Covid-19 Data Analysis

Manpreet Singh

2024-02-28

Covid19 Data Report

1. Dataset

Covid 19 dataset is sourced from John Hopkins University (https://github.com/CSSEGISandData/COVID-19). It contains covid confirmed cases, recovery and deaths per day for all the countries. My analysis will be around countries, hence I'll import the global data and not the US specific data set.

```
global_url = "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_
death_url = "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_c
confirmed_cases = read.csv(global_url)
deaths = read.csv(death_url)
```

2. Data Cleaning

1. Data cleaning involves pivoting the data set to get confirmed cases and death per day as rows

```
confirmed_cases = confirmed_cases %>% pivot_longer(
  col = -c(Country.Region, Province.State, Lat, Long),
  names_to = "date", "values_to" = "confirmed") %>% select(-c(Lat, Long))

deaths = deaths %>% pivot_longer(
  col = -c(Country.Region, Province.State, Lat, Long),
  names_to = "date", "values_to" = "deaths") %>% select(-c(Lat, Long))

global_cases = confirmed_cases %>% full_join(deaths) %>%
  rename(Country_Region = "Country.Region")
```

```
## Joining with `by = join_by(Province.State, Country.Region, date)`
```

2. Converting date of record to date format

```
global_cases = global_cases |>
  # Mutate Date to remove X and convert it to Date
  mutate(across(date, function(x){
    mdy(gsub("X","", x))
  }))
```

3. Analysis

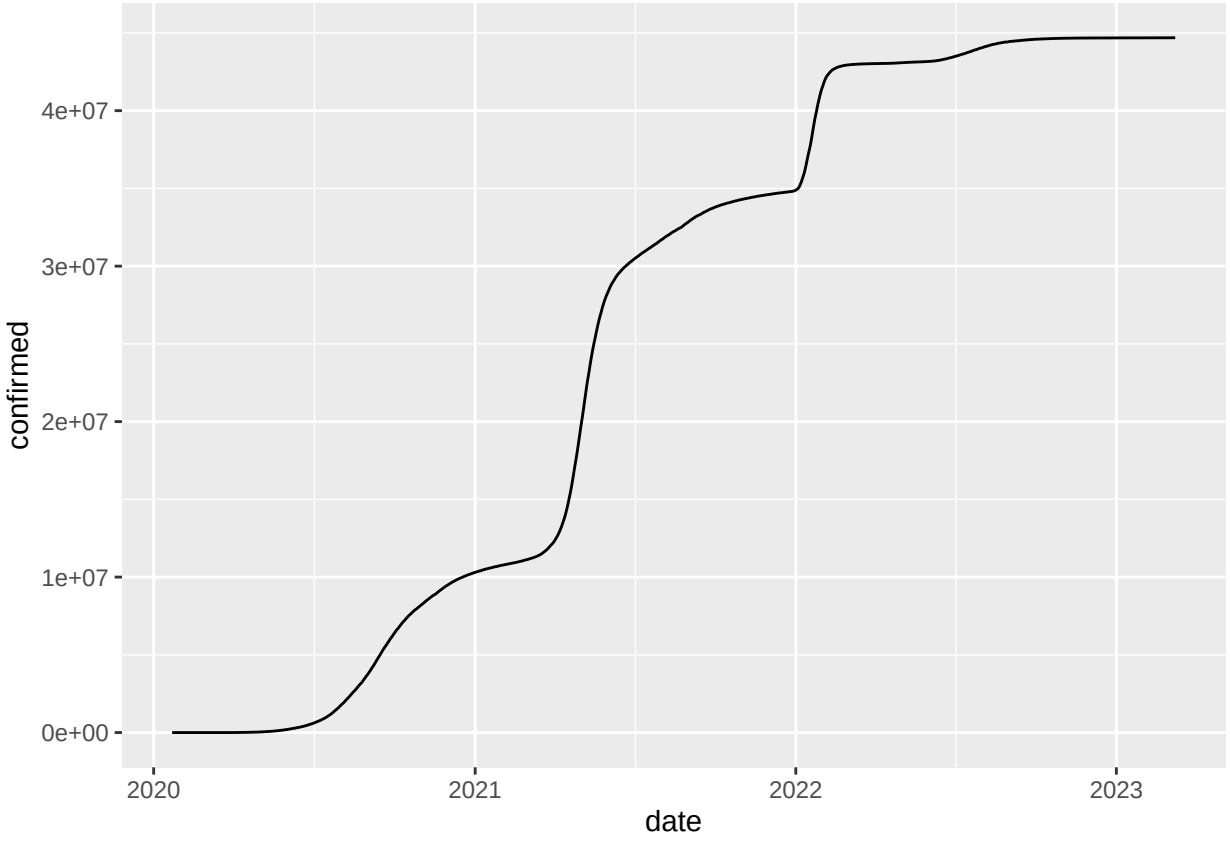
Since I'm from India, I would like to analyze covid situation in India.

Filtering cases for India and adding new cases and new deaths for each day

```
indian_cases = global_cases[global_cases$Country_Region == "India", ]

indian_cases = indian_cases %>% mutate(
  new_cases = confirmed - lag(confirmed),
  new_deaths = deaths - lag(deaths)
)
indian_cases = indian_cases %>% replace_na(list(new_cases = 0, new_deaths = 0))
```

Cases in India

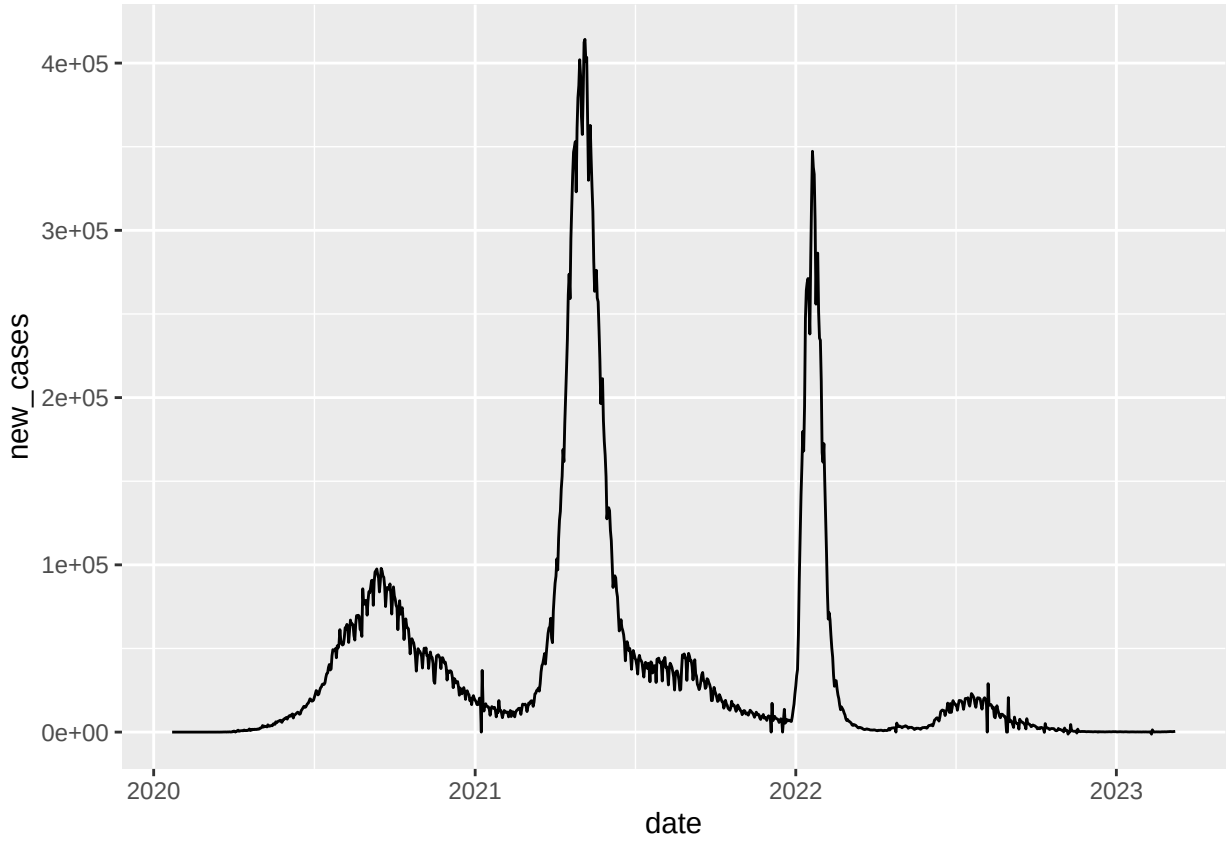
```
ggplot(indian_cases) + aes(x=date, y=confirmed) + geom_line()
```



I also wanted to know how new cases arose, did they come linearly, exponentially?

Plotting news cases shows us cases rises and causes a spike and subsides. These spikes were called waves in media and affected how lock down was implemented.

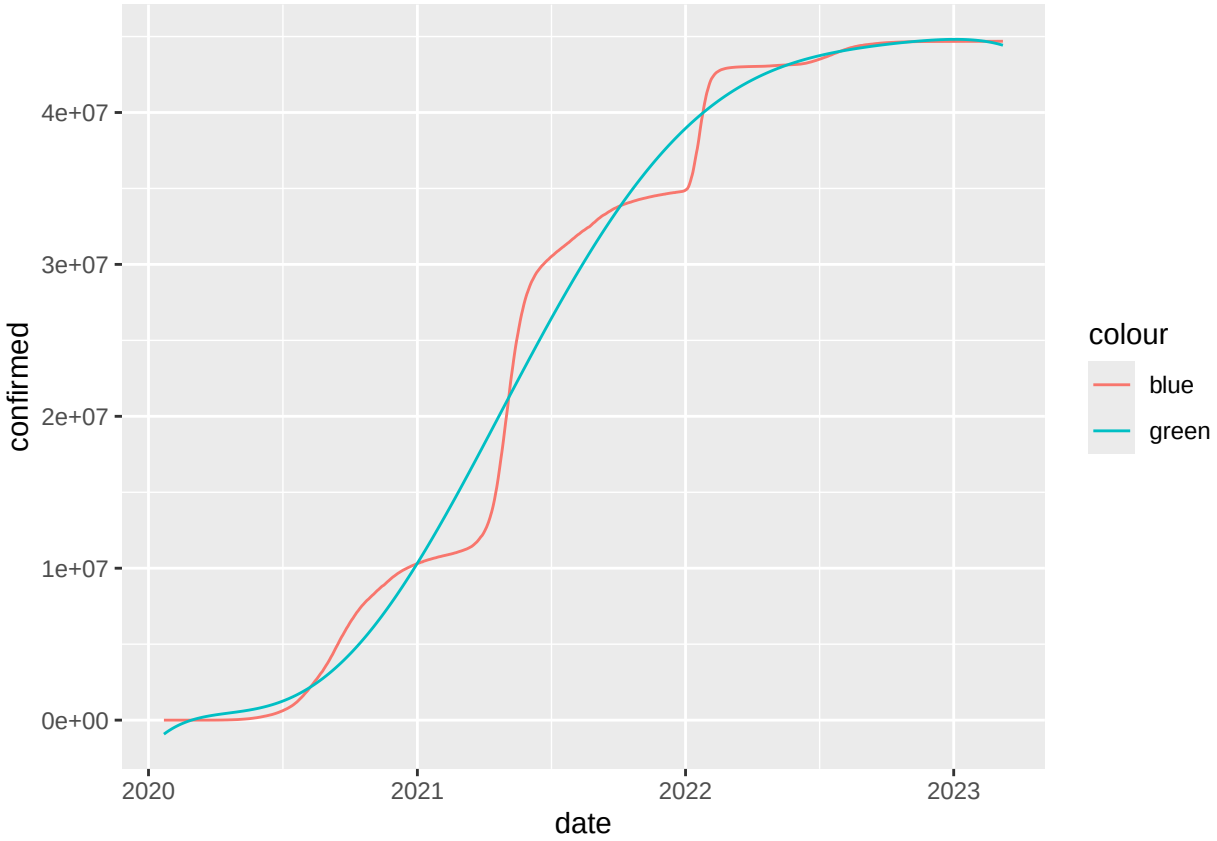
```
ggplot(indian_cases) + aes(x=date, y=new_cases) + geom_line()
```



4. Modelling confirmed cases

Seeing the characteristics of incoming confirmed cases, it looks like a polynomial graph, (more like logistic graph). So I decided to model the confirmed cases as polynomial model

```
indian_cases = indian_cases %>% mutate(day_no = seq(1, dim(indian_cases)[1]))

indian_model = lm(confirmed ~ poly(day_no, 6), data = indian_cases)
indian_cases = indian_cases %>% mutate(predicted_cases = predict(indian_model, indian_cases))

#plot predicted vs actual
ggplot(indian_cases, aes(date)) +
  geom_line(aes(y = confirmed, colour = "blue")) +
  geom_line(aes(y = predicted_cases, colour = "green"))
```



Finding RMSE of the above model

```
sqrt(mean((indian_cases$confirmed - indian_cases$predicted_cases)^2))
```

```
## [1] 1996056
```

Let's see if we can reduce this RMSE by finding an optimal degree for polynomial

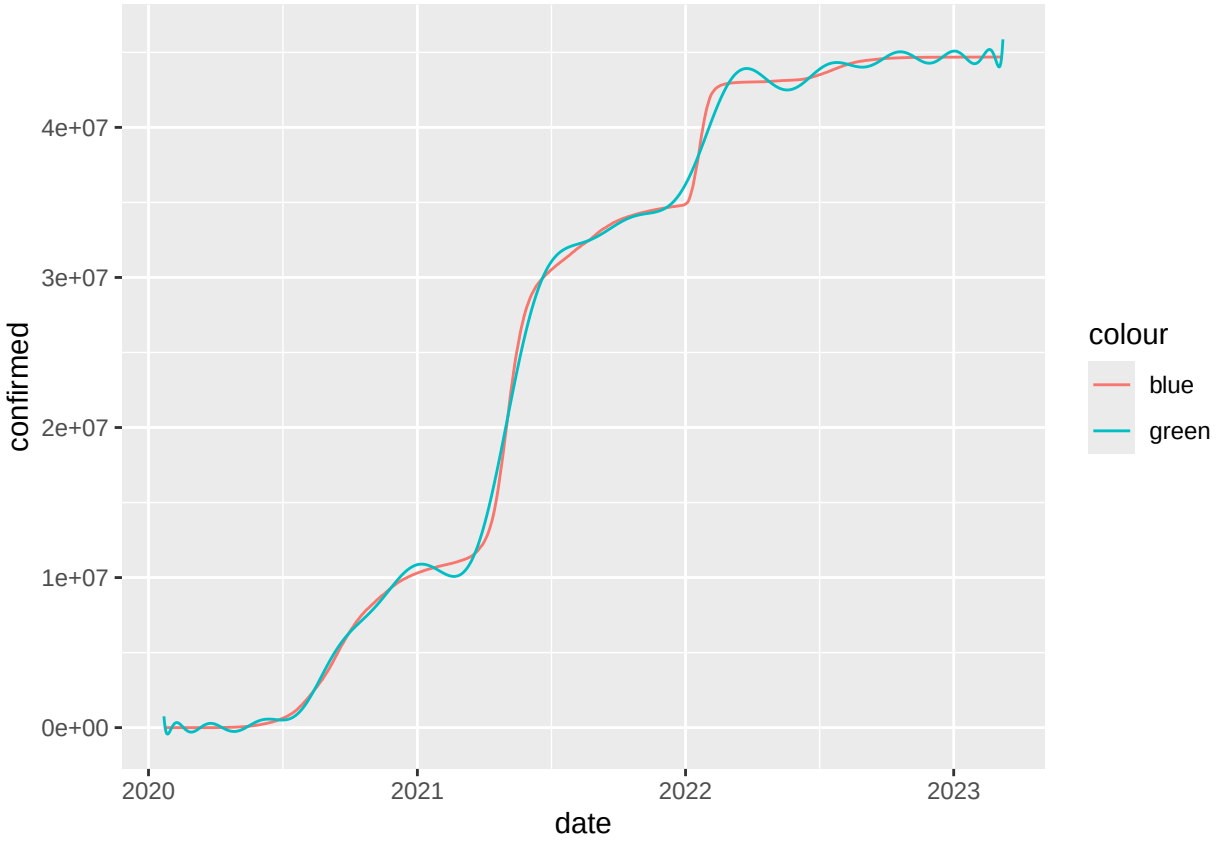
```
n_list = c()
rmse_list = c()
for(i in seq(5,27,1)){
  n_list = append(n_list, i)
  indian_model = lm(confirmed ~ poly(day_no, i), data = indian_cases)
  indian_cases = indian_cases %>% mutate(predicted_cases = predict(indian_model, indian_cases))
  rmse_list = append(rmse_list,sqrt(mean((indian_cases$confirmed - indian_cases$predicted_cases)^2)))
}
data.frame(n_list,rmse_list)
```

```
##    n_list rmse_list
## 1       5 2040837.9
## 2       6 1996055.6
## 3       7 1989362.8
## 4       8 1924910.2
## 5       9 1923958.6
## 6      10 1758543.0
```

```
## 7      11 1691636.6
## 8      12 1488872.1
## 9      13 1292423.7
## 10     14 1154302.7
## 11     15  956470.1
## 12     16  904337.7
## 13     17  825835.4
## 14     18  820829.6
## 15     19  809176.9
## 16     20  803796.6
## 17     21  803407.5
## 18     22  762482.8
## 19     23  762079.9
## 20     24  672765.0
## 21     25  669933.2
## 22     26  554317.2
## 23     27  553854.4
```

For polynomial degree 26, RMSE dropped to ~554000. Lets see how does prediction with 26 degree polynomial looks like

```
indian_cases_optimal = lm(confirmed ~ poly(day_no, 26), data = indian_cases)
indian_cases = indian_cases %>% mutate(predicted_cases = predict(indian_cases_optimal, indian_cases))

#plot predicted vs actual
ggplot(indian_cases, aes(date)) +
  geom_line(aes(y = confirmed, colour = "blue")) +
  geom_line(aes(y = predicted_cases, colour = "green"))
```

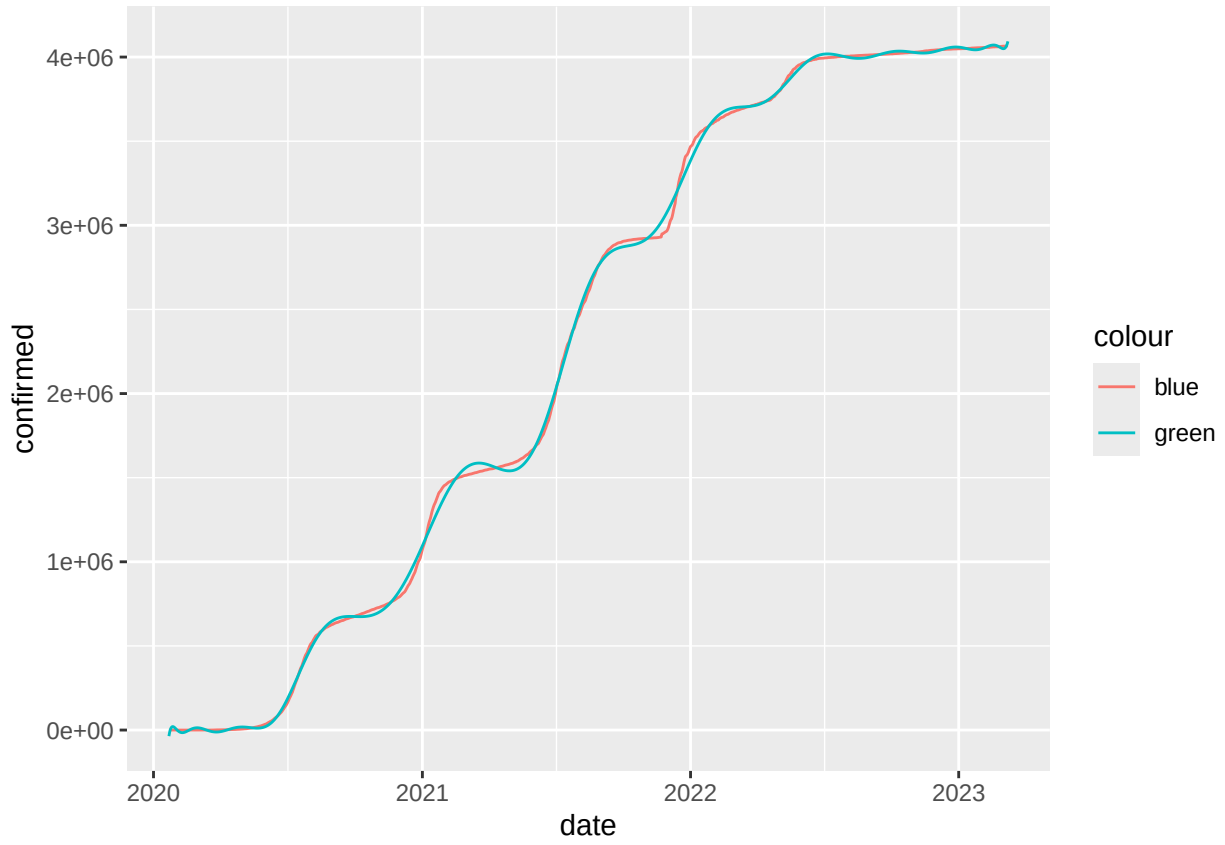


Great! We can model incoming cases with respect to number of days elapsed.
But this raises new questions -
- Does new cases always comes in waves?
- Does confirmed cases always follow polynomial relation?

To check this, I repeated the same exercise on different countries - **South Africa** and **Sweden**.I didn't optimize the degree of polynomial for them.

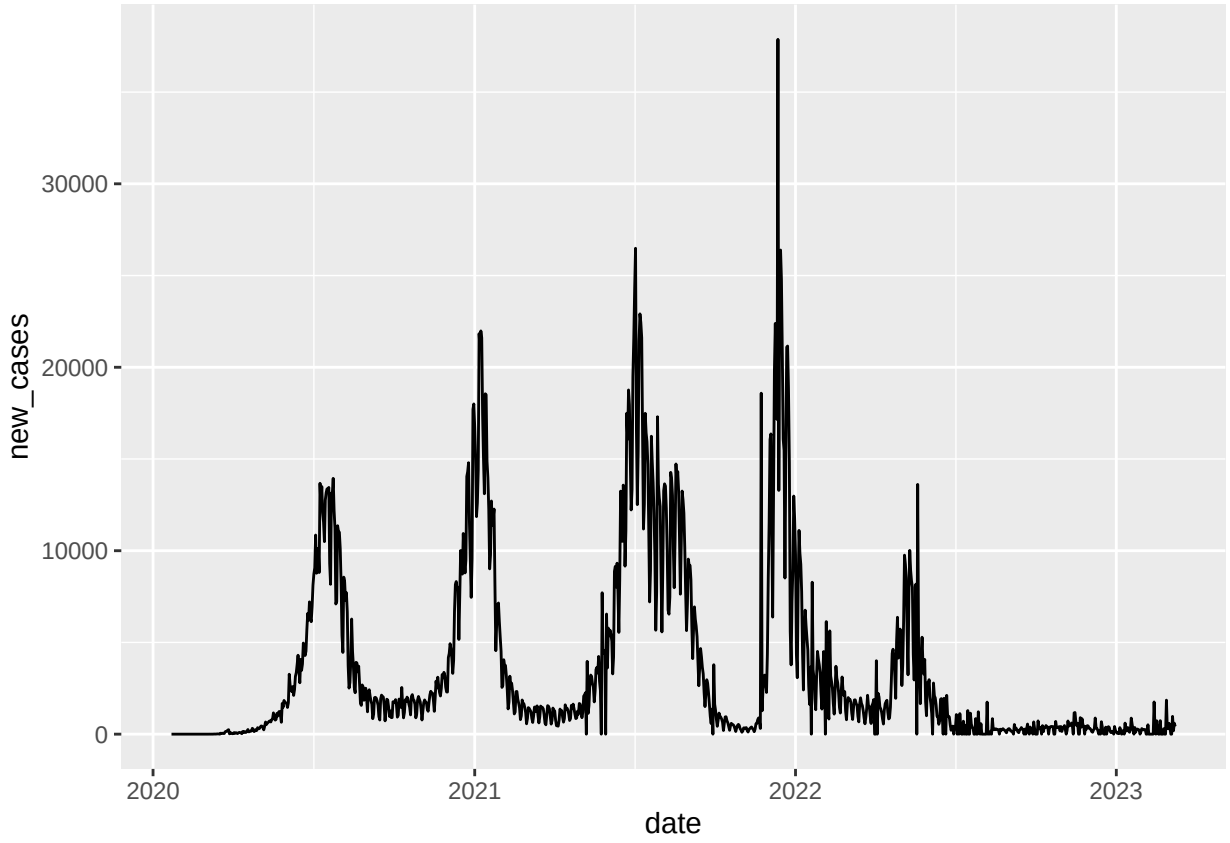
1.a New incoming case model for South Africa

```
sa_cases = global_cases[global_cases$Country_Region == "South Africa", ]

sa_cases = sa_cases %>% mutate(
  new_cases = confirmed - lag(confirmed),
  new_deaths = deaths - lag(deaths)
)
sa_cases = sa_cases %>% replace_na(list(new_cases = 0, new_deaths = 0))

sa_cases = sa_cases %>% mutate(day_no = seq(1, dim(sa_cases)[1]))

sa_model = lm(confirmed ~ poly(day_no, 26), data = sa_cases)
sa_cases = sa_cases %>% mutate(predicted_cases = predict(sa_model, sa_cases))

#plot predicted vs actual
ggplot(sa_cases, aes(date)) +
  geom_line(aes(y = confirmed, colour = "blue")) +
  geom_line(aes(y = predicted_cases, colour = "green"))
```



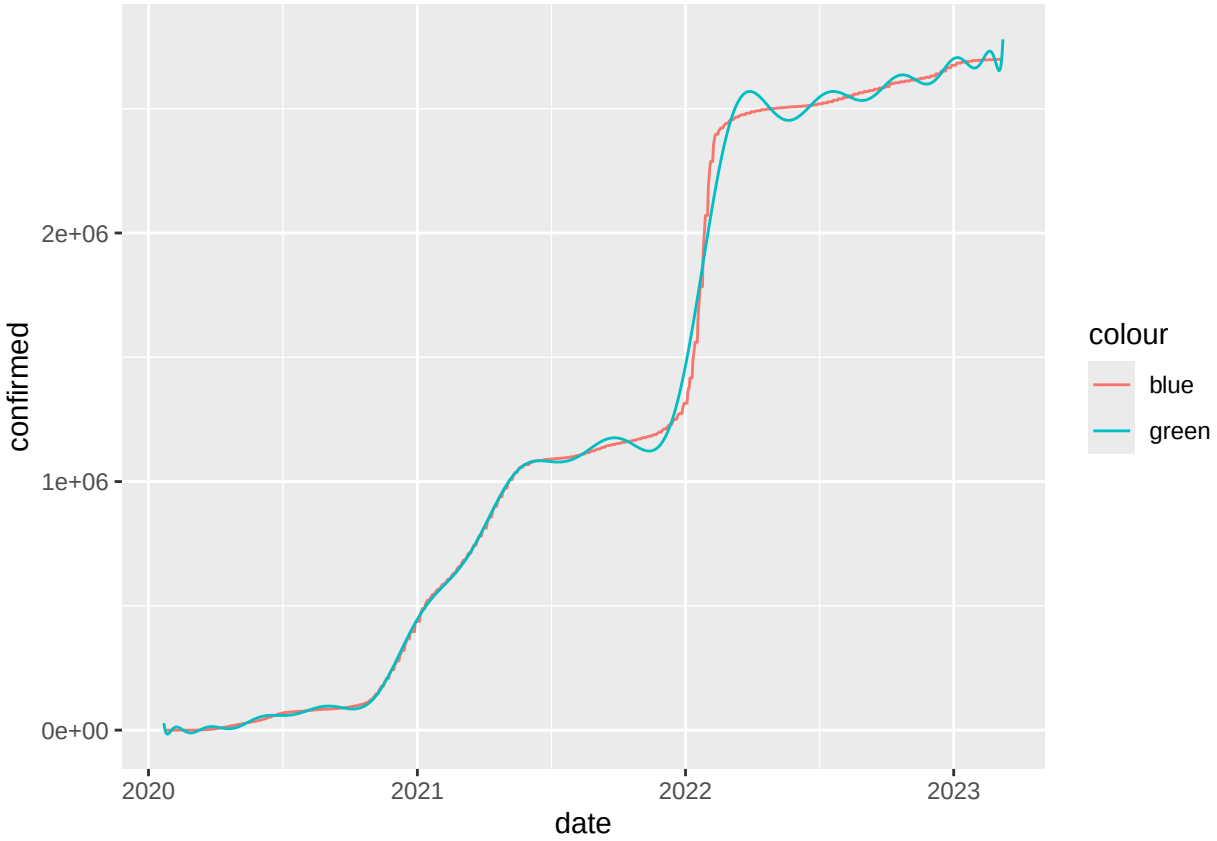
1.b Spike in cases (or corona waves) in South Africa

```
ggplot(sa_cases) + aes(x=date, y=new_cases) + geom_line()
```



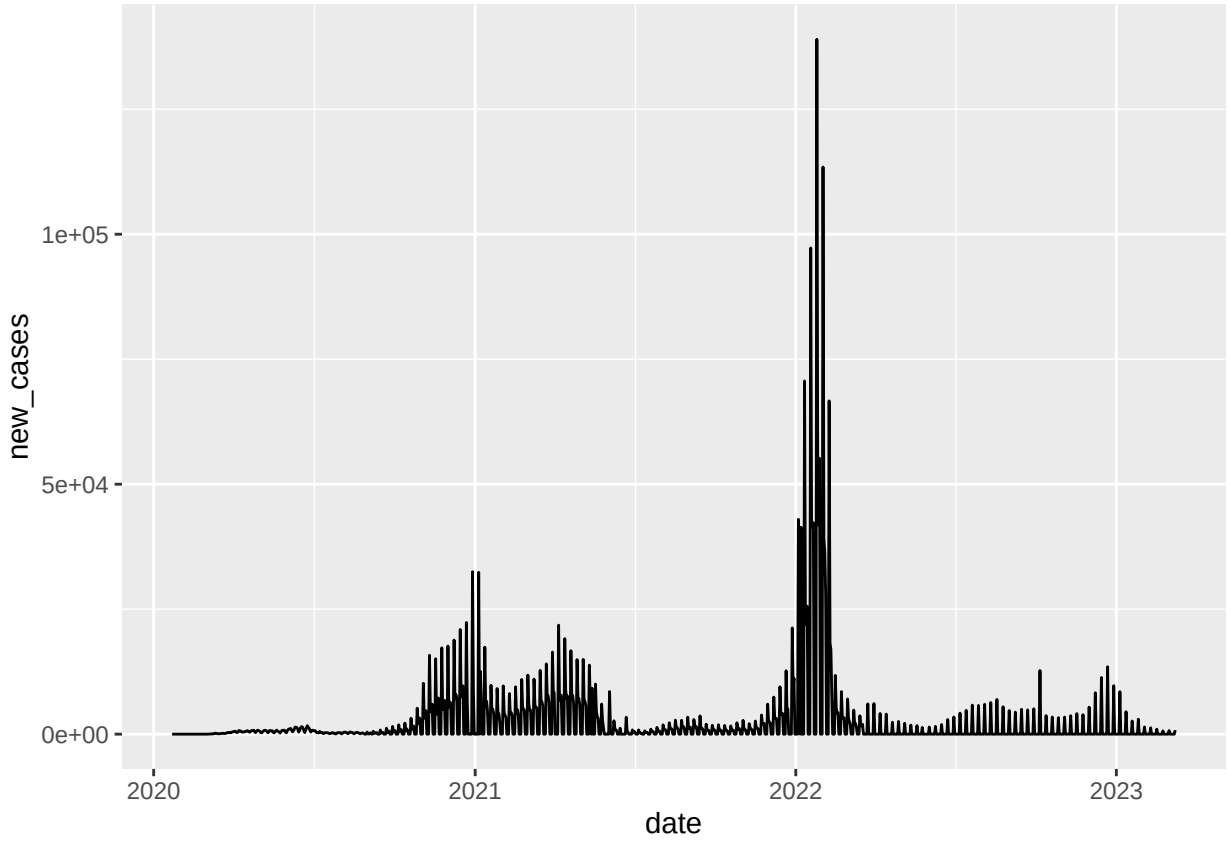
2.a New incoming case model for Sweden

```
sweden_cases = global_cases[global_cases$Country_Region == "Sweden", ]

sweden_cases = sweden_cases %>% mutate(
  new_cases = confirmed - lag(confirmed),
  new_deaths = deaths - lag(deaths)
)
sweden_cases = sweden_cases %>% replace_na(list(new_cases = 0, new_deaths = 0))
sweden_cases = sweden_cases %>% mutate(day_no = seq(1, dim(sweden_cases)[1]))

sweden_model = lm(confirmed ~ poly(day_no, 26), data = sweden_cases)
sweden_cases = sweden_cases %>% mutate(predicted_cases = predict(sweden_model, sweden_cases))

#plot predicted vs actual
ggplot(sweden_cases, aes(date)) +
  geom_line(aes(y = confirmed, colour = "blue")) +
  geom_line(aes(y = predicted_cases, colour = "green"))
```



2.b Spike in cases (or corona waves) in Sweden

```
ggplot(sweden_cases) + aes(x=date, y=new_cases) + geom_line()
```



Interestingly, both South Africa and Sweden also saw spikes of cases (4 and 2 respectively). New confirmed cases can be modeled for both the countries using polynomial model.

5. Bias

1. The source of data may not be 100% reliable since it is collected from various sources across different country.
2. Number of spikes should not used to compare handling of pandemic by different countries as each country has its own unique challenges.

6. Conclusion

1. The nature of new confirmed covid cases was waves potentially because of implementing and relaxing local lock down in respective countries.
2. Confirmed cases can be modeled using polynomial model with varying degree of polynomial.