Final Project

Ri Liu (rl4508)

1/20/2022

1. Introduction

Covid-19 has been raging for more than 2 years, causing millions of people dead, billions of people infected and incalculable economic loss.

In these 2 years, there are several peaks of case surges caused by different variants of virus. I will analyze the trend of Covid-19 in the US from January 2020 to January 2022, including exploring the data in formats of daily new cases & deaths, 7-day average new cases & deaths and accumulated incidence & mortality rate. Plotting is a good way to show a long term trend and the difference across different states.

Public health intervention is a complex system, except for medicine and epidemiology research, it is also affected by politics, economics and publicity. I will test whether political leaning influence pandemic development across different states. The 2020 presidential election happened right amid the pandemic. So, it is a good opportunity to research the connection of epidemic and politics.

2. Packages

```
library(tidyverse) # data manipulation, tidying and visualization
library(cowplot) # organize plots in grid
library(car) # Levene's test for equal variance assumption
```

3. Data

3.1 Data Importing

- Covid-19 Data in the United States https://github.com/nytimes/covid-19-data
- US County Level Election Results https://github.com/tonmcg/US_County_Level_Election_Results_08-20
- US Population Data from GPH-GU 2183
- US Standard State Name Data from R `datasets`

```
us <- read_csv('us.csv')
states <- read_csv('us-states.csv')
counties <- read_csv('us-counties.csv')
population <- read_rds('us_state_populations.rds')
election <- read_csv('2020_US_County_Level_Presidential_Results.csv')
```

3.2 Data Cleaning

- `states` and `counties`

In the `states` and `counties` data set, there are 56 states or territorial governments. But not all of them have other information that is easy to find. So I only use the states that are in the `state.name` dataset and the 'District of Columbia'.

```
all_states <- append(state.name, 'District of Columbia')
#all_states
```

The districts that are temporarily ignored are as follows. This also verifies that other states' names in these data sets are in the same format.

```
except <- unique(counties$state)[!unique(counties$state) %in% all_states]
except
```

```
## [1] "Puerto Rico"              "Virgin Islands"
## [3] "Guam"                     "Northern Mariana Islands"
## [5] "American Samoa"
```

```
states <- states %>%
  inner_join(tibble(state = all_states), by = 'state') %>%
  arrange(date, state)
#head(states, n = 5)
```

```
counties <- counties %>%
  inner_join(tibble(state = all_states), by = 'state') %>%
  arrange(date, state, county)
#head(counties, n = 5)
```

Now, `states` and `counties` both have 51 states.

- Missing value

Missing value occurs in `county` column of `counties` data set when health departments report the patient's county of residence as unknown or pend determination. The editor mark these data as 'Unknown'. For the purpose of county-wise analysis, I filter out those entries with 'Unknown' `county`. Before the filtering, `counties` has 2052869 rows of entries.

```
counties <- counties %>%
  filter(county != 'Unknown')
```

Now, `counties` has 2037229 rows of entries.

To check whether a data set contains standard `NA`.

```
check_na <- function(x) {
  any(is.na(x))
}
map(list(us, states, counties), apply, MARGIN = 2, FUN = check_na)
```

There are `NA`s in the `fips` column. `fips` is a code of standard geographic identifier for states and counties.

```
unique(counties[is.na(counties$fips), c('county', 'state')])
```

```
## # A tibble: 3 x 2
##   county        state
##   <chr>         <chr>
## 1 New York City New York
## 2 Kansas City   Missouri
## 3 Joplin        Missouri
```

The NAs for 'New York City' occurred because the editor combines New York, Kings, Queens, Bronx and Richmond into 'New York City'. 'Kansas City' and 'Joplin' are separated from their origin counties. However those three new "counties" don't have their `fips` in the FIPS database. So I drop the `fips` column.

```
counties <- counties %>%
  select(!fips)
```

- `population`

In the `population` data set, there are population of each states from 1790 to 2011. Only the most recent numbers are needed.

```
population <- population[population$year == max(population$year), c('state', 'population')]
```

Check the consistency of states' names in `all_states` and `population`.

```
population$state[!unique(population$state) %in% all_states]
```

```
## [1] District Of Columbia
## 51 Levels: Alabama Alaska Arizona Arkansas California Colorado ... Wyoming
```

There is a tiny discrepancy in the spelling of 'District of Columbia'. So, I need to unify the format of 'District of Columbia'.

```
population$state <- fct_recode(population$state, 'District of Columbia' = 'District Of Columbia')
```

Check again.

```
length(!unique(population$state) %in% all_states)
```

```
## [1] 51
```

- Data Summary

Now `us` looks like:

```
head(us, 5)
```

```
## # A tibble: 5 x 3
##   date       cases deaths
##   <date>     <dbl>  <dbl>
## 1 2020-01-21     1      0
## 2 2020-01-22     1      0
## 3 2020-01-23     1      0
## 4 2020-01-24     2      0
## 5 2020-01-25     3      0
```

`states` looks like:

```
head(states, 5)
```

```
## # A tibble: 5 x 5
##   date       state      fips  cases deaths
##   <date>     <chr>      <chr> <dbl>  <dbl>
## 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
```

`counties` looks like:

```
head(counties,5)
```

```
## # A tibble: 5 x 5
##   date       county    state      cases deaths
##   <date>     <chr>     <chr>      <dbl>  <dbl>
## 1 2020-01-21 Snohomish Washington     1      0
## 2 2020-01-22 Snohomish Washington     1      0
## 3 2020-01-23 Snohomish Washington     1      0
## 4 2020-01-24 Cook      Illinois       1      0
## 5 2020-01-24 Snohomish Washington     1      0
```

`population` looks like:

```
head(population, 5)
```

```
##          state population
## 192    Alabama    4808274
## 244     Alaska     720004
## 336    Arizona    6509700
## 508   Arkansas    2935019
## 670 California   37607525
```

```
map(list(us = us, states = states, counties = counties), summary)
```

The `us`, `states` and `conties` now have data from 2020-01-21 to 2022-01-11.

`states` reaches its highest cases to 6.447667×10^6 and highest deaths to 7.7447×10^4. `counties` reaches its high cases to 2.046912×10^6 and highest deaths to 3.6089×10^4.

`election` looks like:

```
head(election, 5)
```

```
## # A tibble: 5 x 10
##   state_name county_fips county_name    votes_gop votes_dem total_votes  diff
##   <chr>      <chr>       <chr>              <dbl>     <dbl>       <dbl> <dbl>
## 1 Alabama    01001       Autauga County     19838      7503       27770 12335
## 2 Alabama    01003       Baldwin County     83544     24578      109679 58966
## 3 Alabama    01005       Barbour County      5622      4816       10518   806
## 4 Alabama    01007       Bibb County         7525      1986        9595  5539
## 5 Alabama    01009       Blount County      24711      2640       27588 22071
## # ... with 3 more variables: per_gop <dbl>, per_dem <dbl>, per_point_diff <dbl>
```

`election` has records of 51 different states, and 1887 different counties across the country. It records the votes count, votes difference and votes percent of each party in each county.

4. Exploratory Data Analysis

4.1 Feature Engineering

1. Transform accumulated cases and deaths into daily new cases and new deaths.

For `us`:

```
new_case <- c(us$cases[1]) # result vector for case
new_death <- c(us$deaths[1]) # result vector for death
for (i in 2:nrow(us)){
  new_case = append(new_case, us$cases[i] - us$cases[i-1])
  new_death = append(new_death, us$deaths[i] - us$deaths[i-1])
}
us <- add_column(us, new_case, new_death)
```

For `states`:

```
states <- add_column(states, new_case = c(0)) # result vector for case
states <- add_column(states, new_death = c(0)) # result vector for death
for (state in all_states){
  index = states$state == state
  states$new_case[index][1] = states$cases[index][1]
  states$new_death[index][1] = states$deaths[index][1]
  for (i in 2:nrow(states[index, ])){
    states$new_case[index][i] = states$cases[index][i] - states$cases[index][i-1]
    states$new_death[index][i] = states$deaths[index][i] - states$deaths[index][i-1]
  }
}
```

2. 7-day Average

As each day's test ability is not all on the same level, for example, affected by holidays. Calculating 7-day average number can reduce the fluctuation to some extent.

For `us`:

```
avg7_case <- c(us$cases[1:6]) # result vector for case
avg7_death <- c(us$deaths[1:6]) # result vector for death
for (i in 7:nrow(us)){
  left <- i-6 # the first day of the seven days in a row
  avg7_case = append(avg7_case, mean(us$new_case[left:i]))
  avg7_death = append(avg7_death, mean(us$new_death[left:i]))
}
us <- add_column(us, avg7_case, avg7_death)
```

For `states`:

```
states <- add_column(states, avg7_case = c(0)) # result vector for case
states <- add_column(states, avg7_death = c(0)) # result vector for death
for (state in all_states){
  index = states$state == state # filter out the state
  states$avg7_case[index][1:6] = states$new_case[index][1:6]
  states$avg7_death[index][1:6] = states$new_death[index][1:6]
  for (i in 7:nrow(states[index, ])){
    left <- i-6 # the first day of the seven days in a row
    states$avg7_case[index][i] = mean(states$new_case[index][left:i])
    states$avg7_death[index][i] = mean(states$new_death[index][left:i])
  }
}
```

3. Incidence and Mortality Rate

Transform accumulated cases and deaths into incidence and mortality rate given corresponding population.

For `us`:

```
total_population <- sum(population$population)
us <- us %>%
  mutate(incidence = cases / total_population * 1e6,
         mortality = deaths / total_population * 1e6)
```

For `states`

```
states <- states %>%
  inner_join(population, by = 'state') %>%
  mutate(incidence = cases / population * 1e6,
         mortality = deaths / population * 1e6)
```

Out of `states`, it is not easy to read all states' information at the same time. It is more efficient to analyse a part of the data that is more representative, which is pulling out the states with top 5 and bottom 5 incidence. However, incidence is easily affected in states with smaller population. So finally filtering out states with larger than 25% quantile population, which is 1.714363×10^6, is necessary for analysis on incidence.

```
top5 <- states %>%
  filter(date == as.Date('2022-01-11'), population > 1.7*1e6) %>%
  arrange(desc(incidence)) %>%
  .[1:5, 'state', drop=T]
top5
```

```
## [1] "Utah"           "Florida"        "Tennessee"      "South Carolina"
## [5] "Arizona"
```

```
bottom5 <- states %>%
  filter(date == as.Date('2022-01-11'), population > 1.7e6) %>%
  arrange(incidence) %>%
  .[1:5, 'state', drop=T]
bottom5
```

```
## [1] "Oregon"      "Maryland"    "Washington"  "Virginia"    "Connecticut"
```

Now, `us` looks like

```
head(us, 5)
```

```
## # A tibble: 5 x 9
##   date       cases deaths new_case new_death avg7_case avg7_death incidence
##   <date>     <dbl>  <dbl>    <dbl>     <dbl>     <dbl>      <dbl>     <dbl>
## 1 2020-01-21     1      0        1         0         1          0   0.00321
## 2 2020-01-22     1      0        0         0         1          0   0.00321
## 3 2020-01-23     1      0        0         0         1          0   0.00321
## 4 2020-01-24     2      0        1         0         2          0   0.00642
## 5 2020-01-25     3      0        1         0         3          0   0.00964
## # ... with 1 more variable: mortality <dbl>
```

`states` looks like

```
head(states, 5)
```

```
## # A tibble: 5 x 12
##   date       state      fips  cases deaths new_case new_death avg7_case avg7_death
##   <date>     <chr>      <chr> <dbl>  <dbl>    <dbl>     <dbl>     <dbl>      <dbl>
## 1 2020-01-21 Washington 53        1      0        1         0         1          0
## 2 2020-01-22 Washington 53        1      0        0         0         0          0
## 3 2020-01-23 Washington 53        1      0        0         0         0          0
## 4 2020-01-24 Illinois   17        1      0        1         0         1          0
## 5 2020-01-24 Washington 53        1      0        0         0         0          0
## # ... with 3 more variables: population <dbl>, incidence <dbl>, mortality <dbl>
```

4. 'Election' Data

The `election` data is collected on county-level. I need to summarize it to state-level, and mark the final vote for party of each state as a factor.

```
election <- election %>%
  group_by(state_name) %>%
  summarise(sum_gop = sum(votes_gop), sum_dem = sum(votes_dem)) %>%
  mutate(per_gop = sum_gop / (sum_gop+sum_dem), per_dem = sum_dem / (sum_gop+sum_dem)) %>%
  mutate(diff = round((per_gop - per_dem) * 100, 2)) %>%
  mutate(party = factor(ifelse(diff > 0, 'gop', 'dem'))) %>%
  mutate(diff = abs(diff))
```

Now, `election` looks like:

```
head(election, 5)
```

```
## # A tibble: 5 x 7
##   state_name sum_gop  sum_dem per_gop per_dem  diff party
##   <chr>        <dbl>    <dbl>   <dbl>   <dbl> <dbl> <fct>
## 1 Alabama    1441168   849648   0.629   0.371 25.8  gop
## 2 Alaska      189892   153405   0.553   0.447 10.6  gop
## 3 Arizona    1661686  1672143   0.498   0.502  0.31 dem
## 4 Arkansas    760647   423932   0.642   0.358 28.4  gop
## 5 California 6005961 11109764   0.351   0.649 29.8  dem
```

If the vote difference is not significantly different, the political leaning in that state may not have a uniform characteristic. For my analysis, I filter out the states whose vote difference is larger than 25% quantile of all of the vote difference across the country, and pull out the states who vote for Republican as `gop` and who vote for Democratic as `dem`.

```
quantile(election$diff, seq(0,0.5,0.05))
```

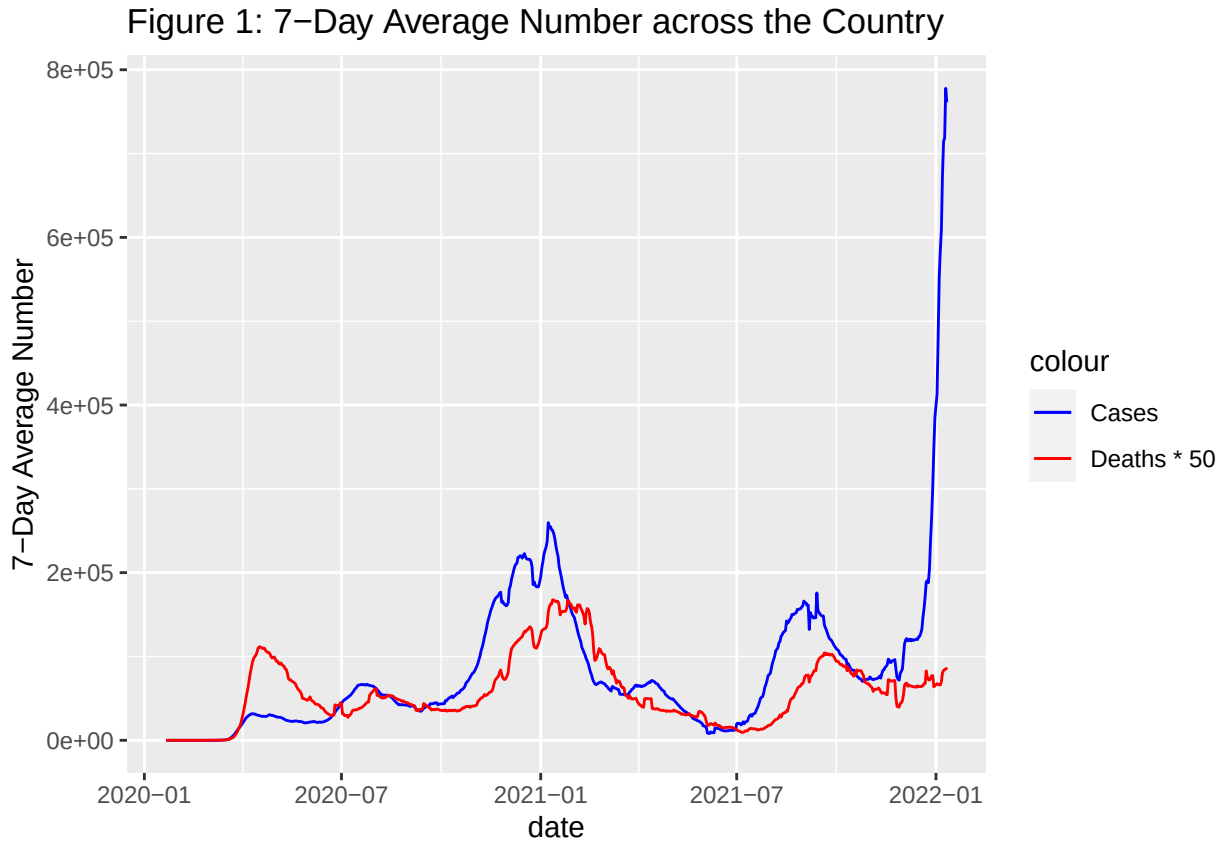
```
##     0%     5%    10%    15%    20%    25%    30%    35%    40%    45%    50%
##  0.240  0.915  2.450  4.525  7.500  8.650 10.630 12.865 15.670 16.505 16.790
```

```
election <- election %>%
  select(state_name, diff, party) %>%
  filter(diff > 8.650)
gop <- election %>%
  filter(diff > 8.650, party == 'gop') %>%
  .$state_name
dem <- election %>%
  filter(diff > 8.650, party == 'dem') %>%
  .$state_name
```

4.2 Trending Analysis

Now, I can examine the trend of COVID-19 on country-level. The scales of cases and deaths are not on the same level, so I multiply deaths by 50 to make a more readible plot.

```
us %>%
  ggplot() +
  geom_line(aes(x = date, y = avg7_case, color = 'Cases')) +
  geom_line(aes(x = date, y = avg7_death*50, color = 'Deaths * 50')) +
  # customize the legend
  scale_color_manual(values = c('Cases' = 'blue', 'Deaths * 50' = 'red')) +
  ylab('7-Day Average Number') +
  ggtitle('Figure 1: 7-Day Average Number across the Country')
```



We can see that from January 2020 to January 2022, there are three major peaks in case number, which are in 2020's winter, 2021's fall and 2021's winter. And there are three major peaks in death number, two of them come with the surges of cases, but the first peak occurs at the early stage when COVID-19 was discovered, which is more severe than the case number. The death surges are mostly later than the case surges, which is easy to understand.

However, out of the outbreak in 2021's winter, even though the case number break through an exceedingly high level compared with previous surges, the death number in this outbreak does not show a corresponding volume of surge. It could be because that the variant virus in this outbreak is less lethal than previous variants. But it could also be because that it is too early to see a surge of death in the development of this outbreak.

Now, let's look at the trend on state-level. As the scale of the outbreak in 2021' winter is too large, I split the plot into two part.

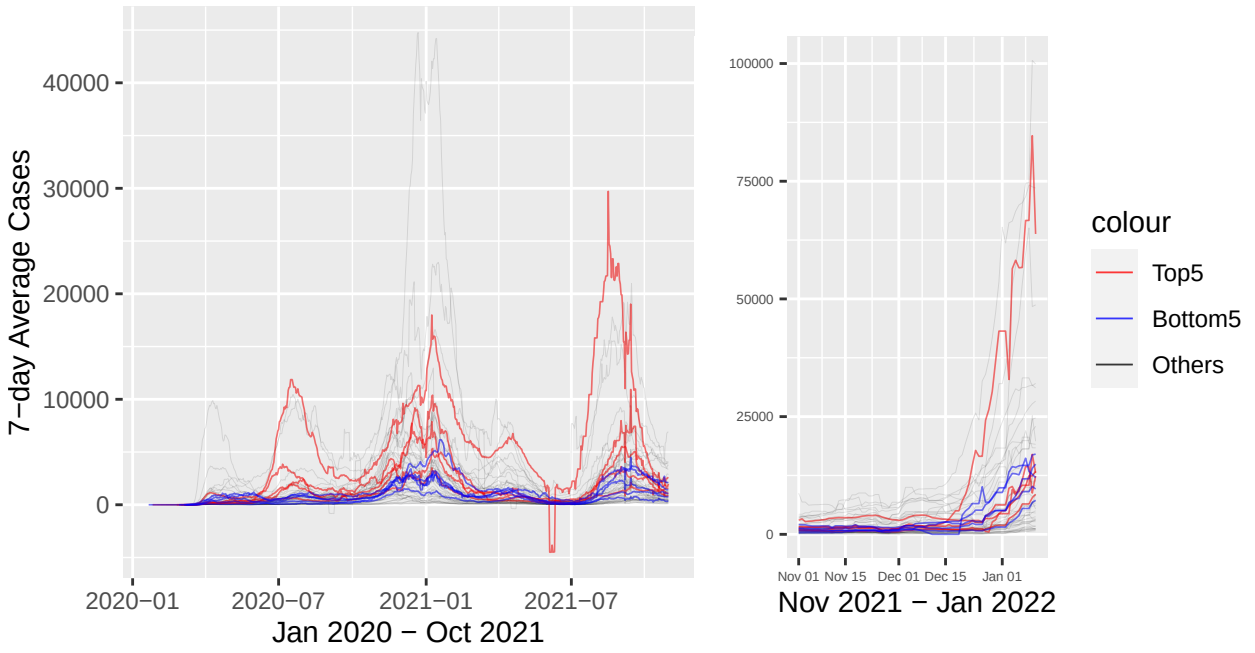
```
# before 2021-10-31
g1 <- ggplot(filter(states, date<as.Date('2021-10-31')), aes(x = date, y = avg7_case, group = state)) +
  geom_line(size = 0.1, alpha = 0.1) +
  geom_line(data = filter(states, state %in% top5, date<as.Date('2021-10-31')),
            aes(date, avg7_case, group = state, color = 'Top5'),
            color = 'red', size = 0.3, alpha = 0.5) +
  geom_line(data = filter(states, state %in% bottom5, date<as.Date('2021-10-31')),
            aes(date, avg7_case, group = state, color = 'Bottom5'),
            color = 'blue', size = 0.3, alpha = 0.5) +
  xlab('Jan 2020 - Oct 2021') +
  ylab('7-day Average Cases') +
  # adjust the margin between the plots
  theme(aspect.ratio = 1, plot.margin = unit(c(0,0,0,0), 'cm'))
```

```
# after 2021-10-31
g2 <- ggplot(filter(states, date>as.Date('2021-10-31')), aes(x = date, y = avg7_case, group = state, co
  geom_line(size = 0.1, alpha = 0.1) +
  geom_line(data = filter(states, state %in% top5, date>as.Date('2021-10-31')),
            aes(date, avg7_case, group = state, color = 'Top5'),
            size = 0.3, alpha = 0.5) +
  geom_line(data = filter(states, state %in% bottom5, date>as.Date('2021-10-31')),
            aes(date, avg7_case, group = state, color = 'Bottom5'),
            size = 0.3, alpha = 0.5) +
  xlab('Nov 2021 - Jan 2022') +
  ylab(label = '') +
  theme(axis.text = element_text(size = 5)) +
  theme(aspect.ratio = 2, plot.margin = unit(c(0,0,0,0), 'cm')) +
  # customize the legend
  scale_color_manual(values = c('Top5' = 'red', 'Bottom5' = 'blue', 'Others' = 'black'))

# make an overall title
title <- ggdraw() + draw_label("Figure 2: 7-Day Average Cases by States", fontface='bold')

# organize the two plot in a grid
p <- plot_grid(plot_grid(g1), plot_grid(g2), rel_widths = c(0.55, 0.45))
plot_grid(title, p, ncol = 1, rel_heights = c(0.1,1))
```

Figure 2: 7−Day Average Cases by States



In these two plot, we can see that the states that have top 5 incidence do have higher 7-day average case number in the left plot, which is between January 2020 to October 2021. However, out of the left plot, we cannot say the same thing. The states that have better defense to the pandemic, relatively speaking, don't

have significantly different performance in the outbreak of 2021's winter.

4.3 Political Analysis

To research whether states with different leaning towards parties have different epidemiology results, I make a T-test.

First, pull out the most recent incidence data and join the presidential election vote result on it.

```
plt_states <- states %>%
   filter(date == as.Date('2022-01-11')) %>%
   right_join(election, by = c('state' = 'state_name')) %>%
   select(date, state, incidence, party)
```

Second, test the equal variance assumption.

```
leveneTest(plt_states$incidence~plt_states$party, center = mean)
```

```
## Levene's Test for Homogeneity of Variance (center = mean)
##       Df F value   Pr(>F)
## group  1  10.823 0.002248 **
##       36
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

It shows that the variance of average incidence in different states grouped by vote result are significantly different.

Next, do a T-test.

H_0 : The mean incidence in states with the two different election vote result are the same. Setting the confidence level as 0.95.

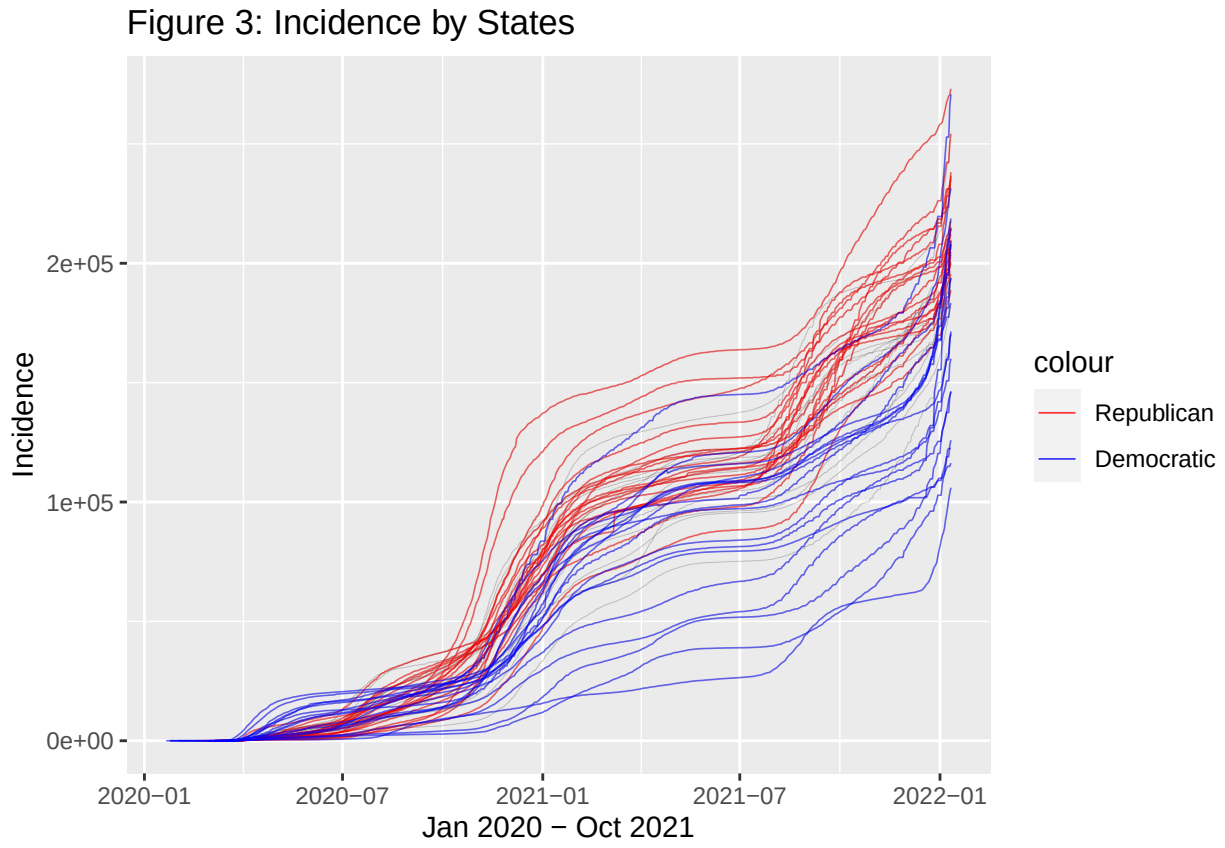
```
t.test(plt_states$incidence~plt_states$party, var.equal = F)
```

```
##
##  Welch Two Sample t-test
##
## data:  plt_states$incidence by plt_states$party
## t = -3.5203, df = 23.559, p-value = 0.001787
## alternative hypothesis: true difference in means between group dem and group gop is not equal to 0
## 95 percent confidence interval:
##  -64548.02 -16805.19
## sample estimates:
## mean in group dem mean in group gop
##          177148.3          217824.9
```

The p-value is far more less than 0.05. So the H_0 is rejected. I can conclude that there is difference in means of incidence between states who vote for Republican and states who vote for Democratic.

In the following plot, we can also have an intuitive look that states who vote for Republican have higher incidence than states who vote for Democratic.

```
ggplot(states, aes(x = date, y = incidence, group = state)) +
  geom_line(size = 0.1, alpha = 0.2) +
  geom_line(data = filter(states, state %in% gop),
            aes(date, incidence, group = state, color = 'Republican'),
            size = 0.3, alpha = 0.5) +
  geom_line(data = filter(states, state %in% dem),
            aes(date, incidence, group = state, color = 'Democratic'),
            size = 0.3, alpha = 0.5) +
  xlab('Jan 2020 - Oct 2021') +
  ylab('Incidence') +
  ggtitle('Figure 3: Incidence by States') +
  scale_color_manual(values = c('Republican' = 'red', 'Democratic' = 'blue'))
```



5. Summary

This report analyze the trend of Covid-19 in the US both on country-level and state-level. With the plots, we can obviously see that there are three major peaks of cases from January 2020 to January 2022. The deaths trend basically accord with the cases trend, but with a little bit of lag. The states that have lower incidence before the outbreak in 2021's winter don't continue to have lower daily new cases in this new outbreak comparing with states that have higher incidence.

The control of epidemic is affected by multiple factors. Given that the presidential election can reflect a political leaning of people in their region, the t-test on Covid-19 cases between states who vote for different

parties is statistically significant. But the deeper reasons and the difference of public health policies between different states need further study.

The data ends on 2022-01-11, on which the newly outbreak is still raging. This outbreak is different with previous outbreaks with exceedingly high infection speed, but the surge of deaths in this outbreak is relatively mild up to now. The further study on this outbreak and this variant virus is really important to get a better sense of the virus and to suppress the pandemic.