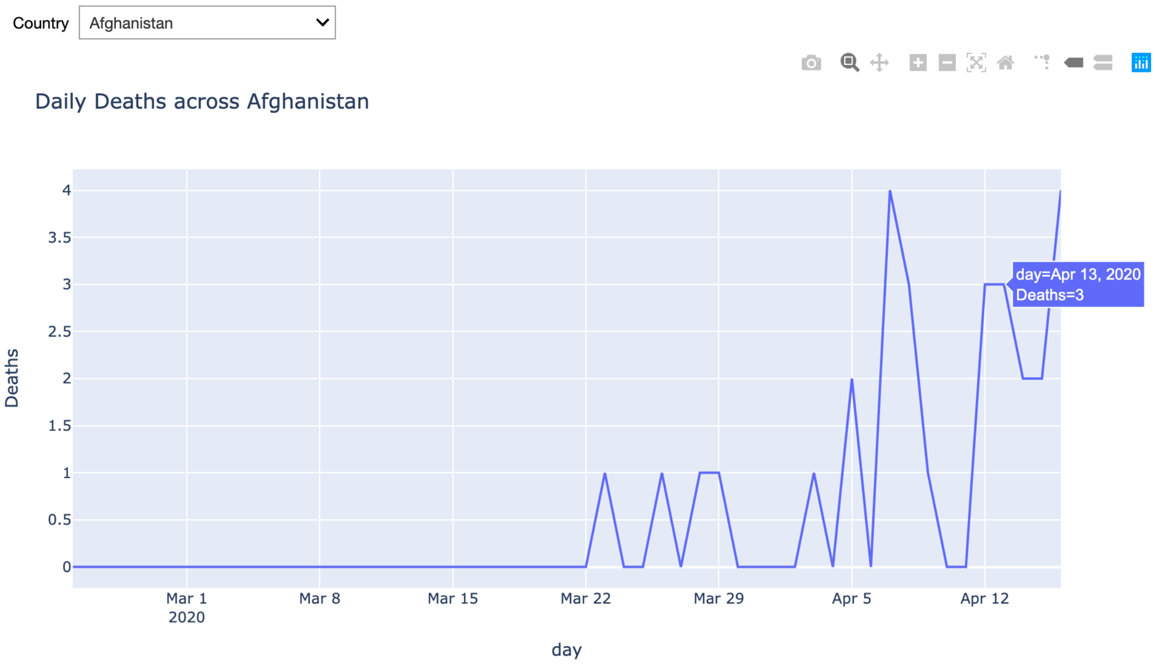


Figure 1.5.6

Figure 1.5.6 shows the interactive plot across all the countries. Using interactive and widget libraries for plotly we provided a panel with dropdown to explore the confirmed cases across all the countries. This is available in the Exploratory Data Analysis section of the code file provided with this report.

Interactive plot for Fatalities:



Country Afghanistan

Figure 1.5.7

Figure 1.5.7 is representing the interactive panel created for fatalities across different countries

High Variability Model (Spain):

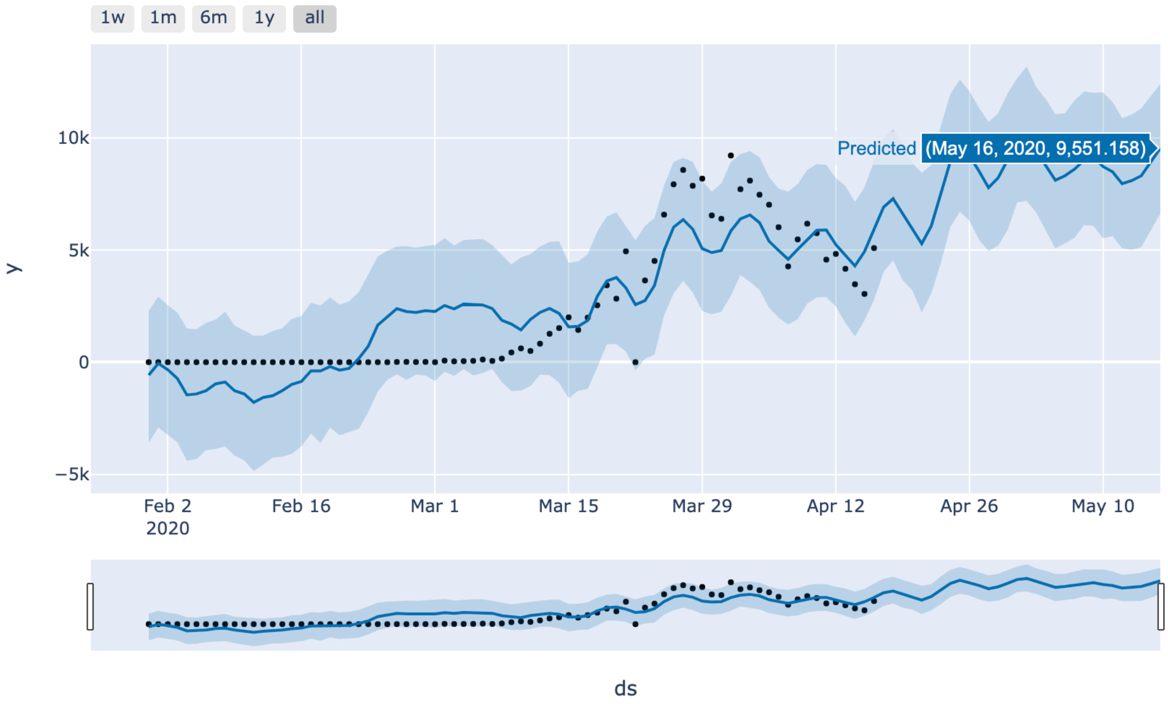


Figure 1.5.8

Figure 1.5.8 shows the progression chart for next 30 days in Spain. An important point to note here is the amount of variability that is present in the data. It can be seen that from 15th March to 20th March, the data is significantly changing its local trajectory.

As a result, the uncertainty of our model is increased (shown with the light blue shaded region) and consequently this had an impact on the accuracy of our predictions as well.

That is, the more uncertain our model is the more our predictions would be off. This is one of the limitations that our data projects. To overcome this, we could train the model again by either scaling our data or by applying exponential smoothening.

2. Some Observations and Limitations:

- The data provided is relatively old (by a month and a half) which means that the analysis performed might not reflect the current situation.
- The models trained for the top 5 countries were trained using monthly seasonality since the data for a whole year is not yet available.
- The models trained will only predict an upward trend even for one year into the future whereas in reality the pandemic might not even be there after one year. This is due to the fact that the data seen by the models so far had only upward trajectory.
- The amount of data used to train the models was less, with high amount of variability. This greatly affects the model performance (it might be accurate up to few days only, in some cases).
- The dips and spikes in the data could be due to the way authorities are collecting the data these sudden spikes and dips increases the uncertainty in the model.
- Finally, the presence of an effective treatment can also influence the predictions.

3. Future Directions:

- Creating a responsive application or a website that continuously takes in the up to date data from the source and displays most recent stats. The code provided with this report is easily convertible to a python flask application with user input points clearly defined.
- The models performance would greatly improve from continuous data integration and the predictions would become more and more accurate over time.

References:

[1] https://www.worldometers.info/coronavirus/coronavirus-death-rate/