Coronavirus Curves in selected countries (Using Python & R)

Dr. Sulove Koirala

5/6/2020

Introduction

We will be using data from JHU CSSE github. Johns Hopkins University has a well maintained data of Coronavirus cases in the world. I find it easy to clean the data using Python. Therefore, the first part of preparing the data will be done in Python.

Python

Importing the libraries

```
import pandas as pd
import numpy as np
```

Loading the Dataset

```
url = 'https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_1
data = pd.read_csv(url)
```

Cleaning the data

```
def cleandata(df_raw):
    df_cleaned=df_raw.melt(id_vars=['Province/State','Country/Region','Lat','Long'],value_name='Cases',
    df_cleaned=df_cleaned.set_index(['Country/Region','Province/State','Date'])
    return df_cleaned

# Clean all datasets
data=cleandata(data)
data.head()
```

```
##                                            Lat     Long  Cases
## Country/Region Province/State Date
## Afghanistan    NaN            1/22/20  33.0000  65.0000      0
## Albania        NaN            1/22/20  41.1533  20.1683      0
## Algeria        NaN            1/22/20  28.0339   1.6596      0
## Andorra        NaN            1/22/20  42.5063   1.5218      0
## Angola         NaN            1/22/20 -11.2027  17.8739      0
```

We have the data which is now easier to proceed doing analysis in R. So, let's export the csv which can be imported in R.

Exporting the CSV

```
data.to_csv(r'C:\Users\sulov\Desktop\PD\covid.csv', index = True)
```

We have exported the file 'covid.csv' to a folder named PD in desktop. Now, further analysis will be done using R.

R

Loading the packages

```
library(dplyr)
library(ggplot2)
library(lubridate)
library(reshape2)
```

Importing the dataset

```
covid <- read.csv("C:/Users/sulov/Desktop/PD/covid.csv")
```

Structuring data and adjusting date

```
str(covid)
```

```
## 'data.frame':    28462 obs. of  6 variables:
##  $ Country.Region: chr  "Afghanistan" "Albania" "Algeria" "Andorra" ...
##  $ Province.State: chr  "" "" "" "" ...
##  $ Date          : chr  "1/22/20" "1/22/20" "1/22/20" "1/22/20" ...
##  $ Lat           : num  33 41.2 28 42.5 -11.2 ...
##  $ Long          : num  65 20.17 1.66 1.52 17.87 ...
##  $ Cases         : int  0 0 0 0 0 0 0 0 0 0 ...
```

It is evident that the Date variable is in character form. We need it in the date form to proceed further. We will be using commands from lubridate package. Here, we also do not require Lat, Long and Provices. We are just focusing on the confirmed cases. Therefore, we are going to remove that columns also.

```
covid$Date = mdy(as.factor(covid$Date))
covid$Province.State = NULL
covid$Lat = NULL
covid$Long = NULL
tail(covid)
```

```
##              Country.Region       Date Cases
## 28457           South Sudan 2020-05-07    74
## 28458        Western Sahara 2020-05-07     6
## 28459 Sao Tome and Principe 2020-05-07   187
## 28460                 Yemen 2020-05-07    25
## 28461               Comoros 2020-05-07     8
## 28462            Tajikistan 2020-05-07   461
```

We are going to group countries based on their number of cases. Since, USA is an outlier we are going to show it separately. We are going to make three groups based on the number of cases.

Arranging Countries into Groups (Based on Number of Cases) Renaming the columns

```
colnames(covid) = c("Country", "Cases", "Date")
```

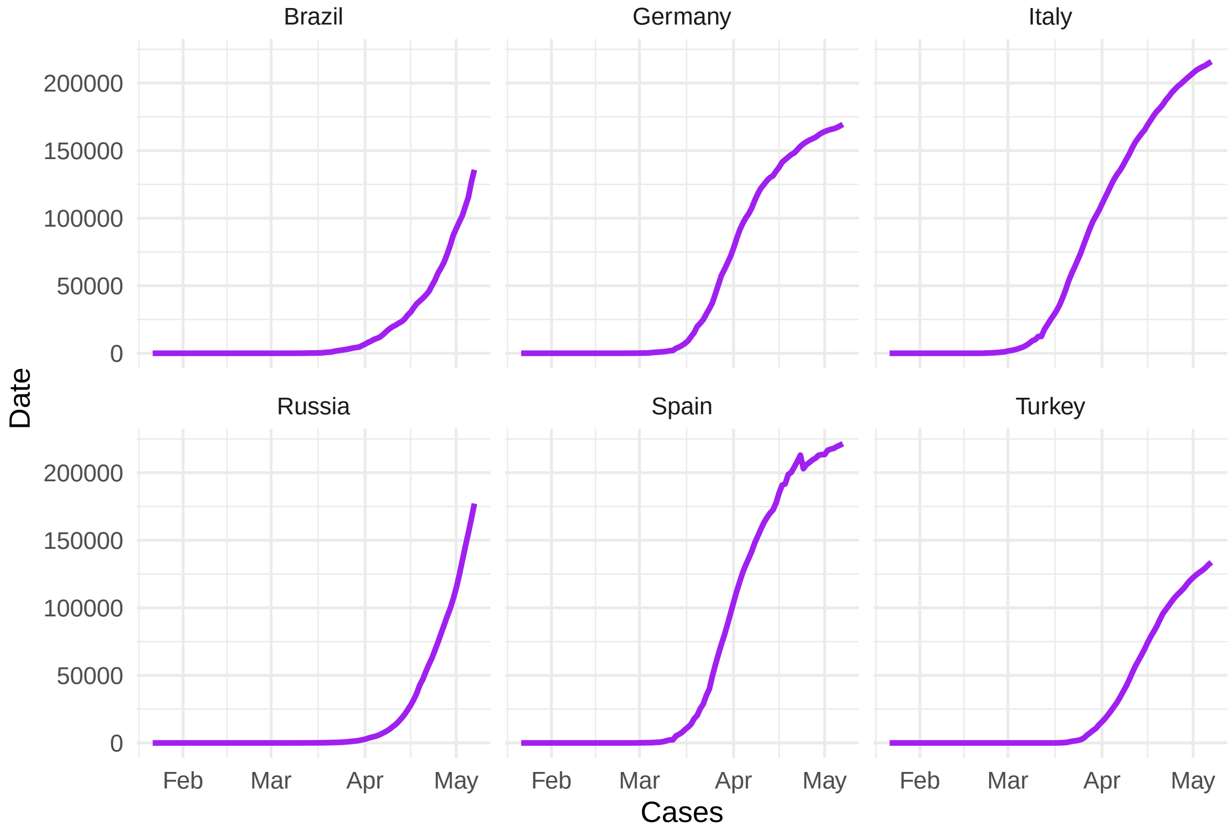
Also, we will be using melt command to restructure data, so that it will easier to use ggplot.

```
#Group A
groupA = filter(covid, Country == "Spain" | Country == "Italy" | Country =="Germany" | Country == "Russi
A = melt(groupA, id=c("Country", "Cases", "Date"))

#Group B
groupB = filter(covid, Country == "Saudi Arabia" | Country == "Switzerland" | Country =="Singapore" | Co
B = melt(groupB, id=c("Country", "Cases", "Date"))

#Group C
groupC = filter(covid, Country == "South Africa" | Country == "Norway" | Country =="Egypt" | Country ==
C = melt(groupC, id=c("Country", "Cases", "Date"))

#Group D
groupD = filter(covid, Country == "Thailand" | Country == "Greece" | Country =="New Zealand" | Country =
D = melt(groupD, id=c("Country", "Cases", "Date"))

#Group E
groupE = filter(covid, Country == "Sri Lanka" | Country == "Uruguay" | Country =="Kenya" | Country == "
E = melt(groupE, id=c("Country", "Cases", "Date"))

#Group F
groupF = filter(covid, Country == "Burma" | Country == "Benin" | Country =="Haiti" | Country == "Nepal"
F = melt(groupF, id=c("Country", "Cases", "Date"))
```

Visualization

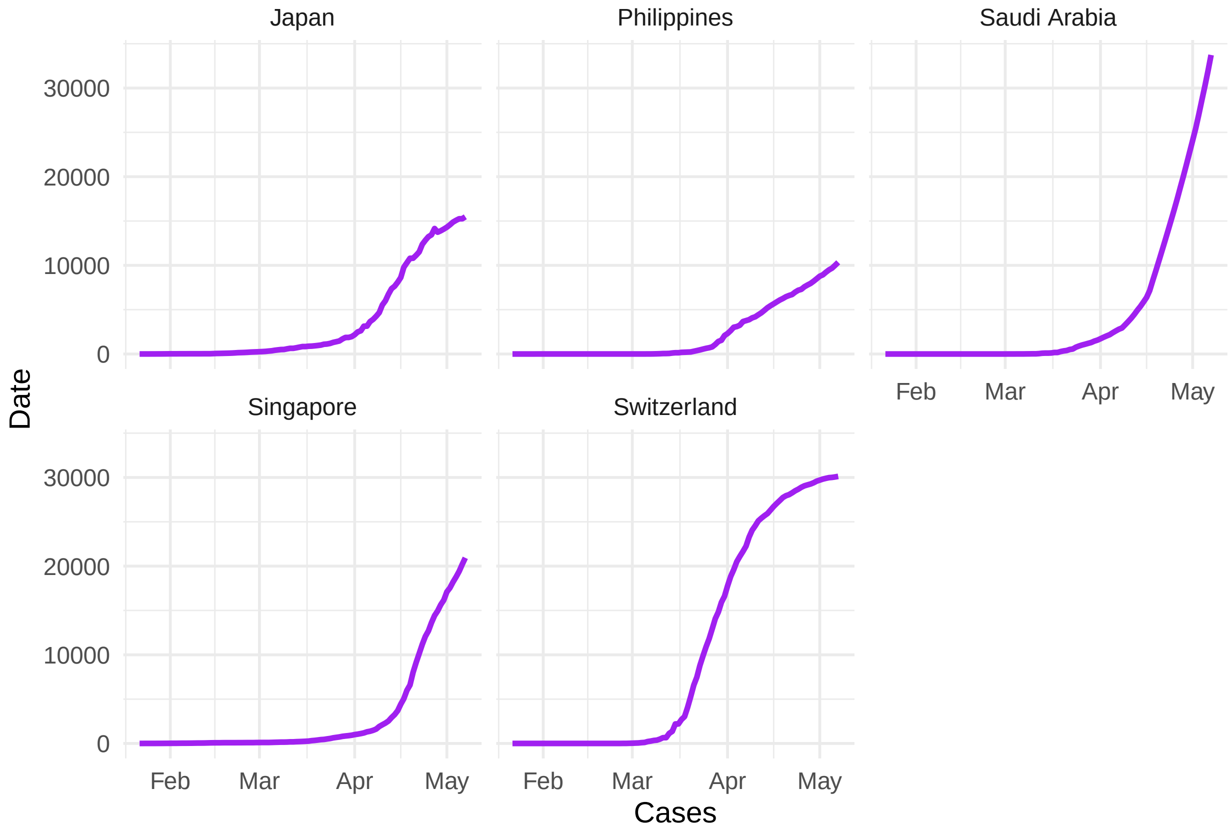
Group A

```
ggplot(A,aes(x = Cases, y = Date)) +
    geom_path(alpha = 2, size = 1, color = "purple") + facet_wrap(~Country)+
    theme_minimal()
```



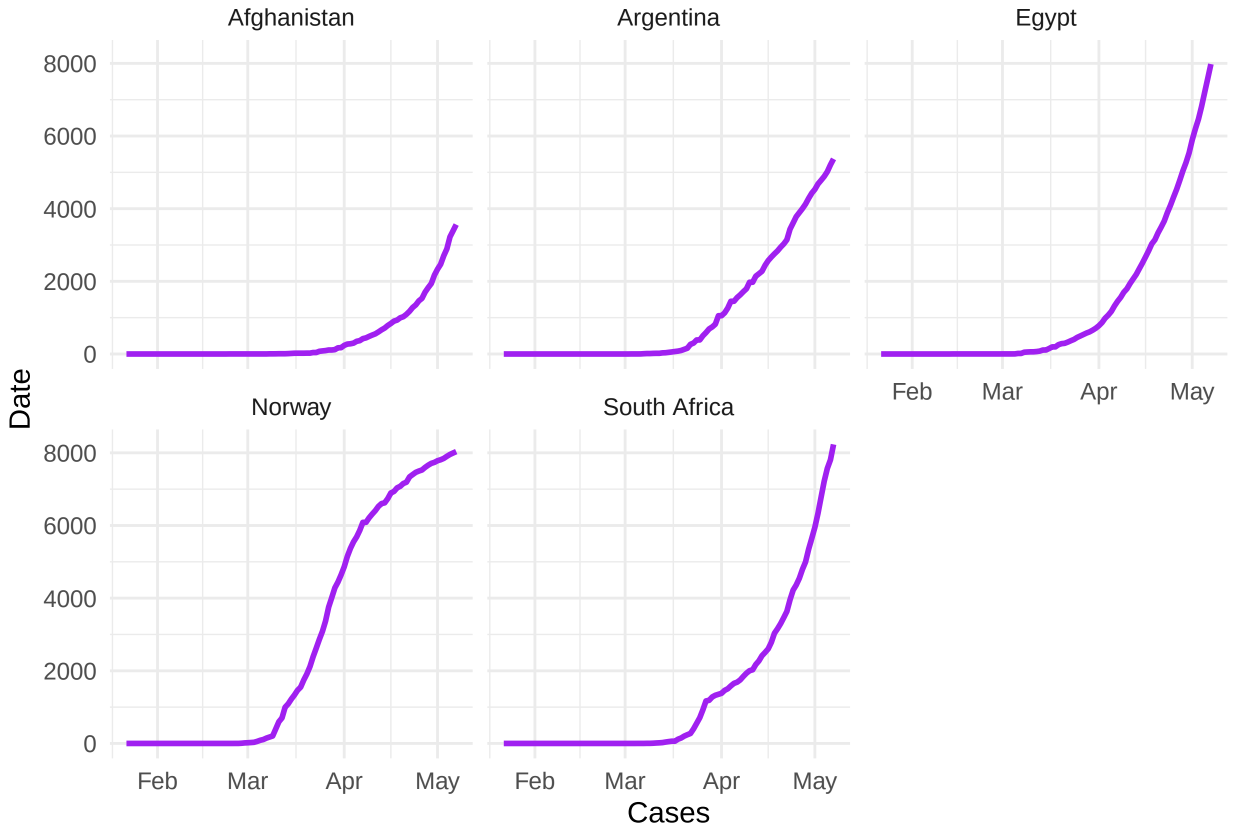
Group B

```
ggplot(B,aes(x = Cases, y = Date)) +
    geom_path(alpha = 2, size = 1, color = "purple") + facet_wrap(~Country)+
    theme_minimal()
```



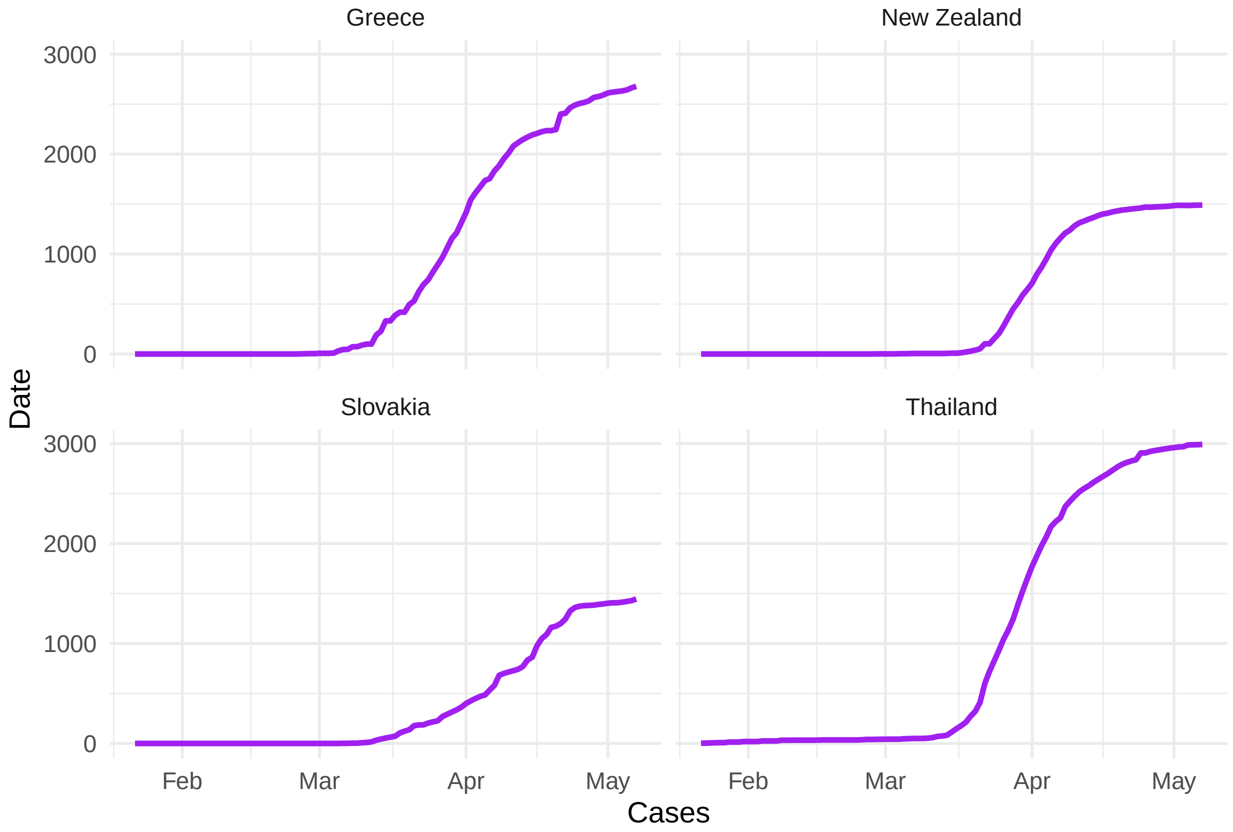
Group C

```r
ggplot(C,aes(x = Cases, y = Date)) +
    geom_path(alpha = 2, size = 1, color = "purple") + facet_wrap(~Country)+
    theme_minimal()
```



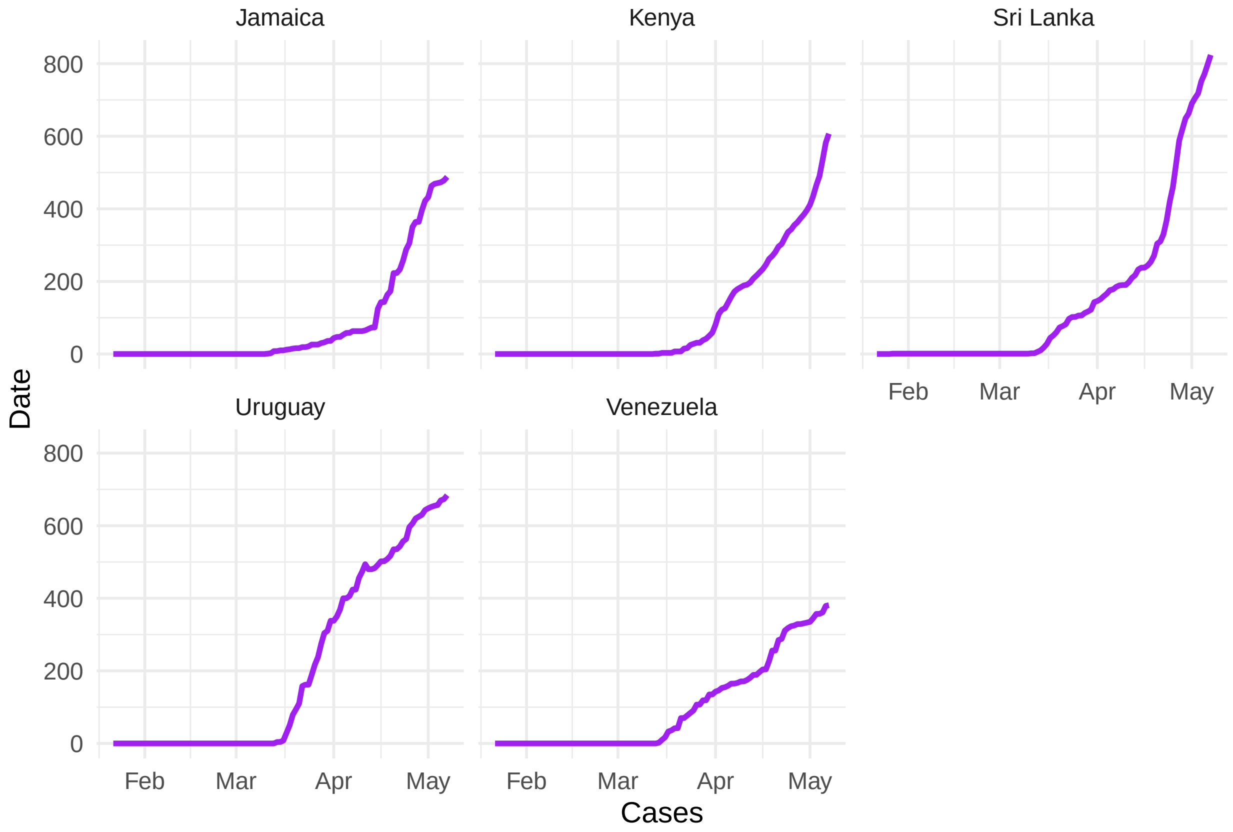
Group D

```
ggplot(D,aes(x = Cases, y = Date)) +
    geom_path(alpha = 2, size = 1, color = "purple") + facet_wrap(~Country)+
    theme_minimal()
```



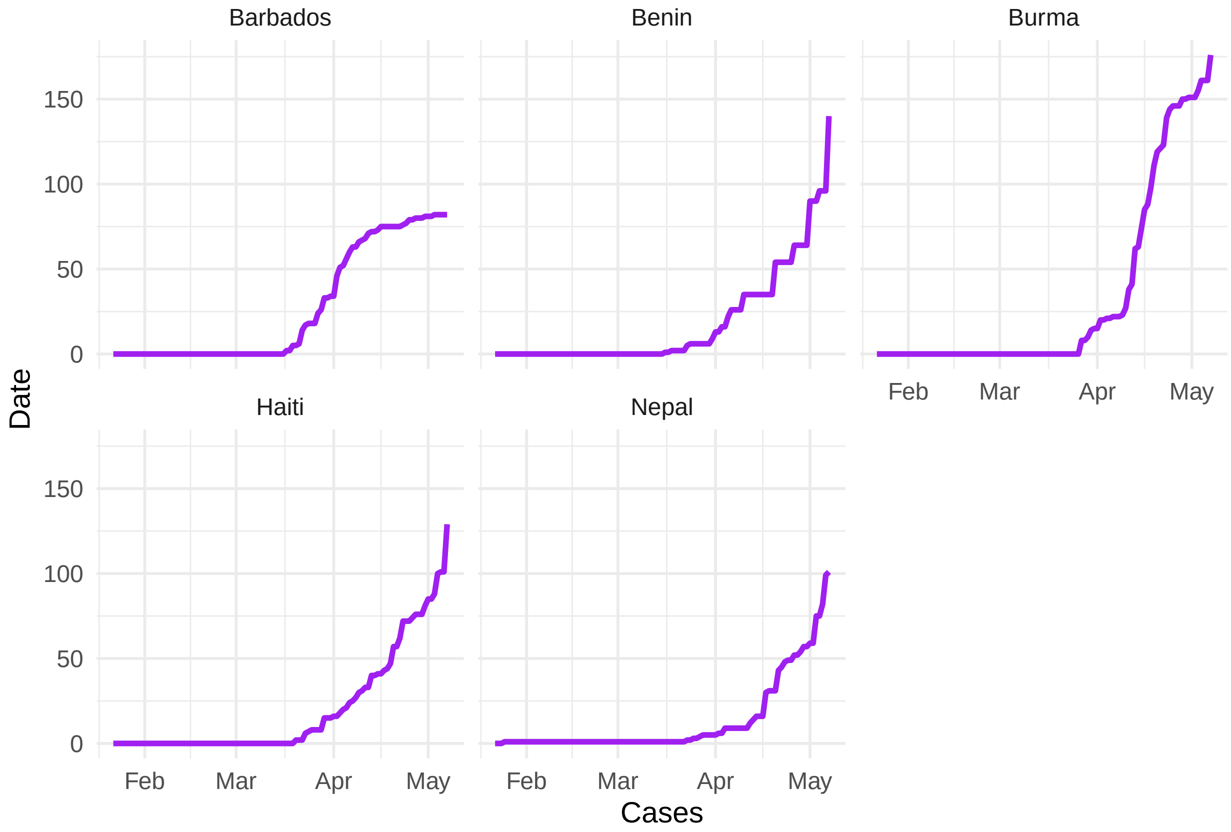
Group E

```
ggplot(E,aes(x = Cases, y = Date)) +
    geom_path(alpha = 2, size = 1, color = "purple") + facet_wrap(~Country)+
    theme_minimal()
```

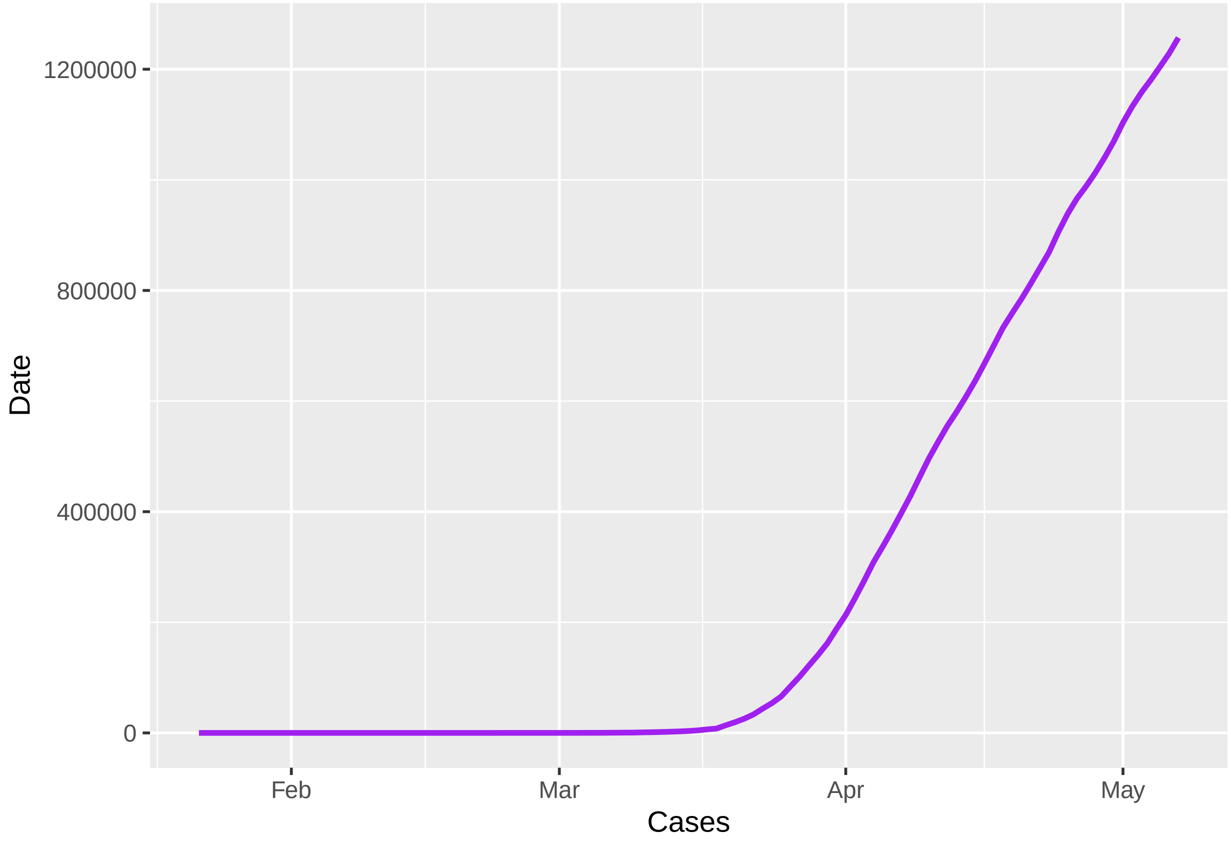


Group F

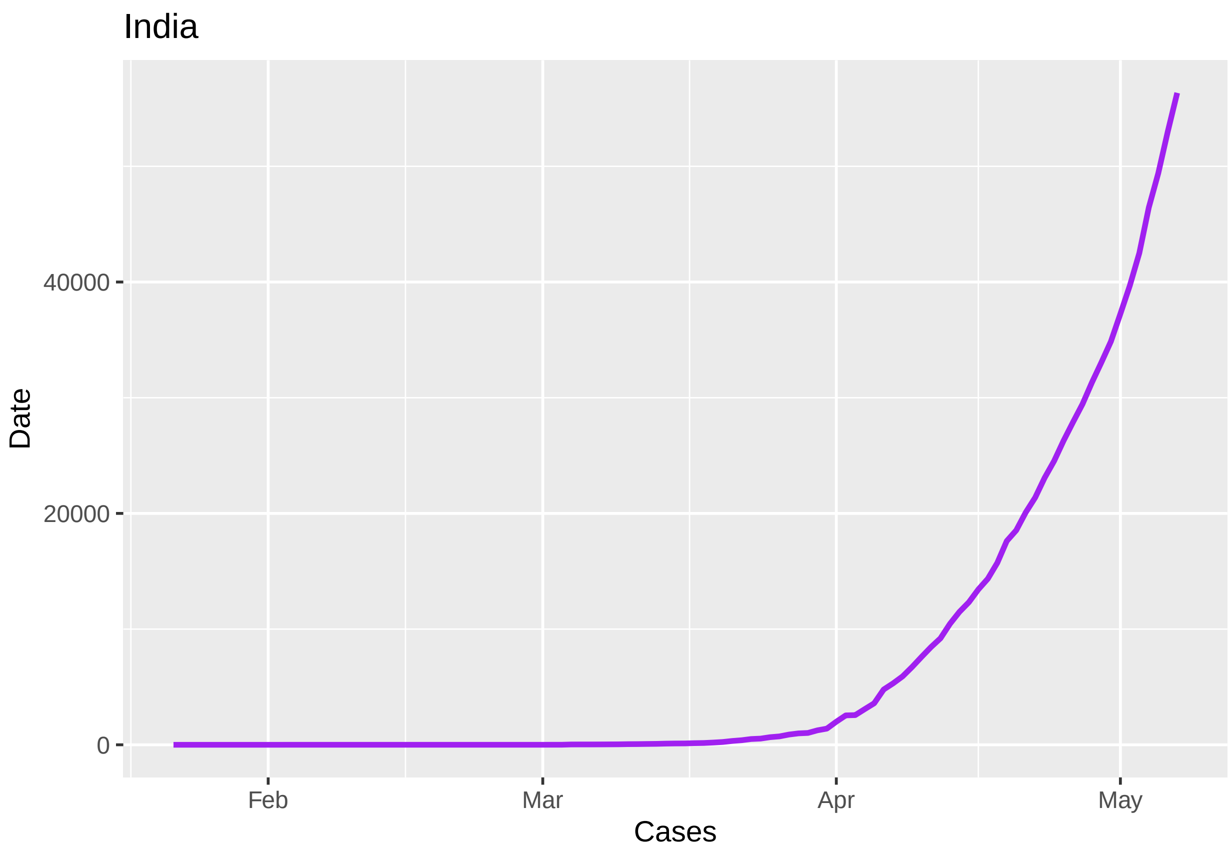
```
ggplot(F,aes(x = Cases, y = Date)) +
    geom_path(alpha = 2, size = 1, color = "purple") + facet_wrap(~Country)+
    theme_minimal()
```


Barbados
Benin
Burma
Haiti
Nepal
Date
150
100
50
0
Feb
Mar
Apr
May
Cases

India and USA



```
## $title
## [1] "United States"
##
## attr(,"class")
## [1] "labels"
```



Bibiliography

1. COVID-19 Dashboard by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University, or the COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University.

Contact Information

For Suggestions and other help Phone: +977-9840018421 Email: sulovekoirala@gmail.com