Cleaning COVID-19 Data from Johns Hopkins

Bill Kayser

I'm Initialize the Notebook loading the following required libraries.

```r
library(tidyverse)
library(lubridate)
library(scales)
library(wpp2019)
# library(tidycensus)
```

Load Datasets

We are pulling in data from the Johns Hopkins COVID-19 repository as well as world population data from wpp2019 and US Census information on county populations

COVID-19 Data

Let's update the CV data directly from the Johns Hopkins git repo. The first thing to do is check for any daily updates by updating the git submodule:

```r
system('git submodule update --remote COVID-19')
theme_set(theme_light())
```

Load the timeseries data. We'll need to convert the table from a wide format where the time series is in columns to a narrow format where each observation is a single datapoint for a day.

This function will process the different timeseries files uniformly.

```r
read_and_clean <- function(infile) {
  # Read the data
  cv.wide <- read.csv(infile)
  if ('Country.Region' %in% names(cv.wide)) {
    # This is the countries data so column names need to be adjusted
    cv.wide <- rename(cv.wide,
                      Province_State=Province.State,
                      Country_Region=Country.Region,
                      Long_=Long) %>%
      mutate(Admin2=NA,
             Combined_Key=str_c(Country_Region, ", ", Province_State))
  }
  # Identify the date columns so we can gather them
  datecols <- grep('^X', names(cv.wide))

  # Gather the date columns into a single pair of columns, "date" and "count"
  # This converts the data from a wide format to a narrow format more amenable to
  # graphing and reshaping
  gather(cv.wide, key='date', value='count', datecols) %>%
    select(State=Province_State, County=Admin2, Country=Country_Region, Lat, Long=Long_, Key=Combined_K
    mutate(Date=mdy(str_sub(Date, 2)),
           County=as.character(County),
           State=as.character(State)) %>%
    as_tibble()
}
```

Load US Timeseries Data

```
confirmed <- read_and_clean('COVID-19/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_
deaths <- read_and_clean('COVID-19/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_dea
```

Load International Timeseries Data

```
countries.confirmed <- read_and_clean('COVID-19/csse_covid_19_data/csse_covid_19_time_series/time_series
countries.deaths <- read_and_clean('COVID-19/csse_covid_19_data/csse_covid_19_time_series/time_series_c
```

Load Census Data

This is how you could get data directly using the US Census bureau APIs but the data appears to be too fine grained. We can use this later if we need more detailed census information.

```
get_estimates(product="population", key=Sys.getenv('CENSUS_KEY'))
```

This is just the static datafile of summary census statistics. Good enough for now.

```
counties <- read.csv('https://www2.census.gov/programs-surveys/popest/datasets/2010-2019/counties/totals
  select(Region=REGION,
         State=STNAME,
         County=CTYNAME,
         Population=POPESTIMATE2019,
         Annual.Deaths=DEATHS2019) %>%
  mutate(County=str_match(County, "^(.*) County")[,2],
         State=as.character(State)) %>%
  filter(!is.na(County))
```

Prepare Data

Join Tables

Join the US confirmed cases and deaths into a single table, `cvdata.us`.

```
cvdata.us <- left_join(confirmed,
                       select(deaths, Key, Date, Count),
                       by=c('Key', 'Date')) %>%
  rename(Confirmed = Count.x,
         Deaths = Count.y)
```

Join the census data to the COVID data table.

```
cvdata.us <- left_join(cvdata.us, counties, by=c('State', 'County'))
```

Join the confirmed and deaths for country data into a single table, `cvdata.i18n`.

```
cvdata.i18n <- left_join(countries.confirmed,
                         select(countries.deaths, Key, Date, Deaths=Count),
                         by=c('Key', 'Date')) %>%
  rename(Confirmed=Count)
```

Add in First Derivative

It will be interesting to study not just the day to day numbers, but the change in the numbers from day to day. I will manually calculate this first derivative and store the values in colums appended with `.Diff`.

```
# sort by locale the date
cvdata.us <- arrange(cvdata.us, Key, Date)
boundaries <- which(cvdata.us$Key[2:nrow(cvdata.us)] != cvdata.us$Key[1:nrow(cvdata.us)-1])

# Calculate differential
X1 <- c(cvdata.us$Confirmed)
X0 <- c(0, cvdata.us$Confirmed)[1:length(X1)]
diff <- X1 - X0

diff[boundaries + 1] <- 0

cvdata.us$Confirmed.Diff <- diff

X1 <- c(cvdata.us$Deaths)
X0 <- c(0, cvdata.us$Deaths)[1:length(X1)]
diff <- X1 - X0
diff[boundaries + 1] <- 0

cvdata.us$Deaths.Diff <- diff
```

Derive Additional Tables

Group data by state and save into table `cvdata.us.by_state`.

```
cvdata.us.by_state <- cvdata.us %>%
  group_by(State, Date) %>%
  summarize(Confirmed = sum(Confirmed),
            Confirmed.Diff = sum(Confirmed.Diff),
            Deaths = sum(Deaths),
            Deaths.Diff = sum(Deaths.Diff),
            Region = first(Region),
            Population = sum(Population),
            Annual.Deaths = sum(Annual.Deaths)) %>%
  ungroup(State, Date)
```

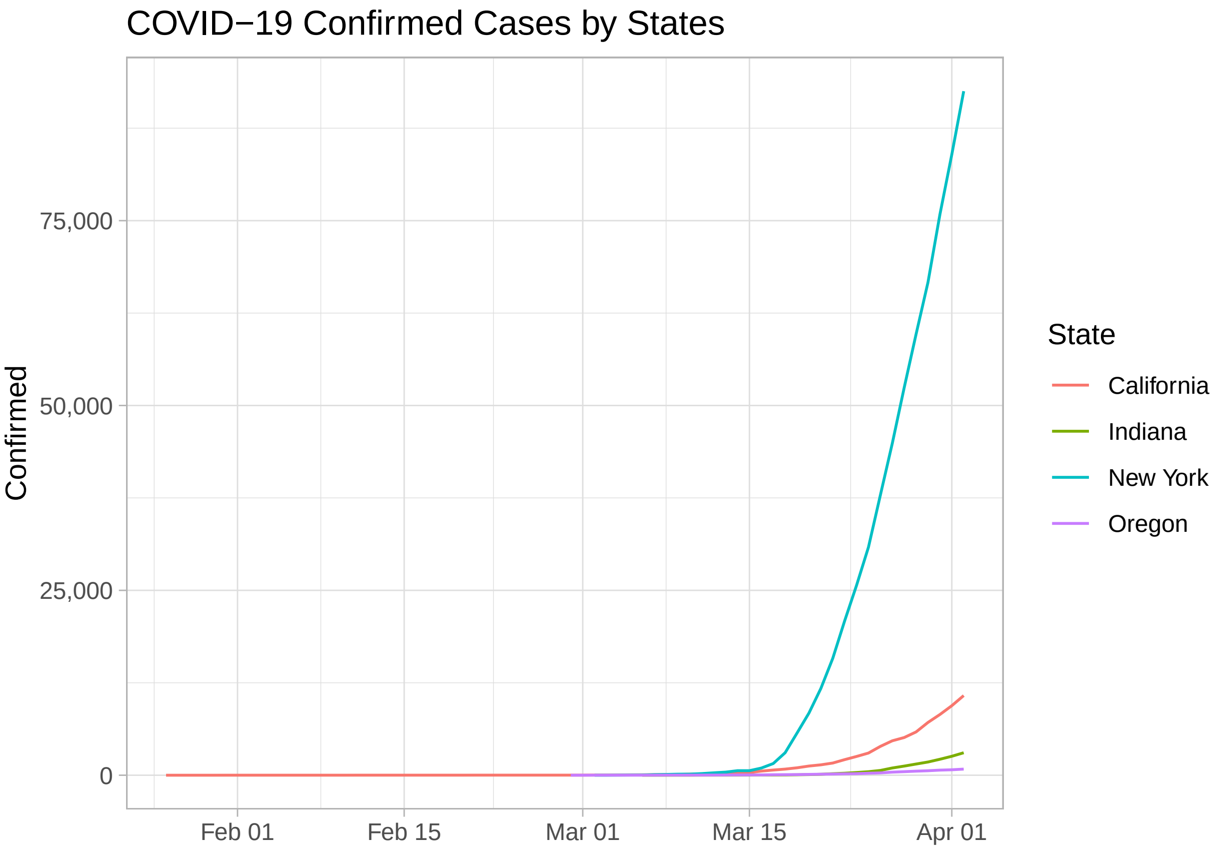
Create a single timeseries table just for Italy called `italy`.

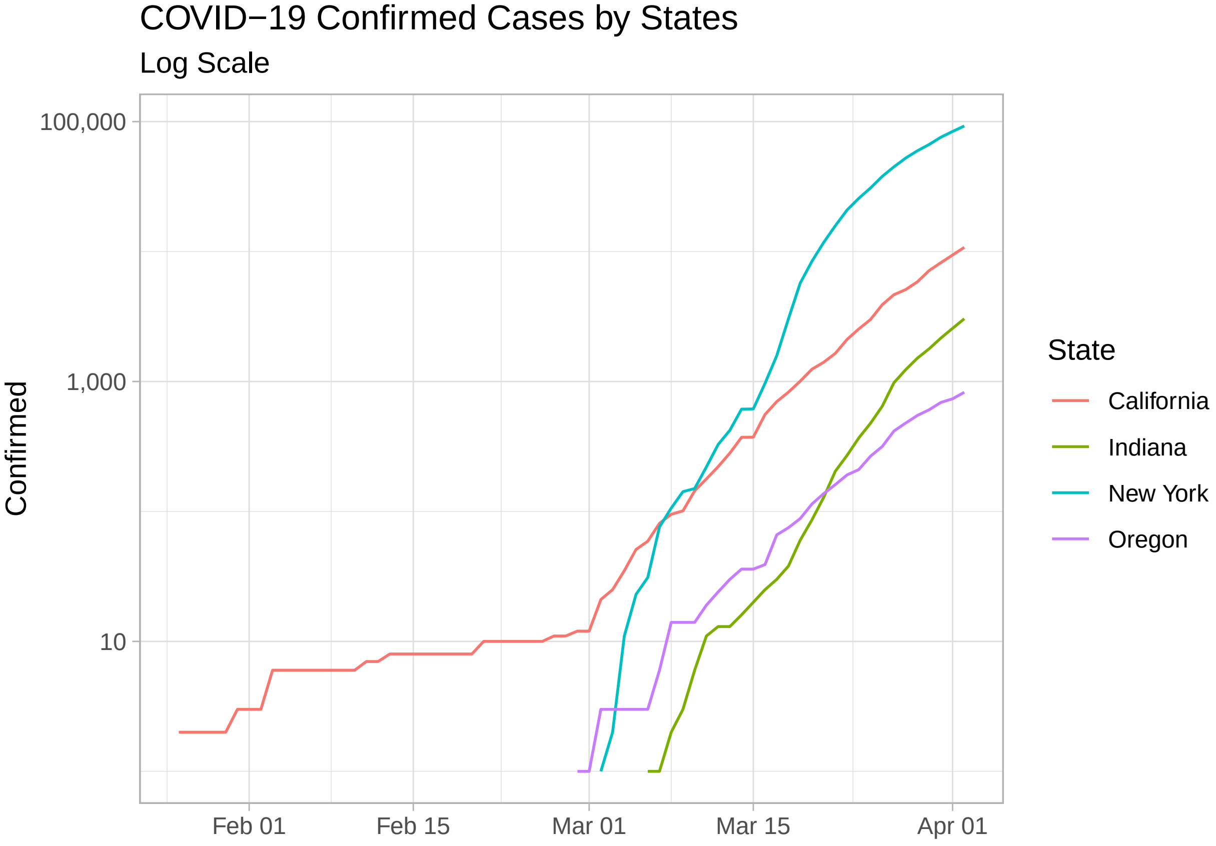
```
italy <- cvdata.i18n %>%
  filter(Country == 'Italy') %>%
  select(-State, -County, -Key, -Lat, -Long, -Country)
```

Data Exploration

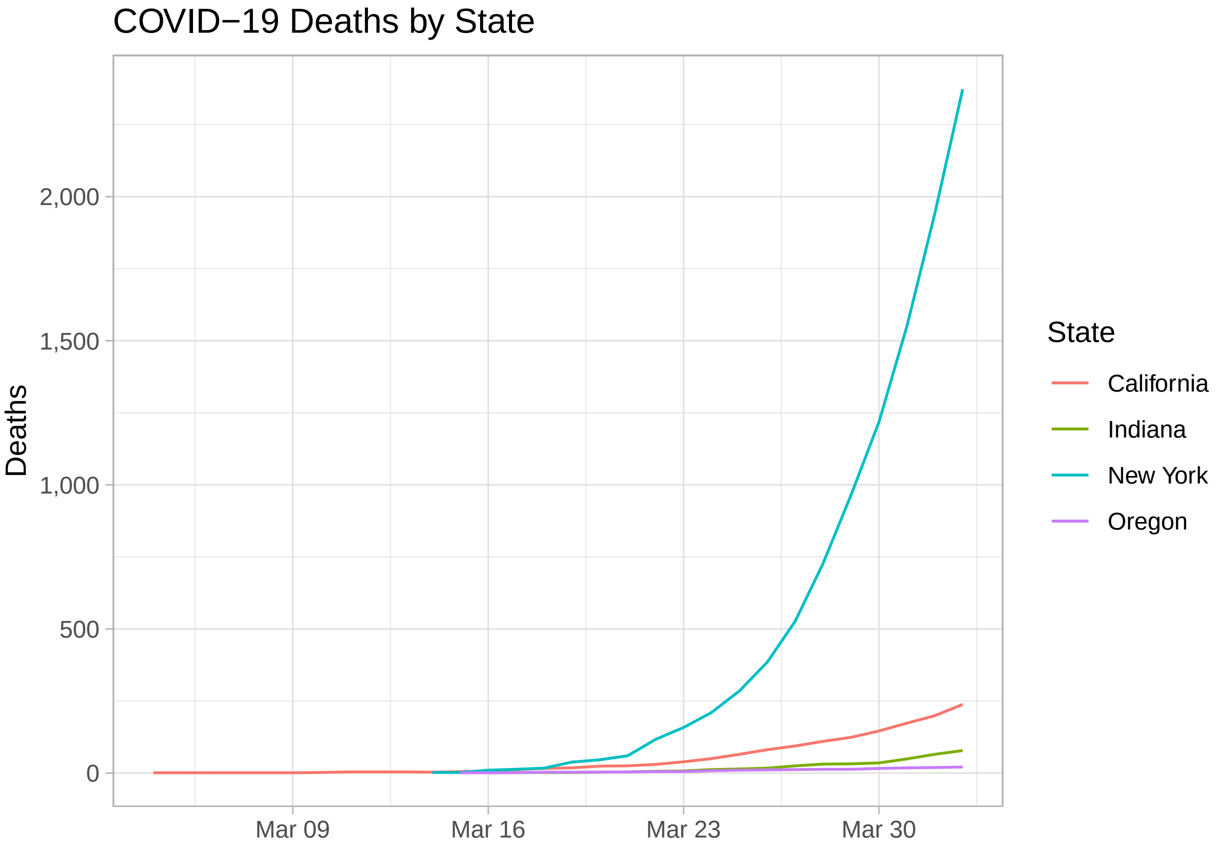
For the US Data I'm just showing a few states so the plots are more readable. This is just preliminary exploration as I get to know the data and clean it for other purposes.

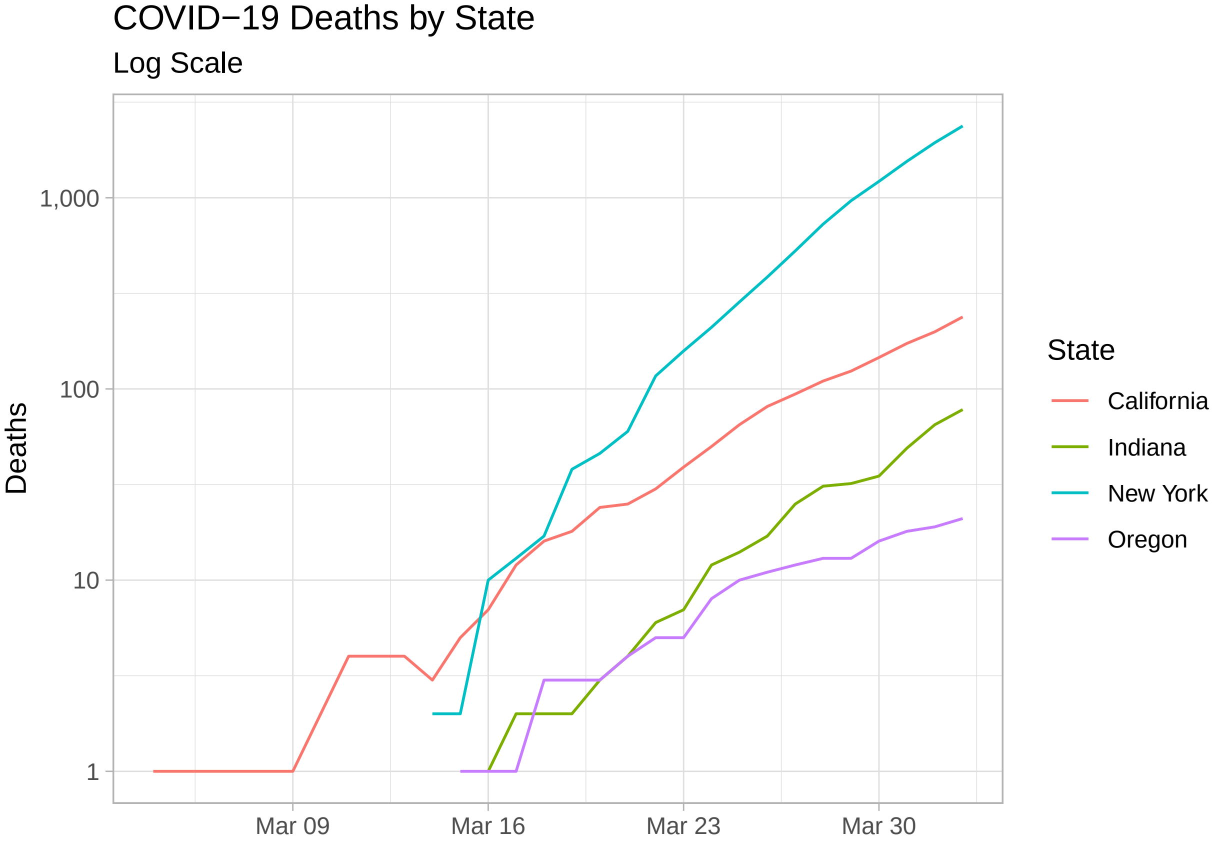
Confirmed Cases Grouped by State





Deaths, grouped by State.


COVID−19 Deaths by State

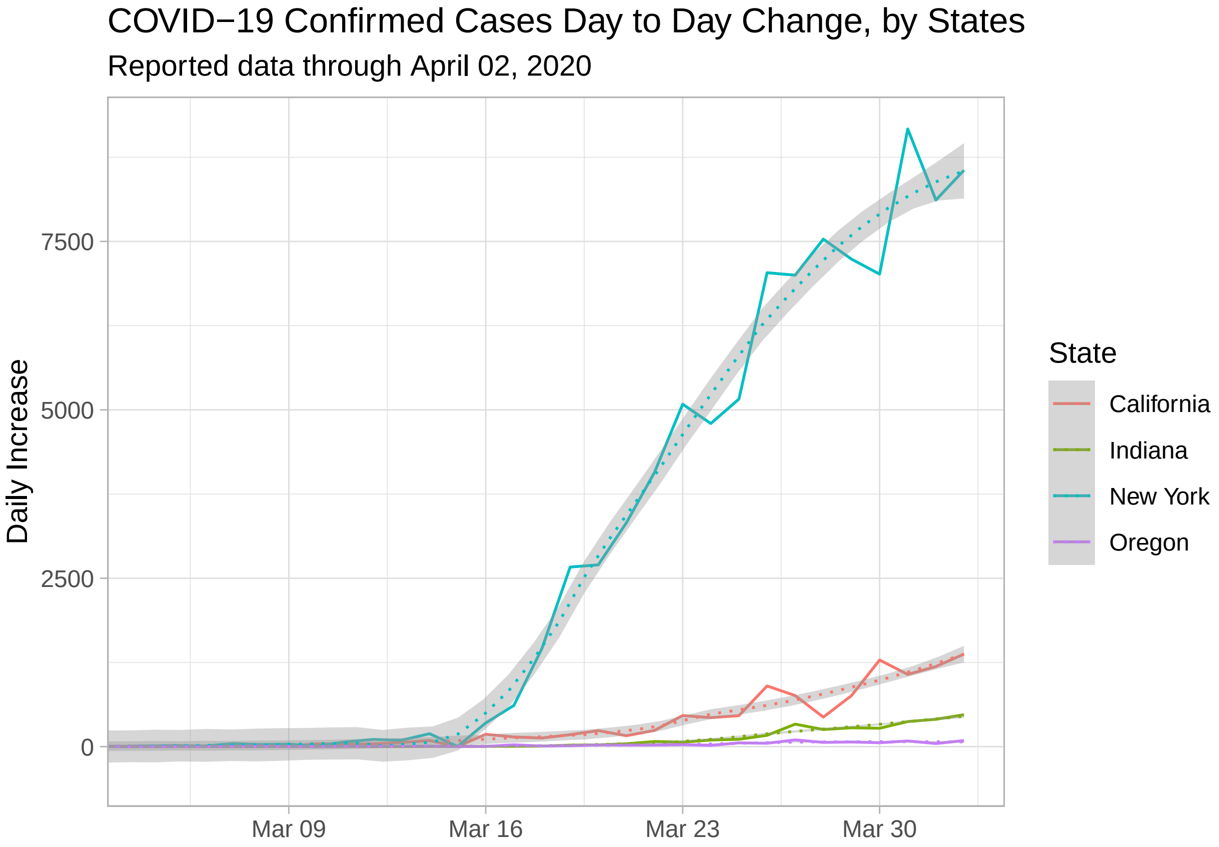

COVID−19 Deaths by State

Log Scale

Change in Confirmed Cases

This plot shows the day to day change in confirmed cases. This is important because it will tell us when we hit that peak in infections, the top of the curve that we are trying to 'flatten'. This peak occurs when the line crosses over the X axis into negative territory.

Right now you can see New York is a long way from that but the LOESS trend line shows it starting to trend downward.



COVID−19 Confirmed Cases Day to Day Change, by States

Reported data through April 02, 2020

Save the Data

```
saveRDS(cvdata.us, 'data/cvdata.us.RDS')
saveRDS(cvdata.i18n, 'data/cvdata.i18n.RDS')
```