Machine Learning Major project report

COVID-19 Forecasting of various Countries in Europe

Shubham Kumar: 2017B4A30712G

Jasdeep Singh: 2017B4A30748G

Praffulla Tripathi: 2017A7PS0933G

Upayan Das: 2017A8PS0617G

30.04.2021

ABSTRACT

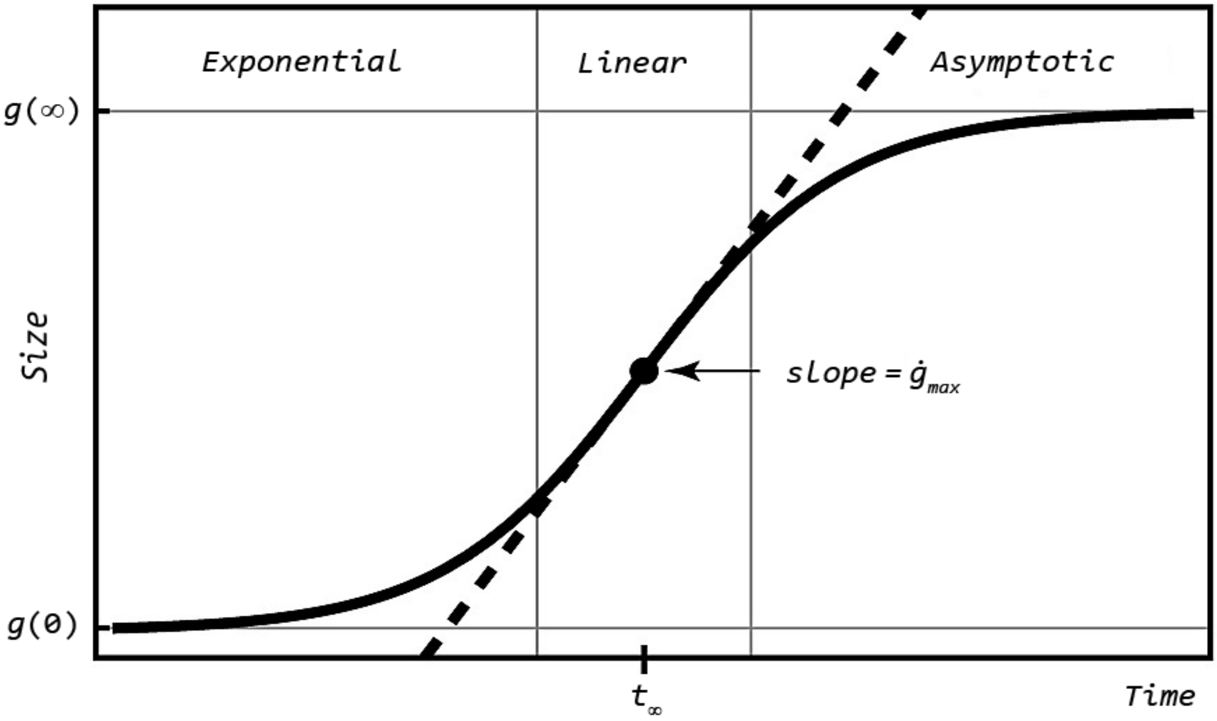
COVID-19 outbreak was first reported in Wuhan, China and then has spread to more than 180 countries. Cases were rising rapidly across the countries, endangering the health of the public and affecting the health management system in that country. So, to keep a track on the cases a platform CoronaTracker was made that provides the latest and best information available as well as statistics and analysis on COVID-19. This report aims to predict and forecast the COVID-19 cases, deaths, recoveries and vaccination available in some countries of Europe such as Denmark, Italy, Austria and Germany by using selected models such as Support vector regression, polynomial regression and used some time forecasting techniques such as Holt's linear model and fbProphet model.

1. INTRODUCTION

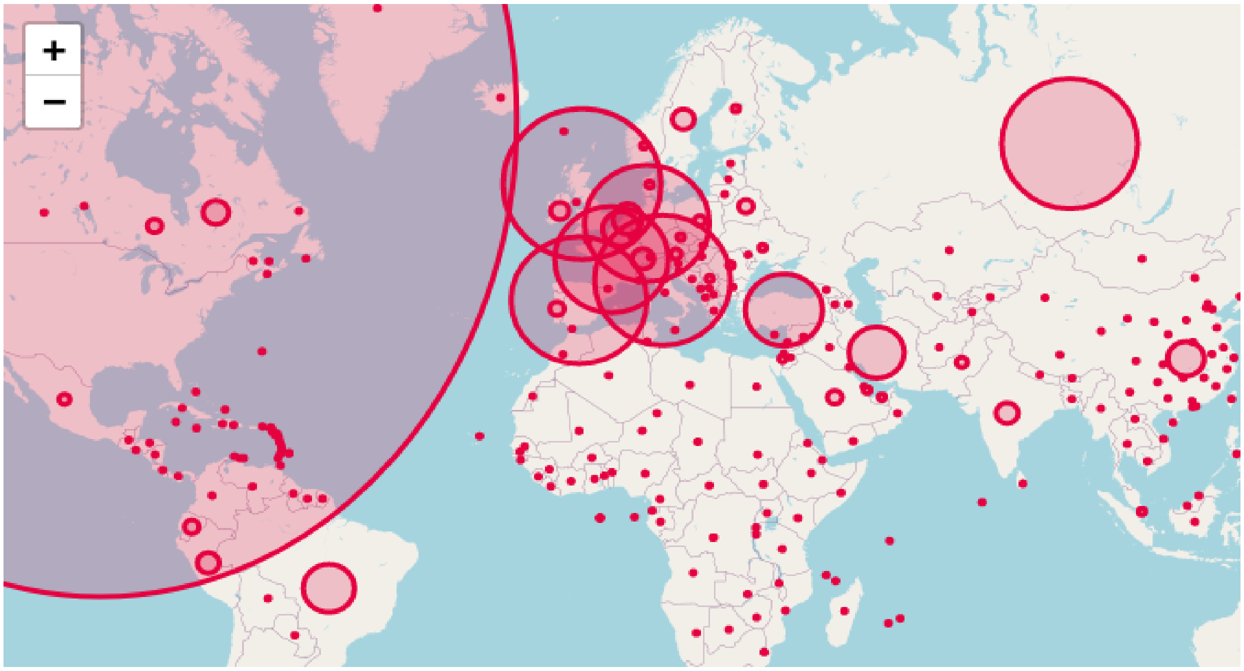
Covid 19 is a highly contagious disease caused by SARS-CoV 2. The first case was identified in 2019 in Wuhan, China and then it spread world wide. The symptoms of Covid 19 are variable and not easy to distinguish and at least one-third of the people who are infected do not have any noticeable symptoms. Transmission of the virus occurs when an infected person is in close contact with another person.

On average, one covid infected person infects more than one person (1.5 to be precise). The spread of the virus can be theorized as the infamous lotus in a pond puzzle which states that it will only take 1 day to fill the pond with lotuses if the reproduction cycle of one lotus is a day. These types of expansions can be modelled using an exponential function but in reality a more precise model is the sigmoid curve as the population gets immune to a disease which is also termed as herd immunity.

The below picture represents a theoretical idea of the total number of infected cases in a country. The slope of the curve first increases and then reaches a maximum which is the representation of the peak value of the new daily cases. After which the count of cases each day will start to decrease and the curve flattens out.



Covid 19 spread curve (https://www.csc2.ncsu.edu/faculty/healey/covid/help.html)[1]



Total infection in the world (https://www.kaggle.com/erikbruin/storytelling-covid-19)

The above figure shows a bubble graph of the total number of cases during May 2020 world wide. The highly infected continents were North America and Europe. In this paper we will try to explore and predict the growth, vaccination and death rates of four countries of Europe namely - Germany, Italy, Austria and Denmark. For most of the European countries the 2nd-3rd wave peaks around March 1, 2021.

In this paper we will extrapolate the data using 3 models namely - Polynomial regression model, Facebook's prophet model and Holt linear model for predicting the total cases, deaths and vaccinations.

In polynomial regression, the relation between independent variable, 'x' and dependent variable, 'y' is modelled using a polynomial of degree n. It is a non linear model to data but is linear in sense that E(y | x) is linear in the unknown parameters that are estimated from the data.

In Holt's regression, we use double exponential smoothing which is a very popular technique to forecast data with trends. The exponential smoothing gives more priority to the more recent data.

Facebook's Prophet is an open source model that is based on season trends in data. It is very useful in seasonal forecasting of data. Prophet is widely being used to forecast endemic and stock market trends.

2. METHODOLOGY

The whole process can be divided into the following stages:

1. Data collection
2. Data processing
3. Data visualization
4. Deciding evaluation factors and model fitting

2.1 Data Collection

Data is extracted from verifiable sources such as John Hopkins University [2], WHO and also from European Centre for Disease Prevention and Control [3]. These sites reported daily confirmed cases , daily confirmed deaths and daily confirmed recoveries for affected countries and regions. Data is updated on a daily basis so we remain up to date with the latest information. Along with it some vaccinations data has also been used.

2.2 Data Processing

The data for our selected countries Germany, Denmark, Austria, Italy was extracted from the Europe dataset [4] by matching the location. The main fields in focus are:

- Total covid cases
- Daily covid cases
- Total covid deaths
- Daily covid deaths
- Total covid recoveries
- Daily covid recoveries
- Total Vaccinations
- Daily Vaccinations
- Active Cases

For Recovery data, a different source of dataset was used.The Null values have been set to 0. The active cases was calculated using the formula:

Active cases = Total cases - Total Deaths - Total Recoveries

2.3 Data visualization

Libraries such as seaborn, matplotlib and plotly have been used to visualize the data in a clean manner. Plots such as Total cases vs Dates, Total deaths vs Dates and Total recoveries vs Dates have been our main focus to analyse the daily increase in cases, deaths or recoveries. Visualization for Countries such as Germany, Italy, Austria and Denmark has been made. Along with it, the vaccination trend has also been explored.

2.4 Deciding evaluation factors and model fitting

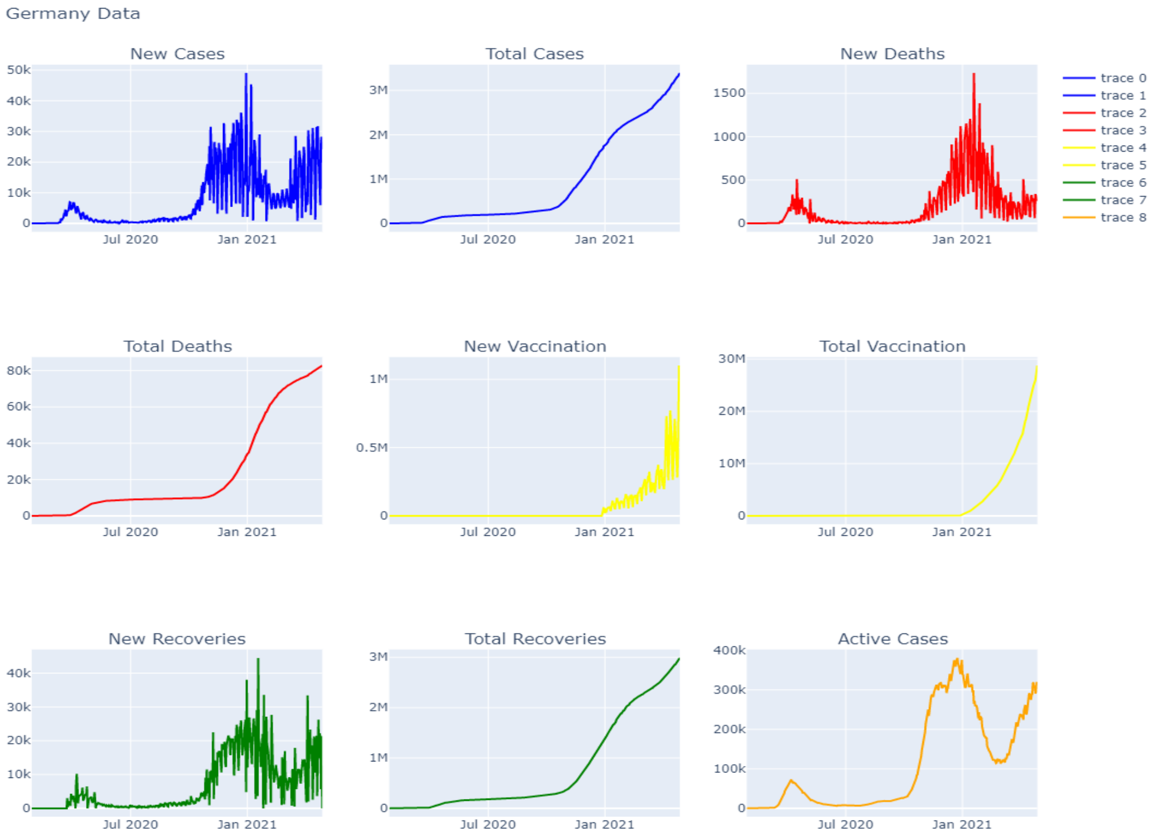
Models such as Support Vector Regression, Linear Regression, Polynomial Regression were first trained with the training data and rmse error was calculated for the testing data . Among these models Polynomial Regression was giving better results then the remaining ones. Some Time Series Forecasting models were also tried such as Holt's LInear model and fbProphet Model. Prophet is open source software released by Facebook's Core Data Science team which is used for forecasting the time series data based on an additive model with non-linear trends

3. Country Wise Analysis

3.1 Germany :

Germany has already suffered 2 waves of covid and the 3rd wave is currently likely in the initial stage.

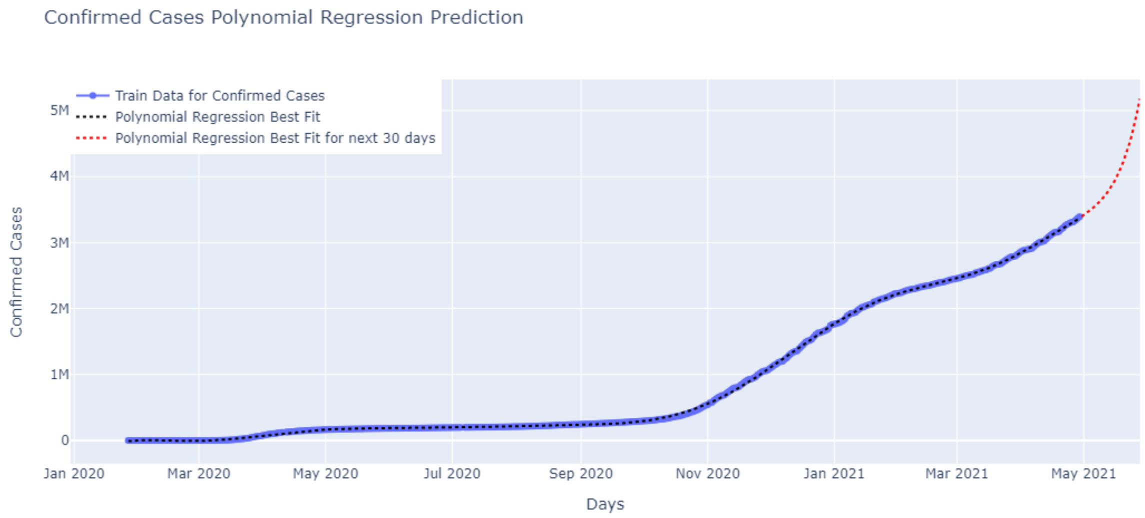
1. 1st wave : March- May 2020
2. 2nd wave : Nov-Dec 2020
3. 3rd wave : April - Present 2021



We can visualize the 3 waves of covid best from the last plot (Active cases) in the above picture. It can be seen from the graph of New Deaths that the 2nd wave had a much higher peak than the 1st wave. This is likely due to the government not taking the virus seriously in the initial days as a result they were not prepared for another wave. Another possible reason could be the different variants of the virus which might have been more deadlier.

The Vaccinations started slowly in January speeded up in recent times as seen from the exponential growth of the curve (Total vaccinations).

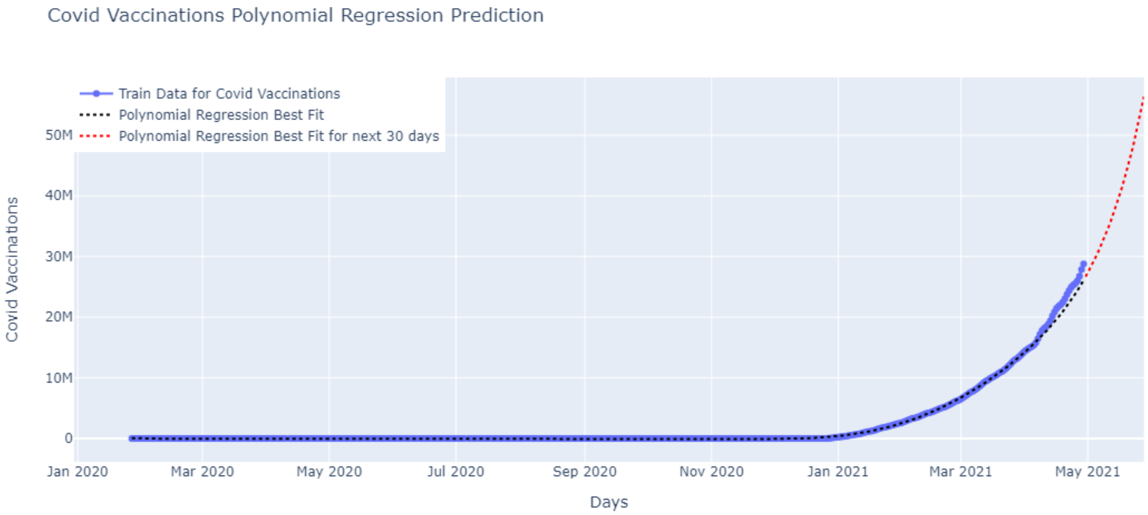
3.1.1 Polynomial Regression Model



For the polynomial regression model, we ran the model to fit the polynomial regression from n=2 to n=14 and chose the best fitting curve based on RMSE scores and the curve.

For the data of Germany, we had the best fitting curve at n=12 for total confirmed cases, n= 8 for total deaths and n=10 for total vaccinations.

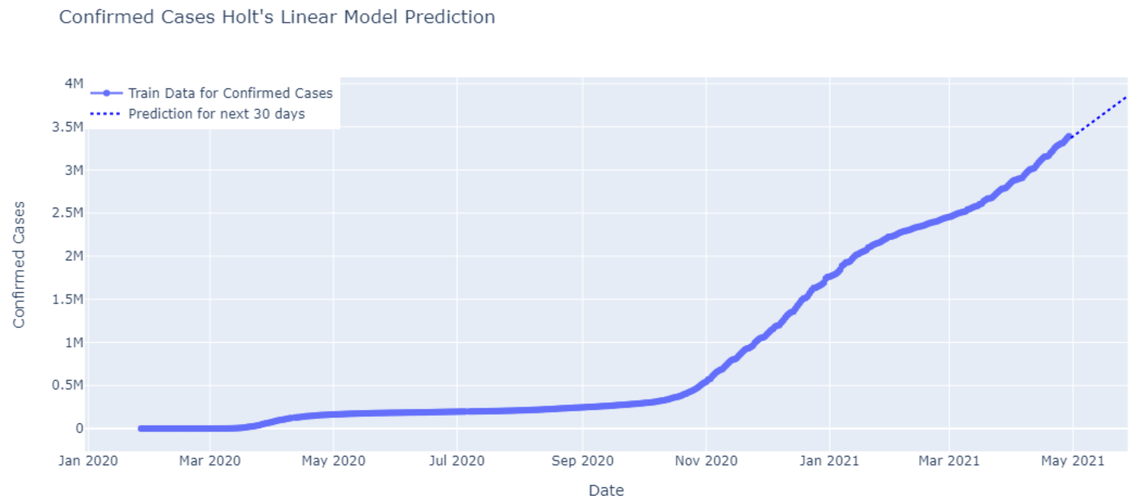
Polynomial regression was most useful for predicting Total Cases and Vaccinations.



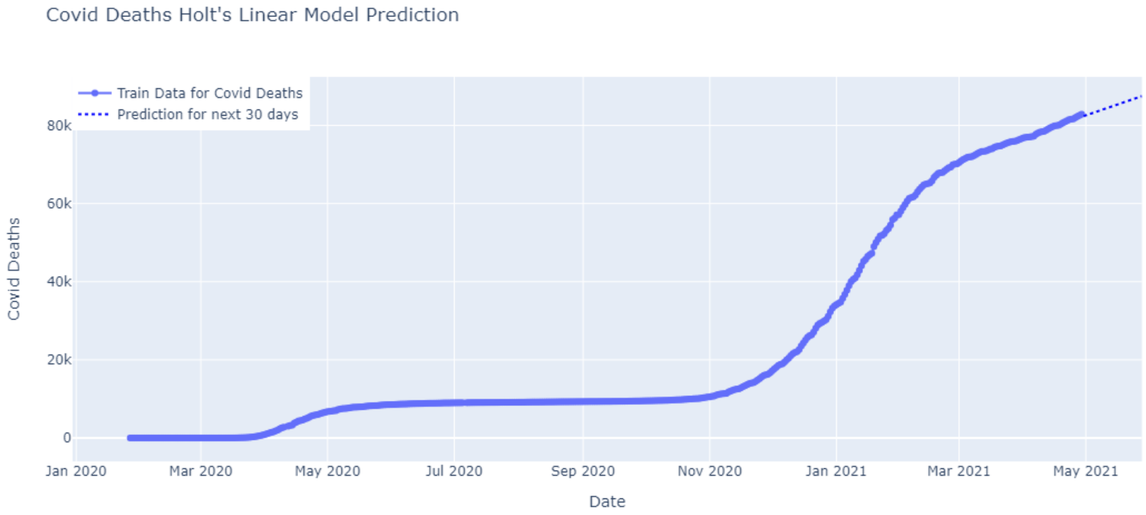
As per the curve, Approximately 50 million people will be vaccinated against covid by the end of May which is approximately 60% of the total population. The similarity with the exponential curve can be seen above.

3.1.2 Holt's Linear model

This model makes linear predictions mostly on the basis of the most recent trend of the curve. Considering the 3rd wave going on it is useful to predict the trend of confirmed cases and deaths in the near future.

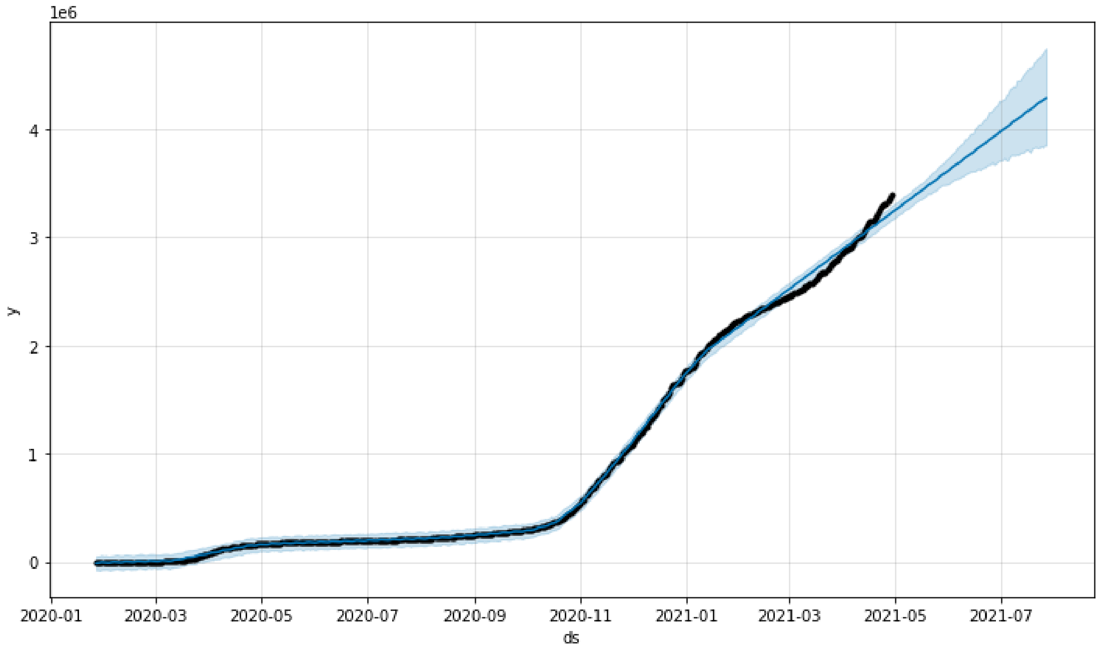


As per the curve, there will be around 450000 new cases in the month of May. But the curve might flat out due to increasing vaccinations.



As can be seen from the slope of the curve, death rate has reduced now as compared to that during december-january.

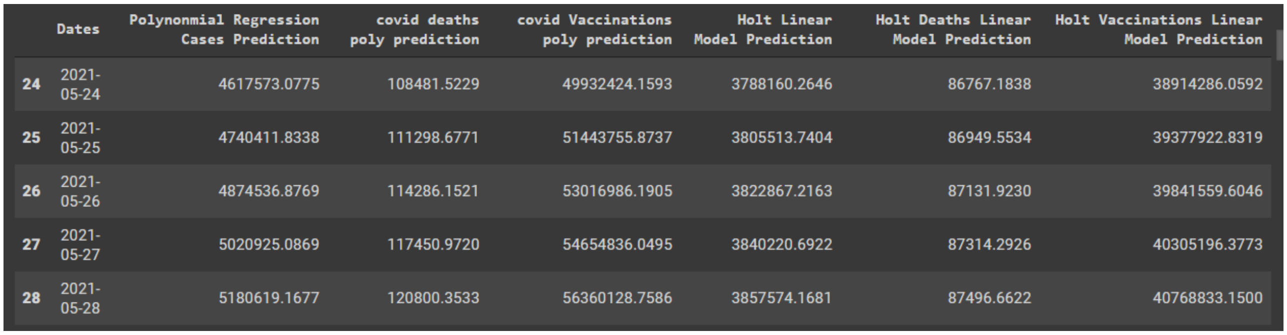
3.1.3 Facebook's Prophet Model



We can get an idea of the most probable area under which the curve is likely to occur in the next 3 months. Without considering the impact of vaccination, there are likely to be 40 million total confirmed cases by the end of june. But as per the earlier exponential vaccination curve, the rate of increases of cases is likely to reduce.

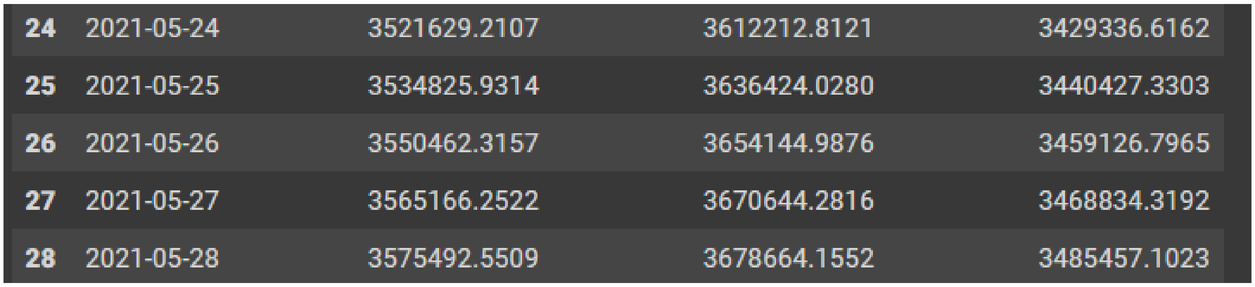
3.1.4 Numerical comparison of the models

To get a general idea of the Numerical difference between the 3 models, the predictions for 5 days towards the end of may are given below.



	Dates	Polynonmial Regression Cases Prediction	covid deaths poly prediction	covid Vaccinations poly prediction	Holt Linear Model Prediction	Holt Deaths Linear Model Prediction	Holt Vaccinations Linear Model Prediction
24	2021-05-24	4617573.0775	108481.5229	49932424.1593	3788160.2646	86767.1838	38914286.0592
25	2021-05-25	4740411.8338	111298.6771	51443755.8737	3805513.7404	86949.5534	39377922.8319
26	2021-05-26	4874536.8769	114286.1521	53016986.1905	3822867.2163	87131.9230	39841559.6046
27	2021-05-27	5020925.0869	117450.9720	54654836.0495	3840220.6922	87314.2926	40305196.3773
28	2021-05-28	5180619.1677	120800.3533	56360128.7586	3857574.1681	87496.6622	40768833.1500

The numbers for prophet model are:

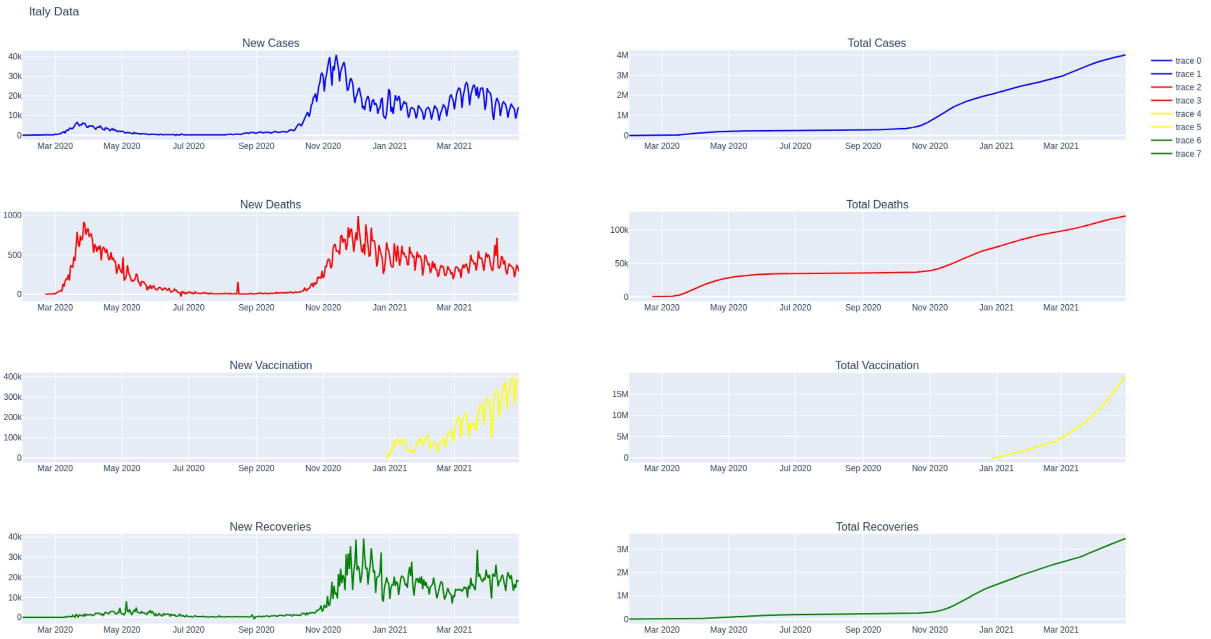


24	2021-05-24	3521629.2107	3612212.8121	3429336.6162
25	2021-05-25	3534825.9314	3636424.0280	3440427.3303
26	2021-05-26	3550462.3157	3654144.9876	3459126.7965
27	2021-05-27	3565166.2522	3670644.2816	3468834.3192
28	2021-05-28	3575492.5509	3678664.1552	3485457.1023

Considering the case numbers on 28th 2021, The polynomial predictions seems way too high compared to the prophet model.

3.2 Italy analysis

Italy suffered from 3 phases of covid 19 which is evident from the plot of new cases. The first wave occurred during March, 2020, second during November and the final wave which occurred in continuation with the second wave occurred during March,2021. From the graphs, we can see that the asymptotic region of covid graph has already been reached in Italy which is a wonderful head towards herd immunity against Coronavirus.



It is interesting to note that the peak of daily deaths due to covid 19 during the 1st wave and 2nd wave were nearly the same which implies a better management of the situation. Italy started its vaccination drive around March 1, 2021 which was the peak of their 2nd wave. From the graph it can be clearly seen that the daily deaths to new cases ratio started to decrease after the vaccination period. The exponential increase in vaccination in Italy suggests the constant hard work deployed in the country.

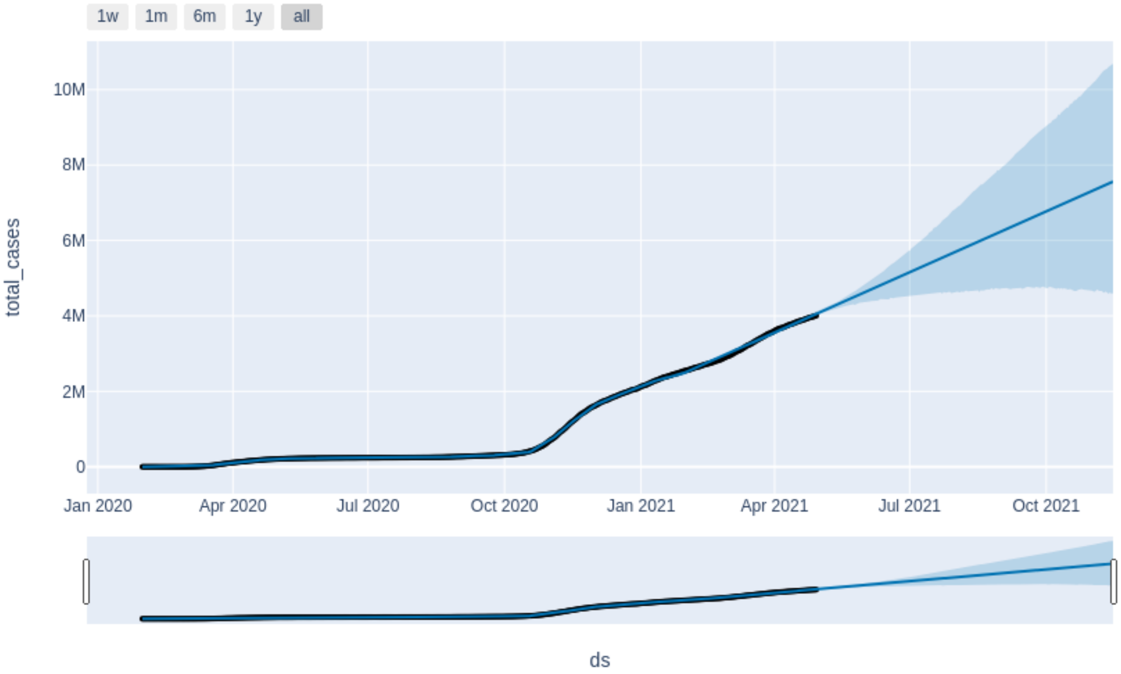
From march 31, 2020 to April 30, 2021 Italy has the following stats -

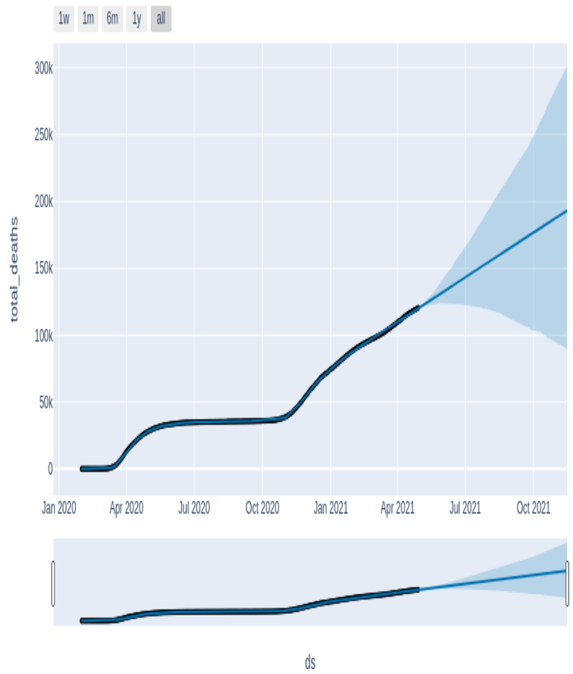
	Mean	Std	Max
New cases	8,812	9,805	40,902
Deaths	278	242	993
Recoveries	7,622	9025	39,266

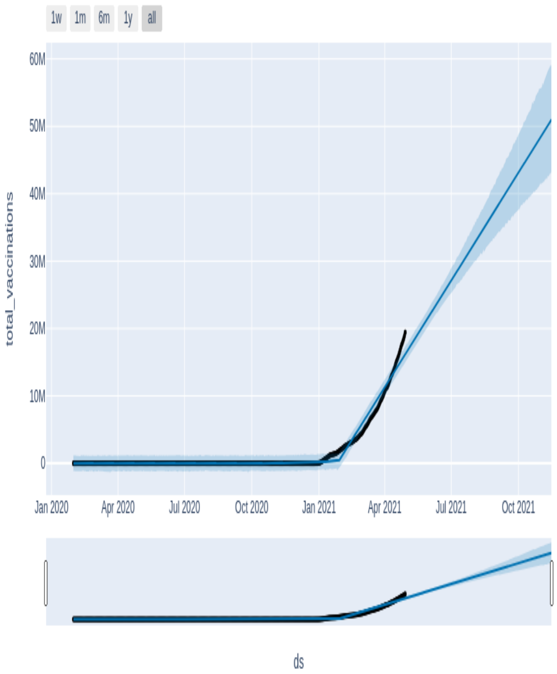
Vaccinations	157,817	58,042	497,993

3.2.1 Facebook’s Prophet Model

The prophet model predicts a linear increase in total number of cases and deaths even though vaccination has an increasing trend. The important thing to observe in the Prophet's model is the confidence interval in case of total deaths data which provides a small hope for a decreased number of deaths in future.







The narrow confidence interval in case of total vaccination provides a very high chance of continuing the trend of people getting vaccinated.

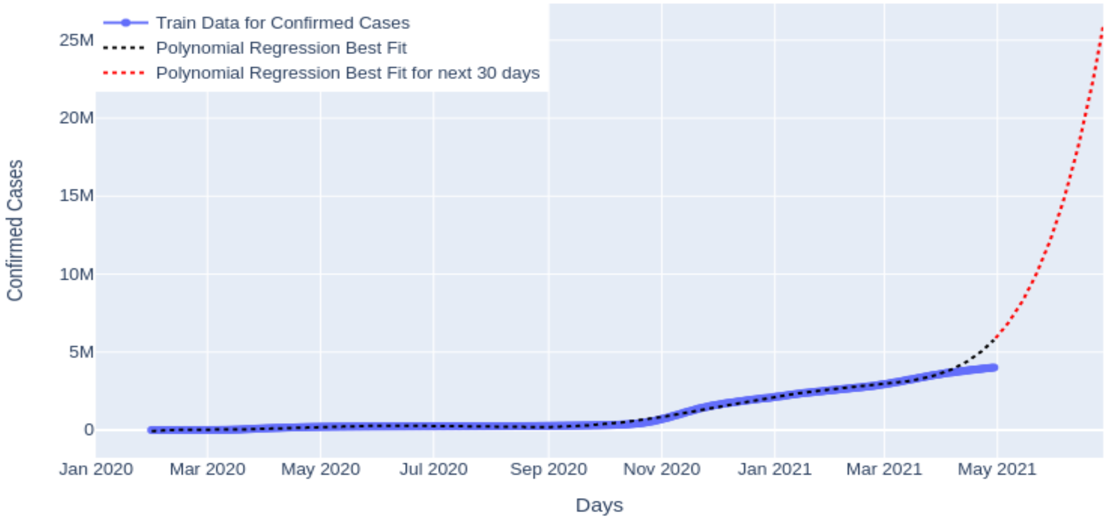
Note - The confidence interval implies a 95% confidence (alpha = 0.05)

3.2.2 Polynomial Regression Model

For the polynomial regression model we ran the model to fit the polynomial regression from n=2 to n=9 and chose the best fitting curve based on RMSE scores.

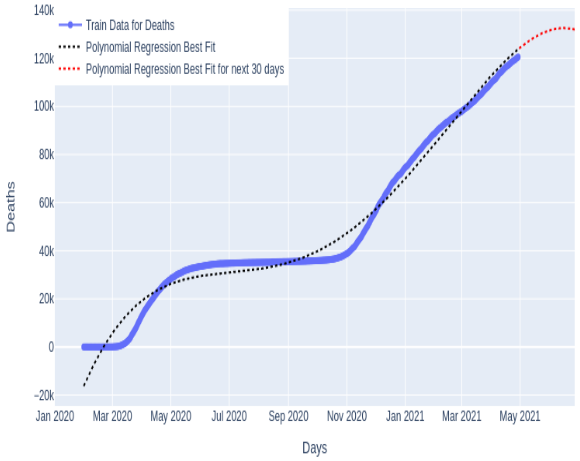
For the data of Italy we had the best fitting curve at n=8 for total confirmed cases, n= 4 for total deaths and n=5 for total vaccinations.

Confirmed Cases Polynomial Regression Prediction

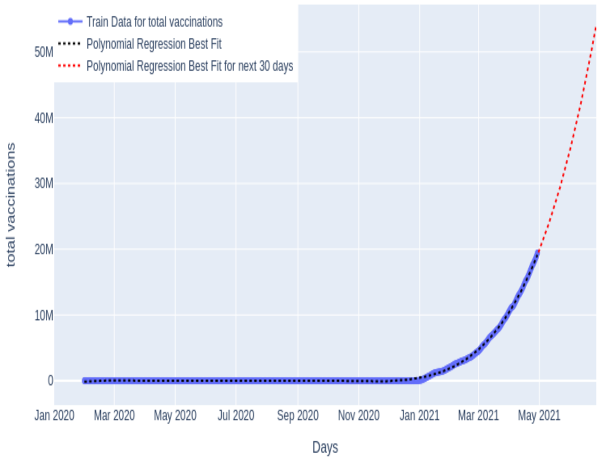


The prediction of the polynomial regression model states that the total number of cases will rise at a very high rate which may be possible and is consistent with the Prophet's prediction. The total number of deaths shows a flattening curve which is again supported by the Prophet's model. Polynomial regression presents a more close to theoretically expected curve of total vaccinations, which is an exponential curve.

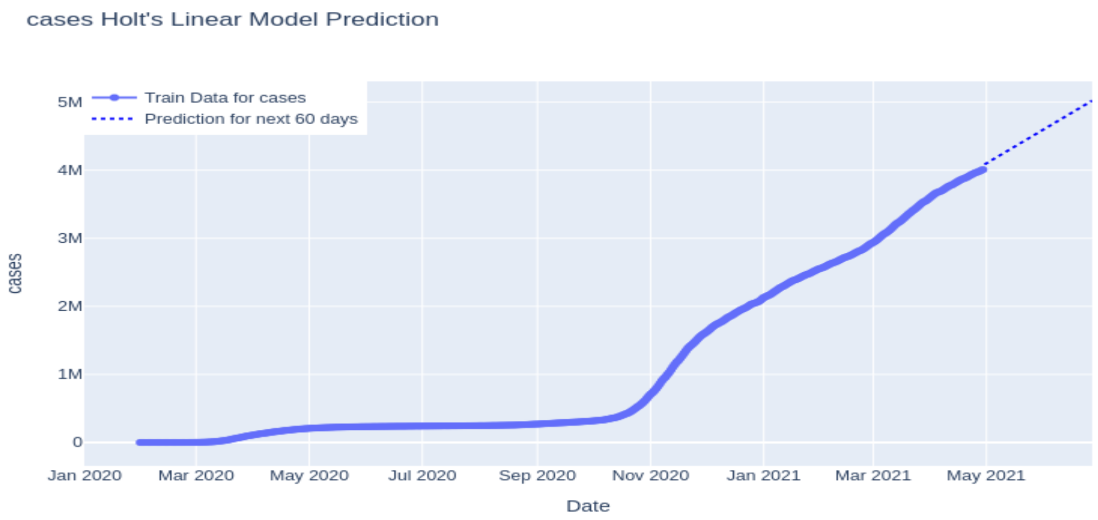
Deaths Polynomial Regression Prediction



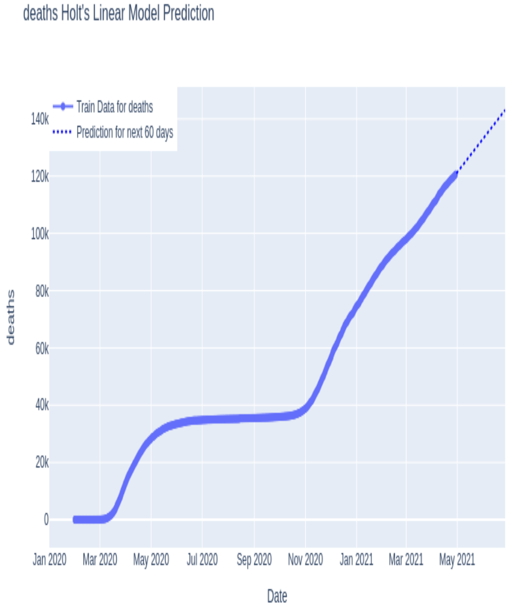
Total Vaccinations Polynomial Regression Prediction



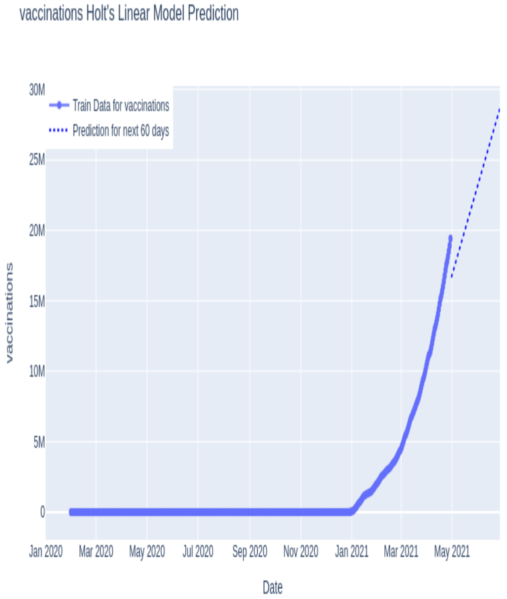
3.2.3 Holt’s Linear Model



As already mentioned, Holt’s model gives more importance to recent observations due to its exponential averages. Predictions using Holt’s methods yield a linear increase in the total number of cases, deaths and vaccination. Holt’s prediction supports Polynomial regression and Prophets in terms of total cases and vaccination but is against the total deaths.



deaths Holt's Linear Model Prediction



vaccinations Holt's Linear Model Prediction

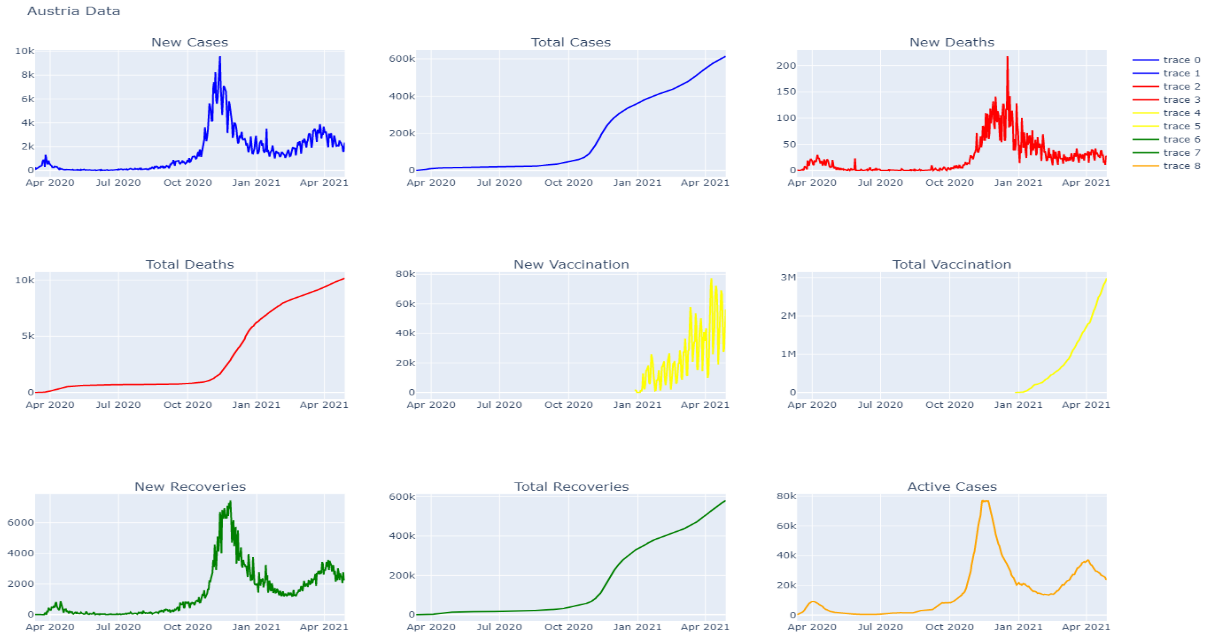
3.3 Austria analysis

3.3.1 Data Analysis

Austria has seen two waves of Covid-19 until now. With the first one during Apr 2020 - May 2020 had very few infections and deaths with maximum cases on a day being lower than 2000 cases per day, the second wave seems to be more widespread with active cases, total cases and deaths spiking at an alarming rate.

The first wave was curbed down by a lockdown from 16 March 2020 to 20 April 2020, resulting in no death on 17th May, first time since 20th March . As a result we can see total active cases being almost constant from May 2020 till starting of October. The second wave was marked from Oct 2020 with cases crossing 1000 per day for the first time and resulted in a second hard lockdown from 17 Nov 2020 - 6 Dec 2020.

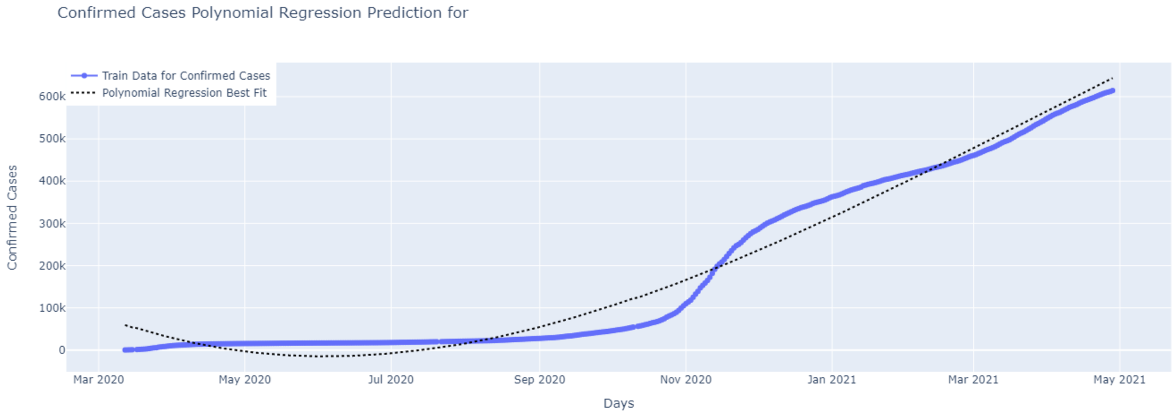
As evident from the daily new cases count, until Feb 2021 the second wave can be seen diminishing, but from March 2021 daily cases again saw a spike, introduction of vaccinations from Jan 2021 also playing an important role in reducing the second wave. But with the virus mutating, becoming airborne as well there is always a threat of cases spiking.

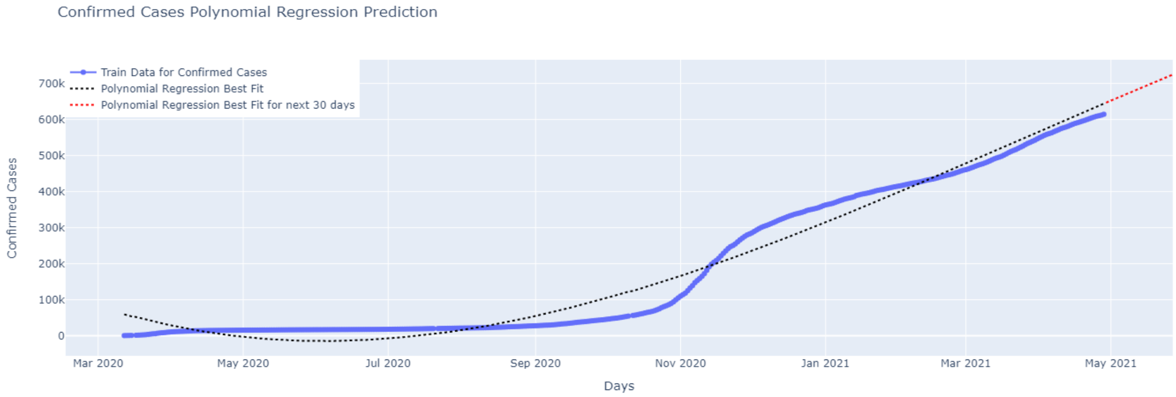


3.3.2 Polynomial Regression Model for Confirmed Cases, Total deaths

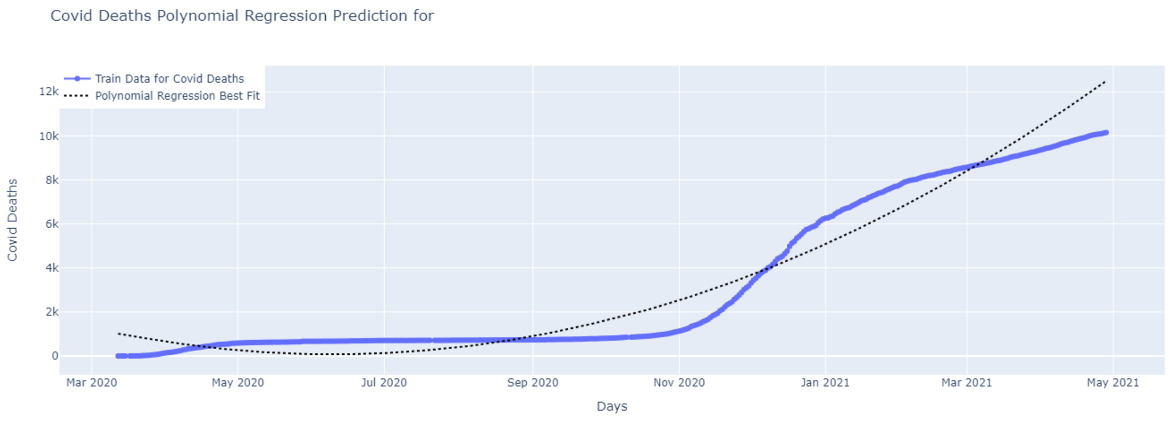
A Polynomial Regression model was run on the confirmed cases data of Austria with n ranging from n=2 to n=10 and the best fitting curve was chosen based upon the RMSE scores.

For modelling confirmed cases of Austria the best fitting curve was observed for n=3 for the Polynomial Regression model. Same model was extended for predicting new cases for the upcoming 30 days. Upward trend is in agreement with the rise in cases in Austria.



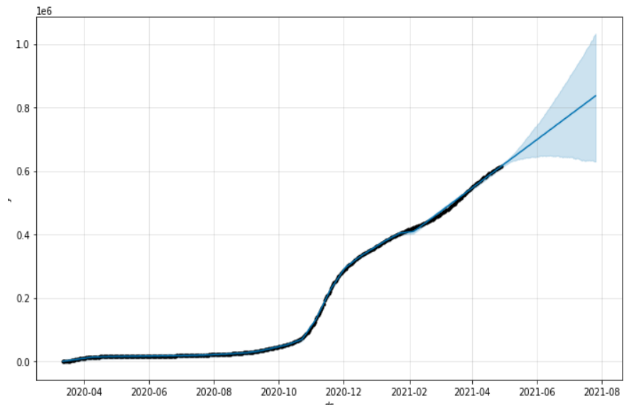


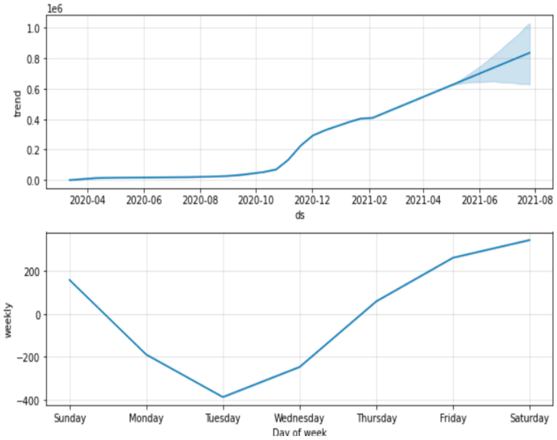
Polynomial Regression model for modelling the Covid Death data was also evaluated based upon the RMSE score and for n=2 the model was best fit and was extended to predict deaths due to covid for the upcoming month. However results were not better compared to the other two models.



3.3.4 Facebook's Prophet Model

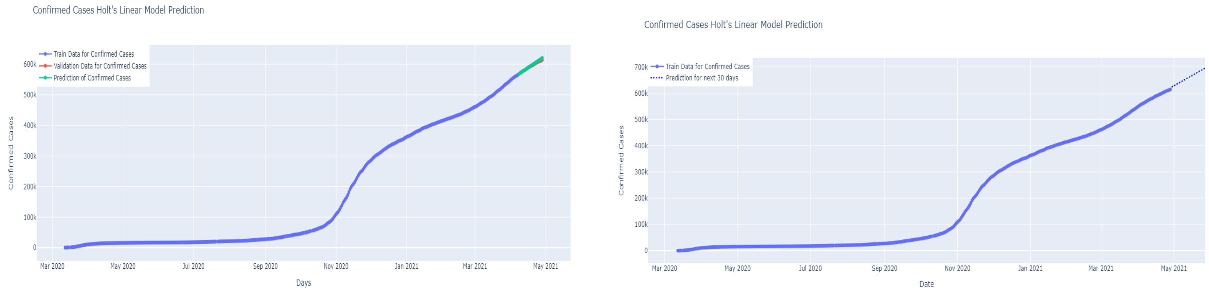
The prophet model predicts a linear increase in total number of cases even though vaccination has an increasing trend. The important thing to observe in the Prophet's model is the confidence interval in case of total deaths data which provides a small hope for a decreased number of deaths in future.



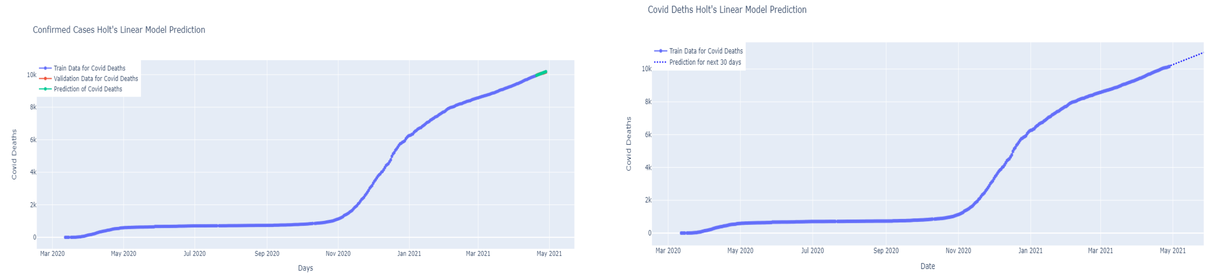


3.3.3 Holt's Linear Model for Confirmed Cases, Covid deaths and Vaccination Data

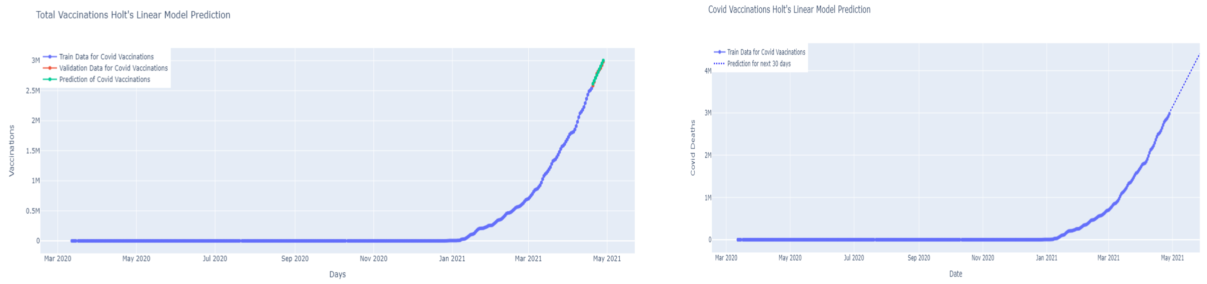
As already mentioned, Holt's model gives more importance to recent observations due to its exponential averages. Predictions using Holt's methods yield a linear increase in the total number of cases, deaths and vaccination. Holt's prediction supports Polynomial regression and Prophets in terms of total cases and vaccination but is against the total deaths.



Holt's model results for confirmed cases and predicted cases for upcoming 30 days



Holt's model results for Covid deaths and predicted deaths for upcoming 30 days

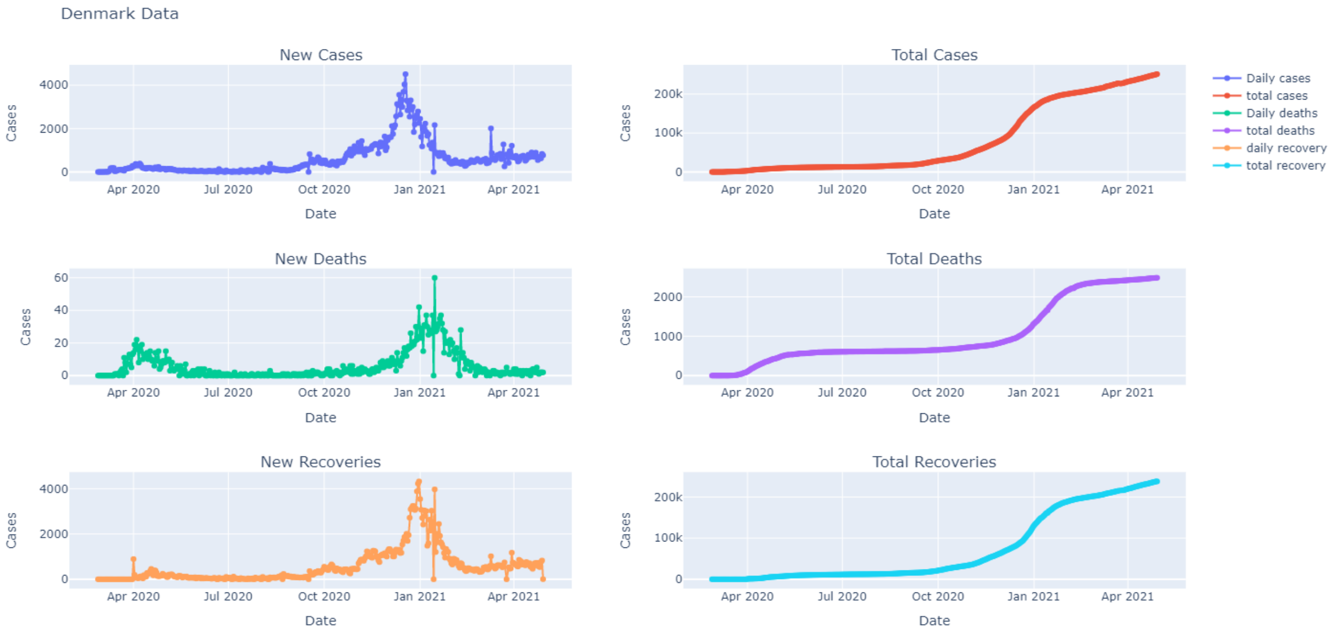


3.4 Denmark analysis

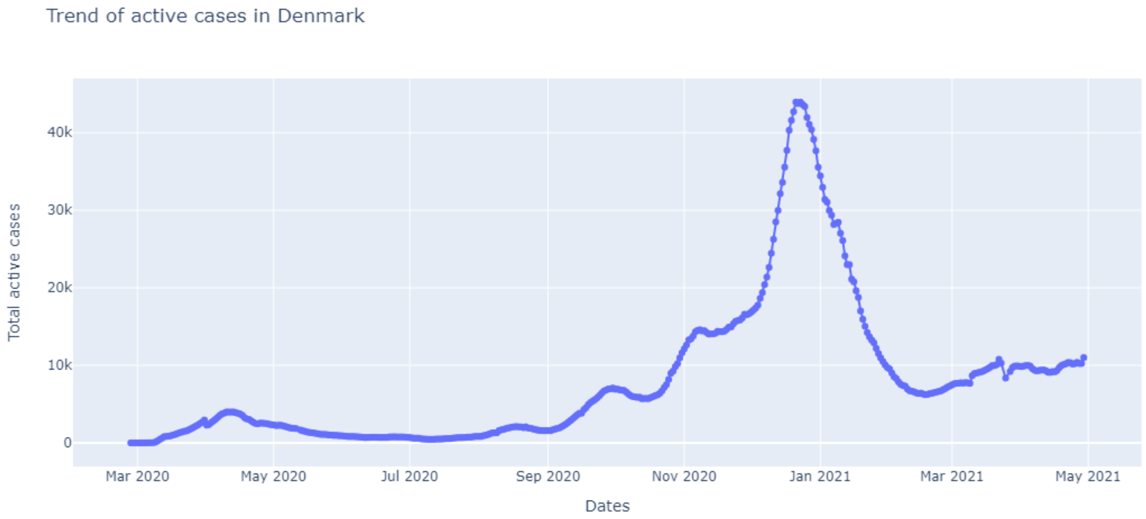
3.4.1 Data Analysis

Denmark has gone through 2 waves of COVID-19, first wave was observed during mid march to april and second wave , and the most deadly wave in Denmark was seen from November to January . Most numbers of Deaths were also reported during the month of January. Denmark imposed its first lockdown on 13th March, 2020 since the new cases were increasing every day . We can see that due to the first lockdown imposed , the cases reduced significantly since people became more aware of how the virus spread and followed social distancing norms. During the months of july to October , new cases was not rising so much , but after october we see a sudden increase in cases which might be due to factor like people were becoming more and more casual , and also because virus was mutating in other forms or may be the seasonality change might be a factor in sudden increase in cases and hence Denmark experienced the second wave of COVID-19.

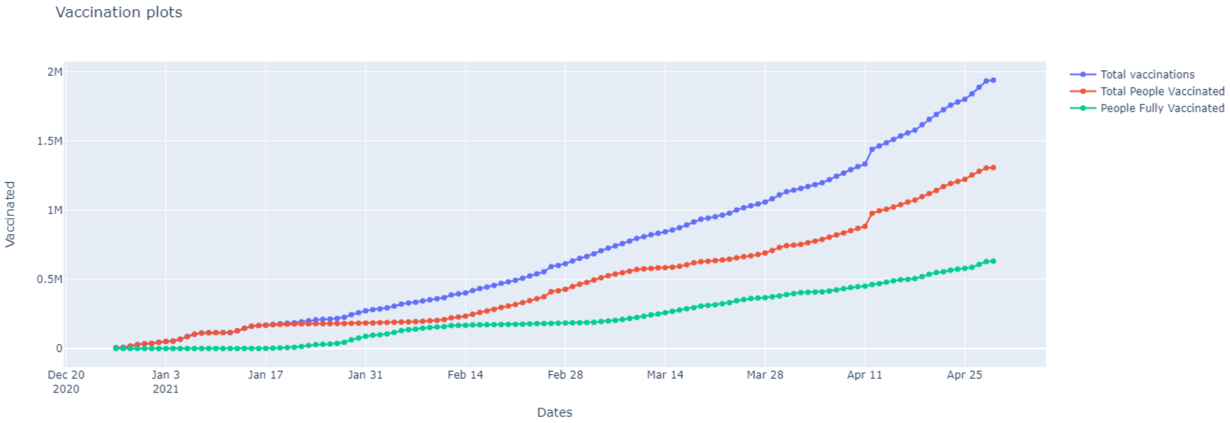
In the current month of april 2021, new cases are again rising and if social distancing rules are not followed then Denmark could soon observe its third wave of covid-19.



As shown in the figure below, we can see that the most number of active cases were seen during the month of December and January, and reporting almost 43.937k active cases on 21st December,2020. By seeing the current trend of active cases , it can be seen that active cases are yet again rising and Denmark could soon experience the Third wave of covid-19.



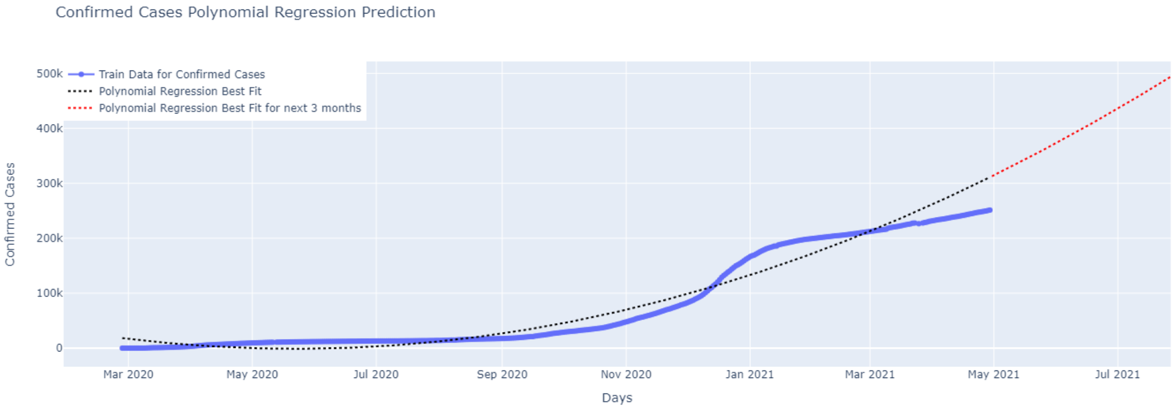
From the vaccination data , we can see that the vaccination drive started in Denmark from 27th December, 2020 and the vaccine used was Pfizer/BioNTech . We can see that Almost 1.3 million people have received the first dose of the vaccine and about 631.26k people have been fully vaccinated.Total Vaccination is on a rising trend which suggests that Denmark is going really great job in providing vaccines to its citizens.



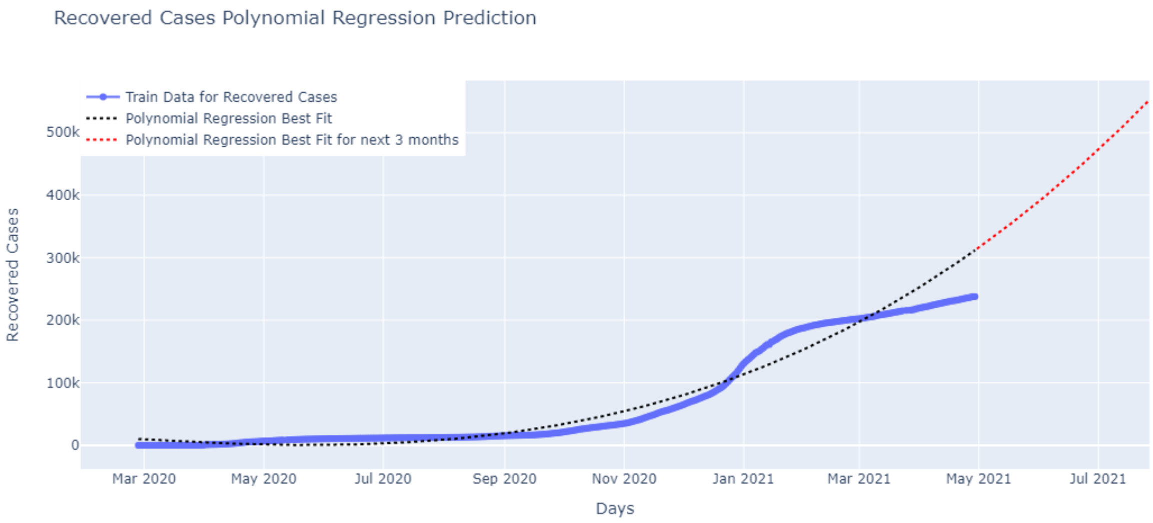
3.4.2 Forecast using Polynomial Regression for Total cases, Total Deaths, Total Recoveries, Total active cases and Vaccination

For the polynomial regression model we ran the model to fit the polynomial regression from n=2 to n=9 and chose the best fitting curve based on RMSE scores. Prediction has been made for next 3 months.

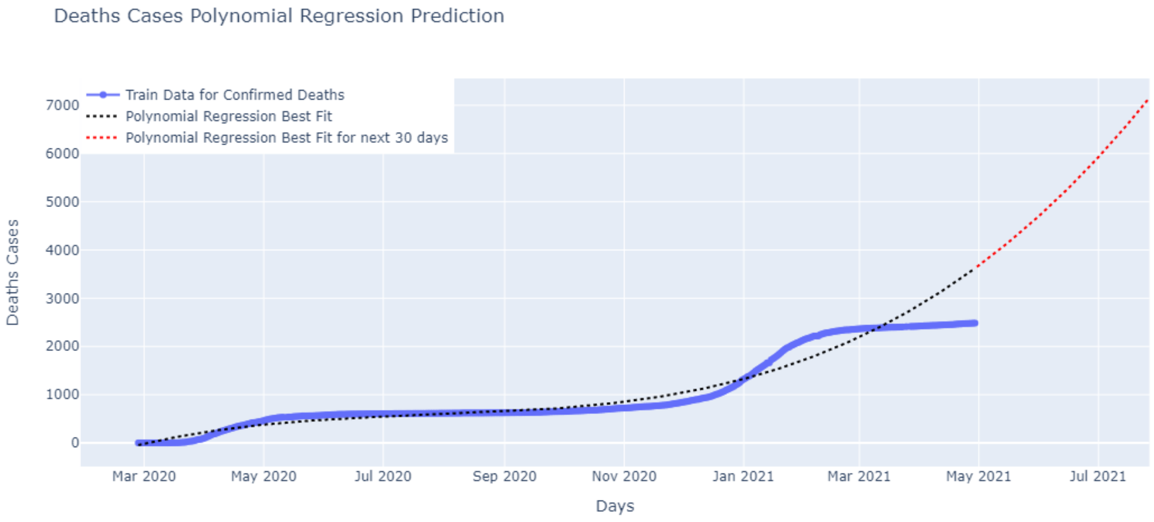
For the total cases , we get the best fitting curve with degrees, n=3.



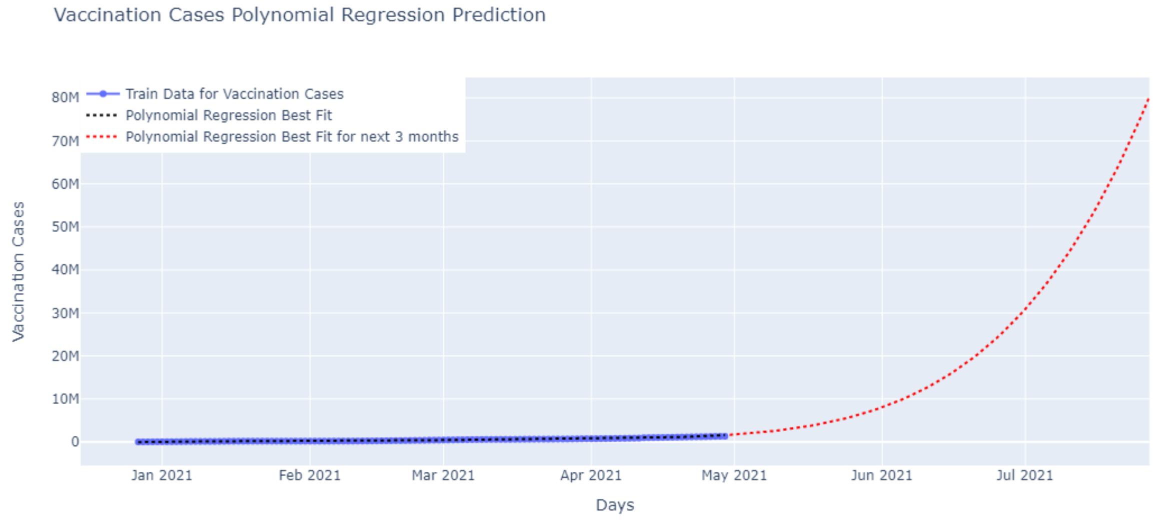
We can see that the total confirmed will increase for next 3 months according to the polynomial model prediction, which is obvious since total confirmed cases will always increase. Although, the prediction is not very accurate since now due to the introduction of vaccination , it was expected that the rate at which confirmed cases are increasing will be reduced but the prediction of polynomial regression shows that the rate of confirmed cases will increase in the near future, which is not a good sign for any country .



Above figure shows the prediction for the recovered cases, due to the introduction of vaccination it was expected that rate of recovery will increase sharply and we can see that polynomial regression also predicted a high increase in rate of recoveries in the future.



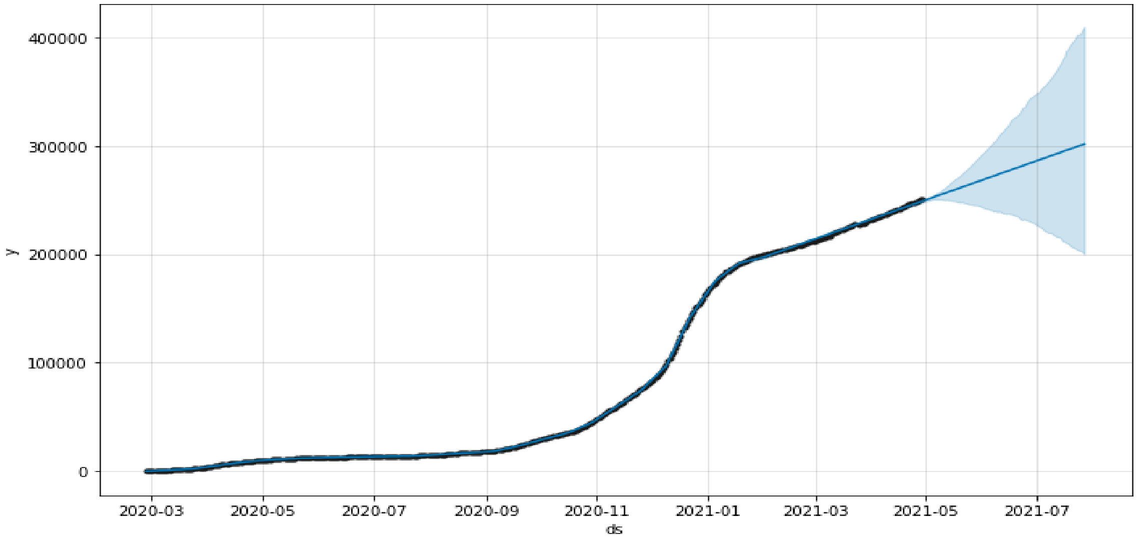
The prediction which we get for the total deaths was not even close to the actual value, since it was expected that recoveries will definitely increase which implies that the death rate should decrease , but polynomial regression predicts the other way round.



The population of Denmark is roughly close to 58.1 lakhs, and according to the polynomial prediction , it is believed that all people will receive a vaccine at the end of month of may, but technically it is not possible to vaccinate all the people within a timeframe of just one month.

3.4.3 Forecast using Facebook's Prophet Model for Total cases, Total Deaths, Total Recoveries, Total active cases and Vaccination

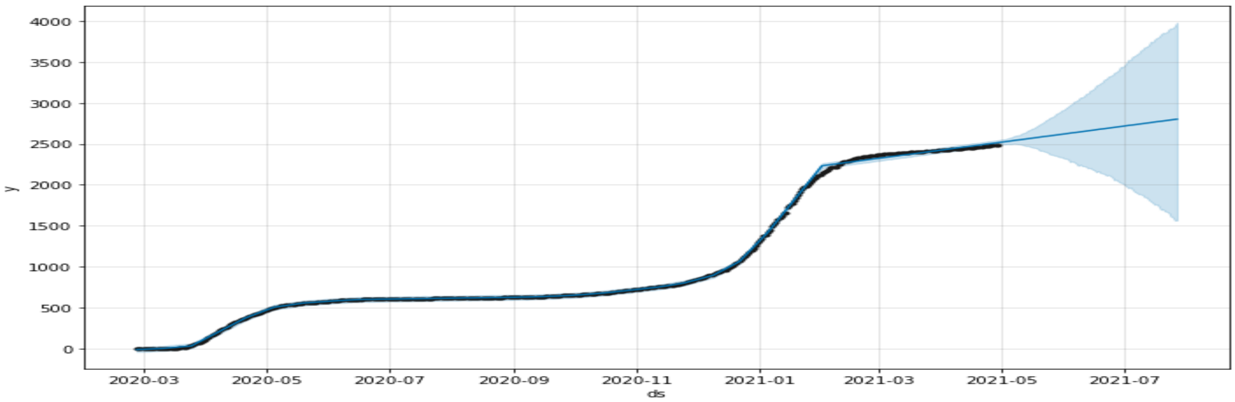
Prediction for Denmark confirmed case:



As we can see from the above graph, that total number of confirmed cases will linearly increase.

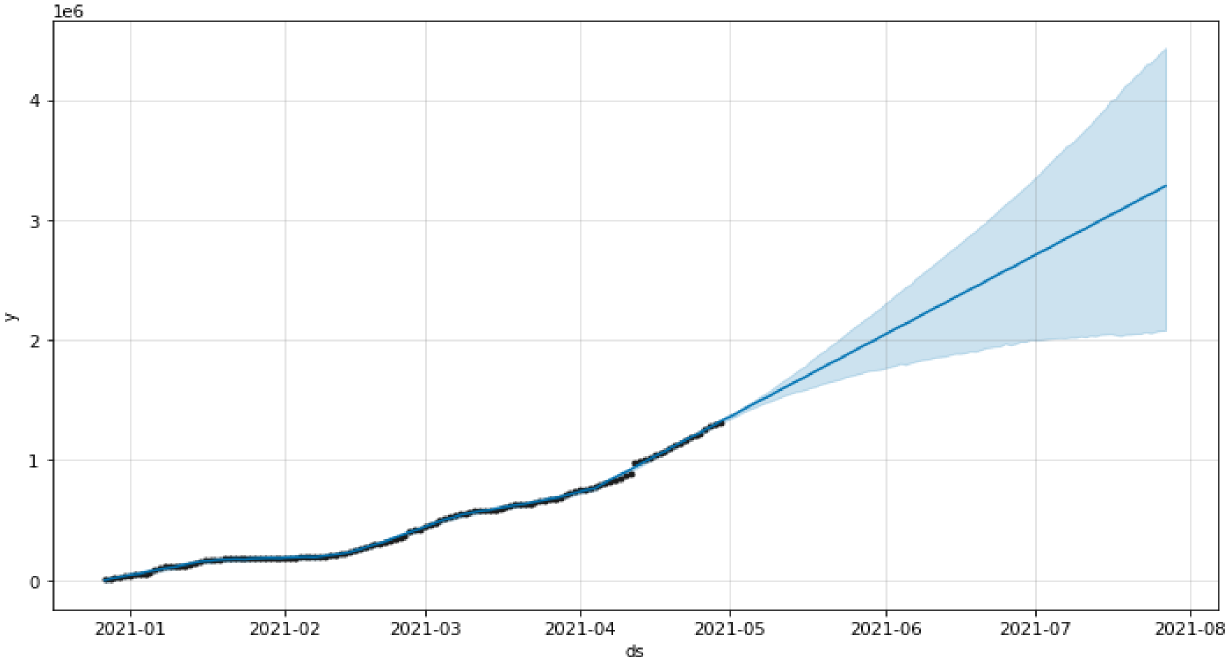
The polynomial model also predicts the increase in confirmed cases but in fbProphet model the rate at which confirmed cases will increase is much less than the rate at which polynomial predicted.

Prediction for Denmark Death case:



From the above graph, we can see that the total number of deaths increases linearly. The polynomial model predicted that deaths will increase at a very high rate, while prophet model says that deaths will increase at a very slow rate.

Prediction for Denmark Vaccinations:



From the above graph, we can state that in a time frame of 3 months, 60% of Denmark's population will be vaccinated while the polynomial model predicted that at the end of march the whole population will receive the vaccination.

CONCLUSION

Based on the general trend (Polynomial Regression), importance of recent data (Holt's Regression model) and seasonality (Facebook's Prophet), the following conclusions can be drawn -

1. Germany - Given the fact that the speedy ongoing Vaccination process, it seems likely that around 60% (majority) of the population will be vaccinated by May end. Although the active cases are rising currently towards the peak of the 3rd wave, it seems likely that the daily cases will start reducing after May (assuming the Vaccine works fine). Now we can't consider the biological factors such as mutants and strains of the virus, but there is likely to be a linear trend of total covid deaths in the near future.
2. Italy vaccination drive is in good momentum and is expected to continue the trend. Even with an increased number of cases as predicted by all the three models, the death ratio is far better than the first wave. Keeping the Polynomial Model in mind, Italy can become Corona free within a month.
3. Austria - So far the country has seen two covid waves with the second wave being more widespread. With vaccination drive in full force the second wave was diminished, but with recent virus mutations, fewer public restrictions the cases are again on a rise.
4. Denmark- So far there have been two waves of covid-19 , with the second wave being more deadlier than the first. With vaccination drive, it believed that total new cases would eventually reduce and if social distancing norms are not followed correctly then Denmark could soon see the third wave of covid-19.

REFERENCES

[1] World Health Organization, Health topics, Coronavirus, https://www.csc2.ncsu.edu/faculty/healey/covid/help.html

[2] John Hopkins University. (2020, march 3). *Coronavirus Map*. Coronavirus Resource Centre. https://coronavirus.jhu.edu/map.html

[3] European Union. (2021, April 30). *Data on COVID-19 vaccination in the EU/EEA*. European Centre for Disease Prevention and Control. https://www.ecdc.europa.eu/en/publications-data/data-covid-19-vaccination-eu-eea

[4] Europe dataset https://raw.githubusercontent.com/owid/covid-19-data/master/public/data/owid-covid-data.csv