Final Project 2 - Reproducible Report on COVID19 Data

Giri Kunche

5/20/2021

Contents

Project Description

COVID-19 or the Coronavirus is a disease caused by SARS-CoV-2 virus. The first known case was identified in Wuhan,China in Dec 2019. This disease is spread all over the world and is currently prevalent in almost all countries.

As part of reproducible report, we will download and analyse COVID-19 data set. There are several websites available for the data set. We will use data set published by Center for Systems Science and Engineering at John Hopkins University. The data is available at github site : https://github.com/CSSEGISandData/COVID-19

Load Libraries

Load tidyverse and lubridate libraries.

Download Data

We will download four daily time series data for the Global confirmed cases, US confirmed cases, Global deaths and US deaths. Both Global and US data set has Province/State, Country, Latitude, Longitude and cases or deaths by date.

```
url_in<-"https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_
file_names<-c("time_series_covid19_confirmed_global.csv",
              "time_series_covid19_deaths_global.csv",
              "time_series_covid19_confirmed_US.csv",
              "time_series_covid19_deaths_US.csv")
urls<-str_c(url_in,file_names)
```

```
global_cases<-read_csv(urls[1])
global_deaths<-read_csv(urls[2])
US_cases<-read_csv(urls[3])
US_deaths<-read_csv(urls[4])
```

Review Raw Data

Let's get glimspe of data from global_cases, global_deaths, US_cases and US_deaths.

```
head(global_cases)
```

```
## # A tibble: 6 x 505
##   `Province/State` `Country/Region`     Lat   Long `1/22/20` `1/23/20` `1/24/20`
##   <chr>            <chr>              <dbl>  <dbl>     <dbl>     <dbl>     <dbl>
## 1 <NA>             Afghanistan         33.9  67.7          0         0         0
## 2 <NA>             Albania             41.2  20.2          0         0         0
## 3 <NA>             Algeria             28.0   1.66         0         0         0
## 4 <NA>             Andorra             42.5   1.52         0         0         0
## 5 <NA>             Angola             -11.2  17.9          0         0         0
## 6 <NA>             Antigua and Barbu~  17.1 -61.8          0         0         0
## # ... with 498 more variables: 1/25/20 <dbl>, 1/26/20 <dbl>, 1/27/20 <dbl>,
## #   1/28/20 <dbl>, 1/29/20 <dbl>, 1/30/20 <dbl>, 1/31/20 <dbl>, 2/1/20 <dbl>,
## #   2/2/20 <dbl>, 2/3/20 <dbl>, 2/4/20 <dbl>, 2/5/20 <dbl>, 2/6/20 <dbl>,
## #   2/7/20 <dbl>, 2/8/20 <dbl>, 2/9/20 <dbl>, 2/10/20 <dbl>, 2/11/20 <dbl>,
## #   2/12/20 <dbl>, 2/13/20 <dbl>, 2/14/20 <dbl>, 2/15/20 <dbl>, 2/16/20 <dbl>,
## #   2/17/20 <dbl>, 2/18/20 <dbl>, 2/19/20 <dbl>, 2/20/20 <dbl>, 2/21/20 <dbl>,
## #   2/22/20 <dbl>, 2/23/20 <dbl>, 2/24/20 <dbl>, 2/25/20 <dbl>, 2/26/20 <dbl>,
## #   2/27/20 <dbl>, 2/28/20 <dbl>, 2/29/20 <dbl>, 3/1/20 <dbl>, 3/2/20 <dbl>,
## #   3/3/20 <dbl>, 3/4/20 <dbl>, 3/5/20 <dbl>, 3/6/20 <dbl>, 3/7/20 <dbl>,
## #   3/8/20 <dbl>, 3/9/20 <dbl>, 3/10/20 <dbl>, 3/11/20 <dbl>, 3/12/20 <dbl>,
## #   3/13/20 <dbl>, 3/14/20 <dbl>, 3/15/20 <dbl>, 3/16/20 <dbl>, 3/17/20 <dbl>,
## #   3/18/20 <dbl>, 3/19/20 <dbl>, 3/20/20 <dbl>, 3/21/20 <dbl>, 3/22/20 <dbl>,
## #   3/23/20 <dbl>, 3/24/20 <dbl>, 3/25/20 <dbl>, 3/26/20 <dbl>, 3/27/20 <dbl>,
## #   3/28/20 <dbl>, 3/29/20 <dbl>, 3/30/20 <dbl>, 3/31/20 <dbl>, 4/1/20 <dbl>,
## #   4/2/20 <dbl>, 4/3/20 <dbl>, 4/4/20 <dbl>, 4/5/20 <dbl>, 4/6/20 <dbl>,
## #   4/7/20 <dbl>, 4/8/20 <dbl>, 4/9/20 <dbl>, 4/10/20 <dbl>, 4/11/20 <dbl>,
## #   4/12/20 <dbl>, 4/13/20 <dbl>, 4/14/20 <dbl>, 4/15/20 <dbl>, 4/16/20 <dbl>,
## #   4/17/20 <dbl>, 4/18/20 <dbl>, 4/19/20 <dbl>, 4/20/20 <dbl>, 4/21/20 <dbl>,
## #   4/22/20 <dbl>, 4/23/20 <dbl>, 4/24/20 <dbl>, 4/25/20 <dbl>, 4/26/20 <dbl>,
## #   4/27/20 <dbl>, 4/28/20 <dbl>, 4/29/20 <dbl>, 4/30/20 <dbl>, 5/1/20 <dbl>,
## #   5/2/20 <dbl>, 5/3/20 <dbl>, ...
```

```
head(global_deaths)
```

```
## # A tibble: 6 x 505
##   `Province/State` `Country/Region`     Lat   Long `1/22/20` `1/23/20` `1/24/20`
##   <chr>            <chr>              <dbl>  <dbl>     <dbl>     <dbl>     <dbl>
## 1 <NA>             Afghanistan         33.9  67.7          0         0         0
## 2 <NA>             Albania             41.2  20.2          0         0         0
## 3 <NA>             Algeria             28.0   1.66         0         0         0
## 4 <NA>             Andorra             42.5   1.52         0         0         0
## 5 <NA>             Angola             -11.2  17.9          0         0         0
## 6 <NA>             Antigua and Barbu~  17.1 -61.8          0         0         0
## # ... with 498 more variables: 1/25/20 <dbl>, 1/26/20 <dbl>, 1/27/20 <dbl>,
## #   1/28/20 <dbl>, 1/29/20 <dbl>, 1/30/20 <dbl>, 1/31/20 <dbl>, 2/1/20 <dbl>,
## #   2/2/20 <dbl>, 2/3/20 <dbl>, 2/4/20 <dbl>, 2/5/20 <dbl>, 2/6/20 <dbl>,
## #   2/7/20 <dbl>, 2/8/20 <dbl>, 2/9/20 <dbl>, 2/10/20 <dbl>, 2/11/20 <dbl>,
## #   2/12/20 <dbl>, 2/13/20 <dbl>, 2/14/20 <dbl>, 2/15/20 <dbl>, 2/16/20 <dbl>,
## #   2/17/20 <dbl>, 2/18/20 <dbl>, 2/19/20 <dbl>, 2/20/20 <dbl>, 2/21/20 <dbl>,
## #   2/22/20 <dbl>, 2/23/20 <dbl>, 2/24/20 <dbl>, 2/25/20 <dbl>, 2/26/20 <dbl>,
## #   2/27/20 <dbl>, 2/28/20 <dbl>, 2/29/20 <dbl>, 3/1/20 <dbl>, 3/2/20 <dbl>,
## #   3/3/20 <dbl>, 3/4/20 <dbl>, 3/5/20 <dbl>, 3/6/20 <dbl>, 3/7/20 <dbl>,
## #   3/8/20 <dbl>, 3/9/20 <dbl>, 3/10/20 <dbl>, 3/11/20 <dbl>, 3/12/20 <dbl>,
## #   3/13/20 <dbl>, 3/14/20 <dbl>, 3/15/20 <dbl>, 3/16/20 <dbl>, 3/17/20 <dbl>,
## #   3/18/20 <dbl>, 3/19/20 <dbl>, 3/20/20 <dbl>, 3/21/20 <dbl>, 3/22/20 <dbl>,
## #   3/23/20 <dbl>, 3/24/20 <dbl>, 3/25/20 <dbl>, 3/26/20 <dbl>, 3/27/20 <dbl>,
## #   3/28/20 <dbl>, 3/29/20 <dbl>, 3/30/20 <dbl>, 3/31/20 <dbl>, 4/1/20 <dbl>,
## #   4/2/20 <dbl>, 4/3/20 <dbl>, 4/4/20 <dbl>, 4/5/20 <dbl>, 4/6/20 <dbl>,
## #   4/7/20 <dbl>, 4/8/20 <dbl>, 4/9/20 <dbl>, 4/10/20 <dbl>, 4/11/20 <dbl>,
## #   4/12/20 <dbl>, 4/13/20 <dbl>, 4/14/20 <dbl>, 4/15/20 <dbl>, 4/16/20 <dbl>,
## #   4/17/20 <dbl>, 4/18/20 <dbl>, 4/19/20 <dbl>, 4/20/20 <dbl>, 4/21/20 <dbl>,
## #   4/22/20 <dbl>, 4/23/20 <dbl>, 4/24/20 <dbl>, 4/25/20 <dbl>, 4/26/20 <dbl>,
## #   4/27/20 <dbl>, 4/28/20 <dbl>, 4/29/20 <dbl>, 4/30/20 <dbl>, 5/1/20 <dbl>,
## #   5/2/20 <dbl>, 5/3/20 <dbl>, ...
```

```
head(US_cases)
```

```
## # A tibble: 6 x 512
##        UID iso2  iso3  code3  FIPS Admin2  Province_State Country_Region   Lat
##      <dbl> <chr> <chr> <dbl> <dbl> <chr>   <chr>          <chr>          <dbl>
## 1 84001001 US    USA     840  1001 Autauga Alabama        US              32.5
## 2 84001003 US    USA     840  1003 Baldwin Alabama        US              30.7
## 3 84001005 US    USA     840  1005 Barbour Alabama        US              31.9
## 4 84001007 US    USA     840  1007 Bibb    Alabama        US              33.0
## 5 84001009 US    USA     840  1009 Blount  Alabama        US              34.0
## 6 84001011 US    USA     840  1011 Bullock Alabama        US              32.1
## # ... with 503 more variables: Long_ <dbl>, Combined_Key <chr>, 1/22/20 <dbl>,
## #   1/23/20 <dbl>, 1/24/20 <dbl>, 1/25/20 <dbl>, 1/26/20 <dbl>, 1/27/20 <dbl>,
## #   1/28/20 <dbl>, 1/29/20 <dbl>, 1/30/20 <dbl>, 1/31/20 <dbl>, 2/1/20 <dbl>,
## #   2/2/20 <dbl>, 2/3/20 <dbl>, 2/4/20 <dbl>, 2/5/20 <dbl>, 2/6/20 <dbl>,
## #   2/7/20 <dbl>, 2/8/20 <dbl>, 2/9/20 <dbl>, 2/10/20 <dbl>, 2/11/20 <dbl>,
## #   2/12/20 <dbl>, 2/13/20 <dbl>, 2/14/20 <dbl>, 2/15/20 <dbl>, 2/16/20 <dbl>,
## #   2/17/20 <dbl>, 2/18/20 <dbl>, 2/19/20 <dbl>, 2/20/20 <dbl>, 2/21/20 <dbl>,
## #   2/22/20 <dbl>, 2/23/20 <dbl>, 2/24/20 <dbl>, 2/25/20 <dbl>, 2/26/20 <dbl>,
## #   2/27/20 <dbl>, 2/28/20 <dbl>, 2/29/20 <dbl>, 3/1/20 <dbl>, 3/2/20 <dbl>,
## #   3/3/20 <dbl>, 3/4/20 <dbl>, 3/5/20 <dbl>, 3/6/20 <dbl>, 3/7/20 <dbl>,
## #   3/8/20 <dbl>, 3/9/20 <dbl>, 3/10/20 <dbl>, 3/11/20 <dbl>, 3/12/20 <dbl>,
## #   3/13/20 <dbl>, 3/14/20 <dbl>, 3/15/20 <dbl>, 3/16/20 <dbl>, 3/17/20 <dbl>,
## #   3/18/20 <dbl>, 3/19/20 <dbl>, 3/20/20 <dbl>, 3/21/20 <dbl>, 3/22/20 <dbl>,
## #   3/23/20 <dbl>, 3/24/20 <dbl>, 3/25/20 <dbl>, 3/26/20 <dbl>, 3/27/20 <dbl>,
## #   3/28/20 <dbl>, 3/29/20 <dbl>, 3/30/20 <dbl>, 3/31/20 <dbl>, 4/1/20 <dbl>,
## #   4/2/20 <dbl>, 4/3/20 <dbl>, 4/4/20 <dbl>, 4/5/20 <dbl>, 4/6/20 <dbl>,
## #   4/7/20 <dbl>, 4/8/20 <dbl>, 4/9/20 <dbl>, 4/10/20 <dbl>, 4/11/20 <dbl>,
## #   4/12/20 <dbl>, 4/13/20 <dbl>, 4/14/20 <dbl>, 4/15/20 <dbl>, 4/16/20 <dbl>,
## #   4/17/20 <dbl>, 4/18/20 <dbl>, 4/19/20 <dbl>, 4/20/20 <dbl>, 4/21/20 <dbl>,
## #   4/22/20 <dbl>, 4/23/20 <dbl>, 4/24/20 <dbl>, 4/25/20 <dbl>, 4/26/20 <dbl>,
## #   4/27/20 <dbl>, 4/28/20 <dbl>, ...
```

```
head(US_deaths)
```

```
## # A tibble: 6 x 513
##        UID iso2  iso3  code3  FIPS Admin2  Province_State Country_Region   Lat
##      <dbl> <chr> <chr> <dbl> <dbl> <chr>   <chr>          <chr>          <dbl>
## 1 84001001 US    USA     840  1001 Autauga Alabama        US              32.5
## 2 84001003 US    USA     840  1003 Baldwin Alabama        US              30.7
## 3 84001005 US    USA     840  1005 Barbour Alabama        US              31.9
## 4 84001007 US    USA     840  1007 Bibb    Alabama        US              33.0
## 5 84001009 US    USA     840  1009 Blount  Alabama        US              34.0
## 6 84001011 US    USA     840  1011 Bullock Alabama        US              32.1
## # ... with 504 more variables: Long_ <dbl>, Combined_Key <chr>,
## #   Population <dbl>, 1/22/20 <dbl>, 1/23/20 <dbl>, 1/24/20 <dbl>,
## #   1/25/20 <dbl>, 1/26/20 <dbl>, 1/27/20 <dbl>, 1/28/20 <dbl>, 1/29/20 <dbl>,
## #   1/30/20 <dbl>, 1/31/20 <dbl>, 2/1/20 <dbl>, 2/2/20 <dbl>, 2/3/20 <dbl>,
## #   2/4/20 <dbl>, 2/5/20 <dbl>, 2/6/20 <dbl>, 2/7/20 <dbl>, 2/8/20 <dbl>,
## #   2/9/20 <dbl>, 2/10/20 <dbl>, 2/11/20 <dbl>, 2/12/20 <dbl>, 2/13/20 <dbl>,
## #   2/14/20 <dbl>, 2/15/20 <dbl>, 2/16/20 <dbl>, 2/17/20 <dbl>, 2/18/20 <dbl>,
## #   2/19/20 <dbl>, 2/20/20 <dbl>, 2/21/20 <dbl>, 2/22/20 <dbl>, 2/23/20 <dbl>,
## #   2/24/20 <dbl>, 2/25/20 <dbl>, 2/26/20 <dbl>, 2/27/20 <dbl>, 2/28/20 <dbl>,
## #   2/29/20 <dbl>, 3/1/20 <dbl>, 3/2/20 <dbl>, 3/3/20 <dbl>, 3/4/20 <dbl>,
## #   3/5/20 <dbl>, 3/6/20 <dbl>, 3/7/20 <dbl>, 3/8/20 <dbl>, 3/9/20 <dbl>,
## #   3/10/20 <dbl>, 3/11/20 <dbl>, 3/12/20 <dbl>, 3/13/20 <dbl>, 3/14/20 <dbl>,
## #   3/15/20 <dbl>, 3/16/20 <dbl>, 3/17/20 <dbl>, 3/18/20 <dbl>, 3/19/20 <dbl>,
## #   3/20/20 <dbl>, 3/21/20 <dbl>, 3/22/20 <dbl>, 3/23/20 <dbl>, 3/24/20 <dbl>,
## #   3/25/20 <dbl>, 3/26/20 <dbl>, 3/27/20 <dbl>, 3/28/20 <dbl>, 3/29/20 <dbl>,
## #   3/30/20 <dbl>, 3/31/20 <dbl>, 4/1/20 <dbl>, 4/2/20 <dbl>, 4/3/20 <dbl>,
## #   4/4/20 <dbl>, 4/5/20 <dbl>, 4/6/20 <dbl>, 4/7/20 <dbl>, 4/8/20 <dbl>,
## #   4/9/20 <dbl>, 4/10/20 <dbl>, 4/11/20 <dbl>, 4/12/20 <dbl>, 4/13/20 <dbl>,
## #   4/14/20 <dbl>, 4/15/20 <dbl>, 4/16/20 <dbl>, 4/17/20 <dbl>, 4/18/20 <dbl>,
## #   4/19/20 <dbl>, 4/20/20 <dbl>, 4/21/20 <dbl>, 4/22/20 <dbl>, 4/23/20 <dbl>,
## #   4/24/20 <dbl>, 4/25/20 <dbl>, 4/26/20 <dbl>, 4/27/20 <dbl>, ...
```

Raw data in all four data sets have each data as column. Also there are some nulls values in Province/State column in global_cases and global_deaths data set. US Data set has additional columns e.g UID, iso2, iso3,code3 etc.

Cleaning and Processing

Global Cases and Deaths Data Set

Both Global cases and Global death dataset has ‘Province/State’, ‘Country/Region’, ‘Lat’, ‘Long’ and data by dates. We will remove ‘Lat’ and ‘Long’ as we will not be using this for data analysis. Also we will move data columns to rows i.e we will have each date in separate rows using pivot_longer method. We will also join cases and death dataset into ‘global’ dataset.

```
global_cases<-global_cases %>%
  pivot_longer(cols = -c('Province/State','Country/Region',Lat,Long),names_to="date",values_to="cases")
global_deaths<-global_deaths %>%
  pivot_longer(cols = -c('Province/State','Country/Region',Lat,Long),names_to="date",values_to="deaths"
global <- global_cases %>% full_join(global_deaths) %>% rename(Country_Region ='Country/Region',Provinc
  mutate(date= mdy(date))
```

```
## Joining, by = c("Province/State", "Country/Region", "date")
```

Below is the output of global dataset post tidying and joining global_cases and global_deaths datasets.

```
head(global)
```

```
## # A tibble: 6 x 5
##   Province_State Country_Region date       cases deaths
##   <chr>          <chr>          <date>     <dbl>  <dbl>
## 1 <NA>           Afghanistan    2020-01-22     0      0
## 2 <NA>           Afghanistan    2020-01-23     0      0
## 3 <NA>           Afghanistan    2020-01-24     0      0
## 4 <NA>           Afghanistan    2020-01-25     0      0
## 5 <NA>           Afghanistan    2020-01-26     0      0
## 6 <NA>           Afghanistan    2020-01-27     0      0
```

Lets summarize the global dataset. We see the earliset case is on 22nd January 2020. Min cases and deaths is zero. There may be possibility of several records with zero cases/deaths.

```
summary(global)
```

```
##  Province_State     Country_Region          date                cases
##  Length:138276      Length:138276      Min.   :2020-01-22   Min.   :       0
##  Class :character   Class :character   1st Qu.:2020-05-26   1st Qu.:      93
##  Mode  :character   Mode  :character   Median :2020-09-28   Median :    1424
##                                        Mean   :2020-09-28   Mean   :  198538
##                                        3rd Qu.:2021-01-31   3rd Qu.:   30361
##                                        Max.   :2021-06-05   Max.   :33357205
##      deaths
##  Min.   :     0
##  1st Qu.:     1
##  Median :    22
##  Mean   :  4698
##  3rd Qu.:   531
##  Max.   :597377
```

We will filter null cases from global dataset, i.e we will consider only positive cases. Below is the summary after removing null cases.

```
#Filter only cases which are positive
global <-global %>% filter(cases >0)
summary(global)
```

```
##  Province_State     Country_Region          date                cases
##  Length:123978      Length:123978      Min.   :2020-01-22   Min.   :       1
##  Class :character   Class :character   1st Qu.:2020-06-25   1st Qu.:     260
##  Mode  :character   Mode  :character   Median :2020-10-21   Median :    2526
##                                        Mean   :2020-10-18   Mean   :  221435
##                                        3rd Qu.:2021-02-12   3rd Qu.:   44416
##                                        Max.   :2021-06-05   Max.   :33357205
##      deaths
##  Min.   :     0.0
##  1st Qu.:     2.0
##  Median :    45.5
##  Mean   :  5239.6
##  3rd Qu.:   752.0
##  Max.   :597377.0
```

US Cases and Deaths Data Set

We will follow similar process for US Cases and Deaths dataset. We will clean data and join into 'US' dataset. This dataset has UID, iso2,iso3,code3,FIPS,Admin2,Province_State, Country_Regon, Lat,Long, Combined_Key and dates. Also date is a character instead of date object. US deaths has population data. We will combined these datasets into 'US' .

```
US_cases<-US_cases %>%pivot_longer(cols=-(UID:Combined_Key),names_to="date",values_to="cases") %>% sele
US_deaths<-US_deaths %>%pivot_longer(cols=-(UID:Population),names_to="date",values_to="deaths") %>% sel
US <- US_cases %>% full_join(US_deaths)
```

```
## Joining, by = c("Admin2", "Province_State", "Country_Region", "Combined_Key", "date")
```

```
head(US)
```

```
## # A tibble: 6 x 8
##   Admin2  Province_State Country_Region Combined_Key date       cases Population
##   <chr>   <chr>          <chr>          <chr>        <date>     <dbl>      <dbl>
## 1 Autauga Alabama        US             Autauga, Al~ 2020-01-22     0      55869
## 2 Autauga Alabama        US             Autauga, Al~ 2020-01-23     0      55869
## 3 Autauga Alabama        US             Autauga, Al~ 2020-01-24     0      55869
## 4 Autauga Alabama        US             Autauga, Al~ 2020-01-25     0      55869
## 5 Autauga Alabama        US             Autauga, Al~ 2020-01-26     0      55869
## 6 Autauga Alabama        US             Autauga, Al~ 2020-01-27     0      55869
## # ... with 1 more variable: deaths <dbl>
```

###World Population Data We have US population data, but we dont have world population data in global dataset. This information is useful to comparative analysis between countries. Lets add population data and variable called combined_key into ‘global' dataset. We will download global population data from https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data from file UID_ISO_FIPS_LookUp_Table.csv. We will add population data to ‘global' data set by joining based on Province_State and Country_Region.

```
global<- global %>% unite("Combined_Key",c(Province_State,Country_Region),sep=", ",na.rm=TRUE,remove=FA
uid_lookup_url<-"https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/UI
uid<-read_csv(uid_lookup_url) %>% select(-c(Lat,Long_,Combined_Key,code3,iso2,iso3,Admin2))
```

```
##
## -- Column specification -------------------------------------------------------
## cols(
##   UID = col_double(),
##   iso2 = col_character(),
##   iso3 = col_character(),
##   code3 = col_double(),
##   FIPS = col_character(),
##   Admin2 = col_character(),
##   Province_State = col_character(),
##   Country_Region = col_character(),
##   Lat = col_double(),
##   Long_ = col_double(),
##   Combined_Key = col_character(),
##   Population = col_double()
## )
```

```
global <- global %>%
  left_join(uid,by=c("Province_State","Country_Region")) %>%
  select(-c(UID,FIPS)) %>%
  select(Province_State,Country_Region,date,cases,deaths,Population,Combined_Key)
```

Data Vizualization

###Covid Cases, Deaths by US and by US States (New York and Alaska) Lets analyse data of United States as a whole and for a given state. We will first group the data by Province_State, Country_Region and date and summarize the data by number of cases,deaths and Population. We will compute covid-19 deaths per million and add under column ‘deaths_per_mill'.

```
US_by_state <- US %>% group_by(Province_State,Country_Region,date) %>%
  summarise(cases=sum(cases),deaths=sum(deaths),Population=sum(Population)) %>%
  mutate(deaths_per_mill =deaths*1000000/Population)%>%
  select(Province_State,Country_Region,date,cases,deaths,deaths_per_mill,Population)%>%
  ungroup()
```

```
## `summarise()` has grouped output by 'Province_State', 'Country_Region'. You can override using the `
```

```
head(US_by_state)
```

```
## # A tibble: 6 x 7
##   Province_State Country_Region date       cases deaths deaths_per_mill
##   <chr>          <chr>          <date>     <dbl>  <dbl>           <dbl>
## 1 Alabama        US             2020-01-22     0      0               0
## 2 Alabama        US             2020-01-23     0      0               0
## 3 Alabama        US             2020-01-24     0      0               0
## 4 Alabama        US             2020-01-25     0      0               0
## 5 Alabama        US             2020-01-26     0      0               0
## 6 Alabama        US             2020-01-27     0      0               0
## # ... with 1 more variable: Population <dbl>
```

Get the US total deaths by summarizing US_by_state data set.

```
US_totals <- US_by_state %>% group_by(Country_Region,date) %>%
  summarise(cases=sum(cases),deaths=sum(deaths),Population=sum(Population)) %>%
  mutate(deaths_per_mill =deaths*1000000/Population) %>%
  select(Country_Region,date,cases,deaths,deaths_per_mill,Population)%>%
  ungroup()
```

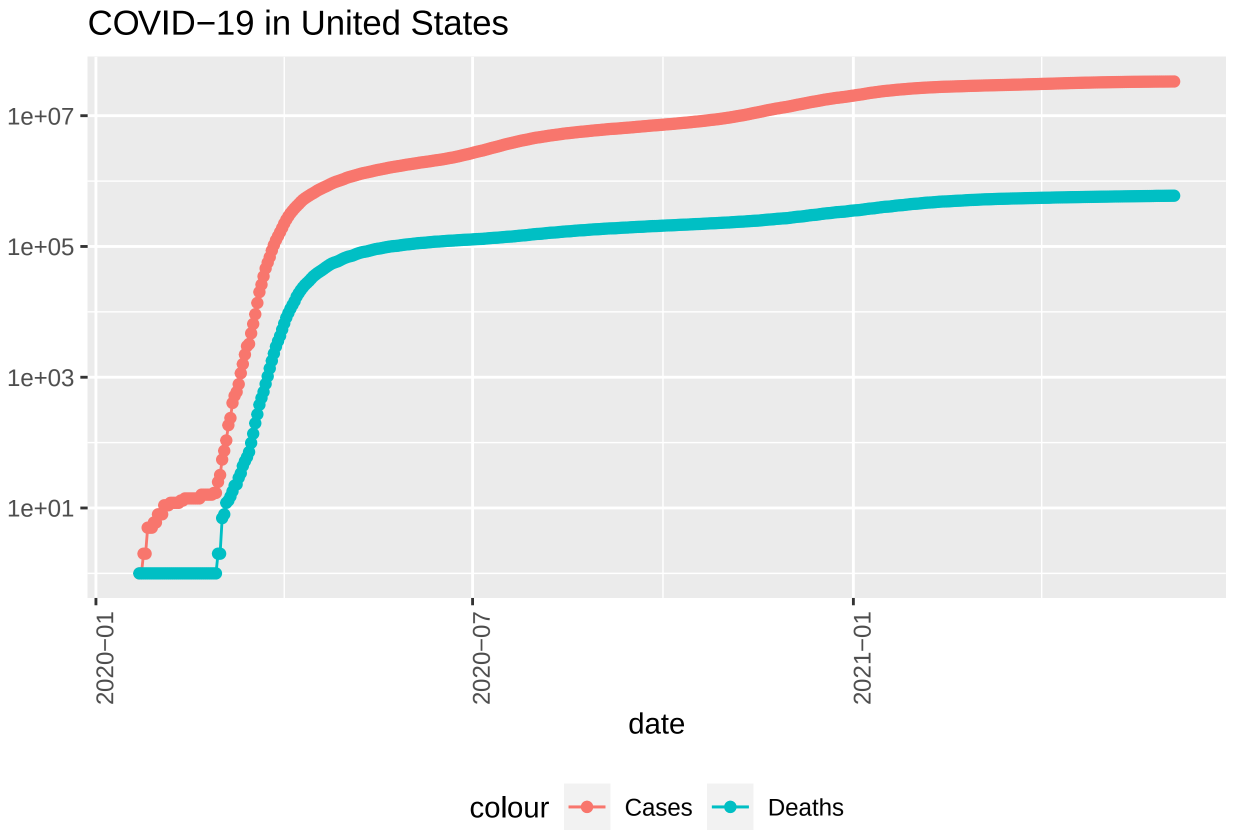
```
## `summarise()` has grouped output by 'Country_Region'. You can override using the `.groups` argument.
```

```
head(US_totals)
```

```
## # A tibble: 6 x 6
##   Country_Region date       cases deaths deaths_per_mill Population
##   <chr>          <date>     <dbl>  <dbl>           <dbl>      <dbl>
## 1 US             2020-01-22     1      1         0.00300  332875137
## 2 US             2020-01-23     1      1         0.00300  332875137
## 3 US             2020-01-24     2      1         0.00300  332875137
## 4 US             2020-01-25     2      1         0.00300  332875137
## 5 US             2020-01-26     5      1         0.00300  332875137
## 6 US             2020-01-27     5      1         0.00300  332875137
```

Below visualization shows the cases and deaths trend in the United States from the start of the reporting of the COVID-19.

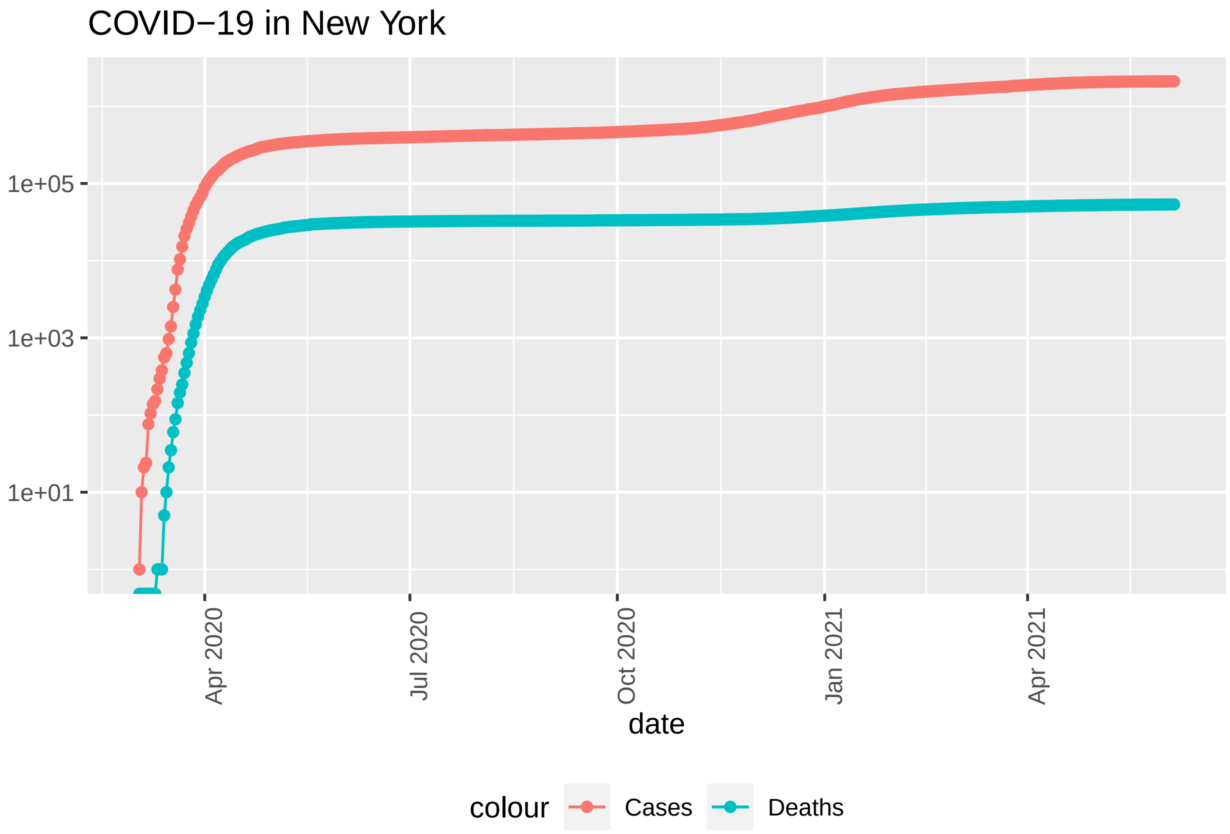
```
US_totals%>%
  filter(cases>0) %>%
  ggplot(aes(x=date,y=cases))+
  geom_line(aes(color="Cases"))+
  geom_point(aes(color="Cases"))+
  geom_line(aes(y=deaths,color="Deaths"))+
  geom_point(aes(y=deaths,color="Deaths"))+
  scale_y_log10()+
  theme(legend.position = "bottom",axis.text.x=element_text(angle=90))+
  labs(title="COVID-19 in United States",y=NULL)
```



Lets visualize cases, and deaths trend by state. We will analyse the trend in New York and Alaska.

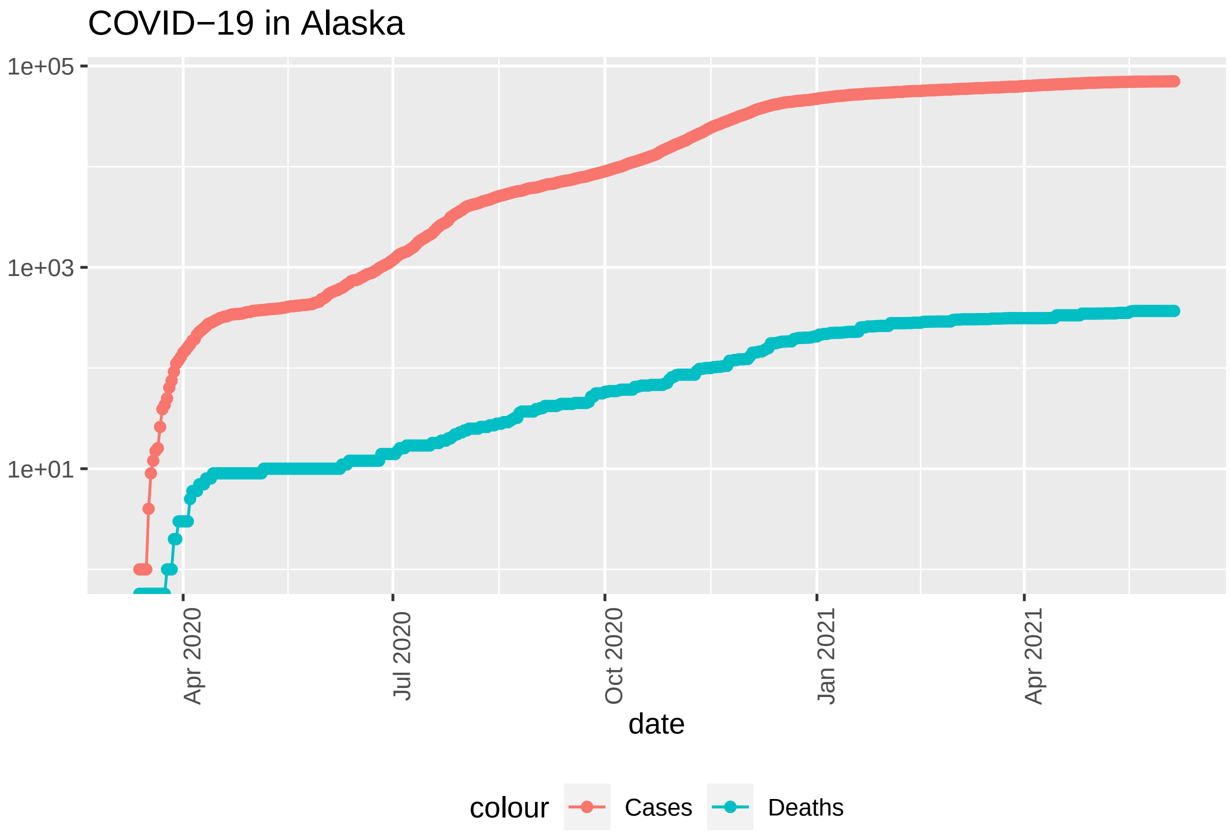
```
state1<-"New York"
US_by_state%>%
  filter(Province_State == state1) %>%
  filter(cases>0) %>%
  ggplot(aes(x=date,y=cases))+
  geom_line(aes(color="Cases"))+
  geom_point(aes(color="Cases"))+
  geom_line(aes(y=deaths,color="Deaths"))+
  geom_point(aes(y=deaths,color="Deaths"))+
  scale_y_log10()+
  theme(legend.position = "bottom",axis.text.x=element_text(angle=90))+
  labs(title=str_c("COVID-19 in ",state1),y=NULL)
```

```
## Warning: Transformation introduced infinite values in continuous y-axis

## Warning: Transformation introduced infinite values in continuous y-axis
```



```
state2<-"Alaska"
US_by_state%>%
  filter(Province_State == state2) %>%
  filter(cases>0) %>%
  ggplot(aes(x=date,y=cases))+
  geom_line(aes(color="Cases"))+
  geom_point(aes(color="Cases"))+
  geom_line(aes(y=deaths,color="Deaths"))+
  geom_point(aes(y=deaths,color="Deaths"))+
  scale_y_log10()+
  theme(legend.position = "bottom",axis.text.x=element_text(angle=90))+
  labs(title=str_c("COVID-19 in ",state2),y=NULL)
```

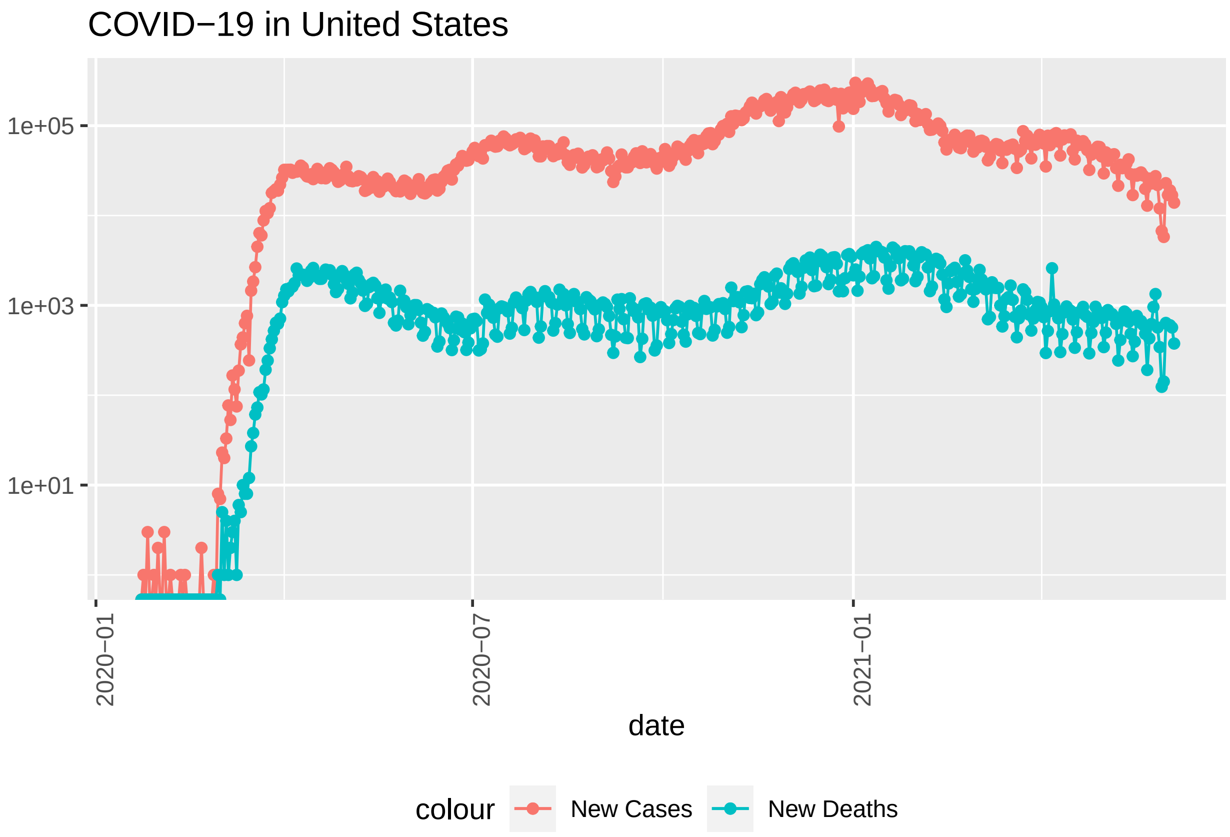
```
## Warning: Transformation introduced infinite values in continuous y-axis

## Warning: Transformation introduced infinite values in continuous y-axis
```



The visualization for both US and states shows that the cases peaked initially and levelled off from Jan 2021 onwards. We will deep dive on the new cases and check if the cases have really leveled off. In our data set we will add two column : new_cases and new_deaths. Below trend shows that the cases are dropping in both US and New York from Jan 2021 onwards. This may be due to Vaccinnation drive by both Federal and State goverments.

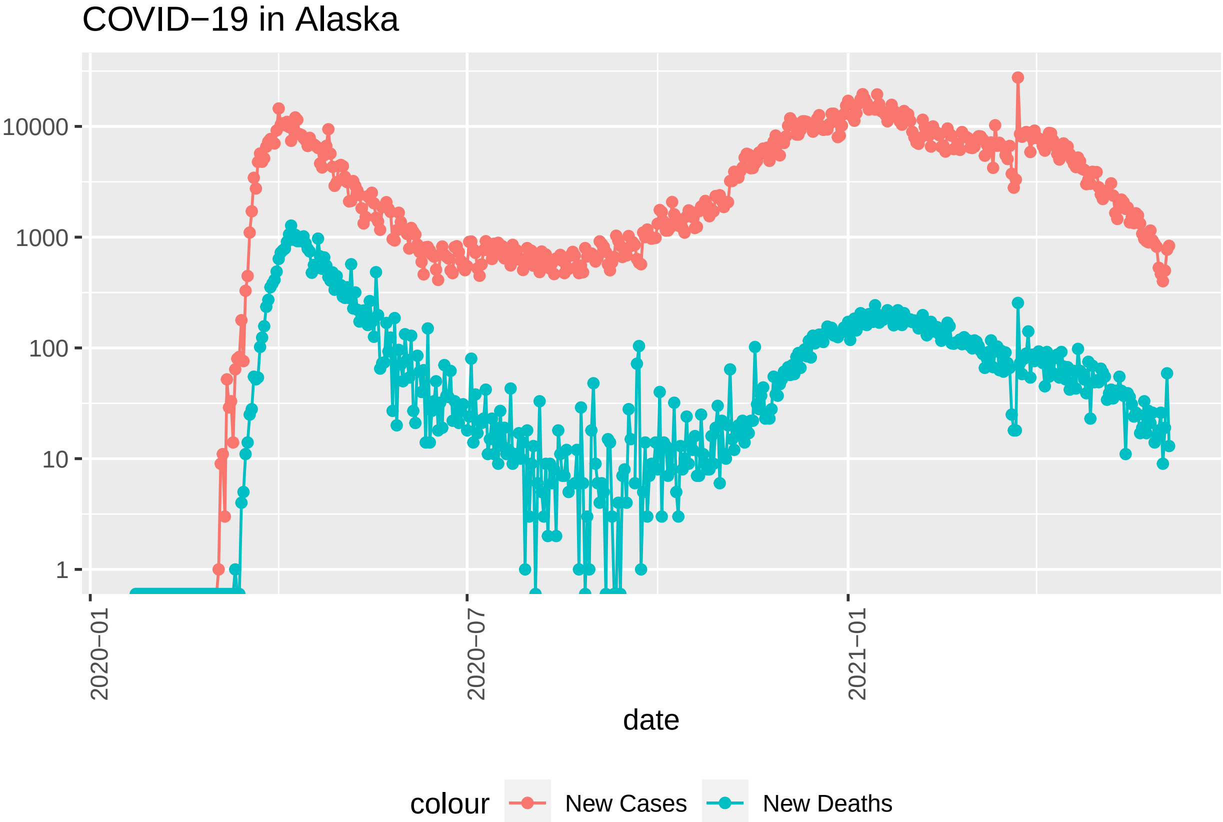
```
US_by_state <- US_by_state %>% mutate(new_cases=cases-lag(cases),new_deaths=deaths-lag(deaths))
US_totals <- US_totals %>% mutate(new_cases=cases-lag(cases),new_deaths=deaths-lag(deaths))

US_totals%>%
  ggplot(aes(x=date,y=new_cases))+
  geom_line(aes(color="New Cases"))+
  geom_point(aes(color="New Cases"))+
  geom_line(aes(y=new_deaths,color="New Deaths"))+
  geom_point(aes(y=new_deaths,color="New Deaths"))+
  scale_y_log10()+
  theme(legend.position = "bottom",axis.text.x=element_text(angle=90))+
  labs(title="COVID-19 in United States",y=NULL)
```

```
## Warning: Transformation introduced infinite values in continuous y-axis

## Warning: Transformation introduced infinite values in continuous y-axis

## Warning: Transformation introduced infinite values in continuous y-axis

## Warning: Transformation introduced infinite values in continuous y-axis
## Warning: Removed 1 row(s) containing missing values (geom_path).
## Warning: Removed 1 rows containing missing values (geom_point).

## Warning: Removed 1 row(s) containing missing values (geom_path).

## Warning: Removed 1 rows containing missing values (geom_point).
```



```
state1<-"New York"
US_by_state%>%
  filter(Province_State == state1) %>%
  ggplot(aes(x=date,y=new_cases))+
  geom_line(aes(color="New Cases"))+
  geom_point(aes(color="New Cases"))+
  geom_line(aes(y=new_deaths,color="New Deaths"))+
  geom_point(aes(y=new_deaths,color="New Deaths"))+
  scale_y_log10()+
  theme(legend.position = "bottom",axis.text.x=element_text(angle=90))+
  labs(title=str_c("COVID-19 in ",state2),y=NULL)
```

```
## Warning in self$trans$transform(x): NaNs produced

## Warning: Transformation introduced infinite values in continuous y-axis

## Warning in self$trans$transform(x): NaNs produced

## Warning: Transformation introduced infinite values in continuous y-axis

## Warning in self$trans$transform(x): NaNs produced

## Warning: Transformation introduced infinite values in continuous y-axis

## Warning in self$trans$transform(x): NaNs produced

## Warning: Transformation introduced infinite values in continuous y-axis
## Warning: Removed 1 row(s) containing missing values (geom_path).

## Warning: Removed 1 rows containing missing values (geom_point).

## Warning: Removed 1 row(s) containing missing values (geom_path).

## Warning: Removed 5 rows containing missing values (geom_point).
```


COVID−19 in Alaska
date
colour
New Cases
New Deaths