Initial EDA - NYT_COVID_CASES

Ruby Han

03/24/2021

```
library(tidyverse)
```

```
## Warning in system("timedatectl", intern = TRUE): running command 'timedatectl'
## had status 1
```

```
library(magrittr)
library(ggplot2)
library(patchwork)
library(sandwich)
library(lmtest)
library(knitr) # kable
theme_set(theme_minimal())
knitr::opts_chunk$set(dpi = 300)

# assemble multiple plots
library(gridExtra)

# read excel format
library(readxl)

# import fread function
library(data.table)
```

```
nyt_covid_data <- fread("https://raw.githubusercontent.com/nytimes/covid-19-data/master/us-states.csv")
# jhu_covid_data <- fread("https://raw.githubusercontent.com/govex/COVID-19/master/data_tables/JHU_USCo
head(nyt_covid_data)
```

```
##          date      state fips cases deaths
## 1: 2020-01-21 Washington   53     1      0
## 2: 2020-01-22 Washington   53     1      0
## 3: 2020-01-23 Washington   53     1      0
## 4: 2020-01-24   Illinois   17     1      0
## 5: 2020-01-24 Washington   53     1      0
## 6: 2020-01-25 California    6     1      0
```

Data Issues

https://github.com/nytimes/covid-19-data/blob/master/PROBABLE-CASES-NOTE.md

- At the beginning of COVID in US, most health departments and CDC only considered acase to be ‘confirmed’ as COVID with a positive lab test result. Recently, ‘probable’ cases are reported which may affect data accuracy as NYT has changed the dataset to start including ‘probable’ cases when available. NYT is working on updating past data to include ‘probable’ cases. Thus, in some states, the data will be revised to show a higher number of cases on past dates.

```
summary(nyt_covid_data) # earliest date record 2020-01-21 and updating on a daily basis
```

```
##       date               state                fips           cases        
##  Min.   :2020-01-21   Length:21409       Min.   : 1.00   Min.   :      1  
##  1st Qu.:2020-06-08   Class :character   1st Qu.:17.00   1st Qu.:   7253  
##  Median :2020-09-13   Mode  :character   Median :31.00   Median :  53940  
##  Mean   :2020-09-12                      Mean   :31.93   Mean   : 186655  
##  3rd Qu.:2020-12-19                      3rd Qu.:46.00   3rd Qu.: 200866  
##  Max.   :2021-03-26                      Max.   :78.00   Max.   :3656693  
##      deaths     
##  Min.   :    0  
##  1st Qu.:  151  
##  Median : 1158  
##  Mean   : 4017  
##  3rd Qu.: 4606  
##  Max.   :58603  
```

```
unique(nyt_covid_data$state) # 55 states - 4 additional US territories (Guam, Northern Mariana Islands,
```

```
##  [1] "Washington"               "Illinois"                
##  [3] "California"               "Arizona"                 
##  [5] "Massachusetts"            "Wisconsin"               
##  [7] "Texas"                    "Nebraska"                
##  [9] "Utah"                     "Oregon"                  
## [11] "Florida"                  "New York"                
## [13] "Rhode Island"             "Georgia"                 
## [15] "New Hampshire"            "North Carolina"          
## [17] "New Jersey"               "Colorado"                
## [19] "Maryland"                 "Nevada"                  
## [21] "Tennessee"                "Hawaii"                  
## [23] "Indiana"                  "Kentucky"                
## [25] "Minnesota"                "Oklahoma"                
## [27] "Pennsylvania"             "South Carolina"          
## [29] "District of Columbia"     "Kansas"                  
## [31] "Missouri"                 "Vermont"                 
## [33] "Virginia"                 "Connecticut"             
## [35] "Iowa"                     "Louisiana"               
## [37] "Ohio"                     "Michigan"                
## [39] "South Dakota"             "Arkansas"                
## [41] "Delaware"                 "Mississippi"             
## [43] "New Mexico"               "North Dakota"            
## [45] "Wyoming"                  "Alaska"                  
## [47] "Maine"                    "Alabama"                 
## [49] "Idaho"                    "Montana"                 
## [51] "Puerto Rico"              "Virgin Islands"          
## [53] "Guam"                     "West Virginia"           
## [55] "Northern Mariana Islands"
```

```
                                              #Virgin Islands and Puerto Rico) compared to CUSP data
names(nyt_covid_data) # 5 columns (date, state, fips, cases, deaths)
```

```
## [1] "date"   "state"  "fips"   "cases"  "deaths"
```

```r
# convert date strings to dates
nyt_covid <- nyt_covid_data %>%
  mutate(
    date = as.Date(date)
  ) %>%
  select(
    date
    ,state
    ,cases
    ,deaths
  )

typeof(nyt_covid_data$date) # R's default date format is in integer
```

```
## [1] "integer"
```

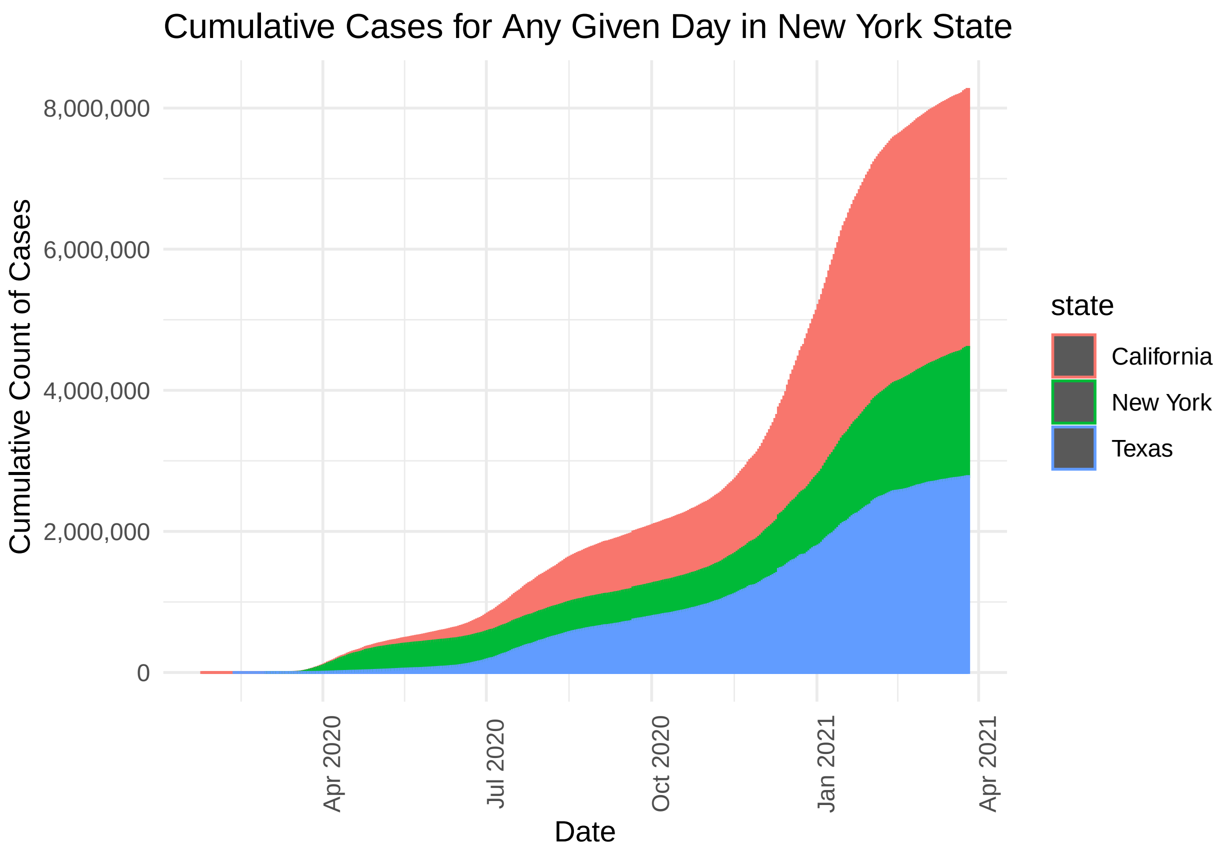
```r
typeof(nyt_covid$date)
```

```
## [1] "double"
```

```r
nyt_covid
```

```
##              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
##    ---
## 21405: 2021-03-26      Virginia 612062  10154
## 21406: 2021-03-26    Washington 362403   5288
## 21407: 2021-03-26 West Virginia 139750   2628
## 21408: 2021-03-26     Wisconsin 633154   7272
## 21409: 2021-03-26       Wyoming  56046    695
```

```r
ny_state_covid <- nyt_covid %>%
  group_by(
    state
    ) %>%
  filter(
    state == 'New York' | state == 'California' | state == 'Texas'
  ) %>%
  ggplot() +
  aes(x = date, y = cases, color=state) +
  geom_col() +
  theme(axis.text.x = element_text(angle = 90)) +
  labs(x = "Date",
       y = " Cumulative Count of Cases",
       title = "Cumulative Cases for Any Given Day in New York State") +
  scale_y_continuous(labels = scales::comma)
ny_state_covid
```



```
date_check <- nyt_covid %>%
  group_by(state) %>%
  summarise(
    data_count = n(),
    min_date = min(date),
    max_date = max(date),
    diff = data_count - (max_date- min_date + 1)
  ) %>%
  arrange(desc(min_date))

date_check
```

```
## # A tibble: 55 x 5
##    state                    data_count min_date   max_date   diff
##    <chr>                         <int> <date>     <date>     <drtn>
##  1 Northern Mariana Islands        364 2020-03-28 2021-03-26 0 days
##  2 West Virginia                   375 2020-03-17 2021-03-26 0 days
##  3 Guam                            377 2020-03-15 2021-03-26 0 days
##  4 Virgin Islands                  378 2020-03-14 2021-03-26 0 days
##  5 Alabama                         379 2020-03-13 2021-03-26 0 days
##  6 Idaho                           379 2020-03-13 2021-03-26 0 days
##  7 Montana                         379 2020-03-13 2021-03-26 0 days
##  8 Puerto Rico                     379 2020-03-13 2021-03-26 0 days
##  9 Alaska                          380 2020-03-12 2021-03-26 0 days
## 10 Maine                           380 2020-03-12 2021-03-26 0 days
## # ... with 45 more rows
```

```
# date_check %$% max(min_date)
```

NYT data

Based on our plotting using NYT COVID dataset, we obtain cumulative cases for any given day in each state. In order to obtain number of new cases per day, we will have to subtract cases from the prior row (day before) as below.

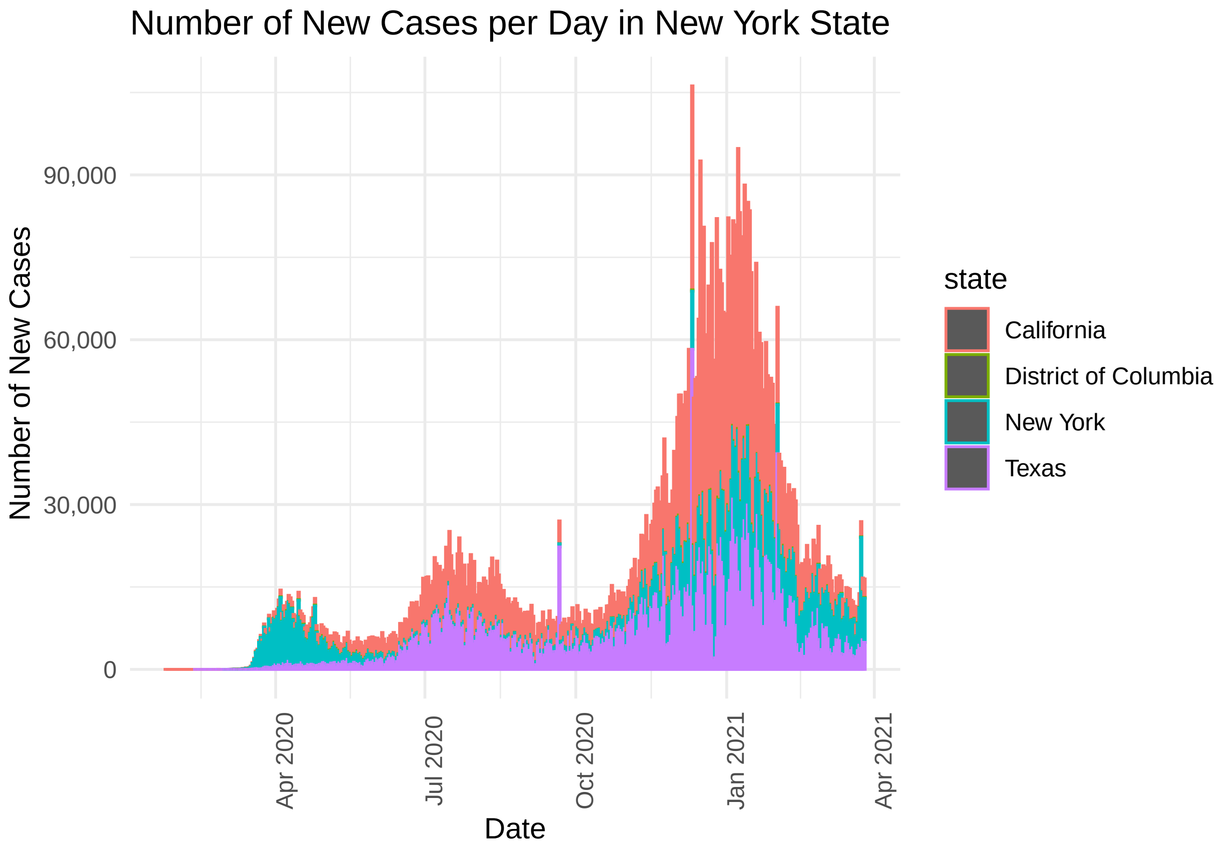
```
nyt_covid_new <- nyt_covid %>%
  group_by(
    state
    ) %>%
  mutate(
    new_cases = cases - lag(cases, default = first(cases), order_by = date)
    ,new_deaths = deaths - lag(deaths, default = first(deaths), order_by = date)
  )
summary(nyt_covid_new)
```

```
##       date               state               cases             deaths     
##  Min.   :2020-01-21   Length:21409       Min.   :      1   Min.   :    0  
##  1st Qu.:2020-06-08   Class :character   1st Qu.:   7253   1st Qu.:  151  
##  Median :2020-09-13   Mode  :character   Median :  53940   Median : 1158  
##  Mean   :2020-09-12                      Mean   : 186655   Mean   : 4017  
##  3rd Qu.:2020-12-19                      3rd Qu.: 200866   3rd Qu.: 4606  
##  Max.   :2021-03-26                      Max.   :3656693   Max.   :58603  
##    new_cases       new_deaths     
##  Min.   :-7757   Min.   :-598.00  
##  1st Qu.:   92   1st Qu.:   1.00  
##  Median :  467   Median :   7.00  
##  Mean   : 1410   Mean   :  25.58  
##  3rd Qu.: 1386   3rd Qu.:  24.00  
##  Max.   :64987   Max.   :2559.00
```

```
nyt_covid_negative <- nyt_covid_new %>%
  filter(
    new_cases < 0 | new_deaths < 0
  )
nyt_covid_negative # negative values are assumed to be corrective adjustments on previous erroneous dat
```

```
## # A tibble: 114 x 6
## # Groups:   state [38]
##    date       state       cases deaths new_cases new_deaths
##    <date>     <chr>       <int>  <int>     <int>      <int>
##  1 2020-04-01 Virginia     1511     18       262         -9
##  2 2020-04-08 Virginia     3644     60       312         -9
##  3 2020-04-12 Georgia     12103    438      -158          6
##  4 2020-04-18 Alabama      4723    147       151         -4
##  5 2020-04-19 Puerto Rico  1213     41        95        -19
##  6 2020-04-21 Puerto Rico   915     43      -337          1
##  7 2020-04-23 Alabama      5832    197       222         -4
##  8 2020-04-25 Colorado    12967    670       712         -2
##  9 2020-05-02 Alabama      7611    288       317         -1
## 10 2020-05-03 Idaho        2059     64        -2          0
## # ... with 104 more rows
```

```r
ny_state_covid_new <- nyt_covid_new %>%
  group_by(
    state
    ) %>%
  filter(
    state == 'New York' | state == 'California' | state == 'Texas' | state == 'District of Columbia'
  ) %>%
  ggplot() +
  aes(x = date, y = new_cases, color=state) +
  geom_col() +
  theme(axis.text.x = element_text(angle = 90)) +
  labs(x = "Date",
       y = " Number of New Cases",
       title = "Number of New Cases per Day in New York State") +
  scale_y_continuous(labels = scales::comma)
ny_state_covid_new
```



```r
write.csv(nyt_covid_new,
          file = paste0("~/W203_RDataHub/lab_2-rbgs/data/interim/",
                        "nyt_covid.csv"))
```