COVID19cases

5/17/2022

Introduction

This data-set records COVID-19 statistics since January 2020 in both the United States and global, operated by the Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE). The data-set gets updated daily.

```
##load library
library(tidyverse)
library(lubridate)
## Get current data in the four files
url_in <- "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covi
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) ##concatenated the url for 4 files
```

```
global_cases <-read_csv(urls[1])
global_deaths <- read_csv(urls[2])
us_cases <- read_csv(urls[3])
us_deaths<-read_csv(urls[4])
```

Tidy data

Global data

Tidy gobal__cases column

```
global_cases<-global_cases %>%
  pivot_longer(cols=-c(`Province/State`,
                      `Country/Region`,
                      Lat, Long),
               names_to="date",
               values_to="cases") %>%
  mutate(date=mdy(date)) %>% ##force to interpret as mdy
  select(-c(Lat,Long))
```

Tidy global_deaths column

```
global_deaths<-global_deaths %>%
  pivot_longer(cols=-c(`Province/State`,
                        `Country/Region`,
                        Lat, Long),
               names_to="date",
               values_to="deaths") %>%
  mutate(date=mdy(date)) %>% ##force to interpret as mdy
  select(-c(Lat,Long))
```

Concatenated global_cases and global_deaths

```
global <- global_cases %>%
  full_join(global_deaths) %>%
  rename(Country_Region = `Country/Region`,
         Province_State =`Province/State`)
```

US data

Tidy US_cases column

```
us_cases <- us_cases %>%
  pivot_longer(cols=-(UID:Combined_Key),
               names_to="date",
               values_to="cases") %>%
  select(Admin2:cases) %>%
  mutate(date=mdy(date)) %>% ##force to interpret as mdy
  select(-c(Lat,Long_))
```

Tidy US_deaths column

```
us_deaths<-us_deaths %>%
  pivot_longer(cols = -(UID:Population),
               names_to="date",
               values_to="deaths") %>%
  select(Admin2:deaths) %>%
  mutate(date=mdy(date))%>%
  select(-c(Lat,Long_))
```

Concatenated us_cases and us_deaths

```
us <- us_cases %>%
  full_join(us_deaths)
```

Create new column

```r
## Create new column called Combine_Key so that Gobal data and US data have the same variables
global<-global %>%
  unite("Combine_Key",
        c(Province_State, Country_Region),
        sep=",",
        na.rm=TRUE,
        remove=FALSE)
```

Look up URL for for global population data

```r
uid_look_up <- "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/UID_
uid <- read_csv(uid_look_up) %>%
  select(-c(Lat,Long_, Combined_Key, code3, iso2, iso3, Admin2))
```

Join dataset so global dataset has population column

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

Data Visualization

US__by__state

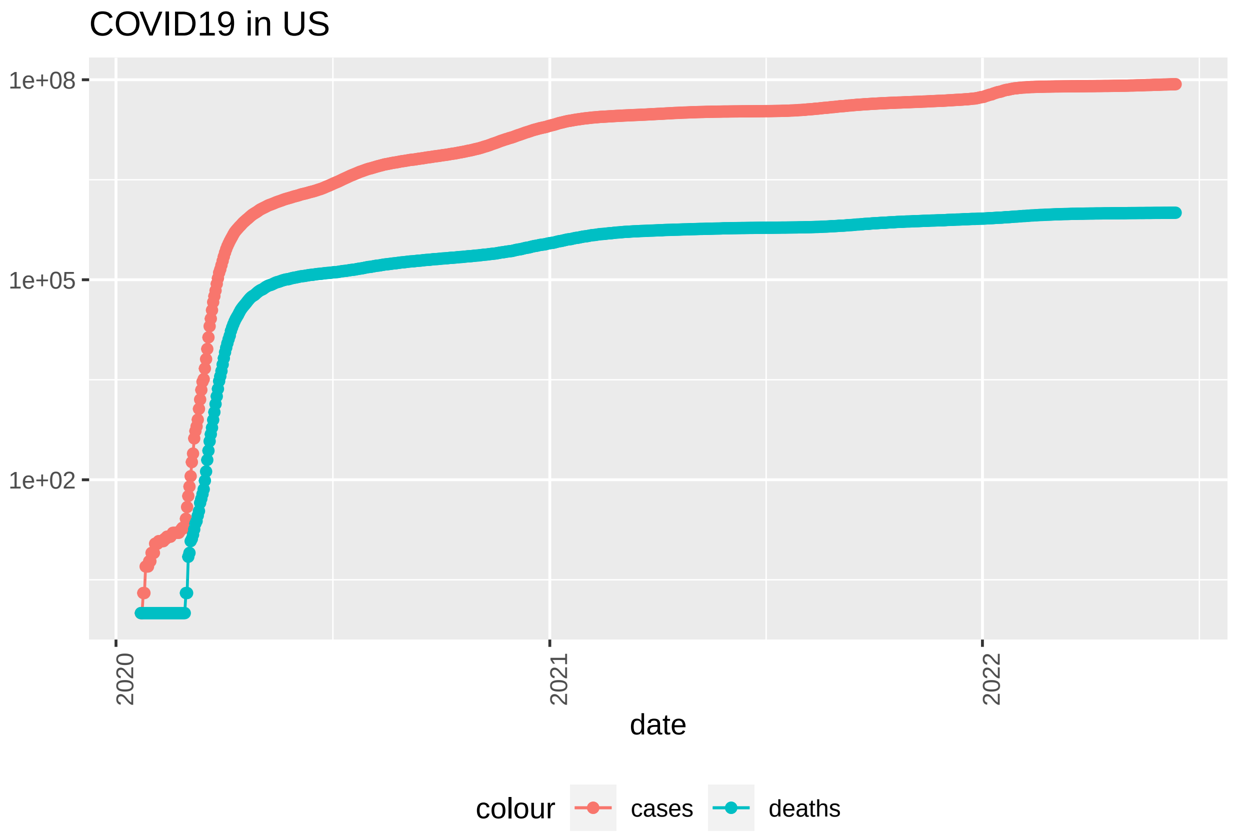
```r
us_by_state <- us %>%
  group_by(Province_State, Country_Region, date) %>%
  summarize(cases=sum(cases), deaths=sum(deaths), Population=sum(Population)) %>%
  mutate(death_per_mill=deaths*1000000/ Population) %>%
  select(Province_State, Country_Region, date, cases, deaths, death_per_mill, Population)%>%
  ungroup()
```

US__totals

```r
us_totals <- us_by_state %>%
group_by(Country_Region, date) %>%
  summarize(cases=sum(cases), deaths=sum(deaths), Population=sum(Population)) %>%
  mutate(death_per_mill=deaths*1000000/ Population) %>%
  select(Country_Region, date, cases, deaths, death_per_mill, Population)%>%
  ungroup()
```

Visualize total us_cases

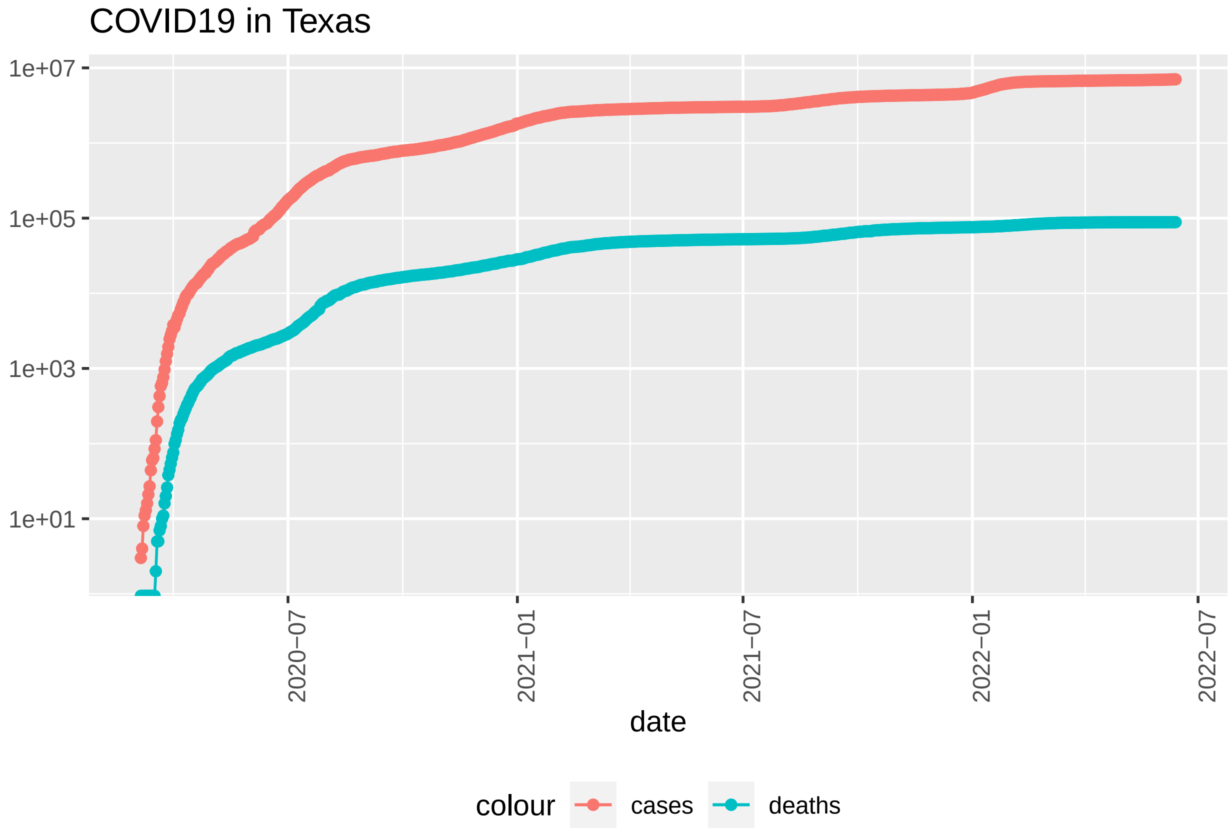
```
us_totals %>%
  filter(cases > 0) %>%
  ggplot(aes(x=date, y=cases)) +
  geom_line(aes(color= "cases")) +
  geom_point(aes(color="cases")) +
  geom_line(aes(y=deaths, color="deaths")) +
  geom_point(aes(y=deaths, color="deaths"))+
  scale_y_log10() +
  theme(legend.position="bottom",
       axis.text.x=element_text(angle = 90))+
  labs(title= "COVID19 in US", y=NULL)
```



Visualize cases in Texas

```
state <- "Texas"
us_by_state %>%
  filter((Province_State == state)) %>%
  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("COVID19 in ", state), y=NULL)
```

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



Data Analysis

Top 10 states with lowest deaths per thousand

```
us_state_totals<- us_by_state %>%
  group_by(Province_State) %>%
  summarize(deaths = max(deaths), cases=max(cases),
            population=max(Population),
            cases_per_thou=1000*cases/population,
            deaths_per_thou=1000*deaths/population) %>%
  filter(cases>0, population >0)
us_state_totals %>%
slice_min(deaths_per_thou, n=10)
```

```
## # A tibble: 10 x 6
##    Province_State        deaths  cases population cases_per_thou deaths_per_thou
##    <chr>                  <dbl>  <dbl>      <dbl>          <dbl>           <dbl>
##  1 American Samoa            31 6.19e3      55641           111.           0.557
##  2 Northern Mariana Isl~     34 1.15e4      55144           208.           0.617
##  3 Hawaii                  1465 2.91e5    1415872           205.           1.03
##  4 Virgin Islands           115 2.05e4     107268           191.           1.07
##  5 Vermont                  673 1.34e5     623989           214.           1.08
##  6 Puerto Rico             4440 7.19e5    3754939           191.           1.18
##  7 Utah                    4793 9.62e5    3205958           300.           1.50
##  8 Washington             13056 1.62e6    7614893           212.           1.71
##  9 Alaska                  1286 2.65e5     740995           357.           1.74
## 10 Maine                   2403 2.66e5    1344212           198.           1.79
```

Top 10 states with highest deaths per thousand

```
us_state_totals<- us_by_state %>%
  group_by(Province_State) %>%
  summarize(deaths = max(deaths), cases=max(cases),
            population=max(Population),
            cases_per_thou=1000*cases/population,
            deaths_per_thou=1000*deaths/population) %>%
  filter(cases>0, population >0)
us_state_totals %>%
slice_max(deaths_per_thou, n=10)
```

```
## # A tibble: 10 x 6
##    Province_State deaths   cases population cases_per_thou deaths_per_thou
##    <chr>           <dbl>   <dbl>      <dbl>          <dbl>           <dbl>
##  1 Mississippi     12481  816572    2976149           274.            4.19
##  2 Arizona         30372 2077346    7278717           285.            4.17
##  3 Oklahoma        16145 1058297    3956971           267.            4.08
##  4 Alabama         19695 1328321    4903185           271.            4.02
##  5 West Virginia    7001  523367    1792147           292.            3.91
##  6 Tennessee       26510 2080690    6829174           305.            3.88
##  7 Arkansas        11526  850535    3017804           282.            3.82
##  8 New Jersey      33859 2438510    8882190           275.            3.81
##  9 New Mexico       7873  547351    2096829           261.            3.75
## 10 Louisiana       17361 1206020    4648794           259.            3.73
```

We can see that American Samoa has the lowest deaths per thousand, and Mississippi has the most deaths per thousand. In fact, top 10 states with the lowest deaths per thousand also have the smallest population, and top 10 states with the highest deaths per thousand have the largest amount of people.

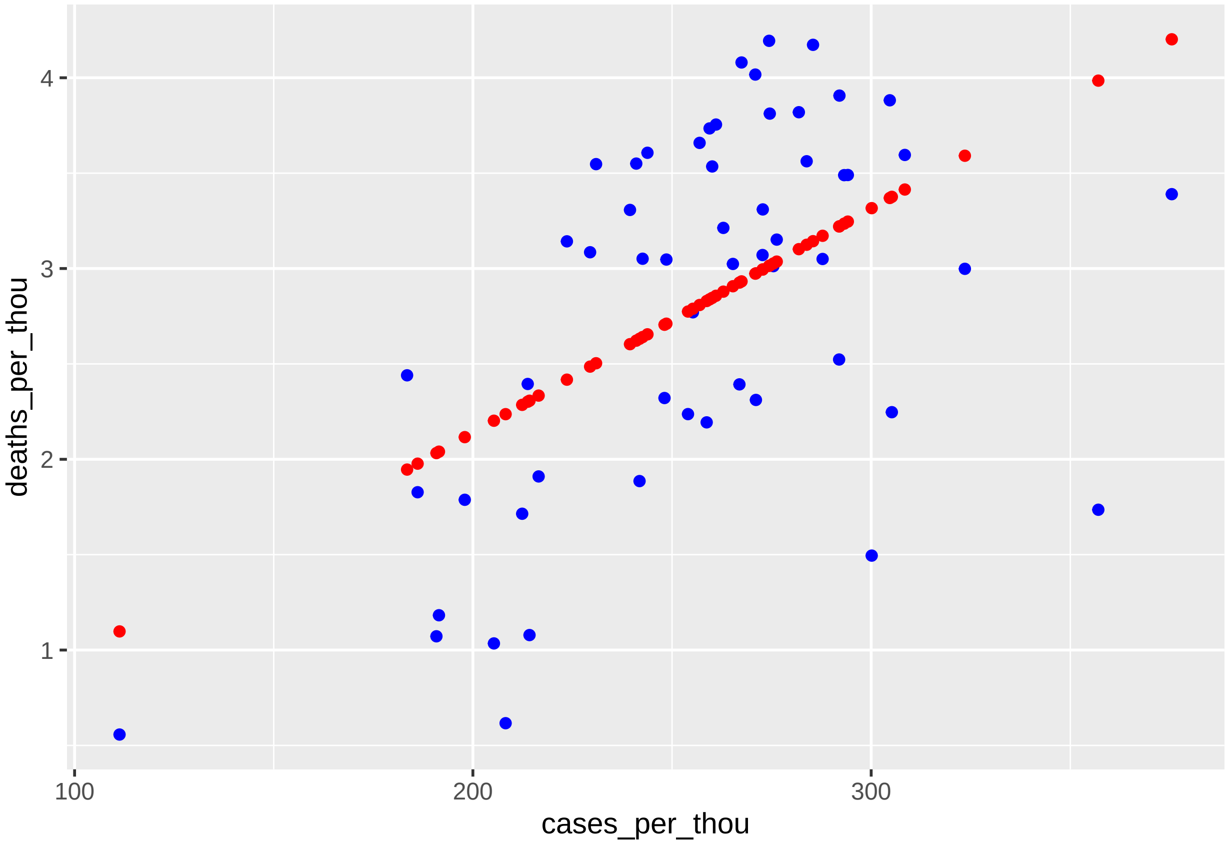
Data Modelling

```
mod <- lm(deaths_per_thou ~cases_per_thou, data= us_state_totals)
summary(mod)
```

```
##
## Call:
## lm(formula = deaths_per_thou ~ cases_per_thou, data = us_state_totals)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.2494 -0.5764  0.1161  0.6926  1.1799
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -0.209713   0.647702  -0.324    0.747
## cases_per_thou  0.011749   0.002489   4.720 1.71e-05 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.8196 on 54 degrees of freedom
## Multiple R-squared:  0.2921, Adjusted R-squared:  0.279
## F-statistic: 22.28 on 1 and 54 DF,  p-value: 1.713e-05
```

```
us_tot_w_pred <- us_state_totals %>% mutate(pred=predict(mod))
```

```
us_tot_w_pred %>% ggplot() +
  geom_point(aes(x= cases_per_thou, y=deaths_per_thou), color= "blue") +
  geom_point(aes(x=cases_per_thou, y=pred), color= "red")
```



Conclusion

From the model and scatter plot, it is shown that there is a positive relationship between cases and deaths. The p-values is 1.675e-05 for the relationship which indicate they are statistically significant. There is clear indication that cases are indication for deaths where the actual cases and deaths follows the predicted model, though the actual model are more scattered than the predicted models. Other factors need to be taken into account to explained why the actual model is more scattered; for example, with the arising COVID vaccine, people recover better which can decreases the deaths rate. ## Bias For my data visualization, I chose Texas to analyze as this is where I live, which I thought would be interested to see the statisical modelling for this state.