Project

tejeshwar

4/21/2020

```r
library(rvest)
library(dplyr)
library(tidyverse)
library(zoo)
library(tidyr)
library(reshape)
library(boot)
library(splines)
```

1) Compare covid19 data from united states with Italy and Spain and also predict the number of cases and deaths based on data from Italy and Spain

a)The link https://github.com/nytimes/covid-19-data/blob/master/us-states.csv contains a breakdown of US cases and deaths of COVID-19 by state.

I have used an ongoing repository of data (GitHub) on coronavirus cases and deaths in the U.S, which is maintained by New York Times. In the repository, we have data on cumulative coronavirus cases and deaths can be found in three files, one for each of these geographic levels: U.S., states, and counties. I have scraped the states level of data from the GitHub since it offers more detail than country level of data. This was little bit straightforward since we have a csv file in the repository.

The scraped data has all its columns as character types. Column data types are changed accordingly i.e., cases column changed to numeric and date column to date format. Since the data is at state level, I have grouped by date and added to get national level data for US. I have stored this tidy version of national data as “usa_data.csv” and state level data as “usa_state_level_data”.

```r
head(usa_data)
```

```
##         date cases deaths daily.cases daily.deaths
## 1 2020-01-21     1      0           0            0
## 2 2020-01-22     1      0           0            0
## 3 2020-01-23     1      0           0            0
## 4 2020-01-24     2      0           1            0
## 5 2020-01-25     3      0           1            0
## 6 2020-01-26     5      0           2            0
```

```r
head(usa_state_data)
```

```
##         date      state cases deaths
## 1 2020-01-21 Washington     1      0
## 2 2020-01-22 Washington     1      0
## 3 2020-01-23 Washington     1      0
## 4 2020-01-24   Illinois     1      0
## 5 2020-01-24 Washington     1      0
## 6 2020-01-25 California     1      0
```

Top 5 states based on total number of cases

```r
head(usa_state_data_total,5)
```

```
## # A tibble: 5 x 3
##   state           cases deaths
##   <chr>           <int>  <int>
## 1 New York      7633815 393681
## 2 New Jersey    2650820 127023
## 3 Massachusetts 1210807  54858
## 4 California    1082131  37878
## 5 Illinois      1011635  40761
```

Top 5 states based on total number of deaths

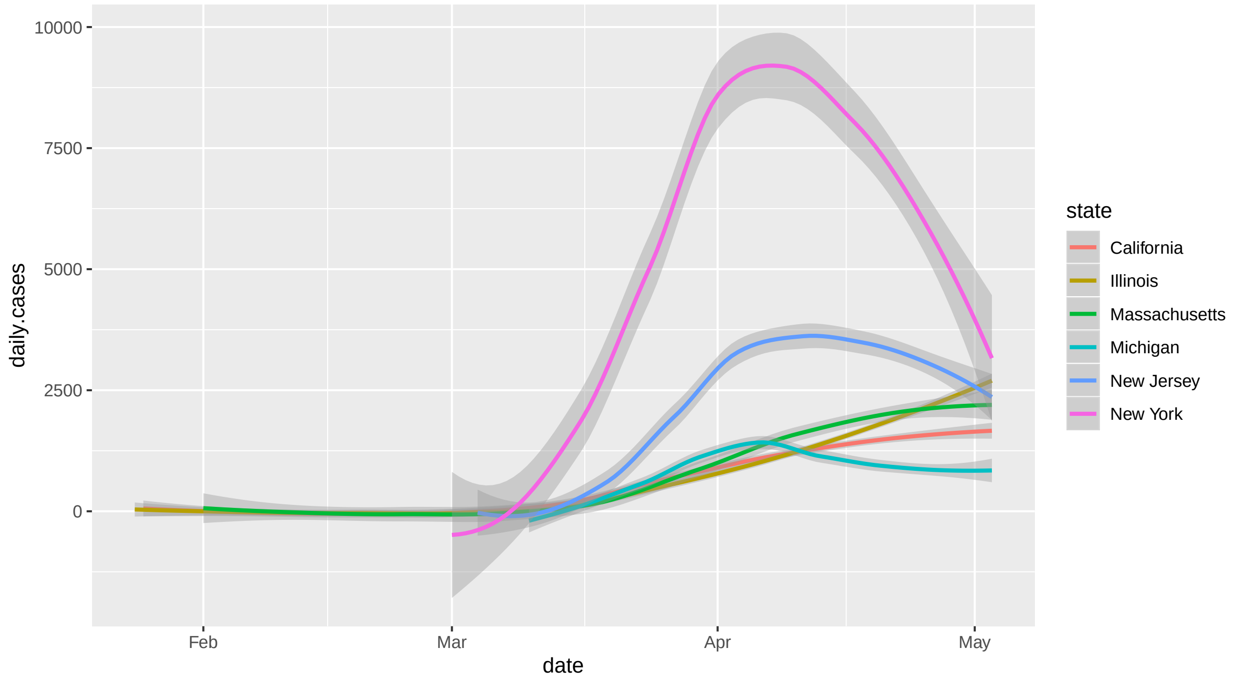
```r
usa_state_data_total %>% arrange(desc(deaths)) %>% head(5)
```

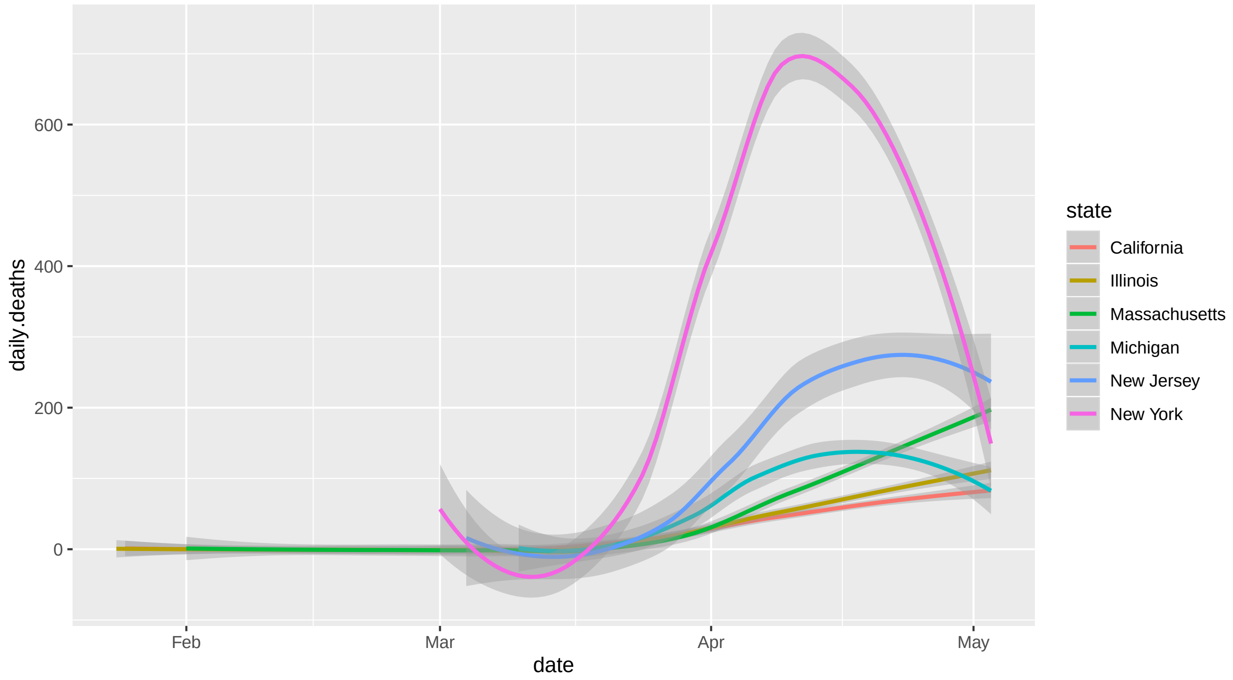
```
## # A tibble: 5 x 3
##   state           cases deaths
##   <chr>           <int>  <int>
## 1 New York      7633815 393681
## 2 New Jersey    2650820 127023
## 3 Michigan       985786  72405
## 4 Massachusetts 1210807  54858
## 5 Illinois      1011635  40761
```

All the top 5 highest daily increases in cases and deaths scenarios are from New York state, so lets look at top 5 state highest daily increases in cases and deaths

```
## # A tibble: 5 x 2
## # Groups:   state [5]
##   state         daily.cases
##   <chr>               <dbl>
## 1 New York            12126
## 2 Massachusetts        4946
## 3 New Jersey           4305
## 4 Illinois             3137
## 5 Louisiana            2726
## # A tibble: 5 x 2
## # Groups:   state [5]
##   state         daily.deaths
##   <chr>                <dbl>
## 1 New York               805
## 2 New Jersey             458
## 3 Pennsylvania           300
## 4 Massachusetts          252
## 5 Michigan               232
```

“New York”, “New Jersey”, “Massachusetts”, “Michigan”, “California” and “Illinois” are the most effected states in the US. Let’s plot the number of cases and deaths of these most affected states to get a better a idea of their situation.





We can see very clearly that in states like New York, New jesrey, the number of cases and deaths have peaked and are in a downward trend, which is a good news. But in other states trend is upwards and it might take few weeks to see a peak and a decreasing trend.

b)The link https://en.wikipedia.org/wiki/2020__coronavirus__pandemic__in__Italy contains a breakdown of Italy cases and deaths of COVID-19 & The link https://en.wikipedia.org/wiki/2020__coronavirus__pandemic__in__Spain contains a breakdown of Spain cases and deaths.

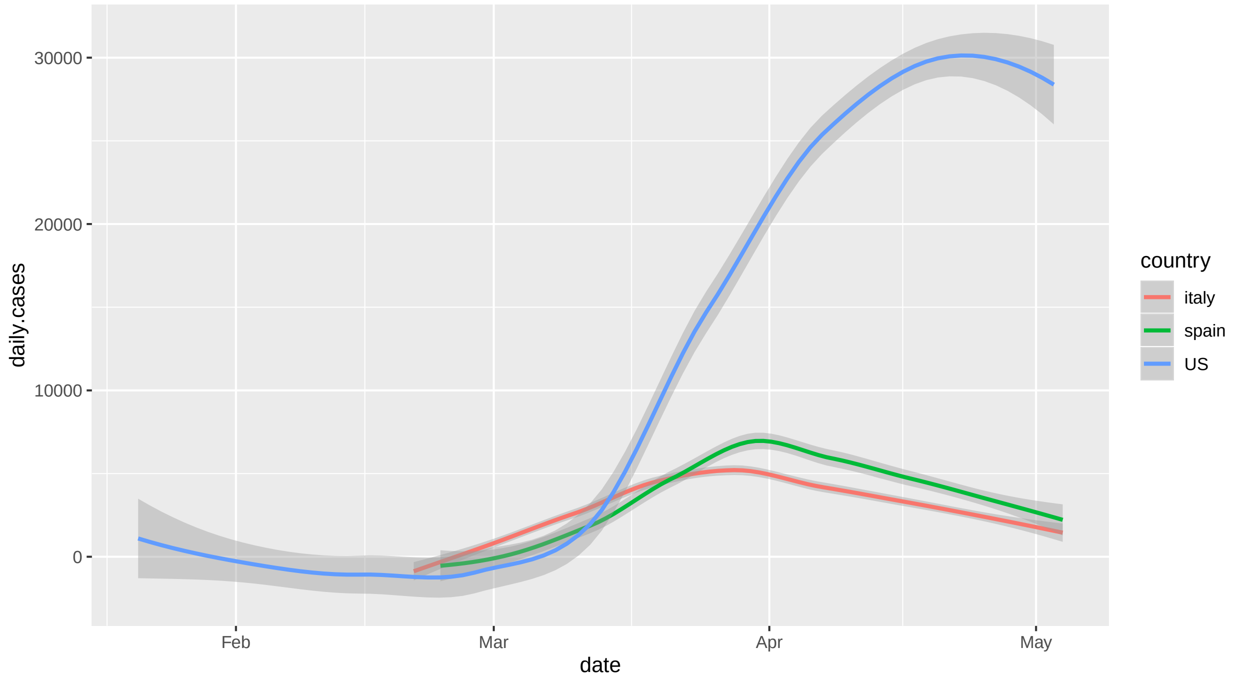
Italy and Spain data: Wikipedia webpages “COVID-19 pandemic in Italy” and “COVID-19 pandemic in Spain” has the data for covid19 cases & deaths for Italy and Spain, respectively. I have found scraping data from Wikipedia to be difficult since here the data is not in a tabular from.

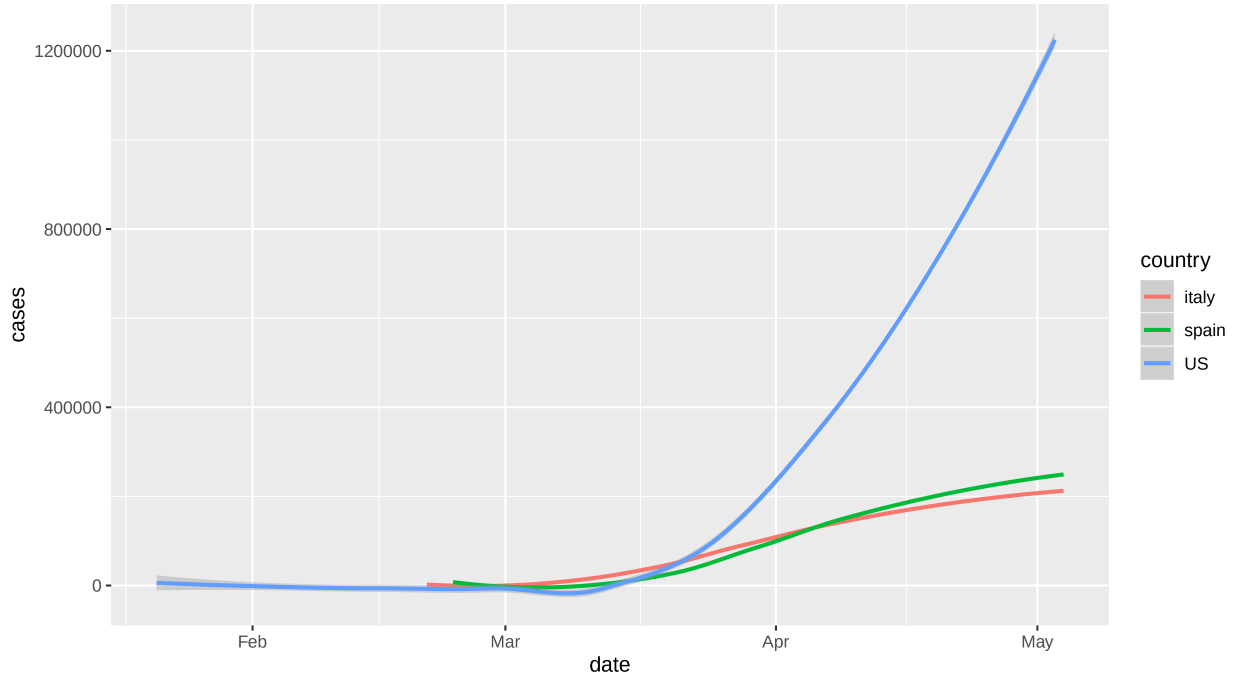
```
## [1] "Highest number of daily cases in italy and spain are 6557 and 8195 respectively"
```

```
## [1] "Highest number of daily deaths in italy and spain are 919 and 961 respectively"
```

c) Comparing the covid19 data of US, Spain and Italy.

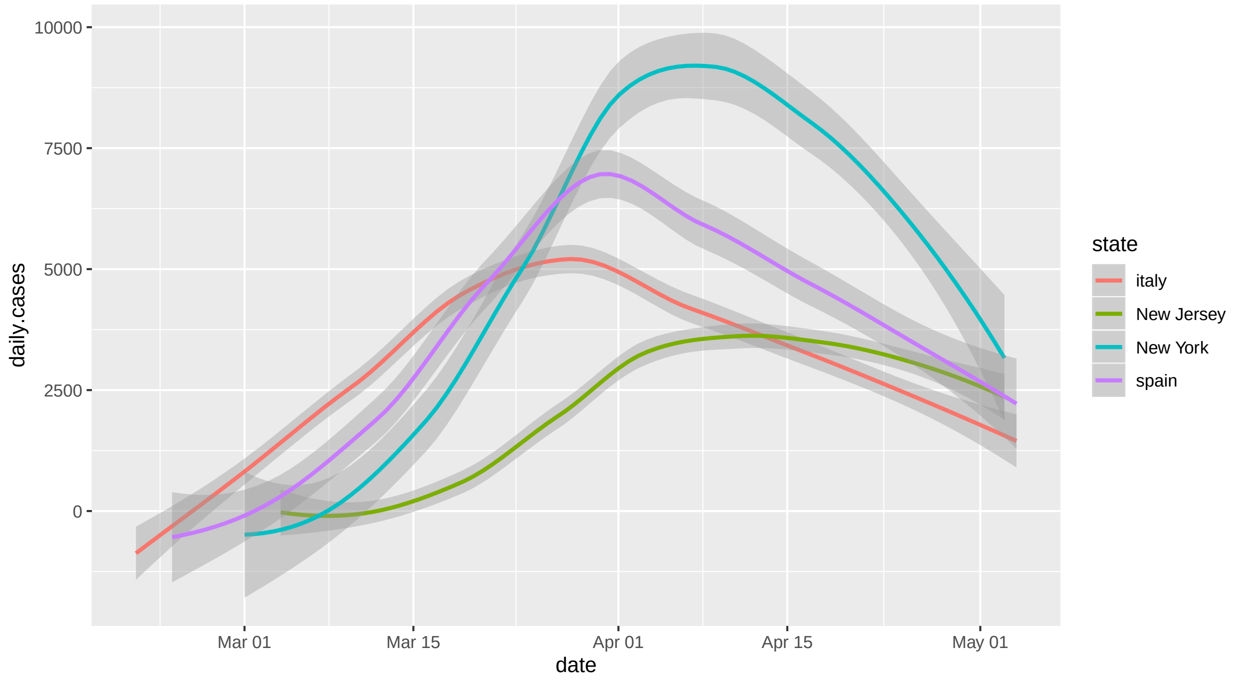
Let us look at the trend of covid19 data of US, Spain and Italy.

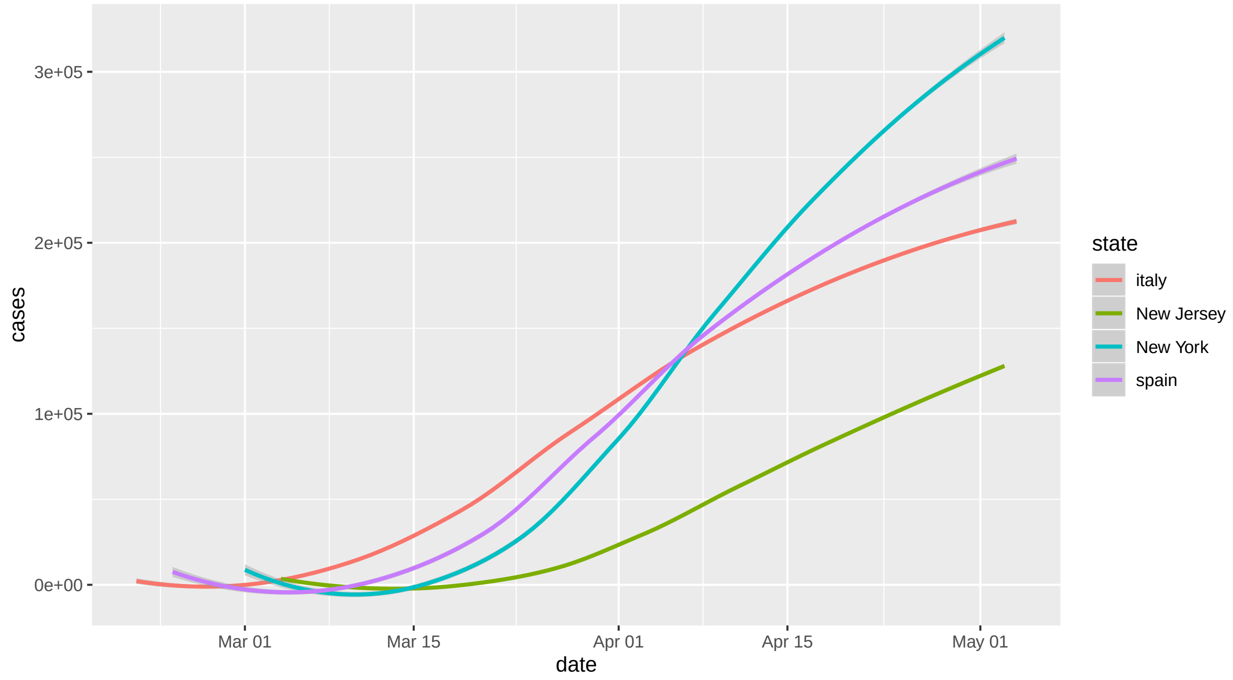




Initially when I proposed to compare italy, spain covid19 data to US, they numbers where similar but since then spain and italy had reached peak and their trend is now downwards. Whereas the US numbers still keep rising and looks like it reached peak recently.

But New York and New Jersey numbers are similar to ltaly and spain. Lets plot their graph and check it.





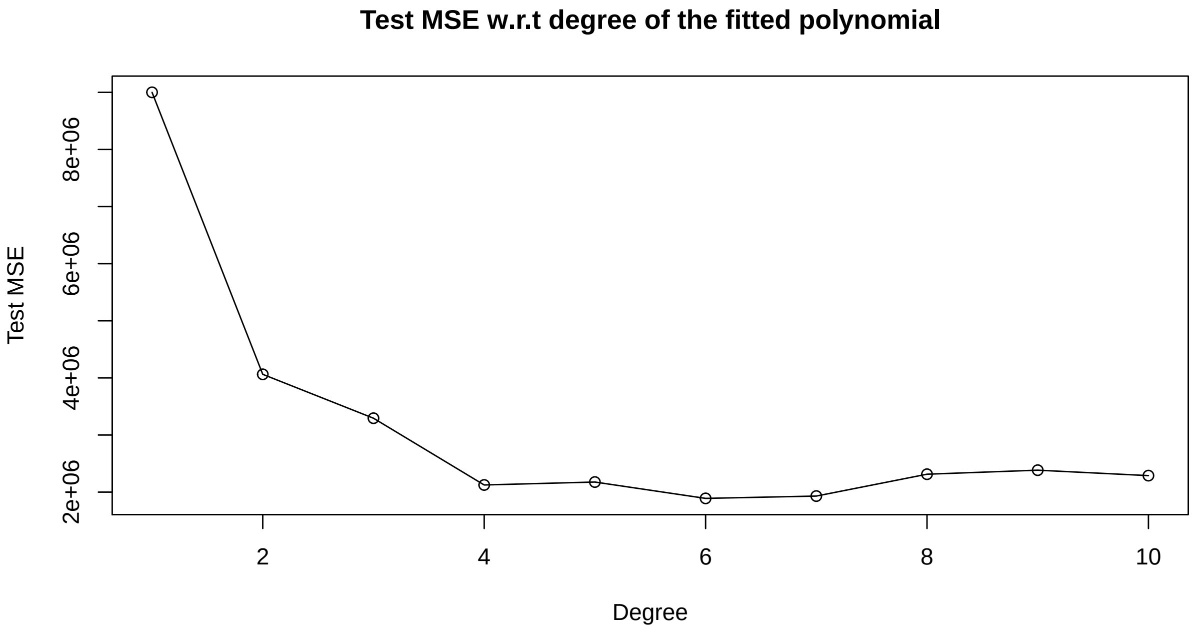
Here we can observe a better similarity in trends of number of cases. In all the cases peak has reached and downward trend has started.

I have mentioned in proposal I will predict number of US cases on Spain and Italy data but since then they have drastically changed. So now instead of what I proposed, I will try to fit a polynomial regression to New York data and predict their future values.

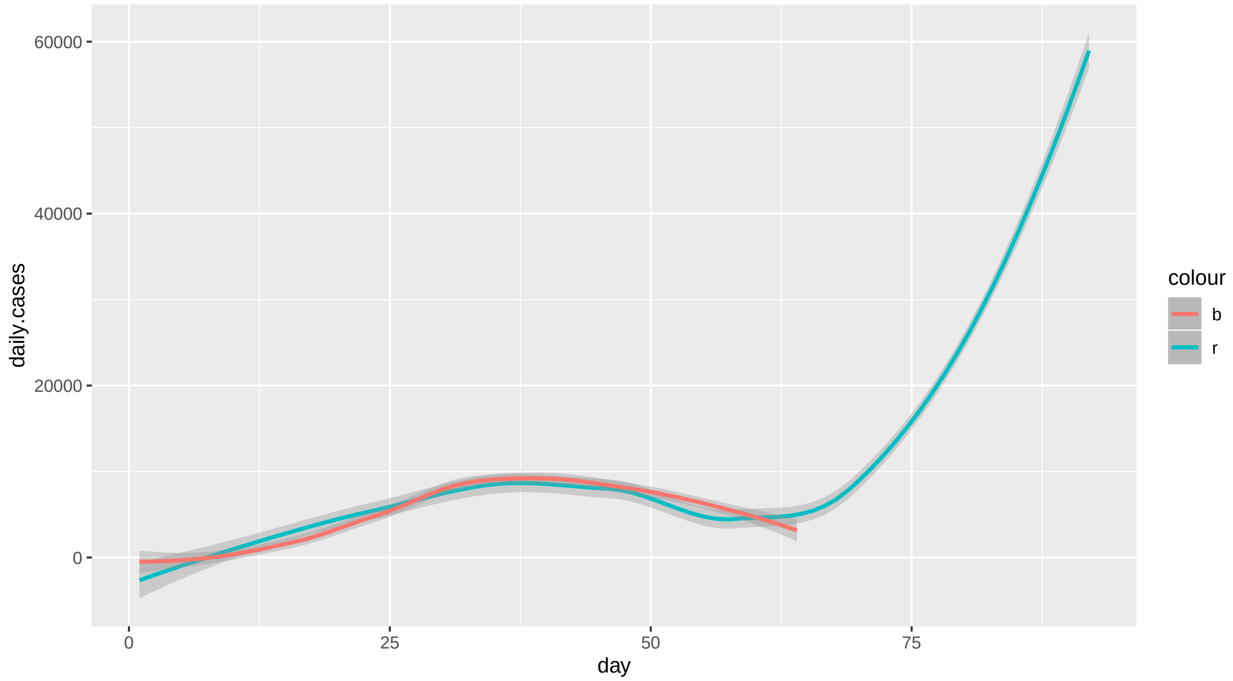
I used cross-validation to select the optimal degree for the polynomial

```
ny_data <- usa_state_data[usa_state_data$state=="New York",]
ny_data$day <- as.numeric(ny_data$date - min(ny_data$date) + 1)

mse <- rep(NA, 10)
for (i in 1:10) {
    fit <- glm(daily.cases ~ poly(day, i), data = ny_data)
    mse[i] <- cv.glm(ny_data, fit, K = 10)$delta[1]
}
plot(1:10, mse, xlab = "Degree", ylab = "Test MSE", type = "o" , main="Test MSE w.r.t degree of the fit
```

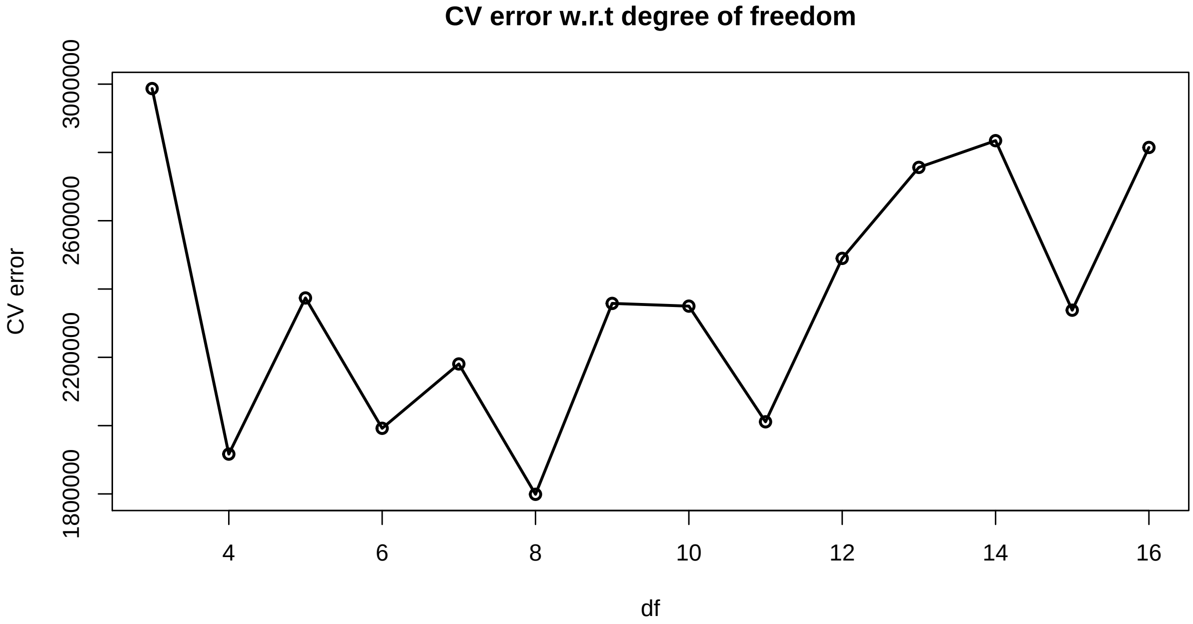


I have plotted test Mean squared error for each degree of polynomial. It shows that the CV error reduces as we increase degree from 1 to 4, stay same till degree 5, and then the starts increasing for higher degrees. We pick the polynomial degree as 4.



We can clearly see that though a polynomial regression fits well, it doesn't predict it well(Expecting a downward trend). Let's build a regression spline and see if it's predicts better.

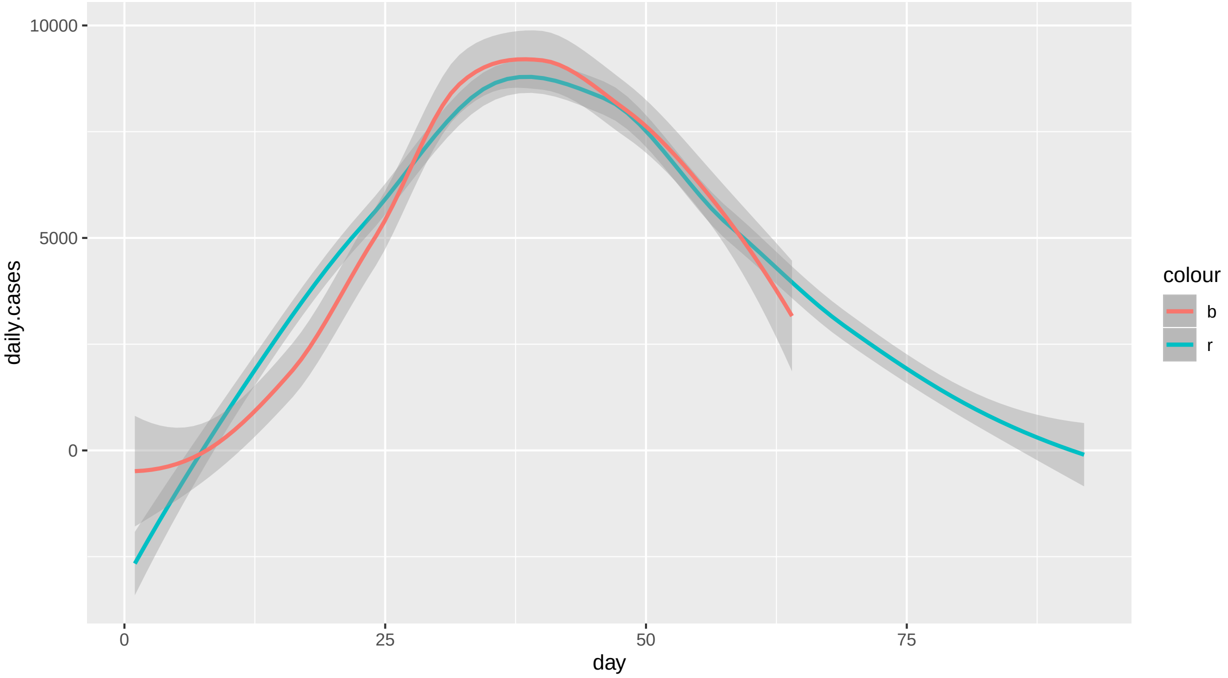
I used cross-validation to select the optimal degree for freedom of B-splines.



There is no visible trend in the plot, but CV error attains minimum at df=6, so we can choose 6 as the optimal degrees of freedom. We need to split day(1 to 64) into 5 parts to attain dof of 6. Using the ggplot of newyork daily cases, we can pick them as (4, 11, 30, 35, 42)

```r
bs.ny <- lm(daily.cases ~ bs(day, knots = c(4, 11, 30, 35, 42)),ny_data)

new.data <- data.frame(day = 1:92)
new.data$daily.cases <- predict(bs.ny, new.data)
ggplot(new.data, mapping = aes(x = day, y = daily.cases, color = "r")) + geom_smooth() +
  geom_smooth(ny_data, mapping = aes(x = day, y = daily.cases, color = "b"))
```



B-spline polynomial spline fits the data well and also predicts it better that the polynomial regression. This is an expected performance.

---

2) The effect of number of coronavirus cases on the economy by looking at S&P 500 stock market index.

a)The link https://finance.yahoo.com/quote/%5EGSPC/history?p=%5EGSPC contains daily values of S&P index.

I have tried using world trading data api but it was not supported in R. So as I suggested in proposal used yahoo finance website instead to get the S&P 500 index.

```r
str(smindex_data)
```

```
## 'data.frame':    100 obs. of  5 variables:
##  $ date : chr  "May 04, 2020" "May 01, 2020" "Apr 30, 2020" "Apr 29, 2020" ...
##  $ Open : num  2815 2869 2931 2918 2910 ...
##  $ High : num  2844 2869 2931 2955 2921 ...
##  $ Low  : num  2798 2822 2892 2912 2861 ...
##  $ close: num  2843 2831 2912 2940 2863 ...
```

But the date column in the scraped data is a different format so had to modify it before converting it to a date type column. I have used month.abb vector to do so.

```
month.abb
```

```
##  [1] "Jan" "Feb" "Mar" "Apr" "May" "Jun" "Jul" "Aug" "Sep" "Oct" "Nov" "Dec"
```

```
str(smindex_data)
```

```
## 'data.frame':    100 obs. of  5 variables:
##  $ date : Date, format: "2020-05-04" "2020-05-01" ...
##  $ Open : num  2815 2869 2931 2918 2910 ...
##  $ High : num  2844 2869 2931 2955 2921 ...
##  $ Low  : num  2798 2822 2892 2912 2861 ...
##  $ close: num  2843 2831 2912 2940 2863 ...
```

```
# saving the tidy version
# write.csv(smindex_data,"smiindex.csv",row.names = FALSE)
```

After modifying the date column, a tidy version is saved and is used in further analysis.

```
## [1] "The highest S&P index rose to is:3393.52 and the lowest value it reached is:2191.86"
```

With in a span of a month, from 21-Feb to 21-March, S&P index dropped from 3300 to 2200 points causing a panic. But since then it recovered and now it is around 2800.

```
## [1] "The highest gain in S&P index in a day is:141.03 and the highest drop is:150.22"
```

b)Data analysis on US covid19 and S&P 500 data

In order to find the effect of covid19 on S&P index, first we need to merge both the dataframes.I have used only Open and close index values while merging since only they were relevant. But merging will generate few NA values in the Open and Close columns since we don't have S&P index on few days(Weekends and public holidays).

```
##         date cases deaths daily.cases daily.deaths    Open   close
## 1 2020-01-21     1      0           0            0 3321.03 3320.79
## 2 2020-01-22     1      0           0            0 3330.02 3321.75
## 3 2020-01-23     1      0           0            0 3315.77 3325.54
## 4 2020-01-24     2      0           1            0 3333.10 3295.47
## 5 2020-01-25     3      0           1            0      NA      NA
## 6 2020-01-26     5      0           2            0      NA      NA
```

So I have used na.locf function which replaced each NA with the most recent non-NA prior to it.

```
df$Open <- na.locf(df$Open)
df$close <- na.locf(df$close)
head(df)
```

```
##         date cases deaths daily.cases daily.deaths    Open   close
## 1 2020-01-21     1      0           0            0 3321.03 3320.79
## 2 2020-01-22     1      0           0            0 3330.02 3321.75
## 3 2020-01-23     1      0           0            0 3315.77 3325.54
## 4 2020-01-24     2      0           1            0 3333.10 3295.47
## 5 2020-01-25     3      0           1            0 3333.10 3295.47
## 6 2020-01-26     5      0           2            0 3333.10 3295.47
```

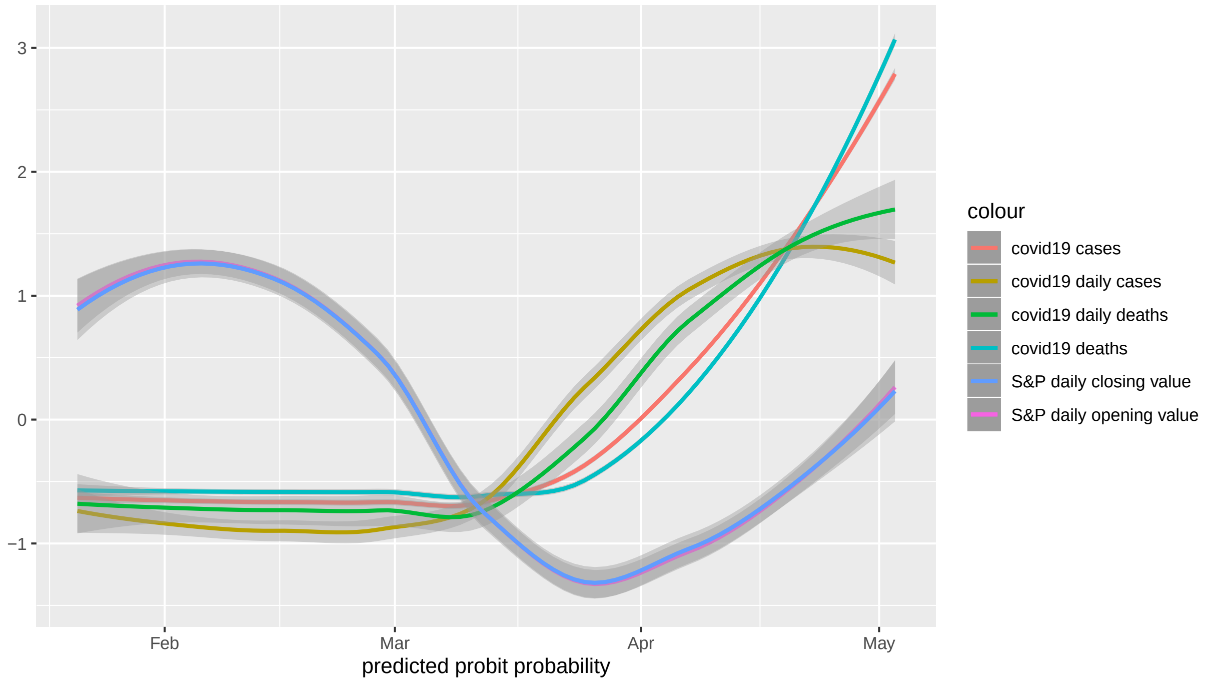
let's look at the corelation between the cases, deaths, S&P index open and close values.

```
cor(df[,c("cases","deaths","Open","close")])
```

```
##             cases     deaths       Open      close
## cases   1.0000000  0.9896053 -0.2759642 -0.2718773
## deaths  0.9896053  1.0000000 -0.2058848 -0.2046184
## Open   -0.2759642 -0.2058848  1.0000000  0.9849152
## close  -0.2718773 -0.2046184  0.9849152  1.0000000
```

Cases and deaths both are negatively correlated with Open and Close. It tells us that increase in cases/deaths has had a negative effect on S&P index.

```
ggplot(df, aes(x = df$date)) +
  geom_smooth(aes(y = scale(df$cases), colour = "covid19 cases")) +
  geom_smooth(aes(y = scale(df$deaths), colour = "covid19 deaths")) +
  geom_smooth(aes(y = scale(df$daily.cases), colour = "covid19 daily cases")) +
  geom_smooth(aes(y = scale(df$daily.deaths), colour = "covid19 daily deaths")) +
  geom_smooth(aes(y = scale(df$Open), colour = "S&P daily opening value")) +
  geom_smooth(aes(y = scale(df$close), colour = "S&P daily closing value")) +
  xlab("predicted probit probability") + ylab("")
```



Though these we don't observe a clear negative trend but we can see that S&P index intially dropped due to rise in covid19 cases but after reaching a low has started to rise. That might due to increased measures taken by govt or other reasons, which might have made the public to reassure about US economy.